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

Đặng Lê Thành Tâm - 22521290
Lê Quang Thiên Phúc - 22521120

Tóm tắt

- Link Github của nhóm: https://github.com/4k4m/CS519.O21.KHTN
- Link YouTube video:
- Ảnh + Họ và Tên của các thành viên

Đặng Lê Thành Tâm

Lê Quang Thiên Phúc

Giới thiệu

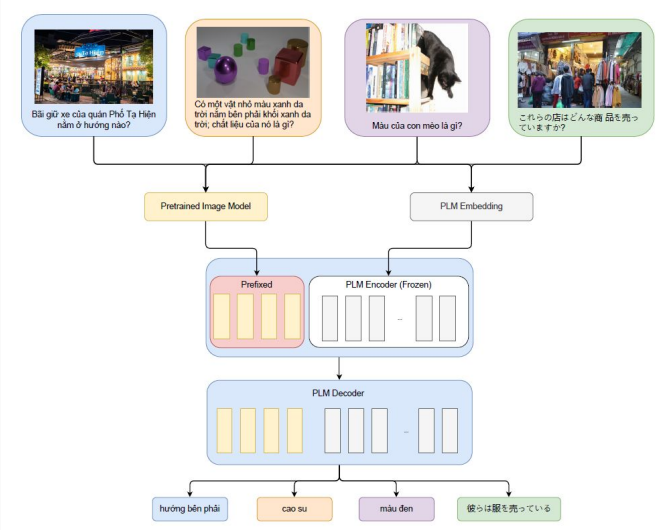
- Sự phát triển của 2 lĩnh vực CV và NLP đã thúc đẩy việc phát triển các mô hình hiểu và trả lời tự động trên hình ảnh (VQA)
- Những năm gần đây đã chứng kiến sự phát triển vượt bậc của bài toán VQA trên tiếng anh với một loạt các mô hình ngôn ngữ lớn, dataset, benchmark
- Bài toán VQA trên tiếng Việt gần đây đã chứng kiến sự phát triển đầu tiên với sự ra đời của các bộ dataset cho bài toán VQA và các mô hình ngôn ngữ tiếng Việt (Language Model)
- Tuy nhiên, vẫn chưa có một bộ tiêu chuẩn chung để đánh giá chính xác hiệu năng của các mô hình cho bài toán VQA tiếng Việt và đó cũng là lí do thúc đẩy chúng tôi thực hiện dự án này

Mục tiêu

- 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.
- 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.
- 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

- Nội dung 1: Phương pháp thực hiện
  - Đầu tiên, trích xuất vector đặc trưng của ảnh
  - Sau đó, đóng băng trọng số của language model để tạo vector đặc trưng ngữ nghĩa
  - Tiếp theo, thiết kế prompt có tham số điều chỉnh được để kết hợp thông tin từ cả hai vector, sau đó tinh chỉnh trên dữ liệu VQA
  - Kết hợp các thành phần trên để tạo câu trả lời và đánh giá dựa trên độ đo

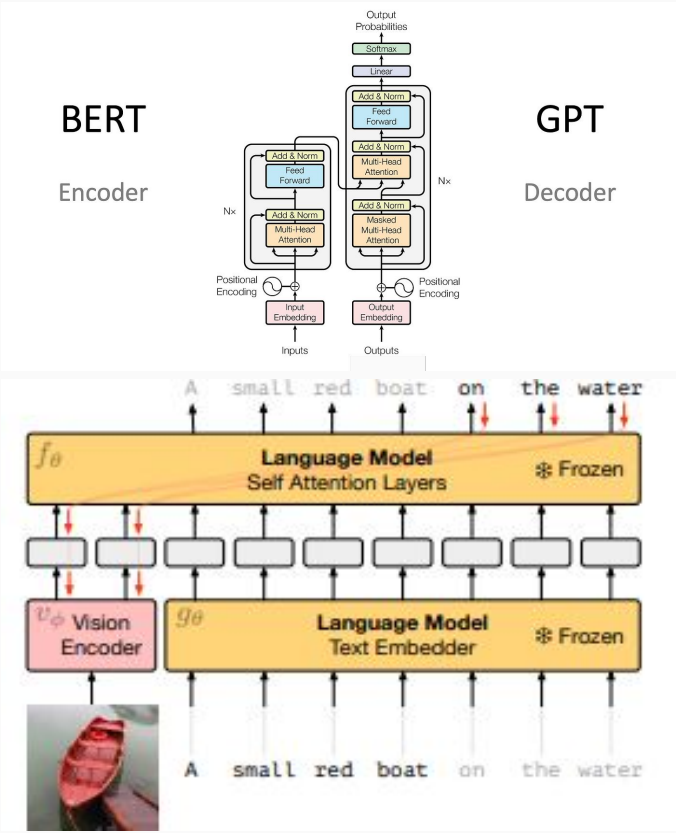


Nội dung và Phương pháp

- Nội dung 2: Các mô hình ngôn ngữ và các tập dataset được sử dụng để đánh giá
  - Mô hình
    - mBERT: mô hình đa ngôn ngữ có khả năng xử lý văn bản bằng nhiều ngôn ngữ trong đó có tiếng Việt
    - PhoBERT: mô hình tiền huấn luyện trên tiếng Việt đầu tiên được công bố bởi VinAI Research
    - ViT5: dựa trên kiến trúc mô hình T5 để huấn luyện trên tiếng Việt
    - BARTPho: mô hình tiền huấn luyện sequence-to-sequence trên tiếng Việt do VinAi Research công bố
  - Các bộ dataset được sử dụng do đội ngũ sinh viên, giảng viên UIT xây dựng cho bài toán VQA: OpenViVQA, ViVQA, EVJVQA, VICLEVR.

Nội dung và Phương pháp

- Nội dung 3: Kiến trúc đề xuất
  - Phương pháp tiếp cận encoder-decoder transformer
  - Triển khai phương pháp prompt base finetuning với xương sống là kỹ thuật FROZEN
  - Sử dụng mô hình Res-net để mã hóa ảnh
  - Sử dụng các mô hình tiền huấn luyện để biểu diễn câu hỏi thành vectơ
  - Cập nhật tham số qua các cặp hình ảnh-chú thích



Nội dung và Phương pháp

- Nội dung 4: Các thang đo được sử dụng:
  - BLEU

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{(ngram \in C)} Count_{clip}(ngram)}{\sum_{C' \in \{Candidates\}} \sum_{(ngram' \in C')} Count(ngram')}$$

$$logBLEU = \min\left(1 - \frac{r}{c}, 0\right) + \sum_{n=1}^{N} w_n log p_n$$

  - BERT-Score

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^T \hat{\mathbf{x}}_j \qquad P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^T \hat{\mathbf{x}}_j \qquad F_{BERT} = 2\frac{P_{BERT}.R_{BERT}}{P_{BERT} + R_{BERT}}$$

  - CIDEr

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})} \log\left(\frac{|I|}{\sum_{I_p \in I} \min\left(1, \sum_q h_k(s_{pq})\right)}\right)$$

$$CIDEr_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{\mathbf{g}^n(c_i).\mathbf{g}^n(s_{ij})}{\|\mathbf{g}^n(c_i)\| \|\mathbf{g}^n(s_{ij})\|}$$

$$CIDEr(c_i, S_i) = \sum_{n=1}^{N} w_n CIDEr_n(c_i, S_i)$$

Kết quả dự kiến

- Trên tất cả các tập dataset:
  - BERT_Score đạt trên 70
  - BLEU đạt trên 85
  - CIDEr đạt trên 90

Tài liệu tham khảo

[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 namese". In: In Findings of the Association for Computational Linguistics. EMNLP.
[7] Hieu Tran1 Hieu Nguyen1-2 Trieu H. Trinh Long Phan1 2. ""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 (2018).