多模态统一框架之BLIP系列工作

这篇文章整理了Salesforce Research在多模态领域提出的NLIP图文统一框架，利用图文数据训练能够解决各类图文任务的统一模型（图文匹配、看图说话等）。共包含3个工作：BLIP、BLIP-2、InstructBLIP。三者的核心店如下：

BLIP：BLIP初步建立了一套多专家网络，用3种不同的文本模型支持多种类型的图文任务。

BLIP-2：BLIP-2提出了Q-Transformer，用来适配预训练图像模型和预训练语言模型，可以在两种模态模型参数不变的情况下实现多模态对齐，解决各类多模态任务。

InstructBLIP：基于BLIP框架，探索使用Instruct Tuning的思路实现多模态中的zero/few-shot learning。

1. BLIP

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

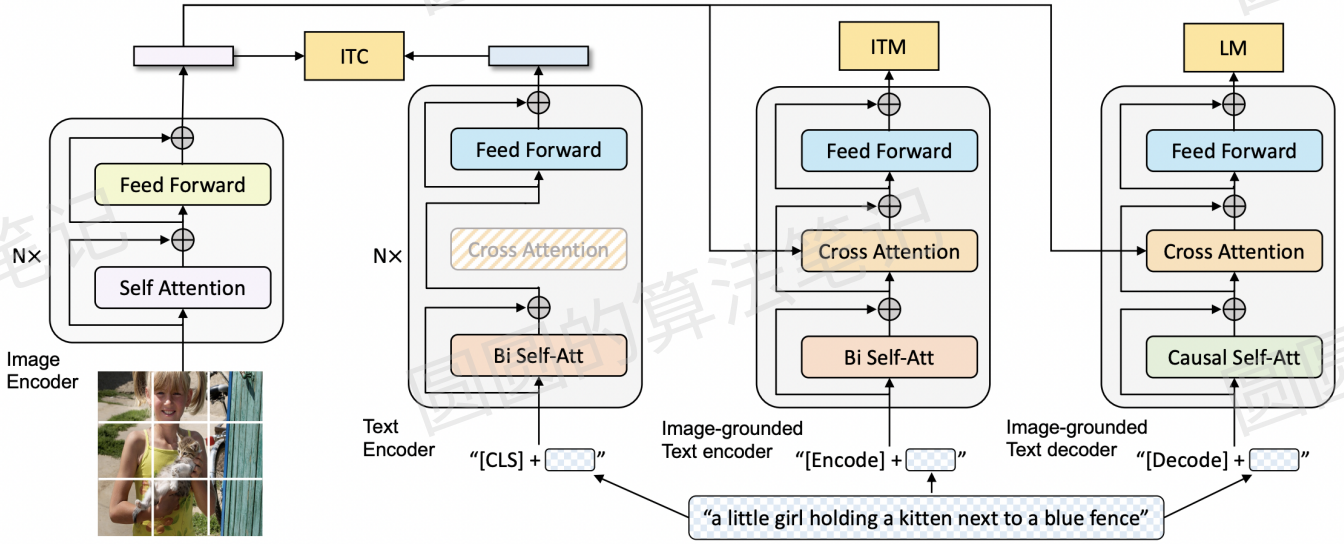
论文标题：BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

下载地址：https://arxiv.org/pdf/2201.12086.pdf

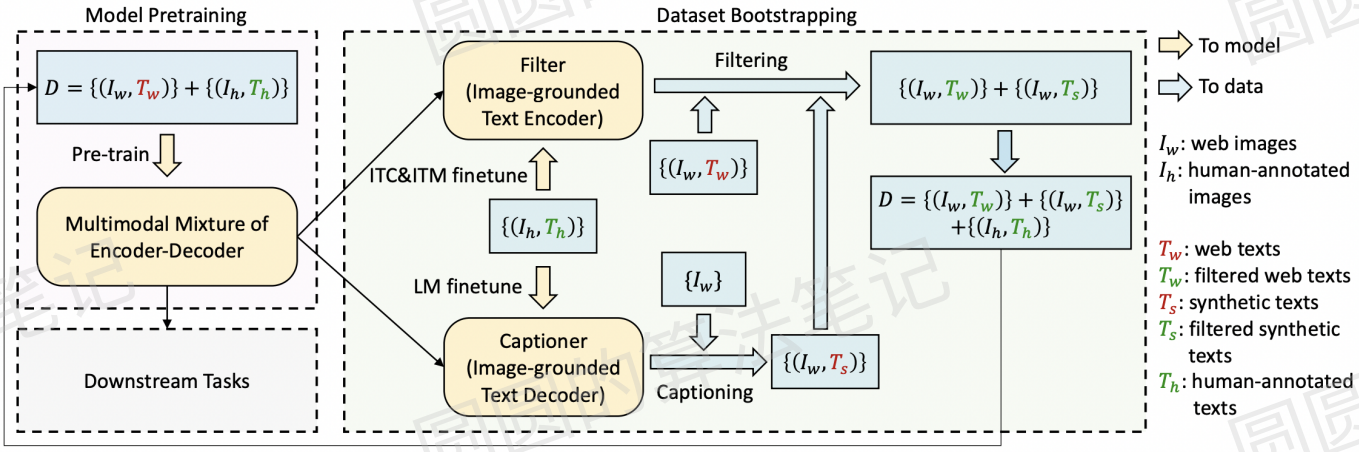
开源代码：https://github.com/salesforce/BLIP

本文提出了一种类似于多专家的多模态模型，以此来兼容多种类型的任务，在多种类型数据上实现训练。同时，在数据维度上，提出了一种CapFilt的方法，生成高质量的图文数据，缓解CLIP等模型中使用Web抓取的数据中噪音较大的问题。

BLIP的整体模型结构如下图，包括1个图像单模态Encoder、1个文本单模态Encoder、1个文本多模态Encoder、1个文本多模态Decoder。图像Encoder采用的是ViT，文本Encoder采用的是BERT。其中多模态的文本Encoder增加了一层cross attention，以图像侧信息作为额外输入，建模图文之间的细粒度关系。文本Decoder采用的是单向attention，主要目的是用于根据图像生成文本的任务。



BLIP的训练任务主要包括ITC、ITM、LM三种。其中ITC是CLIP中的图文对比学习训练方式，拉近相同含义的图像和文本的整体表示。ITM是图文匹配任务，它与图文对比学习的主要区别是，引入了图文之间的cross attention，进行细粒度的图像和文本匹配用来预测，可以理解为单塔模型和双塔模型的区别。LM任务是根据图像生成文本任务，主要用来让Decoder具备文本生成能力。



在数据层面上，文中提出了CapFilt方法，用来生成高质量的图文数据。原来的CLIP使用的图文数据都是Web上自动挖掘的，包含大量噪声。而本文的数据构造方法是，首先从Web上挖掘大量图像，然后用一个看图说话的模型根据图像生成文本。最后再使用图文匹配模型，对图像和生成的文本进行打分，过滤掉打分较低的潜在噪声。

2. BLIP-2

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

Junnan Li Dongxu Li Silvio Savarese Steven Hoi
Salesforce Research
https://github.com/salesforce/LAVIS/tree/main/projects/blip2

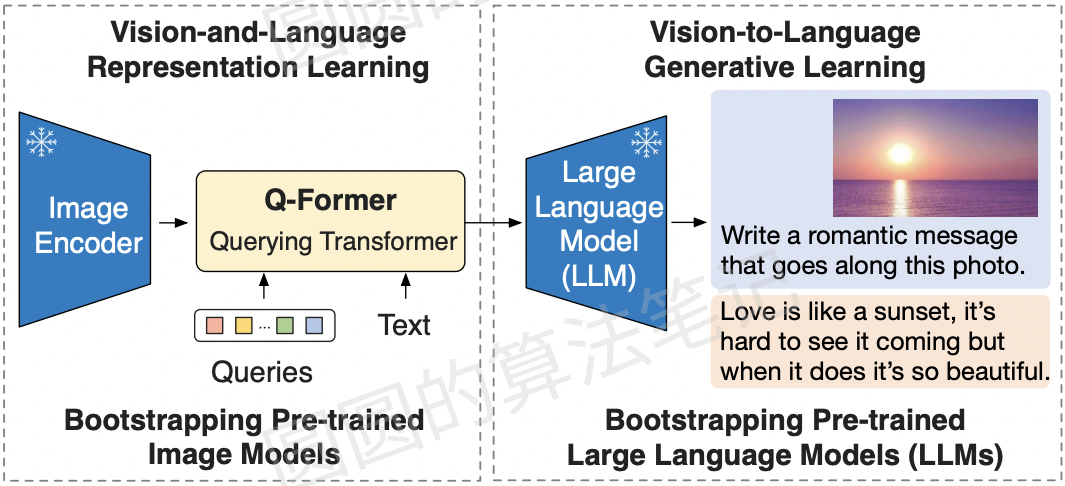
论文标题：BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

下载地址：https://arxiv.org/pdf/2301.12597.pdf

开源代码：https://github.com/salesforce/LAVIS/tree/main/projects/blip2

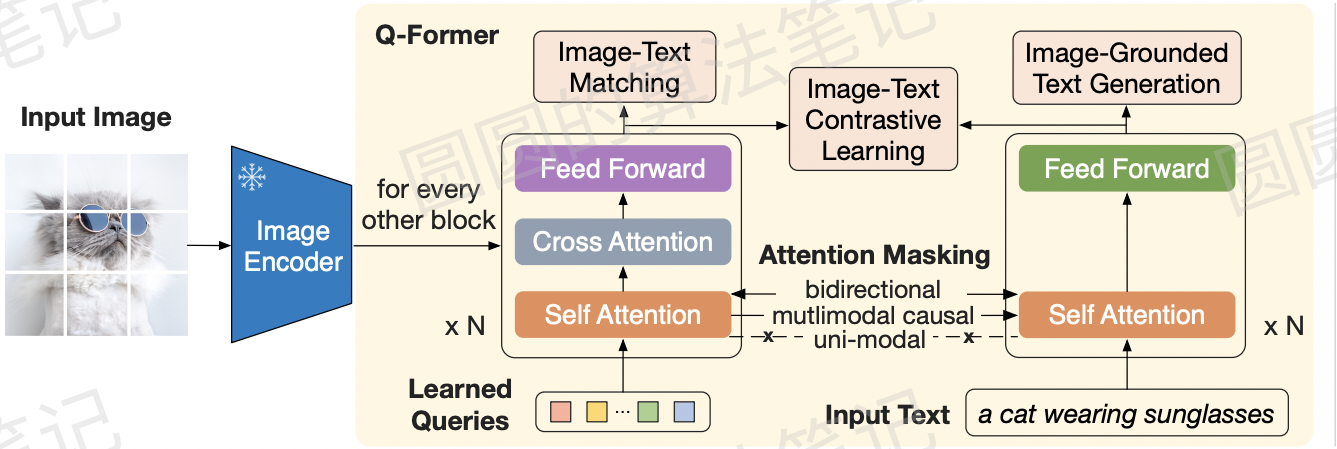
BLIP-2的目标和BLIP相同，也是要打造一个适用于多种任务的统一多模态模型，但是具体的实现方式不同。BLIP-2的核心是如何利用预训练好的图像模型和文本模型。图像模型和文本模型在各自的领域进行了预训练，其单模态的能力非常强，但是由于两个模态之间的空间无法对齐，无法直接在多模态使用。

本文提出的解决思路是，构造一个中间网络，作为预训练图像模型和与训练语言模型得到信息转换媒介，在整个训练过程中，只更新这个中间网络，让预训练对的单模态图文模型参数冻结。这种方式既能直接应用单模态的高质量模型，又能实现更轻量级的模型finetune。

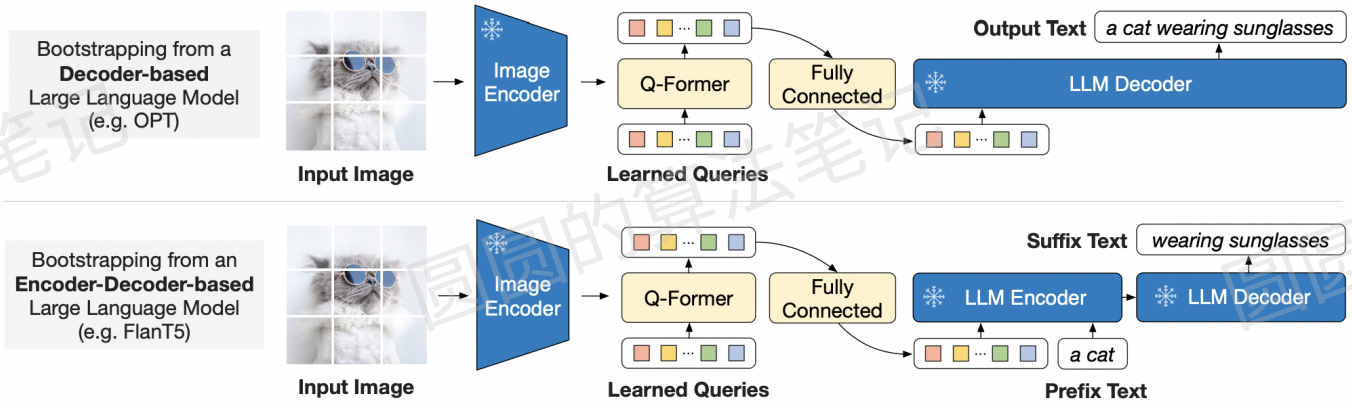


这个中间网络文中称为Q-Former，其训练主要分为两个阶段。第一阶段让Q-Former适配预训练的图像Encoder，第二阶段让Q-Former适配预训练的语言模型，借助语言模型实现更强的文本生成能力。

第一阶段的整体结构图如下，Q-Former包括两个Transformer。左侧Transformer的输入是Queries和预训练图像Encoder生成的图像编码。其中Queries是一些随机初始化的向量，目的是用来和图像进行corss attention，生成相应的转换后的表征。右侧Transformer输入文本。在得到经过Q-Former转换后的图像和文本表征后，使用BLIP中的三类任务进行Q-Former训练。



第二阶段进一步引入预训练的语言模型。将图像经过图像Encoder和Q-Former生成的向量，作为预训练语言模型的前缀，类似于prefix soft prompt（之前的文章中进行过prefix soft prompt的思路，简单来说是在句子前面加一个向量前缀，影响语言模型的后续生成）。预训练语言模型根据soft prompt生成文本。这个过程使用看图说话任务进行优化，让Q-Former能够更好地生成适用于文本生成的图像表征。



整个BLIP-2借助了单模态模型的强大能力，finetune一个中间媒介实现了高效的多模态统一模型训练。

3. InstructBLIP

InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning

Wenliang Dai[†1,2*] **Junnan Li**[†,✉,1] **Dongxu Li**[1] **Anthony Meng Huat Tiong**[1,3]
Junqi Zhao[3] **Weisheng Wang**[3] **Boyang Li**[3] **Pascale Fung**[2] **Steven Hoi**[✉,1]
[1]Salesforce Research [2]Hong Kong University of Science and Technology
[3]Nanyang Technological University, Singapore
https://github.com/salesforce/LAVIS/tree/main/projects/instructblip
[†]Equal contribution [✉]Corresponding authors: {junnan.li,shoi@salesforce.com}

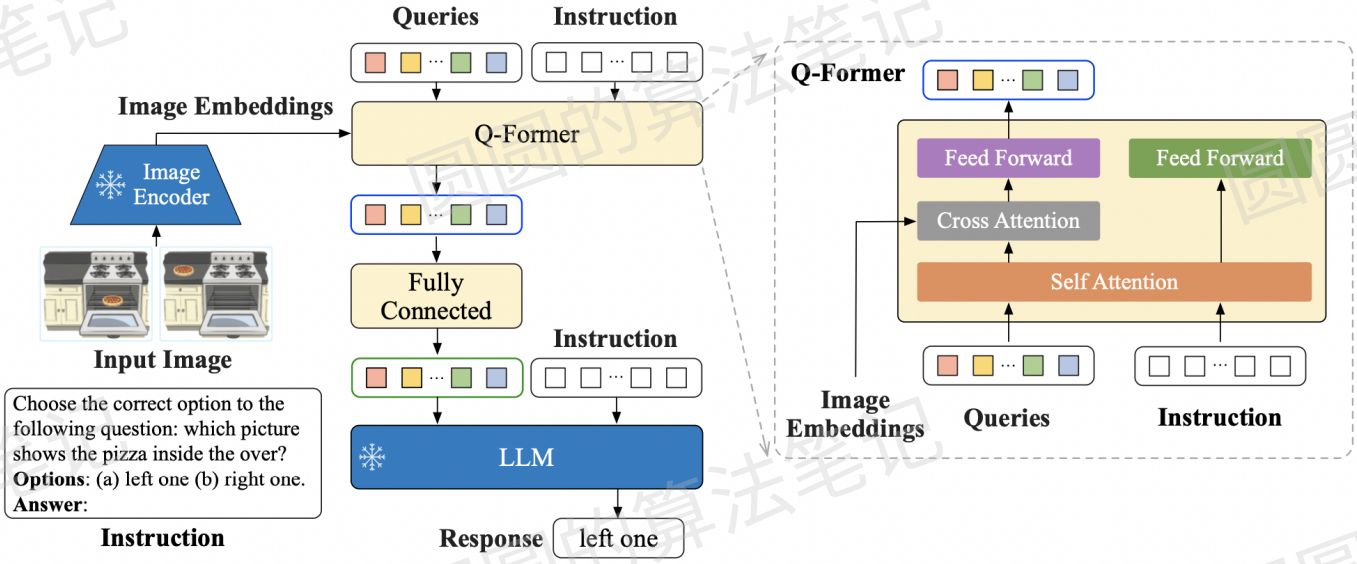
论文标题：InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning

下载地址：https://arxiv.org/pdf/2305.06500.pdf

开源代码：https://github.com/salesforce/LAVIS/tree/main/projects/instructblip

Instruct Tuning在大模型领域取得了非常不错的效果，这种方式在多模态领域也逐渐得到应用，例如BLIP-2就将图像Encoder和Q-Former生成的图像表征作为预训练语言模型的prefix soft prompt。这篇文章进一步探索了多模态领域的Instruct Tuning，相比文本，图像有着更丰富的样式种类，如何实现一种各类图文数据都可用的Instruct Tuning方法非常具有挑战性。

本文既有BLIP-2的框架，提出了InstructBLIP方法。整体结构如下，使用BLIP-2中的Image Encoder和Q-Former生成图像表征，然后做为prefix soft prompt拼接到Instruction前面，整体输入到预训练语言模型中，让预训练语言模型生成预测结果。此外，Instruction也会作为Q-Former的输入，和Queries进行交互，指导从图像中提取相的特征作为prompt。



从多个数据集上的实验结果可以看出，InstructBLIP取得了最新的SOTA效果，比原来的BLIP-2效果提升一大截，验证了InstructBLIP对的优势。

	NoCaps	Flickr 30K	GQA	VSR	IconQA	TextVQA	Visdial	HM	VizWiz	SciQA IMG	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	41.9	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	42.4	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	41.3	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	32.3	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