0610 AI 开发神器来了！

AmusiCVer5 月 17 日

CVer

一个专注侃侃计算机视觉方向的公众号。计算机视觉、图像处理、机器学习、深度学习、C/C++、Python、诗和远方等。

198 篇原创内容

公众号

AI 开发现状

从过去 AlphaGo 在职业围棋中击败世界冠军，到现在大火的自动驾驶，人工智能(AI)在过去几年中取得了许多成就。其中人工智能的成功离不开三要素：**数据、算法和算力。**其中对于算力， 除了训练(train)，AI 实际需要运行在硬件上，也需要推理(inference)，这些都需要强大算力的支撑。

AI 训练硬件平台：GPU、CPU、TPU

常见的模型训练硬件平台主要有：**GPU、CPU 和 TPU。**

- **CPU（Central Processing Unit）** 具有最佳的可编程性，因此，它们为 RNN 实现了最高的 FLOPS 利用率，并且由于内存容量大而支持最大的模型；
- **GPU（Graphical Processing Unit）** 对于不规则计算（例如小批量和 nonMatMul 计算）显示出更好的灵活性和可编程性；
- **TPU（Tensor Processing Unit）** 已针对大型批次和 CNN 进行了高度优化，并且具有最高的训练能力。

截止目前为止，GPU 是 AI 模型训练的硬件主力军，特别是以 NVIDIA 家的 GPU 为主。

AI 推理引擎/部署工具：OpenVINO、TensorRT、Mediapipe

当模型训练结束后，需要对算法模型进行上线部署。这个过程可能会遇到各种问题，比如，模型性能（大小、精度、速度）是否满足线上要求，这些问题都决定着投入产出比。

有许多常见的模型推理部署框架，例如**英特尔的 OpenVINO，NVIDIA 的 TensorRT 和 Google 的 Mediapipe。**

OpenVINO 介绍

OpenVINO 是英特尔针对自家硬件平台开发的一套深度学习工具库，包含推理库，模型优化等等一系列与深度学习模型部署相关的功能。

OpenVINO 是一个 Pipeline 工具集，同时可以兼容各种开源框架训练好的模型，拥有算法模型上线部署的各种能力，只要掌握了该工具，你可以轻松的将预训练模型在 Intel 的 CPU、VPU 等设备上快速部署起来。

TensorRT 介绍

TensorRT 是一个高性能的深度学习推理优化器，可以为深度学习应用提供低延迟，高吞吐率的部署推理。TensorRT 可用于对超大规模数据中心，嵌入式平台或自动驾驶平台进行推理加速。

MediaPipe 介绍

MediaPipe 是个基于图形的跨平台框架，用于构建多模态应用的机器学习 pipiline。MediaPipe 可在移动设备，工作站和服务器上跨平台运行，并支持移动 GPU 加速。使用 MediaPipe，可以将应用的机器学习 pipeline 构建为模块化组件的图形。

上述中，**只有 OpenVINO 具有专属为 CPU 优化的特质**， 提供的 Demo 和 Samples 都很充足，上手比较容易，可以用来快速部署开发，在英特尔的硬件平台上性能超过了大部分开源库，因此本文将重点对 OpenVINO 进行介绍。

OpenVINO

官网：**https://docs.openvinotoolkit.org**

OpenVINO 概述

OpenVINO（Open Visual Inference & Neural Network Optimization，开放视觉推理及神经网络优化）是英特尔基于自身现有的硬件平台开发的一种可以加快高性能计算机视觉和深度学习视觉应用开发速度工具套件，支持各种英特尔平台的硬件加速器上进行深度学习，并且允许直接异构执行。支持在 Windows、Linux、macOS 系统上运行，也支持 Python / C++ 语言。

OpenVINO™ 工具套件：

- 在边界上启用基于卷积神经网络的深度学习推理；

- 支持跨英特尔® CPU、英特尔® 集成显卡、英特尔® 神经电脑棒 2 和搭载英特尔® Movidius™ 视觉处理器的英特尔® Vision Accelerator Design 的异构执行；
- 通过一套易用的计算机视觉功能库和预优化内核库来加速上线时间；
- 包括了针对计算机视觉标准进行优化的调用，包括 OpenCV 和 OpenCL。

以下图表显示了典型的 OpenVINO™ 工作流程：

模型准备，转换和优化

你可以使用你选择的框架来准备和训练深度学习模型，或者从 Open Model Zoo 下载预训练模型。Open Model Zoo 包含针对各种视觉问题的深度学习解决方案，如物体识别、人脸识别、姿态估计、文本检测和动作识别等。

OpenVINO™ 工具套件的一个核心组件是**模型优化器 (Model Optimizer)**，它是一个跨平台命令行工具，可将经过训练的神经网络从源框架转换为与 nGraph 兼容的开源中间表示 (IR)，用于推理运算。

模型优化器导入在 PyTorch、Caffe、TensorFlow、MXNet 和 ONNX 等常用框架中经过训练的模型，并执行几项优化，以尽可能删除过多的层和群运算，以更简单、更快速地形成图表。

推理运行和调优推理

OpenVINO™ 的另一个核心组件是**推理引擎（Inference Engine）**，它管理经过优化的神经网络模型的加载和编译，在输入数据上运行推理运算，并输出结果。推理引擎可以同步或异步执行，其插件架构管理用于在多个英特尔® 设备上执行的适当编译，包括主力 CPU 以及专用显卡和视频处理平台）。

你可以将 OpenVINO™ 调整实用程序与推理引擎一起使用，在模型上试用和测试推理。基准测试实用程序使用输入模型运行迭代测试，以检测吞吐量或延迟，交叉检查实用程序对配置不同的推理的性能进行比较。训练后优化工具集成了一套基于量化和校准的工具，以进一步简化性能。

封装和部署

英特尔® Distribution of OpenVINO™ 工具套件为以下设备输出经过优化的推理运行时：

- 英特尔® CPU
- 英特尔® Processor Graphics
- 英特尔® 神经电脑棒 2
- 采用英特尔® Movidius™ 视觉处理器的英特尔® Vision Accelerator Design

Open Model Zoo

官网：**https://github.com/openvinotoolkit/open_model_zoo**

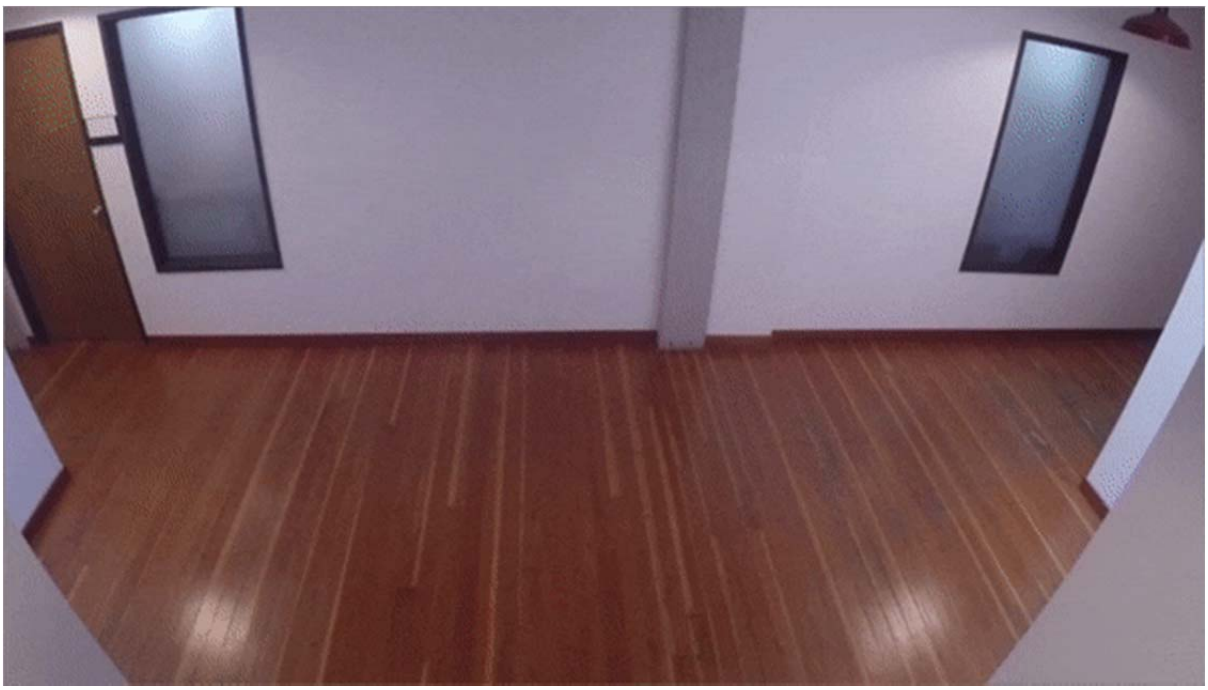
Open Model Zoo 包括丰富的优化后的上百种深度学习模型和一系列演示，有助于加快高性能深度学习推理应用程序的开发。使用这些免费的预训练模型，可以加快开发和生产部署过程。

官方提供的预训练模型包含：目标检测、物体识别、重识别、语义分割、实例分割、人体姿态估计、文本检测、文本识别、行为识别、图像检索、机器翻译等任务。其中目标检测模型：Faster R-CNN、YOLOv2、YOLOv3、SSD 等。

目标检测 Demo 演示：

实例分割 Demo 演示：

行人跟踪 Demo 演示：



3D 人体姿态估计 Demo 演示：

OpenVINO 行业应用

OpenVINO 一经推出就得到行业内的普遍认可和支持，目前在工业、医疗、零售等领域广泛应用。

工业（2D、3D 视觉）

创建安全工作空间：通过将推理和深度学习功能扩展到边缘来帮助预防工作场所的危害和传染病的传播。

制造业的视觉审查：借助英特尔®OpenVINO™工具包优化的自动缺陷检测，并在面向边缘的英特尔®DevCloud 上进行了测试。

医疗（成像、分类、分割）

COVID-19 胸部 X 射线肺炎检测：DarwinAI 开发了由 AI 驱动的解决方案 COVID-Net CT，以快速，准确地检测患者的 COVID-19，并使用 OpenVINO 进行了其他优化。

三星自动化超声测量以改善临床工作流程：通过 OpenVINO™优化的胎儿和产妇测量。

新零售

解决零售困境：实时的购物者流量映射可以帮助零售商在瞬息万变的市场中竞争。

未来展望

OpenVINO 是一个比较成熟而且仍在快速发展的推理库， 将计算机视觉等方向与 AI 模型推理更好地融合在一起，加快赋能产业，助力 AI 开发者来设计具有实际应用价值的解决方案。

如果你还没用过 OpenVINO，赶快来尝试吧！

整理不易，请给 CVer 点赞和在看