

前言

大规模语言模型 (Large Language Model, LLM) 是当前的当红炸子鸡，展现出了强大的逻辑推理，语义理解能力，而视觉作为人类最为主要的感知世界的手段，亟待和LLM进行融合，形成多模态大规模语言模型 (Multimodal LLM, MLLM)，BLIP-2这篇文章利用已经充分训练好的图片编码器和LLM模型，通过Q-Former巧妙地融合在一起，在引入少量待学习参数的同时，取得了显著的效果。本文将对BLIP2进行笔记和笔者个人感想纪录，希望对诸位读者有所帮助。如有谬误请见谅并联系指出，本文遵守CC 4.0 BY-SA版权协议，转载请联系作者并注明出处，谢谢。

▽ 联系方式：

e-mail: FesianXu@gmail.com

github: <https://github.com/FesianXu>

知乎专栏: 计算机视觉/计算机图形理论与应用(https://www.zhihu.com/column/c_1265262560611299328)

微信公众号：机器学习杂货铺3号店

笔者最近忙于工作，已经很久没空更新博客，刚好最近在回顾一些论文，顺便将其进行笔记。BLIP2的目的是希望将现有可用的（预训练好的）视觉编码器和LLM中进行融合得到MLLM，而如何将视觉语义向量和LLM进行融合是一件极具有挑战性的工作。LLM是以文本语义为目标进行训练的，而视觉编码器是以视觉语义为目的进行训练的，视觉语义即便经过了语义对齐，如通过CLIP等方式进行跨模态语义对齐，其语义和LLM之间也会存在较大的区别，如何融合这两种语义信息，是MLLM模型必须解决的问题，而BLIP2 [1]就提出了采用Q-Former的方法进行解决。

不过在深入介绍BLIP2的内容之前，我们不妨先卖个关子，先给自己10分钟思考，如果让我们设计一个简单的融合视觉语义和LLM语义的方法，我们会怎么做呢？笔者能想到的方法，会类似于LLaVA [2]，通过Image/Video Caption模型对图片/视频进行文字描述（场景、事件、实体等等），然后利用LLM，在合适的prompt下对这段文字描述进行总结后，输入到LLM中作为输入，从而间接地引入了视觉语义信息到LLM中，流程可见Fig 1所示。这种通过视觉Captioner间接地将视觉语义转换成文本语义（**语义转换阶段**），然后通过Prompt+LLM的方式更好地适配文本语义（**语义适配阶段**），最后将其作为目标LLM的输入从而构成MLLM的方法（**语义融合阶段**）。这种思路直接且容易操作，但是其缺陷也很明显，这里的每个阶段都存在信息损失，最终的MLLM模型对原始视觉的细粒度信息无法感知，这显然严重限制了MLLM的上限。

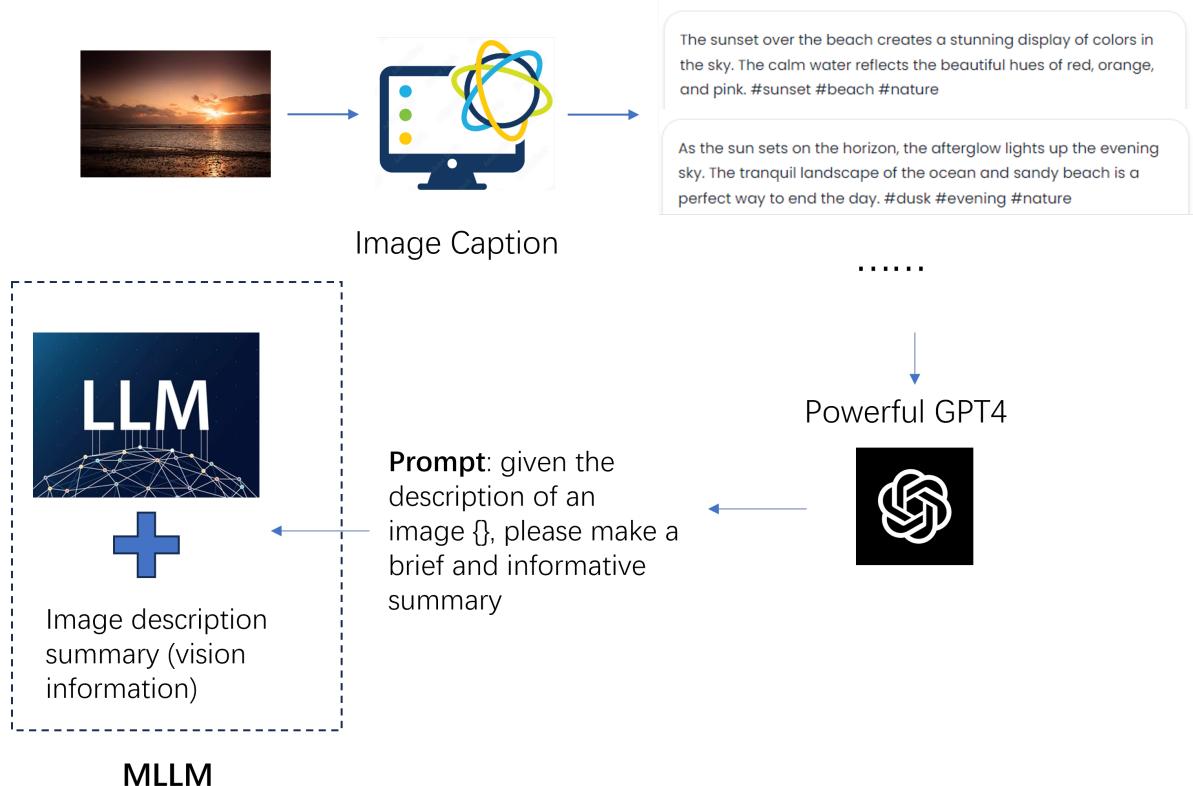


Fig 1. 一种朴素的，通过Captioner间接地将视觉语义转换成文本语义，通过prompt+LLM的方式进行合适的语义适配后，作为目标LLM的输入，从而构成了MLLM。

不过这种思路是可以进行扩展的，其实在以上的三个阶段中，信息损失最为严重的就是语义转换阶段和语义适配阶段，如果我们对视觉语义的转换不是以文字描述的形式，而是语义向量的形式，会不会信息损失控制得最少呢？同时，对于语义适配的过程，我们不采用“硬prompt”，而是可学习的“软prompt”是不是也能进一步提升效果呢？其实这也就是BLIP2中的Q-Former的主要思路了，Q-Former主体如Fig 2. 所示，图片通过预训练好的图片编码器进行特征提取后得到视觉语义 $f_v \in \mathbb{R}^{M \times D}$ 。我们给定 K 个可学习的“软prompt”，在此处称之为“Learnable Queries”，符号表示为 $V_Q \in \mathbb{R}^{K \times D}$ ，这些prompt的作用类似于prompt tuning [3] 中进行下游任务迁移的作用，是为了更好的进行视觉语义到文本语义的迁移，从而产出的Transferred vision representation 可表示为 $E_V \in \mathbb{R}^{K \times D}$ ，则可以作为输入，喂给后续的LLM。为了让 E_V 包含有充分的视觉语义，Q-Former采用了交叉注意力机制（Cross Attention, xattn）融合图片语义和可学习Q，公式如(1)所示。

$$\begin{aligned}
 xattn(Q, K, V) &= \text{softmax}\left(\frac{QK^T + mask}{\sqrt{d_k}}\right)V \\
 K &= V = f_v \in \mathbb{R}^{M \times D} \\
 Q &= V_Q \in \mathbb{R}^{K \times D} \\
 mask &\in \mathbb{R}^{K \times M}
 \end{aligned} \tag{1}$$

总结来看，Q-Former和Fig 1.提到的朴素方法其实可以类比，Learnable Queries可以类比为对文字描述进行总结的prompt词，而产出的Transferred vision representation可以类比经过LLM总结后的文字描述。

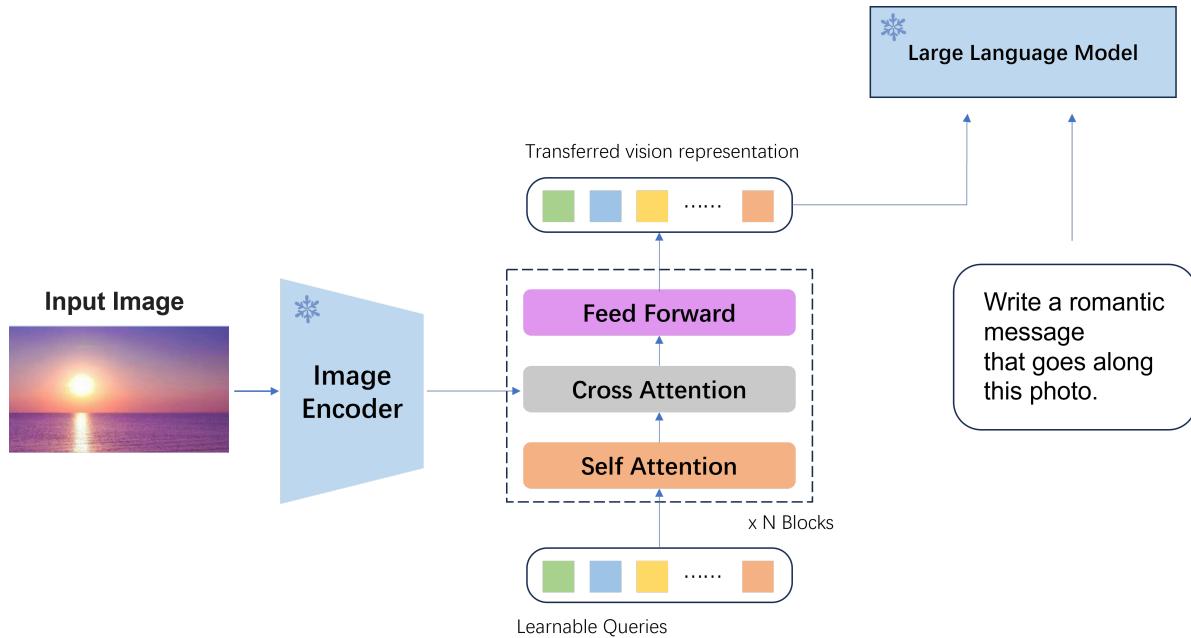


Fig 2. BLIP2中的Q-Former的基本组成。

Q-Former进行视觉语义表征迁移的方式，其实是受启发自Flamingo [4] 中的Perceiver Resampler的设计，此处毕竟不是Flamingo的主场因此不打算对其进行展开介绍，但是笔者觉得有必要对Perceiver Resampler的设计进行简述，会加深读者对Q-Former的理解，如Fig 3所示，由于Flamingo是对图片、视频进行处理的，因此Perceiver Resampler需要将可能变长的视频帧信息转化为固定大小长度的特征，否则过长的视频帧会大大加大后续LLM的计算负担。Perceiver Resampler考虑采用learnable latent queries作为交叉注意力中的Q，而将视频帧/图片帧进行特征提取后展开表示为 X_f ，和Q拼接起来作为交叉注意力中的K和V，通过这种方法将learnable latent queries对应位置的Transformer输出作为视觉特征聚合表示，这样变长的视频帧特征就规整为了固定大小的特征，方便了后续的处理。

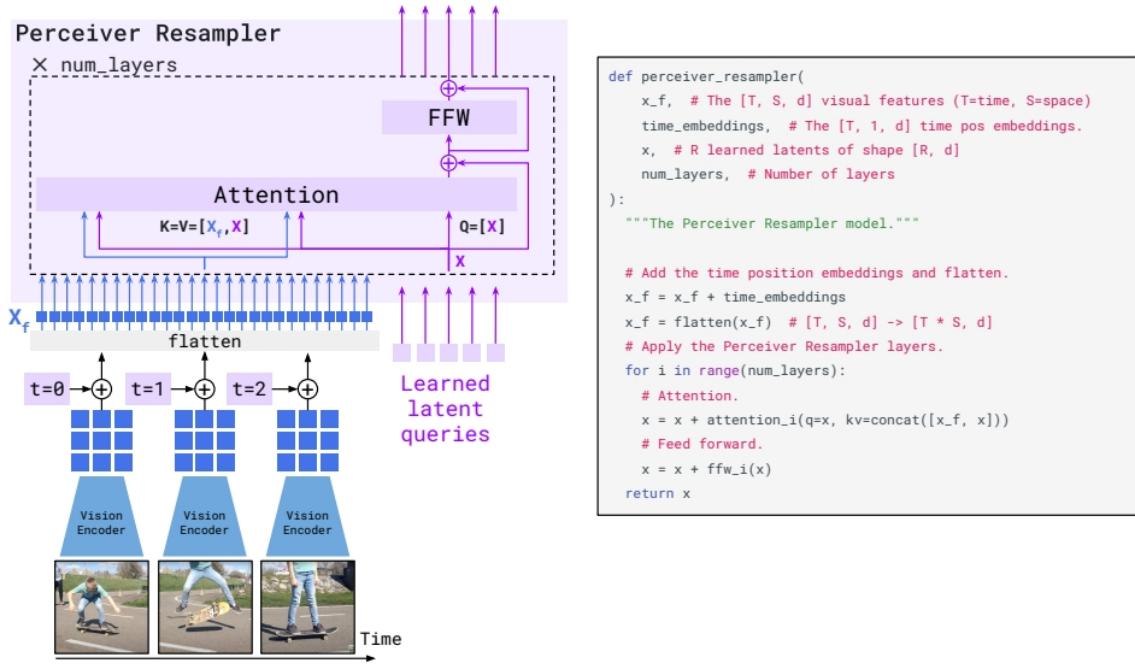


Fig 3. Flamingo中Perceiver Resampler的设计。

到此为止我们算是对Q-Former的设计初衷和设计进行了介绍，接下来才开始对BLIP2这篇工作进行整体介绍，我们要如何训练Q-Former呢？BLIP2的训练分为两个阶段，如下所示。

1. 第一阶段，训练Q-Former实现跨模态表征的语义融合
2. 第二阶段，将训练好的Q-Former和LLM进行结合，实现MLLM

在第一阶段中，为了对Q-Former进行充分训练，作者精妙地设计了三种类型的损失，分别是对比损失 (Image-Text Contrastive Learning, ITC) ，匹配损失(Image-Text Matching, ITM)和生成损失(Image-Grounded Text Generation, ITG)，如Fig 4所示，Q-Former的训练数据是图片——文本对，其中视觉输入Q和文本输入T共用一个自注意层，即是将Q和T拼接起来后输入到自注意层中，通过不同的mask去选择QT之间是否需要注意力交互，mask的具体生效机制见 [5]。让我们回到Q-Former，我们将文本输入[CLS]对应Transformer的输出记为 $E_T \in \mathbb{R}^{N \times D}$ 其中 N 为batch size， $E_V \in \mathbb{R}^{N \times K \times D}$ ，那么以下是各种损失的组成：

1、对比损失 ITC：

```
sim_matrix = matmul(E_V, E_T, transposed_Y=True) # output (N, K, N)
sim_matrix = reduce_max(sim_matrix, axis=1) # output (N, N), maxpool for a best
text-vision matching
dig_label = tensor(list(range(0, batch_size))).reshape(batch_size, 1)
itc_loss = cross_entropy_with_softmax(sim_matrix, dig_label)
```

2、匹配损失 ITM：

```
pos_score = model(Q, T) # positive sample score, output (N, 1)
neg_score = model(Q, T_neg) # negative sample score, output (N, 1), T_neg could
be the hard negative sampled at ITC stage
itm_loss = mse(pos_score, 1) + mse(neg_score, 0)
```

其中的 T_{neg} 为负样本文本，可以参考ALBEF [6] 的工作进行难负样本采样，在此不赘述。

3、生成损失 ITG：

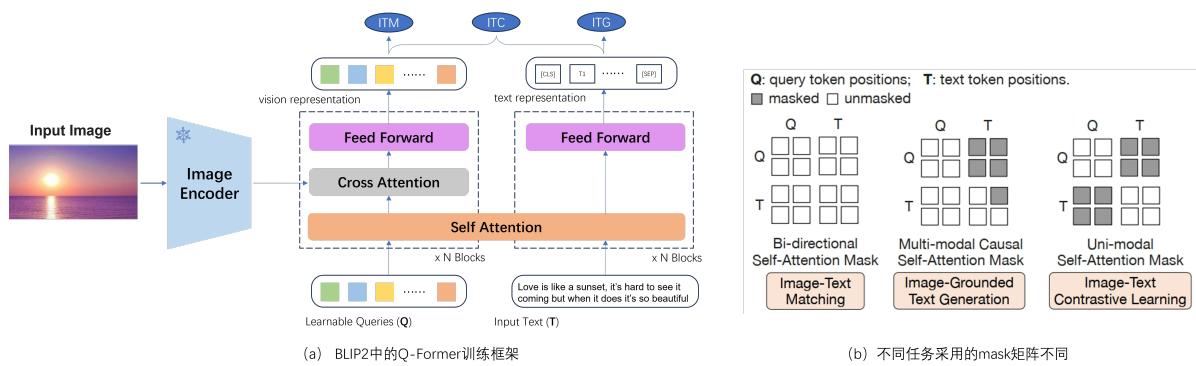


Fig 4. Q-Former的训练过程有ITC、ITM和ITG三种损失构成。

Reference

- [1]. Li, Junnan, Dongxu Li, Silvio Savarese, and Steven Hoi. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." *arXiv preprint arXiv:2301.12597* (2023). aka BLIP2
- [2]. Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. "Visual instruction tuning." *arXiv preprint arXiv:2304.08485* (2023). aka LLaVA
- [3]. Lester, Brian, Rami Al-Rfou, and Noah Constant. "The power of scale for parameter-efficient prompt tuning." *arXiv preprint arXiv:2104.08691* (2021). aka Prompt Tuning.
- [4]. Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc et al. "Flamingo: a visual language model for few-shot learning." *Advances in Neural Information Processing Systems* 35 (2022): 23716-23736. aka Flamingo
- [5]. <https://blog.csdn.net/LoseInVain/article/details/116137177>, 《Transformer代码随记》
- [6]. <https://blog.csdn.net/LoseInVain/article/details/122735603>, 《图文多模态语义融合前的语义对齐——一种单双混合塔多模态模型》
- [7].