大模型面经——大模型训练中超参数的设置与训练数据偏好

原创 喜欢卷卷的瓦力 瓦力算法学研所 2024年05月16日 14:21 广东

◈ 技术总结专栏 ◈

作者：vivida

本篇主要从训练设置（batch size及优化器设置）、训练数据选择两大角度分享大模型训练与微调经验。

本篇开始填大模型面经——超细节大模型训练与微调实操经验总结（上）的坑，继续细节的讲讲大模型中训练和微调的经验。

本篇主要从训练设置（batch size及优化器设置）、训练数据选择两大角度来具体谈谈经验，下面是一个问题的快捷目录。

1. 训练大模型时，batch size如何设置比较合理，可以讲讲自己的思考
2. 如果batch size设置过小或过大分别会怎样?
3. 微调时优化器怎么设置好?
4. 预训练和微调时选择的训练数据分别有什么偏好，有没有一些建议?

batch size如何设置比较合理

我们知道，大模型训练或微调的过程中batch size的设置本质上是取训练效率和模型的最终效果的平衡。

目前一些研究结果表明数据并行程度的临界点是存在的，我们这里先上一个结论：

batch size在一定临界值以内越大越好，超过临界值会开始收益持平或者递减；并且batch size需要跟其他超参数比如学习率、优化器搭配等相适配。

下面先给大家一些不同大模型中batch size的参考值：

- OpenAI的GPT-3模型使用了约3500万个token的batch size；
- 谷歌的PaLM模型使用了2048个样本的batch size；
- llama3至少也使用了上千个样本的batch size。

下面我们基于openAI的一些研究工作，学习一下openAI是如何基于一些研究理论在后续scaling law工作中预测模型的最优batch size的：

1. 最优步长

各种研究结果表明实际是存在一个关于数据并行程度的临界点的，找到这个临界点，就可以有效的平衡训练的效率和模型的最终效果。

OpenAI 发现最优步长公式可以作如下表示:

$$\epsilon_{opt}(B) = argmin_{\epsilon}\mathbb{E}[L(\theta - \epsilon G_{est})] = \frac{\epsilon_{max}}{1+\mathcal{B}_{noise}/B}$$

注：B 为 batch size，Bnoise为 噪声尺度

2. 损失更新

我们基于第1步中最优步长（step size），现在开始改进最优化从含有噪声的梯度中获得的损失：

$$\Delta L_{opt}(B) = \frac{\Delta L_{max}}{1+\mathcal{B}_{noise}/B}$$

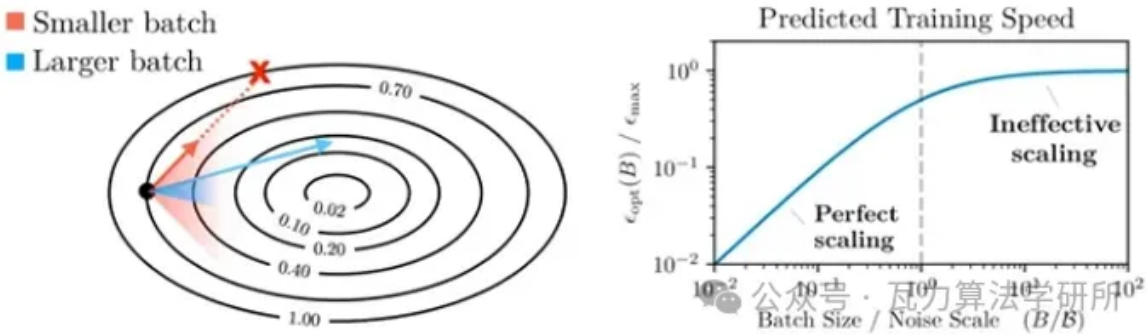
$$\Delta L_{max} = \frac{|G|^4}{2G^THG}$$

上述公式中我们主要观察真实梯度、步长和B的关系，可以得出：

1）无论如何准确地估计真实梯度，总存在一个最大步长

2）批处理大小越大，优化模型的步长就越大（有一个上限）

下面再来看两张比较经典的图,



上面的图说明了更大的批次模型可以取得更多提升。但是当 batch size 太大时，我们会遇到收益递减的问题（因为分母中的 1 开始占主导地位）。

3. 梯度尺度估计

OpenAI 的研究发现，噪声尺度可以通过以下方式估计：

$$\mathcal{B}_{noise} = \frac{tr(H\Sigma)}{G^T H G}$$

其中，H 是参数的真实 Hessian 矩阵，C 是相对于梯度的每个示例的协方差矩阵，g 是真实梯度。

为了进一步简化这个方程，OpenAI作出了一个假设，即优化是完全 well-conditioned 的。
在这种情况下，Hessian 矩阵只是单位矩阵的倍数，噪声尺度简化就可以简化为以下形式：

$$\mathcal{B}_{simple} = \frac{tr(\Sigma)}{G^2}$$

该方程表明噪声尺度等于个别梯度分量的方差之和，除以梯度的 norm。

4. 学习率视作temperature

之前说上述结论合理的前提，是在模型的 LR 是调的比较恰当的情况下。

这是因为 OpenAI 发现噪声尺度基本符合以下规律：

$$T(\epsilon, B) \equiv \frac{\epsilon}{\epsilon_{max}(B)}$$

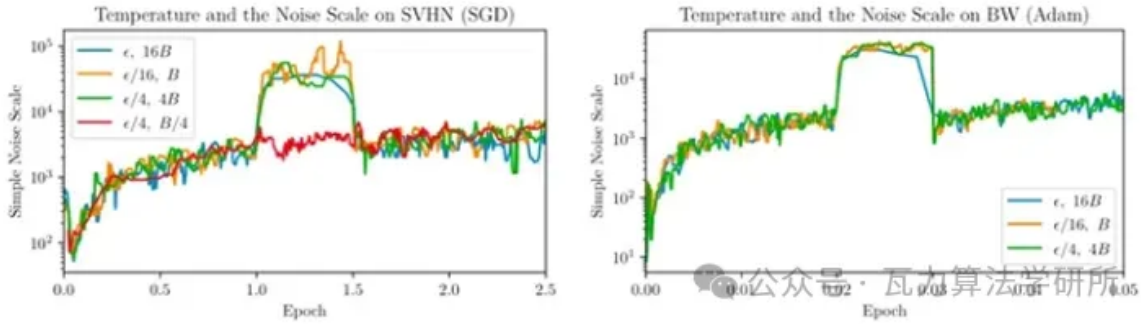
在使用 SGD 和小 batch 进行更新时，可以大概近似为

$$T \approx \frac{\epsilon}{B}.$$

这表明

$$\mathcal{B}_{noise} \propto \mathcal{B}_{simple} \propto \frac{1}{T}$$

下面的图中可以看出规律

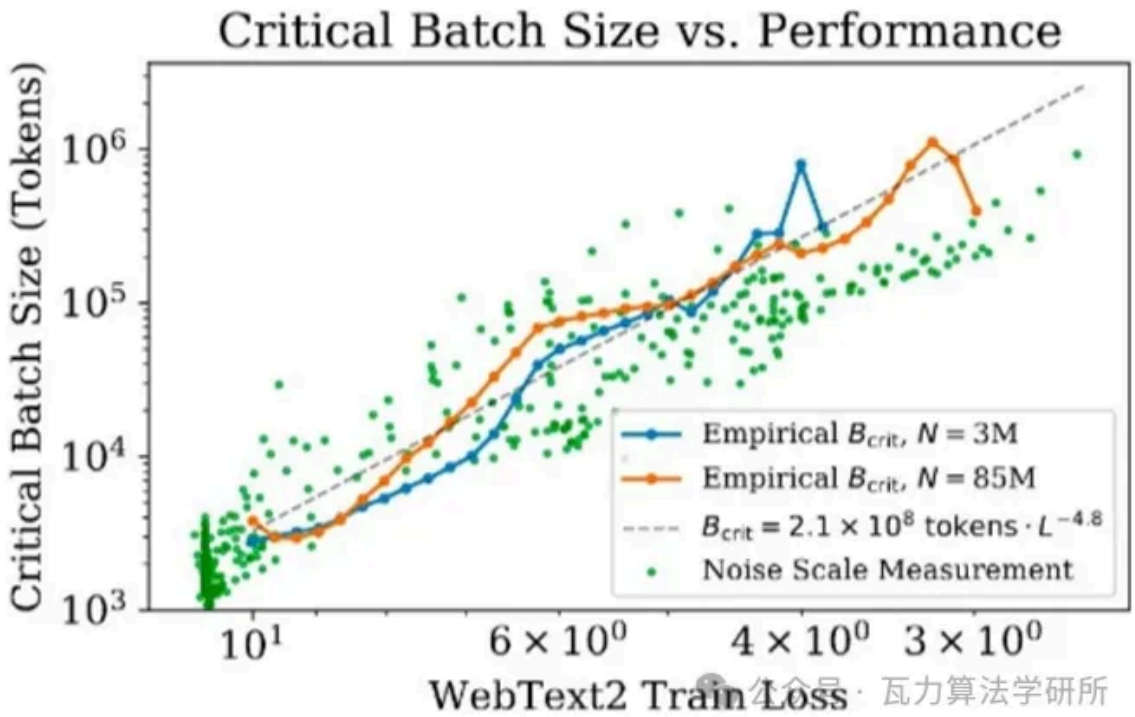


根据以上的公式和图，我们可以得知:

1. 高温度导致较小的噪声尺度。其中的直觉是在高温度下，相对于方差，梯度幅度较大。

2. 当学习率以一个常数因子衰减时，噪声尺度大致以相同的因子增长。因此，如果学习率太小，噪声尺度将被放大。

OpenAI 使用上述结论，在模型训练推理中在后续的 scaling law 工作中预测了模型的最优 batch size 大小，具体如下图。



batch size设置过小或过大分别会怎样

1. 过小

更新方向（即对真实梯度的近似）会具有很高的方差，导致的梯度更新主要是噪声。经过一些更新后，方差会相互抵消。

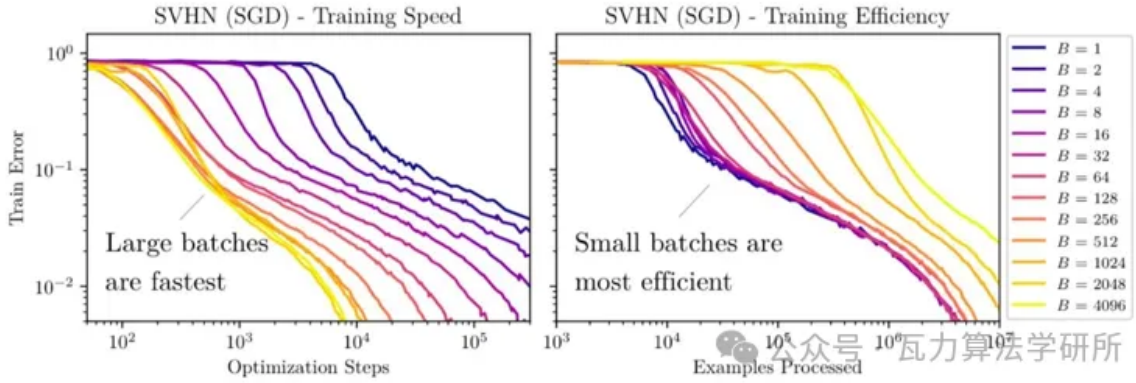
总体上推动模型朝着正确的方向前进，但个别更新可能不太有用，可以使用更大 batch size 进行更新。

2. 过大

当 batch size 非常大时，从训练数据中抽样的任何两组数据都会非常相似（因为它们几乎完全匹配真实梯度）

在这种情况下，增加 batch size 几乎不会改善性能，因为无法改进真实的梯度预测。

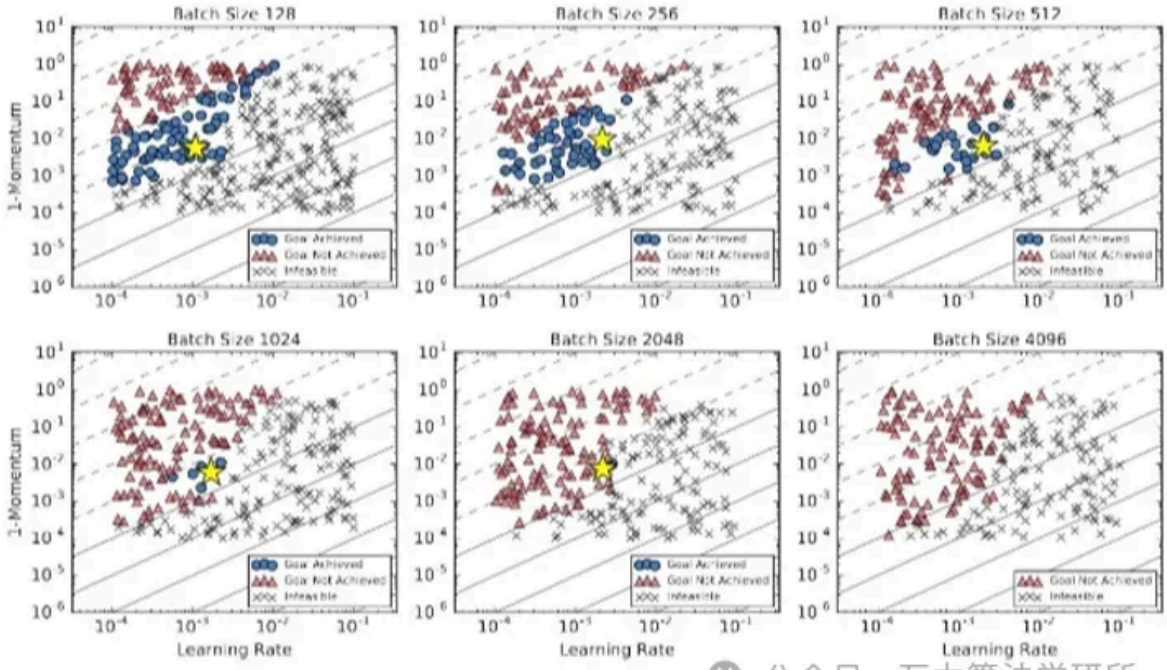
换句话说，需要在每一步中处理更多的数据，但并不能减少整个训练过程中的步数，这表明总体训练时间几乎没有改善，**还增加了总体的 FLOPS。**



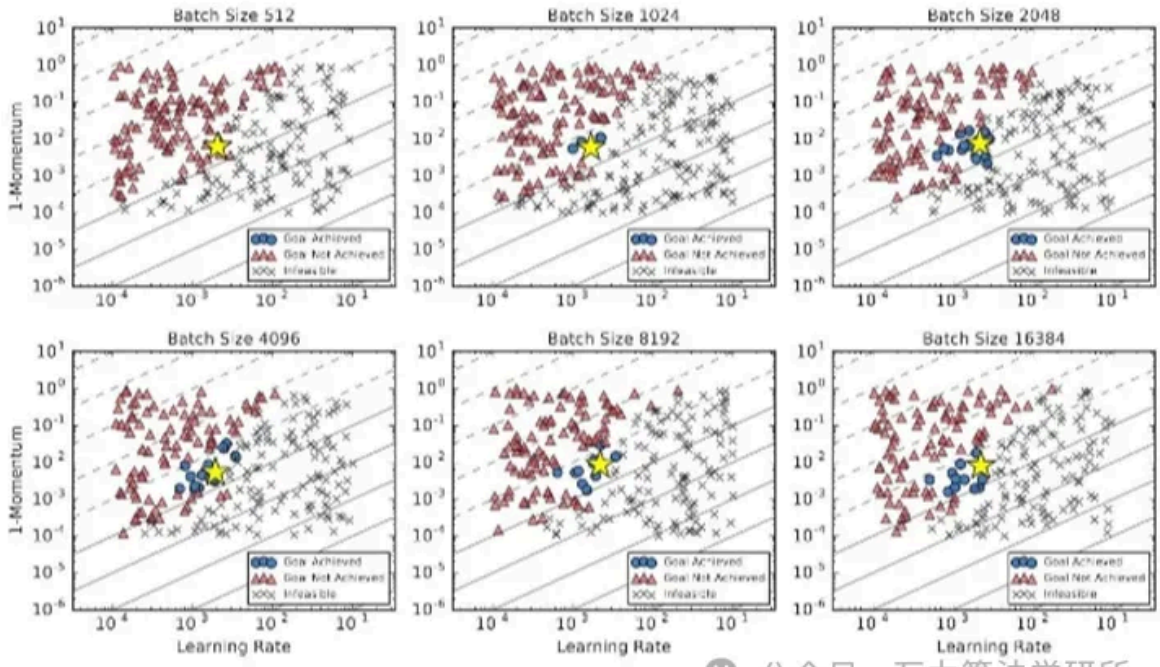
上图，可以观察到更大的 batch size 通常对应较少的训练 step，但相应地需要增加处理的数据。

当 batch size 从 2048 翻倍时，达到同样性能所需要的 step 几乎没有任何改善，但需要花费两倍的计算资源。

再给一张谷歌的实验图：



(a) Transformer on LM1B with a training budget of one epoch.



(b) Transformer on LM1B with a training budget of 25,000 steps.

Google 的经验研究也有类似的观察，即在固定的 epoch budget 下，当 batch size 达到临界值时，模型的性能会 batch size 的增加而降低。

微调时优化器怎么设置好?

在微调 LLM 时，优化器的选择不是影响结果的主要因素。无论是 AdamW、具有调度器 scheduler 的 SGD ，还是具有 scheduler 的 AdamW，对结果的影响都微乎其微。

也可以考虑一个目前比较新的优化器**Sophia**：使用梯度曲率而非方差进行归一化，提高训练效率和模型性能。

预训练和微调时选择的训练数据分别有什么偏好

这里推荐大家先去看一篇短文NLP大语言模型设计的思考笔记（二），对大模型架构设计有一个更深入的了解。

下面简单做个总结。

1. 预训练阶段

首要选择：书籍、论文

数据质量较高，领域相关性比较强，知识覆盖率（密度）较大，此外文本序列较长也有利于增强模型推理能力。

其他：相关领域网站内容、新闻

2. 微调阶段

此阶段数据的质量重要性大于数据的数量的重要性，因此有以下建议：

1）选取的训练数据要干净、并具有代表性。

2）构建的prompt尽量多样化，提高模型的鲁棒性。

3）进行多任务同时进行训练的时候，要尽量使各个任务的数据量平衡

本系列将会持续更新，想要获取面经资料的同学欢迎关注公众号，进群一起交流~