THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/KpejggY7ce4*
- Link slides (dạng .pdf đặt trên Github của nhóm): http://github.com/vunha32/CS2205.CH201-Image-Captioning/ViClipCap

- Họ và Tên: Nguyễn Hà Anh Vũ
- MSSV: 250101077

- Lớp: CS2205.CH201
- Tự đánh giá (điểm tổng kết môn): 9.5/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 5
- Link Github: https://github.com/vunha32/CS2205.CH201-Image-Captioning

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA):

VICLIPCAP: TINH CHỈNH TIỀN TỐ CLIP CHO MÔ HÌNH GPT-2 TRONG BÀI TOÁN SINH MÔ TẢ ẢNH TIẾNG VIỆT

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA):

VICLIPCAP: CLIP-BASED PREFIX TUNING FOR GPT-2 IN VIETNAMESE IMAGE CAPTIONING

TÓM TẮT

Việc chuyển đổi thông tin từ dạng hình ảnh sang ngôn ngữ tự nhiên, hay còn gọi là Image Captioning, đòi hỏi hệ thống phải thấu hiểu đồng thời cả ngữ cảnh thị giác lẫn quy tắc ngôn ngữ. Dù cộng đồng nghiên cứu đã đạt được nhiều bước tiến, các kiến trúc mô hình thế hệ cũ vẫn bộc lộ hạn chế rõ rệt về mặt hiệu suất khi đòi hỏi tài nguyên tính toán quá lớn và quy trình huấn luyện phức tạp. Vì vậy, nhu cầu cấp thiết hiện nay là tìm kiếm một hướng tiếp cận mới tối ưu hơn về mặt chi phí và tài nguyên.

Trong nghiên cứu này, phương pháp sinh mô tả ảnh bằng tiếng Việt được đề xuất hiệu quả dựa trên kiến trúc ClipCap, tập trung tận dụng sức mạnh của các mô hình đã được huấn luyện trước (pre-trained models). Cụ thể, hệ thống vận hành bằng cách sử dụng mạng CLIP (Contrastive Language-Image Pre-training) để trích xuất các đặc trưng thị giác giàu ngữ nghĩa làm "tiền tố" (prefix) định hướng, sau đó đưa qua một Mạng Ánh xạ (Mapping Network) để chuyển đổi không gian đặc trưng trước khi nhập vào mô hình ngôn ngữ GPT-2 đã được huấn luyện trên dữ liệu tiếng Việt (Vietnamese GPT-2). Sự kết hợp này tạo ra một luồng xử lý thông tin liền mạch từ hình ảnh sang ngôn ngữ đích là tiếng Việt.

Phương pháp này mang lại lợi thế kỹ thuật rõ rệt khi cho phép bảo toàn (đóng băng) trọng số của CLIP và GPT-2, chỉ việc huấn luyện mạng ánh xạ đơn giản, giúp giảm thiểu đáng kể thời gian và chi phí tính toán so với các phương pháp train-from-scratch. Nhờ áp dụng chiến lược huấn luyện tinh gọn, mô hình kỳ vọng đạt được khả năng sinh câu mô tả tiếng Việt tự nhiên, đúng ngữ pháp và phản ánh chính xác ngữ cảnh của ảnh đầu vào.

GIỚI THIỆU

1. Bối cảnh và Tầm quan trọng: Bài toán sinh mô tả ảnh (Image Captioning) đóng vai trò then chốt tại giao điểm của Thị giác máy tính và Xử lý ngôn ngữ tự nhiên, nhằm trang bị cho máy móc khả năng thấu hiểu và diễn giải thị giác. Trong bối cảnh dữ liệu số bùng nổ, yêu cầu đặt ra không dừng lại ở việc phát hiện đối tượng đơn lẻ mà còn phải diễn giải được sự tương tác giữa chúng. Điều này có ý nghĩa thiết thực trong việc xây dựng các giải pháp trợ năng (accessibility) cho người khiếm thị hoặc nâng cao khả năng truy xuất hình ảnh thông minh. Chính vì thế, nhiệm vụ xây dựng các hệ thống sinh mô tả sát thực tế là chìa khóa để xóa bỏ rào cản ngữ nghĩa giữa dữ liệu thị giác và ngôn ngữ tự nhiên.

2. Thách thức và Lý do chọn ClipCap: Việc triển khai các mô hình đa phương thức tiên tiến cho tiếng Việt gặp nhiều thách thức, do thiếu hụt dữ liệu gán nhãn quy mô lớn và chi phí phần cứng cao khi tinh chỉnh các mô hình có hàng tỷ tham số. Trong khi các phương pháp truyền thống đòi hỏi huấn luyện từ đầu trên tập dữ liệu khổng lồ, kiến trúc ClipCap (2021) nổi lên như giải pháp tối ưu nhờ tính linh hoạt và kiến trúc mô-đun hóa. Bằng cách tận dụng các mô hình pre-trained sẵn có thay vì xây dựng một kiến trúc nguyên khối (end-to-end). Cách tiếp cận này giúp việc thay đổi module ngôn ngữ trở nên linh hoạt hơn, đồng thời tối ưu hóa được cả hiệu suất lẫn khả năng triển khai thực tế đối với dữ liệu tiếng Việt.

3. Giải pháp đề xuất: Nghiên cứu đề xuất áp dụng kiến trúc ClipCap kết hợp sức mạnh trích xuất đặc trưng của CLIP và khả năng sinh văn bản của mô hình GPT-2 thuần Việt (Vietnamese GPT-2). Điểm mấu chốt của phương pháp là sử dụng Mạng Ánh xạ đóng vai trò cầu nối. Mạng này biến đổi các đặc trưng ảnh sang dạng vector tiền tố (prefix) dùng để gợi ý cho mô hình ngôn ngữ, nhờ đó chúng ta không bắt buộc phải huấn luyện lại (fine-tune) toàn bộ trọng số của CLIP hay GPT-2. Phương pháp "đóng băng" giúp tối ưu hóa chi phí tính toán, tận dụng tri thức sẵn có để sinh ra các câu mô tả tiếng Việt tự nhiên và sát nghĩa.

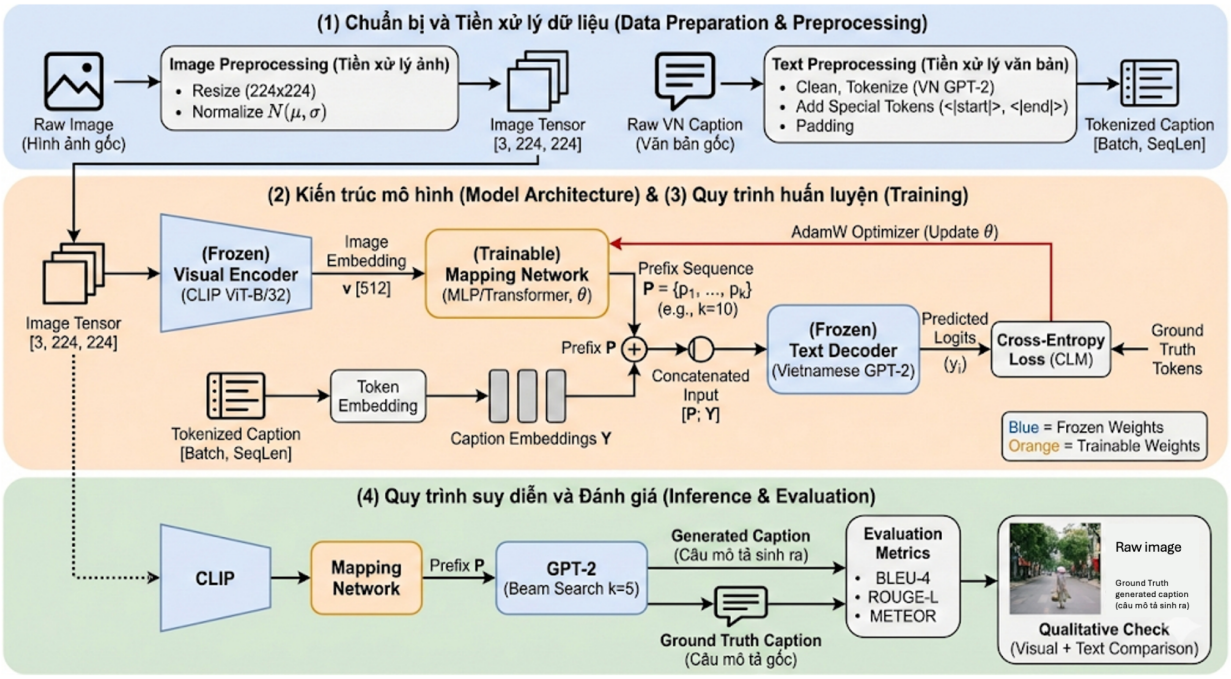
4. Phạm vi và Mục tiêu: Đề tài này chú trọng vào việc triển khai quy trình xử lý, chuyển đổi hình ảnh đầu vào thành câu mô tả đầy đủ bằng tiếng Việt. Thông qua việc training Mạng Ánh xạ với dữ liệu ảnh-caption đã Việt hóa, nhóm mong muốn khẳng định hiệu quả của việc áp dụng mô hình đa phương thức cho các ngôn ngữ hiếm dữ liệu. Thành công của đề tài không chỉ tháo gỡ vấn đề ghép nối kỹ thuật giữa các mô hình mà còn là bước đệm quan trọng để phát triển các ứng dụng AI phục vụ người Việt.

MỤC TIÊU

1. Nắm vững nền tảng lý thuyết: Phân tích chi tiết kiến trúc Transformer, cách thức CLIP trích xuất thông tin ảnh và nguyên lý sinh văn bản tự động của GPT-2.
2. Xây dựng và Huấn luyện mô hình: Cài đặt thuật toán ClipCap áp dụng cho dữ liệu tiếng Việt, thiết kế mạng Mapping Network (sử dụng MLP hoặc Transformer) để kết nối giữa CLIP và Vietnamese GPT-2, và thực hiện huấn luyện trên bộ dữ liệu ảnh-chú thích tiếng Việt (bộ dữ liệu được dịch từ MS-COCO hoặc KTVIC).
3. Thẩm định và Hiệu chỉnh: Tiến hành thực nghiệm trên tập dữ liệu kiểm tra, dùng các độ đo **BLEU, ROUGE, METEOR,...** đánh giá hiệu năng. Các yếu tố định tính (độ trôi chảy, tính đúng đắn của ngữ cảnh) được xem xét nhằm tinh chỉnh tham số để nâng cao hiệu quả.

NỘI DUNG VÀ PHƯƠNG PHÁP

Nghiên cứu được thực hiện theo quy trình tính toán khép kín gồm 4 giai đoạn chính: (1) Chuẩn bị và Tiền xử lý dữ liệu đa phương thức; (2) Thiết kế kiến trúc hệ thống ClipCap-Vi; (3) Chiến lược huấn luyện tối ưu hóa; (4) Quy trình suy diễn và đánh giá.



Hình 1: Pipeline của phương pháp

1. Chuẩn bị dữ liệu (Data Preparation)

Giai đoạn này tập trung vào việc thực hiện tiền xử lý và chuẩn hóa dữ liệu đầu vào nhằm đáp ứng các yêu cầu kỹ thuật của mô hình học sâu.

- **Lựa chọn dữ liệu:** Sử dụng bộ dữ liệu ảnh-chú thích (Image-Captioning Dataset) chứa các cặp (Hình ảnh, Văn bản tiếng Việt). Dữ liệu được tổng hợp từ MS-COCO (phiên bản dịch máy) và bộ dữ liệu KTVIC.
- **Tiền xử lý ảnh (Image Preprocessing):** Hình ảnh đầu vào ***I*** được chuyển đổi kích thước (resize) và chuẩn hóa (normalize) theo chuẩn của mô hình CLIP **(224x224x3)**. Đồng bộ hóa không gian dữ liệu ảnh với chuẩn ImageNet thông qua việc chuẩn hóa vector trung bình (mean) và độ lệch chuẩn (std).
- **Tiền xử lý văn bản (Text Preprocessing):** Các câu chú thích tiếng Việt được làm sạch (loại bỏ ký tự đặc biệt, chuẩn hóa mã unicode) và mã hóa (tokenize) bằng bộ Tokenizer chuyên dụng cho tiếng Việt (tương thích với mô hình **gpt2news**). Chuỗi văn bản được thêm các token đặc biệt (**<|startoftext|>**, **<|endoftext|>**) và đệm (padding) để đảm bảo độ dài cố định cho quá trình huấn luyện theo batch.

2. Kiến trúc mô hình (Model Architecture)

Hệ thống được thiết kế dựa trên mô hình "Prefix Tuning", đặc trưng hình ảnh được xem như tiền tố ngữ nghĩa cung cấp thông tin đầu vào cho mô hình ngôn ngữ. Kiến trúc gồm 3 thành

KẾT QUẢ MONG ĐỢI

1. Sản phẩm thực tế

Phát triển trọn bộ mã nguồn cùng giao diện Web Demo minh họa cho bài toán captioning tiếng Việt. Đồng thời, chia sẻ các file trọng số (checkpoint) của Mạng Ánh xạ đã được huấn luyện tốt nhất để có thể ghép nối hiệu quả với GPT-2.

2. Hiệu năng định lượng

Kỳ vọng rằng kiến trúc đề xuất sẽ ghi nhận kết quả thực nghiệm mang tính cạnh tranh cao, thậm chí tối ưu hơn so với các phương pháp nền tảng (baselines) hiện hành ở các thang đo tiêu chuẩn trên tập dữ liệu kiểm thử:

- ***Độ chính xác:*** *BLEU-4 > 25.0 và METEOR > 20.0.*
- ***Tốc độ:*** *Thời gian sinh mô tả trung bình 200ms/ảnh trên GPU phổ thông, đảm bảo khả năng ứng dụng thực tế.*

3. Giá trị khoa học

Chứng minh tính khả thi, sự hiệu quả của phương pháp "Prefix Tuning" (giữ nguyên CLIP/GPT-2, chỉ train mạng ánh xạ) khi áp dụng cho bài toán đa phương thức tiếng Việt. Bên cạnh đó là phần đánh giá sâu về các trường hợp mô hình dự đoán sai để định hướng cải thiện.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Ron Mokady, Amir Hertz, Amit H. Bermano: ClipCap: CLIP Prefix for Image Captioning. CoRR abs/2111.09734 (2021)

[2]. Anh-Cuong Pham, Van-Quang Nguyen, Thi-Hong Vuong, Quang-Thuy Ha: KTVIC: A Vietnamese Image Captioning Dataset on the Life Domain. CoRR abs/2401.08100 (2024)

[3]. Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, Rita Cucchiara: From Show to Tell: A Survey on Deep Learning-Based Image Captioning. IEEE Trans. Pattern Anal. Mach. Intell. 45(1): 539-559 (2023)