Elettronica e
telecomunicazioni
Anno LXIX - Numero 1/2020

NUMERO MONOGRAFICO

INTELLIGENZA ARTIFICIALE e MONDO RADIOTELEVISIVO

Rai CRITS

Elettronica e telecomunicazioni

La rivista è disponibile su web
alla URL www.crit.rai.it/eletel/

Indice

Anno LXIX
N° 1/2020
Dicembre 2020

RIVISTA PERIODICA A CURA DELLA RAI

Direttore responsabile
Gino Alberico

Redazione
Gemma Bonino
Alberto Ciprian
Roberto Del Pero

Rai Centro Ricerche, Innovazione Tecnologica e Sperimentazione
Via Cavalli, 6 10138 Torino.
+39 011 8103171
redazione.crit@rai.it
www.crit.rai.it

Immagine di copertina creata da Alberto Ciprian

Editoriale

Gino **Alberico**
Direttore del Centro Ricerche, Innovazione Tecnologica e Sperimentazione Rai

L'intelligenza artificiale non è solo una storia di tecnologia. È una trasformazione culturale che sta già cambiando ogni aspetto della nostra vita, dai trasporti e l'assistenza sanitaria alla vendita al dettaglio e all'intrattenimento. Non possiamo fare a meno di essere travolti dall'entusiasmo per questa trasformazione, anche se è difficile prevedere dove ci condurrà. Nel frattempo, mentre procediamo con l'utilizzo di soluzioni abilitate dall'Intelligenza Artificiale, vorremmo essere consapevoli dei rischi e delle sfide che ci attendono.

Cari lettori,
l'incipit dell'editoriale che avete appena letto non è stato scritto da me: si tratta, infatti, di un testo generato da una rete neurale basata sul modello di linguaggio GPT-3, a seguito di una frase ("come l'Intelligenza Artificiale trasformerà i media e l'intrattenimento") fornita come "stimolo". Il GPT-3 (Generative Pre-trained Transformer) è la terza generazione di un gruppo di modelli di linguaggio noti come 'transformer', che nella sua versione completa si basa su una rete neurale con 175 miliardi di parametri.

Stiamo parlando di quella che è ancora una frontiera delle applicazioni di Intelligenza Artificiale (IA) non pienamente esplorata, ossia la generazione di contenuti originali (testi, immagini, …). Tuttavia ci sono altri campi in cui le tecnologie di IA si possono invece considerare indiscutibilmente mature. Tra queste, solo per citare le più diffuse, il riconoscimento e la classificazione di immagini, la trascrizione del parlato e l'analisi semantica del testo. Attraverso tali tecnologie è possibile creare in maniera automatica 'metadati' che possono poi essere utilizzati per aiutare i processi di archiviazione, ricerca, utilizzo e sfruttamento dei contenuti.

Senza dubbio l'IA porterà trasformazioni profonde nel settore dei media e dell'intrattenimento, in quanto impatterà su tutti gli elementi della catena del valore: ad esempio agevolando il processo creativo degli autori, aumentando la produttività degli editori, supportando i processi decisionali e le strategie di pubblicazione e, infine, agevolando gli utenti nel trovare il contenuto che corrisponde ai loro interessi.

Uno dei motivi per cui, fino ad oggi, le applicazioni dell'IA non hanno ancora raggiunto un utilizzo su vasta scala da parte delle 'media company' è che per sfruttare appieno le potenzialità di strumenti sofisticati come gli algoritmi di 'deep learning' occorre disporre di dati di addestramento (dataset) affidabili. Infatti, negli algoritmi di apprendimento supervisionato, i set di dati necessari per addestrare il modello devono essere etichettati manualmente, e questo, per set di dati di grandi dimensioni, rende il processo complesso e costoso. Per contro, di recente si sono sviluppate tecniche che utilizzano una sorta di auto-apprendimento automatico (self-supervised) delle reti, grazie alle quali la bontà del risultato dipende dalla qualità e pertinenza dei dati che sono stati utilizzati nella fase di pre-addestramento e le applicazioni finali possono essere realizzate utilizzando una piccola quantità di dati annotati. Questa seconda famiglia di algoritmi è quindi più promettente poiché, in linea teorica, le media company possono contare su una elevata disponibilità di dati per il pre-addestramento e su un'elevata qualità delle annotazioni. Tutte le architetture di ultima generazione pensate per diversi compiti quali la trascrizione, la traduzione e l'analisi complessa dei dati multimediali si basano su approcci 'self-supervised'.

201_E_1.00

Le tecnologie di IA, oltre a crescere a ritmi vertiginosi, si stanno diffondendo molto rapidamente per via delle loro notevoli prestazioni, ma anche grazie alla facilità di accedervi tramite semplici applicazioni installabili su un comune telefono cellulare. Allo stesso tempo alcuni fenomeni correlati a tale sviluppo costituiscono motivo di preoccupazione, ad esempio in quanto rendono possibile la creazione di contenuti verosimili ma falsi (i cosiddetti 'deep fake') o più in generale poiché studiare e spiegare adeguatamente i comportamenti anomali di tali tecnologie risulta più complesso rispetto alle tecnologie tradizionali. Non a caso il problema della falsificazione dei contenuti e della disinformazione è affrontato a livello mondiale e i filoni di ricerca accademica e applicata sul tema dell'etica e della spiegabilità dell'IA stanno fortemente emergendo.

Volendo chiudere con un messaggio di speranza, come tutte le tecnologie dirompenti, anche l'IA non è connotabile a priori come positiva o negativa, quanto invece lo possono essere i suoi buoni o cattivi utilizzi. Ad esempio, le stesse tecnologie di IA dimostrano di essere degli ottimi strumenti sfruttabili anche per analizzare e smascherare la diffusione di contenuti falsi ed ingannevoli.

In questo numero, curato dai colleghi Giorgio Dimino e Alberto Messina, oltre ad un'introduzione alle moderne tecniche dell'AI e una breve rassegna delle possibili applicazioni nell'ambito dei media, troverete una sintesi di alcuni dei progetti su cui il Centro Ricerche Innovazione Tecnologica e Sperimentazione della Rai sta lavorando, quali ad esempio la sottotitolazione automatica, la meta-datazione negli archivi audiovisivi, la produzione automatica di rassegne stampa, i sistemi di raccomandazione e gli assistenti vocali.

Buona lettura ….

Introduzione alle moderne tecniche dell'Intelligenza Artificiale

Giorgio **Dimino**
Rai - Centro Ricerche, Innovazione Tecnologica e Sperimentazione

L'intelligenza artificiale (o IA, dalle iniziali delle due parole, in italiano) è una disciplina appartenente all'informatica che studia i fondamenti teorici, le metodologie e le tecniche che consentono la progettazione di sistemi hardware e sistemi di programmi software capaci di fornire all'elaboratore elettronico prestazioni che, a un osservatore comune, sembrerebbero essere di pertinenza esclusiva dell'intelligenza umana.

(Marco Sommalvico)
Da Wikipedia, the free encyclopedia

Introduzione

In questo articolo verrà fornito uno spaccato introduttivo sulle *tecniche di Intelligenza Artificiale* allo stato dell'arte, ovvero le *reti neurali profonde*, che in meno di dieci anni hanno profondamente cambiato le aspettative di una tecnologia che sino ad ora non era riuscita, se non in alcuni ambiti verticali, a diventare pervasiva nella vita quotidiana delle persone. I recenti avanzamenti nel campo delle reti neurali profonde hanno portato le prestazioni di sistemi afferenti ai temi classici dell'intelligenza artificiale (*classificazione di testi o immagini*, *riconoscimento di volti*, *traduzione automatica*, *guida autonoma*) a livelli prossimi a quelli richiesti per un utilizzo industriale su larga scala. Questo articolo si propone di fornire le nozioni di base per capire il funzionamento delle reti neurali profonde e le problematiche ad esse connesse, lasciando ad altri articoli in questo stesso numero della rivista il compito di analizzare ambiti applicativi specifici e le relative peculiarità.

Sino a non più di dieci anni fa le reti neurali profonde (DNN) *erano guardate con sospetto nel mondo accademico, poi improvvisamente tutto è cambiato: nel giro di pochissimi anni la situazione si è ribaltata e oggi nelle conferenze di intelligenza artificiale fanno la parte del leone. Che cosa è successo? L'avanzamento tecnologico nel campo della videografica ha reso disponibili piattaforme di calcolo parallelo a bassissimo costo, su cui sono stati realizzati framework open source per la prototipazione di reti neurali, rendendo accessibile anche a piccoli gruppi di ricerca le risorse di calcolo necessarie a sperimentare una tecnologia così computazionalmente intensiva.*

Le reti neurali oggi rappresentano lo stato dell'arte in molti ambiti applicativi nel campo della visione artificiale e del processamento automatico del linguaggio. La pervasività della tecnologia è resa possibile dal fatto che le DNN hanno la caratteristica desiderabile di essere in gran parte indipendenti dal dominio di applicazione, per cui nuove tecniche e ottimizzazioni sperimentate per la risoluzione di un particolare problema spesso risultano immediatamente estendibili ad altri domini. Una rete è caratterizzata da un modello che determina le interconnessioni tra i neuroni che compongono la rete, una funzione di errore che ne permette l'addestramento ed un insieme di iperparametri che ne determinano la dimensione.

Questo contributo fornisce i fondamenti per comprenderne il funzionamento, le principali caratteristiche e le problematiche associate.

201_01.1_2.00

Definizione di Intelligenza Artificiale

In generale parliamo di *Intelligenza Artificiale* riferendoci a una macchina che è in grado di eseguire compiti che richiedono un'elaborazione dei dati non algoritmicamente predefinita allo scopo di prendere delle decisioni ad alto livello (es. giocare a scacchi).

Questi sono alcuni dei campi di applicazione dell'IA:

- Problem solving
- Ragionamento autonomo
- Interpretazione semantica dei dati
- Autoapprendimento
- Modellazione della conoscenza
- Visione artificiale
- Riconoscimento del parlato

Le principali tecniche che fanno capo all'Intelligenza Artificiale possono essere schematicamente raggruppate in due grandi filoni di ricerca. Il primo, particolarmente in auge negli anni '80 e '90 del secolo scorso, si basa sulla *modellazione esplicita della conoscenza* e ha portato alla realizzazione dei cosiddetti *sistemi esperti*, basati su un'organizzazione gerarchica della conoscenza (ad esempio insiemi di ontologie e regole) per trarre conclusioni logiche a partire dai dati. Il secondo, comunemente conosciuto col nome di *Machine Learning*, è focalizzato sull'apprendimento automatico basato sulla ricerca di metodi di regressione che permettano di approssimare un certo fenomeno di interesse con una distribuzione stocastica, stimata a partire da un numero limitato di osservazioni, in modo da poterne predire gli stati futuri.

Le *reti neurali profonde* (*Deep Neural Networks* - ***DNN***) sono una derivazione del machine learning che si basa sulla realizzazione di reti neurali formate da un gran numero di celle elementari identiche, ispirate dalla conformazione del cervello umano.

Cenni storici

I tentativi di impiegare i calcolatori per compiti che emulano il comportamento umano risalgono agli anni '50. Già Alan Turing ipotizzava la realizzazione di una macchina in grado di apprendere autonomamente e definiva il concetto di intelligenza artificiale tramite la famosa metafora dell'*imitation game* [1], detto anche *test di Turing*, secondo cui una macchina può dirsi dotata di intelligenza se un interlocutore dietro ad un tendone non è in grado di distinguere se stia parlando con un essere umano oppure con un automa. La prima formulazione di rete neurale, denominata *multilayer perceptron* [2], risale agli anni '60, ispirata dall'osservazione della anatomia del cervello umano, ma solo negli anni '80 viene introdotto un meccanismo di apprendimento basato sulla propagazione all'indietro del gradiente dell'errore [3], principio fondamentale dell'apprendimento delle reti moderne. Negli stessi anni vengono definite le *reti convoluzionali per la visione artificiale* [4], anch'esse ispirate all'anatomia della corteccia visiva [5].

Gli studi sulle DNN procedono negli anni senza particolare impulso, principalmente ad opera di pochi studiosi tra cui ***Yan LeCun***, sino al 2012, quando un gruppo di ricercatori dell'università di Toronto presenta alla conferenza internazionale **NIPS** (*Neural Information Processing Systems*) un articolo [6] in cui dimostra di essere in grado di surclassare lo stato dell'arte in un noto benchmark di classificazione di immagini, chiamato *ImageNet Classification*, con un sistema basato su reti neurali convoluzionali, denominato **AlexNet**. Tale sistema era a sua volta ispirato da un articolo del 1998 di Bengio e LeCun [7], in cui viene definita l'architettura della prima rete convoluzionale per la classificazione di immagini, chiamata **LeNet**. Da allora, e nel giro di pochissimi anni, le DNN sono state impiegate per migliorare lo stato dell'arte in praticamente tutti i campi dell'IA, dal riconoscimento del parlato, alla traduzione automatica, alla classificazione di immagini.

Tra i fattori che hanno determinato una evoluzione così rapida di una tecnologia rimasta dormiente per così tanti anni vi sono due condizioni esterne fondamentali. La prima è l'introduzione di *acceleratori* per il calcolo parallelo nelle schede grafiche dei personal computer (**GPU**), fondamentali per il rendering in tempo reale degli scenari complessi

dei videogiochi moderni. ***NVIDIA***, uno dei maggiori produttori di chip grafici, intuendo la potenzialità dell'utilizzo delle GPU nella simulazione numerica, ha sviluppato e reso disponibile gratuitamente l'**SDK CUDA** [8] che permette l'utilizzo degli acceleratori per eseguire calcoli vettoriali all'interno di programmi utente, principalmente a scopi di simulazioni scientifiche. Questa tecnologia si sposa perfettamente con le esigenze delle reti neurali ed ha reso possibile ridurre di due ordini di grandezza [9] il tempo necessario ad effettuare l'addestramento di una rete, per di più utilizzando hardware di costo limitato. I ricercatori hanno quindi potuto accedere a piattaforme hardware di basso costo adatte a sperimentare architetture di reti neurali più complesse di quelle utilizzate in precedenza. La seconda condizione è legata alla diffusione del *software open source*. Quando sono apparsi i primi articoli scientifici che dichiaravano risultati oltre lo stato dell'arte sui benchmark di machine learning più popolari, lo scetticismo della comunità scientifica era piuttosto alto. A volte gli articoli venivano addirittura rigettati dai revisori in quanto i risultati riportati erano considerati inverosimili.

La diffusione di librerie open source basaste sul **CUDA** di ***NVIDIA*** per l'esecuzione di reti neurali (ad es. **TensorFlow** di ***Google*** [10] o **PyTorch** di ***Facebook*** [11]) ha permesso ai ricercatori di corredare gli articoli con i programmi software utilizzati per ottenere i risultati dichiarati ed alla comunità scientifica di validare i suddetti risultati e di sperimentare con quanto realizzato da altri [12]. Oggi qualsiasi articolo scientifico sulle reti neurali che aspiri ad un certo grado di credibilità viene corredato da un'implementazione software del modello proposto e ciò ha generato un circolo virtuoso in cui tutti i risultati ottenuti sono immediatamente messi a disposizione dell'intera comunità scientifica che può quindi sperimentare velocemente nuove soluzioni partendo dallo stato dell'arte più recente. Inoltre, l'utilizzo di GPU commerciali, disponibili anche in molte piattaforme di cloud computing commerciali, ha reso la sperimentazione delle reti neurali accessibile a chiunque abbia le conoscenze necessarie, senza necessità di investimenti rilevanti in infrastruttura.

Modello di base

Le prime formulazioni del modello di rete neurale hanno origine dal tentativo di imitare il funzionamento del cervello umano realizzando modelli simili alle reti di neuroni che costituiscono il cervello. Alla base di questo approccio viene postulato che, definito un modello semplificato del cervello, sia possibile addestrarlo ad eseguire dei compiti cognitivi senza dover programmare passo a passo lo svolgimento dell'attività. Sebbene non sia mai stato provato che questi modelli rappresentino un'approssimazione realistica delle facoltà cerebrali, i risultati ottenuti sono stati incoraggianti e hanno spinto la ricerca a sperimentare reti via via più complesse. Il modello si basa su di un elemento detto *perceptron* [13] che implementa la funzionalità di un singolo neurone (Fig. 1). Si costruisce poi una rete composta da un numero elevato di perceptron (da qui in avanti definiti *neuroni*) in comunicazione tra loro.

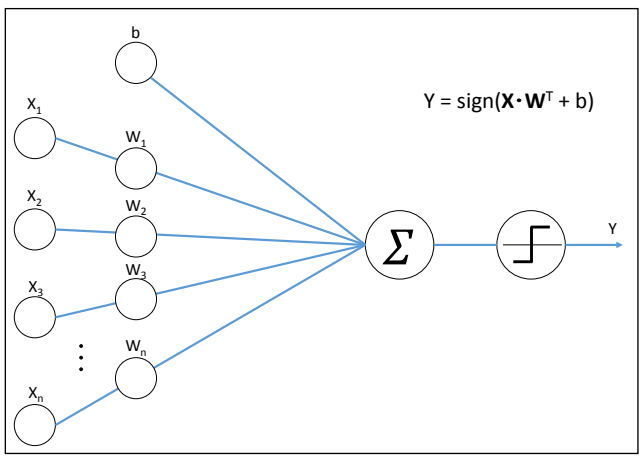


Fig. 1 – Perceptron

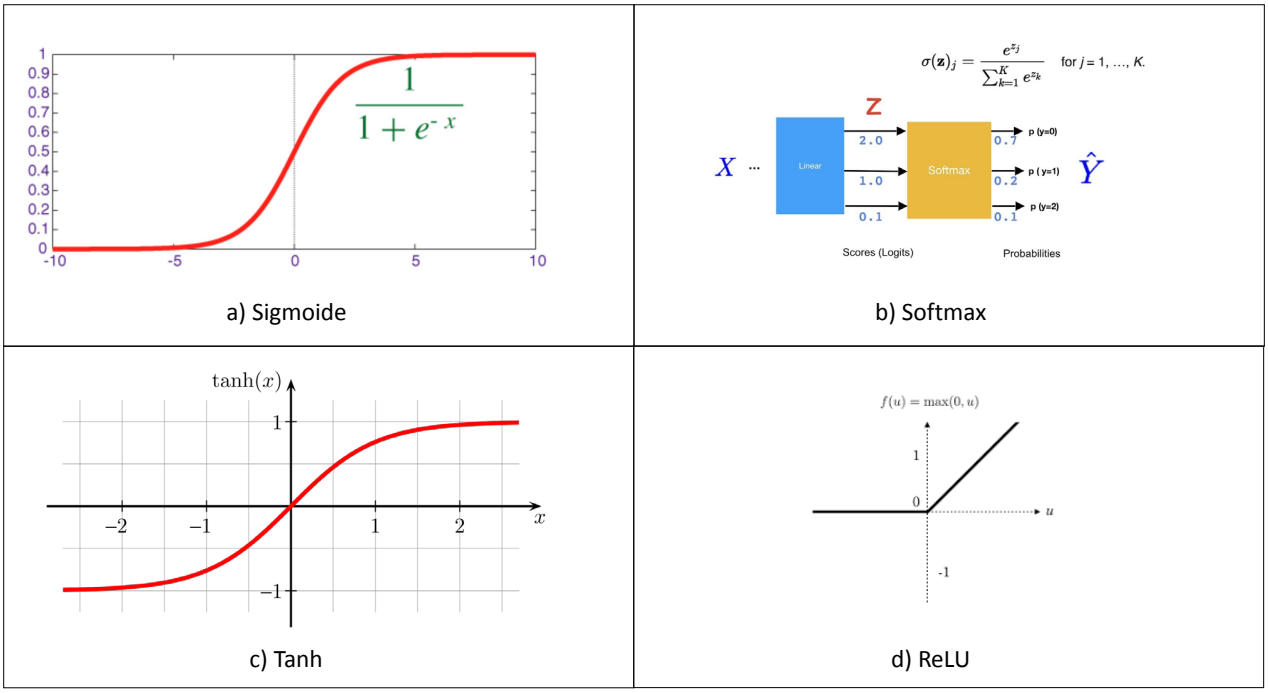
- ***X*** è il vettore di ingresso (es, i pixel di un'immagine oppure l'uscita di altri neuroni)
- ***W*** è un vettore che contiene i parametri del neurone che vengono *appresi* durante il training
- *b* è il termine di bias
- *sign()* è la funzione di attivazione
- *Y* è l'uscita del neurone, detta anche *attivazione*

Quindi la funzione di trasferimento del modello consiste nel prodotto scalare tra il vettore di ingresso ***X*** e il vettore ***W*** (detto anche *campo di ingresso*), modulato dalla *funzione di attivazione* ***f()***, in questo caso realizzata dalla funzione *sign()*, che restituisce *1* se l'input è positivo oppure *–1* se l'input è negativo. Il vettore ***W*** e il *termine di bias* ***b*** (collettivamente chiamati *pesi*) vengono appresi dalla rete autonomamente durante la *fase di training*, che consiste in un processo iterativo in cui si calcola l'errore tra l'uscita attesa e quella prodotta dalla rete su di una serie di esempi il cui risultato è noto, e si utilizza l'errore per modificare progressivamente il valore dei pesi sino a convergere su di un valore di errore minimo. Si noti come nel caso in cui la funzione di attivazione sia un'applicazione lineare del campo di ingresso, il perceptron è analiticamente identico ad un *classificatore lineare*.

La caratteristica che rende potenti le *reti neurali* risiede nell'introduzione di una nonlinearità nel sistema attraverso l'utilizzo di particolari funzioni di attivazione, e nel sovrapporre più livelli di neuroni realizzando così delle reti profonde in analogia col sistema cognitivo umano. Infatti, se il modello del neurone fosse puramente lineare, per il principio della sovrapposizione degli effetti dei sistemi lineari sarebbe sempre possibile collassare la rete in un'unica funzione lineare. Si veda al riguardo il teorema di approssimazione universale che stabilisce che una funzione continua può essere approssimata da una rete del tipo appena descritto di ampiezza o profondità arbitrariamente grande [14][15].

In Fig. 2 sono illustrate le funzioni di attivazione utilizzate più frequentemente nei sistemi.

Fig. 2 – Principali *funzioni di attivazione*:
a) *Sigmoide*: fornisce un valore compreso tra 0 e 1; può quindi essere usata per esprimere una probabilità. Può essere vista come una versione stocastica della funzione gradino utilizzata nei sistemi binari.
b) *Softmax*: agisce su un insieme di attivazioni in ingresso e fornisce un valore tra 0 e 1 ma in più la somma di tutti gli elementi fa sempre 1, quindi è adatta a modellizzare la probabilità di appartenenza ad una classe (ed una sola) fra N classi date.
c) *Tangente iperbolica*: fornisce un valore tra -1 e 1 e viene utilizzata nei livelli interni della rete.
d) *ReLU*: fornisce un valore tra 0 e 1 e viene utilizzata nei livelli interni della rete.



a) Sigmoide

b) Softmax

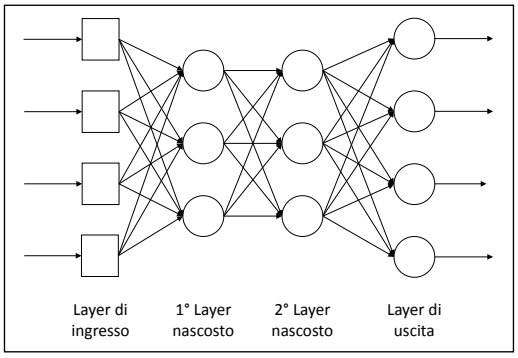
c) Tanh

d) ReLU

Nella configurazione di Fig. 1, il singolo neurone si comporta quindi come un *classificatore binario*, ovvero indica se il vettore di ingresso appartiene o meno ad una data classe. È evidente che è possibile estendere questa rete mettendo più neuroni in parallelo per realizzare un classificatore a *n* classi. Una rete di questo tipo è concettualmente molto simile ai più comuni classificatori utilizzati nel Machine Learning, come **Naive Bayes Classifier** [16] o **Support Vector Machine** [17]. La potenza delle reti neurali però diventa evidente quando vengono sovrapposti più layer di neuroni come schematizzato in Fig. 3.

Si può allora ipotizzare che i layer interni operino delle trasformazioni sui dati di ingresso via via a più alto livello in modo da agevolare il lavoro di classificazione dell'ultimo layer.

Fig. 3 – Multilayer Perceptron

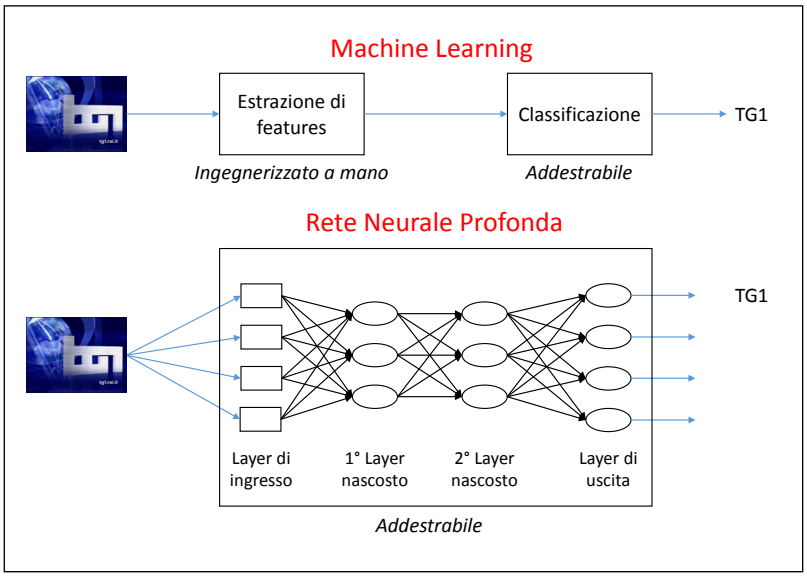


Nel *Machine Learning classico* il processo di progettazione della rete può essere logicamente diviso in due fasi: nella prima vengono individuate, principalmente manualmente, le trasformazioni dei dati di ingresso (chiamate *processi di estrazione di features*) necessarie a portare i dati in uno spazio facilmente separabile, nella seconda invece viene definito un *classificatore* idoneo (che può essere o meno supervisionato). L'efficacia della rete dipende quindi pesantemente dalla scelta delle trasformazioni della prima fase, che solitamente richiedono approfondita conoscenza del dominio dei dati che si devono trattare. Il processo nel suo complesso è scarsamente ingegnerizzabile e non generalizza facilmente. Al contrario in una *DNN*[Nota 1] tutti i layer vengono addestrati contemporaneamente a partire dal set di dati di ingresso e non è richiesta una conoscenza specifica del dominio in quanto la rete *scopre* autonomamente durante il processo di addestramento quali sono le trasformazioni dei dati di ingresso più adatte ad eseguire il compito assegnato.

La Fig. 4 illustra schematicamente il confronto tra questi due approcci.

Nota 1 - Da questo punto in poi useremo il termine *rete neurale* per indicare una DNN

Fig. 4 – Confronto tra *algoritmo di Machine Learning* e *Rete Neurale Profonda* (DNN)

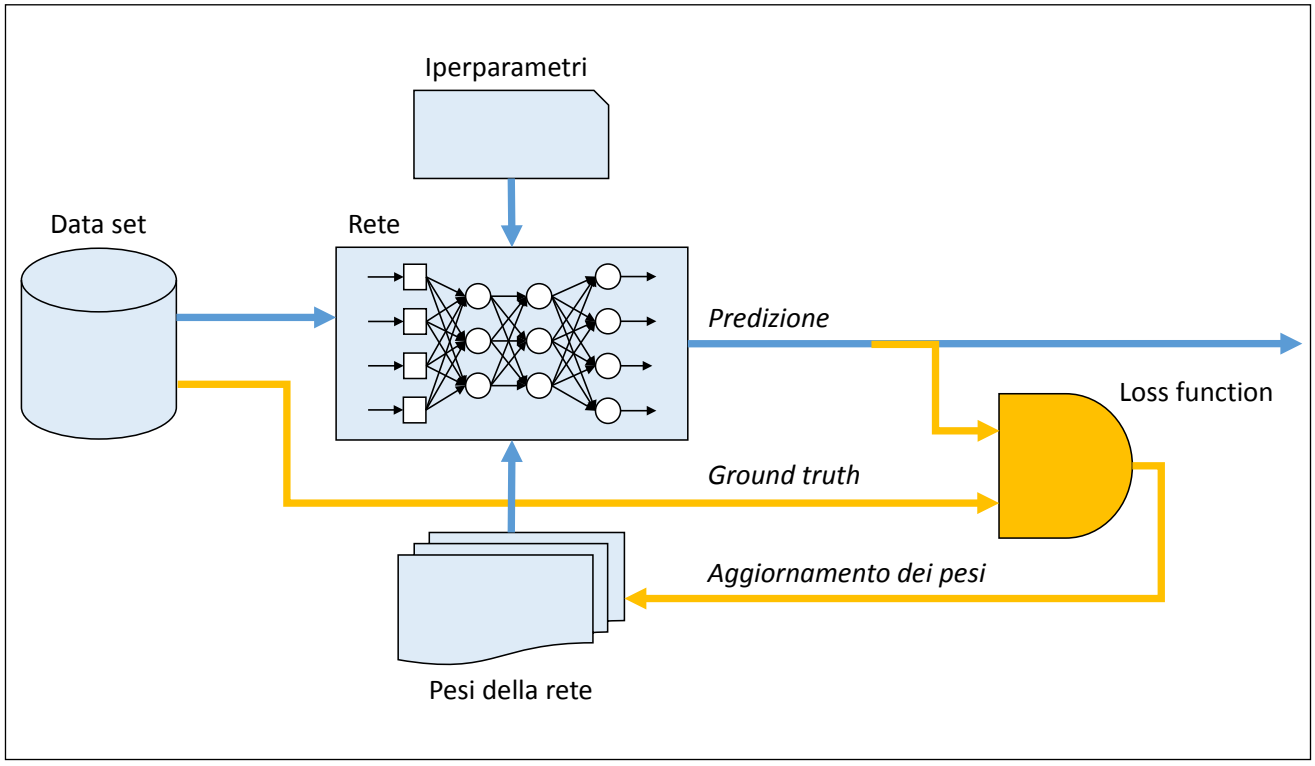


Il progetto di una rete neurale in grado di assolvere un determinato compito richiede la definizione dei seguenti elementi, che verranno illustrati nelle prossime sezioni:

- l'*architettura della rete* ed i relativi *iperparametri*
- il *data set* annotato su cui effettuare l'addestramento
- la *loss function*, ovvero la funzione che misura lo scostamento del comportamento della rete da quello atteso

Una rete neurale ha due modalità di funzionamento: la prima effettua l'addestramento modificando gradualmente i parametri dei neuroni che la compongono sino a minimizzare l'errore calcolato tramite la loss function (Fig. 5), la seconda ne permette l'uso in ambiente di produzione.

Fig. 5 – Schema di principio di addestramento di una Rete Neurale. La parte in giallo sottende all'addestramento della rete e viene rimossa quando la rete viene utilizzata in produzione



Addestramento

Generalità

Le reti neurali fanno parte della famiglia di algoritmi cosiddetti *supervisionati*, ovvero che vengono condizionati ad eseguire un dato task tramite un processo di addestramento in cui i parametri del sistema vengono iterativamente modificati in modo da riprodurre il comportamento atteso su di un insieme di esempi pre-annotati (il *data set di addestramento*). Per effettuare l'addestramento è necessario definire una funzione, detta *loss function*, che ha il compito di stimare l'errore medio effettuato dalla rete nel valutare un certo set di dati di ingresso rispetto ai valori attesi e che funge quindi da guida per l'addestramento. La caratteristica principale richiesta alla *loss function* è quella di essere derivabile rispetto ai pesi della rete, infatti il meccanismo di addestramento si basa sulla propagazione all'indietro (*backpropagation*) del gradiente dell'errore medio [18], calcolato rispetto ai pesi della rete, mediante la funzione

$$w_i \Leftarrow w_i - \lambda \frac{\partial L\left(f(\mathbf{x}; \mathbf{w})\right)}{\partial w_i}$$

o in forma vettoriale

$$\mathbf{w} \Leftarrow \mathbf{w} - \lambda \nabla L(f(\mathbf{x}; \mathbf{w}))$$

Il termine λ è detto *coefficiente di apprendimento* e governa la quantità massima di variazione ammessa per ciascun peso ad ogni iterazione, w_i sono i *pesi della rete*, $f(\boldsymbol{x}; \boldsymbol{w})$ è l'uscita della rete neurale presa a riferimento per l'apprendimento e $L()$ è la *loss function*.

In pratica l'apprendimento avviene nel seguente modo:

- si inizializza la rete inizializzando i parametri secondo un determinato criterio (ad esempio seguendo una distribuzione normale);
- si procede effettuando un certo numero di cicli (detti *epoche*) in cui tutto il training set, suddiviso in mini-batch per ragioni di praticità di calcolo, viene processato dalla rete;
- per ogni mini-batch si calcola l'errore medio della rete tramite la loss function $L()$ e si utilizza il ∇L calcolato rispetto a $\boldsymbol{w}$ per aggiornare ciascun peso w_i;
- l'aggiornamento finisce dopo un numero predefinito di epoche o quando l'errore calcolato su di un data set di validazione (una partizione del data set non utilizzata per il training) non decresce più.

Ottimizzare l'addestramento

La scelta dell'iper-parametro λ è cruciale, infatti valori troppo piccoli determinano una convergenza della rete verso il minimo dell'errore molto lenta e sono quindi necessarie elevate risorse di calcolo, per contro un valore troppo elevato porta ad un andamento irregolare della loss function e questo potrebbe comportare la divergenza della rete e quindi il fallimento dell'addestramento. È evidente che per come il processo di apprendimento è definito, non vi sia alcuna garanzia di riuscire a raggiungere il minimo assoluto della loss function, anzi è molto frequente il caso in cui il processo si arresti in un minimo locale o addirittura non converga verso un valore stabile. Per questo in letteratura vengono proposti numerosi accorgimenti per stabilizzare l'apprendimento [19], detti *di ottimizzazione*. Citeremo qui solo alcuni dei più utilizzati.

Stocastic Gradient Descent (***SGD***): l'aggiornamento dei pesi dovrebbe avvenire al termine di ogni epoca, quando cioè è stata calcolata la media della loss function sull'intero data set di training. In pratica, poiché il data set è solitamente troppo grande per essere contenuto nella memoria dell'unità di calcolo, si preferisce partizionarlo in mini-batch più piccoli, generati campionando in modo casuale il data set, e si aggiornano i pesi ad ogni mini-batch, nell'ipotesi che i dati contenuti nel mini-batch siano sufficienti a rappresentare l'intero set. A questo modo la convergenza della rete è anche più rapida.

Cosine annealing: il coefficiente di apprendimento λ viene modulato nel corso delle epoche secondo una funzione coseno in modo da ridurne progressivamente il valore man mano che il training procede. Più ci si avvicina al minimo più ci si muove per piccoli passi per evitare di scavalcarlo.

Momento del gradiente: in ciascuna iterazione la correzione è data da una media pesata del gradiente attuale e di quello del passo precedente [20]. Di fatto viene inserita un'inerzia nel gradiente per evitare variazioni di direzione troppo repentine magari in una fase in cui la rete è all'inizio del training e ancora instabile.

RMSprop: il gradiente calcolato viene diviso per una quantità proporzionale alla radice quadrata della media pesata del quadrato del gradiente attuale e del quadrato del gradiente precedente. In questo modo si cerca di normalizzare la quantità di variazione tra i diversi pesi della rete.

Weight Decay: nella loss function viene aggiunto un termine che penalizza la complessità della rete, che si riflette nello smorzamento del gradiente di un termine proporzionale al valore del peso. Reti troppo complesse tendono più facilmente a generare *overfitting*.

Con il termine *overfitting* si intende un fenomeno patologico comune nei sistemi supervisionati in cui un addestramento troppo prolungato o realizzato in modo scorretto rende la rete simile a una memoria associativa. Ovvero, la rete anziché costruire un modello che generalizza il comportamento condizionato durante il training, impara così bene a riconoscere gli esempi del training set da rigettare ogni altro input. Pertanto, è necessario prendere ogni precauzione per evitare di generare questa situazione durante l'addestramento. Il primo accorgimento che viene solitamente adottato consiste nel comparare durante il training il valore della loss function applicata al training set col valore calcolato sul set di validazione. Quando il primo risulta essere sensibilmente più basso del secondo (ricordiamo che l'obiettivo del training è la minimizzazione della loss function) vuol dire che la rete sta perdendo la capacità di generalizzare e che non è possibile spingere oltre l'addestramento.

Per mitigare il rischio di overfitting è possibile introdurre varie tecniche dette di regolarizzazione della rete.

Il ***dropout*** è una di queste tecniche in cui parte dell'informazione durante il training viene cancellata di proposito per evitare che la rete apprenda dagli esempi troppi dettagli perdendo in generalità. Si applica solitamente sui dati del training set, mascherando parte dei dati in modo casuale. Ad esempio, se l'ingresso è un'immagine, ne viene mascherata o distorta una porzione. Ovviamente ad ogni epoca il dropout deve agire su dati diversi. È possibile anche mascherare le attivazioni dei layer interni oppure, anche se utilizzato più di rado, eliminare temporaneamente alcuni collegamenti tra i neuroni della rete (ovvero azzerarne i pesi).

Un'altra tecnica utilizzata è la ***data augmentation*** in cui i dati del training set vengono processati per ottenere in modo automatico ulteriori esempi su cui effettuare l'addestramento. Ad esempio, nel caso delle immagini, si possono applicare trasformazioni che ne varino il valore dei pixel senza per questo modificarne il contenuto semantico o generare situazioni inverosimili. Le trasformazioni tipiche sono traslazione, rotazione, zoom, variazione di luminosità e contrasto, distorsione, ma quali siano appropriate o meno dipende dal caso specifico.

Un'altra situazione patologica che vanifica l'addestramento è il cosiddetto *annullamento (o esplosione) del gradiente*. Il gradiente calcolato assume valori così piccoli o così grandi da non poter più essere rappresentato con la notazione floating point di un computer. Per molti anni la profondità delle reti neurali è stata mantenuta bassa a vantaggio del parallelismo dei neuroni nello stesso layer proprio per problematiche legate all'annullamento del gradiente. Per ovviare al problema, solitamente si inseriscono livelli di normalizzazione delle attivazioni tra un layer e l'altro utilizzando una tecnica detta **BatchNorm** [21], che consiste nel normalizzare le attivazioni di uscita di ciascun layer mediante una

formula basata su media e varianza calcolate sul minibatch attuale.

La loss function

Come abbiamo visto la *loss function* è un componente fondamentale della rete, che ha il compito di valutare durante il training quanto il comportamento della rete si discosti da quello atteso. Poiché l'addestramento si basa sulla propagazione del gradiente della loss function, è evidente che questa deve essere derivabile rispetto ai parametri della rete oggetto dell'apprendimento.

A seconda del tipo di uscita della rete, la loss function potrà assumere formulazioni completamente differenti. Un primo caso è quando la rete implementa un classificatore. L'uscita sarà un vettore, la cui dimensione è pari al numero di classi tra cui si deve discriminare, e dove ciascuna componente del vettore può essere associata ad una stima della probabilità che i dati di ingresso appartengano ad una data classe. La misura che viene generalmente utilizzata in questo caso è detta *cross-entropy* ed è espressa dalla formula:

$$CE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\left(y_{i,j}\cdot\log\left(p\left(y_{i,j}\right)\right)+\left(1-y_{i,j}\right)\cdot\log\left(1-p\left(y_{i,j}\right)\right)\right)$$

dove $y_{(i,j)}$ è un indicatore binario che dice se il campione i-esimo appartiene o meno alla classe j-esima e $p(y_{(i,j)})$ è la probabilità stimata dal sistema che il campione i-esimo appartenga alla classe j-esima.

Se invece l'uscita rappresenta una grandezza numerica, viene normalmente usata una misura che esprime una distanza, come l'errore quadratico medio (MSE) oppure la sua radice quadrata (RMSE)

$$MSE = \frac{1}{N}\sum_{1}^{N}(y_i^2 - \hat{y}_i^2)$$

$$RMSE = \frac{1}{N}\sum_{1}^{N}\sqrt{(y_i^2 - \hat{y}_i^2)}$$

dove y_i è il valore vero corrispondente al campione x_i e $\hat{y}_i$ la sua predizione elaborata dalla rete.

Quando è necessario mediare tra più esigenze, è possibile comporre la loss function come somma di più termini, dove ciascun termine esprime un errore da minimizzare. Nei casi più complessi la loss function può contenere una seconda rete neurale addestrata in precedenza o addirittura in tandem con la rete principale, come nel caso delle *reti GAN* che vedremo più avanti.

Il data set per l'addestramento

A seconda della complessità del problema da risolvere il set di dati da utilizzare per l'addestramento, *data set*, può essere più o meno esteso, ma in generale è costituito da un insieme di dati pre-annotati (ad es. immagini ed etichette corrispondenti) scelti in modo da caratterizzare tutti gli aspetti salienti delle classi tra cui si vuole discriminare. Per fare un esempio pratico, il data set più utilizzato in ambiente accademico per effettuare benchmark sulla classificazione di immagini, **ImageNet** [22], contiene circa *14.000.000* di immagini suddivise in *2000* classi diverse. Generalmente una parte del data set (ad esempio il 20%) viene utilizzato come *validation set*, ovvero viene usato come set di riscontro per verificare la capacità della rete di generalizzare, cioè di rispondere correttamente a fronte di dati di ingresso mai visti in precedenza, e non contribuisce al training.

La selezione di un data set appropriato è un'operazione critica nella progettazione di un sistema basato su reti neurali. Innanzi tutto, non esistono formule per calcolare quanto debba essere esteso il data set, in linea di massima più è grande e meglio è. La difficoltà non risiede solo nella possibilità di reperire un gran numero di esempi, ma nel fatto che questi devono essere annotati con la risposta attesa (ad es. la categoria di appartenenza di un'immagine). In casi come la traduzione automatica ciò significa che è necessario disporre di un corpus di testi allineato sulle due lingue tra cui si vuole effettuare la traduzione. Essendo molto oneroso produrre un corpus simile appositamente si ricorre di solito a sorgenti preesistenti, come gli *atti della Commissione Europea*, che sono pubblici e disponibili nelle lingue ufficiali dell'Unione.

Per ottenere dalla rete prestazioni elevate è necessario che i dati annotati che compongono il corpus siano il più possibile attinenti al dominio in cui la rete verrà impiegata, ed è evidente che questo limita notevolmente la disponibilità di dati provenienti da sorgenti esistenti.

Negli ultimi anni si è andata ad affermare la tecnica del *transfer learning*, ovvero del *trasferimento dell'apprendimento*. Questa tecnica si basa sull'ipotesi che in una rete i primi livelli siano sostanzialmente invarianti al variare del compito della rete. Il training viene quindi effettuato partendo da una rete che non viene inizializzata con parametri casuali ma che è già stata addestrata su di un corpus esteso e il più possibile generico. Ad es. nel caso delle immagini, si parte solitamente da modelli pre-addestrati su **ImageNet**.

Il training viene effettuato facendo variare in una prima fase solo i parametri relativi agli ultimi layer che implementano il classificatore vero e proprio, lasciando fissi quelli relativi alla parte convoluzionale, che descriveremo più sotto. Raggiunto un certo grado di stabilità si procede poi ad addestrare anche gli altri layer con un learning rate molto basso. In questo modo si riesce ad addestrare una rete ottimizzata su di un particolare dominio con un numero di iterazioni (epoche) molto basso, utilizzando un data set di diversi ordini di grandezza più ridotto rispetto a quello che sarebbe necessario partendo da una rete inizializzata in modo casuale.

Il problema più grande da affrontare quando si compila un data set di addestramento (ma che risulta evidente solo al momento dell'utilizzo della rete addestrata) è quello del cosiddetto *bias* [23], ovvero *pregiudizio*. Se la distribuzione dei dati nel training set non copre uniformemente il dominio in cui la rete si troverà ad operare, c'è il rischio concreto che la rete risulti affetta da bias, cioè che in presenza di dati afferenti a regioni della distribuzione reale non adeguatamente rappresentati nel data set di addestramento la rete si comporti in modo aprioristico con risultati poco attinenti ai dati puntuali di ingresso. È osservabile che se i dati di ingresso sono affetti da qualche pregiudizio la rete non solo lo emulerà ma tenderà ad amplificarlo. Questo pone molti problemi etici nell'impiego di queste tecnologie in campi che possono avere conseguenze rispetto ai diritti degli individui.

Architetture principali

In questa sezione verranno descritte alcune tra le principali architetture di DNN, comunemente utilizzate sia in ambiente accademico che industriale. Dato lo scopo introduttivo dell'articolo, si illustrerà solo il principio di funzionamento di massima delle reti.

Multilayer perceptron (dense network)

Questa architettura, già descritta sommariamente in precedenza, è conosciuta con diversi nomi tra cui *dense network* e *fully connected network*. È stata la prima architettura proposta di rete neurale ed è una rete di tipo *feed forward*, ovvero in cui il segnale fluisce dagli ingressi verso le uscite senza mai creare recursioni. Si caratterizza per il fatto che ciascun neurone di un layer è connesso a tutti i neuroni del layer successivo. Il numero di neuroni del layer di ingresso è pari alla dimensione del vettore di dati di ingresso, mentre i layer interni possono avere dimensioni diverse a seconda del tipo di utilizzo della rete.

Questo schema viene spesso usato come classificatore. In tal caso il numero di celle nel layer di uscita è uguale al numero di classi tra cui si vuole discriminare. La funzione di attivazione di questo layer è normalmente *softmax* se le scelte sono esclusive oppure *sigmoide* nel caso più classi possano essere valide contemporaneamente. Nel primo caso quindi, il valore di ciascun elemento è associabile alla probabilità che l'insieme dei dati di ingresso appartenga alla classe corrispondente mentre nel secondo caso è assimilabile al grado di appartenenza a ciascuna delle classi. Il numero di layer interni e la loro ampiezza dipende in modo empirico dalla complessità del problema che si vuole risolvere. La funzione di attivazione di questi layer è solitamente *ReLU* oppure *tanh* (si veda la Fig. 2).

Reti convoluzionali

Le reti di tipo *multilayer preceptron* sono computazionalmente troppo pesanti per essere impiegate nell'elaborazione di immagini, in quanto sarebbe necessario prevedere un neurone per ogni pixel dell'immagine. Poiché ciascun neurone viene connesso a tutti i neuroni del layer successivo, il numero di parametri da stimare sarebbe eccessivo per i casi di interesse pratico. Si preferisce quindi utilizzare un'architettura basata su *reti convoluzionali*, che permettono di ridurre pesantemente il numero di parametri da stimare per una data profondità della rete.

L'elemento base è descritto nella Fig. 6 e consiste in una serie di *nuclei di convoluzione* (detti anche *filtri*), tipicamente di dimensione *3x3* o *5x5*, i cui coefficienti vengono appresi durante il training.

Nei casi più in uso, la rete si compone di un numero elevato (sino a qualche decina) di layer di questo tipo intercalati con layer che applicano una funzione non lineare di attivazione (solitamente la *ReLU*) e layer che applicano un sottocampionamento degli elementi del layer attuale per generare un'uscita verso il layer successivo di dimensione ridotta. L'operazione effettuata in questi layer è chiamata *max pooling* e consiste nel condensare un blocchetto *2x2* di dati in un unico valore corrispondente al valore massimo nel blocchetto. Alternando layer convoluzionali a layer di pooling si riduce progressivamente la dimensione dei dati in uscita, sino ad arrivare ad un vettore di dimensione indicativamente dell'ordine delle migliaia di elementi.

La rete in questa operazione impara ad estrarre dall'immagine di ingresso le caratteristiche salienti che utilizzerà per risolvere il problema su cui è stata addestrata (Fig. 7). Visualizzando la forma che assumono i filtri generati durante l'addestramento si nota che questi rappresentano strutture via via più complesse procedendo dall'ingresso verso l'uscita [24].

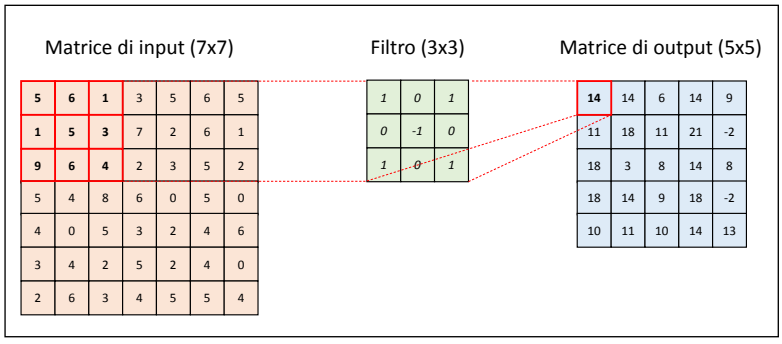


Fig. 6 – Operazione di *convoluzione bidimensionale*. Sull'immagine arancio viene utilizzato un filtro (verde) di dimensione 3x3. Il risultato è la matrice celeste di dimensione 5x5.

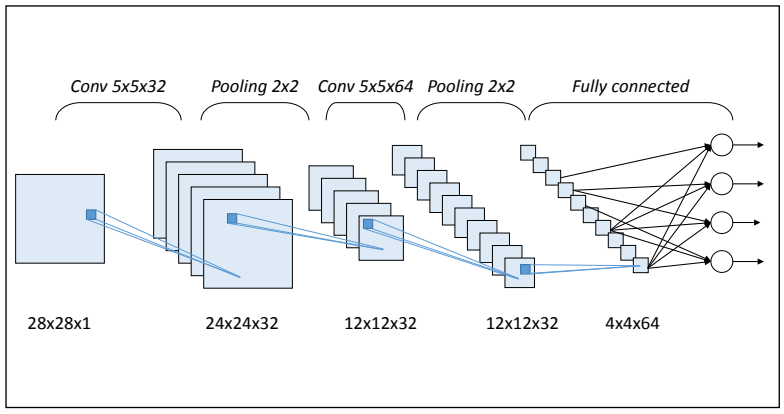


Fig. 7 – Esempio di rete convoluzionale

Se la rete viene impiegata come classificatore di immagini, il layer finale della rete convoluzionale viene connesso ad una *rete dense*, del tipo visto nel paragrafo precedente. Le reti **LeNet** e **AlexNet**, citate in precedenza, sono di questo tipo.

RESNET

Una importante limitazione dell'architettura di rete convoluzionale vista sopra è che è molto difficile addestrare reti con un numero di layer superiore a qualche decina per problemi di annullamento del gradiente già citati.

L'idea alla base della **RESNET** [25] consiste nell'inserire un cortocircuito tra coppie di layer convoluzionali adiacenti, come mostrato in Fig. 8. Di fatto a questo modo viene propagato il residuo tra il segnale di ingresso ed una sua elaborazione effettuata dai layer convoluzionali. Questo schema risulta più stabile e meno soggetto a problemi di regolarizzazione e permette di addestrare agevolmente reti con centinaia di layer. Le reti **RESNET34** e **RESNET50** sono due esempi di configurazioni di **RESNET** spesso utilizzate nei casi pratici, in quanto rappresentano un buon compromesso tra performance e complessità.

L'architettura **RESNET** ha ispirato la realizzazione di diverse altre reti tra cui la **EfficientNet** [26], che allo stato attuale rappresenta tra le reti convoluzionali quella maggiormente efficiente, ovvero quella che riesce ad ottenere risultati allo stato dell'arte nei principali benchmark scientifici con una complessità di calcolo sensibilmente inferiore rispetto alle architetture precedenti.

Reti recursive

Le *reti recursive*, **RNN** (*Recurrent Neural Network*), sono indicate per processare serie ordinate di dati, in cui l'ordine di apparizione dei simboli è importante, come ad es. una sequenza audio oppure dei testi.

La rete ha una struttura molto semplice, basata su di una cella recursiva dove l'uscita di ciascun layer viene rimessa in ingresso al layer stesso. Questa struttura può essere sviluppata come una serie lineare di celle uguali (con gli stessi parametri) che rappresentano lo stato della rete al tempo *t*, *t+1*, *t+2* e così via (si veda l'esempio di Fig. 9).

Nelle reti recursive viene normalmente usata come funzione di attivazione la *tanh* in quanto essendo limitata tra *-1* e *+1*, riduce il problema dell'esplosione del gradiente durante il training. È possibile comporre reti a più layer dove l'uscita di una cella viene posta in ingresso ad un'altra cella dello stesso tipo ma con pesi diversi, così come è possibile creare reti bidirezionali affiancando due celle che vengono alimentate con i dati in ordine inverso l'una rispetto all'altra. L'uscita in questo caso sarà calcolata utilizzando gli stati nascosti (h_t) di entrambe le celle.

Fig. 8 – Elemento base di una architettura RESNET

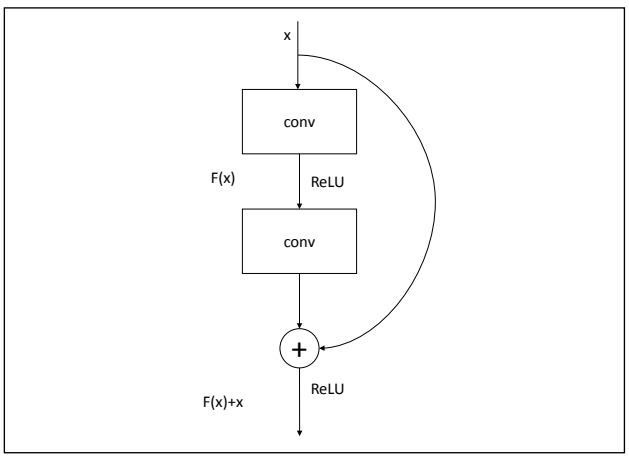
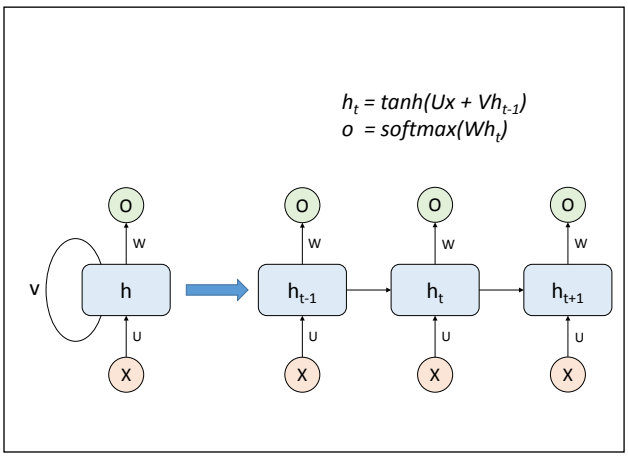


Fig. 9 – Rete Neurale Recursiva



In realtà viene spesso utilizzata una versione modificata della rete descritta qui, chiamata ***Long Short Term Memory*** o ***LSTM*** [27], che risolve alcuni problemi tipici di questa architettura ricorsiva tra cui la difficoltà ad apprendere relazioni tra dati di ingresso che compaiono distanziati nel tempo e il rischio di annullamento del gradiente durante il training dovuto al fatto che la rete, per effetto della recursione, risulta essere molto profonda.

Transformer

Le reti recursive ricevono i dati di ingresso sequenzialmente seguendo l'ordine temporale di arrivo. Per questa ragione le dipendenze tra dati distanti nel tempo risultano difficili da modellare. L'architettura *Transformer* è stata concepita proprio per affrontare questo problema [28]. In questo caso non viene utilizzato uno schema recursivo bensì i dati entrano a pacchetti in una rete di tipo feed forward. Il blocco principale, preposto all'analisi della interdipendenza dei dati, è detto *Attention* ed è costituito da un layer di celle di tipo *dense* che hanno il compito di confrontare ogni dato di ingresso con tutti gli altri del pacchetto e di generare un valore che indica la forza della connessione tra i dati in questione.

Più blocchi di tipo *Attention* possono essere inseriti in parallelo per modellare diversi tipi di dipendenze (*Multi-Head Attention*). L'architettura del Transformer contiene un encoder e un decoder, in quanto viene normalmente utilizzata per risolvere task dove i dati di uscita hanno una struttura simile a quelli di ingresso, ad es. traduzione di un testo tra due lingue o sommarizzazione di un testo, ma è possibile utilizzare solo l'encoder per task che richiedono una risposta globale come la *sentiment analysis* o la *classificazione*. L'encoder si basa sulla sovrapposizione di due blocchi, la *Multi-Head Attention* seguita da un blocco *dense* (chiamato *Feed Forward* in Fig. 10). Si notino in Fig. 10 le connessioni di bypass in ogni blocco analogamente a quanto avviene nelle **RESNET**. Più elementi di questo tipo possono essere sovrapposti.

Il decoder è sostanzialmente simile all'encoder, si differenzia per l'aggiunta di un secondo blocco di *Multi-Head Attention* che permette di utilizzare anche i dati di uscita elaborati dalla rete sino al momento attuale. A differenza dell'encoder che processa i dati in un unico passo, il decoder deve effettuare tante iterazioni quanti sono i dati presenti nel pacchetto dove and ogni iterazione i dati

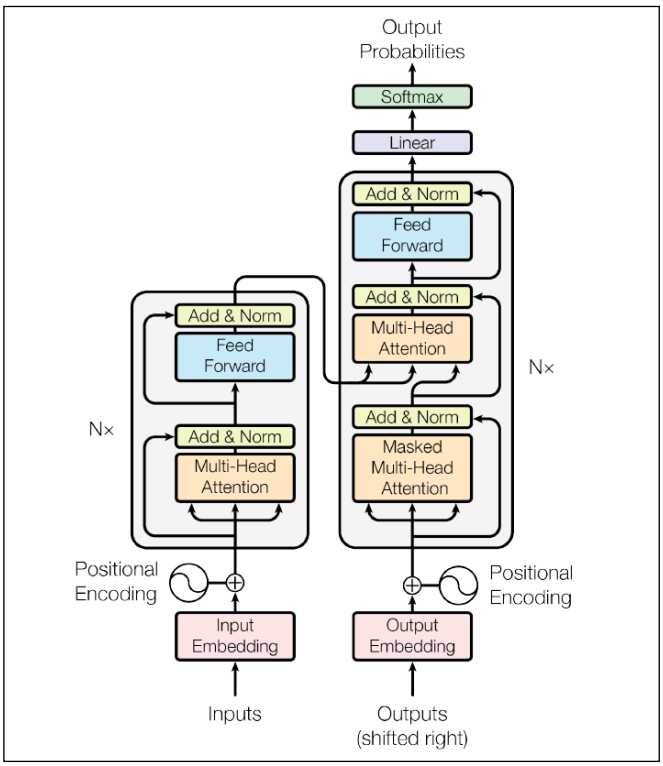


Fig. 10 – Elemento base (encoder e decoder) di una rete Transformer. Tratto da [28]

di uscita generati vengono riportati all'ingresso del decoder scalati di una posizione sino al completamento della sequenza.

Con l'architettura Transformer sono stati migliorati praticamente tutti i benchmark scientifici standard nel campo del *Natural Language Processing*[Nota 2], al costo però di un significativo aumento delle risorse di calcolo richieste rispetto alle soluzioni precedenti. Le reti Transformer più complesse al momento della stesura del presente articolo necessitano dell'apprendimento di un numero di pesi dell'ordine di 10^{11}, non molto distante dal numero di connessioni del cervello umano che è di circa 10^{14}.

Generative Adversarial Network (GAN)

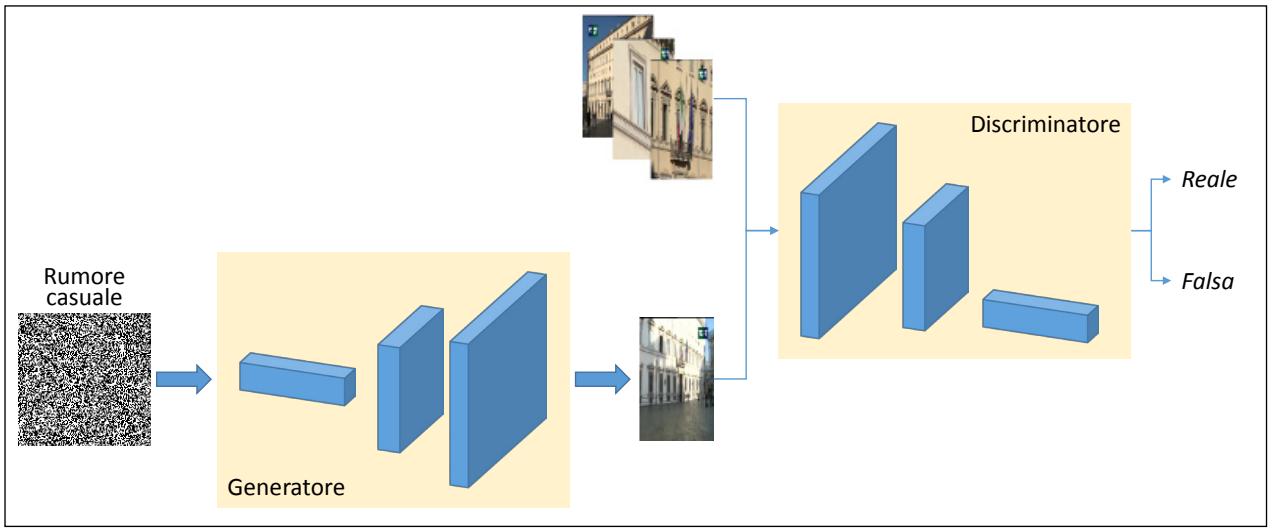
Concludiamo questa breve introduzione alle architetture DNN più diffuse citando le *Generative Adversarial Network*, meglio conosciute con l'acronimo **GAN**. Queste architetture, di cui ne esistono diverse varianti, hanno lo scopo di generare dati originali (cioè non osservati tra i dati di input) ma verosimili.

Durante il training la rete viene condizionata a produrre dati (ad esempio immagini) appartenenti ad una data categoria (ad es. contenenti determinati oggetti) di cui vengono forniti numerosi esempi.

Tecnicamente le **GAN** sono costituite da due reti distinte e speculari che vengono addestrate contestualmente. La prima funge da generatore, la seconda invece ha il compito di discriminare se l'immagine prodotta dal generatore è reale o falsa. La rete discriminante è solitamente di tipo convoluzionale (ad es. una **RESNET**) e viene addestrata a riconoscere le immagini reali contenute nel training set da immagini sintetiche prodotte dalla rete generatrice, la quale viene invece addestrata a produrre immagini che superino il controllo della rete discriminante.

L'architettura della rete generatrice è sostanzialmente inversa a quella di una rete convoluzionale: a partire da un vettore di dati arbitrario, tramite operazioni di convoluzione ed interpolazione si genera la struttura di un'immagine.

Fig. 11 – Generative Adversarial Network



Nota 2 - Il *Natural Language Processing* (NLP) è una branca dell'intelligenza artificiale che si occupa del trattamento automatico di informazioni scritte o parlate espresse in linguaggio naturale. Tra i problemi più comuni affrontati citiamo la trascrizione del parlato in testo, la comprensione di testi scritti, l'analisi grammaticale/sintattica/semantica di testi e la generazione automatica di testi.

L'addestramento avviene utilizzando una loss function di questo tipo [29]:

$$\text{Discriminator Loss} = \text{Max}(D(x) - D(G(z)))$$
$$\text{Generator Loss} = \text{Max}(D(G(z)))$$

Il *discriminatore* cerca di massimizzare la differenza tra la sua uscita in presenza di immagini reali *x* e quella in presenza di immagini generate *G(z)*, mentre il *generatore* cerca di ingannare il discriminatore cercando di generare immagini che ne massimizzino l'uscita. L'addestramento procede a fasi alterne in cui si addestra un componente per volta mantenendo costante l'altro.

Nel caso più semplice di utilizzo, la rete generatrice viene alimentata con un vettore di dati casuali e l'immagine generata viene valutata dalla rete discriminante. Si procede a generare immagini variando il vettore di ingresso sino a che un'immagine generata non ottiene un punteggio sufficientemente alto dalla rete discriminante. In altre varianti di implementazione è invece possibile fornire in ingresso alla rete generatrice un'immagine ed ottenerne in uscita una versione modificata. In questo caso il vettore di ingresso alla rete generatrice non sarà più una sequenza casuale ma l'uscita di una rete convoluzionale applicata all'immagine di ingresso. È possibile a questo modo realizzare reti che effettuano trasformazioni interessanti delle immagini di ingresso, ad es. colorazione di immagini monocromatiche, aumento della risoluzione o rimozione di disturbi.

Conclusioni

In questa trattazione ci si è posti lo scopo di fornire al lettore le nozioni di base per comprendere il funzionamento delle reti neurali profonde, elencando le principali tecniche di ottimizzazione del processo di apprendimento e le principali architetture utilizzate in alcuni dei campi di applicazione più comuni e promettenti. Nei prossimi articoli verranno illustrate alcune applicazioni di particolare interesse per il settore radiotelevisivo e multimediale che, come vedremo, possono ricevere un nuovo impulso da questa promettente tecnologia.

Bibliografia

[1] M. Turing, *Computing Machinery and Intelligence*, in "MIND", vol. LIX, n. 236, 1950, pp. 433-460, **DOI**: 10.1093/mind/LIX.236.433

[2] F. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, Spartan Books, 1961

[3] D. E. Rumelhart, G. E. Hinton e R. J. Williams, *Learning Internal Representations by Error Propagation*, in D. E. Rumelhart e J. L. McClelland (ed.), "Parallel Distributed Processing : Explorations in the Microstructure of Cognition, vol. 1: Foundations", MIT Press,1986, pp. 318-362, https://ieeexplore.ieee.org/servlet/opac?bknumber=6276825

[4] K. Fukushima, *Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position*, in "Biological Cybernetics", vol. 36, 1980, pp. 193-202, **DOI**: 10.1007/BF00344251

[5] D. H. Hubel e T. N. Wiesel, *Receptive fields of single neurones in the cat's striate cortex*, in "The Journal of Physiology", vol. 148, n. 3, 1959, pp. 574-591, **DOI**: 10.1113/jphysiol.1959.sp006308

[6] A. Krizhevsky, I. Sutskever, G. E. Hinton, *ImageNet Classification with Deep Convolutional Networks*, in "Advances in Neural Information Processing Systems 25 (NIPS 2012)", 2012, pp. 1097-1105, https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[7] Y. Lecun ed altri, *Gradient Based Learning Applied to Document Recognition*, in "Proceedings of the IEEE", vol. 86, n. 11, 1998, pp. 2278-2324, **DOI**: 10.1109/5.726791

[8] *CUDA Toolkit home page*, NVIDIA. Developer (web), https://developer.nvidia.com/cuda-toolkit (ultimo accesso 27/10/2020)

[9] *State of AI Report 2020*, stateof.ai (web), https://www.stateof.ai/ (ultimo accesso 27/10/2020)

[10] *TensorFlow home page*, TensorFlow (web), https://www.tensorflow.org/ (ultimo accesso 27/10/2020)

[11] PyTorch home page, PyTorch (web), https://pytorch.org/ (ultimo accesso 27/10/2020)

[12] *The latest in Machine Learning*, paperswithcode.com (web), https://paperswithcode.com/ (ultimo accesso 27/10/2020)

[13] F. Rosenblatt, *The Perceptron, a Perceiving and Recognizing Automaton*, Cornell Aeronautical Laboratory, Report n. 85-460-1, 1957

[14] G. Cybenko, *Approximation by superpositions of a sigmoidal function*, in "Mathematics of Control, Signals and Systems", vol. 2, n. 4, 1989, pp. 303-314, **DOI**: 10.1007/BF02551274

[15] M. Leshno ed altri, *Original Contribution: Multilayer feedforward networks with a nonpolynomial activation function can approximate any function*, in "Neural Networks", vol. 6, n. 6, 1993, pp. 861-867, **DOI**: 10.1016/S0893-6080(05)80131-5

[16] H. Zhang, *The Optimality of Naive Bayes*, in "Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference", 2004

[17] C. Cortes e V. Vapnik, *Support-vector networks*, in "Machine Learning ", vol. 20, n. 3, 1995, pp. 273–297, **DOI**: 10.1007/BF00994018

[18] Robbins e S. Monro, *A Stochastic Approximation Method*, in "Annals of Mathematical Statistics", vol. 22, n. 3, 1951, pp. 400-407, https://projecteuclid.org/euclid.aoms/1177729586

[19] S. Ruder, *An overview of gradient descent optimization algorithms*, arXiv preprint , 2017, arXiv:1609.04747v2

[20] N. Qian, *On the momentum term in gradient descent learning algorithms*, in "Neural networks", vol. 12, n. 1, 1999, pp. 141-151, **DOI**: 10.1016/S0893-6080(98)00116-6

[21] S. Ioffe e C. Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*, in "Proceedings of the 32nd International Conference on Machine Learning", PMLR, vol. 37, 2015, pp. 448-456, http://proceedings.mlr.press/v37/ioffe15.html

[22] *About ImageNet*, ImageNet (web), http://imagenet.stanford.edu/about-overview (ultimo accesso 27/10/2020)

[23] P. Krishnamurthy, *Understanding data bias*, in "towards data science" (web), 2019, https://towardsdatascience.com/survey-d4f168791e57 (ultimo accesso 27/10/2020)

[24] A. Nguyen, J. Yosinski e J. Clune, *Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks*. in ICML Visualization Workshop, 2016, https://icmlviz.github.io/icmlviz2016/assets/papers/5.pdf

[25] Kaiming He, *Deep Residual Learning for Image Recognition*, in "2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)", 2016, **DOI**: 10.1109/CVPR.2016.90

[26] Mingxing Tan e Quoc V. L, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*, in "Proceedings of the 36th International Conference on Machine Learning", PLMR, vol. 97, 2019, pp. 6105-6114, http://proceedings.mlr.press/v97/tan19a.html

[27] S. Hochreiter e J. Schmidhuber, *Long Short-term Memory*, in "Neural Computation", vol. 9, n. 8, 1997, pp. 1735-1780, **DOI**: 10.1162/neco.1997.9.8.1735

[28] A. Vaswani ed altri, *Attention Is All You Need*, in "Advances in Neural Information Processing Systems 30 (NIPS 2017)", 2017, pp. 5998-6008, https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[29] M. Arjovsky, S. Chintala, L. Bottou, *Wasserstein Generative Adversial Networks*, in "Proceedings of the 34th International Conference on Machine Learning", PMLR, vol. 70, 2017, pp. 214-223, http://proceedings.mlr.press/v70/arjovsky17a.html

Le applicazioni dell'Intelligenza Artificiale

Una breve rassegna generale

AA. VV.
Rai - Centro Ricerche, Innovazione Tecnologica e Sperimentazione

Il presente articolo vuole introdurre e illustrare sinteticamente le principali aree applicative dell'Intelligenza Artificiale che hanno un diretto impatto sulla catena del valore dell'industria radiotelevisiva e multimediale.

Si inizia con un contributo che illustra attraverso due mappe e un glossario sintetico il panorama generale dei metodi e domini applicativi dell'IA. Si prosegue con la trattazione di quattro aree fondamentali: la trascrizione del parlato (ASR – Automatic Speech Recognition), la visione artificiale (CV – Computer Vision), l'elaborazione del linguaggio naturale (NLP – Natural Language Processing) e infine la generazione artificiale dei contenuti.

L'introduzione delle tecnologie dell'*Intelligenza Artificiale (IA)* nei processi industriali è una tematica di ampia complessità, che richiede non solo la conoscenza tecnica dei metodi ma anche una profonda conoscenza dei processi di produzione e di business e di come introdurre in essi i necessari elementi di innovazione. In questo contesto è fondamentale da una parte capire quali domini applicativi dell'IA rappresentano la sorgente dell'innovazione e dall'altra quali processi ne possano beneficiare ed in quale misura.

Sulla base di queste considerazioni, il presente articolo vuole introdurre e illustrare sinteticamente le principali aree applicative dell'Intelligenza Artificiale che hanno un diretto impatto sulla catena del valore dell'industria radiotelevisiva e multimediale. Si inizia con un contributo che illustra attraverso due mappe e un glossario sintetico il *panorama generale dei metodi e domini applicativi dell'Intelligenza Artificiale*. Si prosegue con la trattazione di quattro aree fondamentali: la *trascrizione del parlato* (*ASR – Automatic Speech Recognition*), la *visione artificiale* (*CV – Computer Vision*), l'*elaborazione del linguaggio naturale* (*NLP – Natural Language Processing*) e infine la *generazione artificiale dei contenuti*. Ciascuno di questi quattro domini ha impatti trasversali sulla catena del valore poiché i risultati dei metodi illustrati producono i dati fondamentali sui quali si possono costruire le applicazioni e i processi industriali specifici, che saranno invece discussi negli articoli successivi. Dopo una breve contestualizzazione, ciascuno dei contributi di questo articolo dà risalto alle recenti evoluzioni basate sulle architetture delle *reti neurali profonde (DNN)* evidenziando come la rivoluzione tecnologica e metodologica introdotta da esse rappresenti un punto di svolta. Naturalmente, come è nello spirito di questo numero speciale ed in linea con l'approccio editoriale della rivista, i contributi hanno l'intenzione di introdurre al lettore le definizioni, le problematiche e le sfide relative a queste tecniche, rimandando invece gli approfondimenti alle rispettive sezioni bibliografiche.

201_01.2_1.00

Metodi e domini applicativi dell'Intelligenza Artificiale

Contributo a cura di Paolo **Casagranda** e Alberto **Messina**
Rai - Centro Ricerche, Innovazione Tecnologica e Sperimentazione

La ricerca nel campo dell'*Intelligenza Artificiale* (**AI**, *Artificial Intelligence*) è stata estremamente prolifica negli ultimi anni. Accanto a metodi ormai classici se ne sono affermati altri, come le *Reti Neurali Convoluzionali*, che in breve tempo hanno trovato impiego in molte applicazioni.

Le seguenti mappe sintetizzano, in modo necessariamente semplificativo, alcuni dei più importanti *metodi* e *domini applicativi* dell'Intelligenza Artificiale, con la finalità di orientare il lettore alla terminologia e ai riferimenti che saranno fatti nei contributi specifici che seguiranno.

> *A corredo delle mappe, per comodità del lettore, vengono forniti due brevi glossari con alcuni termini in esse utilizzati. Viene fornita anche una bibliografia essenziale nell'ambito dell'Intelligenza Artificiale e dei Learning System.*

Mappa dei metodi

Nella *mappa dei metodi* (Fig. 1) sono presenti algoritmi singoli particolarmente importanti (ad es. le *Support Vector Machines*) e gruppi di algoritmi (*Recommender Systems* o *Clustering*).

Accanto alla divisione in I*ntelligenza Artificiale classica* e *Machine Learning* (**ML**), si sono indicati anche i diversi paradigmi di applicazione degli algoritmi (*Programmed*, *Supervised*, *Reinforcement Learning*, *Unsupervised*) che si applicano trasversalmente.

Non trovano posto nella mappa alcuni metodi fondamentali utilizzati trasversalmente come strumenti (es. *metodi di gradient descent* o *Expectation Maximization*), così come non sono indicati neppure i metodi di preparazione e pulizia dei dati.

Mappa dei domini applicativi

La *mappa dei domini applicativi* (Fig. 2) individua alcuni dei più importanti domini applicativi dell'Intelligenza Artificiale.

La famiglia applicativa della *Knowledge Representation and reasoning* si occupa di tecniche per la rappresentazione strutturata della conoscenza e dell'applicazione di metodi di ragionamento automatico per inferirne di nuova.

La famiglia applicativa del *Language Processing* si occupa delle tecnologie e dei metodi atti alla comprensione, alla traduzione e all'elaborazione del linguaggio naturale.

Le tecnologie di *Computer Vision* sono finalizzate a realizzare metodi di percezione e comprensione automatica delle immagini sia statiche che in movimento.

La famiglia degli *Agenti* include tecniche per lo sviluppo di assistenti software in grado di eseguire azioni e piani per conto di un operatore umano.

Infine, i sistemi di *Information Retrieval and Filtering* comprendono le tecnologie per cercare, filtrare e personalizzare le informazioni.

Bibliografia essenziale

[1] S. J. Russell e P. Norvig, *Artificial Intelligence. A Modern Approach*, Prentice Hall, 2010, 3^ Edizione, **ISBN**: 9780136042594

[2] E. Alpaydin, *Introduction to Machine Learning*, MIT Press, 2014, 3^ Edizione, **ISBN**: 9780262028189

[3] I. Goodfellow, Y. Bengio e A. Courville, *Deep learning*, MIT Press, 2016, **ISBN**: 9780262035613

201_01.2.1_3.00

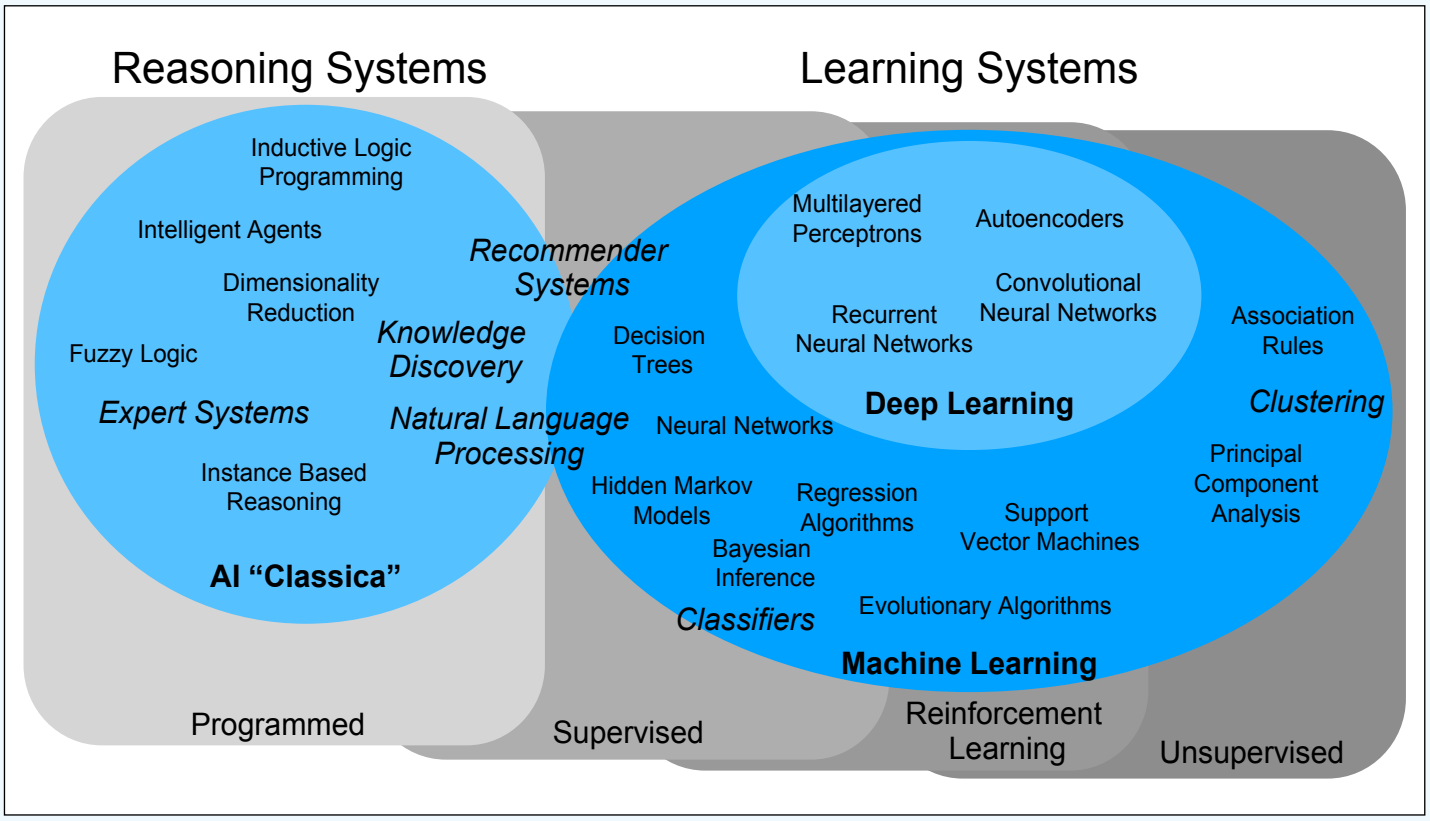


Fig. 1 – Mappa dei metodi utilizzati dall'Intelligenza Artificiale

Glossario - Mappa dei metodi

- ***Reasoning Systems***: sistemi che giungono a conclusioni con metodi induttivi o deduttivi
- ***Learning Systems***: sistemi che generano conclusioni utilizzando grandi moli di dati, e resi possibili dai metodi di Machine Learning
- ***Programmed*** AI: metodi classici che non utilizzano machine learning
- ***Supervised*** ML: metodi che richiedono l'intervento umano per classificare un sottoinsieme (eventualmente molto piccolo) dei dati di input.
- ***Reinforcement Learning***: metodi in cui il risultato viene trovato senza specificare il metodo, utilizzando un obiettivo e un sistema di ricompense per indirizzarlo
- ***Unsupervised*** ML: metodi in grado di trovare pattern o risultati senza l'intervento umano. Ad esempio gli *algoritmi di clustering*.

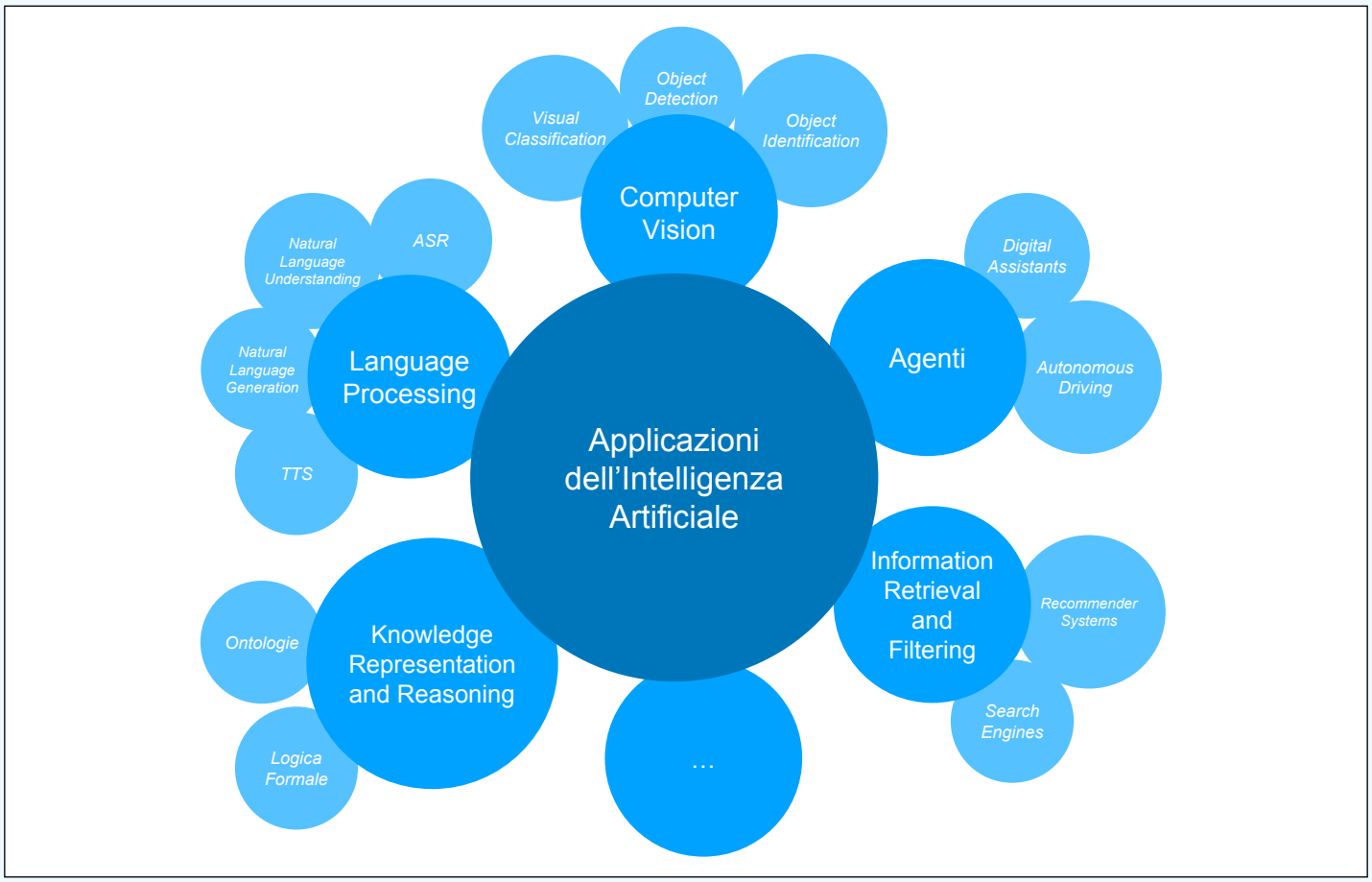


Fig. 2 – Mappa dei domini applicativi dell'Intelligenza Artificiale

Glossario - Mappa dei domini applicativi

- ***Ontologies***: una specifica esplicita e formale di una concettualizzazione condivisa. Ad esempio una lista di oggetti o entità con le loro proprietà e relazioni reciproche.
- ***Digital Assistants***: un agente virtuale che interpreta le richieste di un utente ed esegue di conseguenza un'azione. Necessita di molteplici componenti (ASR, TTS, Natural Language Understanding, Sistemi di raccomandazione...)
- ***Autonomous Driving***: automatizzazione di parte dell'operazione di guida. Esistono diversi livelli di autonomous driving: dal Livello 1 che consiste nell'assistenza automatica di alcune operazioni (ad es. frenata automatica per rischio di collisione) al Livello 5, che specifica una piena autonomia di guida del veicolo.
- ***ASR***: processo che consente la traduzione del linguaggio umano parlato in testo (*Automatic Speech Recognizer)*
- ***TTS***: sintesi vocale a partire da un testo (*Text-to-Speech*)
- ***Natural Language Generation***: generazione automatica di linguaggio umano
- ***Natural Language Understanding***: comprensione del linguaggio umano
- ***Visual Classification***: descrizione e classificazione automatica di elementi visuali
- ***Object Detection and Identification***: riconoscimento automatico di oggetti (volti, insegne, auto)
- ***Recommender Systems***: famiglia di tecnologie volte a suggerire informazioni
- ***Search Engines***: i motori di ricerca

Evoluzione dei sistemi di Automatic Speech Recognition (ASR)

Contributo a cura di Giorgio **Dimino**
Rai - Centro Ricerche, Innovazione Tecnologica e Sperimentazione

L'ambizione di realizzare macchine intelligenti in grado di interagire con le persone secondo le modalità proprie degli esseri umani, come rispondere a comandi vocali e interloquire in linguaggio naturale, è sempre stata molto forte, sin dagli albori dell'informatica. La capacità di trascrivere il parlato in testo è un mattone fondamentale in questa direzione, insieme all'interpretazione del linguaggio naturale e alla sintesi vocale.

I primi tentativi di realizzare macchine in grado di riconoscere il parlato risalgono agli anni '60. ***IBM*** ideò un sistema denominato **Shoebox** che riusciva a riconoscere numeri e semplici comandi come *"plus"* e *"total"* [1]. Questo è stato probabilmente il primo tentativo di individuare dal segnale audio i *formanti* che costituiscono la base del parlato. Altri centri di ricerca nel mondo, principalmente in Giappone e Regno Unito, si dedicarono a studi simili, ma nessuno di questi con la tecnologia disponibile a quei tempi portò alla realizzazione di sistemi che potessero avere un qualche utilizzo pratico, per cui ben presto le ricerche furono messe in pausa. Verso metà degli anni '80 l'intuizione di modellare l'articolazione delle parole tramite *Hidden Markov Models (HMM)* [2], un particolare tipo di grafo di cui parleremo nel seguito di questo articolo, fornì il collante per aggregare le ricerche effettuate sino a quel momento sul riconoscimento dei *fonemi* e finalmente furono realizzati i primi sistemi in grado di trascrivere intere parole e frasi, riaccendendo l'entusiasmo nei ricercatori del settore. Pochi anni dopo, ad inizio anni '90, fu presentato il primo sistema di dettatura commerciale denominato **Dragon Dictate** [3], un sistema di costo elevato rivolto a professionisti, che pochi anni dopo col nome di **Dragon Naturally Speaking** divenne un software alla portata di tutti che poteva essere installato su qualsiasi pc Windows. Nello stesso periodo la ricerca ha fatto passi da gigante nell'ottimizzazione dei sistemi di *Automatic Speech Recognition (ASR)* basati su *HMM*, raggiungendo nel giro di un decennio l'apice dello stato dell'arte della tecnologia, che in condizioni acustiche ideali permette di trascrivere un parlato continuo con una precisione vicina al *95%*. Purtroppo diversi problemi rimangono irrisolti, tra cui l'estrema variabilità della precisione dei sistemi in funzione delle condizioni ambientali di cattura del suono (ad es. rumore o voci sovrapposte) e la difficoltà ad addestrare sistemi allo stato dell'arte per lingue poco diffuse, sia per la scarsità di risorse che per la limitata profittabilità del mercato. Negli anni recenti, poiché l'approccio basato sulle *HMM* ha ormai ampiamente raggiunto i sui limiti, la ricerca si è rivolta alla sperimentazione di sistemi basati sulle *reti neurali profonde (Deep Neural Network, DNN)*, ma solo nel 2020 alcuni sistemi sono riusciti a migliorare ulteriormente lo stato dell'arte, riuscendo anche a mitigare alcune delle problematiche che affliggono i sistemi *HMM*.

Campi di applicazione dei sistemi ASR

Sebbene intuitivamente i campi di applicazione degli *ASR* possano essere innumerevoli, non sempre è possibile realizzare sistemi sufficientemente performanti per essere efficacemente impiegati in un ambiente produttivo.

Un campo di applicazione dove gli *ASR* hanno avuto un certo successo già dalle prime implementazioni è il *riconoscimento di comandi vocali*. I sistemi più semplici si basano sul riconoscimento di un insieme limitato di parole singole o frasi brevi relative ad un contesto ben definito e con un dizionario limitato. Fra questi possiamo citare i risponditori automatici di alcuni call center o le interfacce vocali presenti sui sistemi di infotainment delle auto, sino ai più recenti assistenti vocali come **Alexa** di ***Amazon*** o **Google Assistant**. Questi ultimi sono basati su modelli di riconoscimento allo stato dell'arte ma che richiedono risorse disponibili solo in cloud per funzionare.

201_01.2.2_1.00

Un secondo campo di applicazione è quello della *dettatura automatica*, efficace già dagli anni '90 soprattutto se impiegato in contesti specifici, grazie alla possibilità di personalizzare il modello di riconoscimento sul timbro vocale del parlatore e su un dizionario tarato sul contesto applicativo. Oggi è disponibile in tutti i principali sistemi operativi dei sistemi informatici sia fissi che mobili e grazie agli avanzamenti tecnologici non è più necessario l'adattamento del modello al parlatore. La sua applicazione è comunque limitata dalla necessità di correggere il testo trascritto dai non infrequenti errori.

Un altro campo di applicazione è la *rendicontazione automatica* di sedute e riunioni. Questa applicazione risulta essere particolarmente sfidante in quanto sono presenti tutte le caratteristiche del parlato che ancora oggi rappresentano per i sistemi degli ostacoli insormontabili, ovvero linguaggio spontaneo (meno strutturato di quello scritto), voci sovrapposte, rumore ambientale, inflessioni dialettali e parlatori non di madre lingua [4].

Nel campo multimediale una delle applicazioni principali è sicuramente la *sottotitolazione automatica*, oggetto di un altro contributo di questa raccolta, che presenta parecchie analogie con la rendicontazione.

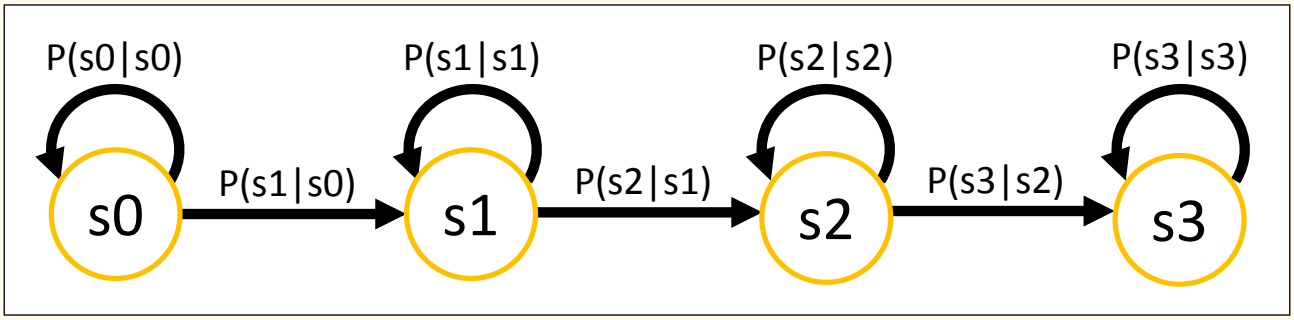
Un'altra applicazione estremamente interessante abilitata dall'*ASR* è la *classificazione e indicizzazione dei contenuti* basata sulla trascrizione del parlato. Infatti per alcuni generi, principalmente nel campo delle news e dei documentari, la maggior parte del contenuto semantico del programma è espressa dalla narrazione. La trascrizione del parlato permette, quindi, l'indicizzazione e la classificazione del contenuto tramite ricerca di parole chiave nel testo trascritto, che a sua volta permette di individuare immediatamente il segmento di video in cui le suddette parole vengono pronunciate. Inoltre, tramite strumenti di analisi *NLP (Natural Language Processing)* del trascritto è possibile estrarre concetti di più alto livello (es. *named entites*, *classificazione*, *argomento*, *sommario*) [5].

Cenni sui principi di funzionamento dei sistemi ASR

Sistemi ASR basati su HMM

La variabilità del suono emesso da persone diverse quando pronunciano una data frase rende la realizzazione di un modello del parlato di validità generale un compito complesso. Infatti il modello deve essere invariante a vari fattori tra cui il timbro vocale e il genere del parlatore, le inflessioni prosodiche e gli accenti dialettali. L'intuizione vincente dei sistemi basasti su *HMM* è stata quella di modellare la sequenza di *emissioni vocali elementari*, che senza addentrarci nei tecnicismi della fonetica chiameremo d'ora in avanti *foni*, con un *modello markoviano*. Un modello si dice markoviano quando la probabilità di transizione da uno stato all'altro dipende solo dallo stato di partenza e non dalla storia passata [6], come schematizzato in Fig. 1. Negli *HMM*, oltre alla probabilità di transizione verso ciascuno stato successivo, a ciascuno stato si associa anche la *probabilità di emissione di variabili* cosiddette *osservabili*. Queste probabilità devono essere apprese dal modello durante una fase di addestramento.

Fig. 1 – Esempio di *modello markoviano*



Nel caso del riconoscimento del parlato lo scopo del modello è trovare la *sequenza di parole* W^* più probabile tra tutte le sequenze W di parole del dizionario data la *sequenza di emissioni vocali* X, espresso matematicamente dalle formule seguenti:

$$W^* = argmax_W P(W|X) \quad (1)$$

ovvero, utilizzando il *teorema di Bayes*:

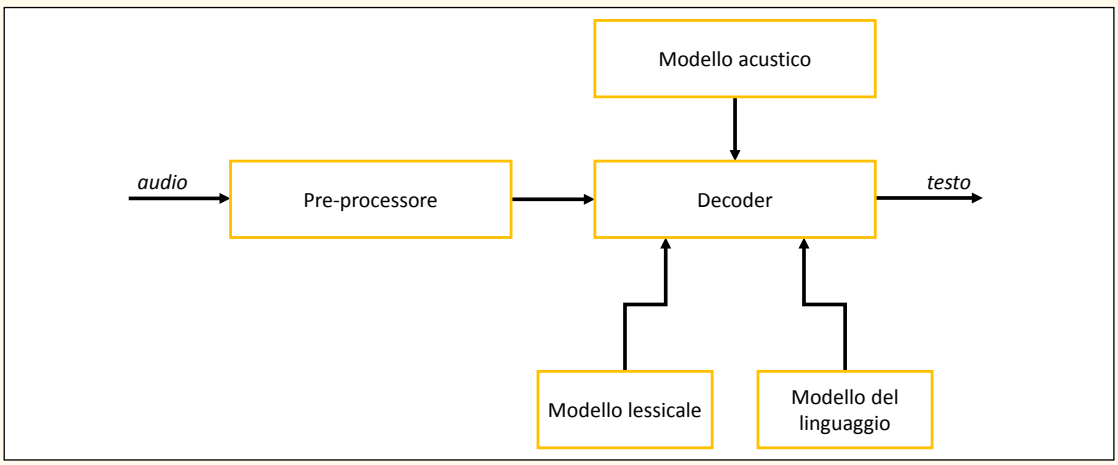
$$W^* = argmax_W P(X|W) * P(W) \quad (2)$$

dove $P(X|W)$ è un *modello acustico* che, data una sequenza di parole del dizionario, ne modella la sequenza di emissioni acustiche corrispondenti e $P(W)$ è un *modello del linguaggio*. È quindi possibile realizzare un sistema di riconoscimento del parlato componendo un *modello acustico* che funga da trasduttore tra porzioni di segnale audio opportunamente pre-processato e i fonemi della lingua in questione, seguito da un *modello lessicale* che metta in relazione i fonemi con le parole contenute nel dizionario ed un *modello del linguaggio* che guidi la selezione tra le varie ipotesi effettuate dal modello acustico verso quella che è statisticamente più probabile da un punto di vista linguistico [7]. L'intero processo è descritto in Fig. 2.

Per distillare dal segnale acustico emesso solo le caratteristiche utili alla comprensione del linguaggio, è necessario applicare un processamento che renda il segnale il più possibile indipendente dal parlatore. Inoltre è utile eliminare, per quanto possibile, suoni estranei che potrebbero essere stati catturati involontariamente dal microfono. Schematicamente tale processamento consiste nella sequenza dei seguenti passi:

- il segnale viene diviso in finestre di analisi da *25 ms*, periodo in cui il segnale può considerarsi stazionario dal punto di vista del linguaggio;
- viene elaborata una finestra di analisi ogni *10 ms*, in modo da avere una sovrapposizione che eviti discontinuità tra una finestra e la successiva;
- viene calcolato lo spettro del segnale contenuto in ciascuna finestra tramite la *Fast Fourier Transform* e applicata una mappatura sulla *scala MEL* [8] che fornisce una rappresentazione spettrale proporzionale alla percezione dell'orecchio umano;
- l'ampiezza di ciascuna riga spettrale viene convertita in scala logaritmica per simulare i processi coinvolti nella percezione umana;
- tramite la *trasformata coseno discreta (DCT)* si calcola una grandezza detta *cepstrum* [9] di cui si utilizzano i primi *12 coefficienti* che sono in stretta relazione con i *formanti*, inviluppi spettrali che caratterizzano le vocali e che sono invarianti tra parlatori;
- le caratteristiche utilizzate dal modello di riconoscimento includono i *12 coefficienti cepstrum* ed una *misura dell'energia complessiva della finestra*. Si aggiungono anche le derivate prima e seconda di questi tredici coefficienti per un totale di *39 valori*.

Fig. 2 – Schema di un *ASR* basato su *HMM*



Il processo appena descritto è mostrato schematicamente in Fig. 3. Il segnale così processato (blocchi di *39 valori* corrispondenti a *25 ms* di segnale) costituisce l'ingresso per il *modello acustico*, che è un trasduttore che ha il compito di convertire il segnale acustico nei foni corrispondenti.

Il *modello acustico* si basa su di un modello markoviano con stati nascosti (*HMM*), ovvero noti ma non osservabili direttamente, che rappresentano i foni. A ciascuno stato, quale variabile osservabile, è associata una *mistura di gaussiane (GMM)* che modella la probabilità di osservazione di una data combinazione di caratteristiche audio calcolate con il metodo del *MEL cepstrum* nello stato attuale. Il modello avanza ad ogni finestra di analisi, ovvero *10 ms*. Poiché il segnale audio non è stabile per l'intera durata dell'emissione di un fono, ciascun fono viene modellato da *3 stati*, inoltre l'emissione acustica dipende anche dal fono che precede e da quello che segue. Ne consegue che il *numero di stati* del modello markoviano è idealmente $3*(n\_foni)^3$, dove *n_foni* è il numero di foni distinti del linguaggio, solitamente da *30* a *50* in dipendenza della lingua in questione. Per ricondurre la complessità del modello a dimensioni accettabili viene effettuata una fusione degli stati le cui misture di gaussiane risultano simili. Sia la composizione delle *GMM* che le probabilità associate ai nodi e agli archi del modello *HMN* devono essere appresi durante una fase di addestramento supervisionato che richiede la disponibilità di un corpus costituito da file audio contenenti esempi di parlato e relativa trascrizione. La dimensione del corpus per ottenere risultati allo stato dell'arte deve essere in genere nell'ordine delle centinaia se non migliaia di ore, in funzione della complessità della lingua e del contesto applicativo. La difficoltà principale nel processo di addestramento risiede nel fatto che l'allineamento del testo con le emissioni vocali a livello di singolo fono non è solitamente disponibile, ma deve essere dedotto contestualmente all'addestramento delle misture di gaussiane. Si procede quindi per fasi, migliorando alternativamente il modello di *GMM* di ciascun fono e l'allineamento tra finestre audio e stati del modello *HMM* secondo l'algoritmo *forward/backward* [2].

Il *modello acustico* genera delle ipotesi relative alla sequenza di foni riconosciuti. Queste ipotesi (o meglio le più probabili) vengono pesate da un *modello lessicale* che ha il compito di mappare sequenze di foni con le parole appartenenti al dizionario di riferimento. Il modello lessicale viene assemblato a mano o tramite regole e la sua complessità dipende principalmente dalle caratteristiche della lingua in oggetto. Nel caso dell'italiano la corrispondenza tra grafemi e foni è molto stretta, quindi il modello lessicale viene derivato dal dizionario tramite sem-

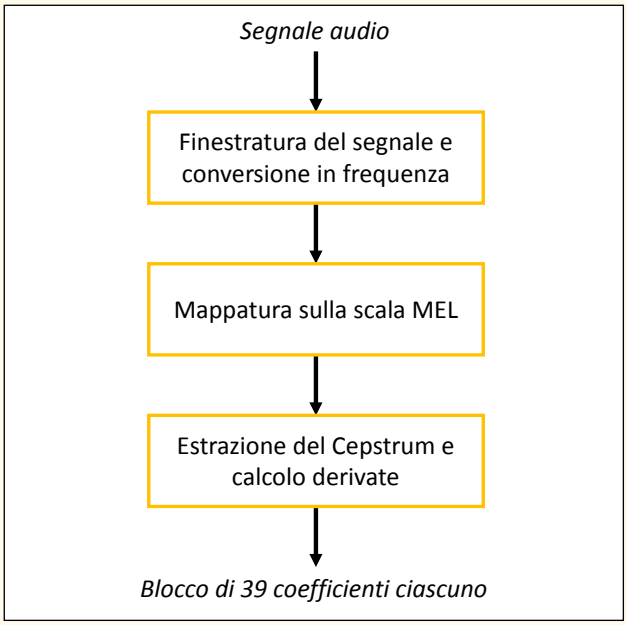


Fig. 3 – Pre-processamento del segnale audio

plici regole, mentre in altre lingue la corrispondenza tra foni e grafemi non è facilmente codificabile e la composizione del modello lessicale è più laboriosa.

Per fare sì che la sequenza di parole riconosciute abbia anche una coerenza a livello semantico si impiega un terzo modello, il *modello linguistico*, che si basa normalmente sull'applicazione di un modello probabilistico di sequenze di parole (*n_grammi*) e fornisce la probabilità condizionata che ciascuna parola del dizionario compaia date le ultime *n-1* parole rilevate. Per costruirlo è necessario solo disporre di una quantità elevata di testi nella lingua desiderata, meglio se inerenti al contesto in cui il sistema verrà impiegato. Anche i modelli lessicale e linguistico sono modellati tramite *HMM*.

Riassumendo, per applicare il modello complessivo ad un'emissione vocale dobbiamo applicare tre livelli di *HMM*: da *misture di gaussiane* a *foni*, da *foni* a *parole*, da *parole* a *n-grammi*. Ciascuno stato e ciascuna possibile transizione da uno stato al successivo ha associata una probabilità. L'uscita del modello sarà la sequenza di parole che ha associata la *probabilità cumulativa massima*.

Per calcolare la probabilità cumulativa massima della sequenza secondo la formula *(2)* viene generalmente utilizzata la *decodifica di Viterbi* [10], ma diverse ottimizzazioni devono essere applicate per rendere il processo computazionalmente efficiente, in quanto il grafo risultante dalla composizione dei tre modelli cresce combinatorialmente. Solitamente si riduce la complessità della ricerca seguendo solo i percorsi con una probabilità associata superiore ad una data soglia (oppure gli *N* percorsi più probabili) secondo l'algoritmo denominato *beam-search* [11].

Come si può facilmente intuire dalla breve descrizione fornita del modello *HMM*, la realizzazione di un sistema di *ASR* allo stato dell'arte con questa tecnologia è un processo complesso che richiede notevoli sforzi sia per il reperimento delle risorse linguistiche necessarie all'addestramento dei modelli sia per le ottimizzazioni richieste al processo per adattarlo alle peculiarità di ciascuna lingua.

Sistemi ASR basati su Reti Neurali

Le *reti neurali (DNN)* hanno la caratteristica di poter essere utilizzate con successo in contesti eterogenei senza necessitare di una conoscenza approfondita del dominio specifico. È quindi naturale che la comunità scientifica si sia rivolta in questa direzione per superare alcune delle limitazioni imposte dall'approccio basato su *HMM*, sia per migliorare la prestazione dei sistemi di *ASR* oltre lo stato dell'arte di quella tecnologia sia per semplificare il processo di addestramento (e quindi ridurne i costi associati) per renderne l'utilizzo più facilmente estendibile a lingue parlate da comunità ristrette e ad ambiti applicativi verticali.

L'introduzione delle reti neurali è avvenuta secondo due filosofie differenti, l'*approccio ibrido*, che consiste nell'inserire alcuni moduli basati su *DNN* nel modello classico *HMM*, e l'*approccio end-to-end* che parte dalla progettazione di una nuova architettura interamente basata su *DNN* e addestrabile complessivamente in un unico ciclo.

Ovviamente l'*approccio ibrido*, partendo da una base già molto ottimizzata, ha dato risultati più rapidamente. Un primo contributo delle reti neurali è stato la sostituzione del *modello linguistico* basato su *n-grammi* con una *rete transformer* [12], che è attualmente l'architettura di riferimento nel campo del *Natural Language Processing*. Un secondo contributo importante proviene dalla sostituzione del *modello acustico GMM/HMM* con una *rete recursiva LSTM* (*Long Short Term Memory*) [13]. In questo caso l'addestramento viene effettuato in due fasi:

1. nella prima si usa un *modello GMM/HMM* classico per generare l'allineamento a livello di singoli foni del segnale audio con il testo corrispondente;
2. nella seconda il data set allineato viene usato per addestrare la *LSTM* tramite una *loss function* di tipo *cross-entropy* [Nota 1].

Nota 1 - Descritta nel contributo *"Introduzione alle moderne tecniche di Intelligenza Artificiale"* in questo stesso numero della rivista.

L'approccio ibrido ha permesso di raggiungere un *Word Error Rate (WER)* sul benchmark accademico *LibriSpeech "other"* [Nota 2] del *10,7%* usando un modello linguistico a *4-grammi* e del *5,7%* con un modello linguistico *transformer* [14].

L'*approccio end-to-end* ha richiesto parecchi anni di sperimentazioni [15] prima di riuscire a mostrare il suo potenziale, ma oggi rappresenta lo stato dell'arte indiscusso. Parecchie architetture sono state proposte in letteratura e apparentemente le prestazioni ottenibili sono simili.

La difficoltà di una rete end-to-end consiste nel riuscire ad addestrare da zero reti profonde con decine o centinaia di milioni di parametri su di un compito, che, come abbiamo visto descrivendo il modello markoviano, richiede diversi livelli di astrazione per trasformare un segnale acustico in un testo scritto. Il fatto di riuscire a farlo senza passi intermedi sembrava sino a pochi anni fa un obiettivo troppo ambizioso, mentre oggi non solo è stato raggiunto ma addirittura è possibile farlo con una quantità di dati annotati di due ordini di grandezza inferiore rispetto ai modelli precedenti.

Uno dei grandi vantaggi che le *DNN* offrono rispetto ad altri approcci di intelligenza artificiale è la sostanziale indipendenza della rete dal contesto applicativo, per cui soluzioni elaborate per risolvere problemi in altri domini, come la visione artificiale o l'*NLP*, possono essere immediatamente impiegate anche qui e viceversa. Il problema principale da risolvere nel progettare una rete in grado di fungere da trasduttore tra un segnale audio e un testo corrispondente è definire una opportuna *loss function*, in quanto è necessario trovare una funzione matematica derivabile che tenga in conto la corrispondenza temporale tra le singole finestre di analisi audio e le lettere (o le sillabe) che andranno a comporre il testo scritto risultante. Una loss function che viene spesso adottata è la *Connectionist Temporal Classifier (CTC)* [16], sviluppata per risolvere il problema di addestrare reti neurali per il riconoscimento di testi scritti. Vediamo il principio di funzionamento della *CTC*. L'architettura utilizzata più spesso per realizzare reti neurali per *ASR* si basa sulla concatenazione di un primo gruppo di livelli convoluzionali che ha il compito di analizzare le finestre di campioni audio, di dimensione di circa *20-25 ms* ciascuna, per analogia con la percezione umana del parlato, seguito da un secondo gruppo di livelli che implementano una *rete recursiva* (oppure un *transformer*), che serve a sfruttare le relazioni temporali tra le finestre audio vicine ed emettere un vettore bidimensionale che rappresenta una stima della probabilità che ciascuna finestra contenga una data lettera [Nota 3] .

All'insieme delle lettere dell'alfabeto viene aggiunto il simbolo speciale "-" che funge da separatore e che verrà poi eliminato nel testo finale, ma che è utile per riconoscere le lettere ripetute (ad es. le doppie consonanti nella lingua italiana). Non conoscendo l'allineamento del testo con le singole finestre di analisi, si calcola la probabilità che la rete abbia rico-

Nota 2 - *LibriSpeech* è un benchmark spesso utilizzato dalla comunità scientifica negli anni recenti per misurare le prestazioni dei modelli *ASR* in modo da potere effettuare delle comparazioni tra approcci diversi. Si tratta di un data set composto da una collezione di audiolibri e testi associati in lingua inglese liberamente scaricabili da Internet e facenti parte del progetto open source **LibriVox**. Il data set contiene circa 1000 ore di parlato con il testo corrispondente allineato temporalmente a livello di frase. Viene solitamente diviso in due partizioni denominate *"clean"* e *"other"* che presentano difficoltà diverse a causa delle diverse condizioni di registrazione (*"other"* è quello più sfidante).

Nota 3 - Si noti che questo approccio si discosta notevolmente da quelli visti sino ad ora in quanto non vi è alcun tentativo di modellare esplicitamente la costruzione del linguaggio tramite foni. L'ingresso della rete è costituito dal segnale audio, suddiviso in finestre di dimensione costante, e in uscita dalla rete si ottiene direttamente la probabilità associata alle lettere che compongono il testo trascritto

nosciuto la sequenza come somma delle probabilità associate a tutti i percorsi che contengono il testo corretto, decodificato secondo lo schema seguente:

supponiamo che il testo da riconoscere sia la parola "casa", *all'interno del segnale audio di lunghezza equivalente a 10 finestre di analisi. Per ciascuna finestra vengono considerate solo le probabilità associate alle lettere contenute nel testo. Poiché a ciascuna lettera per ciascuna finestra è associata una probabilità calcolata dalla rete, è immediato calcolare la probabilità associata a ciascun percorso che viene decodificato con la parola* "casa" *moltiplicando le probabilità di ciascuna lettera. La probabilità finale sarà data dalla somma di tutti i percorsi che si decodificano in* "casa". *La decodifica avviene eliminando tutte le ripetizioni della stessa lettera e successivamente togliendo i simboli "-" rimanenti. Vediamo qualche esempio di percorso:*

-c-aa-ss-a → -c-a-s-a → casa
ccaaa-s-a- → ca-s-a- → casa
eccetera .... Nota 4

La loss function avrà quindi come obiettivo la massimizzazione della probabilità associata al testo $P(W)$ *o, più propriamente, la minimizzazione di* $-\ln P(W)$ *rispetto ai parametri della rete* Nota 5 .

Tra i sistemi end-to-end presentati in letteratura citiamo **Deep Speech 2** [17] della ***Baidu Research***, che nel 2016 raggiungeva su *LibriSpeech "other"* un *WER* del *13,3%*, molto vicino alla prestazione umana stimata del *12%*, effettuando una pesatura delle probabilità generate con un modello linguistico a n-grammi esterno. Nel 2020 il sistema **ContextNet** [18] di ***Google*** ha raggiunto un *WER* di appena *5,5%* sullo stesso benchmark. Nello stesso anno il sistema **wav2vec 2.0** [19] di ***Facebook AI*** dichiara un *WER* del *4,1%*, che può essere ridotto ulteriormente al *3,3%* arricchendo il data set di training con altro materiale audio non trascritto. Ma la cosa più interessante di questo lavoro è il fatto che utilizzando solo un'ora di materiale trascritto il *WER* ottenuto è del *5,8%*, non lontano dallo stato dell'arte. Questo risultato è stato raggiunto tramite una tecnica detta di *pre-training* [20], già utilizzata dal sistema **BERT** [21] di ***Google***.

Il *pre-training* è una tecnica di addestramento non supervisionato dove la rete viene addestrata a raggiungere un obiettivo fittizio in cui è possibile generare algoritmicamente l'uscita attesa della rete e quindi creare un data set di dimensioni grandi a piacere . Ad esempio, nel caso di **BERT**, un sistema per il processamento del testo, l'obiettivo fittizio è la predizione di una o più parole di un testo che vengono mascherate dal sistema.

Nel caso di **wav2vec** la rete si compone di una parte convoluzionale seguita da un transformer (si veda Fig. 4).

Nota 4 - Si noti che invece la sequenza *ccaa-s-ss-a* avrebbe generato la parola *cassa*.

Nota 5 - Per la trattazione matematica relativa si veda [16]

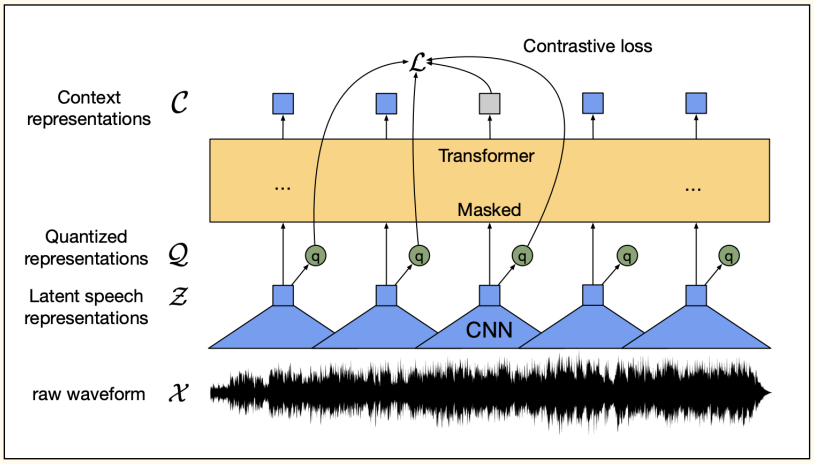


Fig. 4 – Schema di principio **wav2vec 2.0** *(fonte: [19])*

L'ingresso della rete è il segnale audio suddiviso in finestre senza nessuna trasformazione. Il pre-training consiste nel presentare alla rete una grande quantità di materiale audio e di mascherare in modo casuale alcune parti della rappresentazione latente generata della parte convoluzionale. La loss function (di tipo *Contrastive loss*) ha il compito di valutare quanto bene le parti mascherate vengano ricostruite dalla rete. A questo modo si ottiene un modello che, inserito nella rete finale, può essere raffinato per lo scopo in questione della traduzione del parlato in testo con una quantità di testo trascritto decisamente ridotta.

Contemporaneamente, ***Google*** ha presentato un sistema basato sulla nuova architettura **Conformer** [22][23], una variante del transformer in cui viene inserito un modulo convoluzionale tipo **Resnet** [24], che su *LibriSpeech "other"* ottiene un *WER* del *2,6%*, stato dell'arte al momento della scrittura del presente contributo. Si noti che omettendo il modello linguistico il *WER* peggiora di solo lo *0,1%*, per cui si può affermare che la rete, non solo apprende le caratteristiche acustiche del parlato, ma che è in grado di incorporare implicitamente anche le caratteristiche linguistiche senza che sia stato effettuato un addestramento specifico su di un data set testuale. I ricercatori di ***Google*** sono riusciti ad ottenere queste prestazioni, inimmaginabili solo pochi anni prima, utilizzando sia la tecnica di pre-training appena descritta che un'altra tecnica non supervisionata detta **Noisy Student Teacher** (**NST**) [25]. Quest'ultima consiste nei passi seguenti:

- tramite un sistema di *ASR* pre-esistente, detto *Teacher*, si generano delle pseudo-etichette su di un data set non annotato;
- questo data set viene utilizzato per addestrare una nuova rete (*Student*) applicando le tecniche di *dropout* e *data augmentation* per forzare la rete a generalizzare l'apprendimento (da cui il termine *Noisy*);
- infine si effettua un raffinamento del modello con una fase di addestramento su di un data set annotato manualmente.

Il processo può essere ripetuto più volte utilizzando come *Student* reti via via più complesse.

Conclusioni

In questo contributo abbiamo visto come il problema della trascrizione del parlato in testo sia stato affrontato negli anni '90 con successo tramite tecniche che richiedono una modellizzazione dettagliata del processo di analisi del segnale e di sintesi del risultato. Lo sforzo richiesto per l'ottimizzazione di sistemi basati su questo approccio ne ha però limitato l'impiego commerciale per molti anni. Il rapido avanzamento della tecnologia delle reti neurali ha fornito un nuovo impulso alla ricerca nel campo dell'*ASR* migliorando sensibilmente le prestazioni dei modelli e nel contempo riducendo le risorse richieste per la progettazione dei sistemi. Oggi sistemi di riconoscimento del parlato sono disponibili per quasi tutte le lingue del mondo e la precisione della trascrizione è in continuo miglioramento rendendone sempre più efficace l'utilizzo pervasivo in oggetti di uso comune.

Bibliografia

[1] *IBM Shoebox*, IBM Archives (web), https://www.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_7.html (ultimo accesso 30/12/2020)

[2] L. R. Rabiner, *A tutorial on Hidden Markov Models and selected applications in speech recognition*, in "Proceedings of the IEEE ", vol. 77, n. 2, 1989, pp. 257-286, **DOI**: 10.1109/5.18626

[3] *DragonDictate*, in "Wikipedia" (web), https://en.wikipedia.org/wiki/DragonDictate (ultimo accesso 30/12/2020)

[4] D. Monaco, *Le traduzioni in tempo reale al Parlamento europeo sono made in Italy*, in "Wired.it" (web), https://www.wired.it/economia/business/2020/11/11/parlamento-europeo-traduzione-tempo-reale (ultimo accesso 30/12/2020)

[5] A. Messina ed altri, *ANTS: A Complete System for Automatic News Programme Annotation Based on Multimodal Analysis*, in "2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services ", 2008, **DOI**: 10.1109/WIAMIS.2008.15

[6] *Hidden Markov Model*, in "Wikipedia" (web), https://en.wikipedia.org/wiki/Hidden_Markov_model (ultimo accesso 30/12/2020)

[7] J. Hui, *Speech Recognition — GMM, HMM*, in "Medium" (web), https://jonathan-hui.medium.com/speech-recognition-gmm-hmm-8bb5eff8b196 (ultimo accesso 30/12/2020)

[8] S. S. Stevens, J. Volkmann e E. B. Newman, *A scale for the measurement of the psychological magnitude pitch*, in "Journal of the Acoustical Society of America", vol. 8, n. 3, 1937, pp. 185-190, **DOI**: 10.1121/1.1915893

[9] D. G. Childers, D. P. Skinner e R. C. Kemerait, *The Cepstrum: A Guide to Processing*, in "Proceedings of the IEEE", vol. 65, n. 10, 1977, pp. 1428-1443, **DOI**: 10.1109/PROC.1977.10747

[10] A. Viterbi, *Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm*, in "IEEE Transactions on Information Theory", vol. 13, n. 2, 1967, **DOI**: 10.1109/TIT.1967.1054010

[11] Carnegie-Mellon University: Department of Computer Science, *Speech Understanding Systems: A Summary of Results of the Five-Year Research Effort at Carnegie-Mellon University*, 1977, **DOI**: 10.1184/R1/6609821.v1

[12] A. Vaswani ed altri, *Attention Is All You Need*, in "Advances in Neural Information Processing Systems 30 (NIPS 2017)", 2017, pp. 5998-6008, https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[13] S. Hochreiter e J. Schmidhuber, *Long Short-term Memory*, in "Neural Computation", vol. 9, n. 8, 1997, pp. 1735-1780, **DOI**: 10.1162/neco.1997.9.8.1735

[14] C. Lüscher ed altri, *RWTH ASR Systems for LibriSpeech: Hybrid vs Attention*, in "Proceedings of INTERSPEECH 2019", 2019, pp. 231-235, **DOI**: 10.21437/Interspeech.2019-1780

[15] A. Graves, *Sequence Transduction with Recurrent Neural Networks*, in "International Conference of Machine Learning (ICML) 2012 Workshop on Representation Learning", 2012, arXiv:1211.3711

[16] A. Graves ed altri, *Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks*, in "ICML '06: Proceedings of the 23rd international conference on Machine learning", 2006, pp. 369-376, **DOI**: 10.1145/1143844.1143891

[17] D. Amodei ed altri, *Deep Speech 2: End-to-End Speech Recognition in English and Mandarin*, in "ICML'16: Proceedings of the 33rd International Conference on International Conference on Machine Learning", vol. 48, 2016, pp. 173-182, https://dl.acm.org/doi/10.5555/3045390.3045410

[18] Wei Han ed altri, *ContextNet: Improving Convolutional Neural Networks for Automatic Speech Recognition with Global Context*, in "INTERSPEECH 2020", 2020, **DOI**: 10.21437/Interspeech.2020-2059

[19] A. Baevski, *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*, in "34th Conference on Neural Information Processing Systems (NeurIPS 2020)", 2020, https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[20] Qiantong Xu ed altri, *Self-training and Pre-training are Complementary for Speech Recognition*, 2020, arXiv:2010.11430

[21] J. Devlin ed altri, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2019, arXiv:1810.04805

[22] A. Gulati ed altri, *Conformer: Convolution-augmented Transformer for Speech Recognition*, in "INTERSPEECH 2020", 2020, **DOI**: 10.21437/Interspeech.2020-3015

[23] Yu Zhang ed altri, *Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition*, 2020, arXiv:2010.10504

[24] Kaiming He ed altri, *Deep Residual Learning for Image Recognition*, in "2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)", 2016, **DOI**: 10.1109/CVPR.2016.90

[25] D. S. Park ed altri, *Improved Noisy Student Training for Automatic Speech Recognition*, in "INTERSPEECH 2020", 2020, **DOI**: 10.21437/Interspeech.2020-1470

Computer Vision

Contributo a cura di Alberto **Messina**
Rai - Centro Ricerche, Innovazione Tecnologica e Sperimentazione

Questo contributo cerca di offrire al lettore uno spunto alla lettura e comprensione di un dominio applicativo dell'*Intelligenza Artificiale* molto rilevante e allo stesso tempo molto vasto e complesso, quello della *Computer Vision*. Lungi dall'essere una rassegna completa e approfondita dello stato dell'arte, obiettivo che non solo richiederebbe molto più spazio ma risulterebbe ridondante e certamente incompleto e di minor impatto rispetto a molteplici e più autorevoli tentativi già compiuti, il presente contributo è da considerare invece come un inquadramento alla problematica che permetta di valutare, sinteticamente ma ad ampio spettro, la significatività dei vari ambiti della disciplina e presentarne, sotto la medesima luce, le caratteristiche essenziali.

Un modello ispirato alla psicologia cognitiva

Con *Computer Vision* si denotano tradizionalmente le tecnologie e i sistemi che emulano le capacità umane di percezione ed elaborazione cognitiva supportate dal canale sensoriale visivo [1]. Tra i molti approcci di base utili all'inquadramento di questo ambito, il modello di riferimento qui preso ad ispirazione è quello fornito dalla *psicologia cognitiva* [2].

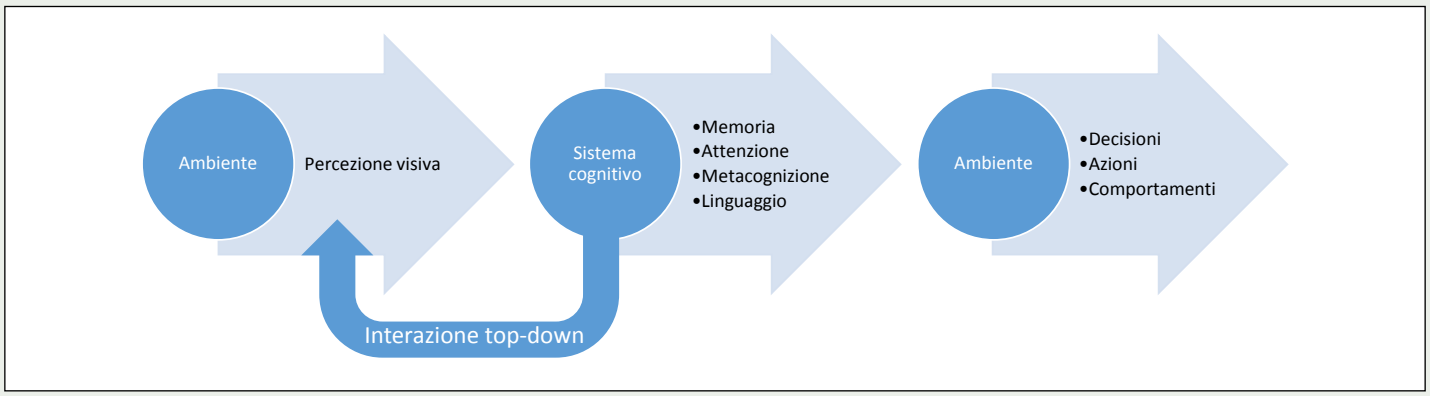
Essa definisce la *percezione* in generale e la visione in particolare come l'insieme di processi attraverso i quali riconosciamo, organizziamo e diamo un significato alle sensazioni che riceviamo dagli stimoli ambientali [3]. In questi processi hanno importanza rilevante non solo i processi *bottom-up* (dalla percezione alla conoscenza, [4]) ma anche i processi *top-down* [5], entrambi influenzati e abilitati dalle abilità cognitive (memoria, attenzione, metacognizione, funzioni esecutive, linguaggio) [6].

Su queste basi e al fine del presente contributo, possiamo riassumere le precedenti considerazioni nel processo semplificato di Fig. 1 [Nota 1], nel quale si introduce anche, come ultimo passo del processo, l'azione di ritorno sull'ambiente in termini di *decisioni*, *azioni* e *comportamenti*.

Il modello di Fig. 1 riportato alla *Computer Vision* è sufficientemente generale da supportare la descrizione di vasti campi applicativi, inclusa la robotica, la videosorveglianza e la guida autonoma [7].

Nota 1 - Nella figura i cerchi rappresentano *entità*, le frecce rappresentano *processi* che hanno come input/attore l'entità alla loro base e forniscono output all'entità puntata.

Fig. 1 – Processo di riferimento per la cognizione visiva (generale)



201_01.2.3_200

Trasferendo e adattando le nozioni di Fig. 1 nel contesto specifico della *comprensione* e *descrizione* di scene visuali campionate attraverso strumenti di cattura multimediali [Nota 2] si perviene al modello di Fig. 2, in cui l'uscita del processo cognitivo si caratterizza in una serie di informazioni fornite seguendo le regole di qualche sistema dichiarativo [Nota 3].

Attraverso il processo di percezione ed evocazione l'osservatore ritrova (o riconosce) nel contenuto fruito una serie di caratteristiche latenti [8], che successivamente vengono organizzate e filtrate dai criteri di verbalizzazione, nonché dalle capacità cognitive e dalla cultura proprie dell'osservatore, e codificati dalle regole del sistema dichiarativo.

Le informazioni prodotte possono essere utilizzate da entità ulteriori (non rappresentate in Fig. 2) al fine di prendere decisioni e compiere azioni nel contesto di un processo di più alto livello. Questo processo di *comprensione e verbalizzazione* della scena è supportato dalle strutture del sistema nervoso centrale dell'osservatore [9].

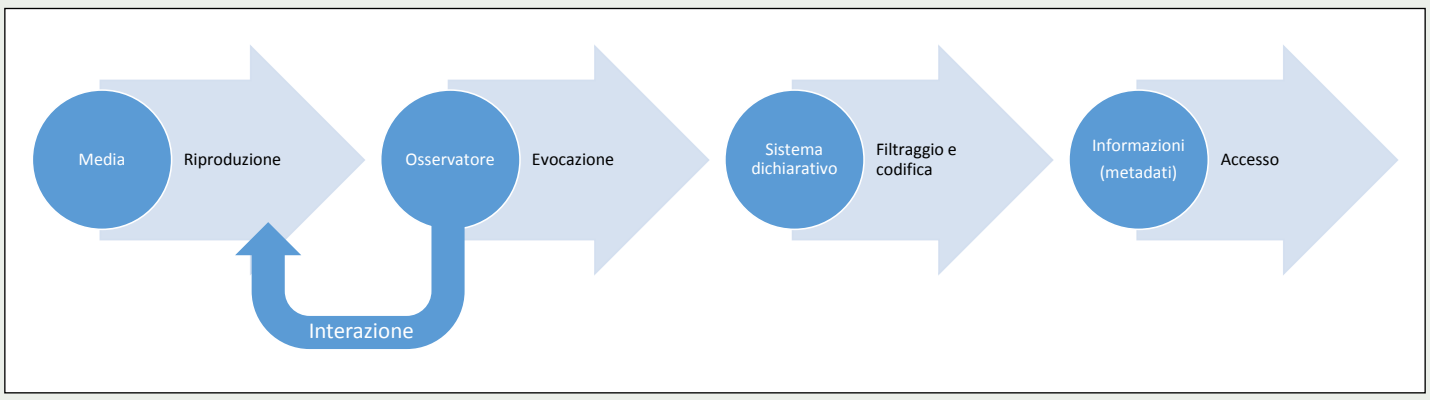
Si noti come il modello di processo rappresentato in Fig. 2 sia indipendente dalla natura dell'osservatore, che può essere un osservatore umano o un qualsivoglia componente di Intelligenza Artificiale [Nota 4] che ne emuli il comportamento e le capacità.

Nota 2 - Si assume quindi che tra la scena naturale e l'osservatore si frapponga un *sistema di cattura* (non evidenziato per motivi di spazio nelle varie figure) che produce un oggetto multimediale (*Media*). Tale sistema di cattura emula le capacità fisiologiche dei sistemi di percezione naturali. Tutti i sistemi di televisione si basano su questo principio di emulazione.

Nota 3 - Un *sistema dichiarativo* è qui inteso come un sistema formale per la rappresentazione dichiarativa della conoscenza (ad esempio una tassonomia o un'ontologia), che ha come oggetto del conoscere l'oggetto visuale. A livello tecnologico, si pensi a *XML*, *RDF* o a un *database relazionale*.

Nota 4 - Un interessante problema è quello di come caratterizzare, per i componenti artificiali, concetti quali *cultura*, *attenzione*, *memoria*, *emozioni*, che giocano un ruolo fondamentale nei processi naturali di visione così come interpretati dalla psicologia cognitiva. Mentre senz'altro i pesi di una *Deep Neural Network* (*DNN*) possono essere assimilati a una nozione di *memoria a lungo termine*, sviluppata attraverso l'elaborazione degli esempi forniti durante la fase di addestramento, e molta letteratura scientifica di settore si è occupata con successo dei modelli attentivi nelle reti neurali, nulla è ancora rintracciabile in merito alla mappatura, nelle architetture delle reti, di caratteristiche cognitive di più alto livello come la *cultura* e le *emozioni*.

Fig. 2 – Processo di riferimento per la comprensione e descrizione di scene visuali



La Fig. 3 esemplifica ulteriormente questo concetto, evidenziando come l'architettura (esemplificata) di base di una *rete multi-blocco convoluzionale* altro non sia che una particolare casistica del modello generale di Fig. 2.

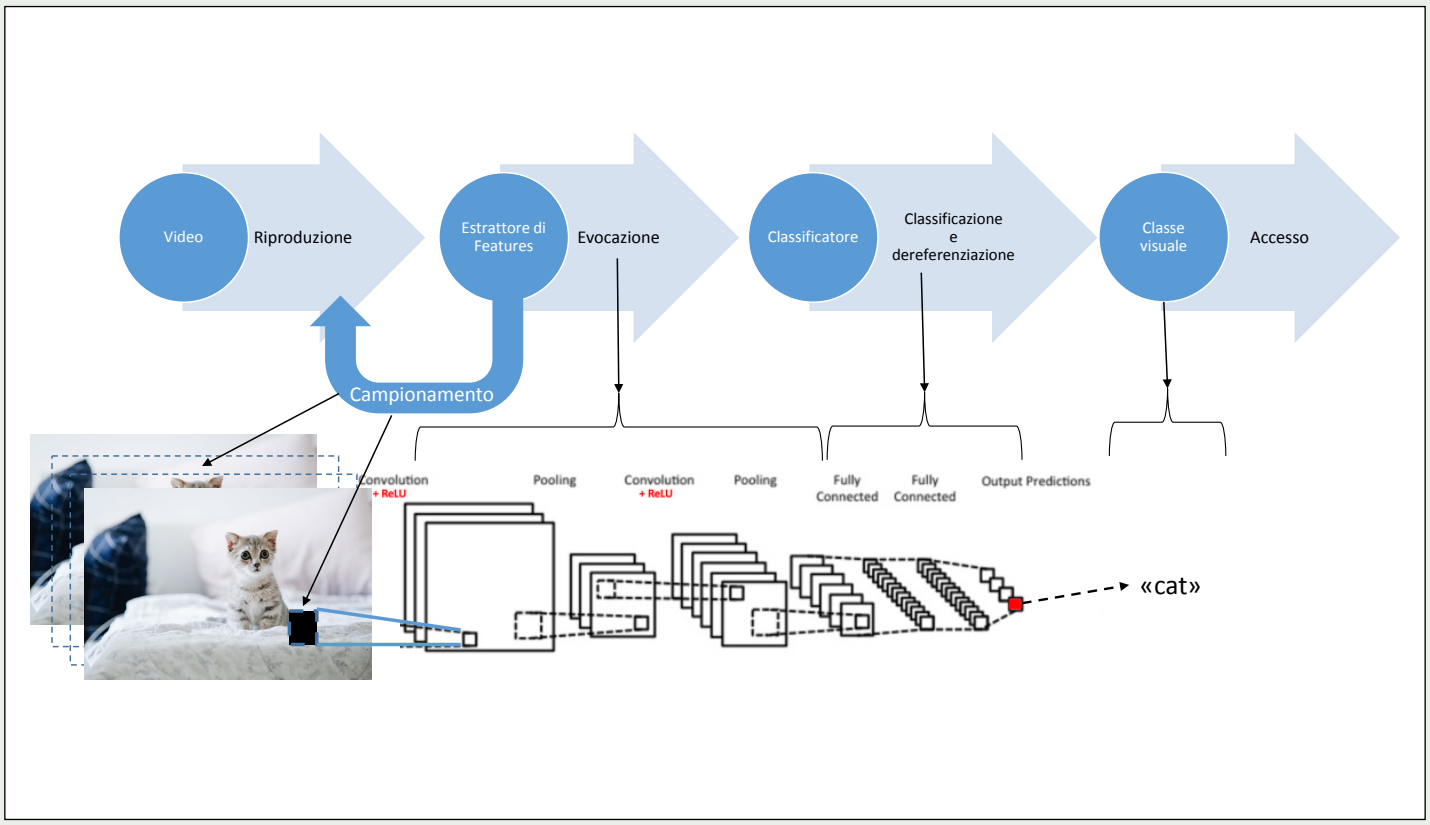
La sezione iniziale della rete funge da campionatore delle informazioni visuali da analizzare nella sezione successiva, dove avviene il processo di evocazione delle caratteristiche latenti.

Infine, le caratteristiche latenti sono combinate e raggruppate da strati di classificazione che riportano in uscita una descrizione conforme ai criteri del sistema dichiarativo prescelto (nel caso in esempio una semplice lista predefinita di etichette).

Generalmente, sebbene le applicazioni di *Computer Vision* siano molteplici e variegate in termini di dominio, complessità e approccio architetturale, esse si possono tutte riferire a questo modello comune.

Nel seguito del presente contributo si illustreranno queste applicazioni, fornendo indicazioni circa la loro applicabilità e le componenti essenziali di cui sono costituite.

Fig. 3 – Reti convoluzionali come caso particolare di osservatori di scene visuali.



Computer Vision: le applicazioni

L'analisi dello stato dell'arte tecnico/scientifico corrente permette di identificare l'insieme di applicazioni riportate in Tabella 1 come le applicazioni principali, vale a dire le applicazioni dove la ricerca di base o applicata ha prodotto negli ultimi anni la maggior parte dei risultati [11].

Esse riflettono, in sostanza, anche buona parte dei compiti a cui un osservatore assolve durante il processo di comprensione di una scena [8] e la loro formulazione permette di generalizzarne le funzionalità ad ampio spettro nei processi di business di molti settori industriali (dall'industria dei media all'industria manifatturiera, a quella automobilistica e a quella alimentare).

Nella prossima sezione verranno analizzati brevemente i componenti principali che abilitano alcune di queste applicazioni.

Tabella 1 – Le principali applicazioni della *Computer Vision*

Applicazione	Definizione	Riferimento al modello generale	Note/esempi
Classificazione di immagini o video	Associazione di una o più etichette di classe ad un immagine o video	L'osservatore produce un a o più etichette di classe	Include il riconoscimento di emozioni/espressioni
Classificazione di immagini o video con localizzazione	Classificazione di immagini o video con l'aggiunta di informazioni di localizzazione spaziotemporale delle classi	L'osservatore produce una o più etichette di classe e segnala una regione spaziotemporale di applicazione di ciascuna classe	Include la classificazione di azioni nel video, il riconoscimento di gesti, di interazioni tra agenti
Rilevamento e identificazione oggetti	Rilevamento Identificazione di specifici oggetti nel video o nell'immagine	L'osservatore identifica una regione dell'immagine come contenente un oggetto e associa ad essa una classe di appartenenza	Include il riconoscimento di testo
Rilevamento e identificazione volti	Rilevamento e Identificazione di volti umani	L'osservatore identifica una regione dell'immagine come contenente un volto umano e associa un nome al volto	Un caso affine è quello della verifica, che consiste nel decidere se o meno un volto corrisponde ad un'identità nota
Segmentazione semantica spazio-temporale	Individuazione e descrizione di una partizione in segmenti spazio-temporali che abbia un valore esplicativo del significato della scena	L'osservatore fornisce una partizione del contenuto visuale percepito e ne fornisce una descrizione dichiarativa	Include la segmentazione temporale basata su azioni, la parsificazione della scena
Descrizione didascalica di immagini o video	Produzione di una frase descrittiva dell'immagine o video	L'osservatore produce una frase in linguaggio naturale che descrive sinteticamente l'immagine o video	

Componenti delle Applicazioni

Classificazione di immagini e video

La *classificazione di immagini* è da sempre uno dei campi più fiorenti della *Computer Vision*, e dove la comunità scientifica si è confrontata in sfide di portata globale [10]. Le tecnologie di ultima generazione disegnate per affrontare questo ambito si basano da anni fortemente su architetture *DNN convoluzionali multi-blocco* per implementare il processo di evocazione di Fig. 2, raggiungendo su dataset standard come **ImageNet** [12] accuratezze anche superiori al livello del giudizio umano [11]. Una DNN multi-blocco è una rete in cui il medesimo blocco si ripete inalterato per un certo numero di volte in sequenza, implementando così un processo di progressiva astrazione di caratteristiche a partire dall'immagine. Esempi di tali reti sono **VGG** [14], **Inception** [13] e **ResNet** [16].

L'architettura stato dell'arte per la classificazione di immagini al momento della scrittura di questo contributo è **EfficientNet** [15], che si basa su alcune idee innovative fondamentali quali:

- l'applicazione della *tecnica dei residui* [Nota 5] e delle *convoluzioni separabili*, inizialmente introdotte con l'architettura **ResNet** [16];
- l'architettura inversa del blocco fondamentale, che la differenzia da **ResNet**;
- l'utilizzo di una particolare funzione di attivazione tra i layer del blocco fondamentale detta *ReLU6*;
- l'utilizzo del cosiddetto *collo di bottiglia lineare* in uscita di ciascun blocco fondamentale.

Gli autori di **EfficientNet** hanno anche proposto uno schema generale per lo scalamento efficiente dell'architettura base della rete (*B0*), individuando sette ulteriori livelli di complessità e basandosi sia su parametri geometrici della rete che su considerazioni di complessità computazionale.

Nota 5 - Tecnica ispirata alla fisiologia delle cellule piramidali della corteccia cerebrale.

La *classificazione di video* aggiunge naturalmente l'ulteriore complessità rappresentata dalla dimensione temporale, che ha reso necessaria l'introduzione di architetture più sofisticate delle reti convoluzionali multi-blocco per modellare esplicitamente questo aspetto, come ad esempio le *Recurrent Neural Networks* e le reti *convoluzionali 3D* [17] [18] . Questi approcci, che fondamentalmente utilizzano sequenze di caratteristiche latenti da ciascun fotogramma estratte con reti multi-blocco stato dell'arte, non hanno tuttavia possibilità di catturare elementi ricorrenti a lunga distanza nel video, che possono essere cruciali per alcuni task di classificazione, come ad esempio la classificazione per genere. Recenti sviluppi utilizzano strutture basate sui grafi per astrarre gradualmente le informazioni partendo dal livello del fotogramma e propagando le informazioni fino a generare una rappresentazione globale del video sulla quale innestare le successive fasi di classificazione [19] o architetture completamente neurali come le *Two-Stream Inflated 3D ConvNet* (**I3D**)[20] nelle quali le caratteristiche latenti di reti convoluzionali standard sono espanse in 3D per apprendere estrattori di funzioni spazio-temporali, o le *Local Global Diffusion* (**LGD**)[21] che estraggono rappresentazioni globali e locali in parallelo modellando le diffusioni tra queste due rappresentazioni.

Rilevamento e localizzazione di oggetti

Un problema di analoga rilevanza rispetto alla classificazione di immagini e video è quello del *rilevamento e localizzazione di oggetti* nella scena. Sebbene la complessità fisiologica del processo di riconoscimento sia tuttora in fase di comprensione [22] , dal punto di vista dell'Intelligenza Artificiale il rilevamento di un oggetto può essere visto come un particolare caso di *classificazione* (attraverso il rilevamento di un'istanza si individua per inferenza una classe). La parte caratterizzante di questo problema è proprio quella dedicata all'indicazione della regione visuale corrispondente all'oggetto rilevato che funge da ulteriore elemento dichiarativo di uscita dall'osservatore. Le architetture stato dell'arte prevedono che il processo di riconoscimento sia abbinato ad un ramo di localizzazione della regione

rettangolare, *bounding box*, che racchiude l'oggetto utilizzando un metodo di regressione [23] o della regione arbitraria, *mask*, attraverso un ulteriore ramo parallelo di analisi [24]. La generalizzazione del problema al video introduce l'ulteriore complessità legata alla dinamica della scena, aggiungendo alla localizzazione statica degli oggetti il problema del loro tracciamento. L'approccio naturale a questo tipo di problema consiste nell'estensione delle architetture disponibili per le immagini con l'aggiunta di un ulteriore componente di tracciamento degli oggetti lungo il video [26] che può basarsi sull'informazione data ad un determinato fotogramma di riferimento [27].

Rilevamento e identificazione di volti

Rilevare la *presenza di un volto* in un'immagine o video, ed eventualmente associare ad esso un nome, è da sempre una delle sfide più ambite della *Computer Vision*. A livello biologico, il rilevamento e il riconoscimento dei volti sono processi distinti che coinvolgono sistemi neurali che non sono probabilmente implicati nel riconoscimento di altri oggetti *non sociali* [28], e studi condotti dagli anni '70 agli anni '90 hanno rivelato che l'elaborazione del viso è collegata a diversi circuiti cerebrali coinvolti nella discriminazione facciale, nel riconoscimento familiare del volto e nel riconoscimento del volto non familiare [29]. In ambito *Computer Vision*, generalmente si separa il compito di rilevare i volti presenti in un'immagine, distinguendoli da altri oggetti nella scena, dal compito di identificare la persona corrispondente al volto rilevato. Gli approcci standard per questi due compiti sono invece in generale ispirati rispettivamente da quelli dedicati al rilevamento di oggetti e alla loro classificazione.

Per quanto riguarda il rilevamento dei volti, *face detection*, al momento della scrittura del presente contributo lo stato dell'arte è rappresentato da **RetinaFace** [30]. Gli autori di questo lavoro propongono una soluzione che unifica la predizione del rettangolo facciale, la localizzazione dei punti chiave facciali (occhi, naso, bocca) e la regressione dallo spazio 3D vincolata ad una topologia di riferimento (*Mesh1k*, un sottoinsieme di [31]). L'unificazione è ottenuta attraverso la definizione di una funzione di perdita globale che tiene in conto linearmente dei tre tipi di perdite. L'architettura fa uso del concetto di *piramide di caratteristiche, feature pyramid*, per estrarre rappresentazioni a diverse scale, che sono poi usate contemporaneamente per le tre regressioni.

Nel campo dell'identificazione e verifica, l'approccio al momento stato dell'arte è **ArcFace** [32] . Gli autori di questo lavoro superano il problema dell'identificazione sotto diverse condizioni di ripresa (età, illuminazione, posa), punto debole quando si usano approcci standard alla classificazione di immagini, con la definizione di una *funzione di perdita* che tiene in conto solo l'angolo formato tra le caratteristiche dell'immagine estratte dalla rete convoluzionale (anche detto *embedding*) e il vettore di pesi dello strato discriminante assieme ad una penalità correlata al margine tra le classi.

Segmentazione semantica spazio-temporale

La *segmentazione semantica spazio-temporale* del video è sicuramente annoverabile tra i compiti di computer vision più complessi e sfidanti. Essa consiste nell'individuare tecniche che consentano di rilevare, classificare e descrivere una partizione in segmenti spazio-temporali (vale a dire, regioni dell'immagine che evolvono nel tempo) che abbia un valore esplicativo del significato della scena. Un sottocaso di questa definizione generale è la *segmentazione puramente temporale*, che considera l'interezza della scena visuale come elemento spaziale e che quindi descrive/classifica l'evoluzione nel tempo a livello globale. In questo ambito spesso le soluzioni proposte in letteratura dipendono fortemente dal genere di contenuti e dall'area applicativa [33] [34] [35]. Mentre, ad esempio, per applicazioni di sorveglianza e monitoraggio il canale visuale può essere considerato dominante per questo genere di analisi, in casi più complessi, come la suddivisione in scene di un film o in unità informative (notizie) di un notiziario, esso rappresenta solo una delle possibili sorgenti informative, sorgente che deve essere complementata con il canale aurale (suono, parlato) e, laddove disponibili, con altri canali informativi.

La quantità di lavori tecnico-scientifici in questo settore è immensa e per ovvie ragioni di spazio ci limiteremo, quindi, a citare e descrivere brevemente alcune tra le ricerche più recenti nel campo della segmentazione temporale di contenuti editoriali.

In questo settore si possono identificare principalmente due problemi fondamentali:

- l'identificazione dei *contenuti editoriali atomici* (programmi) in un flusso continuo di tipo broadcast;
- la suddivisione dei contenuti atomici (anche non provenienti da un flusso broadcast) in *unità semantiche* Nota 6 .

Mentre il primo problema è approcciabile con tecniche generali non dipendenti dal genere dei contenuti, il secondo è, allo stato dell'arte attuale, caratterizzato da approcci verticali. Un lavoro recente che affronta organicamente i due problemi attraverso una modellazione degli stili espressivi delle inquadrature video è presentato in [36], limitando però il problema della segmentazione semantica sostanzialmente al caso news. Gli autori di [37] propongono un'ottimizzazione generica di tecniche di *early fusion* di caratteristiche per il task di segmentazione editoriale, ma proponendo risultati di soli contenuti documentaristico/naturalistici.

Nota 6 - Le unità semantiche dovrebbero essere corrispondenti all'intento editoriale. Ad esempio nel caso news l'identificazione delle singole storie può essere logicamente associata alla scalettatura decisa dalla redazione. Nel caso di contenuti sportivi, la segmentazione è normalmente associabile all'individuazione degli highlights di un evento che farebbe un cronista. Situazioni più sfumate sono invece quelle che riguardano contenuti come talk show, documentari, fiction, dove il punto di vista del fruitore gioca un ruolo non secondario nell'identificazione dei segmenti rilevanti e della loro relazione temporale.

Il recentissimo lavoro proposto in [38] introduce un approccio che integra caratteristiche multimodali su più livelli gerarchici (clip, segmento, intero programma) al fine di generare un supporto alla navigazione top-down nel contenuto. Anche in questo caso, sia i dataset di addestramento che i risultati sperimentali sono verticali su un dominio, quello dei film.

Descrizione didascalica di immagini o video

Questo ambito applicativo si occupa di sviluppare tecnologie e metodi per generare *descrizioni in linguaggio naturale*, semplici frasi di senso compiuto, *di scene visuali*. Può essere considerato come una generalizzazione della classificazione, dove anziché usare etichette statiche e facenti parte di un insieme finito di possibilità definite a priori, si adottano classificazioni strutturate e virtualmente aperte utilizzando il linguaggio naturale come linguaggio dichiarativo (si richiami il processo generale di Fig. 2). Le ultime ricerche in questo campo sfruttano le caratteristiche formali delle architetture basate sul *meccanismo della self-attention* sviluppate nel campo del *Natural Language Processing* ([39] [40]) estendendole organicamente per coprire lo specifico compito visuo-descrittivo [41]. Un recentissimo esempio di tale filone di ricerca è fornito in [42], un lavoro in cui gli autori costruiscono **Oscar**, un metodo e un modello di pre-addestramento per l'apprendimento generalizzato delle cross-correlazioni visuo-linguistiche, che costituisce la base per successive applicazioni specifiche nelle quali il metodo mostra consistenti miglioramenti rispetto allo stato dell'arte. Un altro lavoro analogo per approccio e risultati è **VilBERT** [43].

Come sempre, l'estensione di un task di *Computer Vision* dalle immagini al video introduce una notevole complessità aggiuntiva, e la problematica della *descrizione didascalica* non fa eccezione. In questo campo citiamo il lavoro presentato in [44] in cui gli autori propongono un approccio che modella la dipendenza temporale tra gli eventi in un video in modo esplicito e sfrutta il contesto visivo e linguistico degli eventi precedenti per una narrazione coerente lungo il video stesso.

Conclusioni

La visione umana è uno dei processi più complessi e affascinanti a supporto dell'intelligenza naturale. Con questo presupposto non stupisce che la *Computer Vision* sia tra le branche più complesse e interessanti dell'Intelligenza Artificiale.

Attiva da decenni, ha recentemente visto un'accelerazione esplosiva grazie alle moderne tecnologie delle *reti neurali profonde* (**DNN**) ottenendo in molti campi prestazioni indistinguibili da quelle degli osservatori umani, ma in molti altri essendo ancora distante da risultati davvero sfruttabili in applicazioni pratiche e industriali. La nostra congettura è che laddove i processi sottostanti la percezione naturale, e la conseguente elaborazione cognitiva, non siano ancora adeguatamente compresi e formalizzati, l'analoga controparte artificiale non può raggiungere prestazioni di rilievo.

In questo breve contributo si è tentato di dare un saggio il più possibile coerente delle principali aree applicative e dei componenti tecnologici che sono al giorno d'oggi considerati stato dell'arte in questo campo, fornendone un inquadramento iniziale ispirato al modello percettivo della psicologia cognitiva. La motivazione di questo approccio risiede nella volontà di fornire al lettore non esperto una base comune dove incasellare le tecnologie al fine di individuarne in maniera più immediata il contesto e l'utilità. Nel fare questo lavoro di sintesi estrema, certamente si sono trascurati moltissimi risultati ed approcci per motivi di spazio, ma la speranza è che questo spunto di partenza possa incoraggiare il lettore ad approfondire i dettagli consultando le riviste scientifiche di settore, gli atti delle conferenze specialistiche di punta e le moltissime risorse disponibili in rete.

Bibliografia

[1] D. H. Ballard e C. M. Brown, *Computer Vision*, Prentice Hall, 1982, **ISBN**: 978-0131653160

[2] G.A. Miller, *The cognitive revolution: a historical perspective*, in "Trends in Cognitive Science", vol. 7, n. 3, 2003, pp. 141-144, **DOI**: 10.1016/S1364-6613(03)00029-9

[3] G. B. Vicario, *La percezione visiva*, in G.B. Vicario (ed), "Psicologia sperimentale", 1988, III edizione, CLEUP, Padova, pp. 63-175, **ISBN**: 8871786025

[4] J. Gibson, *The ecological approach to visual perception*, Routledge, 2014, **ISBN**: 9781848725782

[5] R. Gregory, *Eye and Brain: the Psychology of Seeing*, 5^ ed., Oxford University Press, 2015 [1997], **ISBN**: 9780691165165

[6] J.B. Carroll, *Human cognitive abilities: A survey of factor-analytic studies*, Cambridge University Press, 1993, **ISBN**: 0521387124

[7] B. Zhou, P. Krähenbühl e V. Koltun, *Does computer vision matter for action?*, in "Science Robotics", vol. 4, n. 30, maggio 2019, **DOI**: 10.1126/scirobotics.aaw6661

[8] R. Epstein, *The cortical basis of visual scene processing*, in "Visual Cognition", vol. 12, n. 6, 2005, pp. 954–978, **DOI**: 10.1080/13506280444000607

[9] R.A. Epstein e C. I. Baker, *Scene Perception in the Human Brain*, in "Annual review of vision science", vol. 5, 2019, pp. 373-397, **DOI**: 10.1146/annurev-vision-091718-014809

[10] *ImageNet Large Scale Visual Recognition Challenge (ILSVRC)*, ImageNet(web), http://www.image-net.org/challenges/LSVRC/ (ultimo accesso 23/10/2020)

[11] AA.VV., *The AI Index 2019 Annual Report*, AI Index Steering Committee, Human-Centered AI Institute, Stanford University, 2019, https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf (ultimo accesso 23/10/2020)

[12] J. Deng ed altri, *ImageNet: A Large-Scale Hierarchical Image Database*, in "2009 IEEE Conference on Computer Vision and Pattern Recognition", 2009, pp. 248-255, **DOI**: 10.1109/CVPR.2009.5206848

[13] C. Szegedy ed altri, *Going deeper with convolutions*, in "2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)", 2015, pp. 1-9, **DOI**: 10.1109/CVPR.2015.7298594

[14] K. Simonyan e A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, in "3rd International Conference on Learning Representations (ICLR 2015)", 2015, *arXiv:1409.1556*

[15] M. Tan e V. Le Quoc, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,* in "Proceedings of the 36th International Conference on Machine Learning (ICML 2019)", 2019, *http://proceedings.mlr.press/v97/tan19a/tan19a.pdf*

[16] K. He ed altri, *Deep Residual Learning for Image Recognition*, in "2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)", 2016, pp. 770-778, **DOI**: *10.1109/CVPR.2016.90*

[17] Z. Wu ed altri, *Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification*, in "MM '15: Proceedings of the 23rd ACM international conference on Multimedia", 2015, pp. 461-470, **DOI**: *10.1145/2733373.2806222*

[18] F. Yin, ed altri, *Video-based emotion recognition using CNN-RNN and C3D hybrid networks*, in "ICMI '16: Proceedings of the 18th ACM International Conference on Multimodal Interaction", 2016, pp. 445–450, **DOI**: *10.1145/2993148.2997632*

[19] M. Feng ed altri, *Hierarchical Video Frame Sequence Representation with Deep Convolutional Graph Network*, in L. Leal-Taixé e S. Roth (ed), "Computer Vision – ECCV 2018 Workshops", Springer International Publishing, 2018, **ISBN**: 978-3-030-11021-5

[20] J. Carreira e A. Zisserman, *Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset*, in "2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)", 2017, pp. 4724-4733, **DOI**: *10.1109/CVPR.2017.502*

[21] Z. Qiu e altri, *Learning Spatio-Temporal Representation with Local and Global Diffusion*, in "2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)", 2019, pp. 12048-12057, **DOI**: *10.1109/CVPR.2019.01233*

[22] J. J. DiCarlo, D. Zoccolan e N. C. Rust, *How does the brain solve visual object recognition?*, in "Neuron", vol. 73, n. 3, 2012, pp. 415–434, **DOI**: *10.1016/j.neuron.2012.01.010*

[23] J. Redmon ed altri, *You Only Look Once: Unified, Real-Time Object Detection*, in "2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) ", 2016, pp. 779-788, **DOI**: *10.1109/CVPR.2016.91*

[24] K. He ed altri, *Mask R-CNN*, in "2017 IEEE International Conference on Computer Vision (ICCV) ", 2017, pp. 2980-2988, **DOI**: *10.1109/ICCV.2017.322*

[25] X. Lu ed altri, *Grid R-CNN*, in "2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)", 2019, pp. 7355-7364, **DOI**: *10.1109/CVPR.2019.00754*

[26] L. Yang, Y. Fan e N. Xu, *Video Instance Segmentation*, in "2019 IEEE/CVF International Conference on Computer Vision (ICCV)", 2019, pp. 5187-5196, **DOI**: *10.1109/ICCV.2019.00529*

[27] S. Mingjie ed altri, *Fast Template Matching and Update for Video Object Tracking and Segmentation*, in "2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)", 2020, **DOI**: *10.1109/CVPR42600.2020.01080*

[28] D. Y. Tsao e M. S. Livingstone, *Mechanisms of face perception*, in "Annual Review of Neuroscience" vol. 31, 2008, pp. 411–437, **DOI**: *10.1146/annurev.neuro.30.051606.094238*

[29] K. Elgar e R. Campbell, *Annotation: the cognitive neuroscience of face recognition: implications for developmental disorders*, in "The Journal of Child Psychology and Psychiatry", vol. 42, n. 6, 2001, pp 705–717, **DOI**: *10.1111/1469-7610.00767*

[30] J. Deng ed altri, *RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild*, in "2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)", 2020, pp. 5202-5211, **DOI**: *10.1109/CVPR42600.2020.00525*

[31] P. Paysan ed altri, *A 3D face model for pose and illumination invariant face recognition*, in "AVSS '09: Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance", 2009, pp. 296-301, **DOI**: *10.1109/AVSS.2009.58*

[32] J. Deng ed altri, *ArcFace: Additive Angular Margin Loss for Deep Face Recognition*, in "2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)", 2019, pp. 4685-4694, **DOI**: *10.1109/CVPR.2019.00482*

[33] N. Hussein, E. Gavves e A. W. M. Smeulders, *Timeception for Complex Action Recognition*, in "2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)", 2019, pp. 254-263, **DOI**: *10.1109/CVPR.2019.00034*

[34] K. Liu ed altri, *Generalized zero-shot learning for action recognition with web-scale video data*, in "World Wide Web", vol. 22, n. 2, 2019, pp 807–824, **DOI**: *10.1007/s11280-018-0642-6*

[35] A. Cioppa ed altri, *ARTHuS: Adaptive Real-Time Human Segmentation in Sports Through Online Distillation*, in "2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)", 2019, pp. 2505-2514, **DOI**: *10.1109/CVPRW.2019.00306*

[36] R. Kannao e P. Guha, *Segmenting with style: detecting program and story boundaries in TV news broadcast videos*, in "Multimedia Tools and Applications", vol. 78, 2019, pp. 31925–31957, **DOI**: *10.1007/s11042-019-7699-9*

[37] R.M. Kishi, T.H. Trojahn e R. Goularte, *Correlation based feature fusion for the temporal video scene segmentation task*, in "Multimedia Tools and Applications", vol. 78, 2019, pp. 15623–15646, **DOI**: *10.1007/s11042-018-6959-4*

[38] A. Rao ed altri, *A Local-to-Global Approach to Multi-Modal Movie Scene Segmentation*, in "2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)", 2020, pp. 10143-10152, **DOI**: *10.1109/CVPR42600.2020.01016*

[39] J. Devlin ed altri, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, in "Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies", vol. 1, 2019, pp. 4171-4186, **DOI**: *10.18653/v1/N19-1423*

[40] Z. Yang ed altri, *XLNet: Generalized Autoregressive Pretraining for Language Understanding*, in "Advances in Neural Information Processing Systems 32 (NeurIPS 2019)", 2019, *https://papers.nips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf*

[41] A. Burns ed altri, *Language Features Matter: Effective Language Representations for Vision-Language Tasks*, in "2019 IEEE/CVF International Conference on Computer Vision (ICCV)", 2019, pp. 7473-7482, **DOI**: *10.1109/ICCV.2019.00757*

[42] X. Li ed altri, *Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks*, in A. Vedaldi et al. (ed), "Computer Vision – ECCV 2020 ", Springer International Publishing, 2020, **ISBN**: 978-3-030-58576-1

[43] J. Lu ed altri, *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*, in "Advances in Neural Information Processing Systems 32 (NeurIPS 2019)", 2019, *https://papers.nips.cc/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf*

[44] J. Mun ed altri, *Streamlined Dense Video Captioning*, in "2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) ", 2019, pp. 6581-6590, **DOI**: *10.1109/CVPR.2019.00675*

NLP: Natural Language Processing

Contributo a cura di Maurizio **Montagnuolo**
Rai - Centro Ricerche, Innovazione Tecnologica e Sperimentazione

L'elaborazione del linguaggio naturale è la capacità umana di interpretare una sequenza ordinata di parole. Come vedremo, l'emulazione di tale capacità da parte di un elaboratore elettronico risulta particolarmente complessa e difficile da affrontare, a causa delle caratteristiche intrinseche del linguaggio umano, quali ad esempio l'ambiguità, la polisemia e la dipendenza contestuale.

In tale ambito, caratterizzato da un'estrema varietà e quantità di contenuti, l'intelligenza artificiale sta assumendo un ruolo altamente strategico, favorendo lo sviluppo e l'esercizio di soluzioni altamente innovative atte all'elaborazione, comprensione e generazione di testi, dialoghi e conversazioni. In particolare, la crescente capacità di calcolo a disposizione degli sviluppatori, unitamente ai progressi degli algoritmi di *deep learning*, permette oggi di ottenere prestazioni sorprendenti in applicazioni quali la traduzione automatica, l'interazione verbale uomo-macchina, e l'individuazione di informazioni chiave dai documenti testuali.

Questa sezione presenta lo stato attuale della ricerca nel campo dell'analisi del linguaggio naturale, tenuto conto dei più recenti progressi apportati dal deep learning, e gli obiettivi che si stanno perseguendo, con un particolare riferimento alle applicazioni in ambito media e radiotelevisivo.

Cosa si intende per elaborazione del linguaggio naturale

L'elaborazione del linguaggio naturale, detta anche **NLP** dall'inglese *Natural Language Processing*, è l'insieme di algoritmi e procedure che permettono il trattamento e la comprensione del linguaggio mediante tecniche informatizzate. Tale comprensione è determinata dal capire, per poi essere in grado di utilizzare, il linguaggio a vari livelli di astrazione, partendo dai singoli *caratteri*, passando per le *parole* da essi formate, per concludersi con la strutturazione sia di singole *frasi*, sia di *paragrafi* costruiti dalla successione di più frasi.

Questa strutturazione del testo può essere rappresentata in forma piramidale, come illustrato in Fig. 1. Partendo dalla base, e muovendosi verso il vertice della piramide, la semantica dei dati, ossia il significato assunto dal dato stesso, diviene via via più complessa ed articolata.

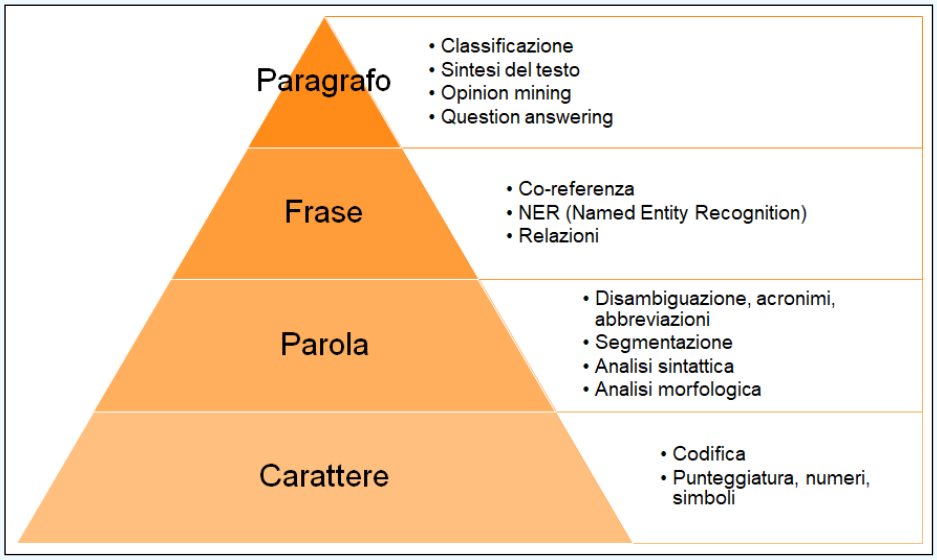


Fig. 1 – Strutturazione e semantica del testo

201_01.2.4_3.00

Alla base della piramide si trovano i singoli caratteri che compongono il testo da analizzare. A questo livello si pongono i problemi atti ad individuare correttamente la codifica utilizzata per la rappresentazione dei caratteri (ad esempio *UTF-8*, *ASCII*, *Latin*, e così via) e, successivamente, distinguere un tipo di carattere da un altro, in modo da essere in grado di riconoscere per ciascun carattere la tipologia di appartenenza, quale ad esempio *punteggiatura*, *numero*, *simbolo*, *lettera alfabetica*. Sebbene all'apparenza possano sembrare banali, queste operazioni sono fondamentali per la corretta esecuzione delle fasi successive.

Una sequenza di caratteri forma una parola. Una parola, che può anche essere una sigla, un acronimo o un'abbreviazione, può assumere significati diversi in contesti diversi (si parla in questo caso di disambiguazione). I processi di elaborazione applicabili ad una parola includono la segmentazione o analisi lessicale, cioè l'individuazione delle singole parole all'interno di una frase, l'analisi sintattica, ovverosia l'arrangiamento delle parole in una struttura sintattica ad albero, e l'analisi morfologica, ossia l'associazione della parola alla corrispondente categoria grammaticale.

Una sequenza di parole forma una frase. All'interno della frase si pongono i problemi di identificare le entità rappresentative (**NER**, dall'inglese *Named Entity Recognition*), quali ad esempio persone, luoghi, organizzazioni, date o valute, la coreferenza delle entità, ovverosia la capacità di individuare le espressioni nel testo che si riferiscono alla stessa entità, e le relazioni causali e/o temporali che intercorrono tra entità diverse.

In ultimo, insiemi di frasi formano un paragrafo. Le elaborazioni applicabili a questo livello includono la classificazione del testo in un insieme di categorie (ad esempio sport, spettacolo, politica, ecc.), la creazione automatica di sommari, l'opinion mining ed il question answering.

L'identificazione delle suddette informazioni, ad ogni livello della piramide, ha un importante risvolto applicativo in diversi ambiti, tra i quali citiamo a titolo esemplificativo, il giornalismo investigativo e la documentazione degli archivi per finalità di ricerca. La Fig. 2 mostra un esempio dell'applicazione del processamento automatico di analisi del linguaggio applicata alla descrizione di una puntata della trasmissione *Techetechetè* pubblicata sul portale **RaiPlay**. Il testo è stato inizialmente suddiviso in frasi (in inglese *Sentence Detection*). Ciascuna frase è stata poi suddivisa in segmenti (in gergo tecnico denominati *token*, dall'inglese *Sentence Tokenisation*) separando le parole dalla punteggiatura, numeri ed altri simboli. Infine, da ciascun segmento sono state individuate le entità rappresentanti organizzazioni (in giallo), persone (in azzurro) e luoghi (in arancione).

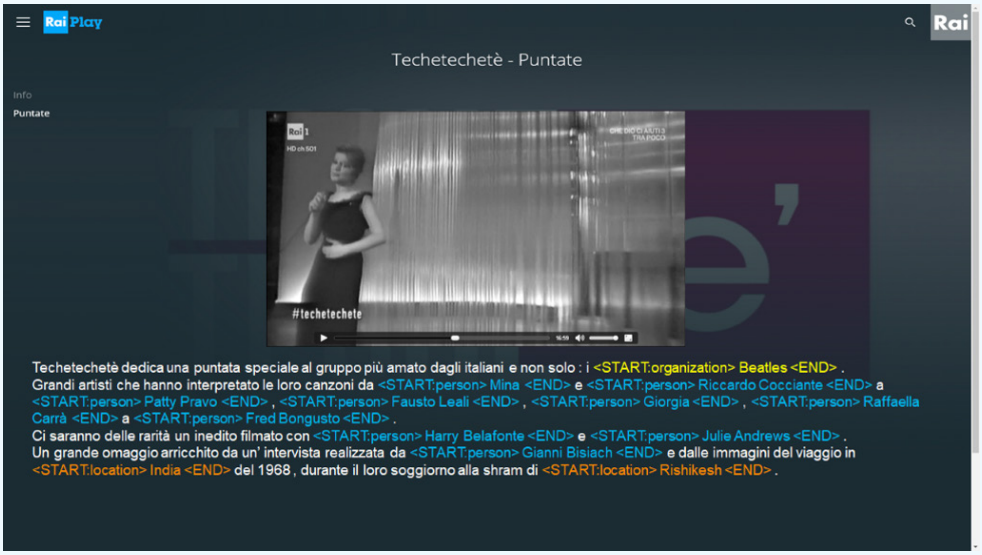


Fig. 2 – Esempio di applicazione dell'estrazione delle entità da una descrizione del portale RaiPlay

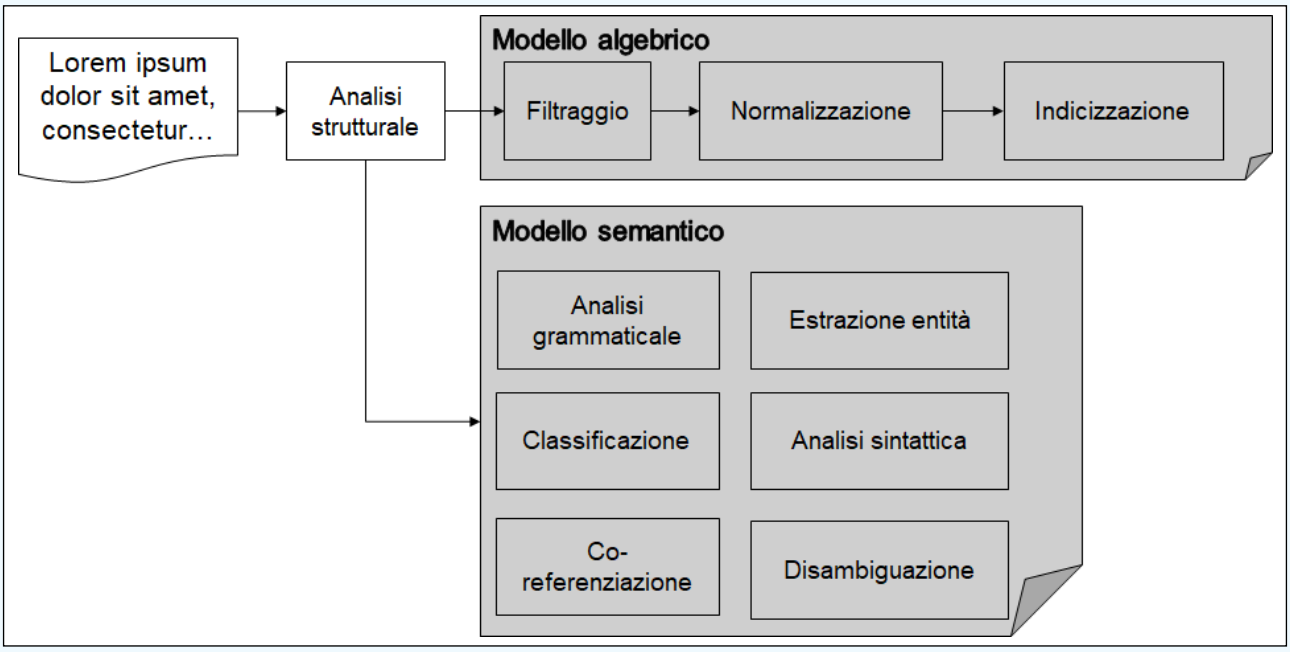


Fig. 3 – Architettura generale per la rappresentazione del testo.

La catena di rappresentazione del testo

La Fig. 3 illustra l'architettura generale della catena di elaborazione di un sistema di analisi del linguaggio.

L'ingresso è costituito da un documento di testo, che può essere originario, come ad esempio una didascalia associata ad un'immagine di archivio, oppure derivato da altre forme di dati, come ad esempio una trascrizione generata in modo automatico con tecniche di riconoscimento del parlato (**ASR,** dall'inglese *Automatic Speech Recognition*).

Attraverso algoritmi di analisi strutturale, si identificano le singole frasi. Ciascuna frase viene successivamente suddivisa nelle singole unità elementari (*token*) sulle quali è possibile adottare due strategie di analisi, basate su due modelli quali quello dell'*algebra* e quello della *semantica* del testo analizzato.

Gli algoritmi del modello algebrico sono alla base del funzionamento dei motori di ricerca full text quali ***Apache Solr*** [1] ed ***Elasticsearch*** [2].

Nel *modello algebrico* i token vengono dapprima filtrati per rimuovere quelli non rilevanti, perché poco significativi o troppo ricorrenti nella lingua analizzata, e quindi inadatti per l'identificazione di concetti e tematiche contenute nel testo. Nel gergo informatico, i token ignorati sono riferiti col termine di *stop words*, ed includono, ad esempio, articoli, congiunzioni, preposizioni e verbi modali. Nella fase successiva viene effettuata la normalizzazione del testo, consistente nell'uniformare spazi, apostrofi ed accenti, nel convertire le parole plurali in singolari o le lettere maiuscole in minuscole, o nel ridurre una parola alla sua radice (in inglese *stemming*). Attraverso i processi di filtraggio e normalizzazione, il testo viene trasformato in un elenco di parole chiave. Ad ogni parola viene associato un valore numerico, normalmente correlato alla frequenza della parola nel testo di riferimento. Si ottiene così una rappresentazione vettoriale (da cui il nome modello algebrico) del documento testuale analizzato. I vettori sono infine memorizzati (fase di indicizzazione) in un'apposita struttura dati, che potrà essere successivamente interrogata per ricercare un particolare documento presente al suo interno.

Gli algoritmi del *modello semantico* sono alla base dei moderni sistemi di assistenza vocale, tra i quali ***Siri*** di **Apple**, ***Alexa*** di **Amazon**, ***Cortana*** di **Microsoft** e **Assistant** di **Google**. A differenza del modello algebrico, nel modello semantico vengono identificate, oltre alla frequenza, altre caratteristiche distintive, quali co-occorrenza, correlazione e posizionamento. La co-occorrenza fa riferimento alla presenza simultanea di due o più parole all'interno del testo. Questa informazione può essere utilizzata per distinguere l'argomento principale a cui il documento si riferisce. Ad esempio, dalla presenza delle parole *classifica*, *pareggio* e *rigore* all'interno di un articolo di agenzia, si potrebbe dedurre che nell'articolo si parla di un evento sportivo. La correlazione fa riferimento alla relazione reciproca o alla corrispondenza tra due o più termini. Ad esempio, le categorie sintattiche che compongono una frase (soggetto, predicato, complemento oggetto e via dicendo) forniscono informazioni circa gli agenti, le modalità e le conseguenze di un'azione descritta nella frase [Nota 1]. Infine, il posizionamento permette di stabilire una gerarchia tra le parole in funzione degli intenti dell'autore del testo analizzato.

Naturalmente, anche nel modello semantico l'elenco di parole che compongono il testo viene trasformato nella corrispondente rappresentazione numerica, in modo che possa essere opportunamente elaborato mediante tecniche automatiche[Nota 2].

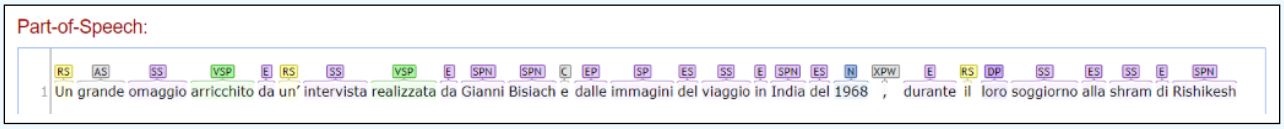
Il modello semantico include diversi tipi di analisi applicabili singolarmente o congiuntamente, a seconda delle necessità dell'utente, brevemente introdotti nel seguito. Per un approfondimento, si rimanda a [3].

Analisi grammaticale

L'analisi grammaticale, anche nota come **POS** (dall'inglese *Part of Speech*) **tagging** consiste nel classificare ciascuna parola secondo la corrispondente categoria grammaticale, quale ad esempio sostantivo, verbo, aggettivo, articolo eccetera. Questa classificazione fornisce un'informazione fondamentale per determinare il ruolo della parola (e di quelle ad essa vicine) all'interno della frase. Ad esempio, sapendo che una parola è un articolo, permette di predire con buona probabilità che la parola successiva sarà un sostantivo. Un algoritmo di POS tagging deve essere in grado di disambiguare, ovverosia assegnare la categoria più appropriata in base al contesto in cui si trova la parola. Ad esempio, nella frase *"Il signor Bianchi vive a Torino"*, occorre riconoscere che la parola *Bianchi* è un nome proprio e non un aggettivo.

Un esempio è mostrato in Fig. 4. Si noti la classificazione degli aggettivi singolari (AS), dei nomi comuni singolari (SS), e dei nomi propri (SPN)[Nota 3].

Fig. 4 – Esempio di POS tagging *(fonte: http://hlt-services2.fbk.eu/textpro-demo/textpro.php ultimo accesso 22/09/2020)*



Nota 1 - Nell'ambito del giornalismo investigativo ci si riferisce a queste informazioni con l'acronimo **5W**, dall'inglese *Who* (chi), *What* (cosa), *Where* (dove), *When* (quando), *Why* (perché).

Nota 2 - A tal fine si possono utilizzare diverse tecniche tra cui le più famose sono gli algoritmi ***word2vec*** (https://code.google.com/archive/p/word2vec/) e ***GloVe*** (https://nlp.stanford.edu/projects/glove/) *(ultimo accesso per entrambi 22/09/2020)*

Nota 3 - L'elenco completo delle etichette grammaticali è consultabile all'indirizzo http://textpro.fbk.eu/modules/tagpro/ita-tagset *(ultimo accesso 22/09/2020)*

Classificazione

La classificazione consiste nell'assegnare una o più categorie facenti parte di un insieme di classi (denominato *tassonomia*) da cui scegliere. La classificazione del testo è utilizzata in numerosi ambiti, quali l'indicizzazione bibliografica attraverso un vocabolario predeterminato, la metadatazione ed archiviazione automatica, od il filtraggio dei documenti a seguito di una ricerca (*faceted search*). Ad esempio, un articolo di agenzia, o la trascrizione di una notizia di telegiornale, possono essere classificati nella corrispondente categoria giornalistica e, successivamente, ricercate e/o filtrate in base a tale categoria.

Co-referenziazione

La co-referenziazione ha l'obiettivo di riconoscere tutte le espressioni (quali ad esempio nomi, pronomi, aggettivi ed acronimi) che si riferiscono alla stessa entità in un testo. Un esempio è mostrato in Fig. 5.

Estrazione delle entità

L'estrazione delle entità consiste nell'individuazione di parole o gruppi di parole che possono corrispondere ad entità semantiche. I gruppi individuati sono normalmente classificati in categorie quali *persone*, *luoghi*, *organizzazioni*, *eventi* od *espressioni temporali*. La principale difficoltà consiste nel fatto che spesso una parola, od un gruppo di parole, può assumere categorie diverse a seconda del contesto. Ad esempio, la parola *Roma* potrebbe essere classificata come persona, luogo od organizzazione, a seconda che si riferisca, rispettivamente, ad un cognome, alla città o alla squadra di calcio.

Analisi sintattica

L'analisi sintattica mira a costruire un grafo (in inglese *dependency parse tree*) rappresentante la struttura di una frase in accordo con determinate forme grammaticali. Nell'esempio di Fig. 6, la parola composta *Julie Andrews* rappresenta il soggetto (SUBJ) della parola *saranno*.

Fig. 5 – Esempio di co-referenziazione *(fonte: https://huggingface.co/coref/ ultimo accesso 22/09/2020)*

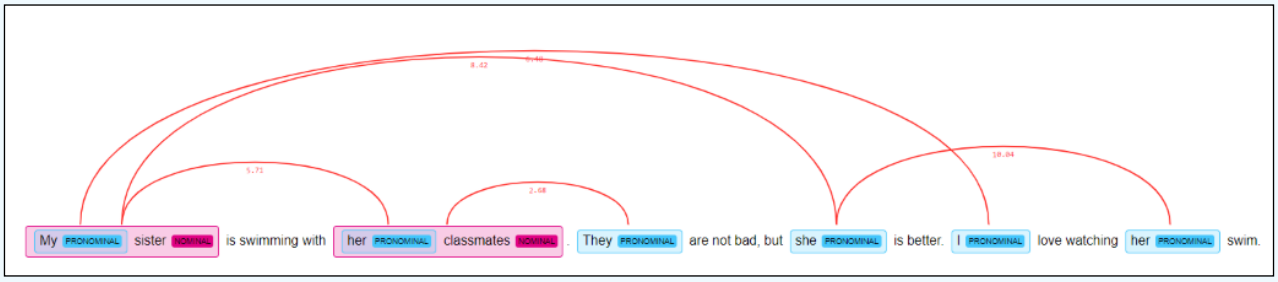
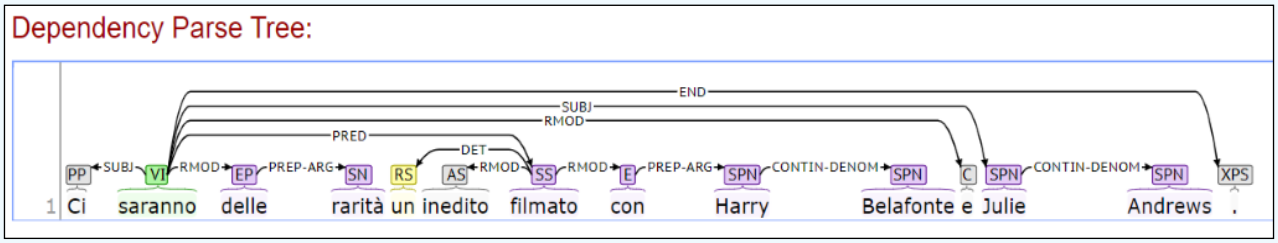


Fig. 6 – Esempio di analisi sintattica *(fonte: http://hlt-services2.fbk.eu/textpro-demo/textpro.php ultimo accesso 22/09/2020)*



Disambiguazione

La disambiguazione (**WSD**, dall'inglese *Word Sense Disambiguation*) è il processo con il quale si precisa il significato di una parola, qualora quest'ultima possa assumere significati diversi a seconda dei contesti. La disambiguazione è particolarmente utile nelle applicazioni di traduzione automatica, metadatazione e ricerca. A titolo di esempio, consideriamo le seguenti frasi:

C. Il *calcio* è uno sport di squadra;
D. Ho dato un *calcio* al pallone;
E. Il simbolo chimico del *calcio* è Ca

Sebbene sia semplice per un essere umano riconoscere che la parola calcio si riferisce (A) ad uno sport, (B) ad un'azione compiuta e (C) ad un elemento della chimica, lo sviluppo di algoritmi in grado di replicare questa capacità umana è particolarmente difficile.

Deep learning per l'elaborazione del linguaggio naturale

I paragrafi precedenti hanno fornito una panoramica sui compiti svolti da un sistema informatico per l'analisi del linguaggio. In questa sezione ci soffermeremo su come questi possano essere efficacemente realizzati, grazie alle opportunità offerte dal deep learning.

I primi studi sull'applicazione di tecniche basate sul deep learning volti alla risoluzione di problemi di NLP risalgono al 2011. Una trattazione dettagliata dell'argomento è disponibile in [4].

Sviluppata nel linguaggio di programmazione ANSI C, **SENNA** (*Semantic/syntactic Extraction using a Neural Network Architecture*) è una rete neurale applicabile a varie attività, tra cui il riconoscimento delle entità ed il POS tagging, che ottenne un sensibile miglioramento delle prestazioni rispetto ad altri approcci rappresentativi dello stato dell'arte dell'epoca [5]. La versatilità di SENNA fu ottenuta sfruttando la capacità di apprendimento e generalizzazione delle reti neurali, basata sull'analisi autonoma dei dati in ingresso, piuttosto che su una complicata ingegnerizzazione delle caratteristiche (*features*) degli stessi, prerogativa dei metodi di machine learning tradizionali.

Successivamente, gli anni dal 2013 al 2017 hanno visto la diffusione di nuove architetture di rete, tra le quali le *reti neurali ricorrenti* (**RNN** - *Recurrent Neural Netwok*), le *reti neurali convoluzionali* (**CNN** - *Convolutional Neural Network*) e le *reti neurali ricorsive* (*Recursive Neural Network*).

Le *reti neurali ricorrenti* sono dotate di connessioni di feedback verso neuroni dello stesso livello e/o verso neuroni dei livelli precedenti, rendendole particolarmente adatte per la gestione di dati temporali, quali sequenze audio, video o testi perché dotate di un effetto memoria che permette di correlare l'informazione ad un determinato istante temporale, con le informazioni riferite agli istanti temporali precedenti. Una rete RNN considera ogni parola di una frase come una variabile di ingresso osservata all'istante temporale *t*. Questa informazione è combinata con quella ottenuta all'istante temporale *t-1* (dunque corrispondente alla parola precedente).

Lo schema a blocchi delle tipiche architetture di una rete RNN per applicazioni NLP è illustrato in Fig. 7 di pagina seguente. Nella prima configurazione (*many to one*) una sequenza di parole in ingresso viene associata ad un unico valore di uscita; applicazioni tipiche sono la classificazione e la sentiment analysis. Nella seconda configurazione (*many to many*) anche l'uscita è costituita da una sequenza di valori che possono essere asincroni (immagine centrale) o sincroni (immagine di destra) rispetto agli ingressi; esempi applicativi includono, rispettivamente, la traduzione automatica (in cui non è necessariamente richiesta una corrispondenza 1:1 tra ingressi ed uscite), ed il POS tagging (in cui, viceversa ciascuna uscita deve corrispondere esattamente a ciascun ingresso).

In origine, le reti RNN erano considerate difficili da addestrare e, di conseguenza, venivano utilizzate raramente. Gli studi in [6] hanno contribuito significativamente al superamento di tali difficoltà,

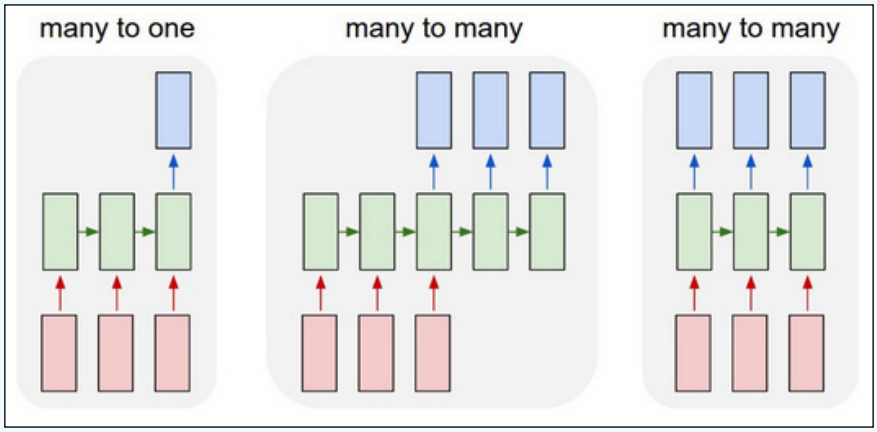


Fig. 7 – Illustrazione delle architetture di rete RNN per l'analisi del linguaggio. I rettangoli rossi, verdi e blu corrispondono, rispettivamente, agli ingressi (parole), unità di elaborazione (hidden layer) ed uscite della rete *(fonte: http://karpathy.github.io/2015/05/21/rnn-effectiveness/ ultimo accesso 22/09/2020)*

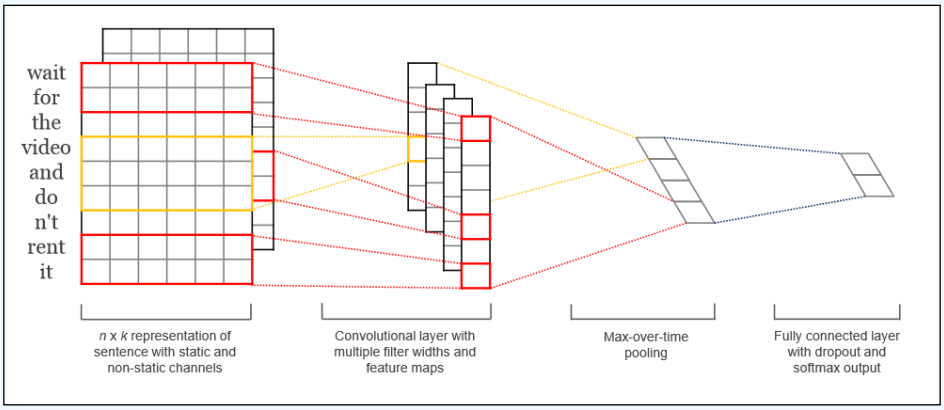
consentendone così il pieno sfruttamento delle potenzialità. Ad oggi le reti ricorrenti sono utilizzate in una moltitudine di applicazioni NLP, tra le quali citiamo le seguenti:

- *classificazione di parole o sequenze di parole* (es. estrazione delle entità);
- *modellazione del linguaggio*, es. POS tagging, riconoscimento del parlato (STT – Speech to Text, traduzione automatica);
- *classificazione di frasi* (es. sentiment analysis)
- *corrispondenza semantica* (es. question answering)

Nelle *reti neurali convoluzionali* lo stato della rete dipende unicamente dallo stato corrente (non si ha cioè propagazione all'indietro delle uscite da un livello di neuroni ai precedenti), richiedendo una minore complessità architetturale rispetto alle reti ricorrenti. Questa caratteristica rende le reti CNN particolarmente adatte ai compiti di visione artificiale, quali la classificazione di oggetti e il riconoscimento dei volti. D'altra parte, la mancanza di informazione contestuale (cioè delle relazioni tra parole), rende questo tipo di reti meno attraente per applicazioni NLP. Tuttavia, in letteratura è possibile trovare alcuni studi volti ad estendere il loro utilizzo anche in ambito linguistico [7][8]. In questo caso il testo da analizzare è rappresentato da una matrice in cui le righe identificano le parole del testo, e le colonne la rappresentazione vettoriale della parola corrispondente. Un esempio è mostrato in Fig. 8.

Per ovviare alla mancanza di contesto, sono inoltre state proposte architetture ibride, composte da un'alternanza di livelli convoluzionali e livelli ricorrenti [9][10]. Gli ambiti applicativi di maggior successo delle reti CNN includono l'estrazione delle entità ed il POS tagging.

Fig. 8 – Esempio di rete neurale convoluzionale per il trattamento del testo *(fonte: [8])*



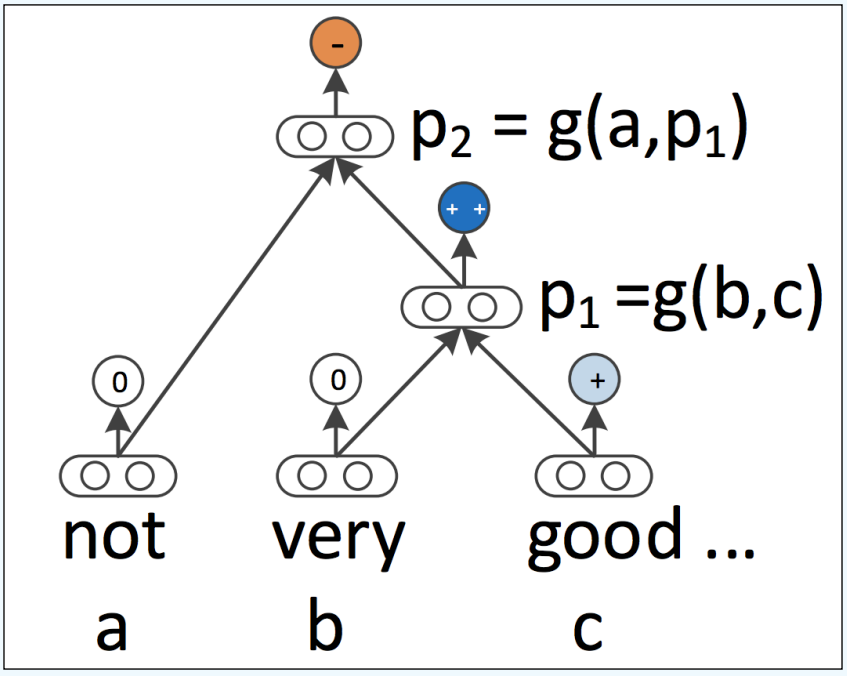
Sia le reti RNN che le CNN basano i propri fondamenti modellando il testo come una sequenza di parole. Tuttavia, da un punto di vista linguistico, il linguaggio naturale presenta anche caratteristiche gerarchiche, oltreché temporali. Infatti, come introdotto precedentemente, le parole sono utilizzate per comporre delle frasi, che a loro volta sono combinate ricorsivamente per formare il testo.

L'idea di trattare un testo con una rappresentazione ad albero, anziché come una lista piatta di parole, ha dato origine alle *reti neurali ricorsive*. Le reti neurali ricorsive costruiscono la rappresentazione del testo in forma gerarchica dal basso verso l'alto. In corrispondenza di ogni nodo dell'albero viene calcolata una nuova rappresentazione componendo la rappresentazione di ciascuno dei nodi figli. Un esempio è mostrato in Fig. 9.

Gli ambiti d'applicazione delle reti neurali ricorsive includono l'analisi sintattica, la co-referenziazione e la sentiment analysis.

La fase di addestramento dei metodi descritti in precedenza necessita di una grande quantità di dati annotati. Si parla, in questo caso, di apprendimento supervisionato. La creazione dei dataset di apprendimento è un'operazione lunga, dispendiosa e non esente da errori; per ovviare a questi problemi sono stati proposti metodi non supervisionati (quindi non necessitanti di dati annotati) basati su modelli linguistici. Le informazioni apprese da un modello linguistico possono essere utilizzate sia per affinare il modello stesso [12], sia per costruire nuovi modelli a partire da quest'ultimo [13]. Questa tecnica, nota con il termine di *transfer learning*, consiste nel trasferire la conoscenza acquisita da un contesto applicativo ad un altro avente caratteristiche simili a quello originario. L'applicazione del transfer learning porta ad un significativo miglioramento delle prestazioni in numerosi ambiti applicativi, come evidenziato in [13]. Tra questi, **BERT** [14], acronimo di *Bidirectional Encoder Representations from Transformers*, è universalmente riconosciuto come uno dei più popolari sistemi per la risoluzione di problemi NLP.

Fig. 9 – Esempio di rete neurale ricorsiva *(fonte: [11])*



Conclusioni

Le recenti evoluzioni nel campo dell'intelligenza artificiale hanno permesso lo sviluppo di nuove architetture di reti neurali per l'apprendimento, valutazione ed interpretazione dei dati multimediali (testi, suoni, immagini, video). La comunità scientifica è oggigiorno consapevole delle possibilità di impiegare tali reti per la risoluzione di compiti complessi nell'ambito dell'analisi del linguaggio, quali ad esempio il question answering o la traduzione automatica. Tali applicazioni stanno permettendo lo sviluppo di nuove tecnologie vocali per l'interazione uomo-macchina basata sul linguaggio e la conversazione, che potranno essere efficacemente impiegate in molteplici ambiti, tra cui media, telecomunicazioni e servizi al cittadino.

Bibliografia

[1] *Solr Home page*, Apache Solr (web), https://lucene.apache.org/solr/ (ultimo accesso 22/09/2020)

[2] *Elasticsearch Home page*, Elastic (web), https://www.elastic.co/elasticsearch/ (ultimo accesso 22/09/2020)

[3] J. Eisenstein, *Introduction to Natural Language Processing*, MIT Press, 2019, ISBN: 9780262042840, https://mitpress.mit.edu/books/introduction-natural-language-processing

[4] Y. Goldberg, *Neural network methods for natural language processing*, in "Synthesis Lectures on Human Language Technologies", vol.10, n° 1, Aprile 2017, pp. 1–309, **DOI**: 10.2200/S00762ED1V01Y201703HLT037

[5] R. Collobert ed altri, *Natural Language Processing (Almost) from Scratch*, in "The Journal of Machine Learning Research", vol. 12, Novembre 2011, pp. 2493–2537, https://dl.acm.org/doi/10.5555/1953048.2078186

[6] I. Sutskever, *Training recurrent neural networks*, Ph.D. Dissertation, Università di Toronto, Canada, Advisor(s) Geoffrey Hinton, 2013, https://www.cs.utoronto.ca/~ilya/pubs/ilya_sutskever_phd_thesis.pdf

[7] N. Kalchbrenner, E. Grefenstette e P. Blunsom, *A Convolutional Neural Network for Modelling Sentences*, in "Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics", Giugno 2014, Baltimora-USA, vol. 1, pp. 655–665, **DOI**: 10.3115/v1/P14-1062

[8] Y. Kim, *Convolutional Neural Networks for Sentence Classification*, in "Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)", Doha-Qatar, Ottobre 2014, pp. 1746-1751, **DOI**: 10.3115/v1/D14-1181

[9] J. Wang ed altri, *Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model*, in "Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)", Agosto 2016, Berlino, vol. 2, pp. 225–230, **DOI**: 10.18653/v1/P16-2037

[10] J. Bradbury ed altri, *Quasi-Recurrent Neural Networks*, "5th International Conference on Learning Representations (ICLR 2017)", Tolone, Aprile 2017, arXiv:1611.01576

[11] R. Socher, A. Perelygin, e J. Wu, *Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank*, in "Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing", Ottobre 2013, Seattle, pp. 1631–1642, https://www.aclweb.org/anthology/D13-1170

[12] P. Ramachandran ed altri, *Unsupervised Pretraining for Sequence to Sequence Learning*, In "Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing", Settembre 2017, Copenhagen, pp. 383-391, **DOI**: 10.18653/v1/D17-1039

[13] M. Peters ed altri, *Deep Contextualized Word Representations*, in "Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies", Giugno 2018, New Orleans, Vol. 1, pp. 2227-2237, **DOI**: 10.18653/v1/N18-1202

[14] J. Devlin ed altri, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, in "Proceedings of NAACL-HLT 2019", Giugno 2019, Minneapolis, vol. 1, pp. 4171-4186, **DOI**: 10.18653/v1/N19-1423

Generazione di contenuti e creatività

Contributo a cura di Sabino **Metta**
Rai - Centro Ricerche, Innovazione Tecnologica e Sperimentazione

Quando si parla di intelligenza artificiale (IA) non si può fare a meno di riportare la nostra memoria ai romanzi o ai film di fantascienza che ci hanno accompagnato fin da quando eravamo bambini. Chi di noi non ha mai letto un libro di Asimov o guardato Frankstein, 2001: Odissea nello spazio, Blade Runner, Terminator, Matrix, ecc.? In generale, le storie sull'IA disegnano scenari distopici in cui i robot, diventati consapevoli delle loro capacità, si rivoltano contro gli esseri umani che li hanno costruiti. Ovviamente, la realtà in cui oggi viviamo è molto diversa e lontana dagli scenari futuristici descritti nei libri o nei film di fantascienza. Pur tuttavia, allo stato attuale, l'IA ha dimostrato di essere in grado di affrontare sfide importanti in diversi ambiti di applicazione. Sono davvero numerosi gli ambiti scientifici (matematica, meteorologia, fisica, medicina, ecc.) ed i settori industriali (media, automotive, sicurezza, agricoltura, ecc.) nei quali l'IA ha dimostrato le sue impressionanti potenzialità. A titolo esemplificativo, riportiamo la recente notizia [Nota 1] dello strabiliante successo ottenuto da **AlphaFold**, il programma di IA basata sull'utilizzo di tecnologie di a*pprendimento profondo (deep learning)* e sviluppata da ***DeepMind*** [Nota 2] (Google). AlphaFold è riuscito a prevedere il ripiegamento delle proteine a partire dalla sequenza di amminoacidi, il cosiddetto protein-folding. Atteso da quasi cinquant'anni, tale traguardo rappresenta una rivoluzione in ambito biologico e medico. La capacità di determinare in maniera molto accurata la forma tridimensionale di una proteina permetterà ai ricercatori di comprendere meglio gli elementi costitutivi delle cellule e delle malattie. Questo risultato potrebbe supportare lo studio di nuove ed avanzate terapie farmacologiche. AlphaFold ha superato i modelli computazionali di oltre cento team di ricerca dimostrando di essere, oltre che molto accurato, anche molto veloce. In alcuni casi ha impiegato circa mezz'ora per determinare la forma di una proteina sulla quale i ricercatori stavano lavorando, senza particolari risultati, da circa dieci anni.

C'è un altro aspetto su cui vogliamo portare l'attenzione del lettore e che costituisce una delle caratteristiche che rendono tali tecnologie davvero promettenti: il tempo di miglioramento. La prima versione del programma risale infatti al 2018, appena due anni fa. Molti esperti concordano nell'affermare che ci troviamo di fronte ad un mutamento di paradigma, ovverosia di fronte ad un cambio delle regole metodologiche e dei criteri di soluzione sinora adottati. In altre parole, l'impiego di tali tecnologie potrebbe rappresentare una vera rivoluzione scientifica.

Ovviamente, c'è ancora molto da studiare ed indagare. Oltre agli aspetti positivi finora descritti, è necessario valutare tutti gli aspetti etici e le ricadute negative che un utilizzo sconsiderato (o semplicemente non pienamente consapevole dei limiti intrinseci esistenti) di tali tecnologie potrebbe comportare.

In analogia con quanto appena descritto, lo sviluppo di tecnologie di IA (in particolare di deep learning) sta investendo anche il mondo dei media. Questo contributo mira in particolare a raccogliere i principali fattori che permettono a tali tecnologie di rivoluzionare il modo con cui i contenuti vengono generati e creati.

Una prima importante sfida di tali tecnologie è quella di automatizzare tutti quei processi editoriali considerati *ripetitivi* e di *routine* (ad es. la produzione di sommari, di riassunti, di reportistiche, ecc.) i quali

Nota 1 - https://www.nature.com/articles/d41586-020-03348-4 *(ultimo accesso 18/12/2020)*

Nota 2 - https://www.deepmind.com/. La stessa DeepMind ha sviluppato il software *AlphaGo* in grado di sconfiggere un maestro umano nel famoso gioco di strategia GO. *(ultimo accesso 18/12/2020)*

201_01.2.5_3.00

impegnano numerose risorse che potrebbero invece essere dirottate su compiti più creativi e pertanto rilevanti. A tal proposito, il contributo introdurrà i principi fondamentali ed alcune importanti applicazioni delle cosiddette **GAN** (*Generative Adversarial Network*), vale a dire l'architettura di deep-learning attualmente di riferimentoper i processi di generazione di contenuti multimediali. A seguire, saranno introdotti alcuni interessanti risultati raggiunti dall'IA nel dominio della *creatività*, un campo ritenuto fino ad oggi di prerogativa dell'essere umano. La parte conclusiva di questo lavoro affronta il tema delle ricadute negative che le tecnologie di IA possono apportare nel mondo dell'informazione. La capacità di generare automaticamente contenuti fotorealistici e sempre più indistinguibili dalla realtà sta infatti acuendo il problema della diffusione di notizie false ed ingannevoli, le cosiddette *fake-news*.

L'architettura di una GAN

In generale, la potenza delle tecnologie di deep learning risiede nella sua abilità di riconoscere *pattern* presenti all'interno di enormi quantità di dati (ad esempio i pixel di una immagine, le transazioni bancarie, le coordinate geografiche, ecc.). Spesso tali pattern non sono riconoscibili dall'occhio umano (in generale dalle limitate capacità cognitive di un essere umano) ma rappresentano in maniera latente le caratteristiche fondamentali di uno specifico insieme di dati. Quando si parla di *generazione* di contenuti nel dominio delle tecnologie di IA si fa abitualmente riferimento ad una particolare famiglia di architetture: la *Rete Generative Avversaria* o *Rete Antagonista Generativa*, dall'inglese *Generative Adversarial Net* (**GAN**). La GAN [1] rappresenta una architettura in grado di riconoscere specifici pattern all'interno di uno specifico contenuto (o da un insieme di contenuti) di riferimento. La stessa architettura è in grado poi di riutilizzare opportunamente i pattern identificati per generare nuovi contenuti. In sintesi, la GAN è una architettura per la stima di modelli generativi (in sostanza i parametri di reti neurali profonde) attraverso un processo avversario, o antagonista, il quale prevede l'allenamento simultaneo di un *modello generativo* e di un *modello discriminativo*. In altre parole, si tratta di due reti neurali profonde che competono tra loro prima di restituire il risultato finale: un contenuto appunto. Di seguito riportiamo i principi fondamentali alla base del funzionamento di una GAN.

Consideriamo, ad esempio, il caso in cui ad una GAN sia assegnato il compito di generare una immagine. Premettiamo che l'obiettivo della GAN è generare una immagine il più possibile fotorealistica. Nella fase di addestramento, un *modello generativo (G)* impara a generare immagini che sembrano reali, mentre un *modello discriminativo (D)* impara a distinguere le immagini reali da quelle false. In altre parole, la rete generativa non è addestrata per minimizzare la *distanza* da una specifica immagine ma piuttosto per *ingannare* in maniera non supervisionata la rete discriminativa. Durante la fase di addestramento, la rete generativa migliora progressivamente creando così immagini sempre più reali. Allo stesso tempo, la fase di addestramento permette alla rete discriminativa di migliorare la prestazione nel distinguere le immagini reali da quelle false. Quando la rete discriminativa non riesce più a distinguere le immagini reali da quelle false allora il processo di generazione ha raggiunto l'equilibrio: *l'immagine generata artificialmente può ritenersi fotorealistica*. Alla base della mutua interazione tra le due reti risiede un caso particolare della teoria dei giochi, il cosiddetto *minimax two-player game* che prevede due giocatori e somma zero.

Entriamo un po' più nel dettaglio. Abbiamo detto che una GAN si compone di due reti profonde, una *Generativa (G)* ed una *Discriminativa (D)*. Prima di addestrare la rete *G*, diamo qualche dettaglio in più su come avviene la generazione di una immagine. Il primo passaggio è quello di generare un segnale z, vale a dire un rumore con una distribuzione normale o uniforme. Tale segnale viene dato come input a *G* per creare una immagine x, ovvero sia $x=G(z)$. Il segnale z rappresenta le caratteristiche latenti (il cosiddetto *latent space*) dell'immagine generata. Ad esempio, il colore e la forma. Tuttavia, all'interno della GAN il significato semantico di z non è sotto controllo. Il processo di addestramento stabilirà quale specifico byte di z è preposto, ad esempio,

al controllo del colore dei capelli all'interno di un'immagine. Questa mancanza di controllo, e la necessità quindi di un addestramento, è tipico di qualunque sistema di classificazione basato sul deep-learning. Da un punto di vista concettuale, *la rete discriminativa guida la rete generativa la quale diversamente genererebbe da sola un rumore casuale*. La rete discriminativa elabora separatamente le immagini reali (che costituiscono il dataset di addestramento) e quelle generate dalla rete generatrice. Dopodiché la rete discriminativa fornisce come output la probabilità *D(x)* che l'immagine di input *x* sia reale (oppure generata). In altre parole, la rete discriminativa si comporta come un classificatore addestrato attraverso una rete neurale profonda. Se la rete discriminativa riceve in input una immagine *x reale*, allora deve essere *D(x)=1*. In caso contrario, vale a dire se la rete discriminativa riceve in input una immagine *x non reale* (perché generata dal generatore e facilmente distinguibile dal dataset di training contenente le immagini reali), allora deve essere *D(x)=0*. Attraverso un processo iterativo, la rete discriminativa è in grado di identificare e catturare le caratteristiche (le cosiddette *features*) che appartengono alle immagini reali. Ricordiamo che l'obiettivo della GAN è quello di generare immagini del tutto indistinguibili da una immagine reale. In sostanza, *la rete generativa deve creare delle immagini che restituiscano un valore di D(x)=1*. A tal proposito, la rete generativa è addestrata *retropropagando* (dall'inglese *backpropagation*) esattamente questo valore di riferimento. Le due reti neurali vengono addestrate iterativamente a passi alterni e costrette a competere mutuamente al fine di migliorare le loro rispettive previsioni. Il processo continuerà fino al raggiungimento del suo equilibrio e, cioè, quando la rete discriminativa non sarà in grado di distinguere tra le immagini reali e quelle generate. A quel punto, il modello GAN sarà in grado di produrre immagini fotorealistiche. Il processo alla base è un problema di ottimizzazione in cui la *funzione di perdita* (*loss function*) è data dalla *cross-entropia* [Nota 3]. In un primo momento, vengono fissati i parametri del *modello G*. Successivamente, viene applicato un algoritmo di *massimizzazione* (generalmente uno *stochastic gradient descent* [Nota 4]) della *funzione di costo* [Nota 5] per la *rete D* usando le immagini reali e quelle generate.

In questa fase la *rete G* non viene addestrata. Dopodiché, il processo si inverte. Viene fissata la *rete discriminatrice D* ed addestrata la *rete generatrice G* attraverso un algoritmo di *minimizzazione* (in questo caso uno *stochastic gradient descent*). Le due reti vengono addestrate alternativamente finché il generatore non è in grado di produrre immagini *buone*, in grado cioè di ingannare la rete discriminatrice. Per semplificare, una GAN è in grado di generare una entità nuova (da lì la necessità di inizializzare la rete generativa con un segnale casuale) a partire da un campione di entità reali il quale funge da insieme di addestramento.

Nella prossima sezione passeremo in rassegna le principali ricerche accademiche condotte sulle GAN. In particolare, ci concentreremo su quelle applicazioni che mirano a supportare il processo di generazione di contenuti.

Nota 3 - Intesa come misura del numero minimo di bit per codificare l'informazione. In formula, *p*log(q)*, dove *p* rappresenta la distribuzione di probabilità dei dati reali e *q* quella delle previsioni calcolate dalla rete neurale

Nota 4 - *Stochastic Gradient Descent*, *Batch Gradient Descent* e *Mini-Batch Gradient Descent* rappresentano algoritmi di ottimizzazione capaci d'individuare il valore minimo di una funzione di costo consentendo di sviluppare un modello previsionale accurato

Nota 5 - la *funzione di perdita* e la *funzione di costo* non sono sinonimi. La prima viene utilizzata per determinare l'errore (la *perdita* appunto) tra l'output dell'algoritmo utilizzato ed il valore target specificato. La funzione di perdita viene utilizzata principalmente su un singolo set di addestramento. La *funzione di costo* può essere calcolata come media delle funzioni di perdita e calcola una *penalità* per un numero maggiore di set di addestramento.

Processo generativo di contenuti

Dal lontano 2014, anno in cui viene pubblicato il primo articolo sulle GAN [1], il numero di pubblicazioni relativo all'impiego di tali architetture è cresciuto in maniera esponenziale (circa 700 pubblicazioni nel 2019) [2]. L'estrema versatilità di tali architetture ha permesso alla ricerca di compiere numerosi passi in avanti migliorando le performance nell'elaborazione dei dati e dando origine a nuove numerose tipologie [Nota 6] di GAN (*3D-ED-GAN*, *ABC-GAN*, *ACtuAL-GAN*, *AL-CGAN*, *AmbientGAN* per nominarne solo alcune).

Il sito web **This X Does Not Exist** [Nota 7] raccoglie una selezione di entità (oggetti, ambientazioni, animali, servizi, ecc.) che *non esistono* nella realtà. Tali entità ***X*** sono state invece generate artificialmente ed in maniera davvero fotorealistica.

In questa sezione passeremo in rassegna alcuni particolari applicazioni. In particolare, l'obiettivo è quello di mettere in evidenza la capacità delle GAN nell'affrontare e nel risolvere alcuni importanti compiti che sono tradizionalmente svolti in maniera manuale all'interno di un processo di produzione di contenuti. Tali architetture dimostrano di essere in grado di identificare e generare in maniera soddisfacente specifici schemi (*pattern*) spesso non visibili o catturabili dall'occhio umano (o in generale dalle capacità cognitive di un essere umano) ed in questa maniera di assolvere alla risoluzione di compiti spesso *ripetitivi* ed assai impegnativi per un essere umano. Da qui si capisce facilmente la portata rivoluzionaria delle GAN che promettono quindi di sollevare il lavoro umano dallo svolgere determinati compiti automatizzabili. Le risorse risparmiate attraverso l'impiego di tali tecnologie possono essere riutilizzate e convogliate verso altre attività, magari caratterizzate da un più alto livello di pensiero creativo.

Un primo esempio di come le tecnologie di deep learning possono entrare nelle vite di tutti i giorni è dato dal nostro TV di casa. Se oggigiorno è relativamente facile (ed economico) avere in casa un televisore 4K, non lo è altrettanto fruire di contenuti che nativamente siano stati prodotti nello stesso formato. A meno di casi particolari (ad esempio utilizzare un lettore Blu-Ray UltraHD oppure avere un abbonamento ad un provider che offra contenuti nativi in 4K) un televisore 4K (o per chi lo possiede un 8K) sfrutta spessissimo l'*upscaling*. Si tratta di un meccanismo di *interpolazione* che permette di generare nuovi pixel o correggere quelli già presenti. Tra i diversi *interpolatori* esistenti (*nearest-neighbor interpolation*, *bilineare*, *bi-cubica*, *Fourier based*, ecc.) troviamo appunto quelli basati sul deep learning.

Un'altra applicazione molto promettente è quella che permette di aumentare la risoluzione delle immagini o dei personaggi *anime* [3]. Nel lavoro [4] i ricercatori dimostrano le potenzialità delle cosiddette **SRGAN** nel migliorare in maniera fotorealistica di un fattore *4x* la risoluzione di una immagine di partenza.

Il processo produttivo di animazioni o di *cartoonizzazione* (creazione di immagini cartoon a partire da foto reali) è molto dispendioso e può richiedere l'impiego di numerosi artisti e disegnatori [5]. Nel lavoro [6], i ricercatori mettono a punto un modello per supportare la creazione di personaggi *anime*. Al fine di assistere tale creazione, gli autori hanno messo anche a disposizione una pagina web [Noya 8] in cui è possibile impostare alcuni parametri (colore dei capelli, gli occhiali, il sorriso, ecc.) e generare così nuovi personaggi.

Nota 6 - https://github.com/hindupuravinash/the-gan-zoo
(ultimo accesso 18/12/2929)

Nota 7 - https://thisxdoesnotexist.com/
(ultimo accesso 18/12/2020)

Nota 8 - https://make.girls.moe/#/
(ultimo accesso 18/12/2020)

Le architetture proposte in [7] e [8] permettono di trasformare una immagine di partenza (possiamo dire quella *reale*) in una immagine *contestualmente* differente. Tali tecnologie permettono di trasformare il soggetto principale di una foto reale in un altro soggetto, come mostrato nell'esempio di Fig. 1 relativo all'architettura **CycleGAN**.

L'architettura **StarGAN** proposta in [9] permette di trasformare l'espressione di un volto all'interno di una immagine di partenza: una faccia sorridente viene trasformata in una arrabbiata, felice o spaventata, ma la trasformazione del dominio di partenza può essere anche più *invasiva* ed interessare, ad esempio, per lo stesso volto, i suoi capelli, il sesso, l'età ed il colore della pelle come mostrato in Fig. 2.

L'utilizzo delle GAN interessa anche il cosiddetto *color grading*. In [10] i ricercatori propongono un algoritmo in grado di trasferire i colori (e di conseguenza le condizioni di illuminamento) tra immagini che condividono strutture semanticamente simili. Un'altra architettura [11] punta alla trasformazione della *texture* delle immagini a partire da un segnale (anche random) di riferimento.

Le GAN hanno dimostrato di generare dipinti [12] e ritratti artistici a partire da foto reali [13]. Viceversa, nel lavoro [14] tali architetture creano fotografie realistiche di volti a partire da semplici bozze o schizzi. Un'altra interessante applicazione, sempre nel dominio artistico, è la creazione di dipinti artistici a partire da semplici *schizzi* [15].

Fig. 1 – *CycleGAN*: trasformazione del dominio di una foto

Fig. 2 – *StarGAN*: trasformazione delle caratteristiche di un volto

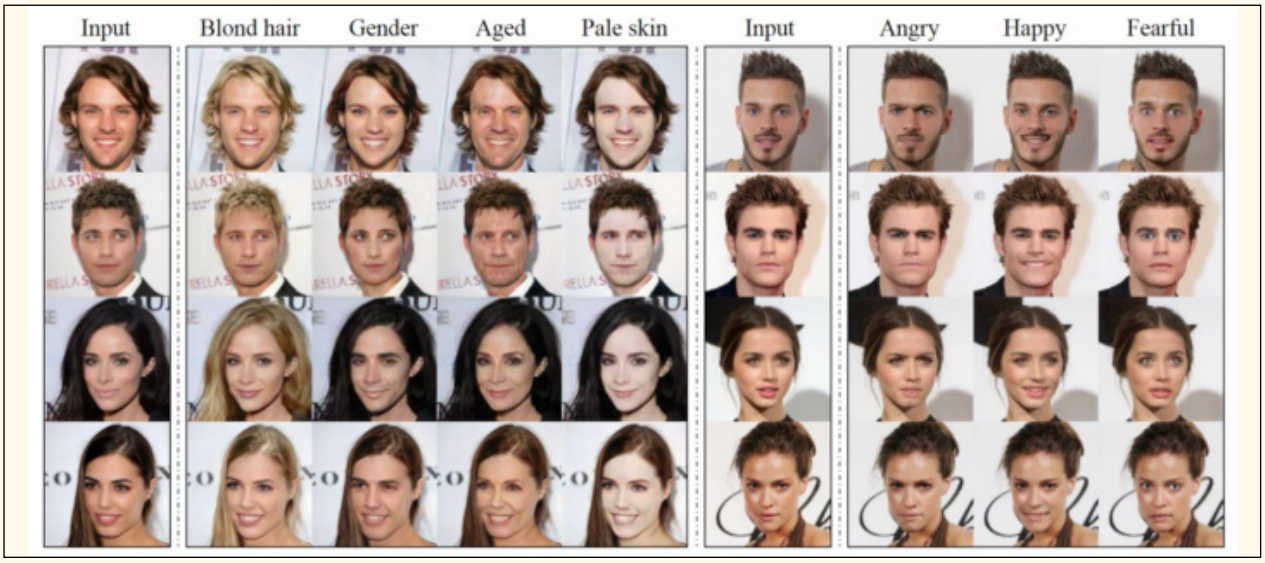


Fig. 3 – Volti (artificiali) da celebrità: GAN addestrata sul dataset *CelebA*

Nel lavoro [16] i ricercatori utilizzano **CelebA** Nota 9, un dataset di addestramento costituito da oltre dieci mila volti di personaggi celebri. Il risultato è la generazione di *nuove* (e *sintetiche*) celebrità in grado di apparire perfettamente a loro agio sul red carpet, vedi Fig. 3.

In [17] invece, l'obiettivo dei ricercatori è quello di generare, per un determinato volto, diverse angolature di ripresa a partire da una singola immagine di partenza. Le architetture proposte in [18] permettono di generare una collezione di immagini fotorealistiche che semanticamente si avvicinano alle parole contenute in un testo fornito come input.

In [19] i ricercatori propongono un'architettura in grado di invecchiare in maniera automatica e realistica un volto di riferimento. A tal proposito, citiamo la recente discussione sul film **The Irishman** prodotto da ***Netflix*** . In questo film, che ha riscosso un certo successo, si è discusso molto sui risultati mediocri del lavoro di *ringiovanimento* compiuto sui due attori principali Robert De Niro e Al Pacino. La discussione Nota 10 verte infatti sul fatto che le metodologie e tecnologie tradizionali (non basate sul deep learning) siano costate svariati milioni di dollari a fronte di un risultato assolutamente mediocre. Ad acuire tale discussione infatti, si inserisce un video *fake* pubblicato dallo Youtuber (e *deepfaker*) ***Shamook*** il quale mette a confronto alcune immagini originali del film con quelle create attraverso le tecnologie di deep learning. Il confronto è assolutamente impressionante, le tecnologie di apprendimento automatico restituiscono un ringiovanimento del volto assolutamente più naturale e verosimile, dimostrando la straordinaria capacità di tali tecnologie nel catturare in maniera incomparabile strutture e caratteristiche fondamentali all'interno dei dati di partenza.

L'architettura proposta nel lavoro [20] si propone di risolvere il problema della generazione e soprattutto della ricostruzione di *oggetti tridimensionali* a partire da una immagine *2D* che costituisce lo spazio di probabilità. In questo lavoro i ricercatori dimostrano la capacità della GAN di catturare implicitamente la struttura di un oggetto e di generare oggetti 3D di alta qualità.

In [21] viene presentata un'architettura in grado di eliminare in maniera automatica la *sfocatura* (*blurring*) all'interno di un'immagine. Parallelamente, la GAN è in grado di sintetizzare immagini sfocate a partire da immagini a fuoco ed in questa maniera migliorare le operazioni di *data augmentation* in caso di ulteriori addestramenti.

Nota 9 - http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
(ultimo accesso 18/12/2020)

Nota 10 - https://www.creativebloq.com/news/the-irishman-deepfake
(ultimo accesso 18/12/2020)

Le GAN hanno dimostrato le loro sorprendenti potenzialità anche nel mondo della musica. Recentemente numerosi studi e sperimentazioni sono stati compiuti nella generazione di melodie musicali [22]. ***OpenAI***, l'organizzazione non profit di ricerca sull'intelligenza artificiale, ha recentemente rilasciato **Jukebox** Nota 11. Tale architettura è in grado di generare canzoni orecchiabili in una varietà di stili diversi (teenybop, country, hip-hop, heavy metal, ecc.) ed a partire da semplici input (un genere, un artista, un testo, o i primi secondi di una canzone).

Anche **IBM Watson Beat** Nota 12 si propone l'obiettivo di supportare gli artisti nella creazione di composizioni originali attraverso l'utilizzo di tecnologie di intelligenza artificiale. Recentemente poi, diverse start-up ed iniziative di ricerca hanno puntato sull'utilizzo dell'intelligenza artificiale per la produzione musicale innovativa: **AIVA** Nota 13, **Amper Music** Nota 14, **Google's Magenta** Nota 15, **Sony's Flow Machines** Nota 16, **Jukedeck** Nota 17, **Humtap** Nota 18 ed altre ancora.

Sebbene le tecnologie di deep learning abbiano dimostrato impressionanti potenzialità, gli esperti sostengono che siamo ancora lontani dal riconoscere in tali tecnologie la *vera arte*. Le creazioni compiute attraverso tali tecnologie necessitano ancora di un input umano (per lo meno in una fase iniziale). L'intelligenza artificiale è utilizzata principalmente per ridurre le risorse (soprattutto in termini di tempo) spese nello svolgimento di compiti ripetitivi che spesso si incontrano all'interno del processo di produzione.

Una grande sfida per l'intelligenza artificiale è quella di catturare e comprendere i pattern relativi alle decisioni artistiche e creative. Su questa aspetto, nemmeno i più famosi esperti (umani) riescono a convergere verso una interpretazione comune. Ad ogni modo, l'avvento delle tecnologie di deep learning sta cominciando a cambiare il modo con cui gli artisti creano. Diversi musicisti e compositori stanno collaborando con esperti di tecnologie dimostrando di essere disposti ad esplorare nuovi processi di creatività, eventualmente intrecciati con le tecnologie di IA.

Nella prossima sezione andremo ad approfondire le potenzialità da parte di una intelligenza artificiale *di esprimere, alla stregua di un artista in carne ed ossa, capacità creative ed artistiche.*

Creatività: quando l'IA diventa artista

Finora abbiamo mostrato la straordinaria capacità delle tecnologie di apprendimento automatico nello svolgimento di compiti complessi anche nel dominio dell'arte. Le architetture di deep learning stanno dimostrando di risolvere compiti che richiederebbero enormi risorse (soprattutto in termini di tempo) anche ad esperti professionisti. Tuttavia, si tratta sempre di compiti che non afferiscono al dominio della creatività e che rispondono direttamente ad una necessità operativa di un essere umano. In questa sezione invece, mostreremo alcuni interessanti sperimentazioni condotte nel dominio della creatività e dell'arte.

La nostra intenzione è quella di mostrare gli sviluppi delle tecnologie di IA ed il modo con cui tali tecnologie stanno modificando la convinzione finora ritenuta inviolabile che un atto creativo ed artistico possa essere compiuto unicamente da un essere umano.

Nota 11 - https://openai.com/blog/jukebox

Nota 12 - https://www.ibm.com/case-studies/ibm-watson-beat

Nota 13 - https://www.aiva.ai/creations

Nota 14 - https://www.ampermusic.com/

Nota 15 - https://research.google/teams/brain/magenta/

Nota 16 - https://www.sonycsl.co.jp/tokyo/2811/

Nota 17 - https://www.jukedeck.com/

Nota 18 - https://www.humtap.com/

(per tutti ultimo accesso 18/12/2020)

Prima di mostrare alcuni lavori artificialmente creativi, vale la pena introdurre alcune criticità note dell'IA.

Le tecnologie di apprendimento profondo sono spesso accusate di opacità e poca interpretabilità. Le loro reti profonde sono costituite da molteplici livelli di neuroni, specifiche funzioni di attivazione e numerosi parametri. Come abbiamo più volte espresso, queste sofisticate architetture permettono il riconoscimento di pattern non visibili da un essere umano. D'altra arte, le stesse architetture non fanno trasparire la catena logica che conduce ai suoi risultati. In alcune situazioni, vedi ad esempio nel supporto di sistemi decisionali, questa mancanza di interpretabilità costituisce un limite serio. Infatti, se una decisione si rivelasse critica, sarebbe doveroso poter risalire alle motivazioni per cui l'IA ha preso quella particolare decisione.

Un altro aspetto critico è dato dalla rappresentatività del dataset iniziale di addestramento. La comunità mondiale sta investigando i rischi seri legati ad eventuali *bias* o meglio *pregiudizi* che tali tecnologie potrebbero mostrare nello svolgimento di importanti compiti a loro assegnati. Un caso emblematico è quello rappresentato dal lavoro di ***Joy Buolamwini***, ricercatrice del **MIT Media Lab**, che nel 2016 ha fondato la ***Algorithmic Justice League*** Nota 19 con lo scopo di identificare ed evidenziare nei codici informatici pregiudizi che possono portare alla discriminazione contro i gruppi sottorappresentati.

A tal proposito, alcune comunità scientifiche si stanno concentrando sulle metodologie e sulle tecnologie per esplorare e visualizzare (e di conseguenza meglio comprendere) i meccanismi con i quali dinamicamente le reti apprendono e di conseguenza generano le loro risposte. Citiamo a titolo esemplificativo le soluzioni **Microscope** Nota 20 lanciato da ***OpenAI***, **TensorBoard** Nota 21 lanciato da ***TensorFlow*** (Google) e **GrandTour** Nota 22 [23].

D'altra parte, la scarsa interpretabilità intrinseca delle tecnologie di apprendimento profondo trova un interessante punto di contatto con il processo creativo umano in grado di creare oggetti *inaspettati*, a volte riconosciuti come artistici.

Il modo dell'arte sta sperimentando da qualche anno l'impiego di tecnologie di IA per creare delle vere e proprie opere. In questo caso si parla di *creatività computazionale*. Per fare qualche esempio riportiamo il caso di un libro Nota 23 selezionato per un premio letterario, alcune poesie Nota 24 , testi poetici Nota 25 ed addirittura un musical Nota 26. Nel 2016, ventinove opere realizzate da **Google AI** Nota 27 sono state vendute all'asta a San Francisco. ***Christie's***, la più grande casa d'aste a livello mondiale, nel 2018 ha battuto all'asta per oltre *quattrocento mila dollari* un'opera creata artificialmente (Fig. 4). L'opera, il ritratto del fantomatico *Edmond de Belamy* Nota 28 è stata generata da una GAN a partire da circa quindicimila dipinti creati, questa volta da artisti umani, tra il quattordicesimo e il ventesimo secolo.

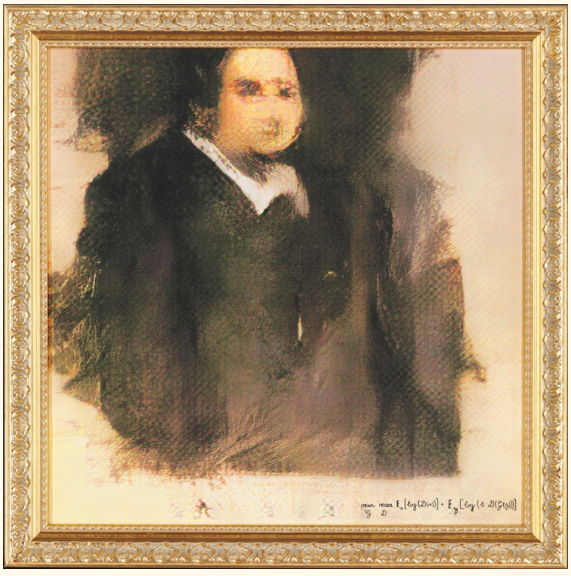


Fig. 4 – *Edmond de Belamy* (2018), creato attraverso una GAN

Nota 19 - https://www.ajl.org/
Nota 20 - https://microscope.openai.com/about
Nota 21 - https://www.tensorflow.org/tutorials
Nota 22 - https://distill.pub/2020/grand-tour/
Nota 23 - https://www.digitaltrends.com/cool-tech/japanese-ai-writes-novel-passes-first-round-nationanl-literary-prize/
Nota 24 - https://www.theguardian.com/technology/2016/may/17/googles-ai-write-poetry-stark-dramatic-vogons
Nota 25 - https://www.gwern.net/GPT-3
Nota 26 - https://www.newscientist.com/article/2079483-beyond-the-fence-how-computers-spawned-a-musical/
Nota 27 - https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html
Nota 28 - https://www.christies.com/Lotfinder/lot_details.aspx?sid=&intObjectID=6166184&T=Lot&language=en
(per tutti ultimo accesso 18/12/2020)

Alcuni artisti stanno esplorando proattivamente la *creatività computazionale* dell'intelligenza artificiale, come ad esempio ***Mario Klingemann***. L'artista tedesco ha addestrato le reti profonde con immagini eterogenee: dipinti di arte classica, selfie di sé stesso e fotografie tratte da Instagram. L'artista si è spinto oltre andando a modificare la struttura di una GAN di partenza e selezionando infine quelle opere che secondo lui meglio esprimevano il concetto di arte. Si veda ad esempio la serie di opere **Neural Glitch** Nota 29 (2018).

Rispetto al tema dell'esplorazione dei meccanismi alla base della creatività, un lavoro molto interessante è quello proposto dai ricercatori dell'***Art and Artificial Intelligence Lab*** della ***Rutgers University***. In questo lavoro [24] i ricercatori hanno studiato una architettura ad-hoc per modellare la creatività computazionale, la cosiddetta *Creative Adversarial Network (CAN)*. I ricercatori sono partiti dagli studi di ***Colin Martindale***, un noto professore di psicologia. Secondo Martindale, esperto di creatività e processi artistici, lo sviluppo artistico nel corso del tempo è il risultato di una continua ricerca di *novità* da parte degli artisti. Ogni artista è continuamente esposto al lavoro di altri artisti e, in generale, ad una larga varietà di arte nel corso della sua vita. Ciò che rimane sconosciuto è come l'artista coniughi la propria conoscenza con la propria abilità di creare nuove opere.

Secondo le teorie di Martindale, un artista cerca di affermare la propria arte per *combattere* l'assuefazione alla particolare arte a cui è esposto. Tale tentativo di affermazione però non deve esagerare per evitare una reazione negativa da parte dei potenziali fruitori. I ricercatori della Rutgers University hanno quindi cercato di modellare il meccanismo attraverso il quale un artista integra la sua conoscenza (esposizione all'arte) con la propria spinta creativa. Basandosi sui principi di Martindale, i ricercatori hanno così creato un'architettura in grado di generare un'opera artistica, vedi Fig. 5. Insomma, la prossima volta che ci emozioneremo di fronte ad un dipinto, potrebbe essere stato merito di una IA.

Generazione di *deep fake*: la disinformazione nei media

Finora abbiamo visto come le tecnologie di deep learning siano estremamente efficaci nel ridurre i costi (ed i tempi) nel risolvere compiti ripetitivi in un processo di produzione di contenuti. Allo stesso tempo, abbiamo mostrato le promettenti sperimentazioni nell'impiego di tali tecnologie nel processo creativo potenziando quindi il lavoro di artisti e performers. In questa sessione ci concentreremo sui rischi associati alla straordinaria capacità da parte delle tecnologie di IA di generare contenuti falsi e sempre più indistinguibili dalla realtà.

Fig. 5 – Esempio di *arte* creata da una CAN

Nota 29 - https://underdestruction.com/2018/10/28/neural-glitch/ *(ultimo accesso 18/12/2020)*

Oggigiorno, con il termine *deep fake* si fa riferimento all'utilizzo di tecnologie di deep learning per generare un contenuto falso (*fake* appunto). Un'operazione comune è quella di sostituire, ad esempio, il soggetto all'interno di un video con un personaggio noto. O viceversa. Uno dei primi lavori di ricerca [25] è stato quello di creare un video *fake* dell'ex Presidente degli Stati Uniti d'America, Barack Obama. Una rete neurale è stata addestrata attraverso numerosi discorsi del Presidente imparando ad associare l'audio alla forma della bocca. In questa maniera, successivamente è stato possibile ricreare, in maniera fotorealistica, il video del Presidente mentre recitava artificialmente discorsi mai pronunciati prima.

In brevissimo tempo, numerose applicazioni sono state sviluppate e messe a disposizione di qualunque utente: **FakeAPP**, **Faceswap** (open-source), **DeepFaceLab** (open-source), **Doublicat**, **DeepFakes**, **NeuralTextures**, **RefaceAI** ed altre ancora. A quel primo (ed innocuo) *deep fake* di Obama sono susseguiti (e continuano ancora adesso) una miriade di *fake*. Youtube raccoglie intere sezioni dedicate alla creazione di tali video. Le tecnologie di IA (vedi ad esempio Lyrebird) permettono di emulare anche la voce, ad esempio di personaggi famosi. In questa maniera è possibile scambiare il timbro di voce tra due diversi soggetti. Al di là dei *meme* che affollano le nostre chat online, una forte eco mediatica si è alzata a causa dell'uso malevolo che si sta facendo di questa tecnologia: la creazione di video pornografici falsi ritraenti celebrità, il *revenge porn*, il *cyber-bullismo* e le *fake news*.

Ad inizio settembre il quotidiano ***The Guardian*** ha utilizzato **GPT-3** Nota 30 (***OpenAI***), l'ultima versione di un software di intelligenza artificiale per la produzione automatica di testi, per scrivere un editoriale sull'utilizzo di AI. In verità, un giornalista (umano) del quotidiano britannico, ha dato a GPT-3 alcune istruzioni scritte e le prime righe del pezzo. A quel punto, il software ha prodotto in pochi secondi otto differenti editoriali sul tema richiesto. Al fine di restituire al lettore le straordinarie capacità dell'intelligenza artificiale, il ***The Guardian*** ha mescolato gli otto pezzi e ricombinati assieme per creare un unico documento. *"Per cominciare, non ho il desiderio di spazzare via la razza umana"*, così inizia la parte dell'editoriale scritta dall'intelligenza artificiale. Ovviamente, le potenzialità sono enormi: GPT-3 è in grado di scrivere, tradurre, comprendere testi, rispondere a domande e scrivere codici informatici. I pericoli legati ad un uso malevolo di questa tecnologia sono facilmente visibili. La stessa OpenAI, azienda proprietaria di GPT-3, ha dichiarato di essere preoccupata dall'abuso che potrebbe essere fatto ed ha pertanto bloccato l'accesso pubblico alle API ai fini dello studio e della ricerca.

In conclusione, le tecnologie di deep learning e le loro prestazioni stanno dimostrando di crescere a ritmi vertiginosi. Allo stesso tempo, tali tecnologie si diffondono molto rapidamente anche grazie alla facilità di reperirle sotto forma di semplici applicazioni per un comune telefono cellulare. Non è un caso che il tema della *falsificazione dei contenuti* (*deep fake*) e della *disinformazione* (*information disorder*) sia un problema affrontato a livello mondiale.

Volendo chiudere con un messaggio di speranza, le stesse tecnologie stanno dimostrando di essere gli strumenti più adatti per smascherare la diffusione di contenuti falsi ed ingannevoli.

Conclusioni

Il cambio di paradigma apportato dalle tecnologie di intelligenza artificiale è sotto gli occhi di tutti.

In questo lavoro abbiamo cercato di mettere in evidenza l'impiego delle tecnologie di apprendimento profondo nel mondo dei media, in particolare nel processo di produzione dei contenuti. Tali tecnologie stanno dimostrando di risolvere in maniera eccellente un grande numero di compiti ripetitivi e dispendiosi e, allo stesso tempo, stanno tentando di catturare ed interpretare i meccanismi alla base della creatività generalmente associata all'essere umano.

Nota 30 - https://github.com/openai/gpt-3
(ultimo accesso 18/12/2020)

Per contro, la rivoluzione apportata da tali strumenti non è a costo zero. La capacità di generare contenuti falsi e sempre più indistinguibili dalla realtà sta agevolando il dilagarsi di contenuti falsi ed ingannevoli. In altre parole, la potenza di tali tecnologie sta acuendo alcuni delicati problemi sociali, tra cui quello della disinformazione. Lo scopo di questo lavoro è quello di stimolare una riflessione sulla necessità, soprattutto da parte di un *Servizio Pubblico*, di sostenere lo studio, la ricerca, la sperimentazione e l'educazione rispetto a queste dirompenti tecnologie.

Bibliografia

[1] I. Goodfellow ed altri, *Generative Adversarial Nets*, in "NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems", vol. 2, 2014, pp. 2672–2680, https://dl.acm.org/doi/10.5555/2969033.2969125

[2] Z. Farou, N. Mouhoub e T. Horvath, *Data Generation Using Gene Expression Generator*, 2020, DOI: 10.13140/RG.2.2.24193.48483/2

[3] Chao Dong ed altri, *Image Super-Resolution Using Deep Convolutional Networks*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", vol. 38, n. 2, 2016, pp. 295-307, **DOI**: 10.1109/TPAMI.2015.2439281

[4] C. Ledig ed altri, *Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network*, in "2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)", 2017, **DOI**: 10.1109/CVPR.2017.19

[5] Yang Chen, Yu-Kun Lai e Yong-Jin Liu, *CartoonGAN: Generative Adversarial Networks for Photo Cartoonization*, in "2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition", 2018, **DOI**: 10.1109/CVPR.2018.00986

[6] Yanghua Jin ed altri, *Towards the Automatic Anime Characters Creation with Generative Adversarial Networks*, in "Comiket 92", 2017, arXiv:1708.05509

[7] Jun-Yan Zhu ed altri, *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*, in "2017 IEEE International Conference on Computer Vision (ICCV)", 2017, **DOI**: 10.1109/ICCV.2017.244

[8] Ming-Yu Liu e Oncel Tuzel, *Coupled Generative Adversarial Networks*, in "NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems", 2016, pp. 469-477, https://dl.acm.org/doi/10.5555/3157096.3157149

[9] Yunjey Choi ed altri, *StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation*, in "2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition ", 2018, **DOI**: 10.1109/CVPR.2018.00916

[10] Mingming He ed altri, *Progressive Color Transfer with Dense Semantic Correspondences*, in "ACM Transactions on Graphics", vol. 38, n. 2, 2019, articolo n. 13, **DOI**: 10.1145/3292482

[11] C. Li e M. Wand, *Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks*, 2016, arXiv:1604.04382

[12] L. Gatys, A. Ecker e M. Bethge, *A Neural Algorithm of Artistic Style*, 2015, arXiv:1508.06576

[13] R. Yi ed altri, *APDrawingGAN: Generating Artistic Portrait Drawings From Face Photos With Hierarchical GANs*, in "2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)", 2019, **DOI**: 10.1109/CVPR.2019.01100

[14] S. Chened altri, *DeepFaceDrawing: Deep Generation of Face Images from Sketches*", in "ACM Transactions on Graphics", vol. 39, n. 4, 2020, articolo n. 72, **DOI**: 10.1145/3386569.3392386

[15] A. Xue, *End-to-End Chinese Landscape Painting Creation Using Generative Adversarial Networks*, pre-print 2020, arXiv:2011.05552

[16] T. Karras ed altri, *Progressive Growing of GANs for Improved Quality, Stability, and Variation*, in "ICLR 2018 - Sixth International Conference on Learning Representations", 2018, https://iclr.cc/Conferences/2018/Schedule?showEvent=204

[17] R. Huang ed altri, *Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis*, in "2017 IEEE International Conference on Computer Vision (ICCV)", 2017, **DOI**: 10.1109/ICCV.2017.267

[18] H. Zhang ed altri, *StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks*, in "2017 IEEE International Conference on Computer Vision (ICCV)", 2017, **DOI**: *10.1109/ICCV.2017.629*

[19] G. Antipov, M. Baccouche e J. Dugelay, *Face aging with conditional generative adversarial networks*, in "2017 IEEE International Conference on Image Processing (ICIP)", 2017, **DOI**: *10.1109/ICIP.2017.8296650*

[20] Jiajun Wu ed altri, *Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling*, in "NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems", 2016, pp. 82-90, *https://dl.acm.org/doi/10.5555/3157096.3157106*

[21] O. Kupyn ed altri, *DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks*, in "2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition", 2018, **DOI**: *10.1109/CVPR.2018.00854*

[22] Li-Chia Yang, Szu-Yu Chou e Yi-Hsuan Yang, *MidiNet: A Convolutional Generative Adversarial Network for Symbolic-Domain Music Generation*, in "Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)", 2017, pp. 324-331, *https://drive.google.com/file/d/0BwDuTuc57K1EN2ZBNExLaHFXZlE/view*

[23] D. Asimov, *The Grand Tour: A Tool for Viewing Multidimensional Data*, in "SIAM Journal on Scientific and Statistical Computing", vol. 6, n. 1, 1985, pp. 128-143, **DOI**: *10.1137/0906011*

[24] A. Elgammal ed altri, *CAN: Creative Adversarial Networks Generating "Art" by Learning About Styles and Deviating from Style Norms*, *arXiv:1706.07068*

[25] S. Suwajanakorn, S. Seitz e I. Kemelmacher-Shlizerman, "*Synthesizing Obama: Learning Lip Sync from Audio*", in "ACM Transactions on Graphics", vol. 36, n. 4, 2017, articolo n. 95, **DOI**: *10.1145/3072959.3073640*

Machine Learning per la Sottotitolazione Automatica

Carmen **Marino**, Mauro **Rossini**
Rai - Centro Ricerche, Innovazione Tecnologica e Sperimentazione

L' Intelligenza Artificiale e le tecnologie di Machine Learning per la Trascrizione Automatica permettono di ipotizzare nuovi scenari applicativi rivolti a servizi per le persone con disabilità cognitive e sensoriali. I contesti dove l'Intelligenza Artificiale incontra le necessità e aspirazioni di un settore di pubblico che desidera esplorare nuove soluzioni per migliorare la propria identità sociale diventano un terreno di sfida per le aziende che erogano servizi al cittadino. In questi scenari si incontrano le richieste per un incremento di contenuti accessibili e comprensibili per tutti a cui il servizio pubblico deve rispondere con innovazione tecnologica e investimenti sulla ricerca.

La Rai *ha voluto accettare questa sfida tecnologica nel settore dei Servizi per le persone con disabilità sensoriali, rispondendo alla forte richiesta di incremento delle ore sottotitolate per le persone sorde, ipoacusiche e anziane.*

Il Centro Ricerche, Innovazione Tecnologica e Sperimentazione *ha attivato una sperimentazione sulla Sottotitolazione Automatica dei Telegiornali Regionali, un progetto innovativo, sfidante e sicuramente non privo di complessità. L'analisi delle prestazioni del sistema di sottotitolazione automatica, seppur richiedendo un minimo intervento di correzione manuale prima della messa in onda, ha evidenziato elevati valori di accuratezza sulle parole trascritte. L'adozione di tale soluzione tecnologica, che prevede l'impiego nei processi di validazione/correzione di figure professionali senza specifiche competenze di stenotipia, potrebbe ridurre i costi del servizio di sottotitolazione e garantire la sostenibilità del progetto.*

Le *tecnologie di Machine Learning per la Trascrizione Automatica* permettono di ipotizzare nuovi scenari applicativi rivolti essenzialmente a servizi per le persone con disabilità cognitiva e sensoriale. I contesti dove l'*Intelligenza Artificiale* incontra le necessità e le aspirazioni di un settore di pubblico che desidera esplorare nuove soluzioni per migliorare la propria identità sociale diventano un terreno di sfida per le aziende che erogano servizi al cittadino cercando di includere tutte le tipologie di pubblico a cui si rivolgono.

In questi scenari si incontrano le richieste per un incremento di contenuti accessibili e comprensibili per tutti a cui il servizio pubblico deve rispondere con innovazione tecnologica e investimenti nella ricerca.

La **Rai** ha voluto accettare questa sfida tecnologica nel settore dei *Servizi per le persone con disabilità sensoriali*, rispondendo alla forte richiesta di incremento delle ore sottotitolate per le persone sorde, ipoacusiche e anziane.

La televisione è uno strumento sia di intrattenimento sia di approfondimento culturale e accompagna la persona nel suo percorso di socializzazione e inclusione nella vita quotidiana.

La possibilità di incrementare le ore di trasmissioni televisive sottotitolate tramite l'adozione di soluzioni innovative che introducano tecnologie atte a produrre contenuti ad oggi non disponibili diventa

201-02.1_3.00

un'opportunità di sperimentare sul campo le potenzialità di questi sistemi, seppur evidenziandone tutte le criticità ancora irrisolte.

Il mondo delle *news* è sicuramente il contesto più sfidante e riveste un ruolo di condivisione delle informazioni vitali in una società globalizzata. L'obbligo di diffondere un'informazione corretta diventa una sfida etica e contemporaneamente tecnologica, dove la *semantica*, ovvero la ricerca del significato del messaggio, riveste un ruolo fondamentale per la comprensione della notizia.

Per questo motivo, la trascrizione automatica e la successiva sottotitolazione applicata alle news non deve solamente focalizzarsi sul riportare correttamente la parola ma deve polarizzarsi nel trasferire il corretto messaggio che si intende veicolare.

In questo ambito è stata attivata una *Sperimentazione Rai sulla Sottotitolazione Automatica delle news regionali*, un progetto innovativo, sfidante e sicuramente non privo di complessità.

Il Progetto Rai

Il nuovo *Contratto di Servizio Rai* [Nota 1] prevede, sui temi dell'Accessibilità e Inclusione Sociale, l'estensione dell'offerta di contenuti sottotitolati e audio.

L'*Articolo 25*, che tocca, tra gli altri, il tema dei servizi rivolti alle *Persone con disabilità*, recita (comma *1*, lettera *h*) che la **Rai**, ai fini dell'attuazione della missione di servizio pubblico, è tenuta a:

> *"[...] estendere progressivamente la sottotitolazione e le audiodescrizioni anche alla programmazione dei canali tematici, con particolare riguardo all'offerta specificamente rivolta ai minori; [...] estendere progressivamente la fruibilità dell'informazione regionale [...]".*

La **Rai** ha pertanto attivato una politica che prevede l'introduzione progressiva della sottotitolazione dei *Telegiornali Regionali* per tutte le regioni. La sede di Bolzano dal mese di marzo 2017 provvede già alla sottotitolazione delle notizie dell'edizione delle ore 20 del *Tagesschau*, il telegiornale locale in lingua tedesca.

Per rispondere a questa indicazione editoriale nasce il progetto *Sperimentazione Sottotitolazione Automatica TGR* che ha l'obiettivo di verificare, in termini tecnici, operativi, qualitativi ed economici, la possibilità di utilizzare una *Piattaforma automatica di Generazione Sottotitoli* nell'ambito della produzione dei sottotitoli dei TG Regionali.

Il **Centro Ricerche, Innovazione Tecnologica e Sperimentazione Rai** ha recentemente condotto un'analisi delle prestazioni di alcuni sistemi commerciali di *trascrizione automatica del parlato* in lingua italiana che, seppur richiedendo un processo di correzione manuale e inserimento della punteggiatura prima della messa in onda, hanno dimostrato elevati valori di accuratezza sulle parole trascritte e di *accuratezza semantica*, indice della correttezza del significato del messaggio trascritto rispetto all'originale. La tesi della sperimentazione è che l'adozione di tale soluzione tecnologica, che prevede l'impiego nei processi di validazione/correzione figure professionali senza specifiche competenze di stenotipia, potrebbe potenzialmente ridurre i costi del servizio di sottotitolazione e garantire la sostenibilità del progetto.

La Sfida

La sfida di utilizzare sistemi di trascrizione e di generazione automatica dei sottotitoli per il campo delle news è avvalorata dalla continua evoluzione in termini di precisione dei sistemi di *machine learning*, con la consapevolezza, però, della criticità del campo specifico di applicazione.

Nota 1 - Il *CONTRATTO NAZIONALE DI SERVIZIO TRA IL MINISTERO DELLO SVILUPPO ECONOMICO E LA RAI-RADIOTELEVISIONE ITALIANA S.P.A. 2018-2022* è disponibile all'indirizzo http://www.rai.it/dl/doc/1521036887269_Contratto%202018%20testo%20finale.pdf

Nel configurare uno scenario applicativo così particolare occorre tener presente alcune considerazioni nodali:

- nel *contesto news* non solo è importante l'*accuratezza sulle parole* (*Word Accuracy-WA*) che il sistema di trascrizione automatica permette di raggiungere, ovvero la percentuale di parole trascritte correttamente, ma riveste un ruolo fondamentale anche l'*accuratezza semantica* (*Semantic Accuracy-SA*) che dà indicazione di quanto il significato del messaggio trascritto sia affine a quello originale. In un servizio giornalistico, anche se si registrasse un valore molto alto di *WA*, un solo errore di trascrizione di una singola parola potrebbe risultare critico e compromettere il significato dell'intero messaggio, portando la Semantic Accuracy a un valore bassissimo;
- la specializzazione del lessico e dei dizionari diventa in questo contesto una attività imprescindibile per poter disambiguare velocemente le parole e utilizzare il lessico specifico per il dominio della notizia.

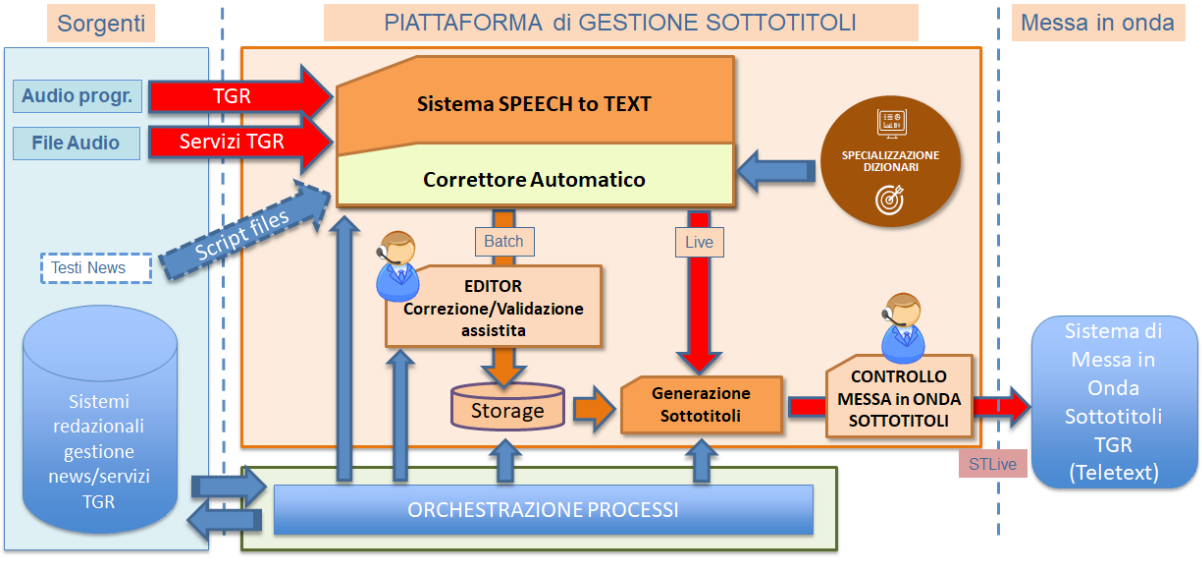
Lo scenario è molto complesso, dovendo anche prevedere l'integrazione della soluzione applicativa in un contesto consolidato di sorgenti e flussi editoriali che rappresentano la struttura portante dell'edizione di un telegiornale.

Tutti questi elementi devono poter essere orchestrati in maniera scrupolosa per poter immaginare una soluzione completa e automatica, orientata alla generazione di un flusso continuo di sottotitoli finalizzato alla corretta comunicazione del messaggio verso l'utente televisivo finale.

Piattaforma di sottotitolazione automatica
Analisi e Progettazione

Il cuore della *Piattaforma di sottotitolazione automatica per le news*, rappresentata in Fig. 1, è il *modulo Speech to Text* che opera la trascrizione dell'audio in testo, affiancato da un modulo che formatta opportunamente il trascritto generando i sottotitoli, corredati di tutte le informazioni di sincronizzazione con il segnale audio/video di riferimento.

Fig. 1 – Piattaforma di sottotitolazione automatica per le news.



Per affrontare la fase di analisi e modellazione di una soluzione per il sistema di sottotitolazione automatica completamente integrato con l'infrastruttura e con le differenti sorgenti necessarie per la messa in onda delle news, è stato necessario affrontare le attività di analisi e progettazione con una metodologia bottom-up.

La metodologia applicata ha consentito di partire da una soluzione che si potesse inserire in un flusso di produzione già attivo, passando poi a soluzioni sempre più complesse per integrare le differenti componenti dell'infrastruttura esistente, incrementandone e modellandone le complessità.

L'attività di analisi e progettazione, pertanto, si è articolata in tre diverse fasi:

- ***Fase 1***: approccio *FULL LIVE*
- ***Fase 2***: approccio *BATCH*
- ***Fase 3***: approccio integrato *TEXT*

Queste tre fasi individuano tre diverse modalità di funzionamento ed erogazione dei sottotitoli della piattaforma e implementano logiche differenti per interfacciarsi e gestire correttamente i sistemi sorgente e i flussi di produzione e di pubblicazione delle news presenti in **Rai**.

Fase 1: approccio FULL LIVE

La *Piattaforma di sottotitolazione automatica per le news* nella modalità detta *FULL LIVE* fornisce in real time il flusso di sottotitoli in uscita.

Questa modalità permette di introdurre la sottotitolazione indipendentemente dall'infrastruttura di gestione dei flussi delle informazioni esistente, dalle sorgenti, con le loro tipicità, e dalle procedure editoriali che governano la messa in onda di un Telegiornale Regionale.

In questa specifica modalità il parlato viene trascritto in tempo reale, con una tolleranza di poche centinaia di millisecondi tra il momento in cui una parola viene pronunciata ed il momento in cui è disponibile la sua trascrizione. La sequenza di più parole così trascritte forma il *Sottotitolo* che, formattato secondo specifiche regole che ne garantiscono la leggibilità (numero di parole per riga, tempo di visualizzazione, dimensione e posizione carattere), è inviato per la visualizzazione tramite il *Servizio Televideo* alla pagina 777.

Questo particolare approccio risulta snello nell'implementazione e di veloce applicabilità in architetture complesse ed eterogenee. Il motore di trascrizione e sottotitolazione non prevede di modificare i flussi produttivi audio e video, ma si può affiancare ad architetture esistenti collocandosi parallelamente alle pipeline di messa in onda del segnale televisivo.

Al sistema, come segnale d'ingresso, è sufficiente fornire un segnale audio normalizzato, tipicamente il segnale audio proveniente dallo studio del TGR.

Il segnale di uscita del sistema di trascrizione e sottotitolazione è rappresentato da un flusso di dati conforme al protocollo *STLive Formatted Input Protocol*, definito da **Rai** per l'erogazione dei servizi di sottotitolazione. STLive è il protocollo con il quale un *Client STLive* invia i sottotitoli, già formattati secondo lo standard teletext, al *Server STLive* che opera come *gestore dei servizi di sottotitolazione*.

Come descritto, l'approccio *FULL LIVE* ha la sua più importante potenzialità nel poter essere inserito in un flusso editoriale di gestione delle news senza introdurre specifiche variazioni e adattamenti.

In questo caso l'analisi delle sorgenti e del workflow di lavoro redazionale non impatta sul risultato finale della sottotitolazione e pertanto non è possibile migliorare le prestazioni del sistema se non migliorando la specializzazione del modulo di trascrizione, inserendo nel dizionario ulteriori informazioni legate, ad esempio, alla geografia dei luoghi, alla toponomastica e alla odonomastica.

Nell'ottica di erogare un servizio di sottotitolazione automatica, è contemplata la possibilità da parte di un operatore di correggere il testo durante la fase di trascrizione e in corrispondenza della messa in onda.

Fase 2: approccio BATCH

La *Piattaforma di sottotitolazione automatica per le news* gestisce una modalità di lavoro complementare, denominata *BATCH*, che prevede il processamento dei contenuti audio relativi ai *servizi chiusi* della redazione TGR, ovvero tutti i servizi già montati che hanno ottenuto la validazione editoriale e che vengono resi disponibili al sistema prima della messa in onda del TGR.

Il sistema in tale modalità esegue *offline* la trascrizione dal parlato in testo e genera il flusso dei sottotitoli che sarà opportunamente richiamato e visualizzato tramite il *Servizio Televideo* alla pagina 777, in corrispondenza della messa in onda dello specifico contenuto.

La gestione della modalità *BATCH*, che si alterna con la modalità *LIVE*, consente di incrementare la qualità del servizio migliorando l'accuratezza della trascrizione del prodotto finale in termini di *word accuracy*, misurata come percentuale di parole trascritte correttamente rispetto al numero totale di parole trascritte dal sistema automatico. In particolare, la trascrizione in modalità *BATCH*, operando su un file audio interamente disponibile al sistema, consente di fornire su un servizio TGR un'accuratezza mediamente superiore del 2% rispetto a quella generata in modalità *LIVE*. Inoltre il sistema è in grado di gestire la punteggiatura fornendo così dei sottotitoli più leggibili e quindi di più facile comprensione.

In aggiunta, il fatto di trascrivere e sottotitolare servizi chiusi prima della messa in onda consente, anche, di prevedere *operazioni di revisione*, ovvero modifica, correzione e validazione, dei sottotitoli generati automaticamente da parte di un operatore, fornendo così una sottotitolazione con un valore di accuratezza pari al 100%.

Nell'ottica di erogare un servizio di sottotitolazione automatica affidabile, è quindi previsto l'impiego di un'applicazione web per effettuare offline questo tipo di operazioni di modifica, correzione e validazione.

Queste funzionalità sono fornite dal modulo *EDITOR*, collocato a valle del sistema *Speech to Text*.

Il modulo EDITOR

Il modulo *EDITOR* è completamente integrato all'interno della piattaforma di generazione automatica dei sottotitoli ed è gestito dal modulo *Orchestrazione Processi* che, durante la messa in onda dell'edizione TGR, accede alle trascrizioni corrette e validate fornendo, così, un servizio di sottotitolazione con un elevatissimo grado di qualità.

Le funzionalità fornite dal modulo *EDITOR* permettono ad un operatore di:

- ascoltare l'audio del servizio giornalistico ed operare con semplicità funzioni di start-stop-repeat (es: tramite pedaliera);
- effettuare delle revisioni del testo generato dal modulo *Speech To Text* e inserire la punteggiatura. L'interfaccia grafica consente una correzione agevole delle parole errate e fornisce dei suggerimenti per una scelta rapida di quelle che, con maggiore probabilità, potrebbero essere corrette;
- salvare i termini maggiormente ricorrenti o specifici di un particolare contesto in un'area *suggerimenti* (es: nomi di personaggi o località).

Fase 3: approccio integrato TEXT

La *Piattaforma di sottotitolazione automatica per le news* contempla una terza modalità operativa, denominata *TEXT*, che non prevede l'attivazione del sistema di trascrizione ma la gestione dei contenuti testuali attinti dai sistemi redazionali TGR, i quali vengono formattati opportunamente e trasformati in sottotitoli.

Tale modalità, che viene attivata tipicamente in corrispondenza dei titoli e dei lanci dei servizi, consente di garantire un'accuratezza della sottotitolazione pari al 100%, di valorizzare fonti dati già disponibili e validate e di mantenere l'integrità del messaggio che il giornalista, nella scrittura del testo, intende comunicare.

Piattaforma di sottotitolazione automatica
Funzionalità

Le principali funzionalità fornite dalla *Piattaforma di sottotitolazione automatica per le news*, rappresentata nella già citata Fig. 1, possono essere sintetizzate in:

- *Acquisizione*. La piattaforma acquisisce i contenuti testo/audio dalle diverse tipologie di sorgenti tramite specifiche interfacce opportunamente definite;
- *Trascrizione*. Il *sistema Speech to Text* è il modulo della piattaforma responsabile della trascrizione automatica del segnale audio. Esso trasforma il parlato in un testo scritto corredato di tutte le informazioni di sincronizzazione con l'audio originale;
- *Correzione Assistita flusso Live*. Per la modalità *Live* è previsto un tool grafico usato da un operatore per correggere eventuali errori di trascrizione e inserire la punteggiatura prima della messa in onda;
- *Correzione mediante editor dei testi*. Per la modalità *Batch* è previsto un editor per correggere e validare il testo generato dalla trascrizione automatica, garantendo così una percentuale di accuratezza pari al 100%;
- *Generazione Sottotitoli*. Formattazione dei sottotitoli, sincronizzazione e interazione con i sistemi di messa in onda.

Le ulteriori funzionalità di *Specializzazione del modello del linguaggio* e di *Arricchimento dei Dizionari* contribuiscono, nel tempo, a migliorare le prestazioni del sistema in termini di accuratezza delle trascrizioni.

Il sistema di trascrizione è dotato a priori di un modello del linguaggio *generalista*, ovvero non specifico di alcun ambito.

Per il progetto *Sottotitolazione Automatica TGR* il modello del linguaggio viene costantemente arricchito, rispetto a specifici argomenti del contesto desiderato (es. politica, sport, meteo, traffico, ecc.) e dello specifico lessico regionale.

Specializzazione del modello del linguaggio

Nel corso della sperimentazione, al fine di migliorare il livello di accuratezza nel riconoscimento del parlato, è prevista la realizzazione di specifici *modelli di linguaggio*, che vengono arricchiti e declinati per, ad esempio, ciascuna regione e/o rispetto a specifici argomenti trattati nel contesto desiderato (es. politica, sport, meteo, traffico, ecc.). È possibile, pertanto, selezionare il modello del linguaggio specializzato in funzione del dominio da trascrivere.

Arricchimento dei Dizionari

Il sistema di trascrizione automatica è dotato di moduli specifici che consentono di aggiornare, migliorare ed ottimizzare il riconoscimento automatico di terminologie speciali ed attualizzate relative ad un ambito specifico:

- *Modulo attualizzazione automatico*: attualizza ed arricchisce automaticamente il dizionario del modello raccogliendo in autonomia nuovi termini da fonti aperte (es. feeds rss) permettendo di riconoscere sempre i nuovi termini;
- *Modulo arricchimento assistito*: permette, attraverso un'avanzata interfaccia utente web-based, l'inserimento e la modifica dei termini, quali, a titolo di esempio:
  - *lista nomi propri*: i nomi e cognomi di persona o di località specifici appartenenti ai contesti locali.
  - *lista parole*: tutte le parole peculiari di uno specifico contesto.

L'Orchestratore Processi

I vantaggi derivanti dalla gestione delle trascrizioni operate nelle modalità *BATCH* e *TEXT* hanno indotto la decisione di sviluppare e introdurre per il progetto *Sottotitolazione Automatica TGR* un orchestratore che, durante la messa in onda dell'edizione TGR, opera una commutazione del sistema di *Trascrizione e Sottotitolazione Automatica* tra le modalità di trascrizione e generazione di sottotitoli *Live*, *Batch* e *Text*, in base all'identificazione della tipologia di contenuto in onda.

L'*Orchestratore Processi* è il modulo software responsabile della gestione dei vari moduli interni alla piattaforma di generazione automatica dei sottotitoli, di tutti i semilavorati generati dai singoli moduli e garantisce il corretto interfacciamento della piattaforma con le differenti tipologie di sistemi sorgente e di sistemi di messa in onda.

Piattaforma di sottotitolazione automatica
Sperimentazione e valutazione del servizio

La sperimentazione prevede un'ultima fase di valutazione della piattaforma per l'erogazione del servizio e del prodotto/servizio finale.

A corredo dell'analisi tecnica delle prestazioni dei singoli moduli della piattaforma di generazione dei sottotitoli, è importante valutare la piattaforma stessa nella sua interezza in termini di usabilità, velocità, impiego di risorse richieste e affidabilità. Tale valutazione sarà effettuata dagli operatori coinvolti nei processi di revisione delle trascrizioni *batch* e di presidio/intervento *live* durante la messa in onda delle diverse edizioni dei TGR.

La valutazione del prodotto finale si articola, invece, in due modalità di analisi:

- una *valutazione tecnico-oggettiva* delle prestazioni del sistema di trascrizione e generazione dei sottotitoli automatico sulla base di analisi sia di tipo *word accuracy*, ovvero basate sulla percentuale di parole trascritte correttamente sul totale delle parole trascritte, sia di tipo *semantic accuracy*, ovvero basate sulla percentuale di concetti comprensibili e corretti sul totale del numero di concetti espressi. È doveroso precisare che, a differenza della word accuracy, la semantic accuracy, basandosi sull'interpretazione semantica dei concetti, intesi come sequenze di parole che veicolano un messaggio di senso compiuto, risente di una minima componente di interpretazione soggettiva;
- una *valutazione soggettiva* che prevede l'attivazione di focus group di potenziali utilizzatori del servizio a cui verranno presentate alcune edizioni dei TGR sottotitolate mediante l'uso del sistema di sottotitolazione automatico e che forniranno un riscontro importante sull'indice di comprensibilità, gradevolezza, leggibilità dei relativi sottotitoli. Una parte consistente degli utenti selezionati saranno le persone alle quali è fortemente indirizzato il servizio di sottotitolazione, in particolare persone con disabilità uditive. È fondamentale pertanto il coinvolgimento delle relative Associazioni e una comunicazione corretta al fine di condividere ambizioni e finalità del progetto, mettendo anche in luce, con la massima trasparenza, i limiti e gli eventuali errori sui sottotitoli che un sistema automatico, anche se in minima misura, potrebbe commettere.

A titolo di esempio, in Fig. 2 pagina seguente è riportato un estratto dell'analisi di una tipica edizione di un TGR. Per ogni tipologia di contenuto viene messa in evidenza la modalità di funzionamento della piattaforma e la relativa *word accuracy (WA)*. Un'accuratezza del 100% si registra in corrispondenza dei sottotitoli generati in modalità *TEXT*, in cui non interviene il sistema di trascrizione. Nelle modalità *LIVE* e *BATCH* le percentuali di accuratezza sono comunque piuttosto elevate, mediamente intorno al 90-95%. Introducendo, per la correzione dei servizi TGR chiusi trascritti in modalità *BATCH*, l'operazione di revisione e correzione del trascritto si raggiunge, come indicato nella quarta colonna, un'accuratezza del 100%.

La valutazione della *Piattaforma di sottotitolazione automatica per le news* fornirà gli elementi per stimare costi e benefici di una soluzione che, grazie all'introduzione di algoritmi di intelligenza artificiale, può innovare e ottimizzare gli attuali processi di produzione dei sottotitoli.

Conclusioni e future evoluzioni

Alla luce di quanto fin qui esposto è lecito pensare che i sistemi di *Machine Learning* per la T*rascrizione Automatica* siano solo all'albore del loro sviluppo, essendo essi una realtà estremamente complessa e

Scaletta TGR	Trascrizione	Word Accuracy	WA con BATCH Corretti
1 – Titoli	TEXT	100 %	100 %
2 – Giornalista	TEXT	100 %	100 %
3 – Servizio	BATCH	94,34 %	100 %
4 – Giornalista	TEXT	100 %	100 %
5 – Collegamento	LIVE	94,34 %	94,34 %
6 – Servizio	BATCH	89,80 %	100 %
7 – Giornalista	TEXT	100 %	100 %
8 – Servizio	BATCH	95,19 %	100 %
9 – Servizio	BATCH	79,66 %	100 %
10 – Giornalista	TEXT	100 %	100 %
11 – Servizio	BATCH	96.97 %	100 %
12 – Giornalista	TEXT	100 %	100 %
13 – Servizio	BATCH	95,36 %	100 %
14 – Giornalista	TEXT	100 %	100 %
15 – Collegamento	LIVE	85,63 %	85,63 %
16 - Saluti	LIVE	99,10 %	99,10 %
		95,65 %	98,69 %

Fig. 2 – Analisi di un'edizione TGR

multiforme in cui coesistono aspetti contradittori, ma anche sviluppi interessanti per innovativi scenari di applicazione.

L'utilizzo sempre più spinto di queste tecnologie indica che tali sistemi, nonostante possibili errori e imprecisioni, sono in grado di fornire una trascrizione di elevata qualità, tale da poter ipotizzare di incentrare sempre più servizi per il cittadino su tali soluzioni.

Nel contesto giornalistico la specializzazione dei modelli del linguaggio riveste un importante tassello per il miglioramento della trascrizione del parlato ed è innegabile che, data l'importanza delle tematiche affrontate e della diffusione del messaggio veicolato, diventano fondamentali anche le attività di post-editing sul testo già trascritto per raggiungere percentuali elevatissime di accuratezza.

Le logiche di gestione del workflow, in relazione alle tipicità delle sorgenti coinvolte nel processo di creazione di una edizione di news, incrementano ulteriormente la qualità dei risultati prodotti dai sistemi automatici in termine di efficacia e di affidabilità del sistema.

Nei prossimi anni i sistemi basati su *Intelligenza Artificiale* saranno un elemento nodale dei processi produttivi coinvolti nella gestione dell'informazione e nell'ottimizzazione delle risorse aziendali.

Intelligenza Artificiale e Archivi Radiotelevisivi

Opportunità e sfide

Alberto **Messina**
Rai - Centro Ricerche, Innovazione Tecnologica e Sperimentazione

Questo articolo introduce le problematiche, le opportunità e le sfide derivanti dall'applicazione delle tecnologie di *Intelligenza Artificiale* (IA) nell'ambito degli *Archivi Radiotelevisivi* e discute il ruolo che gli archivi possono avere come sorgenti di preziosi dati per lo sviluppo di tali tecnologie.

Gli archivi radiotelevisivi sono stati da sempre un ambito di riferimento per la sperimentazione di tecnologie dell'Intelligenza Artificiale (IA), grazie alla varietà, dimensione e ricchezza dei beni in essi conservati. Lo sviluppo esponenziale delle tecnologie moderne dell'IA li pone più che mai al centro dell'attenzione della comunità tecnico-scientifica.

Questo articolo illustra in maniera estremamente sintetica le opportunità derivanti dall'applicazione delle moderne tecniche dell'intelligenza artificiale al mondo degli archivi radiotelevisivi.

Si illustrano due famiglie fondamentali di utilizzo: come supporto ai processi dell'archivio e ai nuovi i processi di pubblicazione e come supporto allo sfruttamento dei dati di archivio per l'addestramento di modelli di IA. Entrambi i casi sono caratterizzati dalla necessità di formalizzare e implementare processi di adeguamento e integrazione dei dati, che possono beneficiare essi stessi di tecnologie di IA. A tal proposito si accenna sinteticamente ad alcuni possibili approcci formali, rimandando il lettore a riferimenti che ne riportano una trattazione più completa.

Cosa sono gli Archivi

Per poter correttamente contestualizzare e discutere le applicazioni dell'*Intelligenza Artificiale* negli archivi bisogna innanzitutto richiamarne una definizione. In generale un *archivio* è un *insieme di risorse e processi deputati alla conservazione e all'accesso a lungo termine di determinati beni d'interesse per una certa comunità di utilizzatori*. Ciò che differenzia un archivio da un magazzino è proprio quindi la connotazione di *lungo termine*. Mentre infatti le informazioni necessarie a descrivere i beni di un magazzino, e quindi a supportarne l'accesso, possono essere considerate conosciute e conoscibili direttamente e completamente dalla comunità di utilizzatori, anche attraverso supporti, convenzioni e relazioni implicite, ciò diventa chiaramente non applicabile quando il tempo intercorrente tra la creazione (es. registrazione, ricezione, ecc.) del bene da archiviare e il suo utilizzo è, in generale, più ampio e superiore alle capacità o disponibilità di memoria individuale o collettiva della comunità stessa. La comunità tecnico-scientifica ha già da tempo formalizzato alcuni di questi aspetti nel modello **OAIS** (*Open Archival Information System*) [1][2].

201_02.2_2.00

Non essendo chiaramente questo articolo una disquisizione critica sugli archivi, ci limitiamo qui a richiamare il concetto di *comunità designata* (*Designated Community*) dal modello OAIS definita come:

> *"Un gruppo identificato di potenziali consumatori che dovrebbero essere in grado di comprendere un particolare insieme di informazioni. La* Comunità Designata *può essere composta da più comunità di utenti. Una comunità designata è definita dall'Archivio e da questo la definizione può cambiare nel tempo."*

Questa definizione introduce delle criticità esplicite e alcune latenti tipiche del contesto intrinsecamente diacronico dell'archivio:

- la comprensione delle informazioni di accesso agli archivi è solo potenziale poiché la conoscenza a priori della capacità interpretativa della comunità non può che essere parziale;
- la designazione è a carico dell'archivio, che però ha una nozione limitata delle potenziali comunità interessate ai propri beni, soprattutto quelle future.

Questo implica che i processi di generazione e conservazione delle informazioni di accesso ai beni di archivio (*metadati*) debbano essere continuativi e riguardare non solo la ricchezza e la completezza delle informazioni in sé ma anche i sottostanti modelli semantici, seguendo precise pratiche di cura del dato [3].

Queste considerazioni svelano, secondo noi, il ruolo chiave dell'IA per gli archivi, ruolo che può declinarsi secondo tre direttive principali:

1. utilizzo dell'IA per supportare e rendere maggiormente convenienti i processi attuali;
2. estensione dei processi e dei modelli informativi con dati generati dall'IA;
3. introduzione di funzionalità e processi completamente nuovi.

IA nel contesto Archivi Radiotelevisivi

Gli *archivi radiotelevisivi*, cioè gli archivi che conservano beni prodotti negli anni dai processi di produzione radiotelevisiva, sono un caso particolare di archivi che non fa tuttavia eccezione alle criticità poco prima evidenziate.

Se all'origine gli archivi radiotelevisivi erano intesi soprattutto come servizi utili alla produzione interna di un broadcaster o come memoria collettiva di una nazione[Nota 1], l'evoluzione del mercato dei media a partire dagli anni 2010 ha visto l'archivio acquisire crescente rilevanza come fonte per la fornitura di nuovi servizi[Nota 2], come ad esempio l'*Over-The-Top* (OTT) ed i *social media* [4][5][6][7]. In termini di modello OAIS, questa evoluzione equivale ad un cambio, anche piuttosto radicale, della *comunità designata* dell'archivio. Idealmente, a questo cambio dovrebbe corrispondere una altrettanto radicale revisione dei modelli descrittivi dei beni di archivio, per assicurare corrispondenza con i criteri di filtro e le aspettative dei *nuovi utenti*. Si veda l'esempio riportato in Fig. 1 di pagina seguente, che riporta la *descrizione originale conservata in archivio* e quella *pubblicata sul servizio RaiPlay* di un medesimo oggetto multimediale. La seconda descrizione è orientata ad una comunità di utenti generici non professionali, a differenza della prima, che, ad esempio, distingue la descrizione della scena video da quella sonora. Ritrovare l'oggetto con la *mentalità* degli utenti generici può essere quindi difficile se esso è annotato con la mentalità dell'archivista.

Nota 1 - Si veda il caso dell'INA francese

Nota 2 - Anche grazie alla loro, in certi casi, notevole dimensione. L'archivio Rai conta ad esempio più di 150 milioni di documenti censiti (tra foto, documenti cartacei e programmi radiofonici e televisivi, di cui 5 milioni in formato digitale)

Contenuto Audio	Angela su storia, stile architettonico e caratteristiche della Basilica di San Vitale, costruito dopo la morte del re degli Ostrogoti Teodorico; descrizione di una vasca presente all'interno della Basilica in cui sono visibili i diversi strati di pavimentazioni costruiti nei secoli per far fronte al fenomeno della subsidenza, ossia lo sprofondamento del terreno e la risalita di acqua sottostante; caratteristiche e simbologia dei mosaici presenti all'interno della basilica con il corteo dell'Imperatore Giustiniano e della moglie Teodora. Melandri su beni di interesse culturale e storico presenti a Ravenna, sua città natale, su tradizione del mosaico a Ravenna e su suoi personali ricordi legati al mosaico
Contenuto Video	Angela cammina all'esterno e all'interno della Basilica di San Vitale. Vedute panoramiche e aeree, riprese con droni, interni e dettagli architettonici, mosaici della Basilica di San Vitale, dettagli vasca presente all'interno della Basilica in cui sono visibili i diversi strati di pavimentazioni costruiti nei secoli. Animazioni grafiche ricostruzioni all'interno della Basilica di San Vitale. Melandri parla in studio con animazione grafica sul fondo

Fig. 1 – Esempio di descrizioni in diversi domini applicativi del medesimo oggetto. Sopra la *descrizione d'archivio*, sotto quella del *servizio RaiPlay*.

Questo avviene perché nel dominio dell'archivio, i documentatori annotano il contenuto secondo un modello di descrizione ben definito, che contiene elementi relativi alle decisioni prese dagli utenti degli archivi (es. se il brano archiviato è utile o meno per una nuova produzione, o se si hanno i diritti di sfruttamento o se la risorsa multimediale è di qualità sufficiente). Nel dominio di pubblicazione OTT gli editori annotano il contenuto secondo un diverso modello di descrizione, che contiene elementi che aiutano gli utenti finali a prendere decisioni sul loro interesse per un programma. Un approccio che preveda sezioni documentative separate per area di business non è tuttavia praticabile, a causa dei costi dei processi di documentazione [Nota 3].

In termini pratici ciò si traduce nella necessità di un'integrazione/adattamento della documentazione dei beni d'archivio, che ha chiaramente dei costi non trascurabili. L'utilizzo dell'IA può quindi, di principio, sopperire alla difficoltà di reperire le risorse adeguate e supportare l'arricchimento della documentazione d'archivio per includere nuove comunità di utilizzatori, nell'ipotesi, tutta da verificare, che i processi di arricchimento basati su IA siano più convenienti dei processi manuali [8][9] . Esempi pratici di tali arricchimenti comprendono il rilevamento e l'identificazione di personaggi storici, la traduzione del parlato in testo, l'annotazione didascalica delle scene, l'identificazione di punti di riferimento o di interesse (ad esempio, monumenti), il rilevamento dell'utilizzo di opere d'autore [10] [11]. Una criticità rilevante nel caso degli archivi è costituita dalla *diacronia* degli stessi, che implica che un medesimo concetto o oggetto possa presentarsi con caratteristiche tecniche o esteriori molto variabili (Fig. 2 risp. (a) e (b)). Al di là dei casi in cui queste tecniche sono di supporto a processi di documentazione manuale, i dati provenienti da questi processi consentono anche di rendere i beni di archivio ricercabili e quindi riutilizzabili in contesti più ampi rispetto a quelli originariamente identificati all'atto della prima documentazione.

Nota 3 - Il costo medio per la documentazione di un'ora di programma televisivo è di circa 30€.

Fig. 2 – Esempi di diacronia dell'archivio Rai.

(a)

(b)

Il recente sviluppo di tecnologie di intelligenza artificiale basate su *reti profonde* (*DNN – Deep Neural Networks*) ha incrementato notevolmente le performance di molte di queste applicazioni, e allo stesso tempo introdotto nuovi approcci, rendendo oggi possibile l'adozione di tali tecniche in molti contesti industriali concreti, tra cui gli archivi radiotelevisivi. La Tabella 1 riporta alcuni esempi di tali applicazioni nell'area relativa all'annotazione degli oggetti di archivio.

Un altro campo applicativo promettente dell'IA negli archivi è quello dell'utilizzo di tecniche finalizzate al rintracciamento di relazioni non esplicite o latenti fra i beni anche appartenenti ad archivi eterogenei. Da un punto di vista qualitativo si può apprezzare come questo tipo di analisi sia più complesso rispetto alla semplice documentazione di singoli beni d'archivio poiché richiede di considerare insiemi di beni nel loro complesso, le loro caratteristiche, e le mutue relazioni che tra questi possono esistere. Le tecnologie di IA utili in questo caso sono quindi quelle che simulano le capacità umane di confronto, aggregazione, astrazione, similarità tra oggetti. Tra i molti approcci sviluppati, citiamo Il lavoro pionieristico di **Rai** nel campo dell'aggregazione multimodale automatica degli archivi giornalistici [12]. Tra i lavori più recenti e interessanti in questo campo spicca anche il tentativo di **BBC** di costruire un palinsesto lineare attingendo al proprio archivio attraverso un agente intelligente che sfrutta le descrizioni estratte automaticamente [13]. Un approccio alternativo è quello di utilizzare tecnologie di ragionamento automatico per esplorare le relazioni semantiche tra i beni di archivio, rappresentate attraverso linguaggi del *Semantic Web* come *OWL* [14], sia esplicite che dedotte dall'applicazione di tecniche di inferenza sui dati [15].

Tabella 1 – Esempi di applicazioni di annotazione dell'archivio abilitate dall'IA.

Segmentazione editoriale	*Suddividere il contenuto in scene dal punto di vista editoriale (ad esempio distinguendo tra introduzione, dibattito e conclusioni in un talk show)*
Sommario testuale	*Creare un sommario testuale a partire da un oggetto d'archivio*
Descrizione didascalica	*Descrivere un contenuto attraverso brevi frasi che descrivono la scena dal punto di vista visuale o sonoro*
Identificazione e caratterizzazione dei personaggi	*Identificare le persone presenti nella scena (in video o in audio) e darne una caratterizzazione anagrafica (età, sesso), emozionale o di contesto (ruolo)*
Riconoscimento elementi di interesse	*Identificare elementi di interesse (opere d'arte, brani musicali, monumenti, oggetti specifici) ripresi in video o in audio*
Indicizzazione e ricerca per contenuti	*Ricercare attraverso esempi visuali o sonori per identificare copie o trasformazioni presenti in archivio*
Classificazione concettuale	*Classificare il contenuto audio o video secondo tassonomie concettuali (es. genere, argomento) o contenutistiche (es. oggetti, azioni)*
Caratterizzazione di qualità	*Classificare i contenuti audio o video in categorie di qualità*
Rilevamento difetti	*Identificare la presenza di difetti di ripresa, artefatti di codifica o di degradazione del supporto di memorizzazione*

Infine, ma non da ultimo, citiamo le applicazioni orientate a supportare i processi di gestione della qualità degli oggetti d'archivio e al loro riutilizzo nei processi di produzione. Il ciclo di vita di un oggetto di archivio è tipicamente caratterizzato da varie fasi, la più importante delle quali è senz'altro la *digitalizzazione*, qualora il supporto di memorizzazione originale sia analogico. Durante questa fase si introducono i tipici difetti intrinseci dei sistemi di codifica [16]. La fase analogica può essere inoltre caratterizzata da fenomeni di degradazione del supporto, che introducono artefatti e difetti che si riflettono nell'oggetto digitalizzato [17]. Il riutilizzo in contesti di produzione moderni, potrebbe inoltre richiedere adattamenti di formato e di risoluzione.

Assieme alle tecniche di annotazione, utili ai processi di ricerca e reperimento di oggetti d'archivio, possiamo quindi annoverare tecniche orientate al miglioramento dei contenuti, utili in fase di riutilizzo degli oggetti stessi. La Tabella 2 riporta alcuni esempi di tali applicazioni.

L'ARCHIVIO COME DATASET: UN PROBLEMA FORMALE

Assieme alle tecniche di IA che supportano direttamente i processi di gestione e riutilizzo in produzione degli archivi, si assiste di recente allo sviluppo di approcci teorici e pratici orientati all'utilizzo degli oggetti d'archivio come fonte dati per l'addestramento [Nota 4] di tecnologie di IA [18][19]. Il problema intrinseco a questo genere di applicazioni è la *rappresentatività* delle informazioni associate agli oggetti d'archivio [20]. Al fine della seguente discussione, possiamo definire un *dataset* come un insieme D di affermazioni S che associano oggetti multimediali, o loro parti spaziotemporali, $m \in M$ a concetti $c \in C$. Tali affermazioni sono enunciate da una serie di osservatori $u \in U$, espresse in qualche forma dichiarativa la cui interpretazione $\mathcal{I}(S)$ risulta vera [Nota 5] in qualche dominio del discorso $\mathcal{D}$. Sinteticamente:

$$D = \{S\}$$
$$S = [(m, c), u]$$
$$m \in M, c \in C, u \in U$$
$$\mathcal{I}((m, c)) \in \mathrm{True}(\mathcal{D})$$

Tabella 2 – Esempi di tecniche di miglioramento della qualità abilitate dall'IA.

Video super resolution	*Incrementare la risoluzione di un video (ad esempio da HD a UHD)*
Restauro digitale	*Rimuovere/mitigare l'effetto degli artefatti di codifica*
Denoising	*Ridurre il livello di rumore presente nelle immagini o nel suono*
Conservazione della grana pellicola	*Conservare l'effetto tipico della grana pellicola*
Colorazione	*Colorare contenuti originariamente prodotti in bianco e nero, oppure ripristinare il colore di contenuti sbiaditi dal tempo*

Nota 4 - Una fonte dati per l'addestramento di una tecnologia di intelligenza artificiale è di solito denotata con il termine *dataset*

Nota 5 - L'associazione di un valore di verità alle dichiarazioni che compongono un dataset può essere intesa come un'operazione di verifica da parte di altri osservatori

Le dichiarazioni che possono essere generate dall'osservatore di un oggetto sono quindi potenzialmente molte, ma solo quelle vere possono far parte di un dataset. Inoltre, dataset diversi possono comprendere solo un sottoinsieme di tutte le possibili dichiarazioni. La Fig. 3 illustra un esempio pratico di tale formalizzazione, nel quale due osservatori *o1* e *o2* generano due diversi insiemi di affermazioni a proposito di un'immagine e viene definito un dataset che ne contiene una particolare combinazione [Nota 6].

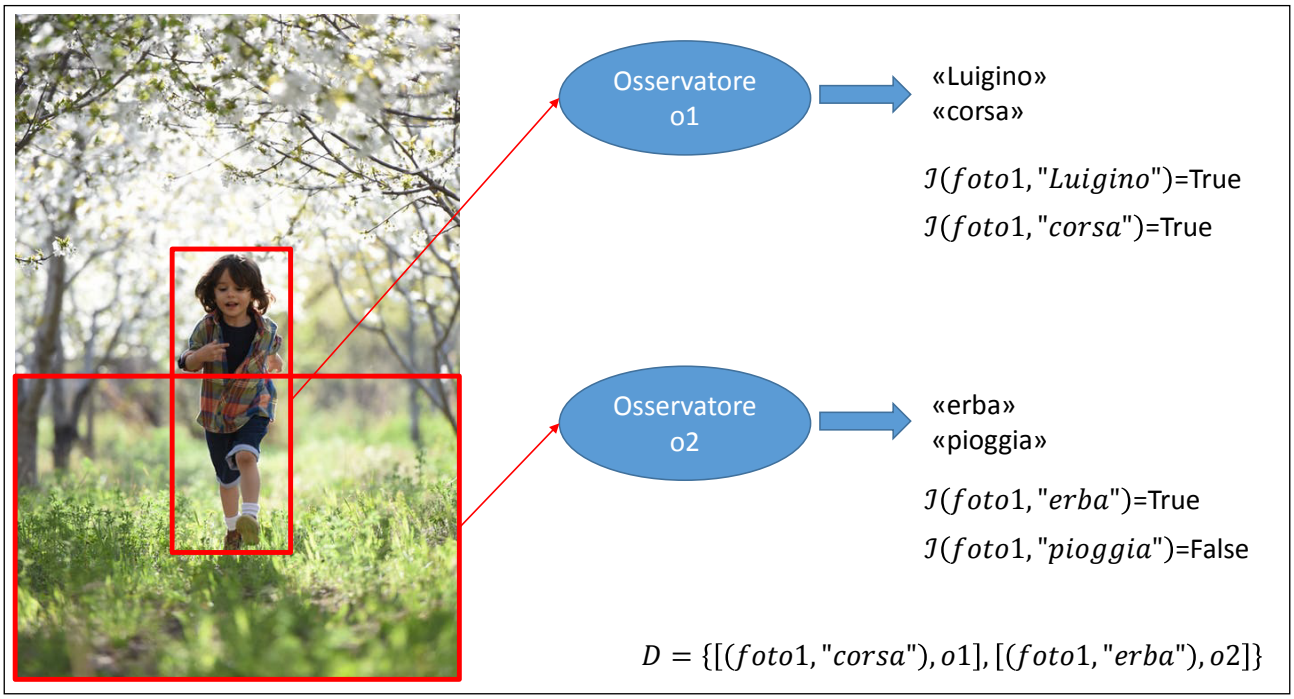
Un dataset utilizzabile per l'addestramento di una tecnologia di intelligenza artificiale dovrebbe essere caratterizzato da almeno i seguenti aspetti:

- sufficiente *copertura* del dominio applicativo di riferimento;
- sufficiente *rappresentatività* degli oggetti d'archivio rispetto al medesimo dominio;
- sufficiente *significatività/qualità* delle associazioni tra concetti del dominio di riferimento e oggetti di archivio (*annotazioni*).

Se ad esempio volessimo costruire un classificatore automatico di immagini o sequenze video di disastri naturali, basato sulle più moderne architetture neurali disponibili dallo stato dell'arte, riferendoci ad una *tassonomia standard* [Nota 7] dovremmo verificare che:

1. l'archivio disponga di un sufficiente numero di esempi di ciascuna delle categorie presenti nella tassonomia;
2. le annotazioni siano *vere*, cioè in linea con le definizioni concettuali della medesima tassonomia.

Fig. 3 – Esempio della caratterizzazione formale di un dataset.



Nota 6 - Si noti la ritenzione dell'informazione relativa all'osservatore nel dataset

Nota 7 - Ad esempio https://iptc.org/standards/media-topics/ (ultimo accesso 10/12/2020)

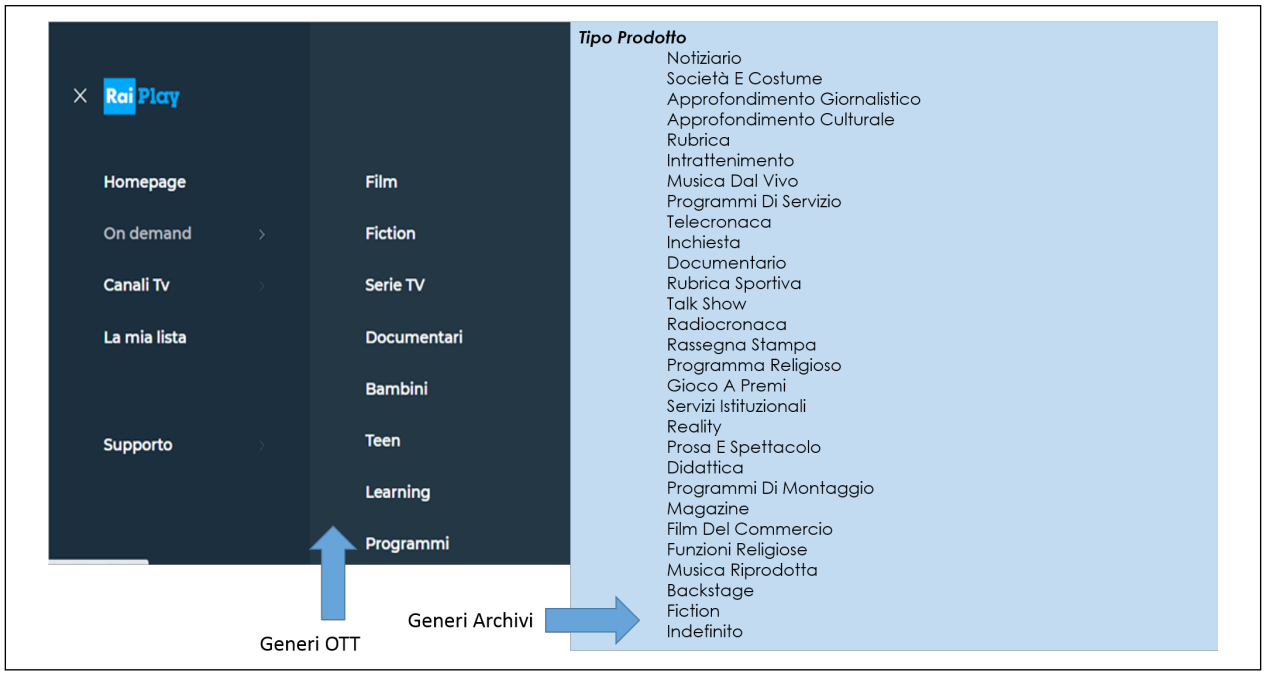
Ad esempio, supponiamo di voler classificare per genere elementi multimediali pubblicati su un servizio OTT, e per fare questo vogliamo riutilizzare un set di dati dagli archivi per addestrare un classificatore basato su IA a riconoscere i generi. Si supponga che i due domini (archivi e OTT) abbiano tassonomie di riferimento diverse e criteri diversi per associare un determinato elemento multimediale a un termine di tassonomia (si veda la Fig. 4). In questo caso il problema dell'adattamento è composto da tre sotto-problemi separati:

1. *mappatura* della tassonomia di genere tra sorgente (archivi) e destinazione (OTT);
2. *integrazione* del set di dati con nuovi elementi multimediali dal dominio OTT;
3. *classificazione* degli elementi multimediali complessivi dagli utenti del dominio di destinazione.

Questo processo di trasformazione/adattamento può essere parzialmente automatizzato ad es. considerando le mappature della tassonomia statica o utilizzando approcci più sofisticati che impiegano tecniche basate sull'intelligenza artificiale. Non c'è dubbio, tuttavia, che questo processo ha un'impronta economica non banale e come tale deve essere adeguatamente progettato.

In generale quindi, al fine di utilizzare in maniera efficace i dati di archivio come dataset, è necessario immaginare, progettare e implementare processi di adattamento/trasformazione dei dati che permettano di sopperire ai difetti evidenziati[Nota 8]. Ed è proprio analizzando questo problema che si rivela necessaria una riflessione fondamentale relativamente a come valutare, possibilmente in maniera oggettiva, l'opportunità di applicare processi di trasformazione di dati esistenti rispetto alla generazione ex-novo del dataset e a come comparare approcci di trasformazione basati essi stessi su IA con approcci di trasformazione manuale. Una teoria relativa a questa problematica è stata sviluppata in [21], lavoro al quale si rimanda per una trattazione dettagliata anche dal punto di vista formale. Nel contesto del presente articolo è sufficiente richiamare il risultato fondamentale.

Fig. 4 – Due differenti tassonomie di genere, rispettivamente RaiPlay e Archivio Rai.



Nota 8 - Le trasformazioni possono essere di natura molto diversa: aggiunta o rimozione di affermazioni, adattamento degli schemi descrittivi, estensione dell'insieme di osservatori

Esistono tre approcci per trasformare un dataset nativo in un dataset adatto ad essere usato come addestramento di strumenti di IA:

1. riscrittura completa delle annotazioni da parte di un team dedicato di osservatori esperti;
2. adattamento dei dati esistenti da parte di un team dedicato di osservatori esperti;
3. adattamento dei dati esistenti con il supporto di strumenti di IA.

La decisione su quale approccio intraprendere dipende fortemente dal caso specifico, ma in generale si può porre un problema di ottimizzazione dei costi delle trasformazioni necessarie per passare dal dataset iniziale a quello finale come segue:

$$\max\left\{\sum_{i=1}^{N}\gamma_{\epsilon}(T_{ii+1}),\sum_{i=1}^{N'}\gamma_{\epsilon}(T'_{ii+1})+\gamma_{\epsilon}{}^{CHK}(T'_{ii+1})\right\}<\gamma_{\epsilon}(T_{0j})$$

Dove N è il numero di trasformazioni manuali necessarie, N' il numero di trasformazioni supportate da IA necessarie, γ_{ε} è una funzione di costo dei processi di trasformazione dipendente dall'errore accettato ε, $\gamma_{\varepsilon}{}^{CHK}$ è una analoga funzione di costo per i processi di correzione, T_{ii+1} è il processo di trasformazione del dataset dallo stato i allo stato $i+1$, T_{0j} è il processo di trasformazione che produce il dataset ex-novo. In pratica questa formulazione indica una regola di selezione di una catena di processi di trasformazione del dataset di partenza (rispettivamente di trasformazione manuale, di trasformazione assistita o di creazione ex-novo) sulla base della stima dei relativi costi medi di processo. La difficoltà pratica di applicare questo approccio risiede tuttavia nell'effettiva misurabilità a priori dei parametri di costo dei processi coinvolti nelle diverse opzioni considerate.

Conclusioni

Questo articolo illustra in maniera estremamente sintetica le opportunità derivanti dall'applicazione delle moderne tecniche dell'intelligenza artificiale al mondo degli archivi radiotelevisivi, intesi come particolare caso di archivi definiti secondo il modello OAIS. Assieme a ciò, si sono anche presentate le problematiche fondamentali relative all'utilizzo dei dati d'archivio come dataset per l'intelligenza artificiale introducendo un possibile approccio oggettivo di soluzione.

Bibliografia

[1] CCSDS, *Reference Model for an Open Archival Information System*, Magenta Book CCSDS 650.0-M-2, 2012, https://public.ccsds.org/pubs/650x0m2.pdf (ultimo accesso 10/12/2020)

[2] ISO 14721:2012, *Space data and information transfer systems - Open archival information system (OAIS) - Reference model*, https://www.iso.org/standard/57284.html (ultimo accesso 10/12/2020)

[3] T. R. Bruce e D. Hillmann, *The Continuum of Metadata Quality: Defining, Expressing, Exploiting*, in D. Hillmann e E. Westbrooks (ed), "Metadata in Practice", ALA Editions, 2004, **ISBN**: 0838908829

[4] European Audiovisual Observatory, *The Exploitation of Film Heritage Works in the Digital Era*, European Commission (web), 2016, https://ec.europa.eu/digital-single-market/en/news/exploitation-film-heritage-works-digital-era (ultimo accesso 10/12/2020)

[5] *Uefa.tv home page*, UEFA.tv (web), https://www.uefa.tv/ (ultimo accesso 10/12/2020)

[6] *Raiplay home page*, RaiPlay (web), https://www.raiplay.it/ (ultimo accesso 10/12/2020)

[7] *BBC Archive home page*, BBC (web), https://www.bbc.co.uk/archive/ (ultimo accesso 10/12/2020)

[8] A. Messina, *Documenting the Archive - Using Content Analysis Techniques*, in "EBU Technical Review" (online), n. 305, 2006, https://tech.ebu.ch/publications/trev_305-messina (ultimo accesso 10/12/2020)

[9] W. Bailer, A. Messina e F. Negro, *Task-based assessment of performance and cost-effectiveness of automatic metadata extraction*, in "2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)", 2014, pp. 1-6, **DOI**: *10.1109/CBMI.2014.6849826*

[10] E. Caimotti, M. Montagnuolo e A. Messina, *An Efficient Visual Search Engine for Cultural Broadcast Archives*, in "Proceedings of the 11th International Workshop on Artificial Intelligence for Cultural Heritage co-located with the 16th International Conference of the Italian Association for Artificial Intelligence (AI*CH@AI*IA 2017)", 2017, pp: 1-8, *http://ceur-ws.org/Vol-2034/paper_1.pdf*

[11] A. Mercier ed altri, *Examples of Uses of Artificial Intelligence in Video Archives*, in "AI4TV '19: Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery", 2019, pp. 49-50, **DOI**: *10.1145/3347449.3357486*

[12] A. Messina ed altri, *Hyper Media News: a fully automated platform for large scale analysis, production and distribution of multimodal news content*, in "Multimedia Tools and Applications", 2013, vol. 63, n. 2, pp. 427-460, **DOI**: *10.1007/s11042-011-0859-1*

[13] *BBC Four announces experimental AI and archive programming*, BBC Media Centre (web), 17/08/2018, *https://www.bbc.co.uk/mediacentre/latestnews/2018/bbc-four-ai* (ultimo accesso 10/12/2020)

[14] *Web Ontology Language (OWL) home page*, W3C Semantic Web (web), *https://www.w3.org/OWL/* (ultimo accesso 10/12/2020)

[15] Y. Raimond ed altri, *Using the Past to Explain the Present: Interlinking Current Affairs with Archives via the Semantic Web*, in "The Semantic Web – ISWC 2013", 2013, Lecture Notes in Computer Science, vol. 8219, pp. 146-161, **DOI**: *10.1007/978-3-642-41338-4_10*

[16] A. Punchihewa, *Artefacts in image and video systems: Classification and mitigation*, in "Image and Vision Computing New Zealand", 2002, *https://www.researchgate.net/publication/264237884_Artefacts_in_image_and_video_systems_Classification_and_mitigation*

[17] A. C. Kokaram, *Motion Picture Restoration: Digital Algorithms for Artefact Suppression in Degraded Motion Picture Film and Video*, Springer Science & Business Media, 2013, **ISBN**: 1447134850

[18] F. Salmon, F. Vallet, *An Effortless Way To Create Large-Scale Datasets For Famous Speakers*, in "Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)", 2014, pp. 348-352, *http://www.lrec-conf.org/proceedings/lrec2014/pdf/32_Paper.pdf*

[19] G. B. Fonseca ed altri, *Hierarchical multi-label propagation using speaking face graphs for multimodal person discovery*, in "Multimed Tools Appl (2020)", **DOI**: *10.1007/s11042-020-09692-x*

[20] Jinfang Niu, *Organisation and description of datasets*, in "Archives and Manuscripts", vol. 44, n. 2, 2016, pp. 73-85, **DOI**: *10.1080/01576895.2016.1179585*

[21] A. Messina, *Dataset Production As A New Process In Future AI-Empowered Media*, in "IBC 2020 Conference", 2020, *https://www.ibc.org/technical-papers/dataset-production-as-a-new-process-in-future-ai-empowered-media/6763.article*

Sistemi a supporto dei giornalisti

L'Intelligenza Artificiale entra nella newsroom

Maurizio **Montagnuolo**
Rai - Centro Ricerche, Innovazione Tecnologica e Sperimentazione

L'*Intelligenza Artificiale* (*IA*) sta diventando una tecnologia pervasiva, utilizzata in molteplici ambiti e servizi. Grazie alla maggiore potenza di calcolo, ed alla disponibilità sempre più ingente di dati a disposizione delle organizzazioni, sono diversi i settori in cui le applicazioni dell'IA stanno rivoluzionando le modalità lavorative. Tra questi ci sono, ad esempio, l'industria dell'intrattenimento, i servizi finanziari e la logistica. Naturalmente, anche l'ambito giornalistico non poteva rimanere insensibile alle tante potenzialità offerte dall'IA, la quale sta avendo un profondo impatto sui processi di reperimento, produzione e distribuzione delle news. Una recente indagine [1] ha rivelato una crescente attenzione da parte dei giornalisti nei confronti dell'IA, in particolare con gli obiettivi di rendere il proprio lavoro più efficiente, efficace ed in grado di soddisfare i bisogni degli utenti.

Questo articolo fornisce una panoramica sull'utilizzo dell'IA in ambito giornalistico, soffermandosi sullo stato dell'arte attuale, ma con un focus rivolto anche alle future prospettive ed evoluzioni.

*L'*Intelligenza Artificiale *sta diventando una tecnologia pervasiva utilizzata anche in ambito giornalistico, fornendo un valido supporto nella raccolta di notizie, dati ed informazioni, nonché nella produzione di report, video e documentazione utili alla narrazione di un evento di interesse mediatico.*

Strumenti quali la ricerca semantica, la trascrizione del parlato e la creazione automatica di contenuti multimediali stanno diventando di uso sempre più frequente nelle redazioni di radio, televisioni e stampa.

Al fine di sfruttare nel modo più efficiente possibile le opportunità offerte dagli strumenti di Intelligenza Artificiale, è tuttavia indispensabile definire una strategia, organizzativa e tecnologica, che promuova la comprensione e la divulgazione dei loro limiti e potenzialità all'interno delle redazioni stesse.

Questo articolo fornisce una panoramica sull'utilizzo dell'Intelligenza Artificiale in ambito giornalistico, soffermandosi sullo stato dell'arte tecnologico, ed i relativi esempi applicativi, ma con uno sguardo rivolto anche agli attuali limiti ed alle future evoluzioni.

Reperimento, produzione e distribuzione dell'informazione

Le prime applicazioni dell'intelligenza artificiale in ambito giornalistico risalgono agli inizi degli anni 2010. Referenziate con il termine *BOT* (dall'inglese *robot journalism*), esse sono volte alla raccolta, analisi ed interpretazione dei dati al fine di redigere un articolo, un report o un approfondimento in maniera automatica.

201_02.3_3.00

Nel 2010 il laboratorio di intelligenza artificiale della **Northwestern University**, negli Stati Uniti, sviluppò il sistema ***Stats Monkey*** in grado di scrivere articoli sportivi sul baseball basandosi su informazioni statistiche reperibili online [2].

Nel 2014, il **Los Angeles Times** usò il sistema ***Quakebot*** per redigere in anteprima la notizia sul terremoto con epicentro nel sud della California pochissimi istanti dopo la fine dello stesso [3].

Il **Washington Post** applicò un BOT chiamato ***Heliograf*** per pubblicare online notizie relative ai *XXXI Giochi Olimpici di Rio 2016* ed alle elezioni presidenziali americane del 2016 [4].

Recentemente, l'agenzia **Reuters** ha presentato un prototipo in grado di catturare gli attimi salienti di una partita di calcio e di generare in maniera completamente automatica una sintesi video con tanto di presentatore [5].

Infine, in Italia, l'agenzia **ANSA** ha introdotto l'uso delle tecniche di generazione automatica del linguaggio (*NLG – Natural Language Generation*) per produrre notizie e grafici in tempo reale sull'evoluzione della pandemia da Covid-19 basandosi sui dati quotidianamente distribuiti dalla Protezione Civile [6].

Gli esempi sopra citati aiutano a capire le direzioni intraprese dall'intelligenza artificiale nel mondo del giornalismo e le opportunità rese disponibili dalla sua applicazione.

A tal proposito, un recente report pubblicato dall'**Associated Press** [7] indica cinque aree di particolare interesse:

- l'*apprendimento automatico* (in inglese *Machine Learning*) per la classificazione (*classification*), segmentazione (*segmentation*) e raggruppamento (*clustering*) dei contenuti;
- il *trattamento del linguaggio* (in inglese *Natural Language Processing*), per il processamento e la generazione di testi;
- l'*analisi acustica* (in inglese *Audio Processing*), per la trasformazione e comprensione di suoni, musica e parlato;
- la *visione artificiale* (in inglese *Computer Vision*), per l'analisi e classificazione di contenuti visuali, quali, ad esempio, il riconoscimento di oggetti e volti nel video o la sottotitolazione di immagini;
- la *robotica* (in inglese *Robotics*), per la realizzazione, ad esempio, di telecamere a controllo automatico o droni a volo autonomo.

Si legge, inoltre, che la sinergia di tali tecnologie:

> *[...]permetterebbe ai giornalisti di analizzare dati; identificare schemi, tendenze ed informazioni da molteplici fonti; individuare particolari non rilevabili a occhio nudo; trasformare i dati e le parole in testo ed il testo in audio e video; comprendere le emozioni; riconoscere oggetti, volti, testo o colori – e molto altro ancora.[...]*

Alcuni esempi applicativi delle suddette tecnologie, *descritte più dettagliatamente nei paragrafi seguenti*, sono illustrati in Fig. 1.

Fig. 1 – Principali compiti dell'*Intelligenza Artificiale* in ambito giornalistico e relativi esempi applicativi.
(icone da https://pixabay.com/)

Apprendimento automatico

L'*apprendimento automatico* è un sottoinsieme dell'intelligenza artificiale che, mediante l'utilizzo di metodi statistici, mira alla creazione di modelli analitici in grado di descrivere un insieme di dati. La particolarità dell'apprendimento automatico consiste nella capacità di un sistema informatico di imparare dai dati analizzati senza l'utilizzo di istruzioni esplicite. Gli algoritmi per l'apprendimento automatico sono normalmente distinti in tre classi:

- *supervisionato* (supervised learning),
- *non supervisionato* (unsupervised learning),
- *per rinforzo* (reinforcement learning).

Nell'*apprendimento supervisionato* il sistema viene istruito fornendo sia l'insieme di dati in ingresso, sia l'informazione desiderata a seguito del processamento, con l'obiettivo di identificare un insieme di regole (cioè una funzione matematica) che colleghi i dati in ingresso con quelli in uscita. L'insieme di regole appreso può quindi essere utilizzato per svolgere compiti simili su insiemi di dati mai elaborati precedentemente. Appartengono alla classe dell'apprendimento supervisionato gli *algoritmi di classificazione* e *di regressione*, quali ad esempio *alberi decisionali* (*decision trees*), *Support Vector Machine* (*SVM*), *Naive Bayes* e *K-Nearest Neighbor* (*KNN*). Un esempio di apprendimento supervisionato è descritto in [8], in cui gli autori hanno addestrato un modello di classificazione per la categorizzazione di contenuti giornalistici di divulgazione scientifica in base alla *tassonomia dei Settori Scientifico Disciplinari* del **MIUR**[Nota 1].

Nell'*apprendimento non supervisionato* al sistema viene fornito solo l'insieme di dati in ingresso senza alcuna indicazione del risultato desiderato. Lo scopo di questo metodo di apprendimento è quello di identificare correlazioni tra i dati senza che questi siano preventivamente etichettati. Appartengono alla classe dell'apprendimento non supervisionato gli *algoritmi di aggregazione* (*clustering*) e di *riduzione della dimensionalità*, quali ad esempio *K-Means*, *misture di Gaussiane* (*Gaussian Mixture Models*) e *Principal Component Analysis* (*PCA*).

A titolo di esempio, ***Hyper Media News*** *(**HMN**)* [9] permette di produrre rassegne stampa giornaliere contenenti le notizie più importanti della giornata, con collegamenti a servizi dei telegiornali ed articoli Web corredati da informazioni statistiche sulle tematiche individuate[Nota 2].

Nell'*apprendimento per rinforzo* il sistema apprende come risolvere un determinato problema basandosi sulla correttezza o meno delle azioni svolte durante la risoluzione del problema stesso. Questa tecnica si basa sul concetto di *ricompensa e penalità*: un'azione corretta comporterà un premio mentre un'azione scorretta porterà ad una penalizzazione. L'obiettivo finale consiste nella risoluzione del problema cui il sistema è preposto massimizzando le ricompense e minimizzando le penalità. Nell'ambito giornalistico, gli algoritmi di apprendimento per rinforzo possono essere usati, ad esempio, nei *sistemi di raccomandazione* per suggerire agli utenti articoli, video, approfondimenti che risultino di loro interesse [10][11].

Trattamento del linguaggio

L'*analisi del linguaggio naturale* è uno degli ambiti più rilevanti dell'intelligenza artificiale. In particolare, due importanti applicazioni che stanno avendo un impatto significativo nel settore giornalistico sono la *generazione* ed il *processamento del linguaggio*, rispettivamente denominate *Natural Language Generation* (*NLG*) e *Natural Language Processing* (*NLP*). Sinteticamente, si potrebbe asserire che la *generazione* è riconducibile alla *scrittura*, mentre il *processamento* alla *lettura*.

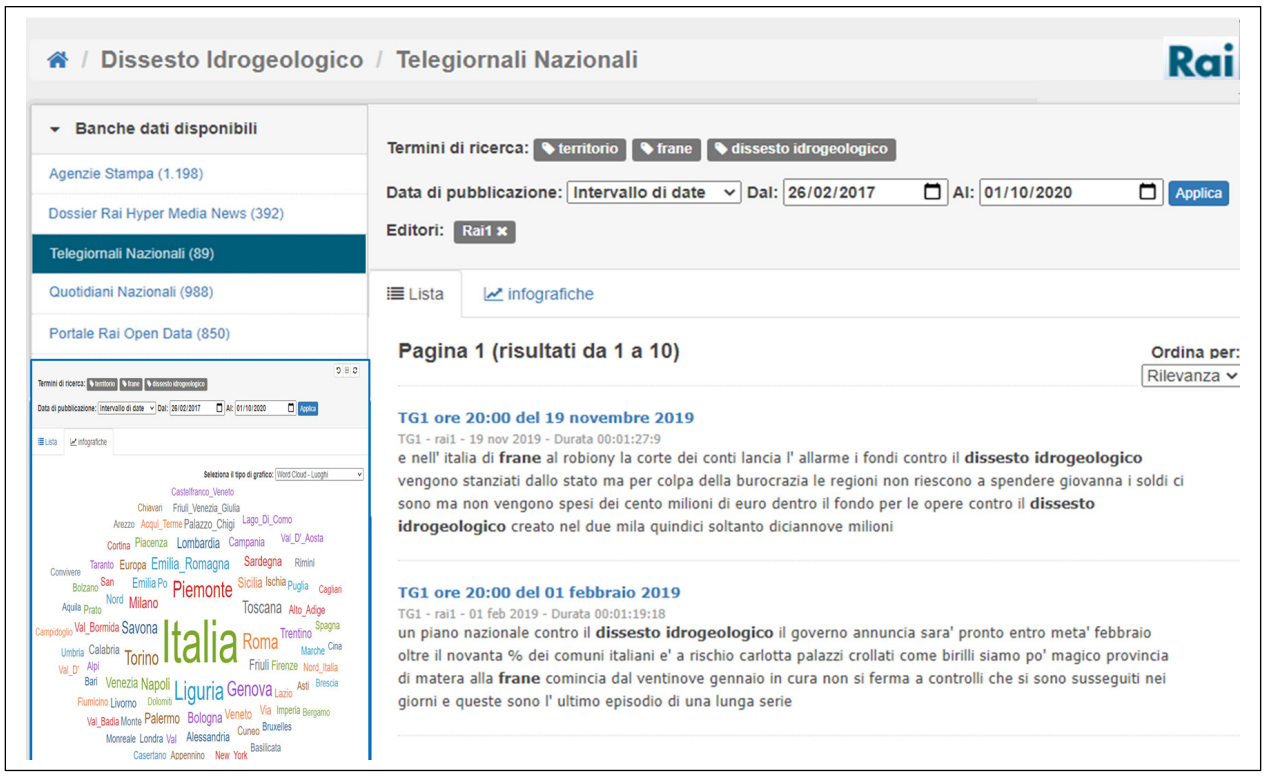
Nota 1 - https://www.miur.gov.it/settori-concorsuali-e-settori-scientifico-disciplinari *(ultimo accesso 27/11/2020)*

Nota 2 - Maggiori informazioni sono disponibili nella sezione dedicata ai progetti ANTS (www.crit.rai.it/CritPortal/progetti/?p=249) ed HMN (www.crit.rai.it/CritPortal/progetti/?p=51) del sito CRITS *(ultimo accesso 27/11/2020)*

L'obiettivo dell'NLG è utilizzare l'IA per produrre narrazioni scritte partendo da un insieme di dati numerici. Tali peculiarità rendono l'NLG un'ottima soluzione per la creazione di contenuti giornalistici a partire da dati strutturati. In letteratura si trovano molti esempi di applicazioni NLG che, basandosi su statistiche quali i risultati elettorali, i punteggi degli eventi sportivi o l'andamento del mercato azionario, permettono di creare approfondimenti testuali sull'argomento in esame [12].

Gli algoritmi NLP sono preposti alla comprensione del testo e del parlato. Tali metodologie sono alla base del funzionamento degli assistenti vocali, i quali mediante tecniche di analisi, contestualizzazione e disambiguazione assolvono al compito di rispondere alle domande degli utenti ed interagire con essi in tempo reale. L'NLP può essere utilizzato per una moltitudine di applicazioni, tra le quali la creazione automatica di riassunti, il riconoscimento delle entità nominali (ad esempio nomi di persone, organizzazioni, luoghi) e la traduzione automatica. Il **CRITS** sta affrontano queste tematiche in modo uniforme ed interdisciplinare. Alla base dell'approccio c'è un insieme di tecniche di processamento del linguaggio e strumenti statistici avanzati mirati all'individuazione di collegamenti semantici tra contenuti multimediali, fornendo agli utenti testi, grafica e notizie video organizzate secondo i propri interessi individuali[Nota 3 (pag. seguente)]. Il risultato di tale studio è un sistema che consente la definizione di profili di ricerca personalizzati aggiornati automaticamente e dinamicamente con i relativi contenuti provenienti dalle sorgenti di informazioni monitorate, tra le quali portali Web, canali televisivi ed altri circuiti specializzati. Un esempio è mostrato in Fig. 2.

Fig. 2 – Esempio di *interfaccia per la ricerca di contenuti multimediali*. Gli algoritmi di NLP sono utilizzati per aggregare i contenuti provenienti da banche dati eterogenee, ed arricchirli con informazioni semantiche. Nell'esempio sono visibili la lista dei telegiornali nazionali per l'argomento "Dissesto Idrogeologico", filtrati per canale di trasmissione e corredati dai luoghi in essi menzionati



Analisi acustica

Il *trattamento di contenuti audio* è un aspetto cruciale nelle attività quotidiane di ogni giornalista. Pertanto, anche in questo ambito sono state sviluppate applicazioni che, mediante l'uso dell'IA, supportano il lavoro giornalistico in maniera efficace ed efficiente. Tra queste, le più rilevanti sono la *trascrizione del parlato*, la *sintesi vocale* e l'*audio fingerprinting*.

La *trascrizione automatica del parlato in testo (STT, Speech-to-Text)*, utilizza dei modelli vocali per trasformare i suoni in caratteri, e successivamente i caratteri in parole. Disporre di applicazioni STT offre diversi vantaggi, tra cui la riduzione dei tempi di sbobinatura delle registrazioni audio (ad esempio appunti o interviste), e la disponibilità degli stessi per attività di archiviazione e ricerca (vedi Fig. 2 pagina precedente).

All'opposto, la *sintesi vocale (TTS, Text-to-Speech)* trasforma i caratteri in suoni, in accordo con le regole fonetiche della lingua sintetizzata. Un esempio è presentato in [13].

L'*audio fingerprint* è una descrizione compatta di una sequenza audio, che ne permette l'identificazione ed il riconoscimento. Un sistema di audio fingerprinting ha lo scopo di identificare duplicati di file audio anche in presenza di disturbi o manipolazioni. Il concetto è simile a quello delle impronte digitali umane, le quali, grazie alla loro unicità, sono usate per individuare l'identità di una persona. Esempi applicativi di interesse giornalistico includono la tutela del diritto d'autore, la localizzazione di un segmento audio all'interno di un file, e la verifica dell'originalità di un contenuto al fine di prevenire e ridurre il fenomeno delle fake news.

Visione artificiale

La *visione artificiale* è l'insieme dei processi che mirano all'interpretazione del contenuto visivo di un'immagine o un video. I sistemi di visione artificiale, utili in svariati settori tra i quali medico, automobilistico ed industriale, sono in grado di identificare rapidamente oggetti e persone, analizzare le azioni, ispezionare le linee produttive e molto altro ancora.

In ambito giornalistico, la visione artificiale può aiutare le redazioni nell'organizzazione di grandi archivi di immagini e video, rendendo conseguentemente il processo editoriale più veloce ed efficiente. Ad esempio, grazie alle tecniche di classificazione automatica è possibile la scrematura e la selezione di contenuti. Casi di successo sono rappresentati dalle applicazioni per la prevenzione e gestione dell'emergenza in occasione di calamità naturali [14] [15], o il riconoscimento di azioni per la creazione degli highlights di un evento sportivo.

Un altro esempio dell'utilizzo della computer vision per l'arricchimento delle news riguarda la possibilità di analizzare in tempo reale eventi di grande impatto mediatico, come nel caso dell'applicazione ***Who's Who*** di **Sky News**[Nota 4], utilizzata in occasione del matrimonio reale tra il Principe Harry e Meghan Markle. Infine, citiamo le applicazioni di sottotitolazione automatica che, lavorando in sinergia con i metodi di analisi del linguaggio, permettono di creare in modo completamente automatizzato didascalie e note relative alle immagini processate [16][17].

L'esempio di Fig. 3 pagina seguente mostra un'immagine tratta dall'***archivio Rai***, la cui didascalia è stata generata automaticamente tramite un servizio di IA disponibile in cloud.

Nota 3 - Maggiori informazioni sono disponibili nella sezione dedicata al *Progetto Data Driven Journalism* del sito CRITS (www.crit.rai.it/CritPortal/progetti/?p=895) *(ultimo accesso 27/11/2020)*

Nota 4 - https://news.sky.com/story/royal-wedding-whos-who-11356656 *(ultimo accesso 27/11/2020)*

Fig. 3 – Esempio di *descrizione automatica* tramite l'utilizzo di reti neurali. Si noti la capacità del sistema di riconoscere il personaggio ritratto nella foto, nonché la sua posa e posizione rispetto ad altri oggetti *"Alberto Sordi indossa un abito e posa di fronte ad un pianoforte"*.

Robotica

Oltre che per l'analisi dei contenuti multimediali, le tecnologie IA possono essere profittevolmente utilizzate per la produzione dei contenuti stessi. In questo caso, gli algoritmi di IA sono direttamente integrati nei dispositivi hardware quali telecamere, microfoni e scanner laser.

Le *telecamere a controllo automatico*, ed i *droni a volo autonomo* sono preziosi strumenti in grado di catturare immagini da un punto di vista completamente diverso rispetto alle tecniche di ripresa tradizionali, fornendo rappresentazioni più immersive e coinvolgenti [18]. Ad esempio, come riportato in [7], un insieme di telecamere robotizzate è stato usato durante i *XXXI Giochi Olimpici di Rio 2016* per il montaggio video di scene normalmente inaccessibili mediante metodi di ripresa tradizionali.

Anche l'audio assume un ruolo di primo piano nella produzione giornalistica. L'uso dei *podcast* ha visto negli anni una crescita sempre maggiore, affiancata dalla diffusione degli *assistenti vocali* e degli *smart speaker*. In questo contesto, l'IA viene impiegata per migliorare la qualità dell'audio catturato dai microfoni (ad esempio rimuovendo il rumore di fondo), e per identificare voci e/o suoni provenienti contemporaneamente da persone/sorgenti audio differenti [19].

Infine, gli *scanner laser*, quali ad esempio il ***LIDAR***, sono strumenti che permettono di analizzare l'ambiente fornendo una rappresentazione tridimensionale della scena analizzata. Di notevole importanza per le applicazioni del campo automobilistico, questi dispositivi stanno riscuotendo interesse anche nell'ambito giornalistico, in quanto abilitanti di servizi quali il *giornalismo immersivo* o la *realtà aumentata*.

Conclusioni

Le tecnologie dell'IA stanno aprendo nuovi orizzonti e cambiando il modo in cui le notizie vengono collezionate, costruite, distribuite e fruite. Nell'era dei social network e dell'informazione globale il giornalista agisce da mediatore tra il pubblico e la storia, occupandosi di rinnovare, aggiornare e perfezionare le notizie in un flusso continuo di informazioni, e seguendo in tempo reale l'evoluzione degli avvenimenti narrati. La mediazione può inoltre avvenire con modalità e tramite canali differenti, includendo non solo testi, ma anche contenuti multimediali e virtuali.

L'IA rappresenta un utile strumento di supporto ai giornalisti nell'assolvimento di tale compito. Tuttavia, alcune questioni tecniche ed organizzative rappresentano ancor oggi sfide da vincere e potenziali ostacoli. Gli algoritmi dell'IA sono complicati ed esistono modalità diverse con cui possono essere implementati; occorre pertanto acquisire conoscenza e consapevolezza delle potenzialità, e ancor più delle limitazioni, insiti in tali strumenti al fine di poterli utilizzare nel modo più proficuo possibile. Inoltre, occorre ridefinire alcuni concetti del processo editoriale, al fine di uniformare le modalità produttive tradizionali con le novità richieste ed apportate dall'IA.

Infine, occorre sempre assicurarsi che gli strumenti IA rispettino le regole etiche e deontologiche del giornalismo al fine di garantire in qualsiasi situazione la diffusione di informazioni precise, corrette ed esaustive.

Bibliografia

[1] C. Beckett, *New powers, new responsibilities. A global survey of journalism and artificial intelligence*, 2019, LSE-Polis (web), https://blogs.lse.ac.uk/polis/2019/11/18/new-powers-new-responsibilities/ (ultimo accesso 27/11/2020)

[2] N. D. Allen ed altri, *StatsMonkey: A Data-Driven Sports Narrative Writer*, in "Computational Models of Narrative: papers from the 2010 AAAI Fall Symposium", Technical Report FS-10-04, 2010, pp. 2-3, https://www.aaai.org/ocs/index.php/FSS/FSS10/paper/view/2305

[3] J. Plucinska, *How an Algorithm Helped the LAT Scoop Monday's Quake*, in "Columbia Journalism Review" (web), 2014, https://www.cjr.org/united_states_project/how_an_algorithm_helped_the_lat_scoop_mondays_quake.php (ultimo accesso 27/11/2020)

[4] M. Ciobanu, *A snapshot of news organisations' reporting on US election day*, in "journalism.co.uk" (web), 2016, https://www.journalism.co.uk/news/a-snapshot-of-news-organisations-reporting-on-us-election-day/s2/a690886/ (ultimo accesso 27/12/2020)

[5] S. Chandler, *Reuters Uses AI To Prototype First Ever Automated Video Reports*, in "Forbes" (web), 2020, https://www.forbes.com/sites/simonchandler/2020/02/07/reuters-uses-ai-to-prototype-first-ever-automated-video-reports/ (ultimo accesso 27/11/2020)

[6] Redazione ANSA, *Coronavirus: Ansa e Applied XLab producono notizie sull'epidemia grazie all'intelligenza artificiale*, in "ANSA.it" (web), 2020, https://www.ansa.it/sito/notizie/cronaca/2020/04/27/coronavirus-ansa-e-applied-xlab-producono-notizie-sullepidemia-grazie-allintelligenza-artificiale_7fecc4c3-8c58-4cbe-815c-a46b51684cd4.html (ultimo accesso 27/11/2020)

[7] F. Marconi, A. Siegman e Machine Journalist, *The future of augmented journalism: A guide for newsrooms in the age of smart machines*, Associated Press, 2017, https://insights.ap.org/industry-trends/report-how-artificial-intelligence-will-impact-journalism

[8] M. Montagnuolo ed altri, *Applying Natural Language Processing to Speech Transcriptions for Automated Analysis of Educational Video Broadcasts*,

in "Proceedings of the 11th International Workshop on Artificial Intelligence for Cultural Heritage co-located with the 16th International Conference of the Italian Association for Artificial Intelligence (AI*CH@ AI*IA 2017)", 2017, pp. 28-29, http://ceur-ws.org/Vol-1983/paper_02.pdf

[9] A. Messina ed altri, *Hyper Media News: a fully automated platform for large scale analysis, production and distribution of multimodal news content*, in "Multimedia Tools and Applications", vol. 63, n. 2, 2013, pp. 427-460, **DOI**: 10.1007/s11042-011-0859-1

[10] A. Coenen, *How The New York Times is Experimenting with Recommendation Algorithms*, in "NYT Open" (web), 2019, https://open.nytimes.com/how-the-new-york-times-is-experimenting-with-recommendation-algorithms-562f78624d26 (ultimo accesso 27/11/2020)

[11] J. Misztal-Radecka ed altri, *Trend-Responsive User Segmentation Enabling Traceable Publishing Insights. A Case Study of a Real-World Large-Scale News Recommendation System*, in "Proceedings of the 7th International Workshop on News Recommendation and Analytics (INRA 2019) in conjunction with 13th ACM Conference on Recommender Systems (RecSys 2019)", 2019, pp. 53-62, http://ceur-ws.org/Vol-2554/paper_08.pdf

[12] L. Leppänen ed altri, *Data-Driven News Generation for Automated Journalism*, in "Proceedings of the 10th International Conference on Natural Language Generation", 2017, pp. 188-197, **DOI**: 10.18653/v1/W17-3528

[13] S. A. K. Weber e X. Bai, *Video translation: weaving synthetic voices into the multilingual production workflow*, in "IBC2016 Conference", 2016, https://www.ibc.org/video-translation-weaving-synthetic-voices-into-the-multilingual-production-workflow/901.article

[14] N. Said ed altri, *Natural disasters detection in social media and satellite imagery: a survey*, in "Multimedia Tools and Applications", vol. 78, n. 22, 2019, pp. 31267-31302, **DOI**: 10.1007/s11042-019-07942-1

[15] Yalong Pi, Nipun D. Nath e Amir H. Behzadan, *Convolutional neural networks for object detection in aerial imagery for disaster response and recovery*, in "Advanced Engineering Informatics", vol. 43, 2020, **DOI**: 10.1016/j.aei.2019.101009

[16] A. Furkan Biten ed altri, *Good News, Everyone! Context Driven Entity-Aware Captioning for News Images*", in "2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)", 2019, pp. 12458-12467, **DOI**: 10.1109/CVPR.2019.01275

[17] A. Tran, A. Mathews e L. Xie, *Transform and Tell: Entity-Aware News Image Captioning*, in "2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)", 2020, pp. 13032-13042, **DOI**: 10.1109/CVPR42600.2020.01305

[18] A. Ntalakas ed altri, *Drone Journalism: Generating Immersive Experiences*, in "Journal of Media Critiques", vol. 3, n. 11, 2017, pp. 187-199, **DOI**: 10.17349/jmc117317

[19] H. Sundar ed altri, *Raw Waveform Based End-to-end Deep Convolutional Network for Spatial Localization of Multiple Acoustic Sources*, in "ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", 2020, pp. 4642-4646, **DOI**: 10.1109/ICASSP40776.2020.9054090

Intelligenza Artificiale e Codifica Video

Una strada per superare i limiti dell'approccio tradizionale.

Roberto **Iacoviello**, Angelo **Bruccoleri**
Rai - Centro Ricerche, Innovazione Tecnologica e Sperimentazione

Nel corso dell'ultimo decennio il consumo di contenuti video sui principali canali online è cresciuto a dismisura fino a far registrare numeri da record nel corso del 2020, assestandosi su una media giornaliera del 65% del totale traffico internet.

Questa tendenza, unita alla richiesta di contenuti ad altissima qualità veicolati via etere, mette a dura prova i canali di trasmissione sia broadband che broadcast, confermando l'importanza e la centralità di un sistema di compressione video efficiente all'interno della catena di distribuzione. Lo sviluppo di uno standard di compressione che unisca qualità visiva e bassi bit-rate è senza dubbio un compito arduo, soprattutto considerando il vincolo della complessità computazionale.

Negli ultimi anni si è assistito ad un interessante cambio di rotta all'interno della comunità scientifica e dell'industria verso algoritmi basati su Intelligenza Artificiale. La costante maturazione delle tecniche di Intelligenza Artificiale e la fiorente attività di ricerca e sviluppo hanno identificato la strada maestra per i possibili sviluppi futuri nel campo della codifica video dando vita a due diversi filoni di studio: il primo prevede il miglioramento dello schema tradizionale di codifica video (Hybrid Coding) mentre il secondo ricerca architetture alternative ad esso (End-to-End Coding).

Negli ultimi anni si è assistito ad una vera e propria rivoluzione nel mondo cinematografico e televisivo, grazie all'avvento di nuovi formati digitali che ha coinvolto l'intera catena di produzione dei prodotti multimediali: **Ultra High Definition** (***UHD***), **High Dynamic Range** (***HDR***), **High Frame Rate** (***HFR***), solo per citarne alcuni. Tale rivoluzione ha impattato l'industria dei *multimedia*, quella della *Consumer Electronics* e quella delle *reti di comunicazione*, aprendo nuove opportunità di convergenza.

La qualità del video è cresciuta in modo esponenziale puntando a raggiungere la ricchezza cromatica, la dinamica e la resa dei dettagli della visione umana, ma al contempo ponendo problemi per quanto riguarda la larghezza di banda nei canali trasmissivi e la memorizzazione su supporti fissi e rendendo necessari nuovi e più performanti *standard di compressione* del segnale video che lascino inalterata la qualità.

L'approccio tradizionale alla *codifica video*, in auge da più di 30 anni, sta oramai raggiungendo il suo limite mentre contemporaneamente stiamo assistendo alla diffusione pervasiva di tecniche di *Intelligenza Artificiale (IA)*, il cui successo è dovuto in buona parte alle prestazioni raggiunte dalle *Reti Neurali Profonde (Deep Neural Networks, DNN)*.

201_02.4_3.00

ISO/IEC JTC1 SC29 MPEG, l' *International Organization for Standardization/International Electrotechnical Commission - Joint Technical Committee 1 - Sub Committee 29 - Moving Picture Experts Group*, è il principale organismo internazionale che da oltre 30 anni si occupa di standard di compressione. Dai gruppi di lavoro costituiti al suo interno sono stati creati standard che hanno ottenuto un consenso universale sia per quanto riguarda l'adozione da parte di diversi settori merceologici (si pensi, ad esempio, alla compressione del genoma umano) sia a livello di copertura geografica. Negli ultimi anni sono comparsi sul mercato altri due gruppi internazionali che mirano a competere con il gruppo ***MPEG*** nel campo della compressione video:

- ***Alliance for Open Media (AOM)***: consorzio fondato nel 2015, è costituito da aziende del calibro di ***Apple***, ***Amazon***, ***ARM***, ***Cisco***, ***Facebook***, ***Google***, ***IBM***, ***Intel***, ***Microsoft***, ***Mozilla***, ***Netflix*** e ***Nvidia***. Il video codec prodotto, chiamato **AV1**, ha dimostrato di avere prestazioni paragonabili a ***MPEG HEVC***.
- ***Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI)***: gruppo fondato nel 2020 con la missione di *sviluppare specifiche di compressione dei dati digitali abilitate dall'intelligenza artificiale* con un chiaro sistema di licenza *IPR (Intellectual Property Rights)*.

C'è un gran fermento attorno ai codec video e nei prossimi anni assisteremo ad una vera e propria battaglia nel campo della compressione del segnale video.

Video codec tradizionali

Nel 2013 è stato ufficialmente pubblicato lo standard **HEVC** *(High Efficiency Video Coding)* definito dal gruppo ***MPEG*** e dal 1° gennaio 2017 è stato reso obbligatorio sui dispositivi di ricezione televisiva in commercio in Italia [1].

Nel 2020 è stato finalizzato il nuovo standard, chiamato **Versatile Video Coding (VVC)**[2]. **VVC** è stato definito per offrire un risparmio di banda fino al *50%* rispetto al suo predecessore **HEVC** a parità di qualità dell'immagine e si presenta quindi come soluzione ideale per la televisione ad altissima definizione (*Ultra High Definition*, **UHD**) e oltre, essendo in grado di gestire risoluzioni che vanno dalla **SD** (*Standard Definition*) fino al **16K** (4 volte i pixel di una risoluzione **8K**) e frame rate fino a *120 fps* (*frames per second*).

Nel complesso, **VVC** propone un compromesso ottimale tra complessità computazionale, tasso di compressione, robustezza agli errori e ritardi di processing e introduce significativi passi avanti sul fronte della qualità dell'immagine espressa in termini di *range dinamico*, *gamut*, *alti frame rate* e *riduzione del rumore*.

Come i suoi predecessori, utilizza l'approccio di *codifica video ibrida* basato sul partizionamento dell'immagine in *blocchi*, un concetto alla base di tutti i principali standard di codifica video a partire dallo standard **MPEG-1** (del 1988). In questo schema, ogni fotogramma di un video viene suddiviso in blocchi e tutti i blocchi vengono elaborati in sequenza.

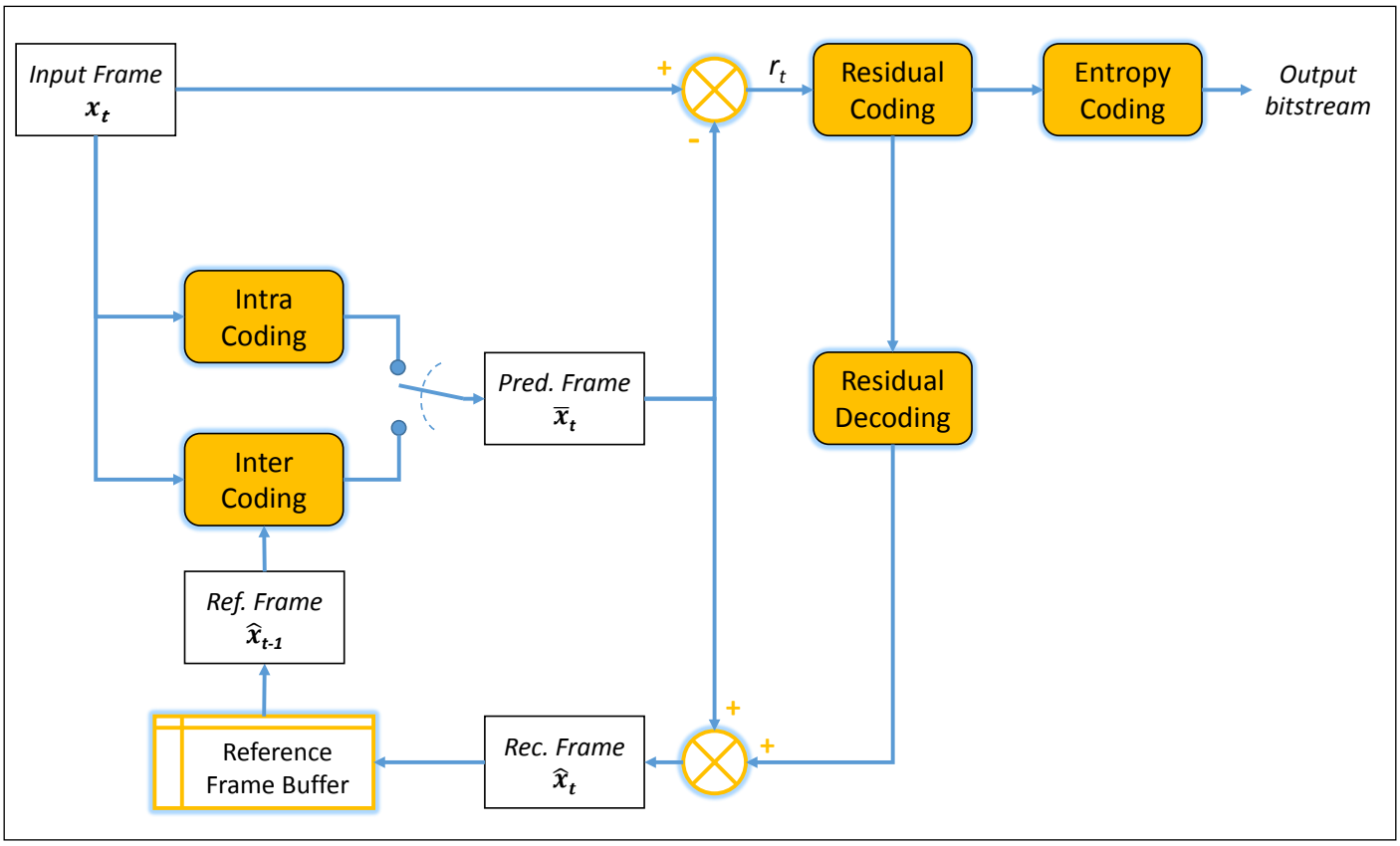
Inoltre, come per i suoi predecessori **MPEG-2**, **AVC** e **HEVC**, **VVC** è un sistema di compressione basato sull'inserimento nel loop di codifica di strumenti per la riduzione della *ridondanza spaziale* e *temporale* che caratterizza il segnale video, in particolare:

- *compressione senza perdita di informazione*, basata sullo sfruttamento della ridondanza spaziale (correlazione tra pixel adiacenti nello stesso frame), della ridondanza temporale (correlazione tra frame diversi nel tempo) e sulla *codifica entropica*;
- *compressione con eliminazione dell'irrilevanza*, ossia di quell'informazione non più ricostruibile dal decodificatore, ma non percepibile dal sistema visivo umano (*codifica psico-visiva*). Tale approssimazione avviene in un dominio diverso da quello dei dati originali, dominio che si ottiene per mezzo di tecniche di codifica basate su trasformate;
- *compressione con perdita di informazione*, legata al processo di quantizzazione.

Il codificatore (Fig. 1) elabora ogni *blocco* in un ciclo:

- al blocco che entra nel ciclo di codifica (*Input Frame*) viene sottratto un segnale di previsione (*Pred. Frame*), generando un segnale di errore r_t nel dominio dei pixel. Esistono due tipi di previsione, *Inter Coding*, che utilizza blocchi da immagini *temporalmente attigue* effettuando una compensazione del movimento, e *Intra Coding*, che utilizza solo le informazioni presenti all'*interno della medesima immagine*;
- il risultato della sottrazione tra il blocco originale e quello predetto, detto *residuo*, viene sottoposto ad un'operazione di *trasformazione* e *quantizzazione* (*Residual Coding*). L'algoritmo più usato è quello chiamato *DCT (Discrete Cosine Transform, trasformata discreta del coseno)* applicato a blocchi di pixel, ma ne sono disponibili anche altri;
- infine, il codec quantizza i coefficienti generati dalla *DCT*, elimina quelli che hanno valore pari a zero ed applica la *codifica entropica* per sfruttare le ridondanze statistiche (*Entropy Coding*);
- all'interno del loop del codificatore, i coefficienti vengono de-quantizzati, ritrasformati nel dominio dei pixel (*Residual Decoding*) e sommati alla previsione ottenendo il blocco ricostruito (*Rec. Frame*) che viene sottoposto ad alcuni filtraggi. Questa fase di solito include un filtro per rimuovere gli artefatti che si verificano ai confini dei blocchi e filtri più avanzati per la ricostruzione dei contorni. Infine, il blocco viene salvato in un buffer (*Reference Frame Buffer*) in modo che possa essere ricostruita l'immagine (*Ref. Frame*) da rendere disponibile al codificatore per lo sfruttamento delle ridondanze temporali e il ciclo può continuare con il blocco successivo.

Fig. 1 – Schema del *codificatore ibrido MPEG*



Approcci basati su Intelligenza Artificiale

L'approccio tradizionale sta ormai mostrando i suoi limiti, infatti sebbene ogni parte del codificatore sia ben progettata per comprimere il video il più possibile, data la non linearità del sistema è molto difficile stimare quale sia la configurazione ottimale in funzione del segnale di ingresso: il codificatore è composto da numerosi tool da utilizzare in alternativa (in **VVC** sono più di venti) e sta diventando complicato armonizzare i loro contributi in un'ottica globale, soprattutto quando il codec deve operare in tempo reale.

Queste limitazioni hanno portato i ricercatori a creare nuovi algoritmi ed in particolare la loro attenzione si è rivolta a quelli basati sulle *reti neurali profonde (DNN)*.

Le reti neurali hanno mostrato prestazioni eccezionali in termini di previsione e classificazione, elementi particolarmente importanti anche nel campo della compressione video. Dunque i ricercatori hanno iniziato a prestare attenzione ad essi come candidati promettenti per un approccio alla codifica video di nuova generazione.

Dal punto di vista dell'architettura, due diversi approcci sono proposti in letteratura: la *codifica ibrida* (*Hybrid Coding*) basata su blocchi con potenziamento tramite reti neurali e la *codifica basata sull'apprendimento End-to-End* (*E2E coding*).

Negli approcci di *codifica ibrida*, le *DNN* sostituiscono alcuni degli strumenti di codifica esistenti o vengono utilizzate come metodi di ottimizzazione, preservando così l'architettura convenzionale basata su blocchi. In questo modo si parte da un sistema già altamente ottimizzato e si ricercano ulteriori strumenti di miglioramento, ma non si affronta il problema dell'ottimizzazione globale del sistema.

Al contrario, nell'*approccio End-to-End*, un'unica *DNN* svolge tutte le funzioni di compressione cercando di ottenere, durante la fase di apprendimento, un'ottimizzazione globale del sistema.

Hybrid Coding

Nella *codifica ibrida*, le *DNN* sono utilizzate per stime quali *previsione intra/inter* e *rimozione di artefatti di compressione*. Nella codifica entropica, si usano le DNN per prevedere la probabilità dei contesti per la *codifica aritmetica binaria adattiva (CABAC)*. Tra gli strumenti di codifica basati su *DNN*, il filtraggio ha mostrato un miglioramento pari al *5.57%* nell'implementazione descritta in [3], mentre in [4] la predizione *intra* fornisce un guadagno pari al *6.05%* rispetto a **VVC**. Il maggior guadagno, pari al *10.05%* sempre rispetto a **VVC,** si ottiene con la *Super Resolution* [5].

Con il termine *Super Resolution (SR)* ci si riferisce ad una classe di tecniche di image processing basate su deep learning aventi l'obiettivo di incrementare la risoluzione delle immagini, trasformando, ad esempio, un segnale **HD** in **4K**. Negli ultimi anni la ricerca scientifica ha prodotto un cospicuo numero di soluzioni e implementazioni, da quelle basate su reti **CNN** [6] o **RESNET** [7] a quelle basate su reti **GAN** [8]. La maggior parte di questi studi si concentra sulla cosiddetta *Single Image Super Resolution (SISR)*, ovvero sul miglioramento della singola immagine partendo dalla sua rappresentazione a bassa risoluzione (*Low Resolution, LR*), mentre la restante parte studia l'applicazione di queste tecniche al mondo video [9] che, sfruttando i dati di movimento e la dipendenza tra frame temporalmente contigui, è in grado di individuare ed estrapolare un maggior numero di informazioni ed allo stesso tempo offrire una qualità, una naturalezza ed un livello di dettaglio del dato generato sicuramente superiore agli approcci *SISR*.

Uno degli approcci ibridi più interessanti è denominato **Deep Learning-Based Video Coding (DLVC)** [10]. È stato proposto in MPEG e ha aggiunto due *DNN* al codificatore **VVC**: *una rete neurale convoluzionale come filtro* e una *rete convoluzionale di super resolution*.

DLVC ha mostrato un guadagno in codifica del *39,6%* rispetto ad **HEVC** e di circa il *10%* rispetto a **VVC**.

Il ***Politecnico di Torino***, nell'ambito di una collaborazione con il ***Centro Ricerche della Rai (CRITS)***, ha realizzato un algoritmo basato su reti neurali profonde con lo scopo di migliorare la predizione all'interno di un codificatore video, rendendo possibile ridurre il bitrate di codifica delle informazioni di residuo. Questo è stato possibile grazie ad una *rete convoluzionale CNN (Convolutional Neural Network)* che elabora le predizioni del frame corrente, combinandole con quelle dei frame precedentemente decodificati. In Fig. 2 viene rappresentato lo schema dell'architettura usata:

- la rete convoluzionale (*Conv.*) elabora sia le informazioni provenienti dal frame predetto (*MC Frame, Motion Compensated Frame*) creato a partire dai vettori di movimento del codificatore video tradizionale, sia due nuove predizioni (*Warped Frame*) ottenute a partire dai frame precedentemente decodificati (*Recon. Frame*) e opportunamente *deformati* in modo da assomigliare al frame corrente. Quest'ultima operazione, detta di *warping*, si rende possibile grazie al calcolo del campo vettoriale di movimento tra i frame decodificati e quello attuale (detto *optical flow*);
- questo processo produce una *stima multipla* del frame corrente e la successiva operazione di fusione (*Merging*) permette di sfruttare anche l'informazione contenuta nei frame passati, migliorando la qualità della previsione.

Tale approccio produce una riduzione del bitrate pari al *10%* confrontato con **HEVC** [11].

End-to-End coding

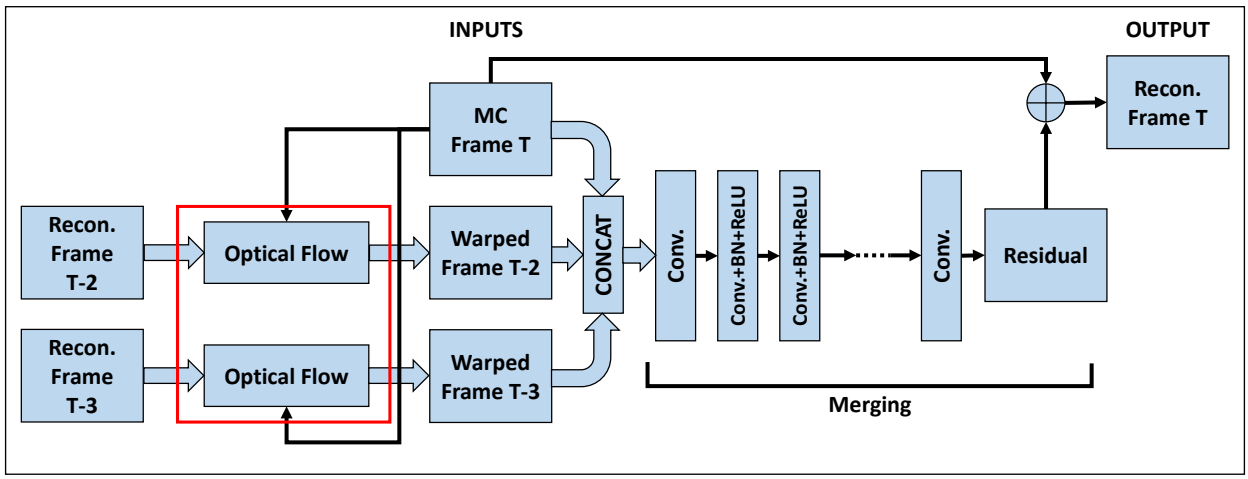
A differenza degli approcci tradizionali, non vi è un consenso comune sulle architetture di codifica basate sull'apprendimento di tipo *End-to-End (E2E)*.

I video codec tradizionali se spinti ad alti livelli di compressione, a causa della loro natura costruttiva, tendono a mostrare i cosiddetti *artefatti da blocchettizzazione*. Al contrario, gli approcci *E2E* riescono a limitare la presenza di artefatti ed in generale di rumore strutturato, se opportunamente addestrati ed ottimizzati.

Gran parte degli approcci *E2E* presenti in letteratura si basano sulla definizione di un framework al cui interno vengono riprodotti e ridefiniti in chiave *IA* tutti i blocchi di un classico schema di codifica video. L'obiettivo è quello di armonizzare questo insieme di componenti e dar vita ad una *black box* da ottimizzare in modalità end to end sfruttando le tecniche note dell'Intelligenza Artificiale.

In questo ambito vale la pena citare [12] che dichiara un miglioramento del *25.06%* rispetto ad **HEVC** e [13] con un notevole guadagno del *38.12%* sempre rispetto ad **HEVC**.

Fig. 2 – Architettura *Multi-frame Enhancement Network*



Un collo di bottiglia fondamentale all'interno di questi framework è rappresentato dalla generazione e compressione delle informazioni sul movimento (*optical flow*), essendo questo uno strumento molto efficace per la codifica video tradizionale, utilizzato per ridurre la ridondanza temporale nelle sequenze video. Per limitare la complessità degli algoritmi di motion estimation i sistemi di codifica tradizionali suddividono i frame in blocchi e da essi calcolano i vettori di movimento. Queste tecniche possono risultare inefficaci e causare un importante degrado della qualità visiva. Gli approcci basati su reti neurali, invece, sfruttano modelli di calcolo dei vettori di movimento a livello di singolo pixel risultando più efficaci ed accurati, sebbene più complessi (Fig. 3).

La letteratura scientifica mette a disposizione un'importante varietà di studi riguardanti questa tematica, ma, come spesso accade con le applicazioni basate sull'Intelligenza Artificiale, l'applicabilità può risultare limitata a causa della inadeguatezza dei dataset di addestramento disponibili. Molti di questi modelli vengono addestrati su dati sintetici, ovvero immagini generate artificialmente in computer grafica, risultando inappropriati e imprecisi una volta impiegati in contesti pratici su immagini *reali*. Spesso vengono poi aggiunti ulteriori blocchi per il refinement dei frame generati, come i filtri per il miglioramento della qualità visiva (*IQE – Image Quality Enhancement*) o i sistemi di *Super-Resolution (SR)* per la generazione di immagini ad alta risoluzione a partire da quelle a bassa risoluzione.

Fig. 3 – Esempio di optical flow

Ad oggi gli approcci *E2E* sono fortemente legati all'architettura del codificatore video tradizionale. Svincolarsi da tale architettura, in auge da più di trent'anni, è senza dubbio un traguardo ambizioso e probabilmente sarà necessario affrontare tale sfida passo dopo passo, cercando magari di incorporare quante più funzioni e funzionalità tipiche dei codec tradizionali all'interno di una singola struttura di rete. Così facendo, la reingegnerizzazione degli elementi e la semplificazione dell'architettura finale della rete renderanno meno complesso e dispendioso il processo di ottimizzazione e apprendimento. Non a caso, la ricerca si sta orientando verso la sperimentazione di architetture alternative, in cui starà alla rete imparare e gestire il trade-off tra bitrate e qualità finale (Fig. 4) [14].

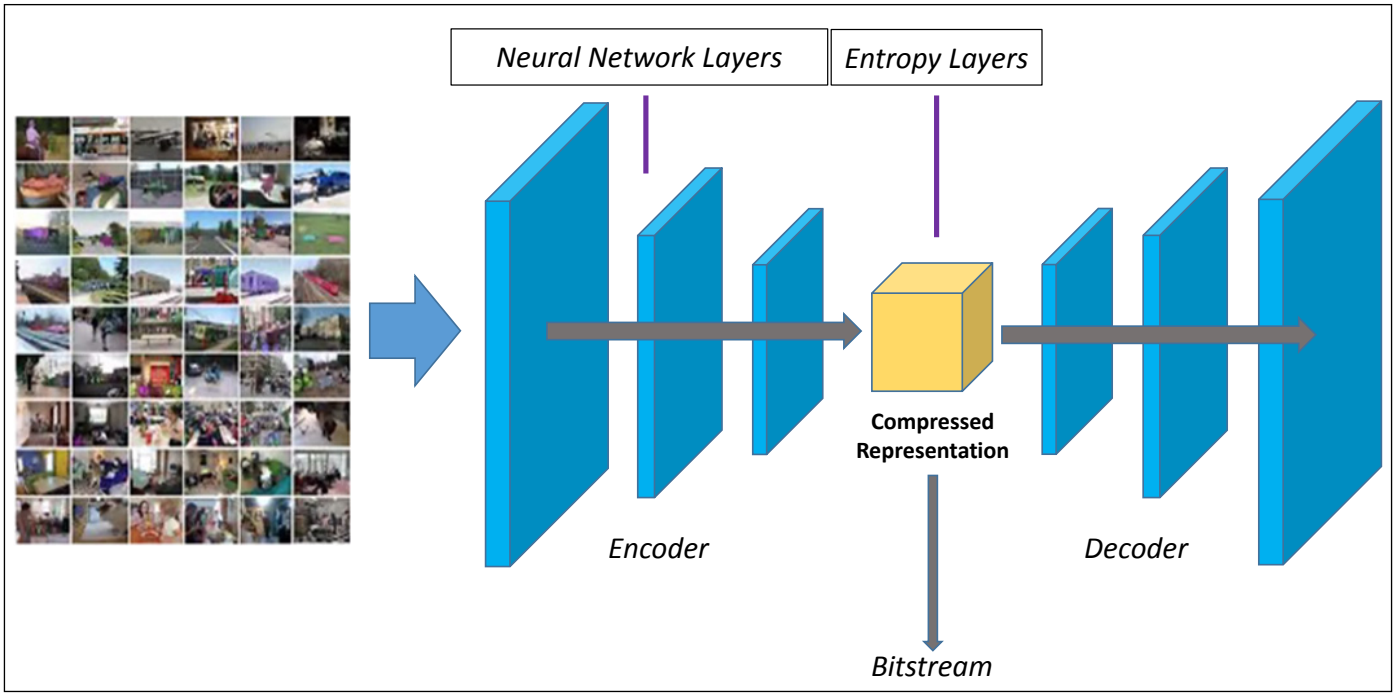
Analisi degli obiettivi comuni

Nei prossimi anni si assisterà alla discussione sulle seguenti questioni rilevanti per l'introduzione, con successo, di tool basati sull'Intelligenza Artificiale nel campo della codifica video: *dataset di addestramento e test*, *ricerca di metriche di valutazione della qualità*, *performance* e, infine, *complessità computazionale*.

Durante la fase di training, per ogni iterazione viene calcolata la *funzione di loss*. Questa è una funzione del tipo $R+\lambda D$ (dove R corrisponde al *bit-rate*, D alla *distorsione* e λ al trade-off tra le due grandezze) che dà una misura dell'*accuratezza dell'output rispetto ai dati in input*. L'obiettivo finale consiste nel minimizzare questa funzione di costo affinché si possa realizzare un modello affidabile ed efficace. La maggior parte delle reti neurali profonde usa la funzione *MSE (Mean Square Error)* che però non tiene conto del fatto che nella visione umana alcuni dettagli sono più rilevanti di altri e, quindi, una media sull'intera immagine non è in grado di stimare la qualità con sufficiente precisione. La scelta di una funzione più appropriata è ancora un problema aperto e la ricerca scientifica spazia dai *meccanismi di self attention*, alla *normalizzazione spettrale* [15].

I metodi oggettivi di valutazione della qualità come *PSNR (Peak Signal-to-Noise Ratio)* e *SSIM (Structural Similarity Index)* non sono adatti a codec basati su IA perché i risultati non sono ben correlati con le valutazioni soggettive. Nelle pubblicazioni accademiche e nelle attività di standardizzazione il *PSNR* rimane il *gold standard* per la valutazione delle prestazioni di codifica, ciononostante è opinione comune che il

Fig. 4 – Architettura *E2E*



PSNR non rifletta accuratamente la percezione visiva umana. Questa, infatti, è un sistema complesso e non lineare difficilmente riproducibile attraverso semplici espressioni matematiche e ciò richiede una continua ricerca di nuove e sempre più efficienti metriche, talvolta basate su approcci *learning-based*, come il caso del *VMAF (Video Multi-Method Assessment Fusion)* sviluppato dai ricercatori di ***Netflix*** [16]. Una delle metriche più interessanti e di ampio utilizzo è la *MS-SSIM (Multi-Scale Structural Similarity Index)* [17]. La filosofia di funzionamento si basa sul concetto di *informazione strutturale delle immagini*. Questa è a tutti gli effetti una evoluzione della *SSIM* e si distingue da quest'ultima per il semplice fatto di estendere il raggio di valutazione a più livelli dell'immagine (da qui *multi-scale*) fornendo così una maggiore flessibilità e capacità nel recepire le variazioni delle condizioni di visualizzazione delle immagini. Entrambe le metriche sfruttano la variazione della luminanza, del contrasto e della struttura dell'immagini per il calcolo delle similarità strutturale.

A causa della varietà di metriche a disposizione e della loro nota inadeguatezza per alcuni scopi, ricercatori e ingegneri sono costretti a corroborare le misurazioni *oggettive* con test visivi *soggettivi* per dar prova della bontà degli studi eseguiti. Sebbene questa metodologia di valutazione abbia funzionato per decenni, non è praticabile per una valutazione su larga scala, soprattutto se il dataset di test copre un'ampia gamma di contenuti (sport, documentari, film, …) e vari intervalli di qualità. Affinché la comunità che studia i codec video possa innovare più rapidamente ed in modo più accurato, è necessario utilizzare misurazioni automatizzate della qualità video riconosciute dall'intera comunità scientifica che riflettano il più possibile la percezione umana.

Uno degli elementi fondamentali per i tool basati sull'Intelligenza Artificiale è la disponibilità di *dataset*, cioè di collezioni di dati adeguati (in formato diverso in base al task di pertinenza) da utilizzare per le fasi di training, oltre che di valutazione finale del modello. Il problema attuale risiede nel fatto che tutti i contributi e gli studi in letteratura scientifica si basano su dataset eterogenei proposti dagli autori stessi, motivo per cui risulta difficile confrontare i risultati. Sarà necessario disporre di un dataset e di un benchmark *standard* a cui fare riferimento per poter valutare e confrontare in maniera universale la *performance* di ogni modello proposto.

Il concetto di *performance* può essere espresso attraverso due fattori: *qualità dell'output* (o *accuratezza*) e *tempo di esecuzione*. Quest'ultimo dipende dalla complessità computazionale del modello proposto che, in termini generali, è correlata alla tipologia e alla profondità della rete oltre che al numero dei parametri. Rispetto ai video codec tradizionali, che non richiedono grosse risorse computazionali e che vengono tipicamente implementati su chip o *DSP (Digital Signal Processor)* semplici e poco costosi, gli approcci basati su reti neurali necessitano di hardware specifico, come le *GPU (Graphic Processing Unit)* o gli *FPGA (Field Programmable Gate Array)*, in grado di offrire capacità di calcolo importanti ma di un costo più elevato. Questo rappresenta ad oggi l'ostacolo più importante che limita lo sviluppo e l'implementazione di questa importante e promettente tecnologia a bordo di dispositivi consumer.

Conclusioni

Le reti neurali si stanno dimostrando capaci di sostituire e migliorare molte delle componenti tecnologiche impiegate nel codificatore tradizionale [3], [4], [5], [18]. L'obiettivo finale della ricerca sui codec video è quello di migliorare l'efficienza di compressione mantenendo allo stesso tempo la complessità computazionale ad un livello ragionevole. Purtroppo, come accennato in precedenza, i tool basati su Intelligenza Artificiale mostrano risultati interessanti ma con impatti ancora troppo elevati in termini di complessità, in particolar modo sul decoder [19].

Da quanto detto è evidente che nel prossimo futuro sarà richiesto uno sforzo congiunto da parte della comunità scientifica e dell'industria affinché questa promettente tecnologia possa trovare largo impiego anche nel mondo della codifica video attraverso le architetture *E2E* o *ibride*.

Bibliografia

[1] [G. J. Sullivan ed altri, *Overview of the High Efficiency Video Coding (HEVC) Standard*, in "IEEE Transactions on Circuits and Systems for Video Technology", vol. 22, n. 12, 2012, pp. 1649-1668, **DOI**: 10.1109/TCSVT.2012.2221191

[2] Jianle Chen, Yan Ye e Seung Hwan Kim, *Algorithm description for Versatile Video Coding and Test Model 7 (VTM 7)*, Documento: JVET-P2002-v1, 2019, https://mpeg.chiariglione.org/standards/mpeg-i/versatile-video-coding/test-model-7-versatile-video-coding-vtm-7

[3] Zhao Wand ed altri, *Preliminary results of Neural Network Loop Filter*, ISO/IEC JTC1/SC29/WG1 doc. m54991, 2020

[4] J. Pfaff ed altri, *Intra prediction modes based on neural networks*, JVET-J0037-v1, 2018

[5] K. Fischer, C. Herglotz e A. Kaup, *On Versatile Video Coding at UHD with Machine-Learning-Based Super-Resolution*, in "2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)", 2020, **DOI**: 10.1109/QoMEX48832.2020.9123140

[6] C. Dong, C. C. Loy e X. Tang, *Accelerating the super resolution convolutional neural network*, in "Computer Vision – ECCV 2016", Springer International Publishing, 2016, **ISBN**: 978-3-319-46474-9

[7] B. Lim ed altri, *Enhanced deep residual networks for single image super-resolution*, in "2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)", 2017, **DOI**: 10.1109/CVPRW.2017.151

[8] X. Wang ed altri, *Esrgan: Enhanced super-resolution generative adversarial networks*, in L. Leal-Taixé e S. Roth (ed), "Computer Vision – ECCV 2018 Workshops", Springer International Publishing, 2018, pp. 63-69, **DOI**: 10.1007/978-3-030-11021-5_5

[9] Y. Jo ed altri, *Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation*, in "2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018)", 2018, **DOI**: 10.1109/CVPR.2018.00340

[10] *Deep Learning-Based Video Coding (DLVC)*, NELBITA DLVC page (web), http://dlvc.bitahub.com/ (ultimo accesso 31/12/2020)

[11] N. Prette ed altri, *Deep Multiframe Enhancement for Motion Prediction in Video Compression*, sottomesso a "2021 IEEE International Conference on Acoustics, Speech and Signal Processing", 2021, *ancora in fase di revisione*

[12] G. Lu ed altri, *An End-to-End Learning Framework for Video Compression*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", **DOI**: 10.1109/TPAMI.2020.2988453

[13] H. Liu ed altri, *Neural video compression using spatio-temporal priors*, 2019, arXiv:1902.07383

[14] A. Jacob ed altri, *Deep Learning Approach to Video Compression*, in "2019 IEEE Bombay Section Signature Conference (IBSSC)", 2019, **DOI**: 10.1109/IBSSC47189.2019.8973035

[15] C. Thomas, *Deep learning image enhancement insights on loss function engineering*, in "towards data science" (web), https://towardsdatascience.com/deep-learning-image-enhancement-insights-on-loss-function-engineering-f57ccbb585d7 (ultimo accesso 30/12/2020)

[16] Zhi Li ed altri, *Toward A Practical Perceptual Video Quality Metric*, in "The Netflix Tech Blog" (web), 2016, https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652 (ultimo accesso 30/12/2020)

[17] Z. Wang, E. P. Simoncelli e A. C. Bovik, *Multi-scale Structural Similarity for Image Quality Assessment*, in "The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003", 2003, **DOI**: 10.1109/ACSSC.2003.1292216

[18] Mingze Wang ed altri, *An Integrated CNN-based Post Processing Filter For Intra Frame in Versatile Video Coding*, in "2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)", 2019, **DOI**: *10.1109/APSIPAASC47483.2019.9023240*

[19] G. Sullivan e J-R. Ohm, *Meeting Report of the 14th Meeting of the Joint Video Experts Team (JVET), Geneva, CH, 19–27 March 2019"*, JVET-N_Notes_dD, 2019, *https://www.itu.int/wftp3/av-arch/jvet-site/2019_03_N_Geneva/JVET-N_Notes_dD.docx*

Sistemi di raccomandazione:

Intelligenza Artificiale, Deep Learning e personalizzazione dei contenuti

Paolo **Casagranda**, Sabino **Metta**
Rai - Centro Ricerche, Innovazione Tecnologica e Sperimentazione

I sistemi di raccomandazione *svolgono l'importante funzione di filtrare e personalizzare in modo automatico le informazioni permettendo in questa maniera di far fronte al sovraccarico informativo a cui ogni utente è quotidianamente esposto.*

Recentemente, la ricerca internazionale sta sperimentando l'utilizzo di tecnologie di apprendimento profondo*, anche conosciuto come* deep learning*, per accrescere le potenzialità dei sistemi tradizionali di raccomandazione.*

L'articolo presenta brevemente il funzionamento, la classificazione ed i limiti di quest'area pervasiva dell'Intelligenza Artificiale.

Anno 2020… al giorno d'oggi i *sistemi di raccomandazione*, i cosiddetti *Recommender Systems* (di seguito *RS*), non hanno più bisogno di presentazioni. Nel nostro precedente articolo [1], abbiamo già introdotto e descritto le tecnologie per raccomandare all'utente finale il contenuto per lei/lui più rilevante. Tali tecnologie, che afferiscono al dominio dell'*Intelligenza Artificiale* (*AI*), sono in continua evoluzione. Se da un lato permettono di ridurre il sovraccarico informativo a cui ogni utente è quotidianamente esposto, dall'altro lato tali tecnologie rappresentano, per il fornitore di contenuti (il cosiddetto *content provider*), lo strumento necessario per valorizzare la propria offerta e quindi difendere il proprio guadagno.

Dal "lontano" 2009, anno in cui **Netflix** ha assegnato un premio da *1 Milione di dollari* al miglior algoritmo in grado di predire le valutazioni degli utenti per i film, la comunità scientifica internazionale ha esplorato nuovi algoritmi e soluzioni per migliorare la previsione delle preferenze degli utenti.

201_02.5_3.00

Secondo la **IDC** (*International Data Corporation*), la prima società mondiale specializzata in ricerche di mercato, la spesa che ruota attorno alle tecnologie di *AI* raggiungerà entro il 2023 i *100 miliardi di dollari*.

Nel 2019 circa *2 miliardi di dollari* sono stati spesi nel dominio dei sistemi di raccomandazione. Tra i vari approcci, negli ultimi anni un fiorente filone di ricerca sta sperimentando l'utilizzo delle tecnologie dell'*apprendimento profondo* (*deep learning*) e delle *reti neurali profonde* (*deep neural networks*) per superare alcuni limiti e alcune criticità dei sistemi tradizionali di raccomandazione.

Sistemi di Raccomandazione tradizionali

In generale, i sistemi di raccomandazione rappresentano un'area dell'Intelligenza Artificiale e comprendono tecniche e strumenti software in grado di suggerire all'utente gli oggetti, e quindi anche i contenuti, più rilevanti tra i tanti a disposizione.

Tali sistemi non intendono necessariamente sostituirsi alla *raccomandazione classica* dove, per fare un esempio a noi più vicino, i suggerimenti vengono scelti e creati manualmente dall'editore del canale televisivo valendosi della competenza di professionisti esperti. In un *RS*, il contenuto più rilevante per un determinato utente viene inferito sulla base dei dati a disposizione ed attraverso l'implementazione di specifiche logiche. Per avere maggiori dettagli, raccomandiamo la lettura di un nostro precedente articolo [1] e, potendo approfondire, del classico di Ricci e altri [2].

Nella sua forma più semplice un sistema di raccomandazione produce, per ogni utente, una lista di oggetti e, per ciascuno di questi oggetti, una *valutazione dell'utilità* (*ranking*). Un oggetto può essere rappresentato da un qualunque contenuto multimediale e/o servizio digitale che possa essere in qualche modo *suggerito* all'utente finale: ad esempio, il prodotto di un supermercato, un video, un programma radiofonico, una canzone, una notizia.

I sistemi di raccomandazione possono essere classificati sulla base della specifica *logica* e dello specifico *algoritmo* attraverso i quali producono la lista finale di oggetti:

- *Collaborative Filtering* (*CF*). I suggerimenti per uno specifico utente vengono generati sulla base di giudizi o valutazioni (*ranking*) che altri utenti, in passato, hanno dato a determinati oggetti a disposizione.
- *Content-based Filtering* (*CB*). Il *RS* suggerisce ad uno specifico utente una lista di oggetti simili ad altri che l'utente ha scelto in passato. In questa categoria vengono confrontati gli oggetti stessi o le loro descrizioni. Rispetto alle metodologie *collaborative filtering*, i metodi *content-based* risultano essere computazionalmente più rapidi ed interpretabili, è cioè più facile capire il motivo di una specifica raccomandazione. Inoltre, gli algoritmi *CB* possono essere facilmente adattati a nuovi oggetti e/o utenti.
- *Demographic*. Il suggerimento del *RS* si basa sulla *caratterizzazione demografica dell'utente*, ad esempio l'età, il sesso, la residenza.
- *Knowledge-based*. I suggerimenti si basano su regole derivate da una conoscenza e descrizione esplicita dell'utente o degli oggetti (ad esempio, avendo conoscenza della mia preferenza verso la tecnologia, mi vengono consigliati articoli di quel tipo).
- *Community-based* o *Social*. I suggerimenti tengono conto delle scelte e delle valutazioni degli amici dell'utente sui *social networks* [2].
- *Hybrid*. Le tecniche descritte sopra vengono combinate per ottenere suggerimenti più precisi, cercando di far fronte alle debolezze di ciascun metodo. Molti *RS* commerciali sono di questo tipo.
- *Context-Aware* (*CARS*). I suggerimenti tengono in conto il *contesto* [3]. Il contesto è definito da svariati fattori quali stato emotivo, attività, posizione geografica, condizioni atmosferiche, e qualsiasi altra cosa concorra alla definizione dello stato dell'ambiente, dell'utente, dell'oggetto o dell'interazione tra essi.

Lo ricordiamo, tali logiche di raccomandazione sono automatiche ed implementate attraverso precisi algoritmi. Recentemente la ricerca internazionale ha intrapreso una promettente sperimentazione delle tecniche di *deep learning* per risolvere alcune criticità presenti negli approcci tradizionali.

Sistemi di raccomandazione basati sul deep learning

Negli ultimi tempi le *reti neurali* e il *deep learning* (*DL*) sono diventati tecnologie emergenti in grado di risolvere compiti molto complessi. Importanti risorse economiche sono state investite per studiare l'applicabilità di tali tecnologie a diversi ambiti e casi d'uso (medico, manifatturiero ecc.). Sono nati, così, i sistemi di raccomandazione basati sul deep learning, i cosiddetti *Deep Learning based Recommender Systems* (*DLRS*). Rispetto ai tradizionali *RS*, i sistemi di raccomandazione basati sul deep learning sono riusciti a raggiungere una notevole precisione e presentano sicuramente enormi potenzialità [3][4][5].

I sistemi di raccomandazione basati sul deep-learning presentano almeno quattro punti di forza rispetto ai sistemi tradizionali:

- *capacità di modellazione delle nonlinearità* eventualmente presenti nei dati attraverso funzioni di attivazione non lineari (rettificatore, sigmoide, tangente iperbolica, ecc.). In altre parole, una *modellazione nonlineare* è in grado di superare l'ipersemplificazione introdotta dai modelli lineari (tra questi la *fattorizzazione di matrice*). Questo permette di catturare in maniera più efficace le relazioni complesse ed intricate tra utenti (*users*) ed oggetti (*items*) nonché di individuare caratteristiche latenti di utenti e oggetti, vedi [4];
- *capacità di apprendimento* di caratteristiche rappresentative ed esplicative nei dati di input. Oggigiorno, la disponibilità di app e servizi in internet fornisce un'ampia base dati dove è possibile recuperare preziose informazioni relative ad utenti e/o ad oggetti (eventualmente raccomandabili). In tal senso, le tecnologie di *DL* possono essere impiegate per estrarre tali informazioni ed in questa maniera potenziare gli algoritmi di raccomandazione. Oltre ad alleviare una eventuale attività manuale di estrazione di informazioni, le tecnologie *DL* permettono di costruire algoritmi di raccomandazione in grado di combinare fonti eterogenee (testo, immagini, audio, video);
- *capacità di modellazione sequenziale*. Le tecnologie *DL* giocano un ruolo fondamentale nell'individuazione di strutture sequenziali presenti all'interno dei dati. Tra le varie applicazioni menzioniamo la *traduzione automatica* e la *comprensione* del *linguaggio naturale*. Per quanto riguarda i *RS*, la modellazione sequenziale è importante per intercettare la dinamica di comportamento dell'utente e l'evoluzione di cambiamento di determinati oggetti;
- *alta flessibilità*. Oggigiorno, è possibile accedere a numerose piattaforme di sviluppo di tecnologie *DL* (citiamo ad es. **TensorFlow**, **Keras**, **PyTorch**, ecc.). Tali piattaforme sono costruite in maniera modulare rendendo così maggiormente efficienti le attività di sviluppo ed integrazione di differenti modelli di raccomandazione.

Di seguito accenniamo brevemente ai *principali metodi di deep learning* che possono essere utilizzati nei sistemi di raccomandazione:

- *Autoencoder*. Il loro scopo è di replicare l'input sull'output passando attraverso una rappresentazione più semplice e possono essere impiegati per ridurre la dimensionalità e la complessità del problema o colmare i vuoti di una matrice sparsa.
- *Restricted Boltzmann Machines* (*RBM*). Le *Macchine di Boltzmann Ristrette* sono rappresentate da grafi non direzionali caratterizzati da un solo layer visibile ed uno nascosto. In ambito accademico tali tecnologie sono state utilizzate con successo per ottenere raccomandazioni di tipo *collaborative filtering*.
- *Recurrent Neural Network* (*RNN*). Le *Reti Neurali Ricorsive* trovano applicazione nella traduzione automatica o nel riconoscimento vocale, e in generale per modellare dati sequenziali dal

momento che elaborano sia un input sia uno stato dipendente dai precedenti dati. Le *RNN* sono state proposte per *RS location-based*, dipendenti cioè dalla sequenza di spostamenti dell'utente;
- *Convolutional Neural Network* (*CNN*). Le *Reti Neurali Convoluzionali* trovano, invece, vasta applicazione nel riconoscimento di immagini e, per questo, nei *RS* possono trovare applicazione nella ricerca di *features degli oggetti* (ad esempio descrizione con parole chiave di immagini da usare nel content-based filtering);
- *Deep Belief Network* (*DBN*). Le *Reti di Credenze Profonde* utilizzano *RBM* come componenti di base e sono state usate per estrarre features della musica, quindi possono trovare utilizzo nel content-based filtering.

In generale, è possibile costruire *RS* che integrino tecnologie classiche e deep learning, oppure *RS* basati interamente su deep learning. La parte di deep learning può, a sua volta, essere basata su un solo tipo di tecnologia in modo monolitico (*deep single model*), oppure integrare diverse tecnologie di deep learning che si completino a vicenda (*deep composite model*).

Nel seguito chiariamo intuitivamente l'utilizzo del *deep learning* nelle diverse classi di sistemi di raccomandazione:

- *Content-Based Filtering*: il *DL* permette di descrivere gli item da raccomandare estraendo caratteristiche salienti e, anche, di catturare relazioni non lineari tra utenti ed items.
- *Collaborative Filtering*: alcuni ricercatori hanno proposto metodi basati sul *DL*, in particolare per far fronte al problema della sparsità della matrice o del *cold start* nel collaborative-filtering. Ad esempio, gli *RNN* sono stati usati per dimostrare la capacità di ottenere raccomandazioni basate non solo su una preferenza dell'utente, ma anche sulla sequenza temporale delle sue preferenze.
- *Hybrid*: viene utilizzata la capacità del *DL* di estrarre caratteristiche salienti dagli item o dagli utenti e potenziare il collaborative-filtering o il content-based filtering alla base dell'hybrid recommender.
- *Context-aware*: i *context-aware recommenders* (*CARS*) hanno avuto recentemente un altissimo interesse per la capacità di incorporare il contesto nelle raccomandazioni. Il *DL* può essere utilizzato per estrarre gli elementi latenti del contesto e usarli per migliorare la raccomandazione.
- *Community based/Social Recommender Systems*: utilizzano la rete sociale degli utenti per migliorare le raccomandazioni, ad esempio le relazioni di fiducia, le reazioni degli utenti e la loro collocazione spazio-temporale. Il *DL* può essere inserito per modellare queste caratteristiche.

Un esempio di *DLRS* molto promettenti riguarda i raccomandatori fra domini diversi (*cross-domain*), in particolare quelli che utilizzano la conoscenza di un dominio per riversarla su un altro, che possono beneficiare della capacità dei metodi di deep learning.

Attualmente la ricerca sui *DLRS* sta cercando di risolvere alcuni problemi. La *scalabilità* del *DLRS* è importante per le applicazioni pratiche: si cerca di ridurre il numero di parametri oppure di comprimere la complessità dei dati. Un altro aspetto critico nel deep learning in generale, ma importante in special modo per i media pubblici, è la *trasparenza dell'algoritmo*, cioè la possibilità di spiegare le ragioni per un certo suggerimento: la ricerca sta lavorando anche su questo.

Oltre alle potenzialità di questi metodi, è necessario tenere in conto che il vantaggio pratico nell'impiego dei *DLRS* è tutt'altro che chiaro e che i risultati della ricerca sono a volte difficilmente riutilizzabili. Nei casi analizzati da un articolo molto discusso presentato alla *Conferenza RecSys 2019* [6], i risultati ottenuti con *DL* spesso non sono ripetibili, cioè non è immediato ricostruire l'esperimento con risultati simili: meno della metà è risultata ripetibile ed inoltre, nei casi rimanenti, il problema poteva essere risolto meglio con algoritmi classici e computazionalmente meno complessi. Le potenzialità dei *DLRS* sono grandi ma la ricerca dovrà lavorare ancora per coglierle appieno.

Conclusioni

L'articolo introduce il concetto di *sistema di raccomandazione* (*RS*), mostrando le ragioni della sua pervasività. Ad una introduzione al funzionamento e alla classificazione dei *RS* segue una panoramica dei più recenti sviluppi nell'applicazione delle *reti neurali* e del *deep learning* (*DL*) ai sistemi di raccomandazione. Mentre i risultati sono promettenti e in futuro le tecnologie di *DL* potrebbero risultare decisive per accrescere l'efficacia o i campi applicativi delle raccomandazioni, alcuni problemi sono tutt'ora aperti e oggetto di ricerca e di un vivace dibattito.

Bibliografia

[1] P. Casagranda e S. Metta, *Leggi questo articolo, una tua amica lo ha trovato interessante*, in "Elettronica e Telecomunicazioni", Anno LXV, n. 2/2016, 2016, pp. 27-37, http://www.crit.rai.it/eletel/2016-2/162-4.pdf

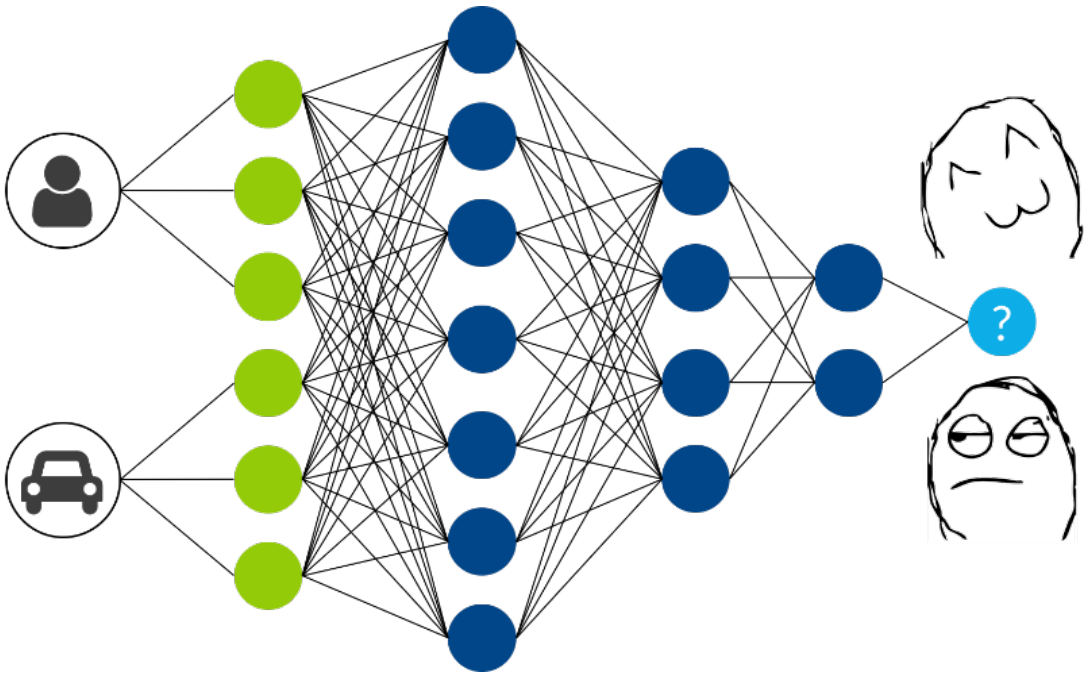
[2] F. Ricci, L. Rokach e B. Shapira, *Recommender systems: introduction and challenges*, in F. Ricci, L. Rokach e B. Shapira (ed) "Recommender Systems Handbook", Springer US, 2015, pp. 1-34, **DOI**: 10.1007/978-1-4899-7637-6_1

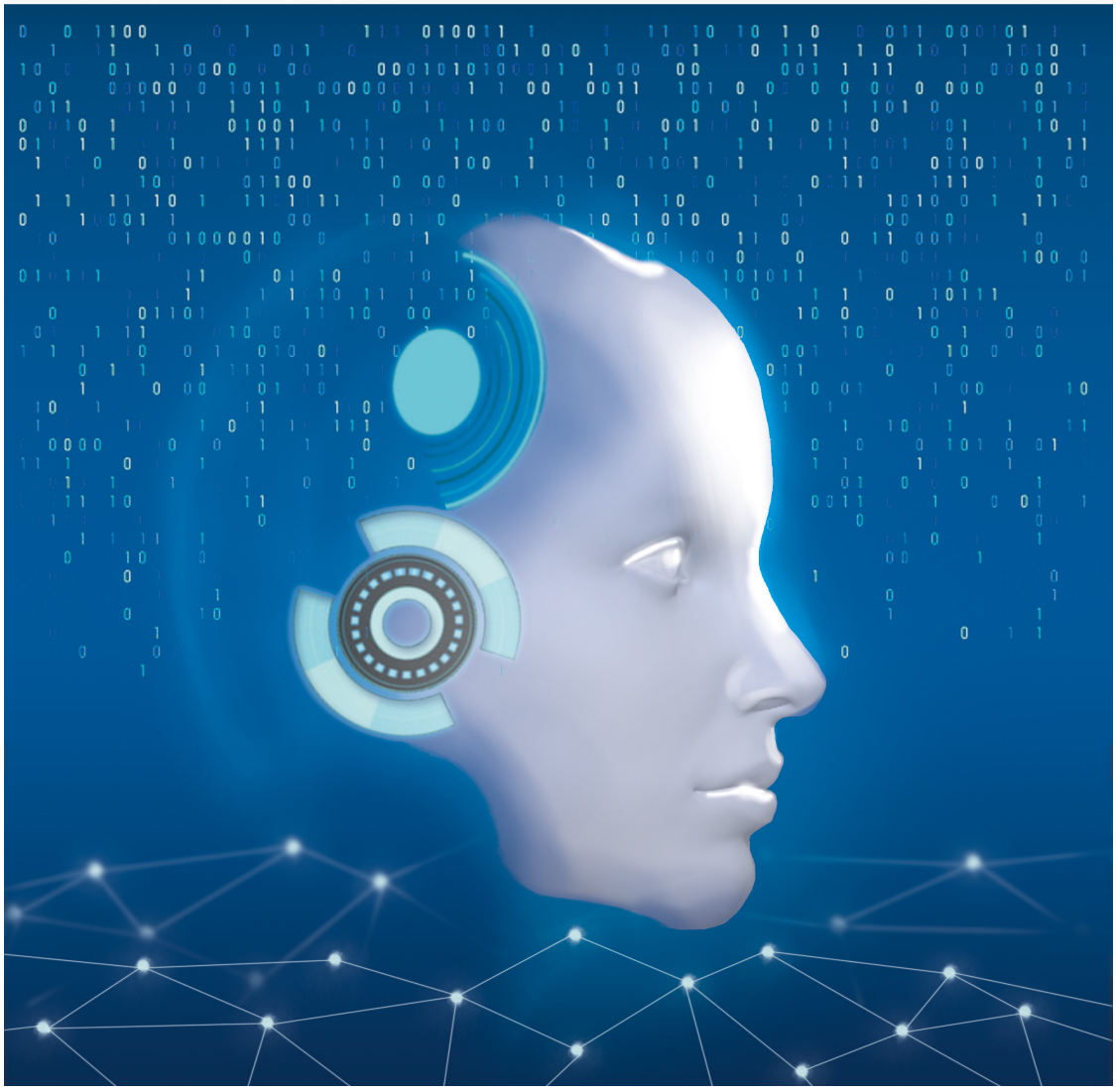
[3] G. Adomavicius ed altri, *Context-Aware Recommender Systems*, in "AI Magazine", vol. 32, n. 3, 2011, pp. 67-80, **DOI**: 10.1609/aimag.v32i3.2364

[4] R. Mu, *A survey of recommender systems based on deep learning*, in "IEEE Access", vol. 6, 2018, pp. 69009-69022, **DOI**: 10.1109/ACCESS.2018.2880197

[5] Z. Batmaz ed altri, *A review on deep learning for recommender systems: challenges and remedies*, in "Artificial Intelligence Review", vol. 52, n. 1, 2018, pp. 1-37, **DOI**: 10.1007/s10462-018-9654-y

[6] M. Ferrari Dacrema, P. Cremonesi e D. Jannach, *Are we really making much progress? A worrying analysis of recent neural recommendation approaches*, in "RecSys '19: Proceedings of the 13th ACM Conference on Recommender Systems", 2019, pp. 101-109, DOI: 10.1145/3298689.3347058





Assistenti vocali:

l'Intelligenza Artificiale a portata di voce

Paolo **Casagranda**, Francesco **Russo**, Raffaele **Teraoni Prioletti**
Rai - Centro Ricerche, Innovazione Tecnologica e Sperimentazione

In questo articolo viene introdotto il concetto di assistente digitale a controllo vocale, chiarendo la situazione del mercato ed i principi generali di funzionamento.

Un assistente vocale è un programma che, opportunamente addestrato, può dialogare con interlocutori umani grazie alla capacità di riconoscere, sintetizzare ed elaborare il linguaggio naturale dei comandi vocali. Gli assistenti vocali intelligenti sono ormai pervasivi ed estremamente rilevanti per la radio, e permettono un utilizzo immediato dell'Intelligenza Artificiale sotto diversi aspetti: riconoscimento e sintesi vocale, riconoscimento della richiesta, attuazione della risposta.

Per meglio comprenderne le applicazioni, al CRITS è stato costruito un prototipo per servizi radiofonici evoluti che è stato valutato anche da un gruppo di utenti. Lo studio ha confermato la rilevanza degli assistenti vocali per la radio, trovandone applicazioni possibili e accennando alcuni fondamentali requisiti sui contenuti.

Gli *assistenti vocali* sono tra i protagonisti della recente evoluzione tecnologica e stanno raggiungendo rapidamente molti ambiti della nostra vita. Un assistente vocale è un programma che, opportunamente addestrato, può dialogare con interlocutori umani grazie alla capacità di riconoscere, sintetizzare ed elaborare il linguaggio naturale dei comandi vocali. Spesso l'assistente vocale è identificato con lo *smart speaker*, il dispositivo fisico che più comunemente lo ospita. Grazie al supporto dell'interfaccia vocale, il mercato degli smart speaker è in rapida crescita e anche le più recenti analisi di mercato condotte negli USA all'inizio del 2020 confermano questo trend [1]. Basti pensare che negli USA il 24% degli americani dai 18 anni in su (60 milioni circa) possiede già uno smart speaker e il 54% ha usato qualche tipo di tecnologia a comando vocale, come gli assistenti vocali su smartphone, smart speaker o altri dispositivi. A partire dal 2014, anno in cui **Amazon** ha introdotto il primo dispositivo, molte altre aziende sono entrate nel mercato con i propri smart speaker e il tasso di adozione di questa tecnologia è risultato più veloce rispetto a qualsiasi altro dispositivo di consumo [2].

Al di là dell'utilizzo degli smart speaker come semplici altoparlanti, la funzionalità principale che questi dispositivi forniscono è l'*Intelligenza Artificiale* (IA). È infatti grazie alle tecnologie di IA che l'assistente vocale costruisce prima un modello per la comprensione del linguaggio parlato e dopo elabora una risposta adeguata, rendendo possibile il dialogo tra utente e macchina.

201_026_3.00

È proprio per evidenziare questa capacità che gli assistenti vocali vengono anche chiamati *assistenti vocali intelligenti*. I dati mostrano, inoltre, che assistenti vocali e smart speaker stanno diventando un veicolo importante per la fruizione della radio e dell'audio in genere, in aggiunta e spesso in sostituzione alle vecchie abitudini di ascolto. Questo scenario li rende estremamente rilevanti per la radiofonia creando sia opportunità che sfide per il mondo dei broadcaster radiofonici.

L'articolo è organizzato nel modo seguente: ad una panoramica dell'attuale offerta di mercato e dei principali utilizzi degli assistenti vocali, seguirà un'introduzione tecnica al loro funzionamento. L'analisi si concluderà con la descrizione delle funzionalità più rilevanti di un prototipo implementato al **CRITS Rai** per valutare le potenzialità di questa tecnologia in ambito radiofonico.

Panoramica dell'offerta di mercato

Comunemente l'assistente vocale è integrato in dispositivi che possono essere prodotti anche da aziende diverse da quelle dello stesso assistente vocale, come smart speaker, telefoni, tablet, PC, TV nonché negli abitacoli di alcune automobili.

Le maggiori aziende che hanno implementato un assistente vocale sono **Amazon** con *Alexa*, **Google** con *Google Assistant*, **Baidu** con *Duer*, **Alibaba** con *Genie*, **Xiaomi** con *Xiao*, **Apple** con *Siri*, **Samsung** con *Bixby*, **Microsoft** con *Cortana* e **Huawei** con *Celia*.

Oltre ai prodotti commerciali sono disponibili progetti open source per l'implementazione degli assistenti vocali. Le tecnologie open source offrono opzioni flessibili a startup e sviluppatori per sperimentare e costruire prodotti orientati al proprio settore di interesse. *Mycroft*, *Kalliope*, *Open Assistant*, *Jasper* e *Leon* sono esempi di assistenti vocali open source.

L'aumento delle vendite di smart speaker degli ultimi anni è da attribuire principalmente al successo degli assistenti vocali, di cui gli smart speaker rappresentano un comodo strumento di accesso. Da alcuni mesi è iniziata anche la vendita di dispositivi con display, i cosiddetti *smart display*, che rendono la risposta dell'assistente vocale più completa grazie al supporto dello schermo (es. *Amazon Echo Show* e *Google Nest Hub* (Fig. 1)) e, in alcuni casi, permettono di effettuare videochiamate.

Col passare del tempo sempre più brand ed aziende hanno lanciato la propria proposta di dispositivi smart con integrazione di uno degli assistenti vocali esistenti, contribuendo a diversificare il mercato. Attualmente, aziende come **JBL** e **Sony** hanno indirizzato la propria scelta verso *Google Assistant* mentre **Yamaha**, **Ultimate Ears** e **Harmand Kardon** hanno optato per *Alexa*. **Marshall**, **Bose** e **Polk** hanno scelto di produrre dispositivi con *Google Assistant* o con *Alexa* mentre **Sonos** offre la possibilità di utilizzare più assistenti vocali nello stesso dispositivo.

Secondo il report di **Canalys** del febbraio 2020 [3], il mercato globale ha avuto una crescita del *52%* nel Q4 del 2019 con *49,2 milioni* di unità vendute. I cinque maggiori produttori di assistenti vocali digitali mondiali, ordinati per numero di unità vendute nel Q4 del 2019, sono stati **Amazon** con *15,6 milioni* di vendite, **Google** con *12,5*, **Baidu** con *5,7*, **Alibaba** con *5,6* e **Xiaomi** con *4,6*. **Apple** si è fermata su un numero di vendite inferiore, in parte dovuto alla differente clientela di riferimento a cui punta l'azienda.

Fig. 1 – *Rai Radio 2* su *Google Nest Hub*

Fig. 2 – Vendite globali di smart speaker e crescita annuale nel 2019 secondo *Canalys* *(fonte [3])*

Spedizioni e crescita annuale degli smart speaker in tutto il mondo
Impulso del mercato degli smart speaker per Canalys: 2019

Venditore	Spedizioni 2019 (milioni)	Quota mercato 2019	Spedizioni 2018 (milioni)	Quota mercato 2018	Crescita annuale
Amazon	37.3	29.9%	24.2	31.1%	+54%
Google	23.8	19.1%	23.4	30.0%	+2%
Baidu	17.3	13.9%	3.6	4.6%	+384%
Alibaba	16.8	13.5%	8.9	11.4%	+89%
Xiaomi	14.1	11.3%	7.1	9.1%	+97%
Altri	15.4	12.3%	10.8	13.8%	+43%
Totale	124.6	100.0%	78.0	100.0%	+60%

Nota: le percentuali potrebbero non arrivare al 100% a causa dell'arrotondamento
Fonte: Analisi degli smart speaker di Canalys (spedizioni), Febbraio 2020

Per l'intero anno 2019 sono stati venduti un totale di circa *125 milioni* di smart speaker con un aumento del *60%* rispetto al 2018 (Fig. 2). Le vendite in Cina sono più che raddoppiate in un anno grazie a **Baidu**, **Alibaba** e **Xiaomi**. In particolare, **Baidu** ha ottenuto un'impressionante crescita nelle vendite passando da *3,6 milioni* nel 2018 a *17,3 milioni* nel 2019.

Alcune aziende hanno adattato gli assistenti vocali per l'utilizzo in contesti specifici, ad esempio alberghi, settore ospedaliero e automotive. Nel settore ricettivo, ad esempio, si stanno implementando attività per migliorare i servizi di concierge, facilitare la riproduzione di musica, il controllo della temperatura o l'illuminazione della camera, la ricerca di servizi locali e persino il check-out (ad es. **Amazon** e **Alibaba**) [4]. Un altro settore da tenere in considerazione è quello ospedaliero dove gli assistenti vocali possono essere utili a pazienti, infermieri e medici. Un rapporto di **IHS** (*Information Handling Services*) descrive l'uso di smart speaker negli ospedali: i comandi vocali possono essere impiegati per controllare televisori e apparati posti nelle stanze dei pazienti, nonché per inoltrare richieste verso i dispositivi mobili utilizzati da medici e infermieri. Un esempio di tale uso è una piattaforma basata su *Alexa* che da febbraio 2019 è impiegata in un progetto pilota presso l'ospedale **Cedars-Sinai Medical Center** di West Hollywood, in California [4]. Inoltre, si ritiene che gli smart speaker possano essere di grande aiuto a persone ipovedenti (almeno 2,2 miliardi di individui a livello globale secondo l'**Organizzazione Mondiale della Sanità**) e agli anziani, rendendo molto più semplice l'interazione con la tecnologia.

L'ultimo settore che citiamo è l'automotive, che sta manifestando un grandissimo interesse nell'implementazione dell'assistente vocale all'interno dei veicoli. *Alexa*, l'assistente vocale di **Amazon**, è già presente nell'abitacolo di alcuni modelli di auto, tra cui Toyota, Audi, Ford [5]. General Motors introdurrà *Android Automotive* di **Google** nei suoi veicoli dal 2021 come già fatto dall'alleanza Renault-Nissan-Mitsubishi, da Fiat Chrysler Automobiles e da Volvo [6]. Attualmente i principali produttori di automobili offrono modelli che già supportano *Apple CarPlay* o hanno in programma di introdurlo [7]. **Alibaba** ha annunciato di aver siglato un accordo, per i veicoli destinati al mercato cinese, con Audi, Renault, Volkswagen e Honda [8]. **Baidu** ha firmato un accordo con Ford, BMW e Volkswagen che entrano a far parte di un consorzio di più di 130 aziende globali per collaborare ad *Apollo*, la piattaforma di tecnologie di *guida autonoma open source* [9]. La piattaforma *Houndify*, fornita da **SoundHound**, è stata utilizzata per lo sviluppo di un proprio assistente vocale da Mercedes, Kia, Honda, Hunday e PSA Group [10]. Infine, altre case automobilistiche come General Motors, FCA, Toyota, Ford, Audi e BMW utilizzeranno, su alcuni veicoli, la tecnologia di **Cerence** che ha da poco presentato il *Cognitive Arbitrator*, un sistema sviluppato per consentire l'utilizzo di più assistenti vocali all'interno dell'auto [11].

Come funziona un assistente vocale

La grande sfida dell'assistenza vocale è l'utilizzo del linguaggio naturale per l'accesso a differenti servizi, tra cui l'ascolto della radio in streaming. L'elaborazione di una richiesta utente prevede l'esecuzione di due fasi logicamente distinte: la *gestione dell'input dell'utente* e la *produzione di un risultato* da restituire come risposta. Di seguito descriveremo l'architettura di funzionamento generale degli assistenti vocali facendo riferimento alle due piattaforme utilizzate per i test: *Google* e *Amazon*.

Il successo di un'applicazione, qualunque sia il dispositivo su cui deve funzionare, è fortemente influenzato dalla sua facilità di utilizzo. Risulta fondamentale, quindi, riuscire a creare un'interfaccia vocale che sia semplice e intuitiva. I comandi vocali, prima di poter essere eseguiti, devono essere compresi e interpretati. La gestione dell'input dell'utente è sicuramente uno dei processi più complessi a causa della varietà e imprevedibilità con cui è possibile fare le richieste. L'uso dell'Intelligenza Artificiale per assolvere a tale scopo risulta fondamentale. I produttori di assistenti vocali mettono a disposizione degli sviluppatori gli strumenti utili per creare applicazioni specifiche di terze parti per il proprio assistente vocale (es. *Google Dialogflow* e *Amazon Alexa Skill Kit*). Tutta la logica usata per l'elaborazione delle richieste si trova, in genere, nel *cloud* proprietario.

In **Google** il fulcro di un'applicazione (*Action*) è la costruzione del *Dialogflow agent*, un agente virtuale incaricato di gestire le conversazioni con l'utente finale e responsabile del riconoscimento della richiesta dell'utente. Per espletare la sua funzione, l'agente virtuale utilizza un modello di comprensione del linguaggio naturale costruito con l'ausilio di avanzate tecniche di *Machine Learning* (**ML**) e *Natural Language Understanding* (**NLU**). La prima operazione che viene eseguita su una nuova richiesta è la trascrizione del parlato in testo attraverso tecniche di *riconoscimento vocale automatico* (**ASR**). L'agente virtuale applica, quindi, il modello costruito al testo della richiesta con lo scopo di comprendere le espressioni dell'utente, associarle agli *intent* (azioni che soddisfano la richiesta dell'utente), estrarre i parametri utili. L'addestramento dell'agente virtuale è fondamentale nella definizione di una *Action*. L'addestramento avviene attraverso un insieme di frasi di training scelte dal progettista che devono essere di qualità e quantità sufficienti da permettere la costruzione di un modello efficace per il riconoscimento dell'intent. Quando un'espressione dell'utente assomiglia ad una delle frasi di training, il corrispondente intent viene innescato. Non è necessario definire ogni possibile esempio poiché le tecniche di machine learning si occuperanno di espandere il modello con frasi simili a quelle di training. Il flusso della conversazione deve essere progettato in modo che la Action sia sempre in grado di interpretare la richiesta.

Anche **Amazon**, l'altro produttore che abbiamo utilizzato per i nostri test, per gestire l'input dell'utente opera con un'architettura del tutto simile a quella appena descritta per Google e che ragionevolmente accomuna anche agli altri principali produttori di assistenti vocali (Fig. 3).

Come già anticipato, ogni volta che l'utente inoltra una richiesta all'assistente virtuale viene attivato il corrispondente *intent* per l'elaborazione della risposta. Nel caso più semplice la risposta che viene restituita all'utente è predefinita e statica ma, più frequentemente, è frutto di un ragionamento più complesso.

La Radio e gli Assistenti Vocali

L'impiego di moderne tecniche di IA permette di effettuare complesse analisi sui dati per rendere le risposte dell'assistente vocale accattivanti e utili per l'utente. Al **CRITS Rai** è stato progettato ed implementato un prototipo che permette un accesso evoluto ai contenuti di *Radio Rai* attraverso gli assistenti vocali. Le piattaforme utilizzate nel corso della sperimentazione sono state *Google* e *Amazon* e, attraverso una user evaluation condotta su più di 130 volontari, sono emersi chiari apprezzamenti verso alcuni dei servizi esposti con il prototipo, che brevemente descriviamo qui di seguito [12].

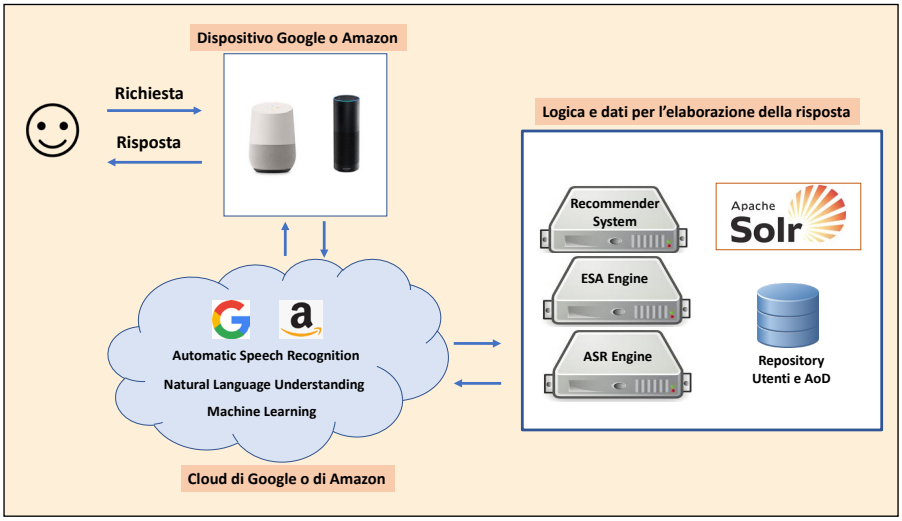


Fig. 3 – Schema di funzionamento del prototipo del CRITS Rai

Ricerca con parola chiave

Questa funzionalità permette la ricerca di parole all'interno del contenuto audio (nel seguito **AoD**, Audio on Demand*). I contenuti trovati vengono ordinati per score decrescente e proposti, uno alla volta, all'ascolto dell'utente che può, a propria discrezione, saltare da un contenuto all'altro. Alla base di questa funzionalità c'è* Apache Solr*, una potente piattaforma open source di ricerca e indicizzazione del testo.*

Ricerca per similarità semantica

Per ogni AoD che l'utente ascolta, c'è la possibilità di chiedere all'assistente vocale una lista di contenuti simili ad esso. Se l'AoD proposto è apprezzato dall'ascoltatore, risulta immediato poterne chiedere altri che potrebbero incontrare il suo interesse.

Contenuti raccomandati

Permette di iniziare l'ascolto di AoD raccomandati. L'utilizzo di questa funzionalità presuppone la presenza di un sistema di raccomandazione che crea una lista di AoD suggeriti e ordinati per score decrescente.

Le funzionalità appena descritte sono state progettate con l'idea di avvicinare il modello di radio tradizionale a forme di fruizione più recenti e di successo, che fanno della flessibilità e della personalizzazione caratteristiche imprescindibili. Anche l'uso di *contenuti atomizzati*, cioè di contenuti suddivisi in atomi semanticamente coerenti, ha contribuito fortemente a migliorare la resa del prototipo grazie alla possibilità di raggiungere direttamente i segmenti di audio rilevanti per l'utente.

Alla base dei servizi esposti dal prototipo c'è l'impiego di differenti tecniche di IA. Solo grazie alla conoscenza profonda degli AoD e delle preferenze dell'utente è possibile creare associazioni rilevanti tra AoD e utente. Nel prototipo, la similarità tra contenuti è calcolata attraverso una comparazione semantica realizzata con tecniche di *Explicit Semantic Analisys* (**ESA**) che utilizzano l'intelligenza artificiale per costruire un interprete semantico che mappa un testo in una sequenza ponderata di concetti [12].

Conclusioni

In questo articolo abbiamo introdotto i concetti di *assistente digitale a controllo vocale* e di *smart speaker*, uno dei dispositivi che più comunemente lo integrano. Gli assistenti intelligenti sono ormai pervasivi ed estremamente rilevanti per la radio, e permettono un utilizzo immediato dell'IA sotto diversi aspetti: riconoscimento e sintesi vocale, riconoscimento della richiesta, attuazione della risposta. Per meglio comprenderne le applicazioni, è stato costruito un prototipo per servizi radiofonici evoluti, valutato da un gruppo di utenti. Lo studio ha confermato la rilevanza degli assistenti vocali per la radio, trovandone applicazioni possibili e individuando alcuni fondamentali requisiti sui contenuti.

BIBLIOGRAFIA

[1] NPR e Edison Research, *The Smart Audio Report*, Edison Research (web), 30/04/2020, https://www.edison-research.com/the-smart-audio-report-2020-from-npr-and-edison-research/ (ultimo accesso 08/10/2020)

[2] *Activate Technology & Media Outlook 2020*, Activate Consulting (web), https://activate.com/outlook/2020/ (ultimo accesso 08/10/2020)

[3] *Canalys: Global smart speaker market to grow 13% in 2020 despite coronavirus disruption*, Canalys (web), 27/02/2020, https://www.canalys.com/static/press_release/2020/pr20200227.pdf (ultimo accesso 08/10/2020)

[4] P. Bajpai, *An Overview Of The Smart Speaker Market*, Nasdaq (web), 20/12/2019, https://www.nasdaq.com/articles/an-overview-of-the-smart-speaker-market-2019-12-20 (ultimo accesso 08/10/2020)

[5] *Vehicles with Alexa*, Amazon (web), https://www.amazon.com/b?node=17744356011&ref=ALEXA_AUTO_VEHICLES (ultimo accesso 08/10/2020)

[6] U. Lawrence, *Android Automotive OS provides the smarts for new Polestar 2 electric sedan*, IEEE Spectrum (web), 04/03/2020, https://spectrum.ieee.org/cars-that-think/transportation/advanced-cars/android-automotive-os-news-polestar-2-electric-sedan (ultimo accesso 08/10/2020)

[7] *CarPlay - Available Models*, Apple (web), https://www.apple.com/ios/carplay/available-models/ (ultimo accesso 08/10/2020)

[8] B. Kinsella, *Alibaba extends Tmall Genie with a new in-car smart speaker pertnership with automakers Audi, Honda, and Renault*, Voicebot.AI (web), 16/06/2019, https://voicebot.ai/2019/06/16/alibaba-extends-tmall-genie-with-a-new-in-car-smart-speaker-partnership-with-automakers-audi-honda-and-renault/ (ultimo accesso 08/10/2020)

[9] Reuters Staff, *VW taps Baidu's Apollo platform to develop self-driving cars in China*, Reuters (web), 02/11/2018, https://www.reuters.com/article/us-volkswagen-autonomous/vw-taps-baidus-apollo-platform-to-develop-self-driving-cars-in-china-idUSKCN1N71J1 (ultimo accesso 08/10/2020)

[10] *Powered By Houndify*, Houndify (web), https://www.houndify.com/powered-by-houndify (ultimo accesso 08/10/2020)

[11] E. H. Schwartz, *FCA will use Cerence voice recognition tech in all vehicles*, Voicebot.AI (web), 20/03/2020. https://voicebot.ai/2020/03/19/fiat-chrysler-will-use-cerence-voice-recognition-tech-in-all-vehicles/ (ultimo accesso 08/10/2020)

[12] P. Casagranda, F. Russo, R. Teraoni Prioletti, *Evolution of Radio Services in the era of Voice-Controlled Digital Assistants*, documento presentato a "IBC 2019 Conference", settembre 2019, Amsterdam, https://www.researchgate.net/publication/335928463_Evolution_of_Radio_Services_in_the_Era_of_Voice-Controlled_Digital_Assistants

Rai CRITS