大模型校招大厂面试题

阿里大模型算法校招面试题（一）

1 自我介绍

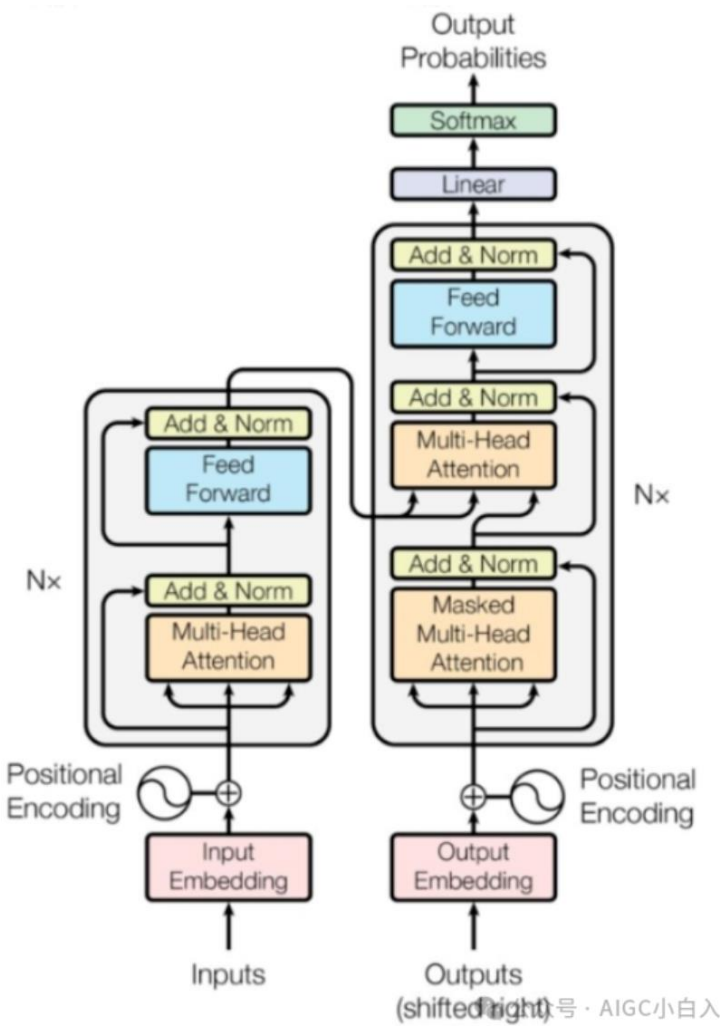
在自我介绍环节，我清晰地阐述了个人基本信息、教育背景、工作经历和技能特长，展示了自信和沟通能力。

2 技术问题回答

2.1 self-attention 的计算方式?

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

2.2 说一下 transformer 的模型架构和细节？



2.3 pre normalization 和 post normalization, layer norm，Batch norm

Layer Norm 篇

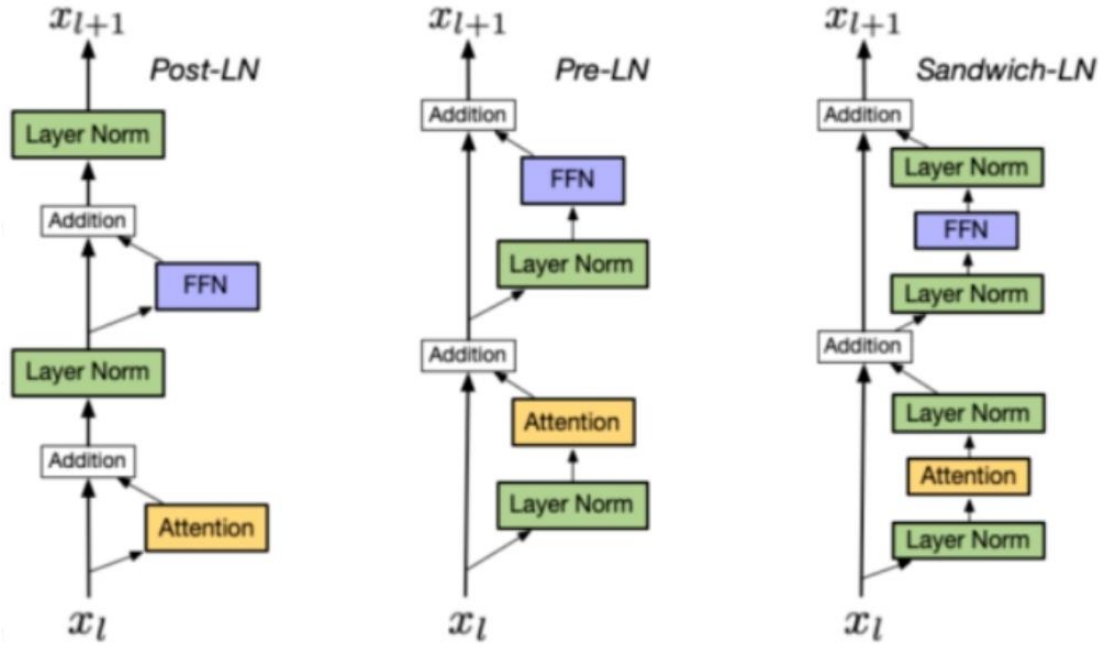
Layer Norm 的计算公式：

$$\mu = E(X) \leftarrow \frac{1}{H}\sum_{i=1}^{H} x_i$$

$$\sigma \leftarrow Var(x) = \sqrt{\frac{1}{H}\sum_{i=1}^{H}(x_i - \mu)^2 + \epsilon}$$

$$y = \frac{x - E(x)}{\sqrt{Var(X) + \epsilon}} \cdot \gamma + \beta$$

gamma: 可训练的再缩放参数
beta: 可训练的再偏移参数



Post LN：

位置：layer norm 在残差链接之后

缺点：Post LN 在深层的梯度范式逐渐增大，导致使用 post-LN 的深层 transformer 容易出现训练不稳定的问题

Pre-LN：

位置：layer norm 在残差链接中

优点：相比于 Post-LN，Pre LN 在深层的梯度范式近似相等，所以使用 Pre-LN 的深层 transformer 训练更稳定，可以缓解训练不稳定问题

缺点：相比于 Post-LN，Pre-LN 的模型效果略差

Sandwich-LN：

位置：在 pre-LN 的基础上，额外插入了一个 layer norm

优点：Cogview 用来避免值爆炸的问题

缺点：训练不稳定，可能会导致训练崩溃。

Layer normalization 对比篇

模型	normalization
GPT3	Pre layer Norm
LLaMA	Pre RMS Norm
baichuan	Pre RMS Norm
ChatGLM-6B	Post Deep Norm
ChatGLM2-6B	Post RMS Norm
Bloom	Pre layer Norm
Falcon	Pre layer Norm

BLOOM 在 embedding 层后添加 layer normalization，有利于提升训练稳定性:但可能会带来很大的性能损失.

2.4 BART、llama、gpt、t5、palm 等主流模型异同点？

BART (bi Encoder+casual Decoder，类 bert 的方法预训练)
T5 (Encoder+Decoder，text2text 预训练)
GPT(Decoder 主打 zero-shot)
GLM (mask 的输入部分是双向注意力，在生成预测的是单向注意力)

目前 主流的开源模型体系 分三种：
第一种：prefix Decoder 系
介绍：输入双向注意力，输出单向注意力
代表模型：ChatGLM、ChatGLM2、U-PaLM
第二种：causal Decoder 系
介绍：从左到右的单向注意力
代表模型：LLaMA-7B、LLaMa 衍生物
第三种：Encoder-Decoder
介绍：输入双向注意力，输出单向注意力
代表模型：T5、Flan-T5、BART

2.5 个人项目中模型的优化点和技术细节？

2.6 个人项目中如何选择最佳的指令策略，以及其对模型效果的影响 ？

2.7 个人项目中模型如何评测、数据集，评测指标等 ？

2.8 在指令微调中，如何设置、选择和优化不同的超参数，以及其对模型效果的影响？ 【涉及项目的问题不展开】

3 Leetcode 题目

类似 【11. 盛最多水的容器】

题目内容：

给定一个长度为 n 的整数数组 height 。有 n 条垂线，第 i 条线的两个端点是 (i, 0) 和 (i, height[i]) 。

找出其中的两条线，使得它们与 x 轴共同构成的容器可以容纳最多的水。

返回容器可以储存的最大水量。

说明：你不能倾斜容器。

示例 1：

输入：[1,8,6,2,5,4,8,3,7]

输出：49

解释：图中垂直线代表输入数组 [1,8,6,2,5,4,8,3,7]。在此情况下，容器能够容纳水（表示为蓝色部分）的最大值为 49。

示例 2：

输入：height = [1,1]

输出：1

题目解答

```
class Solution(object):
    def maxArea(self, height):
        """
        :type height: List[int]
        :rtype: int
        方法：左右指针
        思路：
            1. 定义 左右指针 l,r = 0, len(height)-1
            2. 定义 最大面积 max_area
            3. 计算 当前面积 temp_area = (r-l)*min(height[l],height[r])
            4. 判断 max_area，temp_area
            5. 移动 所指值 最小 的 指针
        """
        l,r = 0, len(height)-1
        max_area = 0
        while l<r:
            temp_area = (r-l)*min(height[l],height[r])
            max_area = max_area if temp_area<max_area else temp_area
            if height[l]<height[r]:
                l = l+1
            else:
                r = r-1
        return max_area
```

阿里大模型算法校招面试题（二）

技术问题回答

1. llama2 中使用的注意力机制是什么?手写实现下分组注意力。
2. 了解 langchain 吗?讲讲其结构。
3. 对位置编码熟悉吗?讲讲几种位置编码的异同
4. RLHF 的具体工程是什么?包含了哪几个模型?
5. 分别讲讲 encoder-only、decoder-only、encoder-decoder 几种大模型的代表作。
6. 具体讲讲 p-tuning、lora 等微调方法，并指出它们与传统 fine-tuning 微调有何不同。
7. 显存不够一般怎么解决的?
8. 几种主流大模型的 loss 了解过吗? 有哪些异同?
9. 了解半精度训练吗?展开讲讲。
10. deepspeed 用过吗? 展开讲讲。

【解答参考对应的 LLM 面试资料】

百度大模型算法校招面试题（一）

技术面

1 self-attention 的公式及参数量，为什么用多头，为什么要除以根号 d?

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

self-attention 模型在对当前位置的信息进行编码时，会过度的将注意力集中于自身的位置，因此作者提出了通过多头注意力机制来解决这一问题。同时，使用多头注意力机制还能够给予注意力层的输出包含有不同子空间中的编码表示信息，从而增强模型的表达能力。
这是因为点积的数量级增长很大，因此将 softmax 函数推向了梯度极小的区域。

2 你能不能介绍一下 BERT 和 GPT 的训练方式(预训练任务训练细节)的区别？

3 简单介绍一下，transformer 架构？

4 大模型的模型架构有哪些？

大模型。
用代码进行预训练。
Prompt/Instruction Tuning。
人类反馈的强化学习 (RLHF)

5 chatGPT 对比 GPT-3 的性能提升主要来源于哪些方面？

1. SFT:生成模型 GPT 的有监督精调(supervised fine-tuning) 。
2. RLHF
3. RM:奖励模型的训练 (reward modeltraining)。
4. PPO:近端策略优化模型 (ProximalPolicy Optimization) 。

6 InstructGPT 和 ChatGPT 模型中使用的关键技术(SFT->RLHF)

7 大模型中常见的位置编码？

8 大模型高效参数微调方法？

面试总结

1. 很多题目非常强调实践，没有做过大模型的项目且没有针对性准备过，很难回答上。
2. 大模型微调是很多公司的考察重点。
3. 几种模型的注意力机制、位置编码要熟悉。
4. RLHF 的几步多熟悉熟悉

百度大模型算法校招面试题（二）

Leetcode 题

【208. 实现 Trie (前缀树)】

技术面

1 结合 GNN 科研项目进行提问

样本构建的流程是怎样的，并且为什么 GCN 相较于其他方法在效果上更胜一筹？
节点特征指的是什么？

2 结合基于 RAG 的医学问答项目进行提问

查询流程？
使用什么向量数据库？
介绍一下 RAG 原理？
RAG 如何解决多实体提问问题？

用户提问：感冒和咳嗽需要吃什么药？

3 结合多模态科研项目进行提问

Prompt 是如何生成的，优化目标是什么，任务是什么？
OCR 抽取效果不好，需要怎么排查问题？

4 技术问题

您是否使用过 Pytorch 提供的预训练模型，例如 torchvision、transformers 以及 OpenAI 开源的 CIIP？对分布式训练有经验么？
回答：学过但是没用过

RNN 与 GNN 之间有哪些区别，以及它们各自适用于哪些场景？
回答：

RNN 与 GNN 的区别：
1. 数据类型：
- RNN 设计用于处理序列数据，即数据点按时间顺序排列，如时间序列分析、语音识别和自然语言处理。
- GNN 专门用于处理图结构数据，图由节点和边组成，代表实体及其关系，如社交网络、交通网络和分子结构。
2. 结构和工作原理：
- RNN 的核心是循环单元，它能够在序列的每个时间步上保持信息的状态，但是长序列会导致梯度消失或梯度爆炸问题，影响学习长期依赖。
- GNN 通过节点和边的特征以及图结构本身的信息，利用特殊的邻居节点更新机制来学习图中的特征表示，更好地捕捉节点间的依赖关系。
3. 长期依赖问题：
- RNN 在处理长序列时存在长期依赖问题，虽然有 LSTM（长短期记忆网络）等变体来缓解这一问题，但本质上是序列模型。
- GNN 通过图结构天然地能够捕捉节点间的依赖关系，因此在处理具有明确关系的数据时更为有效。

各自适用的场景：
- RNN 适用于处理时间序列数据、文本序列等，如股票价格预测、语音识别、机器翻译（序列到序列的任务）。
- GNN 适用于处理结构化数据，如社交网络分析、推荐系统、生物信息学（如蛋白质结构预测）、地理信息系统等，其中实体和关系是数据的核心组成部分。
总的来说，RNN 适合处理时间或顺序上的数据，而 GNN 适合处理具有明确结构关系的数据。两者各有优势，选择哪种模型取决于具体问题和数据的特点。
GPT 和 BERT 在文本表征方面有哪些结构和工作原理上的差异？
回答：BERT 是 Transformer Encoder，属于自监督训练方式，然后两大预训练任务，主要用于下游任务抽特征，GPT 是 Decoder，自回归训练，主要是预测下一个词的分布，依赖大语料库，GPT-3 可以表现出 Few-shot/zero-shot learning

因为说了 BERT 好训练一些，问了为什么？
回答：说了 GPT 任务对简单、比较依赖语料库的大小，BERT 的 MLM 比较直觉且个人能训

练，GPT 只有 openai 等公司有成品

说一说你对 Zero-shot 和 Few-shot 的理解
回答：Few-shot 先给定任务范式描述，Zero-shot 就是直接做

怎么看待计算机网络和操作系统在 DL 中的作用
回答：谈了 DL 研究一些计算机网路的问题，比如网络拓扑、交换机拓扑等，分布式训练时会有通信，也会用到进程相关知识

你来调优一个 BERT 模型适应一个数据集或任务会怎么做
回答：固定 BERT，训练分类头或者使用 Adapter

训练完模型后准确率很低，怎么优化
回答：首先检查代码结构和分类器的网络结构和 BERT 量级是否匹配，学习率+余弦退火调整，改为 Adapter，检查数据集质量，验证阶段代码是否有误

有一批文本数据，来源和质量不太一样，使用时如何处理
回答：反问文本来源不同是否混合或完全分开，结合多模态融合的技术，增加一个学习任务，对不同来源的文本表示进行线性变换投影到相同的特征空间中

腾讯大模型算法校招面试题（一）

1. 自我介绍

在自我介绍环节，我清晰地阐述了个人基本信息、教育背景、工作经历和技能特长，展示了自信和沟通能力。

2. 技术问题回答

2.1 分布式训练框架都了解哪些，能不能简单介绍一下？

2.2 你了解 deepspeed，那介绍 zero1 2 3 分别是什么，分析训练时候显存占用？

2.3 说一下 Transformer 的架构和其内部细节？【必考题】

2.4 介绍大模型推理过程中，可以通过调节哪些参数提高性能？

2.5 你既然做过 RAG，能不能介绍一下 RAG，大模型在里面主要是起到什么作用?

2.6 大模型训练的三种并行是什么?通讯开销比?

模型并行，数据并行，流水线并行

2.7 手撕代码。给一个 md 维度的矩阵，m 代表样本数量，d 是样本的维度。请使用不超过 mm 复杂度的代码求解其两两之间的欧式距离?

理想汽车大模型算法校招面试题（一）

项目面

因为我简历上面写一个 RAG 项目，所以面试官主要围绕 RAG 进行提问

1. 聊一下 RAG 项目总体思路？
2. 使用外挂知识库主要是为了解决什么问题？
3. 如何评价 RAG 项目的效果好坏，即指标是什么？
4. 在做 RAG 项目过程中遇到哪些问题？怎么解决的？
5. RAG 项目里面有哪一些亮点？目前开源的 RAG 项目非常多，你的项目和他们有什么区别？
6. 数据集怎么构建的，什么规模，评估指标是什么，这些指标存在哪些问题？
7. 模型底座是什么，这些不同底座什么区别，什么规模？
8. 使用哪一种训练方法，什么 sft，这些方法有什么不同，有什么优缺点，原理上解释不不同方法的差别？
9. 模型推理是怎么做的，有没有 cot，tot 等等，还是单轮？
10. 大模型可控性如何实现，怎么保证可控性？
11. 模型部署的平台，推理效率怎么样，如何提升推理效率？
12. 项目最后上线了么，上线之后发现什么问题，如何解决？
13. 给一个总的输入输出样例，每一步包含什么 prompt，多轮推理每一步输出什么结果，模拟一下，数据集格式是否要调整成这样，数据形式是什么，怎么拆分成多轮形式？

技术问题回答

1 简单介绍一下大模型存在哪些问题？有什么好的解决方法？

大模型幻觉问题
外挂知识库

大模型微调
强化学习等

2 大模型加速框架了解多少，知不知道原理 如何进行加速优化？

1. vLLM
vLLm 运行大模型非常快主要使用以下方法实现的
先进的服务吞吐量
通过 PageAttention 对 attention key & value 内存进行有效的管理
对于输入请求的连续批处理
高度优化的 CUDA kernels
2. OpenLLM
OpenLLM 运行大模型非常快主要使用以下方法实现的
促进实际生产过程中的大模型的部署，微调，服务和监测.
3. DeepSpeed-MII
DeepSpeed-MII 运行大模型非常快主要使用以下方法实现的
MII(Model Implementations for Inference) 提供加速的文本生成推理通过 Blocked KV Caching, Continuous Batching, Dynamic SplitFuse 和高性能的 CUDA Kernels
4. TensorRT-llm

3 为什么要用大模型做传统结构化解析任务你对用大模型做这些事有什么看法？

某大厂大模型技术面问答

1 介绍一下，现在几种流行的大模型架构?

BART (bi Encoder+casual Decoder，类 bert 的方法预训练)
T5 (Encoder+Decoder，text2text 预训练)
GPT(Decoder 主打 zero-shot)
GLM (mask 的输入部分是双向注意力，在生成预测的是单向注意力)

2 说一下 prefix LM 和 casualLM 的区别？

prefix LM 和 casualLM 所用的 attention mask 不一样：

prefix LM：token 可以相互看到；
casualLM：严格自回归

3 在大模型任务中，你用到 LoRA，讲一下 LoRA 实现原理?

LoRA 的思想很简单:

在原始 PLM (Pre-trained Language Model) 旁边增加一个旁路，做一个降维再升维的操作，来模拟所谓的 intrinsic rank。

训练的时候固定 PLM 的参数，只训练降维矩阵 A 与升维矩阵 B 。而模型的输入输出维度不变，输出时将 BA 与 PLM 的参数叠加。
用随机高斯分布初始化 A ，用 0 矩阵初始化 B ，保证训练的开始此旁路矩阵依然是 0 矩阵。

4 instruction tuning 和 prompt learning 的区别?

instruction tuning 和 prompt learning 的目的都是去挖掘语言模型本身具备的知识。不同的是 Prompt 是激发语言模型的补全能力，例如根据上半句生成下半句，或是完形填空等(few-shot)。Instruct 是激发语言模型的理解能力，它通过给出更明显的指令，让模型去做出正确的行动(zero-shot)。

5 项目中你用到的大模型推理加速工具是什么？简单介绍一下为什么用它？

【具体题目解答参考各专题面试题总结】