ImageNet training in minutes

Wang Jian

2020 年 11 月 12 日

目录

1 Background

- 如果我们能够重复利用计算机的算力，那么训练能够在几秒钟完成
- 在分布式系统中，增大 batch size k 倍, 每个 epoch 迭代次数缩小 k 倍，训练越快，通信开销降低 k 倍，但是可能带来准确率的降低。如何增大 batch size? warmup, LARS
- PS 模型
- Model Parallelism vs Data Parallelism

2 Main Work

2.1 Synchronous SGD

- Synchronous SGD 主要优点：sequential consistency

2.2 warm-up scheme

- Linear Scaling: 增大 batch size k 倍，学习率应该增大 k 倍
- 从较小的学习率开始逐渐增加

2.3 Layer-wise Adaptive Rate Scaling(LARS)

- 在使用 Synchronous SGD 时，通过 LARS 帮助选择学习率

2.4 Summary

- 使用 LARS 和 warmup scheme 算法来扩展 batch size
- Using LARS we efficiently utilized 1024 CPUs to finish the 100-epoch ImageNet training with AlexNet in 11 minutes with 58.6% accuracy (batch size = 32K),
- We got 74.9% top-1 test accuracy in 64 epochs, which only needs 14 minutes.

3 Innovation

- 使用 LARS 和 warmup scheme 算法来扩展 batch size

4 comment

本文从 batch size 对分布式系统的影响入手,提出使用 LARS 和 warmup scheme 算法来扩展 batch size. 重点开始需要有设备才能做出来。