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[1] **Lê Quang Thiên Phúc**[1]

[1] Trường ĐH Công nghệ thông tin,
ĐHQG TP. HCM

What ?

Nghiên cứu tìm hiểu các giải pháp hiện có để giải quyết bài toán Visual question answering (VQA) tiếng Việt, cụ thể:

- Đánh giá các mô hình ngôn ngữ tiếng Việt phổ biến
- Khảo sát và so sánh các phương pháp tinh chỉnh dựa trên prompt
- Phân tích ảnh hưởng của các yếu tố khác nhau đến hiệu quả của mô hình
- Đề xuất hướng nghiên cứu tiếp theo để nâng cao hiệu quả các mô hình

Why ?

- Trong những năm gần đây, 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)
- Tuy nhiên, các nghiên cứu về VQA trên các ngôn ngữ có nguồn tài nguyên hạn hẹp như tiếng Việt vẫn còn rất ít

Overview

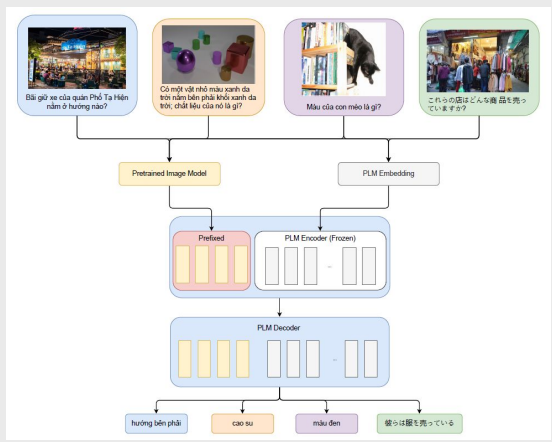
Xây dựng kiến trúc hệ thống với từng mô hình ngôn ngữ → Chạy trên các dataset khác nhau → Đánh giá các kết quả

Description

1. Xây dựng kiến trúc hệ thống với từng mô hình ngôn ngữ tiếng Việt

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

- Các mô hình tiếng Việt được sử dụng:
  - 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



Hình 1. Kiến trúc đề xuất

2. Chạy trên các dataset khác nhau

- Hệ thống được đánh giá trên các dataset khác nhau được xây dựng đội ngũ sinh viên, giảng viên UIT:
  - OpenViVQA
  - ViVQA
  - EVJVQA
  - VICLEVR

3. Đánh giá các kết quả

- Các độ đo được sử dụng:
  - BLEU: ý tưởng chính là so sánh các n-gram giữa câu dự đoán và câu tham chiếu

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

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

  - BERT-Score: dựa trên vector ngữ cảnh hóa và cosine similarity giữa các token, gồm recall, precision và F1 score

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

$$F_{BERT} = 2\frac{P_{BERT}.R_{BERT}}{P_{BERT}+R_{BERT}}$$

  - CIDER: dựa trên sự đồng thuận về mặt ý nghĩa và ngữ cảnh

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

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

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

Đặng Lê Thành Tâm – Trường Đại học Công nghệ thông tin ĐHQG TP.HCM
Email: 22521290@gm.uit.edu.vn