THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/Mfl2F8WsrbU?si=96dzaOfxnSA9taPN*
- Link slides (dạng .pdf đặt trên Github của nhóm): *https://github.com/VietNe/CS2205.CH183/blob/main/slides.pdf*
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*
- *Lớp Cao học, mỗi nhóm một thành viên*

- Họ và Tên: **Hoàng Đình Quang**
- MSSV: **230101018**

- Lớp: **CS2205.CH183**
- Tự đánh giá (điểm tổng kết môn): 9/10
- Số buổi vắng: 1
- Số câu hỏi QT cá nhân: 3
- Link Github: https://github.com/BlueTeamQQ1/CS2205.CH183

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

MÔ HÌNH THỊ GIÁC - NGÔN NGỮ LỚN CHO HỆ THỐNG TRẢ LỜI CÂU HỎI TRỰC QUAN TIẾNG VIỆT

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

VISION - LARGE LANGUAGE MODEL FOR VIETNAMESE VISUAL QUESTION ANSWERING (VQA)

TÓM TẮT *(Tối đa 400 từ)*

Sự phát triển của các mô hình Large Vision-Language Model (VLLM) đã mở ra những tiềm năng lớn trong lĩnh vực Visual Question Answering (VQA), đặc biệt là đối với các ngôn ngữ chưa được nghiên cứu sâu như tiếng Việt. Trong nghiên cứu này, chúng tôi đề xuất fine-tuning mô hình Qwen2VL, một mô hình VLLM hỗ trợ tiếng Việt, nhằm cải thiện khả năng hiểu và trả lời câu hỏi dựa trên hình ảnh trong ngữ cảnh tiếng Việt.

Quá trình fine-tuning sẽ được thực hiện trên một bộ dữ liệu VQA chứa các cặp hình ảnh - câu hỏi - câu trả lời tiếng Việt, kết hợp với kỹ thuật LoRA (Low-Rank Adaptation) để tối ưu hóa việc huấn luyện mô hình với tài nguyên tính toán hạn chế. Kỹ thuật này giúp giảm số lượng tham số cần cập nhật trong quá trình fine-tuning, đồng thời duy trì hiệu suất cao của mô hình gốc.

Nghiên cứu này không chỉ hướng tới nâng cao hiệu suất của mô hình trên tác vụ VQA mà còn đóng góp vào sự phát triển của các hệ thống AI hỗ trợ ngôn ngữ tiếng Việt. Kết quả mong đợi bao gồm việc đánh giá độ chính xác của mô hình trên các tập dữ liệu kiểm thử, so sánh với các phương pháp hiện có và phân tích hiệu quả của kỹ thuật fine-tuning LoRA đối với việc nâng cao khả năng suy luận

thị giác - ngôn ngữ. Bên cạnh đó, nghiên cứu cũng hướng tới khả năng ứng dụng thực tiễn trong các lĩnh vực như giáo dục, trợ lý ảo và hỗ trợ người khiếm thị.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Trong những năm gần đây, lĩnh vực Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) kết hợp với Xử lý thị giác máy tính (Computer Vision - CV) đã chứng kiến sự phát triển mạnh mẽ, đặc biệt trong các bài toán yêu cầu mô hình khai thác thông tin đa phương thức như Visual Question Answering (VQA). VQA đòi hỏi hệ thống có khả năng trích xuất đặc trưng từ hình ảnh, đồng thời phân tích và hiểu ngôn ngữ để trả lời câu hỏi phù hợp với nội dung thị giác.

Các giải pháp ban đầu cho VQA thường sử dụng mạng nơ-ron tích chập (CNN) để trích xuất đặc trưng hình ảnh, kết hợp với mạng nơ-ron hồi tiếp (RNN) hoặc mô hình attention để xử lý câu hỏi. Mặc dù đã đạt được một số thành công trên những bộ dữ liệu phổ biến, các mô hình này vẫn gặp khó khăn khi mở rộng sang ngôn ngữ tiếng Việt, một phần do thiếu dữ liệu chất lượng cao cũng như sự khác biệt về ngữ pháp và cú pháp.

Trong bối cảnh gần đây, hướng tiếp cận mới đang nổi lên với sự xuất hiện của các mô hình ngôn ngữ - thị giác quy mô lớn (Vision-Language Large Language Models - VLLM). Các mô hình này giúp hợp nhất quá trình xử lý dữ liệu dạng hình ảnh và ngôn ngữ vào một khối thống nhất, có thể nắm bắt được mối quan hệ liên phương thức phức tạp hơn. Qwen2VL là một ví dụ điển hình, khi đã được tiền huấn luyện trên khối lượng dữ liệu đa phương thức lớn, đồng thời hỗ trợ nhiều ngôn ngữ, bao gồm tiếng Việt. Nhờ đó, Qwen2VL tạo điều kiện thuận lợi cho việc phát triển hệ thống VQA tiếng Việt trên nền tảng đã có sẵn khả năng hiểu khái quát cả hình ảnh lẫn ngôn ngữ.

Để nâng cao hiệu suất và thích nghi với đặc trưng ngôn ngữ Việt, đề tài này đề xuất áp dụng kỹ thuật LoRA (Low-Rank Adaptation) cho quá trình tinh chỉnh Qwen2VL trên bộ dữ liệu VQA dành riêng cho tiếng Việt. Cách làm này giúp rút gọn thời gian huấn luyện mà vẫn khai thác tối đa khả năng biểu diễn của mô hình. Mục tiêu chính là xây dựng một hệ thống VQA có độ chính xác cao trong việc trả lời các câu hỏi liên quan đến hình ảnh, từ đó đáp ứng nhu cầu thực tế trong các lĩnh vực như giáo dục, thương mại, hay chăm sóc sức khỏe tại Việt Nam. Qua đó, nghiên cứu cũng nhằm thúc đẩy sự phát triển của các giải pháp AI đa phương thức, góp phần khắc phục hạn chế về ngôn ngữ và mở ra các ứng dụng đa dạng hơn cho cộng đồng.

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

Nhóm nghiên cứu đặt ra các mục tiêu sau đây:

- Xây dựng quy trình fine-tuning mô hình Qwen2VL trên dữ liệu VQA tiếng Việt bằng kỹ thuật LoRA, đảm bảo mô hình có thể tiếp thu tri thức mới mà không làm giảm hiệu năng trên các nhiệm vụ gốc
- Đánh giá hiệu quả của mô hình đã được fine-tuning trên các tiêu chí chính của tác vụ VQA (chính xác, độ tin cậy, tốc độ phản hồi), đặc biệt tập trung vào ngôn ngữ tiếng Việt để so sánh với các phương pháp hiện có.
- Xác định các yếu tố then chốt ảnh hưởng đến chất lượng của mô hình (ví dụ: kích thước dữ liệu huấn luyện, cấu trúc mô hình, phương pháp fine-tuning) và đề xuất giải pháp cải tiến nhằm nâng cao hiệu suất cho tác vụ VQA tiếng Việt.

NỘI DUNG VÀ PHƯƠNG PHÁP

1. **Khảo sát mô hình và kỹ thuật**
    - Tìm hiểu kiến trúc Qwen2VL và cách áp dụng **LoRA** (Low-Rank Adaptation).
    - Xem xét các công trình VQA trước đây, đặc biệt với tiếng Việt, để định hướng giải pháp.
2. **Xây dựng và chuẩn bị dữ liệu**
    - Lựa chọn/bổ sung bộ dữ liệu VQA có câu hỏi - câu trả lời bằng tiếng Việt, đảm bảo đa dạng về chủ đề và dạng câu hỏi.
    - Tiền xử lý dữ liệu
3. **Fine-tuning Qwen2VL bằng LoRA**
    - Hạ tầng huấn luyện trên đa GPU
        - Qwen2VL là VLLM có nhiều tham số, đòi hỏi hạ tầng mạnh (multi-GPU/cluster) để rút ngắn thời gian huấn luyện.
        - Dùng cơ chế **data parallel** hoặc **model parallel** (VD: PyTorch Distributed) để chia tải.
    - Giám sát quá trình huấn luyện: Theo dõi các chỉ số (training loss, validation loss, accuracy,...)
4. **Đánh giá mô hình**
    - Dùng các chỉ số đánh giá VQA (Accuracy, BLEU, METEOR...) trên tập kiểm thử.
    - So sánh với Qwen2VL gốc và các mô hình VQA khác (nếu có).
    - Phân tích trường hợp mô hình trả lời sai để rút kinh nghiệm và đề xuất cải tiến.

KẾT QUẢ MONG ĐỢI

- **Hiệu suất VQA tiếng Việt được cải thiện**: Mô hình Qwen2VL sau khi áp dụng LoRA được kỳ vọng đạt độ chính xác cao hơn trên bộ dữ liệu VQA tiếng Việt, nhờ khả năng hiểu ngữ cảnh và diễn giải linh hoạt.
- **Tối ưu tài nguyên**: Kỹ thuật LoRA giúp giảm chi phí và tài nguyên tính toán so với phương pháp fine-tune toàn bộ mô hình, đồng thời vẫn giữ hoặc nâng cao hiệu suất.
- **Khả năng ứng dụng thực tiễn cao**: Mô hình có tiềm năng tích hợp vào các hệ thống trợ lý ảo, mô tả ảnh, hỗ trợ người khuyết tật, kiểm duyệt nội dung, v.v., đáp ứng nhu cầu tương tác tiếng Việt.
- **Đóng góp cho cộng đồng**: Kết quả nghiên cứu (bộ trọng số đã fine-tune và quy trình thực nghiệm) sẽ hỗ trợ phát triển thêm các ứng dụng và nghiên cứu về mô hình thị giác–ngôn ngữ cho tiếng Việt.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., et al.: Qwen2 technical report. arXiv preprint arXiv:2407.10671 (2024)

[2] Nguyen, N.H., Vo, D.T., Nguyen, K.V., Nguyen, N.L.T.: Openvivqa: Task, dataset, and multimodal fusion models for visual question answering in Vietnamese. Information Fusion 100, 101868 (2023)

[3] Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26296–26306 (2024)

[4] Tran, C., Thanh, H.L.: Lavy: Vietnamese multimodal large language model. arXiv preprint arXiv:2404.07922 (2024)

[5] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS

(2023)
[6] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models (2023)
[7] Lin, B., Tang, Z., Ye, Y., Cui, J., Zhu, B., Jin, P., Zhang, J., Ning, M., Yuan, L.: Moe-llava: Mixture of experts for large vision-language models (2024)
[8] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. Available at: https://arxiv.org/abs/2308.12966 (2023)