GENERATING IMAGE CAPTIONS USING A HYBRID CNN-LSTM MODEL

Tran Quang Quy[1*], Nguyen Thi Khuyen[2], Bui Tien Lap[3]
[1]*TNU - University of Information and Communication Technology*
[2]*Tich Luong Secondary School, Thai Nguyen*
[3]*School of Technology - National Economics University*

ARTICLE INFO

Received: 30/6/2025
Revised: 24/8/2025
Published: 26/8/2025

KEYWORDS

Hybrid model
Flickr8k
Image
Caption
InceptionV3
LSTM

ABSTRACT

In the era of rapid advancements in artificial intelligence, image captioning has emerged as a crucial research domain, playing a key role in applications such as image retrieval, assistive technologies for visually impaired individuals, and multimedia content management. This paper proposes an automatic image captioning system based on a hybrid model that integrates an InceptionV3 convolutional neural network with a Long Short-Term Memory network. The model is trained and evaluated on the Flickr8k dataset. To enhance generalization performance and mitigate overfitting, the dropout technique is applied during the training process. Experimental results demonstrate that the proposed model generates image captions with BLEU-1 and BLEU-2 scores of 0.39 and 0.172, respectively. While the hybrid model does not achieve the same level of accuracy in image description as Transformer-based or Vision Transformer – Long Short-Term Memory models, its performance remains acceptable compared to baseline Convolutional Neural Network – Long Short-Term Memory models. These findings highlight the potential of applications in natural language processing and computer vision systems.

TẠO SINH CHÚ THÍCH HÌNH ẢNH BẰNG MÔ HÌNH LAI CNN-LSTM

Trần Quang Quý[1*], Nguyễn Thị Khuyên[2], Bùi Tiến Lập[3]
[1]*Trường Đại học Công nghệ Thông tin và Truyền thông - ĐH Thái Nguyên*
[2]*Trường Trung học Cơ sở Tích Lương, Thái Nguyên*
[3]*Trường Công nghệ - Đại học Kinh tế Quốc dân*

THÔNG TIN BÀI BÁO

Ngày nhận bài: 30/6/2025
Ngày hoàn thiện: 24/8/2025
Ngày đăng: 26/8/2025

TỪ KHÓA

Mô hình lai
Flickr8k
Ảnh
Chú thích
InceptionV3
LSTM

TÓM TẮT

Trong bối cảnh phát triển mạnh mẽ của trí tuệ nhân tạo hiện nay, bài toán tạo sinh chú thích ảnh (image captioning) đã trở thành một lĩnh vực nghiên cứu quan trọng, đóng vai trò chủ đạo trong các ứng dụng như truy xuất hình ảnh, hỗ trợ người khiếm thị và quản lý nội dung đa phương tiện. Bài báo này đề xuất một hệ thống tạo chú thích ảnh tự động dựa trên mô hình lai giữa mạng nơ-ron tích chập InceptionV3 và mạng bộ nhớ dài-ngắn hạn. Mô hình được huấn luyện và đánh giá trên bộ dữ liệu Flickr8k. Để cải thiện hiệu quả tổng quát hóa và giảm thiểu hiện tượng quá khớp, kỹ thuật dropout được áp dụng trong quá trình huấn luyện. Kết quả thực nghiệm cho thấy mô hình đề xuất có khả năng tạo sinh chú thích ảnh với các chỉ số đánh giá BLEU-1 và BLEU-2 lần lượt đạt 0,39 và 0,172. Mặc dù phương pháp sử dụng mô hình lai này chưa thể đảm bảo độ chính xác cao trong việc mô tả nội dung ảnh so với các mô hình Transformer hoặc mô hình biến đổi thị giác – bộ nhớ ngắn dài hạn, nhưng kết quả đạt được có thể chấp nhận được so với mô hình nơ rơn tích chập – bộ nhớ ngắn dài hạn cơ sở. Kết quả của bài báo mở ra tiềm năng ứng dụng trong các hệ thống xử lý ngôn ngữ tự nhiên và thị giác máy tính.

DOI: https://doi.org/10.34238/tnu-jst.13155

* Corresponding author. *Email: tqquy@ictu.edu.vn*

1. Giới thiệu

Việc tạo sinh chú thích hình ảnh (image captioning) đóng vai trò quan trọng trong việc thu hẹp khoảng cách giữa thị giác máy tính và xử lý ngôn ngữ tự nhiên, nhất là trong bối cảnh dữ liệu hình ảnh đang gia tăng với tốc độ nhanh chóng. Một trong những thách thức chủ yếu của bài toán này là sự đa dạng về nội dung, bố cục và ngữ cảnh ảnh đầu vào, từ đó đòi hỏi hệ thống phải đồng thời hiểu thông tin thị giác và ngôn ngữ để tạo ra mô tả chính xác, cô đọng [1]. Các hệ thống sinh chú thích ảnh tự động hiện nay đã được ứng dụng rộng rãi trong nhiều lĩnh vực như hỗ trợ người khiếm thị tiếp cận thông tin trực quan [2], tối ưu hóa tìm kiếm hình ảnh [3], cũng như tăng cường khả năng tổ chức và phân loại dữ liệu đa phương tiện [4].

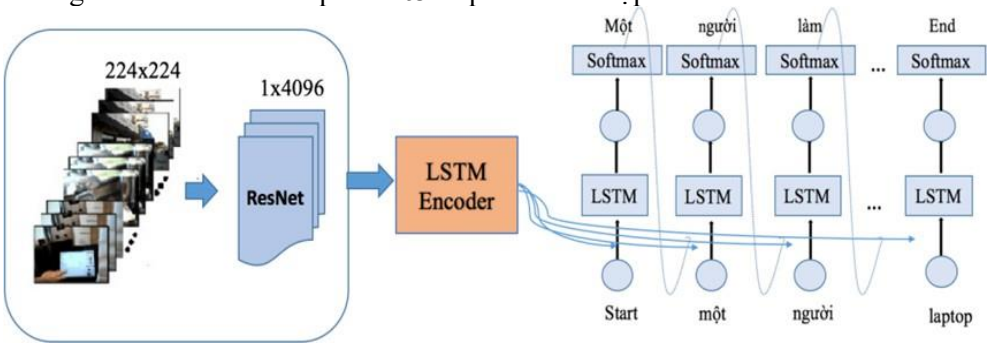
Với sự phát triển mạnh mẽ của các kiến trúc học sâu như mạng nơ ron tích chập (convolutional neural network - CNN), mạng bộ nhớ dài-ngắn hạn (Long Short-Term Memory - LSTM), và mô hình Transformer, việc nghiên cứu các mô hình sinh mô tả ảnh đang trở thành hướng đi chủ đạo trong tiến trình số hóa ngữ nghĩa dữ liệu thị giác. Đặc biệt, các kiến trúc lai giữa CNN và LSTM đã cho thấy hiệu quả vượt trội khi được triển khai trên các tập dữ liệu chuẩn.

Poddar và cộng sự [5] đề xuất một mô hình nơ rơn tích chập – bộ nhớ ngắn dài han (CNN-LSTM cho sinh mô tả ảnh tiếng Hindi theo cấu trúc encoder-decoder, sử dụng các backbone CNN) đã huấn luyện sẵn như AlexNet, VGG16 và DenseNet để trích xuất đặc trưng thị giác, sau đó sinh chuỗi mô tả bằng mạng nơ ron hồi tiếp (recurrent neural network - RNN). Kết quả thực nghiệm cho thấy mô hình này cải thiện BLEU-1 đến 34,64% và BLEU-2 đến 29,13% so với các phương pháp truyền thống.

Trong một hướng tiếp cận khác, Al-Malla và cộng sự [6] đã đề xuất một cơ chế chú ý kết hợp giữa mô hình trích xuất đặc trưng CNN Xception và mô-đun phát hiện đối tượng YOLOv4. Đặc biệt, nhóm nghiên cứu đưa ra khái niệm trọng số ưu tiên đối tượng (object saliency weighting), nhằm nâng cao độ chính xác của mô hình bằng cách tăng cường sự tập trung vào các đối tượng nổi bật ở tiền cảnh, đồng thời giảm thiểu ảnh hưởng từ các chi tiết nền không quan trọng.

Mishra và cộng sự [7] là nhóm nghiên cứu đầu tiên ứng dụng kiến trúc Transformer cho nhiệm vụ sinh mô tả ảnh tiếng Hindi. Nhóm nghiên cứu đã sử dụng CNN để mã hóa đặc trưng ảnh và Transformer làm bộ giải mã (decoder) cho chuỗi mô tả, loại bỏ hoàn toàn sự phụ thuộc vào RNN truyền thống.

Tại Việt Nam, Đặng Thị Phúc và các cộng sự [8] đã triển khai ba mô hình sinh mô tả video tiếng Việt trên bộ dữ liệu MSVD, bao gồm các kiến trúc dựa trên RNN, mô hình hồi quy dùng cổng (Gated Recurrent Unit – GRU) và Transformer. Trong đó, mô hình sequence-to-sequence sử dụng ResNet làm encoder và LSTM làm decoder đã cho thấy khả năng giữ ngữ cảnh qua các khung hình, cho phép mô tả chính xác các hành động đang diễn ra trong video. Hình 1 minh họa kiến trúc tổng thể của mô hình sequence-to-sequence tích hợp ResNet và LSTM:



Hình 1. *Mô hình sequence-to-sequence dựa trên ResNet và LSTM*

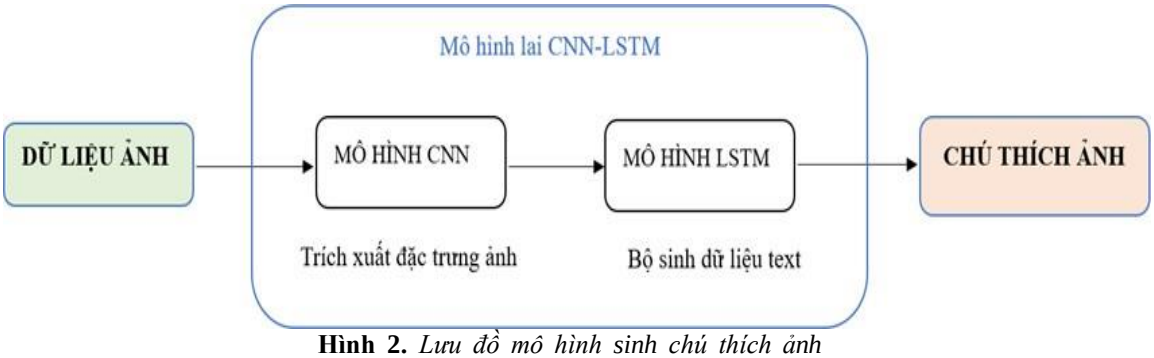
Gần đây, các nghiên cứu tiên tiến đã tập trung vào việc tích hợp các mô hình ngôn ngữ lớn (Large Language Models – LLMs) vào hệ thống sinh mô tả ảnh nhằm nâng cao khả năng biểu đạt ngữ nghĩa và tạo ra các mô tả tự nhiên, phong phú hơn. Nổi bật trong số đó, nghiên cứu của

Radford và cộng sự [9] đã giới thiệu mô hình CLIP (Contrastive Language–Image Pretraining) – một nền tảng thị giác–ngôn ngữ học từ dữ liệu tự nhiên với khả năng học biểu diễn đa phương thức có thể chuyển giao tốt sang nhiều tác vụ thị giác, bao gồm trích xuất đặc trưng ngữ nghĩa giàu thông tin từ ảnh. Dựa trên nền tảng này, nhóm tác giả Mokady và cộng sự [10] đề xuất ClipCap – một phương pháp sinh mô tả ảnh sử dụng CLIP để mã hóa ảnh dưới dạng tiền tố ngôn ngữ (text prefix), sau đó đưa vào bộ giải mã ngôn ngữ GPT-2 để sinh văn bản. Cách tiếp cận này thể hiện sự kết hợp hiệu quả giữa mô hình thị giác học trước và mô hình ngôn ngữ lớn, mở ra hướng phát triển các hệ thống sinh mô tả ảnh có khả năng hiểu ngữ cảnh sâu và tương tác người–máy linh hoạt hơn. Các nghiên cứu trước đây đã chỉ ra rằng việc tích hợp mạng CNN với mạng LSTM là một hướng tiếp cận hiệu quả cho nhiệm vụ tạo mô tả ảnh. CNN có khả năng trích xuất các đặc trưng hình ảnh giàu thông tin, trong khi LSTM đảm nhận vai trò xử lý chuỗi để sinh ra các câu văn tự nhiên phản ánh nội dung hình ảnh. Xuất phát từ nhận định đó, nghiên cứu này tập trung phân tích chi tiết phương pháp kết hợp giữa CNN và LSTM trong bài toán sinh mô tả hình ảnh, đồng thời trình bày các kết quả thực nghiệm nhằm đóng góp thêm dữ liệu thực chứng cho các nghiên cứu liên quan trong lĩnh vực này.

2. Phương pháp nghiên cứu

2.1. Mô hình CNN và LSTM

Trong nghiên cứu này, chúng tôi đề xuất một mô hình tích hợp giữa mạng CNN và mạng LSTM nhằm tự động nhận diện đối tượng trong hình ảnh và sinh chú thích tương ứng. Phương pháp khai thác kỹ thuật học chuyển giao (transfer learning) để chiết xuất đặc trưng và xác định đối tượng thông qua các mạng CNN đã được tiền huấn luyện. Kiến trúc đề xuất thực hiện đồng thời hai tác vụ: (1) phát hiện các đối tượng nổi bật trong ảnh bằng CNN đã huấn luyện trước, và (2) tạo mô tả ngôn ngữ cho ảnh thông qua LSTM. Mục tiêu của nghiên cứu là phát triển và đánh giá hiệu suất của một mô hình mạng nơ-ron sâu, trong đó nhấn mạnh sự phối hợp giữa khả năng chiết xuất thông tin không gian của CNN và năng lực mô hình hóa chuỗi thời gian của LSTM. Cụ thể, mô hình sử dụng hai thành phần chính: CNN để trích xuất đặc trưng từ hình ảnh và LSTM để biến đổi các đặc trưng này thành các câu mô tả. Quy trình sinh chú thích ảnh được minh họa trong Hình 2.



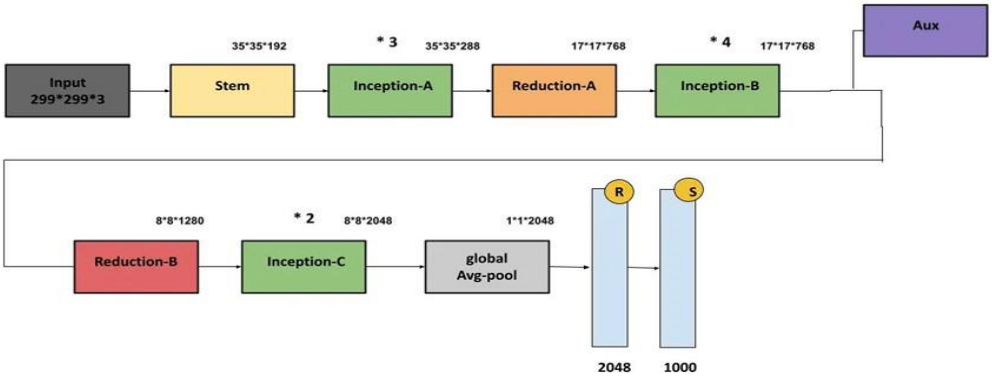
Hình 2. *Lưu đồ mô hình sinh chú thích ảnh*

2.2.1. Mạng nơ-ron tích chập INCEPTIONV3

Bài báo này khai thác kiến trúc InceptionV3, một mạng nơ-ron tích chập sâu (deep convolutional neural network) được tiền huấn luyện trên tập dữ liệu ImageNet với quy mô đa dạng và phức tạp, nhằm mục đích trích xuất đặc trưng hình ảnh đa cấp độ. InceptionV3, do Google phát triển, nổi bật với cấu trúc mô-đun đa nhánh (Inception module) [11], cho phép đồng thời áp dụng các phép tích chập với kích thước kernel khác nhau trong cùng một tầng, từ đó tối ưu hóa khả năng học các đặc trưng không gian đa tỷ lệ và nâng cao hiệu quả biểu diễn.

Mạng bao gồm 48 tầng với các lớp tích chập (convolutional layers), lấy mẫu tối đa (max-pooling) và lấy mẫu trung bình (average-pooling), phối hợp để tái cấu trúc thông tin đầu vào ở nhiều mức trừu tượng hóa khác nhau. Cơ chế thiết kế đa nhánh song song này không chỉ cải thiện khả năng tổng quát hóa của mạng mà còn giúp giảm thiểu hiện tượng gradient biến mất và hiện tượng quá khớp thông qua việc tăng cường đa dạng hóa các phép biến đổi đầu vào trên từng tầng.

Trong hệ thống sinh mô tả ảnh, phần thân mạng (feature extractor backbone) InceptionV3 được giữ nguyên nhằm thu nhận các đặc trưng không gian có độ sâu cao, trong khi lớp fully-connected cuối cùng phục vụ phân loại bị loại bỏ để tránh gây nhiễu đặc trưng do tính chuyên biệt của nhiệm vụ tiền huấn luyện. Đầu ra được lấy từ lớp tích chập sâu nhất tạo thành vector đặc trưng có kích thước 2048 chiều, cung cấp biểu diễn ngữ nghĩa đa chiều và giàu thông tin cho các mô-đun ngôn ngữ phía sau. Cấu trúc của mạng InceptionV3 được mô tả chi tiết trong Hình 3.



Hình 3. *Cấu trúc mạng InceptionV3 [11]*

2.1.2. Mạng hồi quy Long Short-term Memory (LSTM)

Mạng nơ-ron hồi quy LSTM là một dạng của mạng nơ-ron hồi quy RNN. RNN được gọi là mạng nơ-ron hồi quy vì nó áp dụng cùng một phép biến đổi cho mỗi phần tử của một chuỗi theo cách mà đầu ra của RNN phụ thuộc vào kết quả của các lần lặp trước đó. Do đó, RNN duy trì một trạng thái nội bộ để nắm bắt thông tin về các phần tử trước đó trong chuỗi, giống như bộ nhớ. Một đơn vị ẩn RNN sẽ được học dựa vào công thức (1), trong đó đầu ra của RNN, yt, là một phép biến đổi phi tuyến của tổng hai phép nhân ma trận, sử dụng ví dụ như các hàm kích hoạt tanh hoặc ReLU.

$$y_t = g(W_{hh}h_{t-1} + W_{xh}x_t) \quad (1)$$

Trong đó:

- W: ma trận dữ liệu cần tính toán
- W_{hh}: áp dụng cho trạng thái ẩn trước đó h_{t-1}
- W_{xh}: áp dụng cho trạng thái đầu vào hiện tại x_t

Mạng hồi quy LSTM sử dụng bốn biến trung gian $\bar{\imath}, \bar{f}, \bar{o}$ và $\bar{c}$ có kích thước p chiều. Ma trận cập nhật được ký hiệu là $W_{(k)}$ và được sử dụng để nhân trước với vector cột $[\bar{h}_t^{(k-1)}, \bar{h}_{t-1}^{(k)}]^T$. Cập nhật ma trận sẽ được biểu diễn bằng công thức (2):

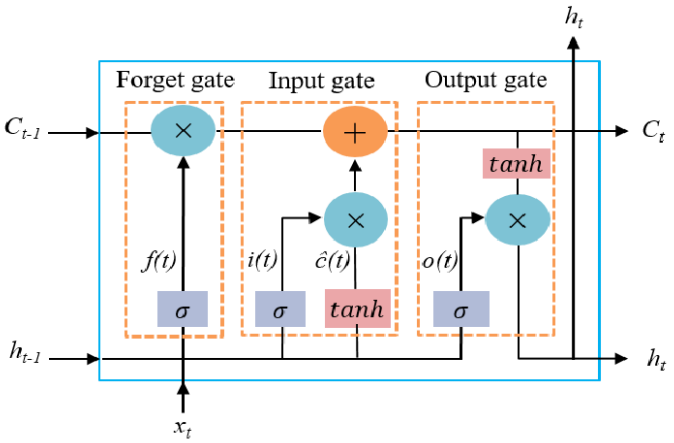
$$\begin{matrix} \text{Cổng đầu vào} \\ \text{Cổng quên} \\ \text{Cổng đầu ra} \\ \text{Trạng thái c mới} \end{matrix} \begin{bmatrix} \bar{\imath} \\ \bar{f} \\ \bar{o} \\ \bar{c} \end{bmatrix} = \begin{pmatrix} sigm \\ sigm \\ sigm \\ tanh \end{pmatrix} W^k \begin{bmatrix} \bar{h}_t^{(k-1)} \\ \bar{h}_{t-1}^{(k)} \end{bmatrix} \text{ [thiết lập tham số]} \quad (2)$$

Sự xác định vector trạng thái ẩn $\bar{h}_t^{(k)}$ và vector trạng thái ô $\bar{c}_t^{(k)}$ sử dụng một quá trình nhiều bước, bắt đầu bằng việc tính toán các biến trung gian tại lần lượt (3) và (4).

$$\bar{c}_t^{(k)} = \bar{f} \odot \bar{c}_{t-1}^{(k)} + \bar{\imath} \odot \bar{c} \text{ [Chọn lọc quên và thêm vào bộ nhớ dài hạn]} \quad (3)$$

$$\bar{h}_t^{(k)} = \bar{o} \odot tanh(\bar{c}_t^{(k)}) \text{ [Chọn lọc rò rỉ bộ nhớ dài hạn vào trạng thái ẩn]} \quad (4)$$

Tại đây, phép nhân từng phần của các vector được ký hiệu bằng "⊙", và ký hiệu "sigm" chỉ hàm sigmoid. Tại lớp thứ nhất (k = 1), h¯(k) trong phương trình trên nên được thay thế bằng x¯t và ma trận W (1) có kích thước p $\times_t$(p + d). Trong các triển khai thực tế, các hệ số điều chỉnh (bias) cũng được sử dụng trong các cập nhật ở trên. Kiến trúc mô hình LSTM được minh họa tại Hình 4.



Hình 4. *Kiến trúc mô hình LSTM*

2.2. Mô hình lai CNN và LSTM

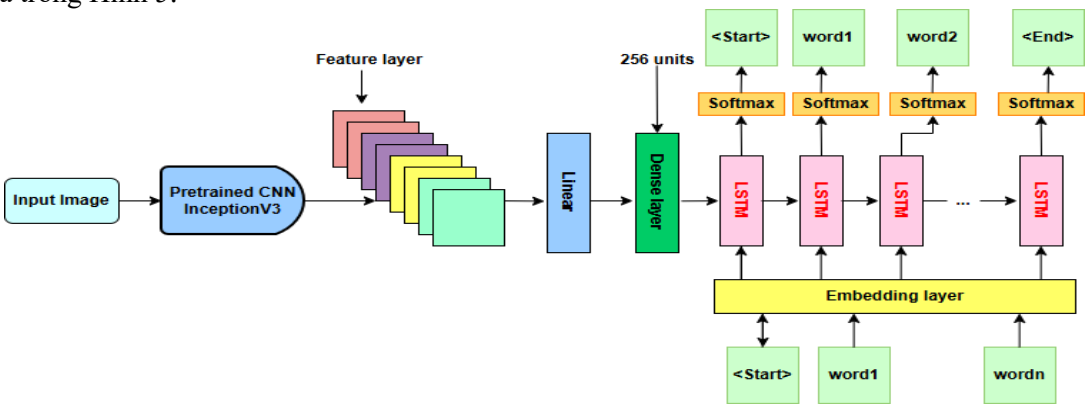
Để sinh mô tả cho một hình ảnh, hệ thống cần kết hợp biểu diễn từ hai miền: thị giác và ngôn ngữ tự nhiên. Mô hình đề xuất sử dụng kiến trúc encoder-decoder đa tầng, trong đó CNN đóng vai trò encoder để trích xuất đặc trưng hình ảnh, còn LSTM đảm nhận vai trò decoder để sinh câu mô tả tương ứng. Toàn bộ hệ thống gồm ba giai đoạn chính: trích xuất đặc trưng, mã hóa thông tin ngữ nghĩa, và giải mã thành chuỗi văn bản.

Trích xuất đặc trưng hình ảnh: Ảnh đầu vào được xử lý bởi mô hình InceptionV3 đã tiền huấn luyện, bỏ lớp fully connected cuối để giữ lại feature maps. Các đặc trưng được nén thành vector 2048 chiều thông qua Global Average Pooling, sau đó chuyển qua lớp Dense với 256 đơn vị và hàm ReLU để thu được vector đặc trưng đầu vào cho decoder.

Mã hóa chuỗi mô tả: Mỗi câu mô tả ảnh (caption) được biểu diễn dưới dạng chuỗi các từ, sau đó được mã hóa thông qua một lớp Embedding với chiều không gian 256. Việc chọn kích thước nhúng là 256 nhằm cân bằng giữa khả năng biểu diễn ngữ nghĩa của từ và chi phí tính toán. Kích thước này đủ lớn để nắm bắt các đặc trưng ngữ nghĩa quan trọng, đồng thời tránh được hiện tượng quá tải mô hình khi làm việc với tập dữ liệu không quá lớn. Sau lớp Embedding, mô hình áp dụng kỹ thuật Dropout với tỷ lệ 0,5 nhằm ngăn ngừa overfitting bằng cách loại ngẫu nhiên một phần các đơn vị trong quá trình huấn luyện. Tiếp theo, chuỗi các vector từ được đưa vào một lớp LSTM gồm 256 đơn vị ẩn (hidden units). Việc lựa chọn số lượng đơn vị này đảm bảo sự đồng nhất về kích thước với lớp Embedding và đủ khả năng học được các phụ thuộc dài hạn trong chuỗi từ. Đầu ra cuối cùng của lớp LSTM chính là vector ngữ cảnh – một biểu diễn tóm lược toàn bộ thông tin ngữ nghĩa của câu mô tả, dùng để kết hợp với đặc trưng ảnh trong bước giải mã mô tả.

Bộ giải mã và sinh văn bản: Hai vector đặc trưng từ ảnh và văn bản được nối với nhau và đưa vào một lớp Dense gồm 256 đơn vị để học các tương tác phi tuyến giữa hai miền thị giác và ngôn ngữ. Lớp này giúp mô hình tạo ra một không gian biểu diễn chung, làm cầu nối giữa đặc trưng ảnh và ngữ cảnh mô tả. Việc chọn kích thước 256 nhằm duy trì nhất quán với các lớp trước đó (Embedding và LSTM), đảm bảo hiệu quả tính toán mà vẫn đủ khả năng biểu diễn mối quan hệ

phức tạp giữa hai miền. Cuối cùng, lớp Softmax được sử dụng để dự đoán xác suất từ tiếp theo, từng bước sinh ra chuỗi mô tả ảnh hoàn chỉnh. Cấu trúc tổng thể của mô hình đề xuất được minh họa trong Hình 5.



Hình 5. *Kiến trúc mô hình lai CNN-LSTM đề xuất*

2.3. Tập dữ liệu và thiết lập siêu tham số

Trước khi đưa vào mô hình huấn luyện, các mô tả văn bản trong tập Flickr8k được xử lý qua nhiều bước nhằm chuẩn hóa và đảm bảo tính nhất quán dữ liệu. Toàn bộ văn bản được chuyển thành chữ thường, loại bỏ dấu câu và ký tự không cần thiết. Mỗi câu mô tả được token hóa thành chuỗi từ, sau đó bổ sung từ đặc biệt "start" ở đầu và "end" ở cuối để xác định rõ ranh giới câu trong quá trình sinh văn bản. Để xây dựng từ điển, chỉ các từ xuất hiện từ 5 lần trở lên trong tập huấn luyện được giữ lại, trong khi các từ hiếm được thay bằng từ "unk" (unknown). Các chuỗi mô tả được đệm (padding) về độ dài cố định là 34 từ – tương ứng với độ dài lớn nhất trong toàn bộ tập. Hình 6 mô tả ví dụ về dữ liệu trong tập Flickr8k.

Hình 6. *Dữ liệu trong tập Flickr8k*

Mô hình được triển khai trên nền tảng Python 3.8 sử dụng thư viện TensorFlow 2.10. Việc huấn luyện được thực hiện trên GPU NVIDIA RTX 3060 với bộ nhớ 12 GB.

Các siêu tham số được điều chỉnh để sử dụng cho mô hình LSTM được đề xuất chi tiết tại Hình 7, một số siêu tham số chính được mô tả như sau:.

- Batch size: 64.
- Epoch: 10.
- Kích thước tập kiểm thử: 25% tập huấn luyện.
- Mode: BLEU score.

Dữ liệu được chia thành ba tập: 6.000 ảnh cho huấn luyện, 1.000 ảnh cho kiểm tra và 1.091 ảnh cho đánh giá, đảm bảo không có sự chồng lặp giữa các tập con. Quy trình này giúp mô hình học hiệu quả hơn, đồng thời đảm bảo khả năng tái lập và so sánh kết quả trong các nghiên cứu liên quan.

Chiến lược huấn luyện: Mô hình sử dụng hàm mất mát sparse categorical crossentropy phù hợp cho các bài toán phân loại đa lớp với nhãn đầu ra ở dạng chỉ số nguyên. Hàm này đánh giá độ lệch giữa phân phối xác suất đầu ra của mô hình và nhãn thực tế, với công thức:

$$\mathcal{L}(y, \hat{y}) = -\log(\hat{y}_y) \#(5)$$

trong đó $y \in \{0,1,\dots,C-1\}$ là nhãn thực tế (dưới dạng chỉ số), và $\hat{y} = [\hat{y}_0, \hat{y}_1, \dots, \hat{y}_{C-1}]$ là

vector xác suất đầu ra từ mô hình (sau softmax). Giá trị mất mát được tính bằng log âm của xác suất tương ứng với nhãn đúng y, từ đó hướng mô hình tối ưu hóa để gia tăng xác suất dự đoán đúng trong quá trình huấn luyện.

Layer (type)	Output Shape	Param #	Connected to
Features_Input (InputLayer)	(None, 2048)	0	-
batch_normalization_94 (BatchNormalization)	(None, 2048)	8,192	Features_Input[0][0]
Sequence_Input (InputLayer)	(None, 33)	0	-
dense (Dense)	(None, 256)	524,544	batch_normalization_9…
embedding (Embedding)	(None, 33, 256)	2,243,584	Sequence_Input[0][0]
not_equal (NotEqual)	(None, 33)	0	Sequence_Input[0][0]
batch_normalization_95 (BatchNormalization)	(None, 256)	1,024	dense[0][0]
lstm (LSTM)	(None, 256)	525,312	embedding[0][0], not_equal[0][0]
add (Add)	(None, 256)	0	batch_normalization_9… lstm[0][0]
dense_1 (Dense)	(None, 256)	65,792	add[0][0]
Output_Layer (Dense)	(None, 8764)	2,252,348	dense_1[0][0]

Total params: 5,620,796 (21.44 MB)
Trainable params: 5,616,188 (21.42 MB)
Non-trainable params: 4,608 (18.00 KB)

Hình 7. *Mô hình LSTM được đề xuất*

Quá trình tối ưu hóa được thực hiện bằng thuật toán Adam với learning rate khởi tạo là 0,001, không sử dụng cơ chế điều chỉnh learning rate theo epoch (scheduler). Mô hình được huấn luyện với kích thước batch là 64 và tổng số epoch là 10. Kích thước batch được đặt là 64 nhằm cân bằng giữa tốc độ huấn luyện và khả năng xử lý của phần cứng. Số epoch là 10, được xác định sau khi quan sát quá trình huấn luyện và nhận thấy mô hình bắt đầu hội tụ mà không bị overfitting. Tập kiểm thử được lấy bằng 25% của tập huấn luyện để theo dõi hiệu suất mô hình trong quá trình đào tạo. Cuối cùng, BLEU score được sử dụng làm thước đo đánh giá chất lượng mô tả sinh ra, vì đây là chỉ số phổ biến trong các bài toán sinh ngôn ngữ tự động. Kết quả cụ thể sẽ được trình bày ở phần sau.

Ngoài kiến trúc CNN–LSTM truyền thống, mô hình được điều chỉnh kỹ thuật để nâng cao hiệu suất: lớp Dense 256 đơn vị giúp giảm chiều và đồng bộ với LSTM; Dropout 0,5 tại lớp Embedding nhằm hạn chế overfitting; và chuẩn hóa chuỗi mô tả với token đặc biệt ("start", "end") cùng padding cố định 34 từ để đảm bảo tính nhất quán. Các điều chỉnh này tuy nhỏ nhưng góp phần cải thiện khả năng tổng quát hóa trong điều kiện tính toán hạn chế.

3. Đánh giá kết quả

BLEU (Bilingual Evaluation Understudy) là chỉ số phổ biến trong đánh giá chất lượng sinh mô tả văn bản trong xử lý ngôn ngữ tự nhiên, đặc biệt với các mô hình như LSTM trong bài toán sinh chú thích ảnh. BLEU-n được tính dựa trên độ chính xác của các n-gram giữa câu sinh và câu tham chiếu, kèm theo hệ số phạt độ dài (brevity penalty) để hạn chế mô hình sinh ra các câu quá ngắn. Chỉ số này giúp phản ánh độ tương đồng ngữ nghĩa giữa mô tả do máy tạo ra và mô tả do con người viết [12].

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (6)$$

Trong đó:

- p_n: độ chính xác của n-gram bậc n,
- w_n: trọng số (với $w_n = 1$, thường $w_n = 1$),
- BP: hệ số phạt độ dài, xác định bởi:

$$BP = \begin{cases} 1 & nếu\ c > r \\ e^{1-\frac{r}{c}} & nếu\ c \leq r \end{cases} \quad (7)$$

Với:

- c: độ dài câu mô hình sinh ra,
- r: độ dài trung bình của câu tham chiếu.

BLEU có thể được tính với các mức độ n-gram khác nhau, trong đó **BLEU-1** và **BLEU-2** là hai biến thể cơ bản, tập trung vào sự so khớp của các unigram (1-gram) và bigram (2-gram).

BLEU-1 chỉ tính độ chính xác của unigrams (1-gram), tức là sự trùng khớp giữa các từ đơn lẻ trong câu mô hình sinh ra và câu tham chiếu. Công thức tính **BLEU-1** là:

$$BLEU - 1 = Precision\ of\ unigrams = \frac{\sum_{i=1}^{N} count(ngram_i)}{N} \quad (8)$$

Trong đó, **N** là số lượng các từ trong câu mô hình sinh ra, và count(ngrami) là số lần xuất hiện của từng unigram trong câu tham chiếu.

BLEU-2 tính độ chính xác của bigrams (2-grams), tức là các cặp từ liên tiếp trong câu. **BLEU-2** không chỉ xem xét các từ đơn lẻ, mà còn chú trọng vào ngữ cảnh của các từ trong câu. Công thức tính **BLEU-2** là:

$$BLEU - 2 = \frac{\sum_{i=1}^{N} count(bigram_i)}{N} \quad (9)$$

Trong đó, count($bigram_i$) là số lần xuất hiện của mỗi bigram trong câu tham chiếu. Mô hình được huấn luyện trong 10 vòng lặp (epoch), và các chỉ số BLEU-1, BLEU-2 được sử dụng để đánh giá mức độ chính xác về từ vựng cũng như khả năng duy trì ngữ cảnh trong quá trình sinh mô tả ảnh. Giá trị BLEU cao cho thấy mô hình tạo ra các câu mô tả gần với mô tả chuẩn do con người viết, phản ánh khả năng hiểu và biểu đạt nội dung hình ảnh một cách hiệu quả. Kết quả chi tiết được thể hiện trong Bảng 1.

Bảng 1. *Kết quả sau huấn luyện mô hình LSTM*

BLEU index	Score
BLEU-1	0,39
BLEU-2	0,172

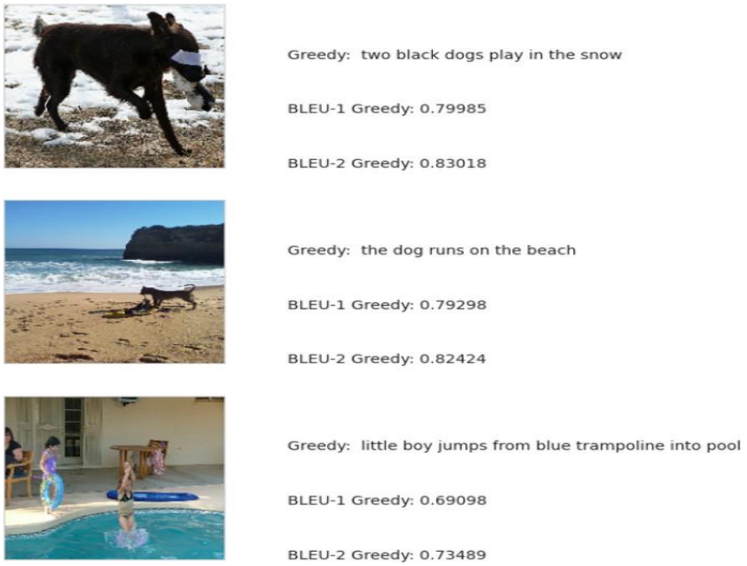
Để đánh giá hiệu quả của mô hình một cách khách quan hơn, chúng tôi tiến hành so sánh kết quả BLEU score với một số mô hình baseline và mô hình tiên tiến trên cùng tập dữ liệu Flickr8k. Cụ thể, theo nghiên cứu của Vinyals và các cộng sự [13], mô hình CNN–LSTM tiêu chuẩn đạt BLEU-1 khoảng 0,63 và BLEU-2 khoảng 0,42 trên Flickr8k. Trong khi đó, mô hình đề xuất trong nghiên cứu này đạt BLEU-1 là 0,39 và BLEU-2 là 0,172.

Bảng 2 trình bày kết quả so sánh mô hình đề xuất với các phương pháp tiêu biểu khác trong cùng tác vụ sinh chú thích ảnh. Có thể nhận thấy mô hình ViT-LSTM (Fine-tuned) đạt giá trị BLEU-1 = 0.42, cao hơn so với CNN-LSTM (0.32) và Attention-based (0.37). Tương tự, chỉ số BLEU-2 của ViT-LSTM cũng nhỉnh hơn so với hai mô hình còn lại. Kết quả này cho thấy việc khai thác đặc trưng hình ảnh bằng Vision Transformer kết hợp với LSTM mang lại hiệu quả vượt trội hơn các kiến trúc truyền thống.

Bảng 2. *So sánh kết quả với các mô hình khác*

Mô hình	BLEU-1	BLEU-2
CNN-LSTM [14]	0,32	0,17
Attention-based [15]	0,37	0,19
ViT-LSTM (Fine-tuned) [16]	0,42	0,23
Mô hình đề xuất (InceptionV3-LSTM)	0,39	0,172

Mặc dù kết quả còn khoảng cách so với các mô hình tối ưu, việc sử dụng kiến trúc đơn giản với số lượng tham số hạn chế và điều kiện tính toán vừa phải cho thấy mô hình vẫn có khả năng sinh ngôn ngữ ở mức cơ bản. Kết quả này mở ra tiềm năng cải tiến mô hình thông qua việc tối ưu hóa siêu tham số hoặc tích hợp thêm các mô hình ngôn ngữ mạnh hơn như Transformer hoặc GPT trong các nghiên cứu tiếp theo. Hình 8 chỉ ra kết quả của mô hình LSTM trên tập dữ liệu kiểm tra, cho thấy độ chính xác cao trong việc sinh mô tả ảnh từ các đặc trưng hình ảnh đầu vào. Các chỉ số BLEU đạt được ở mức đáng kể, phản ánh khả năng của mô hình trong việc tạo ra các mô tả ngữ nghĩa phù hợp với nội dung ảnh. Ví dụ minh họa trong hình này thể hiện rõ cách thức mô hình sinh chú thích dựa trên các đặc trưng hình ảnh đã được mã hóa.



Hình 8. *Kết quả thu được từ tập kiểm tra*

Mặc dù mô hình đạt BLEU-1 và BLEU-2 ở mức chấp nhận được, quá trình đánh giá định tính cho thấy vẫn tồn tại một số lỗi phổ biến, đặc biệt trong các ảnh có cảnh phức tạp hoặc ngữ nghĩa không rõ ràng. Cụ thể, mô hình đôi khi bỏ sót đối tượng chính (object omission), lặp từ không cần thiết (repetition), mô tả sai ngữ nghĩa (semantic drift), và kém hiệu quả với ảnh nền phức tạp. Những hạn chế này cho thấy khả năng mô hình hóa ngữ cảnh còn hạn chế, đồng thời gợi mở hướng tích hợp cơ chế chú ý (attention mechanism) trong các nghiên cứu tiếp theo. Mô hình hoạt động tốt với các ảnh có bối cảnh đơn giản, chủ thể rõ ràng. Tuy nhiên, trong các ảnh có nhiều đối tượng hoặc nền phức tạp, mô hình thường bỏ sót thông tin chính, mô tả sai ngữ nghĩa hoặc lặp từ. Ví dụ, với ảnh "một bé gái chơi bóng trên bãi cỏ", mô hình dự đoán đúng chủ thể và hành động; nhưng với ảnh "nhiều người đi bộ trên đường phố đông đúc", mô hình mô tả thiếu hoặc không rõ hành động. Những quan sát này cho thấy tiềm năng nâng cấp mô hình thông qua tích hợp attention hoặc các kỹ thuật xử lý ngữ cảnh sâu hơn.

4. Kết luận

Bài báo này trình bày phương pháp kết hợp mô hình học sâu lai giữa mạng CNN và LSTM trong tác vụ sinh mô tả ảnh tự động. Trong đó, CNN đảm nhiệm vai trò trích xuất đặc trưng

không gian từ ảnh đầu vào thông qua kiến trúc InceptionV3 có độ sâu và tính đa tỷ lệ cao, trong khi LSTM được triển khai như một bộ giải mã tuần tự (sequential decoder) nhằm sinh chuỗi văn bản mô tả dựa trên vector đặc trưng đầu vào. Cấu hình mô hình phản ánh hướng tiếp cận encoder–decoder cổ điển nhưng có điều chỉnh phù hợp cho miền thị giác–ngôn ngữ.

Kết quả thực nghiệm trên tập dữ liệu chuẩn cho thấy hệ thống có khả năng sinh mô tả ở mức độ chấp nhận được về mặt ngữ nghĩa và cú pháp, với hiệu suất tốt trên các mẫu ảnh có bối cảnh đơn giản đến trung bình. Tuy nhiên, mô hình vẫn cho thấy sự suy giảm đáng kể về độ chính xác trong các trường hợp có cấu trúc cảnh phức tạp hoặc ngữ nghĩa mơ hồ, phản ánh giới hạn trong khả năng mô hình hóa phụ thuộc ngữ cảnh dài hạn và hiểu cảnh ngữ nghĩa sâu. Trong tương lai, hướng nghiên cứu tiếp theo sẽ tập trung vào việc tích hợp cơ chế chú ý (attention mechanism) và khai thác các mô hình ngôn ngữ lớn (LLMs) để tăng cường khả năng hiểu ngữ cảnh sâu và cải thiện chất lượng sinh văn bản. Ngoài ra, việc mở rộng bộ dữ liệu huấn luyện và đa dạng hóa ngôn ngữ mô tả cũng được xem là yếu tố tiềm năng nhằm nâng cao tính tổng quát của mô hình.

TÀI LIỆU THAM KHẢO/REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 3156–3164.

[2] V. Ordonez, G. Kulkarni, and T. L. Berg, “Im2text: Describing images using 1 million captioned photographs,” *Adv. Neural Inf. Process. Syst.*, vol. 24, pp. 1143–1151, 2011.

[3] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, “A comprehensive survey of deep learning for image captioning,” *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, Jan. 2019.

[4] X. Chen and C. L. Zitnick, “Mind's eye: A recurrent visual representation for image caption generation,” *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2422–2431.

[5] A. K. Poddar, S. Ghosh, and P. Bhattacharyya, “Image caption generation in Hindi using deep learning,” *Procedia Comput. Sci.*, vol. 218, pp. 388–395, 2023.

[6] K. Al-Malla, F. Al-Doghman, and M. Abou El-Nasr, “Image captioning with attention mechanism using Xception and YOLOv4,” *IEEE Access*, vol. 9, pp. 112491–112503, 2021.

[7] A. Mishra, A. Shah, and P. Bhattacharyya, “Hindi image captioning using transformers,” in *Proc. Int. Conf. Comput. Vis. Graph.*, 2022, pp. 325–335.

[8] T. P. Dang, Q. T. Tran, V. T. Nguyen, and S. H. Dau, “Video captioning in Vietnamese using deep learning,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 3, pp. 3092–3103, Jun. 2022, doi: 10.11591/ijece.v12i3.pp3092-3103.

[9] J. Radford, K. Narasimhan, and I. Sutskever, “Learning transferable visual models from natural language supervision,” *in Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[10] P. Mokady, A. Hertz, and A. Shamir, “ClipCap: CLIP prefix for image captioning,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 570–588.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception architecture for computer vision,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2818–2826.

[12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” *in Proc. 40th Annu. Meeting Assoc. Comput. Linguist. (ACL)*, Philadelphia, PA, USA, 2002, pp. 311–318.

[13] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 3156–3164.

[14] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” *arXiv preprint arXiv:1412.2306*, Dec. 2014.

[15] 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 preprint arXiv:1502.03044*, Feb. 2015.

[16] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, “Image captioning with semantic attention,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 4651–4659.