



# THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút):

<https://youtu.be/vhXZ0JHxn7w>

- Link slides (dạng .pdf đặt trên Github):

<https://github.com/4k4m/CS519.O21.KHTN/blob/main/VQATiengViet.pdf>

<ul style="list-style-type: none"><li>• Họ và Tên: Đặng Lê Thành Tâm</li><li>• MSSV: 22521290</li></ul> 	<ul style="list-style-type: none"><li>• Lớp: CS519.O21.KHTN</li><li>• Tự đánh giá (điểm tổng kết môn): 9/10</li><li>• Số buổi vắng: 0</li><li>• Số câu hỏi QT cá nhân: 11</li><li>• Link Github: <a href="https://github.com/tamtam24/CS519.O21.KHTN">https://github.com/tamtam24/CS519.O21.KHTN</a></li><li>• Mô tả công việc:<ul style="list-style-type: none"><li>○ Tìm bài báo và đề tài cho đề cương</li><li>○ Viết đề cương</li></ul></li></ul>
<ul style="list-style-type: none"><li>• Họ và Tên: Lê Quang Thiên Phúc</li><li>• MSSV: 22521120</li></ul> 	<ul style="list-style-type: none"><li>• Lớp: CS519.O21.KHTN</li><li>• Tự đánh giá (điểm tổng kết môn): 8.5/10</li><li>• Số buổi vắng: 1</li><li>• Số câu hỏi QT cá nhân: 11</li><li>• Link Github: <a href="https://github.com/4k4m/CS519.O21.KHTN">https://github.com/4k4m/CS519.O21.KHTN</a></li><li>• Mô tả công việc:<ul style="list-style-type: none"><li>○ Làm slide và poster</li><li>○ Quay video báo cáo</li></ul></li></ul>

# ĐỀ CƯƠNG NGHIÊN CỨU

## TÊN ĐỀ TÀI (IN HOA)

NGHIÊN CỨU THỰC NGHIỆM VỀ CÁC MÔ HÌNH NGÔN NGỮ TIỀN HUẤN  
LUYỆN CHO VIỆC HỎI ĐÁP TỰ ĐỘNG TRÊN HÌNH ẢNH SỬ DỤNG CÁC  
PHƯƠNG PHÁP ĐIỀU CHỈNH CHI TIẾT DỰA TRÊN HƯỚNG DẪN

## TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

EMPIRICAL STUDY OF VIETNAMESE PRETRAINED LANGUAGE MODELS  
FOR VISUAL QUESTION ANSWERING USING PROMPT-BASED FINE  
TUNING METHODS

## TÓM TẮT (Tối đa 400 từ)

Visual Question Answering (VQA) là bài toán đòi hỏi máy tính có khả năng hiểu và trả lời câu hỏi dựa trên hình ảnh. Tuy đã có những tiến bộ đáng kể trong VQA cho tiếng Anh, việc thiếu một hệ thống thước đo chuẩn tạo ra một khoảng trống rất lớn cho bài toán VQA trong tiếng Việt. Khoảng trống này tạo ra một rào cản đáng kể đối với các nhà nghiên cứu mong muốn phát triển các hệ thống VQA phù hợp với các đặc điểm đặc biệt của tiếng Việt. Trong nghiên cứu này, mục tiêu của chúng tôi là đánh giá hiệu suất của các mô hình ngôn ngữ tiếng Việt trong tác vụ VQA bằng các phương pháp tinh chỉnh dựa trên prompt [1] (prompt-based fine-tuning). Chúng tôi phân tích và so sánh hiệu quả của các phương pháp này trong việc cải thiện khả năng trả lời câu hỏi của các hệ thống VQA tiếng Việt, đồng thời tìm hiểu ảnh hưởng của các yếu tố như kích thước mô hình, loại dữ liệu huấn luyện và chiến lược huấn luyện. Chúng tôi hy vọng rằng nghiên cứu này sẽ cung cấp cái nhìn toàn diện về hiệu suất của các mô hình và phương pháp tinh chỉnh, từ đó hỗ trợ phát triển các hệ thống VQA tiếng Việt hiệu quả và đề xuất các hướng nghiên cứu tiếp theo.

## GIỚI THIỆU (Tối đa 1 trang A4)

Trong những năm gần đây, sự phát triển vượt bậc về công nghệ và khoa học đã dẫn

đến sự thúc đẩy ở nhiều lĩnh vực, đặc biệt là trí tuệ nhân tạo (AI). Trong đó, các nhà nghiên cứu trong cả 2 lĩnh vực CV (thị giác máy tính) và NLP (xử lý ngôn ngữ tự nhiên) đã tập trung phát triển mô hình đa phương thức kết hợp khả năng hiểu và trả lời các tác vụ liên quan trên cả phương diện hình ảnh và ngôn ngữ. Điều đó đã thúc đẩy sự phát triển nghiên cứu về hỏi đáp trả lời tự động trên hình ảnh (VQA). Trong tiếng Anh, các nghiên cứu về VQA những năm gần đây có số lượng tăng một cách vượt bậc nhờ nguồn tài nguyên về tiếng Anh rất dồi dào; đồng thời phần lớn các LLM (mô hình ngôn ngữ lớn) thường hoạt động tốt trên tiếng Anh. Tuy nhiên, các nghiên cứu về VQA trên các ngôn ngữ có nguồn tài nguyên hạn hẹp như tiếng Việt vẫn còn rất ít. Mặc dù đã có 1 số công trình nghiên cứu về VQA tiếng Việt, nhưng vẫn còn nhiều hạn chế. Nghiên cứu đầu tiên về VQA ở Việt Nam bắt đầu bằng việc xuất bản bộ dữ liệu ViVQA[2], sau đó dần dần các bộ dữ liệu mới được xây dựng và công bố như OpenViVQA[3], UIT-EVJVQA[4], ViCLEVR[5] và các mô hình ngôn ngữ tiếng Việt như PhoBERT[6], ViT5[7], BARTPho[8]. Nhưng hiện nay, vẫn chưa có một bộ tiêu chuẩn đánh giá đủ khái quát và mạnh mẽ để đánh giá chính xác hiệu năng của các mô hình cho bài toán VQA tiếng Việt. Điều này dẫn đến việc các mô hình VQA tiếng Việt hiện tại chưa đạt được hiệu quả cao và khả năng khái quát tốt. Do đó để tìm hiểu khả năng của các mô hình tiền huấn luyện trên các tác vụ VQA trong ngôn ngữ tiếng Việt, chúng tôi đã thực hiện đề tài này trên cả 4 bộ dữ liệu VQA trên tiếng Việt khác nhau với các đặc tính riêng biệt nhằm đánh giá khách quan nhất.

## MỤC TIÊU

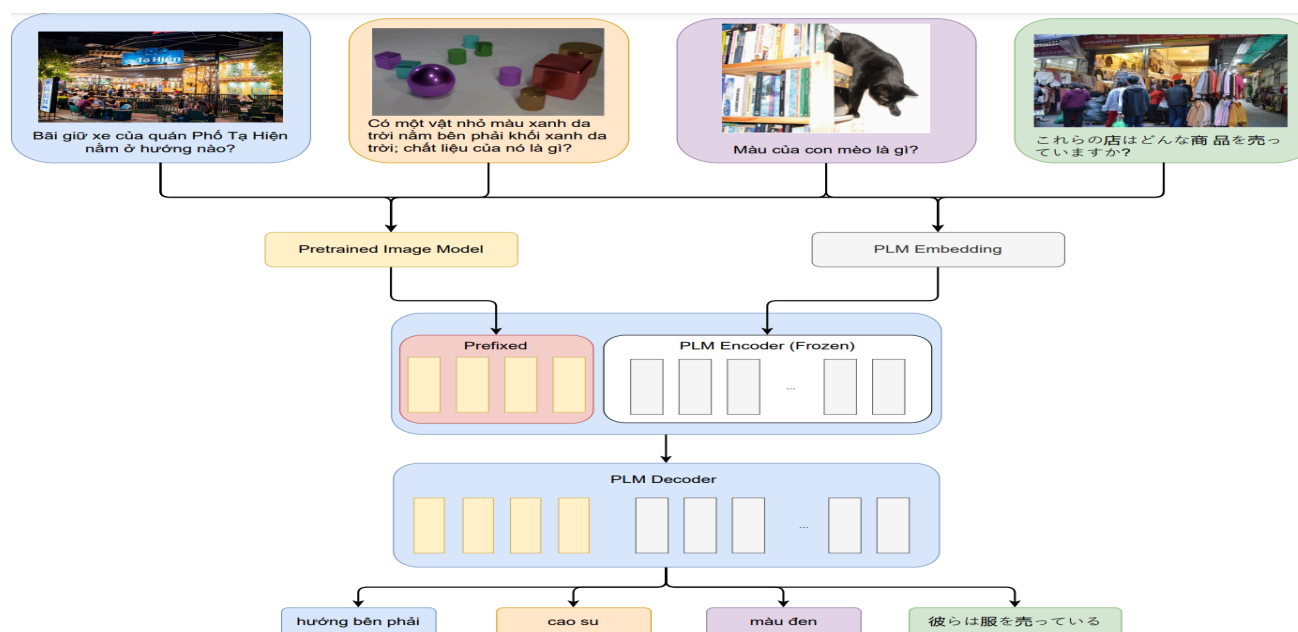
1. Đánh giá hiệu suất của các mô hình ngôn ngữ tiếng Việt trong tác vụ Visual Question Answering (VQA) thông qua việc sử dụng các phương pháp tinh chỉnh dựa trên prompt.
2. Phân tích ảnh hưởng của các yếu tố khác nhau như kích thước mô hình, loại dữ liệu huấn luyện và chiến lược huấn luyện đối với hiệu suất của mô hình VQA tiếng Việt.
3. Đặt ra các hướng nghiên cứu tiếp theo để nâng cao hiệu suất của mô hình VQA

tiếng Việt, nhằm phục vụ cho các ứng dụng thực tế trong tương lai.

## NỘI DUNG VÀ PHƯƠNG PHÁP

### 1. Nội dung

Hệ thống Trả lời Câu hỏi Hình ảnh (VQA) của chúng tôi được thiết kế theo mô hình encoder-decoder, với trọng tâm là cải thiện khả năng hiểu và trả lời câu hỏi dựa trên hình ảnh đầu vào. Chúng tôi sử dụng một mô hình trích xuất đặc trưng hình ảnh đã được huấn luyện trước (ví dụ: ResNet hoặc VGG) để chuyển đổi hình ảnh thành các vector đặc trưng. Sau đó, chúng tôi áp dụng phương pháp FROZEN[9], với mô hình ngôn ngữ (Language Model - LM) được đóng băng trọng số, để xử lý câu hỏi và tạo ra vector đặc trưng ngữ nghĩa. Cả hai vector đặc trưng từ hình ảnh và câu hỏi, được kết hợp thông qua một prompt được thiết kế để tối ưu hóa khả năng kết hợp thông tin. Cuối cùng, hệ thống sinh ra câu trả lời dựa trên một bộ giải mã của mô hình ngôn ngữ tiền huấn luyện.



### 2. Phương pháp nghiên cứu

- Để đánh giá hiệu suất của hệ thống, chúng tôi tiến hành thử nghiệm trên bốn mô hình ngôn ngữ đã được huấn luyện trước cho tiếng Việt: mBERT, PhoBERT, ViT5, và BARTPho. Mỗi mô hình có những đặc điểm riêng biệt trên các tác vụ xử lý ngôn ngữ tự nhiên (NLP).

- **mBERT[10] (Multilingual BERT)**: Được thiết kế để tiền huấn luyện các biểu diễn sâu hai chiều từ văn bản không gán nhãn, mBERT cải thiện hiệu suất trong nhiều tác vụ NLP đặc thù trên nhiều ngôn ngữ trong đó có tiếng Việt
- **PhoBERT[6]**: Phát triển bởi VinAI Research, PhoBERT là mô hình ngôn ngữ quy mô lớn cho tiếng Việt. Nó vượt trội trong các tác vụ như gán nhãn từ loại, phân tích phụ thuộc, nhận diện thực thể, và suy luận ngôn ngữ tự nhiên.
- **ViT5[7]**: Mô hình mã hóa-giải mã dựa trên Transformer, áp dụng phương pháp tự học tự chú ý kiểu T5, huấn luyện trên một tập dữ liệu văn bản tiếng Việt lớn.
- **BARTPho[8]**: Mô hình kết hợp giữa kiến trúc "large" và phương pháp huấn luyện trước của bộ mã hóa-giải mã chuỗi-quà-chuỗi BART, thích hợp cho các nhiệm vụ sinh NLP như tóm tắt văn bản hoặc sinh văn bản
- Trong quá trình huấn luyện, chỉ các tham số của bộ mã hóa hình ảnh được cập nhật. Các tham số của mô hình ngôn ngữ được giữ nguyên (đóng băng), điều này giúp duy trì khả năng của mô hình ngôn ngữ hiện có và tập trung vào việc tối ưu hóa bộ mã hóa hình ảnh. Chúng tôi sử dụng dữ liệu cặp hình ảnh-câu hỏi để huấn luyện, áp dụng lan truyền ngược và tối ưu hóa thông qua thuật toán SGD.
- Trong quá trình dự đoán, mô hình ngôn ngữ tiền huấn luyện được điều kiện hóa trên một gợi ý văn bản hoặc tiền tố tùy ý, sau đó tạo ra chuỗi các từ tiếp theo một cách tự hồi quy. Hình ảnh đầu vào được đưa vào gợi ý bằng cách đặt embedding của hình ảnh cạnh embedding của văn bản, cho phép mô hình xử lý và sinh ra câu trả lời dựa trên cả thông tin từ hình ảnh và câu hỏi.
- Để đánh giá hiệu suất của hệ thống, chúng tôi sử dụng các tiêu chí đánh giá BLEU, BERT-Score và CIDEr:
- Kết quả đánh giá trên các bộ dữ liệu ViVQA, EVJVQA, OpenViVQA, và ViCLEVR cho phép chúng tôi so sánh và đánh giá hiệu suất của các mô hình ngôn ngữ khác nhau đối với tiếng Việt, từ đó rút ra những kết luận về hiệu suất của hệ thống trong các tác vụ VQA đa dạng.

## KẾT QUẢ MONG ĐỢI

- Trên tất cả các tập dataset:
  - BERT\_Score đạt trên 70
  - BLEU đạt trên 85
  - CIDEr đạt trên 90

## TÀI LIỆU THAM KHẢO (Định dạng DBLP)

- [1] Timo Schick and Hinrich Schutze. ““True Few-Shot Learning with Prompts—A Real-World Perspective””. In: (2020).
- [2] An Tran-Hoai Le Kiet Van Nguyen Khanh Quoc Tran An Trong Nguyen. “ViVQA: Vietnamese Visual Question Answering”. In: PACLIC. 2021.
- [3] Kiet Van Nguyen Ngan Luu-Thuy Nguyen Nghia Hieu Nguyen Duong T. D. Vo. “OpenViVQA: Task, Dataset, and Multimodal Fusion Models for Visual Question Answering in Vietnamese”. In: arXiv:2305.04183v1 (2023).
- [4] Duong T. D. Vo Khanh Quoc Tran Kiet Van Nguyen Ngan Luu-Thuy Nguyen Nghia Hieu Nguyen. “VLSP2022-EVJVQACHallenge:Multilingual Visual Question Answering”. In: arXiv:2302.11752v4 (2023).
- [5] Kiet Van Nguyen Ngan Luu Thuy Nguyen Khiem Vinh Tran Hao Phu Phanc. “ViCLEVR: A Visual Reasoning Dataset and Hybrid Multimodal Fusion Model for Visual Question Answering in Vietnamese”. In: arXiv:2310.18046v1 (2023).
- [6] Dat Quoc Nguyen and Anh Tuan Nguyen. “PhoBERT: Pre-trained language models for Viet nameese”. In: In Findings of the Association for Computational Linguistics. EMNLP.
- [7] Hieu Tran<sup>1</sup> Hieu Nguyen<sup>1-2</sup> Trieu H. Trinh Long Phan<sup>1 2</sup>. ““ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation ””. In: arXiv:2205.06457v2
- [8] Dat Quoc Nguyen Nguyen Luong Tran Duong Minh Le. ““BARTpho: Pre-trained Sequence-to Sequence Models for Vietnamese ””. In: arXiv:2109.09701v3 (2022).
- [9] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. “Multimodal Few-Shot Learning with Frozen Language Models”. In: CoRR abs/2106.13884 (2021). arXiv: 2106.13884. url: <https://arxiv.org/abs/2106.13884>.
- [10] Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: arXiv:1810.04805v2

