解决方案简介

Mipsology

XILINX ALVEO

Mipsology Zebra™：神经网络推断的快速迁移

引言

FPGA：理想适用于推断加速

众所周知，将 AI 推断工作负载从 GPU/CPU 迁移到 FPGA 是一项非常困难、耗时且繁琐的任务。Mipsology 可以使迁移到 FPGA 变得快速、简单，而且成本更低，赋能设计者获益于赛灵思 Alveo™ 加速器卡所提供的高吞吐量和确定性低时延等优势。

- **速度最快的推断**
- **支持所有神经网络**
- **非常简单易用**
- **无需修改代码**
- **可扩展、灵活和自适应**

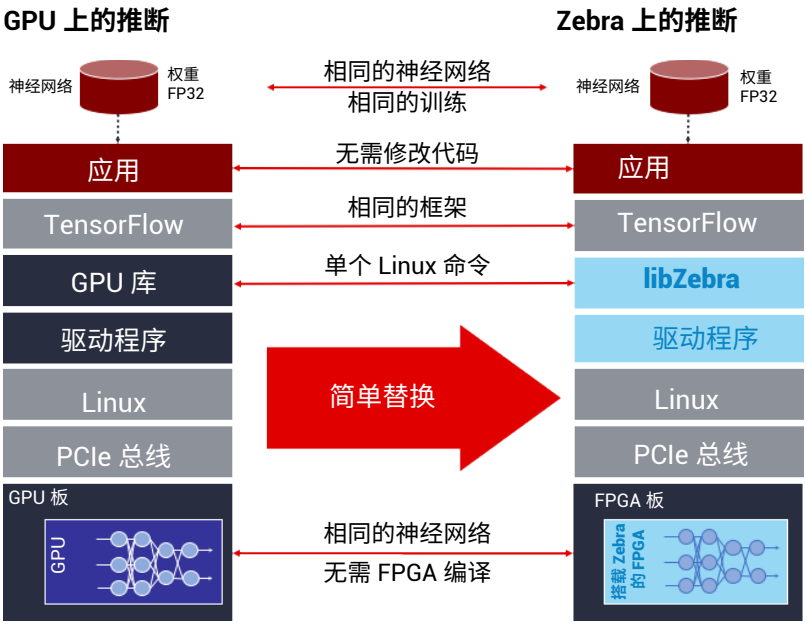
解决方案概览

Zebra 简化了从 CPU/GPU 的迁移。

Mipsology Zebra 是加速卷积神经网络推断的理想计算引擎。Zebra 可无缝替代 CPU/GPU，在 FPGA 上以更快速度计算任何神经网络，同时还能降低功耗与成本。

迁移非常简单，只需输入单个 Linux 命令即可。无需了解 FPGA 技术、编译，也不必对环境或应用进行任何更改。

Zebra 解放了 AI 工程师，使他们能够专注于应用开发，同时获得无与伦比的高性能。



从 GPU/CPU 轻松迁移

- 无需修改神经网络
- 无需新的训练
- 无需添加代码行
- 无需 FPGA 知识
- 无需 FPGA 编译
- 无需进行转换

XILINX

灵活应变.万物智能

解决方案详情

神经网络

- 支持 CNN，无需修改
- 与经过测试的网络一并交付：GoogLeNet V1、Inception V3、Inception V4、VGG16、VGG19、ResNet50、ResNet152、YoloV1、YoloV2、YoloV3、Tiny YoloV2、Tiny YoloV3、VDSR、SSD、MobileNet
- 加速层：卷积、全连接、最大/average pooling、concat、batch norm、scale、add eltwise、reorg、up sampling、depth to space、reduce mean、dilated convolution、squeeze、separable depth wise、clip to value、relu、leaky relu、relu6、sigmoid……
- 自动分割图形
- 最多 32 亿权重
- 最多 100 万层
- 卷积数量不限
- 单个或多个输出
- 最多 1360×1360×3 输入图像
- 最多 24 个独立用户同时运行

支持的框架

- TensorFlow、PyTorch、ONNX、Caffe、MXNet
- 无需修改源代码

精度

- 8 位
- 自动专有量化

从 GPU/CPU 迁移

- 无需修改从 GPU/CPU 训练得到的训练参数
- 无需专有训练或再训练，也无需修剪
- 立即可用
- 准确性与 FP32 相当

功耗和散热

- 从现场的几瓦到数据中心的 140W

结果

Neural Network	Alveo 加速器上的性能					
	大批量			非批量		
	U250	U200	U50_LV	U250	U200	U50_LV
ResNet50[1]	5078	3195	1755	3285	1940	1157
ResNet152[1]	1907	1172	642	1528	892	527
InceptionV3[1]	2452	1463	877	2048	1194	738
InceptionV4[1]	1286	795	450	1154	698	405
VGG16[1]	850	581	369	556	342	219
Yolo-V2[2]	588	411	244	585	411	243
Yolo-V3[3]	254	180	110	253	178	110

注：
性能检测单位为 FPS（每秒帧数）
1. ImageNet 数据集
2. Coco 数据集

建议后续步骤 > 了解有关 Alveo 加速卡的更多信息。
了解有关 Mipsology 的更多信息，请访问 www.mipsology.com
联系 Mipsology 销售