自我介绍

介绍语

Badcase

知识库

文本分解 bge算法 文本分块方法

向量召回 文本召回

召回

排序

对称召回和非对称召回

重排 bge-rerank

召回数据-输入大模型规则

召回相似度不高 如何解决

召回相似度很高 如何解决

信息检索 判别指标

Pg\_embedding pg-vector

项目中遇到什么困难 如何优化改进项目

召回阈值

项目准确率

项目是否上限

Rag-fusion rrf

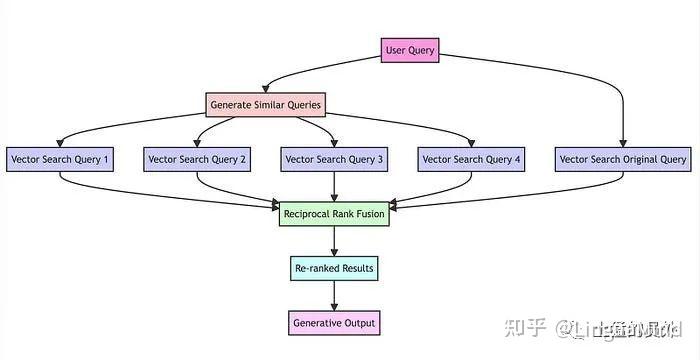
RAG-Fusion’s 工作步骤:

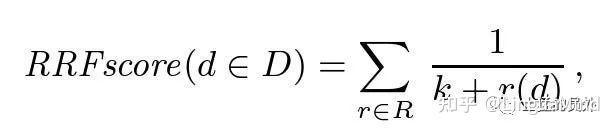
查询语句的相关性复制：通过LLM将用户的查询转换为相似但不同的查询。

并发的向量搜索：对原始查询及其新生成的同级查询执行并发的向量搜索。

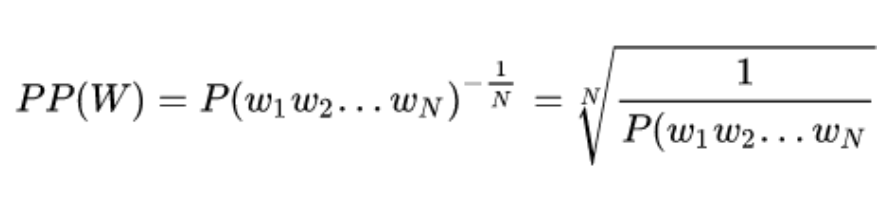
智能重新排名：聚合和细化所有结果使用倒数排序融合(RRF)。

最后优中选优：将精心挑选的结果与新查询配对，引导LLM进行有针对性的查询语句输出，考虑所有查询和重新排序的结果列表。



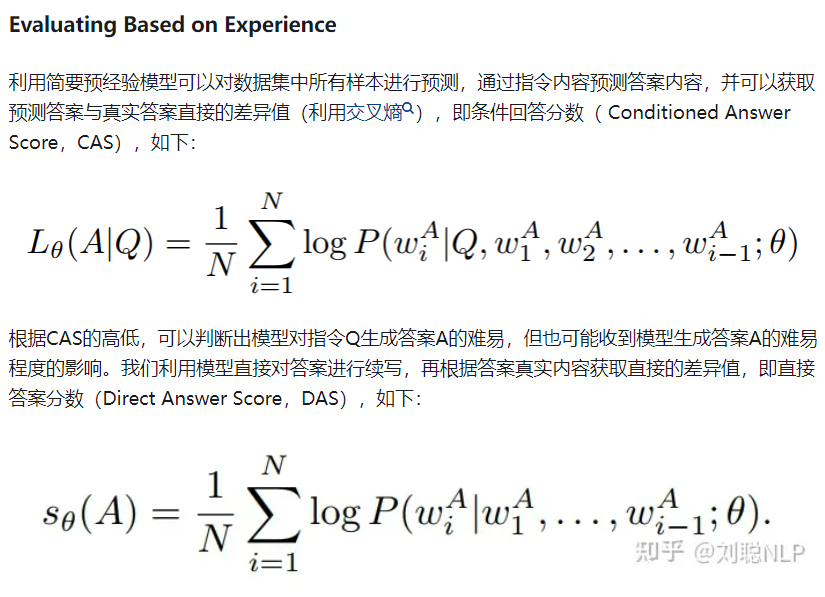


困惑度



主动学习

# 大模型如何在指令微调过程中构造或筛选高质量数据？



NLP

1、Sentencepiece bpe 算法：

最后为了平衡以上两种方法， 又提出了基于 subword 进行分词：它可以较好的平衡词表大小与语义表达能力；常见的子词算法有Byte-Pair Encoding (BPE) / Byte-level BPE（BBPE）、Unigram LM、WordPiece、SentencePiece等。

BPE：即字节对编码。其核心思想是从字母开始，不断找词频最高、且连续的两个token合并，直到达到目标词数。

* BBPE：BBPE核心思想将BPE的从字符级别扩展到子节（Byte）级别。BPE的一个问题是如果遇到了unicode编码，基本字符集可能会很大。BBPE就是以一个字节为一种“字符”，不管实际字符集用了几个字节来表示一个字符。这样的话，基础字符集的大小就锁定在了256（2^8）。采用BBPE的好处是可以跨语言共用词表，显著压缩词表的大小。而坏处就是，对于类似中文这样的语言，一段文字的序列长度会显著增长。因此，BBPE based模型可能比BPE based模型表现的更好。然而，BBPE sequence比起BPE来说略长，这也导致了更长的训练/推理时间。BBPE其实与BPE在实现上并无大的不同，只不过基础词表使用256的字节集。
* WordPiece：WordPiece算法可以看作是BPE的变种。不同的是，WordPiece基于概率生成新的subword而不是下一最高频字节对。WordPiece算法也是每次从词表中选出两个子词合并成新的子词。BPE选择频数最高的相邻子词合并，而WordPiece选择使得语言模型概率最大的相邻子词加入词表。
* Unigram：它和 BPE 以及 WordPiece 从表面上看一个大的不同是，前两者都是初始化一个小词表，然后一个个增加到限定的词汇量，而 Unigram Language Model 却是先初始一个大词表，接着通过语言模型评估不断减少词表，直到限定词汇量。
* SentencePiece：SentencePiece它是谷歌推出的子词开源工具包，它是把一个句子看作一个整体，再拆成片段，而没有保留天然的词语的概念。一般地，它把空格也当作一种特殊字符来处理，再用BPE或者Unigram算法来构造词汇表。SentencePiece除了集成了BPE、ULM子词算法之外，SentencePiece还能支持字符和词级别的分词。

下图是一些主流模型使用的分词算法，比如：GPT-1 使用的BPE实现分词，LLaMA/BLOOM/GPT2/ChatGLM使用BBPE实现分词。BERT/DistilBERT/Electra使用WordPiece进行分词，XLNet则采用了SentencePiece进行分词。

2、持续训练

知识如何注入

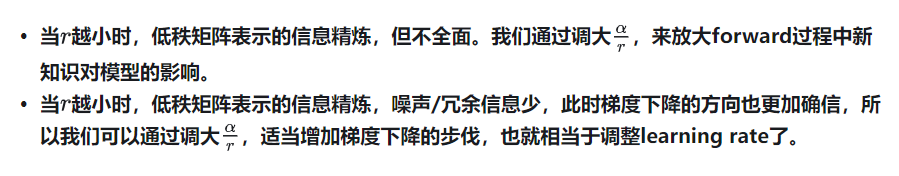
3、精调训练

4、lora lora 显存小原因

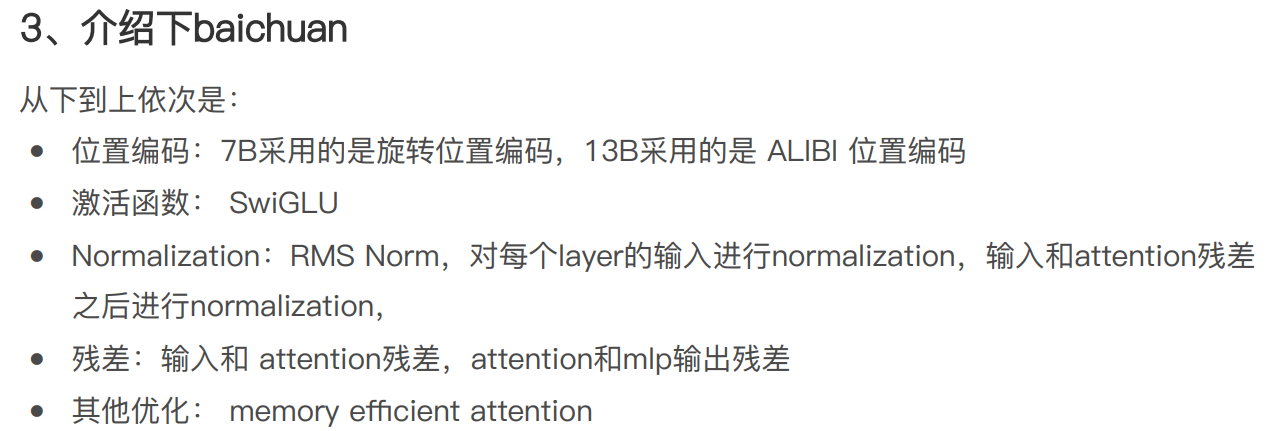
那既然通过数学方法直接做SVD行不通，那就让模型自己去学怎么做SVD吧

Dxd 2xrxd

优化器和梯度



5、chatglm和llama2区别 chatglm掩码



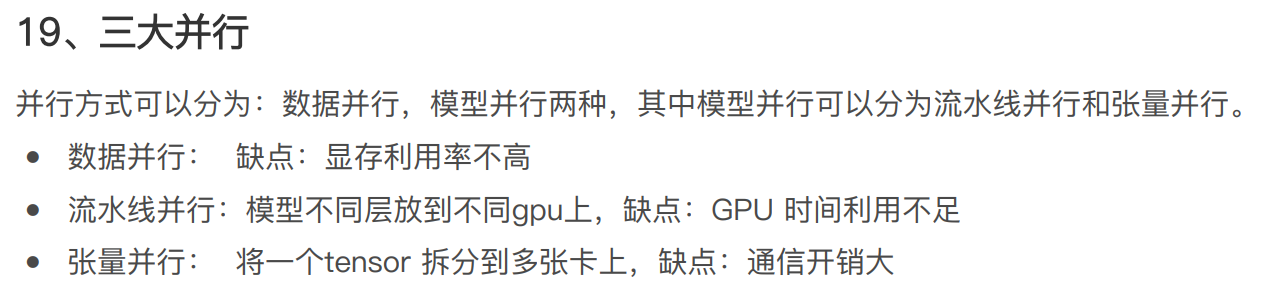
6、rlhf 代码优化 和 ppo输入数据

温度 top\_k top\_p

模型训练

Deepspeed 三阶段

ZeRO-R



COT 思维链 分治思想 复杂问题简单化 step by step

COT评价指标

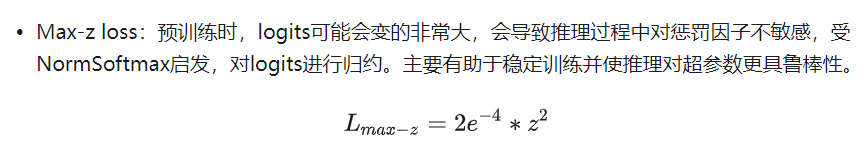
Algent

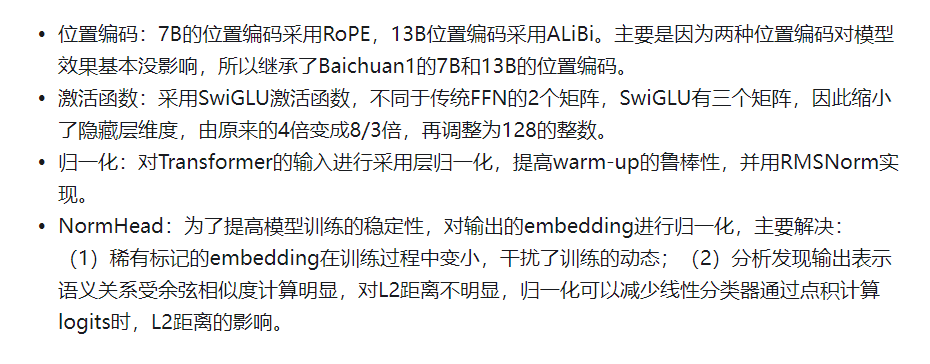
位置编码

正弦位置编码 绝对位置编码 rope alibi

位置编码外推性

ALIBI为啥比rope具有外推性

Baichuan2和llama2的区别



评价指标 准确率

Blue rouge

信息检索（IR）的评价指标介绍 - 准确率、召回率、F1、mAP、ROC、AUC

大模型幻觉 原因

复读机 原因

推理 vllm tensorllm

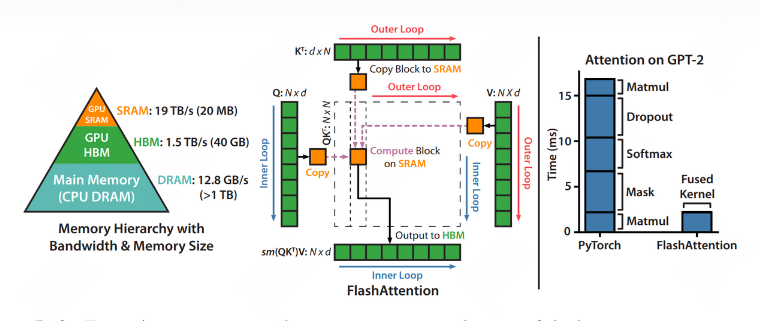
Vllm

推理调度 kuberay

Qkv的意义

Q和K求模式匹配相似度，作为权重加权平均V，作为当前位置的编码。

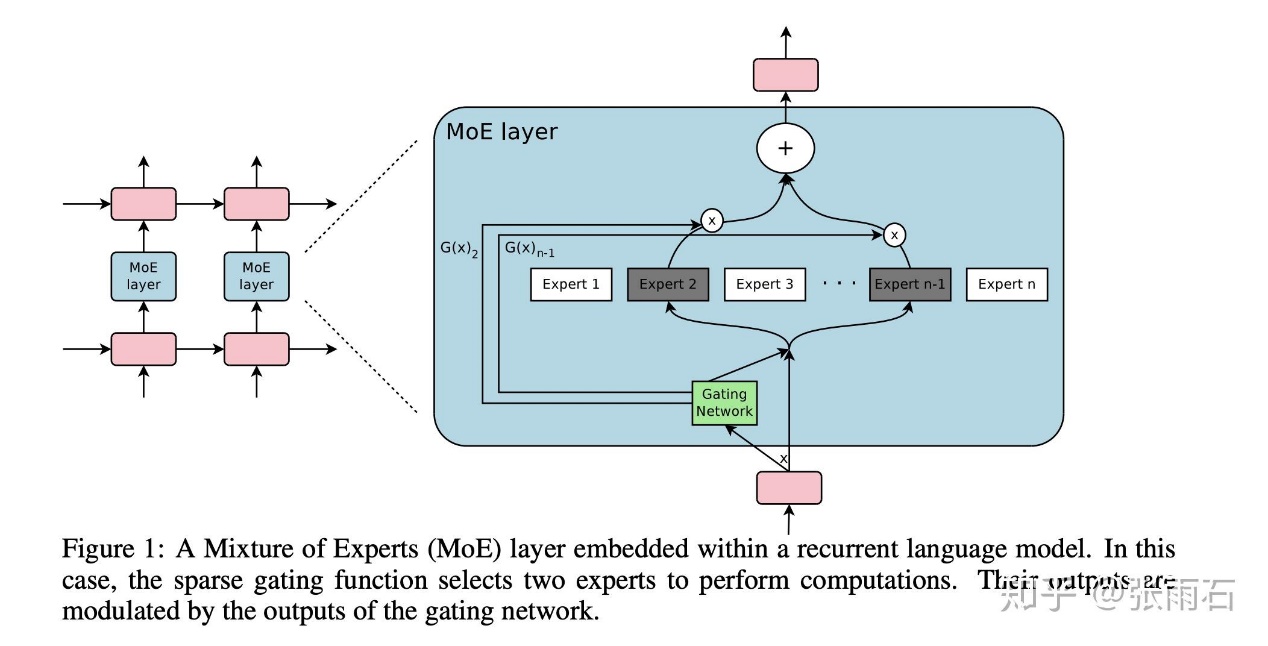
FlashAttention V1



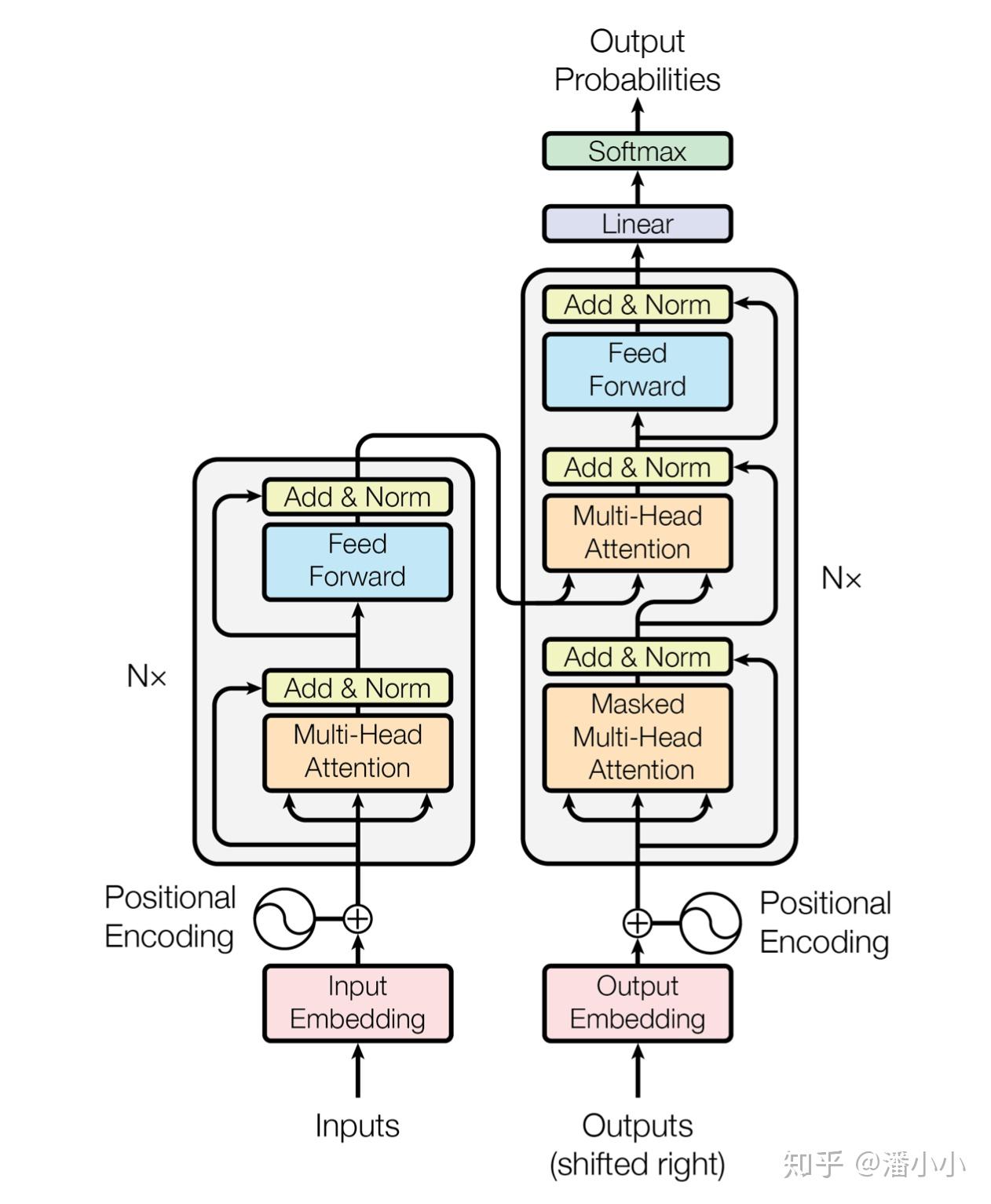
PageAttention

Page调度优化

浅谈MOE结构



Transformer结构



Bert llama区别

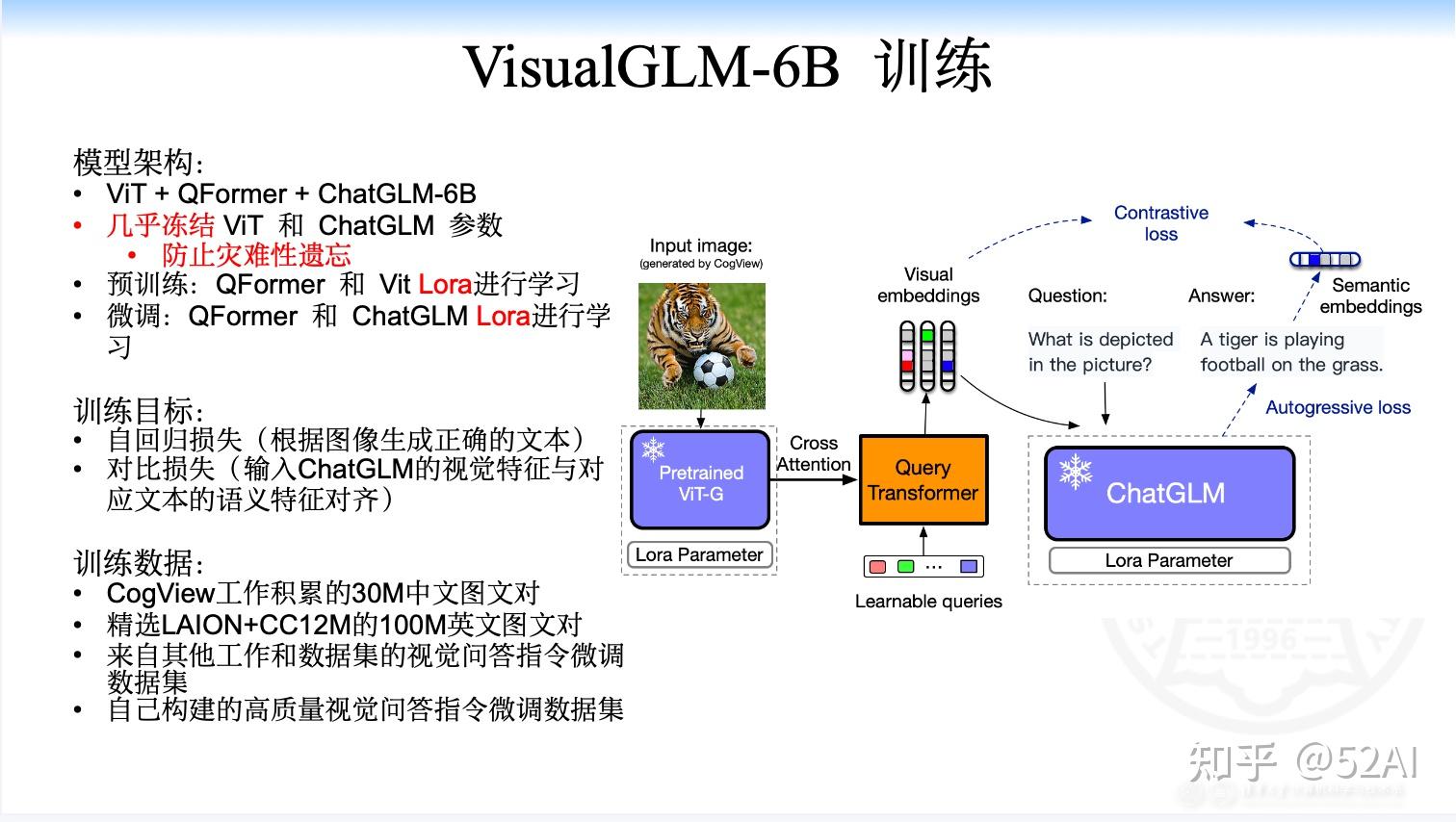
Encoder decoder

多模态

1、CogVLM

* CogVLM模型由四个基本组成部分组成：视觉转换器（ViT）编码器，MLP适配器，预训练大型语言模型（GPT）和视觉专家模块。
* ViT编码器
* 在CogVLM-17B中，我们使用预训练的EVA2-CLIP-E 。ViT编码器的最后一层被移除，因为它专门用于聚合[CLS]特征以进行对比学习。
* MLP适配器
* MLP适配器是一个两层MLP（SwiGLU），用于将ViT的输出映射到与词嵌入的文本特征相同的空间。所有图像特征在语言模型中共享相同的位置ID。
* 预训练大型语言模型
* CogVLM的模型设计与任何现成的GPT风格的预训练大型语言模型兼容。具体而言，CogVLM-17B采用Vicuna-7B-v1.5 进行进一步训练。对所有的注意力操作，包括图像特征之间的注意力，都应用了因果遮盖。
* 视觉专家模块
* 我们在每个层上添加了一个视觉专家模块，以实现深度的视觉-语言特征对齐。具体而言，每个层的视觉专家模块由一个QKV矩阵和一个MLP组成。QKV矩阵和MLP的形状与预训练语言模型中的形状相同，并从预训练模型中初始化。这样设计的动机是语言模型中的每个注意力头部捕捉了一定的语义信息，而可训练的视觉专家可以将图像特征转换为与不同头部对齐的形式，从而实现深度融合。

2、visualGLM-6b



3、CLIP算法

非常经典的就是 CLIP 了，用图像编码器编码 + 文本编码器编码，得到的特征让它们相乘，让成对儿样本的特征的乘积靠近1，让不成对儿的样本特征的乘积靠近0。

4、BLIP-2算法

为了减少计算成本并避免灾难性遗忘的问题，BLIP-2 在预训练时冻结预训练图像模型和语言模型，但是，简单地冻结预训练模型参数会导致视觉特征和文本特征难以对齐，为此BLIP-2提出两阶段预训练 Q-Former 来弥补模态差距：表示学习阶段和生成学习阶段。

