NVIDIA 新架构 GPU 为机器学习应用带来的性能提升的研究与评估

毕业设计答辩

刘子汉
10152130243@stu.ecnu.edu.cn

华东师范大学
计算机科学与软件工程学院
计算机科学与技术系

2019.05.15

大纲

简介

- 2017Q3，NVIDIA 发布新架构 GPU Tesla V100 及其中的张量核心，且宣称矩阵乘加性能提升达 9.3 倍。
- Stefano 等人的研究中，相同情况下其新架构 GPU 性能提升幅度仅有 4-6 倍，如下图所示。

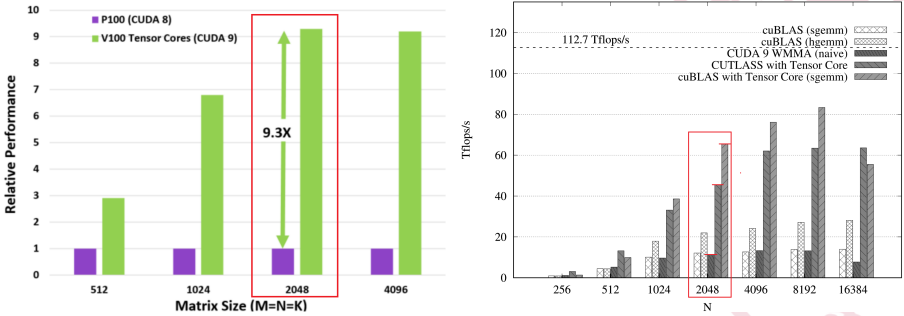

图: 官方白皮书性能与实际研究中性能比较

简介

- 在实际使用框架如 Tensor Flow 搭建的模型中提升幅度更低。在特定结构的网络中开启张量核心仅能带来 60%-80% 的提升。
- **本文将从 Python 源码、CUDA C 源码、PTX 中间代码、SASS 硬件代码的层面，借助卷积神经网络和支持向量机这两种经典的应用，对新架构 GPU 为机器学习应用带来的性能提升进行评估，尝试在代码层面进行优化，并提出设想。**
- 具体评估的应用遵循自底向上的结构：
  - Benchmark 样例 (矩阵乘法、**矩阵乘加**、卷积运算)
  - 基于 CUDA 源码的应用 (卷积神经网络、支持向量机)
  - 基于 Tensor Flow 的应用 (卷积神经网络)
- 除训练过程外，最后使用 TensorRT 以及 Jetson 对部署、推理过程进行优化。

背景

- 机器学习与 GPU：目前绝大部分机器学习应用都需要 GPU 进行加速，而 NVIDIA GPU 长期占据高性能计算的市场。
- NVIDIA GPU 结构：自上而下分为图形处理器簇 (GPC)、纹理处理器簇 (TPC)、流多处理器 (SM)。流多处理器中有若干种处理单元如整数、浮点、逻辑单元。
- 伏特架构/图灵架构：在流处理器中加入了张量核心的新架构，分别对应计算能力 7.0 与 7.5，图灵是消费级芯片，屏蔽了一些硬件。
- 张量核心：专为矩阵乘加设计的硬件，以半精度浮点进行运算 (FP16)，以 wmma 指令批量执行原先整数点积指令与累加指令执行的任务。
- 纹理内存：访问时将二维空间上的周围数据加载进入缓存，其余存储系统为加载一行。
- 线程束：内含 32 个 GPU 线程，作为基本的调度、同步单元。
- TensorRT 与 Jetson：TensorRT 是一个 GPU 推理引擎，用于优化训练完毕的模型，加速推理。Jetson 是 NVIDIA 开发的面向嵌入式应用的芯片。

- GPGPU-SIM：PTX 中间代码执行的软件层面模拟。
- SMart, PerfSIM：SASS 硬件代码执行的软件层面模拟以及 RTL 仿真。
- ThunderSVM：并行支持向量机。
- Leng J.：大型集群的性能、能耗优化。
- Mahmoud K.：访存优化
- ？张量核心

实验平台

表: 实验平台

项目	内容
CPU	AMD Ryzen ThreadRipper 2990WX 32C64T @ 3.0GHz
主板	MSI X399
内存	CORSAIR DDR4 3200 @ 16-15-15-15-34-1T 128GB
GPU	NVIDIA Geforce RTX 2080TI (Turing)
硬盘	INTEL750 NVMe PCIe 1.2TB * 2 @ RAID 0
系统	Windows 10 64-bit build 17763
CUDA	Ver. 10.1, 10.0, 9.2, 9.0
CUTLASS	Ver. 1.2, 1.3
其他	Jetson TX2 $*$

Benchmark:: 矩阵乘加

由于新老架构 GPU 在参数、外围设备等方面均有改进，为了重点研究张量核心的性能，本文中的实验均在 RTX 2080TI 上通过开启/关闭张量核心进行评估。

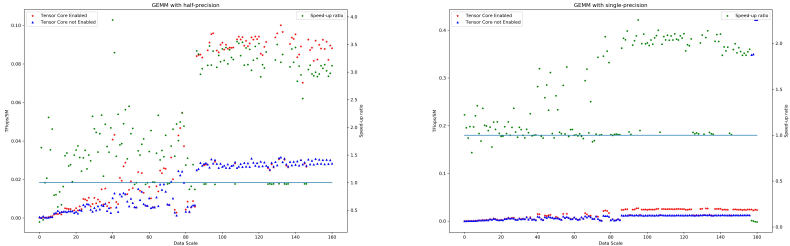

图: 不同计算量下开启和关闭张量核心的性能 (半精度/单精度)

在计算量较大的情况下，开启张量核心后半精度性能提升 3-4 倍，单精度性能提升 2 倍。

Benchmark:: 矩阵乘加

使用 nvprof 和 NSight 进行分析：

表: 开启/关闭张量核心的对比

项目	开启张量核心	关闭张量核心
CUDA 设备同步耗时	**186.15s**	543.51
CUDA 设备同步耗时占比	**79.29%**	91.40%
一次乘加所需计算指令	**一条 wmma**	若干条 idp/idp4a+ 累加指令
每条计算指令延迟	**wmma: 8 时钟周期**	idp/idp4a: 4 时钟周期
上下文切换时间占比	**44.39%**	52.52%

开启张量核心后设备同步、上下文切换等死时间减少，原因为张量核心整合多次计算为一次。

Benchmark:: 矩阵乘加

根据官方文档说明，张量核心对于矩阵裁切形状较为敏感，故将实验结果按加速比排序并按形状特征着色，形状特征分为：能够被 32 整除、能够被 8 整除，无法被整除。

Benchmark:: 矩阵乘法

Benchmark:: 卷积

CUDA C:: 卷积神经网络

CUDA C:: 支持向量机

TensorRT 与 Jetson 优化推理

Tensor Flow-GPU::LeNet-5 卷积神经网络

总结

参考文献


NVIDIA. (2017).
NVIDIA TESLA V100 GPU ARCHITECTURE
NVIDIA Corp., pages 14-15, August, 2017.