

I-TUNING: TUNING FROZEN LANGUAGE MODELS WITH IMAGE FOR LIGHTWEIGHT IMAGE CAPTIONING

Ziyang Luo^{*} Zhipeng Hu^{†‡} Yadong Xi[‡] Rongsheng Zhang[‡] Jing Ma^{*✉}

^{*} Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China

[†] College of Computer Science and Technology, Zhejiang University, Hangzhou, China

[‡] Fuxi AI Lab, NetEase Inc., Hangzhou, China

ABSTRACT

Image Captioning is a traditional vision-and-language task that aims to generate the language description of an image. Recent studies focus on scaling up the model size and the number of training data, which significantly increase the cost of model training. Different to these heavy-cost models, we introduce a lightweight image captioning framework (**I-Tuning**), which contains a small number of trainable parameters. We design a novel I-Tuning cross-attention module to connect the non-trainable pre-trained language decoder GPT2 and vision encoder CLIP-ViT. Since most parameters are not required to be updated during training, our framework is lightweight and fast. Experimental results conducted on three image captioning benchmarks reveal that our framework achieves comparable or better performance than the large-scale baseline systems. But our models contain up to **10 times fewer** trainable parameters and require much fewer data for training compared with state-of-the-art baselines.

Index Terms— Lightweight image captioning, Language models, Transformer, Cross-Modal

1. INTRODUCTION

Image Captioning is a critical task in the field of cross-modal, which focus on natural language generation to depict an image. Recent years have witnessed the success of applying large-scale pre-trained models on the task of image captioning, which generally scale up the number of trainable parameters and training data to achieve state-of-the-art performances [1, 2, 3, 4]. For example, a recent proposed OSCAR model [1] contains more than 135M trainable parameters and requires around 4M images during pre-training. Therefore, in spite of the performances, the heavy demands for extra computational resources and massive data for model training have become an urgent issue.

Recent studies showed that parameter-efficient pre-trained language models (PLMs) tuning [5] can effectively reduce the cost during training, where most parameters are frozen (i.e.,

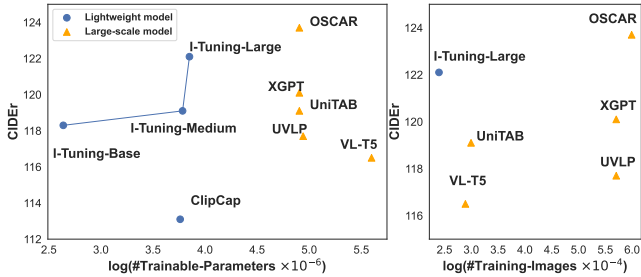


Fig. 1. Image captioning performances on MSCOCO dataset based on the lightweight and large-scale models, where I-Tuning-Base/-Medium/-Large respectively denote our model with base/medium/large number of trainable parameters.

not updated during training) and the rest small set are trainable. [6] recently introduced a ClipCap model that transforms images into fixed-length vectors and prompts a frozen GPT2 for image captioning. However, the learned vectors cannot capture accurate visual information to enhance the caption generation. To overcome these shortcomings, we propose a novel lightweight image captioning framework (**I-Tuning**) to alleviate the cost in terms of computational resource and training data. We design an I-Tuning module to connect the pre-trained vision encoder (i.e., CLIP-ViT [7]) and the language decoder (i.e., GPT2 [8]). To align between the language and vision modals, it serves as a cross-modal filter that automatically picks the visual information from the output of the vision encoder and adjusts the output hidden states of the language decoder. During training, we only update the newly introduced parameters in the I-Tuning module, and the parameters of the two pre-trained models are frozen.

Figure 1 exemplifies the CIDEr scores of our lightweight models and large-scale baselines. In terms of model training, our basic model I-Tuning-Base only contains around 14M trainable parameters, namely **10 times fewer** than the other large-scale models such as OSCAR. In terms of data, even our I-Tuning-Large model can achieve comparable performances with relatively less training data. We evaluate our proposed framework on 3 image captioning benchmarks (i.e.,

I-TUNING: ĐIỀU CHỈNH CÁC MÔ HÌNH NGÔN NGỮ ĐÔNG LẠNH VỚI HÌNH ẢNH CHO CHÚ THÍCH HÌNH ẢNH NHẸ

Tử Dư ơ ng La Hồ Chí Bằng†‡ Yadong Xi‡ Trư ơ ng Dung Thắng‡ Tĩnh Mã

Khoa Khoa học Máy tính, Đại học Baptist Hồng Kông, Khu hành chính đặc biệt Hồng Kông, Trung Quốc

† Khoa Khoa học Máy tính và Công nghệ, Đại học Chiết Giang, Hàng Châu, Trung Quốc

‡ Phòng thí nghiệm AI Fuxi, NetEase Inc., Hàng Châu, Trung Quốc

TÓM TẮT

Chú thích hình ảnh là một tầm nhìn và ngôn ngữ truyền thống nhiệm vụ nhằm mục đích tạo ra mô tả ngôn ngữ của một hình ảnh. Các nghiên cứu gần đây tập trung vào việc mở rộng quy mô mô hình và số lượng dữ liệu đào tạo, làm tăng đáng kể chi phí đào tạo mô hình. Khác với các mô hình tốn kém này, chúng tôi giới thiệu một khuôn khổ chú thích hình ảnh nhẹ (I-Tuning), chứa một số lượng nhỏ các tham số có thể đào tạo. Chúng tôi thiết kế một mô-đun chú ý chéo I-Tuning mới để kết nối bộ giải mã ngôn ngữ được đào tạo trước không thể đào tạo được GPT2 và bộ mã hóa tầm nhìn CLIP-ViT. Vì hầu hết các thông số không bắt buộc phải cập nhật trong quá trình đào tạo, khuôn khổ của chúng tôi nhẹ và nhanh. Kết quả thử nghiệm được tiến hành trên ba tiêu chuẩn chú thích hình ảnh cho thấy khung của chúng tôi đạt được hiệu suất tương đương hoặc tốt hơn hệ thống cơ sở quy mô lớn. Nhưng các mô hình của chúng tôi chứa tối đa hơn 10 lần các thông số có thể đào tạo và yêu cầu ít hơn nhiều dữ liệu đào tạo được so sánh với dữ liệu cơ sở hiện đại.

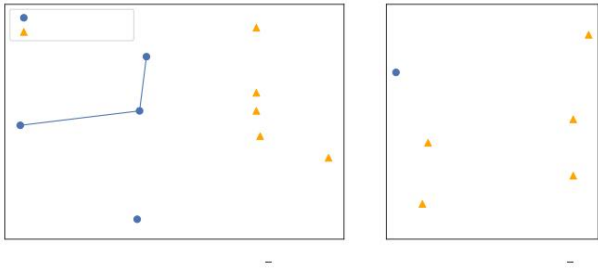
Thuật ngữ chỉ mục— Chú thích hình ảnh nhẹ, Ngôn ngữ mô hình, Biến áp, Đa phương thức

1. GIỚI THIỆU

Chú thích hình ảnh là một nhiệm vụ quan trọng trong lĩnh vực đa phương thức, tập trung vào việc tạo ra ngôn ngữ tự nhiên để mô tả một hình ảnh. Những năm gần đây đã chứng kiến sự thành công của việc áp dụng các mô hình được đào tạo trước quy mô lớn về nhiệm vụ chú thích hình ảnh, thường mở rộng số lượng tham số có thể đào tạo và dữ liệu đào tạo để đạt được hiệu suất tiên tiến [1, 2, 3, 4]. Ví dụ, một OSCAR được đề xuất gần đây mô hình [1] chứa hơn 135M tham số có thể đào tạo và yêu cầu khoảng 4M hình ảnh trong quá trình đào tạo trước. Do đó, trong bất chấp hiệu suất, nhu cầu lớn về các nguồn lực tính toán bổ sung và dữ liệu khổng lồ để đào tạo mô hình đã trở thành vấn đề cấp bách.

Các nghiên cứu gần đây cho thấy rằng các tham số được đào tạo trước hiệu quả việc điều chỉnh các mô hình ngôn ngữ (PLM) [5] có thể làm giảm hiệu quả chi phí trong quá trình đào tạo, trong đó hầu hết các tham số đều bị đóng băng (tức là,

: Tác giả liên hệ (majing@hkbu.edu.hk)



Hình 1. Hiệu suất chú thích hình ảnh trên tập dữ liệu MSCOCO dựa trên các mô hình nhẹ và lớn, trong đó I-Tuning-Base/-Medium/-Large tương ứng biểu thị mô hình của chúng tôi với số lượng tham số có thể đào tạo cơ bản/trung bình/lớn.

không được cập nhật trong quá trình đào tạo) và phần còn lại của tập nhỏ có thể đào tạo được. [6] gần đây đã giới thiệu một mô hình ClipCap chuyển đổi hình ảnh thành các vector có độ dài cố định và nhắc nhở GPT2 đồng bộ để chú thích hình ảnh. Tuy nhiên, các vector đã học không thể nắm bắt thông tin hình ảnh chính xác để nâng cao chú thích thể hệ. Để khắc phục những thiếu sót này, chúng tôi đề xuất một khung chú thích hình ảnh nhẹ mới lạ (I-Tuning) để giảm bớt chi phí về mặt tài nguyên tính toán và dữ liệu đào tạo. Chúng tôi thiết kế một mô-đun I-Tuning để kết nối bộ mã hóa thị giác được đào tạo trước (tức là CLIP-ViT [7]) và bộ giải mã ngôn ngữ (tức là GPT2 [8]). Để cân chỉnh giữa các phương thức ngôn ngữ và thị giác, nó đóng vai trò như một bộ lọc đa phương thức tự động chọn thông tin trực quan từ đầu ra của bộ mã hóa tầm nhìn và điều chỉnh trạng thái ẩn đầu ra của bộ giải mã ngôn ngữ. Trong quá trình đào tạo, chúng tôi chỉ cập nhật các thông số mới được giới thiệu trong mô-đun I-Tuning và các tham số của hai mô hình được đào tạo trước bị đóng băng.

Hình 1 minh họa điểm số CIDEr của trọng lượng nhẹ của chúng tôi mô hình và đường cơ sở quy mô lớn. Về mặt đào tạo mô hình, mô hình cơ bản của chúng tôi I-Tuning-Base chỉ chứa khoảng 14M các thông số có thể đào tạo được, cụ thể là ít hơn 10 lần so với các thông số khác các mô hình quy mô lớn như OSCAR. Về mặt dữ liệu, thậm chí mô hình I-Tuning-Large của chúng tôi có thể đạt được hiệu suất tương đương với dữ liệu đào tạo tương đương đối ít hơn. Chúng tôi đánh giá đề xuất khuôn khổ trên 3 tiêu chuẩn chú thích hình ảnh (tức là,

✉: Corresponding author (majing@hkbu.edu.hk)

MSCOCO [9], Flickr30k [10] and NoCaps [11]). The results show that our **I-Tuning** framework achieves comparable or even better performances than large-scale baselines with up to **10 times fewer** trainable parameters and much fewer cross-modal training data. Moreover, our I-Tuning model is agnostic to the pre-trained language models, suggesting a broadly applicable framework.

2. RELATED WORK

CLIP-ViT and GPT2. CLIP-ViT [7] is the state-of-the-art vision encoder. It is pre-trained with contrastive loss [12] to supervise the vision encoder with language description. GPT2 [8] is the state-of-the-art language decoder, which is pre-trained with large-scale text data. In this work, we propose a lightweight image captioning framework **I-Tuning** to leverage these two off-the-shelf pre-trained models.

Image Captioning. Generating the language descriptions from images is an important task to examine the vision-and-language representation ability of a cross-modal model. The recent works choose to increase the model size and the number of training data to further boost the performance [1, 2, 3, 4, 13]. The training process of these models is heavy. As an alternative, the ClipCap model [6] proposes a lightweight captioning model by connecting the off-the-shelf CLIP-ViT and GPT2. However, their method cannot filter the relevant visual information to adjust the output hidden states of GPT2, leading to poor image captioning performance.

Parameter-efficient PLMs Tuning. Recently, the model size of a pre-trained model becomes larger and larger, which makes us hard to fully fine-tune such models. To make use of them without updating all parameters, researchers propose several great ideas, such as Prefix tuning [14], Adapter tuning [5] and Prompt tuning [15]. However, most of them only focus on the NLP area. Our **I-Tuning** extends the parameter-efficient PLMs tuning idea to the cross-modal setting.

3. THE PROPOSED I-TUNING FRAMEWORK

Overview. Our framework contains three components, the non-trainable vision encoder (CLIP-ViT), the non-trainable language decoder (GPT2), and the trainable I-Tuning Module. During training, our framework is trained with the parallel image-caption data and only updated the parameters of the lightweight I-Tuning Module.

During inference, a frozen visual encoder first generates the visual embeddings V of a given image. Then the **I-Tuning** module serves as a lightweight filter to pick the relevant visual information to tune the output hidden states of the frozen language model. As a result, the language generation is conditioned with the given image.

Visual Encoder and Language Decoder. In our framework, we adopt the state-of-the-art vision pre-trained transformer, CLIP-ViT [7] to generate an image’s visual embed-

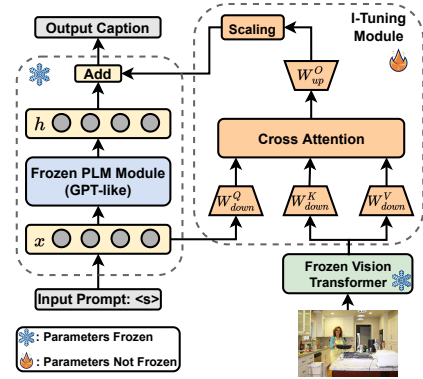


Fig. 2. An overview of our I-Tuning framework for Lightweight Image Captioning.

dings V . Such model takes a sequence of image patches as input and visual representations for each patch as output. For the Language Decoder, we leverage the state-of-the-art auto-regressive pre-trained language model (PLM), GPT2 [8], which is a multi-layer Transformer Decoder model [16] with remarkable language generation ability.

I-Tuning Module. In our framework, the **I-Tuning** module is the key component to extract the relevant visual information from the visual embeddings, which is parallel to a specific PLM module (feedforward) in each Transformer layer. Such module is a bottleneck neural network, sharing a similar structure as the Adapter module [5], but the non-linear activation function is replaced by a cross-attention network (see Figure 2) to filter the visual information from images. The calculation process is as follows:

$$Q_L = W_{down}^Q(X) + b^Q, \quad (1)$$

$$K_V = W_{down}^K(V) + b^K, \quad (2)$$

$$V_V = W_{down}^V(V) + b^V, \quad (3)$$

where X is the input hidden states of a specific PLM module. Then we can get the attention scores across the visual embeddings:

$$S = softmax(Q_L K_V^T). \quad (4)$$

Based on the scores, we can get the final **I-Tuning** output to adjust the output hidden states of the PLM module:

$$\Delta h = \lambda W_{up}^O \left(\sum_i s_i V_{Vi} \right) + b^O, \quad (5)$$

where $\lambda \geq 1$ is a scaling hyper-parameter.

Since the lower layers of PLMs have weaker representation ability, we also propose **I-Tuning Dropping** to remove the **I-Tuning** modules in the first-few layers. As a result, backpropagating through fewer layers can further improve the training efficiency of our models.

Training Objective. The objective is the auto-regressive language modeling conditioned on the visual information:

MSCOCO [9], Flickr30k [10] và NoCaps [11]). Kết quả cho thấy khuôn khổ I-Tuning của chúng tôi đạt được hiệu suất tư ng đư ơ ng hoặc thậm chí tốt hơn so với các đư ờ ng cơ sở quy mô lớn với số lư ợ ng tham số có thể đào tạo ít hơn tới 10 lần và ít dữ liệu đào tạo đa phư ơ ng thức hơn nhiều. Hơn nữa, mô hình I-Tuning của chúng tôi không phụ thuộc vào các mô hình ngôn ngữ đư ợ c đào tạo trư ợ c, cho thấy một khuôn khổ có thể áp dụng rộng rãi.

2. CÔNG TRÌNH LIÊN QUAN

CLIP-ViT và GPT2. CLIP-ViT [7] là bộ mã hóa thị giác tiên tiến. Nó đư ợ c đào tạo trư ợ c với mất mát tư ng đư ợ ng phản [12] để giám sát bộ mã hóa thị giác với mô tả ngôn ngữ. GPT2 [8] là bộ giải mã ngôn ngữ tiên tiến, đư ợ c đào tạo trư ợ c bằng dữ liệu văn bản quy mô lớn. Trong công trình này, chúng tôi đề xuất một khuôn khổ chú thích hình ảnh nhẹ I-Tuning để tận dụng hai mô hình đư ợ c đào tạo trư ợ c có sẵn này.

Chú thích hình ảnh. Việc tạo ra các mô tả ngôn ngữ từ hình ảnh là một nhiệm vụ quan trọng để kiểm tra khả năng biểu diễn ngôn ngữ và tầm nhìn của một mô hình đa phư ơ ng thức. Các công trình gần đây chọn cách tăng kích thước mô hình và số lư ợ ng dữ liệu đào tạo để thúc đẩy hiệu suất hơn nữa [1, 2, 3, 4, 13]. Quá trình đào tạo các mô hình này rất nặng. Một giải pháp thay thế là mô hình ClipCap [6] đề xuất một mô hình chú thích nhẹ bằng cách kết nối CLIP-ViT và GPT2 có sẵn. Tuy nhiên, phư ơ ng pháp của họ không thể lọc thông tin hình ảnh có liên quan để điều chỉnh trạng thái ẩn đầu ra của GPT2, dẫn đến hiệu suất chú thích hình ảnh kém.

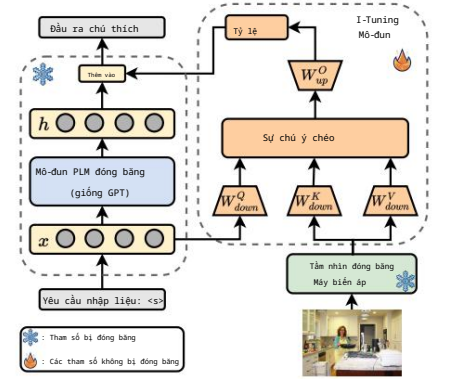
Điều chỉnh PLM hiệu quả về tham số. Gần đây, kích thước mô hình của một mô hình đư ợ c đào tạo trư ợ c ngày càng lớn hơn, khiến chúng ta khó có thể tinh chỉnh hoàn toàn các mô hình như vậy. Để sử dụng chúng mà không cần cập nhật tất cả các tham số, các nhà nghiên cứu đề xuất một số ý tư ờ ng tuyệt vời, chẳng hạn như Điều chỉnh tiền tố [14], Điều chỉnh bộ điều hợp [5] và Điều chỉnh nhắc nhở [15]. Tuy nhiên, hầu hết trong số chúng chỉ tập trung vào lĩnh vực NLP. I-Tuning của chúng tôi mở rộng ý tư ờ ng điều chỉnh PLM hiệu quả về tham số sang cài đặt đa phư ơ ng thức.

3. KHUNG I-TUNING ĐƯ Ợ C ĐỀ XUẤT

Tổng quan. Khung của chúng tôi bao gồm ba thành phần, bộ mã hóa thị giác không thể đào tạo (CLIP-ViT), bộ giải mã ngôn ngữ không thể đào tạo (GPT2) và Mô-đun I-Tuning có thể đào tạo. Trong quá trình đào tạo, khung của chúng tôi đư ợ c đào tạo bằng dữ liệu chú thích hình ảnh song song và chỉ cập nhật các tham số của Mô-đun I-Tuning nhẹ.

Trong quá trình suy luận, một bộ mã hóa hình ảnh đông lạnh đầu tiên tạo ra các nhúng hình ảnh V của một hình ảnh nhất định. Sau đó, mô-đun I-Tuning đóng vai trò như một bộ lọc nhẹ để chọn thông tin hình ảnh có liên quan để điều chỉnh các trạng thái ẩn đầu ra của mô hình ngôn ngữ đông lạnh. Kết quả là, việc tạo ngôn ngữ đư ợ c điều chỉnh với hình ảnh nhất định.

Bộ mã hóa hình ảnh và Bộ giải mã ngôn ngữ. Trong khuôn khổ của chúng tôi, chúng tôi áp dụng bộ chuyển đổi đư ợ c đào tạo trư ợ c về thị giác hiện đại, CLIP-ViT [7] để tạo ra nhúng hình ảnh



Hình 2. Tổng quan về khuôn khổ I-Tuning của chúng tôi dành cho việc tạo chú thích hình ảnh nhẹ.

dings V . Mô hình này lấy một chuỗi các bản vá hình ảnh làm đầu vào và biểu diễn trực quan cho mỗi bản vá làm đầu ra. Đối với Bộ giải mã ngôn ngữ, chúng tôi tận dụng mô hình ngôn ngữ đư ợ c đào tạo trư ợ c tự động hồi quy (PLM) tiên tiến nhất, GPT2 [8], đây là mô hình Bộ giải mã biến áp nhiều lớp [16] có khả năng tạo ngôn ngữ đáng chú ý.

Mô-đun I-Tuning. Trong khuôn khổ của chúng tôi, mô-đun I-Tuning là thành phần chính để trích xuất thông tin trực quan có liên quan từ các nhúng trực quan, song song với mô-đun PLM cụ thể (truyền thẳng) trong mỗi lớp Transformer. Mô-đun như vậy là mạng nơ-ron thất cổ chai, chia sẻ cấu trúc tư ơ ng tự như mô-đun Bộ điều hợp [5], nhưng hàm kích hoạt phi tuyến tính đư ợ c thay thế bằng mạng chú ý chéo (xem Hình 2) để lọc thông tin trực quan từ hình ảnh. Quá trình tính toán như sau:

$$Q_L = W_{down}^Q(X) + b^Q, \quad (1)$$

$$K_V = W_{down}^K(V) + b^K, \quad (2)$$

$$V_V = W_{down}^V(V) + b^V, \quad (3)$$

trong đó X là trạng thái ẩn đầu vào của một mô-đun PLM cụ thể. Sau đó, chúng ta có thể có đư ợ c điểm chú ý trên các nhúng trực quan:

$$S = softmax(Q_L K_V^T). \quad (4)$$

Dựa trên điểm số, chúng ta có thể có đư ợ c đầu ra I-Tuning cuối cùng để điều chỉnh trạng thái ẩn đầu ra của mô-đun PLM:

$$\Delta h = \lambda W_{up}^O \left(\sum_i s_i V_{Vi} \right) + b^O, \quad (5)$$

trong đó $\lambda \geq 1$ là siêu tham số tỷ lệ.

Vì các lớp PLM thấp hơn có khả năng biểu diễn yếu hơn, chúng tôi cũng đề xuất I-Tuning Dropping để loại bỏ các mô-đun I-Tuning trong một vài lớp đầu tiên. Do đó, việc truyền ngư ợ c qua ít lớp hơn có thể cải thiện thêm hiệu quả đào tạo của các mô hình của chúng tôi.

Mục tiêu đào tạo. Mục tiêu là mô hình hóa ngôn ngữ tự hồi quy dựa trên thông tin trực quan:

Model	#Images	#Params	MSCOCO (test)				Flickr (test)			
			CIDEr	BLUE@4	METER	SPICE	CIDEr	BLUE@4	METER	SPICE
Large-scale Cross-Modal Pre-trained Image Captioning Model										
OSCAR _{base} (no tags) [1]	4M	135M	115.6	34.5	29.1	21.9	-	-	-	-
OSCAR _{base} ♠ [1]	4M	135M	123.7	36.5	30.3	23.1	-	-	-	-
Unified VLP [2]	3M	135M	117.7	36.5	28.4	21.3	67.4	30.1	23.0	17.0
XGPT [3]	3M	135M	120.1	37.2	28.6	21.8	70.9	31.8	23.6	17.6
UniTAB [4]	200k	135M	119.1	35.8	28.4	21.5	70.1	30.7	23.7	17.4
VL-T5 [17]	180k	270M	116.5	34.5	28.7	21.9	-	-	-	-
Lightweight Image Captioning Model										
ClipCap(GPT2-Large) [6]	0	43M	113.1	33.5	27.5	21.1	-	-	-	-
Our Lightweight Models w/o VLP										
I-Tuning(GPT2-Base)	0	14M	116.7	34.8	28.3	21.8	61.5	25.2	22.8	16.9
I-Tuning(GPT2-Medium)	0	44M	120.0	35.5	28.8	22.0	72.3	28.8	24.6	19.0
I-Tuning(GPT2-Large)	0	95M	119.4	34.8	29.3	22.4	75.4	29.8	25.1	19.2
Our Lightweight Models w/ VLP										
I-Tuning(GPT2-Base)	110k	14M	118.3	35.2	28.5	22.0	68.4	27.5	24.0	18.4
I-Tuning(GPT2-Medium)	110k	44M	119.1	34.8	29.2	22.2	73.2	29.1	25.2	19.9
I-Tuning(GPT2-Large)	110k	95M	122.2	35.9	29.5	22.6	77.2	30.0	25.5	20.2
Our Lightweight Models w/ I-Tuning Dropping										
I-Tuning(GPT2-Large)	110k	47M	122.1	36.1	29.4	22.6	79.2	31.1	25.3	19.9

Table 1. Evaluations on MSCOCO and Flickr Image Captioning. “-” represents that the model does not report such result in its original paper. **Bold** indicates the best scores of our models. #Images represents the number of distinct images during VLP. #Params represents the number of trainable parameters. ♠: Extra training data are needed to generate the object tags.)

$\mathcal{L} = -\sum_{t=1}^T \log P(x_t|x_{<t}, V)$, where V represents the visual embeddings encoded by the frozen visual encoder, T denotes the length of a sequence and $x_{<t} = (x_0, ..., x_{t-1})$.

4. EXPERIMENT

4.1. Dataset and Setup

We adopts CLIP-ViT B/16 as our visual encoder and GPT2 as language decoder. All of them are frozen during training. We include 3 different GPT2 model sizes, including Base, Medium and Large. For **I-Tuning** modules, the parameters are randomly initialized and updated during training. For VLP, we adopt the cross-modal dataset, Visual Genome [18], which contains 110k distinct images. To evaluate our methods, we use three datasets, namely **MSCOCO** [9], **Flickr30k** [10] and **NoCaps** [11]. For the first two datasets, we follow the Karpathy’s split [19] to split 113.2k/5k/5k and 29.8k/1k/1k images for train/val/test, respectively. We adopt CIDEr [20], BLEU@4 [21], ME-TEOR [22] and SPICE [23] as metrics to evaluate the generated captions. We train our models with the AdamW [24] and 4k batch size. For VLP, our models are pre-trained with 10 epochs. For training on downstream tasks, our models are trained with 30 epochs. For inference, we use the beam search (beam size = 5) to generate captions.

4.2. Result Analysis

Table 1–2 reveal that our lightweight image captioning framework achieves comparable or better performance than all the

large-scale baselines, but contains up to 10 times fewer trainable parameters and/or consume much fewer VLP data.

I-Tuning without VLP. As shown in Table 2, our method outperform the large-scale baselines even without VLP. Especially, the overall CIDEr score of the OSCAR model on the NoCaps even lags behind the frozen GPT2-base with our **I-Tuning** modules by around 2 points, while our model contains around 120M fewer trainable parameters. With the larger GPT2, the performance gap becomes larger. Moreover, our method is also sample efficient. Without any cross-modal pre-training, our **I-Tuning** (GPT2-Medium) already outperforms some baselines with VLP. For example, VL-T5 is pre-trained with 180k distinct cross-modal images, but the CIDEr scores are around 3.4 lower than ours on MSCOCO.

I-Tuning with VLP. Table 1 reveals that after cross-modal pre-training, our **I-Tuning** method achieves better overall performance than all the baseline systems (except OSCAR w/ object tags). Especially, our **I-Tuning** can achieve a CIDEr score of 122.1 on MSCOCO test set, surpassing the XGPT model by 2.0 points, while our method requires less trainable parameters and training data. For the OSCAR model, it requires object tags during pre-training and fine-tuning. Additional supervision is needed to generate these tags. One can find that our lightweight **I-Tuning** framework still reaches comparable performance with only 1.5 CIDEr score lower, while our model requires around 90M less trainable parameters and 30 times less distinct VLP images. Without the help of object tags, OSCAR even lags behind our models without VLP.

I-Tuning with Dropping. Since the lower layers of GPT2 have weaker representation ability, we investigate whether we

Người ời mẫu	#Hình ảnh #Tham số		MSCOCO (kiểm tra)				Flickr (thử nghiệm)			
			CIDEr	BLUE@4	MÉT GIA VỊ	CIDEr	BLUE@4	MÉT GIA VỊ		
Mô hình chú thích hình ảnh được đào tạo trư ớc đa phư ơng thức quy mô lớn										
OSCARbase(không có thẻ) [1]	4M	135M 115,6	OSCAR cơ	34,5	29,1	21,9	-	-	-	-
sở [1]	4M	135M 123.7	VLP thông	36,5	30,3	23,1	-	-	-	-
nhất [2]	3M	135M 117,7	XGPT [3]	36,5	28,4	21,3	67,4	30,1	23.0	17.0
	3M	135M 120.1	UniTAB [4]	37,2	28,6	21,8	70,9	31,8	23.6	17,6
	200k	135M 119.1	VL-T5 [17]	35,8	28,4	21,5	70,1	30,7	23.7	17.4
	180k	270M 116,5		34,5	28,7	21,9	-	-	-	-
Mô hình chú thích hình ảnh nhẹ										
ClipCap(GPT2-Lớn) [6] 0		43 triệu	113,1	33,5	27,5	21.1	-	-	-	-
Các mẫu nhẹ của chúng tôi không có VLP										
I-Tuning (GPT2-Base) 0		14 phút	116,7	34,8	28,3	21,8	61,5	25,2	22,8	16,9
I-Tuning(GPT2-Trung bình) 0		44 triệu	120,0	35,5	28,8	22,0	72,3	28,8	24,6	19.0
I-Tuning(GPT2-Lớn) 0		95 triệu	119,4	34,8	29,3	22,4	75,4	29,8	25,1	19.2
Các mô hình nhẹ của chúng tôi với VLP										
I-Tuning (GPT2-Cơ sở) 110k		14 phút	118,3	35,2	28,5	22.0	68,4	27,5	24.0	18.4
I-Tuning(GPT2-Trung bình) 110k		44 triệu	119,1	34,8	29,2	22.2	73,2	29,1	25.2	19,9
I-Tuning(GPT2-Lớn) 110k		95 triệu	122,2	35,9	29,5	22.6	77,2	30,0	25.5	20.2
Các mô hình nhẹ của chúng tôi với I-Tuning Dropping										
I-Tuning(GPT2-Lớn) 110k		47 triệu	122,1	36,1	29,4	22,6	79,2	31.1	25.3	19,9

Bảng 1. Đánh giá về MSCOCO và Chú thích hình ảnh Flickr. “-” biểu thị rằng mô hình không báo cáo kết quả như vậy trong bài báo gốc của nó. In đậm cho biết điểm số tốt nhất của các mô hình của chúng tôi. #Hình ảnh biểu thị số lượng hình ảnh riêng biệt trong VLP. #Params biểu thị số lượng tham số có thể đào tạo được. : Cần có dữ liệu đào tạo bổ sung để tạo thẻ đối tượng.)

$L = -\sum_{t=1}^T \log P(x_t|x_{<t}, V)$, trong đó V biểu diễn thị giác những được mã hóa bởi bộ mã hóa hình ảnh đồng lạnh, T biểu thị độ dài của một chuỗi và $x_{<t} = (x_0, ..., x_{t-1})$.

4. THÍ NGHIỆM

4.1. Bộ dữ liệu và thiết lập

Chúng tôi áp dụng CLIP-ViT B/16 làm bộ mã hóa hình ảnh và GPT2 như bộ giải mã ngôn ngữ. Tất cả chúng đều bị đóng băng trong quá trình đào tạo. Chúng tôi bao gồm 3 kích thước mô hình GPT2 khác nhau, bao gồm Cơ sở, Trung bình và Lớn. Đối với các mô-đun I-Tuning , các tham số được khởi tạo và cập nhật ngẫu nhiên trong quá trình đào tạo. Đối với VLP, chúng tôi áp dụng tập dữ liệu đa phương thức, Vi-sual Genome [18], chứa 110k hình ảnh riêng biệt. Để đánh giá các phương pháp của chúng tôi, chúng tôi sử dụng ba tập dữ liệu, cụ thể là MSCOCO [9], Flickr30K [10] và NoCaps [11]. Đối với hai tập dữ liệu đầu tiên, chúng tôi theo dõi sự phân chia của Karpathy [19] để chia 113,2k/5k/5k và 29,8k/1k/1k hình ảnh cho đào tạo/đánh giá/kiểm tra, tương ứng. Chúng tôi áp dụng CIDEr [20], BLEU@4 [21], ME-TEOR [22] và SPICE [23] làm số liệu để đánh giá các chú thích được tạo ra. Chúng tôi đào tạo các mô hình của mình bằng AdamW [24] và kích thước lô 4k. Đối với VLP, các mô hình của chúng tôi được đào tạo trước với 10 kỷ nguyên. Để đào tạo về các nhiệm vụ hạ lưu, các mô hình của chúng tôi được đào tạo với 30 kỷ nguyên. Để suy luận, chúng tôi sử dụng chùm tia tìm kiếm (kích thước chùm tia = 5) để tạo chú thích.

4.2. Phân tích kết quả

Bảng 1-2 cho thấy khung chú thích hình ảnh nhẹ của chúng tôi đạt được hiệu suất tương đương hoặc tốt hơn tất cả

đường cơ sở quy mô lớn, nhưng chứa ít hơn tới 10 lần các tham số có thể đào tạo và/hoặc sử dụng ít dữ liệu VLP hơn nhiều.

I-Tuning không có VLP. Như thể hiện trong Bảng 2, phương pháp của chúng tôi vượt trội hơn các đường cơ sở quy mô lớn ngay cả khi không có VLP. Đặc biệt, điểm CIDEr tổng thể của mô hình OSCAR trên NoCaps thậm chí còn tụt hậu so với GPT2-base bị đóng băng với mô-đun I-Tuning của chúng tôi khoảng 2 điểm, trong khi mô hình của chúng tôi chứa khoảng 120M ít hơn các tham số có thể đào tạo. Với GPT2 càng lớn thì khoảng cách hiệu suất càng lớn. Hơn nữa, phương pháp của chúng tôi cũng hiệu quả về mẫu. Không có bất kỳ phương thức chéo nào đào tạo trước, I-Tuning (GPT2-Medium) của chúng tôi đã vượt trội hơn một số đường cơ sở với VLP. Ví dụ, VL-T5 là được đào tạo trước với 180k hình ảnh đa phương thức riêng biệt, nhưng điểm CIDEr thấp hơn điểm MSCOCO của chúng tôi khoảng 3,4.

I-Tuning với VLP. Bảng 1 cho thấy sau khi đào tạo trước đa phương thức, phương pháp I-Tuning của chúng tôi đạt được kết quả tốt hơn hiệu suất tổng thể hơn tất cả các hệ thống cơ sở (trừ OSCAR có thẻ đối tượng). Đặc biệt, I-Tuning của chúng tôi có thể đạt được điểm CIDEr là 122,1 trên bộ thử nghiệm MSCOCO, vượt qua mô hình XGPT bằng 2,0 điểm, trong khi phương pháp của chúng tôi yêu cầu ít tham số có thể đào tạo và dữ liệu đào tạo. Đối với OSCAR mô hình, nó yêu cầu các thẻ đối tượng trong quá trình đào tạo trước và tinh chỉnh. Cần có sự giám sát bổ sung để tạo ra những thẻ. Người ta có thể thấy rằng khung I-Tuning nhẹ của chúng tôi vẫn đạt hiệu suất tương đương chỉ với 1,5 điểm CIDEr thấp hơn, trong khi mô hình của chúng tôi yêu cầu khoảng 90M ít tham số có thể đào tạo hơn và hình ảnh VLP ít hơn 30 lần. Nếu không có sự trợ giúp của thẻ đối tượng, OSCAR thậm chí còn chậm trễ đáng sau các mô hình của chúng tôi mà không có VLP.

I-Tuning với Dropping. Vì các lớp thấp hơn của GPT2 có khả năng đại diện yếu hơn, chúng tôi điều tra xem chúng tôi

Model	#Params	in-domain		near-domain		out-of-domain		Overall	
		CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
OSCAR _{base} [1]	135M	79.6	12.3	66.1	11.5	45.3	9.7	63.8	11.2
ClipCap(GPT2-Large) [6]	43M	84.9	12.1	66.8	10.9	49.1	9.6	65.8	10.9
<i>Our Models</i>									
I-Tuning(GPT2-Base)	14M	83.9	12.4	70.3	11.7	48.1	9.5	67.8	11.4
I-Tuning(GPT2-Medium)	44M	89.6	12.9	77.4	12.2	58.8	10.5	75.4	12.0
I-Tuning(GPT2-Large)	95M	89.6	13.3	80.4	12.6	64.8	11.0	78.5	12.4
<i>Our Models w/ Dropping</i>									
I-Tuning(GPT2-Large)	47M	88.3	12.7	80.8	12.6	66.1	10.8	78.9	12.3

Table 2. Evaluations on NoCaps image captioning. Models are only trained with MSCOCO training set without VLP. #Params represents the number of trainable parameters. **Bold** indicates the best scores.





Image				
Golden Captions	(1) A man with a red helmet on a small moped on a dirt road. (2) Man riding a motor bike on a dirt road on the countryside.	(1) A young girl inhales with the intent of blowing out a candle. (2) A young girl is preparing to blow out her candle.	(1) A man on a bicycle riding next to a train. (2) A person is riding a bicycle but there is a train in the background.	(1) A kitchen is shown with a variety of items on the counters. (2) A kitchen has the windows open and plaid curtains.
Model	Generated Caption			
ClipCap (GPT2-Large)	a man is riding a motorbike 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 kitchen with a sink and a window
OSCAR _{base}	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 kitchen with a sink, dishwasher and a window
I-Tuning (GPT2-Large)	A man riding a motorcycle on a dirt road.	A little girl blowing out a candle on a birthday cake.	A man standing next to a train on a train track.	A kitchen sink sitting under a window next to a window.

Table 3. Examples of our I-Tuning, OSCAR_{base} and ClipCap for the first 4 images in the MSCOCO test set. (Red = inaccurate)

can drop the **I-Tuning** modules in the first few layers, so that we can further reduce the computational overhead during training and inference. Table 1 shows that it is not necessary to include the **I-Tuning** models in all layers. Especially, dropping the **I-Tuning** modules in the first-18 layers can even improve the performance of our model on some evaluation metrics, while the number of trainable parameters is reduced by 50%, improving the efficiency of our models.

Qualitative Evaluation. Table 3 presents the image captioning examples of **I-Tuning**, OSCAR and ClipCap for the first 4 images in the MSCOCO test set. The generated captions of **I-Tuning** depict the image successfully, which can identify the movement of the people in the image. For example, our model can recognize that the little girl is blowing the candles, while ClipCap and OSCAR cannot.

4.3. Cross-Attention Visualization

We visualize the cross-attention maps of **I-Tuning** to examine whether it learns the cross-modal information alignment implicitly. We randomly choose an image in the MSCOCO dataset and present the cross-attention heatmaps in the final **I-Tuning** module of GPT2-Large. Figure 3 shows that our **I-Tuning** module can correctly attend to the corresponding image regions given different tokens. These examples reveal that our method can learn visual grounding implicitly.

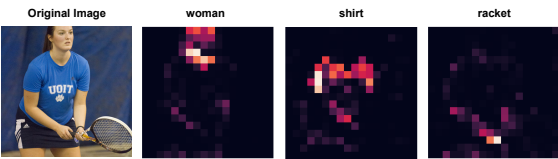


Fig. 3. Visualization of the cross-attention maps of text tokens in the caption “A woman in a blue shirt holding a racket”.

5. CONCLUSION

In this paper, we present a novel lightweight image captioning framework, **I-Tuning**, which efficiently tunes the frozen PLMs with images. Extensive experiments are conducted to verify the effectiveness of our method. Compared with the baseline systems, our method achieves comparable or even better performance, while our models require up to **10 times fewer** trainable parameters and **much fewer** training data.



Acknowledgment

This work is partially supported by National Natural Science Foundation of China Young Scientists Fund (No. 62206233), Hong Kong RGC ECS (No. 22200722), and the Key Research and Development Program of Zhejiang Province (No. 2022C01011).

Người mẫu	#Tham số	trong miền		gần miền		ngoài miền		Tổng thể	
		CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
Cơ sở OSCAR [1]	135 triệu	79,6	12,3	66,1	11,5	45,3	9,7	63,8	11.2
ClipCap(GPT2-Lớn) [6] 43M		84,9	12.1	66,8	10,9	49,1	9,6	65,8	10.9
Các mô hình của chúng tôi									
I-Tuning (GPT2-Cơ sở)	14 phút	83,9	12,4	70,3	11,7	48,1	9,5	67,8	11.4
I-Tuning (GPT2-Trung bình)	44 triệu	89,6	12,9	77,4	12,2	58,8	10,5	75,4	12.0
I-Tuning (GPT2-Lớn)	95 triệu	89,6	13,3	80,4	12,6	64,8	11,0	78,5	12.4
Các mô hình của chúng tôi với Dropping									
I-Tuning (GPT2-Lớn)	47 triệu	88,3	12,7	80,8	12,6	66,1	10.8	78,9	12.3

Bảng 2. Đánh giá về chú thích hình ảnh NoCaps. Các mô hình chỉ được đào tạo với bộ đào tạo MSCOCO mà không có VLP. #Params

biểu thị số lượng tham số có thể đào tạo được. In đậm chỉ ra điểm số tốt nhất.

Hình ảnh				
Vàng	(1) Một người đàn ông đội mũ bảo hiểm màu đỏ có ý định thổi tắt một ngọn nến. (2) Người đàn ông đang lái xe máy trên một đường đất ở nông thôn.	(1) Một cô gái trẻ hít vào với một tách bánh. (2) Một cô gái trẻ đang chuẩn bị thổi tắt ngọn nến của cô ấy.	(1) Một người đàn ông trên một chiếc xe đạp đi tiếp dẫn một chuyến tàu. (2) Một người đi đang đi xe đạp như ng có một chuyến tàu ở phía sau.	(1) Một nhà bếp được hiển thị với nhiều loại của các mặt hàng trên quầy. (2) Một nhà bếp có cửa sổ mở và rèm cửa kẻ caro.
Người mẫu	Chú thích được tạo			
ClipCap (GPT2-Lớn)	một người đàn ông đang đi xe máy một con đường đất.	một cô gái trẻ đang ngồi ở bàn với một tách bánh. một	một người đàn ông đang đứng cạnh một đoàn tàu.	một nhà bếp có bồn rửa và cửa sổ
Cơ sở OSCAR	một người đàn ông đang đi xe máy xuống một con đường đất.	người phụ nữ đang ngồi ở bàn với một đĩa thức ăn.	một người phụ nữ đang đạp xe xuống một con phố bên cạnh tàu hỏa.	một nhà bếp có bồn rửa, máy rửa chén và một cửa sổ
I-Tuning (GPT2-Lớn)	Một người đàn ông đang đi xe máy trên một con đường đất.	Một cô bé thổi ra một nền trên bánh sinh nhật.	Một người đàn ông đứng cạnh một đoàn tàu trên đường ray xe lửa.	Một bồn rửa nhà bếp đặt dưới cửa sổ bên cạnh cửa sổ.

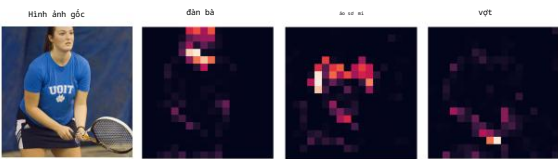
Bảng 3. Ví dụ về I-Tuning, OSCARbase và ClipCap của chúng tôi cho 4 hình ảnh đầu tiên trong bộ thử nghiệm MSCOCO. (Đỏ = không chính xác)

có thể thả các mô-đun I-Tuning trong vài lớp đầu tiên, vì vậy rằng chúng ta có thể giảm thêm chi phí tính toán trong quá trình đào tạo và suy luận. Bảng 1 cho thấy không cần thiết phải đưa các mô hình I-Tuning vào tất cả các lớp. Đặc biệt, thả các mô-đun I-Tuning trong 18 lớp đầu tiên thậm chí có thể cải thiện hiệu suất của mô hình của chúng tôi trên một số đánh giá số liệu, trong khi số lượng các tham số có thể đào tạo được giảm đi tăng 50%, cải thiện hiệu quả của mô hình của chúng tôi.

Đánh giá định tính. Bảng 3 trình bày các ví dụ về chú thích hình ảnh của I-Tuning, OSCAR và ClipCap cho 4 hình ảnh đầu tiên trong bộ thử nghiệm MSCOCO. Các chú thích được tạo ra của I-Tuning mô tả hình ảnh thành công, có thể xác định chuyển động của mọi người trong hình ảnh. Ví dụ, mô hình của chúng tôi có thể nhận ra rằng cô bé đang thổi nến, trong khi ClipCap và OSCAR thì không.

4.3. Hình dung sự chú ý chéo

Chúng tôi hình dung các bản đồ chú ý chéo của I-Tuning để kiểm tra xem nó có học được sự liên kết thông tin đa phương thức hay không ngầm định. Chúng tôi chọn ngẫu nhiên một hình ảnh trong MSCOCO bộ dữ liệu và trình bày các bản đồ nhiệt chú ý chéo trong bản cuối cùng Mô-đun I-Tuning của GPT2-Large. Hình 3 cho thấy rằng Mô-đun I-Tuning có thể tham gia chính xác vào các tư duy ứng vùng hình ảnh được cung cấp các mã thông báo khác nhau. Những ví dụ này tiết lộ rằng phương pháp của chúng tôi có thể học được nền tảng trực quan một cách ngầm định.



Hình 3. Hình ảnh hóa các bản đồ chú ý chéo của các mã thông báo văn bản trong chú thích “Một người phụ nữ mặc áo sơ mi xanh đang cầm vợt”.

5. KẾT LUẬN

Trong bài báo này, chúng tôi trình bày một khuôn khổ chú thích hình ảnh nhẹ mới, I-Tuning, có thể điều chỉnh hiệu quả hình ảnh đông lạnh PLM có hình ảnh. Các thí nghiệm mở rộng được tiến hành để xác minh hiệu quả của phương pháp của chúng tôi. So với hệ thống cơ sở, phương pháp của chúng tôi đạt được hiệu quả tư duy ngang hoặc thậm chí hiệu suất tốt hơn, trong khi các mô hình của chúng tôi yêu cầu lên đến 10 lần ít tham số có thể đào tạo hơn và ít dữ liệu đào tạo hơn nhiều.

Sự thừa nhận

Công trình này được hỗ trợ một phần bởi National Natural Science Quỹ Nhà khoa học trẻ Trung Quốc (Số 62206233), Hồng Kông RGC ECS (Số 22200722) và Chương trình nghiên cứu phát triển trọng điểm của tỉnh Chiết Giang (Số 2022C01011).

6. REFERENCES

[1] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao, “Os-car: Object-semantics aligned pre-training for vision-language tasks,” in *ECCV 2020*.

[2] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao, “Unified vision-language pre-training for image captioning and vqa,” *AAAI*, 2020.

[3] Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, and Ming Zhou, “Xgpt: Cross-modal generative pre-training for image captioning,” in *NLPCC*, 2021.

[4] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang, “Unitab: Unifying text and box outputs for grounded vision-language modeling,” in *ECCV*, 2022.

[5] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly, “Parameter-efficient transfer learning for NLP,” in *ICML*, 2019.

[6] Ron Mokady, Amir Hertz, and Amit H. Bermano, “Clip-cap: Clip prefix for image captioning,” 2021.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, “Learning transferable visual models from natural language supervision,” in *ICML*, 2021.

[8] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, “Language models are unsupervised multitask learners,” 2019.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár, “Microsoft coco: Common objects in context,” 2015.

[10] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik, “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models,” in *ICCV*, 2015.

[11] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson, “Nocaps: novel object captioning at scale,” *ICCV*, 2019.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *CVPR*, 2020.

[13] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao, “SimVLM: Simple visual language model pretraining with weak supervision,” in *ICLR*, 2022.

[14] Xiang Lisa Li and Percy Liang, “Prefix-tuning: Optimizing continuous prompts for generation,” in *ACL*, 2021.

[15] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang, “Gpt understands, too,” *arXiv:2103.10385*, 2021.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in *Neurips*, 2017.

[17] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal, “Unifying vision-and-language tasks via text generation,” in *ICML*, 2021.

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yan-nis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” *IJCV*, vol. 123, pp. 32–73, 2016.

[19] Andrej Karpathy and Fei-Fei Li, “Deep visual-semantic alignments for generating image descriptions,” in *CVPR*, 2015.

[20] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh, “Cider: Consensus-based image description evaluation,” in *CVPR*, 2015.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *ACL*, 2002.

[22] Satanjeev Banerjee and Alon Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.

[23] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould, “Spice: Semantic propositional image caption evaluation,” 2016.

[24] Ilya Loshchilov and Frank Hutter, “Decoupled weight decay regularization,” in *ICLR*, 2019.

6. TÀI LIỆU THAM KHẢO

[1] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi and Jianfeng Gao, “Os-car: Ngữ nghĩa đối tượng được căn chỉnh trước khi đào tạo cho các nhiệm vụ ngôn ngữ thị giác,” trong *ECCV 2020*.

[2] Luowei Chu, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso và Jianfeng Gao, “Đào tạo trước ngôn ngữ thị giác thống nhất cho chú thích hình ảnh và vqa,” *AAAI*, 2020.

[3] Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, và Ming Zhou, “Xgpt: Tiền đào tạo tạo đa phương thức để chú thích hình ảnh,” trong *NLPCC*, 2021.

[4] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu và Li-juan Wang, “Unitab: Thống nhất đầu ra văn bản và hộp cho “Mô hình ngôn ngữ thị giác có cơ sở”, trong *ECCV*, 2022.

[5] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan và Sylvain Gelly, “Học chuyển giao hiệu quả tham số cho NLP,” trong *ICML*, 2019.

[6] Ron Mokady, Amir Hertz và Amit H. Bermano, “Clip-cap: Tiền tố clip để chú thích hình ảnh,” 2021.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger và Ilya Sutskever, “Học tập có thể chuyển giao mô hình trực quan từ giám sát ngôn ngữ tự nhiên,” trong *ICML*, 2021.

[8] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei và Ilya Sutskever, “Mô hình ngôn ngữ là những người học đa nhiệm không có giám sát,” 2019.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick và Piotr Dollár, “Microsoft coco: Các đối tượng phổ biến trong bối cảnh”, 2015.

[10] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier và Svetlana Lazebnik, “Các thực thể Flickr30k: Thu thập sự tương ứng giữa vùng và cụm từ để có hình ảnh và câu phong phú hơn mô hình,” trong *ICCV*, 2015.

[11] Agrawal khắc nghiệt, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee và Peter Anderson, “Nocaps: đối tượng mới lạ “chú thích ở quy mô lớn”, *ICCV*, 2019.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, và Ross B. Girshick, “Độ tương phản động lượng cho việc học biểu diễn trực quan không giám sát”, trong *CVPR*, 2020.

[13] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov và Yuan Cao, “SimVLM: Hình ảnh đơn giản mô hình ngôn ngữ tiền đào tạo với sự giám sát yếu,” trong *ICLR*, 2022.

[14] Xiang Lisa Li và Percy Liang, “Điều chỉnh tiền tố: Tối ưu hóa các lời nhắc liên tục để tạo ra,” trong *ACL*, 2021.

[15] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang và Jie Tang, “Gpt hiểu, cũng vậy,” *arXiv:2103.10385*, 2021.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser và Illia Polosukhin, “Sự chú ý là tất cả những gì bạn cần,” trong *Neurips*, 2017.

[17] Jaemin Cho, Jie Lei, Hao Tan và Mohit Bansal, “Thống nhất các nhiệm vụ về tầm nhìn và ngôn ngữ thông qua việc tạo văn bản,” trong *ICML*, 2021.

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yan-nis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein và Li Fei-Fei, “Bộ gen trực quan: Kết nối ngôn ngữ và tầm nhìn sử dụng hình ảnh dày đặc được cộng đồng đóng góp chú thích,” *IJCV*, tập 123, trang 32-73, 2016.

[19] Andrej Karpathy và Fei-Fei Li, “Ngữ nghĩa thị giác sâu sắc căn chỉnh để tạo mô tả hình ảnh,” trong *CVPR*, 2015.

[20] Ramakrishna Vedantam, C. Lawrence Zitnick và Devi Parikh, “Rủi ro: Mô tả hình ảnh dựa trên sự đồng thuận đánh giá,” trong *CVPR*, 2015.

[21] Kishore Papineni, Salim Roukos, Todd Ward và Wei-Jing Zhu, “Bleu: một phương pháp đánh giá tự động “dịch máy”, trong *ACL*, 2002.

[22] Satanjeev Banerjee và Alon Lavie, “METEOR: Một phép đo tự động để đánh giá MT với mối tương quan được cải thiện với các phán đoán của con người,” trong *Biên bản báo cáo Hội thảo ACL về Đánh giá nội tại và bên ngoài Các biện pháp dịch máy và/hoặc tóm tắt*, 2005.

[23] Peter Anderson, Basura Fernando, Mark Johnson và Stephen Gould, “Spice: Hình ảnh mệnh đề ngữ nghĩa đánh giá chú thích,” 2016.

[24] Ilya Loshchilov và Frank Hutter, “Trọng lượng tách rời “Điều chỉnh phân rã”, trong *ICLR*, 2019.