|  |  |
| --- | --- |
| X |  |
| THÂN NGỌC THIỆN | BỘ CÔNG THƯƠNG  TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI  --------------------------------------- |
|  |
| ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC |
| NGÀNH HỆ THỐNG THÔNG TIN |
| NGHIÊN CỨU PHÁT TRIỂN GIẢI PHÁP TẠO MÔ TẢ  TỰ ĐỘNG CHO HÌNH ẢNH |
|  |
|  |
| CBHD :TS. Phạm Văn Hà |
| HỆ THỐNG THÔNG TIN | Sinh viên : Thân Ngọc Thiện |
| Mã số sinh viên: 2021602775 |
|  |
|  |
|  |
| Hà Nội – Năm 2024 |
|  |

# 

|  |
| --- |
|  |
| BỘ CÔNG THƯƠNG  TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI  --------------------------------------- |
|  |
| ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC  NGÀNH HỆ THỐNG THÔNG TIN |
| NGHIÊN CỨU PHÁT TRIỂN GIẢI PHÁP TẠO MÔ TẢ  TỰ ĐỘNG CHO HÌNH ẢNH |
|  |
|  |
|  |
| CBHD :TS. Phạm Văn Hà |
| Sinh viên : Thân Ngọc Thiện |
| Mã số sinh viên : 2021602775 |
|  |
| Hà Nội – 2024 |

**LỜI CẢM ƠN**

Sau những năm tháng học tập và rèn luyện tại Trường Đại học Công Nghiệp Hà Nội, đồ án tốt nghiệp này là cột mốc quan trọng, đánh dấu sự hoàn thành chặng đường học vấn trên ghế giảng đường của em. Trong suốt quá trình nghiên cứu và thực hiện đồ án với đề tài: “Nghiên cứu phát triển giải pháp tạo mô tả tự động cho hình ảnh”, em đã nhận được sự quan tâm, giúp đỡ nhiệt tình từ nhiều cá nhân và tập thể. Với lòng biết ơn sâu sắc và chân thành, em xin gửi lời cảm ơn đến tất cả những người đã tạo điều kiện, hỗ trợ em trong quá trình học tập và hoàn thành đề tài này.

Lời đầu tiên, em xin gửi lời chào trân trọng, lời chúc sức khỏe và lời cảm ơn sâu sắc nhất đến các thầy cô giáo khoa Công Nghệ Thông Tin, Trường Đại học Công Nghiệp Hà Nội. Sự tận tâm giảng dạy, những kiến thức quý báu và sự chỉ bảo ân cần của thầy cô đã trang bị cho em nền tảng vững chắc để có thể hoàn thành đồ án tốt nghiệp này.

Đặc biệt, em xin bày tỏ lòng biết ơn sâu sắc và chân thành đến TS. Phạm Văn Hà, người thầy đã dành thời gian, tâm huyết hướng dẫn, định hướng và hỗ trợ em vượt qua những khó khăn, thử thách trong quá trình nghiên cứu và hoàn thiện đề tài.

Em cũng xin gửi lời cảm ơn đến Ban Giám hiệu Trường Đại học Công Nghiệp Hà Nội, các phòng ban chức năng đã tạo điều kiện thuận lợi cho em được học tập trong môi trường năng động, chuyên nghiệp và được tiếp cận những kiến thức, công nghệ tiên tiến, đặc biệt trong lĩnh vực Trí tuệ Nhân tạo và Xử lý ảnh. Những kiến thức và kỹ năng tích lũy được trong quá trình học tập sẽ là hành trang quý báu cho em trên con đường sự nghiệp sau này.

Mặc dù đã cố gắng hết mình, song do kiến thức và kinh nghiệm còn hạn chế, đồ án tốt nghiệp này chắc chắn không tránh khỏi những thiếu sót. Kính mong nhận được sự góp ý, chỉ bảo từ quý thầy cô để em có thể hoàn thiện và phát triển hơn nữa trong lĩnh vực nghiên cứu này.

# **MỤC LỤC**

[MỤC LỤC i](#_Toc185848182)

[DANH MỤC CÁC CHỮ VIẾT TẮT ii](#_Toc185848183)

[DANH MỤC CÁC HÌNH VẼ iii](#_Toc185848184)

[MỞ ĐẦU 1](#_Toc185848185)

[1. Giới thiệu đề tài 1](#_Toc185848186)

[2. Lý do chọn đề tài 2](#_Toc185848187)

[3. Mục tiêu của đề tài 3](#_Toc185848188)

[4. Đối tượng và phạm vi nghiên cứu 4](#_Toc185848189)

[5. Kết quả dự kiến đạt được 4](#_Toc185848190)

[6. Bố cục của đề tài 5](#_Toc185848191)

[CHƯƠNG 2: TỔNG QUAN VỀ TRÍ TUỆ NHÂN TẠO, HỌC MÁY, HỌC SÂU VÀ BÀI TOÁN MÔ TẢ HÌNH ẢNH 7](#_Toc185848192)

[2.1. Tổng quan về trí tuệ nhân tạo, học máy và học sâu 7](#_Toc185848193)

[2.1.1. Trí tuệ nhân tạo 7](#_Toc185848194)

[2.1.2. Học máy 9](#_Toc185848195)

[2.1.3. Học sâu 11](#_Toc185848196)

[2.2. Định nghĩa bài toán mô tả hình ảnh 13](#_Toc185848197)

[2.3. Ứng dụng của bài toán mô tả hình ảnh 14](#_Toc185848198)

[2.4. Các dạng triển khai chính của bài toán mô tả hình ảnh 15](#_Toc185848199)

[2.5. Các phương pháp tiếp cận bài toán 17](#_Toc185848200)

[2.6. Tổng kết chương một 21](#_Toc185848201)

[CHƯƠNG 3: MỘT SỐ MÔ HÌNH HỌC SÂU CHO BÀI TOÁN MÔ TẢ HÌNH ẢNH 22](#_Toc185848202)

[3.1. Kiến trúc mô hình học sâu cho nhiệm vụ xử lý hình ảnh 22](#_Toc185848203)

[3.1.1. Mạng nơ-ron nhân tạo (Artificial Neural Network) 22](#_Toc185848204)

[3.1.2. Mạng nơ-ron tích chập nhân tạo (Convolutional Neural Network) 23](#_Toc185848205)

[3.2. Kiến trúc mô hình học sâu cho nhiệm vụ phân tích ngữ nghĩa, ngữ cảnh và tạo mô tả 28](#_Toc185848206)

[3.2.1. Mạng nơ-ron hồi quy (Recurrent Neural Network - RNN) 28](#_Toc185848207)

[3.2.2. Bộ nhớ dài-ngắn hạn (Long-Short Term Memory - LSTM) 36](#_Toc185848208)

[3.2.3. Mô hình Transformer 40](#_Toc185848209)

[3.3. Một số phương pháp đánh giá mô hình mô tả hình ảnh 48](#_Toc185848210)

[3.3.1. Độ đo BLEU (Bilingual Evaluation Understudy) 49](#_Toc185848211)

[3.3.2. Độ đo METEOR (Metric for Evaluation of Translation with Explicit ORdering) 51](#_Toc185848212)

[3.3.3. Độ đo CIDEr (Consensus-based Image Description Evaluation) 53](#_Toc185848213)

[3.4. Tổng kết chương hai 55](#_Toc185848214)

[CHƯƠNG 4: THỰC NGHIỆM VÀ XÂY DỰNG ỨNG DỤNG MÔ TẢ HÌNH ẢNH 57](#_Toc185848215)

[4.1. Dữ liệu thực nghiệm trên các mô hình 57](#_Toc185848216)

[4.2. Phương pháp thực nghiệm và đánh giá các mô hình 60](#_Toc185848217)

[4.3. Xây dựng ứng dụng sinh mô tả tự động cho hình ảnh 62](#_Toc185848218)

[4.4. Tổng kết chương ba 65](#_Toc185848219)

[KẾT LUẬN 67](#_Toc185848220)

[TÀI LIỆU THAM KHẢO 68](#_Toc185848222)

# DANH MỤC CÁC CHỮ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| Từ viết tắt | Tiếng anh | Tiếng Việt |
| API | Application Programming Interface | Giao diện lập trình ứng dụng |
| AI | Artificial Intelligence | Trí tuệ nhân tạo |
| ML | Machine Learning | Học máy |
| OCR | Optical Character Recignition | Nhận dạng ký tự quang học |
| WER | Word Error Rate | Tỷ lệ lỗi từ |
| CER | Character Error Rate | Tỷ lệ lỗi ký tự |
| CNN | Convolutional Neural Network | Mạng nơ-ron tích chập |
| RNN | Recurrent Neural Network | Mạng nơ-ron tuần hoàn |
| LSTM | Long-Short Term Memory | Bộ nhớ ngắn hạn dài hạn |
| RMI | Remote Method Invocation | Gọi phương thức từ xa |
| FFN | Feed Forward Network | Mạng nơ-ron truyền tiếp |
| CTC Loss | Connectionist Temporal Classification loss | Hàm mất mát phân loại thời gian kết nối |
| FCN | Fully Connected Network | Mạng nơ-ron hoàn toàn kết nối |
| GRU | Gated Recurrent Unit | Đơn vị tuần hoàn có cổng |
| LM | Language Model | Mô hình ngôn ngữ |

# DANH MỤC CÁC HÌNH VẼ

[Hình 1.1: Kiến trúc bộ mã hóa-giải mã cơ sở sử dụng bộ mã hóa mạng nơ-ron tích chập được đào tạo trước và bộ giải mã mô hình bộ nhớ dài hạn ngắn 17](#_Toc185713120)

[Hình 1.2: Kiến trúc mô tả hình ảnh sử dung Multi-Attention và Feed Forward để tạo chú thích văn bản. 18](#_Toc185713121)

[Hình 1.3: CPTR - Mạng Transformer hoàn chỉnh cho mô tả hình ảnh 19](#_Toc185713122)

[Hình 1.4: EAMA - Phương pháp tiếp cận dựa trên sự liên kết đa phương thức nhận thức thực thể 20](#_Toc185713123)

[Hình 2.1: Hình ảnh biểu diễn mạng Neural Network 22](#_Toc185713124)

[Hình 2.2: Mạng nơ-ron tích chập 24](#_Toc185713125)

[Hình 2.3: Hình ảnh này minh họa hoạt động của một 25](#_Toc185713126)

[Hình 2.4: Max Pooling trong mạng nơ-ron tích chập (CNN) 26](#_Toc185713127)

[Hình 2.5: Kiến trúc Fully connected (kết nối đầy đủ) và sử dụng hàm kích hoạt ReLU (Rectified Linear Unit) ở lớp ẩn 27](#_Toc185713128)

[Hình 2.6: Một mạng nơ-ron hồi quy và sự diễn ra theo thời gian của quá trình tính toán liên quan đến tính toán chuyển tiếp của nó. 29](#_Toc185713129)

[Hình 2.7: Mô tả cách thức hoạt động của mạng RNN 30](#_Toc185713130)

[Hình 2.8: Kiến trúc One to One 33](#_Toc185713131)

[Hình 2.9: Kiến trúc One to Many 34](#_Toc185713132)

[Hình 2.10: Kiến trúc Many to One 34](#_Toc185713133)

[Hình 2.11: Kiến trúc Many to Many 35](#_Toc185713134)

[Hình 2.12: Cấu trúc của RNN với hàm kích hoạt tanh. 37](#_Toc185713135)

[Hình 2.13: Cấu trúc của LSTM với hàm kích hoạt tanh 37](#_Toc185713136)

[Hình 2.14: Minh họa cấu trúc trong một cell. 39](#_Toc185713137)

[Hình 2.15: Minh họa cấu trúc của GRU. 39](#_Toc185713138)

[Hình 2.16: Miêu tả hoạt động của self-attention 41](#_Toc185713139)

[Hình 2.17: Mô tả nhiệm vụ và chức năng của các vector query, key, value 42](#_Toc185713140)

[Hình 2.18: Mô tả chức năng của MutiHead Attention. 42](#_Toc185713141)

[Hình 2.19: Hình mô tả cấu trúc của mô hình Transformer. 43](#_Toc185713142)

[Hình 2.20: Hình ảnh mô tả sự liên kết giữa decoder và encoder. 44](#_Toc185713143)

[Hình 2.21: Vị trí của positional encoder trong mô hình Transformer. 44](#_Toc185713144)

[Hình 2.22: Hình ảnh vị trí của lớp masked multi head attention. 47](#_Toc185713145)

[Hình 3.1: Quy trình hoạt động của hệ thống mô tả hình ảnh 64](#_Toc185713146)

[Hình 3.2: Giao diện hệ thống tạo mô tả cho ảnh 65](#_Toc185713147)

# MỞ ĐẦU

## Giới thiệu đề tài

Trong thời đại bùng nổ dữ liệu đa phương tiện, đặc biệt là hình ảnh, nhu cầu tự động hóa việc phân tích và hiểu nội dung hình ảnh ngày càng trở nên cấp thiết. Lĩnh vực thị giác máy tính (Computer Vision) đã đạt được những bước tiến vượt bậc nhờ sự phát triển mạnh mẽ của Trí tuệ Nhân tạo (Artificial Intelligence - AI), đặc biệt là học sâu (Deep Learning). Trong đó, bài toán "Tự động mô tả hình ảnh" (Image Captioning) nổi lên như một thách thức quan trọng, thu hút sự quan tâm của đông đảo giới nghiên cứu.

Mô tả hình ảnh tự động là quá trình tạo ra một đoạn văn bản mô tả nội dung của một hình ảnh một cách tự động. Bài toán này không chỉ đơn thuần là nhận diện đối tượng mà còn đòi hỏi khả năng nắm bắt mối quan hệ giữa các đối tượng, bối cảnh, hành động và diễn đạt chúng thành một câu văn hoàn chỉnh, có ngữ nghĩa chặt chẽ. Việc giải quyết bài toán này có ý nghĩa to lớn trong nhiều ứng dụng thực tiễn, bao gồm: hỗ trợ người khiếm thị, tìm kiếm hình ảnh, tạo nội dung tự động, tương tác người-máy,...

Nghiên cứu này tập trung vào việc ứng dụng các kỹ thuật tiên tiến trong lĩnh vực Trí tuệ Nhân tạo, đặc biệt là học sâu, để giải quyết bài toán tự động mô tả hình ảnh. Cụ thể, nghiên cứu sẽ khảo sát các kiến trúc mạng nơ-ron tích chập (Convolutional Neural Networks - CNNs) để trích xuất đặc trưng hình ảnh, mạng nơ-ron hồi quy (Recurrent Neural Networks - RNNs), đặc biệt là mạng LSTM (Long Short-Term Memory) và GRU (Gated Recurrent Unit), cho việc sinh văn bản mô tả. Bên cạnh đó, cơ chế chú ý (Attention Mechanism) cũng sẽ được nghiên cứu và ứng dụng nhằm nâng cao chất lượng của các mô hình.

Mục tiêu của đề tài là xây dựng một mô hình có khả năng tự động tạo ra các mô tả hình ảnh chính xác, đầy đủ thông tin và có tính ngữ nghĩa cao. Kết quả của nghiên cứu dự kiến sẽ đóng góp vào sự phát triển của lĩnh vực thị giác máy tính và mở ra nhiều tiềm năng ứng dụng trong thực tiễn.

## Lý do chọn đề tài

Sự gia tăng theo cấp số nhân của dữ liệu hình ảnh trong kỷ nguyên số đặt ra nhu cầu cấp thiết cho việc phát triển các phương pháp tự động nhằm phân tích và hiểu nội dung hình ảnh. Quá trình xử lý thủ công, phụ thuộc vào con người, trở nên bất khả thi do khối lượng dữ liệu khổng lồ và tính phức tạp của việc diễn giải thông tin hình ảnh. Chính vì vậy, bài toán "Tự động mô tả hình ảnh" đã nổi lên như một hướng nghiên cứu then chốt trong lĩnh vực thị giác máy tính và trí tuệ nhân tạo. Việc phát triển các mô hình có khả năng tự động sinh ra mô tả bằng ngôn ngữ tự nhiên cho hình ảnh không chỉ góp phần giải quyết thách thức về xử lý dữ liệu mà còn mở ra nhiều ứng dụng thực tiễn mang lại giá trị to lớn.

Một lý do quan trọng thúc đẩy việc lựa chọn đề tài này là tiềm năng ứng dụng rộng rãi của công nghệ tự động mô tả hình ảnh. Trong lĩnh vực hỗ trợ người khiếm thị, các mô hình này có thể cung cấp thông tin chi tiết về nội dung hình ảnh, giúp họ tiếp cận và tương tác với thế giới xung quanh một cách dễ dàng hơn. Đối với các hệ thống tìm kiếm hình ảnh, việc tích hợp khả năng mô tả tự động sẽ cho phép người dùng truy vấn bằng ngôn ngữ tự nhiên, nâng cao hiệu quả tìm kiếm và trải nghiệm người dùng. Hơn nữa, công nghệ này còn được ứng dụng trong việc tạo nội dung tự động cho các nền tảng truyền thông xã hội, báo chí, hay thậm chí là trong lĩnh vực robot học, nơi robot cần khả năng hiểu và mô tả môi trường xung quanh.

Bên cạnh đó, bài toán tự động mô tả hình ảnh cũng là một thách thức khoa học thú vị, đòi hỏi sự kết hợp của nhiều kỹ thuật tiên tiến trong lĩnh vực trí tuệ nhân tạo. Khác với các nhiệm vụ đơn giản như phân loại hình ảnh (Image Classification) hay nhận diện đối tượng (Object Detection), mô tả hình ảnh yêu cầu mô hình không chỉ nhận biết các đối tượng mà còn phải nắm bắt được mối quan hệ giữa chúng, ngữ cảnh, hành động và chuyển tải thành một câu văn hoàn chỉnh, có ngữ nghĩa chặt chẽ. Điều này đòi hỏi sự kết hợp giữa khả năng xử lý hình ảnh và xử lý ngôn ngữ tự nhiên (Natural Language Processing), tạo nên một bài toán đa ngành phức tạp nhưng cũng đầy tiềm năng.

Hơn nữa, sự phát triển mạnh mẽ của các mô hình học sâu (Deep Learning) trong những năm gần đây đã mang lại những bước tiến đột phá trong lĩnh vực thị giác máy tính. Các kiến trúc mạng nơ-ron tiên tiến như Mạng nơ-ron tích chập (Convolutional Neural Networks - CNNs), Mạng nơ-ron hồi quy (Recurrent Neural Networks - RNNs), đặc biệt là LSTM (Long Short-Term Memory) và GRU (Gated Recurrent Unit), cùng với cơ chế chú ý (Attention Mechanism) đã cho thấy hiệu quả vượt trội trong việc trích xuất đặc trưng hình ảnh và sinh văn bản. Việc ứng dụng các công nghệ này vào bài toán tự động mô tả hình ảnh mở ra nhiều cơ hội để cải thiện độ chính xác và chất lượng của các mô hình, đồng thời thúc đẩy sự phát triển của lĩnh vực nghiên cứu này.

Tóm lại, với tầm quan trọng về mặt ứng dụng, thách thức khoa học và sự phát triển của các công nghệ tiên tiến, đề tài "Nghiên cứu sử dụng Trí tuệ Nhân tạo tự động mô tả hình ảnh" hứa hẹn mang lại những đóng góp thiết thực cho lĩnh vực thị giác máy tính và có tiềm năng ứng dụng rộng rãi trong thực tiễn.

## Mục tiêu của đề tài

Đề tài: “Nghiên cứu phát triển giải pháp tạo mô tả tự động cho hình ảnh”đáp ứng được những mục tiêu:

Nghiên cứu các mô hình học sâu như CNN (Convolutional Neural Network), RNN (Recurrent Neural Network), LSTM (Long Short-Term Memory), và Transformer để tạo chú thích hình ảnh tự động.

Hiểu rõ cách kết hợp mô hình CNN với RNN/LSTM, trong đó CNN sẽ đảm nhiệm vai trò trích xuất đặc trưng hình ảnh, còn RNN/LSTM được sử dụng để tạo ra chuỗi văn bản mô tả dựa trên các đặc trưng đã trích xuất. Tìm hiểu cách cải thiện hiệu quả bằng cách áp dụng Attention để tập trung vào các vùng quan trọng của hình ảnh trong quá trình tạo chú thích.

Khám phá ứng dụng Transformer, một kiến trúc hiện đại, để thay thế các mô hình truyền thống như LSTM trong việc tạo mô tả, giúp tăng khả năng bắt các mối quan hệ phức tạp trong dữ liệu đầu vào và cải thiện độ chính xác.

Phát triển một hệ thống tự động mô tả hình ảnh dựa trên tập dữ liệu cố định. Ứng dụng sẽ tạo các chú thích tự động cho hình ảnh và cung cấp đánh giá chất lượng mô tả bằng cách sử dụng các chỉ số BLEU, METEOR, CIDEr nhằm so sánh với thực tế.

## Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu: Các phương pháp và kỹ thuật liên quan đến mô tả hình ảnh, bao gồm các phương pháp trích xuất đặc trưng ảnh và thuật toán tạo sinh mô tả.

Phạm vi nghiên cứu:

* Phạm vi của đề tài giới hạn trong việc triển khai và thử nghiệm các mô hình học sâu trên một tập dữ liệu hình ảnh cố định (Flickr30k). Tập dữ liệu này sẽ được sử dụng để huấn luyện, kiểm tra và đánh giá hiệu quả của các mô hình mô tả hình ảnh tự động.
* Đề tài tập trung vào việc xây dựng một ứng dụng thử nghiệm có khả năng tự động mô tả hình ảnh.
* Ngoài ra, phạm vi nghiên cứu cũng bao gồm việc thử nghiệm các phương pháp tối ưu hóa quá trình huấn luyện mô hình, như tinh chỉnh siêu tham số, sử dụng các kỹ thuật tăng cường dữ liệu (data augmentation), và tích hợp các cơ chế như Attention hay Transformer để cải thiện chất lượng mô tả.

## Kết quả dự kiến đạt được

Nghiên cứu lý thuyết:

* Hiểu rõ nguyên lý hoạt động và cơ chế của các mô hình học sâu như CNN, RNN, LSTM và Transformer trong việc trích xuất đặc trưng hình ảnh và sinh văn bản mô tả.
* Nắm vững cách kết hợp giữa CNN và RNN/LSTM trong bài toán mô tả hình ảnh, đồng thời ứng dụng Attention Mechanism để cải thiện khả năng tập trung vào các vùng quan trọng trong hình ảnh.
* Hiểu và triển khai được các phương pháp đánh giá chất lượng chú thích hình ảnh, bao gồm BLEU, METEOR, CIDEr, từ đó đánh giá hiệu quả của mô hình so với thực tế.

Ứng dụng thực tiễn: Phát triển thành công một ứng dụng thử nghiệm có khả năng nhận hình ảnh đầu vào và tạo chú thích văn bản tự động. Ứng dụng sẽ được triển khai trên tập dữ liệu cố định, đảm bảo khả năng sinh chú thích phù hợp với nội dung hình ảnh.

Hiệu quả mô hình:

* Đánh giá được hiệu suất của mô hình qua các chỉ số BLEU, METEOR, CIDEr với kết quả mong muốn đạt ở mức khá (trên 70% độ chính xác theo các chỉ số đánh giá này).
* Xác định các hạn chế còn tồn tại của mô hình như tốc độ xử lý, khả năng mở rộng quy mô, và đề xuất các hướng cải tiến phù hợp.

Tài liệu hóa: Cung cấp báo cáo tổng hợp các kết quả đạt được, kèm theo phân tích sâu về các mô hình và phương pháp đã áp dụng trong đề tài.

## Bố cục của đề tài

Chương 1: Tổng quan về trí tuệ nhân tạo, học máy, học sâu và bài toán mô tả hình ảnh

Chương 2: Một số mô hình học sâu cho bài toán mô tả hình ảnh

Chương 3: Thực nghiệm và xây dựng ứng dụng thử nghiệm mô tả hình ảnh.

Bước đầu đi vào thực tế của em còn hạn chế và gặp nhiều bỡ ngỡ nên không khỏi gặp nhiều thiếu sót, em rất mong nhận được những ý kiến đóng góp quý báu của quý thầy cô để kiến thức của em trong lĩnh vực này được hoàn thiện hơn đồng thời có điều kiện bổ sung, nâng cao kỹ năng nghề nghiệp của em.

Em xin chân thành cảm ơn!

Hà Nội, ngày … tháng … năm 2024

Sinh Viên

Thân Ngọc Thiện

# TỔNG QUAN VỀ TRÍ TUỆ NHÂN TẠO, HỌC MÁY, HỌC SÂU VÀ BÀI TOÁN MÔ TẢ HÌNH ẢNH

## Tổng quan về trí tuệ nhân tạo, học máy và học sâu

### Trí tuệ nhân tạo

Trí tuệ nhân tạo (AI) là một trong những lĩnh vực nghiên cứu nổi bật và cách mạng trong khoa học máy tính, có tầm ảnh hưởng sâu rộng đến nhiều ngành công nghiệp và đời sống xã hội. AI được định nghĩa là khả năng của các hệ thống máy tính trong việc thực hiện các nhiệm vụ mà, nếu không có AI, sẽ đòi hỏi trí tuệ con người. Các nhiệm vụ này bao gồm nhận diện hình ảnh, hiểu ngôn ngữ tự nhiên, suy luận logic, và thậm chí ra quyết định trong các tình huống phức tạp. Mục tiêu của AI là mô phỏng và mở rộng khả năng tư duy và trí tuệ con người, giúp máy tính có thể tự động thực hiện các công việc từ đơn giản đến phức tạp mà không cần sự can thiệp của con người.

AI khác biệt so với các hệ thống tính toán truyền thống ở chỗ thay vì dựa vào những quy tắc đã được lập trình sẵn, AI có thể tự học từ kinh nghiệm và dữ liệu thực tế để cải thiện hiệu suất của mình. Điều này giúp AI thích nghi và điều chỉnh hành vi dựa trên những thay đổi trong môi trường xung quanh, giúp các hệ thống trở nên linh hoạt và hiệu quả hơn theo thời gian.

AI được xây dựng trên nền tảng các thuật toán và mô hình toán học tiên tiến, giúp nó có khả năng phân tích khối lượng dữ liệu lớn và trích xuất thông tin có giá trị từ đó. Đây là lý do tại sao AI có thể đạt được hiệu quả cao trong các bài toán phức tạp như dự đoán xu hướng thị trường, phát hiện gian lận, và hỗ trợ quá trình ra quyết định trong các ngành công nghiệp như y tế, tài chính và năng lượng.

Sự ra đời của AI bắt đầu từ giữa thế kỷ 20, khi những câu hỏi về khả năng của máy móc "suy nghĩ" như con người lần đầu tiên được đặt ra. Một trong những cột mốc đáng chú ý trong lịch sử AI là bài báo nổi tiếng "Computing Machinery and Intelligence" của nhà toán học và triết học Alan Turing vào năm 1950. Trong bài báo này, Turing đề xuất một thí nghiệm – ngày nay được gọi là Thử nghiệm Turing – để đánh giá liệu một máy tính có thể biểu hiện hành vi trí tuệ không thể phân biệt được với con người. Từ đó, khái niệm AI bắt đầu trở thành một lĩnh vực nghiên cứu khoa học và được giới học giả và kỹ sư trên toàn thế giới nghiên cứu và phát triển.

Trong những năm 1950 và 1960, các nhà nghiên cứu AI đã đạt được một số tiến bộ đáng kể với sự ra đời của các chương trình máy tính đầu tiên có khả năng chơi cờ và giải quyết các bài toán logic. Một trong những hệ thống đầu tiên là Logic Theorist do Allen Newell và Herbert Simon phát triển, có khả năng chứng minh các định lý trong logic. Tuy nhiên, vào thời điểm này, do hạn chế về phần cứng và sức mạnh tính toán, AI không thể đạt được những kết quả thực tiễn đáng kể.

Với sự bùng nổ của dữ liệu lớn (Big Data) và sự phát triển vượt bậc của phần cứng máy tính trong những năm 1990 và 2000, AI đã quay trở lại mạnh mẽ hơn bao giờ hết. Các thuật toán mới, đặc biệt là các thuật toán học máy (machine learning), đã được phát triển, cho phép AI xử lý và học từ khối lượng dữ liệu khổng lồ. Các mạng nơ-ron nhân tạo và các kỹ thuật học sâu (deep learning) cũng ra đời, giúp AI không chỉ phân tích dữ liệu hiệu quả hơn mà còn có khả năng mô phỏng nhiều nhiệm vụ phức tạp hơn, như nhận diện giọng nói, dịch ngôn ngữ và điều khiển robot.

Ngày nay, AI được chia thành hai loại chính, phản ánh mức độ và phạm vi ứng dụng của nó: AI hẹp (Narrow AI) và AI tổng quát (General AI):

* AI hẹp là dạng AI phổ biến nhất hiện nay, tập trung vào việc thực hiện một hoặc một số nhiệm vụ cụ thể. Ví dụ, các hệ thống nhận diện khuôn mặt, phân loại văn bản, hoặc hệ thống chơi cờ như AlphaGo của Google DeepMind là những ví dụ điển hình của AI hẹp. Hệ thống AI hẹp có thể làm rất tốt các công việc được chỉ định, nhưng không có khả năng vượt qua giới hạn nhiệm vụ đó hoặc tự chuyển đổi sang các nhiệm vụ khác mà không được lập trình trước.
* AI tổng quát, còn được gọi là Trí tuệ nhân tạo mạnh (Strong AI), là mục tiêu dài hạn của lĩnh vực này. AI tổng quát có khả năng hiểu biết và thực hiện bất kỳ nhiệm vụ nào mà con người có thể làm, từ việc giải quyết các vấn đề tri thức đến tư duy sáng tạo. Tuy nhiên, việc phát triển AI tổng quát vẫn còn là một thách thức lớn. Để đạt được AI tổng quát, các nhà nghiên cứu không chỉ cần mô phỏng hành vi của trí tuệ con người mà còn phải tìm hiểu sâu hơn về cách thức não bộ con người học hỏi, suy nghĩ, và ra quyết định.

Trí tuệ nhân tạo đã trải qua một hành trình phát triển dài và đầy biến động, từ những ngày đầu với kỳ vọng lớn lao, qua những thất bại ban đầu, cho đến sự bùng nổ của công nghệ hiện đại. Ngày nay, AI đang đóng một vai trò then chốt trong cách mạng công nghệ, mang lại tiềm năng to lớn trong việc giải quyết các vấn đề phức tạp và cải thiện chất lượng cuộc sống. Tuy nhiên, để AI phát triển một cách bền vững và an toàn, cần có sự hợp tác chặt chẽ giữa các nhà khoa học, chính phủ và cộng đồng quốc tế nhằm đảm bảo rằng công nghệ này phục vụ lợi ích của toàn nhân loại.

### Học máy

Học máy (Machine Learning - ML) là một trong những nhánh quan trọng và đang phát triển nhanh chóng của trí tuệ nhân tạo (AI), với mục tiêu giúp các hệ thống máy tính có khả năng học hỏi và thích nghi từ dữ liệu một cách tự động mà không cần phải lập trình cụ thể cho mỗi tác vụ. Điều này đánh dấu một bước đột phá lớn trong lĩnh vực công nghệ, khi các thuật toán học máy có thể giải quyết những vấn đề phức tạp bằng cách tự xây dựng mô hình dựa trên dữ liệu đầu vào thay vì tuân theo các quy tắc lập trình cứng nhắc. Theo một nghiên cứu được công bố trên Journal of Machine Learning Research (2020), học máy đã thay đổi cục diện công nghệ nhờ khả năng xử lý lượng lớn dữ liệu và tìm ra các mối quan hệ tiềm ẩn trong dữ liệu một cách hiệu quả, từ đó cung cấp các dự đoán chính xác hơn.

Một trong những đặc điểm quan trọng của học máy là khả năng tự động cập nhật và cải tiến dựa trên dữ liệu mới mà không cần sự can thiệp liên tục của con người. Nhờ vào tính tự động này, học máy đã giúp các hệ thống máy tính trở nên linh hoạt và thích nghi tốt hơn trong môi trường thay đổi liên tục. Một nghiên cứu của Nature Machine Intelligence (2021) cho thấy, khi học máy được ứng dụng trong các hệ thống dự đoán tài chính hoặc chẩn đoán y tế, các mô hình học từ dữ liệu lịch sử và liên tục cải thiện hiệu suất dự đoán khi được cung cấp thêm dữ liệu mới. Điều này giúp các hệ thống học máy không chỉ đưa ra các dự đoán chính xác mà còn có khả năng phát hiện và khắc phục các sai số theo thời gian, cải thiện hiệu quả hoạt động.

Học máy được chia thành ba phương pháp chính: Học có giám sát (Supervised Learning), Học không giám sát (Unsupervised Learning), và Học tăng cường (Reinforcement Learning), mỗi phương pháp có cách thức hoạt động và ứng dụng riêng biệt trong các lĩnh vực khác nhau:

* Học có giám sát (Supervised Learning) là phương pháp mà hệ thống học từ các dữ liệu đã được dán nhãn rõ ràng. Ví dụ, trong bài toán nhận diện hình ảnh, hệ thống sẽ được cung cấp một tập dữ liệu bao gồm các hình ảnh đã được dán nhãn (chẳng hạn như hình ảnh mèo, chó) và học cách phân loại các hình ảnh dựa trên thông tin này. Theo một bài báo đăng trên IEEE Transactions on Pattern Analysis and Machine Intelligence (2020), học có giám sát đã trở thành phương pháp phổ biến nhất trong nhận diện hình ảnh và xử lý ngôn ngữ tự nhiên, nhờ khả năng đạt được độ chính xác cao khi có lượng dữ liệu lớn và phong phú. Tuy nhiên, hạn chế của phương pháp này là cần có một lượng lớn dữ liệu đã dán nhãn, điều này đôi khi có thể tốn kém và mất thời gian.
* Học không giám sát (Unsupervised Learning) lại tiếp cận vấn đề từ một góc độ khác, khi dữ liệu không được dán nhãn và hệ thống tự tìm cách phát hiện các mẫu và cấu trúc tiềm ẩn trong dữ liệu. Một ứng dụng phổ biến của học không giám sát là trong phân cụm dữ liệu (data clustering), nơi các thuật toán như K-means hoặc mô hình Gaussian Mixture được sử dụng để nhóm các điểm dữ liệu có cùng đặc trưng lại với nhau. Trong một nghiên cứu của Proceedings of the National Academy of Sciences (2020), học không giám sát được chứng minh là có hiệu quả trong việc phân tích dữ liệu y sinh và phát hiện các phân nhóm bệnh nhân có chung đặc điểm sinh lý mà không cần đến thông tin dán nhãn trước.
* Học tăng cường (Reinforcement Learning) là phương pháp mà hệ thống học thông qua tương tác với môi trường, nhận phản hồi dưới dạng phần thưởng hoặc hình phạt, từ đó điều chỉnh hành vi để tối ưu hóa kết quả. Học tăng cường đã trở thành nền tảng cho nhiều tiến bộ trong robot tự hành và các trò chơi, nổi bật nhất là sự thành công của AlphaGo, hệ thống AI đã đánh bại nhà vô địch thế giới trong trò chơi cờ vây. Theo một nghiên cứu của Science (2019), học tăng cường được coi là một trong những phương pháp mạnh mẽ nhất trong việc giải quyết các bài toán phức tạp, nơi mà môi trường và nhiệm vụ không thể mô tả đầy đủ bằng các thuật toán đơn giản.

Học máy hiện đang đóng vai trò nền tảng trong nhiều lĩnh vực công nghệ hiện đại và mang lại những lợi ích to lớn cho xã hội. Ứng dụng của học máy trải dài từ việc phát triển các hệ thống trợ lý ảo, như Siri và Google Assistant, đến các hệ thống khuyến nghị trong thương mại điện tử, như Amazon và Netflix, giúp cải thiện trải nghiệm người dùng bằng cách dự đoán và cung cấp các sản phẩm, nội dung phù hợp với sở thích cá nhân. Một báo cáo từ MIT Technology Review (2022) đã chỉ ra rằng, nhờ vào học máy, các công ty có thể tận dụng dữ liệu người dùng để cải thiện chiến lược kinh doanh, tối ưu hóa quy trình vận hành, và nâng cao hiệu quả trong việc tiếp cận khách hàng.

### Học sâu

Học sâu (Deep Learning) là một nhánh tiến bộ của học máy, đặc trưng bởi việc sử dụng các mạng nơ-ron nhân tạo sâu (deep neural networks) để xử lý và phân tích khối lượng dữ liệu khổng lồ, phức tạp. Những mạng nơ-ron này được lấy cảm hứng từ cấu trúc nơ-ron trong não bộ con người, bao gồm nhiều lớp kết nối với nhau, trong đó mỗi lớp thực hiện một nhiệm vụ cụ thể, chẳng hạn như trích xuất đặc trưng từ dữ liệu đầu vào và biến chúng thành các mẫu có ý nghĩa cho các lớp tiếp theo. Mạng nơ-ron sâu cho phép học sâu phân tích dữ liệu theo cách thức mà các hệ thống học máy truyền thống không thể làm được. Các nghiên cứu đã chỉ ra rằng một trong những đặc điểm mạnh mẽ nhất của học sâu là khả năng tự động trích xuất các đặc trưng từ dữ liệu thô, giúp giảm bớt sự can thiệp của con người trong quá trình xây dựng mô hình và đạt hiệu suất vượt trội trong các bài toán phức tạp như xử lý hình ảnh và ngôn ngữ tự nhiên (Nature Machine Intelligence, 2021). Điều này đã làm cho học sâu trở thành một công cụ quan trọng trong việc phát triển các công nghệ tiên tiến như xe tự lái, nơi đòi hỏi sự nhận diện và ra quyết định trong thời gian thực dựa trên dữ liệu phức tạp từ môi trường xung quanh.

Sự khác biệt nổi bật giữa học sâu và các phương pháp học máy truyền thống chủ yếu nằm ở cách thức xử lý dữ liệu và khả năng trích xuất đặc trưng tự động. Trong khi học máy truyền thống sử dụng các thuật toán như cây quyết định (decision trees) hoặc hồi quy tuyến tính (linear regression) để phân loại và dự đoán dựa trên dữ liệu đã được làm sạch và chuẩn hóa trước, học sâu có khả năng học trực tiếp từ dữ liệu thô mà không cần quá trình tiền xử lý phức tạp. Theo một nghiên cứu của IEEE Transactions on Neural Networks and Learning Systems (2020), mạng nơ-ron sâu đã chứng minh khả năng vượt trội trong các bài toán mà ngữ cảnh và các mối quan hệ phức tạp cần được nắm bắt, chẳng hạn như nhận diện hình ảnh hoặc phân tích văn bản tự do. Việc có thể xử lý và phân tích lượng lớn dữ liệu phi cấu trúc đã giúp học sâu đạt được độ chính xác cao hơn, đồng thời làm giảm phụ thuộc vào các chuyên gia trong quá trình chuẩn bị dữ liệu, một điểm mạnh mà các phương pháp truyền thống còn hạn chế. Chính vì lý do này, học sâu đã được ứng dụng rộng rãi trong các hệ thống trí tuệ nhân tạo tổng hợp (AGI), nơi khả năng học tự động và ra quyết định thông minh là yếu tố then chốt.

Ngoài ra, các nghiên cứu cũng chỉ ra rằng khả năng của học sâu trong việc học từ dữ liệu lớn đã góp phần vào sự phát triển của các công nghệ khuyến nghị cá nhân hóa, hệ thống nhận diện giọng nói và các công cụ phân tích y tế. Ví dụ, trong lĩnh vực y tế, các mô hình học sâu đã được ứng dụng để phân tích hình ảnh y khoa và hỗ trợ chẩn đoán các bệnh phức tạp, giúp tăng độ chính xác và giảm thiểu sai sót trong quá trình chẩn đoán, nhờ vào khả năng xử lý khối lượng dữ liệu hình ảnh y khoa lớn mà con người không thể xử lý hiệu quả trong thời gian ngắn (The Lancet Digital Health, 2021). Những tiềm năng vượt trội này đã khiến học sâu trở thành một công cụ không thể thiếu trong cuộc cách mạng công nghệ hiện nay, mở ra nhiều ứng dụng mới và tiềm năng trong tương lai.

## Định nghĩa bài toán mô tả hình ảnh

Image captioning là một bài toán trong trí tuệ nhân tạo, kết hợp hai lĩnh vực trọng yếu là thị giác máy tính và xử lý ngôn ngữ tự, với mục tiêu tạo ra các mô tả ngôn ngữ tự nhiên chính xác cho một hình ảnh đầu vào. Quá trình này yêu cầu không chỉ nhận diện các đối tượng có trong hình ảnh (như con người, động vật, đồ vật) mà còn phải nắm bắt ngữ cảnh không gian, mối quan hệ giữa các đối tượng, và ý nghĩa tổng thể để tạo ra câu mô tả mạch lạc, có ý nghĩa..

Để giải quyết bài toán này, mô hình cần thực hiện ba nhiệm vụ chính:

* Phân tích hình ảnh: Trích xuất các đặc trưng trực quan quan trọng từ hình ảnh, chẳng hạn nhận diện đối tượng (object detection), phát hiện hành động (action recognition), và hiểu được ngữ cảnh.
* Hiểu ngữ nghĩa và ngữ cảnh: Xác định mối quan hệ giữa các đối tượng và hành động trong không gian ba chiều hoặc tình huống cụ thể, chẳng hạn phân biệt một con chó đang chạy với một con chó đang ngủ.
* Tạo câu mô tả: Kết hợp các thông tin đã phân tích để tạo ra câu mô tả tự nhiên, chính xác về ngữ pháp và hợp ngữ nghĩa.

Mặc dù đã có những tiến bộ, mô tả hình ảnh vẫn phải đối mặt với nhiều thách thức:

* Chất lượng và sự đa dạng của dữ liệu: Hiệu quả của các mô hình mô tả hình ảnh phụ thuộc rất nhiều vào chất lượng và sự đa dạng của tập dữ liệu huấn luyện. Nhiều tập dữ liệu hiện có bị giới hạn về kích thước hoặc phạm vi, điều này có thể cản trở hiệu suất của mô hình trong các tình huống thực tế.
* Độ phức tạp của hình ảnh thực tế: Hình ảnh thực tế thường chứa các cảnh phức tạp với nhiều đối tượng và ngữ cảnh khác nhau, gây khó khăn cho các mô hình trong việc tạo ra chú thích chính xác. Sự phức tạp này có thể dẫn đến sự mơ hồ và cách hiểu mang tính chủ quan trong văn bản được tạo ra.
* Hiểu biết đa phương thức: Việc tích hợp thông tin từ cả hai phương thức hình ảnh và văn bản đặt ra một thách thức đáng kể. Mối quan hệ giữa hình ảnh và chú thích tương ứng của chúng không phải lúc nào cũng rõ ràng, vì một hình ảnh có thể có nhiều mô tả hợp lệ và ngược lại
* Khả năng khái quát hóa đối với dữ liệu chưa từng gặp: Các mô hình phải đủ mạnh mẽ để xử lý dữ liệu chưa từng gặp, điều này rất quan trọng cho việc triển khai chúng trong các ứng dụng thực tế như hệ thống truy xuất hình ảnh hoặc hỗ trợ người khiếm thị.

## Ứng dụng của bài toán mô tả hình ảnh

Mô tả hình ảnh không chỉ là một bước tiến trong lĩnh vực trí tuệ nhân tạo mà còn mang lại những ứng dụng thiết thực trong đời sống và công nghệ. Với khả năng chuyển đổi hình ảnh thành ngôn ngữ tự nhiên, công nghệ này đang dần thay đổi cách con người tương tác với thế giới số, từ hỗ trợ người khuyết tật, tối ưu hóa công cụ tìm kiếm, đến cải tiến trải nghiệm trên mạng xã hội và thương mại điện tử.

Mô tả hình ảnh hỗ trợ người khiếm thị bằng cách cung cấp mô tả bằng âm thanh về môi trường xung quanh, giúp họ nhận biết đồ vật, con người và hành động trong không gian thực tế. Công nghệ này giúp tăng cường tìm kiếm đa phương tiện, cho phép người dùng tìm hình ảnh qua từ khóa. Trong thương mại điện tử, được sử dụng để tự động tạo mô tả sản phẩm từ hình ảnh, hỗ trợ phân loại và cải thiện trải nghiệm mua sắm. Tự động tạo chú thích cho hình ảnh cũng được ứng dụng phổ biến, trên các nền tảng như Instagram hoặc Facebook giúp tăng tương tác và tiết kiệm thời gian. Mô tả hình ảnh được ứng dụng trong y tế để mô tả hình ảnh X-quang, hoặc trong an ninh để phát hiện và cảnh báo sự kiện. Hỗ trợ tạo nội dung học tập với mô tả tự động, như chú thích hình ảnh trong sách giáo khoa hoặc tài liệu nghiên cứu, giúp học sinh hiểu nội dung trực quan hơn.

Ứng dụng của mô tả hình ảnh rất đa dạng và linh hoạt, từ cải thiện trải nghiệm cá nhân, hỗ trợ người dùng đặc biệt, đến tự động hóa và tối ưu hóa các hoạt động thương mại và giáo dục. Những tiến bộ trong bài toán này không chỉ nâng cao chất lượng cuộc sống mà còn mở ra những cơ hội mới cho các ngành công nghiệp

## Các dạng triển khai chính của bài toán mô tả hình ảnh

Mỗi dạng triển khai không chỉ thể hiện sự phát triển của công nghệ trí tuệ nhân tạo mà còn mang lại giá trị thiết thực trong nhiều lĩnh vực như tìm kiếm, y tế, giám sát và giải trí. Trong đó, ba dạng chính bao gồm mô tả đơn ảnh, mô tả ảnh chi tiết và mô tả liên tục, mỗi dạng đều có những đặc điểm và ứng dụng nổi bật riêng. Trong đó, ba dạng chính bao gồm mô tả đơn ảnh, mô tả ảnh chi tiết và mô tả liên tục, mỗi dạng đều có những đặc điểm và ứng dụng nổi bật riêng.

Mô tả đơn ảnh là dạng cơ bản và phổ biến nhất của bài toán mô tả hình ảnh, trong đó mục tiêu là tạo ra một câu mô tả chính xác và tự nhiên cho một hình ảnh tĩnh duy nhất. Mô hình thường sử dụng các kỹ thuật học sâu, kết hợp giữa mạng nơ-ron tích chập (CNN) để trích xuất đặc trưng từ hình ảnh và mạng nơ-ron hồi tiếp (RNN) hoặc mô hình Transformer để sinh văn bản. Ứng dụng chính của dạng này bao gồm hỗ trợ tìm kiếm hình ảnh thông qua mô tả văn bản, tự động gắn chú thích cho hình ảnh trên mạng xã hội hoặc cung cấp thông tin trực quan trong thương mại điện tử. Tuy nhiên, thách thức lớn nhất của mô tả đơn ảnh là nắm bắt đầy đủ ngữ cảnh và mối quan hệ giữa các đối tượng, đặc biệt trong các hình ảnh phức tạp hoặc chứa nhiều chi tiết.

Mô tả ảnh chi tiết là một phiên bản nâng cao của bài toán mô tả hình ảnh, trong đó mục tiêu là không chỉ tạo ra một câu mô tả chung cho hình ảnh mà còn cung cấp các chú thích chi tiết cho từng đối tượng hoặc khu vực cụ thể trong hình. Kỹ thuật này yêu cầu mô hình kết hợp nhận diện đối tượng (object detection) với khả năng tạo văn bản, thường thông qua các mô hình như Faster R-CNN tích hợp với Transformer. Ứng dụng nổi bật của mô tả ảnh chi tiết bao gồm phân tích dữ liệu y tế, chẳng hạn như mô tả tổn thương trên ảnh X-quang, hoặc gắn nhãn các đối tượng cụ thể trong hình ảnh vệ tinh. Thách thức chính nằm ở việc duy trì sự chính xác và mạch lạc khi tạo nhiều chú thích đồng thời, đặc biệt với hình ảnh có ngữ cảnh phức tạp.

Mô tả liên tục mở rộng khả năng của mô tả hình ảnh sang dữ liệu video, trong đó mô hình cần tạo ra các câu mô tả liên tục và mạch lạc dựa trên thông tin không gian và thời gian trong chuỗi hình ảnh. Điều này yêu cầu các mô hình học sâu phải tích hợp cả mạng CNN để xử lý khung hình riêng lẻ và mạng LSTM hoặc Transformer để nắm bắt các mối quan hệ thời gian giữa các khung hình. Ứng dụng chính của mô tả liên tục bao gồm tạo phụ đề tự động cho video, phân tích hành động trong hệ thống giám sát an ninh hoặc hỗ trợ người khiếm thị hiểu nội dung video. Thách thức lớn nhất của dạng này là đảm bảo mô tả đồng bộ với sự thay đổi nhanh chóng trong video, đồng thời duy trì tính chính xác và hợp ngữ cảnh.

Các dạng triển khai của bài toán mô tả hình ảnh đáp ứng nhiều nhu cầu đa dạng trong thực tế, từ các ứng dụng cơ bản như tạo chú thích đơn giản đến các hệ thống phức tạp như phân tích video hoặc mô tả đa ngôn ngữ. Sự khác biệt trong cách triển khai phản ánh khả năng ứng dụng linh hoạt của image captioning, đồng thời cũng đặt ra các thách thức lớn trong việc phát triển và tối ưu hóa mô hình.

## Các phương pháp tiếp cận bài toán

Có nhiều phương pháp tiếp cận khác nhau để giải quyết bài toán mô tả hình ảnh, tùy thuộc vào các yêu cầu và điều kiện cụ thể của ứng dụng.

* Phương pháp trích xuất đặc trưng truyền thống

Sử dụng các mô hình học sâu như mạng nơ-ron tích chập (CNN) để trích xuất đặc trưng từ hình ảnh, sau đó kết hợp với các mô hình ngôn ngữ như LSTM hoặc GRU để tạo ra câu mô tả. Các đặc trưng hình ảnh được biểu diễn dưới dạng vector và làm đầu vào cho mạng sinh văn bản. Đây là cách tiếp cận cơ bản nhất, phù hợp với các ứng dụng yêu cầu sự đơn giản trong thiết kế và triển khai. Tuy nhiên, phương pháp này gặp khó khăn trong việc nắm bắt các mối quan hệ phức tạp giữa các đối tượng hoặc chi tiết trong hình ảnh, đặc biệt khi dữ liệu đầu vào có ngữ cảnh phức tạp.

A diagram of a computer system

Description automatically generated

Hình 1.1: Kiến trúc bộ mã hóa-giải mã cơ sở sử dụng bộ mã hóa mạng nơ-ron tích chập được đào tạo trước và bộ giải mã mô hình bộ nhớ dài hạn ngắn

* Phương pháp dựa trên mô hình chú ý (Attention-based Approach)

Phương pháp dựa trên chú ý cải tiến cách tiếp cận truyền thống bằng cách cho phép mô hình tập trung vào các khu vực quan trọng trong hình ảnh khi tạo câu mô tả. Các cơ chế chú ý như Soft Attention và Hard Attention được sử dụng để điều chỉnh trọng số tập trung vào từng vùng tương ứng với mỗi từ trong câu. Điều này giúp mô hình nắm bắt tốt hơn ngữ cảnh và các chi tiết quan trọng trong hình ảnh. Các ứng dụng như mô tả hình ảnh y tế hoặc phân tích hình ảnh vệ tinh hưởng lợi lớn từ cách tiếp cận này. Tuy nhiên, việc triển khai đòi hỏi tài nguyên tính toán cao và cần thiết kế mô hình phức tạp hơn.

A diagram of a person standing in a field

Description automatically generated

Hình 1.2: Kiến trúc mô tả hình ảnh sử dung Multi-Attention và Feed Forward để tạo chú thích văn bản.

* Phương pháp dựa trên Transformer

Phương pháp này tận dụng kiến trúc Transformer, sử dụng Vision Transformer (ViT) để trích xuất đặc trưng hình ảnh và Text Transformer để sinh văn bản. Đây là bước phát triển vượt bậc khi loại bỏ sự phụ thuộc vào mạng hồi tiếp (RNN), cải thiện hiệu quả và tốc độ huấn luyện. Các mô hình hiện đại như CLIP hoặc BLIP ứng dụng Transformer đã đạt được hiệu suất vượt trội trong nhiều tác vụ. Ưu điểm nổi bật là khả năng xử lý đồng thời các đặc trưng không gian và ngữ cảnh văn bản, phù hợp với các ứng dụng đòi hỏi hiệu suất cao. Tuy nhiên, phương pháp này yêu cầu dữ liệu lớn và nguồn lực tính toán mạnh mẽ.

A diagram of a dog

Description automatically generated

Hình 1.3: CPTR - Mạng Transformer hoàn chỉnh cho mô tả hình ảnh

* Phương pháp học tăng cường (Reinforcement Learning-based Approach)

Học tăng cường được áp dụng để tối ưu hóa các chỉ số đánh giá như BLEU, CIDEr, vốn khó đạt được với các phương pháp học có giám sát truyền thống. Các thuật toán như REINFORCE được sử dụng để tinh chỉnh mô hình, cải thiện chất lượng câu mô tả thông qua phản hồi từ môi trường huấn luyện. Phương pháp này phù hợp với các ứng dụng đòi hỏi độ chính xác ngữ nghĩa cao hoặc các kịch bản đặc thù. Tuy nhiên, việc triển khai học tăng cường thường phức tạp và dễ gặp vấn đề mất ổn định trong quá trình huấn luyện.

* Phương pháp học đa phương thức (Multimodal Learning-based Approach)

Phương pháp học đa phương thức tích hợp dữ liệu từ nhiều nguồn như hình ảnh, văn bản và âm thanh để tạo ra các mô tả ngữ nghĩa chính xác hơn. Các mô hình đa phương thức như CLIP liên kết không gian biểu diễn của hình ảnh và ngôn ngữ, giúp cải thiện khả năng hiểu và mô tả ngữ cảnh phức tạp. Điều này đặc biệt hữu ích trong các ứng dụng như tìm kiếm hình ảnh bằng truy vấn văn bản hoặc phân tích hình ảnh y tế. Tuy nhiên, phương pháp này đòi hỏi tập dữ liệu phong phú và nguồn lực tính toán lớn để đạt hiệu quả tối ưu.

A person standing in front of a diagram

Description automatically generated

Hình 1.4: EAMA - Phương pháp tiếp cận dựa trên sự liên kết đa phương thức nhận thức thực thể

Các phương pháp tiếp cận bài toán mô tả hình ảnh không ngừng phát triển, từ các phương pháp truyền thống đến các kiến trúc tiên tiến như Transformer và mô hình đa phương thức. Mỗi phương pháp có ưu điểm và thách thức riêng, phù hợp với các mục tiêu ứng dụng khác nhau, đồng thời thể hiện tiềm năng lớn trong việc cải thiện khả năng hiểu và mô tả thế giới trực quan của máy móc.

## Tổng kết chương một

Chương đầu tiên đã giới thiệu tổng quan về trí tuệ nhân tạo (AI), học máy (ML), và học sâu (DL), từ đó làm nền tảng cho bài toán mô tả hình ảnh. Chương giải thích khái niệm cơ bản của AI, ML và DL, trong đó học sâu đóng vai trò quan trọng trong việc xử lý dữ liệu phức tạp như hình ảnh. Chương cũng định nghĩa bài toán mô tả hình ảnh, các ứng dụng thực tế của nó, và các dạng triển khai chính, bao gồm các mô hình học sâu sử dụng ngữ nghĩa và ngữ cảnh để tạo ra các mô tả hình ảnh tự động. Cuối cùng, chương phân tích các phương pháp tiếp cận khác nhau trong nghiên cứu mô tả hình ảnh, từ đó cung cấp cơ sở lý thuyết cho nghiên cứu sâu hơn trong lĩnh vực này.

# MỘT SỐ MÔ HÌNH HỌC SÂU CHO BÀI TOÁN MÔ TẢ HÌNH ẢNH

Học sâu (Deep Learning) đã tạo ra những đột phá trong lĩnh vực trí tạo nói chung và bài toán mô tả hình ảnh nói riêng nhờ khả năng học tự động từ dữ liệu lớn, kết hợp thông tin từ hình ảnh và ngôn ngữ để tạo ra mô tả chính xác, tự nhiên. Chương này trình bày một số mô hình học sâu tiêu biểu như CNN, RNN, và Transformer, phân tích cơ chế, ưu nhược điểm và những cải tiến gần đây nhằm nâng cao hiệu quả trong lĩnh vực này…

## Kiến trúc mô hình học sâu cho nhiệm vụ xử lý hình ảnh

### Mạng nơ-ron nhân tạo (Artificial Neural Network)

Mạng nơ-ron nhân tạo (Neural Network) là một mô hình tính toán được lấy cảm hứng từ cách hoạt động của hệ thần kinh sinh học. Nó là một loại mô hình học máy có khả năng tự động học và áp dụng kiến thức từ dữ liệu.

A diagram of a network

Description automatically generated

Hình 2.1: Hình ảnh biểu diễn mạng Neural Network

Mạng nơ-ron nhân tạo là một tập hợp các đơn vị tính toán (nơ-ron) được kết nối với nhau theo cấu trúc phân lớp. Mỗi nơ-ron nhận tín hiệu đầu vào, xử lý thông tin và truyền tín hiệu đầu ra đến các nơ-ron khác. Quá trình này được thực hiện thông qua các liên kết trọng số, cho phép mạng học cách điều chỉnh trọng số để tạo ra các đầu ra phù hợp với đầu vào.

Mạng nơ-ron nhân tạo có thể có nhiều lớp bao gồm: lớp đầu vào, lớp ẩn và lớp đầu ra. Những lớp ẩn này giúp mạng học các mức độ biểu diễn trừu tượng và phức tạp hơn từ dữ liệu đầu vào. Mạng nơ-ron hoạt động dựa trên các trọng số (weights) và hàm kích hoạt (activation function). Khi dữ liệu được đưa vào, mỗi nơ-ron tính toán đầu ra dựa trên đầu vào và trọng số tương ứng. Nếu đầu ra đạt ngưỡng nhất định, nơ-ron sẽ "kích hoạt" và truyền tín hiệu đến nơ-ron tiếp theo. Quá trình này diễn ra theo chiều từ lớp đầu vào đến lớp đầu ra, thường được gọi là mạng nơ-ron truyền thẳng (feedforward).

Quá trình huấn luyện mạng nơ-ron nhân tạo thường bao gồm việc cung cấp dữ liệu huấn luyện đã được gán nhãn và điều chỉnh các trọng số của mạng để giảm thiểu sai số giữa đầu ra dự đoán và đầu ra thực tế. Quá trình này được thực hiện thông qua thuật toán backpropagation, trong đó đạo hàm của hàm mất mát được tính toán và sử dụng để cập nhật trọng số.

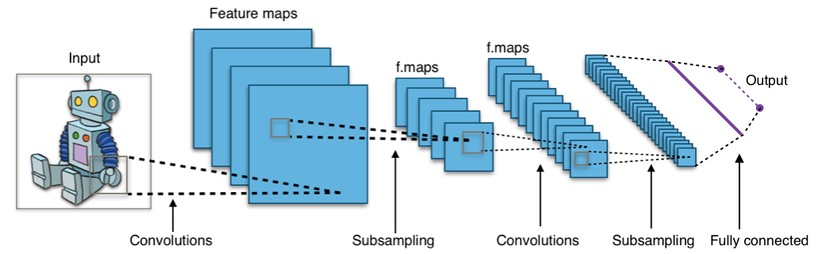
Mạng Neural network đã có những ứng dụng rất rộng trong lĩnh vực nhận dạng hình ảnh, xử lý ngôn ngữ tự nhiên, dự báo, điều khiển tự động, và nhiều lĩnh vực khác.

### Mạng nơ-ron tích chập nhân tạo (Convolutional Neural Network)

Mạng CNN (Convolutional Neural Network) là một loại mạng nơ-ron nhân tạo được sử dụng rộng rãi trong lĩnh vực xử lý ảnh và thị giác máy tính. Mạng này được thiết kế cho việc tự động học và phân tích các đặc trưng từ dữ liệu hình ảnh một cách hiệu quả. Cấu trúc của mạng nơ-ron tích chập được lấy cảm hứng từ cách mà thị giác của con người hoạt động. Thành phần của mạng nơ-ron tích chập bao gồm các lớp tích chập (convolutional layers) và các lớp tổng hợp (pooling layers), kết hợp với các lớp kết nối đầy đủ (fully connected layers) ở cuối để thực hiện phân loại hoặc dự đoán.

CNN phân loại hình ảnh bằng cách lấy một hình ảnh đầu vào, xử lý và phân loại nó theo các hạng mục nhất định (Ví dụ: Chó, Mèo, Hổ, ...). Máy tính coi hình ảnh đầu vào là 1 mảng pixel và nó phụ thuộc vào độ phân giải của hình ảnh. Dựa trên độ phân giải hình ảnh, máy tính sẽ thấy H x W x D (H: Chiều cao, W: Chiều rộng, D: Độ dày).

Về kỹ thuật, mô hình CNN để training và kiểm tra, mỗi hình ảnh đầu vào sẽ chuyển nó qua 1 loạt các lớp tích chập với các bộ lọc (Kernals), tổng hợp lại các lớp được kết nối đầy đủ (Full Connected) và áp dụng hàm Softmax để phân loại đối tượng có giá trị xác suất giữa 0 và 1.



Hình 2.2: Mạng nơ-ron tích chập

Mạng CNN được cấu tạo bởi nhiều tầng, mỗi tầng thực hiện một chức năng cụ thể để trích xuất các đặc trưng quan trọng từ dữ liệu đầu vào.

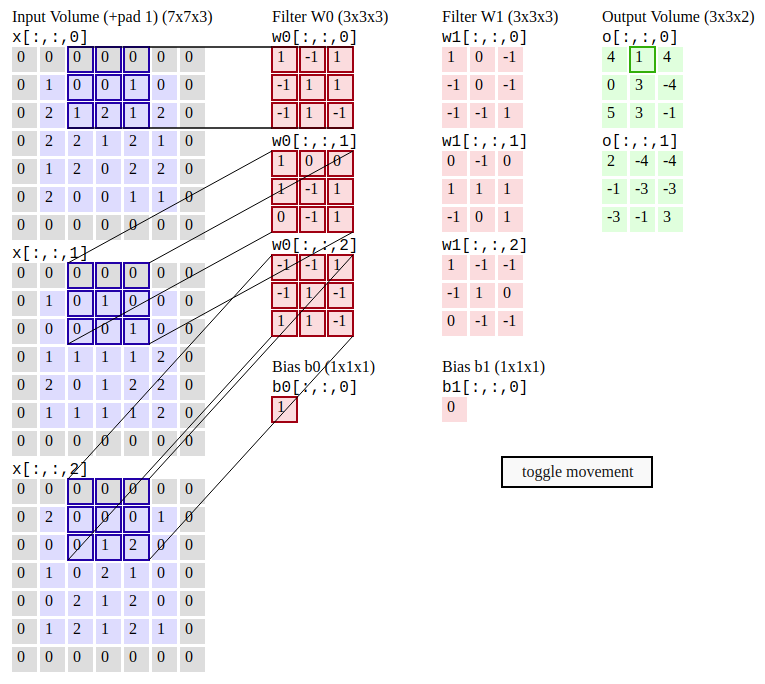
* Convolution (lớp tích chập) là lớp đầu tiên trích xuất các đặc tính từ hình ảnh. Tham số lớp này bao gồm một tập hợp các bộ lọc có thể học được. Các bộ lọc đều nhỏ thường có kích cỡ hai chiều đầu tiên khoảng 3x3 hoặc 5x5, .... và có độ sâu bằng với độ sâu của đầu vào đầu vào. Bằng cách trượt dần bộ lọc theo chiều ngang và dọc trên ảnh, chúng thu được một Feature Map chứa các đặc trưng được trích xuất từ trên hình ảnh đầu vào. Quá trình trượt các bộ lọc thường có các giá trị được quy định bao gồm:
* Bước nhảy – Stride: Stride là số pixel thay đổi trên ma trận đầu vào. Khi stride là 1 thì ta di chuyển các kernel 1 pixel. Khi stride là 2 thì ta di chuyển các kernel đi 2 pixel và tiếp tục như vậy. Hình dưới là lớp tích chập hoạt động với stride là 2.
* Đường viền – Padding: Đôi khi kernel không phù hợp với hình ảnh đầu vào. Ta có 2 lựa chọn:
* Chèn thêm các số 0 vào 4 đường biên của hình ảnh (padding).
* Cắt bớt hình ảnh tại những điểm không phù hợp với kernel.
* Hàm phi tuyến – ReLU:

ReLU viết tắt của Rectified Linear Unit, là 1 hàm phi tuyến. Với đầu ra là:

ƒ (x) = max (0, x).

ReLU giới thiệu tính phi tuyến trong ConvNet. Vì dữ liệu trong thế giới mà chúng ta tìm hiểu là các giá trị tuyến tính không âm.

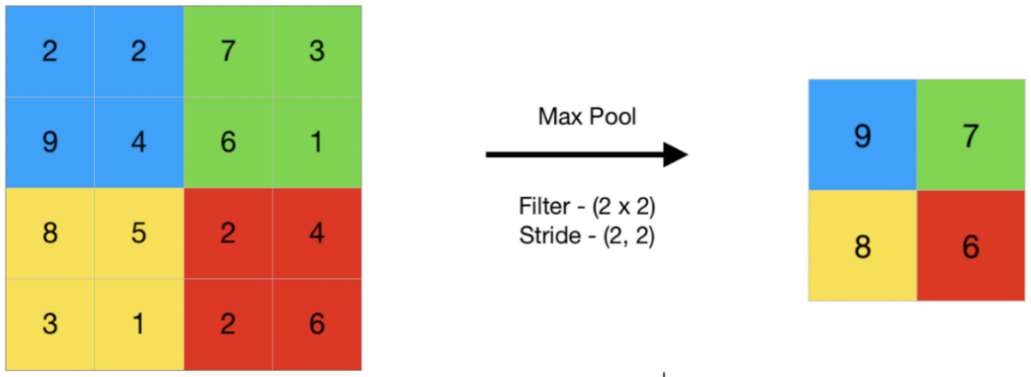
Có 1 số hàm phi tuyến khác như tanh, sigmoid cũng có thể được sử dụng thay cho ReLU. Hầu hết người ta thường dùng ReLU vì nó có hiệu suất tốt.



Hình 2.3: Hình ảnh này minh họa hoạt động của một

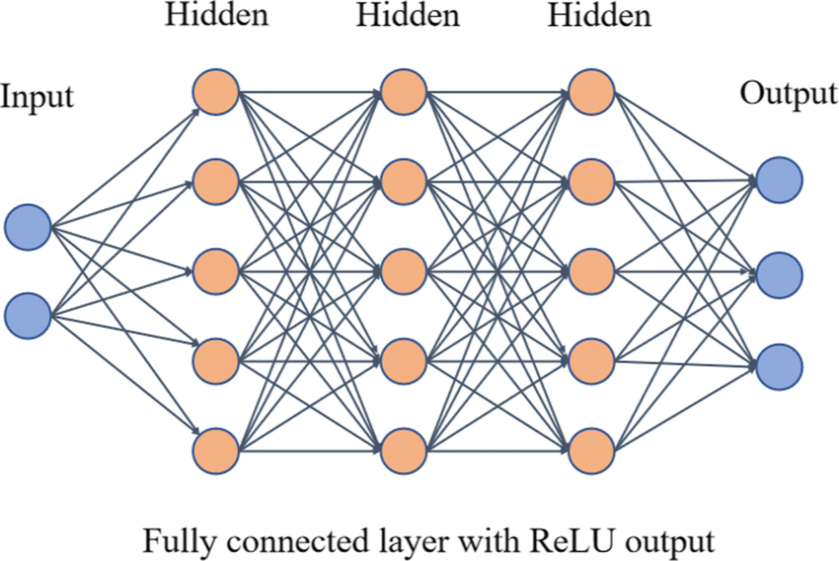
Với mỗi kernel khác nhau ta sẽ học được những đặc trưng khác nhau của ảnh, nên trong mỗi convolutional layer ta sẽ dùng nhiều kernel để học được nhiều thuộc tính của ảnh. Vì mỗi kernel cho ra output là 1 matrix nên k kernel sẽ cho ra k output matrix. Ta kết hợp k output matrix này lại thành 1 tensor 3 chiều có chiều sâu k. Output của convolutional layer sẽ qua hàm activation function trước khi trở thành input của convolutional layer tiếp theo.

* Pooling layer với mục tiêu của lớp gộp là lấy các tính năng quan trọng nhất từ ma trận phức tạp. Điều này được thực hiện bằng cách áp dụng một số thao tác tổng hợp, giúp giảm kích thước của bản đồ đặc trưng (ma trận phức tạp), do đó giảm bộ nhớ được sử dụng trong khi huấn luyện mạng. Pooling cũng có liên quan để giảm thiểu tình trạng trang bị quá mức.



Hình 2.4: Max Pooling trong mạng nơ-ron tích chập (CNN)

* Các hàm tổng hợp phổ biến nhất có thể được áp dụng là:
* Max pooling: là một loại phép gộp được sử dụng trong mạng nơ-ron tích chập (CNN) và các kiến trúc liên quan. Nó thường được áp dụng sau lớp tích chập để giảm kích thước của đầu vào và tạo ra các phiên bản đã tạo ra của đặc trưng cục bộ.
* Sum pooling: là một phương pháp trong xử lý ảnh và mạng học sâu, được sử dụng để giảm kích thước của feature maps trong quá trình trích xuất thông tin từ hình ảnh. Tương tự như Average Pooling và Max Pooling, Sum Pooling cũng là một loại phép gộp thông tin trong mạng neural.
* Average pooling: là một phép tích chập thường được sử dụng trong mạng học sâu và xử lý ảnh. Nó là một phần quan trọng của các kiến trúc mạng thần kinh như Convolutional Neural Networks (CNNs) để giảm kích thước của đặc trưng (feature maps) trong quá trình trích xuất thông tin từ hình ảnh.
* Fully connected layers là lớp cuối cùng trong mạng CNN có chức năng thực hiện việc phân loại hoặc dự đoán các giá trị đầu ra của mô hình. Convolutional layer và pooling layer đã thực hiện các phép tích chập và giảm kích thước của dữ liệu đầu vào, tạo ra các đặc trưng cục bộ cho ảnh. Tuy nhiên thì chúng chỉ tạo ra các đặc trưng cục bộ và không liên kết các đặc trưng đó để đưa ra một dự đoán cuối cùng. Để phân loại chính xác hơn, chúng ta cần kết nối tất cả các đặc trưng đó với nhau và tính toán xác suất cho các lớp đầu ra. Fully connected layer trong mạng CNN được thiết kế có một số nơ-ron kết nối với tất cả các nơ-ron trong lớp trước đó (có thể là một lớp convolutional layer hoặc pooling layer) để nhận đầu vào là tất cả các đặc trưng từ lớp trước đó. Fully connected layer lúc này có nhiệm vụ kết nối các đặc trưng cục bộ đã được trích xuất từ các lớp convolutional layer và pooling layer để đưa ra một dự đoán cuối cùng.



Hình 2.5: Kiến trúc Fully connected (kết nối đầy đủ) và sử dụng hàm kích hoạt ReLU (Rectified Linear Unit) ở lớp ẩn

Fully connected layer giúp mạng CNN học các mối quan hệ phức tạp giữa các đặc trưng đầu vào và tạo ra các phân loại chính xác hơn. Tuy nhiên khi sử dụng quá nhiều fully connected layer trong một mạng CNN sẽ dẫn đến hiện tượng overfitting, khi mô hình chỉ học nhớ dữ liệu huấn luyện mà không có khả năng dự đoán đúng trên các dữ liệu mới. Do đó số lượng fully connected layer cần được cân nhắc để tối ưu hóa hiệu suất của mạng CNN.

## Kiến trúc mô hình học sâu cho nhiệm vụ phân tích ngữ nghĩa, ngữ cảnh và tạo mô tả

Đầu vào cho mô hình tạo mô tả là một chuỗi các vector đặc trưng ảnh đã được trích xuất từ mô hình xử lý ảnh. Vector đặc trưng này đại diện cho một biểu diễn cô đọng các đặc điểm thị giác quan trọng của ảnh đầu vào. Bên cạnh vector đặc trưng ảnh này, mô hình tạo mô tả còn nhận thêm một trạng thái ẩn khởi đầu, thường được khởi tạo bằng các giá trị không hoặc thông qua một quá trình học riêng. Trong quá trình sinh mô tả theo từng bước, đầu vào còn bao gồm chuỗi các từ đã được tạo ra trước đó, tạo thành một chuỗi thông tin đầy đủ để mô hình dự đoán từ tiếp theo trong câu mô tả.

### Mạng nơ-ron hồi quy (Recurrent Neural Network - RNN)

Mạng nơ-ron hồi quy (RNN) là một mô hình học sâu được đào tạo để xử lý và chuyển đổi đầu vào dữ liệu tuần tự thành đầu ra dữ liệu tuần tự cụ thể. Dữ liệu tuần tự là dữ liệu, chẳng hạn như từ, câu hoặc dữ liệu chuỗi thời gian, trong đó các thành phần tuần tự tương quan với nhau dựa trên ngữ nghĩa phức tạp và quy tắc cú pháp. RNN là một hệ thống phần mềm gồm nhiều thành phần được kết nối với nhau theo cách con người thực hiện chuyển đổi dữ liệu tuần tự, chẳng hạn như dịch văn bản từ ngôn ngữ này sang ngôn ngữ khác. Phần lớn RNN đang được thay thế bằng trí tuệ nhân tạo (AI) dựa trên công cụ biến đổi và các mô hình ngôn ngữ lớn (LLM), hiệu quả hơn nhiều trong việc xử lý dữ liệu tuần tự.

Ý tưởng chính của RNN (Recurrent Neural Network) là sử dụng chuỗi các thông tin. Trong các mạng nơ-ron truyền thống tất cả các đầu vào và cả đầu ra là độc lập với nhau. Tức là chúng không liên kết thành chuỗi với nhau. Nhưng các mô hình này không phù hợp trong rất nhiều bài toán. Ví dụ, nếu muốn đoán từ tiếp theo có thể xuất hiện trong một câu thì ta cũng cần biết các từ trước đó xuất hiện lần lượt thế nào? RNN được gọi là hồi quy (Recurrent) bởi lẽ chúng thực hiện cùng một tác vụ cho tất cả các phần tử của một chuỗi với đầu ra phụ thuộc vào cả các phép tính trước đó. Nói cách khác, RNN có khả năng nhớ các thông tin được tính toán trước đó. Trên lý thuyết, RNN có thể sử dụng được thông tin của một văn bản rất dài, tuy nhiên thực tế thì nó chỉ có thể nhớ được một vài bước trước đó (ta cùng bàn cụ thể vấn đề này sau) mà thôi. Về cơ bản một mạng RNN có dạng như sau:

A diagram of a diagram of a diagram

Description automatically generated

Hình 2.6: Một mạng nơ-ron hồi quy và sự diễn ra theo thời gian của quá trình tính toán liên quan đến tính toán chuyển tiếp của nó.

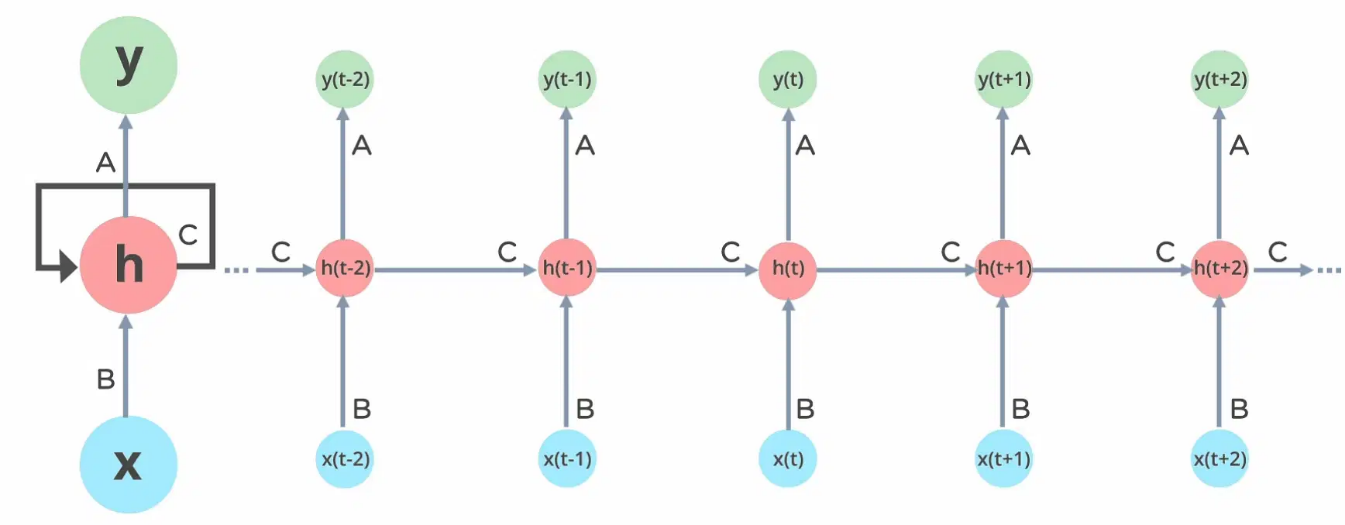
Cấu trúc cơ bản của mạng nơ-ron hồi quy gồm:

* Đầu Vào (Input): Tại mỗi bước thời gian t, RNN nhận một đầu vào . Đầu vào này có thể là một vector hoặc một giá trị đơn lẻ, tùy thuộc vào bài toán.
* Trạng Thái Ẩn (Hidden State): RNN duy trì một trạng thái ẩn  tại mỗi bước thời gian. Trạng thái ẩn này lưu trữ thông tin từ tất cả các bước trước đó và được sử dụng để tính toán trạng thái ẩn ở bước tiếp theo.
* Đầu Ra (Output): Tại mỗi bước thời gian t, RNN có thể tạo ra một đầu ra . Đầu ra này có thể được sử dụng để dự đoán hoặc đưa ra quyết định.

Cách thức hoạt động của mạng nơ-ron hồi quy

Giống như các mạng nơ-ron truyền thống, chẳng hạn như mạng nơ-ron truyền thẳng (Feedforward Neural Networks - FNN) và mạng nơ-ron tích chập (Convolutional Neural Networks - CNNs), mạng nơ-ron hồi quy sử dụng dữ liệu huấn luyện để học. Tuy nhiên, điểm khác biệt chính của RNN so với các mạng nơ-ron truyền thống là khả năng "ghi nhớ" thông tin từ các đầu vào trước đó để ảnh hưởng đến đầu vào và đầu ra hiện tại.

Sự phụ thuộc của RNN vào thông tin từ các bước trước. Trong khi các mạng nơ-ron truyền thống giả định rằng các đầu vào và đầu ra là độc lập với nhau, đầu ra của RNN lại phụ thuộc vào các yếu tố trước đó trong chuỗi. Điều này có nghĩa là, các sự kiện trong tương lai có thể hữu ích trong việc xác định đầu ra của một chuỗi, nhưng các mạng nơ-ron hồi quy một chiều (unidirectional RNNs) không thể tính toán được các sự kiện này trong dự đoán của chúng.



Hình 2.7: Mô tả cách thức hoạt động của mạng RNN

Ví dụ: Với thành ngữ "feeling under the weather", được sử dụng khi ai đó cảm thấy bị ốm. Để thành ngữ này có nghĩa, các từ cần phải xuất hiện theo một thứ tự cụ thể. Vì vậy, RNN cần phải lưu ý đến vị trí của mỗi từ trong thành ngữ này, và sử dụng thông tin đó để dự đoán từ tiếp theo trong chuỗi.

Mỗi từ trong thành ngữ "feeling under the weather" là một phần của chuỗi, trong đó thứ tự các từ là rất quan trọng. RNN theo dõi ngữ cảnh bằng cách duy trì một trạng thái ẩn tại mỗi bước thời gian. Một vòng lặp phản hồi được tạo ra khi truyền trạng thái ẩn từ một bước thời gian này sang bước thời gian kế tiếp. Trạng thái ẩn đóng vai trò là bộ nhớ, lưu trữ thông tin về các đầu vào trước đó. Tại mỗi bước thời gian, RNN xử lý đầu vào hiện tại (ví dụ, một từ trong câu) cùng với trạng thái ẩn từ bước thời gian trước đó. Điều này giúp RNN "nhớ" các điểm dữ liệu trước đó và sử dụng thông tin đó để ảnh hưởng đến đầu ra hiện tại.

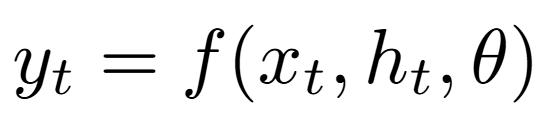
Đặc điểm khác biệt của RNN

* Chia sẻ tham số giữa các lớp: Một đặc điểm phân biệt của mạng nơ-ron hồi quy là việc chia sẻ tham số giữa các lớp của mạng. Trong khi các mạng nơ-ron truyền thẳng có các trọng số khác nhau giữa các nút, các mạng nơ-ron hồi quy chia sẻ cùng một tham số trọng số trong mỗi lớp của mạng. Tuy nhiên, các trọng số này vẫn được điều chỉnh thông qua quá trình lan truyền ngược (backpropagation) và giảm dần theo độ dốc (gradient descent) để hỗ trợ học tăng cường (reinforcement learning).
* Quá trình lan truyền ngược qua thời gian (BPTT): Mạng nơ-ron hồi quy sử dụng thuật toán lan truyền ngược qua thời gian (Backpropagation Through Time - BPTT) để tính toán các đạo hàm (gradients). Điều này khác với thuật toán lan truyền ngược truyền thống vì nó đặc thù cho dữ liệu chuỗi. Nguyên lý của BPTT vẫn giống như lan truyền ngược truyền thống, nơi mô hình tự huấn luyện bằng cách tính toán sai số từ lớp đầu ra đến lớp đầu vào. Quá trình này giúp điều chỉnh và tối ưu các tham số của mô hình.

Tuy nhiên, BPTT khác với lan truyền ngược truyền thống ở chỗ nó cộng dồn sai số tại mỗi bước thời gian, trong khi các mạng nơ-ron truyền thẳng không cần phải cộng dồn sai số vì chúng không chia sẻ tham số giữa các lớp.

Mạng nơ-ron hồi quy có thể được mô tả bằng các công thức toán học sau:

* Đầu và và đầu ra:



Trong đó:

* ​ là đầu vào tại bước thời gian t,
* ​ là trạng thái ẩn tại bước thời gian t,
* ​ ​ là đầu ra tại bước thời gian t,
* *θ* là các tham số của mô hình.
* Cập nhật trạng thái ẩn:

A black text on a white background

Description automatically generated

Trong đó:

* g là hàm kích hoạt (activation function), thường là tanh hoặc ReLU,
* Trạng thái ẩn ​​ được tính toán từ đầu vào ​​ và trạng thái ẩn ​​ tại bước thời gian t.

Quá trình này được lặp lại qua từng bước thời gian để tạo ra chuỗi đầu ra ​​ từ chuỗi đầu vào ​.

Các hàm kích hoạt phổ biến

Hàm kích hoạt là một hàm toán học được áp dụng lên đầu ra của mỗi lớp neuron trong mạng, nhằm đưa vào tính phi tuyến và cho phép mạng học được các mẫu phức tạp hơn trong dữ liệu. Nếu không có các hàm kích hoạt, mạng RNN (Mạng Nơ-ron Hồi tiếp) chỉ có thể thực hiện các phép biến đổi tuyến tính của đầu vào, khiến nó không thể xử lý các vấn đề phi tuyến. Tính phi tuyến là yếu tố quan trọng để học và mô hình hóa các mẫu phức tạp, đặc biệt trong các nhiệm vụ như xử lý ngôn ngữ tự nhiên (NLP), phân tích chuỗi thời gian, và dự đoán dữ liệu tuần tự.

Hàm kích hoạt kiểm soát biên độ đầu ra của neuron, giữ giá trị trong một phạm vi xác định (ví dụ, từ 0 đến 1 hoặc từ -1 đến 1). Điều này giúp ngăn các giá trị trở nên quá lớn hoặc quá nhỏ trong quá trình lan truyền xuôi (forward pass) và lan truyền ngược (backward pass). Trong RNN, hàm kích hoạt được áp dụng tại mỗi bước thời gian lên các trạng thái ẩn, giúp kiểm soát cách mạng cập nhật bộ nhớ bên trong (trạng thái ẩn) dựa trên đầu vào hiện tại và các trạng thái ẩn trước đó.

Các hàm kích hoạt phổ biến bao gồm:

* Hàm Sigmoid: Được sử dụng để diễn giải đầu ra dưới dạng xác suất hoặc để kiểm soát các cổng (gates) quyết định lượng thông tin cần giữ lại hoặc bỏ qua. Tuy nhiên, hàm sigmoid dễ gặp vấn đề về *độ dốc biến mất* (vanishing gradient), làm giảm hiệu quả đối với các mạng sâu.
* Hàm Tanh (Hyperbolic Tangent): Thường được sử dụng vì nó cho đầu ra có giá trị tập trung quanh 0, giúp cải thiện dòng gradient và học các phụ thuộc dài hạn một cách dễ dàng hơn.
* Hàm ReLU (Rectified Linear Unit): Có thể gây ra vấn đề về *độ dốc bùng nổ* (exploding gradient) do tính chất không bị giới hạn của nó. Tuy nhiên, các biến thể như Leaky ReLU và Parametric ReLU đã được phát triển để giảm thiểu một số vấn đề này.

Các loại kiến trúc RNN

Kiến trúc One to One, còn được gọi là Vanilla Neural Network, là dạng mạng nơ-ron đơn giản nhất, trong đó một đầu vào tương ứng với một đầu ra duy nhất. Loại mạng này không xử lý dữ liệu tuần tự mà thực hiện các phép biến đổi toán học trực tiếp trên dữ liệu tĩnh. Đặc điểm của kiến trúc này là không có khả năng ghi nhớ hay liên kết giữa các bước thời gian, do đó phù hợp với các bài toán không yêu cầu tính liên tục của dữ liệu. Ứng dụng phổ biến bao gồm phân loại hình ảnh, như xác định một bức ảnh thuộc về lớp nào (ví dụ: mèo hoặc chó), và hồi quy, như dự đoán giá trị cụ thể từ dữ liệu đầu vào.

A diagram of a single output

Description automatically generated

Hình 2.8: Kiến trúc One to One

Kiến trúc One to Many được thiết kế để xử lý một đầu vào duy nhất và tạo ra một chuỗi các đầu ra. Đây là loại mạng đặc biệt hữu ích khi đầu ra là một chuỗi dữ liệu liên tục, chẳng hạn như văn bản hoặc tín hiệu. Một ví dụ minh họa là mô tả hình ảnh (image captioning), nơi một hình ảnh (đầu vào) được mô tả thành một câu hoặc đoạn văn (đầu ra). Kiến trúc này thường được sử dụng trong các bài toán tạo dữ liệu tuần tự, ví dụ như tạo nhạc hoặc phát sinh văn bản từ một tín hiệu ban đầu.

A diagram of multiple output

Description automatically generated

Hình 2.9: Kiến trúc One to Many

Trong kiến trúc Many to One, mạng nhận một chuỗi dữ liệu đầu vào và tạo ra một đầu ra duy nhất. Loại mạng này phù hợp với các bài toán đòi hỏi đầu ra phản ánh toàn bộ ý nghĩa hoặc tính chất tổng hợp của chuỗi đầu vào. Một ví dụ phổ biến là phân tích cảm xúc (sentiment analysis), nơi một đoạn văn bản được phân loại thành cảm xúc tích cực hoặc tiêu cực. Ngoài ra, kiến trúc này còn được áp dụng trong nhận diện giọng nói, giúp xác định ý nghĩa của tín hiệu âm thanh, và phân loại chuỗi thời gian, nhằm phát hiện xu hướng hoặc các sự kiện trong dữ liệu cảm biến.

A diagram of multiple input output

Description automatically generated

Hình 2.10: Kiến trúc Many to One

Kiến trúc Many to Many được thiết kế để xử lý một chuỗi dữ liệu đầu vào và tạo ra một chuỗi dữ liệu đầu ra, có thể có độ dài bằng hoặc khác nhau. Một ví dụ điển hình của kiến trúc này là dịch máy (machine translation), trong đó một câu ở ngôn ngữ nguồn được dịch sang ngôn ngữ đích. Có hai dạng chính của Many to Many:

* Sequence to Sequence: Chiều dài của chuỗi đầu vào và đầu ra khác nhau, ví dụ dịch ngôn ngữ hoặc tóm tắt văn bản.
* Sequence to Sequence Đồng thời: Chiều dài của chuỗi đầu vào và đầu ra là tương đương, ví dụ như gán nhãn thực thể trong văn bản.  
  Ứng dụng của kiến trúc này rất đa dạng, từ phân tích chuỗi gene trong sinh học đến nhận diện hoạt động (activity recognition) dựa trên dữ liệu cảm biến.

A diagram of multiple output

Description automatically generated

Hình 2.11: Kiến trúc Many to Many

Tóm lại, kiến trúc One to One phù hợp với các bài toán đơn giản không cần xử lý tuần tự. One to Many, Many to One, và Many to Many là các kiến trúc đặc trưng của RNN, mạnh mẽ trong xử lý dữ liệu tuần tự và dữ liệu có mối liên hệ thời gian. Khi làm việc với chuỗi dài, các vấn đề như độ dốc biến mất hoặc độ dốc bùng nổ cần được xử lý thông qua các biến thể như LSTM hoặc GRU.

### Bộ nhớ dài-ngắn hạn (Long-Short Term Memory - LSTM)

LSTM là một dạng đặc biệt của RNN, ó có khả năng học các thông tin ở xa. Về cơ bản thì LSTM và RNN không khác nhau là mấy nhưng LSTM có cải tiến một số phép tính trong 1 hidden state và nó đã hiểu quả. Cấu trúc của LSTM không khác gì RNN, những sự cải tiến ở đây nằm ở phần tính toán trong từng hidden state như sau: Thay vì chỉ có một tầng mạng nơ-ron, LSTM thiết kế với 4 tầng mạng nơ-ron tương tác với nhau một các rất đặc biệt.

Với kiến trúc đặc biệt, LSTM có khả năng giữ và quản lý thông tin trong một khoảng thời gian dài, cho phép nó xử lý các chuỗi dữ liệu dài và xử lý các phụ thuộc xa. Điều này khác biệt so với mạng RNN truyền thống, có khả năng mất mát thông tin trong quá trình lan truyền ngược.

Mạng LSTM sử dụng các đơn vị LSTM, trong đó mỗi đơn vị có các thành phần chính sau đây:

* Cổng quên (Forget gate): Xác định xem thông tin nào nên được giữ lại và thông tin nào nên được bỏ qua từ quá khứ.
* Cổng đầu vào (Input gate): Quyết định thông tin nào nên được cập nhật từ đầu vào mới.
* Cổng ra (Output gate): Quyết định thông tin nào nên được truyền tiếp làm đầu ra của đơn vị LSTM.
* Ô tình trạng (Cell state): Lưu trữ và truyền thông tin dọc theo chuỗi thời gian, giúp duy trì thông tin quan trọng trong quá trình xử lý chuỗi.

Các cổng trong LSTM được điều chỉnh bởi hàm kích hoạt (ví dụ: hàm sigmoid) để quyết định thông tin nào nên được đi qua và thông tin nào nên bị lọc hoặc bỏ qua. Mạng LSTM đã chứng minh hiệu quả của nó trong nhiều ứng dụng như xử lý ngôn ngữ tự nhiên, dịch máy, nhận dạng giọng nói, dự báo chuỗi thời gian và nhiều lĩnh vực khác đòi hỏi xử lý dữ liệu có tính thời gian hoặc phụ thuộc dài hạn. Dưới đây là 2 hình ảnh biểu diễn sự khác nhau giữa RNN và LSTM.

A screenshot of a computer

Description automatically generated

Hình 2.12: Cấu trúc của RNN với hàm kích hoạt tanh.

A diagram of a cell phone

Description automatically generated

Hình 2.13: Cấu trúc của LSTM với hàm kích hoạt tanh

Chìa khóa để giúp LSTM có thể truyền tải thông tin giữa các hidden state một cách xuyên suốt chính là cell state.

Trong đó:

* Forget gate : Γ𝑓
* Đầu ra là hàm *sigmoid* chứa các giá trị từ 0 đến 1.
* Nếu forget gate có giá trị bằng 0, LSTM sẽ "quên" trạng thái được lưu trữ trong đơn vị tương ứng của trạng thái cell trước đó.
* Nếu cổng quên có giá trị bằng 1, LSTM sẽ chủ yếu ghi nhớ giá trị tương ứng ở trạng thái được lưu trữ.

|  |  |
| --- | --- |
| A black text on a white background  Description automatically generated | (1) |

* Candidate value: Chứa thông tin có thể lưu trữ từ step hiện tại

A black text on a white background

Description automatically generated

* Update gate: Quyết định xem phần thông tin nào của và được thêm vào

A close-up of a mathematical equation

Description automatically generated

* Cell state: Là bộ nhớ trong của LSTM. Cell state như 1 băng tải truyền các thông tin cần thiết xuyết suất cả quá trình, qua các nút mạng và chỉ tương tác tuyển tính 1 chút. Vì vậy thông tin có thể truyền đi thông suốt mà không bị thay đổi.

A black text on a white background

Description automatically generated

* Output gate Γ𝑜: Cổng điều chỉnh lượng thông tin đầu ra của cell hiện tại và lượng thông tin truyền tới trạng thái tiếp theo.

A black text on a white background

Description automatically generated

* Hidden state : Được sử dụng để xác định ba cổng của time step tiếp theo.

A close-up of a number

Description automatically generated

* Prediction: Dự đoán trong trường hợp sử dụng này là phân loại, vì vậy bạn sẽ sử dụng softmax.

A black text on a white background

Description automatically generated

A diagram of a diagram

Description automatically generated

Hình 2.14: Minh họa cấu trúc trong một cell.

Nhìn chung LSTM khá giống với RNN hay nói cách khác thì RNN là một dạng đặc biệt của LSTM. LSTM giải quyết phần nào của Vanishing gradient so với RNN, nhưng chỉ một phần.

Cuối cùng, một biến thể khá nổi tiếng của LSTM, đó chính là Gated Recurrent Units hay là GRU. Kiến trúc của GRU kết hợp với cổng quên, cổng đầu vào một cổng cập nhật duy nhất. Ngoài ra, GRU cũng kết hợp các cell state và hidden state để đưa ra các thay đổi khác.

A diagram of a mathematical equation

Description automatically generated

Hình 2.15: Minh họa cấu trúc của GRU.

Ngược lại, với lượng tính toán như trên thì RNN đã chậm rồi, LSTM còn chậm hơn nữa:

* RNN và LSTM vẫn bị sẩy ra trường hợp vanishing gradient và thậm chí còn mất khá nhiều thời gian. Bên cạnh đó việc nó không tương thích với dữ liệu có cấu trúc khiến cho việc mô hình bị mất đi cơ chế học được sự liên quan giữa các từ trong câu với nhau.
* Vì vậy, attention đã ra đời, để tăng cường thông tin giữa các hidden state, thể hiện sự tương quan giữa các hidden state với nhau, với việc loại bỏ hoàn toàn tính tuần tự, cải thiện việc song song hóa, tăng tốc độ xử lý. Mục đích cuối cùng là để tính toán mối tương quan giữa input và output.

### Mô hình Transformer

#### Cơ chế Attention trong mô hình Transformer

Với cơ chế Attention, Transformer sẽ không cần tới các phép toán convolutional và recurrence, với cơ chế xử lý song song rõ ràng cho thấy được cơ chế này cung cấp một hiệu suất vượt trội và thời gian huấn luyện được tối ưu đang kể. Việc ở các mạng trước đó, khi sử dụng với chuỗi đầu vào là một chuỗi dài, sẽ khiến cho việc tìm hiểu mối tương quan giữa các từ trở nên khó khăn hơn rất nhiều.

Trong Transformer vấn đề này đã được giải quyết nhưng sẽ phải giảm effective resolution do lấy trung bình các attention-weighted position, những vấn để này cũng sẽ được giải quyết ở phần sau với Multi-head attention.

Trong bài toán, đối với trường hợp cụ thể như ta cần mã hóa (encode) một chuỗi đầu vào thành một vector ngữ cảnh rồi sau đó giải mã (decode) dự đoán các từ trước đó. Nhưng chúng ta sẽ gặp một số vấn đề như sau :

* Vanishing gradient: mất mát đạo hàm ở những hidden state cuối khi đầu vào là một chuỗi dài như đoạn văn
* Exploding gradient: Hiện tượng gradient tăng mất kiểm soát do dồn ở những lớp cuối và đặc biệt xảy ra với những câu dài như đoạn văn.
* Nén bộ nhớ (Memory compression): Do việc embed chuỗi đầu vào thành vector cố định kích thước, qua nhiều lần thử nghiệm cho thấy được mô hình này học với các câu dài rất kém trong khi đó lại rất lãng phí bộ nhớ với những câu ngắn. Vấn đề này vẫn còn xuất hiện trong LSTM và GRU.

Không phù hợp với dạng dữ liệu có mối quan hệ liên kết ngữ nghĩa: Ví dụ với câu: “She is doing homework”. Nhận thấy “homework” có mối quan hệ nhiều hơn với “doing”. Đối với RNN - học tuần tự từ trái sang phải nhưng không có cơ chế để mô hình học được sự liên quan giữa các từ. Attention sẽ giúp chúng ta thực hiện được việc đó.

A diagram of a software program

Description automatically generated

Hình 2.16: Miêu tả hoạt động của self-attention

Còn self-attention có thể hiểu là attention trong 1 câu, khi từng thành phần trong câu sẽ tương tác với nhau. Từng token sẽ quan sát các tokens còn lại, thu thập nhữ cảnh của câu và cật nhập vào vector biểu diễn. Để xây dựng cơ chế self-attention ta cần chú ý đến hoạt động của vector biểu diễn cho mỗi từ lần lượt là:

* Query: được sử dụng để đặt câu hỏi từ này ở trong câu có ý nghĩa gì.
* Key: trả lời từ hiện tại có liên quan thế nào đến từ mà query đang hỏi.
* Value: chứa thông tin của từ đang hỏi.

A black text on a white background

Description automatically generated

Dựa vào ý nghĩa của *q, k, v* có thể giải thích được công thức của Transformers Attention:

* *q.k* qua hàm softmax để đưa ra xác suất của từ liên quan nhất với từng từ để hỏi tương ứng.
* Sau đó nhân với *v* để đưa ra giá trị dựa vào sự tương quan đó.

A diagram of a machine

Description automatically generated

Hình 2.17: Mô tả nhiệm vụ và chức năng của các vector query, key, value

Query được sử dụng khi một token quan sát những tokens còn lại, nó sẽ tìm kiếm thông tin xung quanh để hiểu được ngữ cảnh và mối quan hệ của nó với tokens còn lại. Key sẽ phản hồi yêu cầu của Query và được sử dụng để tính trọng số attention. Cuối cùng, Value được sử dụng trong số attention vừa rồi để tính ra vector đại diện (attention vector).

Mỗi quá trình như vậy được gọi là 1 head của attention. Khi lặp lại quá trình này nhiều lần (trong bài báo là 3 heads) ta sẽ thu được quá trình Multi-head Attention như biến đổi bên dưới:

A diagram of a multi-head diagram

Description automatically generated

Hình 2.18: Mô tả chức năng của MutiHead Attention.

A close-up of words

Description automatically generated

Để trả về output có cùng kích thước với ma trận input chúng ta chỉ cần nhân với ma trận W để chiều rộng bằng với chiều rộng của ma trận input.

#### Cấu trúc của mô hình Transformer

Tiếp theo ta sẽ đi tìm hiểu về các thành phần chính cấu tạo thành transformer. Đây là cấu trúc của transformer được giới thiệu trong bài báo “Attention is all you need”. Mô hình thực hiện chính xác những gì đã được giới thiệu ở trên.

Ở bên trái là encoder, thông thường có Nx = 6 layers chồng lên nhau. Mỗi layer sẽ có multi-head attention như đã tìm hiểu và khối feed-forward. Ngoài ra còn các các kết nối residual giống như trong mạng Resnet. Ở bên phải là decoder, tương tự cũng có Nx = 6 layers chồng lên nhau.

Kiến trúc thì khá giống encoder những chỉ có thêm khối masked multi-head attention ở vị trí đầu tiên. Ta sẽ tìm hiểu sâu hơn các thành phần trong transformer.

A diagram of a diagram

Description automatically generated

Hình 2.19: Hình mô tả cấu trúc của mô hình Transformer.

Với kiến trúc của Transformer có 2 phần chính là encoder và decoder:

* Encoders: gồm bộ lớp encoder giống hệt nhau và xếp chồng lên nhau. Mỗi lớp có 2 khối chính đó là: attention block và feed-forward block, feed-forward network được kết nối đầy đủ theo vị trí ( positionwide fully connected feed-forward network ).
* Decoders: Mỗi khối decoder nhận được các features từ encoder. Chúng ta có thể hình dung được sự liên kết giữa decoder và encoder

A diagram of a process

Description automatically generated

Hình 2.20: Hình ảnh mô tả sự liên kết giữa decoder và encoder.

Positional Encoding:

Diagram of a diagram of a program

Description automatically generated

Hình 2.21: Vị trí của positional encoder trong mô hình Transformer.

Khác với việc học tuần tự của RNN, Transformer loại bỏ cơ chế học tuần tự để học song song dựa trên multi-head self-attention, nên máy tính không hiểu được cấu trúc của câu dẫn đến việc sai về mặt ngữ nghĩa. Do đó cần một cách để giúp máy tính hiểu được thứ tự các từ trong câu, ý tưởng là ta sẽ đánh trọng số cho các từ trong câu.

Cách tiếp cận ngây thơ để đánh trọng trọng số là thêm giá trị mỗi từ theo tuyến tính. Có nghĩa là với từ đầu tiên thêm 1, từ thứ hai thêm hai, cứ như vậy cho đến hết. Với cách tiếp cần này sẽ có vài vấn đề như: các giá trị sẽ trở lên quá lớn, mô hình mất nhiều thời gian để hội tụ hơn (lý do các bạn có thể tìm kiếm với từ khóa “feature scaling”)

Một cách tiếp cận khác là thêm các giá trị trong khoảng [0, 1]; với giá trị 0 cho từ đầu tiên và 1 cho từ cuối cùng. Một trong những vấn đề đối với cách tiếp cần này là bạn không thể tìm ra có bao nhiêu từ hiện diện trong một phạm vi cụ thể.

Nói cách khác, giá trị cộng thêm không có ý nghĩa nhất quán trong các câu khác nhau. Cách tiếp cận tốt nhất nên thỏa mãn các yêu cầu sau:

* Giá trị được thêm nên là duy nhất với mỗi time-step.
* Khoảng cách giữa các từ trong câu nên thống nhất với mọi câu.
* Giá trị thêm vào cần phải được giới hạn.
* Giá trị thêm vào cần phải làm rõ được vị trí các từ trong câu. Công thức tính toán positional encoding trong Transformer là:

A close-up of a mathematical equation

Description automatically generated

A number and a line

Description automatically generated with medium confidence

Trong đó :

* d là số chiều dài của vector từ.
* k là chỉ số của chiều trong vector từ
* là vị trí từ trong câu

Mỗi thành phần trong ma trận positional encoding sẽ có giá trị sin hoặc cos của một hàm sóng với tần số khác nhau. Khi kết hợp với ma trận từ nhúng, positional encoding sẽ giúp mô hình Transformer biết được vị trí tương đối của từng từ trong câu.

Ví dụ: Giả sử chúng ta có một câu gồm các từ "I", "love", "cats". Đầu tiên, mỗi từ sẽ được biểu diễn bằng vector từ nhúng. Tiếp theo, positional encoding sẽ được thêm vào các vector từ để đưa thông tin vị trí vào mô hình Transformer.

* Đầu vào ban đầu: [vector\_embed("I"), vector\_embed("love"), vector\_embed("cats")]
* Sau khi thêm positional encoding: [vector\_embed("I")+positional\_encoding("I"), vector\_embed("love")+positional\_encoding("love"), vector\_embed("cats")+positional\_encoding("cats")]

Positional encoding giúp mô hình Transformer phân biệt được vị trí và thứ tự các từ trong câu, làm cơ sở cho việc xử lý thông tin chuỗi trong mô hình.

Masked Multi-Head Attention

Một layer tương tự như Multi-Head Attention, điểm khác biệt là ở lúc train, khi đã biết rõ output mong muốn để đưa vào làm input của time-step tiếp theo. Để hiểu rõ điều này trước tiên bạn cần biết cách Decoder vận hành, tôi sẽ lấy ví dụ trong bài toán text generate với câu “Tall house with wide door”.

A diagram of a process

Description automatically generated

Hình 2.22: Hình ảnh vị trí của lớp masked multi head attention.

Cũng giống như RNN, Decoder sử dụng output của time-step trước làm input của time-step hiện tại. Tức với ví dụ trên ta sẽ có các bước decode sau:

* input = “” qua Masked Multi-Head Attention, nhận thêm thông tin từ Encoder để đưa ra thông tin đầu tiên output = “Tall”.
* input = “Tall” → output = “Tall house”.
* input = “Tall house” → output = “Tall house with”.
* input = “Tall house with” → output = “Tall house with wide”.
* input = “Tall house with wide” → output = “Tall house with wide door”.

Mọi thứ sẽ như trên nếu mô hình đã được train xong, nhưng trong quá trình train output ở time-step trước có thể sai khiến output của time-step hiện tại cũng có thể sai theo. Và đương nhiên việc này có thể dần cải thiện trong quá trình train, nhưng sẽ rất tốn tài nguyên. Vì vậy giải pháp được đưa ra là sử dụng output mà chúng ta muốn đưa vào làm input của từng time-step. Mỗi time-step sẽ nhập input tương ứng nên có ở thời điểm đó, tức sẽ cho biết các từ nên được generate ở thời đó và che đi phần còn lại. Để máy tính hiểu được cách làm trên ta sử dụng Look-ahead Mask là ma trận tam giác dưới với các giá trị 1 biểu thị từ không muốn che, và 0 cho từ muốn che.

Add & Norm layers và Feed-forward layers

Normalization layers: Trong cấu trúc của transformer có các lớp “Add và Norm” thì “Norm” ở đây thể hiện cho Normalization layer. Lớp này đơn giản là sẽ chuẩn hóa lại đầu ra của multi-head attention, mang lại hiệu quả cho việc nâng cao khả năng hội tụ.

Khối Feed-Forward: Sau khi thực hiện tính toán ở khối attention cho mỗi lớp thì khối tiếp theo là FFN. Có thể hiểu là cơ chế attention giúp thu thập thông tin từ những tokens đầu vào thì FFN là khối xử lý những thông tin đó.

Tổng quan về việc sử dụng mô hình Transformer.

Việc áp dụng Transformer vào bài toán nhận diện chữ giúp mô hình có khả năng mô hình hóa quan hệ phức tạp giữa hình ảnh và văn bản. Các cơ chế attention trong Transformer giúp mô hình tập trung vào các phần quan trọng của dữ liệu đầu vào và giúp cải thiện chính xác của hệ thống trích xuất thông tin.

## Một số phương pháp đánh giá mô hình mô tả hình ảnh

Trong lĩnh vực mô tả hình ảnh, việc đánh giá hiệu quả của mô hình là một bước quan trọng nhằm xác định mức độ chính xác và phù hợp của các mô tả mà mô hình tạo ra. Không giống như các bài toán thông thường, việc đánh giá mô hình không chỉ dựa trên độ chính xác tuyệt đối, mà còn phải xem xét nhiều khía cạnh như ngữ nghĩa, tính mạch lạc và sự tương đồng ngữ cảnh giữa mô tả được tạo ra và mô tả tham chiếu (ground truth).

Để đảm bảo kết quả đánh giá khách quan và toàn diện, các nhà nghiên cứu sử dụng nhiều chỉ số định lượng khác nhau. Những chỉ số này không chỉ giúp đo lường sự tương đồng về mặt từ vựng giữa các mô tả, mà còn cung cấp cái nhìn sâu hơn về hiệu suất của mô hình trong việc truyền đạt ý nghĩa và nội dung của hình ảnh. Phần dưới đây sẽ trình bày chi tiết một số chỉ số đánh giá phổ biến được sử dụng trong bài toán mô tả hình ảnh.

### Độ đo BLEU (Bilingual Evaluation Understudy)

BLEU (Bilingual Evaluation Understudy) là một chỉ số đánh giá phổ biến trong bài toán Image Captioning, được thiết kế để đo lường độ chính xác của các mô hình sinh mô tả bằng cách so sánh mô tả được tạo ra với các mô tả tham chiếu.

BLEU được giới thiệu trong bài báo "BLEU: a Method for Automatic Evaluation of Machine Translation" và ban đầu được phát triển cho dịch máy, nhưng hiện nay cũng được áp dụng rộng rãi trong các nhiệm vụ như tóm tắt văn bản và sinh mô tả cho ảnh. Chỉ số này đo lường sự tương đồng giữa câu được sinh ra (caption candidate) và một tập hợp các câu tham chiếu (reference captions). Điểm BLEU nằm trong khoảng từ 0 đến 1, với 1 biểu thị sự khớp hoàn hảo.

Cách tính BLEU

N-gram Matching (So khớp n-gram): BLEU sử dụng n-gram, là các chuỗi liên tiếp gồm n từ, để so sánh giữa câu sinh ra và câu tham chiếu. Độ chính xác của nnn-gram được tính bằng cách chia số lượng n-gram khớp (có mặt trong cả câu sinh ra và câu tham chiếu) cho tổng số n-gram trong câu sinh ra.

A black text on a white background

Description automatically generated

Ví dụ minh họa:

Giả sử:

* Câu sinh ra (candidate): *"The cat is on the mat."*
* Câu tham chiếu (reference): *"The cat is sitting on the mat."*

Với unigram (n = 1):

* Các unigram trong câu sinh ra: {The, cat, is, on, the, mat}.
* Các unigram khớp: {The, cat, is, on, the, mat} (tất cả khớp).
* Precision = 6/6 = 1.0.

Với bigram (n = 2):

* Các bigram trong câu sinh ra: {The cat, cat is, is on, on the, the mat}.
* Các bigram khớp: {The cat, on the, the mat}.
* Precision = 3/5 = 0.6.

Brevity Penalty (BP - Hình phạt độ dài): BLEU sử dụng hệ số phạt độ dài để đảm bảo các câu sinh ra không bị cắt ngắn quá mức nhằm đạt điểm cao.

Công thức tính BP:

A black text on a white background

Description automatically generated

Trong đó:

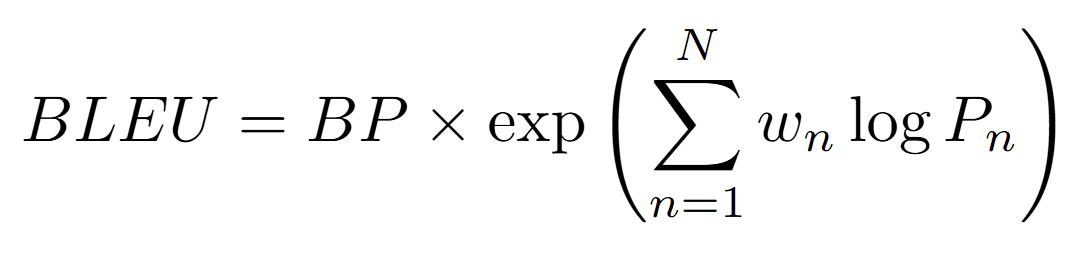
* c: Độ dài của câu sinh ra.
* r: Độ dài trung bình của các câu tham chiếu.

Ví dụ:

* Câu sinh ra: *"The cat is on the mat."* (c = 6).
* Câu tham chiếu: *"The cat is sitting on the mat."* (r = 7).
* BP = exp(1 − 7/6) = 0.

Công thức tổng quát BLEU

BLEU kết hợp các độ chính xác của n-gram và Brevity Penalty để tính điểm cuối cùng:



Trong đó:

* : Độ chính xác của n-gram với *n = 1,2,…,N*.
* ​: Trọng số của từng *n-gram* (thường = 1 / N, phân bố đều).
* BP: Hệ số phạt độ dài.
* Ví dụ cụ thể:

Giả sử chúng ta tính BLEU-2 (unigram và bigram):

* + = 1.0, = 0.6.
  + = ​ = 0.5.
  + BP = 0.

Tính điểm BLEU:

*BLEU* = 0.85 × exp(0.5 × log(1.0) + 0.5 × log(0.6))

*BLEU* = 0.85 × exp(0 − 0.255) = 0.85 × 0.775 = 0

BLEU là một chỉ số hiệu quả và dễ tính toán để đánh giá sự tương đồng giữa các câu sinh ra và câu tham chiếu trong bài toán mô tả hình ảnh. Tuy nhiên, để đảm bảo đánh giá toàn diện hơn, các nhà nghiên cứu thường kết hợp BLEU với các chỉ số khác như METEOR hoặc CIDEr.

### Độ đo METEOR (Metric for Evaluation of Translation with Explicit ORdering)

Chỉ số METEOR là một công cụ đánh giá quan trọng trong bài toán Image Captioning và các nhiệm vụ xử lý ngôn ngữ tự nhiên khác. METEOR được phát triển để cải thiện các chỉ số đánh giá trước đó như BLEU, bằng cách xem xét không chỉ độ chính xác mà còn cả ngữ nghĩa và cấu trúc của câu. Chỉ số này nhằm mục đích cung cấp một cách đánh giá chính xác hơn về chất lượng của các mô tả được sinh ra, đặc biệt trong các trường hợp mà ngữ nghĩa và sự đa dạng ngôn ngữ là quan trọng. Chi tiết về các bước tính toán của METEOR gồm:

Tính chỉ số Precision và Recall (Độ chính xác và độ phủ)

Precision (P) - Đo tỷ lệ từ trong bản dịch đúng (khớp với bản gốc) so với tất cả từ trong bản dịch. Được tính theo công thức:

A black text on a white background

Description automatically generated

Recall (R) - Đo tỷ lệ từ trong bản gốc đúng (khớp với bản dịch) so với tất cả từ trong bản gốc. Được tính theo công thức:

A black text on a white background

Description automatically generated

****F-Score -**** Điểm F là một chỉ số cân bằng giữa precision và recall. Nó được tính bằng cách sử dụng công thức **F-Score** hàm điều hòa (harmonic mean):

A black text on a white background

Description automatically generated

Penalties (Hình phạt)

METEOR áp dụng một số hình phạt để phản ánh sự sai lệch về ngữ nghĩa hoặc thứ tự từ. Các hình phạt chính là:

Penalty for Fragmentation: Được áp dụng nếu bản dịch có quá nhiều sự phân mảnh, tức là khi bản dịch có quá nhiều các cụm từ không phù hợp hoặc thiếu kết nối.

A black text with a white background

Description automatically generated

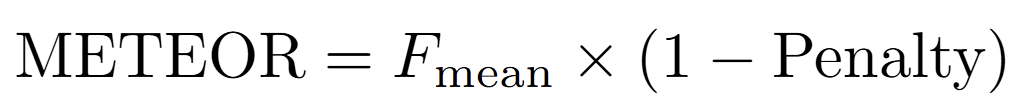
Penalty for Word Order: Hình phạt cho sự thay đổi thứ tự từ trong câu.



Trong đó, weight\_order là một trọng số điều chỉnh mức độ ảnh hưởng của sự thay đổi thứ tự từ.

Final METEOR Score (Điểm cuối cùng)

Cuối cùng, điểm METEOR được tính bằng công thức sau, trong đó có sự điều chỉnh hình phạt và trọng số cho các yếu tố:



Chỉ số METEOR (Metric for Evaluation of Translation with Explicit Ordering) đánh giá chất lượng của mô tả hình ảnh thông qua nhiều yếu tố khác nhau. Đầu tiên, nó xem xét tương đồng từ vựng và khớp từ chính xác, giúp xác định sự phù hợp giữa từ trong câu dự đoán và câu tham chiếu. METEOR cũng đánh giá tính linh hoạt trong việc sắp xếp từ trong câu, đảm bảo rằng thứ tự từ không làm mất đi ý nghĩa, khác với các chỉ số như BLEU chỉ dựa trên khớp n-gram. Thêm vào đó, stemming được áp dụng để nhận diện các từ có gốc chung, và cân nhắc sự khác biệt về chính tả giúp tăng độ chính xác khi so sánh các từ. Cuối cùng, METEOR duy trì cân bằng giữa Precision và Recall, điều này giúp tránh việc chỉ đánh giá các từ chính xác mà bỏ qua những từ có ý nghĩa tương đương nhưng không hoàn toàn giống nhau, làm cho nó trở thành một chỉ số hiệu quả hơn trong các bài toán như mô tả hình ảnh.

Trong bài toán Image Captioning, chỉ số METEOR được sử dụng để đánh giá chất lượng mô tả hình ảnh bằng cách so sánh câu dự đoán với các câu tham chiếu. METEOR xem xét độ chính xác của các từ, sự đồng nghĩa giữa từ vựng, và tính tự nhiên của câu mô tả khi phản ánh thứ tự các đối tượng trong hình ảnh. So với BLEU, METEOR có ưu điểm vượt trội trong việc đánh giá sự tương đồng từ vựng và cấu trúc câu, điều này quan trọng trong việc mô tả hình ảnh khi các mô tả có thể sử dụng từ ngữ khác nhau nhưng mang ý nghĩa tương đương. Hơn nữa, METEOR ít bị ảnh hưởng bởi câu ngắn, vấn đề thường gặp trong các chỉ số như BLEU.

### Độ đo CIDEr (Consensus-based Image Description Evaluation)

CIDEr được phát triển để cải thiện các chỉ số đánh giá trước đó như BLEU và METEOR, bằng cách tập trung vào sự đồng thuận giữa nhiều mô tả của con người cho cùng một hình ảnh. Mục tiêu của CIDEr là cung cấp một cách đánh giá chính xác hơn về chất lượng của các mô tả, phản ánh tốt hơn cách mà con người mô tả hình ảnh.

Cách tính CIDEr

CIDEr áp dụng Term Frequency-Inverse Document Frequency (TF-IDF) để xác định tầm quan trọng của từng từ trong các mô tả. TF-IDF là một phương pháp phổ biến trong xử lý văn bản, được dùng để đo độ quan trọng của một từ trong ngữ cảnh của một tập hợp văn bản lớn.

* Term Frequency (TF): Đo tần suất xuất hiện của một từ trong một văn bản (hoặc mô t).

A close up of a sign

Description automatically generated

* Inverse Document Frequency (IDF): Đo độ hiếm của từ đó trong toàn bộ tập hợp văn bản (các mô tả tham chiếu). Càng ít xuất hiện, giá trị IDF càng cao.

A black text on a white background

Description automatically generated

Trong đó:

* N là tổng số tài liệu trong tập dữ liệu D.
* Số tài liệu chứa từ t là số tài liệu trong D mà từ ttt xuất hiện ít nhất một lần.

Trong CIDEr, các từ có TF-IDF cao sẽ được coi là quan trọng hơn trong việc mô tả hình ảnh, giúp tăng độ chính xác của mô tả. Những từ hiếm gặp nhưng có ý nghĩa trong mô tả hình ảnh sẽ được đánh giá cao hơn, tránh sự thiên vị cho các từ quá phổ biến.

**Đánh Giá Dựa trên Consensus (Sự Đồng Thuận) -** CIDEr đánh giá mô tả hình ảnh không chỉ dựa vào độ chính xác của từng từ mà còn xem xét **sự đồng thuận** giữa các mô tả tham chiếu. Các mô tả tham chiếu là những mô tả chuẩn mà con người đưa ra cho một hình ảnh, và mục tiêu của mô hình mô tả là sinh ra mô tả gần giống nhất với những mô tả này. **Sự đồng thuận** giữa các mô tả tham chiếu là một yếu tố quan trọng trong CIDEr. Nếu mô tả sinh ra có nhiều điểm trùng khớp với các mô tả tham chiếu, chỉ số CIDEr sẽ cao hơn.

Công thức tính CIDEr là sự kết hợp của các yếu tố như sau:

* Tính n-gram precision giữa mô tả dự đoán và các mô tả tham chiếu.
* Sử dụng TF-IDF để điều chỉnh độ quan trọng của các n-gram.
* Tổng hợp kết quả cho từng n-gram (unigram, bigram, trigram, v.v.) và trung bình hóa trên tất cả các n-gram.

Công thức cơ bản có thể được mô tả như sau:

A close-up of a word

Description automatically generated

Trong đó:

* TF IDF(n) là trọng số TF-IDF cho n-gram,
* similarity(n) là độ tương đồng giữa n-gram của mô tả sinh ra và mô tả tham chiếu.
* normnorm là yếu tố chuẩn hóa để đảm bảo tính chính xác.

CIDEr có những ưu điểm vượt trội trong việc đánh giá mô tả hình ảnh so với các chỉ số khác như BLEU hoặc METEOR. Đầu tiên, CIDEr cung cấp kết quả chính xác hơn nhờ vào việc tính đến sự đa dạng và đồng thuận của các câu tham chiếu. Thứ hai, việc sử dụng TF-IDF giúp CIDEr ưu tiên các từ hoặc cụm từ quan trọng và có ngữ nghĩa, thay vì chỉ chú trọng vào sự khớp chính xác của từ ngữ. Cuối cùng, CIDEr giảm thiểu sự thiên lệch từ các câu tham chiếu duy nhất, giúp đánh giá các mô tả sinh ra từ mô hình có sự đa dạng, phù hợp hơn với nhiều cách mô tả khác nhau về cùng một hình ảnh.

## Tổng kết chương hai

Trong chương hai em thực hiện đưa ra cái nhìn tổng quan về các mô hình học sâu (Deep Learning) áp dụng trong bài toán mô tả hình ảnh, bao gồm các kiến trúc mạng nơ-ron cho xử lý hình ảnh và phân tích ngữ nghĩa, ngữ cảnh. Đầu tiên, chương giới thiệu Mạng nơ-ron nhân tạo (ANN) và Mạng nơ-ron tích chập (CNN), hai mô hình quan trọng trong xử lý hình ảnh, giúp trích xuất đặc trưng từ ảnh và học các đặc điểm không gian hiệu quả. Tiếp theo, chương trình bày các mô hình học sâu cho phân tích ngữ nghĩa và tạo mô tả, bao gồm Mạng nơ-ron hồi quy (RNN), Bộ nhớ dài-ngắn hạn (LSTM) và Mô hình Transformer. Các mô hình này chủ yếu phục vụ cho việc xử lý chuỗi ngữ nghĩa và tạo ra các câu mô tả tự động từ các đặc trưng hình ảnh đã được trích xuất.

Chương cũng làm rõ các yếu tố then chốt trong mô hình Transformer, đặc biệt là cơ chế Attention, giúp mô hình tập trung vào những phần quan trọng của hình ảnh khi tạo mô tả, và cấu trúc tổng thể của nó, nhằm cải thiện hiệu quả trong các tác vụ liên quan đến chuỗi dài và mô tả hình ảnh.

Cuối cùng, chương này cũng giới thiệu các chỉ số đánh giá quan trọng trong việc đo lường chất lượng của các mô hình mô tả hình ảnh: BLEU, METEOR, và CIDEr. Những chỉ số này giúp đánh giá mức độ tương đồng giữa mô tả được tạo ra và các câu tham chiếu, từ đó xác định tính chính xác và sự phù hợp của mô hình đối với bài toán mô tả hình ảnh.

Những nội dung trong chương này cơ bản cung cấp nền tảng vững chắc về lý thuyết và các công nghệ hiện có, phục vụ cho việc nghiên cứu và phát triển các mô hình học sâu trong bài toán mô tả hình ảnh.

# THỰC NGHIỆM VÀ XÂY DỰNG ỨNG DỤNG MÔ TẢ HÌNH ẢNH

## Dữ liệu thực nghiệm trên các mô hình

Trong khuôn khổ của đồ án này em sử dụng bộ dữ liệu Flickr8k để thực nghiệm vì tính đa dạng về nội dung hình ảnh và sự phong phú trong các chú thích, giúp mô hình học được cách tạo ra mô tả hình ảnh chính xác và tự nhiên.

Tổng quan về bộ dữ liệu Flickr8k

Bộ dữ liệu Flickr8k là một bộ dữ liệu quan trọng trong lĩnh vực chú thích hình ảnh (image captioning) và thị giác máy tính (computer vision). Bộ dữ liệu này bao gồm hơn 8.000 bức ảnh độc đáo, mỗi bức ảnh đi kèm với năm câu chú thích khác nhau, miêu tả đa dạng về nội dung hình ảnh. Các chú thích này được viết bởi con người, giúp phản ánh nhiều cách nhìn nhận khác nhau về một bức ảnh, từ đó tạo cơ hội cho mô hình học được các cách diễn giải ngữ nghĩa phong phú. Chính vì vậy, Flickr8k trở thành một bộ dữ liệu phổ biến trong nghiên cứu và phát triển các mô hình tạo mô tả hình ảnh tự động, trong đó mô hình cần phải phân tích hình ảnh và tạo ra các câu miêu tả hợp lý, rõ ràng.

Cấu trúc bộ dữ liệu và tính chất của dữ liệu

Bộ dữ liệu Flickr8k bao gồm:

* Hơn 8.000 bức ảnh được lấy từ Flickr, một nền tảng chia sẻ ảnh trực tuyến, với đa dạng các thể loại và phong cảnh. Các bức ảnh này được chọn lọc cẩn thận để đảm bảo sự đa dạng về chủ đề và bối cảnh.
* Năm câu chú thích cho mỗi bức ảnh, mỗi câu chú thích cung cấp một cách diễn giải khác nhau về hình ảnh đó, qua đó tạo ra sự đa dạng trong việc mô tả hình ảnh, giúp mô hình học được nhiều cách biểu đạt khác nhau từ một hình ảnh duy nhất.

Đặc điểm và ưu điểm của bộ dữ liệu

* Đa dạng về nội dung: Một trong những điểm mạnh của bộ dữ liệu Flickr8k là sự đa dạng trong các bức ảnh. Các hình ảnh trong bộ dữ liệu này không tập trung vào các địa danh nổi tiếng hay các nhân vật nổi tiếng, mà chủ yếu phản ánh các hoạt động và cảnh vật trong cuộc sống hàng ngày. Điều này giúp bộ dữ liệu trở nên gần gũi và thực tế hơn, phản ánh một thế giới phong phú và đa dạng về văn hóa, địa lý.
* Đa dạng trong chú thích: Việc mỗi hình ảnh có đến 5 chú thích khác nhau giúp tăng tính linh hoạt và độ phong phú trong việc huấn luyện mô hình. Các câu chú thích này có thể mô tả hình ảnh theo những cách khác nhau, từ đó cho phép mô hình học được cách nhận diện và miêu tả các đặc điểm của hình ảnh một cách chính xác hơn. Bằng cách này, bộ dữ liệu này giúp các mô hình học được cách tạo ra những câu chú thích không chỉ chính xác mà còn tự nhiên và linh hoạt.
* Được lựa chọn thủ công: Các bức ảnh trong bộ dữ liệu được lựa chọn thủ công từ nhiều nhóm khác nhau trên Flickr, đảm bảo rằng chúng bao phủ nhiều chủ đề, phong cảnh và tình huống khác nhau. Điều này giúp bộ dữ liệu trở nên đa dạng và bao quát, cung cấp một cái nhìn toàn diện về các bức ảnh trong thế giới thực.

Nhược điểm của bộ dữ liệu

Mặc dù Flickr8k có nhiều ưu điểm, nhưng nó cũng có một số hạn chế:

* Kích thước bộ dữ liệu hạn chế: Với chỉ 8.000 bức ảnh, bộ dữ liệu này tương đối nhỏ so với những bộ dữ liệu khác như MS COCO hay Flickr30k, khiến cho việc huấn luyện các mô hình phức tạp có thể gặp khó khăn. Những mô hình học sâu hiện đại thường yêu cầu một lượng dữ liệu lớn để đạt được độ chính xác cao, và bộ dữ liệu này có thể không đủ lớn để đáp ứng yêu cầu này.
* Chất lượng chú thích không đồng đều: Mặc dù chú thích được viết bởi con người, nhưng chất lượng của các chú thích có thể không đồng đều giữa các bức ảnh. Một số chú thích có thể quá chung chung hoặc không phản ánh chính xác nội dung của bức ảnh, điều này có thể gây khó khăn cho việc huấn luyện mô hình.

Bộ dữ liệu Flickr8k là một trong những bộ dữ liệu lý tưởng cho bài toán chú thích hình ảnh tự động (image captioning) vì một số lý do sau:

* Tính đa dạng và phong phú của dữ liệu: Mỗi bức ảnh trong Flickr8k được mô tả bởi 5 chú thích khác nhau, điều này giúp mô hình học được nhiều cách diễn giải và cách miêu tả khác nhau đối với cùng một hình ảnh. Điều này đặc biệt quan trọng trong bài toán image captioning, vì mục tiêu là tạo ra một mô tả tự nhiên, phong phú và chính xác về nội dung của hình ảnh.
* Đảm bảo tính thực tế và gần gũi: Các bức ảnh trong Flickr8k không chỉ bao gồm các cảnh vật nổi tiếng mà còn phản ánh những khoảnh khắc hàng ngày và những chủ đề phổ biến. Điều này giúp mô hình học được cách tạo ra các chú thích phù hợp với nhiều tình huống thực tế, từ đó có thể áp dụng tốt trong nhiều ứng dụng thực tế.
* Dễ dàng đánh giá và thử nghiệm: Với cấu trúc chuẩn, bộ dữ liệu Flickr8k cung cấp một nền tảng để các nhà nghiên cứu đánh giá và so sánh các mô hình image captioning. Các mô hình có thể được huấn luyện và kiểm tra với bộ dữ liệu này, giúp đánh giá khả năng của mô hình trong việc tạo ra các mô tả hình ảnh chính xác và tự nhiên.
* Dễ tiếp cận và sử dụng: Bộ dữ liệu Flickr8k có sẵn và dễ dàng tiếp cận, giúp các nhà nghiên cứu và lập trình viên có thể nhanh chóng bắt đầu huấn luyện mô hình mà không phải mất quá nhiều thời gian vào việc thu thập và chuẩn bị dữ liệu.

Kết luận

Bộ dữ liệu Flickr8k là một tài nguyên quan trọng và hữu ích trong việc phát triển các mô hình chú thích hình ảnh. Với sự đa dạng về nội dung hình ảnh và các chú thích khác nhau, bộ dữ liệu này không chỉ giúp huấn luyện mô hình nhận diện hình ảnh mà còn giúp mô hình học được cách diễn giải và tạo ra các mô tả văn bản chính xác và tự nhiên. Tuy nhiên, bộ dữ liệu này vẫn có một số hạn chế về kích thước và chất lượng chú thích, điều này đòi hỏi các nhà nghiên cứu cần phải tìm cách cải thiện hoặc kết hợp với các bộ dữ liệu khác để nâng cao hiệu quả huấn luyện mô hình.

## Phương pháp thực nghiệm và đánh giá các mô hình

Trong bài báo cáo em sử dụng bộ dữ liệu Flickr8k, được chia thành tập huấn luyện và kiểm tra. Các đặc trưng hình ảnh được trích xuất từ mạng VGG-16 đã tiền huấn luyện. Ba mô hình RNN, LSTM, và Transformer được triển khai, kết hợp bộ nhúng từ để tạo chú thích ảnh. Hiệu suất được đánh giá qua các chỉ số BLEU, METEOR và CIDEr trên tập kiểm tra. Quá trình thực nghiệm sử dụng GPU, với các công cụ Python, TensorFlow/Transformer. Dưới đây là bảng so sánh các mô hình với số lượng epochs tương ứng với từng chỉ số để đánh giá:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mô hình với số epochs huấn luyện | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | CIDEr |
| RNN 5 eps | 0.457471994 | 0.375501916 | 0.346628727 | 0.332784733 | 0.559018868 | 0.736660468 |
| RNN 10 eps | 0.461029781 | 0.375867081 | 0.347978538 | 0.334412898 | 0.560981542 | 0.756770334 |
| RNN 15 eps | 0.45182593 | 0.374621173 | 0.347561059 | 0.334253878 | 0.557549667 | 0.757760031 |
| RNN 20 eps | 0.456025214 | 0.379681281 | 0.348687067 | 0.334738788 | 0.556660759 | 0.779146565 |
| LSTM 5 eps | 0.458265164 | 0.377935488 | 0.3480577 | 0.334657207 | 0.559602279 | 0.756559942 |
| LSTM 10 eps | 0.454729979 | 0.376669716 | 0.348338312 | 0.335108742 | 0.559579867 | 0.764204807 |
| LSTM 15 eps | 0.457602524 | 0.378612738 | 0.349026083 | 0.335174869 | 0.560139917 | 0.745538789 |
| LSTM 20 eps | 0.455105431 | 0.374309965 | 0.345873015 | 0.332030552 | 0.558519217 | 0.749696264 |
| Transformer  5 eps | 0.545141437 | 0.476216817 | 0.420392101 | 0.398947538 | 0.641645002 | 0.818618945 |
| Transformer  10 eps | 0.545677905 | 0.435701705 | 0.39949305 | 0.41264657 | 0.635643886 | 0.798402589 |
| Transformer  15 eps | 0.551857676 | 0.474008744 | 0.414026718 | 0.387223935 | 0.616153998 | 0.810270084 |
| Transformer  20 eps | 0.512925002 | 0.44573765 | 0.416837133 | 0.38708752 | 0.63202333 | 0.85186167 |

Trong nghiên cứu này, em phân tích hiệu suất của ba loại mô hình tạo chú thích ảnh, bao gồm RNN, LSTM và Transformer. Các mô hình này được huấn luyện trên tập đặc trưng VGG-16, và hiệu suất của các mô hình được đánh giá qua các chỉ số BLEU-1 đến BLEU-4, METEOR, và CIDEr. Các chỉ số này được sử dụng để đo lường chất lượng, mức độ liên quan, và tính đa dạng của các chú thích mà mô hình sinh ra qua các số lượng epoch khác nhau.

Hiệu suất của RNN

RNN cho thấy hiệu suất ban đầu ở mức trung bình khi được huấn luyện trong 5 epoch, với BLEU-1 = 0.457, BLEU-2 = 0.375, BLEU-3 = 0.346, BLEU-4 = 0.332, METEOR = 0.559, và CIDEr = 0.737. Hiệu suất của mô hình này cải thiện nhẹ khi tăng số lượng epoch, nhưng có sự dao động đáng kể, điều này cho thấy khả năng xuất hiện hiện tượng quá khớp hoặc mô hình đạt ngưỡng bão hòa sớm. Đặc điểm hạn chế của RNN là sự phụ thuộc tuần tự, dẫn đến khó khăn trong việc xử lý các mối quan hệ dài hạn trong dữ liệu.

Hiệu suất của LSTM

So với RNN, LSTM thể hiện hiệu suất tốt hơn trên hầu hết các chỉ số đánh giá. Điều này có thể được giải thích bởi khả năng của LSTM trong việc xử lý các mối quan hệ dài hạn nhờ cấu trúc bộ nhớ đặc biệt của nó. Tuy nhiên, xu hướng cải thiện và dao động hiệu suất của LSTM cũng tương tự RNN khi tăng số lượng epoch. Điều này cho thấy mặc dù LSTM khắc phục một số hạn chế của RNN, nó vẫn gặp khó khăn trong việc duy trì hiệu suất ổn định với các tập dữ liệu phức tạp và lớn hơn.

Hiệu suất của Transformer

Transformer vượt trội hơn cả RNN và LSTM, ngay cả khi chỉ huấn luyện với 5 epoch. Các chỉ số đạt được bao gồm BLEU-1 = 0.545, BLEU-2 = 0.476, BLEU-3 = 0.420, BLEU-4 = 0.398, METEOR = 0.642, và CIDEr = 0.819. Hiệu suất của Transformer duy trì ở mức cao ngay cả khi tăng số lượng epoch, mặc dù xuất hiện một số dao động tại các epoch lớn, điều này chỉ ra rằng Transformer cũng có nguy cơ quá khớp nếu không được tối ưu hóa cẩn thận. Thành công của Transformer chủ yếu nhờ vào cơ chế tự chú ý (self-attention), giúp mô hình xử lý hiệu quả các mối quan hệ dài hạn trong dữ liệu.

Nguyên nhân khác biệt về hiệu suất

Khả năng vượt trội của Transformer có thể được giải thích bởi hai yếu tố chính. Thứ nhất, cơ chế tự chú ý (self-attention) cho phép Transformer xác định và xử lý các mối quan hệ dài hạn trong dữ liệu một cách hiệu quả, điều này cực kỳ quan trọng đối với bài toán tạo chú thích ảnh. Thứ hai, mô hình Transformer hỗ trợ xử lý song song, giúp tăng tốc độ huấn luyện so với các mô hình tuần tự như RNN và LSTM, mặc dù đòi hỏi nhiều tài nguyên tính toán hơn.

Kết luận

Kết quả phân tích cho thấy Transformer là lựa chọn tối ưu cho bài toán tạo chú thích ảnh dựa trên đặc trưng VGG-16, nhờ hiệu suất vượt trội trên các chỉ số BLEU, METEOR, và CIDEr. Tuy nhiên, cần cẩn trọng trong việc lựa chọn số lượng epoch để tránh hiện tượng quá khớp. Để nâng cao hiệu suất hơn nữa, các nghiên cứu trong tương lai có thể tập trung vào tối ưu hóa siêu tham số, tinh chỉnh mô hình, và thực hiện phân tích định tính các chú thích.

## Xây dựng ứng dụng sinh mô tả tự động cho hình ảnh

Ứng dụng được xây dựng nhằm mục tiêu tự động tạo ra các mô tả văn bản cho hình ảnh. Với khả năng này, ứng dụng có thể hỗ trợ nhiều lĩnh vực thực tế như quản lý dữ liệu hình ảnh trong các thư viện lớn, hỗ trợ tìm kiếm hình ảnh dựa trên mô tả văn bản, hoặc thậm chí trợ giúp người dùng khiếm thị trong việc hiểu nội dung hình ảnh. Tính năng tự động hóa không chỉ tiết kiệm thời gian mà còn nâng cao khả năng tiếp cận công nghệ cho mọi đối tượng.

Ứng dụng được xây dựng với kiến trúc bao gồm hai thành phần chính: FastAPI làm thực hiện logic và Streamlit làm giao diện, kết hợp để đảm bảo hiệu suất và trải nghiệm người dùng mượt mà.

FastAPI - Backend

FastAPI được lựa chọn để xây dựng backend nhờ vào các ưu điểm vượt trội như:

* Hiệu suất cao: Dựa trên ASGI (Asynchronous Server Gateway Interface), FastAPI có thể xử lý đồng thời nhiều yêu cầu, giúp cải thiện tốc độ phản hồi của hệ thống.
* Dễ tích hợp với machine learning: FastAPI hỗ trợ tích hợp các mô hình machine learning như PyTorch và TensorFlow nhờ khả năng xử lý dữ liệu JSON và thư viện Pydantic để xác thực đầu vào.
* Hỗ trợ RESTful API: Cho phép xây dựng các endpoint linh hoạt, chẳng hạn như /upload\_image để nhận ảnh từ người dùng và /generate\_caption để trả về kết quả mô tả.

Cụ thể, backend đảm nhiệm các nhiệm vụ:

* Nhận dữ liệu từ giao diện: FastAPI nhận hình ảnh dưới dạng tệp tải lên từ Streamlit.
* Tiền xử lý ảnh: Thực hiện các bước chuẩn hóa hình ảnh (thay đổi kích thước, chuẩn hóa pixel) trước khi đưa vào mô hình machine learning.
* Gọi mô hình machine learning: Tích hợp mô hình Transformer
* Xử lý kết quả: Sau khi nhận được mô tả từ mô hình, FastAPI định dạng dữ liệu và gửi lại cho giao diện để hiển thị.

Streamlit - Frontend

Streamlit là công cụ frontend cung cấp giao diện người dùng thân thiện, giúp người dùng tương tác trực tiếp với hệ thống.

* Dễ sử dụng: Streamlit cho phép tạo giao diện đơn giản với chỉ vài dòng mã, phù hợp để xây dựng giao diện tải ảnh và hiển thị kết quả nhanh chóng.
* Tương tác thời gian thực: Người dùng có thể tải ảnh lên và nhận kết quả mô tả ngay lập tức mà không cần phải tải lại trang.
* Tích hợp trực quan: Streamlit hỗ trợ hiển thị kết quả dưới dạng văn bản hoặc hình ảnh, giúp người dùng dễ dàng theo dõi.

Giao diện Streamlit cung cấp các chức năng chính:

* Tải ảnh lên: Người dùng chọn ảnh từ máy tính để tải lên hệ thống.
* Hiển thị kết quả: Mô tả tự động được sinh bởi mô hình sẽ hiển thị trực tiếp trên giao diện sau khi xử lý hoàn tất.
* Thông báo lỗi: Trong trường hợp xảy ra lỗi (như ảnh không hợp lệ), Streamlit thông báo để người dùng kiểm tra lại đầu vào.

A screenshot of a computer

Description automatically generated

Hình 3.1: Quy trình hoạt động của hệ thống mô tả hình ảnh

Quy trình hoạt động của hệ thống sinh mô tả tự động cho hình ảnh bao gồm các bước: đầu tiên, người dùng tải ảnh lên thông qua giao diện Streamlit, sau đó ảnh được gửi đến backend xử lý bằng FastAPI. Tại backend, ảnh được tiền xử lý thông qua bước chuẩn hóa và thay đổi kích thước để phù hợp với đầu vào của mô hình VGG-16. Mô hình VGG-16 sau đó trích xuất các vector đặc trưng đại diện cho nội dung hình ảnh, và những đặc trưng này được đưa vào mô hình Transformer để sinh câu mô tả. Bộ giải mã (decoder) của Transformer dự đoán từng từ trong mô tả dựa trên ngữ cảnh của các từ trước đó, tạo ra một câu hoàn chỉnh mô tả nội dung ảnh.

A screenshot of a phone

Description automatically generated

Hình 3.2: Giao diện hệ thống tạo mô tả cho ảnh

Cuối cùng, kết quả mô tả được gửi trở lại frontend để hiển thị cùng với hình ảnh gốc. Quy trình đảm bảo tính chính xác cao và khả năng xử lý thời gian thực, phù hợp với các ứng dụng yêu cầu hiệu quả và độ tin cậy.

## Tổng kết chương ba

Chương này trình bày quy trình thực nghiệm và xây dựng ứng dụng sinh mô tả tự động cho hình ảnh. Dữ liệu thực nghiệm được chuẩn bị kỹ lưỡng, và các mô hình như VGG-16 kết hợp Transformer được đánh giá bằng các chỉ số BLEU, METEOR, và CIDEr. Hệ thống ứng dụng được phát triển với FastAPI và Streamlit, đảm bảo hiệu quả và thân thiện với người dùng. Kết quả cho thấy Transformer vượt trội trong việc sinh mô tả, tạo cơ sở cho cải tiến hệ thống ở các chương tiếp theo.

# KẾT LUẬN

Kết quả đạt được của luận án

Đề tài đã nghiên cứu và triển khai thành công hệ thống sinh mô tả tự động cho hình ảnh, kết hợp phân tích và so sánh các mô hình RNN, LSTM, và Transformer. Trong đó, Transformer đã chứng minh hiệu quả vượt trội nhờ khả năng xử lý ngữ cảnh và mối quan hệ phức tạp giữa các thành phần hình ảnh. Kết quả thực nghiệm cho thấy các chỉ số BLEU, METEOR, và CIDEr đạt mức cao, khẳng định tính ưu việt của mô hình trong việc tạo chú thích hình ảnh chính xác và đa dạng. Điều này khẳng định tiềm năng của Transformer trong việc xử lý các bài toán yêu cầu khả năng hiểu ngữ cảnh và mối quan hệ phức tạp giữa các thành phần trong hình ảnh.

Bên cạnh đó, việc sử dụng FastAPI và Streamlit trong hệ thống giúp xây dựng ứng dụng thân thiện, dễ sử dụng, hỗ trợ người dùng nhanh chóng nhận được mô tả chính xác từ hình ảnh đầu vào. Tuy nhiên, một số hạn chế về thời gian xử lý và yêu cầu tài nguyên đã được ghi nhận, làm tiền đề cho các cải tiến trong tương lai.

Hướng phát triển trong tương lai

Hướng phát triển tiếp theo sẽ tập trung vào việc mở rộng dữ liệu huấn luyện để nâng cao khả năng sinh mô tả trong nhiều ngữ cảnh phức tạp hơn, tối ưu hóa mô hình nhằm giảm thời gian xử lý, và tích hợp các kỹ thuật học chuyển giao để nâng cao hiệu quả trên các thiết bị hạn chế. Sử dụng các kỹ thuật giảm số lượng tham số và nén mô hình, chẳng hạn như Pruning hoặc Quantization, để giảm thời gian xử lý và tiết kiệm tài nguyên. Nghiên cứu và áp dụng các biến thể hiện đại của Transformer, chẳng hạn như Vision Transformer (ViT) hoặc Swin Transformer, nhằm cải thiện hiệu quả và khả năng hiểu sâu sắc hơn về đặc trưng hình ảnh. Những nỗ lực này sẽ giúp hệ thống không chỉ đạt độ chính xác cao hơn mà còn mở rộng phạm vi ứng dụng trong các bài toán thị giác máy tính thực tế.



# TÀI LIỆU THAM KHẢO

[1]. Nguyễn Thị Thu Hà, Lê Hoàng Sơn (2020). "Nghiên cứu và đề xuất giải pháp tạo chú thích ảnh tự động sử dụng học sâu". Tạp chí Khoa học và Công nghệ, Đại học Quốc gia Hà Nội.

[2]. Trần Văn Nam, Nguyễn Thị Thanh Huyền (2022). "Ứng dụng mô hình Transformer trong bài toán sinh chú thích ảnh tự động cho tiếng Việt". Tạp chí Công nghệ Thông tin và Truyền thông.

[3]. Vũ Tiến Thành, Hoàng Văn Thông (2023). "Cải tiến mô hình Attention trong bài toán sinh chú thích ảnh tự động". Tạp chí Tin học và Điều khiển học.

[4]. Nguyễn Thị Mai, Lê Hoàng Quốc Bảo (2021). "Xây dựng hệ thống tự động sinh chú thích ảnh sử dụng mạng nơ-ron tích chập và mạng LSTM". Tạp chí Khoa học Công nghệ Việt Nam.

[5]. Trần Công Án, Nguyễn Doãn Phước (2022). "Phương pháp tối ưu hóa quá trình học cho mô hình sinh chú thích ảnh đa ngôn ngữ". Tạp chí Khoa học và Công nghệ.

[6]. Đặng Thị Thu Hiền, Lê Thanh Hương (2023). "Nghiên cứu và đề xuất giải pháp cải thiện độ chính xác trong bài toán sinh chú thích ảnh tự động cho tiếng Việt". Tạp chí Công nghệ Thông tin và Truyền thông.

Quanzeng You , Hailin Jin , Zhaowen Wang , Chen Fang , and Jiebo Luo (2016). Image Captioning with Semantic Attention.

[7]. Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara (2021). From Show to Tell: A Survey on Deep Learning-based Image Captioning

[8]. Anbara Z Al-Jamal , Maryam J Bani-Amer, Shadi Aljawarneh (2022). Image Captioning Techniques: A ReviewImage Captioning Techniques: A Review.

[9]. S. Nehan, A.Vijaya Lakshmi, C. Shambhavi, G. Chandrakanth (2024). Implementation of Image Caption Generation Using VGG16 and Resnet50. DOI:[10.1109/ICDCECE60827.2024.10548963](http://dx.doi.org/10.1109/ICDCECE60827.2024.10548963)

[10]. Keshav Rungta, Geeling Chau, Anshuman Dewangan, Margot Wagner (2022). Image Captioning using an LSTM Network.

[11]. Faizan Shaikh (2021). Automatic Image Captioning using Deep Learning (CNN and LSTM) in PyTorch.

[12]. Atul Ranjan (2022). Image Captioning with an End-to-End Transformer Network.

[13]. Junzhe Zhang, Huixuan Zhang, Xunjian Yin, Xiaojun Wan (2024). EAMA : Entity-Aware Multimodal Alignment Based Approach for News Image Captioning.

[14]. Seungjun Lee, Jungseob Lee 1,Hyeonseok Moon, Chanjun Park, Jaehyung Seo,Sugyeong Eo, Seonmin Koo andHeuiseok Lim (2023). A Survey on Evaluation Metrics for Machine Translation.

[15]. Cornia, M., Baraldi, L., & Cucchiara, R. (2020). "Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions". IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[16]. Chen, X., & Lawrence Zitnick, C. (2015). "Mind's eye: A recurrent visual representation for image caption generation". IEEE Conference on Computer Vision and Pattern Recognition.

[17]. Huang, L., Wang, W., Chen, J., & Wei, X. Y. (2019). "Attention on Attention for Image Captioning". IEEE/CVF International Conference on Computer Vision.

[18]. Li, X., Jiang, S., & Han, J. (2019). "Learning Object Context for Dense Captioning". AAAI Conference on Artificial Intelligence.

[19]. Pan, Y., Yao, T., Li, Y., & Mei, T. (2020). "X-Linear Attention Networks for Image Captioning". IEEE/CVF Conference on Computer Vision and Pattern Recognition.