OpenVINO™ 工具套件

入门指南

操作步骤

指南

资源

性能信息

API 参考

英特尔® 深度学习部署工具套件简介

本文档

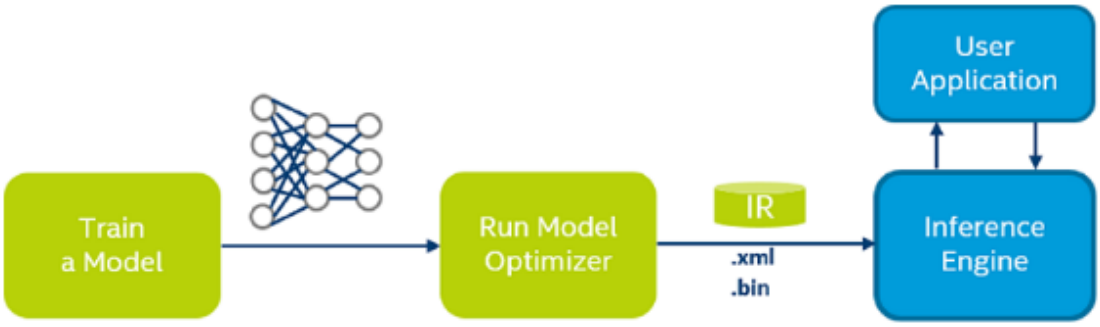
部署挑战

将深度学习网络从训练环境部署到嵌入式平台进行推理是一项复杂的任务，带来了许多技术挑战：

- 业界广泛使用了许多深度学习框架，例如 Caffe *、TensorFlow*、MXNet* 和 Kaldi* 等。
- 深度学习网络的训练通常在数据中心或服务器群内执行，推理可能在面向性能和功耗优化的嵌入式平台内执行。从软件角度（编程语言、第三方依赖关系、内存消耗、支持的操作系统）和硬件角度（不同的数据类型、有限的功耗）来看，此类平台一般都存在限制，因此通常不建议（有时甚至不可能 ）使用原始训练框架进行推理。一种替代解决方案是使用面向特定硬件平台优化的专用推理 API。
- 部署工作包括支持各种越来越复杂的层和网络，这进一步增加了难度。显然，确保转换网络的准确性并非易事。

部署工作流程

该流程假定您拥有使用其中一种支持框架进行训练的网络模型。以下方案展示了部署经过训练的深度学习模型的典型工作流程：



步骤包括：

1. 针对特定框架（用于训练模型）**配置模型优化器**。
2. 运行模型优化器，根据经过训练的网络拓扑、权重和偏差值以及其他可选参数，生成模型的**中间表示 (IR)**文件。
3. 使用目标环境中的推理引擎和提供的**推理引擎示例应用**，测试 IR 格式的模型。
4. **将推理引擎集成**到您的应用中，在目标环境中部署模型。

模型优化器

模型优化器是一种跨平台命令行工具，可促进训练和部署环境之间的过渡，执行静态模型分析并自动调整深度学习模型，以在端点目标设备上实现最佳执行。

模型优化器旨在支持多种深度学习支持框架和格式。

运行模型优化器时，您无需考虑要使用的目标设备，可以在所有目标中使用相同的 MO 输出。

模型优化器工作流程

该流程假定您拥有使用其中一种支持框架进行训练的网络模型。模型优化器的工作流程可描述如下：

- 为用于训练模型的其中一个支持的深度学习框架**配置模型优化器**。

- 以经过训练的网络作为输入，该网络包括特定的网络拓扑以及调整后的权重和偏差（带有一些可选参数）。
- **运行模型优化器**，以执行特定的模型优化（例如，某些网络层的水平融合）。优化是针对特定框架的，请参考相应的文档页面：**转换 Caffe 模型**，**转换 TensorFlow 模型**，**转换 MXNet 模型**，**转换 Kaldi 模型**，**转换 ONNX 模型**。
- 模型优化器生成网络的**中间表示 (IR)**文件作为输出，用作所有目标上的推理引擎的输入。

支持的框架和格式

- Caffe*（大多数公共分支）
- TensorFlow*
- MXNet*
- Kaldi*
- ONNX*

支持的模型

如欲获取支持模型的列表，请参见面向特定框架或格式的页面：

- **支持的 Caffe* 模型**
- **支持的 TensorFlow* 模型**
- **支持的 MXNet* 模型**
- **支持的 ONNX* 模型**
- **支持的 Kaldi* 模型**

中间表示

描述深度学习模型的中间表示在连接 OpenVINO™ 工具套件组件方面起着重要作用。IR 是一对文件：

- `.xml`：拓扑文件 - 一种描述网络拓扑的 XML 文件
- `.bin`：训练数据文件 – 一种包含权重和偏差二进制数据的 .bin 文件

可以使用推理引擎读取、加载和推理中间表示 (IR) 文件。推理引擎 API 在许多支持的英特尔® 平台上提供了一个统一的 API。训练后优化工具还可以使用、修改和写入 IR，该工具可提供量化功能。

更多详细信息请参见有关**中间表示和操作集**的专用描述。

nGraph 集成

OpenVINO 工具套件基于 nGraph 功能，用于图形构建 API、图形转换引擎和 Reshape，可替代以前的 NN Builder API 产品。nGraph 函数用作 CNNNetwork API 下运行时模型的中间表示。如果使用某些常规 API 方法时要求向后兼容，则 CNN 网络的常规表示形式仍然可用。请参见**nGraph Flow 概述**，详细了解 nGraph 与推理引擎的集成以及与传统表示形式的共存。

推理引擎

推理引擎是一种提供统一 API，将推理和应用逻辑集成起来的运行时。

- 以模型作为输入。该模型以模型优化器生成的特定形式的**中间表示 (IR)**来表示。
- 优化目标硬件的推理执行。
- 提供推理解决方案，节省嵌入式推理平台的空间。

推理引擎支持多种图像分类网络的推理，包括 AlexNet、GoogLeNet、VGG 和 ResNet 网络系列，用于图像分割的完全卷积网络（如 FCN8）以及对象检测网络（如 Faster R-CNN）。

如欲获取受支持硬件的完整列表，请参见**支持的设备**部分。

推理引擎包**包含**头文件、运行时库**和****示例控制台应用**，这些示例演示了如何在应用中使用推理引擎。

另请参阅

- 推理引擎示例
- 英特尔® 深度学习部署工具套件网页

优化声明

如欲了解有关编译器优化的更多完整信息，请参见优化声明。

有关编译器优化的更多完整信息，请参阅我们的优化声明

支持

英特尔® OpenVINO™ 工具套件分发版的英特尔® 开发人员专区论坛

Cookie

英特尔 Cookie 和类似技术声明