Mô Hình Ngôn Ngữ Thị Giác Và Ứng Dụng

Phan Hữu Đoàn Anh, Lưu Minh Hoàng, Nguyễn Mạnh Hùng[1]

[1]Đại học Khoa học Tự nhiên TP. Hồ Chí Minh

Tháng 2 năm 2024

Tóm tắt nội dung

Lĩnh vực giao thoa giữa Trí tuệ nhân tạo và Thị giác máy tính đã lâu đấu tranh với thách thức về nhận dạng hình ảnh và văn bản, đánh dấu một phương pháp cơ bản để giải quyết các vấn đề phức tạp liên quan đến thị giác và mở ra các ứng dụng mới như lái xe tự động và giám sát an ninh. Trong số đó, nghiên cứu nhận dạng hình ảnh truyền thống thường phụ thuộc vào các tập dữ liệu gán nhãn cộng đồng để huấn luyện các Mạng Nơ-ron Sâu (DNNs). Hơn nữa, việc huấn luyện DNNs thường đòi hỏi nỗ lực dành riêng để giải quyết các vấn đề cụ thể, dẫn đến việc tạo ra các mô hình nhận dạng hình ảnh tốn nhiều công sức và thời gian. Do đó, để giải quyết các vấn đề này, Mô hình Ngôn ngữ Hình ảnh (VLM) đã và đang được nghiên cứu một cách sâu sắc với mục tiêu là tạo ra một mô hình có thể hiểu được sự tương quan giữa hình ảnh và văn bản từ các nguồn ảnh đầu vào là các cặp ảnh-văn bản có được trên mạng, loại nguồn mà hầu như là vô tận, và đồng thời cho phép các thực hiện các dự đoán nhận dạng zero-shot trên nhiều bài toán nhận dạng hình ảnh khác nhau chỉ bằng một mô hình VLM. Tuy với nhiều nỗ lực nghiên cứu hiện tại, ta vẫn mắc phải một số giới hạn khác nhau khi nghiên cứu VLM như: học hiểu đa phương tiện, truy xuất hình ảnh từ văn bản tự nhiên, khả năng mở rộng, chuẩn đoán kết quả hình ảnh. Đối với một số vấn đề nêu trên thì đã có một số giải pháp như sử dụng các mô hình học đa ngữ như BERT, sử dụng các mô hình visual đã được huấn luyện trước, thiết kế các kiến trúc mới nhẹ hơn. Qua bài báo này, nhóm chúng tôi sẽ trình bày những nghiên cứu tìm hiểu được qua các bài báo mới hiện nay về chủ đề VLM, sau đó mô tả các mô hình, giải thích kết quả và đưa ra các kết luận liên quan.

1 Giới Thiệu

Trong bối cảnh của cuộc cách mạng dữ liệu và AI hiện đại, Vision-Language Models (VLMs)[23] đã nổi lên như một lĩnh vực nghiên cứu hứa hẹn, đánh dấu sự hội tụ giữa thị giác máy tính và xử lý ngôn ngữ tự nhiên. Động lực chính cho sự phát triển này bắt nguồn từ nhu cầu giải quyết các bài toán phức tạp như chuyển đổi hình ảnh thành văn bản, phân đoạn ngữ nghĩa hình ảnh, và nhiều tác vụ nhận dạng thị giác khác. Điều này không chỉ mở rộng khả năng của máy móc trong việc hiểu và tương tác với thế giới xung quanh, mà còn giúp tạo ra các giải pháp mới cho các vấn đề thực tế, như hỗ trợ người khiếm thị, phát triển trợ lý ảo thông minh hơn, và tự động hóa quy trình làm việc. VLMs đặc biệt quan trọng trong việc mô phỏng cách con người tương tác và hiểu thế giới qua cả hình ảnh và ngôn ngữ. Mô hình hóa sự tương quan giữa dữ liệu hình ảnh và văn bản không chỉ yêu cầu một hiểu biết sâu sắc về cả hai miền này, mà còn đòi hỏi khả năng tích hợp và xử lý thông tin đa modal một cách mạch lạc. VLMs tiên phong trong việc khám phá cách các mô hình có thể tự động tạo ra mô tả hình ảnh chi tiết, dịch văn bản thành cảnh quan hình ảnh, hoặc thậm chí phát hiện và phân loại đối tượng trong các tình huống phức tạp, giúp máy móc hiểu và phản ứng với thế giới thị giác giống như

con người.

Trong thời đại công nghệ thông tin phát triển vượt bậc, việc hợp nhất hình ảnh và ngôn ngữ trở nên quan trọng, điển hình qua các mô hình Vision-Language (VLMs). Khác biệt rõ ràng với Generative Adversarial Networks (GANs), VLMs đưa ra phương pháp tiếp cận hợp nhất độc đáo, không chỉ sinh ra hình ảnh mới mà còn có khả năng hiểu và diễn đạt thông tin thị giác một cách chính xác. Quá Trình Mã Hóa Thông Tin Thị Giác và Ngôn Ngữ VLMs bắt đầu với hai luồng xử lý chính: xử lý hình ảnh và xử lý ngôn ngữ. Trong quá trình "Image Processing", hình ảnh được chuẩn hóa và mã hóa thông tin thị giác thông qua phương pháp "Visual Encoding", sử dụng từ các tính năng CNN toàn cục cho đến các kỹ thuật tiên tiến hơn như chú ý cộng hưởng và tự chú ý. Mặt khác, "Language Processing" đề cập đến việc chia nhỏ và mã hóa ngôn ngữ, từ LSTM đơn giản đến các mô hình Transformer phức tạp. Kết Hợp Đa Phương Tiện và Không Gian Nhúng Chung Tiếp theo, "Multimodal Fusion" là bước quan trọng nơi thông tin từ "Visual Encoding" và "Text Encoding" được hợp nhất, tạo ra một không gian đặc trưng chung "Joint Embedding Space". Quá trình này đảm bảo sự tương tác chặt chẽ giữa hai loại thông tin, tối ưu cho các nhiệm vụ như tạo chú thích hình ảnh hoặc phân loại. Chiến Lược Huấn Luyện và Tối Ưu Hóa Mô Hình Phần "Training Strategies" của framework thể hiện sự đa dạng trong việc lựa chọn các chiến lược huấn luyện, từ hàm mất mát đối sánh đến mất mát chú thích, nhằm tối ưu hóa hiệu suất của mô hình. "Loss Function and Optimization" không chỉ giúp mô hình hóa sự hiểu biết ngôn ngữ mà còn tăng cường khả năng kết hợp giữa thông tin thị giác và ngôn ngữ. Đầu Ra và Các Chỉ Số Đánh Giá Cuối cùng, "Output" của mô hình là kết quả của quá trình học, được đánh giá thông qua "Evaluation Metrics" để đảm bảo chất lượng và độ chính xác. Các chỉ số này cho phép đánh giá hiệu quả của mô hình trong các tác vụ cụ thể và so sánh với các phương pháp tiếp cận khác như GANs.

2 Công Trình Nghiên Cứu Liên Quan

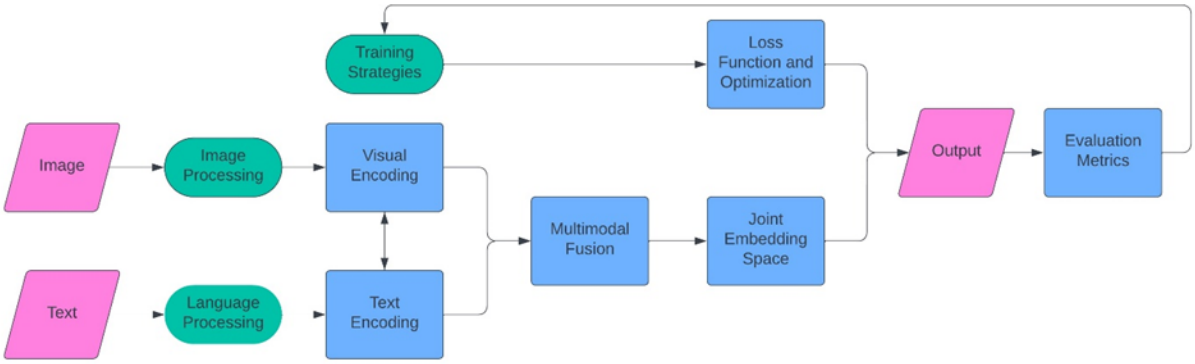
Phát triển các mô hình hợp nhất hình ảnh và ngôn ngữ (VLMs) là một trong những lĩnh vực nghiên cứu năng động nhất trong trí tuệ nhân tạo hiện đại. Nhiều công trình gần đây đã chú trọng vào việc mở rộng khả năng của VLMs thông qua việc tối ưu hóa cách thức huấn luyện và kiến trúc mô hình để xử lý các tác vụ phức tạp về ngôn ngữ và thị giác. Dưới đây, chúng tôi giới thiệu một số công trình quan trọng trong lĩnh vực này.

2.1 One-Shot Learning và SNNs: Cơ sở lý thuyết và Ứng dụng

One-shot learning và Siamese Neural Networks (SNNs)[6] đại diện cho một hướng tiếp cận học máy độc đáo, nơi một mô hình chỉ yêu cầu một ví dụ duy nhất từ mỗi lớp để học cách phân loại hoặc nhận dạng các thể hiện mới. Đây là một kỹ thuật quan trọng trong các tình huống có lượng dữ liệu giám sát hạn chế. Triplet Loss Function và kiến trúc SNN là cơ chế tối ưu cho One-shot Learning, giúp điều chỉnh các trọng số của mạng để phân biệt hiệu quả giữa các thể hiện dữ liệu.

2.1.1 Triplet Loss Function:

Cơ chế Tối Ưu cho One-shot Learning Một trong những phương pháp hiệu quả nhất trong việc thực hiện one-shot learning là sử dụng Triplet Loss Function, một hàm mất mát được thiết kế để huấn luyện một Siamese Neural Network (SNN). Phương pháp này liên quan đến việc sử dụng bộ ba hình ảnh: một hình ảnh "anchor", một hình ảnh "positive" tương đồng với anchor, và một hình ảnh "negative" khác biệt. Mục tiêu của hàm mất mát là điều chỉnh các trọng số của SNN sao cho khoảng cách giữa các đặc trưng của anchor và positive nhỏ hơn so với khoảng cách giữa anchor và negative.



Hình 1: Enter Caption

2.1.2 Siamese Neural Networks:

Kiến Trúc và Ứng dụng SNN[11] là kiến trúc mạng gồm hai hoặc nhiều mạng con giống hệt nhau, mỗi mạng có cùng cấu hình và chia sẻ trọng số, cho phép mô hình hóa sự tương đồng giữa các cặp dữ liệu. Quá trình này bao gồm việc chọn một cặp dữ liệu, đưa mỗi dữ liệu qua mạng để sinh ra các vector nhúng, và sau đó tính toán khoảng cách Euclidean giữa chúng. Một giá trị khoảng cách gần với 1 ám chỉ sự giống nhau cao giữa hai đầu vào, và ngược lại.

2.2 Few-Shot Learning: Kỹ Thuật và Thách Thức

Few-shot learning[15] là một nhánh của học máy mà ở đó mô hình được thiết kế để học từ một số lượng rất nhỏ ví dụ cho mỗi lớp. Điều này không chỉ giảm bớt nhu cầu về dữ liệu chú thích mà còn đặt ra các thách thức mới liên quan đến khả năng của mô hình để tổng quát hóa từ những thông tin hạn chế. Kỹ thuật trong Few-Shot Learning thường dựa vào các mô hình meta-learning hoặc transfer learning để "học cách học" từ dữ liệu, cải thiện hiệu suất trên các tác vụ với ít dữ liệu hơn.

2.3 Domain Adaptation và Out-of-Distribution Generalizations:

Domain Adaptation (DA) tập trung vào việc điều chỉnh mô hình được huấn luyện trên miền nguồn để nó có thể hoạt động hiệu quả trên miền mục tiêu, thường khác biệt so với miền nguồn về phân phối dữ liệu. Mục tiêu là giảm thiểu sự khác biệt giữa miền nguồn và miền mục tiêu, giúp mô hình đạt được hiệu suất tốt trên cả hai. Phương pháp này đặc biệt quan trọng trong các ứng dụng thực tế nơi mà việc thu thập dữ liệu được gán nhãn từ miền mục tiêu có thể khó khăn hoặc tốn kém.

Trong khi đó, OOD Generalization nhắm đến việc phát triển các mô hình có khả năng tổng quát hóa tốt trên dữ liệu từ phân phối không nhìn thấy trước đó trong quá trình huấn luyện. Điều này bao gồm việc thiết kế các mô hình và kỹ thuật huấn luyện có khả năng đối phó với sự không chắc chắn và sự biến đổi trong dữ liệu đầu vào, giúp mô hình duy trì hiệu suất cao ngay cả khi gặp dữ liệu mới hoặc bất ngờ.

2.4 Domain Generalization (DG)

Domain Generalization (DG) nhằm cải thiện khả năng tổng quát hóa của các mô hình đối với các miền mới không gặp trong quá trình huấn luyện [3, 2]. DG thường áp dụng các kỹ thuật như tìm kiếm các cực tiểu phẳng để cải thiện khả năng tổng quát hóa[2] và

học các đặc trưng gần với biểu diễn chuẩn nhất từ quan điểm của một người chuyên gia [3].

2.5 Unsupervised Domain Adaptation (UDA)

Trong kịch bản UDA, dữ liệu không được gán nhãn từ miền mục tiêu có sẵn trong giai đoạn huấn luyện cùng với dữ liệu được gán nhãn từ miền nguồn [24, 8]. Phương pháp này bao gồm cách tiếp cận như cải thiện căn chỉnh đặc trưng và tăng cường độ bền vững đối với nhãn giả , tạo ra một miền trung gian hiệu quả , và giới thiệu tổn thất nhầm lẫn lớp chung để tránh căn chỉnh rõ ràng giữa miền nguồn và mục tiêu.

2.6 Show and Tell: A Neural Image Caption Generator

Phát triển của các mô hình sinh chú thích hình ảnh đã trải qua nhiều giai đoạn đổi mới, từ các hệ thống dựa trên quy tắc đến việc áp dụng học sâu để tự động hóa quy trình này. Trong số những bước đột phá ban đầu, công trình "Show and Tell: A Neural Image Caption Generator" của Vinyals và cộng sự [18] đã định hình lại cách thức chúng ta tiếp cận nhiệm vụ này. Sử dụng mạng nơ-ron LSTM kết hợp với CNN, mô hình của họ đã thành công trong việc học cách sinh chú thích mô tả một cách mạch lạc và liên quan đến nội dung hình ảnh. Công trình này đã mở ra cánh cửa cho nhiều nghiên cứu sau này, nhấn mạnh sự cần thiết của việc tích hợp chặt chẽ giữa thông tin thị giác và ngôn ngữ.

2.7 Show, Attend and Tell

Nối tiếp xu hướng tích hợp sâu giữa xử lý hình ảnh và ngôn ngữ tự nhiên, mô hình "Show and Tell" của Vinyals và cộng sự [18] đã đặt nền móng vững chắc cho ngành. Mở rộng thêm từ công trình này, Xu và cộng sự đã đề xuất mô hình "Show, Attend and Tell" [21], đánh dấu sự tiến bộ lớn trong lĩnh vực bằng việc áp dụng cơ chế chú ý vào sinh chú thích hình ảnh. Cơ chế này cho phép mô hình tập trung vào các phần cụ thể của hình ảnh khi sinh từng từ của chú thích, cung cấp một cách tiếp cận linh hoạt và mạnh mẽ hơn để tạo ra chú thích có liên quan mật thiết với nội dung thị giác.

2.8 Vision-and-Language Transformer

Một trong những bước tiến gần đây trong lĩnh vực Vision-and-Language Models là sự ra đời của ViLT (Vision-and-Language Transformer) [10], mô hình này làm việc mà không cần đến các phép tích chập hoặc giám sát vùng. ViLT được thiết kế để xử lý cả hình ảnh và ngôn ngữ mà không cần phụ thuộc vào CNN truyền thống hay các phương pháp dựa trên vùng như các mô hình trước đây. Nhờ vào kiến trúc Transformer mạnh mẽ, ViLT học được cách trích xuất và tổng hợp các đặc trưng từ dữ liệu hình ảnh và văn bản một cách hiệu quả, mở ra khả năng áp dụng trong nhiều tình huống đòi hỏi sự hiểu biết liên modal mà không cần đến việc phân vùng hoặc đánh dấu các đối tượng cụ thể trong hình ảnh.

2.9 Tiền Huấn Luyện Cho Hiểu Ngôn Ngữ Sâu

Một trong những đóng góp quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên là mô hình BERT (Bidirectional Encoder Representations from Transformers) [5], được tiền huấn luyện để hiểu ngôn ngữ sâu rộng với cơ chế Transformer hai chiều. Mô hình BERT đã thiết lập một tiêu chuẩn mới cho nhiều tác vụ NLP bằng cách sử dụng tiền huấn luyện trên một lượng lớn văn bản không được gán nhãn, cho phép nó nắm bắt được một biểu diễn ngôn ngữ phong phú và đa dạng. Phương pháp này đã ảnh hưởng lớn đến sự phát triển của các mô hình sau này trong VLMs, nơi sự hiểu biết về ngôn ngữ và hình ảnh đang ngày càng trở nên quan trọng.

3 Hướng Tiếp Cận

Trước khi khám phá cách tiếp cận cụ thể của chúng tôi đối với mô hình Vision-Language Models (VLMs), chúng ta cần xác định và hiểu rõ các khái niệm chủ chốt mà mô hình của chúng tôi sẽ dựa trên. Một trong những khái niệm cốt yếu là việc "Aligning" -

quá trình căn chỉnh không gian đặc trưng chung giữa dữ liệu hình ảnh và văn bản, sao cho mô hình có thể hiểu và xử lý thông tin đồng thời từ cả hai nguồn đầu vào. Quá trình này yêu cầu mô hình phải khám phá và học cách liên kết các đặc trưng thị giác với ngữ cảnh ngôn ngữ, một bước tiến quan trọng hướng tới khả năng tổng quát hóa tốt hơn và ứng dụng thực tế mạnh mẽ hơn.

Khái niệm thứ hai là "Zero-shot learning", một phương pháp cho phép mô hình thực hiện nhận dạng hoặc tác vụ liên quan mà không cần dữ liệu huấn luyện được gán nhãn cụ thể cho nhiệm vụ đó. Trong lĩnh vực VLMs, điều này đặc biệt có giá trị bởi vì nó mở ra khả năng áp dụng mô hình trên một phạm vi rộng lớn các miền và tác vụ mà không cần tinh chỉnh rộng rãi hoặc cung cấp dữ liệu huấn luyện cho từng trường hợp cụ thể.

Với việc thiết lập nền tảng vững chắc về những khái niệm này, phần tiếp theo sẽ mô tả chi tiết hơn về cách tiếp cận mà chúng tôi áp dụng trong việc phát triển mô hình VLMs, hướng tới việc giải quyết các thách thức liên quan đến việc hiểu và sinh nội dung đa phương tiện.

3.1 Kỹ thuật Aligning

Aligning là kỹ thuật giúp mô hình có thể hiểu được nội dung ngữ nghĩa của hình ảnh bằng ngôn ngữ tự nhiên. Nhìn chung có ba phương pháp cổ điển mà các mô hình VLM thực hiện việc học mối liên kết giữa ảnh và text:

3.1.1 Attention Alignment

Thay vì đi theo các phương pháp cũ như liên kết text và ảnh thông qua đặc trưng chủ thể (object-centric features) hay tập trung vào bức tranh toàn cục (coarse-grained features). Phương pháp Attention Alignment[7] thực hiện việc học các mối liên kết giữa ảnh và ngôn ngữ bằng hai tiêu chí lần lượt bao gồm: định vị các thông tin thị giác trong hình ảnh đã cho bằng cách kết hợp hàm mất mát hồi quy khung giới hạn và mất mát IoU, căn chỉnh các văn bản với các thông tin thị giác trước đó thông qua hàm mất mát đối lập (contrastive loss), mất mát phù hợp (matching loss) và mất mát mô hình ngôn ngữ được che khuất (masked language modeling loss) trong đó các sự căn chỉnh là ở nhiều cấp độ khác nhau. Phương pháp này được sử dụng thông qua mô hình X-VLM với vision encoding được xây dựng dựa trên mạng vision transformer, text encoder và cross-modal modeling - mạng dùng để kết hợp giữa text token và embedded vector ảnh lại với nhau thông qua chiến lược contrastive learning. Hình 2

Chú thích thêm về contrastive learning: Cho một cặp (V, T), trong đó T là ví dụ positive cho V, và chúng ta xem xét các $(N-1)$ văn bản khác trong mini-batch là các ví dụ negative. Tính sự tương đồng cosin theo công thức như sau: $s(V,T) = g_v(v_{\text{cls}}) \cdot g_w(w_{\text{cls}})$. Trong đó, w_{cls} là embedding đầu ra của bộ mã hóa văn bản. g_v và g_w là các biến đổi ánh xạ các embedding sang biểu diễn có chiều thấp được chuẩn hóa. Sau đó, ta tính in-batch vision-to-text similarity theo công thức:

$$p^{v2t}(V) = \frac{\exp(s(V,T)/\tau)}{\sum_{i=1}^{N} \exp(s(V,T^i)/\tau)}, \tag{1}$$

Và the text-to-vision similarity:

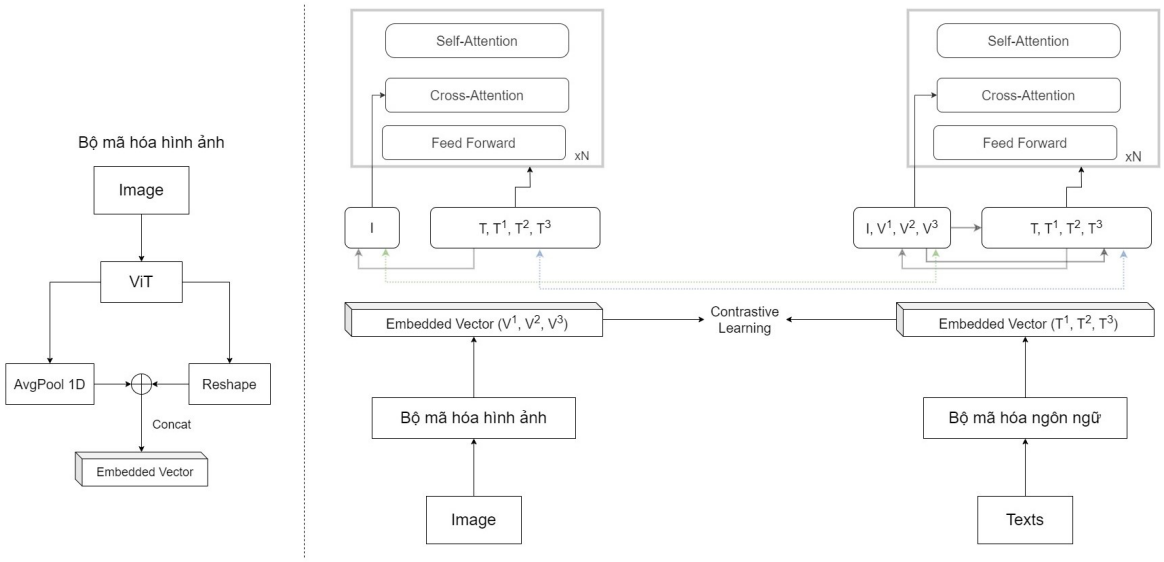
$$p^{t2v}(T) = \frac{\exp(s(V,T)/\tau)}{\sum_{i=1}^{N} \exp(s(V^i,T)/\tau)}, \tag{2}$$

Với τ là một tham số nhiệt độ có thể học được. Đặt $y^{v2t}(V)$ và $y^{t2v}(T)$ là sự tương đồng one-hot thực tế, trong đó chỉ có cặp tích cực có xác suất là một. Contrastive loss được định nghĩa theo hàm mất mát cross-entropy H giữa p và y:

$$\mathcal{L}_{cl} = -\frac{1}{2}\mathbb{E}_{V,T\sim\mathcal{D}}\left[H(y^{v2t}(V), p^{v2t}(V)) + H(y^{t2v}(T), p^{t2v}(T))\right] \tag{3}$$

3.1.2 Cross-Modal Alignment

Cross-Modal alignment[9] được phát triển dựa trên cơ chế học từ vựng từ ngữ cảnh (cross-situational learning - XLS), máy sẽ được học một từ có liên quan đến một đối tượng hình ảnh trong một hoàn cảnh không rõ ràng, và việc ánh xạ từ và các đối tượng hình ảnh trở nên chính xác hơn nếu đủ thông tin được thu thập thông qua sự tiếp xúc liên tục với các



Hình 2: Contrastive Learning

tình huống khác nhau. Cụ thể, với tập dữ liệu gồm ảnh I chứa có M vật thể được gán nhãn thông qua tập B và tập các đoạn text đã được mã hóa S mô tả các nội dung tương ứng với ảnh I. Mô hình sẽ được học liên tục đồng thời toàn bộ các vật thể trong ảnh và từ ngữ tương ứng với nhau nhằm xây dựng đồ thị quan hệ giữa B và S và các trọng số về object và từ ngữ của các cạnh trong đồ thị sẽ được liên tục cập nhật trong quá trình học.

Xuyên suốt quá trình học, sẽ có hai đồ thị được xây dựng lần lượt là đồ thị về ngôn ngữ và vật thể. Đối với đồ thị ngôn ngữ $G_W(V_W, E_W)$ với V_W là các nút và E_W là các cạnh. Các nút của đồ thị sẽ bao gồm các token mã hóa vwi và số lượng từ dwi đã gặp từ dữ liệu đầu vào. Với mỗi local context từ dữ liệu đầu vào $Q = \{q_1, q_2, q_3, \ldots, q_k\}$, nếu mô hình gặp một từ mới chưa có trong đồ thị, một nút mới sẽ được lập ta và d_{word} cho nút đó sẽ được khởi tạo bằng 0. Sau đó, hệ số d_{q_k} được tăng lên một đơn vị cho toàn bộ các nút có liên quan đến q_k trong Q.

Mỗi cạnh (vwi, vwj) có số lần xuất hiện chung $c(wi, wj)$ của hai từ wi và wj. Đối với mỗi input ngữ cảnh cục bộ $Q_t = \{q_1, q_2, \ldots, q_k\}$, khi một nút mới vw_{new} được tạo ra, các cạnh (vw_{new}, vw) kết nối nút đó vw_{new} và tất cả các nút khác vw được tạo ra, và $c(w_{\text{new}}, w)$ được khởi tạo với giá trị 0. Sau đó, đối với mọi cặp từ (q_i, q_j) trong Q_t, số lần xuất hiện chung $c(q_i, q_j)$ được tăng lên 1. Thông qua các quy trình này, trọng số cạnh e_{wiwj} giữa hai nút vwi và vwj được định nghĩa thông qua công thức không rõ ràng. Lúc này, trọng số cạnh e_{wiwj} có giá trị bằng 1 cho hai từ wi và wj luôn xuất hiện cùng nhau, và càng thấp tần suất xuất hiện chung của họ, trọng số càng gần với 0. Điều này giúp học được phân phối và mối quan hệ của các từ, và mạng lưới đồ thị có trọng số không hướng được liên tục hình thành.

$$e_{w_i w_j} = \frac{c^2_{w_i w_j}}{\langle d_{w_i}, d_{w_j} \rangle}, \tag{4}$$

Tương tự với từ, trong thị giác, các đối tượng có liên quan về ý nghĩa thường xuất hiện cùng nhau hơn trong một cảnh. Dựa trên giả định này, chúng tôi cũng xây dựng các mạng lưới đồ thị quan hệ $G_O(V_O, E_O)$ của các đối tượng theo cùng cách như quá trình hình thành mạng lưới đồ thị quan hệ của

các từ. Ở đây, V_O là tập hợp các nút đối tượng và E_O là tập hợp các cạnh giữa các nút. Để hình thành các mạng lưới đồ thị quan hệ cho các đối tượng, ta sử dụng tập hợp các đối tượng B trong luồng đầu vào $B = \{b_1, b_2, \ldots, b_m\}$, nếu một đối tượng mới o_{new} chưa có bao gồm trong tập hợp O được quan sát, một nút mới $v_{O_{\text{new}}}$ được tạo ra, và $d_{O_{\text{new}}}$ và $c(o_{\text{new}}, b)$ được khởi tạo với giá trị 0. Sau đó, d_{b_m} được tăng lên 1 cho mỗi nút v_{b_m} tương ứng với mỗi đối tượng trong B, và số lần xuất hiện chung $c(b_i, b_j)$ cho cạnh (v_{b_i}, v_{b_j}) cho mỗi cặp đối tượng (b_i, b_j) trong B cũng được tăng lên 1. Cuối cùng, bằng cách sử dụng số lượng đầu vào tích lũy d_{O_i} và d_{O_j} của hai nút đối tượng v_{O_i} và v_{O_j} và số lần xuất hiện chung $c(o_i, o_j)$ của chúng, trọng số cạnh $e_{o_i o_j}$ của các nút đối tượng được định nghĩa theo công thức sau:

$$e_{o_i o_j} = \frac{c^2_{o_i o_j}}{\langle d_{o_i}, d_{o_j} \rangle}, \tag{5}$$

Sau khi xây dựng 2 đồ thị về text và object, một phương pháp học cross-modal representation nhằm học các biểu diễn ý nghĩa của các đối tượng và từ dựa trên các mạng lưới đồ thị quan hệ chéo phương thức đã được xây dựng. Phương pháp này cho phép các thực thể của các phương thức khác nhau với cùng một ý nghĩa khái niệm có cùng một vector biểu diễn. Mỗi nút đối tượng và từ đều có vector biểu diễn ý nghĩa riêng của mình roi hoặc rwi, mỗi một vector này được khởi tạo theo thuật toán Neighborhood aggregation như hình 3.

Các vector được huấn luyện thông qua mô hình multi-layer perceptron với chiến thuật self-supervised learning để phân loại semantic của các nút và ghép chúng lại với nhau về mặt ngữ nghĩa giữa 2 đồ thị thông qua hàm softmax.

3.1.3 Contrastive and generative alignment

Việc huấn luyện visual adapter không căn chỉnh hiệu quả giữa các phương thức ngôn ngữ-hình ảnh. Các đặc trưng của các token văn bản thường có độ tương đồng cosin lớn với hầu hết các đặc trưng của các patch hình ảnh, cho thấy sự căn chỉnh yếu giữa các phương thức ngôn ngữ-hình ảnh. Hơn nữa, sự căn chỉnh yếu này làm cho mô hình rất đòi hỏi dữ liệu cho việc điều chỉnh fine-tuning theo hướng dẫn hình ảnh. Do đó, cần có một mô hình ngôn ngữ-hình ảnh hiệu quả về dữ liệu điều chỉnh hướng dẫn hình ảnh, đạt được hầu hết hiệu suất ngay cả khi chỉ được huấn luyện với 10% lượng dữ liệu gốc. Bảo tồn nhiều khả năng văn bản từ mô hình ngôn ngữ lớn ban đầu và đạt được kết quả tốt hơn trên hướng dẫn hoặc câu hỏi chỉ văn bản. Phương pháp **CG-Alignment**[13] giải quyết vấn đề trên.

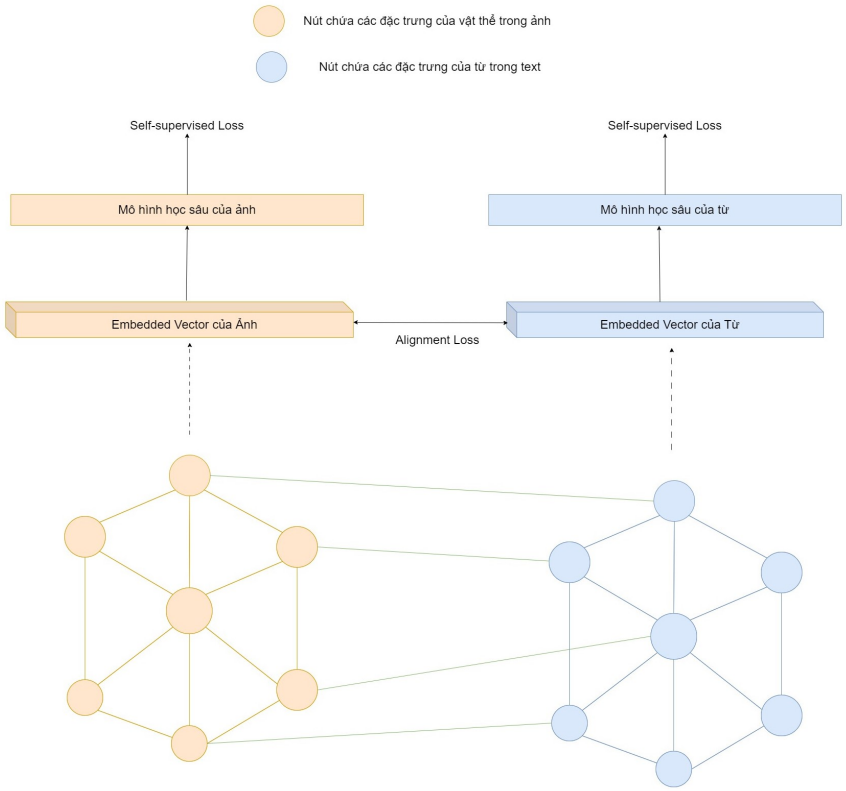
Phương pháp **CG-VLM** bao gồm hai giai đoạn chính để trang bị cho mô hình ngôn ngữ lớn (hình4) đã được huấn luyện trước với khả năng tuân thủ hướng dẫn ngôn ngữ-hình ảnh:

1. **Giai Đoạn Căn Chỉnh Hình Ảnh-Ngôn Ngữ CG**: Ở giai đoạn này, thực hiện kết hợp cả các object trái chiều và tạo sinh. Điều này giúp huấn luyện một bộ chuyển đổi hình ảnh để căn chỉnh biểu diễn của bộ biến đổi hình ảnh đã được huấn luyện trước và mô hình ngôn ngữ lớn trên tập dữ liệu mô tả hình ảnh.

2. **Giai Đoạn Điều Chỉnh Hướng Dẫn Hình Ảnh**: Sau giai đoạn căn chỉnh, chúng tôi tiếp tục điều chỉnh mô hình ngôn ngữ lớn bằng cách sử dụng tập dữ liệu theo hướng dẫn. Mục tiêu ở đây là nâng cao khả năng của mô hình tuân thủ hướng dẫn, đặc biệt là khi được cung cấp nội dung hình ảnh làm tiền tố.

Vision-language generative loss: Vì không gian ngữ nghĩa của các mô hình thị giác và ngôn ngữ không căn chỉnh, chúng ta cần căn chỉnh biểu diễn của chúng trước khi lan truyền thông tin ngôn ngữ-hình ảnh. Để đạt được điều này, các phương pháp hiện có thường căn chỉnh ViT và LLM với dữ liệu mô tả hình ảnh, tức là dữ liệu cặp hình ảnh-văn bản như sau. Cho trước các đặc trưng thị giác và vector nhúng của từ chuỗi prefix $x_{<}j$ làm dữ liệu đầu vào. Mô hình LLM sẽ dự đoán xác suất của từ kế tiếp x_j thông qua công thức:

$$p(x_j | u_{1:N}, x_{<j}) = g\left(\bigoplus_{i=1}^{N}[z_i] \oplus \bigoplus_{k=1}^{j-1}[e_k]\right) \tag{6}$$



Hình 3: Sơ đồ workflow Cross Model.

Với $\oplus$ là phép toán nối các đặc trưng hình ảnh z_i và vector nhúng của từ e_k của chuỗi câu. Sau đó, sự liên kết ngôn ngữ thị giác đạt được bằng cách tối ưu hóa mất dự đoán từ tiếp theo tổng quát như sau

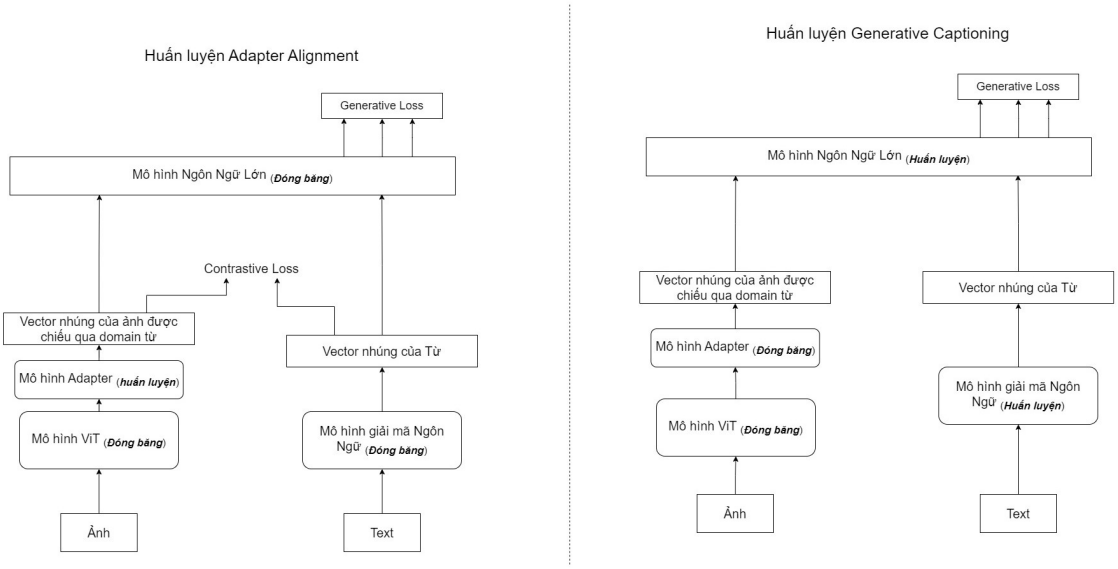
$$\mathcal{L}_{\text{align}}^{\text{gen}} = -\frac{1}{M}\sum_{j=1}^{M}\log p(x_j|u_{1:N}, x_{<j}). \tag{7}$$

Vision-language contrastive loss: Có ba khó khăn khi căn chỉnh giữa ViT và LLM. Thứ nhất, các đặc trưng hình ảnh trong một mô hình ngôn ngữ-hình ảnh thường ở mức độ patch để cung cấp tín hiệu hình ảnh mật độ cao. Thứ hai, thường có nhiều hơn một embedding cho một chú thích văn bản. Thứ ba, không có sự gắn kết rõ ràng giữa các patch hình ảnh và các token văn bản trong tập dữ liệu mô tả hình ảnh tiêu chuẩn. Để giải quyết các vấn đề trên, tác giả đã đề xuất tối đa hóa sự tương đồng trung bình giữa các đặc trưng hình ảnh gom lại và các nhúng của chú thích tương ứng

$$s^{i,j} = \frac{1}{M}\sum_{m=1}^{M}\varepsilon(\hat{Z}^i, E_m^j) \tag{8}$$

Trong đó $\varepsilon(u,v) = \tau \cdot \frac{u^T v}{\|u\|_2\|v\|_2}$ biểu thị hàm tương đồng cosin có tỉ lệ, và τ là một hệ số có thể học. Sau đó, chúng ta rút ra hàm mất mát căn chỉnh đối lập của mình như sau:

$$\mathcal{L}_{align}^{con} = \frac{1}{B}\sum_{b=1}^{B}\frac{\exp(s^{b,b})}{\exp(s^{b,b}) + \sum_{b'\neq b}\exp(s^{b,b'})} \tag{9}$$



Hình 4: Sơ đồ workflow của Contrastive Generative Alignment.

Visual Instruction Tuning: Với các token và embedded căn chỉnh, giai đoạn điều chỉnh hướng dẫn hình ảnh nhằm mục đích điều chỉnh mô hình ngôn ngữ lớn để tuân thủ hướng dẫn ngôn ngữ-hình ảnh tốt hơn. Tương tự như quá trình căn chỉnh tạo ra, các giai đoạn điều chỉnh hướng dẫn hình ảnh cũng sử dụng dữ liệu cặp hình ảnh-văn bản, nhằm hướng đến một cuộc trò chuyện tự nhiên giữa con người và một trợ lý về mặt ngôn ngữ. Cho ảnh đầu vào được phân chia thành các patch $u_{1:N}$, các nhúng từ của câu truy vấn văn bản $x^q_{1:Q}$ và tiền tố được tạo ra của phản hồi văn bản $x^r_{<j}$, mất mát tạo ra liên quan đến phản hồi văn bản $x^r_{1:M}$ được tính như sau:

$$\mathcal{L}_{\text{tune}} = -\frac{1}{M}\sum_{j=1}^{M} \log p(x^r_j | u_{1:N}, x^q_{1:Q}, x^r_{<j}); \qquad (10)$$

Với p là xác suất của từ tiếp theo trong phản hồi.

3.2 Zero-shot learning:

Zero-shot learning [19] là phương pháp xác định lớp của các object chưa được quan sát thấy trong quá trình huấn luyện mô hình. Ý tưởng đằng sau Zero-shot learning là huấn luyện máy khả năng nhận dạng, tương tự như việc con người có thể tìm thấy sự tương đồng giữa các lớp dữ liệu một cách tự nhiên, bằng việc dựa trên việc chuyển vector đặc trưng trong không gian ngữ nghĩa từ các nhãn đã học được sang nhãn mới.

3.2.1 Hướng tiếp cận

Tiếp cận dựa trên model: Mục tiêu chính của phương pháp này là biến đổi các đặc trưng hình ảnh và các thuộc tính ngữ nghĩa vào một không gian vector. Hình ảnh đầu vào ban đầu được chuyển qua mạng trích xuất đặc trưng (Feature Extractor) để có được vectơ đặc trưng N-chiều cho hình ảnh. Vectơ này đóng vai trò là đầu vào của mạng encoder và mạng trả về kết quả của một vectơ đầu ra có kích thước D-chiều.

Mục tiêu cuối cùng là tính toán trọng số của mạng chiếu để ánh xạ đầu vào N chiều thành đầu ra D chiều. Để có được điều này, hàm loss giữa đầu ra D-chiều và thuộc tính ngữ nghĩa sự thật cơ bản để tính toán độ lỗi và cập nhật lại các trọng số của mạng sao cho đầu ra D-chiều càng gần với dữ liệu chân thực cơ bản càng tốt.

Giả sử có các đặc trưng như sau: bốn chân, ăn cá, ăn cỏ, có vằn, có đuôi, giả sử vector thuộc tính của con ngựa sẽ là [1, 0, 1, 0, 1]. Với giả sử mô hình đủ tốt để nhận diện ra các đặc tính cảu bức ảnh, ta đưa đầu vào là bức ảnh con ngựa vằn vào, mô hình trả ra cho ta vector thuộc tính dạng như con ngựa nhưng ở thuộc tính có vằn thì lại không được đề cập đến trong vector nhúng. (2 stages model).

Ngoài ra, có một số cải tiến cho pp này như việc học kết nối trực tiếp giữa miền đặc trưng ảnh và miền ngữ nghĩa một cách tuyến tính, song tuyến tính hoặc phi tuyến tính thông qua các hàm loss.

Hạn chế chính của phương pháp encoder là chúng gặp phải vấn đề sai lệch và dịch chuyển miền. Điều này có nghĩa là vì mô hình mạng học sâu chỉ được học bằng cách sử dụng các lớp được nhìn thấy trong quá trình đào tạo, kết quả của mô hình sẽ thiên về dự đoán các nhãn danh mục đã nhìn thấy. Cũng không có gì chắc chắn rằng chức năng mã hóa được đào tạo sẽ ánh xạ chính xác các đối tượng ảnh danh mục không quan sát được vào không gian ngữ nghĩa tương ứng ở giai đoạn thực nghiệm.

Cách tiếp cận dựa trên mô hình tạo sinh: Mục tiêu của phương pháp tạo sinh là tạo ra các đặc điểm hình ảnh cho các danh mục không được quan sát bằng cách sử dụng các thuộc tính ngữ nghĩa. Nói chung, điều này được thực hiện bằng cách sử dụng mạng cGAN nhằm tạo ra các đặc điểm hình ảnh có điều kiện dựa trên thuộc tính ngữ nghĩa của một danh mục nhất định như Hình 5. Giống với phương pháp trước, ảnh được đưa qua mạng trích xuất đặc trưng để có được vectơ đặc trưng N chiều. Sau đó, thay vì được đối chiếu trực tiếp với đặc trưng của ảnh đầu vào, chúng sẽ được đưa vào một mô hình tạo sinh (Generative Model) như Hình 6. Ở giai đoạn này, vector thuộc tính thay vị được so sánh trực tiếp với vector đặc trưng thì chúng được đưa qua khối Generator để điều chỉnh vector thuộc tính sao cho giống với vector đặc trưng nhất có thể. Như vậy, khác với cách tiếp cận ban đầu, thay vì cố gắng biến đổi vector đặc trưng sao cho sát nhất với vector thuộc tính, ở cách tiếp cận này, cả 2 vector đều phải biến đổi sao cho giống nhau, khó phân biệt.

Sau khi mô hình sinh đã được huấn luyện, lúc này trong quá trình Inference, Ảnh đầu vào sẽ dược đưa vào mô hình, đi kèm với đó là các lớp có trong train dataset và các lớp mới được đưa vào khối Generator, tiếp tục sử dụng khối Discriminative để phân biệt, trong đó việc phân biệt vector đặc trưng với đầu ra của khối Generator của thuộc tính nhìn thấy được phân biệt với trọng số cao hơn so với thuộc tính không nhìn thấy. Đơn giản vì với dữ liệu mới, nguyên lý là chủ yếu dựa trên các thuộc tính nhìn thấy, còn các thuộc tính mới là bổ sung nên trọng số phụ thuộc sẽ thấp hơn. Cuối cùng trả ra Classification score, từ đó có thể trả ra các kết quả mong muốn.

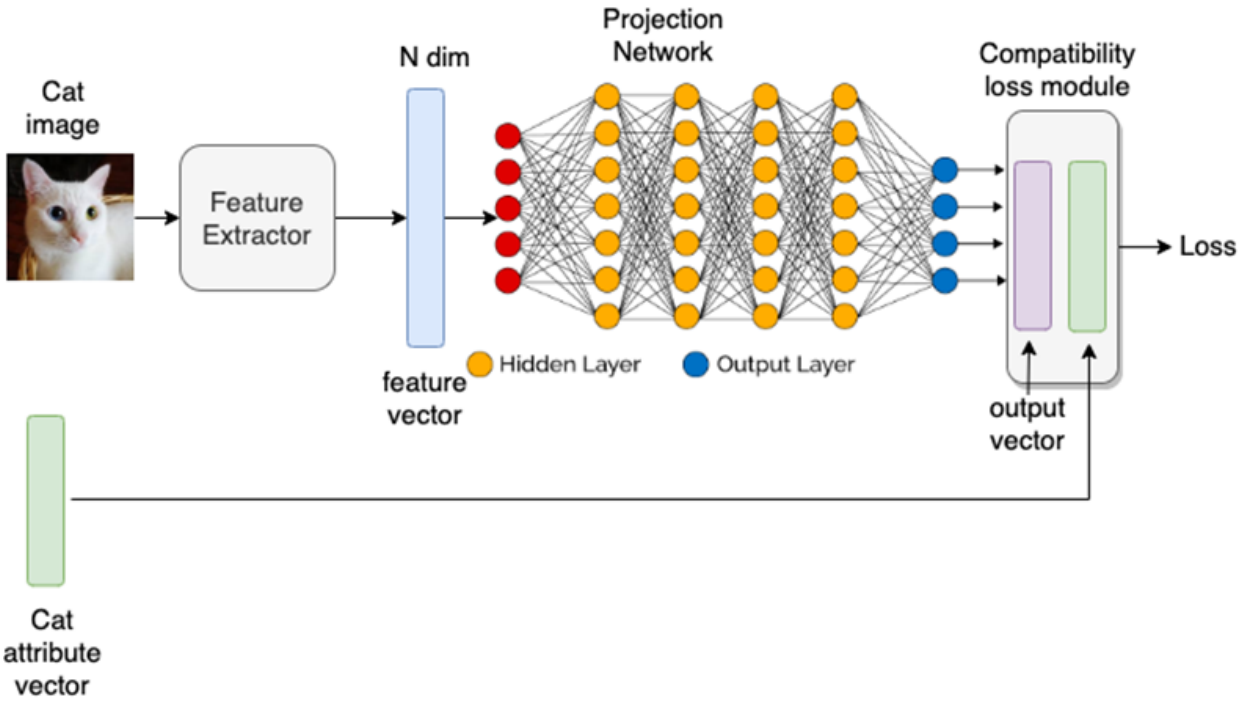
Chỉ số đánh giá cho phương pháp học Zero-shot: Nếu độ chính xác được tính trung bình cho tất cả các hình ảnh, thì việc tính toán sẽ trở nên dễ dàng và hiệu quả cho các lớp có nhiều dữ liệu nhưng lại kém hiệu quả cho các lớp có quá ít dữ liệu. Tuy nhiên, việc quan tâm đến có được hiệu suất cao trên lớp dân cư thưa thớt là điều cần thiết. Vì vậy, ta sẽ thay đổi cách tính trung bình giá trị dự đoán độc lập cho từng lớp theo top 1 accuracy theo cách sau

$$\text{acc}_y = \frac{1}{\|Y\|} \sum_{c=1}^{\|Y\|} \frac{\#\text{ correct predictions in } c}{\#\text{ samples in } c} \tag{11}$$

Sau đó, tính toán độ chính xác cho từng lớp riêng lẻ và tính trung bình trên tất cả các danh mục khác nhau. Điều này giúp tăng hiệu suất ở cả các lớp có ít dữ liệu. Mục tiêu chính của phương pháp này là đạt được độ chính xác cao trên cả các lớp đã quan sát (seen) và các lớp chưa quan sát (unseen). Do đó, chỉ số hiệu suất được định nghĩa là giá trị trung bình hài hòa (harmonic mean) của hiệu suất trên các lớp seen và unseen.

$$H = \frac{2 \times \text{acc}_{y_{\text{tr}}} \times \text{acc}_{y_{\text{ts}}}}{\text{acc}_{y_{\text{tr}}} + \text{acc}_{y_{\text{ts}}}} \tag{12}$$

Zero-Shot learning trong VLM: Phần lớn các mô hình VLM đều sử dụng chiến lược Zero-shot learning



Hình 5: Cách tiếp cận mã hóa các đặc trưng

trong quá trình đào tạo của mình như Hình 7. Cụ thể, ở mô hình lớn như Sim VLM, thực hiện chiến lược zero-shot learning trong khâu huấn luyện trực tiếp cho các encoder và decoder của ảnh và ngôn ngữ đối với các dữ liệu nhiễu và zero-shot cross modal transfer khi huấn luyện trên nhiều tập dữ liệu ảnh và ngôn ngữ khác nhau. Ngoài ra, zero-shot learning còn tạo ra tiền đề cho mô hình VLM mới cho phép sử dụng hiệu quả dữ liệu web và dự đoán zero-shot mà không cần tinh chỉnh theo nhiệm vụ cụ thể thông qua cấu trúc contrastive learning, ví dụ như CLIP và COCA.

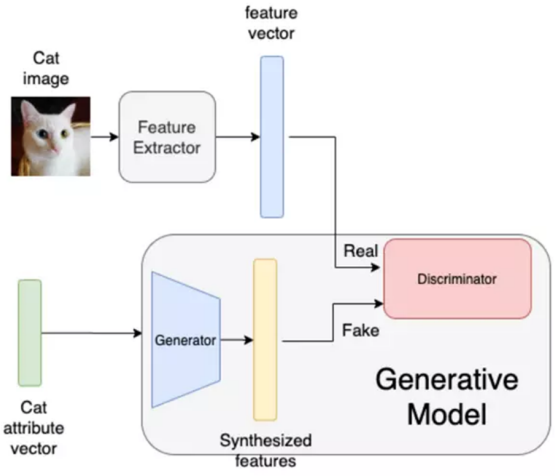
Những phân khúc cần lưu ý khi dùng zero-shot: *Phụ thuộc vào thông tin phụ trợ chính xá*c: phụ thuộc nặng nề vào chất lượng và độ chính xác của thông tin phụ trợ được cung cấp cho các lớp chưa được học và huấn luyện.

Giới hạn trong việc kết hợp các thuộc tính: Nếu chúng ta lấy các thuộc tính từ các nguồn khác nhau hoặc nếu số lượng thuộc tính rất lớn, thì chúng ta cần đào tạo một số lượng rất lớn các bộ phân loại cho các thuộc tính.

Khả năng tổng quát hóa của mô hình: Các mô hình ZSL thường giới hạn trong việc nhận biết các lớp mà chúng không được đào tạo trên. Điều này có thể gây ra vấn đề trong các tình huống thực tế, nơi có thể không có đủ hình ảnh được gán nhãn cho tất cả các lớp trong quá trình đào tạo.

Yêu cầu dữ liệu đào tạo lớn: Mặc dù ZSL không yêu cầu dữ liệu đào tạo cho các lớp chưa thấy, nhưng nó vẫn cần một lượng lớn dữ liệu đào tạo cho các lớp đã thấy để học được các biểu diễn đa phương tiện.

Khả năng suy luận không gian hạn chế: Các mô hình ZSL có thể gặp khó khăn trong việc suy luận về không gian hoặc các môi trường không thực tế về mặt hình ảnh lẫn ngữ nghĩa do có domain shift trong các lớp encoder và decoder.



Hình 6: Cách tiếp cận theo GAN.

4 Chi tiết các mô hình:

Mô hình hợp nhất ngôn ngữ và hình ảnh (VLMs) như ClipCap, BLIP, CoCa và SimVLM đang mở ra kỷ nguyên mới trong việc xử lý và sinh mô tả cho hình ảnh. Với khả năng hiểu và tạo ra văn bản mô tả phong phú, chúng không chỉ giải quyết thách thức trong việc nhận dạng đối tượng mà còn trong việc hiểu cảnh và mối liên hệ với ngữ cảnh rộng lớn hơn. Bài báo này sẽ tập trung vào việc khám phá tiềm năng của chúng trong lĩnh vực Image Captioning, nơi chúng không chỉ cung cấp chú thích mà còn kể chuyện và giải thích hình ảnh một cách tự nhiên và trực quan.

4.1 SIMVLM: SIMPLE VISUAL LANGUAGE MODEL PRE-TRAINING WITH WEAK SUPERVISION

Các nghiên cứu trước đây thường phụ thuộc vào việc sử dụng hai loại tập dữ liệu được gán nhãn bởi con người từ đa dạng nguồn, qua đó thực hiện tiền huấn luyện thông qua hai bước chính. Đầu tiên, sử dụng tập dữ liệu phát hiện đối tượng để huấn luyện bộ phát hiện đối tượng giám sát, như R-CNN và Faster R-CNN, cho phép trích xuất các đặc trưng vùng quan tâm (ROI) từ ảnh. Tiếp theo, các tập dữ liệu chứa cặp hình ảnh-văn bản được căn chỉnh để tiền huấn luyện mô hình bằng cách sử dụng Masked Language Modeling, kết hợp đặc trưng ROI và văn bản.

Tuy nhiên, giới hạn về quy mô dữ liệu được gán nhãn bởi con người và việc sử dụng các mất mát phụ trợ cho mỗi nhiệm vụ cụ thể đã khiến quy trình tiền huấn luyện trở nên phức tạp, cản trở quá trình cải thiện chất lượng mô hình. Các phương pháp này thường không hỗ trợ học zero-shot một cách hiệu quả do thiếu khả năng tổng quát hóa giữa các modal khác nhau.

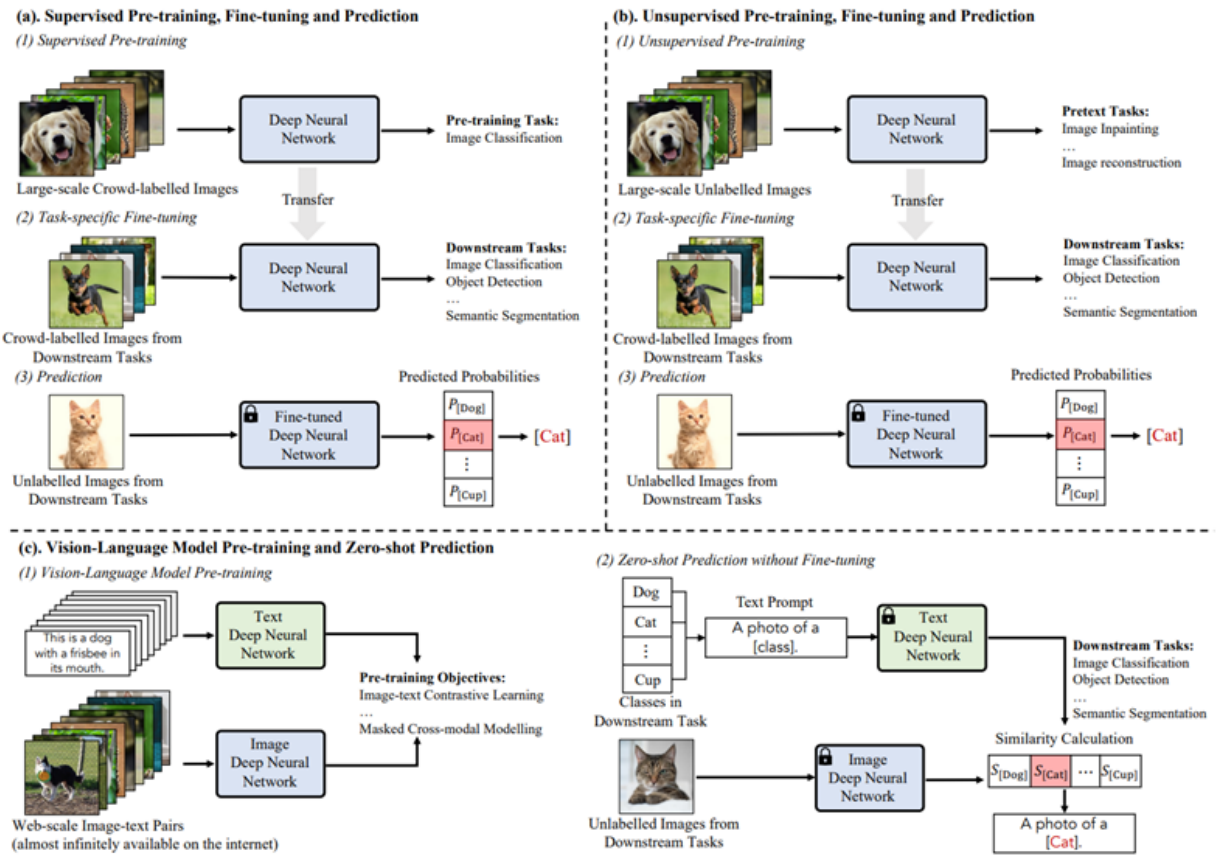
Do đó, sự cần thiết của một mô hình VLM mới là rõ ràng, với các yêu cầu cụ thể như sau:

- Tích hợp mượt mà vào quy trình tiền huấn luyện và điều chỉnh, đạt hiệu suất cao trên các bộ dữ liệu Vision-Language.
- Giảm thiểu yêu cầu về tiền huấn luyện phức tạp so với các phương pháp trước đó.
- Nâng cao khả năng tổng quát hóa trong học zero-shot thông qua việc tận dụng dữ liệu cross-modal.

SimVLM [20] xuất hiện như một giải pháp đáp ứng các yêu cầu trên bằng cách tiếp cận đơn giản nhất: sử dụng trực tiếp hình ảnh thô và tập trung vào mất mát mô hình ngôn ngữ, loại bỏ nhu cầu về tiền huấn luyện phát hiện đối tượng và các mất mát phụ trợ khác. Không chỉ giảm thiểu độ phức tạp, SimVLM còn cho thấy hiệu suất vượt trội so với các mô hình VLP hiện tại, đạt được kết quả hàng đầu trên 6 bộ kiểm tra VL mà không yêu cầu dữ liệu bổ sung hoặc tinh chỉnh cụ thể cho từng nhiệm vụ.

Cách Tiếp Cận của SimVLM

Image Encoder SimVLM sử dụng một hệ thống mã hóa hình ảnh lấy cảm hứng từ mạng ViT và CoAtNet. Mô hình này nhận đầu vào là hình ảnh thô $x \in \mathbb{R}^{H \times W \times C}$ và chuyển đổi nó thành một chuỗi một chiều của các patch hình ảnh được làm phẳng. Kích thước của mỗi patch là $T_i \times D$, trong đó D là kích thước ẩn cố định của các lớp biến đổi và $T_i = \frac{H \times W}{P^2}$ là tổng số token hình ảnh cho một kích thước patch



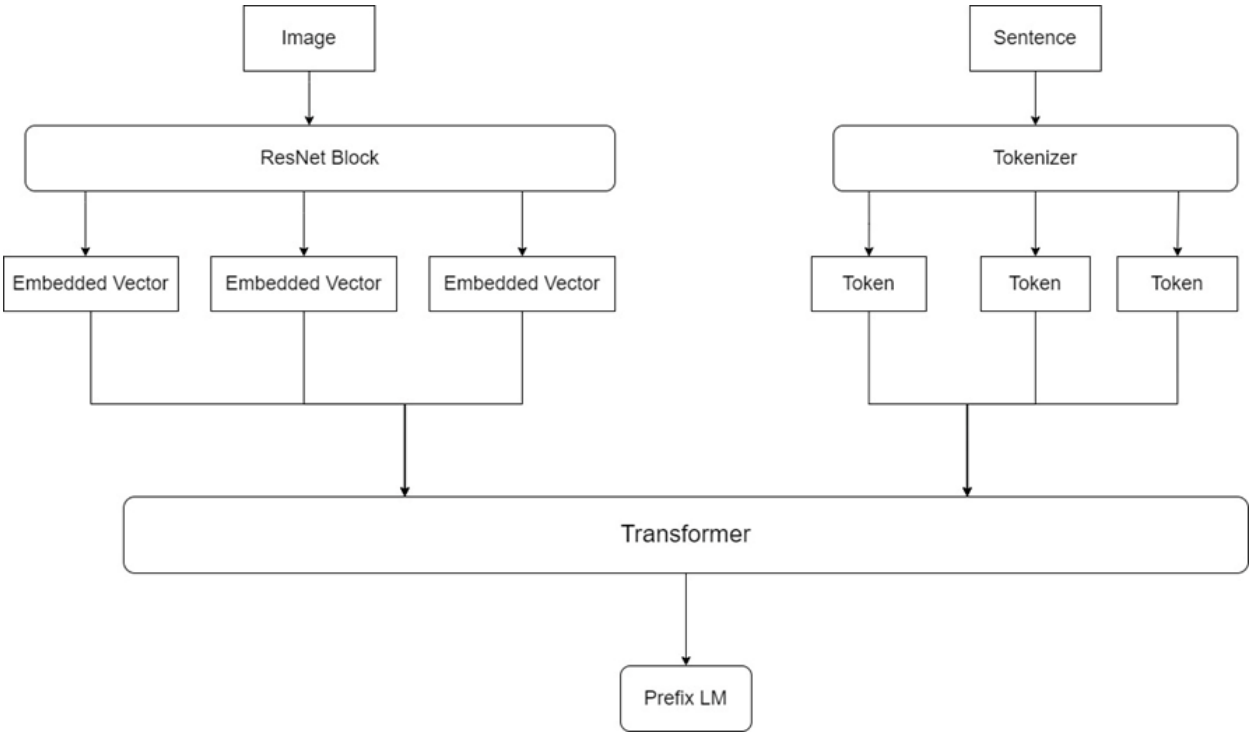
Hình 7: Các hướng tiếp cận của mô hình mạng học sâu

P đã chọn. Các patch này sau đó được xử lý bởi ba khối đầu tiên của một mạng ResNet để trích xuất các vector nhúng có bối cảnh.

Language Encoder Đối với việc mã hóa ngôn ngữ, mô hình tuân theo phương pháp tiêu chuẩn trong việc tách câu thành các token con và sử dụng các token được học cho từ vựng cố định. Để duy trì thông tin về vị trí, mô hình thêm vào các embedding vị trí 1D có thể huấn luyện được cho cả đầu vào hình ảnh và văn bản. Ngoài ra, 2D relative attention được áp dụng cho các patch hình ảnh trong lớp Transformer để nâng cao khả năng chú ý tới mối quan hệ không gian giữa các patch.

Transformer Trái tim của SimVLM là một Transformer dựa trên mô hình PrefixLM, được chọn vì khả năng của nó trong việc xử lý hiệu quả cả nhiệm vụ ngôn ngữ và thị giác. Khác biệt so với một Language Model (LM) tiêu chuẩn, PrefixLM hỗ trợ chú ý hai chiều trong chuỗi tiền tố, cho phép mô hình được áp dụng cho các tác vụ mã hóa-decode mà chỉ cần bộ giải mã. Thử nghiệm sơ bộ của chúng tôi cho thấy rằng cấu trúc mã hóa-giải mã, với việc tách biệt quá trình mã hóa khỏi quá trình sinh sản, mang lại lợi ích đáng kể cho việc cải thiện các nhiệm vụ xuôi.

Prefix Language Modeling trong SimVLM Mô hình SimVLM áp dụng một hàm mất mát Prefix Language Model (PrefixLM), như được biểu diễn



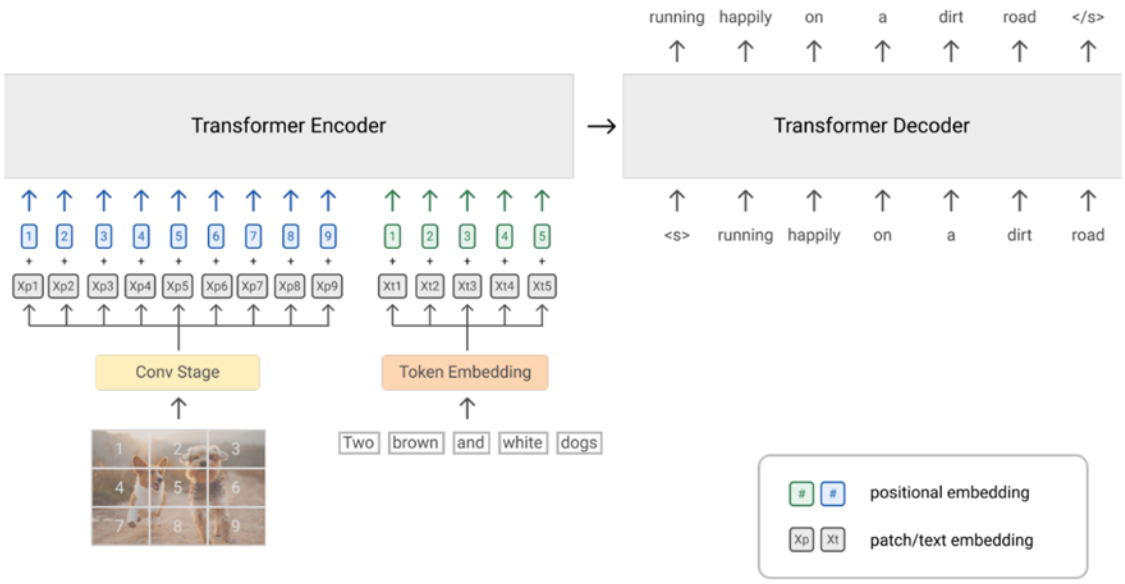
Hình 8: Sơ đồ cấu trúc của SimVLM.

$$\begin{aligned}\mathcal{L}_{\text{PrefixLM}}(\theta) &= -\mathbb{E}_{X\sim\mathcal{D}}\left[\log P_\theta(X_{>T_p} \mid X_{\leq T_p})\right] \\ &= -\mathbb{E}_{X\sim\mathcal{D}}\left[\sum_{t=T_p+1}^{T} \log P_\theta(X_t \mid X_{[T_p,t)}, X_{\leq T_p})\right].\end{aligned} \tag{13}$$

trong Phương trình 13 :

PrefixLM đặc biệt ở chỗ nó cho phép sự chú ý hai chiều trên chuỗi tiền tố ($X_{\leq T_p}$), trong khi vẫn tiếp tục phân tích tương tự như một chuỗi sau ($X_{>T_p}$). Trong quá trình tiền huấn luyện, một chuỗi tiền tố của các token với độ dài ngẫu nhiên T_p được cắt ra từ chuỗi đầu vào, và mục tiêu của quá trình huấn luyện là tối ưu hóa hàm mất mát để dự đoán chuỗi còn lại của token. Điều này giúp mô hình hiểu được cấu trúc và ngữ cảnh của cả dữ liệu hình ảnh và ngôn ngữ một cách đồng bộ.

Một cách tự nhiên, hình ảnh có thể được xem xét như là tiền tố cho các mô tả văn bản của chúng vì chúng thường xuất hiện trước văn bản trong một tài liệu web. Do đó, đối với một cặp hình ảnh-văn bản đã cho, chúng tôi thêm vào đầu chuỗi đặc trưng hình ảnh có độ dài T_i vào trước chuỗi văn bản, và bắt mô hình lấy mẫu một tiền tố có độ dài $T_p \geq T_i$ để tính toán mất mát LM chỉ trên dữ liệu văn bản, như minh họa trong Hình 9. So với các phương pháp VLP theo kiểu MLM trước đó, mô hình PrefixLM thực hiện theo seq2seq framework không chỉ biểu diễn bối cảnh hai chiều như trong MLM, mà còn có thể thực hiện việc tạo ra văn bản tương tự như LM.



Hình 9: Quá trình training PrefixLM.

4.2 Tổng kết và ứng dụng của SimVLM:

Mô hình SimVLM có thể đáp ứng được nhu cầu giải quyết được các bài toán về chú thích nội dung ảnh và hỏi đáp thông tin thị giác. Cụ thể, quá trình training vẫn được tiếp cận theo hướng zero-shot learning đối với 99% các dữ liệu huấn luyện, đồng thời thực hiện điều chỉnh lại chiến lược học của mô hình trên 1% dữ liệu còn lại của tập huấn luyện và sử dụng chúng để huấn luyện mô hình trong vòng 5 epoch. Ngoài ra, kết quả của mô hình trên tác vụ này luôn có câu tiền tố là "A picture of " nhằm giúp cho mô hình đạt được kết quả tốt hơn trong suốt quá trình thực nghiệm. Đối với tác vụ VQA, Sim VLM có thể trả lời tốt các câu hỏi nằm ngoài nội dung của toàn bộ tập dữ liệu huấn luyện với độ chính xác cao so với các hiểu biết và đánh giá của người dùng. Qua đó, mô hình có thể cho thấy được quá trình tự học và tự tích lũy các kiến thức mới dựa trên những gì đã học trong tập dữ liệu. Từ đó, SimVLM đã hoàn toàn tiếp cận theo hướng zero-shot learning mà không hề có bất kỳ chỉnh sửa hay thay đổi nào về mô hình hay dữ liệu cho tác vụ hỏi đáp thông tin thị giác.

4.3 ClipCap: CLIP Prefix for Image Captioning

ClipCap[14] là một phương pháp mới trong lĩnh vực Image Captioning, nơi nó cố gắng giải quyết hai thách thức chính: hiểu ngữ nghĩa và đa dạng trong cách miêu tả một hình ảnh. Mô hình này sử dụng kỹ thuật mã hóa hình ảnh và giải mã văn bản để tạo ra các chú thích mô tả hợp lý, nhằm cầu nối giữa ngôn ngữ tự nhiên và hình ảnh. ClipCap tập trung vào việc hiểu các đối tượng và mối quan hệ giữa chúng trong hình ảnh, cũng như xử lý đa dạng cách diễn đạt mà tập dữ liệu huấn luyện đề xuất. Được đào tạo trên tập dữ liệu Conceptual Captions, nó cho thấy khả năng sinh chú thích chính xác và chi tiết mà không yêu cầu thời gian huấn luyện quá lâu, số lượng tham số lớn hay dữ liệu phụ trợ như kết quả phát hiện, từ đó nâng cao tính ứng dụng thực tế.

4.3.1 Tổng quan về ClipCap:

ClipCap tiếp cận nhiệm vụ tạo chú thích hình ảnh bằng cách kết hợp sức mạnh của bộ mã hoá CLIP với khả năng sinh văn bản của mô hình ngôn ngữ GPT-2, tạo ra một phương thức đơn giản nhưng hiệu quả. Nó giải quyết hai thách thức chính là hiểu biết ngữ nghĩa sâu sắc của hình ảnh và đa dạng cách mô tả. CLIP đóng vai trò tiền tố, chuyển hình ảnh thành các vector nhúng qua mạng ánh xạ tinh tế, được huấn luyện để tối ưu hóa việc ánh xạ này, giúp giảm yêu cầu dữ liệu và thời gian huấn luyện. Prefix được ghép nối với mô hình GPT-2 để sinh ra chú thích, đưa ra cái nhìn sâu sắc và mô tả phong phú, mang lại hiệu quả trong việc sinh ra các chú thích có ý nghĩa cho hình ảnh. (Hình 10)

4.3.2 Chi tiết phương pháp:

Chúng ta bắt đầu với bài toán đặt ra là tạo ra một chú thích ý nghĩa từ một hình ảnh đầu vào chưa từng được nhìn thấy. Chúng ta có thể xem các chú thích như là một chuỗi các tokens c_i, và chúng ta thêm các tokens để đạt đến độ dài lớn nhất l. Mục tiêu huấn luyện được biểu diễn như sau:

$$\max_{\theta} \sum_{i=1}^{N} \log p_\theta(c_1^i, \ldots, c_l^i | x^i), \tag{14}$$

Để đi sâu vào cách mô hình học cách tạo ra chú thích, thay vì dự đoán toàn bộ chú thích cùng một lúc, mô hình được huấn luyện để dự đoán từng token một cách tuần tự, từ đầu đến cuối chú thích, mỗi lần một token. Mục tiêu là để tối đa hóa xác suất dự đoán đúng mỗi token tiếp theo dựa vào các token đã biết và hình ảnh đang được xem xét:

$$\max_{\theta} \sum_{i=1}^{N} \sum_{j=1}^{l} \log p_\theta(c_j^i | x^i, c_1^i, \ldots, c_{j-1}^i). \tag{15}$$

Sử dụng GPT-2 (cỡ lớn) làm mô hình ngôn ngữ, và sử dụng bộ mã hoá từ ngữ (tokenizer) để chiếu chú thích thành chuỗi vector nhúng. Để trích xuất thông tin hình ảnh từ một hình ảnh x_i, chúng ta sử dụng bộ mã hoá hình ảnh tiền huấn luyện từ mô hình CLIP. Tiếp theo, ta sử dụng một mạng ánh xạ nhẹ ký hiệu là F, để ánh xạ nhúng CLIP thành k vector nhúng:

$$p_1^i, \ldots, p_k^i = F(\text{CLIP}(x^i)). \tag{16}$$

Nơi mà vector $p_(ij)$ có cùng kích thước với một vector nhúng từ ngữ c_i. Sau đó chúng ta nối các vector nhúng hình ảnh đã thu được với vector nhúng của chú thích c_i:

$$Z'^i = p_1^i, \ldots, p_k^i, c_1^i, \ldots, c_l^i. \tag{17}$$

Trong quá trình huấn luyện, chúng ta cung cấp cho mô hình ngôn ngữ bằng cách nối tiền tố chú thích Z_i. Mục tiêu huấn luyện của chúng ta là dự đoán các token chú thích dựa tên tiền tố theo các tự hoàn thiện. Để mục đạt mục đích, chúng ta huấn luyện thành phần ánh xạ F sử dụng hàm mất mát cross-entrophy.

$$L_x = -\sum_{i=1}^{N} \sum_{j=1}^{\ell} \log p_\theta(c_j^i | p_1^i, \ldots, p_k^i, c_1^i, \ldots, c_{j-1}^i). \tag{18}$$

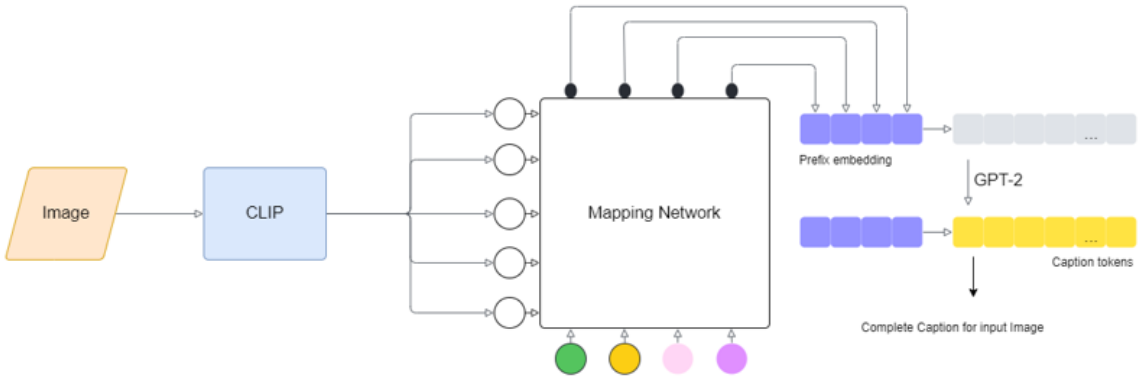
4.3.3 Fine-tuning:

Một thách thức đặc biệt trong việc chuyển đổi giữa các biểu diễn của CLIP và mô hình ngôn ngữ là việc chúng hoạt động trong không gian tiềm ẩn độc lập và không được huấn luyện cùng nhau. Để giải quyết vấn đề này, việc tinh chỉnh mô hình ngôn ngữ trở nên cần thiết, nhất là khi các bộ dữ liệu chú thích yêu cầu các phong cách khác nhau mà mô hình ngôn ngữ tiền huấn luyện có thể chưa được "hiểu" một cách đầy đủ.

Tinh chỉnh mô hình ngôn ngữ mang lại lợi ích qua việc cung cấp sự linh hoạt hơn và tạo ra kết quả biểu đạt mạnh mẽ hơn. Tuy nhiên, nó cũng đồng nghĩa với việc tăng đáng kể số lượng tham số cần được huấn luyện. Một phương án thay thế được đề xuất là giữ nguyên mô hình ngôn ngữ trong suốt quá trình huấn luyện và chỉ tập trung vào việc huấn luyện mạng ánh xạ.

Kiến trúc Mạng Ánh xạ

- *Vai trò chính:* Mạng ánh xạ có trách nhiệm chuyển đổi vector nhúng của CLIP sang không gian mà GPT-2 có thể "hiểu" được.



Hình 10: Mô hình chung của ClipCap

- *Khi có Tinh chỉnh:* Nếu mô hình ngôn ngữ được tinh chỉnh cùng lúc, quá trình ánh xạ trở nên đơn giản và có thể sử dụng MLP đơn giản.
- *Khi không Tinh chỉnh:* Khi mô hình ngôn ngữ đóng băng, cần sử dụng kiến trúc Transformer phức tạp hơn để giảm số lượng tham số và cho phép sự chú ý toàn cầu giữa các token đầu vào.
- *Kết quả:* Kiến trúc Transformer có khả năng cải thiện kết quả bằng cách tăng kích thước tiền tố và cho phép điều chỉnh mô hình ngôn ngữ đóng băng với dữ liệu mới mà không cần tinh chỉnh.

4.4 Tổng kết và ứng dụng của CLIPCAP:

Mô hình CLIPCAP tập trung giải quyết bài toán captioning ảnh, sử dụng sự kết hợp giữa CLIP và một mạng lưới ánh xạ đơn giản để sinh ra các caption có ý nghĩa từ hình ảnh đầu vào. Điểm mạnh chính của phương pháp này là khả năng tận dụng các mô hình tiền huấn luyện như CLIP, giảm thiểu nhu cầu cho dữ liệu gán nhãn bổ sung và thời gian huấn luyện, trong khi vẫn duy trì hiệu suất cao trên các tập dữ liệu lớn và đa dạng. Điểm yếu có thể là việc phụ thuộc vào chất lượng và đa dạng của dữ liệu mà mô hình CLIP đã được huấn luyện trước đó, có thể hạn chế khả năng của mô hình trong việc hiểu và tạo ra các caption cho những tình huống cụ thể chưa từng thấy.

Về Vision-Language Models (VLMs), mô hình CLIPCAP là một ví dụ cho thấy cách mà các mô hình tiền huấn luyện có thể được áp dụng linh hoạt để giải quyết các nhiệm vụ khác nhau liên quan đến hiểu và sinh sản nội dung giữa hình ảnh và văn bản. Điểm mạnh của VLMs trong trường hợp này bao gồm khả năng hiểu sâu sắc các mối liên hệ giữa hình ảnh và văn bản, cũng như khả năng áp dụng tri thức đã học vào các nhiệm vụ mới mà không cần huấn luyện từ đầu. Điểm yếu có thể bao gồm sự phức tạp của mô hình và nhu cầu về tài nguyên tính toán cao cho việc tiền huấn luyện và tinh chỉnh.

4.5 BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

4.5.1 Tổng quan về BLIP

BLIP (Bootstrapping Language-Image Pre-training) [12] đưa ra một khung công tác mới cho Vision-Language Pre-training (VLP) qua hai góc nhìn đột phá:

Góc nhìn về mô hình Các mô hình hiện tại thường dựa trên encoder hoặc encoder-decoder. Các mô hình dựa trên encoder có thể đơn giản hơn nhưng kém hiệu quả trong tác vụ tạo văn bản, trong khi các mô hình encoder-decoder lại gặp khó khăn trong bài toán truy xuất văn bản-hình ảnh.

Góc nhìn dữ liệu BLIP đề xuất một cách thức để hiểu và sản sinh ngôn ngữ hình ảnh một cách thống nhất, mở rộng khả năng thực hiện nhiều nhiệm vụ từ góc độ mô hình và dữ liệu.

- *Multimodal Mixture Encoder-Decoder (MED):* MED có thể hoạt động như một bộ mã hóa đơn mô hình, mã hóa văn bản dựa trên hình ảnh, hoặc giải mã văn bản dựa trên hình ảnh, huấn luyện chung với ba mục tiêu thị giác-ngôn ngữ.
- *Captioning and Filtering (CapFilt):* Quy trình này tinh chỉnh MED để tạo ra chú thích tổng hợp và lọc bỏ chú thích ồn ào từ dữ liệu web, nâng cao chất lượng dữ liệu huấn luyện.

4.5.2 Chi tiết phương pháp

Chúng tôi giới thiệu kiến trúc Multimodal Encoder-Decoder (MED) mới và mục tiêu huấn luyện của nó, đồng thời mô tả chi tiết về Captioning and Filtering (CapFilt) như một phương pháp cho việc khởi động tập dữ liệu.

Kiến trúc mô hình Mô hình sử dụng bộ mã hoá hình ảnh transformer trực quan, phân chia hình ảnh đầu vào thành các patches và mã hoá chúng thành chuỗi vector nhúng. Đặc trưng toàn cục của hình ảnh được biểu diễn bởi token [CLS] bổ sung, tạo ra sự khác biệt so với việc sử dụng bộ phát hiện đối tượng tiền huấn luyện. Việc sử dụng Vision Transformer (ViT) đem lại lợi ích từ góc độ hiệu quả tính toán và đã trở thành xu hướng tiêu chuẩn trong các nghiên cứu gần đây. (Hình 11)

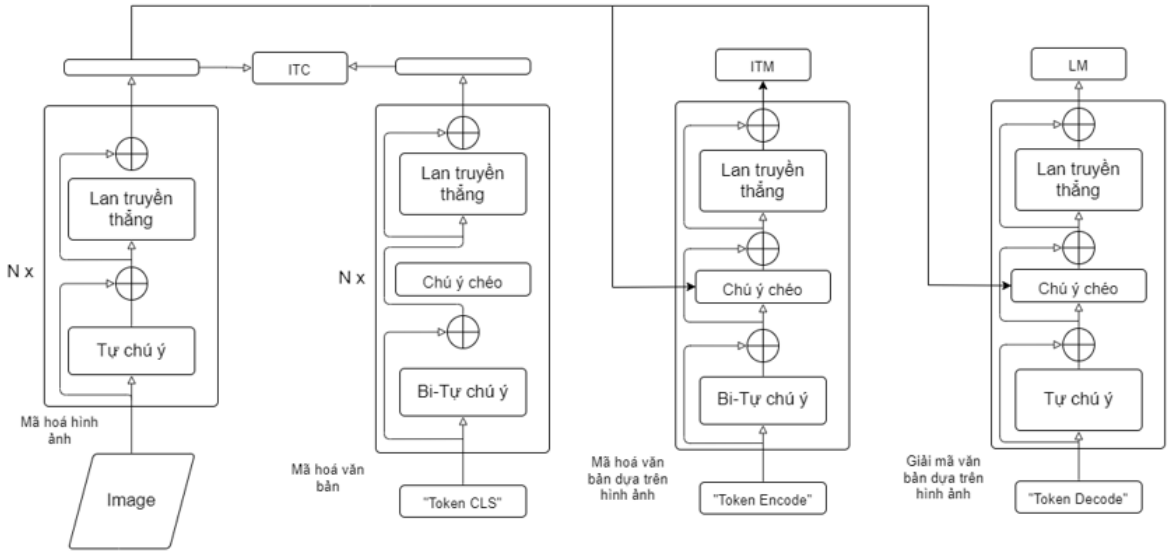
Để tiền huấn luyện một mô hình thống nhất với cả khả năng hiểu và tạo ram chúng tôi đề xuất multimodal mixture of encoder-decoder (MED), một mô hình đa nhiệm vụ có thể hoạt động một trong 4 chức năng:

1. **Bộ Mã Hóa Hình Ảnh:** Hình ảnh được phân chia thành các patch và sau đó mã hóa thành chuỗi các vector nhúng bằng cách sử dụng kiến trúc transformer với quá trình tự chú ý (Self-Attention).
2. **Bộ Mã Hóa Văn Bản:** Sử dụng cơ chế tự chú ý hai chiều (Bi-directional Self-Attention) để tối ưu hóa việc hiểu ngữ cảnh toàn cầu của văn bản đầu vào, với token [CLS] đặt ở đầu để tóm tắt thông tin.
3. **Bộ Mã Hóa Văn Bản Dựa Trên Hình Ảnh:** Cung cấp thông tin hình ảnh bằng cách thêm một lớp chú ý chéo (Cross-Attention) vào mỗi khối transformer của bộ mã hóa văn bản để tạo ra một biểu diễn đa phương tiện.
4. **Bộ Giải Mã Văn Bản Dựa Trên Hình Ảnh:** Thay thế tự chú ý hai chiều bằng tự chú ý nguyên nhân (Causal Self-Attention) và sử dụng token [Decode] để kết thúc chuỗi.

4.5.3 Mục Tiêu Huấn Luyện

Chúng tôi xác định các mục tiêu huấn luyện chính thông qua mất mát tương phản hình ảnh-văn bản (ITC), mất mát khớp hình ảnh-văn bản (ITM), và mất mát mô hình ngôn ngữ (LM) như sau:

- **ITC (Image-Text Contrastive Loss)**: Mục tiêu là điều chỉnh không gian đặc trưng để các cặp hình ảnh-văn bản tích cực có biểu diễn gần nhau hơn so với các cặp tiêu cực.



Hình 11: Framework của BLIP

- **ITM (Image-Text Matching Loss)**: Học biểu diễn đa phương tiện nhằm nắm bắt mối quan hệ chặt chẽ giữa hình ảnh và ngôn ngữ.
- **LM (Language Modeling Loss)**: Tạo ra mô tả văn bản chính xác dựa trên hình ảnh và tối ưu hóa mô hình dựa trên tổn thất cross entropy.

Bộ mã hóa và bộ giải mã văn bản trong mô hình MED chia sẻ tham số, trừ lớp tự chú ý (SA), để phản ánh sự khác biệt giữa mã hóa và giải mã, đồng thời cải thiện hiệu quả huấn luyện.

4.5.4 Kết luận và ứng dụng của BLIP:

BLIP (Bootstrap your own Latent for Image-Text Pre-training) được thiết kế để khắc phục các hạn chế của các mô hình Vision-Language Pre-training (VLP) truyền thống, chủ yếu là khả năng thích ứng kém với các nhiệm vụ dựa trên hiểu biết và tạo sinh nội dung. BLIP đặc biệt mạnh mẽ trong việc xử lý dữ liệu ồn ào từ web, nhờ vào quy trình bootstrapping mà nó sử dụng để lọc ra các caption không chính xác từ dữ liệu thu thập được. Mô hình này đã thiết lập kỷ lục mới trên một loạt các nhiệm vụ như image-text retrieval, image captioning, và visual question answering (VQA), chứng minh sự linh hoạt và hiệu quả của nó trong cả việc hiểu và sinh ra văn bản dựa trên hình ảnh.

Điểm mạnh của BLIP không chỉ nằm ở khả năng xử lý dữ liệu ồn ào mà còn ở khả năng tổng quát hóa mạnh mẽ của nó, có thể chuyển giao trực tiếp sang các nhiệm vụ liên quan đến video mà không cần huấn luyện lại từ đầu. Điều này làm nổi bật khả năng của BLIP trong việc đối mặt với thách thức về đa dạng dữ liệu và tình huống, một điểm hạn chế lớn của nhiều mô hình VLP khác. Mặc dù BLIP đem lại những cải thiện đáng kể, nhưng cũng cần lưu ý rằng hiệu suất của nó phụ thuộc vào lượng dữ liệu pre-training và cách thức xử lý dữ liệu ồn ào, đặt ra những thách thức về việc thu thập và xử lý dữ liệu hiệu quả.

BLIP trong Image Captioning: Trong lĩnh vực tạo caption cho ảnh, BLIP đạt được tiến bộ đáng kể thông qua việc tinh chỉnh mô hình trên các tập dữ liệu như NoCaps và COCO, với việc sử dụng mất mát ngôn ngữ (LM loss). Đặc biệt, phương pháp của BLIP bao gồm việc thêm một câu mở đầu "một bức

ảnh của" vào trước mỗi caption, giúp cải thiện kết quả. Sự tiếp cận này, kết hợp với khả năng lọc và cải thiện dữ liệu web ồn ào, đã cho phép BLIP vượt trội so với các phương pháp sử dụng lượng dữ liệu tiền huấn luyện tương đương, đồng thời cạnh tranh chặt chẽ với các mô hình sử dụng nhiều dữ liệu tiền huấn luyện hơn. BLIP không chỉ chứng minh được sự hiệu quả trong việc sinh ra các caption có ý nghĩa từ hình ảnh mà còn thể hiện khả năng tổng quát hóa mạnh mẽ khi được chuyển giao trực tiếp và không cần chỉnh sửa vào các nhiệm vụ liên quan đến video và ngôn ngữ một cách zero-shot. Điều này làm nổi bật tiềm năng của BLIP trong việc đóng góp vào sự phát triển của các ứng dụng đa phương tiện tương lai, mở rộng khả năng của VLP trong việc hiểu và tạo ra nội dung phong phú giữa hình ảnh và văn bản.

4.6 CoCa: Contrastive Captioners are Image-Text Foundation Models

Mô hình CoCa [22, 17] gần đây đã phát triển từ nhu cầu tổng hợp giữa các phương pháp dual-encoder và encoder-decoder trong tiền huấn luyện mô hình hình ảnh-ngôn ngữ. Các mô hình dual-encoder trước đây đã được huấn luyện trên dữ liệu web lớn với contrastive loss để tạo ra các vector nhúng thị giác có khả năng phân loại hình ảnh và truy xuất hình ảnh-văn bản mà không cần dữ liệu huấn luyện. Tuy nhiên, chúng không trực tiếp hỗ trợ các nhiệm vụ hiểu hình ảnh-ngôn ngữ phức tạp như trả lời câu hỏi hình ảnh (VQA), do thiếu khả năng kết hợp biểu diễn hình ảnh và văn bản.

Một hướng nghiên cứu khác đã đề xuất sử dụng các mô hình encoder-decoder trong tiền huấn luyện để học biểu diễn chung hình ảnh và ngôn ngữ. Trong quá trình này, mô hình sử dụng hình ảnh làm đầu vào cho bộ mã hóa và áp dụng mất mát mô hình ngôn ngữ (LM loss) hoặc PrefixLM trên đầu ra của bộ giải mã. Đối với các tác vụ sau tiền huấn luyện, các đầu ra của bộ giải mã có thể diễn giải về mặt ngữ nghĩa cho cả hình ảnh và ngôn ngữ. Mặc dù mô hình encoder-decoder đã cho thấy hiệu quả trong việc hiểu hình ảnh-ngôn ngữ, chúng không thể hiện khả năng căn chỉnh token văn bản với tensor hình ảnh một cách toàn diện, làm giảm khả năng áp dụng cho các nhiệm vụ cross-modal alignment.

Do đó, CoCa được phát triển như một mô hình thống nhất, kết hợp các ưu điểm của cả hai phương pháp trên, nhằm tạo ra một mô hình nền tảng ngôn ngữ - thị giác duy nhất có khả năng thực hiện cả hai loại nhiệm vụ hiệu quả.

4.6.1 Các tiếp cận của CoCa:

Phương Pháp Single-Encoder Cổ Điển Trong phương pháp tiền huấn luyện single-encoder cổ điển, một bộ mã hóa hình ảnh được huấn luyện thông qua nhiệm vụ phân loại hình ảnh trên một tập dữ liệu chú thích hình ảnh lớn, được thu thập từ cộng đồng như ImageNet, Instagram hoặc JFT. Trong cách tiếp cận này, từ vựng của văn bản chú thích thường được cố định và các chú thích hình ảnh được ánh xạ vào các vectơ lớp rời rạc. Mô hình được huấn luyện sử dụng mất mát cross-entropy, được biểu diễn như sau:

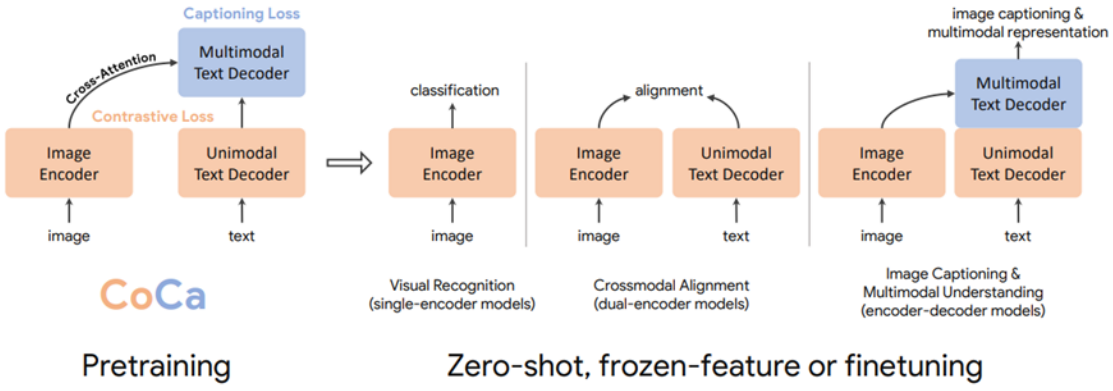
$$\mathcal{L}_{\text{Cls}} = -p(y) \log q_\theta(x), \tag{19}$$

trong đó $p(y)$ đại diện cho các nhãn one-hot, multi-hot hoặc smoothed labels từ ground truth y, và $q_\theta(x)$ là xác suất được dự đoán bởi bộ mã hóa cho lớp y. Bộ mã hóa hình ảnh đã huấn luyện sau đó có thể được sử dụng như một bộ trích xuất biểu diễn hình ảnh chung cho các nhiệm vụ downstream.

Encoder-Decoder Captioning Trong kiến trúc Encoder-Decoder, phương pháp tạo sinh, hay còn gọi là captioning, tập trung vào việc tạo ra các văn bản chi tiết và tỉ mỉ. Mục tiêu là yêu cầu mô hình dự đoán chính xác các token của văn bản đã được mã hóa từ y. Bộ mã hóa hình ảnh cung cấp các đặc trưng mã hóa ẩn, ví dụ, sử dụng một Vision Transformer hoặc ConvNets, trong khi bộ giải mã văn bản học cách tối đa hóa xác suất có điều kiện của chuỗi văn bản y thông qua hàm mất mát sau:

$$\mathcal{L}_{\text{Cap}} = -\sum_{t=1}^{T} \log P_\theta(y_t | y_{<t}, x). \tag{20}$$

Khác với các phương pháp trước đó, phương pháp tạo sinh ra văn bản này đã mang lại một phương thức biểu diễn hình ảnh-văn bản chung có khả năng hỗ trợ hiểu biết hình ảnh-ngôn ngữ, đồng thời cũng mở rộng



Hình 12: Sơ đồ mô hình của CoCa.

khả năng ứng dụng cho việc tạo sinh mô tả hình ảnh bằng văn bản tự nhiên.

4.6.2 Contrastive Captioners Pretraining (CoCa)

Mô hình CoCa (Hình 13) mô tả một kiến trúc encoder-decoder đơn giản nhưng hiệu quả, được tối ưu hóa thông qua ba hàm mất mát tiền huấn luyện khác nhau. Mô hình mã hóa hình ảnh sử dụng một bộ mã hóa mạng nơ-ron như vision transformer (ViT), hoặc các loại bộ mã hóa khác như ConvNets, và sau đó giải mã văn bản thông qua một bộ masking transformer decoder.

Không giống như các mô hình decoder transformer trước đây, CoCa loại bỏ sự chú ý chéo (cross-attention) trong nửa đầu của các lớp decoder để xử lý các biểu diễn văn bản thông qua unimodal. Qua các lớp decoder còn lại, sự chú ý chéo được áp dụng giữa các biểu diễn ngôn ngữ và thị giác, tạo ra biểu diễn hình ảnh-văn bản theo cách multimodal. Điều này cho phép bộ giải mã CoCa tạo ra biểu diễn văn bản từ cả hai quá trình đơn mô hình và đa mô hình, đồng thời hướng đến việc tối ưu hóa mục tiêu tương phản và tạo sinh.

Hàm mất mát tổng hợp của CoCa được định nghĩa như sau:

$$\mathcal{L}_{\text{CoCa}} = \lambda_{\text{Con}} \cdot \mathcal{L}_{\text{Con}} + \lambda_{\text{Cap}} \cdot \mathcal{L}_{\text{Cap}}, \qquad (21)$$

trong đó λ_{Con} và λ_{Cap} là các hệ số trọng số tương ứng với mất mát tương phản và tạo sinh. Chú ý rằng mục tiêu phân loại entropy chéo của một bộ mã hóa đơn có thể được coi là một trường hợp đặc biệt của mục tiêu tạo sinh áp dụng cho dữ liệu chú thích hình ảnh với một tập từ vựng bao gồm tất cả các nhãn.

4.6.3 Decoupled Text Decoder trong Kiến Trúc CoCa

Trong CoCa, chúng tôi tối ưu hóa xác suất có điều kiện của văn bản thông qua một phương pháp mô tả và một biểu diễn văn bản không điều kiện thông qua phương pháp contrastive. Để hợp nhất cả hai phương pháp vào một mô hình duy nhất, chúng tôi đề xuất một thiết kế bộ giải mã tách rời đơn giản. Bộ giải mã chia thành hai phần: phần đơn mô hình và phần đa mô hình. Các lớp giải mã đơn mô hình ở dưới cùng (nuni) mã hóa văn bản đầu vào thành các vectơ ẩn sử dụng masked self-attention mà không cần cross-attention. Các lớp giải mã đa mô hình (nmulti) phía trên áp dụng masked self-attention cùng với cross-attention tới đầu ra của bộ mã hóa hình ảnh.

Cả hai phần của bộ giải mã đều ngăn không cho các token hiện tại nhìn thấy các token tương lai. Việc sử dụng đầu ra của bộ giải mã đa mô hình giúp cho việc tối ưu hóa mục tiêu tạo sinh $\mathcal{L}_{\text{Cap}}$, trong khi đối với mục tiêu tương phản $\mathcal{L}_{\text{Con}}$, một token [CLS] học được được thêm vào cuối chuỗi văn bản đầu vào và đầu ra của bộ giải mã đơn mô hình được sử dụng làm text embedding.

4.6.4 Attentional Poolers trong Kiến Trúc CoCa

Trái ngược với việc sử dụng một single embedding đại diện cho mỗi hình ảnh trong phương pháp contrastive, bộ giải mã trong mô hình encoder-decoder thông thường hướng tới việc sinh ra một chuỗi token. Thực nghiệm cho thấy việc tổng hợp các image embeddings giúp cải thiện nhận dạng hình ảnh thông qua đặc trưng toàn cục, trong khi sử dụng nhiều token hình ảnh hơn lại hữu ích cho các nhiệm vụ đòi hỏi đặc trưng cục bộ và hiểu đa dạng.

CoCa sử dụng task-specific attentional pooling để tùy chỉnh biểu diễn hình ảnh cho các loại mục tiêu huấn luyện khác nhau. Bộ pooler là một lớp multi-head attention với các nquery có thể học được, nơi đầu ra của bộ mã hóa được sử dụng làm cả khóa và giá trị. Mô hình này cho phép học cách tổng hợp các pool embeddings với độ dài khác nhau cho hai mục tiêu huấn luyện. Bộ pooler attentional cụ thể theo nhiệm vụ không chỉ giải quyết nhu cầu khác nhau cho các nhiệm vụ khác nhau mà còn giới thiệu bộ chọn như là một cơ chế chuyển đổi nhiệm vụ tự nhiên. Trong quá trình tiền huấn luyện, chúng tôi sử dụng các bộ chọn chú ý với nquery = 256 cho mục tiêu tạo sinh và nquery = 1 cho mục tiêu tương phản.

4.6.5 Kết luận và ứng dụng của CoCa:

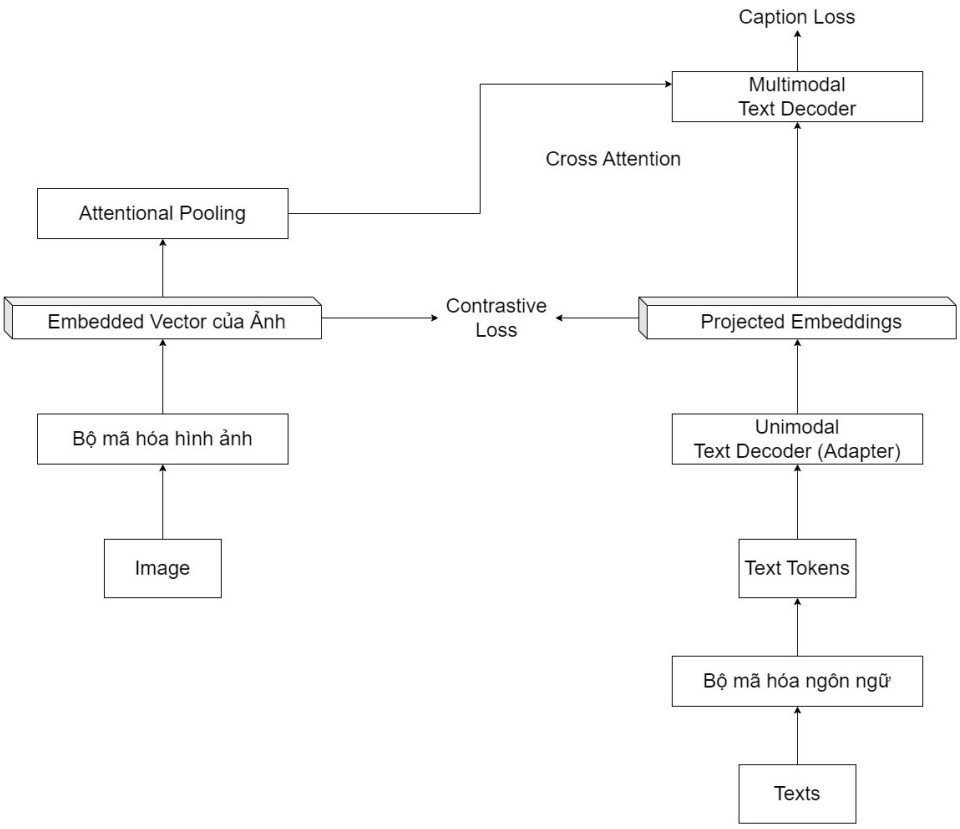
Mô hình CoCa là mô hình encoder decoder có thể làm được các tác vụ khá phổ biến trong lĩnh vực VLM như truy vấn thông tin thị giác (VQA), chú thích văn bản cho sẵn bằng hình ảnh tương ứng (VE – visual entailment), tư duy hình ảnh (VR – visual reasoning) và chú thích hình ảnh bằng văn bản (IC – image captioning). Về các tác vụ VQA, VE và VR, mô hình đã vượt bậc hơn các mô hình đương thời với trở thành mô hình tiên tiến bậc nhất trong cả 3 lĩnh vực trên. Đối với tác vụ IC, mô hình áp dụng hàm lỗi Lcap duy nhất cho tập MSCOCO và đánh giá mô hình trên tập MSCOCO và NoCaps [1, **?**]. Kết quả đạt được của mô hình CoCa hoàn toàn vượt trội so với các mô hình tiền huấn luyện với độ đo CIDEr [16]

5 Giải Quyết Shift Domain thông qua Cross-domain Generalization

Trong bối cảnh nghiên cứu hiện nay, việc áp dụng mô hình học sâu cho các ứng dụng thị giác máy tính đã đạt được những thành tựu đáng kể, nhưng điều này chỉ đúng khi dữ liệu huấn luyện và kiểm thử đến từ cùng một phân phối. Sự dịch chuyển miền (domain shift) — khi phân phối dữ liệu thay đổi từ miền huấn luyện đến miền kiểm thử — là một thách thức đáng kể, gây nên bởi sự khác biệt về phân phối giữa hai miền này. Để giải quyết vấn đề này, học xuyên miền (cross-domain learning) được đề xuất như một hướng tiếp cận nhằm khai thác kiến thức có thể áp dụng qua các miền khác nhau. Trong số các phương pháp học xuyên miền, phát triển giải pháp cho bài toán *domain adaptation* được xem là chìa khóa để giúp mô hình tổng quát hóa tốt trên dữ liệu mới, như việc một xe tự hành cần phải xử lý được các điều kiện thời tiết khác nhau mà nó không được huấn luyện trước đó.

Mặc dù có nhiều phương pháp được đề xuất để giải quyết sự dịch chuyển miền, mô hình cross-domain và các phương pháp học Vision-and-Language (VLMs) như CLIP chủ yếu tập trung vào việc giải quyết *domain adaptation*. Các mô hình này nhằm mục tiêu giảm thiểu khoảng cách giữa miền nguồn và mục tiêu thông qua việc học các đặc trưng mà không thay đổi qua các miền, từ đó cải thiện khả năng tổng quát hóa của mô hình trên dữ liệu không nhìn thấy trong quá trình huấn luyện. Tuy nhiên, việc học các đặc trưng không thay đổi qua các miền từ dữ liệu hình ảnh-văn bản ghép đôi lớn là một thách thức do sự chênh lệch giữa các miền có thể rất lớn.

Cụ thể, *học xuyên miền* không chỉ bao gồm *Domain Generalization (DG)* và *Unsupervised Domain Adaptation (UDA)* mà còn cả các giải pháp cho *domain*



Hình 13: CoCa workflow

adaptation. DG và UDA là những phương pháp nhằm giảm bớt sự thay đổi miền bằng cách học các đặc trưng không phụ thuộc vào miền. Tuy nhiên, trong khi DG tập trung vào việc huấn luyện mô hình từ nhiều miền nguồn và kiểm thử trên miền mục tiêu không được xem trước, UDA tận dụng dữ liệu không nhãn từ miền mục tiêu để cải thiện sự căn chỉnh giữa miền nguồn và mục tiêu. Cả hai phương pháp này đều quan trọng nhưng không thể hoàn toàn giải quyết được vấn đề dịch chuyển miền mà không có sự tập trung vào domain adaptation.

5.1 Tổng quan về VLLaVO

VLLaVO [4] là một phương pháp mới nhằm tích hợp các mô hình ngôn ngữ lớn (LLMs) vào học xuyên miền hình ảnh, qua các bước sau:

- Bước 1: Sử dụng VLMs như CLIP và BLIP để chuyển đổi hình ảnh thành mô tả văn bản chi tiết, từ đó phát triển bộ dữ liệu văn bản cho phân loại.

- Bước 2: Thiết kế các mẫu hướng dẫn để LLMs có thể phân loại hình ảnh dựa trên mô tả văn bản. Sự chênh lệch miền của mô tả vẫn tồn tại, làm giảm hiệu suất xuyên miền.

- Bước 3: Tinh chỉnh LLMs với dữ liệu Q&A, nơi câu hỏi và câu trả lời đề cập đến token lớp của hình ảnh tương ứng. Sử dụng dữ liệu miền mục tiêu với nhãn giả để tinh chỉnh thêm, cải thiện UDA.

Kết quả cho thấy, mặc dù LLMs có khả năng zero-shot ấn tượng cho các nhiệm vụ dựa trên văn bản, việc mở rộng thành công sang các nhiệm vụ hình ảnh và ngôn ngữ là không trực tiếp, chủ yếu do sự khác biệt về modalities và cấu trúc nhiệm vụ. Chưa có công trình nào trước đây tích hợp LLMs với visual cross-domain learning, điều này mở ra một hướng tiếp cận mới trong nghiên cứu.

5.2 Chi tiết kiến trúc (Hình 14):

Trong bối cảnh của học xuyên miền, chúng ta xét tập hợp $\mathcal{S}$ gồm các miền nguồn $\{D_i\}_{i=1}^{s}$, trong đó mỗi D_i bao gồm các cặp dữ liệu và nhãn (x_{ij}, y_{ij}) và n_i là số lượng mẫu trong D_i. Tập hợp $\mathcal{D}$ là tổng hợp dữ liệu huấn luyện từ tất cả các miền nguồn. Mục tiêu của học xuyên miền là huấn luyện một mô hình trên các miền nguồn và sau đó tổng quát hóa nó sang miền mục tiêu $\mathcal{T}$ với phân phối dữ liệu khác biệt. Trong DG, dữ liệu từ miền mục tiêu không được sử dụng trong quá trình huấn luyện, còn trong UDA, dữ liệu không có nhãn từ miền mục tiêu được cho phép sử dụng. Đặc biệt, đối với UDA, $\mathcal{S} = 1$.

VLLaVo bắt đầu bằng cách tạo ra mô tả văn bản cho hình ảnh sử dụng các VLMs đã được huấn luyện từ trước. Những mô tả này có thể thể hiện sự chuyển dịch miền (domain gap). Do đó, chúng tôi tinh chỉnh LLM sử dụng một mẫu hướng dẫn câu hỏi được thiết kế để tích hợp mô tả và nhãn loại hình ảnh. Quá trình tinh chỉnh này cho phép LLM theo dõi mẫu và tập trung vào các thông tin bất biến miền có liên quan đến nhiệm vụ phân loại.

5.2.1 Bước 1: Tạo Mô Tả Văn Bản Từ Hình Ảnh

Sử dụng các mô hình VLM như CLIP và BLIP, chuyển hình ảnh thành mô tả văn bản bao gồm tags, attributes, và captions. CLIP được sử dụng để mã hóa hình ảnh thành vector nhúng và xác định tags dựa trên độ tương đồng cosin. GPT-3 sau đó được sử dụng để bổ sung thông tin đặc điểm và tạo chú thích. Cuối cùng, BLIP kết hợp tất cả thông tin để tạo ra mô tả văn bản chi tiết cho hình ảnh.

5.2.2 Bước 2: Thiết Kế Câu Hỏi Truy Vấn

Từ mô tả văn bản rút trích, LLMs được truy vấn để phân loại hình ảnh. Một mẫu câu hỏi được thiết kế để LLMs có thể định danh mục từ mô tả. $D(x)$ biểu diễn mô tả văn bản của hình ảnh x, và các danh mục ứng cử viên được xem xét. LLM lựa chọn danh mục dựa trên mô tả này, nhằm mục tiêu giảm thiểu ảnh hưởng của domain shift.

5.2.3 Bước 3: Fine-tuning LLM cho Phân Loại

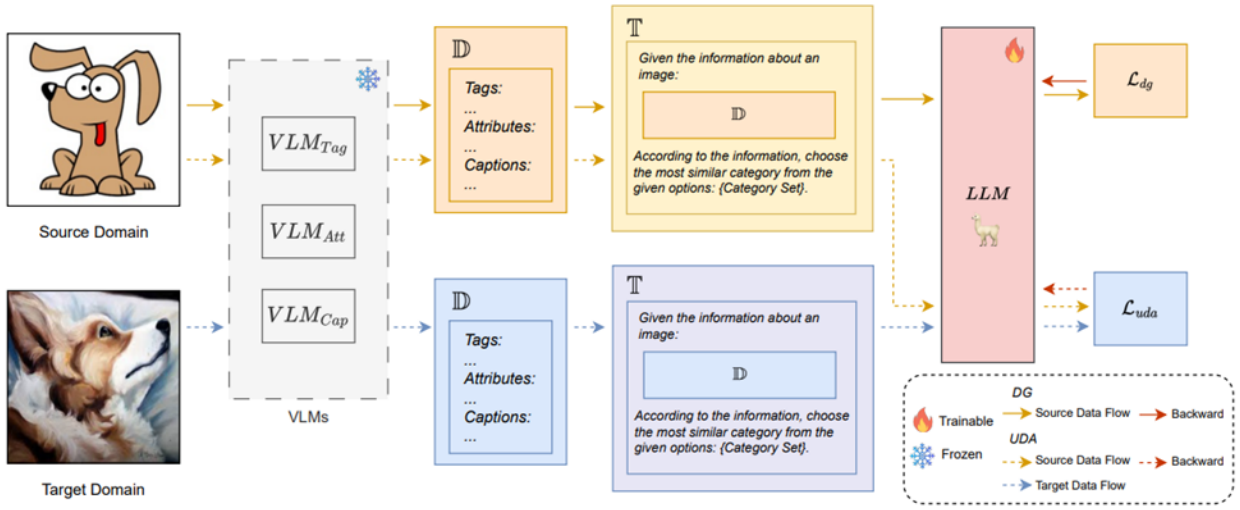
Khi gặp vấn đề phân loại xuyên lĩnh vực, LLM Zero-Shot có thể gặp khó khăn do sự khác biệt trong mô tả văn bản. Cách tiếp cận đề xuất là chuyển vấn đề phân loại thành nhiệm vụ câu hỏi-đáp sử dụng LLM, với nhãn lớp được biểu diễn bởi các token tương ứng, tận dụng khả năng của LLM cho phân loại xuyên lĩnh vực.

5.3 Fine-tuning:

Tinh chỉnh cho Domain Generalization: Domain Generalization đòi hỏi sự tinh chỉnh trên tập hợp dữ liệu $\mathcal{D}$ từ tất cả các miền nguồn. Tuy nhiên, LLM với hàng tỷ tham số cần phương pháp tinh chỉnh hiệu quả. LoRA là phương pháp được chọn, và hàm loss được định nghĩa là:

$$\mathcal{L}_{dg}(\theta) = -\sum_{(x,y)\in\mathcal{D}} \log p_\theta(y \mid \mathcal{T}(x)). \tag{22}$$

Trong đó, $\mathcal{T}(x)$ là mô tả văn bản được tạo ra từ hình ảnh x và p_θ là mô hình dự đoán có tham số θ.
Tinh chỉnh cho UDA: Trong kỹ thuật gán nhãn giả, dữ liệu không được dán nhãn từ miền mục tiêu được kết hợp với dữ liệu đã được dán nhãn từ miền nguồn để tinh chỉnh Large Language Model (LLM). Quá trình này bao gồm việc tinh chỉnh LLM với dữ liệu có sẵn từ miền nguồn và sau đó sử dụng mô hình đã tinh chỉnh để dự đoán nhãn giả cho dữ liệu không



Hình 14: Framework của Cross-domain.

nhãn từ miền mục tiêu. Mục tiêu cuối cùng là giảm thiểu sai số trong quá trình huấn luyện và cải thiện khả năng phân loại của mô hình trên dữ liệu mới.

Hàm mất mát sử dụng trong kỹ thuật gán nhãn giả được biểu diễn như sau:

$$\mathcal{L}_{\text{uda}}(\theta) = -\sum_{(x,y)\in\mathcal{D}} \log p_\theta(y \mid \mathcal{T}(x)) - \sum_{x\in\mathcal{T}} \log p_\theta(\hat{y} \mid \mathcal{T}(x)), \tag{23}$$

trong đó $\mathcal{T}(x)$ là mô tả văn bản từ hình ảnh x, y là nhãn thực tế, và $\hat{y}$ là nhãn giả định dự đoán bởi mô hình.

Hàm mất mát này gồm hai phần:

- Tính tổng log-likelihood của nhãn thực tế y so với dự đoán của mô hình dựa trên câu hỏi $\mathcal{T}(x)$.
- Tính tổng log-likelihood của nhãn giả định $\hat{y}$ so với dự đoán của mô hình cho dữ liệu không có nhãn.

6 Experiments:

6.1 Thông tin về các Datasets:

Trong phần này, nhóm sẽ đề cập đến các tập dữ liệu mà các mô hình nổi tiếng sử dụng.

Clip Prefix

- **Tập CoCo:** Tập dữ liệu nổi tiếng bao gồm các hình ảnh được dán nhãn và chú thích mô tả.
- **Tập Conceptual Captions:** Tập dữ liệu của Google, chứa hơn 3 triệu hình ảnh với hơn 300.000 ảnh được chú thích mô tả.

Dù không được công bố nhưng phương thức gán nhãn và chú thích của 2 tập dữ liệu trên đều kết hợp từ các công đoạn thủ công và tự động.

Open CLIP

- **Tập LAION-400M:** Một tập dữ liệu được qua công đoạn CLIP-filtered với 400 triệu cặp dữ liệu ảnh-văn bản, thêm vào đó là các CLIP embeddings và các kNN indices giúp cho việc truy vấn ảnh tương ứng dễ dàng và hiệu quả hơn. Các

hình ảnh và văn bản được trích từ dữ liệu web Common Crawl và đến từ các trang web ngẫu nhiên được crawl từ năm 2014-2021.

- **Tập LAION-2B:** Một tập dữ liệu bao gồm 5.85 tỷ cặp ảnh-văn bản được trải qua CLIP-filtered, với 2.32 tỷ gồm có ngôn ngữ Anh.
- **Tập Datacomp-1B:** Tập hợp các hình ảnh được lấy từ Common Crawl, bao gồm 1.4 tỷ cặp hình ảnh-văn bản.

Tương tự như trước thì trong các cặp tập dữ liệu trên, công việc gán nhãn được thực hiện thủ công và tự động nhờ vào các script để giảm thời gian và nguồn lực.

BLIP

- **Tập COCO:** Tương tự như các mô hình trên, là một tập dữ liệu nổi tiếng.
- **Tập NoCaps:** Bao gồm 166.100 chú thích thủ công cho 15.100 hình ảnh từ OpenImages và được thông qua bởi các bộ phận kiểm tra và nhận dạng của OpenImages.

Nhóm sẽ thực nghiệm trên nhiều mã nguồn khác nhau để kiểm tra độ chính xác của kết quả của các mô hình đã trình bày như CLIP, BLIP, COCA, SIMVLM. Tuy nhiên trên tìm hiểu thực tế của nhóm thì có quá ít các mã nguồn của COCA và SimVLM để nhóm có thể chạy thực nghiệm và so sánh, cho nên các kết quả so sánh được trong phần này sẽ chỉ giới hạn ở các thông số huấn luyện. Còn lại, đối với CLIP và BLIP có một số mã nguồn chạy thành công được, với CLIP đạt được độ chính xác cao hơn so với mô hình còn lại. Ảnh đầu vào của toàn bộ các mô hình trong lúc thực nghiệm đều sử dụng Hình 15

6.2 CLIP Interrogator.

CLIP Interrogator là một công cụ được phát triển để tạo ra các prompt tự động bằng cách sử dụng cả hai mô hình CLIP của OpenAI và BLIP của Salesforce. Công cụ này tối ưu hóa các từ khóa phù hợp với hình ảnh đầu vào. Các thông số ban đầu cho công cụ này bao gồm:

Hình 15: Hình ảnh núi Phú Sĩ, Nhật Bản.

- Mô hình chú thích (Caption Model): blip-base.
- Mô hình CLIP (Clip Model): ViT-L-14 cung cấp bởi OpenAI.

Ví dụ, khi đưa vào một bức ảnh của núi Phú Sĩ, CLIP Interrogator sẽ tạo ra các từ khóa mô tả hình ảnh đó. Để tìm hiểu thêm và sử dụng công cụ này, bạn có thể truy cập mã nguồn tại `https://github.com/pharmapsychotic/clip-interrogator`.

Và như ta thấy thì các từ khóa ta có được khá chính xác so với bức ảnh đầu vào. Các từ khóa như "mount fuji in the background", "at snowy fuji mountain sunrise" khớp với đầu vào là ảnh núi Phú sĩ.

6.3 CLIP prefix captioning:

CLIP Prefix Captioning là công cụ dùng mô hình CLIP cho việc mã hóa hình ảnh và tạo chú thích (caption) tự động. Mô hình này được tiền huấn luyện sử dụng bộ dữ liệu COCO. Thông số cấu hình của mô hình bao gồm:

- Mô hình chú thích (Caption Model): Clip caption model.
- Mô hình CLIP (Clip Model): ViT-B/32.

Trong quá trình thử nghiệm với một hình ảnh của núi Phú Sĩ, công cụ đã sinh ra chú thích phản ánh nội dung của hình ảnh. Để tham khảo mã nguồn và

thực hiện thử nghiệm, truy cập https://github.com/rmokady/CLIP_prefix_caption.

Như vậy ta thấy là mã nguồn này chưa thể nhận dạng được cảnh trong bức ảnh trên là cảnh của núi Phú sĩ. Nhưng nó vẫn có khả năng nhận dạng được các chi tiết chung chung như là núi và hồ.

6.4 OpenCLIP + CoCa

Mã nguồn OpenCLIP được sử dụng để tạo chú thích cho hình ảnh sử dụng một bộ encoder-decoder tiên tiến. Cấu hình được đề cập trong mã nguồn này như sau:

- Mô hình chú thích (Caption Model): coca-ViT-L-14.
- Mô hình tiền huấn luyện (Pretrained Model): mscoco_finetuned_laion2B-s13B-b90k.

Với cấu hình này, mô hình sử dụng bộ mã hóa và giải mã từ thư viện OpenCLIP để sinh ra chú thích cho hình ảnh. Khi thực hiện thử nghiệm với hình ảnh của núi Phú Sĩ, mô hình đã tạo ra chú thích tương ứng. Mã nguồn có thể được truy cập tại https://github.com/mlfoundations/open_clip.

6.5 Salesforce's BLIP

Mã nguồn BLIP, phát triển bởi Salesforce, cung cấp khả năng thực hiện một loạt các tác vụ liên quan đến hình ảnh và văn bản, bao gồm chú thích hình ảnh (image captioning), trả lời câu hỏi hình ảnh (VQA), trích xuất đặc trưng (feature extraction), và so khớp hình ảnh-văn bản (image-text matching). Mã nguồn này sử dụng một bộ giải mã BLIP với một mô hình tiền huấn luyện cố định, cụ thể là *model_base_capfilt_large*. Để khám phá mã nguồn và tìm hiểu thêm về các tác vụ mà BLIP có thể thực hiện, bạn có thể truy cập tại https://github.com/salesforce/BLIP.

6.6 SimVLM

Mặc dù không có nhiều mã nguồn mở có sẵn trình bày mô hình SimVLM có thể chạy một cách hoàn thiện, nhóm nghiên cứu của chúng tôi đã tìm thấy một số kho lưu trữ GitHub liên quan đến mô hình này. Điểm chung của các mã nguồn này là tác giả của chúng đã huấn luyện mô hình SimVLM sử dụng các mô hình tiền huấn luyện (pretrained models) có sẵn:

- https://github.com/Yuhei-Handa/SimVLM
- https://github.com/FerryHuang/SimVLM
- https://github.com/YulongBonjour/SimVLM

Tuy nhiên, mã nguồn không được cung cấp bởi các tác giả để có thể chạy đầy đủ và hoàn chỉnh, do đó nhóm chưa thể kiểm tra toàn diện khả năng của các mã nguồn này.

6.7 Đánh giá khả năng:

Tất cả mã nguồn đều sẽ được thực nghiệm trên hình ảnh núi Phú sĩ. Sau đây là một bảng tổng hợp lại các kết quả dự đoán caption và hiệu năng của các mô hình trên 16:

Mã nguồn	Thông số đầu vào	Loại kết quả	Kết quả	Thời gian chạy	Ưu và nhược điểm
CLIP Interrogator	- Caption Model: blip-base. - Clip Model: ViT-L-14/openai	Prompt	A mountain in the distance, mount fuji on the background, at snowy fuji mountain sunrise, mountain fuji on the background, mount fuji in the background, japan mountains, japan nature,...	0.8 giây *(sử dụng phương thức "fast" của bộ generator)*	Ưu điểm: - Có thể chạy nhiều ảnh cùng một lúc Nhược điểm: - Chỉ tối ưu khi chạy trên GPU
CLIP prefix captioning	- Caption model: Clip caption model - Clip model: ViT-B/32	Caption	A lake with a large mountain in the background.	3 giây	Ưu điểm: - Tối ưu khi chạy trên CPU và GPU Nhược điểm:
OpenCLIP + COCA	- Caption model: coca-ViT-L-14 - Pretrained model: mscoco_finetuned_laion2B-s13B-b90k.		A view of a mountain with a lake in the foreground	25 giây	Ưu điểm: Nhược điểm: - Chỉ tối ưu khi chạy trên GPU
BLIP	- Caption model: blip - Pretrained model: model_base_capfilt_large		A mountain	11 giây	Ưu điểm: Nhược điểm: - Chỉ tối ưu khi chạy trên GPU
SimVLM	N/A	N/A	N/A	N/A	N/A

Hình 16: Bảng đánh giá khả năng

Tài liệu

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2019.

[2] J. Cha, S. Chun, K. Lee, C. Ho-Choi, Y. Park, S. Lee, and S. Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:24025–24128, 2021.

[3] J. Cha, K. Lee, S. Park, and S. Chun. Domain generalization by mutual-information regularization with pre-trained models. In *European Conference on Computer Vision*, pages 440–457. Springer, 2022.

[4] Shuhao Chen, Yulong Zhang, Weisen Jiang, Jiangang Lu, and Yu Zhang. Vllavo: Mitigating visual gap through llms, 2024.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186, 2019.

[6] Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

[7] Li Hang, Zeng Yan, and Zhang Xinsong. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *Name of the Conference or Journal*, page Page Range. Organization or Publisher, 2021.

[8] Y. Jin, M. Xu, M. Wang, M. Long, and J. Wang. Minimum class confusion for versatile domain adaptation. In *European Conference on Computer Vision*, pages 464–480. Springer, 2020.

[9] Taehyeong Kim, Hyeonseop Song, and Byoung-Tak Zhang. Cross-modal alignment learning of vision-language conceptual systems. *arXiv preprint arXiv:2208.01744*, 2022.

[10] Wonjae Kim, Bokyung Son, Ildoo Lee, Minhwan Kang, and Jinwoo Lee. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[11] Gregory R. Koch. Siamese neural networks for one-shot image recognition. 2015.

[12] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[13] Lizhao Liu, Xinyu Sun, Tianhang Xiang, Zhuangwei Zhuang, Liuren Yin, and Mingkui Tan. Contrastive vision-language alignment makes efficient instruction learner, 2023.

[14] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: Clip prefix for image captioning, 2021.

[15] Archit Parnami and Minwoo Lee. Learning from few examples: A summary of approaches to few-shot learning, 2022.

[16] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning, 2017.

[17] Research Research, Scientists Google and Brain Team. Image-text pre-training with contrastive captioners.

[18] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[19] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019.

[20] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision, 2022.

[21] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[22] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022.

[23] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[24] Y. Zhang, X. Wang, J. Liang, L. Zhang, R. Wang, T. Jin, and T. Tan. Free lunch for domain adversarial training: Environment label smoothing. In *International Conference on Learning Representations*, 2023.