解决方案简介

DeepAI：面向边缘的集成式 AI 训练与推断

引言

AI 训练会占用大量计算资源，通常要依赖昂贵而耗电较高的 GPU。因此，深度学习通常在云端或本地大型数据中心上进行。训练新模型往往耗时数日乃至数周才能完成，而推断查询也由于往返云端的高时延而效果不理想。

DeepAI 的解决方案基于赛灵思 Alveo U50 加速器卡提供集成式 ML 训练和推断解决方案，从而有效应对上述挑战。该解决方案部署在边缘，能提供高吞吐量性能，实现低时延。

解决方案概览

升级训练模型和推断查询所需的数据主要在边缘生成，这包括商店、工厂、终端、写字楼、医院、城市设施、5G 单元站点、车辆、农场、家庭和手持移动设备等。迅速增长的数据在云端或数据中心之间相互传输，可能导致网络带宽难以为继，成本高昂，响应迟缓，此外也会影响个人数据隐私和安全性，并降低设备自动化水平和应用可靠性。

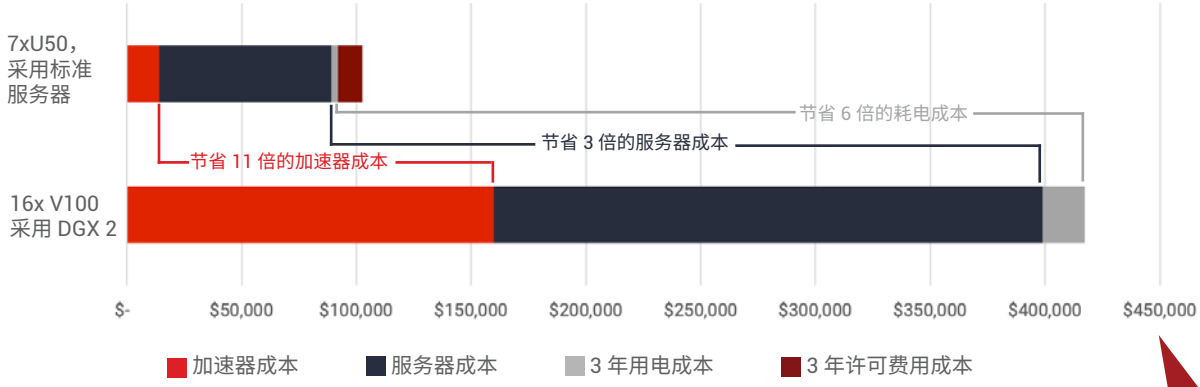
DeepAI 独特而高效的边缘深度学习解决方案，支持集成式 ML 训练和推断工作负载 Alveo U50 上，不再需要去云端或数据中心通过高端 GPU 进行再训练。

DeepAI 训练解决方案的性能达到 NVIDIA V100 GPU 的 2 倍，功耗却仅为其三分之一，成本仅为其四分之一。由于训练在边缘进行，因此客户能使用同一个系统并行开展训练和在线推断。

特性和优势

- 8 位定点量化训练
- 高稀疏比训练
- 同一款硬件加速器支持训练和推断使用
- 比 Nvidia v100 训练快 2 倍
- 比 Nvidia v100 总拥有成本低 4 倍
- 训练输出支持推断
- 训练和推断之间无缝转换
- 可扩展、安全可靠

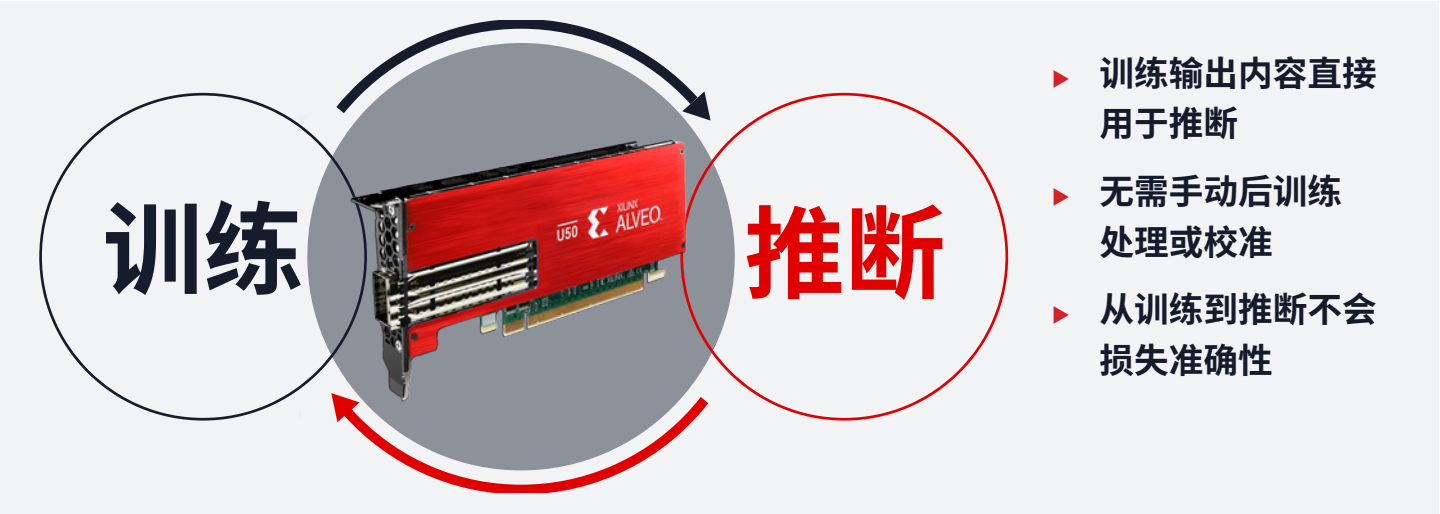
DeepAI 与 Nvidia V100 总拥有成本对比分析



解决方案详情

DeepAI 的解决方案通过赛灵思 Alveo U50 数据中心加速器卡运行。该加速器卡可用于推断和深度学习模型的再训练，能根据不断生成的新数据持续迭代更新模型。

DeepAI 软件解决方案能够确保底层 Alveo U50 加速器对设计 AI 应用的数据科学家和开发者完全透明。



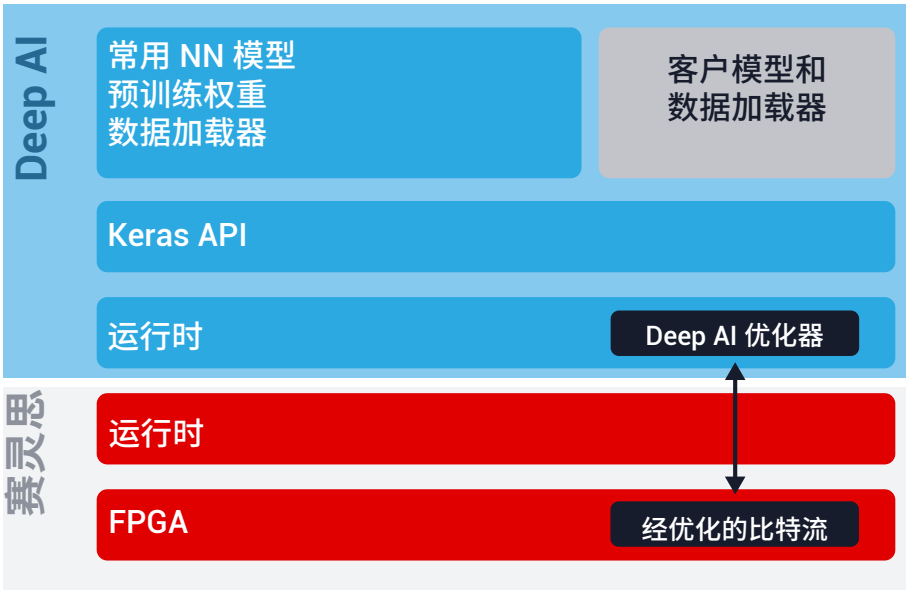
- 训练输出内容直接用于推断
- 无需手动后训练处理或校准
- 从训练到推断不会损失准确性

支持的框架

- Tensorflow、PyTorch、Keras

支持的神经网络

- CNN 用于图像应用常用的神经层、卷积、最大/ average pooling、residual shortcut、batch norm 等操作
- Resnet 和 Mobilenet 用于分类
- Yolo、TinyYolo 和 SSD 用于对象检测
- 多层感知 (MLP)



注：
性能基于用 ImageNet 数据集训练 ResNet50
V100 训练性能 = 360 图像/秒，FP32，基准为 AWS
U50 训练性能 = 800 图像/秒
假定标准服务器最多能支持 8 个 U50 卡
假定所有加速器为公平市场定价
标准服务器使用匹配硬件（CPU、内存、存储等）配置为 DGX2

建议后续步骤 > 如欲了解更多信息或申请演示，请访问：www.deep-aitech.com
或联系 DeepAI 销售

XILINX

灵活应变.万物智能