



VICLIPCAP: TINH CHỈNH TIỀN TỐ CLIP CHO MÔ HÌNH GPT-2 TRONG BÀI TOÁN SINH MÔ TẢ ẢNH TIẾNG VIỆT

Nguyễn Hà Anh Vũ - 250101077

Tóm tắt

Lớp: CS2205.CH201



Link Github: <https://github.com/vunha32/CS2205.CH201-Image-Captioning>

Link YouTube video: <https://youtu.be/KpeiggY7ce4>



Nguyễn Hà Anh Vũ -250101077



THÁCH THÚC:

Khoảng cách ngôn ngữ:

- Thế giới (Tiếng Anh): Đa dạng giải pháp SOTA.
- Việt Nam: Ngôn ngữ ít tài nguyên (Low-resource).
- Thực trạng: Thiếu dữ liệu & Giải pháp tối ưu hóa.



Rào cản tài nguyên

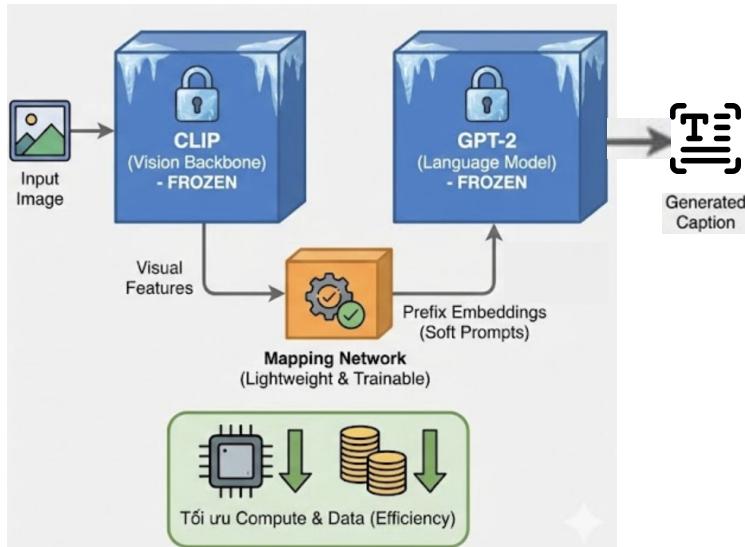
- Huấn luyện từ đầu (Scratch): Tốn kém thời gian & dữ liệu.
- Phần cứng: Yêu cầu GPU đắt đỏ.
- Nhu cầu: Cần phương pháp nhẹ (Lightweight).



→ "Làm thế nào để đạt HIỆU SUẤT CAO cho tiếng Việt với CHI PHÍ THẤP NHẤT?"



GIẢI PHÁP ĐỀ XUẤT



Công nghệ lõi: **Prefix Tuning** (Tinh chỉnh tiền tố).
Chiến lược: "Đứng trên vai người khổng lồ".

→ Tận dụng **CLIP** (Thị giác) & **GPT-2** (Ngôn ngữ).

Cơ chế hoạt động:

- **Đóng băng (Frozen):** Toàn bộ Backbone (Không tốn chi phí train lại).
- **Chỉ huấn luyện:** Mạng ánh xạ (Mapping Network) siêu nhẹ.

Hiệu quả: Tối ưu tài nguyên tính toán & Dữ liệu.

Mục tiêu



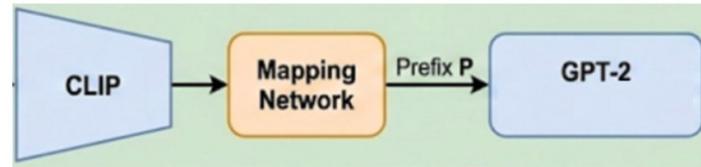
1. Cơ sở lý thuyết:

- Nghiên cứu kiến trúc Transformer.
- Phân tích cơ chế CLIP (Thị giác) & GPT-2 (Ngôn ngữ).



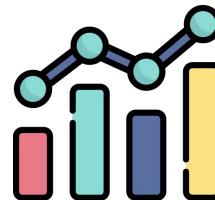
2. Xây dựng & Huấn luyện:

- Thiết kế **Mapping Network** (MLP/Transformer) kết nối CLIP và VN GPT-2.
- Huấn luyện mô hình ClipCap trên dữ liệu tiếng Việt (**KTVIC**).



3. Đánh giá & Tối ưu hóa:

- **Định lượng:** Các chỉ số BLEU, METEOR, ROUGE.
- **Định tính:** Độ tự nhiên & Chính xác ngữ nghĩa.



Kết quả



Metric	CNN+LSTM	ViCLIPCap
BLEU-4	0.2572	0.3431
ROUGE-L	0.4895	0.5204
CIDEr	0.6282	0.8127
METEOR	0.2995	0.3194
SPICE	0.0782	0.0829

Kết quả: Vượt trội ở **mọi chỉ số** trên tập dataset KTVIC

Kết luận: Chất lượng cao với chi phí huấn luyện tối thiểu (Prefix Tuning).



```
{"caption": "có hai tô phở cùng một đĩa quẩy xuất hiện ở trên bàn",  
"segment_caption": "có hai tô phở cùng một đĩa quẩy xuất hiện ở trên bàn"},  
  
{"caption": "có một người đang cầm trên tay một cái thìa",  
"segment_caption": "có một người đang cầm trên tay một cái thìa"},  
  
{"caption": "có một cái muỗng xuất hiện ở trên tay của một người",  
"segment_caption": "có một cái muỗng xuất hiện ở trên tay của một người"},  
  
 {"caption": "có một đĩa quẩy được đặt ở bên cạnh hai bát phở",  
"segment_caption": "có một đĩa quẩy được đặt ở bên cạnh hai bát phở"},  
  
 {"caption": "có hai bát phở được bày ra ở trên bàn",  
"segment_caption": "có hai bát phở được bày ra ở trên bàn"},  
  
"image_id": 10954
```

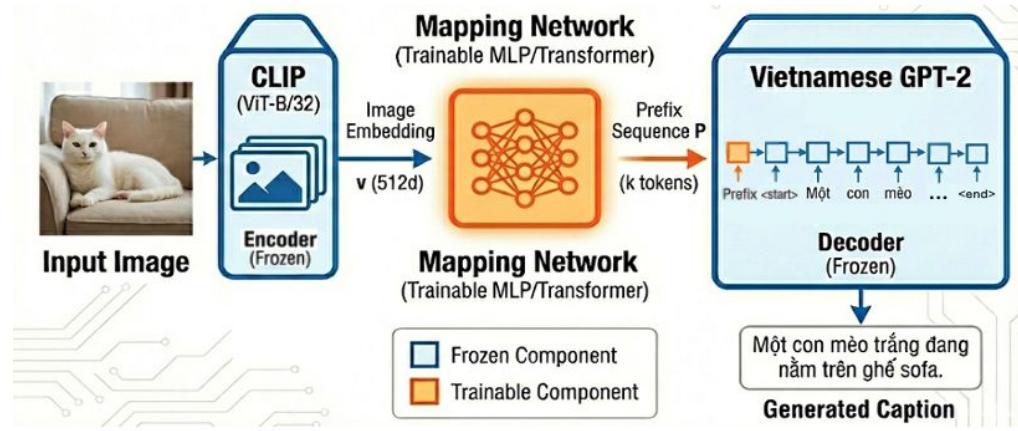
Its five annotated captions

Bộ dữ liệu: KTVIC

- **Lựa chọn:** Ưu tiên **Life Domain** (Đời sống) hơn Thể thao (UIT-ViC).
- **Mục tiêu:** Phản ánh hoạt động đa dạng hàng ngày của người Việt.
- **Quy mô (Scale):**
 - **4,327** Hình ảnh.
 - **21,635** Caption (~5 câu/ảnh).
- **Ý nghĩa:** Giải quyết thách thức thiếu dữ liệu (Low-resource).

Nội dung và Phương pháp

KIẾN TRÚC TỔNG QUAN CỦA HỆ THỐNG ViClipCap



Encoder: CLIP ViT-B/32 (Frozen).

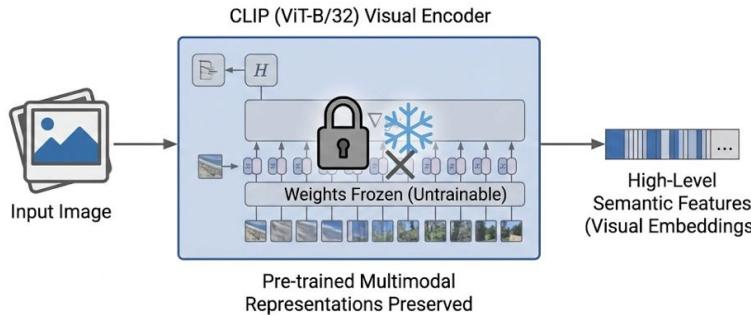
Bridge: Mapping Network (Trainable).

Decoder: VN GPT-2 (Frozen).

Luồng: $\text{Ảnh} \xrightarrow{\text{CLIP}} v \xrightarrow{\text{Mapping Net}} P \xrightarrow{\text{VN GPT-2}} \text{VN Caption}$

Nội dung và Phương pháp

BỘ MÃ HÓA CLIP (ViT-B/32)



Backbone (Sử dụng CLIP (ViT-B/32)):

- **Chức năng:** Trích xuất đặc trưng ngữ nghĩa mức cao (High-level semantic features).

Cơ chế (Mechanism):

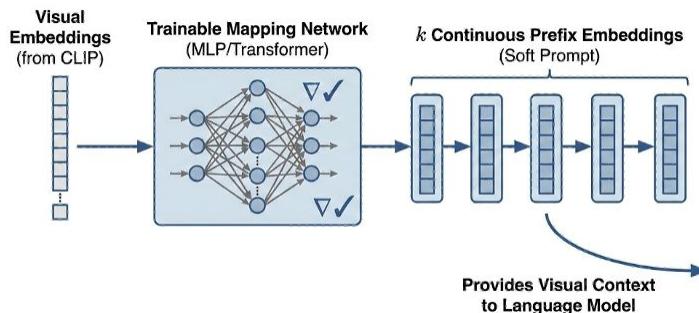
- **Đóng băng hoàn toàn (Frozen Weights).**
- Không cập nhật tham số trong quá trình huấn luyện.

Lợi ích:

- Bảo toàn tri thức tiền huấn luyện (Pre-trained knowledge).
- Giảm chi phí tính toán (Compute costs) ↓



MẠNG ÁNH XẠ (MLP/Transformer)



Vai trò: Cầu nối ngữ nghĩa (Semantic Bridge).

Luồng xử lý:

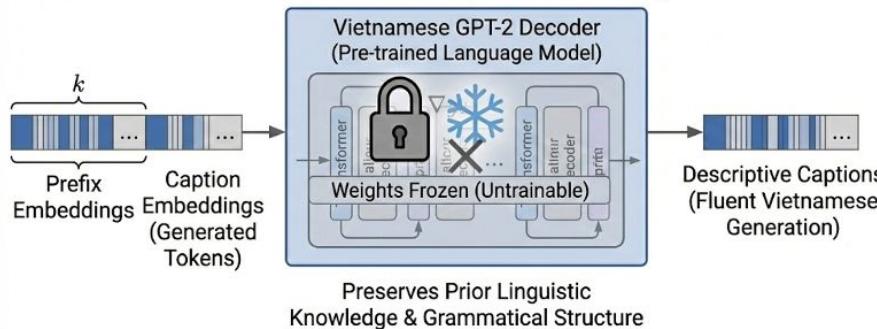
- Input: Visual Embeddings (từ CLIP).
- Model: MLP / Transformer (Nhẹ & Được huấn luyện).
- Output: Prefix Embeddings ("Soft Prompts").

Chức năng:

- Chuyển đổi đặc trưng **Thị giác** Gọi ý **Ngôn ngữ**.
- Dẫn dắt GPT-2 sinh văn bản mà không cần cập nhật trọng số.

Nội dung và Phương pháp

BỘ GIẢI MÃ: VN GPT-2 (FROZEN)



Mô hình: Vietnamese GPT-2 (Đã huấn luyện trước).

Cấu trúc đầu vào (Input):

- Chuỗi nối kết (Concatenated Sequence):
- [Prefix Embeddings, Caption Embeddings]

Lợi ích:

- Bảo toàn tri thức ngôn ngữ (Linguistic Knowledge).
- Sinh câu tiếng Việt trôi chảy (Fluent) & đúng ngữ pháp.

CHIẾN LƯỢC HUẤN LUYỆN

Hàm mục tiêu (Objective):

- Tối thiểu hóa Cross-Entropy Loss.

$$\mathcal{L} = - \sum_{t=1}^T \log P_\theta(y_t | y_{<t}, \mathbf{p})$$

Tối ưu hóa (Optimization):

- Optimizer: AdamW.
- Scheduler: Linear Warmup.

Hiệu quả (Efficiency):

- Chỉ cập nhật tham số θ (Mapping Network).
- \rightarrow Ngăn chặn hiện tượng **Quên kiến thức (Catastrophic Forgetting)**.



Tài liệu tham khảo



- [1]. Ron Mokady, Amir Hertz, Amit H. Bermano: [ClipCap: CLIP Prefix for Image Captioning](#). CoRR abs/2111.09734 (2021)
- [2]. Anh-Cuong Pham, Van-Quang Nguyen, Thi-Hong Vuong, Quang-Thuy Ha: [KTVIC: A Vietnamese Image Captioning Dataset on the Life Domain](#). CoRR abs/2401.08100 (2024)
- [3]. Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, Rita Cucchiara: [From Show to Tell: A Survey on Deep Learning-Based Image Captioning](#). IEEE Trans. Pattern Anal. Mach. Intell. 45(1): 539-559 (2023)