THE DEEP LEARNING AI REVOLUTION

JEN-HSUN HUANG, FOUNDER & CEO, NVIDIA | GTC JAPAN 2016
A NEW ERA OF COMPUTING

PC INTERNET
WinTel, Yahoo!
1 billion PC users

MOBILE-CLOUD
iPhone, Amazon AWS
2.5 billion mobile users

AI & IOT
Deep Learning, GPU
100s of billions of devices

1995
2005
2015
THE STAGE IS SET FOR THE AI REVOLUTION

2012: GPU Deep Learning Big Bang

2015: ImageNet — Deep Learning achieves superhuman image recognition

2016: Microsoft’s Deep Learning system achieves milestone in speech recognition

Abstract

Deep learning is a subset of machine learning in artificial intelligence (AI) that is designed to mimic the way the human brain operates to identify patterns within data. When compared to other AI methods, deep learning has shown exceptional results in fields like object recognition and speech recognition. The TITAN X is the latest technology in deep learning, and it is capable of achieving new benchmarks in both areas.

The TITAN X is a high-performance computing engine that provides exceptional power for deep learning applications. With its parallel processing capabilities, it is able to achieve high speeds and accuracy in object recognition and speech recognition tasks.


Microsoft, Google
3.5% Error Rate

Deep Learning

Hand-coded CV

Human

7.6%

96%

2010 2011 2012 2013 2014 2015

Year of model evaluation

Error Rate

Microsoft 09/13/16

THE STAGE IS SET FOR THE AI REVOLUTION

Abstract

Deep learning is a subset of machine learning in artificial intelligence (AI) that is designed to mimic the way the human brain operates to identify patterns within data. When compared to other AI methods, deep learning has shown exceptional results in fields like object recognition and speech recognition. The TITAN X is the latest technology in deep learning, and it is capable of achieving new benchmarks in both areas.

The TITAN X is a high-performance computing engine that provides exceptional power for deep learning applications. With its parallel processing capabilities, it is able to achieve high speeds and accuracy in object recognition and speech recognition tasks.


Microsoft, Google
3.5% Error Rate

Deep Learning

Hand-coded CV

Human

7.6%

96%

2010 2011 2012 2013 2014 2015

Year of model evaluation

Error Rate

Microsoft 09/13/16

THE STAGE IS SET FOR THE AI REVOLUTION

Abstract

Deep learning is a subset of machine learning in artificial intelligence (AI) that is designed to mimic the way the human brain operates to identify patterns within data. When compared to other AI methods, deep learning has shown exceptional results in fields like object recognition and speech recognition. The TITAN X is the latest technology in deep learning, and it is capable of achieving new benchmarks in both areas.

The TITAN X is a high-performance computing engine that provides exceptional power for deep learning applications. With its parallel processing capabilities, it is able to achieve high speeds and accuracy in object recognition and speech recognition tasks.


Microsoft, Google
3.5% Error Rate

Deep Learning

Hand-coded CV

Human

7.6%

96%

2010 2011 2012 2013 2014 2015

Year of model evaluation

Error Rate

Microsoft 09/13/16
NVIDIA – “THE AI COMPUTING COMPANY”

GPU Computing

Visual Computing

Artificial Intelligence
SCI-FI NO LONGER THE NEAR FUTURE WITH VR, AR AND AI
GTC — 25X GROWTH IN GPU DL DEVELOPERS

4X Attendees

3X GPU Developers

25X Deep Learning Developers
WHY DID AI RESEARCHERS ADOPT GPUs FOR DEEP LEARNING?
BRAIN IS LIKE A GPU

BRAIN CREATES MENTAL IMAGES WHEN WE THINK
GPU IS LIKE A BRAIN
GPU DEEP LEARNING IS A NEW COMPUTING MODEL
AI — THE ULTIMATE COMPUTING CHALLENGE

**IMAGE RECOGNITION**

- **2012**
  - AlexNet
  - 8 Layers
  - 1.4 GFLOP
  - ~16% Error

- **2015**
  - ResNet
  - 152 Layers
  - 22.6 GFLOP
  - ~3.5% Error

**SPEECH RECOGNITION**

- **2014**
  - Deep Speech 1
  - 80 GFLOP
  - 7,000 hrs of Data
  - ~9% Error

- **2015**
  - Deep Speech 2
  - 465 GFLOP
  - 12,000 hrs of Data
  - ~5% Error

**TRAINING DEEP NEURAL NETS**

- Billions of Operations * Billions of Iterations
PASCAL — GPU OPTIMIZED FOR DEEP LEARNING

- Pascal — 5 Miracles
- NVIDIA DGX-1 Supercomputer
- 65X in 4 yrs

Chart: Relative speed-up of images/sec vs K40 in 2013. AlexNet training throughput based on 20 iterations. CPU: 1x E5-2680v3 12 Core 2.5GHz. 128GB System Memory. 8x M40 datapoint: 8x M40 GPUs in a node. P100: 8x P100 NVLink-enabled.

PaddlePaddle
Baidu Deep Learning

Caffe
Chainer
Caffe
Caffe
MXNet
Chainer
PaddlePaddle
Theano
Torch

Accelerate Every Framework
GPU DEEP LEARNING IS A NEW COMPUTING MODEL

10s of billions of image, voice, video queries per day

GPU inference for fast response, maximize datacenter throughput

DATACENTER INFERENCING
TESLA P4 & P40
INFERENCING ACCELERATORS

Pascal Architecture | INT8
P4: 50W | 40X Energy Efficient versus CPU
P40: 250W | 40X Performance versus CPU
TensorRT
PERFORMANCE OPTIMIZING INFERENCING ENGINE

FP32, FP16, INT8 | Vertical & Horizontal Fusion | Auto-Tuning
VGG, GoogLeNet, ResNet, AlexNet & Custom Layers
Available Today: developer.nvidia.com/tensorrt
AI DEVELOPMENT IN JAPAN

- Deep Learning for IOT
- Deep Learning for Recommendations
- Deep Learning for Trading Strategies
- Deep Learning for Retail Analytics
AI POWERS BILLIONS OF INTELLIGENT DEVICES

“Billions of intelligent devices will take advantage of DNNs to provide personalization and localization as GPUs become faster and faster over the next several years.”

— Tractica
JETSON TX1
EMBEDDED AI SUPERCOMPUTER

10W | 1 TF FP16 | >20 images/sec/W
FANUC BUILDING THE FACTORY OF THE FUTURE ON NVIDIA AI PLATFORM
FANUC BUILDING THE FACTORY OF THE FUTURE ON NVIDIA AI PLATFORM

- FANUC AI Brain
- Training GPU
- FIELD System Inferencing GPU
- GPU for Operations for Every Machine
FANUC FACTORY OF THE FUTURE
FANUC BUILDING THE FACTORY OF THE FUTURE ON NVIDIA AI PLATFORM

Dr. Kiyonori Inaba
Member of the Board, Executive Managing Officer and General Manager of FANUC
AI TRANSPORTATION – $10T INDUSTRY
AUTONOMOUS DRIVING
Perception, Reasoning, Driving
AI Supercomputing, AI Algorithms, Software
Scalable Architecture

NVIDIA DRIVE PX 2
AUTOCRUISE TO FULL AUTONOMY
ONE ARCHITECTURE

AutoCruise
AutoChauffeur
Full Autonomy

AutoChauffeur
AutoCruise
Full Autonomy
NVIDIA DRIVE PX 2 AUTOCRUISE

10W AI Supercomputer | Passive Cooling | Automotive IO
AI Highway Driving | Localization & Mapping
DRIVEWORKS ALPHA 1
OS FOR SELF-DRIVING CARS

- DriveNet
- OpenRoadNet
- Localization
- PilotNet

- Occupancy Grid
- Traffic Prediction
- Path Planning
- Action Engine

- Path Planning
- Traffic Prediction
- Action Engine
DRIVEWORKS ALPHA 1
OS FOR SELF-DRIVING CARS

- DriveNet
- OpenRoadNet
- Localization
- PilotNet

- Occupancy Grid
- Traffic Prediction
- Path Planning
- Action Engine

- 8:03
- 57
- 3.9
TESLA FOR CLOUD HD MAP PROCESSING

DRIVE PX 2 FOR IN-CAR HD MAP PROCESSING

OPEN PLATFORM—"CLOUD-TO-CAR"
HD MAP, AI ALGORITHMS, AI SUPERCOMPUTER

BAIDU & TOMTOM SELECT NVIDIA FOR SELF-DRIVING CARS
NVIDIA AI COMPUTING FOR INTELLIGENT MACHINES

AI

VISUAL COMPUTING

HPC
INTRODUCING XAVIER
AI SUPERCOMPUTER SOC

7 Billion Transistors 16nm FF
8 Core Custom ARM64 CPU
512 Core Volta GPU
New Computer Vision Accelerator
Dual 8K HDR Video Processors
Designed for ASIL C Functional Safety
INTRODUCING XAVIER

AI SUPERCOMPUTER SOC

DRIVE PX 2
- 2 PARKER + 2 PASCAL GPU
- 20 TOPS DL
- 120 SPECINT
- 80W

XAVIER
- 20 TOPS DL
- 160 SPECINT
- 20W
NVIDIA AI COMPUTING FOR EVERY INDUSTRY

- END-TO-END DEEP LEARNING
- AI CAR
- AI MANUFACTURING
- AI FUTURE
AI will Revolutionize Society

AI will Revolutionize Healthcare

AI will Revolutionize Transportation