Hello AI World

0. **Introduction**

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

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

TEnsorRT provides INT8 and FP16 optimization for production deployment of deep learning reasoning applications, such as video streaming, speech recognition, recommendation and natural language processing. Reducing precision inference can significantly reduce application latency, which is a requirement for many real-time services, automated, and embedded applications.

Hello AI World can run completely on Jetson, including reasoning with TensorRT and transfer learning with PyTorch. The reasoning part of Hello AI World -- including writing your own image classification and object detection applications for Python or C++, as well as live camera demonstration -- can be run on Jetson in about two hours or less, and transfer learning is best to let it run overnight