北京交通大学
BEIJING JIAOTONG UNIVERSITY

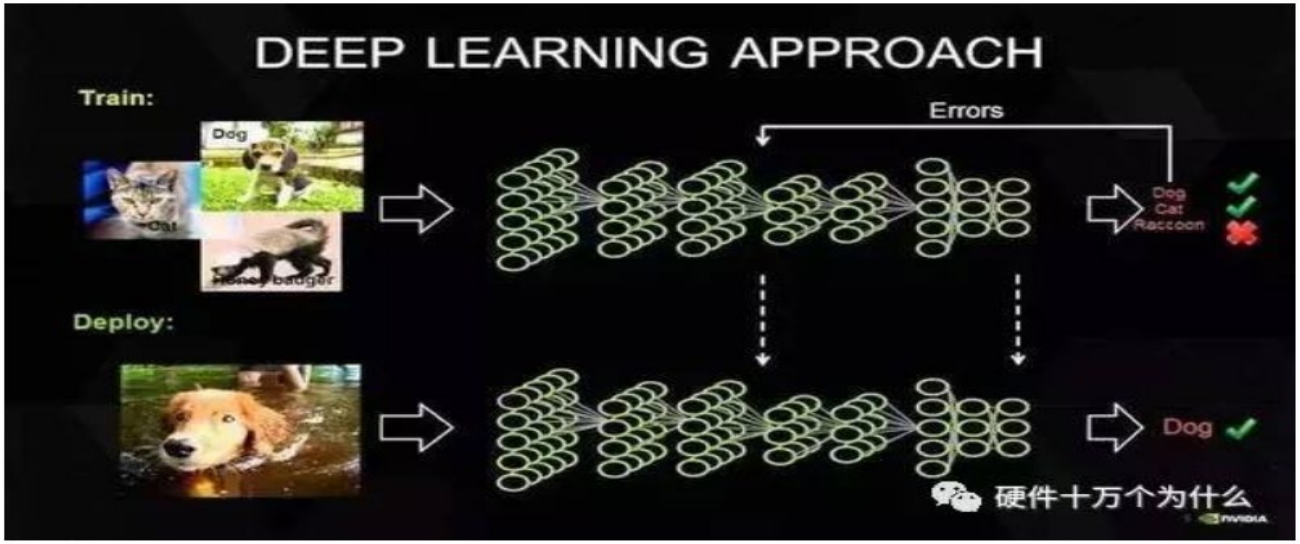
大数据计算处理加速技术

目录

- GPU
- TPU
- FPGA

深度学习的种硬件方案

- 百度的硅谷人工智能实验室（SVAIL）已经为深度学习硬件提出了DeepBench基准，这一基准着重衡量的是基本计算的硬件性能。
- 现在的深度学习算法主要包括卷积神经网络（CNN）和循环神经网络（RNN）。基于这些算法，DeepBench提出以下四种基本运算：



深度学习的种硬件方案

- 矩阵相乘（MatrixMultiplication）——几乎所有的深度学习模型都包含这一运算，它的计算十分密集。
- 卷积（Convolution）——这是另一个常用的运算，占用了模型中大部分的每秒浮点运算（浮点 / 秒）。
- 循环层（RecurrentLayers）——模型中的反馈层，并且基本上是前两个运算的组合。
- AllReduce——这是一个在优化前对学习到的参数进行传递或解析的运算序列。在跨硬件分布的深度学习网络上执行同步优化时（如AlphaGo的例子），这一操作尤其有效。

GPU

图形处理器(graphics processing unit,GPU),又称显示核心、视觉处理器、显示芯片,是一种专门在个人电脑、工作站、游戏机和一些移动设备(如平板电脑、智能手机等)上进行图像运算工作的微处理器。

GPU加速计算是指同时利用图形处理器(GPU)和CPU,加快科学分析、工程、消费和企业应用程序的运行速度。

GPU

GPU加速计算可以提供非凡的应用程序性能,能将应用程序计算密集部分的工作负载转移到GPU,同时仍由CPU运行其余程序代码。理解GPU和CPU之间区别的一种简单方式是比较它们如何处理任务。

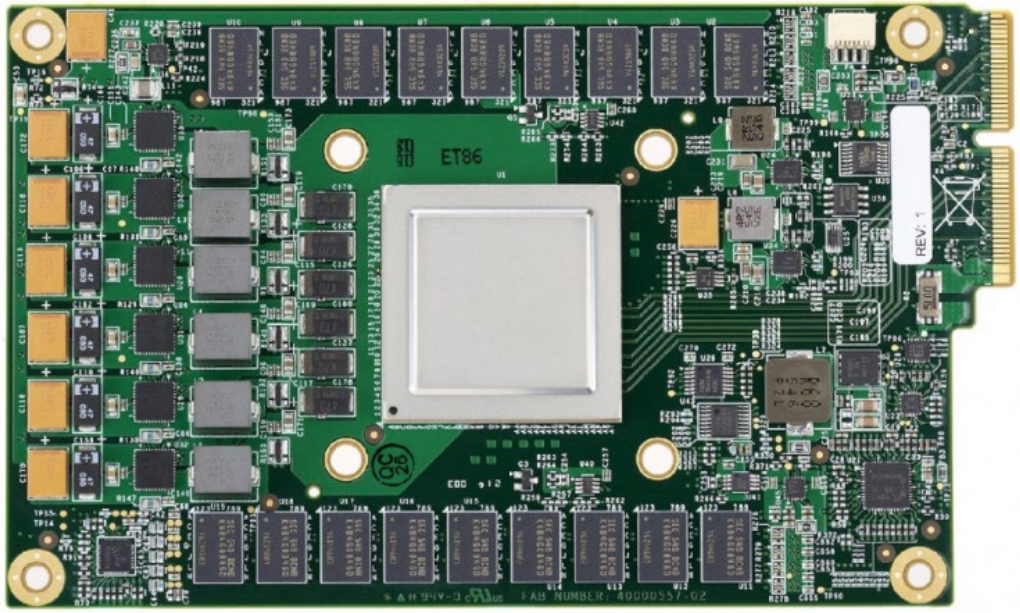
- CPU由专为顺序串行处理而优化的几个核心组成,而GPU则拥有一个由数以千计的更小、更高效的核心(专为同时处理多重任务而设计)组成的大规模并行计算架构。

GPU

- GPU和CPU的浮点计算能力差异的原因是:GPU是特别为计算密集、高并行度计算(如同图像渲染)设计的,因此将更多的晶体管用于数据处理而不是数据缓存和流控。特别地,GPU非常适合处理那些能够表示为数据并行计算(同一程序在多个数据上并行执行)的问题数据并行计算的算术计算密度(算术操作和存储器操作的比例)非常高。由于同一程序在每个元素上执行,因此对复杂流控的要求非常少,更因个元素上执行和高计算密度,访存延迟可以被计算隐藏,因此无数据缓存。

TPU

谷歌资深硬件工程师Norman Jouppi刊文表示，谷歌的专用机器学习芯片TPU处理速度要比GPU和CPU快15-30倍（和TPU对比的是英特尔Haswell CPU以及Nvidia Tesla K80 GPU），而在能效上，TPU更是提升了30到80倍。

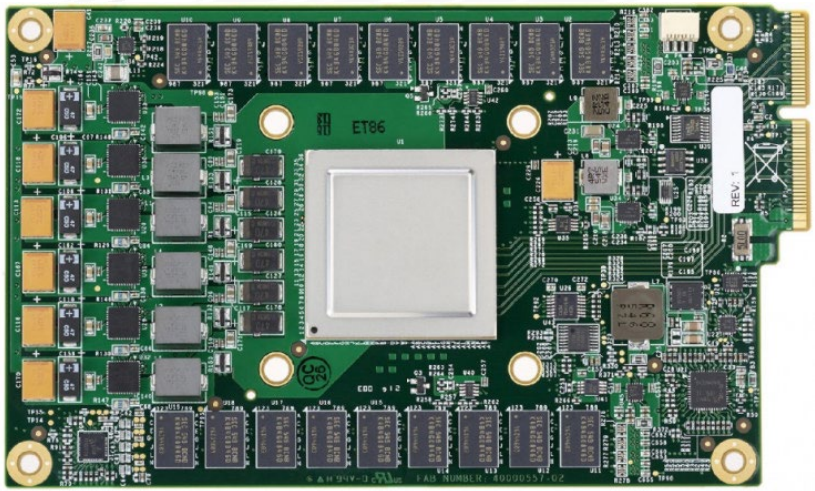


TPU

张量处理器(tensor processing unit,TPU)是为机器学习定制的专用芯片(ASIC),专为深度学习框架 Tensorflow而设计。与图形处理器(GPU)相比,TPU采用低精度(8位)计算,以降低每步操作使用的晶体管数量。降低精度对于深度学习的准确度影响很小,但却可以大幅降低功耗、加快运算速度。同时,TPU使用了脉动阵列的设计,用来优化矩阵乘法与卷积运算,减少I/O操作。此外,TPU还采用了更大的片上内存,以此减少对DRAM的访问,从而更大程度地提升性能。

TPU

与CPU和GPU相比，由于引入了ache、乱序执行、多线程和预取等造成的行时间不确定相比,TPU的确定性执行模块能够满足神经网络应用上99%相应时间需求。CPU/GPU的结构特性对平均吞吐更有效，而TPU针对响应延迟设计。



FPGA

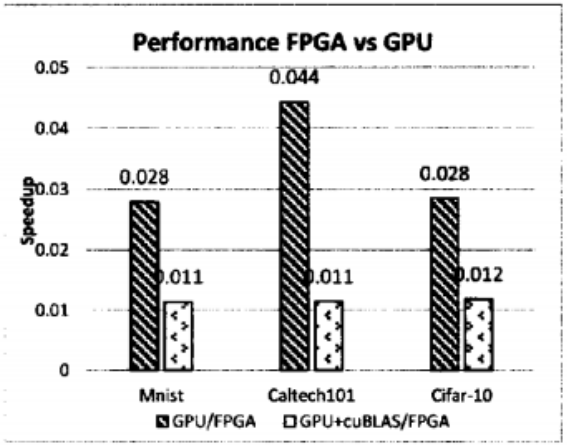
现场可编程逻辑闸阵列(field programmable gate array,FPGA)。是在PAL、GAL、CPLD等可编程逻辑器件的基础上进一步发展的产物。它是作为专用集成电路领域中的一种半定制电路的不足,又克服了原有可编程有限的缺点目前以硬件描述语言(Verilog或VHDL)描述的逻辑电路,可以利用逻辑综合和布局、布线工具软件,快速地烧录至FPGA上进行测试,这一过程是现代集成电路设计验证的技术主流。

FPGA

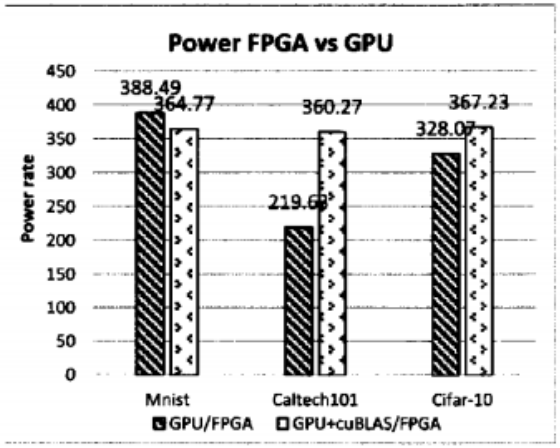
FPGA一般来说比专用集成电路(ASIC)的速度要慢,无法完成更复杂的设计,并且会消耗更多的电能。但是,FPGA具有很多优点,比如可以快速成品,而且其内部逻辑可以被设计者反复修改,从而改正程序中的错误,此外,使用FPGA进行除错的成本较低。

CPU GPU & FPGA

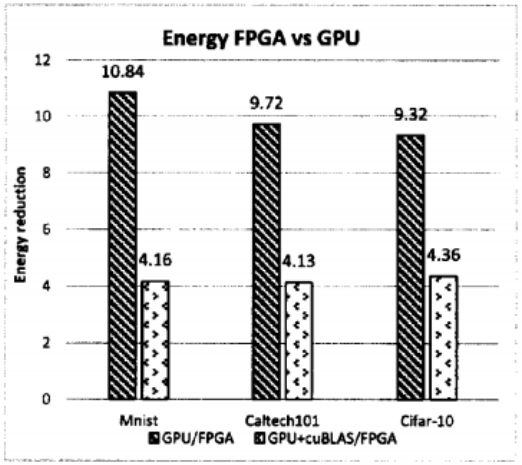
- 以NVIDIA Tesla K40c 为基准，采用6种不同的神经网络结构分别测试了深度学习的预测过程、本地预训练过程和全局训练过程下FPGA(Zedboard)和GPU(K40c)的性能、功率和能耗数据。
- 对深度神经网络下分别使用OpenBLAS和cuBLAS优化对应的CPU程序和GPU程序做对比测试。



（a）性能加速比



（b）功率比



（c）能耗比

CPU GPU & FPGA

- 在异构处理器中,“CPU+GPU”是一个重要选项。GPU采用SIMD(单指令流多数据流)的力式让多个执行单元以同样的步调处理不同的数据,大大提升了并行数据处理的能力,在计算密集型任务中可堪重用。不过GPU有一个“硬伤”,就是延迟比较高。这是因为GPU虽可实现数据并行但是其流水线深度受限,每个计算单元处理不同的数据包时,需要按照统一的步调做相同的事,这就使得输人输出的延迟增加,通常GPU的延迟会达到毫秒级。
- 从数据吞吐能力上看,新一代FPGA的数据处理加速能力理论上已经可以与GPU比肩。同时由于半导体工艺的不断进步,FPGA器件的功率也控制得很好。所以CPU+FPGA这种异构处理器组合被越来越多的人所看好。