Recent Trend in GPU Computing: Deep Learning

7 Oktober 2016 – FMIPA, Universitas Andalas

Muhammad Teguh Satria, MSc

NovaGlobal Pte Ltd
Green & Scientific Computing Solutions

NVIDIA PREFERRED SOLUTION PROVIDER

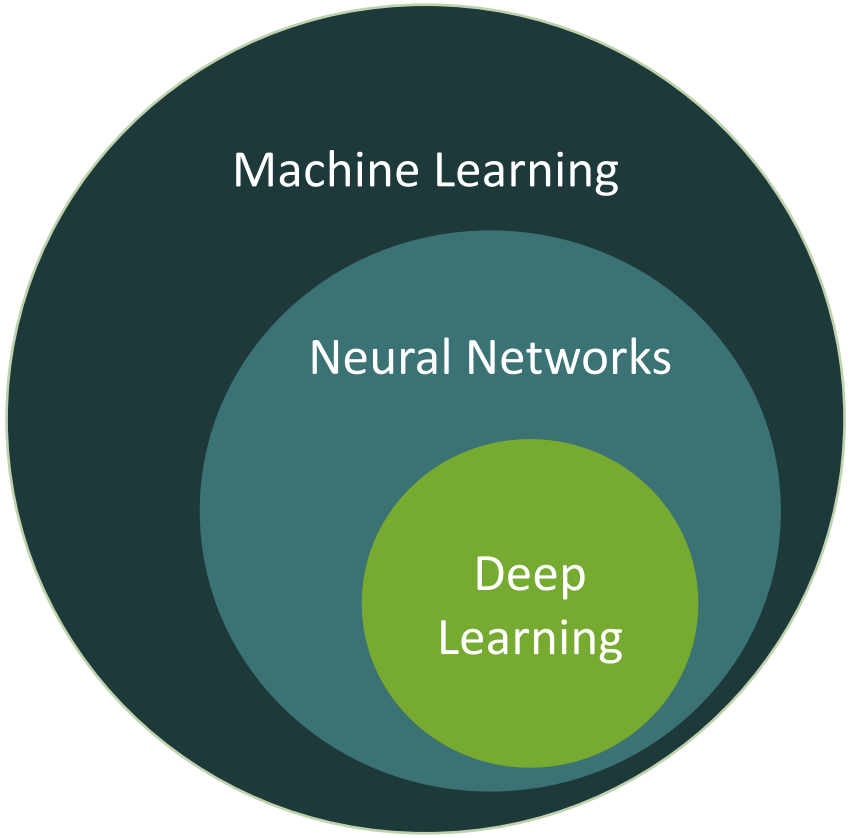
Outline

- What is Deep Learning ?
- GPU and Deep Learning
- Deep Learning Frameworks

Acknowledgement: most of slide contents credit to NVIDIA

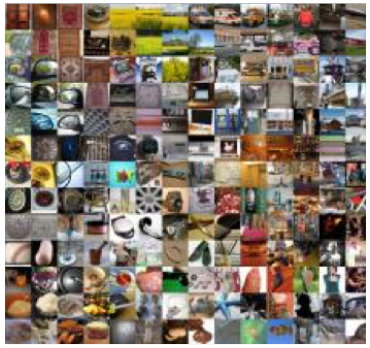
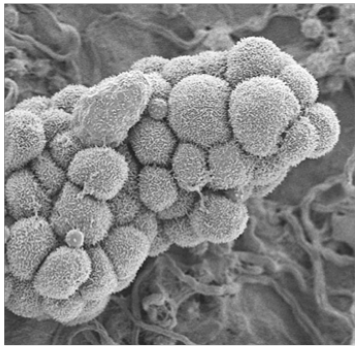
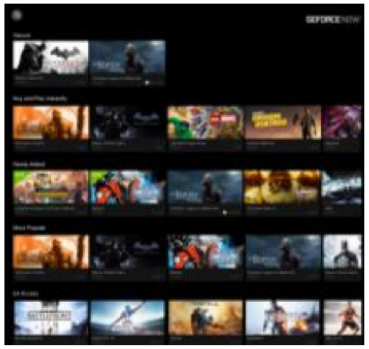
What is Deep Learning ?

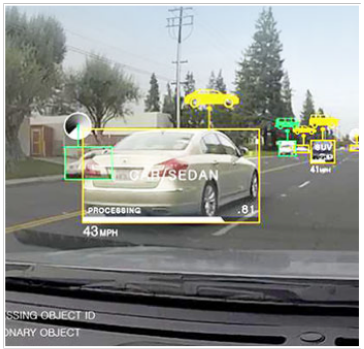
- Deep learning is a subset of machine learning that refers to artificial neural networks that are composed of many layers.
- Neural Networks inspired by human brain.



Where it's applied ?

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

Artificial Intelligence Case - Dataset

"dog"

"cat"

Images source: http://awesomealgorithm.blogspot.sg/2015/08/machine-learning-image-detection-cats.html

Artificial Intelligence Case - Prediction

What is this ?

Traditional Approach

Raw Images

Feature Extraction

Needs hand-tuning of feature parameters

SVM, shallow neural net, …

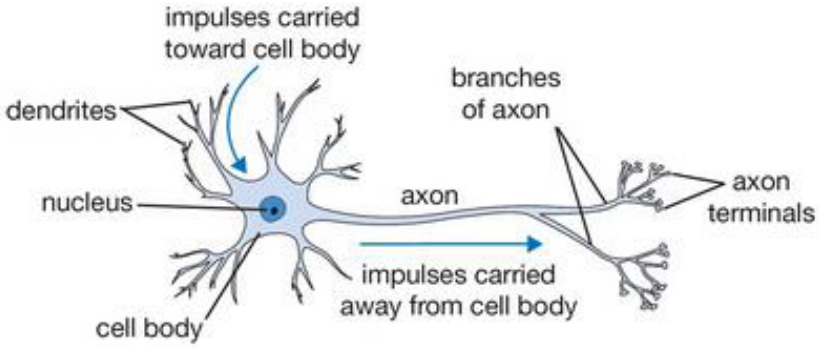
Classifier / Detector

Result

Deep Learning Approach

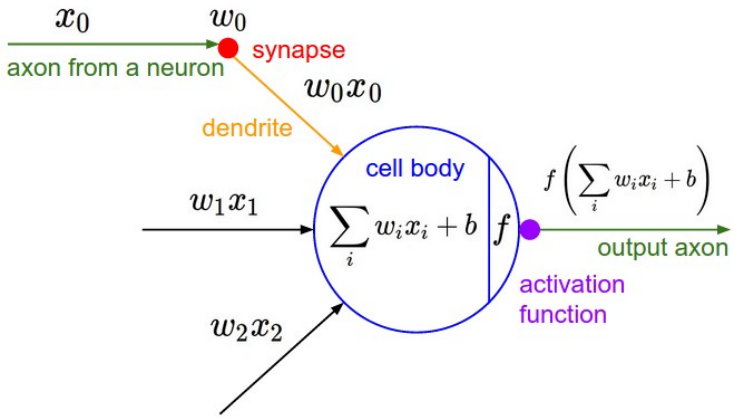
- Artificial Neural Networks inspired by human brain.

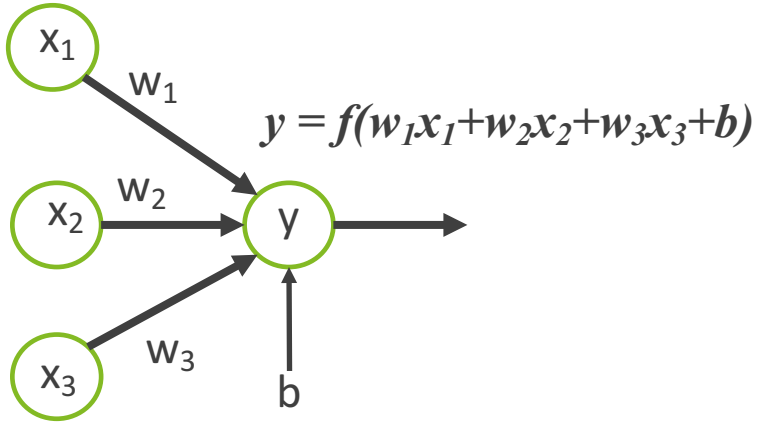
Biological Neuron



From stanford cs231n lecture notes

Artificial Neuron

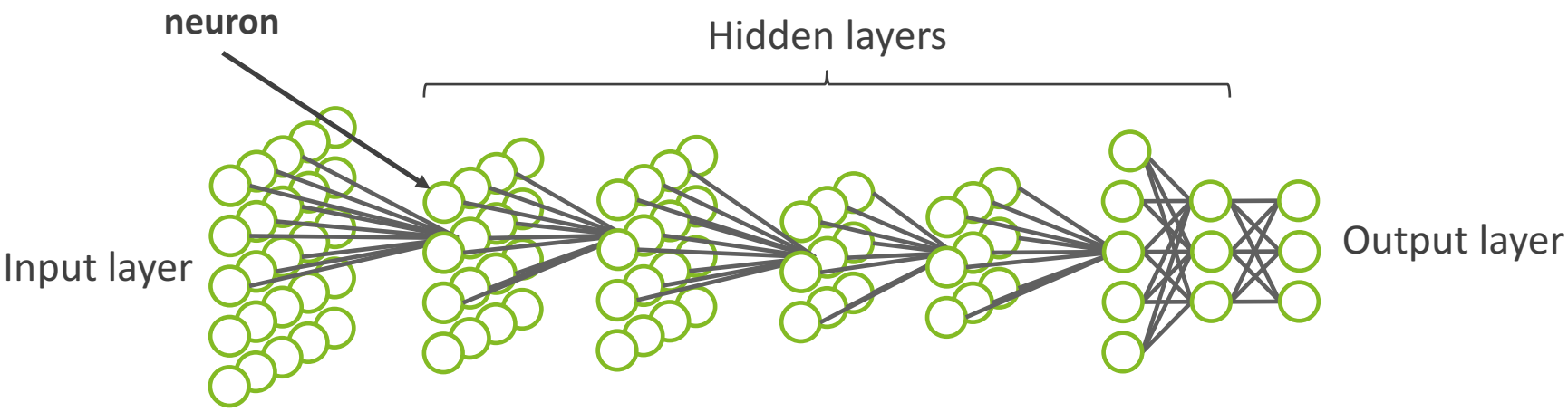




Artificial neuron = a simple mathematical unit

Deep Learning Approach

- Neural Network: A collection of simple, trainable mathematical units that collectively learn complex functions



- Deep means many hidden layers
- More layers tend to give better accuracy

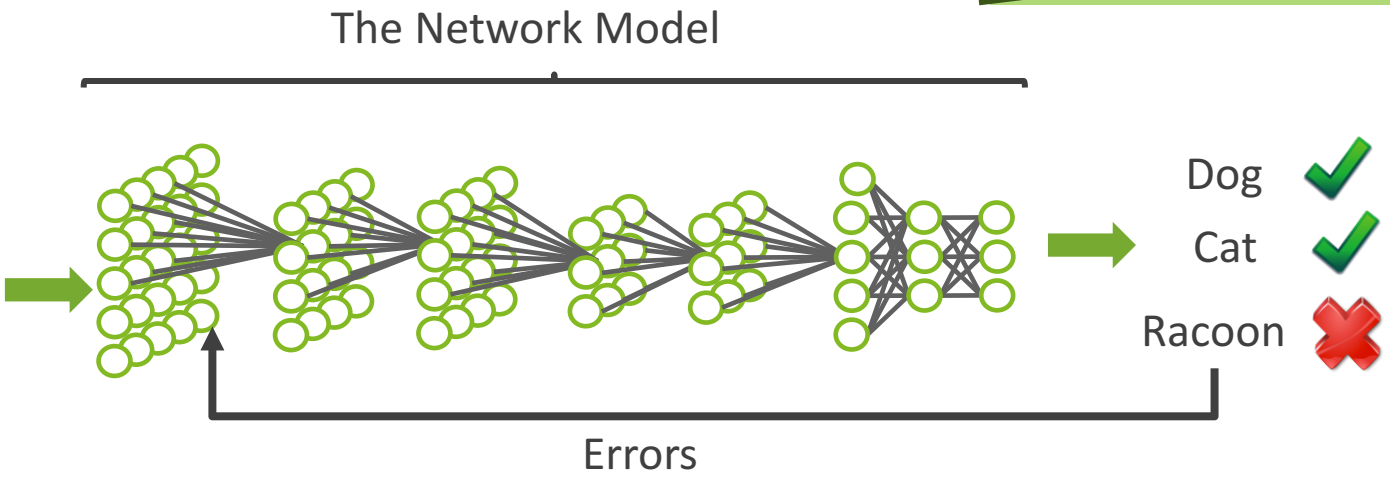
Deep Learning Approach

Train Dataset:

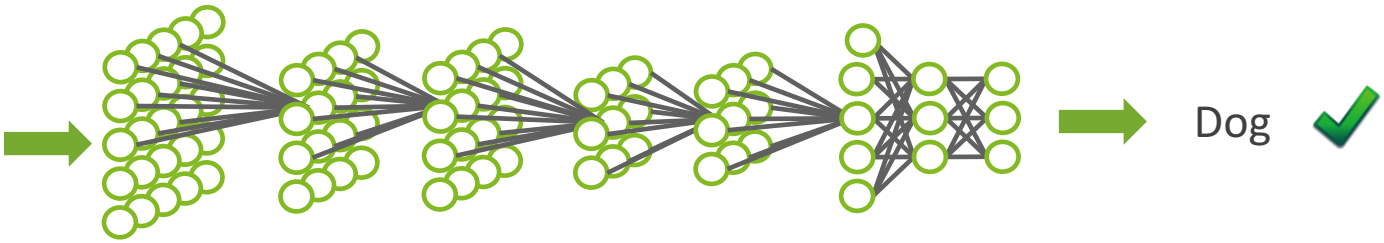
dog

cat

honey badger



Deploy The Network Model:



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

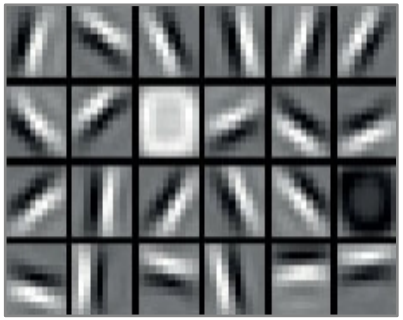
Three Kinds of Network Model

- DNN: Deep Neural Network
  - All fully connected layers
- CNN: Convolutional Neural Network
  - Some convolutional layers
- RNN: Recurrent Neural Network
  - Connections between units form a directed cycle

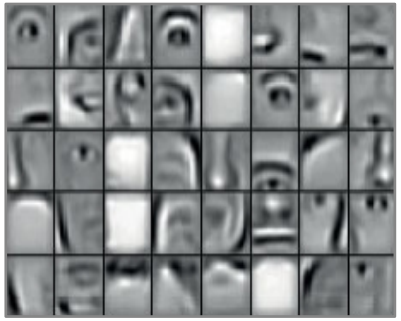
Deep Neural Network

Raw data

Low-level features

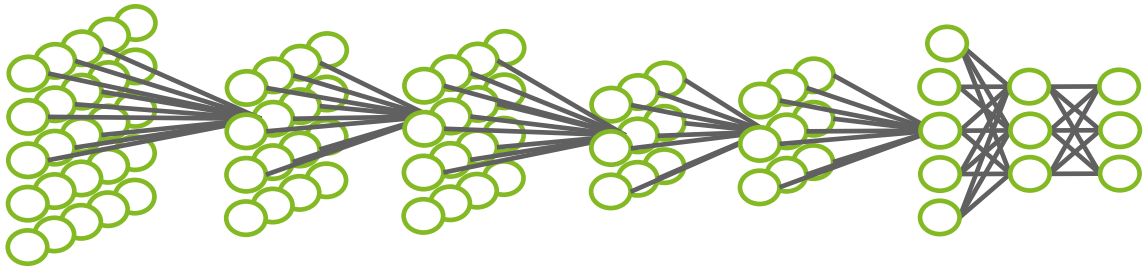


Mid-level features



High-level features

input



result

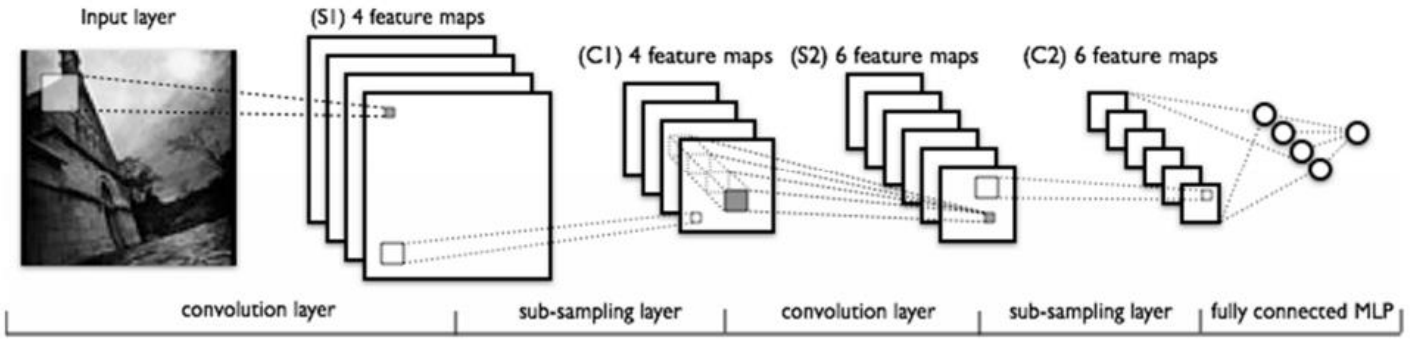
Typical Network

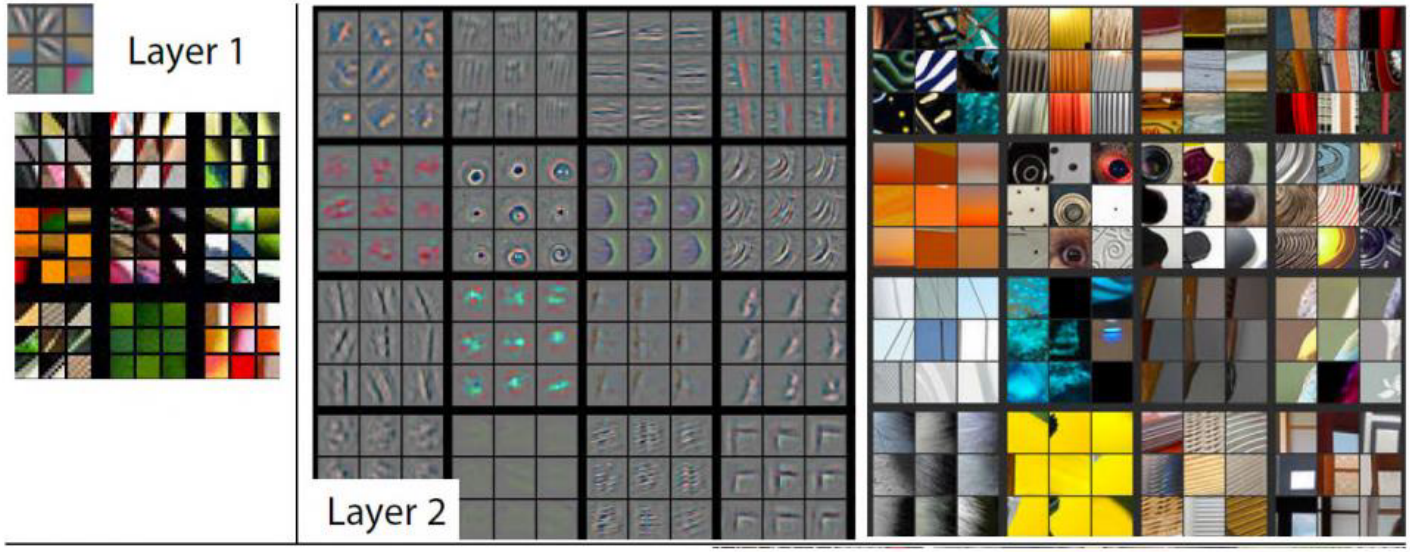
Task objective
e.g. Identify face
Training data
10-100M images
Network architecture
10 layers
1B parameters
Learning algorithm
~30 Exaflops
~30 GPU days

Convolutional Neural Network

- Inspired by the human visual cortex
- Learns a hierarchy of visual features
- Local pixel level features are scale and translation invariant
- Learns the “essence” of visual objects and generalizes well

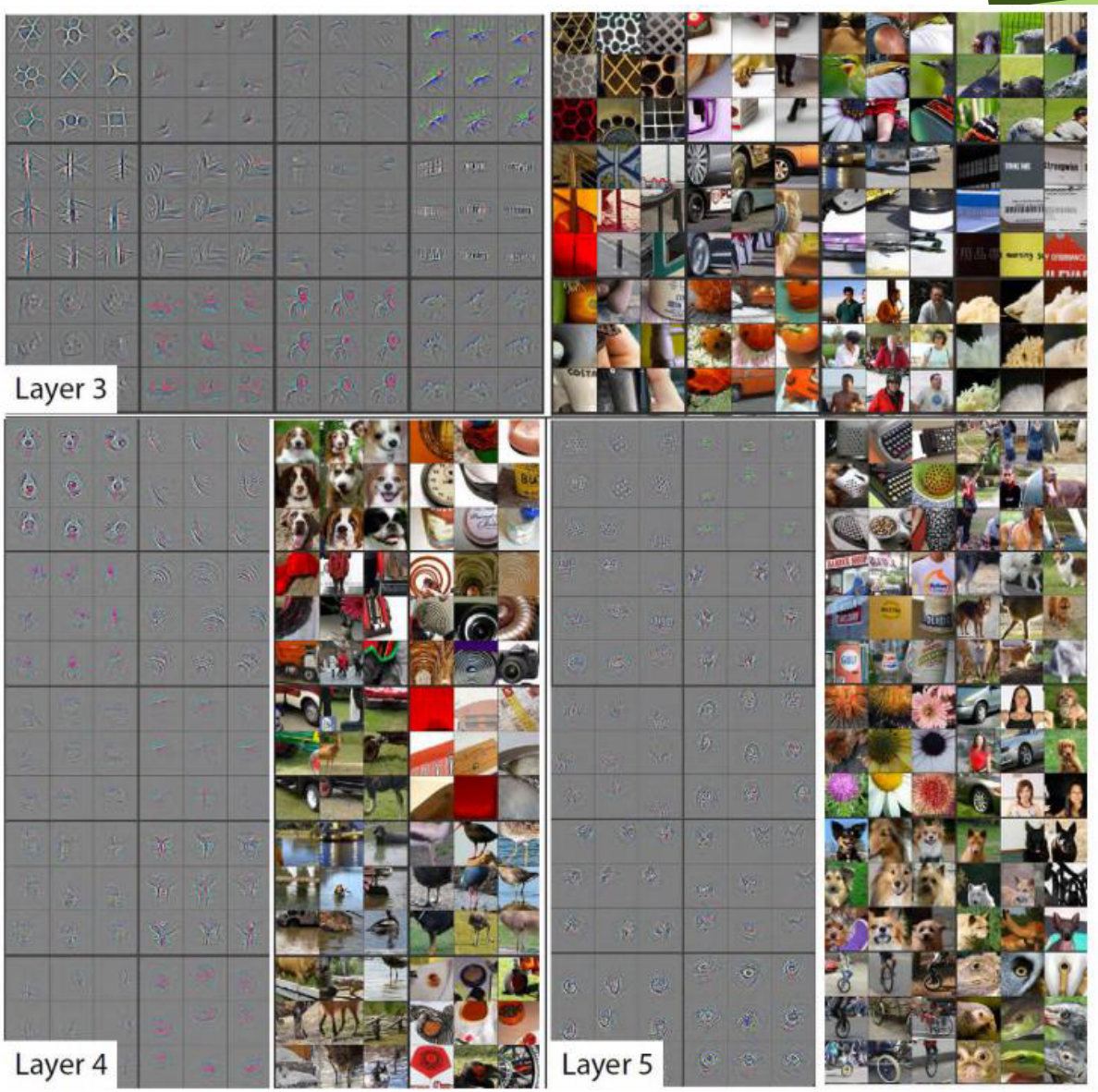
M.D. Zeiler, R. Fergus, "Visualizing and Understanding Convolutional Networks", ECCV 2014



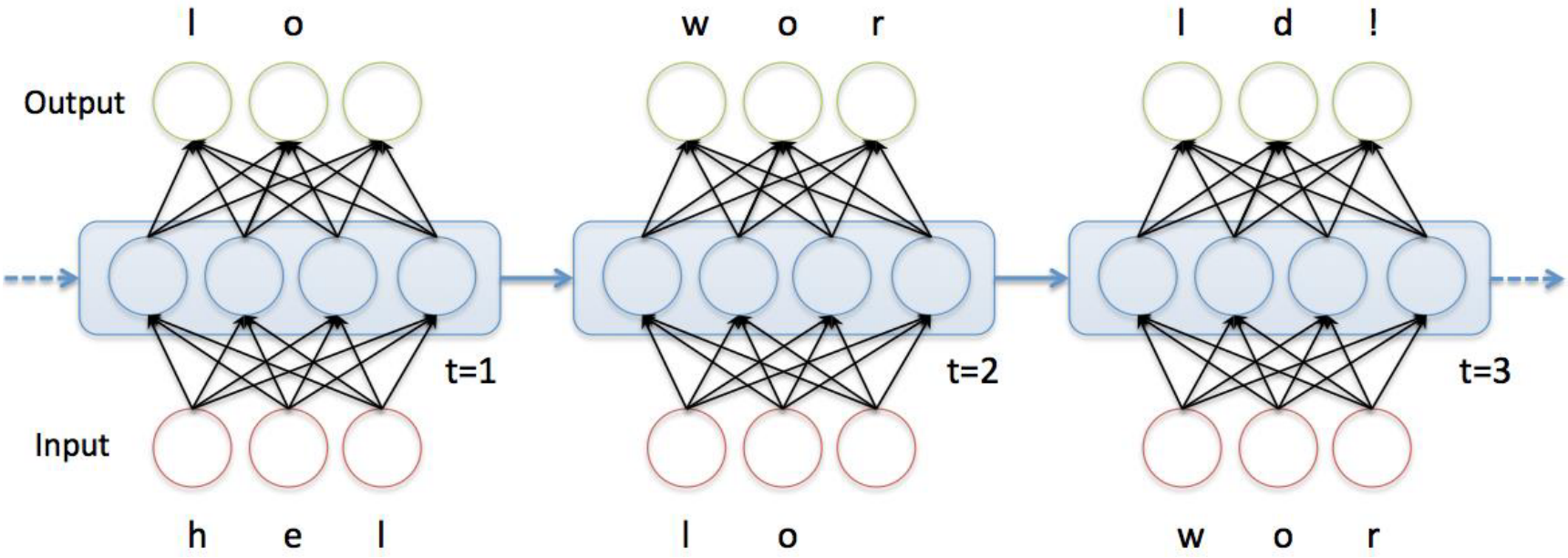


[Matthew Zeiler & Rob Fergus]

Convolutional Neural Network



Recurrent Neural Network



DNNs Dominate in Perceptual Tasks

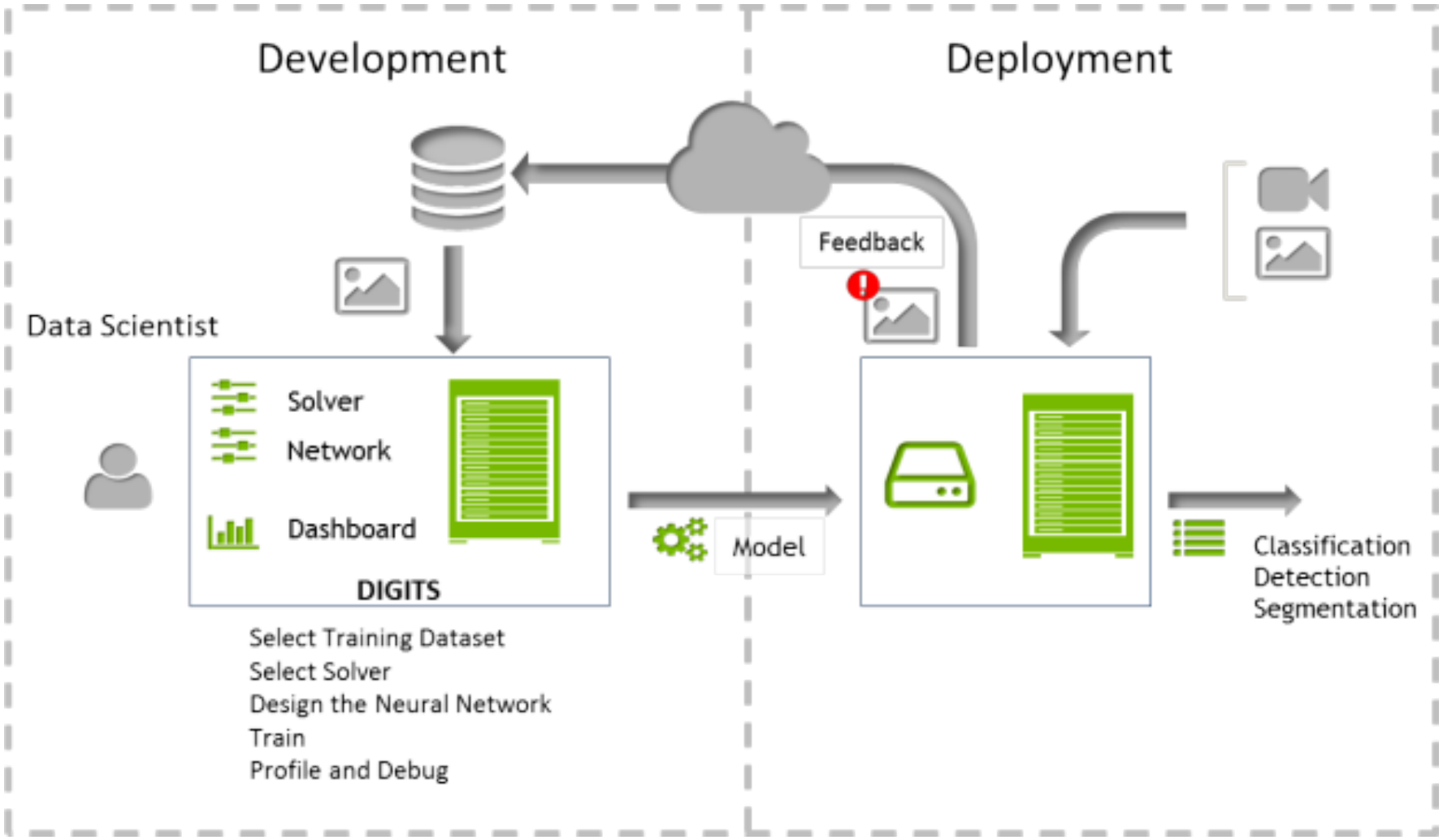
- Handwriting recognition MNIST (many), Arabic HWX (IDSIA)
- OCR in the Wild [2011]: StreetView House Numbers (NYU and others)
- Traffic sign recognition [2011] GTSRB competition (IDSIA, NYU)
- Asian handwriting recognition [2013] ICDAR competition (IDSIA)
- Pedestrian Detection [2013]: INRIA datasets and others (NYU)
- Volumetric brain image segmentation [2009] connectomics (IDSIA, MIT)
- Human Action Recognition [2011] Hollywood II dataset (Stanford)
- Object Recognition [2012] ImageNet competition (Toronto)
- Scene Parsing [2012] Stanford bgd, SiftFlow, Barcelona datasets (NYU)
- Scene parsing from depth images [2013] NYU RGB-D dataset (NYU)
- Speech Recognition [2012] Acoustic modeling (IBM and Google)
- Breast cancer cell mitosis detection [2011] MITOS (IDSIA)

Slide credit: YannLecun, Facebook & NYU

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

Deep Learning Development Cycle



AlphaGo

Google DeepMind

First Computer Program to Beat a Human Go Professional

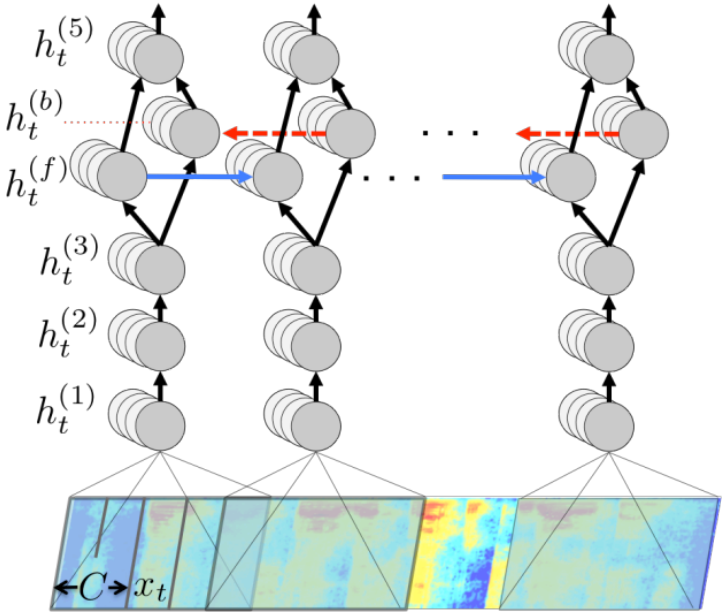
- Training DNNs: 3 weeks, 340 million training steps on 50 GPUs
- Play: Asynchronous multi-threaded search
  - Simulations on CPUs, policy and value DNNs in parallel on GPUs
  - Single machine: 40 search threads, 48 CPUs, and 8 GPUs
  - Distributed version: 40 search threads, 1202 CPUs and 176 GPUs

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

Baidu Deep Speech 2

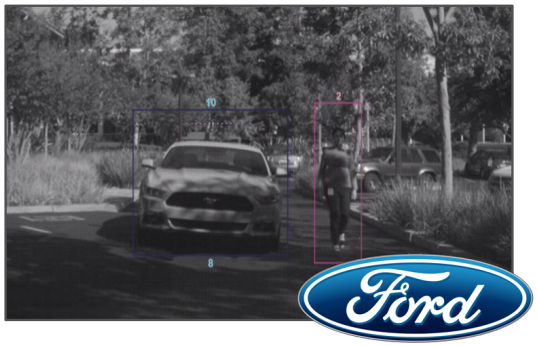
End-to-end Deep Learning for English and Mandarin Speech Recognition

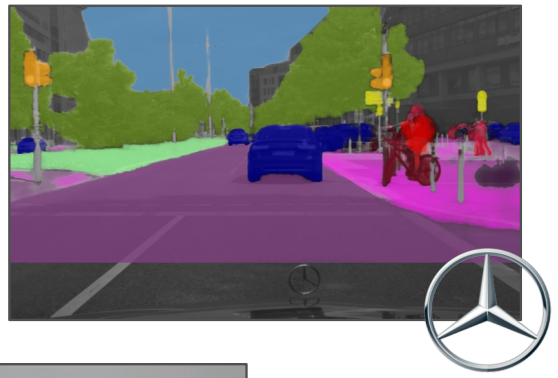
- English and Mandarin speech recognition
- Transition from English to Mandarin made simpler by end-to-end DL
  - No feature engineering or Mandarin-specifics required
- More accurate than humans
  - Error rate 3.7% vs. 4% for human tests
- http://svail.github.io/mandarin/
- http://arxiv.org/abs/1512.02595

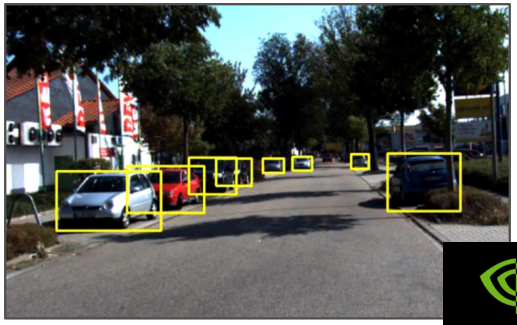


Autonomous Vehicles

Audi







NVIDIA

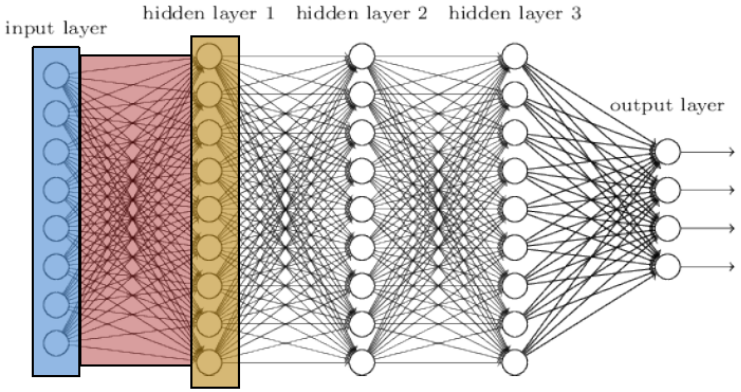
Audi

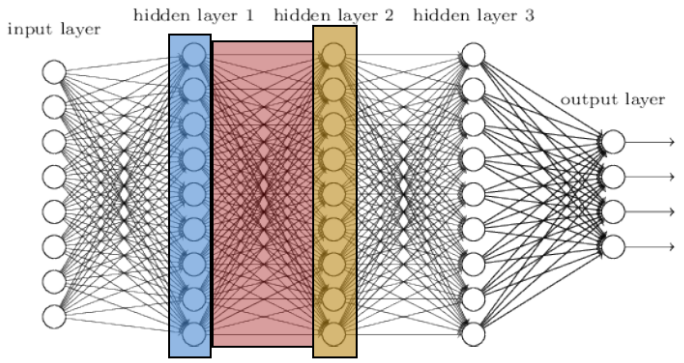
GPU and Deep Learning

Why Are GPUs Good For Deep Learning ?

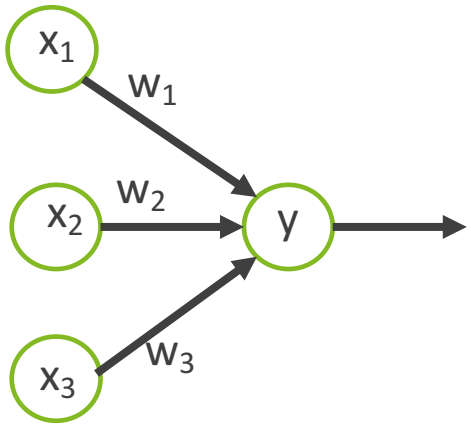
Why Are GPUs Good For Deep Learning ?

- Matrix calculation

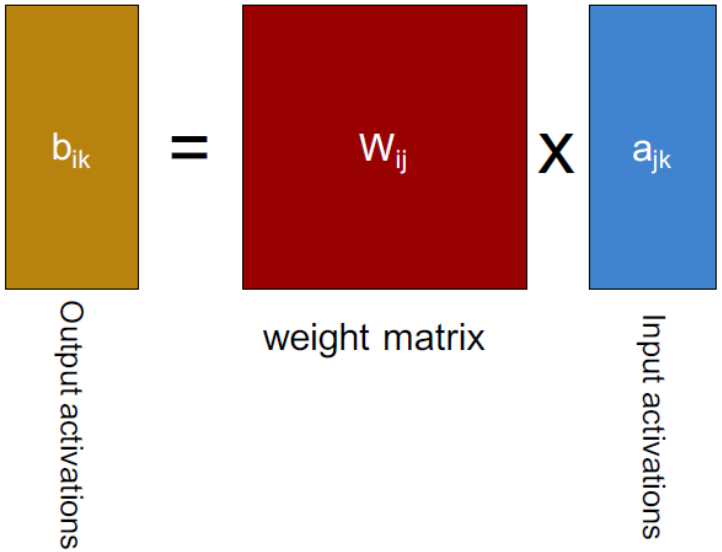




Repeat for each layer



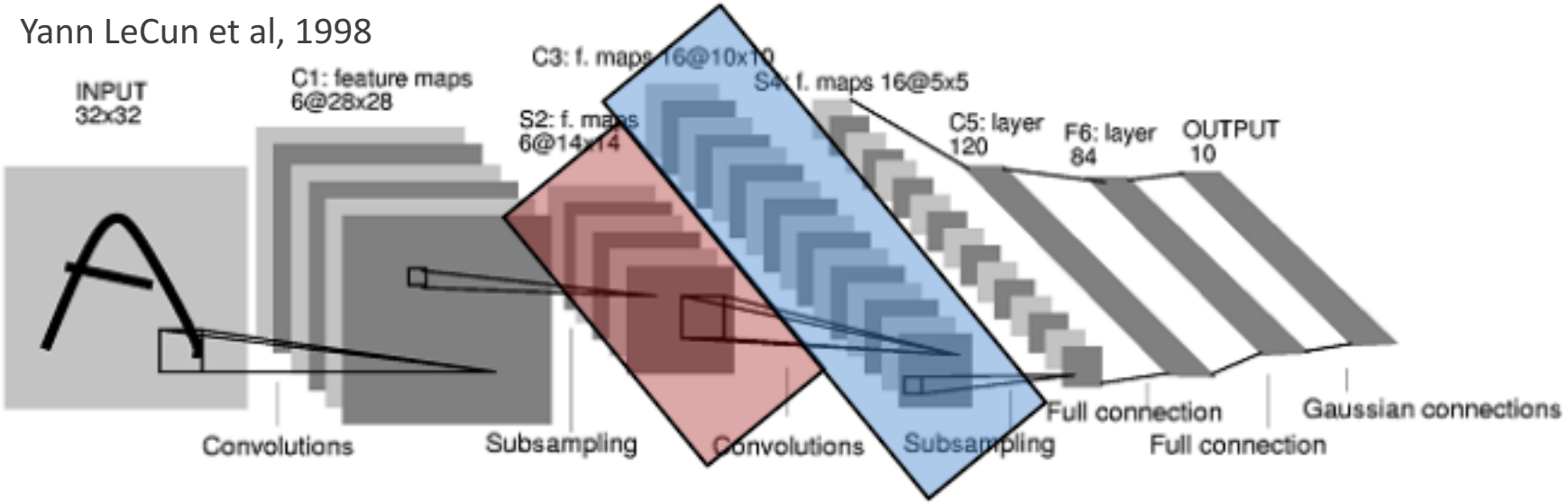
$$y = f(w_1x_1 + w_2x_2 + w_3x_3)$$



Why Are GPUs Good For Deep Learning ?

- Lots of Parallelism Available

Yann LeCun et al, 1998

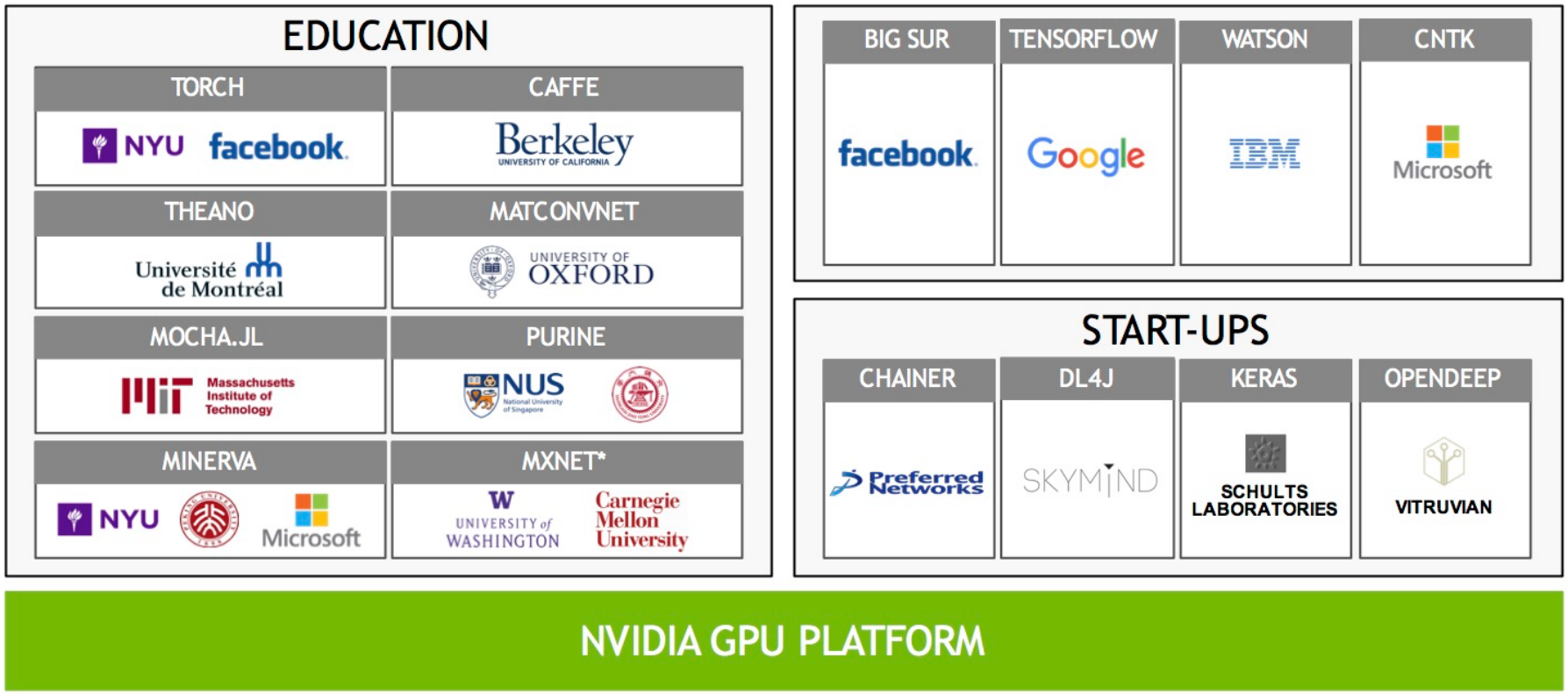


- Inputs
- Points of a feature map
- Filters
- Elements within a filter

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

Deep Learning Frameworks

The Engine of Modern AI

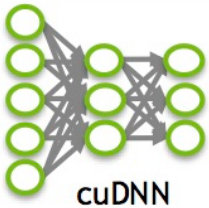


*U. Washington, CMU, Stanford, TuSimple, NYU, Microsoft, U. Alberta, MIT, NYU Shanghai

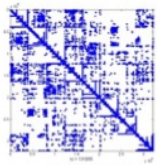
CUDA for Deep Learning Development

DEEP LEARNING SDK

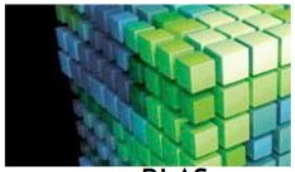
DIGITS



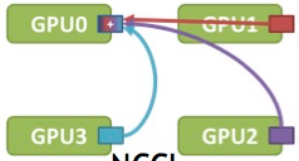
cuDNN



cuSPARSE



cuBLAS



NCCL

TITAN X

DEVBOX

GPU CLOUD

amazon web services™

Microsoft Azure

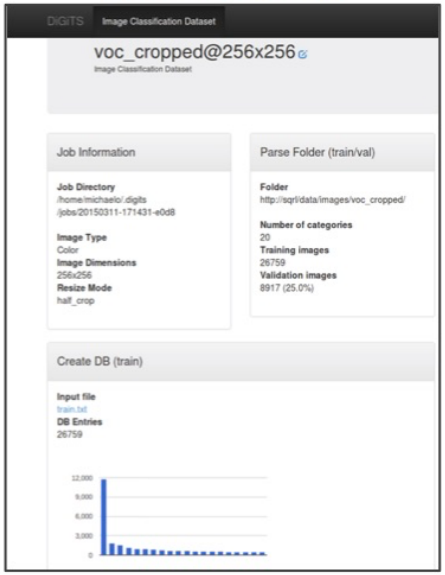
IBM SOFTLAYER

阿里云计算 Alibaba Cloud Computing

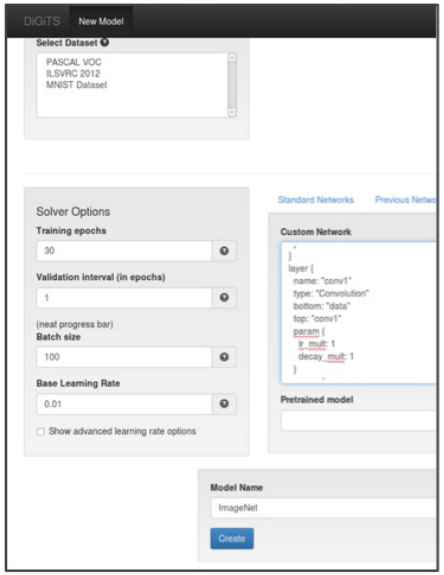
DIGITS

Interactive Deep Learning GPU Training System

Process Data

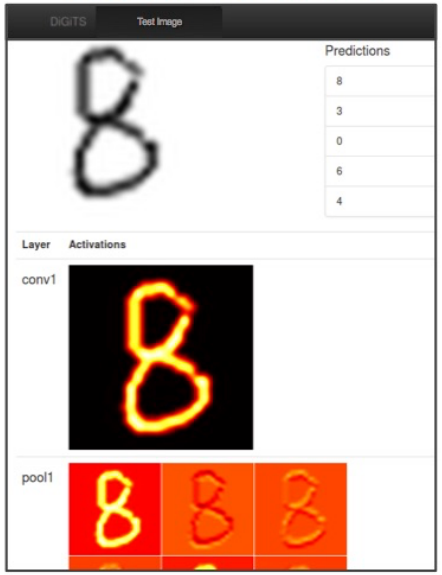


Configure DNN



Monitor Progress

Visualize Layers



http://developer.nvidia.com/digits

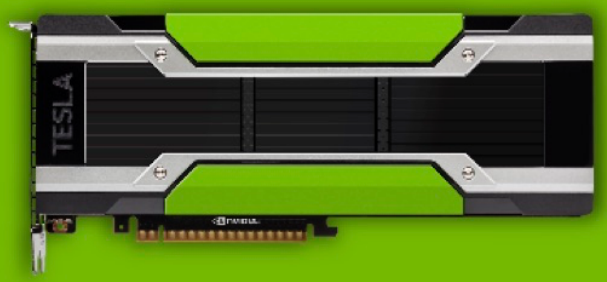
DIGITS DevBox

- Four TITAN X GPUs with 12GB of memory per GPU
- 64GB DDR4
- Asus X99-E WS motherboard with 4-way PCI-E Gen3 x16 support
- Core i7-5930K 6 Core 3.5GHz desktop processor
- Three 3TB SATA 6Gb 3.5” Enterprise Hard Drive in RAID5
- 512GB PCI-E M.2 SSD cache for RAID
- 250GB SATA 6Gb Internal SSD
- 1600W Power Supply Unit
- Ubuntu 14.04
- NVIDIA-qualified driver
- NVIDIA® CUDA® Toolkit 7.0
- NVIDIA® DIGITS™ SW
- Caffe, Theano, Torch, BIDMach

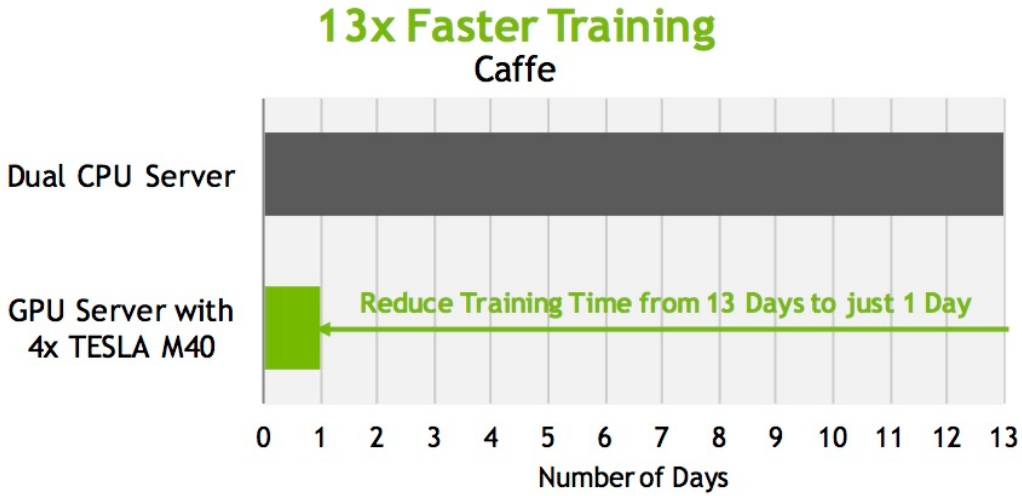
Tesla M40 for Deep Learning Training

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

DGX-1

World's First Deep Learning Supercomputer

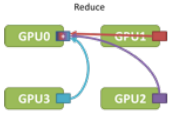
- Engineered for deep learning
- 170 TF FP16
- 8x Tesla P100 in hybrid cube mesh
- NVLink technology
- Accelerates major AI frameworks

DGX-1 Software Stack

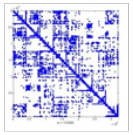
One Product: Two Use Modes – Base OS & Cloud Managed

Accelerated Deep Learning

cuDNN

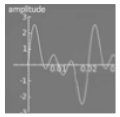


NCCL



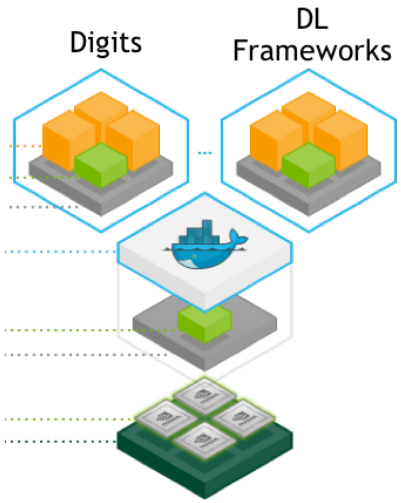
cuSPARSE

cuBLAS



cuFFT

Container Based Applications



NVIDIA Cloud Management

AI Researchers ⟷ Enterprise Data Scientists

Looking for Deep Learning Solutions ?

support@novaglobal.com.sg

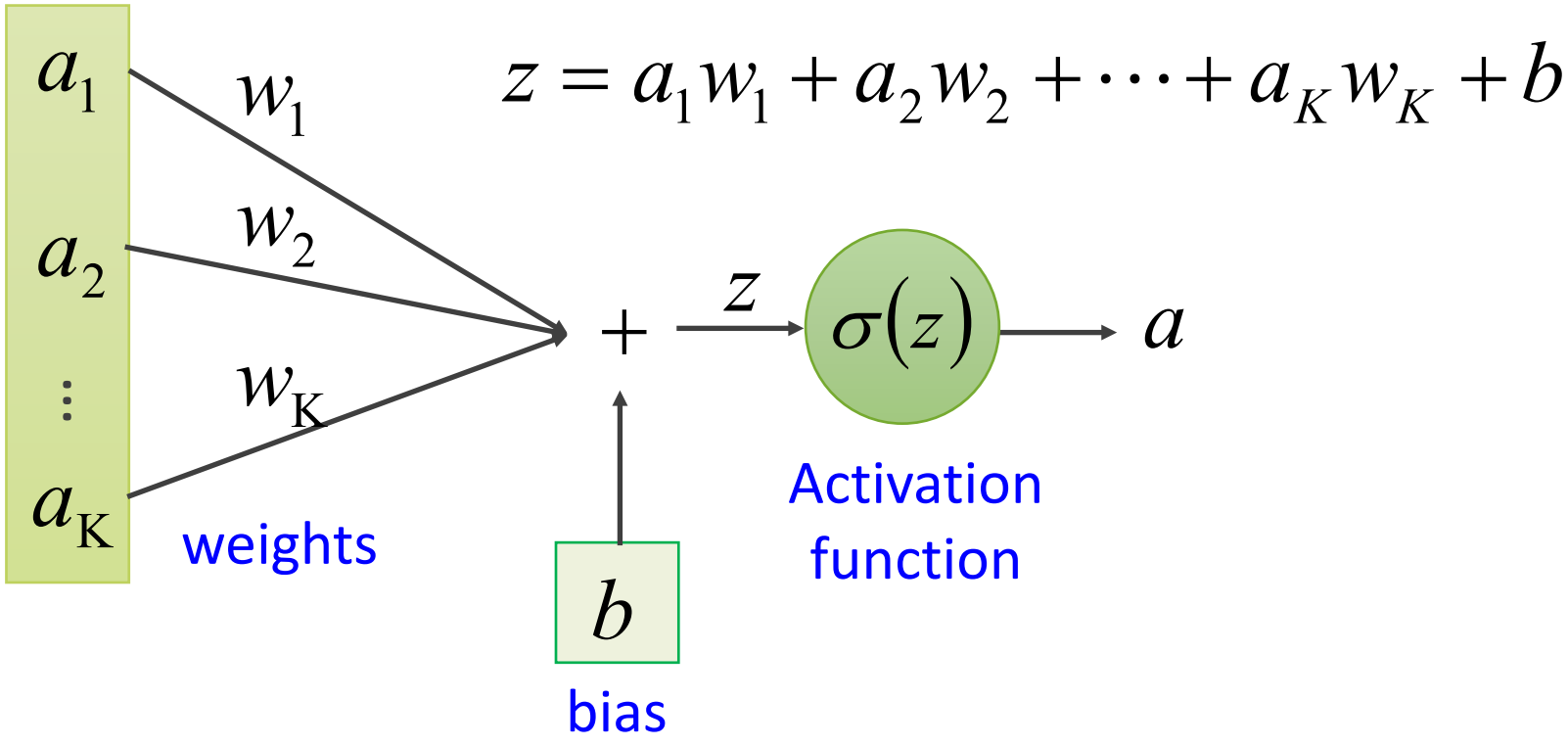
NovaGlobal Pte Ltd
Green & Scientific Computing Solutions
http://novaglobal.com.sg

- Established in 1996
- Operating in ASEAN region
  - Primarily in Singapore, Malaysia, Indonesia, Thailand and Vietnam
- Solution provider for
  - High Performance Computing
  - Accelerated GPU Computing
  - Cloud/Virtualization
- Platform-Independent
- Partnering with Technology Leaders

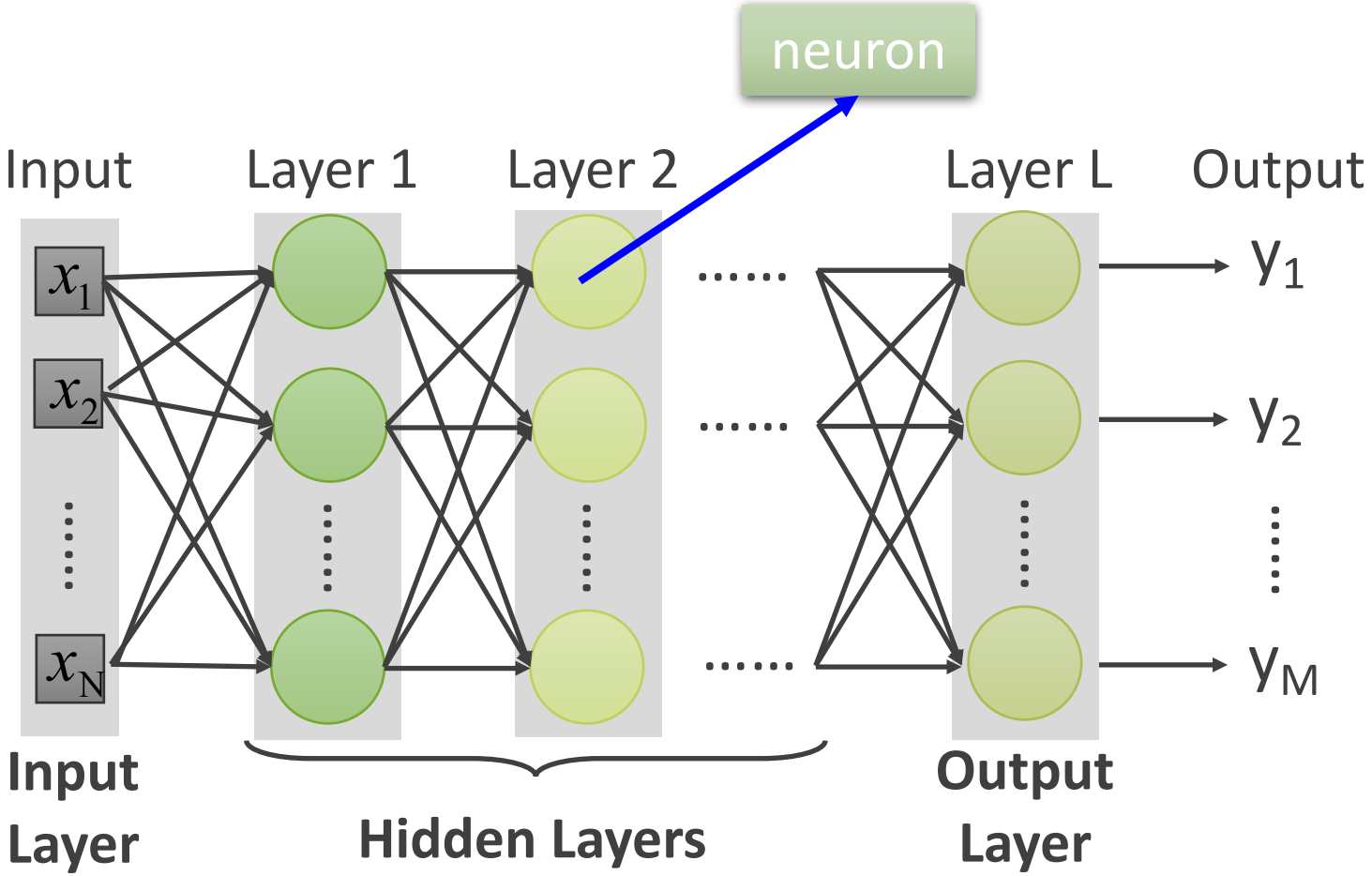
Thank you for your attention !

Element of Neural Network

Neuron $f: R^K \rightarrow R$

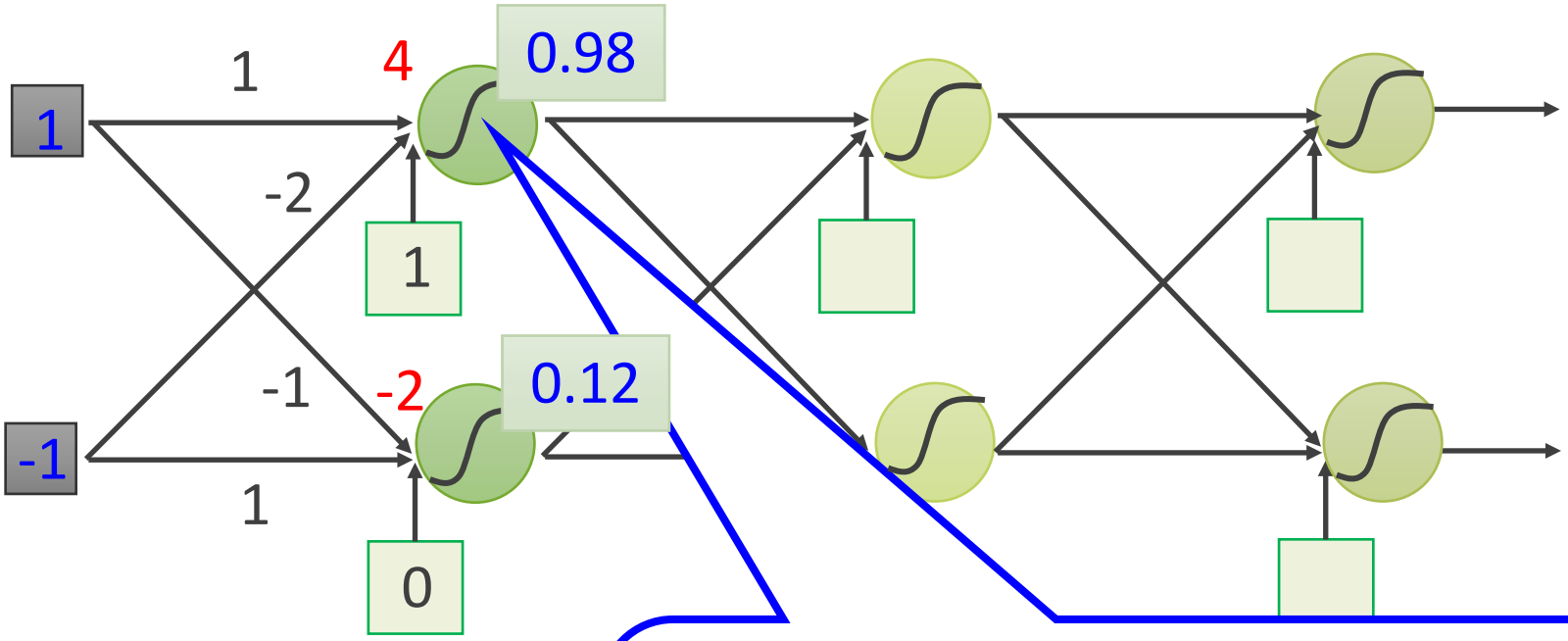


Neural Network



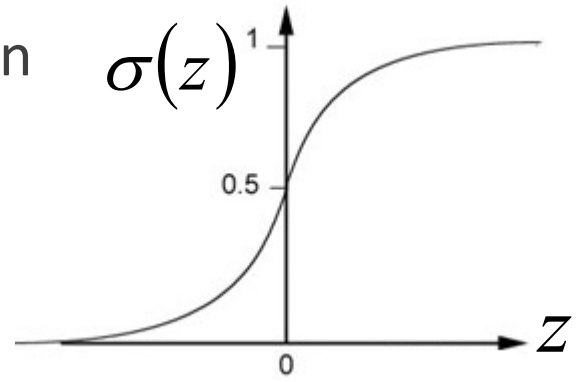
Deep means many hidden layers

Example of Neural Network

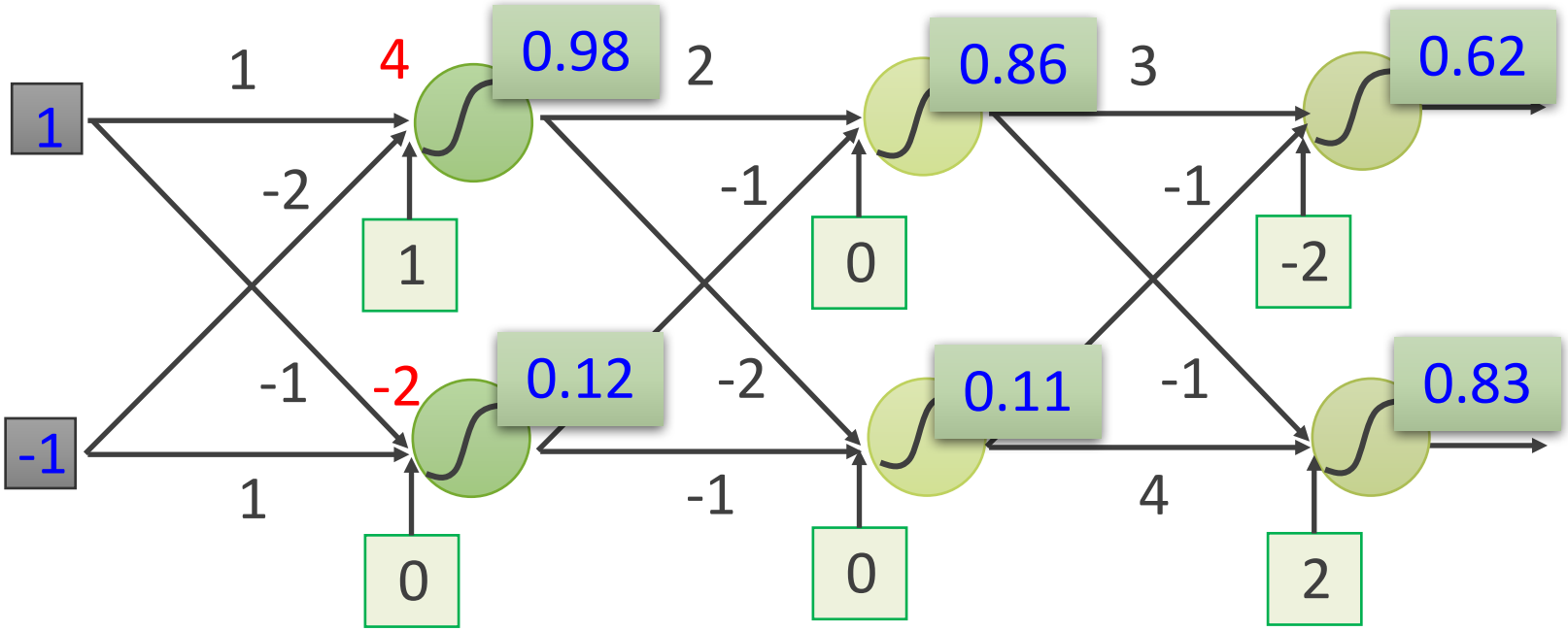


Sigmoid Function

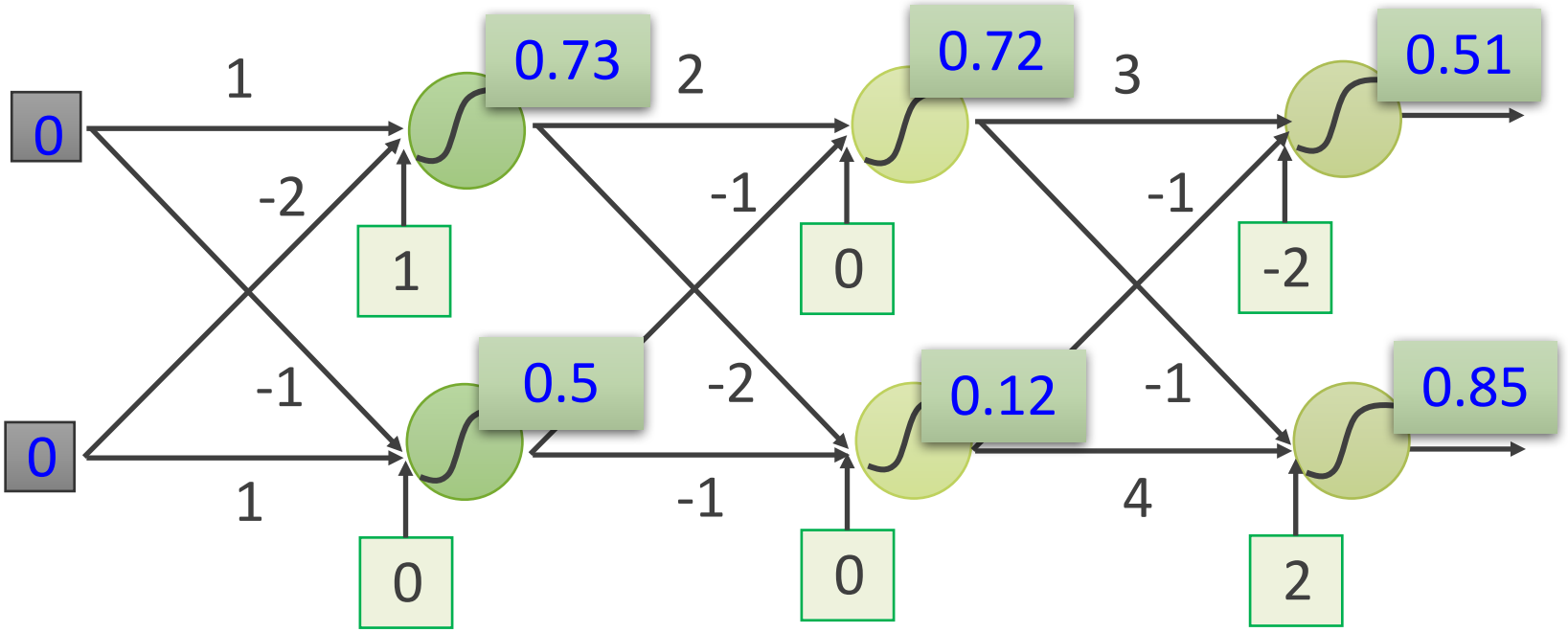
$$\sigma(z) = \frac{1}{1+e^{-z}}$$



Example of Neural Network



Example of Neural Network

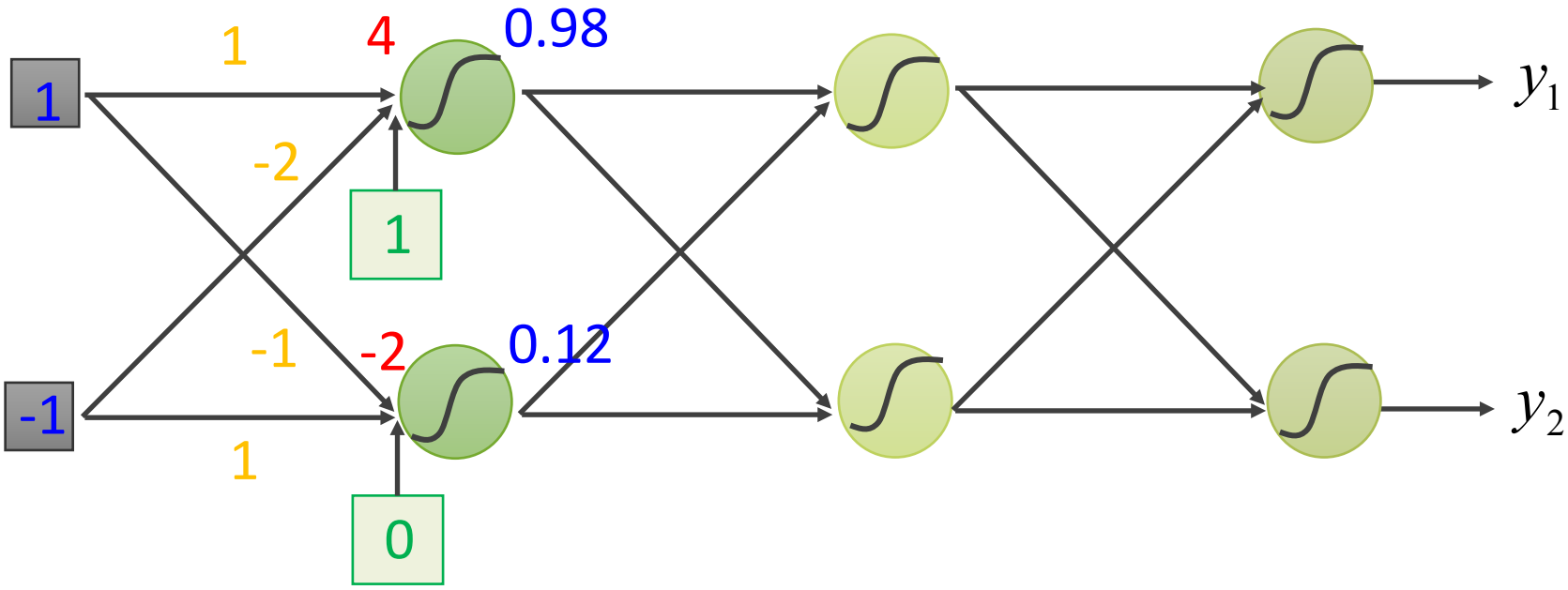


$f: R^2 \rightarrow R^2$

$$f\left(\begin{bmatrix}1\\-1\end{bmatrix}\right) = \begin{bmatrix}0.62\\0.83\end{bmatrix} \quad f\left(\begin{bmatrix}0\\0\end{bmatrix}\right) = \begin{bmatrix}0.51\\0.85\end{bmatrix}$$

Different parameters define different function

Matrix Operation



$$\sigma\left(\begin{bmatrix}1 & -2\\ -1 & 1\end{bmatrix}\begin{bmatrix}1\\ -1\end{bmatrix}+\begin{bmatrix}1\\ 0\end{bmatrix}\right)=\begin{bmatrix}0.98\\ 0.12\end{bmatrix}$$

$$\begin{bmatrix}4\\ -2\end{bmatrix}$$