ACCELERATING AI

GPU POWERED DEEP LEARNING TRAINING

Welcome to the era of AI and intelligent machines. An era fueled by big data, driven by deep learning, and powered by GPUs. Deep learning training is the process by which a neural network learns from data in the form of images, video, text, speech, and transactions, converting it into intelligence.

AI IS ONLY USEFUL IF IT'S FAST

A neural network can be trained to understand natural conversation, monitor hundreds of live video streams, or navigate a vehicle safely through a city. However, inference needs to be fast to deliver that learned intelligence to users.

The combination of TensorRT with NVIDIA GPUs delivers the world's fastest inference for AI-enabled services, with latency under that seven-millisecond mark.

ACCELERATING AI

GPU Deep Learning with the NVIDIA TensorRT Programmable Inference Accelerator

Inference is when a trained neural network is deployed into a product or application so that it can do things such as recognize images, understand conversational speech, or make a shopping recommendation.

THE EXPLOSION OF AI

The uses for deep learning inference are becoming more complex and widespread. People have come to expect fast and natural interactions with their devices. At the same time, unprecedented growth is dramatically increasing the number and variety of AI-powered applications and products.

THE POWER OF NVIDIA TensorRT

TensorRT is a high-performance optimizing compiler and runtime engine for production deployment of AI applications. It can rapidly optimize, validate, and deploy trained neural networks for inference to hyperscale data centers, embedded, or automotive GPU platforms.

INFERENCING PERFORMANCE

40X faster inference with TensorRT 3 on NVIDIA Tesla® V100 compared to CPU-only inference.

INFERENCING COST SAVINGS

The introduction of TensorRT on GPUs enables an AI inferencing solution that sharply boosts performance and slashes the cost of inferencing from the data center to cloud to edge devices, including self-driving cars and robots.

ONE UNIFIED SOLUTION FOR AI

With the introduction of TensorRT on GPUs, NVIDIA offers an AI inferencing solution that sharply boosts performance and slashes the cost of inferencing from the data center to cloud to edge devices, including self-driving cars and robots.

Faster AI. Lower cost. www.nvidia.com/inference

Deep learning is now used to build AI into everything from kitchen appliances to cars to robots. With every new use case, the cost of supporting these applications and products increases as well.

More features, more products, more applications, more complexity, more usage.

More...

Insight

Versatility

Efficiency

Performance

Accuracy

Latency

Cost savings

Speech recognition

Drones

Autonomous cars

Virtual assistants

Mobile phones

Medical devices

Smart cities

Personalized content

Drones

Understanding

Interpreting

Analyzing

Strategy

Personalizing

Recognizing

Understanding

Predicting

Avoiding

Drone

Autonomous cars

Virtual assistants

Mobile phones

Medical devices

Smart cities

ANALYZING

STRATEGY

INTERPRETING

DATA

RECOGNIZING

BEHAVIOR

PREDICTING

EVENTS

AVOIDING

COLLISIONS

DRONES

AUTONOMOUS CARS

SMART CITIES

VIRTUAL ASSISTANTS

MOBILE PHONES

MEDICAL DEVICES