eIQ® Inference with TensorFlow™ Lite

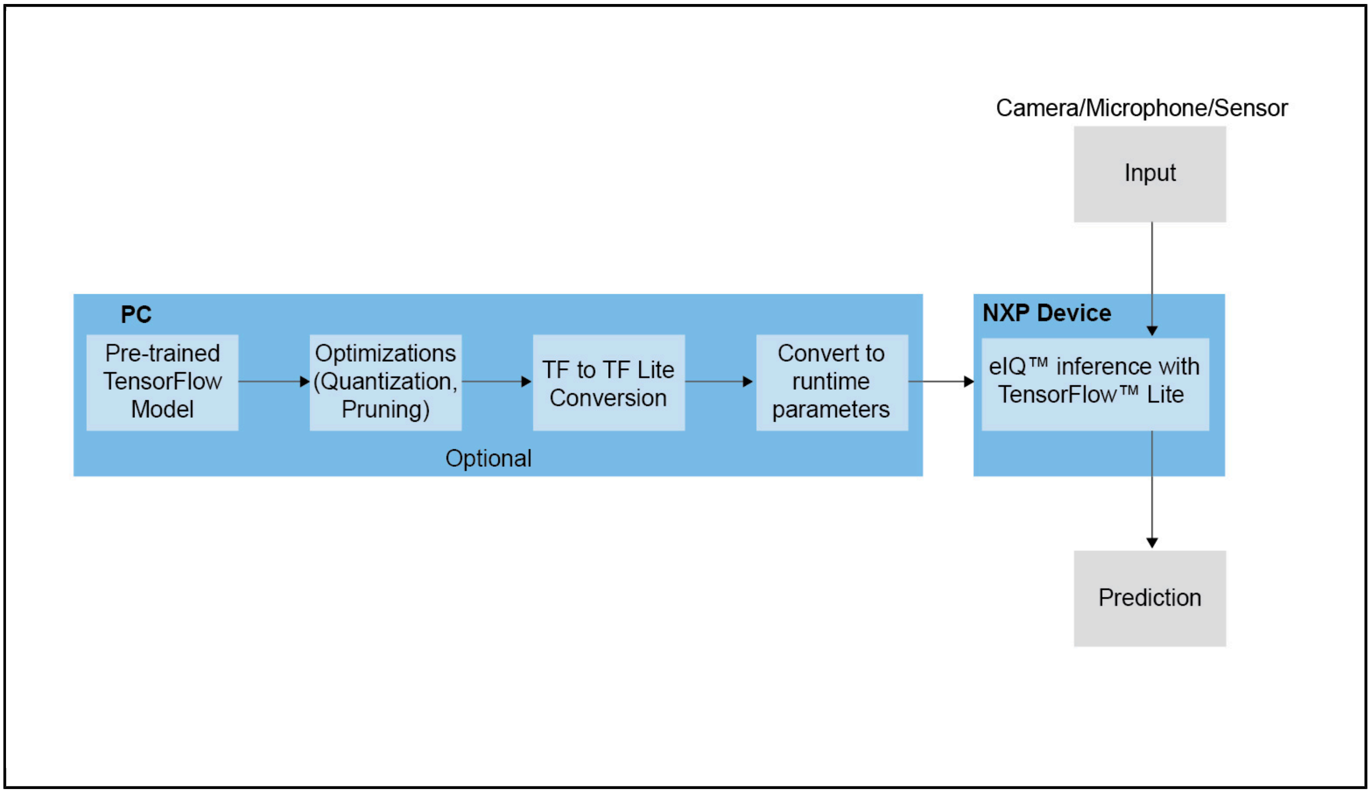
eIQTensorFlowLite

Last Updated: Apr 16, 2024

Integrated into NXP's Yocto development environment, eIQ software delivers TensorFlow Lite for NXP’s MPU platforms. Developed by Google to provide reduced implementations of TensorFlow (TF) models, TF Lite uses many techniques for achieving low latency such as pre-fused activations and quantized kernels that allow smaller and (potentially) faster models. Furthermore, like TensorFlow, TF Lite utilizes the Eigen library to accelerate matrix and vector arithmetic.

TF Lite defines a model file format, based on FlatBuffers. Unlike TF’s protocol buffers, FlatBuffers have a smaller memory footprint allowing better use of cache lines, leading to faster execution on NXP devices. TF Lite supports a subset of TF neural network operations, and also supports recurrent neural networks (RNNs) and long short-term memory (LSTM) network architectures.

eIQ® TensorFlow Lite Block Diagram


Camera/Microphone/Sensor
Input
PC
Pre-trained TensorFlow Model
Optimizations (Quantization, Pruning)
TF to TF Lite Conversion
Convert to runtime parameters
Optional
NXP Device
eIQ™ inference with TensorFlow™ Lite
Prediction

View additional information for eIQ® Inference with TensorFlow™ Lite.

Note: The information on this document is subject to change without notice.

www.nxp.com