MeaCap: Chú thích hình ảnh Zero-shot được tăng cường bộ nhớ

Zequn Zeng*, Yan Xie*, Hao Zhang† , Chiyu Chen, Bo Chen
Phòng thí nghiệm trọng điểm quốc gia về xử lý tín hiệu radar, Đại học Xidian, Tây An, 710071, Trung Quốc
{zzequn99, yanxie0904, zhanghao xidian}@163.com, {chenchiyu, bchen}@mail.xidian.edu.cn

Vương Chính Giác
Phòng thí nghiệm trọng điểm nhà nước về mạng lưới dịch vụ tích hợp, Đại học Xidian, Tây An, 710071, Trung Quốc
wangzhengjue@xidian.edu.cn

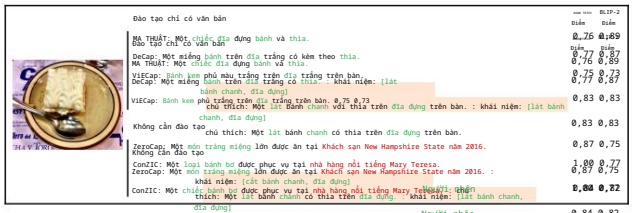
Tóm tắt

Chú thích hình ảnh không cần chụp (IC) mà không cần ghép nối tốt dữ liệu hình ảnh-văn bản có thể được chia thành hai loại, đào tạo không cần đào tạo và đào tạo chỉ có văn bản. Nhìn chung, hai các loại phương pháp thực hiện IC không bắn bằng cách tích hợp các mô hình ngôn ngữ thị giác được đào tạo trước như CLIP cho hình ảnh-văn bản đánh giá sự tương đồng và mô hình ngôn ngữ được đào tạo trước (LM) để tạo chú thích. Sự khác biệt chính giữa chúng là liệu có sử dụng ngữ liệu văn bản để đào tạo LM. Mặc dù đạt được hiệu suất hấp dẫn wrt một số số liệu, các phương pháp hiện có thường thể hiện một số nhược điểm chung. Các phương pháp không cần đào tạo có xu hướng tạo ra ảo giác, trong khi đào tạo chỉ bằng văn bản thường mất khả năng khái quát hóa. Để tiến lên, trong bài báo này, chúng tôi đề xuất một hình ảnh zero-shot tăng cường bộ nhớ mới lạ Khung chú thích (MeaCap). Cụ thể, được trang bị với bộ nhớ văn bản, chúng tôi giới thiệu một phương pháp lấy-rồi-lọc mô-đun để có được các khái niệm chính có liên quan cao đến hình ảnh. Bằng cách triển khai bộ nhớ tăng cường được đề xuất của chúng tôi điểm số hợp nhất liên quan đến hình ảnh trong LM từ khóa thành câu, MeaCap có thể tạo ra các chú thích tập trung vào khái niệm giúp duy trì tính nhất quán cao với hình ảnh với ít ảo giác hơn và nhiều kiến thức về thế giới hơn. Khung của Mea-Cap đạt được hiệu suất tiên tiến trên một loạt các thiết lập IC không có cú đánh. Mã của chúng tôi có sẵn tại

https://github.com/joeyz0z/MeaCap.

1. Giới thiệu

Chú thích hình ảnh (IC) nhằm mục đích hiểu nội dung trực quan và tạo ra các mô tả văn bản. Sử dụng chú thích tốt cặp hình ảnh-văn bản, mô hình có giám sát [7, 17, 23, 34, 39, 41, 48, 49, 56] đã đạt được kết quả khả quan trên IC điển hình chuẩn mực [1, 25, 32, 63]. Do chi phí cao của ký hiệu, các bộ đào tạo của các chuẩn mực này thường liên quan đến các kiểu/nội dung hình ảnh hạn chế, đây là một trở ngại khó khăn



(a) Hiện tượng ảo giác.

(b) Hình ảnh chứa đựng kiến thức thế giới.

Hình 1. Động lực của MeaCap được đề xuất của chúng tôi, nơi màu đỏ là không chính xác và màu xanh lá cây là đúng. (a) Các phương pháp không cần đào tạo liên kết hình tròn với thông tin vị trí không chính xác, thực tế là đạt điểm cao trong CLIPscore. Điều này có thể là do thực tế là CLIP được đào tạo trên dữ liệu hình ảnh-văn bản nhiễu theo quy mô web. (b) Các phương pháp đào tạo chỉ văn bản (ToT) hiện có không tạo ra được người nhện như một số phương pháp không cần đào tạo thì có, nhưng phiên bản ToT của phương pháp của chúng tôi (MeaCapToT) cũng có thể làm được điều đó.

cle cho những mô hình được giám sát đó được khái quát thành hình ảnh trong tự nhiên. Để hiện thực hóa IC mà không cần cặp hình ảnh-văn bản do con người chú thích, gần đây, IC zero-shot đã thu hút sự chú ý ngày càng tăng. Các tác phẩm hiện có chủ yếu có thể được chia thành hai nhóm, phương pháp không cần đào tạo và phương pháp đào tạo chỉ có văn bản. Các phương pháp không cần đào tạo [53, 54, 64] thực hiện zero-shot tạo hình ảnh thành văn bản bằng cách sử dụng các mô hình được đào tạo trước mà không cần tinh chỉnh. Cụ thể, họ sử dụng một mô hình ngôn ngữ thị giác được đào tạo trước như CLIP để hướng dẫn một ngôn ngữ được đào tạo trước mô hình (LM), chẳng hạn như BERT [8] hoặc GPT-2 [44], để tạo ra các câu phù hợp với hình ảnh đã cho. Với lặp đi lặp lại suy ra, công việc này không đòi hỏi bất kỳ sự đào tạo nào. Mặc dù đã đạt được khả năng khái quát hóa cao hơn và điểm CLIPscore cao hơn [16], các phương pháp này cho thấy các yếu tố bên ngoài

*Đóng góp ngang nhau. †Tác giả liên hệ

hiện tượng lucination, tức là, họ có xu hướng tạo ra một câu chuyện chứa thông tin tưởng tượng có thể không tồn tại trong hình ảnh cho sẵn, như thể hiện trong Hình 1a.

Để giảm bớt vấn đề này, một tuyến đường sắt khác sẽ đào tạo hoặc tinh chỉnh bộ giải mã văn bản dựa trên dữ liệu văn bản chất lượng cao không có hình ảnh tương ứng, được gọi là đào tạo chỉ có văn bản phương pháp [12, 29, 40, 51, 55]. Để kiểm tra hình ảnh chứa các đối tượng được mô tả trong ngữ liệu đào tạo, các phương pháp đào tạo chỉ có văn bản tạo ra chú thích một cách khách quan, đạt được cải thiện đáng kể về điểm số dựa trên tham chiếu như như BLEU [42], METEOR [4] và CIDEr [57]. Tuy nhiên, do khối lượng đào tạo hạn chế, kiến thức chứa đựng trong LM được đào tạo trước dần dần bị lãng quên trong quá trình đào tạo, dẫn đến suy giảm hiệu suất nghiêm trọng trên dữ liệu ngoài miền, như thể hiện trong Hình 1b. Mặc dù đào tạo trên ngữ liệu chất lượng cao trên web là một giải pháp tiềm năng, do đó tạo ra chi phí tính toán cực kỳ cao.

Để duy trì khả năng khái quát tốt cho hình ảnh trong hoang dã và để thoát khỏi trí tưởng tượng vô lý, bài báo này đề xuất một khuôn khổ chú thích hình ảnh zero-shot tăng cường trí nhớ mới , cụ thể là MeaCap, dựa trên cơ chế hướng dẫn bộ nhớ, cung cấp một giải pháp thay thế kế hoạch sử dụng ngữ liệu phụ đề thay vì sử dụng nó để đào tạo LM. Cụ thể, từ bộ nhớ văn bản bên ngoài, chúng tôi phát triển một mô-đun lấy-sau-lọc để tìm khóa các khái niệm có liên quan cao đến hình ảnh đã cho. Giới thiệu về hình ảnh liên quan đến trí nhớ được tăng cường của chúng tôi điểm hợp nhất cho từ khóa thành câu LM, CBART [15], MeaCap có thể tạo ra các chú thích tập trung vào khái niệm giúp duy trì tính nhất quán cao với hình ảnh. Hình ảnh mới này liên quan đến điểm không chỉ xem xét sự tương đồng giữa hình ảnh và văn bản giữa các phương thức như hầu hết các phương pháp IC không có cú đánh nào [51, 53–55, 64] thực hiện CLIP nhưng cũng xem xét sự tương đồng trong phương thức văn bản-văn bản bằng đánh giá sự giống nhau giữa chú thích và nội dung đã lấy được bộ nhớ liên quan đến hình ảnh. MeaCap được đề xuất của chúng tôi có thể là MeaCapTF không cần đào tạo hoặc đào tạo chỉ bằng văn bản được đặt tên là MeaCapToT bằng cách tinh chỉnh CBART.

Những đóng góp của chúng tôi được tóm tắt như sau:

- Chúng tôi sử dụng tập hợp chú thích chỉ có văn bản làm bộ nhớ ngoài để tăng cường IC không cần đào tạo. Để Với mục đích này, chúng tôi giới thiệu một mô-đun truy xuất rồi lọc để trích xuất các khái niệm chính từ bộ nhớ và thực hiện các thế hệ tập trung vào khái niệm bằng CBART để giảm bớt vấn đề ảo giác của các phương pháp không cần đào tạo trước đây.
- Dựa trên bộ nhớ văn bản đã thu thập được, chúng tôi phát triển một điểm số hợp nhất liên quan đến thị giác được tăng cường trí nhớ vào CBART, cải thiện mối tương quan giữa hình ảnh và tạo chú thích trong khi vẫn giữ lại kiến thức của thế giới.
- Các thí nghiệm mở rộng dưới điều kiện không có phát bắn, trong miền, và các kịch bản liên miền chứng minh đề xuất của chúng tôi thiết kế tăng cường bộ nhớ có thể cải thiện đáng kể sự nhất quán với nội dung hình ảnh trong cả hai chương trình đào tạo miễn phí và cài đặt đào tạo chỉ có văn bản.

2. Công việc liên quan

2.1. Chú thích hình ảnh có giám sát

IC được giám sát thường sử dụng các cặp hình ảnh-văn bản được căn chỉnh tốt và đào tạo một mô hình mã hóa-giải mã. Ví dụ, một số những nỗ lực ban đầu [9, 13, 58, 60] xây dựng bộ mã hóa dựa trên CNN để trích xuất các tính năng trực quan và bộ giải mã dựa trên RNN/LSTM để tạo ra câu đầu ra. Để hiểu trực quan hơn, một số phương pháp [3, 7, 17, 18, 26, 43, 59] sử dụng máy dò đối tượng để trích xuất các vùng hình ảnh chú ý. Để khuyến khích nhiều tương tác hơn giữa hai phương thức, cơ chế chú ý [7, 17, 39, 41, 48, 49] và đồ thị thần kinh mạng [61, 62] đã được áp dụng rộng rãi.

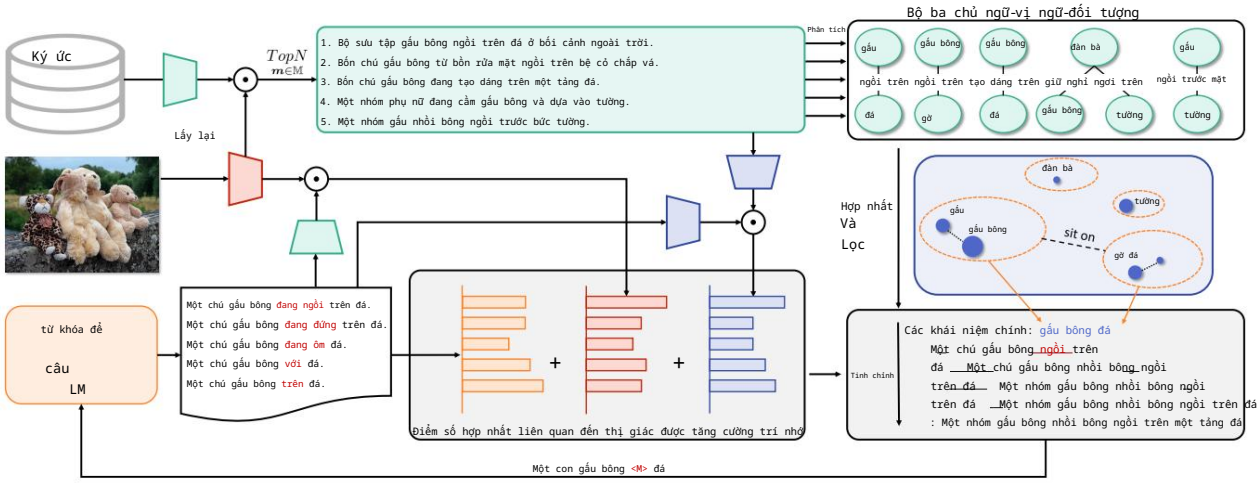
2.2. Chú thích hình ảnh Zero-shot

Gần đây, IC zero-shot ngày càng được chú ý nhiều hơn, nhằm mục đích tạo chú thích cho hình ảnh trong hai trường hợp: i) không có dữ liệu nào để đào tạo tên là IC zero-shot không đào tạo; ii) chỉ sử dụng văn bản từ tập dữ liệu chú thích để huấn luyện LM có tên là text-only-training zero-shot IC.

Các phương pháp không cần đào tạo hiện thực hóa IC không bắn thông qua mô hình ngôn ngữ thị giác được đào tạo trước [45], để hướng dẫn việc tạo ra LM được đào tạo trước. Cụ thể, ZeroCap [54] và phần mở rộng của nó [53] cho phụ đề video được đề xuất dựa trên trên vòng lặp tìm kiếm gradient. Để tạo IC zero-shot có thể kiểm soát được, ConZIC [64] được đề xuất bằng cách kết hợp lấy mẫu Gibbs với LM không tự hồi quy, cải thiện tính đa dạng và tốc độ suy luận của IC. Mặc dù chúng đạt được khả năng khái quát hóa cao hơn với CLIPscore cao hơn [16], họ có thể tạo ra một số mô tả không xuất hiện trong hình ảnh, được gọi là ảo giác như thể hiện trong Hình 1a.

Phương pháp đào tạo chỉ văn bản đào tạo hoặc tinh chỉnh văn bản bộ giải mã chỉ sử dụng ngữ liệu từ tập dữ liệu chú thích. Cụ thể, sau khi tinh chỉnh simCTG có sẵn [52] trực tiếp trên ngữ liệu cụ thể, MAGIC [51] và ZERO-GEN [55] được đề xuất bằng cách giới thiệu một CLIP được tạo ra điểm để điều chỉnh quá trình tạo ra simCTG, làm cho chú thích có liên quan về mặt ngữ nghĩa với một hình ảnh nhất định. Bằng cách liên quan đến câu gốc hoặc câu nhúng như lời nhắc nhở để đào tạo một LM, DeCap [29], CapDec [40] và ViECap [12] được phát triển bằng cách lập bản đồ đặc điểm trực quan đến tính năng văn bản, sau đó được đưa vào LM này để tạo chú thích.

MeaCap được đề xuất của chúng tôi có thể thực hiện cả hai chức năng không cần đào tạo và IC zero-shot đào tạo chỉ có văn bản. Đối với đào tạo không có thiết lập, vì chúng tôi giới thiệu cơ chế bộ nhớ cho xác định khái niệm chính (Phần 3.1) và hướng dẫn cho LM thế hệ (Phần 3.3), phương pháp của chúng tôi có thể tạo ra các chú thích chính xác hơn với ít ảo giác hơn. Đối với đào tạo chỉ có văn bản thiết lập, việc sử dụng ngữ liệu như bộ nhớ ngoài trong chúng ta phương pháp có thể làm giảm bớt vấn đề của các phương pháp hiện có quên đi kiến thức thế giới được học bởi LM được đào tạo trước để điều chỉnh chính xác từng ngữ liệu.



Hình 2. Tổng quan về MeaCap được đề xuất của chúng tôi. Luồng dữ liệu tổng thể theo chiều kim đồng hồ. i) Cho một hình ảnh, trước tiên chúng tôi lấy các mô tả có liên quan Top-N từ bộ nhớ, được chuyển đổi thành bộ ba chủ ngữ-vị ngữ-đối tượng; chúng tôi hợp nhất và lọc các nút để có được các khái niệm chính Mục 3.1. ii) Với điểm số hợp nhất liên quan đến hình ảnh được tăng cường trí nhớ (Mục 3.3), bắt đầu từ các khái niệm chính, LM từ khóa thành câu có thể hoàn thiện mô tả hình ảnh bằng cách tinh chỉnh lặp lại (Mục 3.2). Ei, Et, St là bộ mã hóa hình ảnh CLIP, bộ mã hóa văn bản CLIP, biểu thị độ tương tự BERT, tương ứng là điểm số lưu loát trong Công thức (7), độ tương tự giữa các phương thức hình ảnh và bộ nhớ trong Công thức (8) và độ tương tự trong phương thức của chú thích-bộ nhớ Công thức (9), tương ứng.

2.3. Bộ nhớ ngoài trong chú thích hình ảnh

Người ta đã chứng minh rằng việc đưa bộ nhớ ngoài vào rất hữu ích cho nhiều tác vụ ngôn ngữ và thị giác khác nhau, như quá trình ngôn ngữ tự nhiên [6, 14, 21, 22, 37], nhận dạng thị giác [33, 35], tổng hợp hình ảnh [5, 10], trả lời câu hỏi miền mở [20, 27] và bao gồm IC [29, 46]. Ví dụ, SmallCap [46] là phương pháp IC có giám sát sử dụng CLIP để truy xuất một số chú thích có liên quan, sau đó lấy các chú thích này làm lời nhắc cho LM, chứng minh rằng bộ nhớ có thể giúp LM tạo ra các chú thích chính xác với ít tham số đào tạo hơn. Trong chú thích không cảnh, De-Cap [29] đào tạo LM để đảo ngược nhúng văn bản CLIP thành câu tương ứng. Nó chiếu nhúng trực quan CLIP thành tổng có trọng số của nhúng bộ nhớ văn bản và lấy nhúng văn bản cuối cùng làm lời nhắc mềm cho LM để hướng dẫn tạo chú thích.

So với DeCap và SmallCap, tận dụng toàn bộ câu nhớ như một lời nhắc để hướng dẫn tạo ra, chúng tôi đề xuất một bộ lọc không cần đào tạo loại bỏ thông tin nhiễu để lấy các khái niệm chính từ bộ nhớ văn bản đã lấy được. Không giống như bộ nhớ trong DeCap chỉ được thiết kế cho các phương pháp đào tạo chỉ có văn bản, thiết kế bộ nhớ rõ ràng của chúng tôi có thể được áp dụng cho cả các tình huống đào tạo không cần đào tạo và chỉ có văn bản và cho thấy khả năng vượt trội trong việc tạo ra các chú thích chính xác hơn.

3. Thuốc MeaCap

Để có IC không bắn tốt hơn với ít ảo giác hơn và lưu giữ nhiều kiến thức thế giới hơn, như thể hiện trong Hình 2, chúng tôi đề xuất một khuôn khổ mới gọi là MeaCap. i) Để giải quyết vấn đề

lem của các phương pháp không cần đào tạo hiện có [53, 54, 64] có thể gây ảo giác trong chú thích, MeaCap xác định một số khái niệm chính từ bộ nhớ văn bản được truy xuất có liên quan cao đến hình ảnh và thực hiện chú thích tập trung vào khái niệm (Phần 3.1). ii) Chúng tôi phát triển một điểm số hợp nhất liên quan đến hình ảnh được tăng cường trí nhớ (Phần 3.3), xem xét cả sự tương đồng giữa hình ảnh và văn bản và sự tương đồng giữa văn bản và văn bản trong phương thức (giữa bộ nhớ văn bản và chú thích), được giới thiệu trong LM từ khóa đến câu, CBART [15]
(Phần 3.2), cải thiện mối tương quan giữa hình ảnh và chú thích.

3.1. Lấy-sau-lọc để có được các khái niệm chính

Các phương pháp IC zero-shot huấn luyện chỉ văn bản hiện có [12, 29, 40, 51, 55] thường huấn luyện hoặc tinh chỉnh LM trên các văn bản từ tập dữ liệu chú thích, mang lại các mô tả phù hợp hơn với ít ảo giác hơn. Tuy nhiên, các phương pháp như vậy làm cho các chú thích được tạo ra quá khớp với một ngữ liệu cụ thể, thiếu khái quát hóa ngoài phân phối. Được thúc đẩy bởi hiện tượng này, thay vì huấn luyện hoặc tinh chỉnh LM trên các văn bản, chúng tôi chỉ xây dựng một bộ nhớ văn bản tăng cường để có được các khái niệm chính, sau đó có thể hướng dẫn IC zero-shot.

Xây dựng bộ nhớ tăng cường. Để đạt được mục đích này, trước tiên chúng ta xây dựng một bộ nhớ văn bản lớn M chứa nhiều câu liên quan đến hình ảnh với nhiều khái niệm hình ảnh.
Ký ức này có ý nghĩa quan trọng trong việc loại bỏ ảo giác đối với trường hợp không được đào tạo và có thể làm giảm bớt kiến thức bị quên đối với trường hợp chỉ được đào tạo bằng văn bản.

Lấy lại các mô tả liên quan đến hình ảnh. Sau khi có được bộ nhớ, cho một hình ảnh I, chúng tôi sử dụng CLIP để đánh giá độ tương đồng giữa hình ảnh và văn bản để lấy lại Top-N hình ảnh-

mô tả liên quan từ bộ nhớ như {mn} $_{n=1}^{Nd}$:

(1)

trong đó Ei(·) và Et(·) biểu thị bộ mã hóa hình ảnh và văn bản trong CLIP, tương ứng; cos(·, ·) là độ tương tự cosin.

Bộ ba chủ ngữ-vị ngữ-đối tượng. Để giảm thêm tác động của một số từ ít thông tin trong hình ảnh- Nd mô tả liên quan {mn} $_{n=1}$, chẳng hạn như mạo từ và giới từ vị trí, chúng tôi sử dụng một trình phân tích cú pháp văn bản có sẵn, TextGraph-Parser [31], để chuyển đổi từng mô tả mn thành một đồ thị văn bản gn bao gồm nhiều chủ ngữ-vị ngữ-đối tượng bộ ba, trong đó chủ ngữ và tân ngữ là các nút trong khi các vị ngữ là mối quan hệ. Các nút này được coi là các khái niệm ứng viên sẽ được lọc và hợp nhất để tạo thành một tập hợp các khái niệm chính. Các mối quan hệ sẽ quyết định thứ tự giữa hai khái niệm. Chúng tôi định nghĩa {vn} $_{n=1}^{Nc}$ như tập hợp của tất cả các nút từ tất cả các đồ thị văn bản Nd {gn} $_{n=1}^{Nd}$.

Kết hợp và lọc để có được các khái niệm chính. Như đã hiển thị trong Hình 2, một số nút biểu thị các khái niệm có thể biểu diễn cùng một vật thể trong hình ảnh (ví dụ, "gấu" và "gấu bông"), trong khi một số có thể không liên quan đến hình ảnh (ví dụ, "phụ nữ"), cần được hợp nhất và lọc trước khi lấy được các khái niệm chính.

i) Hợp nhất. Với sự trợ giúp của bộ mã hóa văn bản từ Sentence-BERT [47], St(·), chúng ta có thể thu được tập nhúng khái niệm {v $_{N}^{c}$} $_{n=1}^{Nc}$ như f_{N}^{c} = St(vn). Sau đó chúng ta đánh giá độ tương đồng giữa bất kỳ hai khái niệm nhúng nào như

(2)

Sau đó, chúng tôi đặt siêu tham số τ làm ngưỡng, trong đó dij > τ biểu thị rằng khái niệm thứ i và khái niệm thứ j thuộc về đến cùng một cụm. Sau đó, tổng cộng chúng ta có Nv con- Ncn chấp nhận các cụm như cn = {vi} $_{tôi=1}^{Nv}$ nơi Ncn biểu thị số lượng nút trong cụm khái niệm thứ n cn.

2) Lọc. Trong bước này, chúng ta cần quyết định xem cụm khái niệm thứ n cn có bị xóa hay được giữ lại không. Vì mục đích này, một giả định hợp lý là từ không liên quan đến hình ảnh có vẻ ngoài thấp hơn trong các mô tả được lấy lại {phút} $_{n=1}^{Nd}$ trong Phương trình (1). Do đó, chúng tôi tính toán tần số cụm khái niệm CF(cn) bằng cách dần dần xem liệu vi từ cn xuất hiện trong mk như

trong đó CF(cn) biểu thị tần số của cụm thứ n xuất hiện trong các mô tả đã truy xuất {mn} $_{n=1}^{Nd}$. Theo kinh nghiệm, nếu CF(cn) > 0,5, chúng tôi sẽ giữ lại cụm này cn và nếu không thì xóa nó. Cuối cùng, chúng tôi lọc ra các cụm khái niệm khóa nv từ những bản Nv gốc , có liên quan chặt chẽ đến hình ảnh.

3) Tìm các khái niệm chính. Có được khái niệm chính nv cụm {cn} $_{n=1}^{nv}$ trong đó mỗi cụm có thể chứa nhiều các khái niệm tương tự, chúng ta cần xác định một khái niệm để đại diện cho cụm này. Đối với mục tiêu này, chúng ta sử dụng CLIP để chọn một khái niệm từ một cụm bằng cách tìm ra sự tương đồng tối đa giữa hình ảnh và khái niệm

(4)

nơi c $_{N}^{chìa khóa}$ là khái niệm được chọn cho cụm cn.

Sau ba bước này, chúng ta có tập hợp các khái niệm chính như sau { phút c} $_{n=1}^{nv}$ có liên quan rất nhiều đến thị giác. Trước khi sử dụng những các khái niệm để tạo chú thích bằng các từ khóa sau thành câu LM, chúng ta cần quyết định thứ tự của chúng, đó là được thực hiện bởi các mối quan hệ trong bộ ba chủ ngữ-vị ngữ-tân ngữ.

3.2. Từ khóa thành câu LM

Để tạo chú thích liên quan đến hình ảnh trôi chảy bắt đầu từ các khái niệm chính {cN key} $_{n=1,}^{nv}$ chúng tôi sử dụng một mô hình ngôn ngữ bị ràng buộc về mặt từ vựng được đào tạo trước, CBART [15]. Cụ thể, CBART được phát triển để tạo ra một câu S = (x1, ..., xn) cho K từ khóa được sắp xếp {ci} $_{tôi=1}^{K}$ bởi max- mô phỏng xác suất có điều kiện.

(5)

trong đó x1, ...,xn là các từ. Để đạt được mục đích này, CBART có một bộ mã hóa hành động và một bộ giải mã ngôn ngữ để tinh chỉnh lặp đi lặp lại câu bắt đầu từ các từ khóa. Tại lần lặp t, bộ mã hóa có trách nhiệm dự đoán hành động nào ở cấp độ từ (sao chép, thay thế và chèn) nên được thực hiện. Trong nói cách khác, bộ mã hóa lấy một câu không đầy đủ St có n ′ các từ làm đầu vào và đầu ra tương ứng ac-, lt,n′}, trong trình tự tion Lt = {lt,1, · · · đó lt,i biểu thị hành động của từ thứ i ở lần lặp thứ t.

i) Sao chép. Sao chép có nghĩa là từ hiện tại vẫn không thay đổi.

ii) Thay thế. Thay thế gợi ý từ hiện tại nên được thay thế. Cụ thể, CBART sử dụng mặt nạ to-ken < M > để thay thế từ hiện tại và lấy mẫu từ mới dựa trên xác suất có điều kiện p lm(x<M>|x <M>), trong đó x <M> biểu thị các mã thông báo không được che dấu.

iii) Chèn. Chèn chỉ ra bộ giải mã nên chèn một từ trước từ hiện tại. Tương tự như hành động thay thế, CBART chèn một mã thông báo < M > trước từ hiện tại và sau đó lấy mẫu một từ từ plm (x<M>|x <M>).

Theo đó, bộ giải mã có thể tinh chỉnh câu từ St. đến St+1. Do đó, câu mã hóa-giải mã hoàn chỉnh sự tinh chỉnh của CBART ở lần lặp thứ t có thể được xây dựng như sau

Sau một vài lần lặp lại, CBART sẽ chấm dứt quá trình tinh chỉnh khi bộ mã hóa đưa ra chuỗi hành động sao chép đầy đủ.

Theo phần giới thiệu ở trên, CBART hiện tại không đáp ứng nhu cầu của chúng tôi vì đối với các hành động thay thế và chèn, từ chỉ được rút ra từ xác suất theo LM p lm(x<M>|x <M> được đào tạo trước), điều này chỉ đảm bảo sự trôi chảy nhưng không xem xét mối quan hệ giữa hình ảnh và văn bản.

3.3. Điểm số hợp nhất liên quan đến thị giác tăng cường trí nhớ

Để làm cho các chú thích có liên quan cao đến hình ảnh đã cho Tôi, chúng ta cần một hướng dẫn trực quan để tạo ra các từ trong hành động chèn và thay thế. Được thúc đẩy bởi điểm số tương phản CLIP được sử dụng rộng rãi để đánh giá sự tương đồng giữa hình ảnh và văn bản, chúng tôi phát triển một bộ nhớ tăng cường điểm số hợp nhất liên quan đến thị giác để điều chỉnh phân phối dự đoán từ gốc của CBART để liên kết với hình ảnh đã cho, xem xét cả i) sự tương đồng giữa hình ảnh và văn bản giữa các phương thức và ii) sự tương đồng trong phương thức giữa văn bản-văn bản.

Cụ thể, khi lấy mẫu[1] một từ xi ở vị trí i, CBART đầu tiên dự đoán một xác suất có điều kiện $p^{\text{làm sao}}$ Và chọn từ ứng viên top-Kw {xik} $_{k=1}^{Kw}$ với điểm trôi chảy tương ứng như sau:

(7)

Sau đó ứng cử viên Kw {sk Kw câu = (x1, ..., xik, ..., xn)} $_{k=1}$ được hình thành bằng cách kết hợp từ ứng viên xik với ngữ cảnh x i .

i) sự tương đồng giữa hình ảnh và văn bản. Sự tương đồng này là được biểu thị là $p^{NÓ}$ s, có thể được tính toán bằng cách lấy ứng cử viên câu ngày tháng {sk} $_{k=1}^{Kw}$ và hình ảnh I làm đầu vào để tính toán sự tương đồng giữa các phương thức CLIP như

(8)

i) sự tương đồng trong phương thức văn bản-văn bản. Lưu ý rằng đã lấy được bộ nhớ {mn} $_{n=1}^{Nd}$ trong Eq. (1) cũng liên quan đến hình ảnh. Do đó, chúng tôi giới thiệu một mô phỏng liên quan đến hình ảnh được tăng cường trí ilarity như p^{TT} là nhớ để cải thiện hơn nữa sự tương quan giữa hình ảnh và chú thích lation bằng cách sử dụng bộ mã hóa văn bản Sentence-BERT St để đánh giá sự giống nhau giữa {sk} $_{k=1}^{Kw}$ và {mn} $_{n=1}^{Nd}$ BẰNG

(9)

Cuối cùng, sau tổng có trọng số của Phương trình (7), Phương trình (8) và Phương trình (9), chúng ta có điểm số hợp nhất liên quan đến thị giác được tăng cường trí nhớ là

(10)

Kết quả là, khi lấy mẫu từ thứ i để thay thế hoặc chèn vào CBART cho mô hình của chúng tôi, chúng tôi chọn ứng viên từ có điểm hợp nhất cao nhất là

(11)

Cho đến nay, MeaCap mà chúng tôi đề xuất có thể đạt được IC không cần đào tạo với ít ảo giác hơn, được gọi là MeaCapTF trong các thí nghiệm.

Hơn nữa, giống như hầu hết các IC zero-shot đào tạo chỉ có văn bản các mô hình [51, 55] chỉ sử dụng văn bản để tinh chỉnh mô hình ngôn ngữ, trước tiên chúng ta cũng có thể tinh chỉnh CBART và sau đó thực hiện IC zero-shot chỉ có văn bản, được gọi là MeaCapToT trong các thí nghiệm.

4. Thí nghiệm

Để chứng minh rằng MeaCap có thể đạt được hiệu suất ấn tượng một cách hiệu quả trong các cài đặt zero-shot khác nhau, chúng tôi thực hiện theo các công trình trước đây [12, 29] để tiến hành toàn diện các thí nghiệm về Nhiệm vụ Một: IC không bắn trong Mục 4.1 và Nhiệm vụ hai: IC không ghép nối trong Mục 4.2. Đối với mỗi cài đặt, chúng tôi báo cáo cả hai kết quả của phiên bản không đào tạo MeaCapTF và phiên bản đào tạo chỉ có văn bản MeaCapToT. Trong Phần 4.3, chúng tôi đánh giá thêm tính hợp lệ của bộ nhớ dựa trên đề xuất của chúng tôi khung IC zero-shot với LM khác. Trong Phần 4.4, chúng tôi tiến hành các nghiên cứu cắt bỏ chi tiết cho MeaCap.

Bộ dữ liệu. Chúng tôi tiến hành thí nghiệm trên ba sử dụng các tiêu chuẩn chú thích hình ảnh, tức là MSCOCO [32], Flickr30K [63] và NoCaps [1]. Đối với MSCOCO và Bộ dữ liệu Flickr30K, chúng tôi theo dõi các công trình trước đây [7, 11, 12, 29] và sử dụng Karpathy split [19]. Chúng tôi sử dụng bộ xác thực của NoCaps để đánh giá khả năng chuyển giao của các mô hình IC được đào tạo trên các tập dữ liệu khác. Bên cạnh đó, đối với Nhiệm vụ Một, chúng tôi tuân theo các công trình trước đó [29] chuyển mô hình từ quy mô web corpus CC3M [50] cho MSCOCO và NoCaps. CC3M chứa ba triệu cặp hình ảnh-mô tả được thu thập từ web và chúng tôi chỉ sử dụng văn bản để xây dựng bộ nhớ hoặc tinh chỉnh LM.

Chi tiết triển khai. Có nhiều chương trình được đào tạo trước các mô-đun được sử dụng trong MeaCap. i) CLIP: chúng tôi sử dụng CLIP VIT-B/32 được đào tạo trước. ii) Sentence-BERT: chúng tôi sử dụng mô hình được đào tạo trước từ HuggingFace[2] . iii) CBART: chúng tôi sử dụng mô hình được đào tạo trước trên ngữ liệu Một tỷ từ[3] . iv) TextGraphParser: chúng tôi sử dụng văn bản có sẵn trình trích xuất đồ thị cảnh [31]. Đối với phiên bản không cần đào tạo MeaCapTF, chúng tôi nối một tiền tố "Hình ảnh trên mô tả điều đó" ở vị trí bắt đầu của câu. Đối với phiên bản đào tạo chỉ văn bản MeaCapToT, chúng tôi tiếp tục tinh chỉnh CBART trên tập dữ liệu đào tạo tương ứng với AdamW [24] trình tối ưu hóa. Đối với Nhiệm vụ Một, chúng tôi sử dụng CC3M để đóng vai trò là bộ nhớ, trong khi đối với Nhiệm vụ Hai, chúng tôi sử dụng ngữ liệu đào tạo của tập dữ liệu nguồn làm bộ nhớ. Nhiều thí nghiệm hơn với bộ nhớ văn bản khác được trình bày trong Phụ lục C. Chúng tôi đặt ngưỡng tương đồng khái niệm τ = 0,55

[1] Bất kể thay thế hay chèn vào thì bản chất vẫn như nhau, tức là, lấy mẫu một từ để thay thế mặt nạ.

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[3] https://www.statmt.org/lm-benchmark/

Phương pháp	Văn bản Corpus	MSCOCO					NoCap giá trị (CIDEr)			
	Rèn luyện trí nhớ B@4 MC		S	CLIP-S	BLIP2-S	Trong	Gần	Ra	Tổng	Thể
Không có nắp [54]		2.6	11,5	14,6	5,5	0,87	0,70	13,3	14,9	19,7	16,6
Tewel và cộng sự [53]		2.2	12,7	17,2	7,3	0,74	0,68	13,7	15,8	18,3	16,9
ConZIC [64]		1.3	11.2	13.3	5.0	1,00	0,76	15,4	16,0	20,3	17,5
CLIPRe [29]	CC3M 4.6		13.3	25.6	9.2	0,84	0,70	23,3	26,8	36,5	28.2
DeCap [29]	CC3M CC3M 8.8		16.0	42.1	10.9	0,76	-	34,8	37,7	49,9	39,7
MeaCapTF	CC3M 7.1		16,6	42,5	11,8	0,84	0,81	35,3	39,0	45,1	40,2
MeaCapToT	CC3M CC3M 9.0		17,8	48,3	12,7	0,79	0,75	38,5	43,6	50,0	45,1

Bảng 1. Kết quả chú thích Zero-shot trên MSCOCO Karpathy-test split và NoCaps được thiết lập. In, Near và Out biểu thị trong miền, gần miền và ngoài miền. MeaCapTF là phiên bản không đào tạo và MeaCapToT là phiên bản đào tạo chỉ có văn bản.

Phương pháp	MSCOCO				Flickr30K			
	B@4	MC	S	B@4	MC	S	
	Đào tạo về cặp hình ảnh-văn bản						
Từ dưới lên [3]	36,2	27,0	113,5	20,3	27,3	21,7	56,6	16,0
OSCAR [30]	36,5	30,3	123,7	23,1	40,9	-	-	-
VinVL [65]	30,9	140,6	25,1	33,5	27,5	-	-	-
ClipCap [38]	113,1	21,1	37,0	27,9	119,7	-	-	-
Vốn hóa nhỏ [46]	21,3	34,8	28,3	116,7	21,8	-	-	-
Điều chỉnh I [36]	25,2	22,8	61,5	16	.9			
	Đào tạo chỉ văn bản, suy luận không có cú đánh nào							
ZeroCap† [54]	15,4	17,8	7,9	11,8	16,8	6,2	9.2	5.4
PHÉP THUẬT [51]	12.9	17.4	49.3	13.1	20.4	7.1	11.3	6.4
KHÔNG GÂY RA [55]	15,5	18,7	55,4	12,1	13,1	15,2	26,4	8,3
CLIPRe [29]	12,4	20,4	53,4	14,8	18,2	31,7	12,0	9.8
MeaCapTF	9,1	20,6	56,9	17,8	36,5	19,5	7.2	
MeaCapToT	17,7	24,3	84,8	18,7	15,3	20,6	50,2	14,5

Bảng 2. Kết quả chú thích trong miền trên phân tách thử nghiệm Karpathy của MSCOCO và phân tách thử nghiệm Karpathy của Flickr30K. † có nghĩa là phiên bản chỉ có văn bản được triển khai lại từ [51].

đối với bộ nhớ CC3M và τ = 0,6 đối với các bộ nhớ khác.
Nd, Kw, α, β, γ được đặt là 5, 200, 0,1, 0,4, 0,2 trong số tất cả thí nghiệm. Tất cả các thí nghiệm được tiến hành trên một GPU RTX3090. Chúng tôi xử lý trước văn bản thành văn bản nhúng bằng bộ mã hóa văn bản CLIP và lưu trữ nhúng văn bản làm bộ nhớ của chúng tôi để truy xuất nhanh. Ví dụ, truy xuất lại trên CC3M tốn trung bình 0,05 giây trên RTX3090 GPU hoặc trung bình 1 giây trên CPU. Phân tích chi phí tính toán chi tiết hơn được trình bày trong Phụ lục E.

Số liệu. Để đánh giá độ chính xác của dữ liệu được tạo ra chú thích, chúng tôi sử dụng số liệu giám sát truyền thống BLEU (B@n) [42], METEOR (M) [4], CIDEr (C) [57], và SPICE (S) [2] tính toán sự giống nhau giữa các câu ứng viên và các tham chiếu của con người. Đối với các phương pháp không cần đào tạo, chúng tôi sử dụng CLIPScore (CLIP-S) [16] để đo độ tương đồng giữa hình ảnh và văn bản. Ngoài ra, xét đến việc CLIP-S không nhạy cảm với ảo giác của các phương pháp dựa trên CLIP đó như thể hiện trong Hình 1b, chúng tôi sử dụng một lớn được đào tạo trước khác mô hình BLIP-2 [28] để đánh giá độ tương đồng giữa hình ảnh và văn bản, tức là BLIP2Score (BLIP2-S). Chi tiết hơn có trong Phụ lục D.

Phương pháp	MSCOCO				Flickr30k				Flickr30k				MSCOCO			
	B@4	MC	S		B@4	MC	S									
PHÉP THUẬT [51]	6.2		12.2	17.5	5.9				5.2		12,5	18,3	5,7			
CLIPRe [29]	9.8		16.7	30.1	10.3				6.0		16.0	26.5	10.2			
MeaCapTF			16,6	34,4	11,4	7,4					16,2	46,4	11,2			
MeaCapToT	13,4	18,5	40,3	12,1					9,8	17,4	51,7	12,0				

Bảng 3. Kết quả chú thích miền chéo trên MSCOCO và Flickr30K Phân chia thử nghiệm Karpathy.

4.1. Chú thích hình ảnh Zero-shot

Trong phần này, chúng tôi tiến hành các thí nghiệm IC không bắn để đánh giá khả năng của các mô hình chuyển từ một tổng thể ngữ liệu được thu thập trên web cho các tập dữ liệu IC hạ lưu khác nhau.

Đường cơ sở. Trong nghiên cứu này, chúng tôi so sánh hai loại đường cơ sở. i) Các phương pháp không cần đào tạo: ZeroCap [54], Tewel et al. [53] và ConZIC [64]. Các phương pháp đó tận dụng CLIP được đào tạo trước và LM đông lạnh (BERT hoặc GPT-2) để đạt được IC không bắn. ii) Phương pháp đào tạo chỉ bằng văn bản4 : DeCap [29], đây cũng là phương pháp dựa trên trí nhớ được thảo luận ở Mục 2.3. Thay vì sử dụng LM được đào tạo trước, DeCap đào tạo một ngôn ngữ bộ giải mã từ đầu. Bên cạnh đó, tác giả của Decap thiết lập một đường cơ sở được gọi là CLIPRe, tạo ra các mô tả hình ảnh bằng cách lấy các văn bản có liên quan nhất từ bộ nhớ trực tiếp. Tiếp theo Decap, đối với MeaCapTF, chúng ta chỉ sử dụng CC3M là bộ nhớ và đối với MeaCapToT, chúng tôi sử dụng CC3M như bộ nhớ và cũng điều chỉnh CBART. Tab. 1 hiển thị kết quả trên MSCOCO và NoCaps, và MeaCap đạt được kết quả mới nhất. Kết quả không cần đào tạo. Cụ thể, phiên bản không cần đào tạo MeaCapTF của chúng tôi đã cho thấy hiệu suất vượt trội trên số liệu tham chiếu (B@4, M, C, S) hơn tất cả các đường cơ sở không cần đào tạo trước đó, ZeroCap, Tewel et al. và ConZIC trên cả MSCOCO và NoCaps bộ dữ liệu với biên độ lớn, chứng minh tính hiệu quả của thiết kế tăng cường bộ nhớ của chúng tôi. Đối với các số liệu không tham chiếu (CLIP-S và BLIP2-S), MeaCapTF đạt được kết quả tốt hơn trên BLIP2-S và kém hơn trên CLIP-S. Như đã thảo luận trong phần giới thiệu, các phương pháp không cần đào tạo trước đây được CLIP-S ưa chuộng vì hiện tượng ảo giác. Bên cạnh đó, MeaCapTF của chúng tôi cũng vượt trội hơn dựa trên truy xuất

4Các phương pháp đào tạo chỉ có văn bản khác ngoại trừ DeCap không có kinh nghiệm được đề cập trong bối cảnh này, chúng tôi đã so sánh chúng trong Nhiệm vụ Hai (Phần 4.2)

Hình 3. Ví dụ về IC zero-shot so với các đường cơ sở zero-shot khác. GT biểu thị Ground Truth. ConZIC và ZeroCap là không cần đào tạo, trong khi MAGIC và DeCap chỉ đào tạo văn bản. MeaCap hiển thị các khái niệm đã trích xuất bằng màu xanh lá cây và chú thích được tạo.

Hình 4. Ví dụ về kiến thức thực tế. MeaCapToT có thể giảm bớt vấn đề kiến thức thế giới bị lãng quên của các phương pháp đào tạo chỉ bằng văn bản hiện có, chẳng hạn như "batman" trong hình ảnh thứ ba.

đường cơ sở CLIPRe với biên độ lớn, chỉ ra rằng chỉ việc lấy lại chú thích có liên quan nhất là thiếu chính xác. Hơn nữa, thậm chí so với đào tạo chỉ có văn bản phương pháp Decap, MeaCapTF cho thấy vượt trội hoặc tương đương hiệu suất trên cả MSCOCO và NoCap.

Kết quả đào tạo chỉ văn bản. Để khám phá tiềm năng của MeaCap của chúng tôi với việc đào tạo chỉ văn bản thêm trên kho dữ liệu quy mô web sau DeCap, chúng tôi cũng tinh chỉnh CBART trên tập hợp CC3M, tức là MeaCapToT. Có thể quan sát thấy rằng MeaCapToT cải thiện đáng kể hiệu suất, đặc biệt là trên NoCap. Cụ thể, dưới cùng một quá trình đào tạo và tình trạng bộ nhớ, MeaCapToT vượt trội hơn DeCap ở cả hai tập dữ liệu MSCOCO và tập dữ liệu NoCaps, cho thấy tính ưu việt của phương pháp sử dụng bộ nhớ ngoài của chúng tôi.

Kết quả định tính. Bên cạnh việc so sánh định lượng, chúng tôi hình dung các chú thích được tạo ra trong Hình 1, 3 và 4. Rõ ràng, MeaCap có thể tạo ra phụ đề tốt hơn với nhiều kiến thức hơn và ít ảo giác hơn. Thêm nhiều kết quả có trong Phụ lục F.

4.2. Nhiệm vụ thứ hai: Chú thích hình ảnh không ghép đôi

4.2.1 Chú thích trong miền

Để khám phá thêm tiềm năng của MeaCap cho việc thiết lập trong miền, nơi dữ liệu đào tạo, bộ nhớ và bộ kiểm tra được từ cùng một tập dữ liệu, nhưng không sử dụng cặp hình ảnh-văn bản để xây dựng mô hình và bộ nhớ.

Đường cơ sở. Trong nghiên cứu này, chúng tôi so sánh với các phương pháp đào tạo chỉ có văn bản khác là ZeroCap† [54], MAGIC [51] và ZEROGEN [55] và phương pháp tiếp cận dựa trên truy xuất CLIPRe. ZeroCap là một phương pháp không cần đào tạo được mở rộng thành phiên bản đào tạo chỉ có văn bản ZeroCap† [54]. Các phương pháp đó đóng băng CLIP và tinh chỉnh LM trên đào tạo tương ứng văn bản. Theo thiết lập trong miền, chúng tôi cũng báo cáo cả phiên bản không cần đào tạo MeaCapTF, chỉ sử dụng văn bản đào tạo như bộ nhớ, và phiên bản đào tạo chỉ có văn bản MeaCapToT sử dụng văn bản đào tạo để tinh chỉnh CBART cũng có chức năng như bộ nhớ.

Kết quả. Như thể hiện trong Tab. 2, MeaCapTF vượt trội hơn CLIPRe và các cơ sở đào tạo chỉ có văn bản khác về C và S điểm. So với điểm B@4 và M, điểm C và S điểm số chú ý nhiều hơn đến độ chính xác của các thực thể và mối quan hệ. Hiệu suất vượt trội trên hai điểm số này chứng minh chất lượng cao của phương pháp truy xuất sau đó lọc dựa trên bộ nhớ mà chúng tôi đề xuất để có được các khái niệm chính. Hơn nữa, MeaCapToT vượt trội hơn tất cả các đường cơ sở một cách đáng kể biên độ, cho thấy phương pháp chúng tôi đề xuất có tiềm năng lớn hơn với quá trình đào tạo trong miền sâu hơn.

4.2.2 Chú thích liên miền

Chúng tôi đánh giá MeaCap cho IC đa miền với dữ liệu đào tạo và thử nghiệm từ các tập dữ liệu khác nhau. Chúng tôi sử dụng văn bản từ bộ đào tạo làm bộ nhớ cho MeaCapTF và MeaCapToT và tinh chỉnh CBART cho MeaCapToT.

Phương pháp	MSCOCO				Flickr30K			
	B@4	M	C	S	B@4	M	C	S
DeCap [29]	24,7	25,0	91,2	18,7	21,2	21,8	56,7	15,2
CapDec [40]	26,4	25,1	91,8	-	ViECap		17,7	20,0	39,1	-
[12]	27,2	24,8	92,9	18,2	21,4	20,1	47,9	13,6				
MeaCapInvLM	27,2	25,3	95,4	19,0	22,3	22,3	59,4	15,6				
	MSCOCO		Flickr30K				Flickr30K		MSCOCO	
DeCap [29]	16,3	17,9	35,7	11,1	12,1	CapDec [40]			18.0	44.4	10.9
17,3	18,6	35,7	17,4	18,0	38,4	11,2	-	9.2	16.3	27.3	-
ViECap [12]	12,6	19,3	54,2	12,5							
MeaCapInvLM	18,5	19,5	43,9	12,8	13,1	19,7	56,4	13,2			

Bảng 4. Kết quả chú thích trong miền và xuyên miền với Bộ giải mã ngôn ngữ đảo ngược CLIP.

Phương pháp	ReF IT TTs	MSCOCO B@4 M C S
MeaCapTF	5,0 1,8 5,7 7,1	13.3 31.1 5.6 9,7 12,7 4,8 13,6 38,6 8,5 16,6 42,5 11,8
MeaCapToT	7,9 3,2 8,1 9,0	14,9 37,1 10,4 9,9 17,3 5,2 15,6 44,7 11,1 17,8 48,3 12,7

Bảng 5. Nghiên cứu cắt bỏ trên IC không bắn. ReF, IT, TT biểu thị mô-đun lấy và lọc, IT (8) và TT (1) là hình ảnh-văn bản và độ tương đồng giữa văn bản và văn bản từ điểm số liên quan đến hình ảnh được tăng cường trí nhớ.

Kết quả. Chúng tôi so sánh MeaCap với MAGIC cơ sở đào tạo chỉ có văn bản (điều chỉnh GPT-2) và CLIPRe. cơ sở truy xuất dựa trên. Kết quả trong Tab. 3 hiển thị MAGIC bị suy giảm hiệu suất trên dữ liệu mục tiêu, thậm chí còn tệ hơn phương pháp dựa trên truy xuất CLIPRe. Được trang bị thiết kế tăng cường bộ nhớ được đề xuất, MeaCapTF vượt trội hơn CLIPRe về hầu hết các số liệu và MeaCapToT vượt trội hơn tất cả các đường cơ sở, chứng minh hiệu quả của thiết kế tăng cường trí nhớ được đề xuất.

4.3. Tính linh hoạt của MeaCap với các LM khác

Cơ chế bộ nhớ được đề xuất của chúng tôi để tìm các khái niệm chính trong Mục 3.1 là một mô-đun cắm và chạy để cải thiện hơn nữa hầu hết các phương pháp SOTA đào tạo chỉ bằng văn bản hiện có [12, 29, 40]. Vì mục đích này, chúng ta chỉ cần thay thế CBART (Mục 3.2) trong MeaCap với một LM khác được sử dụng trong các phương pháp này (không cần điểm hợp nhất trong Mục 3.3) được mô tả như sau.

Đường cơ sở. DeCap [29], CapDec [40] và ViECap [12] đào tạo LM từ đầu để đảo ngược bộ mã hóa văn bản CLIP, được biểu thị là InvLM trong phần sau. Chúng chiếu hình ảnh nhúng được trích xuất bởi bộ mã hóa hình ảnh CLIP vào văn bản nhúng không gian của bộ mã hóa văn bản CLIP. Sau đó, họ sử dụng InvLM để tái tạo văn bản từ các văn bản nhúng. Để tạo ra các mô tả dựa trên các khái niệm chính đã trích xuất của chúng tôi, đầu tiên chúng ta sử dụng mẫu nhắc nhở như "Có [c1, c2, ..., cn] trong hình ảnh" để đưa các khái niệm vào một khái niệm nhận thức câu theo sau ViECap, trong đó cn là khái niệm thứ n. Sau khi mã hóa câu nhận biết khái niệm thành văn bản nhúng bằng bộ mã hóa văn bản CLIP, chúng ta sẽ nhận được lời nhắc nhận biết khái niệm. Chúng tôi kết hợp lời nhắc nhận biết khái niệm với các đoạn nhúng văn bản làm đầu vào của InvLM, được đặt tên là MeaCapInvLM.

Kết quả. Tab. 4 cho thấy MeaCapInvLM vượt trội hơn tất cả các đường cơ sở trên tất cả các số liệu trong các kịch bản trong miền và liên miền, chứng minh tính hiệu quả của chúng tôi đề xuất các khái niệm chính dựa trên bộ nhớ và cũng chỉ ra tính linh hoạt của nó đối với nhiều LM và các thiết lập zero-shot khác nhau, với phân tích chi tiết trong Phụ lục A.

4.4. Nghiên cứu cắt bỏ

Để khám phá tác động của từng mô-đun chính trong MeaCap, tức là mô-đun lấy-rồi-lọc (ReF), hình ảnh-văn bản điểm tương đồng (ITs) và điểm tương đồng văn bản-văn bản (TTs), chúng tôi tiến hành các nghiên cứu cắt bỏ toàn diện trên Bộ dữ liệu MSCOCO dựa trên Nhiệm vụ Một của việc thiết lập không có cú đánh. Chúng tôi đánh giá cả phiên bản không đào tạo MeaCapTF và phiên bản đào tạo chỉ có văn bản MeaCapToT có kết quả được cung cấp trong Tab. 5. Như chúng ta có thể thấy, chỉ kết hợp với ReF và LM ban đầu (hàng đầu tiên) có thể vượt qua chỉ có kết quả của IT ở hàng thứ hai (IT là kết quả trực quan duy nhất hướng dẫn của các phương pháp đào tạo miễn phí trước đây của CLIP), chỉ ra các khái niệm chính được trích xuất bởi mô-đun ReF là quan trọng đối với IC zero-shot. Hàng thứ ba cho thấy việc kết hợp ReF với IT mang lại nhiều cải tiến hơn so với các mô-đun riêng lẻ. Cuối cùng, bằng cách kết hợp các TT, hiệu suất được cải thiện hơn nữa, làm nổi bật hiệu quả của điểm số hợp nhất liên quan đến thị giác được tăng cường trí nhớ. Chúng tôi tiến hành phân tích tác động của trí nhớ trong Phụ lục B.

5. Kết luận

Trong bài báo này, chúng tôi đề xuất một khuôn khổ IC zero-shot tăng cường bộ nhớ mới, MeaCap. Chúng tôi giới thiệu một mô-đun truy xuất rồi lọc để trích xuất các khái niệm chính từ văn bản bên ngoài bộ nhớ. Dựa trên bộ nhớ văn bản đã thu thập được, chúng tôi tiếp tục phát triển một điểm số hợp nhất liên quan đến thị giác được tăng cường trí nhớ để hướng dẫn việc tạo phụ đề. Kết hợp với CBART, chúng ta có thể tạo ra các mô tả tập trung vào khái niệm để giảm bớt ảo giác của các phương pháp không đào tạo trước đây và tăng cường độ chính xác của các phương pháp đào tạo chỉ có văn bản. Các thí nghiệm mở rộng trên nhiều cài đặt chú thích không có cảnh quay chứng minh rằng MeaCap vượt trội hơn các phương pháp trước đây.

6. Lời cảm ơn

H. Zhang ghi nhận sự hỗ trợ của NSFC (62301384), và Quỹ Nhà khoa học trẻ xuất sắc (ở nước ngoài). Z. Wang ghi nhận sự hỗ trợ của NSFC (62301407). B. Chen ghi nhận sự hỗ trợ của NSFC (U21B2006); Dự án Đội sáng tạo thanh niên Thiểm Tây; Khoản tài trợ cho dự án 111 (B18039); Quỹ nghiên cứu cơ bản cho Trung ương Đại học QTZX22160.

Tài liệu tham khảo

[1] Agrawal khắc nghiệt, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Ste-fan Lee và Peter Anderson. Nocaps: Chú thích đối tượng mới lạ ở quy mô lớn. Trong Biên bản báo cáo của IEEE/CVF quốc tế hội nghị về tầm nhìn máy tính, trang 8948–8957, 2019. 1, 5

[2] Peter Anderson, Basura Fernando, Mark Johnson và Stephen Gould. Spice: Đánh giá chú thích hình ảnh mệnh đề ngữ nghĩa. Trong hội nghị châu Âu về tầm nhìn máy tính, trang 382–398. Springer, 2016. 6

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould và Lei Zhang. Sự chú ý từ dưới lên và từ trên xuống cho chú thích hình ảnh và trả lời câu hỏi trực quan. Trong Biên bản báo cáo của hội nghị IEEE về tầm nhìn máy tính và nhận dạng mẫu, các trang 6077–6086, 2018. 2, 6

[4] Satanjeev Banerjee và Alon Lavie. Thiên thạch: Một thước đo để đánh giá mt với mối tương quan được cải thiện với các phán đoán của con người. Trong Biên bản hội thảo acl về các biện pháp đánh giá bên trong và bên ngoài để dịch máy và/hoặc tóm tắt, trang 65–72, 2005. 2, 6

[5] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Muller và Bj orn Ommer. Các mô hình khuếch tán tăng cường truy xuất. Những tiến bộ trong hệ thống xử lý thông tin thần kinh, 35:15309–15324, 2022. 3

[6] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Cải thiện các mô hình ngôn ngữ bằng cách lấy từ hàng nghìn tỷ mã thông báo. Trong hội nghị quốc tế về học máy, trang 2206–2240. PMLR, 2022. 3

[7] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, và Rita Cucchiara. Bộ biến đổi bộ nhớ dạng lưới để chú thích hình ảnh. Trong Biên bản báo cáo của hội nghị IEEE/CVF về tầm nhìn máy tính và nhận dạng mẫu, trang 10578–10587, 2020. 1, 2, 5

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee và Kristina Tổng cộng. Bert: Đào tạo trước về song hướng sâu máy biến áp để hiểu ngôn ngữ. bản in trước arXiv arXiv:1810.04805, 2018. 1

[9] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, và Trevor Darrell. Các mạng lưới tích chập hồi quy dài hạn để nhận dạng và mô tả trực quan. Trong Biên bản báo cáo của hội nghị IEEE về thị giác máy tính và mẫu nhận dạng, trang 2625–2634, 2015. 2

[10] Zhibin Duan, Lv Zhiyi, Chaojie Wang, Bo Chen, Bo An, và Mingyuan Zhou. Tạo ra ít ảnh thông qua việc nhớ lại trí nhớ ngữ nghĩa theo từng giai đoạn lấy cảm hứng từ não. Trong Hội nghị lần thứ ba mươi bảy về Hệ thống xử lý thông tin thần kinh, 2023. 3

[11] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang và Zichen Liu. tiêm chích các khái niệm ngữ nghĩa vào chú thích hình ảnh đầu cuối. Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Tầm nhìn máy tính và Nhận dạng Mẫu, trang 18009–18019, 2022. 5

[12] Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang và Feng Zheng. Giải mã có thể chuyển giao bằng hình ảnh các thực thể cho chú thích hình ảnh không có cảnh quay. Trong Biên bản Hội nghị quốc tế IEEE/CVF về tầm nhìn máy tính, trang 3136–3146, 2023. 2, 3, 5, 8

[13] Jiuxiang Gu, Gang Wang, Jianfei Cai và Tsuhan Chen. MỘT nghiên cứu thực nghiệm về ngôn ngữ cnn cho chú thích hình ảnh. Trong Biên bản báo cáo hội nghị quốc tế IEEE về tầm nhìn máy tính, trang 1222–1231, 2017. 2

[14] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, và Mingwei Chang. Đào tạo trước mô hình ngôn ngữ tăng cường truy xuất. Trong hội nghị quốc tế về học máy, trang 3929–3938. PMLR, 2020. 3

[15] Xingwei He. Cải tiến song song cho từ vựng bị hạn chế tạo văn bản với bart. bản in trước arXiv arXiv:2109.12487, 2021. 2, 3, 4

[16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, và Yejin Choi. Clipscore: Một thước đo đánh giá không tham chiếu cho chú thích hình ảnh. Bản in trước arXiv arXiv:2104.08718, 2021. 1, 2, 6

[17] Luân Hoàng, Wenmin Wang, Jie Chen, và Xiao-Yong Wei. Chú ý đến chú ý cho chú thích hình ảnh. Trong Biên bản của hội nghị quốc tế IEEE/CVF về tầm nhìn máy tính, trang 4634–4643, 2019. 1, 2

[18] Luân Hoàng, Văn Dân Vương, Yaxian Xia, và Jie Chen. Chú thích hình ảnh được căn chỉnh thích ứng thông qua sự chú ý thích ứng thời gian. Những tiến bộ trong hệ thống xử lý thông tin thần kinh, 32, 2019. 2

[19] Andrej Karpathy và Li Fei-Fei. Sự liên kết ngữ nghĩa thị giác sâu sắc để tạo ra các mô tả hình ảnh. Trong Biên bản hội nghị IEEE về tầm nhìn máy tính và nhận dạng mẫu, trang 3128–3137, 2015. 5

[20] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen và Wen-tau Yih. Truy xuất đoạn văn dày đặc cho câu hỏi miền mở trả lời. bản in trước arXiv arXiv:2004.04906, 2020. 3

[21] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettle-moyer và Mike Lewis. Tổng quát hóa thông qua ghi nhớ: Mô hình ngôn ngữ láng giềng gần nhất. Bản in trước arXiv arXiv:1911.00172, 2019. 3

[22] Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettle-moyer và Mike Lewis. Bản dịch máy láng giềng gần nhất. Bản in trước arXiv arXiv:2010.00710, 2020. 3

[23] Wonjae Kim, Bokyung Son và Ildoo Kim. Vilt: Bộ chuyển đổi tầm nhìn và ngôn ngữ không có tích chập hoặc giám sát vùng. Trong Hội nghị quốc tế về học máy, trang 5583–5594. PMLR, 2021. 1

[24] Diederik P Kingma và Jimmy Ba. Adam: Một phương pháp cho tối ưu hóa ngẫu nhiên. bản in trước arXiv arXiv:1412.6980, 2014. 5

[25] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalan-tidis, Li-Jia Li, David A Shamma, và những người khác. Bộ gen thị giác: Kết nối ngôn ngữ và tầm nhìn bằng cách sử dụng crowdsourced thick chú thích hình ảnh. Tạp chí quốc tế về thị giác máy tính, 123(1):32–73, 2017. 1

[26] Chia-Wen Kuo và Zsolt Kira. Ngoài một trình phát hiện đối tượng được đào tạo trước: Ngữ cảnh trực quan và văn bản liên phương thức cho chú thích hình ảnh. Trong Biên bản báo cáo Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu, trang 17969–17979, 2022. 2 [27] Kenton Lee, Ming-Wei Chang và Kristina

Toutanova. Truy xuất trễ để trả lời câu hỏi miền mở được giám sát yếu. Bản in trước arXiv arXiv:1906.00300, 2019. 3 [28] Junnan Li, Dongxu Li, Silvio Savarese và Steven Hoi.

Blip-2: Khởi động quá trình đào tạo trước ngôn ngữ-hình ảnh với bộ mã hóa hình ảnh đông lạnh và các mô hình ngôn ngữ lớn. Bản in trước arXiv arXiv:2301.12597, 2023. 6 [29] Wei Li, Linchao

Zhu, Longyin Wen và Yi Yang. Decap: Giải mã các đoạn clip tiềm ẩn cho chú thích không có cảnh quay nào thông qua đào tạo chỉ có văn bản. Bản in trước arXiv arXiv:2303.03032, 2023. 2, 3, 5, 6, 8

[30] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Đào tạo trước theo ngữ nghĩa đối tượng cho các tác vụ ngôn ngữ thị giác. Trong Hội nghị Châu Âu về Thị giác máy tính, trang 121–137. Springer, 2020. 6 [31] Zhuang Li, Yuyang Chai, Terry Zhuo Yue, Lizhen Qu, Gholamreza Haffari,

Fei Li, Donghong Ji và Quan Hung Tran. Factual: Điểm chuẩn để phân tích cú pháp đồ thị cảnh văn bản trung thực và nhất quán. Bản in trước arXiv arXiv:2305.17497, 2023. 4, 5 [32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays,

Pietro Perona, Deva Ramanan, Piotr Dollar và C Lawrence Zitnick. Microsoft coco: Các

đối tượng chung trong ngữ cảnh. Trong hội nghị châu Âu về thị giác máy tính, trang 740–755.

Springer, 2014. 1, 5

[33] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong và Stella X Yu. Nhận dạng đuôi dài quy mô lớn trong thế giới mở. Trong Biên bản báo cáo hội nghị IEEE/CVF về thị giác máy tính và nhận dạng mẫu, trang 2537–2546, 2019. 3

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin và Baining Guo. Bộ biến đổi Swin: Bộ biến đổi thị giác phân cấp sử dụng cửa sổ dịch chuyển. Trong Kỷ yếu Hội nghị quốc tế IEEE/CVF về thị giác máy tính, trang 10012–10022, 2021. 1 [35] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi

Garg, Alan Blair, Chunhua Shen và Anton van den Hengel. Phân loại tăng cường truy xuất để nhận dạng hình ảnh đuôi dài. Trong Kỷ yếu Hội nghị IEEE/CVF về thị giác máy tính và nhận dạng mẫu, trang 6959–6969, 2022. 3 [36] Ziyang Luo, Zhipeng Hu, Yadong Xi, Rongsheng Zhang và Jing Ma. I-tuning: Điều chỉnh các mô hình ngôn ngữ đông lạnh với hình ảnh để thêm chú thích cho hình ảnh nhẹ. Trong ICASSP 2023- 2023 IEEE

International Conference on Acoustics, Speech and Signal Processing (ICASSP), trang 1–5. IEEE, 2023. 6 [37] Yuxian Meng, Shi Zong, Xiaoya Li, Xiaofei Sun, Tianwei Zhang, Fei Wu và Jiwei Li. Gnn-lm: Mô hình hóa ngôn ngữ dựa trên ngữ cảnh toàn cầu thông qua gnn. arXiv preprint arXiv:2110.08743, 2021. 3

[38] Ron Mokady, Amir Hertz và Amit H Bermano. Clip-cap: Tiền tố clip để chú thích hình ảnh. Bản in trước arXiv arXiv:2111.09734, 2021. 6 [39] Van-Quang Nguyen, Masanori Suganuma

và Takayuki Okatani. Grit: Bộ chuyển đổi chú thích hình ảnh nhanh hơn và tốt hơn bằng cách sử dụng các tính năng trực quan kép. Trong Hội nghị Châu Âu về Thị giác Máy tính, trang 167–184. Springer, 2022. 1, 2 [40] David Nukrai, Ron Mokady và Amir Globerson. Đào tạo chỉ văn bản để chú

thích hình ảnh bằng cách sử dụng clip có chèn nhiễu. Bản in trước arXiv arXiv:2211.00575, 2022. 2, 3, 8

[41] Yingwei Pan, Ting Yao, Yehao Li, và Tao Mei. Mạng chú ý tuyến tính X cho chú thích hình ảnh. Trong Biên bản báo cáo của hội nghị IEEE/CVF về thị giác máy tính và nhận dạng mẫu, trang 10971–10980, 2020. 1, 2 [42] Kishore Papineni, Salim Roukos, Todd Ward, và Wei-

Jing Zhu. Bleu: một phương pháp đánh giá tự động bản dịch máy. Trong Biên bản báo cáo của cuộc họp thường niên lần thứ 40 của Hiệp hội Ngôn ngữ học tính toán, trang 311–318, 2002. 2, 6

[43] Yu Qin, Jiajun Du, Yonghua Zhang và Hongtao Lu. Nhìn lại và dự đoán về phía trước trong chú thích hình ảnh. Trong Biên bản báo cáo của hội nghị IEEE/CVF về thị giác máy tính và nhận dạng mẫu, trang 8367–8375, 2019. 2

[44] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Các mô hình ngôn ngữ là những người học đa nhiệm không được giám sát. Blog OpenAI, 1(8):9, 2019. 1

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Học các mô hình trực quan có thể chuyển giao từ tầm nhìn siêu ngôn ngữ tự nhiên. Trong Hội nghị quốc tế về học máy, trang 8748–8763. PMLR, 2021. 2

[46] Rita Ramos, Bruno Martins, Desmond Elliott và Yova Ke-mentchedjhieva. Smallcap: chú thích hình ảnh nhẹ được nhắc đến với sự tăng cường truy xuất. Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu, trang 2840–2849, 2023. 3, 6 [47] Nils Reimers và Iryna Gurevych. Sentence-bert: Nhúng câu bằng mạng

siamese bert. Trong Biên bản báo cáo của Hội nghị năm 2019 về Phương pháp thực nghiệm trong Xử lý ngôn ngữ tự nhiên. Hiệp hội Ngôn ngữ học tính toán, 2019. 4

[48] Idan Schwartz, Alexander Schwing và Tamir Hazan. Các mô hình chú ý bậc cao để trả lời câu hỏi trực quan. Những tiến bộ trong Hệ thống xử lý thông tin thần kinh, 30, 2017. 1, 2

[49] Idan Schwartz, Alexander G Schwing và Tamir Hazan. Một đường cơ sở đơn giản cho đối thoại nhận biết cảnh bằng âm thanh-hình ảnh. Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu, trang 12548–12558, 2019. 1, 2

[50] Piyush Sharma, Nan Ding, Sebastian Goodman và Radu Soricut. Chú thích khái niệm: Một tập dữ liệu văn bản thay thế hình ảnh được làm sạch, có siêu ẩn danh, để chú thích hình ảnh tự động. Trong Biên bản báo cáo của Hội nghị thường niên lần thứ 56 của Hiệp hội Ngôn ngữ học tính toán (Tập 1: Bài báo dài), trang 2556–2565, 2018. 5

[51] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yo-gatama, Yan Wang, Lingpeng Kong và Nigel Collier. Các mô hình ngôn ngữ có thể thấy: Cắm các điều khiển trực quan vào quá trình tạo văn bản. bản in trước arXiv arXiv:2205.02655, 2022. 2, 3, 5, 6, 7

[52] Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong và Nigel Collier. Một khuôn khổ tương phản cho thần kinh tạo văn bản. Tiến bộ trong xử lý thông tin thần kinh Hệ thống, 35:21548–21561, 2022. 2

[53] Yoad Tewel, Yoav Shalev, Roy Nadler, Idan Schwartz và Lior Wolf. Chú thích video Zero-shot với các mã thông báo giả đang phát triển. Bản in trước arXiv arXiv:2207.11100, 2022. 1, 2, 3, 6

[54] Yoad Tewel, Yoav Shalev, Idan Schwartz và Lior Wolf. Zerocap: Tạo ảnh thành văn bản Zero-shot cho số học ngữ nghĩa thị giác. Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu, các trang 17918–17928, 2022. 1, 2, 3, 6, 7

[55] Haoqin Tu, Bowen Yang, và Xianfeng Zhao. Không tạo ra: Tạo văn bản đa phương thức có thể điều khiển bằng Zero-shot với nhiều oracle. Bản in trước arXiv arXiv:2306.16649, 2023. 2, 3, 5, 6, 7

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko-reit, Llion Jones, Aidan N Gomez, Łukasz Kaiser và Illia Polosukhin. Sự chú ý là tất cả những gì bạn cần. Những tiến bộ trong thần kinh hệ thống xử lý thông tin, 30, 2017. 1

[57] Ramakrishna Vedantam, C Lawrence Zitnick và Devi Parikh. Cider: Đánh giá mô tả hình ảnh dựa trên sự đồng thuận. Trong Biên bản báo cáo của hội nghị IEEE về máy tính tầm nhìn và nhận dạng mẫu, trang 4566–4575, 2015. 2, 6

[58] Oriol Vinyals, Alexander Toshev, Samy Bengio và Du-mitru Erhan. Hiển thị và kể: Một máy tạo chú thích hình ảnh thần kinh. Trong Biên bản hội nghị IEEE về máy tính tầm nhìn và nhận dạng mẫu, trang 3156–3164, 2015. 2

[59] Li Wang, Zechen Bai, Yonghua Zhang, và Hongtao Lu. Hiển thị, nhớ lại và kể lại: Chú thích hình ảnh với cơ chế nhớ lại. Trong Biên bản báo cáo của hội nghị AAAI về nhân tạo trí thông minh, trang 12176–12183, 2020. 2

[60] Kelvin Xu, Jimmy Ba, Ryan Kiros, Cho Kyunghyun, Aaron Courville, Ruslan Salakhudinov, Rich Zemel và Yoshua Bengio. Hiển thị, tham dự và kể: Tạo chú thích hình ảnh thần kinh với sự chú ý trực quan. Trong hội nghị quốc tế về học máy, trang 2048–2057. PMLR, 2015. 2

[61] Xu Yang, Kaihua Tang, Hanwang Zhang, và Jianfei Cai. Tự động mã hóa đồ thị cảnh để chú thích hình ảnh. Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Tầm nhìn máy tính và Nhận dạng Mẫu, trang 10685–10694, 2019. 2

[62] Ting Yao, Yingwei Pan, Yehao Li và Tao Mei. Khám phá mối quan hệ trực quan cho chú thích hình ảnh. Trong Biên bản Hội nghị châu Âu về tầm nhìn máy tính (ECCV), các trang 684–699, 2018. 2

[63] Peter Young, Alice Lai, Micah Hodosh và Julia Hocken-maier. Từ mô tả hình ảnh đến biểu thị trực quan: Mới số liệu tương tự cho suy luận ngữ nghĩa trên các mô tả sự kiện. Giao dịch của Hiệp hội tính toán Ngôn ngữ học, 2:67–78, 2014. 1, 5

[64] Zequn Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, và Zhengjue Wang. Conzic: Có thể kiểm soát được cú đánh không chú thích hình ảnh bằng cách đánh bóng dựa trên mẫu. Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Tầm nhìn máy tính và Nhận dạng mẫu, trang 23465–23476, 2023. 1, 2, 3, 6

[65] Bành Xuyên Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lôi Chương, Lijuan Wang, Yejin Choi và Jianfeng Gao. Vinvl: Xem lại các biểu diễn trực quan trong ngôn ngữ thị giác mô hình. Trong Biên bản Hội nghị IEEE/CVF về Tầm nhìn máy tính và nhận dạng mẫu, trang 5579–5588, 2021. 6