


# THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):  
<https://youtu.be/8waYSED-wjQ>
- Link slides (dạng .pdf đặt trên Github của nhóm):  
<https://github.com/ngocthien705b/CS519.P11/blob/main/Thi%E1%BB%87n%20Tr%E1%BA%A7n%20Ng%E1%BB%8Dc%20-%20CS519.P11.DeCuong.FinalReport.Template.Slide.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

<ul style="list-style-type: none"><li>● Họ và Tên: Trần Ngọc Thiện</li><li>● MSSV: 21521465</li></ul> 	<ul style="list-style-type: none"><li>● Lớp: CS519.P11</li><li>● Tự đánh giá (điểm tổng kết môn): 8/10</li><li>● Số buổi vắng: 0</li><li>● Số câu hỏi QT cá nhân: 0</li><li>● Số câu hỏi QT của cả nhóm: 0</li><li>● Link Github: <a href="https://github.com/ngocthien705b/CS519.P11">https://github.com/ngocthien705b/CS519.P11</a></li><li>● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none"><li>○ Lên ý tưởng</li><li>○ Viết đề cương</li><li>○ Làm video YouTube</li></ul></li></ul>
---	--

# ĐỀ CƯƠNG NGHIÊN CỨU

## TÊN ĐỀ TÀI (IN HOA)

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

## TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

ENHANCE THE PERFORMANCE OF VISUAL DESCRIPTION GENERATION MODELS BY COMBINATION OF LARGE LANGUAGE MODEL AND VISUAL MODEL

## TÓM TẮT (Tối đa 400 từ)

Trong thời đại công nghệ 4.0, việc phát triển các hệ thống tự động có khả năng hiểu và tạo ra ngôn ngữ tự nhiên từ hình ảnh đang trở thành một đề tài nghiên cứu quan trọng trong lĩnh vực trí tuệ nhân tạo. Nghiên cứu này tập trung vào việc ứng dụng mô hình ngôn ngữ lớn (Large Language Models - LLMs) kết hợp với các mô hình thị giác (Vision Models) để giải quyết bài toán image captioning, một nhiệm vụ yêu cầu hệ thống phải sinh ra các chú thích miêu tả chính xác và phong phú cho các hình ảnh.

## GIỚI THIỆU (Tối đa 1 trang A4)

Trong thời đại công nghệ 4.0, việc phát triển các hệ thống tự động có khả năng hiểu và tạo ra ngôn ngữ tự nhiên từ hình ảnh đang trở thành một đề tài nghiên cứu quan trọng trong lĩnh vực trí tuệ nhân tạo. Bài toán image captioning yêu cầu một hệ thống không chỉ nhận diện các đối tượng và hoạt động trong hình ảnh mà còn phải hiểu ngữ cảnh để sinh ra các chú thích chính xác và phong phú. Đầu vào của bài toán là một hình ảnh (input), trong khi đầu ra là một câu mô tả (output) về nội dung hình ảnh.

Trong nghiên cứu hiện tại, nhiều phương pháp đã được phát triển để giải quyết bài toán image captioning, trong đó hai phương pháp phổ biến nhất là mô hình encoder-decoder và transformer-based.

1. Mô hình Encoder-Decoder: Phương pháp này sử dụng một mạng nơ-ron tích chập (CNN) như phần encoder để trích xuất đặc điểm từ hình ảnh và một mạng LSTM (Long Short-Term Memory) như phần decoder để sinh ra câu mô tả. Mặc dù mô hình này đã đưa ra những kết quả khả quan, nhưng nó thường gặp khó khăn trong việc duy trì ngữ cảnh và sinh ra các câu mô tả phong phú. Hạn

chế chính của phương pháp này là khả năng tiếp nhận thông tin liên tục từ một chuỗi đầu vào dài, dẫn đến việc thiếu tính chính xác trong các mô tả.

2. Mô hình Transformer-based: Nền tảng của mô hình này là kiến trúc transformer, vốn nổi bật với khả năng xử lý thông tin theo cách tự chú ý (self-attention). Mô hình transformer đã cải thiện đáng kể hiệu suất trong nhiều nhiệm vụ NLP, nhưng trong image captioning, nó vẫn gặp khó khăn trong việc khai thác đầy đủ các đặc điểm hình ảnh do không gian thể hiện hình ảnh và văn bản khác nhau.

Để khắc phục những hạn chế trên, nghiên cứu này đề xuất một phương pháp kết hợp mô hình ngôn ngữ lớn (Large Language Models - LLMs) và mô hình thị giác (Vision Models) nhằm tăng cường hiệu quả sinh mô tả cho hình ảnh. Mô hình ngôn ngữ lớn sẽ cung cấp khả năng sinh ngôn ngữ tự nhiên phong phú và sâu sắc, trong khi mô hình thị giác sẽ giúp cải thiện khả năng phân tích hình ảnh. Nghiên cứu sẽ tập trung vào việc phát triển và tối ưu hóa mô hình, từ đó cải thiện chất lượng các chú thích hình ảnh trong các ứng dụng thực tế.

#### **MỤC TIÊU** *(Viết trong vòng 3 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 MONG ĐỢI

- 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 (*Định dạng DBLP*)

- [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.