

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO CUỐI KỲ HỌC SÂU

MÔ TẢ MÓN ĂN TỪ HÌNH ẢNH SỬ DỤNG MÔ HÌNH MANG KIẾN TRÚC TRANSFORMER

Người hướng dẫn: **PGS.TS. LÊ ANH CƯỜNG**

Người thực hiện: **TRẦN MINH TRÍ – 52000815**

TRẦN QUỐC VINH – 52000823

LÊ VĂN VIỆT - 52000822

Lớp : 200503401

Khóa : 24

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO CUỐI KỲ HỌC SÂU

MÔ TẢ MÓN ĂN TỪ HÌNH ẢNH SỬ DỤNG MÔ HÌNH MANG KIẾN TRÚC TRANSFORMER

Người hướng dẫn: **PGS.TS. LÊ ANH CƯỜNG**

Người thực hiện: **TRẦN MINH TRÍ – 52000815**

TRẦN QUỐC VINH – 52000823

LÊ VĂN VIỆT - 52000822

Lớp : 200503401

Khóa : 24

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024

LỜI CẢM ƠN

Nhóm xin cảm ơn thầy Lê Anh Cường đã giảng dạy một cách dễ hiểu về môn học sâu. Thông qua những bài tập trên lớp, giúp chúng em hiểu sâu hơn về cách hoạt động đằng sau những mô hình kinh điển và hiện đại trong lĩnh vực này. Trong quá trình làm bài báo cáo cuối kỳ này sẽ không tránh khỏi những sai sót, mong thầy thông cảm bỏ qua.

BÁO CÁO ĐƯỢC HOÀN THÀNH TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong Khóa luận/Đồ án tốt nghiệp còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung Báo cáo cuối kỳ của mình. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày 18 tháng 5 năm 2024

Tác giả

Trần Minh Trí

Trần Quốc Vinh

Lê Văn Việt

TÓM TẮT

Với sự phát triển của AI (Artificial Intelligence) hiện nay, các mô hình Computer Vision (CV) hay Natural Language Processing (NLP) đang ngày càng dễ học, dễ làm và dễ ứng dụng hơn bao giờ hết. Không dừng ở đó, ta còn có thể kết hợp những mô hình kể trên để giải quyết bài toán multimodal như Vision-Language. Tuy nhiên, việc huấn luyện end-to-end sẽ vô cùng tốn kém do phải huấn luyện cùng lúc nhiều loại mô hình khác nhau, hệ quả là hiệu suất ứng dụng sẽ không đạt kỳ vọng. Liệu có cách nào để chúng ta có thể tận dụng những mô hình có sẵn mà không cần huấn luyện lại hay không? Làm sao để tận dụng sức mạnh xử lý hình ảnh của Vision Transformer hay khả năng tạo sinh văn bản của Large Language Models mà không cần “học” lại? Để trả lời cho câu hỏi đó, ta sẽ cùng tìm hiểu về BLIP-2 và cách mô hình này đã dùng ViT và LLM để giải quyết bài toán Vision-Language.

MỤC LỤC

CHƯƠNG 1 – TỔNG QUAN VỀ ĐỀ TÀI.....	1
1.1 Mục đích chọn đề tài	1
1.2 Mô hình ngôn ngữ thị giác	1
CHƯƠNG 2 – MÔ HÌNH BLIP-2	3
2.1 Sơ lược về mô hình họ BLIP	3
2.2 Mô hình BLIP-2	4
2.3 Cấu trúc của Query Transformer	5
2.3.1 Giai đoạn 1: Representation learning.....	5
2.3.2 Giai đoạn 2: Generative learning	7
2.4 Kết quả thử nghiệm của mô hình.....	9
CHƯƠNG 3 – CẢI TIẾN BLIP-2 BẰNG PHƯƠNG PHÁP INSTRUCTION TUNING.....	10
3.1 Sơ lược về InstuctBLIP	10
3.2 Kiến trúc InstructBLIP	10
3.3 Kết quả đánh giá	10
3.4 Kết luận	11
CHƯƠNG 4 – THỰC NGHIỆM.....	12
4.1 Xây dựng dữ liệu.....	12
4.2 Huấn luyện mô hình.....	12
4.3 Đánh giá.....	14
TÀI LIỆU THAM KHẢO	2

DANH MỤC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

Hình 2.1 Kiến trúc của BLIP	3
Hình 2.2 BLIP sử dụng Captioner để tạo caption cho ảnh và dung Filter để lọc nhiễu. 4	
Hình 2.3 Kiến trúc tổng quát của BLIP-2.....	5
Hình 2.4 Kiến trúc Q-Former trong giai đoạn đầu của pre-training.....	6
Hình 2.5 Attention mask cho từng mục tiêu huấn luyện.....	7
Hình 2.6 Kiến trúc cho giai đoạn 2 của pre-training, sử dụng LLM decoder (trên) hoặc LLM encoder-decoder (dưới)	8
Hình 2.7 Một số ví dụ về khả năng xử lý hình ảnh và ngôn ngữ của BLIP-2.....	Error!
Bookmark not defined.	
 Hình 3.1 Kiến trúc tổng quát của InstructedBLIP	 10
Hình 3.2 Kết quả zero-shot evaluation	11
 Hình 4.1: Bộ dữ liệu món ăn	 12
Hình 4.2: Sử dụng mô hình đã được pretrained	12
Hình 4.3: Cấu hình LoRA cho mô hình	13
Hình 4.4: Huấn luyện mô hình	13
Hình 4.5: Kết quả thử nghiệm với ảnh món bún bò Huế	14
Hình 4.6: Kết quả thử nghiệm với ảnh món bún đậu mắm tôm	14
Hình 4.7: Kết quả thử nghiệm với ảnh món bánh bèo	15

CHƯƠNG 1 – TỔNG QUAN VỀ ĐỀ TÀI

1.1 Mục đích chọn đề tài

Trong lĩnh vực thị giác máy tính, ta xoay quanh các bài toán như phát hiện đối tượng (*object detection*), phân loại hình ảnh (*image classification*), phân đoạn hình ảnh (*image segmentation*),... Cùng với sự phát triển vượt bậc của nhân loại và công nghệ thì các mô hình ngày càng đạt độ hiệu quả chính xác hơn (song vẫn chưa đủ tốt để so sánh với mắt người) và đem lại nhiều lợi ích trong việc phục vụ đời sống, xã hội loài người. Trong số đó, không thể không nhắc tới các mô hình thuộc nhóm *vision-language*, một số bài toán điển hình của các mô hình này có thể kể tới như: chú thích hình ảnh (*image captioning*), trả lời câu hỏi trực quan (*visual question answering*), truy xuất văn bản-hình ảnh (*image-text retrieval*),...

Đối với mỗi bài toán, ta sẽ có những lợi ích riêng ứng dụng vào mọi khía cạnh của cuộc sống. Ví dụ, đối với *image captioning* thì mô hình sẽ có khả năng sinh các chú thích mô tả về bức ảnh nhằm hỗ trợ những người khiếm khuyết có khả năng hiểu được nội dung của bức tranh, hình ảnh,... ngoài ra, mô hình sẽ giúp cải thiện việc lập chỉ mục và truy xuất nội dung trong cơ sở dữ liệu multi-media. Con người sẽ có khả năng tương tác với máy bằng cách hỏi chúng những câu liên quan tới ảnh đầu vào, và mô hình sẽ cung cấp câu trả lời liên quan đến ngữ cảnh của câu hỏi, đây là một bước tiến trong việc xây dựng trợ lý ảo cho nhiều lĩnh vực (ví dụ: giáo dục), cải thiện *interactive search engines*.

Những ưu điểm thực tế mà các mô hình *vision language* thể hiện tính linh hoạt, hữu ích trên nhiều lĩnh vực khác nhau, góp phần đưa ra giải pháp cho nhiều thách thức trong việc hiểu, tương tác, và tận dụng thông tin từ hình ảnh một cách hiệu quả.

1.2 Mô hình ngôn ngữ thị giác

Mô hình ngôn ngữ thị giác (Vision Language Models: VLM), là mô hình trí tuệ nhân tạo cho phép máy tính có khả năng hiểu cả thông tin hình ảnh và ngữ cảnh văn bản. Theo truyền thống, các mô hình thị giác máy tính và xử lý ngôn ngữ hoàn toàn độc lập với nhau, một bên chỉ quan tâm đến giải hình ảnh, một bên chỉ chú trọng phân tích ngôn ngữ. Song, mô hình ngôn ngữ thị giác đã phá vỡ rào cản này, chúng có thể vừa

phân tích hình ảnh vừa có thể hiểu được văn bản liên quan đến hình ảnh, cho phép nắm rõ toàn diện hơn về nội dung trực quan của hình ảnh.

Trọng tâm cốt lõi của mô hình ngôn ngữ thị giác chính là học tập đa mô hình (multimodal learning), đây là một phương pháp phức tạp vì đòi hỏi mô hình phải xử lý và hiểu thông tin từ nhiều phương diện khác nhau từ độ chi tiết của hình ảnh tới mức ngữ nghĩa của văn bản. Bằng cách tiếp cận này, các mô hình VLM có thể nắm bắt được các mối quan hệ sắc thái giữa các yếu tố hình ảnh và ngôn ngữ mô tả tương ứng về chúng.

Một vài lĩnh vực mà mô hình ngôn ngữ thị giác được ứng dụng như:

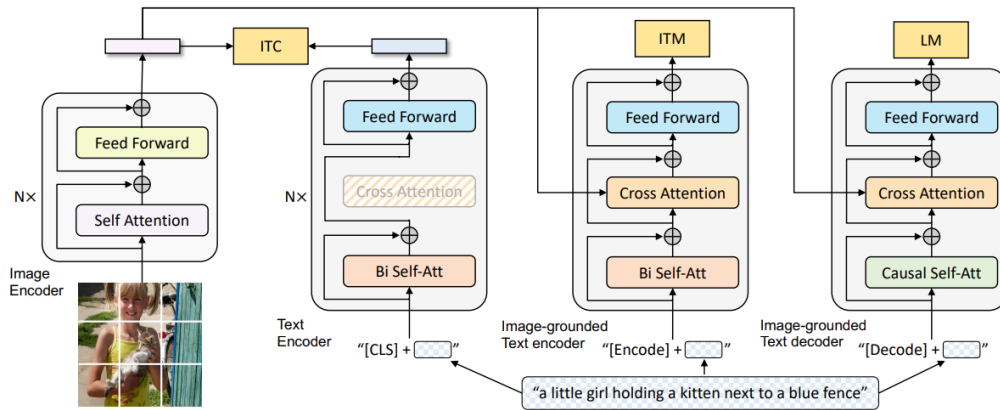
- Y tế: Trong lĩnh vực y tế, VLM có thể hỗ trợ bác sĩ giải thích các hình ảnh y tế và hiểu các báo cáo phức tạp. Bằng cách phân tích cả dữ liệu trực quan như tia X và các văn bản y tế liên quan, VLM có thể hỗ trợ chẩn đoán và lập kế hoạch điều trị chính xác.
- Giải trí: VLM đang chuyển đổi ngành công nghiệp giải trí, cho phép người tạo nội dung tạo ra trải nghiệm đa phương tiện, phong phú về mặt tương tác. Từ trò chơi điện tử có các đoạn hội thoại đến môi trường thực tế ảo sống động, VLM nâng cao mức độ tương tác và trải nghiệm của người dùng.
- Giáo dục: Trong giáo dục, VLM có thể tạo ra trải nghiệm học tập toàn diện. Ví dụ, chúng có thể giúp học sinh khiếm thị hiểu nội dung trực quan trong sách giáo khoa bằng cách cung cấp các mô tả chi tiết về hình ảnh bằng lời nói.
- Thương mại điện tử: VLM nâng cao, cải tiến hệ thống đề xuất sản phẩm bằng cách phân tích cả hình ảnh sản phẩm và đánh giá của khách hàng. Điều này cho phép đưa ra các đề xuất chính xác và được cá nhân hóa hơn dựa trên sở thích của người dùng được trích xuất từ hình ảnh và văn bản.

CHƯƠNG 2 – MÔ HÌNH BLIP-2

2.1 Sơ lược về mô hình họ BLIP

Thông thường, các phương pháp Vision-language Pre-training (VLP) sẽ tận dụng bộ dữ liệu cào được trên mạng theo dạng ảnh kèm alt-text làm cặp dữ liệu hình ảnh – văn bản để huấn luyện mô hình. Tuy nhiên, dữ liệu như vậy sẽ chứa rất nhiều nhiễu, khiến cho mô hình đầu ra không đạt hiệu quả mong muốn. Để giải quyết điều này, nhóm nghiên cứu đã đề xuất một framework VLP mới mang tên Bootstrapping Language-Image Pre-training (BLIP).

Về mặt kiến trúc, mô hình BLIP kết hợp 3 khối Image Encoder, Text Encoder và Text Decoder để thực hiện những tác vụ xử lý đồng thời hình ảnh và ngôn ngữ (Hình 2.1). Hiểu đơn giản thì các khối sẽ chia sẻ “kiến trúc” với nhau, cho phép mô hình có thể hiểu được dữ liệu (do Image Encoder và Text Encoder đảm nhiệm) và trả lại đầu ra tương ứng (do Text Decoder đảm nhiệm).



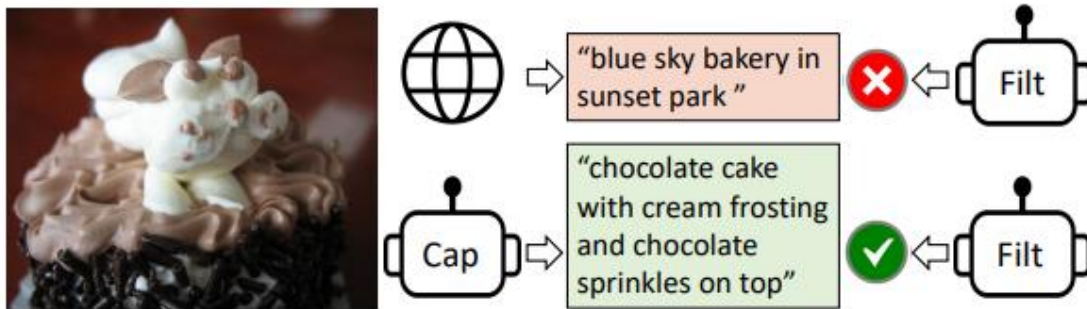
Hình 2.1 Kiến trúc của BLIP

Về mặt dữ liệu, BLIP sẽ sử dụng chính các khối kể trên để làm 2 mô hình Captioner và Filter nhằm xử lý nhiễu trong bộ dữ liệu cào được trên mạng.

- Captioner: mô hình miêu tả trong ảnh có gì
- Filter: mô hình kiểm tra xem ảnh và caption có liên quan không.

Lấy ví dụ như trong hình 2.2, Alt-Text (Alt Text là văn bản thay thế sẽ hiển thị trong trường hợp hình ảnh không thể hiển thị trên trang web) thu thập được từ web không hề

liên quan đến ảnh (bầu trời với bánh kem). Lúc này Filter sẽ từ chối alt-text trên và nhiệm vụ của Captioner sẽ tạo ra một alt-text khác phù hợp hơn.



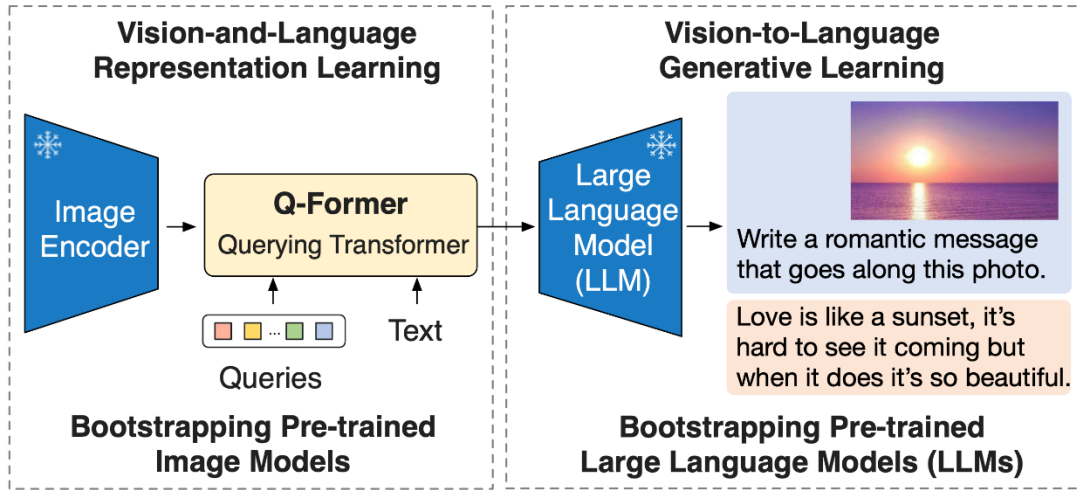
Hình 2.2 BLIP sử dụng Captioner để tạo caption cho ảnh và dung Filter để lọc nhiễu

Tóm lại, BLIP đặc biệt ở chỗ có khả năng tận dụng bộ dữ liệu khổng lồ và kết hợp các kiến trúc chuyên môn trong việc “hiểu” hình ảnh-văn bản và tạo sinh. Kết quả là mô hình đầu ra sẽ có khả năng xử lý những tác vụ Vision-Language.

2.2 Mô hình BLIP-2

Đột phá là vậy nhưng để đem lại hiệu suất mong muốn, mô hình huấn luyện theo phương pháp BLIP sẽ cần học end-to-end, tức mọi khối đều sẽ trải qua quá trình huấn luyện. Việc huấn luyện end-to-end như vậy thực sự rất “đau ví”, đặc biệt khi ta muốn scale mô hình lớn hơn. Ở thời điểm bấy giờ, những mô hình chuyên xử lý hình ảnh (các mô hình ViT) và chuyên xử lý văn bản (các mô hình LLM) đã trở nên vô cùng mạnh mẽ. Lúc này, nhóm nghiên cứu tại Salesforce nảy ra ý tưởng dùng luôn 2 ông kể trên, còn mục tiêu của mình chỉ là làm sao để chúng hiểu nhau thôi. Và thế là BLIP-2 ra đời. Khác với BLIP, BLIP-2 sử dụng Frozen Image Encoders và Frozen LLMs, tức là 2 ông này sẽ “nằm im”, không “học” gì cả trong toàn bộ quá trình huấn luyện. Thay vào đó, nhóm nghiên cứu sẽ sử dụng kiến trúc mới “Query Transformer” như là cầu nối giữa Image Encoder và LLM.

Kiến trúc tổng quát của BLIP-2 như hình 2.3, với việc huấn luyện diễn ra trong 2 giai đoạn. Giai đoạn 1, mô hình sẽ học các diễn giải đặc trưng ảnh thành đặc trưng văn bản. Trong giai đoạn 2, mô hình sẽ chú trọng vào học cách tạo câu trả lời từ đặc trưng trích xuất được.



Hình 2.3 Kiến trúc tổng quát của BLIP-2

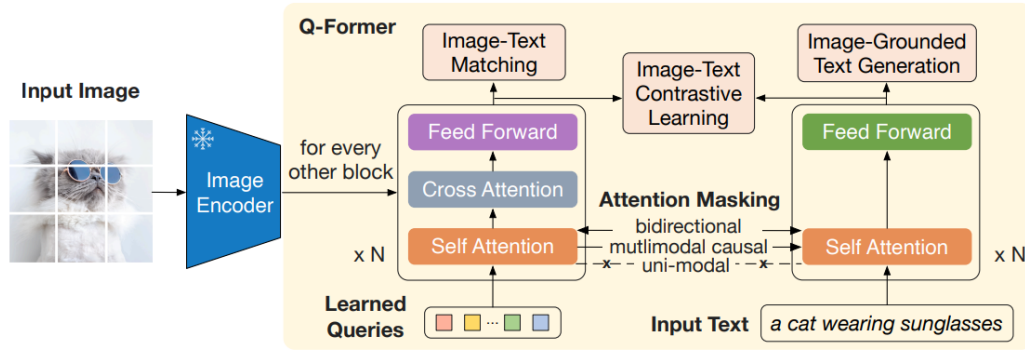
2.3 Cấu trúc của Query Transformer

Query Transformer (gọi tắt là Q-Former) có cấu tạo bao gồm 2 khối Image Transformer (đề tương tác với Frozen Image Encoder) và Text Transformer (tương đương với 2 khối Text Encoder và Text Decoder). Text Transformer sẽ tiếp nhận văn bản làm đầu vào còn Image Transformer sẽ tiếp nhận bộ queries. Hiểu nôm na thì bộ queries này sẽ tương tự như loạt câu hỏi “Ai? Cái gì? Chuyện gì?” vậy. Qua quá trình huấn luyện thì dần dần bộ queries sẽ “biết” nhất những đặc trưng cần thiết và liên quan đến văn bản. Một điểm đặc biệt khác là Image Transformer và Text Transformer sẽ dùng chung lớp self-attention, và nhóm nghiên cứu sẽ sử dụng attention masks để kiểm soát tương tác giữa queries và văn bản thông qua lớp self-attention kể trên. Q-Former được huấn luyện trong 2 giai đoạn: Representation Learning và Generative Learning.

2.3.1 Giai đoạn 1: Representation learning

Ở giai đoạn 1, Q-Former sẽ học liên kết đặc trưng giữa hình ảnh và ngôn ngữ, sao cho phần “Query” hiểu được đặc trưng nào của ảnh sẽ tương đương với đặc trưng của văn bản.

Lấy ví dụ như thế này: cho một cặp ảnh chú mèo và văn bản “mèo đang đeo kính râm” (Hình 2.4), Image Encoder sẽ trả về một lượng đặc trưng đa dạng và có kích thước lớn. Nhiệm vụ của Queries lúc này là chỉ nhặt ra đặc trưng tương ứng với đặc trưng của văn bản (đặc trưng nào tương ứng “mèo” và đặc trưng nào tương ứng với “kính râm”).



Hình 2. 4 Kiến trúc Q-Former trong giai đoạn đầu của pre-training

Trong giai đoạn 1, BLIP-2 có 3 mục tiêu huấn luyện, mỗi mục tiêu sẽ áp dụng một attention masking khác nhau cho lớp self-attention. Cụ thể ba mục tiêu này gồm:

1. Image-Text Contrastive Learning (ITC)
2. Image-Grounded Text Generation (ITG)
3. Image-Text Matching (ITM)

2.3.1.1 Image-Text Contrastive Learning (ITC)

Mục tiêu của ITC là để mô hình học được sự liên kết giữa ảnh và văn bản, sao cho với cặp ảnh và văn bản tương đồng thì phải cho ra Embedding tương đồng, và cặp khác nhau thì phải cho ra Embedding khác nhau. Trong trường hợp này, ta sẽ tính chỉ số tương đồng dựa trên đầu ra của khối Image Transformer (query embedding) và của khối Text Transformer (text embedding). Cặp nào càng giống thì chỉ số càng cao và ngược lại.

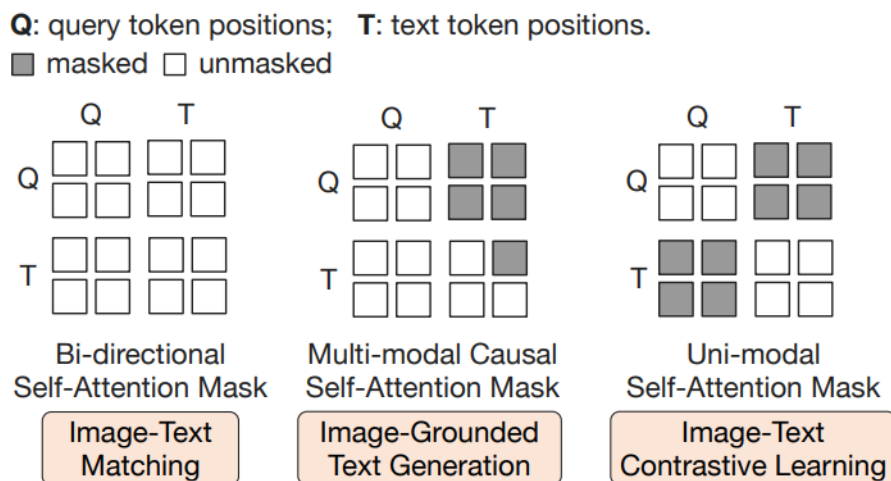
Ví dụ: ảnh một chú mèo đeo kính với văn bản “mèo đeo kính râm” sẽ phải cho ra 2 embedding có chỉ số tương đồng cao. Mặt khác, với văn bản “bãi biển mùa hè” thì chỉ số tương đồng giữa 2 embedding bắt buộc phải thấp. Để tránh rò rỉ thông tin, nhóm nghiên cứu sử dụng uni-modal self-attention mask, tức không cho phép tương tác giữa queries và văn bản.

2.3.1.2 Image-Grounded Text Generation (ITG)

Mục tiêu của ITG là tạo văn bản (caption) dựa trên ảnh cho trước. Để làm được vậy thì Text Transformer sẽ cần thông tin từ bức ảnh, nhưng Q-Former lại không cho phép khối này dùng trực tiếp thông tin của ảnh mà phải thông qua queries. Hệ quả là queries sẽ bắt buộc chỉ nhận thông tin hữu ích nhất cho Text Transformer mà thôi. Ở đây nhóm nghiên cứu áp dụng multimodal causal self-attention mask, tức chỉ cho phép text được tương tác một chiều với queries (để lấy thông tin).

2.3.1.3 Image-Grounded Text Generation (ITG)

Image-Text Matching (ITM): Mục tiêu của ITM là làm sao để mô hình học được sự liên kết giữa ảnh và caption tương ứng (tương tự với ITC nhưng học chi tiết hơn). Nói cách khác, đây là bài toán phân loại nhị phân mà mô hình sẽ phải đoán xem ảnh và văn bản có liên quan đến nhau không. Để làm được điều đó thì khối Image Encoder sẽ cần thông tin từ cả bức ảnh và văn bản. Ở đây, nhóm nghiên cứu sử dụng Bi-directional self-attention mask, tức cho phép queries và text tương tác với nhau hoàn toàn. Hệ quả là đầu ra Query Embedding sẽ chứa thông tin của cả bức ảnh và văn bản, từ đó sẽ chạy qua một lớp phân loại nhị phân để phân loại.



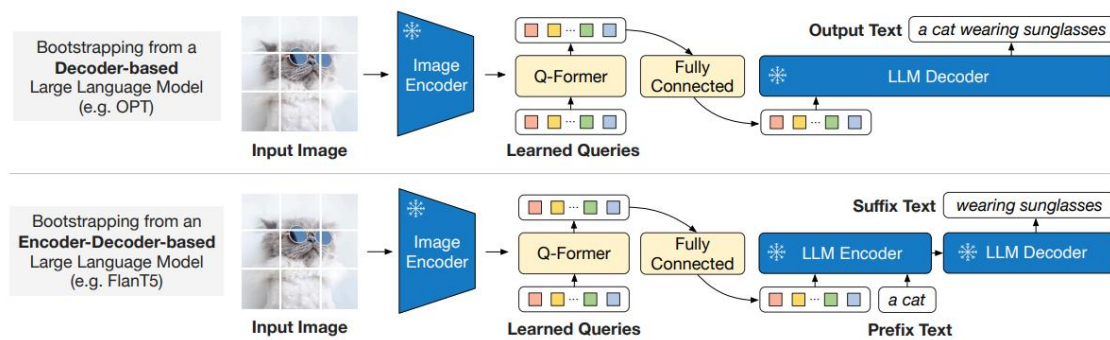
Hình 2.5 Attention mask cho từng mục tiêu huấn luyện

2.3.2 Giai đoạn 2: Generative learning

Sau giai đoạn 1, ta sẽ thu được một Q-Former có khả năng “mô tả bức ảnh”. Hiểu nôm na là bây giờ ông Q-Former đã sẵn sàng “hiểu và dịch” cho ông LLM những gì ông Image Encoder “nói”. Tất nhiên là đầu ra của Q-Former không dùng trực tiếp cho

LLM được (vì khác biệt trong chiều của data). Thay vào đó, query embedding sẽ phải chạy qua một khối Fully Connected (FC) để biến đổi tuyến tính, sao cho phù hợp làm đầu vào của LLM. Nói cách khác thì query embedding sau khi chạy qua lớp FC sẽ hoạt động tương tự như mô tả của bức ảnh (soft visual prompt) để đưa vào LLM.

Ở giai đoạn 2 này, việc “học” sẽ diễn ra chủ yếu với khối FC, mục tiêu là để LLM tạo ra câu trả lời mong muốn dựa trên query embeddings thu được từ Q-Former (vì LLM đã frozen nên chỉ có thể huấn luyện lớp FC). Ở trong hình 2.6, nhóm tác giả có thử nghiệm với 2 loại LLM: decoder-based và encoder-decoder-based. Đối với decoder-based, đầu ra của khối FC sẽ làm đầu vào của LLM. Đối với encoder-decoder-based, nhóm nghiên cứu “cắt” phần đầu của output mong muốn (prefix) và kết hợp với query embeddings để làm đầu vào của LLM (đầu ra tất nhiên là phần còn lại của output).















Hình 2.6 Kiến trúc cho giai đoạn 2 của pre-training, sử dụng LLM decoder (trên) hoặc LLM encoder-decoder (dưới)

Tóm lại, việc pre-train của Q-Former diễn ra thành 2 giai đoạn và có thể hiểu như sau: Trong giai đoạn 1, Q-Former sẽ học cách trích xuất từ bộ đặc trưng của ảnh thành queries embedding sao cho thông tin thu thập được liên quan đến văn bản nhất. Sau đó, trong giai đoạn 2, queries embedding sẽ được “dịch” thành prompt (tương tự như việc mô tả bức ảnh). Trọng tâm của giai đoạn 2 là để khối FC học cách “dịch” queries embedding sao cho LLM tạo được đầu ra mong muốn. Chi tiết hơn về cách hoạt động, xin mời đọc thêm về bài báo này tại mục tham khảo.

2.4 Kết quả thử nghiệm của mô hình

Khác với các phương pháp khác, BLIP-2 tập trung huấn luyện một mô hình trung gian thay vì huấn luyện toàn bộ. Bằng cách này, mô hình BLIP-2 vừa đạt được hiệu suất State-of-the-art (SOTA) mà lại tiết kiệm, giúp cắt giảm phần lớn chi phí ứng dụng.

Sau đây là một số ví dụ về instructed zero-shot image-to-text generation sử dụng mô hình BLIP-2 kết hợp ViT-g và FlanT5XXL, trong đó mô hình cho thấy nó có thể xử lý nhiều tác vụ bao gồm trao đổi, lý luận kiến thức bằng hình ảnh, kể chuyện,...

 <p>Explain the advantages of this product.</p> <p>The audi e-tron quattro concept is a plug-in hybrid electric sports car that has a range of 310 miles and can accelerate from 0 to 60 mph in just 4 seconds.</p>	 <p>Tell me something about the history of this place.</p> <p>The great wall of china was built in 221 bc by emperor qin shi Huang to protect the emperor's capital from invaders from the north.</p>	 <p>Write down the facts that you know about this flower.</p> <p>It is a genus of flowering plants in the orchid family, orchidaceae. It is native to eastern asia and south east asia. It is a perennial flowering plant that grows up to a height of 1 m.</p>
 <p>Is this photo unusual?</p> <p>Yes, it's a house that looks like it's upside down.</p> <p>How could someone get out of the house?</p> <p>It has a slide on the side of the house.</p>	 <p>What are shown in the photo?</p> <p>A man and a chicken.</p> <p>What does the man feel and why?</p> <p>He is scared of the chicken because it is flying at him.</p>	 <p>What are the ingredients I need to make this?</p> <p>Pepperoni, mozzarella cheese, pizza sauce, olive oil, salt, pepper, basil.</p> <p>What is the first step?</p> <p>Place the pizza dough on a baking sheet, brush with olive oil, sprinkle with salt, pepper, and basil.</p>
 <p>Which city is this?</p> <p>Singapore.</p> <p>Why do you think so?</p> <p>The city has a statue of a merlion.</p>	 <p>What happened at the end of this movie?</p> <p>The titanic sank.</p> <p>Did Leonardo Dicaprio's character survive?</p> <p>No, he drowned.</p>	 <p>What is in the photo?</p> <p>A pizza that looks like a cat.</p> <p>What is the nose made of?</p> <p>A slice of pepperoni.</p>
 <p>Write a romantic message that goes along this photo.</p> <p>Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.</p>	 <p>Tell us about the photo you took for Darren and Jade.</p> <p>Darren and Jade are a couple who have a passion for horses, so we decided to take a photo of them with a horse in the desert on their wedding day.</p>	 <p>Write a conversation between the two animals.</p> <p>cat: hey dog, can i ride on your back? dog: sure, why not? cat: i'm tired of walking in the snow.</p>

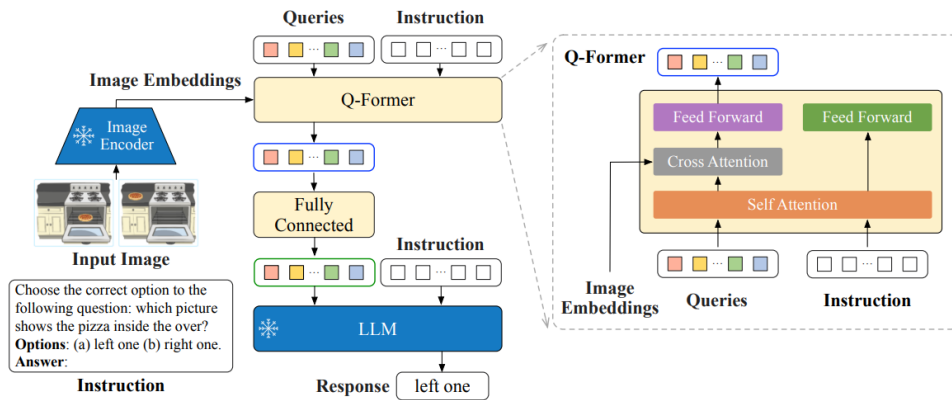
CHƯƠNG 3 – CẢI TIẾN BLIP-2 BẰNG PHƯƠNG PHÁP INSTRUCTION TUNING

3.1 Sơ lược về InstructBLIP

Như đã đề cập ở chương 2, một đặc điểm mà BLIP-2 vẫn cải thiện được nằm ở khả năng trích xuất đặc trưng ảnh của Q-Former. Việc chỉ huấn luyện Q-Former với cặp ảnh và văn bản miêu tả khiến cho nhiều thông tin trích xuất không liên quan đến prompt (hay còn gọi là instruction) của người dùng, dẫn đến việc câu trả lời có thể không đúng trọng tâm. Bởi vậy, InstructBLIP đã ra đời với mục tiêu cải tiến BLIP-2 bằng phương pháp instruction tuning.

3.2 Kiến trúc InstructBLIP

Để giải quyết vấn đề thì InstructBLIP cho phép prompt cũng được đưa vào Q-Former để trích xuất thông tin ảnh. Điều này giúp cho thông tin LLM nhận được “chất lượng” hơn, qua đó khiến cho câu trả lời nhắm đúng vào nhu cầu của người dùng.



Hình 3. 1 Kiến trúc tổng quát của InstructBLIP

Ý tưởng InstructBLIP chỉ có vậy, tuy rất đơn giản nhưng việc fine tune dựa trên instruction đã cải thiện đáng kể hiệu suất của BLIP-2. Tiếp theo, ta sẽ xem qua một số đánh giá, kết quả thu được của mô hình.

3.3 Kết quả đánh giá

Đối với zero-shot evaluation, nhóm tác giả thực hiện so sánh với những mô hình thuộc họ BLIP-2 và Flamingo. Tập data sử dụng để đánh giá cũng là tập validation data đề cập ở trên. Trong bảng 1, ta có thể thấy InstructBLIP vượt qua BLIP-2 và Flamingo ở mọi tập data, minh chứng cho tính hiệu quả của instruction tuning. Cụ thể, InstructBLIP FlanT5XL cải thiện 15% so với BLIP-2 FLanT5XL. Thậm chí với một số

tác vụ mà InstructBLIP không được huấn luyện (ví dụ như video QA, tập MSRVT QA), InstructBLIP cũng cải thiện khoảng 47.1% so với các mô hình SOTA trước đó. Bên cạnh đó, mô hình InstructBLIP FlanT5 nhỏ nhất (4 tỉ tham số) cũng vượt qua mô hình Flamingo-80B (80 tỉ tham số) với mức độ cải thiện trung bình 24.8%.

	NoCaps	Flickr 30K	GQA	VSR	IconQA	TextVQA	Visdial	HM	VizWiz	SciQA image	MSVD QA	MSRVTT QA	iVQA
Flamingo-3B [4]	-	60.6	-	-	-	30.1	-	53.7	28.9	-	27.5	11.0	32.7
Flamingo-9B [4]	-	61.5	-	-	-	31.8	-	57.0	28.8	-	30.2	13.7	35.2
Flamingo-80B [4]	-	67.2	-	-	-	35.0	-	46.4	31.6	-	35.6	17.4	40.7
BLIP-2 (FlanT5 _{XL}) [20]	104.5	76.1	44.0	60.5	45.5	43.1	45.7	53.0	29.8	54.9	33.7	16.2	40.4
BLIP-2 (FlanT5 _{XXL}) [20]	98.4	73.7	44.6	68.2	45.4	44.1	46.9	52.0	29.4	64.5	34.4	17.4	45.8
BLIP-2 (Vicuna-7B)	107.5	74.9	38.6	50.0	39.7	40.1	44.9	50.6	25.3	53.8	18.3	9.2	27.5
BLIP-2 (Vicuna-13B)	103.9	71.6	41.0	50.9	40.6	42.5	45.1	53.7	19.6	61.0	20.3	10.3	23.5
InstructBLIP (FlanT5 _{XL})	119.9	84.5	48.4	64.8	50.0	46.6	46.6	56.6	32.7	70.4	43.4	25.0	53.1
InstructBLIP (FlanT5 _{XXL})	120.0	83.5	47.9	65.6	51.2	46.6	48.5	54.1	30.9	70.6	44.3	25.6	53.8
InstructBLIP (Vicuna-7B)	123.1	82.4	49.2	54.3	43.1	50.1	45.2	59.6	34.5	60.5	41.8	22.1	52.2
InstructBLIP (Vicuna-13B)	121.9	82.8	49.5	52.1	44.8	50.7	45.4	57.5	33.4	63.1	41.2	24.8	51.0

Hình 3. 2 Kết quả zero-shot evaluation






3.4 Kết luận

BLIP, BLIP-2 và InstructBLIP đều là những phương pháp tận dụng sức mạnh từ những mô hình đơn nhiệm để phát triển khả năng đa nhiệm. So với BLIP, BLIP-2 tiết kiệm chi phí bằng cách huấn luyện một ông “biên dịch” Q-Former để tận dụng 2 mô hình Visual Transformer và Large Language Model. InstructBLIP lại cải tiến BLIP-2 thêm một bước xa hơn nữa bằng cách huấn luyện Q-Former trích xuất thông tin chất lượng và liên quan đến prompt, qua đó cải thiện khả năng trả lời đúng trọng tâm của mô hình. Nhờ vào việc tiết kiệm chi phí huấn luyện mà lại cải thiện hiệu suất mà những mô hình kể trên thường được tin dùng cho bài toán Vision-Language.

CHƯƠNG 4 – THỰC NGHIỆM

4.1 Xây dựng dữ liệu

Bộ dữ liệu mà nhóm sử dụng là các cặp ảnh – mô tả về 3 món ăn của Việt Nam gồm: bánh bèo, bún bò Huế và bún đậu mắm tôm. Nhóm thực hiện việc cào dữ liệu bằng thư viện BeautifulSoup4, Request,... từ các trang như foody.vn, cooky.vn. Sau đó tạo metadata cho bộ dữ liệu và đưa lên hugging face.

	A plate of banh beo, a Vietnamese rice cake dish, is topped with dried shrimp, crispy fried shallots, and chopped scallions. A small bowl of dipping sauce sits next to the plate.
	A platter of small bowls filled with Banh Beo, a Vietnamese rice cake dish, each topped with crispy pork rinds and green onions. A side of dipping sauce completes this delicious dish.
	A close-up of multiple small white bowls filled with banh beo, a Vietnamese rice cake dish. Each bowl is topped with crispy pork rinds, green onions, and a sprinkle of chili powder.
	A red tray filled with small white bowls of Banh Beo, a Vietnamese rice cake dish. Each bowl is topped with crispy pork rinds, dried shrimp, and scallions, and a dipping sauce is served.
	A round bamboo tray filled with small bowls of Banh Beo, a Vietnamese rice cake dish, each topped with crispy pork rinds and green onions.

Hình 4.1: Bộ dữ liệu món ăn

Bộ dữ liệu gồm 90 cặp ảnh và mô tả (mỗi món là 30 cặp) được nhóm thực hiện tạo sinh bằng công cụ AI (gemini advanced) được thực hiện rà soát lại về ngữ nghĩa.

4.2 Huấn luyện mô hình

Nhóm sử dụng mô hình blip2 được pretrained với 2.7 tỷ tham số bởi đội ngũ Salesforce. Trong đó các tham số gồm:

- load_in_8bit: Giảm kích thước và chi phí tính toán của mô hình.
- llm_int8_threshold: Ngưỡng thực hiện quantize.

```
processor = AutoProcessor.from_pretrained("Salesforce/blip2-opt-2.7b")
quantization_config = BitsAndBytesConfig(
    load_in_8bit=True, # Enable 8-bit quantization
    llm_int8_threshold=6.0 # Optional: set a threshold for which layers to quantize (default is 6.0)
)
model = Blip2ForConditionalGeneration.from_pretrained("ybelkada/blip2-opt-2.7b-fp16-sharded", device_map="auto", quantization_config=quantization_config)
```

Hình 4.2: Sử dụng mô hình đã được pretrained

Tiếp theo, ta sẽ tạo LoraConfig. LoRA (Low-Rank Adaption) là một phương pháp giúp fine-tuning mô hình hiệu quả hơn bằng cách thêm các low-rank adapter vào một số lớp của mô hình.

```

config = LoraConfig(
    r=16,
    lora_alpha=32,
    lora_dropout=0.05,
    bias="none",
    target_modules=["q_proj", "k_proj"]
)

model = get_peft_model(model, config)
model.print_trainable_parameters()

```

trainable params: 5,242,880 || all params: 3,749,922,816 || trainable%: 0.1398

Hình 4.3: Cấu hình LoRA cho mô hình

Trong đó, các tham số gồm:

- `r`: Giá trị càng cao mô hình học được nhiều thông tin hơn và ngược lại.
- `lora_alpha`: Giá trị điều khiển learning rate của mô hình.
- `lora_dropout`: Giá trị giúp mô hình tránh mô hình bị overfitting.
- `bias`: Giá trị bằng none tức không có bias
- `target_modules`: Các module được thực hiện LoRA gồm query projection và key projection.

Nhóm thực hiện huấn luyện mô hình dựa trên tập dữ liệu đã xây dựng như sau:

```

train_dataset = ImageCaptioningDataset(dataset, processor)
train_dataloader = DataLoader(train_dataset, shuffle=True, batch_size=3, collate_fn=collate_fn)
optimizer = torch.optim.Adam(model.parameters(), lr=5e-4)
device = "cuda" if torch.cuda.is_available() else "cpu"
model.train()

for epoch in range(100):
    print("Epoch:", epoch)
    for idx, batch in enumerate(train_dataloader):
        input_ids = batch.pop("input_ids").to(device)
        pixel_values = batch.pop("pixel_values").to(device, torch.float16)

        outputs = model(input_ids=input_ids,
                        pixel_values=pixel_values,
                        labels=input_ids)

        loss = outputs.loss

        print("Loss:", loss.item())

        loss.backward()

        optimizer.step()
        optimizer.zero_grad()

```

Hình 4.4: Huấn luyện mô hình

Các tham số huấn luyện gồm:

- `batch_size`: Số sample mỗi batch.
- `optimizer`: Thuật toán Adam với `learning_rate=0.0004`.
- `device`: Sử dụng GPU để đẩy nhanh quá trình huấn luyện.
- `epoch`: 100.

4.3 Đánh giá

BLEU score (trên tập train):

BLEU Score: {'bleu': 0.7939087381333401, 'precisions': [0.9897959183673469, 0.9869816779170685, 0.98393574
ROUGE Score: {'rouge1': AggregateScore(low=Score(precision=0.9585405131980893, recall=0.8156379593132727,

Sau đây là một số kết quả thử nghiệm từ các ảnh bất kỳ trên mạng:

Text(0.5, 1.0, 'Generated caption: A steaming bowl of bún bò Huế, a popular Vietnamese noodle soup, featuring tender slices of beef, pork knuckle, and a flavorful broth garnished with fresh herbs and chili peppers.')
Generated caption: A steaming bowl of bún bò Huế, a popular Vietnamese noodle soup, featuring tender slices of beef, pork knuckle, and a flavorful broth garnished with fresh herbs and chili peppers.



Hình 4.5: Kết quả thử nghiệm với ảnh món bún bò Huế

Text(0.5, 1.0, 'Generated caption: A top-down view of a Vietnamese dish called "bún đậu mắm tôm" featuring vermicelli noodles, fried tofu, boiled pork, Vietnamese sausage, chả cốm (fried rice cake with meat), fresh herbs, and shrimp paste, served on banana leaves.')
Generated caption: A top-down view of a Vietnamese dish called "bún đậu mắm tôm" featuring vermicelli noodles, fried tofu, boiled pork, Vietnamese sausage, chả cốm (fried rice cake with meat), fresh herbs, and shrimp paste, served on banana leaves.



Hình 4.6: Kết quả thử nghiệm với ảnh món bún đậu mắm tôm

Text(0.5, 1.0, 'Generated caption: A hand holds a bowl of banh beo, small steamed rice cakes, topped with crispy fried shallots, dried shrimp, and chopped scallions.')
Generated caption: A hand holds a bowl of banh beo, small steamed rice cakes, topped with crispy fried shallots, dried shrimp, and chopped scallions.



Hình 4.7: Kết quả thử nghiệm với ảnh món bánh bèo

TÀI LIỆU THAM KHẢO

Tiếng Việt:

[1]: Mô hình BLIP - Nguyễn Thành Đạt – 03/08/2023

Tiếng Anh:

[1]: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, by Junnan Li, Dongxu Li, Caiming Xiong, Steven Hoi, at Salesforce Research, publish by arxiv:2201.12086 – 15 Feb 2022.

[2]: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, by Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi, at Salesforce Research, publish by arxiv:2301.12597v3 – 15 Jun 2023.

[3]: InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, by Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, Steven Hoi, at Salesforce Research, Hong Kong University of Science and Technology, Nanyang Technological University (Singapore), publish by arxiv:2305.06500v2 – 15 Jun 2023.