WORKSHOP

Open Data, Big Data e IA: le nuove sfide sui dati ambientali

Breve storia dell'Intelligenza Artificiale Multimodale e sue Applicazioni

Vittorio Murino

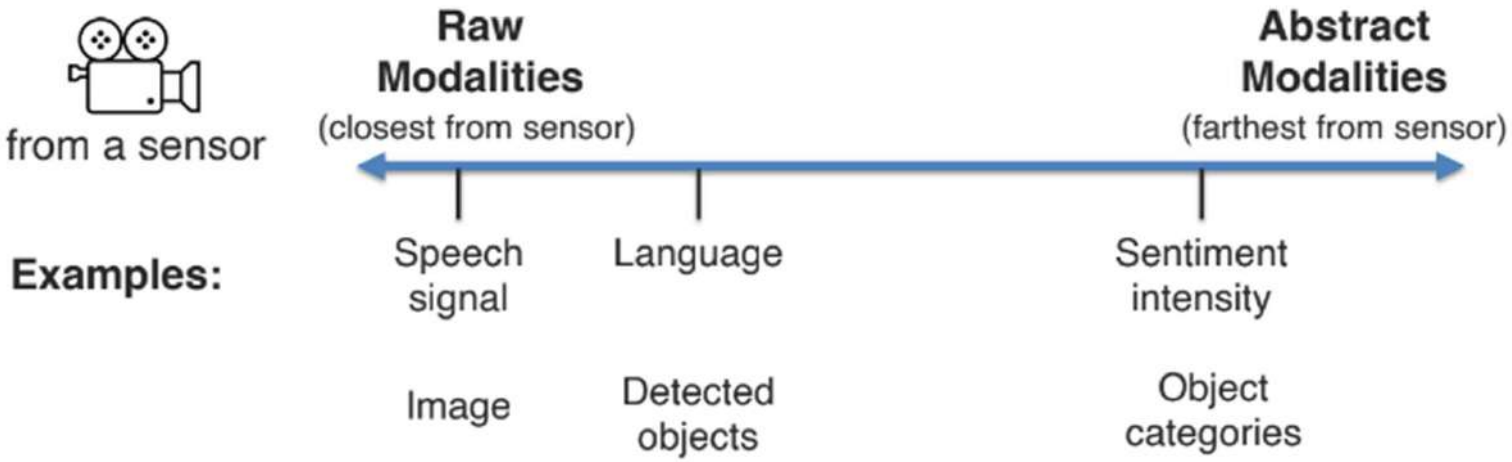
AI for Good (AIGO), Istituto Italiano di Tecnologia

Dipartimento di Informatica, Università di Verona

Genova, 24 Settembre 2025

(Multi-)Modalità

Modalità si riferisce al modo nel quale l'informazione è espressa o percepita, in particolare attraverso diverse modalità sensoriali

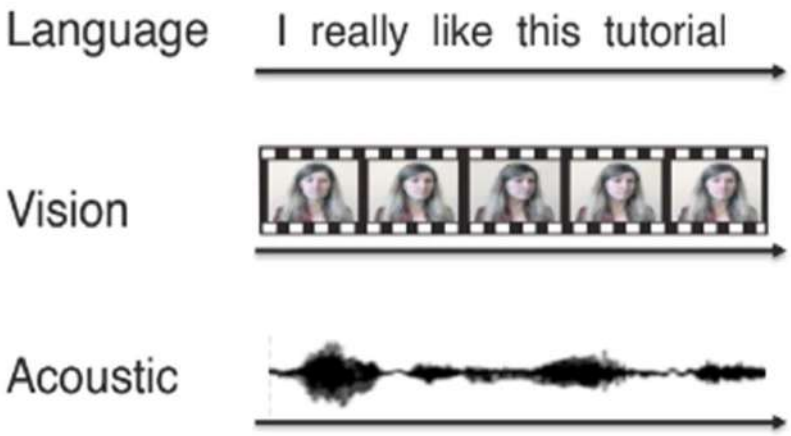


Cos’è il *Multimodal Learning*

- Integrazione di più modalità: immagini, testo, audio, diversa sensoristica, grafi
- Benefici: robustezza, prestazioni, generalizzazione, interpretabilità
- Sfide: allineamento spazio-temporale, sincronizzazione, dati non etichettati

Esempi di Modalità

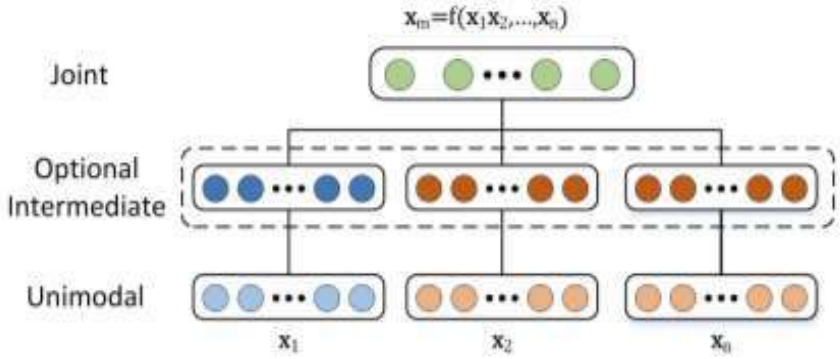
- Natural language – spoken, written
- Visual – images, videos, RGB, infrared (thermal), depth, …
- Audio – voice, sound, music
- Biological signals – Electroencephalogram (EEG), Electrocardiogram (ECG) and Galvanic Skin Response (GSR)
- Haptics – touch
- Motion capture system's data
- Dati multispettrali (remote sensing)
- …



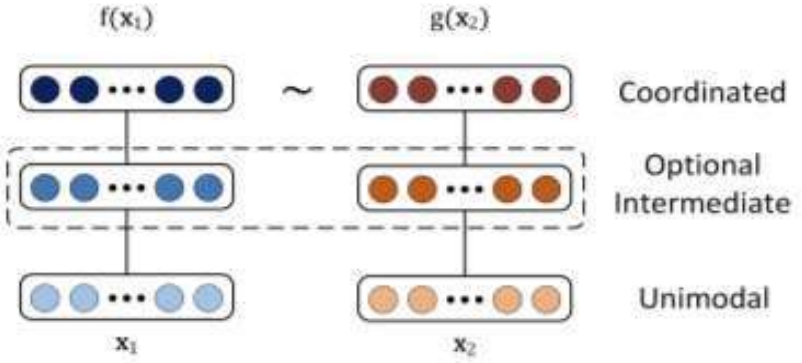
Rappresentazioni

- Modalità eterogenee, diverse per qualità e struttura → *rappresentazioni*
- "Imparare" rappresentazioni che riflettono interazioni cross-modali tra elementi di differenti modalità
- Come rappresentare e riassumere dati multimodali in modo da sfruttare la complementarietà e ridondanza di tali modalità multiple.
- Proiezione di tutte le modalità in un unico spazio, OR in diversi spazi "coordinati" usando una metrica di similarità (vincolo).

Rappresentazioni unite ("joint")



Rappresentazioni coordinate



T. Baltrusaitis, C. Ahuja, and L.-P. Morency, Multimodal Machine Learning: A Survey and Taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. 2019.

Allineamento

- Cercare **relazioni** e **corrispondenze** tra le sotto-componenti delle istanze da 2 o più modalità.
- Identificare le connessioni tra gli elementi di modalità multiple.
- *Esplicito* → segnale/dato VS. *implicito* → spazio (intermedio) latente
- Implica un **meccanismo di attenzione**, a.k.a. **self-attention** e **transformer**

T. Baltrusaitis et al., Multimodal Machine Learning: A Survey and Taxonomy. TPAMI, 2019.

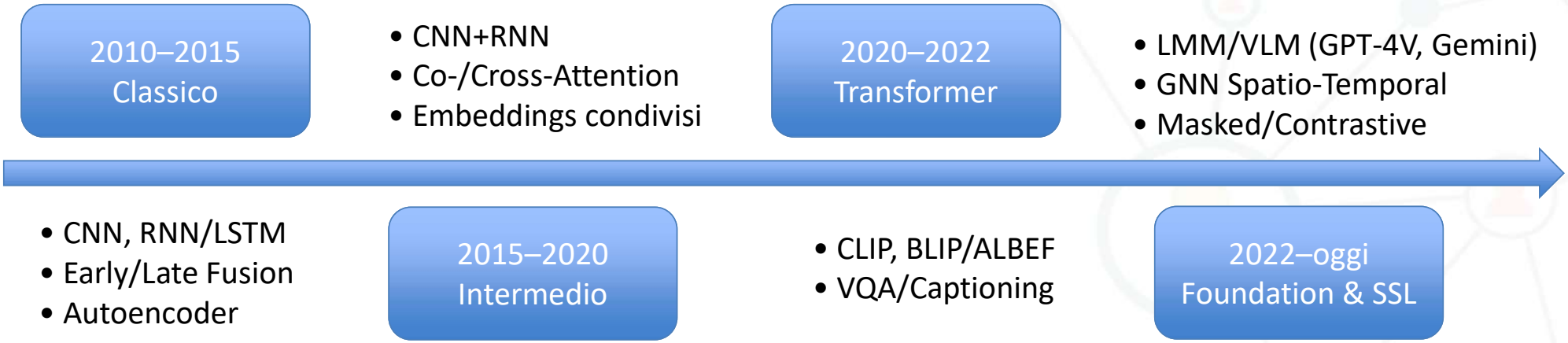
Meccanismi di Attenzione nei Dati Multimodali

- Consente a un modello di **assegnare pesi diversi a diverse parti dei dati** di input.
  - Il modello può **concentrarsi maggiormente sulle parti rilevanti di ciascuna modalità**, ignorando le informazioni meno importanti.
  - I meccanismi di attenzione aiutano ad allineare queste modalità imparando **quali parti di una modalità corrispondono a quali parti di un'altra modalità**

- Esempio: in un'attività che coinvolge sia testo che immagini, il modello può imparare a concentrarsi sulle parti dell'immagine che corrispondono a determinate parole o frasi nel testo.

- Efficace nella **gestione di sequenze di lunghezza variabile**.
  - Importante per i dati multimodali in cui, ad esempio, la lunghezza della descrizione del testo può differire dalla durata dell'audio corrispondente o dal numero di oggetti in un'immagine.

Timeline


2010–2015
Classico
• CNN, RNN/LSTM
• Early/Late Fusion
• Autoencoder
2015–2020
Intermedio
• CNN+RNN
• Co-/Cross-Attention
• Embeddings condivisi
2020–2022
Transformer
• CLIP, BLIP/ALBEF
• VQA/Captioning
2022–oggi
Foundation & SSL
• LMM/VLM (GPT-4V, Gemini)
• GNN Spatio-Temporal
• Masked/Contrastive

Tecniche classiche: Early vs Late Fusion

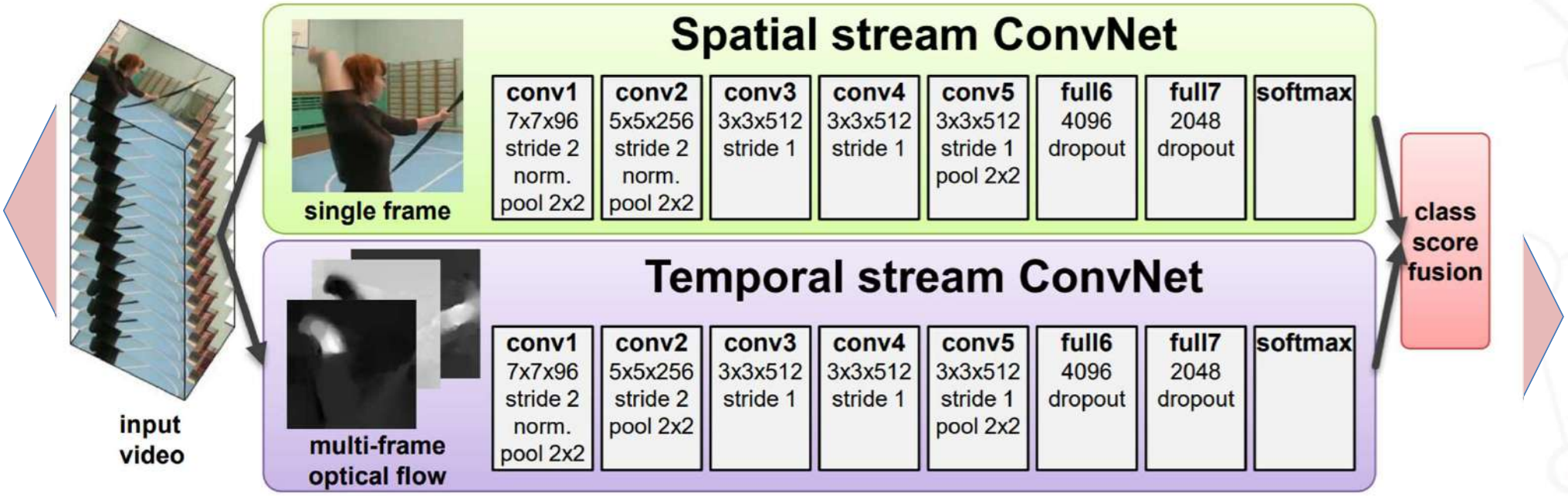
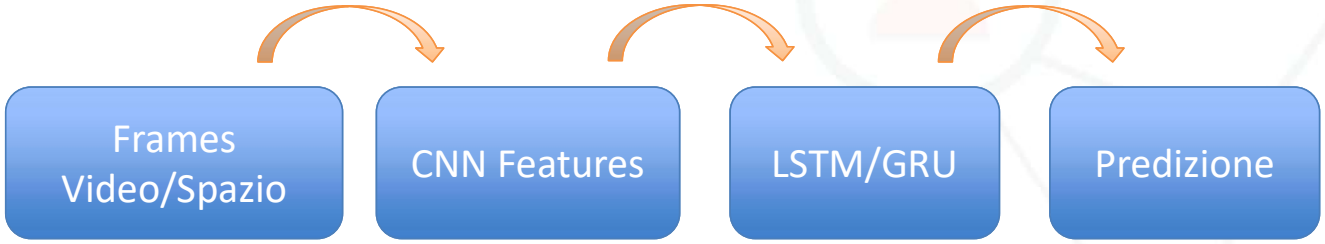
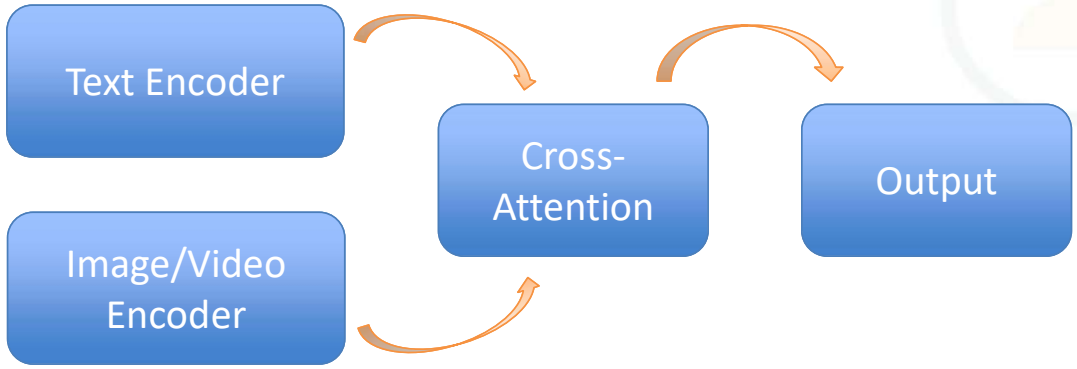


Figure 1: **Two-stream architecture for video classification.**

Approcci intermedi: Ibridi CNN+RNN e Cross-Attention





Transformer Multimodali

- **Self-/Cross-attention**; **pre-training** contrastivo/generativo su coppie immagine–testo
- Esempi: CLIP, BLIP/BLIP-2, ALBEF, Flamingo, Kosmos
- Task: *retrieval*, *captioning*, *VQA*, ragionamento multimodale

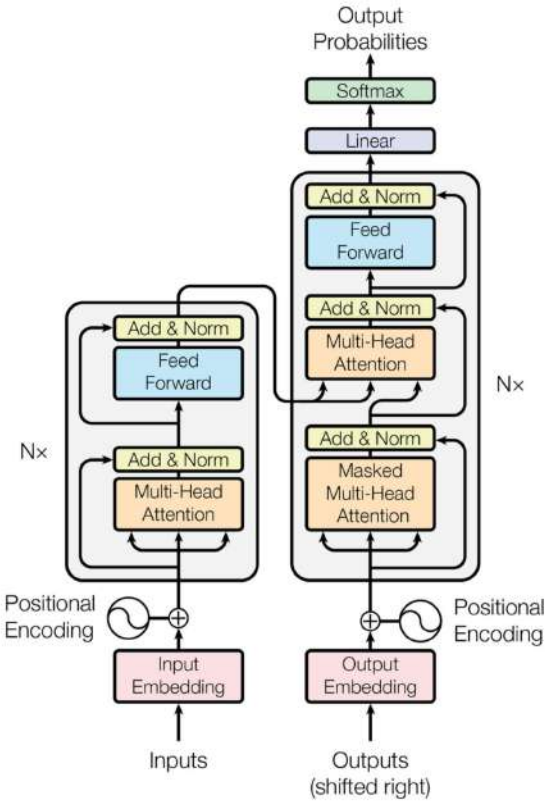
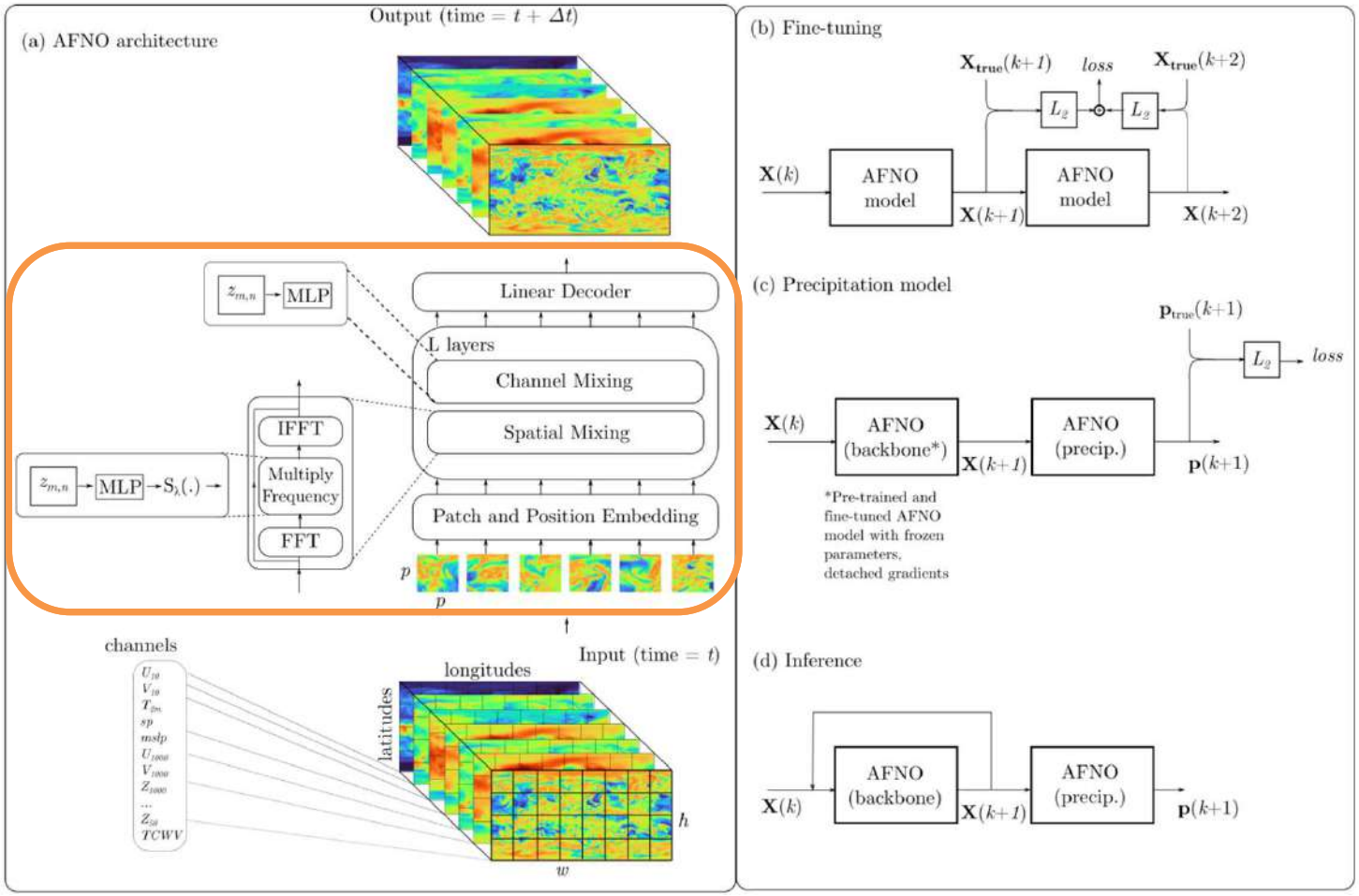


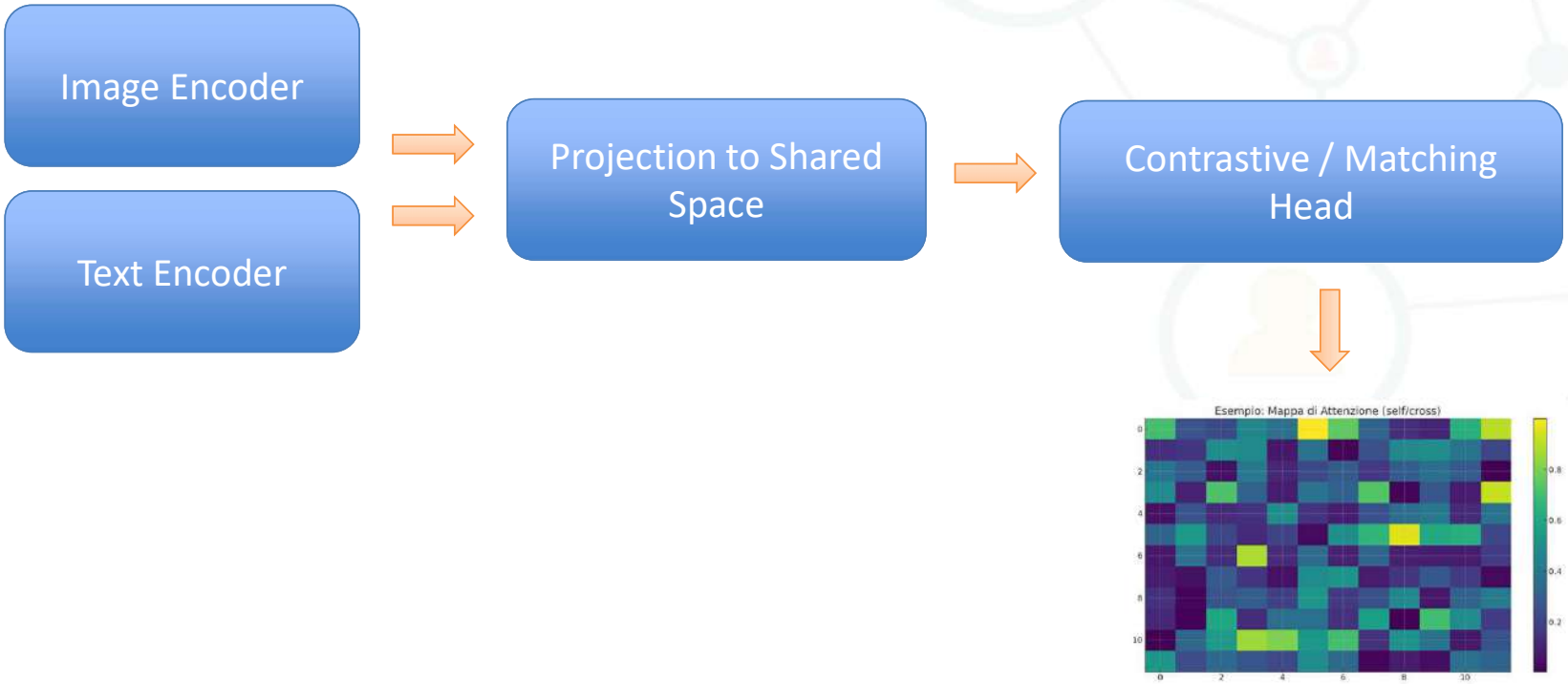
Figure 1: The Transformer - model architecture.

FourCastNet



Nvidia & Caltech, 2022

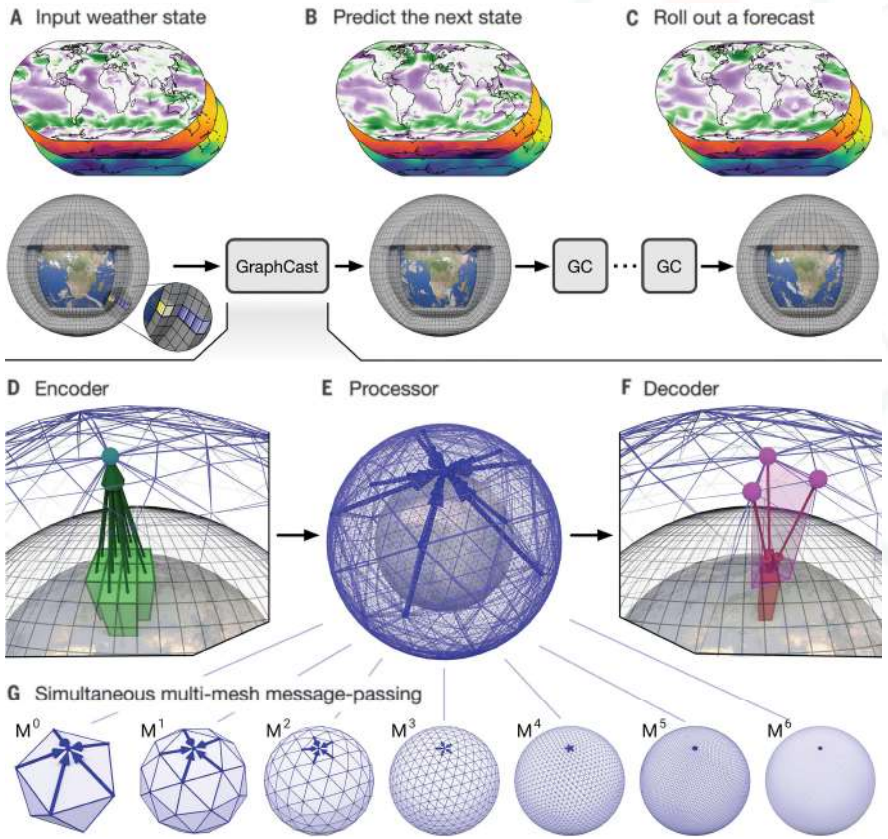
Architettura VLM (stile CLIP) + Attenzione



Graph Neural Networks e dati strutturati

- *Message passing*, grafi dinamici spatio-temporali
- Fusione su grafo: nodi multimodali, edge dipendenti dal contesto
- Applicazioni: meteo (griglie/GraphCast), traffico, reti biologiche

GraphCast



Google DeepMind (2023)

Modelli Fondazionali Multimodali

- **Pre-training** su dataset eterogenei e massivi (VLM/LMM)
- **Adattamento**: fine-tuning, LoRA, prompt/in-context
- Capacità: ragionamento multimodale, tool-use, generalizzazione zero/few-shot

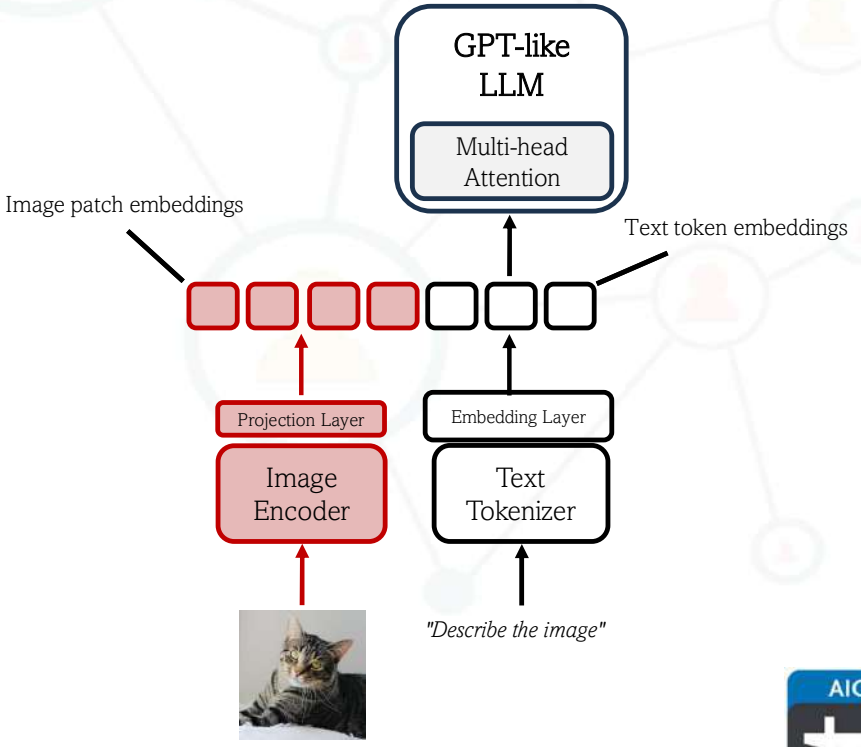
FM Multimodali

- Se si considerano i **vision-language FMs**, possiamo identificare 3 blocchi principali: *Image Encoder*, *Text Tokenizer*, *GPT-like LLM*

- **Usa un singolo decoder**, come un'architettura standard pre-addestrata LLM, e.g., GPT-2.

- **Immagini sono convertite in *tokens*** (i.e., vettori) con la stessa dimensione degli *embedding* dei token di testo

- Questo permette all'**LLM** di **processare i token sia di testo che di immagini insieme dopo averli concatenati**.
  Ex: LLaVA, MiniGPT-4

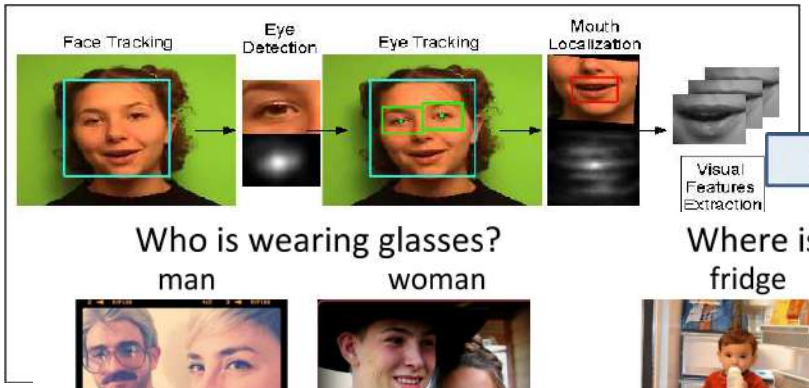


Esempi di Applicazioni Multimodali

- Vision+Language: *captioning*, VQA, *retrieval* multi/cross-modale
- Medicina: immagini + referti testuali
- Autonomous driving: sensor fusion (video, lidar, radar)
- Meteo classico: *nowcasting* radar+satellite, FourCastNet/GraphCast
- Meteo V+L: generazione report, *captioning* satellitare, *retrieval*

Examples

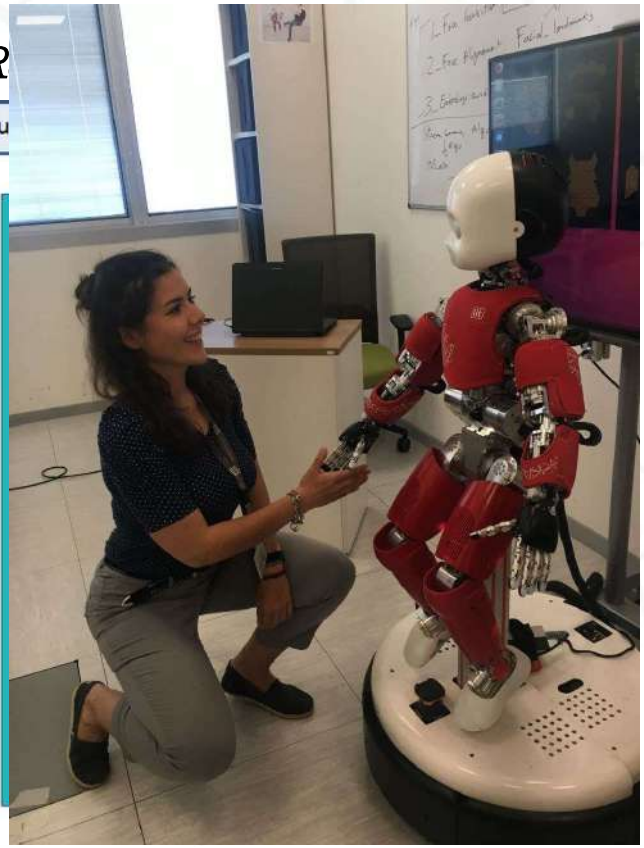
Audio-visual speech recognition



Event/Behavior R

VQA

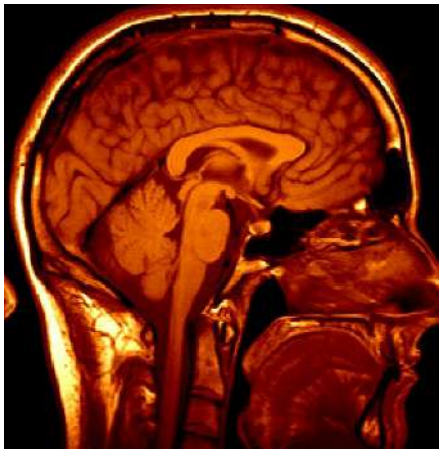
Affective Computing

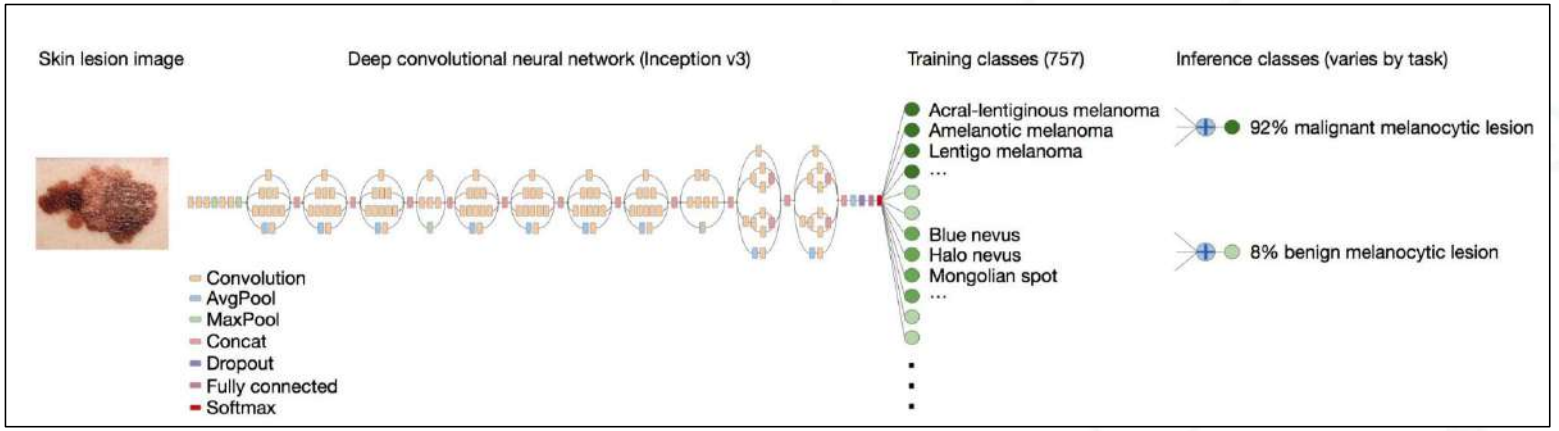


Human-Human/Human-Robot Interaction

Medical imaging

3D imaging (MRI, CT)





Skin cancer classification with deep learning
https://cs.stanford.edu/people/esteva/nature/

Weather Forecasting using Deep Learning

Problematiche

- Gestione dati spazio-temporali 4D, multi-canale, a diversa risoluzione e struttura sia nello spazio che nel tempo.
- Generalizzazione su aree diverse
- *Scope* temporale e spaziale
- Interpretabilità

Modelli DL

- **CNNs**: Efficace per riconoscimento di pattern spaziali.
- **LSTMs**: gestisce dipendenze temporali nei dati.
- **Transformers**: modellazione avanzata di dati sequenziali.
- **Generative Models**:
  - Adversarial models (e.g., DGMR).
  - Diffusion models (e.g., Prediff).

Weather Forecasting using Deep Learning

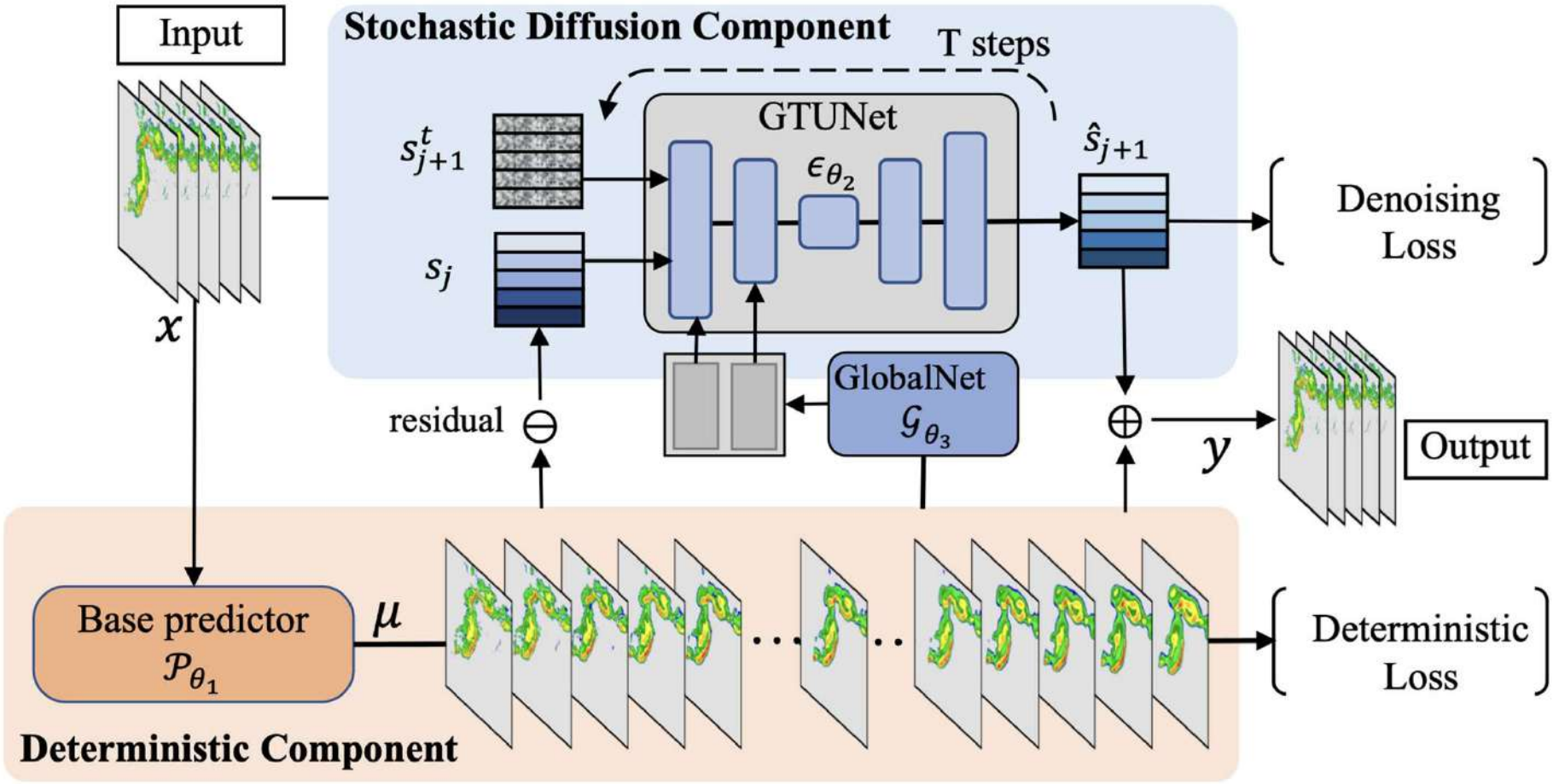
Vantaggi DL rispetto ai metodi tradizionali

- **Velocità ed efficienza:** elabora grandi set di dati meteorologici più velocemente dei modelli numerici.
- **Riconoscimento di pattern:** identifica relazioni complesse e non lineari nei dati.
- **Automazione:** consente un'elaborazione rapida, dalla raccolta dei dati alla generazione delle previsioni.

Sfide nelle predizioni mediante DL

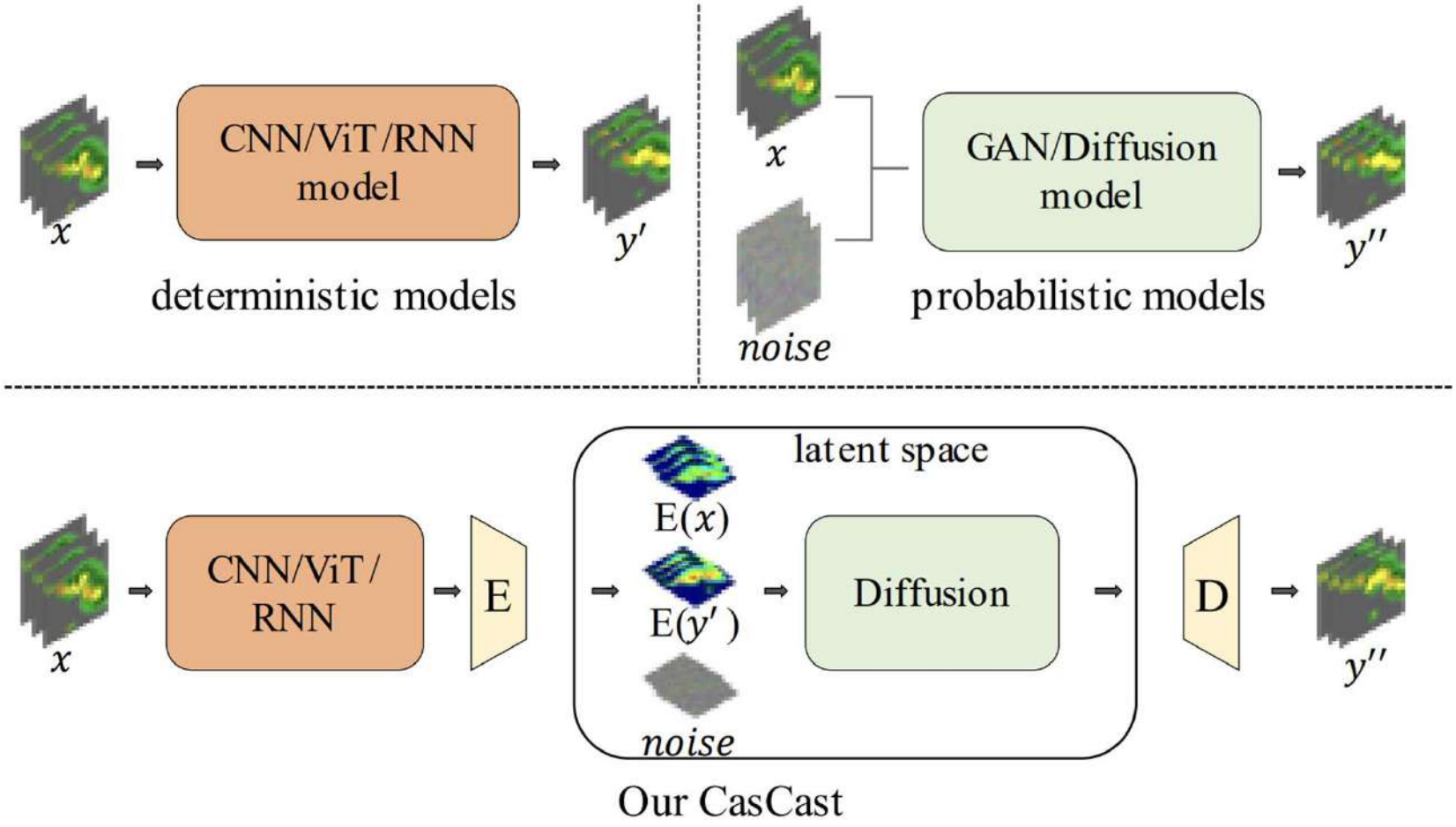
- Qualità dei dati e costi computazionali.
- Generalizzazione in diverse aree geografiche.
- Difficoltà nel prevedere eventi estremi.
- Assenza di vincoli fisici.

DiffCast (Diffusion Model)



D. Yu, et al. “DiffCast: A Unified Framework via Residual Diffusion for Precipitation Nowcasting”, CVPR 2024.

CasCast (Diffusion Model & Transformer)



J. Gong, et al. “CasCast: Skillful High-resolution Precipitation Nowcasting via Cascaded Modelling”, ICML 2024.

Earthformer

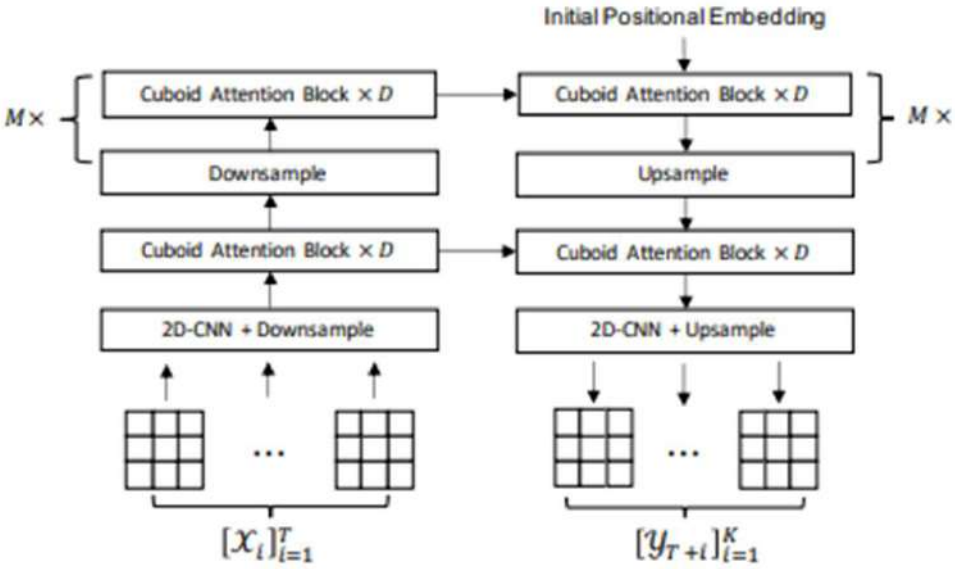


Figure 2: Illustration of the Earthformer architecture. It is a hierarchical Transformer encoder-decoder based on cuboid attention. The input sequence has length T and the target sequence has length K. "$\times D$" means to stack D cuboid attention blocks with residual connection. "$M\times$" means to have M layers of hierarchies.

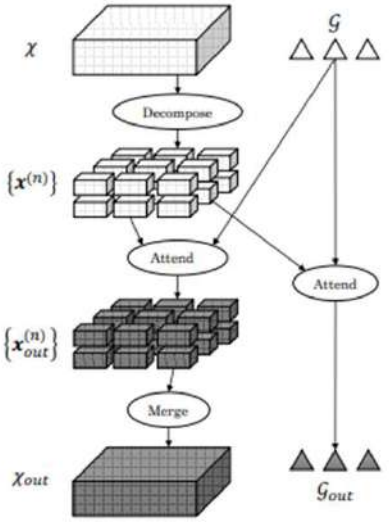


Figure 3: Illustration of the cuboid attention layer with global vectors.

(a) strategy="local" shift=(0, 0, 0)

(b) strategy="dilated" shift=(0, 0, 0)

(c) strategy="local" shift = (0, 1, 1)

Figure 4: Illustration of cuboid decomposition strategies when the input shape is $(T, H, W) = (6, 4, 4)$, and cuboid size $(b_T, b_H, b_W) = (3, 2, 2)$. Cells that have the same color belong to the same cuboid and will attend to each other. shift $= (0, 1, 1)$ shifts the cuboid decomposition by 1 pixel along height and width dimensions. strategy = "`local`" means to aggregate contiguous (b_T, b_H, b_W) pixels as a cuboid. strategy = "`dilated`" means to aggregate pixels every $\lceil \frac{T}{b_T} \rceil$ ($\lceil \frac{H}{b_H} \rceil$, $\lceil \frac{W}{b_W} \rceil$) steps along time (height, width) dimension. (Best viewed in color).

Z. Gao, et al. "Earthformer: Exploring Space-Time Transformers for Earth System Forecasting", NeurIPS 2022.

NowcastingGPT

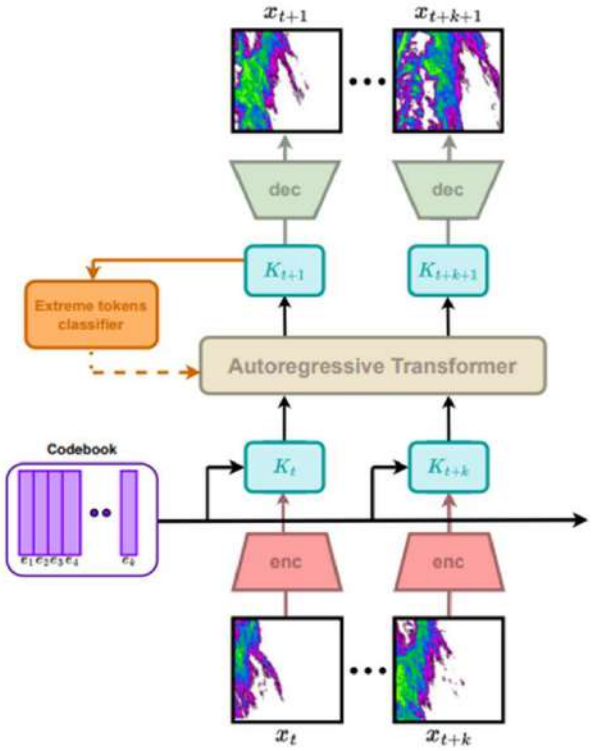


Figure 2: The image shows the NowcastingGPT-EVL model architecture. The VQ-VAE Encoder and Decoder are depicted in red and green respectively. The Extreme tokens classifier is depicted in orange, it takes the predicted tokens as input from the transformer and outputs the probabilities u_t used in the EVL loss. The dashed line indicates that the output of the Classifier is only used to optimize the transformer and not as input.

C. Meo, et al. “Extreme Precipitation Nowcasting using Transformer-based generative model”, ICLR 2024.

Sfide aperte e prospettive future

- Multimodalità è il paradigma di riferimento
- Scalabilità, efficienza e latenza; streaming/online fusion
- Valutazione casi rari/estremi; robustezza, bias
- Interpretabilità, auditing multimodale, sicurezza e affidabilità

Direzioni future

- Integrare modelli fisici con tecniche IA.
- Sviluppo di modelli fondazionali per i diversi task meteo.
- Far leva sui *big data analytics* per migliorare le predizioni