TẠP CHÍ KHOA HỌC TRƯỜNG ĐẠI HỌC HOA LƯ
ISSN 2615-9538
Website: http://hluv.edu.vn/vi/tckh

NGHIÊN CỨU VỀ CÁC KỸ THUẬT, BỘ DỮ LIỆU VÀ ĐỘ ĐO CỦA SINH CHÚ THÍCH CHO ẢNH

Phạm Thị Thanh[1]

Ngày nhận bài: 14/10/2023 *Ngày chấp nhận đăng: 21/12/2023*

Tóm tắt: *Bài báo này nghiên cứu tổng quan về các kỹ thuật để sinh chú thích cho ảnh như chú thích ảnh dựa trên truy xuất thông tin, dựa trên khuôn mẫu và đặc biệt chú thích học dựa trên kỹ thuật học sâu đã mang lại cuộc cách mạng trong sinh chú thích cho ảnh. Ngoài cập nhật các nghiên cứu mới ra, bài báo còn giới thiệu các tập dữ liệu phục vụ huấn luyện và kiểm thử hệ thống sinh chú thích, các loại độ đo phổ biến để đánh giá hiệu quả sinh chú thích cho ảnh. Phần kết luận bài báo đề xuất một số hướng nghiên cứu về lĩnh vực chú thích ảnh mà các nhà nghiên cứu có thể đi sâu tìm hiểu.*

Từ khóa*: Chú thích cho ảnh, truy xuất thông tin, khuôn mẫu, học sâu, huấn luyện, kiểm thử, bộ dữ liệu, độ đo.*

IMAGE CAPTIONING: A SURVEY OF METHODS, DATASETS, EVALUATION METRICS

Abstract: *This article studies an overview of techniques to generate captions for images such as image captioning based on information retrieval, based on templates and especially based on deep learning, which has brought a revolution in generating captions for photos. In addition to updating the new studies, the study also introduces datasets for training and testing the image captioning system, common metrics to evaluate the efficiency of images captioning. The conclusion of the article proposes some research directions in the field of image captioning that researchers can study further.*

Keywords: *image captioning, information retrieval, template, deep learning, test, datasets, metrics*

GIỚI THIỆU

"Một bức tranh đáng giá hơn ngàn lời nói" là một câu ngạn ngữ trong nhiều ngôn ngữ, có nghĩa là những ý tưởng phức tạp, đôi khi là nhiều ý tưởng có thể được truyền tải bằng một hình ảnh tĩnh duy nhất, truyền đạt ý nghĩa hoặc bản chất của nó hiệu quả hơn là chỉ mô tả bằng lời nói.

Trong cuộc sống chúng ta bắt gặp rất nhiều hình ảnh từ các nguồn khác nhau như sách, báo, sơ đồ, tài liệu, hình ảnh đa phương tiện, … con người có thể nhận dạng và hiểu hình ảnh đó một cách tự nhiên dễ dàng hơn rất nhiều khi máy tính làm công việc nhận dạng đó.

Chú thích ảnh – IC (Image Captioning) là một lĩnh vực mới kết hợp giữa thị giác máy tính – CV (Computer Vision) và xử lý ngôn ngữ tự nhiên (Natural Language Processing). Chú thích ảnh là việc máy tính tự động tạo ra các cụm từ hoặc câu hợp lý về mặt ngôn ngữ và trung thực về

[1] Trung tâm Ngoại ngữ - Tin học, Trường Đại học Hoa Lư; Email: ptthanh@hluv.edu.vn

mặt ngữ nghĩa mô tả nội dung của một hình ảnh cho trước. Chú thích ảnh là lĩnh vực được các nhà khoa học quan tâm nghiên cứu gần đây và đã đạt được nhiều kết quả quan trọng.

Bài báo này chúng tôi nghiên cứu những thành tựu của các nhà khoa học trên thế giới hiện nay về lĩnh vực Chú thích ảnh đồng thời chỉ ra những hạn chế và thách thức hiện tại và trong tương lai.

NỘI DUNG

1. Các phương pháp sinh chú thích ảnh

Các phương pháp sinh chú thích ảnh có thể được chia thành 3 loại như dưới đây:

1.1. Chú thích ảnh dựa trên truy xuất thông tin (retrieval-based method)

Một loại phương pháp chú thích hình ảnh phổ biến ban đầu là dựa trên truy xuất. Đưa ra một hình ảnh truy vấn, các phương thức dựa trên truy xuất sẽ tạo chú thích bằng cách truy xuất một hoặc một tập hợp các câu từ nhóm câu được chỉ định trước. Chú thích được tạo có thể là một câu đã tồn tại hoặc một câu được tạo từ những câu đã truy xuất.

Trong nghiên cứu [1], Farhadi thiết lập một bộ ba (đối tượng, hành động, ngữ cảnh) tạo thành không gian liên kết các hình ảnh và câu. Đưa ra một hình ảnh truy vấn, mô hình ánh xạ nó vào không gian liên kết bằng cách giải Trường ngẫu nhiên Markov và sử dụng phép đo độ tương tự Lin để xác định khoảng cách ngữ nghĩa giữa hình ảnh này và mỗi câu hiện có được trình phân tích cú pháp Curran phân tích cú pháp. Câu gần nhất với hình ảnh truy vấn được lấy làm chú thích.

Trong nghiên cứu [2], để tạo chú thích cho một hình ảnh, trước tiên tác giả sử dụng các bộ mô tả hình ảnh toàn cục để truy xuất một tập hợp các hình ảnh từ một bộ sưu tập các bức ảnh có chú thích ở quy mô web. Sau đó, họ sử dụng nội dung ngữ nghĩa của các hình ảnh được truy xuất để thực hiện xếp hạng lại và sử dụng chú thích của hình ảnh trên cùng làm mô tả của truy vấn.

Trong nghiên cứu [3], tác giả xem chú thích hình ảnh như một nhiệm vụ xếp hạng. Các tác giả sử dụng kỹ thuật phân tích sự tương quan chính tắc hàm nhân (Kernel Canonical Correlation Analysis) để chiếu các mục hình ảnh và văn bản vào một không gian chung, nơi các hình ảnh và chú thích tương ứng của chúng có mối tương quan tối đa. Trong không gian chung mới, độ tương đồng cosine giữa hình ảnh và câu được tính toán để chọn các câu được xếp hạng cao nhất đóng vai trò mô tả hình ảnh truy vấn.

Trong [4], Gupta sử dụng bộ công cụ Stanford CoreNLP để xử lý các câu trong bộ dữ liệu ảnh - chú thích để lấy danh sách các cụm từ cho mỗi hình ảnh. Để tạo mô tả cho một hình ảnh, việc truy xuất hình ảnh trước tiên được thực hiện dựa trên các tính năng hình ảnh toàn cục để truy xuất một tập hợp các hình ảnh cho truy vấn. Sau đó, một mô hình được huấn luyện để xác định các cụm từ liên quan được sử dụng để chọn các cụm từ từ những cụm từ được liên kết với hình ảnh được truy xuất. Cuối cùng, một câu mô tả được tạo ra dựa trên các cụm từ có liên quan đã chọn.

1.2. Chú thích ảnh dựa trên khuôn mẫu (template-based method)

Trong các phương pháp dựa trên mẫu, chú thích hình ảnh được tạo thông qua một quy trình bị ràng buộc về mặt cú pháp và ngữ nghĩa. Thông thường, để sử dụng phương pháp dựa trên mẫu, trước tiên cần phát hiện một tập hợp các khái niệm trực quan cụ thể. Sau đó, được kết nối thông qua các mẫu câu hoặc quy tắc ngữ pháp ngôn ngữ cụ thể hoặc thuật toán tối ưu hóa tổ hợp để soạn câu.

Trong nghiên cứu [5], Yang sử dụng bộ tứ (danh từ-động từ-cảnh-giới từ) làm mẫu câu. Để mô tả một ảnh, đầu tiên nhóm tác giả sử dụng các thuật toán dò tìm để ước lượng các đối tượng và cảnh trong ảnh. Sau đó, sử dụng một mô hình ngôn ngữ được huấn luyện qua Gigaword corpus3 để xác định động từ, cảnh và giới từ có thể được sử dụng để soạn câu. Với xác suất của tất cả các phần tử được tính toán, bộ tứ tốt nhất thu được bằng cách sử dụng suy luận Mô hình Markov ẩn. Cuối cùng, mô tả hình ảnh được tạo ra bằng cách điền vào cấu trúc câu được cung cấp bởi bộ tứ.

Kulkarni sử dụng Trường ngẫu nhiên có điều kiện để xác định nội dung hình ảnh sẽ được hiển thị trong chú thích hình ảnh [6], [7]. Theo đó, các nút của đồ thị tương ứng với các đối tượng, thuộc tính đối tượng và mối quan hệ không gian giữa các đối tượng. Nội dung hình ảnh cần mô tả được xác định bằng cách thực hiện suy luận Trường ngẫu nhiên có điều kiện. Đầu ra của suy luận được sử dụng để tạo mô tả dựa trên mẫu câu.

Li sử dụng các mô hình trực quan để trích xuất thông tin ngữ nghĩa bao gồm các đối tượng, thuộc tính và các mối quan hệ không gian [8]. Sau đó, họ xác định một bộ ba định dạng ((adj1, obj1), prep, (adj2, obj2)) để mã hóa kết quả nhận dạng. Để tạo mô tả với bộ ba, dữ liệu n-gram quy mô web, có khả năng cung cấp số lượng tần suất của các chuỗi n-gram, được sử dụng để thực hiện lựa chọn cụm từ, thu thập các cụm từ ứng viên có thể tạo nên bộ ba. Sau đó, hợp nhất cụm từ được triển khai để sử dụng lập trình động nhằm tìm ra tập hợp cụm từ tương thích tối ưu để đóng vai trò mô tả hình ảnh truy vấn.

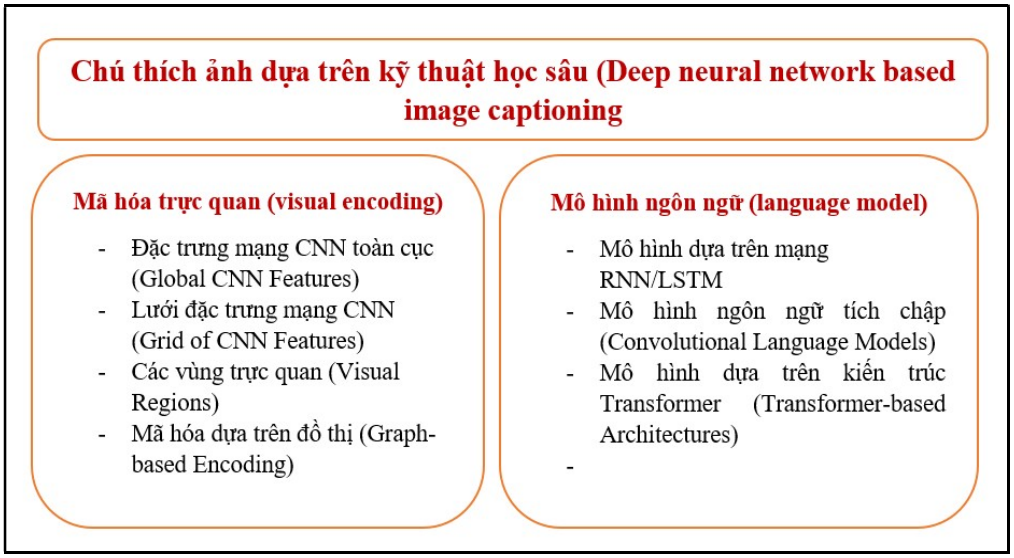
Mitchell sử dụng các thuật toán thị giác máy tính để xử lý một hình ảnh và thể hiện hình ảnh này bằng cách sử dụng bộ ba (đối tượng, hành động, mối quan hệ không gian) [9]. Sau đó, họ xây dựng mô tả hình ảnh như một quá trình tạo cây dựa trên kết quả nhận dạng hình ảnh. Thông qua cụm danh từ đối tượng và sắp xếp thứ tự, tác giả xác định nội dung hình ảnh cần miêu tả. Sau đó, cây con được tạo cho danh từ đối tượng, được sử dụng thêm để tạo cây đầy đủ. Cuối cùng, một mô hình ngôn ngữ bát quái được sử dụng để chọn một chuỗi từ các cây đầy đủ được tạo làm mô tả của hình ảnh.

Ushiku trình bày một phương pháp gọi là không gian con chung cho mô hình và tính tương tự để học các bộ phân loại cụm từ trực tiếp cho các hình ảnh chú thích [10]. Cụ thể, các tác giả trích xuất các từ liên tục từ chú thích huấn luyện dưới dạng cụm từ. Sau đó, họ ánh xạ các đặc điểm hình ảnh và đặc điểm cụm từ vào cùng một không gian con, trong đó phân loại dựa trên mô hình và dựa trên sự tương đồng được tích hợp để tìm hiểu bộ phân loại cho từng cụm từ. Trong giai đoạn suy luận, các cụm từ ước tính từ một hình ảnh truy vấn được kết nối bằng cách sử dụng tìm kiếm chùm nhiều ngăn để tạo mô tả.

Chú thích hình ảnh dựa trên mẫu có thể tạo ra các câu đúng về mặt cú pháp và các mô tả thường phù hợp hơn với nội dung hình ảnh so với các mô tả dựa trên truy xuất. Tuy nhiên, do việc tạo mô tả dựa trên mẫu với số lượng mô hình có sẵn thường nhỏ, nên hạn chế phạm vi bao phủ, tính sáng tạo và độ phức tạp của các câu được tạo ra. Hơn nữa, so với chú thích do con người viết, việc sử dụng các mẫu cứng nhắc làm cấu trúc chính của câu sẽ khiến các mô tả được tạo ra kém tự nhiên hơn.

1.3. Chú thích ảnh dựa trên kỹ thuật học sâu (Deep neural network based image captioning)

Nhóm chú thích ảnh dựa trên kỹ thuật học sâu được phân chia làm hai nhóm chính: Mã hóa trực quan (visual encoding) và Mô hình ngôn ngữ (Language Model) (xem hình 1). Những nghiên cứu đã cải thiện đáng kể hiệu năng mô hình: từ các đề xuất dựa trên học sâu đầu tiên áp dụng mạng nơ ron hồi quy (RNN - Recurrent Neural Network) cho đến những đột phá của mô hình transfomer và phương pháp tiếp cận tựa BERT (Bidirectional Encoder Representations from Transformers) với cơ chế "tự chú ý". Đồng thời, đã giải được quyết thách thức trong việc xây dựng các giao thức và chỉ số đánh giá phù hợp để so sánh kết quả với các chú thích do con người tạo ra.



Hình 1. Các phương pháp mô tả ảnh dựa trên học sâu

1.3.1. Mã hóa trực quan (Visual encoding)

1.3.1.1. Đặc trưng mạng CNN (Convolutional Neural Network) toàn cục (Global CNN Features)

Với sự ra đời của CNN, các mô hình sử dụng đầu vào trực quan đã được cải thiện về mặt hiệu suất. Bước mã hóa trực quan của chú thích hình ảnh cũng không ngoại lệ. Trong công thức đơn giản nhất, việc sử dụng một trong những lớp cuối cùng của CNN để trích xuất các biểu diễn cấp cao, sau đó được sử dụng làm đầu vào cho mô hình ngôn ngữ. Đây là cách tiếp cận được sử dụng trong bài báo "Show and Tell" [12], trong đó đầu ra của GoogleNet được đưa vào trạng thái ẩn ban đầu của mô hình ngôn ngữ. Karpathy đã sử dụng các tính năng toàn cầu được trích xuất từ AlexNet làm đầu vào cho một mô hình ngôn ngữ. Mao và Donahue thêm các tính năng toàn cầu được trích xuất từ mạng VGG (Visual Geometry Group) tại mỗi bước thời gian của mô hình ngôn ngữ.

Các đặc trưng mạng CNN toàn cục sau đó đã được sử dụng trong rất nhiều mô hình chú thích hình ảnh. Đáng chú ý, Rennie [13] đã giới thiệu mô hình SC (Self-Criticle), trong đó hình ảnh được mã hóa bằng ResNet-101, giữ nguyên kích thước ban đầu của chúng. Các cách tiếp cận khác [14], [15] tích hợp các thuộc tính hoặc thẻ cấp cao, được biểu diễn dưới dạng phân phối xác suất trên các từ phổ biến nhất của chú thích huấn luyện.

Ưu điểm chính của việc sử dụng các đặc trưng mạng CNN toàn cầu nằm ở tính đơn giản và nhỏ gọn của biểu diễn, bao gồm khả năng trích xuất và cô đọng thông tin từ toàn bộ đầu vào và xem xét bối cảnh tổng thể của hình ảnh. Tuy nhiên, mô hình này cũng dẫn đến việc nén thông tin quá mức và thiếu độ chi tiết, khiến mô hình khó tạo ra mô tả chi tiết.

1.3.1.2. Lưới đặc trưng mạng CNN (Grid of CNN Features)

Do các hạn chế của biểu diễn toàn cục, hầu hết các cách tiếp cận sau đây đã làm tăng mức độ chi tiết của mã hóa hình ảnh. Dai và cộng sự [16] đã sử dụng các bản đồ kích hoạt 2D thay cho các vectơ đặc trưng toàn cầu 1D để đưa trực tiếp cấu trúc không gian vào mô hình ngôn ngữ. Cảm hứng từ mô hình dịch máy, cơ chế chú ý được sử dụng cung cấp cho mô hình khả năng mã hóa các đặc điểm hình ảnh thay đổi theo thời gian, cho phép tính linh hoạt cao và độ chi tiết tốt hơn.

1.3.1.3. Các vùng trực quan (Visual Regions)

Trực giác sử dụng sự nổi bật bắt nguồn từ khoa học thần kinh, điều này gợi ý rằng bộ não của chúng ta tích hợp quá trình lập luận từ trên xuống với luồng tín hiệu hình ảnh từ dưới lên. Lộ trình từ trên xuống bao gồm dự đoán đầu vào giác quan sắp tới bằng cách tận dụng kiến thức và khuynh hướng quy nạp của chúng ta, trong khi lộ trình từ dưới lên cung cấp các kích thích thị

giác điều chỉnh các dự đoán trước đó. Sự chú ý bổ sung có thể được coi là một hệ thống từ trên xuống. Trong cơ chế này, mô hình ngôn ngữ dự đoán từ tiếp theo trong khi tham gia vào lưới tính năng, có dạng hình học không phụ thuộc vào nội dung hình ảnh.

Theo cách tiếp cận này, mạng Faster R-CNN [17] được sử dụng để phát hiện các đối tượng, thu được một vectơ đặc trưng gộp cho từng vùng. Một trong những yếu tố chính của phương pháp này nằm trong chiến lược huấn luyện trước của nó, trong đó một bộ hỗ trợ tính toán mất mát được thêm vào để học cách dự đoán các lớp thuộc tính cùng với các lớp đối tượng trên bộ dữ liệu Visual Genome. Điều này cho phép mô hình dự đoán một tập hợp phát hiện dày đặc và phong phú, bao gồm cả đối tượng nổi bật và vùng theo ngữ cảnh, đồng thời hỗ trợ việc học các biểu diễn tính năng tốt hơn.

1.3.1.4. Mã hóa dựa trên đồ thị (Graph-based Encoding)

Để cải thiện hơn nữa việc mã hóa các vùng hình ảnh và mối quan hệ của chúng, một số nghiên cứu xem xét sử dụng các đồ thị được tạo trên các vùng hình ảnh để làm phong phú thêm biểu diễn bằng cách bao gồm các kết nối ngữ nghĩa và không gian.

Nỗ lực đầu tiên theo nghĩa này là của Yao [18], tiếp theo là Guo [19], người đã đề xuất sử dụng mạng tích chập đồ thị (GCN) để tích hợp cả mối quan hệ ngữ nghĩa và không gian giữa các đối tượng. Biểu đồ mối quan hệ ngữ nghĩa thu được bằng cách áp dụng một trình phân loại được huấn luyện trước trên Visual Genome để dự đoán một hành động hoặc tương tác giữa các cặp đối tượng.

Tập trung vào việc mô hình hóa các mối quan hệ ngữ nghĩa, Yang [20] đã đề xuất tích hợp các tiền đề ngữ nghĩa đã học được từ văn bản trong mã hóa hình ảnh bằng cách khai thác biểu diễn dựa trên đồ thị của cả hình ảnh và câu. Biểu diễn được sử dụng là biểu đồ cảnh, tức là đồ thị có hướng kết nối các đối tượng, thuộc tính và quan hệ của chúng. Shi [21] biểu diễn hình ảnh dưới dạng đồ thị quan hệ ngữ nghĩa nhưng đề xuất huấn luyện mô-đun chịu trách nhiệm dự đoán các nút vị ngữ trực tiếp trên chú thích thực thay vì trên bộ dữ liệu bên ngoài.

Mã hóa đồ thị mang đến cơ chế tận dụng mối quan hệ giữa các đối tượng được phát hiện, cho phép trao đổi thông tin trong các nút liền kề và do đó theo cách cục bộ. Hơn nữa, nó cho phép tích hợp thông tin ngữ nghĩa bên ngoài. Mặt khác, việc xây dựng cấu trúc đồ thị theo cách thủ công có thể hạn chế sự tương tác giữa các đặc trưng trực quan. Đây là nơi tự chú ý tỏ ra thành công hơn bằng cách kết nối tất cả các yếu tố với nhau trong một biểu diễn đồ thị hoàn chỉnh.

1.3.2. Mô hình ngôn ngữ (Language Model)

1.3.2.1. Mô hình dựa trên mạng RNN/LSTM (Recurrent Neural Network / Long Short-Term Memory)

Vì ngôn ngữ có cấu trúc tuần tự, RNN phù hợp để tạo ra các câu. LSTM là một biến thể khắc phục một số hạn chế của RNN, với cơ chế chú ý bổ sung, được sử dụng phổ biến trong mô hình hóa ngôn ngữ.

Để nối các từ vào các vùng hình ảnh, Lu [22] đã kết hợp một mạng trỏ giúp điều chỉnh cơ chế chú ý dựa trên nội dung. Đặc biệt, trong quá trình tạo câu, mạng dự đoán các vị trí trong chú thích, sau đó các vị trí này sẽ được lấp đầy bằng các lớp vùng hình ảnh. Đối với những từ không trực quan, một "lính canh trực quan" (visual sentinel) được sử dụng làm nền tảng giả.

Ke [23] đã giới thiệu hai mô-đun phản xạ: mô-đun thứ nhất tính toán mức độ liên quan giữa các trạng thái ẩn từ tất cả các từ được dự đoán trong quá khứ và từ hiện tại. Mô-đun thứ hai cải thiện cấu trúc cú pháp của câu bằng cách hướng dẫn quá trình tạo câu bằng thông tin vị trí chung của từ.

Huang [24] đã đề xuất một cơ chế thời gian chú ý thích ứng, trong đó bộ giải mã có thể thực hiện một số bước chú ý tùy ý cho mỗi từ được tạo, được xác định bởi một mạng tin cậy trên LSTM lớp thứ hai.

Một số nghiên cứu khác đã sử dụng tự chú ý thay cho chú ý cộng thêm trong các mô hình ngôn ngữ dựa trên LSTM, LSTM tăng cường bằng "chú ý trên chú ý", tính toán một bước chú ý

khác trên sự chú ý trực quan, tăng cường khả năng tự chú ý với các tương tác thứ hai và cải thiện cả mã hóa hình ảnh và mô hình ngôn ngữ.

1.3.2.2. Mô hình ngôn ngữ tích chập (Convolutional Language Models)

Mô hình ngôn ngữ tích chập được đề xuất bởi Aneja [25], sử dụng một vectơ đặc trưng hình ảnh toàn cầu được kết hợp với các từ nhúng và được cung cấp cho CNN, hoạt động trên tất cả các từ song song trong quá trình đào tạo và suy luận tuần tự. Convolutions được che bên phải để ngăn mô hình sử dụng thông tin của các từ trong tương lai. Mặc dù có lợi thế rõ ràng của huấn luyện song song, việc sử dụng toán tử tích chập trong các mô hình ngôn ngữ đã không trở nên phổ biến do hiệu suất kém và sự ra đời của kiến trúc Transformer.

1.3.2.3. Mô hình dựa trên kiến trúc Transformer (Transformer-based Architectures)

Mô hình chú ý đầy đủ (fully-attentive paradigm) được đề xuất bởi Vaswani và cộng sự [26] đã thay đổi hoàn toàn quan điểm về mô hình ngôn ngữ. Ngay sau đó, mô hình Transformer đã trở thành nền tảng của những bước đột phá khác trong NLP, chẳng hạn như BERT [27] và GPT [28], và kiến trúc tiêu chuẩn cho nhiều tác vụ "hiểu ngôn ngữ". Vì chú thích hình ảnh có thể được xem là một vấn đề theo trình tự, nên kiến trúc Transformer cũng đã được sử dụng cho nhiệm vụ này. Bộ giải mã Transformer ban đầu đã được sử dụng trong một số mô hình chú thích hình ảnh mà không cần sửa đổi kiến trúc đáng kể. Bên cạnh đó, một số biến thể đã được đề xuất để cải thiện việc tạo ngôn ngữ và mã hóa tính năng trực quan.

Li [29] đề xuất cơ chế tạo cổng cho toán tử chú ý chéo, điều khiển luồng thông tin hình ảnh và ngữ nghĩa bằng cách kết hợp và điều chỉnh các biểu diễn vùng hình ảnh với các thuộc tính ngữ nghĩa đến từ một trình gắn thẻ bên ngoài. Ji [30] tích hợp một cơ chế kiểm tra ngữ cảnh để điều chỉnh ảnh hưởng của biểu diễn hình ảnh toàn cục đối với mỗi từ được tạo, được mô hình hóa thông qua sự chú ý của nhiều đầu attention (MultiHead). Cornia [31] đề xuất tính đến tất cả các lớp mã hóa thay vì chỉ thực hiện chú ý chéo trên lớp cuối cùng. Để đạt được mục tiêu này, họ đã phát minh ra bộ giải mã dạng lưới, chứa toán tử dạng lưới điều chỉnh phần đóng góp.

H. Pavin (2023) [32] Thiết kế mạng sinh dựa trên transformer như một hướng mức từ (word-level) để sinh ra từ tiếp theo dựa trên trạng thái hiện tại. Sau đó huấn luyện không gian ngầm để học cách sắp xếp chú thích và ảnh vào cùng một không gian nhúng, rút ra mối quan hệ giữa ảnh và chú thích. Cuối cùng thiết kế mạng lựa chọn như là 1 hướng mức câu (sentence-level) để đánh giá từ tiếp theo bằng cách gán điểm thích hợp cho phần chú thích thông qua không gian nhúng. Nhóm nghiên cứu bổ sung thêm ba thành phần Self-Eliminator Module (SEM), Mask Attention Weight (MAW), Multiple Attention Distribution (MAD) để tăng hiệu quả mô hình. Nhóm nghiên cứu thử nghiệm trên hai bộ dữ liệu chuẩn là Microsoft COCO và Flickr30K.

2. Tập dữ liệu và phương pháp đánh giá

2.1. Tập dữ liệu

Dữ liệu là nền tảng của trí tuệ nhân tạo. Con người ngày càng phát hiện ra rằng có thể tìm thấy nhiều định luật khó tìm từ một lượng lớn dữ liệu. Trong tác vụ tạo chú thích hình ảnh, cần sử dụng các bộ dữ liệu để chạy thử nghiệm. Nhóm nghiên cứu có thể sử dụng một hoặc hai bộ để thử nghiệm và đánh giá mô hình của mình. Một số bộ dữ liệu phổ biến hiện nay: MSCOCO, Flickr8k, Flickr30k, PASCAL 1K, AI Challenger Dataset và STAIR Captions.

Bảng 1. Một số tập dữ liệu chú thích ảnh

Dataset	Train	Valid	Test
MSCOCO	82783	40504	40775
Filckr8k	6000	1000	1000
Filckr30k	32783	1000	100
PASCAL 1K			1000
AIC	210000	30000	30000
STAIR	82783	40504	40775

Trong các bộ dữ liệu trên thì hai bộ được sử dụng nhiều hơn cả là:

Flickr30k: bao gồm 31783 ảnh từ Flickr bao phủ rộng các hành động của con người, mỗi ảnh có 5 chú thích. Bộ dữ liệu Flickr30k do cộng đồng các nhà nghiên cứu về thị giác máy tính xây dựng và phát triển.

MS-COCO: tập dữ liệu này phức tạp hơn do chứa nhiều đối tượng, nhiều nền và mối quan hệ đa dạng. Nó chứa 82783, trong đó 40504 ảnh cho huấn luyện và xác thực, 40775 ảnh cho thử nghiệm. Bộ dữ liệu MS-COCO do nhóm nghiên cứu thuộc tập đoàn Microsoft tạo ra và chia sẻ với cộng đồng nghiên cứu.

2.2. Phương pháp đánh giá

Trên thực tế, cách trực quan nhất để xác định mức độ một câu được tạo ra mô tả nội dung của hình ảnh tốt như thế nào là dựa vào phán đoán trực tiếp của con người. Tuy nhiên, vì đánh giá của con người đòi hỏi một lượng lớn nỗ lực không thể tái sử dụng, nên rất khó để mở rộng quy mô. Hơn nữa, đánh giá của con người vốn mang tính chủ quan khiến nó có sự khác nhau của người dùng khác nhau. Do đó, các chỉ số, phương pháp đánh giá được đưa ra nhằm tự động đánh giá kết quả tạo câu mô tả của các mô hình.

Khi đánh giá kết quả tạo câu, BLEU (BiLingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit Ordering), ROUGE (Recall-Oriented Understudy for Gisting Evaluation) và CIDer (Consensus-based Image Description Evaluation) thường được sử dụng làm chỉ số đánh giá. Đối với năm chỉ báo, BLEU và METEOR dành cho bản dịch máy, ROUGE dành cho tóm tắt tự động và CIDEr dành cho chú thích hình ảnh. Chúng đo tính nhất quán của n-gram giữa các câu được tạo, điều này bị ảnh hưởng bởi tầm quan trọng và độ hiếm của n-gram. Đồng thời, cả bốn chỉ số này đều có thể được tính toán trực tiếp bằng công cụ đánh giá MSCOCO. (Bộ công cụ đánh giá MSCOCO là COCO Evaluation Toolkit. Công cụ này cung cấp các phương pháp để đánh giá chất lượng của các mô hình mô tả ảnh dựa trên các tiêu chí như BLEU, METEOR, ROUGE, và CIDEr. COCO Evaluation Toolkit giúp cung cấp một cách tiêu chuẩn để đo lường hiệu suất của các mô hình mô tả ảnh, làm cho quá trình so sánh và đánh giá kết quả trở nên công bằng và đồng nhất trong cộng đồng nghiên cứu.)

BLEU [33] là sử dụng các cụm từ có độ dài thay đổi của một câu ứng viên để khớp với các câu tham chiếu để đo lường mức độ gần gũi của chúng. Nói cách khác, số liệu BLEU được xác định bằng cách so sánh một câu ứng viên với các câu tham chiếu tính bằng n-gam. Điểm số unigram (BLEU-1) thể hiện mức độ phù hợp, trong khi điểm số n-gram cao hơn thể hiện mức độ lưu loát của câu.

METEOR [34] là thước đo đánh giá bản dịch máy tự động. Trước tiên, nó thực hiện khớp unigram tổng quát giữa câu ứng cử viên và câu tham chiếu, sau đó tính điểm dựa trên kết quả khớp. Việc tính toán liên quan đến độ chính xác, thu hồi và sắp xếp các từ phù hợp. Trường hợp có nhiều câu đối chiếu thì điểm cao nhất trong số các câu được tính độc lập được lấy làm kết quả đánh giá cuối cùng. Việc giới thiệu chỉ số này là để giải quyết điểm yếu của chỉ số BLEU, chỉ được tạo ra dựa trên độ chính xác của n-gram phù hợp.

ROUGE [35] là viết tắt của Recall-Oriented Understudy for Gisting Evaluation, được sử dụng để tự động xác định chất lượng của một bản tóm tắt văn bản bằng cách so sánh nó với các bản tóm tắt (lý tưởng) khác do con người tạo ra. Các biện pháp đếm số lượng đơn vị như n-gram, chuỗi từ và cặp từ giữa bản tóm tắt do máy tính tạo ra và bản tóm tắt lý tưởng do con người tạo ra. ROUGE gồm bốn thước đo ROUGE khác nhau: ROUGE-N, ROUGE-L, ROUGE-W và ROUGE-S.

CIDer [36] là một mô hình sử dụng sự đồng thuận để đánh giá chất lượng của chú thích hình ảnh. Số liệu này đo lường mức độ giống nhau của một câu do phương pháp chú thích hình ảnh tạo ra với phần lớn các câu do con người tạo ra. Nó đạt được điều này bằng cách mã hóa tần suất của n-gam trong câu ứng cử viên để xuất hiện trong các câu tham chiếu, trong đó sử dụng

trọng số Tần suất tài liệu nghịch đảo tần số (Frequency Inverse Document Frequency) cho mỗi n-gam. Số liệu này được thiết kế để đánh giá các câu được tạo theo các khía cạnh về ngữ pháp, mức độ nổi bật, tầm quan trọng và độ chính xác.

KẾT LUẬN

Chú thích ảnh là sự kết nối của hai lĩnh vực thị giác máy tính (CV) và xử lý ngôn ngữ tự nhiên (NLP), đang thu hút sự quan tâm của các nhà nghiên cứu. Chú thích ảnh/video phải giải quyết các vấn đề của cả hai lĩnh vực CV và NLP. Vì vậy, mặc dù đã đạt được nhiều kết quả quan trọng, nó vẫn là một vấn đề khá thách thức. Đối với thị giác máy tính, mạng CNN đã đem lại thành công lớn như phân loại ảnh, phát hiện đối tượng… Trong lĩnh vực xử NLP, mô hình Transformer với cơ chế Attention đã đem lại các kết quả SOTA (State-of-the-art) trong nhiều lĩnh vực như chatbot, dịch máy… Mặc dù áp dụng các tiến bộ công nghệ của cả hai lĩnh vực, Chú thích ảnh vẫn chưa đạt được các kết quả như mong muốn. Một vấn đề gây cản trở rất lớn là dữ liệu. Để có chất lượng nhận dạng và chú thích ảnh cao thường yêu cầu một lượng rất lớn dữ liệu được chú thích, đây là một công việc rất khó khăn và tốn kém. Khác với NLP có thể dễ dàng thu thập lượng lớn dữ liệu từ internet trong nhiều lĩnh vực. Hơn nữa dữ liệu file ảnh thường có dung lượng lớn, nên thường không đủ tài nguyên để huấn luyện các mô hình rất lớn như GPT trong NLP được huấn luyện trên hàng TB (45TB) dữ liệu văn bản với 175 tỷ tham số. Việc khai thác các công nghệ SOTA nhất của cả hai lĩnh vực như đã nêu trên, hoặc giải pháp giảm dung lượng file ảnh để tăng cường dữ liệu huấn luyện vẫn là một chủ đề khá tiềm năng cho sinh chú thích ảnh.

Ngoài ra, dữ liệu ảnh biến thiên đa dạng hơn rất nhiều so với dữ liệu tiếng nói hoặc văn bản. Ngoài việc ghép các đối tượng, cảnh, màu sắc, góc quay, tư thế…, chỉ một từ "cái ghế", nhưng có rất nhiều kiểu ghế khác nhau trong thế giới đối tượng ảnh. Số lượng từ vựng trong NLP là xác định, nhưng số lượng đối tượng trong thế giới xung quanh gần như là vô hạn. Do vậy việc nhận dạng để chú thích các đối tượng phong phú trong ảnh trong thế giới thực là rất khó khăn. Do đó cần nhiều nghiên cứu và thử nghiệm hơn nữa để sinh chú thích ảnh có thể áp dụng được trong thế giới thực một cách chân thực hơn.

Một số hướng nghiên cứu cho các nhà khoa học là việc khai thác các mô hình của công nghệ học sâu tiên tiến hiện nay, đặc biệt là công nghệ CNN trong phân loại/ phát hiện đối tượng và cơ chế Attention cho mô hình ngôn ngữ, áp dụng cho Chú thích ảnh với các đối tượng phong phú là một hướng đi thiết thực và nhiều hứa hẹn, không chỉ dưới góc độ phương pháp luận mà còn cả trong các ứng dụng trong thế giới thực với muôn vàn các đối tượng khám phá trong môi trường xung quanh. Thêm vào đó, chúng ta cũng thấy rằng, để đạt hiệu quả tốt hơn, cần xem xét các vấn đề trong bối cảnh tổng thể hơn, từ dữ liệu huấn luyện tới kiến trúc mô hình. Kiến trúc transformer đa nhiệm với các nhiệm vụ nhận dạng và mô tả cảnh với các đối tượng mới, các nhiệm vụ mô tả các câu khác nhau trong mô tả đoạn hình ảnh, có lẽ sẽ là mô hình phù hợp để xem xét và gắn kết các lĩnh vực SOTA của CV và NLP với các dữ liệu dễ dàng thu thập được từ internet. Trong các nghiên cứu về chú thích ảnh, nếu xem xét các đặc trưng ảnh mà không tính đến ngữ cảnh có thể dẫn đến mô tả sai lầm. Mặt khác, trong các mô hình, mô tả còn yếu về xử lý thời gian thực, đó là hướng mở cho các nghiên cứu sau này.

TÀI LIỆU THAM KHẢO

[1] Ali Farhadi et al., "Every Picture Tells a Story: Generating Sentences from Images," *Computer Vision – ECCV 2010. Lecture Notes in Computer Science,* vol. 6314, no. Springer, Berlin, Heidelberg., 2010.

[2] V. Ordonez, G. Kulkarni, T.L. Berg., "Im2Text: describing images using 1 million captioned photographs," *Proceedings of the Advances in Neural Information Processing Systems,* p. 1143–1151, 2011.

[3] M. Hodosh, P. Young, J. Hockenmaier, "Framing image description as a ranking task: data, models and evaluation metrics," *J. Artif. Intell. Res. ,* vol. 47, p. 853–899, 2013.

[4] A. Gupta, Y. Verma, C.V. Jawahar, "Choosing linguistics over vision to describe images," *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 5, 2012.

[5] Y. Yang, C.L. Teo, H. Daume, Y. Aloimono, "Corpus-guided sentence generation of natural images," *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* p. 444–454, 2011.

[6] G. Kulkarni et al., "Baby talk: understanding and generating simple image descriptions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2011.

[7] G. Kulkarni et al, "BabyTalk: understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.,* p. 2891–2903, 2013.

[8] S. Li et al., "Composing simple image descriptions using web-scale n-grams," *Proceedings of the Fifteenth Conference on Computational Natural Language Learning,* 2011.

[9] M. Mitchell at al., "Generating image descriptions from computer vision detections," *Proceedings of the Thirteenth Conference of the European Chapter of the Association for Computational Linguistics,* 2012.

[10] Y. Ushiku et al., "Common subspace for model and similarity: phrase learning for caption generation from images," *IEEE International Conference on Computer Vision,* p. 2668–2676, 2015.

[11] Matteo Stefanini et al., "From Show to Tell: A Survey on Deep Learning-based Image Captioning," *arXiv:2107.06912v3 [cs.CV],* 2021.

[12] Oriol Vinyals et al., "Show and Tell: A Neural Image Caption Generator," *arXiv:1411.4555 [cs.CV, 2015].*

[13] S. J. Rennie et al., "Selfcritical sequence training for image captioning," *in CVPR,* 2017.

[14] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," *in ICCV,* 2017.

[15] Z. Gan et al, "Semantic Compositional Networks for Visual Captioning," *in CVPR,* 2017.

[16] B. Dai, D. Ye, and D. Lin, "Rethinking the form of latent states in image captioning," *in ECCV,* 2018.

[17] Shaoqing Ren et al., "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans. PAMI, vol. 39, no. 6,* p. pp. 1137–1149, 2017.

[18] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring Visual Relationship for Image Captioning," *in ECCV,* 2018.

[19] L. Guo et al., "Aligning linguistic words and visual semantic units for image captioning," *in ACM Multimedia,,* 2019.

[20] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-Encoding Scene Graphs for Image Captioning," *in CVPR,* 2019.

[21] Z. Shi, X. Zhou, X. Qiu, and X. Zhu, "Improving Image Captioning with Better Use of Captions," *in ACL,* 2020.

[22] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural Baby Talk," *in CVPR,* 2018.

[23] L. Ke, W. Pei, R. Li, X. Shen, and Y.-W. Tai, "Reflective Decoding Network for Image Captioning," *in ICCV,* 2019.

[24] L. Huang, W. Wang, Y. Xia, and J. Chen, "Adaptively Aligned Image Captioning via Adaptive Attention Time," *in NeurIPS,* 2019.

[25] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," *in CVPR,* 2018.

[26] Ashish Vaswani et al., "Attention Is All You Need," *31st Conference on Neural Information Processing Systems (NIPS 2017),* no. Long Beach, CA, USA, 2017.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," *NAACL,* 2018.

[28] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[29] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled Transformer for Image Captioning," *in ICCV,* 2019.

[30] J. Ji et al., "Improving Image Captioning by Leveraging Intra- and Interlayer Global Representation in Transformer Network," *in AAAI,* 2021.

[31] M. Cornia et al, "MeshedMemory Transformer for Image Captioning," *In CVPR,* 2020.

[32] Hashem Parvin, Ahmad Reza Naghsh-Nilchi, Hossein Mahvash Mohammadi, "Transformer-based local-global guidance for image captioning," *Expert Systems with Applications, vol. 223,* 2023.

[33] K. Papineni, S. Roukos, T. Ward, W. Zhu, "BLEU: a method for automatic evaluation of machine translation," *in: Proceedings of the Meeting on Association for Computational Linguistics, vol. 4,* 2002.

[34] A. Lavie, A. Agarwal, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," *in: Proceedings of the Second Workshop on Statistical Machine Translation,* p. pp. 228–231, 2007.

[35] C.-Y. Lin, "ROUGE: a package for automatic evaluation of summaries, in Proceedings of the Text Summarization Branches Out," *Workshop on Text Summarization Branches Out, Barcelona, Spain,* 2004.

[36] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: consensus-based image description evaluation," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA,* p. pp. 4566–4575, 2015.

[37] [Online]. Available: https://en.wikipedia.org/wiki/A_picture_is_worth_a_thousand_words.