Có sẵn trực tuyến tại www.sciencedirect.com

Có sẵn trực tuyến tại www.sciencedirect.com
Có sẵn trực tuyến tại www.sciencedirect.com

Khoa học trực tiếp

Khoa học máy tính Procedia 00 (2022) 000-000

Khoa học máy tính Procedia 00 (2023) 000–000
Khoa học máy tính Procedia 218 (2023) 686–696

Computer Science

www.elsevier.com/locate/procedia

www.elsevier.com/locate/procedia

Hội nghị quốc tế về học máy và kỹ thuật dữ liệu Hội nghị quốc tế về học máy và kỹ thuật dữ liệu

Kiến trúc lai sử dụng CNN và LSTM để chú thích hình ảnh trong Kiến trúc lai sử dụng CNN và LSTM để chú thích hình ảnh bằng tiếng Hindi Tiếng Hindi Ayush Kumar Poddara, Tiến sĩ Rajneesh Ranib Ayush Kumar Poddara, Tiến sĩ

Rajneesh Ranib

aTiến sĩ BR Viện Công nghệ Quốc gia Ambedkar, Jalandhar , Punjab, Ấn Độ bTiến sĩ BR Viện Công nghệ Quốc gia Ambedkar, Jalandhar, Punjab, Ấn Độ aTiến sĩ BR Viện Công nghệ Quốc gia Ambedkar, Jalandhar, Punjab, Ấn Độ bTiến sĩ BR Viện Công nghệ Quốc gia Ambedkar, Jalandhar, Punjab, Ấn Độ

Tóm tắt

Với sự ra đời của học sâu trong những năm gần đây, việc tích hợp thị giác máy tính với xử lý ngôn ngữ tự nhiên Tóm tắt đã thu hút được rất nhiều sự chú ý. Tạo ảnh là một trong những lĩnh vực hấp dẫn và được chú ý nhất Với sự ra đời của học sâu trong những năm gần đây, việc tích hợp thị giác máy tính với xử lý ngôn ngữ tự nhiên của Học máy, lĩnh vực này gặp phải một số trở ngại, đặc biệt là khi mô tả hình ảnh bằng các ngôn ngữ khác, đã thu hút được rất nhiều sự chú ý. Tạo mô tả từ hình ảnh là một trong những lĩnh vực hấp dẫn và được chú ý nhất so với tiếng Anh. Trong chú thích hình ảnh, máy tính được đào tạo để hiểu thông tin trực quan của hình ảnh và sau đó Học máy, lĩnh vực này gặp phải một số trở ngại, đặc biệt là khi mô tả hình ảnh bằng các ngôn ngữ khác, tạo mô tả dựa trên các đặc điểm hình ảnh và câu tham chiếu. Có một ứng dụng tự động hơn tiếng Anh. Trong chú thích hình ảnh, máy tính được đào tạo để hiểu thông tin trực quan của hình ảnh và sau đó mô tả các sự kiện trong môi trường của chúng và sau đó dịch chúng thành chú thích hoặc tin nhắn có thể tạo ra một mô tả dựa trên các đặc điểm hình ảnh và câu tham chiếu đáng kể. Có một ứng dụng tự động đóng góp cho xã hội. Bài báo này trình bày một mô hình mạng nơ-ron CNN-LSTM nhiều lớp được sử dụng để mô tả các sự kiện trong môi trường của chúng và sau đó dịch chúng thành chú thích hoặc thông điệp có thể nhận dạng và tạo chú thích tiếng Hindi đáng kể cho các đối tượng trong hình ảnh. Ngoài ra, nhiều mô hình khác nhau đã được đào tạo bằng cách đóng góp cho xã hội. Bài báo này trình bày một mô hình mạng nơ-ron CNN-LSTM nhiều lớp được sử dụng để điều chỉnh các siêu tham số và số lớp ẩn để tìm mô hình tối ưu và tối đa hóa khả năng nhận dạng và tạo chú thích tiếng Hindi cho các đối tượng trong hình ảnh. Ngoài ra, nhiều mô hình khác nhau đã được đào tạo bằng mô tả tiếng Hindi kết quả. Hơn nữa, sau khi kiểm tra hiệu quả của các mô hình của chúng tôi, người ta quan sát thấy tăng việc điều chỉnh các siêu tham số và số lớp ẩn để tìm mô hình tối ưu và tối đa hóa mô hình khả năng đã cho thấy sự gia tăng lần lượt là 34,64% và 29,13% về điểm BLEU (Unigram) và điểm BLEU (bigram) của mô tả tiếng Hindi kết quả. Hơn nữa, sau khi kiểm tra hiệu quả của các mô hình của chúng tôi, người ta quan sát thấy tăng khi so sánh với công trình hiện có đã thực hiện trong lĩnh vực này. Chú thích hình ảnh bằng tiếng Hindi có thể có nhiều mô hình đã cho thấy sự gia tăng 34,64% và 29,13% về điểm BLEU (Unigram) và điểm BLEU (bigram) của các ứng dụng trong xã hội ngày nay và nó cũng có thể cung cấp giao diện thân thiện với người dùng cho người nói tiếng Hindi. tương ứng khi so sánh với công việc hiện có được thực hiện trong lĩnh vực này. Chú thích hình ảnh bằng tiếng Hindi có thể có nhiều ứng dụng trong xã hội ngày nay và nó cũng có thể cung cấp giao diện thân thiện với người dùng cho người nói tiếng Hindi. © 2023 Các tác giả. Xuất bản bởi Elsevier BV

Mạng nơ-ron.

Từ khóa: Chú thích hình ảnh; Khung mã hóa giải mã; Học sâu; Mạng nơ-ron tích chập; Mạng nơ-ron hồi quy.

1. Giới thiệu

1. Giới thiệu Khái niệm tự động tạo ra các từ mô tả cho hình ảnh đã thu hút được rất nhiều sự quan tâm trong vài năm trở lại đây. Tạo chú thích cho hình ảnh là một nỗ lực quan trọng đòi hỏi phải có ngữ nghĩa Khái niệm tự động tạo ra các từ mô tả cho hình ảnh đã thu hút được rất nhiều sự quan tâm trong vài năm trở lại đây. Tạo chú thích cho hình ảnh là một nỗ lực quan trọng đòi hỏi phải có ngữ nghĩa

Tác giả liên hệ. Điện thoại: +0-000-000-0000 ; fax: +0-000-000-0000.
Địa chỉ email: author@institute.xxx
Tác giả liên hệ. Điện thoại: +0-000-000-0000 ; fax: +0-000-000-0000.
Địa chỉ email: author@institute.xxx

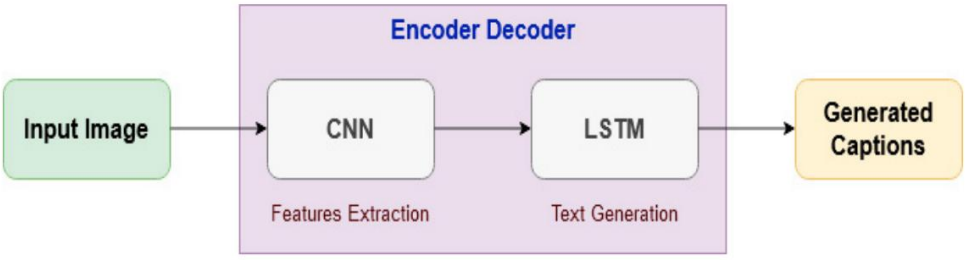
1877-0509 © 2023 Các tác giả. Xuất bản bởi Elsevier BV

hiểu biết về hình ảnh cũ ng như khả năng xây dựng chú thích có liên quan. Nó có thể có lợi cho nhiều lĩnh vực từ hỗ trợ người khiếm thị đến hiểu nội dung web. Nó cũ ng cho phép các công ty phân tích lượng dữ liệu hình ảnh khổng lồ. Những tiến bộ gần đây trong nhận dạng đối tượng và mô hình ngôn ngữ đã cho phép tạo ra chú thích hình ảnh có ý nghĩa.

Những nỗ lực ban đầu dựa vào các phương pháp mẫu, trong đó thông tin có nguồn gốc từ hình ảnh được sử dụng để điền vào các mẫu văn bản đã chỉ định [2]. Các kỹ thuật học sâu được sử dụng trong các hệ thống hiện tại, tận dụng sức mạnh tính toán có sẵn. Một trong những phương pháp tốt nhất để tạo chú thích hình ảnh là sử dụng kiến trúc mã hóa-giải mã. Trong phương pháp này, trước tiên các đặc điểm hình ảnh được trích xuất bằng các mô hình CNN được đào tạo trước như Alexnet, VGG16, DenseNet, v.v., sau đó các chú thích cho hình ảnh nhất định được tạo bằng mạng nơ-ron hồi quy [13]. Nhiều nhà nghiên cứu cũ ng đã đề xuất các giải pháp để cải thiện kiến trúc mã hóa-giải mã hiện có. Một trong số đó là mô hình lấy cảm hứng từ U-net [10] sử dụng kiến trúc mã hóa-giải mã đối xứng. Mô hình này sử dụng CNN nhiều lớp và đã chứng minh được sự cải thiện đáng kể trong quá trình phát hiện và phân đoạn hình ảnh.

Bài báo này giới thiệu một mô hình CNN-LSTM để tự động nhận dạng các đối tượng trong hình ảnh và tạo chú thích có liên quan cho chúng. Bài báo sử dụng các mô hình dựa trên Transfer Learning để nhận dạng các đối tượng bằng các phương pháp học sâu. Mô hình này có khả năng thực hiện hai nhiệm vụ. Nhiệm vụ đầu tiên là nhận dạng các đối tượng chính trong hình ảnh bằng các mô hình CNN được đào tạo trước và nhiệm vụ thứ hai là tạo mô tả cho hình ảnh bằng RNN. Mục tiêu của nghiên cứu này là cung cấp một mô hình mạng nơ-ron tích chập nhiều lớp và đánh giá kết quả của các mô hình này. Luồng của bài báo như sau:

Phần 2 của bài báo sẽ tập trung vào các công trình hiện có đã thực hiện trong lĩnh vực này. Phần 3 phác thảo phương pháp tiếp cận của chúng tôi, bao gồm mô hình đề xuất cũ ng như giai đoạn xử lý dữ liệu trước cho dữ liệu hình ảnh và văn bản. Phần 4 thảo luận về các mô hình đề xuất cho thử nghiệm. Trong phần 5, thiết kế và kết quả thử nghiệm được trình bày chi tiết cùng với so sánh với các cuộc điều tra trước đó. Cuối cùng, phần 6 kết luận bài báo cùng với một số công trình trong tương lai có thể được triển khai trên công trình hiện có.



Hình 1. Sơ đồ mô hình tạo chú thích hình ảnh

2. Các tác phẩm liên quan

Phần này xem xét tình trạng chú thích hình ảnh hiện đại cũ ng như các phương pháp khác nhau được các học giả sử dụng. Nhiều học giả gần đây đã làm việc trong lĩnh vực này và đưa ra nhiều phương pháp khác nhau để tạo chú thích chất lượng cao cho hình ảnh.

Trong giai đoạn đầu của chú thích hình ảnh, các phương pháp tiếp cận dựa trên mẫu đã được xây dựng. Trong phương pháp tiếp cận này, có một số lượng mẫu cố định và chúng được điền vào dựa trên phát hiện đối tượng, nhận dạng cảnh, phân loại thuộc tính hoặc các đặc điểm khác.

Farhadi et al. [2] đã đề xuất một phương pháp tiếp cận dựa trên mẫu, trong đó một bộ ba gồm đối tượng, hành động và cảnh được sử dụng để nắm bắt các yếu tố của cảnh. Các thuật toán phát hiện đối tượng được sử dụng ban đầu để ước tính các cảnh và đối tượng, sau đó là việc sử dụng các mô hình ngôn ngữ được đào tạo trước để xác định giới từ, động từ và nhiều tình huống. Cuối cùng, một chú thích mô tả được tạo ra.

Kulkarni et al. [6] đã đề xuất một kỹ thuật dựa trên mẫu sử dụng trường ngẫu nhiên có điều kiện (CRF) để dự đoán mô tả hình ảnh tốt nhất. Trong phương pháp tiếp cận của họ, trước tiên, một số lượng lớn các máy dò đối tượng được sử dụng để quét hình ảnh và thu thập tập hợp các phát hiện có điểm cao. Sau đó, CRF được sử dụng để tạo mô tả cho hình ảnh.

Hạn chế của các phương pháp dựa trên mẫu này là chúng chỉ có thể tạo ra các chú thích có độ dài cụ thể. Điều này dẫn đến sự ra đời của khuôn khổ mã hóa-giải mã [15]. Với sự thành công của khuôn khổ mã hóa-giải mã, nhiều kỹ thuật dựa trên mạng nơ-ron đã xuất hiện. Phương pháp này đã thu hút được rất nhiều sự ưa chuộng trong số các nhà nghiên cứu. Một số công trình liên quan đến các phương pháp dựa trên mạng nơ-ron bao gồm:

Kiros et al. [5] đã trình bày một mô hình mã hóa-giải mã sử dụng mạng nơ-ron truyền thẳng. Phương pháp của họ dự đoán từ mục tiêu dựa trên các đặc điểm được trích xuất của hình ảnh và từ trước đó bằng cách sử dụng mô hình log-bilinear đa phương thức. Trong phương pháp của họ, bộ mã hóa cho phép bạn xếp hạng ảnh và từ bằng cách sử dụng các mô hình CNN (Mạng nơ-ron tích chập) được đào tạo trước, trong khi bộ giải mã có thể tạo chú thích mới từ đầu bằng cách sử dụng LSTM.

Xiao et al. [16] đã đề xuất một mô hình mã hóa-giải mã LSTM ba lớp có hiệu quả hợp nhất hình ảnh và dữ liệu văn bản để tạo ra các mô tả có liên quan. Họ đã nghiên cứu cách đầu ra từ lớp giữa của LSTM có thể được sử dụng để cải thiện mô hình tạo ngôn ngữ của LSTM trên cùng. Hơn nữa, họ cũ ng sử dụng chiến lược tối ưu hóa độ dốc chính sách để tăng hiệu suất của mô hình.

Trong chú thích hình ảnh, các cơ chế chú ý gần đây đã được đưa vào các khuôn khổ thần kinh mã hóa-giải mã hiện có. Sau đây là một số tác phẩm sử dụng kỹ thuật chú thích hình ảnh dựa trên sự chú ý:

Xu et al. [17] đã đề xuất một mô hình để tạo ra các mô tả hình ảnh kết hợp sự chú ý trực quan với trạng thái ẩn của một lớp LSTM duy nhất. Họ đã sử dụng các kỹ thuật truyền ngược điển hình để đào tạo mô hình của họ theo cách xác định. Họ cũ ng chứng minh cách mô hình có thể tự động học cách tập trung sự chú ý vào các mục quan trọng trong khi tạo ra các câu thích hợp trong chuỗi đầu ra thông qua trực quan hóa.

Al-Malla và cộng sự [1] đã đề xuất một mô hình chú thích dựa trên sự chú ý. Mô hình chú thích hình ảnh của họ sử dụng hai kỹ thuật trích xuất đặc điểm: Xception, một mô hình CNN được đào tạo trước và YOLOv4, là một mô-đun chú ý được sử dụng để phát hiện các đối tượng trong hình ảnh. Họ cũ ng minh họa cách "yếu tố quan trọng" cải thiện độ chính xác của mô hình bằng cách ưu tiên các đối tượng lớn ở phía trước hơn các đối tượng nhỏ ở phía sau.

Mishra et al. [7] là những người đầu tiên thiết kế kiến trúc dựa trên bộ chuyển đổi để chú thích hình ảnh bằng tiếng Hindi. Họ đã sử dụng tập dữ liệu MSCOCO, sau đó được dịch sang tiếng Hindi bằng trình dịch.
Mô hình được đề xuất sử dụng CNN làm bộ mã hóa để trích xuất đặc điểm và mô hình biến đổi làm bộ giải mã. Mô hình này không có bất kỳ mô hình RNN nào, thay vào đó, nó sử dụng mô hình biến đổi. Hơn nữa, họ cũ ng đã so sánh mô hình được đề xuất của mình với nhiều đường cơ sở. Trong mỗi đường cơ sở, các CNN và RNN khác nhau có sự chú ý về không gian được sử dụng làm bộ mã hóa và giải mã tương ứng.

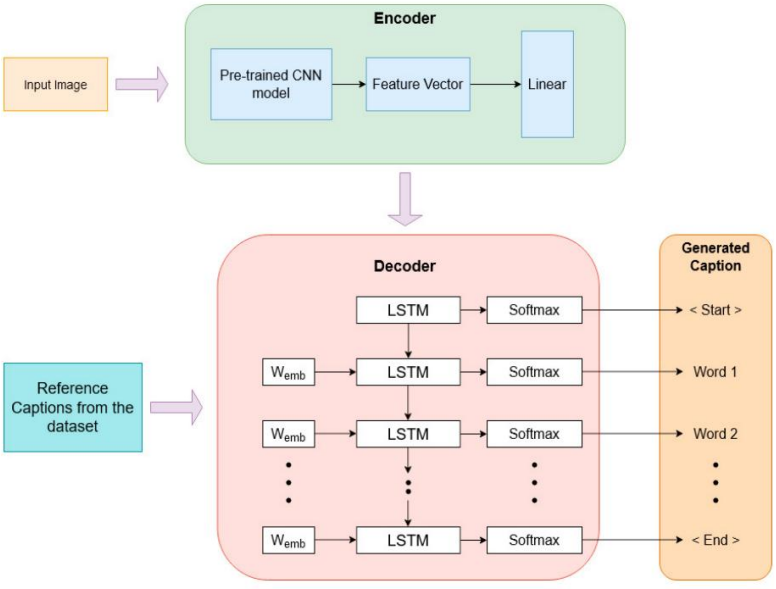
Rastogi et al. [9] đã phát triển một phương pháp độc đáo để phân loại bạch cầu bằng cách sử dụng mô hình trích xuất đặc điểm VGG16 tùy chỉnh. Mô hình trích xuất đặc điểm này được đào tạo theo 2 bước. Ở bước đầu tiên, một tập dữ liệu tăng cường với các lớp được đào tạo trước ở trạng thái đóng băng được sử dụng để đảm bảo rằng chỉ có lớp được kết nối đầy đủ mới được thêm vào mới nhận được các bản cập nhật trọng số và các lớp khác vẫn bị đóng băng. Ở bước thứ hai, tất cả các lớp đều được mở khóa và mô hình được đào tạo lại bằng cùng một tập dữ liệu nhưng với tốc độ học thấp. Các thí nghiệm cho thấy kỹ thuật dựa trên học sâu này có thể được sử dụng để trích xuất nhiều tính năng khác nhau với hiệu suất được cải thiện.

3. Phương pháp chi tiết

Phần này giải thích về thiết lập môi trường và phương pháp chi tiết để triển khai mô hình chú thích hình ảnh. Chúng ta hãy cùng khám phá chúng trong các tiểu mục tiếp theo:

3.1. Thiết lập môi trường

Dự án được triển khai trên máy tính chạy Windows có bộ xử lý Intel Core i7.
Tập lệnh được viết bằng Python 3.9 trong Jupyter Notebook. Hơn nữa, mô hình chú thích hình ảnh yêu cầu một lượng bộ nhớ đáng kể trong giai đoạn đào tạo của mô hình. Do đó, tất cả các thí nghiệm đều được chạy trên 16 GB RAM. Đối với mô hình ngôn ngữ, dự án cũ ng tận dụng các API cấp cao như Keras. Keras giúp triển khai hầu hết các mạng nơ-ron như CNN và RNN hoặc kết hợp cả CNN và RNN một cách dễ dàng. Hơn nữa, một số thư viện python được sử dụng để chạy mô hình mạng nơ-ron bao gồm Tensorflow, Numpy, Matplotlib, TQDM, Pillow và NLTK.



Hình 2. Khung giải mã bộ mã hóa

3.2. Các giai đoạn của mô hình Encoder-Decoder Để

tạo chú thích cho một hình ảnh nhất định, hai thông tin được lấy làm đầu vào: dữ liệu hình ảnh và dữ liệu văn bản. Nghiên cứu trước đây đã chứng minh rằng việc kết hợp CNN và RNN có thể cung cấp mô tả trực quan phong phú.
Kết quả là, mô hình mã hóa-giải mã trong nghiên cứu đề xuất sử dụng mô hình CNN-RNN. Công trình đề xuất sử dụng hai mô hình mạng nơ-ron: CNN được sử dụng để trích xuất các đặc điểm hình ảnh và LSTM được sử dụng để dịch các đặc điểm hình ảnh thành câu. Cấu trúc của mô hình tạo chú thích hình ảnh được thể hiện trong Hình 2.

Mô hình mã hóa-giải mã để chú thích hình ảnh có thể được xây dựng theo ba giai đoạn chính:

• Trích xuất tính năng •
Tiền xử lý văn bản • Mô hình
hóa ngôn ngữ

Trước khi đưa dữ liệu hình ảnh và văn bản vào mô hình, cả dữ liệu hình ảnh và văn bản đều phải được xử lý trước và chuyển đổi. Trong các tiểu mục tiếp theo, chúng ta sẽ xem xét cách dữ liệu hình ảnh và văn bản được xử lý trước trong các phương pháp của chúng tôi trước khi được đưa vào mô hình mã hóa-giải mã.

3.2.1. Trích xuất tính năng

Quá trình chú thích hình ảnh bắt đầu bằng việc trích xuất tính năng. Nó làm giảm chiều của kênh RGB thành biểu diễn không gian tiềm ẩn. Ban đầu, hình ảnh được chuyển đổi thành định dạng vectơ trước khi được đưa vào mạng nơ-ron. Một CNN được đào tạo trước [3] được sử dụng để chuyển đổi hình ảnh thành các vectơ có kích thước cố định. VGG-16, DenseNet, MobileNet và nhiều mô hình CNN được đào tạo trước khác có sẵn để trích xuất tính năng.

Đối với thí nghiệm này, chúng tôi đã sử dụng mô hình VGG-16 để kiểm tra một số cấu trúc mô hình. Ngoài ra, CNN đã được đào tạo trước để tránh quá khớp trong các mô hình chú thích hình ảnh. Lớp cuối cùng của mô hình CNN được sử dụng để dự đoán các đặc điểm hình ảnh. Do đó, lớp cuối cùng của CNN bị loại bỏ và mô hình hiện được triển khai bằng cách sử dụng lớp thứ hai từ cuối, lớp này trả về đặc điểm hình ảnh ở định dạng vector [14].

3.2.2. Tiền xử lý văn bản

Xử lý văn bản cũ ng là một trong những bước chính được sử dụng để tạo ra các chú thích chất lượng cao. Nó là cần thiết để xử lý trước dữ liệu văn bản và chuyển đổi nó thành dạng số trước khi đưa vào mạng nơ-ron.

Đối với quá trình tiền xử lý văn bản, trước tiên, tất cả các từ số, chữ cái đặc biệt và dấu câu đều bị loại bỏ trong giai đoạn tiền xử lý văn bản. Ngoài ra, một số từ không hữu ích cũ ng bị loại khỏi từ điển. Thứ hai, dữ liệu văn bản được chú thích bằng thẻ bắt đầu và kết thúc. Điều này giúp máy xác định câu bắt đầu và kết thúc ở đâu. Dưới đây là ví dụ về chú thích có thẻ bắt đầu và kết thúc.

startedseq endseq

Hơn nữa, văn bản không thể được xử lý trực tiếp bởi mạng nơ-ron. Văn bản phải được chuyển đổi thành dạng số. Điều này có thể dễ dàng thực hiện bằng cách sử dụng trình phân tích Keras.

3.2.3. Mô hình hóa ngôn ngữ

Trong bước thứ ba của chú thích hình ảnh, văn bản được xử lý trước và các đặc điểm hình ảnh được trích xuất được đưa vào mô hình chú thích hình ảnh để tạo ra các mô tả. Để đạt được điều này, các mô hình ngôn ngữ khác nhau như LSTM, GRU hoặc sự kết hợp của các mô hình hiện có khác thường được sử dụng trong tài liệu. Đối với bài báo này, chúng tôi sử dụng các mạng hồi quy LSTM [13].

Lớp Bộ nhớ dài hạn ngắn cũ ng là một loại RNN. Nó chủ yếu được sử dụng để truyền dữ liệu từ ô này sang ô khác và để tạo ra một từ hoàn chỉnh. Nó có thể xử lý nhiều loại dữ liệu, bao gồm video và âm thanh, ngoài hình ảnh. Một LSTM thông thường bao gồm ba cổng, tức là cổng vào, cổng ra và cổng quên. Các cổng này của LSTM kiểm soát chuyển động của thông tin. Đối với việc phân loại, xử lý và dự đoán, dữ liệu chuỗi thời gian là lý tưởng cho các mạng LSTM.

Hãy cùng xem cách LSTM được sử dụng để tạo chú thích hình ảnh cho hình ảnh đã cho:

Hình 3. Hình ảnh một con mèo

Caption ->

Từ vựng cho chú thích này sẽ là:

Từ vựng -> , , , , ngày , , startedeq, endseq

Chú thích đầu ra:

Từ bảng 1, rõ ràng là trước tiên, từ mục tiêu được dự đoán bằng LSTM, sau đó là từ mục tiêu từ được nối vào phần chú thích.

4. Các mô hình đề xuất cho thử nghiệm của chúng tôi

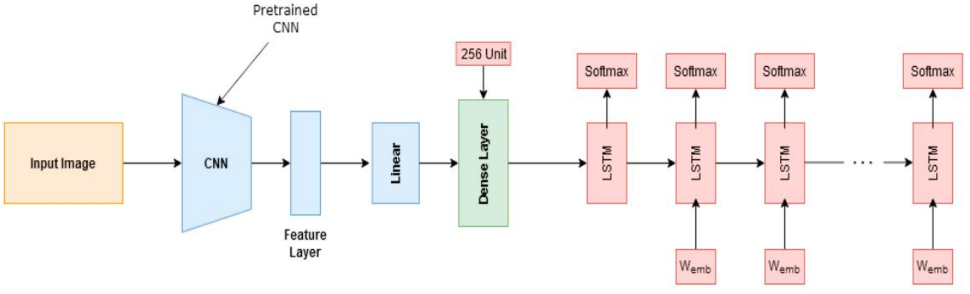
Trong thử nghiệm của chúng tôi, nhiều mô hình khác nhau được triển khai có các lớp và thông số khác nhau. Chúng tôi có đã sử dụng VGG-16 làm mô hình CNN được đào tạo trước. Chúng ta hãy xem xét một số mô hình mã hóa-giải mã:

4.1. Mô hình

1 Cấu trúc của mô hình 1 có thể được giải thích như sau:

Bảng 1. Các chú thích một phần được tạo ra trong mỗi lần lặp lại

Lặp lại	Tính năng hình ảnh Vector	Chú thích một phần	mục tiêu từ
1	hình ảnh vector	startedeq	
2	hình ảnh vector	startedeq	
3	hình ảnh vector	startedeq	
4	hình ảnh vector	startedeq	
5	hình ảnh vector	startedeq	ngày
6	hình ảnh vector	startedeq	
7	hình ảnh vector		
8	hình ảnh vector hình ảnh vector	startedeq Bắt đầu công việc của bạn.	kết thúc



Hình 4. Mô hình 1

1. Đầu vào tính năng hình ảnh: Ban đầu, VGG-16 được sử dụng để xây dựng một vectơ tính năng hình ảnh, đó là đưa vào mô hình mã hóa. Với điều này, hình dạng đầu vào của bộ mã hóa hình ảnh được đặt thành kích thước của trích xuất đặc điểm hình ảnh. Ngoài ra, một lớp bỏ qua với tỷ lệ bỏ qua là 0,3 được thêm vào. Cuối cùng, sử dụng Hàm kích hoạt ReLU trong lớp dày đặc, vectơ đặc trưng hình ảnh được nén thành 256 phần tử.

2. Đầu vào tính năng văn bản: Mô tả hình ảnh được sử dụng làm đầu vào thứ hai trong bộ mã hóa văn bản mô hình. Nó dự đoán một chuỗi đầu vào, được truyền tiếp qua lớp mã hóa. Sau đây rằng, có một lớp bỏ học với tỷ lệ bỏ học là 0,3. Ở giai đoạn cuối, lớp LSTM được sử dụng, có 256 đơn vị bộ nhớ và tạo ra đầu ra vector 256 phần tử.

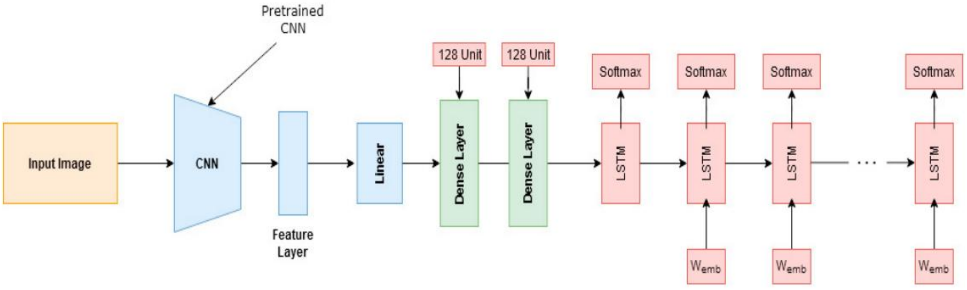
3. Bộ giải mã: Cuối cùng, bộ trích xuất tính năng và mô hình trình tự được đưa vào mô hình bộ giải mã. Cả hai các mô hình đầu vào tạo ra một vectơ 256 phần tử, sau đó được cộng lại với nhau. Thông tin sau đó được được truyền đến một lớp dày đặc 256 tế bào thần kinh. Sau đó, nó tạo ra một dự báo softmax của một từ ở cuối quá trình. Sau đó, mô hình được chạy trong 25 kỷ nguyên và biểu đồ Mất mát so với Kỷ nguyên cho mô hình 1 là được tạo ra. Đồ thị được mô tả trong hình 8.

4.2. Mô hình 2

Cấu trúc của mô hình 2 có thể được giải thích như sau:

1. Đầu vào tính năng hình ảnh: Ban đầu, VGG-16 được sử dụng để xây dựng một vectơ tính năng hình ảnh, đó là đưa vào mô hình mã hóa. Với điều này, hình dạng đầu vào của bộ mã hóa hình ảnh được đặt thành kích thước của trích xuất đặc điểm hình ảnh. Ngoài ra, một lớp bỏ qua với tỷ lệ bỏ qua là 0,5 được thêm vào. Cuối cùng, sử dụng Hàm kích hoạt ReLU trong lớp dày đặc, vectơ đặc trưng hình ảnh được nén thành 128 phần tử.

2. Đầu vào tính năng văn bản: Chú thích tham chiếu của hình ảnh được sử dụng làm đầu vào trong bộ mã hóa văn bản mô hình. Nó dự đoán một chuỗi đầu vào, được truyền tiếp qua lớp mã hóa. Sau đây rằng, có một lớp bỏ học với tỷ lệ bỏ học là 0,5. Ở giai đoạn cuối, lớp LSTM được sử dụng,

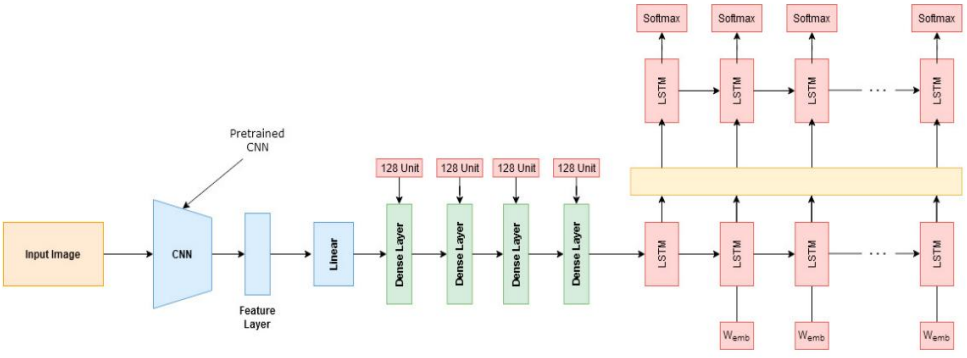


Hình 5. Mô hình 2

có 128 đơn vị bộ nhớ và tạo ra đầu ra vector gồm 128 phần tử.

3. Bộ giải mã: Cuối cùng, bộ trích xuất tính năng và các mô hình trình tự được đưa vào mô hình bộ giải mã. Cả hai mô hình đầu vào đều tạo ra một vectơ 128 phần tử, sau đó được cộng lại với nhau. Thông tin sau đó được truyền đến một lớp dày đặc 128 nơ-ron. Sau đó, nó tạo ra một dự báo softmax của một từ ở cuối quá trình. Sau đó, mô hình được chạy trong 25 kỷ nguyên và biểu đồ Mất mát so với Kỷ nguyên cho mô hình 2 được tạo ra. Biểu đồ được mô tả trong hình 8.

4.3. Mô hình 3



Hình 6. Mô hình 3

Cấu trúc của mô hình 3 có thể được giải thích như sau:

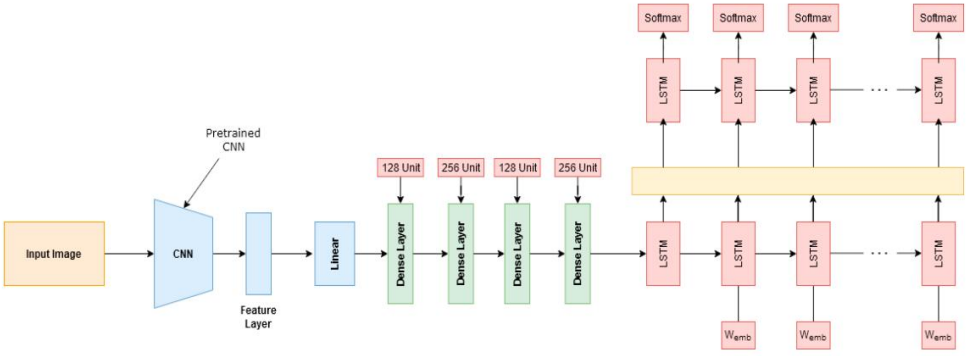
1. Đầu vào tính năng hình ảnh: Ban đầu, VGG-16 được sử dụng để xây dựng một vectơ tính năng hình ảnh, được đưa vào mô hình mã hóa. Với điều này, hình dạng đầu vào của bộ mã hóa hình ảnh được đặt thành kích thước của tính năng hình ảnh được trích xuất. Ngoài ra, một lớp bỏ qua với tỷ lệ bỏ qua là 0,5 được thêm vào. Cuối cùng, sử dụng hàm kích hoạt ReLU trong lớp dày đặc, vectơ tính năng hình ảnh được nén thành 128 phần tử trong 4 lớp dày đặc tiếp theo.

2. Đầu vào tính năng văn bản: Các chú thích tham chiếu của hình ảnh được sử dụng làm đầu vào trong mô hình mã hóa văn bản. Nó dự đoán một chuỗi đầu vào, được truyền tiếp qua lớp mã hóa. Tiếp theo đó, có một lớp bỏ qua với tỷ lệ bỏ qua là 0,5. Ở giai đoạn cuối, hai lớp LSTM song song được sử dụng, có 128 đơn vị bộ nhớ và tạo ra đầu ra vectơ gồm 128 phần tử.

3. Bộ giải mã: Cuối cùng, bộ trích xuất tính năng và các mô hình trình tự được đưa vào mô hình bộ giải mã. Cả hai mô hình đầu vào đều tạo ra một vectơ 128 phần tử, sau đó được cộng lại với nhau. Thông tin sau đó được truyền đến một lớp dày đặc 128 nơ-ron. Sau đó, nó tạo ra một dự báo softmax của một từ ở cuối

quá trình. Sau đó, mô hình được chạy trong 25 kỷ nguyên và biểu đồ Mất mát so với Kỷ nguyên cho mô hình 3 được tạo ra. Biểu đồ được mô tả trong hình 8.

4.4. Mô hình 4



Hình 7. Mô hình 4

Cấu trúc của mô hình 4 có thể được giải thích như sau:

1. Đầu vào tính năng hình ảnh: Ban đầu, VGG-16 được sử dụng để xây dựng một vectơ tính năng hình ảnh, được đưa vào mô hình mã hóa. Với điều này, hình dạng đầu vào của bộ mã hóa hình ảnh được đặt thành kích thước của tính năng hình ảnh được trích xuất. Ngoài ra, một lớp bỏ qua với tỷ lệ bỏ qua là 0,5 được thêm vào. Cuối cùng, sử dụng hàm kích hoạt ReLU trong lớp dày đặc, vectơ tính năng hình ảnh được nén thành 128 và 256 phần tử sau đó trong 4 lớp dày đặc.

2. Đầu vào tính năng văn bản: Các chú thích tham chiếu của hình ảnh được sử dụng làm đầu vào trong mô hình mã hóa văn bản. Nó dự đoán một chuỗi đầu vào, được truyền tiếp qua lớp mã hóa. Tiếp theo là một lớp bỏ qua với tỷ lệ bỏ qua là 0,5. Ở giai đoạn cuối, hai lớp LSTM song song được sử dụng, có 256 đơn vị bộ nhớ và tạo ra đầu ra vectơ 256 phần tử.

3. Bộ giải mã: Cuối cùng, bộ trích xuất tính năng và các mô hình trình tự được đưa vào mô hình bộ giải mã. Cả hai mô hình đầu vào đều tạo ra một vectơ 128 phần tử, sau đó được cộng lại với nhau. Thông tin sau đó được truyền đến một lớp dày đặc 256 nơ-ron. Sau đó, nó tạo ra một dự báo softmax của một từ ở cuối quá trình. Sau đó, mô hình được chạy trong 25 kỷ nguyên và biểu đồ Mất mát so với Kỷ nguyên cho mô hình 4 được tạo ra. Biểu đồ được mô tả trong hình 8.

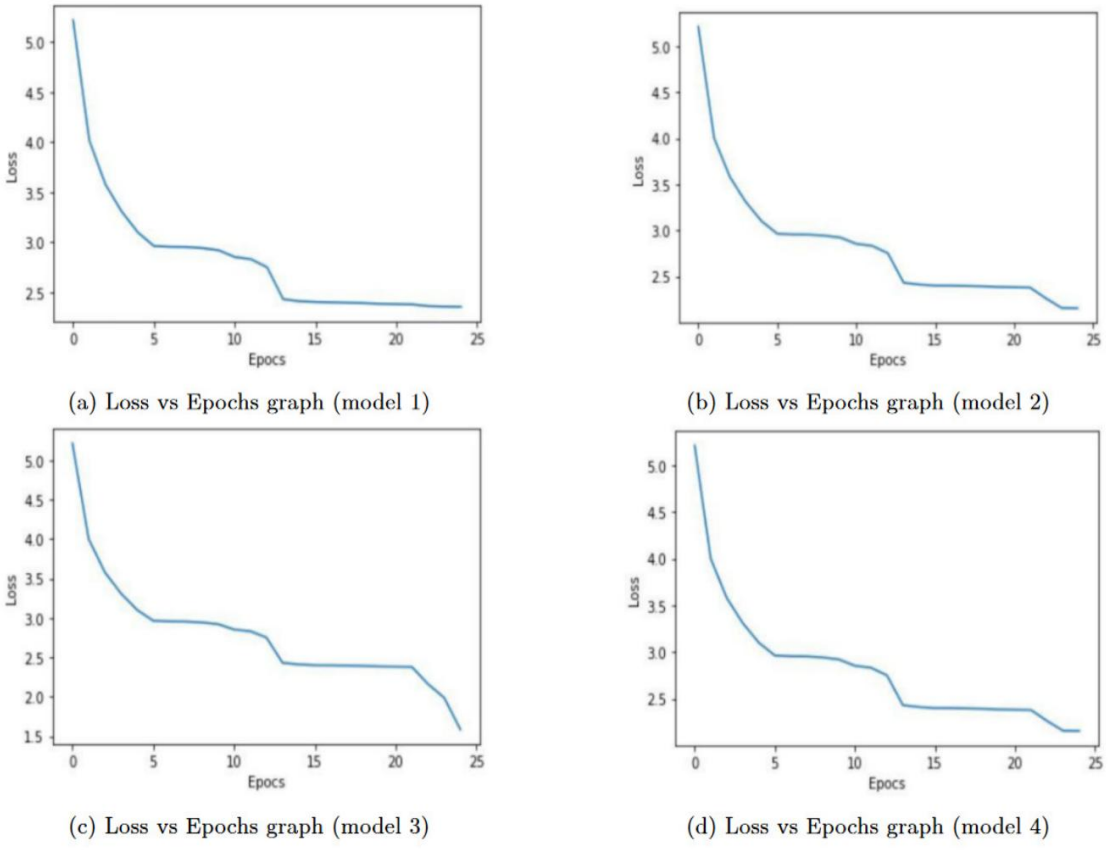
5. Kết quả thực nghiệm và phân tích

Phần này giới thiệu cho bạn tập dữ liệu được xem xét cho thí nghiệm của chúng tôi và phân tích của nó. Sau đó, chúng tôi chứng minh hiệu quả của mô hình đề xuất của chúng tôi trên tập dữ liệu đã cho và trình bày so sánh với các công trình tiên tiến khác.

5.1. Kết quả thực nghiệm

Các thí nghiệm đã được thực hiện trên tập dữ liệu Flickr8k Hindi. Tập dữ liệu Flickr8k Hindi bao gồm 5 chú thích tiếng Hindi cho mỗi hình ảnh. Tập dữ liệu này bao gồm 8.000 hình ảnh đào tạo và 1.000 hình ảnh được dành riêng để thử nghiệm và xác thực. Hơn nữa, vì mục đích thử nghiệm, 25 kỷ nguyên đã được chạy cho mỗi mô hình, sau đó kết quả được tạo ra. Biểu đồ mất mát so với kỷ nguyên được tạo ra được mô tả trong hình 8.

Để kiểm tra hiệu quả của mô hình, chúng tôi đã so sánh kết quả với mô hình Encoder-Decoder của A. Rathi cho tập dữ liệu tiếng Hindi Flickr8k. So với kết quả của A. Rathi, người ta thấy rằng mô hình của chúng tôi đạt được điểm BLEU cao hơn [8] đối với dữ liệu hình ảnh và văn bản được xử lý trước. Do đó, thí nghiệm của chúng tôi đã



(a) Loss vs Epochs graph (model 1)

(b) Loss vs Epochs graph (model 2)

(c) Loss vs Epochs graph (model 3)

(d) Loss vs Epochs graph (model 4)

Hình 8. Biểu đồ Tổn thất được tạo ra so với Epochs cho các mô hình khác nhau

đạt được kết quả tiên tiến nhất về điểm BLEU. Bảng 2 mô tả chú thích hình ảnh được đề xuất
Điểm BLEU của mô hình so với điểm thu được bởi A. Rathi trên tập dữ liệu tiếng Hindi Flickr8k.

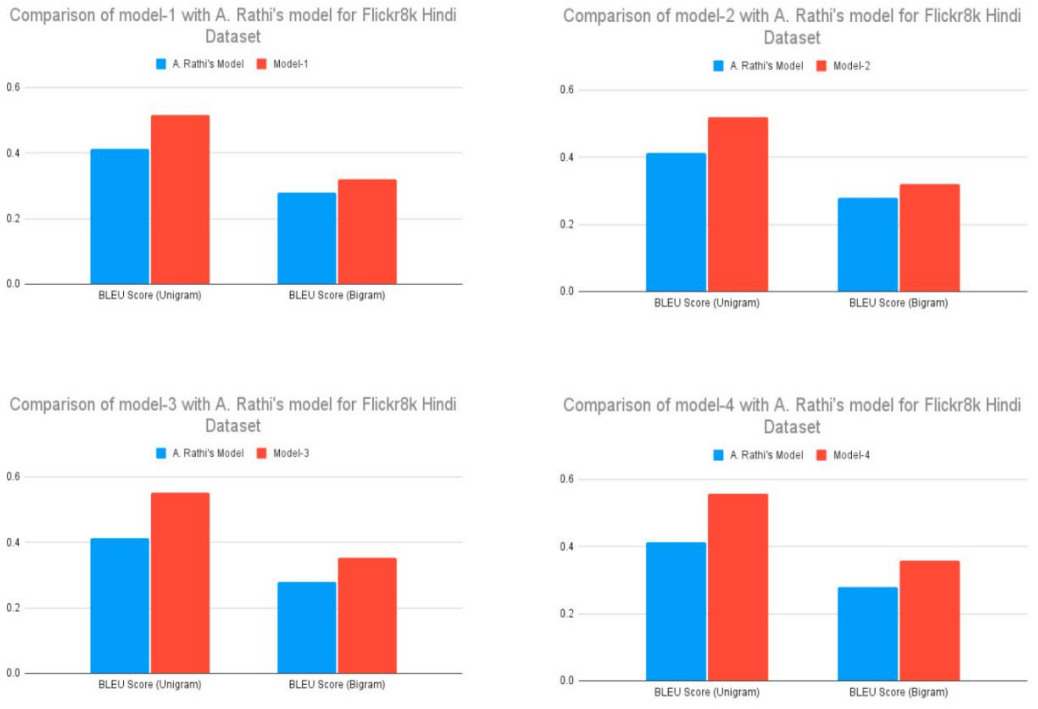
Bảng 2. So sánh công việc đề xuất với công việc hiện có

Người mẫu	Bộ dữ liệu	Điểm BLEU (Unigram)	Điểm BLEU (Bigram)
Rathi A. [11]	Bộ dữ liệu tiếng Hindi Flickr8k	0,4136	0,278
Mô hình đề xuất 1 Mô	Bộ dữ liệu tiếng Hindi Flickr8k	0,51688	0,29309
hình đề xuất 2 Mô hình	Bộ dữ liệu tiếng Hindi Flickr8k	0,52022	0,31973
đề xuất 3 Mô hình đề	Bộ dữ liệu tiếng Hindi Flickr8k	0,55257	0,35166
xuất 4	Bộ dữ liệu tiếng Hindi Flickr8k	0,55698	0,35914

5.2. So sánh kết quả

Trong phần này, các kết quả thu được từ công trình đề xuất được so sánh bằng đồ họa với các kết quả hiện có. công việc tức là mô hình của A. Rathi. Từ hình 9, có thể thấy rằng các mô hình chúng tôi đề xuất tức là mô hình-1, mô hình-2, mô hình-3 và mô hình-4 đã cho thấy những cải thiện đáng kể về điểm BLEU. Bây giờ, chúng ta hãy phân tích đồ thị kết quả theo điểm BLEU cho Bộ dữ liệu tiếng Hindi Flickr8k.

- So với công trình hiện có, có thể thấy mô hình 1 của chúng tôi đã cho thấy mức tăng 24,95% đối với điểm BLEU (Unigram) và 5,39% đối với điểm BLEU (Bigram).



Hình 9. So sánh các mô hình được đề xuất với mô hình của A. Rathi cho Flickr8k Hindi Dataset

- So với công trình hiện có, có thể thấy rằng mô hình-2 của chúng tôi đã cho thấy sự gia tăng 25,77% đối với điểm BLEU (Unigram) và 14,74% đối với điểm BLEU (Bigram). • So với công trình hiện có, có thể thấy rằng mô hình-3 của chúng tôi đã cho thấy sự gia tăng 33,58% đối với điểm BLEU (Unigram) và 26,47% đối với điểm BLEU (Bigram). • So với công trình hiện có, có thể thấy rằng mô hình-4 của chúng tôi đã cho thấy sự gia tăng 34,64% đối với điểm BLEU (Unigram) và 29,13% đối với điểm BLEU (Bigram).

Như vậy, từ phân tích so sánh trên, có thể kết luận rằng đối với Bộ dữ liệu tiếng Hindi Flickr8k đã cho, mô hình có số lớp dày đặc và LSTM cao hơn đã cho thấy điểm BLEU được cải thiện so với mô hình mã hóa-giải mã dựa trên CNN-LSTM truyền thống.

6. Kết luận và phạm vi tương lai

Bài báo này đề xuất một mô hình Encoder-Decoder trên tập dữ liệu tiếng Hindi Flickr8k sử dụng CNN được đào tạo trước (VGG16) để trích xuất tính năng và sử dụng LSTM để mô hình hóa ngôn ngữ. Để chứng minh tính hiệu quả của khuôn khổ Encoder-Decoder cho tập dữ liệu chú thích tiếng Hindi, điểm BLEU đã được tính toán trên nhiều mô hình chú thích hình ảnh khác nhau. Các mô hình này được tối ưu hóa bằng cách điều chỉnh các siêu tham số và thay đổi các lớp ẩn trong khuôn khổ hiện có. Rõ ràng từ các kết quả thử nghiệm và phân tích rằng mô hình mạng nơ-ron CNN-LSTM nhiều lớp đã cho thấy sự cải thiện đáng kể về điểm BLEU trái ngược với mô hình mạng nơ-ron CNN-LSTM truyền thống. Những phát hiện của thử nghiệm này có thể đóng vai trò là chuẩn mực cho các nghiên cứu trong tương lai trong lĩnh vực học sâu.

Công trình được cung cấp ở đây mở đường cho nghiên cứu sâu hơn trong lĩnh vực này. Đây chỉ là giải pháp cắt đầu tiên và có nhiều cách để cải thiện hơn nữa. Trong tương lai, nghiên cứu này có thể được mở rộng bằng cách triển khai mô hình VGG16 tùy chỉnh được tinh chỉnh [9] và sử dụng một số tập dữ liệu lớn. Ngoài ra, có thể triển khai tích hợp nhận dạng đối tượng với mô hình tích chập để làm cho mô hình mạnh mẽ hơn. Hiệu quả của mô hình cũ ng có thể được cải thiện bằng cách thêm mô-đun chú ý vào bộ mã hóa-giải mã hiện có

kiến trúc. Do đó, mặc dù chú thích hình ảnh đã có những tiến bộ vượt bậc trong những năm gần đây, nhưng vẫn luôn có chỗ để cải thiện trong tương lai.

Tài liệu tham khảo

[1] Al-Malla, MA, Jafar, A., Ghneim, N. (2022) "Mô hình chú thích hình ảnh sử dụng sự chú ý và các đặc điểm của đối tượng để mô phỏng khả năng hiểu hình ảnh của con người" Tạp chí Dữ liệu lớn 9.1: 1-16.

[2] Farhadi, A., Hejrati, M., Sadeghi, MA, Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D. (2010) "Mỗi bức tranh kể một câu chuyện: Tạo câu từ hình ảnh" Hội nghị châu Âu về thị giác máy tính. Springer, Berlin, Heidelberg.

[3] Gu, J., Wang, G., Cai, J., Chen, T. (2017) "Một nghiên cứu thực nghiệm về ngôn ngữ cnn cho chú thích hình ảnh" Biên bản báo cáo Hội nghị quốc tế IEEE về thị giác máy tính.

[4] Kaur, J., Josan, GS (2020) "Dịch chú thích hình ảnh đa phương thức từ tiếng Anh sang tiếng Hindi" Tạp chí nghiên cứu khoa học 64.2.

[5] Kiros, R., Salakhutdinov,R., Zemel, RS (2014) "Thống nhất nhúng ngữ nghĩa thị giác với ngôn ngữ thần kinh đa phương thức mô hình" bản in trước arXiv arXiv:1411.2539.

[6] Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, AC, Berg, TL (2013) "Babytalk: Hiểu và tạo ra các mô tả hình ảnh đơn giản" Giao dịch IEEE về phân tích mẫu và trí tuệ máy móc 35.12: 2891-2903.

[7] Mishra, SK, Dhir, R., Saha, S., Bhattacharyya, P., Singh, AK (2021) "Chú thích hình ảnh bằng tiếng Hindi sử dụng trans- "mạng cũ " Máy tính và Kỹ thuật Điện 92 (2021): 107114.

[8] Papineni, K., Roukos, S., Ward, T., Zhu, W. (2002) "Bleu: một phương pháp đánh giá tự động bản dịch máy" Biên bản cuộc họp thường niên lần thứ 40 của Hiệp hội Ngôn ngữ học tính toán.

[9] Rastogi, P., Khanna, K., Singh, V. (2022) "LeuFeatx: Công cụ trích xuất tính năng dựa trên học sâu để chẩn đoán bệnh bạch cầu cấp tính từ hình ảnh hiển vi của vết máu ngoại vi" Máy tính trong Sinh học và Y học 142 (2022): 105236.

[10] Rastogi, P., Khanna, K., Singh, V. (2022) "Phân đoạn tuyến trong hình ảnh mô bệnh học ung thư trực tràng bằng U-net mạng tích chập lấy cảm hứng từ" Máy tính nơ-ron và ứng dụng 34.7: 5383-5395.

[11] Rathi, A. (2020) "Phương pháp học sâu để chú thích hình ảnh bằng tiếng Hindi" Hội nghị quốc tế năm 2020 về Kỹ thuật máy tính, điện và truyền thông (ICCECE), IEEE.

[12] Srinivasan, L., Sreekanthan, D., Amutha, AL (2018) "Chú thích hình ảnh-một phương pháp học sâu" Int. J. Appl. Eng. Res 13.9 : 7239-7242.

[13] Tanti, M., Gatt, A., Camilleri, KP (2017) "Vai trò của mạng nơ-ron hồi quy (rnns) trong chú thích hình ảnh là gì? máy phát điện?." Bản in trước của arXiv arXiv:1708.02043.

[14] Tanti, M., Gatt, A., Camilleri, KP (2018) "Nên đặt hình ảnh ở đâu trong trình tạo chú thích hình ảnh" Kỹ thuật ngôn ngữ tự nhiên 24.3: 467-489.

[15] Vinyals, O., Toshev, A., Bengio, S., Erhan, D. (2015) "Trình bày và kể: Một trình tạo chú thích hình ảnh thần kinh" Biên bản báo cáo Hội nghị IEEE về thị giác máy tính và nhận dạng mẫu.

[16] Xiao, X., Wang, L., Ding, K., Xiang, S., Pan, C. (2019) "Mạng mã hóa-giải mã phân cấp sâu để chú thích hình ảnh" Giao dịch IEEE về đa phương tiện 21.11: 2942-2956.

[17] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y. (2015) "Trình bày, tham dự và kể: Tạo chú thích hình ảnh bằng nơ-ron với sự chú ý trực quan" Hội nghị quốc tế về máy học. PMLR.