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
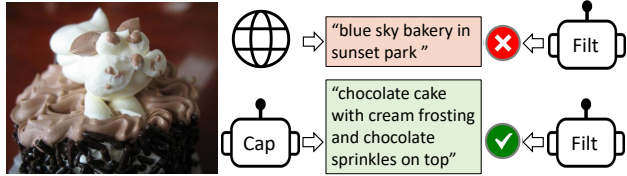

Figure 1. We use a Captioner (Cap) to generate synthetic captions for web images, and a Filter (Filt) to remove noisy captions.

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.

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 bootstrapping 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-
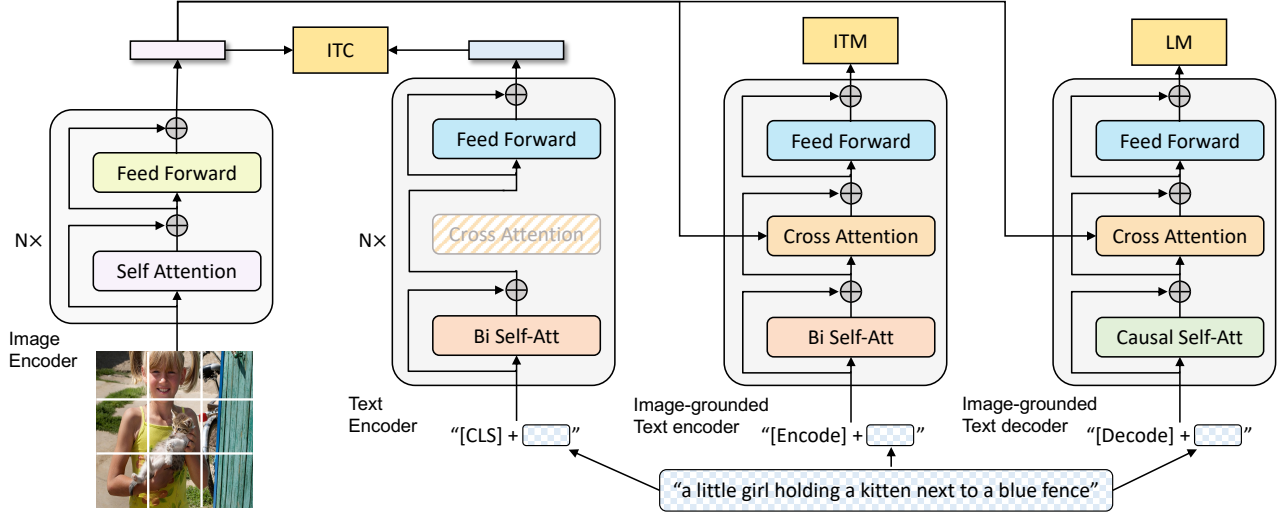
Figure 2. Pre-training model architecture and objectives of BLIP (same parameters have the same color). We propose multimodal mixture of encoder-decoder, a unified vision-language model which can operate in one of the three functionalities: (1) Unimodal encoder is trained with an image-text contrastive (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

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 trans-

former 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.
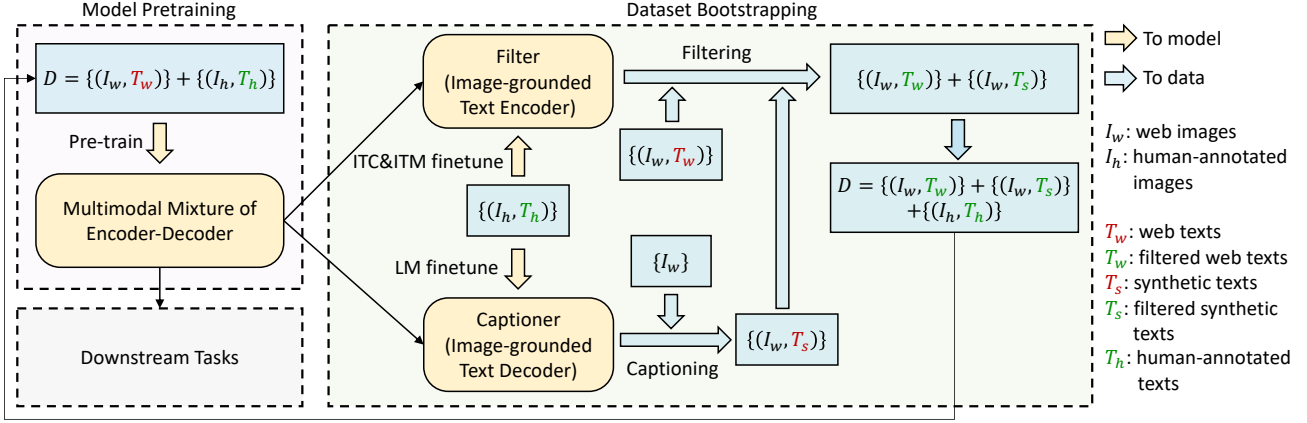
Figure 3. Learning framework of BLIP. We introduce a captioner to produce synthetic captions for web images, and a filter to remove noisy image-text pairs. The captioner and filter are initialized from the same pre-trained model and finetuned individually on a small-scale human-annotated dataset. The 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 BERT$_{base}$ (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.

Pre-train dataset	Bootstrap C	Bootstrap F	Vision backbone	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
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.
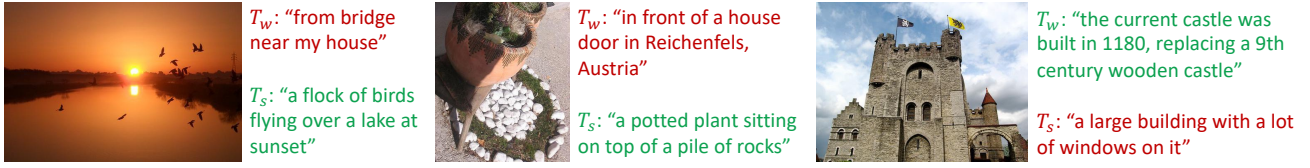
T_w: "from bridge near my house"

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)		Retrieval-ZS (Flickr)		Caption-FT (COCO)		Caption-ZS (NoCaps)	
		TR@1	IR@1	TR@1	IR@1	B@4	CIDEr	CIDEr	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)		Retrieval-ZS (Flickr)		Caption-FT (COCO)		Caption-ZS (NoCaps)	
		TR@1	IR@1	TR@1	IR@1	B@4	CIDEr	CIDEr	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

Captioner & Filter	Noise ratio	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
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)						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
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**
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**
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. 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			IR		
		R@1	R@5	R@10	R@1	R@5	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
BLIP$_{\text{CapFilt-L}}$	129M	**96.0**	**99.9**	**100.0**	**85.5**	**96.8**	**98.7**
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

---

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

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: No-Caps (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

Method	Pre-train #Images	in-domain		near-domain		out-domain		overall		COCO Caption Karpathy test	
		C	S	C	S	C	S	C	S	B@4	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
LEMON$_{base}$† (Hu et al., 2021)	12M	104.5	14.6	100.7	14.0	96.7	12.4	100.4	13.8	-	-
LEMON$_{base}$† (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
BLIP$_{CapFilt-L}$	129M	**111.8**	**14.9**	**108.6**	**14.8**	**111.5**	**14.2**	**109.6**	**14.7**	39.7	**133.3**
LEMON$_{large}$† (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
SimVLM$_{huge}$ (Wang et al., 2021)	1.8B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
BLIP$_{ViT-L}$	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. BLIP$_{CapFilt-L}$ 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. SimVLM$_{huge}$ uses 13× more training data and a larger vision backbone than ViT-L.
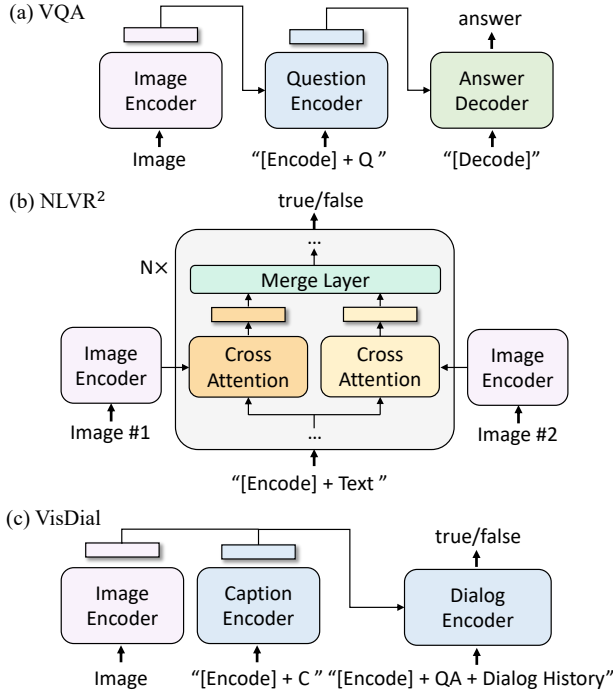
(a) VQA

(b) NLVR^2

(c) VisDial

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;

Method	Pre-train #Images	VQA		NLVR^2	
		test-dev	test-std	dev	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
SimVLM$_{base}$†	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**
BLIP$_{CapFilt-L}$	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.

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.

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-

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

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.

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.

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.

A. Downstream Task Details

Table 14 shows the hyperparameters that we use for fine-tuning 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 No-Caps 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.

T_w: "a week spent at our rented beach house in Sandbridge"

T_s: "an outdoor walkway on a grass covered hill"

T_w: "that's what a sign says over the door"

T_s: "the car is driving past a small old building"

T_w: "hand held through the glass in my front bedroom window"

T_s: "a moon against the night sky with a black background"

T_w: "stunning sky over walney island, lake district, july 2009"

T_s: "an outdoor walkway on a grass covered hill"

T_w: "living in my little white house"

T_s: "a tiny white flower with a bee in it"

T_w: "the pink rock from below"

T_s: "some colorful trees that are on a hill in the mountains"

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.