**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN CUỐI KÌ**

**COMPUTER VISION**

**AUTOMATIC CAPTION GENERATION FOR FASHION IMAGES USING TRANSFORMER-BASED METHODS**

*Người thực hiện*: **VÕ QUỐC HUY – 20081001**

**NGUYỄN ĐÌNH NGUYÊN BẮC – 21115711**

*Người hướng dẫn*: **LƯU GIANG NAM**

**TRẦN TẤN THÀNH**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN CUỐI KÌ**

**COMPUTER VISION**

**AUTOMATIC CAPTION GENERATION FOR FASHION IMAGES USING TRANSFORMER-BASED METHODS**

Người thực hiện: **VÕ QUỐC HUY- 20081001**

**NGUYỄN ĐÌNH NGUYÊN BẮC- 21115711**

Người hướng dẫn: **LƯU GIANG NAM**

**TRẦN TẤN THÀNH**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

LỜI CẢM ƠN

Chúng tôi xin gửi lời cảm ơn chân thành nhất đến quý thầy Lưu Giang Nam và thầy Trần Tấn Thành – Giảng viên đã tận tình hướng dẫn cho chúng tôi trong suốt quá trình học tập môn Thị giác máy tính và tạo điều kiện cho chúng tôi làm Đồ án này. Dưới sự hướng dẫn của Thầy chúng tôi được tiếp cận với những kiến thức chuyên môn sâu rộng trong lĩnh vực thị giác máy tính. Qua đó có thể hoàn thành Đồ án này một cách thuận lợi.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của thầy Lưu Giang Nam và thầy Trần Tấn Thành. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Công nghiệp TP Hồ Chí Minh không liên quan đến những vi phạm tác quyền, bản quyền do chúng tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Võ Quốc Huy*

*Nguyễn Đình Nguyên Bắc*

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT NỘI DUNG

**Tiêu đề:** Sinh caption tự động cho hình ảnh thời trang sử dụng các phương pháp học transformer.

**Tóm tắt:**

Với sự phát triển của Thương mại điện tử và nhu cầu tự động hóa trong Công nghệ thông tin ngày càng tăng, đặc biệt trong ngành công nghiệp Thời trang, nơi các công ty quần áo cung cấp các danh mục trực tuyến tràn ngập hình ảnh và chú thích để khách hàng có thể tìm thấy món đồ mong muốn. Chú thích hình ảnh thời trang nhằm mục đích tự động tạo mô tả sản phẩm cho các mặt hàng thời trang, thu hút sự chú ý của khách hàng, giúp người mua hiểu rõ hơn về hàng hóa, từ đó tăng doanh thu cho công ty may mặc. Tuy nhiên, việc scan, render, tạo chú thích các mặt hàng thời trang này hầu hết vẫn được thực hiện thủ công và đòi hỏi rất nhiều thời gian để thực hiện. Trong nghiên cứu này, chúng tôi đề xuất một mô hình tạo chú thích tự động cho hình ảnh thời trang dựa trên các kiến trúc Transformer là BEiT và GPT2. BEiT được sử dụng để trích xuất đặc trưng hình ảnh từ hình ảnh thời trang, trong khi GPT2 được sử dụng để tạo ra chú thích mô tả hình ảnh một cách chính xác và sáng tạo. Chúng tôi thực hiện đánh giá trên bộ dữ liệu Fashion Captioning Dataset (FACAD) bằng các độ đo như ROUGE, BLEU, METEOr. Kết quả các thử nghiệm cho thấy phương pháp đề xuất của chúng tôi có hiệu suất đáng kể hơn so với các phương pháp trước đây.

**Từ khóa:** Tạo chú thích tự động, Transformer, BEiT, GPT2, Chú thích ảnh thời trang

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc167561588)

[PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc167561589)

[TÓM TẮT NỘI DUNG iv](#_Toc167561590)

[MỤC LỤC 1](#_Toc167561591)

[DANH MỤC CHỮ VIẾT TẮT 3](#_Toc167561592)

[DANH MỤC CÁC HÌNH VẼ 4](#_Toc167561593)

[DANH MỤC CÁC BẢNG 5](#_Toc167561594)

[CHƯƠNG 1: GIỚI THIỆU 6](#_Toc167561595)

[1.1 Tổng quan 6](#_Toc167561596)

[1.2 Ý nghĩa bài toán 6](#_Toc167561597)

[1.3 Sự đóng góp 7](#_Toc167561598)

[1.4 Cấu trúc đồ án 8](#_Toc167561599)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 9](#_Toc167561600)

[2.1 Phân tích yêu cầu bài toán 9](#_Toc167561601)

[2.2 Các nghiên cứu liên quan 9](#_Toc167561602)

[2.2.1 Image Captioning 9](#_Toc167561603)

[2.2.2 Transformer Based Method 11](#_Toc167561604)

[2.3 Phương pháp đề xuất giải quyết bài toán 12](#_Toc167561605)

[2.4 Mô hình tổng quát 13](#_Toc167561606)

[2.5 Đặc trưng của mô hình đề xuất 13](#_Toc167561607)

[2.5.1 Encoder 13](#_Toc167561608)

[2.5.1.1 BEiT 13](#_Toc167561609)

[2.5.1.2 ViT 15](#_Toc167561610)

[2.5.2 Decoder 16](#_Toc167561611)

[2.5.2.1 BERT 16](#_Toc167561612)

[2.5.2.2 GPT-2 17](#_Toc167561613)

[CHƯƠNG 3: THỰC NGHIỆM 19](#_Toc167561614)

[3.1 Dữ liệu 19](#_Toc167561615)

[3.2 Cách đánh giá 20](#_Toc167561616)

[3.3 Kết quả 21](#_Toc167561617)

[CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 24](#_Toc167561618)

[4.1 Kết quả đạt được 24](#_Toc167561619)

[4.2 Hạn chế 24](#_Toc167561620)

[4.3 Hướng phát triển 24](#_Toc167561621)

[TÀI LIỆU THAM KHẢO 26](#_Toc167561622)

DANH MỤC CHỮ VIẾT TẮT

|  |  |
| --- | --- |
| AI | Artificial Intelligence |
| GPT-2 | Generative Pre-trained Transformer 2 |
| BeiT | Bidirectional Encoder Representation from Image Transformers |
| ViT | Vision Transformer |
| BERT | Bidirectional Encoder Representations from Transformers |
| ROUGE | Recall-Oriented Understudy for Gisting Evaluation |
| BLEU | Bilingual Evaluation Understudy |
| METEOR | Metric for Evaluation of Translation with Explicit Ordering |

DANH MỤC CÁC HÌNH VẼ

[**Hình 1**: Mô hình tổng quát 10](#_Toc167557423)

[**Hình 2**: Kiến trúc BEiT 12](#_Toc167557424)

[**Hình 3**: Kiến trúc ViT 14](#_Toc167557425)

[**Hình 4**: Kiến trúc BERT 15](#_Toc167557426)

[**Hình 5**: Kiến trúc GPT-2 17](#_Toc167557427)

DANH MỤC CÁC BẢNG

[**Bảng 1**: Mô tả dữ liệu 18](#_Toc167557430)

[**Bảng 2**: Kết quả so sánh giữa các mô hình 20](#_Toc167557431)

[**Bảng 3**: Kết quả so sánh giữa mô hình đề xuất và các mô hình tiên tiến 21](#_Toc167557432)

[**Bảng 4**: Kết quả thực nghiệm 22](#_Toc167557433)

# GIỚI THIỆU

## Tổng quan

Sự tiến bộ nhanh chóng của các kỹ thuật học sâu đã làm cách mạng hóa lĩnh vực thị giác máy tính, tạo ra những tiến bộ đáng kể trong nhiều ứng dụng khác nhau như nhận dạng hình ảnh, phát hiện vật thể và chú thích hình ảnh. Trong số những ứng dụng này, việc tạo chú thích tự động đã thu hút sự quan tâm đáng kể bởi tiềm năng nâng cao trải nghiệm người dùng trên nhiều lĩnh vực, bao gồm mạng xã hội, thương mại điện tử và hỗ trợ người khiếm thị. Trong những năm gần đây, công nghệ thị giác máy tính và các kỹ thuật học sâu với khả năng trích xuất đặc trưng mạnh mẽ đã được sử dụng rộng rãi trong các ứng dụng thời trang, bao gồm phát hiện quần áo [1], phân tích quần áo [2], trích xuất quần áo [3], thử quần áo [4], ...

Chú thích hình ảnh thời trang đóng một vai trò quan trọng trong môi trường kỹ thuật số, đặc biệt là trong lĩnh vực thương mại điện tử và mạng xã hội. Các chú thích chất lượng cao không chỉ cải thiện và tối ưu hóa các công cụ tìm kiếm mà còn tăng cường sự tương tác của người dùng bằng cách cung cấp các mô tả có ý nghĩa, đáp ứng nhu cầu mua hàng của mọi người nhằm gia tăng doanh số bán hàng của doanh nghiệp. Các phương pháp tạo chú thích truyền thống, thường dựa trên các cách tiếp cận dựa trên mẫu hoặc quy tắc, không thể nắm bắt được bản chất tinh tế và năng động của các xu hướng và phong cách thời trang. Do đó, cần có các phương pháp tinh vi hơn có thể thích ứng với sự phát triển của ngành thời trang và tạo ra các chú thích mô tả, chính xác và hấp dẫn.

## Ý nghĩa bài toán

Chú thích hình ảnh thời trang là một thử thách hấp dẫn và mới mẻ trong lĩnh vực thị giác máy tính và xử lý ngôn ngữ tự nhiên. Mục tiêu của nó là tạo ra những mô tả văn bản ngắn gọn, chính xác và thu hút cho các hình ảnh thời trang. Điều này đòi hỏi mô hình AI phải hiểu được nội dung của hình ảnh, bao gồm các yếu tố như trang phục, phụ kiện, phong cách, màu sắc, và bối cảnh, sau đó chuyển đổi thông tin này thành ngôn ngữ tự nhiên. Ứng dụng của bài toán này rất đa dạng, từ việc cải thiện trải nghiệm mua sắm trực tuyến, tự động tạo mô tả sản phẩm, hỗ trợ người khiếm thị, đến việc tạo nội dung sáng tạo trên mạng xã hội. Tuy nhiên, để giải quyết bài toán này một cách hiệu quả, mô hình cần phải vượt qua nhiều thách thức, bao gồm sự đa dạng của thời trang, tính chủ quan của nhận thức thẩm mỹ, và khả năng diễn đạt ngôn ngữ phong phú.

## Sự đóng góp

Một trong những phương pháp tiên tiến nhất hiện nay để tạo ra chú thích tự động là sử dụng các mô hình Transformer dựa trên kiến trúc mã hóa-giải mã (encoder-decoder). Các mô hình này đã chứng tỏ được khả năng vượt trội trong việc xử lý ngôn ngữ tự nhiên và tạo văn bản, nhờ vào khả năng học sâu và xử lý song song hiệu quả. Đặc biệt, trong lĩnh vực thời trang, nơi mà hình ảnh thường phức tạp và đa dạng về màu sắc, kiểu dáng, chất liệu và các phụ kiện, các phương pháp này hứa hẹn mang lại những tiến bộ đáng kể. Trong đồ án này chúng tôi tập trung vào việc ứng dụng phương pháp mã hóa-giải mã dựa trên Transformer để tự động tạo ra các chú thích cho hình ảnh thời trang. Chúng tôi sẽ trình bày chi tiết về kiến trúc mô hình, phương pháp huấn luyện, và đánh giá hiệu suất của chúng trên các bộ dữ liệu thực tế. Mục tiêu của chúng tôi là cung cấp một cái nhìn toàn diện về tiềm năng của các phương pháp này và định hướng cho các nghiên cứu trong tương lai. Tóm lại, sau đây là những đóng góp của chúng tôi trong đồ án này:

* Chúng tôi đề xuất mô hình cải tiến dựa trên transformer encoder-decoder. Mô hình này được thiết kế để tối ưu hóa việc hiểu ngữ cảnh và mô tả chi tiết các đặc điểm thời trang.
* Chúng tôi tiến hành các thực nghiệm và đánh giá đầy đủ trên tập dữ liệu FACAD nhằm chứng minh tính hiệu quả và độ ổn định của mô hình đề xuất của mình.
* Chúng tôi so sánh, đánh giá phương pháp của chúng tôi với các phương pháp hiện tại, qua đó chứng minh tính ưu việt trong phương pháp của chúng tôi

## Cấu trúc đồ án

Ngoài phần giới thiệu, các phần còn lại của đồ án sẽ bao gồm những nội dung sau: Cơ sở lí thuyết sẽ được trình bày ở Phần 2. Thực nghiệm và đánh giá sẽ nằm trong Phần 3. Phần 4 sẽ đưa ra những kết luận về nghiên cứu của chúng tôi và hướng phát triển của nghiên cứu trong tương lai.

# CƠ SỞ LÝ THUYẾT

## Phân tích yêu cầu bài toán

Bài toán này yêu cầu một hệ thống AI có khả năng phân tích hình ảnh thời trang ở mức độ chi tiết, nhận diện và hiểu được các yếu tố như kiểu dáng trang phục (váy, áo khoác, quần jeans), chất liệu (ren, lụa, denim), họa tiết (hoa, kẻ sọc, chấm bi), màu sắc (đỏ tươi, xanh navy, pastel), phụ kiện (túi xách, giày, mũ),... Sau đó, hệ thống phải diễn đạt thông tin này thành một đoạn văn ngắn, súc tích, sử dụng ngôn ngữ tự nhiên, mô tả chính xác và hấp dẫn nội dung của hình ảnh. Đoạn mô tả này không chỉ liệt kê các yếu tố thời trang mà còn thể hiện được sự kết hợp hài hòa giữa chúng, tạo nên một bức tranh tổng thể về phong cách và ấn tượng của bộ trang phục. Ngoài ra, hệ thống cần phải linh hoạt để thích ứng với sự đa dạng của thời trang, từ phong cách thường ngày đến các xu hướng mới nhất, và có khả năng tùy chỉnh mô tả theo nhu cầu cụ thể của người dùng hoặc ứng dụng.

## Các nghiên cứu liên quan

### Image Captioning

Chú thích hình ảnh là một tác vụ phức tạp liên quan đến việc tạo mô tả cho hình ảnh, đã được Matteo Stefanini et al. [5] tổng hợp và đánh giá toàn diện về các phương pháp, chỉ số đánh giá và bộ dữ liệu thực nghiệm. Để tạo chú thích hình ảnh một cách chính xác, cần xác định các đối tượng, thuộc tính và mối quan hệ của chúng trong ảnh, đồng thời cần phải đảm bảo tính đúng đắn về ngữ pháp và ngữ nghĩa của câu được tạo ra. Đã có nhiều nghiên cứu về chú thích hình ảnh, chẳng hạn như [6,7,8,9].

Sự phát triển của học sâu đã cải thiện đáng kể độ chính xác trong chú thích hình ảnh. Các thuật toán học sâu có khả năng tự động học các đặc trưng hình ảnh từ dữ liệu huấn luyện, cho phép xử lý một lượng lớn hình ảnh và giải quyết các thách thức của nhiệm vụ này. Mô hình chú thích hình ảnh dựa trên học sâu thường kết hợp hai thành phần chính là bộ mã hóa CNN [10] và bộ giải mã RNN [11]. Trong khi CNN trích xuất đặc trưng biểu diễn của toàn bộ hình ảnh, RNN tạo ra chuỗi từ mô tả. Sự kết hợp giữa CNN và RNN đã được chứng minh tính hiệu quả trong các nghiên cứu của Simonyan và Zisserman [12], Karpathy và Fei-Fei [13], Vinyals et al. [14] trong việc giúp xác định các mẫu trong hình ảnh và tạo các mô tả tương ứng. Mặc dù sự kết hợp giữa CNN và RNN đã đem lại những bước tiến quan trọng trong lĩnh vực chú thích hình ảnh, tuy nhiên các hạn chế về khả năng tổng quát hóa và độ chính xác vẫn còn tồn tại. Chẳng hạn như khi CNN trích xuất vectơ đặc trưng từ hình ảnh, biểu diễn bên trong chứa quá nhiều thông tin để RNN giải mã thành các câu mô tả. Một vấn đề khác là sự phụ thuộc của mô hình vào các cặp dữ liệu hình ảnh-chú thích có sẵn, điều này có thể giới hạn khả năng mô hình hóa và tạo ra các mô tả mới và sáng tạo. Hơn nữa, các mô hình hiện tại thường thiếu khả năng hiểu ngữ cảnh sâu rộng và thường không thể tạo ra các mô tả mang tính ngữ nghĩa cao hoặc phản ánh đầy đủ các chi tiết trong hình ảnh.

Để khắc phục những vấn đề này, các nghiên cứu gần đây đã chuyển sang các phương pháp tiên tiến hơn, trong đó nổi bật là việc sử dụng cơ chế attention [15]. Cơ chế này cho phép mô hình tập trung vào các phần quan trọng của hình ảnh khi tạo mô tả, từ đó cải thiện chất lượng và độ chính xác của chú thích. Những nghiên cứu gần đây, như của Nguyen Bao Thien et al. (2020) [16], đã áp dụng thành công cơ chế attention vào mô hình chú thích hình ảnh, cho thấy sự tiến bộ vượt bậc so với các phương pháp truyền thống. Tuy nhiên, sự xuất hiện của các mô hình Transformers đã nổi lên như một giải pháp tiên tiến, cung cấp khả năng xử lý thông tin vượt trội và khả năng tự học từ dữ liệu một cách hiệu quả đã mở ra một hướng tiếp cận mới và mạnh mẽ hơn, đặc biệt là trong các nhiệm vụ liên quan đến ngôn ngữ và hình ảnh.

### Transformer Based Method

Các mô hình Transformers được giới thiệu bởi Vaswani et al. (2017) [17], với cấu trúc attention mạnh mẽ, đã chứng minh hiệu quả vượt trội trong nhiều tác vụ liên quan đến ngôn ngữ tự nhiên và gần đây đã được áp dụng thành công vào nhiệm vụ chú thích hình ảnh. Không giống như các mô hình truyền thống dựa trên CNN và RNN, Transformers có khả năng xử lý toàn bộ chuỗi dữ liệu cùng một lúc, giúp nắm bắt được các mối quan hệ dài hạn và các tương tác phức tạp trong dữ liệu hình ảnh và văn bản. Trong lĩnh vực chú thích hình ảnh, các nghiên cứu như của Cornia et al. (2020) [18] đã áp dụng mô hình Transformers để cải thiện khả năng tạo mô tả hình ảnh. Mô hình này sử dụng một kiến trúc kết hợp giữa CNN để trích xuất đặc trưng hình ảnh và Transformer để tạo chuỗi từ mô tả, cho phép khai thác các mối quan hệ ngữ cảnh mạnh mẽ hơn giữa các đối tượng trong ảnh. Nghiên cứu của Li et al. (2019) [19] với mô hình VisualBERT đã chứng minh khả năng kết hợp thông tin hình ảnh và ngôn ngữ một cách liền mạch. VisualBERT sử dụng các đặc trưng từ hình ảnh và văn bản đồng thời trong cùng một mô hình Transformer, cải thiện khả năng hiểu và tạo ra các mô tả ngữ nghĩa chính xác và tự nhiên hơn. Ngoài ra, Dosovitskiy et al. (2020) [20] với mô hình Vision Transformer (ViT) đã cho thấy rằng Transformers có thể thay thế hoàn toàn CNN trong việc trích xuất đặc trưng hình ảnh, mở ra một hướng đi mới cho việc tích hợp sâu hơn giữa hình ảnh và ngôn ngữ. ViT học các đặc trưng từ các mảng nhỏ của hình ảnh thay vì toàn bộ hình ảnh, giúp mô hình tập trung vào các chi tiết quan trọng và cải thiện khả năng chú thích. Nhìn chung, các phương pháp dựa trên Transformers đã chứng minh tiềm năng lớn trong việc cải thiện chất lượng và độ chính xác của chú thích hình ảnh. Sự linh hoạt và mạnh mẽ của kiến trúc này cho phép mô hình hóa các mối quan hệ phức tạp và tạo ra các mô tả tự nhiên và phù hợp hơn, vượt trội so với các phương pháp truyền thống.

## Phương pháp đề xuất giải quyết bài toán

Từ những nghiên cứu đã có trước đây, trong đồ chúng tôi đề xuất một mô hình cho nhiệm vụ chú thích hình ảnh thời trang dựa trên kiến trúc Transformers encoder – decoder. Cụ thể, chúng tôi sử dụng mô hình BEiT (Bidirectional Encoder Representation from Image Transformers) [21] làm bộ mã hóa để trích xuất các đặc trưng từ hình ảnh thời trang. BEiT được sử dụng vì khả năng mạnh mẽ trong việc học các đặc trưng ngữ nghĩa từ hình ảnh, vượt trội so với các mô hình CNN truyền thống. Tiếp theo, chúng tôi sử dụng GPT-2 (Generative Pre-trained Transformer 2) [22] làm bộ giải mã để tạo ra các mô tả văn bản từ các đặc trưng hình ảnh này. GPT-2 đã được chứng minh là có khả năng tạo ra các văn bản tự nhiên và mạch lạc, nhờ vào khả năng học sâu các mối quan hệ ngữ cảnh trong văn bản. Mô hình đề xuất của chúng tôi không chỉ tận dụng sức mạnh của BEiT trong việc hiểu các đặc trưng của hình ảnh mà còn khai thác GPT-2 để tạo ra các chú thích phong phú và chính xác. Bằng cách kết hợp hai thành phần mạnh mẽ này, mô hình đề xuất có thể vượt qua các hạn chế của các phương pháp trước đây, cải thiện cả về mặt độ chính xác lẫn tính ngữ nghĩa của các chú thích hình ảnh thời trang. Mô hình tổng quát của chúng tôi sẽ được trình bày trong phần tiếp theo.

## Mô hình tổng quát

Chúng tôi đề xuất một mô hình cho nhiệm vụ chú thích hình ảnh thời trang, sử dụng kiến trúc Transformers encoder – decoder (BEiT - GPT2). Mô hình này bao gồm hai phần chính: bộ mã hóa BEiT và bộ giải mã GPT-2. Mô hình tổng quát của chúng tôi được trình bày ở Hình 1.

Input Image

Decoder

Output Caption

BERT

Encoder

ViT

BEiT

GPT2

**Hình 1**: Mô hình tổng quát

## Đặc trưng của mô hình đề xuất

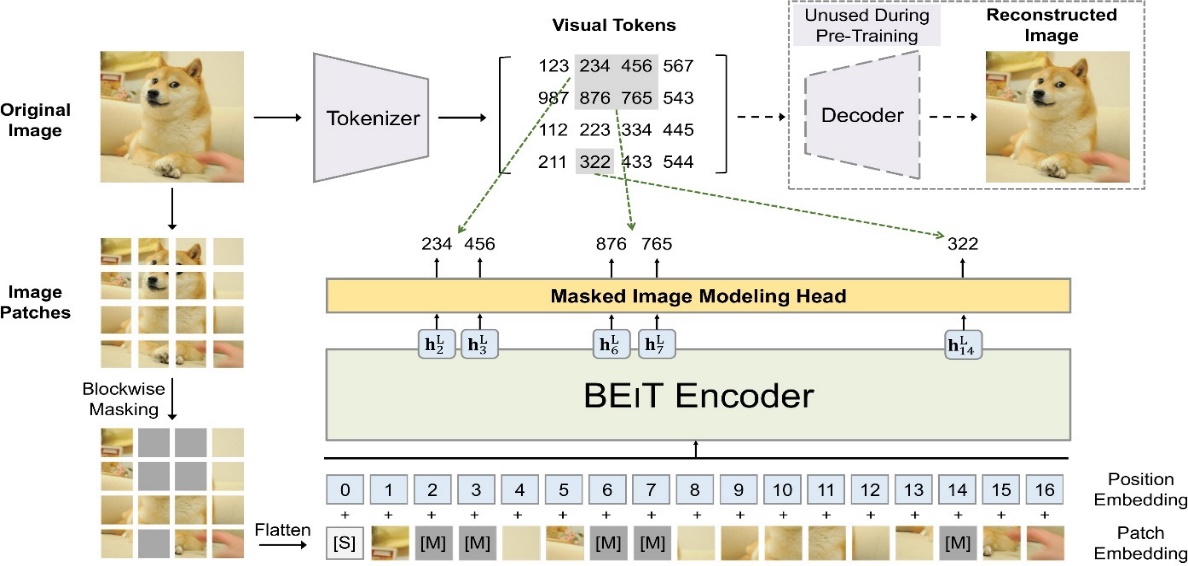
### Encoder

#### BEiT

BEiT (Bidirectional Encoder representation from Image Transformers) là một mô hình học sâu tiên tiến dựa trên kiến trúc Transformer, được thiết kế để xử lý các nhiệm vụ liên quan đến hình ảnh. BEiT được phát triển nhằm mục đích chuyển đổi và cải thiện khả năng hiểu và xử lý thông tin hình ảnh tương tự như cách BERT đã cách mạng hóa xử lý ngôn ngữ tự nhiên (NLP). BEiT sử dụng một kiến trúc tương tự như Transformer encoder, nhưng được điều chỉnh để làm việc với dữ liệu hình ảnh. Thay vì làm việc với các từ hoặc token, BEiT chia nhỏ hình ảnh thành các mảnh nhỏ gọi là "patches" và mã hóa chúng thành các vectơ đầu vào cho mô hình.

Mỗi patch của hình ảnh được coi như một token trong mô hình NLP và được xử lý bởi các lớp Transformer encoder. Các lớp này bao gồm các lớp self-attention, giúp mô hình tập trung vào các phần quan trọng của hình ảnh và học cách các phần này tương tác với nhau. Quá trình mã hóa này giúp mô hình hiểu rõ hơn về cấu trúc và nội dung của hình ảnh

BEiT được huấn luyện trước (pre-trained) trên một tập dữ liệu hình ảnh lớn bằng cách sử dụng một phương pháp tương tự như masked language modeling (MLM) trong NLP. Trong quá trình này, một phần các patch của hình ảnh được che đi và mô hình phải dự đoán các patch bị che dựa trên ngữ cảnh từ các patch không bị che. Phương pháp này giúp mô hình học được các đặc trưng quan trọng của hình ảnh mà không cần nhãn cụ thể cho từng ảnh. Sau quá trình pre-training, BEiT có thể được tinh chỉnh (fine-tuned) cho các nhiệm vụ cụ thể như phân loại hình ảnh, phát hiện đối tượng, và phân đoạn hình ảnh. Quá trình tinh chỉnh này sử dụng một tập dữ liệu được gán nhãn cho nhiệm vụ cụ thể, giúp mô hình cải thiện hiệu suất trên nhiệm vụ đó.



**Hình 2**: Kiến trúc BEiT

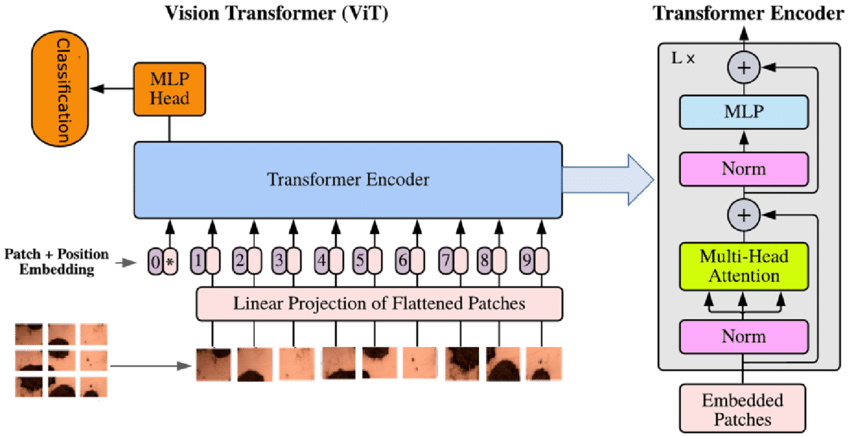
#### ViT

CNNs đã được sử dụng làm công nghệ tiêu chuẩn trong các tác vụ thị giác máy tính kể từ AlexNet (2012). Khi xây dựng một mô hình học máy xử lý dữ liệu hình ảnh, CNNs tránh sử dụng các đặc trưng được thiết kế thủ công mà thay vào đó học trực tiếp các đặc trưng hình ảnh từ dữ liệu. Các kiến trúc được xây dựng dựa trên CNNs hiệu quả trong xử lý ảnh, nhưng được thiết kế riêng cho loại dữ liệu này và có thể đòi hỏi nhiều tài nguyên tính toán khi mục tiêu là xây dựng các mô hình thị giác có khả năng mở rộng.

Khi xử lý dữ liệu văn bản, mô hình Transformer được đề xuất trong bởi Vaswani et al. (2017) hiệu quả về mặt tính toán và có khả năng mở rộng, cho phép huấn luyện các mô hình lớn trên lượng dữ liệu văn bản khổng lồ. Từ năm 2020, các nhà nghiên cứu bắt đầu tự hỏi liệu có thể khai thác một kiến trúc giống Transformer để xử lý dữ liệu hình ảnh hay không.

Vision Transformer (ViT) là một mô hình thị giác dựa trên kiến trúc Transformer. Mục tiêu của mô hình là xử lý một hình ảnh đầu vào như thể nó là một đầu vào văn bản, thực hiện ít thay đổi nhất có thể đối với mô hình Transformer cổ điển. Các kết quả đã chứng minh rằng ViT đạt được hiệu suất tuyệt vời khi được đào tạo trên dữ liệu đủ lớn, vượt trội hơn CNN tiên tiến tương đương trong khi đồng thời hiệu quả hơn (nó yêu cầu ít hơn bốn lần tài nguyên tính toán).

ViT khai thác một chồng các khối mã hóa, mỗi khối có cấu trúc được trình bày trong mô hình gốc. Ảnh đầu vào được chia thành một lưới vuông các “patches”; mỗi “patches” được làm phẳng thành một vectơ duy nhất bằng cách nối các kênh màu và sau đó chiếu vectơ lên chiều mong muốn thông qua một lớp tuyến tính. Một vectơ vị trí có thể học được (learnable positional embedding) được thêm vào mỗi “patches”, giúp mô hình có thể học cấu trúc của ảnh thông qua thông tin về vị trí của các “patches”. Chuỗi vectơ cuối cùng được đưa vào bộ mã hóa Transformer tiêu chuẩn. Để thực hiện phân loại các ảnh đầu vào, một mã thông báo [class] có thể học được được thêm vào trước chuỗi các “patches” được nhúng; lớp cuối cùng là đầu ra của một tầng dự đoán trên khối mã hóa cuối cùng. ViT được huấn luyện trước trên các tập dữ liệu lớn và tinh chỉnh cho các tác vụ nhỏ hơn; bước tinh chỉnh yêu cầu loại bỏ tầng dự đoán đã được huấn luyện trước bằng một tầng feedforward mới được khởi tạo.



**Hình 3**: Kiến trúc ViT

### Decoder

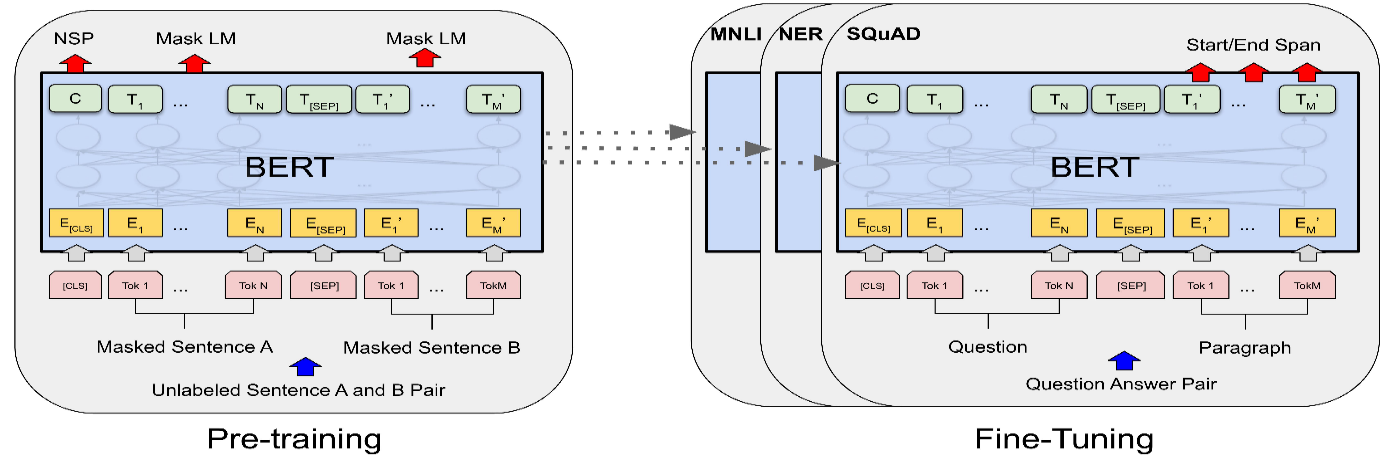
#### BERT

Với sự ra đời của kiến trúc Transformer, cộng đồng nghiên cứu đã bắt đầu phân tích các biến thể của mô hình ban đầu, và sự hiểu biết về cách biểu diễn câu và từ đã nhanh chóng phát triển. Bidirectional Encoder Representations from Transformers (BERT), là một phương pháp tiền huấn luyện biểu diễn ngôn ngữ tiên tiến. BERT học một mô hình Transformer đa dụng được huấn luyện trên một tập dữ liệu khổng lồ. Kiến trúc của BERT chỉ bao gồm ngăn mã hóa của Transformer, sử dụng nhiều đầu chú ý và khối mã hóa hơn so với mô hình Transformer ban đầu.

Giai đoạn tiền huấn luyện là bước đầu tiên trong quá trình học của mô hình. Trong giai đoạn này, mô hình học cách biểu diễn ngôn ngữ, trong đó các từ được điều kiện hóa bởi cả ngữ cảnh trước và sau. BERT đạt được điều này bằng cách huấn luyện các khối mã hóa của nó trên nhiệm vụ “masked language modeling”, 15% các token đầu vào của mỗi câu được thay thế bằng token [MASK]. Mô hình sau đó cố gắng dự đoán các giá trị ban đầu dựa trên ngữ cảnh được cung cấp bởi các token không bị che. Do hàm mất mát chỉ được tính trên các vị trí bị che, mô hình hội tụ chậm hơn so với các mô hình khác.

Ngoài ra, trong quá trình tiền huấn luyện, BERT nhận các cặp câu và học cách dự đoán xem câu thứ hai có phải là câu tiếp theo trong văn bản gốc hay không. Đầu vào được định dạng sao cho một token [SEP] được chèn vào cuối mỗi câu; thêm vào đó, mô hình sử dụng một vector bổ sung để chỉ ra câu nguồn của mỗi token tương ứng.

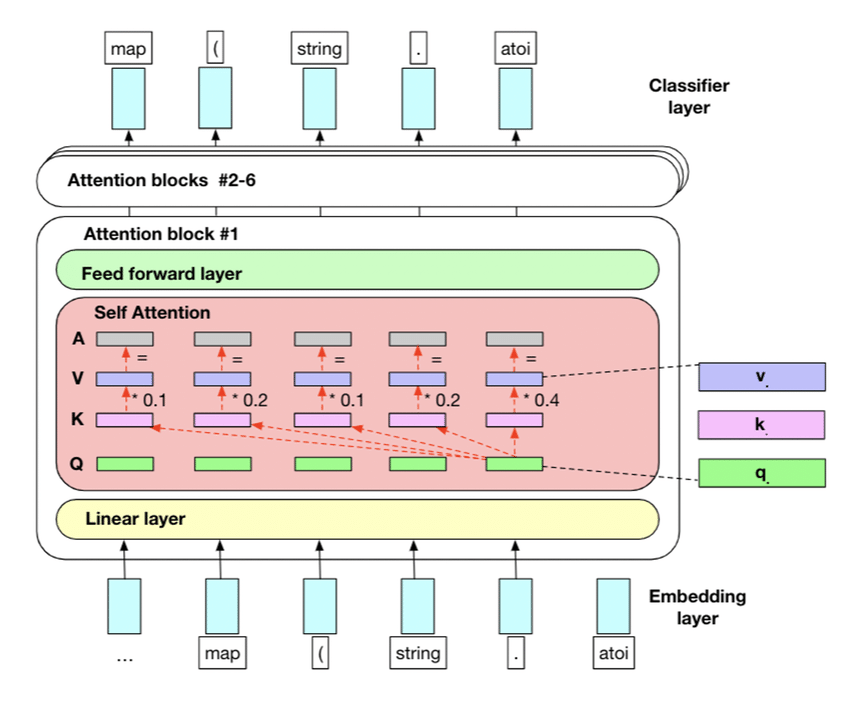
Bằng cách tận dụng quy trình tiền huấn luyện này, người dùng có thể tải xuống các tham số đã được tính toán trước, tinh chỉnh mô hình, tiết kiệm thời gian, kiến thức và tài nguyên, đồng thời đạt được hiệu suất cao hơn so với việc xây dựng một mô hình xử lý ngôn ngữ từ đầu. BERT đạt được kết quả tiên tiến nhất trên nhiều nhiệm vụ xử lý ngôn ngữ tự nhiên (NLP).



**Hình 4**: Kiến trúc BERT

#### GPT-2

GPT-2 là một mô hình ngôn ngữ dựa trên Transformer, vượt trội hơn các kiến trúc đã phát triển trước đây về khả năng tạo văn bản. Mô hình này được huấn luyện trên một tập dữ liệu rất lớn khoảng 40GB văn bản với mục tiêu đơn giản là dự đoán từ tiếp theo, dựa trên tất cả các từ trước đó trong văn bản. Không giống như BERT [23], kiến trúc của GPT-2 chỉ bao gồm các khối Transformer decoder: nó xuất ra một token tại một thời điểm, thêm token vừa dự đoán vào chuỗi các đầu vào, làm cho mô hình có tính chất “tự hồi quy” (auto-regressive) theo bản chất.Các khối tạo thành ngăn xếp của kiến trúc GPT-2 bao gồm một lớp masked self-attention và một mạng nơ-ron feed-forward. Lớp self-attention đạt được tính chất “tự hồi quy” bằng cách che đi các token liên quan đến các từ tương lai, ngăn chặn thông tin từ các token ở bên phải của vị trí đang được tính toán. Trong quá trình huấn luyện, câu đầu vào hoàn toàn lấy từ batch, trong khi ở thời điểm suy luận, token cuối cùng được dự đoán sẽ được thêm vào đầu vào hiện tại để dự đoán token tiếp theo; quá trình suy luận tiếp tục cho đến khi token kết thúc chuỗi được dự đoán, hoặc đạt đến độ dài tối đa của văn bản tạo ra. Các kiến trúc Transformer chỉ sử dụng decoder không chỉ được sử dụng cho mô hình ngôn ngữ, mà còn trong các nhiệm vụ NLP khác: nhờ vào cách tiền huấn luyện đơn giản nhưng hiệu quả và lượng mẫu huấn luyện khổng lồ, GPT-2 cho phép đạt được hiệu suất cao thông qua việc tinh chỉnh trên các nhiệm vụ như tóm tắt văn bản, trả lời câu hỏi và những nhiệm vụ khác.



**Hình 5**: Kiến trúc GPT-2

# THỰC NGHIỆM

## Dữ liệu

Mô hình được huấn luyện và đánh giá trên tập FAshion CAptioning Dataset (FACAD)[24], là tập dữ liệu lớn nhất hiện có cho việc chú thích hình ảnh thời trang. Cụ thể, FACAD chứa 993 nghìn hình ảnh có độ phân giải cao được mô tả bằng 130 nghìn chú thích liên quan đến thời trang chi tiết, với 6 ∼ 7 hình ảnh cho mỗi mặt hàng quần áo. Hơn nữa, tập dữ liệu chứa một danh sách gồm 990 thuộc tính và 78 danh mục để dán nhãn cho mỗi mặt hàng bằng thông tin chi tiết. Trong đồ án này, chúng tôi sử dụng 20.000 cặp hình ảnh văn bản cho cả việc huấn luyện và xác thực mô hình. Dưới đây là một số hình ảnh về tập dữ liệu này.

|  |  |
| --- | --- |
| Image | Caption |
| A black and white striped tank top  Description automatically generated | Jersey top with narrow shoulder straps. |
|  | Short-sleeved top in jersey with sewn-in turn-ups on the sleeves. |

**Bảng 1**: Mô tả dữ liệu

## Cách đánh giá

Hiệu quả của phương pháp đề xuất của chúng tôi cho nhiệm vụ chú thích hình ảnh thời trang được đánh giá thông qua các thí nghiệm được thực hiện trên tập dữ liệu FACAD bằng các chỉ số đo lường tiêu chuẩn, chẳng hạn như ROUGE [26], BLEU [27], METEOR [28].

**ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**: ROUGE là một trong những số liệu phổ biến nhất để đánh giá chất lượng tóm tắt văn bản. Nó đo lường sự trùng lặp n-gram (cụm từ gồm n từ) giữa bản tóm tắt được tạo ra và bản tóm tắt tham chiếu. ROUGE được tính toán cho các n-gram khác nhau (1-gram, 2-gram, 3-gram) và các điểm số được kết hợp để có được điểm số tổng thể.

**BLEU (Bilingual Evaluation Understudy):** BLEU là một thước đo được sử dụng để đánh giá chất lượng của việc dịch máy từ ngôn ngữ này sang ngôn ngữ khác. Tuy nhiên, nó cũng có thể được sử dụng để đánh giá chất lượng của việc tạo chú thích ảnh. Thước đo BLEU so sánh chú thích được tạo ra với chú thích tham chiếu và gán điểm dựa trên sự trùng lặp từ giữa hai chú thích này. Điểm BLEU nằm trong khoảng từ 0 đến 1. Càng gần 1, chất lượng của chú thích càng tốt. Để tính điểm BLEU, chúng ta sử dụng một công thức tính đến lượng từ trùng lặp giữa chú thích được tạo và chú thích tham chiếu. Càng nhiều từ chung, điểm BLEU càng cao.

**METEOR (Metric for Evaluation of Translation with Explicit Ordering):** Độ đo này đã được áp dụng để bù đắp cho những hạn chế của chỉ số BLEU vì nó không tính đến sự thu hồi và chỉ cho phép khớp chính xác n-gram. Nó xem xét các cặp từ đơn có thể giữa câu được tạo ra và câu tham chiếu, ngay cả khi chúng không hoàn toàn giống nhau. METEOR đặt trọng tâm như nhau vào độ chính xác và khả năng thu hồi: độ chính xác đề cập đến mức độ chính xác của các từ trong chú thích được tạo so với các chú thích tham chiếu của chúng, trong khi việc thu hồi đánh giá độ chính xác tương tự nhưng đối với các chú thích tham chiếu xuất hiện với các chú thích tương ứng của chúng.

## Kết quả

Từ những đề xuất ở trên chúng tôi tiến hành thực nghiệm, huấn luyện trên các mô hình khác nhau. Cuối cùng tiến hành so sánh và đánh giá kết quả giữa các mô hình này dựa trên các phương pháp đánh giá đã nêu ở phần trên. Từ đó lựa chọn mô hình có kết quả tốt nhất để dự đoán các mẫu trên tập kiểm tra.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ViT – Bert | ViT – GPT2 | BEiT – Bert | BEiT – GPT2 |
| BLEU 1 | 63.12 | 57.19 | 15.13 | **68.15** |
| BLEU 2 | 53.84 | 46.39 | 1.03 | **59.86** |
| ROUGE 1 | 69.28 | 63.98 | 19.80 | **74.03** |
| ROUGE 2 | 51.86 | 44.62 | 0.56 | **58.82** |
| ROUGE L | 64.81 | 58.74 | 12.58 | **69.89** |
| METEOR | 64.91 | 59.26 | 15.80 | **70.34** |

**Bảng 2**: Kết quả so sánh giữa các mô hình

Mặt khác, để chứng minh tính hiệu quả của mô hình đề xuất chúng tôi đã lựa chọn các mô hình đã được giới thiệu và phát triển trước đây đạt kết quả tốt đối với bài toán này để tiến hành so sánh và đánh giá với phương pháp đề xuất của chúng tôi. Bảng 3 so sánh phương pháp được đề xuất với các mô hình hiện đại khác được huấn luyện trên tập dữ liệu FACAD. Đối với các số liệu dùng để so sánh, chúng tôi sử dụng các kết quả có sẵn được đưa ra bởi Nicholas Moratelli at el. (2023) [30].

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **B-1** | **B-4** | **M** | **R** |
| **Show, Attend, and Tell [31]** | - | 4.3 | 9.5 | 19.1 |
| **Up–Down [32]** | - | 4.4 | 9.7 | 19.6 |
| **LBPF [33]** | - | 4.5 | 9.5 | 19.1 |
| **SRFC [34]** | - | 4.4 | 9.8 | 20.2 |
| **SCST [35]** | - | 5.6 | 11.8 | 22.0 |
| **Transformer [36]** | 24.5 | 6.8 | 10.1 | 19.7 |
| **CaMEL [37]** | 25.0 | 7.0 | 10.7 | 20.4 |
| **OpenCLIP ViT-L/14 [30]** | 27.3 | 10.6 | 11.5 | 22.3 |
| **Ours** | **68.15** | **49.05** | **70.34** | **69.89** |

**Bảng 3**: Kết quả so sánh giữa mô hình đề xuất và các mô hình tiên tiến

Có thể thấy, mô hình của chúng tôi đạt hiệu suất vượt trội so với tất cả các mô hình khác được so sánh trong Bảng. Các phương pháp khác như OpenCLIP ViT-L/14 và CaMEL cũng đạt được kết quả tốt, nhưng vẫn thua kém đáng kể so với phương pháp của chúng tôi. Với chỉ số cao nhất ở tất cả các tiêu chí đánh giá, mô hình của chúng tôi chứng minh khả năng tạo caption chính xác, chi tiết và phù hợp hơn hẳn so với các mô hình trước đây.

Từ những kết quả trên, chúng tôi tiến hành dự đoán một số mẫu trên tập kiểm tra sử dụng mô hình đề xuất. Các kết quả ở Bảng 4 cho thấy mô hình đưa ra các chú thích khá chính xác và phù hợp cho từng hình ảnh.

|  |  |  |  |
| --- | --- | --- | --- |
| A pink and white shirt  Description automatically generated | A pair of pants on a white background  Description automatically generated | A black shirt with a white stripe on it  Description automatically generated | A black and white shoe  Description automatically generated |
| **Actual Caption:**  Long-sleeved, polo-neck tops in organic cotton jersey.  **Predicted Caption:**  Long-sleeved, polo-neck tops in soft jersey. | **Actual Caption:** CONSCIOUS Joggers in soft organic cotton sweatshirt fabric with an elasticated drawstring waist and ribbed hems.  **Predicted Caption:**  Trousers in soft organic cotton jersey with elasticated ribbing at the waist, side pockets and ribbed hems. | **Actual Caption:**  Short-sleeved sports top in airy, organic cotton jersey with slits in the sides. Slightly longer at the back.  **Predicted Caption:** Round-necked T-shirt in soft jersey. | **Actual Caption:**  Mesh and scuba trainers with imitation leather details, a hook and loop tab with a text print, and a loop at the back. Mesh insoles and rubber soles.  **Predicted Caption:** Trainers in imitation leather with lacing at the front, mesh linings and insoles and rubber soles. |

**Bảng 4**: Kết quả thực nghiệm

Tổng hợp các kết quả ở trên cho thấy sự kết hợp giữa mô hình BEiT và GPT2 là một phương pháp tiềm năng cho bài toán chú thích hình ảnh thời trang với độ chính xác khá cao. Với việc cải thiện thêm dữ liệu và kỹ thuật, mô hình BeiT – GPT2 có thể đạt được hiệu suất cao hơn nữa trong tương lai.

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## Kết quả đạt được

Về mặt lý thuyết, Đồ án đã tìm hiểu về các phương pháp giải quyết bài toán chú thích hình ảnh thời trang, đồng thời Đồ án cũng đề xuất các phương pháp transformers để giải quyết bài toán này.

Kết quả thực nghiệm của nghiên cứu đã chứng minh tính hiệu quả và tiềm năng của phương pháp đề xuất trong bài toán này. Mô hình có thể đạt độ chính xác cao trong việc tạo chú thích cho các hình ảnh thời trang.

## Hạn chế

Mặc dù mô hình đã đạt được kết quả thực nghiệm khá tốt, tuy nhiên mô hình vẫn còn một số hạn chế như:

* Ngôn ngữ diễn đạt: Việc sử dụng ngôn ngữ tự nhiên để mô tả thời trang đòi hỏi sự tinh tế và sáng tạo. Mô hình cần phải có khả năng diễn đạt phong phú, sử dụng từ vựng đa dạng và phù hợp với từng phong cách thời trang, để tạo ra những mô tả hấp dẫn và thu hút người đọc.
* Độ phức tạp của hình ảnh: Hình ảnh thời trang thường có độ phức tạp cao, với nhiều chi tiết nhỏ, các lớp trang phục chồng lên nhau, hoặc các yếu tố gây nhiễu. Điều này gây khó khăn cho việc nhận diện và phân tích chính xác các yếu tố thời trang trong hình ảnh.

## Hướng phát triển

Trong tương lai, để cải thiện và phát triển khả năng cho mô hình tạo chú thích hình ảnh thời trang, một số hướng phát triển được đề xuất như:

* Tiếp tục cải thiện chất lượng dữ liệu: Đầu tư vào việc thu thập và gán nhãn dữ liệu huấn luyện chất lượng cao, đa dạng và phong phú hơn để giúp mô hình học hỏi và cải thiện khả năng mô tả thời trang
* Tích hợp thông tin: Kết hợp thông tin từ nhiều nguồn khác nhau như hình ảnh, văn bản, âm thanh, video để tạo ra mô tả thời trang phong phú và chính xác hơn.
* Tạo mô tả sáng tạo và đa dạng: Mô hình sẽ không chỉ mô tả một cách khách quan mà còn có khả năng tạo ra những mô tả sáng tạo, sử dụng ngôn ngữ biểu cảm, so sánh, ẩn dụ, để thể hiện được phong cách và cảm xúc của bộ trang phục.

# TÀI LIỆU THAM KHẢO

1. Ge, Y., Zhang, R., Wang, X., Tang, X., & Luo, P. (2019). Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5337-5345).
2. Li, J., Zhao, J., Wei, Y., Lang, C., Li, Y., Sim, T., ... & Feng, J. (2017). Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206*.
3. Huang, Q., Han, X., Lu, T., & Liu, G. (2021, May). Clothing Image Retrieval Based on Parts Detection and Segmentation. In *Proceedings of the 2021 3rd International Conference on Image Processing and Machine Vision* (pp. 53-59).
4. Santesteban, I., Otaduy, M. A., & Casas, D. (2019, May). Learning‐based animation of clothing for virtual try‐on. In *Computer Graphics Forum* (Vol. 38, No. 2, pp. 355-366).
5. Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., & Cucchiara, R. (2022). From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, *45*(1), 539-559.
6. Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, *51*(6), 1-36.
7. Wang, C., Yang, H., Bartz, C., & Meinel, C. (2016, October). Image captioning with deep bidirectional LSTMs. In *Proceedings of the 24th ACM international conference on Multimedia* (pp. 988-997).
8. Chohan, M., Khan, A., Mahar, M. S., Hassan, S., Ghafoor, A., & Khan, M. (2020). Image captioning using deep learning: A systematic. *image*, *11*(5).
9. Castro, R., Pineda, I., Lim, W., & Morocho-Cayamcela, M. E. (2022). Deep learning approaches based on transformer architectures for image captioning tasks. *IEEE Access*, *10*, 33679-33694.
10. Ji, Y., Zhang, H., Zhang, Z., & Liu, M. (2021). CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances. *Information Sciences*, *546*, 835-857.
11. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 7, 12 2015.
12. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
13. Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3128-3137).
14. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156-3164).
15. Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, *452*, 48-62.
16. Nguyen, B. T., Prakash, O., & Vo, A. H. (2021). Attention mechanism for fashion image captioning. In *Computational Intelligence Methods for Green Technology and Sustainable Development: Proceedings of the International Conference GTSD2020 5* (pp. 93-104). Springer International Publishing.
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.
18. Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10578-10587).
19. Li, L. H., Yatskar, M., Yin, D., Hsieh, C. J., & Chang, K. W. (2019). Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.
20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
21. Bao, H., Dong, L., Piao, S., & Wei, F. (2021). Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.
22. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.
23. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
24. Yang, X., Zhang, H., Jin, D., Liu, Y., Wu, C. H., Tan, J., ... & Wang, X. (2020). Fashion captioning: Towards generating accurate descriptions with semantic rewards. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16* (pp. 1-17). Springer International Publishing.
25. Tato, A., & Nkambou, R. (2018). Improving adam optimizer.
26. Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74-81).
27. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
28. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the Annual Meeting on Association for Computational Linguistics Workshops, Ann Arbor, MI, USA, 25–30 June 2005; pp. 65–72.
29. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021, July). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning* (pp. 10347-10357). PMLR.
30. Moratelli, N., Barraco, M., Morelli, D., Cornia, M., Baraldi, L., & Cucchiara, R. (2023). Fashion-oriented image captioning with external knowledge retrieval and fully attentive gates. *Sensors*, *23*(3), 1286.
31. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
32. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6077–6086.
33. Qin, Y.; Du, J.; Zhang, Y.; Lu, H. Look Back and Predict Forward in Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 8367–8375.
34. Yang, X.; Zhang, H.; Jin, D.; Liu, Y.; Wu, C.H.; Tan, J.; Xie, D.; Wang, J.; Wang, X. Fashion Captioning: Towards Generating Accurate Descriptions with Semantic Rewards. arXiv 2020, arXiv:2008.02693v2.
35. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-Critical Sequence Training for Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7008–7024.
36. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-Memory Transformer for Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 10578–10587.
37. Barraco, M.; Stefanini, M.; Cornia, M.; Cascianelli, S.; Baraldi, L.; Cucchiara, R. CaMEL: Mean Teacher Learning for Image Captioning. In Proceedings of the International Conference on Pattern Recognition, Montreal, QC, Canada, 21–25 August 2022; pp. 4087–4094.