Intelligenza Artificiale e Codifica Video

Una strada per superare i limiti dell'approccio tradizionale.

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

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

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

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

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

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

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

201_02.4_3.00

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

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

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

Video codec tradizionali

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

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

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

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

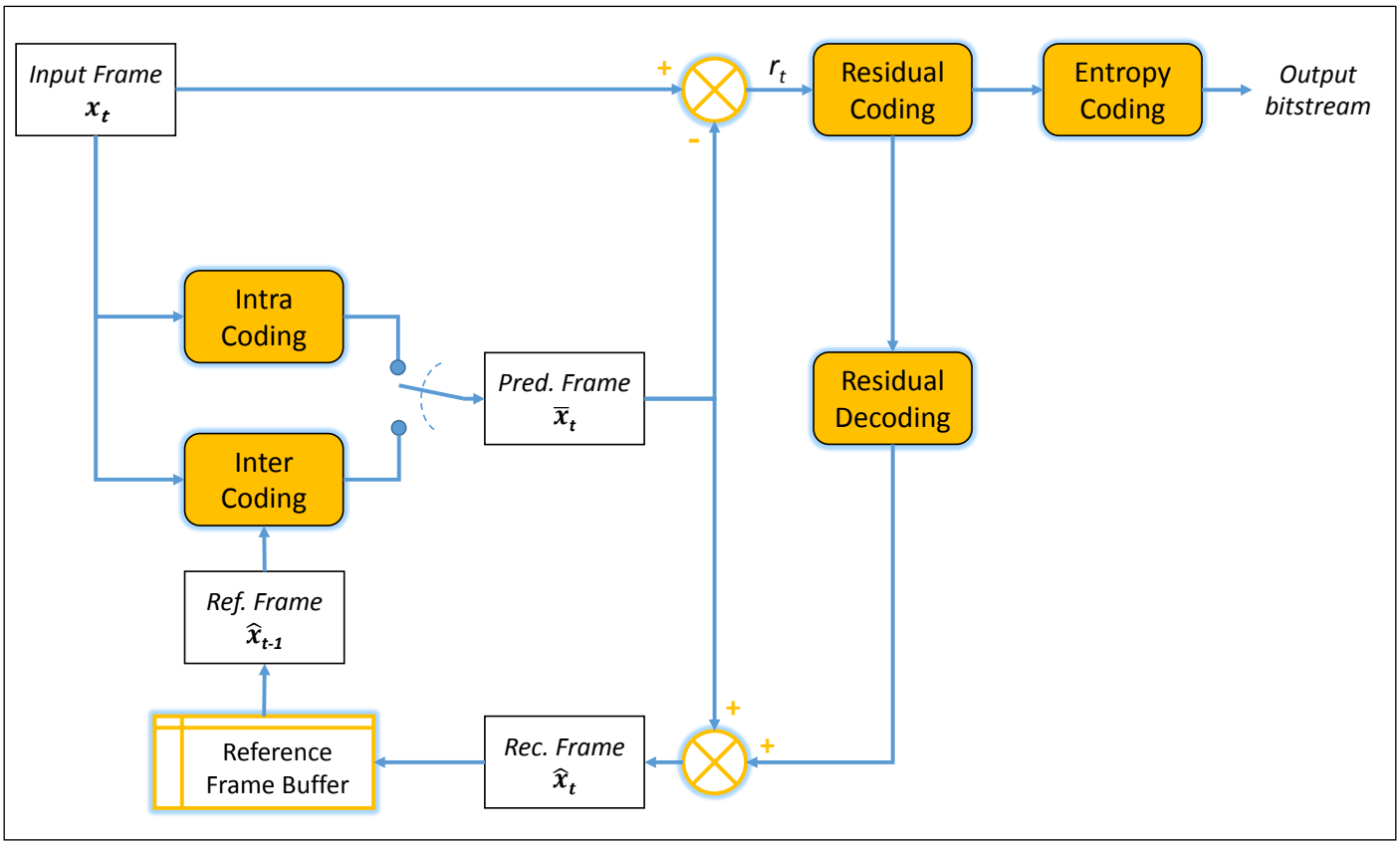
Inoltre, come per i suoi predecessori **MPEG-2**, **AVC** e **HEVC**, **VVC** è un sistema di compressione basato sull'inserimento nel loop di codifica di strumenti per la riduzione della *ridondanza spaziale* e *temporale* che caratterizza il segnale video, in particolare:

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

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

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

Fig. 1 – Schema del *codificatore ibrido MPEG*



Approcci basati su Intelligenza Artificiale

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

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

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

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

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

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

Hybrid Coding

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

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

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

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

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

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

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

End-to-End coding

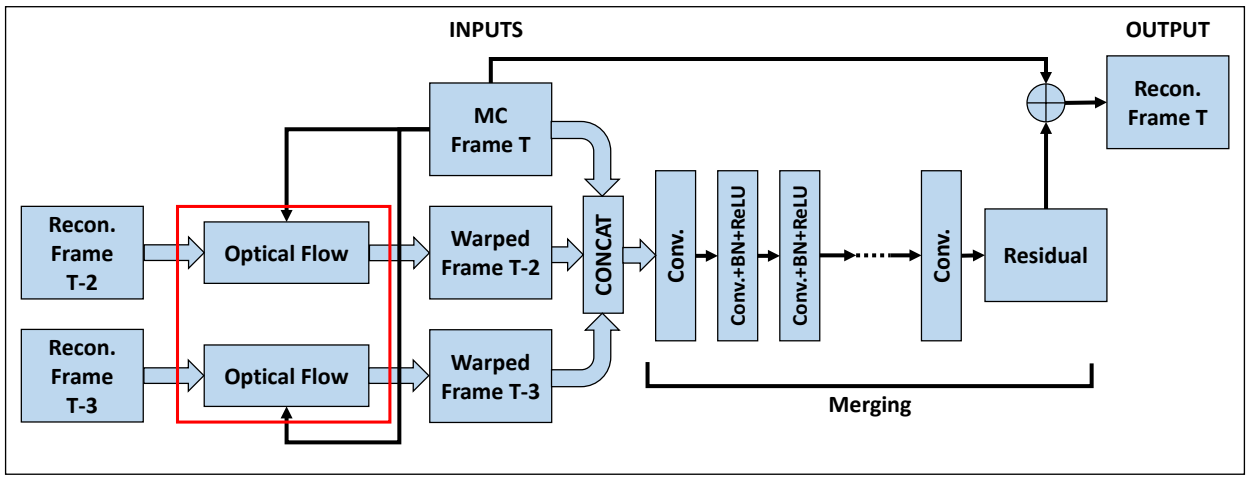
A differenza degli approcci tradizionali, non vi è un consenso comune sulle architetture di codifica basate sull'apprendimento di tipo *End-to-End (E2E)*.

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

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

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

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



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

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

Fig. 3 – Esempio di optical flow

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

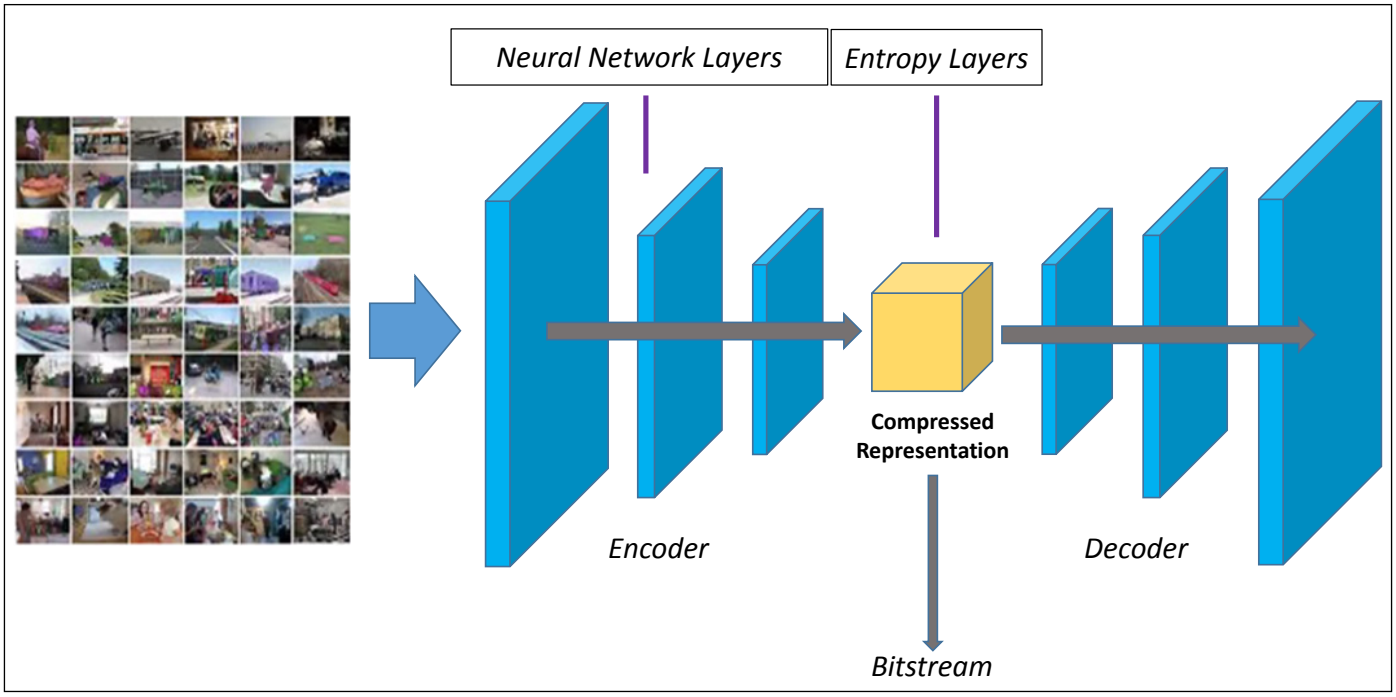
Analisi degli obiettivi comuni

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

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

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

Fig. 4 – Architettura *E2E*



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

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

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

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

Conclusioni

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

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

Bibliografia

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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