**SINH CAPTION TỰ ĐỘNG CHO HÌNH ẢNH THỜI TRANG SỬ DỤNG CÁC PHƯƠNG PHÁP HỌC TRANSFORMER**

VÕ QUỐC HUY1\*, NGUYỄN ĐÌNH NGUYÊN BẮC1, BÙI THANH HÙNG1

*1 Khoa công nghệ Thông tin, Trường Đại học Công nghiệp Thành phố Hồ Chí Minh*

*\*huyvo8500@gmail.com, nguyendinhnguyenbac@gmail.com, buithanhhung@iuh.edu.vn*

**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

**Lĩnh vực:** Điện – Điện tử - Công nghệ thông tin

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

**Abstract:** As E-commerce continues to burgeon and the demand for automation in Information Technology escalates, particularly in the Fashion industry, where clothing companies curate online catalogs replete with images and captions to facilitate customer discovery of desired items, the significance of fashion image captions becomes pronounced. These captions serve to automatically generate product descriptions, capture customer attention, enhance understanding of merchandise, and ultimately augment revenue for clothing enterprises. However, the process of scanning, rendering, and captioning these fashion items predominantly remains a manual endeavor, entailing substantial time investment. In this study, we propose a model for the automated generation of fashion image captions leveraging Transformer architectures, specifically BEiT and GPT2. BEiT is employed to extract features from fashion images, while GPT2 is tasked with generating precise and engaging descriptions. Our evaluation conducted on the Fashion Captioning Dataset (FACAD), employing metrics such as ROUGE, BLEU, METEOr, demonstrates that our approach surpasses prior methodologies significantly, underscoring its efficacy in fashion image captioning task.

**Keywords: Automatic caption generation**, Transformer, BEiT, GPT2, Fashion image captioning

**1 GIỚI THIỆU**

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], ...

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.

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 nghiên cứu 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 bài báo 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.

Ngoài phần giới thiệu, các phần còn lại của bài báo sẽ bao gồm những nội dung sau: Các nghiên cứu liên quan sẽ được trình bày ở Phần 2. Phần 3 là mô hình đề xuất của chúng tôi. Thực nghiệm và đánh giá sẽ nằm trong Phần 4. Phần 5 sẽ đưa ra những kết luận về nghiên cứu của chúng tôi và hướng phát triển của nghiên cứu trong tương lai.

**2 CÁC NGHIÊN CỨU LIÊN QUAN**

**2.1 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 T 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.

**2.2 Transformers Based Methods**

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.

Từ những nghiên cứu đã có trước đây, trong bài nghiên cứu 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.

**3 PHƯƠNG PHÁP**

**3.1 Mô hình đề xuất**

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.

Transformer

Input Image

Output Caption

Encoder

BEiT

Decoder

GPT2

**Hình 1**: Mô hình tổng quát

**3.2 Các thành phần mô hình**

**3.2.1 BEiT**

Chúng tôi chọn BEiT (Bidirectional Encoder representation from Image Transformers) làm khối mã hóa trong mô hình chú thích hình ảnh thời trang vì những lợi ích nổi bật mà nó mang lại trong việc xử lý thông tin hình ảnh. BEiT 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.

Một lý do quan trọng khác là 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.

**3.2.2 GPT-2**

GPT-2 được chọn làm khối giải mã trong mô hình chú thích hình ảnh thời trang nhờ vào khả năng vượt trội trong việc tạo văn bản tự nhiên và mạch lạc. 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. Hơn nữa, tính tự hồi quy của GPT-2 cho phép nó tạo ra các chuỗi văn bản liên tục, điều này rất quan trọng trong việc mô tả các chi tiết phức tạp và phong phú của các sản phẩm thời trang. Khả năng này giúp đảm bảo rằng mỗi mô tả không chỉ dừng lại ở việc liệt kê các đặc điểm của sản phẩm mà còn có thể kể một câu chuyện hoặc cung cấp thêm thông tin hữu ích cho người dùng.

Sự kết hợp giữa BEiT và GPT-2 trong mô hình đề xuất không chỉ tối ưu hóa quá trình hiểu và xử lý thông tin hình ảnh mà còn đảm bảo rằng các mô tả văn bản được tạo ra chính xác, tự nhiên. Điều này làm cho mô hình đề xuất trở thành một giải pháp mới và hiệu quả trong nhiệm vụ chú thích hình ảnh thời trang.

**4 THỰC NGHIỆM**

**4.1 Dữ liệu**

**4.1.1 FAshion CAptioning Dataset**

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 993k 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 nghiên cứu này, chúng tôi sử dụng 103k cặp hình ảnh văn bản cho cả việc huấn luyện và xác thực mô hình. Với 83k hình ảnh để huấn luyện, 10k hình ảnh để xác thực và 10k hình ảnh để thử nghiệm. 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. |
| A yellow t-shirt on a white background  Description automatically generated | Short-sleeved top in jersey with sewn-in turn-ups on the sleeves. |

**Bảng 1**: Mô tả bộ dữ liệu FACAD

**4.1.2 Fashion-Gen**

Fashion-Gen [25] bao gồm 293.008 hình ảnh, được chia thành tập huấn luyện gồm 260.480 hình ảnh, tập xác thực gồm 32.528 hình ảnh và tập kiểm tra gồm 32.528 hình ảnh. Các hình ảnh mô tả các mặt hàng thời trang được chụp từ nhiều góc độ khác nhau, tùy thuộc vào loại mặt hàng. Bao gồm các chú thích mô tả chi tiết do các chuyên gia viết cho mỗi mặt hàng thời trang. Mặc dù ban đầu được tạo ra để tạo hình ảnh thời trang, tập dữ liệu này cũng có thể được sử dụng để chú thích hình ảnh thời trang. Do đó, chúng tôi đã thực hiện các bước tiền xử lý bao gồm làm sạch dữ liệu, chuẩn hóa dữ liệu và loại bỏ các từ không thường xuyên không liên quan đến các thuộc tính thời trang. Cuối cùng, chúng tôi thu được mô tả thuộc tính của các mặt hàng quần áo này cho việc huấn luyện nhiệm vụ chú thích hình ảnh thời trang trong nghiên cứu của chúng tôi. Bảng 2 mô tả một số hình ảnh về tập dữ liệu này.

|  |  |
| --- | --- |
| Image | Caption |
|  | long sleeve over long wool shirt in red and black. plaid pattern throughout. button closure and ap pocket at front. logo print at front hem and center back in white. shirt tail hem. tonal stitching. single button barrel cuff |
|  | long sleeve cotton and linen blend denim shirt in blue. fading and distressing throughout. spread collar. press stud closure at front. ap pocket at chest. dropped shoulder. single button barrel cuff. tortoiseshell hardware. Contrast stitching in tan. |

**Bảng 2**: Mô tả bộ dữ liệu Fashion-Gen

**4.2 Hyperparameters**

Mô hình đề xuất của chúng tôi được triển khai bằng Pytorch và được đào tạo bằng Kaggle GPU T4x2. Mô hình của chúng tôi được huấn luyện trên 10 epochs, hàm tối ưu hóa AdamW [26] được sử dụng với learning rate 5e-5 và sử dụng batch size bằng 16 trong cả giai đoạn huấn luyện và đánh giá. Các mô hình Trong quá trình huấn luyện, chúng tôi sử dụng Seq2SeqTrainer để tự động quản lý các quá trình training và evaluation, tận dụng khả năng parallelization và tích hợp với các công cụ theo dõi hiệu suất như TensorBoard.

**4.3 Phương pháp 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 [27], BLEU [28], METEOR [29].

* **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.

**4.4 Kết quả**

Dựa trên phương pháp đã được đề xuất ở trên chúng tôi tiến hành thực nghiệm và đánh giá mô hình của mình trên tập dữ liệu FACAD. Phương pháp được đề xuất của chúng tôi cho ra kết quả ấn tượng trên các chỉ số đánh giá. **Hình** 2 là kết quả trực quan kết quả mô hình chúng tôi trên các phương pháp đánh giá. Nhìn vào kết quả ta có thể thấy các chỉ số BLEU, ROUGE và METEOR đều cho thấy sự cải thiện liên tục qua các epoch, chỉ ra rằng mô hình ngày càng hiệu quả hơn trong việc tạo ra các đầu ra tương tự với các văn bản tham khảo khi quá trình huấn luyện kéo dài.

A graph of different colored lines

Description automatically generated

**Hình 2**: Trực quan hóa kết quả

Tiếp theo, chúng tôi tiến hành lựa chọn một số mô hình Transformers khác để thực nghiệm, so sánh và đánh giá với mô hình đề xuất của chúng tôi là: ViT – BERT, ViT – GPT2, DEiT [30] – BERT, DEiT -GPT2. **Bảng** 2 cho thấy kết quả đánh giá của mô hình đề xuất và các mô hình khác trên các chỉ số đánh giá như ROUGE, BLEU, METEOR. Cụ thể mô hình của chúng tôi có hiệu suất cao nhất trong tất cả các chỉ số, với Rouge-1 là 74.03, Rouge-2 là 58.82, Rouge-L là 69.89, Bleu-1 là 68.15, Bleu-4 là 49.05, và Meteor là 70.34. Điều này cho thấy mô hình của chúng tôi tạo ra các câu mô tả sát với câu mô tả tham chiếu hơn so với các mô hình khác.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **R-1** | **R-2** | **R-L** | **B-1** | **B-4** | **M** |
| ViT – BERT [31] | 69.28 | 51.86 | 64.81 | 63.12 | 42.06 | 64.91 |
| ViT – GPT2 [32] | 63.98 | 44.62 | 58.74 | 57.19 | 33.39 | 59.26 |
| DEiT – BERT [33] | 70.13 | 53.25 | 65.78 | 64.02 | 43.64 | 66.08 |
| DEiT -GPT2 [34] | 69.67 | 52.38 | 65.07 | 63.21 | 41.45 | 65.34 |
| **Mô hình đề xuất** | **72.31** | **54.98** | **67.51** | 65.35 | **43.47** | **68.15** |

**Bảng 2:** Kết quả so sánh giữa các mô hình Transformers khác trên bộ dữ liệu FACAD

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 mô hình được đưa ra bởi Nicholas Moratelli at el. (2023) [35] và sau đó tiến hành thực nghiệm lại trên bộ dữ liệu của chúng tôi.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **B-1** | **B-4** | **M** | **R** |
| LBPF [36] | - | 25.43 | 29.67 | 38.3 |
| SRFC [37] | - | 26.42 | 29.45 | 35.27 |
| Transformer [38] | 37.56 | 39.24 | 42.13 | 39.74 |
| CaMEL [39] | 54.72 | 38.53 | 37.48 | 42.16 |
| OpenCLIP ViT-L/14 [35] | 57.31 | 41.43 | 46.82 | 56.55 |
| **Mô hình đề xuất** | **65.35** | **43.47** | **68.15** | **67.51** |

**Bảng 3**: Kết quả so sánh giữa mô hình đề xuất và các mô hình cơ sở trên bộ dữ liệu FACAD

Để đảm bảo về hiệu quả của mô hình, chúng tôi cũng thực hiện đánh và xác thực trên tập dữ liệu Fashion-Gen. Kết quả của chúng tôi được so sánh với các phương pháp SOTA khác như: encoder-decoder model [14], channel-wise attention [40], spatial attention [41], channel and spatial attention [42]. Các kết quả so sánh được hiển thị ở **Bảng 4**. Cụ thể, mô hình của chúng tôi thu được các kết quả 55.7, 46.3, 38.8, 32.4 cho BLEU-1, BLEU-2, BLEU-3, BLEU-4 tương ứng và ROUGE-L là 54.2. Điều này cho thấy phương pháp của chúng tôi vượt trội hơn đáng kể so với các phương pháp SOTA khác trong việc tạo chú thích chính xác và phù hợp cho hình ảnh thời trang.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | B-1 | B-2 | B-3 | B-4 | R-L |
| Encoder-Decoder [14] | 29.2 | 23.5 | 16.4 | 12.4 | 45.5 |
| Channel-wise attention [40] | 28 | 20.5 | 14.4 | 10.9 | 34.5 |
| Spatial attention [41] | 40.8 | 33.3 | 26.7 | 22.1 | 50.2 |
| Channel and spatial attention [42] | 45.6 | 36.4 | 28.8 | 22.8 | 50.6 |
| **Mô hình đề xuất** | **55.7** | **46.3** | **38.8** | **32.4** | **54.2** |

**Bảng 4**: Kết quả so sánh giữa mô hình đề xuất và các mô hình cơ sở trên bộ dữ liệu Fashion-Gen

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 3** và **Bảng 4**. Sự vượt trội này đến từ sự kết hợp mạnh mẽ giữa BEiT và GPT2 trong nhiệm vụ chú thích hình ảnh. BEiT Encoder sử dụng kỹ thuật Masked Image Modeling (MIM) để huấn luyện mô hình, cho phép chuyển đổi hình ảnh thành các biểu diễn ngữ nghĩa chi tiết và chính xác, cung cấp thông tin phong phú về hình ảnh cho quá trình sinh văn bản. Mặt khác, GPT-2 với thiết kế unidirectional chỉ dựa vào ngữ cảnh từ trái sang phải và mục tiêu huấn luyện sinh từ tiếp theo, phù hợp hơn cho việc sinh văn bản tự nhiên và mạch lạc, giúp tận dụng hiệu quả các biểu diễn hình ảnh từ BEiT để tạo ra các mô tả hình ảnh chính xác và phong phú. Mô hình mới được đề xuất có khả năng nắm bắt cả thông tin vùng và các đặc trưng hình ảnh của các mặt hàng thời trang. Các kết quả cho thấy sự cải thiện đáng kể về hiệu suất của nhiệm vụ chú thích hình ảnh thời trang khi sử dụng mô hình đề xuất của chúng tôi.

Từ những kết quả trên, chúng tôi tiến hành dự đoán một số mẫu trên tập dữ liệu FACAD sử dụng mô hình đề xuất. Các kết quả ở **Bảng** **5** cho thấy mô hình đưa ra các chú thích khá chính xác và phù hợp cho từng hình ảnh. Tuy nhiên, khi phân tích chi tiết các kết quả này, chúng tôi nhận thấy có sự khác biệt rõ rệt giữa các trường hợp chú thích tốt và chưa tốt. Đối với các trường hợp tốt, mô hình BEiT-GPT2 có khả năng tạo ra các chú thích không chỉ chính xác mà còn chi tiết, phản ánh đúng các đặc điểm nổi bật của hình ảnh. Ví dụ, chú thích dự đoán cho hình ảnh về "Long-sleeved polo-neck tops in soft jersey" gần như tương đồng với chú thích thực tế "Long-sleeved polo-neck tops in organic cotton jersey", chỉ có sự khác biệt nhỏ về chất liệu. Điều này cho thấy mô hình có khả năng nhận diện tốt các yếu tố chính của trang phục. Tuy nhiên, trong các trường hợp chưa tốt, mô hình thường bỏ sót hoặc không nhận diện chính xác một số chi tiết quan trọng của hình ảnh. Ví dụ, đối với hình ảnh về "Short-sleeved sports top in airy organic cotton jersey with slits in the sides", chú thích dự đoán "Round-necked T-shirt in soft jersey" thiếu nhiều thông tin quan trọng như đặc điểm "airy", "organic cotton", và "slits in the sides". Điều này cho thấy mô hình gặp khó khăn trong việc nắm bắt các chi tiết tinh tế hơn của hình ảnh.

Nguyên nhân của sự khác biệt này có thể do nhiều yếu tố, bao gồm độ phức tạp của hình ảnh, sự đa dạng của tập dữ liệu huấn luyện, và khả năng tổng quát hóa của mô hình. Để cải thiện hiệu suất của mô hình, cần có các biện pháp như mở rộng tập dữ liệu huấn luyện với các hình ảnh đa dạng hơn và tinh chỉnh các tham số của mô hình để tăng cường khả năng nhận diện chi tiết.

|  |  |  |  |
| --- | --- | --- | --- |
| A pink and white shirt  Description automatically generated | A pair of pants on a white background  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 5**: Kết quả thực nghiệm trên bộ dữ liệu FACAD

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

**5 KẾT LUẬN**

Trong nghiên cứu này, chúng tôi đã đề xuất một mô hình chú thích hình ảnh thời trang dựa trên kiến trúc Transformers encoder-decoder (BEiT - GPT2). Mô hình sử dụng BEiT để trích xuất đặc trưng hình ảnh và GPT-2 để tạo các mô tả văn bản, giúp cải thiện đáng kể chất lượng và độ chính xác của chú thích hình ảnh. Kết quả thực nghiệm cho thấy mô hình của chúng tôi không chỉ vượt trội so với các phương pháp truyền thống mà còn tạo ra các chú thích phong phú, ngữ nghĩa và phù hợp với ngữ cảnh. Sự kết hợp giữa BEiT và GPT-2 đã mở ra những hướng đi mới trong việc ứng dụng học sâu vào các tác vụ chú thích hình ảnh, đặc biệt là trong lĩnh vực thời trang.

Trong tương lai, để khắc phục các hạn chế và tiếp tục nâng cao hiệu suất của mô hình, chúng tôi đề xuất một số hướng nghiên cứu như tiếp tục cải thiện chất lượng dữ liệu đầu vào, 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. 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 hơn, 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. Rostamzadeh, N., Hosseini, S., Boquet, T., Stokowiec, W., Zhang, Y., Jauvin, C., & Pal, C. (2018). Fashion-gen: The generative fashion dataset and challenge. arXiv preprint arXiv:1806.08317.
26. Tato, A., & Nkambou, R. (2018). Improving adam optimizer.
27. Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74-81).
28. 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.
29. 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.
30. 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.
31. Li, Q., Gong, B., Cui, Y., Kondratyuk, D., Du, X., Yang, M. H., & Brown, M. (2021). Towards a unified foundation model: Jointly pre-training transformers on unpaired images and text. *arXiv preprint arXiv:2112.07074*.
32. Vasireddy, I., HimaBindu, G., & Ratnamala, B. (2023). Transformative Fusion: Vision Transformers and GPT-2 Unleashing New Frontiers in Image Captioning within Image Processing. *International Journal of Innovative Research in Engineering & Management*, *10*(6), 55-59.
33. Jaknamon, T., & Marukatat, S. (2022, November). ThaiTC: thai transformer-based image captioning. In *2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)* (pp. 1-4). IEEE.
34. Basak, D., Srijith, P. K., & Desarkar, M. S. (2024, April). Transformer based Multitask Learning for Image Captioning and Object Detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 260-272). Singapore: Springer Nature Singapore.
35. 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.
36. 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.
37. 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.
38. 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.
39. 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.
40. S. Li and K. Yamaguchi, “Attention to describe products with attributes,” in 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), May 2017, pp. 215–218, doi: 10.23919/MVA.2017.7986839.
41. K. Xu et al., “Show, attend and tell: Neural image caption generation with visual attention,” Prepr. arXiv.1502.03044, Feb. 2015
42. Nguyen, B. T., Nguyen, S. T., & Vo, A. H. (2023). Channel and spatial attention mechanism for fashion image captioning. *International Journal of Electrical & Computer Engineering (2088-8708)*, *13*(5).