mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections

Chenliang Li*, **Haiyang Xu***, **Junfeng Tian**, **Wei Wang**, **Ming Yan**†, **Bin Bi**†, **Jiabo Ye**, **Hehong Chen**, **Guohai Xu**, **Zheng Cao**, **Ji Zhang**, **Songfang Huang**, **Fei Huang**, **Jingren Zhou**, **Luo Si**
DAMO Academy, Alibaba Group
{lcl193798, shuofeng.xhy, tjf141457, hebian.ww, ym119608, b.bi, yejiabo.yjb, hehong.chh, guohai.xgh, zhengzhi.cz, zj122146, songfang.hsf, f.huang, jingren.zhou, luo.si}@alibaba-inc.com

Abstract

Large-scale pretrained foundation models have been an emerging paradigm for building artificial intelligence (AI) systems, which can be quickly adapted to a wide range of downstream tasks. This paper presents mPLUG, a new vision-language foundation model for both cross-modal understanding and generation. Most existing pre-trained models suffer from the problems of low computational efficiency and information asymmetry brought by the long visual sequence in cross-modal alignment. To address these problems, mPLUG introduces an effective and efficient vision-language architecture with novel cross-modal skip-connections, which creates inter-layer shortcuts that skip a certain number of layers for time-consuming full self-attention on the vision side.

mPLUG is pre-trained end-to-end on large-scale image-text pairs with both discriminative and generative objectives. It achieves state-of-the-art results on a wide range of vision-language downstream tasks, such as image captioning, image-text retrieval, visual grounding and visual question answering. mPLUG also demonstrates strong zero-shot transferability when directly transferred to multiple video-language tasks.

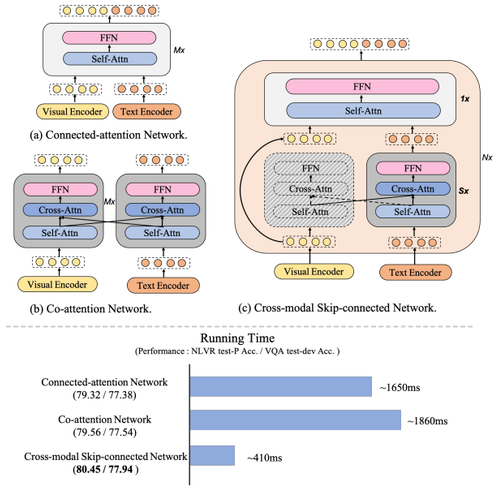


Figure 1: Illustration of two conventional cross-modal fusion networks and our proposed cross-modal skip-connected network. We compare the running time and performance of different fusion networks, where the total fusion layers, image encoder and text encoder are all kept the same. The running time is the total forward time of 100 samples in different fusion networks.

1 Introduction

Large-scale pre-training of vision-language models have recently received tremendous success on a wide range of cross-modal tasks [1, 2, 3, 4, 5, 6, 7]. Such vision-language models learn cross-modal representations from a quantity of image-text pairs by aligning the visual and linguistic modalities. A great challenge of learning vision-language models is to find a good alignment between the two modalities to close the semantic gap in-between.

To discover a cross-modal alignment, prior studies [4, 8, 9] employ a pre-trained object detector to extract salient regions from images, which are then aligned with language counterparts. Such an architecture, however, is generally limited by the power of the object detector, the pre-defined visual semantics it can represent, and the quantity of annotations available. Besides, it is also computationally expensive to extract region-based visual features from high-resolution (e.g. 600×1000) images. More recent work [3, 7, 6, 10, 11], which scales and performs better on many vision-language tasks, drops the requirement of pre-trained object detection and enables a direct alignment between the image and text representations in an end-to-end manner. These models extract finer-grained visual representation with a long sequence of image patches or grids for good vision understanding [11].

* Equal contribution
† Corresponding authors

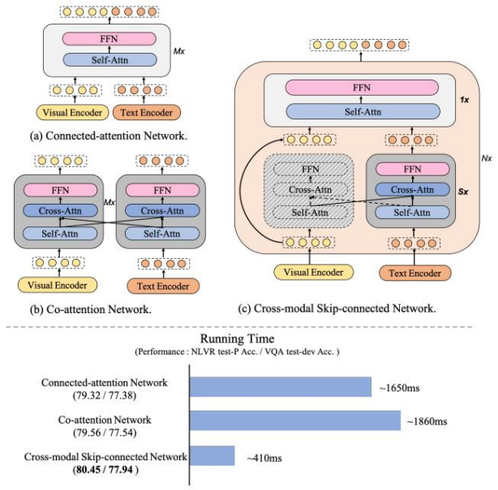
mPLUG: Học ngôn ngữ bằng thị giác hiệu quả và hiệu suất Kết nối bỏ qua đa phư ơng thức

Trần Lư ơng Lý, Haiyang Xu , Junfeng Tian, Wei Wang, Ming Yan† ,Bân Bí†,Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Chu, Luo Si
Học viện DAMO, Tập đoàn Alibaba
{lcl193798, shuofeng.xhy, tjf141457, hebian.ww, ym119608, b.bi, yejiabo.yjb, hehong.chh, guohai.xgh, zhengzhi.cz, zj122146, songfang.hsf, f.huang, jingren.zhou, luo.si}@alibaba-inc.com

Tóm tắt

Các mô hình nền tảng đư ợc đào tạo trư ớc quy mô lớn đã là một mô hình mới nổi để xây dựng hệ thống trí tuệ nhân tạo (AI), có thể có thể nhanh chóng thích ứng với nhiều nhiệm vụ hạ nguồn. Bài báo này trình bày mPLUG, một mô hình nền tảng ngôn ngữ-tầm nhìn mới cho cả sự hiểu biết và thế hệ liên phư ơng thức. Hầu hết các mô hình đư ợc đào tạo trư ớc hiện có đều gặp vấn đề từ các vấn đề về hiệu quả tính toán thấp và sự bất đối xứng thông tin do chuỗi hình ảnh dài trong sự sắp xếp chéo phư ơng thức. Để giải quyết những vấn đề này, mPLUG giới thiệu một kiến trúc ngôn ngữ thị giác hiệu quả và hiệu suất cao với phư ơng thức liên phư ơng thức mới bỏ qua các kết nối, tạo ra lớp liên kết các phím tắt bỏ qua một số lớp nhất định dành thời gian để chú ý hoàn toàn vào bản thân phía tầm nhìn.

mPLUG đư ợc đào tạo trư ớc từ đầu đến cuối trên các cặp hình ảnh-văn bản quy mô lớn với cả phân biệt và các mục tiêu tạo ra. Nó đạt đư ợc kết quả tiên tiến trên nhiều nhiệm vụ hạ nguồn ngôn ngữ thị giác, chẳng hạn như hình ảnh chú thích, tìm kiếm hình ảnh-văn bản, nền tảng trực quan và trả lời câu hỏi trực quan. mPLUG cũng chứng minh khả năng chuyển giao mạnh mẽ khi chuyển trực tiếp sang nhiều nhiệm vụ ngôn ngữ video.



Hình 1: Minh họa hai phư ơng thức liên phư ơng thức thông thư ờng mạng lư ới hợp nhất và mạng lư ới kết nối bỏ qua đa phư ơng thức đư ợc đề xuất của chúng tôi. Chúng tôi so sánh thời gian chạy và hiệu suất của các mạng lư ới hợp nhất khác nhau, trong đó các lớp hợp nhất tổng thể, bộ mã hóa hình ảnh và bộ mã hóa văn bản là tất cả đều giữ nguyên. Thời gian chạy là tổng thời gian chuyển tiếp thời gian của 100 mẫu trong các mạng lư ới tổng hợp khác nhau.

1 Giới thiệu

Đào tạo trư ớc quy mô lớn các mô hình ngôn ngữ thị giác gần đây đã nhận đư ợc thành công to lớn trên một nhiều nhiệm vụ liên phư ơng thức [1, 2, 3, 4, 5, 6, 7]. Các mô hình ngôn ngữ thị giác như vậy học liên phư ơng thức biểu diễn từ một số lư ợng cặp hình ảnh-văn bản bằng cách sắp xếp các phư ơng thức trực quan và ngôn ngữ. A Thách thức lớn nhất của việc học các mô hình ngôn ngữ thị giác là tìm ra sự liên kết tốt giữa hai phư ơng thức để thu hẹp khoảng cách ngữ nghĩa ở giữa.

Để khám phá sự liên kết đa phư ơng thức, các nghiên cứu trư ớc đây [4, 8, 9] sử dụng một máy dò đối tư ợng đư ợc đào tạo trư ớc để trích xuất các vùng nổi bật từ hình ảnh, đó là sau đó căn chỉnh với các đối tác ngôn ngữ. Một tuy nhiên, kiến trúc nói chung bị giới hạn bởi sức mạnh của máy dò đối tư ợng, hình ảnh đư ợc xác định trư ớc ngữ nghĩa mà nó có thể biểu diễn và số lư ợng chú thích có sẵn. Bên cạnh đó, việc trích xuất các đặc điểm trực quan dựa trên vùng từ hình ảnh có độ phân giải cao (ví dụ: 600×1000) cũng tốn kém về mặt tính toán . Công trình gần đây hơn [3, 7, 6, 10, 11], trong đó có quy mô và thực hiện tốt hơn nhiều nhiệm vụ ngôn ngữ thị giác, loại bỏ yêu cầu phát hiện đối tư ợng đư ợc đào tạo trư ớc và cho phép căn chỉnh trực tiếp giữa hình ảnh và văn bản biểu diễn trong một end-to-end cách. Các mô hình này trích xuất biểu diễn hình ảnh chi tiết hơn với một chuỗi hình ảnh dài các miếng vá hoặc lư ới để có tầm nhìn tốt [11].

Đóng góp ngang nhau
† Tác giả liên hệ

However, there exist two significant problems in modeling long visual sequences: 1) *efficiency*: full self-attention on long visual sequences requires much more computation than that on textual sequences, and 2) *information asymmetry*: the caption text in widely-used image-text pre-training data is usually short and highly abstract while more detailed and diverse information can be extracted from the image. This asymmetry presents challenges for effective multi-modal fusion between the modalities.

One straightforward way of multi-modal fusion is the connected-attention network as shown in Figure 1 (a). It adopts a single Transformer [12] network for early fusion of vision and language by simply taking the concatenation of visual and linguistic features as input [13]. This paradigm allows self-attention to discover alignments between the modalities from the bottom level, and requires full self-attention on the concatenation of cross-modal sequences, which is rather time-consuming. Besides, this type of methods process information from both modalities equally, which may suffer from the information asymmetry especially when there is a big difference in information density or sequence lengths between the modalities.

Another line of work keeps separate Transformer networks for both textual and visual features, and uses techniques such as cross-attention to enable cross-modal interaction [11], as shown in Figure 1 (b). This architecture design conducts multi-modal fusion on both modalities independently, which can help alleviate the information asymmetry problem. However, it still suffers from computation inefficiency for full self-attention on long visual sequences, and it is not that parameter-efficient with two separate Transformer networks.

In this work, we propose mPLUG, a unified Multi-modal Pre-training framework for both vision-Language Understanding and Generation. mPLUG performs effective and efficient vision-language learning with novel cross-modal skip-connections to address the fundamental information asymmetry problem. Instead of fusing visual and linguistic representations at the same levels, the cross-modal skip-connections enables the fusion to occur at disparate levels in the abstraction hierarchy across the modalities. It creates inter-layer shortcuts that skip a certain number of layers for visual representations to reflect the semantic richness of language compared to vision. As shown in Figure 1 (c), in each block of our cross-modal skip-connected network, mPLUG first adopts an asymmetric co-attention architecture at the first few layers for efficiency, by removing the co-attention on vision side. It is then followed by one layer of connected-attention, by concatenating the original visual representation and the co-attention output on the language side as input. In addition to the modeling efficacy due to the asymmetry, the cross-modal skip-connections ease the model training by alleviating vanishing gradients with the inserted shortcuts. Figure 1 shows that the new cross-modal skip-connected network achieves superior performance with at least four times speeding-up than other cross-modal fusion networks.

Our key contributions can be summarized as follows:

- We propose a unified vision-language pre-trained model mPLUG of cross-modal understanding and generation for both effectiveness and efficiency in cross-modal learning.
- We introduce a new asymmetric vision-language architecture with novel cross-modal skip-connections, to address two fundamental problems of information asymmetry and computation inefficiency in multi-modal fusion.
- mPLUG achieves state-of-the-art performance on a wide range of vision-language tasks, including image captioning, image-text retrieval, visual grounding and visual question answering. mPLUG also demonstrates strong zero-shot transferability when directly transferred to a wide range of vision-language and video-language tasks.

2 Related Work

2.1 Vision-Language Pre-training

Vision-Language pre-training (VLP) has recently received tremendous success and achieved state-of-the-art results across a variety of vision-language tasks [14, 15, 16]. In terms of how information from different modalities are aggregated, typical approaches to VLP [1, 2, 3, 5, 6, 17, 18] can be roughly divided into two categories: *dual encoder* and *fusion encoder*. Dual encoder approach utilizes two single-modal encoders to encode images and text separately, and then uses simple functions such as dot product to model the instance-level cross-modal interaction between image and text. The

Machine Translated by Google

Tuy nhiên, tồn tại hai vấn đề quan trọng trong
mô hình hóa các chuỗi hình ảnh dài: 1) hiệu quả: đầy đủ
sự chú ý của bản thân vào các chuỗi hình ảnh dài đòi hỏi
nhiều tính toán hơn so với trên các chuỗi văn bản và 2)
thông tin không đối xứng: văn bản chú thích trong quá trình
đào tạo trư ớc hình ảnh-văn bản đư ợc sử dụng rộng rãi
dữ liệu thư ờng ngắn và trừu tư ợng trong khi nhiều hơn
thông tin chi tiết và đa dạng có thể đư ợc trích xuất
từ hình ảnh. Sự bất đối xứng này đặt ra những thách thức
cho sự kết hợp đa phư ơng thức hiệu quả giữa
các phư ơng thức.

Một cách đơn giản để kết hợp đa phư ơng thức
là mạng lư ới chú ý đư ợc kết nối như thể hiện trong Hình
1 (a). Nó sử dụng một mạng lư ới Transformer [12] duy nhất
để hợp nhất sớm thị giác và ngôn ngữ bằng
chỉ cần lấy sự kết hợp của các đặc điểm trực quan và ngôn
ngữ làm đầu vào [13]. Mô hình này cho phép
tự chú ý để khám phá sự liên kết giữa
các phư ơng thức từ cấp độ thấp nhất và đòi hỏi đầy đủ
sự chú ý của bản thân vào sự kết nối của phư ơng thức chéo
trình tự, khá tốn thời gian. Bên cạnh đó, loại phư ơng pháp
này xử lý thông tin
từ cả hai phư ơng thức như nhau, điều này có thể bị ảnh hư ởng
từ sự bất đối xứng thông tin đặc biệt là khi
có một sự khác biệt lớn về mật độ thông tin hoặc
độ dài chuỗi giữa các phư ơng thức.

Một dòng công việc khác giữ riêng Transformer
mạng lư ới cho cả tính năng văn bản và hình ảnh, và
sử dụng các kỹ thuật như chú ý chéo để cho phép
tư ơng tác liên phư ơng thức [11], như thể hiện trong Hình 1
(b). Thiết kế kiến trúc này thực hiện đa phư ơng thức
sự kết hợp của cả hai phư ơng thức một cách độc lập,
có thể giúp làm giảm vấn đề bất đối xứng thông tin . Tuy
nhiên, nó vẫn gặp phải vấn đề về tính toán
sự kém hiệu quả đối với sự chú ý hoàn toàn vào các chuỗi
hình ảnh dài và nó không hiệu quả về mặt tham số với
hai mạng lư ới biến áp riêng biệt.

Trong công trình này, chúng tôi đề xuất mPLUG, một khuôn
khổ tiền đào tạo đa phư ơng thức thống nhất cho cả
tầm nhìn-Hiểu ngôn ngữ và tạo ra ngôn ngữ.
mPLUG thực hiện việc học ngôn ngữ thị giác hiệu quả và hiệu
suất cao với các kết nối bỏ qua đa phư ơng thức mới để giải
quyết thông tin cơ bản
vấn đề bất đối xứng. Thay vì hợp nhất hình ảnh
và các biểu diễn ngôn ngữ ở cùng cấp độ,
kết nối bỏ qua đa phư ơng thức cho phép hợp nhất
xảy ra ở các cấp độ khác nhau trong hệ thống phân cấp trừu
tư ợng trên các phư ơng thức. Nó tạo ra các lớp liên lớp
các phím tắt bỏ qua một số lớp nhất định cho
biểu diễn trực quan để phản ánh sự phong phú về mặt ngữ
nghĩa của ngôn ngữ so với thị giác. Như đã trình bày
trong Hình 1 (c), trong mỗi khối của phư ơng thức liên phư ơng thức của chúng tôi
mạng bỏ qua kết nối, mPLUG đầu tiên áp dụng một
kiến trúc đồng chú ý không đối xứng ở vài đầu tiên
các lớp để đạt hiệu quả, bằng cách loại bỏ sự chú ý đồng thời
về phía tầm nhìn. Sau đó là một lớp
sự chú ý đư ợc kết nối, bằng cách nối kết bản gốc
biểu diễn trực quan và đầu ra đồng chú ý
về phía ngôn ngữ như đầu vào. Ngoài ra
hiệu quả mô hình hóa do tính bất đối xứng, các kết nối bỏ qua
đa phư ơng thức giúp việc đào tạo mô hình dễ dàng hơn bằng cách
làm giảm sự biến mất của gradient bằng cách chèn
phím tắt. Hình 1 cho thấy phư ơng thức liên phư ơng thức mới
mạng kết nối bỏ qua đạt đư ợc hiệu suất vư ợt trội với tốc
độ tăng tốc ít nhất gấp bốn lần so với
các mạng lư ới kết hợp đa phư ơng thức khác.

Những đóng góp chính của chúng tôi có thể đư ợc tóm tắt như sau
sau đây:

- Chúng tôi đề xuất một mô hình mPLUG đư ợc đào tạo
  trư ớc bằng ngôn ngữ thị giác thống nhất của
  đứng và thế hệ cho cả hiệu quả
  và hiệu quả trong học tập đa phư ơng thức.
- Chúng tôi giới thiệu một kiến trúc ngôn ngữ tầm nhìn
  bất đối xứng mới với phư ơng thức liên phư ơng thức mới
  bỏ qua các kết nối, để giải quyết hai vấn đề cơ bản
  các vấn đề về thông tin bất đối xứng và tính toán
  kém hiệu quả trong quá trình kết hợp đa phư ơng thức.
- mPLUG đạt đư ợc hiệu suất tiên tiến
  trên nhiều nhiệm vụ ngôn ngữ thị giác, bao gồm chú
  thích hình ảnh, truy xuất văn bản hình ảnh,
  nền tảng trực quan và trả lời câu hỏi trực quan .
  mPLUG cũng chứng minh khả năng chuyển đổi mạnh mẽ
  khi đư ợc chuyển trực tiếp
  cho nhiều nhiệm vụ ngôn ngữ thị giác và ngôn ngữ
  video.

2 Công trình liên quan

2.1 Đào tạo trư ớc về Tầm nhìn-Ngôn ngữ

Đào tạo trư ớc về Ngôn ngữ Tầm nhìn (VLP) gần đây đã
đã đạt đư ợc thành công to lớn và đạt đư ợc trạng thái
kết quả nghệ thuật trên nhiều ngôn ngữ tầm nhìn
nhiệm vụ [14, 15, 16]. Về mặt thông tin
từ các phư ơng thức khác nhau đư ợc tổng hợp, điển hình
các cách tiếp cận VLP [1, 2, 3, 5, 6, 17, 18] có thể là
đư ợc chia thành hai loại: bộ mã hóa kép
và bộ mã hóa hợp nhất. Phư ơng pháp mã hóa kép sử dụng
hai bộ mã hóa đơn phư ơng thức để mã hóa hình ảnh và
văn bản riêng biệt, và sau đó sử dụng các chức năng đơn giản như
như tích vô hư ớng để mô hình hóa tư ơng tác đa phư ơng
thức ở cấp độ thể hiện giữa hình ảnh và văn bản.

advantage of dual encoder models like CLIP [17] and ALIGN [18] is that images and text can be pre-computed and cached, which is quite computation-efficient and more appropriate for retrieval tasks. However, they tend to fail in handling more complicated VL understanding tasks that require complex reasoning, such as visual question answering [14]. In contrast, fusion encoder approach uses deep fusion functions such as multi-layer self-attention and cross-attention networks to model the fine-grained cross-modal interaction between image and text sequences. Representative methods of this category include the single-stream architecture such as UNITER [2] and OSCAR [4], and two-stream architecture such as LXMERT [1], ALBEF [6] and ERNIE-ViL [5]. This kind of methods can better capture the underlying association between image and text for vision-language understanding tasks, while it needs to jointly encode all possible image-text pairs, which leads to a relatively slow inference speed.

To improve the inference speed, some recent work such as Pixel-BERT [3], E2E-VLP [19] and ViLT [10] removes the complicated object detector in feature extraction, and conducts end-to-end VL learning with CNN-based grid features and linearly projected patched embeddings, respectively. To combine the benefits of both categories of architectures, VLMo [20] further unifies the dual encoder and fusion encoder modules with shared mixture-of-modality-experts Transformer. In this work, mPLUG introduces a new cross-modal fusion mechanism with cross-modal skip-connections, to enables the fusion to occur at disparate levels in the abstraction hierarchy across the modalities. It achieves superior performances in effectiveness and efficiency across a wide range of VL tasks.

2.2 Skip-connection

Skip-connection is a popular technique to bypass the gradient exploding or vanishing problem for model optimization in deep neural networks, which is widely-used in CV and NLP architectures such as ResNet [21] and Transformer [12]. A variety of skip connection methods have been proposed in recent years [22, 21, 12, 23, 24, 25]. ResNet [21] introduces summed shortcut connections between different layers using simple identity mapping, while highway network [22] designs a transform gating function to control the balance of the input and the transformed input. DenseNet [23] designs new architectures with concatenated skip-connections, allowing the subsequent layers to re-use all the middle representations of previous layers. Layer Normalization and recursive skip connection are further used in combination with plain skip connection for further stablizing model optimization and better incorporating the transformed input [12, 25]. In this work, mPLUG proposes a new cross-modal skip connection method to address cross-modal fusion problem, and combines the concatenated skip-connection and summed skip-connection for choosing whether to attend to all the concatenated representations of different modalities or just focus on the cross-modal interaction part at each layer.

3 mPLUG

In this section, we will first introduce our new model architecture with the key module of the cross-modal skip-connected network, and then give the details of the pre-training objectives and scalable training infrastructure.

3.1 Model Architecture

As shown in Figure 2, mPLUG consists of two unimodal encoders for image and text independently, a cross-modal skip-connected network and a decoder for text generation. To better model the inherent modality bias information, we first use two unimodal encoders to encode image and text separately. Following [11, 26], we use a visual transformer [27] directly on the image patches as the visual encoder, which is more computation-friendly than using pre-trained object detectors for visual feature extraction [8, 9]. The visual encoder divides an input image into patches and encodes them as a sequence of embeddings $\{v_{cls}, v_1, v_2, ..., v_M\}$ with an additional $[CLS]$ token. The input text is fed to the text encoder and represented as a sequence of embeddings $\{l_{cls}, l_1, l_2, ..., l_N\}$, where l_{cls} is the embedding of the $[CLS]$ token and used to summarize the input text. Then, the visual and linguistic representations are fed into a cross-modal skip-connected network, which consists of multiple skip-connected fusion blocks. In each skip-connected fusion block, we adopt connected cross-modal fusion to each of S *asymmetric co-attention* layers where S is a fixed stride value. The aim of this network is to take advantage of the effectiveness of the connected cross-modal fusion and the efficiency of the asymmetric co-attention for

Machine Translated by Google

lợi thế của các mô hình mã hóa kép như CLIP [17] và ALIGN [18] là hình ảnh và văn bản có thể được tính toán trước và lưu vào bộ nhớ đệm, điều này khá hiệu quả về mặt tính toán và phù hợp hơn cho các tác vụ truy xuất. Tuy nhiên, chúng có xu hướng thất bại trong việc xử lý các nhiệm vụ hiểu VL phức tạp hơn đòi hỏi sự phức tạp lý luận, chẳng hạn như trả lời câu hỏi trực quan [14]. Ngược lại, phương pháp mã hóa hợp nhất sử dụng các chức năng hợp nhất sâu như tự chú ý nhiều lớp và mạng lưới chú ý chéo để mô hình hóa các chi tiết nhỏ tương tác đa phương thức giữa hình ảnh và văn bản trình tự. Các phương pháp tiêu biểu của thể loại này bao gồm kiến trúc luồng đơn như như UNITER [2] và OSCAR [4], và hai luồng kiến trúc như LXMERT [1], ALBEF [6] và ERNIE-ViL [5]. Phương pháp này có thể tốt hơn nắm bắt mối liên hệ cơ bản giữa hình ảnh và văn bản cho các nhiệm vụ hiểu ngôn ngữ thị giác, trong khi nó cần phải mã hóa chung tất cả các cặp hình ảnh-văn bản có thể, dẫn đến suy luận tương đối chậm tốc độ.

Để cải thiện tốc độ suy luận, một số gần đây công việc như Pixel-BERT [3], E2E-VLP [19] và ViLT [10] loại bỏ bộ dò đối tượng phức tạp trong quá trình trích xuất tính năng và tiến hành học VL đầu cuối với các tính năng lưới dựa trên CNN và nhúng vá chiếu tuyến tính, tương ứng . Để kết hợp lợi ích của cả hai loại của kiến trúc, VLMo [20] thống nhất hơn nữa mô-đun mã hóa kép và mã hóa hợp nhất với chia sẻ hỗn hợp các chuyên gia phương thức Transformer. Trong công trình này, mPLUG giới thiệu một cơ chế hợp nhất đa phương thức mới với các kết nối bỏ qua đa phương thức, cho phép hợp nhất diễn ra ở các cấp độ khác nhau trong hệ thống phân cấp trừu tượng trên toàn bộ phương thức. Nó đạt được hiệu suất vượt trội trong hiệu quả và hiệu suất trên một phạm vi rộng Nhiệm vụ VL.

2.2 Bỏ qua kết nối

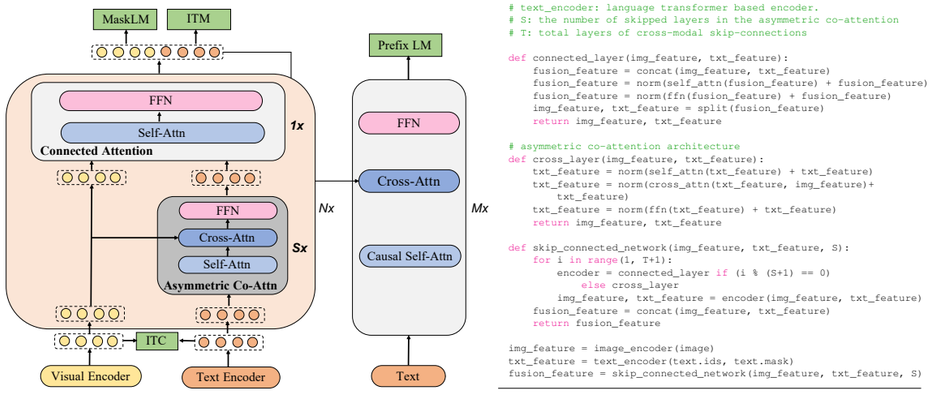
Kết nối bỏ qua là một kỹ thuật phổ biến để bỏ qua vấn đề bùng nổ hoặc biến mất của gradient tối ưu hóa mô hình trong mạng nơ-ron sâu, trong đó được sử dụng rộng rãi trong các kiến trúc CV và NLP như như ResNet [21] và Transformer [12]. Một loạt các các phương pháp kết nối bỏ qua đã được đề xuất trong những năm gần đây [22, 21, 12, 23, 24, 25]. ResNet [21] giới thiệu các kết nối tắt được tổng hợp giữa các các lớp khác nhau sử dụng ánh xạ danh tính đơn giản, trong khi mạng lưới đường cao tốc [22] thiết kế một cổng chuyển đổi chức năng kiểm soát sự cân bằng của đầu vào và đầu vào được chuyển đổi. DenseNet [23] thiết kế mới kiến trúc với các kết nối bỏ qua được nối tiếp, cho phép các lớp tiếp theo tái sử dụng tất cả biểu diễn ở giữa của các lớp trước đó. Lớp Chuẩn hóa và kết nối bỏ qua đệ quy là được sử dụng kết hợp với kết nối bỏ qua đơn giản để ổn định hơn nữa quá trình tối ưu hóa mô hình và kết hợp tốt hơn đầu vào đã chuyển đổi [12, 25]. Trong công trình này, mPLUG đề xuất một phương thức liên phương thức mới bỏ qua phương pháp kết nối để giải quyết vấn đề đa phương thức vấn đề hợp nhất, và kết hợp các nối tiếp bỏ qua kết nối và bỏ qua kết nối tổng hợp cho lựa chọn có nên tham dự tất cả các sự kiện được nối tiếp hay không biểu diễn các phương thức khác nhau hoặc chỉ tập trung về phần tương tác liên phương thức ở mỗi lớp.

3 mPHÍ CH CẤM

Trong phần này, trước tiên chúng tôi sẽ giới thiệu sản phẩm mới của chúng tôi kiến trúc mô hình với mô-đun chính của mạng lưới kết nối bỏ qua đa phương thức, và sau đó cung cấp chi tiết về mục tiêu đào tạo trước và cơ sở hạ tầng đào tạo có thể mở rộng.

3.1 Kiến trúc mô hình

Như thể hiện trong Hình 2, mPLUG bao gồm hai bộ mã hóa đơn phương thức cho hình ảnh và văn bản độc lập, một mạng lưới kết nối bỏ qua đa phương thức và một bộ giải mã để tạo văn bản. Để mô hình hóa tốt hơn thông tin thiên vị phương thức vốn có, trước tiên chúng tôi sử dụng hai bộ mã hóa đơn thức để mã hóa hình ảnh và văn bản riêng biệt. Tiếp theo [11, 26], chúng tôi sử dụng bộ chuyển đổi hình ảnh [27] trực tiếp trên các bản vá hình ảnh làm bộ mã hóa trực quan, thân thiện hơn với tính toán hơn là sử dụng các máy dò đối tượng được đào tạo trước cho hình ảnh trích xuất tính năng [8, 9]. Bộ mã hóa hình ảnh chia hình ảnh đầu vào thành các bản vá và mã hóa chúng như một chuỗi nhúng {vcls, v1, v2, ..., vM} với một mã thông báo [CLS] bổ sung . Văn bản đầu vào được đưa vào bộ mã hóa văn bản và được biểu diễn dưới dạng một chuỗi nhúng {lcls, l1, l2, ..., lN }, trong đó lcls là nhúng của mã thông báo [CLS] và được sử dụng để tóm tắt văn bản đầu vào. Sau đó, hình ảnh và các biểu diễn ngôn ngữ được đưa vào một phương thức chéo mạng kết nối bỏ qua, bao gồm nhiều khối hợp nhất kết nối bỏ qua. Trong mỗi khối hợp nhất kết nối bỏ qua, chúng tôi áp dụng hợp nhất chéo phương thức được kết nối cho mỗi S đồng chú ý không đối xứng các lớp trong đó S là giá trị bước cố định. Mục đích của mạng lưới này là tận dụng hiệu quả của sự kết hợp đa phương thức được kết nối và hiệu quả của sự chú ý đồng thời không đối xứng cho



Algorithm 1: Pseudocode of Cross-modal Skip-connected Network.

```
# image, text.ids, text.mask: paired (image, text) pairs.
# image_encoder: vision transformer based encoder.
# text_encoder: language transformer based encoder.
# S: the number of skipped layers in the asymmetric co-attention
# T: total layers of cross-modal skip-connections

def connected_layer(img_feature, txt_feature):
    fusion_feature = concat(img_feature, txt_feature)
    fusion_feature = norm(self_attn(fusion_feature) + fusion_feature)
    fusion_feature = norm(ffn(fusion_feature) + fusion_feature)
    img_feature, txt_feature = split(fusion_feature)
    return img_feature, txt_feature

# asymmetric co-attention architecture
def cross_layer(img_feature, txt_feature):
    txt_feature = norm(self_attn(txt_feature) + txt_feature)
    txt_feature = norm(cross_attn(txt_feature, img_feature)+
        txt_feature)
    txt_feature = norm(ffn(txt_feature) + txt_feature)
    return img_feature, txt_feature

def skip_connected_network(img_feature, txt_feature, S):
    for i in range(1, T+1):
        encoder = connected_layer if (i % (S+1) == 0)
            else cross_layer
        img_feature, txt_feature = encoder(img_feature, txt_feature)
    fusion_feature = concat(img_feature, txt_feature)
    return fusion_feature

img_feature = image_encoder(image)
txt_feature = text_encoder(text.ids, text.mask)
fusion_feature = skip_connected_network(img_feature, txt_feature, S)
```

Figure 2: The model architecture and objectives of mPLUG, which consists of two unimodal encoders for images and text separately, a cross-modal skip-connected network and a decoder for text generation. An image-text contrastive loss is first applied to align the unimodal representations from the visual encoder and text encoder. Then, we use a novel cross-modal skip-connected network to fuse the visual and linguistic representations effectively and efficiently. We adopt connected cross-modal fusion to every S *asymmetric co-attention* layers, where S is a fixed stride value. Based on the connected representation of the image and prefix sub-sequence, the decoder is trained with a prefix language modeling (Prefix LM) loss by generating the remaining caption.

enhanced cross-modal fusion in a recursive manner. Finally, the output cross-modal representations are fed into a transformer decoder for sequence to sequence learning, which equips mPLUG with both understanding and generation capabilities.

3.2 Cross-modal Skip-connected Network

The cross-modal skip-connected network consists of N skip-connected fusion blocks. In each skip-connected fusion block, we adopt *connected-attention* layer to each of S *asymmetric co-attention* layers where S is a fixed stride value. We first pass the text feature and image feature from unimodal encoders through the S asymmetric co-attention layers, and then connect the output text feature and image feature to one connected-attention layer. We repeat the skip-connected fusion block N times for the final connected image and text representation.

Specifically, the asymmetric co-attention is composed of the self-attention (SA) layer, cross-attention (CA) layer and the feed-forward network (FFN). The input text feature l^{n-1} is first fed to the self-attention layer, and then the visual feature v^{n-1} is injected into the text feature l^n_{SA} by the cross-attention layer which gives l^n_{CA}. The output of self-attention l^n_{SA} and cross-attention l^n_{SA} are added up and fed to the FFN layer for the visual-aware text representation l^n:

$$l^n_{SA} = LN(SA(l^{n-1}) + l^{n-1}) \quad (1)$$

$$l^n_{CA} = LN(CA(l^n_{SA}, v^{n-1}) + l^n_{SA}) \quad (2)$$

$$l^n = LN(FFN(l^n_{CA}) + l^n_{CA}) \quad (3)$$

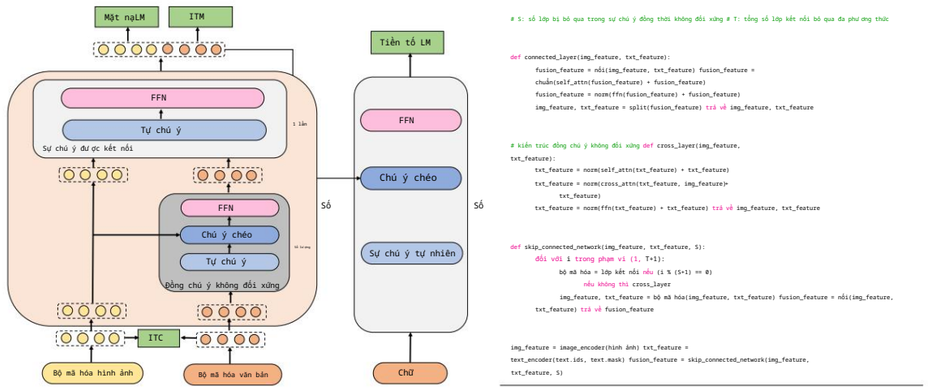
where LN is short for layer normalization.

The connected-attention layer is composed of the self-attention (SA) layer and the feed-forward network (FFN). We connect the image feature v^{n-1} and input text feature l^{n-1}, where l^{n-1} is the output of S asymmetric co-attention layers. The connected image and text feature $[v^{n-1}; l^{n-1}]$ are fed to the self-attention layer and FFN layer:

$$[v^n_{SA}; l^n_{SA}] = LN(SA([v^{n-1}; l^{n-1}]) + [v^{n-1}; l^{n-1}]) \quad (4)$$

$$[v^n; l^n] = LN(FFN([v^n_{SA}; l^n_{SA}]) + [v^n_{SA}; l^n_{SA}]) \quad (5)$$

Then $[v^n; l^n]$ is fed into the next cross-modal skip-connected network repeatedly to get the final connected image and text representation. Finally, the connected output is fed into a Transformer decoder for sequence to sequence learning.



Thuật toán 1: Mã giả của mạng kết nối bỏ qua đa phư ơng thức.

```
# image, text.ids, text.mask: các cặp (image, text) đư ợc ghép nối. # image_encoder: bộ mã hóa dựa
trên bộ chuyển đổi thị giác. # text_encoder: bộ mã hóa dựa trên bộ chuyển đổi ngôn ngữ.

# S: số lớp bị bỏ qua trong sự chú ý đồng thời không đối xứng # T: tổng số lớp kết nối bỏ qua đa phư ơng thức


def connected_layer(img_feature, txt_feature):
        fusion_feature = nối(img_feature, txt_feature) fusion_feature =
        chuẩn(self_attn(fusion_feature) + fusion_feature)
        fusion_feature = norm(ffn(fusion_feature) + fusion_feature)
        img_feature, txt_feature = split(fusion_feature) trả về img_feature, txt_feature


# kiến trúc đồng chú ý không đối xứng def cross_layer(img_feature,
txt_feature):
        txt_feature = norm(self_attn(txt_feature) + txt_feature)
        txt_feature = norm(cross_attn(txt_feature, img_feature)+
            txt_feature)
        txt_feature = norm(ffn(txt_feature) + txt_feature) trả về img_feature, txt_feature


def skip_connected_network(img_feature, txt_feature, S):
        đối với i trong phạm vi (1, T+1):
            bộ mã hóa = lớp kết nối nếu (i % (S+1) == 0)
                nếu không thì cross_layer
            img_feature, txt_feature = bộ mã hóa(img_feature, txt_feature) fusion_feature = nối(img_feature,
        txt_feature) trả về fusion_feature


img_feature = image_encoder(hình ảnh) txt_feature =
text_encoder(text.ids, text.mask) fusion_feature = skip_connected_network(img_feature,
txt_feature, S)
```

Hình 2: Kiến trúc mô hình và mục tiêu của mPLUG, bao gồm hai bộ mã hóa đơn thức cho hình ảnh và văn bản riêng biệt, một mạng bỏ qua kết nối chéo phư ơng thức và một bộ giải mã để tạo văn bản. Đầu tiên, một mất mát tư ơng phản hình ảnh-văn bản đư ợc áp dụng để căn chỉnh các biểu diễn đơn thức từ bộ mã hóa hình ảnh và bộ mã hóa văn bản. Sau đó, chúng tôi sử dụng một mạng bỏ qua kết nối chéo phư ơng thức mới để hợp nhất các biểu diễn hình ảnh và ngôn ngữ một cách hiệu quả và hiệu quả. Chúng tôi áp dụng hợp nhất liên phư ơng thức kết nối cho mọi lớp đồng chú ý bất đối xứng S, trong đó S là giá trị bư ớc cố định. Dựa trên biểu diễn kết nối của hình ảnh và chuỗi con tiền tố, bộ giải mã đư ợc đào tạo với mất mát mô hình ngôn ngữ tiền tố (Tiền tố LM) bằng cách tạo chú thích còn lại.

tăng cư ờng sự kết hợp đa phư ơng thức theo cách đệ quy. Cuối cùng, các biểu diễn đa phư ơng thức đầu ra đư ợc đư a vào bộ giải mã biến áp để học chuỗi, giúp mPLUG có cả khả năng hiểu và khả năng tạo.

3.2 Mạng kết nối bỏ qua đa phư ơng thức

Mạng kết nối bỏ qua đa phư ơng thức bao gồm N khối hợp nhất kết nối bỏ qua. Trong mỗi khối hợp nhất kết nối bỏ qua, chúng tôi áp dụng lớp chú ý đư ợc kết nối cho mỗi lớp đồng chú ý không đối xứng trong S , trong đó S là giá trị bư ớc cố định. Đầu tiên, chúng tôi truyền tính năng văn bản và tính năng hình ảnh từ bộ mã hóa đơn thức qua các lớp đồng chú ý bất đối xứng S , sau đó kết nối tính năng văn bản và tính năng hình ảnh đầu ra với một lớp chú ý đư ợc kết nối. Chúng tôi lặp lại khối hợp nhất kết nối bỏ qua N lần cho biểu diễn hình ảnh và văn bản đư ợc kết nối cuối cùng .

Cụ thể, sự đồng chú ý bất đối xứng bao gồm lớp tự chú ý (SA), lớp chú ý chéo (CA) và mạng truyền thẳng (FFN). Tính năng văn bản đầu vào l đầu tiên đư ợc đư a vào lớp tự chú ý, sau đó tính năng trực quan đư ợc đư a vào tính năng văn bản l n SA bởi lớp chú ý chéo, cung cấp l n Đầu ra CA. của sự tự chú ý l n và sự chú ý chéo l n SA SA lần n 1 là

đư ợc thêm vào và đư a vào lớp FFN để biểu diễn văn bản có nhận thức trực quan l n

$$l^n_{SA} = LN(SA(l^{n-1}) + l^{n-1}) \quad (1)$$

$$l^n_{CA} = LN(CA(l^n_{SA}, v^{n-1}) + l^n_{SA}) \quad (2)$$

$$l^n = LN(FFN(l^n_{CA}) + l^n_{CA}) \quad (3)$$

trong đó LN là viết tắt của chuẩn hóa lớp.

Lớp chú ý đư ợc kết nối bao gồm lớp tự chú ý (SA) và lớp truyền tiếp mạng (FFN). Chúng tôi kết nối tính năng hình ảnh v $^{n-1}$ và tính năng văn bản l $^{n-1}$, nơi tôi là ra-đầu vào l đặt của các lớp đồng chú ý không đối xứng S. Tính năng hình ảnh và văn bản [v $^{n-1}_{:tôi}$; l $^{n-1}_{:tôi}$] đư ợc cho ăn đư ợc kết nối [v với lớp tự chú ý và lớp FFN:

$$[v^n_{SA}; l^n_{SA}] = LN(SA([v^{n-1}_{:tôi}; l^{n-1}_{:tôi}]) + [v^{n-1}_{:tôi}; l^{n-1}_{:tôi}]) \quad (4)$$

$$[v^n; l^n] = LN(FFN([v^n_{SA}; l^n_{SA}]) + [v^n_{SA}; l^n_{SA}]) \quad (5)$$

Sau đó, [v n ; l n] đư ợc đư a vào mạng bỏ qua liên phư ơng thức tiếp theo nhiều lần để có đư ợc hình ảnh và biểu diễn văn bản đư ợc kết nối cuối cùng. Cuối cùng, đầu ra đư ợc kết nối đư ợc đư a vào bộ giải mã Transformer để học chuỗi sang chuỗi.

3.3 Pre-training Tasks

We perform four pre-training tasks including three understanding tasks (Image-Text Contrastive Learning, Image-Text Matching, Masked Language Modeling) and one generation task (Prefix Language Modeling). These pre-training tasks are optimized jointly.

Image-Text Contrastive (ITC): Following [6], we employ the task to align the image features and the text features from the unimodal encoders. Specifically, we calculate the softmax-normalized image-to-text and text-to-image similarity, and take two dynamic memory queues (text, image) to increase the number of negative examples as MoCo [28].

Image-Text Matching (ITM): This task aims to predict whether an image and a sentence match with each other on the cross-modal representation. We also select hard negative image-text pairs based on the contrastive text-image similarity as [6].

Masked Language Modeling (MLM): The task setup is basically the same as in BERT [29], where we randomly mask 15% of tokens in text and the model is asked to predict these masked words with the cross-modal representations.

Prefix Language Modeling (PrefixLM): This task aims to generate the caption given an image and predict the text segment subsequent to the cross-modal context as [30]. It optimizes a cross entropy loss by maximizing the likelihood of text in an autoregressive manner.

4 Distributed Learning on a Large Scale

Training a big model like mPLUG on large-scale datasets faces many efficiency challenges. We increase the throughput from the perspective of reducing memory usage and computation time, thereby accelerating the training of the model.

The memory usage during model training is mainly composed of two aspects: the static memory usage composed of parameters/optimizer states/gradients, etc., and the runtime memory usage caused by intermediate variables like activation values. For static memory overhead, we use the ZeRO [31] technique to partition parameters/optimizer states/-gradients into the entire data-parallel group, so that the static memory overhead of a single GPU can be approximately reduced to $1/N$, where N denotes the number of GPU cards. We use gradient checkpointing [32] for the runtime memory cost, which greatly reduces the runtime memory usage at the expense of increasing forward time by recomputing part of the activation values during backward pass without keeping them in memory.

To reduce the computation time, we use BF16 precision training. BF16 is a new data type supported by NVIDIA's new Ampere architecture GPU like A100. Compared with the previously widely used mixed-precision training of FP16 and FP32, BF16 has the same representation range as FP32, thereby reducing the risk of numerical overflow and ensuring model convergence stability, and at the same time has the same fast computing speed as FP16.

5 Experiments

5.1 Data & Setup

Following the previous work [6], we use the same pre-training dataset with 14M images with texts, which includes two in-domain datasets (MS COCO [36] and Visual Genome [37]), and three web out-domain datasets (Conceptual Captions [38], Conceptual 12M [39], SBU Captions [40].

We pretrain the model for 30 epochs with the total batch size of 1024 on 16 NVIDIA A100 GPUs. We use a 6-layer Transformer for both the text encoder and the cross-modal skip-connected network, and a 12-layer Transformer for the decoder. The text encoder is initialized using the first 6 layers of the BERT_{base} [29] model and the skip-connected network is initialized using the last 6 layers of the BERT_{base}. We initialize the visual encoder by CLIP-ViT [17] pretrained on 400M noisy image-text pairs. The visual transformer with ViT-B/16 is used as our base architecture, the one with ViT-L/14 as the large architecture. We use the AdamW [41] optimizer with a weight decay of 0.02. The learning rate is warmed-up to 1e-5 (ViT-B/16) and 1e-4 (BERT_{base}) for $\text{mPLUG}_{\text{ViT-B}}$, and 5e-6 (ViT-L/14) and 5e-5 (BERT_{base}) for $\text{mPLUG}_{\text{ViT-L}}$ in the first 1000 iterations, and decayed to 1e-6 following a cosine schedule. During pre-training, we take random image crops of resolution 256×256 (ViT-B/16)/224×224 (ViT-L/14) as input, and also apply RandAugment [42] to improve the generalization of vision encoders. For VQA and image captioning tasks, we do an additional continue pre-training on 4M image-text pairs. We increase the image resolution during finetuning. For image-text contrastive learning, the queue size is set as 65,536 and the momentum coefficient is set as 0.995.

3.3 Nhiệm vụ trư ớc khi đào tạo

Chúng tôi thực hiện bốn nhiệm vụ tiền đào tạo bao gồm ba nhiệm vụ hiểu (Học đối chiếu hình ảnh-văn bản, Ghép hình ảnh-văn bản, Mô hình hóa ngôn ngữ che giấu) và một nhiệm vụ tạo (Mô hình hóa ngôn ngữ tiền tố). Các nhiệm vụ tiền đào tạo này đư ợc tối ư u hóa chung.

Tư ơng phản hình ảnh-văn bản (ITC): Tiếp theo [6], chúng tôi sử dụng nhiệm vụ căn chỉnh các đặc điểm hình ảnh và các đặc điểm văn bản từ bộ mã hóa đơn thức.

Cụ thể, chúng tôi tính toán độ tư ơng đồng giữa hình ảnh với văn bản và văn bản với hình ảnh đư ợc chuẩn hóa bằng softmax và sử dụng hai hàng đợi bộ nhớ động (văn bản, hình ảnh) để tăng số lư ợng ví dụ tiêu cực như MoCo [28].

So khớp hình ảnh-văn bản (ITM): Nhiệm vụ này nhằm mục đích dự đoán liệu hình ảnh và câu có khớp với nhau trên biểu diễn đa phư ơng thức hay không.

Chúng tôi cũng chọn các cặp hình ảnh-văn bản tiêu cực cứng dựa trên sự tư ơng đồng văn bản-hình ảnh tư ơng phản như [6].

Mô hình hóa ngôn ngữ bị che giấu (MLM): Thiết lập nhiệm vụ về cơ bản giống như trong BERT [29], trong đó chúng tôi che giấu ngẫu nhiên 15% mã thông báo trong văn bản và mô hình đư ợc yêu cầu dự đoán các từ bị che giấu này bằng các biểu diễn đa phư ơng thức.

Mô hình hóa ngôn ngữ tiền tố (PrefixLM): Nhiệm vụ này nhằm mục đích tạo chú thích cho một hình ảnh và dự đoán phân đoạn văn bản tiếp theo ngữ cảnh đa phư ơng thức như [30]. Nó tối ư u hóa mất mát entropy chéo bằng cách tối đa hóa khả năng của văn bản theo cách tự hồi quy.

4 Học tập phân tán trên quy mô lớn

Việc đào tạo một mô hình lớn như mPLUG trên các tập dữ liệu quy mô lớn phải đối mặt với nhiều thách thức về hiệu quả. Chúng tôi tăng thông lư ợng theo quan điểm giảm sử dụng bộ nhớ và thời gian tính toán, do đó đẩy nhanh quá trình đào tạo mô hình.

Việc sử dụng bộ nhớ trong quá trình đào tạo mô hình chủ yếu bao gồm hai khía cạnh: việc sử dụng bộ nhớ tĩnh bao gồm các tham số/trạng thái tối ư u hóa/gradient , v.v. và việc sử dụng bộ nhớ thời gian chạy do các biến trung gian như giá trị kích hoạt gây ra. Đối với chi phí bộ nhớ tĩnh, chúng tôi sử dụng kỹ thuật ZeRO [31] để phân vùng các tham số/trạng thái tối ư u hóa/ gradient thành toàn bộ nhóm dữ liệu song song, do đó chi phí bộ nhớ tĩnh của một GPU đơn có thể giảm xuống còn khoảng 1/N, trong đó N biểu thị số lư ợng thẻ GPU. Chúng tôi sử dụng kiểm tra điểm kiểm tra gradient [32] cho chi phí bộ nhớ thời gian chạy, giúp giảm đáng kể việc sử dụng bộ nhớ thời gian chạy tại chi phí tăng thời gian chuyển tiếp bằng cách tính toán lại một phần giá trị kích hoạt trong quá trình chuyển ngư ợc mà không lư u chúng trong bộ nhớ.

Để giảm thời gian tính toán, chúng tôi sử dụng huấn luyện độ chính xác BF16. BF16 là kiểu dữ liệu mới đư ợc hỗ trợ bởi GPU kiến trúc Ampere mới của NVIDIA như A100. So với huấn luyện độ chính xác hỗn hợp đư ợc sử dụng rộng rãi trư ớc đây của FP16 và FP32, BF16 có cùng phạm vi biểu diễn như FP32, do đó giảm nguy cơ tràn số và đảm bảo tính ổn định hội tụ của mô hình, đồng thời có cùng tốc độ tính toán nhanh như FP16.

5 Thí nghiệm

5.1 Dữ liệu & Thiết lập

Tiếp theo công trình trư ớc đó [6], chúng tôi sử dụng cùng một tập dữ liệu tiền đào tạo với 14M hình ảnh có văn bản, bao gồm hai tập dữ liệu trong miền (MS COCO [36] và Visual Genome [37]) và ba tập dữ liệu ngoài miền web (Chú thích khái niệm [38], Chú thích khái niệm 12M [39], Chú thích SBU [40].

Chúng tôi đào tạo trư ớc mô hình trong 30 kỷ nguyên với tổng kích thư ớc lô là 1024 trên 16 GPU NVIDIA A100.

Chúng tôi sử dụng một Transformer 6 lớp cho cả bộ mã hóa văn bản và mạng bỏ qua kết nối chéo phư ơng thức, và một Transformer 12 lớp cho bộ giải mã. Bộ mã hóa văn bản đư ợc khởi tạo bằng cách sử dụng 6 lớp đầu tiên của mô hình BERTbase [29] và mạng bỏ qua kết nối đư ợc khởi tạo bằng cách sử dụng 6 lớp cuối cùng của BERTbase . Chúng tôi khởi tạo bộ mã hóa trực quan bằng CLIP-ViT [17] đư ợc đào tạo trư ớc trên 400M cặp hình ảnh-văn bản nhiễu. Bộ biến đổi trực quan với ViT-B/16 đư ợc sử dụng làm kiến trúc cơ sở của chúng tôi, một bộ biến đổi với ViT- L/14 là kiến trúc lớn. Chúng tôi sử dụng trình tối ư u hóa AdamW [41] với trọng số suy giảm là 0,02. Tốc độ học đư ợc làm ấm lên đến 1e-5 (ViT-B/16) và 1e-4 (BERTbase) cho $\text{mPLUG}_{\text{ViT-B}}$ L/14) và 5e-5 (BERTbase) cho $\text{mPLUG}_{\text{ViT-L}}$ trong 1000 lần lặp đầu tiên và giảm , và 5e-6 (ViT-xuống 1e-6 theo lịch trình cosin. Trong quá trình đào tạo trư ớc, chúng tôi lấy các hình ảnh cắt ngẫu nhiên có độ phân giải 256×256 (ViT- B/16)/224×224 (ViT-L/14) làm đầu vào và cũng áp dụng RandAugment [42] để cải thiện khả năng khái quát hóa của bộ mã hóa thị giác. Đối với các tác vụ VQA và chú thích hình ảnh, chúng tôi thực hiện thêm một lần đào tạo trư ớc liên tục trên 4M cặp hình ảnh-văn bản. Chúng tôi tăng độ phân giải hình ảnh trong quá trình tinh chỉnh. Đối với việc học tư ơng phản hình ảnh-văn bản , kích thư ớc hàng đợi đư ợc đặt là 65.536 và hệ số động lư ợng đư ợc đặt là 0,995.

Models	Data	COCO Caption								NoCaps	
		Cross-entropy Optimization				CIDEr Optimization					
		B@4	M	C	S	B@4	M	C	S	C	S
Encoder-Decoder	CC12M	-	-	110.9	-	-	-	-	-	90.2	12.1
E2E-VLP [19]	4M	36.2	-	117.3	-	-	-	-	-	-	-
VinVL [9]	5.65M	38.5	30.4	130.8	23.4	41.0	31.1	140.9	25.2	97.3	13.8
OSCAR [4]	6.5M	-	-	-	-	41.7	30.6	140.0	24.5	83.4	11.4
SimVLM$_{large}$ [7]	1.8B	40.3	**33.4**	**142.6**	**24.7**	-	-	-	-	-	-
LEMON$_{large}$ [33]	200M	40.6	30.4	135.7	23.5	42.3	31.2	144.3	25.3	113.4	**15.0**
BLIP [34]	129M	40.4	-	136.7	-	-	-	-	-	113.2	14.8
OFA [35]	18M	-	-	-	-	43.5	31.9	149.6	**26.1**	-	-
mPLUG	14M	**43.1**	31.4	141.0	24.2	**46.5**	**32.0**	**155.1**	26.0	**114.8**	14.8

Table 1: Evaluation Results on COCO Caption "Karpathy" test split and NoCaps validation set. B@4: BLEU@4, M: METEOR, C: CIDEr, S: SPICE.

Models	Data	Test-dev	Test-std
Pretrained on COCO, VG, SBU and CC datasets			
VLBERT [43]	4M	71.16	-
E2E-VLP [19]	4M	73.25	73.67
VL-T5 [44]	4M	-	71.30
UNITER[2]	4M	72.70	72.91
OSCAR[4]	4M	73.16	73.44
CLIP-ViL[26]	4M	76.48	76.94
METER[11]	4M	77.68	77.64
ALBEF[6]	4M	74.54	74.70
mPLUG$_{\text{ViT-B}}$	4M	**77.94**	**77.96**
Models Pretrained on More Data			
ALBEF [6]	14M	75.84	76.04
BLIP [34]	129M	78.25	78.32
SimVLM [7]	1.8B	80.03	80.34
Florence [45]	0.9B	80.16	80.36
OFA [35]	18M	79.87	80.02
VLMo [20]	-	79.94	79.98
mPLUG$_{\text{ViT-B}}$	14M	79.79	79.81
mPLUG$_{\text{ViT-L}}$	14M	**81.27**	**81.26**

Table 2: Evaluation Results on VQA test set.

5.2 Evaluation on Vision-Language Tasks

We compare our pre-trained model against other VLP models on the six downstream V+L tasks. We introduce each task and our fine-tuning strategy below. Details of the datasets and fine-tuning hyperparameters are in Appendix.

5.2.1 Visual Question Answering

The VQA task [14] requires the model to answer natural language questions given an image. Most methods [1, 20, 4, 7] deal with visual question answering tasks as multi-label classification on predefined answer sets. This strategy achieves strong performance, but it is not suitable for real-world open scenarios. We treat VQA as an answer generation task and directly use unconstrained open-vocab generation during inference, which is different from constrained close-vocab generation models [6, 35]. Following [4, 35], we concatenate the question with the object labels and OCR tokens extracted from image. As shown in Table 2, mPLUG achieves 81.27 on Test-std split and outperforms the SOTA models including SimVLM and Florence, which use 100X and 60X more pre-training image-text pairs, respectively. Based on the same 4M pre-training data, mPLUG outperforms CLIP-ViL and METER, which also use CLIP [17] as the visual encoder. Besides, under the same settings, mPLUG always significantly outperforms ALBEF and BLIP which only rely on co-attention from images to text for cross-modal fusion. The gain can derive from the network design of cross-modal skip-connections specifically for information asymmetry of the two modalities. Neither ALBEF nor BLIP addresses this problem well, with bias towards the language modality.

5.2.2 Image Captioning

The image captioning task requires a model to generate an appropriate and fluent caption for a given image. We evaluate image captioning on two datasets COCO Caption [47] and NoCaps [48]. mPLUG finetuned with training data of COCO Caption is tested on both of the datasets. We train

Mô hình	Dữ liệu	Chú thích COCO								Không có chữ viết hoa	
		Tối ưu hóa entropy chéo				Tối ưu hóa CIDEr					
		B@4	M	C	S	B@4	M	C	S	C	S
Bộ mã hóa-giải mã	CC12M	-	-	110,9	-	-	-	-	-	90,2	12,1
E2E-VLP [19]	4M	36,2	-	117,3	-	-	-	-	-	-	-
VinVL [9]	5,65M	38,5	30,4	130,8	23,4	41,0	31,1	140,9	25,2	97,3	13,8
OSCAR [4]	6,5 triệu	-	-	-	-	41,7	30,6	140,0	24,5	83,4	11,4
SimVLMlarge [7]	1.8B	40.3	33.4	142.6	24.7	LEMONlarge [33]				-	-
	200M	40.6	30.4	135.7	23.5	42.3	31.2	144.3	25.3	113.4	15.0
BLIP [34]	129M	40,4	-	136,7	-	OFA [35]				-	-
	18 triệu	-	-	-	-	43,5	31,9	149,6	26,1	113.2	14.8
mPLUG	14M	43,1	31,4	141,0	24,2	46,5	32,0	155,1	26,0	114,8	14,8

Bảng 1: Kết quả đánh giá trên COCO Caption "Karpathy" test split và NoCaps validation set. B@4: BLEU@4, M: METEOR, C: CIDER, S: SPICE.

Mô hình	Kiểm tra dữ liệu-phát triển	Kiểm tra-std
Được đào tạo trước trên các tập dữ liệu COCO, VG, SBU và CC		
VLBERT [43]	4M	71,16
E2E-VLP [19]	4M	73,25
VL-T5 [44]	4M	-
UNITER[2]	4M	72,70
OSCAR[4]	4M	73,16
CLIP-ViL[26]	4M	76,48
METER[11]	4M	77,68
ALBEF[6]	4M	74,54
mPLUGViT-B	4M	77,94
Các mô hình được đào tạo trước trên nhiều dữ liệu hơn		
ALBEF [6]	14M	75,84
BLIP [34]	129M	78,25
SimVLM [7]	1,8B	80,03
Florence [45]	0,9B	80,16
OFA [35]	18M	79,87
VLMo [20]	-	79,94
mPLUGViT-B	14M	79,79
mPLUGViT-L	14M	81.27

Bảng 2: Kết quả đánh giá trên bộ kiểm tra VQA.

5.2 Đánh giá về các nhiệm vụ ngôn ngữ thị giác

Chúng tôi so sánh mô hình được đào tạo trước của chúng tôi với các mô hình khác
Mô hình VLP trên sáu tác vụ V+L hạ lưu.

Chúng tôi giới thiệu từng nhiệm vụ và chiến lược tinh chỉnh
của chúng tôi bên dưới. Chi tiết về các tập dữ liệu và tinh chỉnh
siêu tham số được nêu trong Phụ lục.

5.2.1 Trả lời câu hỏi trực quan

Nhiệm vụ VQA [14] yêu cầu mô hình phải trả lời
câu hỏi ngôn ngữ tự nhiên cho một hình ảnh. Hầu hết
phương pháp [1, 20, 4, 7] xử lý các nhiệm vụ trả lời câu hỏi trực
quan như phân loại đa nhãn trên các tập câu trả lời được xác định
trước. Chiến lược này đạt được hiệu quả mạnh mẽ
hiệu suất, nhưng nó không phù hợp với thế giới thực
các kịch bản mở. Chúng tôi xử lý VQA như một nhiệm vụ tạo
câu trả lời và sử dụng trực tiếp việc tạo từ vựng mở không
bị ràng buộc trong quá trình suy luận, điều này khác với các
mô hình tạo từ vựng đóng bị ràng buộc [6, 35]. Theo [4, 35],
chúng tôi nối các
câu hỏi với các nhãn đối tượng và mã thông báo OCR được
trích xuất từ hình ảnh. Như thể hiện trong Bảng 2, mPLUG
đạt 81,27 trên Test-std split và vượt trội
các mô hình SOTA bao gồm SimVLM và Flo- rence, sử
dụng nhiều hơn 100X và 60X quá trình đào tạo trước
cặp hình ảnh-văn bản, tương ứng. Dựa trên cùng một
Dữ liệu đào tạo trước 4M, mPLUG vượt trội
hơn CLIP- ViL và METER, cũng sử dụng CLIP [17] làm
bộ mã hóa hình ảnh. Bên cạnh đó, trong cùng một thiết lập,
mPLUG luôn vượt trội hơn ALBEF một cách đáng kể
và BLIP chỉ dựa vào sự chú ý chung từ
hình ảnh thành văn bản để kết hợp đa phương thức. Sự gia tăng
có thể bắt nguồn từ thiết kế mạng lưới của đa phương thức
kết nối bỏ qua đặc biệt cho sự bất đối xứng thông tin của
hai phương thức. Cả ALBEF và
BLIP giải quyết vấn đề này rất tốt, thiên về phương
thức ngôn ngữ.

5.2.2 Chú thích hình ảnh

Nhiệm vụ chú thích hình ảnh yêu cầu một mô hình để
tạo ra một chú thích phù hợp và trôi chảy cho một
hình ảnh đã cho. Chúng tôi đánh giá chú thích hình ảnh trên
hai tập dữ liệu COCO Caption [47] và NoCaps [48].
mPLUG được tinh chỉnh với dữ liệu đào tạo của COCO
Chú thích được thử nghiệm trên cả hai tập dữ liệu. Chúng tôi đào tạo

Models	# Pretrain data	MSCOCO (5K test set)						Flickr30K (1K test set)					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
E2E-VLP [19]	4M	-	-	-	-	-	-	86.2	97.5	98.92	73.6	92.4	96.0
UNITER [2]	4M	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
OSCAR [4]	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
UNIMO [46]	4M	-	-	-	-	-	-	89.4	98.9	99.8	78.0	94.2	97.1
VLMo [20]	4M	78.2	94.4	97.4	60.6	84.4	91.0	95.3	99.9	100.0	84.5	97.3	98.6
ALIGN [18]	1.8B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100.0	84.9	97.4	98.6
ALBEF [6]	14M	77.6	94.3	97.2	60.7	84.3	90.5	95.9	99.8	100.0	85.6	97.5	98.9
Florence [45]	0.9B	81.8	95.2	-	63.2	85.7	-	97.2	99.9	-	87.9	98.1	-
BLIP [34]	14M	80.6	95.2	97.6	63.1	85.3	91.1	96.6	99.8	100.0	87.2	97.5	98.8
BLIP [34]	129M	82.4	95.4	97.9	65.1	86.3	91.8	97.4	99.8	99.9	87.6	97.7	99.0
mPLUG	14M	**82.8**	**96.1**	**98.3**	**65.8**	**87.3**	**92.6**	**97.6**	**100.0**	**100.0**	**88.4**	**97.9**	**99.1**

Table 3: Image-text retrieval results on Flickr30K and COCO datasets.

mPLUG on the MS COCO Caption and test on the same Karpathy split [4, 7] and NoCaps validation set. Following [4, 35], we first fine-tune mPLUG with cross-entropy loss and then with CIDEr optimization [49] for extra 5 epochs. As shown in Table 1, mPLUG with only 14M pre-training images can outperform the SOTA models including LEMON and SimVLM on both COCO Caption and Nocaps datasets, which uses more than $10X$ and $100X$ pre-training data, respectively. For the COCO Caption, mPLUG performs the best on CIDEr evaluation and surpasses the SOTA model by a large margin of 5.5 on Karpathy test set. We use the best checkpoint on COCO Caption and predict on the Nocaps validation set directly.

5.2.3 Image-Text Retrieval

We conduct experiments for both image-to-text retrieval (TR) and text-to-image retrieval (IR) on COCO [36] and Flickr30K [53] datasets. Following [6, 34], we jointly optimize the ITC loss and the ITM loss during fine-tuning. During inference, we first select top-k candidates by computing the dot-product similarity between the image and text encoder features, and then rerank the selected candidates based on their ITM scores. We set $k = 256$ for COCO and $k = 128$ for Flickr30K. As shown in Table 3, mPLUG outperforms all existing methods on both datasets. Using 14M images, mPLUG achieves better performance than BLIP with 129M and Florence with 0.9B pre-training data. Using the same 14M pre-training images, mPLUG substantially outperforms the previous best model BLIP by +2.7% in TR recall@1 on COCO and +1.0 % in TR recall@1 on Flickr30K.

5.2.4 Visual Grounding

Given a query in plain text and an image, visual grounding requires models to localize the referred object in the image. Instead of regressing the bounding boxes directly, we concatenate visual features and attended textual features and feed them into the decoder to predict the coordinates. Table 4 shows that mPLUG outperforms all the SOTA methods. We observe that in RefCOCO testB the images often contain arbitrary objects and in RecCOCOg test-u the expressions are longer than other datasets. Compared with the previous best model OFA, mPLUG achieves 3.16% absolute improvement on RefCOCO testB and 1.22% absolute improvement on RefCOCOg test-u. It demonstrates that mPLUG learns better multi-modal interaction from cross-modal skip-connections and is better at handling complex images and long queries.

5.2.5 Visual Reasoning

We consider two datasets for visual reasoning: NLVR2 [54] and SNLI-VE [55]. The NLVR2 [54] task requires the model to predict whether a sentence describes a pair of images. Following [34], we use two cross-attention layers to process the two input images, and their outputs are merged and fed to the FFN. An MLP classifier is then applied on the output embedding of the language [CLS] token. The SNLI-VE [55] task requires the model to evaluate how the given image and text are semantically correlated, i.e., entailment, neutral, or contradiction. Following [35], the image premise, text premise and text hypothesis are fed to the encoder. While we remove the decoder, and only use the encoder modules for three-way classification, which can save nearly half of the total computation cost. We predict the class probabilities using

Mô hình	# Dữ liệu huấn luyện trư ớc	MSCOCO (bộ đề thi 5K) TR						Flickr30K (bộ thử nghiệm 1K) TR					
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
E2E-VLP [19]	4M	-	-	-	-	-	-	86,2	97,5	98,92	73,6	92,4	96,0
UNITER [2]	4M	65,7	88,6	93,8	52,9	79,9	88,0	87,3	98,0	99,2	75,6	94,1	96,8
OSCAR [4]	4M	70,0	91,1	95,5	54,0	80,8	88,5	-	-	-	-	-	-
UNIMO [46]	4M	-	-	-	-	-	-	89,4	98,9	99,8	78,0	94,2	97,1
VLMo [20]	4M	78,2	94,4	97,4	60,6	84,4	91,0	95,3	99,9	100,0	84,5	97,3	98,6
CĂN CHỈNH [18]	1,8B	77,0	93,5	96,9	59,9	83,3	89,8	95,3	99,8	100,0	84,9	97,4	98,6
ALBEF [6]	14M	77,6	94,3	97,2	60,7	84,3	90,5	95,9	99,8	100,0	85,6	97,5	98,9
Florentia [45]	0,9B	81,8	95,2	-	63,2	85,7	-	97,2	99,9	-	87,9	98,1	-
BLIP [34]	14M	80,6	95,2	97,6	63,1	85,3	91,1	96,6	99,8	100,0	87,2	97,5	98,8
BLIP [34]	129M	82,4	95,4	97,9	65,1	86,3	91,8	97,4	99,8	99,9	87,6	97,7	99,0
mPLUG	14M	82,8	96,1	98,3	65,8	87,3	92,6	97,6	100,0	100,0	88,4	97,9	99,1

Bảng 3: Kết quả truy xuất hình ảnh-văn bản trên tập dữ liệu Flickr30K và COCO.

mPLUG trên MS COCO Chú thích và thử nghiệm trên
cùng một bộ phân tách Karpathy [4, 7] và bộ xác thực NoCaps.
Tiếp theo [4, 35], trư ớc tiên chúng tôi tinh chỉnh
mPLUG với mất mát entropy chéo và sau đó với
Tối ư u hóa CIDEr [49] cho 5 kỷ nguyên bổ sung. Như
đư ợc thể hiện trong Bảng 1, mPLUG chỉ với 14M hình ảnh đư ợc
đào tạo trư ớc có thể vư ợt trội hơn các mô hình SOTA
bao gồm LEMON và SimVLM trên cả COCO
Bộ dữ liệu chú thích và Nocaps, sử dụng nhiều hơn
Dữ liệu đào tạo trư ớc 10X và 100X tư ơng ứng. Đối với
COCO Caption, mPLUG hoạt động tốt nhất trên
Đánh giá CIDEr và vư ợt qua mô hình SOTA
với biên độ lớn là 5,5 trên bộ kiểm tra Karpathy. Chúng tôi
sử dụng điểm kiểm tra tốt nhất trên COCO Caption và
dự đoán trực tiếp trên tập xác thực Nocaps.

5.2.3 Truy xuất hình ảnh-văn bản

Chúng tôi tiến hành thí nghiệm cho cả hình ảnh thành văn bản
truy xuất (TR) và truy xuất văn bản thành hình ảnh (IR) trên
COCO [36] và Flickr30K [53] tập dữ liệu. Tiếp theo
[6, 34], chúng tôi cùng nhau tối ư u hóa tổn thất ITC và
mất mát ITM trong quá trình tinh chỉnh. Trong quá trình suy luận,
đầu tiên chúng tôi chọn ứng viên top-k bằng cách tính toán
sự tư ơng đồng giữa hình ảnh và văn bản
các tính năng mã hóa, sau đó xếp hạng lại các ứng viên đư ợc
chọn dựa trên điểm ITM của họ. Chúng tôi đặt k = 256
cho COCO và k = 128 cho Flickr30K. Như đã hiển thị
trong Bảng 3, mPLUG vư ợt trội hơn tất cả các phư ơng pháp
hiện có trên cả hai tập dữ liệu. Sử dụng 14M hình ảnh, mPLUG
đạt hiệu suất tốt hơn BLIP với 129M
và Florence với dữ liệu đào tạo trư ớc 0,9B. Sử dụng
cùng 14M hình ảnh tiền đào tạo, mPLUG vư ợt trội hơn đáng kể
so với mô hình BLIP tốt nhất trư ớc đó
tăng +2,7% trong TR recall@1 trên COCO và +1,0% trong
TR recall@1 trên Flickr30K.

5.2.4 Cơ sở thị giác

Đư a ra một truy vấn bằng văn bản thuần túy và một hình ảnh, trực quan
việc tiếp đất đòi hỏi các mô hình phải định vị đư ợc tham chiếu
đối tư ợng trong hình ảnh. Thay vì hồi quy
hộp giới hạn trực tiếp, chúng tôi nối các đặc điểm trực quan
và các đặc điểm văn bản đư ợc chú ý và đư a chúng vào
vào bộ giải mã để dự đoán tọa độ. Bảng
4 cho thấy mPLUG vư ợt trội hơn tất cả các SOTA
phư ơng pháp. Chúng tôi quan sát thấy rằng trong thử nghiệm RefCOCO B
hình ảnh thư ờng chứa các đối tư ợng tùy ý và trong Rec-
COCOg test-u các biểu thức dài hơn các biểu thức khác
bộ dữ liệu. So với mô hình tốt nhất trư ớc đó
OFA, mPLUG đạt đư ợc sự cải thiện tuyệt đối 3,16% trên thử
nghiệm RefCOCO B và sự cải thiện tuyệt đối 1,22%
cải tiến trên thử nghiệm RefCOCOg-u. Nó chứng minh
mPLUG học tư ơng tác đa phư ơng thức tốt hơn
từ các kết nối bỏ qua đa phư ơng thức và tốt hơn ở
xử lý hình ảnh phức tạp và truy vấn dài.

5.2.5 Lý luận trực quan

Chúng tôi xem xét hai tập dữ liệu để lý luận trực quan:
NLVR2 [54] và SNLI-VE [55]. NLVR2 [54]
nhiệm vụ yêu cầu mô hình dự đoán xem một câu có mô tả một cặp
hình ảnh hay không. Sau [34],
chúng tôi sử dụng hai lớp chú ý chéo để xử lý
hai hình ảnh đầu vào và đầu ra của chúng đư ợc hợp nhất và
đư ợc đư a vào FFN. Sau đó, một bộ phân loại MLP đư ợc áp dụng
về nhúng đầu ra của ngôn ngữ [CLS]
token. Nhiệm vụ SNLI-VE [55] yêu cầu mô hình
để đánh giá cách hình ảnh và văn bản đư ợc đư a ra có mối
tư ơng quan ngữ nghĩa như thế nào, tức là sự hàm ý, trung lập hoặc
mâu thuẫn. Tiếp theo [35], tiền đề hình ảnh,
tiền đề văn bản và giả thuyết văn bản đư ợc đư a vào bộ mã
hóa . Trong khi chúng ta loại bỏ bộ giải mã và chỉ sử dụng
các mô-đun mã hóa cho phân loại ba chiều,
có thể tiết kiệm gần một nửa tổng chi phí tính toán. Chúng
tôi dự đoán xác suất lớp bằng cách sử dụng

Model	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val-u	test-u
VLBERT [43]	-	-	-	72.59	78/57	62.30	-	-
UNITER [2]	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77
VILLA [50]	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
MDETR [51]	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
UNICORN [52]	88.29	90.42	83.06	80.30	85.05	71.88	83.44	83.93
OFA [35]	90.05	92.93	85.26	84.49	90.10	77.77	84.54	85.20
mPLUG	**92.40**	**94.51**	**88.42**	**86.02**	**90.17**	**78.17**	**85.88**	**86.42**

Table 4: Visual grounding results (Acc@0.5) on ReferCOCO, ReferCOCO+, and ReferCOCOg.

Model	NLVR2		SNLI-VE	
	dev	test-P	dev	test
LXMERT[1]	74.90	74.50	-	-
VL-T5[44]	-	73.6	-	-
UNITER[2]	79.12	79.98	79.39	79.38
CLIP-ViL[26]	-	-	80.61	80.20
METER[11]	82.33	83.05	80.86	81.19
UNIMO[46]	-	-	81.11	80.63
ALBEF[6]	82.55	83.14	80.80	80.91
BLIP[34]	82.67	82.30	-	-
SimVLM$_{large}$[7]	84.13	84.84	85.68	85.62
VLMo[20]	**85.64**	**86.86**	-	-
OFA[35]	-	-	**90.30**	**90.20**
mPLUG	84.58	84.95	89.45	89.29

Table 5: Evaluation Results on NLVR2 and SNLI-VE.

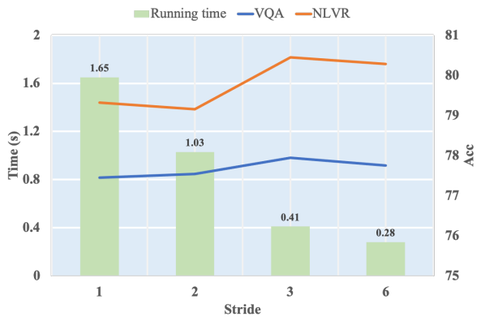


Figure 3: Results w.r.t different stride values in cross-modal skip-connected network on running time and performance of VQA test-dev and NLVR2 test-P, where the running time is the total forward time of 100 samples.

the multimodal encoder's output representation of the language [CLS] token. As shown in Table 5, mPLUG can obtain competitive performances to the SOTA models [1] in both visual reasoning tasks, and even outperform SimVLM [7] and BLIP [34], which use far more pre-training data.

5.3 Effectiveness and Efficiency

To validate the effectiveness and efficiency of our proposed cross-modal skip-connected network, we conduct in-depth analysis on different stride values and various cross-modal fusion methods.

5.3.1 Analysis of Stride for Skip

The stride S is the key factor to control the effectiveness and efficiency tradeoff. Therefore, we further compare the running time and performance of different stride value S in cross-modal skip-connected network on VQA and NLVR2 tasks. Specifically, we test four different stride values, which can be divisible by the total number of cross-modal fusion layers. The model is chosen as mPLUG$_{\text{ViT-B}}$ and all the other experiment settings are kept the same. As shown in Figure 3, we can see that the larger S is, the more efficient cross-modal fusion is, where the running time can be largely reduced from skipping the vision co-attention layers by $5X$ times from $S = 1$ to $S = 6$. The performances of mPLUG on both datasets gradually increases when $S = 3$, and slightly decreases later on. Compared with $S = 3$, mPLUG can achieve comparable performance at $S = 6$, while speeding up by nearly 30%. Therefore, we set $S = 6$ on mPLUG$_{\text{ViT-L}}$ for faster pre-training.

5.3.2 Analysis of Cross-modal Fusion

We compare the effectiveness and efficiency of different cross-modal fusion variants in terms of run-

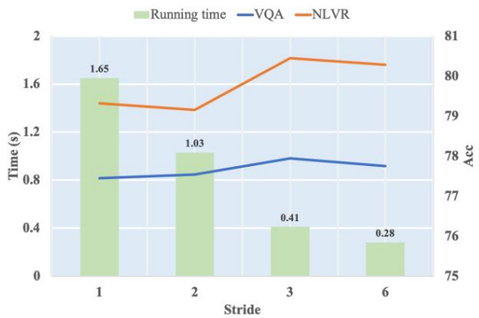
[1]The SOTA models such as OFA and VLMo both add large-scale text-only and image-only pre-training data for improving the reasoning ability.

Ngư ời mẫu	RefCOCO			Tham khảoCOCO+			Tham khảoCOCOg	
	giá trị	testA	testB	giá trị	testA	testB	val-u	test-u
VLBERT [43]	-	-	-	72,59	78/57	62,30	-	-
UNITER [2]	81,41	87,04	74,17	75,90	81,45	66,70	74,86	75,77
BIỆT THỰ [50]	82,39	87,48	74,84	76,17	81,54	66,84	76,18	76,71
MDETR [51]	86,75	89,58	81,41	79,52	84,09	70,62	81,64	80,89
KỲ LÂN [52]	88,29	90,42	83,06	80,30	85,05	71,88	83,44	83,93
OFA [35]	90,05	92,93	85,26	84,49	90,10	77,77	84,54	85,20
mPLUG	92,40	94,51	88,42	86,02	90,17	78,17	85,88	86,42

Bảng 4: Kết quả tiếp địa trực quan (Acc@0,5) trên ReferCOCO, ReferCOCO+ và ReferCOCOg.

Ngư ời mẫu	NLVR2		SNLI-VE	
	kiểm tra phát triển	P kiểm tra phát triển		
LXMERT[1]	74,90	74,50	-	-
VL-T5[44]	-	73,6	-	-
UNITER[2]	79,12	79,98	79,39	79,38
CLIP-ViL[26]	-	-	80,61	80,20
MÉT[11]	82,33	83,05	80,86	81,19
ĐỘC QUYỀN[46]	-	-	81,11	80,63
ALBEF[6]	82,55	83,14	80,80	80,91
BLIP[34]	82,67	82,30	-	-
SimVLMlarge[7]	84,13	84,84	85,68	85,62
VLMo[20]	85,64	86,86	-	-
OFA[35]	-	-	90.30	90.20
mPLUG	84,58	84,95	89,45	89,29

Bảng 5: Kết quả đánh giá trên NLVR2 và SNLI-VE.



Hình 3: Kết quả liên quan đến các giá trị bư ớc tiến khác nhau trong mạng kết nối bỏ qua đa phư ơng thức về thời gian chạy và hiệu suất của VQA test-dev và NLVR2 test-P, trong đó thời gian chạy là tổng thời gian chuyển tiếp của 100 mẫu.

biểu diễn đầu ra của bộ mã hóa đa phư ơng thức mã thông báo ngôn ngữ [CLS]. Như đư ợc hiển thị trong Bảng 5, mPLUG có thể đạt đư ợc hiệu suất cạnh tranh để các mô hình SOTA trong[1] cả hai nhiệm vụ lý luận trực quan, và thậm chí còn vư ợt trội hơn SimVLM [7] và BLIP [34], sử dụng nhiều dữ liệu đào tạo trư ớc hơn.

5.3 Hiệu quả và Hiệu suất

Để xác nhận tính hiệu quả và hiệu suất của chúng tôi đề xuất mạng lư ới kết nối bỏ qua đa phư ơng thức, chúng tôi tiến hành phân tích chuyên sâu về các giá trị sải chân khác nhau và nhiều phư ơng pháp kết hợp đa phư ơng thức khác nhau.

5.3.1 Phân tích bư ớc tiến để bỏ qua

Bư ớc tiến S là yếu tố chính để kiểm soát sự đánh đổi hiệu quả và hiệu suất. Do đó, chúng tôi tiếp tục so sánh thời gian chạy và hiệu suất của diff- giá trị bư ớc tiến khác nhau S trong kết nối bỏ qua đa phư ơng thức mạng lư ới trên các nhiệm vụ VQA và NLVR2. Cụ thể, chúng tôi kiểm tra bốn giá trị sải chân khác nhau, có thể là chia hết cho tổng số lớp hợp nhất đa phư ơng thức. Mô hình đư ợc chọn là mPLUGViT-B và tất cả các thiết lập thử nghiệm khác đư ợc giữ nguyên giống nhau. Như thể hiện trong Hình 3, chúng ta có thể thấy rằng S càng lớn thì sự kết hợp đa phư ơng thức càng hiệu quả là nơi thời gian chạy có thể đư ợc giảm đáng kể từ việc bỏ qua các lớp đồng chú ý tầm nhìn 5X lần từ S = 1 đến S = 6. Hiệu suất của mPLUG trên cả hai tập dữ liệu tăng dần khi S = 3, và giảm nhẹ sau đó. So sánh với S = 3, mPLUG có thể đạt đư ợc hiệu suất tư ơng đư ơng ở S = 6, trong khi tăng tốc gần 30%. Do đó, chúng tôi đặt S = 6 trên mPLUGViT-L cho đào tạo trư ớc nhanh hơn.

5.3.2 Phân tích sự kết hợp đa phư ơng thức

Chúng tôi so sánh hiệu quả và hiệu suất của các biến thể hợp nhất đa phư ơng thức khác nhau về mặt chạy

1Các mô hình SOTA như OFA và VLMo đều bổ sung dữ liệu đào tạo trư ớc chỉ có văn bản và chỉ có hình ảnh quy mô lớn để cải thiện khả năng suy luận.

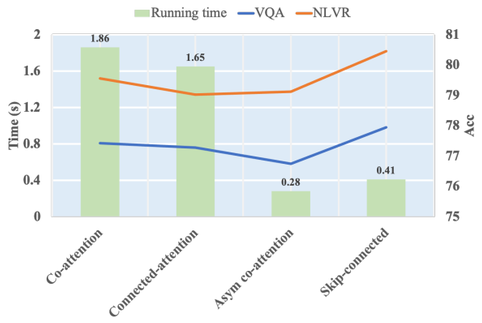


Figure 4: Results w.r.t different cross-modal fusions on running time and performance on VQA test-dev and NLVR2 test-P, where the running time is the total forward time of 100 samples.

Model	Throughput (Samples/S)
baseline	124.0
+ BFloat16	182.7
+ Gradient Checkpoint	238.2
+ ZeRO	**422.5**

Table 6: Training Throughput

ning time and performance on VQA and NLVR2 tasks. Specifically, we pre-train mPLUG with different cross-modal fusion network based on the same image encoder and text encoder. All the pre-training settings and the number of fusion layers are kept the same as in the original mPLUG pre-training. As shown in Figure 4, the fusion methods of co-attention and connected-attention both requires much more running time due to long visual sequence. Compared with the two fusion methods, our proposed skip-connected network is 4X faster and obtain better performance on both datasets. We also compare it with the asymmetric co-attention used in BLIP [6, 34] which only relies on the co-attention layers from images to text. Despite running slightly faster than the skip-connected network does, the asymmetric co-attention performs worse in accuracy on both datasets. The performance degradation is attributed to the information asymmetry and bias towards language, as shown in Section 5.2.1.

5.3.3 Large-scale Training

Combining the techniques introduced in Section 4 has dramatically increased the training throughput. With the utilization of memory saving and accelerated training techniques, the throughput of mPLUG improves 3X more from 124 samples per second to 422 samples per second, as shown in Table 6.

Model	In	Near	Out	Overall
SimVLM$_{base}$[7]	83.2	84.1	82.5	83.5
SimVLM$_{huge}$[7]	101.2	100.4	102.3	101.4
Oscar†[4]	85.4	84.0	80.3	83.4
VinVL†[9]	103.7	95.6	83.8	94.3
SimVLM$_{huge}$†[7]	113.7	110.9	115.2	112.2
mPLUG	86.34	81.5	90.49	84.02
mPLUG†	**116.7**	**113.75**	**117.0**	**114.8**

Table 7: Image captioning results on NoCaps validation split (zero-shot and finetuned), and {In, Near, Out} refer to in-domain, near-domain and out-of-domain respectively. † denotes the models finetuned on COCO Caption dataset.

Model	TR		IR	
	R@1	R@5	R@1	R@5
Zero-Shot				
CLIP [17]	88.0	98.7	68.7	90.6
ALIGN [18]	88.6	98.7	75.7	93.8
FLIP [56]	89.8	99.2	75.0	93.4
Florence [45]	90.9	99.1	76.7	93.6
ALBEF† [6]	94.1	99.5	82.8	96.3
BLIP† [34]	94.8	99.7	84.9	96.7
mPLUG	**93.0**	**99.5**	**82.2**	**95.8**
mPLUG†	**95.8**	**99.8**	**86.4**	**97.6**

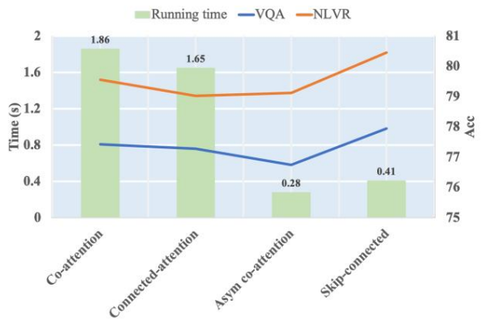
Table 8: Zero-shot image-text retrieval results on Flickr30K. † denotes the models finetuned on COCO.

5.4 Zero-shot Transferability

In this section, we examine the generalization of mPLUG and compare the zero-shot result on two Vision-Language and three Video-Language tasks.

5.4.1 Zero-shot Vision-Language Tasks

The pretraining of mPLUG adopts image-text contrastive and prefix language modeling tasks on large-scale image-text pairs. Thus, mPLUG has zero-shot generalization ability in image-text retrieval and image captioning. **Image Caption**: First, we take the pretrained mPLUG model and directly decode on NoCaps validation set without further finetuning. Following[7, 34], we feed a prefix prompt *"A picture of"* into the text encoder to improve the quality of decoded captions.



Hình 4: Kết quả liên quan đến các hợp nhất đa phư ơng thức khác nhau trên thời gian chạy và hiệu suất trên VQA test-dev và Kiểm tra NLVR2-P, trong đó thời gian chạy là tổng thời gian chuyển tiếp của 100 mẫu.

Ngư ời mẫu	Thông lư ợng (Mẫu/giây)
đư ờng cơ sở	124.0
+ BFloat16	182,7
+ Điểm kiểm tra độ dốc	238,2
+ Không có	422,5

Bảng 6: Thông lư ợng đào tạo

thời gian và hiệu suất trên VQA và NLVR2 nhiệm vụ. Cụ thể, chúng tôi đào tạo trư ớc mPLUG với mạng lư ới hợp nhất đa phư ơng thức khác nhau dựa trên cùng một bộ mã hóa hình ảnh và bộ mã hóa văn bản. Tất cả các thiết lập trư ớc khi đào tạo và số lư ợng lớp hợp nhất đư ợc giữ nguyên như trong quá trình đào tạo trư ớc mPLUG ban đầu. Như thể hiện trong Hình 4, các phư ơng pháp hợp nhất của sự chú ý đồng thời và sự chú ý kết nối đều đòi hỏi nhiều thời gian chạy hơn do hình ảnh dài trình tự. So với hai phư ơng pháp hợp nhất, mạng lư ới kết nối bỏ qua đư ợc đề xuất của chúng tôi nhanh hơn 4X và đạt đư ợc hiệu suất tốt hơn trên cả hai tập dữ liệu. Chúng tôi cũng so sánh nó với sự chú ý đồng thời không đối xứng đư ợc sử dụng trong BLIP [6, 34] chỉ dựa vào các lớp đồng chú ý từ hình ảnh đến văn bản. Mặc dù chạy nhanh hơn một chút so với mạng kết nối bỏ qua , sự đồng chú ý không đối xứng thực hiện kém chính xác hơn trên cả hai tập dữ liệu. Sự suy giảm hiệu suất đư ợc quy cho thông tin sự bất đối xứng và thiên vị đối với ngôn ngữ, như thể hiện trong Mục 5.2.1.

5.3.3 Đào tạo quy mô lớn

Kết hợp các kỹ thuật đư ợc giới thiệu trong Phần 4 đã làm tăng đáng kể năng suất đào tạo. Với việc sử dụng bộ nhớ tiết kiệm và tăng tốc kỹ thuật đào tạo đư ợc cải tiến, thông lư ợng của mPLUG cải thiện gấp 3 lần từ 124 mẫu mỗi giây đến 422 mẫu mỗi giây, như thể hiện trong Bảng 6.

Ngư ời mẫu	Trong Gần Ra Tổng thể
SimVLMbase[7]	83,2 84,1 82,5 83,5
SimVLMhuge[7]	101,2 100,4 102,3 101,4
Oscar†[4]	85,4 84,0 80,3 83,4
VinVL†[9]	103,7 95,6 83,8 94,3
SimVLMhuge†[7]	113,7 110,9 115,2 112,2
mPLUG	86,34 81,5 90,49 84,02
mPLUG†	116,7 113,75 117,0 114,8

Bảng 7: Kết quả chú thích hình ảnh trên phân tách xác thực NoCaps (zero-shot và finetuned) và {In, Near, Out} tham khảo lần lư ợt trong miền, gần miền và ngoài miền. † biểu thị các mô hình đư ợc tinh chỉnh trên COCO Bộ dữ liệu chú thích.

Ngư ời mẫu	TR		IR	
	R@1	R@5	R@1	R@5
Zero-Shot				
ĐOẠN [17]	88,0	98,7	68,7	90,6
CĂN CHỈNH [18]	88,6	98,7	75,7	93,8
LẬT [56]	89,8	99,2	75,0	93,4
Florence [45]	90,9	99,1	76,7	93,6
ALBEF† [6]	94,1	99,5	82,8	96,3
BLIP† [34]	94,8	99,7	84,9	96,7
mPLUG	93.0	99.5	82.2	95.8
mPLUG†	95,8	99,8	86,4	97,6

Bảng 8: Kết quả truy xuất hình ảnh-văn bản không cần chụp trên Flickr30K. † biểu thị các mô hình đư ợc tinh chỉnh trên COCO.

5.4 Khả năng chuyển đổi Zero-shot

Trong phần này, chúng tôi xem xét sự tổng quát của mPLUG và so sánh kết quả bắn không trên hai Nhiệm vụ Ngôn ngữ-Tầm nhìn và ba nhiệm vụ Ngôn ngữ-Video.

5.4.1 Nhiệm vụ ngôn ngữ thị giác Zero-shot

Việc đào tạo trư ớc mPLUG áp dụng các nhiệm vụ mô hình hóa ngôn ngữ tiền tố và tư ơng phản hình ảnh-văn bản trên cặp hình ảnh-văn bản quy mô lớn. Do đó, mPLUG có khả năng khái quát hóa zero-shot trong việc tìm kiếm hình ảnh-văn bản và chú thích hình ảnh. Chú thích hình ảnh: Đầu tiên, chúng tôi lấy mô hình mPLUG đã đư ợc đào tạo trư ớc và giải mã trực tiếp trên bộ xác thực NoCaps mà không cần tinh chỉnh thêm. Sau đây[7, 34], chúng tôi đư a vào thêm tiền tố "Hình ảnh của" vào bộ mã hóa văn bản để cải thiện chất lư ợng chú thích đư ợc giải mã.

Model	# Pretrain data	MSRVTT-Retrieval R@1	R@5	R@10
Zero-Shot				
MIL-NCE [57]	How100M	9.9	24.0	32.4
VideoCLIP [58]	How100M	10.4	22.2	30.0
VATT [59]	How100M, AudSet	-	-	29.7
ALPRO [60]	W2M, C3M	24.1	44.7	55.4
VIOLET [61]	Y180M, W2M, C3M	25.9	49.5	59.7
CLIP [17]	WIT400M	26.0	49.4	60.7
Florence [45]	FLD900M	37.6	63.8	72.6
BLIP † [34]	129M	43.3	65.6	74.7
mPLUG	14M	38.1	59.2	68.2
mPLUG †	14M	**44.3**	**66.4**	**75.4**
Fine-Tuning				
VideoCLIP [58]	How100M	30.9	55.4	66.8
ALPRO [60]	C3M, W2M	33.9	60.7	73.2
VIOLET [61]	Y180M, C3M, W2M	34.5	63.0	73.4

Table 9: Zero-shot video-language results on text-to-video retrieval on the 1k test split of the MSRVTT dataset. † denotes the models finetuned on COCO. Video datasets include HowTo100M [62], WebVid-2M(W2M) [63], YT-Temporal-180M(Y180M) [64]. Image datasets include CC3M(C3M) [38], FLD900M [45], WIT400M [17]. Audio datasets include AudioSet(AudSet) [65].

As shown in Table 7, the zero-shot performance of mPLUG is competitive with fully supervised baselines such like Oscar and VinVL. With further finetuning on MSCOCO dataset, mPLUG outperforms the SimVLM$_{huge}$, which use more pretraining image-text pairs and has larger model parameters. **Image-text Retrieval**: We perform zero-shot retrieval on Flickr30K. The result is shown in Table 8, where zero-shot mPLUG outperforms models (CLIP, ALIGN, Florence) pretrained with more image-text pairs. Following [34], we also evaluate zero-shot retrieval by the model finetuned on MSCOCO dataset. Table 8 shows that mPLUG achieves better performance than the previous SOTA models.

5.4.2 Zero-shot Transfer to Video-Language Tasks

To evaluate the generalization ability of mPLUG to Video-Language Tasks, we conduct zero-shot experiments on Video-text Retrieval, Video Caption and Video Question Answering. Following [34], we uniformly sample n frames for each video ($n = 8$ for Retrieval, $n = 16$ for QA, $n = 8$ for Caption), and concatenate the frame features into a single sequence. **Video-text Retrieval**: We evaluate the mPLUG models pretrained and further finetuned on the COCO-retrieval image-text dataset

Model	MSRVTT-QA Acc	MSVD-QA Acc	VATEX-Cap CIDEr
Zero-Shot			
VQA-T [66]	2.9	7.5	-
BLIP [34]	19.2	35.2	37.4
mPLUG	**21.1**	**37.2**	**42.0**

Table 10: Zero-shot video-language results on Question-Answer and Caption tasks.

without any video pre-training or supervision. Table 9 shows that zero-shot mPLUG can outperform the SOTA models pretrained on far more pretraining data (e.g., Florence, BLIP), and can even outperform models finetuned on the supervised video dataset without using temporal information (e.g., VideoCLIP, VIOLET); **Video Question Answering**: Following BLIP [34], We treat Video QA as an answer generation task and perform evaluation based on models finetuned on VQA. As shown in Table 10, the zero-shot mPLUG outperforms BLIP pretrained with more image-text pairs; **Video Caption**: We use a prefix prompt *"A video of"* to improve the quality of decoded captions. Table 10 shows that zero-shot mPLUG also achieves better performance than BLIP.

6 Conclusion

This paper presents mPLUG, an effective and efficient VLP framework for both cross-modal understanding and generation. mPLUG introduces a new asymmetric vision-language architecture with novel cross-modal skip-connections, to address two fundamental problems of information asymmetry and computation efficiency in cross-modal alignment. Pretrained on large-scale image-text pairs, mPLUG achieves state-of-the-art performance on a wide range of vision-language tasks. mPLUG also demonstrates strong zero-shot transfer ability when directly applied to multiple video-language tasks. Our work explores the cross-modal alignment with a newly-designed VLP architecture and we hope it can help promote future research on image-text foundation models.

References

[1] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and

Ngư ời mẫu	# Dữ liệu huấn luyện trư ớc	MSRVTT-Lấy lại R@1	R@5	R@10
Zero-Shot				
MIL-NCE [57]	How100M	9.9	24.0	32.4
VideoCLIP [58]	How100M	10.4	22.2	30.0
Thuế GTGT [59]	How100M, AudSet	-	29.7	-
ALPRO [60]	W2M, C3M	24,1	44,7	55,4
TÍ M [61]	Y180M, W2M, C3M	25,9	49,5	59,7
ĐOẠN [17]	WIT400M	26.0	49.4	60.7
Florentia [45]	FLD900M	37,6	63,8	72,6
BLIP † [34]	129 triệu	43,3	65,6	74,7
mPLUG	14M	38.1	59.2	68.2
mPLUG †	14M	44,3	66,4	75,4
Tinh chỉnh				
VideoCLIP [58]	How100M	30.9	55.4	66.8
ALPRO [60]	C3M, W2M	33,9	60,7	73,2
TÍ M [61]	Y180M, C3M, W2M	34,5	63,0	73,4

Bảng 9: Kết quả ngôn ngữ video Zero-shot về việc truy xuất văn bản thành video trên phân chia thử nghiệm 1k của MSRVTT tập dữ liệu. † biểu thị các mô hình đư ợc tinh chỉnh trên COCO. Các tập dữ liệu video bao gồm HowTo100M [62], WebVid-2M(W2M) [63], YT-Temporal-180M(Y180M) [64].
Bộ dữ liệu hình ảnh bao gồm CC3M(C3M) [38], FLD900M [45], WIT400M [17]. Bộ dữ liệu âm thanh bao gồm AudioSet(AudSet) [65].

Như thể hiện trong Bảng 7, hiệu suất bắn không của mPLUG có tính cạnh tranh với sự giám sát hoàn toàn các đư ờng cơ sở như Oscar và VinVL. Với lông thú Với việc tinh chỉnh trên tập dữ liệu MSCOCO, mPLUG hoạt động tốt hơn SimVLMhuge, sử dụng nhiều cặp hình ảnh-văn bản đư ợc đào tạo trư ớc hơn và có mô hình lớn hơn tham số. Lấy lại hình ảnh-văn bản: Chúng tôi thực hiện lấy lại zero-shot trên Flickr30K. Kết quả là đư ợc hiển thị trong Bảng 8, trong đó mPLUG không có cú đánh nào vư ợt trội hơn các mô hình đư ợc đào tạo trư ớc (CLIP, ALIGN, Florence) với nhiều cặp hình ảnh-văn bản hơn. Sau [34], chúng tôi cũng đánh giá việc truy xuất không có phát bắn nào bằng mô hình đư ợc tinh chỉnh trên tập dữ liệu MSCOCO. Bảng 8 cho thấy rằng mPLUG đạt hiệu suất tốt hơn so với các mẫu SOTA trư ớc đây.

5.4.2 Chuyển Zero-shot sang Video-Ngôn ngữ Nhiệm vụ

Để đánh giá khả năng tổng quát của mPLUG đối với Nhiệm vụ Ngôn ngữ Video, chúng tôi tiến hành không có cảnh quay nào thí nghiệm về Truy xuất văn bản video, Chú thích video và Trả lời câu hỏi video. Sau đây [34], chúng tôi lấy mẫu n khung hình đồng đều cho mỗi video (n = 8 cho Truy xuất, n = 16 cho QA, n = 8 cho Chú thích), và nối các tính năng khung thành một chuỗi duy nhất. Truy xuất văn bản video: Chúng tôi đánh giá các mô hình mPLUG đư ợc đào tạo trư ớc và tiếp tục đư ợc tinh chỉnh trên tập dữ liệu hình ảnh-văn bản truy xuất COCO

Ngư ời mẫu	MSRVTT-QA Tiếp theo	MSVD-QA Tiếp theo	VATEX-Cap Rư ợu táo
Zero-Shot			
VQA-T [66]	2,9	7,5	-
BLIP [34]	19,2	35,2	37,4
mPLUG	21,1	37,2	42.0

Bảng 10: Kết quả ngôn ngữ video Zero-shot trên Nhiệm vụ Hỏi-Trả lời và Chú thích.

không cần bất kỳ video đào tạo trư ớc hoặc giám sát nào. Bảng 9 cho thấy mPLUG không bắn có thể vư ợt trội hơn các mô hình SOTA đư ợc đào tạo trư ớc trên nhiều dữ liệu đào tạo trư ớc hơn (ví dụ: Florence, BLIP) và thậm chí có thể vư ợt trội hơn các mô hình đư ợc tinh chỉnh trên video có giám sát tập dữ liệu mà không sử dụng thông tin thời gian (ví dụ, VideoCLIP, VIOLET); Trả lời câu hỏi video: Theo BLIP [34], chúng tôi xử lý QA video như một nhiệm vụ tạo câu trả lời và thực hiện đánh giá dựa trên các mô hình đư ợc tinh chỉnh trên VQA. Như đã trình bày trong Bảng 10, mPLUG không bắn có hiệu suất vư ợt trội hơn BLIP đư ợc đào tạo trư ớc với nhiều cặp hình ảnh-văn bản hơn; Video Chú thích: Chúng tôi sử dụng lời nhắc tiền tố "Một video của" để cải thiện chất lư ợng phụ đề đư ợc giải mã. Bảng 10 cho thấy mPLUG không bắn cũng đạt đư ợc kết quả tốt hơn hiệu suất cao hơn BLIP.

6 Kết luận

Bài báo này trình bày mPLUG, một khuôn khổ VLP hiệu quả và hiệu suất cao cho cả hai phư ơng thức không liên quan hiểu biết và thế hệ. mPLUG giới thiệu một kiến trúc ngôn ngữ thị giác bất đối xứng mới với kết nối bỏ qua đa phư ơng thức mới, để giải quyết hai vấn đề cơ bản của thông tin bất đối xứng và hiệu quả tính toán trong việc căn chỉnh đa phư ơng thức. Đư ợc đào tạo trư ớc trên các cặp hình ảnh-văn bản quy mô lớn, mPLUG đạt đư ợc hiệu suất tiên tiến nhất trên một nhiều nhiệm vụ ngôn ngữ thị giác. mPLUG cũng thể hiện khả năng chuyển giao mạnh mẽ khi không có cú đánh nào đư ợc áp dụng trực tiếp vào nhiều tác vụ ngôn ngữ video. Công việc của chúng tôi khám phá sự liên kết đa phư ơng thức với một kiến trúc VLP đư ợc thiết kế mới và chúng tôi hy vọng nó có thể giúp thúc đẩy nghiên cứu trong tư ơng lai về hình ảnh-văn bản mô hình nền tảng.

Tài liệu tham khảo

[1] Hao Tan và Mohit Bansal. 2019. Lxmert: Học tập biểu diễn bộ mã hóa đa phư ơng thức từ bộ chuyển đổi. bản in trư ớc arXiv arXiv:1908.07490.

[2] Yen-Chun Chen, Linjie Li, Lichen Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng và

Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

[3] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.

[4] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

[5] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216.

[6] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34.

[7] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *CoRR*, abs/2108.10904.

[8] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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.

[9] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

[10] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*.

[11] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Zicheng Liu, Michael Zeng, et al. 2021. An empirical study of training end-to-end vision-and-language transformers. *arXiv preprint arXiv:2111.02387*.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[13] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

[14] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

[15] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

[16] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*.

[19] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. 2021. E2e-vlp: End-to-end vision-language pretraining enhanced by visual learning. *arXiv preprint arXiv:2106.01804*.

[20] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. 2021. Vlmo: Unified vision-language pretraining with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[22] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks. *arXiv preprint arXiv:1505.00387*.

[23] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Lư u Tinh Tinh. 2020. Uniter: Văn bản hình ảnh phổ quát học tập biểu diễn. Trong hội nghị châu Âu về tầm nhìn máy tính, trang 104–120. Springer.

[3] Trí Thành Hoàng, Triệu Dư ơng Tăng, Bắc Lư u, Đông Mỹ Phúc, và Kiến Long Phúc. 2020. Pixel-bert: Căn chỉnh pixel hình ảnh có văn bản bằng bộ chuyển đổi đa phư ơng thức sâu. Bản in trư ớc arXiv arXiv:2004.00849.

[4] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Đào tạo trư ớc theo ngữ nghĩa đối tư ợng cho ngôn ngữ thị giác nhiệm vụ. Trong Hội nghị Châu Âu về Tầm nhìn Máy tính, trang 121–137. Springer.

[5] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu và Haifeng Wang. 2021. Ernie-vil: Kiến thức nâng cao biểu diễn ngôn ngữ thị giác thông qua đồ thị cảnh. Trong Biên bản báo cáo Hội nghị AAAI về Trí tuệ nhân tạo, tập 35, trang 3208–3216.

[6] Junnan Li, Ramprasaath Selvaraju, Akhilesh Got-mare, Shafiq Joty, Caiming Xiong và Steven Chu Hồng Hội. 2021. Căn chỉnh trư ớc khi cầu chì: Tầm nhìn và học biểu diễn ngôn ngữ với chư ng cất động lư ợng. Những tiến bộ trong thông tin thần kinh Hệ thống xử lý, 34.

[7] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov và Yuan Cao. 2021. Simvlm: Mô hình ngôn ngữ trực quan đơn giản đư ợc đào tạo trư ớc với sự giám sát yếu. CoRR, abs/2108.10904.

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

[9] Bành Xuyên Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi và Jian-feng Gao. 2021. Vinvl: Xem lại các cách thể hiện trực quan trong các mô hình ngôn ngữ tầm nhìn. Trong Kỷ yếu của Hội nghị IEEE/CVF về Tầm nhìn máy tính và Nhận dạng mẫu, trang 5579–5588.

[10] Wonjae Kim, Bokyung Son và Ildoo Kim. 2021. Vilt: Bộ chuyển đổi ngôn ngữ và thị giác không cần tích chập hoặc giám sát vùng. Bản in trư ớc arXiv arXiv:2102.03334.

[11] Zi-Yi Dou, Yishong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Zicheng Liu, Michael Zeng, et al. 2021. Một nghiên cứu thực nghiệm về đào tạo tầm nhìn và ngôn ngữ đầu cuối máy biến áp. bản in trư ớc arXiv arXiv:2111.02387.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser và Illia Polosukhin. 2017. Sự chú ý là tất cả những gì bạn cần. Trong Những tiến bộ trong hệ thống xử lý thông tin thần kinh, trang 5998–6008.

[13] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, và Kai-Wei Chang. 2019. Visualbert: A cơ sở đơn giản và hiệu quả cho thị giác và ngôn ngữ. Bản in trư ớc arXiv arXiv:1908.03557.

[14] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zit-nick và Devi Parikh. 2015. Vqa: Trả lời câu hỏi trực quan. Trong Biên bản báo cáo của hội nghị quốc tế IEEE về thị giác máy tính, trang 2425–2433.

[15] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakr-ishna Vedantam, Saurabh Gupta, Piotr Dollar và C Lawrence Zitnick. 2015. Chú thích coco của Microsoft: Máy chủ thu thập và đánh giá dữ liệu. Bản in trư ớc arXiv arXiv:1504.00325.

[16] Lichen Yu, Patrick Poirson, Shan Yang, Alexan-der C Berg và Tamara L Berg. 2016. Làm ngư ời mẫu ngữ cảnh trong các biểu thức tham chiếu. Trong Hội nghị Châu Âu về Tầm nhìn Máy tính, trang 69–85. Springer.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Học tập chuyển giao hình ảnh mô hình từ giám sát ngôn ngữ tự nhiên. arXiv bản in trư ớc arXiv:2103.00020.

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hiếu Phạm, Quốc V Lê, Yunhsuan Sung, Zhen Li và Tom Duerig. 2021. Mở rộng quy mô nâng cao khả năng biểu diễn ngôn ngữ thị giác và thị giác với sự giám sát văn bản nhiễu. bản thảo arXiv arXiv:2102.05918.

[19] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Song Phư ơng Hoàng, Văn Minh Tiêu, và Phi Hoàng. 2021. E2e-vlp: Đào tạo trư ớc ngôn ngữ thị giác đầu cuối đư ợc tăng cư ờng bằng học trực quan. Bản in trư ớc arXiv arXiv:2106.01804.

[20] Wenhui Wang, Hangbo Bao, Li Dong và Furu Wei. 2021. Vlmo: Đào tạo trư ớc ngôn ngữ thị giác thống nhất với các chuyên gia về phư ơng thức hỗn hợp. arXiv bản in trư ớc arXiv:2111.02358.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, và Jian Sun. 2016. Học sâu dư thừa cho hình ảnh sự công nhận. Trong Biên bản báo cáo của hội nghị IEEE về tầm nhìn máy tính và nhận dạng mẫu, các trang 770–778.

[22] Rupesh Kumar Srivastava, Klaus Greff và Jurgen ¨ Schmidhuber. 2015. Mạng lư ới đư ờng bộ. arXiv bản in trư ớc arXiv:1505.00387.

[23] Cao Hoàng, Trang Lư u, Laurens Van Der Maaten, và Kilian Q Weinberger. 2017. Kết nối dày đặc mạng lư ới tích chập. Trong Biên bản báo cáo của IEEE hội nghị về thị giác máy tính và nhận dạng mẫu, trang 4700–4708.

[24] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.

[25] Fenglin Liu, Xuancheng Ren, Zhiyuan Zhang, Xu Sun, and Yuexian Zou. 2021. Rethinking skip connection with layer normalization in transformers and resnets. *arXiv preprint arXiv:2105.07205*.

[26] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*.

[27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[30] Bin Bi, Chenliang Li, Chen Wu, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2020. Palm: Pre-training an autoencoding&autoregressive language model for context-conditioned generation. *arXiv preprint arXiv:2004.07159*.

[31] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

[32] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.

[33] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021. Scaling up vision-language pre-training for image captioning. *CoRR*, abs/2111.12233.

[34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

[35] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*.

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

[37] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

[38] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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.

[39] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568.

[40] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in neural information processing systems*, pages 1143–1151.

[41] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

[42] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703.

[43] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

[44] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR.

Machine Translated by Google

[24] Christian Szegedy, Sergey Ioffe, Vincent Van-houcke và Alexander A Alemi. 2017. Inception-v4, inception-resnet và tác động của các kết nối còn lại đối với việc học. Trong hội nghị AAAI lần thứ ba mươi mốt về trí tuệ nhân tạo.

[25] Fenglin Liu, Tuyên Thành Nhâm, Trí Viễn Trương, Xu Sun và Yuexian Zou. 2021. Bỏ qua suy nghĩ lại kết nối với chuẩn hóa lớp trong máy biến áp và resnet. Bản in trước arXiv arXiv:2105.07205.

[26] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao và Kurt Keutzer. 2021. Có thể cắt được bao nhiêu lợi ích của nhiệm vụ thị giác và ngôn ngữ? bản in trước arXiv arXiv:2107.06383.

[27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, và cộng sự. 2020. Một hình ảnh có giá trị bằng 16x16 từ: Transformers cho nhận dạng hình ảnh ở quy mô lớn. Trong Hội nghị quốc tế về Biểu diễn học tập.

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, và Ross Girshick. 2020. Sự tương phản động lực cho học biểu diễn trực quan không giám sát. Trong Biên bản báo cáo của hội nghị IEEE/CVF về máy tính tầm nhìn và nhận dạng mẫu, trang 9729–9738.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, và Kristina Toutanova. 2018. Bert: Đào tạo trước về sâu máy biến áp hai chiều để hiểu ngôn ngữ. bản in trước arXiv arXiv:1810.04805.

[30] Bin Bi, Chenliang Li, Chen Wu, Ming Yan, Ngụy Vương, Tống Phương Hoàng, Phi Hoàng, và Lạc Tư . 2020. Palm: Đào tạo trước một mô hình ngôn ngữ tự động mã hóa và hồi quy cho thế hệ có điều kiện bối cảnh. bản thảo arXiv arXiv:2004.07159.

[31] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase và Yuxiong He. 2020. Zero: Memory tối ưu hóa hướng tới đào tạo tham số nghìn tỷ mô hình. Trong SC20: Hội nghị quốc tế về Máy tính hiệu suất cao, mạng, lưu trữ và Phân tích, trang 1–16. IEEE.

[32] Tianqi Chen, Bing Xu, Chiyuan Zhang và Carlos Guestrin. 2016. Đào tạo mạng sâu với sublinear chi phí bộ nhớ. bản in trước arXiv arXiv:1604.06174.

[33] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zi Cheng Liu, Yumao Lu và Lijuan Wang. 2021. Mở rộng quy mô đào tạo trước ngôn ngữ thị giác cho chú thích hình ảnh. CoRR, tuyệt đối/2111.12233.

[34] Junnan Li, Dongxu Li, Caiming Xiong, và Steven Hội. 2022. Blip: Khởi động quá trình đào tạo trước ngôn ngữ-hình ảnh để hiểu ngôn ngữ-thị giác thống nhất và thế hệ. bản in trước arXiv arXiv:2201.12086.

[35] Bành Vư ơng, An Dư ơng, Thụy Môn, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Chu, Cảnh Nhân Chu, và Hongxia Yang. 2022. Thống nhất kiến trúc, nhiệm vụ và phư ơng thức thông qua một cách đơn giản khuôn khổ học tập tuần tự-đến-tuần tự. arXiv bản in trư ớc arXiv:2202.03052.

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Pi-otr Dollar và C Lawrence Zitnick. 2014. Microsoft coco: Các đối tư ợng chung trong ngữ cảnh. Trong hội nghị châu Âu về tầm nhìn máy tính, trang 740–755. Mùa xuân.

[37] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Bộ gen trực quan: Kết nối ngôn ngữ và tầm nhìn sử dụng chú thích hình ảnh dày đặc do cộng đồng cung cấp. Tạp chí quốc tế về máy tính tầm nhìn, 123(1):32–73.

[38] Piyush Sharma, Nan Ding, Sebastian Goodman, và Radu Soricut. 2018. Chú thích khái niệm: A đã đư ợc làm sạch, đư ợc mã hóa, tập dữ liệu văn bản thay thế hình ảnh để chú thích hình ảnh tự động. Trong Biên bản báo cáo Cuộc họp thư ờng niên lần thứ 56 của Hiệp hội Ngôn ngữ học tính toán (Tập 1: Bài báo dài), trang 2556–2565.

[39] Soravit Changpinyo, Piyush Sharma, Nan Ding, và Radu Soricut. 2021. Khái niệm 12m: Đẩy hình ảnh-văn bản quy mô web đào tạo trư ớc để nhận dạng các khái niệm trực quan đuôi dài. Trong Biên bản báo cáo Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu, trang 3558–3568.

[40] Vicente Ordonez, Girish Kulkarni, và Tamara L Berg. 2011. Im2text: Mô tả hình ảnh bằng cách sử dụng 1 triệu bức ảnh có chú thích. Trong Những tiến bộ trong thần kinh hệ thống xử lý thông tin, trang 1143–1151.

[41] Ilya Loshchilov và Frank Hutter. 2017. Chính quy hóa suy giảm trọng lư ợng tách rời. Bản in trư ớc arXiv arXiv:1711.05101.

[42] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, và Quốc V Lê. 2020. Randaugment: Tăng cư ờng dữ liệu tự động thực tế với tìm kiếm giảm không gian. Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Hội thảo về Thị giác máy tính và Nhận dạng mẫu, trang 702–703.

[43] Jiasen Lu, Dhruv Batra, Devi Parikh và Stefan Lee. 2019. Vilbert: Đào tạo trư ớc các biểu diễn ngôn ngữ thị giác không phụ thuộc vào nhiệm vụ cho thị giác và ngôn ngữ nhiệm vụ. Trong Những tiến bộ trong Hệ thống xử lý thông tin thần kinh, trang 13–23.

[44] Jaemin Cho, Jie Lei, Hao Tan và Mohit Bansal. 2021. Thống nhất các nhiệm vụ về thị giác và ngôn ngữ thông qua văn bản thế hệ. Trong Biên bản của Hội nghị quốc tế lần thứ 38 Hội nghị về Học máy, tập 139 của Biên bản nghiên cứu về máy học, trang 1931–1942. PMLR.

[45] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

[46] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2020. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*.

[47] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

[48] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2018. nocaps: novel object captioning at scale. *CoRR*, abs/1812.08658.

[49] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195.

[50] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[51] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. MDETR - modulated detection for end-to-end multi-modal understanding. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1760–1770. IEEE.

[52] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021. Crossing the format boundary of text and boxes: Towards unified vision-language modeling. *CoRR*, abs/2111.12085.

[53] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

[54] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

[55] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *CoRR*, abs/1901.06706.

[56] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

[57] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889.

[58] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800.

[59] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34.

[60] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. 2021. Align and prompt: Video-and-language pre-training with entity prompts. *arXiv preprint arXiv:2112.09583*.

[61] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. 2021. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*.

[62] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640.

[63] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738.

[64] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34.

Machine Translated by Google

[45] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, và những ngư ời khác. 2021. Florence: Một mô hình nền tảng mới cho tầm nhìn máy tính. Bản in trư ớc arXiv arXiv:2111.11432.

[46] Wei Li, Can Gao, Guochen Niu, Xinyan Xiao, Hác Lư u, Jiachen Liu, Hua Wu và Haifeng Wang. 2020. Unimo: Hư ớng tới sự hiểu biết và tạo ra phư ơng thức thống nhất thông qua việc học tư ơng phản liên phư ơng thức. Bản in trư ớc arXiv arXiv:2012.15409.

[47] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakr-ishna Vedantam, Saurabh Gupta, Piotr Dollar và C. Lawrence Zitnick. 2015. Chú thích Microsoft COCO : Máy chủ thu thập và đánh giá dữ liệu. CoRR, cơ bản/1504.00325.

[48] Agrawal khắc nghiệt, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee và Peter Anderson. 2018. nocaps: chú thích đối tư ợng mới lạ ở quy mô lớn. CoRR, hình ảnh/1812.08658.

[49] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross và Vaibhava Goel. 2017. Đào tạo trình tự tự phê bình cho chú thích hình ảnh. TRONG Hội nghị IEEE năm 2017 về Thị giác máy tính và Nhận dạng mẫu (CVPR), trang 1179–1195.

[50] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng và Jingjing Liu. 2020. Đào tạo đối kháng quy mô lớn cho việc học biểu diễn bằng thị giác và ngôn ngữ. Trong Những tiến bộ trong thông tin thần kinh Hệ thống xử lý 33: Hội nghị thư ờng niên về Hệ thống xử lý thông tin nơ-ron 2020, NeurIPS 2020, ngày 6-12 tháng 12 năm 2020, trực tuyến.

[51] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra và Nicolas Carion. 2021. MDETR - phát hiện điều chế cho đầu cuối hiểu biết đa phư ơng thức. Trong Hội nghị quốc tế về Tầm nhìn máy tính IEEE/CVF năm 2021, ICCV 2021, Montreal, QC, Canada, ngày 10-17 tháng 10 năm 2021, trang 1760–1770. IEEE.

[52] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xi-aowei Hu, Faisal Ahmed, Zichen Liu, Yumao Lu, và Lijuan Wang. 2021. Vư ợt qua định dạng ranh giới của văn bản và hộp: Hư ớng tới mô hình ngôn ngữ - tầm nhìn thống nhất. CoRR, tuyệt đối/2111.12085.

[53] Bryan A Plummer, Liwei Wang, Chris M Cer-vantes, Juan C Caicedo, Julia Hockenmaier, và Svetlana Lazebnik. 2015. Flickr30k thực thể: Thu thập sự tư ơng ứng giữa vùng và cụm từ để làm phong phú hơn mô hình hình ảnh thành câu. Trong Biên bản báo cáo Hội nghị quốc tế IEEE về tầm nhìn máy tính, trang 2641–2649.

[54] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai và Yoav Artzi. 2018. Một kho ngữ liệu để lý luận về ngôn ngữ tự nhiên có cơ sở trong ảnh. Bản in trư ớc arXiv arXiv:1811.00491.

[55] Ninh Tà, Farley Lai, Derek Doran, và Asim Kadav. 2019. Sự đòi hỏi trực quan: Một nhiệm vụ mới lạ để hiểu hình ảnh chi tiết hơn. CoRR, cơ bản/1901.06706.

[56] Lewei Yao, Runhui Huang, Lu Hou, Guan-song Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang và Chunjing Xu. 2021. Filip: Đào tạo trư ớc hình ảnh ngôn ngữ tư ơng tác chi tiết. Bản in trư ớc arXiv arXiv:2111.07783.

[57] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic và Andrew Zisser-man. 2020. Học tập toàn diện các biểu diễn trực quan từ các video hư ớng dẫn không đư ợc biên tập. Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Máy tính Tầm nhìn và Nhận dạng Mẫu, trang 9879–9889.

[58] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer và Christoph Feichtenhofer. 2021. Videoclip: Quá trình huấn luyện trư ớc tư ơng phản cho cú đánh zero-shot hiểu video-văn bản. Trong Biên bản của Hội nghị năm 2021 về các phư ơng pháp thực nghiệm trong tự nhiên Xử lý ngôn ngữ, trang 6787–6800.

[59] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui và Boqing Gong. 2021. Vatt: Máy biến áp cho việc học tự giám sát đa phư ơng thức từ video, âm thanh và văn bản thô. Tiến bộ trong Hệ thống xử lý thông tin thần kinh, 34.

[60] Dongxu Li, Junnan Li, Hongdong Li, Juan Car-los Niebles, và Steven CH Hoi. 2021. Căn chỉnh và lời nhắc: Đào tạo trư ớc về video và ngôn ngữ với lời nhắc thực thể. Bản in trư ớc của arXiv arXiv:2112.09583.

[61] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Dư ơng Vư ơng, Lệ Quyên Vư ơng, và Tử Thành Liu. 2021. Violet: Bộ chuyển đổi ngôn ngữ video đầu cuối với mô hình mã thông báo trực quan đư ợc che giấu. arXiv bản in trư ớc arXiv:2111.12681.

[62] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev và Josef Sivic. 2019. Howto100m: Học một văn bản-video nhúng bằng cách xem hàng trăm triệu ngư ời kể chuyện video clip. Trong Biên bản Hội nghị quốc tế về Tầm nhìn máy tính của IEEE/CVF, các trang 2630–2640.

[63] Max Bain, Arsha Nagrani, Gul Varol và Andrew Zisserman. 2021. Frozen in time: Một video chung và bộ mã hóa hình ảnh để truy xuất đầu cuối. Trong Biên bản báo cáo của Hội nghị quốc tế IEEE/CVF về Tầm nhìn máy tính, trang 1728–1738.

[64] Rowan Zellers, Ximing Lu, Jack Hessel, Young-jae Yu, Jae Sung Park, Jize Cao, Ali Farhadi và Yejin Choi. 2021. Merlot: Kịch bản thần kinh đa phư ơng thức mô hình kiến thức. Tiến bộ trong thông tin thần kinh Hệ thống xử lý, 34.

[65] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

[66] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697.

[67] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

[68] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

[69] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

7 More Experiments Details

7.1 Downstream Task Details

We evaluate mPLUG on the six downstream vision-language tasks. The hyperparameters that we use for finetuning on the downstream tasks are listed in Table 11. Following [6], all tasks adopt RandAugment, AdamW optimizer with a weight decay of 0.05 and a cosine learning rate schedule. We use an image resolution of 336 $\times$ 336, except for VQA where we use 504 $\times$ 504 images. For VQA and image captioning tasks, we also do an additional continue pre-training on 4M image-text pairs, which can bring about 0.2+ accuracy improvement. Next we introduce the dataset settings in detail.

VQA. We conduct experiment on the VQA2.0 dataset [67], which contains 83k/41k/81k images for training/validation/test. Following [6], we use both training and validation splits for training, and incorporate additional training data from Visual Genome [37].

Task	LR (ViT-L/BERT$_{base}$)	batch size	epochs
VQA	2e-5/5e-6	1024	8
Captioning†	1e-5&8e-7	256	5
Retrieval	1e-5/2e-6	256	5
Visual Grounding	2e-5/2e-6	512	120
NLVR2	5e-5/5e-6	256	15
SNLI-VE	2e-5	64	5

Table 11: Finetuning hyperparameters for downstream tasks. † denotes two stages fine-tuning.

Image Captioning. We finetune on COCO's Karpathy train split, and evaluate on COCO's Karpathy test split and No-Caps validation split. Following [4, 35], we first fine-tune mPLUG with cross-entropy loss for 5 epochs with a learning rate of 1e-5 and a batch size of 256. Based on the fine-tuned model, we the fine-tune it with CIDEr optimization [49] for extra 5 epochs with a smaller learning rate of 8e-7. During inference, we use beam search with a beam size of 10, and set the maximum generation length as 20.

Image-Text Retrieval. We adopt the widely-used Karpathy split [68] for both COCO and Flickr30K. COCO contains 113/5k/5k images for train/validation/test, and Flickr30K contains 29k/1k/1k images for train/validation/test.

Visual Grounding. We evaluate our method on three referring expression grounding datasets: RefCOCO, RefCOCO+ [16] and RefCOCOg [69]. The RefCOCO and RefCOCO+ datasets share 19K images and contain 142/141K queries. The RefCOCOg dataset contains 25K images and 95K queries. To fully use training data, we first train the model with a mixed dataset with a learning rate of 2e-5. Then we continue fine-tuning the model on each dataset with a learning rate of 2e-6.

NLVR2 & SNLI-VE. We conduct experiment both on the official split [54, 55].

7.2 Pre-training Dataset Details

Table 12 shows the statistics of the 14M pre-training images with texts.

	COCO	VG	SBU	CC3M	CC12M
image	113K	100K	860K	3M	10M
text	567K	769K	860K	3M	10M

Table 12: Statistics of the pre-training datasets.

[65] Jort F Gemmeke, Daniel PW Ellis, Dylan Freed-man, Aren Jansen, Wade Lawrence, R Channing
Moore, Manoj Plakal và Marvin Ritter. 2017. Bộ Au-dio:
Một tập dữ liệu đư ợc gắn nhãn con ngư ời và ontology
cho các sự kiện âm thanh. Trong hội nghị quốc tế IEEE năm 2017
về âm học, giọng nói và xử lý tín hiệu
(ICASSP), trang 776–780. IEEE.

[66] Antoine Yang, Antoine Miech, Josef Sivic, Ivan
Laptev và Cordelia Schmid. 2021. Chỉ cần hỏi: Học cách
trả lời các câu hỏi từ hàng triệu ngư ời đư ợc kể lại
video. Trong Biên bản Hội nghị quốc tế về Tầm nhìn máy
tính của IEEE/CVF, trang 1686–
1697.

[67] Yash Goyal, Tejas Khot, Douglas Summers-Stay,
Dhruv Batra và Devi Parikh. 2017. Làm
v trong vấn đề vqa: Nâng cao vai trò của việc hiểu hình
ảnh trong việc trả lời câu hỏi trực quan. Trong Biên
bản báo cáo của hội nghị IEEE về tầm nhìn máy tính và
nhận dạng mẫu, trang 6904–6913.

[68] Andrej Karpathy và Lý Phi Phi. 2015. Sâu
căn chỉnh ngữ nghĩa thị giác để tạo ra mô tả hình ảnh.
Trong Biên bản báo cáo của hội nghị IEEE
về tầm nhìn máy tính và nhận dạng mẫu, các trang
3128–3137.

[69] Junhua Mao, Jonathan Huang, Alexander Toshev,
Oana Camburu, Alan L Yuille và Kevin Murphy.
2016. Tạo và hiểu các mô tả đối tư ợng rõ ràng. Trong
Biên bản báo cáo của IEEE
hội nghị về thị giác máy tính và nhận dạng mẫu, trang
11–20.

7 Thí nghiệm khác Chi tiết

7.1 Chi tiết nhiệm vụ hạ lư u

Chúng tôi đánh giá mPLUG trên sáu nhiệm vụ ngôn ngữ-tầm nhìn
hạ lư u. Các siêu tham số mà chúng tôi sử dụng
để tinh chỉnh các nhiệm vụ hạ lư u đư ợc liệt kê trong
Bảng 11. Sau đây [6], tất cả các tác vụ đều áp dụng RandAug
-ment, trình tối ư u hóa AdamW với sự suy giảm trọng số là
0,05 và một lịch trình tốc độ học cosin. Chúng tôi sử dụng
độ phân giải hình ảnh là 336 × 336, ngoại trừ VQA
nơi chúng tôi sử dụng hình ảnh 504 × 504. Đối với VQA và các
tác vụ chú thích hình ảnh, chúng tôi cũng thực hiện thêm một
quá trình đào tạo trư ớc tiếp tục trên 4M cặp hình ảnh-văn bản,
có thể mang lại sự cải thiện độ chính xác 0,2+. Tiếp theo
chúng tôi giới thiệu chi tiết về cài đặt tập dữ liệu.

VQA. Chúng tôi tiến hành thử nghiệm trên VQA2.0
tập dữ liệu [67], chứa 83k/41k/81k hình ảnh
để đào tạo/xác thực/kiểm tra. Sau [6], chúng tôi sử dụng
cả hai phân chia đào tạo và xác thực cho đào tạo, và
kết hợp dữ liệu đào tạo bổ sung từ Visual
Bộ gen [37].

Nhiệm vụ	Các thời kỳ kích thư ớc lô LR (ViT-L/BERTbase)		
VQA	2e-5/5e-6	1024	8
Phụ đề†	1e-5&8e-7	256	5
Lấy lại	1e-5/2e-6	256	5
Nền tảng thị giác	2e-5/2e-6	512	120
NLVR2	5e-5/5e-6	256	15
SNLI-VE	2e-5	64	5

Bảng 11: Tinh chỉnh siêu tham số cho hạ lư u
nhiệm vụ. † biểu thị hai giai đoạn tinh chỉnh.

Chú thích hình ảnh. Chúng tôi hoàn thiện trên COCO
Đoàn tàu Karpathy tách ra và đánh giá trên COCO
Phân tách thử nghiệm Karpathy và phân tách xác thực No-Caps.
Tiếp theo [4, 35], đầu tiên chúng tôi tinh chỉnh mPLUG với
mất entropy chéo trong 5 thời kỳ với tốc độ học tập
của 1e-5 và kích thư ớc lô là 256. Dựa trên
mô hình tinh chỉnh, chúng tôi tinh chỉnh nó với CIDEr
tối ư u hóa [49] cho 5 kỷ nguyên bổ sung với một
tốc độ học tập của 8e-7. Trong quá trình suy luận, chúng tôi sử dụng
tìm kiếm chùm tia với kích thư ớc chùm tia là 10 và đặt
Độ dài thế hệ tối đa là 20.

Truy xuất hình ảnh-văn bản. Chúng tôi áp dụng phân chia
Karpathy đư ợc sử dụng rộng rãi [68] cho cả COCO và
Flickr30K. COCO chứa 113/5k/5k hình ảnh
để đào tạo/xác thực/kiểm tra và Flickr30K chứa
29k/1k/1k hình ảnh để đào tạo/xác thực/kiểm tra.

Nền tảng trực quan. Chúng tôi đánh giá phư ơng pháp của chúng tôi trên
ba tập dữ liệu nền tảng biểu thức tham chiếu: Re- fCOCO,
RefCOCO+ [16] và RefCOCOg [69].
Các tập dữ liệu RefCOCO và RefCOCO+ chia sẻ 19K
hình ảnh và chứa 142/141K truy vấn. Bộ dữ liệu Re-fCOCOg
chứa 25K hình ảnh và 95K
truy vấn. Để sử dụng đầy đủ dữ liệu đào tạo, trư ớc tiên chúng tôi đào tạo
mô hình với tập dữ liệu hỗn hợp có tốc độ học tập là
2e-5. Sau đó chúng tôi tiếp tục tinh chỉnh mô hình trên
mỗi tập dữ liệu có tốc độ học tập là 2e-6.

NLVR2 & SNLI-VE. Chúng tôi tiến hành thí nghiệm
cả hai đều chia tách chính thức [54, 55].

7.2 Chi tiết về tập dữ liệu trư ớc khi đào tạo

Bảng 12 hiển thị số liệu thống kê của 14M hình ảnh đư ợc
đào tạo trư ớc có chứa văn bản.

	COCO	VG	SBU	CC3M	CC12M
hình ảnh	113K	100K	860K	3M	10M
chữ	567K	769K	860K	3M	10M

Bảng 12: Thống kê của các tập dữ liệu tiền đào tạo.