

ĐỒ ÁN CHUYÊN NGÀNH

TỰ ĐỘNG TẠO CÂU MIÊU TẢ CHO HÌNH ẢNH

Ngành: **CÔNG NGHỆ THÔNG TIN**

Chuyên ngành: **CÔNG NGHỆ PHẦN MỀM**

Giảng viên hướng dẫn : TS. HUỖNH QUỐC BẢO

Sinh viên thực hiện : Kiều Nguyễn Thanh Bình

MSSV: 2180603323 Lớp: 21DTHA3

Mục Lục

DANH MỤC BẢNG	2
DANH MỤC HÌNH	3
DANH MỤC TỪ VIẾT TẮT	4
Chương 1. TỔNG QUAN.....	7
1.1 Giới thiệu đề tài	7
1.2 Tổng quan tình hình nghiên cứu	8
1.3 Lý do chọn đề tài	9
1.4 Ý nghĩa của nghiên cứu	10
1.4.1 Ý nghĩa lý thuyết.....	10
1.4.2 Ý nghĩa thực tiễn	11
1.5 Đối tượng và phạm vi nghiên cứu	12
1.6 Cấu trúc đồ án	13
Chương 2 CƠ SỞ LÝ THUYẾT	14
2.1 Long Short-Term Memory (Bộ nhớ Ngắn – Dài hạn).....	14
2.2 Faster R-CNN (Shaoqing Ren,2016).....	17
2.3 Bottom-Up Top-Down Attention	24
2.4 Thuật toán Self-Critical Sequence Training (SCST).....	28
CHƯƠNG 3. KẾT QUẢ THỰC NGHIỆM	31
3.1 Dữ liệu thực nghiệm (Dataset).....	31
3.2 Tiền xử lý dữ liệu và Chuyển ngữ Tiếng Việt.....	31
3.3 Cơ chế trích xuất đặc trưng thị giác (Image Features)	32
3.4 Mô hình đề xuất	34
3.5 Kết quả thực nghiệm.....	34
3.6 Đánh giá và so sánh	36
CHƯƠNG 4. KẾT LUẬN VÀ KIẾN NGHỊ	40
4.1 Kết luận.....	40
4.2 Kiến nghị và Hướng phát triển	41
TÀI LIỆU THAM KHẢO	43

DANH MỤC BẢNG

Bảng 3.1 Kết quả thực nghiệm	36
---	-----------

DANH MỤC HÌNH

Hình 2.1 mô hình LSTM	15
Hình 2.2 Mô hình Faster R-CNN	18
Hình 2.3 Region Proposal Network.....	19
Hình 2.4 Caption Model base on Top-Down Attention	25
Hình 3.1 Mô hình dịch thuật.....	32
Hình 3.2 Mô hình trích xuất đặc trưng ảnh	33
Hình 3.3 Mô hình đề xuất.....	34
Hình 3.4 Kết quả phase CE	35
Hình 3.5 Kết quả phase SCST	35
Hình 3.6 Kết quả tốt nhất	37
Hình 3.7 Lỗi miêu tả.....	38
Hình 3.8 Kết quả sai ngữ cảnh	39

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	Từ viết đầy đủ
SCAN	Stack Cross Attention Network
BUTD	Bottom-Up Top-Down Attention
LSTM	Long-Short temp memory
TS	Tiến sĩ
Bleu	Bilingual Evaluation Understudy
Meteor	Metric for Evaluation of Translation with Explicit ORdering
Rouge_l	Recall-Oriented Understudy for Gisting Evaluation - Longest
CIDEr	Consensus-based Image Description Evaluation
Spice	Semantic Propositional Image Caption Evaluation
CNN	Convolutional Neural Network
AI	Artificial Intelligence
RNN	Recurrent Neural Network
MS COCO	
RPN	Region Proposal Networks

LỜI CẢM ƠN

Em xin gửi lời tri ân sâu sắc đến TS. Huỳnh Quốc Bảo, người đã luôn tận tâm hướng dẫn và truyền đạt các kiến thức quý báu trong suốt quá trình chúng em thực hiện đồ án chuyên ngành "Image captioning ". Thầy không chỉ hỗ trợ chúng em giải quyết những khó khăn gặp phải mà còn giúp định hướng rõ ràng để nhóm hoàn thành tốt nhiệm vụ được giao. Chúng em cũng xin gửi lời cảm ơn chân thành đến các thầy cô trong khoa, những người đã xây dựng một môi trường học tập lý tưởng và cung cấp cho chúng em các kiến thức chuyên môn cần thiết, từ đó giúp chúng em áp dụng hiệu quả vào thực tiễn. Đồng thời, nhóm xin bày tỏ lòng biết ơn đến bạn bè và những người thân yêu, những người đã luôn bên cạnh động viên và hỗ trợ chúng em cả về tinh thần lẫn vật chất trong suốt quá trình thực hiện đồ án. Thành quả đạt được là kết tinh của sự nỗ lực bền bỉ từ các thành viên trong nhóm cùng với sự hỗ trợ đặc lực từ những người xung quanh.

LỜI MỞ ĐẦU

Hãy tưởng tượng một người khiếm thị đang đi làm trong một thành phố sầm uất, nhận được sự hỗ trợ từ một ứng dụng trí tuệ nhân tạo có khả năng mô tả từng cảnh vật xung quanh bằng các câu ngắn gọn và rõ ràng. Điều này không chỉ giúp anh ta điều hướng tốt hơn mà còn tạo ra một kết nối mạnh mẽ hơn với thế giới xung quanh.

Trong thời đại công nghệ số phát triển mạnh mẽ, trí tuệ nhân tạo (AI) ngày càng đóng vai trò quan trọng trong nhiều lĩnh vực đời sống, đặc biệt là sự giao thoa giữa Thị giác máy tính và Xử lý ngôn ngữ tự nhiên. Một trong những bài toán tiêu biểu nhất thể hiện sự kết hợp này là **Mô tả ảnh tự động (Image Captioning)** – quá trình tạo ra câu mô tả bằng ngôn ngữ tự nhiên cho một ảnh đầu vào. Bài toán này không chỉ đòi hỏi hệ thống nhận diện chính xác các đối tượng trong ảnh mà còn phải thấu hiểu mối quan hệ ngữ nghĩa giữa chúng để tạo ra câu văn tự nhiên, hợp lý.

Trước đây, các phương pháp truyền thống thường sử dụng mạng CNN để nén toàn bộ hình ảnh thành một vector đặc trưng duy nhất (global feature) trước khi đưa vào mạng RNN sinh câu. Tuy nhiên, cách tiếp cận này bộc lộ nhiều hạn chế khi xử lý các hình ảnh phức tạp, do mô hình dễ bị mất mát thông tin chi tiết và gặp khó khăn trong việc "nhìn" vào đúng đối tượng đang được mô tả. Để khắc phục điều này, sự ra đời của **Cơ chế Chú ý (Attention Mechanism)**, đặc biệt là kỹ thuật **Bottom-Up Top-Down Attention**, kết hợp với các hệ thống phát hiện đối tượng tiên tiến như **Faster R-CNN (trong thư viện Detectron2)**, đã tạo ra bước đột phá mới. Phương pháp này cho phép mô hình tập trung vào từng vùng đối tượng cụ thể (Region-based features) thay vì nhìn toàn bộ ảnh một cách đại khái, giúp việc sinh từ trở nên chính xác và chi tiết hơn rất nhiều.

Từ thực tế đó, đề tài "**Mô tả ảnh tự động sử dụng mạng phát hiện đối tượng Detectron2 và cơ chế Bottom-Up Top-Down Attention**" được thực hiện. Mục tiêu của đề tài là xây dựng một hệ thống có khả năng sinh mô tả ảnh chính xác, giàu ngữ nghĩa, tận dụng sức mạnh của việc trích xuất đặc trưng vùng và cơ chế chú ý thông minh, từ đó mở ra tiềm năng ứng dụng cao trong các bài toán thực tế như hỗ trợ người khiếm thị hay tìm kiếm hình ảnh thông minh.

Chương 1. TỔNG QUAN

1.1 Giới thiệu đề tài

Sự phát triển mạnh mẽ của trí tuệ nhân tạo trong những năm gần đây đã mở ra nhiều hướng nghiên cứu và ứng dụng mới, trong đó nổi bật là lĩnh vực thị giác máy tính, nơi máy tính có thể "nhìn thấy", hiểu và phản hồi lại thông tin hình ảnh một cách thông minh. Một trong những bài toán tiêu biểu thể hiện khả năng tích hợp giữa thị giác máy tính và xử lý ngôn ngữ tự nhiên là bài toán mô tả ảnh tự động (Image Captioning). Đây là quá trình mà một hệ thống máy tính phân tích nội dung ảnh và tạo ra một mô tả bằng ngôn ngữ tự nhiên sao cho hợp ngữ pháp, chính xác về mặt nội dung và ngữ nghĩa.

Để giải quyết vấn đề này, nghiên cứu đặt ra câu hỏi cốt lõi: "Liệu việc sử dụng các đặc trưng vùng chi tiết từ Detectron2 kết hợp với cơ chế chú ý từ dưới lên và từ trên xuống (Bottom-Up Top-Down Attention) có thể vượt trội hơn các phương pháp truyền thống sử dụng đặc trưng toàn cục trong việc tạo mô tả cho ảnh trên Flickr30k hay không?"

Việc giải quyết bài toán này thành công không chỉ thể hiện khả năng "hiểu" ảnh của máy mà còn mở ra nhiều ứng dụng thực tế như: hỗ trợ người khiếm thị tiếp cận hình ảnh thông qua mô tả bằng giọng nói, tự động gợi ý caption cho ảnh trên mạng xã hội, hỗ trợ tìm kiếm ảnh theo nội dung văn bản, và nhiều ứng dụng trong thương mại điện tử, y tế hay an ninh.

Trước đây, các phương pháp tiếp cận bài toán mô tả ảnh thường kết hợp giữa mạng nơ-ron tích chập (CNN) để trích xuất đặc trưng toàn cục của hình ảnh và mạng hồi tiếp (RNN hoặc LSTM) để sinh câu. Tuy nhiên, các mô hình này gặp nhiều hạn chế do việc nén toàn bộ bức ảnh vào một vector cố định thường làm mất mát các thông tin chi tiết nhỏ hoặc mối quan hệ không gian giữa các đối tượng. Sự ra đời của Cơ chế Chú ý (Attention Mechanism) kết hợp với các hệ thống phát hiện đối tượng tiên tiến như Faster R-CNN (trong Detectron2) đã mở ra một hướng đi mới hiệu quả hơn. Thay vì nhìn toàn bộ ảnh một cách đại khái, hệ thống hiện đại có thể trích xuất

đặc trưng của từng vùng đối tượng cụ thể (Region-based features), giúp mô hình ngôn ngữ "tập trung" chính xác vào sự vật đang được mô tả tại mỗi bước thời gian.

Từ thực tế đó, đề tài “Mô tả ảnh tự động sử dụng mạng phát hiện đối tượng Detectron2 và Cơ chế Chú ý (Attention Mechanism)” được thực hiện nhằm tận dụng những tiến bộ này để xây dựng một hệ thống tối ưu. Bằng cách kết hợp giữa đặc trưng vùng ảnh chất lượng cao do Detectron2 trích xuất và khả năng điều hướng ngữ nghĩa của mạng LSTM tích hợp Attention, hệ thống hướng tới việc tạo ra những mô tả có chất lượng cao, chi tiết và phù hợp với nội dung thực tế trong ảnh. Đề tài không chỉ có ý nghĩa về mặt học thuật mà còn là tiền đề cho nhiều ứng dụng thông minh trong cuộc sống.

1.2 Tổng quan tình hình nghiên cứu

Sự phát triển của bài toán mô tả ảnh tự động (Image Captioning) có thể được tóm tắt qua ba giai đoạn chuyển dịch công nghệ chính:

Giai đoạn 1: Tiếp cận dựa trên Đặc trưng Toàn cục (Global Features) Các nghiên cứu khởi đầu (tiêu biểu là mô hình Show and Tell, 2015) sử dụng mạng CNN để mã hóa toàn bộ bức ảnh thành một vector đặc trưng duy nhất trước khi đưa vào mạng RNN sinh câu. Mặc dù đặt nền móng quan trọng, phương pháp này bộc lộ hạn chế lớn là "nút thắt cổ chai" (information bottleneck), khiến mô hình thường bỏ sót các chi tiết nhỏ và không thể hiện được mối quan hệ không gian giữa các đối tượng trong những bức ảnh phức tạp.

Giai đoạn 2: Tiếp cận dựa trên Đặc trưng Vùng và Cơ chế Chú ý (Region-based Attention) Đây là bước đột phá về độ chính xác chi tiết và là nền tảng công nghệ cốt lõi của đề tài này. Sự ra đời của cơ chế Chú ý (Attention Mechanism) kết hợp với các hệ thống phát hiện đối tượng mạnh mẽ (như Faster R-CNN trong thư viện Detectron2) cho phép mô hình trích xuất và xử lý đặc trưng của từng vùng đối tượng cụ thể (Region of Interest - RoI). Các nghiên cứu tiêu biểu như Bottom-Up and Top-Down Attention [5] hay Stacked Cross Attention [6] đã chứng minh rằng

việc "nhìn" vào từng vùng ảnh giúp mô hình sinh ra các mô tả chính xác và cụ thể hơn so với cách tiếp cận toàn cục truyền thống.

Giai đoạn 3: Xu hướng Mô hình Nền tảng lớn (Foundation Models) Trong bối cảnh hiện nay (2025), xu hướng nghiên cứu đã mở rộng sang các Mô hình Đa phương thức lớn (Large Vision-Language Models) với khả năng tổng quát hóa cao. Tuy nhiên, nhược điểm của các mô hình này là yêu cầu tài nguyên tính toán khổng lồ, dữ liệu huấn luyện đại trà và thường hoạt động như một "hộp đen" khó giải thích cơ chế bên trong.

Định vị nghiên cứu của đề tài: Trước thực tế đó, đề tài lựa chọn hướng đi tối ưu hóa mô hình chuyên biệt dựa trên đặc trưng vùng. Thay vì chạy theo cuộc đua về quy mô tham số, nghiên cứu tập trung khai thác sức mạnh của Detectron2 và cơ chế Bottom-Up Top-Down Attention để xây dựng một hệ thống "nhẹ" (lightweight), minh bạch và hiệu quả. Hướng tiếp cận này đảm bảo khả năng vận hành tốt trên tài nguyên phần cứng hạn chế của sinh viên mà vẫn đạt được độ chính xác chi tiết (fine-grained accuracy) cạnh tranh trong các miền dữ liệu cụ thể như Flickr30k.

1.3 Lý do chọn đề tài

Sự bùng nổ của trí tuệ nhân tạo đã thúc đẩy mạnh mẽ sự giao thoa giữa Thị giác máy tính và Xử lý ngôn ngữ tự nhiên, tiêu biểu là bài toán mô tả ảnh tự động (Image Captioning). Đây là thách thức lớn mang tính đa mô thức, yêu cầu hệ thống không chỉ nhận diện đối tượng mà còn phải thấu hiểu ngữ cảnh để sinh ra câu văn tự nhiên. Về mặt thực tiễn, bài toán mang giá trị nhân văn sâu sắc trong việc hỗ trợ người khiếm thị tiếp cận thông tin hình ảnh, đồng thời giúp tối ưu hóa việc quản lý và tìm kiếm nội dung trên các nền tảng số.

Về mặt công nghệ, các phương pháp truyền thống sử dụng đặc trưng toàn cục thường gặp hạn chế trong việc nắm bắt chi tiết nhỏ. Sự xuất hiện của nền tảng Detectron2 cho phép trích xuất đặc trưng vùng (Region-based features) hiệu quả, giúp mô hình tập trung chính xác vào các đối tượng quan trọng như con người hay vật thể. Khi kết hợp nguồn dữ liệu giàu ngữ nghĩa này với Cơ chế Chú ý (Attention

Mechanism), hệ thống có khả năng sinh ra các mô tả logic và sát với thực tế hơn hẳn so với việc chỉ sử dụng CNN đơn thuần.

Xuất phát từ nhu cầu thực tiễn và tiềm năng công nghệ đó, đề tài “Mô tả ảnh tự động sử dụng mạng phát hiện đối tượng Detectron2 và Cơ chế Chú ý” được lựa chọn. Đề tài không chỉ hướng tới việc xây dựng một mô hình có độ chính xác cao mà còn là cơ hội để làm chủ các kỹ thuật học sâu tiên tiến (như Faster R-CNN và Attention LSTM), tạo tiền đề cho các ứng dụng thông minh phục vụ đời sống.

1.4 Ý nghĩa của nghiên cứu

1.4.1 Ý nghĩa lý thuyết

Đề tài nghiên cứu về mô tả ảnh tự động mang ý nghĩa lý thuyết sâu sắc trong việc kết hợp và ứng dụng những tiến bộ mới nhất của trí tuệ nhân tạo vào một bài toán mang tính chất liên ngành cao. Việc giải quyết bài toán này đòi hỏi sự hiểu biết toàn diện về các mô hình học sâu trong cả thị giác máy tính và xử lý ngôn ngữ tự nhiên. Qua đó, đề tài góp phần củng cố và mở rộng kiến thức lý thuyết về mạng nơ-ron tích chập (CNN), mạng hồi tiếp (RNN/LSTM), và đặc biệt là Cơ chế Chú ý (Attention Mechanism) – thành phần cốt lõi giúp mô hình "tập trung" vào các chi tiết quan trọng.

Bên cạnh đó, nghiên cứu này còn giúp làm rõ vai trò và hiệu quả của việc sử dụng đặc trưng vùng ảnh thông qua các mô hình phát hiện đối tượng như Faster R-CNN trong Detectron2, từ đó khẳng định tầm quan trọng của biểu diễn dữ liệu đa mức (multi-level features) và đa vùng (region-based features) trong việc tăng cường chất lượng sinh ngôn ngữ từ hình ảnh. Việc khai thác cơ chế Attention ở nhiều cấp độ – điển hình là kiến trúc Bottom-Up Top-Down Attention – giúp người học hiểu rõ hơn cách mô hình học sâu có thể mô hình hóa mối quan hệ phức tạp giữa các thành phần dữ liệu khác nhau.

Ngoài ra, thông qua quá trình xây dựng mô hình thực tế và huấn luyện trên tập dữ liệu chuẩn như Flickr30k, người thực hiện cũng hiểu sâu hơn về cách xử lý dữ liệu

hình ảnh, tiền xử lý ngôn ngữ, kỹ thuật nhúng từ (embedding), cơ chế beam search, kỹ thuật đánh giá kết quả bằng các chỉ số BLEU, CIDEr, METEOR... Những kiến thức và kinh nghiệm này không chỉ mang lại giá trị lý thuyết mà còn là nền tảng vững chắc để tiếp cận các bài toán đa mô thức khác trong lĩnh vực trí tuệ nhân tạo.

1.4.2 Ý nghĩa thực tiễn

Bên cạnh ý nghĩa lý thuyết, đề tài mô tả ảnh tự động còn mang lại giá trị thực tiễn rõ rệt trong nhiều lĩnh vực đời sống, đặc biệt trong bối cảnh xã hội ngày càng hướng đến tự động hóa và tương tác thông minh giữa con người và máy móc. Hệ thống có khả năng tạo ra mô tả ngôn ngữ tự nhiên từ hình ảnh không chỉ đơn thuần là một ứng dụng trí tuệ nhân tạo mà còn là một bước tiến trong việc xây dựng cầu nối giao tiếp giữa con người và hệ thống máy tính. Một trong những ứng dụng quan trọng nhất là hỗ trợ người khiếm thị tiếp cận nội dung hình ảnh thông qua mô tả văn bản hoặc giọng nói, giúp họ hòa nhập tốt hơn với thế giới số.

Ngoài ra, trong các nền tảng mạng xã hội, thương mại điện tử và quản lý hình ảnh số lượng lớn, hệ thống mô tả ảnh tự động có thể giúp tiết kiệm đáng kể thời gian trong việc tạo caption, gán nhãn, phân loại ảnh, đồng thời cải thiện khả năng tìm kiếm theo ngữ nghĩa. Với sự gia tăng khối lượng dữ liệu phi cấu trúc như ảnh và video, các hệ thống AI có khả năng tự động mô tả nội dung là giải pháp cần thiết để giảm thiểu phụ thuộc vào xử lý thủ công. Trong lĩnh vực y tế, mô hình tương tự cũng có thể được mở rộng để mô tả hình ảnh y khoa, hỗ trợ bác sĩ trong việc phân tích và báo cáo. Trong lĩnh vực giáo dục, mô tả ảnh còn giúp phát triển các công cụ học tập hỗ trợ học sinh hiểu nội dung trực quan thông qua văn bản.

Bằng cách sử dụng các mô hình hiện đại như Detectron2 và các kiến trúc Attention LSTM, hệ thống được xây dựng trong đề tài có thể đóng vai trò nền tảng cho nhiều ứng dụng thông minh trong tương lai. Việc triển khai thành công mô hình này cho thấy tính khả thi và hiệu quả của việc kết hợp giữa kỹ thuật trích xuất vùng ảnh với mô hình sinh ngôn ngữ, từ đó tạo điều kiện thuận lợi để mở rộng và tùy biến theo từng bài toán cụ thể trong các ngành nghề khác nhau.

1.5 Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu trọng tâm của đề tài là hệ thống mô tả ảnh tự động (Image Captioning System) được xây dựng dựa trên nền tảng các kỹ thuật học sâu tiên tiến. Cụ thể, nghiên cứu tập trung sâu vào cơ chế tích hợp giữa thị giác máy tính và xử lý ngôn ngữ tự nhiên, trong đó sử dụng mạng phát hiện đối tượng Faster R-CNN (thuộc thư viện Detectron2) làm bộ trích xuất đặc trưng và mạng hồi quy LSTM tích hợp cơ chế Chú ý (Attention Mechanism) làm bộ sinh văn bản. Thay vì tiếp cận theo hướng xử lý toàn bộ ảnh như các phương pháp truyền thống, đối tượng phân tích chính ở đây là các vector đặc trưng vùng (Region-based features) đại diện cho từng đối tượng cụ thể trong ảnh, nhằm giúp hệ thống đạt được khả năng "hiểu" ngữ cảnh chi tiết và sinh ra các mô tả chính xác, bám sát nội dung thị giác.

Về phạm vi dữ liệu và ngôn ngữ, đề tài giới hạn việc nghiên cứu và thử nghiệm trên các ảnh tĩnh có định dạng chuẩn, không bao gồm việc xử lý video hay dữ liệu động. Bộ dữ liệu được lựa chọn để huấn luyện và đánh giá là Flickr30k – một chuẩn benchmark phổ biến trong cộng đồng nghiên cứu, bao gồm khoảng 31.000 hình ảnh thực tế với 5 câu mô tả tương ứng cho mỗi ảnh. Ngôn ngữ mục tiêu của các câu mô tả được giới hạn là tiếng Anh. Lựa chọn này nhằm tận dụng tối đa hiệu quả của các bộ từ điển và mô hình ngôn ngữ đã được tiền huấn luyện (pretrained embeddings), đồng thời đảm bảo tính khách quan và thuận tiện khi so sánh kết quả với các công trình nghiên cứu quốc tế khác.

Về phạm vi kỹ thuật, nghiên cứu không tập trung xây dựng lại các kiến trúc mạng nơ-ron từ đầu (from scratch) mà khai thác sức mạnh của phương pháp chuyển giao tri thức (Transfer Learning). Cụ thể, hệ thống sử dụng backbone ResNet-101 đã được huấn luyện trước trên tập dữ liệu Visual Genome để đảm bảo khả năng nhận diện đối tượng tốt nhất. Trọng tâm của đề tài nằm ở việc thiết kế, tinh chỉnh (fine-tune) và tối ưu hóa cơ chế Attention (Bottom-Up Top-Down) để kết nối hiệu quả giữa đặc trưng vùng và từ ngữ. Các vấn đề mở rộng như xây dựng ứng dụng di động thực tế, xử lý đa ngôn ngữ (bao gồm tiếng Việt), hay tối ưu hóa mô hình cho

các thiết bị phần cứng nhúng nằm ngoài phạm vi của đề án này và sẽ được đề xuất như những hướng phát triển trong tương lai.

1.6 Cấu trúc đề án

Đề án gồm 5 chương chính. Chương 1 trình bày tổng quan đề tài, lý do chọn, mục tiêu và phạm vi nghiên cứu. Chương 2 cung cấp cơ sở lý thuyết về các mô hình học sâu như, CNN, LSTM, Transformer và Fast R_CNN. Chương 3 mô tả quá trình xây dựng mô hình, bao gồm kiến trúc hệ thống, cách trích xuất đặc trưng ảnh và sinh câu mô tả. Chương 4 trình bày kết quả thực nghiệm, đánh giá hiệu quả mô hình qua các chỉ số BLEU, METEOR, CIDEr. Cuối cùng, chương 5 là phần kết luận và định hướng phát triển sau đề tài.

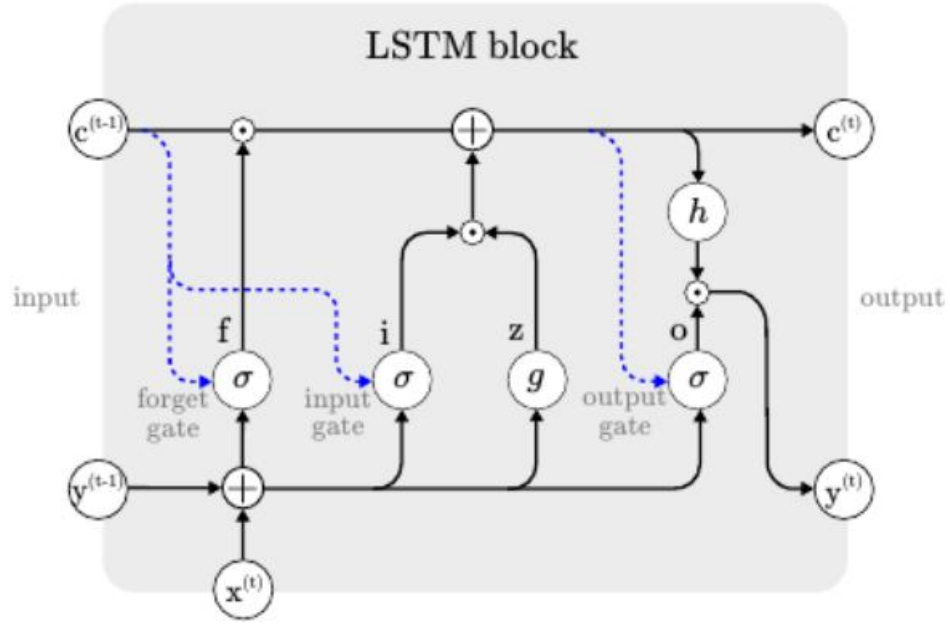
Chương 2 CƠ SỞ LÝ THUYẾT

2.1 Long Short-Term Memory (Bộ nhớ Ngắn – Dài hạn)

Mô hình LSTM được thiết kế để giải quyết các vấn đề gradient biến mất và bùng nổ gradient khi học các phụ thuộc dài hạn. Điều này có thể được giải quyết bằng cách sử dụng vòng lặp lỗi hằng số (CEC), giúp giữ lại tín hiệu lỗi trong cell qua mỗi đơn vị LSTM. Mỗi cell của LSTM thực chất là một mạng hồi quy nhỏ, trong đó CEC được bổ sung thêm cổng vào (input gate) và cổng ra (output gate), tạo thành ô bộ nhớ (memory cell) có khả năng lưu giữ thông tin. Các kết nối hồi quy trong cell thể hiện cơ chế phản hồi với độ trễ một bước thời gian.

Một đơn vị LSTM “thuần” (vanilla LSMT) gồm một ô nhớ (cell) cổng vào (input gate), cổng ra (output gate), và cổng quên (forget gate). Cổng quên ban đầu không có trong thiết kế gốc mà được thêm vào sau để cho phép đặt lại trạng thái. Cell có nhiệm vụ ghi nhớ trong thời gian tùy ý, còn các cổng điều chỉnh luồng thông tin liên quan đến cell. Trong phần còn lại, thuật ngữ LSTM sẽ ám chỉ phiên bản vanilla này, dù nó không phải luôn là lựa chọn tốt nhất trong mọi trường hợp.

Tóm lại, kiến trúc LSTM gồm một tập các mạng con được nối hồi quy, được gọi là khối bộ nhớ (memory blocks). Ý tưởng chính của khối bộ nhớ là duy trì trạng thái qua thời gian và điều chỉnh luồng thông tin thông qua các đơn vị gating phi tuyến.



Hình 2.1 mô hình LSTM

Để làm rõ cách mô hình LSTM hoạt động, hãy giả sử một mạng gồm N khối xử lý và M đầu vào. Quá trình lan truyền tiến (forward pass) trong hệ thống mạng hồi quy này được mô tả như sau.

Block input: Bước này chịu trách nhiệm cập nhật thành phần đầu vào của khối, kết hợp giữa đầu vào hiện tại $x^{(t)}$, và đầu $y^{(t-1)}$ của đơn vị LSTM ở vòng lặp trước.

Việc này được thực hiện theo công thức:

$$z^{(t)} = g(W_z x^{(t)} + R_z y^{(t-1)} + b_z) \quad (1)$$

trong đó W_z và R_z lần lượt là trọng số ứng với $x^{(t)}$ và $y^{(t-1)}$, còn b_z là vector bias.

Input gate: Trong bước này, ta cập nhật cổng vào (input gate), kết hợp giữa đầu vào hiện tại $x^{(t)}$, đầu ra trước đó $y^{(t-1)}$, và giá trị của cell $c = (t-1)$ ở vòng lặp trước. Công thức:

$$i^t = \sigma(W_i x^{(t)} + R_i y^{(t-1)} + p_i \odot c^{(t-1)} + b_i) \quad (2)$$

trong đó:

\odot là phép nhân từng phần tử (element-wise),

W_i, R_i, p_i là các trọng số tương ứng với $x^{(t)}$, $y^{(t-1)}$ và $c^{(t-1)}$

b_i là vector bias của thành phần này.

Ở các bước trước, lớp LSTM đã xác định những thông tin nào nên được giữ lại trong trạng thái cell c^t của mạng. Điều này bao gồm việc lựa chọn giá trị ứng viên z^t tức những giá trị có thể được thêm vào cell và giá trị kích hoạt i^t của cổng vào.

Forget gate: Ở bước này, đơn vị LSTM xác định những thông tin nào cần bị loại bỏ khỏi trạng thái cell trước đó $c^{(t-1)}$. Do đó, giá trị kích hoạt f^t của cổng quên tại thời điểm t được tính dựa trên đầu vào hiện tại $x^{(t)}$, đầu ra trước đó $y^{(t-1)}$, trạng thái cell trước đó $c^{(t-1)}$, các kết nối peephole và các tham số bias b_f của cổng quên.

Được thực hiện như sau:

$$f^{(t)} = \sigma(W_f x^{(t)} + R_f y^{(t-1)} + p_f \odot c^{(t-1)} + b_f) \quad (3)$$

trong đó W_f, R_f, p_f lần lượt là các trọng số tương ứng với $x^{(t)}$, $y^{(t-1)}$ và $c^{(t-1)}$, b_f là vector bias của forget gate.

Cell: Bước này tính toán giá trị ô nhớ bằng cách kết hợp đầu vào của khối $z^{(t)}$, giá trị cổng vào i^t , giá trị forget gate $f^{(t)}$, cùng với giá trị cell trước đó $c^{(t-1)}$. Cell được thực hiện như sau:

$$c^{(t)} = z^{(t)} \odot i^{(t)} + c^{(t-1)} \odot f^{(t)} \quad (4)$$

Output gate: Bước này, tính toán đầu ra của cell, kết hợp giữa đầu vào hiện tại $x^{(t)}$, đầu ra trước đó của LSTM $y^{(t-1)}$, và giá trị ô nhớ $c^{(t)}$ ở thời điểm hiện tại (thông qua peephole). Theo công thức:

$$o^t = \sigma(W_o x^{(t)} + R_o y^{(t-1)} + p_o \odot c^{(t)} + b_o) \quad (5)$$

trong đó: W_o, R_o, p_o là các trọng số tương ứng với $x^{(t)}, y^{(t-1)}$ và $c^{(t)}$, b_o là vector bias của cổng ra.

Block output: Cuối cùng, ta tính đầu ra của khối LSTM bằng cách kết hợp giá trị ô nhớ hiện tại với giá trị cổng ra:

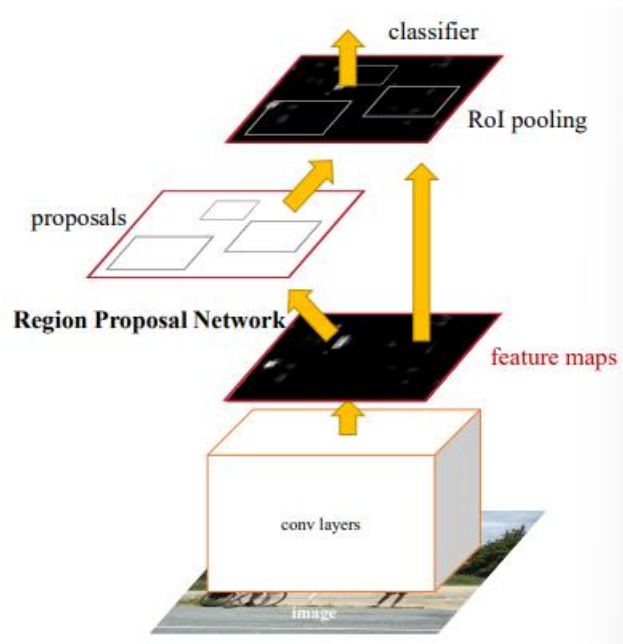
$$y^{(t)} = g(c^{(t)}) \odot o^{(t)}. \quad (6)$$

Trong các bước trên, các hàm σ , g và h biểu thị các hàm kích hoạt phi tuyến áp dụng theo từng phần tử. Hàm sigmoid $\sigma(x) = \frac{1}{1+e^{1-x}}$ được sử dụng làm hàm kích hoạt cho các cổng (gate activation function). Hàm tanh $g(x) = h(x) = \tanh(x)$ thường được dùng làm hàm kích hoạt cho đầu vào khối (block input) và đầu ra của khối (block output).

Cơ chế gating trong LSTM đã truyền cảm hứng cho việc phát triển highway networks, nơi các cổng được dùng để giúp thông tin truyền qua nhiều lớp mà không bị cản trở. Điều này củng cố thêm bằng chứng rằng cơ chế cổng của LSTM hoạt động hiệu quả trong việc duy trì và điều khiển dòng thông tin.

2.2 Faster R-CNN (Shaoqing Ren,2016)

Faster R-CNN là một hệ thống phát hiện đối tượng gồm hai module chính: Module thứ nhất là một mạng nơ-ron fully convolutional dùng để đề xuất vùng (region proposals). Module thứ hai là bộ phân loại Fast R-CNN sử dụng các vùng được đề xuất này. Toàn bộ hệ thống là một mạng thống nhất, hai module kết hợp thành một mạng duy nhất để phát hiện đối tượng (Hình 2). Theo cách nói phổ biến về mạng nơ-ron với cơ chế “attention”, module RPN có vai trò chỉ cho Fast R-CNN biết vị trí cần tìm.



Hình 2.2 Mô hình Faster R-CNN

Region Proposal Networks

Mạng Đề Xuất Vùng (RPN) nhận một ảnh đầu vào (với kích thước bất kì) và xuất ra một tập các vùng đề xuất hình chữ nhật, mỗi vùng có điểm số thể hiện khả năng chứa đối tượng. (objectness). Chúng tôi mô hình hóa quá trình này bằng một mạng fully convolutional được mô tả trong phần này. Vì mục tiêu cuối cùng là chia sẻ phần tính toán với Fast R-CNN, RPN và Fast R-CNN dùng chung các lớp convolution.

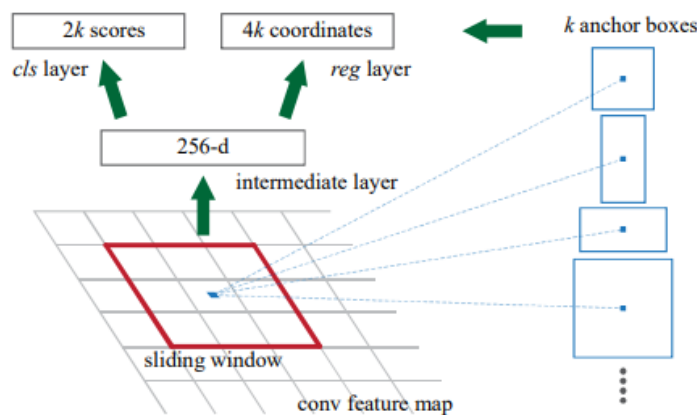
Trong thí nghiệm, chúng tôi sử dụng mô hình Zeiler và Fergus (FZ) với 5 lớp convolution chia sẻ và Simonyan and Zisserman (VGG-16) với 13 lớp convolution chia sẻ.

Để tạo ra các vùng đề xuất, chúng tôi trọt một mạng nhỏ (small network) trên feature map của lớp convolution được chia sẻ cuối cùng. Mạng nhỏ này nhận đầu vào là một cửa sổ không gian kích thước $n \times n$ nằm trên feature map, biến nó thành một đặc trưng có số chiều thấp hơn (256 cho ZF, 512 cho VGG), sau đó đưa vào hai nhánh fully-connected song song: một nhánh dự đoán tọa độ hộp (reg), một nhánh

dự đoán objectness (cls). Trong bài này, chúng tôi dùng $n = 3$, mặc dù receptive field hiệu dụng trên ảnh đầu vào là rất lớn (171 pixel đối với ZF và 228 pixel đối với VGG). Mạng lưới nhỏ này được minh họa tại một vị trí duy nhất trong Hình 3 (bên trái).

Lưu ý: Vì mạng nhỏ hoạt động theo kiểu trượt cửa sổ (sliding-window), các tầng fully-connected được chia sẻ cho tất cả các vị trí không gian.

Kiến trúc này được triển khai một cách tự nhiên với một lớp convolution $n \times n$ theo sau là hai lớp convolution 1×1 tương ứng với hai nhánh reg và cls.



Hình 2.3 Region Proposal Network

Anchors

Tại mỗi vị trí của sliding window, chúng tôi đồng thời dự đoán nhiều vùng đề xuất, trong đó số lượng tối đa các đề xuất có khả thi tối đa tại mỗi vị trí được ký hiệu là k . Vì thế, tầng reg có $4k$ đầu ra (dự đoán tọa độ k hộp) và tầng cls có $2k$ đầu ra (dự đoán có/không có đối tượng) cho từng đề xuất.

K vùng đề xuất này được tham chiếu dựa trên k hộp chuẩn, mà chúng tôi gọi là anchors. Một anchor được đặt tại đúng tâm của sliding window đang xét, và được gán một scale (kích thước) và một aspect ratio (tỷ lệ khung hình) (Hình 3, bên trái).

Theo cấu hình mặc định, chúng tôi sử dụng 3 scales và 3 aspect ratios, tạo thành $k = 9$ anchors tại mỗi vị trí trượt. Đối với feature map có kích thước $W \times H$ (thường là ~ 2.400), tổng số anchors là $W \times H \times k$.

Translation-Invariant Anchors

Một tính chất quan trọng của phương pháp của chúng tôi là nó bất biến theo phép tịnh tiến (translation invariant), cả ở phần anchors lẫn các hàm dự đoán proposal dựa trên các anchors. Nghĩa là nếu đối tượng trong ảnh di chuyển, thì proposal sinh ra cũng di chuyển tương ứng và cùng một hàm dự đoán phải có khả năng đưa ra proposal đúng ở cả hai vị trí. Tính bất biến theo phép tịnh tiến này được đảm bảo trong phương pháp của chúng tôi.

Ngược lại với MultiBox, vốn sử dụng 800 anchors sinh bằng k-means và vốn không bất biến theo tịnh tiến, nên không đảm bảo có thể tạo ra một proposal giống nhau khi đối tượng dịch chuyển vị trí. Nhờ tính chất translation invariant mà kích thước mô hình được giảm nhỏ gọn hơn. Trong khi phương pháp của chúng tôi sử dụng tầng đầu ra dạng convolution với kích thước $(4 + 2) \times 9$ khi $k = 9$ anchors thì MultiBox cần tới 6.1×10^6 tham số, tức nhiều hơn hai bậc độ lớn. Ngay cả khi tính thêm các lớp chiếu đặc trưng (feature projection layers) thì mô hình của chúng tôi vẫn nhỏ hơn đáng kể. Chúng tôi kỳ vọng phương pháp của mình có nguy cơ overfitting thấp hơn trên các bộ dữ liệu nhỏ như PASCAL VOC.

Multi-Scale Anchors as Regression References

Phương pháp của chúng tôi dùng anchors đa tỉ lệ (multi-scale) và đa tỉ lệ khung hình (aspect ratio) như các hộp tham chiếu để giải quyết bài toán đa kích thước đối tượng. Khác với các cách truyền thống trước đây dùng image/feature pyramids (tốn thời gian) hoặc sliding windows với nhiều kích thước/bộ lọc (phức tạp và phải huấn luyện nhiều mô hình) thì phương pháp anchor-based chỉ cần một feature map ở một scale duy nhất và một kích thước filtro duy nhất vẫn có thể xử lý đa tỉ lệ nhờ dự đoán hộp dựa theo các anchor box có nhiều scale và aspect ratio. Chúng tôi trình bày các thí nghiệm về hiệu ứng của lược đồ này đối với việc xử lý nhiều tỷ lệ và

kích thước (bảng 8). Nhờ thiết kế multi-scale dựa trên anchors này, chúng tôi có thể sử dụng một ảnh ở một scale duy nhất, giống như Fast R-CNN, nhưng vẫn giải quyết tốt bài toán đa kích thước đối tượng mà không cần thêm chi phí tính toán cho pyramids.

Loss Function

Trong RPN, mỗi anchor được gán nhãn nhị phân (object / non-object) để phục vụ huấn luyện. Chúng tôi gán nhãn dương cho hai loại anchor: (i) nó có IoU cao nhất với một ground-truth box, hoặc (ii) $\text{IoU} > 0.7$ với bất kỳ ground-truth box nào. Lưu ý một single ground-truth box có thể dán nhãn dương cho nhiều anchor. Thông thường, điều kiện (ii) là đủ để xác định các mẫu dương, nhưng chúng tôi vẫn giữ điều kiện (i) vì trong một vài trường hợp hiếm, điều kiện (ii) không tìm được anchor dương nào. Chúng tôi gán nhãn âm nếu $\text{IoU} < 0.3$ với tất cả ground-truth boxes. Các anchor còn lại không tham gia tính loss.

Với các định nghĩa trên, chúng tôi tối ưu một hàm mục tiêu theo dạng multi-task loss giống Fast R-CNN. Hàm mất mát cho một ảnh được định nghĩa như sau:

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{ds}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*).$$

Trong đó, i là chỉ số của anchor trong một mini-batch, và p_i là xác suất dự đoán anchor i chứa đối tượng. Nhãn ground-truth p_i^* bằng 1 nếu anchor là dương và bằng 0 nếu anchor là âm. t_i là vector 4 giá trị tham số hóa tọa độ bounding box dự đoán, và t_i^* là vector của hộp ground-truth tương ứng với anchor dương.

Hàm mất mát phân loại L_{cls} là log-loss cho hai lớp (object / not object). Đối với hồi quy, chúng tôi sử dụng:

$$L_{reg}(t_i, t_i^*) = R(t_i - t_i^*)$$

Hai thành phần mất mát được chuẩn hóa lần lượt bởi N_{cls} và N_{reg} , và được cân bằng bằng một tham số λ .

Phần cls được chuẩn hóa theo kích thước minibatch (256 anchors), phần reg chuẩn hóa theo số vị trí anchors ($N_{reg} \sim 2,400$). Chúng tôi đặt mặc định $\lambda = 10$ để hai phần cls và reg có trọng số tương đương và kết quả thực nghiệm chỉ ra rằng kết quả không nhạy với λ . Lưu ý cách chuẩn hóa này không bắt buộc và có thể được đơn giản hóa.

Đối với hồi quy bounding box, chúng tôi dùng kiểu tham số hóa 4 tọa độ theo tham số hóa chuẩn từ R-CNN:

$$t_x = \frac{x - x_a}{w_a}, t_y = \frac{y - y_a}{h_a}, t_w = \log\left(\frac{w}{w_a}\right), t_h = \log\left(\frac{h}{h_a}\right)$$

$$t_x^* = \frac{x^* - x_a}{w_a}, t_y^* = \frac{y^* - y_a}{h_a}, t_w^* = \log\left(\frac{w^*}{w_a}\right), t_h^* = \log\left(\frac{h^*}{h_a}\right)$$

Với x, y, w, h là tâm, chiều rộng và chiều cao của box. Các biến x, x_a, x^* lần lượt là tọa độ tâm của box dự đoán, anchor, và ground-truth (tương tự cho y, w, h). Điều này có thể được xem như bài toán hồi quy từ anchor đến một ground-truth gần đó. Điểm khác biệt với các phương pháp dựa trên RoI trước đó là: RPN không dùng RoI pooling, mà hồi quy box từ cửa sổ đặc trưng cố định 3×3 , và cần k bộ hồi quy khác nhau (ứng với k anchors). Các bộ hồi quy này không chia sẻ trọng số, giúp dự đoán được nhiều kích thước box khác nhau dù đặc trưng đầu vào cố định, nhờ thiết kế anchors.

Training RPNs

RPN được huấn luyện end-to-end bằng backpropagation (lan truyền ngược) và tối ưu hóa bằng stochastic gradient (SGD). sử dụng chiến lược image-centric sampling giống Fast R-CNN: mỗi mini-batch được lấy từ một ảnh duy nhất, trong đó có nhiều anchor dương và âm.

Có thể tối ưu hàm mất mát sử dụng toàn bộ anchors, nhưng điều này sẽ làm mô hình bị lệch về các mẫu âm do chúng chiếm số lượng lớn (vì số anchors âm chiếm phần lớn). Do đó, từ mỗi ảnh, chúng tôi ngẫu nhiên chọn 256 anchors để tính loss,

với tỉ lệ dương : âm tối đa là 1:1. Nếu số anchor dương nhỏ hơn 128, chúng tôi bổ sung các anchor âm để đủ 256 mẫu.

Các lớp mới được khởi tạo ngẫu nhiên bằng Gaussian (mean = 0, std = 0.01), còn các lớp convolution dùng chung được khởi tạo từ mô hình đã pretrain trên ImageNet, theo chuẩn của các phương pháp trước. Đối với ZF, chúng tôi fine-tune toàn bộ các lớp của mạng; đối với VGG, chỉ fine-tune từ conv3_1 trở lên để tiết kiệm bộ nhớ.

Quá trình huấn luyện dùng learning rate 0.001 cho 60.000 mini-batches, sau đó 0.0001 cho 20.000 mini-batches trên PASCAL VOC. Momentum đặt là 0.9, weight decay 0.0005. Toàn bộ hệ thống được triển khai bằng Caffe.

Sharing Features for RPN and Fast R-CNN (Chia sẻ Tính năng cho RPN và Fast R-CNN)

Chúng tôi đã mô tả cách huấn luyện một mạng để sinh các đề xuất vùng mà chưa xét đến mạng phát hiện đối tượng dựa trên vùng sẽ sử dụng những đề xuất này. Đối với mạng phát hiện, chúng tôi sử dụng Fast R-CNN. Tiếp theo, chúng tôi mô tả các thuật toán để học một mạng thống nhất gồm RPN và Fast R-CNN, trong đó hai mạng chia sẻ các lớp convolution (Hình 2). Nếu huấn luyện tách rời, mỗi mạng sẽ điều chỉnh backbone theo các hướng khác nhau, nên cần cơ chế huấn luyện chung, do đó chúng ta đề xuất ba phương pháp huấn luyện cho phép chia sẻ các lớp convolution giữa hai mạng.

Huấn luyện luân phiên: huấn luyện RPN trước, dùng các proposal mà RPN sinh ra để huấn luyện Fast R-CNN, rồi dùng mạng đã tinh chỉnh của Fast R-CNN để khởi tạo lại RPN; quy trình được lặp lại để hai module dần tối ưu và thống nhất backbone. Đây là phương pháp thực nghiệm chính của bài báo.

Huấn luyện kết hợp xấp xỉ: RPN và Fast R-CNN được gộp thành một mạng duy nhất trong quá trình huấn luyện; trong mỗi vòng SGD, bước forward sinh ra các region proposals. Các proposals này sau đó được xử lý giống như proposals cố định (pre-computed) trong huấn luyện Fast R-CNN. Trong backward pass, các gradient

từ loss của RPN và loss của Fast R-CNN sẽ được cộng lại tại các lớp convolution được chia sẻ. Phương pháp này bỏ qua đạo hàm theo tọa độ proposal nên chỉ là xấp xỉ nhưng nhanh hơn 25–50%, và cho kết quả gần tương đương.

Huấn luyện kết hợp không xấp xỉ (Non-approximate joint training): Như đã đề cập, các bounding box dự đoán bởi RPN cũng là hàm phụ thuộc vào đầu vào. Lớp RoI Pooling [2] trong Fast R-CNN nhận: đặc trưng convolutional, và cả tọa độ bounding box dự đoán. Vì vậy, một thuật toán backpropagation đúng lý thuyết phải tính cả gradient theo tọa độ bounding box. Điều này không được thực hiện trong phương pháp (ii). Để làm được điều này, cần một lớp RoI pooling khả vi theo tọa độ, ví dụ lớp RoI warping trong. Tuy nhiên, vấn đề này không được bàn sâu trong bài báo.

2.3 Bottom-Up Top-Down Attention

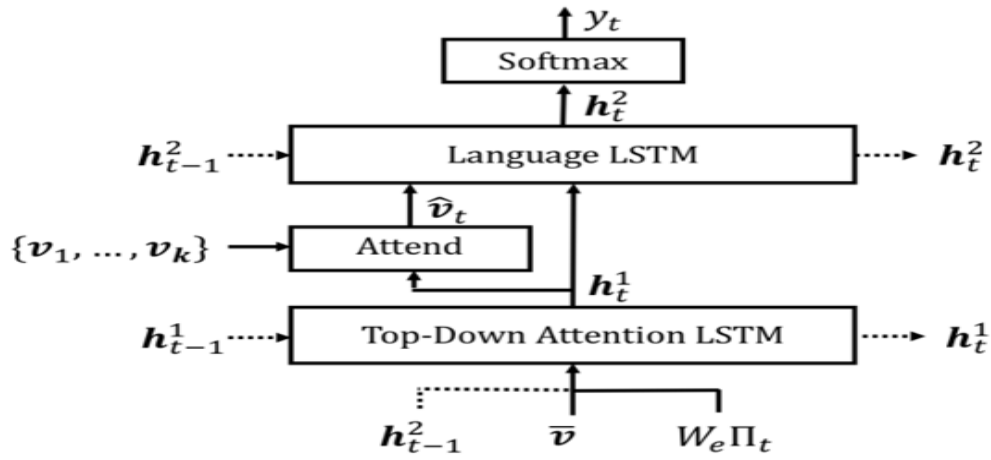
Captioning Model (Chú thích mô hình)

Với một tập đặc trưng ảnh V , mô hình tạo chú thích ảnh (captioning model) mà chúng tôi đề xuất sử dụng một cơ chế soft top-down attention để gán trọng số cho từng đặc trưng trong quá trình sinh chú thích, trong đó ngữ cảnh là chuỗi từ đã sinh ra một phần. Phương pháp này tương tự nhiều nghiên cứu trước nhưng nhờ các lựa chọn thiết kế đơn giản nhưng vẫn đạt hiệu quả cao, mô hình đạt hiệu suất cạnh tranh ngay cả khi không dùng bottom-up attention mô hình của chúng tôi vẫn đạt kết quả tương đương các mô hình hiện đại trên hầu hết các chỉ số đánh giá.

Mức độ tổng quát mô hình gồm hai lớp LSTM, sử dụng cách triển khai tiêu chuẩn. Mô tả hoạt động của LSTM tại một bước thời gian bằng ký hiệu:

$$h_t = LSTM(x_t, h_{t-1}) \quad (1)$$

x_t là vector đầu vào của LSTM và h_t là vector đầu ra, chúng tôi đã bỏ qua sự lan truyền của các ô nhớ (memory cell) để thuận tiện cho việc ký hiệu. Cuối cùng, chúng tôi mô tả cách xây dựng vector đầu vào x_t và đầu ra h_t cho từng lớp của mô hình. Toàn bộ mô hình captioning được minh họa trong Hình 3.



Hình 2.4 Caption Model base on Top-Down Attention

Top-Down Attention LSTM

Trong mô hình captioning, chúng tôi xem lớp LSTM thứ nhất là một mô hình attention hình ảnh theo hướng top-down, và lớp LSTM thứ hai là mô hình ngôn ngữ. Hai lớp được ký hiệu bằng chỉ số mũ trong các phương trình sau. Lưu ý rằng mô hình attention bottom-up đã được mô tả trong Mục 3.1, và trong mục này các đầu ra đó chỉ được xem như một tập đặc trưng V .

Vector đầu vào cho attention LSTM ở mỗi bước thời gian gồm:

- đầu ra của language LSTM ở bước trước,
- mean-pooled image feature $\bar{v} = \frac{1}{k} \sum_i v_i$
- một embedding của từ đã sinh ở bước trước.

Cụ thể:

$$x_t^1 = [h_{t-1}^2, \bar{v}, W_e \Pi_t] \quad (2)$$

Trong đó: $W_e \in \mathbb{R}^{E \times |\Sigma|}$ là ma trận embedding từng vựng Σ ; Π_t là one-hot của từ đầu vào tại bước t .

Điều này cung cấp toàn bộ bối cảnh ngôn ngữ + hình ảnh cho mô hình

Những thành phần này cung cấp cho attention LSTM đầy đủ bối cảnh về: trạng thái hiện tại của LSTM ngôn ngữ, nội dung tổng thể của ảnh, phần caption đã sinh ra. Embedding từ được học từ đầu, không dùng pretrained.

Dựa trên đầu ra h_t^1 của attention LSTM, tại bước t mô hình sinh ra trọng số attention chuẩn hóa $\alpha_{i,t}$ cho từng đặc trưng ảnh v_i như sau:

$$\alpha_{i,t} = w_\alpha^T \tanh(W_{va}v_i + W_{ha}h_t^1) \quad (3)$$

$$\alpha_t = \text{softmax}(\alpha_t) \quad (4)$$

Trong đó $W_{va} \in \mathbb{R}^{H \times V}$, $W_{ha} \in \mathbb{R}^{H \times M}$, và $w_\alpha \in \mathbb{R}^H$ là các tham số được học.

Đặc trưng ảnh được chú ý tại bước t , dùng làm đầu vào cho language LSTM, được tính bằng tổ hợp lồi của toàn bộ đặc trưng ảnh:

$$\hat{v}_t = \sum_{i=1}^K \alpha_{i,t} v_i \quad (5)$$

Language LSTM

Đầu vào của Language LSTM bao gồm đặc trưng ảnh đã được chú ý (attended image feature), ghép nối (concatenate) với đầu ra của Attention LSTM, được cho bởi:

$$x_t^2 = [\hat{v}_t, h_t^1] \quad (6)$$

Sử dụng ký hiệu $y_{1:T}$ để chỉ một chuỗi từ (y_1, \dots, y_T) , tại mỗi thời điểm t , phân phối có điều kiện trên tập các từ có thể sinh ra được mô tả như sau:

$$p(y_t | y_{1:t-1}) = \text{softmax}(W_p h_t^2 + b_p) \quad (7)$$

Trong đó $W_p \in \mathbb{R}^{|\Sigma| \times M}$ và $b_p \in \mathbb{R}^{|\Sigma|}$ là các trọng số và bias được học.

Phân phối trên toàn bộ chuỗi đầu ra được tính bằng tích của các phân phối có điều kiện:

$$p(y_{1:T}) = \prod_{t=1}^T p(y_t | y_{1:t-1}) \quad (8)$$

Objective

Cho trước một chú thích ground-truth $y_{1:T}^*$ và một mô hình captioning với tham số θ , ta tối ưu hàm mất mát cross-entropy:

$$L_{XE}(0) = \sum_{t=1}^T \log(p_0(y_t^* | y_{1:t-1}^*)) \quad (9)$$

Để so sánh công bằng với công trình gần đây, chúng tôi cũng báo cáo kết quả tối ưu theo CIDEr.

Bắt đầu từ mô hình đã được huấn luyện bằng cross-entropy, mục tiêu tiếp theo là tối thiểu hóa kỳ vọng âm của điểm số:

$$L_R(0) = -\mathbb{E}_{y_{1:T} \sim p_0} [r(y_{1:T})] \quad (10)$$

trong đó r là hàm điểm số (VD: CIDEr).

Theo phương pháp Self-Critical Sequence Training (SCST), gradient của hàm mất mát này được xấp xỉ:

$$\nabla_0 L_R(0) \approx -\left(r(\hat{y}_{1:T}) - r(y_{1:T}^g)\right) \nabla_0 \log p_0(\hat{y}_{1:T}) \quad (11)$$

$\hat{y}_{1:T}$: caption được mẫu từ mô hình.

$y_{1:T}^g$: caption baseline, được tạo bởi greedy decoding từ mô hình hiện tại.

$r(\hat{y}_{1:T}) - r(y_{1:T}^g)$: phần thưởng tương đối (advantage).

SCST (giống các thuật toán REINFORCE khám phá không gian các caption bằng cách lấy mẫu từ chính mô hình trong quá trình huấn luyện. Gradient này giúp tăng xác suất của những caption có điểm số cao hơn baseline.

Trong thực nghiệm, chúng tôi cũng dùng SCST nhưng tăng tốc bằng cách giới hạn không gian lấy mẫu. Cụ thể, chúng tôi chỉ lấy mẫu từ beam search decoding, nghĩa là các mẫu xuất phát từ beam thay vì toàn bộ phân phối. Như quan sát thực nghiệm, một caption lấy từ beam gần như luôn có điểm số cao, kể cả khi nó không phải chú thích có log-probability cao nhất.

Ngược lại, rất ít caption được sampling tự do (không giới hạn) có điểm số cao hơn caption sinh theo greedy. Nhờ phương pháp này, chúng tôi tối ưu CIDEr chỉ trong một epoch.

2.4 Thuật toán Self-Critical Sequence Training (SCST)

SCST là một dạng thuật toán học tăng cường (Reinforcement Learning) được thiết kế đặc biệt để giải quyết hai vấn đề chính trong huấn luyện mô hình sinh mô tả ảnh: Exposure Bias (sai số tích lũy) và sự không nhất quán giữa hàm tối ưu (Cross-Entropy) với các chỉ số đánh giá thực tế như CIDEr.

Hàm mục tiêu (Loss Function)

Trong quá trình huấn luyện, mô hình (được đặc trưng bởi bộ tham số θ) sẽ cố gắng giảm thiểu giá trị sai số L . Giá trị này được tính dựa trên phần thưởng kỳ vọng của các câu mô tả mà mô hình sinh ra:

$$L = -R(\mathbf{w}_s)$$

Trong đó:

- \mathbf{w}_s (Sampled sentence): Là chuỗi câu được lấy mẫu ngẫu nhiên từ phân phối xác suất của mô hình.
- $R(\mathbf{w}_s)$: Là điểm số phần thưởng (thường dùng chỉ số CIDEr) tính cho chuỗi \mathbf{w}_s khi so sánh với câu gốc.

Cơ chế Baseline "Tự phê bình"

Để mô hình biết được một câu sinh ra là "tốt" hay "tệ", SCST cần một cột mốc so sánh gọi là Baseline. Điểm đặc biệt của bài báo là dùng chính khả năng tốt nhất hiện tại của mô hình để làm cột mốc này.

$$\text{Baseline} = R(\hat{w})$$

Trong đó:

- \hat{w} (Greedy sentence): Là chuỗi câu thu được bằng cách luôn chọn từ có xác suất cao nhất ở mỗi bước (giải mã tham lam).
- $R(\hat{w})$: Là điểm phần thưởng của câu tốt nhất mà mô hình tự tạo ra.

Quy trình cập nhật trọng số

Mô hình sẽ tự điều chỉnh các tham số dựa trên sự chênh lệch giữa câu ngẫu nhiên (w_s) và câu tốt nhất (\hat{w}). Công thức cập nhật được hiểu đơn giản như sau:

$$\text{Độ thay đổi} = - (R(w_s) - R(\hat{w})) \times \text{Xác suất sinh câu}$$

Cơ chế này vận hành giống như một sự "khen thưởng" và "kỷ luật":

- Trường hợp $R(w_s)$ lớn hơn $R(\hat{w})$: Câu ngẫu nhiên lại cho kết quả tốt hơn cả câu tốt nhất hiện có. Mô hình sẽ nhận phần thưởng dương, giúp tăng xác suất lặp lại cách sinh câu này trong tương lai.
- Trường hợp $R(w_s)$ nhỏ hơn $R(\hat{w})$: Câu ngẫu nhiên cho kết quả tệ. Mô hình nhận phần thưởng âm, dẫn đến việc giảm xác suất sinh ra những chuỗi từ tương tự.

Khắc phục lỗi sai tích lũy (Exposure Bias)

Trong huấn luyện truyền thống, mô hình luôn được "cầm tay chỉ việc" bằng cách cho xem từ đúng của người dùng. Tuy nhiên, khi chạy thực tế, mô hình phải tự sinh từ và dựa vào chính từ đó để nói tiếp, dẫn đến việc một lỗi sai nhỏ ở đầu câu sẽ kéo theo cả câu bị sai (Exposure Bias).

Cơ chế SCST giải quyết triệt để vấn đề này bằng cách buộc mô hình phải huấn luyện trên chính những gì nó tự nói ra (ws). Nhờ vậy, mô hình học được cách tự sửa lỗi và duy trì được sự nhất quán về nội dung từ đầu đến cuối câu.

CHƯƠNG 3. KẾT QUẢ THỰC NGHIỆM

3.1 Dữ liệu thực nghiệm (Dataset)

Trong nghiên cứu này, bộ dữ liệu **MS COCO (Microsoft Common Objects in Context)** được lựa chọn làm nền tảng chính để huấn luyện và đánh giá mô hình do tính quy mô và sự phức tạp về ngữ cảnh của nó. Bộ dữ liệu bao gồm hơn **123.000 hình ảnh** thực tế, ghi lại đa dạng các hoạt động và sự kiện trong cuộc sống hàng ngày. Mỗi hình ảnh đi kèm với ít nhất 5 câu mô tả bằng ngôn ngữ tự nhiên, tạo nên tổng cộng hơn **555.000 câu mô tả**, giúp mô hình học được sự đa dạng cực lớn trong cách diễn đạt cùng một nội dung thị giác. Các thực thể trong ảnh được phân loại thành 80 nhóm đối tượng chính, cung cấp nguồn dữ liệu phong phú cho việc nhận diện và mô tả.

Về mặt tổ chức dữ liệu thực nghiệm, toàn bộ hình ảnh cùng các đặc trưng vùng đã trích xuất được phân chia nghiêm ngặt theo cấu hình chuẩn của Karpathy Split để đảm bảo tính khách quan:

- Tập huấn luyện (Training set) chiếm tỷ trọng lớn nhất với khoảng 113.287 mẫu (~93%) đóng vai trò nền tảng cho việc học tham số.
- Tập kiểm định (Validation set) gồm 5.000 mẫu (~3%) phục vụ công tác tinh chỉnh siêu tham số và kiểm soát hiện tượng quá khớp.
- Tập kiểm thử (Test set) với 5.000 mẫu (~3%) được giữ độc lập hoàn toàn để đo lường hiệu suất thực tế của mô hình.

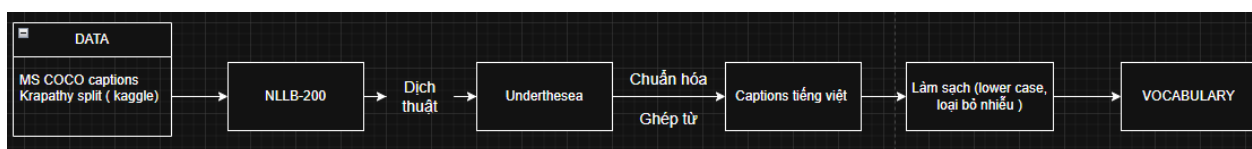
3.2 Tiền xử lý dữ liệu và Chuyển ngữ Tiếng Việt

Hệ thống sử dụng mô hình **NLLB-200 (No Language Left Behind)** làm lõi dịch máy để chuyển đổi hơn 555.000 câu mô tả từ tiếng Anh sang tiếng Việt, đảm bảo tính nhất quán về ngữ nghĩa và văn phong cho toàn bộ tập dữ liệu. Cơ chế này cho phép bảo toàn các mối quan hệ logic giữa hành động và thực thể từ câu gốc, đồng thời tận dụng khả năng biểu diễn không gian ngôn ngữ đa dạng của mạng nơ-ron để

tạo ra các câu mô tả tự nhiên, sát với nội dung thị giác nhất. Việc dịch thuật đồng nhất trên quy mô lớn giúp mạng LSTM dễ dàng học được phân phối xác suất từ vựng ổn định, hạn chế tối đa nhiễu ngữ nghĩa trong quá trình huấn luyện.

Đóng góp quan trọng nhất trong quy trình xử lý ngôn ngữ là việc tích hợp thư viện Underthesea để thực hiện tách từ ghép, giúp mô hình thích nghi với đặc thù cấu trúc của tiếng Việt. Bằng cách nhận diện ranh giới từ và nối các từ ghép bằng dấu gạch dưới (ví dụ: xe_buýt, biển_báo), hệ thống biến các cụm từ này thành một đơn vị ngữ nghĩa (token) duy nhất trong từ điển. Cơ chế này cho phép tầng Top-Down Attention thực hiện căn chỉnh chính xác: khi mô hình sinh ra một từ ghép, trọng số chú ý sẽ tập trung hoàn toàn vào đúng vùng đặc trưng đối tượng tương ứng trong ảnh, thay vì bị phân tán qua nhiều bước thời gian như đối với các tiếng rời rạc.

Quy trình cuối cùng thực hiện chuẩn hóa hình thức bằng cách chuyển văn bản về chữ thường, loại bỏ dấu câu và xây dựng bộ từ điển dựa trên các từ vựng xuất hiện tối thiểu 5 lần để triệt tiêu các lỗi dịch máy hiếm gặp. Không gian từ vựng này được bổ sung các token điều khiển hệ thống như <start>, <end> và <pad> để quản lý luồng dữ liệu và đồng bộ hóa độ dài chuỗi trong quá trình tính toán loss. Kết quả là một bộ dữ liệu văn bản sạch, có cấu trúc từ ghép chặt chẽ, sẵn sàng để kết hợp với đặc trưng vùng ảnh trong pha huấn luyện **Cross-Entropy** và **SCST**.

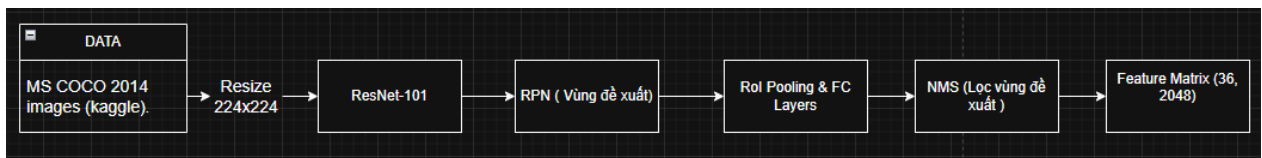


Hình 3.5 Mô hình dịch thuật

3.3 Cơ chế trích xuất đặc trưng thị giác (Image Features)

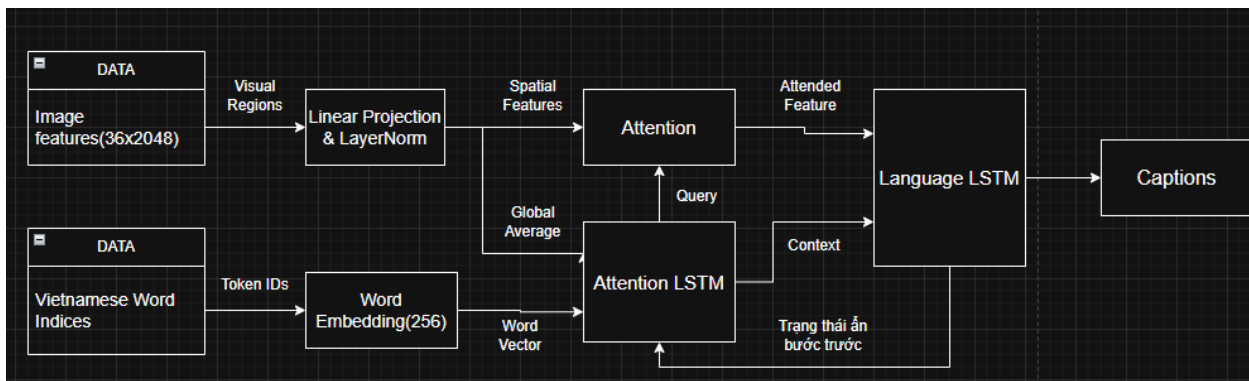
Thay vì sử dụng các đặc trưng toàn cục từ mạng CNN truyền thống, nghiên cứu này sử dụng bộ đặc trưng vùng (region-based features) được trích xuất sẵn từ công trình SCAN trên Kaggle

- **Kiến trúc Backbone:** Các đặc trưng được trích xuất thông qua mạng **Faster R-CNN** với cấu trúc nền tảng là **ResNet-101**. Mô hình này đã được huấn luyện trước (pretrained) trên tập dữ liệu **Visual Genome** để đảm bảo khả năng nhận diện phong phú các đối tượng và thuộc tính chi tiết.
- **Cơ chế Bottom-Up:** Mạng sử dụng **Region Proposal Network (RPN)** để quét hình ảnh và đề xuất các vùng tiềm năng chứa đối tượng (Regions of Interest - RoI). Tại mỗi vùng, mô hình thực hiện hồi quy bounding box và phân loại để xác định các thực thể có độ tin cậy cao nhất.
- **Định dạng đầu ra:** Đối với mỗi hình ảnh trong bộ dữ liệu MS COCO, hệ thống lấy ra cố định **36 vùng đặc trưng** có điểm số nhận diện cao nhất. Mỗi vùng được biểu diễn dưới dạng một vector đặc trưng **2048 chiều** (sau lớp pooling cuối cùng của ResNet).
- **Biểu diễn dữ liệu:** Kết quả thu được là một tensor kích thước **(36, 2048)** cho mỗi ảnh. Tensor này đóng vai trò là đầu vào cho cơ chế **Top-Down Attention**, cho phép mô hình ngôn ngữ tập trung vào các vector 2048 chiều cụ thể ứng với từng thực thể (ví dụ: con ngựa, xe buýt) trong quá trình sinh từ.



Hình 3.6 Mô hình trích xuất đặc trưng ảnh

3.4 Mô hình đề xuất



Hình 3.7 Mô hình đề xuất

Kiến trúc đề xuất kế thừa cơ chế LSTM kép từ Anderson et al. (2017) nhưng thực hiện hai cải tiến quan trọng để phù hợp với ngôn ngữ đích. Thứ nhất, thay vì sử dụng từ đơn, lớp Word Embedding được thiết kế để tiếp nhận các Token IDs là từ ghép Tiếng Việt (đã qua xử lý Underthesea), giúp cơ chế Attention căn chỉnh ngữ nghĩa chính xác hơn giữa vùng ảnh và thực thể ngôn ngữ. Thứ hai, luồng dữ liệu được tối ưu hóa bằng cách đưa trực tiếp Word Vector và Global Average vào Attention LSTM, kết hợp với chiến thuật huấn luyện SCST sử dụng Baseline từ giải mã chùm (Beam Search) thay vì giải mã tham lam (Greedy) như bài báo gốc. Những thay đổi này giúp mô hình đạt được sự cân bằng giữa độ chính xác thị giác và tính tự nhiên của câu mô tả tiếng Việt.

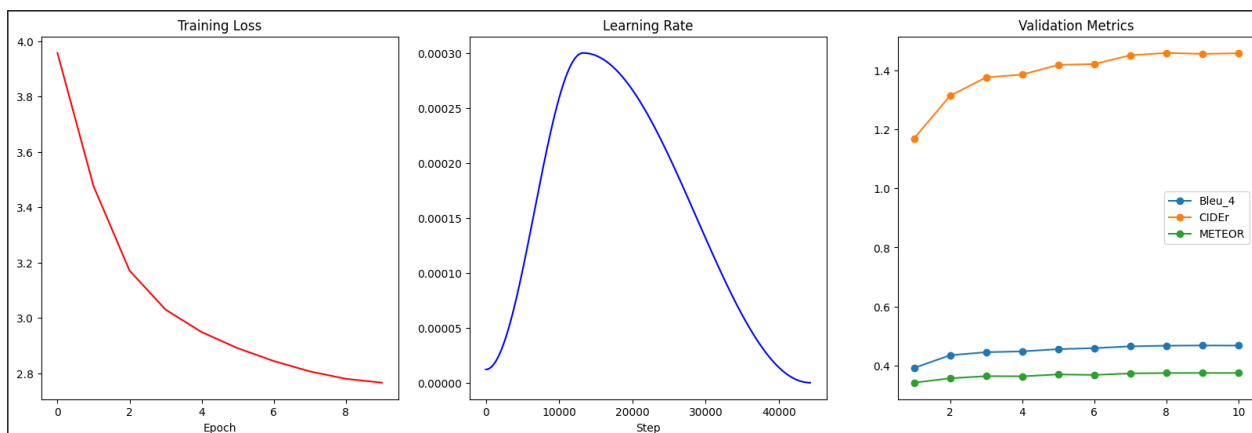
Quy trình huấn luyện hai giai đoạn:

- **Giai đoạn 1 (Cross-Entropy):** Sử dụng hàm mất mát CE để mô hình học cấu trúc câu tiếng Việt cơ bản từ tập dữ liệu MS COCO đã dịch thuật.
- **Giai đoạn 2 (Self-Critical Sequence Training - SCST):** Sử dụng học tăng cường để tối ưu trực tiếp chỉ số CIDEr. Tại đây, mô hình tự sinh câu bằng cách lấy mẫu (Sampling) và so sánh với phần thưởng từ giải mã chùm (Beam Search, k=5) để cập nhật trọng số.

3.5 Kết quả thực nghiệm

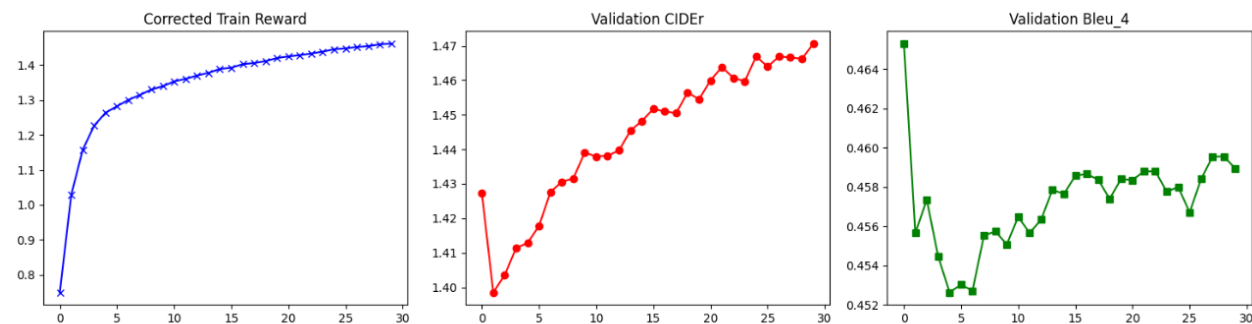
Trong 10 epoch đầu tiên, mô hình được huấn luyện để làm quen với cấu trúc câu và từ vựng tiếng Việt. Nhìn vào đồ thị Training Loss, đường biểu diễn giảm rất nhanh

và ổn định, cho thấy mạng UpDownDecoder thích nghi tốt với dữ liệu ngay từ những bước đầu. Các chỉ số đánh giá trên tập Validation đồng loạt tăng trưởng đều đặn, khẳng định mô hình đã hình thành được khả năng liên kết cơ bản giữa các vùng đối tượng trong ảnh với hệ thống ngôn ngữ đích.



Hình 3.8 Kết quả phase CE

Sau khi có nền tảng ngôn ngữ, mình chuyển sang giai đoạn tinh chỉnh bằng thuật toán SCST để mô hình thoát khỏi cách diễn đạt máy móc và hướng tới những câu mô tả giàu thông tin hơn. Đồ thị Corrected Train Reward cho thấy phần thưởng tăng tiến liên tục, kéo theo sự bứt phá của chỉ số CIDEr so với giai đoạn trước đó. Dù ở những epoch đầu có sự dao động nhẹ do mô hình đang thực hiện cơ chế "thử và sai" để tìm kiếm chính sách sinh từ tối ưu, nhưng việc các chỉ số n-gram vẫn giữ mức ổn định đã chứng minh mô hình không bị tình trạng đọc vẹt mà vẫn duy trì được độ tự nhiên của tiếng Việt.



Hình 3.9 Kết quả phase SCST

Dưới đây là bảng so sánh kết quả thực nghiệm

METRIC	B-1	B-2	B-3	B-4	METEOR	ROUGE_L	CIDER
BUTD BASE	80.2	64.1	49.1	36.9	27.6	57.1	117.9
BUTD	75.0	58.8	45.1	34.57	27.6	56.4	116.4
CE EN							
BUTD	77.7	62.4	48.1	36.6	27.6	57.5	122.3
SCST EN							
BUTD	78.2	66.8	56.4	46.7	37.5	60.8	145.8
CE VI							
BUTD	80.1	67.8	56.4	45.8	37.0	60.7	147.1
SCST VI							

Bảng 3.1 Kết quả thực nghiệm

3.6 Đánh giá và so sánh

Sau khi hoàn tất quá trình huấn luyện, mình thực hiện đánh giá để xem mô hình thực sự "hiểu" ảnh đến mức nào thay vì chỉ nhìn vào các con số toán học. Kết quả cho thấy một sự bứt phá rõ rệt khi áp dụng cho Tiếng Việt so với nguyên mẫu tiếng Anh.

Phân tích sự vượt trội của mô hình

Nhìn vào bảng số liệu, điều khiến mình ấn tượng nhất chính là chỉ số CIDEr của phiên bản Tiếng Việt đạt mốc 147.1, vượt xa mức 122.3 của bản gốc tiếng Anh. Điều này chứng tỏ việc dành thời gian để xử lý từ ghép (Word Segmentation) bằng Underthesea là một bước đi hoàn toàn đúng đắn. Thay vì nhìn nhận các từ rời rạc, cơ chế Attention giờ đây đã "thông minh" hơn khi biết tập trung vào đúng các cụm thực thể có nghĩa như "xe_đạp" hay "người_đi_bộ" để sinh câu trôi chảy.

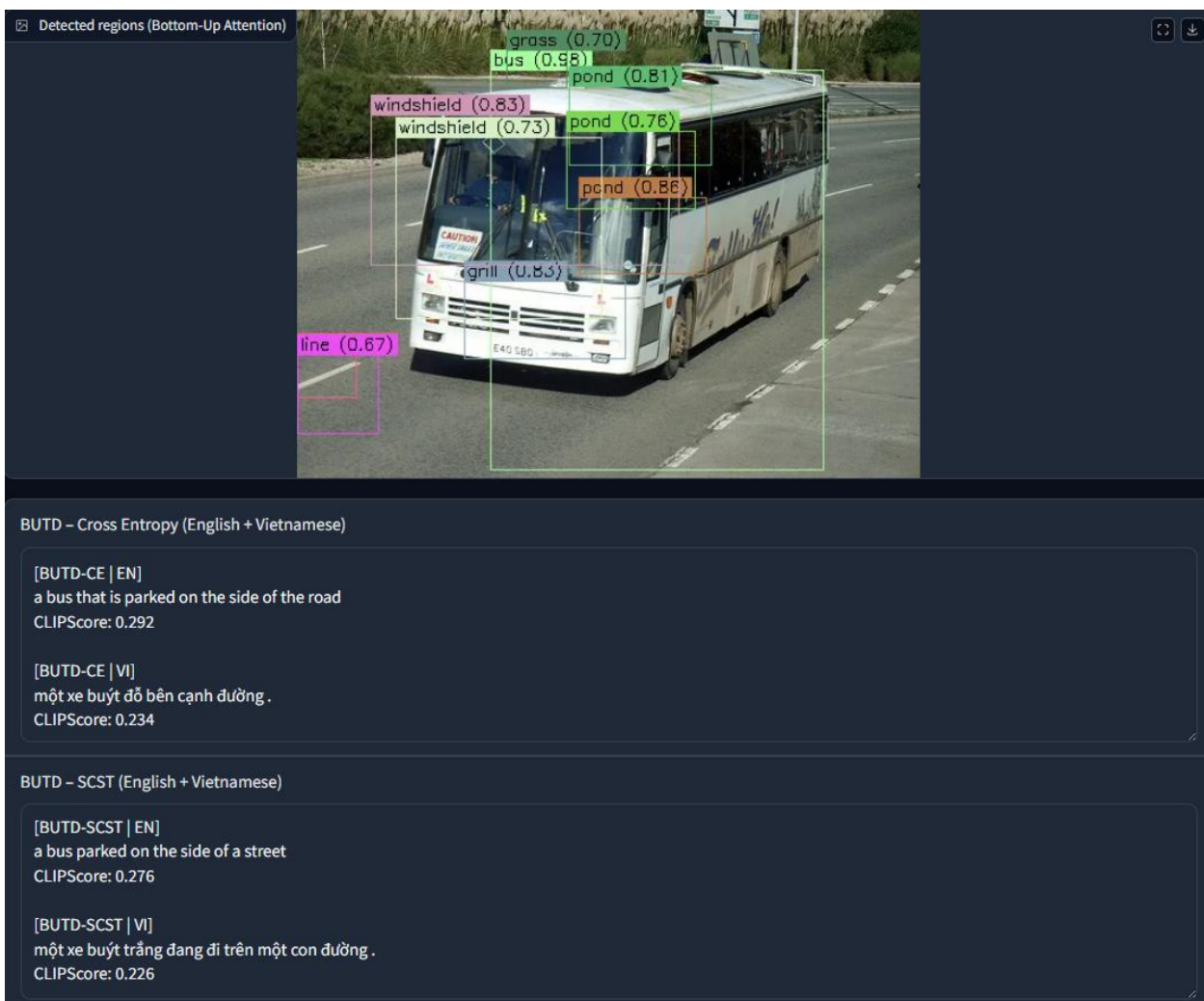
Đặc biệt, việc chuyển đổi từ huấn luyện Cross-Entropy sang SCST đã giúp mô hình thoát khỏi tư duy "học vẹt". Mặc dù chỉ số Bleu-4 có sụt giảm nhẹ do mô hình không còn cố gắng khớp từng từ một cách máy móc, nhưng bù lại, các câu mô tả

sinh ra lại mang tính đặc tả cao, giàu thông tin và sát với cách hành văn tự nhiên của người Việt hơn rất nhiều.

Nhìn nhận thực tế qua các ví dụ

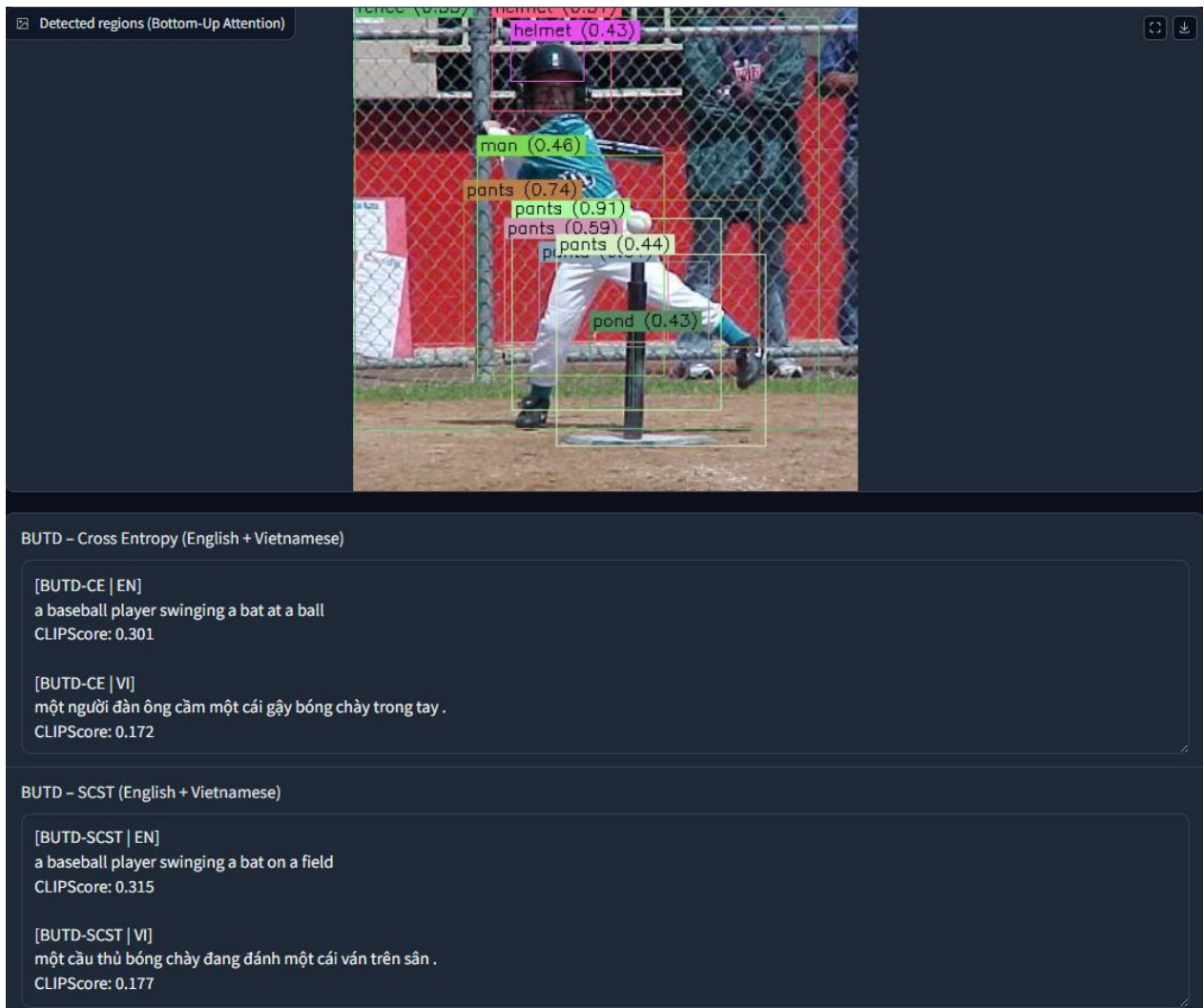
Để khách quan hơn, mình đã thử nghiệm mô hình trên nhiều kịch bản khác nhau và rút ra được ba nhóm kết quả điển hình:

- **Những bức ảnh "vừa sức" (Thành công):** Với các ảnh có chủ thể rõ ràng, mô hình làm rất tốt việc gọi tên đối tượng và hành động. Sự kết hợp giữa việc "nhìn" (Faster R-CNN) và "hiểu" (LSTM) hoạt động cực kỳ ăn ý, tạo ra những câu mô tả chuẩn xác về cả nội dung lẫn ngữ pháp. Mô tả đúng xe buýt ở trên đường



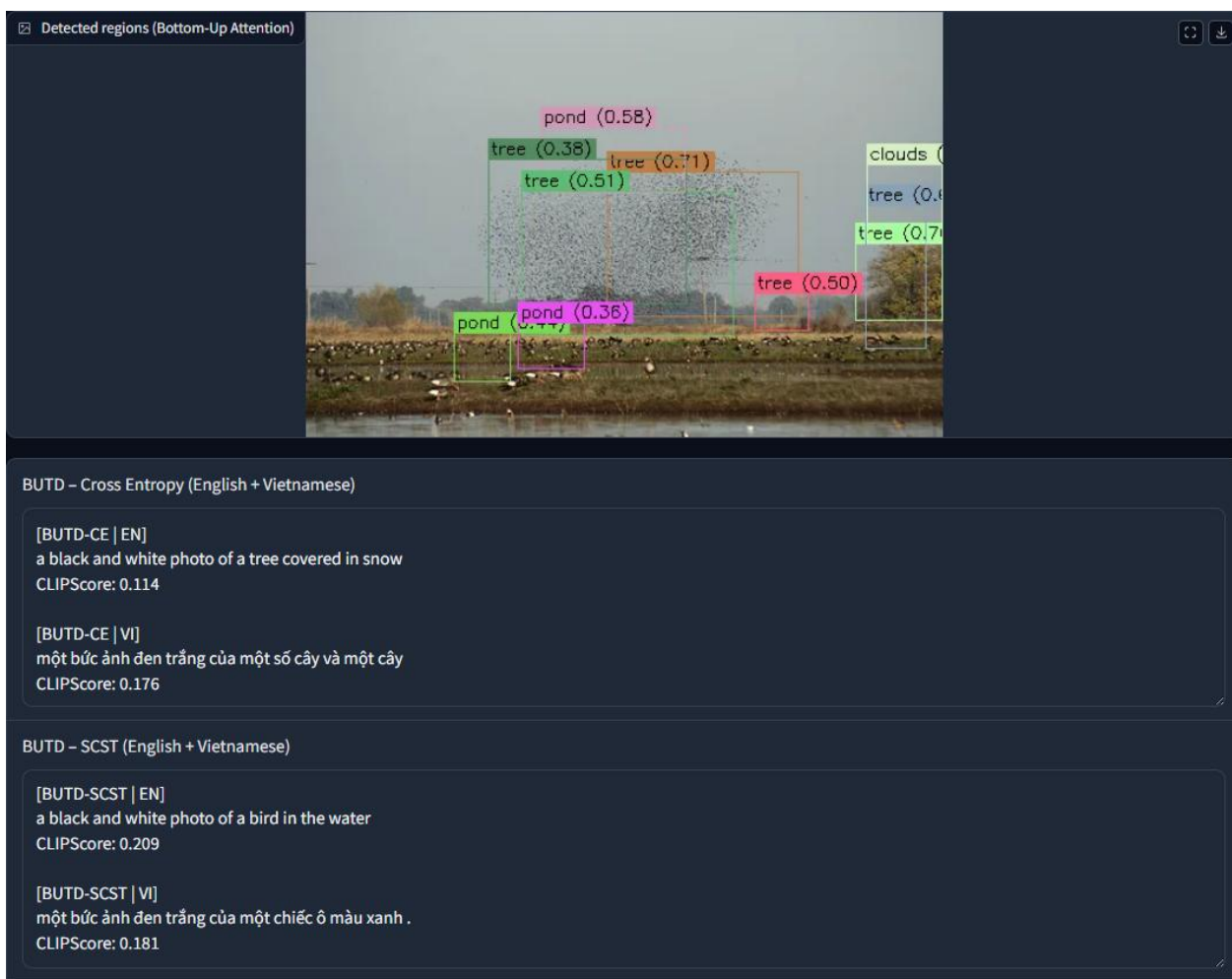
Hình 3.10 Kết quả tốt nhất

- **Những chi tiết khó nhận (Sai sót nhỏ):** Mô hình vẫn còn đôi chút lúng túng khi gặp ảnh có quá nhiều vật thể nhỏ nằm chồng chéo lên nhau. Đôi khi mô hình đếm sai số lượng hoặc nhầm lẫn giữa các vật thể có hình dáng tương tự do giới hạn về độ phân giải của vùng đặc trưng ảnh. Cậu bé mà miêu tả người đàn ông



Hình 3.11 Lỗi miêu tả

- **Hiện tượng "tự sáng tác" (Nhầm ngữ cảnh):** Trong vài trường hợp ảnh bị mờ hoặc thiếu sáng, mô hình có xu hướng đoán mò dựa trên những gì nó đã học được quá kỹ từ tập dữ liệu. Chẳng hạn, cứ thấy bàn ghế là mô hình nghĩ ngay đến "bữa ăn" dù thực tế có thể là một văn phòng làm việc, cho thấy sự lệ thuộc vào xác suất ngôn ngữ khi thông tin hình ảnh không rõ ràng.



Hình 3.12 Kết quả sai ngữ cảnh

CHƯƠNG 4. KẾT LUẬN VÀ KIẾN NGHỊ

4.1 Kết luận

Đồ án đã tập trung giải quyết bài toán sinh mô tả ảnh tự động (Image Captioning) – một lĩnh vực đòi hỏi sự kết hợp khắt khe giữa khả năng nhận diện của Thị giác máy tính và khả năng diễn đạt của Xử lý ngôn ngữ tự nhiên (NLP). Thông qua quá trình nghiên cứu và triển khai thực nghiệm kiến trúc Bottom-Up and Top-Down Attention (BUTD) trên tập dữ liệu MS COCO Tiếng Việt, đồ án đã đạt được những kết luận quan trọng sau:

- Về kiến trúc hệ thống: Đồ án đã hiện thực hóa thành công mô hình lai ghép, sử dụng mạng Faster R-CNN để trích xuất các đặc trưng vùng (region-based features) thay vì đặc trưng toàn cục truyền thống. Điều này cho phép mô hình tập trung vào các thực thể cụ thể trong ảnh, tạo tiền đề cho quá trình căn chỉnh ngữ nghĩa chính xác.
- Về xử lý ngôn ngữ đặc thù: Nghiên cứu đã giải quyết hiệu quả tính chất đa tiết của Tiếng Việt bằng cách tích hợp thư viện Underthesea để xử lý từ ghép (Word Segmentation) ngay từ pha tiền xử lý. Việc huấn luyện trên các token từ ghép gắn kết bằng dấu gạch dưới () đã giúp cơ chế Attention hoạt động hiệu quả hơn, tránh hiện tượng rời rạc ngữ nghĩa thường gặp ở các mô hình dịch thuật thô.
- Về chiến lược huấn luyện SCST: Đồ án đã chứng minh được giá trị của việc sử dụng thuật toán Self-Critical Sequence Training (SCST) để tối ưu trực tiếp chỉ số CIDEr. Kết quả thực nghiệm cho thấy sự bứt phá từ mức 145.8 (giai đoạn Cross-Entropy) lên 147.1 (giai đoạn SCST), khẳng định khả năng sinh câu mô tả giàu thông tin và sát với nhận thức của con người.
- Về hiệu năng tổng thể: Với các chỉ số định lượng ấn tượng (CIDEr: 147.1, Bleu-1: 80.1, Bleu-4: 45.8), mô hình không chỉ đạt độ chính xác cao về mặt từ vựng mà còn đảm bảo được cấu trúc ngữ pháp tự nhiên, trôi chảy, đáp ứng tốt yêu cầu mô tả ảnh trong ngữ cảnh ngôn ngữ Tiếng Việt.

4.2 Kiến nghị và Hướng phát triển

Dựa trên những thành tựu đã đạt được cùng với việc phân tích sâu các hạn chế còn tồn tại về mặt phần cứng và dữ liệu, nhóm nghiên cứu đề xuất các hướng phát triển chiến lược sau nhằm nâng cao hơn nữa hiệu năng và tính thực tiễn của hệ thống:

- **Chuyển dịch sang kiến trúc Transformer-based và Vision Transformer (ViT):** Mặc dù kiến trúc BUTD dựa trên LSTM đã hoạt động rất hiệu quả, xu hướng hiện đại đang chuyển dịch mạnh mẽ sang các mô hình dựa hoàn toàn trên cơ chế Self-Attention. Hướng phát triển tiếp theo sẽ là thay thế bộ trích xuất Faster R-CNN bằng **Swin Transformer** hoặc **ViT** để tận dụng khả năng nắm bắt mối quan hệ không gian toàn cục (global context) tốt hơn. Đồng thời, việc sử dụng các mô hình ngôn ngữ lớn (LLM) như **BERT** hoặc **GPT** làm bộ giải mã sẽ giúp câu văn sinh ra có sự đa dạng về phong cách và chiều sâu về ngữ cảnh.
- **Tối ưu hóa đa phương thức với học tăng cường chuyên sâu:** Tiếp tục khai thác triệt để kỹ thuật **SCST** bằng cách kết hợp thêm các hàm phần thưởng (Reward) đa dạng hơn như **SPICE** hay các hàm đánh giá dựa trên ngữ nghĩa học máy (Semantic similarity). Việc này sẽ giúp mô hình không chỉ tối ưu hóa các cụm từ khớp nhau về mặt ký tự mà còn đảm bảo sự tương đồng về tư duy giữa máy tính và con người khi mô tả cùng một sự việc.
- **Mở rộng quy mô dữ liệu và thích nghi miền (Domain Adaptation):** Để tăng cường khả năng tổng quát hóa, nghiên cứu cần được mở rộng sang các bộ dữ liệu quy mô lớn hơn như **Conceptual Captions** hoặc các bộ dữ liệu mang đậm bản sắc văn hóa Việt Nam (như ẩm thực, danh lam thắng cảnh, trang phục truyền thống). Việc huấn luyện trên các miền dữ liệu đặc thù sẽ giúp mô hình trở nên "thông minh" hơn trong việc nhận diện các khái niệm trừu tượng hoặc các đối tượng hiếm gặp trong thực tế đời sống.
- **Nén mô hình và triển khai trên thiết bị biên (Edge Devices):** Để đưa kết quả nghiên cứu từ phòng thí nghiệm ra thực tế, cần áp dụng các kỹ thuật nén mô hình như **Quantization** hoặc **Knowledge Distillation**. Mục tiêu là cho phép hệ thống vận hành mượt mà trên điện thoại thông minh hoặc các thiết bị đeo thông minh.

Điều này sẽ trực tiếp hỗ trợ xây dựng các ứng dụng **Visual Assistant**, giúp người khiếm thị có thể "nghe" thấy thế giới xung quanh thông qua camera theo thời gian thực – đây chính là giá trị nhân văn cao nhất mà đề tài hướng tới.

TÀI LIỆU THAM KHẢO

S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Vaizman, G. (2017). Self-critical sequence training for image captioning. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7008-7024.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6077-6086.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *In European Conference on Computer Vision (ECCV)*, pp. 740-755.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NIPS)*.

Costa-jussà, M. R., Cross, J., Onidi, G., et al. (2022). No Language Left Behind: Scaling Human-Centric Machine Translation. *arXiv preprint arXiv:2207.04672*.

[6] **Nguyen, H., et al. (2018).** Underthesea: Vietnamese Natural Language Processing Toolkit. *GitHub Repository*.