LLaMA 常见面试题篇

来自： AiGC面试宝典

宁静致远

2024年01月27日 19:44

一、相比较于llama而言，llama2有哪些改进，对于llama2是应该如何finetune?

llama和llama2都是一种大型语言模型（Large Language Model，LLM），它们可以用于多种自然语言处理的任务，如文本生成、文本摘要、机器翻译、问答等。

llama是一种基于Transformer的seq2seq模型，它**使用了两种预训练任务，一种是无监督的Span级别的mask，另一种是有监督的多任务学习**。llama将所有的下游任务都视为文本到文本的转换问题，即给定一个输入文本，生成一个输出文本。llama使用了一个干净的大规模英文预料C4，包含了约750GB的文本数据。llama2是llama的改进版本，它在以下几个方面有所提升：

- 数据量和质量：llama2使用了比llama1多40%的数据进行预训练，其中包括更多的高质量和多样性的数据，例如来自Surge和Scale等数据标注公司的数据。
- 上下文长度：llama2的上下文长度是llama1的两倍，达到了4k个标记，这有助于模型理解更长的文本和更复杂的逻辑。
- 模型架构：llama2在训练34B和70B参数的模型时使用了分组查询注意力（Grouped-Query Attention，GQA）技术，可以提高模型的推理速度和质量。
- 微调方法：llama2使用了监督微调（Supervised Fine-Tuning，SFT）和人类反馈强化学习（Reinforcement Learning from Human Feedback，RLHF）两种方法来微调对话模型（llama2-chat），使模型在有用性和安全性方面都有显著提升。

对llama2进行微调有以下步骤：

- 准备训练脚本：你可以使用Meta开源的llama-recipes项目，它提供了一些快速开始的示例和配置文件，以及一些自定义数据集和策略的方法。
- 准备数据集：你可以选择一个符合你目标任务和领域的数据集，例如GuanacoDataset，它是一个多语言的对话数据集，支持alpaca格式。你也可以使用自己的数据集，只要按照alpaca格式进行组织即可。
- 准备模型：你可以从Hugging Face Hub下载llama2模型的权重，并转换为Hugging Face格式。
- 启动训练：你可以使用单GPU或多GPU来进行训练，并选择是否使用参数高效微调（Parameter-Efficient Fine-Tuning，PEFT）或量化等技术来加速训练过程。具体命令可以参考这里。