AI学习笔记--Tensorflow--TensorFlow models on the Edge TPU

本页描述了模型与TPU兼容的平台，以及如何创建它们，或者训练一个您自己的TensorFlow模型，或者通过迁移学习对现有模型进行再训练。TPU能够实现卷积神经网络（CNN）等深层前向神经网络。它只支持完全8位量化的TensorFlow Lite模型，然后专门为Edge TPU编译。

如果您不熟悉TensorFlow Lite，它是为移动和嵌入式设备设计的TensorFlow的轻量级版本。它在较小的二进制大小下实现了低延迟的推理，TensorFlow Lite模型和解释器内核都要小得多。TensorFlow Lite模型可以通过量化，将32位参数数据转换为8位表示（这是TPU所要求的），从而使模型变得更小、更有效。

不能使用TensorFlow Lite直接训练模型；相反，必须使用TensorFlow Lite转换器将模型从TensorFlow文件（如.pb文件）转换为TensorFlow Lite文件（即.tflite文件）。

图1说明了创建与Edge TPU兼容的模型的基本过程。大多数工作流使用标准的TensorFlow工具。一旦有了TensorFlow Lite模型，就可以使用我们的Edge TPU编译器创建一个与Edge TPU兼容的.tflite文件。

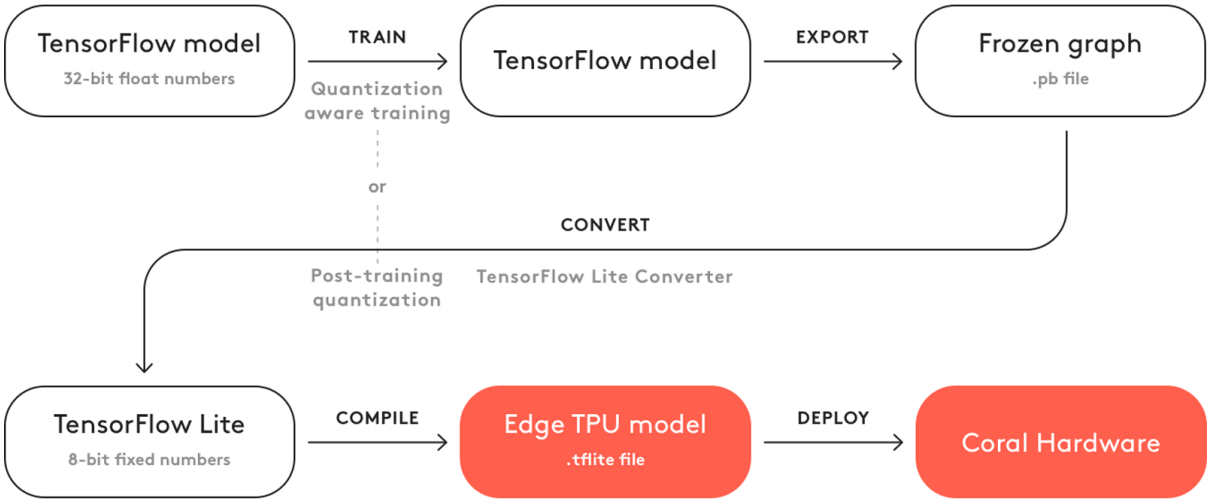

Figure 1. The basic workflow to create a model for the Edge TPU

然而，您不需要遵循这整个过程来为Edge TPU创建一个好的模型。相反，您可以利用与您自己的数据集重新训练的现有TysFunes模型来调整与边缘TPU兼容的现有模型。例如，MobileNet是一种流行的图像分类/检测模型体系结构，它与Edge TPU兼容。我们已经创建了此模型的多个版本，您可以将其用作创建可识别不同对象的自己模型的起点。首先，请参阅下一节关于如何用迁移学习重新训练现有模型。

如果您已经设计或计划设计自己的模型体系结构，那么您应该阅读下面关于模型需求的部分。

- 迁移学习

在没有建立模型，并且从头开始训练的，我们称迁移学习（transfer learning），这是

使用一种叫做转移学习的技术（有时也称为“微调”）重新训练一个已经与边缘TPU兼容的现有模型。

从头开始训练一个神经网络（当它没有计算出的权重或偏差时）可能需要几天的计算时间，并且需要大量的训练数据。但是转移学习允许你从一个已经为相关任务训练过的模型开始，然后使用一个较小的训练数据集执行进一步的训练来教授模型新的分类。您可以通过重新训练整个模型（调整整个网络的权重）来实现这一点，但您也可以通过简单地删除执行分类的最后一层，并在上面训练一个新层来识别新类，从而获得非常精确的结果。

使用这个过程，需要足够的训练数据和对超参数的一些调整，您可以在一次坐姿中创建一个高精度的TensorFlow模型。一旦您对模型的性能满意，只需将其转换为TensorFlow Lite，然后为Edge TPU编译它。而且，由于模型架构在迁移学习期间不会改变，因此您知道它将完全编译为Edge TPU（假设您从兼容的模型开始）。

如果您已经熟悉转移学习，请查看我们的Edge TPU兼容模型，您可以将其用作创建自己的模型的起点。只需点击下载“所有模型文件”即可获得TensorFlow模型和开始转移学习所需的预训练检查点。

如果您对这项技术还不熟悉，并且希望快速看到一些结果，请尝试以下教程，以简化使用新类重新训练MobileNet模型的过程：

- Retrain an image classification model
- Retrain an object detection model