白皮书

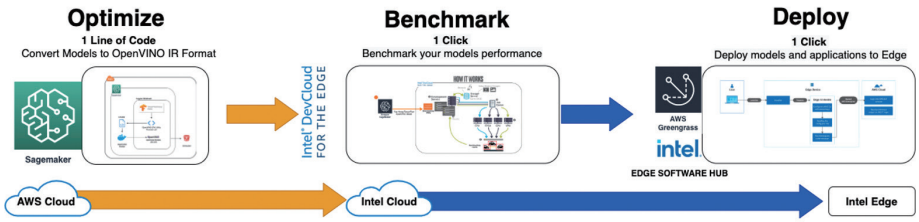
利用 AWS SageMaker 与英特尔® 软硬件技术加快 AI 推理速度的步骤

助力云开发人员无缝开启云端到边缘测的开发旅程

作者

Vibhu Bithar

Devang Aggarwal



为了支持云开发人员从云端到边缘测的旅程，我们构建了多个开发工具以加速开发。我们将在本博文中介绍其中三个开发工具。您可以使用 AWS SageMaker 在 AWS 云中构建和训练模型，然后使用 OpenVINO™ 工具套件模型优化器优化这些模型。优化后，您将能够在英特尔® DevCloud 中，基于各类英特尔® 硬件对模型进行性能指标评测。最后，我们将介绍如何搭建基于英特尔® OpenVINO™ 工具套件分发版和 AWS Greengrass 的边缘环境，以及如何使用 AWS Greengrass Python Lambda 服务，在边缘测部署利用英特尔® OpenVINO™ 工具套件分发版执行的图像分类和对象检测应用。

目录

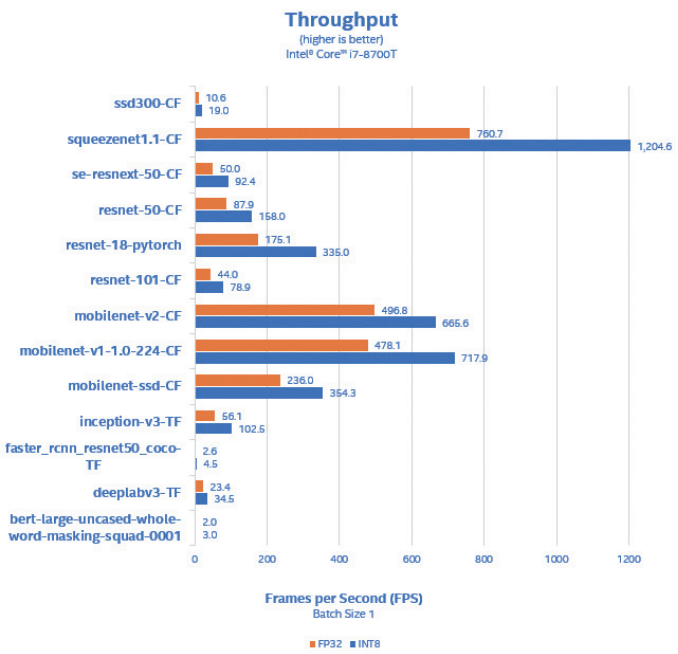
英特尔® OpenVINO™ 工具套件分发版概述

得益于 AI 领域的最新发展，开发人员如今在框架、模型和硬件方面有多种选择。然而，开发人员需要使用正确的硬件及其相关软件，才能够充分利用底层硬件算力，提高 AI 推理性能。英特尔® OpenVINO™ 工具套件分发版就是这样一款加速工具，借助预优化的模型，它可帮助开发人员最大限度提高推理性能。

具体而言，英特尔® OpenVINO™ 工具套件分发版是一款全面的工具套件，支持开发人员快速开发可模拟人类视觉的 AI 应用和解决方案。基于卷积神经网络（CNN），该工具套件可将计算机视觉工作负载扩展至各类英特尔® 硬件，并实现卓越性能。更多信息请访问 OpenVINO™ 工具套件概述。

本文要点

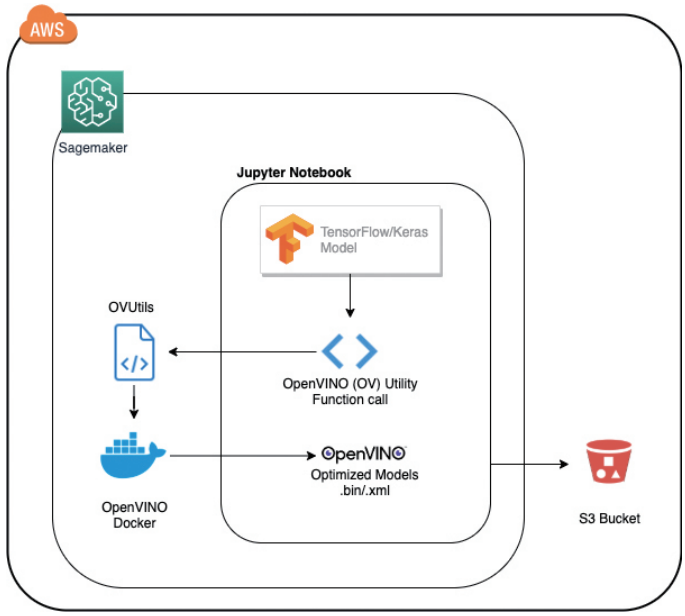
- 了解如何借助 1 行代码，在 AWS SageMaker 中使用 OpenVINO™ 工具套件模型优化器将 TensorFlow 和 Keras 模型转换为 OpenVINO IR 中间表示格式。
- 了解如何在英特尔® DevCloud 中，使用 Jupyter notebook 中的性能评测应用，在多个英特尔硬件上对模型进行一键性能指标评测。
- 了解如何使用英特尔® 边缘软件中心一键将 AI 推理应用和 OpenVINO 模型部署到边缘。



硬件方面，英特尔提供 CPU、VPU 和 FPGA 等可扩展的产品组合，能够充分满足您 AI 推理解决方案的需求。表 1 向我们展示了英特尔® 酷睿® i7 处理器的高性能输出。您可以看到使用英特尔® OpenVINO™ 工具套件分发版后，对于某些模型，可实现高达每秒 1200 帧的性能。更多信息请访问系统配置和更多性能指标评测。

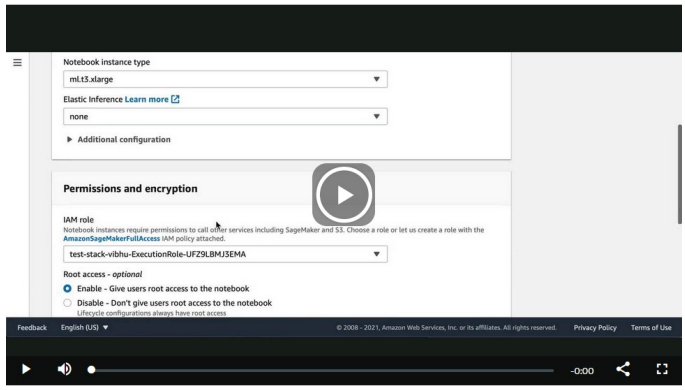
了解如何在 AWS SageMaker 中使用 OpenVINO™ 工具套件模型优化器

现在，我们将介绍如何在 AWS SageMaker 中使用 OpenVINO™ 工具套件模型优化器轻松优化模型。为帮助您轻松进行模型优化，我们开发了 python 函数，该函数简化并实现了内联模型转换。它使用 OpenVINO™ 工具套件 docker 容器来转换 TensorFlow 和 Keras 模型。利用 OpenVINO 中间表示格式 IR，您只需编写一次推理代码，即可使用这些从不同框架转换的 IR 格式模型。为方便起见，我们提供了支持的 TFHub 模型及其 input shape。

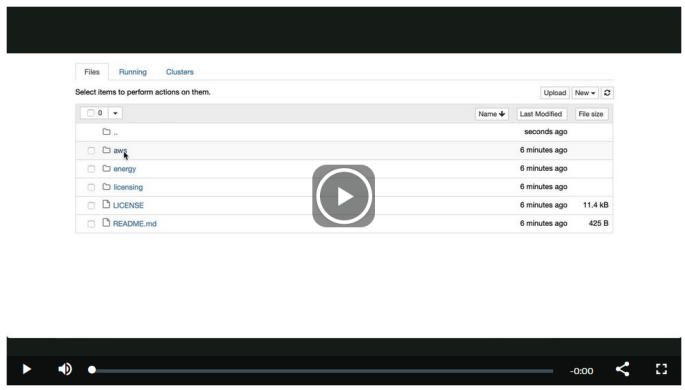


立即开始

1. 创建一个 SageMaker Notebook 并将 Github repo 克隆到您的 SageMaker Notebook 实例



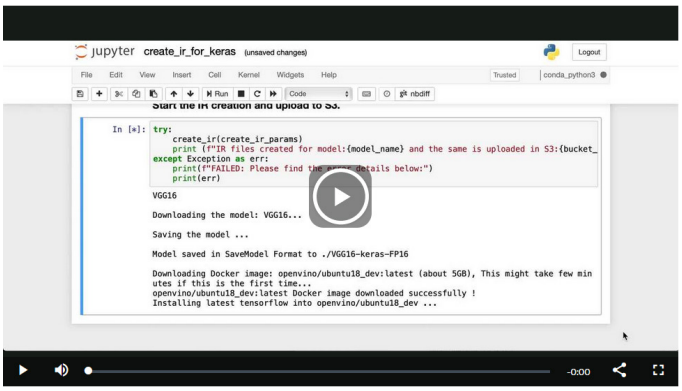
2. 打开 SageMaker Notebook，转向 aws / mo-utility 目录



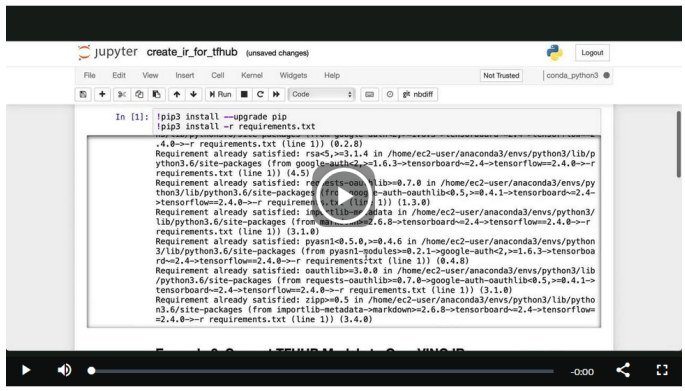
3. 转向 aws / mo-utility 目录后，您将看到以下文件：

文件名	描述
create_ir_for_keras.ipynb	示例 notebook，演示如何将 Keras App 模型转换为 OpenVINO IR 格式
create_ir_for_tfhub.ipynb	示例 notebook，演示如何将 Tensorflow Hub 模型转换为 OpenVINO IR 格式
create_ir_for_obj_det.ipynb	示例 notebook，演示如何将对象检测模型转换为 OpenVINO IR 格式
ov_utils.py	实用程序代码，支持模型转换
TFHub-SupportedModelList.md	来自 Tensorflow Hub、支持 TF1/TF2 的模型和其 input shape 列表
Keras-SupportedModelList.md	支持的 Keras App 模型列表
ObjDet-SupportedModelList.md	支持的对象检测模型列表
TFHub-TF1-SupportedModelList.pdf	来自 Tensorflow Hub、支持 pdf 格式的 TF1 模型和相关 input shape 列表
TFHub-TF2-SupportedModelList.pdf	来自 Tensorflow Hub、支持 pdf 格式的 TF2 模型和相关 input shape 列表
Keras-SupportedModelList.pdf	支持 pdf 格式的 Keras App 模型列表
ObjDet-SupportedModelList.pdf	支持的对象检测模型列表
requirements.txt	带有 pip、安装到 Jupyter Notebook 的 python 库列表
README.md	README file
ov-utils-arch.png	架构图

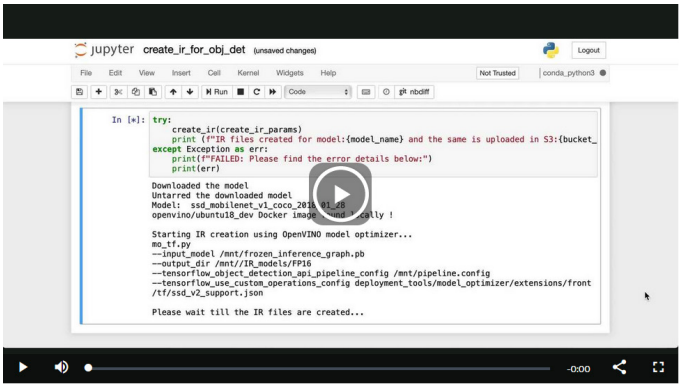
将 Keras App 模型转换为 OpenVINO™ IR 的流程一览



将 Tensorflow Hub 模型转换为 OpenVINO™ IR 的流程一览



将对象检测模型转换为 OpenVINO™ IR 的流程一览



下一步：

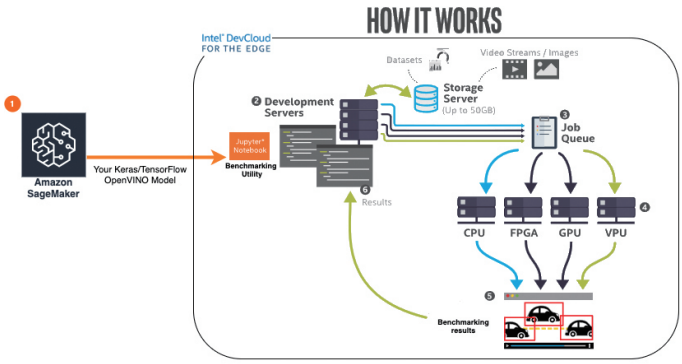
在下一节中，我们将探讨如何使用面向边缘的英特尔® DevCloud 对模型在各类英特尔® 硬件上的性能指标进行评测。

英特尔® DevCloud，一键使用/零成本，评估深度学习模型在各类英特尔® 硬件上的推理性能

您是否想知道您的模型在不同英特尔® 硬件上的性能如何？英特尔提供了设备沙盒英特尔® DevCloud，可帮助您在最新的英特尔® 硬件系列中免费开发、测试和运行您的工作负载。在一站式访问所有最新英特尔® 硬件的同时，您也需要了解哪款英特尔® 硬件最适合您的应用。为此，我们提供英特尔® DevCloud 这一强大的工具，帮助您明确哪款硬件最适合您的深度学习模型。

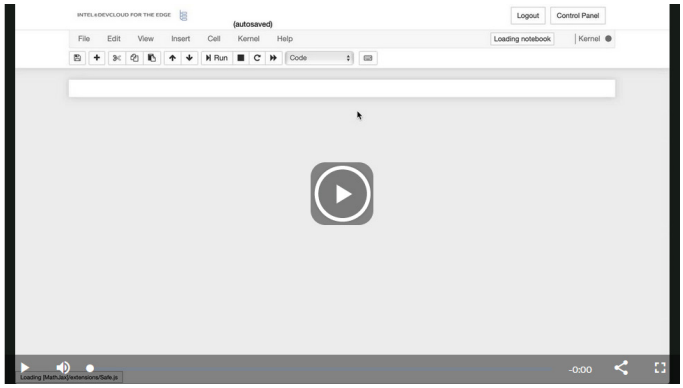
在上一节中，您已经了解了如何将 TensorFlow 和 Keras 图像分类模型，以及 TensorFlow 对象检测模型，转换为 OpenVINO IR 中间表示格式，并将其存储在 S3 存储桶中。

在本节中，您将了解如何直接从 S3 存储桶中获取 OpenVINO IR 模型，并使用提供的 Jupyter notebook 示例，借助英特尔® DevCloud 在不同硬件上对您的模型进行一键式性能指标评测。



想要基于您的模型进行尝试？请查看英特尔® DevCloud 性能指标评测示例。您只需提供 AWS 凭证和 S3 存储桶，我们就会为您从 S3 存储桶中提取模型。

Jupyter Notebook 示例概览



在运行了 Jupyter Notebook 中的所有单元之后，您将通过详细的输出表格，了解到模型在哪种硬件上可实现最佳性能，类似于以下内容：

Model Name: efficientnet-b1-tfhub-FP16

Best Device(Based on Throughput) : MULTI:HDDL,CPU

Buy Now: www.intel.com

Benchmark Results

	Throughput (FPS)	Load network time (ms)	Read network time (ms)	First inference time (ms)	Total execution time (ms)	Total number of iterations	Model Precision
NCS2	17.05	7075.43	400.90	95.74	703.94	12	FP16
Core	72.41	1146.47	324.40	81.77	165.72	12	FP16
GPU	99.05	53456.97	294.74	17.04	121.15	12	FP16
MULTI:CPU,GPU	100.55	49475.29	305.33	88.95	127.92	10	FP16
HDDL-R	95.05	47213.83	286.80	107.48	336.65	32	FP16
MULTI:HDDL,CPU	107.58	48378.02	276.71	103.43	334.62	36	FP16
XeonE3	95.50	1649.48	263.69	90.10	125.66	12	FP16

下一步：

在下一节中，我们将讨论如何使用英特尔® 边缘软件中心部署部署推理应用和 OpenVINO™ 模型。

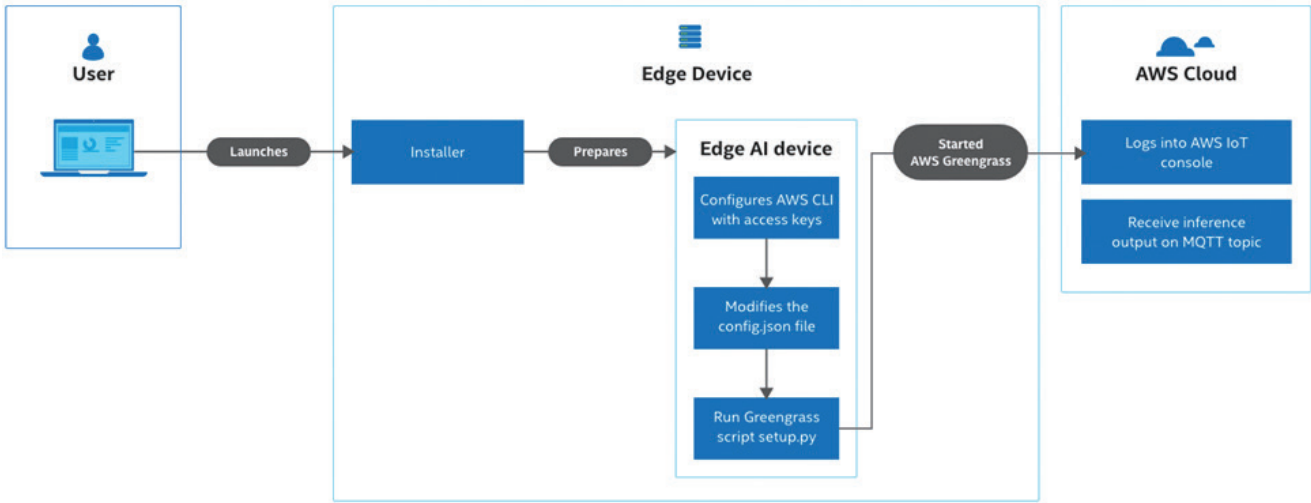
使用英特尔® 边缘软件中心部署推理应用和 OpenVINO™ 模型

现在，您已经在多个英特尔® 硬件上对模型进行了性能指标评测，接下来，您需要在边缘端部署模型，这正是英特尔® 边缘软件中心的用武之地。英特尔® 边缘软件中心能够帮助开发人员更快速、更放心地配置、验证和部署针对特定用例的解决方案。

英特尔的边缘软件中心提供丰富的参考实现与用例，能够为您带来便捷简易的开发体验。Amazon Web Services (AWS) * 云端到边缘管道就是这样的一个用例。该用例可帮助您一键部署及使用从云端到边缘端的推理管道。该管道在边缘测使用 AWS IoT Greengrass 和 OpenVINO™ 工具套件，在云端使用 AWS IoT 服务。通过使用 AWS Greengrass 中包含的功能，您可以将应用部署到多个边缘设备。该用例还包括用于图像分类和对象检测的 AWS IoT Greengrass Lambda 示例。

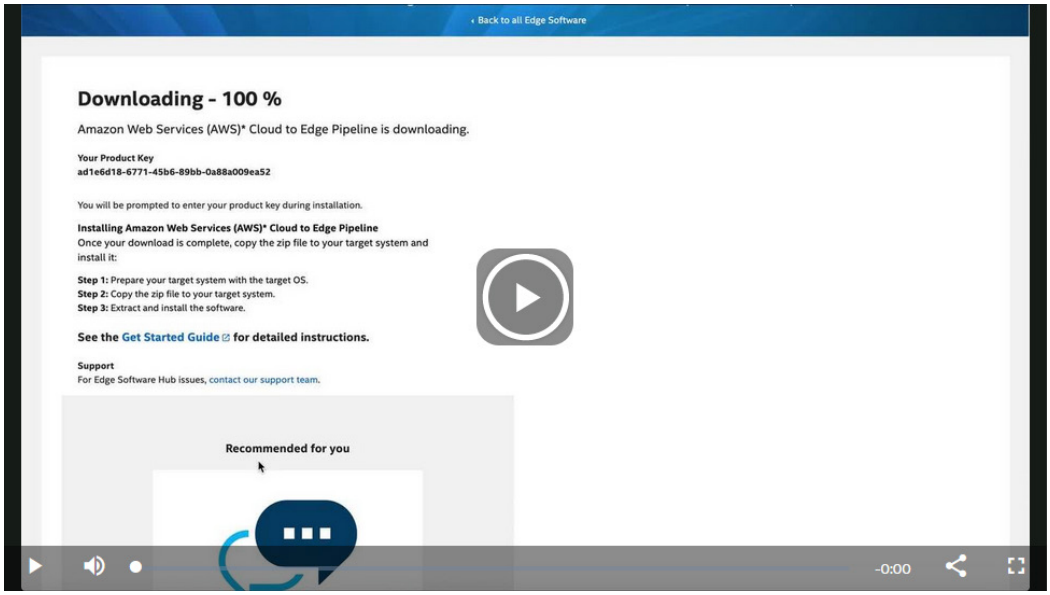
工作原理

该用例使用了英特尔® OpenVINO™ 工具套件分发版中包含的推理引擎，能够帮助云开发人员在英特尔边缘设备上部署 AI 推理应用并获得优化性能。同时，利用 AWS Greengrass，可以将视觉分析负载从云端安全无缝地迁移到边缘。



立即开始

准备好在边缘测部署 AI 推理？从英特尔® 边缘软件中心下载 Amazon Web Services (AWS) * 云端到边缘管道。下载此用例后，请遵循文档建立云端到边缘的管道并进行边缘推理。



intel®

脚注

[1] 当前，英特尔® OpenVINO™ 工具套件分发版仅支持部分 TFHub 和 Keras 模型。

通知和免责声明

在性能测试过程中使用的软件及工作负载可能仅针对英特尔® 微处理器进行了性能优化。

性能测试 (如 SYSmark 和 MobileMark) 使用特定的计算机系统、组件、软件、操作和功能进行测量。上述任何要素的变动都有可能导致测试结果的变化。您应该参考其他信息和性能测试以帮助您全面评估正在考虑的采购，包括产品在与其他产品结合使用时的性能。有关详细的完整信息，请访问：www.intel.com/benchmarks

性能结果基于截至配置中所示日期的测试，可能并不反映所有公开发布的安全更新。有关配置详细信息，请参阅备份页。没有任何产品或组件是绝对安全的。

您的成本或结果可能有所差异。

英特尔技术可能需要启用硬件、软件或激活服务。

英特尔编译器针对英特尔微处理器的优化程度可能与针对非英特尔微处理器的优化程度不同。这些优化包括 SSE2、SSE3 和 SSSE3 指令集和其他优化。对于在非英特尔制造的微处理器上进行的优化，英特尔不对相应的可用性、功能或有效性提供担保。此产品中依赖于处理器的优化仅适用于英特尔微处理器。某些不是专门面向英特尔微体系结构的优化保留专供英特尔微处理器使用。请参阅相应的产品用户和参考指南，以了解关于本通知涉及的特定指令集的更多信息。

英特尔尊重人权，坚决与侵犯人权的行为划清界限。请参阅英特尔《全球人权原则》。英特尔的产品和软件仅限用于不会导致违反国际公认人权或成为侵权推手的应用。