LLMs 训练经验帖

来自： AiGC面试宝典

宁静致远

2023年09月28日 22:03

分布式训练框架选择?

多用 DeepSpeed，少用 Pytorch 原生的 torchrun。在节点数量较少的情况下，使用何种训练框架并不是特别重要；然而，一旦涉及到数百个节点，DeepSpeed显现出其强大之处，其简便的启动和便于性能分析的特点使其成为理想之选。

LLMs 训练时 有哪些有用的建议?

1. 弹性容错和自动重启机制

大模型训练不是以往那种单机训个几小时就结束的任务，往往需要训练好几周甚至好几个月，这时候你就知道能稳定训练有多么重要。**弹性容错能让你在机器故障的情况下依然继续重启训练**；自动重启能让你在训练中断之后立刻重启训练。毕竟，大模型时代，节约时间就是节约钱。

1. 定期保存模型

训练的时候每隔一段时间做个checkpointing，这样如果训练中断还能从上次的断点来恢复训练。

1. 想清楚再开始训练

训练一次大模型的成本很高的。**在训练之前先想清楚这次训练的目的，记录训练参数和中间过程结果，少做重复劳动。**

1. 关注GPU使用效率

有时候，即使增加了多块 A100 GPU，大型模型的训练速度未必会加快，这很可能是因为GPU使用效率不高，尤其在多机训练情况下更为明显。仅仅依赖nvidia-smi显示的GPU 利用率并不足以准确反映实际情况，因为即使显示为100%，实际GPU利用率也可能不是真正的 100%。**要更准确地评估GPU利用率，需要关注TFLOPS和吞吐率等指标，这些监控在DeepSpeed框架中都得以整合。**

1. 不同的训练框架 对 同一个模型 影响不同

对于同一模型，选择不同的训练框架，对于资源的消耗情况可能存在显著差异（比如使用Huggingface Transformers和DeepSpeed训练OPT-30相对于使用Alpa对于资源的消耗会低不少）。

1. 环境问题

针对已有的环境进行分布式训练环境搭建时，一定要注意之前环境的python、pip、virtualenv、setuptools的版本。不然创建的虚拟环境即使指定对了Python版本，也可能会遇到很多安装依赖库的问题（GPU服务器能够访问外网的情况下，建议使用Docker相对来说更方便）。

1. 升级GLIBC等底层库问题

遇到需要升级GLIBC等底层库需要升级的提示时，一定要慎重，不要轻易升级，否则，可能会造成系统宕机或很多命令无法操作等情况。

模型大小如何选择?

进行大模型模型训练时，先使用小规模模型（如：OPT-125m/2.7b）进行尝试，然后再进行大规模模型（如：OPT-13b/30b...）的尝试，便于出现问题时进行排查。目前来看，业界也是基于相对较小规模参数的模型（6B/7B/13B）进行的优化，同时，13B模型经过指令精调之后的模型效果已经能够到达GPT4的90%的效果。

加速卡如何选择?

于一些国产AI加速卡，目前来说，坑还比较多，如果时间不是时间非常充裕，还是尽量选择Nvidia的AI加速卡。