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

Junnan Li Dongxu Li Caiming Xiong Steven Hoi
Salesforce Research
https://github.com/salesforce/BLIP

Abstract

Vision-Language Pre-training (VLP) has advanced the performance for many vision-language tasks. However, most existing pre-trained models only excel in either understanding-based tasks or generation-based tasks. Furthermore, performance improvement has been largely achieved by scaling up the dataset with noisy image-text pairs collected from the web, which is a suboptimal source of supervision. In this paper, we propose BLIP, a new VLP framework which transfers flexibly to both vision-language understanding and generation tasks. BLIP effectively utilizes the noisy web data by bootstrapping the captions, where a captioner generates synthetic captions and a filter removes the noisy ones. We achieve state-of-the-art results on a wide range of vision-language tasks, such as image-text retrieval (+2.7% in average recall@1), image captioning (+2.8% in CIDEr), and VQA (+1.6% in VQA score). BLIP also demonstrates strong generalization ability when directly transferred to video-language tasks in a zero-shot manner. Code, models, and datasets are released.

1. Introduction

Vision-language pre-training has recently received tremendous success on various multimodal downstream tasks. However, existing methods have two major limitations:

(1) Model perspective: most methods either adopt an encoder-based model (Radford et al., 2021; Li et al., 2021a), or an encoder-decoder (Cho et al., 2021; Wang et al., 2021) model. However, encoder-based models are less straightforward to directly transfer to text generation tasks (*e.g.* image captioning), whereas encoder-decoder models have not been successfully adopted for image-text retrieval tasks.

(2) Data perspective: most state-of-the-art methods (*e.g.*, CLIP (Radford et al., 2021), ALBEF (Li et al., 2021a), SimVLM (Wang et al., 2021)) pre-train on image-text pairs collected from the web. Despite the performance gain obtained by scaling up the dataset, our paper shows that the noisy web text is suboptimal for vision-language learning.

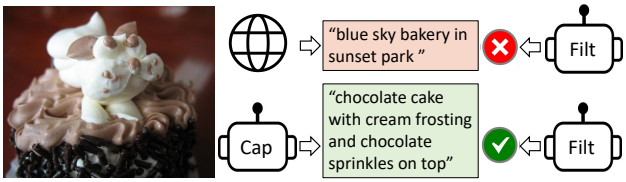


Figure 1. We use a Captioner (Cap) to generate synthetic captions for web images, and a Filter (Filt) to remove noisy captions.

To this end, we propose BLIP: Bootstrapping Language-Image Pre-training for unified vision-language understanding and generation. BLIP is a new VLP framework which enables a wider range of downstream tasks than existing methods. It introduces two contributions from the model and data perspective, respectively:

(a) Multimodal mixture of Encoder-Decoder (MED): a new model architecture for effective multi-task pre-training and flexible transfer learning. An MED can operate either as a unimodal encoder, or an image-grounded text encoder, or an image-grounded text decoder. The model is jointly pre-trained with three vision-language objectives: image-text contrastive learning, image-text matching, and image-conditioned language modeling.

(b) Captioning and Filtering (CapFilt): a new dataset boostrapping method for learning from noisy image-text pairs. We finetune a pre-trained MED into two modules: a *captioner* to produce synthetic captions given web images, and a *filter* to remove noisy captions from both the original web texts and the synthetic texts.

We perform extensive experiments and analysis, and make the following key observations.

- We show that the captioner and the filter work together to achieve substantial performance improvement on various downstream tasks by bootstrapping the captions. We also find that more diverse captions yield larger gains.
- BLIP achieves state-of-the-art performance on a wide range of vision-language tasks, including image-text re-

BLIP：用于统一视觉-语言理解和生成的引导语言-图像预训练

李俊楠 李 Dongxu 李 Caim Xiong Steven Hoi
Salesforce 研究
https://github.com/salesforce/BLIP

抽象

视觉语言预训练 （VLP） 提高了许多视觉语言任务的性能。然而，大多数现有的预训练模型仅在基于理解的任务或基于生成的任务中表现出色。此外，性能改进主要是通过使用从 Web 收集的嘈杂图像-文本对来扩展数据集来实现的，这是一个次优的监督来源。在本文中，我们提出了 BLIP，一种新的 VLP 框架，可灵活地转移到视觉语言理解和生成任务中。BLIP 通过引导字幕来有效地利用嘈杂的 Web 数据，其中字幕编写者生成合成字幕，过滤器删除嘈杂的字幕。我们在广泛的视觉语言任务上取得了最先进的结果，例如图像文本检索（平均 recall@1 +2.7%）、图像字幕（CIDEr 中 +2.8%）和 VQA（+VQA 分数 +1.6%）。当以零镜头方式直接传输到视频语言任务时，BLIP 还表现出很强的泛化能力。发布代码、模型和数据集。

1. 引言

视觉语言预训练最近在各种多模态下游任务上取得了巨大成功。但是，现有方法有两个主要限制：

（1） 模型视角：大多数方法要么采用基于编码器的模型（Radford et al.， 2021;Li等人，2021a）或编码器-解码器（Cho等人，2021 年;Wang et al.， 2021） 模型。然而，基于编码器的模型不太容易直接转移到文本生成任务（例如图像字幕），而编码器-解码器模型尚未成功用于图像-文本检索任务。

（2） 数据视角：大多数最先进的方法（例如，CLIP （Radford et al.， 2021）、ALBEF （Li et al.， 2021a）、SimVLM （Wang et al.， 2021））对图像-文本对进行预训练

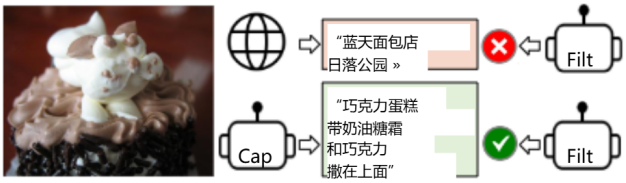


图 1.我们使用 Captioner （Cap） 为 Web 图像生成合成字幕，并使用过滤器 （Filt） 删除杂色字幕。

从 Web 收集。尽管通过扩大数据集获得了性能提升，但我们的论文表明，嘈杂的 Web 文本对于视觉语言学习来说并不理想。

为此，我们提出了 BLIP： Bootstrapping LanguageImage Pre-training 用于统一视觉-语言的理解和生成。BLIP 是一种新的 VLP 框架，与现有方法相比，它支持更广泛的下游任务。它分别从模型和数据的角度介绍了两个贡献：

（a） 编码器-解码器（MED）的多模态混合：一种用于有效多任务预训练和灵活迁移学习的新模型架构。MED 可以用作单模编码器、图像接地文本编码器或图像接地文本解码器。该模型与三个视觉-语言目标联合预训练：图像文本对比学习、图像-文本匹配和图像条件语言建模。

（b） 字幕和过滤 （CapFilt）：一种新的数据集 boostrapping 方法，用于从嘈杂的图像-文本对中学习。我们将预训练的 MED 微调为两个模块：一个用于生成给定 Web 图像的合成字幕的字幕，以及一个用于从原始 Web 文本和合成文本中去除噪点字幕的过滤器。

我们进行了广泛的实验和分析，并做出了以下关键观察。

- 我们展示了字幕生成器和过滤器协同工作，通过引导字幕在各种下游任务上实现实质性的性能改进。我们还发现，字幕越多样化，收益就越大。

- BLIP 在广泛的视觉语言任务上实现了最先进的性能，包括图像-文本重新

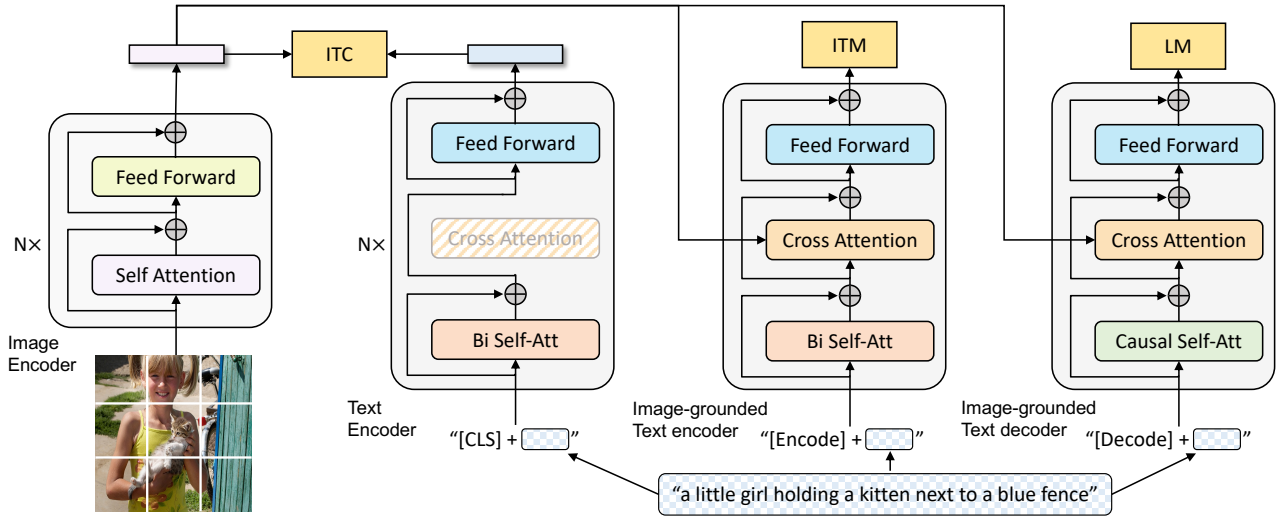


Figure 2. Pre-training model architecture and objectives of BLIP (same parameters have the same color). We propose multimodal mixture of encoder-decoder, a unified vision-language model which can operate in one of the three functionalities: (1) Unimodal encoder is trained with an image-text contrastive (ITC) loss to align the vision and language representations. (2) Image-grounded text encoder uses additional cross-attention layers to model vision-language interactions, and is trained with a image-text matching (ITM) loss to distinguish between positive and negative image-text pairs. (3) Image-grounded text decoder replaces the bi-directional self-attention layers with causal self-attention layers, and shares the same cross-attention layers and feed forward networks as the encoder. The decoder is trained with a language modeling (LM) loss to generate captions given images.

trieval, image captioning, visual question answering, visual reasoning, and visual dialog. We also achieve state-of-the-art zero-shot performance when directly transferring our models to two video-language tasks: text-to-video retrieval and videoQA.

2. Related Work

2.1. Vision-language Pre-training

Vision-language pre-training (VLP) aims to improve performance of downstream vision and language tasks by pre-training the model on large-scale image-text pairs. Due to the prohibitive expense of acquiring human-annotated texts, most methods (Chen et al., 2020; Li et al., 2020; 2021a; Wang et al., 2021; Radford et al., 2021) use image and alt-text pairs crawled from the web (Sharma et al., 2018; Changpinyo et al., 2021; Jia et al., 2021), Despite the use of simple rule-based filters, noise is still prevalent in the web texts. However, the negative impact of the noise has been largely overlooked, shadowed by the performance gain obtained from scaling up the dataset. Our paper shows that the noisy web texts are suboptimal for vision-language learning, and proposes CapFilt that utilizes web datasets in a more effective way.

There have been many attempts to unify various vision and language tasks into a single framework (Zhou et al., 2020; Cho et al., 2021; Wang et al., 2021). The biggest challenge is to design model architectures that can perform both understanding-based tasks (*e.g.* image-text retrieval) and generation-based tasks (*e.g.* image captioning). Neither encoder-based models (Li et al., 2021a;b; Radford et al., 2021) nor encoder-decoder models (Cho et al., 2021; Wang et al., 2021) can excel at both types of tasks, whereas a single unified encoder-decoder (Zhou et al., 2020) also limits the model's capability. Our proposed multimodal mixture of encoder-decoder model offers more flexibility and better performance on a wide range of downstream tasks, in the meantime keeping the pre-training simple and efficient.

2.2. Knowledge Distillation

Knowledge distillation (KD) (Hinton et al., 2015) aims to improve the performance of a student model by distilling knowledge from a teacher model. Self-distillation is a special case of KD where the teacher and student have equal sizes. It has been shown to be effective for image classification (Xie et al., 2020), and recently for VLP (Li et al., 2021a). Different from mostly existing KD methods which simply enforce the student to have the same class predictions as the teacher, our proposed CapFilt can be interpreted as a more effective way to perform KD in the context of VLP, where the captioner distills its knowledge through semantically-rich synthetic captions, and the filter distills its knowledge by removing noisy captions.

2.3. Data Augmentation

While data augmentation (DA) has been widely adopted in computer vision (Shorten & Khoshgoftaar, 2019), DA for language tasks is less straightforward. Recently, generative language models have been used to synthesize examples for various NLP tasks (Kumar et al., 2020; Anaby-Tavor

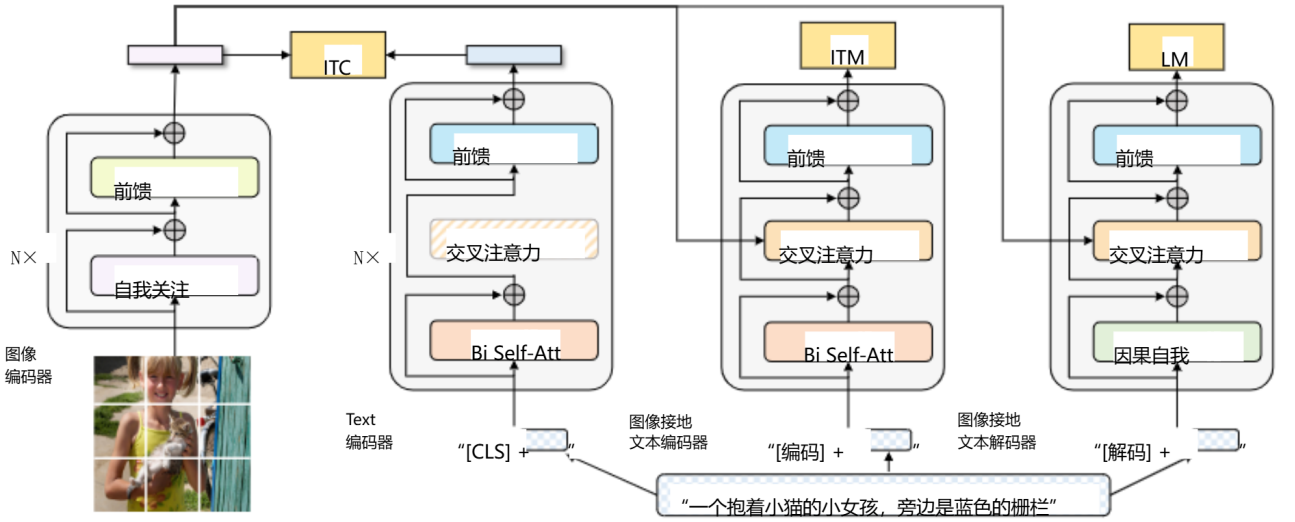


图 2.BLIP 的预训练模型架构和目标（相同的参数具有相同的颜色）。我们提出了编码器-解码器的多模态混合，这是一种统一的视觉-语言模型，可以在以下三种功能之一中运行：（1） 单模编码器用图像-文本对比 （ITC） 损失进行训练，以对齐视觉和语言表示。（2） 图像基于的文本编码器使用额外的交叉注意力层来模拟视觉-语言交互，并使用图像-文本匹配 （ITM） 损失进行训练，以区分正和负的图像-文本对。（3） 图像基于的文本解码器用因果自注意力层代替双向自注意力层，并与编码器共享相同的交叉注意力层和前馈网络。解码器使用语言建模 （LM） 损失进行训练，以生成给定图像的字幕。

Trieval、图像字幕、视觉问答、视觉推理和视觉对话。当将我们的模型直接传输到两个视频语言任务时，我们还实现了最先进的零镜头性能：文本到视频检索和 videoQA。

2. 相关工作

2.1. 视觉语言预训练

视觉-语言预训练 （VLP） 旨在通过在大规模图像-文本对上预训练模型来提高下游视觉和语言任务的性能。由于获取人工注释文本的费用高昂，大多数方法（Chen等人，2020 年;Li et al.， 2020;2021 年a;Wang et al.， 2021;Radford et al.， 2021）使用从网络上抓取的图像和替代文本对（Sharma et al.， 2018;Changpinyo等人，2021 年;Jia et al.， 2021），尽管使用了简单的基于规则的过滤器，但噪声在网络文本中仍然普遍存在。然而，噪声的负面影响在很大程度上被忽视了，被扩大数据集获得的性能提升所掩盖。我们的论文表明，嘈杂的网络文本对于视觉语言学习来说是次优的，并提出了以更有效的方式利用网络数据集的 CapFilt。

已经有许多尝试将各种愿景和语言任务统一到一个框架中（周 et al.， 2020;Cho et al.， 2021;Wang et al.， 2021）。最大的挑战是设计能够同时执行基于理解的任务（例如图像文本检索）和基于生成的任务（例如图像描述）的模型架构

基于编码器的模型（Li等人，2021a;b;Radford等人，2021 年）或编码器-解码器模型（Cho et al.， 2021;Wang et al.， 2021）可以在这两种类型的任务中表现出色，而单个统一的编码器-解码器（周 et al.， 2020）也限制了模型的能力。我们提出的编码器-解码器多模态混合模型在广泛的下游任务上提供了更大的灵活性和更好的性能，同时保持了预训练的简单和高效。

2.2. 知识蒸馏

知识蒸馏 （KD） （Hinton et al.， 2015） 旨在通过从教师模型中提炼知识来提高学生模型的表现。自蒸馏是 KD 的一种特殊情况，其中教师和学生的大小相等。它已被证明对图像分类有效（Xie等人，2020 年），最近对 VLP （Li等人，2021a） 有效。与大多数现有的 KD 方法不同，这些方法只是简单地强制学生与教师进行相同的班级预测，我们提出的 CapFilt 可以解释为在 VLP 上下文中执行 KD 的更有效方法，其中字幕作者通过语义丰富的合成字幕来提炼其知识，而过滤器通过删除嘈杂的字幕来提炼其知识。

2.3. 数据增强

虽然数据增强（DA）已在计算机视觉中广泛采用（Shorter & Khoshgoftaar，2019），但语言任务的DA并不那么简单。最近，生成语言模型已被用于合成各种 NLP 任务的示例（Kumar et al.， 2020;阿纳比-塔沃尔

et al., 2020; Puri et al., 2020; Yang et al., 2020). Different from these methods which focus on the low-resource language-only tasks, our method demonstrates the advantage of synthetic captions in large-scale vision-language pre-training.

3. Method

We propose BLIP, a unified VLP framework to learn from noisy image-text pairs. This section first introduces our new model architecture MED and its pre-training objectives, and then delineates CapFilt for dataset bootstrapping.

3.1. Model Architecture

We employ a visual transformer (Dosovitskiy et al., 2021) as our image encoder, which divides an input image into patches and encodes them as a sequence of embeddings, with an additional `[CLS]` token to represent the global image feature. Compared to using pre-trained object detectors for visual feature extraction (Chen et al., 2020), using a ViT is more computation-friendly and has been adopted by the more recent methods (Li et al., 2021a; Kim et al., 2021).

In order to pre-train a unified model with both understanding and generation capabilities, we propose multimodal mixture of encoder-decoder (MED), a multi-task model which can operate in one of the three functionalities:

(1) **Unimodal encoder**, which separately encodes image and text. The text encoder is the same as BERT (Devlin et al., 2019), where a `[CLS]` token is appended to the beginning of the text input to summarize the sentence.

(2) **Image-grounded text encoder**, which injects visual information by inserting one additional cross-attention (CA) layer between the self-attention (SA) layer and the feed forward network (FFN) for each transformer block of the text encoder. A task-specific `[Encode]` token is appended to the text, and the output embedding of `[Encode]` is used as the multimodal representation of the image-text pair.

(3) **Image-grounded text decoder**, which replaces the bi-directional self-attention layers in the image-grounded text encoder with causal self-attention layers. A `[Decode]` token is used to signal the beginning of a sequence, and an end-of-sequence token is used to signal its end.

3.2. Pre-training Objectives

We jointly optimize three objectives during pre-training, with two understanding-based objectives and one generation-based objective. Each image-text pair only requires one forward pass through the computational-heavier visual transformer, and three forward passes through the text transformer, where different functionalities are activated to compute the three losses as delineated below.

Image-Text Contrastive Loss (ITC) activates the unimodal encoder. It aims to align the feature space of the visual transformer and the text transformer by encouraging positive image-text pairs to have similar representations in contrast to the negative pairs. It has been shown to be an effective objective for improving vision and language understanding (Radford et al., 2021; Li et al., 2021a). We follow the ITC loss by Li et al. (2021a), where a momentum encoder is introduced to produce features, and soft labels are created from the momentum encoder as training targets to account for the potential positives in the negative pairs.

Image-Text Matching Loss (ITM) activates the image-grounded text encoder. It aims to learn image-text multimodal representation that captures the fine-grained alignment between vision and language. ITM is a binary classification task, where the model uses an ITM head (a linear layer) to predict whether an image-text pair is positive (matched) or negative (unmatched) given their multimodal feature. In order to find more informative negatives, we adopt the hard negative mining strategy by Li et al. (2021a), where negatives pairs with higher contrastive similarity in a batch are more likely to be selected to compute the loss.

Language Modeling Loss (LM) activates the image-grounded text decoder, which aims to generate textual descriptions given an image. It optimizes a cross entropy loss which trains the model to maximize the likelihood of the text in an autoregressive manner. We apply a label smoothing of 0.1 when computing the loss. Compared to the MLM loss that has been widely-used for VLP, LM enables the model with the generalization capability to convert visual information into coherent captions.

In order to perform efficient pre-training while leveraging multi-task learning, the text encoder and text decoder share all parameters except for the SA layers. The reason is that the differences between the encoding and decoding tasks are best captured by the SA layers. In particular, the encoder employs *bi-directional* self-attention to build representations for the *current* input tokens, while the decoder employs *causal* self-attention to predict *next* tokens. On the other hand, the embedding layers, CA layers and FFN function similarly between encoding and decoding tasks, therefore sharing these layers can improve training efficiency while benefiting from multi-task learning,

3.3. CapFilt

Due to the prohibitive annotation cost, there exist a limited number of high-quality human-annotated image-text pairs $\{(I_h, T_h)\}$ (e.g., COCO (Lin et al., 2014)). Recent work (Li et al., 2021a; Wang et al., 2021) utilizes a much larger number of image and alt-text pairs $\{(I_w, T_w)\}$ that are automatically collected from the web. However, the alt-texts often do not accurately describe the visual content of the images, making them a noisy signal that is suboptimal for learning vision-language alignment.

等人，2020 年;Puri等人，2020 年;Yang et al.， 2020）。与这些专注于低资源语言任务的方法不同，我们的方法展示了合成字幕在大规模视觉语言预训练中的优势。

3. 方法

我们提出了 BLIP，这是一个统一的 VLP 框架，用于从嘈杂的图像-文本对中学习。本节首先介绍了我们的新模型架构 MED 及其预训练目标，然后描述了用于数据集引导的 CapFilt

3.1. 模型架构

我们使用视觉转换器 （Dosovitskiy et al.， 2021） 作为我们的图像编码器，它将输入图像分成块并将它们编码为嵌入序列，并带有一个额外的 [CLS] 标记来表示全局图像特征。与使用预先训练的对象检测器进行视觉特征提取相比（Chen 等人，2020 年），使用 ViT 对计算更友好，并且已被最近的方法采用（Li et al.， 2021a;Kim et al.， 2021）。

为了预先训练一个具有理解和生成能力的统一模型，我们提出了编码器-解码器（MED）的多模态混合，这是一种可以在以下三种功能之一中运行的多任务模型：

（1） 单模编码器，分别对图像和文本进行编码。文本编码器与 BERT （Devlin et al.， 2019） 相同，其中 [CLS] 标记附加到文本输入的开头以总结句子。（2） 图像接地文本编码器，通过在文本编码器的每个转换器块的自注意力（SA） 层和前馈网络 （FFN） 之间插入一个额外的交叉注意力 （CA） 层来注入视觉信息。特定于任务的 [编码] 令牌将追加到文本中，并且 [编码] 的输出嵌入用作图像-文本对的多模式表示形式。

（3） 图像基于文本解码器，它用因果自注意力层替换了基于图像的文本编码器中的双向自注意力层。[Decode] 令牌用于表示序列的开始，序列结束令牌用于表示其结束。

3.2. 训练前目标

我们在预培训期间共同优化了三个目标，两个基于理解的目标和一个基于世代的目标。每个图像-文本对只需要通过计算较重的视觉转换器的一次正向传递，以及通过文本转换器的三次正向传递，其中激活不同的功能来计算如下所示的三种损失。

图像-文本对比损失 （ITC） 激活单峰编码器。它旨在调整视觉转换的特征空间

前者和文本转换器，方法是鼓励正图像-文本对与负对具有相似的表示形式。它已被证明是提高视力和语言理解的有效目标（Radford 等人，2021 年;Li et al.， 2021a）。我们遵循 Li 等人 （2021a） 的 ITC 损失，其中引入了动量编码器来产生特征，并从动量编码器创建软标签作为训练目标，以解释负对中的潜在正值。

图像-文本匹配损失 （ITM） 激活图像锁定文本编码器。它旨在学习图像-文本多模态表示，以捕捉视觉和语言之间的细粒度对齐。ITM 是一项二元分类任务，其中模型使用 ITM 头（线性层）来预测给定其多模态特征的图像-文本对是正（匹配）还是负（不匹配）。为了找到更多信息丰富的负数，我们采用了 Li et al. （2021a） 的硬负挖掘策略，其中更有可能选择一批中对比相似性较高的负数对来计算损失。

语言建模损失 （LM） 激活基于图像的文本解码器，该解码器旨在为给定图像生成文本描述。它优化了交叉熵损失，训练模型以自回归方式最大化文本的可能性。在计算损失时，我们应用 0.1 的标签平滑。与广泛用于 VLP 的 MLM 损失相比，LM 使模型具有泛化能力，可以将视觉信息转换为连贯的标题。

为了在利用多任务学习的同时执行高效的预训练，文本编码器和文本解码器共享除 SA 层之外的所有参数。原因是编码和解码任务之间的差异最好由 SA 层捕获。特别是，编码器采用双向自我注意来构建当前输入标记的表示，而解码器则采用因果自我注意来预测下一个标记。另一方面，嵌入层、CA 层和 FFN 在编码和解码任务之间的功能相似，因此共享这些层可以提高训练效率，同时受益于多任务学习。

3.3. CapFilt

由于注释成本高昂，存在有限数量的高质量人工注释图像文本对 {（I， T）}（例如，COCO（Lin et al.， 2014））。最近的工作（Li et al.， 2021a;Wang et al.， 2021）利用了大量从网络上自动收集的图像和替代文本对 {（I， T）}。然而，替代文本通常不能准确描述图像的视觉内容，使它们成为一个嘈杂的信号，对于学习视觉-语言对齐来说不是最佳选择。

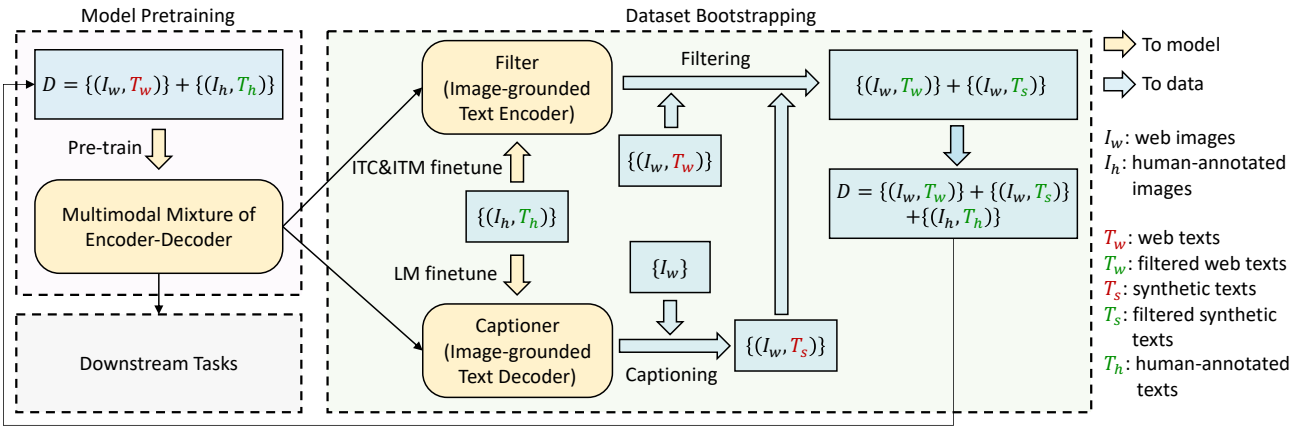


Figure 3. Learning framework of BLIP. We introduce a captioner to produce synthetic captions for web images, and a filter to remove noisy image-text pairs. The captioner and filter are initialized from the same pre-trained model and finetuned individually on a small-scale human-annotated dataset. The bootstrapped dataset is used to pre-train a new model.

We propose Captioning and Filtering (CapFilt), a new method to improve the quality of the text corpus. Figure 3 gives an illustration of CapFilt. It introduces two modules: a *captioner* to generate captions given web images, and a *filter* to remove noisy image-text pairs. Both the captioner and the filter are initialized from the same pre-trained MED model, and finetuned individually on the COCO dataset. The finetuning is a lightweight procedure.

Specifically, the *captioner* is an image-grounded text decoder. It is finetuned with the LM objective to decode texts given images. Given the web images I_w, the captioner generates synthetic captions T_s with one caption per image. The *filter* is an image-grounded text encoder. It is finetuned with the ITC and ITM objectives to learn whether a text matches an image. The filter removes noisy texts in both the original web texts T_w and the synthetic texts T_s, where a text is considered to be noisy if the ITM head predicts it as unmatched to the image. Finally, we combine the filtered image-text pairs with the human-annotated pairs to form a new dataset, which we use to pre-train a new model.

4. Experiments and Discussions

In this section, we first introduce pre-training details. Then we provide a detailed experimental analysis on our method.

4.1. Pre-training Details

Our models are implemented in PyTorch (Paszke et al., 2019) and pre-trained on two 16-GPU nodes. The image transformer is initialized from ViT pre-trained on ImageNet (Touvron et al., 2020; Dosovitskiy et al., 2021), and the text transformer is initialized from $\text{BERT}_{\text{base}}$ (Devlin et al., 2019). We explore two variants of ViTs: ViT-B/16 and ViT-L/16. Unless otherwise specified, all results reported in this paper as "BLIP" uses ViT-B. We pre-train the model for 20 epochs using a batch size of 2880 (ViT-B) / 2400 (ViT-L). We use AdamW (Loshchilov & Hutter, 2017) optimizer with a weight decay of 0.05. The learning rate is warmed-up to $3e$-4 (ViT-B) / $2e$-4 (ViT-L) and decayed linearly with a rate of 0.85. We take random image crops of resolution 224×224 during pre-training, and increase the image resolution to 384×384 during finetuning. We use the same pre-training dataset as Li et al. (2021a) with 14M images in total, including two human-annotated datasets (COCO and Visual Genome (Krishna et al., 2017)), and three web datasets (Conceptual Captions (Changpinyo et al., 2021), Conceptual 12M (Changpinyo et al., 2021), SBU captions (Ordonez et al., 2011)). We also experimented with an additional web dataset, LAION (Schuhmann et al., 2021), which contains 115M images with more noisy texts[1]. More details about the datasets can be found in the appendix.

4.2. Effect of CapFilt

In Table 1, we compare models pre-trained on different datasets to demonstrate the efficacy of CapFilt on downstream tasks, including image-text retrieval and image captioning with finetuned and zero-shot settings.

When only the captioner or the filter is applied to the dataset with 14M images, performance improvement can be observed. When applied together, their effects compliment each other, leading to substantial improvements compared to using the original noisy web texts.

CapFilt can further boost performance with a larger dataset and a larger vision backbone, which verifies its scalability in both the data size and the model size. Furthermore, by using a large captioner and filter with ViT-L, performance of the base model can also be improved.

[1]We only download images whose shorter edge is larger than 256 pixels from the original LAION400M. Due to the large size of LAION, we only use $1/5$ of it each epoch during pre-training.

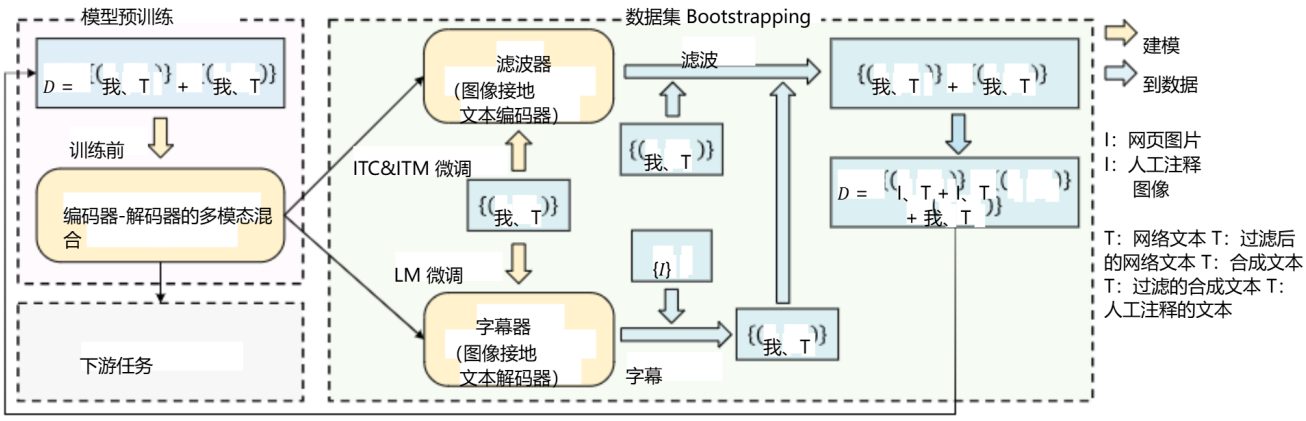


图 3.BLIP 的学习框架。我们引入了一个字幕生成器来为 Web 图像生成合成字幕，以及一个过滤器来删除干扰的图像-文本对。字幕器和过滤器从相同的预训练模型初始化，并在小规模人工注释数据集上单独微调。引导的数据集用于预训练新模型。

我们提出了字幕和过滤 （CapFilt），这是一种提高文本语料库质量的新方法。图 3 给出了 CapFilt 的图示。它介绍了两个模块：一个用于为给定的 Web 图像生成标题的字幕器，以及一个用于删除干扰图像-文本对的过滤器。字幕器和过滤器都是从同一个预先训练的 MED 模型初始化的，并在 COCO 数据集上单独微调。

微调是一个轻量级的过程。

具体来说，字幕器是一种基于图像的文本解码器。它与 LM 物镜进行了微调，以解码给定图像的文本。给定 Web 图像 I，字幕器会生成合成字幕 T，每张图片一个字幕。筛选器是图像基于的文本编码器。它根据 ITC 和 ITM 目标进行微调，以了解文本是否与图像匹配。过滤器会删除原始 Web 文本 T 和合成文本 T 中的干扰文本，如果 ITM 负责人预测文本与图像不匹配，则认为该文本是干扰文本。最后，我们将过滤后的图像-文本对与人工注释对相结合，形成一个新的数据集，用于预训练新模型。

4. 实验和讨论

在本节中，我们首先介绍训练前的详细信息。然后，我们对我们的方法进行详细的实验分析。

4.1. 训练前细节

我们的模型是在 PyTorch 中实现的（Paszke et al.，2019），并在两个 16-GPU 节点上进行了预训练。图像转换器是从 ImageNet 上预训练的 ViT 初始化的（Touvron等人，2020 年;Dosovitskiy et al.， 2021），文本转换器是从 BERT 初始化的（Devlin et al.， 2019）。我们探讨了 ViTs 的两种变体：ViT-B/16 和 ViT-L/16。除非另有说明，否则本文中报告为 “BLIP” 的所有结果均使用 ViT-B

我们使用 2880 （ViT-B） / 2400 （ViT-L） 的批量大小对模型进行了 20 个 epoch 的预训练。我们使用 AdamW （Loshchilov & Hutter， 2017） 优化器，权重衰减为 0.05。学习率预热到 3e-4 （ViT-B） / 2e-4 （ViT-L），并以 0.85 的速率线性衰减。在预训练期间，我们随机裁剪分辨率为 224 × 224 的图像，并在微调期间将图像分辨率提高到 384 × 384。我们使用与 Li 等人（2021a）相同的预训练数据集，总共有 14M 图像，包括两个人工注释的数据集（COCO 和 Visual Genome（Krishna et al.， 2017））和三个网络数据集（概念标题（Changpinyo et al.， 2021）、概念 12M（Changpinyo et al.， 2021）、SBU 标题（Ordonez et al.， 2011））。我们还试验了一个额外的网络数据集 LAION （Schuhmann et al.， 2021），其中包含 115M 图像和更多嘈杂的文本。有关数据集的更多详细信息，请参阅附录。

4.2. CapFilt 的影响

在表 1 中，我们比较了在不同数据集上预先训练的模型，以证明 CapFilt 在下游任务上的有效性，包括具有微调和零镜头设置的图像文本检索和图像字幕。

当仅将 captioner 或过滤器应用于具有 14M 图像的数据集时，可以观察到性能改进。当一起使用时，它们的效果会相辅相成，与使用原始的嘈杂 Web 文本相比，可以带来实质性的改进。

CapFilt 可以通过更大的数据集和更大的视觉主干网进一步提高性能，这验证了它在数据大小和模型大小方面的可扩展性。此外，通过在 ViT-L 中使用大型字幕器和过滤器，还可以提高基本模型的性能。

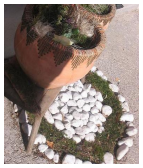
[1]我们只下载原始LAION400M中较短边大于 256 像素的图像。由于 LAION 的体积很大，我们在预训练期间每个 epoch 只使用 1/5。

Pre-train dataset	Bootstrap C	Bootstrap F	Vision backbone	Retrieval-FT (COCO) TR@1	Retrieval-FT (COCO) IR@1	Retrieval-ZS (Flickr) TR@1	Retrieval-ZS (Flickr) IR@1	Caption-FT (COCO) B@4	Caption-FT (COCO) CIDEr	Caption-ZS (NoCaps) CIDEr	Caption-ZS (NoCaps) SPICE
COCO+VG +CC+SBU (14M imgs)	✗	✗	ViT-B/16	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
	✗	✓$_B$		79.1	61.5	94.1	82.8	38.1	128.2	102.7	14.0
	✓$_B$	✗		79.7	62.0	94.4	83.6	38.4	128.9	103.4	14.2
	✓$_B$	✓$_B$		80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4
COCO+VG +CC+SBU +LAION (129M imgs)	✗	✗	ViT-B/16	79.6	62.0	94.3	83.6	38.8	130.1	105.4	14.2
	✓$_B$	✓$_B$		81.9	64.3	96.0	85.0	39.4	131.4	106.3	14.3
	✓$_L$	✓$_L$		81.2	64.1	96.0	85.5	39.7	133.3	109.6	14.7
	✗	✗	ViT-L/16	80.6	64.1	95.1	85.5	40.3	135.5	112.5	14.7
	✓$_L$	✓$_L$		82.4	65.1	96.7	86.7	40.4	136.7	113.2	14.8

Table 1. Evaluation of the effect of the captioner (C) and filter (F) for dataset bootstrapping. Downstream tasks include image-text retrieval and image captioning with finetuning (FT) and zero-shot (ZS) settings. TR / IR@1: recall@1 for text retrieval / image retrieval. ✓$_{B/L}$: captioner or filter uses ViT-B / ViT-L as vision backbone.

T_w: "from bridge near my house"

T_s: "a flock of birds flying over a lake at sunset"



T_w: "in front of a house door in Reichenfels, Austria"

T_s: "a potted plant sitting on top of a pile of rocks"

T_w: "the current castle was built in 1180, replacing a 9th century wooden castle"

T_s: "a large building with a lot of windows on it"

Figure 4. Examples of the web text T_w and the synthetic text T_s. Green texts are accepted by the filter, whereas red texts are rejected.

Generation method	Noise ratio	Retrieval-FT (COCO) TR@1	Retrieval-FT (COCO) IR@1	Retrieval-ZS (Flickr) TR@1	Retrieval-ZS (Flickr) IR@1	Caption-FT (COCO) B@4	Caption-FT (COCO) CIDEr	Caption-ZS (NoCaps) CIDEr	Caption-ZS (NoCaps) SPICE
None	N.A.	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
Beam	19%	79.6	61.9	94.1	83.1	38.4	128.9	103.5	14.2
Nucleus	25%	80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4

Table 2. Comparison between beam search and nucleus sampling for synthetic caption generation. Models are pre-trained on 14M images.

Layers shared	#parameters	Retrieval-FT (COCO) TR@1	Retrieval-FT (COCO) IR@1	Retrieval-ZS (Flickr) TR@1	Retrieval-ZS (Flickr) IR@1	Caption-FT (COCO) B@4	Caption-FT (COCO) CIDEr	Caption-ZS (NoCaps) CIDEr	Caption-ZS (NoCaps) SPICE
All	224M	77.3	59.5	93.1	81.0	37.2	125.9	100.9	13.1
All except CA	252M	77.5	59.9	93.1	81.3	37.4	126.1	101.2	13.1
All except SA	252M	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
None	361M	78.3	60.5	93.6	81.9	37.8	127.4	101.8	13.9

Table 3. Comparison between different parameter sharing strategies for the text encoder and decoder during pre-training.

In Figure 4, we show some example captions and their corresponding images, which qualitatively demonstrate the effect of the captioner to generate new textual descriptions, and the filter to remove noisy captions from both the original web texts and the synthetic texts. More examples can be found in the appendix.

4.3. Diversity is Key for Synthetic Captions

In CapFilt, we employ nucleus sampling (Holtzman et al., 2020) to generate synthetic captions. Nucleus sampling is a stochastic decoding method, where each token is sampled from a set of tokens whose cumulative probability mass exceeds a threshold p ($p = 0.9$ in our experiments). In Table 2, we compare it with beam search, a deterministic decoding method which aims to generate captions with the highest probability. Nucleus sampling leads to evidently better performance, despite being more noisy as suggested by a higher noise ratio from the filter. We hypothesis that the reason is that nucleus sampling generates more diverse and surprising captions, which contain more new information that the model could benefit from. On the other hand, beam search tends to generate safe captions that are common in the dataset, hence offering less extra knowledge.

4.4. Parameter Sharing and Decoupling

During pre-training, the text encoder and decoder share all parameters except for the self-attention layers. In Table 3, we evaluate models pre-trained with different parameter sharing strategies, where pre-training is performed on the 14M images with web texts. As the result shows, sharing all

训练前数据	Bootstrap C	F	愿景骨干	Retrieval-FT （COCO） TR@1	IR@1	Retrieval-ZS （Flickr） TR@1	IR@1	标题-FT （COCO） B@4	CIDEr	标题-ZS （NoCaps） CIDEr	SPICE
COCO+VG 系列 +CC+SBU （14M 图像）	✗	✗	维生素 B/16	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
	✓	✗		79.1	61.5	94.1	82.8	38.1	128.2	102.7	14.0
	✗	✓		80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4
COCO+VG 系列 +CC+SBU +拉昂 （129M 图像）	✗	✗	维生素 B/16	79.6	62.0	94.3	83.6	38.8	130.1	105.4	14.2
	✓	✗		81.9	64.3	95.6	85.0	39.4	131.4	106.3	14.3
	✗	✓		81.2	64.1	96.0	85.5	39.7	133.3	109.6	14.7
	✗	✗	维生素 L/16	80.6	64.1	95.1	85.5	40.3	135.5	112.5	14.7
	✗	✓		82.4	65.1	96.7	86.7	40.4	136.7	113.2	14.8

表 1.评估字幕器 （C） 和过滤器 （F） 对数据集引导的影响。下游任务包括图像文本检索和具有微调 （FT） 和零样本 （ZS） 设置的图像字幕。TR / IR@1：用于文本检索/图像检索的recall@1。3：字幕器或滤镜使用 ViT-B / ViT-L 作为视觉支柱。

T： “从桥在我家附近”

T： “日落时分一群鸟飞过湖”

T： “在奥地利赖兴费尔斯的一扇房子门前”

T： “一盆盆栽植物坐在一堆石头上”

T： “现在的城堡建于 1180 年，取代了一座 9 世纪的木制城堡”

T： “一栋有很多窗户的大建筑”

图 4.Web 文本 T 和合成文本的示例 T. 过滤器接受绿色文本，而拒绝红色文本。

代方法	噪声率	检索-FT （COCO） TR@1	IR@1	检索-ZS （Flickr） TR@1	IR@1	标题-FT （COCO） B@4	CIDEr	标题-ZS （NoCaps） CIDEr	SPICE
无	不适用	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
光束	19%	79.6	61.9	94.1	83.1	38.4	128.9	103.5	14.2
核心	25%	80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4

表 2.光束搜索和原子核采样在合成标题生成中的比较。模型在 14M 图像上进行预训练。

图层共享 #parameters	检索-FT （COCO） TR@1	IR@1	检索-ZS （Flickr） TR@1	IR@1	标题-FT （COCO） B@4	CIDEr	标题-ZS （NoCaps） CIDEr	SPICE
所有 224M	77.3	59.5	93.1	81.0	37.2	125.9	100.9	13.1
除 CA 外的所有 252M	77.5	59.9	93.1	81.3	37.4	126.1	101.2	13.1
除 SA 外的所有 252M	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
无 361M	78.3	60.5	93.6	81.9	37.8	127.4	101.8	13.9

表 3.预训练期间文本编码器和解码器的不同参数共享策略之间的比较。

在图 4 中，我们展示了一些示例字幕及其相应的图像，它们定性地展示了字幕生成器生成新文本描述的效果，以及过滤器从原始 Web 文本和合成文本中删除干扰字幕的效果。更多示例可在附录中找到。

4.3. 多样性是合成字幕的关键

在 CapFilt 中，我们采用细胞核采样（Holtzman et al.，2020）来生成合成字幕。核采样是一种随机解码方法，其中每个标记都是从一组累积概率质量超过阈值 p 的标记中采样的（在我们的实验中为 p = 0.9）。在表 2 中，我们将其与光束搜索进行了比较一种确定性解码方法，旨在生成具有最高概率的字幕。Nucleus 采样明显可以带来更好的性能，尽管滤波器的噪声比更高，因此噪声更大。我们假设原因是细胞核采样生成了更多样化和令人惊讶的标题，其中包含更多模型可以从中受益的新信息。另一方面，光束搜索往往会生成数据集中常见的安全字幕，因此提供的额外知识较少。

4.4. 参数共享和解耦

在预训练期间，文本编码器和解码器共享除自注意力层之外的所有参数。在表 3 中，我们评估了使用不同参数共享策略进行预训练的模型，其中对带有 Web 文本的 14M 图像进行预训练。如结果所示，共享所有

Captioner & Filter	Noise ratio	Retrieval-FT (COCO) TR@1	Retrieval-FT (COCO) IR@1	Retrieval-ZS (Flickr) TR@1	Retrieval-ZS (Flickr) IR@1	Caption-FT (COCO) B@4	Caption-FT (COCO) CIDEr	Caption-ZS (NoCaps) CIDEr	Caption-ZS (NoCaps) SPICE
Share parameters	8%	79.8	62.2	94.3	83.7	38.4	129.0	103.5	14.2
Decoupled	25%	80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4

Table 4. Effect of sharing parameters between the captioner and filter. Models are pre-trained on 14M images.

Method	Pre-train # Images	COCO (5K test set) TR R@1	TR R@5	TR R@10	IR R@1	IR R@5	IR R@10	Flickr30K (1K test set) TR R@1	TR R@5	TR R@10	IR R@1	IR R@5	IR R@10
UNITER (Chen et al., 2020)	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
VILLA (Gan et al., 2020)	4M	-	-	-	-	-	-	87.9	97.5	98.8	76.3	94.2	96.8
OSCAR (Li et al., 2020)	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
UNIMO (Li et al., 2021b)	5.7M	-	-	-	-	-	-	89.4	98.9	99.8	78.0	94.2	97.1
ALIGN (Jia et al., 2021)	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 (Li et al., 2021a)	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
BLIP	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	129M	**81.9**	95.4	97.8	**64.3**	85.7	91.5	**97.3**	**99.9**	**100.0**	87.3	97.6	**98.9**
$\text{BLIP}_{\text{CapFilt-L}}$	129M	81.2	**95.7**	**97.9**	64.1	**85.8**	**91.6**	97.2	**99.9**	**100.0**	**87.5**	**97.7**	**98.9**
$\text{BLIP}_{\text{ViT-L}}$	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

Table 5. Comparison with state-of-the-art image-text retrieval methods, finetuned on COCO and Flickr30K datasets. $\text{BLIP}_{\text{CapFilt-L}}$ pre-trains a model with ViT-B backbone using a dataset bootstrapped by captioner and filter with ViT-L.

Method	Pre-train # Images	Flickr30K (1K test set) TR R@1	TR R@5	TR R@10	IR R@1	IR R@5	IR R@10
CLIP	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN	1.8B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	14M	94.1	99.5	99.7	82.8	96.3	98.1
BLIP	14M	94.8	99.7	**100.0**	84.9	96.7	98.3
BLIP	129M	**96.0**	**99.9**	**100.0**	85.0	**96.8**	98.6
$\text{BLIP}_{\text{CapFilt-L}}$	129M	**96.0**	**99.9**	**100.0**	**85.5**	**96.8**	**98.7**
$\text{BLIP}_{\text{ViT-L}}$	129M	96.7	100.0	100.0	86.7	97.3	98.7

Table 6. Zero-shot image-text retrieval results on Flickr30K.

layers except for SA leads to better performance compared to not sharing, while also reducing the model size thus improveing training efficiency. If the SA layers are shared, the model's performance would degrade due to the conflict between the encoding task and the decoding task.

During CapFilt, the captioner and the filter are end-to-end finetuned individually on COCO. In Table 4, we study the effect if the captioner and filter share parameters in the same way as pre-training. The performance on the downstream tasks decreases, which we mainly attribute to *confirmation bias*. Due to parameter sharing, noisy captions produced by the captioner are less likely to be filtered out by the filter, as indicated by the lower noise ratio (8% compared to 25%).

5. Comparison with State-of-the-arts

In this section, we compare BLIP to existing VLP methods on a wide range of vision-language downstream tasks[2]. Next we briefly introduce each task and finetuning strategy. More details can be found in the appendix.

5.1. Image-Text Retrieval

We evaluate BLIP for both image-to-text retrieval (TR) and text-to-image retrieval (IR) on COCO and Flickr30K (Plummer et al., 2015) datasets. We finetune the pre-trained model using ITC and ITM losses. To enable faster inference speed, we follow Li et al. (2021a) and first select k candidates based on the image-text feature similarity, and then rerank the selected candidates based on their pairwise ITM scores. We set $k = 256$ for COCO and $k = 128$ for Flickr30K.

As shown in Table 5, BLIP achieves substantial performance improvement compared with existing methods. Using the same 14M pre-training images, BLIP outperforms the previous best model ALBEF by +2.7% in average recall@1 on COCO. We also perform zero-shot retrieval by directly transferring the model finetuned on COCO to Flickr30K. The result is shown in Table 6, where BLIP also outperforms existing methods by a large margin.

5.2. Image Captioning

We consider two datasets for image captioning: NoCaps (Agrawal et al., 2019) and COCO, both evaluated using the model finetuned on COCO with the LM loss. Similar as Wang et al. (2021), we add a prompt "a picture of" at the beginning of each caption, which leads to slightly better results. As shown in Table 7, BLIP with 14M pre-training images substantially outperforms methods using a similar amount of pre-training data. BLIP with 129M images achieves competitive performance as LEMON with

[2] we omit SNLI-VE from the benchmark because its test data has been reported to be noisy (Do et al., 2020)

字幕员 & 滤波器	噪声率	检索-FT （COCO） TR@1	检索-ZS （Flickr） IR@1	标题-FT （COCO） B@4	标题-ZS （NoCaps） CIDEr	CIDEr	SPICE	TR@1	IR@1
份额参数	8%	79.8	62.2	94.3	83.7	38.4	129.0	103.5	14.2
脱钩	25%	80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4

表 4.在字幕器和滤镜之间共享参数的效果。模型在 14M 图像上进行预训练。

方法	预训练	COCO （5K 测试集）				Flickr30K （1K 测试集）				# 图片	TR		IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10					
UNITER （Chen et al.， 2020）	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		
VILLA （Gan et al.， 2020）	4M	-	-	-	-	-	-	87.9	97.5	98.8	76.3	94.2	96.8		
OSCAR （Li et al.， 2020）	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-		
UNIMO （Li et al.， 2021b）	5.7M	-	-	-	-	-	-	89.4	98.9	99.8	78.0	94.2	97.1		
ALIGN （Jia et al.， 2021）	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 （Li et al.， 2021a）	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		
点点	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		
点点	129M	81.9	95.4	97.8	64.3	85.7	91.5	97.3	99.9	100.0	87.3	97.6	98.9		
BLIPCapFilt-L	129M	81.2	95.7	97.9	64.1	85.8	91.6	97.2	99.9	100.0	87.5	97.7	98.9		
BLIP	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		

表 5.与最先进的图像文本检索方法进行比较，在 COCO 和 Flickr30K 数据集上进行了微调。BLIP使用由字幕器引导的数据集和ViT-L过滤器预训练具有ViT-B主干的模型。

方法	预训练	Flickr30K （1K 测试集）					# 图片	TR		IR		
		R@1	R@5	R@10	R@1	R@5	R@10					
剪辑	400M	88.0	98.7	99.4	68.7	90.6	95.2					
对齐	1.8B	88.6	98.7	99.7	75.7	93.8	96.8					
ALBEF	14M	94.1	99.5	99.7	82.8	96.3	98.1					
点	14M	94.8	99.7	100.0	84.9	96.7	98.3					
点点	129M	96.0	99.9	100.0	85.0	96.8	98.6					
点点	129M	96.0	99.9	100.0	85.5	96.8	98.7					
BLIP	129M	96.7	100.0	100.0	86.7	97.3	98.7					

表 6.Flickr30K 上的零样本图像文本检索结果。

与不共享相比，除 SA 以外的层可以带来更好的性能，同时还可以减小模型大小，从而提高训练效率。如果 SA 层是共享的，则由于编码任务和解码任务之间的冲突，模型的性能会下降。

在 CapFilt 期间，字幕器和过滤器在 COCO 上单独进行端到端微调。在表 4 中，我们研究了 captioner 和 filter 以与预训练相同的方式共享参数的影响。下游任务的性能下降，我们主要将其归因于确认偏差。由于参数共享，字幕制作者生成的嘈杂字幕不太可能被滤镜过滤掉，如较低的噪声率（8% 对 25%）所示。

5. 与最先进的比较

在本节中，我们将 BLIP 与现有的 VLP 方法在广泛的视觉语言下游任务上进行了比较。下一个

[2] 我们从基准测试中省略了 SNLI-VE，因为据报道其测试数据有噪声（Do等人，2020 年）

我们简要介绍了每个任务和微调策略。更多详细信息可在附录中找到。

5.1. 图像文本检索

我们在 COCO 和 Flickr30K （Plummer et al.， 2015） 数据集上评估了图像到文本检索 （TR） 和文本到图像检索 （IR） 的 BLIP 。我们使用 ITC 和 ITM 损失对预训练模型进行微调。为了实现更快的推理速度，我们遵循 Li et al. （2021a） 的做法，首先根据图像文本特征相似性选择 k 个候选者，然后根据成对 ITM 分数对选定的候选者进行重新排序。

我们为 COCO 设置 k = 256，为 Flickr30K 设置 k = 128。

如表 5 所示，与现有方法相比，BLIP 实现了显著的性能改进。使用相同的 14M 预训练图像，BLIP 在 COCO 上的平均 recall@1比之前的最佳模型 ALBEF 高出 +2.7%。我们还通过将 COCO 上微调的模型直接传输到 Flickr30K 来执行零样本检索。结果如表 6 所示，其中 BLIP 的性能也大大优于现有方法。

5.2. 图像字幕

我们考虑了两个用于图像描述的数据集：NoCaps（Agrawal 等人，2019 年）和 COCO，两者都使用在 COCO 上微调的模型进行了评估，并带有 LM 损失。与 Wang et al. （2021） 类似，我们在每个标题的开头添加了一个提示 “a picture of”，这会导致结果略好。如表 7 所示，具有 14M 预训练图像的 BLIP 的性能大大优于使用类似数量的预训练数据的方法。具有 129M 图像的 BLIP 具有与 LEMON 一样具有竞争力的性能

Method	Pre-train #Images	NoCaps validation in-domain C	in-domain S	near-domain C	near-domain S	out-domain C	out-domain S	overall C	overall S	COCO Caption Karpathy test B@4	COCO Caption Karpathy test C
Enc-Dec (Changpinyo et al., 2021)	15M	92.6	12.5	88.3	12.1	94.5	11.9	90.2	12.1	-	110.9
VinVL† (Zhang et al., 2021)	5.7M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3
$\text{LEMON}_{\text{base}}$† (Hu et al., 2021)	12M	104.5	14.6	100.7	14.0	96.7	12.4	100.4	13.8	-	-
$\text{LEMON}_{\text{base}}$† (Hu et al., 2021)	200M	107.7	14.7	106.2	14.3	107.9	13.1	106.8	14.1	**40.3**	**133.3**
BLIP	14M	111.3	15.1	104.5	14.4	102.4	13.7	105.1	14.4	38.6	129.7
BLIP	129M	109.1	14.8	105.8	14.4	105.7	13.7	106.3	14.3	39.4	131.4
$\text{BLIP}_{\text{CapFilt-L}}$	129M	**111.8**	**14.9**	**108.6**	**14.8**	**111.5**	**14.2**	**109.6**	**14.7**	39.7	**133.3**
$\text{LEMON}_{\text{large}}$† (Hu et al., 2021)	200M	116.9	15.8	113.3	15.1	111.3	14.0	113.4	15.0	40.6	135.7
$\text{SimVLM}_{\text{huge}}$ (Wang et al., 2021)	1.8B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
$\text{BLIP}_{\text{ViT-L}}$	129M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	40.4	136.7

Table 7. Comparison with state-of-the-art image captioning methods on NoCaps and COCO Caption. All methods optimize the cross-entropy loss during finetuning. C: CIDEr, S: SPICE, B@4: BLEU@4. $\text{BLIP}_{\text{CapFilt-L}}$ is pre-trained on a dataset bootstrapped by captioner and filter with ViT-L. VinVL† and LEMON† require an object detector pre-trained on 2.5M images with human-annotated bounding boxes and high resolution (800×1333) input images. $\text{SimVLM}_{\text{huge}}$ uses 13× more training data and a larger vision backbone than ViT-L.

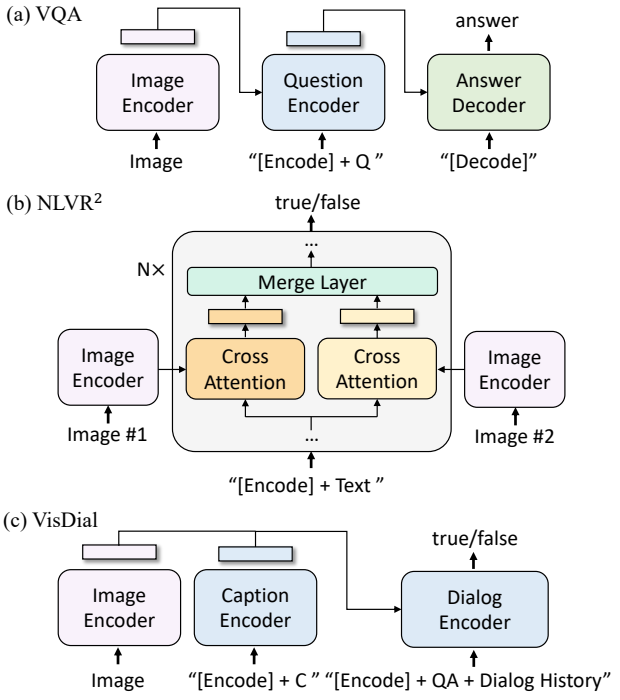


Figure 5. Model architecture for the downstream tasks. Q: question; C: caption; QA: question-answer pair.

200M images. Note that LEMON requires a computational-heavy pre-trained object detector and higher resolution (800×1333) input images, leading to substantially slower inference time than the detector-free BLIP which uses lower resolution (384×384) input images.

5.3. Visual Question Answering (VQA)

VQA (Antol et al., 2015) requires the model to predict an answer given an image and a question. Instead of formulating VQA as a multi-answer classification task (Chen et al., 2020; Li et al., 2020), we follow Li et al. (2021a) and consider it as an answer generation task, which enables open-ended VQA. As shown in Figure 5(a), during finetuning, we rearrange the pre-trained model, where an image-question is first encoded into multimodal embeddings and then given to an answer decoder. The VQA model is finetuned with the LM loss using ground-truth answers as targets.

Method	Pre-train #Images	VQA test-dev	VQA test-std	NLVR^2 dev	NLVR^2 test-P
LXMERT	180K	72.42	72.54	74.90	74.50
UNITER	4M	72.70	72.91	77.18	77.85
VL-T5/BART	180K	-	71.3	-	73.6
OSCAR	4M	73.16	73.44	78.07	78.36
SOHO	219K	73.25	73.47	76.37	77.32
VILLA	4M	73.59	73.67	78.39	79.30
UNIMO	5.6M	75.06	75.27	-	-
ALBEF	14M	75.84	76.04	82.55	83.14
$\text{SimVLM}_{\text{base}}$†	1.8B	77.87	78.14	81.72	81.77
BLIP	14M	77.54	77.62	**82.67**	82.30
BLIP	129M	78.24	78.17	82.48	**83.08**
$\text{BLIP}_{\text{CapFilt-L}}$	129M	**78.25**	**78.32**	82.15	82.24

Table 8. Comparison with state-of-the-art methods on VQA and NLVR^2. ALBEF performs an extra pre-training step for NLVR^2. SimVLM† uses 13× more training data and a larger vision backbone (ResNet+ViT) than BLIP.

The results are shown in Table 8. Using 14M images, BLIP outperforms ALBEF by +1.64% on the test set. Using 129M images, BLIP achieves better performance than SimVLM which uses 13× more pre-training data and a larger vision backbone with an additional convolution stage.

5.4. Natural Language Visual Reasoning (NLVR^2)

NLVR^2 (Suhr et al., 2019) asks the model to predict whether a sentence describes a pair of images. In order to enable rea-

方法	训练前 #Images	COCO Caption 域内 Karpathy 测试 B@4	C	S	近域 nocaps 验证 B@4	C	S	域外 nocaps 验证 B@4	C	S	整体 Karpathy 测试 B@4	C	S
Enc-Dec （Changpinyo et al.， 2021）	15M	92.6	12.5	88.3	12.1	94. 5	11.9	90.2	12.1	-	110.9		
VinVL† （Zhang et al.， 2021）	5.7M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3		
LEMON† （胡 et al.， 2021）	12M	104.5	14.6	100.7	14.0	96.7	12.4	100.4	13.8	-	LEMON† （胡 et al.， 2021）	200M	107.7	14.7	106.2	14.3	107.9	13.1	106.8	14.1	40.3	133.3
点点	14M	111.3	15.1	104.5	14.4	102.4	13.7	105.1	14.4	38.6	129.7	点点	129M	109.1	14.8	105.8	14.4	105.7	13.7	106.3	14.3	39.4	131.4
BLIPCapFilt-L	129M	111.8	14.9	108.6	14.8	111.5	14.2	109.6	14.7	39.7	133.3												
LEMON† （胡 et al.， 2021）	200M	116.9	15.8	113.3	15.1	111.3	14.0	113.4	15.0	40.6	135.7		
SimVLM（Wang et al.， 2021）	1.8B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3		
BLIP	129M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	40.4	136.7		

表 7.与 NoCaps 和 COCO Caption 上最先进的图像字幕方法的比较。所有方法都优化了微调期间的交叉熵损失。C： CIDEr， S： 香料， B@4： BLEU@4.BLIPCapFilt-L 在由字幕器引导的数据集上进行预训练，并使用 ViT-L 进行过滤器。VinVL† 和 LEMON †需要一个对象检测器，该检测器在 2.5M 图像上进行了预训练，带有人工注释的边界框和高分辨率 （800×1333） 输入图像。SimVLM使用的训练数据比 ViT-L 多 13× 个，并且视觉主干更大。

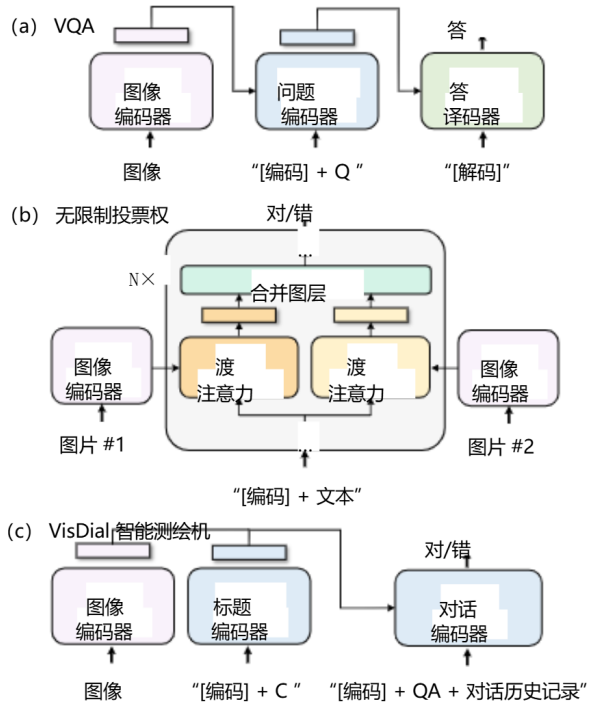


图 5.下游任务的模型架构。Q：问题;C： 标题;QA：问答对。

200M 图像。请注意，LEMON 需要计算量大的预训练对象检测器和更高分辨率 （800×1333） 的输入图像，这导致推理时间比使用较低分辨率 （384×384） 输入图像的无检测器 BLIP 慢得多。

方法	训练前 #Images	VQA test-dev	VQA test-std	NLVR dev	NLVR test-P
LXMERT	180K	72.（英语：LXMERT	180K	72.）	42	72.54	74.90	74.50
UNITER	4M	72.70	72.91	77.18	77.85
VL-T5/BART	180K	-	71.3	-	73.6
OSCAR	4M	73.16	73.44	78.07	78.36
SOHO	219K	73.25	73.47	76.37	77.32
别墅	4M	73.59	73.67	78.39	79.30
UNIMO	5.6M	75.06	75.27	-	
ALBEF	14M	75.84	76.04	82.55	83.14
SimVLM†	1.8B	77.87	78.14	81.72	81.77
点	14M	77.54	77.62	82.67	82.30
点	129M	78.24	78.17	82.48	83.08
BLIP	129M	78.25	78.32	82.15	82.24

表 8.与 VQA 和 NLVR 的最新方法进行比较。ALBEF 为 NLVR 执行额外的预训练步骤。SimVLM† 使用的训练数据比 BLIP 多 13× 和更大的视觉主干 （ResNet+ViT）。

5.3. 视觉问答 （VQA）

VQA （Antol et al.， 2015） 要求模型在给定图像和问题的情况下预测答案。与其将 VQA 表述为多答案分类任务 （Chen et al.， 2020；Li et al.， 2020），我们遵循 Li et al. （2021a） 并将其视为一个答案生成任务，从而实现开放式 VQA。如图 5（a） 所示，在微调过程中，我们重新排列预训练模型，其中图像问题首先被编码为多模态嵌入，然后提供给答案解码器。VQA 模型使用真实答案作为目标，使用 LM 损失进行微调。

结果如表 8 所示。使用 14M 张图像，BLIP 在测试集中的表现比 ALBEF 高出 +1.64%。使用 129M 图像，BLIP 实现了比 SimVLM 更好的性能，SimVLM 使用多 13× 的预训练数据和更大的视觉主干和一个额外的卷积阶段。

5.4. 自然语言视觉推理 （NLVR）

NLVR（Suhr et al.， 2019） 要求模型预测一个句子是否描述了一对图像。为了启用 rea-

Method	MRR↑	R@1↑	R@5↑	R@10↑	MR↓
VD-BERT	67.44	54.02	83.96	92.33	3.53
VD-ViLBERT†	69.10	55.88	85.50	93.29	3.25
BLIP	**69.41**	**56.44**	**85.90**	**93.30**	**3.20**

Table 9. Comparison with state-of-the-art methods on VisDial v1.0 validation set. VD-ViLBERT† (Murahari et al., 2020) pre-trains ViLBERT (Lu et al., 2019) with additional VQA data.

soning over two images, we make a simple modification to our pre-trained model which leads to a more computational-efficient architecture than previous approaches (Li et al., 2021a; Wang et al., 2021). As shown in Figure 5(b), for each transformer block in the image-grounded text encoder, there exist two cross-attention layers to process the two input images, and their outputs are merged and fed to the FFN. The two CA layers are intialized from the same pre-trained weights. The merge layer performs simple average pooling in the first 6 layers of the encoder, and performs concatenation followed by a linear projection in layer 6-12. An MLP classifier is applied on the output embedding of the `[Encode]` token. As shown in Table 8, BLIP outperforms all existing methods except for ALBEF which performs an extra step of customized pre-training. Interestingly, performance on NLVR^2 does not benefit much from additional web images, possibly due to the domain gap between web data and downstream data.

5.5. Visual Dialog (VisDial)

VisDial (Das et al., 2017) extends VQA in a natural conversational setting, where the model needs to predict an answer not only based on the image-question pair, but also considering the dialog history and the image's caption. We follow the discriminative setting where the model ranks a pool of answer candidates (Gan et al., 2019; Wang et al., 2020; Murahari et al., 2020). As shown in Figure 5(c), we concatenate image and caption embeddings, and pass them to the dialog encoder through cross-attention. The dialog encoder is trained with the ITM loss to discriminate whether the answer is true or false for a question, given the entire dialog history and the image-caption embeddings. As shown in Table 9, our method achieves state-of-the-art performance on VisDial v1.0 validation set.

5.6. Zero-shot Transfer to Video-Language Tasks

Our image-language model has strong generalization ability to video-language tasks. In Table 10 and Table 11, we perform zero-shot transfer to *text-to-video retrieval* and *video question answering*, where we directly evaluate the models trained on COCO-retrieval and VQA, respectively. To process video input, we uniformly sample n frames per video ($n = 8$ for retrieval and $n = 16$ for QA), and concatenate the frame features into a single sequence. Note that this simple approach ignores all temporal information.

Method	R1↑	R5↑	R10↑	MdR↓
zero-shot				
ActBERT (Zhu & Yang, 2020)	8.6	23.4	33.1	36
SupportSet (Patrick et al., 2021)	8.7	23.0	31.1	31
MIL-NCE (Miech et al., 2020)	9.9	24.0	32.4	29.5
VideoCLIP (Xu et al., 2021)	10.4	22.2	30.0	-
FiT (Bain et al., 2021)	18.7	39.5	51.6	10
BLIP	**43.3**	**65.6**	**74.7**	**2**
finetuning				
ClipBERT (Lei et al., 2021)	22.0	46.8	59.9	6
VideoCLIP (Xu et al., 2021)	30.9	55.4	66.8	-

Table 10. Comparisons with state-of-the-art methods for text-to-**video** retrieval on the 1k test split of the MSRVTT dataset.

Method	MSRVTT-QA	MSVD-QA
zero-shot		
VQA-T (Yang et al., 2021)	2.9	7.5
BLIP	19.2	35.2
finetuning		
HME (Fan et al., 2019)	33.0	33.7
HCRN (Le et al., 2020)	35.6	36.1
VQA-T (Yang et al., 2021)	41.5	46.3

Table 11. Comparisons with state-of-the-art methods for **video** question answering. We report top-1 test accuracy on two datasets.

Despite the domain difference and lack of temporal modeling, our models achieve state-of-the-art performance on both video-language tasks. For text-to-video retrieval, zero-shot BLIP even outperforms models finetuned on the target video dataset by +12.4% in recall@1. Further performance improvement can be achieved if the BLIP model is used to initialize a video-language model with temporal modeling (*e.g.* replace our ViT with a TimeSformer (Bertasius et al., 2021)) and finetuned on video data.

6. Additional Ablation Study

In this section, we provide additional ablation experiments on CapFilt.

Improvement with CapFilt is not due to longer training. Since the bootstrapped dataset contains more texts than the original dataset, training for the same number of epochs takes longer with the bootstrapped dataset. To verify that the effectiveness of CapFilt is not due to longer training, we replicate the web text in the original dataset so that it has the same number of training samples per epoch as the bootstrapped dataset. As shown in Table 12, longer training using the noisy web texts does not improve performance.

A new model should be trained on the bootstrapped dataset. The bootstrapped dataset is used to pre-train a new model. We investigate the effect of continue training

方法	MRR↑	R@1↑	R@5↑	R@10↑	MR↓
VD-BERT	67.44	54.02	83.96	92.33	3.53
VD-ViLBERT†	69.10	55.88	85.50	93.29	3.25
闪光点	69.41	56.44	85.90	93.30	3.20

表 9.与 VisDial v1.0 验证集上的最新方法进行比较。VD-ViLBERT†（Murahari et al.， 2020） 使用额外的 VQA 数据对 ViLBERT (Lu et al.， 2019) 进行预训练

在两张图像上，我们对预训练模型进行了简单的修改，从而获得了比以前的方法更具计算效率的架构（Li et al.， 2021a;Wang et al.， 2021）。如图 5（b） 所示，对于图像接地文本编码器中的每个 transformer 模块，存在两个交叉注意力层来处理两个输入图像，它们的输出被合并并馈送到 FFN。这两个 CA 层从相同的预训练权重初始化。合并层在编码器的前 6 层中执行简单的平均池化，并在第 6-12 层中执行串联，然后执行线性投影。MLP 分类器应用于 [Encode] 令牌的输出嵌入。如表 8 所示，BLIP 的性能优于所有现有方法，但 ALBEF 除外，ALBEF 执行额外的自定义预训练步骤。有趣的是，NLVR的性能并没有从额外的Web图像中受益匪浅，这可能是由于Web数据和下游数据之间的域差距。

5.5. 可视对话框 （VisDial）

VisDial （Das et al.， 2017） 在自然对话环境中扩展了 VQA，其中模型不仅需要根据图像-问题对预测答案，还需要考虑对话历史和图像的标题。我们遵循判别性设置，其中模型对一组答案候选人进行排名（Gan et al.， 2019;Wang et al.， 2020;Murahari et al.， 2020）。如图 5（c） 所示，我们将图像和标题嵌入连接起来，并通过交叉注意将它们传递给对话编码器。对话编码器使用 ITM 损失进行训练，以区分问题的答案是真还是假，给定整个对话历史记录和图像标题嵌入。如表 9 所示，我们的方法在 VisDial v1.0 验证集上实现了最先进的性能。

5.6. 零镜头传输到视频语言任务

我们的图像语言模型对视频语言任务具有很强的泛化能力。在表 10 和表 11 中，我们执行了对文本到视频检索和视频问答的零镜头转移，分别直接评估了在 COCO 检索和 VQA 上训练的模型。为了处理视频输入，我们对每个视频的 n 帧进行统一采样（n = 8 用于检索，n = 16 用于 QA），并将帧特征连接到一个序列中。请注意，这种简单的方法会忽略所有时态信息。

方法	R1↑	R5↑	R10↑	MdR↓
零点				
ActBERT （Zhu & Yang， 2020）	8.6	23.4	33.1	36
SupportSet （Patrick et al.， 2021）	8.7	23.0	31.1	31
MIL-NCE （Miech et al.， 2020）	9.9	24.0	32.4	29.5
VideoCLIP （Xu et al.， 2021）	10.4	22.2	30.0	
FiT （Bain et al.， 2021）	18.7	39.5	51.6	10
BLIP	43.3	65.6	74.7	2
微调				
ClipBERT （Lei et al.， 2021）	22.0	46.8	59.9	6
VideoCLIP （Xu et al.， 2021）	30.9	55.4	66.8	-

表 10.与 MSRVTT 数据集的 1k 测试拆分上文本到视频检索的最新方法的比较。

方法	MSRVTT-QA	MSVD-QA
零点		
VQA-T（Yang et al.， 2021）	2.9	7.5
条目	19.2	35.2
微调		
HME（Fan等人，2019 年）	33.0	33.7
HCRN（Le 等人，2020 年）	35.6	36.1
VQA-T（Yang 等人，2021 年）	41.5	46.3

表 11.与最先进的视频问答方法的比较。我们报告了两个数据集的顶级测试准确性。

尽管存在领域差异且缺乏时间建模，但我们的模型在两个视频语言任务上都实现了最先进的性能。对于文本到视频的检索，zeroshot BLIP 的性能甚至比在目标视频数据集上微调的模型高出 +12.4% recall@1。如果使用 BLIP 模型来初始化具有时间建模的视频语言模型（例如，用 TimeSformer 替换我们的 ViT（Bertasius et al.， 2021））并对视频数据进行微调，则可以进一步提高性能。

6. 额外的消融研究

在本节中，我们提供了关于 CapFilt 的其他烧蚀实验。

CapFilt 的改进不是由于更长的训练时间。
由于引导数据集包含的文本比原始数据集多，因此使用引导数据集训练相同数量的 epoch 需要更长的时间。为了验证 CapFilt 的有效性不是由于更长的训练时间，我们在原始数据集中复制了 Web 文本，以便它每个时期的训练样本数量与引导数据集相同。如表 12 所示，使用嘈杂的 Web 文本进行较长时间的训练并不能提高性能。

新模型应该在 bootstrapped 上训练
数据。引导的数据集用于预训练新模型。我们调查了持续培训的效果

CapFilt	#Texts	Retrieval-FT (COCO)		Retrieval-ZS (Flickr)		Caption-FT (COCO)		Caption-ZS (NoCaps)	
		TR@1	IR@1	TR@1	IR@1	B@4	CIDEr	CIDEr	SPICE
No	15.3M	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
No	24.7M	78.3	60.5	93.7	82.2	37.9	127.7	102.1	14.0
Yes	24.7M	80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4

Table 12. The original web texts are replicated to have the same number of samples per epoch as the bootstrapped dataset. Results verify that the improvement from CapFilt is not due to longer training time.

Continue	Retrieval-FT (COCO)		Retrieval-ZS (Flickr)		Caption-FT (COCO)		Caption-ZS (NoCaps)	
	TR@1	IR@1	TR@1	IR@1	B@4	CIDEr	CIDEr	SPICE
Yes	80.6	63.0	94.5	84.6	38.5	129.9	104.5	14.2
No	80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4

Table 13. Continue training the pre-trained model offers less gain compared to training a new model with the bootstrapped dataset.

from the previous pre-trained model, using the bootstrapped dataset. Table 13 hows that continue training does not help. This observation agrees with the common practice in knowledge distillation, where the student model cannot be initialized from the teacher.

7. Conclusion

We propose BLIP, a new VLP framework with state-of-the-art performance on a wide range of downstream vision-language tasks, including understanding-based and generation-based tasks. BLIP pre-trains a multimodal mixture of encoder-decoder model using a dataset bootstrapped from large-scale noisy image-text pairs by injecting diverse synthetic captions and removing noisy captions. Our bootstrapped dataset are released to facilitate future vision-language research.

There are a few potential directions that can further enhance the performance of BLIP: (1) Multiple rounds of dataset bootstrapping; (2) Generate multiple synthetic captions per image to further enlarge the pre-training corpus; (3) Model ensemble by training multiple different captioners and filters and combining their forces in CapFilt. We hope that our paper motivates future work to focus on making improvements in both the model aspect and the data aspect, the bread and butter of vision-language research.

References

Agrawal, H., Anderson, P., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., and Lee, S. nocaps: novel object captioning at scale. In *ICCV*, pp. 8947–8956, 2019.

Anaby-Tavor, A., Carmeli, B., Goldbraich, E., Kantor, A., Kour, G., Shlomov, S., Tepper, N., and Zwerdling, N. Do not have enough data? deep learning to the rescue! In *AAAI*, pp. 7383–7390, 2020.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. VQA: visual question answering. In *ICCV*, pp. 2425–2433, 2015.

Bain, M., Nagrani, A., Varol, G., and Zisserman, A. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021.

Bertasius, G., Wang, H., and Torresani, L. Is space-time attention all you need for video understanding? In *ICML*, 2021.

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

Chen, Y., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. UNITER: universal image-text representation learning. In *ECCV*, volume 12375, pp. 104–120, 2020.

Cho, J., Lei, J., Tan, H., and Bansal, M. Unifying vision-and-language tasks via text generation. *arXiv preprint arXiv:2102.02779*, 2021.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M. F., Parikh, D., and Batra, D. Visual dialog. In *CVPR*, pp. 1080–1089, 2017.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *NAACL*, pp. 4171–4186, 2019.

Do, V., Camburu, O.-M., Akata, Z., and Lukasiewicz, T. e-snli-ve: Corrected visual-textual entailment with natural language explanations. *arXiv preprint arXiv:2004.03744*, 2020.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

CapFilt #Texts	检索-FT （COCO） 检索-ZS （Flickr） 标题-FT （COCO） 标题-ZS （NoCaps） TR@1 IR@1 TR@1 IR@1 B@4 CIDEr CIDEr SPICE
否 15.3M 78.4 60.7 93.9 82.1 38.0 127.8 102.2 13.9 否 24.7M 78.3 60.5 93.7 82.2 37.9 127.7 102.1 14.0 是 24.7M 80.6 63.1 94.8 84.9 38.6 129.7 105.1 14.4	

表 12.原始 Web 文本将被复制，以在每个 epoch 中具有与引导数据集相同的样本数。结果验证了 CapFilt 的改进不是由于更长的训练时间。

继续	检索-FT （COCO） 检索-ZS （Flickr） 标题-FT （COCO） 标题-ZS （NoCaps） TR@1 IR@1 TR@1 IR@1 B@4 CIDEr CIDEr SPICE
是 80.6 63.0 94.5 84.6 38.5 129.9 104.5 14.2 否 80.6 63.1 94.8 84.9 38.6 129.7 105.1 14.4	

表 13.继续训练 与使用引导数据集训练新模型相比，预训练模型的增益较低。

从前面的预训练模型中，使用 bootstrapped 数据集。表 13 继续训练如何无济于事。这一观察结果与知识提炼中的常见做法一致，即学生模型不能从教师那里初始化。

7. 总结

我们提出了 BLIP，这是一个新的 VLP 框架，在广泛的下游视觉语言任务上具有最先进的性能，包括基于理解和基于生成的任务。BLIP 使用从大规模嘈杂图像文本对引导的数据集，通过注入各种合成字幕和删除嘈杂字幕，预先训练编码器-解码器模型的多模态混合。我们的 bootstrap 数据集发布是为了促进未来的视觉语言研究。

有几个潜在的方向可以进一步提高 BLIP 的性能：（1） 多轮数据集引导;（2） 为每张图像生成多个合成字幕，以进一步放大预训练语料;（3） 通过在 CapFilt 中训练多个不同的字幕和过滤器并结合它们的力量来建模集成。我们希望我们的论文能激励未来的工作，专注于在模型方面和数据方面进行改进，这是视觉语言研究的基本要素。

引用

Agrawal，H.，Anderson，P.，Desai，K.，Wang，Y.，Chen，X.，Jain，R.，Johnson，M.，Batra，D.，Parikh，D.，Lee，S.，和 Anderson，P. nocaps：大规模的新奇对象标题。在 ICCV 中，第 1 页。8947–8956, 2019.

Anaby-Tavor，A.，Carmeli，B.，Goldbraich，E.，Kantor，A.，Kour，G.，Shlomov，S.，Tepper，N.，和 Zwerdling，N.没有足够的数据？深度学习来拯救！在 AAAI，第 7383-7390 页，2020 年。

安托尔，S.，阿格拉瓦尔，A.，卢，J.，米切尔，M.，巴特拉，D.，Zitnick，CL 和 Parikh，D. VQA：视觉问答。在 ICCV，第 2425-2433 页，2015 年。

Bain，M.，Nagrani，A.，Varol，G.，和 Zisserman，A. 时间冻结：用于端到端检索的联合视频和图像编码器。在 ICCV，2021 年。

Bertasius，G.，Wang，H.，和 Torresani，L.时空注意力是理解视频所需要的吗？在 ICML 中，2021 年。

Changpinyo，S.，Sharma，P.，Ding，N.，和 Soricut，R. 概念 12M：推动网络规模的图像文本预训练以识别长尾视觉概念。在 CVPR，2021 年。

Chen，Y.，Li，L.，Yu，L.，Kholy，AE，Ahmed，F.，Gan，Z.，Cheng，Y.，and Liu，J. UNITER：通用图像文本表示学习。ECCV，第 12375 卷，第 1 页。

104–120, 2020.

Cho，J.，Lei，J.，Tan，H.，and Bansal，M. 通过文本生成统一视觉和语言任务。arXiv 预印本 arXiv：2102.02779,2021 年。

Das，A.，Kottur，S.，Gupta，K.，Singh，A.，Yadav，D.，Moura，JMF，Parikh，D.，和 Batra，D. 视觉对话框。在 CVPR，第 1080-1089 页，2017 年。

Devlin，J.，Chang，M.，Lee，K.，和 Toutanova，K. BERT：用于语言理解的深度双向转换器的预训练。在 Burstein，J.、Doran，C. 和 Solorio，T.（编辑）中，NAACL，第 4171-4186 页，2019 年。

Do，V.，Camburu，O.-M.，Akata，Z.，and Lukasiewicz，T. esnli-ve：用自然修正的视觉文本蕴涵

语言解释。arXiv 预印本 arXiv：2004.03744,2020 年。

Dosovitskiy，A.，Beyer，L.，Kolesnikov，A.，Weissenborn，D.，Zhai，X.，Unterthiner，T.，Dehghani，M.，Minderer，M.，Heigold，G.

一张图像胜过 16x16 个单词：用于大规模图像识别的 Transformers。在 ICLR，2021 年。

Fan, C., Zhang, X., Zhang, S., Wang, W., Zhang, C., and Huang, H. Heterogeneous memory enhanced multimodal attention model for video question answering. In *CVPR*, pp. 1999–2007, 2019.

Gan, Z., Cheng, Y., Kholy, A. E., Li, L., Liu, J., and Gao, J. Multi-step reasoning via recurrent dual attention for visual dialog. In Korhonen, A., Traum, D. R., and Màrquez, L. (eds.), *ACL*, pp. 6463–6474, 2019.

Gan, Z., Chen, Y., Li, L., Zhu, C., Cheng, Y., and Liu, J. Large-scale adversarial training for vision-and-language representation learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *NeurIPS*, 2020.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pp. 6325–6334, 2017.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration. In *ICLR*, 2020.

Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., and Wang, L. Scaling up vision-language pre-training for image captioning, 2021.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

Karpathy, A. and Li, F. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pp. 3128–3137, 2015.

Kim, J., Jun, J., and Zhang, B. Bilinear attention networks. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *NIPS*, pp. 1571–1581, 2018.

Kim, W., Son, B., and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*, 2021.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

Kumar, V., Choudhary, A., and Cho, E. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*, 2020.

Le, T. M., Le, V., Venkatesh, S., and Tran, T. Hierarchical conditional relation networks for video question answering. In *CVPR*, pp. 9972–9981, 2020.

Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T. L., Bansal, M., and Liu, J. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, pp. 7331–7341, 2021.

Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S., Xiong, C., and Hoi, S. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021a.

Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., and Wang, H. UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *ACL*, pp. 2592–2607, 2021b.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., and Gao, J. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, pp. 121–137, 2020.

Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: common objects in context. In Fleet, D. J., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *ECCV*, volume 8693, pp. 740–755, 2014.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *NeurIPS*, pp. 13–23, 2019.

Miech, A., Alayrac, J.-B., Smaira, L., Laptev, I., Sivic, J., and Zisserman, A. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, pp. 9879–9889, 2020.

Murahari, V., Batra, D., Parikh, D., and Das, A. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J. (eds.), *ECCV*, pp. 336–352, 2020.

Ordonez, V., Kulkarni, G., and Berg, T. L. Im2text: Describing images using 1 million captioned photographs. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. C. N., and Weinberger, K. Q. (eds.), *NIPS*, pp. 1143–1151, 2011.

Fan，C.，Zhang，X.，Zhang，S.，Wang，W.，Zhang，C.，and Huang，H. 用于视频问答的异构记忆增强多模态注意力模型。在 CVPR，第 1999-2007 页，2019 年。

甘 Z.，程 Y.，Kholy，A. E.，李 L.，刘 J. 和高 J.

通过重复的双重注意进行多步骤推理，以实现视觉对话。在 Korhonen，A.、Traum，DR 和 M 'arquez，L.（编辑）中，ACL，第 6463-6474 页，2019 年。

甘 Z.，陈 Y.，李 L.，朱 C.，程 Y. 和刘 J.
用于视觉和语言表示学习的大规模对抗性训练。在 Larochelle，H.、Ranzato，M.、Hadsell，R.、Balcan，M. 和 Lin，H.（编辑）中，NeurIPS，2020 年。

Goyal，Y.，Khot，T.，Summers-Stay，D.，Batra，D.，和 Parikh，D. 让 VQA 中的 V 很重要：提升图像理解在视觉问答中的作用

在 CVPR，第 6325-6334 页，2017 年。

Hinton，G.、Vinyals，O. 和 Dean，J. 在神经网络中提炼知识。arXiv 预印本 arXiv：1503.02531,2015 年。

Holtzman，A.，Buys，J.，Du，L.，Forbes，M. 和 Choi，Y.
神经文本退化的奇特案例。在 ICLR，2020 年。

胡，X.，甘，Z.，王，J.，杨，Z.，刘，Z.，卢，Y.，和王，L. 扩大图像描述的视觉语言预训练，2021.

Jia，C.，Yang，Y.，Xia，Y.，Chen，Y.-T.，Parekh，Z.，Pham，H.，Le，Q. V.，Sung，Y.，Li，Z.，and Duerig，T. 扩大视觉和视觉语言表征学习

嘈杂的文本监督。arXiv 预印本 arXiv：2102.05918,2021 年。

Karpathy，A. 和 Li，F. 用于生成图像描述的深度视觉语义对齐。CVPR，第 3128-3137 页，2015 年。

Kim，J.，Jun，J.，和 Zhang，B. 双线性注意力网络。

在 Bengio，S.、Wallach，HM、Larochelle，H.、Grauman，K.、Cesa-Bianchi，N. 和 Garnett，R.（编辑），NeurIPS，2018.

Kim，W.，Son，B. 和 Kim，I. Vilt：没有卷积或区域监督的视觉和语言转换器。
arXiv 预印本 arXiv：2102.03334,2021 年。

克里希纳，R.，朱，Y.，格罗斯，O.，约翰逊，J.，哈塔，K.，克拉维茨，J.，陈，S.，卡兰蒂迪斯，Y.，李，L.，沙玛，DA，Bernstein，MS 和 Fei-Fei，L. 视觉基因组：使用众包密集图像注释连接语言和视觉。IJCV，123（1）：32-73 2017 年

Kumar，V.，Choudhary，A.，和 Cho，E. 使用预训练的 transformer 模型进行数据增强。arXiv 预印本 arXiv：2003.02245 2020 年

Le，TM，Le，V.，Venkatesh，S.，和 Tran，T. 用于视频问答的分层条件关系网络。CVPR，第 9972-9981 页，2020 年。

Lei，J.，Li，L.，周，L.，Gan，Z.，Berg，T. L.，Bansal，M.，和 Liu，J. 少即是多：Clipbert 通过稀疏采样进行视频和语言学习。在 CVPR 中，第 7331-7341 页，2021 年。

Li，J.，Selvaraju，RR，Gotmare，AD，Joty，S.，Xiong，C.，and Hoi，S. 融合前对齐：使用动量蒸馏进行视觉和语言表示学习。在 NeurIPS 中，2021a。

Li，W.，Gao，C.，Niu，G.，Xiao，X.，Liu，H.，Liu，J.，Wu，H.，and Wang，H. UNIMO：通过跨模态对比学习实现统一模态理解和生成。在 Zong，C.、Xia，F.、Li，W. 和 Navigli，R.

（编辑），ACL，第 2592-2607 页，2021b。

李晓东，尹晓东，李俊杰，张平，胡晓明，张丽玲，王俊杰，L.，胡，H.，Dong，L.，Wei，F.，Choi，Y.，和 Gao，J. Oscar：视觉语言任务的对象语义对齐预训练。在 ECCV，第 121-137 页，2020 年

Lin，T.，Maire，M.，Belongie，SJ，Hays，J.，Perona，P.，Ramanan，D.，Doll 'ar，P.，and Zitnick，CL Microsoft COCO：上下文中的常见对象。在 Fleet，DJ、Pajdla，T.、Schiele，B. 和 Tuytelaars，T.（编辑）中，ECCV，第 8693 卷，第 740-755 页，2014 年。

Loshchilov，I. 和 Hutter，F. 解耦的权重衰减正则化。arXiv 预印本 arXiv：1711.05101,2017 年。

Lu，J.，Batra，D.，Parikh，D.，and Lee，S. Vilbert：用于视觉和语言任务的预训练任务不可知的视觉语言表示。在 Wallach，HM、Larochelle，H.、Beygelzimer，A.、d'Alch 'e-Buc，F.、Fox，EB 和 Garnett，R.（编辑）中，NeurIPS，第 13-23 页，2019 年。

Miech，A.，Alayrac，J.-B.，Smaira，L.，Laptev，I.，Sivic，J.，和 Zisserman，A. 从未策划的教学视频中端到端学习视觉表示。在 CVPR，第 9879-9889 页，2020 年。

Murahari，V.，Batra，D.，Parikh，D.，和 Das，A. 视觉对话的大规模预训练：一个简单的最先进的基线。在 Vedaldi，A.、Bischof，H.、Brox，T. 和 Frahm，J.（编辑）中，ECCV，第 336-352 页，2020 年。

Ordonez，V.、Kulkarni，G. 和 Berg，TL Im2text：使用 100 万张带标题的照片描述图像。在

Shawe-Taylor，J.、Zemel，R. S.、Bartlett，P. L.、Pereira，F. C. N. 和 Weinberger，KQ（eds.），NIPS，第 1143-1151 页，

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32:8026–8037, 2019.

Patrick, M., Huang, P.-Y., Asano, Y., Metze, F., Hauptmann, A. G., Henriques, J. F., and Vedaldi, A. Support-set bottlenecks for video-text representation learning. In *ICLR*, 2021.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pp. 2641–2649, 2015.

Puri, R., Spring, R., Shoeybi, M., Patwary, M., and Catanzaro, B. Training question answering models from synthetic data. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *EMNLP*, pp. 5811–5826, 2020.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Gurevych, I. and Miyao, Y. (eds.), *ACL*, pp. 2556–2565, 2018.

Shorten, C. and Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *J. Big Data*, 6:60, 2019.

Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., and Artzi, Y. A corpus for reasoning about natural language grounded in photographs. In Korhonen, A., Traum, D. R., and Màrquez, L. (eds.), *ACL*, pp. 6418–6428, 2019.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

Wang, Y., Joty, S. R., Lyu, M. R., King, I., Xiong, C., and Hoi, S. C. H. VD-BERT: A unified vision and dialog transformer with BERT. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *EMNLP*, pp. 3325–3338, 2020.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

Xie, Q., Luong, M., Hovy, E. H., and Le, Q. V. Self-training with noisy student improves imagenet classification. In *CVPR*, pp. 10684–10695, 2020.

Xu, H., Ghosh, G., Huang, P.-Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., and Feichtenhofer, C. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, pp. 6787–6800, 2021.

Yang, A., Miech, A., Sivic, J., Laptev, I., and Schmid, C. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, pp. 1686–1697, 2021.

Yang, Y., Malaviya, C., Fernandez, J., Swayamdipta, S., Bras, R. L., Wang, J., Bhagavatula, C., Choi, Y., and Downey, D. G-daug: Generative data augmentation for commonsense reasoning. In Cohn, T., He, Y., and Liu, Y. (eds.), *EMNLP Findings*, pp. 1008–1025, 2020.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. Vinvl: Making visual representations matter in vision-language models. *arXiv preprint arXiv:2101.00529*, 2021.

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J. J., and Gao, J. Unified vision-language pre-training for image captioning and VQA. In *AAAI*, pp. 13041–13049, 2020.

Zhu, L. and Yang, Y. Actbert: Learning global-local video-text representations. In *CVPR*, pp. 8746–8755, 2020.

Paszke，A.，Gross，S.，Massa，F.，Lerer，A.，Bradbury，J.，Chanan，G.，Killeen，T.，Lin，Z.，Gimelshein，N.，Antiga，L.，et al. Pytorch：命令式高性能深度学习库。神经IPS，32：8026–8037,2019 年。

Patrick，M.，Huang，P.-Y.，Asano，Y.，Metze，F.，Hauptmann，AG，Henriques，JF，and Vedaldi，A. 支持视频文本表示学习的瓶颈。在 ICLR，2021 年。

Plummer，BA，Wang，L.，塞万提斯，CM，凯塞多，JC，

Hockenmaier，J. 和 Lazebnik，S. Flickr30k 实体：为更丰富的图像到句子模型收集区域到短语的对应关系。在 ICCV，第 2641-2649 页，2015 年。

Puri，R.，Spring，R.，Shoeybi，M.，Patwary，M.，和 Catanzaro，B. 从合成数据中训练问答模型。在 Webber，B.，Cohn，T.，He，Y. 和 Liu，Y.（编辑），EMNLP，第 5811-5826 页，2020 年。

Radford，A.，Kim，JW，Hallacy，C.，Ramesh，A.，Goh，G.，Agarwal，S.，Sastry，G.，Askell，A.，Mishkin，P.，Clark，J.，et al.从自然中学习可转移语言监督。arXiv 预印本 arXiv：2103.00020,2021 年。

Schuhmann，C.，Vencu，R.，Beaumont，R.，Kaczmarczyk，R.，Mullis，C.，Katta，A.，Coombes，T.，Jitsev，J.，and Komatsuzaki，A. Laion-400m：剪辑过滤的 4 亿个图像文本对的开放数据集。arXiv 预印本 arXiv：2111.02114,2021 年。

Sharma，P.，Ding，N.，Goodman，S.，和 Soricut，R. 概念字幕：用于自动图像字幕的清理、上位词状图像替代文本数据集。在 Gurevych，I. 和 Miyao，Y.（eds.），ACL，第 2556-2565 页，2018 年。

Shorten，C. 和 Khoshgoftaar，TM深度学习的图像数据增强调查。大数据杂志，6：60，2019.

Suhr，A.，周，S.，张，A.，张，I.，白，H.，和 Artzi，Y.一个基于照片的自然语言推理语料库。在 Korhonen，A.、Traum，DR 和 M'arquez，L.（编辑）中，ACL，第 6418-6428 页，2019 年。

Touvron，H.，Cord，M.，Douze，M.，Massa，F.，Sablayrolles，A.，和 J 'egou，H. 通过注意力训练数据高效的图像转换器和蒸馏。arXiv 预印本 arXiv：2012.12877，2020。

Wang，Y.，Joty，S. R.，Lyu，MR，King，I.，Xiong，C.，and Hoi，S. C. H. VD-BERT：带有 BERT 的统一视觉和对话转换器。在 Webber，B.、Cohn，T.、He，Y. 和 Liu，Y.（编辑）中，EMNLP，第 3325-3338 页，2020 年。

Wang，Z.，Yu，J.，Yu，AW，Dai，Z.，Tsvetkov，Y.，and Cao，Y. Simvlm：简单的视觉语言模型预训练监督薄弱。arXiv 预印本 arXiv：2108.10904,2021 年。

Xie，Q.，Luong，M.，Hovy，EH，and Le，Q. V.与嘈杂的学生进行自我训练可以提高图像网络分类。在 CVPR，第 10684-10695 页，2020 年。

Xu，H.，Ghosh，G.，Huang，P.-Y.，Okhonko，D.，Aghajanyan，A.，Metze，F.，Zettlemoyer，L.，和 Feichtenhofer，C. 视频剪辑：用于零镜头视频文本理解的对比预训练。在 EMNLP，第 6787-6800 页，2021 年。

Yang，A.，Miech，A.，Sivic，J.，Laptev，I.，和 Schmid，C. 只需提问：学习回答数百万个旁白视频中的问题。在 ICCV，第 1686-1697 页，2021 年。

Yang，Y.，Malaviya，C.，Fernandez，J.，Swayamdipta，S.，Bras，RL，Wang，J.，Bhagavatula，C.，Choi，Y.，and Downey，D. G-daug：用于常识推理的生成数据增强。在 Cohn，T.、He（编辑），EMNLP 调查结果，第 1008-1025 页，2020 年。

Zhang，P.，Li，X.，胡，X.，Yang，J.，Zhang，L.，Wang，L.，Choi，Y.，and Gao，J. Vinvl：使视觉表示在视觉语言模型中变得重要。arXiv 预印本 arXiv：2101.00529,2021 年。

周，L.，Palangi，H.，张，L.，胡，H.，Corso，J. J.，和 Gao，J. 图像字幕和 VQA 的统一视觉语言预训练。在 AAAI，第 13041-13049 页，2020 年。

Zhu，L. 和 Yang，Y. Actbert：学习全局-局部视频文本表示。CVPR，第 8746-8755 页，2020 年。

A. Downstream Task Details

Table 14 shows the hyperparameters that we use for finetuning on the downstream vision-language tasks. All tasks uses AdamW optimizer with a weight decay of 0.05 and a cosine learning rate schedule. We use an image resolution of 384×384, except for VQA where we follow Wang et al. (2021) and use 480×480 images. Next we delineate the dataset details.

Image-Text Retrieval. We use the Karpathy split (Karpathy & Li, 2015) 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.

Image Captioning. We finetune on COCO's Karpathy train split, and evaluate on COCO's Karpathy test split and NoCaps validation split. During inference, we use beam search with a beam size of 3, and set the maximum generation length as 20.

VQA. We experiment with the VQA2.0 dataset (Goyal et al., 2017), which contains 83k/41k/81k images for training/validation/test. Following Li et al. (2021a), we use both training and validation splits for training, and include additional training samples from Visual Genome. During inference on VQA, we use the decoder to rank the 3,128 candidate answers (Li et al., 2021a; Kim et al., 2018).

NLVR^2. We conduct experiment on the official split (Suhr et al., 2019).

VisDial. We finetune on the training split of VisDial v1.0 and evaluate on its validation set.

Task	init LR (ViT-L)	batch size	#epoch
Retrieval	$1e^{-5}$ ($5e^{-6}$)	256	6
Captioning	$1e^{-5}$ ($2e^{-6}$)	256	5
VQA	$2e^{-5}$	256	10
NLVR^2	$3e^{-5}$	256	15
VisDial	$2e^{-5}$	240	20

Table 14. Finetuning hyperparameters for downstream tasks.

B. Additional Examples of Synthetic Captions

In Figure 6, we show additional examples of images and texts where the web captions are filtered out, and the synthetic captions are kept as clean training samples.

C. Pre-training Dataset Details

Table 15 shows the statistics of the pre-training datasets.

	COCO	VG	SBU	CC3M	CC12M	LAION
# image	113K	100K	860K	3M	10M	115M
# text	567K	769K	860K	3M	10M	115M

Table 15. Statistics of the pre-training datasets.

Figure 6. Examples of the web text T_w and the synthetic text T_s. Green texts are accepted by the filter, whereas red texts are rejected.

A. 下游任务详细信息

表 14 显示了我们用于微调下游视觉语言任务的超参数。所有任务都使用权重衰减为 0.05 和余弦学习率计划的 AdamW 优化器。我们使用 384 × 384 的图像分辨率，除了 VQA，我们遵循 Wang 等人（2021 年）并使用 480 × 480 图像。接下来，我们描述数据集的详细信息。

图像文本检索。我们使用Karpathy分割（Karpathy & Li，2015）来表示COCO和Flickr30K。COCO 包含 113/5k/5k 图像用于训练/验证/测试，Flickr30K 包含 29k/1k/1k 图像用于训练/验证/测试。

图像字幕。我们对 COCO 的 Karpathy 列车拆分进行了微调，并评估了 COCO 的 Karpathy 测试拆分和 NoCaps 验证拆分。在推理过程中，我们使用 beam 大小为 3 的 beam 搜索，并将最大生成长度设置为 20。

VQA 的。我们用 VQA2.0 数据集（Goyal et al.， 2017）进行实验，其中包含 83k/41k/81k 图像用于训练/验证/测试。按照 Li 等人（2021a）的说法，我们使用训练和验证拆分进行训练，并包括来自 Visual Genome 的额外训练样本。在 VQA 的推理过程中，我们使用解码器对 3,128 个候选答案进行排名（Li et al.， 2021a;Kim et al.， 2018）。

Task	init LR （ViT-L） 批量大小 #epoch
检索 1e（5e） 256 6 字幕 1e（2e） 256 5 VQA 2e 256 10 NLVR 3e 256 15 VisDial 2e 240 20	

表 14.微调下游任务的超参数。

B. 合成字幕的其他示例

在图 6 中，我们展示了图像和文本的其他示例，其中 Web 字幕被过滤掉，合成字幕被保存为干净的训练样本。

C. 预训练数据集详细信息

表 15 显示了预训练数据集的统计数据。

	可可 VG SBU CC3M CC12M LAION
# 图片 113K 100K 860K 3M 10M 115M # 文字 567K 769K 860K 3M 10M 115M	

表 15.预训练数据集的统计信息。

NLVR 的。我们对官方分裂进行了实验（Suhr et al.，2019）。

VisDial 的。我们对 VisDial v1.0 的训练拆分进行了微调，并对其验证集进行了评估。

图 6.Web 文本 T 和合成文本的示例 T. 过滤器接受绿色文本，而拒绝红色文本。