

ClipCap: CLIP Prefix for Image Captioning

Ron Mokady* Amir Hertz* Amit H. Bermano
The Blavatnik School of Computer Science, Tel Aviv University

Abstract

Image captioning is a fundamental task in vision-language understanding, where the model predicts a textual informative caption to a given input image. In this paper, we present a simple approach to address this task. We use CLIP encoding as a prefix to the caption, by employing a simple mapping network, and then fine-tunes a language model to generate the image captions. The recently proposed CLIP model contains rich semantic features which were trained with textual context, making it best for vision-language perception. Our key idea is that together with a pre-trained language model (GPT2), we obtain a wide understanding of both visual and textual data. Hence, our approach only requires rather quick training to produce a competent captioning model. Without additional annotations or pre-training, it efficiently generates meaningful captions for large-scale and diverse datasets. Surprisingly, our method works well even when only the mapping network is trained, while both CLIP and the language model remain frozen, allowing a lighter architecture with less trainable parameters. Through quantitative evaluation, we demonstrate our model achieves comparable results to state-of-the-art methods on the challenging Conceptual Captions and nocaps datasets, while it is simpler, faster, and lighter. Our code is available in https://github.com/rmokady/CLIP_prefix_caption.

1. Introduction

In image captioning, the task is to provide a meaningful and valid caption for a given input image in a natural language. This task poses two main challenges. The first is semantic understanding. This aspect ranges from simple tasks such as detecting the main object, to more involved ones, such as understanding the relations between depicted parts of the image. For example, in the top-left image of Fig. 1, the model understands that the object is a gift. The second challenge is the large number of possible ways to describe a single image. In this aspect, the training dataset typically dictates the preferable option for a given image.

*Equal contribution.

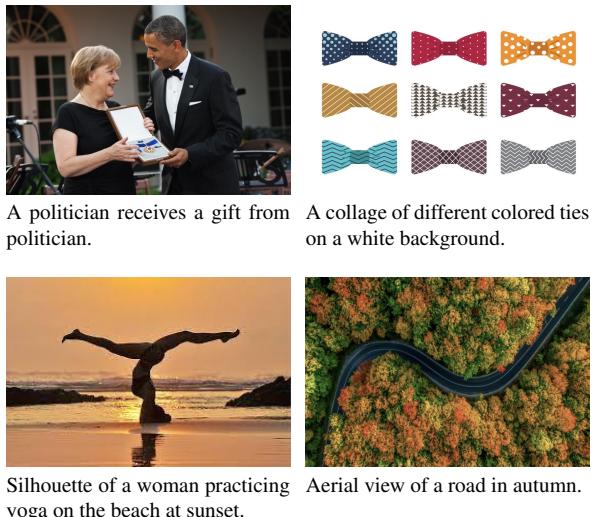


Figure 1. Our ClipCap model produces captions depicting the respective images. Here, the results are of a model that was trained over the Conceptual Captions dataset.

Many approaches have been proposed for image captioning [4, 9, 13, 19, 34, 35, 42, 44, 47]. Typically, these works utilize an encoder for visual cues and a textual decoder to produce the final caption. Essentially, this induces the need to bridge the challenging gap between the visual and textual representations. For this reason, such models are resource hungry. They require extensive training time, a large number of trainable parameters, a massive dataset, and in some cases even additional annotations (such as detection results), which limit their practical applicability.

Excessive training time is even more restrictive for applications that require several training procedures. For instance, training multiple captioning models over various datasets could provide different users (or applications) with different captions for the same image. Additionally, given fresh samples, it is desirable to update the model routinely with the new data. Therefore, a *lightweight* captioning model is preferable. Specifically, a model with faster training times and fewer trainable parameters would be beneficial, especially if it does not require additional supervision.

ClipCap: Tiền tố CLIP để thêm chú thích cho hình ảnh

Ron Mokady* Amir Hertz* Amit H. Bermano
Trường Khoa học Máy tính Blavatnik, Đại học Tel Aviv

Tóm tắt

Chú thích hình ảnh là một nhiệm vụ cơ bản trong việc hiểu ngôn ngữ thị giác, trong đó mô hình dự đoán chú thích thông tin dạng văn bản cho một hình ảnh đầu vào nhất định. Trong Trong bài báo này, chúng tôi trình bày một cách tiếp cận đơn giản để giải quyết nhiệm vụ này. Chúng tôi sử dụng mã hóa CLIP làm tiền tố cho chú thích, bằng cách sử dụng mạng ánh xạ đơn giản, sau đó tinh chỉnh mô hình ngôn ngữ để tạo chú thích hình ảnh. Mô hình CLIP được đề xuất gần đây chứa các tính năng ngữ nghĩa phong phú được đào tạo bằng ngữ cảnh văn bản, làm cho nó tốt nhất cho nhận thức ngôn ngữ thị giác. Ý tưởng chính của chúng tôi là cùng với mô hình ngôn ngữ được đào tạo trước (GPT2), chúng tôi có được sự hiểu biết rộng rãi về cả dữ liệu trực quan và dữ liệu văn bản.

Do đó, cách tiếp cận của chúng tôi chỉ yêu cầu đào tạo khá nhanh để tạo ra một mô hình chú thích có thẩm quyền. Không cần chú thích bổ sung hoặc đào tạo trước, nó tạo ra hiệu quả

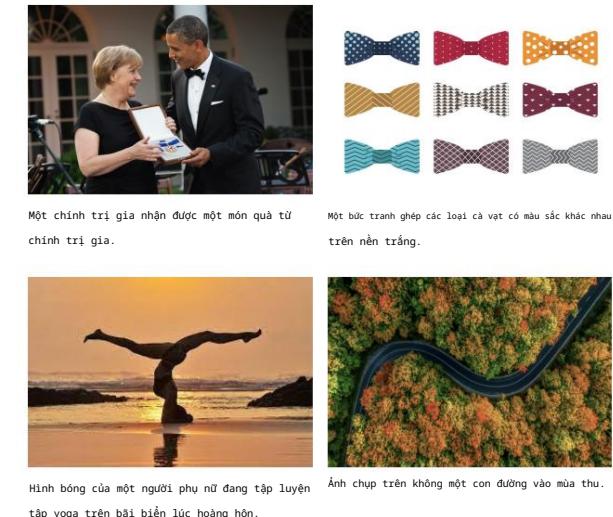
chú thích có ý nghĩa cho các tập dữ liệu đa dạng và quy mô lớn. Thật ngạc nhiên, phương pháp của chúng tôi hoạt động tốt ngay cả khi chỉ có mạng lưới lập bản đồ được đào tạo, trong khi cả CLIP và mô hình ngôn ngữ vẫn bị đóng băng, cho phép có một kiến trúc nhẹ hơn với các tham số ít có thể đào tạo hơn. Thông qua đánh giá định lượng, chúng tôi chứng minh mô hình của chúng tôi đạt được kết quả tương đương với các phương pháp tiên tiến trên các tập dữ liệu Conceptual Captions và nocaps đầy thách thức, trong khi nó đơn giản hơn, nhanh hơn, và nhẹ hơn. Mã của chúng tôi có sẵn tại https://github.com/rmokady/CLIP_prefix_caption.

1. Giới thiệu

Trong chú thích hình ảnh, nhiệm vụ là cung cấp chú thích có ý nghĩa và hợp lệ cho một hình ảnh đầu vào nhất định theo cách tự nhiên ngôn ngữ. Nhiệm vụ này đặt ra hai thách thức chính. Thách thức đầu tiên là sự hiểu biết ngữ nghĩa. Khía cạnh này bao gồm từ đơn giản các nhiệm vụ như phát hiện đối tượng chính, để tham gia nhiều hơn những cái như hiểu được mối quan hệ giữa các hình ảnh được mô tả các phần của hình ảnh. Ví dụ, trong hình ảnh trên cùng bên trái của Hình 1, mô hình hiểu rằng đối tượng là một món quà.

Thách thức thứ hai là số lượng lớn các cách có thể để mô tả một hình ảnh duy nhất. Trong khía cạnh này, tập dữ liệu đào tạo thường chỉ ra lựa chọn ưu tiên cho một hình ảnh nhất định.

*Đóng góp ngang nhau.



Một chính trị gia nhận được một món quà từ chính trị gia.

Một bức tranh ghép các loại cà vạt có màu sắc khác nhau trên nền trắng.



Hình bóng của một người phụ nữ đang tập luyện yoga trên bãi biển lúc hoàng hôn.



Ảnh chụp trên không một con đường vào mùa thu.

Hình 1. Mô hình ClipCap của chúng tôi tạo ra các chú thích mô tả các hình ảnh tương ứng. Ở đây, kết quả là của một mô hình đã được đào tạo trên tập dữ liệu Chú thích khái niệm.

Nhiều cách tiếp cận đã được đề xuất cho việc chú thích hình ảnh [4, 9, 13, 19, 34, 35, 42, 44, 47]. Thông thường, những tác phẩm này sử dụng bộ mã hóa cho các tín hiệu trực quan và bộ giải mã văn bản để tạo ra chú thích cuối cùng. Về cơ bản, điều này tạo ra nhu cầu để thu hẹp khoảng cách đầy thách thức giữa các biểu diễn trực quan và văn bản. Vì lý do này, các mô hình như vậy rất cần tài nguyên. Chúng đòi hỏi thời gian đào tạo mở rộng, một số lượng các tham số có thể đào tạo, một tập dữ liệu lớn và trong một số trường hợp thậm chí còn có thêm chú thích (chẳng hạn như phát hiện kết quả), điều này hạn chế khả năng ứng dụng thực tế của chúng.

Thời gian đào tạo quá mức thậm chí còn hạn chế hơn đối với các ứng dụng yêu cầu nhiều quy trình đào tạo. Ví dụ, đào tạo nhiều mô hình chú thích trên nhiều các tập dữ liệu có thể cung cấp cho người dùng khác nhau (hoặc các ứng dụng) các chú thích khác nhau cho cùng một hình ảnh. Ngoài ra, đưa ra mẫu mới, nên cập nhật mô hình thường xuyên với dữ liệu mới. Do đó, một chú thích nhẹ mô hình được ưa chuộng hơn. Cụ thể, một mô hình có thời gian đào tạo nhanh hơn và ít tham số đào tạo hơn sẽ có lợi, đặc biệt là nếu nó không yêu cầu giám sát bổ sung.

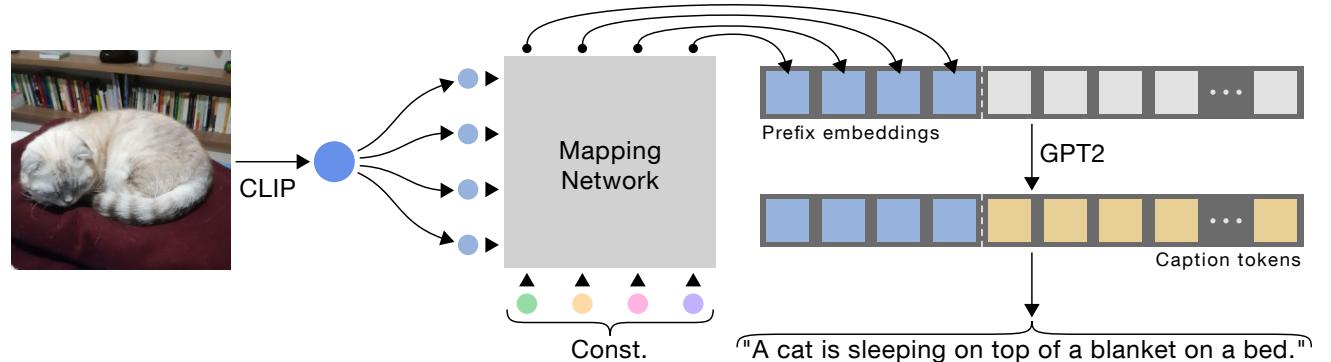
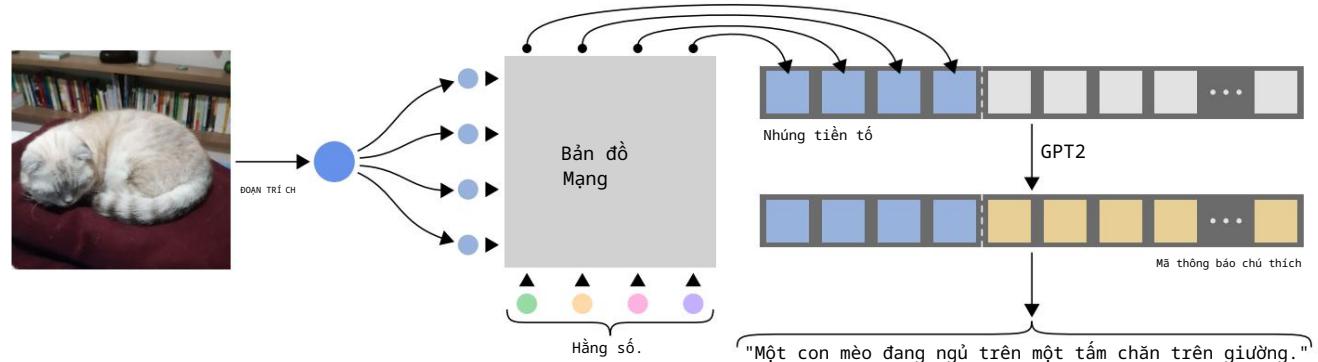


Figure 2. Overview of our transformer-based architecture, enabling the generation of meaningful captions while both CLIP and the language model, GPT-2, are frozen. To extract a fixed length prefix, we train a lightweight transformer-based mapping network from the CLIP embedding space and a learned constant to GPT-2. At inference, we employ GPT-2 to generate the caption given the prefix embeddings. We also suggest a MLP-based architecture, refer to Sec. 3 for more details.

In this paper, we leverage powerful vision-language pre-trained models to simplify the captioning process. More specifically, we use the CLIP (Contrastive Language-Image Pre-Training) encoder, recently introduced by Radford et al. [29]. CLIP is designed to impose a shared representation for both images and text prompts. It is trained over a vast number of images and textual descriptions using a contrastive loss. Hence, its visual and textual representations are well correlated. As we demonstrate, this correlation saves training time and data requirements.

As illustrated in Fig. 2, our method produces a prefix for each caption by applying a mapping network over the CLIP embedding. This prefix is a fixed size embeddings sequence, concatenated to the caption embeddings. These are fed to a language model, which is fine-tuned along with the mapping network training. At inference, the language model generates the caption word after word, starting from the CLIP prefix. This scheme narrows the aforementioned gap between the visual and textual worlds, allowing the employment of a simple mapping network. To achieve even a lighter model, we introduce another variant of our method, where we train only the mapping network, while both CLIP and the language model are kept frozen. By utilizing the expressive transformer architecture, we successfully produce meaningful captions, while imposing substantially less trainable parameters. Our approach is inspired by Li et al. [20], which demonstrates the ability to efficiently adapt a language model for new tasks by concatenating a learned prefix. We use GPT-2 [30] as our language model, which has been demonstrated to generate rich and diverse texts.

As our approach exploits the rich visual-textual representation of CLIP, our model requires significantly lower training time. For instance, we train our model on a single Nvidia GTX1080 GPU for 80 hours over the three million samples of the massive Conceptual Captions dataset. Nevertheless, our model generalizes well to complex scenes, as



Hình 2. Tổng quan về kiến trúc dựa trên bộ chuyển đổi của chúng tôi, cho phép tạo ra các chủ thích có ý nghĩa trong khi cả CLIP và mô hình ngôn ngữ, GPT-2, bị đóng băng. Để trích xuất một tiền tố có độ dài cố định, chúng tôi đào tạo một mạng lưới ánh xạ dựa trên bộ biến đổi nhẹ từ không gian nhúng CLIP và hàng số học được dồi với GPT-2. Khi suy luận, chúng tôi sử dụng GPT-2 để tạo chủ thích cho tiền tố nhúng. Chúng tôi cũng đề xuất kiến trúc dựa trên MLP, tham khảo Mục 3 để biết thêm chi tiết.

Trong bài báo này, chúng tôi tận dụng các mô hình ngôn ngữ thị giác được đào tạo trước mạnh mẽ để đơn giản hóa quá trình chủ thích.Thêm cụ thể, chúng tôi sử dụng CLIP (Ngôn ngữ tương phản-Hình ảnh Bộ mã hóa Pre-Training), mới được giới thiệu gần đây bởi Radford et al. [29]. CLIP được thiết kế để áp đặt một biểu diễn chung cho cả hình ảnh và lời nhắc văn bản. Nó được đào tạo qua một số lượng lớn hình ảnh và mô tả văn bản sử dụng một mắt mèo tương phản. Do đó, các biểu diễn trực quan và văn bản của nó có mối tương quan tốt. Như chúng tôi chứng minh, mối tương quan này tiết kiệm thời gian đào tạo và yêu cầu dữ liệu.

Như minh họa trong Hình 2, phương pháp của chúng tôi tạo ra một tiền tố cho mỗi chủ thích bằng cách áp dụng mạng lưới ánh xạ trên nhúng CLIP. Tiền tố này là một nhúng có kích thước cố định chuỗi, được nối với các nhúng chủ thích. Những nhúng này được đưa vào một mô hình ngôn ngữ, được tinh chỉnh cùng với đào tạo mạng lưới lập bản đồ. Khi suy luận, ngôn ngữ mô hình tạo ra chủ thích từ sau từ, bắt đầu từ tiền tố CLIP. Số đó này thu hẹp những điều đã đề cập ở trên khoảng cách giữa thế giới trực quan và văn bản, cho phép sử dụng mạng lưới bản đồ đơn giản. Để đạt được ngay cả một mô hình nhẹ hơn, chúng tôi giới thiệu một biến thể khác của phương pháp của chúng tôi, nơi chúng tôi chỉ đào tạo mạng lưới lập bản đồ, trong khi cả CLIP và mô hình ngôn ngữ được giữ nguyên. Bằng cách sử dụng kiến trúc biến đổi biểu cảm, chúng tôi đã thành công trong việc tạo ra các chủ thích có ý nghĩa, đồng thời áp đặt ít hơn đáng kể các thông số có thể đào tạo được. Cách tiếp cận của chúng tôi được lấy cảm hứng từ Li et al. [20], chứng minh khả năng thích ứng hiệu quả một mô hình ngôn ngữ cho các nhiệm vụ mới bằng cách nối kết một tiền tố. Chúng tôi sử dụng GPT-2 [30] làm mô hình ngôn ngữ của chúng tôi, đã được chứng minh là có thể tạo ra các văn bản phong phú và đa dạng.

Vì cách tiếp cận của chúng tôi khai thác biểu diễn văn bản trực quan phong phú của CLIP, mô hình của chúng tôi yêu cầu thấp hơn đáng kể thời gian đào tạo. Ví dụ, chúng tôi đào tạo mô hình của mình trên một GPU Nvidia GTX1080 trong 80 giờ trên ba triệu mẫu của tập dữ liệu hình ảnh và văn bản. Nhiều tác phẩm đã sử dụng CLIP thành công cho các tác vụ thị giác máy tính đòi hỏi sự hiểu biết về một số văn bản phụ trợ, chẳng hạn như tạo hoặc chỉnh sửa hình ảnh dựa trên điều kiện ngôn ngữ tự nhiên [5, 14, 28]. Trong bài báo này, chúng tôi sử dụng mô hình CLIP cho nhiệm vụ chủ thích hình ảnh. Lưu ý rằng phương pháp không sử dụng bộ mã hóa văn bản của CLIP, vì không có văn bản đầu vào và văn bản đầu ra được tạo ra bởi một

có thể được nhìn thấy trong Hình 1 (ví dụ, tập yoga trên bãi biển tại hoàng hôn). Chúng tôi đánh giá phương pháp của mình một cách rộng rãi, chứng minh chủ thích thực tế và có ý nghĩa thành công. Mặc dù mô hình của chúng tôi cần ít thời gian đào tạo hơn, nó vẫn đạt được kết quả tương đương với các phương pháp tiên tiến trên các tập dữ liệu đầy thách thức là Chủ thích khái niệm [33] và nocaps [1], và thấp hơn một chút đối với COCO hạn chế hơn [7, 22] chuẩn mực. Ngoài ra, chúng tôi cung cấp một phân tích toàn diện về độ dài tiền tố cần thiết và hiệu ứng của việc tinh chỉnh mô hình ngôn ngữ, bao gồm cả việc giải thích sản phẩm của chúng tôi tiền tố. Nhìn chung, những đóng góp chính của chúng tôi như sau:

- Một phương pháp chủ thích nhẹ sử dụng các mô hình đóng băng được đào tạo trước để xử lý cả hình ảnh và văn bản.
- Ngay cả khi mô hình ngôn ngữ được tinh chỉnh, cách tiếp cận của chúng tôi vẫn đơn giản và đào tạo nhanh hơn, đồng thời chứng minh kết quả tương đương với công nghệ tiên tiến trên các tập dữ liệu đầy thách thức.

2. Các tác phẩm liên quan

Gần đây, Radford et al. [29] đã trình bày một cách tiếp cận mới, được gọi là CLIP, để cùng nhau biểu diễn hình ảnh và mô tả văn bản. CLIP bao gồm hai bộ mã hóa, một cho hình ảnh tín hiệu và một cho văn bản. Nó đã được đào tạo qua hơn 400 hàng triệu cặp hình ảnh-văn bản được hướng dẫn bởi sự tương phản không giám sát, dẫn đến không gian tiềm ẩn ngữ nghĩa phong phú được chia sẻ bởi cả hai dữ liệu hình ảnh và văn bản. Nhiều tác phẩm đã sử dụng CLIP thành công cho các tác vụ thị giác máy tính đòi hỏi sự hiểu biết về một số văn bản phụ trợ, chẳng hạn như tạo hoặc chỉnh sửa hình ảnh dựa trên điều kiện ngôn ngữ tự nhiên [5, 14, 28]. Trong bài báo này, chúng tôi sử dụng mô hình CLIP cho nhiệm vụ chủ thích hình ảnh. Lưu ý rằng phương pháp không sử dụng bộ mã hóa văn bản của CLIP, vì không có văn bản đầu vào và văn bản đầu ra được tạo ra bởi một

language model.

Commonly, image captioning [34] models first encode the input pixels as feature vectors, which are then used to produce the final sequence of words. Early works utilize the features extracted from a pre-trained classification network [6, 9, 13, 42], while later works [4, 19, 47] exploit the more expressive features of an object detection network [31]. Though a pre-trained object detection network is available for the popular COCO benchmark [7, 22], it is not necessarily true for other datasets. This implies that most methods would require additional object detection annotations to operate over new and diverse datasets. To further leverage the visual cues, an attention mechanism is usually utilized [4, 6, 42] to focus on specific visual features. Moreover, recent models apply self-attention [16, 43] or use an expressive visual Transformer [12] as an encoder [23]. Our work uses the expressive embedding of CLIP for visual representation. Since CLIP was trained over an extremely large number of images, we can operate on any set of natural images without additional annotations.

To produce the caption itself, a textual decoder is employed. Early works have used LSTM variants [8, 38, 39], while recent works [16, 26] adopted the improved transformer architecture [36]. Built upon the transformer, one of the most notable works is BERT [11], demonstrating the dominance of the newly introduced paradigm. With this paradigm, the language model is first pre-trained over a large data collection to solve an auxiliary task. Then, the model is fine-tuned for a specific task, where additional supervision is used. As our visual information resides in the prefix, we utilize a powerful auto-regressive language model, GPT-2 [30]. Considering the training loss term, earlier works adopt the effective cross-entropy, while contemporary methods also apply self-critical sequence training [15, 32, 45]. That is, an additional training stage to optimize the CIDEr metric. We deliberately refrain from this optimization to retain a quick training procedure.

Most close to ours, are works that employ vision-and-language pre-training to create a shared latent space of both vision and text [19, 25, 35, 46, 47]. Zhou et al. [47] use visual tokens extracted from object detector as a prefix to caption tokens. The entire model is then pre-trained to perform prediction utilizing the BERT [11] architecture. Li et al. [19] and Zhang et al. [46] also utilize BERT, but require the additional supervision of object tags. Hence, these methods are limited to datasets in which such object detectors or annotations are available. The approach of Wang et al. [40] mitigate the need for supplementary annotations, but still perform an extensive pre-train process with millions of image-text pairs, resulting in a lengthy training time. This exhaustive pre-training step is required to compensate for the lack of joint representation of language and vision, which we inherently obtained by employing CLIP.

3. Method

We start with our problem statement. Given a dataset of paired images and captions $\{x^i, c^i\}_{i=1}^N$, our goal is to learn the generation of a meaningful caption for an unseen input image. We can refer to the captions as a sequence of tokens $c^i = c_1^i, \dots, c_\ell^i$, where we pad the tokens to a maximal length ℓ . Our training objective is then the following:

$$\max_{\theta} \sum_{i=1}^N \log p_{\theta}(c_1^i, \dots, c_\ell^i | x^i), \quad (1)$$

where θ denotes the model's trainable parameters. Our key idea is to use the rich semantic embedding of CLIP, which contains, virtually, the essential visual data, as a condition. Following recent works [47], we consider the condition as a prefix to the caption. Since the required semantic information is encapsulated in the prefix, we can utilize an autoregressive language model that predicts the next token without considering future tokens. Thus, our objective can be described as:

$$\max_{\theta} \sum_{i=1}^N \sum_{j=1}^{\ell} \log p_{\theta}(c_j^i | x^i, c_1^i, \dots, c_{j-1}^i) \quad (2)$$

3.1. Overview

An illustration of our method is provided in Fig. 2. We use GPT-2 (large) as our language model, and utilize its tokenizer to project the caption to a sequence of embeddings. To extract visual information from an image x^i , we use the visual encoder of a pre-trained CLIP [29] model. Next, we employ a light mapping network, denoted F , to map the CLIP embedding to k embedding vectors:

$$p_1^i, \dots, p_k^i = F(\text{CLIP}(x^i)). \quad (3)$$

Where each vector p_j^i has the same dimension as a word embedding. We then concatenate the obtained visual embedding to the caption c^i embeddings:

$$Z^i = p_1^i, \dots, p_k^i, c_1^i, \dots, c_\ell^i. \quad (4)$$

During training, we feed the language model with the prefix-caption concatenation $\{Z^i\}_{i=1}^N$. Our training objective is predicting the caption tokens conditioned on the prefix in an autoregressive fashion. To this purpose, we train the mapping component F using the simple, yet effective, cross-entropy loss:

$$\mathcal{L}_X = - \sum_{i=1}^N \sum_{j=1}^{\ell} \log p_{\theta}(c_j^i | p_1^i, \dots, p_k^i, c_1^i, \dots, c_{j-1}^i). \quad (5)$$

We now turn to discuss two variants of our method regarding the additional fine-tuning of the language model and their implications.

mô hình ngôn ngữ.

Thông thường, các mô hình chú thích hình ảnh [34] trước tiên mã hóa các pixel đầu vào dưới dạng các vectơ đặc điểm, sau đó được sử dụng để tạo ra chuỗi từ cuối cùng. Các tác phẩm ban đầu sử dụng các đặc điểm được trích xuất từ mạng phân loại được đào tạo trước [6, 9, 13, 42], trong khi các tác phẩm sau đó [4, 19, 47] khai thác các đặc điểm biểu cảm hơn của mạng phát hiện đối tượng [31]. Mặc dù mạng phát hiện đối tượng được đào tạo trước có sẵn cho chuẩn mực COCO phổ biến [7, 22], nhưng nó không nhất thiết đúng với các tập dữ liệu khác. Điều này ngụ ý rằng hầu hết các phương pháp sẽ yêu cầu chú thích phát hiện đối tượng bổ sung để vận hành trên các tập dữ liệu mới và đa dạng. Để tận dụng thêm các tín hiệu trực quan, một cơ chế chú ý thường được sử dụng [4, 6, 42] để tập trung vào các đặc điểm trực quan cụ thể. Hơn nữa, các mô hình gần đây áp dụng sự tự chú ý [16, 43] hoặc sử dụng bộ chuyển đổi trực quan biểu cảm [12] làm bộ mã hóa [23]. Tác phẩm của chúng tôi sử dụng những biểu cảm của CLIP để biểu diễn trực quan. Vì CLIP được đào tạo trên số lượng hình ảnh cực lớn nên chúng ta có thể xử lý bất kỳ tập hợp hình ảnh tự nhiên nào mà không cần chú thích bổ sung.

3. Phương pháp

Chúng tôi bắt đầu với tuyên bố vấn đề của chúng tôi. Cho một tập dữ liệu hình ảnh và chú thích được ghép nối (x mục tiêu của chúng tôi) $\{x, c\}$ cách tạo ra chú thích có ý nghĩa cho hình ảnh đầu vào chưa được nhìn thấy. Chúng ta có thể tham chiếu đến chú thích như một chuỗi các token $c = c_1, \dots, c_\ell$, nơi chúng tôi đệm các mã thông báo đến mức tối đa chiều dài. Mục tiêu đào tạo của chúng tôi sau đó là:

$$\max_{\theta} \sum_{i=1}^N \log p_{\theta}(c_1^i, \dots, c_\ell^i | x^i), \quad (1)$$

trong đó θ biểu thị các tham số có thể đào tạo của mô hình. Ý tưởng chính của chúng tôi là sử dụng những ngữ nghĩa phong phú của CLIP, chưa dữ liệu trực quan cần thiết về mặt áo, như một điều kiện. Tiếp theo các công trình gần đây [47], chúng tôi coi điều kiện là tiền tố cho chú thích. Vì thông tin ngữ nghĩa cần thiết được đóng gói trong tiền tố, chúng tôi có thể sử dụng mô hình ngôn ngữ hồi quy tự động dự đoán mã thông báo tiếp theo mà không cần xem xét các mã thông báo trong tương lai. Do đó, mục tiêu của chúng tôi có thể được mô tả như sau:

$$\max_{\theta} \sum_{i=1}^N \log p_{\theta}(c_j^i | x^i, c_1^i, \dots, c_{j-1}^i) \quad (2)$$

3.1. Tổng quan

Mình họa về phương pháp của chúng tôi được cung cấp trong Hình 2. Chúng tôi sử dụng GPT-2 (lớn) làm mô hình ngôn ngữ của mình và sử dụng tokenizer của nó để chiếu chú thích vào một chuỗi nhúng. Chúng tôi sử dụng để trích xuất thông tin trực quan từ bộ mã hóa hình ảnh x của mô hình CLIP được đào tạo trước [29]. Tiếp theo, chúng tôi sử dụng mạng ánh xạ ánh sáng, được ký hiệu là F , để ánh xạ nhúng CLIP thành k vectơ nhúng: $1, \dots, p_k$.

$$tối đa \log p_{\theta}(c_j^i | x^i, c_1^i, \dots, c_{j-1}^i) \quad (3)$$

Trong đó mỗi vectơ p_j nhúng có cùng kích thước với một từ. Sau đó, chúng tôi nối nhúng trực quan thu được vào nhúng chú thích c :

$$Z^i = p_1^i, \dots, p_k^i, c_1^i, \dots, c_\ell^i. \quad (4)$$

Trong quá trình đào tạo, chúng tôi cung cấp cho mô hình ngôn ngữ nội tiền tố-iêu đề (Z^i). Mục tiêu đào tạo của chúng tôi là dự đoán các mã thông báo chú thích có điều kiện trên tiền tố theo cách tự hồi quy. Với mục đích này, chúng tôi đào tạo thành phần ánh xạ F bằng cách sử dụng mắt mèo entropy chéo đơn giản nhưng hiệu quả:

$$L_X = \sum_{i=1}^N \sum_{j=1}^{\ell} \log p_{\theta}(c_j^i | p_1^i, \dots, p_k^i, c_1^i, \dots, c_{j-1}^i) \quad (5)$$

Bây giờ chúng ta sẽ thảo luận về hai biến thể của phương pháp liên quan đến việc điều chỉnh thêm mô hình ngôn ngữ và ý nghĩa của chúng.

3.2. Language model fine-tuning

Our main challenge during training is to translate between the representations of CLIP and the language model. Even though both models develop a rich and diverse representation of text, their latent spaces are independent, as they were not jointly trained. Moreover, each captioning dataset incorporates a different style, which may not be natural for the pre-trained language model. Hence, we propose fine-tuning the language model during the training of the mapping network. This provides additional flexibility for the networks and yields a more expressive outcome.

However, fine-tuning the language model naturally increases the number of trainable parameters substantially. Thus, we present an additional variant of our approach, in which we keep the language model fixed during training. Our attempt to adjust a frozen language model is inspired by the work of Li and Liang [20]. In their work, they accommodate such a pre-trained model to an unfamiliar task by learning only a prefix. Such prefix is automatically optimized to steer the language model towards the new objective during a standard training procedure. Following this approach, we suggest avoiding the fine-tuning to realize an even lighter model, where only the mapping network is trained. As presented in Section 4, our model not only produces realistic and meaningful captions, but also achieves superior results for some of the experiments without fine-tuning the language model. Note that fine-tuning CLIP does not benefit resulting quality, but does increase training time and complexity. We hence postulate that the CLIP space already encapsulates the required information, and adapting it towards specific styles does not contribute to flexibility.

3.3. Mapping Network Architecture

Our key component is the mapping network, which translates the CLIP embedding to the GPT-2 space. When the language model is simultaneously fine-tuned, the mapping is less challenging, as we easily control both networks. Therefore, in this case, we can employ a simple Multi-Layer Perceptron (MLP). We have achieved realistic and meaningful captions even when utilizing only a single hidden layer, as CLIP is pre-trained for a vision-language objective.

Nevertheless, when the language model is frozen, we propose utilizing the more expressive transformer [36] architecture. The transformer enables global attention between input tokens while reducing the number of parameters for long sequences. This allows us to improve our

results by increasing prefix size, as shown in Section 4. We feed the transformer network with two inputs, the visual encoding of CLIP and a learned constant input. The constant has a dual role, first, to retrieve meaningful information from CLIP embedding through the multi-head attention. Second, it learns to adjust the fixed language model to the new data. This is demonstrated in Section 4, where

we offer interpretability for our generated prefix. As can be seen, when the language model is fixed, the transformer mapping network learns a meticulous set of embeddings without any textual meaning. These are optimized to tame the language model.

3.4. Inference

During inference, we extract the visual prefix of an input image x using the CLIP encoder and the mapping network F . We start generating the caption conditioned on the visual prefix, and predict the next tokens one by one, guided by the language model output. For each token, the language model outputs probabilities for all vocabulary tokens, which are used to determine the next one by employing a greedy approach or beam search.

4. Results

Datasets. We use the *COCO-captions* [7, 22], *nocaps* [1], and *Conceptual Captions* [33] datasets. We split the former according to the Karpathy et al. [17] split, where the training set contains 120,000 images and 5 captions per image. Since COCO is limited to 80 classes, the nocaps dataset is designed to measure generalization to unseen classes and concepts. It contains only validation and test sets, with the training utilizing COCO itself. The nocaps dataset is divided to three parts — *in-domain* contains images portraying only COCO classes, *near-domain* contains both COCO and novel classes, and *out-of-domain* consists of only novel classes. As suggested by Li et al. [19], we evaluate the model using only the validation set. Though some methods utilize object tags of the novel classes, we only consider the setting of no additional supervision, as we find it more applicable in practice. Therefore, we do not employ a constrained beam search [2].

The Conceptual Captions dataset consists of 3M pairs of images and captions, harvested from the web and post-processed. It is considered to be more challenging than COCO due to the larger variety of styles of both the images and the captions, while not limited to specific classes. To focus on the concepts, specific entities in this dataset are replaced with general notions. For example, in Fig. 1, the names are replaced with “politician”. For evaluation, we use the validation set, consisting of 12.5K images, as the test set is not publicly available. Consequently, we did not use this set for validation.

Baselines. We compare our method to the state-of-the-art works of Li et al. [19] (known as Oscar), Vision-Language Pre-training model (VLP) [47], and the eminent work of Anderson et al. [4], denoted BUTD. These models first produce visual features using an object detection network [31]. BUTD then utilizes an LSTM to generate the captions, while VLP and Oscar employ a transformer, trained simi-

3.2. Tinh chỉnh mô hình ngôn ngữ

Thách thức chính của chúng tôi trong quá trình đào tạo là chuyển đổi giữa các biểu diễn của CLIP và mô hình ngôn ngữ. Mặc dù cả hai mô hình đều phát triển một cách biểu diễn văn bản phong phú và đa dạng, nhưng không gian tiềm ẩn của chúng lại độc lập, vì họ không được đào tạo chung. Hơn nữa, mỗi phụ đề tập dữ liệu kết hợp một phong cách khác, có thể không phải là tự nhiên đối với mô hình ngôn ngữ được đào tạo trước. Do đó, chúng tôi đã xuất tinh chỉnh mô hình ngôn ngữ trong quá trình đào tạo mạng lưới lập bản đồ. Điều này cung cấp thêm tính linh hoạt cho các mạng lưới và mang lại kết quả biểu cảm hơn.

Tuy nhiên, việc tinh chỉnh mô hình ngôn ngữ tự nhiên sẽ làm tăng đáng kể số lượng tham số có thể đào tạo được. Vì vậy, chúng tôi trình bày một biến thể bổ sung của cách tiếp cận của chúng tôi, trong mà chúng tôi giữ nguyên mô hình ngôn ngữ trong quá trình đào tạo. Nỗ lực của chúng tôi nhằm điều chỉnh một mô hình ngôn ngữ đóng băng được truyền cảm hứng bằng công trình của Li và Liang [20]. Trong công trình của họ, họ điều chỉnh một mô hình được đào tạo trước như vậy cho một nhiệm vụ không quen thuộc bằng cách chỉ học một tiền tố. Tiền tố như vậy được tự động tối ưu hóa để điều hướng mô hình ngôn ngữ hướng tới mục tiêu mới trong quá trình đào tạo chuẩn. Sau đây

cách tiếp cận, chúng tôi đã xuất tránh việc tinh chỉnh để nhận ra

một mô hình thậm chí còn nhẹ hơn, trong đó chỉ có mạng lưới lập bản đồ là

được đào tạo. Như đã trình bày trong Phần 4, mô hình của chúng tôi không chỉ tạo ra các chủ thích thực tế và có ý nghĩa mà còn đạt được

kết quả vượt trội cho một số thí nghiệm mà không cần tinh chỉnh mô hình ngôn ngữ. Lưu ý rằng việc tinh chỉnh CLIP

không mang lại lợi ích về chất lượng nhưng làm tăng thời gian đào tạo

và sự phức tạp. Do đó, chúng tôi đưa ra giả thuyết rằng không gian CLIP

đã bao gồm thông tin cần thiết và việc điều chỉnh

việc hướng tới những phong cách cụ thể không góp phần tạo nên sự linh hoạt.

3.3. Lập bản đồ kiến trúc mạng

Thành phần chính của chúng tôi là mạng lưới lập bản đồ, dịch nhúng CLIP sang không gian GPT-2. Khi mô hình ngôn ngữ được tinh chỉnh đồng thời, map-ping ít thách thức hơn vì chúng ta có thể dễ dàng kiểm soát cả hai mạng. Do đó, trong trường hợp này, chúng ta có thể sử dụng một Multi-Layer đơn giản Perceptron (MLP). Chúng tôi đã đạt được các chủ thích thực tế và có ý nghĩa ngay cả khi chỉ sử dụng một lớp ẩn duy nhất, vì CLIP được đào tạo trước cho mục tiêu ngôn ngữ thị giác.

Tuy nhiên, khi mô hình ngôn ngữ bị đóng băng, chúng ta

đã xuất sử dụng kiến trúc biến đổi biểu cảm hơn [36]. Biến đổi cho phép

chú ý toàn cầu giữa các mã thông báo đầu vào trong khi giảm số lượng tham số cho các chuỗi dài. Điều này cho phép chúng tôi cải thiện

kết quả bằng cách tăng kích thước tiền tố, như thể hiện trong Phần 4.

Chúng tôi cung cấp cho mạng lưới máy biến áp hai đầu vào, mã hóa hình ảnh của CLIP và đầu vào hàng số đã học.

hàng số có vai trò kép, đầu tiên là thu thập thông tin có ý nghĩa từ những CLIP thông qua sự chú ý nhiều đầu. Thứ hai, nó học cách điều chỉnh mô hình ngôn ngữ cố định để

dữ liệu mới. Điều này được chứng minh trong Phần 4, trong đó

chúng tôi cung cấp khả năng giải quyết tiền tố được tạo ra của chúng tôi. Như có thể nhìn thấy, khi mô hình ngôn ngữ được cố định, bộ chuyển đổi mạng lưới lập bản đồ học một tập hợp nhung tí mi không có bất kỳ ý nghĩa văn bản nào. Chúng được tối ưu hóa để thuần hóa mô hình ngôn ngữ.

3.4. Suy luận

Trong quá trình suy luận, chúng tôi trích xuất tiền tố trực quan của một đầu vào hình ảnh x sử dụng bộ mã hóa CLIP và mạng lưới ánh xạ F. Chúng tôi bắt đầu tạo chủ thích dựa trên hình ảnh tiền tố và dự đoán các mã thông báo tiếp theo từng cái một, được hướng dẫn bởi đầu ra của mô hình ngôn ngữ. Đối với mỗi mã thông báo, ngôn ngữ mô hình đưa ra xác suất cho tất cả các mã thông báo từ vựng, được sử dụng để xác định cái tiếp theo bằng cách sử dụng một tham lam tiếp cận hoặc tìm kiếm chùm tia.

4. Kết quả

Bộ dữ liệu. Chúng tôi sử dụng *COCO-captions* [7, 22], *nocaps* [1], và các tập dữ liệu Chủ thích khái niệm [33]. Chúng tôi chia tách các tập trước theo Karpathy et al. [17] phân chia, trong đó bộ đào tạo chứa 120.000 hình ảnh và 5 chủ thích cho mỗi hình ảnh. Vì *COCO* bị giới hạn ở 80 lớp nên tập dữ liệu *nocaps* là được thiết kế để do lưỡng sự khai quát hóa đối với các lớp chưa biết và các khái niệm. Nó chỉ chứa các bộ xác thực và kiểm tra, với đào tạo sử dụng chính *COCO*. Bộ dữ liệu *nocaps* được chia thành ba phần – trong miền chứa hình ảnh chỉ mô tả các lớp *COCO*, gần miền chứa cả *COCO* và các lớp mới lạ, và ngoài miền chỉ bao gồm các lớp mới lạ các lớp. Theo đề xuất của Li et al. [19], chúng tôi đánh giá mô hình chỉ sử dụng bộ xác thực. Mặc dù một số phương pháp sử dụng thê đổi tượng của các lớp mới, chúng tôi chỉ xem xét cài đặt không có giám sát bổ sung, vì chúng tôi thấy nó

áp dụng nhiều hơn trong thực tế. Do đó, chúng tôi không sử dụng tìm kiếm chùm tia bị hạn chế [2]. Các chủ thích khái niệm tập dữ liệu bao gồm 3M cặp hình ảnh và chủ thích, được thu thập từ web và xử lý hậu kỳ. Nó được coi là trở nên khó khăn hơn *COCO* do có nhiều phong cách đa dạng hơn về cả hình ảnh và chủ thích, trong khi không giới hạn ở các lớp cụ thể. Để tập trung vào các khái niệm, các thực thể cụ thể trong tập dữ liệu này được thay thế bằng các khái niệm chung. Ví dụ, trong Hình 1, các tên được thay thế bằng “chính trị gia”. Để đánh giá, chúng tôi sử dụng bộ xác thực, bao gồm 12,5K hình ảnh, vì bộ thử nghiệm không được công khai. Do đó, chúng tôi không sử dụng bộ này để xác thực.

Đường cơ sở. Chúng tôi so sánh phương pháp của chúng tôi với phương pháp tiên tiến nhất tác phẩm của Li et al. [19] (được gọi là Oscar), Vision-Language Mô hình tiền đào tạo (VLP) [47] và công trình nổi bật của Anderson et al. [4], được ký hiệu là BUTD. Các mô hình này đều tiên tạo ra các đặc điểm trực quan bằng cách sử dụng mạng phát hiện đối tượng [31]. BUTD sau đó sử dụng LSTM để tạo ra các chủ thích, trong khi VLP và Oscar sử dụng một máy biến áp, được đào tạo simi-

(A) Conceptual Captions									
Model	ROUGE-L ↑	CIDEr ↑	SPICE ↑	#Params (M) ↓	Training Time ↓				
VLP	24.35	77.57	16.59	115	1200h (V100)				
Ours; MLP + GPT2 tuning	26.71	87.26	18.5	156	80h (GTX1080)				
Ours; Transformer	25.12	71.82	16.07	43	72h (GTX1080)				

(B) nocaps										
Model	in-domain		near-domain		out-of-domain		Overall		Time ↓	
	CIDEr↑	SPICE↑	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE		
BUTD [4]	74.3	11.5	56.9	10.3	30.1	8.1	54.3	10.1	52	960h
Oscar [19]	79.6	12.3	66.1	11.5	45.3	9.7	63.8	11.2	135	74h
Ours; MLP + GPT2 tuning	79.73	12.2	67.69	11.26	49.35	9.7	65.7	11.1	156	7h
Ours; Transformer	84.85	12.14	66.82	10.92	49.14	9.57	65.83	10.86	43	6h

(C) COCO									
Model	B@4 ↑	METEOR ↑	CIDEr ↑	SPICE ↑	#Params (M) ↓	Training Time ↓			
BUTD [4]	36.2	27.0	113.5	20.3	52	960h (M40)			
VLP [47]	36.5	28.4	117.7	21.3	115	48h (V100)			
Oscar [19]	36.58	30.4	124.12	23.17	135	74h (V100)			
Ours; Transformer	33.53	27.45	113.08	21.05	43	6h (GTX1080)			
Ours; MLP + GPT2 tuning	32.15	27.1	108.35	20.12	156	7h (GTX1080)			

(D) Ablation									
Ours; Transformer + GPT2 tuning	32.22	27.79	109.83	20.63	167	7h (GTX1080)			
Ours; MLP	27.39	24.4	92.38	18.04	32	6h (GTX1080)			

Table 1. Quantitative evaluation. As can be seen, our method achieves comparable results for both nocaps and Conceptual Captions with much faster training time.

Ground Truth	A man with a red helmet on a small moped on a dirt road	A young girl inhales with the intent of blowing out a candle	A man on a bicycle riding next to a train	a wooden cutting board topped with sliced up food	A kitchen is shown with a variety of items on the counters
Oscar	a man riding a motorcycle down a dirt road	a woman sitting at a table with a plate of food	a woman riding a bike down a street next to a train	a woman sitting at a table with a plate of food	a kitchen with a sink, dishwasher and a window
Ours; MLP + GPT2 tuning	a man riding a motorcycle on a dirt road	a woman is eating a piece of cake with a candle	a man is standing next to a train	a row of wooden cutting boards with wooden spoons	a kitchen with a sink, stove, and window
Ours; Transformer	a man is riding a motorcycle on a dirt road	a young girl sitting at a table with a cup of cake	a man is standing next to a train	a wooden table with a bunch of wood tools on it	a kitchen with a sink and a window

Figure 3. Uncurated results of the first five images in the COCO test set (Karpathy et al. [17] split).

(A) Chú thích khái niệm									
Người mẫu	ROUGE-L	CIDEr	SPICE	#Params (M)	Thời gian đào tạo				
VLP	24.35	77.57	16.59	115	1200 giờ (V100)				
Của chúng tôi; MLP + GPT2 điều chỉnh	26.71	87.26	18.5	156	80 giờ (GTX1080)				
Của chúng tôi; Máy biến áp	25.12	71.82	16.07	43	72 giờ (GTX1080)				

(B) không có nắp										
Người mẫu	trong miền		gần miền ngoài miền		Tổng thể					
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	Tham số	Thời gian		
NHƯNG [4]	74,3	11,5	56,9	10,3	30,1	8,1	54,3	10,1	52	960 giờ
Oscar [19]	79,6	12,3	66,1	11,5	45,3	9,7	63,8	11,2	135	74 giờ
Của chúng tôi; MLP + GPT2 điều chỉnh	79,73	12,2	67,69	11,26	49,35	9,7	65,7	11,1	156	7 giờ
Của chúng tôi; Máy biến áp	84,85	12,14	66,82	10,92	49,14	9,57	65,83	10,86	43	6 giờ

(C) COCO									
Người mẫu	B@4	METEOR	CIDEr	SPICE	#Params (M)	Thời gian đào tạo			
NHƯNG [4]	36,2	27,0	113,5	20,3	52	960 giờ (M40)			
VLP [47]	36,5	28,4	117,7	21,3	115	48 giờ (V100)			
Oscar [19]	36,58	30,4	124,12	23,17	135	74 giờ (V100)			
Của chúng tôi; Máy biến áp	33,53	27,45	113,08	21,05	43	6 giờ (GTX1080)			
Của chúng tôi; Điều chỉnh MLP + GPT2	32,15	27,1	108,35	20,12	156	7 giờ (GTX1080)			

(D) Phá hủy									
Của chúng tôi; Biến áp + điều chỉnh GPT2	32,22	27,79	109,83	20,63	167	7 giờ (GTX1080)			

Bảng 1. Đánh giá định lượng. Như có thể thấy, phương pháp của chúng tôi đạt được kết quả tương đương cho cả nocaps và chú thích khái niệm với thời gian đào tạo nhanh hơn nhiều.

<tbl_r cells="5" ix="1" maxcspan="1" maxrspan="1" usedcols="5



Ground Truth	A life in photography – in pictures.	Photograph of the sign being repaired by brave person.	Globes : the green 3d person carrying in hands globe.	The player staring intently at a computer screen.	The - bedroom stone cottage can sleep people.
VLP	Actors in a scene from the movie.	The sign at the entrance.	Templates: green cartoon character holding the earth globe.	Person works on a video.	The master bedroom has a king - sized bed with a queen size bed.
Ours; MLP + GPT2 tuning	Actor sits in a hotel room.	The sign at the entrance.	3d render of a man holding a globe.	Person, a student, watches a video on his laptop.	The property is on the market for £ 1.
Ours; Transformer	person sitting on a chair in a room.	a sign is seen at the entrance to the store.	stock image of a man holding the earth.	portrait of a young boy playing video game.	one of the bedrooms in the house has been converted into a living room.

Figure 4. Uncurated results of the first five images in our test set for Conceptual Captions [33].

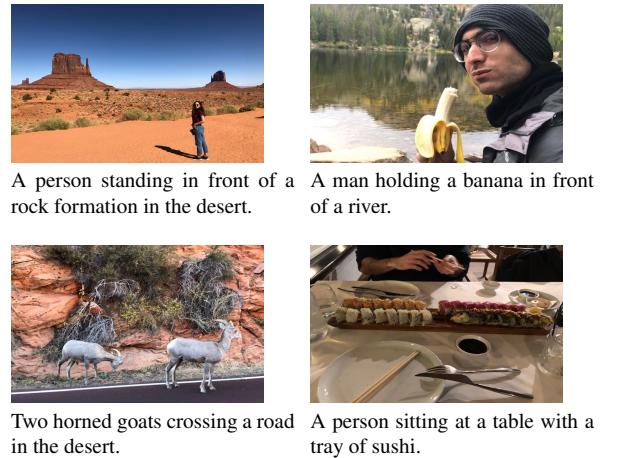


Figure 5. Results over smartphone photos. Top: using our Conceptual Captions model. Bottom: COCO model. As demonstrated, our approach generalizes well to newly photographed images.

larly to BERT [11]. Both VLP and Oscar exploit an extensive pre-trained procedure over millions of image-text pairs. Oscar [19] also uses additional supervision compared to our setting, in the form of object tags for each image.

Our default configuration employs the transformer mapping network, without fine-tuning the language model, denoted **Ours; Transformer**. Additionally, we also evaluate our variant that utilizes the MLP mapping network, and fine-tunes the language model, denoted **Ours; MLP + GPT2 tuning**. Other configurations are evaluated in Tab. 1(D).

Evaluation metrics. Similar to Li et al. [19], we validate our results over the COCO dataset using the common metrics BLEU [27], METEOR [10], CIDEr [37] and SPICE [3], and for the nocaps dataset using CIDEr and SPICE. For the Conceptual Captions, we report the ROUGE-L [21], CIDEr, and SPICE, as suggested by the authors [33].

Furthermore, we measure the training time and the number of trainable parameters to validate the applicability of our method. Reducing the training time allows to quickly obtain a new model for new data, create an ensemble of models, and decrease energy consumption. Similar to other works, we report training time in GPU hours, and the GPU model used. The number of trainable parameters is a popular measure to indicate model feasibility.

Quantitative evaluation. Quantitative results for the challenging Conceptual Captions dataset are presented in Tab. 1(A). As can be seen, we surpass the results of VLP, while requiring orders of magnitude less training time. We note that our lightweight model, which does not fine-tune GPT-2, achieves an inferior result for this dataset. We hypothesize that due to the large variety of styles, a more expressive model is required than our light model, which induces a significantly lower parameter count. We compare only to VLP, as the other baselines haven't published results nor trained models for this dataset.

Tab. 1(B) presents results for the nocaps dataset, where we achieve comparable results to the state-of-the-art method Oscar. As can be seen, Oscar achieves a slightly better SPICE score and we attain a slightly better CIDEr score. Still, our method uses a fraction of training time and trainable parameters with no additional object tags required, hence it is much more useful in practice.



Sự thật cơ bản	Một cuộc đời trong nhiếp ảnh - trong những bức ảnh.	Bức ảnh chụp biển báo đang được sửa chữa bởi một người dùng cảm.	Quả địa cầu: hình ảnh 3 chiều màu xanh lá cây đang cầm quả địa cầu trên tay.	Người chơi đang chăm chú nhìn vào máy tính màn hình.	Ngôi nhà đã có phòng ngủ đủ sức chứa mọi người.
VLP	Các diễn viên trong một cảnh phim.	Biển báo ở lối vào. Mẫu: nhân vật hoạt hình màu xanh lá cây cầm quả địa cầu.	Người này đang làm việc trên một video.	Phòng ngủ chính có một giường cờ King và một giường cờ Queen.	
Của chúng tôi; MLP + Điều chỉnh GPT2	Diễn viên ngồi trong khách sạn phòng.	Biển báo ở lối vào. Hình ảnh 3D về một người đàn ông đang cầm quả địa cầu.	Một người, một học sinh, đang xem video trên máy tính xách tay của mình.	Bất động sản này đang được rao bán với giá 1 bảng Anh.	

Của chúng tôi;	Máy biến áp	người ngồi trên ghế một dấu hiệu được nhìn thấy ở en-trong một căn phòng.	hình ảnh có sẵn về một người đàn ông chân dung một cậu bé đang cầm trái đất, đang chơi trò chơi điện tử.	một trong những phòng ngủ trong nhà đã được chuyển đổi thành một cuộc sống phòng.
----------------	-------------	---	--	---

Hình 4. Kết quả chưa được kiểm duyệt của năm hình ảnh đầu tiên trong bộ thử nghiệm của chúng tôi dành cho Chủ thích khái niệm [33].



Số liệu đánh giá. Tương tự như Li et al. [19], chúng tôi xác thực kết quả của mình trên tập dữ liệu COCO bằng cách sử dụng các số liệu chung BLEU [27], METEOR [10], CIDEr [37] và SPICE [3], và đối với tập dữ liệu nocaps bằng cách sử dụng CIDEr và SPICE. Đối với Chủ thích khái niệm, chúng tôi báo cáo ROUGE-L [21], CIDEr và SPICE, theo đề xuất của tác giả [33].

Hơn nữa, chúng tôi đo thời gian đào tạo và số lượng tham số có thể đào tạo để xác thực tính khả thi của phương pháp của chúng tôi. Giảm thời gian đào tạo cho phép nhanh chóng có được một mô hình mới cho dữ liệu mới, tạo ra một tập hợp các mô hình và giảm mức tiêu thụ năng lượng. Tương tự như các công trình khác, chúng tôi báo cáo thời gian đào tạo theo giờ GPU và mô hình GPU được sử dụng. Số lượng tham số có thể đào tạo là một biện pháp phổ biến để chỉ ra tính khả thi của mô hình.

Hình 5. Kết quả trên ảnh chụp bằng điện thoại thông minh. Trên cùng: sử dụng mô hình Chủ thích khái niệm của chúng tôi. Dưới cùng: mô hình COCO. Như đã chứng minh, cách tiếp cận của chúng tôi tổng quát hóa tốt đối với những hình ảnh mới chụp.

chủ yếu là BERT [11]. Cả VLP và Oscar đều khai thác một quy trình được đào tạo trước mở rộng trên hàng triệu cặp hình ảnh-văn bản. Oscar [19] cũng sử dụng sự giám sát bổ sung so với thiết lập của chúng tôi, dưới dạng thẻ đối tượng cho mỗi hình ảnh.

Cấu hình mặc định của chúng tôi sử dụng mạng map-ping biến áp, không tinh chỉnh mô hình ngôn ngữ, được ký hiệu là Ours; Transformer. Ngoài ra, chúng tôi cũng đánh giá biến thể của mình sử dụng mạng ánh xạ MLP và tinh chỉnh mô hình ngôn ngữ, được ký hiệu là Ours; MLP + điều chỉnh GPT2. Các cấu hình khác được đánh giá trong Tab. 1(D).

Bảng 1(B) trình bày kết quả cho tập dữ liệu nocaps, trong đó chúng tôi đạt được kết quả tương đương với phương pháp tiên tiến Oscar. Như có thể thấy, Oscar đạt được điểm SPICE tốt hơn một chút và chúng tôi đạt được điểm CIDEr tốt hơn một chút. Tuy nhiên, phương pháp của chúng tôi sử dụng một phần nhỏ thời gian đào tạo và các tham số có thể đào tạo mà không cần thêm thẻ đối tượng, do đó, nó hữu ích hơn nhiều trong thực tế.

Figure 6. Prefix Interpretability. We present both the generated caption and our prefix interpretation. Upper: Ours; MLP + GPT2 tuning. Bottom: Ours; Transformer.

Tab. 1(C) present the results for the COCO dataset. Oscar reaches the best results, however, it uses additional input in the form of object tags. Our results are closed to VLP and BUTD which utilize considerably more parameters and training time. Note that the training time of VLP and Oscar does not include the pre-training step. For instance, pre-training of VLP requires training over Conceptual Captions which consumes 1200 GPU hours.

Both Conceptual Captions and nocaps are designed to model a larger variety of visual concepts than COCO. Therefore, we conclude our method is preferable for generalizing to diverse data using a quick training procedure. This originates from utilizing the already rich semantic representations of both CLIP and GPT-2.

Qualitative evaluation. Visual results of the uncurated first examples in our test sets of both Conceptual Captions and COCO datasets are presented in Figs. 3 and 4 respectively. As can be seen, our generated captions are meaningful and depict the image successfully for both datasets. We present additional examples collected from the web in Fig. 1. As can be seen, our Conceptual Captions model generalizes well to arbitrary unseen images as it was trained over a sizable and diverse set of images. We also

demonstrate generalization to new scenarios. Moreover, our model successfully identifies uncommon objects even when trained only over COCO. For example, our method recognizes the wooden spoons or the cake with a candle better than Oscar in Fig. 3, since CLIP is pre-trained over a diverse set of images. However, our method still fails in some cases, such as recognizing the bicycle next to the train in Fig. 3. This is inherited from the CLIP model, which does

				
Chú thích một chiếc xe máy đang được trưng bày trong phòng trưng bày.	một nhóm người đang ngồi quanh một cái bàn.	một phòng khách đầy đồ nội thất và giá sách chưa đầy sách.	một vòi cứu hỏa đang ở trong giữa phố.	tủ trưng bày đầy rất nhiều loại khác nhau bánh rán.
Tiền tố com showcase xe máy-cle Một tia xe máy-đặt ra những gì đánh bóng Mục	món ăn chay của người tóc vàng ăn uống mong đợi mỉm cười nhóm bạn bè gần như	ghế sofa tt ghế Bart sách cửa ra vào hiện đại phòng ngủ	phố neon Da alley putis-tan đêm đầy màu sắc	hộp đựng bánh mì thủy tinh trưng bày bánh sandwich2 In mứt
Chú thích chiếc xe máy đang được trưng bày tại một buổi triển lãm.	một nhóm người đang ngồi cùng ngồi chung một bàn.	một phòng khách với một ghế sofa và giá sách.	một vòi cứu hỏa ở phía trước phố thành phố.	tủ trưng bày đầy đủ các loại bánh rán khác nhau.
Tiền tố cover voi Sniper^ A^ A^ A^ sự kết hợp xe máy chắc chắn đạt được\n	amic Con voi ngon SukActionCode nghiệp ánh gia có thể hoán đổi cho nhau không thể phủ nhận đạt được	orianclassic con voi ~ Cameroon^ A^ A^ Aroom sự kết hợp án ý đạt được thành công đáng kính ngạc\n	ockets Đê con voi Sniper^ A^ A^ Một chiếc xe đạp sự hiệp lực không thể phủ nhận\da đạt được\n	món tráng miệng đậm phòng ele-phbm^ A^ A^ A^ Một món ăn gặt kinh ngạc\n

Hình 6. Khả năng diễn giải tiền tố. Chúng tôi trình bày cả chủ thích được tạo ra và diễn giải tiền tố của chúng tôi. Phía trên: Của chúng tôi; Điều chỉnh MLP + GPT2. Bên dưới: Của chúng tôi; Máy biến áp.

Tab. 1(C) trình bày kết quả cho tập dữ liệu COCO. Os-car đạt được kết quả tốt nhất, tuy nhiên, nó sử dụng thêm đầu vào dưới dạng thẻ đối tượng. Kết quả của chúng tôi gần với VLP và BUTD sử dụng nhiều tham số hơn đáng kể và thời gian đào tạo. Lưu ý rằng thời gian đào tạo của VLP và Oscar không bao gồm bước đào tạo trước. Ví dụ, đào tạo trước của VLP yêu cầu đào tạo qua Chú thích khái niệm tiêu tốn 1200 giờ GPU.

Cả chủ thích khái niệm và nocaps đều được thiết kế để tạo ra nhiều khái niệm trực quan đa dạng hơn COCO. Do đó, chúng tôi kết luận rằng phương pháp của chúng tôi thích hợp hơn để tổng quát hóa dữ liệu đa dạng bằng cách sử dụng quy trình đào tạo nhanh. Điều này bắt nguồn từ việc sử dụng các biểu diễn ngũ nghĩa phong phú của cả CLIP và GPT-2.

Đánh giá định tính. Kết quả trực quan của các ví dụ đầu tiên trong bộ thử nghiệm của chúng tôi về cả Cap-tions khái niệm và bộ dữ liệu COCO được trình bày trong Hình 3 và

4 tương ứng. Như có thể thấy, chú thích được tạo ra của chúng tôi có ý nghĩa và mô tả hình ảnh thành công cho cả hai bộ dữ liệu. Chúng tôi trình bày các ví dụ bổ sung được thu thập từ trang web trong Hình 1. Như có thể thấy, Chú thích khái niệm của chúng tôi mô hình tổng quát hóa tốt với các hình ảnh không nhìn thấy tùy ý như nó đã từng được đào tạo trên một tập hợp hình ảnh đa dạng và có kích thước lớn. Chúng tôi cũng hiện diện trong Hình 5 kết quả trên hình ảnh điện thoại thông minh, để tiếp tục chứng minh sự khái quát hóa cho các kịch bản mới. Hơn nữa, mô hình xác định thành công các đối tượng không phổ biến ngay cả khi chỉ được đào tạo qua COCO. Ví dụ, phương pháp của chúng tôi nhận dạng thia gõ hoặc bánh có nén tốt hơn

hơn Oscar trong Hình 3, vì CLIP được đào tạo trước trên một tập hợp hình ảnh đa dạng. Tuy nhiên, phương pháp của chúng tôi vẫn thất bại trong một số các trường hợp, chẳng hạn như nhận ra chiếc xe đẹp bên cạnh tàu hỏa trong Hình 3. Điều này được kể thừa từ mô hình CLIP, mô hình này

hông nhận ra chiếc xe đẹp ngay từ đầu. Chúng tôi kết luận rằng hình của chúng tôi sẽ được hưởng lợi từ việc cải thiện khả năng phát hiện đối tượng CLIP, nhưng hãy để hướng này cho công việc trong tương lai. Đối với chúng thích khai niêm, phương pháp của chúng tôi chủ yếu tạo ra chính xác chú thích, chẳng hạn như nhận thức người 3D màu xanh lá cây trong Hình 4. Nhìn dự kiến, phương pháp của chúng tôi vẫn bị ảnh hưởng bởi sự thiên vị dữ liệu. Ví dụ, nó mô tả hình ảnh phòng ngủ trong Hình 4 là "Bất động sản này đang được rao bán với giá 1 bảng Anh" sau khi chúng kiểm những chú thích như vậy. Ở quảng cáo bất động sản trong quá trình đào tạo.

inh chính mô hình ngôn ngữ. Như đã mô tả trong Phần 3, việc tinh chỉnh mô hình ngôn ngữ dẫn đến một mô hình biểu đạt hơn nhiều, nhưng cũng dễ bị quá khớp hơn khi số lượng tham số có thể đảo tạo tăng lên.

đó thể được nhìn thấy trong Tab. 1, hai biến thể – có và không có hình ngón ngữ tinh chỉnh – có thể so sánh được. Trên dữ liệu chú thích khái niệm cực kỳ phức tạp, chúng tôi nhận được kết quả vượt trội với việc tinh chỉnh. Trong khi trên tập dữ liệu COCO không biến, việc tránh tinh chỉnh đạt được kết quả tốt hơn kết quả. Về tập dữ liệu nocaps, kết quả là gần đúng bằng nhau, do đó mô hình nhẹ hơn sẽ được ưu chuộng hơn. Chúng tôi cho giả thuyết rằng các tập dữ liệu cực kỳ phức tạp hoặc những tập dữ liệu hiển hiện một phong cách đặc đáo đòi hỏi sự biểu cảm hơn, và do khả năng được hưởng lợi từ việc điều chỉnh chính xác càng cao.

tiền tố Khả năng diễn giải. Để hiểu rõ hơn
nhưngh pháp và kết quả, chúng tôi đề xuất giải thích các kết quả được tạo ra
tiền tố như một chuỗi các từ. Vì tiền tố và từ
nhưng chia sẻ cùng một không gian tiệm ẩn, chúng có thể được xử lý
tương tự như vậy. Chúng tôi xác định cách giải thích của mỗi k
hông tiền tố như là mã thông báo từ vựng gần nhất, bên dưới
để tương tự cosin. Hình 6 cho thấy các ví dụ về hình ảnh,
gọ ra các chủ thích và cách diễn giải tiền tố của chúng.

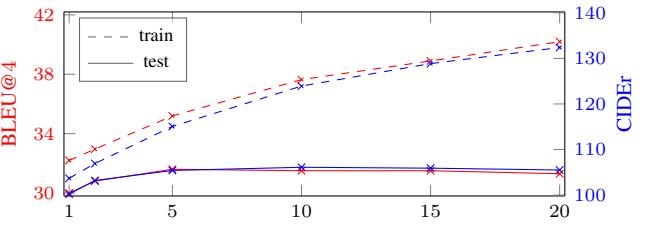
interpretation is meaningful when both the mapping network and GPT-2 are trained. In this case, the interpretation contains salient words that associate with the content of the image. For instance, motorcycle and showcase in the first example. However, when we only train the mapping network, the interpretation becomes essentially unreadable since the network is also charged with maneuvering the fixed language model. Indeed, a considerable part of the prefix embeddings is shared across different images for the same model, as it performs the same adjustment to GPT-2.

Prefix length. Li and Liang [20] showed that increasing the size of the prefix length, up to a certain value, improves the performance of the model in an underlying task. Moreover, the saturation length might differ between tasks. For the image captioning task, we conduct an ablation over the prefix lengths using the COCO dataset over two configurations of our method: **Ours; Transformer** and **Ours; MLP + GPT2 tuning**. The results are summarized in Fig. 7. For each prefix size and configuration, we train the network for 5 epochs and report the BLEU@4 and CIDEr scores over the test and train sets.

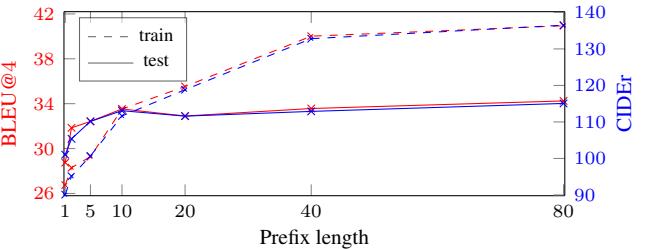
As can be seen in Fig. 7a, increasing the prefix size while allowing tuning of the language model results in overfitting to the training set, due to the large number of trainable parameters. However, when the language model is frozen, we experience improvement for both the training and test evaluations, as can be seen in Fig. 7b. Naturally, extremely small prefix length yields inferior results as the model is not expressive enough. In addition, we point out that the MLP architecture is inherently more limited as it is not scalable for a long prefix. For example, a prefix size of 40 implies a network with over 450M parameters, which is unfeasible for our single GPU setting. The transformer architecture allows increasing the prefix size with only marginal increment to the number of the parameters, but only up to 80 — due to the quadratic memory cost of the attention mechanism.

Mapping network. An ablation study for the mapping network architecture is shown in Tab. 1(C),(D). As can be seen, with language model fine-tuning, the MLP achieves better results. However, the transformer is superior when the language model is frozen. We conclude that when employing the fine-tuning of the language model, the expressive power of the transformer architecture is unnecessary.

Implementation details. We used the prefix length of $K = 10$ for the MLP mapping networks, where the MLP contains a single hidden layer. For the transformer mapping network, we set the CLIP embedding to $K = 10$ constants tokens and use 8 multi-head self-attention layers with 8 heads each. We train for 10 epochs using a batch size of 40. For optimization, we use AdamW [18] with weight



(a) MLP mapping network with fine-tuning of the language model.



(b) Transformer mapping network with frozen language model.

Figure 7. Effect of the prefix length on the captioning performance over the COCO-captions dataset. For each prefix length, we report the BLEU@4 (red) and CIDEr (blue) scores over the test and train sets.

decay fix as introduced by Loshchilov et al. [24], with a learning rate of $2e^{-5}$ and 5000 warm-up steps. For GPT-2 we employ the implementation of Wolf et al. [41].

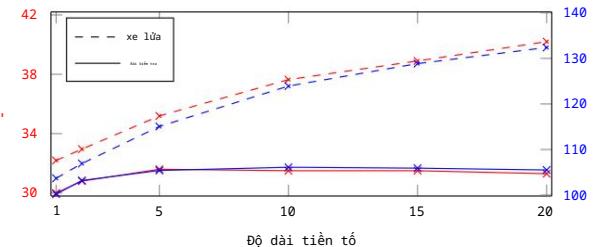
5. Conclusion

Overall, our CLIP-based image-captioning method is simple to use, doesn't require any additional annotations, and is faster to train. Even though we propose a simpler model, it demonstrates more merit as the dataset becomes richer and more diverse. We consider our approach as part of a new image captioning paradigm, concentrating on leveraging existing models, while only training a minimal mapping network. This approach essentially learns to adapt existing semantic understanding of the pre-trained models to the style of the target dataset, instead of learning new semantic entities. We believe the utilization of these powerful pre-trained models would gain traction in the near future. Therefore, the understanding of how to harness these components is of great interest. For future work, we plan to incorporate pre-trained models (e.g., CLIP), to other challenging tasks, such as visual question answering or image to 3D translation, through the utilization of mapping networks.

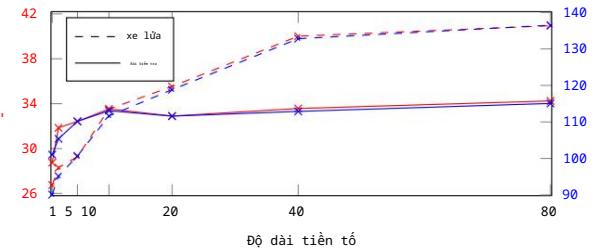
References

- [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 caption-

diễn giải có ý nghĩa khi cả mạng lưới ánh xạ và GPT-2 đều được đào tạo. Trong trường hợp này, diễn giải chứa các từ nỗi bật liên quan đến nội dung của hình ảnh. Ví dụ, xe máy và showcase trong ví dụ đầu tiên. Tuy nhiên, khi chúng ta chỉ đào tạo mạng lưới ánh xạ, diễn giải về cơ bản trở nên không thể đọc được vì mạng lưới cũng chưa giao nhiệm vụ điều khiển mô hình ngôn ngữ cố định. Thật vậy, một phần đáng kể các nhung tiền tố được chia sẻ trên các hình ảnh khác nhau cho cùng một mô hình, vì nó thực hiện cùng một điều chỉnh đối với GPT-2.



(a) Mạng ánh xạ MLP với sự tinh chỉnh của mô hình ngôn ngữ.



(b) Mạng ánh xạ biến đổi với mô hình ngôn ngữ đóng băng.

Độ dài tiền tố. Li và Liang [20] đã chỉ ra rằng việc tăng kích thước của độ dài tiền tố, lên đến một giá trị nhất định, sẽ cải thiện hiệu suất của mô hình trong một tác vụ cơ bản. Hơn nữa, độ dài bao hàm có thể khác nhau giữa các tác vụ. Đối với tác vụ chú thích hình ảnh, chúng tôi tiến hành cắt bỏ độ dài tiền tố bằng cách sử dụng tập dữ liệu COCO qua hai câu hỏi của phương pháp của chúng tôi: Ours; Transformer và Ours; điều chỉnh MLP + GPT2. Kết quả được tóm tắt trong Hình 7. Đối với mỗi kích thước và câu hỏi tiền tố, chúng tôi đào tạo mạng trong 5 kỳ nguyên và báo cáo điểm BLEU@4 và CIDEr qua các tập kiểm tra và đào tạo.

Như có thể thấy trong Hình 7a, việc tăng kích thước tiền tố trong khi cho phép điều chỉnh mô hình ngôn ngữ dẫn đến quá khớp với tập huấn luyện, do số lượng lớn các tham số có thể huấn luyện. Tuy nhiên, khi mô hình ngôn ngữ bị đóng băng, chúng ta thấy sự cải thiện cho cả đánh giá huấn luyện và kiểm tra, như có thể thấy trong Hình 7b. Đường nhiên, độ dài tiền tố cực nhỏ mang lại kết quả kém hơn vì mô hình không đủ biểu đạt. Ngoài ra, chúng tôi chỉ ra rằng kiến trúc MLP vốn bị hạn chế hơn vì nó không thể mở rộng quy mô cho một tiền tố dài. Ví dụ: kích thước tiền tố là 40 ngay cả một mạng có hơn 450M tham số, điều này không khả thi đối với thiết lập GPU đơn của chúng tôi. Kiến trúc máy biến áp cho phép tăng kích thước tiền tố chỉ với mức tăng biến cho số lượng tham số, nhưng chỉ tối đa là 80 — do chi phí bộ nhớ bậc hai của cơ chế chia.

Mạng lập bản đồ. Một nghiên cứu cắt bỏ cho kiến trúc mạng lập bản đồ được thể hiện trong Tab. 1(C),(D). Như có thể thấy, với việc tinh chỉnh mô hình ngôn ngữ, MLP đạt được kết quả tốt hơn. Tuy nhiên, bộ biến đổi vượt trội hơn khi mô hình ngôn ngữ bị đóng băng. Chúng tôi kết luận rằng khi sử dụng tinh chỉnh mô hình ngôn ngữ, sức mạnh biểu đạt của kiến trúc bộ biến đổi là không cần thiết.

Chi tiết triển khai. Chúng tôi đã sử dụng độ dài tiền tố $K = 10$ cho các mạng ánh xạ MLP, trong đó MLP chứa một lớp ẩn duy nhất. Đối với mạng mapping biến áp, chúng tôi đặt những CLIP thành $K = 10$ mà thông báo hàng số và sử dụng 8 lớp tự chú ý nhiều đầu với 8 đầu mỗi lớp. Chúng tôi đào tạo trong 10 kỳ nguyên bằng cách sử dụng kích thước lô là 40. Để tối ưu hóa, chúng tôi sử dụng AdamW [18] với trọng số

sửa lỗi phân rã như được giới thiệu bởi Loshchilov et al. [24], với tốc độ học là $2e^{-5}$ và 5000 bước khởi động. Đối với GPT-2, chúng tôi sử dụng cách triển khai của Wolf et al. [41].

5. Kết luận

Nhìn chung, phương pháp chú thích hình ảnh dựa trên CLIP của chúng tôi dễ sử dụng, không yêu cầu bắt kỳ chú thích bổ sung nào và đào tạo nhanh hơn. Mặc dù chúng tôi đề xuất một mô hình đơn giản hơn, nhưng nó chứng minh được nhiều giá trị hơn khi tập dữ liệu trờ nên phong phú và đa dạng hơn. Chúng tôi coi cách tiếp cận của mình là một phần của mô hình chú thích hình ảnh mới, tập trung vào việc tận dụng các mô hình hiện có, trong khi chỉ đào tạo một mạng lưới ánh xạ tối thiểu. Về cơ bản, cách tiếp cận này học cách điều chỉnh hiểu ngữ nghĩa hiện có của các mô hình được đào tạo trước theo phong cách của tập dữ liệu mục tiêu, thay vì học các thực thể ngữ nghĩa mới. Chúng tôi tin rằng việc sử dụng các mô hình được đào tạo trước mạnh mẽ này sẽ đạt được sức hút trong tương lai gần.

Do đó, việc hiểu cách khai thác các thành phần này rất được quan tâm. Đối với công việc trong tương lai, chúng tôi có kế hoạch kết hợp các mô hình được đào tạo trước (ví dụ: CLIP) vào các nhiệm vụ đầy thách thức khác, chẳng hạn như trả lời câu hỏi trực quan hoặc chuyển đổi hình ảnh sang 3D, thông qua việc sử dụng mạng lưới ánh xạ lập bản đồ.

Tài liệu tham khảo

- [1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: chú thích đối tượng mới la-

- ing at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957, 2019. 2, 4
- [2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. *arXiv preprint arXiv:1612.00576*, 2016. 4
- [3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016. 6
- [4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 1, 3, 4, 5
- [5] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021. 2
- [6] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667, 2017. 3
- [7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2, 3, 4
- [8] Xinpeng Chen, Lin Ma, Wenhao Jiang, Jian Yao, and Wei Liu. Regularizing rnns for caption generation by reconstructing the past with the present. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7995–8003, 2018. 3
- [9] Xinlei Chen and C Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv:1411.5654*, 2014. 1, 3
- [10] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 6
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 6
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3
- [13] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015. 1, 3
- [14] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021. 2
- [15] Junlong Gao, Shiqi Wang, Shanshe Wang, Siwei Ma, and Wen Gao. Self-critical n-step training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6300–6308, 2019. 3
- [16] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *arXiv preprint arXiv:1906.05963*, 2019. 3
- [17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 4, 5
- [18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 8
- [19] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 1, 3, 4, 5, 6
- [20] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 2, 4, 8
- [21] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, 2004. 6
- [22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 3, 4
- [23] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804*, 2021. 3
- [24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 8
- [25] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 3
- [26] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-level collaborative transformer for image captioning. *arXiv preprint arXiv:2101.06462*, 2021. 3
- [27] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6
- [28] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 2
- ở quy mô lớn. Trong Biên bản báo cáo của IEEE/CVF International Hội nghị về Tâm nhìn máy tính, trang 8948–8957, 2019. 2, 4
- [2] Peter Anderson, Basura Fernando, Mark Johnson và Stephen Gould. Chú thích hình ảnh từ vựng mở có hướng dẫn với tìm kiếm chùm tia bị hạn chế. Bản in trước arXiv arXiv:1612.00576, 2016. 4
- [3] 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 tâm nhìn máy tính, trang 382–398. Springer, 2016. 6
- [4] 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 tâm nhìn máy tính và nhận dạng mẫu, các trang 6077–6086, 2018. 1, 3, 4, 5
- [5] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva và Antonio Torralba. Về bối cảnh. Bản in trước arXiv arXiv:2103.10951, 2021. 2
- [6] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu và Tat-Seng Chua. Sca-cnn: Không gian và sự chú ý theo từng kênh trong mạng tích chập để quan tâm chú thích độ tuổi. Trong Biên bản báo cáo của hội nghị IEEE về tâm nhìn máy tính và nhận dạng mẫu, trang 5659–5667, 2017. 3
- [7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár và C Lawrence Zitnick. Tiêu đề Microsoft coco: Thu thập và đánh giá dữ liệu máy chủ. Bản in trước arXiv arXiv:1504.00325, 2015. 2, 3, 4
- [8] Xinpeng Chen, Lin Ma, Wenhao Jiang, Jian Yao và Wei Liu. Chuẩn hóa rnn để tạo chú thích bằng cách tái tạo quá khứ với hiện tại. Trong Biên bản của IEEE Hội nghị về Thị giác máy tính và Nhận dạng mẫu, trang 7995–8003, 2018. 3
- [9] Xinlei Chen và C Lawrence Zitnick. Học một biểu diễn trực quan để tạo chú thích hình ảnh. arXiv bản in trước arXiv:1411.5654, 2014. 1, 3
- [10] Michael Denkowski và Alon Lavie. Meteor universal: Đánh giá bản dịch theo ngôn ngữ cụ thể cho bất kỳ ngôn ngữ đích nào. Trong Biên bản hội thảo lần thứ chín về máy thông kê bản dịch, trang 376–380, 2014. 6
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee và Kristina Tόng cộng. Bert: Đào tạo trước về song hướng sâu máy biến áp để hiểu ngôn ngữ. Bản in trước arXiv arXiv:1810.04805, 2018. 3, 6
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. Một hình ảnh có giá trị bằng 16x16 từ: Trans-formers để nhận dạng hình ảnh ở quy mô lớn. Bản in trước arXiv arXiv:2010.11929, 2020. 3
- [13] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Sri-vastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. Từ chú thích đến các khái niệm trực quan và ngược lại. Trong Biên bản báo cáo của hội nghị IEEE về tâm nhìn máy tính và nhận dạng mẫu, các trang 1473–1482, 2015. 1, 3
- [14] Rinon Gal, Hoặc Patashnik, Haggai Maron, Gal Chechik, và Daniel Cohen-Or. Stylegan-nada: Chuyển thể doomain được hướng dẫn bằng clip của trình tạo hình ảnh. Bản in trước arXiv arXiv:2108.00946, 2021. 2
- [15] Junlong Gao, Shiqi Wang, Shanshe Wang, Siwei Ma, và Wen Gao. Đào tạo n-step tự phân để chú thích hình ảnh. Trong Biên bản Hội nghị IEEE về Máy tính Tâm nhìn và Nhận dạng Mẫu, trang 6300–6308, 2019. 3
- [16] Simao Herdade, Armin Kappeler, Kofi Boakye và Joao Soares. Chú thích hình ảnh: Biến đổi các đối tượng thành từ ngữ. Bản in trước arXiv arXiv:1906.05963, 2019. 3
- [17] 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 tâm nhìn máy tính và nhận dạng mẫu, trang 3128–3137, 2015. 4, 5
- [18] Diederik P. Kingma và Jimmy Ba. Adam: Một phương pháp tối ưu hóa ngẫu nhiên. CoRR, abs/1412.6980, 2015. 8
- [19] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Đổi tượng-ngữ nghĩa được căn chỉnh trước khi đào tạo cho nhiệm vụ ngôn ngữ thị giác. Trong Hội nghị máy tính châu Âu Tâm nhìn, trang 121–137. Springer, 2020. 1, 3, 4, 5, 6
- [20] Xiang Lisa Li và Percy Liang. Điều chỉnh tiền tố: Tối ưu hóa các lời nhắc liên tục để tạo ra. Bản in trước arXiv arXiv:2101.00190, 2021. 2, 4, 8
- [21] Chin-Yew Lin và Franz Josef Och. Đánh giá tự động chất lượng dịch máy sử dụng chuỗi con chung dài nhất và thống kê bỏ qua bigram. Trong Biên bản báo cáo Cuộc họp thường niên lần thứ 42 của Hiệp hội tính toán Ngôn ngữ học (ACL-04), trang 605–612, 2004. 6
- [22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár và C Lawrence Zitnick. Microsoft coco: Các đổi tượng phổ biến trong ngữ cảnh. Trong Hội nghị châu Âu về tâm nhìn máy tính, trang 740–755. Springer, 2014. 2, 3, 4
- [23] Ngụy Lưu, Sihan Chen, Longteng Guo, Xinxin Zhu, và Jing Liu. Cptr: Mạng biến áp đầy đủ để chú thích hình ảnh. Bản in trước arXiv arXiv:2101.10804, 2021. 3
- [24] Ilya Loshchilov và Frank Hutter. Giảm trọng lượng tách rời chính quy hóa. Bản in trước arXiv arXiv:1711.05101, 2017. 8
- [25] Jiasen Lu, Dhruv Batra, Devi Parikh và Stefan Lee. Vilbert: Đào tạo trước các biểu diễn ngôn ngữ thị giác không phụ thuộc vào nhiệm vụ cho các nhiệm vụ thị giác và ngôn ngữ. Bản in trước của arXiv arXiv:1908.02265, 2019. 3
- [26] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin và Rongrong Ji. Bộ chuyển đổi công tác hai cấp để chú thích hình ảnh. Bản in trước arXiv arXiv:2101.06462, 2021. 3
- [27] Kishore Papineni, Salim Roukos, Todd Ward và Wei-Jing Zhu. Bleu: một phương pháp đánh giá tự động máy móc bản dịch. Trong Biên bản cuộc họp thường niên lần thứ 40 của Hiệp hội Ngôn ngữ học tính toán, trang 311–318, 2002. 6
- [28] Hoặc Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, và Dani Lischinski. Styleclip: Thảo tác điều khiển bằng văn bản của hình ảnh stylegan. Trong Biên bản báo cáo của Hội nghị quốc tế IEEE/CVF về Tâm nhìn máy tính, trang 2085–2094, 2021. 2

- [29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2, 3
- [30] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2, 3
- [31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 3, 4
- [32] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017. 3
- [33] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 2, 4, 6
- [34] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on image captioning. *arXiv preprint arXiv:2107.06912*, 2021. 1, 3
- [35] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 1, 3
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3, 4
- [37] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6
- [38] 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. 3
- [39] Yufei Wang, Zhe Lin, Xiaohui Shen, Scott Cohen, and Garrison W Cottrell. Skeleton key: Image captioning by skeleton-attribute decomposition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7272–7281, 2017. 3
- [40] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 3
- [41] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chau-mond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. 8
- [42] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 1, 3
- [43] Xu Yang, Hanwang Zhang, and Jianfei Cai. Learning to collocate neural modules for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4250–4260, 2019. 3
- [44] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018. 1
- [45] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M Hospedales. Actor-critic sequence training for image captioning. *arXiv preprint arXiv:1706.09601*, 2017. 3
- [46] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 3
- [47] Luwei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020. 1, 3, 4, 5
- [29] 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. Bản in trước arXiv arXiv:2103.00020, 2021. 2, 3
- [30] 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. 2, 3
- [31] Nhâm Thanh Thanh, Hà Minh Hà, Ross Girshick và Jian Sun. R-cnn nhanh hơn: Hướng tới phát hiện đối tượng theo thời gian thực với vùng mạng lưới dễ xuất. Những tiến bộ trong hệ thống xử lý thông tin thần kinh, 28:91–99, 2015. 3, 4
- [32] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross và Vaibhava Goel. Đào tạo trình tự phê bình cho chủ thích hình ảnh. Trong Biên bản báo cáo của hội nghị IEEE về tầm nhìn máy tính và nhận dạng mẫu, trang 7008–7024, 2017. 3
- [33] Piyush Sharma, Nan Ding, Sebastian Goodman và Radu Soricut. Chủ thích khái niệm: Một tập dữ liệu văn bản thay thế hình ảnh được làm sạch, có siêu ẩn danh, để tạo chủ thích hình ảnh tự động. Trong Biên bản báo cáo của Cuộc họp thường niên lần thứ 56 của Hiệp hội Ngôn ngữ học tính toán (Tập 1: Bài báo dài), trang 2556–2565, 2018. 2, 4, 6
- [34] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni và Rita Cucchiara. Từ điển thị dẻ kèm: Một cuộc khảo sát về chủ thích hình ảnh. Bản in trước arXiv arXiv:2107.06912, 2021. 1, 3
- [35] Hao Tan và Mohit Bansal. Lxmert: Học các biểu diễn mã hóa đa phương thức từ bộ biến đổi. arXiv bản in trước arXiv:1908.07490, 2019. 1, 3
- [36] 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. Trong những tiến bộ trong thần kinh hệ thống xử lý thông tin, trang 5998–6008, 2017. 3, 4
- [37] 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. 6
- [38] Oriol Vinyals, Alexander Toshev, Samy Bengio và Dumitru Erhan. Hiển thị và kèm: Một máy tạo chủ thích hình ảnh thần kinh. Trong Biên bản hội nghị IEEE về máy tính tầm nhìn và nhận dạng mẫu, trang 3156–3164, 2015. 3
- [39] Yufei Wang, Zhe Lin, Xiaohui Shen, Scott Cohen và Garrison W Cottrell. Khóa xương: Chủ thích hình ảnh bằng cách phân tích thuộc tính xương. Trong Biên bản báo cáo của hội nghị IEEE về thị giác máy tính và nhận dạng mẫu, các trang 7272–7281, 2017. 3
- [40] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov và Yuan Cao. Simvlm: Ngôn ngữ hình ảnh đơn giản mô hình đào tạo trước với sự giám sát yếu. Bản in trước arXiv arXiv:2108.10904, 2021. 3
- [41] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chau-mond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest và Alexander M. Rush. Trans-formers: Xử lý ngôn ngữ tự nhiên hiện đại. Trong Biên bản Hội nghị năm 2020 về Phương pháp Thực nghiệm trong Xử lý ngôn ngữ tự nhiên: Trình diễn hệ thống, trang 38–45, Trực tuyến, tháng 10 năm 2020. Hiệp hội Ngôn ngữ học tính toán. 8
- [42] Kelvin Xu, Jimmy Ba, Ryan Kiros, Cho Kyunghyun, Aaron Courville, Ruslan Salakhudinov, Rich Zemel và Yoshua Bengio. Hiển thị, tham dự và kể: Tạo chủ thích hình ảnh thần kinh với sự chú ý trực quan. Trong hội nghị quốc tế về học máy, trang 2048–2057. PMLR, 2015. 1, 3
- [43] Xu Yang, Hanwang Zhang và Jianfei Cai. Học cách định vị các mứ-dun thần kinh để chủ thích hình ảnh. Trong Biên bản của Hội nghị quốc tế về tầm nhìn máy tính của IEEE/CVF, trang 4250–4260, 2019. 3
- [44] Ting Yao, Yingwei Pan, Yehao Li và Tao Mei. Khám phá mối quan hệ trực quan cho chủ thích hình ảnh. Trong Biên bản Hội nghị châu Âu về tầm nhìn máy tính (ECCV), các trang 684–699, 2018. 1
- [45] Lý Chương, Lữ Tống, Phùng Lưu, Đào Tương, Thiếu Cường Gong, Yongxin Yang và Timothy M Hospedales. Đào tạo trình tự diễn viên-nhà phê bình cho chủ thích hình ảnh. Bản in trước arXiv arXiv:1706.09601, 2017. 3
- [46] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lý 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. 3
- [47] Luwei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso và Jianfeng Gao. Đào tạo trước ngôn ngữ thị giác nhất cho chủ thích hình ảnh và vqa. Trong Biên bản báo cáo Hội nghị AAAI về Trí tuệ nhân tạo, tập 34, trang 13041–13049, 2020. 1, 3, 4, 5