Enhancing Multimodal Large Language Models with Multi-instance Visual

Prompt Generator for Visual Representation Enrichment

5 Giugno 2024

Nel paper che stai analizzando, gli autori si occupano di migliorare le **Multimodal Large Language Models (MLLMs)**, cioè quei modelli di intelligenza artificiale che possono gestire simultaneamente informazioni provenienti da diverse modalità, come immagini e testo. Questi modelli sono usati in compiti dove è necessario combinare informazioni visive e linguistiche, come la **descrizione automatica di immagini (Image Captioning)** o la **risposta a domande basate su immagini (Visual Question Answering, VQA)**.

**Contesto e problema iniziale**

Molti dei modelli MLLM all'avanguardia utilizzano una componente chiamata **Visual Prompt Generator (VPG)**. Il compito del VPG è prendere un’immagine, trasformarla in un insieme di rappresentazioni numeriche (chiamate **"visual tokens"**) e far sì che queste possano essere usate dal modello di linguaggio per svolgere compiti complessi. Tra i VPG più comuni c'è il **Q-former**, un modulo che utilizza meccanismi di attenzione per estrarre informazioni rilevanti dalle immagini e trasferirle al modello di linguaggio.

Il **Q-former** funziona bene in molti contesti, ma gli autori notano una limitazione importante: questo modulo tratta le immagini o le parti di un’immagine in modo **indipendente**, senza tenere conto delle relazioni o delle correlazioni tra le immagini che rappresentano lo stesso oggetto. Nella realtà, spesso dobbiamo considerare più immagini o porzioni della stessa immagine come collegate tra loro. Ad esempio:

* **In un sito di e-commerce**, un prodotto può avere diverse immagini che mostrano angolazioni diverse (ad esempio la vista frontale, laterale e dettagli dei materiali). Tutte queste immagini si riferiscono allo stesso prodotto, ma il Q-former non riesce a cogliere questa relazione.
* **In ambito medico**, immagini di grande dimensione, come quelle di biopsie digitali (Whole Slide Images, WSI), sono spesso suddivise in più porzioni (patches) per essere analizzate. Anche in questo caso, ogni porzione è una parte della stessa immagine, ma il Q-former tratta ogni patch in modo separato, ignorando il collegamento tra le diverse sezioni del tessuto.

**Soluzione proposta: MIVPG (Multi-instance Visual Prompt Generator)**

Per risolvere questa limitazione, gli autori propongono un nuovo componente chiamato **Multi-instance Visual Prompt Generator (MIVPG)**. Questo nuovo approccio non si limita a trattare le immagini o i patch individualmente, ma tiene conto delle **relazioni e correlazioni** tra le diverse immagini o porzioni della stessa immagine. Il MIVPG si ispira a una tecnica di machine learning chiamata **Multiple Instance Learning (MIL)**.

**Cosa fa il MIVPG e come si differenzia dal Q-former**

Nel **Multiple Instance Learning (MIL)**, anziché fare previsioni su ogni singola immagine o porzione (patch), si considera un insieme di immagini o patch come un "bag" (un insieme di istanze). Il modello fa previsioni a livello di "bag" piuttosto che a livello di singola istanza, aggregando le informazioni in modo da tenere conto delle relazioni tra le diverse immagini o patch che appartengono allo stesso campione.

Il **MIVPG** estende questo concetto di MIL e aggiunge una componente che permette di:

* **Riconoscere le correlazioni tra immagini o patch** all'interno dello stesso insieme (bag), quindi non considera le immagini come isolate, ma cerca di capire come sono collegate tra loro.
* **Aggregare l'informazione da diverse dimensioni**, che può essere l'aggregazione delle immagini (nel caso di un prodotto e-commerce con più immagini) o l'aggregazione dei patch (nel caso di immagini mediche gigapixel).

In altre parole, il **MIVPG** è in grado di cogliere il contesto globale di ciò che sta elaborando, combinando le informazioni in modo più ricco rispetto al Q-former, che tratta ogni elemento visivo come separato.

**Esempi concreti**

Per comprendere meglio il valore di questo approccio, possiamo vedere alcuni esempi pratici:

1. **Scenario e-commerce**: Immagina di cercare un paio di scarpe su un sito di e-commerce. Tipicamente, un prodotto ha diverse immagini che mostrano il prodotto da angolazioni diverse: vista laterale, vista frontale, dettagli del materiale e così via. Un modello tradizionale con il Q-former tratterebbe ogni immagine come un’entità separata, senza collegarle tra loro. Il **MIVPG**, invece, è progettato per capire che tutte queste immagini si riferiscono allo stesso paio di scarpe, combinando le informazioni da ogni immagine per ottenere una rappresentazione più completa e coerente del prodotto.
2. **Immagini mediche (Whole Slide Images)**: In ambito medico, le immagini di una biopsia possono essere troppo grandi per essere elaborate direttamente, quindi vengono suddivise in tanti piccoli frammenti (patch). Il Q-former elaborerebbe ogni patch come se fosse un'immagine indipendente. Tuttavia, poiché tutte le patch appartengono alla stessa immagine (ad esempio, parti di un tessuto), trattarle separatamente porta a una perdita di informazioni contestuali. Il **MIVPG** è in grado di correlare i diversi patch, ricostruendo una visione globale del tessuto e migliorando la comprensione del campione medico.

**Esperimenti e risultati**

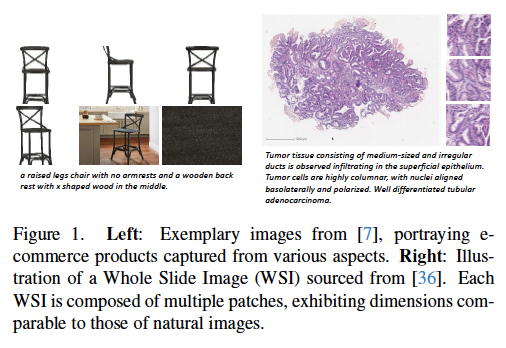
Per dimostrare l'efficacia del MIVPG, gli autori hanno condotto esperimenti su tre dataset pubblici che coprono scenari molto diversi:

* **Immagini naturali**, come quelle che troviamo in contesti comuni (ad esempio, scene di vita quotidiana).
* **Immagini gigapixel in ambito medico**, come le biopsie (Whole Slide Images).
* **Immagini di prodotti e-commerce**, dove ci sono più foto dello stesso prodotto da diverse angolazioni.

In tutti questi casi, il **MIVPG ha dimostrato di superare il Q-former**, mostrando che è capace di aggregare le informazioni visive da diverse fonti e migliorare la comprensione complessiva dell'immagine. In particolare, ha ottenuto migliori risultati nei compiti di **Image Captioning** (dove il modello genera descrizioni testuali per le immagini) e **Visual Question Answering** (dove il modello risponde a domande su ciò che appare nelle immagini).

**Conclusioni**

In conclusione, il contributo principale di questo lavoro è l'introduzione del **MIVPG**, un nuovo componente che migliora l’integrazione delle informazioni visive nei modelli di linguaggio multimodali. Il MIVPG non solo permette di gestire input visivi complessi come più immagini dello stesso oggetto o porzioni di immagini di grandi dimensioni, ma fa anche in modo che il modello possa cogliere le correlazioni tra queste immagini o porzioni. Questo approccio porta a un miglioramento significativo delle prestazioni rispetto al Q-former, dimostrando che è un metodo più robusto e generalizzabile per gestire compiti complessi che coinvolgono sia il testo che le immagini.



Nel secondo capitolo del paper, gli autori discutono il contesto del loro lavoro, descrivendo i contributi e le limitazioni degli approcci precedenti sia nel campo dell'apprendimento multimodale che nell'**Multiple Instance Learning (MIL)**.

**2.1 Apprendimento multimodale**

Gli autori iniziano spiegando l'evoluzione e lo stato attuale dei modelli di apprendimento multimodale (vision-language models, VLMs), che combinano immagini e testo per risolvere compiti complessi. Questi modelli mirano a migliorare la fusione delle informazioni visive e testuali. Alcuni esempi chiave includono:

* **TCL (Triplet Contrastive Learning)**: Un modello che utilizza l'apprendimento contrastivo triplet per imparare simultaneamente da immagini e testo, cercando di allineare meglio le rappresentazioni visive e linguistiche.
* **FROMAGe e LLaVA**: Questi modelli utilizzano un approccio più semplice, impiegando una proiezione lineare come Visual Prompt Generator (VPG) per convertire le immagini in rappresentazioni utilizzabili dai modelli di linguaggio.
* **Flamingo**: Introduce un approccio innovativo utilizzando il **Perceiver Resampler**, che incorpora meccanismi di attenzione incrociata e query learnable, migliorando la capacità del modello di gestire input complessi.
* **BLIP2 e MiniGPT-4**: Questi modelli utilizzano un componente chiamato **Q-former** per migliorare l'allineamento tra immagini e testo. BLIP2 è innovativo nell'uso del Q-former, mentre MiniGPT-4 usa una versione congelata del Q-former con strati addizionali apprendibili per migliorare le prestazioni.

Nonostante il successo di questi modelli, gli autori sottolineano una limitazione comune: la maggior parte di questi approcci presuppone una relazione **uno-a-uno** tra immagini e testo (cioè, ogni immagine ha un solo testo corrispondente). Tuttavia, nella realtà, queste relazioni possono essere più complesse, come nei casi di **uno-a-molti** (una descrizione per più immagini) o **molti-a-molti** (più descrizioni per più immagini). Applicare i modelli multimodali in questi scenari complessi è ancora una sfida aperta.

**2.2 Apprendimento Multi-Istanze (MIL)**

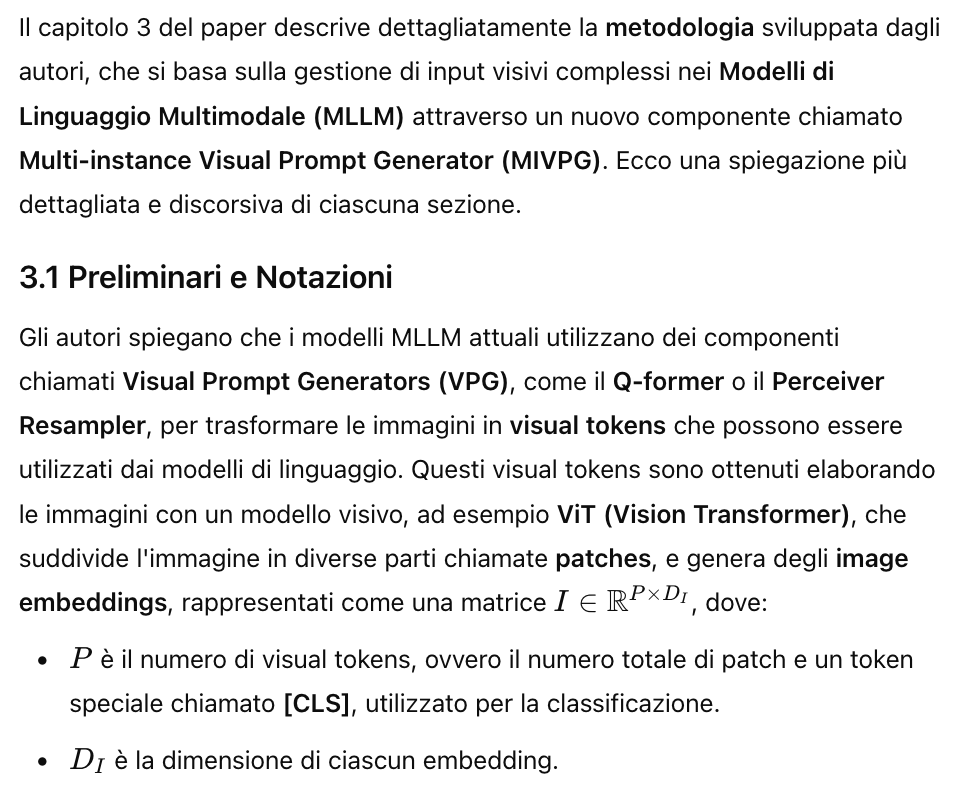
Gli autori passano poi a discutere l'**Apprendimento Multi-Istanze (Multiple Instance Learning, MIL)**, che è una tecnica in cui un insieme di istanze (per esempio, immagini o porzioni di immagini) viene trattato come un "bag" (un insieme), e il modello effettua previsioni basate sull'intero bag, piuttosto che sulle singole istanze. Esistono due principali approcci:

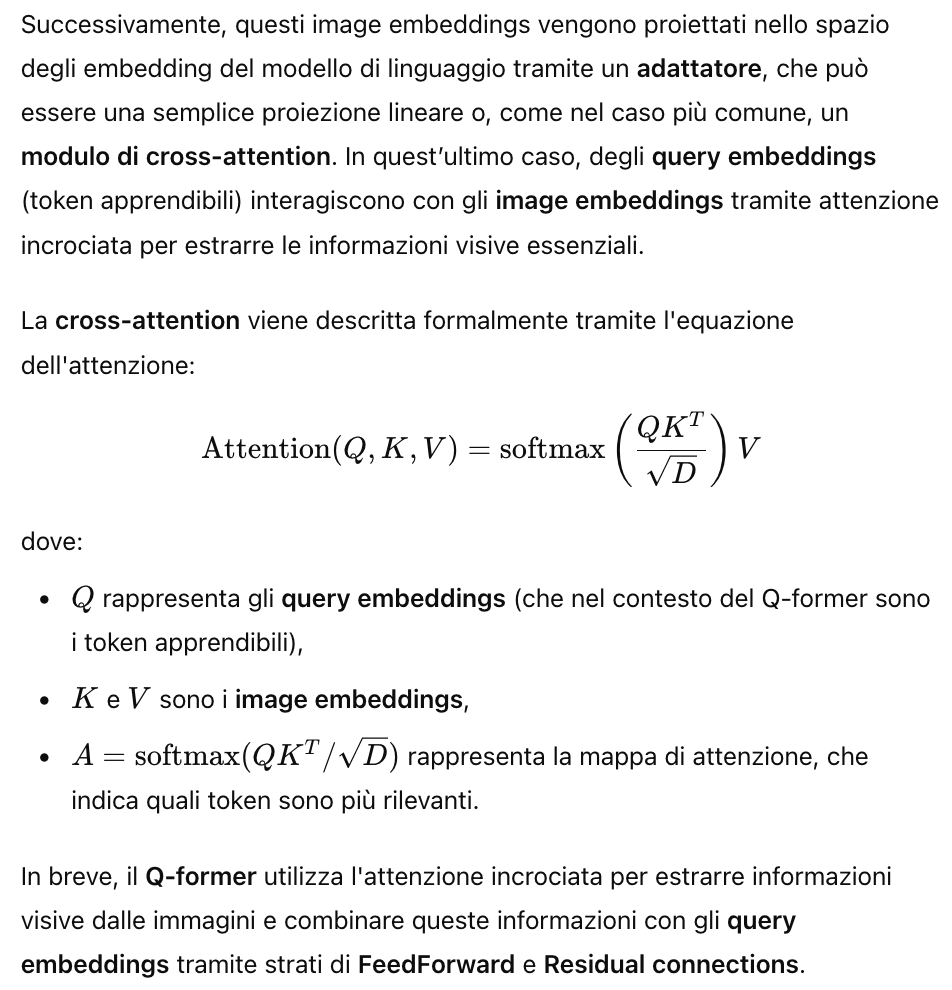
1. **Approccio a livello di istanza**: In questo caso, il modello fa previsioni per ciascuna istanza nel bag, e poi aggrega queste previsioni a livello di bag (per esempio, usando operatori come il max pooling o il mean pooling). Tuttavia, questi operatori di pooling predefiniti spesso non riescono a cogliere le informazioni più complesse.
2. **Approccio a livello di embedding**: Qui, il modello crea un embedding a livello di bag, che rappresenta tutte le istanze, e usa questo embedding per fare previsioni. Questo approccio, sebbene più complesso, di solito porta a risultati migliori poiché permette di aggregare le caratteristiche di tutte le istanze in modo più efficace. Per esempio:
   * **MI-Net** utilizza un livello completamente connesso (fully connected layer) per effettuare il pooling.
   * **AB-MIL** introduce un meccanismo di attenzione durante il pooling, che permette al modello di dare pesi diversi alle istanze in base alla loro importanza.
   * **DS-MIL** va oltre, applicando l'attenzione per considerare non solo le relazioni tra le singole istanze, ma anche le relazioni tra ciascuna istanza e l'intero bag.

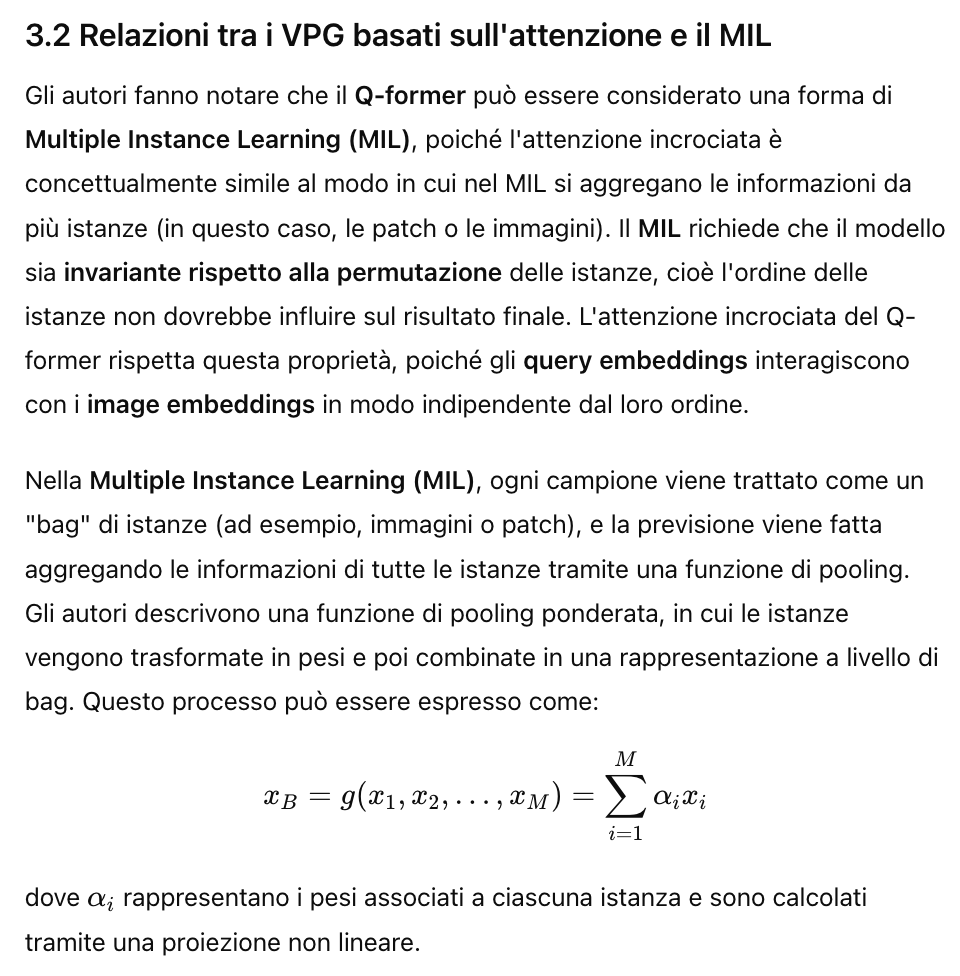
Gli autori segnalano che il campo del MIL ha visto molte innovazioni per quanto riguarda il pooling delle istanze e la gestione delle relazioni tra le stesse. Tuttavia, la maggior parte di queste ricerche si è concentrata su una singola modalità (di solito immagini), mentre l'estensione del MIL al **contesto multimodale** (dove si combinano immagini e testo) è ancora poco esplorata.

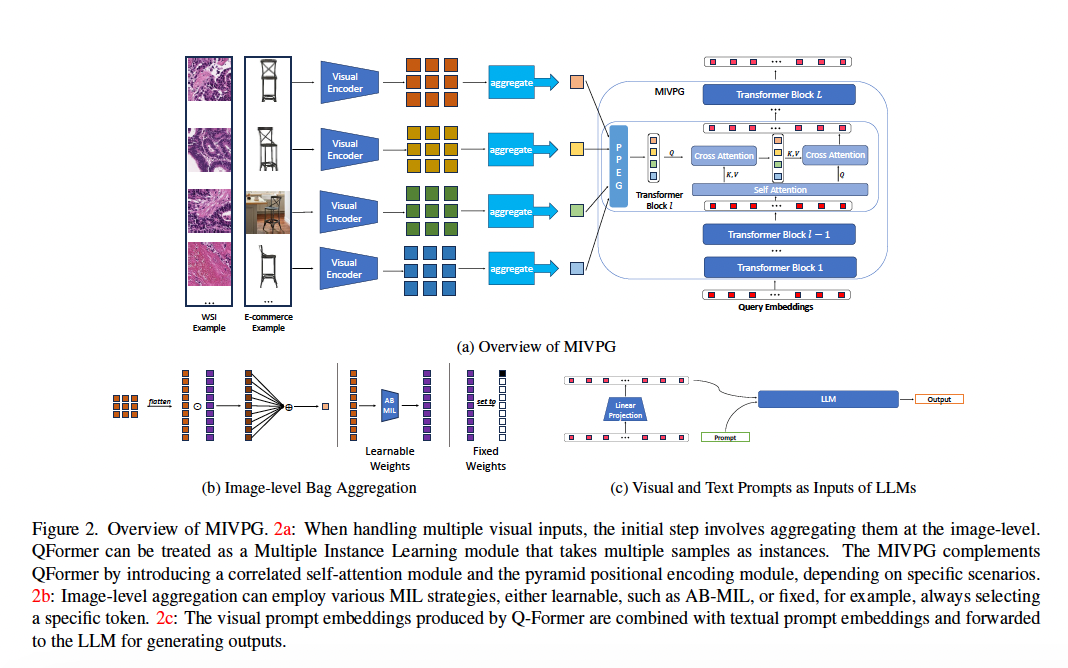
**Conclusione del "Related Work"**

La sezione si chiude evidenziando che, mentre ci sono stati progressi significativi sia nell'apprendimento multimodale che nel MIL, nessuno dei modelli attuali riesce veramente a gestire efficacemente le relazioni complesse tra testo e immagini in contesti reali, dove queste relazioni possono essere uno-a-molti o molti-a-molti. Questo pone le basi per il loro lavoro, che cerca di combinare i vantaggi del MIL con l'apprendimento multimodale per affrontare queste sfide.







****

**3.3 MIVPG per Input Visivi Multipli**

Gli autori propongono il **MIVPG** come estensione del Q-former per gestire casi in cui un singolo campione contiene **più immagini** o **più patch**. Mentre approcci precedenti, come il **Perceiver Resampler**, concatenano semplicemente le patch di più immagini in una singola sequenza, il **MIVPG** tratta ciascuna immagine come un bag distinto e ciascuna patch come un'istanza. Questo approccio più strutturato evita errori nella distribuzione dei pesi tra le istanze e migliora l'aggregazione delle informazioni.

Quando un campione contiene più immagini, il **MIVPG** deve aggregare le informazioni da due prospettive:

* **A livello di immagine**, dove ogni immagine è un "bag" e ogni patch all'interno dell'immagine è un'istanza.
* **A livello di campione**, dove l'intero campione è un "bag" e ogni immagine è un'istanza.

Questo approccio multilivello consente al **MIVPG** di gestire correttamente situazioni complesse in cui ci sono più immagini correlate all'interno di un singolo campione, come nel caso di analisi di immagini mediche o e-commerce.

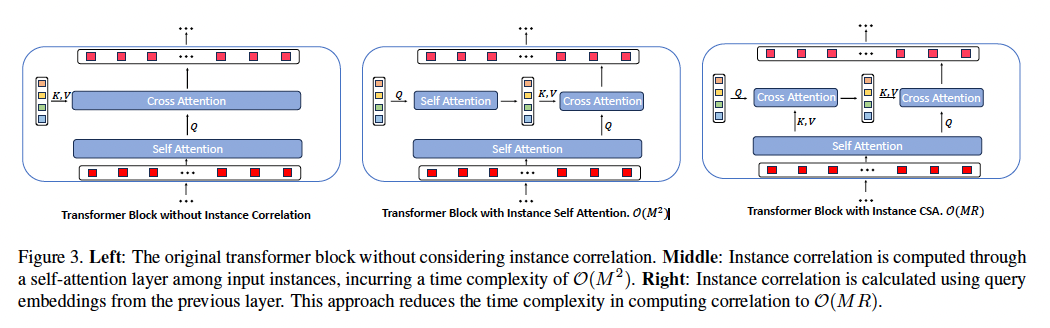
**3.4 Correlazione tra Istanze nel MIVPG**

Gli autori introducono un concetto avanzato chiamato **Correlated Self Attention (CSA)** per migliorare ulteriormente il MIVPG nei casi in cui ci sono correlazioni tra le immagini o le patch di un campione. Nei contesti pratici, come l'analisi di immagini mediche gigapixel, ci possono essere migliaia di istanze (patch). Calcolare direttamente l'auto-attenzione su tutte le istanze comporta una complessità computazionale elevata O(M2). Per ridurre questa complessità, gli autori propongono una **proiezione a basso rango** che riduce il costo computazionale utilizzando una matrice apprendibile di rango inferiore L. Questo consente di apprendere le correlazioni tra istanze in modo più efficiente.

Il **MIVPG** mantiene le proprietà del MIL anche quando viene dotato del modulo **CSA**, garantendo che il modello rispetti l'invarianza alla permutazione e che riesca ad aggregare in modo efficace le informazioni dalle diverse istanze correlate.

**Conclusione**

In sintesi, il **Q-former** può essere visto come una forma specializzata di **Multiple Instance Learning (MIL)**, ma con alcune limitazioni. Il **MIVPG** proposto dagli autori estende questa capacità, permettendo una gestione più sofisticata di input visivi multipli e correlati, come immagini e patch. Questo componente migliora l'integrazione delle informazioni visive nei modelli multimodali, garantendo un’aggregazione più efficiente e accurata delle informazioni visive provenienti da più fonti o immagini.



Il capitolo 4 del paper descrive gli **esperimenti** condotti per valutare l'efficacia del nuovo componente **MIVPG** in vari scenari di apprendimento multimodale. Gli autori esplorano tre scenari principali, che rappresentano diverse configurazioni di input visivi e modalità di elaborazione. Ecco una descrizione dettagliata e discorsiva di ciascun esperimento:

### 4.1 Configurazione Generale

Gli esperimenti sono stati condotti utilizzando il modello **BLIP2** come base, in combinazione con **FLAN-T5-XL** come modello di linguaggio. Il componente MIVPG è stato inizializzato con i pesi del **QFormer**. È importante notare che il **modello di linguaggio** e il **modello visivo** (basato su **ViT-G**) sono mantenuti fissi durante il training, e viene aggiornata solo la parte del MIVPG. Le immagini sono ridimensionate a 224x224 pixel, e le patch generate da queste immagini sono utilizzate come input visivo.

Gli autori sottolineano che, in dataset di piccole dimensioni, lasciare libero il modello visivo non porta miglioramenti significativi, quindi hanno deciso di non farlo nei loro esperimenti.

### 4.2 Scenario 1: Campioni con Singola Immagine

Nel primo scenario, gli autori si concentrano su campioni che contengono una **singola immagine**. In questo caso, le patch generate dall'immagine vengono trattate come **istanze**. Il dataset scelto per questo scenario è **MSCOCO**, un dataset comunemente utilizzato per valutare modelli di visione e linguaggio. Il training set originale contiene circa 560.000 coppie di immagini e testo.

L'obiettivo era dimostrare che il MIVPG, con l'aggiunta del **Pyramid Positional Encoding Generator (PPEG)**, è in grado di migliorare le prestazioni nei compiti di apprendimento con poche istanze. Come mostrato dai risultati, l'uso del PPEG diventa più efficace nei dataset di dimensioni ridotte, poiché moduli più sofisticati come MIVPG possono scoprire relazioni latenti che i modelli standard non riescono a rilevare. Tuttavia, con dataset di grandi dimensioni, la differenza tra il metodo proposto e i modelli tradizionali tende a diminuire, poiché la disponibilità di una grande quantità di dati aiuta a compensare l'assenza di moduli più complessi.

### 

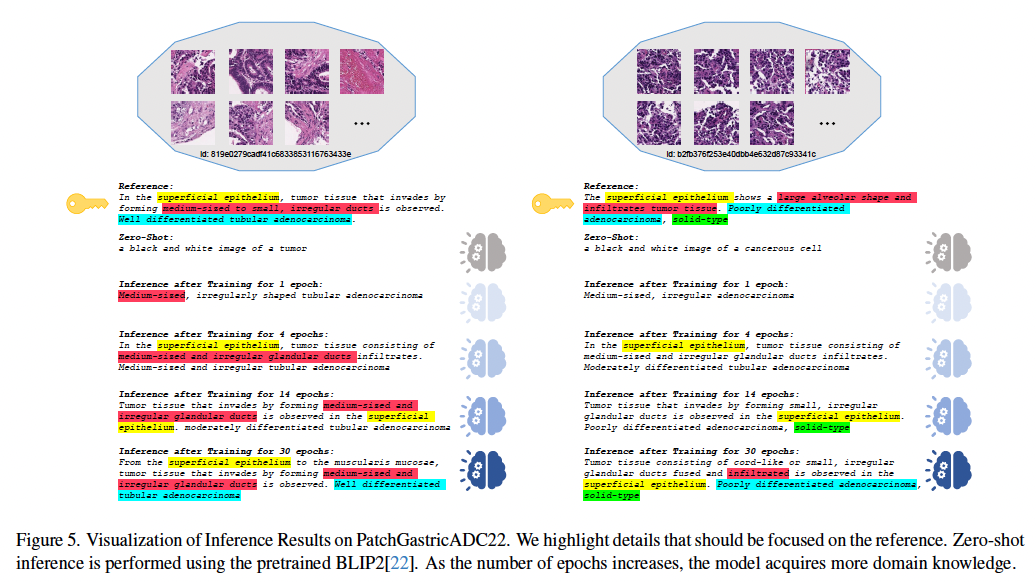
### 

### 4.3 Scenario 2: Campioni con Multiple Immagini, Ogni Immagine come Embedding Generale

Il secondo scenario riguarda campioni che contengono **multiple immagini**, dove ciascuna immagine è rappresentata da un singolo embedding generale. Questo approccio è testato utilizzando il dataset **PatchGastricADC22**, che contiene immagini **Whole Slide Image (WSI)** di campioni di adenocarcinoma gastrico, con circa 262.777 patch estratte da 991 WSIs.

In questo caso, ogni patch è rappresentata dal token [CLS] generato dal **ViT**. L'obiettivo è ridurre il carico computazionale, poiché i WSIs possono contenere fino a 1860 patch. I risultati mostrano che il metodo proposto supera significativamente i modelli basati su DenseNet121 o EfficientNetB3 combinati con LSTM, suggerendo che il MIVPG è particolarmente efficace nei compiti di classificazione medica.

Gli autori evidenziano anche che, man mano che il modello acquisisce più conoscenze specifiche del dominio durante il training, le descrizioni generate dalle immagini mediche diventano sempre più dettagliate. Questo conferma che il modello migliora nel cogliere informazioni rilevanti anche in domini che inizialmente non sono familiari.

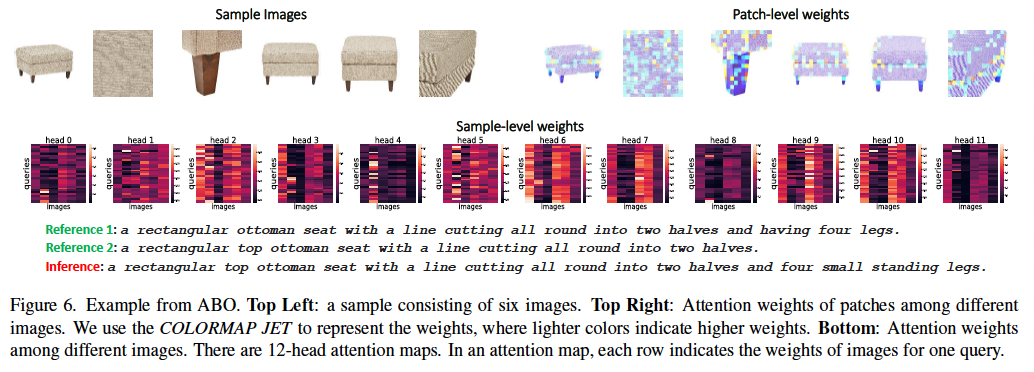


### 4.4 Scenario 3: Campioni con Multiple Immagini, Ogni Immagine con Multiple Patch

Nel terzo scenario, vengono valutati campioni che contengono **multiple immagini**, e ciascuna immagine è composta da **multiple patch**. Il dataset utilizzato è **Amazon Berkeley Objects (ABO)**, che contiene immagini di prodotti di e-commerce da diverse prospettive. In questo caso, è fondamentale considerare sia i dettagli delle singole immagini che le informazioni aggregate di tutte le immagini.

In questo scenario, gli autori utilizzano il **MIL** su due dimensioni: a livello di immagine (per generare gli embedding delle immagini) e a livello di patch (per analizzare i dettagli all'interno delle immagini). I risultati mostrano che l'approccio proposto supera il modello **BLIP2** in diversi set di configurazioni, dimostrando che considerare più immagini e patch migliora notevolmente la qualità delle descrizioni generate dal modello.

Un aspetto interessante che emerge da questo scenario è la visualizzazione delle **mappe di attenzione** del modello: il modello presta particolare attenzione a dettagli rilevanti, come i bordi o le gambe di un oggetto (nell'esempio dell'ottomano), riuscendo così a generare descrizioni più accurate.

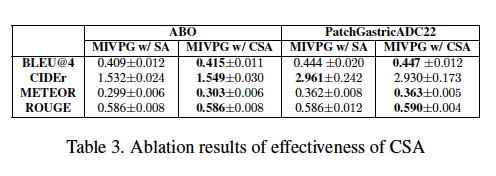


### 4.5 Studio di Casi

Gli autori hanno condotto ulteriori esperimenti per esplorare l'impatto della correlazione tra istanze, confrontando l'uso della **self-attention (SA)** e della **correlated self-attention (CSA)**. I risultati mostrano che nel dataset PatchGastricADC22, SA e CSA danno risultati simili, mentre nel caso del dataset ABO, la CSA supera la SA, suggerendo che la correlazione tra immagini non è sempre esplicita in contesti di e-commerce, ma può essere sfruttata in modo efficace attraverso la CSA.

### Conclusione

Gli esperimenti dimostrano che il componente **MIVPG** proposto dagli autori offre un significativo miglioramento rispetto ai metodi esistenti nei compiti multimodali, soprattutto nei casi in cui ci sono input visivi complessi, come più immagini o immagini con molte patch. Il modello riesce a gestire meglio la correlazione tra diverse istanze visive, migliorando così le prestazioni nei task downstream, in particolare quando i dati disponibili sono limitati o complessi.



Nella **conclusione** del paper, gli autori riassumono il contributo principale del loro lavoro, che consiste nell'introduzione del **Multi-instance Visual Prompt Generator (MIVPG)**. Questo componente è descritto come **flessibile, generale e potente**, e serve a migliorare la fusione delle rappresentazioni visive nei **Modelli di Linguaggio Multimodali (MLLMs)**. Attraverso gli esperimenti, gli autori dimostrano che il MIVPG offre prestazioni superiori rispetto ai metodi esistenti in diversi scenari, soprattutto quando si tratta di gestire input visivi complessi, come immagini multiple o immagini composte da molte patch.

Gli autori evidenziano che il **QFormer**, un componente usato comunemente in molti MLLM all'avanguardia, è in realtà una versione limitata del MIVPG. Hanno fornito supporto teorico per dimostrare l'efficacia di quest'ultimo, evidenziando come le tecniche avanzate basate sul **Multiple Instance Learning (MIL)** e la capacità del MIVPG di incorporare segnali visivi arricchiti possano contribuire allo sviluppo futuro dei MLLM.

In sintesi, il paper presenta il MIVPG come un passo in avanti nella fusione di informazioni visive e linguistiche, migliorando la capacità dei modelli di affrontare compiti complessi che coinvolgono dati visivi.