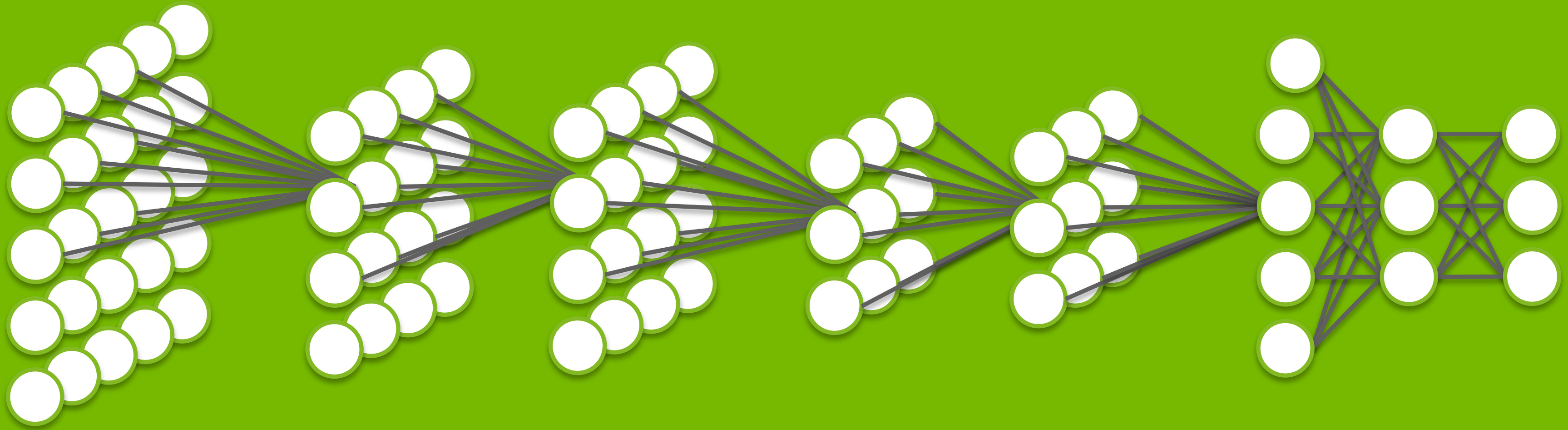
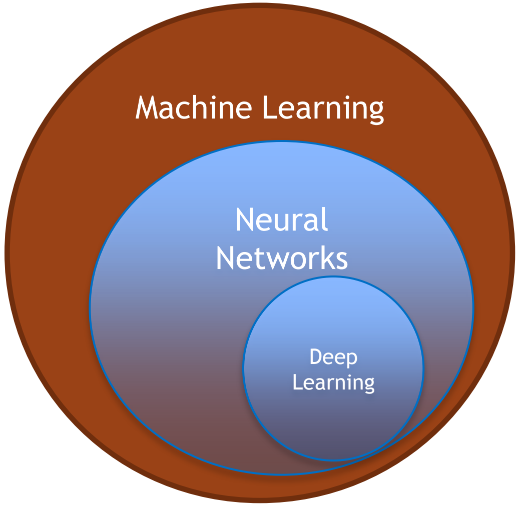
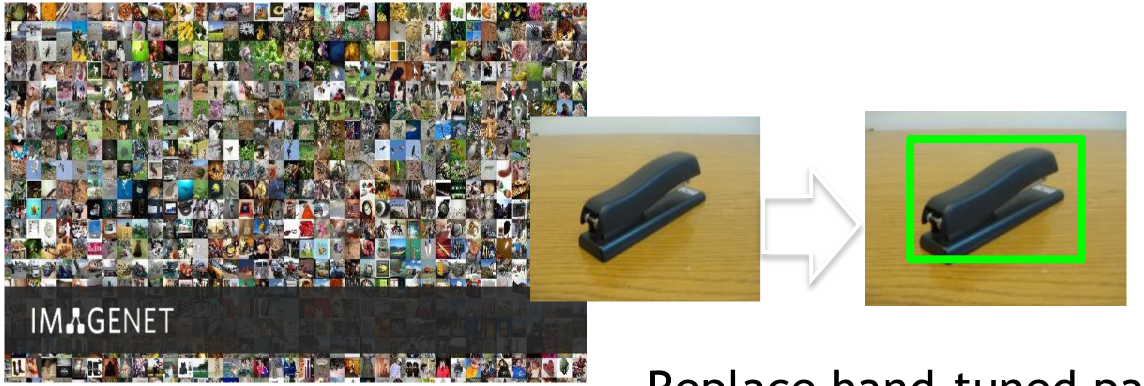
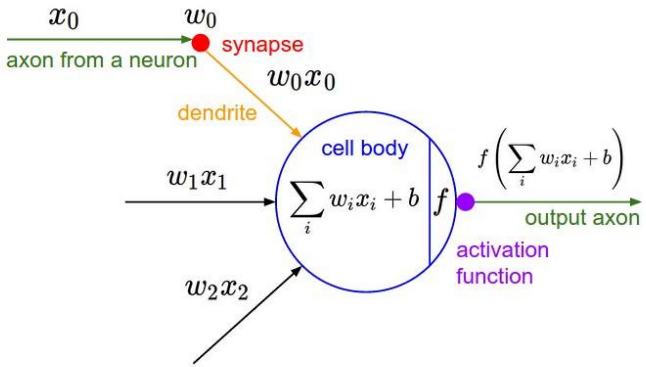
Deep Learning

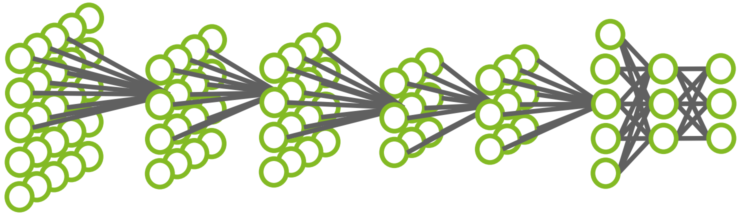


GPUS IN ARTIFICIAL INTELLIGENCE









Replace hand-tuned parameters of the feature extraction steps (e.g. in voice and image recognition)

Artificial Neural Networks inspired by human brain and need lots of training data (ideal for BigData).

Deep learning is a subset of machine learning that refers to artificial neural networks that are composed of many layers.

NVIDIA GPUs and cuDNN software broadly adopted for machine learning.

Neural Networks

Modeling Neurons

Activation function-

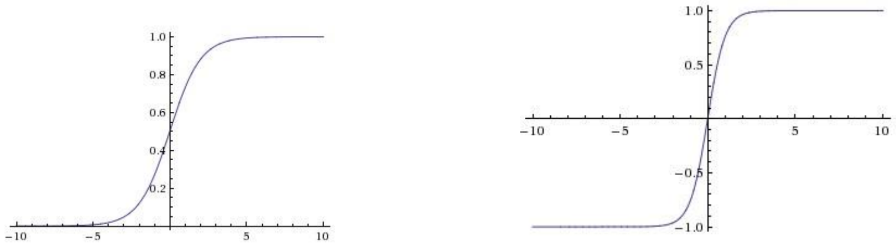
Sigmoid $\sigma(x) = \frac{1}{(1+e^{-x})}$

Tanh - squashes a real-valued number to the range [-1, 1].

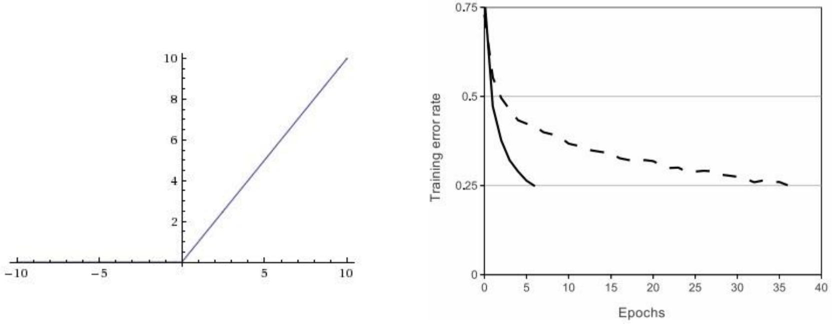
ReLU - Computes the function $f(x) = max(0, x)$

Many more in literature & more details at

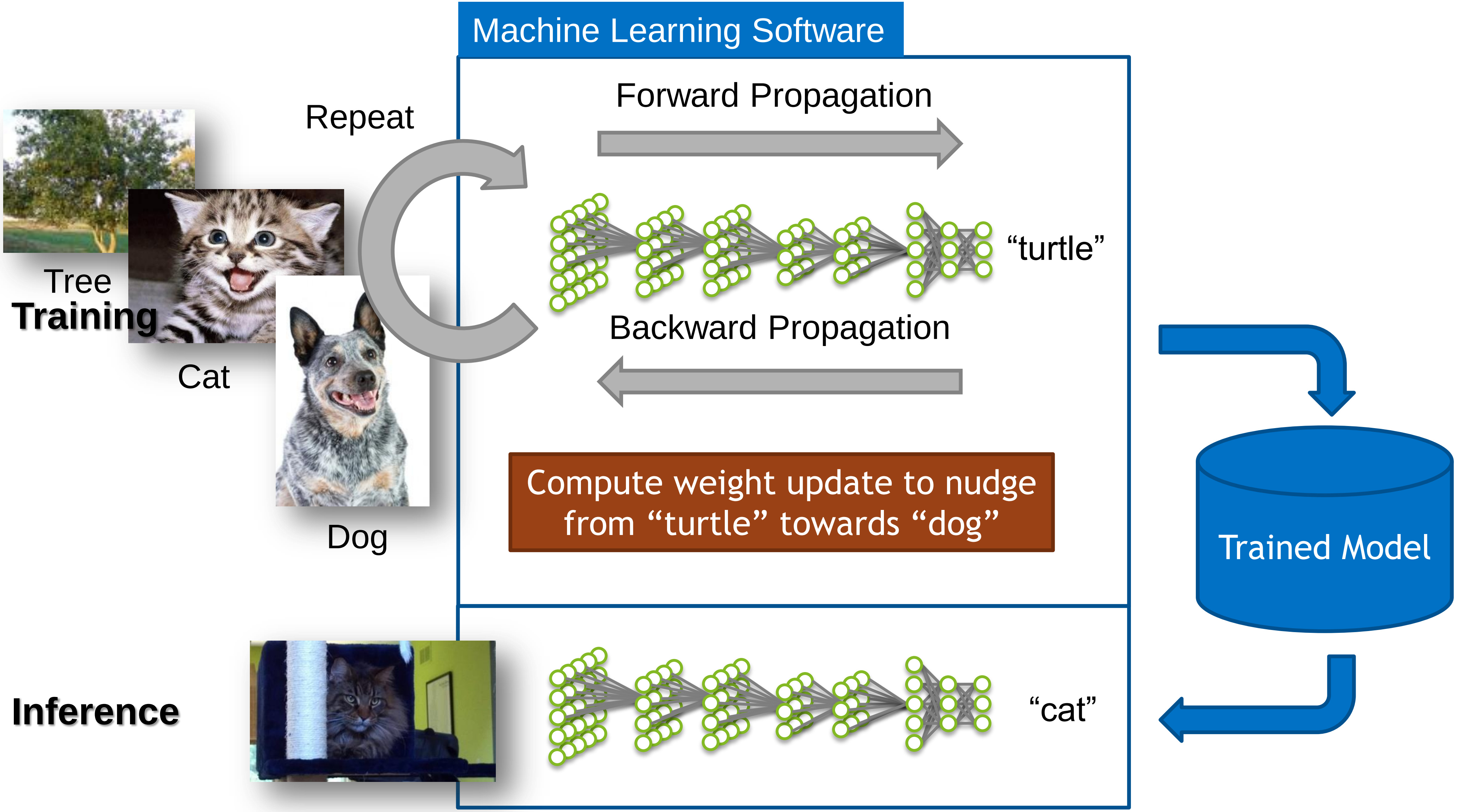
References in beginning.



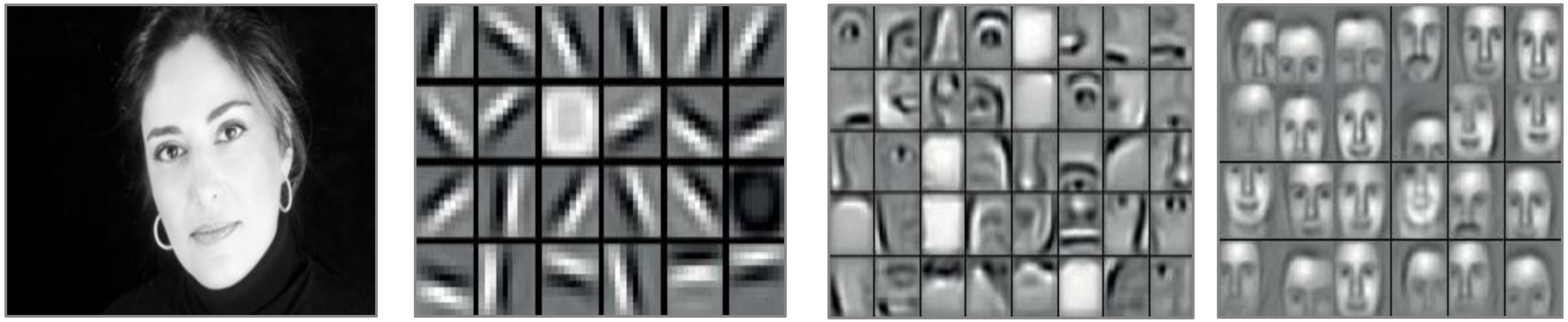
Left: *Sigmoid non-linearity squashes real numbers to range between [0,1]* ***Right:*** *The tanh non-linearity squashes real numbers to range between [-1,1].*

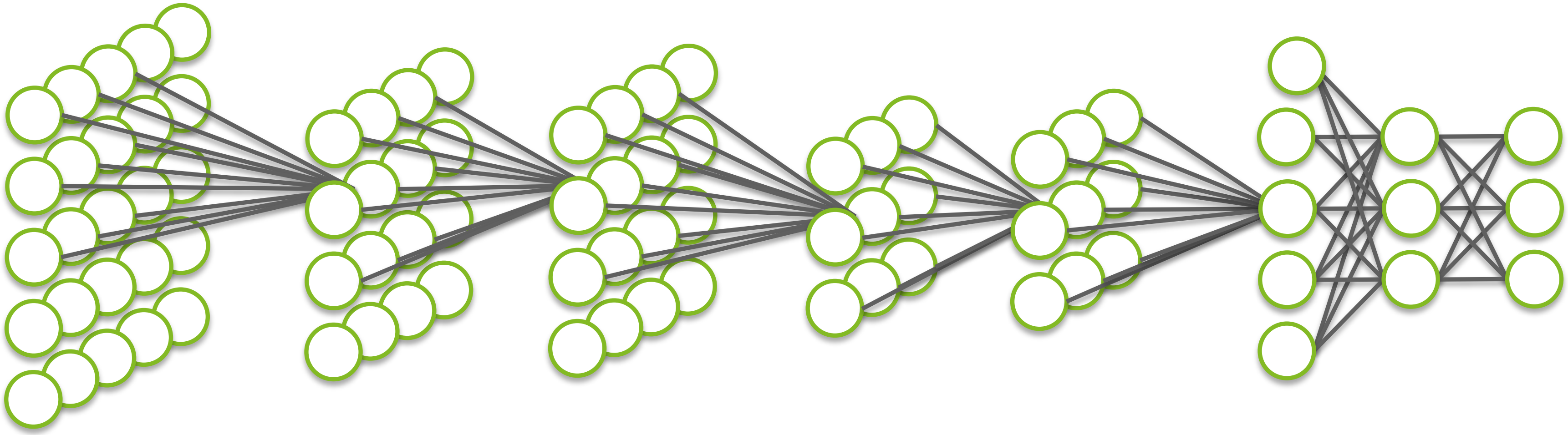


Left: *Rectified Linear Unit (ReLU) activation function, which is zero when x < 0 and then linear with slope 1 when x > 0.* ***Right:*** *A plot from [Krizhevsky et al.](pdf) paper indicating the 6x improvement in convergence with the ReLU unit compared to the tanh unit.*


Machine Learning Software
Repeat
Forward Propagation
Tree
Training
Cat
Dog
“turtle”
Backward Propagation
Compute weight update to nudge from “turtle” towards “dog”
Trained Model
Inference
“cat”

What Is Deep Learning?





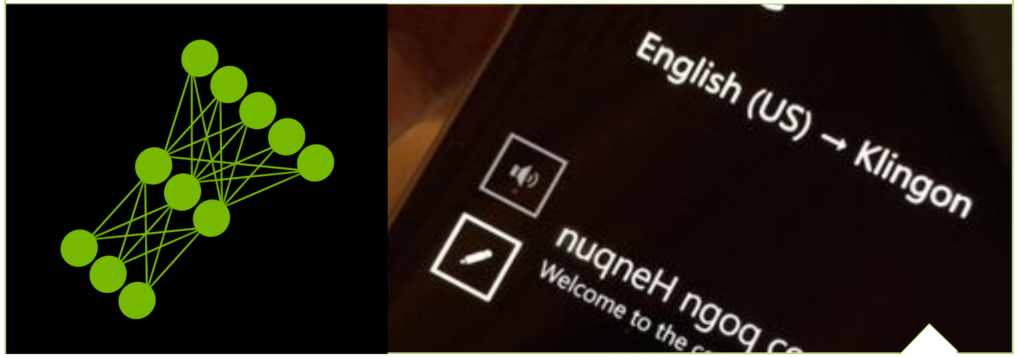
Typical Network

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

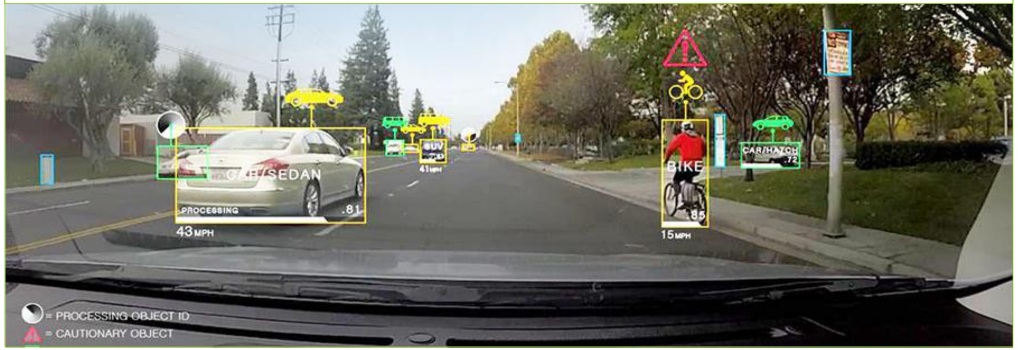
Practical Examples of Deep Learning

Image Classification, Object Detection, Localization, Action Recognition

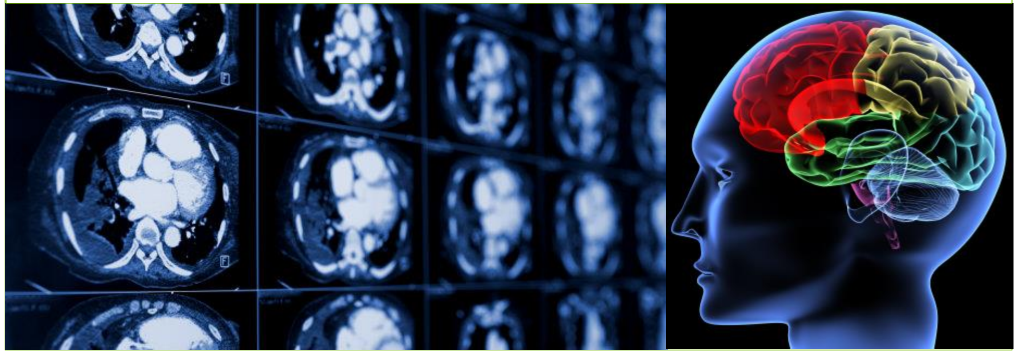
Speech Recognition, Speech Translation, Natural Language Processing



Pedestrian Detection, Lane Detection, Traffic Sign Recognition



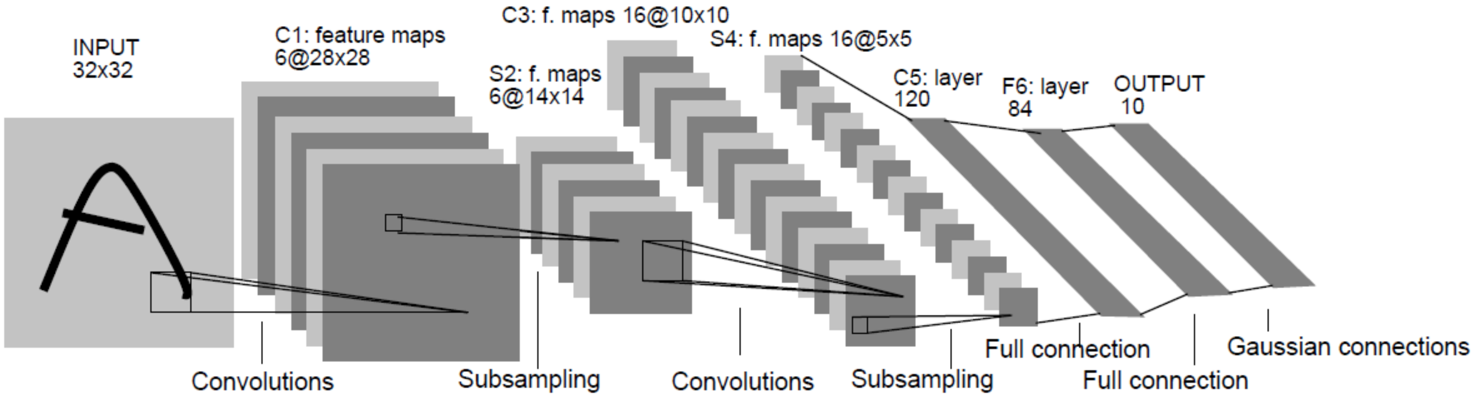
Breast Cancer Cell Mitosis Detection, Volumetric Brain Image Segmentation

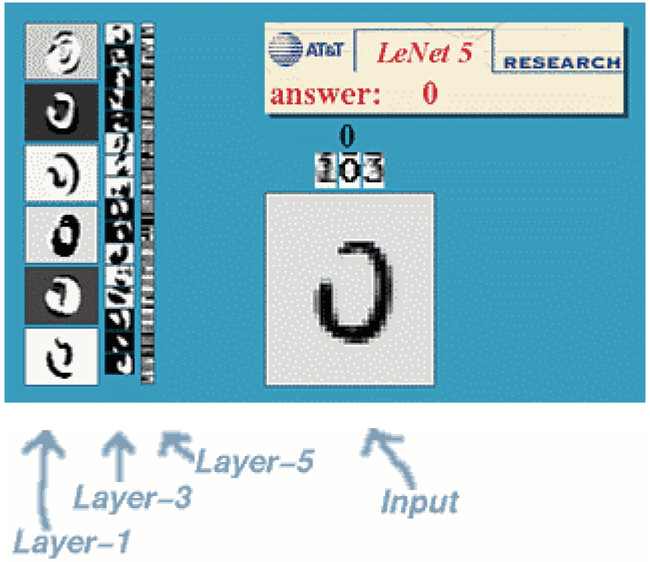


Convolutional Networks Used Case

Local receptive field + weight sharing

Yann LeCun et al, 1998





- MNIST: 0.7% error rate

"Gradient-Based Learning Applied to Document Recognition", Proceedings of the IEEE 1998, http://yann.lecun.com/exdb/lenet/index.html

CNNs- Stacked Repeating Triplets

OverFeat Network 2014

Layer	1	2	3	4	5	6	7	8	Output 9
Stage	conv + max	conv + max	conv	conv	conv	conv + max	full	full	full
# channels	96	256	512	512	1024	1024	4096	4096	1000
Filter size	7x7	7x7	3x3	3x3	3x3	3x3	-	-	-
Conv. stride	2x2	1x1	1x1	1x1	1x1	1x1	-	-	-
Pooling size	3x3	2x2	-	-	-	3x3	-	-	-
Pooling stride	3x3	2x2	-	-	-	3x3	-	-	-
Zero-Padding size	-	-	1x1x1x1	1x1x1x1	1x1x1x1	1x1x1x1	-	-	-
Spatial input size	221x221	36x36	15x15	15x15	15x15	15x15	5x5	1x1	1x1

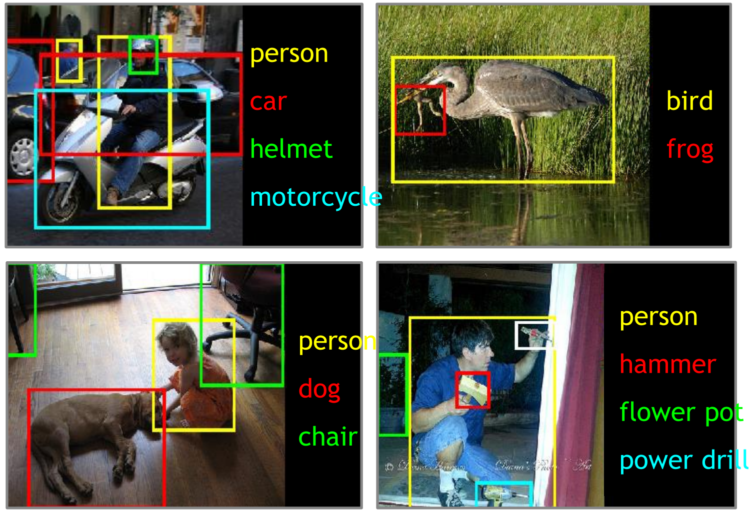
GPUs – the hot platform for Machine Learning

Image Recognition CHALLENGE

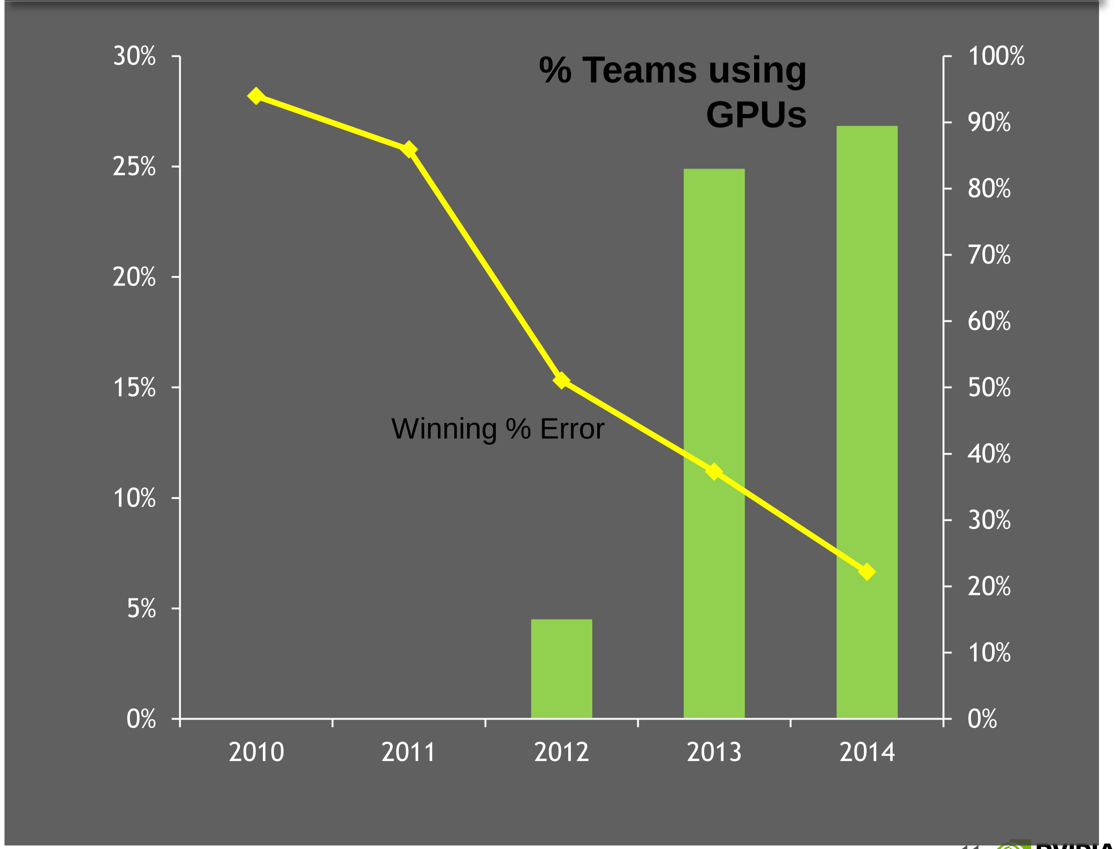
1.2M training images • 1000 object categories

Hosted by

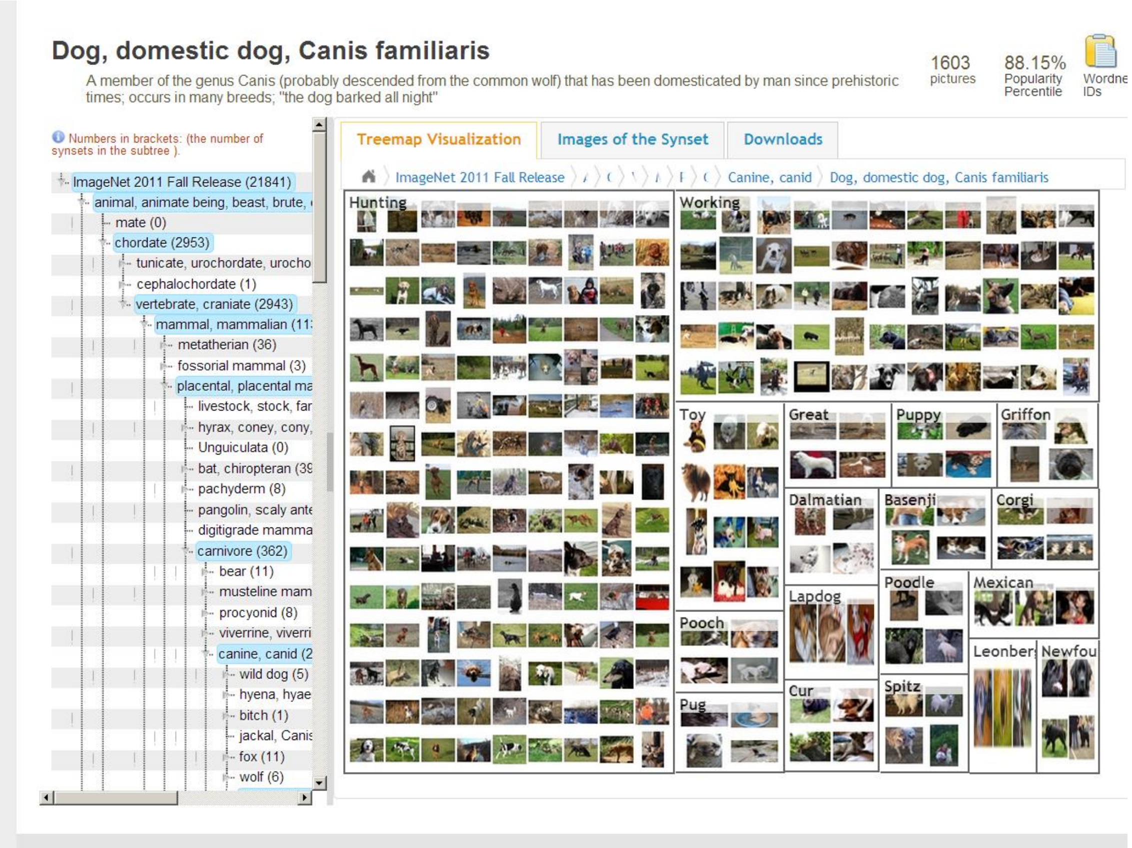
IMAGENET



GPU usage for ILSVRC



Imagenet Database



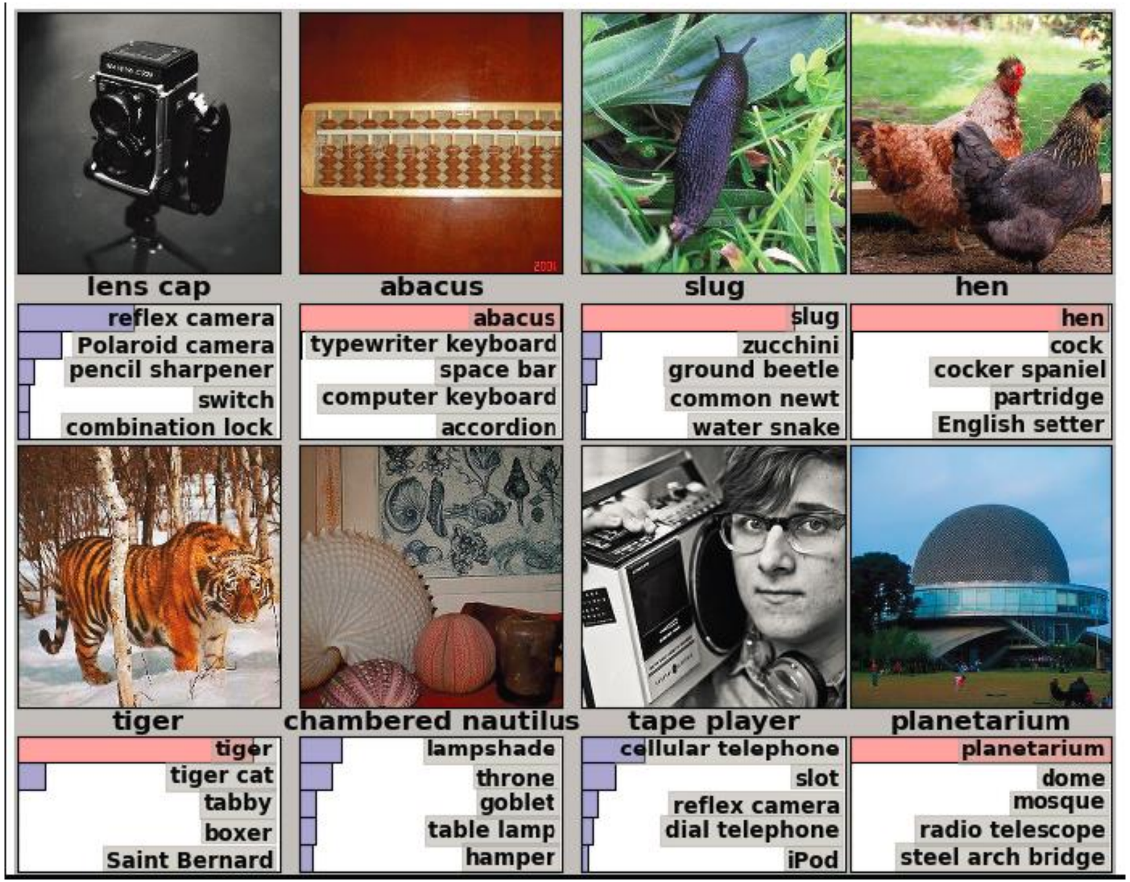
ILSVRC

- **Classification over 1000 categories:**
  - 1.2 million training images, 50,000 validation images, 150,000 testing images
  - Assign to each image label 5 guesses
- **Classification and Localization**
  - 200 classes.
  - Top-5 guesses: label + bounding
- **Detection:**
  - any number of objects in image (including zero). False positives are penalized

http://www.image-net.org/challenges/LSVRC/2015/

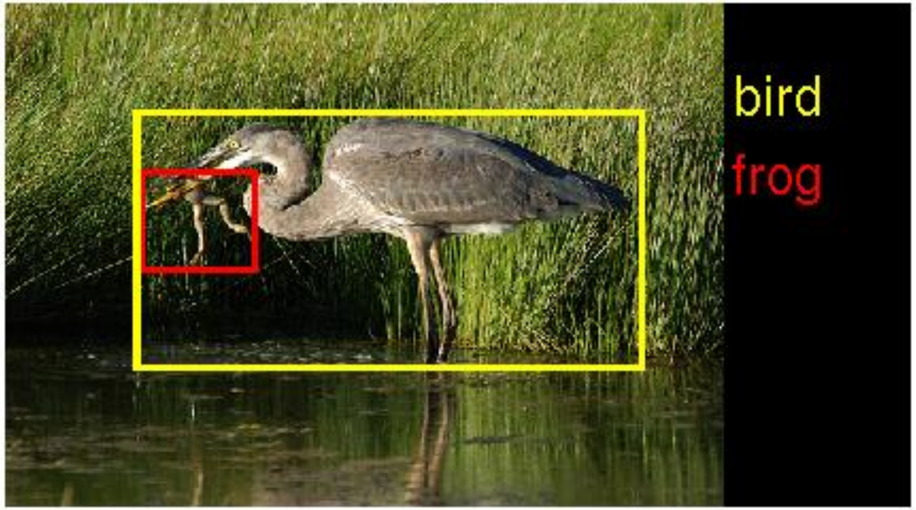
ILSVRC: Classification

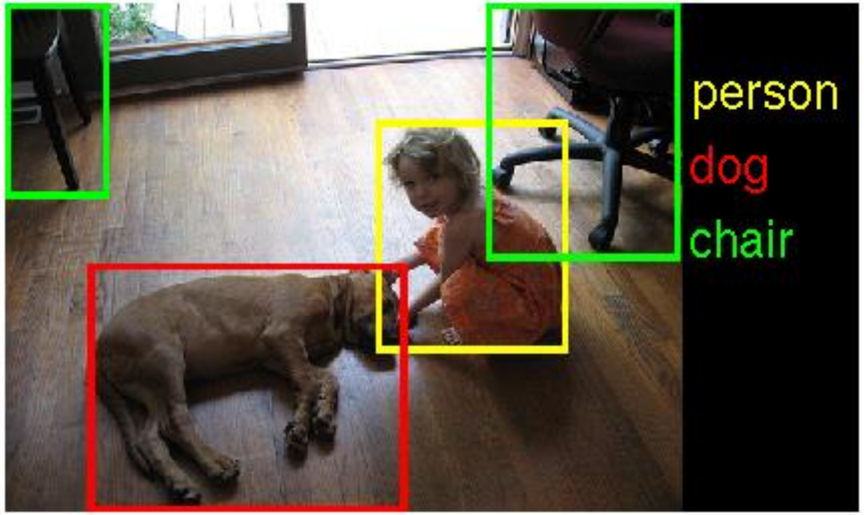
- top-5 labels



http://www.image-net.org/challenges/LSVRC/2015/

ILSVRC: Classification & Localization





top-5 labels + bounding box

ILSVRC 2014 vs PASCAL 2012

		PASCAL 2012	ILSVRC 2013	ILSVRC 2014
	# classes	20	200	200
Training	# images	5,717	395,909	**456,567**
	# objects	13,609	345,854	**478,807**
Validation	# images	5,823	20,121	20,121
	# objects	13,841	55,502	55,502
testing	# images	10,991	40,152	40,152
	# objects			

- http://image-net.org/challenges/LSVRC/2015/ http://pascallin.ecs.soton.ac.uk/challenges/VOC/

ILSVRC: Classification

(a) Siberian husky

(b) Eskimo dog

ILSVRC: Classification

Groundtruth: **????**

ILSVRC: Classification

Groundtruth: **coffee mug**

ILSVRC: Classification

<u>Groundtruth</u>: **coffee mug**

<u>Top-5</u>:
- **table lamp**
- **lamp shade**
- **printer**
- **projector**
- **desktop computer**

Example applications

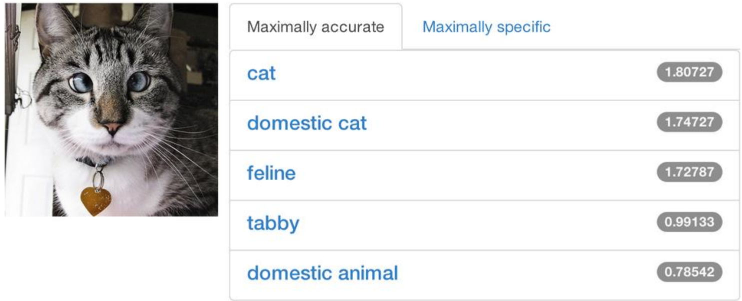
Use case 1: classification of images

Object

http://demo.caffe.berkeleyvision.org/

Open source demo code:

$CAFFE_ROOT/examples/web_demo



Scene

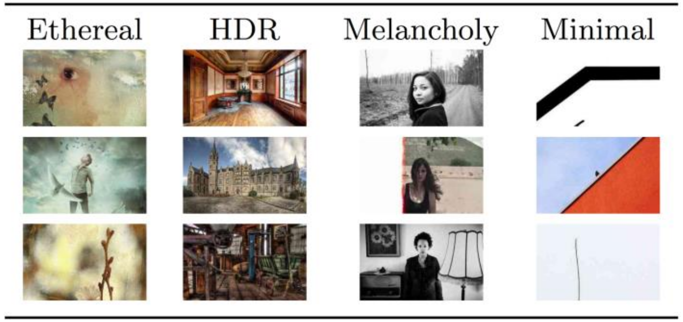
http://places.csail.mit.edu/

B. Zhou et al. NIPS 14

Style

http://demo.vislab.berkeleyvision.org/

Karayev et al. *Recognizing Image Style*. BMVC14

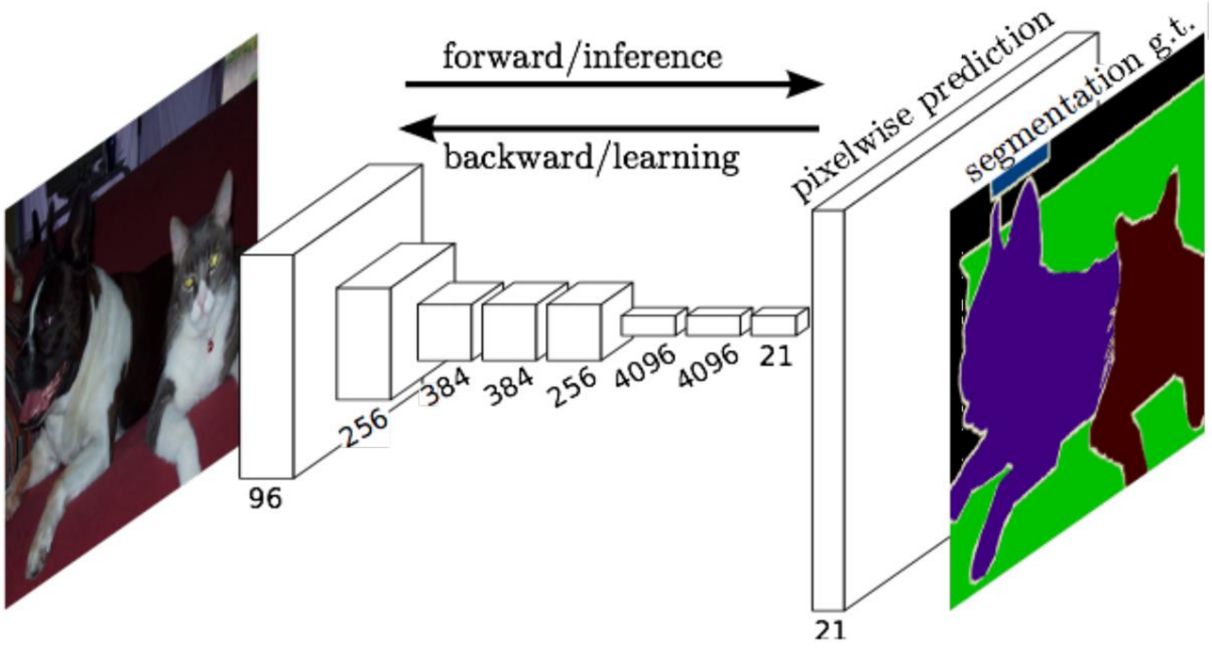


CNN used case: Semantic Segmentation

[Farabet et al. 2013]

CNN used case

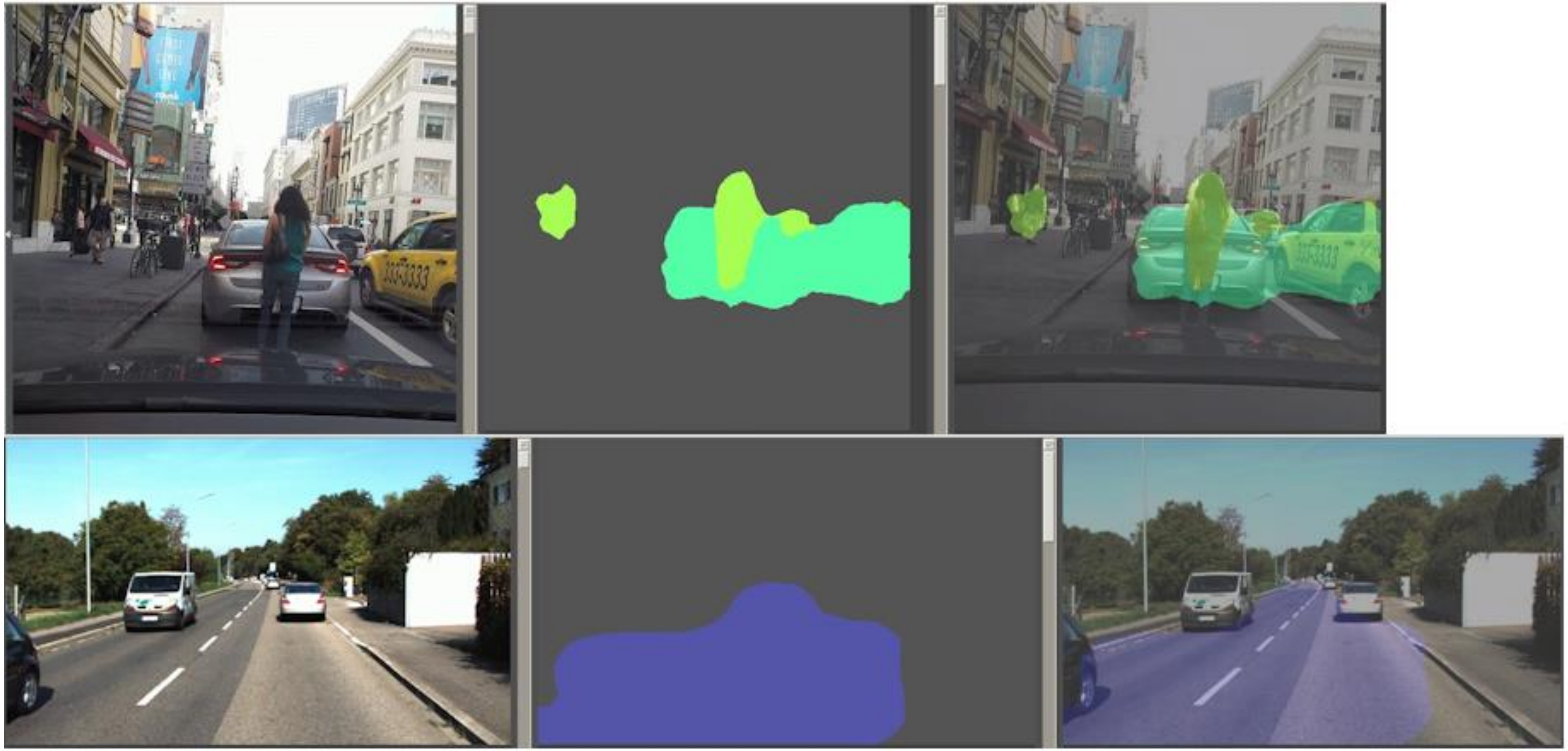
Fully convolutional pixel segmentation : level classification and segmentation



http://fcn.berkeleyvision.org

Long, Shelhamer, Darrell, *Fully convolutional networks for semantic segmentation*, CVPR 2015

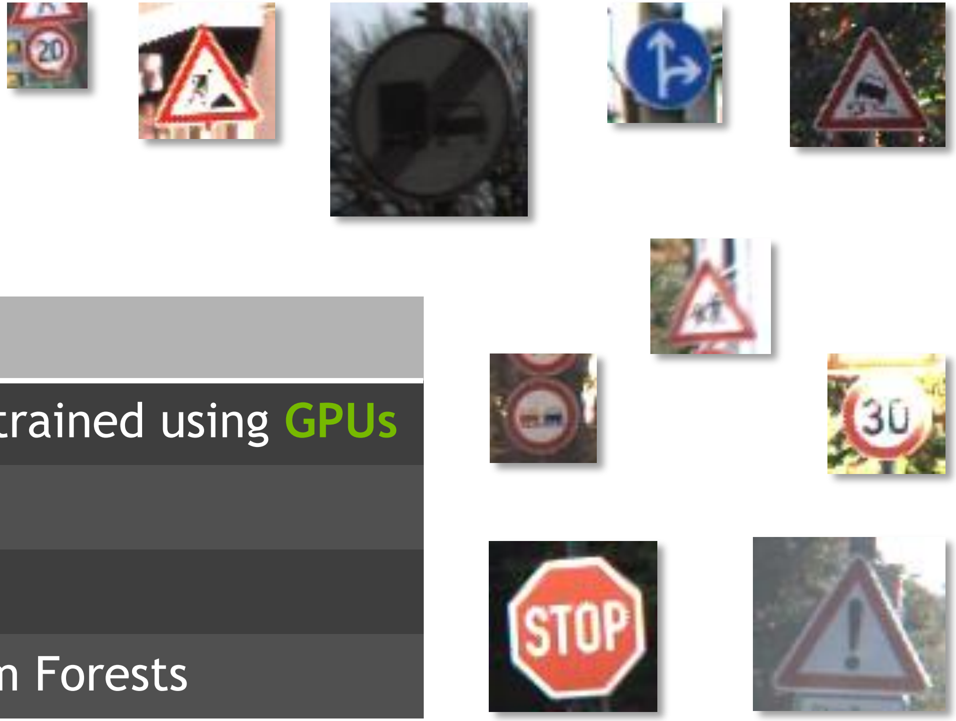
CNN used case : Fully Convolutional Segmentation



CNN used case: Traffic sign recognition

GTSRB

The German Traffic Sign Recognition Benchmark, 2011



Rank	Team	Error rate	Model
1	IDSIA, Dan Ciresan	0.56%	CNNs, trained using GPUs
2	Human	1.16%	
3	NYU, Pierre Sermanet	1.69%	CNNs
4	CAOR, Fatin Zaklouta	3.86%	Random Forests

http://benchmark.ini.rub.de/?section=gtsrb

Transfer Learning

Dogs vs. Cats

Dogs vs. Cats, 2014

Train model on one dataset – ImageNet

Re-train the last layer only on a new dataset – Dogs and Cats

Rank	Team	Error rate	Model
1	Pierre Sermanet	1.1%	CNNs, model transferred from ImageNet, GPUs
2-4	…		

https://www.kaggle.com/c/dogs-vs-cats

Caffee Example applications

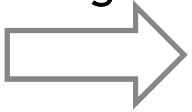
Use case 5: Transfer learning

Lots of data

CNN

Object Classifier

Transfer weights



New data

© kaggle.com

Dog vs. Cat

Top 10 in 10 mins after finetuning

- Just change a few lines in the model prototxt file

```
layer {
  name: "data"
  type: "Data"
  data_param {
    source: "ilsvrc12_train"
    …
  }
  …
}
…
layer {
  name: "fc8"
  type: "InnerProduct"
  inner_product_param {
    num_output: 1000
    …
  }
}
```

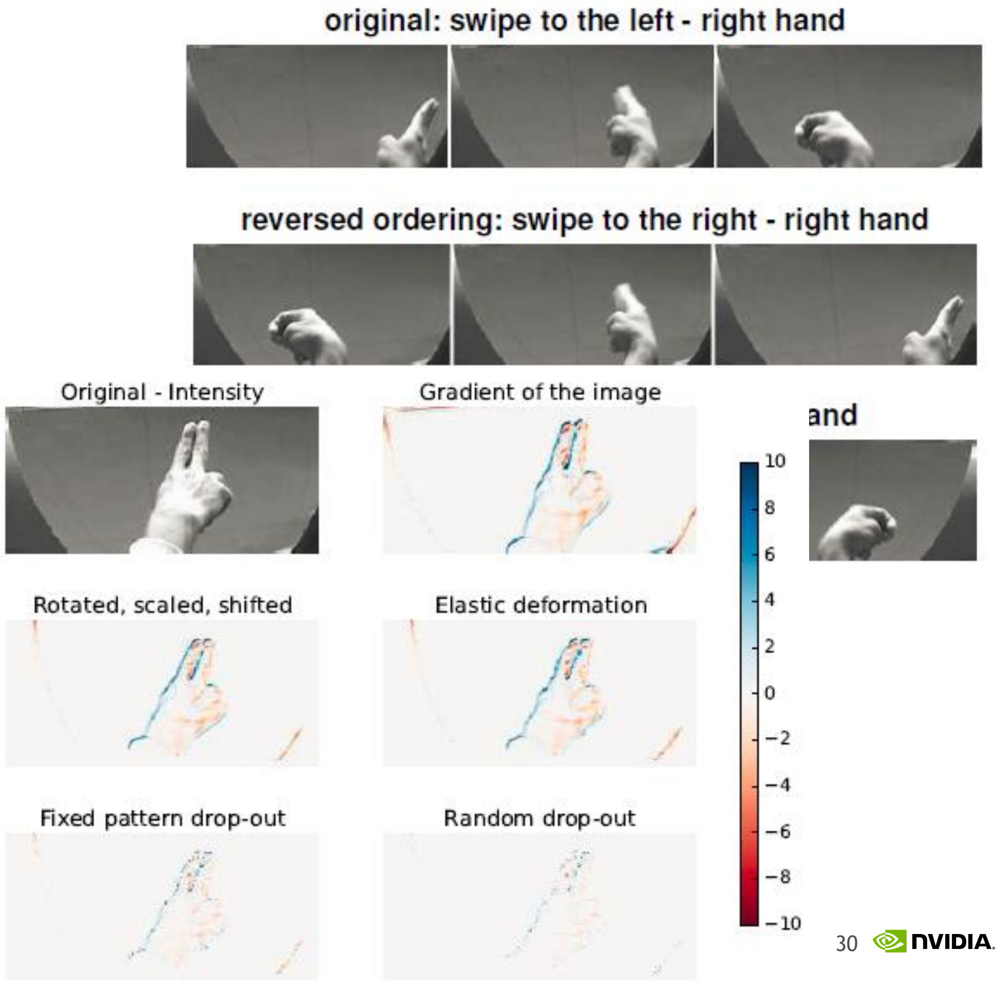
```
layer {
  name: "data"
  type: "Data"
  data_param: {
    source: "dogcat_train"
    …
  }
  …
}
…
layer {
  name: "fc8-dogcat"
  type: "InnerProduct"
  inner_product_param {
    num_output: 2
    …
  }
}
```

Hand Gesture Recognition with 3D CNN

NVIDIA research paper presented at CVPR 2015

Abstract

Touchless hand gesture recognition systems are becoming important in automotive user interfaces as they improve safety and comfort. Various computer vision algorithms have employed color and depth cameras for hand gesture recognition, but robust classification of gestures from different subjects performed under widely varying lighting conditions is still challenging. We propose an algorithm for drivers' hand gesture recognition from challenging depth and intensity data using 3D convolutional neural networks. Our solution combines information from multiple spatial scales for the final prediction. It also employs spatio-temporal data augmentation for more effective training and to reduce potential overfitting. Our method achieves a correct classification rate of 77.5% on the VIVA challenge dataset.



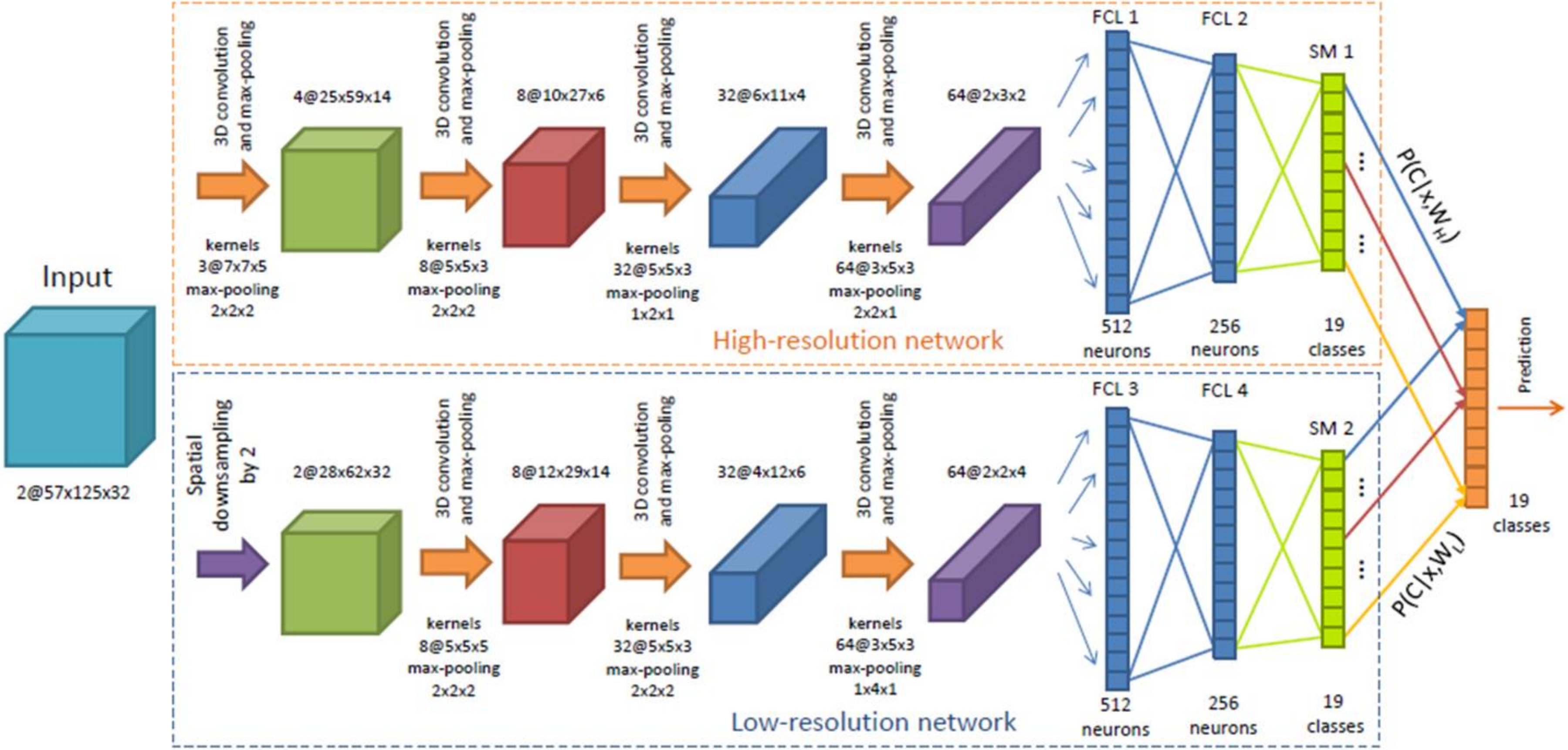


Figure 1: **Overview of our CNN classifier** We used a CNN-based classifier for hand gesture recognition. The inputs to the classifier were $57 \times 125 \times 32$ sized volumes of image gradient and depth values. The classifier consisted of two sub-networks: a high-resolution network (HRN) and a low-resolution network (LRN). The outputs of the sub-networks were class-membership probabilities $P(C|x, \mathcal{W}_H)$ and $P(C|x, \mathcal{W}_L)$, respectively. The two networks were fused by multiplying their respective class-membership probabilities element-wise.

Speech recognition

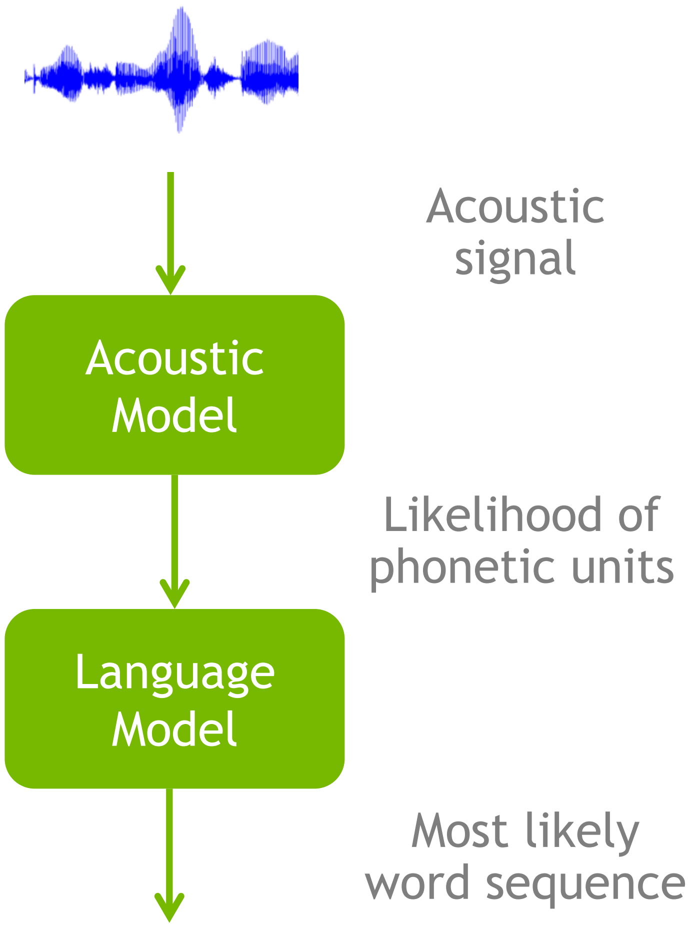
Acoustic model

Acoustic model is DNN

- Usually fully-connected layers
- Some try using convolutional layers with spectrogram used as input
- Both fit GPU perfectly

Language model is weighted Finite State Transducer (wFST)

- Beam search runs fast on GPU



http://devblogs.nvidia.com/parallelforall/cuda-spotlight-gpu-accelerated-speech-recognition/
http://devblogs.nvidia.com/parallelforall/understanding-natural-language-deep-neural-networks-using-torch/
http://devblogs.nvidia.com/parallelforall/deep-speech-accurate-speech-recognition-gpu-accelerated-deep-learning

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

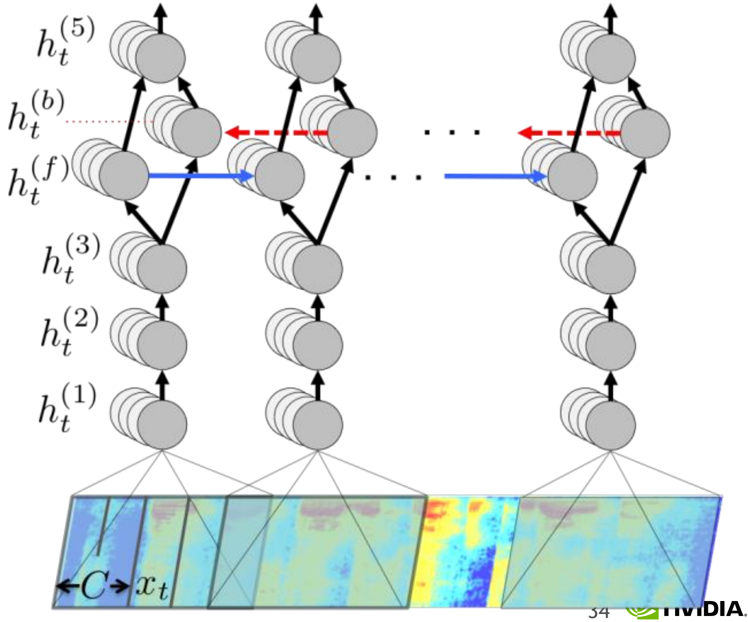
- No feature engineering or Mandarin-specifics required

More accurate than humans

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

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

http://arxiv.org/abs/1512.02595

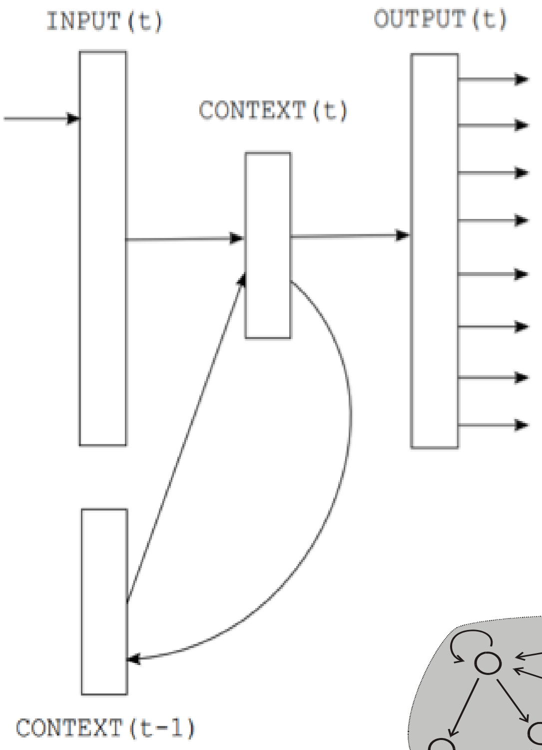


Simple Recurrent (Elman) NN

Recurrent: Activation of a neuron depends on its previous output.

Cyclic path of synaptic connections

- all biological neural networks are recurrent
- RNNs implement dynamical systems



Forward pass (identical to Feedforward Propagation)

Load input to the input neurons x(t)

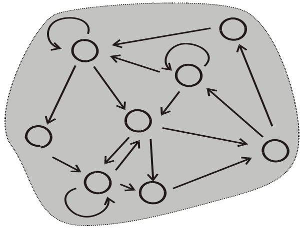
Hidden layer activations :

applying a sigmoidal to the sum of the weighted inputs :

from input units and from copy layer z(t-1)

Output layer activations as usual y(t)

Copy new hidden layer to copy (context) layer



History (Memory!) increases for bigger time intervals.

LSTM long-short term memory

Simple Recurrent (Elman) NN

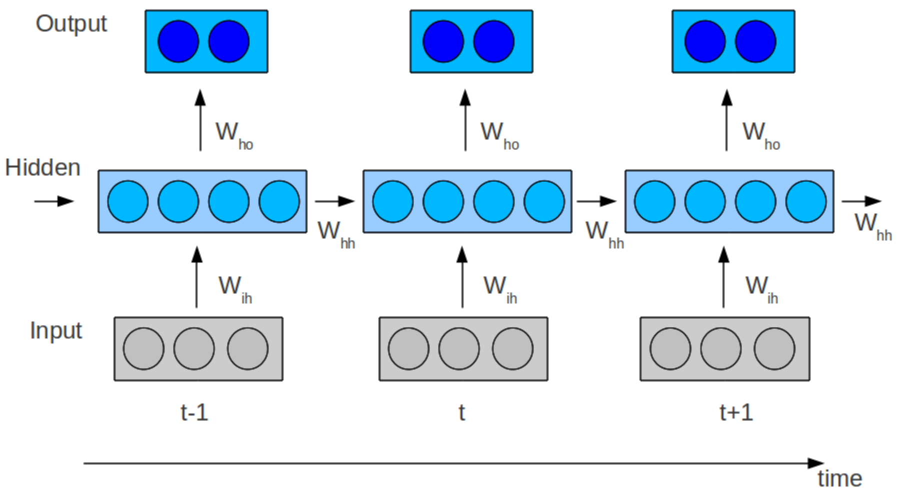
Back Propagation Through Time by Network unfolding

Training approximately with straight Backpropagation.

Each network copy corresponds to a time step.

Backpropagation with weight constraints :

- The set of weights for each copy (time step) always remain the same.
- The weight changes calculated for each network copy are summed before individual weights are adapted.



The problem with RNN

- Exploding gradient – large increase of the gradient's norm due to long term dependencies. Results in large increase of the cost function during training.
- Address with Rectified-linear (ReLu) activation function
- Vanishing gradient – opposite behaviour, long term components go exponentially fast to 0. Results in bad prediction of long term dependencies.

Long short-term memory (LSTM)

Hochreiter (1991) analysed vanishing gradient “*LSTM falls out of this almost naturally*”

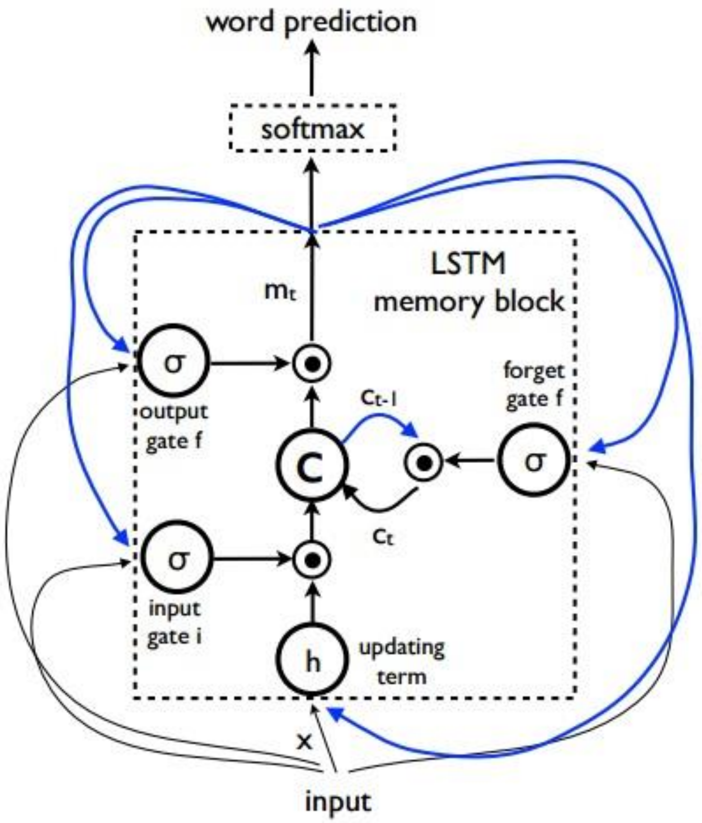
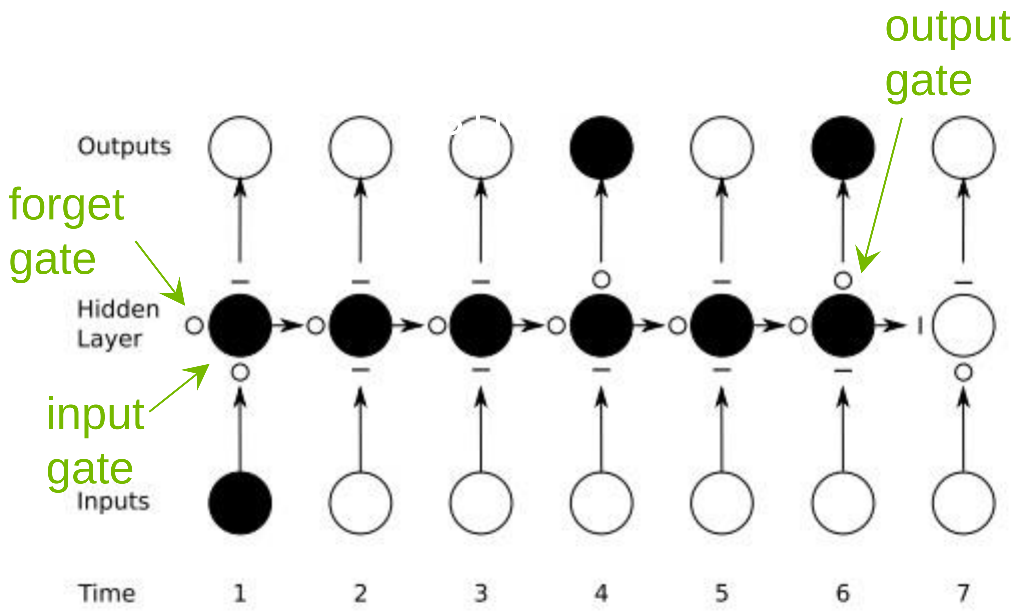


Fig from Vinyals et al, Google April 2015 NIC Generator

Training via backprop unfolded in time



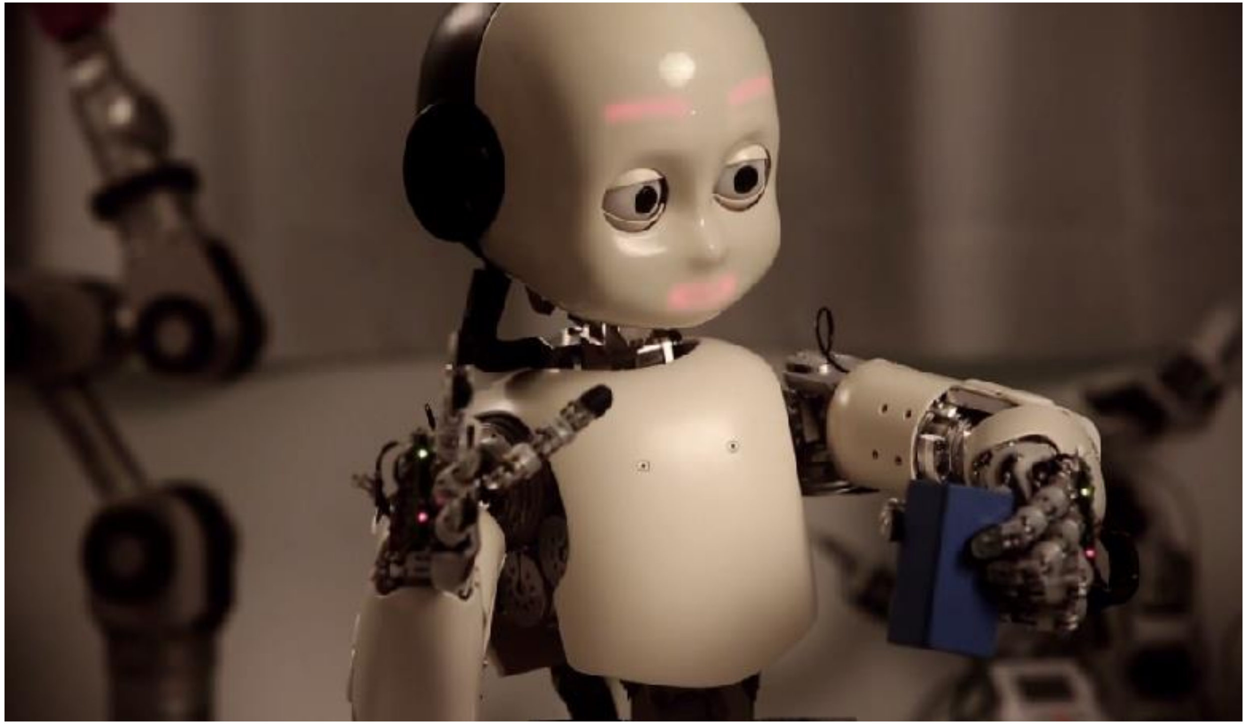
Long time dependencies are preserved until
input gate is closed (-) and forget gate is open (O)

Gates control importance of the corresponding activations

Fig from Graves, Schmidhuber et al, Supervised Sequence Labelling with RNNs

LSTM Use Cases

“The RNN's memory is necessary to deal with ambiguous sensory inputs from repetitively visited states”



iCub (Intelligent Cub) http://robotics.idsia.ch/robots/

Metalearning - general purpose learning algorithms that can learn better learning algorithms! Applying Genetic Programming to itself.

Courtesy of J Schmidhuber, IDSIA

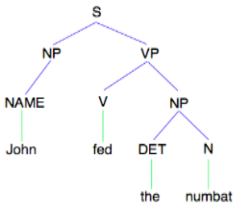
LSTM used case

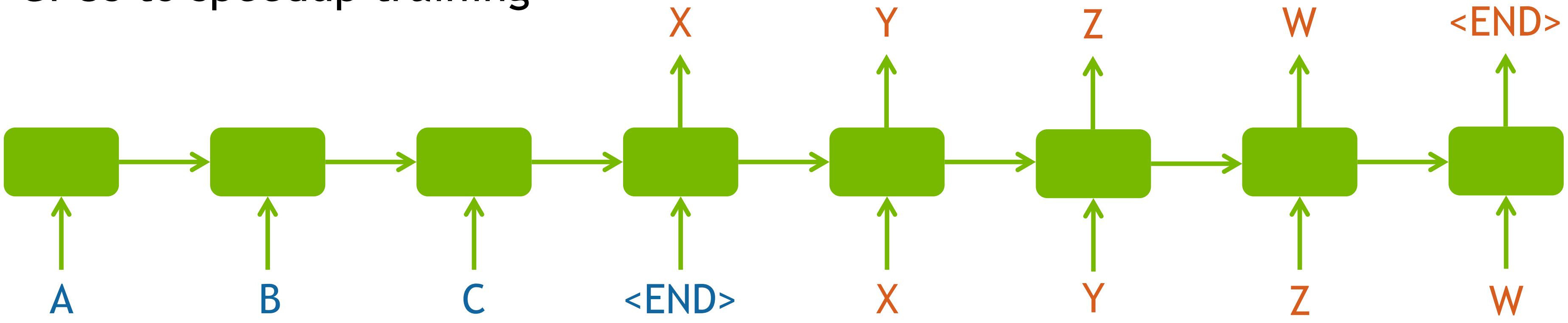
Text translation

Ilya Sutskever et al (Google), 2014 English to French translation

LSTM, variation of RNN, almost achieved state of the art (36.5 and 37.0)

Used 8 GPUs to speedup training



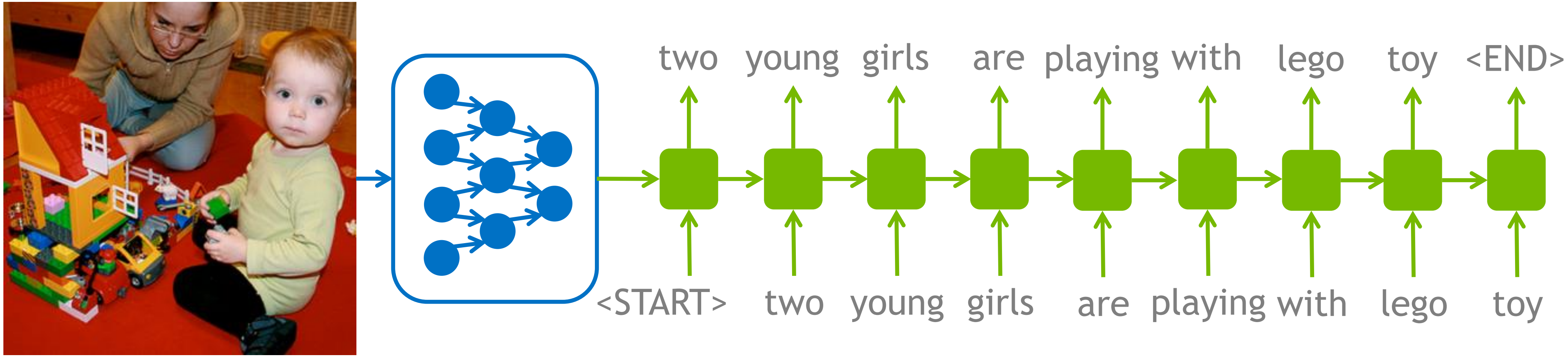


Fan et al., (Microsoft), Text-to-speech synthesis
Bluche et al., (DAS) Arabic handwriting recognition

Image to Text translation

ConvNets + RNN/LSTM

Multiple papers (Stanford, Google, UC Berkley, and others) in Nov 14 and at ICML 15

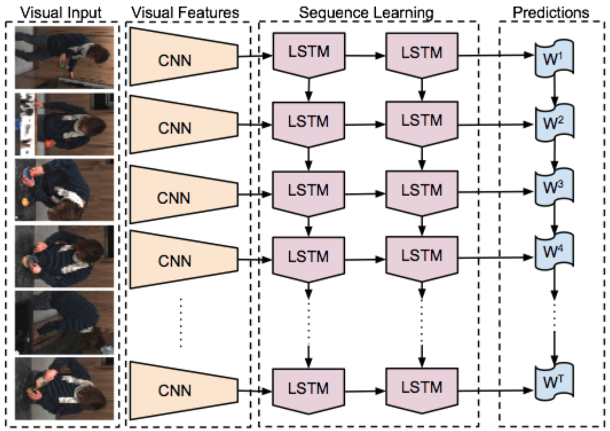


http://cs.stanford.edu/people/karpathy/deepimagesent/
http://arxiv-web3.library.cornell.edu/pdf/1411.4389v1.pdf

Example applications

Use case 4: Sequence learning

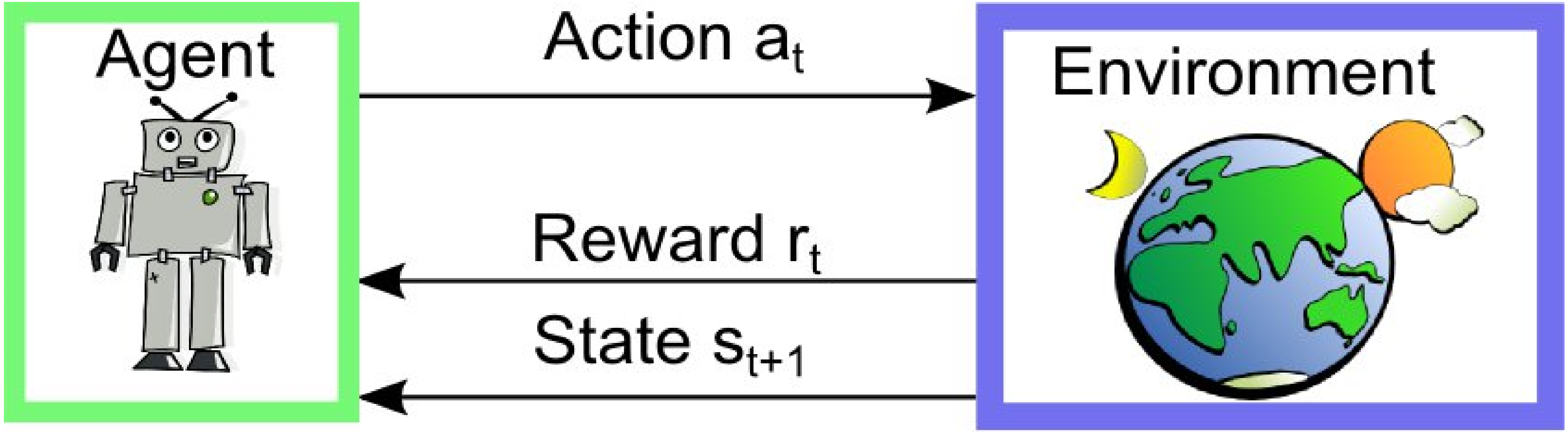
- Recurrent Neural Networks (RNNs) and Long Short Term Memory (LSTM)
  - Video
  - Language
  - Dynamic data
- Current Caffe pull request to add support
  - https://github.com/BVLC/caffe/pull/1873
  - http://arxiv.org/abs/1411.4389



A group of young men playing a game of soccer.

Jeff Donahue et al.

Reinforcement Learning



Reinforcement Learning Setup

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

Google DeepMind

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

THANK YOU!

- Developer Zone: https://developer.nvidia.com/deeplearning
- GPU Technology Conference: http://www.gputechconf.com/
- cuDNN Download: https://developer.nvidia.com/cuDNN
- DIGITS Download: https://developer.nvidia.com/digits
- DIGITS Source: https://www.github.com/nvidia/digits

gunterr@nvidia.com