EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MATHEMATISCH-
NATURWISSENSCHAFTLICHE FAKULTÄT

Deep Neural Networks / Deep Learning

Andreas Zell

0. Organizational Issues

- Textbook:
  - Ian Goodfellow, Yoshua Bengio, Aaron Courville: Deep Learning, MIT Press, 2016
  - Online version of this textbook available at http://www.deeplearningbook.org/
- Scriptum of the course (which heavily draws on the above textbook, but also uses other sources)
- I will put the scriptum online before classes
- My office hours are
  - Wed. 13:30 -15:00 in A310, Sand 1
- Email: andreas.zell@uni-tuebingen.de

0. Organizational Issues

- Research assistants:
  - Kevin Laube (kevin.laube@uni-tuebingen.de, A311)
  - Maximus Mutschler (maximus.mutschler@uni-tuebingen.de, A318)
  - Faranak Shamsafar, (faranak.shamsafar@uni-tuebingen.de, A304)
  - Benjamin Kiefer (benjamin.kiefer@uni-tuebingen.de, A306)
- Lecture and Tutoring
  - Wednesdays, 10:15 – 12:00, F119 Lecture
  - Tuesdays, 12:15 – 14:00, F119 Tutoring session

Overview

1. Introduction and Motivation
2. Linear Algebra for Deep Learning
3. Probability and Information Theory
4. Machine Learning Basics
5. Backpropagation
6. Deep Feedforward Networks
7. Regularization for Deep Learning
8. Optimization for Training Deep Models
9. Convolutional Networks

Overview (continued)

10. Sequence Modeling: Recurrent Networks
11. Practical Methodology
12. Popular Network Architectures and Applications

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MATHEMATISCH-
NATURWISSENSCHAFTLICHE FAKULTÄT

Deep Neural Networks

Chapter 1: Introduction and Motivation

Brains of different animals

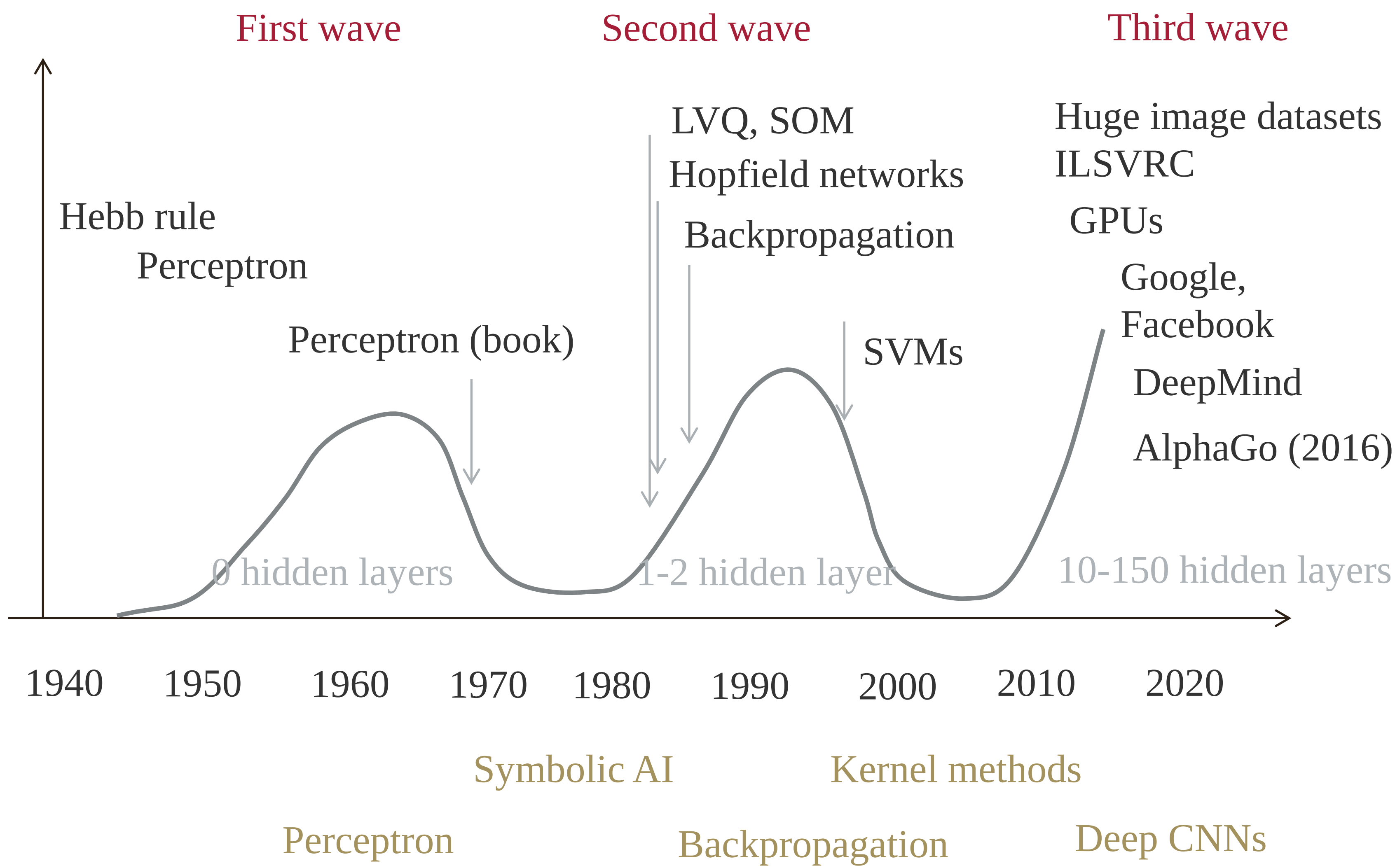
Name	Neurons in the brain	Synapses
Caenorhabditis elegans	302	~ 7,500
Fruit fly	250,000	
Ant	~ 250,000	
Honey bee	960,000	~ 10^9
Zebrafish	~ 10,000,000	
House mouse	71,000,000	~ 10^{11}
Brown rat	200,000,000	~ 5×10^{11}
Cat	760,000,000	~ 10^{13}
Rhesus macaque	6,376,000,000	
Human	86,000,000,000	adult: 10^{14}–10^{15}
African elephant	257,000,000,000	*

Source: https://en.wikipedia.org/wiki/List_of_animals_by_number_of_neurons

Neurons in the Cerebral Cortex

Name	Neurons in Cerebral Cortex	
Mouse	4,000,000	Mus musculus
Rat	~ 20,000,000	Rattus
Dog	160,000,000	Canis lupus familiaris
Cat	300,000,000	Felis catus
Domesticated pig	450,000,000	Sus scrofa
Horse	1,200,000,000	Equus ferus caballus
Gorilla	4,300,000,000	Gorilla
Chimpanzee	6,200,000,000	Pan
African elephant	11,000,000,000	Loxodonta
Fin whale	15,000,000,000	Balaenoptera physalus
Human	16,300,000,000	Homo sapiens
Long-finned pilot whale	37,200,000,000	Globicephala melas

The new hype about neural networks



First wave of neural networks

Year	Researchers	***Paper or Book* (Main topic)**
1943	Warren McCulloch, Walter Pitts	*A logical calculus of the ideas immanent in nervous activity* (neurological networks)
1947	Walter Pitts, Warren McCulloch	*How we know universals* (recognition of spacial patterns)
1949	Donald O. Hebb	*The organization of behaviour* (Hebb's rule, concept of cell groups)
1958	Selfridge	Pandemonium (dynamical, interactive mechanisms)
1959	Frank Rosenblatt	*Principles of Neurodynamics* (Perceptron, Perceptron convergence theorem)
1960	Bernard Widrow, Marcian E. Hoff	*Adaptive switching circuits* (Adaline, Delta-Rule)
1969	Marvin Minsky, Seymour Papert	*Perceptrons* (mathematical analysis of the perceptron)

Between first and second wave

Year	Researchers	***Paper or Book*** **(Main topic)**
1973	J. A. Anderson	(Distributed representation, concept learning)
1974	Paul Werbos, Harvard	(Backpropagation) … but nobody cared
1976 1980	Stephen Grossberg	*Adaptive pattern classif. and univ. recoding,* *How does the brain build a cognitive code*?
1976	David Marr, Tomaso Poggio	*Cooperative computation of stereo disparity* (Stereo Vision and early vision processing)
1981	Geoffrey E. Hinton, J. A. Anderson	Parallel Models of Associative Memory
1982	Jerome Feldman, Steven Ballard	*Connectionist models and their properties*

Second wave of neural networks

Year	Researchers	***Paper or Book*** **(Main topic)**
1982	Teuvo Kohonen	*Self-organized formation of topologically correct feature maps*
1982	John J. Hopfield	*Neural networks and physical systems with emergent collective computational abilities*
1983	K. Fukushima, S. Miyake, T. Ito	*Neocognitron: a neural network model for a mechanism of visual pattern recognition*
1983	A. Barto, R. Sutton, C.W. Anderson	*Neuronlike adaptive elements that can solve difficult learning control problems*
1986	D. Rumelhart, G. Hinton, R. Williams	*Learning internal representations by error propagation*
		Learning representations by back-propagating errors, Nature 323:533-536
1997	J. Hochreiter, J. Schmidhuber	Long short-term Memories (LSTMs)

Third wave of neural networks

Year	Researchers	*Paper or Book* (Main topic)
2006	G. E. Hinton	*Training of deep belief networks*
2007	Y. Bengio et al.	*Greedy layer-wise training of deep networks*
2012	Krizhevski, Sutskever, Hinton	*ImageNet Classification with Deep Convolutional Neural Networks*
2014	Ioffe, Szegedy	*Batch Normalization: Accelerating Deep Netw. Training by Reducing Covariate Shift*
2014	N. Srivastava et al.	*Dropout. A Simple Way to Prevent Neural Networks from Overfitting*
2015	Long, Shelhamer, Darrell	*Fully Convolutional Networks for Semantic Segmentation*
2016	He, Zhang, Ren, Sun	*Deep Residual Learning for Image Recognition*
2017	G. Huang, Z. Liu, K. Weinberger, L. Maaten	*Densely Connected Convolutional Networks*

History of Neural Networks

- First phase (1950 – 1970)
  - Theoretical concepts of artificial neural networks
  - Perceptron, Adaline, Madaline, …
  - Single layer networks (one layer of trainable weights)
  - First learning algorithms, interesting initial sucesses, then unclear, why they don’t learn further
  - Few (1-100) training patterns
  - Few industrial applications, but successfully used in Modem patent (B. Widrow)
  - Phase ended with publication “Perceptron” of Minsky & Rosenblatt, which critically analyzed Perceptrons

History of Neural Networks

- Second phase (1985 – 1996)
  - Backpropagation (originally published by P. Werbos in 1974)
  - Self-organizing Map (SOM),
  - Learning Vector Quantization (LVQ)
  - Radial Basis function networks (RBF)
  - Multi-layer networks, but gradient loss
  - In practice mostly 2-3 layers, usually less than 5
  - 100 – 10.000 training patterns
  - Medium industrial interest, many applications

History of Neural Networks

- Third phase (2012 – 2020+ ?)
  - Convolutional neural networks
  - Gradient descent for many layers
  - Recurrent LSTM networks for sequence processing
  - ReLu (rectified linear unit) preserves gradient
  - Better regularization (dropout, batch normalization)
  - GPU computing (Nvidia Cuda) speeds NN training
  - Some scientific breakthroughs
    - Google Deepmind AlphaGo
    - ILSVRC challenge
    - Google Translate
  - Huge industrial interest (startup companies frenzy)

Deep Learning Companies

- Google
  - Deepmind: Atari games, AlphaGo,
  - Translate
  - Youtube
- Amazon: Echo, Echo dot
- Facebook: Image search
- IBM: Watson
- Netflix: Recommender systems
- Nvidia: GPUs, GPU servers for Deep learning

- Daimler: Auton. driving
- Bosch: Auton. driving
- GM: Auton. driving
- Apple: Auton. driving
- Tesla: Auton. driving
- Uber: Auton. driving
- Audi: Auton. driving
- BMW: Auton. driving
- VW: Auton. Driving
- Ford: Auton. Driving
- Waymo: Auton. Driving
- …

Deep Learning Hardware

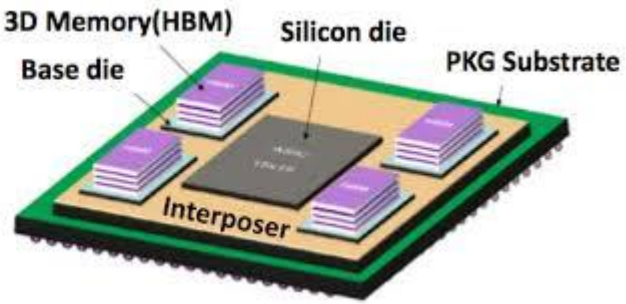
Nvidia Quadro GP100 with Pascal
P100 Chip, 16 GByte HBM2,
Ca. 6.400 €, Feb. 2017

5 Gflop (FP 64)
10 Gflop (FP 32)
20 Gflop (FP 16)

Facebook CNN server with 8 Nvidia Tesla GPUs

Movidius (now Intel)
Fathom Neural Compute Stick

Nervana Systems (now Intel)
Deep Learning Hardware

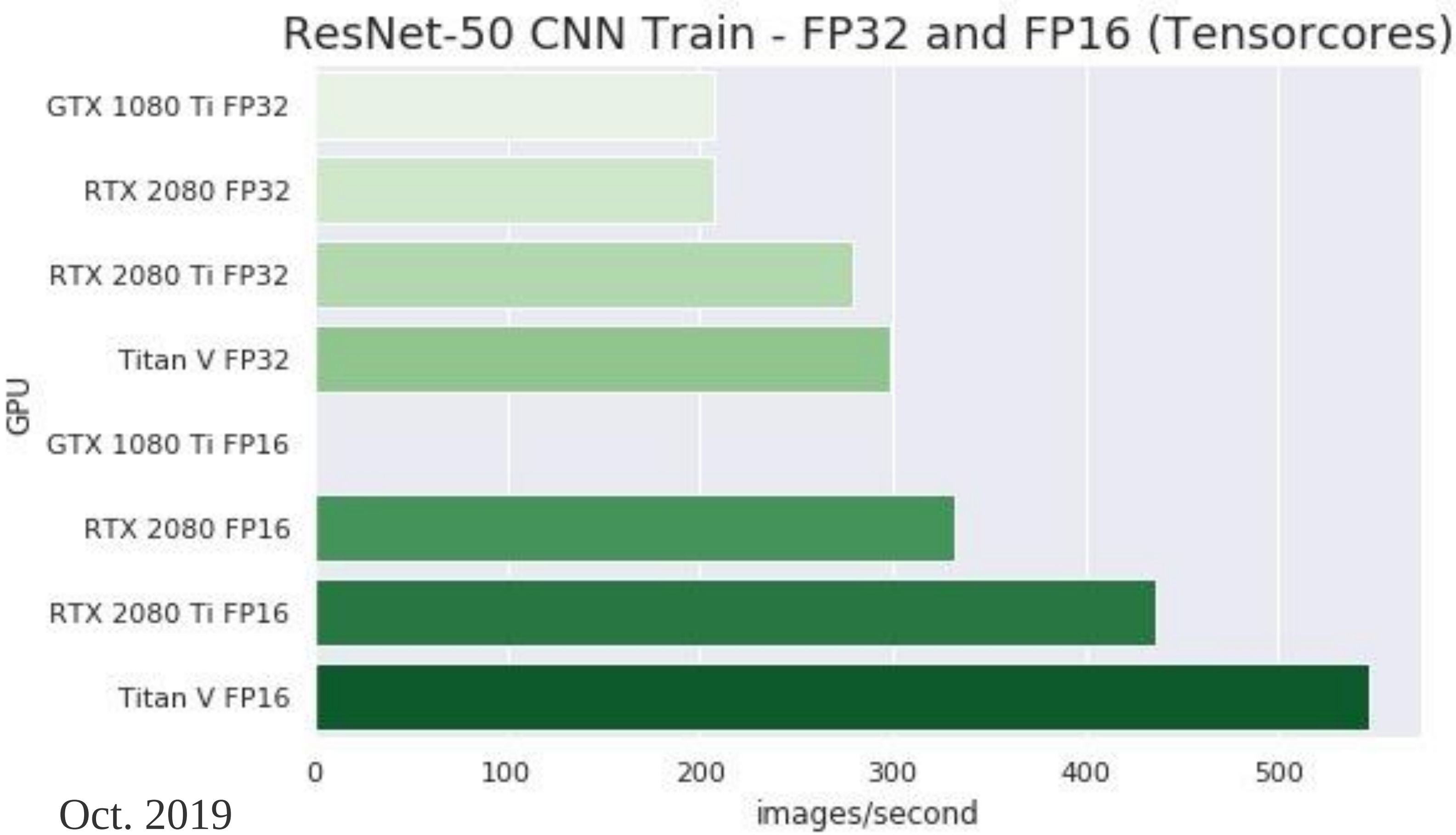


Deep Learning GPUs for Deep NNs

Nvidia	GeForce GTX 1080 Ti	Nvidia Titan X	GeForce GTX 1080
GPU	GP102	GP102	GP104
Fertigung	16 nm FinFET	16 nm FinFET	16 nm FinFET
Transistoren	12 Mrd.	12 Mrd.	7,2 Mrd.
Shader-Rechenkerne	3584	3584	2560
Textureinheiten	224	224	160
Rasterendstufen	88	96	64
GPU-/Turbo-Takt	n.a. / 1600 MHz	1417 / 1531 MHz	1607 / 1733 MHz
Rechenleistung (SP)	11,47 TFlops	10,16 TFlops	8,23 TFlops
Speicher	11 GByte GDDR5X	12 GByte GDDR5X	8 GByte GDDR5X
Speicher-Takt (R/W)	2750 MHz	2502 MHz	2754 MHz
Speicher-Anbindung	352 Bit	384 Bit	256 Bit
Datentransferrate	484 GByte/s	480 GByte/s	320 GByte/s
Stromanschlüsse	1 × 6-pin, 1 × 8-pin	1 × 6-pin, 1 × 8-pin	1 × 8-pin
TDP	250 Watt	250 Watt	180 Watt
Preis ab	ca. 835 Euro	1300 Euro	600 Euro

https://www.heise.de/newsticker/meldung/Nvidia-GeForce-GTX-1080-Ti-Vega-Konkurrent-mit-Titan-X-Leistung-fuer-700-US-Dollar-3640723.html , 1.3.2017

Deep Learning GPUs for Deep NNs



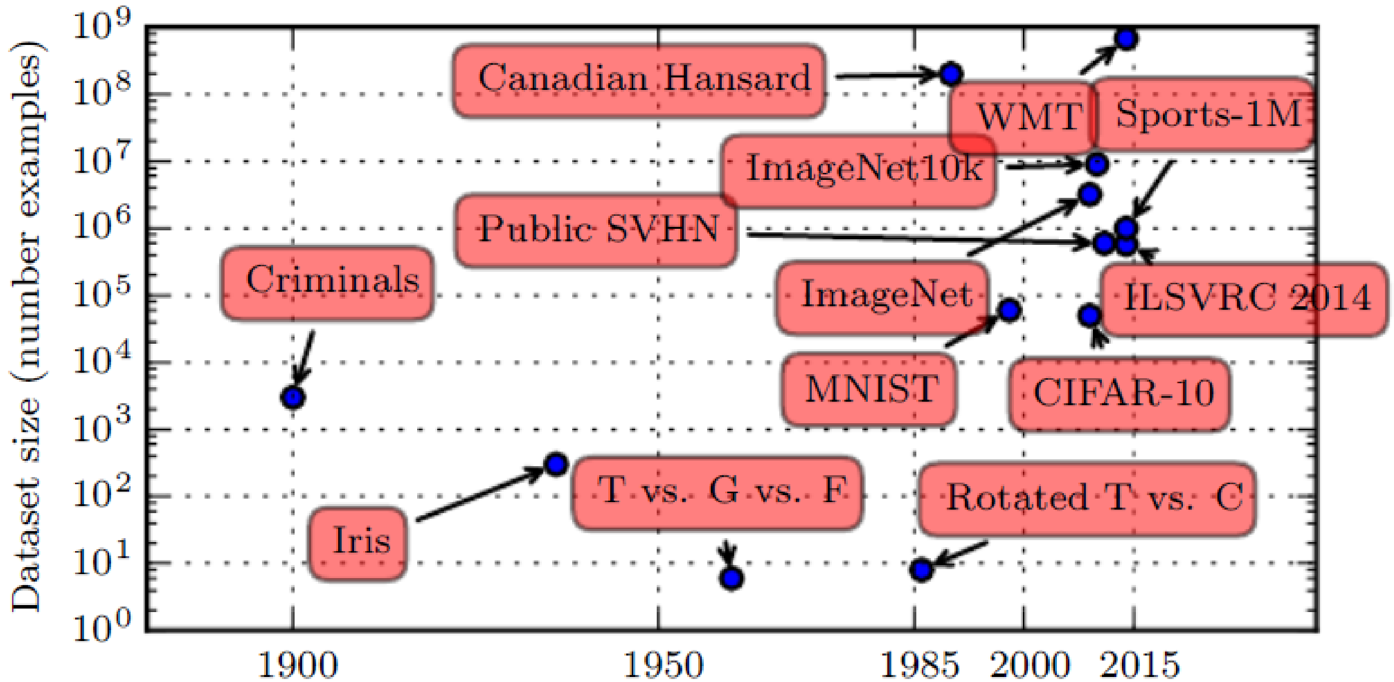
Oct. 2019

https://www.pugetsystems.com/labs/hpc/NVIDIA-RTX-2080-Ti-vs-2080-vs-1080-Ti-vs-Titan-V-TensorFlow-Performance-with-CUDA-10-0-1247/

What has Changed with Deep NNs?

1. New techniques for training deep networks
2. Convolutional networks are now mainstream
3. Increasing dataset sizes
4. Increasing model sizes
5. Increasing accuracy
6. Increasing community size
7. Increasing real-world impact

Increasing Dataset Sizes

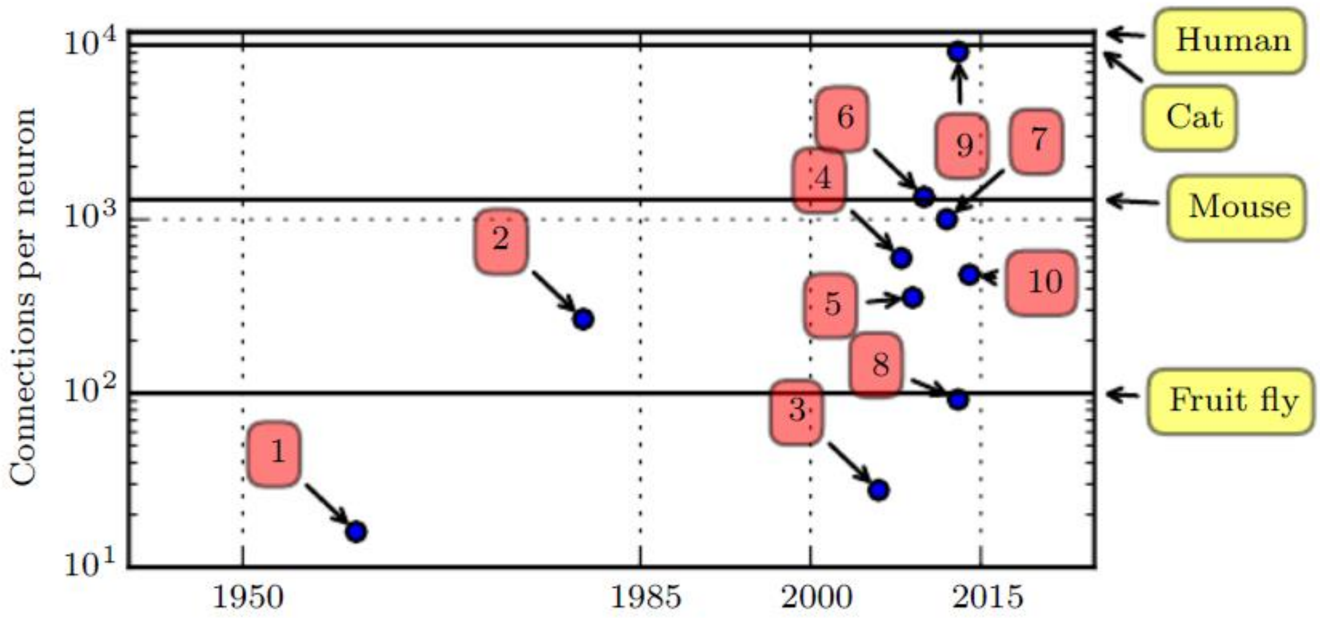


Goodfellow et al.: Deep Learning, Fig. 1.8, p. 19

- MNIST: modified scans of handwritten digits
- CIFAR-10
- ImageNet
- ImageNet10k
- ILSVRC 2014: ImageNet Large-Scale Visual Recognition Challenge

- Canadian Hansard (IBM): translated sentences
- WMT 2014: English to French translation

Increasing # of connections/neuron

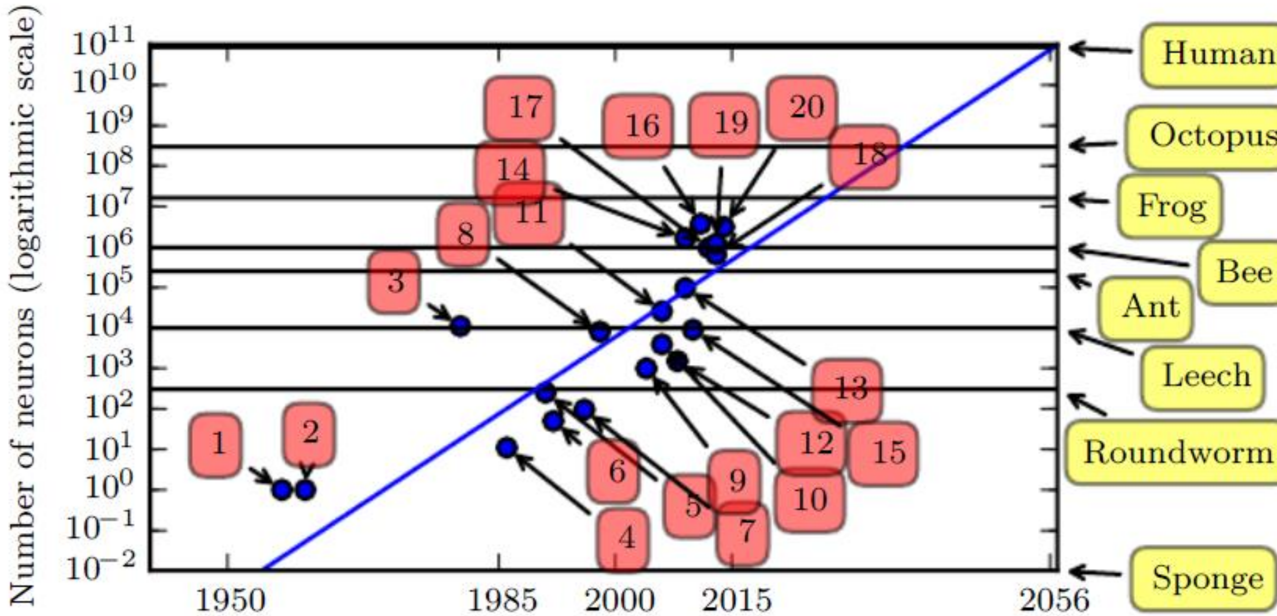


Goodfellow et al.: Deep Learning, Fig. 1.10, p. 22

1. Adaline (Widrow and Hoff, 1960)
2. Neocognitron (Fukushima, 1980)
3. GPU-accelerated convolutional network (Chellapilla *et al.*, 2006)
4. Deep Boltzmann machine (Salakhutdinov and Hinton, 2009a)
5. Unsupervised conv. network (Jarrett *et al.*, 2009)
6. GPU-accelerated MLP (Ciresan *et al.*, 2010)
7. Distributed autoencoder (Le *et al.*, 2012)
8. Multi-GPU conv. network (Krizhevsky *et al.*, 2012)
9. COTS HPC unsupervised conv. network (Coates *et al.*, 2013)
10. GoogLeNet (Szegedy *et al.*, 2014a)

Increasing Neural Network Model Sizes

1. Perceptron (Rosenblatt, 1
2. Adaline (Widrow & Hoff, 1
3. Neocognitron (Fukushima, 1980)
4. Early BP network (Rumelhart *et al.*, 1986)
5. Recurr. NN for speech re (Robinson *et al.* , 1991)
6. MLP for speech recog. (Bengio *et al.*, 1991)
7. Mean field sigm. belief ne
8. LeNet-5 (LeCun *et al.*, 19
9. Echo state network (Jaeger and Haas, 2004)
10. Deep belief network (Hinton *et al.*, 2006)
11. GPU-acc. conv. network
12. Deep Boltzmann machine
13. GPU-acc. deep belief network
14. Unsupervised conv. Network
15. GPU-acc. MLP (Ciresan *et al.*, 2010)

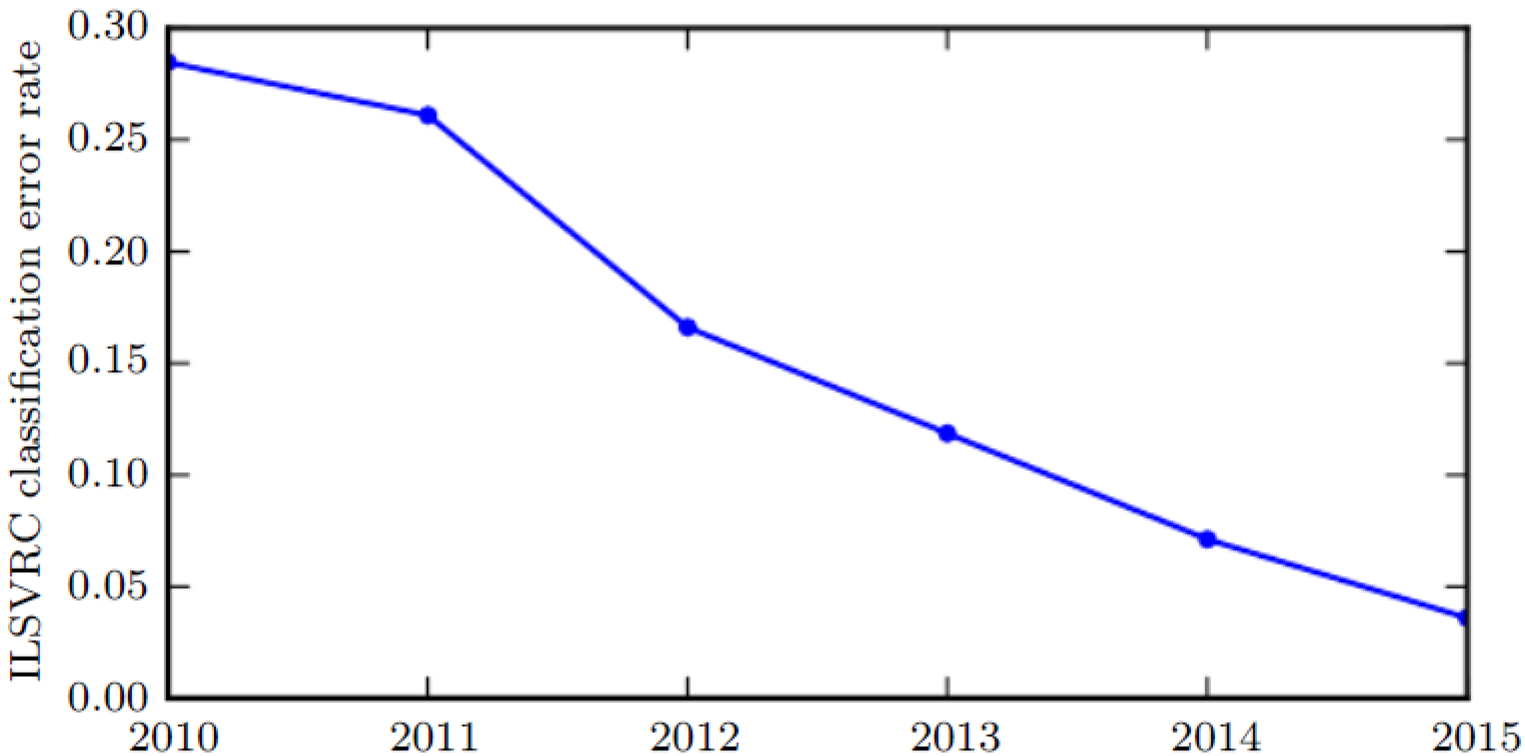


Goodfellow et al.: Deep Learning, Fig. 1.11, p. 23

16. OMP-1 network (Coates and Ng, 2011)
17. Distributed autoencoder (Le *et al.*, 2012)
18. Multi-GPU conv. network (Krizhevsky *et al.*, 2012)
19. COTS HPC unsupervised conv. network (Coates *et al.*, 2013)
20. GoogLeNet (Szegedy *et al.*, 2014a)

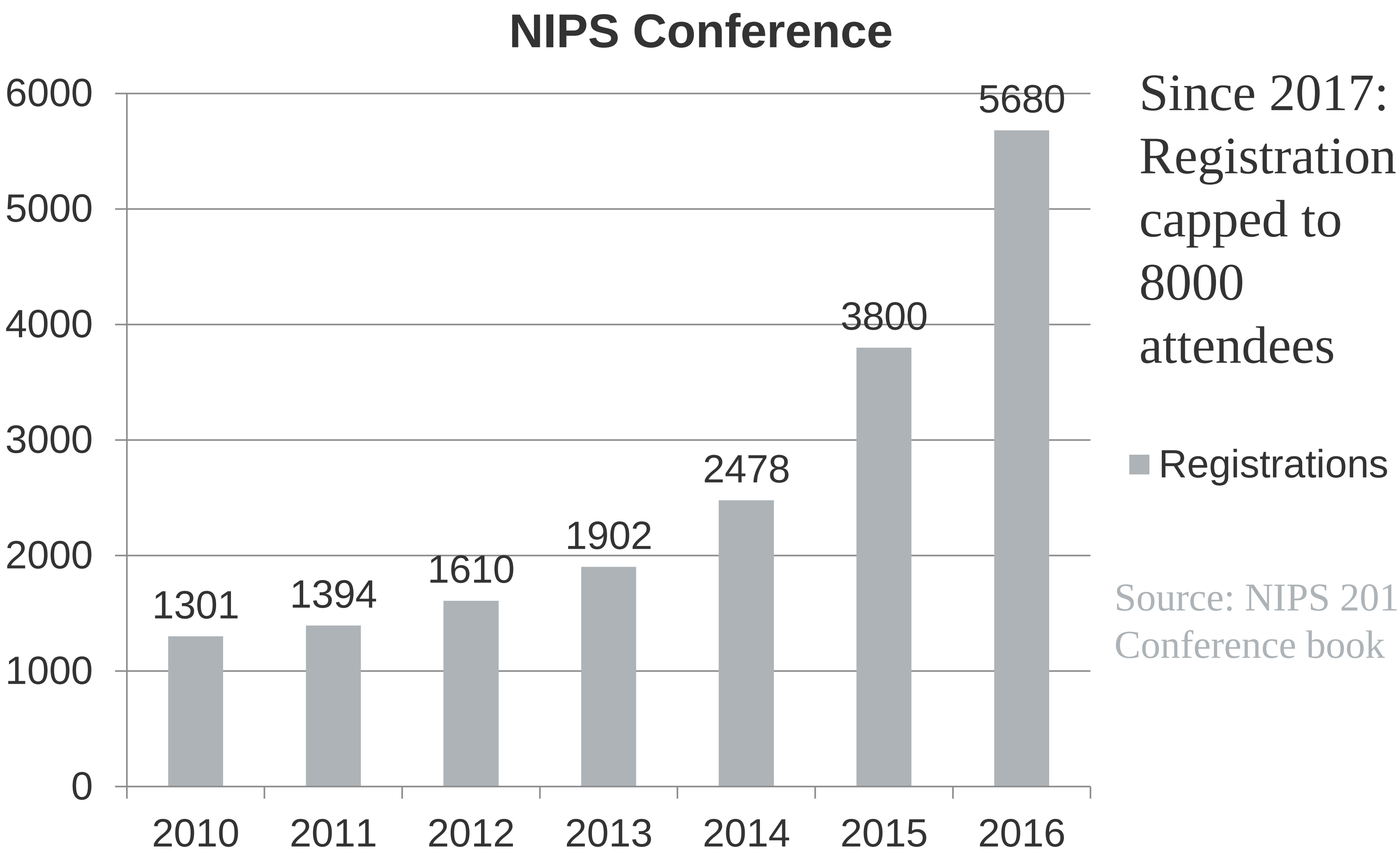
Increasing Accuracy

- ILSVRC top-5 recognition error rate
  (ImageNet Large Scale Visual Recognition Challenge)



Goodfellow et al.: Deep Learning, Fig. 1.12, p. 24

Increasing Community Size



Since 2017: Registration capped to 8000 attendees

Source: NIPS 2016 Conference book

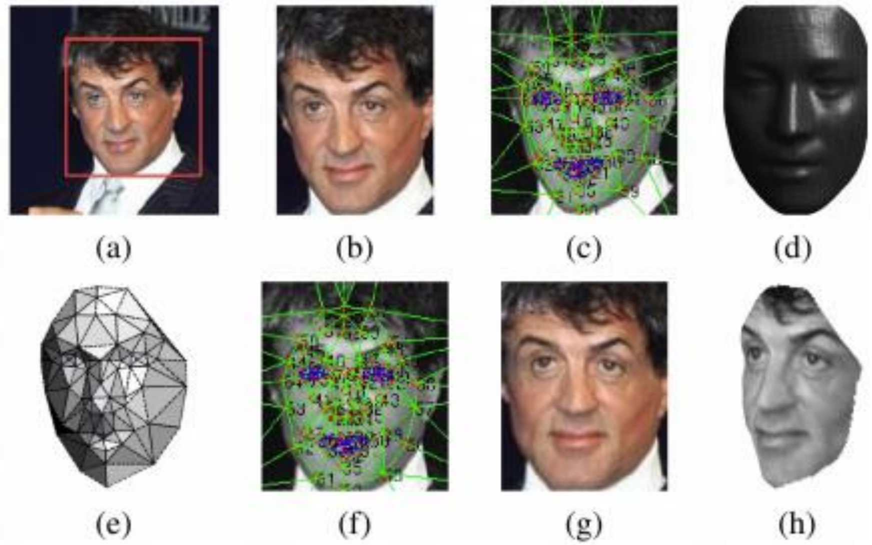
Face Recognition: DeepFace

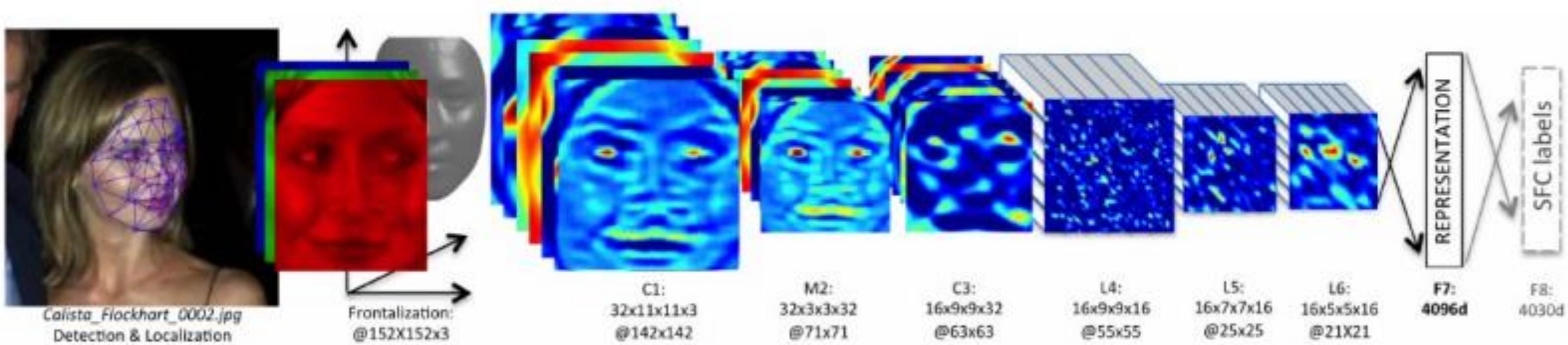
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Face Recognition:DeepFace (Facebook AI Research)

Y LeCun

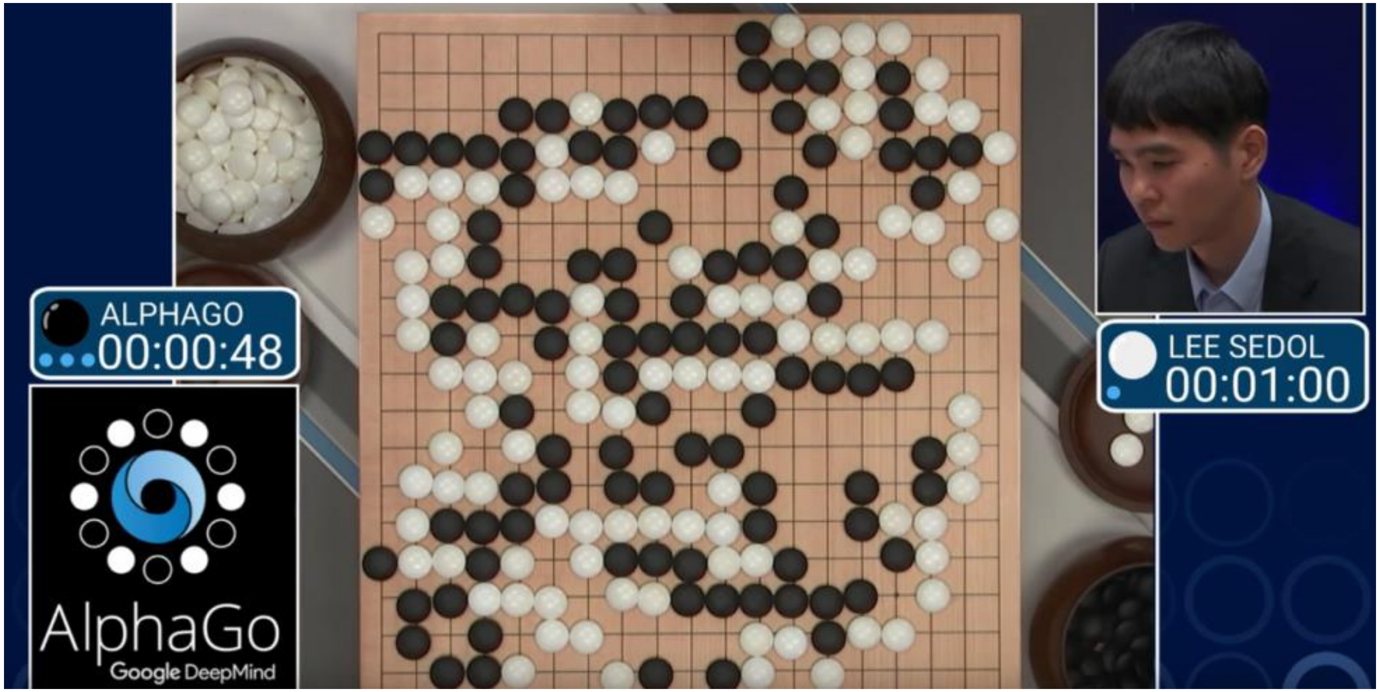
- [Taigman et al. CVPR 2014]
  - Alignment
  - Convnet
  - Close to human performance on frontal views
  - Can now look for a person among 800 millions in 5 seconds
  - Uses 256-bit "compact binary codes"
  - [Gong et al CVPR 2015]





Game Playing: DeepMind AlphaGo

- Go: AlphaGo
  - AlphaGo, developed by Google DeepMind (London) defeated the best professional Go player Lee Sedol (9-dan, Korea) 4:1 in 5 matches, March 2016



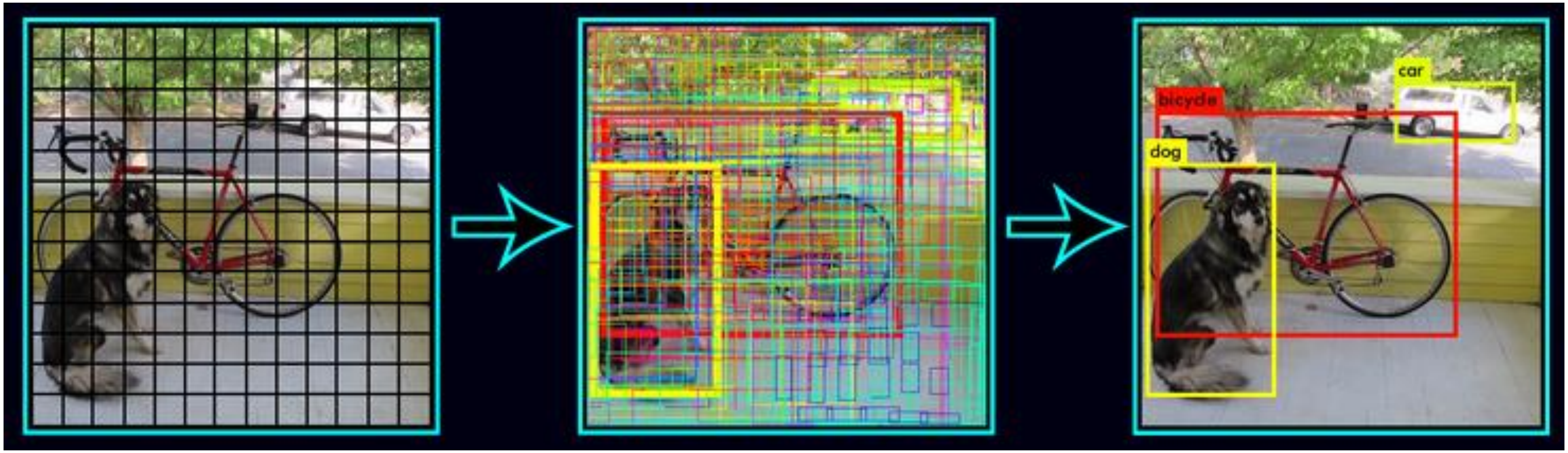
Game Playing: DeepMind AlphaGo

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- AlphaGo used 1,920 CPUs and 280 GPUs in the match against Lee Sedol
- It used Monte Carlo tree search, guided by a value network and a policy network, both implemented using deep neural network technology
- AlphaGo was initially trained on a database of around 30 million moves of expert human players
- Then it was trained further by playing large numbers of games against other instances of itself, using reinforcement learning to improve its play
- https://en.wikipedia.org/wiki/AlphaGo

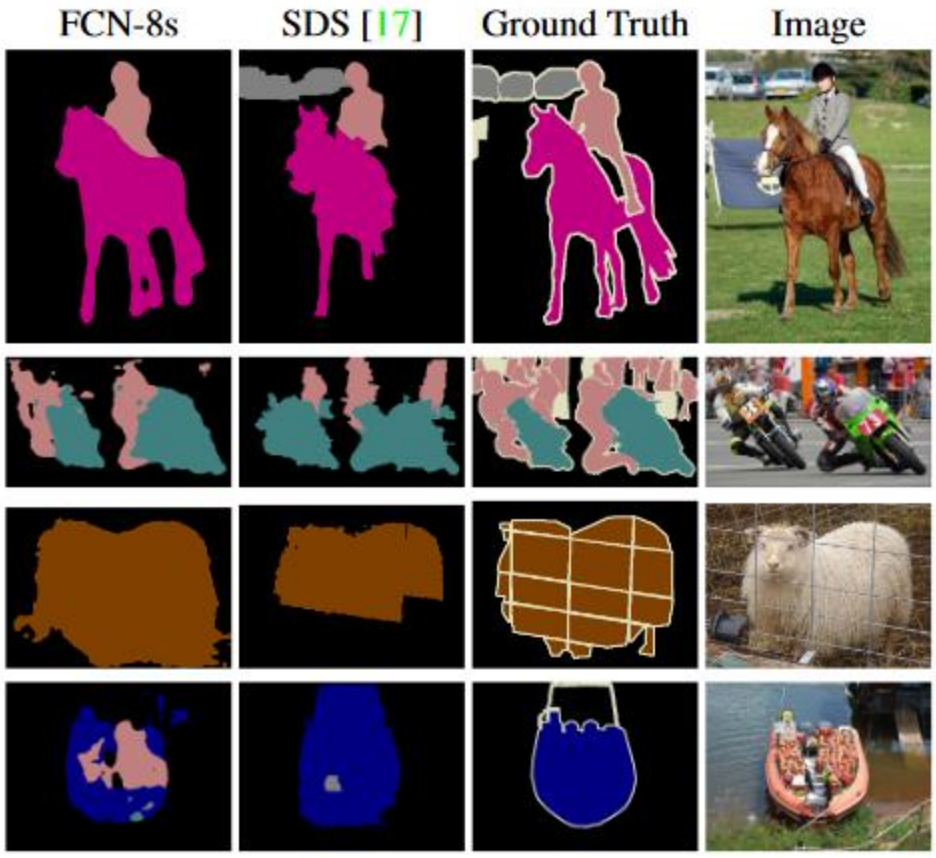
Object recognition: Yolo and Yolo2

- Yolo applies a single neural network to the full image. This network divides the image into regions and predicts bounding boxes and probabilities for each region. These bounding boxes are weighted by the predicted probabilities.
- https://www.youtube.com/watch?v=VOC3huqHrss



Semantic Object Segmentation

- Fully convolutional networks for semantic segmentation
- Perform pixel-wise classification
- SegNet road scene segmentation: https://www.youtube.com/watch?v=CxanE_W46ts
- Full resolution residual networks (FRRNs): https://www.youtube.com/watch?v=PNzQ4PNZSzc

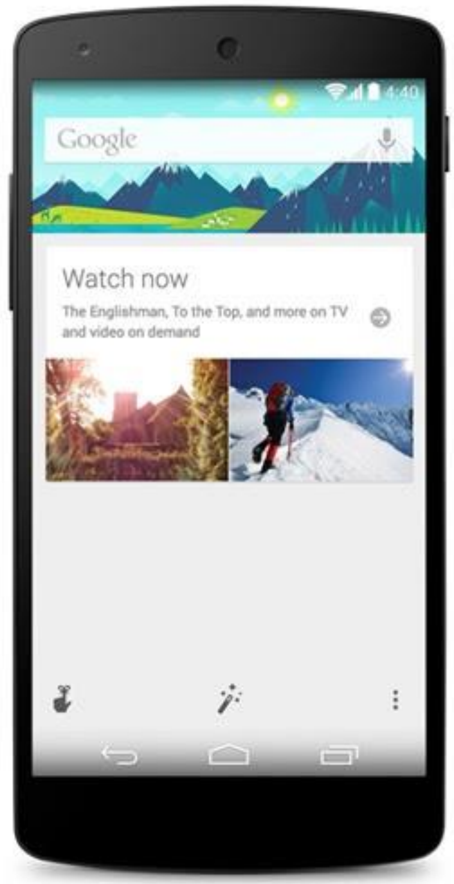


Speech Recognition

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Apple Siri

Google Now



Windows Cortana

Deep Learning Software Libraries

- Theano (Bergstra et al. 2010)
- Caffe (Jia, 2013)
- Torch (Collbert et al. 2011b)
- PyLearn2 (Goodfellow et al., 2013)
- Google Tensorflow (Abadi et al. 2015)
- Microsoft CNTK (Microsoft, 2016)
- Google Tensorflow 2.0 (Google, 2017)
- PyTorch (Paszke, 2017)
- Comparison of Deep Learning Software:
  https://en.wikipedia.org/wiki/Comparison_of_deep_learning_software