Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour

Wang Jian

2020 年 11 月 14 日

目录

1 Background **1**

2 Motivation **1**

3 Main Work **2**

4 Innovation **3**

5 Summary **3**

1 Background

深度学习训练的时间开销

2 Motivation

随着 mini-batch size 的增加，训练的准确率会受到影响，在 64-8k 还是平稳的。

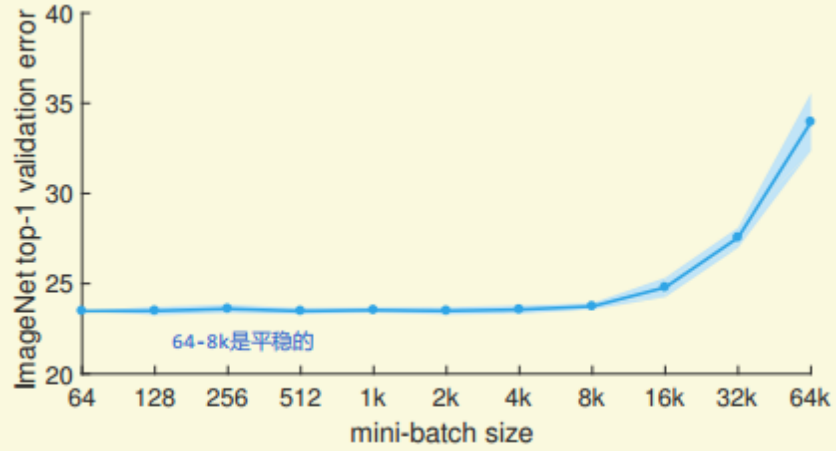

Figure 1. **ImageNet top-1 validation error *vs.* minibatch size.**

如何在保证准确率的基础上进行分布式训练?

3 Main Work

- 使用 Linear Scaling Rule 适应不同 batch size 的学习率
- 采用 warm-up 方法逐渐提高学习率
- Batch Normalization
  - BN 打破每个样本 loss 的独立性，这里的独立性不利于训练。使用 BN 后，单个样本的 loss 依赖所在 batch 的其他样本
  - 各个 batch 可以看做独立的样本，保持 n 不变，那么训练集中独立的样本个数不变。因此只改变 batch size，loss 不会受影响
  - 在整个训练集中，每个 batch 是独立的，假设总共有 n 个 batch，如果 n 发生变化，则训练集的 loss 也会受到影响
  - 但是在分布式系统，单机为 n，则整个系统为 kn，kn 可以看作 k 个 batch，因此分布式不会对训练集的 loss 产生影响
  - 把 n 看做 BN 的超参数，只要固定 n，则分布式训练就不会对系统的 loss 产生影响，文中 n=32

  - 固定 n，通过改变 k 实现 mini-batch size 的改变，通过这种方法，本文实现了分布式上训练达到与单机相同的训练效果

- Communication
  - binary blocks algorithm, 梯度的 aggregation

4 Innovation

在保证准确率的前提下进行分布式训练，对 batch size 进行探讨。

5 Summary

- 行文十分流畅，从中学习到了分布式深度学习一些通用细节