Image Captioning

*Investigate image captioning models, select one model to conduct surveys, evaluations, and feedback, and propose improvement directions.

21120071 - Nguyễn Thị Thanh Hoa
University of Science, HCMC
nguyenhoatp29@gmail.com
0342579403

21120175 - Tô Ngọc Hân
University of Science, HCMC
tongochan230@gmail.com
0835813863

21120184 - Lê Thị Minh Thư
University of Science, HCMC
21120184@student.hcmus.edu.vn
0888392122

***Tóm tắt nội dung*—Tài liệu này trình bày về các phương pháp để tạo chú thích tự động cho ảnh (image captioning). Dựa trên các phương pháp đó, nhóm đã đề xuất, xây dựng và thử nghiệm với mô hình CNN-LSTM, sử dụng cơ chế Encoder-Decoder.**

***Index Terms*—Image Captioning, Text Generation, Encoder-Decoder Captioning.**

I. Giới thiệu

- Image Captioning là quá trình tự động sinh ra mô tả văn bản cho các hình ảnh đầu vào. Mô tả này thường phản ánh các đặc điểm quan trọng của hình ảnh và có thể bao gồm thông tin về các đối tượng, hành động, và ngữ cảnh.
- **Input**: Đầu vào của bài toán Image Captioning là một hình ảnh. Mỗi hình ảnh sẽ được biểu diễn dưới dạng vector đặc trưng bằng các phương pháp trích xuất đặc trưng như CNN (Convolutional Neural Networks).
- **Dataset**: Các tập dữ liệu thường được sử dụng trong nghiên cứu chú thích hình ảnh bao gồm COCO, Flickr8k, Flickr30k, và MSCOCO. Mỗi tập dữ liệu có các đặc điểm và mục tiêu sử dụng riêng.
- **General system architecture**: Cấu trúc tổng quát của hệ thống Image Captioning bao gồm hai phần chính: mô hình trích xuất đặc trưng hình ảnh và mô hình sinh mô tả văn bản. Mô hình trích xuất đặc trưng thường sử dụng CNN để biểu diễn hình ảnh dưới dạng vector. Sau đó, vector đặc trưng này được đưa vào mô hình sinh mô tả, thường là một mạng NLP (Natural Language Processing) như RNN (Recurrent Neural Networks) hoặc Transformer.
- **Metrics**: Các chỉ số đánh giá phổ biến cho chú thích hình ảnh bao gồm BLEU, METEOR, CIDEr, và ROUGE. Các chỉ số này đánh giá chất lượng của mô tả ngôn ngữ so với mô tả thực tế, dựa trên các tiêu chí như độ chính xác, độ tương đồng, và sự liên quan giữa mô tả.

II. Các công trình nghiên cứu liên quan

"Visuals to text: A comprehensive review on automatic image captioning" của Y. Ming, N. N. Hu, C. X. Fan, F. Feng, J. W. Zhou, và H. Yu công bố tại IEEE/CAA Journal of Automatica Sinica, pp. 1339–1365 năm 2022. [1] tổng hợp các phương pháp truyền thống và các kỹ thuật dựa trên học sâu (deep learning) trong lĩnh vực tạo chú thích hình ảnh.

A. Phương pháp truyền thống

"Retrieval-augmented Image Captioning" của Rita Ramos, Desmond Elliott, và Bruno Martins [2] đề xuất một cách tiếp cận mới về chú thích hình ảnh, tận dụng các kỹ thuật truy xuất.

1) Retrieval-Based Image Captioning:

- **Nguyên lý**: Sử dụng cơ sở dữ liệu có chứa các cặp hình ảnh-chú thích để tạo ra mô tả cho hình ảnh mới.
- **Phương pháp**:
  - **So sánh Tương tự**: Đầu tiên, mô hình tìm kiếm các hình ảnh tương tự với hình ảnh đầu vào từ cơ sở dữ liệu bằng cách so sánh độ tương tự (có thể dùng nguyên tác Tree-F1)
  - **Chọn Chú thích**: mô hình chọn chú thích tốt nhất từ các hình ảnh được truy xuất, dựa trên độ tương tự và các tiêu chí khác như độ chính xác và đặc thù.
  - **Xử lý chú thích**: Một số phương pháp đơn giản chỉ sử dụng chú thích đã được truy xuất mà không có xử lý thêm. Tuy nhiên, một số phương pháp khác sử dụng các kỹ thuật xử lý để tạo ra một chú thích mới cho hình ảnh dựa trên các phần của các chú thích đã được truy xuất. Bao gồm sử dụng luật láng giềng gần nhất k-NN (nearest neighbor), tính toán sự tương tự toàn cục giữa các hình ảnh, sử dụng thông tin về cú pháp để tìm kiếm hành động và chủ đề, và sử dụng kỹ thuật ước lượng mật độ phi tham số để giảm thiểu ảnh hưởng của nhiễu ước lượng hình ảnh.
- **Giải quyết được thách thức gì trong bài toán image captioning**: Tổng quát hóa các mô tả dựa trên những gì đã được học từ dữ liệu huấn luyện, xây dựng được mối tương quan giữa văn bản-hình ảnh, có thể tận dụng được cấu trúc của dữ liệu huấn luyện để học cách tập trung vào các đặc trưng quan trọng của hình ảnh khi tạo ra mô tả
- **Nhược điểm còn tồn đọng khi giải bài toán image captioning**: Phụ thuộc quá nhiều vào cơ sở dữ liệu chú thích hiện có, giới hạn việc tạo ra các câu chú thích mới và không thể chứa được các đối tượng hoặc cảnh mới.

2) Template-Based Image Captioning:

- **Nguyên lý**: Phương pháp dùng các ràng buộc cú pháp và ngữ nghĩa để tạo ra các mô tả cho hình ảnh. Thông thường, đây là một mô hình dựa trên dữ liệu, đã xác định

trước các quy tắc cú pháp. Nó nhận diện và trích xuất các yếu tố liên quan như đối tượng, hành động, cảnh quan, mối quan hệ và chuyển chúng thành biểu diễn ngữ nghĩa để dự đoán nhãn ngôn ngữ, sau đó điền các nhãn vào mẫu đã xác định trước để tạo ra các mô tả.

- **Phương pháp**: Mô hình này thường sử dụng các phương pháp như máy vector hỗ trợ SVM (công trình được trình bày bởi Farhadi et al. [3]), mô hình Markov ẩn HMM(Công trình được trình bày bởi Rahman et al. [4]), trường ngẫu nhiên có điều kiện CRF (Công trình được trình bày bởi Kulkarni et al. [5]), cấu trúc cây phụ thuộc (được trình bày bởi Xu et al [6]) và các mô hình kết hợp khác để trích xuất các yếu tố từ hình ảnh và điền vào mẫu để tạo ra các mô tả.
- **Giải quyết được thách thức gì trong bài toán image captioning**: Phương pháp này giúp giải quyết vấn đề về tính chính xác cú pháp của câu mô tả hình ảnh. Nó giúp đảm bảo rằng mô tả được tạo ra từ hình ảnh là hợp lý cú pháp và ngữ nghĩa.
- **Nhược điểm còn tồn đọng khi giải bài toán image captioning**: Mặc dù phương pháp này giúp tạo ra các câu mô tả có cú pháp đúng, nhưng nó có một số hạn chế. Mô hình phụ thuộc nhiều vào mẫu đã xác định trước, làm cho độ dài của các mô tả tạo ra không thể thay đổi và nội dung mô tả tương đối đơn giản. Ngoài ra, phương pháp này cần chú thích nhiều đối tượng, thuộc tính và mối quan hệ của hình ảnh, làm cho việc xử lý dữ liệu quy mô lớn và không thể áp dụng cho tất cả các loại hình ảnh.

B. Phương pháp Deep Learning

"DEEP LEARNING APPROACHES ON IMAGE CAPTIONING: A REVIEW" của tác giả Taraneh Ghandi, Hamidreza Pourreza và Hamidreza Mahyar [7] cung cấp một cái nhìn tổng quan toàn diện về các kỹ thuật học sâu áp dụng cho chú thích hình ảnh.

1) Encoder-Decoder Framework:

- **Encoder-Decoder Framework (Bộ khung mã hóa-giải mã)** là một nền tảng quan trọng trong lĩnh vực học sâu, được áp dụng rộng rãi trong các mô hình chú thích hình ảnh.
- **Nguyên lý**: Nguyên lí cơ bản của bộ khung này là kết hợp hai phần chính: Mã hóa (Encoder) và Giải mã (Decoder). Mã hóa chịu trách nhiệm trích xuất và biểu diễn thông tin từ hình ảnh, trong khi giải mã tạo ra câu chú thích dựa trên thông tin đã được mã hóa. Điều này tạo ra một cơ chế mạnh mẽ để tự động tạo ra chú thích cho hình ảnh.
- **Phương pháp**:
  - **Mã hóa (Encoder)**: Bộ phận mã hóa là một mạng nơ-ron tích chập (CNN) được sử dụng để phân tích hình ảnh và trích xuất các đặc trưng hình ảnh. Các đặc trưng này bao gồm thông tin về các đối tượng, cấu trúc, màu sắc, và các yếu tố quan trọng khác trong hình ảnh.
  - **Giải mã (Decoder)**: Bộ phận giải mã là một mạng nơ-ron hồi quy (RNN) hoặc các biến thể của nó như LSTM (Long Short-Term Memory). Mục tiêu của bộ phận giải mã là tạo ra một chuỗi từ, hay cụ thể hơn là một câu chú thích, dựa trên các đặc trưng đã được mã hóa. Mô hình giải mã sẽ học cách tạo ra từng từ một cách tuần tự, tạo thành một câu chú thích mô tả nội dung của hình ảnh.

Bộ khung (framework) này xử lý chú thích hình ảnh như một tác vụ dịch, với hình ảnh là đầu vào và các câu là đầu ra. Để tạo ra một câu giống con người hơn, các nhà nghiên cứu đã thực hiện nhiều cải tiến sáng tạo dựa trên khuôn khổ cơ bản. Phần này sẽ tóm tắt các công trình phát triển này từ góc độ mã hóa hình ảnh và giải mã ngôn ngữ tương ứng:

- **Mã hóa hình ảnh (Visual Encoding)**:
  - Học biểu diễn tích chập (Convolutional representation learning): Phương pháp này liên quan đến việc sử dụng CNN để trích xuất các đặc điểm từ hình ảnh. Các cách tiếp cận khác nhau sử dụng các biến thể trong cách trích xuất và sử dụng các tính năng, chẳng hạn như sử dụng thông tin kích hoạt từ các lớp khác nhau hoặc kết hợp các thành phần đa phương thức.
  - Học biểu diễn đồ thị (Graph representation learning): Sử dụng đồ thị cảnh để nắm bắt các mối quan hệ ngữ nghĩa cấp cao giữa các đối tượng trong ảnh. Sau đó, mạng tích chập đồ thị (GCN) được sử dụng để tìm hiểu cách biểu diễn đồ thị, cho phép trích xuất thông tin ngữ nghĩa có cấu trúc chặt chẽ hơn.
  - Học biểu diễn sự chú ý (Attention representation learning): Các mô hình gần đây thay thế kiến trúc CNN-LSTM truyền thống bằng Transformers dựa trên cơ chế tự chú ý. Những mô hình này mã hóa hình ảnh thành các đặc điểm gây chú ý, nắm bắt bối cảnh hình ảnh tổng thể mà không bị phức tạp.
- **Giải mã ngôn ngữ (Language Decoding)**:
  - Giới thiệu thông tin trực quan (Introducing visual information): Các phương pháp như thêm thông tin ngữ nghĩa trực quan vào đơn vị LSTM hoặc sử dụng vectơ "trực giác" bên cạnh các trạng thái ẩn LSTM để tăng cường sự đóng góp của thông tin hình ảnh vào việc tạo từ.
  - Các biến thể LSTM định hướng và sâu (Directional and deep LSTM variants): LSTM hai chiều và cấu trúc LSTM sâu hơn được đề xuất để nắm bắt cả bối cảnh quá khứ và tương lai, làm phong phú thêm quá trình giải mã. Các biến thể LSTM hai giai đoạn và ba luồng khám phá thêm bối cảnh trực quan và thông tin ngữ nghĩa, dẫn đến chú thích mang tính mô tả và đa dạng hơn.

- **Giải quyết được thách thức gì trong bài toán image captioning**:
  - Bài toán chú thích ảnh đòi hỏi kết hợp thông tin từ mô hình ngôn ngữ và mô hình hình ảnh. Phương pháp Encoder-Decoder giúp kết hợp hai loại thông tin này

bằng cách sử dụng mô hình hình ảnh (encoder) để trích xuất đặc trưng từ hình ảnh và mô hình ngôn ngữ (decoder) để tạo ra mô tả văn bản dựa trên đặc trưng hình ảnh đã trích xuất.
  - Một số phương pháp Encoder-Decoder cho phép kiểm soát chú thích ảnh bằng cách thêm các tín hiệu điều khiển. Điều này giúp người dùng có khả năng tùy chỉnh mô tả ảnh theo ý muốn, ví dụ như chỉ định mức độ chi tiết hoặc thay đổi phong cách văn bản.
  - Phương pháp này cho phép huấn luyện mô hình từ dữ liệu gán nhãn tự động và tự động tạo ra mô tả ảnh mà không cần bước trung gian.
- **Nhược điểm còn tồn đọng khi giải bài toán image captioning**:
  - Mặc dù phương pháp Encoder-Decoder đã đạt được nhiều tiến bộ, nhưng việc hiểu biết chính xác về nội dung hình ảnh vẫn còn hạn chế. Mô hình có thể không hiểu được các chi tiết phức tạp hoặc ngữ cảnh ẩn sau hình ảnh.
  - Hiệu suất của phương pháp phụ thuộc vào khả năng trích xuất đặc trưng hình ảnh. Nếu mô hình không thể trích xuất thông tin quan trọng từ hình ảnh, mô tả văn bản sẽ bị ảnh hưởng.

2) Attention Mechanism:

- **Nguyên lý**: Attention Mechanism đặt chú thích cho hình ảnh bằng cách tập trung vào các vùng cụ thể có liên quan đến từng phần của câu mô tả hay nói cách khác, nó tập trung vào các chi tiết quan trọng trong ảnh để tạo ra câu mô tả. Cơ chế này có thể kết hợp với các phương pháp khác nhau trong đó có Encoder – Decode.
- **Phương pháp**:
  - **Cross-Modal Attention**:
    - Cơ chế chú ý qua các modal khác nhau trong tạo chú thích hình ảnh nhằm tạo ra sự phù hợp giữa thông tin hình ảnh và ngôn ngữ ở các cấp độ khác nhau. Cụ thể, nó giúp mô hình tập trung vào các đặc điểm quan trọng của hình ảnh và các khía cạnh ngữ nghĩa của ngôn ngữ, giúp tạo ra các chú thích chính xác và mô tả đầy đủ cho hình ảnh.
    - Các loại cơ chế chú ý khác nhau được sử dụng để cải thiện sự tương quan giữa thông tin hình ảnh và ngôn ngữ:
      - **Global-local attention** (chú ý toàn cầu-địa phương): Kết hợp thông tin cục bộ và toàn cục từ hình ảnh để tạo ra chú thích chi tiết và đồng thời phù hợp với bối cảnh toàn cảnh.
      - **Semantic attention** (chú ý ngữ nghĩa): Tích hợp ngữ nghĩa vào việc tạo chú thích, giúp mô hình tập trung vào các khía cạnh quan trọng và đa dạng của hình ảnh.
      - **Spatial and channel-wise attention** (chú ý theo không gian và theo kênh): Tùy chỉnh sự chú ý dựa trên ngữ cảnh của câu, giúp chú trọng vào các thuộc tính ngữ nghĩa quan trọng nhất.
      - **Adaptive attention** (chú ý thích ứng): Quyết định khi nào nên sử dụng thông tin hình ảnh và khi nào chỉ nên sử dụng mô hình ngôn ngữ, tùy thuộc vào nhu cầu của câu.
      - **Context attention** (chú ý bối cảnh): Tập trung vào các phần khác nhau của hình ảnh dựa trên ngữ cảnh của câu và trạng thái hiện tại.
  - **Intra-Modal Attention**:
    - Cơ chế chú ý trong cùng một modal tập trung vào việc tạo ra sự tương tác giữa các phần của dữ liệu trong cùng một loại modal, ví dụ như giữa các phần của hình ảnh hoặc các từ trong văn bản, tạo ra sự tương tác giữa các phần của hình ảnh và từ trong câu chú thích.
    - Cụ thể, cơ chế này thường dựa trên cơ chế tự chú ý (self-attention) được giới thiệu trong mô hình Transformer. Trong mô hình Transformer, cơ chế tự chú ý cho phép mô hình tập trung vào các phần quan trọng của dữ liệu đầu vào trong cùng một chuỗi, giúp nắm bắt các mối quan hệ dài hạn và tương tác giữa các phần khác nhau của dữ liệu.
    - Quá trình phát triển của học tương tác trong cùng một modal có thể chia thành ba giai đoạn chính:
      - **Kết hợp Transformer với CNN-LSTM**: Ở giai đoạn đầu tiên, các nghiên cứu đã kết hợp cấu trúc Transformer với mô hình CNN-LSTM truyền thống. Một số nghiên cứu đã sử dụng mô hình Transformer để thay thế phần giải mã LSTM, giúp giải quyết vấn đề của chuỗi liên thời gian trong mô hình LSTM.
      - **Sử dụng toàn bộ cấu trúc Transformer với các đặc điểm phát hiện đối tượng**: Các nghiên cứu tiếp theo đã sử dụng cấu trúc Transformer hoàn chỉnh với các đặc điểm phát hiện đối tượng từ hình ảnh để hướng dẫn quá trình tạo chú thích. Các biến thể của Transformer như Object relation Transformer (ORT) và Geometry-aware self-attention (G-SAN) đã được giới thiệu để mô hình hóa các mối quan hệ hình học giữa các đối tượng trong hình ảnh.
      - **Tiếp cận không sử dụng tích chập**: Một số nghiên cứu đã tiếp cận không sử dụng tích chập, thay vào đó sử dụng mô hình Transformer từ đầu để xử lý hình ảnh. Điều này giúp mô hình có thể mô hình hóa ngữ cảnh toàn cầu từ đầu đến cuối mà không cần phải sử dụng tích chập hoặc tuần tự.
- **Giải quyết được thách thức gì trong bài toán image captioning**: Attention mechanism cho phép mô hình chú ý đến các phần của hình ảnh có ảnh hưởng lớn nhất đến việc tạo ra mô tả, giúp mô hình tập trung vào các đặc trưng quan trọng của hình ảnh, thay vì chỉ sử dụng toàn bộ đặc trưng hình ảnh một cách đồng nhất. Giúp tạo ra các mô tả một cách tự nhiên và phong phú hơn. Điều này giúp cải thiện đáng kể chất lượng của các mô hình image

captioning, tăng khả năng chính xác và đa dạng của các mô tả được tạo ra.

- **Nhược điểm còn tồn đọng khi giải bài toán image captioning**: Attention Mechanism còn tồn tại một số nhược điểm như tính phức tạp của mô hình, khả năng tính toán cao, khả năng tổng quát hóa không tốt và khó khăn trong quá trình huấn luyện. Mặc dù đã đạt được sự tiến bộ đáng kể, nhưng việc cải thiện và áp dụng cơ chế chú ý vẫn đang là một hướng nghiên cứu tiềm năng trong tương lai để tạo ra các mô hình image captioning hiệu quả hơn.

3) Training Strategies:

- **Cross-entropy loss**:
  - **Nguyên lý**: tối ưu hóa mô hình bằng cách tính toán sự khác biệt giữa phân phối xác suất của các từ mục tiêu và các từ được dự đoán để định hình mô hình.
  - **Phương pháp**: Sử dụng mất mát cross-entropy để tính toán sự chênh lệch giữa phân phối xác suất của các từ mục tiêu và các từ dự đoán. Được sử dụng trong các mô hình như NIC, Show, Attend and Tell, Semantic Attention, SCA-CNN và Adaptive Attention. Mô hình được huấn luyện thông qua việc tối ưu hóa hàm mất mát cross-entropy loss thông qua việc điều chỉnh các tham số.
  - **Giải quyết được thách thức gì trong bài toán image captioning**: giúp giải quyết vấn đề về tính chính xác cú pháp của các câu mô tả hình ảnh. Nó giúp mô hình học được phân phối xác suất của các từ trong câu mô tả thực tế, tạo ra các mô tả ngữ nghĩa và cú pháp hợp lý cho hình ảnh.
  - **Nhược điểm còn tồn đọng khi giải bài toán image captioning**: phương pháp này có thể bỏ qua các phụ thuộc xa hơn giữa các từ được tạo ra, dẫn đến việc các mô tả có thể thiếu chi tiết và tự nhiên. Ngoài ra, bias trong việc tiếp cận dữ liệu có thể làm cho mô hình tạo ra các mô tả an toàn nhưng ít chi tiết.
- **Reinforcement learning**:
  - **Nguyên lý**: được thiết kế để vượt qua các hạn chế của việc huấn luyện ở mức từ bằng cách sử dụng huấn luyện ở mức chuỗi. Phương pháp này tập trung vào việc tối ưu hóa các chuỗi các hành động dựa trên các chỉ số đánh giá như BLEU, CIDEr, ROUGE, SPICE, và các phương pháp đánh giá khác.
  - **Phương pháp**: Tận dụng tìm kiếm bằng đám và giải mã theo cách tham lam để tính toán độ dốc của mất mát.Các phần thưởng thường dựa trên các chỉ số đánh giá như BLEU, CIDEr, ROUGE, SPICE,...Các công trình khác nhau khám phá các chiến lược RL sử dụng các chỉ số mức chuỗi khác nhau làm phần thưởng.
  - **Giải quyết được thách thức gì trong bài toán image captioning**: giúp tạo ra các mô tả tự nhiên và phong phú hơn bằng cách tập trung vào việc tối ưu hóa các chuỗi các hành động dựa trên phần thưởng từ các chỉ số đánh giá.
  - **Nhược điểm còn tồn đọng khi giải bài toán image captioning**: phương pháp này có thể đòi hỏi việc tiền huấn luyện kỹ lưỡng và có thể phức tạp hơn trong việc điều chỉnh các tham số của mô hình.
- **Pre-training model**:
  - **Nguyên lý**: sử dụng các mô hình đã được huấn luyện trước trên các dữ liệu lớn để cải thiện hiệu suất của mô hình cụ thể.
  - **Phương pháp**: Tập trung vào việc cân bằng văn bản-hình ảnh và mất mát thông tin từ ngữ ngữ cảnh được che dấu.
  - **Giải quyết được thách thức gì trong bài toán image captioning**: Phương pháp này giúp cải thiện hiệu suất của mô hình và giảm thiểu thời gian huấn luyện bằng cách sử dụng các mô hình đã được huấn luyện trước trên dữ liệu lớn.
  - **Nhược điểm còn tồn đọng khi giải bài toán image captioning**: việc sử dụng các mô hình đã được tiền huấn luyện có thể đòi hỏi lượng dữ liệu lớn và có thể gặp phải vấn đề về đa dạng và khối lượng của dữ liệu.

III. Phương pháp

Trong đồ án này, chúng tôi đề xuất mô hình CNN-LSTM để giải quyết vấn đề image captioning, sử dụng cơ chế Encoder-Decoder. Trong cấu trúc này, một mạng CNN (Convolutional Neural Network) được sử dụng như một phần mã hóa để trích xuất các đặc trưng từ hình ảnh, trong khi một mạng LSTM (Long Short-Term Memory) được sử dụng như một phần giải mã để tạo ra chuỗi mô tả hình ảnh.

- CNN Encoder: Được huấn luyện trước sử dụng mô hình ResNet-152 trên tập dữ liệu phân loại hình ảnh ILSVRC-2012-CLS. Mục đích của CNN Encoder là trích xuất các đặc trưng từ hình ảnh, tạo ra một biểu diễn trung gian của thông tin trong hình ảnh.
- LSTM Decoder: Nhận vào cả vector đặc trưng từ hình ảnh và chuỗi từ nguồn, sau đó được huấn luyện để dự đoán chuỗi từ đích. Trong quá trình kiểm tra, LSTM được sử dụng trong một vòng lặp để sinh ra từng từ tiếp theo dựa trên từ đã được sinh ra trước đó.

Mô hình được thử nghiệm và đánh giá trên bộ dữ liệu khác Flickr8k. Mô hình có thể được đánh giá dựa trên các thông số: BLEU.

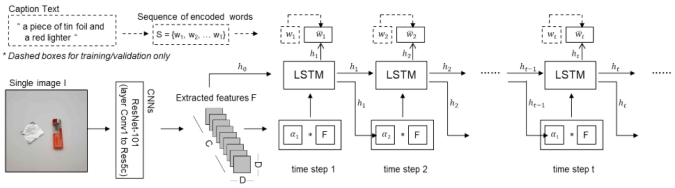
Để cải thiện khả năng tập trung vào các phần cụ thể của hình ảnh khi tạo ra chú thích, chúng tôi cũng đề xuất kết hợp mô hình này với cơ chế chú ý (attention mechanism), giúp mô hình tập trung vào các phần quan trọng của hình ảnh trong quá trình sinh ra chuỗi mô tả.

A. Kiến trúc mô hình CNN-LSTM

Mô hình CNN-LSTM là một mạng nơ-ron sâu chứa hai mạng nơ-ron thành phần:

- Mạng nơ-ron tích chập (CNN) dùng để rút trích véc-tơ đặc trưng ảnh.

- Mạng bộ nhớ dài-ngắn (LSTM) dùng để phát sinh ra các từ của câu mô tả từ véc-tơ đặc trưng ảnh của CNN và từ các từ đã phát sinh ra.



Hình 1. Minh họa kiến trúc mô hình CNN-LSTM.

Ảnh đầu vào I đưa vào CNN để rút trích đặc trưng F, sau đó đặc trưng này được dùng để khởi tạo các trạng thái ẩn h_{-1} và trạng thái lưu giữ c_{-1}. Tiếp theo x_0, h_{-1} và c_{-1} được đưa vào tầng ẩn của LSTM tại bước đầu tiên $(t = 0)$. Tại các bước sau, LSTM dựa trên trạng thái ẩn, trạng thái lưu giữ trước đó h_{t-1}, c_{t-1} và "distributed feature vector" α_t được đưa vào LSTM để phát sinh ra câu mô tả tương ứng.

Mô hình nhận đầu vào là ảnh I và phát sinh câu mô tả $\hat{S} = \{\hat{w_1}, \hat{w_2}, ..., \hat{w_T}\}, \hat{w_i} \in \mathbb{R}^{|V|}$. Với $\hat{w_i}$ là các từ được biểu diễn dưới dạng "one-hot vector". V là bộ từ vựng và $|V|$ là kích thước bộ từ vựng.

Ta sử dụng mô hình CNN để trích đặc trưng của ảnh đầu vào I; sau đó đặc trưng ảnh I này được đưa vào LSTM để phát sinh ra câu mô tả tương ứng. Chi tiết cách đưa đặc trưng ảnh này vào LSTM để phát sinh câu mô tả được thể hiện trong hình 1 và các bước thực hiện như sau:

- Đầu tiên, mô hình lan truyền ảnh đầu vào I qua các tầng của CNN; sau đó, ta thu được đặc trưng F ở tầng cuối cùng của CNN.

$$F = CNN(I)(3.1)$$

- Ở bước $t = -1$, trạng thái ẩn h_{-1} và trạng thái lưu giữ c_{-1} được tính từ đặc trưng F bằng hai hàm tuyến tính có ma trận tham số lần lượt là K_h và K_c.

$$h_{-1} = K_h^T F$$

$$c_{-1} = K_c^T F(3.2)$$

- Ở bước đầu tiên $(t = 0)$, từ đầu tiên w_0 (kí hiệu <START>) được chuyển về "distributed feature vector" α_0 bằng ma trận chiếu P. Sau đó, x_t cùng với trạng thái ẩn h_{-1} và trạng thái lưu giữ c_{-1} được đưa vào tầng ẩn đầu tiên của LSTM để tính véc-tơ dự đoán xác suất cho từ kế tiếp $\hat{y}_0$.

$$\alpha_0 = P^T w_0(3.3)$$

$$\hat{y}_0 = LSTMcell(\alpha_0, h_{-1}, c_{-1})(3.4)$$

- Từ véc-tơ $\hat{y}_0$ này ta lấy vị trí i có giá trị xác suất lớn nhất trong $\hat{y}_0$, nghĩa là $\hat{y}_{0,i} = max_{1 \le j \le |V|} \hat{y}_{0,j}$, từ được dự đoán ở bước này $\hat{w}_1$ là từ thứ i trong tập từ vựng V. $\hat{w}_1$ dùng để làm đầu vào tiếp theo cho bước $t = 1$.
- Tương tự, ở các bước tiếp theo, ta tiếp tục đưa vào tầng ẩn của LSTM từ được phát sinh trước đó $\hat{w}_t$ (bằng cách chuyển về "distributed feature vector" α_t như ở công thức (3.3)), trạng thái ẩn h_{t-1} và trạng thái lưu giữ c_{t-1}. Quá trình này được lặp lại qua các tầng còn lại của LSTM để tính véc-tơ dự đoán xác suất cho từ kế tiếp $\hat{y}_t$. Dựa trên $\hat{y}_t$ ta tính được từ dự đoán kế tiếp là $\hat{w}_t$.

$$\hat{y}_t = LSTMcell(\alpha_t, h_{t-1}, c_{t-1}), t \in \{1, ..., T\}(3.5)$$

- Cuối cùng, ta thu được câu mô tả dự đoán tương ứng $\hat{S} = \{\hat{w_1}, \hat{w_2}, ..., \hat{w_T}\}$.

Bộ tham số θ cần học của mô hình:

$$\theta = \{\theta_{CNN}, P, K_h, K_c, \theta_{LSTM}\}$$

Với θ_{CNN} là bộ tham số của CNN và θ_{LSTM} là bộ tham số của LSTM.

B. Huấn luyện mô hình CNN-LSTM

Mô hình được huấn luyện dựa trên tập dữ liệu gồm các cặp (input, output), trong đó mỗi cặp gồm input là một ảnh I và output là câu mô tả đúng S tương ứng của ảnh.

Đặc trưng ảnh được trích xuất từ ảnh I thông qua mạng CNN, sau đó đưa vào mạng LSTM để dự đoán mô tả $\hat{S}$ cho ảnh I.

Mục tiêu huấn luyện là tìm ra bộ tham số θ tối ưu để cực đại hóa xác suất $p(S|I;\theta)$ phát sinh mô tả cho ảnh:

$$\theta^* = \text{argmax}_\theta \, p(S|I;\theta)$$

Trong quá trình huấn luyện ta sử dụng một hàm mất mát để đánh giá bộ tham số θ của mô hình. Ta cần tìm bộ tham số θ để cực tiểu hóa hàm mất mát cross-entropy là:

$$L(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T} \sum_{j=1}^{|V|} y_{t,j} \log(\hat{y_{t,j}})$$

Trong đó:

- N: số cặp dữ liệu huấn luyện.
- T: số lượng từ trong mỗi câu mô tả.
- |V|: kích thước của tập từ vựng, tức là số lượng từ khác nhau có thể xuất hiện trong mô tả.
- $y_{t,j}$:phần tử thứ j của vector nhãn tại bước thứ t trong chuỗi được biểu dưới dạng "one-hot vector".
- $\hat{y_{t,j}}$: phần tử thứ j của vector dự đoán phân bố xác suất của từ tại bước thứ t.

Để cực tiểu hóa hàm chi phí $L(\theta)$ ta có thể sử dụng thuật toán "Mini-batch Gradient Descent" có thể giúp quá trình hội tụ nhanh hơn. Do phương pháp này hoạt động bằng cách cập nhật các tham số của mô hình bằng cách sử dụng gradient của hàm mất mát tính trên một số lượng nhỏ các mẫu dữ liệu, gọi là "mini-batch", thay vì sử dụng toàn bộ dữ liệu (batch gradient descent) hoặc một mẫu dữ liệu duy nhất (stochastic gradient descent).Trong "Mini-batch Gradient Descent", mỗi lần cập nhật, thuật toán này lấy ra một "mini-batch" từ dữ liệu huấn luyện để tính toán gradient của hàm mất mát. Sau đó, nó cập nhật các tham số của mô hình bằng cách di chuyển theo hướng ngược của gradient với một khoảng nhất định, được

điều chỉnh bởi một hệ số học (η). Công thức cập nhật trong "Mini-batch Gradient Descent" có thể được biểu diễn như sau:

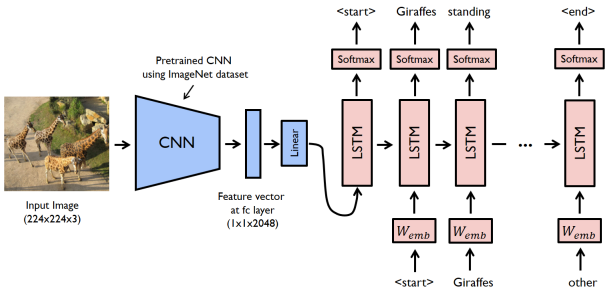
$$\theta = \theta - \eta \nabla_\theta L(\theta; I_{(i:i+n)}; S_{(i:i+n)})$$

Trong đó:

- η: là hệ số học (learning rate), quyết định tốc độ học của thuật toán.
- ∇_θ: gradient của hàm mất mát theo các tham số θ
- $I_{(i:i+n)}; S_{(i:i+n)}$: tập hợp các ảnh và các câu mô tả tương ứng từ chỉ số i đến $i+n-1$.

C. Phát sinh câu mô tả từ ảnh với mô hình CNN-LSTM đã được huấn luyện.

Sau khi hoàn thành việc huấn luyện, ta có bộ tham số θ và có thể sử dụng mô hình để phát sinh câu mô tả của một ảnh I bất kỳ. Quá trình phát sinh câu mô tả từ mô hình CNN-LSTM đã được huấn luyện được trình bày ở phần A và hình minh họa 2.



Hình 2. Minh họa quá trình mô hình CNN-LSTM phát sinh câu mô tả ảnh. Ta đưa ảnh đầu vào I qua CNN để trích xuất đặc trưng ảnh, sau đó đưa đặc trưng ảnh này vào LSTM để phát sinh ra câu mô tả nội dung ảnh

Tuy nhiên, việc chọn từ có xác suất dự đoán cao nhất cho bước tiếp theo không đảm bảo rằng chuỗi phát sinh $\hat{S}$ có thể cực đại hóa $p(\hat{S}|I;\theta)$. Để giải quyết vấn đề này, ta có thể duyệt tất cả các trường hợp để chọn câu dự đoán tốt nhất, nhưng cách này không khả thi khi từ điển có kích thước lớn.

Để giảm tính toán và chọn được câu dự đoán tốt, ta sử dụng phương pháp tìm kiếm "beam" (Beam Search). Tại mỗi bước, thay vì chọn tất cả các từ, ta chọn *k* từ (kích thước của "beam") sao cho cực đại hóa:

$$\prod_{t=1}^{tc} p(\hat{w}_t | I, \hat{w}_{t-1}, ..., \hat{w}_1, w_0; \theta)$$

với *tc* là bước hiện tại. Điều này có nghĩa là tại mỗi bước, ta mong muốn phát sinh từ sao cho chuỗi các từ tốt nhất, thay vì chỉ tìm từ tốt nhất ở bước hiện tại. Khi trong *k* từ được chọn có ký hiệu kết thúc câu (<END>), ta đã phát sinh được một câu ứng viên và tiếp tục tìm kiếm *k-1* câu ứng viên còn lại. Quy trình này lặp lại cho đến khi phát sinh được *k* câu ứng viên và cuối cùng, ta chọn câu ứng viên có $p(\hat{S}|I)$ lớn nhất.

IV. Mô hình CNN-LSTM-Attention

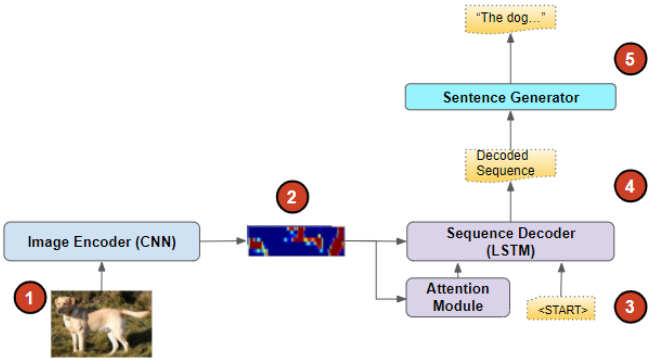
A. Vấn đề của mô hình CNN-LSTM và khắc phục bằng cơ chế Attention

Trong mô hình CNN-LSTM, sau khi CNN trích xuất đặc trưng từ ảnh, thông tin này được truyền vào LSTM để tạo câu mô tả. Tuy nhiên, thông tin về ảnh có thể mất dần qua các bước của LSTM, vì nó chỉ được xem một lần ở bước đầu. Để khắc phục vấn đề này, ta có thể cung cấp thông tin về ảnh vào từng bước của LSTM để mô hình "nhìn" vào ảnh nhiều lần, giúp cải thiện việc ghi nhớ thông tin. Tuy nhiên, đối với các ảnh phức tạp, việc chọn thông tin phù hợp cho từng bước có thể là thách thức. Do đó, thay vì đưa vào toàn bộ ảnh, ta có thể chỉ chọn các vùng ảnh quan trọng để cung cấp thông tin cho mỗi bước, giúp mô hình tạo câu mô tả hiệu quả hơn.

Đây chính là ý tưởng trong nghiên cứu "Show, attend and tell: Neural image caption generation with visual attention" của Kelvin Xu và đồng nghiệp [8], nghiên cứu đã đề xuất một mô hình học sâu tương tự như mô hình đã được đề cập,và đã kết hợp thêm cơ chế Attention. Cơ chế này cho phép mô hình tập trung vào các vùng quan trọng của ảnh ở mỗi bước trong quá trình sinh câu mô tả, thay vì xem toàn bộ ảnh một lần. Tại mỗi bước, thông tin từ ảnh và từ các từ trước đó trong LSTM được cung cấp cho cơ chế Attention, từ đó mô hình chọn ra thông tin cần thiết từ ảnh để sinh từ tiếp theo. Với cơ chế này, mô hình đã đạt được kết quả tốt nhất tại thời điểm nghiên cứu trên các tập dữ liệu thử nghiệm. Ngoài việc cải thiện độ chính xác, cơ chế Attention còn giúp hiểu rõ hơn về quá trình học của mô hình bằng cách quan sát các vùng quan trọng mà mô hình chú ý trong quá trình sinh câu mô tả.

B. Kiến trúc mô hình CNN-LSTM- Attention

Mô hình CNN-LSTM-Attention bổ sung cơ chế Attention vào kiến trúc cơ bản của CNN-LSTM, cho phép mô hình tập trung vào các vùng quan trọng của ảnh trong quá trình sinh câu mô tả. Kiến trúc mô hình bao gồm ba phần chính:



Hình 3. Minh họa kiến trúc mô hình CNN-LSTM-Attention.

- Mạng nơ-ron tích chập (CNN): : Đây là phần đầu tiên của mô hình, được sử dụng để xử lý và trích xuất đặc trưng từ ảnh. CNN làm việc bằng cách áp dụng một loạt các bộ lọc (kernels) để trích xuất các đặc trưng của ảnh. Kết quả là một tập hợp các véc-tơ đặc trưng, mỗi véc-tơ đại

diện cho một vùng của ảnh. Các véc-tơ này chứa thông tin về các đặc điểm quan trọng trong ảnh như cạnh, góc, hoặc đặc điểm địa hình.

- Mạng bộ nhớ dài-ngắn (LSTM): Phần thứ hai của mô hình, LSTM, được sử dụng để tạo ra câu mô tả từ các đặc trưng đã được trích xuất từ CNN. LSTM là một loại mạng nơ-ron đặc biệt có khả năng nhớ và quản lý thông tin trong thời gian dài. Nó nhận các đặc trưng từ ảnh và từ các từ đã được sinh ra trước đó để dự đoán từ tiếp theo trong câu mô tả. LSTM cung cấp khả năng xử lý chuỗi dữ liệu, làm cho nó trở thành lựa chọn phổ biến cho các tác vụ như sinh mô tả ảnh.
- Cơ chế Attention: Đây là phần mới được thêm vào mô hình để cải thiện việc sinh câu mô tả. Cơ chế này cho phép mô hình tập trung vào các vùng quan trọng của ảnh trong quá trình sinh câu mô tả, thay vì xem toàn bộ ảnh một lần. Cụ thể, ở mỗi bước trong quá trình sinh câu, cơ chế Attention tập trung vào các phần của ảnh có liên quan đến từ đang được dự đoán. Điều này giúp mô hình tạo ra câu mô tả chính xác hơn và cải thiện hiểu biết về nội dung của ảnh.

C. Huấn luyện mô hình CNN-LSTM-Attention

Mô hình được huấn luyện trên tập dữ liệu gồm N mẫu, mỗi mẫu bao gồm một cặp ảnh và câu mô tả tương ứng. Độ lỗi của mỗi mẫu huấn luyện được tính bằng hàm mất mát, bao gồm hai phần: đo lường sự sai khác giữa từ nhãn và từ dự đoán, cùng với một phần kiểm soát dựa trên trọng số của cơ chế Attention.

Công thức độ lỗi cho mỗi mẫu huấn luyện:

$$L_{\text{total}} = \frac{1}{N} \sum_{i=1}^{N} \left(L_{\text{cross-entropy}}^{(i)} + \lambda L_{\text{attention}}^{(i)} \right)$$

Trong đó:

- L_{total}: Độ lỗi tổng của toàn bộ tập dữ liệu huấn luyện.
- N: Số lượng mẫu trong tập dữ liệu huấn luyện.
- $L_{\text{cross-entropy}}^{(i)}$: Độ lỗi cross-entropy của mẫu thứ i.
- λ: Tham số kiểm soát để điều chỉnh sự quan trọng giữa hai phần tử trong hàm mất mát.
- $L_{\text{attention}}^{(i)}$: Độ lỗi kiểm soát dựa trên trọng số của cơ chế Attention của mẫu thứ i.

Hàm mất mát $L_{\text{cross-entropy}}^{(i)}$ được tính bằng công thức:

$$L_{\text{cross-entropy}}^{(i)} = -\frac{1}{T^{(i)}} \sum_{t=1}^{T^{(i)}} \sum_{j=1}^{|V|} y_{t,j}^{(i)} \log(\hat{y}_{t,j}^{(i)})$$

Trong đó $T^{(i)}$ là số lượng từ trong câu mô tả tương ứng với mẫu thứ i, $|V|$ là kích thước của tập từ vựng, $y_{t,j}^{(i)}$ là phần tử thứ j của vector nhãn tại bước thứ t trong chuỗi câu mô tả của mẫu thứ i, $\hat{y}_{t,j}^{(i)}$ là phần tử thứ j của vector dự đoán phân bố xác suất của từ tại bước thứ t trong câu mô tả của mẫu thứ i.

D. Phát sinh câu mô tả từ ảnh với mô hình CNN-LSTM-Attention đã được huấn luyện:

Sau khi mô hình CNN-LSTM-Attention đã được huấn luyện, ta sử dụng nó để tạo ra câu mô tả cho một ảnh bất kỳ. Quá trình này bao gồm các bước sau:

Đầu tiên, ảnh đầu vào được đưa qua mạng CNN để trích xuất ra các đặc trưng của ảnh. Mạng CNN đã được huấn luyện trước đó để nhận diện các đặc trưng quan trọng của ảnh.

Sau khi có các đặc trưng của ảnh, một trạng thái ẩn ban đầu cho mạng LSTM được khởi tạo. Trạng thái ẩn này sẽ giúp mô hình bắt đầu sinh ra các từ đầu tiên trong câu mô tả.

Sử dụng cơ chế Attention để tạo câu mô tả: Quá trình sinh câu mô tả sử dụng cơ chế Attention, cho phép mô hình tập trung vào các phần quan trọng của ảnh trong quá trình tạo câu mô tả. Cơ chế này giúp cải thiện khả năng tạo ra các câu mô tả chính xác và tự nhiên.

Tại mỗi bước thời gian, mô hình LSTM dự đoán từ tiếp theo trong câu mô tả dựa trên đặc trưng của ảnh và từ đã được dự đoán trước đó. Quá trình này diễn ra cho đến khi gặp từ kết thúc câu mô tả hoặc đạt đến giới hạn số từ cho mỗi câu.

Lặp lại quá trình cho đến khi hoàn thành câu mô tả: Quá trình dự đoán từ tiếp theo và sử dụng cơ chế Attention được lặp lại cho đến khi câu mô tả hoàn chỉnh. Mỗi từ được dự đoán dựa trên thông tin từ ảnh và từ đã được dự đoán trước đó.

Cuối cùng, sau khi đã có một số câu mô tả khác nhau, ta chọn câu mô tả có xác suất cao nhất được dự đoán bởi mô hình là kết quả cuối cùng. Điều này giúp đảm bảo rằng câu mô tả được tạo ra là tự nhiên và phù hợp nhất với nội dung của ảnh.

Quá trình này giúp mô hình CNN-LSTM-Attention tạo ra các câu mô tả chất lượng cao cho các ảnh đầu vào, đồng thời tận dụng cơ chế Attention để cải thiện khả năng tập trung vào các đặc trưng quan trọng của ảnh trong quá trình tạo câu mô tả.

V. Thực nghiệm

A. Mục tiêu thực nghiệm

Mục tiêu của thực nghiệm về chú thích hình ảnh (image captioning) có thể là:

Đánh giá hiệu suất bằng chỉ số BLEU và METEOR: nhằm đo lường chất lượng của các mô tả được tạo ra của mô hình bằng các chỉ số BLEU và METEOR. BLEU đánh giá độ trùng lặp ngữ cảnh giữa mô tả được tạo ra và mô tả thực tế, trong khi METEOR đo lường sự tương đồng giữa hai chuỗi văn bản.

Đánh giá tác động của yếu tố khác nhau: xác định tác động của các yếu tố khác nhau đối với chất lượng của các mô tả được tạo ra. Các yếu tố này có thể bao gồm kiến trúc mạng CNN, cơ chế attention, kích thước bộ dữ liệu huấn luyện, và cách tiền xử lý dữ liệu.

So sánh phương pháp: so sánh hiệu suất của hai mô hình CNN-LSTM và CNN-LSTM-Attention trong việc tạo ra các mô tả hình ảnh. Thông qua các thực nghiệm, ta có thể so sánh và đánh giá hai phương pháp này để xác định mô hình nào tốt hơn hoặc phù hợp hơn cho bài toán image captioning.

Cải thiện hiệu suất: nhằm cải thiện hiệu suất của mô hình image captioning thông qua việc thực hiện các thực nghiệm và nghiên cứu các phương pháp, kỹ thuật mới. Kết quả thực nghiệm có thể được sử dụng để cải thiện mô hình để tạo ra các mô tả chất lượng cao hơn cho hình ảnh.

Kiểm tra tính khả thi: nhằm xác định tính khả thi của mô hình trong việc tạo ra các mô tả hình ảnh trong các tình huống thực tế. Thực nghiệm có thể đưa ra đánh giá về độ chính xác, sự ổn định và khả năng áp dụng của các mô hình trong các bối cảnh thực tế khác nhau.

B. Môi trường thực nghiệm

Thực nghiệm trên môi trường Kaggle và ngôn ngữ lập trình Python.

C. Các thiết kế thực nghiệm

Để cài đặt mô hình, nhóm sử dụng GPU P100 của Nvidia có bộ nhớ trong là 16GB để tính toán song song và tăng tốc quá trình huấn luyện mô hình. Ngoài việc cài đặt mô hình CNN-LSTM, nhóm cũng tiến hành cài đặt mô hình CNN-LSTM-Attention để thử cải tiến mô hình CNN-LSTM, thể hiện so sánh kết quả giữa hai mô hình xem liệu mô hình CNN-LSTM-Attention có tốt hơn so với mô hình CNN-LSTM.

Với kích thước ảnh là 224×224 sau khi được đưa qua mạng "ResNet-50" thu được 512 feature maps kích thước 14×14. Từ các feature maps, ta có tập a gồm 196 véc-tơ đặc trưng ứng với các vùng ảnh khác nhau, mỗi véc-tơ có kích thước là 196×512. Nhóm thiết lập kích thước của "word embedding" là 512 chiều.

Trong quá trình huấn luyện, nhóm sử dụng thuật toán "Adam" để tối ưu hóa. Chúng tôi chọn kích thước "mini-batch" là 32. Sử dụng kỹ thuật học chuyển tiếp với bộ trọng số của mạng ResNet-50 đã được huấn luyện trước. Huấn luyện mô hình LSTM với hệ số học 4×10^{-4}.

Trong quá trình kiểm tra, nhóm sử dụng hai độ đo đánh giá tự động đó là: BLEU và METEOR. Độ đo BLEU là độ đo được sử dụng phổ biến trong bài toán dịch máy, nhưng độ đo này vẫn còn tồn tại một số hạn chế. Do đó, nhóm sử dụng thêm độ đo METEOR. Ý tưởng chung của các độ đo này đó là đánh giá câu phát sinh bởi mô hình bằng mức độ so khớp với các câu đúng được viết bởi con người. Độ đo BLEU và METEOR có miền giá trị $[0, 1]$; với giá trị của độ đo càng cao thì kết quả câu mô tả càng tốt.

D. Dữ liệu thực nghiệm

Sử dụng bộ dữ liệu Flickr8k trong thực nghiệm vì nó cung cấp một bộ sưu tập đa dạng và toàn diện gồm khoảng 8000 hình ảnh, mỗi hình ảnh kèm theo chú thích chi tiết. Tập dữ liệu bao gồm hình ảnh của các cảnh quan, vật thể và chủ đề, vì thế nó rất phù hợp để giúp đánh giá hiệu suất của các mô hình chú thích hình ảnh trên nhiều tình huống trong thế giới thực. Bằng cách kết hợp Flickr8k vào thực nghiệm, nhóm nhắm đến việc tận dụng sự phong phú và đa dạng của nó để đánh giá kỹ lưỡng khả năng khái quát hóa của các kỹ thuật chú thích hình ảnh được sử dụng. Sự sẵn có của hình ảnh chất lượng cao với chú thích tương ứng trong tập dữ liệu đã tạo điều kiện thuận lợi cho các giai đoạn đào tạo, đánh giá và kiểm thử mô hình, đưa ra một đánh giá mạnh mẽ về hiệu suất của các mô hình trên một phạm vi rộng của nội dung hình ảnh. Nhìn chung, sự lựa chọn bộ dữ liệu Flickr8k phù hợp với mục tiêu đánh giá toàn diện và thực tế về mô hình chú thích hình ảnh được cài đặt đại diện trong thực nghiệm trong khi vẫn có hiệu quả tính toán cho phạm vi của dự án so với các bộ dữ liệu khác trong thực tế.

E. Kết quả thực nghiệm

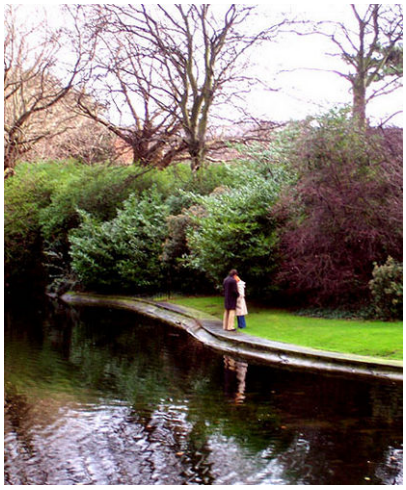
1) Kết quả so sánh hai mô hình dựa trên điểm đánh giá BLUE-4 và METEOR: Mỗi kết quả được tạo ra dựa trên 4 mẫu câu chú thích tham chiếu so với 1 chú thích đầu ra được tạo bởi mô hình đã huấn luyện. 5 hình ảnh 4 5 6 7 8 ví dụ thử nghiệm đã được sử dụng để đánh giá cả hai mô hình được chọn như dưới đây:

Hình 4. Ảnh ví dụ 1.

Hình 5. Ảnh ví dụ 2.

Hình 6. Ảnh ví dụ 3.

Hình 7. Ảnh ví dụ 4.



Hình 8. Ảnh ví dụ 5.

Kết quả đánh giá hai mô hình dựa trên độ đo BLUE-4 và METEOR được thể ở bảng I và II.

Bảng I
ĐIỂM BLUE-4

	Ví dụ 1	***Ví dụ 2***	***Ví dụ 3***	***Ví dụ 4***	***Ví dụ 5***
Không dùng Attention	0.037	0.086	0.092	0.086	0.021
Dùng Attention	0.056	0.132	0.089	0.499	0.023

Bảng II
ĐIỂM METEOR

	Ví dụ 1	***Ví dụ 2***	***Ví dụ 3***	***Ví dụ 4***	***Ví dụ 5***
Không dùng Attention	0.168	0.422	0.463	0.216	0.187
Dùng Attention	0.221	0.603	0.326	0.763	0.106

Dựa trên kết quả đánh giá của độ đo BLUE-4 và METEOR sau khi thực nghiệm, có thể thấy mô hình CNN-LSTM-Attention có kết quả tốt hơn trên hầu hết tất cả các ví dụ thực nghiệm (ngoại trừ ví dụ 3). Mô hình CNN-LSTM-Attention được cung cấp thông tin những vùng ảnh cần tập trung tại từng bước trong quá trình phát sinh câu mô tả. Do đó, các từ chú thích được phát sinh dựa vào thông tin ảnh nhiều hơn và chính xác hơn.

2) Kết quả trực quan hóa những vùng ảnh mô hình CNN-LSTM-Attention tập trung để phát sinh từ kế tiếp: Trong phần này, nhóm thực hiện trực quan hóa những vùng ảnh mô hình tập trung để phát sinh từ kế tiếp. Nhomd đưa ảnh vào mô hình đã được huấn luyện để phát sinh ra câu mô tả nội dung ảnh. Các bước trực quan hóa được nhóm thực hiện như sau:

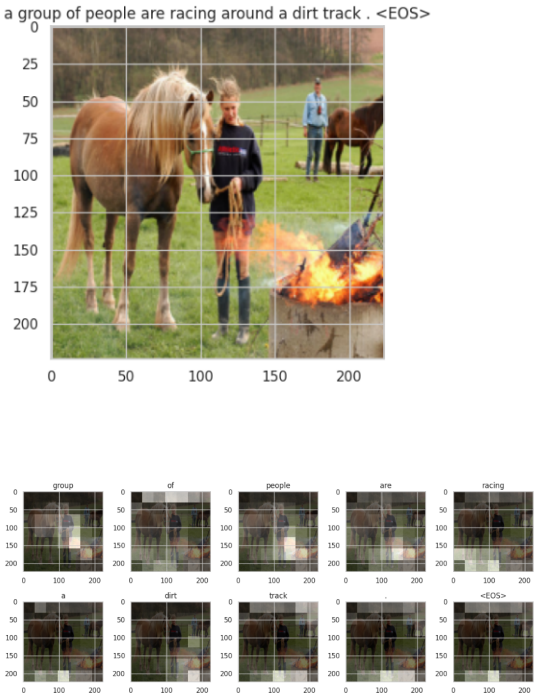
- Ảnh đầu vào được giảm kích thước về (224x224) .
- Tại mỗi bước LSTM phát sinh từ, ngoài việc trả về "context vector" sau khi đưa các véc-tơ đặc trưng ảnh và trạng thái ẩn trước đó vào mô hình Attention, chúng tôi cũng trả về bộ trọng số a_t của bước đó.
- Từ bộ trọng số a_t tại mỗi bước ta tiến hành "upscale" 16 lần thành ảnh xám với những vùng có giá trị a_{ti} càng lớn thì sẽ càng sáng. Như vậy, ta được ảnh xám kích thước (224x224).

Hình 9. Ảnh xám sau khi đã được "upscale"

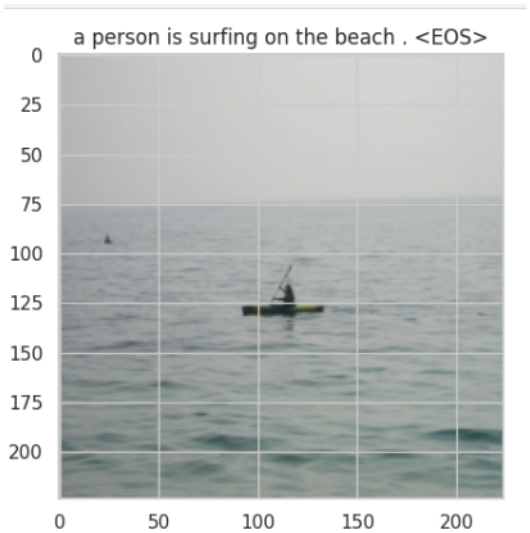
- Ta chồng ảnh xám này lên ảnh đầu vào tại mỗi bước để thấy những vùng cần tập trung (vùng sáng) là những vùng nào.

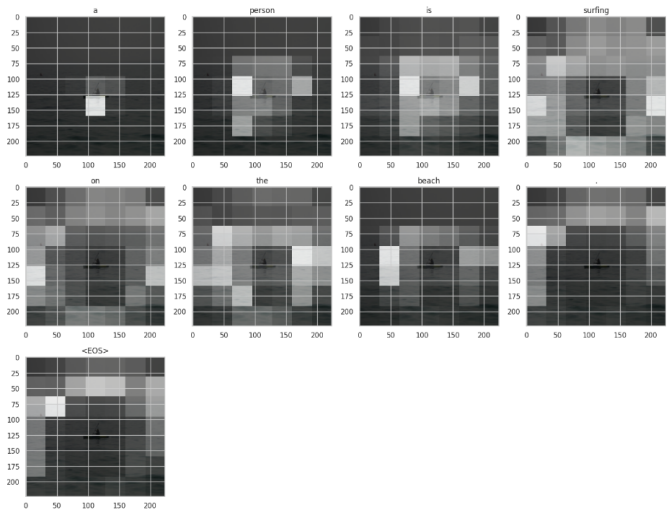
Hình 10. Ảnh thể hiện những vùng sáng cần tập trung

Các hình ảnh 11 12 dưới đây là kết quả trực quan hóa của mô hình CNN-LSTM-Attention:



Hình 11. Trực quan hóa những vùng ảnh mô hình CNN-LSTM-Attention tập trung để phát sinh từ kế tiếp.





Hình 12. Trực quan hóa những vùng ảnh mô hình CNN-LSTM-Attention tập trung để phát sinh từ kế tiếp. Câu mô tả được phát sinh ở đây là *"a person is surfing on the beach"*

TÀI LIỆU

[1] Y. Ming, N. N. Hu, C. X. Fan, F. Feng, J. W. Zhou, and H. Yu, "Visuals to text: A comprehensive review on automatic image captioning," IEEE/CAA Journal of Automatica Sinica, pp. 1339–1365, Aug 2022.

[2] S. Ramos, D. Elliott, and B. Martins, "Retrieval-augmented image captioning," 2023.

[3] A. Farhadi, M. Hejrati, A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," Proc. European Conf. Computer Vision, Springer, pp. 15–29, 2010.

[4] I. U. Rahman, Z. Wang, W. Liu, B. Ye, M. Zakarya, and X. Liu, "An n-state markovian jumping particle swarm optimization algorithm," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 11, pp. 6626–6638, 2020.

[5] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.

[6] R. Xu, C. Xiong, W. Chen, and J. Corso, "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 29, no. 1, 2015.

[7] T. Ghandi, H. Pourreza, and H. Mahyar, "Deep learning approaches on image captioning: A review," 2023.

[8] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," arXiv:1502.03044 [cs.LG], 2016.