

**BỘ GIÁO DỤC VÀ ĐÀO TẠO  
ĐẠI HỌC CẦN THƠ  
TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**



**LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC  
NGÀNH CÔNG NGHỆ THÔNG TIN**

**Đề tài**

**SINH CÂU MÔ TẢ CHO HÌNH ẢNH  
DỰA TRÊN MÔ HÌNH SMALLCAP VỚI DEEP ATTENTION**

**IMAGE CAPTIONING BASED ON SMALLCAP  
WITH DEEP ATTENTION**

**Sinh viên: Ngô Đức Hiếu**

**Mã số: B1910223**

**Khóa: K45**

**Cần Thơ, 12/2023**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO  
ĐẠI HỌC CẦN THƠ  
TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**



**LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC  
NGÀNH CÔNG NGHỆ THÔNG TIN**

**Đề tài**

**SINH CÂU MÔ TẢ CHO HÌNH ẢNH  
DỰA TRÊN MÔ HÌNH SMALLCAP VỚI DEEP ATTENTION**

**IMAGE CAPTIONING BASED ON SMALLCAP  
WITH DEEP ATTENTION**

**Giảng viên hướng dẫn:  
TS. Lâm Nhựt Khang**

**Sinh viên: Ngô Đức Hiếu  
Mã số: B1910223  
Khóa: K45**

**Cần Thơ, 12/2023**

**XÁC NHẬN CHỈNH SỬA LUẬN VĂN  
THEO YÊU CẦU CỦA HỘI ĐỒNG**

Tên luận văn (tiếng Việt và tiếng Anh):

SINH CÂU MÔ TẢ CHO HÌNH ẢNH DỰA TRÊN MÔ HÌNH SMALLCAP VỚI DEEP  
ATTENTION

IMAGE CAPTION BASED ON SMALLCAP WITH DEEP ATTENTION

Họ tên sinh viên: Ngô Đức Hiếu

MASV: B1910223

Mã lớp: DI19V7A3

Đã báo cáo tại hội đồng ngành: Công Nghệ Thông Tin

Ngày báo cáo: 09/12/2023.

Luận văn đã được chỉnh sửa theo góp ý của Hội đồng.

*Cần Thơ, ngày 27 Tháng 12 năm 2023*

**Giáo viên hướng dẫn**

*(Ký và ghi họ tên)*

# NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Cần Thơ, ngày 9 tháng 12 năm 2023

Giảng viên hướng dẫn

TS. Lâm Nhật Khang

# LỜI CẢM ƠN

Luận văn tốt nghiệp đại học “**Sinh câu mô tả cho hình dựa trên mô hình SmallCap với deep attention**” là kết quả của quá trình phấn đấu và nghiên cứu và được sự giúp đỡ tận tình của quý Thầy Cô. Qua đây em xin gửi lời cảm ơn chân thành đến những người đã giúp đỡ em trong thời gian học tập, nghiên cứu vừa qua.

Em xin gửi lời cảm ơn chân thành và sâu sắc đến giáo viên hướng dẫn TS. Lâm Nhựt Khang, người đã trực tiếp hướng dẫn, hỗ trợ em tìm kiếm nguồn tài liệu tham khảo, cung cấp các thông tin khoa học cần thiết trong suốt quá trình em thực hiện đề tài để em có thể hoàn thành đề tài nghiên cứu của mình một cách tốt nhất.

Em xin chân thành cảm ơn lãnh đạo, ban giám hiệu cùng toàn thể các thầy cô giáo Trường Công Nghệ Thông Tin và Truyền Thông, trường Đại Học Cần Thơ đã truyền thụ những kiến thức quý báu tạo tiền đề cho em hoàn thành tốt công việc nghiên cứu khoa học của mình.

Cuối cùng, em xin chân thành cảm ơn gia đình, bạn bè đã luôn động viên, khích lệ và tạo điều kiện giúp đỡ trong suốt quá trình thực hiện để tôi có thể hoàn thành bài luận văn một cách tốt nhất.

Cần Thơ, ngày 9 tháng 12 năm 2023

Sinh viên thực hiện

Ngô Đức Hiếu

# MỤC LỤC

CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN .....	2
1.1 TỔNG QUAN .....	2
1.2 NHỮNG NGHIÊN CỨU LIÊN QUAN .....	3
1.3. MỤC TIÊU ĐỀ TÀI.....	4
1.4. ĐỐI TƯỢNG NGHIÊN CỨU .....	4
1.5. PHƯƠNG PHÁP NGHIÊN CỨU .....	4
CHƯƠNG 2. CƠ SỞ LÝ THUYẾT .....	5
2.1. MÔ HÌNH TRANSFORMER.....	5
2.2. CƠ CHẾ SELF ATTENTION .....	6
2.3. BỘ MÃ HÓA ENCODER .....	7
2.3.1 CƠ CHẾ RETRIEVING CAPTION VỚI PROMT.....	11
2.3.2 BỘ GIẢI MÃ GENERATIVE PRE-TRAINED TRANSFORMER .....	12
2.3.3 LỚP CROSS ATTENTION .....	13
2.4. PHƯƠNG PHÁP ĐÁNH GIÁ.....	14
CHƯƠNG 3. PHƯƠNG PHÁP THỰC HIỆN.....	15
3.1. MÔ TẢ BÀI TOÁN.....	15
3.1.1. DEEP ATTENTION .....	17
CHƯƠNG 4. THỰC NGHIỆM .....	18
4.1. TẬP DỮ LIỆU .....	18
4.2. TIỀN XỬ LÝ DỮ LIỆU .....	19
4.3. KẾT QUẢ THỰC NGHIỆM .....	19
4.4 ĐÁNH GIÁ ĐỘ CHÍNH XÁC .....	21
4.5 THẢO LUẬN KẾT QUẢ THỰC NGHIỆM .....	21
CHƯƠNG 5. KẾT LUẬN.....	25
5.1. KẾT QUẢ ĐẠT ĐƯỢC.....	25
5.2 HƯỚNG PHÁT TRIỂN.....	25
TÀI LIỆU THAM KHẢO.....	26

# DANH MỤC HÌNH ẢNH

Hình 1 . Mô hình Transformer [6] .....	5
Hình 2 . Cấu trúc self attention Nguồn : <a href="https://arxiv.org/abs/1706.03762">https://arxiv.org/abs/1706.03762</a> .....	6
Hình 3. Ví dụ biểu diễn từ đầu vào .....	7
Hình 4. Tổng quan kiến trúc mô hình SmallCap [7].....	9
Hình 5. Tổng quan mô hình CLIP [8] .....	10
Hình 6. Minh hoạ cách chia patch Nguồn <a href="https://medium.com/geekculture/vision-transformer-tensorflow-82ef13a9279">https://medium.com/geekculture/vision-transformer-tensorflow-82ef13a9279</a> .....	11
Hình 7. Tổng quan mô hình GPT [15] .....	12
Hình 8. Kiến trúc cross attention.....	13
Hình 9. Quy trình thực hiện xây dựng câu mô tả cho hình ảnh .....	15
Hình 10.Tổng quan mô hình sinh câu mô tả cho hình ảnh dựa trên mô hình SmallCap .....	16
Hình 11. Self-attention và deep attention [17] .....	17
Hình 12. Tập dữ liệu Flickr8k.....	18

# DANH MỤC BẢNG BIỂU

Bảng 1. Bảng cấu hình máy thực nghiệm .....	19
Bảng 2 . Bảng tham số huấn luyện mô hình .....	20
Bảng 3. Bảng kết quả đánh giá BLEU cho mô hình xây dựng câu mô tả cho hình ảnh .....	21
Bảng 4. Mô tả hình ảnh với mô hình SmalCap modify deep attention và SmallCap Base .....	24



## DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	Mô tả
vector	Véc-tơ
inference	Là quá trình suy luận của mô hình
vector embedding	Véc-tơ nhúng từ hoặc ảnh
cosine similarity	Độ đo giữa véc-tơ tính bằng tích vô hướng giữa 2 véc tơ
CLIP	Contrastive Language-Image Pre-Training
GPT	Chat Generative Pre-training Transformer
BLEU	Bilingual Evaluation Understudy

## TÓM TẮT

Sinh câu mô tả cho hình ảnh là một trong những bài toán quan trọng trong lĩnh vực hiểu hình ảnh liên quan đến thị giác máy tính và xử lý ngôn ngữ tự nhiên. Trong luận văn này, mô hình học sâu dựa được sử dụng để tạo câu mô tả cho hình ảnh tiếng Việt và tiếng Anh. Kiến trúc hợp kết hợp đặc trưng hình ảnh được rút trích từ mô hình CLIP, cùng với các câu mô tả được mã hoá và đặc trưng ảnh đưa vào deep attention kết hợp mô hình GPT kết để sinh câu mô tả cho hình ảnh. Mô hình được huấn luyện trên tập dữ liệu Flickr8k. Kết quả đánh giá mô hình sử dụng tập dữ liệu Flickr8k tiếng Việt BLEU-1: 72.39, BLEU-2: 59.79, BLEU-3: 49.2, BLEU-4: 39.51. Và trên tập dữ liệu Flickr8k tiếng Anh BLEU-1: 73.9 , BLEU-2: 56.69, BLEU-3: 41.54, BLEU-4: 30.57 .

# ABSTRACT

Image captioning is a significant problem in the field of computer vision and natural language processing, where the objective is to generate descriptive sentences for images. In this thesis, a deep learning model is utilized for creating captions in both Vietnamese and English. The architecture combines image features extracted by the CLIP model and encoded descriptive sentences are fed into a deep attention with GPT model for generating image captions. The model is trained on the Flickr8k dataset. Evaluation results on the Flickr8k dataset show the following scores: For Vietnamese, BLEU-1: 72.39, BLEU-2: 59.79, BLEU-3: 49.2, and BLEU-4: 39.51. For English, the scores are BLEU-1: 73.9, BLEU-2: 56.69, BLEU-3: 41.54, and BLEU-4: 30.57.

# CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN

Chương đầu tiên của luận văn sẽ giới thiệu một cách tổng quát về đề tài cũng như sự cần thiết của vấn đề mà luận văn hướng tới: bài toán xây dựng câu mô tả cho hình ảnh nói chung, bài toán xây dựng câu mô tả cho hình ảnh ở tiếng Việt nói riêng, các nghiên cứu liên quan, mục tiêu đề tài và phương pháp nghiên cứu cũng được trình bày trong chương này.

## 1.1 TỔNG QUAN

Trong những năm gần đây, với sự phát triển nhanh chóng của trí tuệ nhân tạo, vấn đề xây dựng câu mô tả cho hình ảnh đã thu hút sự quan tâm của nhiều nhà nghiên cứu trong lĩnh vực trí tuệ nhân tạo. Xây dựng mô tả cho hình ảnh liên quan đến thị giác máy tính và xử lý ngôn ngữ tự nhiên để nhận ra được ngữ cảnh liên quan đến thị giác máy tính và xử lý ngôn ngữ tự nhiên để nhận ra được ngữ cảnh của hình ảnh, nội dung hình ảnh có thể được mô tả một phần thông qua đối tượng và vị trí của chúng. Từ đó, tiến tới mô tả chúng bằng các ngôn ngữ tự nhiên. Nói cách khác, một hệ thống sinh câu mô tả cho hình ảnh thường gồm 2 thành phần là mô hình thị giác máy tính dùng để rút trích đặc trưng ảnh và mô hình ngôn ngữ để sinh câu mô tả cho ảnh. Với một hình ảnh đầu vào cho trước, mô hình thị giác máy tính sẽ rút trích các đặc trưng quan trọng của ảnh, cố gắng “hiểu” hình ảnh đó; các đặc trưng đã rút trích được sẽ làm đầu vào cho mô hình ngôn ngữ để sinh ra câu mô tả cho hình ảnh đó. Ứng dụng của việc xây dựng mô tả cho hình ảnh rất rộng rãi và quan trọng, chẳng hạn như việc thực hiện tương tác giữa người và máy tính, áp dụng cho trợ lý ảo, tích hợp vào các công cụ chỉnh sửa, lập chỉ mục hình ảnh và hỗ trợ người khuyết tật.

Trong những nghiên cứu gần đây, vấn đề xây dựng câu mô tả cho hình ảnh được nhiều nhà nghiên cứu quan tâm vận dụng trên nhiều ngôn ngữ khác nhau. Luận văn “**Xây dựng câu mô tả cho hình ảnh dựa trên mô hình SmallCap với deep attention**” sẽ trình bày một phương pháp xây dựng câu mô tả cho hình ảnh trên ngôn ngữ Tiếng Anh và Tiếng Việt.

## 1.2 NHỮNG NGHIÊN CỨU LIÊN QUAN

Trong lĩnh vực “image captioning”, nhiều nghiên cứu tiên tiến đã được thực hiện để khám phá và giải quyết những thách thức cơ bản liên quan đến việc tạo câu mô tả cho hình ảnh. Một số nhóm nghiên cứu hàng đầu đã đưa ra các phương pháp sáng tạo để nâng cao chất lượng và hiệu suất của các mô hình trong nhiệm vụ này.

Zhongliang Yang và cộng sự [1] đã sử dụng phương pháp mạng nơ-ron đa mô hình (Multi-Model Neural Network) để xây dựng câu mô tả cho hình ảnh. Đầu tiên, mô hình Faster R-CNN [2] được sử dụng để trích xuất thông tin của các đối tượng và mối quan hệ không gian của chúng trong ảnh tương ứng; bên cạnh đó, một mạng RNN dựa trên các đơn vị LSTM với kiến trúc Attention để tạo câu.

Nghiên cứu của Dosovitskiy [3] đã đưa ra một cái nhìn mới về khả năng của mô hình Vision Transformer trong việc rút trích đặc trưng ảnh, vượt xa so với ResNet. Trên cơ sở này, Liu và đồng nghiệp [8] đã giới thiệu mô hình sử dụng Transformer, chia ảnh thành các patch để rút trích đặc trưng, và sử dụng encoder-decoder để tạo câu mô tả cho hình ảnh.

Lam và cộng sự [4] đã kết hợp VGG16 và LSTM để sinh câu mô tả ảnh tiếng Việt, và họ đã áp dụng mô hình YOLO [5] để dò tìm màu sắc trong ảnh và điều chỉnh câu mô tả để phản ánh chính xác màu sắc.

Tất cả những nghiên cứu này đều đóng góp vào sự phát triển và hiểu biết sâu sắc trong lĩnh vực đầy thách thức của “image captioning”. Các phương pháp và kỹ thuật được đề cập mở ra những triển vọng mới trong việc hiểu và mô tả hình ảnh bằng ngôn ngữ tự nhiên.

### **1.3. MỤC TIÊU ĐỀ TÀI**

Trong phạm vi luận văn này sẽ hướng đến mục tiêu xây dựng một giải pháp sinh câu mô tả ảnh cho phép tạo ra những câu có nội dung sao cho giống nhất với nội dung ảnh và có thể áp dụng được trong những nội dung thực tế. Giải pháp này sẽ ứng dụng những kỹ thuật, tiến bộ mới, đặc biệt là những thành quả nổi bật trong lĩnh vực học sâu để giải quyết chất lượng nội dung câu mô tả. Đồng thời, giải pháp cũng được chứng minh khả năng hoạt động ổn định trong điều kiện thiếu thôn dữ liệu huấn luyện. Ngoài ra, luận văn cũng sẽ đưa ra chi tiết những thử nghiệm, đánh giá trên những mô hình phương pháp có sử dụng trong quá trình nghiên cứu để phân tích những ưu, nhược điểm của chúng khi áp dụng vào bài toán này.

### **1.4. ĐỐI TƯỢNG NGHIÊN CỨU**

Đối tượng nghiên cứu của đề tài là nghiên cứu các mô hình học sâu như Transformer từ đó tìm kiếm giải pháp cho bài toán xây dựng câu mô tả cho hình ảnh.

### **1.5. PHƯƠNG PHÁP NGHIÊN CỨU**

Phương pháp nghiên cứu của luận văn chủ yếu dựa vào nguồn tài liệu tham khảo (sách, bài báo khoa học, trang web, ...) từ Internet, các trang web của cộng đồng công nghệ thông tin. Từ đó, vận dụng các kiến thức, phương pháp đã học và nghiên cứu được vào xây dựng các mô hình thích hợp nhằm giải quyết vấn đề đặt ra ban đầu.

### **1.6. BỐ CỤC QUYỂN LUẬN VĂN**

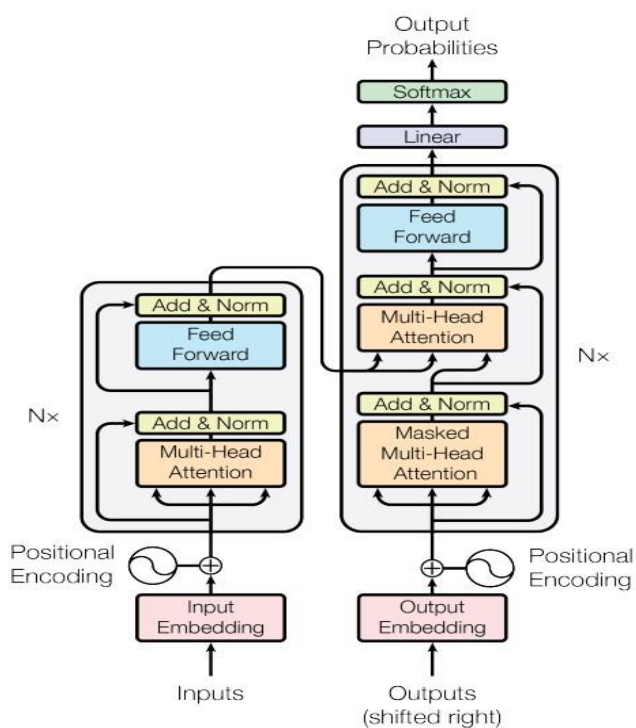
Nội dung của quyển luận văn bao gồm 05 chương. Chương 1 giới thiệu tổng quan về đề tài, trình bày nội dung, mục tiêu đề tài, phạm vi và phương pháp nghiên cứu đề tài. Bên cạnh đó, còn trình bày các nghiên cứu liên quan đến đề tài. Chương 2 trình bày các cơ sở lý thuyết có liên quan để giải quyết bài toán sinh câu mô tả cho hình ảnh. Chương 3 thảo luận chi tiết các bước được thực hiện để giải quyết bài toán được đề tài đặt ra. Chương 4 mô tả các bước tiến hành thực nghiệm, xử lý dữ liệu, huấn luyện mô hình và đánh giá kết quả. Chương 5 tổng kết kết quả đạt được của đề tài và đề xuất hướng phát triển cho đề tài.

## CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

Chương 2 sẽ trình bày nội dung các lý thuyết liên quan được sử dụng để giải quyết bài toán xây dựng câu mô tả cho hình ảnh như mô hình Transformer và mô hình Vision Transformer.

### 2.1. MÔ HÌNH TRANSFORMER

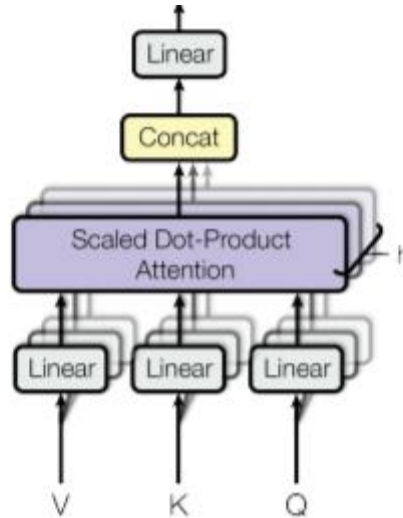
Mô hình Transformer [6] bao gồm 2 phần lớn là bộ mã hóa (Encoder) và bộ giải mã (Decoder) với bộ mã hóa dùng để học các véc-tơ biểu diễn của câu với mong muốn rằng véc-tơ này mang thông tin của câu đó. Bộ giải mã thực hiện chức năng chuyển vector biểu diễn kia về từ ở ngôn ngữ đích. Transformer có thể nhận song song chuỗi đầu vào giúp đẩy nhanh quá trình xử lý.



Hình 1 . Mô hình Transformer [6]

## 2.2. CƠ CHẾ SELF ATTENTION

Self-attention là thành phần quan trọng nhất của Transformer nó là cơ chế giúp Transformer “hiểu” được sự “liên quan” giữa các từ trong một câu bằng cách tính các khoảng cách cosine similarity từng từ trong câu với các từ còn lại, self-attention có thể hiểu là mức độ chú ý của các từ trong câu với nhau.



Hình 2 . Cấu trúc self attention

Nguồn : <https://arxiv.org/abs/1706.03762>

+ Xây dựng self attention cần có 3 vector  $Q, K, V$  tương ứng với *Query, Key, Value* trong đó :

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

- *Query vector (Q)*: chứa thông tin của từ được tìm kiếm.
- *Key vector (K)*: biểu diễn thông tin các từ được so sánh với các từ tìm kiếm.
- *Value vector (V)*: biểu diễn nội dung, ý nghĩa của từ.
- $d_k$  là số chiều của key vector.



## 2.3. BỘ MÃ HÓA ENCODER

Bộ mã hóa Encoder sử dụng cơ chế self-attention để mã hóa. Bộ mã hóa gồm 2 lớp chính là self-attention và feed forward. Trong mô hình Transformer, bộ mã hóa gồm  $N$  encoder xếp chồng lên nhau. Output của encoder cuối cùng sẽ được nối sang bộ giải mã, tất cả các encoder đều không chia sẻ trọng số.

**Word Embedding:** lớp dùng để chuyển đổi chuỗi đầu vào thành vector mà máy có thể hiểu được, do mỗi từ với vị trí khác nhau sẽ có nghĩa khác nhau với kỹ thuật này các vector biểu diễn có sự liên hệ về ngữ nghĩa.

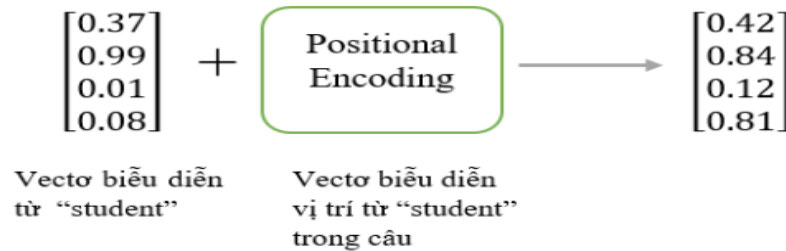
**Input Embedding:** câu đầu vào sẽ được mã hóa thành các vector bằng cách sử dụng lớp word embedding.

**Positional Encoding:** do mô hình Transformer sử dụng cơ chế self-attention và không có mối tương quan giữa các vị trí trong chuỗi, do đó nó có thể mất đi tính tuần tự trong câu. Lớp Position Encoding sẽ được thêm vào vector đầu ra của Embedding để tạo tính tuần tự cho câu nhưng vẫn sử dụng được Self-attention. Giá trị được tính như sau:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$
$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \quad (2)$$

+Trong đó,  $pos$  là vị trí của từ trong chuỗi và  $PE$  là giá trị thứ  $i$  trong Embedding có độ dài là  $d_{model}$ .

Trong đó,  $pos$  là vị trí của từ trong câu và  $PE$  là giá trị phần thứ  $i$  trong Embedding có độ dài  $d_{model}$ . Như vậy bộ mã hóa sẽ nhận ma trận biểu diễn của các từ đã được cộng với thông tin vị trí thông qua Positional Encoding (Hình 3).



Hình 3. Ví dụ biểu diễn từ đầu vào

+ Positional Encoding sử dụng hàm sin và cosin với tần suất khác nhau. Với mỗi chỉ số lẻ trên chuỗi dữ liệu đầu vào sẽ sử dụng hàm  $\sin$  và với mỗi chỉ số chẵn tương ứng sẽ sử dụng hàm  $\cos$  như trình bày ở (1).

**Multi-head Attention:** Thực chất là sử dụng nhiều khối self-attention. Multi-head attention tính cùng lúc với mỗi *Query*, *Key*, *Value* là một “Head” việc này giúp cho Transformer có thể thực hiện song song sau đó kết hợp lại để cho kết quả cuối cùng.

**Residual Connection:** sau khi nhận kết quả đầu ra của self-attention thì kết quả sẽ được chuyển sang các lớp Add & Normallize lớp này chính là Normalize sử dụng thêm Residual Connection. Residual Connection sử dụng kết quả đầu ra cộng với kết quả đầu ra của các lớp trước đó để đưa vào lớp Normalize nhằm hạn chế những đặc trưng không mang nhiều lợi ích và tránh việc bị overfit. Ngoài ra nó còn giúp kết nối.

**Feed Forward:** sau khi thực hiện tính toán ở các lớp thì kết quả đầu ra của các lớp sẽ được chuyển đến lớp feed forward. Lớp này sẽ chịu trách nhiệm xử lý thông tin trước khi truyền đi tới các giai đoạn giải mã sau.

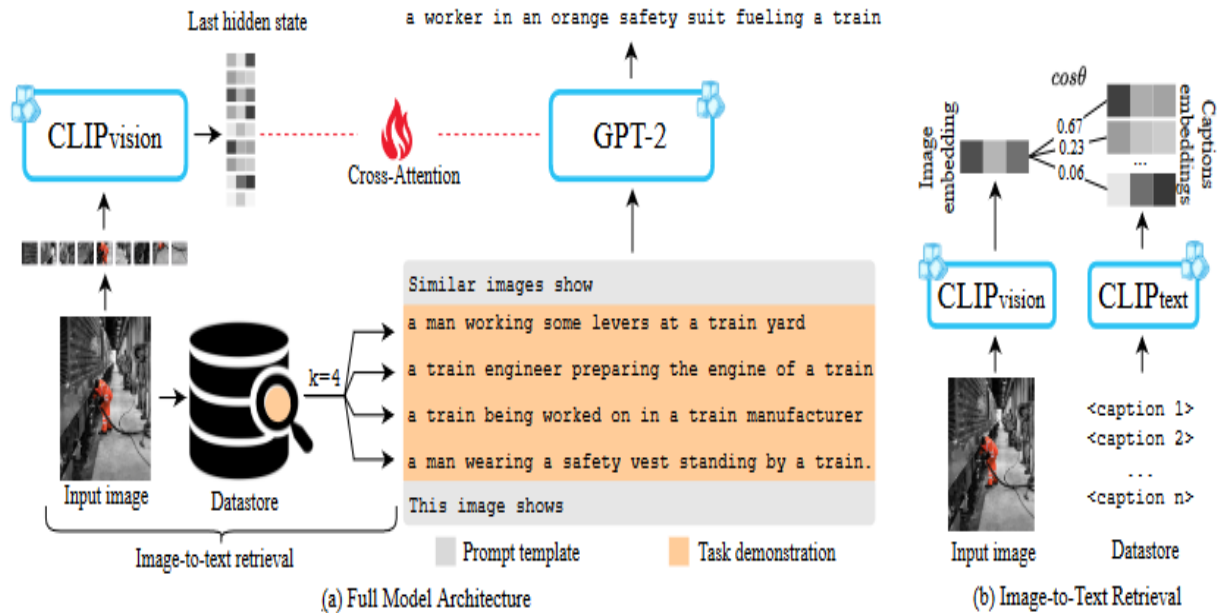
## 2.4. BỘ GIẢI MÃ DECODER

Về cơ bản bộ giải mã Decoder giống với bộ mã hóa Encoder nhưng có thêm thành phần Masked Multi-head attention. Mục đích của Maskead multi-head dùng để tính toán attention đối với những từ mà đằng sau ta đang muốn tìm. Bộ giải mã Decoder cũng gồm  $N$  decoder xếp chồng lên nhau.

**Cơ chế hoạt động của bộ giải mã:** bộ mã hóa tiến hành xử lý chuỗi đầu vào. Đầu ra của bộ mã hóa được chuyển thành một tập hợp các vector chú ý  $K$  và  $V$ . Các vector này sẽ được sử dụng bởi bộ giải mã trong lớp encoder giúp bộ giải mã tập trung vào những vị trí trích hợp trong chuỗi đầu vào. Thực hiện tính toán giống như bộ mã hóa. Tương tự như cách làm với các đầu vào của bộ mã hóa nhưng thêm mã hóa vị trí vào các đầu vào của bộ giải mã đó để chỉ ra vị trí của một từ. Masked MultiHead Attention là Multi-Head Attention đã nói ở trên, tuy nhiên các từ có một số từ sẽ được che (mask) lại.

## 2.5. MÔ HÌNH SMALLCAP

Phần còn lại của Chương sẽ trình bày chi tiết mô hình xây dựng câu mô tả cho hình ảnh sử dụng mô hình SmallCap [7] (Hình 4).

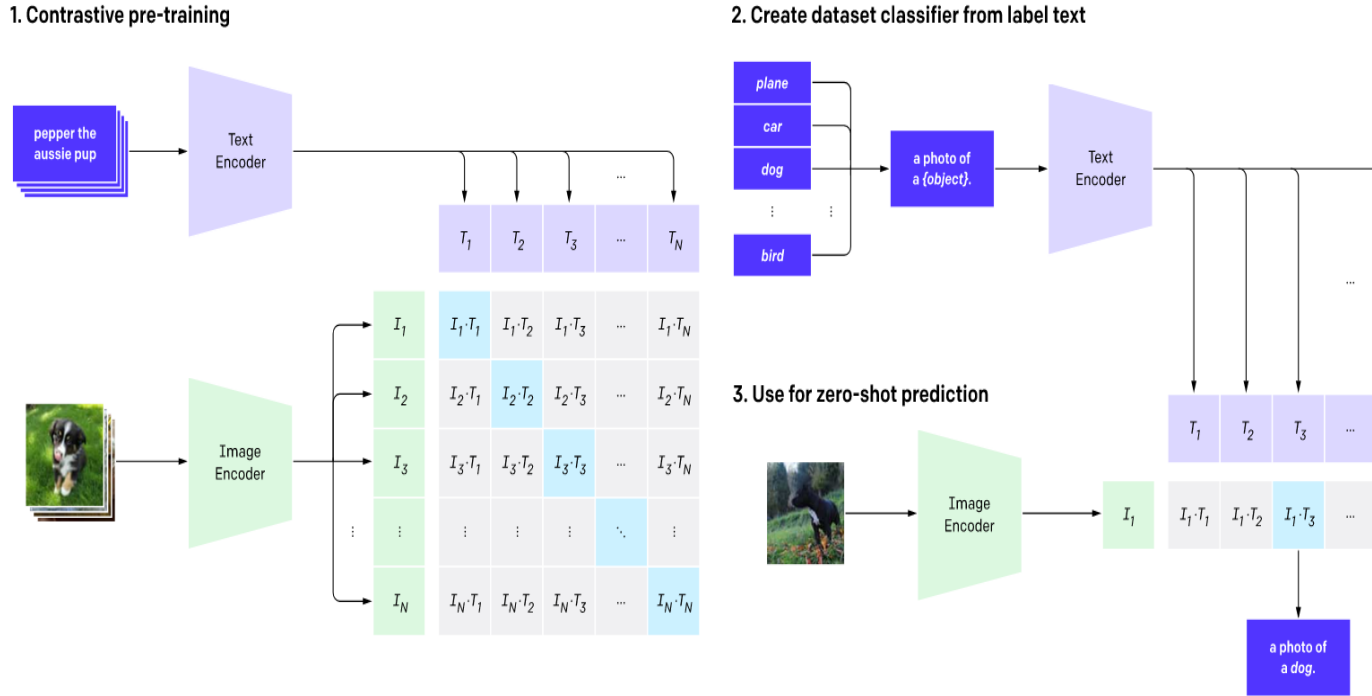


Hình 4. Tổng quan kiến trúc mô hình SmallCap [7]

Mô hình SmallCap [7] là một “light-weight” Model xây dựng dựa trên Transformer. Khối encoder sử dụng mô hình CLIP [8] nhận đầu vào là một cặp hình ảnh sau khi mã hóa thành vec-tơ từ chúng được đem so sánh với các vec-tơ bằng phép đo cosin similarity từ lấy từ ở datastore thông qua cơ chế Retrieval Caption [9] từ đó cung cấp thông tin về hình ảnh đầu vào cho khối decoder.

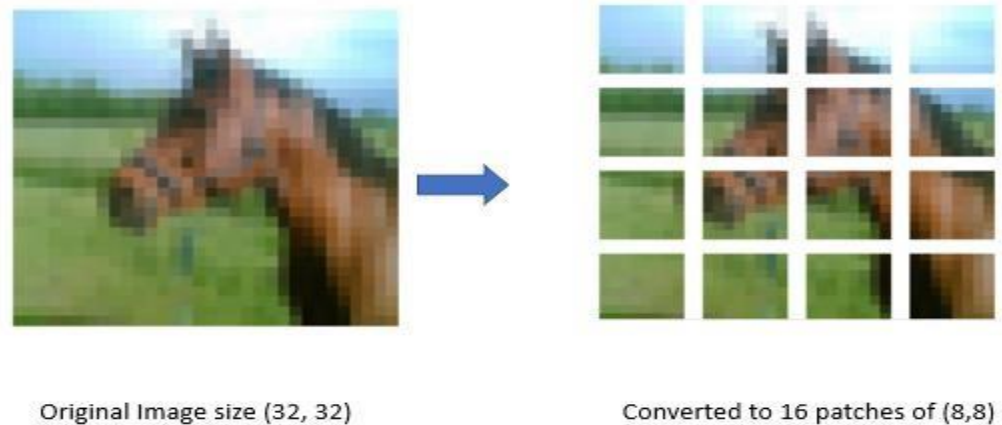
## 2.5.1 BỘ MÃ HÓA CONTRASTIVE LANGUAGE–IMAGE PRE-TRAINING

*Contrastive Language–Image Pre-training* hay còn được gọi là CLIP [8] là bộ giải mã được sử dụng cho mô hình SmallCap [7] (hình 5).



Hình 5. Tổng quan mô hình CLIP [8]

Hình ảnh trước khi được đưa vào mô hình CLIP cần được chia thành một chuỗi các patch có kích thước cố định để cho phù hợp với đầu vào Transformer. Chúng tôi thay đổi kích thước hình ảnh đầu vào thành độ phân giải cố định  $X \in \mathbb{R}^{H \times W \times 3}$ , sau đó chia hình ảnh đã thay đổi kích thước thành  $N$  mảng, trong đó  $N = HW/P$  và  $P$  là số lượng patch. Sau đó làm phẳng từng patch và định hình lại chúng thành chuỗi patch 1D. Lớp linear embedding để ánh xạ chuỗi bản vá được làm phẳng thành không gian tiềm ẩn và thêm vị trí 1D có thể học được nhúng vào các tính năng của patch, kết quả nhận được là đầu vào của CLIP Vision Encoder được ký hiệu là  $Pa = [p_1, \dots, p_N]$ .



Hình 6. Minh họa cách chia patch

Nguồn <https://medium.com/geekculture/vision-transformer-tensorflow-82ef13a9279>

### 2.3.1 CƠ CHẾ RETRIEVING CAPTION VỚI PROMT

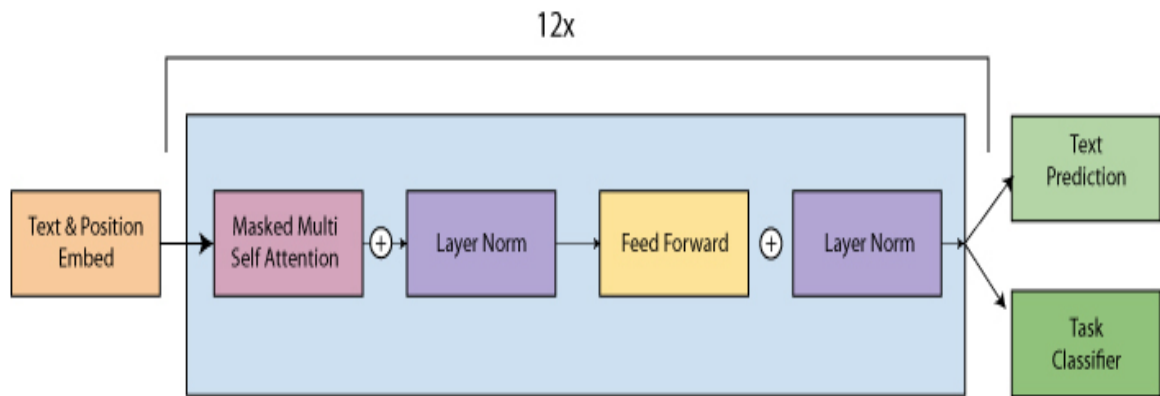
Kết hợp với mô hình CLIP [8] (hình 5) khi ảnh đầu vào sẽ được phân tách thành các patches và biến đổi thành các vector embedding (image embedding qua CLIP Vision và word embedding qua CLIP text) sau đó sử dụng phép đo cosine similarity để tính khoảng cách giữa image vector embedding và text vector embedding để chọn ra  $k$  (thực nghiệm  $k=4$ ) caption tương đồng nhất điền vào chỗ trống template dạng như sau để được prompt [10] hoàn chỉnh (3):

$$\begin{aligned} &\text{"Similar images show} \\ &\quad \{caption-1\}...\{caption-k\}. \\ &\text{This image shows."} \end{aligned} \tag{3}$$

Cơ chế retrieve caption [9] [11] [12] [13] [14] cho phép dữ liệu bổ sung có thể được thêm vào trong quá trình inference để cải thiện khái quát hóa. Toàn bộ dữ liệu có thể được hoán đổi cho dữ liệu mới tại thời điểm inference, cho phép chuyển miền dữ liệu mà không cần đào tạo lại (re-training). Cách tiếp cận này cho phép SmallCap thích ứng với các yêu cầu dữ liệu khác nhau và chuyển sang các miền dữ liệu mới một cách hiệu quả.

### 2.3.2 BỘ GIẢI MÃ GENERATIVE PRE-TRAINED TRANSFORMER

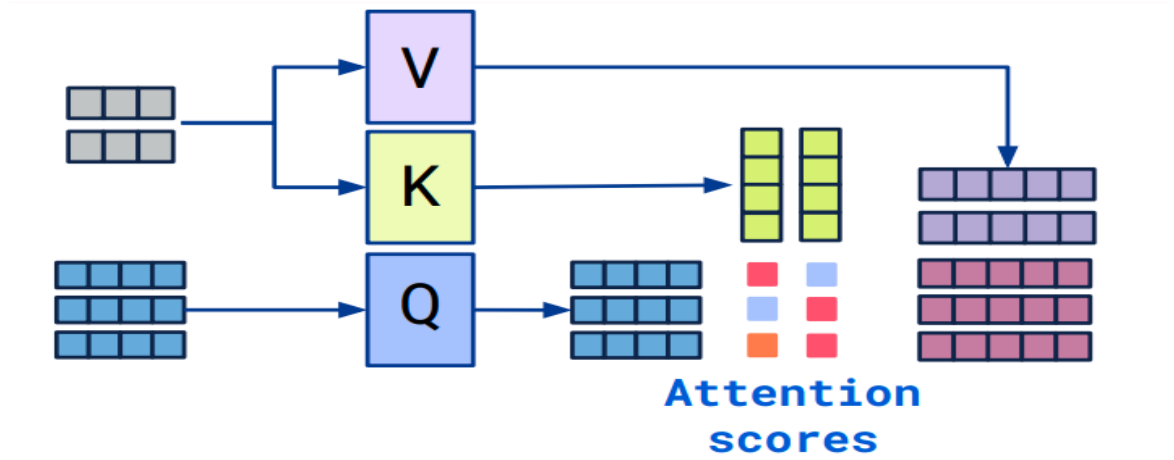
Mô hình Generative Pre-trained Transformer hay được gọi là GPT [15] được xây dựng dựa trên kiến trúc của mô hình Transformer [6]. Mô hình GPT được cấu thành từ 12 attention heads ở decoder của Transformer [6] xếp chồng lên nhau và hoạt động theo “tự hồi quy” (autoregressive). Tự hồi quy là kỹ thuật khi đầu ra của vòng hồi quy trước đó được đưa vào thành đầu vào của vòng hồi quy hiện tại.



Hình 7. Tổng quan mô hình GPT [15]

- Ở mô hình SmallCap<sub>base</sub> sử dụng GPT [11] với kích thước  $d_{model} = 768$  hidden layer và 12 attention heads,  $d$  mặc định là 64 ( $d_{model}/h$ ) theo Vaswani và cộng sự [6].

### 2.3.3 LỚP CROSS ATTENTION



Hình 8. Kiến trúc cross attention

Để kết nối giữa encoder CLIP [8] và decoder GPT [15] cả 2 mô hình hoạt động 2 không gian vector khác nhau nên chúng được kết nối thông qua context-layer gồm các lớp cross-attentions.

Theo [7] Rita và cộng sự nghiên cứu thì layer cross-attention có thể nằm trước hoặc trong và thay thế self-attention ở khối decoder trong SmallCap tùy vào cách fine-tune mô hình.

Lớp cross-attentions Gồm bộ vector  $Q, K, V$  xử lí như mutlihead-cross-attention (MHA) với tham số như sau:

$$MHA(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W_o \quad (4)$$

$$\text{Head}_i = \text{Att}(QW_i^Q, KW_i^K, VW_i^V)$$

$$\text{Att}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

+ Với  $W_i^K \in \mathbb{R}^{d_{\text{encoder}} \times d}$ ,  $W_i^V \in \mathbb{R}^{d_{\text{encoder}} \times d}$ ,  $W_i^Q \in \mathbb{R}^{d_{\text{decoder}} \times d}$ ,  $W_o \in \mathbb{R}^{h \times d \times d_{\text{decoder}}}$  là tham số được học của mô hình với kích thước mô hình  $d$ .

## 2.4. PHƯƠNG PHÁP ĐÁNH GIÁ

BLEU [16] là viết tắt của Bilingual Evaluation Understudy. BLEU thường được sử dụng trong bài toán dịch máy (Machine Translation) như một độ đo hay một hệ số điểm khi so sánh một bản dịch (candidate translation) với một hay nhiều bản dịch tham khảo (reference translation). Mặc dù vậy, BLEU cũng có thể được sử dụng để đánh giá cho các bài toán sinh văn bản (text generation) như: image caption, Text summarization... Công thức tính BLEU như sau:

$$p_i = \frac{\sum_j NR_j}{\sum_j NT_j} \quad (5)$$

$$Score = \exp \left\{ \sum_1 w_i \log(p_i) - \max \left\{ \frac{L_{ref}}{L_{tra}} - 1.0 \right\} \right\} \quad (6)$$

Trong đó

- $NR_j$ : là số lượng N-grams trong phân đoạn  $j$  của bản dịch tham chiếu
- $NT_j$ : là số lượng của các N-grams trong phân đoạn  $j$  của bản dịch máy.
- $L_{ref}$ : là số lượng của các từ trong bản dịch tham chiếu, độ dài của nó thường bằng độ dài của bản dịch máy.
- $W_i = N^{-I}$
- $L_{tra}$ : là số lượng của các từ trong bản dịch máy.

Giá trị BLEU [16] đánh giá mức độ tương ứng giữa hai bản dịch và nó được thực hiện trên từng phân đoạn, ở đây phân đoạn được hiểu là đơn vị tối thiểu trong các bản dịch, thông thường mỗi phân đoạn là một câu hoặc một đoạn. Việc thống kê độ trùng khớp của các N-grams dựa trên tập hợp các N-grams trên các phân đoạn trước hết là nó được tính trên từng phân đoạn, sau đó là giá trị này trên tất cả các phân đoạn. Trong luận văn này BLEU-1, BLEU-2, BLEU-3 và BLEU-4 được sử dụng để đánh giá mô hình sinh câu mô tả cho hình ảnh.



## CHƯƠNG 3. PHƯƠNG PHÁP THỰC HIỆN

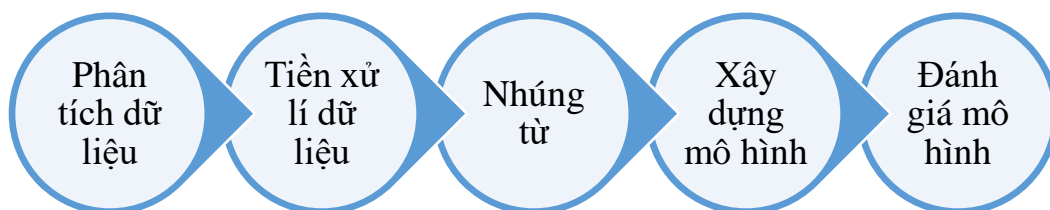
Chương này trình bày chi tiết các bước giải quyết bài toán đã đặt ra trong đề tài ở chương 1. Đầu tiên chúng tôi sẽ mô tả cụ thể lại bài toán mô tả ảnh, kế tiếp là phương pháp để xây dựng câu mô tả tổng quát, và cuối cùng là phương pháp mà chúng tôi đề xuất để xây dựng câu mô tả cho hình ảnh.

### 3.1. MÔ TẢ BÀI TOÁN

Xây dựng mô tả cho hình ảnh là một trong những bài toán thuộc lĩnh vực trí tuệ nhân tạo kết hợp giữa hai khía cạnh thị giác máy tính để nhận dạng hình ảnh và xử lý ngôn ngữ tự nhiên để xây dựng câu mô tả. Mô hình thực hiện rút trích đặc trưng hình ảnh mã hóa chúng tạo nên các vector đặc trưng hình ảnh. Bên cạnh đó, mô hình cũng thực hiện mã hóa các mô tả thành vector từ. Tiếp theo, vector đặc trưng hình ảnh và vector từ đồng thời được kết hợp và “giải mã” chúng để được câu mô tả hoàn chỉnh.

### 3.2. TỔNG QUÁT QUY TRÌNH XÂY DỰNG CÂU MÔ TẢ CHO HÌNH ẢNH

Tổng quát quá trình triển khai và đánh giá mô hình xây dựng câu mô tả cho hình ảnh được thực hiện theo các bước như sau (hình 8):



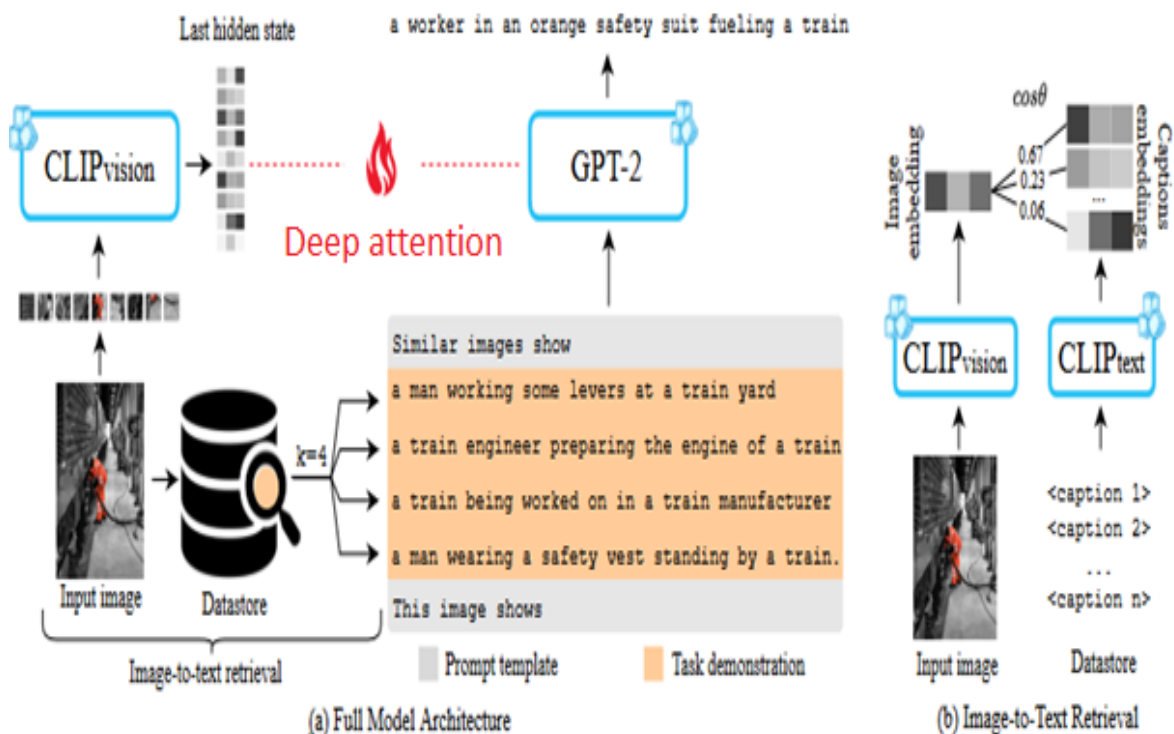
Hình 9. Quy trình thực hiện xây dựng câu mô tả cho hình ảnh

- Bước 1 - Phân tích dữ liệu: dataset dùng để huấn luyện bài toán sinh câu mô tả cho hình ảnh thường có dạng: mỗi hình sẽ có khoảng 5 câu mô tả khác nhau được đánh số từ 0 đến 4.

- Bước 2 - Tiền xử lý dữ liệu: làm sạch dữ liệu loại bỏ các ký tự dư thừa nhằm cải thiện hiệu suất của mô hình.
- Bước 3 – Nhúng từ là một cách biểu diễn đã học cho văn bản trong đó các có nghĩa có biểu diễn tương tự. Nhúng từ trên thực tế là một lớp kỹ thuật trong đó các từ riêng lẻ được biểu diễn dưới dạng vector có giá trị thực trong không gian vector được xác định trước. Mỗi từ được ánh xạ tới một vector và các giá trị vector được học theo cách tương tự như mạng nơ-ron.
- Bước 4 - Xây dựng mô hình: được trình bày chi tiết trong phần tiếp theo.
- Bước 5 - Đánh giá mô hình: đánh giá mô hình bằng điểm BLEU [16].

### 3.3 MÔ HÌNH XÂY DỰNG CÂU MÔ TẢ CHO HÌNH ẢNH

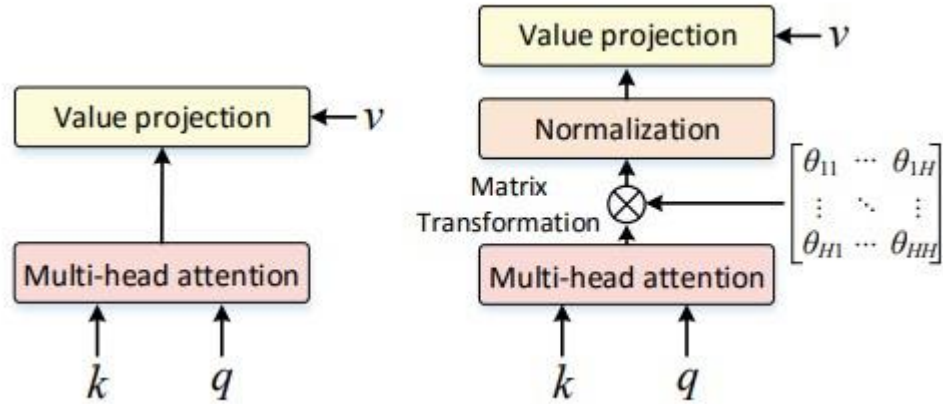
Phần còn lại của Chương sẽ trình bày chi tiết mô hình xây dựng câu mô tả cho hình ảnh sử dụng kiến trúc mô hình Transformer với CLIP Encoder [8] , GPT [15] decoder kết hợp với deep attention ở context layer.



Hình 10. Tổng quan mô hình sinh câu mô tả cho hình ảnh dựa trên mô hình SmallCap với deep attention

### 3.1.1. DEEP ATTENTION

Zhou và các cộng sự [17] nhận thấy bảng đồ attention (attention maps) trong Vision Transformer rất giống nhau sau một vài lớp. Nói cách khác, attention map có xu hướng giống nhau ở các lớp trên cùng trong mô hình ViT, điều này cũng có nghĩa là cơ chế attention đã thất bại trong việc thực thi ở các lớp trên cùng. Zhou và các cộng sự [17] đã cải tiến mô hình bằng cách cải tiến lớp self-attention thành lớp deep attention (hình 10).



Hình 11. Self-attention và deep attention [17]

$$Deep\ Attention(Q, K, V) = Norm(\theta^T(Softmax(\frac{QK^T}{\sqrt{d_k}})))V, \quad (7)$$

Trong đó, ma trận Transformer  $\theta^T$  là ma trận có thể học được (learnable)  $\theta^T \in R^{H \times H}$  được nhân với bảng đồ self-attention và head dimension,  $d_k$  là số chiều véc-tơ Key. Việc chia cho  $\sqrt{d_k}$  là nhằm mục đích tránh tràn luồng nếu số mũ trở nên quá lớn.

## CHƯƠNG 4. THỰC NGHIỆM

Chương này sẽ tiến hành đánh giá thực nghiệm thu được khi triển khai mô hình xây dựng câu mô tả cho hình ảnh được thực hiện ở Chương 3.

### 4.1. TẬP DỮ LIỆU

Chúng tôi sử dụng tập dữ liệu Flickr8k để huấn luyện và đánh giá mô hình xây dựng câu mô tả cho hình ảnh. Tập dữ liệu Flickr8k được lấy từ trang web của Kaggle để xây dựng và đánh giá mô hình xây dựng mô tả cho hình ảnh. Tập dữ liệu chứa 8.091 hình ảnh và mỗi hình ảnh chứa tương ứng với 5 câu mô tả bằng tiếng Anh. Ở tiếng Việt chúng tôi thực nghiệm với tập dữ liệu Flickr8k của Lam và các cộng sự cung cấp. Tên hình ảnh được xem là ID dùng để phân biệt nhau và được dùng để liên giữa các file train và test. Ví dụ về tập dữ liệu được minh hoạ ở (Hình 11).



a little girl in a pink dress going into a wooden cabin .  
a little girl climbing the stairs to her playhouse .  
a little girl climbing into a wooden playhouse .  
a girl going into a wooden building .  
a child in a pink dress is climbing up a set of stairs in an entry way .



two dogs on pavement moving toward each other .  
two dogs of different breeds looking at each other on the road .  
a black dog and a white dog with brown spots are staring at each other in the street .  
a black dog and a tri-colored dog playing with each other on the road .  
a black dog and a spotted dog are fighting



young girl with pigtails painting outside in the grass .  
there is a girl with pigtails sitting in front of a rainbow painting .  
a small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it .  
a little girl is sitting in front of a large painted rainbow .  
a little girl covered in paint sits in front of a painted rainbow with her hands in a bowl .



man laying on bench holding leash of dog sitting on ground  
a shirtless man lies on a park bench with his dog .  
a man sleeping on a bench outside with a white and black dog sitting next to him .  
a man lays on the bench to which a white dog is also tied .  
a man lays on a bench while his dog sits by him .



the man with pierced ears is wearing glasses and an orange hat .  
a man with glasses is wearing a beer can crocheted hat .  
a man with gauges and glasses is wearing a blitz hat .  
a man wears an orange hat and glasses .  
a man in an orange hat starring at something .

Hình 12. Tập dữ liệu Flickr8k

Nguồn: <https://www.kaggle.com/datasets/adityajn105/flickr8k>

## 4.2. TIỀN XỬ LÝ DỮ LIỆU

Tập dữ liệu Flickr8k có 8.091 với mỗi hình 5 câu mô tả vì thế tương ứng ta sẽ thu được 40.455 câu mô tả tiếng Việt. Tập dữ liệu được chia làm 3 phần với tỉ lệ 75% training (6000 ảnh và 30000 caption ), 12,5% validation (1000 ảnh và 5000 caption) và 12% test (1000 ảnh và 5000 caption). Ở bước tách từ, chúng tôi sử dụng phương pháp tách từ cho tiếng anh đó là sử dụng mô hình Tokenizer. Với tiếng việt chúng tôi sử dụng thư viện Underthesea tách từ trước khi đưa vào đưa vào mô hình Tokenizer. Đối với ảnh đầu vào chúng tôi chuyển về kích thước 224x224 px.

## 4.3. KẾT QUẢ THỰC NGHIỆM

Thực nghiệm trên Google Colaboratory với cấu hình 12GB Ram, GPU NVIDIA TESLA T4 và các thư viện mã nguồn mở của python.

CẤU HÌNH	CHỈ SỐ
CPU	Intel ® Xeon ® @ 2.2Ghz
GPU	Nvidia Tesla T4 15GB Vram
RAM	12 GB
OS	Linux

Bảng 1. Bảng cấu hình máy thực nghiệm

- Các tham số của các thực nghiệm được mô tả ở các Bảng 2 như sau:

THAM SỐ	GIÁ TRỊ
Encoder	CLIP Vit/B-32
Decoder	GPT-2 <sub>base</sub> (head = 12)
Batch size	32
Retrieval encoder	ResNet 50x64
Max caption length	100
Beam search	7
Learning rate	1e-4
Epoch	20
Number of caption	4
Optimizer	AdamW
Image size	224x224 px

*Bảng 2 . Bảng tham số huấn luyện mô hình*

## 4.4 ĐÁNH GIÁ ĐỘ CHÍNH XÁC

Chỉ số BLEU đánh giá được tính bằng cách đánh giá chỉ số BLEU-1, BLEU-2, BLEU-3, BLEU-4 trung bình trên 1.000 hình ảnh của tập test, so sánh giữa câu mô tả được dự đoán từ mô hình vừa xây dựng và 5 câu mô tả trong tập dữ liệu ban đầu của ảnh đó, kết quả đo thể hiện ở bảng 3.

MÔ HÌNH THỰC NGHIỆM	BLEU-1	BLEU-2	BLEU-3	BLEU-4
SmallCap <sub>BASE</sub> tiếng Anh	59,26	38,62	24,28	14,63
SmallCap <sub>BASE</sub> tiếng Việt	64,88	51,87	41,4	32,13
SmallCap modify deep-attention tiếng Anh	73,9	56,69	41,54	30,57
SmallCap modify deep-attention tiếng Việt	72,39	59,79	49,2	39,51



*Bảng 3. Bảng kết quả đánh giá BLEU cho mô hình xây dựng câu mô tả cho hình ảnh*

## 4.5 THẢO LUẬN KẾT QUẢ THỰC NGHIỆM




Từ kết quả thực nghiệm ở Bảng 1, Bảng 2: Bảng kết quả đánh giá BLEU cho mô hình sinh câu mô tả ảnh trên tập Flickr8k tiếng Việt và tiếng Anh, chúng tôi nhận thấy rằng trên các phương pháp thực hiện mô hình SmallCap với deep attention giúp đạt điểm BLEU tốt hơn trên cả tập dữ liệu Flickr8k tiếng Anh và Tiếng Việt so với mô hình với cross attention trong điều kiện tương tự. Ngoài ra khi so với mô hình của chúng tôi với mô hình VGG16 + LTSM trên tập Flickr8k tiếng Việt của Lam [4] và cộng sự. Hiện tại, các điểm BLEU-1,2,3 và 4 của mô hình tốt nhất, sử dụng deep attention thay cho cross attention lần lượt là :72.39 , 59.79 , 49.2, 39.51; trong khi các điểm BLEU của Lam và cộng sự lần lượt là 62.9, 42.6, 28.1 và 17.5. Cả mô hình chưa tinh chỉnh và đã tinh chỉnh



của chúng tôi đều có thể xây dựng câu mô tả cho hình ảnh nhưng đối với các trường hợp ảnh có nhiều đối tượng, có nhiều chi tiết khó nhận biết thì mô hình đã được tinh chỉnh mô tả chi tiết hơn, được thể hiện ở một số hình ảnh minh họa kết quả sinh mô tả hình ảnh của mô hình như sau (bảng 4 ):

IMAGE	MODEL	VIETNAMESE	ENGLISH
	SmallCap <sub>base</sub>	Một vận động viên trượt tuyết đang bay trên tuyết	a skier is skiing down a snowy hill
	SmallCap Modify deep-attention	Một người đàn ông mặc áo khoác xanh và đội mũ bảo hiểm đang trượt tuyết	a person in a blue jacket is skiing
	SmallCap <sub>base</sub>	Một cô bé leo xung quanh một tòa nhà bằng gỗ	a little girl climbing into a wooden cabin
	SmallCap Modify deep-attention	Một cô bé mặc váy hồng đi vào một tòa nhà bằng gỗ	a little girl climbing the stairs to her home



	SmallCap <sub>base</sub>	Một cô gái trẻ đang ngồi trên cát ở bãi biển	a young girl sitting in the sand on a beach
	SmallCap Modify deep-attention	Một cô gái trẻ ngồi trên cát trên bãi biển	a little girl wearing a blue bikini sits on a sand beach
	SmallCap <sub>base</sub>	Một người đàn ông đang đứng trên đỉnh núi	a man stands on a mountaintop looking into the distance
	SmallCap Modify deep-attention	Một người đàn ông đứng trên đỉnh núi	a man is standing on a rock overlooking a sunset
	SmallCap <sub>base</sub>	Một người đàn ông đứng cạnh bức tường vẽ bậy	a man walks through the graffiti wall
	SmallCap Modify deep-attention	Người đàn ông đang đứng trên một bức tường bị vẽ bậy	a man in a white t shirt is standing by a wall of colorful graffiti

	SmallCap <sub>base</sub>	Một chiếc xe lớn trên núi	a mountain landscape
	SmallCap Modify deep-attention	Một nhóm người với thiết bị cắm trại trên núi	a road with mountains in the background
	SmallCap <sub>base</sub>	Một nhóm trẻ em chơi trong bong bóng	children play in bubbles
	SmallCap Modify deep-attention	Trẻ em chơi với những quả bóng màu xanh lá cây trên cánh đồng	children playing with bubbles
	SmallCap <sub>base</sub>	Một số người đang đứng trên một mỏm đá vào lúc hoàng hôn	the sky above the ocean
	SmallCap Modify deep-attention	Những người trên một chiếc tàu lượn siêu tốc	a jet is flying over the ocean

Bảng 4. Mô tả hình ảnh với mô hình SmalCap modify deep attention và SmallCap Base

## CHƯƠNG 5. KẾT LUẬN

Trong chương này sẽ trình bày tóm tắt kết quả nghiên cứu đã đạt được đồng thời nêu lên hướng phát triển tiếp theo trong tương lai nhằm mang lại hiệu quả tốt hơn cho đề tài nghiên cứu.

### 5.1. KẾT QUẢ ĐẠT ĐƯỢC

Sau khi tham khảo các nguồn tài liệu, nghiên cứu các phương pháp giải quyết đề tài đã đặt ra ban đầu, chúng tôi đã nắm bắt được các nội dung cơ sở lý thuyết về mô hình CLIP dùng để rút trích đặc trưng hình ảnh cũng như mô hình GPT cho xây dựng mô tả ảnh. Chúng tôi đã tiến hành xây dựng và đánh giá thực nghiệm mô hình xây dựng mô tả ảnh trên tập dữ liệu Flickr8k tiếng Anh và tiếng Việt. Kết quả đánh giá tương đối khả quan huấn luyện trên tập dữ liệu Flickr8k tiếng Việt và tiếng Anh khá tương đồng.

### 5.2 HƯỚNG PHÁT TRIỂN

Hiện tại, mô hình xây dựng mô tả ảnh còn một số điểm hạn chế, chưa thật sự tạo nên câu mô tả hoàn hảo. Trong thời gian tới, chúng tôi sẽ tiến hành nghiên cứu áp dụng phương pháp mới vào việc rút trích đặc trưng hình ảnh để cải thiện chất lượng dữ liệu đầu vào tạo tiền đề tốt hơn cho việc nhận dạng các đối tượng trong hình ảnh. Ngoài ra, chúng tôi sẽ nâng cao chất lượng của bộ dữ liệu đào tạo bằng cách cải tiến mô hình dịch và sẽ hướng tới nhiều tập dữ liệu khác nhau chẳng hạn như MSCOCO [19]. Từ đó giúp cho câu mô tả dự đoán ra được sẽ bám sát vào nội dung hình ảnh đề cập.

## TÀI LIỆU THAM KHẢO

- [1] Xu Yang, Kaihua Tang, Hanwang Zhang, Jianfei Cai, "Auto-Encoding Scene Graphs for Image Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019*, 2019.
- [2] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Advances in neural information processing systems* 28, 2015.
- [3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [4] Khang Nhut Lam, Kim-Ngoc Thi Nguyen, Loc Huu Nguy, Jugal Kalita, "Facial Expression Recognition and Image Description Generation in Vietnamese," *Fuzzy Systems and Data Mining VII: Proceedings of FSDM 2021* 340 , 2021.
- [5] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *arXiv:1506.02640* , 2016.
- [6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention Is All You Need," *arXiv:1706.03762*, 2017.

- [7] Ramos, Rita and Martins, Bruno and Elliott, Desmond and Kementchedjieva, Yova, "SmallCap: Lightweight Image Captioning Prompted with Retrieval Augmentation," *CVPR*, 2023.
- [8] 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 Proceedings of the International Conference on Machine Learning*, 2021.
- [9] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave, "Few shot learning with retrieval augmented language models.," *arXiv preprint arXiv:2208.03299*, 2022.
- [10] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, Xiang Ren, "A Good Prompt Is Worth Millions of Parameters: Low-resource Prompt-based Learning for Vision-Language Models," *arXiv:2110.08484* , 2019.
- [11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *In Proceedings of the Annual Conference on Neural Information Processing Systems*, 2020.
- [12] Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu, "A survey on retrieval-augmented text generation," *arXiv preprint arXiv:2202.01110*, 2022.

- [13] Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng, "Training data is more valuable than you think: A simple and effective method by retrieving from training data," *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2022.
- [14] Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cuc-, "Retrieval-augmented transformer for image caption," *ARPV*, 2022.
- [15] Radford, Alec and Wu, Jeff and Child, Rewon and Luan, David and Amodei, Dario and Sutskever, Ilya, "Language Models are Unsupervised Multitask Learners," *radford2019language*, 2019.
- [16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," *ACL*, 2002.
- [17] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, Jiashi Feng, "DeepViT: Towards Deeper Vision Transformer," *arXiv:2103.11886*, 2021.
- [18] Ilya Loshchilov and Frank Hutter. , "Fixing weight decay regularization in adam," *arXiv preprint arXiv:1711.05101*, 2018.
- [19] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollár, "Microsoft COCO: Common Objects in Context," in *Computer Vision–ECCV 2014: 13th European Conference*, Zurich, Switzerland.