羿阁 编译整理

量子位 | 公众号 QbitAI

Batch大小不一定是2的n次幂？

是否选择2的n次幂在运行速度上竟然也相差无几？

有没有感觉常识被颠覆？

这是威斯康星大学麦迪逊分校助理教授Sebastian Raschka（以下简称R教授）的最新结论。

在神经网络训练中，2的n次幂作为Batch大小已经成为一个标准惯例，即64、128、256、512、1024等。

一直有种说法，是这样有助于提高训练效率。

但R教授做了一番研究之后，发现并非如此。

在介绍他的试验方法之前，首先来回顾一下这个惯例究竟是怎么来的？

2的n次幂从何而来？

一个可能的答案是：因为CPU和GPU的内存架构都是由2的n次幂构成的。

或者更准确地说，根据内存对齐规则，cpu在读取内存时是一块一块进行读取的，块的大小可以是2，4，8，16（总之是2的倍数）。

因此，选取2的n次幂作为batch大小，主要是为了将一个或多个批次整齐地安装在一个页面上，以帮助GPU并行处理。

其次，矩阵乘法和GPU计算效率之间也存在一定的联系。

假设我们在矩阵之间有以下矩阵乘法A和B：

当A的行数等于B的列数的时候，两个矩阵才能相乘。

其实就是矩阵A的第一行每个元素分别与B的第一列相乘再求和，得到C矩阵的第一个数，然后A矩阵的第一行再与B矩阵的第二列相乘，得到第二个数，然后是A矩阵的第二行与B矩阵的第一列……

因此，如上图所示，我们拥有2×M×N×K个每秒浮点运算次数（FLOPS）。

现在，如果我们使用带有Tensor
Cores的GPU，例如V100时，当矩阵尺
寸（M，N以及K）
与16字节的倍数对齐，在FP16混合精度训练中，8的倍数的运算效率最为理想。

因此，假设在理论上，batch大小为8倍数时，对于具有Tensor Cores和FP16混合精度训练的GPU最有效，那么让我们调查一下这一说法在实践中是否也成立。

不用2的n次幂也不影响速度

为了了解不同的batch数值对训练速度的影响，R教授在CIFAR-10上运行了一个简单的基准测试训练——MobileNetV3（大）
——图像的大小为224×224，以便达到适当的GPU利用率。

R教授用16位自动混合精度训练在V100卡上运行训练，该训练能更高效地使用GPU的Tensor Cores。

如果你想自己运行，该代码可在此GitHub存储库中找到（链接附在文末）。

该测试共分为以下三部分：

小批量训练

从上图可以看出，以
样本数量128为参考点，将样本数量减少
1（127）或增加1（129）
，的确会导致训练速度略慢，但这种差异几乎可以忽略不计。

而将样本数量减少28（100）
会导
致训练速
度明显放缓，这可
能是因为模型现在需要处理的批次比以前更多（50,000/100=500与50,000/128=390）。

同样的原理，当我们将样本数量增加28（156）时，运行速度明显变快了。

最大批量训练

鉴于MobileNetV3架构和输入映像大小，上一轮中样本数量相对较小，因此GPU利用率约为70%。

为了调查GPU满载时的训练速度，本轮把样本数量增加到512，使GPU的计算利用率接近100%。

△由于GPU内存限制，无法使用大于515的样本数量

可以看出，跟上一轮结果一样，不管样本数量是否是2的n次幂，训练速度的差异几乎可以忽略不计。

多GPU训练

基于前两轮测试评估的都是单个GPU的训练性能，而如今多个GPU上的深度神经网络训练更常见。为此，这轮进行的是多GPU培训。

正如我们看到的，2的n次幂（256）的运行速度并不比255差太多。

测试注意事项

在上述3个基准测试中，需要特别声明的是：

所有基准测试的每个设置都只运行过一次，理想情况下当然是重复运行次数越多越好，最好还能生成平均和标准偏差，但这并不会影响到上述结论。

此外，虽然R教授是在同一台机器上运行的所有基准测试，但两次运营之间没有特意相隔很长时间，因此，这可能意味着前后两次运行之间的GPU基本温度可能不同，并可能稍微影响到运算时间。

结论

可以看出，选择2的n次幂或8的倍数作为batch大小在实践中不会产生明显差异。

然而，由于在实际使用中已成为约定俗成，选择2的n次幂作为batch大小，的确可以帮助运算更简单并且易于管理。

此外，如果你有兴趣发表学术研究论文，选择2的n次幂将使你的论文看上去不那么主观。

尽管如此，R教授仍然认为，batch的最佳大小在很大程度上取决于神经网络架构和损失函数。

例如，在最近使用相同ResNet架构的研究项目中，他发现batch的最佳大小可以在16到256之间，具体取决于损失函数。

因此，R教授建议始终把调整batch大小，作为超参数优化的一部分。

但是，如果你由于内存限制而无法使用512作为batch大小，那么则不必降到256，首先考虑500即可。

作者Sebastian Raschka

Sebastian Raschka，是一名机器学习和 AI 研究员。

他在UW-Madison（威斯康星大学麦迪逊分校）
担任统计学助理教授，专注于深度学习和机器学习研究，同时也是Lightning AI的首席 AI 教育家。

另外他还写过一系列用Python和Scikit-learn做机器学习的教材。

基准测试代码链接：
https://github.com/rasbt/b3-basic-batchsize-benchmark
参考链接：
https://sebastianraschka.com/blog/2022/batch-size-2.html

— 完 —

量子位 QbitAI · 头条号签约

关注我们，第一时间获知前沿科技动态