大模型面经——MoE混合专家模型总结

原创 喜欢卷卷的瓦力 瓦力算法学研所 2024年09月08日 17:45 安徽

面试总结专栏

本篇将介绍MoE（Mixture of Experts，混合专家模型）相关面试题。

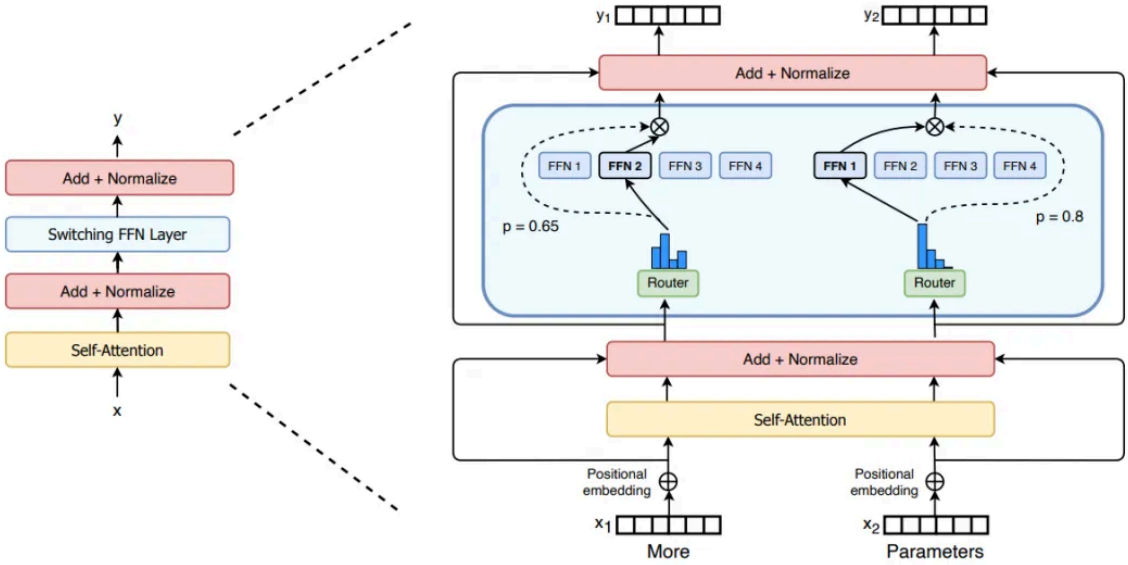
以下是一个快捷目录：

回答

一、MoE介绍

"Mixture of Experts"（MoE）是一种机器学习模型，特别是在深度学习领域中，它属于集成学习的一种形式。MoE模型由多个专家（experts）和一个门控网络（gating network）组成。每个专家负责处理输入数据的不同部分或不同特征，而门控网络则负责决定每个输入应该由哪个专家来处理。

例如，在下图中，“More”这个 token 被发送到第二个专家，而“Parameters”这个 token 被发送到第一个专家。



二、MoE出现的背景

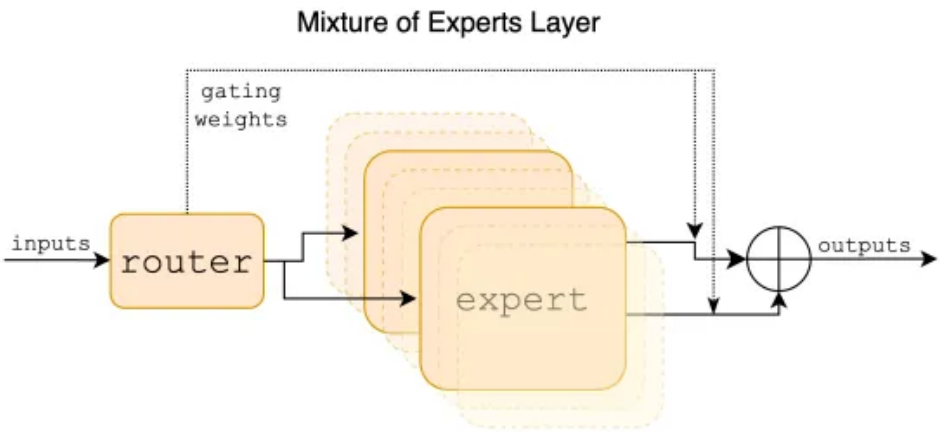
本质上来说就是一种高效的 scaling 技术，用较少的 compute 实现更大的模型规模，从而获得更好的性能。

三、有哪些MoE模型

Switch Transformers、Mixtral、GShard、DBRX、Jamba DeepSeekMoE 等等。

以Mixtral为例

Mixtral 是一个稀疏的专家混合网络。它是一个decoder-only的模型，其中前馈块从一组 8 个不同的参数组中选择。在每一层，对于每个令牌，路由器网络选择其中两个组（“专家”）来处理令牌并附加地组合他们的输出。



混合专家层

这种技术在控制成本和延迟的同时增加了模型的参数数量，因为模型只使用每个令牌总参数集的一小部分。具体来说，Mixtral 总共有 46.7B 个参数，但每个令牌只使用 12.9B 个参数。因此，它以与 12.9B 型号相同的速度和相同的成本处理输入和生成输出。

Mixtral 基于从开放 Web 中提取的数据进行预训练——同时培训专家和路由器。

四、介绍稀疏 MoE 层

稀疏 MoE 层一般用来替代传统 Transformer 模型中的前馈网络 (FFN) 层。MoE 层包含若干“专家” (例如 8 个)，每个专家本身是一个独立的神经网络。在实际应用中，这些专家通常是前馈网络 (FFN)，但它们也可以是更复杂的网络结构，甚至可以是 MoE 层本身，从而形成层级式的 MoE 结构。

五、介绍门控网络或路由

门控网络接收输入数据并执行一系列学习的非线性变换。这一过程产生了一组权重，这些权重表示了每个专家对当前输入的贡献程度。通常，这些权重经过softmax等函数的处理，以确保它们相加为1，形成了一个概率分布。这样的分布表示了在给定输入情境下每个专家被激活的概率。一个典型的门控函数通常是一个带有 softmax 函数的简单的网络。

六、为什么门控网络要引入噪声呢

为了专家间的负载均衡。也即防止一句话中的大部分token都只有一个专家来处理，剩下的七个专家（假设一共八个专家）“无所事事”。

七、如何均衡专家间的负载

引入噪声、引入辅助损失（鼓励给予所有专家相同的重要性）、引入随机路由、设置一个专家能处理的token数量上限

八、“专家”指什么

一个“专家”通常是前馈网络 (FFN)。数据经过门控网络选择后进入每个专家模型，每个专家根据其设计和参数对输入进行处理。每个专家产生的输出是对输入数据的一种表示，这些表示将在后续的步骤中进行加权聚合。或者通过单个专家模型进行处理。

九、专家的数量对预训练有何影响?

增加更多专家可以提升处理样本的效率和加速模型的运算速度，但这些优势随着专家数量的增加而递减 (尤其是当专家数量达到 256 或 512 之后更为明显)。同时，这也意味着在推理过程中，需要更多的显存来加载整个模型。值得注意的是，Switch Transformers 的研究表明，其在大规模模型中的特性在小规模模型下也同样适用，即便是每层仅包含 2、4 或 8 个专家。

十、什么是topK门控

选择前k个专家。为什么不仅选择最顶尖的专家呢？最初的假设是，需要将输入路由到不止一个专家，以便门控学会如何进行有效的路由选择，因此至少需要选择两个专家。

十一、MoE模型的主要特点：

- 灵活性：每个专家可以是不同类型的模型，例如全连接层、卷积层或者递归神经网络。
- 可扩展性：通过增加专家的数量，模型可以处理更复杂的任务。
- 并行处理：不同的专家可以并行处理数据，这有助于提高模型的计算效率。

- 动态权重分配：门控网络根据输入数据的特点动态地为每个专家分配权重，这样模型可以更加灵活地适应不同的数据。
- 容错性：即使某些专家表现不佳，其他专家的表现也可以弥补，从而提高整体模型的鲁棒性。

十二、moe和稠密模型的对比

1、预训练

相同计算资源，MoE 模型理论上可以比密集模型更快达到相同的性能水平。

2、推理

moe：高显存，高吞吐量；

稠密模型：低显存，低吞吐量

十三、moe的优势

1、训练优势：预训练速度更快；

2、推理优势：推理速度更快

十四、moe的挑战

1、训练挑战：微调阶段，泛化能力不足，容易过拟合

2、推理挑战：对显存的要求更高

十五、微调moe的方法

1、冻结所有非专家层的权重，专门只训练专家层

2、只冻结moe层参数，训练其它层的参数

十六、moe的并行计算

图片来源：b站 算法猪立业

参考资料

https://blog.csdn.net/2201_75499313/article/details/136412787

https://www.zhihu.com/question/634844209/answer/3467132890

https://zhuanlan.zhihu.com/p/674698482

https://b23.tv/jCL0r4N

想要获取技术资料的同学欢迎关注公众号，进群一起交流~

上一篇
多模态大模型中，多模态融合后怎样知道最终结果受哪种模态影响更大?

下一篇
大模型面经——以医疗领域为例，整理RAG基础与实际应用中的痛点