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

Hoàng Đình Quang - 230101018

Tóm tắt

- Lớp: CS2205.CH183
- Link Github của nhóm: https://github.com/BlueTeamQQ1/CS2205.CH183
- Link YouTube video: https://youtu.be/Mfl2F8WsrbU?si=Sw90uaIAEY-S66bO
- Ảnh + Họ và Tên: Hoàng Đình Quang - 230101018

Giới thiệu

- Visual Question Answering (VQA) kết hợp xử lý ngôn ngữ và thị giác để trả lời câu hỏi từ hình ảnh.
- Tiếng Việt trong VQA gặp nhiều thách thức do hạn chế dữ liệu và mô hình chưa được tối ưu.
- Nghiên cứu này tinh chỉnh Qwen2VL bằng LoRA để cải thiện độ chính xác, mở rộng ứng dụng AI đa phương thức.

Mục tiêu

- 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.

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

Nội dung và Phương pháp

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,...)

Nội dung và Phương pháp

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 dự kiến

Cải thiện hiệu suất VQA tiếng Việt

- Mô hình Qwen2VL sau khi áp dụng LoRA được kỳ vọng đạt độ chính xác cao hơn.
- Nâng cao khả năng hiểu ngữ cảnh và diễn giải câu hỏi linh hoạt.

Tối ưu tài nguyên

- LoRA giúp giảm chi phí tính toán so với fine-tune toàn bộ mô hình.
- Đảm bảo hiệu suất cao mà vẫn tiết kiệm tài nguyên.

Đóng góp cho cộng đồng

- Cung cấp bộ trọng số fine-tune và quy trình thực nghiệm.
- Hỗ trợ nghiên cứu và phát triển AI thị giác – ngôn ngữ cho tiếng Việt.

Tài liệu tham khảo

[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)