

OpenVINO™ 与 TensorFlow 集成简介

作者： Arindam, Yamini, Mustafa, Ritesh, Priya, Chandrakant, Surya, Amar, Sesh

翻译： 李翊玮

技术的传播采用通常是由用户体验的飞跃引发的。例如，iPhone 促使智能手机和“应用商店”的快速普及。或者，最近，TensorFlow 中的易用性启动了人工智能的大规模增长，几乎触及了我们今天日常生活的方方面面。

[OpenVINO™ 工具包](#)重新定义了采用英特尔技术的设备上的 AI 推理，并获得了前所未有的开发人员采用。如今，成千上万的开发人员使用 OpenVINO™ 工具包来加速几乎所有可以想象的用例的 AI 推理，从人类视觉仿真，自动语音识别，自然语言处理，推荐系统等等。该工具包基于最新一代的人工神经网络，包括卷积神经网络（CNN）、循环和基于注意力的网络，可跨英特尔®硬件（英特尔 CPU、英特尔集成显卡®、英特尔®神经计算棒 2 和英特尔®视觉加速器设计与英特尔® Movidius™ VPU）扩展计算机视觉和非视觉工作负载，从而最大限度地提高性能。它通过从边缘部署到云的高性能、AI 和深度学习推理来加速应用。

我们很荣幸能与客户合作，为他们的成功做出贡献。我们不断倾听和创新，以满足他们不断变化的需求，同时也致力于提供世界一流的用户体验。因此，根据客户反馈，并在 OpenVINO™ 工具包的成功基础上，我们将 [OpenVINO™ 与 TensorFlow* 集成](#)。这种集成使 TensorFlow 开发人员只需 2 行额外的代码即可在部署中加速其 TensorFlow 模型的推理。

对 TensorFlow 开发人员的好处：

[OpenVINO™ 与 TensorFlow* 的集成](#)提供了增强 TensorFlow 兼容性所需的 [OpenVINO™ 工具包](#) 内联优化和运行时。它专为希望体验使用 OpenVINO™ 工具包的好处的开发人员而设计 - 帮助提高其推理应用程序的性能 - 只需最少的代码修改。它可以加速各种 英特尔® 芯片上 [许多 AI 模型](#) 的推理，例如：

- 英特尔中央®处理器
- 英特尔®集成显卡
- 英特尔® Movidius™ 视觉处理单元 - 称为 VPU
- 采用 8 个英特尔® Movidius™ MyriadX VPC 的英特尔视觉加速器设计 - 称为 VAD-M 或 HDDL

利用此集成的开发人员可以预期以下好处：

- 与本机 TensorFlow 相比，性能加速（取决于底层硬件配置）。
- **精度** – 保持与原始模型几乎相同的精度。
- **简单性** – 继续使用 TensorFlow API 进行推理。无需重构代码。只需导入，启用和设置设备。
- **健壮性** – 旨在支持各种操作系统/Python 环境中的各种 TensorFlow 模型和运算符。
- **无缝的内联模型转换** – 无需显式模型转换。
- **轻量级占用空间** – 所需的增量内存和磁盘占用空间极小。
- **支持广泛的英特尔供电设备** – CPU、iGPU、VPU（Myriad-X）。

注意：为了获得最佳性能、效率、工具定制和硬件控制，我们建议超越此组件，采用本机 OpenVINO™ API 及其运行时。

它是如何工作的？

开发人员可以通过将以下两行代码添加到他们的 Python 代码或 Jupyter Notebooks 中来大大加快 TensorFlow 模型的推理。

1. `import openvino_tensorflow`
2. `openvino_tensorflow.set_backend('<backend_name>')`

支持的后端<backend_name>包括“CPU”，“GPU”，“MYRIAD”和“VAD-M”。参见图 1。

示例代码：

以下是 OpenVINO™ 与 TensorFlow* 集成的示例：

```

1 # Installation steps
2 # more details : https://github.com/openvinotoolkit/openvino\_tensorflow GitHub
3 #pip3 install -U pip
4 #pip3 install -U tensorflow==2.x.x
5 #pip3 install openvino-tensorflow
6
7 # Import package and set backend
8 import openvino_tensorflow Line 1
9 openvino_tensorflow.set_backend('GPU') Line 2
10
11 # Load a TF Saved Model
12 model = tf.keras.models.load_model('resnet50_saved_model')
13
14 # Get the input size of the model
15 network_input_size = saved_model_loaded.input.shape()
16
17 # Resize the input image
18 resized_image = resize(input_image, network_input_size)
19
20 # Run inference
21 model.predict(resized_image)

```

CPU
GPU
MYRIAD
VAD-M

图1

它是如何在引擎盖下真正工作的？

[OpenVINO™ 与 TensorFlow* 的集成](#)通过将 TensorFlow 图有效地划分为多个子图来提供加速的 TensorFlow 性能，然后将这些子图调度到 TensorFlow 运行时或 OpenVINO™ 运行时以实现最佳加速推理。最终组合结果以提供最终的推理结果。

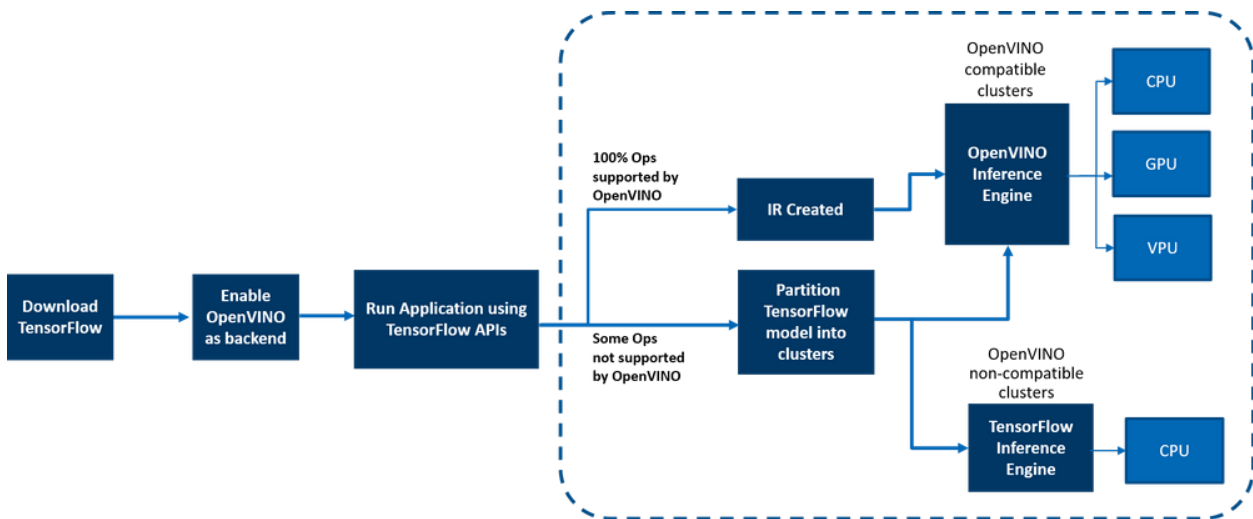


图2: 工作流的端到端概述

[下面是一个详细的体系结构图。](#)

在边缘和云端部署

OpenVINO™ 与 TensorFlow 的集成适用于从云到边缘的各种环境，只要底层硬件是英特尔平台即可。例如，附加组件适用于以下云平台：

- [面向边缘的英特尔® DevCloud](#)
- [AWS Deep Learning AMI Ubuntu 18 和 Ubuntu 20 on EC2 C5 实例，针对推理进行了优化](#)
- [Azure ML](#)
- [谷歌实验室](#)

支持任何基于 AI 的边缘设备。

示例在 [gitrepo](#) 的示例/目录中提供。

这与使用原生 OpenVINO™ 工具包有何不同：

OpenVINO™ 与 TensorFlow* 的集成使 TensorFlow 开发人员能够以非常快速简便的方式加速其 TensorFlow 模型推理 - 只需 2 行代码。OpenVINO™ 模型优化器可加速推理性能，以及丰富的集成开发人员工具和高级功能，但如前所述，为了获得最佳性能，效率，工具定制和硬件控制，我们建议使用本机 OpenVINO™ API 及其运行时。

客户采用

客户正在将 OpenVINO™ 集成用于 TensorFlow 用于各种用例。以下是一些示例

1. [Extreme Vision \(极视角\)](#)：极视角的 CV MART 等专用 AI 云 可帮助数十万开发人员提供丰富的服务、模型和框架目录，从而在各种英特尔平台（如 CPU 和 iGPU）上进一步优化其 AI 工作负载。与 AI 框架（如 OpenVINO™ 与 TensorFlow* 的集成）正确集成的易于使用的开发人员工具包可加速模型，从而提供两全其美的优势 - 提高推理速度以及以最小的更改重用已创建的 AI 推理代码的能力。Extreme Vision 团队正在测试 OpenVINO™ 与 TensorFlow* 的集成，目标是在 Extreme Vision 平台上为 TensorFlow 开发人员提供支持。
2. [由博德研究所开发的基因组分析工具包 \(GATK\)](#) 是世界上使用最广泛的变体调用开源工具包之一。Terra 是一个更安全，可扩展的开源平台，供生物医学研究人员访问数据，运行分析工具和协作。基于云的平台由麻省理工学院博德研究所与哈佛大学，微软和 Verily 共同开发。Terra 平台包括 GATK 工具和管道，供研究界运行其分析。[CNNScoreVariants](#) 是 GATK 中包含的深度学习工具之一，它应用卷积神经网络来过滤带注释的变体。在一篇[博客](#)中，Broad Institute 展示了如何使用 OpenVINO™ 与 TensorFlow* 集成来进一步加速 CNNScoreVariants 的推理性能。

结论

现在，您已经更好地了解了其优势、工作原理、部署环境以及 OpenVINO 与 TensorFlow 的集成与使用原生 OpenVINO API 的不同之处，我们迫不及待地想亲自尝试将 OpenVINO 与

TensorFlow 集成，并开始在所有英特尔平台上体验 AI 模型的推理性能提升。与往常一样，我们很乐意听到您对此集成的反馈，请[通过 OpenVINO-tensorflow@intel.com](mailto:OpenVINO-tensorflow@intel.com) 与我们联系或在 gitrepo 中提出问题。谢谢！

资源

以下资源可帮助您了解更多信息：

- [Github repository](#)
- [解决方案简介](#)
- [动态回退漫画](#)
- [此处提供了详细的常见问题解答。](#)
- [英特尔 DevCloud for the Edge 上的示例笔记本可在此处获取。](#)
- [Google Colab 笔记本示例可在此处获取。](#)