TĂNG CƯỜNG HIỆU QUẢ CỦA MÔ HÌNH SINH MÔ TẢ CHO HÌNH ẢNH BẰNG CÁCH KẾT HỢP MÔ HÌNH NGÔN NGỮ LỚN VÀ MÔ HÌNH THỊ GIÁC

Trần Ngọc Thiện - 21521465

Tóm tắt

- Lớp: CS519.P11
- Link Github của nhóm: https://github.com/ngocthien705b/CS519.P11
- Link YouTube video:
- Các thành viên

Trần Ngọc Thiện

Giới thiệu

- Image captioning là việc tự động tạo ra mô tả văn bản cho hình ảnh. Các phương pháp truyền thống như đã đạt được những tiến bộ nhất định nhưng vẫn còn những hạn chế về khả năng nắm bắt ngữ cảnh, tạo mô tả phong phú và xử lý thông tin hình ảnh.
- Để khắc phục những hạn chế trên, đề xuất nghiên cứu này đưa ra một giải pháp mới là kết hợp mô hình ngôn ngữ lớn (LLM) và mô hình thị giác. Việc kết hợp này nhằm tận dụng khả năng sinh văn bản tự nhiên của LLM và khả năng phân tích hình ảnh của mô hình thị giác để tạo ra các mô tả chất lượng cao hơn.
- Nghiên cứu tập trung vào việc phát triển và tối ưu hóa mô hình kết hợp này, nhằm cải thiện chất lượng các chú thích hình ảnh và mở rộng ứng dụng của công nghệ này trong thực tế.

Mục tiêu

- Xây dựng mô hình kết hợp mô hình ngôn ngữ lớn (Large Language Models - LLMs) và mô hình thị giác (Vision Models)
- Huấn luyện và đánh giá mô hình được đề xuất trên các thang đo BLEU, METEOR, ROUGE
- So sánh mô hình được đề xuất với các phương pháp baseline

Nội dung và Phương pháp

- Khảo sát các mô hình thị giác và mô hình ngôn ngữ lớn phù hợp.
- Xây dựng mô hình:
  - Mô hình thị giác: Trích xuất các đặc trưng hình ảnh từ hình đầu vào.
  - Mô hình ngôn ngữ lớn: Sử dụng các đặc trưng hình ảnh này để tạo ra mô tả văn bản.
  - Kết hợp: Áp dụng các kỹ thuật attention để kết hợp thông tin từ cả hai mô hình, giúp mô hình tập trung vào các vùng quan trọng trong hình ảnh và tạo ra các mô tả phù hợp.
  - Áp dụng các kỹ thuật tinh chỉnh như prompt learning, prefix tuning
- Huấn luyện mô hình trên bộ dữ liệu KTVIC.
- Đánh giá mô hình trên các thang đo BLEU, METEOR, ROUGE.
- So sánh hiệu suất của mô hình với các phương pháp baseline.

Kết quả dự kiến

- Mô hình được tạo ra sẽ có khả năng sinh ra các chú thích hình ảnh một cách tự động với độ chính xác và độ mượt mà cao hơn so với các mô hình hiện có.
- Kết quả đánh giá mô hình sẽ cho thấy sự cải thiện so với các phương pháp baseline.
- Mô hình tích hợp sẽ cho ra các chú thích chính xác hơn, phong phú hơn và có khả năng truyền tải tốt hơn ngữ cảnh hình ảnh.

Tài liệu tham khảo

- [1]. Bucciarelli, D., Moratelli, N., Cornia, M., Baraldi, L., & Cucchiara, R. (2024). Personalizing Multimodal Large Language Models for Image Captioning: An Experimental Analysis. arXiv preprint arXiv:2412.03665.
- [2]. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).
- [3]. Xu, Kelvin. "Show, attend and tell: Neural image caption generation with visual attention." arXiv preprint arXiv:1502.03044 (2015).
- [4]. Pham, A. C., Nguyen, V. Q., Vuong, T. H., & Ha, Q. T. (2024). KTVIC: A Vietnamese Image Captioning Dataset on the Life Domain. arXiv preprint arXiv:2401.08100.