Hello AI World

0. Introduction

NVIDIA TensorRT™ is a high-performance deep learning inference platform. It includes a deep learning inference optimizer and runtime that delivers low latency and high throughput for deep learning inference applications. During inference, TensorRT-based applications execute up to 40 times faster than CPU-only platforms. With TensorRT, you can optimize neural network models trained in all major frameworks, calibrate low precision to high accuracy, and finally deploy to hyperscale data centers, embedded or automotive product platforms.

TensorRT is built on CUDA, NVIDIA's parallel programming model, enabling you to leverage the libraries, development tools, and technologies in CUDA-X AI to optimize inference for all deep learning frameworks for artificial intelligence, autonomous machines, high-performance computing, and graphics.

TensorRT provides INT8 and FP16 optimizations for production deployment of deep learning inference applications, such as video streaming, speech recognition, recommendations, and natural language processing. Reduced precision inference significantly reduces application latency, a requirement for many real-time services, autonomous, and embedded applications.

Hello AI World can run entirely on Jetson, including inference using TensorRT and transfer learning using PyTorch. The inference portion of Hello AI World — which includes writing your own image classification and object detection applications for Python or C++, as well as a live camera demo — can run on Jetson in about two hours or less, while transfer learning is best left running overnight.