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

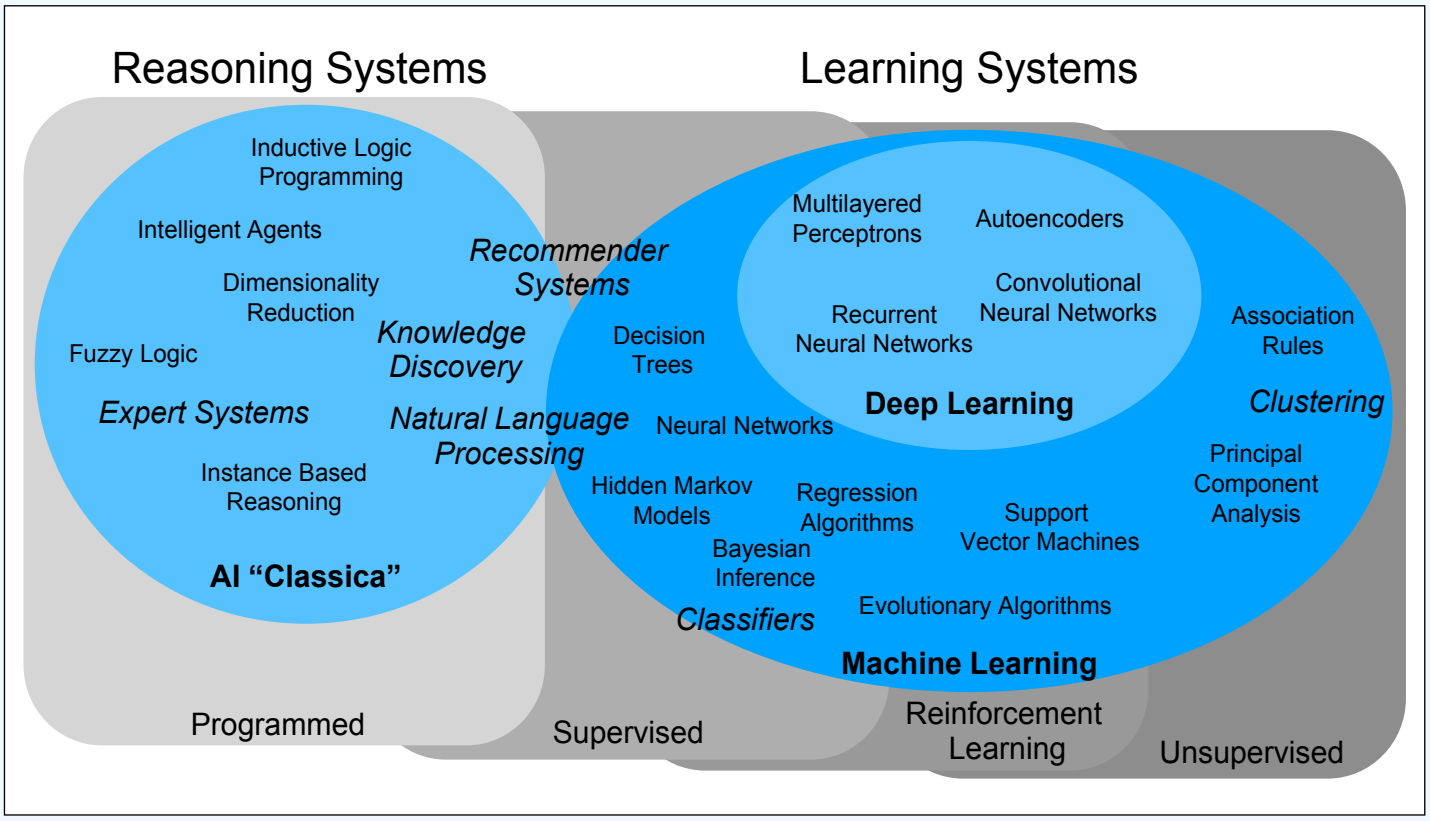


Fig. 1 – Mappa dei metodi utilizzati dall'Intelligenza Artificiale

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*.

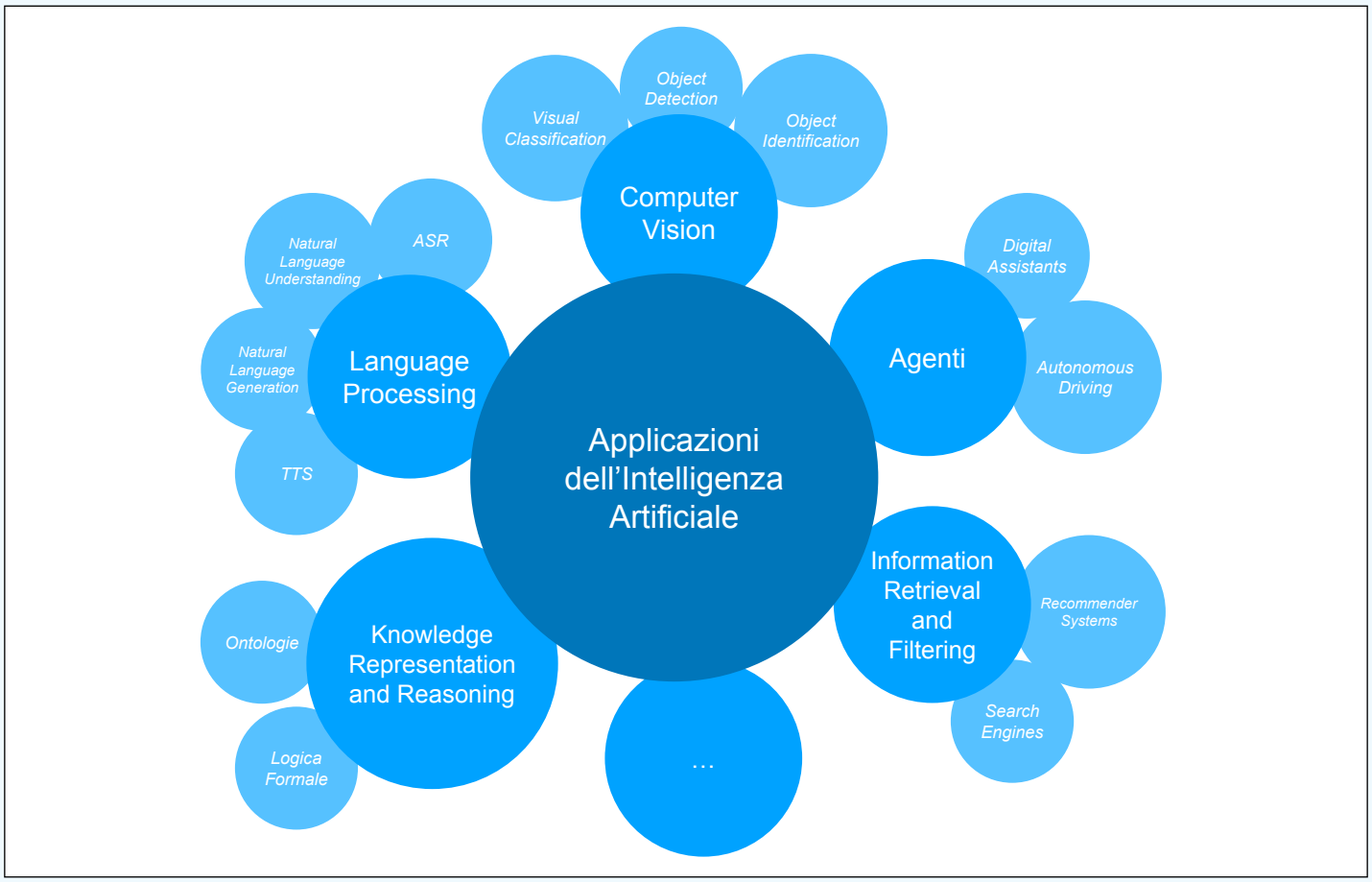


Fig. 2 – Mappa dei domini applicativi dell'Intelligenza Artificiale

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.

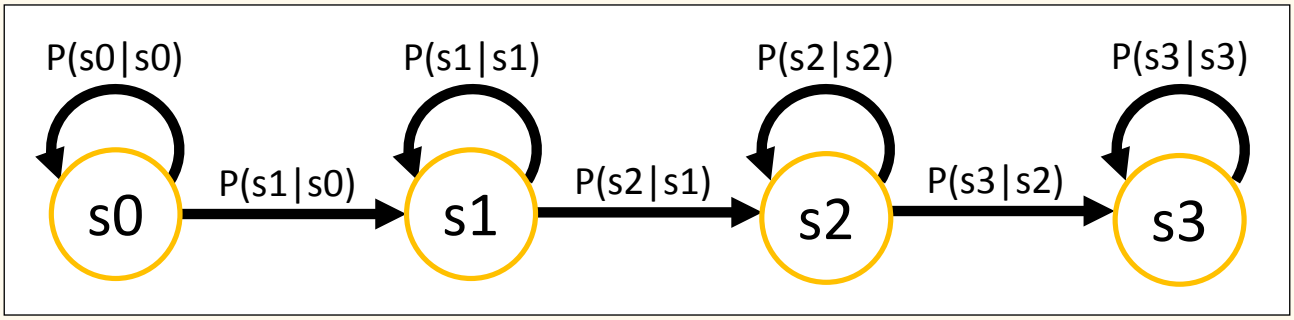
Un'altra applicazione estremamente interessante abilitata dall'*ASR* è la *classificazione e indicizzazione dei contenuti* basata sulla trascrizione del parlato. Infatti per alcuni generi, principalmente nel campo delle news e dei documentari, la maggior parte del contenuto semantico del programma è 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*:

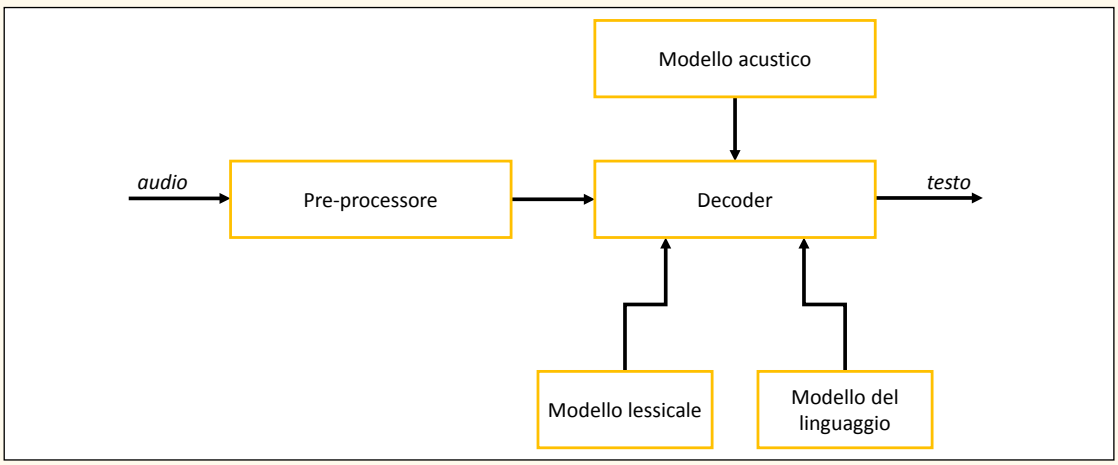
$$W^* = argmax_W P(X|W) * P(W) \quad (2)$$

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-

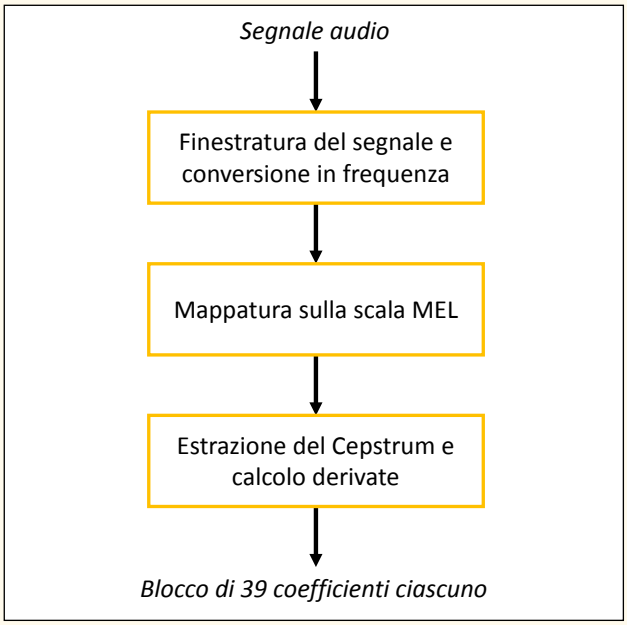


Fig. 3 – Pre-processamento del segnale audio

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]

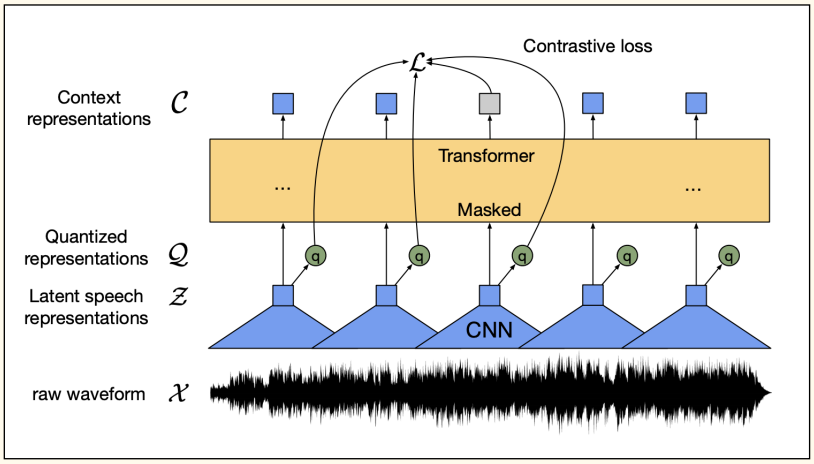


Fig. 4 – Schema di principio **wav2vec 2.0** *(fonte: [19])*

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

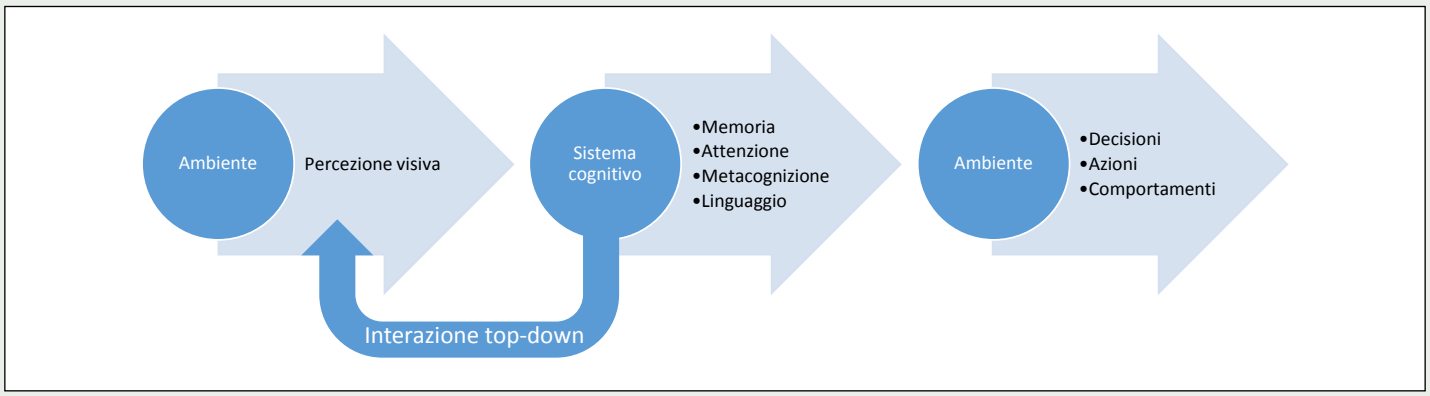
Con *Computer Vision* si denotano tradizionalmente le tecnologie e i sistemi che emulano le capacità umane di percezione ed elaborazione cognitiva supportate dal canale sensoriale visivo [1]. Tra i molti approcci di base utili all'inquadramento di questo ambito, il modello di riferimento qui preso ad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

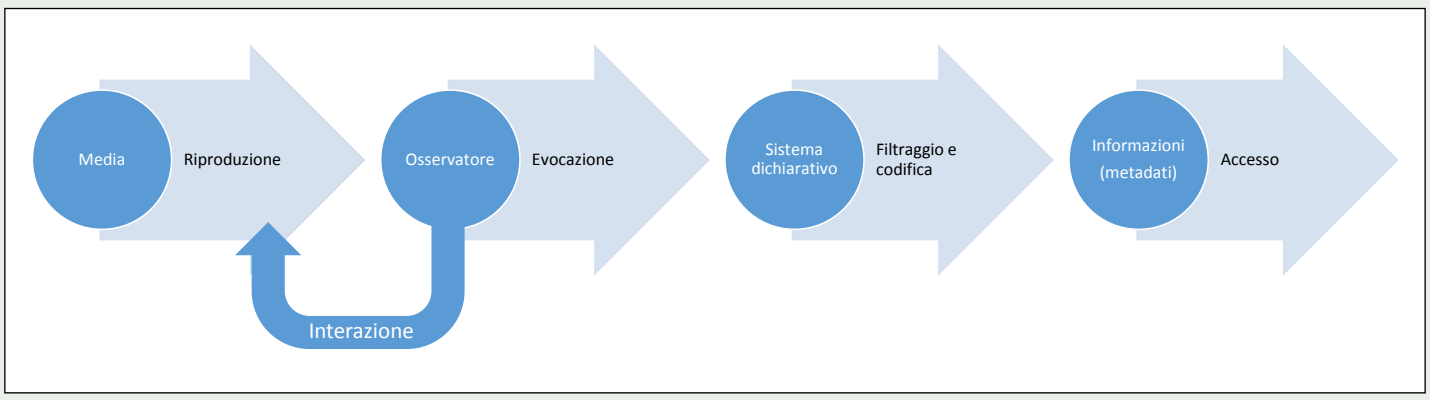
Si noti come il modello di processo rappresentato in Fig. 2 sia indipendente dalla natura dell'osservatore, che può essere un osservatore umano o un qualsivoglia componente di Intelligenza Artificiale [Nota 4] che ne emuli il comportamento e le capacità.

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.

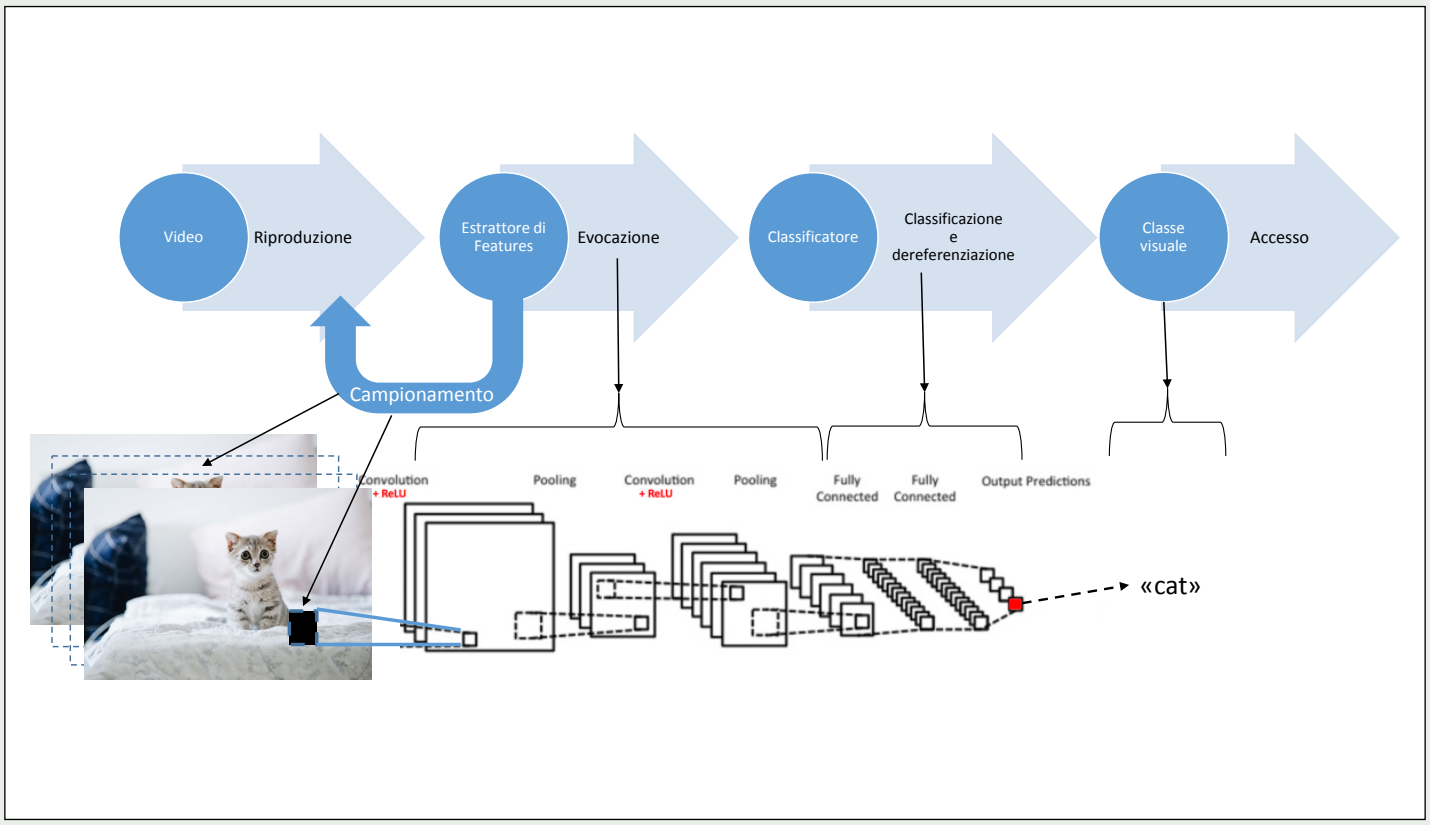
La sezione iniziale della rete funge da campionatore delle informazioni visuali da analizzare nella sezione successiva, dove avviene il processo di evocazione delle caratteristiche latenti.

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.

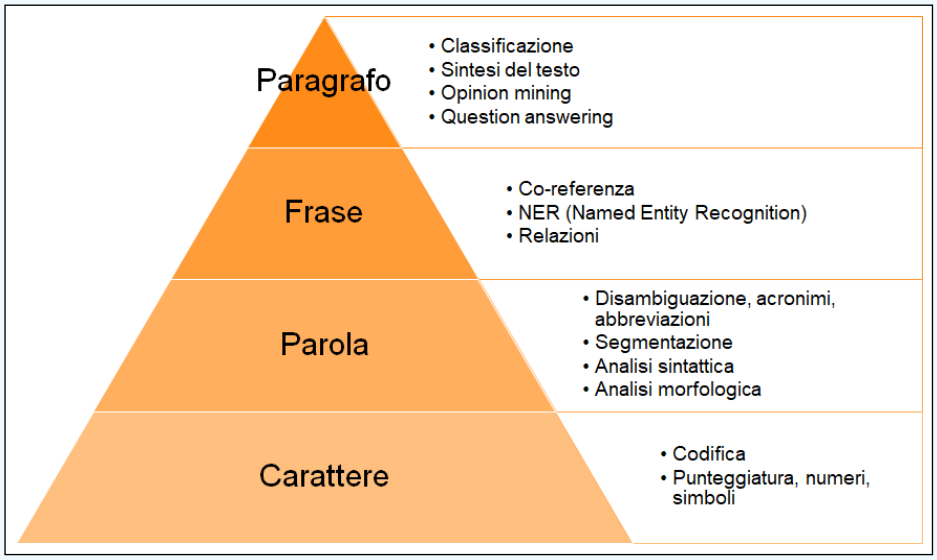


Fig. 1 – Strutturazione e semantica del testo

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).

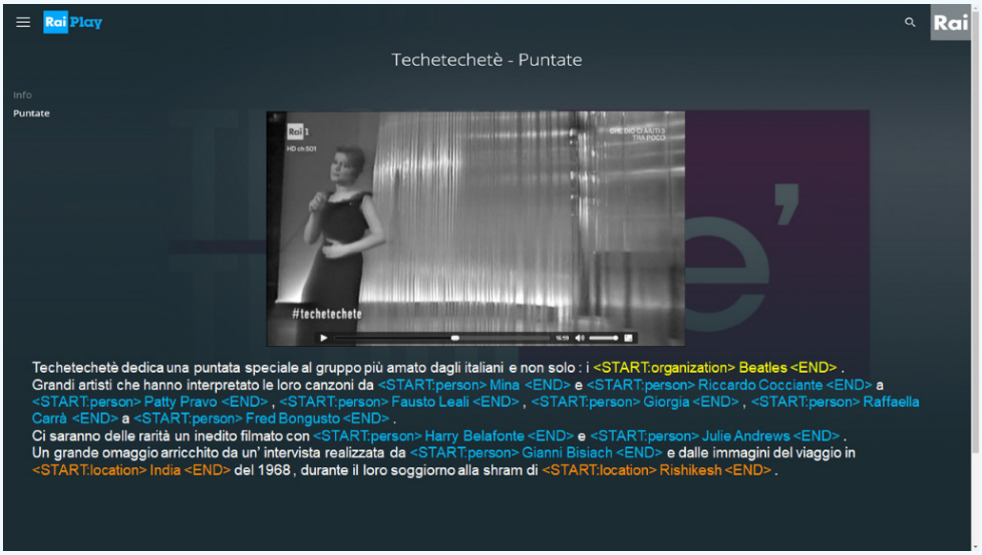


Fig. 2 – Esempio di applicazione dell'estrazione delle entità da una descrizione del portale RaiPlay

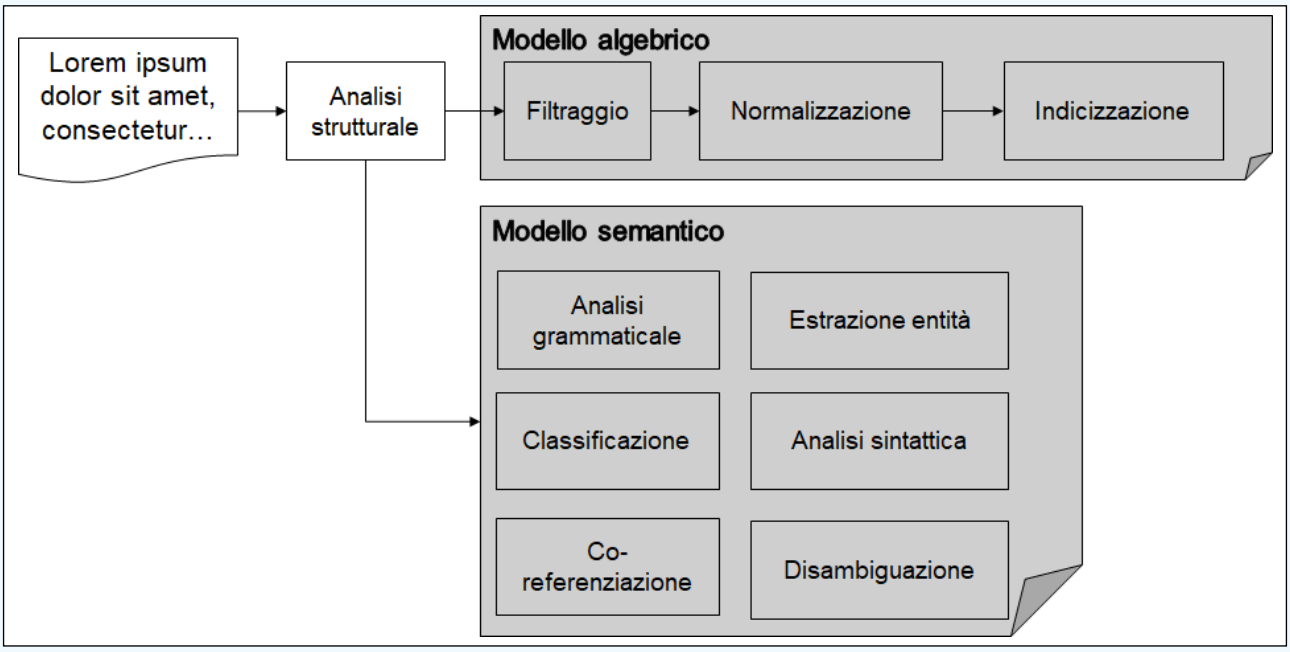


Fig. 3 – Architettura generale per la rappresentazione del testo.

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].

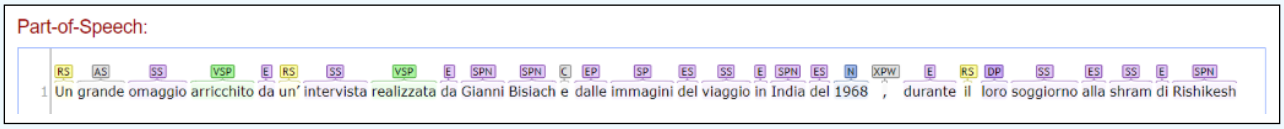
Il modello semantico include diversi tipi di analisi applicabili singolarmente o congiuntamente, a seconda delle necessità dell'utente, brevemente introdotti nel seguito. Per un approfondimento, si rimanda a [3].

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)*

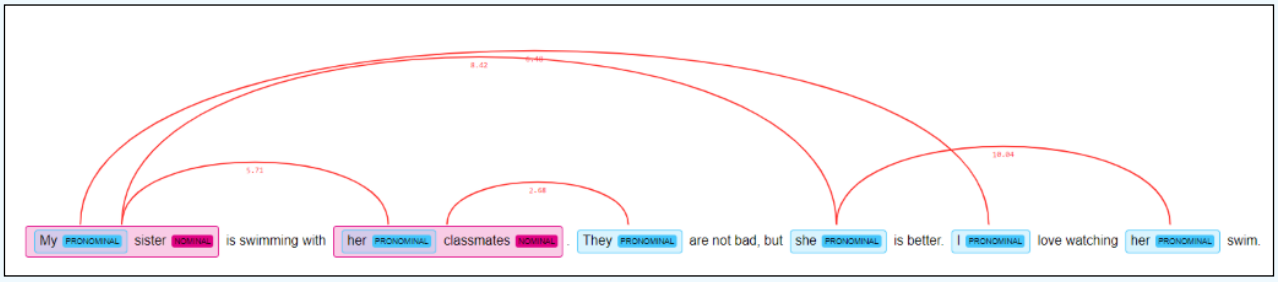
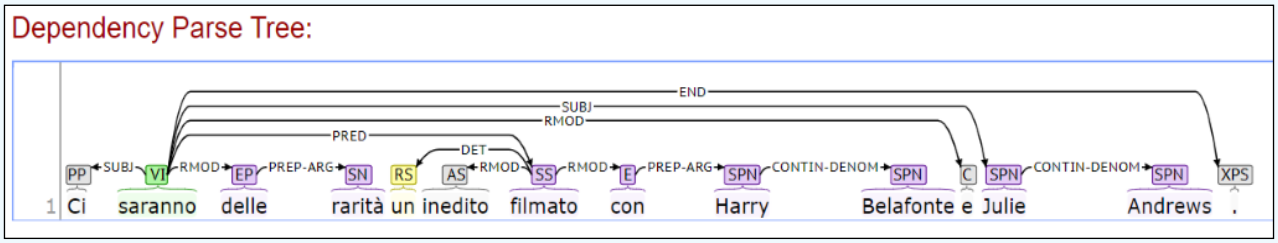


Fig. 6 – Esempio di analisi sintattica *(fonte: http://hlt-services2.fbk.eu/textpro-demo/textpro.php ultimo accesso 22/09/2020)*



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,

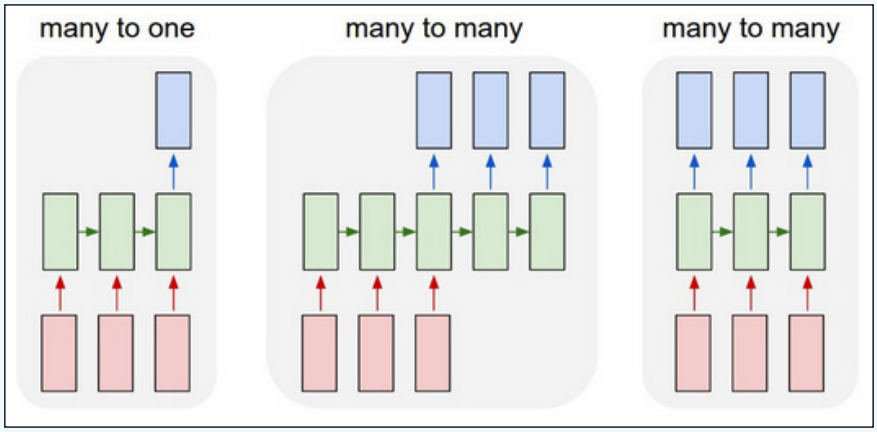


Fig. 7 – Illustrazione delle architetture di rete RNN per l'analisi del linguaggio. I rettangoli rossi, verdi e blu corrispondono, rispettivamente, agli ingressi (parole), unità di elaborazione (hidden layer) ed uscite della rete *(fonte: http://karpathy.github.io/2015/05/21/rnn-effectiveness/ ultimo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

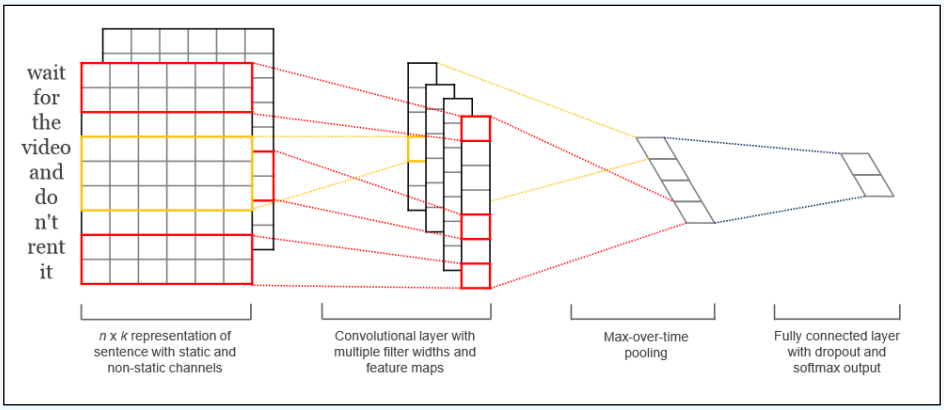
consentendone così il pieno sfruttamento delle potenzialità. Ad oggi le reti ricorrenti sono utilizzate in una moltitudine di applicazioni NLP, tra le quali citiamo le seguenti:

- *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])*



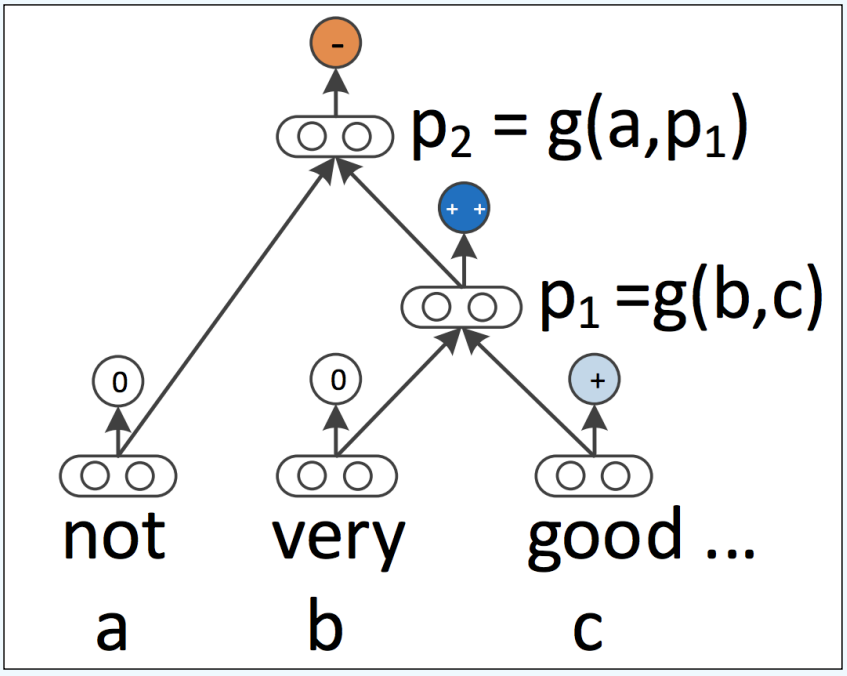
Sia le reti RNN che le CNN basano i propri fondamenti modellando il testo come una sequenza di parole. Tuttavia, da un punto di vista linguistico, il linguaggio naturale presenta anche caratteristiche gerarchiche, oltreché temporali. Infatti, come introdotto precedentemente, le parole sono utilizzate 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

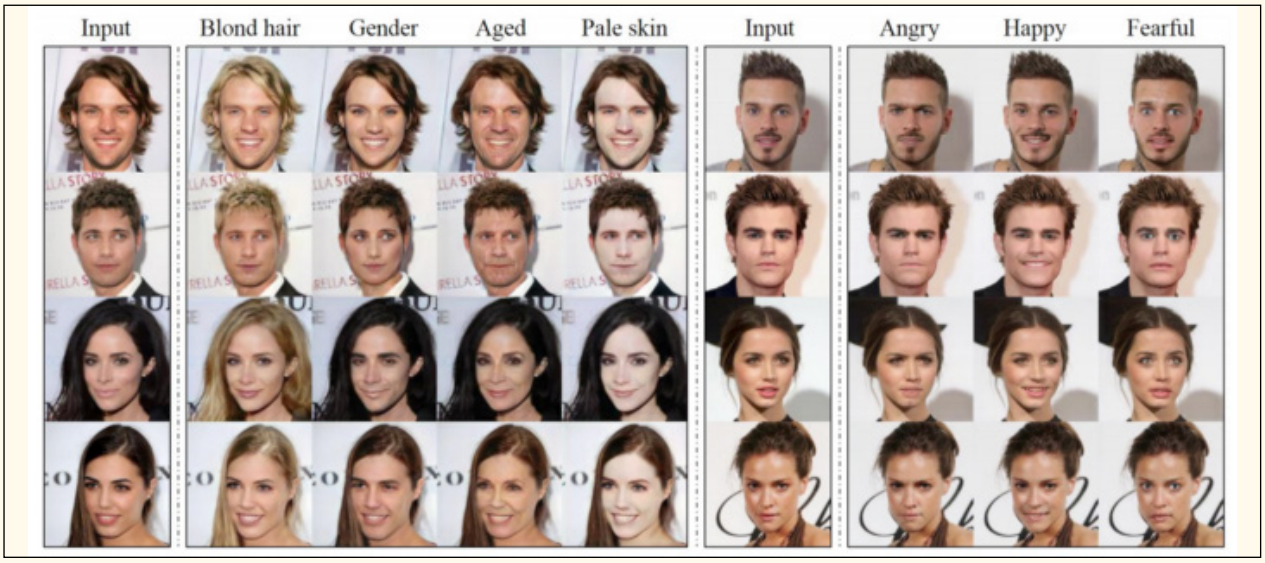


Fig. 3 – Volti (artificiali) da celebrità: GAN addestrata sul dataset *CelebA*

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.

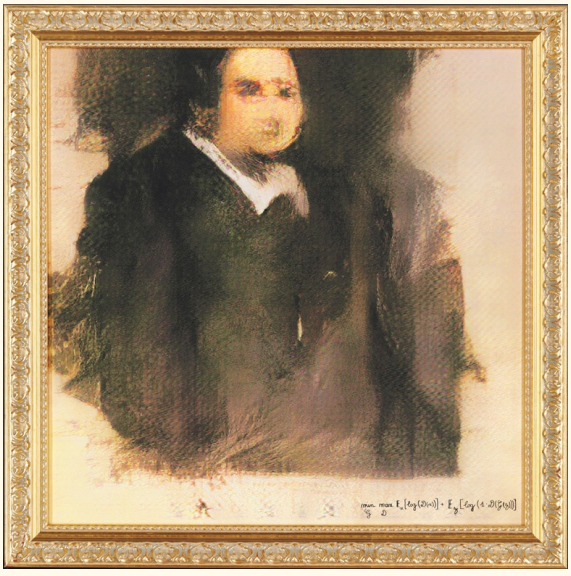


Fig. 4 – *Edmond de Belamy* (2018), creato attraverso una GAN

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*