NODES – Nord Ovest Digitale e Sostenibile

STATO DELL'ARTE DEI MODELLI DI INTELLIGENZA ARTIFICIALE MULTIMODALI [BiTe]

SPOKE 3 – Industria del turismo e cultura

DELIVERABLE D 1.0

Version history

No.	Date	Details	Author(s)
0.1			
0.5	Jul 16, 2024	Initial draft	Pietro Ruiu; Andrea Lagorio Università degli Studi di Sassari
0.9	Jul 16, 2024	Draft	Pietro Ruiu; Andrea Lagorio; Enrico Grosso Università degli Studi di Sassari
1	Jul 31, 2024	First Version	Pietro Ruiu; Andrea Lagorio; Enrico Grosso Università degli Studi di Sassari

This document is part of the project NODES which has received funding from the MUR – Missione 4, Componente 2, Investimento 1.5 – Creazione e rafforzamento di "Ecosistemi dell'innovazione", costruzione di "leader territoriali di R&S" – del PNRR with grant agreement no. ECS00000036

STATO DELL'ARTE DEI MODELLI DI INTELLIGENZA ARTIFICIALE MULTIMODALI [BiTe] © 2024 by OverApp srl; Università degli Studi di Sassari; Pietro Ruiu; Andrea Lagorio; Enrico Grosso is licensed under CC BY-SA 4.0

overapp

UNISS UNIVERSITÀ DEGLI STUDI DI SASSARI

Contents

Contents 2
A) Introduzione 4
B) Computer Vision 4
Recenti sviluppi legati all'emergere delle tecnologie basate su AI 4
Transformer 5
I vision Transfomers (ViT) 6
C) Modelli di apprendimento basati sul linguaggio 8
Large Language Model (LLM) 9
Vision-Language Models (VLM) 10
Large Language Model Multimodali (MLLM) 11
Esempi di MLLM 15
Flamingo (2022, DeepMind) 15
BLIP-2 (2023, Salesforce) 15
Kosmos-2 (2023, Microsoft) 16
LLaVA (2023, University of Wisconsin-Madison) 16
Gemini 1.5 (2024, Google) 16
Qwen-VL (2024, Alibaba Cloud) 17
Claude 3.5 Sonnet (2024, Anthropic) 17
GPT-4o (2024, OpenAI) 17
Conclusioni 19
Bibliografia 20

STATO DELL'ARTE DEI MODELLI DI INTELLIGENZA ARTIFICIALE MULTIMODALI [BiTe] © 2024 by OverApp srl; Università degli Studi di Sassari; Pietro Ruiu; Andrea Lagorio; Enrico Grosso is licensed under CC BY-SA 4.0

Glossary

	Definition
Hub Coordinator (HC)	The Hub Coordinator represents the single point of contact for the implementation of the innovation ecosystem towards the MUR. It carries out the management and coordination activities of the innovation ecosystem, receives the fundings, verifies, and transmits to the MUR the reporting of the activities carried out by the Spoke and their affiliates.
National Recovery and Resilience Plan (NRRP)	This document uses the Italian acronym for the NRRP, which is PNRR (Piano Nazionale della Ripresa e Resilienza)
Research Program Manager	The person who will be the responsible for the overall scientific contents of the NODES project. The NODES will appoint the Research Program Manager. It refers to "Responsabile del Programma di Ricerca" in the MUR's Call of proposal for "Ecosistemi di Innovazione"
NODES' Research and innovation program	NODES' Research and Innovation program is articulated in specific programs for each Spoke, with the aim to promote and support applied research on topics consistent with the Intelligent Specialization Strategy, with the guidelines of the 2021-2027 partnership agreement scheme, with regional operational plans and regional and national research and innovation priorities. Although NODES' Spokes are concentrated on different themes, they will organize their activities and actions within a common framework – NODES' Booster Methodoloy
Spoke Coordinator	The University in charge of coordinating the Spoke's ecosystem. It refers to "Spoke" in the MUR's Call of proposal for "Ecosistemi di Innovazione"
Spoke Data Manager	The person who will be the responsible for the monitoring and management of data generated at the Spoke level. The Spoke Coordinator will appoint the Spoke Data Manager.
Spoke Partner	The entity associated to the Spoke Coordinator. It can be an Innovation Cluster, Competence Center, Research Center related to the Spoke's ecosystem and contributes to achieve objectives and impact under the Spoke' leadership and management. It refers to "soggetti affiliati" in the MUR's Call of proposal for "Ecosistemi di Innovazione".
Spoke Project manager	The person who will be the responsible for the management, coordination and progress of the project at the Spoke level. The Spoke Coordinator will appoint the Spoke Project Manager.
Spoke research and innovation program	NODES' Research and Innovation program is articulated in specific programs for each Spokes. The spoke will leverage a consolidated collaboration with leading private and public companies and will focus the applied research activity on technological domains and applications that can favour the integration of SMEs into new value chains.
Spoke Scientific and Technical Manager	The person who will be the responsible for the overall scientific contents of the project at the Spoke level. The Spoke Coordinator will appoint the Spoke Scientific and Technical Manager.
Spoke Stakeholders Committee (SC)	Consultation structure formed by relevant stakeholders (Government, universities, companies, civil society, third sector, etc.)
Spoke Thematic	General target focus and domain of the Spoke research.
Spoke Topics	Specific areas/lines of development within the Spoke.
Spoke Work Package Leader	At the Spoke level, Work Packages (WPs) will be organized by WP leaders, who will be responsible for performance evaluation and reporting.
Flagship Project	Main research project at the Spoke level with the goal of prototyping, testing, demonstrating the research activities towards higher TRLs.

A) Introduzione

BITE è un progetto che prevede lo sviluppo di App fruibili su dispositivi mobili capaci di fornire dati sui flussi turistici e servizi agli utenti attraverso notifiche proattive. Attraverso la raccolta, la condivisione e l'analisi di immagini di siti d'interesse turistico, vuole contribuire all'ottimizzazione dell'allocazione delle risorse e a ridurre la congestione e l'usura delle destinazioni turistiche. Vuole inoltre partecipare alla creazione di esperienze turistiche immersive e sostenibili, grazie all'interazione responsabile tra imprese turistiche e visitatori, favorendo la conservazione della natura e il benessere degli ecosistemi locali. Uno degli obiettivi del progetto è mettere a disposizione una serie di funzioni atte ad elaborare con tecniche evolute i dati raccolti ed integrati nella piattaforma BITE. In particolare, tali funzioni saranno dedicate all'analisi del contenuto semantico delle immagini.

Il presente report tecnico esplora i recenti sviluppi nel campo della Computer Vision e dell'Intelligenza Artificiale, con un focus particolare sull'implementazione e l'evoluzione delle architetture di rete neurale, in particolare dei Transformer, nel contesto della visione artificiale.

La Computer Vision, che consente ai computer di interpretare e comprendere le informazioni visive in modo simile agli esseri umani, ha subito trasformazioni significative grazie all'adozione di tecniche avanzate di Deep Learning e alla crescente disponibilità di hardware specializzato. Questo report analizza le tappe fondamentali dello sviluppo tecnologico in questo ambito, illustrando il panorama scientifico e tecnologico di diverse tecnologie per la Computer Vision. Si descriveranno in particolare gli impatti dei Vision Transformers (ViT), sulle prestazioni e applicazioni dei modelli di visione artificiale e il ruolo dei modelli multimodali e dei Large Language Models (LLM) nell'integrazione di input testuali e visivi, evidenziando le sfide e i progressi recenti nell'ottimizzazione delle capacità di riconoscimento e generazione delle immagini.

B) Computer Vision

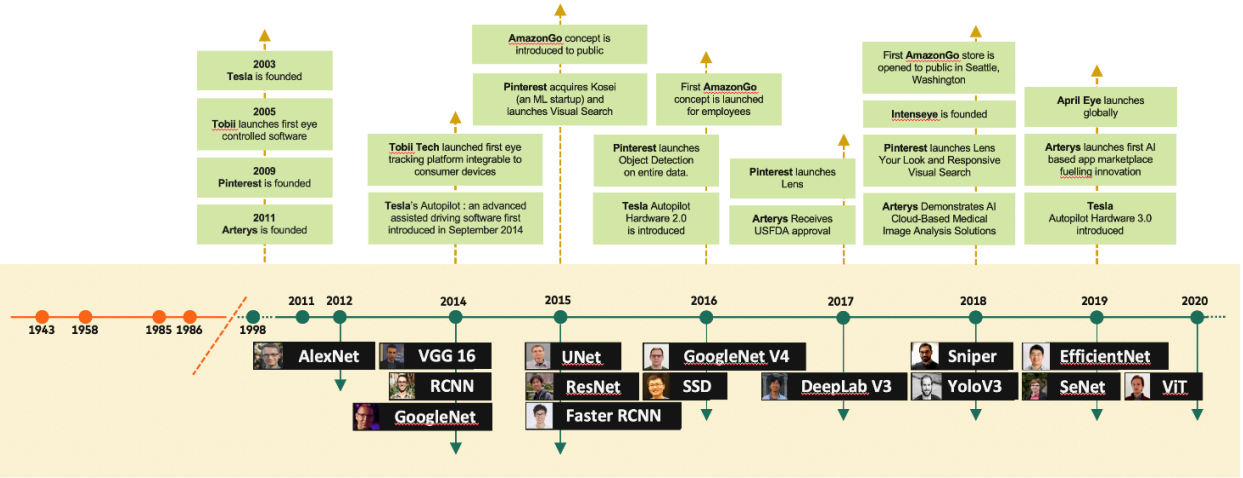
L'essere umano possiede la capacità innata e straordinaria di localizzare, riconoscere e classificare gli oggetti che vede. Fin dalla nascita, il cervello è in grado di elaborare informazioni visive complesse e di identificarne rapidamente le caratteristiche salienti. Questo processo, che avviene in modo naturale e senza sforzo, ci consente di interagire efficacemente con l'ambiente che ci circonda. Da decenni i ricercatori stanno cercando di sviluppare metodi che consentano alle macchine di imitare questa capacità umana e la branca della scienza che si occupa di questo tema è chiamata Computer Vision (CV). La CV è un campo dell'informatica, che sfrutta anche tecniche di intelligenza artificiale (AI), che consente ai computer e ai sistemi automatici di ricavare informazioni da immagini digitali e video e di intraprendere azioni o generare segnalazioni sulla base di tali informazioni. Se l'AI permette ai computer di "pensare", la CV permette loro di "vedere".

Recenti sviluppi legati all'emergere delle tecnologie basate su AI

Le radici della CV risalgono agli anni '50 e '60, quando i ricercatori iniziarono a esplorare l'idea di insegnare ai computer a comprendere e interpretare i dati visivi. Nel corso del tempo si è passati da implementare semplici algoritmi per il riconoscimento di elementi salienti di una immagine fino ad arrivare alla rivoluzione del deep learning.

La svolta si ebbe con la diffusione delle tecniche di Deep Learning (DL), intorno al 2010, ed in particolare con l'avvento delle Convolutional Neural Networks (CNN), una classe di reti neurali artificiali che utilizzano operazioni di convoluzione per estrarre caratteristiche rilevanti dai dati di input attraverso diversi strati. Questo approccio consente di ridurre la dimensionalità dei dati, preservando al contempo le informazioni essenziali per il compito di apprendimento, come il riconoscimento di pattern visivi. Architetture innovative ed efficienti come AlexNet [Krizhevsky et al., 2012], VGGNet [Simonyan et. al., 2014] e ResNet [He et al.,

2016] aprirono la strada a una vasta gamma di applicazioni, tra cui il rilevamento e il riconoscimento di oggetti e volti. In Figura 1 vengono riportate, in ordine cronologico, le tappe fondamentali dello sviluppo di modelli di AI per la CV.



[1]

Al fine di realizzare modelli addestrati per compiti specifici e ottenere migliori prestazioni vengono comunemente usate tecniche di transfer learning, come il fine-tuning dei modelli pre-addestrati. Ad esempio, è comune utilizzare reti nate per il riconoscimento di oggetti e specializzarle mediante fine-tuning per riconoscere uno specifico insieme di oggetti.
Una categoria particolare di reti neurali sono quelle generative che hanno l'obiettivo di generare nuovi dati. Ad esempio, le Generative Adversarial Networks (GAN) [Goodfellow et al., 2014] sono state sviluppate e impiegate per generare immagini e video realistici.

Più di recente, i meccanismi di attenzione, come quelli nei modelli Transformer (ViT), sono stati applicati ai compiti di visione artificiale.

La grande diffusione delle nuove tecniche di intelligenza artificiale si deve anche allo sviluppo di hardware specializzato, come le Graphics Processing Units (GPU) e le Tensor Processing Units (TPU) ha portato ad una accelerazione nel processo di addestramento delle reti neurali consentendo l'utilizzo di grandi quantità di dati e portando alla generazione di modelli più complessi e performanti.

Transformer

Un Transformer è un tipo di architettura di rete neurale progettata per gestire dati sequenziali (ad esempio il testo) ed è particolarmente efficace nei compiti di elaborazione del linguaggio naturale (NLP, Natural Language Processing) [Vaswani et al. 2017]. La caratteristica distintiva dei Transformers è il meccanismo di attenzione, che consente al modello di attribuire pesi diversi a diverse parti della sequenza di input, migliorando la capacità di catturare le dipendenze a lungo raggio e le relazioni tra le parole. Un Transformer è costituito da strati di encoder e decoder, ognuno dei quali contiene meccanismi di attenzione e reti neurali feed-forward. Questo design consente una parallelizzazione efficiente e ha portato a significativi miglioramenti nelle performance rispetto ai modelli sequenziali tradizionali come LSTM (Long Short-Term Memory) e GRU (Gated Recurrent Unit).

Il meccanismo di auto-attenzione (self-attention) costituisce l'elemento centrale dei modelli Transformer. Questa funzionalità è particolarmente importante per i compiti di comprensione del linguaggio, dove è fondamentale considerare il contesto

[1] Figura 1. Timeline dell'evoluzione delle reti neurali per la computer vision#

complessivo. Infatti, esso consente al modello di determinare quanto sia rilevante ogni singola parte dell'input in relazione a tutte le altre parti. La formula per l'attenzione è:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

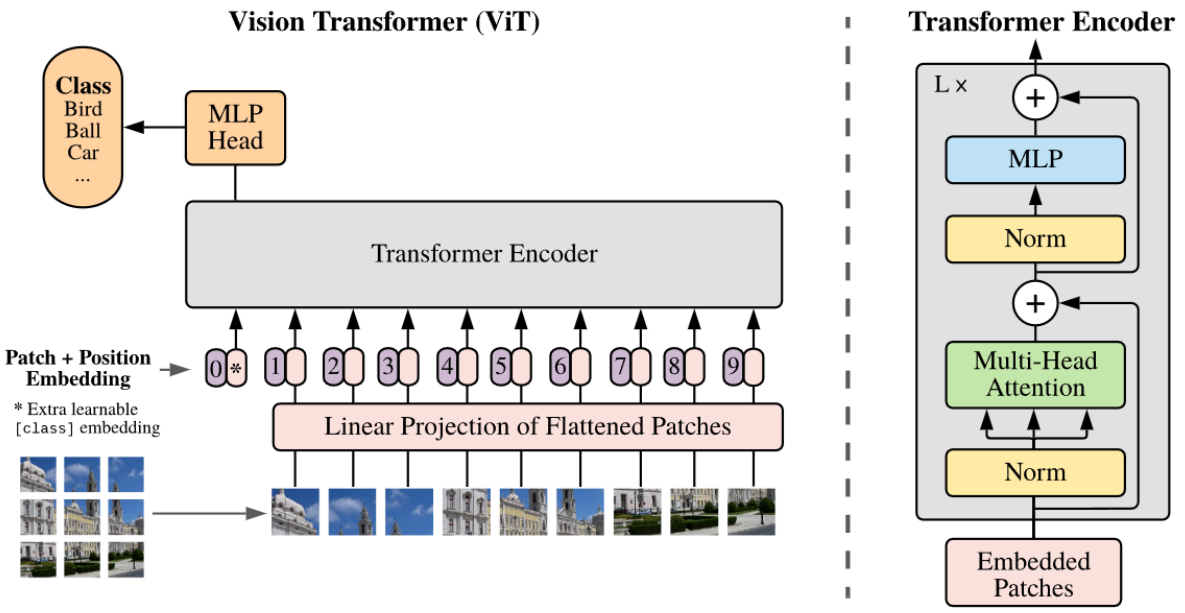
dove Q (query), K (key), e V (value) sono matrici derivanti dall'input, d_k è la dimensione delle key. Il meccanismo di self-attention consente al modello di esaminare simultaneamente tutte le parole di una sequenza, migliorando la capacità di apprendere relazioni a lungo raggio. Questo rappresenta un avanzamento significativo rispetto alle tradizionali reti neurali ricorsive, che elaborano le parole in modo sequenziale.

I vision Transfomers (ViT)

Il modello Vision Transformer (ViT) è stato proposto nel famoso lavoro "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" [Dosovitskiy et al. 2020]. Gli autori hanno addestrato un encoder Transformer utilizzando ImageNet (un grande database di immagini utilizzato per addestrare reti neurali) , ottenendo risultati molto buoni rispetto alle ben note architetture convoluzionali.

Sebbene l'architettura Transformer sia diventata lo standard de facto per i compiti di elaborazione del linguaggio naturale, le sue applicazioni alla visione artificiale rimangono limitate. Nella visione, l'attenzione viene applicata in combinazione con le reti convoluzionali oppure utilizzata per sostituire alcuni componenti delle reti convoluzionali mantenendo la loro struttura complessiva. Quando pre-addestrato su grandi quantità di dati un modello ViT raggiunge risultati eccellenti rispetto alle reti convoluzioni allo stato dell'arte richiedendo significativamente meno risorse computazionali per l'addestramento.

La Figura 2 illustra la struttura del ViT, mostrando come le immagini vengono suddivise in piccole porzioni o segmenti chiamati patch, trasformate in rappresentazioni numeriche chiamate embedding lineari tramite una trasformazione lineare che proietta i patch delle immagini in uno spazio di dimensioni fisse, facilitando l'elaborazione da parte del modello. Questi embedding lineari sono combinati con embedding posizionali, che sono vettori aggiunti per fornire informazioni sulla posizione di ciascun elemento nella sequenza. Successivamente, le immagini vengono elaborate da un Transformer standard per la classificazione.



[2]

Gli autori dell'articolo [Liu et al., 2021] propongono un modello ViT chiamato Swin (Shifted Windows) per affrontare le sfide derivanti dall'adattare il Transformer dal linguaggio alla visione. Le difficoltà derivano principalmente dalle differenze che esistono tra i due domini, come le variazioni nella scala degli elementi presenti nelle immagini rispetto alle parole nel testo. Per affrontare queste differenze, il modello Swin propone un Transformer gerarchico la cui rappresentazione è calcolata con finestre mobili. Questa soluzione porta ad una maggiore efficienza limitando il calcolo dell'auto-attenzione a finestre locali non sovrapposte, consentendo allo stesso tempo la connessione tra finestre. Questa architettura gerarchica ha la flessibilità di modellare a varie scale e ha una complessità computazionale lineare rispetto alla dimensione.

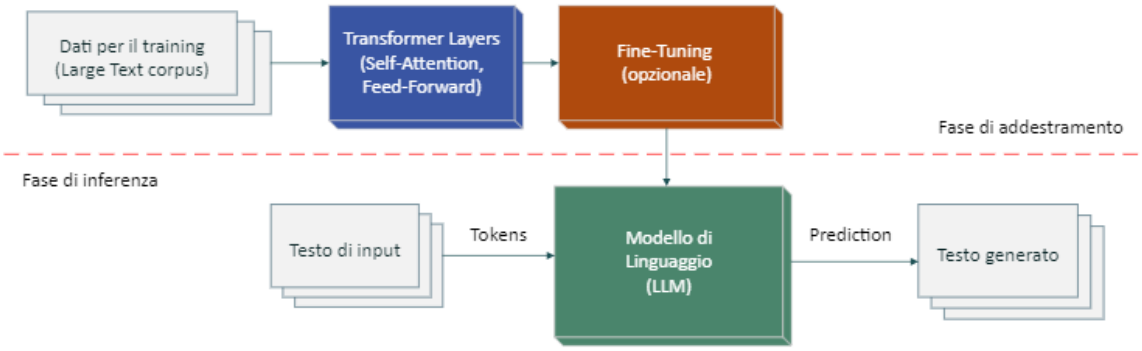
[2] Figura 2 - Modello architetturale del ViT [Dosovitskiy et al. 2021]

C) Modelli di apprendimento basati sul linguaggio

La famiglia dei modelli di apprendimento basati sul linguaggio rappresenta un avanzamento significativo nell'intelligenza artificiale, combinando capacità di comprensione e generazione di diverse modalità di dati. All'interno di questa famiglia, i modelli di linguaggio di grandi dimensioni o Large Language Model (LLM), come GPT-4, si concentrano sull'elaborazione del linguaggio naturale, eccellendo nella generazione di testo e nella comprensione del contesto linguistico. I modelli che combinano visione e linguaggio, chiamati Vision-Language Model (VLM), d'altra parte, sono una categoria specializzata di modelli multimodali che integrano input testuali e visivi e generano output testuali. Infine, i modelli multimodali di grandi dimensioni, detti Multimodal Large Language Model (MLLM), hanno la capacità di lavorare con più modalità, non solo testo e immagini, e hanno performance superiori nelle abilità di ragionamento. Questi modelli possono, ad esempio, descrivere immagini in linguaggio naturale o generare immagini a partire da descrizioni testuali, ampliando notevolmente le possibilità applicative dell'intelligenza artificiale in settori come l'analisi dei media, la realtà aumentata e la robotica.

Large Language Model (LLM)

I modelli di linguaggio di grandi dimensioni sono sistemi sofisticati che mirano a comprendere, interpretare e generare testo che somiglia strettamente al linguaggio umano. Questi modelli hanno subito significativi avanzamenti, passando da semplici strutture basate su regole a complesse reti neurali capaci, elaborando vasti set di dati di addestramento, di produrre testo quasi indistinguibile da quello umano.



[3]

I componenti rappresentati in Figura 3 costituiscono i pilastri fondamentali di qualsiasi architettura di modelli di linguaggio di grandi dimensioni (LLM). Il sistema può essere suddiviso in due parti: la fase di addestramento (training) che produce il modello e la fase di inferenza, che utilizza il modello prodotto per specifici task.

I dati di addestramento, costituiti da ampi corpus di testo, sono utilizzati per l'addestramento del modello. Questi dataset, che contengono una vasta gamma di esempi testuali, aiutano il modello ad apprendere e generalizzare le strutture e i contenuti linguistici necessari.

Il blocco Transformer Layers costituisce il nucleo dell'architettura del LLM. Questi strati utilizzano meccanismi di self-attention e reti neurali feed forward per elaborare le informazioni. Gli strati di auto-attenzione consentono al modello di integrare e valutare informazioni provenienti da diverse posizioni del testo, migliorando così la coerenza e la comprensione complessiva del contenuto.

Il fine-tuning è un processo opzionale che consente di adattare ulteriormente il modello per applicazioni specifiche. Questo passaggio, che avviene dopo l'addestramento iniziale, permette di specializzare il modello su compiti o domini particolari, migliorando così la sua performance in contesti mirati.

Dalla fase di addestramento viene ricavato un modello (LLM) che incorpora la conoscenza del sistema, basata sui dati di addestramento utilizzati per generarlo attraverso il Transformer Layer. Generalmente, questo modello viene salvato in un file con un formato specifico, che può essere trasferito e caricato su un sistema differente.

Nella fase di inferenza il modello pre-addestrato viene utilizzato per risolvere task specifici.

Il testo di input, dopo essere stato tokenizzato, rappresenta il prompt fornito al modello. La tokenizzazione è un passaggio essenziale che trasforma il testo in una sequenza di unità più piccole, chiamate "token", che il modello può elaborare. Questo processo è cruciale per la comprensione e l'analisi del testo da parte del modello.

La generazione dell'output rappresenta la fase in cui il modello produce una prediction, ovvero una risposta, una previsione o un risultato in risposta al prompt fornito. Utilizzando le informazioni elaborate, il modello genera un testo che soddisfa le esigenze specifiche del prompt.

Le funzioni principali dei LLM riguardano la generazione, la traduzione e la classificazione di diverse tipologie di testo. Per esempio, questi modelli possono scrivere testi originali oppure suggerire modifiche per migliorare contenuto e stile. Gli LLM possono anche rispondere a domande specifiche dopo aver analizzato archivi digitali passati in input. Inoltre, possono classificare

[3] Figura 3 - Componenti di alto livello di un modello di linguaggio

testi con significati o sentimenti simili. Gli usi di questa tecnica includono la sentiment analysis, la determinazione delle relazioni tra i testi e la ricerca di documenti. Un'altra interessante funzione degli LLM è quella della generazione di codice in diversi linguaggi di programmazione a partire da istruzioni fornite in linguaggio naturale.

Vision-Language Models (VLM)

I Vision-Language Model sono una classe di modelli di LLM che trattano non solo il testo, ma, come suggerisce il nome, comprendono anche le immagini. Questi modelli apprendono le correlazioni tra visione e linguaggio a partire da coppie immagine-testo reperibili su vasta scala dal web, permettendo previsioni zero-shot[4], senza fine-tuning, su vari compiti di riconoscimento visivo utilizzando un singolo modello pre-addestrato. La Figura 4 illustra il paradigma di "apprendimento" utilizzato per la generazione dei VLM.

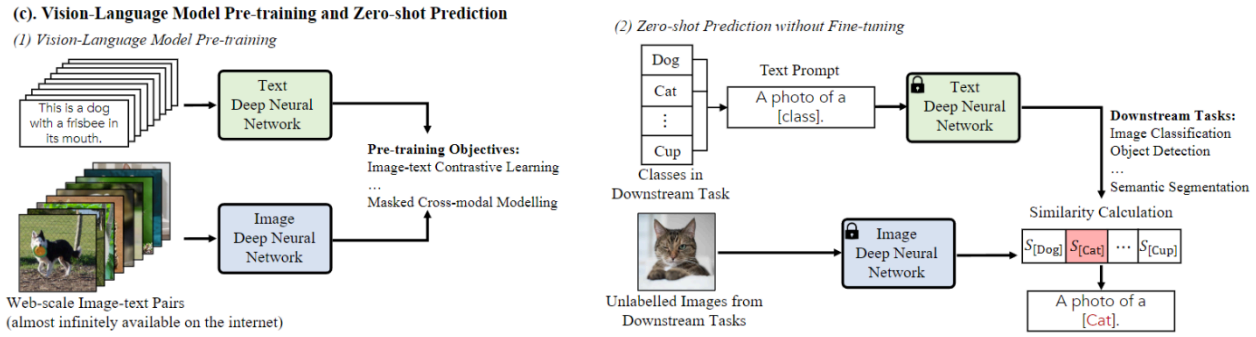


Figura 4 – Rappresentazione grafica del paradigma di addestramento dei VLM che utilizza previsioni zero-shot senza bisogno di un fine-tuning specifico per il compito. [Zhang J. et al. 2024][5]

Esistono diverse famiglie di VLM ed è possibile classificare i modelli differenziando in base al metodo di addestramento. È possibile identificare quattro diversi paradigmi di addestramento. L'addestramento contrastivo è una strategia comunemente utilizzata che impiega coppie di esempi positivi e negativi. Il modello VLM viene addestrato per prevedere rappresentazioni simili per le coppie positive e rappresentazioni diverse per quelle negative. Il masking è un'altra strategia che può essere sfruttata per addestrare i VLM, ricostruendo le patch mancanti a partire da una didascalia di testo non mascherata. Allo stesso modo, mascherando parole in una didascalia, è possibile addestrare un VLM a ricostruire quelle parole a partire da un'immagine non mascherata. Mentre la maggior parte di questi approcci sfrutta rappresentazioni intermedie o ricostruzioni parziali, i VLM generativi sono addestrati in modo tale da poter generare intere immagini o didascalie molto lunghe. Data la natura di questi modelli, essi sono spesso i più costosi da addestrare. I VLM basati su backbone preaddestrati spesso sfruttano LLM open-source per apprendere una mappatura tra un encoder di immagini (che potrebbe essere anch'esso preaddestrato) e l'LLM. È importante sottolineare che questi paradigmi non sono mutuamente esclusivi; molti approcci si basano su una combinazione di criteri contrastivi, masking e generativi.

Sebbene diversi studi abbiano già esteso i LLM alla visione, il collegamento tra linguaggio e visione non è ancora completamente risolto. Ad esempio, la maggior parte dei modelli fatica a comprendere le relazioni spaziali o a contare. Molti VLM mancano anche di una comprensione degli attributi e dell'ordine. Spesso ignorano parte del prompt di input, richiedendo significativi sforzi di

[4] Lo zero-shot è una tecnica di apprendimento automatico in cui un modello è in grado di eseguire un compito, come la classificazione o il riconoscimento, senza aver visto alcun esempio di addestramento specifico per quel compito. Invece di apprendere direttamente dai dati annotati per il compito target, il modello utilizza conoscenze acquisite da altri compiti correlati o da un'ampia base di dati generale. Questo approccio consente al modello di generalizzare meglio e di affrontare nuovi compiti senza la necessità di dati di addestramento specifici.

[5] Figura 4 – Rappresentazione grafica del paradigma di addestramento dei VLM che utilizza previsioni zero-shot senza bisogno di un fine-tuning specifico per il compito. [Zhang J. et al. 2024]

ingegneria del prompt per produrre il risultato desiderato. Alcuni di essi possono anche generare contenuti inesistenti o non pertinenti. Di conseguenza, lo sviluppo di modelli affidabili è ancora un'area di ricerca molto attiva [Bordes et al. 2024].

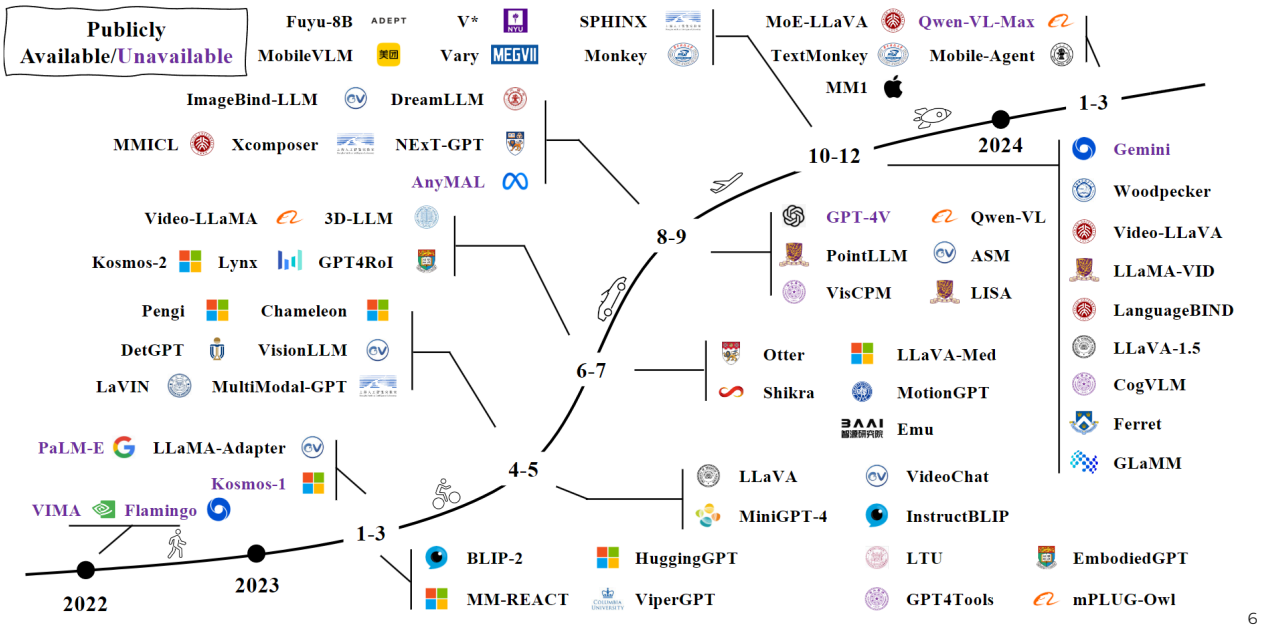
Uno dei più noti VLM è Contrastive Language–Image Pre-training (CLIP) [Radford et al. 2021]. Il modello originale, addestrato su 400 milioni di coppie di didascalie e immagini raccolte dal web, ha dimostrato notevoli capacità di trasferimento zero-shot nella classificazione.

I modelli MLLM hanno un'ampia gamma di applicazioni grazie alla loro capacità di gestire e interpretare dati provenienti da diverse modalità, come testo, immagini, audio e video. Essi possono migliorare significativamente le capacità degli assistenti virtuali, integrando la comprensione del linguaggio naturale con il riconoscimento visivo e sonoro per fornire risposte più precise e contestualizzate. In ambito medico, gli MLLM possono combinare immagini mediche con descrizioni testuali per supportare diagnosi più accurate e decisioni cliniche. Inoltre, questi modelli sono fondamentali per i veicoli a guida autonoma, dove integrano dati da sensori visivi e audio per migliorare la percezione e l'interazione con l'ambiente. Infine, nei sistemi di raccomandazione, gli MLLM possono unire informazioni visive e testuali per offrire suggerimenti più rilevanti e personalizzati.

Large Language Model Multimodali (MLLM)

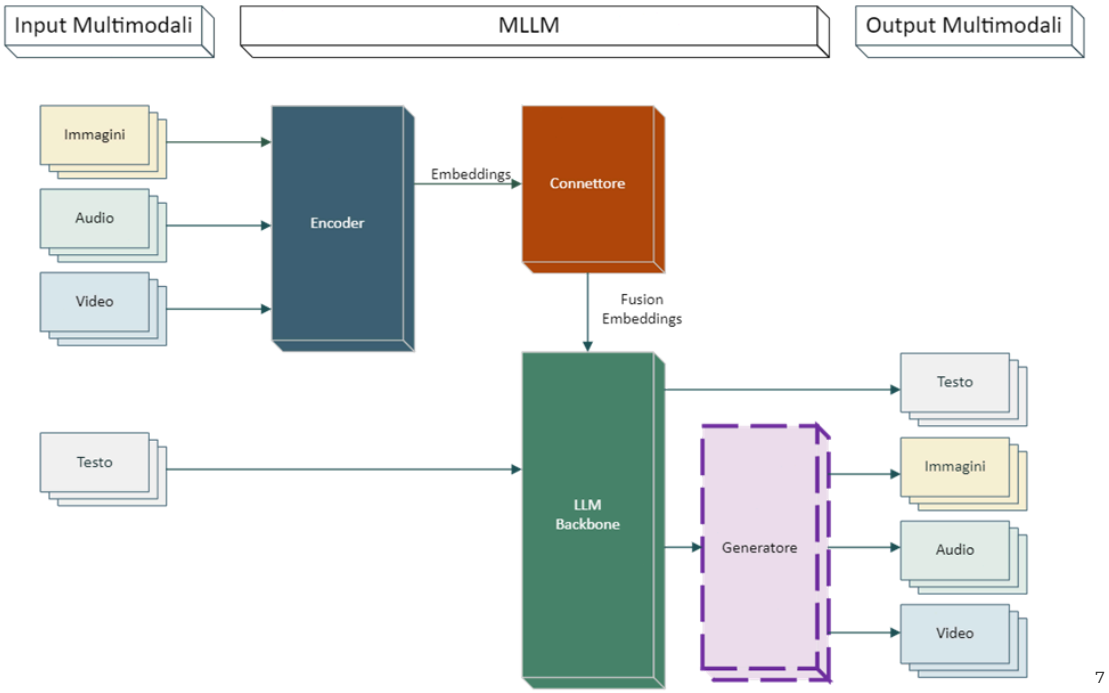
I MLLM sono modelli basati su LLM con la capacità di ricevere, elaborare e generare output utilizzando informazioni multimodali. Questi modelli riducono i costi computazionali dell'addestramento da zero sfruttando efficacemente la conoscenza pre-addestrata di ciascuna modalità. Gli MLLM ereditano le capacità cognitive dei LLM, evidenziando numerose caratteristiche di rilievo, tra cui una robusta capacità di generazione del linguaggio e abilità avanzate di transfer learning. Inoltre, stabilendo forti connessioni tra modelli basati su diverse modalità, gli MLLM possono elaborare input provenienti da input differenti, ampliando in modo significativo il loro ambito applicativo.

I MLLM sono l'ultima frontiera dell'intelligenza artificiale su cui c'è grande fervore dal punto di vista industriale e scientifico, in particolare a seguito del debutto di GPT-4V [Achiam et al. 2023] e Gemini [Team et al. 2023], che hanno dimostrato impressionanti capacità di comprensione e generazione multimodale. La Figura 5 mostra una timeline dello sviluppo dei principali MLLM dal 2022 al 2024, evidenziando l'evoluzione e le date di introduzione dei vari modelli.



Inizialmente la ricerca si è concentrata sulla comprensione dei contenuti multimodali e sulla generazione di testo, includendo compiti come la comprensione immagine-testo, esemplificata da progetti come BLIP-2 [Li et al., 2023], LLaVA [Liu et al. 2024], MiniGPT-4 [Zhu et al. 2023]; la comprensione video-testo e la comprensione audio-testo. Successivamente, le capacità dei MLLM sono state ampliate per supportare output specifici per ogni modalità. Questo comprende compiti che producono output immagine-testo e voce/audio-testo. Le recenti iniziative di ricerca si sono concentrate sull'imitazione della conversione tra diverse modalità in modo simile a quello umano, aprendo la strada verso l'intelligenza artificiale generale [Zhang D. et al. 2024].
La struttura di un MMLM può essere suddivisa in 3 blocchi principali, come descritto in Figura 6: l'Encoder, l'LLM Backbone e il Connettore [Yin et a. 2023]. Se si desidera ottenere un output in una forma diversa da quella testuale, è possibile aggiungere in coda un Generatore.

[6] Figura 5 - Timeline dell'evoluzione dei principali MLLM [Yin et al. 2023]



[7]

L'encoder è il componente che si relaziona direttamente con il mondo reale. Il suo compito è trasformare gli input in una forma che sia efficiente per l'elaborazione dei blocchi successivi. I dati forniti in input possono essere di diverso tipo, come immagini, file audio o video. La funzione principale dell'encoder è quella di tradurre questi dati in una rappresentazione compatta e altamente rappresentativa, spesso chiamata features o anche embeddings.

Tipicamente, gli encoder non vengono creati da zero, ma vengono utilizzati modelli pre-addestrati. È stato dimostrato infatti che il transfer learning[8] è un approccio che consente di ottenere performance comparabili se non migliori rispetto a quelli di modelli addestrati da zero [Kornblith et al. 2019].

In letteratura, esistono numerosi modelli adatti a essere utilizzati come encoder. Per quanto riguarda l'elaborazione delle immagini, la maggior parte dei sistemi è basata su modelli di tipo CLIP (Contrastive Language Image Pre-training), che sono in grado di comprendere e generare associazioni tra testo e immagini [Radford et al. 2021]. Utilizzando un approccio contrastivo, CLIP allinea rappresentazioni testuali e visive in uno spazio di embedding comune, apprendendo a distinguere coppie di testo e immagini corrette da quelle errate. Questo metodo sfrutta grandi dataset non supervisionati, migliorando la capacità del modello di eseguire una vasta gamma di compiti multimodali, come il riconoscimento delle immagini, la generazione di didascalie e la ricerca visiva basata su testo. Negli anni sono state sviluppate diverse varianti di CLIP che ne migliorano le performance, come EVA-CLIP [Sun et al. 2023] e Minigpt [Zhu et al. 2023]. Diversi studi hanno inoltre dimostrato che le capacità di tali sistemi dipendono dalla risoluzione dell'input. È perciò possibile migliorare le performance, aumentando la risoluzione degli input [Li et al., 2024]. Per questo motivo, la risoluzione rappresenta un fattore cruciale nella scelta dell'encoder da utilizzare. Alcuni esempi di encoders e le relative risoluzioni sono riportati in Tabella 1.

Encoder	Risoluzione
OpenCLIP-ConvNext-L [Cherti et al., 2023]	320
CLIP-ViT-L/14 [Radford et al., 2021]	224/336
EVA-CLIP-ViT-G/14 [Sun et al., 2023]	224

[7] Figura 6 - Architettura generica e componenti principali di un MLLM

[8] Il transfer learning è una tecnica di apprendimento automatico in cui un modello addestrato su un compito viene riutilizzato come punto di partenza per un modello su un altro compito correlato. Questa tecnica sfrutta conoscenze acquisite da un dominio per migliorare le performance in un altro, riducendo il tempo e le risorse necessarie per l'addestramento e migliorando l'efficacia complessiva, soprattutto quando i dati disponibili sono limitati.

OpenCLIP-ViT-G/14 [Cherti et al., 2023]	224
OpenCLIP-ViT-bigG/14 [Cherti et al., 2023]	224

Tabella 1 - Esempi di modelli usati come encoders per l'elaborazione di immagini

L'LLM backbone è il cuore del sistema è ha il compito più complesso, ovvero quello di capire e ragionare sugli input ricevuti.
Anche nel caso degli LLM l'approccio più comune è quello del transfer learning. Grazie all'ampio pre-addestramento su vasti corpus web, i LLM hanno acquisito una conoscenza approfondita del mondo e dimostrano notevoli capacità di generalizzazione e ragionamento.
Un parametro importante da considerare nella scelta dell'LLM consiste nel numero di parametri utilizzati. Il numero di parametri influisce infatti sulle performance, ed è stato dimostrato che aumentando il numero di parametri è possibile migliorare le performance in diversi benchmark [Liu et al. 2024]. Alcuni esempi dei più utilizzati LLM e il range di dimensioni disponibili, sono riportati in Tabella 2.

LLM	Parametri [B]
Flan-T5-XL/XXL [Chung et al., 2024]	3/ 11
LLaMA [Touvron et al., 2023]	7/ 13/ 33/ 65
Vicuna [Chiang et al., 2023]	7/ 13/ 33
LLaMA-2 [Touvron et al., 2023]	7/ 13/ 70
Qwen [Bai et al., 2023]	1.8 / 7/ 14/ 72

Tabella 2 - Esempio di LLM open source e relative dimensioni

Dato che gli LLM restituiscono output in formato testuale, è possibile aggiungere un modulo Generatore in coda al sistema per ottenere i risultati in un formato diverso. Tale modulo è capace di trasformare gli embedding prodotti dall'LLM in un formato specifico, utilizzando appositi decoder.

Il terzo componente di un MLLM è il Connettore, il quale ha il compito di facilitare la comunicazione tra l'encoder e l'LLM. Sebbene l'encoder possa processare diverse modalità di input, l'LLM è in grado di elaborare solo input di tipo testuale. Pertanto, il connettore ha la funzione di trasformare gli embedding estratti dall'encoder in una forma adatta all'input dell'LLM, fondendo tra loro informazioni con modalità differenti.

Esistono due tipologie di connettori, token-level e feature-level, che differiscono nel metodo utilizzato per la fusione.
Nel caso della fusione a livello di token, le caratteristiche generate dagli encoder vengono trasformate in token e concatenate con i token testuali prima di essere inviate agli LLM. Esistono due approcci per il concatenamento, il primo riduce la dimensione degli embedding delle immagini utilizzando metodi query-based [Li et al. 2023], il secondo, utilizza strutture basate su MLP[9] [Liu et al. 2024].

La fusione a livello di feature inserisce moduli aggiuntivi che consentono una profonda interazione e fusione tra le caratteristiche testuali e quelle visive [Alayrac et al. 2022, Wang et al. 2023]. Tale approccio può essere realizzato inserendo strati di cross-attention tra gli strati dei Transformer congelati[10] degli LLM, arricchendo così le caratteristiche linguistiche con segnali visivi

[9] Un MLP (Multi-Layer Perceptron) è un tipo di rete neurale artificiale composta da più strati di nodi (neuroni) completamente connessi. Gli MLP sono usati per apprendere e modellare relazioni complesse tra input e output, grazie alla loro capacità di apprendere funzioni non lineari. Sono particolarmente utili in problemi di classificazione e regressione.
[10] ovvero che non modifica i pesi e i parametri dell'encoder durante l'addestramento del modello principale

esterni. Un altro approccio consiste nell'integrare un modulo esperto visivo in ciascun livello del Transformer, permettendo un'interazione e una fusione duale tra caratteristiche visive e linguistiche. Per migliorare le performance, la matrice dei pesi del modulo introdotto viene inizializzata a partire dall'LLM pre-addestrato. Inoltre, è possibile introdurre prompt apprendibili negli strati del Transformer, che vengono prima incorporati con conoscenze visive e poi concatenati alle caratteristiche testuali come prefissi.

Esempi di MLLM

Di seguito sono elencati alcuni dei principali e più diffusi MLLM. È stata dedicata una sezione a parte a questo argomento in quanto i MLLM rappresentano l'avanguardia dell'intelligenza artificiale e attirano un notevole interesse da parte della comunità scientifica. In questa sezione sono descritti i modelli, includendo dettagli relativi alle loro caratteristiche tecniche, agli sviluppatori e ai diritti di utilizzo.

Flamingo (2022, DeepMind)

- **Input**: testo, immagini
- **Output**: testo
- **Dettagli**: I modelli Flamingo sfruttano due modelli pre-addestrati: un modello di visione che può "percepire" scene visive e un LLM che esegue una forma basilare di ragionamento. Tra questi modelli vengono aggiunti nuovi componenti architetturali per connetterli in modo da preservare le conoscenze accumulate durante il pre-addestramento computazionalmente intensivo.
- **Sito Web**: https://github.com/mlfoundations/open_flamingo
- **Termini di utilizzo**: Licenza MIT, si concede il permesso, gratuitamente, a qualsiasi persona ottenga una copia di questo software e dei file di documentazione associati (il "Software"), di trattare il Software senza restrizioni, inclusi, senza limitazione, i diritti di usare, copiare, modificare, unire, pubblicare, distribuire, concedere in sublicenza e/o vendere copie del Software, e di permettere alle persone a cui il Software è fornito di fare lo stesso, a condizione che: L'avviso di copyright di cui sopra e questo avviso di permesso siano inclusi in tutte le copie o porzioni sostanziali del Software.
- **Biblio**: [Alayrac et al. 2022]

BLIP-2 (2023, Salesforce)

- **Input**: testo, immagini
- **Output**: testo, immagini
- **Dettagli**: BLIP-2 è una strategia di pre-addestramento versatile ed efficiente che utilizza encoder di immagini e modelli di linguaggio pre-addestrati e "congelati", ovvero che non modifica i pesi e i parametri dell'encoder durante l'addestramento del modello principale. Questo metodo colma il divario tra visione e linguaggio attraverso un Querying Transformer leggero, pre-addestrato in due fasi. Nella prima fase, viene avviato l'apprendimento della rappresentazione visione-linguaggio utilizzando un encoder di immagini congelato. Nella seconda fase, viene avviato l'apprendimento generativo visione-linguaggio utilizzando un modello di linguaggio congelato. Nonostante abbia un numero significativamente inferiore di parametri addestrabili rispetto ad altri metodi, BLIP-2 raggiunge risultati di eccellenza su diversi compiti che combinano visione e linguaggio.
- **Sito Web**: https://github.com/salesforce/LAVIS/tree/main/projects/blip2
- **Termini di utilizzo**: Licenza MIT, si concede il permesso, gratuitamente, a qualsiasi persona ottenga una copia di questo software e dei file di documentazione associati (il "Software"), di trattare il Software senza restrizioni, inclusi, senza limitazione, i diritti di usare, copiare, modificare, unire, pubblicare, distribuire, concedere in sublicenza e/o vendere copie

del Software, e di permettere alle persone a cui il Software è fornito di fare lo stesso, a condizione che: L'avviso di copyright di cui sopra e questo avviso di permesso siano inclusi in tutte le copie o porzioni sostanziali del Software.

- **Biblio**: [Li et al. 2023]

Kosmos-2 (2023, Microsoft)

- **Input**: testo, immagini
- **Output**: testo, immagini
- **Dettagli**: Oltre alle capacità classiche dei MLLM (ad esempio, percepire modalità generali, seguire istruzioni e apprendere nel contesto), KOSMOS-2 permette all'utente di indicare direttamente l'oggetto o la regione di interesse nell'immagine piuttosto che inserire descrizioni testuali dettagliate per riferirsi ad esso. Il modello è in grado di generare risposte visive (cioè, riquadri di delimitazione), che assolvono meglio i compiti di visione-linguaggio, le risposte visive sono più accurate e risolvono l'ambiguità della coreferenza rispetto alle risposte testuali.
- **Sito Web**: https://github.com/microsoft/unilm/tree/master/kosmos-2
- **Termini di utilizzo:** Il modello è destinato a scopi accademici e di ricerca. (https://huggingface.co/spaces/ydshieh/Kosmos-2)
- **Biblio**: [Peng et al. 2023]

LLaVA (2023, University of Wisconsin-Madison)

- **Input**: testo, immagini.
- **Output**: testo, immagini.
- **Dettagli**: LLaVA (Large Language and Vision Assistant) è un modello AI multimodale open-source che eguaglia le capacità di chat di GPT-4 e stabilisce un nuovo record in Science QA, dimostrando una comprensione avanzata visivo-linguistica. LLaVA è stato addestrato utilizzando una tecnica chiamata "instruction tuning", in cui GPT-4 è stato utilizzato per generare compiti multimodali sintetici che coinvolgono testo e immagini (novità del 2023). LLaVA ha imparato da questi diversi esempi generati da GPT-4 senza supervisione umana diretta.
- **Sito web**: https://llava-vl.github.io
- **Termini di utilizzo**: I dati, il codice e il checkpoint sono destinati e concessi in licenza solo per uso di ricerca. (https://llava-vl.github.io/)
- **Biblio**: [Liu et al. 2024]

Gemini 1.5 (2024, Google)

- **Input**: testo, immagini.
- **Output**: testo, immagini.
- **Dettagli**: Gemini è una famiglia di modelli linguistici di grandi dimensioni sviluppata da Google. Presentato per la prima volta nel dicembre 2023 è disponibile in tre varianti ottimizzate: Gemini Ultra (il più grande), Gemini Pro (per scalabilità) e Gemini Nano (per attività su dispositivi di limitata capacità). Il nome Gemini è un riferimento al segno zodiacale dei Gemelli, che rappresenta i "Gemelli" nella mitologia greca. Questo è appropriato data la natura duale di Gemini come modello linguistico altamente capace che può anche elaborare e generare dati multimodali come immagini, audio e video.
- **Sito web:** https://gemini.google.com/, https://deepmind.google/technologies/gemini/

- **Termini di utilizzo:** Sebbene conceda all'utente l'autorizzazione per l'utilizzo dei suoi servizi, Google mantiene tutti i diritti di proprietà intellettuale che detiene in merito ai propri servizi. (https://policies.google.com/terms, https://ai.google.dev/gemini-api/terms?hl=it)
- **Biblio:** [Team et al. 2023]

Qwen-VL (2024, Alibaba Cloud)

- **Input**: testo, immagini.
- **Output**: testo, immagini.
- **Dettagli**: Qwen-VL è un modello AI multimodale open-source che combina capacità linguistiche e visive. È un'estensione del modello linguistico Qwen, progettato per superare le limitazioni nella generalizzazione multimodale. Le versioni recentemente aggiornate (Qwen-VL-Plus e Qwen-VL-Max) presentano un miglioramento nel ragionamento visivo, una migliore analisi dei dettagli nelle immagini e nel testo, e supporto per immagini ad alta risoluzione. Dopo il suo lancio, Qwen-VL è rapidamente salito in cima alla classifica OpenVLM ma è stato superato da altri modelli più potenti, in particolare GPT-4o.
- **Sito Web**: https://qwenlm.github.io/blog/qwen-vl/
- **Termini di utilizzo:** Il codice sorgente è concesso con Licenza Apache 2.0. I ricercatori e gli sviluppatori sono liberi di utilizzare i codici e i modelli sia di Qwen che di Qwen-Chat. Per il loro utilizzo commerciale è necessario consultare il Contratto di Licenza allegato a ciascun modello. (https://github.com/QwenLM/Qwen)
- **Biblio**: [Bai et al., 2023]

Claude 3.5 Sonnet (2024, Anthropic)

- **Input**: testo, immagini.
- **Output**: testo, immagini.
- **Dettagli**: Claude 3.5 Sonnet è un sistema multimodale che può comprendere e generare testo, immagini, audio e altri formati di dati. Eccelle nell'analisi approfondita, nella ricerca, nella generazione di ipotesi e nell'automazione dei compiti in vari settori come la finanza, le scienze della vita e l'ingegneria del software. Anthropic utilizza una tecnica chiamata "recursive reward modeling" che prevede l'uso di una versione precedente di Claude per fornire feedback e ricompense per gli output del modello.
- **Sito web**: https://claude.ai
- **Termini di utilizzo**: la versione di prova non può essere utilizzata per scopi commerciali (https://www.anthropic.com/legal/consumer-terms)

GPT-4o (2024, OpenAI)

- **Input**: testo, immagini, audio (beta), video (beta).
- **Output**: testo, immagini.
- **Descrizione**: GPT-4o sta per "GPT-4 Omni", con "Omni" che si riferisce alle sue capacità multimodali attraverso le modalità di testo, visione e audio. Si tratta di un unico modello unificato che può comprendere e generare qualsiasi combinazione di input/output di testo, immagini, audio e video. GPT-4o impiega un approccio "multi-modal chain of thought", in cui prima valuta come suddividere un problema in una serie di passaggi attraverso diverse modalità (testo, visione, audio), e poi esegue quei passaggi per arrivare alla soluzione finale.
- **Sito web:** https://chatgpt.com/

- **Termini di utilizzo:** i termini di utilizzo si applicano all'uso di ChatGPT, DALL·E e degli altri servizi di OpenAI, insieme a qualsiasi applicazione software associata, tecnologia e siti web, inclusi l'uso personale e non commerciale dei servizi da parte dei consumatori. (https://openai.com/policies/terms-of-use/)
- **Biblio:** [Achiam et al. 2023]

Nome	Sviluppatori	Anno	Termini di utilizzo
Flamingo	DeepMind	2022	Open (Licenza MIT)
BLIP-2	Salesforce	2023	Open (Licenza MIT)
Kosmos-2	Microsoft	2023	Open (Scopi di ricerca)
LLaVA	University of Wisconsin-Madison	2023	Open (Scopi di ricerca)
Gemini 1.5	Google	2024	Close (solo utilizzo)
Qwen-VL	Alibaba Cloud	2024	Open (Licenza Apache 2.0)
Claude 3.5 Sonnet	Anthropic	2024	Open (No scopi commerciali)
GPT-4°	OpenAI	2024	Close (solo utilizzo)

Tabella 3: Tabella riassuntiva dei modelli MLLM

Conclusioni

Il documento discute l'evoluzione e lo stato attuale del campo della CV, con particolare attenzione all'influenza dell'AI e dei modelli di deep learning. La CV mira a dotare i computer della capacità di interpretare e analizzare le immagini e i video, simile alla visione umana. Storicamente, le tecniche di CV sono passate da algoritmi di base a metodi avanzati come le CNN, che hanno rivoluzionato il campo grazie alla loro capacità di estrarre e riconoscere pattern visivi complessi.

Il documento introduce anche i Transformer e i Vision Transformer. I Transformer, originariamente sviluppati per il processamento del linguaggio naturale, utilizzano il meccanismo di self-attention per gestire dati sequenziali, migliorando la capacità di comprendere contesti complessi. I Vision Transformer estendono questi principi alla visione artificiale, dimostrando eccellenti risultati nella classificazione delle immagini con minori risorse computazionali rispetto alle CNN.

L'avvento dei LLM ha segnato una svolta significativa nel settore dell'intelligenza artificiale, grazie alla loro capacità di comprendere e generare linguaggio naturale con una precisione senza precedenti. Questo progresso ha avuto un impatto profondo anche nella computer vision, poiché i LLM sono stati integrati con modelli visivi per migliorare l'interpretazione e l'analisi delle immagini, facilitando una sinergia più avanzata tra testo e visione e aprendo nuove possibilità per applicazioni intelligenti e interattive.

Inoltre, i modelli multimodali, come i VLM e i MLLM, rappresentano un'evoluzione significativa. Questi modelli integrano dati da più modalità (testo, immagini, audio), migliorando le capacità di comprensione e generazione multimodale. I VLM, ad esempio, possono effettuare previsioni zero-shot su compiti di riconoscimento visivo, mentre gli MLLM ampliano queste capacità, combinando conoscenze visive e testuali per applicazioni in vari settori, come la medicina e i veicoli autonomi.

Nonostante i VLM rappresentino una soluzione altamente sofisticata e completa per le esigenze del progetto, offrendo una vasta gamma di funzionalità per l'integrazione e l'analisi di dati visivi e testuali, è stata presa la decisione strategica di concentrare lo studio sui modelli multimodali di tipo MLLM. Questa scelta è motivata dalla superiore capacità di "ragionamento" che i MLLM dimostrano rispetto ai VLM tradizionali.

I MLLM si distinguono per la loro abilità di gestire e integrare informazioni provenienti da diverse modalità (testo e immagine) in modo più sofisticato e coeso. Questa capacità di "ragionamento" si manifesta attraverso una maggiore abilità nel comprendere e correlare concetti complessi, nel generare risposte più pertinenti e contestualmente accurate e nel migliorare le prestazioni in compiti che richiedono un'analisi profonda e una sintesi di informazioni multi-sensoriali.

La ragione principale per la quale si è deciso di focalizzarsi sugli MLLM è che questi modelli offrono un vantaggio significativo nella risoluzione di problemi complessi che coinvolgono la comprensione e la manipolazione simultanea di dati visivi e testuali. Questo approccio permette di ottenere risultati più robusti e adattabili in contesti variabili, migliorando l'efficacia complessiva del sistema in scenari reali e applicazioni pratiche. Pertanto, sebbene i VLM siano strumenti avanzati e ben equipaggiati, i MLLM offrono un potenziale superiore per le esigenze specifiche del progetto, giustificando ampiamente la decisione di concentrare lo studio su questi modelli.

Bibliografia

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. and Avila, R., 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.
- Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M. and Ring, R., 2022. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35, pp.23716-23736.
- Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C. and Zhou, J., 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966.
- Cherti, Mehdi, et al. "Reproducible scaling laws for contrastive language-image learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
- Chiang, Wei-Lin, et al. "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality." See https://vicuna. lmsys. org (accessed 14 April 2023) 2.3 (2023): 6.
- Chung, Hyung Won, et al. "Scaling instruction-finetuned language models." Journal of Machine Learning Research 25.70 (2024): 1-53.
- Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. Advances in neural information processing systems, 27.
- He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
- Kornblith, S., Shlens, J., Le, Q.V.: Do better imagenet models transfer better? In: Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 2661–2671 (2019). IEEE
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.
- Li, J., Li, D., Savarese, S. and Hoi, S., 2023, July. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning (pp. 19730-19742). PMLR.
- Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y. and Bai, X., 2024. Monkey: Image resolution and text label are important things for large multi-modal models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 26763-26773).
- Liu, H., Li, C., Li, Y. and Lee, Y.J., 2024. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 26296-26306).
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).
- Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S. and Wei, F., 2023. Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824.
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021, July. Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.
- Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

- Sun, Quan, et al. "Eva-clip: Improved training techniques for clip at scale." arXiv preprint arXiv:2303.15389 (2023).
- Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A. and Millican, K., 2023. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. and Bikel, D., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.
- Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X. and Xu, J., 2023. Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079.
- Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T. and Chen, E., 2023. A survey on multimodal large language models. arXiv preprint arXiv:2306.13549.
- Zhang, D., Yu, Y., Li, C., Dong, J., Su, D., Chu, C. and Yu, D., 2024. Mm-llms: Recent advances in multimodal large language models. arXiv preprint arXiv:2401.13601.
- Zhang, J., Huang, J., Jin, S. and Lu, S., 2024. Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence.
- Zhu, D., Chen, J., Shen, X., Li, X. and Elhoseiny, M., 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592.