BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

目录 CONTENTS

1. 研究背景
2. 模型结构
3. 模型训练
4. 实验结果
5. 后续工作

01 研究背景

研究背景

视觉语言预训练模型为各种下游任务提供一个基础模型，并且在零样本迁移上表现出色。

过去很多工作使用大量图片文本对数据集对模型进行**端到端训练**，这种方式随着模型的不断增大，所需要的计算资源也不断变大。而且无法利用视觉领域和文本领域已有的单模态预训练大模型。

本文的方法**冻结预训练的视觉模型和语言模型**，预训练的过程只调节连接视觉模型和语言模型的模块（Q-Former），从而大大减少可训练的参数数量（比Flamingo80B少54倍的参数数量）。

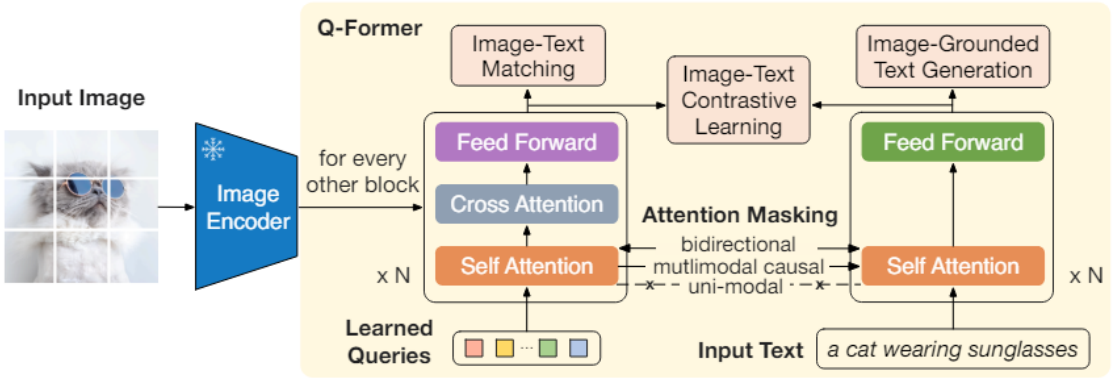
这种方法的核心挑战在于对齐视觉特征和语言特征。相关工作有Frozen（微调视觉模型，将视觉模型的特征作为语言模型的soft prompts）、Flamingo（在LLM中插入cross-attention层，注入视觉特征）。这些工作都使用text conditioned language modeling loss进行预训练。

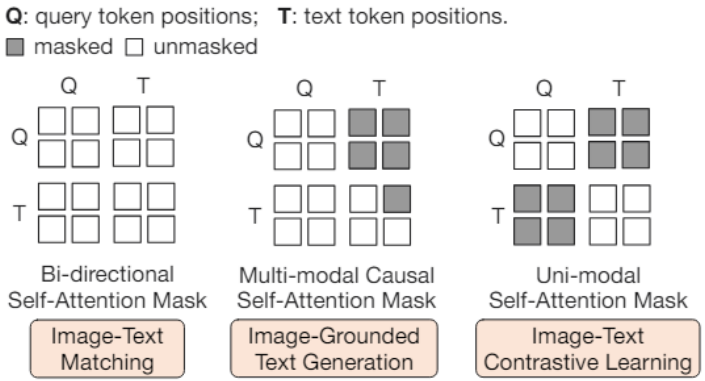
BLIP-2借鉴了BLIP，使用多种loss进行预训练：ITM、ITC、ITG。

02 模型结构

模型结构

BLIP-2的连接模块使用的是Querying Transformer（Q-Former）：





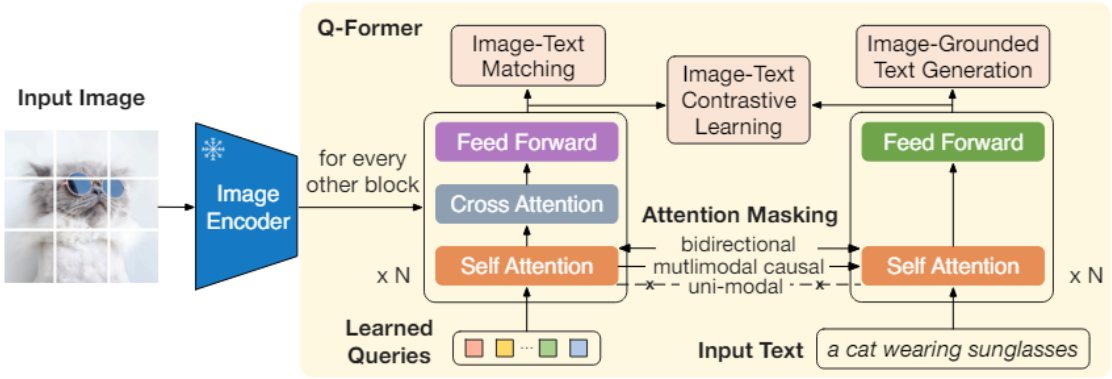
包括一个Image Transformer和Text Transformer。Image Transformer和Text Transformer共享相同的Self Attention层。根据训练目标不同，采用不同的Attention Masking控制Query和Text的交互。

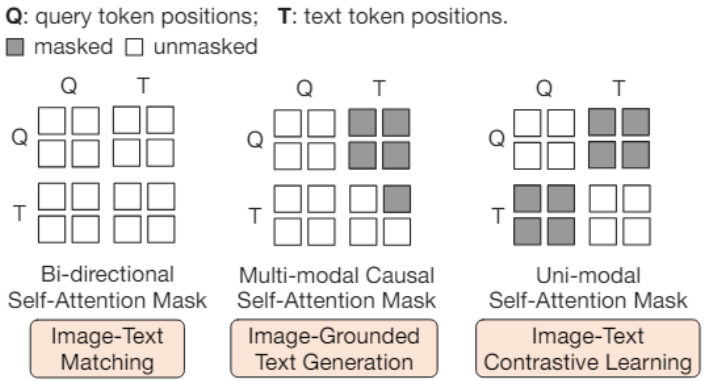
Image Transformer和视觉特征进行交互，提取出和文本最相关的视觉信息。

Text Transformer既可以作为text encoder，也可以作为text decoder。

模型结构

BLIP-2的连接模块使用的是Querying Transformer（Q-Former）：





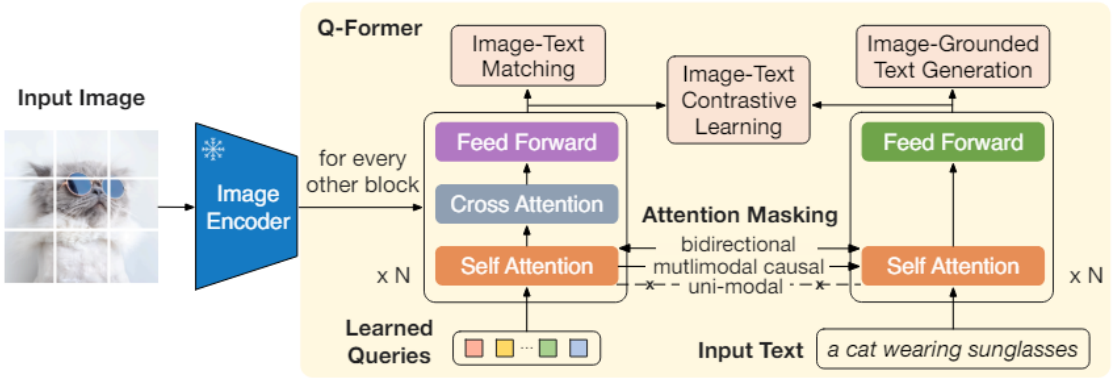
将Queries作为参数，查询出视觉特征中和文本最相关的视觉信息。

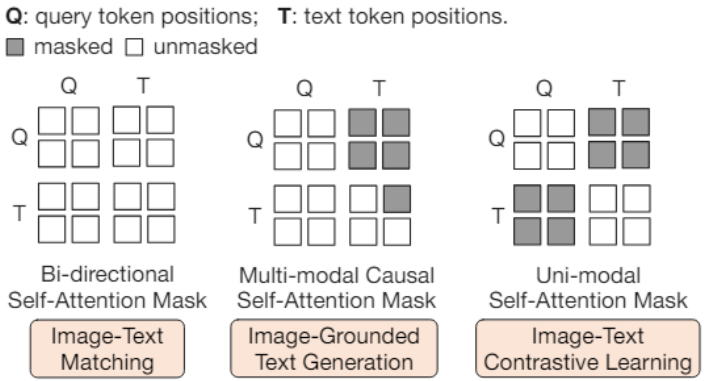
无论图片的分辨率多大，提取出固定数量的视觉信息。

在文章的实验中，使用32 queries，每个query 768维，比图像257 * 1024（ViT-L/14）的特征小很多。

模型结构

BLIP-2的连接模块使用的是Querying Transformer（Q-Former）：





使用$BERT_{base}$对Q-Former进行初始化，cross attention layer随机初始化。

总共188M可训练参数。

03 模型训练

模型训练

使用图像文本对数据集进行预训练，分为两个阶段：

- 阶段一：vision-language representation learning stage with a frozen image encoder
- 阶段二：vision-to-language generative learning stage with a frozen LLM

模型训练

阶段一：vision-language representation learning stage with a frozen image encoder

训练Q-Former，从视觉特征中导出和文本最相关的信息。使用三种预训练目标：ITC、ITG、ITM。

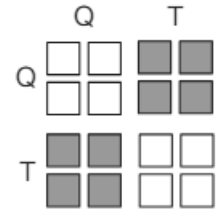
- ITC（Image-Text Contrastive Learning）

使image representation和text representation的表示在向量空间对齐。

配对的图像和文本对相似度尽可能大，不配对的图像和文本对相似度尽可能小。

采用Uni-modal Self-Attention Mask，query和text不能互相看到。

相似度计算：每个query的输出向量和文本的[CLS]输出向量计算相似度，选取最高的作为图像和文本相似度。



模型训练

阶段一：vision-language representation learning stage with a frozen image encoder

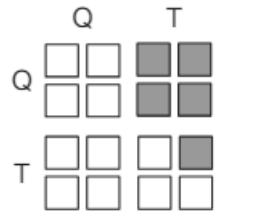
训练Q-Former，从视觉特征中导出和文本最相关的信息。使用三种预训练目标：ITC、ITG、ITM。

- ITG（Image-grounded Text Generation）

训练Q-Former在图片条件下生成文本。

采用Multi-modal Causal Self-Attention Mask。

Text transformer无法直接和视觉编码器交互，生成文本需要的信息必须首先由queries从视觉特征中提取出。因此queries被训练提取出生成文本需要的所有视觉信息。



模型训练

阶段一：vision-language representation learning stage with a frozen image encoder

训练Q-Former，从视觉特征中导出和文本最相关的信息。使用三种预训练目标：ITC、ITG、ITM。

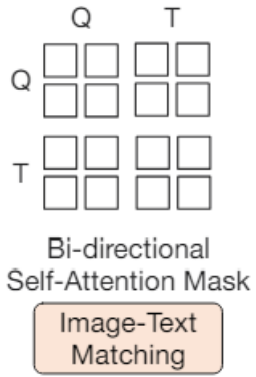
- ITM（Image-Text Matching）

使图片和文本表示细粒度对齐。

二分类任务，模型预测一个图片文本对是否匹配。

采用Bi-directional Self-Attention Mask。

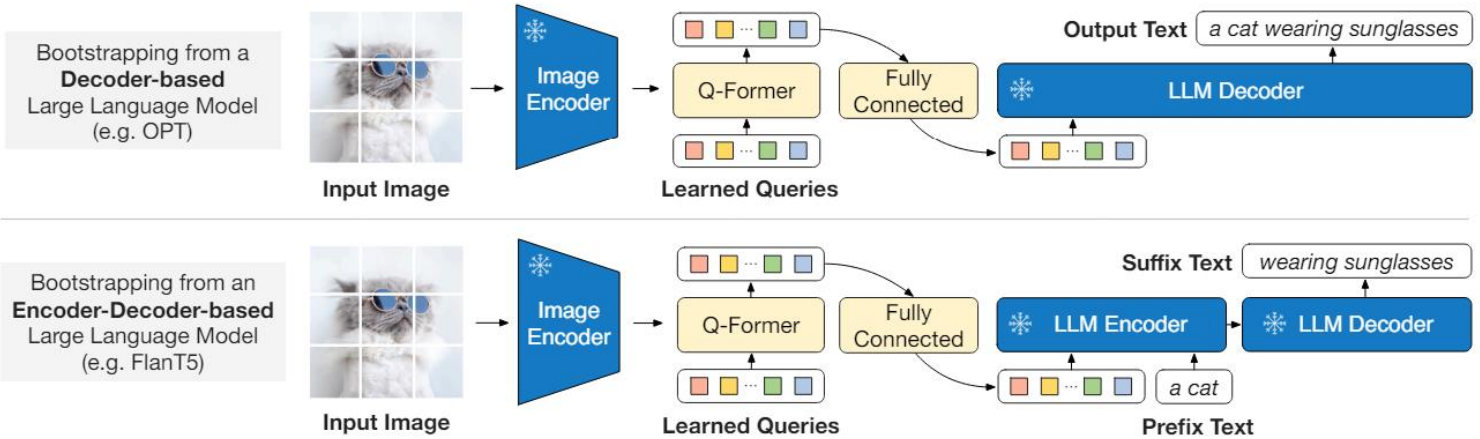
把每个query的输出embedding送入一个二分类的线性分类头得到logit，将所有query的logits相加作为最终的匹配分数。



模型训练

阶段二： vision-to-language generative learning stage with a frozen LLM

训练Q-Former，加入一个FC和冻结的LLM连接，使用language modeling loss。



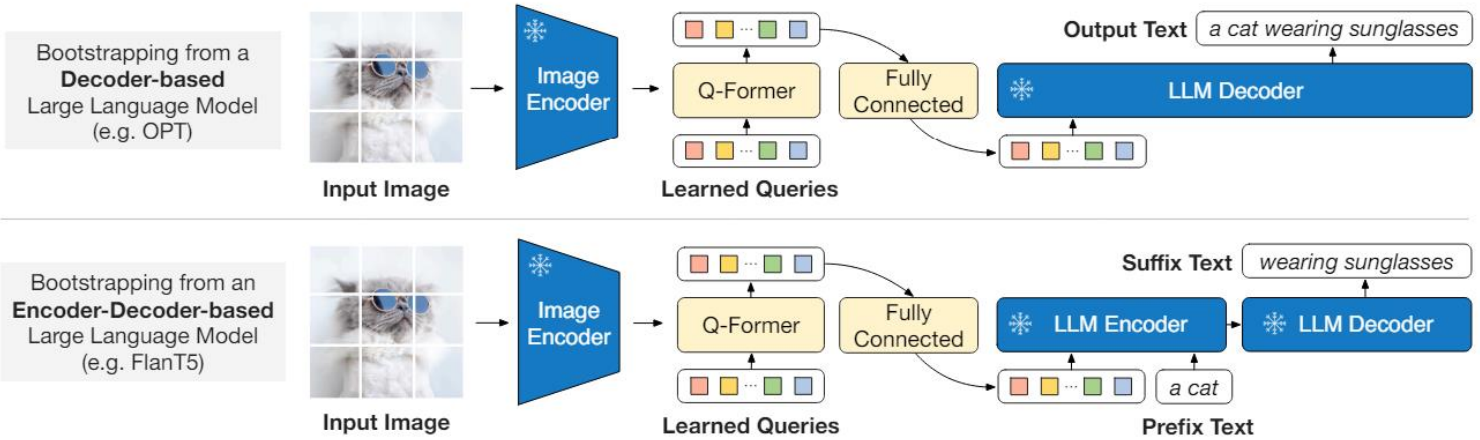
将Q-Former导出的视觉信息使用FC映射到和LLM embedding层相同的dimension，这些信息作为soft visual prompts拼接到输入的text embeddings之前。

经过阶段一的训练，Q-Former学习了抽取最有用的视觉信息，进行了视觉表示和文本表示的对齐，减轻了第二阶段将视觉信息映射到LLM embedding空间的负担。

模型训练

阶段二： vision-to-language generative learning stage with a frozen LLM

训练Q-Former，加入一个FC和冻结的LLM连接，使用language modeling loss。



尝试了两种LLM，decoder-based LLM：OPT和encoder-decoder-based LLM：FlanT5。

对于OPT，根据Q-Former的视觉信息和已生成的文本预测下一个token。对于FlanT5，将文本分为两部分，一部分和视觉信息拼接送给编码器，另一部分作为decoder的生成目标。

04 实验结果

实验结果

University of Chinese Academy of Sciences

Zero-shot VQA

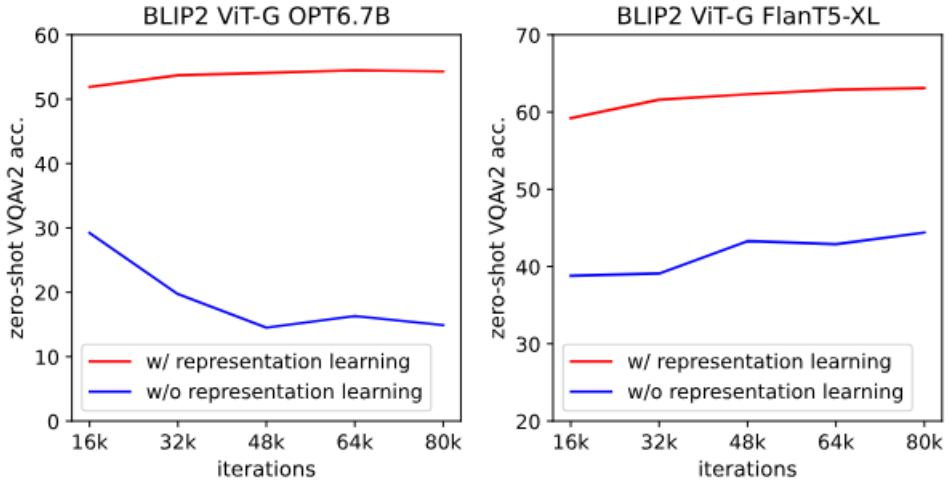
Models	#Trainable Params	#Total Params	VQAv2 val	VQAv2 test-dev	OK-VQA test	GQA test-dev
VL-T5$_{\text{no-vqa}}$	224M	269M	13.5	-	5.8	6.3
FewVLM (Jin et al., 2022)	740M	785M	47.7	-	16.5	29.3
Frozen (Tsimpoukelli et al., 2021)	40M	7.1B	29.6	-	5.9	-
VLKD (Dai et al., 2022)	406M	832M	42.6	44.5	13.3	-
Flamingo3B (Alayrac et al., 2022)	1.4B	3.2B	-	49.2	41.2	-
Flamingo9B (Alayrac et al., 2022)	1.8B	9.3B	-	51.8	44.7	-
Flamingo80B (Alayrac et al., 2022)	10.2B	80B	-	56.3	**50.6**	-
BLIP-2 ViT-L OPT$_{2.7B}$	104M	3.1B	50.1	49.7	30.2	33.9
BLIP-2 ViT-g OPT$_{2.7B}$	107M	3.8B	53.5	52.3	31.7	34.6
BLIP-2 ViT-g OPT$_{6.7B}$	108M	7.8B	54.3	52.6	36.4	36.4
BLIP-2 ViT-L FlanT5$_{XL}$	103M	3.4B	62.6	62.3	39.4	44.4
BLIP-2 ViT-g FlanT5$_{XL}$	107M	4.1B	63.1	63.0	40.7	44.2
BLIP-2 ViT-g FlanT5$_{XXL}$	108M	12.1B	**65.2**	**65.0**	45.9	**44.7**

在VQAv2上，比Flamingo80B高8.7%，尽管比Flamingo80B少54倍的参数。

另外实验发现：更强大的image encoder和更强大的LLM都可以使实验结果变得更好。

实验结果

Vision-Language Representation Learning消融实验



当去掉第一阶段的pretraining时，两种类型的LLM在zero-shot VQA上的性能都大幅下降。

对于OPT，产生灾难性遗忘，随着training的进行，性能下降。

消融实验证明了第一阶段pretraining的重要性。

实验结果

Image Captioning

Models	#Trainable Params	NoCaps Zero-shot (validation set)								COCO Fine-tuned Karpathy test	
		in-domain		near-domain		out-domain		overall			
		C	S	C	S	C	S	C	S	B@4	C
OSCAR (Li et al., 2020)	345M	-	-	-	-	-	-	80.9	11.3	37.4	127.8
VinVL (Zhang et al., 2021)	345M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3
BLIP (Li et al., 2022)	446M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	40.4	136.7
OFA (Wang et al., 2022a)	930M	-	-	-	-	-	-	-	-	**43.9**	145.3
Flamingo (Alayrac et al., 2022)	10.6B	-	-	-	-	-	-	-	-	-	138.1
SimVLM (Wang et al., 2021b)	~1.4B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
BLIP-2 ViT-g $OPT_{2.7B}$	1.1B	123.0	15.8	117.8	15.4	123.4	**15.1**	119.7	15.4	43.7	**145.8**
BLIP-2 ViT-g $OPT_{6.7B}$	1.1B	**123.7**	15.8	119.2	15.3	124.4	14.8	121.0	15.3	43.5	145.2
BLIP-2 ViT-g $FlanT5_{XL}$	1.1B	**123.7**	**16.3**	**120.2**	**15.9**	**124.8**	**15.1**	**121.6**	**15.8**	42.4	144.5

使用prompt：a photo of作为LLM的初始文本输入，在COCO上微调Q-Former和image encoder。当零样本迁移到NoCaps数据集上时，比之前的方法性能上有很大提升。证明了模型对于out-domain images的泛化能力。

实验结果

Visual Question Answering

在VQA数据集上（VQAv2的训练集和验证集、 Visual Genome的训练集）微调Q-Former和image encoder的参数，使用open-ended答案生成损失。

将问题也输入进Q-Former，从而指导Q-Former注意和问题更相关的图像区域。

在open-ended generation models中BLIP-2取得了SOTA的性能。

Models	#Trainable Params	VQAv2 test-dev	VQAv2 test-std
Open-ended generation models			
ALBEF (Li et al., 2021)	314M	75.84	76.04
BLIP (Li et al., 2022)	385M	78.25	78.32
OFA (Wang et al., 2022a)	930M	82.00	82.00
Flamingo80B (Alayrac et al., 2022)	10.6B	82.00	82.10
BLIP-2 ViT-g FlanT5$_{XL}$	1.2B	81.55	81.66
BLIP-2 ViT-g OPT$_{2.7B}$	1.2B	81.59	81.74
BLIP-2 ViT-g OPT$_{6.7B}$	1.2B	**82.19**	**82.30**
Closed-ended classification models			
VinVL	345M	76.52	76.60
SimVLM (Wang et al., 2021b)	~1.4B	80.03	80.34
CoCa (Yu et al., 2022)	2.1B	82.30	82.30
BEIT-3 (Wang et al., 2022b)	1.9B	**84.19**	**84.03**

实验结果

Image-Text Retrieval

Model	#Trainable Params	Flickr30K Zero-shot (1K test set)						COCO Fine-tuned (5K test set)					
		Image → Text			Text → Image			Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Dual-encoder models													
CLIP (Radford et al., 2021)	428M	88.0	98.7	99.4	68.7	90.6	95.2	-	-	-	-	-	-
ALIGN (Jia et al., 2021)	820M	88.6	98.7	99.7	75.7	93.8	96.8	77.0	93.5	96.9	59.9	83.3	89.8
FILIP (Yao et al., 2022)	417M	89.8	99.2	99.8	75.0	93.4	96.3	78.9	94.4	97.4	61.2	84.3	90.6
Florence (Yuan et al., 2021)	893M	90.9	99.1	-	76.7	93.6	-	81.8	95.2	-	63.2	85.7	-
BEIT-3(Wang et al., 2022b)	1.9B	94.9	99.9	**100.0**	81.5	95.6	97.8	84.8	96.5	98.3	67.2	**87.7**	**92.8**
Fusion-encoder models													
UNITER (Chen et al., 2020)	303M	83.6	95.7	97.7	68.7	89.2	93.9	65.7	88.6	93.8	52.9	79.9	88.0
OSCAR (Li et al., 2020)	345M	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
VinVL (Zhang et al., 2021)	345M	-	-	-	-	-	-	75.4	92.9	96.2	58.8	83.5	90.3
Dual encoder + Fusion encoder reranking													
ALBEF (Li et al., 2021)	233M	94.1	99.5	99.7	82.8	96.3	98.1	77.6	94.3	97.2	60.7	84.3	90.5
BLIP (Li et al., 2022)	446M	96.7	**100.0**	**100.0**	86.7	97.3	98.7	82.4	95.4	97.9	65.1	86.3	91.8
BLIP-2 ViT-L	474M	96.9	**100.0**	**100.0**	88.6	97.6	**98.9**	83.5	96.0	98.0	66.3	86.5	91.8
BLIP-2 ViT-g	1.2B	**97.6**	**100.0**	**100.0**	**89.7**	**98.1**	**98.9**	**85.4**	**97.0**	**98.5**	**68.3**	**87.7**	92.6

不涉及文本生成，因此只使用第一阶段预训练的模型。在COCO数据集上使用第一阶段相同的3个训练目标进行微调，测试COCO上的性能和Flickr30K上零样本的性能。

实验结果显示，相比于已有的方法，BLIP-2在零样本检索上性能有很大提升。

05 后续工作

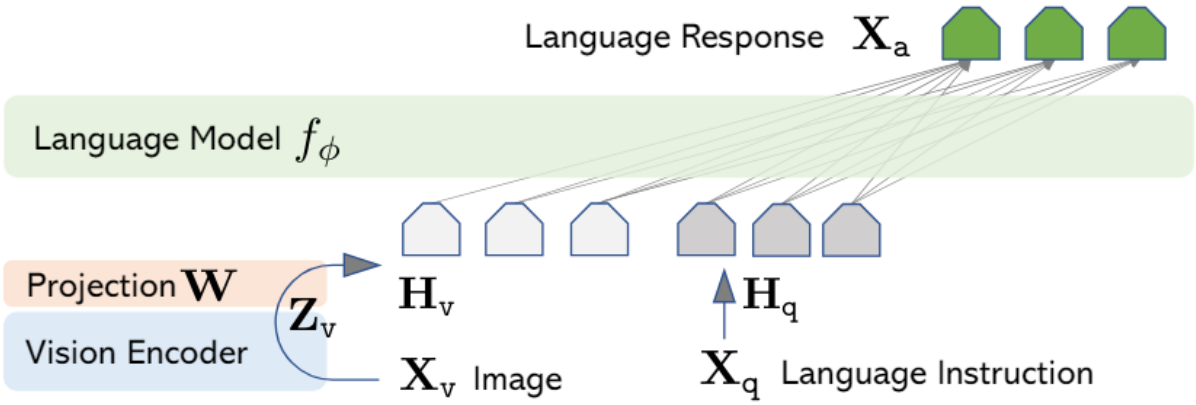
后续工作

Vision-language Instruction Tuning

- LLaVA
- InstructBLIP

后续工作

- LLaVA



冻结视觉模型和语言模型，通过一个简单的线性映射层进行连接视觉模型和语言模型。

视觉模型选用ViT-L/14，语言模型选用Vicuna。

将视觉模型的grid features经过一个线性映射层转换为一系列视觉token。

后续工作

- LLaVA

$$\mathbf{X}_{\text{system-message}} \texttt{<STOP>}$$
$$\text{Human}: \mathbf{X}^1_{\text{instruct}} \texttt{<STOP>}\ \text{Assistant}: \mathbf{X}^1_{\text{a}} \texttt{<STOP>}$$
$$\text{Human}: \mathbf{X}^2_{\text{instruct}} \texttt{<STOP>}\ \text{Assistant}: \mathbf{X}^2_{\text{a}} \texttt{<STOP>} \cdots$$

使用GPT-4，由图片文本对生成了multimodal instruction-following data，包括Conversation、Detailed description、Complex reasoning三类，总共158K个样本。

采用语言模型的损失，只有绿色的部分计算损失。

采用两个阶段的预训练：

- Stage 1: Pre-training for Feature Alignment
- Stage 2: Fine-tuning End-to-End

后续工作

- LLaVA

Stage 1: Pre-training for Feature Alignment

使用595K图片文本对，每个文本对是一个单轮的对话，随机采样一个让模型描述图片的问题，ground-truth是原始的caption。在训练时冻结视觉模型和语言模型，仅训练线性映射层的参数。这个阶段可以理解为为LLM训练一个兼容的visual tokenizer。

Stage 2: Fine-tuning End-to-End

冻结视觉模型的参数，使用构建的158K instruction-following data，对映射层的参数和LLM的参数进行更新。

后续工作

- LLaVA

手工设计了LLaVA-Bench (In-the-Wild)数据集，用来评测模型的Instruction-following能力，使用GPT4进行打分。

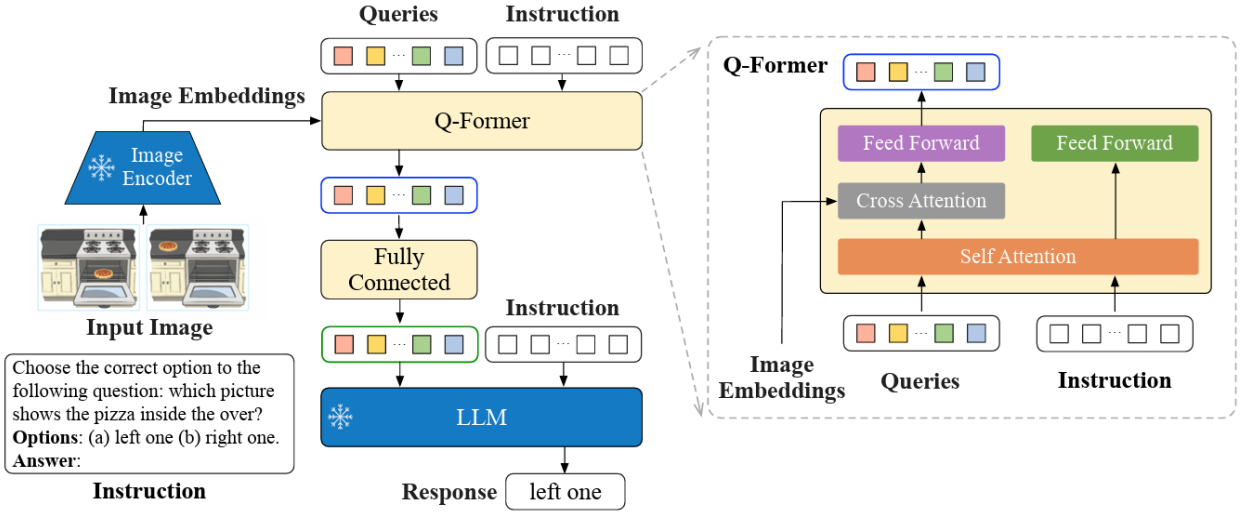
	Conversation	Detail description	Complex reasoning	All
OpenFlamingo [5]	19.3 ± 0.5	19.0 ± 0.5	19.1 ± 0.7	19.1 ± 0.4
BLIP-2 [28]	54.6 ± 1.4	29.1 ± 1.2	32.9 ± 0.7	38.1 ± 1.0
LLaVA	57.3 ± 1.9	52.5 ± 6.3	81.7 ± 1.8	67.3 ± 2.0
LLaVA†	58.8 ± 0.6	49.2 ± 0.8	81.4 ± 0.3	66.7 ± 0.3

Table 5: Instruction-following capability comparison using relative scores on LLaVA-Bench (In-the-Wild). The results are reported in the format of *mean* $\pm$ *std*. For the first three rows, we report three inference runs. LLaVA performs significantly better than others. † For a given set of LLaVA decoding sequences, we evaluate by querying GPT-4 three times; GPT-4 gives a consistent evaluation.

可以看到LLaVA比OpenFlamingo和BLIP-2表现好很多。

后续工作

- InstructBLIP



在BLIP-2的基础上进行指令微调，Q-Former除了接收Queries和Image Embeddings，还接收instruction text tokens，instruction和Query embeddings在self-attention层进行交互，指导Q-Former提取出instruction相关的视觉信息。

后续工作

- InstructBLIP

使用模板的方式由多个数据集构造指令微调数据集，并使用了LLaVA-Instruct-150K数据集。

Task	Instruction Template
Image Captioning	<Image>A short image caption: <Image>A short image description: <Image>A photo of <Image>An image that shows <Image>Write a short description for the image. <Image>Write a description for the photo. <Image>Provide a description of what is presented in the photo. <Image>Briefly describe the content of the image. <Image>Can you briefly explain what you see in the image? <Image>Could you use a few words to describe what you perceive in the photo? <Image>Please provide a short depiction of the picture. <Image>Using language, provide a short account of the image. <Image>Use a few words to illustrate what is happening in the picture.
VQA	<Image>{Question} <Image>Question: {Question} <Image>{Question} A short answer to the question is <Image>Q: {Question} A: <Image>Question: {Question} Short answer: <Image>Given the image, answer the following question with no more than three words. {Question} <Image>Based on the image, respond to this question with a short answer: {Question}. Answer: <Image>Use the provided image to answer the question: {Question} Provide your answer as short as possible: <Image>What is the answer to the following question? "{Question}" <Image>The question "{Question}" can be answered using the image. A short answer is
VQG	<Image>Given the image, generate a question whose answer is: {Answer}. Question: <Image>Based on the image, provide a question with the answer: {Answer}. Question: <Image>Given the visual representation, create a question for which the answer is "{Answer}". <Image>From the image provided, craft a question that leads to the reply: {Answer}. Question: <Image>Considering the picture, come up with a question where the answer is: {Answer}. <Image>Taking the image into account, generate an question that has the answer: {Answer}. Question:

后续工作

- InstructBLIP

	NoCaps	Flickr 30K	GQA	VSR	IconQA	TextVQA	Visdial	HM	VizWiz	SciQA image	MSVD QA	MSRVTT QA	iVQA
Flamingo-3B [4]	-	60.6	-	-	-	30.1	-	53.7	28.9	-	27.5	11.0	32.7
Flamingo-9B [4]	-	61.5	-	-	-	31.8	-	57.0	28.8	-	30.2	13.7	35.2
Flamingo-80B [4]	-	67.2	-	-	-	35.0	-	46.4	31.6	-	35.6	17.4	40.7
BLIP-2 (FlanT5$_{XL}$) [20]	104.5	76.1	44.0	60.5	45.5	43.1	45.7	53.0	29.8	54.9	33.7	16.2	40.4
BLIP-2 (FlanT5$_{XXL}$) [20]	98.4	73.7	44.6	68.2	45.4	44.1	46.9	52.0	29.4	64.5	34.4	17.4	45.8
BLIP-2 (Vicuna-7B)	107.5	74.9	38.6	50.0	39.7	40.1	44.9	50.6	25.3	53.8	18.3	9.2	27.5
BLIP-2 (Vicuna-13B)	103.9	71.6	41.0	50.9	40.6	42.5	45.1	53.7	19.6	61.0	20.3	10.3	23.5
InstructBLIP (FlanT5$_{XL}$)	119.9	**84.5**	48.4	64.8	50.0	46.6	46.6	56.6	32.7	70.4	43.4	25.0	53.1
InstructBLIP (FlanT5$_{XXL}$)	120.0	83.5	47.9	**65.6**	**51.2**	46.6	**48.5**	54.1	30.9	**70.6**	**44.3**	**25.6**	**53.8**
InstructBLIP (Vicuna-7B)	**123.1**	82.4	49.2	54.3	43.1	50.1	45.2	**59.6**	**34.5**	60.5	41.8	22.1	52.2
InstructBLIP (Vicuna-13B)	121.9	82.8	**49.5**	52.1	44.8	**50.7**	45.4	57.5	33.4	63.1	41.2	24.8	51.0

Table 1: Zero-shot results on the held-out datasets. Here, Visdial, HM and SciQA denote the Visual Dialog, HatefulMemes and ScienceQA datasets, respectively. For ScienceQA, we only evaluate on the set with image context. Following previous works [4, 49, 32], we report the CIDEr score [42] for NoCaps and Flickr30K, iVQA accuracy for iVQA, AUC score for HatefulMemes, and Mean Reciprocal Rank (MRR) for Visual Dialog. For all other datasets, we report the top-1 accuracy (%).

可以看到InstructBLIP在各个数据集上都比BLIP-2性能要好，证明了指令微调的有效性。

参考文献

- BLIP：https://arxiv.org/abs/2201.12086
- BLIP-2：https://arxiv.org/abs/2301.12597
- LLaVA：https://arxiv.org/abs/2304.08485
- InstructBLIP：https://arxiv.org/abs/2305.06500

附录

附录

Salesforce/blip2-flan-t5-xxl

a clock tower

Salesforce/blip2-opt-2.7b

a large building with a clock tower

附录

llava-hf/llava-v1.6-mistral-7b-hf

The image shows a tall clock tower with a distinctive design, featuring a series of blue and white circular patterns along its sides. The tower is part of a larger building with a red brick facade. The sky is partly cloudy, suggesting it might be a pleasant day. There are trees in the foreground, and a person on a bicycle is visible on the street below. The architecture and the presence of the clock tower suggest this could be a significant landmark or a central part of a city or town.

附录

Salesforce/instructblip-vicuna-13b

In the image, there is a large brick building with a clock tower on top of it. The building is situated in the middle of a city street, surrounded by trees and other buildings. There is also a car parked on the side of the road near the building.

附录

GPT-4o

The image shows a tall brick tower with a clock on top. There are multiple blue and white circular emblems and what appears to be text in blue running vertically down the tower's façade. The building, likely part of an educational institution, has multiple windows and is situated on a street with a few vehicles and a person riding a bicycle nearby. The sky is partly cloudy, and there are trees framing the image on the right side.

感谢