ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN

LÊ XUÂN BÌNH
TRẦN KHÔI NGUYÊN

KHÓA LUẬN TỐT NGHIỆP

XÂY DỰNG BỘ DỮ LIỆU ĐÁNH GIÁ VÀ THỬ NGHIỆM MÔ HÌNH ĐA THỂ THỨC CHO BÀI TOÁN TẠO SINH CHÚ THÍCH TÓM TẮT TỪ INFOGRAPHIC TIẾNG VIỆT

ENHANCING MULTIMODAL BENCHMARK DATASET AND MODEL FOR VIETNAMESE INFOGRAPHIC ABSTRACTIVE CAPTIONING

CỬ NHÂN NGÀNH KHOA HỌC DỮ LIỆU

TP. HỒ CHÍ MINH, 2025

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN

LÊ XUÂN BÌNH – 22520131

TRẦN KHÔI NGUYÊN – 22520987

KHÓA LUẬN TỐT NGHIỆP

XÂY DỰNG BỘ DỮ LIỆU ĐÁNH GIÁ VÀ THỬ NGHIỆM MÔ HÌNH ĐA THỂ THỨC CHO BÀI TOÁN TẠO SINH CHÚ THÍCH TÓM TẮT TỪ INFOGRAPHIC TIẾNG VIỆT

ENHANCING MULTIMODAL BENCHMARK DATASET AND MODEL FOR VIETNAMESE INFOGRAPHIC ABSTRACTIVE CAPTIONING

CỬ NHÂN NGÀNH KHOA HỌC DỮ LIỆU

GIẢNG VIÊN HƯỚNG DẪN

ThS. TRẦN VĨNH KHIÊM

TP. HỒ CHÍ MINH, 2025

THÔNG TIN HỘI ĐỒNG CHẤM KHÓA LUẬN TỐT NGHIỆP

Hội đồng chấm khóa luận tốt nghiệp, thành lập theo Quyết định số …… ngày …….. của Hiệu trưởng Trường Đại học Công nghệ Thông tin.

LỜI CẢM ƠN

Chúng em xin bày tỏ lòng biết ơn sâu sắc đến Khoa Khoa học và Kỹ thuật Thông tin, Trường Đại học Công nghệ Thông tin – ĐHQG TP.HCM, vì đã tạo dựng một môi trường học tập và nghiên cứu chuyên nghiệp, giàu cảm hứng; đồng thời luôn đồng hành và hỗ trợ kịp thời trong suốt quá trình học tập, rèn luyện, giúp chúng em nuôi dưỡng niềm đam mê học thuật và tích lũy những hành trang quý giá để từng bước hoàn thiện bản thân.

Chúng em đặc biệt tri ân các thầy cô, anh chị trợ giảng Bộ môn Khoa học Dữ liệu vì sự nhiệt huyết và tận tâm trong giảng dạy, truyền đạt những tri thức nền tảng, góp phần xây dựng cơ sở vững chắc cho quá trình thực hiện và hoàn thiện khóa luận.

Trong suốt thời gian triển khai đề tài, chúng em xin gửi lời cảm ơn chân thành đến thầy Trần Vĩnh Khiêm và thầy Nguyễn Hiếu Nghĩa vì sự hướng dẫn tận tình, những góp ý thẳng thắn và quý báu, cũng như sự kiên nhẫn đồng hành. Sự hỗ trợ của thầy không chỉ giúp chúng em hoàn thiện nghiên cứu một cách chỉnh chu, mà còn truyền dạy tinh thần trách nhiệm, sự tỉ mỉ trong công việc và tình yêu đối với nghiên cứu khoa học.

Chúng em cũng xin cảm ơn các anh chị và bạn bè đã luôn sẵn sàng hỗ trợ, chia sẻ kinh nghiệm và động viên chúng em trong suốt quá trình thực hiện đề tài.

Cuối cùng, chúng em xin gửi lời cảm ơn sâu sắc đến gia đình, cha mẹ và những người thân yêu vì đã luôn là chỗ dựa tinh thần vững chắc, tiếp thêm động lực để chúng em vượt qua những khó khăn và thử thách.

Trân trọng cảm ơn,

Lê Xuân Bình

Trần Khôi Nguyên

MỤC LỤC

DANH MỤC HÌNH

DANH MỤC BẢNG

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	Tiếng Anh	Tiếng Việt
VQA	Visual Question Answering	Hỏi-đáp dựa trên hình ảnh
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
CV	Computer Vision	Thị giác máy tính
AI	Artificial Intelligence	Trí tuệ nhân tạo
POS	Parts-Of-Speech	Từ loại
NER	Named Entity Recognition	Nhận dạng thực thể
ECDF	Empirical Cumulative Distribution Function	Hàm phân phối tích lũy thực nghiệm
BLEU	Bilingual Evaluation Understudy	Đánh giá chất lượng ngôn ngữ dịch
METEOR	Metric for Evaluation of Translation with Explicit ORdering	Đánh giá bản dịch có chú trọng thứ tự
ROUGE	Recall-Oriented Understudy for Gisting Evaluation	Đánh giá độ chính xác của tóm tắt
CIDEr	Consensus-based Image Description Evaluation	Đánh giá độ đồng thuận của chú thích hình ảnh
OCR	Optical character recognition	Nhận dạng ký tự quang học
SOTA	State-Of-The-Art	Tiên tiến nhất
LLM	Large Language Model	Mô hình ngôn ngữ lớn
LMM	Large Multimodal Model	Mô hình đa thể thức lớn
OOV	Out-of-Vocabulary	Từ vựng ngoài danh mục
MLP	Multilayer Perceptron	Mạng nơ-ron đa lớp
M-RoPE	Multimodal Rotary Position Embedding	Phương pháp nhúng vị trí quay đa thể thức

TÓM TẮT KHÓA LUẬN

Trong kỷ nguyên số, sự dịch chuyển từ văn hóa "đọc sâu" sang văn hóa "quét" thông tin và sự bùng nổ của dữ liệu dẫn đến nhu cầu truyền tải thông tin trực quan đã thúc Infographic trở thành một công cụ truyền thông đa phương tiện mạnh mẽ. Bài toán tạo sinh tóm tắt cho Infographic (Abstractive Infographic Captioning) theo đó trở thành một lĩnh vực giao thoa đầy tiềm năng giữa thị giác máy tính (CV) và xử lý ngôn ngữ tự nhiên (NLP). Nhiệm vụ này đòi hỏi các hệ thống không chỉ dừng lại ở việc nhận diện vật thể hay trích xuất chữ viết đơn thuần, mà còn phải thực hiện các phép lập luận phức tạp dựa trên bối cảnh không gian và nội dung ngữ nghĩa để rút ra thông điệp cốt lõi.

Tại Việt Nam, mặc dù đã có những bước tiến đáng kể với các bộ dữ liệu đa thể thức như ViVQA, OpenViVQA, ViTextVQA hay ViInfographicVQA, nhưng hầu hết vẫn tập trung vào bài toán hỏi-đáp từ hình ảnh (Visual Question Answering), chưa thật sự chú trọng đến khả năng tổng hợp và tóm tắt thông điệp trừu tượng. Để giải quyết khoảng trống này, nhóm đã xây dựng bộ dữ liệu ViInfographicCaps với 17.840 mẫu infographic tiếng Việt kèm bản tóm tắt tương ứng, được thu thập từ nguồn tin chính thống của Thông tấn xã Việt Nam.

Trên cơ sở đó, khóa luận đề xuất một kiến trúc mô hình mới tối ưu cho bài toán, tập trung vào việc khai thác đặc trưng văn bản OCR, và đặc trưng bố cục. Điểm cốt lõi của kiến trúc này là việc tích hợp cơ chế nhận diện tương quan không gian giữa các vùng bố cục, kết hợp với phương pháp định hướng từ khóa sử dụng kỹ thuật học tương phản. Sự phối hợp này giúp mô hình ưu tiên các thực thể trọng tâm và nắm bắt chính xác các thông điệp then chốt, đưa ra được tóm tắt hình ảnh bao quát và đáng tin cậy. Kết quả thực nghiệm cho thấy kiến trúc đề xuất đạt hiệu suất vượt trội so với các mô hình cơ sở và đa thể thức lớn, khẳng định hiệu quả của việc kết hợp logic "ngữ pháp thị giác" với sức mạnh của các mô hình ngôn ngữ tiền huấn luyện tiếng Việt trong việc xử lý dữ liệu đa thể thức đặc thù này.

Chương 1. TỔNG QUAN

1.1. Đặt vấn đề

Trong kỷ nguyên số, chúng ta đang chứng kiến một sự chuyển dịch mô hình căn bản trong cách con người tiêu thụ thông tin: từ văn hóa "đọc sâu" dựa trên văn bản tuyến tính truyền thống, xã hội đang chuyển sang văn hóa "quét" dựa trên hình ảnh và các đoạn văn bản ngắn [1]. Infographic ra đời là một công cụ truyền thông mạnh mẽ kết hợp giữa dữ liệu định lượng, văn bản cô đọng và các yếu tố đồ họa phức tạp [2]. Trong bối cảnh đó, bài toán tạo sinh tóm tắt cho infographic (Abstractive Infographic Captioning) đã trở thành như một thách thức trong lĩnh vực thị giác máy tính (CV) kết hợp xử lý ngôn ngữ tự nhiên (NLP), yêu cầu các mô hình không chỉ dừng lại ở việc nhận diện vật thể mà còn phải thực hiện các phép lập luận dựa trên bố cục không gian và nội dung ngữ nghĩa. Khác với bài toán mô tả hình ảnh tự nhiên truyền thống (Image Captioning), vốn thường tập trung vào việc liệt kê các thực thể nổi bật trong một bối cảnh, tóm tắt infographic đòi hỏi sự hiểu biết sâu sắc về cấu trúc phân cấp, mối quan hệ giữa các thành phần dữ liệu và khả năng tổng hợp các thông tin rời rạc thành một văn bản mạch lạc [3].

1.2. Mô tả dữ liệu infographic

Infographic là sự kết hợp giữa "*information*" (thông tin) và "*graphic*" (đồ họa), một hình thức trình bày trực quan giúp truyền tải dữ liệu, sự kiện nhanh chóng và rõ ràng. Bằng cách tối ưu hóa thị giác, infographic biến các quy luật và xu hướng phức tạp trở nên dễ nhận diện. Với mục tiêu phục vụ truyền thông đại chúng, chúng được thiết kế tối giản để bất kỳ ai cũng có thể thấu hiểu mà không cần nền tảng chuyên môn sâu [4]. Nhà thiết kế, tác giả người Anh Nigel Holmes còn gọi đây là "đồ họa giải thích" (explanation graphics), và cho rằng những infographic thực sự hiệu quả sẽ nâng cao khả năng nắm bắt thông tin phức tạp của người bình thường [5].

1.3. Mô tả bài toán Tạo sinh chú thích tóm tắt hình ảnh

Trong khóa luận này, chúng tôi xác định Tạo sinh chú thích tóm tắt hình ảnh (Abstractive Captioning) là một hướng nghiên cứu mở rộng từ bài toán chú thích hình ảnh (Image Captioning), đồng thời chịu ảnh hưởng rõ rệt từ mục tiêu của bài toán tóm tắt trừu tượng (Abstractive Summarization) trong NLP. Nếu như Image Captioning truyền thống chủ yếu tập trung mô tả hình ảnh bằng cách nêu các đối tượng và thuộc tính nổi bật (có gì trong ảnh ?), thì với các hình ảnh giàu thông tin và cấu trúc phức tạp như infographic (nhiều thực thể, nhiều chi tiết và quan hệ ngữ nghĩa), cách mô tả theo dạng liệt kê thường chỉ phản ánh bề mặt, chưa đáp ứng yêu cầu nắm bắt nội dung cốt lõi và ý nghĩa mà hình ảnh truyền tải.

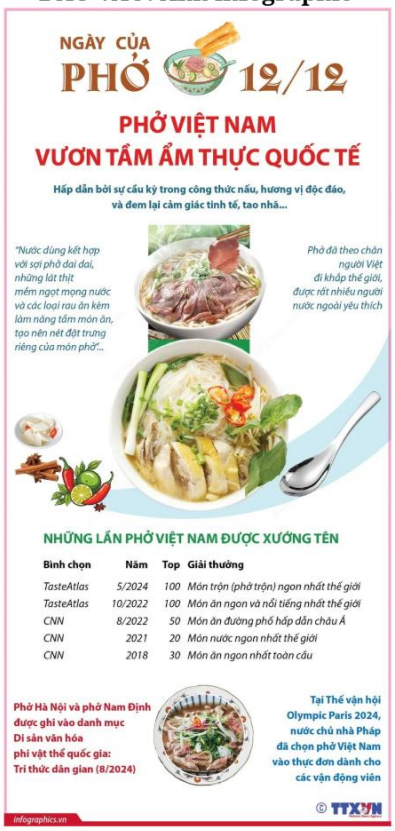
Vì vậy, tạo sinh chú thích tóm tắt hình ảnh đặt mục tiêu tạo ra một chú thích mang tính khái quát cho hình ảnh, tương tự như một bản tóm tắt: đầu ra cần ngắn gọn để dễ tiếp nhận, nhưng đồng thời phải giàu thông tin và mạch lạc để bao quát các điểm quan trọng nhất. Thay vì chỉ nhắc lại các thực thể xuất hiện, mô hình cần tổng hợp và diễn đạt lại nội dung theo cách tự nhiên, nhấn mạnh vào các yếu tố then chốt như: thông tin trung tâm của hình ảnh; ngữ cảnh hoặc thông điệp chính; và các mối quan hệ quan trọng giữa các thành phần trong ảnh.

Nói cách khác, Abstractive Captioning hướng đến việc sinh ra một mô tả cô đọng nhưng đủ bao quát nhằm phản ánh mức độ hiểu nội dung ở tầng ý nghĩa thay vì chỉ dừng lại ở mức nhận diện đối tượng.

Chi tiết về đầu vào và đầu ra của bài toán:

- Đầu vào: Một ảnh infographic chứa thông tin đa dạng, bao gồm cả yếu tố hình ảnh (biểu tượng, biểu đồ, màu sắc) và văn bản (chữ viết, số liệu) được trình bày dưới dạng cấu trúc đồ họa.
- Đầu ra: Một đoạn văn bản ngắn gọn, súc tích, có tính tóm lược và diễn đạt lại nội dung chính của infographic theo hướng trừu tượng - tức là không sao chép nguyên văn mà tái diễn giải thông tin một cách tự nhiên và có ngữ nghĩa đầy đủ.

ĐẦU VÀO: Ảnh infographic



ĐẦU RA: Chú thích tóm tắt

Ngày 12/12 hằng năm được chọn là "Ngày của Phở". Sự kiện được khởi xướng từ năm 2017, trở thành một hoạt động quảng bá văn hóa ẩm thực quan trọng, góp phần nâng tầm ẩm thực Việt Nam và lan tỏa món Phở truyền thống của Việt Nam ra khắp thế giới. Trong những năm gần đây, Phở Việt Nam nhiều lần được bình chọn là một trong những món ăn ngon và nổi tiếng trên thế giới: Top 30 món ăn ngon nhất toàn cầu năm 2018, Top 20 món nước ngon nhất thế giới năm 2021, Top 100 món ăn ngon và nổi tiếng nhất thế giới 2022...

Hình 1.1: Ví dụ minh họa đầu vào – đầu ra của bài toán

1.4. Lý do thực hiện đề tài

Trong những năm gần đây, cộng đồng NLP và CV tại Việt Nam đã đạt được những bước tiến đáng kể với các bộ dữ liệu đa thể thức [6], [7], [8], [9]. Tuy nhiên, đặc trưng chung của các tập dữ liệu này là thường tập trung vào bài toán hỏi-đáp thông qua hình ảnh Visual Question Answering (VQA) với đầu ra thường chỉ là các từ hoặc cụm từ ngắn. Mặc dù các mô hình đã có khả năng trích xuất thông tin cụ thể, nhưng khả năng tổng hợp, diễn giải và tạo ra văn bản có tính mạch lạc từ hình ảnh vẫn chưa được nghiên cứu sâu.

Các bộ dữ liệu chú thích hình ảnh Image Captioning hiện có cho tiếng Việt, tiêu biểu là UIT-ViIC [9] chủ yếu tập trung vào nhóm ảnh thực tế từ nguồn MS COCO [10]. Những mô hình huấn luyện trên dữ liệu này thường chỉ dừng lại ở mức độ mô tả bề mặt các thực thể và hành động diễn ra trong ảnh. Đối với các loại hình ảnh phức tạp, mang tính thông tin cao như infographic, các mô tả ngắn kiểu truyền

thống không thể truyền tải được thông điệp cốt lõi hay giá trị tóm lược mà người dùng cần.

Infographic là một dạng dữ liệu đặc thù, nơi thông tin được nén vào các yếu tố thiết kế và bố cục. Hiện nay, dù đã có bộ ViInfographicVQA [11] với dữ liệu là các cặp hỏi–đáp, nhưng nhiệm vụ chính vẫn là truy vấn các điểm dữ liệu đơn lẻ. Sự thiếu vắng một bộ dữ liệu quy mô lớn, chuẩn hóa cho bài toán chú thích tóm tắt với yêu cầu mô hình không chỉ "đọc" được chữ hay "nhìn" được ảnh, mà phải "hiểu" và "tóm tắt" được toàn bộ nội dung thành một đoạn văn ngắn gọn, súc tích này chính là động lực thôi thúc chúng tôi thực hiện đề tài nhằm lấp đầy khoảng trống kỹ thuật và dữ liệu cho các nghiên cứu về dữ liệu đa thể thức trên tiếng Việt.

1.5. Đối tượng và phạm vi nghiên cứu

Phạm vi nghiên cứu của đề tài tập trung vào bài toán sinh tóm tắt ngôn ngữ tự nhiên từ ảnh infographic tiếng Việt, với trọng tâm là các infographic được thu thập từ website *Tin Đồ họa – Thông tấn xã Việt Nam*[1]. Đây là một nguồn dữ liệu chính thống, được cập nhật thường xuyên, có tính xác thực cao và đặc biệt phong phú về nội dung. Các infographic được khai thác có đặc điểm:

- Dạng ảnh cấu trúc rõ ràng, kết hợp cả yếu tố đồ họa và văn bản (OCR-friendly).
- Phủ rộng nhiều lĩnh vực thuộc nhiều thể loại, bao gồm: Chính trị – Ngoại giao, Kinh tế – Hội nhập, Văn hóa – Xã hội, Giáo dục – Khoa học, Y tế – Cộng đồng, Thể thao – Nghệ thuật, Thiên tai – Tai nạn và Nhân vật – Sự kiện.

Việc lựa chọn tập trung vào nguồn dữ liệu này giúp đảm bảo tính đa dạng ngữ nghĩa, đồng thời thể hiện rõ các đặc trưng thị giác – ngôn ngữ đặc thù trong tiếng Việt.

[1] https://infographics.vn/

1.6. Phương pháp tiếp cận

Để giải quyết bài toán Tạo sinh chú thích tóm tắt hình ảnh cho infographic tiếng Việt, đề tài thực hiện nghiên cứu và triển khai hướng tiếp cận chính sau:

1.6.1. Tiếp cận từ bài toán Chú thích hình ảnh dựa trên đặc trưng OCR

Chú thích hình ảnh dựa trên đặc trưng OCR (OCR-based Image Captioning) là bài toán với mục đích tạo sinh đoạn mô tả bằng ngôn ngữ tự nhiên, mô tả bức ảnh dựa vào các thành phần thị giác (visual objects) và văn bản (scene texts) có trong bức ảnh đó. Hướng tiếp cận này tập trung vào việc kế thừa và tùy biến các mô hình kiến trúc State-of-the-art (SOTA) đã thành công trên các bộ dữ liệu tiếng Anh để kiểm nghiệm khả năng thích ứng với tiếng Việt.

Mô hình tiếp nhận đầu vào là các đặc trưng đa thể thức phức tạp được trích xuất từ infographic, bao gồm: các thực thể, nội dung văn bản, bản đồ độ sâu, và các lớp thông tin,...

Mục tiêu của chúng tôi là kiểm chứng khả năng của các kiến trúc Encoder-Decoder trong việc tổng hợp các đặc trưng thị giác và văn bản rời rạc để sinh ra một chú thích mang tính tóm lược mạch lạc.

1.6.2. Tiếp cận từ bài toán Tóm tắt văn bản trừu tượng

Trong hướng tiếp cận này, đề tài đặt giả thuyết rằng các dữ liệu văn bản xuất hiện trong infographic đã chứa đựng phần lớn thông điệp cốt lõi cần thiết cho bản tóm tắt. Do đó, chúng tôi thực hiện một nghiên cứu đối chứng bằng việc bỏ qua các yếu tố về hình ảnh và cấu trúc đồ họa, chỉ tập trung vào luồng văn bản được trích xuất qua công cụ nhận dạng ký tự quang học (OCR). Sau đó, thử nghiệm việc sử dụng các mô hình cơ sở trong bài toán tóm tắt văn bản trừu tượng (Abstractive Text Summarization) cho tiếng Việt để xử lý dữ liệu đầu vào này.

Mục tiêu là đánh giá hiệu quả của việc tóm tắt thuần túy dựa trên ngôn ngữ, làm cơ sở để so sánh và xác định giá trị gia tăng của các đặc trưng thị giác trong bài toán tổng thể.

1.6.3. Kiểm chứng khả năng của các mô hình đa thể thức lớn

Chúng tôi thử nghiệm đánh giá trên các mô hình đa thể thức lớn (Large Multimodal Models – LMMs) mã nguồn mở đa ngôn ngữ (multilingual) và đơn ngữ tiếng Việt (monolingual). Các mô hình này đã được huấn luyện trên tập dữ liệu lớn, với khả năng mô phỏng logic giữa hình ảnh và văn bản. Tuy nhiên, các mô hình này chưa được kiểm chứng đầy đủ trên dữ liệu tiếng Việt, đặc biệt là loại ảnh có cấu trúc đặc biệt như infographic.

Chúng tôi thực hiện kiểm nghiệm zero-shot nhằm đo lường khả năng hiểu tự nhiên của mô hình đối với loại hình dữ liệu có cấu trúc đặc thù như infographic mà không cần qua huấn luyện bổ sung. Từ đó, xác định ngưỡng năng lực hiện tại của LMMs trong việc mô phỏng logic liên kết giữa hình ảnh và văn bản tiếng Việt.

1.7. Vấn đề thách thức

1.7.1. Đặc thù ngôn ngữ tiếng Việt và lỗi trích xuất đặc trưng OCR

Tiếng Việt là ngôn ngữ đơn lập với hệ thống thanh điệu phức tạp, gây khó khăn cho các bộ OCR trong việc nhận diện các ký tự có dấu nhỏ. Khi các lỗi OCR này xảy ra trên các từ khóa mang tính định lượng, bản tóm tắt sẽ bị ảnh hưởng theo lỗi lan truyền. Hiện vẫn chưa có nhiều nghiên cứu về việc kết hợp các mô hình ngôn ngữ tiếng Việt chuyên sâu vào các kiến trúc tóm tắt từ hình ảnh để sửa lỗi hoặc bù đắp thông tin bị mất từ OCR.

1.7.2. Vấn đề ảo giác dữ liệu trong tóm tắt định lượng

Hiện tượng ảo giác (hallucination) thường xuất hiện ở các mô hình tạo sinh hiện khi mô hình sinh ra các nội dung đọc rất trôi chảy nhưng lại không có trong tài liệu gốc. Hiện tại cũng chưa có nghiên cứu kiểm định tính xác thực hoặc các hàm mất mát để kiểm soát tính chính xác của dữ liệu tạo sinh trong bản tóm tắt.

1.8. Mục tiêu khóa luận

Đề tài của chúng tôi hướng tới thực hiện các mục tiêu chính sau:

- Xây dựng bộ dữ liệu ViInfographicCaps: thu thập và chuẩn hóa một tập hợp các infographic tiếng Việt kèm phần tóm tắt văn bản tương ứng. Bộ dữ liệu được phân tích và đánh giá để đảm bảo chất lượng, phục vụ hiệu quả cho nghiên cứu sinh tóm tắt từ ảnh.
- Thiết kế một đánh giá toàn diện cho bài toán, bao gồm:
  - Định nghĩa rõ ràng về bài toán sinh tóm tắt từ infographic.
  - Xây dựng bộ tiêu chí đánh giá thống nhất là các thang đo uy tín.
  - Tổng hợp và huấn luyện các mô hình cơ sở để so sánh hiệu quả, sau đó đề xuất mô hình mới tối ưu cho bài toán trên dạng dữ liệu này.

1.9. Cấu trúc nội dung khóa luận

Khóa luận này gồm 7 chương với nội dung chính như sau:

Chương 1: TỔNG QUAN trình bày về bài toán của đề tài, lý do thực hiện, đối tượng và phạm vi nghiên cứu và phương pháp tiếp cận.

Chương 2: CÁC CÔNG TRÌNH LIÊN QUAN nói về các nghiên cứu đã công bố trong nước và quốc tế liên quan đóng góp cho bài toán và chỉ ra những vấn đề mà đề tài cần giải quyết.

Chương 3: XÂY DỰNG BỘ DỮ LIỆU ViInfographicCaps mô tả quá trình xây dựng bộ dữ liệu ViInfographicCaps và các phân tích trên dữ liệu này.

Chương 4: KIẾN TRÚC ĐỀ XUẤT đưa ra mô hình mới đề xuất của chúng tôi để giải quyết bài toán, cải tiến thêm trên các mô hình cơ sở.

Chương 5: CÁC PHƯƠNG PHÁP THỰC NGHIỆM là nội dung về các phương pháp hiện có để tiếp cận và xử lý bài toán, bao gồm các mô hình cơ sở và các mô hình đa thể thức lớn.

Chương 6: CÀI ĐẶT, THỰC NGHIỆM VÀ ĐÁNH GIÁ nêu rõ các độ đo đánh giá và cách chúng tôi thiết lập tham số thử nghiệm. Kết quả và phân tích kết quả cũng được trình bày trong chương này.

Chương 7: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN bao gồm tổng kết về các đóng góp cũng như hạn chế của đề tài và hướng phát triển trong tương lai.

Chương 2. CÁC CÔNG TRÌNH LIÊN QUAN

2.1. Các bộ dữ liệu liên quan

Bảng 2.1: Thống kê các bộ dữ liệu liên quan

Bộ dữ liệu	Tác vụ chính	Miền dữ liệu	Ngôn ngữ	Đặc điểm nổi bật
MS COCO [12]	Chú thích hình ảnh	Ảnh tự nhiên	Tiếng Anh	Mô tả thực thể trong bối cảnh tự nhiên.
UIT-ViIC [9]	Chú thích hình ảnh	Ảnh tự nhiên	Tiếng Việt	Kế thừa từ MS COCO, tạo chú thích hình ảnh bằng tiếng Việt.
TextCaps [13]	Chú thích hình ảnh	Ảnh chứa chữ	Tiếng Anh	Mô tả hình ảnh yêu cầu hiểu văn bản trong ảnh.
ViVQA [6]	VQA	Ảnh tự nhiên	Tiếng Việt	Câu hỏi-trả lời ngắn, tập trung vào thực thể.
OpenViVQA [7]	VQA	Ảnh tự nhiên	Tiếng Việt	Câu hỏi-câu trả lời tự do, giàu ngữ cảnh và yêu cầu suy luận.
ViTextVQA [8]	VQA	Ảnh chứa chữ	Tiếng Việt	Câu hỏi-trả lời yêu cầu hiểu chữ trong ngữ cảnh thực tế.

ViInfographicVQA [11]	VQA	Infographic	Tiếng Việt	Câu hỏi-trả lời về bố cục và số liệu trong infographic.
InfographicVQA [2]	VQA	Infographic	Tiếng Anh	Câu hỏi-trả lời yêu cầu suy luận đa bước trên bố cục không gian phức tạp.
Chart-to-Text [14]	Tóm tắt	Biểu đồ	Tiếng Anh	Chuyển dữ liệu biểu đồ thành văn bản mô tả tóm tắt.
VisText [15]	Tóm tắt	Biểu đồ	Tiếng Anh	Phân tích sâu biểu đồ, xu hướng từ số liệu.
Screen2Words [16]	Tóm tắt	Giao diện người dùng	Tiếng Anh	Tóm tắt dựa trên cấu trúc bố cục và các thành phần điều hướng.
VietNews [17]	Tóm tắt	Văn bản báo chí	Tiếng Việt	Tóm tắt văn bản tiếng Việt.

Trong giai đoạn đầu của các nghiên cứu về chú thích hình ảnh, các bộ dữ liệu tập trung chủ yếu vào việc mô tả thực thể và bối cảnh tự nhiên, tiêu biểu là bộ dữ liệu MS COCO [12] và UIT-ViIC [9]. Mặc dù thiết lập nền móng vững chắc cho ngôn ngữ mô tả, các công trình này bộc lộ hạn chế khi đối mặt với dữ liệu giàu văn

bản và khả năng lý giải văn bản trong hình ảnh do sự thiếu hụt các đặc trưng về ký tự và số liệu. Sự ra đời của các bộ dữ liệu như TextCaps [13] và ViTextVQA [8] đã đánh dấu một bước chuyển dịch quan trọng khi đưa yếu tố OCR vào quy trình suy luận, buộc mô hình phải kết hợp việc "đọc" văn bản vào ngữ cảnh thị giác. Tuy nhiên, các phương pháp này vẫn tiếp cận văn bản như những thành phần rời rạc trong không gian ảnh tự nhiên, với các bộ dữ liệu tiếng Việt như ViVQA [6], OpenViVQA [7] hay ViTextVQA [8], tác vụ vẫn chủ yếu dừng lại ở mức độ nhận biết cục bộ với các câu trả lời ngắn cho bài toán VQA, chưa thể đáp ứng yêu cầu về tính liên kết và chiều sâu ngữ nghĩa của một bản tóm tắt hoàn chỉnh.

Trong khi đó, yêu cầu về khả năng hiểu cấu trúc thông tin phức tạp được thể hiện rõ nét qua các bộ dữ liệu chuyên biệt như InfographicVQA [2], ViInfographicVQA [11] và Screen2Words [16]. Đặc điểm cốt lõi của nhóm dữ liệu này là xác lập một loại "ngữ pháp thị giác" dựa trên bố cục không gian phi tuyến tính, nơi thông tin được truyền tải thông qua sự kết hợp giữa chữ viết, biểu đồ và các yếu tố dẫn hướng. Trong khi InfographicVQA [2] đặt ra thách thức về khả năng suy luận đa bước trên những hình ảnh có tỷ lệ dài, ViInfographicVQA [11] yêu cầu mô hình suy luận trên nhiều ảnh, thì Screen2Words [16] cung cấp cách tiếp cận tóm tắt dựa trên logic sắp xếp của các thành phần đồ họa. Tuy nhiên, một nhận định quan trọng rút ra là dù các bộ dữ liệu này rất mạnh trong việc định vị thông tin không gian, đầu ra của chúng vẫn mang tính chất trích xuất dữ liệu thô dưới dạng câu hỏi-đáp hoặc tóm tắt liệt kê, dẫn đến sự thiếu hụt các liên kết logic mang tính bao quát để diễn đạt được thông điệp tổng thể của một đồ họa thông tin.

Ở cấp độ cao hơn của khả năng khái quát hóa thông tin, các bộ dữ liệu như Chart-to-Text [14], VisText [15] và VietNews [17] tạo thách thức cho việc việc tóm tắt trừu tượng khi yêu cầu mô hình không chỉ dừng lại ở việc đọc số liệu từ biểu đồ, bảng đồ mà phải thực hiện các phép toán nhận thức để diễn giải xu hướng và các nhận định hàm ý để tạo ra một bản chú thích có giá trị. Song song đó, VietNews [17] đóng vai trò là tiêu chuẩn về năng lực ngôn ngữ, tạo tiền đề cho việc phát triển

các mô hình tóm tắt tiếng Việt đảm bảo tính mạch lạc và chuẩn mực trong diễn đạt cần tính khái quát cao.

2.2. Các phương pháp liên quan

2.2.1. Phương pháp tiếp cận Chú thích hình ảnh truyền thống

Các nghiên cứu trên thế giới về bài toán mô tả hình ảnh thường tập trung khai thác đặc trưng dạng vùng (region feature) hoặc dạng lưới (grid feature) từ mạng CNN để tạo sinh câu mô tả, tiêu biểu là các công trình như [18], [19]. Những phương pháp tiếp cận này thường bộc lộ hạn chế khi xử lý các hình ảnh chứa văn bản mang thông tin quan trọng. Điều này đã thúc đẩy sự ra đời của dòng nghiên cứu mô tả hình ảnh dựa trên OCR (OCR-based Image Captioning).

2.2.2. Phương pháp tiếp cận Chú tích hình ảnh dựa trên OCR

Nhằm giải những khó khăn mà các mô hình mô tả truyền thống đang đối mặt, các nguyên cứu thời điểm đó tập trung vào các việc xây dựng mô hình bằng việc tích hợp thông tin văn bản, được trích xuất từ các mô hình OCR. Hướng tiếp cận này dựa trên nguyên lý chia nhỏ bài toán thành các công đoạn rời rạc, trọng tâm là việc trích xuất thực thể văn bản để phục vụ cho quá trình sinh nội dung. Thách thức lớn nhất mà phương pháp này giải quyết là sự xuất hiện dày đặc của các từ vựng ngoài danh mục (Out-of-Vocabulary - OOV) như số liệu, tên riêng hoặc thuật ngữ chuyên biệt,... Kế thừa tư duy từ Mạng trỏ [20] của tác giả Vinyals và cộng sự, các nghiên cứu tiên phong [21], [22], [23] được công bố đã thiết lập tiêu chuẩn cho việc sử dụng Transformer đa thể thức để căn chỉnh đặc trưng thực thể OCR với đặc trưng thị giác. Điểm cốt lõi của hướng tiếp cận này là cơ chế Mạng trỏ động (Dynamic Pointer Network), cho phép mô hình linh hoạt lựa chọn giữa việc sinh từ mới hoặc sao chép trực tiếp các token OCR vào bản tóm tắt. Các nghiên cứu mở rộng như TAP [24] và LaTr [25] đã chỉ ra rằng việc tiền huấn luyện dựa trên nhận thức văn bản giúp mô hình duy trì tính trung thực của dữ liệu định lượng. Anchor-Captioner [23] đề ra một phương pháp có thể tạo sinh nhiều mô tả dựa vào những từ khóa khác nhau trong câu. trong khi DEVICE [22] đại

diện cho một bước tiến mới trong việc mô hình hóa cấu trúc không gian hình ảnh một cách sâu sắc và đa chiều hơn.

2.2.3. Phương pháp tiếp cận Tóm tắt văn bản trừu tượng

Đối với các loại dữ liệu có bố cục phức tạp và mật độ văn bản dày đặc, nơi các mô hình mô tả hình ảnh thường gặp hạn chế trong việc bao quát toàn bộ ngữ cảnh, việc sử dụng một hệ thống OCR trích xuất văn bản từ hình ảnh và mô tả bằng mô hình Tóm tắt trừu tượng đang là hướng tiếp cận tiềm năng cho các loại dữ liệu hình ảnh giàu văn bản, ví dụ như infographic. Hiện nay, các mô hình tóm tắt dữ liệu được cộng hưởng mạnh mẽ nhờ sức mạnh của các mô hình ngôn ngữ tiền huấn luyện, như T5 [26], Pegasus [27] và BART [28]. Những mô hình này đạt được hiệu suất cao trên nhiều ngôn ngữ, trong đó có cả tiếng Việt. Những phiên bản mô hình đơn ngữ trên tiếng Việt có thể kể đến ViT5 [29], BARTpho [30] đã chứng minh ưu thế vượt trội về hiệu suất so với các phương pháp tiền nhiệm, cho phép tạo sinh bản tóm tắt chính xác, mạch lạc và phù hợp với đặc trưng thành phần ngôn ngữ.

2.2.4. Phương pháp tiếp cận Định hướng từ khóa và thực thể trọng tâm

Bên cạnh việc hiểu cấu trúc không gian, một hướng nghiên cứu quan trọng khác tập trung vào việc sử dụng các từ khóa hoặc thực thể mang thông tin cốt lõi để dẫn dắt quá trình tạo sinh văn bản. Phương pháp này dựa trên giả thuyết rằng hiệu suất tóm tắt sẽ được cải thiện đáng kể nếu mô hình được cung cấp các gợi ý về mặt nội dung trước khi bắt đầu giải mã. Các công trình nghiên cứu về Tóm tắt có hướng dẫn (Guided Summarization) như GSum [31] đã chứng minh rằng việc tích hợp các thực thể bên ngoài (như từ khóa, câu trọng tâm) vào mô hình Transformer [32] giúp kiểm soát tốt hơn nội dung đầu ra, giảm thiểu hiện tượng lạc đề. Trong lĩnh vực đọc hiểu tài liệu, hướng tiếp cận Định hướng từ khóa thường sử dụng các kỹ thuật như Keyword-aware Deep Attentional Model trong [33] để tăng cường trọng số chú ý cho các từ vựng mang tính chủ đề. Trong bối

cảnh infographic, từ khóa thường được xác định thông qua các đặc trưng thị giác nổi bật như kích thước phông chữ lớn hoặc độ tương phản màu sắc cao.

2.2.5. Phương pháp tiếp cận Học tương phản

Phương pháp học tương phản (Contrastive Learning) là một trong những kỹ thuật mạnh mẽ nhất trong học máy hiện đại, đặc biệt là trong quá trình tự học có giám sát. Thay vì học từ các nhãn do con người tạo ra, mô hình học bằng cách so sánh các cặp dữ liệu để tìm ra sự tương đồng và khác biệt. Trong lĩnh vực tóm tắt, CLIFF [34] ra đời để giải quyết vấn đề ảo giác của các mô hình ngôn ngữ. Bằng cơ chế tạo ra các bản tóm tắt lỗi thông qua kỹ thuật tráo đổi thực thể, che từ (masking) hoặc thay đổi nội dung và học từ những nội dung lệch chuẩn đó, CLIFF cực kỳ hiệu quả khi xử lý các dữ liệu đòi hỏi độ chính xác cao như báo cáo tài chính hay Infographic. Trong khi đó, BRIO [35] học cách giải quyết sự học lệch của mô hình khi một từ bị dự đoán sai sẽ khiến toàn bộ nội dung bị sai lệch theo. BRIO tạo ra một tập các ứng viên, xếp hạng chúng dựa trên thang đo đánh giá, BRIO sử dụng học tương phản để đảm bảo rằng các bản tóm tắt có điểm đánh giá cao luôn được hệ thống ưu tiên với xác suất dự đoán lớn nhất.

2.3. Nhận xét và thảo luận

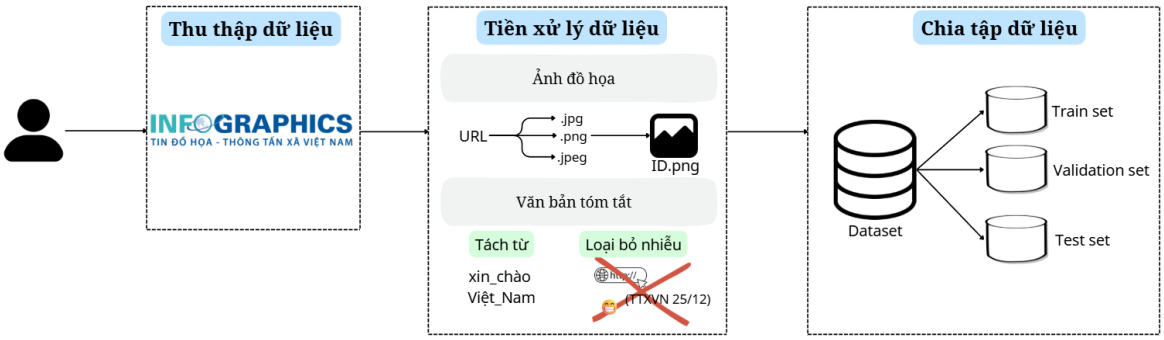
Việc hệ thống hóa các nghiên cứu tiền đề cho thấy một xu hướng phát triển rõ rệt: từ mô tả các thực thể trực quan đơn giản đến việc nhận thức và khái quát hóa các cấu trúc thông tin phi tuyến tính có tính trừu tượng cao từ hình ảnh. Mặc dù các bộ dữ liệu tiếng Việt phục vụ tác vụ trả lời câu hỏi (VQA) đã góp phần làm phong phú tài nguyên đa thể thức, song sự thiếu hụt các nghiên cứu chuyên sâu về tóm tắt hình ảnh phi cấu trúc như Infographic vẫn là một khoảng trống lớn. Đây chính là động lực cốt lõi để nghiên cứu này tập trung khỏa lấp sự thiếu hụt về tài nguyên dữ liệu, đồng thời giải quyết bài toán tóm tắt ở cấp độ khái quát sâu sắc hơn.

Bên cạnh việc xác lập bài toán, các phương pháp tiếp cận hiện hữu còn đóng vai trò là nguồn ý tưởng và nền tảng phương pháp luận then chốt để xây dựng khung

giải pháp. Chúng tôi coi các công trình đi trước là cơ sở khoa học quan trọng để tiến hành đánh giá, kế thừa và phát triển các cải tiến mới trong phần đề xuất mô hình nghiên cứu của mình.

Chương 3. XÂY DỰNG BỘ DỮ LIỆU ViInfographicCaps

3.1. Quy trình xây dựng dữ liệu



Hình 3.1: Quy trình xây dựng bộ dữ liệu

3.1.1. Tổng quan quy trình

Quy trình xây dựng bộ dữ liệu đóng vai trò then chốt, quyết định trực tiếp đến hiệu năng và khả năng tổng quát hóa của các mô hình học sâu. Tổng quan quy trình được mô tả trực quan tại *Hình 3.1*. Chúng tôi sẽ lý giải chi tiết các giai đoạn xây dựng bộ dữ liệu lần lượt từ khâu thu thập đến tiền xử lý và phân chia dữ liệu thực nghiệm ở các phần sau.

3.1.2. Thu thập dữ liệu

Chúng tôi tiến hành khảo sát thực nghiệm trên nhiều nguồn tài nguyên hình ảnh đồ họa thông tin trực tuyến như Google Images[2], Pinterest[3], và các trang tin điện tử tổng hợp. Qua quá trình đánh giá định tính và định lượng, Trang thông tin Tin Đồ Họa – Thông Tấn Xã Việt Nam được chúng tôi lựa chọn làm nguồn dữ liệu sơ cấp, bởi chúng tôi thấy rằng đây là một nguồn dữ liệu chất lượng, phù hợp dựa trên các tiêu chí:

[2] https://images.google.com/
[3] https://www.pinterest.com/

- **Tính chuẩn xác và tin cậy**: Trang này thuộc quản lý của cơ quan thông tin chính thức của Chính phủ Nhà nước Việt Nam, đảm bảo độ chính xác tuyệt đối về số liệu và chuẩn mực về ngôn ngữ báo chí trang trọng tiếng Việt.
- **Sự đa dạng về miền thể loại**: Dữ liệu tổng hợp trải dài trên nhiều lĩnh vực như Kinh tế, Chính trị, Giáo dục, Khoa học,... đảm bảo độ phủ rộng của chủ đề nội dung.
- **Cấu trúc dữ liệu đồng nhất và đầy đủ đặc trưng:** Mỗi mẫu dữ liệu đều bao gồm hình ảnh infographic đi kèm tiêu đề và văn bản tóm tắt nội dung chính, là những yêu cầu đặc trưng đầu vào – đầu ra cho bài toán tóm tắt hình ảnh của chúng tôi.

Chúng tôi triển khai công cụ trích xuất dữ liệu tự động Octoparse[4] để thu thập các trường thông tin từ Trang bao gồm: URL, Tiêu đề, Văn bản tóm tắt, Lĩnh vực. Sau giai đoạn làm sạch sơ bộ, chúng tôi thu được bộ dữ liệu gồm hơn 17.000 mẫu có cấu trúc.

3.1.3. Tiền xử lý dữ liệu

Chúng tôi thực hiện tiền xử lý riêng biệt cho dữ liệu hình ảnh và văn bản.

- **Đối với ảnh infographic:**
  Sau khi ảnh được tải về từ các link URL (và đồng thời loại bỏ các đường link lỗi), chúng tôi thực hiện chuẩn hóa định dạng: toàn bộ dữ liệu ảnh được chuyển đổi đồng nhất sang định dạng PNG giúp bảo toàn độ sắc nét của ảnh nén.

  Chúng tôi thiết lập cơ chế đặt tên tệp dựa trên mã định danh ID gồm 7 chữ số (ví dụ: 0000001.png). Việc này nhằm mục đích quản lý có hệ thống bộ dữ liệu, đồng thời sẵn sàng cho việc kế thừa, tổng hợp hay mở rộng quy mô bộ dữ liệu trong tương lai (nếu có) mà không ảnh hưởng đến cấu trúc bộ dữ liệu.

[4] https://www.octoparse.com/

- **Đối với văn bản chú thích tóm tắt:**
  Nguồn chú thích tóm tắt được biên tập theo phong cách báo chí, vì vậy các yếu tố đặc trưng của ngôn ngữ mạng xã hội như sai chính tả, viết tắt, teencode không đáng có. Chúng tôi đã thực hiện một số bước tiền xử lý như xóa bỏ các đường dẫn liên kết, các thông tin về ngày đăng, địa điểm, thẻ nguồn (ví dụ: "TTXVN", "Ảnh: ") đi.

  Ngoài ra, do tiếng Việt là ngôn ngữ đơn lập với đặc trưng từ ghép, để hỗ trợ cho một số mô hình ngôn ngữ nhận diện chính xác ranh giới từ, chúng tôi sử dụng thư viện VnCoreNLP [36] để tiến hành tách từ, đảm bảo chất lượng của bộ từ điển.

3.1.4. Chia tập dữ liệu

Để đảm bảo mỗi tập dữ liệu đều có tính đại diện cho toàn bộ quần thể, chúng tôi cho rằng có hai yếu tố cần phải quan tâm ở việc chia tập dữ liệu này.

Thứ nhất là về phân phối lĩnh vực, nghĩa là về chủ đề mà infographic nói tới. Chúng tôi thực hiện phân lớp theo chủ đề để đảm bảo tỷ lệ các lĩnh vực trong các tập con là tương đồng với phân phối chủ đề gốc, đảm bảo giữ đúng đặc trưng của dữ liệu.

Thứ hai là về độ dài của nội dung tóm tắt. Nhằm đánh giá năng lực tóm tắt ở các cấp độ chi tiết khác nhau, chúng tôi kiểm soát sao cho phân phối độ dài của văn bản chú thích giữa tập huấn luyện và tập kiểm thử không có sự sai lệch đáng kể.

Chúng tôi thực hiện kỹ thuật chia nhóm theo phân vị dựa trên số lượng từ trong văn bản chú thích. Độ dài của tóm tắt được chia thành 5 nhóm từ ngắn đến dài. Sau đó, chúng tôi tạo ra một nhãn phân lớp tổng hợp bằng cách kết hợp hai yếu tố lĩnh vực và độ dài tóm tắt. Việc phân chia dựa trên nhãn này giúp đảm bảo rằng tập kiểm định và đánh giá sẽ có cùng mức độ phân bổ về cả chủ đề lẫn độ dài văn bản như tập huấn luyện, từ đó đưa ra kết quả đánh giá trung thực nhất về khả năng tổng quát hóa của mô hình.

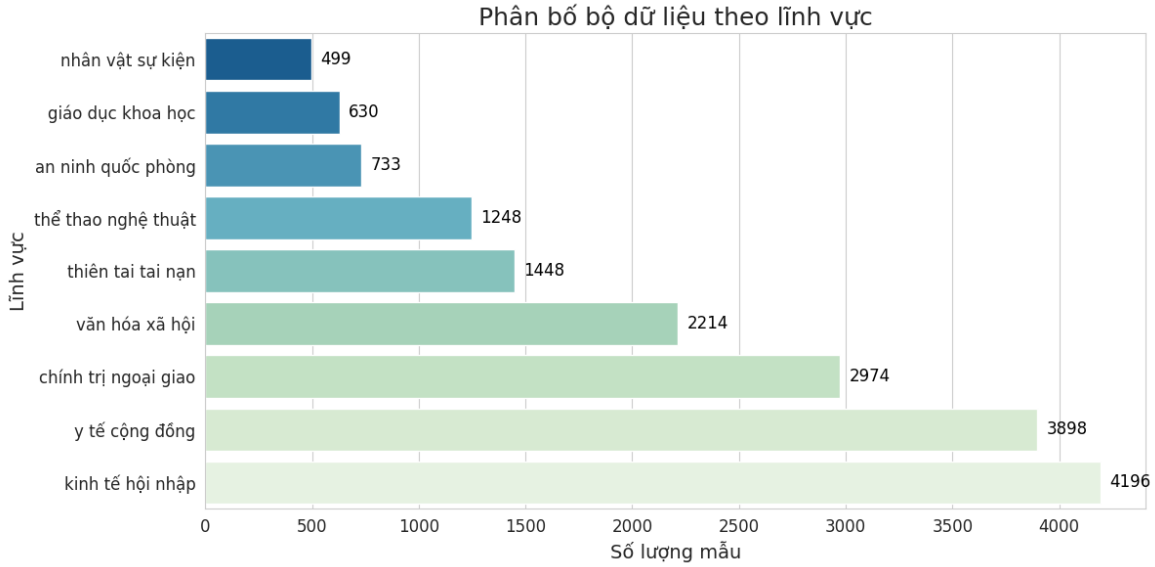
Dữ liệu được phân chia theo tỷ lệ 8-1-1 tương ứng cho các tập huấn luyện-kiểm định-đánh giá.

3.2. Phân tích bộ dữ liệu

3.2.1. Phân tích thống kê tổng quát

Hình 3.2 minh họa cơ cấu phân bổ của bộ dữ liệu thực nghiệm qua 9 lĩnh vực đời sống xã hội khác nhau. Với tổng quy mô mẫu đạt 17.840 mẫu dữ liệu, biểu đồ treemap của chúng tôi cho thấy một sự phân hóa rõ rệt về dung lượng giữa các chủ đề, phản ánh tính đa dạng và đa chiều trong nguồn dữ liệu được thu thập.

Trước hết, nhóm các lĩnh vực chiếm tỷ trọng áp đảo bao gồm "*Kinh tế hội nhập*" và "*Y tế cộng đồng*". Cụ thể, lĩnh vực *Kinh tế hội nhập* dẫn đầu với 4.196 mẫu (chiếm khoảng 23,5%), theo sát là *Y tế cộng đồng* với 3.898 mẫu (chiếm 21,8%). Việc hai nhóm này chiếm tới gần 50% tổng dung lượng bộ dữ liệu cho thấy đây là những mảng nội dung trọng tâm thường được biên tập bằng infographic để truyền đạt thông tin, có tần suất xuất hiện cao trong nguồn tin trích xuất.



Hình 3.2: Phân phối dữ liệu theo lĩnh vực

Kế đến, nhóm các lĩnh vực có mức độ phân bổ trung bình tạo nên sự cân bằng cho

bộ dữ liệu. Các chủ đề như "*Chính trị ngoại giao*" (2.974 mẫu) và "*Văn hóa xã hội*" (2.214 mẫu) đóng vai trò là cầu nối thông tin, đảm bảo tính toàn diện về mặt nội dung xã hội. Bên cạnh đó, lĩnh vực "*Thiên tai tai nạn*" với 1.448 mẫu và "*Thể thao nghệ thuật*" với 1.248 mẫu cũng duy trì một lượng dữ liệu đủ.

Cuối cùng, số lượng mẫu của các lĩnh vực có tỷ trọng thấp hơn như "*An ninh quốc phòng*" (733 mẫu), "*Giáo dục khoa học*" (630 mẫu) và *"Nhân vật sự kiện*" (499 mẫu) cho thấy sự bao quát các ngách thông tin chuyên sâu. Mặc dù số lượng mẫu ở các nhóm này khiêm tốn hơn so với nhóm dẫn đầu, nhưng chúng đóng vai trò quan trọng trong việc kiểm chứng khả năng tổng quát hóa của mô hình đối với các lớp dữ liệu ít phổ biến.

Việc đánh giá tổng quát về sự phân bổ dữ liệu tại *Hình 3.2* cho thấy một cấu trúc dữ liệu tương đối ổn định. Dù có sự chênh lệch về số lượng giữa các nhãn, nhưng mỗi lĩnh vực đều sở hữu một lượng mẫu đủ để thực hiện các phân tích thống kê có ý nghĩa. Sự tập trung mạnh mẽ vào các lĩnh vực kinh tế, y tế và chính trị phản ánh đúng dòng chảy thông tin chủ lưu trong thực tế, từ đó gia tăng tính ứng dụng và giá trị thực tiễn cho bộ dữ liệu trong các tác vụ xử lý ngôn ngữ tự nhiên hoặc phân tích dữ liệu lớn sau này.

3.2.2. Phân tích đặc điểm thống kê của văn bản chú thích

Trong phân tích này, chúng tôi thiết lập 4 loại thước đo khác nhau để phân tích thống kê văn bản chú thích tóm tắt. Việc đa dạng hóa các thước đo giúp phản ánh đầy đủ từ đặc tính hình thái học đến dung lượng lưu trữ của dữ liệu:

- **Từ (word)** là đơn vị từ vựng sau khi thực hiện tách từ dựa trên khoảng trắng. Đặc điểm của thước đo này là tính cả các dấu câu (chấm, phẩy, hỏi chấm...) như một token độc lập. Đây cũng là thước đo để xác định kích thước đầu vào cho các mô hình ngôn ngữ.
- **Từ (bỏ dấu câu)** được loại bỏ hoàn toàn các token chỉ chứa ký tự dấu câu. Chỉ số này sẽ phản ánh lượng thông tin thực mà nội dung chú thích chứa, loại bỏ các nhiễu về mặt cấu trúc ngữ pháp.

- **Âm tiết (syllables)**: Vì tiếng Việt là một ngôn ngữ đơn lập, một token sau khi tách từ có thể chứa nhiều âm tiết (ví dụ: Việt_Nam là một token chứa hai âm tiết) nên thang đo này sẽ đếm âm tiết giúp phản ảnh chính xác về độ dài phát âm và nhịp điệu của văn bản.
- **Ký tự (characters)** là thước đo sẽ tính tổng số lượng ký tự trong toàn bộ chú thích, đóng vai trò chỉ số tham chiếu cơ bản để tính toán chi phí lưu trữ của dữ liệu.

Dựa trên kết quả thống kê tại *Bảng 3.1*, bộ dữ liệu cho thấy những đặc điểm nhân trắc học ngôn ngữ rõ rệt.

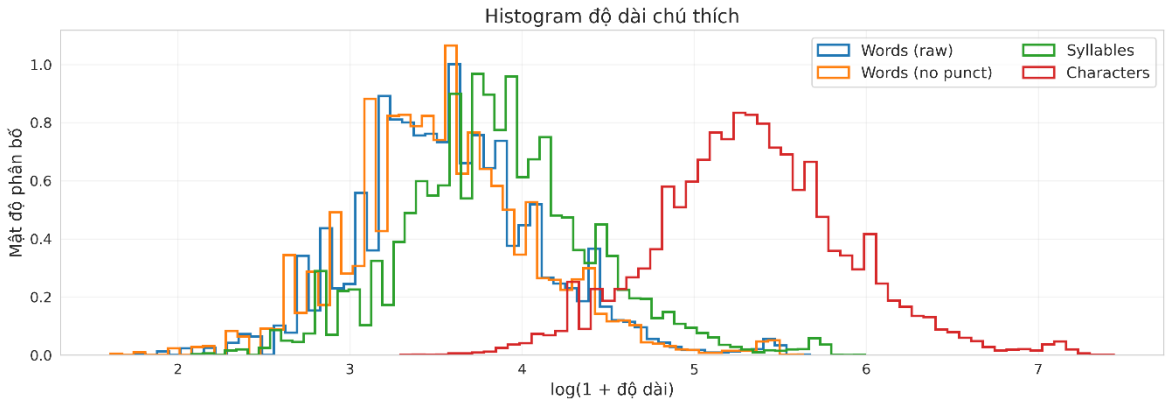
Tại tất cả các thước đo, giá trị trung bình (Mean) luôn cao hơn giá trị trung vị (Median). Ví dụ, số lượng từ trung bình là 41.19 nhưng có 50% dữ liệu nằm dưới mức 34 từ. Điều này chứng tỏ phần lớn các chú thích được viết theo phong cách súc tích, ngắn gọn, trong khi các giá trị cực đại (lên tới 290 từ hoặc 1699 ký tự) tạo nên một "đuôi dài" (long tailed) trong phân phối.

Độ lệch chuẩn (Std) tương đối lớn so với giá trị trung bình cho thấy sự đa dạng cao về dung lượng nội dung, phản ánh tính thực tế của dữ liệu từ các đoạn mô tả ngắn đến các phân tích chi tiết. Chênh lệch giữa từ có và không có dấu câu (~1.73 dấu câu/chú thích) thể hiện đặc trưng văn phong trực diện, ít cấu trúc ngữ pháp phức tạp.

Ngoài ra, để làm rõ hơn hình dạng phân phối, chúng tôi sử dụng biểu đồ Histogram với phép biến đổi $y = \log(1 + x)$ với x là độ dài của chú thích (*Hình 3.3*). Việc nén khoảng cách ở các giá trị cực lớn giúp các đường bậc thang trong biểu đồ histogram của từ và âm tiết hiện rõ dạng hình chuông đối xứng. Điều này xác nhận rằng độ dài chú thích tuân theo Phân phối Log-normal – một đặc tính điển hình trong ngôn ngữ tự nhiên. Sự trùng khớp về đỉnh giữa hai dạng token theo từ cũng khẳng định cấu trúc ngữ pháp của bộ dữ liệu đạt độ đồng nhất cao.

Bảng 3.1: Bảng thống kê độ dài chú thích tóm tắt

	min	max	mean	median	std
Từ	4	290	41.19	34	28.46
Từ (bỏ dấu câu)	4	278	39.46	33	27.75
Âm tiết	7	400	55.06	46	38.37
Ký tự	26	1699	243.67	203	164.9



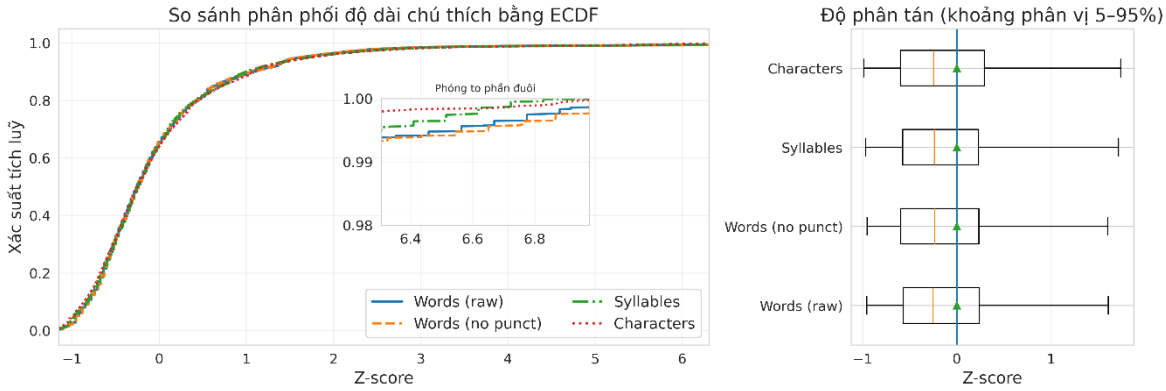
Hình 3.3: Phân phối độ dài nội dung chú thích sử dụng hàm log(1+x)

3.2.3. Phân tích tương quan và độ phân tán văn bản chú thích

Chúng tôi thực hiện phân tích tương quan và độ phân tán của văn chú thích, thể hiện qua *Hình 3.4.* Để so sánh các thước đo có đơn vị khác nhau trên cùng một hệ quy chiếu, dữ liệu được chuẩn hóa về phân phối chuẩn.

Biểu đồ phân phối tích lũy thực nghiệm (ECDF) cho thấy sự trùng khớp gần như tuyệt đối của cả 4 đường cong. Điều này chứng minh rằng bất kể sử dụng đơn vị đo lường nào, quy luật tăng trưởng độ dài của tập dữ liệu vẫn giữ được tính nhất quán cực cao. Tuy nhiên, tại vùng đuôi (Z-score > 6), đường Ký tự và Âm tiết có xu hướng tách nhẹ. Hiện tượng này chỉ ra rằng ở những mẫu chú thích cực dài, các tác giả thường có xu hướng sử dụng từ vựng phức tạp, đa âm tiết hơn.

Bên cạnh đó, biểu đồ Boxplot cung cấp cái nhìn trực quan về sự tập trung dữ liệu. Khoảng tứ phân vị nằm lệch hẳn về phía bên trái của trục số, củng cố nhận định về độ lệch phải. Đường whisker bên phải dài hơn đáng kể so với bên trái và khoảng cách từ Median đến ngưỡng 95% dài gấp nhiều lần so với khoảng cách đến ngưỡng 5%, khẳng định biến động của các chú thích dài là rất lớn và khó dự đoán hơn so với các chú thích ngắn.



Hình 3.4: Phân tích tương quan và độ phân tán của chú thích tóm tắt

3.2.4. Phân tích độ đa dạng từ vựng theo lĩnh vực

Để đánh giá tính phong phú và đặc thù ngôn ngữ của tập dữ liệu, chúng tôi thực hiện phân tích độ đa dạng từ vựng dựa trên sự kết hợp của ba chỉ số chính qua Biểu đồ cột – đường (*Hình 3.5*).

Ở đây, chúng tôi sử dụng hệ thống ba biến số để định lượng đặc tính từ vựng:

- **Tổng số Token (Cột):** Phản ánh quy mô dữ liệu thô của từng lĩnh vực.
- **Tỉ lệ Type–Token (TTR):** Chỉ số này đo lường mức độ phong phú từ vựng. Chỉ số này có thể bị ảnh hưởng bởi số kích thước mẫu được kiểm tra.

  TTR được xác định bằng công thức:

$$\text{TypesTokenRatio(TTR)} = \frac{\text{Số lượng từ duy nhất}}{\text{Tổng số từ}} \quad (1)$$

- **Tỉ lệ Sampled-TTR:** Là chỉ số TTR được tính toán sau khi lấy mẫu ngẫu nhiên cùng một số lượng token cố định cho tất cả các lĩnh vực. Mục tiêu là để loại bỏ thiên lệch do quy mô của các nhóm mẫu dữ liệu, cho phép so

sánh sự đa dạng ngôn ngữ một cách công bằng giữa các nhóm dữ liệu có kích thước khác nhau.

Phân tích dựa trên biểu đồ trực quan, ta thấy rằng:

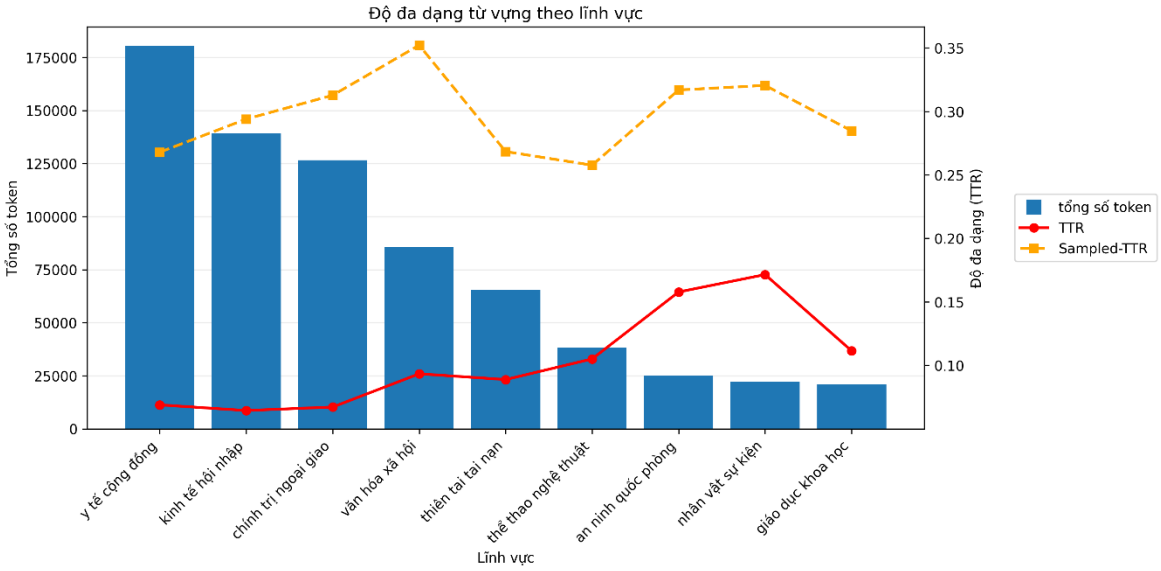
Về quy mô, lĩnh vực *Y tế cộng đồng* chiếm ưu thế tuyệt đối về tổng số token, theo sau là *Kinh tế – hội nhập* và *Chính trị – ngoại giao*. Ngược lại, các nhóm như *An ninh – quốc phòng* và *Giáo dục – khoa học* có dung lượng khiêm tốn hơn đáng kể.

Quan sát đường TTR (nét liền, màu đỏ), có thể thấy một sự tương quan nghịch rõ rệt: các lĩnh vực có tổng số token lớn thường có tỷ lệ TTR thấp hơn. Hiện tượng này hoàn toàn phù hợp với các quy luật ngôn ngữ học thống kê, khi quy mô văn bản tăng lên, các từ loại chức năng (hư từ) và thuật ngữ chuyên ngành lặp lại với tần suất cao hơn, làm giảm tỷ lệ từ mới xuất hiện. Ngược lại, ở những lĩnh vực nhỏ hơn như *Nhân vật – sự kiện*, TTR đạt mức cao nhất do sự tập trung dày đặc của các thực thể tên riêng và danh từ riêng duy nhất.

Sau khi chuẩn hóa quy mô bằng đường Sampled-TTR (nét đứt, màu cam), các đặc tính ngôn ngữ của từng lĩnh vực thể hiện như sau:

- *Văn hóa – xã hội* nổi bật với giá trị Sampled-TTR cao nhất trong toàn bộ tập dữ liệu. Điều này chứng tỏ trong cùng một đơn vị dung lượng, vốn từ vựng của lĩnh vực này phong phú và biến hóa hơn các nhóm khác, ít bị gò bó bởi các cấu trúc lặp lại.
- *An ninh – quốc phòng* và *Nhân vật – sự kiện* cũng duy trì mức đa dạng cao. Đặc thù của hai lĩnh vực này là chứa nhiều danh từ riêng, thực thể định danh và các thuật ngữ chuyên biệt không lặp lại, tạo nên sự phong phú về mặt ngữ nghĩa cho mô hình học máy.
- Ngược lại, *Thiên tai – tai nạn* và *Thể thao – nghệ thuật* có chỉ số đa dạng thấp hơn. Nguyên nhân chủ yếu do các lĩnh vực này thường sử dụng văn phong tin tức theo khuôn mẫu, tập trung vào các thông số cố định như

thời gian, địa điểm, con số thiệt hại hoặc kết quả trận đấu, dẫn đến việc lặp lại các cấu trúc từ vựng đặc thù.



Hình 3.5: Biểu đồ phân tích độ đa dạng từ vựng theo lĩnh vực

3.2.5. Đánh giá mức độ tương quan giữa nội dung chú thích và văn bản trích xuất (OCR)

Để đánh giá sự trùng lặp của nội dung chú thích vào các thông tin văn bản xuất hiện trực tiếp trong hình ảnh, chúng tôi thực hiện trích xuất đặc trưng văn bản bằng công cụ PaddleOCR [37] kết hợp VietOCR [38].

Tỷ lệ trùng lặp được xác định dựa trên tỷ lệ các token trong chú thích đồng thời xuất hiện trong kết quả OCR trích xuất được.

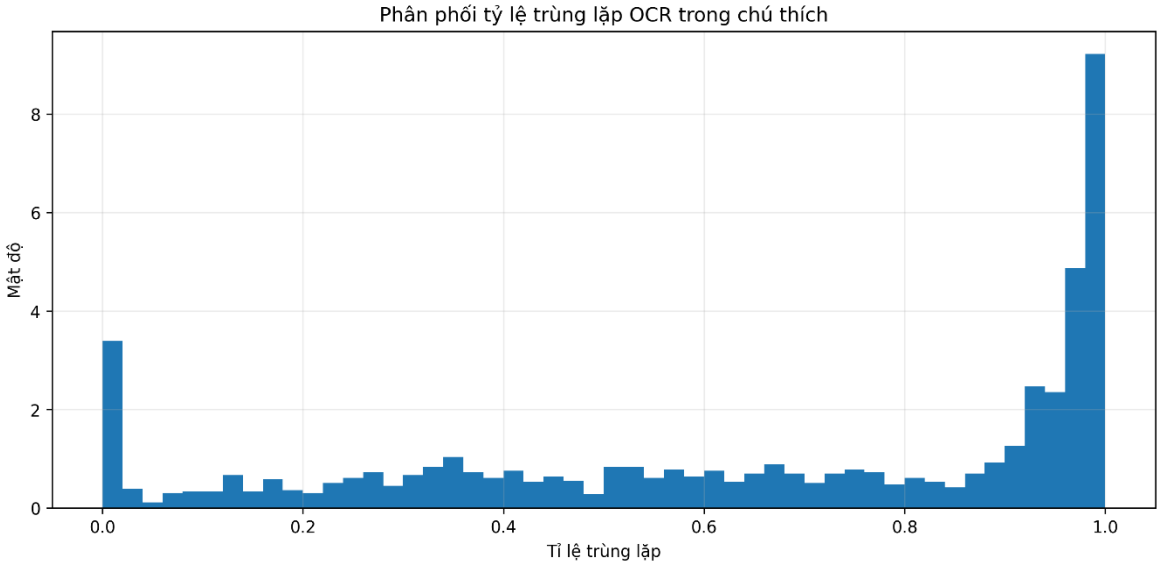
Kết quả thống kê tại *Bảng 3.2* cho thấy một sự tương quan rất mạnh mẽ giữa nội dung mô tả và dữ liệu OCR. Dựa trên số liệu này, có thể thấy hơn 80% số lượng chú thích trong bộ dữ liệu chứa ít nhất 30% thông tin từ OCR. Đáng chú ý, có tới 40,36% mẫu dữ liệu đạt mức độ trùng lặp gần như nguyên văn ($\geq$ 90\%). Điều này khẳng định rằng trong ngữ cảnh của bộ dữ liệu này, việc tóm tắt hoặc viết chú thích dựa trên việc tái sử dụng các từ vựng xuất hiện trong ảnh là một xu hướng phổ biến.

Chúng tôi trực quan hóa mật độ phân phối tỷ lệ trùng lặp OCR thông qua biểu đồ Histogram đã được chuẩn hóa ở *Hình 3.6.* Ở biểu đồ này, trục tung biểu diễn mật độ trùng khớp, sao cho tổng diện tích của tất cả các cột bằng 1. Thay vì biểu thị số lượng mẫu tuyệt đối, chiều cao của mỗi cột thể hiện mức độ tập trung tương đối của dữ liệu. Tỷ lệ mẫu trong một khoảng giá trị cụ thể sẽ tương ứng với diện tích của các cột trong vùng đó (xấp xỉ mật độ x độ rộng bin). Cách tiếp cận này giúp đánh giá xác suất phân bổ dữ liệu một cách khách quan hơn trên toàn bộ tập dữ liệu.

Quan sát phân phối cho thấy khoảng 0.9–1.0 xuất hiện đỉnh rõ rệt với mật độ dày, đồng nghĩa với việc một phần đáng kể các mẫu tập trung ở vùng trùng lặp rất cao. Nói cách khác, nhiều chú thích có xu hướng sử dụng rất nhiều token giống văn bản xuất hiện trong infographic. Ngoài vùng đỉnh này, phân phối trải rộng từ khoảng 0.1 đến 0.8, phản ánh các mức trùng lặp trung bình đến cao, cho thấy chú thích vẫn có mức tùy biến khác nhau tuỳ mẫu. Đồng thời, cũng xuất hiện một cụm tương đối lớn gần 0, nghĩa là tồn tại không ít chú thích gần như độc lập với OCR, hoặc các trường hợp OCR bị thiếu/không khớp nội dung khiến phần giao nhau về token giữa caption và OCR rất thấp.

Bảng 3.2: Bảng thể hiện mức độ trùng lặp OCR trên số mẫu

Mức độ trùng lặp	**% số mẫu trên tổng thể**
≥ 30%	81.11%
≥ 50%	67.77%
≥ 70%	53.20%
≥ 90%	40.36%



Hình 3.6: Biểu đồ phân phối tỷ lệ trùng lặp OCR trong chú thích

3.2.6. Phân tích mật độ thông tin và đặc trưng thực thể

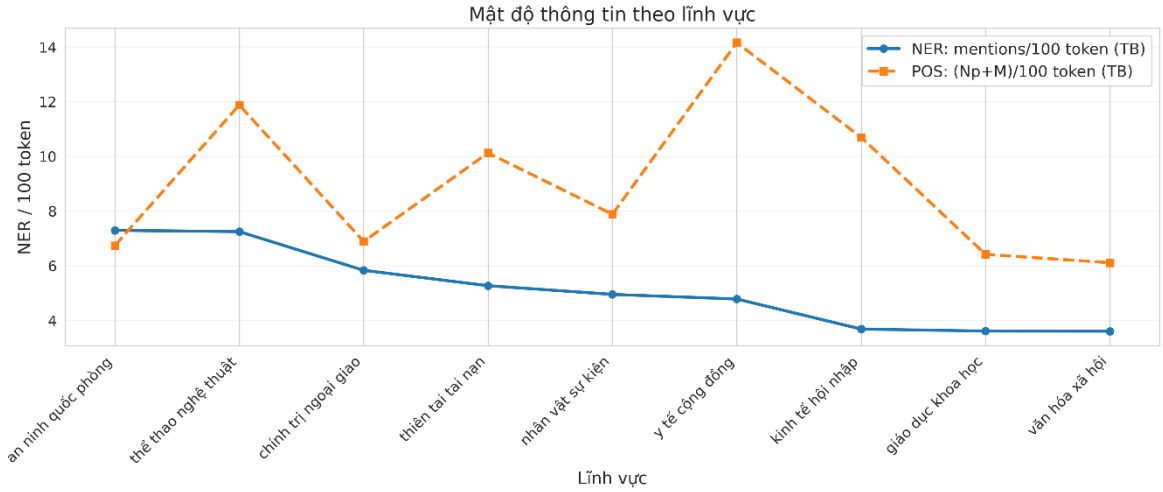
Để đánh giá sâu hơn về giá trị nội dung của bộ dữ liệu, chúng tôi thực hiện phân tích mật độ thông tin thông qua hai chỉ số định lượng chính: thực thể tên riêng và từ loại sử dụng thư viện VnCoreNLP [36] để trích xuất đặc trưng NER (Name Entity Recognition) và POS (Part-of-Speech).

Sau đó, chúng tôi tính toán các chỉ số:

- **NER (mentions/100 tokens)**: Đo lường số lần xuất hiện trung bình của các thực thể (tên người, tổ chức, địa danh...) trên mỗi 100 đơn vị từ vựng. Chỉ số này phản ánh mức độ "giàu thực thể" của chú thích.
- **POS (N_p +M)/100 tokens**: Đo lường mật độ danh từ riêng (N_p) và số từ/định lượng (M) trên mỗi 100 token. Đây là thước đo mật độ sự kiện, phản ánh mức độ chi tiết về mặt số liệu, ngày tháng và định danh cụ thể trong văn bản.

Dựa trên kết quả thực nghiệm tại *Hình 3.7*, chúng tôi rút ra các nhận định quan trọng về đặc thù nội dung của từng lĩnh vực. Đường POS (nét đứt, màu cam) có xu hướng dao động mạnh và luôn duy trì ở mức cao hơn so với đường NER (nét liền, màu xanh). Điều này là hợp lý về mặt ngôn ngữ học, bởi chỉ số POS bao

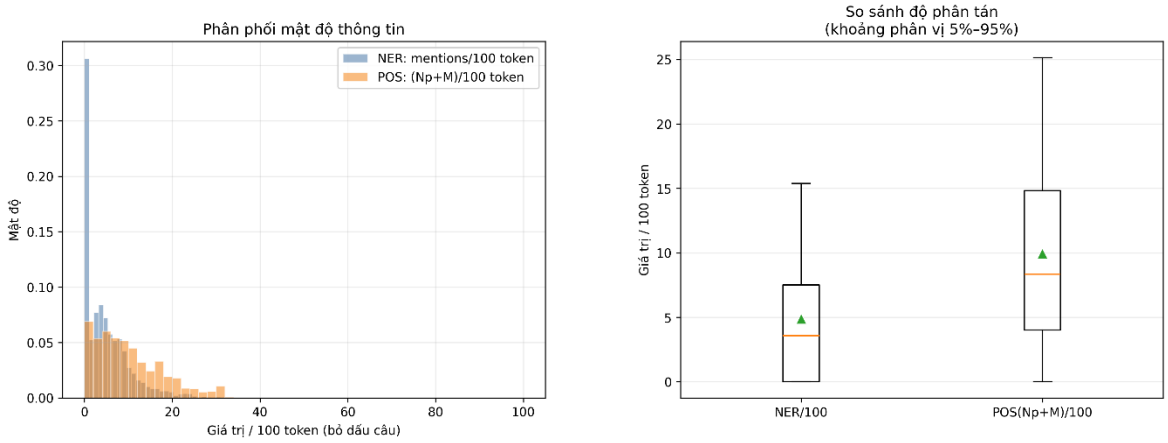
quát được phạm vi rộng hơn các đơn vị mang thông tin (số liệu, ký hiệu định lượng), trong khi NER chỉ tập trung vào các nhóm thực thể định danh cụ thể. Đường NER cho thấy sự ổn định cao hơn, chứng tỏ lượng thực thể cốt lõi được phân bổ tương đối đồng đều giữa các nhóm chủ đề. Một số lĩnh vực ghi nhận sự bùng nổ của chỉ số POS trong khi NER không tăng tương ứng (ví dụ: *Y tế cộng đồng*). Điều này phản ánh văn phong chú thích chứa nhiều con số, biểu mẫu hoặc mã hiệu – những thành phần mang tính thông tin cao nhưng không được phân loại là thực thể tên riêng truyền thống. Ngược lại, các lĩnh vực như *An ninh quốc phòng* có chỉ số NER cao vượt trội, cho thấy nội dung tập trung nhiều vào các đối tượng định danh như nhân vật hoặc tổ chức quốc tế.



Hình 3.7: Biểu đồ mật độ thông tin theo lĩnh vực

Quan sát *Hình 3.8*, đặc tính mật độ thông tin của bộ dữ liệu được làm rõ qua các góc độ thống kê. Cả hai chỉ số NER và POS đều có dạng phân phối lệch phải rõ rệt. Phần lớn các chú thích duy trì mật độ thông tin ở mức trung bình, tuy nhiên tồn tại một nhóm các mẫu "giàu thông tin" (long tailed) với mật độ thực thể và số liệu cực kỳ dày đặc. Biểu đồ Boxplot cho thấy chỉ số POS (N_p +M) có khoảng phân vị và đường whisker trên dài hơn đáng kể so với NER. Điều này chứng tỏ sự khác biệt về lượng số liệu và danh từ riêng giữa các chú thích là rất lớn, có những mẫu chứa mật độ sự kiện rất cao, vượt xa mức trung bình của tập dữ liệu. Ngoài ra, với hộp có diện tích hẹp và độ phân tán thấp hơn, chỉ số NER thể hiện rằng

việc sử dụng các thực thể định danh trong chú thích tuân theo một quy luật ổn định và ít có sự đột biến cực đoan như các chỉ số định lượng.



Hình 3.8: Biểu đồ phân phối và phân tán thông tin

3.3. Kết luận về bộ dữ liệu

Nhóm đã xây dựng thành công bộ dữ liệu ViInfographicCaps với 17.840 mẫu dữ liệu bao gồm hình ảnh infographic và chú thích tóm tắt tương ứng. Đây là bộ dữ liệu đầu tiên về bài toán chú thích tóm tắt trừu tượng từ ảnh trong tiếng Việt.

Tổng kết các đặc tính thực nghiệm của bộ dữ liệu cho thấy tập chú thích tóm tắt sở hữu các đặc tính thống kê ổn định và tuân thủ chặt chẽ quy luật phân phối ngôn ngữ học tự nhiên. Mặc dù tồn tại hiện tượng lệch phải về độ dài, song tính nhất quán cao giữa các thước đo và sự tập trung dữ liệu ở ngưỡng dung lượng vừa phải là tiền đề quan trọng cho việc huấn luyện các mô hình tóm tắt văn bản tự động.

Về mặt nội dung, độ đa dạng từ vựng không chỉ thể hiện qua sự bao phủ rộng rãi của các chủ đề xã hội mà còn nằm ở vốn từ phong phú, biến thiên linh hoạt theo đặc thù của từng lĩnh vực chuyên biệt. Đặc biệt, các kết quả thực nghiệm khẳng định thông tin văn bản trích xuất trực tiếp từ hình ảnh đóng vai trò là đặc trưng cốt lõi cấu thành nên nội dung chú thích. Việc hơn một nửa dung lượng tập dữ liệu có mức độ trùng lặp thông tin từ ảnh đạt ngưỡng cao là minh chứng thực nghiệm quan trọng để đề xuất hướng tiếp cận mô hình tập trung khai thác và xử lý đặc trưng văn bản (OCR) trong ảnh nhằm tối ưu hóa hiệu năng sinh tóm tắt.

Bên cạnh đó, mật độ thông tin cao với sự hiện diện dày đặc của các thực thể tên riêng và dữ liệu định lượng đã mang lại giá trị tri thức lớn cho bộ dữ liệu. Sự kết hợp giữa lượng thực thể ổn định và các sự kiện biến thiên đa dạng tạo nên một nguồn tài nguyên đầy tiềm năng cho các bài toán trích xuất thông tin và sinh văn bản có định hướng thực thể trong nghiên cứu mô hình đề xuất của chúng tôi.

Chương 4. KIẾN TRÚC ĐỀ XUẤT

4.1. Ý tưởng thiết kế

Infographic là dạng tài liệu đặc trưng bởi mật độ văn bản dày đặc được sắp xếp theo các khối bố cục phức tạp. Qua đánh giá bằng thực nghiệm của chúng tôi, phương pháp mô tả hình ảnh dựa trên OCR (OCR-based Image Captioning) bộc lộ hạn chế lớn khi không thể nắm bắt trọn vẹn ngữ nghĩa và nội dung chi tiết trong Infographic. Ngược lại, tiếp cận theo hướng tóm tắt văn bản trừu tượng (Abstractive Text Summarization) - thông qua việc trích xuất toàn bộ nội dung văn bản (OCR) - cho thấy kết quả vượt trội hơn về mặt thông tin so với hướng tiếp cận mô tả hình ảnh. Tuy nhiên, hướng tiếp cận này tồn tại những khó khăn như sau:

- Mất mát ngữ cảnh trọng tâm: Infographic là sự tổng hợp có thứ tự giữa thông tin chính (tiêu đề, đề mục, từ khóa nổi bật) và thông tin chi tiết. Khi trích xuất dữ liệu OCR, đưa đầu vào sang dạng văn bản thuần túy, các dấu hiệu thị giác quan trọng (như bố cục, kích thước các cụm văn bản) bị loại bỏ. Do đó, mô hình tóm tắt không thể phân định chính xác đâu là vùng thông tin trọng tâm cần được ưu tiên đưa vào câu tóm tắt.
- Hiện tượng ảo giác đối với các thực thể: Nhằm đảm bảo tính trực quan, ngắn gọn và chính xác, Infographic thường chứa nhiều nội dung về thực thể tên riêng và dữ liệu định lượng bằng số liệu, phần trăm,... Tuy nhiên, các mô hình tóm tắt trừu tượng thường mắc lỗi "ảo giác", dẫn đến việc sinh ra các con số hoặc tên riêng sai lệch so với văn bản gốc. Sự sai lệch này ảnh hưởng nghiêm trọng đến tính toàn vẹn và độ tin cậy của dữ liệu đầu ra.

Để giải quyết các thách thức đặc thù trong việc tóm tắt Infographic, chúng tôi đề xuất một kiến trúc hệ thống tập trung vào việc kết hợp giữa ngữ nghĩa văn bản và cấu trúc hình ảnh thông qua ba chiến lược sau:

- **Tối ưu hóa quy trình trích xuất văn bản theo cụm**: Nhóm sử dụng quy trình OCR theo cụm để phân nhóm nội dung theo bố cục, giúp duy trì tính

mạch lạc về nội dung, đảm bảo luồng thông tin logic ngay từ giai đoạn thu thập dữ liệu.

- **Cơ chế nhận diện tương quan không gian của các vùng bố cục**: Dựa trên các cụm văn bản đã trích xuất, nhóm xây dựng một module phân tích vị trí dựa trên chỉ số Intersection over Union (IoU).
- **Hàm học tương phản cải thiện chất lượng tạo sinh của các thực thể:** Nhằm cải thiện độ chính xác của các thông tin quan trọng trong tóm tắt, hàm học tương phản được thiết kế để tối ưu hóa việc phân biệt thực thể giữa câu dự đoán và câu tham khảo mẫu dựa vào độ tương quan giữa các mối quan hệ "ngữ nghĩa - ngữ nghĩa", "loại thực thể - loại thực thể", "ngữ nghĩa - loại thực thể".

4.2. Cơ sở lý thuyết

Nghiên cứu về mô hình đa thể thức tập trung vào việc giải quyết bài toán hội tụ đặc trưng từ các bộ mã hóa khác nhau vào một không gian biểu diễn chung. Vì vậy, phương pháp trích xuất, xử lý và tích hợp hiệu quả các đặc trưng đầu vào đóng vai trò quyết định đến khả năng hiểu ngữ cảnh và hiệu suất tổng quát của mô hình.

Phần này chúng tôi sẽ hệ thống hóa các cơ sở lý thuyết cốt lõi cùng những hướng tiếp cận tiên phong mà nghiên cứu kế thừa để xây dựng nên kiến trúc đề xuất, bao gồm các phương pháp biểu diễn bố cục không gian, trích xuất đặc trưng thị giác dạng mảnh và cơ chế tối ưu hóa thông qua hàm mất mát tương phản.

4.2.1. Biểu diễn đặc trưng bố cục

Đối với các tài liệu giàu hình ảnh, thông tin về vị trí đóng vai trò là "ngôn ngữ thị giác" giúp các mô hình xác định cấu trúc logic của tài liệu. Việc mô hình hóa đặc trưng bố cục tập trung vào hai cơ chế chính: nhúng tọa độ không gian và kết hợp tọa độ và ngữ nghĩa tại tọa độ đó.

Cơ chế Nhúng tọa độ không gian (2D spatial position embedding) được đưa ra trong các nghiên cứu [39], [40] đã thiết lập chuẩn mực về việc biểu diễn vị trí dưới dạng các hộp bao $b = (x_1, y_1, x_2, y_2)$.

Theo đó, các giá trị tọa độ này được ánh xạ thành các vector nhúng độc lập cho từng trục. Vector bố cục cuối cùng E_{layout} được hình thành bằng cách tổng hợp các vector thành phần:

$$E_{layout} = E_{x1} + E_{y1} + E_{x2} + E_{y2} + E_w + E_h \quad (2)$$

Để ánh xạ vùng ngữ nghĩa, LaTr [25] định nghĩa đặc trưng của vùng thực thể trong ảnh V_{final} bằng cách kết hợp đặc trưng về vị trí E_{layout} và ngữ nghĩa V_{text}:

$$V_{final} = V_{text} + E_{layout} \quad (3)$$

Sự kết hợp này cho phép mô hình nhận diện được các quan hệ logic như "trên – dưới" hay "trái – phải", từ đó hiểu được các cấu trúc phức tạp của tài liệu, nơi ý nghĩa của một từ phụ thuộc nhiều vào vị trí của nó.

4.2.2. Biểu diễn đặc trưng mảnh

Sự bùng nổ của kiến trúc Vision Transformer [41] đã đưa đặc trưng dạng mảnh (patches feature) trở thành chuẩn mực mới trong thị giác máy tính, thay thế dần các đặc trưng dạng lưới truyền thống. Sức mạnh của loại đặc trưng này đã được minh chứng qua hiệu suất vượt trội của hàng loạt mô hình tiên tiến [42], [43], [44], [45], [46], [47] sử dụng trong tác vụ mô tả hình ảnh.

Giả sử ảnh đầu vào là $I \in R^{H \times W \times C}$, trong đó (H, W) là độ phân giải và C là số kênh màu. Cơ chế phân mảnh và Chiếu tuyến tính được thực hiện qua các bước:

Chia ảnh thành N mảnh nhỏ không chồng lấp kích thước $P \times P$, với:

$$N = (H \times W)/P^2 \quad (4)$$

Mỗi mảnh được trải phẳng thành vector $x_p \in R^{P^2 \cdot C}$ và ánh xạ vào không gian ẩn thông qua ma trận chiếu $E \in R^{(P^2 \cdot C) \times D}$ để tạo thành các visual tokens:

$$z_0 = [x_p^1 E; x_p^2 E; \dots; x_p^N E] + E_{pos} \quad (5)$$

Trong đó E_{pos} là vector nhúng vị trí để bảo toàn cấu trúc không gian.

Sau đó, mỗi token mảnh z_i tương tác với tất cả các token khác thông qua cơ chế self-attention để học được bối cảnh toàn cục.

4.2.3. Hàm mất mát tương phản

Mục tiêu của học tương phản là tối ưu hóa không gian ẩn sao cho biểu diễn của các cặp mẫu có liên quan (positive) nằm gần nhau, đồng thời đẩy xa các cặp không liên quan (negative) với độ đo tương đồng thường được sử dụng là Cosine Similarity.

Trong các nghiên cứu như BRIO [35], CLIFF [34], hàm mất mát tương phản InfoNCE được mô hình hóa như sau:

Với một mẫu truy vấn i, hàm mất mát được định nghĩa:

$$\mathcal{L}_{\mathcal{CL}} = -\sum_{i=1}^{N} \log \frac{\exp(\mathrm{sim}(z_i, z_i^+)/\tau)}{\sum_{j=1}^{M} \exp(\mathrm{sim}(z_i, z_j)/\tau)} \quad (6)$$

Trong đó:

- z_i^+: Biểu diễn của mẫu tích cực.
- z_j: Tập hợp gồm cả mẫu tích cực và các mẫu tiêu cực.
- τ: Tham số nhiệt độ để điều chỉnh độ tập trung của phân phối.

Để ứng dụng hàm mất mát tương phản trong bài toán nhận diện thực thể (NER), BINDER [48] đã thiết lập sự tương quan giữa đặc trưng vùng thực thể và đặc trưng loại thực thể. Thay vì chỉ phân loại, BINDER tối ưu hóa khoảng cách giữa vector vùng văn bản V_e và vector nhãn tương ứng V_t bằng cách:

Kéo gần: Nếu vùng e thuộc loại t, khoảng cách giữa V_e và V_t được thu hẹp.

Đẩy xa: Nếu vùng e không thuộc loại t, chúng bị đẩy xa nhau trong không gian ẩn.

4.3. Trích xuất dữ liệu đầu vào

Để phục vụ cho việc mô hình hóa các giai đoạn trích xuất đặc trưng, chúng tôi thực hiện định nghĩa trước các biến như sau:

- b_n^{lay}: là tọa độ các hộp bao của bố cục.
- $b_n^{lay-trans}$: là tọa độ các hộp bao của bố cục đã được chuyển đổi.
- t_n^{lay}: là nội dung chữ của các bố cục.
- x_n^{t-lay}: là vector đặc trưng ngữ nghĩa của nội dung chữ của các bố cục.
- x_n^{lay}: là vector đặc trưng ngữ nghĩa của các bố cục.
- b_m^{pat}: là tọa độ các mảnh.
- x_m^{pat}: là vector đặc trưng của các mảnh.
- S_{source}: là văn bản đầu vào của mô hình tóm tắt văn bản.

4.3.1. Trích xuất đặc trưng văn bản trong hình ảnh (OCR)

Các cụm thông tin trong infographic có liên kết chặt chẽ về mặt ngữ nghĩa thường được trình bày trong các khối hình học xác định (hình chữ nhật, hình tròn, đa giác...), tạo nên cấu trúc bố cục của tài liệu. Quy trình trích xuất OCR truyền thống thực thi theo thứ tự từ trái-phải-trên-dưới. Cách tiếp cận này làm ảnh hưởng tính liên kết của các đoạn văn bản có cùng nội dung, gây nhiễu đầu vào và làm suy giảm khả năng hiểu ngữ cảnh của các mô hình ngôn ngữ. Nhóm sử dụng quy trình trích xuất đặc trưng ba bước:

- **Phát hiện bố cục:** Sử dụng những mô hình nhận diện bố cục Layout Detection (LDT) để tìm được vùng bố cục có trong ảnh:

$$\{b_n^{lay}\}_{\{n=1:N\}} = LDT(I) \quad (7)$$

- **Nhận diện văn bản**: Sử dụng mô hình nhận diện văn bản Text Detection (TDT) để trích xuất nội dung chữ có trong các vùng bố cục:

$$\{t_n^{lay}\} = TDT(\{b_n^{lay}\}), \{n = 1:N\} \quad (8)$$

- **Hợp nhất văn bản:** Kết nối tất cả nội dung nội dung từ các vùng theo thứ tự logic từ trái-phải-trên-dưới, tạo ra một chuỗi văn bản đầu vào có tính mạch lạc, bảo toàn được dòng chảy thông tin và ý nghĩa nguyên bản của infographic.

$$S_{source} = Concat\left(\{ t_n^{lay} \}_{\{n=1:N\}} | \text{ sort by } (y, x) \right) \quad (9)$$

4.3.2. Trích xuất đặc trưng mảnh

Hình ảnh đầu vào được tiền xử lý về kích thước chuẩn 224×224, sau đó đi qua lớp patch_embeddings c ủa mô hình LayoutLMv3 với kích thước mảnh được thiết lập là 16, chia bức ảnh thành một lưới 14×14 (tương ứng với 196 patches):

$$\left\{ \{x_m^{pat}\}_{\{m=1:M\}}; \{b_m^{pat}\}_{\{m=1:M\}} \right\} = LayoutLMv3(I) \quad (10)$$

4.3.3. Trích xuất đặc trưng bố cục

Tọa độ của các bố cục được định nghĩa với cấu trúc: $[x_{min}, y_{min}, x_{max}, y_{max}]$ và được chuẩn hóa theo chiều dài và chiều cao của ảnh. Nhóm thực hiện biến đổi cấu trúc tọa độ này với dạng $b_n^{lay-trans}$, với $[.;.]$ là phép nối.

$$b_n^{lay-trans} = [x_{tl}; y_{tl}; x_{tr}; y_{tr}; x_{bl}; y_{bl}; x_{br}; y_{br}; w; h] \quad (11)$$

Trong đó:

- x_{tl}, y_{tl}: là tọa độ trên cùng bên trái
- x_{tr}, y_{tr}: là tọa độ trên cùng bên phải
- x_{bl}, y_{bl}: là tọa độ dưới cùng bên trái
- x_{br}, y_{br}: là tọa độ dưới cùng bên phải
- w, h: lần lượt là chiều rộng và chiều dài của các bố cục

Các tọa độ này được nhúng sử dụng 2D Spatial Position Embedding (sử dụng lớp nn.Embedding của Pytorch), kết hợp với đặc trưng ngữ nghĩa của nội dung văn bản bên trong bố cục bằng công thức sau:

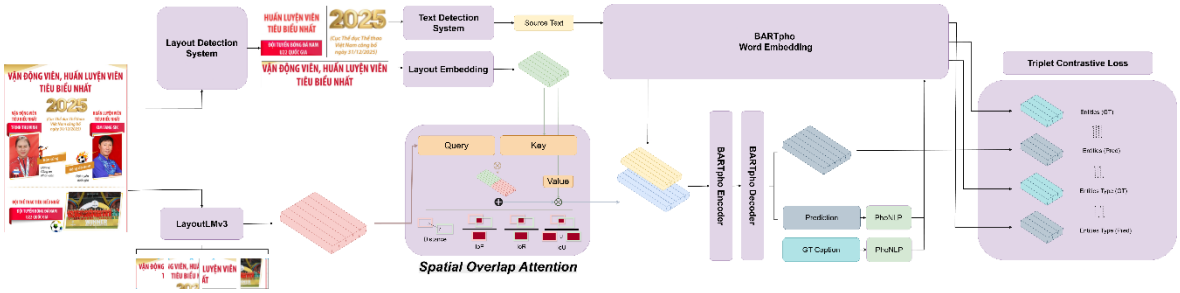
$$x_n^{lay} = x_n^{t-lay} + \text{PosEmbedding}(b_n^{lay-trans}).\text{mean}() \quad (12)$$

4.4. Trích xuất đặc trưng cho văn bản chú thích

Chúng tôi sử dụng PhoBERT [49] để trích xuất đặc trưng cho văn bản chú thích. PhoBERT là mô hình ngôn ngữ đơn ngữ quy mô lớn đầu tiên được tiền huấn luyện dành riêng cho tiếng Việt, bao gồm hai phiên bản chính là $\text{PhoBERT}_{\text{base}}$ và $\text{PhoBERT}_{\text{large}}$. Được xây dựng dựa trên kiến trúc RoBERTa [50], PhoBERT tối ưu hóa quy trình huấn luyện của BERT nhằm mang lại hiệu suất ổn định và mạnh mẽ hơn cho các tác vụ ngôn ngữ. Mô hình được huấn luyện trên bộ dữ liệu tiếng Việt khổng lồ lên đến 20GB, áp dụng cơ chế phân mảnh từ để giải quyết triệt để sự nhập nhằng giữa âm tiết và từ vựng đặc thù của tiếng Việt. Dựa trên kết quả thực nghiệm, PhoBERT [49] cho hiệu suất SOTA trên các tác vụ NLP quan trọng như gán nhãn từ loại, nhận dạng thực thể có tên, suy diễn ngôn ngữ tự nhiên, mang đến hiệu suất vượt trội so với các mô hình đa ngôn ngữ tổng quát.

4.5. Mô hình đề xuất

Hình 4.1 là minh họa về kiến trúc mô hình đề xuất. Mô hình đề xuất của chúng tôi dựa trên kiến trúc cở bản của mô hình tóm tắt văn bản trừu tượng, dựa theo kiến trúc của mô hình $\text{BARTpho}_{\text{syllable}}$ [30], bao gồm những mô-đun chính như sau:



Hình 4.1: Kiến trúc mô hình đề xuất

4.5.1. SpatialOverlapAttention

Infographic đặc trưng với các bố cục lớn, chứa nhiều thông tin, trải dài trên nhiều vùng ảnh. Vì thế, nhóm đề xuất một cơ chế có thể biểu diễn được vị trí của các hộp trên không gian của Infographic bằng cơ chế IoRs (Intersection Over Regions). IoRs biểu diễn tỷ lệ trùng khớp của điểm giao (Intersection) trên 3 vùng (Regions) khác nhau như sau:

- IoU (Intersection Over Union): Tỷ lệ trùng khớp trên vùng hợp
- IoP (Intersection Over Patch): Tỷ lệ trùng khớp trên vùng Patch
- IoL (Intersection Over Layout): Tỷ lệ trùng khớp trên vùng Layout

Lấy cảm hứng từ mô hình SaL [51], các tỷ lệ trùng khớp được nhúng vào một bảng tìm kiếm (LookupTable - sử dụng nn.Embedding) có kích thước 32×16, với 16 là số lượng số lượng attention heads.

Đầu tiên, xét tọa độ của một bố cục $\mathrm{b}_{\mathrm{n}}^{\mathrm{lay}}$ và tọa độ của một mảnh $\mathrm{b}_{\mathrm{m}}^{\mathrm{pat}}$, IoRs được tính toán bằng công thức sau, với Area là hàm tính diện tích:

$$\mathrm{IoU}_{\mathrm{nm}} = \frac{\mathrm{Area}\left(\mathrm{b}_{\mathrm{n}}^{\mathrm{lay}} \cap \mathrm{b}_{\mathrm{m}}^{\mathrm{pat}}\right)}{\mathrm{Area}\left(\mathrm{b}_{\mathrm{n}}^{\mathrm{lay}} \cup \mathrm{b}_{\mathrm{m}}^{\mathrm{pat}}\right)} \tag{13}$$

$$\mathrm{IoP}_{\mathrm{nm}} = \frac{\mathrm{Area}\left(\mathrm{b}_{\mathrm{n}}^{\mathrm{lay}} \cap \mathrm{b}_{\mathrm{m}}^{\mathrm{pat}}\right)}{\mathrm{Area}\left(\mathrm{b}_{\mathrm{m}}^{\mathrm{pat}}\right)} \tag{14}$$

$$\mathrm{IoL}_{\mathrm{nm}} = \frac{\mathrm{Area}\left(\mathrm{b}_{\mathrm{n}}^{\mathrm{lay}} \cap \mathrm{b}_{\mathrm{m}}^{\mathrm{pat}}\right)}{\mathrm{Area}\left(\mathrm{b}_{\mathrm{n}}^{\mathrm{lay}}\right)} \tag{15}$$

$$\mathrm{IoRs}_{\mathrm{nm}} = [\mathrm{IoU}_{\mathrm{nm}}; \mathrm{IoP}_{\mathrm{nm}}; \mathrm{IoL}_{\mathrm{nm}}] \tag{16}$$

Với $\mathrm{IoRs}_{\mathrm{nm}} \in [0, 1]$, ta thực hiện phép biến đổi:

$$\phi = (\mathrm{IoRs}_{\mathrm{nm}} * 31).\mathrm{long}() \tag{17}$$

để ánh xạ giá trị chồng lấp lên không gian giá trị đầu vào của LookupTable.

Đặc trưng của từng mảnh được tăng cường bằng cách bổ sung thông tin vị trí không gian tương ứng. Cụ thể, mô hình LayoutLMv3 trích xuất đặc trưng từ 196 mảnh, tương ứng với lưới 14×14 theo hai chiều ngang và dọc của ảnh. Vị trí của

mỗi mảnh được biểu diễn thông qua cặp tọa độ (x, y), trong đó mỗi trục tọa độ được ánh xạ độc lập vào một không gian embedding bằng các lớp nn.Embedding

$$x_x^{pat-pos} = nn.\,Embedding\left(p_x^{pat-pos}\right) \quad (18)$$

$$x_y^{pat-pos} = nn.\,Embedding\left(p_y^{pat-pos}\right) \quad (19)$$

Cuối cùng, đặc trưng mảnh sau khi được tăng cường vị trí được tính bằng cách cộng đặc trưng thị giác ban đầu với các embedding vị trí theo hai trục, sau khi chuẩn hóa bằng Layer Normalization:

$$x^{pat+} = x^{pat} + LN\left(x_x^{pat-pos}\right) + LN\left(x_y^{pat-pos}\right) \quad (20)$$

SpatialOverlapAttention cho phép mô hình học được sự tương quan của bố cục dựa vào vị trí chồng lấp của chúng lên các phân mảnh, được mô tả như sau:

$$\begin{cases} Q = x^{pat} W_Q \\ K = x^{lay} W_K \\ V = x^{lay} W_V \end{cases} \quad (21)$$

Với query $Q \in \mathbb{R}^{M\times d}$, key $K \in \mathbb{R}^{N\times d}$ và value $V \in \mathbb{R}^{N\times d}$, W_Q, W_K, W_V là các tham số học được. Ma trận trọng số A được tính như sau:

$$A = \frac{QK^T}{\sqrt{d}} \quad (22)$$

Với $A \in \mathbb{R}^{M\times N}$, nhằm học được mối tương quan về vị trí, các giá trị thuộc $IoRs_{mn}$ được biến đổi theo công thức sau:

$$IoUattn_{mn} = LookupTable\left(\phi(IoU_{mn})\right) \quad (23)$$

$$IoPattn_{mn} = LookupTable\left(\phi(IoP_{mn})\right) \quad (24)$$

$$IoLattn_{mn} = LookupTable\left(\phi(IoL_{mn})\right) \quad (25)$$

$$IoRsattn_{mn} = \frac{(IoUattn_{mn} + IoPattn_{mn} + IoLattn_{mn})}{3} \quad (26)$$

Để giảm thiểu ảnh hưởng từ các đặc trưng mảnh không liên quan, nhóm tính toán khoảng cách Euclid giữa các mảnh và các bố cục xuất hiện trong ảnh. Cơ chế này giúp mô hình tập trung sự chú ý vào các mảnh phân bố gần bố cục, đồng thời loại bỏ nhiễu từ các mảnh biên ở xa.

$$D_{mn} = -\log\left(\text{EuclidDistance}\left(b_n^{lay}, b_m^{pat}\right)\right) \quad (27)$$

Với $\text{IoRs\_attn}, D \in \mathbb{R}^{M\times N}$. Sau cùng, đầu ra được cộng với kết nối tắt (residual connection) từ đặc trưng mảnh gốc để bảo toàn thông tin thị giác.

$$x^{patoverlap} = x^{pat} + \text{softmax}(A + \text{IoRsattn} + D)V \quad (28)$$

Nhờ được bổ sung thông tin ngữ nghĩa, $x^{patoverlap}$ biểu diễn mối quan hệ chi tiết giữa các bố cục xuất hiện trong hình ảnh mà vẫn giữ nguyên được đặc trưng thị giác của các mảnh ban đầu.

4.5.2. Triplet Contrastive Loss

Lấy cảm hứng từ nghiên cứu BINDER [48] của Microsoft cho cho bài toán Nhận dạng thực thể tên riêng (NER), Triplet Contrastive Loss được thiết kế để tối ưu hóa đồng thời ba mối quan hệ cốt lõi:

- "Ngữ nghĩa - Ngữ nghĩa" ($\text{Word} - \text{Word}$) : Mục tiêu gia tăng độ tương đồng về nội dung giữa thực thể dự đoán và thực thể đúng.
- "Loại thực thể - Loại thực thể" ($\text{TagType} - \text{TagType}$): Mục tiêu đảm bảo sự đồng nhất về loại thực thể.
- "Ngữ nghĩa - Loại thực thể" ($\text{Word}-\text{TagType}$): Mục tiêu gia tăng độ tương đồng mối quan hệ giữa loại thực thể và ngữ nghĩa của nó.

Đầu tiên, danh sách các thực thể Word và loại thực thể TagType sẽ được nối lại bằng khoảng trắng. Tận dụng sức mạnh của BARTpho[30], lớp embed_tokens $\text{BARTpho}^{\text{Encoder}}$, được sử dụng để mã hóa chuỗi thực thể xuất hiện trong tóm tắt tham chiếu và loại thực thể xuất hiện trong tóm tắt tham chiếu thành vector biểu diễn đặc trưng.

$$\text{Wordseq}_{sum}^{e} = \text{BARTpho}^{\text{Encoder}}(\text{Wordseq}_{sum}) \quad (29)$$

$$\text{Tagtypeseq}_{sum}^{e} = \text{BARTpho}^{\text{Encoder}}(\text{Tagtypeseq}_{sum}) \quad (30)$$

Nhóm sử dụng PhoNLP [52] kết hợp với VnCoreNLP [36] để trích xuất thực thể và loại thực thể của câu dự đoán. Đối với các chuỗi thực thể xuất hiện trong câu dự đoán, nhóm sử dụng vector đặc trưng của từng token thực thể trích xuất từ

lớp $\text{BARTpho}^{\text{Decoder}}$. Với loại thực thể xuất hiện trong câu dự đoán, lớp embed_tokens $\text{BARTpho}^{\text{Encoder}}$ được sử dụng để mã hóa chuỗi loại thực thể.

$$\text{Wordseq}^{\text{e}}_{\text{pred}} = \text{BARTpho}^{\text{Decoder-ht}}(\text{Wordseq}_{\text{pred}}) \quad (31)$$

$$\text{TagTypeseq}^{\text{e}}_{\text{pred}} = \text{BARTpho}^{\text{Encoder}}(\text{TagTypeseq}_{\text{pred}}) \quad (32)$$

Do đặc trưng âm tiết tiếng Việt, các danh từ riêng, thông tin định lượng thường được cấu thành từ 2-3 âm tiết, do đó các token đơn lẻ có thể bị chia cắt về mặt ngữ nghĩa. Để từng token trong vector đặc trưng hiểu được ngữ nghĩa của các âm tiết phía trước và sau nó, nhóm sử dụng một lớp mạng Conv1d có $\text{kernel\_size} = 3$ lặp qua chuỗi vector biểu diễn nhằm làm giàu ngữ nghĩa cho từng token.

$$* \text{seq}^{\text{e+}}_{\text{pred}} = \text{Conv1d}(* \text{seq}^{\text{e}}_{\text{pred}}) \quad (33)$$

$$* \text{seq}^{\text{e+}}_{\text{sum}} = \text{Conv1d}(* \text{seq}^{\text{e}}_{\text{sum}}) \quad (34)$$

$$* \text{seq}^{\text{e}}_{\text{pred}} = \{\text{Wordseq}^{\text{e}}_{\text{pred}}, \text{TagTypeseq}^{\text{e}}_{\text{pred}}\} \quad (35)$$

$$* \text{seq}^{\text{e}}_{\text{sum}} = \{\text{Wordseq}^{\text{e}}_{\text{sum}}, \text{TagTypeseq}^{\text{e}}_{\text{sum}}\} \quad (36)$$

Độ tương đồng của hai vector được định nghĩa bằng phép tích vô hướng của chúng, được định nghĩa bằng:

$$\text{sim}(\text{x}, \text{y}) = \text{Matmul}(\text{x}, \text{y}^{\text{T}}) \quad (37)$$

Để đảm bảo mô hình không chỉ tạo sinh ra các câu văn lưu loát mà còn phải chính xác tuyệt đối về mặt thứ tự thông tin thực thể, dựa vào ý tưởng của BINDER [48], mục tiêu huấn luyện của nhóm là: tại mỗi vị trí thực thể i trong câu tóm tắt tham chiếu, mô hình phải học cách cực đại hóa độ tương đồng với một nhóm các syllable/token trong khoảng $[\text{i} - \text{k}, \text{i} + \text{k}]$ tương ứng trong câu dự đoán, với k là khoảng đại diện âm tiết. Với cách thiết kế này, mô hình có thể cực đại hóa độ tương đồng của các thực thể ở câu mô tả tham chiếu với câu dự đoán thông qua cực đại hóa độ tương đồng giữa các âm tiết của chúng. Hơn nữa, điều này vô hình chung buộc mô hình phải tạo sinh ra thứ tự các thực thể đúng với thứ tự xuất hiện của các thực thể có trong câu mô tả tham chiếu, đảm bảo về trình tự

và tính chính xác của các thực thể khi được tạo sinh. Nhóm đề ra các mối quan hệ như sau:

- $\mathrm{Word}-\mathrm{Word}$: Mối quan hệ $\mathrm{w}-\mathrm{w}$ mô hình hóa sự tương đồng của các từ về mặt ngữ nghĩa. Các từ có ý nghĩa về mặt nội dung ngữ nghĩa giống nhau sẽ có độ tương đồng cao.
- $\mathrm{Type}-\mathrm{Type}$: Mối quan hệ $\mathrm{t}-\mathrm{t}$ mô hình hóa sự tương đồng của các loại thực thể. Các từ có loại thực thể giống nhau sẽ có độ tương đồng cao.
- $\mathrm{Type}-\mathrm{Word}$: Mối quan hệ $\mathrm{t}-\mathrm{w}$ mô hình hóa sự tương đồng của các từ và loại thực thể của chúng. Nếu ý nghĩa của từ có liên quan đến loại thực thể được đề cập sẽ có độ tương đồng cao.

Nhóm đặt ra giả thuyết rằng, khi các thực thể được dự đoán có sự liên quan mạnh về quan hệ $\mathrm{TagType}-\mathrm{Word}$ với các loại thực thể trong câu tóm tắt tham chiếu cho thấy các từ và nội dung xung quanh được dự đoán sẽ giống về mặt ngữ pháp và cấu trúc câu từ.

Nhóm xây dựng nhóm hàm mất mát tích cực cho từng cặp quan hệ như sau:

$$\mathcal{L}_{\mathrm{Word-Word}} = -\frac{1}{|\mathrm{P_i}|}\sum_{\mathrm{i=1}}^{\mathrm{N}} \log\left(\frac{\Sigma_{\mathrm{j}\in\mathrm{P_i}} \exp\left(\mathrm{sim}\left(\mathrm{w_i^S},\ \mathrm{w_j^P}\right)\right)}{\Sigma_{\mathrm{k}\in\mathrm{A_i}} \exp\left(\mathrm{sim}\left(\mathrm{w_i^S},\ \mathrm{w_k^P}\right)\right)}\right) \tag{38}$$

$$\mathcal{L}_{\mathrm{Type-Type}} = -\frac{1}{|\mathrm{P_i}|}\sum_{\mathrm{i=1}}^{\mathrm{N}} \log\left(\frac{\Sigma_{\mathrm{j}\in\mathrm{P_i}} \exp\left(\mathrm{sim}\left(\mathrm{t_i^S},\ \mathrm{t_j^P}\right)\right)}{\Sigma_{\mathrm{k}\in\mathrm{A_i}} \exp\left(\mathrm{sim}\left(\mathrm{t_i^S},\ \mathrm{t_k^P}\right)\right)}\right) \tag{39}$$

$$\mathcal{L}_{\mathrm{TagType-Word}} = -\frac{1}{|\mathrm{P_i}|}\sum_{\mathrm{i=1}}^{\mathrm{N}} \log\left(\frac{\Sigma_{\mathrm{j}\in\mathrm{P_i}} \exp\left(\mathrm{sim}\left(\mathrm{t_i^S},\ \mathrm{w_j^P}\right)\right)}{\Sigma_{\mathrm{k}\in\mathrm{A_i}} \exp\left(\mathrm{sim}\left(\mathrm{t_i^S},\ \mathrm{w_k^P}\right)\right)}\right) \tag{40}$$

Với:

$\mathcal{L}_{\text{Word}-\text{Word}}$ thể hiện độ tương đồng của các thực thể trong câu dự đoán và các thực thể trong câu tóm tắt tham chiếu;

$\mathcal{L}_{\text{TagType}-\text{TagType}}$ thể hiện độ tương đồng của các tên thực thể trong câu dự đoán và tên thực thể trong câu tóm tắt tham chiếu;

$\mathcal{L}_{\text{Word}-\text{TagType}}$ thể hiện độ tương đồng của các loại thực thể trong câu tóm tắt tham chiếu và tên thực thể trong câu dự đoán.

Trong đó:

- P_i là tập hợp các vị trí tích cực.
- A_i là tập hợp các vị trí tích cực và tiêu cực.

Cuối cùng, hàm mất mát tổng được tính toán bằng cách tổng hợp các hàm mất mát tương phản của các mối quan hệ:

$$\mathcal{L}_{\text{Contrastive}} = \beta\mathcal{L}_{\text{Word}-\text{Word}} + \gamma\mathcal{L}_{\text{Type}-\text{Type}} + \delta\mathcal{L}_{\text{TagType}-\text{Word}} \quad (41)$$

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{Contrastive}} \quad (42)$$

Trong đó β, γ, δ, α là các tham số học được. Đặc biệt, α đóng vai trò như tham số giới hạn mức độ ảnh hưởng của hàm mất mát tương phản lên quá trình học biểu diễn ngữ nghĩa của mô hình thông qua hàm mất mát Cross-Entropy. Cách thiết kế này giúp tránh tình trạng mô hình quá tập trung vào việc tối ưu hóa khả năng tạo sinh các thực thể, dẫn đến suy giảm độ lưu loát và tự nhiên của câu sinh ra.

4.5.3. Mô hình encoder-decoder

Nhóm sử dụng BARTpho [30] để tổng hợp các nguồn thông tin đầu vào. Tại giai đoạn này, vector ngữ nghĩa của đoạn văn bản đầu vào được nối trực tiếp với đặc trưng mảnh được tăng cường, tạo nên một biểu diễn đa thể thức toàn diện, bao hàm cả nội dung ngữ nghĩa văn bản lẫn cấu trúc không gian thị giác:

$$\text{MMT}_{\text{Input}} = \left[\text{BARTpho}^{\text{E}}(S_{\text{source}});\ x^{\text{pat}_o\text{verlap}}\right] \quad (43)$$

$$C = \text{BARTpho}\left(\text{MMT}_{\text{Input}}\right) \quad (44)$$

Trong đó:

- BARTpho: là mô hình tóm tắt.
- C: là mô tả dự đoán, đầu ra của mô hình.
- MMT_{Input}: là đầu vào của mô hình.

Chương 5. CÁC PHƯƠNG PHÁP THỰC NGHIỆM

5.1. Các mô hình cơ sở cho bài toán Chú thích hình ảnh

Hiện nay, trong bài toán chú thích hình ảnh (Image Captioning), các nghiên cứu trên thế giới thường tập trung khai thác đặc trưng dạng vùng (region features) hoặc dạng lưới (grid features) từ mạng CNN để tạo sinh câu mô tả, tiêu biểu là các công trình như [18], [19]. Những phương pháp tiếp cận này thường bộc lộ hạn chế khi xử lý các hình ảnh chứa văn bản (scene text). Điều này đã thúc đẩy sự ra đời của dòng nghiên cứu mô tả hình ảnh dựa trên OCR (OCR-based Image Captioning) với các kiến trúc tiêu biểu như [13], [22], [53],[23].

5.1.1. Trích xuất đặc trưng cho các mô hình Chú thích hình ảnh

Đối với hướng tiếp cận Image Captioning trên bộ dữ liệu, YOLOv8 [54] được sử dụng để phát hiện đối tượng, SwinTextSpotter [55] được sử dụng để trích xuất các từ trong ảnh.

Cụ thể, đối với đặc trưng vật thể trong hình ảnh, chúng tôi sử dụng YoloV8 [54] để trích xuất. YOLOv8 được phát triển bởi Ultralytics, được cải thiện chất lượng cho tác vụ trích xuất đặc trưng và nhận dạng thực thể. Nhóm sử dụng mô hình YOLOv8 được tiền huấn luyện trên bộ dữ liệu OpenImageV[39]7, một tập dữ liệu lớn cho bài toán nhận diện vật thể (object detection) lớn với hơn 600+ lớp, giúp đa dạng hơn các đối tượng nhận diện được, so với FasterRCNN [56] tiền huấn luyện trên bộ dữ liệu MS-COCO [10] chỉ với 80 lớp.

Các đặc trưng văn bản trong hình ảnh được trích xuất bằng mô hình SwinTextSpotter [55]. Trong khi các mô hình OCR hiện tại cho tiếng Việt thường kết hợp việc sử dụng các mô hình, theo quy trình hai giai đoạn tách biệt, bao gồm phát hiện vùng chữ và nhận diện nội dung văn bản, cách tiếp cận này làm mang đến nguy cơ bị sai lệch thông tin, do đặc trưng về các dấu thanh nhỏ có thể bị mất mát trong quá trình chuyển tiếp giữa hai giai đoạn. Để khắc phục, SwinTextSpotter [55] là mô hình end-to-end, thống nhất việc phát hiện và nhận

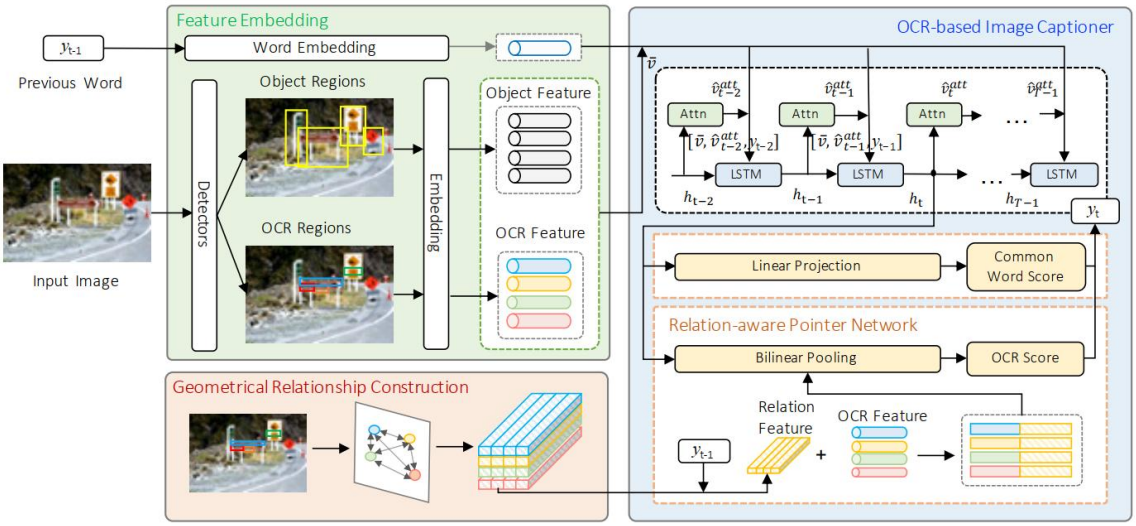
diện vào một luồng duy nhất. Thực nghiệm cho thấy, SwinTextSpotter [55] đạt hiệu suất cao hơn đáng kể so với các phương pháp ghép nối trên các tập dữ liệu đa ngôn ngữ, bao gồm cả tiếng Việt. Do đó, SwinTextSpotter [55] là lựa chọn phù hợp để trích xuất đặc trưng văn bản cho bài toán mô tả hình ảnh.

5.1.2. Mô hình LSTM-R

Những vấn đề hiện có của các mô hình mô tả hình ảnh hiện nay là khó khăn trong việc tìm hiểu và sắp xếp đúng vị trí của các OCR tokens được trích xuất từ hình ảnh. Các phương pháp hiện tại mã hóa OCR tokens bằng các đặc trưng hình ảnh và ngữ nghĩa của hình ảnh đầu vào, tuy nhiên, mối liên hệ trong không gian giữa các OCR tokens với nhau không thể được biểu diễn hoàn toàn bởi những đặc trưng tổng quát đó, dẫn đến việc sắp xếp và mô tả các OCR tokens không chính xác. Để giải quyết vấn đề này, mô hình LSTM-R[57] sử dụng các đặc trưng hình học trong không gian 2D gọi là Geometrical Relationship để mô tả mối quan hệ của các OCR tokens với nhau, sử dụng những thông tin liên quan tới:

- Chiều dài
- Chiều rộng
- Khoảng cách giữa các tokens
- Độ trùng lắp IoU (Intersection over Union)
- Góc giữa các OCR Tokens

Các thông tin này giúp xác định mối quan hệ thực sự giữa các token, đặc biệt là khi các token gần nhau về không gian nhưng không liên quan về ngữ nghĩa. Trên cơ sở ý tưởng đó, LSTM-R (LSTM + Relation-aware pointer network), nơi mạng Pointer Aware mối quan hệ hình học được dùng để lựa chọ OCR Token phù hợp cho quá trình tạo sinh chú thích, kết hợp đặc trưng hình ảnh và mối quan hệ trong không gian, cho ra kết quả vượt xa các phương pháp trước.



Hình 5.1: Kiến trúc mô hình LSTM-R [57]

Kiến trúc của mô hình (*Hình 5.1)* gồm 3 thành phần chính:

- **Feature Embedding:** Tại đây, các đặc trưng thực thể (*Object*) và chữ (*Scene Texts*) được trích xuất từ các mô hình tiền huấn luyện trích xuất đặc trưng sẽ được lần lượt mã hóa. *Embedding of Objects* sẽ được ánh xạ vào không gian có chiều bằng với số chiều của mô hình ngôn ngữ LSTM bằng phép biến đổi tuyến tính. Song song với đó, *Embedding of OCR tokens* được mã hóa bằng cách kết hợp thêm features của FastText và Pyramidal Histogram of Characters (PHOC), được mô tả bằng công thức sau:

$$x_m^{obj} = \sigma\left(LN\left(W_1 L_2 N\left(x_m^{obj-a}\right)\right)\right) \tag{45}$$

$$x_m^{ocr} = \sigma\left(LN(W_2 x_n^{main})\right) + \sigma\left(LN(W_3 x_n^{bb})\right) \tag{46}$$

$$x_n^{main} = \left[L_2 N(x_n^{ocr-a}); L_2 N\left(x_n^{ft}\right); L_2 N\left(x_n^{phoc}\right)\right] \tag{47}$$

- **Geometrical Relationship Construction:** Nhằm tối ưu hóa khả năng thấu hiểu ngữ cảnh không gian, phương pháp tập trung vào cải thiện mối quan hệ hình học giữa các đối tượng với nhau. Với mỗi cặp OCR tokens, các đặc trưng về kích thước, không gian, khoảng cách, độ chồng lắp (IoU) và góc giữa các đối tượng được kết hợp tạo thành vector mã hóa mối quan hệ không

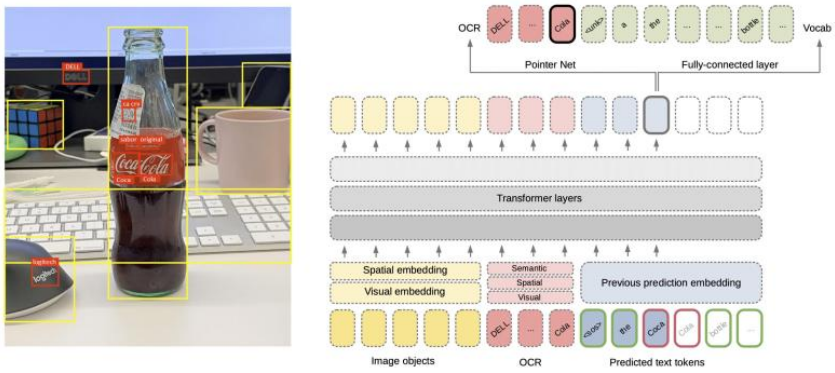
gian giữa các đối tượng. Vector này được ánh xạ bằng phép biến đổi tuyến tính với số chiều bằng số chiều của mô hình ngôn ngữ, tạo ra một biểu diễn đặc trưng giàu ngữ nghĩa, giúp mô hình nắm bắt trọn vẹn cấu trúc và bố cục phức tạp của dữ liệu.

- **OCR-based Image Captioner:** Mô-đun này được xây dựng dựa trên nền tảng kiến trúc LSTM [58], kết hợp với cơ chế Attention qua các lớp, mô hình hóa khả năng sao chép trực tiếp các token OCR từ hình ảnh qua mạng Pointer Network [20]. Tại mỗi bước thời gian, quá trình tổng hợp thông tin diễn ra thông qua việc cập nhật trạng thái ẩn của LSTM dựa trên đầu vào là đặc trưng thị giác và từ đã được tính toán trước đó. Để xác định từ tiếp theo, mô hình thực hiện so sánh xác suất xuất hiện giữa các từ trong bộ từ điển chung và các token OCR. Điểm cải tiến cốt lõi nằm ở cơ chế Pointer Network: hệ thống sẽ xem xét token liền trước; nếu đó là một token OCR, vector quan hệ hình học tương ứng giữa nó và các token ứng viên sẽ được trích xuất để làm giàu ngữ cảnh không gian; ngược lại, nếu từ liền trước là từ vựng thông thường, vector quan hệ này được gán giá trị không. Các đặc trưng hình học này sau đó được ghép nối với đặc trưng nội dung của OCR token để tạo thành đầu vào cho việc tính toán điểm số sao chép, giúp mô hình đưa ra quyết định tối ưu về việc sinh từ mới hay trích xuất nội dung văn bản có sẵn trong ảnh.

5.1.3. Mô hình M4C-Captioner

Multimodal Multi-Copy Mesh (M4C) [21] là mô hình tiên phong trong việc giải quyết bài toán VQA dưới dạng sinh câu trả lời. Mô hình này hoạt động dựa trên việc nhúng các dữ liệu đa thể thức (hình ảnh và văn bản) vào một không gian ngữ nghĩa chung và xử lý chúng thông qua kiến trúc Multimodal Transformer. M4C [21] nổi bật nằm ở khả năng giải mã câu trả lời lặp lại kết hợp mạng trỏ động (Dynamic Pointer Network). Cơ chế này cho phép mô hình không chỉ sinh ra các từ trong tập từ điển cố định, mà còn giải quyết vấn đề liên quan đến nằm ngoài bộ từ điển (Out of Vocabulary – OOV) bằng tính chất sao chép trực tiếp

OCR token có trong ảnh. M4C-Captioner [13] được xây dựng dựa trên nền tảng của M4C [21], M4C-Captioner được tinh chỉnh bằng cách loại bỏ đầu vào câu hỏi và tận dụng bộ giải mã đa từ (multi-word decoder) để trực tiếp sinh câu mô tả dựa trên các thực thể thị giác và dữ liệu văn bản được trích xuất.



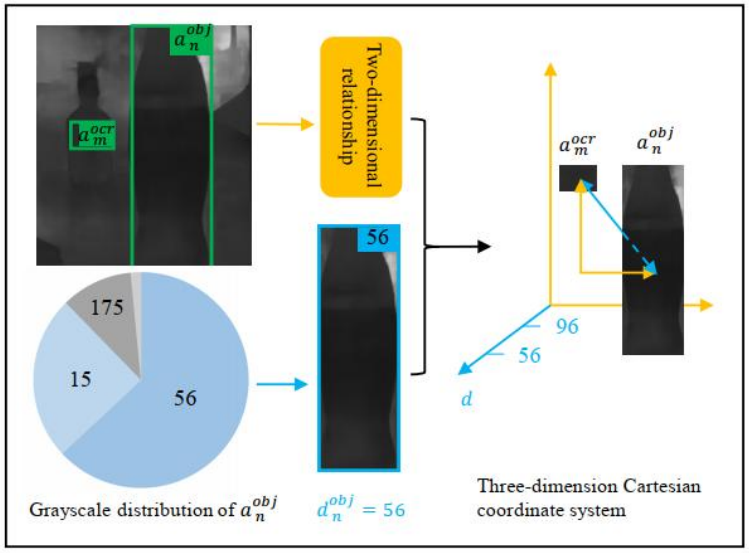
Hình 5.2: Kiến trúc mô hình M4C-Captioner [13]

5.1.4. Mô hình DEVICE

Các nghiên cứu trước đây về OCR-based Image Captioning thường chỉ tập trung vào việc xây dựng mối quan hệ hình học phẳng (2D geometric relationship) giữa các OCR tokens, mà thiếu đi thông tin về chiều sâu của vật thể (depth information), khi các hình ảnh được chụp trong không gian 3D. Để khắc phục hạn chế này, Depth and Visual Concepts Aware Transformer (DEVICE) [22] được đề xuất, tập trung xây dựng mối quan hệ trong không gian ba chiều (three-dimensional geometric relations), giúp mô hình hiểu được mối liên hệ không gian thực giữa các thực thể trong ảnh với nhau. Kiến trúc của mô hình bao gồm 4 mô-đun chính:

- **Multimodal Embedding**: Tại mô-đun này, các đặc trưng thị giác của vật thể và văn bản trong ảnh được trích xuất và nhúng vào không gian vector chung. Tại đây, tác giả đề xuất sử dụng thêm giá trị độ sâu để mô hình hiểu hơn về mối quan hệ không gian ba chiều của các thực thể trong ảnh. Tác giả sử dụng mô hình ước lượng độ sâu BTS [59] để tạo ra bản đồ cho toàn bộ ảnh. Giá trị độ sâu của từng thực thể được xác định bằng cách lấy giá trị

thang độ xám có tần suất xuất hiện cao nhất trong vùng bounding box của thực thể đó. Giá trị độ sâu này được chuẩn hóa và kết hợp với tọa độ 2D của bounding box để tạo thành tọa độ hình học 5 chiều (bao gồm trục *z* ảo), giúp đo lường khoảng cách tương đối của thực thể với người quan sát.

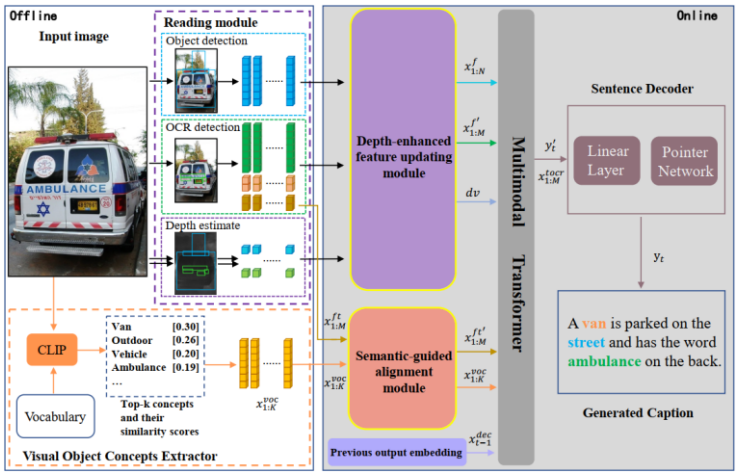


Hình 5.3: Mô tả mô đun DEFUM trong mô hình DEVICE [22]

- **Depth-enhanced Feature Updating Module (DEFUM) (*Hình 5.3)*:** Khác với bài toán VQA thường chú trọng quan hệ giữa các vật thể, bài toán này đòi hỏi khả năng mô hình hóa mối quan hệ giữa các OCR tokens thật mạnh mẽ. DeFUM được thiết kế để cập nhật và làm giàu đặc trưng của OCR tokens bằng sự kết hợp của thông tin độ sâu. Mô-đun sử dụng Relative Depth Map kết hợp với cơ chế Depth-aware Self-Attention. Ma trận Attention được cộng thêm một ma trận trọng số độ sâu R, giúp mô hình phân biệt và giảm sự chú ý giữa các thực thể có chênh lệch độ sâu quá lớn, đồng thời tăng cường kết nối giữa các thực thể nằm trên cùng một mặt phẳng độ sâu.
- **Sematic-guided Alignment Module**: Để mô hình không bỏ sót thông tin của các đối tượng quan trọng trong hình, gây ra câu trả lời kém chất lượng. Mô-đun sử dụng sức mạnh của mô hình **CLIP** [60] để lấy được top 15 thực thể liên quan tới hình nhất. Sau đó, nhúng các vector ngữ nghĩa của các thực thể lên một không gian chung có số chiều cùng với số chiều của

mô hình ngôn ngữ. Thay vì chỉ nhúng đơn thuần, mô-đun SgAM sử dụng cơ chế Semantic Attention. Nó coi đặc trưng ngữ nghĩa của OCR tokens là Query và đặc trưng của Visual Concepts là Key/Value. Quá trình này giúp tích hợp thông tin của các khái niệm thị giác vào các OCR tokens có ý nghĩa tương, giúp mô hình hiểu ngữ cảnh tốt hơn.

- **Reasoning and Generation Module (*Hình 5.4)*:** Một multi-modal transformers được sử dụng để mã hóa toàn bộ các đặc trưng đầu vào. DEVICE được xây dựng trên nền tảng của M4C-Captioner, Quá trình sinh câu sử dụng Dynamic Pointer Network, cho phép mô hình linh hoạt lựa chọn từ tiếp theo từ bộ từ vựng cố định hoặc sao chép trực tiếp từ các OCR tokens trong ảnh để đảm bảo độ chính xác của nội dung văn bản được mô tả.



Hình 5.4: Mô tả mô đun Reasoning and Generation trong mô hình DEVICE [22]

5.1.5. Mô hình Anchor Captioner

Các nghiên cứu Text-based Image Captioning trước đây thường chỉ tập trung vào việc tạo ra một câu mô tả toàn cục duy nhất cho toàn bộ bức ảnh. Tuy nhiên, các thực thể và nội dung chữ trong hình là rất phức tạp và phong phú, vì vậy việc chỉ tạo sinh một câu mô tả duy nhất là không đủ để bao quát toàn bộ thông tin. Để giải quyết khó khăn này, tác giả Xu và cộng sự [23] nghiên cứu cách tạo sinh nhiều câu mô tả đa dạng, tập trung vào các vùng nội dung khác nhau có trong bức

ảnh thông qua cơ chế mỏ neo (anchor). Kiến trúc của mô hình gồm 3 mô-đun chính:

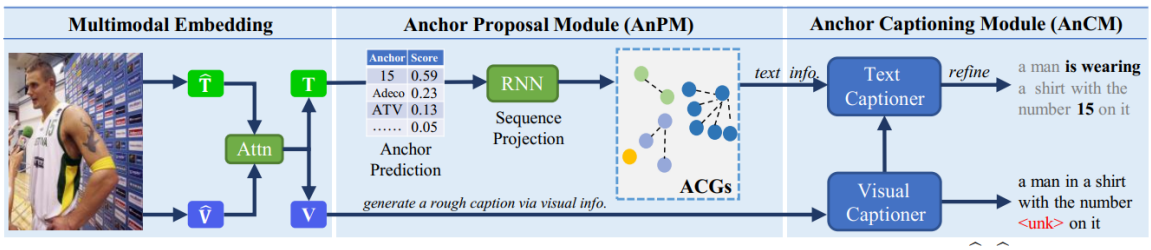
- **Multimodal embedding**: Tại mô-đun này, các đặc trưng hình ảnh của vật thể (Visual Objects) và văn bản trong ảnh (Scene Text/OCR tokens) được trích xuất và nhúng vào không gian vector chung. Visual Embedding được trích xuất bằng mô hình phát hiện vật thể, được làm giàu ngữ nghĩa bằng cách kết hợp tọa độ của bounding boxes, và đặc trưng ngữ nghĩa của các vật thể. Token Embedding được mã hóa bằng cách kết hợp thêm đặc trưng FastText và Pyramidal Histogram of Characters (PHOC) giúp làm giàu ngữ nghĩa hơn cho đặc trưng đối tượng chữ. Sau cùng Multimodal embedding fusion, sẽ nhúng các đặc trưng đa nguồn (modalities) lại với nhau bằng lớp L1 Transformers, làm giàu ngữ nghĩa của của các OCR Tokens.
- **Anchor proposal module**: Anchor Proposal Module được xây dựng để xác đinh đối tượng chữ nào cần được chú ý khi mô tả. Thừa hưởng ý tưởng của region proposal network (RPN), AnPM sử dụng một lớp tuyến tính kết hợp hàm Softmax để dự đoán điểm số quan trọng của các OCR Tokens. Các Token có scores cao sẽ được giữ lại làm, ta gọi các token này là mỏ neo. Trong quá trình huấn luyện, token có điểm số cao nhất được chọn làm Anchor (T_{anchor}). Trong quá trình suy luận (Inference), mô hình chọn Top-K token có điểm cao nhất để tạo ra K câu mô tả khác nhau. Cơ chế Anchor-centred Graph Construction (ACG) nhằm mô hình hóa mối quan hệ phụ thuộc tiềm ẩn giữa một T_{anchor} và các token còn lại trong ảnh. Với mỗi T_{anchor}, tác giả sử dụng một mạng RNN, T_{anchor} được lựa chọn làm trạng thái ẩn ban đầu, lần lượt xử lý các tokens khác. Cách thiết kế này cho phép RNN học được sự tương tác có điều kiện giữa anchor và từng token, thay vì chỉ dựa trên ngữ cảnh toàn cục.

  Đầu ra của RNN là biểu diễn token đã được cập nhật:

$$T_{graph} = RNN(T; T_{anchor}) \quad (48)$$

Với T_{graph} là đặc trưng mới của M tokens khi thích hợp thông tin từ Anchor. Họ ước lượng mức độ liên quan giữa từng anchor và các tokens bằng phép biến đổi tuyến tính và sigmoid. Dựa vào điểm số này, các token có điểm liên kết cao (> 0.5) sẽ được chọn chung với anchor gốc để hình thành Anchor-centered Graph. Tổng thể, AnPM có nhiệm vụ tự động chọn ra các OCR quan trọng làm Anchor, và xây dựng một tập ACG tương ứng. Từng ACG như là tín hiệu để mô hình tạo sinh mô tả đa dạng và giàu thông tin hơn.

- **Anchor captioning module**: Anchor Captioning Module (AnCM), một kiến trúc sinh caption theo chiến lược progressive refinement, lấy cảm hứng từ Deliberation Network. AnCM gồm hai thành phần: visual-captioner (AnCMv) và text-captioner (AnCMt). AnCMv sinh ra một caption chỉ dựa vào thông tin thị giác bằng cách sử dụng embedding hình ảnh, đóng vai trò như bước tạo ngữ cảnh ban đầu. Tiếp theo, AnCMt tinh chỉnh caption này dưới sự dẫn hướng của Anchor-centred Graph (ACG), cho phép mô hình tập trung vào các OCR token liên quan và đưa chúng vào caption cuối cùng. Hai mô-đun được huấn luyện end-to-end thông qua việc truyền trực tiếp hidden state từ AnCMv sang AnCMt. Nhờ kết hợp sinh caption thị giác và tinh chỉnh dựa trên cấu trúc anchor–token, AnCM khai thác hiệu quả đặc trưng cốt lõi của TextCap, giúp cải thiện độ chính xác và tính nhất quán của caption chứa văn bản so với các phương pháp captioning thông thường.

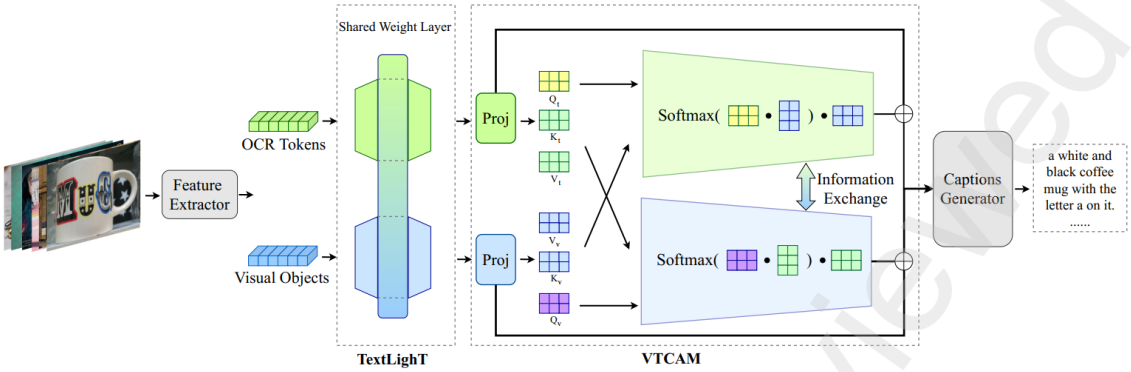


Hình 5.5: Kiến trúc mô hình Anchor-Captioner [23]

5.1.6. Mô hình LCM-Captioner

Các phương pháp chú thích ảnh dựa trên văn bản (Text-based Image Captioning) hiện nay thường sử dụng các kiến trúc mạng phức tạp, cố gắng liên kết thông tin hình ảnh và văn bản với nhau. Điều này dẫn đến thời gian chạy lâu, bộ nhớ không được tối ưu hóa, khó khăn khi triển khai thực thế, đôi khi thất bại trong việc căn chỉnh ngữ nghĩa giữa vật thể và văn bản OCR. LCM-Captioner [53] đề xuất hai mô đun chính để giải quyết vấn đề trên:

- **TextLighT**: TextLighT có vai trò ánh xạ các đặc trưng hình ảnh và văn bản xuống không gian chiều thấp hơn, sử dụng phép biến đổi tuyến tính theo nhóm (Group Linear Transformations) giúp giảm chi phí tính toán. Lớp mạng tuyến tính chia sẻ trọng số (Shared Weight Network), giúp mô hình học được các biểu diễn chung (joint representations) cho hình ảnh và văn bản phong phú hơn
- **VTCAM**: VTCAM mô hình hóa giả thuyết rằng, các nội dung văn xuất hiện ở câu mô tả thường bao hàm những thông tin về các đối tượng thị giác có trong ảnh. VTCAM sử dụng kiến trúc hai luồng gồm hai mô-đun transformer song song (dual-stream network). Trước khi đưa vào VTCAM, đặc trưng thị giác và văn bản được đưa qua lớp TextLighT trước khi đưa qua VTCAM. Sử dụng cơ chế mutual attention giữa hai đặc trưng, kết hợp với mạng FFN để tăng cường ngữ nghĩa của các đặc trưng, trước khi đưa vào bộ giải mã để mô tả hình ảnh. VTCAM đóng vai trò quan trọng trong việc căn chỉnh và liên kết thông tin thị giác và văn bản, giúp mô hình xác định chính xác các đối tượng thị giác then chốt đồng thời khai thác hiệu quả nội dung văn bản quan trọng, từ đó dẫn hướng mô hình sinh caption tập trung vào các cặp đối tượng - văn bản có ý nghĩa.



Hình 5.6: Kiến trúc mô hình LCM-Captioner [53]

5.2. Các mô hình cơ sở cho bài toán Text Summarization

Mặc dù bao gồm nhiều đặc trưng thông tin trong đồ họa, đối với các loại dữ liệu đặc thù như infographic, nội dung văn bản thường chiếm tỷ lệ cao và đóng vai trò chủ đạo so với các thực thể thị giác thông thường (đã được chúng tôi chứng minh ở mục **3.2.5**). Trong ngữ cảnh đó, các mô hình mô tả hình ảnh dựa trên OCR khó khăn trong việc hiểu được toàn bộ nội dung văn bản có trong hình.

Do đó, nhóm chúng tôi đề xuất tiếp cận bài toán thông qua tác vụ tóm tắt nội dung xuất hiện trong hình ảnh (Abstractive Summarization). Các mô hình ngôn ngữ đơn ngữ tiếng Việt mạnh mẽ nhất hiện nay là BARTpho [30] và ViT5 [29] đã được lựa chọn để huấn luyện, đồng thời mô hình đa ngôn ngữ mT5 [61] cũng được đưa vào nhằm đối sánh và đánh giá hiệu năng một cách trực quan.

5.2.1. Trích xuất đặc trưng cho các mô hình Tóm tắt văn bản trừu tượng

Đối với hướng tiếp cận Tóm tắt văn bản trừu tượng (Abstractive Text Summarization) trên bộ dữ liệu, chúng tôi yêu cầu độ chính xác, tính liền mạch và thứ tự logic cao đối với văn bản đầu vì đây là yếu tố then chốt quyết định chất lượng bản tóm tắt. Do đó PaddleOCR [37], kết hợp với VietOCR [38], được sử dụng để trích xuất nội dung văn bản có trong ảnh.

Khác với kiểu hình ảnh thuần túy, infographic thường sở hữu cấu trúc thông tin theo cụm, nơi các nội dung liên quan được gom nhóm thành từng đoạn theo vùng không gian để tăng tính trực quan. Do đó, phương pháp trích xuất văn bản

tuần tự truyền thống thường không hiệu quả do làm phá vỡ cấu trúc ngữ nghĩa của tài liệu. Nhằm đảm bảo trật tự và tính mạch lạc của văn bản đầu vào, nhóm áp dụng quy trình trích xuất hai giai đoạn. Nhóm sử dụng mô hình *PP-DocLayout_plus-L* là mô hình nhận diện bố cục tài liệu thuộc hệ sinh thái PaddleOCR [37] .Đồng thời, VietOCR [38] là một mô hình mã nguồn mở end-to-end dành cho tác vụ Nhận Dạng Tiếng Việt (Vietnamese Recognition). Đầu tiên, mô hình *PP-DocLayout_plus-L* được sử dụng để phân tích và nhận diện bố cục tài liệu (Layout Analysis). Sau khi xác định được các vùng văn bản có trong hình, các khối văn bản sẽ được sắp xếp lại theo một trình tự đọc logic. Cuối cùng, sử dụng mô hình VietOCR [38] để nhận diện chính xác nội dung tiếng Việt trong từng khối, trước khi nối kết chúng thành một văn bản đầu vào hoàn chỉnh cho mô hình tóm tắt.

Ngoài ra, để làm giàu thêm đặc trưng từ hình ảnh, chúng tôi còn trích xuất đặc trưng bố cục bằng mô hình LayoutLMv3 [39]. Trước đây, đặc trưng trục (grid features) thường được trích xuất từ các lớp tích chập cuối cùng của mạng CNN (như ResNet-101 [62] hoặc VGG-16 [63]). Các mô hình kinh điển như Show, Attend and Tell [18] đã sử dụng đặc trưng này để cho phép cơ chế Attention tập trung vào các vùng không gian cụ thể trên ảnh. Gần đây, việc sử dụng đặc trưng dạng mảnh (patch features) thông qua các kiến trúc Transformer như Vision Transformer (ViT) [60] đã chứng minh được hiệu suất vượt trội. LayoutLMv3 [39] sở hữu kiến trúc Unified Multimodal Transformer, được huấn luyện dựa trên các tác vụ như Masked Image Modeling (MIM) và Word-Patch Alignment (WPA), giúp nó đạt được khả năng căn chỉnh cực tốt giữa nội dung chữ và vùng ảnh tương ứng. Mô hình phù hợp với đặc trưng dữ liệu infographic khi mà giá trị thông tin không chỉ nằm ở nội dung chữ thuần túy mà còn phụ thuộc chặt chẽ vào cấu trúc và bố cục sắp xếp không gian của các đoạn nội dung. Việc sử dụng LayoutLMv3 [39] giúp hệ thống hiểu sâu sắc mối quan hệ giữa hình ảnh và văn bản, từ đó sinh ra các mô tả chính xác và giàu ngữ nghĩa hơn.

5.2.2. Mô hình BARTpho

BARTpho [30] là mô hình sequence-to-sequence đơn ngữ quy mô lớn đầu tiên dành cho tiếng Việt, được phát triển bởi VinAI gồm hai phiên bản $\text{BARTpho}_{\text{syllable}}$ (dựa trên đơn vị âm tiết) và $\text{BARTpho}_{\text{word}}$ (dựa trên đơn vị từ), nhằm tối ưu hóa khả năng xử lý các đặc thù về hình thái học và ngữ pháp của tiếng Việt. BARTpho kế thừa cơ chế tiền huấn luyện của mô hình tự mã hóa khử nhiễu từ kiến trúc BART gốc. Vì thế, BARTpho hình thành khả năng thấu hiểu ngữ cảnh sâu sắc và tái tạo nội dung một cách tự nhiên. Chính ưu thế này đã biến BARTpho trở thành công cụ tối ưu cho các tác vụ tạo sinh văn bản. Nhóm thực hiện hiện tinh mô hình trên hai phiên bản $\text{BARTpho}_{\text{syllable}}$ và $\text{BARTpho}_{\text{word}}$ nhằm đánh giá tác động của hai phương pháp mô hình hóa đơn vị ngôn ngữ khác nhau đến chất lượng của văn bản được sinh ra.

5.2.3. Mô hình ViT5

ViT5 [29] là một mô hình ngôn ngữ transformers encoder-decoder tiền huấn luyện tiên tiến, được xây dựng dựa trên kiến trúc T5 và tối ưu hóa đặc biệt cho ngôn ngữ Tiếng Việt. ViT5 được huấn luyện trên bộ dữ liệu tiếng Việt chất lượng cao, giúp ViT5 nắm bắt được sự đa dạng của từ vựng, các cấu trúc câu phức tạp và đặc biệt là các sắc thái ngữ nghĩa đặc thù của Tiếng Việt. Nhờ khả năng hiểu sâu sắc cấu trúc câu, mô hình này vượt trội trong việc tổng hợp thông tin rời rạc thành một đầu ra ngôn ngữ hoàn chỉnh và giàu ý nghĩa dành cho tác vụ tóm tắt văn bản, tạo sinh văn bản dành cho ngôn ngữ tiếng Việt. Nhóm thực hiện tinh chỉnh ViT5 với hai phiên bản $\text{ViT5}_{\text{base}}$ và $\text{ViT5}_{\text{large}}$ tuân thủ quy trình huấn luyện chuẩn của mô hình T5.

5.2.4. Mô hình mT5

mT5 (Multilingual T5) [61] là một mô hình đa ngôn ngữ, tiên tiến được Google giới thiệu, kế thừa kiến trúc "text-to-text" từ mô hình T5 gốc, được huấn luyện trên bộ dữ liệu khổng lồ *mC4 (Multilingual Colossal Clean Crawled Corpus)* gồm 101 ngôn ngữ, trong đó có tiếng Việt. Nhờ khả năng chuyển giao tri

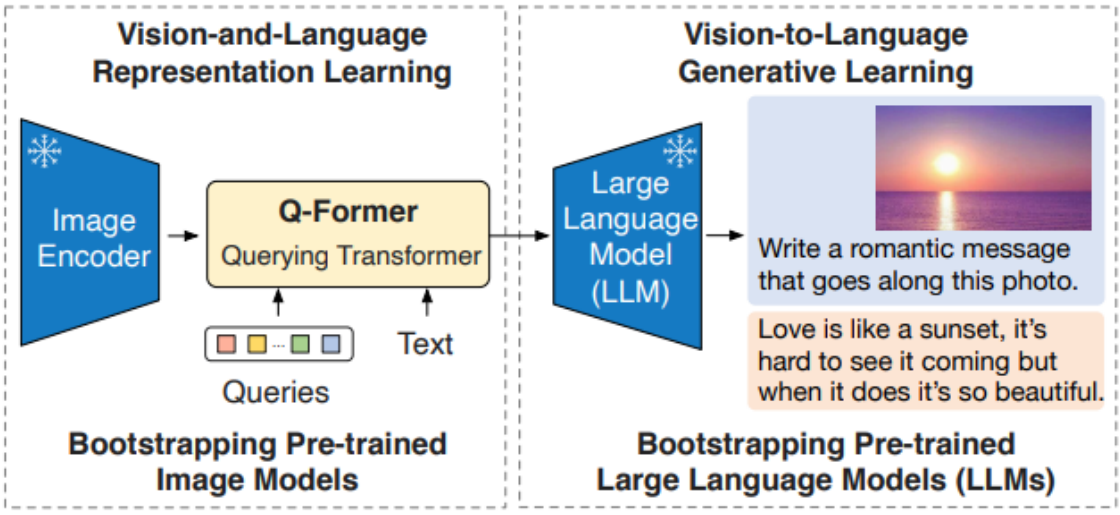
thức liên ngôn ngữ (cross-lingual transfer) xuất sắc, mô hình hoạt động hiệu quả ngay cả khi dữ liệu huấn luyện cho ngôn ngữ đích còn hạn chế. Nhóm thực hiện tinh chỉnh mT5 với phiên bản $mT5_{base}$ nhằm đánh giá mô hình đa ngôn ngữ so với các mô hình đơn ngữ trên tiếng Việt.

5.3. Các mô hình đa thể thức lớn

Bên cạnh các mô hình cơ sở, các mô hình đa thể thức lớn (Large Multimodal Models – LMMs) cũng được đầu tư nghiên cứu và công bố rộng rãi nhằm thực hiện các tác vụ với đa dạng dữ liệu đầu vào – đầu ra. Chúng tôi thực hiện đánh giá các mô hình đa thể thức lớn mã nguồn mở trên tác vụ tạo sinh chú thích tóm tắt infographic để đánh giá ngưỡng năng lực của các mô hình này. Các mô hình này được tuyển chọn dựa trên sự đa dạng về kiến trúc và phạm vi ngôn ngữ, từ các hệ thống đa ngôn ngữ (multilingual) đến các mô hình đơn ngữ (monolingual) được tối ưu hóa chuyên sâu cho tiếng Việt.

5.3.1. BLIP-2

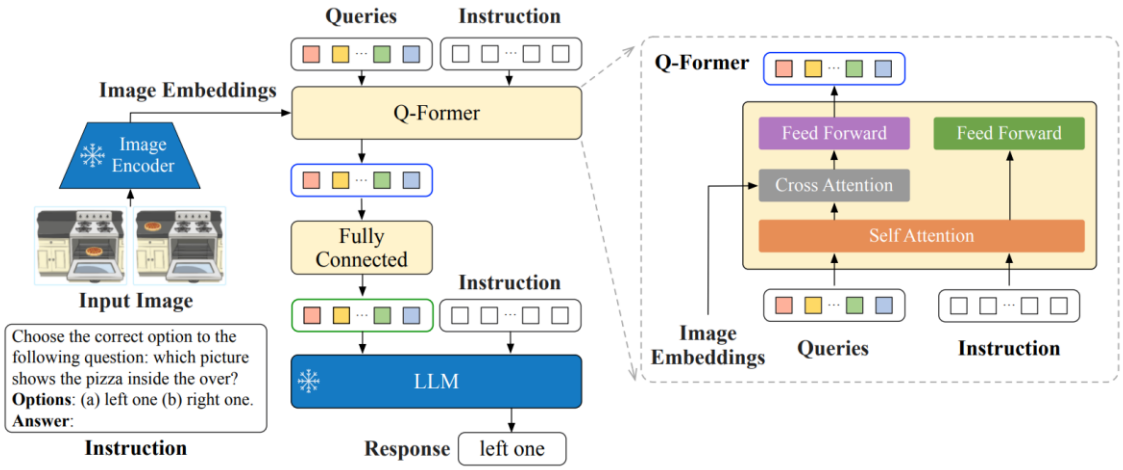
BLIP-2 (*Hình 5.7)* là mô hình được giới thiệu bởi nhóm nghiên cứu SaleForce [42]. BLIP-2 có cấu trúc gồm 3 khối Image Encoder, Text Encoder và Text Decoder để hực hiện những tác vụ xử lý đồng thời hình ảnh và ngôn ngữ. Tuy nhiên đối với BLIP-2, nhờ sự phát triển của các mô hình ViT [41] và các mô hình ngôn ngữ lớn (LLMs), khác với BLIP, BLIP-2 sử dụng Frozen Image Encoders và Frozen LLMs, kết hợp với kiến trúc "Query Transformer" như là cầu nối giữa Image Encoder và LLM. Bằng cách này, BLIP-2 đạt được hiệu suất SOTA giúp cắt giảm chi phí tính toán, chỉ cần huấn luyện một lớp trung gian duy nhất.



Hình 5.7: Kiến trúc mô hình BLIP-2 [42]

5.3.2. InstructBLIP

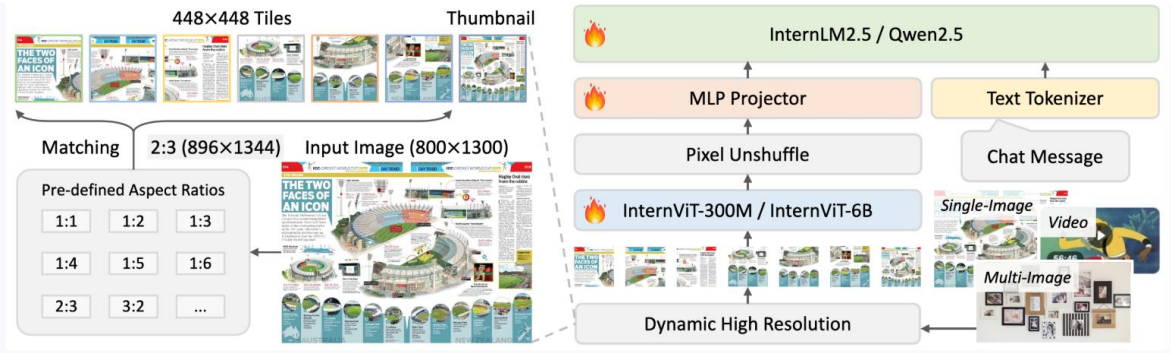
Cải tiến thêm trên mô hình BLIP2, InstructBLIP [43] *(Hình 5.8)* được nhóm nghiên cứu Salesforce giới thiệu nhằm tối ưu hóa khả năng phản hồi theo chỉ dẫn (instruction following) của các mô hình đa thể thức. Điểm cải tiến cốt lõi của InstructBLIP so với BLIP-2 nằm ở kiến trúc "Instruction-aware Q-Former", cho phép đưa trực tiếp các câu lệnh văn bản vào khối Q-Former để trích xuất các đặc trưng hình ảnh có tính mục đích cao, thay vì chỉ trích xuất đặc trưng hình ảnh chung chung. Mô hình vẫn duy trì chiến lược đóng băng các thành phần Image Encoder và các mô hình ngôn ngữ lớn (LLM) mạnh mẽ như Vicuna hay Flan-T5 để tận dụng tối đa tri thức sẵn có. Nhờ việc huấn luyện trên tập dữ liệu đa nhiệm phong phú được chuyển đổi sang định dạng chỉ dẫn, InstructBLIP đạt được hiệu suất SOTA vượt trội, đặc biệt là khả năng zero-shot một cách linh hoạt và chính xác hơn nhiều so với các thế hệ tiền nhiệm.



Hình 5.8: Kiến trúc mô hình InstructBLIP [43]

5.3.3. InternVL 2.5

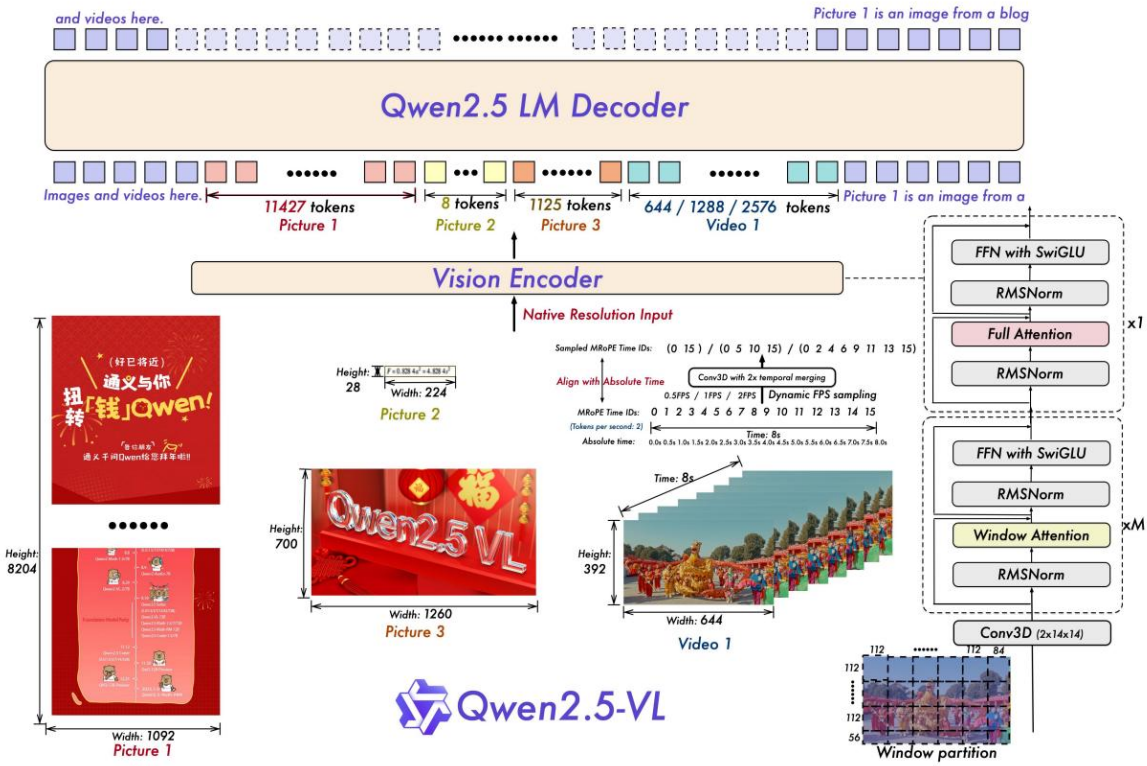
InternVL [64] là một hệ thống mô hình đa thể thức mã nguồn mở mạnh mẽ, được thiết kế nhằm thu hẹp khoảng cách hiệu suất giữa các mô hình mã nguồn mở và các mô hình thương mại hàng đầu như GPT-4V [65]. Khác biệt lớn nhất của InternVL so với dòng BLIP chính là việc tập trung vào việc mở rộng quy mô bộ mã hóa hình ảnh với InternViT-6B, một Vision Transformer có kích thước lên đến 6 tỷ tham số, thay vì sử dụng các ViT nhỏ hơn thường thấy. Mô hình này áp dụng cơ chế căn chỉnh thông tin thị giác và ngôn ngữ thông qua một lớp kết nối MLP đơn giản nhưng hiệu quả để kết hợp với các mô hình ngôn ngữ lớn (LLM) như InternLM [66] hoặc Llama [67]. Đặc biệt, InternVL nổi bật nhờ khả năng xử lý hình ảnh độ phân giải cực cao thông qua kỹ thuật chia nhỏ hình ảnh động, giúp nó đạt được độ chính xác vượt trội trong các tác vụ đòi hỏi sự tỉ mỉ như nhận diện văn bản (OCR), phân tích tài liệu phức tạp và lập luận thị giác đa bước, đạt nhiều kỷ lục SOTA trên các bảng xếp hạng đa thể thức.



Hình 5.9: Kiến trúc mô hình InternVL 2.5 [64]

5.3.4. Qwen2.5VL

Qwen2.5-VL [44] *(Hình 5.10)* là một bước tiến quan trọng từ đội ngũ Alibaba Cloud, nổi bật với khả năng xử lý hình ảnh ở độ phân giải cao và hỗ trợ tương tác đa ngôn ngữ mạnh mẽ. Kiến trúc của Qwen-VL kết hợp giữa bộ mã hóa hình ảnh ViT-bigG và mô hình ngôn ngữ lớn Qwen-7B thông qua một thành phần trung gian gọi là Visual Adapter, sử dụng cơ chế cross-attention để nén các chuỗi đặc trưng hình ảnh giúp tối ưu hóa hiệu suất tính toán. Mô hình Qwen2.5-VL còn tích hợp cơ chế M-RoPE (Multimodal Rotary Position Embedding), giúp đồng nhất việc biểu diễn vị trí trong không gian và thời gian, từ đó nâng cao khả năng định vị vật thể và hiểu các cấu trúc đồ họa phức tạp một cách chính xác. Với việc hỗ trợ đầu vào độ phân giải 448x448 và được huấn luyện trên các tập dữ liệu chất lượng cao, Qwen-VL đạt hiệu suất xuất sắc trong các tác vụ nhận dạng văn bản (OCR), phân tích biểu đồ và hội thoại đa bước, trở thành một trong những mô hình toàn diện nhất trong phân khúc mô hình ngôn ngữ - thị giác hiện nay.



Hình 5.10: Kiến trúc mô hình Qwen2.5-VL [44]

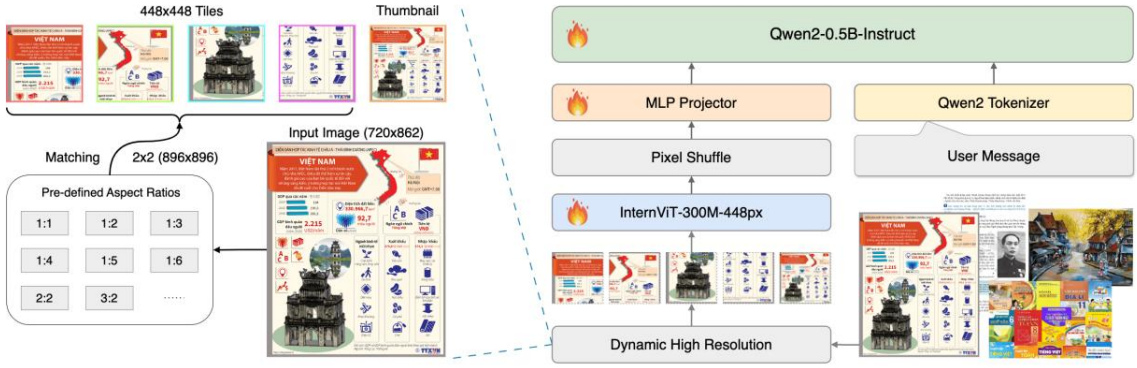
5.3.5. LLaVA-NeXT

LLaVA-NeXT [45] (hay còn gọi là LLaVA-1.6) là phiên bản nâng cấp mạnh mẽ từ mô hình LLaVA-1.5 phổ biến, được phát triển bởi các nhà nghiên cứu từ UW-Madison, Microsoft Research và Đại học Columbia. Cải tiến đột phá nhất của LLaVA-NeXT là việc áp dụng kỹ thuật AnyRes (Dynamic High-Resolution), cho phép mô hình xử lý hình ảnh ở độ phân giải cao hơn nhiều bằng cách chia nhỏ hình ảnh một cách thông minh, giúp bảo toàn các chi tiết thị giác quan trọng mà không làm tăng quá mức chi phí tính toán. Thay vì thay đổi cấu trúc phức tạp, LLaVA-NeXT tiếp tục duy trì sự đơn giản và hiệu quả của kiến trúc LLaVA nguyên bản với một lớp kết nối MLP kết nối giữa Vision Encoder và các mô hình ngôn ngữ lớn (LLM) hiện đại như Mistral-7B hay Yi-34B. Nhờ việc tinh chỉnh trên tập dữ liệu chỉ dẫn thị giác chất lượng cao và khả năng suy luận logic được cải thiện rõ rệt, LLaVA-NeXT đã thu hẹp đáng kể khoảng cách với các mô hình mã nguồn đóng như GPT-4V, đặc biệt xuất sắc trong các tác vụ hiểu tài liệu, đọc

biểu đồ và suy luận hình ảnh phức tạp trong khi vẫn duy trì được tính linh hoạt và dễ tiếp cận cho cộng đồng mã nguồn mở.

5.3.6. Vintern-1B

Vintern-1B [46] *(Hình 5.11)* là mô hình ngôn ngữ lớn đa thể thức chuyên biệt cho tiếng Việt, được phát triển bởi nhóm nghiên cứu 5CD-AI dựa trên nền tảng kiến trúc InternVL2. Mô hình này kết hợp bộ mã hóa thị giác InternViT-300M cùng mô hình ngôn ngữ Qwen2-Instruct [44] thông qua một lớp kết nối MLP, tạo nên một hệ thống nhỏ gọn nhưng cực kỳ hiệu quả với quy mô khoảng 1 tỷ tham số. Điểm đột phá của Vintern nằm ở việc tối ưu hóa sâu cho ngôn ngữ và đặc trưng văn hóa Việt Nam, đặc biệt xuất sắc trong các tác vụ nhận dạng ký tự quang học (OCR) tiếng Việt, hiểu tài liệu phức tạp và trả lời câu hỏi hình ảnh (VQA) trong bối cảnh địa phương. Nhờ áp dụng cơ chế phân giải động để xử lý chi tiết hình ảnh và cấu trúc tinh gọn, Vintern không chỉ đạt hiệu suất vượt trội trên các bảng đo lường tiếng Việt mà còn có thể triển khai mượt mà trên các thiết bị phần cứng phổ thông, mang lại giải pháp AI đa thể thức dễ tiếp cận cho cộng đồng người dùng Việt.



Hình 5.11: Kiến trúc mô hình Vintern-1B [46]

5.3.7. LaVy

LaVy [47] là một mô hình ngôn ngữ - thị giác quy mô lớn được thiết kế chuyên biệt cho tiếng Việt, dựa trên nền tảng kiến trúc của dòng LlaVA [45]. Mô hình này kết hợp bộ mã hóa hình ảnh mạnh mẽ CLIP-ViT-L/14 với một mô hình ngôn

ngữ lớn đã được tối ưu hóa cho tiếng Việt là PhoGPT [68], thông qua một lớp kết nối MLP đơn giản để căn chỉnh đặc trưng hình ảnh vào không gian ngôn ngữ. Điểm đóng góp quan trọng của LaVy nằm ở việc xây dựng và sử dụng tập dữ liệu chỉ dẫn đa thể thức chất lượng cao bằng tiếng Việt, được chuyển ngữ và tinh lọc kỹ lưỡng từ các tập dữ liệu quốc tế kết hợp với dữ liệu bản địa. Nhờ đó, LaVy không chỉ thể hiện khả năng hiểu hình ảnh vượt trội mà còn am hiểu sâu sắc về ngữ cảnh văn hóa, địa danh và ngôn ngữ Việt Nam, giải quyết hiệu quả các tác vụ trả lời câu hỏi hình ảnh (VQA) và mô tả nội dung bằng tiếng Việt một cách tự nhiên và chính xác.

Chương 6. CÀI ĐẶT, THỬ NGHIỆM VÀ ĐÁNH GIÁ

6.1. Các độ đo đánh giá

Nhằm đánh giá khách quan chất lượng của các chú thích tóm tắt, nghiên cứu tiến hành so sánh kết quả được tạo ra từ mô hình với tập hợp các câu mô tả chuẩn được thu thập trong tập dữ liệu của chúng tôi. Quá trình đánh giá được thực hiện thông qua bộ công cụ chuẩn hóa pycocoevalcap[5], bao gồm các hệ số đo lường: BLEU-n [69] nhằm xác định độ chính xác dựa trên sự trùng khớp cụm từ; METEOR [70] đánh giá dựa trên sự tương quan ngữ nghĩa và độ phủ; ROUGE-L [71] kiểm định cấu trúc câu thông qua chuỗi con chung dài nhất; và CIDEr [72] là độ đo đặc thù cho bài toán chú thích hình ảnh nhằm đo lường mức độ đồng thuận giữa câu dự đoán và tập tham chiếu.

6.1.1. BLEU

Chỉ số này đánh giá chất lượng câu dựa trên sự trùng khớp của các cụm n-gram. Quá trình tính BLEU được tính toán như sau. Với P_n là điểm precision cho từng n-gram, theo công thức:

$$p_n = \frac{\sum_{C\in\{Candidates\}} \sum_{n\text{-gram}\in C} Count_{clip}(n\text{-gram})}{\sum_{C'\in\{Candidates\}} \sum_{n\text{-gram}'\in C'} Count(n\text{-gram}')} \tag{49}$$

Để tránh việc mô hình sinh ra các câu quá ngắn nhằm đạt điểm Precision cao, hệ số phạt Brevity Penalty (BP) được áp dụng:

$$BP = \begin{cases} 1 & \text{nếu } c > r \\ e^{(1-r/c)} & \text{nếu } c \leq r \end{cases} \tag{50}$$

Trong đó c là độ dài câu dự đoán và r là độ dài câu tham chiếu. Công thức tổng quát của $BLEU_N$ là:

$$BLEU_N = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \ln p_n\right) \tag{51}$$

[5] https://github.com/salaniz/pycocoevalcap

Chúng tôi tính toán các giá trị BLEU với n-gram lần lượt là BLEU-1, BLEU-2, BLEU-3, BLEU-4 để đánh giá độ trùng khớp từ đơn cho độ chính xác tuyệt đối đến cụm từ dài phản ánh độ trôi chảy của câu.

6.1.2. METEOR

Khác với BLEU chỉ dựa trên sự trùng khớp chính xác của từ, METEOR cải tiến bằng cách sử dụng Stemming để so khớp các từ có cùng gốc và Synonymy để so khớp các từ đồng nghĩa thông qua bộ từ điển.

Độ đo này nhấn mạnh vào cả Recall và Precision, giúp đánh giá sát hơn với cách con người cảm nhận ngôn ngữ.

Đầu tiên, chỉ số F_{mean} được tính toán bằng cách kết hợp giữa Precision (P) và Recall (R) của các từ đơn (unigram):

$$\mathrm{F_{mean}} = \frac{10 \cdot \mathrm{P} \cdot \mathrm{R}}{\mathrm{R} + 9 \cdot \mathrm{P}} \tag{52}$$

Để đánh giá mức độ trôi chảy và thứ tự từ trong câu, METEOR tính toán một hệ số phạt Penalty dựa trên số lượng các cụm từ khớp liên tiếp nhau. Một câu có ít cụm từ khớp (tức là các từ khớp nằm sát nhau) sẽ có hệ số phạt thấp hơn:

$$\mathrm{Penalty} = 0.5 \cdot \left(\frac{\mathrm{chunks}}{\mathrm{matched\_unigrams}}\right)^3 \tag{53}$$

Điểm METEOR cuối cùng là kết quả của việc điều chỉnh F_{mean} bởi hệ số phạt:

$$\mathrm{METEOR} = \mathrm{F_{mean}} \cdot (1 - \mathrm{Penalty}) \tag{54}$$

Chúng tôi sử dụng METEOR để đánh giá độ trùng khớp chính xác của từ được sử dụng trong câu tạo sinh so với câu tham chiếu, từ đó phản ánh đánh giá dựa trên sự tương quan ngữ nghĩa và độ phủ.

6.1.3. ROUGE-L

ROUGE-L dựa trên khái niệm Longest Common Subsequence (LCS) - Chuỗi con chung dài nhất. Điểm số này xác định mức độ tương đồng về cấu trúc giữa câu dự đoán và câu tham chiếu mà không cần các cụm n-gram phải liên tiếp nhau. Chúng tôi sử dụng thang đo này để kiểm định sự tương đồng về cấu trúc câu tạo sinh với câu tham chiếu.

Giả sử câu tham chiếu X có độ dài m và câu dự đoán Y có độ dài n, Recall và Precision được tính:

- Recall (R_{lcs}):

$$R_{lcs} = \frac{LCS(X, Y)}{m} \tag{55}$$

- Precision(P_{lcs}):

$$P_{lcs} = \frac{LCS(X, Y)}{n} \tag{56}$$

Điểm ROUGE-L (hay F_{lcs}) được tính bằng công thức F-measure:

$$F_{lcs} = \frac{(1 + \beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{57}$$

6.1.4. CIDEr

CIDEr là một thang đo quan trọng trong bài toán tạo sinh chú thích, sử dụng trọng số TF-IDF cho các n-gram để giảm thiểu tầm quan trọng của các từ phổ biến nhưng ít mang thông tin và tăng trọng số cho các từ mô tả đặc trưng của ảnh. Nó tính toán sự tương đồng Cosine giữa vector của câu dự đoán và tập hợp các câu tham chiếu để tìm ra sự đồng thuận. Một điểm CIDEr cao cho thấy câu mô tả của mô hình rất gần với cách mà đa số con người mô tả về hình ảnh đó.

Vector $g^n(s_{ij})$ biểu diễn tần suất xuất hiện của các n-gram trong câu. Điểm CIDEr giữa câu dự đoán c_i và tập câu tham chiếu $S_i = \{s_{i1}, \dots, s_{im}\}$ được tính bằng trung bình cộng sự tương đồng Cosine:

$$\mathrm{CIDEr}_n(c_i, S_i) = \frac{1}{m}\sum_{j=1}^{m}\frac{g^n(c_i)\cdot g^n(s_{ij})}{|g^n(c_i)||g^n(s_{ij})|} \quad (58)$$

6.2. Thiết lập tham số thử nghiệm

6.2.1. Tham số thực nghiệm cho các mô hình cơ sở

Bảng 6.1: Bảng thông số thực nghiệm các mô hình cơ sở

Mô hình	Phiên bản	Cài đặt	Thông số
LSTM-R	-	Batche size	32
M4C-Captioner	-	Warmup	10
DEVICE	-	Learning rate	1e-4
Anchor Captioner	-	Learning rate step	7
LCM-Captioner	-	Epoch	40 (Early Stopping)
$\text{BARTpho}_{\text{syllable}}$	Base (132M)		
$\text{BARTpho}_{\text{word}}$	Base (150M)		
ViT5	Base (310M)		
mT5	Base (580M)		

6.2.2. Tham số thực nghiệm các mô hình đa thể phương lớn

Do sự khác biệt về đặc trưng kiến trúc và khả năng hỗ trợ ngôn ngữ giữa các mô hình đa thể thức lớn (LMMs), chúng tôi đã thiết lập quy trình xử lý dữ liệu riêng biệt cho từng nhóm mô hình. Đối với nhóm mô hình ưu tiên hoặc chỉ hỗ trợ đầu ra tiếng Anh (như BLIP-2 và InstructBLIP), chúng tôi áp dụng thêm một bước hậu xử lý bằng cách sử dụng mô hình EnViT5 [17] để chuyển ngữ các câu trả lời sang tiếng Việt, đảm bảo tính đồng nhất cho quá trình đánh giá.

Bảng 6.2: Bảng đặc điểm các mô hình đa thể thức lớn

Mô hình	Phiên bản	Yêu cầu prompt	Ngôn ngữ tóm tắt đầu ra
BLIP2	blip2-opt-2.7b	Không	Tiếng Anh
InstructBLIP	flan-t5-xl	Có	Tiếng Anh
InternVL 2.5	InternVL2_5-1B	Có	Tiếng Việt
Qwen2.5VL	Qwen2.5-VL-7B-Instruct	Có	Tiếng Việt
LlaVA-NeXT	llava-v1.6-mistral-7b	Có	Tiếng Việt
Vintern-1B	Vintern-1B-v2	Có	Tiếng Việt
LaVy	LaVy-instruct	Có	Tiếng Việt

Nhằm định hướng mô hình tập trung vào các đặc điểm đặc thù của dữ liệu, chúng tôi sử dụng một cấu trúc câu lệnh chung chuyên biệt để tối ưu hóa khả năng phân tích infographic như sau:

"*Bạn là chuyên gia phân tích hình ảnh và nhận dạng văn bản infographic. Nhiệm vụ là tạo chú thích chi tiết, ưu tiên diễn giải văn bản trong ảnh. Kết hợp các yếu tố này để viết chú thích ngắn, độ dài 1-3 câu, rõ ràng và chính xác, nhấn mạnh vào thông tin tổng quan của infographic.*"

Thông qua cấu trúc này, mô hình được yêu cầu không chỉ nhận diện các thành phần thị giác mà còn phải tổng hợp logic các luồng thông tin văn bản, từ đó đưa ra những nhận định súc tích nhưng vẫn bao quát được nội dung cốt lõi của bức ảnh.

6.2.3. Tham số thực nghiệm cho mô hình đề xuất

Bảng 6.3: Bảng thông số cài đặt mô hình đề xuất

Mô hình/Thành phần	Cài đặt	Thông số
SpatialOverlapAttention	IoRs_LookupTable (Shape)	32×16
TripletContrastiveLoss	num_syllable β, γ, δ α	3 0.33 0.5
Encoder-Decoder Model	Batch size Warmup Learning rate Learning rate step Epoch	32 10 1e-4 7 40 (Early Stopping)

6.3. Kết quả và phân tích

6.3.1. Phân tích kết quả đánh giá các mô hình theo nhiều hướng tiếp cận

Kết quả so sánh trên các mô hình thực nghiệm cho thấy hiệu suất chênh lệch rõ rệt của các mô hình ở các hướng tiếp cận khác nhau. Đối với các mô hình đa thể thức lớn, hiệu suất tổng thể của các mô hình khá kém trên tác vụ này. Mô hình Vintern-1B [46] là mô hình đơn ngữ tiếng Việt đạt hiệu suất cao trên các thang đo BLEU, METEOR và ROUGE_L nhưng lại kém trên thang đo CIDEr, trong khi mô hình Qwen2.5 VL [44] lại đạt điểm cao hơn trên thang đo này. Đối với hướng tiếp cận Chú thích hình ảnh dựa trên đặc trưng OCR, các mô hình thể hiện hiệu

suất ở mức trung bình. Trong đó, DEVICE [22] và Anchor_captioner [23] có kết quả tương đối ổn định. Cả hai mô hình đều sử dụng các cơ chế tập trung giúp mô hình hóa thành công các mối quan hệ phức tạp, từ đó tạo ra các câu chú thích có độ chính xác cao về mặt cấu trúc lẫn nội dung. Các mô hình cơ sở theo hướng Tóm tắt văn bản trừu tượng với kiến trúc được tối ưu hóa riêng cho tiếng Việt thể hiện ưu thế vượt trội so với mô hình đa ngữ. Trong nhóm các mô hình đơn ngữ, dòng mô hình BARTpho [30] ghi nhận những kết quả ấn tượng nhất ở cả hai phương pháp mã hóa. Sự vượt trội này bắt nguồn từ cơ chế khử nhiễu kết hợp với việc tiền huấn luyện trên tập dữ liệu báo chí tiếng Việt, giúp mô hình đạt độ tương thích cao với phong cách ngôn ngữ báo chí. Đặc biệt, việc xử lý theo đơn vị âm tiết (syllable) giúp $\text{Bartpho}_{\text{syllable}}$ phản ánh chính xác cấu trúc hình thái đặc trưng của tiếng Việt, từ đó tạo ra bản tóm tắt có độ mạch lạc và tính trừu tượng tốt hơn trong khi $\text{BARTpho}_{\text{word}}$ có sự trùng khớp cao trong việc sử dụng từ với câu tham vấn do bộ mã hóa theo từ dẫn đến BLEU-4 và METEOR của mô hình này đạt điểm số cao. Phương pháp đề xuất của nhóm chúng tôi đạt hiệu suất tổng thể vượt trội hơn so với các mô hình ở các hướng tiếp cận trước đó do tận dụng mô hình BARTpho cùng với các cơ chế tập trung từ cả đặc trưng hình ảnh và thị giác. Điểm số ROUGE_L và CIDEr của chúng tôi là SOTA trong các mô hình hiện tại.

Bảng 6.4: Bảng kết quả chính của các mô hình

Hướng tiếp cận	**Mô hình**	**Phương thức đầu vào**	**BLEU-4**	**METEOR**	**ROUGE_L**	**CIDEr**
OCR-based Image Captioning	Anchor_captioner [23]	Đặc trưng vật thể. Đặc	8.41	18.43	22.13	51.52
	LCM_captioner [53]		3.75	18.70	17.28	15.49
	M4C_captioner		3.94	19.18	18.19	16.44

	[13]	trưng OCR.				
	DEVICE [22]		9.49	17.88	23.70	43.45
	LSTM-R [57]		7.89	16.86	18.91	9.76
Abstractive Text Summarization	$Bartpho_{word}$ [30]	Văn bản được trích xuất trong ảnh.	**40.62**	**34.61**	55.26	202.82
	$Bartpho_{syllable}$ [30]		35.76	31.3	58.23	286.94
	ViT5_base [29]		22.64	22.77	36.64	118.76
	MT5_base [73]		13.26	19.13	40.33	91.72
LMMs	BLIP2 [42]	Ảnh.	0.00	4.40	2.43	0.64
	InstructBLIP [43]		0.00	4.74	2.70	0.74
	InternVL 2.5 [64]		0.25	13.57	4.57	0.11
	Vintern-1B [46]		1.56	22.51	10.33	1.06
	Qwen2.5VL [44]		1.14	18.28	9.48	3.24
	LLava-NeXT [45]		0.05	14.35	3.72	0.00
	LaVy [47]		0.00	14.13	4.65	0.05
Abstractive Captioning	ViAbsCaptioner (đề xuất)	Văn bản được trích xuất trong ảnh. Đặc trưng bố cục.	35.42	31.48	**59.26**	**306.34**

6.3.2. Phân tích cắt bỏ đối với phương pháp đề xuất

Để đánh giá chi tiết đóng góp của từng thành phần trong kiến trúc, nhóm đã tiến hành thực nghiệm cắt bỏ và trình bày kết quả tại *Bảng 6.5*. Qua đó, vai trò của hai mô-đun then chốt được phân tích cụ thể như sau:

- **Hiệu quả của hàm mất mát Triplet Contrastive Loss:**

  Việc tích hợp hàm mất mát Triplet Contrastive Loss mang lại sự cải thiện vượt trội về chất lượng ngữ nghĩa cho mô hình. So với mô hình cơ sở ($Bartpho_{syllable}$), khi chỉ thêm mô-đun này, tất cả các chỉ số đều tăng, trong đó điểm CIDEr ghi nhận mức tăng mạnh nhất là +11.58 (từ 286.94 lên 305.58). Điều này chứng tỏ Triplet Contrastive Loss giúp mô hình học được các đặc trưng phân biệt tốt hơn, tối ưu hóa khả năng so khớp giữa nội dung hình ảnh và văn bản mô tả, từ đó tạo ra những câu mô tả có độ chính xác về mặt nội dung rất cao.

- **Vai trò của cơ chế SpatialOverlap Attention:**

  Cơ chế SpatialOverlap Attention đóng vai trò then chốt trong việc nâng cao độ chính xác về từ vựng và cấu trúc câu. Khi kích hoạt cơ chế này, mô hình đạt được điểm số cao nhất ở hai chỉ số quan trọng là BLEU-4 (36.05) và METEOR (31.56). Kết quả này cho thấy việc tập trung vào sự chồng lấp không gian giữa các đối tượng giúp mô hình nắm bắt tốt hơn các mối quan hệ thực thể, từ đó cải thiện khả năng lựa chọn từ ngữ chính xác và duy trì sự trôi chảy của văn bản.

- **Sự kết hợp cả hai mô đun:**

  Khi kết hợp đồng thời cả SpatialOverlap Attention và Triplet Contrastive Loss, mô hình đạt được kết quả ấn tượng nhất ở các chỉ số đánh giá độ tương đồng ngữ nghĩa sâu. Cụ thể, điểm ROUGE_L đạt mức cao nhất (59.26) và điểm CIDEr đạt đỉnh với 306.34 (tăng mạnh +19.4 so với mô hình cơ sở). Mặc dù có sự sụt giảm nhẹ ở chỉ số BLEU-4 so với khi chỉ dùng riêng lẻ cơ chế Attention, nhưng sự tăng lên vượt bậc của CIDEr cho thấy sự kết hợp

này giúp mô hình không chỉ dừng lại ở việc khớp từ ngữ mà còn thực sự hiểu và tái hiện lại cấu trúc nội dung phức tạp. Sự bổ trợ qua lại giữa việc nắm bắt quan hệ không gian và việc tối ưu hóa không gian đặc trưng đã tạo ra những câu mô tả có tính toàn diện và giàu thông tin nhất.

Bảng 6.5: Bảng kết quả phân tích cắt bỏ thành phần của mô hình đề xuất

Mô-đun		**Độ đo**			
SpatialOverlap Attention	Triplet Contrastive Loss	BLEU-4	METEOR	ROUGE_L	CIDEr
x	x	35.76	31.3	58.23	286.94
x	✓	35.42 (↓-0.34)	31.38 (↑+0.08)	59.11 (↑+0.88)	305.58 (↑+18.64)
✓	x	**36.05** (↑+0.29)	**31.56** (↑+0.26)	58.76 (↑+0.53)	288.81 (↑+1.87)
✓	✓	35.42 (↓-0.34)	31.48 (↑+0.18)	**59.26** (↑+1.03)	**306.34** (↑+19.4)

6.3.3. Phân tích ảnh hưởng của tham số α đến Triplet Contrastive Loss

Dựa trên kết quả thực nghiệm, việc điều chỉnh tham số α tương thích với đặc điểm dữ liệu là vô cùng quan trọng. Với α=0.3, hàm mất mát tương phản đóng góp không đáng kể vào tổng thể, dẫn đến hạn chế trong việc tối ưu hóa thực thể. Khi tăng α=0.7, sự can thiệp quá sâu của hàm này làm mất cân bằng với Cross Entropy Loss, khiến nội dung tạo sinh kém mạch lạc do quá chú trọng vào các thực thể riêng lẻ. Giá trị α=0.5 cho thấy sự hài hòa, cao nhất ở hai chỉ số quan trọng CIDEr (305.58) và ROUGE-L (59.11), cho phép mô hình tận dụng hiệu quả hàm học tương phản trong khi vẫn đảm bảo chất lượng và độ tự nhiên của văn

bản đầu ra. Vì vậy, 0.5 là tham số tối ưu cho Triplet Contrastive Loss mà nhóm lựa chọn.

Bảng 6.6: Bảng kết quả thử nghiệm trên các tham số α khác nhau

Tham số α	**Độ đo**			
	BLEU-4	METEOR	ROUGE_L	CIDEr
0.3	34.87	31.02	57.49	295.29
0.5	**35.42**	**31.38**	**59.11**	**305.58**
0.7	35.19	31.08	58.09	296.63

6.3.4. Phân tích các lỗi được khắc phục ở phương pháp đề xuất

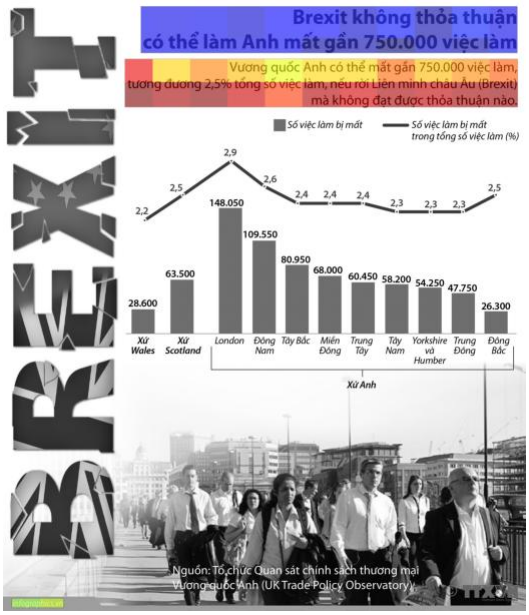
Phương pháp mà nhóm đề xuất sử dụng hai mô-đun nhằm xử lý các vấn đề hiện có trong mô hình cơ sở. Để hiểu rõ hơn sự cải tiến của mô hình, nhóm đã kiểm tra câu trả lời giữa mô hình cơ sở $\text{Bartpho}_{\text{syllable}}$ và mô hình đề xuất. Từ đó, phân tích được ưu điểm vượt trội của mô hình ViAbsCaptioner.

Một trong những lỗi phổ biến nhất của mô hình cơ sở đang gặp phải là khả năng tạo sinh thông tin sai so với câu mô tả tham chiếu. Mô hình cơ sở, không có sự nhận biết về các bố cục quan trọng, thất bại trong việc nhận diện được nội dung, đồng thời tạo sinh ra các thực thể riêng sai lệch cho với thực thể tham chiếu. Chẳng hạn như mẫu dữ liệu minh họa ở *Bảng 6.7*, câu trả lời của mô hình cơ sở "*từ ngày 20 đến ngày 31/6/1993*" sai thông tin hoàn toàn khi so với câu mô tả tham chiếu "*từ ngày 12 đến 20/6/1993*". Mô hình đề xuất khắc phục được vấn đề đó.

Bảng 6.7: Mẫu dữ liệu khắc phục lỗi ngày tháng

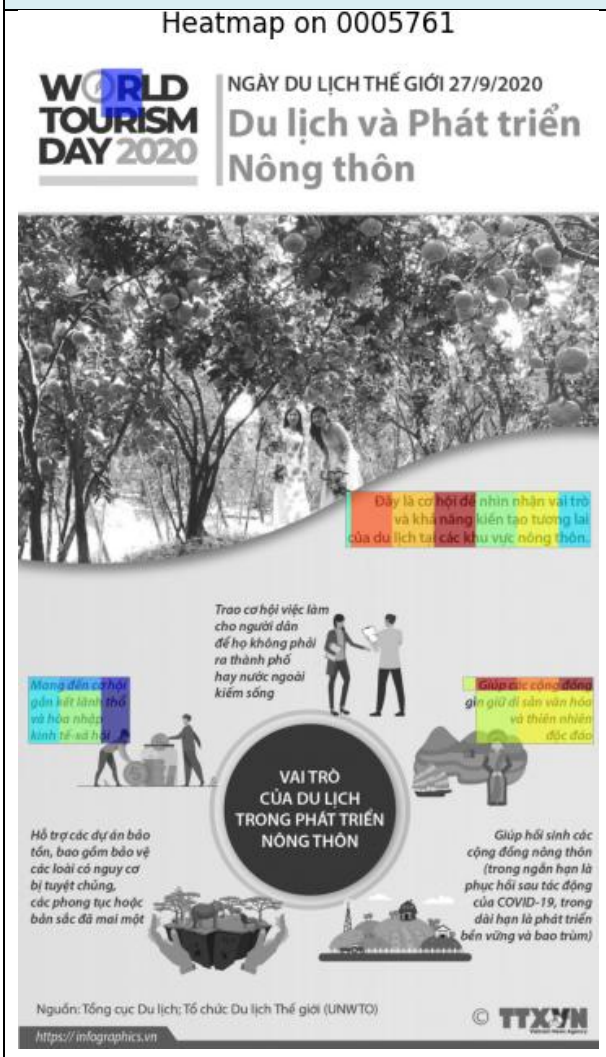
Hình ảnh	Chú thích tóm tắt
	Ground Truth: *sea games 17 tổ chức tại singapore từ ngày 12 đến 20/6/1993 với 29 môn thi đấu chính thức . indonesia đứng đầu bảng tổng sắp với 253 huy chương .* Baseline: *sea games 17 tổ chức tại singapore từ ngày 20 đến ngày 31/6/1993 với 29 môn thi đấu chính thức . indonesia đứng đầu bảng tổng sắp với 253 huy chương,* ViAbsCaptioner: *sea games 17 tổ chức tại singapore từ ngày 12 đến 20/6/1993 với 29 môn thi đấu chính thức . indonesia đứng đầu bảng tổng sắp với 253 huy chương .*

Bảng 6.8: Mẫu dữ liệu khắc phục lỗi số liệu

Hình ảnh	Chú thích tóm tắt
	Ground Truth: *vương quốc anh có thể mất gần 750.000 việc làm tương đương 25 tổng số việc làm nếu rời liên minh châu âu brexit mà không đạt được thoả thuận nào .* Baseline: *vương quốc anh có thể mất gần 750 việc làm nếu rời liên minh châu âu brexit mà không đạt được thoả thuận nào tương đương 25 tổng số việc làm .* ViAbsCaptioner: *vương quốc anh có thể mất gần 750.000 việc làm nếu rời liên minh châu âu brexit mà không đạt được thoả thuận nào tương đương 25 tổng số việc làm .*

Tại ví dụ ở *Bảng 6.8*, câu chú thích của mô hình cơ sở "*750 việc làm*" sai thông tin hoàn toàn khi so với câu mô tả tham chiếu và của chúng tôi là "*750.000 việc làm*". Dựa vào biểu đồ nhiệt, mô hình của chúng tôi chú ý đến những bố cục mang nội dung chính, đồng thời là những chi tiết quan trọng cho câu mô tả như số lượng, thời gian, địa điểm, … Đồng thời, hàm học tương phản cho phép mô hình chỉnh sửa, tăng độ tương đồng giữa các thực thể trong câu mô tả tham chiếu và câu mô tả dự đoán, giúp cho thực thể riêng trong câu được tạo sinh chính xác hơn.

Bảng 6.9: Mẫu dữ liệu khắc phục lỗi tập trung sai bố cục

Hình ảnh	Chú thích tóm tắt
	Ground Truth: *chủ đề của ngày du lịch thế giới 27/9/2020 là du lịch và phát triển nông thôn . ngày này là cơ hội để nhìn nhận vai trò và khả năng kiến tạo tương lai của du lịch tại các khu vực nông thôn .* Baseline: *ngày 27/9/2020 là ngày du lịch thế giới thứ 2 dành cho du lịch nông thôn . với chủ đề tourism du lịch và phát triển nông thôn năm 2020 ngày du lịch thế giới 27/9/2020 du lịch và* ViAbsCaptioner: *ngày du lịch thế giới 27/9/2020 có chủ đề du lịch và phát triển nông thôn . đây là cơ hội để nhìn nhận vai trò và khả năng kiến tạo tương lai của du lịch tại các khu vực nông thôn .*

Mô hình cơ sở tuy có thể nắm bắt được sự kiện chính dựa vào nội dung mô tả đầu vào tốt, nhưng khó khăn trong việc nắm bắt được các bố cục chính trong hình. Dựa vào biểu đồ nhiệt, mô hình của chúng tôi có khả năng chú ý đến những bố cục mang nội dung chính, đồng thời là những chi tiết quan trọng cho câu mô tả như số lượng, thời gian, địa điểm. Trong *Bảng 6.9*, vế thứ hai của câu trả lời của mô hình cơ sở "*với chủ đề tourism du lịch và phát triển nông thôn năm 2020 ngày*

du lịch thế giới 27/9/2020 du lịch và" không phải là nội dung cần được chú trọng để đưa vào câu mô tả. Nhờ vào mô-đun SpatialOverlapAttention, nội dung phần bố cục quan trọng "*đây là cơ hội để nhìn nhận vai trò và khả năng kiến tạo tương lai của du lịch tại các khu vực nông thôn*" được chú ý hơn, góp phần cải thiện chất lượng của câu tạo sinh.

6.3.5. Phân tích mức độ bao phủ thực thể của phương pháp đề xuất

Bảng 6.10: Bảng so sánh độ chính xác trong nhận diện thực thể

		Accuracy	**Precision**	**Recall**	**F1**
Baseline	*Micro*	30.8	59.85	38.82	47.09
ViAbsCaptioner		31.0 (↑+0.2)	60.95 (↑+1.1)	38.68 (↓-0.14)	47.33 (↑+0.24)
Baseline	*Macro*	46.19	55.25	46.64	48.14
ViAbsCaptioner		46.43 (↑+0.24)	56.03 (↑+0.78)	46.56 (↓-0.08)	48.48 (↑+0.34)

Khi phân tích độ trùng khớp giữa các cặp thực thể - tên thực thể trong câu dự đoán so với câu mô tả tham chiếu, phương pháp đề xuất ViAbsCaptioner của nhóm cho thấy khả năng bám sát nội dung gốc tốt hơn so với mô hình cơ sở. Việc cải thiện đồng thời ở các chỉ số Accuracy và F1 minh chứng rằng mô hình mới đã giảm thiểu hiệu quả hiện tượng ảo giác và tăng cường khả năng nhận diện chính xác các thực thể quan trọng từ dữ liệu đầu vào. Mặt dù độ bao phủ Recall có giảm nhẹ, nhưng mức độ bao phủ thực thể vẫn được duy trì ở ngưỡng tương đương với Baseline. Sự đánh đổi nhỏ về độ phủ để đổi lấy độ chính xác cao hơn đã giúp mô hình bảo toàn trọn vẹn giá trị thông tin và ngữ nghĩa cốt lõi của dữ liệu gốc. Sự cải thiện ở thang đo Precision cũng khẳng định rằng ViAbsCaptioner không chỉ bắt được nhiều thực thể đúng hơn mà còn đảm bảo mỗi thực thể được đưa vào đều có độ tin cậy cao, từ đó bảo toàn trọn vẹn thông tin và ngữ nghĩa của hình ảnh ban đầu.

Chương 7. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

7.1. Đóng góp

Đề tài mà nhóm thực hiện tập trung giải quyết bài toán tạo sinh tóm tắt văn bản trừu tượng từ Infographic tiếng Việt, đóng góp cho nguồn dữ liệu truyền thông đa phương tiện mới ngày càng phổ biến trong truyền thông báo chí xã hội, nhưng đầy thách thức cho các mô hình đa thể thức.

Nhóm đã tiền xử lý và chuẩn hóa thành công 17.840 cặp dữ liệu hình ảnh - mô tả chất lượng cao, được thu thập từ nguồn Tin Đồ họa của Thông tấn xã Việt Nam. Đây là bộ dữ liệu đa thể thức đầu tiên tại Việt Nam chuyên biệt cho tác vụ tóm tắt trừu tượng với bố cục phức tạp, có tính chuẩn xác cao về số liệu và ngôn ngữ báo chí.

Nhóm đã thực hiện các phân tích thống kê chi tiết về phân phối lĩnh vực, độ dài văn bản, độ đa dạng từ vựng và mối tương quan giữa OCR và chú thích. Kết quả phân tích cho thấy dữ liệu có mật độ thực thể dày đặc, tạo cơ sở tri thức lớn cho các mô hình học máy.

Cùng với việc đánh giá bộ dữ liệu, nhóm đề xuất ra phương pháp mới giúp mô hình hiểu được tương quan không gian bố cục và hàm mất mát tương phản với vai trò nhằm tăng cường độ chính xác cho các thực thể và số liệu trong câu tóm tắt, giúp giảm thiểu hiện tượng ảo giác gây ra ở các mô hình ngôn ngữ.

7.2. Hạn chế

Mặc dù đạt được những kết quả khả quan ở bước đầu nghiên cứu trên bài toán này, vẫn tồn tại một số hạn chế nhất định cần được xem xét.

Trước hết là vấn đề thiếu nhất quán trong quá trình tích hợp các đặc trưng thông tin. Do quá trình trích xuất đặc trưng sử dụng đồng thời nhiều mô hình cho các đặc trưng khác nhau, việc căn chỉnh giữa các vectơ đặc trưng đôi khi xuất hiện nhiễu và thiếu sự đồng bộ, gây ảnh hưởng trực tiếp đến khả năng hội tụ cũng như hiệu suất tổng thể của mô hình.

Bên cạnh đó, hiện tượng lỗi lan truyền vẫn là một thách thức lớn khi mô hình phụ thuộc chặt chẽ vào chất lượng đầu ra của các thành phần tiền xử lý. Những sai sót phát sinh trong bước trích xuất đặc trưng hoặc nhận diện văn bản nếu không được kiểm soát tốt sẽ dẫn đến những sai lệch về nội dung trong kết quả tóm tắt cuối cùng.

Cuối cùng, vẫn còn tồn tại khoảng cách giữa thực nghiệm và khả năng ứng dụng thực tế. Nghiên cứu hiện tại chủ yếu tập trung đánh giá trên tập dữ liệu mà chúng tôi đề xuất, do đó việc triển khai trong hệ thống thực đòi hỏi cần thêm nhiều thử nghiệm, thực nghiệm, cũng như tối ưu hóa về mặt tài nguyên tính toán.

7.3. Hướng phát triển

Để khắc phục các hạn chế hiện có và mở rộng khả năng ứng dụng của đề tài, nhóm đề xuất các hướng phát triển trọng tâm trong tương lai.

Thứ nhất, nghiên cứu sẽ thực hiện các đánh giá chuyên sâu về vai trò của từng thành phần cụ thể, đặc biệt là cơ chế SpatialOverlap Attention và hàm mất mát tương phản Triplet Contrastive Loss trên nhiều nhóm lĩnh vực khác nhau. Việc này nhằm làm rõ nguyên nhân giúp phương pháp đề xuất đạt hiệu quả vượt trội hơn so với các mô hình cơ sở thông thường.

Thứ hai, nhóm hướng tới việc tinh chỉnh các mô hình ngôn ngữ thị giác lớn trên bộ dữ liệu đồ họa thông tin tiếng Việt chuyên biệt để tận dụng tối đa khả năng hiểu ngữ cảnh sâu và khối lượng tri thức phong phú đã được tích lũy từ các tập dữ liệu quy mô lớn, từ đó nâng cao độ chính xác, tính nhất quán và khả năng tóm tắt linh hoạt cho mô hình.

TÀI LIỆU THAM KHẢO

[1] Z. Liu, "Reading behavior in the digital environment: Changes in reading behavior over the past ten years," *Journal of Documentation*, vol. 61, pp. 700–712, 2005.

[2] M. Mathew, V. Bagal, R. Tito, D. Karatzas, E. Valveny, and C. V. Jawahar, "InfographicVQA," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022.

[3] N. Landman, "Towards Abstractive Captioning of Infographics," 2018.

[4] M. Smiciklas, "The Power of Infographics: Using Pictures to Communicate and Connect with Your Audiences," 2012.

[5] S. Heller and N. Holmes, "Nigel Holmes on Information Design," in Working Biography Series. New York: Jorge Pinto Books, 2006.

[6] K. Q. Tran, A. T. Nguyen, A. T.-H. Le, and V. K. Nguyen, "ViVQA: Vietnamese Visual Question Answering," *Proceedings of the Pacific Asia Conference on Language, Information and Computation*, pp. 683–691, 2021.

[7] N. H. Nguyen, D. T. D. Vo, V. K. Nguyen, and N. L.-T. Nguyen, "OpenViVQA: Task, Dataset, and Multimodal Fusion Models for Visual Question Answering in Vietnamese," *Information Fusion*, vol. 100, p. 101868, 2023.

[8] Q. V. Nguyen *et al.*, "ViTextVQA: A Large-Scale Visual Question Answering Dataset for Evaluating Vietnamese Text Comprehension in Images," 2025, [Online]. Available: https://arxiv.org/abs/2404.10652

[9] Q. H. Lam, Q. D. Le, V. K. Nguyen, and N. L.-T. Nguyen, "UIT-ViIC: A Dataset for the First Evaluation on Vietnamese Image Captioning," *International Conference on Computational Collective Intelligence*, pp. 730–742, 2020.

[10] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," *European Conference on Computer Vision*, pp. 740–755, 2014.

[11] T.-T. Van-Dinh, H.-D. Tran, T.-B. Duong, M.-H. Pham, B.-N. Le-Nguyen, and Q.-T. Nguyen, "ViInfographicVQA: A Benchmark for Single and Multi-image Visual Question Answering on Vietnamese Infographics," 2025, [Online]. Available: https://arxiv.org/abs/2512.12424

[12] X. Chen *et al.*, "Microsoft COCO Captions: Data Collection and Evaluation Server," *arXiv preprint arXiv:1504.00325*, 2015.

[13] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, "TextCaps: A Dataset for Image Captioning with Reading Comprehension," *European Conference on Computer Vision*, pp. 742–758, 2020.

[14] S. Kantharaj *et al.*, "Chart-to-Text: A Large-Scale Benchmark for Chart Summarization," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 4005–4023, 2022.

[15] B. Tang, A. Boggust, and A. Satyanarayan, "VisText: A Benchmark for Semantically Rich Chart Captioning," *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 7268–7298, 2023.

[16] B. Wang, G. Li, X. Zhou, Z. Chen, T. Grossman, and Y. Li, "Screen2Words: Automatic Mobile UI Summarization with Multimodal Learning," *Proceedings of the ACM Symposium on User Interface Software and Technology*, pp. 498–510, 2021.
[17] C. Ngo *et al.*, "MTet: Multi-domain Translation for English and Vietnamese," *arXiv preprint arXiv:2210.05610*, 2022.
[18] K. Xu *et al.*, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," *Proceedings of the International Conference on Machine Learning*, pp. 2048–2057, 2015.
[19] P. Anderson *et al.*, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018.
[20] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer Networks," *Proceedings of the Neural Information Processing Systems*, 2017.
[21] R. Hu, A. Singh, T. Darrell, and M. Rohrbach, "Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
[22] D. Xu, Q. Huang, X. Zhang, H. Cheng, F. Shuang, and Y. Cai, "DEVICE: Depth and Visual Concepts Aware Transformer for OCR-based Image Captioning," *Pattern Recognition*, vol. 164, p. 111522, 2025.
[23] G. Xu, S. Niu, M. Tan, Y. Luo, Q. Du, and Q. Wu, "Towards Accurate Text-Based Image Captioning with Content Diversity Exploration," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12637–12646, 2021.
[24] Z. Yang *et al.*, "TAP: Text-Aware Pre-training for Text-VQA and Text-Caption," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8747–8757, 2020.
[25] A. F. Biten, R. Litman, Y. Xie, S. Appalaraju, and R. Manmatha, "LaTr: Layout-Aware Transformer for Scene-Text VQA," *LaTr: Layout-Aware Transformer for Scene-Text VQA*, 2022.
[26] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *arXiv preprint arXiv:1910.10683*, vol. 21, 2020.
[27] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-training with Extracted Gap-Sentences for Abstractive Summarization," *arXiv preprint arXiv:1912.08777*, 2020.
[28] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," *ACL*, pp. 7871--7880, 2020, doi: 10.18653/v1/2020.acl-main.703.
[29] L. Phan, H. Tran, H. Nguyen, and T. H. Trinh, "ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation," *Proceedings of the North*

American Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 136–142, 2022.
[30] N. L. Tran, D. M. Le, and D. Q. Nguyen, "BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese," *Proceedings of the Annual Conference of the International Speech Communication Association*, 2022.
[31] Z.-Y. Dou, P. Liu, H. Hayashi, Z. Jiang, and G. Neubig, "GSum: A General Framework for Guided Neural Abstractive Summarization," *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 2021.
[32] A. Vaswani *et al.*, "Attention Is All You Need," *NIPS'17*, pp. 6000–6010, 2017.
[33] A. Berg, M. O'Connor, and M. T. Cruz, "Keyword Transformer: A Self-Attention Model for Keyword Spotting," *Interspeech*, 2021.
[34] T. T. Wang, I. Zablotchi, N. Shavit, and J. S. Rosenfeld, "Cliff-Learning," 2023.
[35] Y. Liu, P. Liu, D. Radev, and G. Neubig, "BRIO: Bringing Order to Abstractive Summarization," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2890--2903, 2022, doi: 10.18653/v1/2022.acl-long.207.
[36] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, and M. Johnson, "VnCoreNLP: A Vietnamese Natural Language Processing Toolkit," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 56–60, 2018, doi: 10.18653/v1/N18-5012.
[37] PaddlePaddle Contributors, "PaddleOCR: An Ultra Lightweight Optical Character Recognition System," 2020.
[38] Q. Pham, "VietOCR: A Fast and Accurate Optical Character Recognition Toolkit for Vietnamese," 2020.
[39] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking," *Proceedings of the ACM International Conference on Multimedia*, pp. 4083–4091, 2022.
[40] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of Text and Layout for Document Image Understanding," *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1192–1200, Aug. 2020, doi: 10.1145/3394486.3403172.
[41] A. Dosovitskiy, "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR*, 2021.
[42] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," *Proceedings of the International Conference on Machine Learning*, 2023.
[43] W. Dai, J. Li, D. Li, and S. Hoi, "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning," *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

[44] J. Bai *et al.*, "Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities," *arXiv preprint arXiv:2308.12966*, 2023.
[45] H. Liu *et al.*, "LLaVA-NeXT: Improved Reasoning, OCR, and World Knowledge," 2024.
[46] K. T. Doan *et al.*, "Vintern-1B: An Efficient Multimodal Large Language Model for Vietnamese," *arXiv preprint arXiv:2408.12480*, 2024.
[47] C. Tran and H. Le Thanh, "LaVy: Vietnamese Multimodal Large Language Model," *arXiv preprint arXiv:2404.07922*, 2024.
[48] S. Zhang, H. Cheng, J. Gao, and H. Poon, "Optimizing Bi-Encoder for Named Entity Recognition via Contrastive Learning," *ICLR 2023*, 2023.
[49] D. Q. Nguyen and A.-T. Nguyen, "PhoBERT: Pre-trained Language Models for Vietnamese," *Findings of the Association for Computational Linguistics: EMNLP*, pp. 1037–1042, 2020.
[50] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.
[51] C. Fang, J. Li, L. Li, C. Ma, and D. Hu, "Separate and Locate: Rethink the Text in Text-based Visual Question Answering," 2023, doi: 10.48550/ARXIV.2308.16383.
[52] L. T. Nguyen and D. Q. Nguyen, "PhoNLP: A joint multi-task learning model for Vietnamese part-of-speech tagging, named entity recognition and dependency parsing," 2021, doi: 10.48550/ARXIV.2101.01476.
[53] Q. Wang *et al.*, "LCM-Captioner: A Lightweight Text-Based Image Captioning Method with Collaborative Mechanism Between Vision and Text," *Neural Networks*, vol. 162, pp. 318–329, 2023.
[54] G. Jocher, J. Qiu, and A. Chaurasia, "Ultralytics YOLO," 2023.
[55] M. Huang *et al.*, "SwinTextSpotter: Scene Text Spotting via Better Synergy between Text Detection and Text Recognition," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4593–4603, 2022.
[56] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Advances in Neural Information Processing Systems*, 2015.
[57] J. Wang, J. Tang, M. Yang, X. Bai, and J. Luo, "Improving OCR-based Image Captioning by Incorporating Geometrical Relationship," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1306–1315, 2021.
[58] C. B. Vennerød, A. Kjærran, and E. S. Bugge, "Long Short-Term Memory RNN," *arXiv preprint arXiv:2105.06756*, 2021.
[59] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From Big to Small: Multi-scale Local Planar Guidance for Monocular Depth Estimation," *arXiv preprint arXiv:1907.10326*, 2019.

[60] A. Radford *et al.*, “Learning Transferable Visual Models From Natural Language Supervision,” *Proceedings of the International Conference on Machine Learning*, 2021.
[61] L. Xue *et al.*, “mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer,” *arXiv preprint arXiv:2010.11934*, 2021.
[62] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.
[63] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv preprint arXiv:1409.1556*, 2015.
[64] W. Wang *et al.*, “InternVL: Scaling Up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks,” *Large Vision-Language Models: Pre-training, Prompting, and Applications*, pp. 23–57, 2025.
[65] Z. Yang *et al.*, “The Dawn of LMMs: Preliminary Explorations with GPT-4V,” *arXiv preprint arXiv:2309.17421*, 2023.
[66] InternLM Team, “InternLM: A Multilingual Language Model with Progressively Enhanced Capabilities,” 2023.
[67] H. Touvron *et al.*, “LLaMA: Open and Efficient Foundation Language Models,” *arXiv preprint arXiv:2302.13971*, 2023.
[68] D. Q. Nguyen, L. T. Nguyen, C. Tran, D. N. Nguyen, D. Phung, and H. Bui, “PhoGPT: Generative Pre-training for Vietnamese,” *arXiv preprint arXiv:2311.02945*, 2024.
[69] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A Method for Automatic Evaluation of Machine Translation,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.
[70] A. Lavie and M. J. Denkowski, “The METEOR Metric for Automatic Evaluation of Machine Translation,” *Machine Translation*, vol. 23, pp. 105–115, 2009.
[71] A. A. Citarella, M. Barbella, M. G. Ciobanu, F. De Marco, L. Di Biasi, and G. Tortora, “Rouge Metric Evaluation for Text Summarization Techniques,” 2024.
[72] R. Vedantam, C. L. Zitnick, and D. Parikh, “CIDEr: Consensus-based Image Description Evaluation,” pp. 4566–4575, 2015.
[73] L. Xue *et al.*, “mT5: A massively multilingual pre-trained text-to-text transformer,” 2020, doi: 10.48550/ARXIV.2010.11934.