**MỤC LỤC**

Trang

[MỞ ĐẦU 1](#_Toc134533467)

[Chương 1 7](#_Toc134533468)

[TỔNG QUAN VỀ TÌM KIẾM ĐỐI TƯỢNG TỪ CAMERA GIÁM SÁT SỬ DỤNG CÂU MÔ TẢ NGÔN NGỮ TỰ NHIÊN TIẾNG VIỆT 7](#_Toc134533469)

[1.1. Giới thiệu chung về hệ thống tìm kiếm 7](#_Toc134533470)

[1.1.1. Sơ đồ khối chung 7](#_Toc134533471)

[1.1.2. Các bài toán cần giải quyết và thách thức có liên quan 8](#_Toc134533472)

[1.1.3. Ứng dụng trong công tác công an 12](#_Toc134533473)

[1.2. Các mô hình học sâu sử dụng cho tìm kiếm hình ảnh đối tượng từ câu mô tả ngôn ngữ tự nhiên tiếng Việt 13](#_Toc134533474)

[1.2.1. Các mô hình cho xử lý hình ảnh 13](#_Toc134533475)

[1.2.2. Các mô hình cho xử lý ngôn ngữ tự nhiên 18](#_Toc134533476)

[1.2.3. Các mô hình so khớp ảnh và câu mô tả 22](#_Toc134533477)

[Chương 2 30](#_Toc134533478)

[MÔ HÌNH HỌC SÂU TRANSFORMER CHO XỬ LÝ ẢNH VÀ XỬ LÝ NGÔN NGỮ 30](#_Toc134533479)

[2.1. Mô hình học sâu Transformer cho xử lý ảnh 30](#_Toc134533480)

[2.1.1. Kiến trúc của mô hình Vision Transformer (ViT) 31](#_Toc134533481)

[2.1.2. Ứng dụng của ViT trong xử lý ảnh 34](#_Toc134533482)

[2.2. Mô hình học sâu Transformer cho xử lý ngôn ngữ 36](#_Toc134533483)

[2.2.1. Biểu diễn đầu vào và đầu ra của BERT 38](#_Toc134533484)

[2.2.2. Quá trình huấn luyện trước BERT (pre-training BERT) 39](#_Toc134533485)

[2.2.3. Quá trình hiệu chỉnh và các thử nghiệm của BERT 41](#_Toc134533486)

[Chương 3 45](#_Toc134533487)

[HỆ THỐNG ĐỀ XUẤT DỰA TRÊN KIẾN TRÚC TRANSFORMER CHO TÌM KIẾM HÌNH ẢNH ĐỐI TƯỢNG SỬ DỤNG CÂU MÔ TẢ NGÔN NGỮ TỰ NHIÊN TIẾNG VIỆT 45](#_Toc134533488)

[3.1. Sơ đồ khối của hệ thống 45](#_Toc134533489)

[3.1.1. Sơ đồ khối 45](#_Toc134533490)

[3.1.2. Trích chọn đặc trưng ảnh người 46](#_Toc134533491)

[3.1.3. Trích chọn đặc trưng câu mô tả 47](#_Toc134533492)

[3.1.4. Huấn luyện mô hình cặp ảnh – câu mô tả 56](#_Toc134533493)

[3.2. Thử nghiệm và kết quả 57](#_Toc134533494)

[3.2.1. Cơ sở dữ liệu thử nghiệm và độ đo đánh giá 57](#_Toc134533495)

[3.2.2. Kết quả thử nghiệm 58](#_Toc134533496)

[KẾT LUẬN 62](#_Toc134533497)

[TÀI LIỆU THAM KHẢO 63](#_Toc134533498)

**DANH MỤC CHỮ VIẾT TẮT**

| **STT** | **Viết tắt** | **Viết đầy đủ tiếng nước ngoài** | **Viết đầy đủ tiếng Việt** |
| --- | --- | --- | --- |
| 1 | ANCT |  | An ninh chính trị |
| 2 | ANTT |  | An ninh trật tự |
| 3 | BERT | Bidirectional Encoder Representations from Transformers | Biểu diễn bộ mã hóa hai chiều từ các mạng biến đổi |
| 4 | Bi-LSTM | Bidirectional long-short term memory | Bộ nhớ ngắn dài hai chiều |
|  | CNN | Convolutional Neuron Network | Mạng nơ-ron tích chập |
| 5 | CNTT |  | Công nghệ thông tin |
| 6 | GNA-RNN | Gated Neural Attention Recurrent Neural Network | Mạng nơ-ron tái diễn với sự chú ý nơ-ron được điều chỉnh |
| 7 | LSTM | Long short-term memory | Mạng trí nhớ ngắn hạn định hướng dài hạn |
| 8 | RNN | Recurrent Neural Network | Mạng nơ-ron tái diễn |
| 9 | VGG | Visual Geometry Group | Nhóm hình học ảnh |
| 10 | ViTAA | Visual-Textual Attributes Alignment | Căn chỉnh các thuộc tính ảnh-văn bản |

**DANH MỤC BẢNG BIỂU**

Trang

[Bảng 1.1: Kết quả của mô hình trên bộ dữ liệu CUHK-PEDES với câu mô tả tiếng Việt so với kết quả mô hình khi sử dụng câu mô tả tiếng Anh [4] 27](#_Toc134482052)

[Bảng 1.2: Kết quả của mô hình trên bộ dữ liệu VnPersonSearch [4] 27](#_Toc134482053)

[Bảng 3.1: Kết quả thực nghiệm trên tập dữ liệu 3000VnPersonSearch 60](#_Toc134482054)

|  |  |
| --- | --- |
| **DANH MỤC HÌNH** |  |

Trang

[Hình 1.1: Sơ đồ chung của hệ thống tìm kiếm đối tượng từ camera giám sát sử dụng câu mô tả ngôn ngữ tự nhiên tiếng Việt 8](#_Toc134542661)

[Hình 1.2: Có nhiều người mặc cùng kiểu dáng trang phục giống nhau 9](#_Toc134542662)

[Hình 1.3: Hình ảnh người đi bộ được thu hình trong những điều kiện 9](#_Toc134542663)

[cường độ ánh sáng khác nhau 9](#_Toc134542664)

[Hình 1.4: Hình ảnh cùng một người nhưng được chụp ở nhiều góc 10](#_Toc134542665)

[Hình 1.5: Hình ảnh trong một khung hình camera chụp sẽ có nhiều đối tượng 11](#_Toc134542666)

[Hình 1.6: Kiến trúc CNN 14](#_Toc134542667)

[Hình 1.7: Kiến trúc mạng VGG-16 16](#_Toc134542668)

[Hình 1.8: Kiến trúc mạng ResNet-50 17](#_Toc134542669)

[Hình 1.9: Sơ đồ kiến trúc mạng nơ-ron hồi quy (RNN) 20](#_Toc134542670)

[Hình 1.10: Kiến trúc một ô nhớ trong LSTM 20](#_Toc134542671)

[Hình 1.11: Kiến trúc mạng Bi-LSTM gồm 2 nhánh forward và backward LSTM 22](#_Toc134542672)

[Hình 1.12: Cấu trúc mạng GNA-RNN để so khớp ảnh với câu mô tả tiếng Việt [3] 23](#_Toc134542673)

[Hình 1.13: Cấu trúc mạng ViTAA [15] 25](#_Toc134542674)

[Hình 1.14: Quy trình tổng thể quá trình huấn luyện mô hình tìm kiếm người bằng câu truy vấn tiếng Việt [4] 26](#_Toc134542675)

[Hình 2.1: Kiến trúc của mô hình Vision Transformer [5] 31](#_Toc134542676)

[Hình 2.2: Mô tả phần Linear Projection 32](#_Toc134542677)

[Hình 2.3: Kiến trúc của Transformer Encoder [5] 33](#_Toc134542678)

[Hình 2.4: Quy trình tổng thể của việc tiền huấn luyện và hiệu chỉnh cho BERT [16] 38](#_Toc134542679)

[Hình 2.5: Biểu diễn đầu vào của BERT [16] 38](#_Toc134542680)

[Hình 2.6: Kết quả thử nghiệm của BERT với tệp dữ liệu SquAD [16] 41](#_Toc134542681)

[Hình 2.7: Kết quả thử nghiệm của BERT với tệp dữ liệu GLUE [16] 42](#_Toc134542682)

[Hình 2.8: Kết quả thử nghiệm của BERT với SWAG [16] 43](#_Toc134542683)

[Hình 3.1: Mô hình tìm kiếm ảnh người dựa trên câu mô tả tiếng Việt. 46](#_Toc134542684)

[Hình 3.2: Trích chọn đặc trưng ảnh người 47](#_Toc134542685)

[Hình 3.3: Trích chọn đặc trưng câu mô tả 48](#_Toc134542686)

[Hình 3.4: Mã hoá câu mô tả tiếng Việt 49](#_Toc134542687)

[Hình 3.5: Kết quả POS Tag thực hiện trên website Underthesea.com 49](#_Toc134542688)

[Hình 3.6: Mô hình mạng TabTransformer 51](#_Toc134542689)

[Hình 3.7: Tạo vector truy vấn, khóa và giá trị 52](#_Toc134542690)

[Hình 3.8: Tính điểm của self-attention 53](#_Toc134542691)

[Hình 3.9: Tính tổng các vector giá trị 54](#_Toc134542692)

[Hình 3.10: Multi-Head Attention 55](#_Toc134542693)

[Hình 3.11: Bộ dữ liệu 3000VnPersonSearch 58](#_Toc134542694)

[Hình 3.12: Sự thay đổi của giá trị mất mát qua các vòng lặp 59](#_Toc134542695)

# MỞ ĐẦU

**1. Tính cấp thiết của đề tài**

Ngày nay sự phát triển nhanh chóng của cuộc cách mạng công nghiệp lần thứ 4 với xu hướng phát triển mạnh của công nghệ Trí tuệ nhân tạo (AI- Artificial Intelligence) và Internet kết nối vạn vật (IoT-Internet of Thing) đang làm thay đổi thế giới từng giờ từng phút. Sự phát triển của cuộc cách mạng lần này tác động mạnh mẽ đến tất cả các lĩnh vực của đời sống xã hội. Việc ứng dụng các thành tựu khoa học công nghệ vào giải quyết công việc là tất yếu để tạo ra sức mạnh mới trong thời đại hiện nay. Đối với công tác bảo vệ an ninh quốc gia và bảo đảm trật tự an toàn xã hội cũng tuân theo quy luật tất yếu này. Nhằm bắt kịp xu thế công nghệ và nâng cao hiệu quả các mặt công tác Công an trong thời kỳ mới của cách mạng công nghiệp 4.0, Bộ Công an đã ban hành nhiều văn bản chỉ đạo Công an các đơn vị, địa phương tăng cường ứng dụng, phát triển CNTT, đặc biệt là trí tuệ nhân tạo vào công tác Công an. Một trong các mục tiêu quan trọng của việc ứng dụng CNTT đối với công tác công an hiện nay là phục vụ việc số hóa dữ liệu dân cư, cải cách hành chính phục vụ người dân và doanh nghiệp, quản lý địa bàn, tuyến lĩnh vực trọng điểm, các đối tượng đặc biệt phục vụ các công tác nghiệp vụ của lực lượng công an.

Nghiên cứu của đề tài xuất phát từ chủ trương về đẩy mạnh nghiên cứu, ứng dụng thành tựu khoa học công nghệ mới là trí tuệ nhân tạo, học máy và học sâu (một nhánh tiên tiến của trí tuệ nhân tạo) vào phục vụ công tác phòng ngừa, phát hiện, đấu tranh với tội phạm của lực lượng Công an. Đề tài tập trung nghiên cứu nghiên cứu ứng dụng kiến trúc mạng học sâu mới cho tìm kiếm đối tượng từ hệ thống camera giám sát sử dụng câu mô tả ngôn ngữ tự nhiên tiếng Việt. Hiện nay, hệ thống camera giám sát ANTT đã được triển khai rộng khắp ở nhiều địa phương trên cả nước. Hệ thống này đã hỗ trợ rất nhiều cho lực lượng chức năng trong công tác đấu tranh bảo vệ an ninh quốc gia và bảo đảm trật tự an toàn xã hội, đặt biệt là trong công tác ngăn chặn, truy xét, truy tìm các đối tượng chạy trốn trên các tuyến, địa bàn trọng điểm. Với những lợi ích to lớn mà hệ thống camera an ninh này đem lại, cần nghiên cứu phát triển các công cụ hỗ trợ tìm kiếm tự động hình ảnh đối tượng từ hệ thống camera giám sát. Thực tiễn công tác cho thấy các thông tin ban đầu mà cơ quan công an nhận được thường là các tin báo tố giác của quần chúng nhân dân, với rất nhiều thông tin mô tả nhận dạng về đối tượng như trang phục, chiều cao, dáng người, độ tuổi, ..). Vì vậy nếu có sự giúp sức của công cụ tìm kiếm hình ảnh đối tượng từ những mô tả qua ban đầu đó qua hệ thống camera an ninh trên địa bàn sẽ giúp cơ quan công an nhanh chóng tìm được đối tượng tình nghi để xác minh và tiến hành các biện pháp nghiệp vụ khác.

Từ những phân tích nêu trên ta thấy được tính ứng dụng cao và lợi ích rất lớn mà đề tài nghiên cứu đem lại cho công tác của lực lượng chức năng hiện nay. Vì vậy nhóm chúng em đề xuất đề tài nghiên cứu “Nghiên cứu mô hình học sâu dựa trên kiến trúc mạng Transfomer ứng dụng trong tìm kiếm đối tượng từ camera giám sát sử dụng câu mô tả ngôn ngữ tự nhiên tiếng Việt”.

**2. Tình hình nghiên cứu liên quan đến đề tài**

Bài toán tìm kiếm hình ảnh người dựa trên trên câu mô tả ngôn ngữ tự nhiên trong những năm trở lại đây đã có rất nhiều các nghiên cứu được công bố trên thế giới. Điểm mấu chốt để giải quyết bài toán là việc thực hiện trích chọn đặc trưng của ảnh người và câu mô tả. Trước đây những nghiên cứu về vấn đề này tập trung vào việc trích chọn đặc trưng toàn cục (global feature), tuy nhiên gần đây có rất nhiều nghiên cứu tập trung vào việc trích chọn đặc trưng cục bộ (local feature) đã nâng cao hiệu quả tìm kiếm một cách đáng kể.

Trong bài báo “Person Search with Natural Language Description” [1] tác giả đã đề xuất mô hình mạng học sâu GNA-RNN (Gated Neural Attention Recurrent Neural Network) để trích chọn đặc trưng và học những đặc trưng của ảnh với các phần liên quan của câu truy vấn, tuy nhiên mô hình mạng này chỉ chú trọng nhiều đến trích chọn đặc trưng ảnh mà không chú trọng về câu truy vấn vì vậy tỉ lệ chính xác top-1 chỉ đạt khoảng 19,05%. Trong bài báo “Pose-Guided Multi-Granularity Attention Network for Text-Based Person Search” [2] nhóm tác giả đã đề xuất mô hình mạng PMA (Pose-guided Multi-granularity Attention), kết hợp với phân ngưỡng thích nghi (Adaptive Threshold) để so khớp các vùng ảnh và từ trong câu mô tả tìm ra sự giống nhau giữa các phần này. Mô hình này cho khả năng chính xác top-1 đạt 47,02% mạng VGG-16 và lên tới 53,81% đối với Res-50.

Những nghiên cứu trên đã góp phần không nhỏ cho việc phát triển các mô hình so khớp hình ảnh và câu mô tả, những mô hình trên đã được áp dụng với các câu mô tả tiếng Anh và đem lại hiệu quả rất tốt. Ngoài ra những mô hình sử dụng câu mô tả ngôn ngữ tiếng Việt cũng đã được phát triển công bố trên các diễn đàn khoa học. Trong bài báo “Person Search by Queried Description in Vietnamese Natural Language” [3], tác giả Phạm Thị Thanh Thủy và các cộng sự đã đề trình bày về một mô hình mạng học sâu mới cho bài toán so khớp hình ảnh và câu mô tả bằng ngôn ngữ tiếng Việt sử dụng mạng GNA-RNN (Gated Neural Attention Recurrent Neural Network) [1] để huấn luyện và cho về kết quả với độ chính xác khoảng 26,85% trên bộ dữ liệu VnPersonSearch do tác giả thu thập và xây dựng và khoảng 18,40% đối với bộ dữ liệu CUHK-PEDES với câu mô tả đã được dịch sang tiếng Việt. Trong bài báo “Person search by natural language description in Vietnamese using pre-trained visual-textual attributes alignment model” [4] Các tác giả đề xuất sử dụng kiến trúc mạng ViTAA để học cách trích xuất thông tin từ các mô tả về người trong tiếng Việt và kết hợp với thông tin hình ảnh để tìm kiếm người. Kết quả thực nghiệm trên các bộ dữ liệu cũng được cải tiến hơn với độ chính xác đạt 61,57% với bộ dữ liệu VnPersonSearch và 52,40% đối với bộ dữ liệu CUHK-PEDES-VN và 61,57% với bộ dữ liệu VnPersonSearch.

Hơn nữa tại Việt Nam hiện nay, mô hình camera giám sát an ninh đã được triển khai trên hầu hết các địa bàn trên cả nước, hỗ trợ rất nhiều cho lực lượng chức năng trong công tác quản lý địa bàn cũng như đấu tranh phòng chống tội phạm. Tuy nhiên với việc lắp đặt một lượng lớn camera trên địa bàn như vậy thì việc trích xuất thông tin từ camera một cách thủ công cũng là một vấn đề không hề dễ dàng, bởi thời gian tìm kiếm lâu, hiệu quả không cao và dễ bị bỏ sót đối tượng. Vì vậy việc nghiên cứu chuyên sâu về tìm kiếm ảnh dựa trên câu mô tả ngôn ngữ tiếng Việt phục vụ công tác đấu tranh phòng chống tội phạm của lực lượng Công an cần được quan tâm hơn nữa.

Những năm gần đây, kiến trúc Transformer được công bố với tốc độ và hiệu quả vượt chội so với các mô hình cũ. Kiến trúc này đã và đang được nghiên cứu và phát triển mạnh mẽ để giải quyết các bài toán liên quan đến trích trọn đặc trưng ngôn ngữ và hình ảnh. Trong bài báo [5] nhóm tác giả đã đề xuất mô hình mạng Vision Transformer (ViT) sử dụng trong việc xử lý hình ảnh có ưu điểm là có thể học được mối quan hệ giữa các đặc trưng và quan hệ giữa chúng một cách riêng biệt nhờ vào sử dụng các attention. Mô hình ViT cho hiệu quả tốt hơn về tốc độ xử lý cũng như chất lượng kết quả đầu ra so với các mô hình CNNs (Convolutional Neural Networks) cổ điển như VGG hay Res-50. Do đó việc các áp dụng các mô hình Transformer vào nghiên cứu mô hình mạng so khớp hình ảnh với câu mô tả tiếng Việt là một ý tưởng mới nhưng hứa hẹn sẽ đem lại một kết quả tốt hơn.

**3. Mục tiêu và nhiệm vụ nghiên cứu**

- Mục tiêu nghiên cứu: Nghiên cứu ứng dụng mạng học sâu dựa trên kiến trúc Transformer cho trích chọn và so khớp đặc trưng ảnh người và câu mô tả, ứng dụng trong tìm kiếm người từ camera giám sát sử dụng câu mô tả ngôn ngữ tự nghiên tiếng Việt.

- Nhiệm vụ nghiên cứu: Để đạt được mục tiêu nghiên cứu nêu trên, các nhiệm vụ nghiên cứu chính được đặt ra trong đề tài như sau:

+ Nghiên cứu các mô hình học sâu dựa trên kiến trúc Transformer cho trích chọn đặc trưng ảnh người và câu mô tả.

+ Nghiên cứu đề xuất mô hình học sâu dựa trên kiến trúc Transformer cho ước lượng tự động sự tương đồng giữa câu mô tả và hình ảnh đối tượng.

+ Đánh giá thử nghiệm mô hình đề xuất cho tìm kiếm hình ảnh người từ câu mô tả dựa trên các bộ cơ sở dữ liệu tiêu chuẩn.

**4. Đối tượng, phạm vi nghiên cứu**

***4.1. Đối tượng nghiên cứu***

- Việc trích chọn và biểu diễn đặc trưng ảnh và đặc trưng câu mô tả ngôn ngữ tự nhiên.

- Sự tương đồng giữa đặc trưng ảnh và đặc trưng câu mô tả ngôn ngữ tự nhiên.

***4.2. Phạm vi nghiên cứu***

- Trích chọn và biểu diễn đặc trưng cho ảnh người được khoanh vùng từ các khung hình thu nhận bởi camera giám sát.

- Trích chọn và biểu diễn đặc trưng câu mô tả ngôn ngữ tự nhiên tiếng Việt. Câu mô tả về diện mạo của đối tượng như đầu tóc, mũ, quần áo, giày dép, phụ kiện.

- Các mô hình học sâu dựa trên kiến trúc Transformer cho trích chọn và biểu diễn đặc trưng ảnh người, câu mô tả tiếng Việt và so khớp giữa vector đặc trưng ảnh và vector đặc trưng câu mô tả.

- Ngôn ngữ lập trình Python, thư viện học máy Pytorch, Tensorflow, Keras dùng cho xây dựng và huấn luyện, kiểm thử các mạng học sâu đề xuất;

- Các bộ cơ sở dữ liệu tiêu chuẩn có chứa hình ảnh người trích xuất từ camera giám sát và câu mô tả tiếng Việt tương ứng để thử nghiệm hệ thống.

**5. Phương pháp nghiên cứu**

- Phương pháp nghiên cứu lý thuyết:

+ Nghiên cứu tài liệu có liên quan đến bài toán tìm kiếm hình ảnh người dựa trên câu mô tả ngôn ngữ tự nhiên.

+ Nghiên cứu lý thuyết chuyên sâu về các mô hình mạng học sâu ứng dụng trong các bài toán xử lý ảnh và xử lý ngôn ngữ tự nhiên có liên quan đến bài toán nghiên cứu của đề tài.

- Phương pháp nghiên cứu thực nghiệm:

+ Sử dụng các công cụ, ngôn ngữ lập trình, thư viện mã nguồn mở để xây dựng mô hình học sâu dựa trên kiến trúc Transformer cho tìm kiếm hình ảnh người theo câu mô tả ngôn ngữ tiếng Việt.

+ Chạy thử nghiệm mô hình đề xuất trên các bộ cơ sở dữ liệu thử nghiệm và đánh giá kết quả thu được.

**6. Những đóng góp mới và ý nghĩa khoa học**

Các nghiên cứu liên quan đã có sử dụng mạng học sâu như VGG, Resnet... để giải quyết bài toán và đã đạt được một số kết quả nhất định. Tuy nhiên hiện nay các mạng học sâu dựa trên kiến trúc Transformer đang thu hút nhiều sự quan tâm nghiên cứu và đã chứng minh tính hiệu quả vượt trội hơn so với các mô hình CNN ở cả bài toán liên quan đến xử lý ảnh và xử lý ngôn ngữ tự nhiên. Việc nghiên cứu và vận dụng các mạng học sâu dựa trên kiến trúc Transformer cho bài toán tìm kiếm hình ảnh người theo câu mô tả ngôn ngữ tự nhiên tiếng Việt là đóng góp mới.

Việc nghiên cứu mô hình học sâu dựa trên kiến trúc Transformer cho bài toán tìm kiếm hình ảnh người dựa trên câu mô tả ngôn ngữ tiếng Việt là cơ sở quan trọng để phát triển hệ thống tìm kiếm tự động đối tượng từ camera giám sát theo câu mô tả về đối tượng. Các kết quả thử nghiệm đối với mô hình đề xuất trên các bộ cơ sở dữ liệu thử nghiệm tiêu chuẩn là căn cứ khoa học có giá trị để tiếp tục nghiên cứu, phát triển hệ thống ứng dụng thực tiễn phục vụ công tác Công an, góp phần tìm kiếm nhanh, cải thiện lớn về mặt thời gian, chi phí và công sức; ứng phó kịp thời với tình hình tội phạm diễn biến phức tạp hiện nay, gắn liền với cuộc phát triển của cuộc cách mạng khoa học kỹ thuật trên thế giới nói chung và khoa học Công an nói riêng.

**7. Bố cục đề tài nghiên cứu**

Ngoài phần mở đầu, kết luận, kiến nghị, danh mục tài liệu tham khảo, đề tài bố cục gồm 3 chương:

Chương 1: Tổng quan về tìm kiếm đối tượng từ hệ thống camera giám sát sử dụng câu mô tả ngôn ngữ tự nhiên tiếng Việt.

Chương 2: Mô hình học sâu Transformer cho xử lý ảnh và xử lý ngôn ngữ.

Chương 3: Hệ thống đề xuất cho tìm kiếm đối tượng từ camera giám sát sử dụng câu mô tả ngôn ngữ tự nhiên tiếng Việt.

# Chương 1

# TỔNG QUAN VỀ TÌM KIẾM ĐỐI TƯỢNG TỪ CAMERA GIÁM SÁT SỬ DỤNG CÂU MÔ TẢ NGÔN NGỮ TỰ NHIÊN TIẾNG VIỆT

## 1.1. Giới thiệu chung về hệ thống tìm kiếm

### 1.1.1. Sơ đồ khối chung

Một hệ thống tìm kiếm đối tượng từ camera giám sát sử dụng câu mô tả ngôn ngữ tự nhiên tiếng Việt trong thực tế cần có những thành phần chính sau đây:

- Đầu tiên, thu thập các video từ các hệ thống camera sau đó thực hiện phân tích và xử lý chúng để phát hiện và theo vết đối tượng người trong video. Quá trình này cho chúng ta biết được vị trí xuất hiện và khung hình ảnh của người đó trong video. Các hình ảnh trích xuất được từ bước này sẽ được lưu vào trong cơ sở dữ liệu hình ảnh phục vụ cho các yêu cầu truy vấn sau này.

- Khi cần tìm kiếm một đối tượng cụ thể người dùng sẽ nhập một câu truy vấn mô tả về đối tượng cần tìm. Hệ thống sẽ xử lý câu mô tả đó sau đó tìm kiếm và trả về các ảnh đối tượng phù hợp với câu mô tả.

- Hệ thống thực hiện so khớp câu mô tả và các hình ảnh trong kho dữ liệu hình ảnh thông qua các bước: (1) trích chọn đặc trưng của ảnh người; (2) trích chọn đặc trưng của câu mô tả; (3) so khớp giữa hai đặc trưng vừa trích chọn được.

Sơ đồ sau thể hiện được mô hình chung của hệ thống tìm kiếm người dựa trên câu mô tả trong thực tế. Qua sơ đồ ta đã hiểu về các nguyên lý hoạt động cơ bản của hệ thống tìm kiếm người bằng câu mô tả, tiếp theo để phục vụ nghiên cứu chúng ta sẽ tìm hiểu về các bài toán cần giải quyết và những vấn đề khi xây dựng hệ thống này.

Diagram

Description automatically generated

Hình 1.1: Sơ đồ chung của hệ thống tìm kiếm đối tượng từ camera giám sát sử dụng câu mô tả ngôn ngữ tự nhiên tiếng Việt

Để xây dựng hệ thống trên tìm kiếm đối tượng từ camera giám sát sử dụng câu mô tả ngôn ngữ tự nhiên tiếng Việt, chúng ta cần giải quyết những bài toán cơ bản sau:

- Phát hiện và theo vết đối tượng người trong video thu nhận từ các camera. Hiệu quả phát hiện và theo vết phụ thuộc vào độ phức tạp của dữ liệu video, sự đa dạng về kích thước và hình dạng của đối tượng, đối tượng bị che khuất và bị nhiễu…

- So khớp ảnh trong cơ sở dữ liệu với câu mô tả. Nhiều bài toán con liên quan đến vấn đề so khớp giữa ảnh và câu mô tả cần được giải quyết: Các bài toán liên quan đến xử lý và trích chọn đặc trưng ảnh; xử lý và trích chọn đặc trưng câu mô tả; biểu diễn các đặc trưng ảnh và câu mô tả trên cùng một không gian để đo sự giống nhau giữa chúng.

### 1.1.2. Các bài toán cần giải quyết và thách thức có liên quan

Từ những bài toán đặt ra bên trên chúng ta sẽ phân tích và làm rõ các thách thức đặt ra đối với việc giải quyết những bài toán này từ đó sẽ giúp định hướng cho việc xây dựng hệ thống tìm kiếm đối tượng từ camera giám sát sử dụng câu mô tả ngôn ngữ tự nhiên tiếng Việt.

Vấn để thách thức đặt ra trong phát hiện và theo vết đối tượng và xử lý ảnh:

- Khả năng thích ứng với đa dạng các tình huống: Một thách thức khác đối với các bài toán này là khả năng thích ứng với đa dạng các tình huống khác nhau, có thể có nhiều đối tượng có ngoại hình giống nhau về hình dạng màu sắc, tương tự nhau xuất hiện trong cùng một khung ảnh.

A collage of a person

Description automatically generated with low confidence

Hình 1.2: Có nhiều người mặc cùng kiểu dáng trang phục giống nhau

- Điều kiện ánh sáng và môi trường: Các bài toán phân tích và xử lý video đòi hỏi điều kiện ánh sáng và môi trường tốt để đảm bảo độ chính xác của kết quả. Tuy nhiên, trong thực tế, điều kiện ánh sáng và môi trường thường không hoàn hảo như yếu tố ánh sáng thay đổi theo khung giờ trong ngày hay do điều kiện thời tiết làm ảnh hưởng đến chất lượng ảnh mà camera thu thập được gây khó cho việc xác định đối tượng trong ảnh.

A picture containing text, person, group, people

Description automatically generated

Hình 1.3: Hình ảnh người đi bộ được thu hình trong những điều kiện

cường độ ánh sáng khác nhau

- Ngoài ra việc nhận dạng những đối tượng che khuất cũng là thách thức lớn trong việc xử lý, phát hiện và theo vết sự che khuất xảy ra khi đối tượng người bị che khuất bởi các vật thể khác trong khung hình, ví dụ như tường, cây, bức tường, xe cộ, v.v. Hơn nữa góc quan sát cũng ảnh hưởng đến phạm vi quan sát và độ chi tiết của hình ảnh; một góc quan sát rộng sẽ cho phép camera ghi được nhiều hình ảnh hơn, giảm thiểu điểm mù và tăng khả năng giám sát tuy nhiên độ chi tiết về hình ảnh sẽ giảm và ngược lại.

A collage of a person

Description automatically generated with medium confidence

Hình 1.4: Hình ảnh cùng một người nhưng được chụp ở nhiều góc

- Chuyển động nhanh đặt ra thách thức trong bài toán phát hiện và theo vết đối tượng người trong video. Khi đối tượng di chuyển quá nhanh, nó sẽ xuất hiện ở nhiều vị trí khác nhau trong các khung hình liên tiếp, và việc phát hiện và theo dõi đối tượng sẽ trở nên phức tạp hơn. Đặc biệt, khi chuyển động nhanh kết hợp với sự che khuất, thách thức sẽ trở nên cao hơn nữa.

- Độ chính xác: Một trong những thách thức lớn nhất đối với các bài toán này là độ chính xác của kết quả. Các thuật toán và công nghệ phải đảm bảo độ chính xác cao để tránh việc phát hiện nhầm hoặc bỏ sót đối tượng. Trong một tình huống giám sát nếu đối tượng cần tìm kiếm nằm trong một khung cảnh lộn xộn, bừa bộn các chi tiết của đối tượng chồng chéo vào nhau thì cần thuật toán riêng để tách đối tượng khỏi nền đó và so sánh đối chiếu về các đặc điểm của đối tượng là người bỏ đi khung cảnh không có liên quan ở xung quanh.

A collage of people walking

Description automatically generated with medium confidence

Hình 1.5: Hình ảnh trong một khung hình camera chụp sẽ có nhiều đối tượng

- Độ phức tạp tính toán: Các bài toán phân tích và xử lý video đòi hỏi tính toán rất lớn bởi dữ liệu video có kích thước lớn, phải xử lý theo từng khung hình (frame) để phát hiện vè theo dõi đối tượng. Vì vậy, cần phải sử dụng các máy tính có cấu hình mạnh để đảm bảo tốc độ xử lý đủ nhanh.

Thách thức đặt ra với việc xử lí ngôn ngữ tự nhiên tiếng Việt:

- Xử lý câu mô tả ngôn ngữ tiếng Việt là một trong những phần khó nhất trong việc xây dựng hệ thống này, vì để biểu diễn câu mô tả này theo một dạng mà máy tính có thể hiểu được đầy đủ ngữ nghĩa của nó rất khó khăn.

- Để máy tính có thể hiểu được nội dung của câu mô tả một cách rõ ràng nhất vấn để đầu tiên phải giải quyết đó là bước tiền xử lý văn bản:

+ Những câu mô tả có thể có chứa nhiễu (những lý tự thừa, lỗi không liên quan đến việc mô tả đối tượng). Làm sạch nhiễu sẽ giúp nâng cao hiệu quả của việc tìm kiếm sau này. Ngoài những từ nhiễu chúng ta cũng cần loại bỏ những từ dừng (stopwords) vì chúng cũng rất hay xuất hiện trong văn bản tuy nhiên lại không mang nhiều ý nghĩa đối với mô tả đối tượng trong bài toán.

+ Tiếp theo ta cần phải phân đoạn từ vì tiếng Việt là một ngôn ngữ đặc biệt, các từ ghép trong tiếng Việt được ghép từ nhiều từ đơn với nhau và mang ý nghĩa không giống với các từ đơn thành phần. Do đó, nếu ta không phân đoạn từ ngữ trong câu một cách chính xác thì có thể làm máy hiểu sai ý nghĩa của câu mô tả. Ngoài ra, để đảm bảo tính thống nhất, chính xác trong xử lý câu mô tả chúng ta cần phải chuẩn hóa từ ngữ trong câu mô tả (tên riêng, từ ngữ địa phương, từ viết tắt) về một từ ngữ chung cho tất cả các câu mô tả.

+ Ngoài ra, xử lý những từ đồng nghĩa cũng là một thách thức đối với những câu mô tả ngôn ngữ tiếng Việt bởi đây là một ngôn ngữ rất phong phú với cùng một đối tượng nhưng những người khác nhau sẽ có những mô tả khác nhau. Ví dụ như “áo thun màu nâu”, “áo cộc tay màu nâu”, “áo tay ngắn sẫm màu” đều có thể là mô tả cùng một đối tượng cần tìm kiếm.

Thách thức đặt ra với việc so khớp hình ảnh tìm kiếm và câu mô tả ngôn ngữ tự nhiên tiếng Việt:

- Để thực hiện tìm kiếm người dựa trên câu mô tả ngôn ngữ tự nhiên chúng ta cần phải tìm được một không gian chung để biểu diễn đặc trưng của câu mô tả và hình ảnh từ đó xác định độ liên quan của chúng với nhau.

- Đối với câu mô tả chúng ta phải xác định xem nội dung câu mô tả mô tả đặc điểm gì của đối tượng, đặc điểm bề ngoài (quần, áo, kiểu tóc, phụ kiện,…) hay hoạt động đối tượng (đang chơi bóng, đi bộ,…). Xác định mức độ quan trọng của từ ngữ mô tả đối tượng, xây dựng bộ mô tả quan hệ giữa người và ảnh.

- Việc xác định vị trí đặc trưng trên ảnh đúng với từ ngữ trong câu mô tả cũng là một công việc khó khăn, có thể mỗi đặc trưng của ảnh có trùng khớp với một đặc trưng của một vùng ảnh nhất định tuy nhiên khi ghép tất cả các đặc trưng này lại thì lại không trùng khớp giữa ảnh và câu mô tả nữa.

- Bên cạnh đó kích thước của cơ sở dữ liệu cũng đặt ra một vấn đề đối với việc so khớp hình ảnh và câu mô tả bởi nếu cơ sở dữ liệu quá lớn, việc tìm kiếm sẽ trở nên phức tạp và tốn nhiều thời gian. Tuy nhiên, nếu cơ sở dữ liệu quá nhỏ, độ chính xác của kết quả tìm kiếm cũng sẽ bị giảm.

### 1.1.3. Ứng dụng trong công tác công an

Hiện nay trong lĩnh vực đảm bảo an ninh trật tự nói chung và ngành khoa học an ninh của Công an nói riêng thì việc ứng dụng sự tìm kiếm của máy tính vào công tác điều tra, nhận dạng để khoanh vùng và truy vết đối tượng nghi vấn lẩn trốn, phương tiện giao thông đi lại đang là nhu cầu hết sức cần thiết và cấp bách.

Cùng với sự phát triển vượt bậc của khoa học hiện nay, số lượng camera giám sát đã và đang được trang bị rất nhiều về số lượng và chất lượng trên mọi nẻo đường, tuyến phố, ngóc ngách của từng ngõ ra vào…phục vụ cho nhu cầu giám sát an ninh trật tự, giao thông, truy tìm các đối tượng phạm tội được camera ghi hình ảnh lại…Để tối ưu hóa dữ liệu từ camera giám sát, rút ngắn thời gian trích xuất video, thu gọn thời gian tìm kiếm phục vụ cho bài toán truy tìm đối tượng phạm tội qua câu mô tả của lực lượng Công an thì việc tích hợp trí tuệ nhân tạo vào là vô cùng cần thiết.

Đối tượng phạm tội có chiều hướng ngày một gia tăng, xu hướng hoạt động và quá trình lẩn trốn ngày càng tinh vi hơn nên việc ứng dụng camera giám sát vào là vô cùng hiệu quả có thể thực hiện cả về ngày đêm, nắng mưa…, thu được hình ảnh nghi vấn về đối tượng tình nghi là tội phạm, đảm bảo cho sự nhanh chóng mất ít thời gian và kinh phí, phục vụ cho công tác truy bắt đối tượng đạt hiệu quả cao, đem lại niềm tin cho quần chúng nhân dân, đảm bảo tình hình an ninh trật tự luôn được giữ vững.

Hệ thống tìm kiếm hình ảnh đối tượng bằng câu mô tả tiếng Việt sẽ rất tiện dụng trong tình huống: (1) hình ảnh của đối tượng cần truy tìm không sẵn có nhưng tin báo của quần chúng, hay của người bị hại về đối tượng lại sẵn có hơn; (2) Khả năng đối tượng di chuyển qua nhiều địa bàn khác nhau trên phạm vị rộng. Hệ thống giúp truy vết nhanh và chính xác là rất cần thiết nhằm giúp ngắn thời gian điều tra, phá án của lực lượng chức năng.

## 1.2. Các mô hình học sâu sử dụng cho tìm kiếm hình ảnh đối tượng từ câu mô tả ngôn ngữ tự nhiên tiếng Việt

### 1.2.1. Các mô hình cho xử lý hình ảnh

Các kỹ thuật học sâu đã trở thành một trong những công cụ quan trọng trong lĩnh vực trí tuệ nhân tạo và đã được áp dụng rộng rãi trong các bài toán phân lớp, nhận dạng ảnh, xử lý ngôn ngữ tự nhiên, và nhiều lĩnh vực khác. Các kiến trúc mạng học sâu dựa trên mạng nơ-ron tích chập CNN (Convolutional Neural Network) như VGG16, ResNet50 và nhiều mô hình khác đã đạt được kết quả ấn tượng trong nhiều bài toán phức tạp và đã trở thành tiêu chuẩn cho việc xây dựng các mô hình học sâu hiệu quả.

CNN là một loại mạng nơ-ron sử dụng trong lĩnh vực thị giác máy tính, đặc biệt là trong xử lý ảnh. CNN bao gồm nhiều lớp nơ-ron tích chập (convolutional layer), tầng gộp (pooling layer) và tầng kết nối đầy đủ (fully connected layer). Tầng tích chập giúp trích xuất các đặc trưng từ ảnh đầu vào, tầng gộp giảm kích thước của đầu ra từ tầng tích chập và giúp giảm thiểu overfitting, tầng kết nối đầy đủ giúp phân loại đầu ra dựa trên các đặc trưng đã được trích xuất. Tuy nhiên, kiến trúc CNN có thể khác nhau tùy thuộc vào bài toán cụ thể và được điều chỉnh để đạt được hiệu quả cao nhất.

Diagram, engineering drawing

Description automatically generated

Hình 1.6: Kiến trúc CNN

Một mô hình nhận dạng phân loại ảnh sử dụng kiến trúc CNN được thực hiện trong 3 giai đoạn chính như sau:

- Tầng tích chập (convolutional layer): là quá trình tính toán trong đó sử dụng nhiều bộ lọc đặc trưng trượt trên các ma trận điểm ảnh để trích xuất các đặc trưng tương ứng của dữ liệu ảnh. Với mỗi tầng, giá trị tích chập cho một điểm ảnh đầu ra được tính theo công thức (1.1), trong đó 𝑌[𝑚, 𝑛] là giá trị đầu ra cho tầng tiếp theo, 𝑚 và 𝑛 là chỉ số của hàng và cột của ma trận đầu ra, 𝑓 biểu thị cho ma trận điểm ảnh đầu vào, ma trận lọc được biểu diễn là ℎ và ∗ là phép tích chập.

𝑌[𝑚, 𝑛] = (𝑓 ∗ ℎ)[𝑚, 𝑛] = ∑ ∑𝑘 ℎ[𝑗, 𝑘]𝑓[𝑚 − 𝑗, 𝑛 − 𝑘]. 𝑗 (1.1)

Sau mỗi quá trình tích chập, ta thu được ma trận các giá trị đặc trưng. Sau đó, cần sử dụng một hàm kích hoạt (activation function) để phi tuyến hóa các giá trị đó. Trong hầu hết các nghiên cứu trước đây, **sigmoid** và **tanh** là những hàm kích hoạt phi tuyến được sử dụng phổ biến nhất. Tuy nhiên, gần đây ReLU (rectified linear unit) [6] đã được sử dụng rộng rãi và phổ biến hơn. ReLU có tốc độ tính toán nhanh hơn nhờ đạo hàm chỉ có 2 giá trị {0, 1} và không có lũy thừa cơ số e như hàm sigmoid. Hàm ReLU được tính theo công thức (1.2).

𝑅𝑒𝐿𝑈(𝑥) = max(0, 𝑥) (1.2)

với điều kiện: 𝑅𝑒𝐿𝑈(𝑥) = {1 𝑖𝑓 𝑥 > 0; 0 𝑜𝑡ℎ𝑒𝑟𝑤𝑖𝑠𝑒}.

* Tầng tổng hợp (pooling layer): giảm số chiều của bản đồ đặc trưng bằng phép toán lấy mẫu con (subsampling), thường sử dụng phép toán max pooling hoặc average pooling.
* Tầng kết nối đầy đủ (fully connected layer): làm phẳng ma trận đặc trưng thành một vector đặc trưng và sử dụng các kết nối đầy đủ giữa các tầng để phân loại đầu ra dựa trên hàm kích hoạt Softmax.

Mô hình phân lớp sử dụng kiến trúc CNN thực hiện trích xuất đặc trưng và phân loại ảnh. Tầng tích chập và tầng gộp trích xuất đặc trưng của ảnh, trong khi các tầng kết nối đầy đủ hoạt động như một bộ phân loại trên các đặc trưng này. Điều quan trọng là CNN có khả năng giảm số lượng tham số trong mạng nơ-ron và tăng độ phức tạp phát hiện hình ảnh, giúp phát hiện các đặc trưng bậc cao của dữ liệu ảnh. Càng ở những tầng tích chập về sau càng có khả năng phát hiện các đường nét phức tạp, đã rõ ràng hình thù và thậm chí là cấu thành vật thể, đây được gọi là những đặc trưng bậc cao (high level) của dữ liệu ảnh. Các lớp được kết nối đầy đủ sẽ học cách sử dụng kết hợp những đặc trưng này để phân loại hình ảnh một cách chính xác.

Một số kiến trúc mạng CNN được dùng phổ biến:

* VGG (Visual Geometry Group): Là một mạng nơ-ron tích chập (CNN) được giới thiệu bởi nhóm nghiên cứu VGG của Đại học Oxford vào năm 2014. Mô hình này được phát triển để giải quyết vấn đề nhận dạng hình ảnh trong cuộc thi ImageNet Large Scale Visual Recognition Challenge (ILSVRC) và đã đạt được kết quả rất tốt. Mô hình VGG sử dụng nhiều lớp tích chập với kích thước nhân lớn và các lớp max pooling với bước nhảy cố định. Mô hình này có số lượng tham số lớn, tuy nhiên độ chính xác của nó rất cao, vượt trội so với các mô hình trước đó.

Một số phiên bản của mô hình VGG được đề xuất, có kiến trúc khác nhau, tuy nhiên phiên bản VGG-16 là phiên bản phổ biến nhất. Kiến trúc mô hình VGG-16 được Karen Simonyan và Andrew Zisserman giới thiệu vào năm 2014 trong [7]. VGG-16 bao gồm 16 tầng, trong đó 13 là tầng tích chập và 3 tầng kết nối đầy đủ. Phiên bản này sử dụng các khối tích chập (convolutional block) được hình thành bởi việc kết hợp các tầng tích chập và tầng gộp cực đại (max pooling) để giảm kích thước bản đồ đặc trưng. Mỗi khối tích chập có các bộ lọc kích thước nhỏ 3x3, giúp giảm số lượng tham số và khối lượng tính toán. Đầu vào của mô hình là một hình ảnh có kích thước cố định và đầu ra là một vector dài 1000 phần tử, tương ứng với 1000 lớp đối tượng khác nhau trong tập dữ liệu ImageNet. VGG-16 đã được sử dụng trong nhiều ứng dụng học máy và thị giác máy tính, và là một trong những kiến trúc CNN phổ biến nhất hiện nay.

Diagram

Description automatically generated

Hình 1.7: Kiến trúc mạng VGG-16

* Inception: Kiến trúc mô hình Inception (hay còn gọi là GoogLeNet) được giới thiệu bởi Christian Szegedy và đồng nghiệp tại Google vào năm 2014 trong bài báo [8]. Mục tiêu của Inception là tối ưu hóa việc chọn kích thước bộ lọc cho các tầng tích chập của mô hình. Thay vì sử dụng các bộ lọc có kích thước cố định, Inception sử dụng các bộ lọc với nhiều kích thước khác nhau (ví dụ: 1x1, 3x3, 5x5) trong mỗi tầng tích chập. Sau đó, các bản đồ đặc trưng được gộp lại với nhau (có thể thông qua việc sử dụng cả max pooling và average pooling) để tạo ra các bản đồ đặc trưng có kích thước khác nhau.
* ResNet: Là một kiến trúc mạng neural convolutional (CNN) rất sâu được giới thiệu bởi Kaiming He và các đồng nghiệp vào năm 2015 [9]. Kiến trúc ResNet có khả năng giải quyết vấn đề mất mát thông tin (vanishing gradient [10]) trong các mô hình sâu bằng cách sử dụng các khối residual (residual blocks) cho phép dữ liệu truyền qua các tầng mà không bị ảnh hưởng bởi mất mát thông tin. ResNet sử dụng kiến trúc residual block được hình thành bởi việc kết hợp các tầng tích chập và tầng kết nối đầy đủ với việc bỏ qua một số tầng trung gian, tạo ra các shortcut connection giữa các tầng để truyền dữ liệu trực tiếp. Kiến trúc ResNet đã đạt được nhiều kết quả ấn tượng trong các bài toán học máy và thị giác máy tính, và được sử dụng rộng rãi trong nhiều ứng dụng thực tế. Có nhiều biến thể của kiến trúc ResNet với số lớp khác nhau như ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152, …Các chỉ số theo sau ResNet dùng để chỉ kiến trúc ResNet với số lớp nhất định.

ResNet-50 có 50 lớp, bao gồm 49 lớp tích chập và một lớp kết nối đầy đủ để phân loại. Kiến trúc của ResNet-50 được hình thành bởi các khối residual (residual blocks) được kết hợp với các tầng gộp cực đại (max pooling) và tầng kết nối đầy đủ. Với sự kết hợp này, ResNet-50 có khả năng học được các đặc trưng phức tạp của ảnh và đạt được kết quả tốt trong các bài toán học máy và thị giác máy tính, bao gồm cả nhận dạng ảnh và phân loại ảnh. ResNet-50 là một trong những kiến trúc CNN phổ biến nhất hiện nay và thường được sử dụng như một mô hình tham chiếu để so sánh với các kiến trúc khác.

Chart, bar chart

Description automatically generated

Hình 1.8: Kiến trúc mạng ResNet-50

* MobileNet: Kiến trúc mạng CNN được giới thiệu bởi Andrew G. Howard và đồng nghiệp vào năm 2017, với mục tiêu thiết kế một mô hình nhẹ, nhanh và có khả năng chạy trên các thiết bị di động với tài nguyên tính toán hạn chế. MobileNet sử dụng các kỹ thuật như Depthwise Separable Convolution để giảm số lượng tham số và tính toán của mô hình. Để giảm số lượng tham số, MobileNet sử dụng kiến trúc Depthwise Separable Convolution thay cho Convolution thông thường. Trong Depthwise Separable Convolution, một bước tích chập được chia thành hai bước: đầu tiên là tích chập riêng cho từng kênh (depthwise convolution) và sau đó là một tích chập tuyến tính đầy đủ (pointwise convolution) để kết hợp các đặc trưng được trích xuất từ các kênh. Sử dụng kiến trúc này giúp giảm số lượng tham số và tính toán của mô hình. MobileNet cũng sử dụng các kỹ thuật như Global Average Pooling và hàm kích hoạt ReLU6 để tối ưu hóa mô hình. MobileNet đã đạt được kết quả tốt trong nhiều bài toán như phân loại hình ảnh và nhận dạng vật thể. Do tính nhẹ và tốc độ nhanh, MobileNet được sử dụng rộng rãi trong các ứng dụng thực tế yêu cầu tính toán trên các thiết bị có tài nguyên hạn chế như điện thoại di động và máy tính nhúng.
* DenseNet (Densely Connected Convolutional Networks): Kiến trúc mạng CNN được giới thiệu bởi Gao Huang, Zhuang Liu và đồng nghiệp vào năm 2016 trong bài báo [12]. Kiến trúc DenseNet có đặc điểm kết nối mạng gần như đầy đủ giữa các lớp, cho phép truyền dữ liệu đồng thời qua các lớp với số lượng tham số và bộ nhớ ít hơn so với các mô hình CNN khác. Trong kiến trúc DenseNet, các lớp được kết nối với nhau thông qua việc ghép nối (concatenation) đầu vào và đầu ra của các lớp. Việc ghép nối này giúp mô hình học được một số đặc trưng ở các lớp trung gian của kiến trúc. Điều này cũng giúp cho mô hình trở nên nhỏ gọn hơn vì mỗi lớp chỉ cần học một số đặc trưng mới thay vì phải học lại tất cả các đặc trưng đã học ở các lớp trước đó. Một ưu điểm nổi bật của DenseNet là giúp khắc phục vấn đề vanishing gradient khi huấn luyện mạng sâu, cũng như giúp mô hình đạt kết quả tốt trong các bài toán phân loại ảnh. Do tính năng kết nối đầy đủ và hiệu quả, DenseNet đã được áp dụng rộng rãi trong nhiều ứng dụng như nhận dạng vật thể, phân loại ảnh, xử lý ngôn ngữ tự nhiên và xử lý âm thanh.

### 1.2.2. Các mô hình cho xử lý ngôn ngữ tự nhiên

Các mô hình học sâu phổ biến trong xử lý ngôn ngữ tự nhiên bao gồm mạng tích chập CNN và mạng nơ-ron hồi quy RNN. CNN được sử dụng để trích xuất các đối tượng cấp cao hơn trong câu thông qua bộ lọc và max-pooling, trong khi RNN được sử dụng để mô hình hóa các phụ thuộc ngữ cảnh trong dữ liệu câu thông qua chuỗi các thông tin. Đầu vào của RNN thường là one-hot vector hoặc word embedding. So sánh với các mô hình truyền thống như SVM và hồi quy logistic, các mô hình học sâu đã đạt được kết quả vượt trội trong các tác vụ liên quan đến ngôn ngữ.

Mạng tích chập CNN (Convolutional Neural Network): CNN một kiến trúc mạng nơ-ron sử dụng rộng rãi trong thị giác máy tính, xử lý ảnh, âm thanh và cả xử lý ngôn ngữ tự nhiên. CNN áp dụng các bộ lọc để trích xuất đặc trưng từ dữ liệu đầu vào và sử dụng các lớp tích chập và lớp pooling để giảm kích thước đầu ra của các lớp. Một trong những đặc điểm nổi bật của CNN là khả năng học được các đặc trưng cấp cao và tổng hợp các đặc trưng đó lại để đưa ra kết quả chính xác.

Trong xử lý ngôn ngữ tự nhiên, CNN thường được sử dụng để mô hình hóa các dữ liệu có cấu trúc như câu hoặc văn bản. Đầu vào của CNN thường là ma trận nhúng từ, được tạo ra bằng cách biểu diễn mỗi từ dưới dạng vector. Sau đó, CNN sử dụng các bộ lọc để trích xuất các đặc trưng từ ma trận này, qua đó tạo ra một ma trận đặc trưng. Cuối cùng, kết quả đầu ra của CNN được tạo ra bằng cách áp dụng lớp pooling để thu nhỏ kích thước của ma trận đặc trưng và sau đó truyền qua các lớp nơ-ron để tạo ra dự đoán. Một số ứng dụng của CNN trong xử lý ngôn ngữ tự nhiên bao gồm phân loại văn bản, trích xuất thông tin và dịch máy. So với các phương pháp truyền thống, CNN cho kết quả tốt hơn và đạt được kết quả tốt nhất trong các bài toán yêu cầu xử lý dữ liệu có cấu trúc và cả không có cấu trúc.

Mạng nơ-ron hồi quy RNN (Recurrent Neuron Network): RNN là một loại mạng nơ-ron được sử dụng phổ biến trong lĩnh vực xử lý ngôn ngữ tự nhiên và dữ liệu chuỗi. Mạng RNN có khả năng ghi nhớ thông tin trạng thái trước đó và sử dụng nó để tính toán các đầu ra tiếp theo. Điều này làm cho mô hình RNN phù hợp để mô hình hóa các phụ thuộc ngữ cảnh với các đầu vào có độ dài tùy ý.

Một trong những ứng dụng phổ biến của mạng RNN là mô hình ngôn ngữ, trong đó mô hình được huấn luyện trên một tập dữ liệu văn bản để dự đoán từ tiếp theo trong một chuỗi. Mô hình RNN cũng được sử dụng trong nhiều ứng dụng khác nhau như dịch máy, nhận dạng giọng nói, phân tích tín hiệu, và nhiều lĩnh vực khác. Các đầu vào của mô hình RNN thường là các vectơ one-hot hoặc word embedding. Ngoài ra, có nhiều biến thể của mô hình RNN, bao gồm LSTM (Long Short-Term Memory) và GRU (Gated Recurrent Unit), được thiết kế để giải quyết vấn đề vanishing gradient và exploding gradient trong quá trình huấn luyện mạng RNN.

A picture containing text, clock

Description automatically generated

Hình 1.9: Sơ đồ kiến trúc mạng nơ-ron hồi quy (RNN)

Mạng LSTM (Long Short-Term Memory): là một dạng mạng nơ-ron hồi quy RNN (Recurrent Neural Network) được thiết kế để xử lý các dữ liệu có mối quan hệ phụ thuộc thời gian dài, đặc biệt là trong lĩnh vực xử lý ngôn ngữ tự nhiên. Một trong những vấn đề chính của mạng RNN truyền thống là hiện tượng vanishing gradient, khi các đạo hàm giảm dần theo chiều sâu của mạng, làm cho việc huấn luyện mạng trở nên khó khăn và kém hiệu quả. Để giải quyết vấn đề này, mạng LSTM được ra đời với kiến trúc có khả năng tự điều chỉnh các trọng số và bộ nhớ dài hạn (long-term memory) và ngắn hạn (short-term memory). Các phép tính trong mạng LSTM được xây dựng từ các cấu trúc nhỏ hơn gọi là LSTM cell. Mỗi LSTM cell bao gồm ba cổng (gate) quan trọng: cổng quên (forget gate), cổng đầu vào (input gate) và cổng đầu ra (output gate), và một bộ nhớ trung gian gọi là cell state. Các cổng được điều khiển bởi hàm sigmoid để quyết định xem thông tin nào sẽ được lưu lại hoặc bỏ qua.

Quá trình huấn luyện mạng LSTM giúp cho mô hình học được cách lưu trữ và sử dụng thông tin từ quá khứ để dự đoán kết quả trong tương lai. Mạng LSTM đã được ứng dụng rất hiệu quả trong các bài toán xử lý ngôn ngữ tự nhiên như dịch máy, tạo tiêu đề cho ảnh và văn bản, phân loại cảm xúc, và nhiều bài toán khác. Kiến trúc cụ thể của mỗi ô nhớ được biểu diễn như trong Hình 1.10.

Diagram

Description automatically generated

Hình 1.10: Kiến trúc một ô nhớ trong LSTM

Trong đó:

\* 𝜎, 𝑡𝑎𝑛ℎ: hàm kích hoạt sigmoid, tanh;

\* ×: phép nhân ma trận (element – wise multiplication);

\* +: phép cộng ma trận;

\* : vector đầu vào tại mỗi bước thời gian t;

\* giá trị kích hoạt các cổng quên, cổng vào, cổng ra;

\* , vector đại diện cho trạng thái bên trong ô nhớ và giá trị tiềm năng;

\* : giá trị đầu ra của tế bào LSTM (hidden state);

\* : hệ số bias;

\* 𝑊, 𝑈: ma trận các hệ số của mạng.

Ý tưởng của LSTM là bổ sung thêm trạng thái bên trong ô nhớ (cell internal state) và ba cổng sàng lọc các thông tin đầu vào cho ô nhớ Hình 1.10biểu diễn trạng thái thứ 𝑡, các cổng đều lần lượt nhận giá trị đầu vào (đại diện cho một phần tử chuỗi đầu vào) và trạng thái ẩn , trạng thái ô có được từ đầu ra trạng thái 𝑡 - 1 trước đó.

Mạng Bi-LSTM: Mặc dù mạng LSTM có thể giữ được thông tin trong thời gian dài, tuy nhiên, nó chỉ xử lý thông tin trong một chiều thời gian. Để giải quyết vấn đề này, mạng Bi-LSTM (Bidirectional LSTM) được giới thiệu, cho phép mô hình học được cả thông tin của quá khứ và tương lai. Mạng Bi-LSTM kết hợp hai mạng LSTM chạy song song với nhau, một mạng LSTM chạy theo chiều thuận và một mạng LSTM chạy theo chiều ngược lại. Đầu ra của mỗi lớp LSTM được kết hợp với nhau để tạo ra đầu ra của mạng Bi-LSTM. Điều này cho phép mạng Bi-LSTM mô hình hóa được tất cả thông tin trong chuỗi dữ liệu và tăng khả năng dự đoán đầu ra.

Mạng Bi-LSTM thường được sử dụng trong các bài toán xử lý ngôn ngữ tự nhiên như dịch máy, phân loại văn bản, và nhận dạng tên thực thể. Nó cũng được sử dụng trong các ứng dụng khác như nhận dạng giọng nói và nhận dạng hình ảnh. Tuy nhiên, mạng Bi-LSTM có số lượng tham số lớn hơn so với mạng LSTM thông thường, do đó cần nhiều dữ liệu huấn luyện hơn để đạt được kết quả tốt. Một kiến trúc Bi-LSTM thường chứa 2 mạng LSTM đơn được sử dụng đồng thời và độc lập để mô hình hoá chuỗi đầu vào theo 2 hướng: từ trái sang phải (forward LSTM) và từ phải sang trái (backward LSTM) như Hình 1.11.

Background pattern

Description automatically generated

Hình 1.11: Kiến trúc mạng Bi-LSTM gồm 2 nhánh forward và backward LSTM

### 1.2.3. Các mô hình so khớp ảnh và câu mô tả

Để giải quyết bài toán tìm kiếm người từ câu mô tả ngôn ngữ tự nhiên thì sau khi giải quyết bài toán về xử lý ảnh và bài toán xử lý ngôn ngữ tự nhiên, chúng ta cần phải thực hiện việc so khớp giữa ảnh và câu mô tả. Những năm gần đây việc nghiên cứu và xây dựng những mô hình so khớp ảnh với câu mô tả đã được rất nhiều nhà khoa học quan tâm nghiên cứu. Những mô hình mới được đề xuất đối với bài toán ngày càng đạt được độ chính xác cao, trong đó cũng có những mô hình được xây dựng riêng cho so khớp hình ảnh với câu mô tả ngôn ngữ tiếng Việt được phát triển dựa trên các mô hình mạng học sâu nổi tiếng như GNA-RNN (Gated Neural Attention - Recurrent Neural Network), ViTAA (Visual-Textual Attributes Alignment),…

Trong bài báo [4] các tác giả trình bày về một mô hình mạng học sâu mới cho bài toán so khớp hình ảnh và câu mô tả bằng ngôn ngữ tiếng Việt. Bài báo trình bày cụ thể về quá trình xây dựng và huấn luyện mô hình mạng học sâu gồm hai giai đoạn là giai đoạn huấn luyện để huấn luyện mô hình với những cặp ảnh – câu mô tả và giai đoạn tìm kiếm hình ảnh đối tượng từ câu mô tả bằng cách ước lượng sự tương đồng của ảnh và câu mô tả.

Giai đoạn huấn luyện sẽ giúp mô hình học được sự tương đồng của câu mô tả và ảnh. Trong bài báo, tác giả đã đề xuất sử dụng mạng GNA-RNN (Gated Neural Attention Recurrent Neural Network) [1] để huấn luyện.

Diagram

Description automatically generated

Hình 1.12: Cấu trúc mạng GNA-RNN để so khớp ảnh với câu mô tả tiếng Việt [3]

Mô hình mạng này gồm hai phần là:

- Mạng con hình ảnh (visual-sub network): mạng con này có chức năng thu thập các thông tin về diện mạo người trong ảnh. Với mỗi bức ảnh mạng con này sẽ tính toán ra các tế bào thị giác (visual neural) **v** đại diện cho các đặc trưng ngoại hình của người trong ảnh. Trong bài báo gốc có 512 tế bào được ký hiệu **v** **=** [*v1, … , v512*]Tđược tạo ra dựa trên mạng VGG-16 và hai lớp kết nối đầy đủ (fully connected layers). Ngoài ra chúng ta cũng có thể sử dụng mạng Resnet [9] cho mạng con này.

- Mạng con ngôn ngữ (Language sub-network): nhằm mục đích học được sự quan trọng của các visual neurons thông qua mạng attention (attention network) và vai trò của các từ khác nhau trong một câu thông qua mạng LSTM. Trong mạng này, mỗi cặp “câu mô tả - hình ảnh” sẽ được tính toán một véc-tơ nhúng (embedding) từ đặc trưng từ ngữ và đặc trưng hình ảnh sau đó kết nối chúng với nhau tạo thành véc-tơ đầu vào [ , ]­­T của mạng LSTM. Mạng LSTM gồm có một tế bào nhớ () , tầng ẩn () và 3 cổng điều khiển: cổng vào , cổng ra , cổng quên được định nghĩa như sau:

Text, letter

Description automatically generated

(1.3)

Mỗi tầng ẩn được đưa vào trong mạng attention để lấy giá trị attention (attention value) cho mỗi từ ngữ. Mỗi từ ngữ trong câu sẽ mang lượng thông tin khác nhau đối với việc so khớp câu mô tả với hình ảnh, vì vậy cổng cấp độ (word-level gate) sẽ tính toán trọng số của mỗi từ trong câu. Để làm được việc này, sẽ được ánh xạ thông qua một lớp kết nối đầy đủ với hàm phi tuyến sigmoid.

= σ( + ) (1.4)

Sự tương đồng giữa từ thứ t trong câu và hình ảnh được tính như sau:

(1.5)

Sau đó mạng sẽ được huấn luyện bằng cách sử dụng hàm mất mát entropy chéo như sau:

(1.6)

Ở đây, đại diện cho sự tương đồng dự đoán cho mẫu thứ i, và đại diện cho nhãn thực tế của nó: = 1 cho các cặp câu-hình ảnh tương ứng và = 0 cho những cặp không tương ứng.

Với mô hình mạng trên bài báo cho về kết quả với độ chính xác với top-10 hình ảnh đầu tiên khoảng 70% trên bộ dữ liệu VnPersonSearch do tác giả thu thập và xây dựng và khoảng 55% đối với bộ dữ liệu CUHK-PEDES với câu mô tả đã được dịch sang tiếng Việt. Kết quả thực nghiệm cho thấy hiệu suất của phương pháp tìm kiếm người dựa trên các câu truy vấn tiếng Việt là rất đáng kỳ vọng. Tuy nhiên, tập dữ liệu về mô tả và hình ảnh người dùng thu thập bằng tiếng Việt trong bài báo vẫn còn khá nhỏ và còn cần cải thiện thêm về độ chính xác.

Tiếp tục sự phát triển của các mô hình tìm kiếm hình ảnh dựa trên câu mô tả ngôn ngữ Tiếng Việt, chúng ta có mô hình tìm kiếm người bằng câu mô tả tiếng Việt sử dụng mô hình máy học visual-textual attributes alignment (ViTAA) được huấn luyện sẵn. Mô hình này được đề xuất trong bài báo [4], trong đó các tác giả đề xuất sử dụng kiến trúc mạng ViTAA để học cách trích xuất thông tin từ các mô tả về người trong tiếng Việt và kết hợp với thông tin hình ảnh để tìm kiếm người. Cuối cùng, các tác giả sử dụng trọng số mô hình đã được huấn luyện từ một tập dữ liệu lớn để tìm kiếm người bằng mô tả ngôn ngữ tự nhiên bằng tiếng Anh để áp dụng vào tập dữ liệu VnPersonSearch dành cho tìm kiếm người bằng mô tả ngôn ngữ tự nhiên bằng tiếng Việt.

Mạng ViTAA (Visual-Textual Attributes Alignment) được đề xuất tại [15] được đăng trên hội nghị IEEE International Conference on Computer Vision (ICCV) năm 2019. Mạng Visual-Textual Attributes Alignment (ViTAA) là một kiến trúc mạng được sử dụng trong tìm kiếm người bằng ngôn ngữ tự nhiên (Person Search by Natural Language). Mục đích của mạng ViTAA là giải quyết bài toán tìm kiếm người bằng câu mô tả.

Diagram

Description automatically generated

Hình 1.13: Cấu trúc mạng ViTAA [15]

Mạng ViTAA sử dụng kiến trúc CNN và RNN để học các biểu diễn của hình ảnh và mô tả người dựa trên các đặc trưng thị giác và ngôn ngữ. Trong quá trình huấn luyện, mạng ViTAA được tối ưu hóa để đảm bảo rằng các biểu diễn của hình ảnh và mô tả được đồng bộ hóa và liên kết với nhau.

Mạng ViTAA có cấu trúc gồm hai phần chính là mô hình trích xuất đặc trưng hình ảnh và mô hình xử lý ngôn ngữ tự nhiên, kết hợp với một mô hình thực hiện việc cân bằng giữa thông tin hình ảnh và thông tin văn bản. Bottom of Form

Mạng ViTAA đã được đánh giá trong các tập dữ liệu tìm kiếm người bằng ngôn ngữ tự nhiên, và cho thấy kết quả tốt hơn so với các phương pháp khác.

Trong bài báo [4], để thực hiện huấn luyện mô hình mạng này tác giả sử dụng bộ dữ liệu tìm kiếm người CUHK-PEDES. Đây là một bộ dữ liệu chứa nhiều hình ảnh người và các mô tả tương ứng, được sử dụng để huấn luyện mạng học sâu. Các ảnh sẽ được xử lý thông qua mạng Resnet50 để tạo ra véc- tơ đặc trưng của ảnh. Tuy nhiên, các mô tả này được viết bằng tiếng Anh. Để sử dụng cho bài toán tìm kiếm người bằng mô tả bằng tiếng Việt, các câu tiếng Anh được dịch sang tiếng Việt và được tiền xử lý trước khi đi vào quá trình trích xuất đặc trưng toàn cục và địa phương.

Diagram

Description automatically generated

Hình 1.14: Quy trình tổng thể quá trình huấn luyện mô hình tìm kiếm người bằng câu truy vấn tiếng Việt [4]

Đầu tiên các câu mô tả sẽ được phân đoạn thành các từ (hay token). Các từ đều được ghép lại từ một hoặc nhiều âm tiết, được phân cách bằng khoảng trắng. Sau đó, các câu mô tả tiếng Anh được dịch sang tiếng Việt bằng cách sử dụng Google API. Các câu mô tả tiếng Việt này sẽ được sử dụng để trích xuất đặc trưng văn bản toàn cục. Tiếp theo đó, mô tả sẽ được phân tích và chia thành các phần nhỏ tương ứng với năm thuộc tính của một người. Đối với mỗi phần, sử dụng phân tích cú pháp (POS) để xác định vai trò ngữ pháp của từ trong câu và trích xuất các chuỗi danh từ. Các chuỗi danh từ được chia thành 5 phần tương ứng với 5 thuộc tính hình ảnh (đầu, phần thân trên, phần thân dưới, giày và phụ kiện). Sau đó, sử dụng mạng Bi-LSTM để trích xuất đặc trưng toàn cục và đặc trưng cục bộ của các từ đã được phân tích trước đó.

Các từ được mã hóa thành vector one-hot và đưa vào mạng Bi-LSTM để trích xuất đặc trưng. Kết quả đầu ra sẽ được chuyển qua một mạng liên kết tuyến tính để tạo ra véc-tơ đặc trưng toàn cục và cục bộ tương ứng cho mỗi phần. Cuối cùng, các véc-tơ đặc trưng được sử dụng để tìm kiếm người dựa trên mô tả trong tập dữ liệu CUHK-PEDES.

Mô hình trong bài báo sau khi được huấn luyện đạt được kết quả tương đối tốt trên các bộ dữ liệu tìm kiếm người như bộ dữ liệu VnPersonSearch (bộ dữ liệu đầu tiên xây dựng cho tìm kiếm người bằng câu mô tả tiếng Việt) CUHK-PEDES (bộ dữ liệu xây dựng cho tìm kiếm người bằng câu mô tả bằng tiếng Anh) với độ chính xác với câu mô tả tiếng Việt gần xấp xỉ so với câu mô tả tiếng Anh.

|  |  |  |  |
| --- | --- | --- | --- |
| **Ranking** | **English** | **Vietnamese** | |
| **Setting A** | **Setting B** |
| Top-1 | 55,97% | 53,18% | 52,40% |
| Top-5 | 74,84% | 73,29% | 72,28% |
| Top-10 | 83,52% | 81,03% | 80,78% |

Bảng 1.1: Kết quả của mô hình trên bộ dữ liệu CUHK-PEDES với câu mô tả tiếng Việt so với kết quả mô hình khi sử dụng câu mô tả tiếng Anh [4]

|  |  |  |
| --- | --- | --- |
| **Ranking** | **Method** | **Our method** |
| Top-1 | 26,85% | **61,57%** |
| Top-5 | 52,47% | **83,93%** |
| Top-10 | 66,40% | **90,67%** |

Bảng 1.2: Kết quả của mô hình trên bộ dữ liệu VnPersonSearch [4]

Trên đây là những thông tin cơ bản về các mô hình so khớp ảnh và câu mô tả ngôn ngữ tiếng Việt. Những mô hình này đã đạt được những kết quả tương đối tốt so với các mô hình so khớp hình ảnh câu mô tả bằng tiếng Anh, điều này cho ta thấy được tiềm năng phát triển của những mô hình này trong tương lai.

**KẾT LUẬN CHƯƠNG 1**

Chương 1 trình bày khái quát về sơ đồ của hệ thống tìm kiếm người dựa trên câu mô tả, về các bài toán cần giải quyết cũng như các thách thức có liên quan khi tiến hành xây dựng một hệ thống tìm kiếm người dựa trên câu mô tả. Trong bối cảnh khoa học công nghệ ngày càng phát triển mạnh như ngày nay, việc ứng dụng hệ thống này vào trong công tác của lực lượng công an là hết sức cần thiết, nó sẽ góp phần hỗ trợ cho công tác nắm tình hình, truy tìm đối tượng của lực lượng công an, giảm thiểu nguy cơ xảy ra tội phạm giúp đảm bảo an ninh trật tự ở các địa phương. Chương này cũng trình bày các mô hình học sâu sử dụng cho tìm kiếm hình ảnh đối tượng từ câu mô tả ngôn ngữ tự nhiên tiếng Việt: mô hình xử lý hình ảnh VGG16, Resnet; mô hình xử lý ngôn ngữ tự nhiên LSTM, BiLSTM; Mô hình so khớp ảnh với câu mô tả GNA-RNN sử dụng kiến trúc mạng GNA-RNN để so khớp giữa hình ảnh và câu mô tả tiếng Việt, ViTAA sử dụng mô hình mạng Resnet50 để trích chọn đặc trưng ảnh và mạng BiLSTM để trích chọn đặc trưng câu mô tả. Những mô hình này là những mô hình tiêu biểu được sử dụng cho bài toán tìm kiếm người dựa trên câu mô tả ngôn ngữ tự nhiên tiếng Việt. Những kiến thức tổng quan về bài toán nghiên cứu của đề tài được nêu ở Chương 1 sẽ giúp chúng tôi có những luận cứ khoa học có giá trị để tiếp tục nghiên cứu đề xuất giải pháp phù hợp và hiệu quả cho bài toán nghiên cứu đặt ra trong đề tài.

# Chương 2

# MÔ HÌNH HỌC SÂU TRANSFORMER CHO XỬ LÝ ẢNH VÀ XỬ LÝ NGÔN NGỮ

## 2.1. Mô hình học sâu Transformer cho xử lý ảnh

Hiện nay mô hình học sâu Transformer được xem là một trong những mô hình thành công nhất trong lĩnh vực xử lý ngôn ngữ tự nhiên, với khả năng xử lý dữ liệu dạng chuỗi. Tuy nhiên, với sự phát triển của lĩnh vực thị giác máy tính, các nhà nghiên cứu đã áp dụng mô hình Transformer vào xử lý ảnh và đưa ra mô hình học sâu Transformer cho xử lý ảnh (ViT-Vision Transformer).

ViT là một kiến trúc mạng học sâu có tính đột phá trong lĩnh vực thị giác máy tính, được các nhà nghiên cứu từ Google Research Brain Team giới thiệu trong hội nghị tại ICLR 2021 [5]. Thay vì sử dụng các lớp tích chập như các kiến trúc CNN truyền thống, ViT sử dụng các lớp transformer để xử lý ảnh.Trong ViT, ảnh được chia thành các phần nhỏ (patch) và được xem như là một chuỗi các vector. Mỗi patch được đưa vào một lớp Linear Projection để tạo ra một vector đại diện cho patch đó. Sau đó, các vector này được đưa vào các lớp transformer để học các mối quan hệ giữa chúng. Cuối cùng, vector đại diện của toàn bộ ảnh được đưa qua một hoặc nhiều lớp kết nối đầy đủ (fully-connected layers) để phân loại.ViT đã đạt được những kết quả tốt trong các bài toán nhận dạng ảnh và xử lý ảnh, và hiện đang được sử dụng rộng rãi trong cộng đồng học máy và thị giác máy tính.

ViT giúp giải quyết một số vấn đề của các mô hình CNN truyền thống trong việc xử lý ảnh. Những vấn đề này bao gồm:

- Vấn đề xác định miền con: Trong các mô hình CNN truyền thống, việc định hình miền con của ảnh là một vấn đề quan trọng. Điều này có nghĩa là mô hình sẽ chỉ tập trung vào một số vùng của ảnh, dẫn đến việc bỏ qua các vùng khác quan trọng. Tuy nhiên, ViT sử dụng cơ chế phân giải trên toàn ảnh để phân tách ảnh thành các miền con đầu vào, giúp cho mô hình có thể tập trung vào toàn bộ ảnh và bao quát hơn.

- Vấn đề cục bộ và toàn cục: Một vấn đề khác của các mô hình CNN truyền thống là chúng chỉ tập trung vào các thông tin cục bộ của ảnh, mà bỏ qua các thông tin toàn cục. Điều này làm giảm khả năng của mô hình trong việc hiểu được bối cảnh của ảnh. Tuy nhiên, ViT sử dụng cơ chế phân tích toàn bộ ảnh và tạo ra các miền con đầu vào không chồng lấn để tìm ra các đặc trưng toàn cục của ảnh.

- Vấn đề ảnh hưởng của kích thước ảnh: Trong các mô hình CNN truyền thống, kích thước của ảnh đầu vào phải được cố định, điều này làm giảm tính linh hoạt của mô hình. Tuy nhiên, ViT không bị giới hạn bởi kích thước ảnh đầu vào, giúp cho mô hình có thể xử lý các ảnh có kích thước khác nhau.

- Vấn đề về mức độ trực quan của kết quả: Với các mô hình CNN truyền thống, việc giải thích kết quả dự đoán của mô hình có thể rất khó khăn. Tuy nhiên, với ViT, việc giải thích trở nên dễ dàng hơn bằng cách sử dụng các công cụ hỗ trợ giải thích như bản đồ thể hiện vùng quan trọng trong ảnh (attention map).

### 2.1.1. Kiến trúc của mô hình Vision Transformer (ViT)A picture containing diagram Description automatically generated

Hình 2.1: Kiến trúc của mô hình Vision Transformer [5]

Kiến trúc của ViT bao gồm hai phần chính: tầng nhúng (Embedding layer) và tầng mã hóa Transformers (Transformer Encoder layer).

#### 2.1.1.1. Embedding layer

Embedding layer trong kiến trúc Vision Transformer (ViT) là phần đầu tiên của mô hình, được sử dụng để biến đổi ảnh thành các vector để đưa vào các lớp Transformer Encoder để xử lý.

Các bước thực hiện của Embedding layer như sau:

- Chia ảnh thành các phần nhỏ (patch): Trước tiên, ảnh sẽ được chia thành các phần nhỏ bằng cách chia toàn bộ ảnh thành các ô vuông cùng kích thước. Kích thước của mỗi ô vuông (patch size) được xác định trước đó.

- Biến đổi các patch thành vector: Mỗi patch sẽ được biến đổi thành một vector dựa trên một lớp linear projection. Quá trình này được thực hiện bằng cách sử dụng một ma trận trọng số để ánh xạ từng patch thành một vector đặc trưng (feature vector). Thực chất Linear Projection là một lớp Dense với đầu vào là flatten vector của các patches, đầu ra sẽ là embedding vector tương ứng với từng patch.

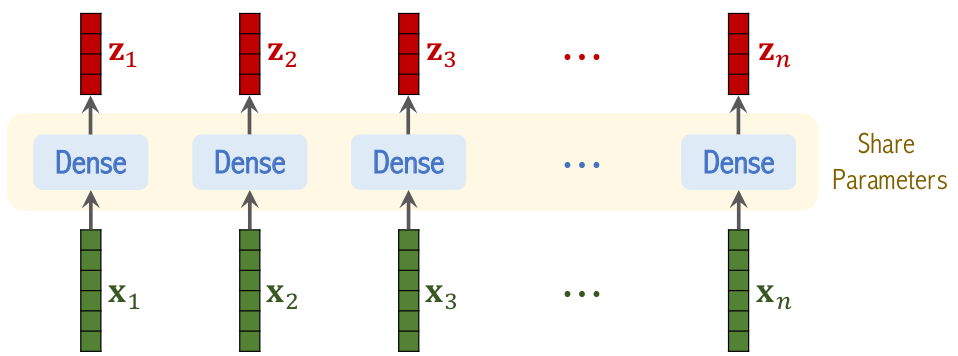
(2.1)

Trong đó:

: là flatten vector của patch thứ i.

: là output tương ứng của khi qua Linear Projection.

W được gọi là ma trận embeeding.

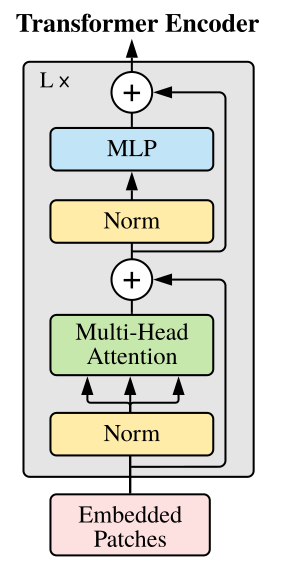


Hình 2.2: Mô tả phần Linear Projection

- Tạo chuỗi đầu vào: Sau khi tất cả các patch được biến đổi thành các vector, chúng sẽ được sắp xếp thành một chuỗi đầu vào cho phần sau của mô hình. Mỗi vector đại diện cho một patch sẽ là một phần tử trong chuỗi này.

Với Embedding layer, ViT đã chuyển đổi ảnh thành một chuỗi các vector số mà mỗi vector đại diện cho một patch. Điều này cho phép ViT sử dụng lớp Transformer Encoder để học các mối quan hệ giữa các patch và phân tích ảnh để giải quyết các bài toán nhận dạng và xử lý ảnh.

#### 2.1.1.2. Tầng mã hóa Transformers (Transformer Encoder layer)



Hình 2.3: Kiến trúc của Transformer Encoder [5]

Tầng mã hóa Transformers là phần chính trong kiến trúc mạng Vision Transformer (ViT), được sử dụng để xử lý chuỗi các vector đại diện cho các phần nhỏ (patch) của ảnh.

Mỗi tầng mã hóa Transformers trong ViT bao gồm hai phần chính:

- Multi-head Attention: Trong lớp Multi-head Attention, các vector đại diện cho các patch của ảnh sẽ được đưa vào để học các mối quan hệ giữa chúng. Cụ thể, mỗi vector đại diện cho một patch sẽ được truyền qua ba lớp linear projection để tạo ra ba vector: query, key và value. Sau đó, các vector này sẽ được sử dụng để tính toán các trọng số chú ý (attention scores), từ đó tính toán ra một vector kết quả mới.

- Feedforward Neural Network (Mạng nơ-ron truyền thẳng): Sau khi tính toán được các vector mới từ lớp Multi-head Attention, chúng sẽ được đưa qua một lớp Feedforward Neural Network để tạo ra các vector đại diện mới. Lớp Feedforward Neural Network bao gồm hai lớp linear projection và một hàm kích hoạt phi tuyến để tạo ra các vector đại diện mới.

Sau khi các vector đại diện mới được tạo ra từ lớp Feedforward Neural Network, chúng sẽ được chuẩn hóa bằng một lớp Layer Normalization trước khi được đưa tiếp vào lớp Transformer Encoder tiếp theo hoặc đưa vào lớp linear projection cuối cùng để phân loại. ViT sử dụng nhiều lớp Transformer Encoder được xếp chồng lên nhau để xử lý chuỗi các vector đại diện cho các patch của ảnh. Mỗi lớp Transformer Encoder sẽ học các mối quan hệ giữa các patch để giải quyết các bài toán nhận dạng và xử lý ảnh. Các mô hình ViT với số lượng lớp Transformer Encoder khác nhau có thể được sử dụng tùy thuộc vào loại bài toán và tài nguyên tính toán.

### 2.1.2. Ứng dụng của ViT trong xử lý ảnh

ViT có thể được sử dụng trong nhiều ứng dụng khác nhau trong xử lý ảnh, bao gồm:

- Phân loại ảnh: Trong ứng dụng phân loại ảnh, ViT đã chứng tỏ khả năng vượt trội của mình so với các mô hình CNN truyền thống. ViT có thể học được các tính năng quan trọng của ảnh thông qua cơ chế phân tích toàn bộ ảnh mà không cần định hình miền con như các mô hình CNN truyền thống. Điều này giúp cho ViT có khả năng bao quát và xử lý thông tin ảnh tốt hơn, đặc biệt là đối với các ảnh có kích thước lớn. ViT cũng cho thấy khả năng học tập ở mức độ cao, khi có thể học được các đặc trưng phức tạp của ảnh và tăng hiệu suất phân loại.

- Nhận diện đối tượng: ViT cũng có thể được sử dụng trong ứng dụng nhận diện đối tượng. Điều này đặc biệt hữu ích khi xử lý các ảnh có độ phân giải cao và số lượng đối tượng lớn. ViT có khả năng phân tích toàn bộ ảnh, giúp cho việc nhận diện đối tượng trở nên dễ dàng và chính xác hơn. Trong lĩnh vực nhận dạng khuôn mặt, ViT cũng có thể được sử dụng để phân loại khuôn mặt theo độ tuổi, giới tính, biểu cảm và các đặc trưng khác. Ngoài ra, ViT cũng có khả năng học được các đặc trưng phức tạp của đối tượng, giúp tăng độ chính xác của việc nhận diện đối tượng.

- Phát hiện vật thể: ViT có thể được sử dụng để phát hiện vật thể trong hình ảnh, tức là xác định vị trí và định hình của vật thể. Với kiến trúc transformer, ViT có thể học được các đặc trưng của vật thể và tạo ra một biểu diễn trừu tượng cho chúng. ViT có thể sử dụng một mô hình phân loại đầu ra hoặc một mô hình hồi quy đầu ra để xác định vị trí và định hình của vật thể trong hình ảnh.

- Trích xuất thông tin từ hình ảnh: ViT có thể được sử dụng trong trích xuất thông tin từ hình ảnh, đặc biệt là trong việc trích xuất văn bản từ hình ảnh. Trong quá trình chuyển đổi hình ảnh thành một vector, ViT không chỉ mã hóa thông tin hình ảnh mà còn mã hóa cả thông tin về văn bản nếu có trên hình ảnh đó. ViT có thể học được những kết nối đồng thời giữa văn bản và hình ảnh trong quá trình huấn luyện, từ đó giúp cho việc trích xuất văn bản từ hình ảnh trở nên chính xác hơn. Một trong những ứng dụng của ViT trong trích xuất thông tin từ hình ảnh là trích xuất dữ liệu từ hình ảnh chứa các biểu đồ. Trong các báo cáo khoa học hoặc kinh doanh, các biểu đồ thường được sử dụng để trình bày và truyền đạt thông tin. Tuy nhiên, việc trích xuất dữ liệu từ các biểu đồ thường gặp nhiều khó khăn do độ phức tạp của biểu đồ và định dạng dữ liệu. ViT có thể giúp giải quyết vấn đề này bằng cách học cách mã hóa đồng thời cả hình ảnh và văn bản trên biểu đồ, từ đó trích xuất thông tin về giá trị và đơn vị của các trục trong biểu đồ. Ngoài ra, ViT cũng có thể được sử dụng để trích xuất thông tin từ các hình ảnh chứa vật thể hoặc khuôn mặt. ViT có thể học được các đặc trưng của vật thể hoặc khuôn mặt, từ đó giúp cho việc nhận dạng chính xác hơn. ViT cũng có thể học được các mối quan hệ giữa các vật thể trong cùng một bức ảnh, từ đó giúp cho việc phân tích ảnh trở nên chính xác hơn.

- Phân tích dữ liệu y tế: ViT có thể được sử dụng trong các ứng dụng phân tích dữ liệu y tế, đặc biệt là trong bài toán phân loại ảnh y tế. ViT có khả năng học được các đặc trưng của các hình ảnh y tế, giúp cho việc phát hiện các bất thường và dự đoán các bệnh lý trở nên dễ dàng hơn. Ví dụ, ViT có thể được sử dụng để phát hiện ung thư qua hình ảnh siêu âm, CT hoặc MRI, giúp cho việc chẩn đoán sớm và điều trị hiệu quả hơn. ViT cũng có thể được sử dụng để phân tích hình ảnh từ các thiết bị y tế thông minh như đồng hồ thông minh, vòng đeo tay, hoặc các thiết bị y tế khác để đánh giá sức khỏe và tiên đoán nguy cơ bệnh lý.

## 2.2. Mô hình học sâu Transformer cho xử lý ngôn ngữ

Ngoài mô hình học sâu phục vụ cho xử lý ảnh thì kiến trúc mạng Transformer còn được ứng dụng để xây dựng các mạng học sâu để xử lý ngôn ngữ tự nhiên, một trong những mô hình tiêu biểu sử dụng kiến trúc mạng này đó chính là mô hình mạng học sâu BERT (Bidirectional Encoder Representations from Transformers).

Mô hình mạng học sâu BERT được giới thiệu lần đầu vào năm 2018 trong bài báo [16] của các nhà nghiên cứu làm việc tại Google AI là Jacob Devlin, Ming-Wei Chang, Kenton Lee và Kristina Toutanova. Trong bài báo các tác giả giới thiệu BERT là một mô hình được huấn luyện sẵn (pre-train model) trên một tập dữ liệu lớn, và sau đó có thể được hiệu chỉnh (fine-tuning) trên các tác vụ cụ thể như phân loại văn bản, truy xuất thông tin, và dịch máy.

BERT là một mô hình mạng học sâu được huấn luyện sẵn trên một lượng lớn những văn bản từ nhiều nguồn khác nhau như Wikipedia và BookCorpus. Điều này khiến BERT có khả năng hiểu được ngữ cảnh và cấu trúc câu văn bản bằng cách xem xét các từ trong câu và quan hệ giữa chúng. Ngoài ra với việc huấn luyện trên một lượng dữ liệu lớn có thể giúp cho BERT học được nhiều kiến thức thông qua các ngữ cảnh khác nhau của từ ngữ, các kiến thức này có thể được sử dụng lại trong quá trình hiệu chỉnh (fine-tuning) cho các tác vụ cụ thể mà không phải huấn luyện lại từ đầu, giúp tiết kiệm được thời gian và tăng độ chính xác.

BERT được xây dựng dựa trên kiến trúc Transformer. với L là số lớp Transformer (blocks), H là số tầng ẩn, A là số heads ở lớp attention. Có hai mô hình BERT là BERTBASE (L=12, H=768, A=12, Total Parameters=110M), và mô hình BERTLARGE (L=24, H=1024, A=16, Total Parameters=340M). Kiến trúc của BERT gồm:

- Một lớp ánh xạ đầu vào (input embedding) để ánh xạ các từ đầu vào thành các véc tơ số (vector embedding), giúp biểu diễn các chuỗi văn bản đầu vào của mô hình.

- Một số lượng lớn các lớp Transformer tạo thành bộ mã hóa (Encoder) để biểu diễn đầu vào dưới dạng các vector đặc trưng. Các lớp Transformer trong bộ mã hóa của BERT bao gồm nhiều lớp tự chú ý (self-attention) và các mạng truyền thẳng (feedforward networks).

+ Lớp tự chú ý (self-attention) là một phần quan trọng trong kiến trúc mạng transformer, được sử dụng trong nhiều nhiệm vụ xử lý ngôn ngữ tự nhiên. Trong lớp self-attention, mỗi véc-tơ đầu vào sẽ được sử dụng để tính toán trọng số cho tất cả các véc-tơ đầu vào khác. Tức là mỗi véc-tơ đầu vào đều tham gia tính toán trọng số cho các véc-tơ khác. Các trọng số này sau đó được sử dụng để tính toán trung bình có trọng số của các véc-tơ đầu vào, tạo ra một véc-tơ đầu ra mới cho mỗi vector đầu vào. Lớp self-attention cho phép mô hình tập trung vào các phần quan trọng của đầu vào, đồng thời giúp cho mô hình có khả năng xử lý các câu văn bản có chiều dài khác nhau.

+ Các mạng truyền thẳng (feedforward networks) là mạng nơ-ron trong đó các đơn vị tính toán được sắp xếp thành các lớp và thông tin được truyền theo một chiều từ lớp đầu vào đến lớp đầu ra mà không có các kết nối phản hồi. Mỗi đơn vị tính toán trong một lớp sử dụng đầu vào của nó để tính toán đầu ra của nó và truyền nó đến lớp tiếp theo.

- Huấn luyện trước: Trước khi được hiệu chỉnh cho một tác vụ cụ thể, BERT được huấn luyện trước (pre-trained) trên một tập dữ liệu lớn để học các đại diện ngữ nghĩa của ngôn ngữ tự nhiên. Quá trình huấn luyện trước này được thực hiện bằng cách sử dụng hai tác vụ:

+ Tác vụ dự đoán từ bị che giấu (Masked Language Model - MLM): BERT được huấn luyện để dự đoán từ bị che giấu trong câu đầu vào, để mô hình có thể hiểu được ngữ cảnh của các từ.

+ Tác vụ dự đoán từ tiếp theo (Next Sentence Prediction - NSP): BERT được huấn luyện để học cấu trúc ngữ pháp và mối quan hệ giữa hai câu liên tiếp.

Sau khi được huấn luyện trước trên một lượng lớn dữ liệu, mô hình BERT được hiệu chỉnh (fine-tune) trên các tác vụ xử lý ngôn ngữ tự nhiên cụ thể. Việc này cho phép mô hình học các đặc trưng ngữ nghĩa cụ thể của các tác vụ này và đạt được hiệu suất tốt trong các tác vụ đó.

Graphical user interface, diagram, application

Description automatically generated

Hình 2.4: Quy trình tổng thể của việc tiền huấn luyện và hiệu chỉnh cho BERT [16]

### 2.2.1. Biểu diễn đầu vào và đầu ra của BERT

Để giúp BERT có thể thực hiện được nhiều tác vụ khác nhau, các nhà nghiên cứu đã thực hiện việc biểu diễn đầu vào của mô hình theo một câu hoặc một cặp câu văn bản (ví dụ: một câu hỏi và một câu trả lời) được tạo thành từ một chuỗi từ ngữ, hoặc có thể là một đoạn văn bản bất kỳ chứ không nhất thiết là một câu hoàn chỉnh. Trong bài báo tác giả sử dụng nhúng WordPiece[17] của tác giả Wu và cộng sự năm 2016 với một từ điển 30.000 từ và sử dụng ## làm dấu phân tách. Ví dụng từ playing được tách thành play##ing.

**A screenshot of a computer

Description automatically generated with medium confidence**

Hình 2.5: Biểu diễn đầu vào của BERT [16]

Hình 2.5 thể hiện chi tiết biểu diễn đầu vào của mô hình BERT một cách trực quan. Trong đó, các câu đầu vào được chia thành các véc-tơ và được gắn các token. Token đầu tiên là một token đặc biệt có giá trị là [CLS] thông báo bắt đầu chuỗi. Trong trường hợp có hai câu được gộp lại thì chúng ta có thể phân biệt hai câu bằng cách tách chúng ra bằng token đặc biệt [SEP], và thực hiện nhúng phân đoạn cho mỗi câu như trên (những từ thuộc câu A sẽ có được đánh dấu EA, những từ thuộc câu B là EB). Như vậy biểu diễn đầu vào của một chuỗi trong mô hình BERT được tính bằng tổng của các token, phân đoạn và vị trí tương ứng của các từ trong chuỗi.

### 2.2.2. Quá trình huấn luyện trước BERT (pre-training BERT)

Đây là quá trình để huấn luyện cho BERT trước khi nó được hiệu chỉnh để thực hiện các tác vụ cụ thể khác, trong quá trình này các nhà nghiên cứu đã không thực hiện việc huấn luyện theo cách thông thường như các mô hình trước đây mà sử dụng hai tác vụ học không giám sát (unsupervised task) như trong Hình 2.4.

- Tác vụ dự đoán từ bị che giấu (Masked Language Model - MLM): Áp dụng tác vụ này cho phép chúng ta huấn luyện một mô hình học được ngữ cảnh 2 chiều. Với MLM, mô hình sẽ có thể dự đoán các từ bị che giấu trong chuỗi đầu vào. Cụ thể, một tỷ lệ nhất định các từ trong chuỗi đầu vào được chọn ngẫu nhiên và thay thế bằng một token đặc biệt là [MASK]. Sau đó, mô hình phải dự đoán lại các từ đã được thay thế này. Quá trình này được thực hiện trên toàn bộ các chuỗi đầu vào. Sau đó, các véc-tơ ẩn (hidden vector) cuối cùng tương ứng với các token đã được [MASK] được đưa vào một lớp softmax trên toàn bộ bộ từ vựng, tương tự như một ngôn ngữ mô hình điều kiện tiêu chuẩn để dự đoán. Các nhà nghiên cứu trong bài báo đã thử nghiệm giấu 15% tất cả các token lấy từ từ điển của WordPiece trong câu một cách ngẫu nhiên và cho mô hình chỉ dự đoán các từ được giấu mà không cần phục hồi toàn bộ chuỗi đầu vào.

Mặc dù điều này cho phép chúng ta có được một mô hình huấn luyện 2 chiều, nhưng có 2 nhược điểm tồn tại. Đầu tiên là nó tạo ra một sự không phù hợp giữa quá trình huấn luyện và hiệu chỉnh (fine-tune) khi một số từ trong quá trình hiệu chỉnh sẽ không bị giấu. Để giảm thiểu điều này, chúng ta không nên lúc nào cũng thay thế các từ được giấu đi bằng token [MASK]. Thay vào đó, trình tạo dữ liệu huấn luyện chọn 15% tokens một cách ngẫu nhiên và thực hiện các bước như sau:

Ví dụ với câu: "chiếc\_áo của bạn thật đẹp", từ được chọn để giấu là từ "đẹp":

+ Thay thế 80% từ được chọn trong dữ liệu huấn luyện thành token [MASK] " chiếc\_áo của bạn thật [MASK]"

+ 10% các từ được chọn sẽ được thay thế bởi 1 từ ngẫu nhiên " chiếc\_áo của bạn thật sách"

* 10% còn lại được giữ không thay đổi " chiếc\_áo của bạn thật đẹp"

Khi làm như vậy Transformer không hề biết được từ nào sẽ được yêu cầu dự đoán hoặc từ nào đã được thay thế bằng một từ ngẫu nhiên, do đó, nó buộc phải giữ một biểu diễn theo ngữ cảnh của mỗi token đầu vào giúp giảm sự không liên quan giữa quá trình huấn luyện và quá trình hiệu chỉnh của BERT.

- Tác vụ dự đoán từ tiếp theo (Next Sentence Prediction - NSP): tác vụ này đảm nhiệm việc huấn luyện cho mô hình BERT hiểu và phân tích mối quan hệ giữa hai câu trong một văn bản, thông qua việc dự đoán xem hai câu liền kề trong văn bản có phải là câu tiếp theo của nhau hay không. Để thực hiện nhiệm vụ này, khi huấn luyện, BERT sẽ chọn hai câu liên tiếp từ tập dữ liệu và phân loại chúng thành hai loại: "IsNext" và "NotNext". Nếu câu thứ hai là câu tiếp theo của câu đầu tiên, thì nó được gán nhãn là "IsNext", còn nếu không phải, nó sẽ được gán nhãn là "NotNext". Cụ thể, khi chọn câu A và câu B cho mỗi mẫu huấn luyện thì có 50% khả năng câu B là câu tiếp theo sau câu A và 50% câu A là một câu không liên quan. Ví dụ:

+ Input: [CLS] Cái\_áo kia [MASK] quá [SEP] tôi muốn [MASK] nó [SEP] Câu này có nhãn là: isNext

+ Input: [CLS] Cái\_áo kia [MASK] quá [SEP] anh ấy đang đi ăn tối [SEP] Câu này có nhãn là: notNext

Mặc dù ta chọn những câu notNext một cách ngẫu nhiên tuy nhiên mô hình cuối cùng cũng đạt được độ chính xác lên đến 97 - 98% đối với tác vụ này. Qua đó, BERT học được các mối quan hệ giữa các câu trong văn bản, tạo ra một mô hình ngôn ngữ tự động có khả năng hiểu được ngữ cảnh và phân tích được cấu trúc ngữ pháp trong các văn bản tự nhiên.

Sau khi huấn luyện với hai tác vụ trên, mô hình BERT sẽ được hiệu chỉnh để thực hiện các tác vụ cụ thể như phân loại, trích xuất thông tin, trả lời câu hỏi, và các tác vụ liên quan đến xử lý và phân tích ngôn ngữ (NLP - Natural Language Processing) khác.

### 2.2.3. Quá trình hiệu chỉnh và các thử nghiệm của BERT

Fine-tuning của BERT là quá trình huấn luyện lại mô hình với một tập dữ liệu nhỏ hơn và một tác vụ cụ thể, nhằm đạt được hiệu suất tốt hơn trên tác vụ đó. Quá trình này được gọi là fine-tuning vì mô hình đã được huấn luyện sẵn trên một tập dữ liệu lớn, và việc huấn luyện lại chỉ cần điều chỉnh một số trọng số của mô hình để phù hợp với tác vụ cụ thể. Kết quả hiệu chỉnh của BERT đã đạt được kết quả tốt hơn nhiều so với các mô hình huấn luyện trước đó. Sau đây, chúng ta cùng xem qua một vài nhiệm vụ mà BERT đã thực hiện.

Trên tập dữ liệu SQuAD (Stanford Question Answering Dataset), mô hình BERT đạt được độ chính xác rất cao cho tác vụ trả lời câu hỏi. Trong bài báo gốc về BERT, mô hình này đã đạt được độ chính xác F1-Score trên tập SQuAD với kết quả là 93,2%, phá vỡ kỷ lục trước đó của các mô hình trước đó. Sau đó, các nghiên cứu khác cũng đã tiếp tục cải tiến và nâng cao độ chính xác của BERT trên tập dữ liệu này, ví dụ như BERT-large đạt được kết quả F1-Score lên đến 94,6%. Đây là một kết quả rất ấn tượng và cho thấy BERT là một trong những mô hình NLP hiệu quả nhất hiện nay cho tác vụ trả lời câu hỏi.

Table

Description automatically generated

Hình 2.6: Kết quả thử nghiệm của BERT với tệp dữ liệu SquAD [16]

Đối với công việc phân loại văn bản, các nhà nghiên cứu đã tiến hành thử nghiệm trên tập dữ liệu lớn như GLUE (General Language Understanding Evaluation). Trong đó, với tập dữ liệu GLUE, mô hình BERT-base đã đạt được kết quả tốt nhất cho tất cả các tác vụ, với độ chính xác trung bình là 80,5%, BERT đạt được kết quả tốt nhất cho tác vụ phân loại văn bản như SST-2, MRPC, QQP, STS-B. Mô hình này đạt được kết quả tốt hơn cả so với các mô hình đã được đào tạo đặc biệt cho dịch máy và các mô hình dựa trên mạng nơ-ron tích chập (CNN). Đây là các kết quả rất ấn tượng cho thấy mô hình BERT là một trong những mô hình NLP hiệu quả nhất hiện nay cho các tác vụ phân loại văn bản và hiểu ngôn ngữ tổng quát.

Table

Description automatically generated

Hình 2.7: Kết quả thử nghiệm của BERT với tệp dữ liệu GLUE [16]

Ngoài ra các tác giá cũng đa hiệu chỉnh cho BERT để tiến hành nhiều tác vụ khác như dự đoán từ, câu tiếp theo với tệp dữ liệu SWAG (Situations With Adversarial Generations). Tập dữ liệu này chứa 113 nghìn ví dụ hoàn thành câu mà bạn phải đánh giá, suy luận để hoàn thành chúng. Với một câu từ tập dữ liệu, nhiệm vụ là quyết định xem trong bốn lựa chọn đó thì lựa chọn nào là hợp lý nhất. Ví dụ :

A girl is going across a set of monkey bars. She

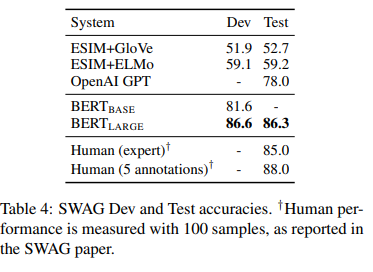
(1) jump up across the monkey bars.

(2) struggles onto the bars to grab her head.

(3) gets to the end and stands on a wooden plank

(4) jump up an does a back flip

Thử nghiệm cho thấy BERT đạt được độ chính xác lên tới 81,6% với BERTBASE.



Hình 2.8: Kết quả thử nghiệm của BERT với SWAG [16]

Như vậy chúng ta đã tìm hiểu cơ bản về mô hình BERT, đây là mô hình học sâu Transformers cho xử lý ngôn ngữ khá nổi tiếng trong thời gian gần đây. Mô hình BERT đã cho thấy hiệu quả cao trên nhiều bộ dữ liệu xử lý ngôn ngữ khác nhau và đã vượt qua nhiều mô hình trước đó trong nhiều tác vụ NLP. Một trong những đóng góp lớn nhất của BERT là khả năng học được biểu diễn từ và câu từ phụ thuộc vào ngữ cảnh, giúp cải thiện đáng kể hiệu quả của các tác vụ phụ thuộc vào ngữ cảnh. Tuy nhiên, BERT cũng có nhược điểm nhất định. Vì mô hình rất lớn và phức tạp, việc huấn luyện và triển khai BERT đòi hỏi nhiều tài nguyên tính toán. Với các ứng dụng thực tế, việc sử dụng BERT có thể gặp rào cản về tốc độ và hiệu quả tính toán. Ngoài ra, do BERT chỉ được huấn luyện trên các tập dữ liệu bằng tiếng Anh nên không thể hiểu được đặc thù của ngôn ngữ khác, đặc biệt với ngôn ngữ phức tạp như tiếng Việt thì việc hiệu chỉnh BERT gặp rất nhiều khó khăn.

**KẾT LUẬN CHƯƠNG 2**

Tóm lại, trong Chương 2 chúng ta đã tìm hiểu về kiến trúc của mô hình học sâu Transformer cho xử lý ảnh Vision Transformers (ViT), các thành phần chính trong kiến trúc Transformers là: tầng nhúng (Embedding layer) và tầng mã hóa Transformers, thành phần của các tầng trong mô hình ViT và các ứng dụng của ViT trong xử lý hình ảnh như: phân loại ảnh, nhận diện đối tượng, phát hiện vật thể, trích xuất thông tin từ hình ảnh, phân tích dữ liệu y tế. Kiến trúc mô hình học sâu Transformer cho xử lý ngôn ngữ BERT, quá trình xử lý đầu vào của mô hình, bộ mã hóa (Transformers Encoder), các tác vụ huấn luyện trước cho BERT (MLM và NSP) giúp BERT có khả năng hiểu được ngữ cảnh và phân tích được cấu trúc ngữ pháp của văn bản đầu vào, sau khi được huấn luyện trước, BERT sẽ được hiệu chỉnh cho phù hợp để thực hiện từng công việc xử lý văn bản cụ thể như phân loại văn bản, trả lời câu hỏi, dự đoán từ ngữ, dịch máy,… với độ chính xác rất cao, tuy nhiên mô hình BERT mới chỉ được huấn luyện cho ngôn ngữ tiếng Anh nên việc hiệu chỉnh để thực hiện các công việc xử lý ngôn ngữ Tiếng Việt sẽ gặp rất nhiều khó khăn.

Với những kiến thức đã tìm hiểu được về các mô hình học sâu Transformers cho xử lý ảnh và ngôn ngữ, chúng tôi đề xuất phát triển một hệ thống sử dụng Transformer cho tìm kiếm hình ảnh đối tượng từ camera sử dụng câu mô tả ngôn ngữ tự nhiên tiếng Việt. Chi tiết mô hình đề xuất và các thử nghiệm có liên quan sẽ được trình bày trong chương tiếp theo.

# Chương 3

# HỆ THỐNG ĐỀ XUẤT DỰA TRÊN KIẾN TRÚC TRANSFORMER CHO TÌM KIẾM HÌNH ẢNH ĐỐI TƯỢNG SỬ DỤNG CÂU MÔ TẢ NGÔN NGỮ TỰ NHIÊN TIẾNG VIỆT

## 3.1. Sơ đồ khối của hệ thống

### 3.1.1. Sơ đồ khối

Sơ đồ khối hệ thống tìm kiếm người qua câu mô tả được thể hiện ở Hình 3.1. Hệ thống gồm 2 pha chính là huấn luyện mô hình (End-to-end training) và đánh giá mô hình (Testing) sử dụng bộ dữ liệu 3000VnPersonSeach [11].

- Pha huấn luyện gồm 3 phần chính:

+ Trích chọn đặc trưng ảnh người sử dụng mạng học sâu ResNet-50, được chia thành 2 nhánh:

\* Trích chọn đặc trưng trên toàn bộ ảnh người (đặc trưng toàn cục).

\* Trích chọn đặc trưng trên từng bộ phận của cơ thể người (đặc trưng cục bộ) gồm 5 bộ phận: đầu (head), thân trên (upperbody), thân dưới (lowerbody), giày (shoe), khác (túi, ba lô). Tách bộ phận cơ thể người bằng cách huấn luyện mạng phân tách vùng ảnh người (Human Parsing Network).

+ Trích chọn đặc trưng câu mô tả sử dụng mạng học sâu Bi-LSTM và Tabtransformer, được chia thành 2 nhánh:

\* Trích chọn đặc trưng trên toàn bộ câu mô tả (đặc trưng toàn cục).

\* Trích chọn đặc trưng cục bộ câu mô tả (đặc trưng cục bộ) dựa vào từng cụm danh từ tương ứng với các bộ phận trên cơ thể người.

+ Huấn luyện đồng thời cặp ảnh – câu mô tả dựa trên các đặc trưng đã trích chọn

- Pha đánh giá, các ảnh người và câu mô tả được tiền xử lý trước khi đưa vào mô hình đã huấn luyện để đưa ra vector đặc trưng tương ứng với ảnh và câu mô tả. Tiến hành so khớp vector đặc trưng ảnh và câu mô tả để đưa ra ảnh phù hợp nhất với câu mô tả. Sử dụng phép đo Consin để so khớp vector đặc trưng ảnh và câu mô tả.

A screenshot of a diagram

Description automatically generated with low confidence

Hình 3.1: Mô hình tìm kiếm ảnh người dựa trên câu mô tả tiếng Việt.

### 3.1.2. Trích chọn đặc trưng ảnh người

Hình 3.2 thể hiện các bước trích chọn đặc trưng toàn cục và cục bộ của ảnh người. Để trích xuất đặc trưng từ ảnh người đầu vào I, sử dụng mạng học sâu ResNet-50 với kích thước chuẩn hóa ảnh là 384x128 pixels. Sau khi mã hóa, ảnh được biểu diễn dưới dạng vector đặc trưng Fimg với kích thước (1024x24x8). Từ vector Fimg, sử dụng hai nhánh để trích xuất đặc trưng: nhánh toàn cục (Global branch) dùng để thu được vector đặc trưng toàn cục với kích thước (1x256), và nhánh cục bộ (Local branch) dùng để thu được vector đặc trưng cục bộ với kích thước (1x256) cho từng thuộc tính ảnh tương ứng với từng bộ phận người.

Để tách ra 5 thuộc tính ảnh (Đầu, Thân trên, Thân dưới, Giày, Khác (Other: túi xách, ba lô…)), sử dụng mạng phân tách vùng người (Human Parsing Network) để gán nhãn cho từng vùng ảnh tương ứng. Sau đó, sử dụng những thông tin này để phân tích và đưa ra quyết định về các thuộc tính ảnh của người trong ảnh đầu vào.

A diagram of a person walking

Description automatically generated with low confidence

Hình 3.2: Trích chọn đặc trưng ảnh người

### 3.1.3. Trích chọn đặc trưng câu mô tả

Hình 3.3 thể hiện các bước trích chọn đặc trưng toàn cục và cục bộ từ câu mô tả.

A picture containing text, screenshot, font, diagram

Description automatically generated

Hình 3.3: Trích chọn đặc trưng câu mô tả

Vector đặc trưng toàn cục được trích chọn từ toàn bộ câu mô tả được thực hiện qua các bước:

- Xử lý câu mô tả thành dạng bảng gồm các trường là bộ phận cơ thể người: đầu (head), thân trên (upperbody), thân dưới (lowerbody), giày (shoe), khác (túi, ba lô..) và trọng số Flnweight.

- Đưa bảng ở trên qua mạng TabTransformer cho ra vector đặc trưng có kích thước 1x193. Tiếp tục đưa vector trước đó qua nhánh toàn cục (Global branch) cho ra vector đặc trưng toàn cục tglob có kích thước (1×256).

Vector đặc trưng cục bộ được được thực hiện trên từng cụm danh từ đã được tách ra từ câu mô tả qua các bước sau:

- Mã hóa các cụm danh từ dưới dạng one-hot vector dựa trên từ điển được xây dựng từ bộ dữ liệu 3000VnPersonSearch. Từ điển này gồm 2555 token được xây dựng từ 12604 câu mô tả tiếng Việt.

- Mỗi token trong từ điển được mã hoá thành một số nguyên thể hiện vị trí của nó trong từ điển sau khi được sắp xếp. Khi token hoá câu mô tả, nếu không có trong bộ từ điển đã xây dựng, thì giá trị token tại đó sẽ được gán giá trị bằng 0.

A picture containing text, screenshot, font, line

Description automatically generated

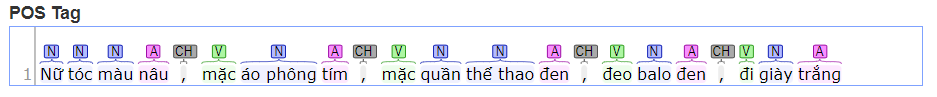
Hình 3.4: Mã hoá câu mô tả tiếng Việt

- Đưa one-hot vector qua mạng Bi-LSTM cho đầu ra là vector đặc trưng Ftext có kích thước (1024×1×1).

- Vector đặc trưng Ftext được đưa qua nhánh cục bộ (Local branch) cho ra vector đặc trưng cục bộ tloc có kích thước (1×256).

#### 3.1.3.1. Tiền xử lý câu mô tả

Trong bước tiền xử lý, thực hiện loại bỏ từ dừng, chuyển chữ hoa thành chữ thường. Sau đó thực hiện phân đoạn từ, gán nhãn từ loại, phân cụm từ và trích ra các cụm danh từ sử dụng công cụ Underthesea [13]. Underthesea là một bộ dữ liệu mô-đun Python mã nguồn mở và các hướng dẫn hỗ trợ nghiên cứu và phát triển trong xử lý ngôn ngữ tự nhiên tiếng Việt. Trong công cụ Underthesea, sử dụng module POS Tag để tách các thẻ đánh dấu phân loại từ loại trong tiếng Việt. Ví dụ: danh từ (N), tính từ (A), động từ (V), giới từ (P), liên từ (C), dấu câu (CH), … Hình 3.4 minh họa kết quả POS Tag thực hiện trực tiếp trên website của Underthesea:



Hình 3.5: Kết quả POS Tag thực hiện trên website Underthesea.com

Từ kết quả POS Tag như ở hình trên, tiến hành trích xuất ra các cụm danh từ mô tả các thuộc tính trên cơ thể người dựa trên các danh từ (N) và tính từ (A). Cấu trúc chung của một cụm danh từ thường là: các danh từ + các tính từ. Sau khi trích xuất ra được các cụm danh từ, thực hiện phân loại các cụm danh từ theo các từ khóa tương ứng với các thuộc tính: Person, Head, Upperbody, Lowerbody, Shoe, Other. Cụ thể như sau:

- Nếu cụm danh từ chứa các từ như “đàn ông”, “phụ nữ”, “nam”, “bé gái”,… thì sẽ được phân loại vào thuộc tính “Person”.

- Nếu cụm danh từ chứa các từ như “mái tóc”, “kính”, “mũ”, … thì sẽ được phân loại vào thuộc tính “Head”.

- Nếu cụm danh từ chứa các từ như “áo”, “áo phông”, “váy”,… thì sẽ được phân loại vào thuộc tính “Upperbody”.

- Nếu cụm danh từ chứa các từ như “quần”, “quần bò”,… thì sẽ được phân loại vào thuộc tính “Lowerbody”.

- Nếu cụm danh từ chứa các từ như “giày”, “dép”, … thì sẽ được phân loại vào thuộc tính “Shoe”.

- Nếu cụm danh từ chứa các từ như “ví”, “balo”, “túi xách”, … thì sẽ được phân loại vào thuộc tính “Other”.

Từ việc phân tích câu mô tả thành các cụm danh từ chỉ từng bộ phận, tiến hành thống kê số lần các danh từ thể hiện bộ phận người xuất hiện. Ví dụ: “tóc” xuất hiện 2055 lần, “áo phông” xuất hiện 4562 lần, “áo khoác” xuất hiện 1777 lần, "quần bò" xuất hiện 1442 lần, "giày thể thao" xuất hiện 2718 lần, “dép” xuất hiện 656 lần, …Tiến hành tính trọng số Flnweight cho mỗi câu mô tả bằng cách tính tổng số lần xuất hiện của các danh từ thể hiện các bộ phận người mà người đó được mô tả. Ví dụ, câu: “nam thanh niên, mặc áo phông xanh, mặc quần bò dài, đi dép lê” sẽ có trọng số Flnweight = 4562 + 1442 + 656 = 6660.

#### 3.1.3.2. Kiến trúc mạng TabTransformer

TabTransformer [14] là một kiến trúc mạng học sâu dùng để giải quyết bài toán dự đoán và phân loại trên dữ liệu bảng (tabular data). Dữ liệu bảng thường là các bảng dữ liệu có cấu trúc với các cột tương ứng với các thuộc tính và các hàng tương ứng với các mẫu dữ liệu.

Câu mô tả được xử lý thành dạng bảng khi đưa vào mạng TabTransformer sẽ được chia làm hai phần chính bao gồm:

+ Column Embedding là các mô tả của từng bộ phận cơ thể người.

+ Flnweight là giá trị trọng số của câu mô tả.

A picture containing text, diagram, plan, technical drawing

Description automatically generated

Hình 3.6: Mô hình mạng TabTransformer

Mỗi khối Transformer gồm 2 thành phần chính: Multi-Head Attention và Feedforward network, ngoài ra còn có cả normalization layer. Transformer đầu tiên sẽ nhận ma trận biểu diễn của các từ, sau đó ma trận này sẽ được xử lý bởi Multi-Head Attention. Multi-Head Attention thật chất là self-attention, nhưng mà để mô hình có thể có chú ý nhiều khuôn mẫu khác nhau, để đơn giản thì sử dụng nhiều self-attention.

Để tính self-attention, bước đầu tiên là tạo ra ba vector từ mỗi véc tơ đầu vào của encoder. Với mỗi từ, sẽ tạo một vector truy vấn (Query), một vector khóa (Key), và một vector giá trị (Value). Các vector này được tạo ra bằng cách nhân embedding với ba ma trận được cập nhật trong quá trình huấn luyện (WQ, WK, WV).

A picture containing text, screenshot, diagram, design

Description automatically generated

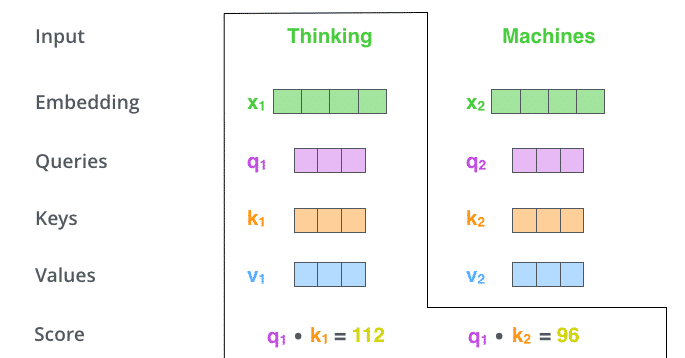
**mặc**

**đàn ông**

Hình 3.7: Tạo vector truy vấn, khóa và giá trị

Ví dụ với 2 từ đầu vào “đàn ông” và “mặc” tương ứng 2 vector embeding X1 và X2: Nhân vector X1 với ma trận trọng số WQ ta được vector truy vấn đi với từ “đàn ông” là q1. Tương tự đối với vector khóa và vector giá trị ta được bộ ba vector truy vấn, khóa, giá trị tương ứng với từ “đàn ông” và “mặc” là (q1, k1, v1), (q2, k2, v2).

Bước thứ hai để tính self-attention là tính điểm. Giả sử tính self-attention cho từ đầu tiên trong câu là “đàn ông”. Ta cần tính điểm cho mỗi từ trong câu đầu vào so với từ này. Điểm sẽ quyết định cần chú ý bao nhiêu vào các phần khác của câu đầu vào khi ta đang mã hóa một từ cụ thể. Điểm được tính bằng phép nhân vô hướng giữa vector truy vấn với vector khóa của từ mà đang tính điểm. Nếu tiến hành self-attention cho từ ở vị trí thứ nhất, điểm đầu tiên sẽ là tích vô hướng của q1 và k1. Điểm thứ hai là tích vô hướng của q1 và k2.



**mặc**

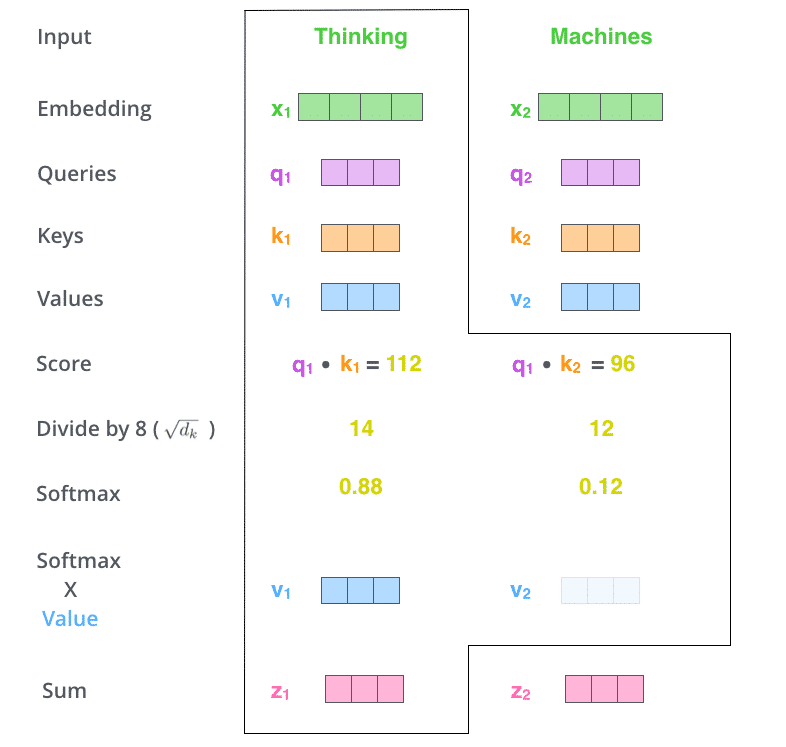
**đàn ông**

Hình 3.8: Tính điểm của self-attention

Bước thứ ba là chia điểm cho căn bậc hai của số chiều của vector khóa. và truyền kết quả qua một phép softmax. Softmax chuẩn hóa các điểm để chúng là các số dương có tổng bằng 1. Điểm softmax sẽ quyết định mỗi từ sẽ được thể hiện nhiều hay ít tại vị trí hiện tại. Rõ ràng là từ tại vị trí này sẽ có điểm softmax cao nhất, nhưng đôi khi, chú ý đến các từ khác là cần thiết để hiểu từ hiện tại.

Bước thứ tư là nhân mỗi vector giá trị với điểm softmax (trước khi cộng chúng lại). Việc này bảo toàn giá trị của các từ mà ta muốn chú ý và bỏ qua các từ không liên quan.

Bước thứ năm là cộng các vector giá trị đã được nhân trọng số. Kết quả chính là đầu ra của lớp self-attention tại vị trí hiện tại.

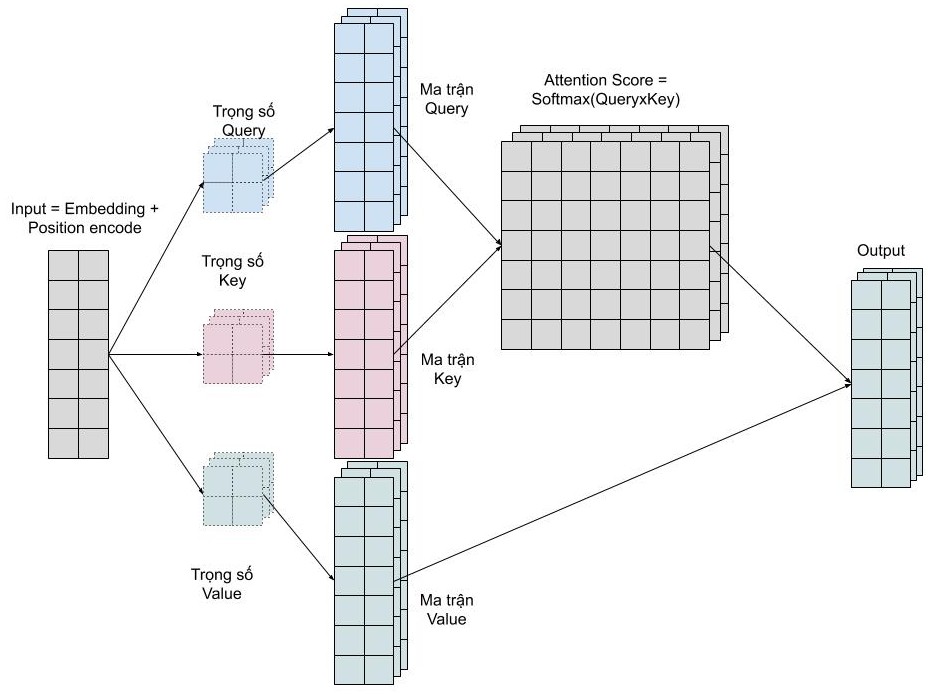


**mặc**

**đàn ông**

Hình 3.9: Tính tổng các vector giá trị

Đến đây là kết thúc việc tính toán self-attention. Vector kết quả có thể được gửi tới mạng truyền thẳng. Tuy nhiên, với mong muốn mô hình có thể học nhiều kiểu mối quan hệ giữa các từ với nhau Multi-Head Attention đã thêm nhiều self-attention. Tức là cần nhiều ma trận truy vấn, khóa và giá trị.



Hình 3.10: Multi-Head Attention

Do đang thực hiện trên ma trận, nên có thể gộp bước từ 2 đến 5 trong một công thức tính self- attention để tính đầu ra của lớp self-attention.

Attention(Q,K,V)=softmax()V

Với dk là số chiều vector giá trị.

Các vector sau khi đi qua bước Add & Normalize (sẽ được trình bày phần sau) sẽ được gửi tới Feed-Forward Networks (FFN). Lớp này bao gồm 2 tầng biến đổi thông tin và 1 hàm ReLU (các giá trị < 0 được gán lại = 0) ở giữa. Dropout với tỉ lệ 0.1 cũng được áp dụng ở lần biến đổi thứ nhất sau khi các vector qua hàm ReLU.

FFN =

Sau khi qua lớp Feed-Forward các vector cũng phải qua bước Add & Normalize trước khi đi vào khối Transformer kế tiếp. Lớp Feed-Forward này dùng để học mối quan hệ tiềm ẩn giữa các vector độc lập mà chưa được mô tả rõ ràng. Khác với mối quan hệ giữa các từ được khuếch đại qua lớp self-attention, vẫn còn những mối quan hệ tiềm ẩn khác không thể diễn giải bằng công thức toán học sẽ được học thông qua lớp này.

Lớp Add & Normalize được sử dụng sau lớp Multi-Head Attention và Feed-Forward để chuẩn hóa giá trị đầu ra của các lớp này. Các vector đầu ra từ lớp con (Multi-Head Attention và Feed-Forward) qua bước dropout với tỉ lệ 0.1, rồi cộng thêm vector đầu vào (vector trước khi bị biến đổi), lớp chuẩn hóa này tính toán giá trị trung bình và độ lệch chuẩn của đầu vào và điều chỉnh giá trị đầu ra để có độ lệch trung bình bằng 0 và độ sai lệch chuẩn bằng 1, rồi chuyển vào layer kế tiếp. Ý nghĩa của bước này là để bổ sung thêm thông tin nguyên bản, tránh bị mất mát quá nhiều thông tin sau khi qua các phép biến đổi ở các layer multi-head self-attention và feed forward.

### 3.1.4. Huấn luyện mô hình cặp ảnh – câu mô tả

Sau khi lựa chọn các đặc trưng toàn cục và cục bộ từ cả ảnh và câu mô tả, chúng ta thực hiện huấn luyện mô hình end-to-end với các cặp đặc trưng ảnh - câu mô tả. Mục đích của giai đoạn huấn luyện là xây dựng một không gian nhúng chung để có thể ánh xạ giữa các đặc trưng thuộc tính của ảnh và văn bản. Để đạt được mục tiêu này, chúng ta sử dụng phương pháp học tương phản (contrastive learning) với hàm mất mát entropy chéo (Cross-entropy loss) để hỗ trợ việc học các đặc trưng phân biệt của từng mô hình ảnh hoặc văn bản.

Hàm mất mát (loss function) thường được sử dụng trong quá trình huấn luyện của một mạng học sâu với mục tiêu là tối thiểu hóa hàm mất mát. Hàm mất mát cross-entropy được tính toán trên cả hai nhánh tương ứng với đặc trưng được trích chọn từ ảnh (công thức 3.1) và câu mô tả (công thức 3.2) như sau:

(3.2)

(3.1)

với B là kích cỡ lô (batch size), *vi, ti* là các đặc trưng của ảnh và câu mô tả thứ *i* tương ứng; *k* là phân lớp của ảnh hay câu mô tả thứ *i* với đặc trưng tương ứng là *vi* và *ti*; Wk hay Wj là các vector trọng số tương ứng với lớp thứ k và lớp thứ j; là tham số độ lệch (bias) tương ứng với lớp k của ảnh hay mô tả thứ i; C là tổng số lớp của cặp ảnh và câu mô tả.

Hàm mất mát cross-entropy cũng được sử dụng để phân loại các bộ phận thuộc tính cơ thể người trong ảnh (Đầu, Thân trên, Thân dưới, Giày, Other):

(3.3)

với: là đặc trưng của thuộc tính ảnh thứ i và m là số lượng thuộc tính ảnh;

M là tổng số lớp thuộc tính ảnh.

## 3.2. Thử nghiệm và kết quả

### 3.2.1. Cơ sở dữ liệu thử nghiệm và độ đo đánh giá

#### 3.2.1.1. Cơ sở dữ liệu

Đề tài sử dụng bộ dữ liệu 3000VnPersonSearch. Bộ dữ liệu bao gồm 3000 ID người, 6302 ảnh và 12604 câu mô tả. Bộ dữ liệu ảnh được thu thập từ camera giám sát tại những khu vực đông người như hồ Gươm, trường đại học. Các địa điểm được ghi hình vào ban ngày, lúc tối, trong đó thời gian ghi hình chủ yếu diễn ra vào ban ngày và trong điều kiện thời tiết bình thường với ánh sáng ổn định. Việc trích chọn khu vực có ảnh người được xác định qua khung hình chữ nhật bao quanh người đó được tiến hành thủ công qua công cụ LabelImg hoặc công cụ phát hiện người tự động YOLOv8. Tuy nhiên việc trích chọn khu vực ảnh người được tiến hành chủ yếu tự động. Đối với mỗi ảnh sẽ được mô tả bởi hai người khác nhau để đảm bảo sự đa dạng về ngôn ngữ tuy nhiên cần tập trung thể hiện vào bề ngoài của người trong ảnh.

A picture containing text, line, parallel, font

Description automatically generated

Hình 3.11: Bộ dữ liệu 3000VnPersonSearch

#### 3.2.1.2. Độ đo đánh giá

Hệ thống được đánh giá dựa trên khả năng xếp hạng đúng K kết quả đầu tiên, được gọi là Rank K (R@K). Đối với mỗi câu mô tả được đưa vào, hệ thống sẽ xử lý để đưa ra được một vector đặc trưng cho câu mô tả, sau đó tiến hành so sánh với các vector đặc trưng của ảnh trong tập dữ liệu ảnh cần tìm kiếm sử dụng độ đo cosine. Độ đo đánh giá cosine là một phương pháp đo độ tương đồng giữa hai vector dựa trên khoảng cách giữa chúng trong không gian đa chiều. Độ đo cosine càng nhỏ thì câu mô tả càng giống với ảnh. Trong các kết quả trả về, nếu lấy 1 ảnh có độ cosine lớn nhất với câu mô tả đầu vào thì ta được K = 1, nếu lấy 2 ảnh có độ cosine lớn nhất với câu mô tả đầu vào thì ta được K = 2, ….Trong các thử nghiệm của đề tài, chúng em lựa chọn K = 1, 5, 10, là những chỉ số thường được chọn để biểu diễn kết quả trong các thử nghiệm trên thế giới.

### 3.2.2. Kết quả thử nghiệm

#### 3.2.2.1. Phân bổ dữ liệu

Để huấn luyện và kiểm định mô hình trên bộ dữ liệu 3000VnPersonSearch, thực hiện chia bộ dữ liệu thành 3 phần:

+ Tập huấn luyện (train): gồm 2370 người, 10080 ảnh và 20160 câu mô tả.

+ Tập kiểm định (valid): gồm 314 người, 1258 ảnh và 2516 câu mô tả.

+ Tập đánh giá (test): gồm 316 người, 1262 ảnh và 2524 câu mô tả.

#### 3.2.2.2. Các tham số huấn luyện

Hình ảnh được đưa về kích thước (384x128) pixels.

- Trọng số tiêu biến (weight decay) được đặt bằng 4 x 10-5.

- Số lượng ảnh trên một lô (batch) là 64 ảnh.

- Tốc độ học (learning rate) được đặt bằng 1 x 10-4.

A picture containing line, plot, diagram

Description automatically generated- Tổng số vòng lặp (epoch) thực hiện huấn luyện là 20 epoch: dựa trên quan sát giá trị mất mát (loss) trong quá trình huấn luyện cho thấy mất mát trung bình cho các epoch từ 17 trở đi gần như thay đổi rất ít nên các thử nghiệm sẽ sử dụng tham số huấn luyện epoch=20 (Hình 3.11).

Hình 3.12: Sự thay đổi của giá trị mất mát qua các vòng lặp

#### 3.2.2.3. Kết quả thử nghiệm

Các thử nghiệm được thực hiện trên máy tính Intel(R) Xeon(R) CPU @ 2.00GHz, RAM 12.7GB, GPU 15GB.

Kết quả thử nghiệm trên bộ dữ liệu vn3k được thực hiện qua 2 kịch bản:

+ Kịch bản 1: Dữ liệu dạng bảng phân tách bộ phận cơ thể người không đưa qua mạng TabTransformer mà mã hóa thành 1 one-hot vector.

+ Kịch bản 2: Dữ liệu dạng bảng phân tách bộ phận cơ thể người đưa qua mạng TabTransformer, thu được 1 vector đặc trưng toàn cục.

|  |  |  |  |
| --- | --- | --- | --- |
| **Kịch bản** | **R@1** | **R@5** | **R@10** |
| Kịch bản 1 | 46,197% | 73,851% | 83,677% |
| Kịch bản 2 | 53.724% | 82,448% | 89,540% |

Bảng 3.1: Kết quả thực nghiệm trên tập dữ liệu 3000VnPersonSearch

Kết quả thực nghiệm ở hai kịch bản cho thấy độ chính xác kịch bản 2 cao hơn kịch bản 1. Độ chênh lệch ở R@1, R@5 và R@10 tương đối lớn dao động trong khoảng từ 5,8% đến 8,597%.

**KẾT LUẬN CHƯƠNG 3**

Chương 3 trình bày mô hình đề xuất cho tìm kiếm hình ảnh người từ câu mô tả ngôn ngữ tự nhiên tiếng Việt. Mạng học sâu dựa trên kiến trúc Transformer (mạng TabTransformer) được đề xuất để trích chọn đặc trưng câu mô tả. Dữ liệu đầu vào cho mạng được cấu trúc dưới dạng bảng. Dữ liệu gồm các cụm danh từ (NP-Noun Phrase) được sắp xếp dưới dạng cột theo 5 thuộc tính (Đầu, Thân trên, Thân dưới, Giày, Khác). Đây là lần đầu tiên dữ liệu câu mô tả được biểu diễn dưới dạng bảng, làm đầu vào cho mạng TabTransformer để biểu diễn đặc trưng toàn cục cho câu mô tả. Trích chọn đặc trưng ảnh và đặc trưng cục bộ của câu mô tả sử dụng mạng ResNet50 và mạng Bi-LSTM tương ứng. Kết quả tìm kiếm dựa trên hệ thống đề xuất đạt hiệu quả khá khả quan trên bộ cơ sở dữ liệu tiêu chuẩn. Về cơ bản kết quả đạt được đã hoàn thành mục tiêu nghiên cứu đề ra của đề tài.

# KẾT LUẬN

Đề tài đã hoàn thành các mục tiêu và nhiệm vụ nghiên cứu chính được đặt ra trong thời gian nghiên cứu đề tài như sau:

- Tìm hiểu về cấu trúc của hệ thống tìm kiếm người dựa trên câu mô tả tiếng Việt; nghiên cứu các mô hình mạng học sâu sử dụng cho bài toán tìm kiếm người bằng câu mô tả.

- Nghiên cứu kiến trúc Transformer và ứng dụng mạng học sâu TabTransformer cho trích chọn và so khớp đặc trưng câu mô tả ngôn ngữ tự nhiên tiếng Việt.

- Đánh giá thử nghiệm mô hình đề xuất cho tìm kiếm hình ảnh người từ câu mô tả dựa trên các bộ cơ sở dữ liệu tiêu chuẩn.

Tuy nhiên, do thời gian nghiên cứu còn hạn chế để có thể triển khai các sản phẩm của đề tài trong thực tiễn, trong thời gian tới, nhóm thực hiện đề tài sẽ tiếp tục nghiên cứu và hoàn thiện sản phẩm theo các hướng sau:

- Nghiên cứu cách áp dụng mạng học sâu Transformer cho toàn bộ mô hình tìm kiếm người bằng câu mô tả ngôn ngữ tiếng Việt bằng cách kết hợp mạng ViT (Vision Transformer [1]) và mạng TabTransformer.

- Thực hiện huấn luyện mô hình tìm kiếm người dựa trên câu mô tả tiếng Việt trên các bộ dữ liệu lớn hơn như bộ CUHK-PEDES gồm 40206 ảnh của 13003 người để nâng cao độ chính xác của mô hình.

- Sử dụng bộ từ điển word2vec của Google gồm 3 triệu từ và cụm từ để vector hóa câu mô tả góp phần nâng cao hiệu quả xử lý từ ngữ của mô hình trong quá trình huấn luyện giúp mô hình có thể hiểu được ngôn ngữ trong nhiều ngữ cảnh khác nhau.

# TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| 1. | A. Dosovitskiy *và c.s.*, “An image is worth 16x16 words: Transformers for image recognition at scale”, *arXiv preprint arXiv:2010.11929*, 2020. |
| 2. | A. F. Agarap, “Deep learning using rectified linear units (relu)”, *arXiv preprint arXiv:1803.08375*, 2018. |
| 3. | C. Szegedy *và c.s.*, “Going deeper with convolutions”, trong *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, tr 1–9. |
| 4. | G. Huang, Z. Liu, L. van der Maaten, và K. Q. Weinberger, “Densely connected convolutional networks”, trong *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, tr 4700–4708. |
| 5. | H. H. Tan và K. H. Lim, “Vanishing gradient mitigation with deep learning neural network optimization”, trong *2019 7th international conference on smart computing & communications (ICSCC)*, IEEE, 2019, tr 1–4. |
| 6. | J. Devlin, M.-W. Chang, K. Lee, và K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, *arXiv preprint arXiv:1810.04805*, 2018. |
| 7. | K. He, X. Zhang, S. Ren, và J. Sun, “Deep residual learning for image recognition”, trong *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, tr 770–778. |
| 8. | K. Simonyan và A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, *arXiv preprint arXiv:1409.1556*, 2014. |
| 9. | S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, và X. Wang, “Person search with natural language description”, trong *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, tr 1970–1979. |
| 10. | T. T. T. Pham *và c.s.*, “Person search by natural language description in Vietnamese using pre-trained visual-textual attributes alignment model”, trong *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*, IEEE, 2021, tr 1–6. |
| 11. | T. T. T. Pham *và c.s.*, “Towards a large-scale person search by vietnamese natural language: dataset and methods”, *Multimed Tools Appl*, vol 81, số p.h 19, tr 27569–27600, 2022. |
| 12. | T. T. T. Pham, D.-D. Nguyen, B. H. P. Ta, T.-B. Nguyen, T.-N.-D. Do, và T.-L. Le, “Person search by queried description in vietnamese natural language”, trong *Intelligent Information and Database Systems: 12th Asian Conference, ACIIDS 2020, Phuket, Thailand, March 23–26, 2020, Proceedings 12*, Springer, 2020, tr 469–480. |
| 13. | Vietnamese NLP Research Group, Word\_tokenize, <https://github.com/undertheseanlp/word_tokenize>, 17/04/2019. |
| 14. | X. Huang, A. Khetan, M. Cvitkovic, và Z. Karnin, “Tabtransformer: Tabular data modeling using contextual embeddings”, arXiv preprint arXiv:2012.06678, 2020. |
| 15. | Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, và T. Tan, “Pose-guided multi-granularity attention network for text-based person search”, trong *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, tr 11189–11196. |
| 16. | Y. Wu *và c.s.*, “Google’s neural machine translation system: Bridging the gap between human and machine translation”, *arXiv preprint arXiv:1609.08144*, 2016. |
| 17. | Z. Wang, Z. Fang, J. Wang, và Y. Yang, “Vitaa: Visual-textual attributes alignment in person search by natural language”, trong *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, Springer, 2020, tr 402–420. |