Machine Learning Models Optimized for Google Tensor's Edge TPU

Requirements

Overview

An illustration of NAS to find Edge TPU optimized models, Each column represents a stage in the natural network, with dots indicating different options, and each color representing a different type of building block. A path from inputs (e.g., an image) to outputs (e.g., per-pixel label predictions) through the matrix represents a candidate neural network. In each iteration of the search, a neural network is formed using the blocks chosen at every stage, and the search algorithm aims to find neural networks that jointly minimize TPU latency and/or energy and maximize accuracy.

This repository contains machine learning models optimized for the Edge TPU in Pixel 6's SoC, Google Tensor. We use Neural Architecture Search (NAS) to automate the process of designing ML models and incentivize the search algorithms to discover models that achieve higher quality as well as better latency and computing efficiency. This automation also allows us to scale the development of ML models for a variety of on-device tasks. We're making these ML models publicly available through the Tensorflow model garden and Tensorflow Hub to enable researchers and developers to bootstrap further use case development on Pixel 6.

Image Classification

Object Detection

Semantic Segmentation

Natural Language Understanding