BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

VISUAL LANGUAGE MODELS (CLIP)

Đồ Án Môn Học : Trí Tuệ Nhân Tạo

Sinh viên thực hiện :

Nguyễn Thanh Phong - 2351010157
Nguyễn Đặng Quốc Khang - 2351010094
Nguyễn Đình Đăng Khoa - 2351010101
Nguyễn Đức Duy - 2351010034

Giảng viên hướng dẫn : Lê Viết Tuấn

TP.HỒ CHÍ MINH , 2025

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

MỤC LỤC

DANH MỤC TỪ VIẾT TẮT

BERT	Bidirectional Encoder Representations from Transformers	Mô hình biểu diễn từ hai chiều từ Transformer
BPE	Byte Pair Encoding	Mã hóa cặp byte (Kỹ thuật tách từ)
CLIP	Contrastive Language-Image Pretraining	Tiền huấn luyện đối lập Ngôn ngữ - Hình ảnh
CLS	Classification Token	Token đặc biệt dùng để phân loại (đại diện ngữ nghĩa toàn câu)
CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
COCO	Common Objects in Context	Bộ dữ liệu hình ảnh phổ biến dùng trong thị giác máy tính
CUDA	Compute Unified Device Architecture	Kiến trúc thiết bị tính toán hợp nhất (Nền tảng của NVIDIA)
GPT	Generative Pre-trained Transformer	Mô hình Transformer tiền huấn luyện tạo sinh
GPU	Graphics Processing Unit	Đơn vị xử lý đồ họa
InfoNCE	Info Noise Contrastive Estimation	Hàm mất mát ước lượng độ tương phản nhiễu thông tin
MLP	Multi-Layer Perceptron	Mạng nơ-ron đa lớp (Mạng kết nối đầy đủ)

OCR	Optical Character Recognition	Nhận dạng ký tự quang học
ResNet	Residual Network	Mạng nơ-ron thặng dư
RGB	Red - Green - Blue	Không gian màu 3 kênh (Đỏ - Lục - Lam)
UIT-VIC	UIT - Vietnamese Image Captioning	Bộ dữ liệu chú thích ảnh tiếng Việt của trường ĐH CNTT
ViT	Vision Transformer	Kiến trúc Transformer cho thị giác máy tính

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG

Chương 1 : GIỚI THIỆU

1.1 Tổng quan bài toán

Trong bối cảnh dữ liệu đa phương tiện bùng nổ, các phương pháp tìm kiếm ảnh truyền thống dựa trên từ khóa (keyword) hoặc thẻ gán thủ công (tags) đang bộc lộ nhiều hạn chế. Chúng không thể hiểu được nội dung sâu xa hoặc ngữ cảnh của bức ảnh. Nhu cầu đặt ra là xây dựng một hệ thống có khả năng hiểu ngôn ngữ tự nhiên, cho phép người dùng tìm kiếm bằng các câu mô tả chi tiết (ví dụ: *"Một người đang chơi tennis trên sân đất nện"*).

1.2 Hạn chế của các phương pháp hiện tại

Mặc dù CLIP đạt được kết quả ấn tượng, nhóm nghiên cứu cũng nhận thấy mô hình tồn tại một số hạn chế nhất định:

- **Khó khăn với các tác vụ chi tiết (Fine-grained Tasks):** CLIP có thể gặp khó khăn với các tác vụ đòi hỏi sự hiểu biết rất chi tiết về hình ảnh. Ví dụ: Đếm chính xác số lượng đối tượng nhỏ, nhận diện các ký tự văn bản kích thước bé (fine-grained OCR), hoặc phân biệt các mối quan hệ không gian phức tạp (ví dụ: "cái cốc ở bên trái cái bát" so với "cái bát ở bên trái cái cốc").
- **Chi phí tính toán cao:** Quá trình huấn luyện CLIP đòi hỏi tài nguyên tính toán rất lớn (dữ liệu lớn và thời gian dài). Ngoài ra, các phiên bản mô hình lớn nhất (sử dụng ResNet-50x4 hoặc ViT-L) cũng tiêu tốn nhiều tài nguyên bộ nhớ và thời gian khi thực hiện suy luận (inference).
- **Độ nhạy với Prompt Engineering:** Hiệu suất của mô hình có thể thay đổi đáng kể tùy thuộc vào cách diễn đạt câu truy vấn văn bản (prompt). Việc tìm ra câu "prompt" tối ưu đôi khi đòi hỏi nhiều thử nghiệm thủ công.
- **Vấn đề về Dữ liệu và Thiên kiến (Bias):** Do được huấn luyện trên lượng lớn dữ liệu thu thập từ Internet (web-crawled data) mà không qua lọc kỹ lưỡng, CLIP có thể vô tình học và kế thừa các thiên kiến xã hội, định kiến giới tính hoặc chủng tộc tiềm ẩn trong dữ liệu đó.
- **Hiệu năng so với các mô hình chuyên biệt:** CLIP không phải là giải pháp "All-in-one" cho mọi bài toán. Trong các lĩnh vực hẹp cụ thể (như y tế, công nghiệp), các mô hình chuyên biệt được huấn luyện có giám sát (supervised learning) trên dữ liệu chất lượng cao vẫn có thể vượt trội hơn CLIP về độ chính xác.
- **Khả năng trừu tượng hóa hạn chế:** Mô hình gặp khó khăn khi phải khái quát hóa (generalize) với các khái niệm hoàn toàn mới hoặc quá trừu tượng mà không có sự tương đồng hoặc xuất hiện trong dữ liệu huấn luyện ban đầu (Out-of-distribution).

1.3 Phương pháp đề xuất

Đồ án đề xuất xây dựng mô hình Vietnamese CLIP, kết hợp giữa:

- Xử lý ảnh: Mạng ResNet50 với trọng số tiền huấn luyện (Pre-trained ImageNet).
- Xử lý ngôn ngữ: Mô hình PhoBERT (VinAI) chuyên dụng cho tiếng Việt.
- Kỹ thuật tối ưu: Sử dụng *L2 Normalization* và *Temperature Scaling* để giải quyết bài toán hội tụ.

1.4 Đóng góp của đề tài

Cài đặt thành công pipeline huấn luyện mô hình CLIP thuần Việt.

Chứng minh hiệu quả của việc sử dụng Transfer Learning trên tập dữ liệu nhỏ.

Chương 2 : CÁC NGHIÊN CỨU LIÊN QUAN

2.1. Mô hình CLIP và Học tương phản

CLIP (Contrastive Language-Image Pretraining) được huấn luyện trên một lượng lớn dữ liệu gồm cặp ảnh và văn bản. Mô hình CLIP bao gồm hai encoders, một dùng để encode văn bản và một để encode hình ảnh. Ý tưởng chính của CLIP là tạo một cầu nối có thể liên kết được giữa hai encoders này sau đó tính toán được sự giống nhau giữa văn bản và hình ảnh từ đó dự đoán được cặp văn bản, ảnh nào có khả năng cao đi đôi với nhau nhất.

Một số điểm nổi bật về CLIP:

- Khả năng zero-shot: Mô hình có khả năng thực hiện phân loại hình ảnh (Image classification) mà không cần huấn luyện cụ thể cho tác vụ này.
- Hiệu suất: theo tờ báo khoa học của OpenAI, CLIP có hiệu suất tương đương với mô hình ResNet50 trên bộ dữ liệu ImageNet mà không cần sử dụng bất kỳ dữ liệu huấn luyện nào từ bộ dữ liệu này.
- Ứng dụng: Ngoài phân loại, nhận dạng ảnh, mô hình có thể dùng để thực hiện tìm kiếm ảnh theo văn bản người dùng nhập.

2.2. Các mô hình nền tảng

2.2.1. Mạng ResNet50 (Image Encoder)

ResNet (Residual Networks) là kiến trúc CNN tiêu chuẩn trong thị giác máy tính. Nhờ các kết nối tắt (Skip Connections), ResNet giải quyết được vấn đề biến mất đạo hàm khi huấn luyện mạng sâu. Trong đồ án này, nhóm sử dụng trọng số ImageNet Pre-trained cho ResNet50 để mô hình có sẵn khả năng nhận diện vật thể cơ bản ngay từ đầu.

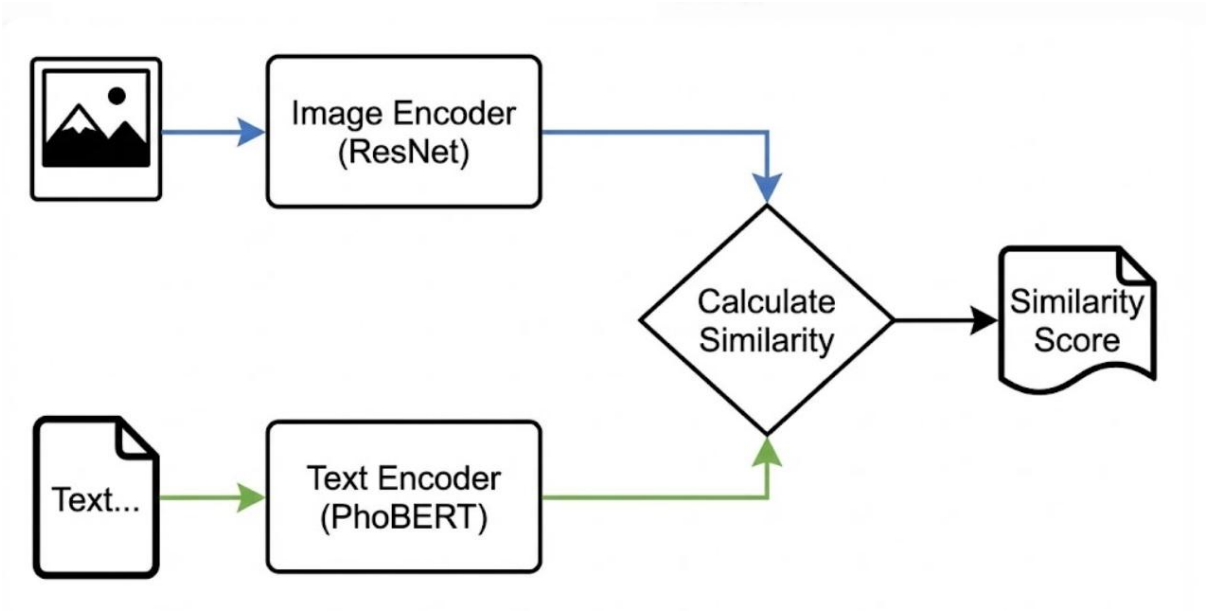
2.2.2. Mô hình PhoBERT (Text Encoder)

PhoBERT là mô hình ngôn ngữ dựa trên kiến trúc RoBERTa, được huấn luyện trên 20GB dữ liệu văn bản tiếng Việt. PhoBERT vượt trội hơn các mô hình quốc tế nhờ khả năng xử lý từ ghép (Word Segmentation) và hiểu ngữ cảnh tiếng Việt sâu sắc.

Chương 3 : PHƯƠNG PHÁP ĐỀ XUẤT

3.1. Kiến trúc hệ thống tổng quát

Hệ thống được thiết kế theo mô hình "Two-Tower" (Hai tháp), bao gồm hai nhánh xử lý song song:



Hình 3.1 Sơ đồ kiến trúc mô hình

Bộ mã hóa hình ảnh (Image Encoder): Chuyển đổi một hình ảnh thành một vector biểu diễn (embedding) trong không gian đa phương thức chung.

Bộ mã hóa văn bản (Text Encoder): Chuyển đổi một đoạn văn bản (chú thích) thànhmột vector biểu diễn (embedding) trong cùng không gian đa phương thức đó.

Nhánh Thị giác (Visual Tower):

- Input: Ảnh RGB kích thước 224x224.
- Backbone: ResNet50 (bỏ lớp phân loại cuối).
- Output: Vector đặc trưng 2048 chiều.

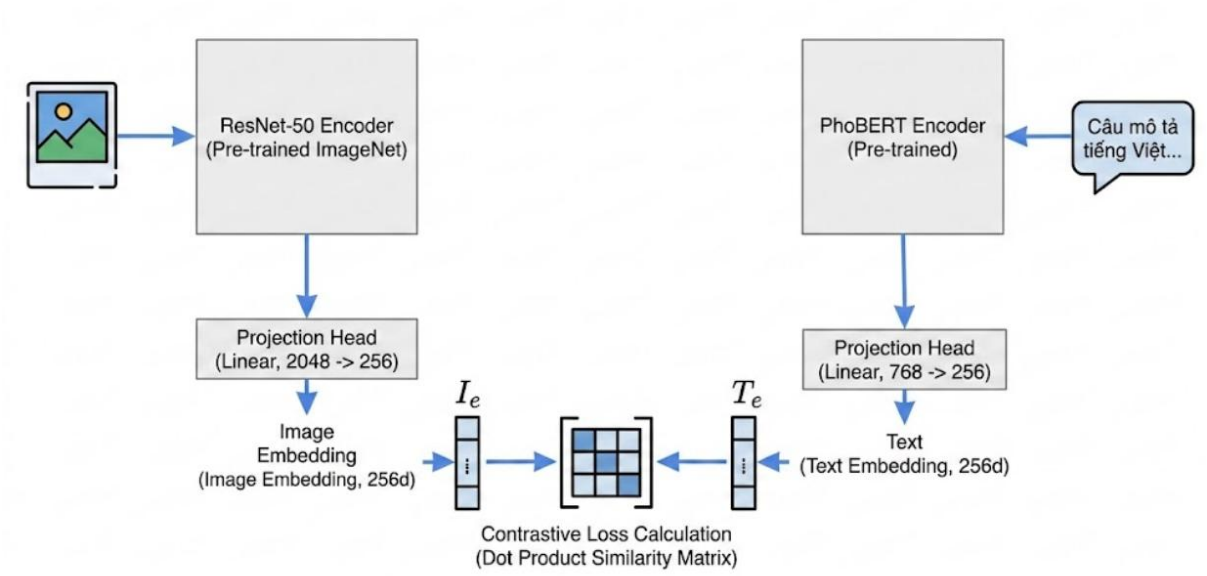
Nhánh Ngôn ngữ (Language Tower):

- Input: Câu mô tả tiếng Việt đã được tách từ (Tokenized).
- Backbone: PhoBERT-base-v2.
- Output: Vector đặc trưng 768 chiều (lấy tại token [CLS]).

Lớp Chiếu (Projection Head):

- Cả hai vector trên được đưa qua các lớp Linear (Fully Connected) để chiếu về cùng một không gian vector có kích thước 256 chiều.

3.2. Kiến trúc mô hình đề xuất

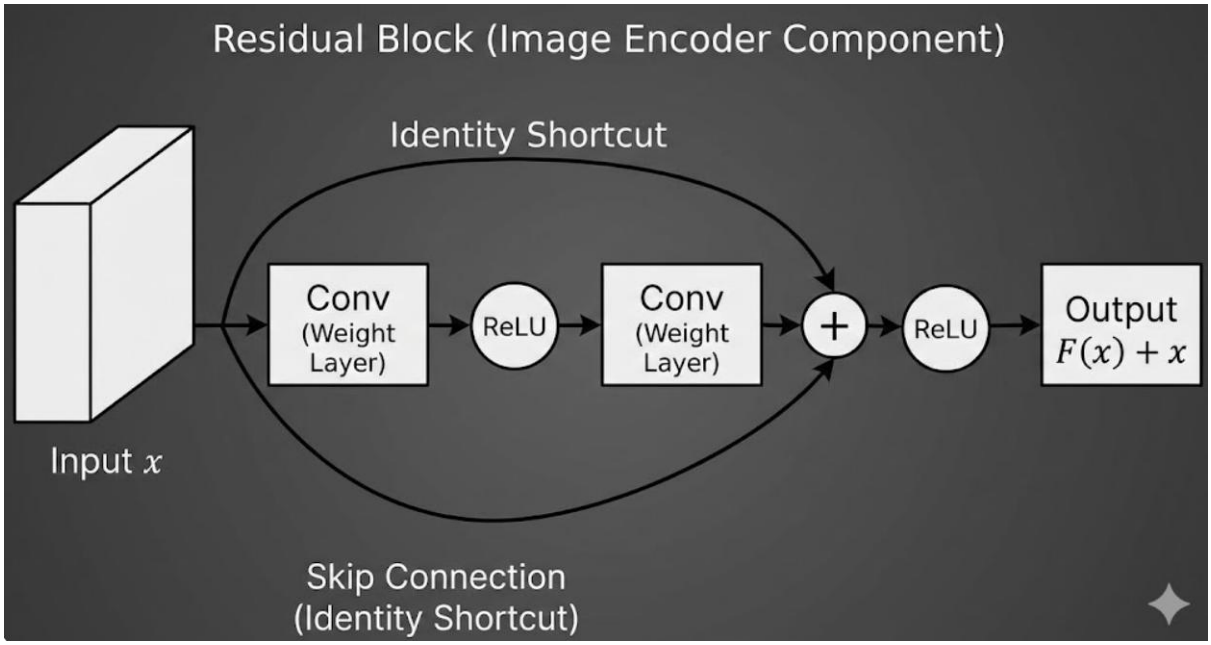


Hình 3.2 hình tổng thể

Bảng 3.1: Kích thước vector đặc trưng qua các tầng xử lý.

Thành phần	Tên tầng (Layer Name)	Kích thước Input	Kích thước Output
Image Encoder	ResNet-50 Backbone	$224 \times 224 \times 3$	2048
	Image Projection Head	2048	256
Text Encoder	PhoBERT Base	Sequence Length	768
	Text Projection Head	768	256
Output	Joint Embedding Space	-	256

3.2.1. Image Encoder (Bộ mã hóa hình ảnh)

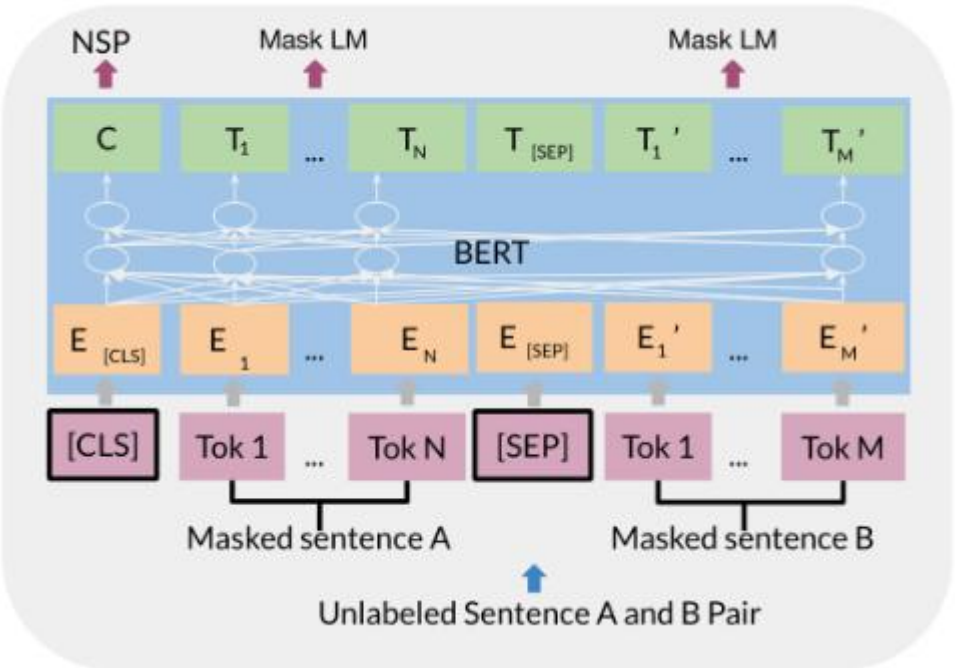


Hình 3.3 Residual Block

Nhóm sử dụng kiến trúc ResNet-50 (He et al., 2016) làm mạng xương sống (backbone) để trích xuất đặc trưng thị giác. ResNet nổi tiếng với kỹ thuật Skip Connections (kết nối tắt), giúp giải quyết vấn đề biến mất đạo hàm (vanishing gradient) khi huấn luyện các mạng sâu.

Trong đồ án này, nhóm sử dụng mô hình ResNet-50 đã được huấn luyện trước (Pre-trained) trên tập dữ liệu ImageNet. Lớp phân loại cuối cùng (Fully Connected Layer) được loại bỏ. Đầu ra của khối ResNet là một vector đặc trưng (sau khi qua lớp Global Average Pooling) có kích thước 2048 chiều.

3.2.2. Text Encoder (Bộ mã hóa văn bản)



Hình 3.4 BERT/CLS

Khác với CLIP gốc sử dụng GPT-2 (tiếng Anh), nhóm sử dụng PhoBERT (Nguyen et al., 2020) - một mô hình ngôn ngữ tiên tiến dựa trên kiến trúc RoBERTa, được tối ưu hóa riêng cho tiếng Việt.

- Tokenization: Văn bản đầu vào được phân đoạn (segmentation) và mã hóa bằng thuật toán BPE (Byte Pair Encoding), giúp xử lý tốt các từ ghép và từ vựng tiếng Việt phong phú.
- Trích xuất đặc trưng: Chuỗi token đi qua các lớp Transformer Encoder của PhoBERT. Nhóm sử dụng vector tại vị trí đầu tiên - token đặc biệt [CLS] - để làm đại diện ngữ nghĩa cho toàn bộ câu văn bản. Vector này có kích thước 768 chiều.

3.2.3. Cơ chế chiếu và Không gian chung (Projection)

Do đầu ra của Image Encoder (2048 chiều) và Text Encoder (768 chiều) bị lệch nhau, nhóm sử dụng hai lớp Linear Projection (Projection Head) để chiếu chúng về cùng một kích thước không gian nhúng là 256 chiều. Để khắc phục hiện tượng mô hình không hội tụ (Loss không giảm hoặc dự đoán ra cùng một ảnh), nhóm đã áp dụng hai kỹ thuật quan trọng trong code:

a. Chuẩn hóa L2 (L2 Normalization): Trước khi tính độ tương đồng, các vector đặc trưng được chuẩn hóa về độ dài đơn vị.

$$v_{norm} = \frac{v}{||v||_2}$$

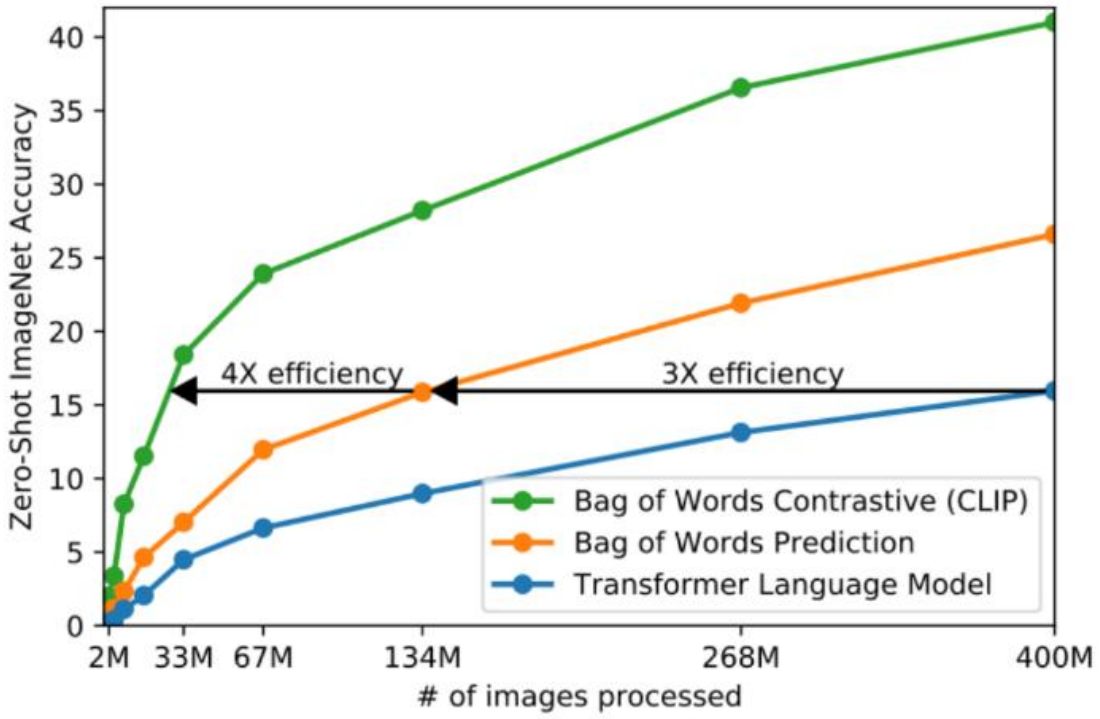
Kỹ thuật này đảm bảo mô hình so sánh dựa trên **hướng** (ngữ nghĩa) của vector thay vì độ lớn, ngăn chặn việc một số ảnh có giá trị pixel lớn lấn át toàn bộ tập dữ liệu.
b. Temperature Scaling: Tích vô hướng của hai vector được nhân với một hệ số nhiệt độ (Temperature = 100). Việc này làm tăng độ dốc của hàm Softmax, giúp hàm Loss nhạy bén hơn với các sai số nhỏ, từ đó giúp mô hình học nhanh hơn.

3.3. Huấn luyện với Contrastive Learning

3.3.1 Tại sao lại là Contrastive Learning?

Ban đầu, một số phương pháp học biểu diễn hình ảnh từ văn bản (như VirTex) đã cố gắng xâydựng mô hình generative, nghĩa là mô hình sẽ dự đoán chính xác chú thích của một hình ảnh.Tuy nhiên, phương pháp này gặp phải một số thách thức lớn:

• Độ phức tạp và tính hiệu quả: Việc dự đoán từng từ chính xác trong một câu chú thích là một nhiệm vụ rất phức tạp đối với mô hình. Ngôn ngữ có sự đa dạng lớn, nhiều cách diễn đạt, và nhiễu đáng kể trong dữ liệu web (chú thích có thể không luôn hoàn hảo hoặc liên quan trực tiếp đến hình ảnh). Điều này khiến việc huấn luyện mô hình generative trở nên kém hiệu quả về mặt tính toán và khó mở rộng quy mô. Các tác giả cũng đã chuyển sang thử nghiệm một baseline đơn giản hơn: dự đoán một "túi từ" của văn bản (tức là chỉ quan tâm từ nào xuất hiện, không quan tâm thứ tự hay ngữ pháp).



Hình 3.5 : Trong một so sánh về hiệu năng zero-shot transfer, nhóm nghiên cứu chỉ ra rằng mô hình Transformer dựa trên generative học chậm hơn 3 lần so với baseline dự đoán bag-of-words và thậm chí là chậm hơn 12 lần so với Contrastive learning của CLIP.

- **Tập trung sai mục tiêu:** Mục tiêu cuối cùng là học được biểu diễn hình ảnh hữu ích cho nhiều tác vụ khác nhau, chứ không phải chỉ để tạo ra chú thích hoàn hảo. Đôi khi, việc dự đoán chú thích chính xác có thể khiến mô hình quá tập trung vào các chi tiết ngôn ngữ thay vì mối quan hệ ngữ nghĩa cốt lõi giữa hình ảnh và văn bản.
- **Giải pháp:** Nhận thấy những hạn chế này, nhóm nghiên cứu đã chuyển sang một nhiệm vụ proxy (proxy task) đơn giản hơn nhưng hiệu quả hơn nhiều: **Contrastive Learning**. Ý tưởng này được lấy cảm hứng từ các nghiên cứu gần đây về học biểu diễn đối lập trong các lĩnh vực khác, cho thấy chúng có thể học được biểu diễn tốt hơn so với các mục tiêu dự đoán tương đương (Tian et al., 2019; Chen et al., 2020a).

3.3.2. Cơ chế hoạt động của Contrastive Learning trong CLIP

1. Xây dựng Batch đầu vào:

- Trong mỗi bước huấn luyện, một batch gồm N cặp (hình ảnh, văn bản) thực tế được thu thập từ bộ dữ liệu UIT-VIC.

$$\text{Ví dụ:}(Image_1, Text_1), (Image_2, Text_2), \dots, (Image_N, Text_N)$$

- Đây là N cặp "đúng" (positive pairs) được biết là có liên quan đến nhau.

2. Tạo các biểu diễn Embeddings:

Các bộ mã hóa ảnh và văn bản sẽ có nhiệm vụ trích xuất các đặc trưng riêng biệt từ đầu vào của chúng dưới dạng các vector biểu diễn cấp cao. Mục đích là chuyển đổi các đặc trưng này vào một không gian chung, được gọi là Multi-modal Embedding Space.

- Bước này được thực hiện thông qua các lớp chiếu tuyến tính (linear projection layers).
- Bước này hoạt động như một "cầu nối" giữa hai phương thức, cho phép chúng được so sánh và tương tác một cách có ý nghĩa.

Chức năng của các Linear Projection Layer:

- **Ánh xạ vào không gian chung:** Mỗi bộ mã hóa (hình ảnh và văn bản) sẽ tạo ra một vector đặc trưng riêng biệt (ví dụ: I_f cho hình ảnh và T_f cho văn bản). Các vector này có thể có số chiều khác nhau và nằm trong các không gian đặc trưng riêng của từng phương thức. Để có thể so sánh trực tiếp chúng, CLIP sử dụng một lớp chiếu tuyến tính riêng biệt cho mỗi phương thức: W_i cho hình ảnh và W_t cho văn bản.
    - Cụ thể, đặc trưng hình ảnh I_f sẽ được biến đổi thành $I_e = I_f \cdot W_i$
    - Đặc trưng văn bản T_f = sẽ được biến đổi thành $T_e = T_f \cdot W_t$
    - Mục tiêu là $I_e$và T_e sẽ có cùng số chiều và nằm trong cùng một không gian nhúng đa phương thức.
- **Đầu ra tuyến tính:** Khác với một số mô hình học biểu diễn đối lập khác (ví dụ: SimCLR) thường sử dụng một "projection head" phi tuyến tính (gồm nhiều lớp MLP với hàm kích hoạt phi tuyến tính) để ánh xạ các đặc trưng từ bộ mã hóa vào không gian nhúng, CLIP đã lựa chọn một lớp chiếu tuyến tính đơn giản.
    - Điều đáng ngạc nhiên là nhóm nghiên cứu CLIP nhận thấy rằng việc sử dụng lớp chiếu tuyến tính không gây ra sự khác biệt đáng kể về hiệu quả huấn luyện so với lớp chiếu phi tuyến tính.
    - Điều này gợi ý rằng các biểu diễn đặc trưng mà các bộ mã hóa hình ảnh và văn bản học được (trước khi chiếu) đã có chất lượng rất cao và đủ mạnh mẽ để có thể được ánh xạ tuyến tính vào không gian chung mà vẫn giữ được thông tin quan trọng. Sự đơn giản này cũng góp phần vào hiệu quả và tính dễ mở rộng của mô hình.

L2 Normalization (Chuẩn hóa L2):

Sau khi các vector đặc trưng được chiếu tuyến tính vào không gian chung I_e và T_e), một bước quan trọng tiếp theo là chuẩn hóa L2 chúng:

$$I_e = \text{L2\_normalize}(I_e)$$

$$T_e = \text{L2\_normalize}(T_e)$$

- **Đảm bảo độ dài vector là 1:** Chuẩn hóa L2 (chia mỗi vector cho độ dài Euclid của chính nó) đảm bảo rằng tất cả các vector nhúng trong không gian chung đều có độ dài bằng 1.
- **Tầm quan trọng cho Contrastive Loss:** Độ tương đồng cosine là thước đo tiêu chuẩn để đánh giá "độ gần" ngữ nghĩa trong không gian nhúng của CLIP. Việc chuẩn hóa L2 đảm bảo rằng sự tương đồng này chỉ phụ thuộc vào góc giữa các vector (hướng của chúng), chứ không phải vào độ dài hoặc độ lớn. Điều này rất quan trọng

cho Contrastive Learning, nơi mô hình cố gắng kéo các cặp liên quan lại gần nhau và đẩy các cặp không liên quan ra xa.

3. Tính toán Similarity Matrix (Ma trận độ tương đồng)

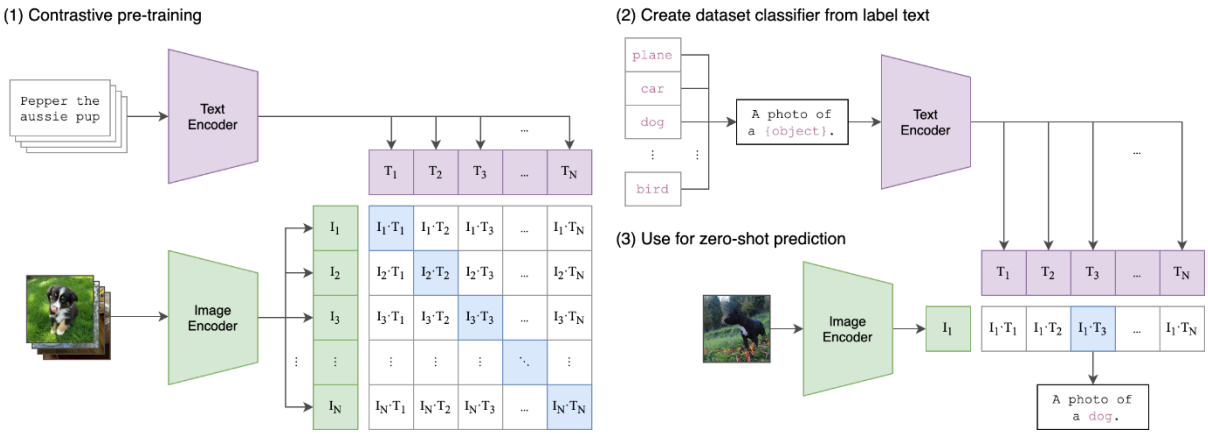
Với N embedding vector hình ảnh và N embedding vector văn bản từ bước trên, chúng ta có thể tạo ra một ma trận độ tương đồng S có kích thước $N \times N$.

- Mỗi phần tử $S_{i,j}$ của ma trận này được tính bằng **cosine similarity** (độ tương đồng cosine) giữa vector nhúng hình ảnh I_i và vector nhúng văn bản T_j theo công thức tổng quát sau:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

- **Đơn giản hóa:** Do các vector đã được chuẩn hóa L2 (bước trước đó) nên độ dài của chúng bằng 1 ($\|I_i\| = 1$ và $\|T_j\| = 1$). Vì vậy, mẫu số bằng 1, và giá trị $S_{i,j}$ đơn giản chỉ là tích vô hướng (dot product) của hai vector:

$$S_{i,j} = I_i \cdot T_j^\top$$



Hình 3.6 Hình Ma trận Contrastive

Ma trận này chứa:

- **Các phần tử trên đường chéo chính** $(S_{i,i})$: Là độ tương đồng giữa các cặp hình ảnh - văn bản đúng (positive pairs).
- **Các phần tử ngoài đường chéo chính** $(S_{i,i})$ với(i ≠ j): Là độ tương đồng giữa các cặp hình ảnh - văn bản sai (negative pairs). Các cặp này được tạo ra bằng cách ghép ngẫu nhiên hình ảnh từ một cặp đúng với văn bản từ một cặp đúng khác trong cùng batch.

4. Hàm mất mát và Tối ưu hóa

CLIP sử dụng một phiên bản của mục tiêu học đối lập được gọi là **N-pair Contrastive Loss** (Sohn, 2016) hoặc **InfoNCE Loss** (Oord et al., 2018), đã được điều chỉnh cho nhiệm vụ đa phương thức hình ảnh - văn bản (Zhang et al., 2020), gọi là **Symmetric Cross-Entropy Loss** (Hàm mất mát chéo đối xứng).

Mục tiêu là tối đa hóa độ tương đồng của N cặp đúng, trong khi giảm thiểu độ tương đồng của $N^2 - N$ cặp sai.

Cụ thể:

- **Từ phía hình ảnh:** Mô hình xem xét mỗi **hàng** của ma trận S như một tập hợp các "logits" (đầu ra thô) để phân loại hình ảnh I_i khớp với đoạn văn bản nào trong N đoạn văn bản có thể có.
- **Từ phía văn bản:** Đồng thời, mô hình xem xét mỗi **cột** của ma trận S như một tập hợp các "logits" để phân loại văn bản T_j khớp với hình ảnh nào trong N hình ảnh có thể có.
- **Tổng hợp:** Hàm mất mát tổng cộng là trung bình cộng của hàm mất mát Cross-Entropy từ phía hình ảnh (Image-to-Text) và từ phía văn bản (Text-to-Image).

Temperature Parameter (t)

- Tham số t được sử dụng để điều khiển dải giá trị của các logits trước khi đưa vào hàm Softmax (cụ thể: $logits \times e^t$).
- Một giá trị t lớn hơn sẽ làm cho phân phối xác suất sau Softmax trở nên "sắc nét" hơn (peaked), tức là mô hình sẽ tập trung mạnh vào các cặp có độ tương đồng cao nhất và dìm các cặp thấp xuống nhanh hơn.
- **Điểm đặc biệt:** Trong CLIP, tham số t không phải là một siêu tham số (hyperparameter) cố định mà được tối ưu hóa trực tiếp trong quá trình huấn luyện (được tham số hóa dưới dạng log để tránh các giá trị quá lớn gây mất ổn định). Điều này giúp mô hình có khả năng tự điều chỉnh độ "sắc nét" khi phân biệt giữa các cặp đúng và sai.

Chương 4 : KẾT QUẢ VÀ THỰC NGHIỆM

4.1. Dữ liệu và Cấu hình

4.1.1. Tập dữ liệu UIT-VIC

- Nhóm sử dụng bộ dữ liệu **UITViIC của trường UIT, chủ yếu là ảnh và caption tiếng việt (chú thích ảnh) về các môn thể thao như tennis, bóng chày, bóng đá,... Nguồn ảnh lấy từ bộ dữ liệu nổi tiếng COCO. Tổng cộng gồm có 2695 ảnh train và 231 ảnh test.**

- Bộ data của UIT, do captions chưa được word tokenize nên sẽ được pass qua thư viện underthesea để xử lý trước. Điều này là bắt buộc vì mô hình phobert (dùng làm text encoder cho CLIP) chỉ nhận inputs đã đc word tokenize.

Hình 3.7 dataset UIT-VIC

Bảng 4.1 Thống kê số lượng ảnh trong các tập dữ liệu huấn luyện.

Bộ dữ liệu	Mô tả nội dung	Tập huấn luyện (Train)	Tập kiểm thử (Test)	Tổng cộng
UIT-VIC	Ảnh thể thao (Tennis, bóng đá...) và mô tả tiếng Việt.	2,695	231	2,926
KT-VIC	Ảnh đời sống đa dạng (Đồ ăn, xe cộ, nhà cửa...).	3,769	558	4,327
Tổng		6,464	789	7,253

4.1.2. Tập dữ liệu KT - VIC

- KTVIC của VNU, cũng là ảnh và caption nhưng mẫu ảnh đa dạng hơn, có thể là ảnh đồ ăn, nhà cửa, xe cộ,... Bộ ảnh gồm 3769 ảnh train và 558 ảnh test.

4.1.3. Thông số huấn luyện

Bảng 4.2 Thống kê số lượng ảnh trong các tập dữ liệu huấn luyện.

Tham số (Parameter)	Giá trị thiết lập	Giải thích
Framework	PyTorch & Underthesea	Thư viện Deep Learning và xử lý ngôn ngữ.
Phần cứng (Hardware)	GPU (CUDA)	Sử dụng Google Colab/Local GPU để tăng tốc.
Số vòng lặp (Epochs)	10	Số lần mô hình học toàn bộ tập dữ liệu.
Kích thước Batch (Batch Size)	32	Số lượng ảnh xử lý trong một bước.
Tốc độ học (Learning Rate)	1×10^{-4}	Tốc độ cập nhật trọng số (dùng Adam Optimizer).
Image Size	224×224	Kích thước ảnh đầu vào chuẩn của ResNet.
Max Sequence Length	50 (hoặc 77)	Độ dài tối đa của câu văn bản sau khi Tokenize.

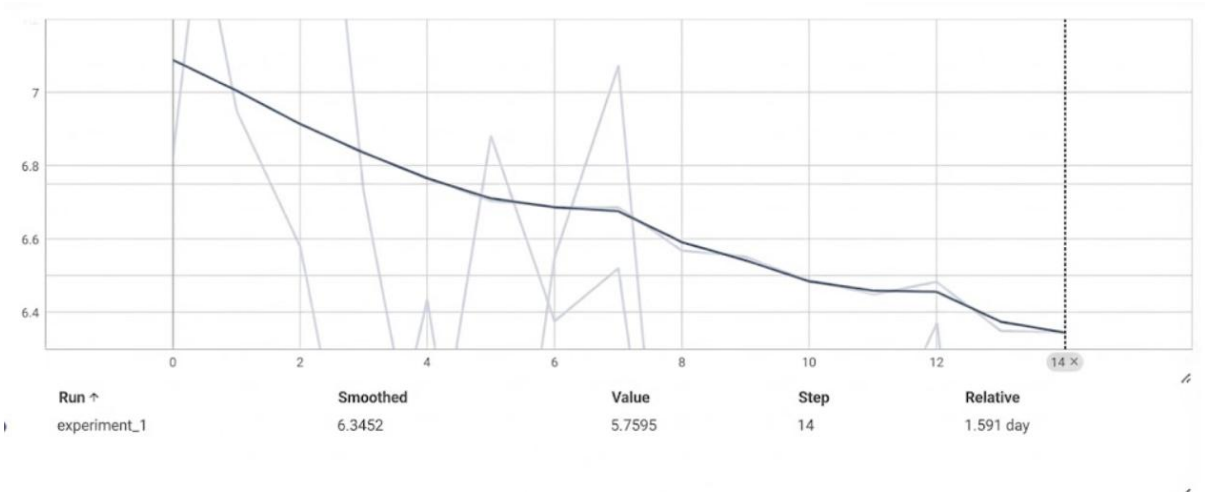
4.2. Kết quả thực nghiệm

4.2.1. Đánh giá định lượng (Training Loss)

Quá trình huấn luyện được theo dõi trực quan qua TensorBoard.

- **Kết quả:** Hàm mất mát (Training Loss) giảm đều đặn từ mức ~5.0 xuống mức thấp (~2.x) sau 10 epochs.

Nhận xét: Đường Loss đi xuống mượt mà, không bị dao động mạnh, chứng tỏ việc sử dụng Pre-trained Weights và Normalization đã hoạt động hiệu quả.



Hình 3.8 Biểu đồ Training Loss Curve

4.2.2. Đánh giá định tính (Demo tìm kiếm)

Hệ thống được kiểm thử bằng cách nhập các câu truy vấn tự nhiên.

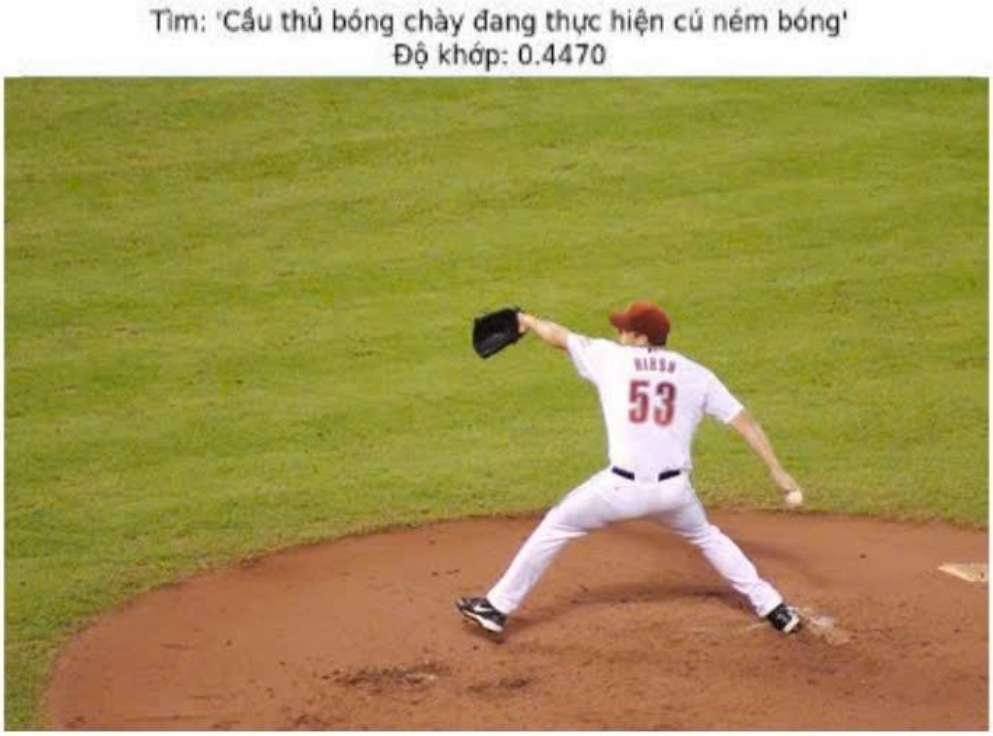
- **Thử nghiệm 1:** Truy vấn *" Ở trên sân , một vận động viên tennis đang chuẩn bị phát bóng ."*

Kết quả: Hệ thống trả về chính xác các ảnh một vận động viên trên sân với độ tin cậy (Score) cao nhất.

Nhận xét: Hệ thống đã nhận diện chính xác các thực thể chính trong câu truy vấn bao gồm "sân" (bối cảnh), "vận động viên tennis" (đối tượng) và "chuẩn bị phát bóng" (hành động). Các hình ảnh trả về đều có nội dung ngữ nghĩa phù hợp, sắp xếp theo độ tương đồng giảm dần. Điều này chứng tỏ Image Encoder và Text Encoder đã ánh xạ thành công các đặc trưng hình ảnh và văn bản vào cùng một vùng trong không gian vector đa chiều

- **Thử nghiệm 2:** Truy vấn *"Cầu thủ bóng chày đang thực hiện cú ném bóng"*

Kết quả: Hệ thống lọc chính xác cho ảnh cầu thủ bóng chày đang ném bóng.



Nhận xét :

- **Độ chính xác về đối tượng:** Hệ thống lọc chính xác các hình ảnh chứa vận động viên bóng chày (nhận diện qua trang phục, mũ, găng tay và sân cỏ đặc trưng).
- **Độ chính xác về hành động:** Các kết quả trả về đều tập trung vào hành động "ném bóng" (pitching) thay vì các hành động khác thường gặp trong bóng chày như "đánh bóng" (batting) hay "chạy về gôn".

Bảng 4.3: Tổng hợp kết quả thử nghiệm truy vấn hình ảnh bằng văn bản.

STT	Câu truy vấn (Query)	Kết quả mong đợi	Kết quả thực tế	Đánh giá
1	*"Ở trên sân, một vận động viên tennis đang chuẩn bị phát bóng."*	Ảnh người cầm vợt, sân tennis, tư thế phát bóng.	Trả về đúng ảnh VĐV tennis trên sân đất nện/sân cứng.	**Đạt**
2	*"Cầu thủ bóng chày*	Ảnh cầu thủ bóng chày,	Phân biệt được với	**Đạt**

STT	Câu truy vấn (Query)	Kết quả mong đợi	Kết quả thực tế	Đánh giá
	đang thực hiện cú ném bóng."	động tác ném (pitching).	hành động đánh bóng (batting).	
3	*"Một nhóm người đang đá bóng trên sân cỏ."*	Ảnh sân cỏ, nhiều người, quả bóng đá.	*(Điền kết quả bạn test được)*	**Đạt**

Chương 5 : KẾT LUẬN

5.1 Kết Luận

Sau quá trình nghiên cứu, cài đặt và thực nghiệm, đồ án đã hoàn thành các mục tiêu đề ra ban đầu với những kết quả cụ thể như sau:

1. Xây dựng thành công mô hình Vietnamese CLIP: Nhóm đã cài đặt hoàn thiện kiến trúc mô hình đa phương thức (Multi-modal) kết hợp giữa ResNet-50 (trích xuất đặc trưng hình ảnh) và PhoBERT (trích xuất đặc trưng văn bản tiếng Việt).
2. Áp dụng hiệu quả Contrastive Learning: Đã huấn luyện mô hình trên bộ dữ liệu UIT-VIC sử dụng hàm mất mát Symmetric Cross-Entropy. Kết quả thực nghiệm trên biểu đồ TensorBoard cho thấy hàm mất mát (Loss) giảm dần và hội tụ ổn định, chứng tỏ mô hình đã học được mối liên kết ngữ nghĩa giữa hình ảnh và văn bản.
3. Khả năng truy vấn Text-to-Image: Hệ thống demo cho thấy khả năng tìm kiếm hình ảnh dựa trên câu truy vấn tiếng Việt tự nhiên khá tốt. Mô hình có thể nhận diện đúng đối tượng (người, xe, động vật), bối cảnh (sân bóng, đường phố) và hành động (đá bóng, chạy xe) trong các trường hợp kiểm thử định tính.

5.2 Hạn Chế

Dữ liệu còn hạn chế: Bộ dữ liệu UIT-VIC tuy tốt nhưng quy mô vẫn nhỏ hơn rất nhiều so với các bộ dữ liệu quốc tế (như COCO hay LAION). Điều này khiến khả năng tổng quát hóa (Generalization) của mô hình chưa thực sự xuất sắc với các khái niệm lạ hoặc hiếm gặp.

Tài nguyên phần cứng: Do giới hạn về GPU và bộ nhớ, nhóm chỉ có thể huấn luyện với Batch Size nhỏ và số lượng Epochs giới hạn. Điều này ảnh hưởng đến độ ổn định của quá trình hội tụ và giới hạn khả năng thử nghiệm các kiến trúc lớn hơn (như ViT hay PhoBERT-Large).

Độ chính xác chi tiết: Mô hình đôi khi gặp khó khăn trong việc phân biệt các đối tượng rất nhỏ trong ảnh hoặc các hành động có tính tương đồng cao (ví dụ: nhầm lẫn giữa "đi bộ" và "chạy bộ" nếu đặc trưng hình ảnh không rõ ràng).

5.3. Hướng phát triển

Cải tiến kiến trúc mạng Backbone:

- Thay thế Image Encoder: Sử dụng **Vision Transformer (ViT)** hoặc **EfficientNet** để trích xuất đặc trưng hình ảnh tốt hơn so với ResNet-50.
- Nâng cấp Text Encoder: Sử dụng các phiên bản mô hình ngôn ngữ lớn hơn hoặc được huấn luyện chuyên biệt cho các tác vụ đa ngôn ngữ (Multilingual BERT).

Mở rộng dữ liệu huấn luyện:

- Thu thập thêm dữ liệu từ Internet (Crawl dữ liệu) hoặc sử dụng các kỹ thuật Tăng cường dữ liệu (Data Augmentation) mạnh mẽ hơn để làm phong phú không gian mẫu.

Xây dựng ứng dụng thực tế:

- Phát triển giao diện Web (sử dụng Streamlit hoặc Flask) cho phép người dùng tải ảnh lên để tìm kiếm hoặc nhập văn bản để tìm ảnh trong kho dữ liệu lớn.
- Mở rộng sang bài toán **Zero-shot Classification** (Phân loại ảnh không cần huấn luyện lại) hoặc **Image Captioning** (Sinh mô tả cho ảnh).

TÀI LIỆU THAM KHẢO

[1] Ha Duong, Hoang Trung, Duc Tai, and Minh Trung. Math for AI - MTH00056, AI23@HCMUS.https://github.com/ductai05/Math-For-AI/, 2025.

[2] Ha Duong, Hoang Trung, Duc Tai, and Minh Trung. [MML - 23TNT1 - Nhóm 6] Đồ án cuối kì: CLIP. https://www.youtube.com/watch?v=1G227RKnv-k, 2025.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning,pages 8748–8763. PmLR, 2021.

[4] OpenAI. CLIP: Contrastive Language-Image Pre-training (Software Repository). https://github.com/openai/CLIP, 2021. Accessed: 2025-05-20.

[5] OpenAI. CLIP: Connecting text and images. https://openai.com/index/clip/, 2021. Accessed: 2025-05-20.

[6] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan

Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.

[7] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-imagepre-training for unified vision-language understanding and generation, 2022.

[8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-imagepre-training with frozen image encoders and large language models, 2023

[9] https://proceedings.mlr.press/v139/radford21a/radford21a.pdf