Deep Learning on GPUs

March 2016

NVIDIA®

AGENDA

What is Deep Learning?

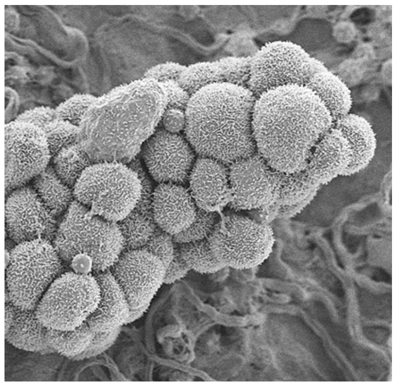
GPUs and DL

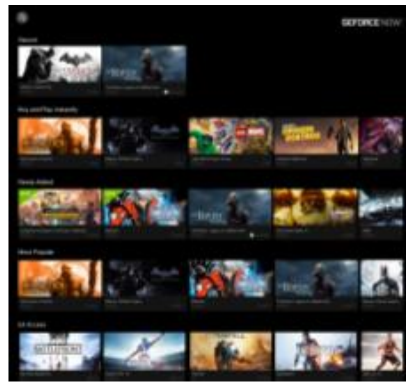
DL in practice

Scaling up DL

What is Deep Learning?

DEEP LEARNING EVERYWHERE





INTERNET & CLOUD

Image Classification
Speech Recognition
Language Translation
Language Processing
Sentiment Analysis
Recommendation

MEDICINE & BIOLOGY

Cancer Cell Detection
Diabetic Grading
Drug Discovery

MEDIA & ENTERTAINMENT

Video Captioning
Video Search
Real Time Translation

SECURITY & DEFENSE

Face Detection
Video Surveillance
Satellite Imagery

AUTONOMOUS MACHINES

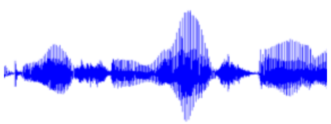
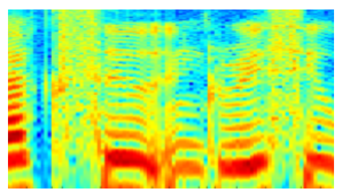
Pedestrian Detection
Lane Tracking
Recognize Traffic Sign

Traditional machine perception

Hand crafted feature extractors

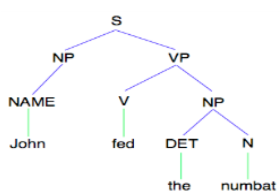
Raw data | Feature extraction | Classifier/ detector | Result


SVM,
shallow neural net,
…

HMM,
shallow neural net,
…

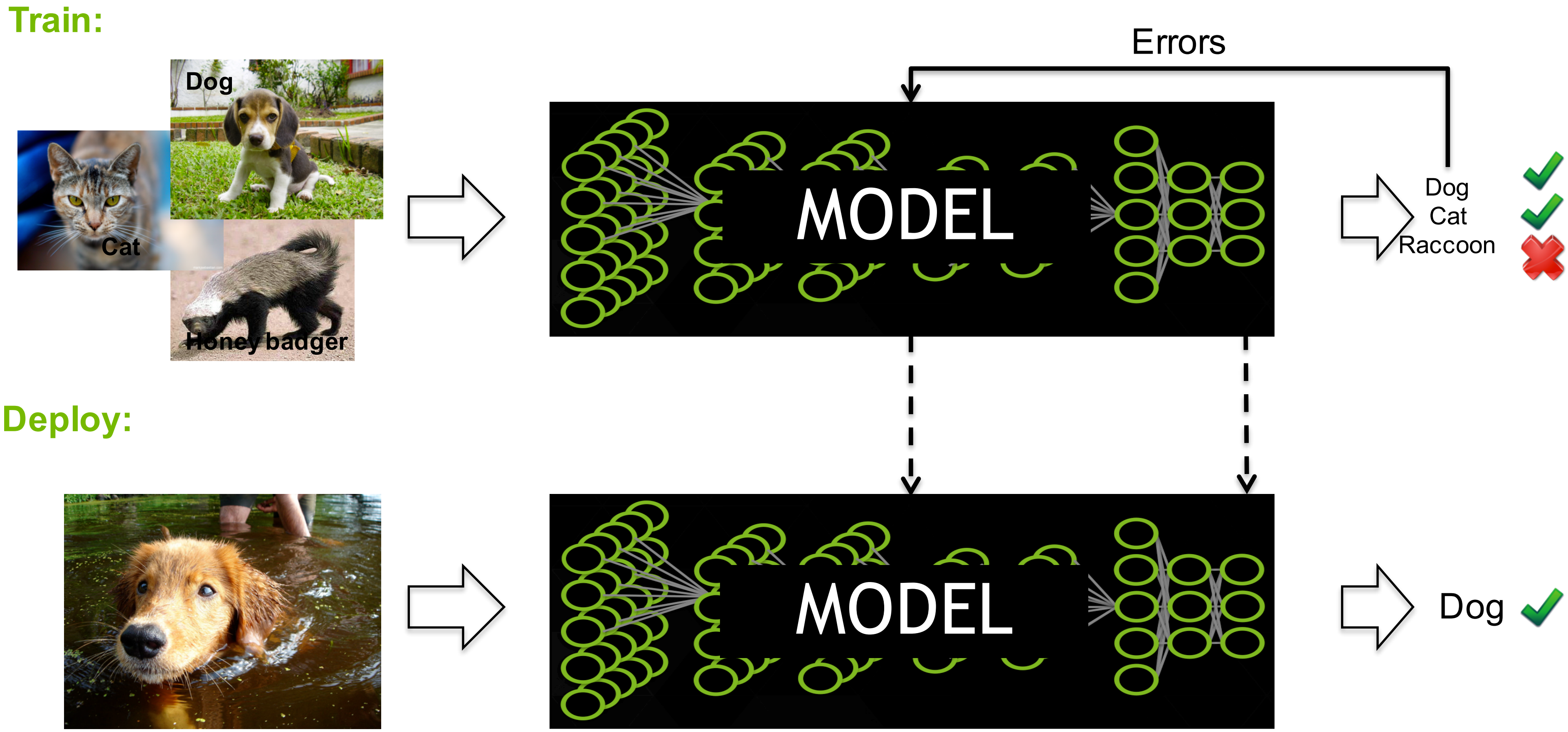
Speaker ID,
speech transcription, …

Clustering, HMM,
LDA, LSA
…

Topic classification,
machine translation,
sentiment analysis…

Deep learning approach

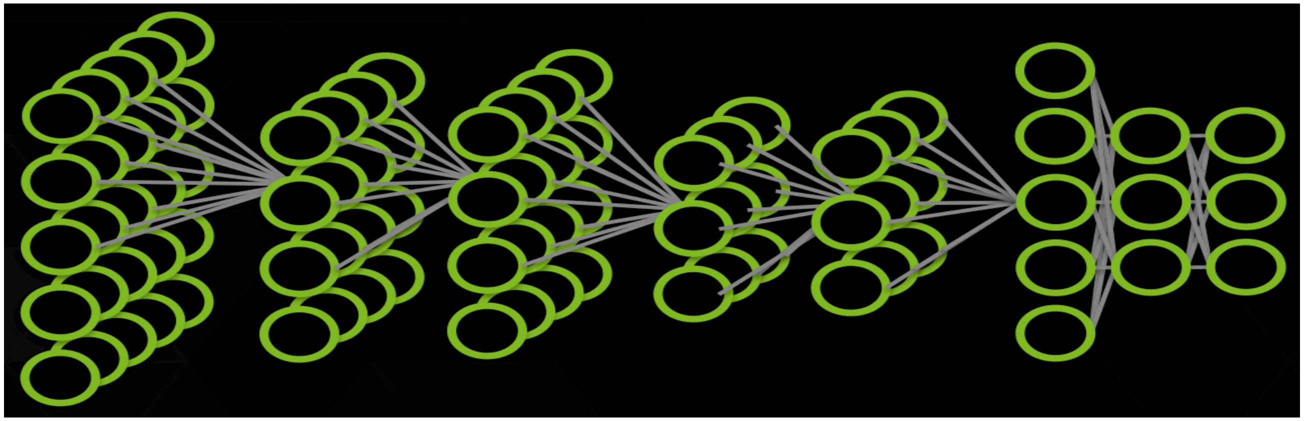


Artificial neural network

A collection of simple, trainable mathematical units that collectively learn complex functions

Hidden layers

Input layer

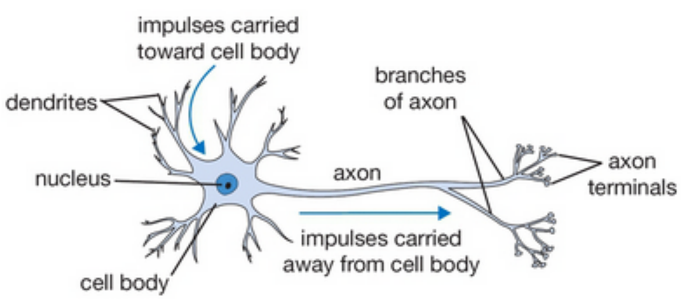


Output layer

Given sufficient training data an artificial neural network can approximate very complex functions mapping raw data to output decisions

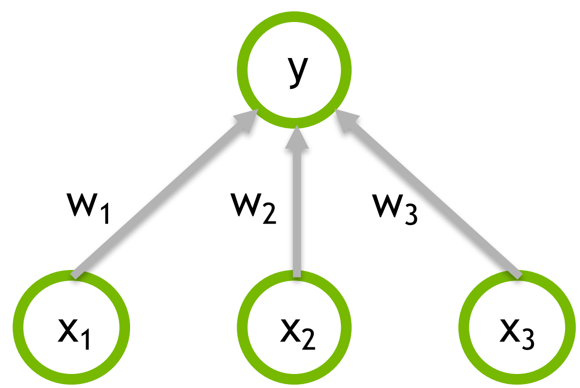
Artificial neurons

Biological neuron



From Stanford cs231n lecture notes

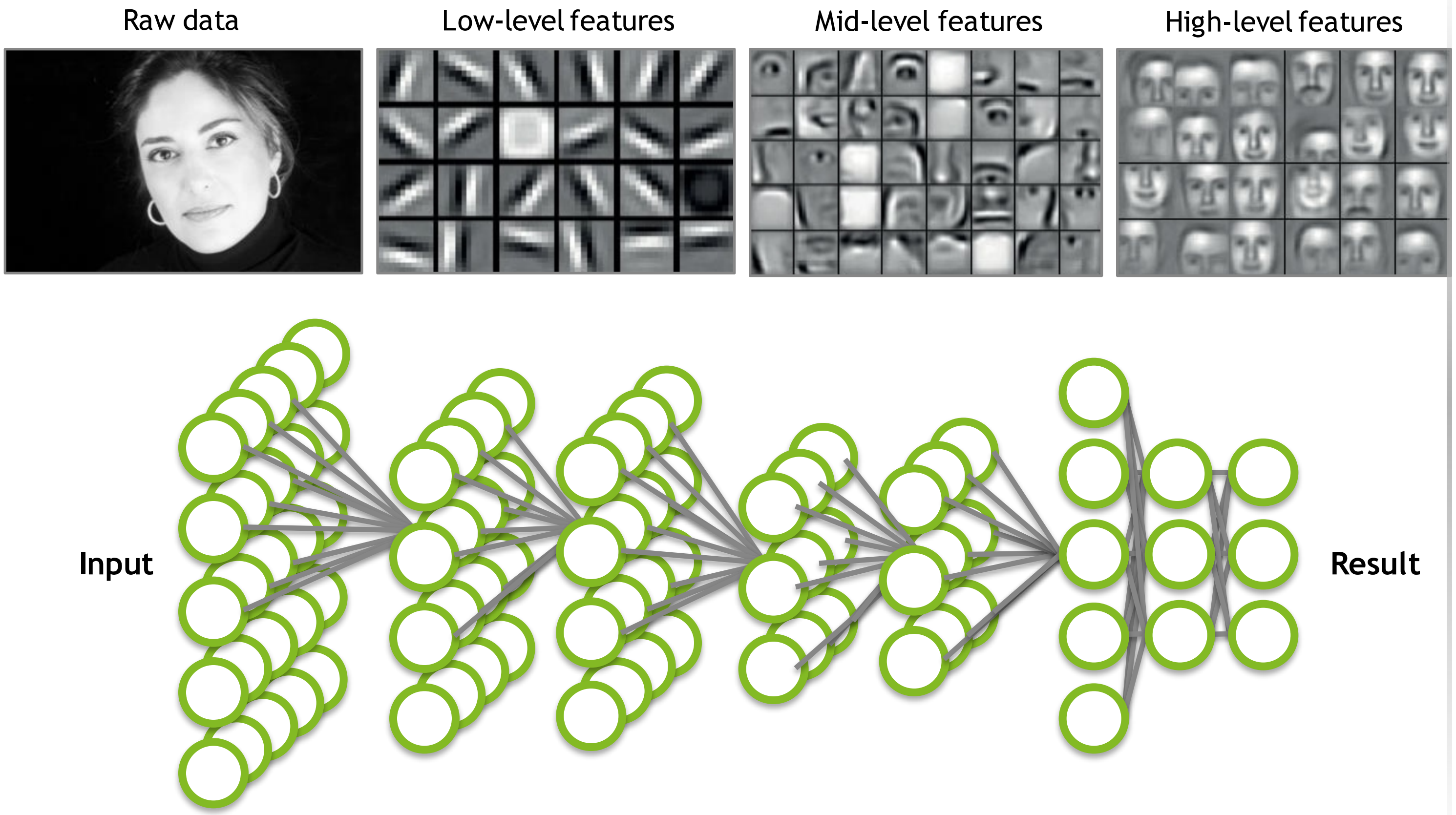
Artificial neuron



$$y=F(w_1x_1+w_2x_2+w_3x_3)$$

$$F(x)=max(0,x)$$

Deep neural network (dnn)



Application components:

Task objective
e.g. Identify face
Training data
10-100M images
Network architecture
~10 layers
1B parameters
Learning algorithm
~30 Exaflops
~30 GPU days

Deep learning benefits

- Robust
  - No need to design the features ahead of time – features are automatically learned to be optimal for the task at hand
  - Robustness to natural variations in the data is automatically learned
- Generalizable
  - The same neural net approach can be used for many different applications and data types
- Scalable
  - Performance improves with more data, method is massively parallelizable

Baidu Deep Speech 2

End-to-end Deep Learning for English and Mandarin Speech Recognition

English and Mandarin speech recognition

Transition from English to Mandarin made simpler by end-to-end DL

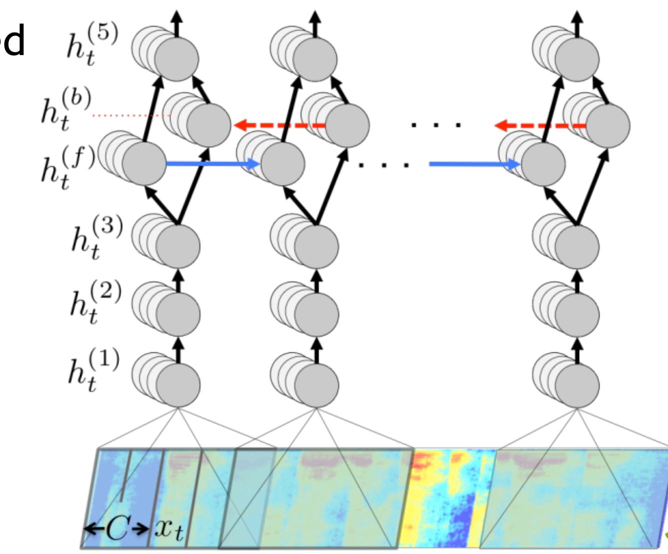
- No feature engineering or Mandarin-specifics required

More accurate than humans

- Error rate 3.7% vs. 4% for human tests

http://svail.github.io/mandarin/

http://arxiv.org/abs/1512.02595



AlphaGo

First Computer Program to Beat a Human Go Professional

Training DNNs: 3 weeks, 340 million training steps on 50 GPUs

Play: Asynchronous multi-threaded search

Simulations on CPUs, policy and value DNNs in parallel on GPUs

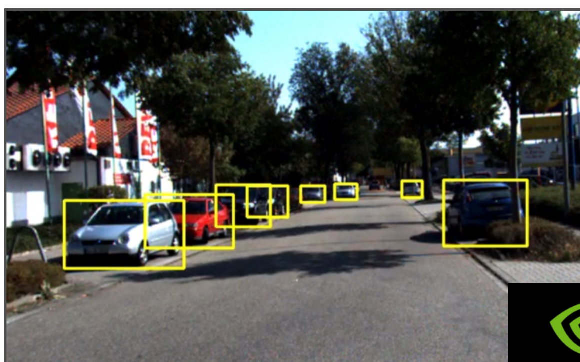
Single machine: 40 search threads, 48 CPUs, and 8 GPUs

Distributed version: 40 search threads, 1202 CPUs and 176 GPUs

Outcome: Beat both European and World Go champions in best of 5 matches

http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html
http://deepmind.com/alpha-go.html

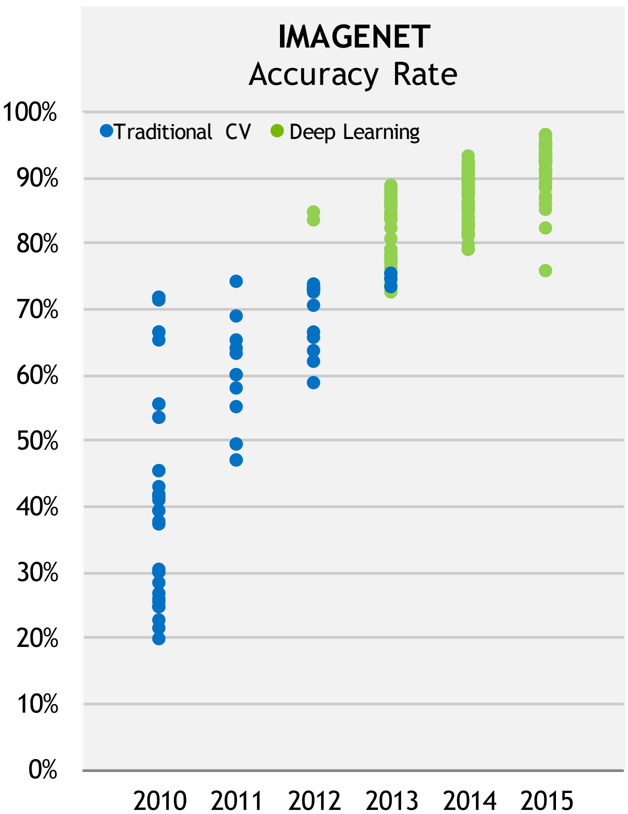
Deep Learning for Autonomous vehicles

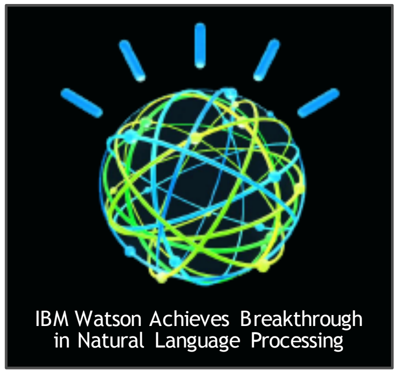


Deep Learning Synthesis

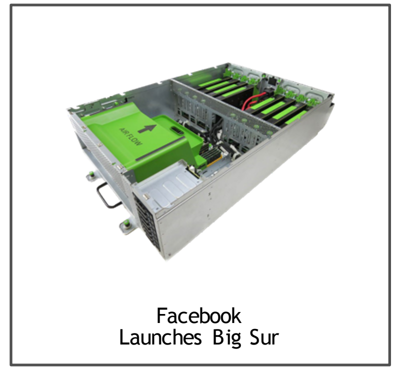
Texture synthesis and transfer using CNNs. Timo Aila et al., NVIDIA Research

THE AI RACE IS ON





IBM Watson Achieves Breakthrough in Natural Language Processing



Facebook Launches Big Sur

Baidu Deep Speech 2 Beats Humans

TensorFlow

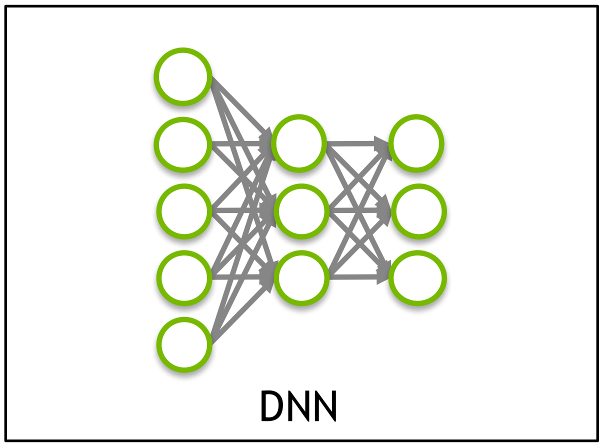
Google Launches TensorFlow

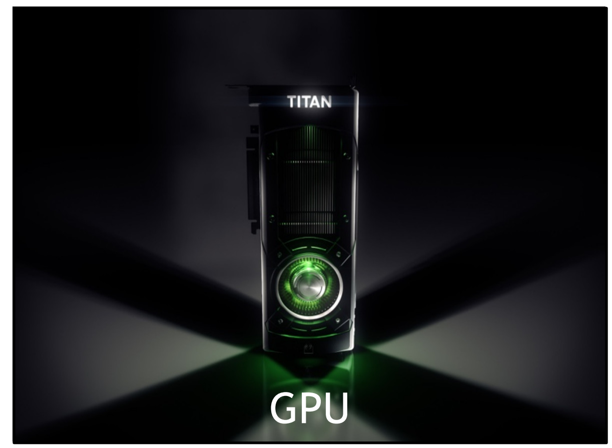
Toyota Invests $1B in AI Labs

Microsoft

Microsoft & U. Science & Tech, China Beat Humans on IQ

The Big Bang in Machine Learning





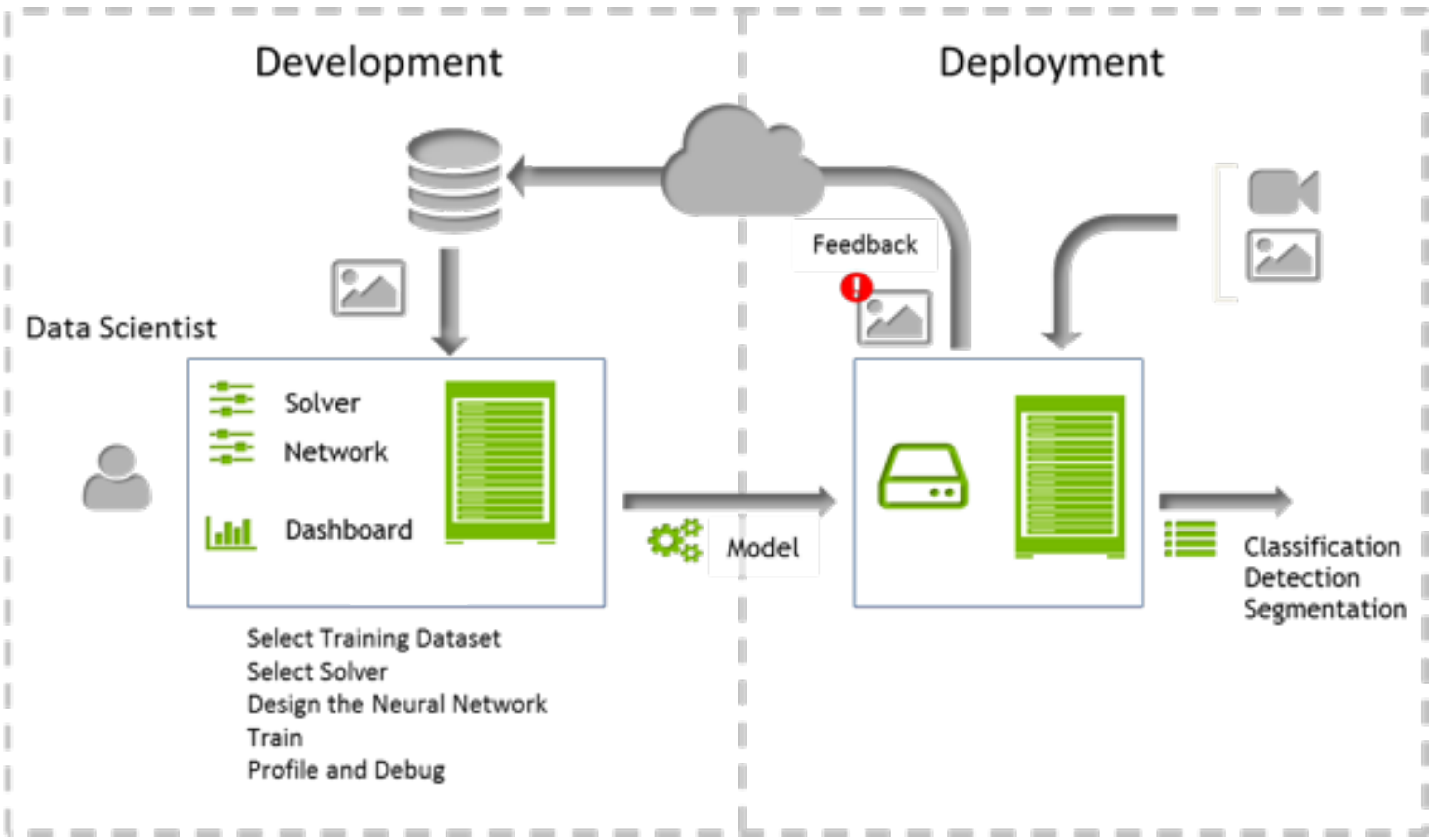
"Google's AI engine also reflects how the world of computer hardware is changing. (It) depends on machines equipped with GPUs… And it depends on these chips more than the larger tech universe realizes."

WIRED

GPUs and DL

USE MORE PROCESSORS TO GO FASTER

Deep learning development cycle

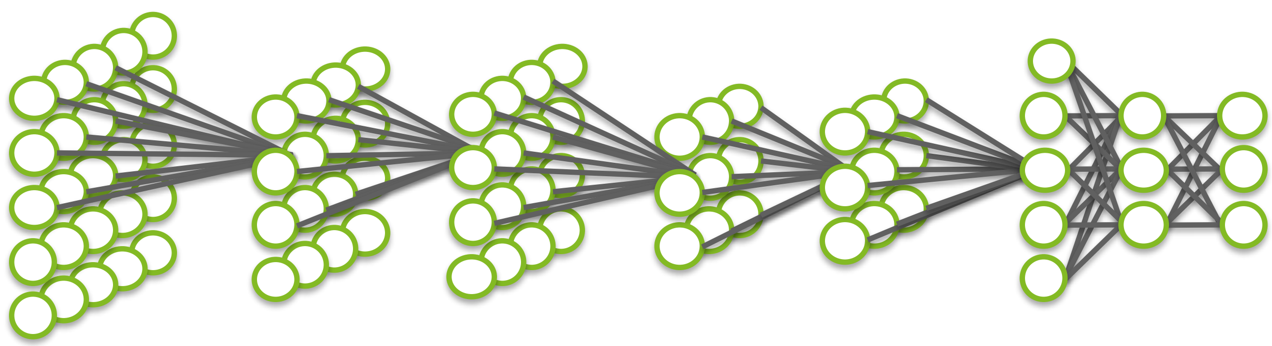


Three Kinds of Networks

DNN – all fully connected layers

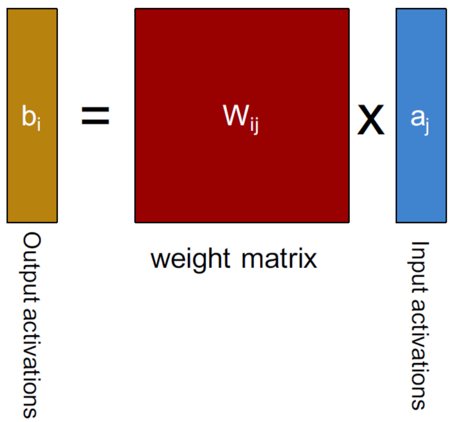
CNN – some convolutional layers

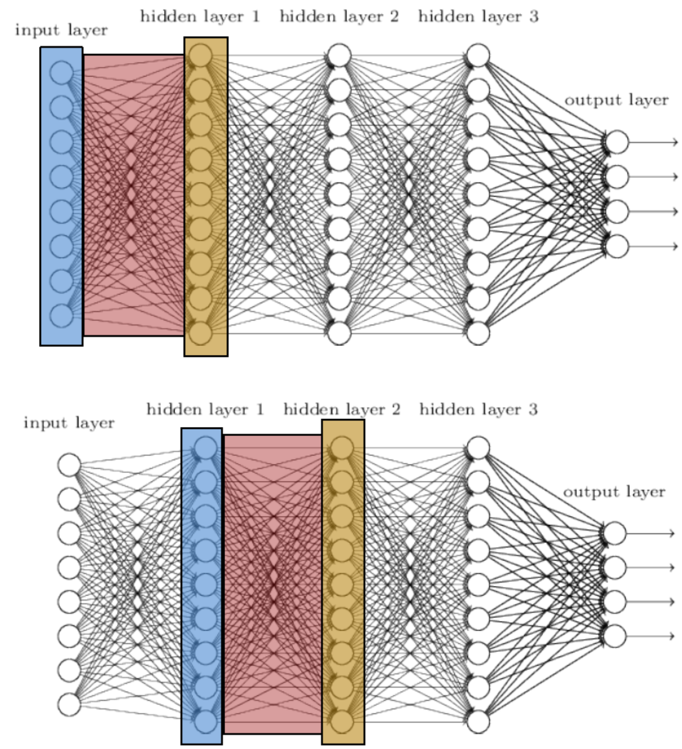
RNN – recurrent neural network, LSTM



DNN

Key operation is dense M x V



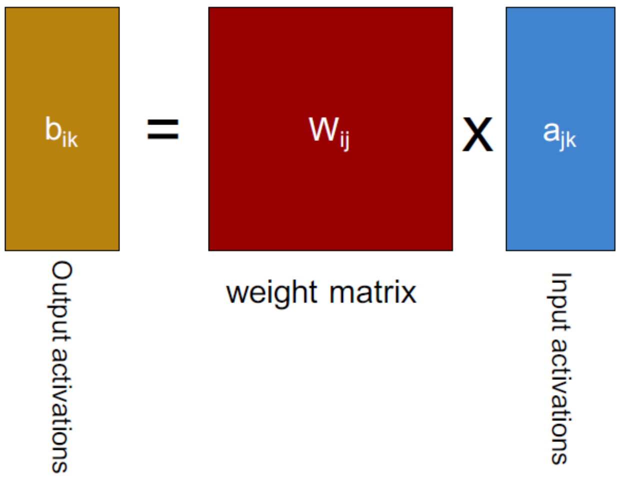


Repeat for each layer

Backpropagation uses dense matrix-matrix multiply starting from softmax scores

DNN

Batching for training and latency insensitive.
M x M



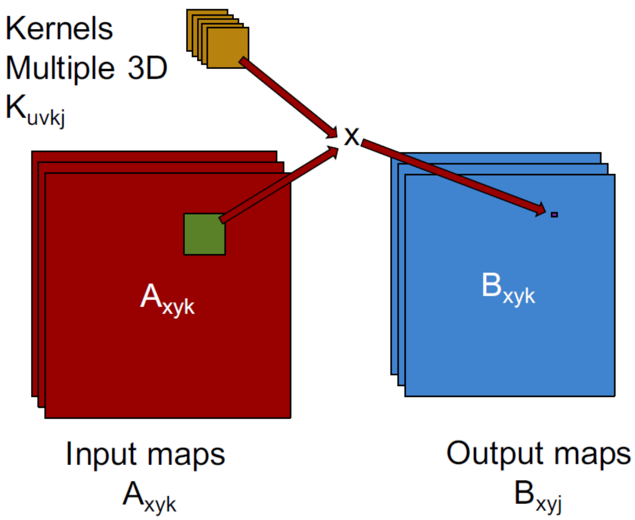
Batched operation is M x M – gives re-use of weights.

Without batching, would use each element of Weight matrix once.

Want 10-50 arithmetic operations per memory fetch for modern compute architectures.

CNN

Requires convolution and M x V



Filters conserved through plane

Multiply limited – even without batching.

6D Loop

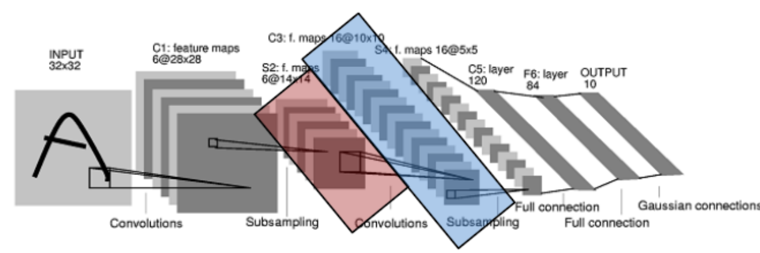
For each output map j

  For each input map k

    For each pixel x,y

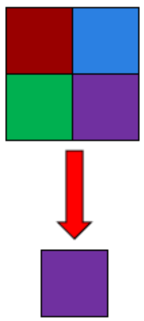
      For each kernel element u,v

$$B_{xyj} \mathrel{+}= A_{(x-u)(y-v)k} \times K_{uvkj}$$



Other Operations

To finish building a DNN



Pooling

ReLU
(or other non-linear function)

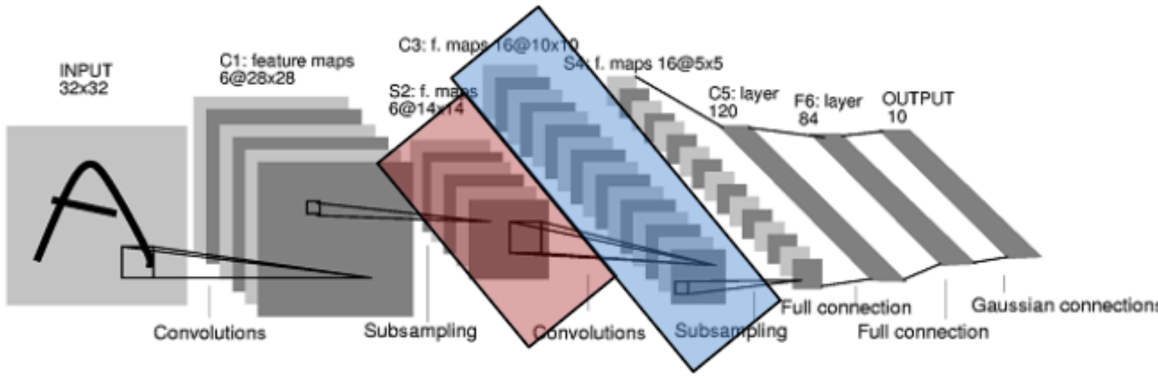
$$w_{ij} \mathrel{+}= \alpha a_j g_i$$

Weight Update

These are not limiting factors with appropriate GPU use

Complex networks have hundreds of millions of weights.

Lots of Parallelism Available in a DNN

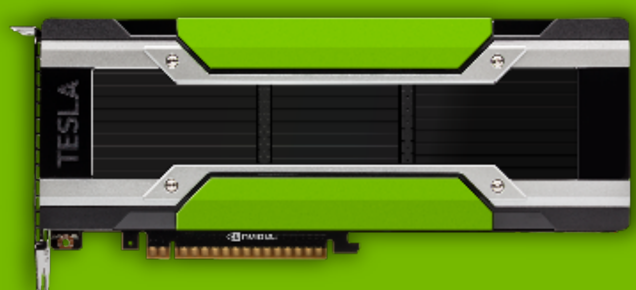


- Inputs
- Points of a feature map
- Filters
- Elements within a filter

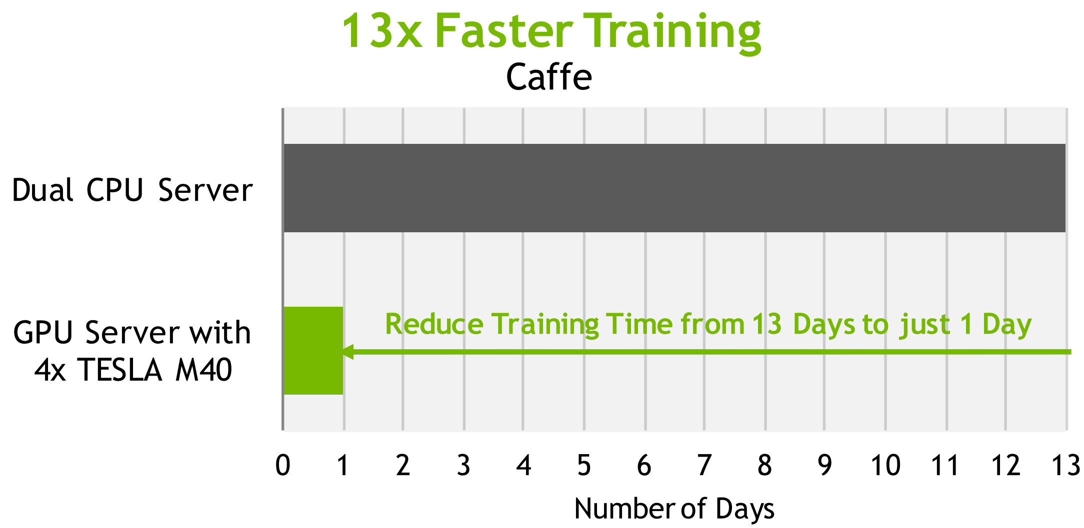
- Multiplies within layer are independent
- Sums are reductions
- Only layers are dependent
- No data dependent operations
  => can be statically scheduled

TESLA M40

World's Fastest Accelerator for Deep Learning Training



28 Gflop/W

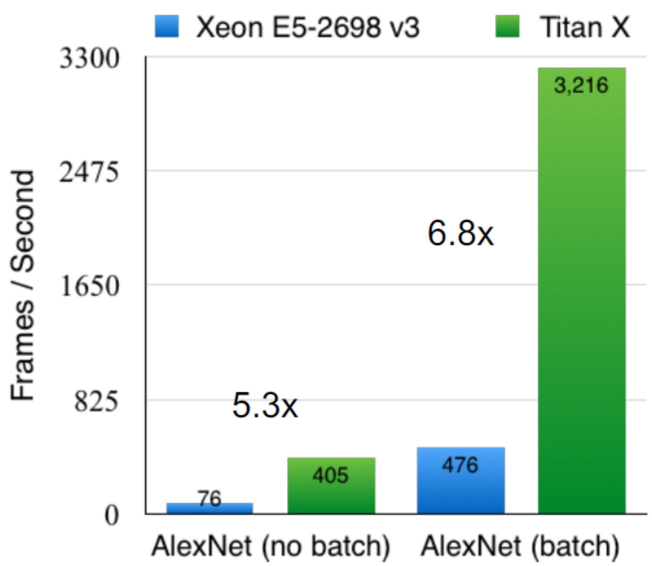


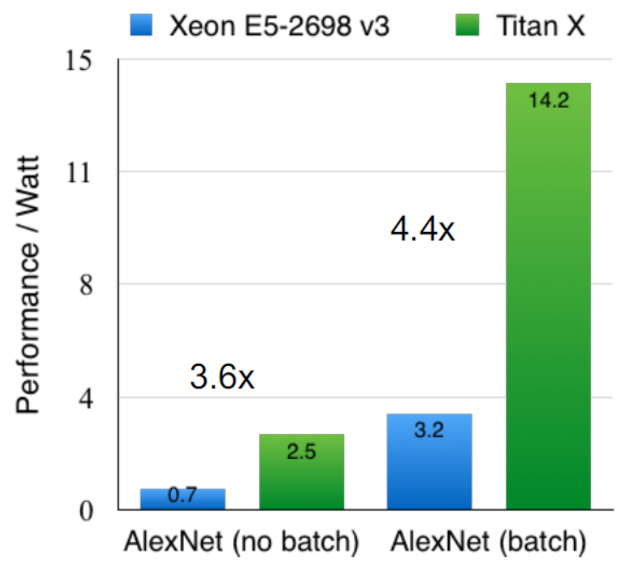
CUDA Cores	3072
Peak SP	7 TFLOPS
GDDR5 Memory	12 GB
Bandwidth	288 GB/s
Power	250W

Note: Caffe benchmark with AlexNet,
CPU server uses 2x E5-2680v3 12 Core 2.5GHz CPU, 128GB System Memory, Ubuntu 14.04

Comparing CPU and GPU – server class

Xeon E5-2698 and Tesla M40





NVIDIA Whitepaper “GPU based deep learning inference: A performance and power analysis.”

DL in practice

The Engine of Modern AI

EDUCATION

- TORCH: NYU, facebook
- CAFFE: Berkeley, University of California
- THEANO: Université de Montréal
- MATCONVNET: University of Oxford
- MOCHA.JL: Massachusetts Institute of Technology
- PURINE: NUS National University of Singapore
- MINERVA: NYU, Microsoft
- MXNET*: University of Washington, Carnegie Mellon University

- BIG SUR: facebook
- TENSORFLOW: Google
- WATSON: IBM
- CNTK: Microsoft

START-UPS

- CHAINER: Preferred Networks
- DL4J: SKYMIND
- KERAS: SCHULTS LABORATORIES
- OPENDEEP: VITRUVIAN

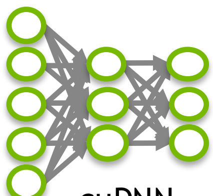
NVIDIA GPU PLATFORM

*U. Washington, CMU, Stanford, TuSimple, NYU, Microsoft, U. Alberta, MIT, NYU Shanghai

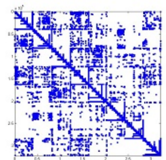
CUDA for Deep Learning Development

DEEP LEARNING SDK

DIGITS

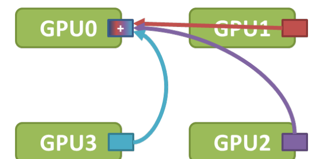


cuDNN



cuSPARSE

cuBLAS



NCCL

TITAN X

DEVBOX

GPU CLOUD

amazon web services™

Microsoft Azure

IBM SOFTLAYER®

阿里云计算 Alibaba Cloud Computing

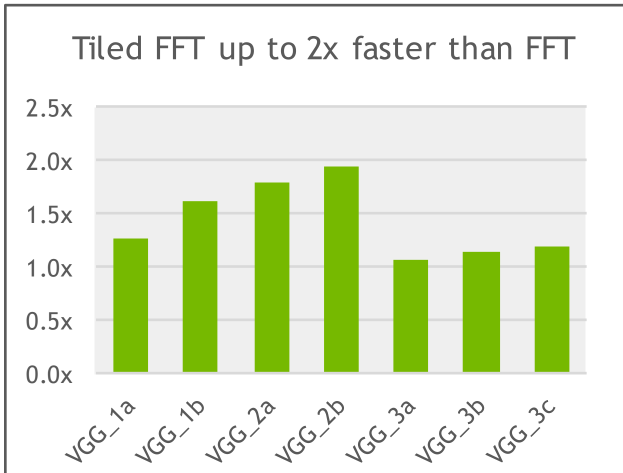
cuDNN

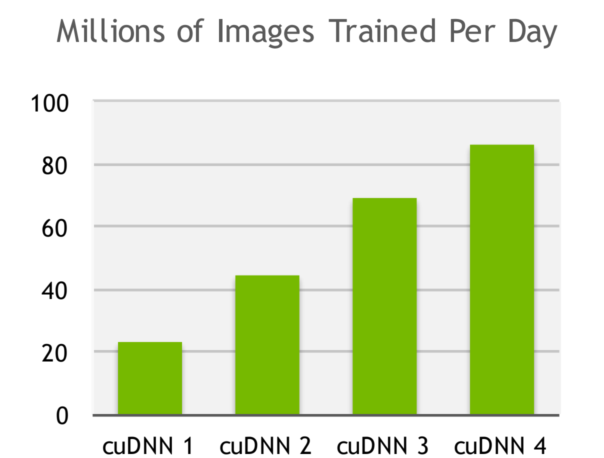
Deep Learning Primitives

Accelerating Artificial Intelligence

developer.nvidia.com/cudnn

- GPU-accelerated Deep Learning subroutines
- High performance neural network training
- Accelerates Major Deep Learning frameworks: Caffe, Theano, Torch, TensorFlow
- Up to 3.5x faster AlexNet training in Caffe than baseline GPU

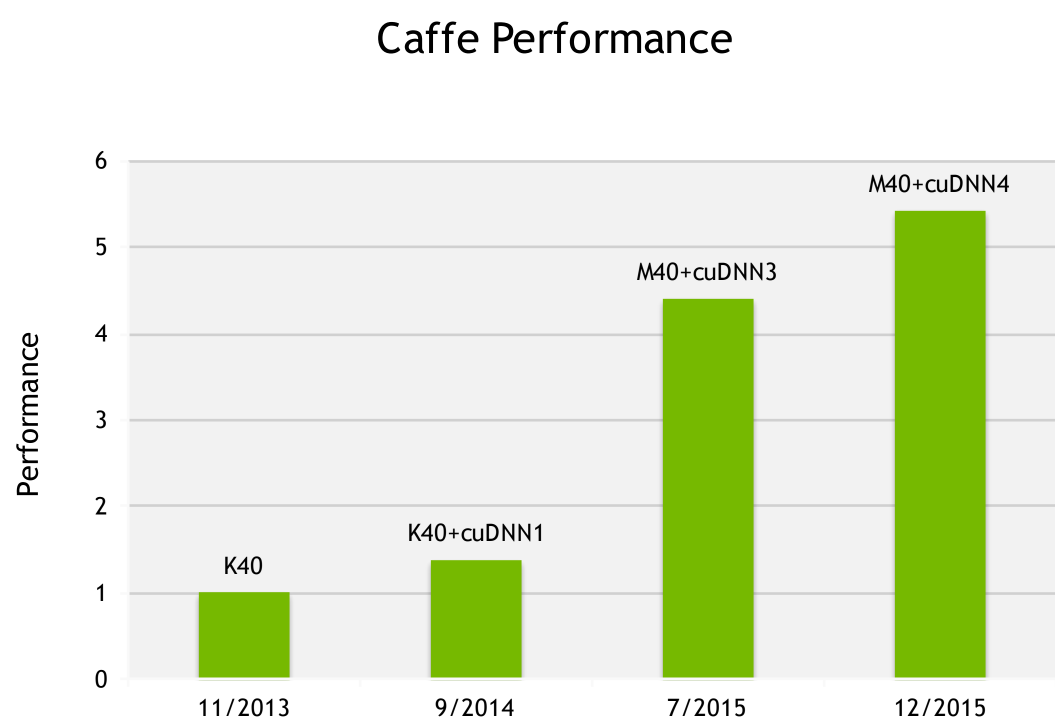




TensorFlow
torch
Caffe
Microsoft CNTK
mxnet
KERAS
theano
Chainer
MINERVA
DL4J Deeplearning4j
MatConvNet

CUDA BOOSTS DEEP LEARNING

5X IN 2 YEARS

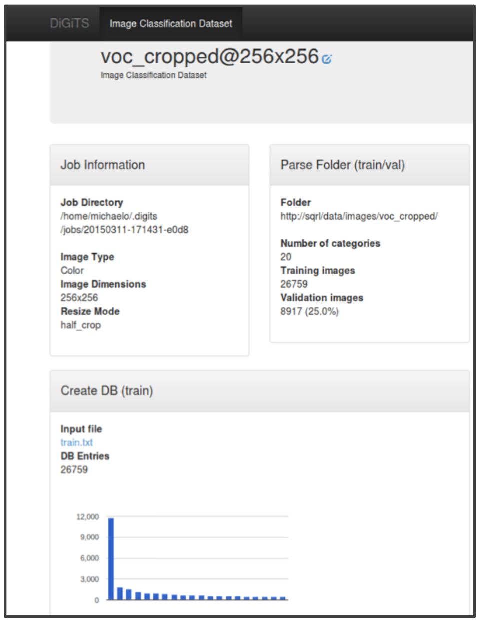


AlexNet training throughput based on 20 iterations,
CPU: 1x E5-2680v3 12 Core 2.5GHz. 128GB System Memory, Ubuntu 14.04

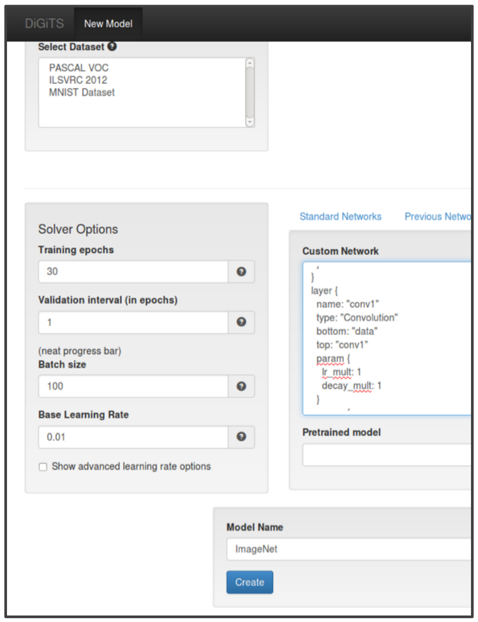
NVIDIA DIGITS

Interactive Deep Learning GPU Training System

Process Data

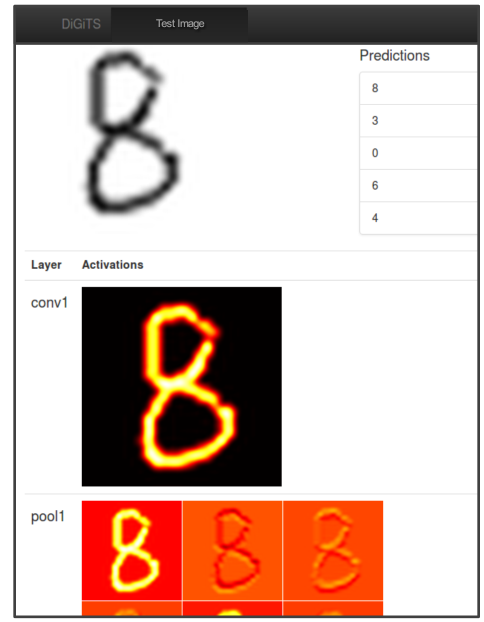


Configure DNN



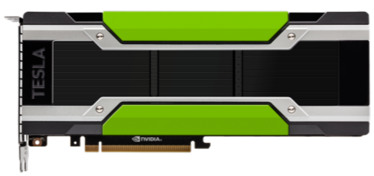
Monitor Progress

Visualize Layers



developer.nvidia.com/digits

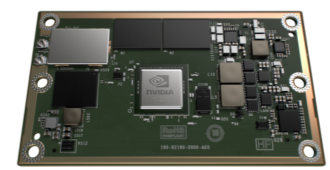
ONE ARCHITECTURE — END-TO-END AI



Tesla
for Cloud

Titan X
for PC

DRIVE PX
for Auto

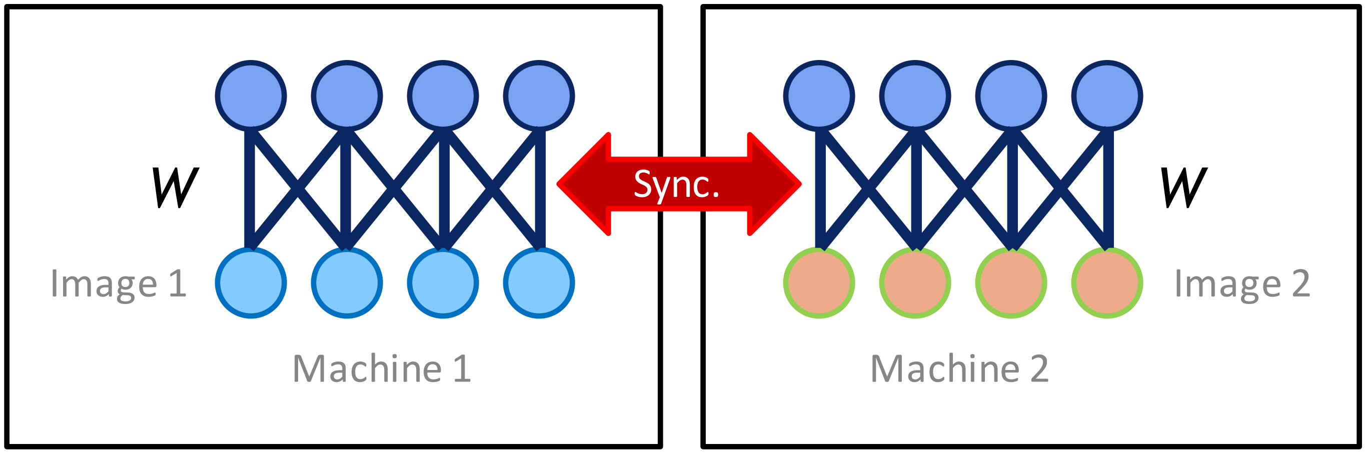


Jetson
for Embedded

Scaling DL

Scaling Neural Networks

Data Parallelism

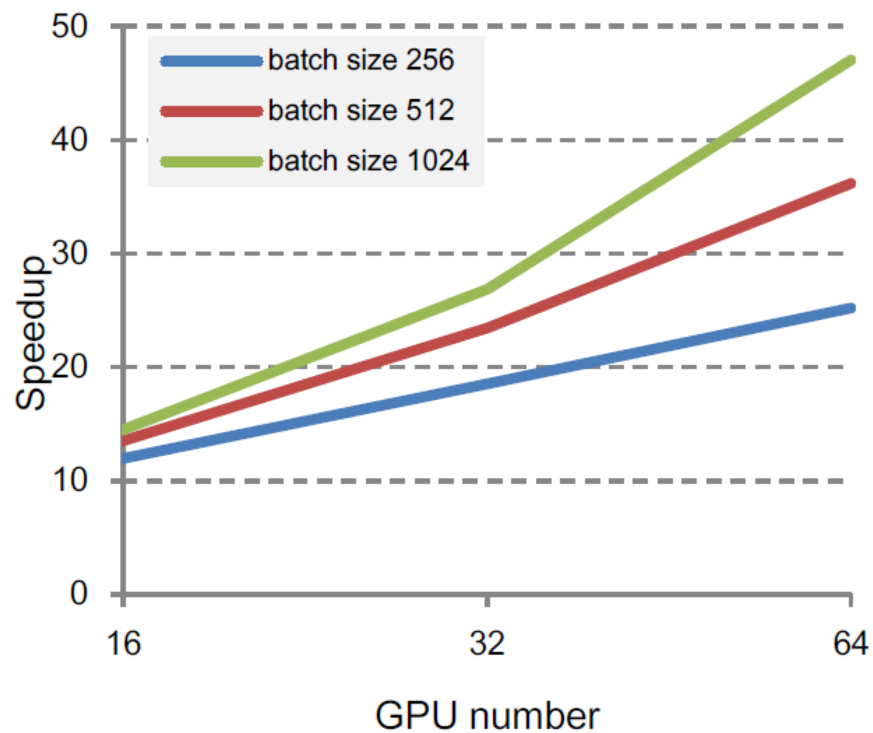


Notes:

Need to sync model across machines.
Largest models do not fit on one GPU.
Requires P-fold larger batch size.
Works across many nodes – parameter server approach – linear speedup.

Multiple GPUs

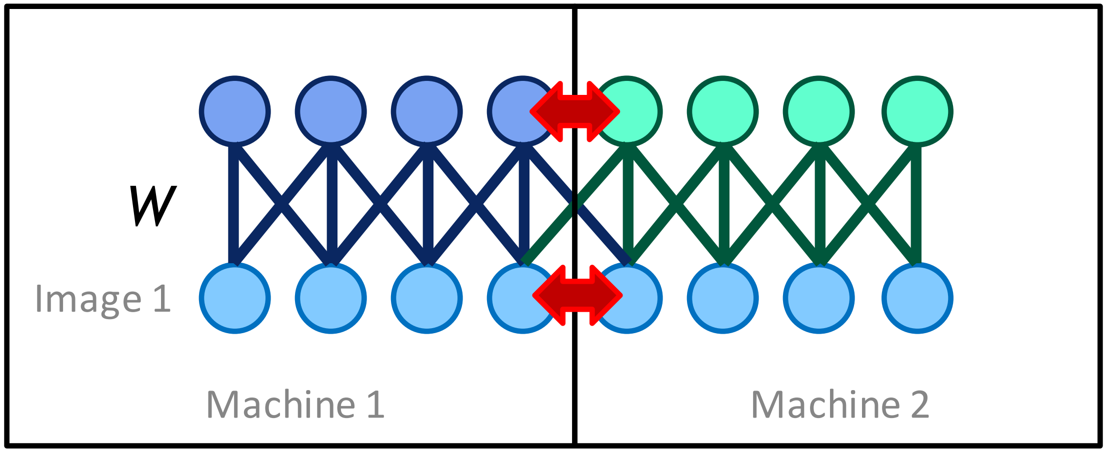
Near linear scaling – data parallel.



Ren Wu et al, Baidu, “Deep Image: Scaling up Image Recognition.” arXiv 2015

Scaling Neural Networks

Model Parallelism



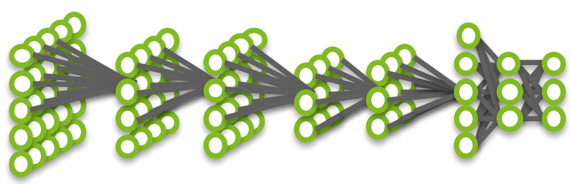
Notes:

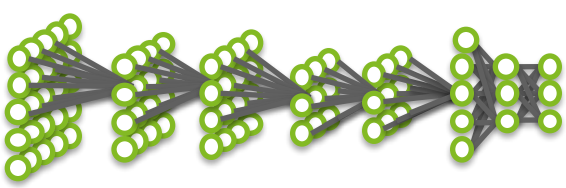
Allows for larger models than fit on one GPU.
Requires much more frequent communication between GPUs.
Most commonly used within a node – GPU P2P.
Effective for the fully connected layers.

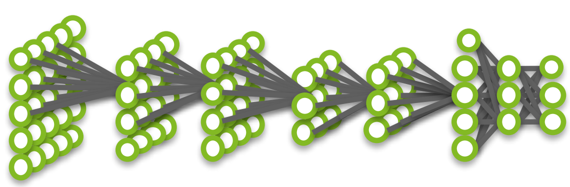
Scaling Neural Networks

Hyper Parameter Parallelism

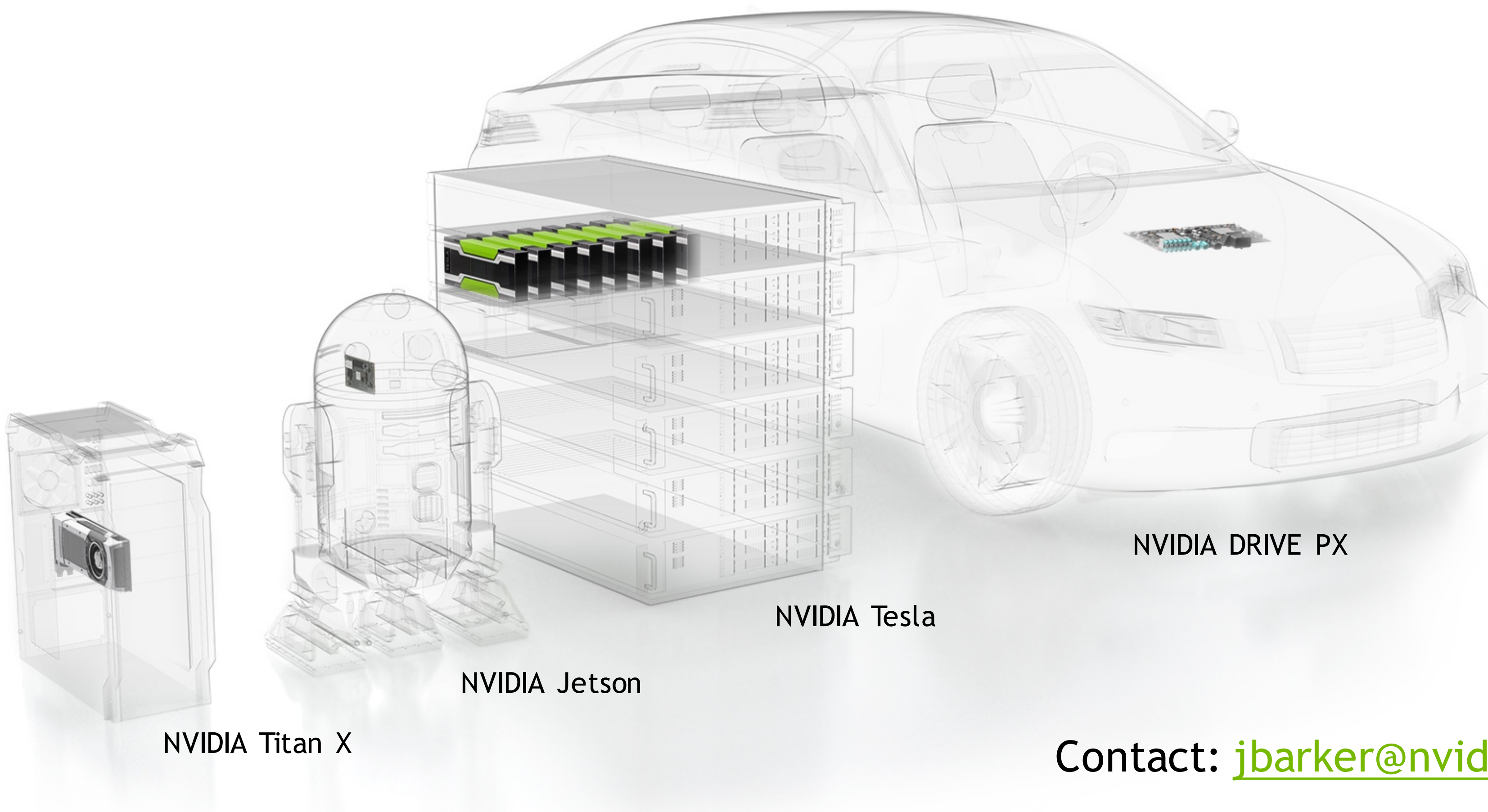
Try many alternative neural networks in parallel – on different CPU / GPU / Machines.
Probably the most obvious and effective way!







Deep Learning Everywhere



Contact: jbarker@nvidia.com