ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN CUỐI KỲ

MÔN XỬ LÝ ẢNH VÀ ỨNG DỤNG (CS406.P12.KHCL)

Đề tài: Image captioning

Giảng viên hướng dẫn:

ThS. CÁP PHẠM ĐÌNH THĂNG

Sinh viên thực hiện:

MAI PHÚC MINH (MSSV 21521127)
TRƯƠNG HỮU THỌ (MSSV 21521479)
NGUYỄN TÔ THIÊN BẢO (MSSV 21521857)
ĐỖ BÁ HUY (MSSV 21522137)

MỤC LỤC

TÓM TẮT ĐỒ ÁN

Đồ án này nghiên cứu và triển khai hệ thống Image Captioning sử dụng bộ dữ liệu KTVIC, với mục tiêu tự động tạo ra các mô tả ngôn ngữ tự nhiên cho hình ảnh. Hệ thống được xây dựng trên nền tảng kết hợp giữa CLIP làm encoder và các mô hình ngôn ngữ lớn (LLMs) cùng kiến trúc transformer làm decoder. CLIP, với khả năng học đồng thời từ cả hình ảnh và văn bản, sẽ trích xuất đặc trưng hình ảnh và tạo ra các biểu diễn chung có thể kết nối với ngữ nghĩa của các mô tả. Sau đó, các mô hình LLMs và transformer sẽ sử dụng những đặc trưng này để tạo ra các câu mô tả chính xác, tự nhiên và phù hợp với ngữ cảnh của hình ảnh. Hệ thống sẽ được triển khai lên web, cho phép người dùng tải lên hình ảnh và nhận lại mô tả tự động, ứng dụng trong các lĩnh vực như trợ lý ảo, tìm kiếm hình ảnh, hoặc hỗ trợ người khiếm thị. Đồ án này không chỉ khám phá khả năng ứng dụng của CLIP và LLMs trong bài toán Image Captioning mà còn cung cấp một giải pháp thực tiễn qua giao diện web, giúp người dùng dễ dàng tương tác với mô hình.

Chương 1 – TỔNG QUAN

1.1. Giới thiệu bài toán

Bài toán Image Captioning là một nhiệm vụ quan trọng trong lĩnh vực trí tuệ nhân tạo, nhằm tạo ra các mô tả ngôn ngữ tự nhiên cho hình ảnh đầu vào. Nhiệm vụ này yêu cầu hệ thống phải không chỉ nhận diện các đối tượng và cảnh vật trong hình ảnh mà còn phải diễn đạt những nhận diện đó thành những câu mô tả có nghĩa, trôi chảy và tự nhiên. Trong bối cảnh phát triển nhanh chóng của các mô hình học sâu, một nhóm nghiên cứu gần đây đã đề xuất một phương pháp tiếp cận mới mẻ và hiệu quả, kết hợp các mô hình mạnh mẽ để xử lý hình ảnh và ngôn ngữ, sử dụng CLIP [1] làm encoder và LLMs [2] (các mô hình ngôn ngữ lớn) cùng transformers làm decoder.

CLIP (Contrastive Language-Image Pretraining) [1] là một mô hình mạnh mẽ do OpenAI phát triển, có khả năng xử lý đồng thời hình ảnh và văn bản trong không gian biểu diễn chung. CLIP có thể hiểu mối quan hệ giữa các đặc trưng hình ảnh và ngữ nghĩa của văn bản, cho phép mô hình tạo ra các đại diện hình ảnh phong phú có thể tương tác trực tiếp với các mô tả ngôn ngữ. Với CLIP làm encoder, hệ thống có thể mã hóa thông tin hình ảnh vào một không gian biểu diễn chung, nơi thông tin hình ảnh và văn bản có thể được đối chiếu và liên kết một cách hiệu quả.

Ở phần decoder, nhóm sử dụng các mô hình ngôn ngữ lớn (LLMs) và transformers có hỗ trợ tiếng Việt, như PhoGPT [3] hoặc mT5 [4], bên cạnh LLMs nhóm sử dụng các biến thể khác của transformers như mBART [5], để tạo ra các mô tả ngôn ngữ tự nhiên. Những mô hình này có khả năng xử lý ngữ nghĩa phức tạp và tạo ra các câu văn mạch lạc, giàu ngữ nghĩa. Việc kết hợp CLIP và LLMs giúp hệ thống không chỉ tạo ra các mô tả chính xác về hình ảnh mà còn có thể sáng tạo ra các mô tả phong phú, phù hợp với ngữ cảnh và đa dạng hơn so với các phương pháp truyền thống.

1.2. Mô tả bài toán

Input:

Ảnh I có kích thước $\{C, H, W\}$. Trong đó:

- C : Số kênh của ảnh (ví dụ: 3 cho ảnh màu RGB)
- H : Chiều cao của ảnh (tính bằng pixel)

- W : Chiều rộng của ảnh (tính bằng pixel)

Output:

Câu caption y_i

Với $y_i = c_1, c_2, \ldots, c_r$

Sao cho chuỗi từ $(c_1, c_2, \ldots, c_r)$ có xác suất điều kiện $P(c_1, c_2, \ldots, c_r | I)$ là lớn nhất đối với ảnh I

1.3. Bảng phân công

Công việc	Nguyễn Tô Thiên Bảo	Đỗ Bá Huy	Mai Phúc Minh	Trương Hữu Thọ
Code (CLIP+mBART)	✓			
Code (CLIP+PhoGPT)		✓		
Code (CLIP+mT5)			✓	
Viết ứng dụng demo				✓
Làm slide thuyết trình	✓	✓	✓	✓
Viết báo cáo	✓	✓	✓	✓

Chương 2 – CÁC HỆ THỐNG VÀ NGHIÊN CỨU LIÊN QUAN

2.1. Bài toán image captioning

Bài toán image captioning là một bài toán trong lĩnh vực thị giác máy tính (Computer Vision) và xử lý ngôn ngữ tự nhiên (Natural Language Processing). Mục tiêu của bài toán là tạo ra một mô tả văn bản chính xác và tự nhiên cho một hình ảnh. Điều này đòi hỏi mô hình phải hiểu nội dung hình ảnh và diễn đạt lại thông qua ngôn ngữ.

2.2. Các nghiên cứu liên quan

Chú thích ảnh, nhiệm vụ tạo ra các mô tả ngôn ngữ tự nhiên cho hình ảnh, đã có những tiến bộ lớn nhờ sự phát triển của học sâu. Các mô hình ban đầu sử dụng các đặc trưng thủ công và kỹ thuật học máy đơn giản, nhưng với sự ra đời của Mạng Nơ-ron Tích Chập (CNNs) và Mạng Nơ-ron Hồi Quy (RNNs), các mô hình như [6] đã đạt được những kết quả đáng kể, kết hợp CNNs để trích xuất đặc trưng và LSTMs để tạo ra văn bản.

Sự xuất hiện của cơ chế attention [7], như trong mô hình [8], đã cải thiện khả năng mô hình tập trung vào các phần quan trọng trong hình ảnh khi tạo ra mô tả. Những cải tiến này đã giúp tạo ra các mô tả chính xác hơn và tự nhiên hơn.

Gần đây, việc sử dụng các mô hình transformer [7] và các mô hình ngôn ngữ lớn (LLMs) [2] như GPT-3 [9] để tạo ra các mô tả tự nhiên từ hình ảnh đã mở ra hướng đi mới cho bài toán này. Các mô hình như VisualBERT [10] và UNITER [11] giúp học các đại diện chung cho hình ảnh và văn bản, cải thiện chất lượng và độ chính xác của mô tả.

Các bộ dữ liệu như MS COCO [12] và Flickr30k [13] đã cung cấp tài nguyên phong phú cho việc nghiên cứu, trong khi các bộ dữ liệu đa ngôn ngữ như Multi30K [14] đã hỗ trợ phát triển các mô hình chú thích ảnh cho nhiều ngôn ngữ khác nhau. Bên cạnh đó, sự phát triển của các bộ dữ liệu chú thích ảnh tiếng Việt như bộ dữ liệu UIT-ViIC [15] một trong những bộ dữ liệu đầu tiên cho bài toán image captioning trên ngôn ngữ tiếng Việt. Bên cạnh đó còn có bộ dữ liệu VieCap4H [16], đây là một bộ dữ liệu image captiong tiếng Việt, với các ảnh và captions trong bộ dữ liệu đề thuộc một chủ đề duy nhất là y khoa. Nhằm mục đích tăng sự đa dạng cho các câu caption bộ dữ liệu gần đây nhất KTVIC [17], các câu caption trong bộ dữ liệu này thuộc nhiều lĩnh vực khác nhau như ăn uống, phong cảnh, động vật,... bộ dữ liệu KTVIC cho ra được các ảnh và các câu caption đa dạng mà không phụ thuộc vào một lĩnh vực cụ thể.

Chương 3 – DỮ LIỆU

3.1. **KTVIC dataset**

Bộ dữ liệu được sử dụng trong đồ án được lấy từ bộ dữ liệu KTVIC [17]. Bộ dữ liệu KTVIC gồm 4,327 tấm ảnh với 21,635 câu captions. Với số lượng ảnh và captions được chia cụ thể thành tập train và test tại Bảng 3.1.

	Train	Test
Image	3,759	558
Caption	18,845	2,790

Bảng 3.1. Số lượng dữ liệu trong tập train và test

3.1.1. Images

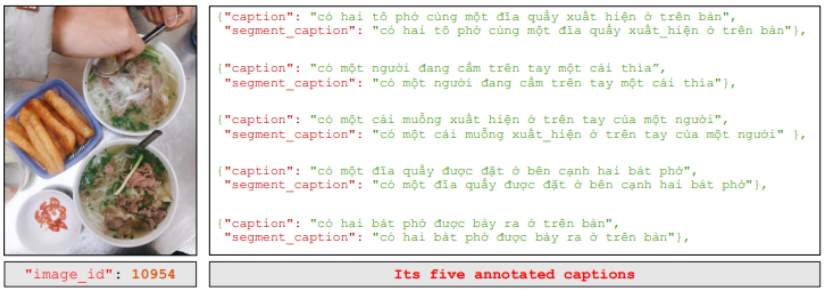
Các tấm ảnh từ bộ dữ liệu KTVIC được thu thập từ bộ dữ liệu UIT-EVJVQA [18]. Nhằm tăng sự đa dạng cho các câu caption các bức ảnh được lấy từ nhiều lĩnh vực như ăn uống, động vật, phong cảnh,... Mỗi ảnh sẽ gồm có 5 captions khác nhau bằng tiếng Việt.

3.1.2. Captions

Các câu caption được viết bởi 3 người đã được huấn luyện trước. Tuân theo 9 quy tắc MS-COCO annotation, trừ quy tắc đầu. Vì quy tắc đầu có khả năng làm hạn chế đi sự đa dạng khi viết ra các caption cho ảnh. Captions trong bộ dữ liệu KTVIC có hai loại:

- ‘caption’: Là chuỗi văn bản mô tả nội dung của bức ảnh, viết bằng tiếng Việt, không được chuẩn hóa hay dùng bất kỳ phương pháp xử lý nào.
- ‘segment_caption’: Là phiên bản của `caption` nhưng được xử lý thành các phân đoạn hoặc dạng chuẩn hóa bằng RDRSegmenter.

Hình 3.1 sẽ cho thấy ví dụ về một mẫu dữ liệu bao gồm cả hình ảnh lẫn câu caption

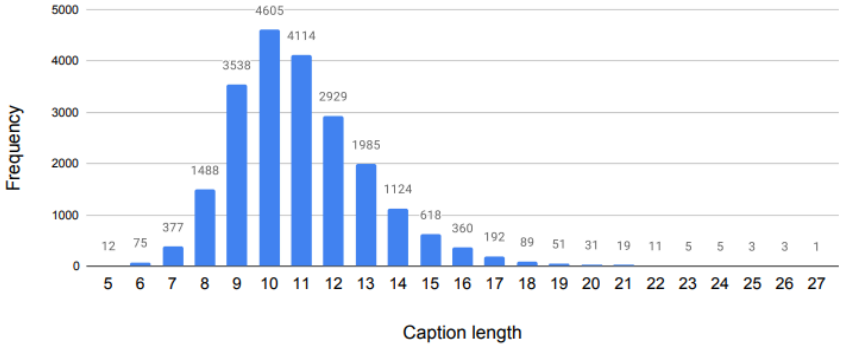


Hình 3.1: Ví dụ về một mẫu dữ liệu

Khi phân tích các câu captions của bộ dữ liệu từ Hình 3.2 có thể thấy được sự đa dạng của bộ dữ liệu cụm từ "xuất hiện" là từ lớn nhất, cho thấy đây là từ được nhắc đến nhiều nhất. Nó có thể ám chỉ việc ghi nhận sự có mặt của người, vật thể hoặc sự kiện trong một tập dữ liệu (ví dụ: hình ảnh, video). Ngoài ra các từ như "phụ nữ" và "gái", các từ này cho thấy tập dữ liệu có liên quan đến phụ nữ, có thể là mô tả về họ hoặc các đặc điểm liên quan đến giới tính. Các cụm từ như "mặc áo", "áo trắng", "áo vàng", "màu xanh", nhấn mạnh sự mô tả về trang phục và màu sắc, cho thấy tập trung vào nhận dạng trang phục hoặc thời trang. Các từ liên quan đến ngữ cảnh không gian và địa điểm cũng có xuất hiện như "đường", "siêu thị", "chợ", "khu vực", "biển", "xung quanh". Ngoài ngữ cảnh, còn có một số từ liên quan đến các chủ đề như hành động, trạng thái, mô tả vật thể và cảnh quan, cụ thể là "đứng", "đi", "quay", "dừng", "xe máy", "hàng", "siêu thị", "màu xanh", "màu đen"

Hình 3.2: Tần suất xuất hiện của các từ trong tập dữ liệu

Bên cạnh sự đa dạng, một số đặc trưng của các câu captions của bộ dữ liệu KTVIC có thể thấy tại Hình 3.3. Các captions có độ dài 10 từ là phổ biến nhất (4605 lần), cho thấy xu hướng tối ưu hóa độ dài caption ở mức vừa phải. Các độ dài khác như 9 từ (3538) và 11 từ (4114) cũng rất phổ biến. Phạm vi tập trung của các câu captions có thể thấy đa phần các câu captions có độ dài từ 8 đến 13 từ, với tần suất cao, tạo thành một "đỉnh" tập trung. Từ Hình 3.3 có thể thấy biểu đồ có dạng phân phối hình chuông lệch phải, với phần đỉnh tập trung ở các chú thích trung bình (8–13 từ) và phần "đuôi" kéo dài về các chú thích dài hơn. Do đó có thể kết luận dữ liệu gợi ý rằng độ dài 8–13 từ là khoảng tối ưu để viết chú thích, vừa đủ để truyền tải ý nghĩa mà không gây quá dài dòng.



Hình 3.3: Độ dài các câu captions và tần suất xuất hiện

Chương 4 – PHƯƠNG PHÁP TIẾP CẬN

Phương pháp tiếp cận trong bài toán image captioning này kết hợp hai thành phần chính: một encoder mạnh mẽ là CLIP và một decoder sử dụng mô hình ngôn ngữ lớn (LLMs) và bên cạnh đó kiến trúc transformer cũng được sử dụng để kết hợp với CLIP. Đây là một sự kết hợp giữa thị giác máy tính và xử lý ngôn ngữ tự nhiên, giúp tạo ra các mô tả ngôn ngữ tự nhiên từ hình ảnh.

4.1. CLIP + PhoGPT

Phương pháp tiếp cận bài toán image captioning trong đồ án này sử dụng sự kết hợp giữa CLIP và PhoGPT, một mô hình ngôn ngữ mạnh mẽ được phát triển để xử lý ngôn ngữ tiếng Việt. Phương pháp này tận dụng sức mạnh của CLIP trong việc hiểu mối quan hệ giữa hình ảnh và văn bản, kết hợp với khả năng sinh ngôn ngữ tự nhiên của PhoGPT để tạo ra các mô tả ngôn ngữ chính xác và tự nhiên từ hình ảnh.

4.1.1 Encoder: CLIP

CLIP là một mô hình học sâu được huấn luyện trên một lượng dữ liệu lớn bao gồm hình ảnh và văn bản, giúp mô hình học được mối quan hệ giữa các đối tượng trong hình ảnh và các mô tả ngữ nghĩa tương ứng. CLIP sử dụng một không gian biểu diễn chung cho hình ảnh và văn bản, cho phép mô hình nhận diện các đặc trưng hình ảnh và kết hợp chúng với thông tin ngữ nghĩa từ văn bản. Khi hình ảnh được đưa vào CLIP, mô hình sẽ chuyển nó thành một vector đặc trưng, giúp mô tả nội dung của hình ảnh dưới dạng một biểu diễn ngữ nghĩa có thể so sánh với các mô tả văn bản.

4.1.2 Decoder: PhoGPT

PhoGPT là một mô hình ngôn ngữ lớn được huấn luyện cho ngôn ngữ tiếng Việt, có khả năng sinh ra các đoạn văn tự nhiên và mạch lạc. Tại Hình 4.1 có thể thấy được mô hình PhoGPT đạt hiệu suất cao nhất trên ngôn ngữ tiếng Việt. Sau khi CLIP mã hóa hình ảnh thành một vector đặc trưng, vector này sẽ được đưa vào PhoGPT, nơi mô hình sẽ sử dụng thông tin hình ảnh để sinh ra các câu mô tả ngữ nghĩa trong ngữ cảnh tiếng Việt. PhoGPT được huấn luyện để hiểu ngữ pháp, ngữ nghĩa và các mối quan hệ giữa các từ trong văn bản, giúp tạo ra các câu văn không chỉ chính xác mà còn tự nhiên, trôi chảy và phù hợp với ngữ cảnh của hình ảnh.

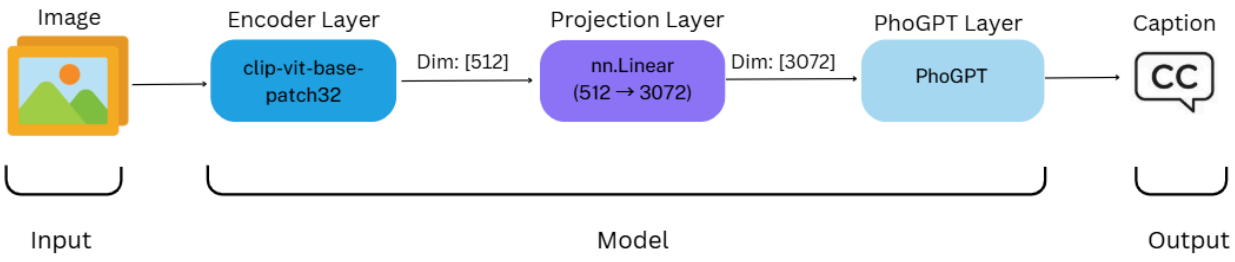
Model	All truthful questions	Vietnam-specific
PhoGPT-4B-Chat	41.7 (83 / 199)	**43.5 (64 / 147)**
GPT-4-0125-preview	**44.7 (89 / 199)**	39.5 (58 / 147)
GPT-3.5-turbo	29.1 (58 / 199)	22.4 (33 / 147)
Gemini Pro 1.0	39.7 (79 / 199)	34.7 (51 / 147)
Vistral-7B-Chat	41.2 (82 / 199)	42.9 (63 / 147)
Sailor-7B-Chat	28.6 (57 / 199)	27.9 (41 / 147)
Sailor-4B-Chat	15.6 (31 / 199)	14.3 (21 / 147)
SeaLLM-7B-v2	20.6 (41 / 199)	13.6 (20 / 147)
VBD-Llama2-7B-50B-Chat	15.6 (31 / 199)	10.9 (16 / 147)
Vinallama-7B-Chat	11.1 (22 / 199)	8.2 (12 / 147)
Gemma-7B-it	8.0 (16 / 199)	6.1 (9 / 147)

Hình 4.1: Kết quả của mô hình PhoGPT so sánh với các mô hình SOTA về LLMs hiện nay như GPT-4 hay GPT-3.5

4.1.3 Kiến trúc tổng thể của mô hình

Mô hình được chia thành ba phần chính, có thể thấy được tại Hình 4.2:

- **Input:** Ảnh đầu vào được sử dụng làm dữ liệu gốc để mô hình sinh chú thích.
- **Model:**
  - **Encoder Layer:** Sử dụng CLIP-ViT-Base-Patch32 để trích xuất đặc trưng ảnh, đầu ra là vector 512 chiều.
  - **Projection Layer:** Một tầng Linear chuyển vector đặc trưng từ không gian 512 chiều sang 3072 chiều (đầu vào cho PhoGPT). Chuyển đổi này đảm bảo tính tương thích với mô hình ngôn ngữ tiếp theo.
  - **PhoGPT Layer:** Mô hình ngôn ngữ PhoGPT nhận vector 3072 chiều để sinh chú thích tiếng Việt.
- **Output**: Câu chú thích tiếng Việt phù hợp với nội dung ảnh.



Hình 4.2: Kiến trúc của tổng quan của mô hình khi kết hợp CLIP và PhoGPT

4.2. CLIP + mBART

Phương pháp tiếp cận bài toán image captioning này sử dụng sự kết hợp giữa CLIP và mBART, một mô hình ngôn ngữ mạnh mẽ đa ngữ, trong đó tiếng Việt là một trong những ngôn ngữ được mBART huấn luyện chuyên sâu. Kỹ thuật này tận dụng khả năng mã hóa hình ảnh của CLIP và khả năng sinh văn bản tự nhiên của mBART, giúp tạo ra các mô tả ngôn ngữ chính xác và tự nhiên từ hình ảnh.

4.2.1 Encoder: CLIP

CLIP, như đã mô tả ở phần trên, là một mô hình học sâu được thiết kế để hiểu và ánh xạ mối quan hệ ngữ nghĩa giữa hình ảnh và văn bản vào cùng một không gian biểu diễn. Mô hình mã hóa hình ảnh thành vector đặc trưng, giúp cô đọng nội dung ngữ nghĩa của hình ảnh, tạo tiền đề cho mBART xử lý.

4.2.2 Decoder: mBART

mBART là một mô hình ngôn ngữ lớn dựa trên kiến trúc BART, được huấn luyện đa ngôn ngữ, trong đó tiếng Việt chiếm vị trí quan trọng với lượng dữ liệu huấn luyện lớn (*Hình 4.3*).

Sau khi nhận vector đặc trưng từ CLIP, mBART tận dụng thông tin này để sinh ra các mô tả ngôn ngữ ngữ nghĩa chính xác, tự nhiên và phù hợp với ngữ cảnh tiếng Việt. mBART được huấn luyện với bài toán khôi phục văn bản từ dữ liệu bị làm nhiễu, giúp mô hình hiểu rõ ngữ pháp, ngữ nghĩa và các mối quan hệ phức tạp giữa các từ trong văn bản, tạo ra các câu văn chính xác, tự nhiên, trôi chảy và phù hợp với ngữ cảnh của hình ảnh.

Code	Language	Tokens/M	Size/GB
En	English	55608	300.8
Ru	Russian	23408	278.0
Vi	Vietnamese	24757	137.3
Ja	Japanese	530 (*)	69.3
De	German	10297	66.6
Ro	Romanian	10354	61.4
Fr	French	9780	56.8
Fi	Finnish	6730	54.3
Ko	Korean	5644	54.2
Es	Spanish	9374	53.3
Zh	Chinese (Sim)	259 (*)	46.9
It	Italian	4983	30.2
Nl	Dutch	5025	29.3
Ar	Arabic	2869	28.0
Tr	Turkish	2736	20.9
Hi	Hindi	1715	20.2
Cs	Czech	2498	16.3
Lt	Lithuanian	1835	13.7
Lv	Latvian	1198	8.8
Kk	Kazakh	476	6.4
Et	Estonian	843	6.1
Ne	Nepali	237	3.8
Si	Sinhala	243	3.6
Gu	Gujarati	140	1.9
My	Burmese	56	1.6

Hình 4.3 - Kích thước bộ dữ liệu của 25 ngôn ngữ trong dataset;

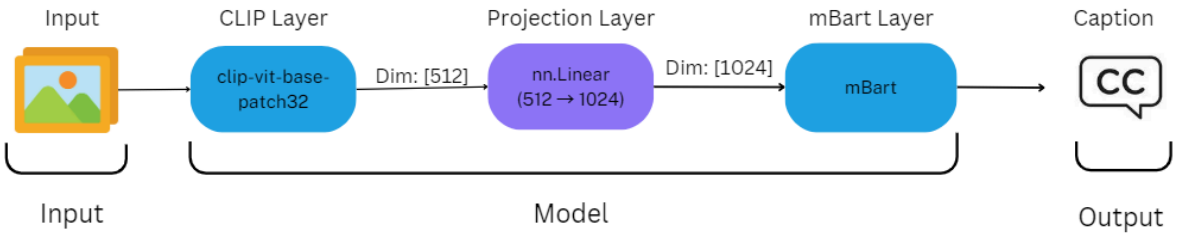
4.2.3 Kiến trúc tổng thể của mô hình

Mô hình được chia thành ba phần chính, có thể thấy được tại *Hình 4.4*:

- **Input:** Ảnh đầu vào được sử dụng làm dữ liệu gốc để mô hình sinh chú thích.
- **Model:**
    - **Encoder Layer:** Sử dụng CLIP-ViT-Base-Patch32 để trích xuất đặc trưng ảnh, đầu ra là vector 512 chiều.
    - **Projection Layer:** Một tầng Linear chuyển vector đặc trưng từ không gian 512 chiều sang 1024 chiều (đầu vào cho mBART layer). Chuyển đổi này đảm bảo tính tương thích với mô hình ngôn ngữ tiếp theo.
    - **mBART Layer:** Mô hình ngôn ngữ mbart-large-50-many-to-many-mmt

nhận vector 1024 chiều để sinh chú thích tiếng Việt.

- **Output**: Câu chú thích tiếng Việt phù hợp với nội dung ảnh.



Hình 4.4 - Kiến trúc của tổng quan của mô hình khi kết hợp CLIP và mBART

4.3. CLIP + mT5

Nhóm không chỉ thử nghiệm với những phương pháp tiếp cho bài toán image captioning như trên. Nhóm cũng thiết kế model kết hợp giữa CLIP và mT5, sử dụng tính năng mã hóa hình ảnh của CLIP và sử dụng mT5 để giải mã ra thành caption cho hình ảnh được nhập vào.

4.3.1 Encoder: CLIP

Model CLIP ở đây cũng là clip-vit-base-patch32 và cũng được sử dụng như 2 model trên, nên ở đây sẽ không nhắc lại nữa.

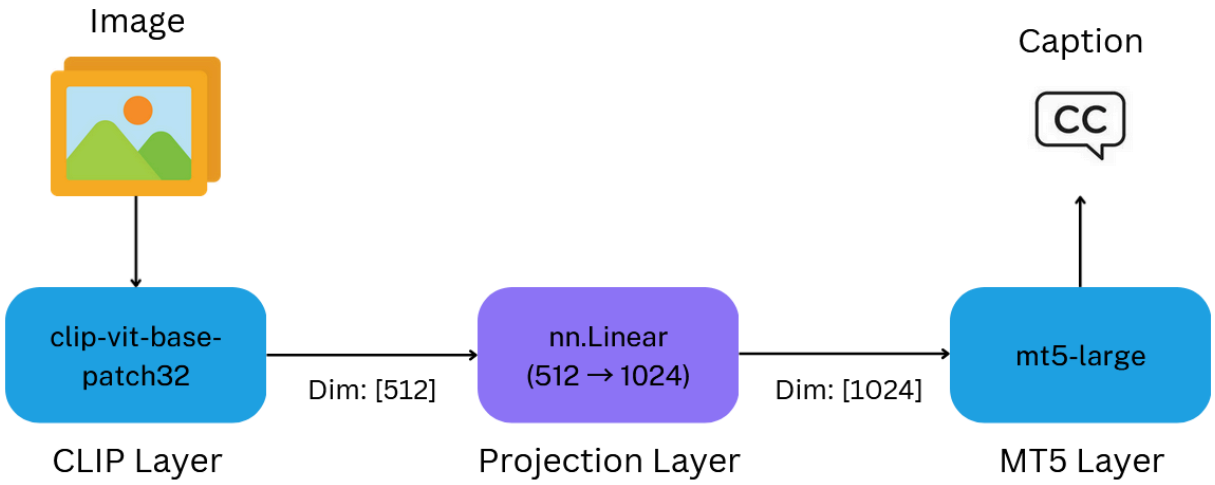
4.3.2 Decode: mT5

mT5 (Multilingual Text-to-Text Transfer Transformer) là model LLM được Google Research tạo ra. mT5 được dựa trên model LLM T5, cũng bởi Google Research tạo ra. mT5 được cải tiến từ T5, được chỉnh sửa cấu trúc và train lại trên 101 ngôn ngữ, trong đó có cả tiếng Việt. mT5 được chọn chủ yếu vì 2 lý do:

- mT5 đã được pretrain trên nhiều ngôn ngữ, trong đó có tiếng Việt
- mT5, cụ thể là mt5-large được chọn để finetune, vừa không quá to, vừa không quá nhỏ, tầm 1 triệu tham số, vừa tầm nhóm có thể finetune.

Sau khi nhận vector đặc trưng từ CLIP, mT5 sẽ tiếp tục tạo ra những embedding có thể tả hình ảnh được nhập vào, sử dụng thông tin từ vector đặc trưng của ảnh, những thông tin, ngữ cảnh được train bởi Google và được finetine bởi nhóm. Và từ những embedding được tạo ra, model chuyển thành những từ ngữ hy vọng tạo thành các câu văn chính xác, tự nhiên, trôi chảy và phù hợp với ngữ cảnh của hình ảnh.

4.3.2 Kiến trúc tổng thể của mô hình



Hình 4.5 - Kiến trúc của tổng quan của mô hình khi kết hợp CLIP và mT5

Model CLIP + mT5 này cũng khá tượng tự như 2 model trên:

- **Input:** Ảnh đầu vào được sử dụng làm dữ liệu gốc để model tạo caption.
- **Model:**
  - **Encoder Layer:** Sử dụng CLIP-ViT-Base-Patch32 để trích xuất đặc trưng ảnh, đầu ra là vector 512 chiều.
  - **Projection Layer:** Một tầng Linear chuyển vector đặc trưng từ không gian 512 chiều sang 1024 chiều (đầu vào cho mt5-large). Chuyển đổi này đảm bảo tính tương thích với mt5-large.
  - **mT5 Layer:** Mô hình ngôn ngữ mT5 nhận vector 1024 chiều để sinh chú thích tiếng Việt.
- **Output**: Câu chú thích tiếng Việt phù hợp với nội dung ảnh.

Chương 5 – KẾT QUẢ THỰC NGHIỆM

5.1. Đánh giá hiệu suất các mô hình trên bài toán mô tả ảnh

Tại Bảng 5, kết quả so sánh cho bài toán mô tả ảnh (image captioning) giữa mô hình được sử dụng trong bài báo của bộ dữ liệu KTVIC và các mô hình PhoGPT, mBART, mT5-large cho thấy sự khác biệt rõ rệt về hiệu suất trên các thước đo. KTVIC dẫn đầu ở các chỉ số BLEU-1 (74.7), BLEU-4 (40.6) và CIDEr (136.0), thể hiện khả năng tạo mô tả chính xác và bám sát tham chiếu. Trong khi đó, mT5-large nổi bật với điểm METEOR (53.94) và ROUGE-L (61.07), cho thấy sự linh hoạt trong việc tạo các mô tả tự nhiên và bao quát hơn. PhoGPT có hiệu suất thấp hơn đáng kể, đặc biệt ở BLEU-4 và CIDEr, cho thấy khả năng mô tả còn hạn chế. Mô hình trong bài báo KTVIC phù hợp hơn khi ưu tiên độ chính xác cao, trong khi mT5-large có thể được sử dụng nếu cần các mô tả tự nhiên và tổng quát.

Nguyên nhân giúp cho mô hình trong bài báo KTVIC đạt kết quả cao hơn các mô hình như PhoGPT, mT5 hay mBART do tác giả trong bài báo KTVIC sử dụng mô hình GRIT [19], một mô hình được huấn luyện duy nhất cho bài toán image captioning. Ngoài ra, những mô hình nhóm sử dụng đa phần đều có số lượng tham số lớn trong đó thấp nhất là mT5-large với số lượng tham số gần 900 triệu, tại chương 3 có thể thấy số lượng của bộ dữ liệu KTVIC chưa đủ lớn để huấn luyện các mô hình có lượng tham số lớn mà nhóm sử dụng điều này có thể dẫn đến các mô hình nhóm sử dụng bị overfit do đó không thể đạt hiệu suất cao hơn so với mô hình được sử dụng trong bài báo KTVIC

	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr
Paper (GRIT)	**74.7**	**40.6**	36.6	59.7	**136.0**
PhoGPT	53.90	22.73	38.73	50.80	50.98
mBART	37.50	13.65	32.37	41.90	101.87
mT5-large	68.67	34.90	**53.94**	**61.07**	94.70

Bảng 5: So sánh các mô hình nhóm sử dụng với mô hình GRIT của bài báo KTVIC

5.2. Đánh giá hiệu suất các phiên bản mT5, PhoGPT và mBART trong bài toán mô tả ảnh

5.2.1 mT5

Bên cạnh việc so sánh với mô hình được sử dụng trong bài báo KTVIC, với mỗi mô hình nhóm tạo ra các biến thể bằng cách thay đổi các tham số chẳng hạn như số epoch, batch_size,... để để chọn ra mô hình tối ưu nhất. Tại Bảng 6a, nhóm đã tạo ra các biến thể trong việc kết hợp hai mô hình CLIP và mT5 trong đó tại một vài lần thử nghiệm nhóm cho freeze CLIP hoặc mT5 hoặc đôi khi freeze và unfreeze cả hay mô hình. Kết quả từ Bảng 6a so sánh các phiên bản của mô hình mT5 cho bài toán image captioning cho thấy chiến lược huấn luyện ảnh hưởng đáng kể đến hiệu suất.

- **MT5-large (freeze CLIP+MT5)**:
  + Tất cả các chỉ số đều bằng **0.00**, cho thấy chiến lược "đóng băng" toàn bộ CLIP và MT5 không hiệu quả. Điều này có thể xảy ra vì mô hình không được tối ưu hóa hoặc không học được bất kỳ thông tin mới nào từ dữ liệu.
- **MT5-large (No freeze)**:
  + BLEU-1 đạt **55.28** và BLEU-4 đạt **17.20**, cho thấy mô hình có khả năng học nhưng hiệu suất còn hạn chế so với các phiên bản khác.
  + METEOR (37.66) và ROUGE-L (50.14) cũng khá thấp, thể hiện khả năng tạo ra mô tả tự nhiên và bao quát chưa tốt.
  + CIDEr (31.21) cho thấy độ khớp giữa mô tả của mô hình và tham chiếu rất thấp, mô hình cần cải thiện sự tương thích.
- **MT5-base (freeze CLIP)**:
  + Hiệu suất khá tốt với BLEU-1 đạt **63.01** và BLEU-4 đạt **26.75**, dù thấp hơn phiên bản mT5-large (freeze CLIP).
  + METEOR (46.50) và ROUGE-L (55.56) cho thấy mô hình có khả năng tạo mô tả mượt mà và bám sát tham chiếu.
  + CIDEr (59.96) cũng khá ổn, mặc dù thấp hơn mT5-large (freeze CLIP).
- **MT5-large (freeze CLIP)**:
  + Phiên bản này đạt kết quả tốt nhất trên tất cả các thước đo, với BLEU-1 (**68.67**), BLEU-4 (**34.90**), METEOR (**53.94**), ROUGE-L (**61.07**) và CIDEr (**94.70**).
  + Kết quả này cho thấy việc "đóng băng" CLIP trong khi huấn luyện MT5-large giúp mô hình học hiệu quả hơn, cân bằng giữa tính tự nhiên và độ chính xác của mô tả.

MT5-large (freeze CLIP) là phiên bản tối ưu nhất, cho thấy rằng chiến lược "đóng băng" CLIP nhưng vẫn huấn luyện MT5-large giúp đạt hiệu suất cao trong bài toán.

MT5-base (freeze CLIP) là lựa chọn tốt nếu tài nguyên hạn chế nhưng vẫn cần hiệu quả khá cao. Việc không "đóng băng" hoặc "đóng băng" toàn bộ CLIP và MT5 dẫn đến hiệu suất kém, cho thấy tầm quan trọng của việc chọn chiến lược huấn luyện phù hợp.

.	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr
MT5-large (freeze CLIP+MT5)	0.00	0.00	0.00	0.00	0.00
MT5-large (No freeze)	55.28	17.20	37.66	50.14	31.21
MT5-base (freeze CLIP)	63.01	26.75	46.50	55.56	59.96
MT5-large (freeze CLIP)	**68.67**	**34.90**	**53.94**	**61.07**	**94.70**

Bảng 6a: Kết quả của các phiên bản mô hình mT5

5.2.2 PhoGPT

Bảng 6b cho thấy kết quả sau khi huấn luyện và đánh giá của các phiên bản PhoGPT trên tập dữ liệu KTVIC

- **PhoGPT (30 epochs)**:
  + Hiệu suất rất thấp trên tất cả các chỉ số, với BLEU-1 (12.93), BLEU-4 (2.09), METEOR (23.92), ROUGE (29.92), và CIDEr (7.03).
  + Điều này cho thấy rằng huấn luyện mô hình trong quá nhiều epoch có thể dẫn đến hiện tượng overfitting, làm giảm khả năng tạo ra mô tả chính xác.
- **PhoGPT (15 epochs)**:
  + Hiệu suất cải thiện đáng kể so với 30 epochs, với BLEU-1 (48.02) và BLEU-4 (13.72), nhưng vẫn chưa đạt mức tối ưu.
  + Các chỉ số khác như METEOR (29.16), ROUGE (42.20), và CIDEr (29.69) cũng tăng lên, cho thấy mô hình đang dần học được cách tạo mô tả tốt hơn.
- **PhoGPT (3 epochs)**:
  + Đây là phiên bản đạt hiệu suất tốt nhất, với BLEU-1 (53.90), BLEU-4 (22.73), METEOR (38.73), ROUGE (50.80), và CIDEr (50.98).
  + Kết quả này cho thấy rằng mô hình PhoGPT hoạt động hiệu quả nhất khi được huấn luyện trong thời gian ngắn, tránh được hiện tượng overfitting và giữ được khả năng tổng quát hóa.

PhoGPT (3 epochs) là phiên bản tối ưu nhất trong ba thử nghiệm, đạt kết quả cao nhất trên tất cả các thước đo. Số lượng epoch lớn hơn (15 và 30) không giúp cải thiện hiệu suất, mà ngược lại làm giảm độ chính xác và tự nhiên của mô tả do overfitting. Kết quả này nhấn mạnh tầm quan trọng của việc lựa chọn số lượng epoch phù hợp trong huấn luyện để đạt được sự cân bằng giữa việc học và tổng quát hóa.

	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr
PhoGPT (30 epoch)	12.93	2.09	23.92	29.92	7.03
PhoGPT (15 epochs)	48.02	13.72	29.16	42.20	29.69
PhoGPT (3 epoch)	**53.90**	**22.73**	**38.73**	**50.80**	**50.98**

Bảng 6b: Kết quả của các phiên bản mô hình PhoGPT

5.2.3 mBART

Đánh giá kết quả của các phiên bản của mô hình mBART trên tập dữ liệu KTVIC có thể xem được tại Bảng 6c.

- **mBART (10 epochs)**:
  + BLEU-1 đạt **36.73** và BLEU-4 đạt **12.47**, cho thấy khả năng tái tạo các mô tả ngắn chính xác nhưng còn hạn chế với các mô tả dài.
  + METEOR (**31.08**) và ROUGE (**41.31**) ở mức trung bình, thể hiện sự cải thiện về tính tự nhiên nhưng chưa thực sự tốt.
  + CIDEr đạt **95.40**, cho thấy sự tương đồng với mô tả tham chiếu ở mức chấp nhận được.
- **mBART (3 epochs)**:
  + Hiệu suất cải thiện so với phiên bản 10 epochs ở tất cả các chỉ số, với BLEU-1 (**37.52**) và BLEU-4 (**13.65**) tăng nhẹ, thể hiện khả năng mô tả tốt hơn.
  + METEOR (**32.37**) và ROUGE (**41.90**) tăng nhẹ, cho thấy các mô tả trở nên mượt mà và bao quát hơn.
  + CIDEr đạt **101.87**, cao hơn rõ rệt so với phiên bản 10 epochs, cho thấy khả năng tương thích với tham chiếu được nâng cao.

Phiên bản mBART huấn luyện với **3 epochs** đạt hiệu suất cao hơn phiên bản 10 epochs trên tất cả các thước đo, đặc biệt là CIDEr, chứng tỏ thời gian huấn luyện ngắn hơn giúp mô hình tránh overfitting và học tốt hơn từ dữ liệu. Kết quả này nhấn mạnh rằng việc lựa chọn số lượng epoch phù hợp là yếu tố quan trọng để tối ưu hóa hiệu suất của mô hình trong bài toán mô tả ảnh.

.	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr
mBART (10 epoch)	36.73	12.47	31.08	41.31	95.40
mBART (3 epoch)	**37.52**	**13.65**	**32.37**	**41.90**	**101.87**

Bảng 6c: Kết quả của các phiên bản mô hình mBART

Chương 6 – TRIỂN KHAI

6.1 Mục tiêu triển khai

Việc triển khai hệ thống **Image Captioning** lên nền tảng web không chỉ đơn thuần là bước tiếp theo trong quá trình phát triển ứng dụng mà còn mang lại giá trị thiết thực trong cả lĩnh vực nghiên cứu và ứng dụng thực tiễn. Các mục tiêu cụ thể bao gồm:

6.1.1 Cải thiện khả năng tiếp cận của người dùng

Trước đây, việc sử dụng mô hình **Image Captioning** thường yêu cầu người dùng phải có kiến thức cơ bản về lập trình hoặc làm việc trên môi trường coding. Điều này trở thành rào cản lớn đối với những người không có nền tảng kỹ thuật.

Triển khai hệ thống trên nền tảng web với giao diện đồ họa trực quan sẽ loại bỏ rào cản này, giúp người dùng cuối dễ dàng tương tác với mô hình chỉ bằng các thao tác đơn giản như tải ảnh, chọn mô hình và nhận kết quả.

6.1.2 Tạo nền tảng tích hợp các mô hình đa dạng

Việc triển khai ba phương pháp ***(CLIP + mBart, CLIP + mT5, CLIP + PhoGPT)*** trên cùng một nền tảng web cho phép người dùng dễ dàng lựa chọn, so sánh và đánh giá hiệu quả của từng mô hình trên cùng một ảnh đầu vào.

Điều này không chỉ giúp người dùng cuối có cái nhìn toàn diện, trực quan hơn về hệ thống mà còn hỗ trợ các nhà nghiên cứu trong việc tối ưu hóa và cải tiến các phương pháp.

6.1.3 Ứng dụng thực tiễn trong các lĩnh vực đa ngành

Một hệ thống Image Captioning trên web có tiềm năng ứng dụng trong nhiều lĩnh vực như:

- **Thương mại điện tử:** Tự động tạo mô tả sản phẩm từ hình ảnh.
- **Y tế:** Phân tích và mô tả hình ảnh chẩn đoán, mô tả tên và công dụng các loại thuốc.
- **Giáo dục:** Hỗ trợ học sinh, sinh viên trong việc hiểu ngữ cảnh từ hình minh họa.
- **Hỗ trợ người khiếm thị:** Mô tả hình ảnh giúp người khiếm thị tiếp cận thông tin thị giác.

6.1.4 Khả năng mở rộng và phát triển lâu dài

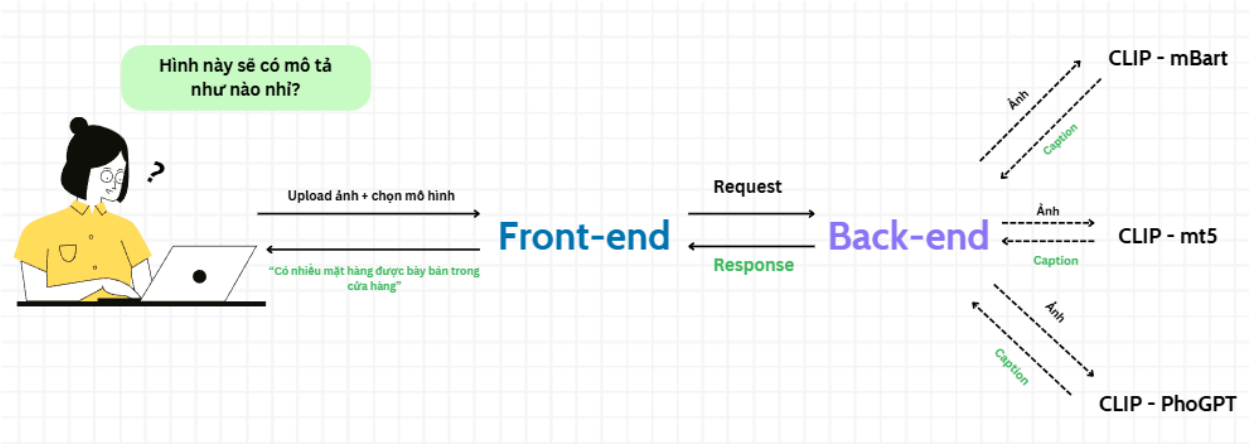
Triển khai trên nền tảng web không chỉ giới hạn trong việc xử lý hình ảnh mà còn mở ra các hướng phát triển khác như tích hợp thêm các chức năng chỉnh sửa ảnh, dịch caption sang nhiều ngôn ngữ, hoặc kết hợp các tính năng liên quan đến tìm kiếm hình ảnh dựa trên nội dung.

6.1.5 Thu thập phản hồi thực tế từ người dùng

Hệ thống web cung cấp một môi trường lý tưởng để thu thập phản hồi từ người dùng về chất lượng caption được sinh ra. Dữ liệu thu thập này không chỉ giúp đánh giá khách quan hiệu quả của từng mô hình mà còn là cơ sở để cải thiện và tối ưu hệ thống trong tương lai.

6.2 Kiến trúc hệ thống

Hệ thống được thiết kế dựa trên kiến trúc ***client-server***, trong đó Frontend đảm nhiệm giao diện và tương tác người dùng, còn Backend thực hiện xử lý dữ liệu và vận hành các mô hình AI. Mô hình này đảm bảo phân tách rõ ràng giữa giao diện và logic xử ly, giúp dễ dàng mở rộng và bảo trì hệ thống. Sơ đồ kiến trúc được thể hiện ở *Hình 6.1*.



Hình 6.1 - Kiến trúc hệ thống client-server

6.2.1 Frontend

Nhóm quyết định sử dụng **ReactJS** là một thư viện JavaScript phổ biến được sử dụng để xây dựng giao diện người dùng. Lý do chọn ReactJS:

- **Tính chất xây dựng giao diện dựa theo chia các component (Component-based architecture)**: giúp dễ dàng tái sử dụng, giúp dễ dàng quản lý và phát triển giao diện phức tạp.
- **Hiệu năng cao nhờ việc React sử dụng Virtual DOM** để tối ưu hóa quá trình cập nhất giao diện, đảm bảo hiệu năng tốt ngay khi có nhiều tương tác đồng thời.
- **Hỗ trợ sử dụng thư viện React-query**, một thư viện mạnh mẽ được sử dụng để quản lý trạng thái và thực hiện các request API một cách hiệu quả, hỗ trợ làm mới dữ liệu và tối ưu hóa việc giao tiếp với Backend.

6.2.2 Backend

Flask là một framework web nhẹ của Python, được sử dụng rộng rãi trong các ứng dụng AI/ML. Chẳng những thế, Flask có cúc trúc đơn giản, dễ dàng tích hợp các mô hình AI được xây dựng bằng Python, hỗ trợ xây dựng API RESTful mạnh mẽ và dễ dàng, giúp giao tiếp giữa Frontend và Backend diễn ra hiệu quả, ngoài ra còn tương thích tốt với nhiều thư viện Python như Pillow để xử lý ảnh, transformers để tải mô hình AI,... Với những lý do trên thì Flask chính là lựa chọn thích phù hợp nhất cho nhóm trong việc xây dựng Backend.

Về mô hình dịch vụ, Backend sẽ cung cấp dịch vụ chính thông qua API RESTful **POST /<tên mô hình> (ví dụ: POST /clip-mbart, POST /clip-mt5)**: nhận ảnh từ Frontend, thực hiện xử lý với mô hình mà người dùng đã chọn và trả về caption.

6.3 Quy trình triển khai

- **Bước 1: Chuẩn bị mô hình**

Tất cả các mô hình trước đó sẽ được huấn luyện sao cho phù hợp với mục tiêu, yêu cầu và đưa vào server. Flask sẽ sử dụng thư viện transformers để tải mô hình và xử lý dữ liệu đầu vào.

- **Bước 2: Xây dựng Backend (Flask)**

Ở bước này, nhóm tiến hành xây dựng các endpoint API RESTful để phục vụ các chức năng như như nhận ảnh, gửi trả caption được sinh ra và quản lý mô hình, cụ thể là dùng các thư viện như Pillow để xử lý ảnh, transformers để gọi mô hình ngôn ngữ.

- **Bước 3: Phát triển Frontend (ReactJS)**

Xây dựng giao diện trực quan cho người dùng với các chức năng chính như là: upload ảnh, dropdown menu để người dùng có thể chọn mô hinh, hiển thị caption và ảnh đã xử lý, sử dụng ***React-query*** để quản lý các request API và thêm hiệu ứng render chữ theo kiểu typing để nâng cao trải nghiệm người dùng.

- **Bước 4: Thử nghiệm hệ thống**

Kiểm tra toàn bộ quy trình từ upload ảnh đến sinh ra caption trên các mô hình, qua đó đánh giá giao diện, hiệu năng và độ chính xác của hệ thống (hiện tại nhóm chỉ có thể triển khai 2 mô hình lên web vì lý do hạn chế tài nguyên, cụ thể là hai mô hình **CLIP-mBART** và **CLIP-mT5**).

6.4 Kết quả triển khai

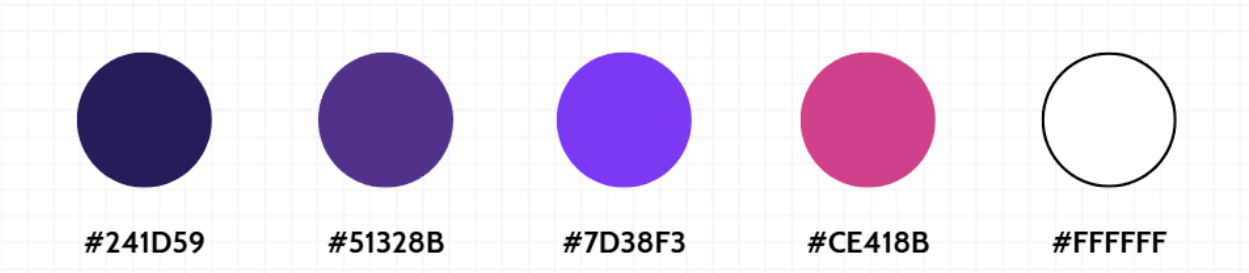
- *Về mặt hiệu năng,* tốc độ sinh ra caption khá nhanh, trung bình **3-4s** cho mỗi ảnh đối với mô hình CLIP + mtT5 và trung bình **5-7s** cho mỗi ảnh đối với mô hình CLIP + mBART.
- *Về độ chính xác,* tùy rằng hai mô hình cho ra các captions khác nhau, tuy nhiên các captions được sinh ra đều phù hợp với ngữ cảnh của ảnh đầu vào.
- *Về trải nghiệm người dùng,* giao diện trực quan và dễ sử dụng, phù hợp cho cả người không có kiến thức chuyên môn.

6.5 Kết luận

Việc triển khai hệ thống **Image Captioning** lên nền tảng web đã đáp ứng được các mục tiêu đề ra, bao gồm cải thiện khả năng tiếp cận và khả năng ứng dụng thực tiễn. Tuy nhiên hệ thống vẫn còn một khuyết điểm chưa khắc phục được và có thể tìm cách để cải thiện trong tương lai là có thể triển khai các mô hình to lớn như **CLIP + PhoGPT** và có thể tạo ra các câu captions mô tả từ tổng quát đến chi tiết nhỏ nhặt.

6.6 Demo

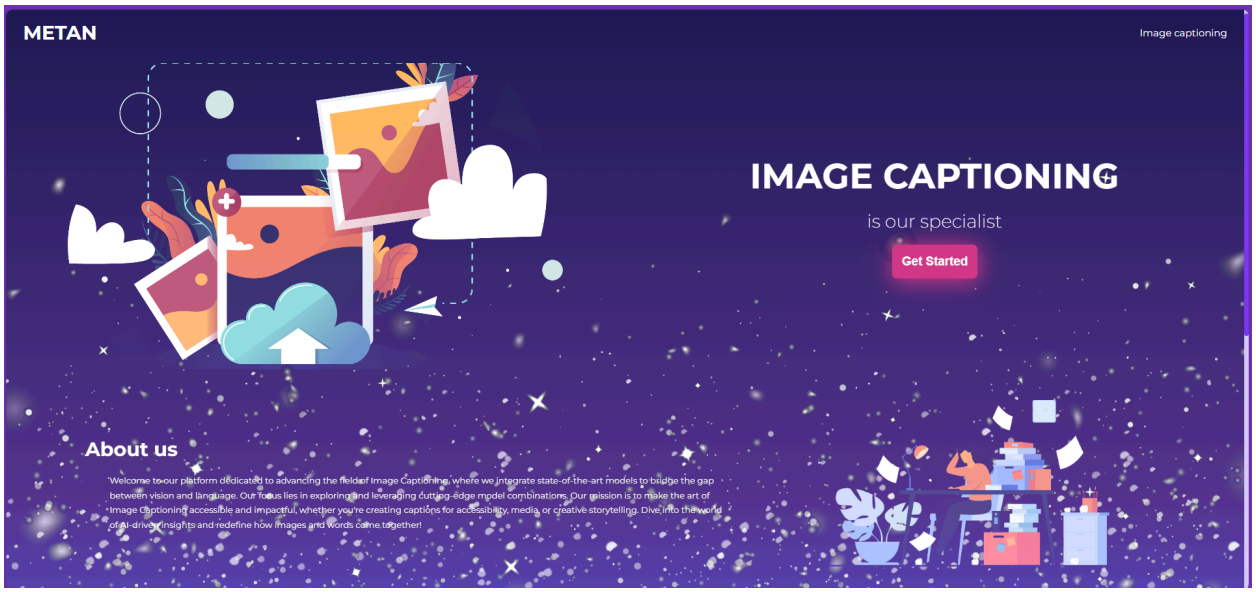
Về màu sắc, nhóm quyết định sử dụng palette màu sau: **#241D59, #51328B, #7D38F3, #CE418B, #FFFFFF**. Palette này chủ yếu sử dụng các tông tím và hồng để tạo cảm giác trẻ trung, ưa nhìn giúp gây ấn tượng cho người dùng khi sử dụng.



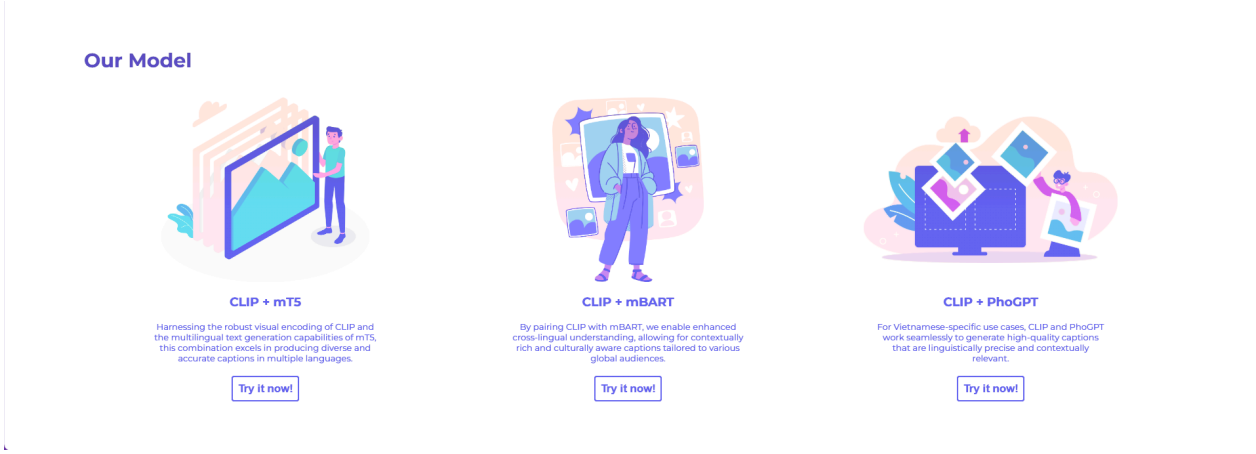
Hình 6.2 - Palette màu của trang web

Về giao diện, sẽ gồm có 2 trang chính: **Home và Upload.**

- **Trang Home** đóng vai trò là phần giới thiệu tổng quan về hệ thống cho người dùng, cung cấp thông tin giới thiệu về nhóm phát triển và các mô hình được sử dụng trong nghiên cứu Image Captioning. Trang này sẽ tạo cảm giác chuyên nghiệp, giúp người dùng hiểu rõ về dự án và những mô hình hiện có trước khi sử dụng.



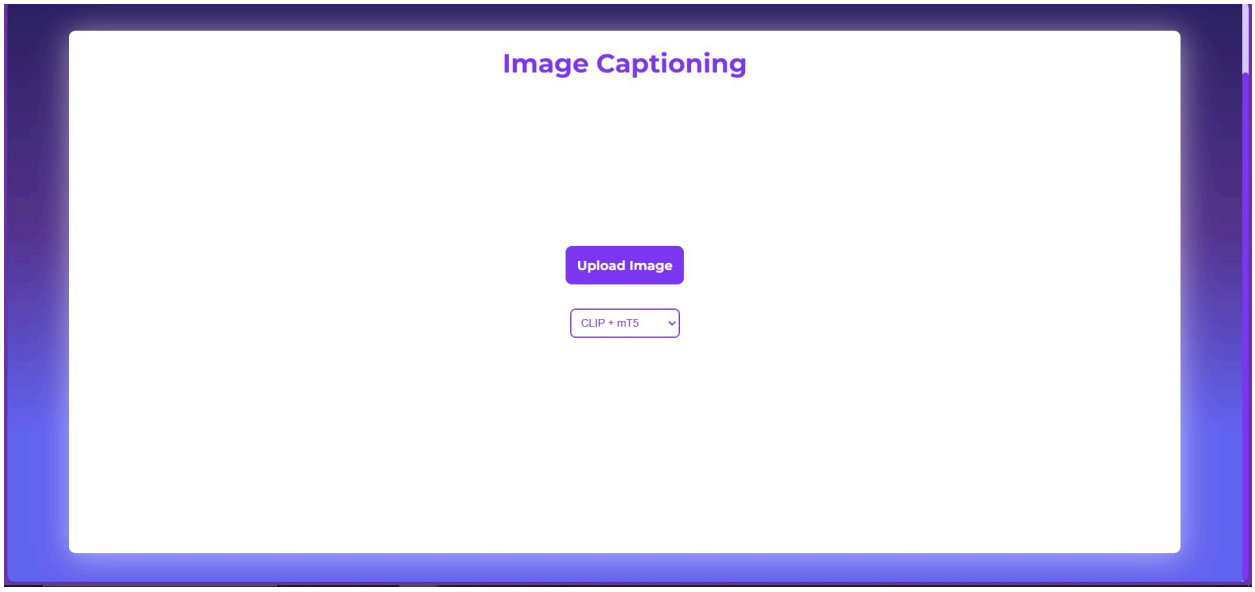
Hình 6.3 - Trang Home giới thiệu tổng quan về nhóm và hệ thống



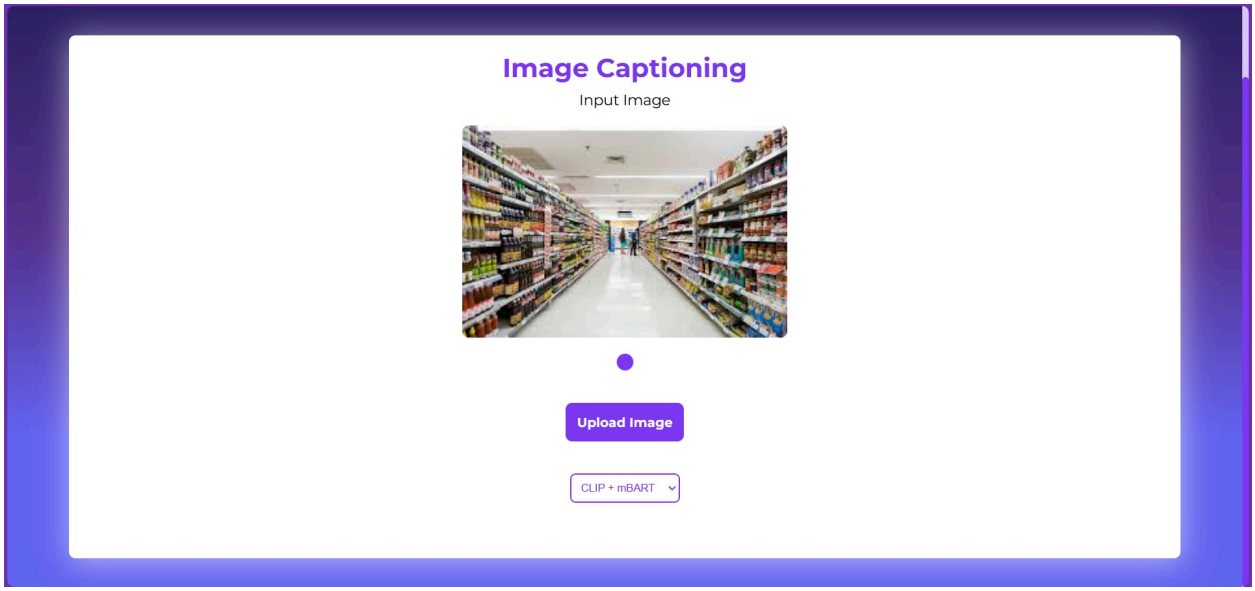
Hình 6.4 - Trang Home giới thiệu sơ lược về các mô hình

- **Trang Image Captioning** sẽ cung cấp giao diện cho người dùng upload ảnh, chọn mô hình và tự động sinh caption cho ảnh vừa được upload. Trang này sẽ cung cấp quy trình rõ ràng và đơn giản để người dùng không phải đọc tài liệu hướng dẫn

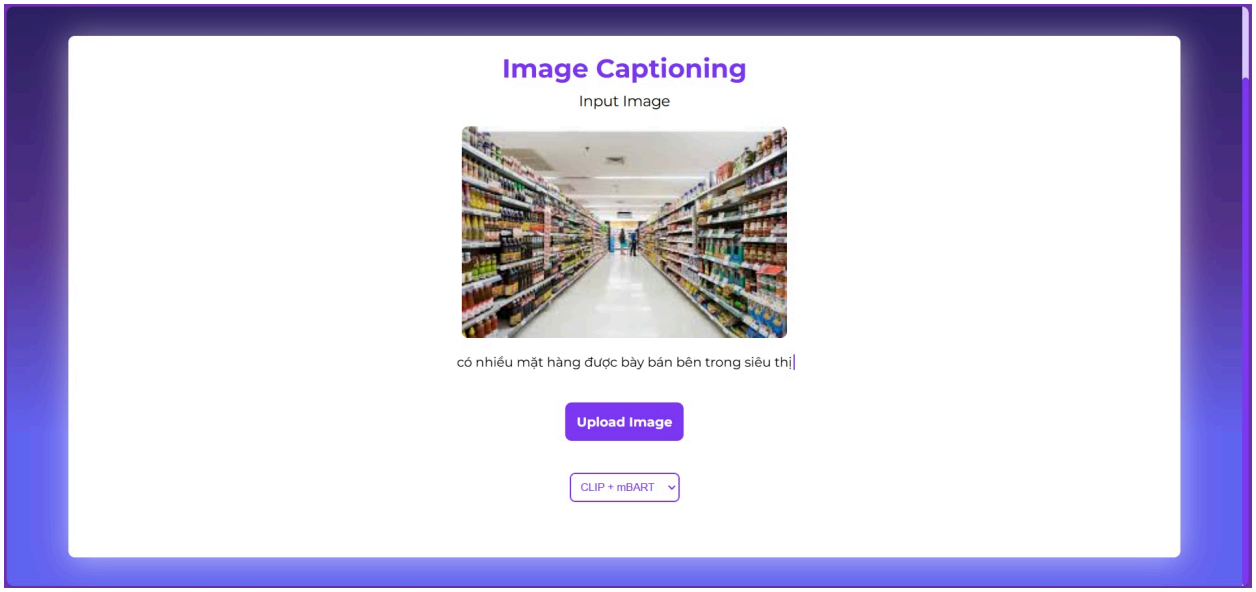
trước mà vẫn có thể sử dụng, thao tác dễ dàng - *"Chỉ cần upload ảnh, mọi thứ cứ để chúng tôi"*.



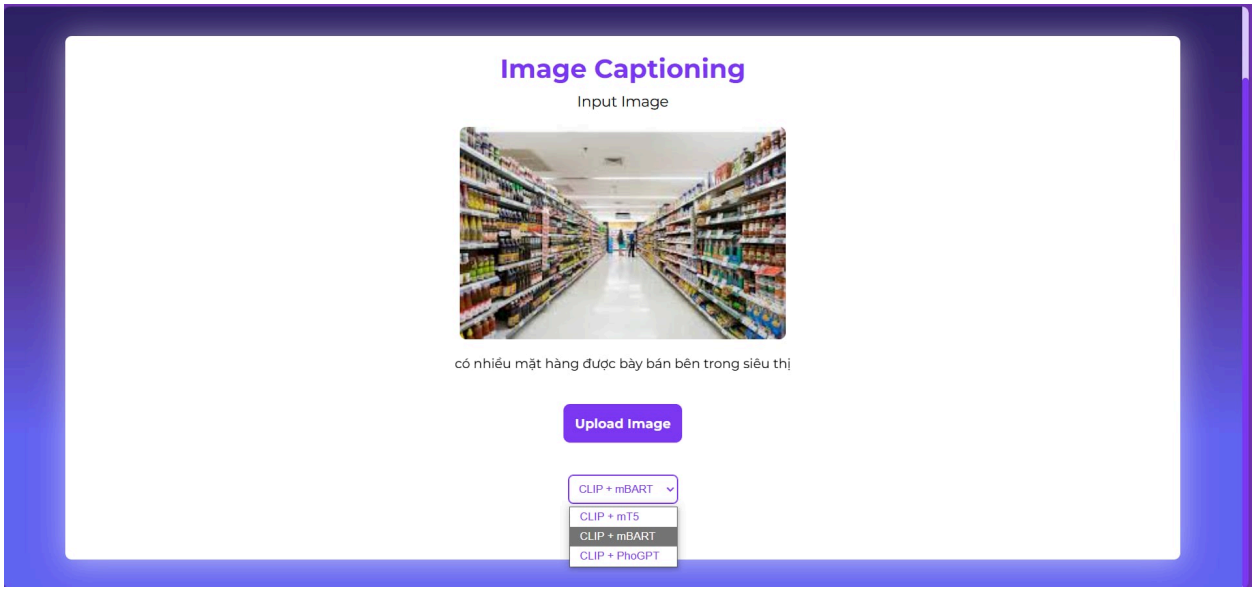
Hình 6.5 - Trang Upload



Hình 6.6 - Khi người dùng upload ảnh, hệ thống sẽ hiển thị màn hình loading cho đến khi bên phía backend trả caption



Hình 6.7 - Sau khi bên Backend trả caption về cho Frontend, kết quả sẽ được hiển thị lên cho người dùng



Hình 6.8 - Dropdown menu cho phép người dùng lựa chọn mô hình muốn sử dụng

Chương 7 – KẾT LUẬN

Kết quả thực nghiệm trên bài toán mô tả ảnh (image captioning) cho thấy hiệu suất của các mô hình phụ thuộc lớn vào chiến lược huấn luyện, cấu hình mô hình, và số lượng epoch.

- **KTVIC**, mô hình từ bài báo, đạt hiệu suất cao nhất trên các thước đo BLEU-1, BLEU-4, và CIDEr, chứng minh khả năng tạo mô tả chính xác và phù hợp với tham chiếu. Tuy nhiên, các mô hình khác như **mT5-large (freeze CLIP)**, **PhoGPT (3 epochs)** và **mBART (3 epochs)** cho thấy sự cạnh tranh với KTVIC ở các thước đo METEOR và ROUGE, chứng minh sự linh hoạt và tính tự nhiên trong các mô tả được tạo ra.
- Đối với **mT5**, việc "đóng băng" CLIP trong mô hình mT5-large giúp đạt hiệu suất tối ưu trên hầu hết các thước đo, trong khi việc "đóng băng" toàn bộ CLIP và MT5 (freeze CLIP+MT5) lại làm mô hình không học được. Điều này nhấn mạnh vai trò của chiến lược huấn luyện hợp lý.
- Với **PhoGPT**, phiên bản được huấn luyện trong 3 epochs đạt hiệu suất tốt nhất, trong khi việc huấn luyện quá lâu (15 hoặc 30 epochs) làm giảm hiệu quả do overfitting.
- Tương tự, **mBART** cũng cho thấy hiệu suất tốt hơn khi huấn luyện trong 3 epochs so với 10 epochs, với sự cải thiện đáng kể ở thước đo CIDEr, nhấn mạnh tầm quan trọng của việc tối ưu hóa số lượng epoch.

Tóm lại, các mô hình như mT5-large (freeze CLIP), PhoGPT (3 epochs) và mBART (3 epochs) nổi bật nhờ khả năng cân bằng giữa độ chính xác và tính tự nhiên, cạnh tranh tốt với baseline KTVIC. Kết quả cũng nhấn mạnh tầm quan trọng của việc lựa chọn số lượng epoch phù hợp và chiến lược huấn luyện hiệu quả để đạt được hiệu suất tối ưu trong bài toán mô tả ảnh.

TÀI LIỆU THAM KHẢO

[1] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," Feb. 26, 2021, *arXiv*: arXiv:2103.00020. doi: 10.48550/arXiv.2103.00020.

[2] H. Naveed *et al.*, "A Comprehensive Overview of Large Language Models," Oct. 17, 2024, *arXiv*: arXiv:2307.06435. doi: 10.48550/arXiv.2307.06435.

[3] D. Q. Nguyen, L. T. Nguyen, C. Tran, D. N. Nguyen, D. Phung, and H. Bui, "PhoGPT: Generative Pre-training for Vietnamese," Mar. 22, 2024, *arXiv*: arXiv:2311.02945. doi: 10.48550/arXiv.2311.02945.

[4] L. Xue *et al.*, "mT5: A massively multilingual pre-trained text-to-text transformer," Mar. 11, 2021, *arXiv*: arXiv:2010.11934. doi: 10.48550/arXiv.2010.11934.

[5] Y. Liu *et al.*, "Multilingual Denoising Pre-training for Neural Machine Translation," Jan. 23, 2020, *arXiv*: arXiv:2001.08210. doi: 10.48550/arXiv.2001.08210.

[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," Apr. 20, 2015, *arXiv*: arXiv:1411.4555. doi: 10.48550/arXiv.1411.4555.

[7] A. Vaswani *et al.*, "Attention Is All You Need," Aug. 02, 2023, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.

[8] K. Xu *et al.*, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," Apr. 19, 2016, *arXiv*: arXiv:1502.03044. doi: 10.48550/arXiv.1502.03044.

[9] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," Jul. 22, 2020, *arXiv*: arXiv:2005.14165. doi: 10.48550/arXiv.2005.14165.

[10] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A Simple and Performant Baseline for Vision and Language," Aug. 09, 2019, *arXiv*: arXiv:1908.03557. doi: 10.48550/arXiv.1908.03557.

[11] Y.-C. Chen *et al.*, "UNITER: UNiversal Image-TExt Representation Learning," Jul. 17, 2020, *arXiv*: arXiv:1909.11740. doi: 10.48550/arXiv.1909.11740.

[12] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," Feb. 21, 2015, *arXiv*: arXiv:1405.0312. doi: 10.48550/arXiv.1405.0312.

[13] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models," Sep. 19, 2016, *arXiv*: arXiv:1505.04870. doi: 10.48550/arXiv.1505.04870.

[14] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30K: Multilingual English-German Image Descriptions," May 02, 2016, *arXiv*: arXiv:1605.00459. doi: 10.48550/arXiv.1605.00459.

[15] Q. H. Lam, Q. D. Le, K. V. Nguyen, and N. L.-T. Nguyen, "UIT-ViIC: A Dataset for

the First Evaluation on Vietnamese Image Captioning," Feb. 01, 2020, *arXiv*: arXiv:2002.00175. doi: 10.48550/arXiv.2002.00175.

[16] T. T. Nguyen *et al.*, "vieCap4H-VLSP 2021: Vietnamese Image Captioning for Healthcare Domain using Swin Transformer and Attention-based LSTM," *VNU J. Sci. Comput. Sci. Commun. Eng.*, vol. 38, no. 2, Dec. 2022, doi: 10.25073/2588-1086/vnucsce.369.

[17] A.-C. Pham, V.-Q. Nguyen, T.-H. Vuong, and Q.-T. Ha, "KTVIC: A Vietnamese Image Captioning Dataset on the Life Domain," Jan. 16, 2024, *arXiv*: arXiv:2401.08100. doi: 10.48550/arXiv.2401.08100.

[18] N. L.-T. Nguyen, N. H. Nguyen, D. T. D. Vo, K. Q. Tran, and K. V. Nguyen, "EVJVQA Challenge: Multilingual Visual Question Answering," *J. Comput. Sci. Cybern.*, pp. 237–258, Sep. 2023, doi: 10.15625/1813-9663/18157.

[19] J. Wu *et al.*, "GRiT: A Generative Region-to-text Transformer for Object Understanding," Dec. 01, 2022, *arXiv*: arXiv:2212.00280. doi: 10.48550/arXiv.2212.00280.