IMAGE CAPTIONING IN VIETNAMESE CONTEXT

Môn học: CS420 – Các vấn đề chọn lọc trong Thị giác máy tính

Giảng viên hướng dẫn:

TS. Mai Tiến Dũng

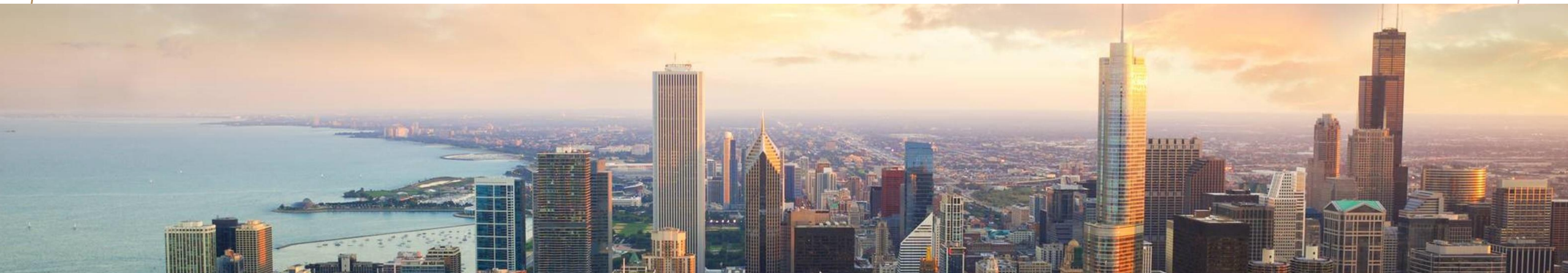
ThS. Đỗ Văn Tiến

Danh sách thành viên

- Phạm Quốc Việt – 21522792
- Trần Hoài An – 21520553
- Lường Đại Phát – 21522443

Mục lục

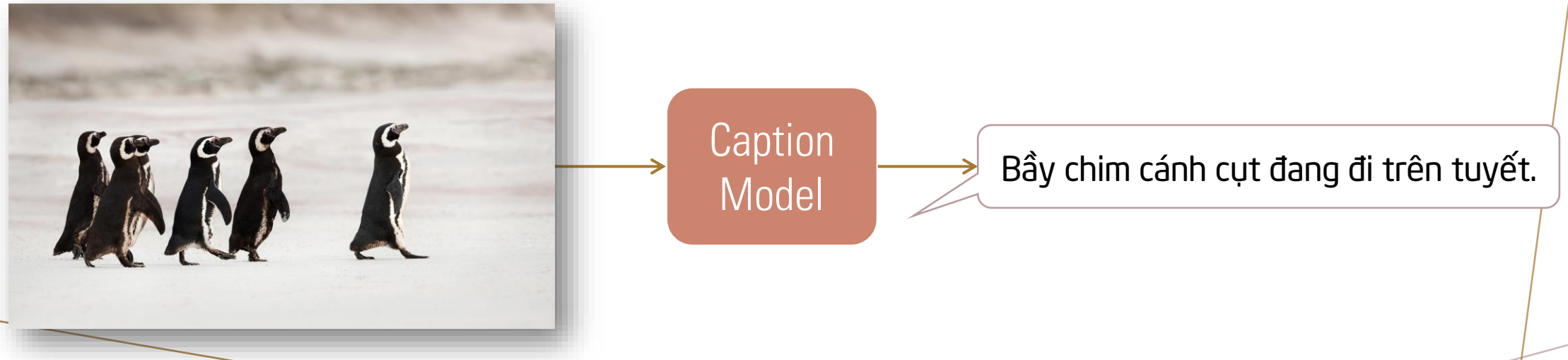
I. GIỚI THIỆU



I. Giới thiệu

Image captioning là gì?

Image captioning là quá trình tự động tạo ra mô tả văn bản cho hình ảnh. Nó kết hợp các kỹ thuật trong xử lý ngôn ngữ tự nhiên (NLP) và thị giác máy tính (Computer Vision) để phân tích nội dung hình ảnh và diễn đạt thành câu hoàn chỉnh, thường mô tả các đối tượng, hành động, hoặc ngữ cảnh trong ảnh.



I. GIỚI THIỆU

Ứng dụng của Image Captioning:

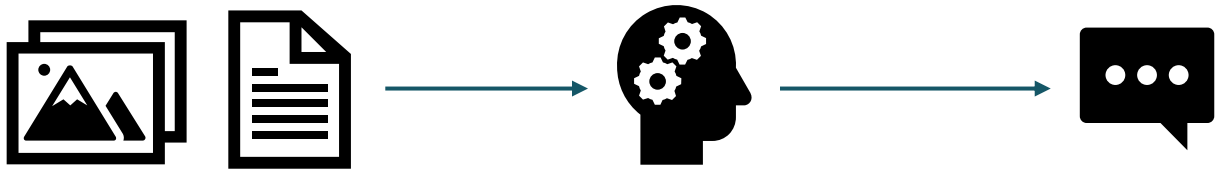
- Cải thiện khả năng tìm kiếm hình ảnh qua văn bản bằng cách dựa vào thông tin mô tả của bức ảnh thay vì dựa vào tên ảnh, thẻ hoặc nội dung trang chứa hình ảnh đó.

- Giúp cho người khiếm thị hiểu được nội dung của hình ảnh thông qua mô tả (mô tả được biểu diễn qua chữ nổi hoặc qua âm thanh).

I. Giới thiệu

Input và Output của bài toán

Input	Output
- Dataset gồm n ảnh và k caption cho mỗi ảnh (k >= 1). - Một ảnh đầu vào I là một ma trận 3 chiều $I \in R^{H \times W \times C}$ Trong đó: - H: Chiều cao của ảnh - W: Chiều rộng của ảnh - C: Số kênh màu của ảnh (C = 3)	Chuỗi từ $S = w_1, w_2, w_3, \ldots, w_T$, trong đó: - T: độ dài của chuỗi S. - w_i: Từ thứ i trong chuỗi, được lấy từ một tập từ vựng V cố định.



I. GIỚI THIỆU

Mục tiêu của đề tài:

- Tìm hiểu kiến thức cơ bản về Image Captioning
- Xây dựng mô hình Image Captioning hiệu quả
- Ứng dụng thực tế

2. DATASET VÀ KIỂM THỬ

2. DATASET VÀ KIỂM THỬ

Dataset được sử dụng trong quá trình huấn luyện mô hình:

UIT-ViIC:

- Bộ dataset với chú thích tiếng Việt cho các ảnh từ dataset Microsoft COCO có nội dung liên quan đến các môn thể thao sử dụng bóng.
- Chứa 19250 chú thích cho 3850 ảnh khác nhau.

Phân chia các tập Train, Val, Test:

- Train dataset: gồm 2695 ảnh và 13481 caption.
- Val dataset: gồm 924 ảnh và 4620 caption.
- Test dataset: gồm 231 ảnh và 1155 caption.

Caption
Người đàn ông áo đang chuẩn bị phát bóng ở trên sân tennis.
Người áo đen đang quan sát động tác phát bóng của người đàn ông áo.
Hai vận động viên tennis đang luyện tập trên sân.
Một người đàn ông và một người khác đang chơi tennis ở trên sân.
Một người đàn ông đang tung quả bóng tennis lên cao để chuẩn bị phát.

2. DATASET VÀ KIỂM THỬ

Quá trình tiền xử lí dữ liệu:

Đối với ảnh: chuyển đổi về một kích thước phù hợp với input của các mô hình (224x224x3).

Đối với các câu mô tả:

- Loại bỏ kí tự đặc biệt.
- Lowercasing (VD: Bóng đá -> bóng đá).
- Word segmentation: tách một chuỗi liên tiếp các kí tự thành các từ riêng lẻ.
  (VD: “người đàn ông” -> [“người”, “đàn”, “ông”]
- Loại bỏ khoảng trắng thừa.
- Thêm các token đặc biệt vào đầu và cuối câu (VD: “quần vợt” -> “startseq quần vợt endseq”.
- Loại bỏ các dữ liệu bị rỗng (VD: có ảnh nhưng không có caption).

2. DATASET VÀ KIỂM THỬ

Phương pháp đánh giá BLEU score

BLEU (Bilingual Evaluation Understudy) là một chỉ số được sử dụng rộng rãi để đánh giá chất lượng của các hệ thống xử lý ngôn ngữ tự nhiên. BLEU đo lường mức độ tương đồng giữa văn bản do máy sinh ra (machine-generated text) và một hoặc nhiều văn bản tham chiếu (reference text) được con người viết.

BLEU so sánh n-grams (chuỗi n từ liên tiếp) giữa văn bản sinh ra và văn bản tham chiếu. Điểm số BLEU cao hơn biểu thị rằng văn bản sinh ra gần giống hơn với văn bản tham chiếu.

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

2. DATASET VÀ KIỂM THỬ

Phương pháp đánh giá ROUGE score

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) là một tập hợp các chỉ số được sử dụng để đánh giá chất lượng của văn bản sinh ra bởi các mô hình xử lý ngôn ngữ tự nhiên. Nó đo lường mức độ tương đồng giữa văn bản sinh ra và một hoặc nhiều văn bản tham chiếu. Các biến thể: ROUGE-N, ROUGE-L, ROUGE-W,…

Công thức ROUGE-L:

$$P = \frac{LCS(X,Y)}{|X|}$$

$$R = \frac{LCS(X,Y)}{|Y|}$$

$$F = \frac{(1+\beta^2).P.R}{\beta^2.P+R}$$

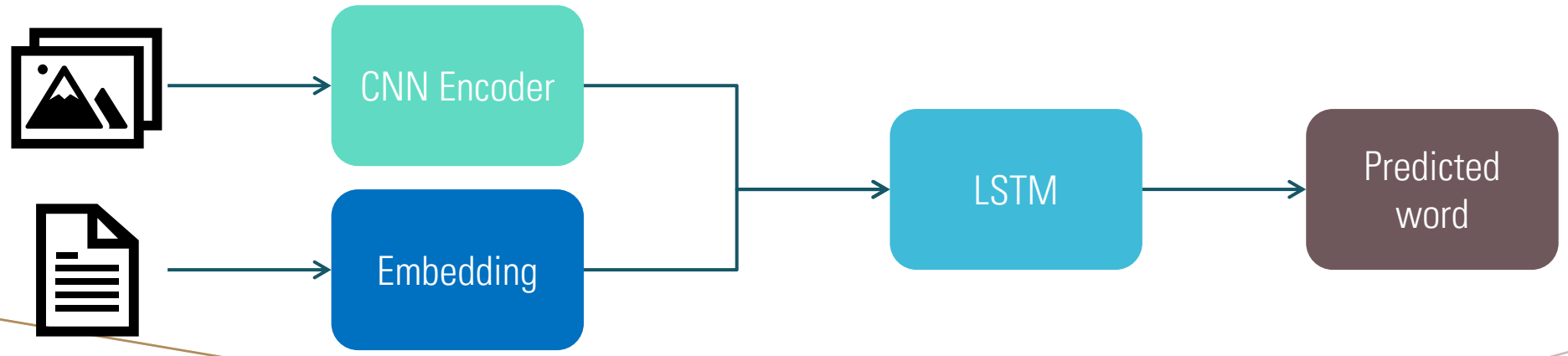
3. PHƯƠNG PHÁP

3. Phương pháp

CNN + LSTM

Mô hình gồm 2 phần chính:

- CNN Encoder: Sử dụng một mô hình CNN (VGG, DenseNet, EfficientNet) để trích xuất đặc trưng của ảnh.
- LSTM Decoder: Mã hóa các từ trong ngữ cảnh thành các vector embedding. Kết hợp đặc trưng ảnh (từ CNN) và đặc trưng câu (từ Embedding) dọc theo trục thời gian. Bước thời gian đầu tiên đại diện cho ảnh, và các bước tiếp theo đại diện cho các từ trong câu. LSTM nhận input chuỗi thời gian, tạo ra một vector ẩn biểu diễn thông tin ngữ cảnh từ cả hình ảnh và câu đầu vào. Mô hình sẽ sử dụng thông tin ngữ cảnh này để dự đoán từ tiếp theo của câu.

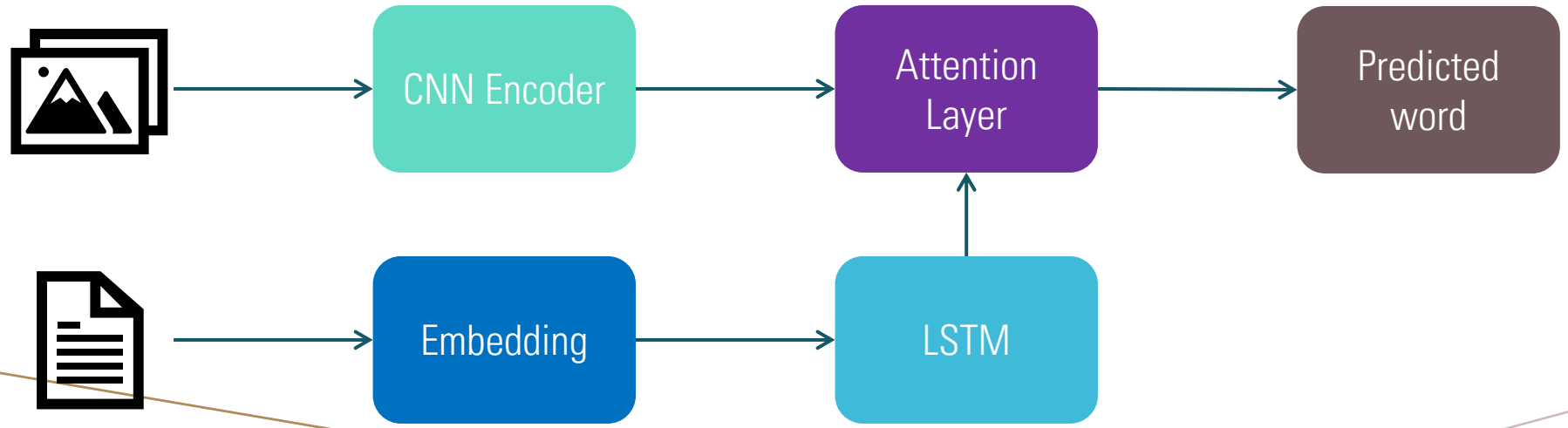


3. Phương pháp

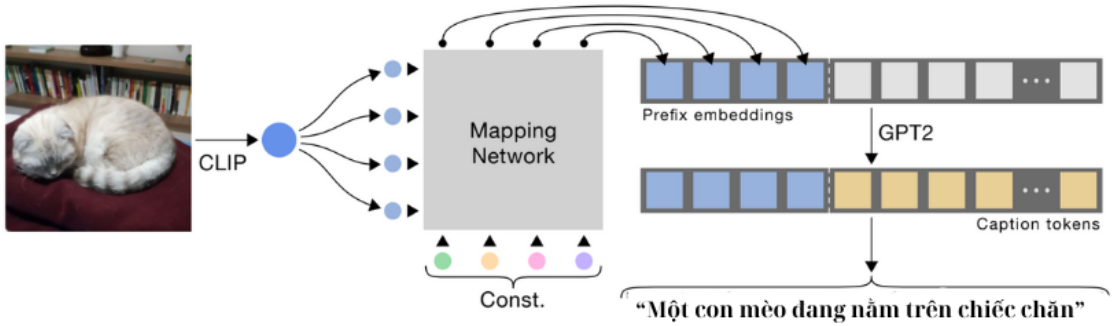
CNN + LSTM with Attention

Mô hình gồm 3 phần chính:

- CNN Encoder: Sử dụng một mô hình CNN (VGG, DenseNet, EfficientNet) để trích xuất đặc trưng của ảnh.
- LSTM Decoder: Mã hóa các từ trong ngữ cảnh thành các vector embedding. LSTM nhận input là chuỗi các vector embedding, tạo ra một vector ẩn biểu diễn thông tin ngữ cảnh từ câu đầu vào.
- Bahdanau Attention: áp dụng để mô hình tập trung vào các phần quan trọng của đặc trưng ảnh dựa trên ngữ cảnh của câu.



3. Phương pháp



ClipCap

Mô hình ClipCap sử dụng ViT (Vision Transformer) cho việc trích xuất đặc trưng hình ảnh + GPT-2 cho việc đọc hiểu embedding ảnh và chuyển đổi sang caption tiếng Việt. Cầu nối giữa encoder và decoder là một Mapping Network (mạng lưới chuyển đổi đầu ra của ViT sao cho phù hợp với GPT-2).

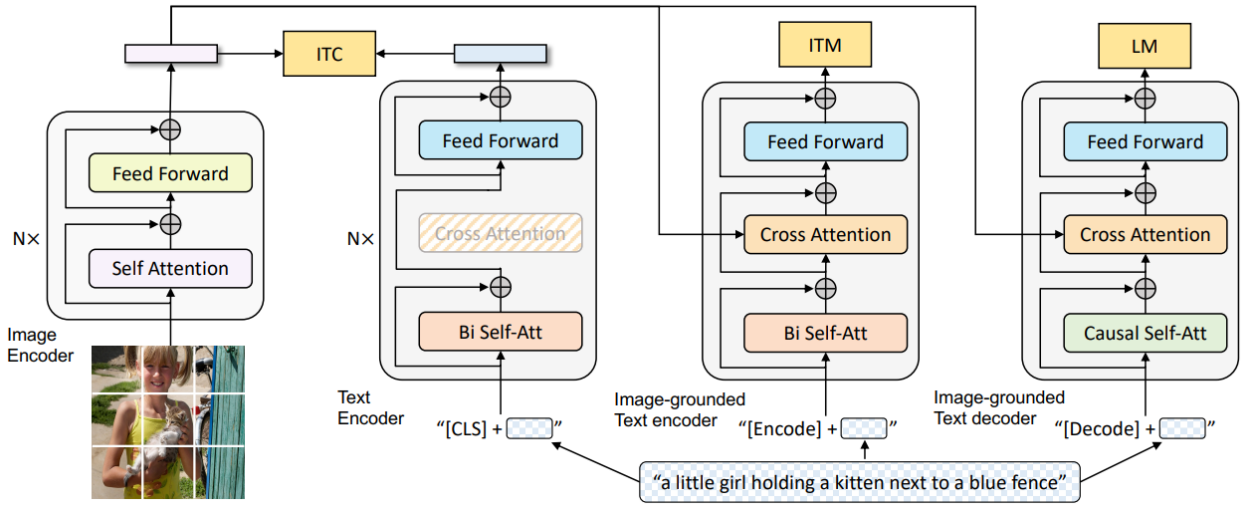
Encoder: Vision Transformer (ViT)

- ViT chia hình ảnh đầu vào thành các patch (mảnh nhỏ).
- Biến đổi các patch này thành các vector đặc trưng thông qua các lớp transformer.
- Kết quả của ViT là một vector đặc trưng biểu diễn toàn bộ hình ảnh.

Decoder: GPT-2-vietnamesee

- GPT-2-vietnamese là một mô hình ngôn ngữ đã được pretrained dành cho tiếng việt dựa trên transformer, được tối ưu để sinh văn bản tự nhiên.
- GPT-2-vietnamese nhận các đặc trưng từ ViT (qua các vector đầu ra của encoder) và sinh một chuỗi văn bản (caption) mô tả hình ảnh.

3. Phương pháp



BLIP

Sử dụng ViT (Vision Transformer) cho phần Image Encoder và phần còn lại là hỗn hợp đa phương thức của Encoder-Decoder (MED), bao gồm 3 phần sau:

- Unimodal encoder: mã hóa riêng biệt các phần hình ảnh và văn bản, cấu trúc phần text encoder tương tự như BERT.

- Image-grounded text encoder: mã hóa văn bản dựa trên hình ảnh, chèn từng transformer block vào lớp Cross-attention (CA) giữa Self-attention (SA) và Feed Forward Network (FFN).

- Image-grounded text decoder: giải mã văn bản dựa trên hình ảnh, thay thế lớp bi-directional Self-attention thành casual Self-attention.

3. Phương pháp

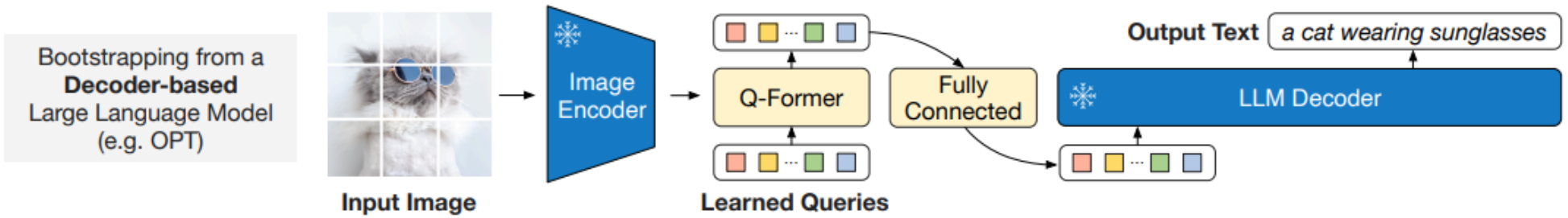
BLIP-2 (PEBT)

Image Encoder: sử dụng ViT cho phần mã hóa thông tin hình ảnh. ViT-g và ViT-L được sử dụng trong bài báo.

Q-Former: Là cầu nối giữa Image Encoder và LLM Decoder, sử dụng 2 phần transformer: một image transformer để xử lý thông tin từ Image Encoder, một text transformer cho text encoder và text decoder.

LLM Decoder: sử dụng mô hình ngôn ngữ lớn cho phần chuyển đổi thông tin mã hóa thành văn bản phù hợp với hình ảnh. OPT và FlanT5 được sử dụng trong bài báo.

PEFT: Vì mô hình BLIP-2 có tham số rất lớn (3.4 tỉ tham số) và tài nguyên của nhóm có hạn, nhóm sử dụng PEFT (Parameter-Efficient Fine-Tuning) cụ thể LoRA nhằm giảm thiểu các tham số huấn luyện để dễ dàng huấn luyện và sử dụng mô hình với tài nguyên hiện có, nhưng đồng thời phải đánh đổi độ chính xác của mô hình.



4. THỰC NGHIỆM

4. THỰC NGHIỆM

Kết quả thực nghiệm trên tập dataset ViIC (learning rate default):

	Type	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
CNN+LSTM	DenseNet201	67.21	52.49	42.76	35.39	47.21
	VGG16	61.67	45.3	35.53	28.17	45.13
	EfficientNetV2	56.84	40.06	31.68	24.68	41.6
CNN+LSTM (attention)	DenseNet201	65.98	52.58	44.92	39.04	48.0
	VGG16	56.84	44.93	38.22	32.69	41.6
	EfficientNetV2	56.84	44.93	38.22	32.69	41.6
ClipCap	ViT-g + GPT2	43.42	40.25	37.24	35.63	42.12
BLIP	ViT-b + BERT-b	63.5	52.06	43.5	36.06	49.01
BLIP-2	ViT-g + OPT 2.7B + PEFT (int8)	53.78	43.2	35.9	29.47	48.06

4. THỰC NGHIỆM

Kết quả thực nghiệm trên tập dataset ViIC (learning rate = 1e-4):

	Type	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
CNN+LSTM	DenseNet201	69.85	56.03	45.97	37.93	48.73
	VGG16	63.11	47.11	37.68	30.1	45.64
	EfficientNetV2	56.84	40.06	31.68	24.68	41.6
CNN+LSTM (attention)	DenseNet201	69.24	57.44	49.47	42.99	48.03
	VGG16	62.31	49.93	42.74	36.97	45.08
	EfficientNetV2	56.84	44.93	38.22	32.69	41.6
ClipCap	ViT-g + GPT2	43.42	40.25	37.24	35.63	42.12
BLIP	ViT-b + BERT-b	0	0	0	0	0
BLIP-2	ViT-g + OPT 2.7B + PEFT (int8)	23.77	20.6	17.78	14.75	39.88

4. THỰC NGHIỆM

Kết quả thực nghiệm trên tập dataset ViIC (learning rate = 5e-4):

	Type	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
CNN+LSTM	DenseNet201	68.0	54.82	46.04	38.77	48.13
	VGG16	63.13	47.11	37.45	29.76	45.22
	EfficientNetV2	56.84	40.06	31.68	24.68	41.6
CNN+LSTM (attention)	DenseNet201	65.68	52.62	45.19	39.29	48.18
	VGG16	54.75	42.66	36.06	30.85	41.43
	EfficientNetV2	56.84	44.93	38.22	32.69	41.6
ClipCap	ViT-g + GPT2	43.42	40.25	37.24	35.63	42.12
BLIP	ViT-b + BERT-b	0	0	0	0	0
BLIP-2	ViT-g + OPT 2.7B + PEFT (int8)	53.78	43.2	35.9	29.47	48.06

4. THỰC NGHIỆM

Thời gian thực nghiệm trên tập dataset ViIC:

	Type	Training time	Trainable Params
CNN+LSTM	DenseNet201	38m	1,570,119
	VGG16	38m	1,570,119
	EfficientNetV2	36m	1,570,119
CNN+LSTM (attention)	DenseNet201	18m40s	1,603,143
	VGG16	17m47s	1,603,143
	EfficientNetV2	23m42s	1,603,143
ClipCap	ViT-g + GPT2	6h44m	124,439,808
BLIP	ViT-b + BERT-b	4h17m	161,323,580
BLIP-2	ViT-g + OPT 2.7B + PEFT (int8)	2h55m	5,242,880

5. KẾT LUẬN

5. KẾT LUẬN

Hiệu quả mô hình trên ViIC

- Các mô hình sử dụng DenseNet201 kết hợp với LSTM hoặc LSTM (attention) đạt hiệu suất cao nhất, đặc biệt khi tinh chỉnh lại learning rate, cho thấy khả năng trích xuất đặc trưng hình ảnh hiệu quả của DenseNet201 trong Image Captioning.
- Mô hình sử dụng VGG16 và EfficientNet cho kết quả thấp hơn, đặc biệt với chỉ số BLEU-4, phản ánh rằng chúng không phù hợp bằng DenseNet201 trong việc biểu diễn dữ liệu hình ảnh cho nhiệm vụ này.
- Các mô hình dựa trên ViT (Vision Transformer) như BLIP và ClipCap đạt hiệu suất tốt hơn ở một số khía cạnh nhưng chưa vượt qua CNN+LSTM trong tác vụ tạo chú thích chính xác.

Ảnh hưởng của Attention

Việc bổ sung Attention vào kiến trúc LSTM giúp cải thiện các chỉ số BLEU và ROUGE-L, nhấn mạnh tầm quan trọng của việc tập trung vào các vùng quan trọng trong hình ảnh khi sinh chú thích.

Các mô hình tiên tiến

Mặc dù BLIP-2 (sử dụng ViT + OPT) cho thấy cải tiến trong một số chỉ số, cho thấy tiềm năng của các kiến trúc Transformer. Tuy nhiên, hiệu suất tổng thể vẫn chưa vượt qua CNN+LSTM khi không có tối ưu hóa cụ thể.

5. KẾT LUẬN

Khó khăn gặp phải

- Tối ưu hóa mô hình: Việc lựa chọn kiến trúc, tham số phù hợp để tối ưu hóa mô hình là một thử thách, đặc biệt với các mô hình phức tạp như ViT.
- Giới hạn phần cứng: Khi huấn luyện các mô hình lớn sử dụng Transformer, thời gian huấn luyện và dự đoán lâu hơn so với mô hình thông thường.
- Khó khăn với dữ liệu lớn: Khi thử nghiệm với bộ dataset lớn hơn như Flickr 8k, mô hình DenseNet gặp khó khăn trong việc huấn luyện khi sử dụng batch_size nhỏ dẫn đến việc ước tính gradient trở nên kém chính xác, gradient dao động nhiều hơn. Điều này có thể khiến mô hình khó hội tụ hoặc cần nhiều thời gian hơn để đạt được trạng thái tối ưu. Nguyên nhân của việc sử dụng batch_size nhỏ đến từ giới hạn phần cứng khi phải huấn luyện trên tập dữ liệu lớn.

THANK YOU