

Megatron-LM，又一大模型训练神器

原创 喜欢瓦力的卷卷 瓦力算法学研所 2024年08月14日 15:55 广东

◇◇ 技术总结专栏 ◇◇

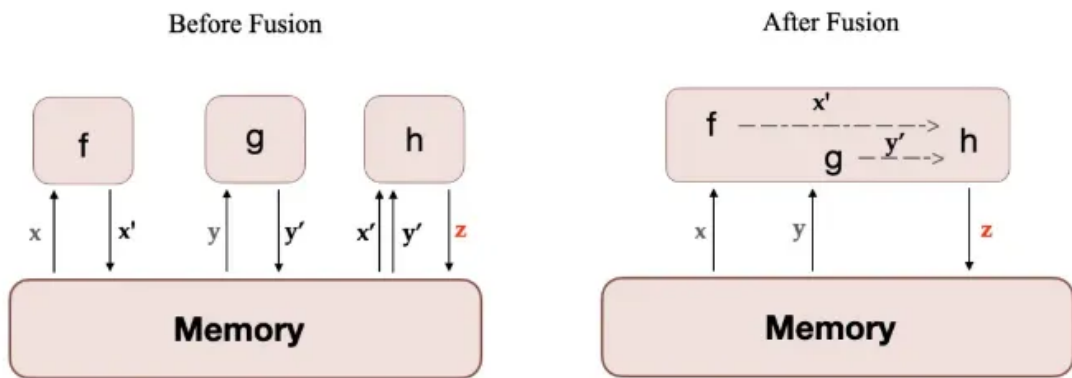
本文介绍了Megatron-LM的相关知识。

NVIDIA Megatron-LM 是一个建立在 PyTorch 之上的分布式训练框架，旨在培养基于 Transformer 的大型语言模型。Megatron-LM 结合了数据并行、张量并行和流水线并行技术，用以实现类似于 GPT-3 的训练效果。

Megatron-LM优势

数据加载：Megatron-LM 提供了一个高效的 DataLoader，它在训练前对数据进行 tokenize 和 shuffle。此外，它还可以将数据拆分为带有索引的编号序列，并将这些索引存储起来，这样在后续的训练过程中就无需再次进行 tokenize 操作。为了实现这一点，首先需要根据训练参数计算出每个 epoch 的数据量，然后创建一个排序，接着对数据进行 shuffle 操作。这与大多数情况下的处理方式不同，通常情况下我们会遍历整个数据集直到其用尽，然后再开始第二个 epoch。这种处理方式可以平滑学习曲线并节省训练时间。

融合 CUDA 内核：当计算任务在GPU上进行时，所需的数据会从内存中提取并加载到GPU上，随后计算结果会被写回内存。融合内核的基本概念是将通常由PyTorch分开执行的相似操作合并为一个单独的硬件操作。这样做可以将多个分散的计算任务整合成一个任务，进而减少在这些分散计算中的内存传输次数。下图展示了内核融合的概念。



指导原则

Megatron的开发者对于各种不同的并行模式及其相互作用进行了深入研究，并总结出了一些分布式训练的指导原则：

- **并行模式相互作用：**不同的并行策略会以复杂方式相互作用。所选的并行模式会影响通信量、计算核心的效率，以及因流水线刷新引起的工作节点空闲时间。例如，虽然张量模型并行在多GPU服务器上表现良好，但对于大型模型而言，采用流水线模型并行更为理想。
- **流水线并行调度影响：**在流水线并行中，使用的调度策略会影响通信量、流水线中的空隙大小及存储激活所需的内存。Megatron引入了一种新的交错调度方法，该方法在略微增加内存使用的前提下，能提高高达10%的吞吐量。
- **超参数影响：**诸如微批处理大小这类超参数的值，会影响内存占用、工作节点上核心的执行效率及流水线空隙的大小。
- **通信密集性：**分布式训练是一个通信密集型的过程。使用速度较慢的节点连接或更通信密集的分区会限制性能表现。


想要获取技术资料的同学欢迎关注公众号，进群一起交流~



瓦力算法学研所

我们是一个致力于分享人工智能、机器学习和数据科学方面理论与应用知识的公众号。我...
117篇原创内容

公众号



喜欢卷卷的瓦力

扫一扫上面的二维码图案，加我为朋友。

添加瓦力微信

算法交流群 · 面试群

大咖分享 · 学习打卡

👤 公众号 · 瓦力算法学研所

学术理论解析 53

学术理论解析 · 目录

上一篇

从大模型推理极限理论最优值谈谈推理优化

下一篇

大语言模型也有mbti?