



CENG 4480

Embedded System Development & Applications

Lec 02: Deep Learning Basis

Bei Yu

CSE Department, CUHK

byu@cse.cuhk.edu.hk

(Latest update: September 22, 2025)

2025 Fall



① CNN Architecture Overview

② CNN Energy Efficiency



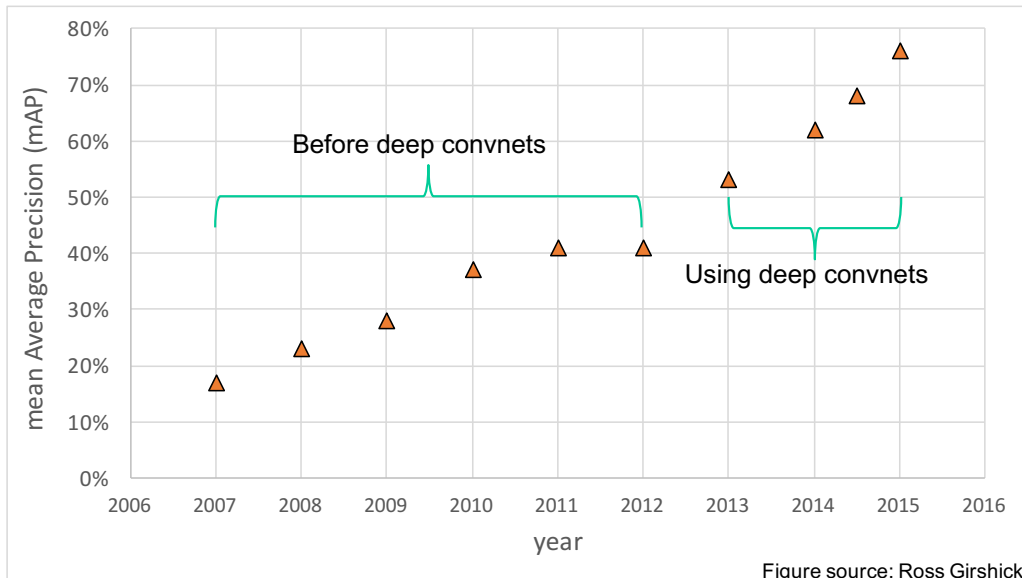
1 CNN Architecture Overview

2 CNN Energy Efficiency

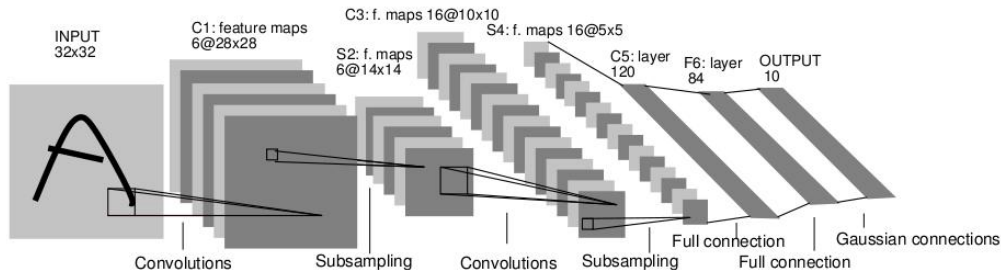
What happened to Object Detection



Object Detection: PASCAL VOC mean Average Precision (mAP)

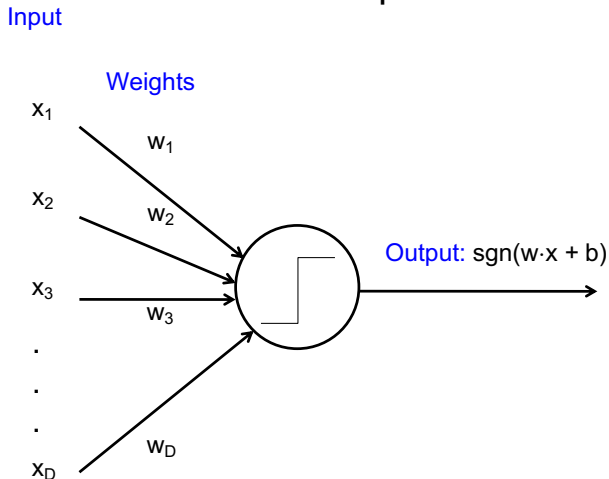


LeNet 5



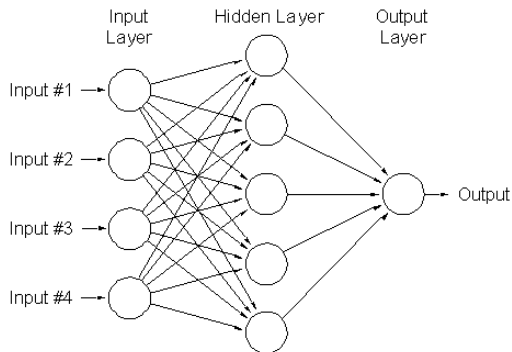
Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, [Gradient-based learning applied to document recognition](#), Proc. IEEE 86(11): 2278–2324, 1998.

The Perceptron

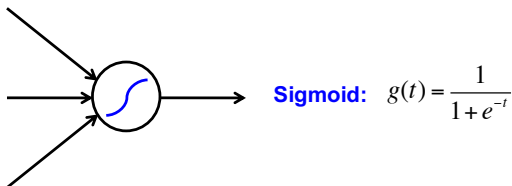


Rosenblatt, Frank (1958), The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Cornell Aeronautical Laboratory, Psychological Review, v65, No. 6, pp. 386–408.

Two-layer neural network



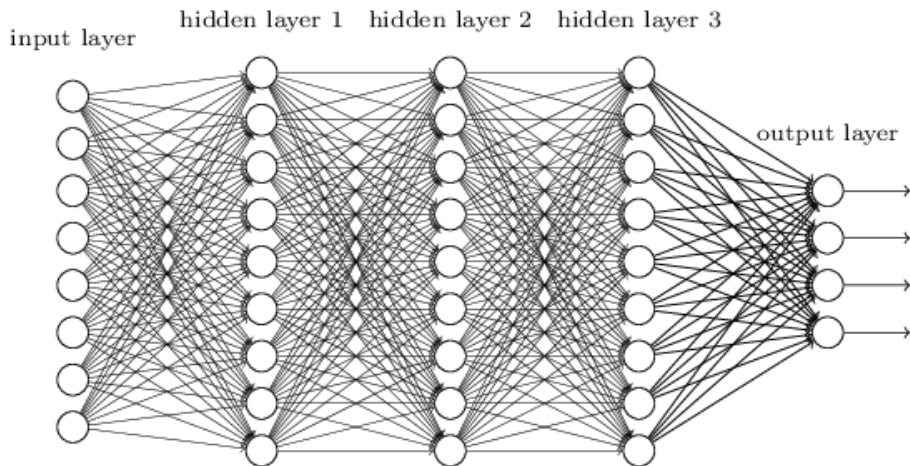
- Can learn nonlinear functions provided each perceptron has a differentiable nonlinearity





What is the value range of sigmoid activation?

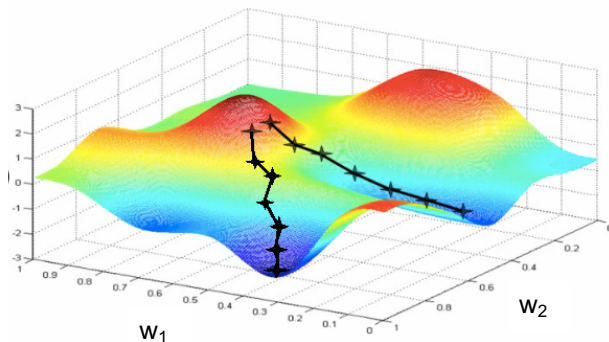
- $[-1, 1]$
- $[-\infty, +\infty]$
- $[0, 1]$
- $[0, +\infty]$



- Find network weights to minimize the *training error* between true and estimated labels of training examples, e.g.:

$$E(\mathbf{w}) = \sum_{i=1}^N (y_i - f_{\mathbf{w}}(\mathbf{x}_i))^2$$

- Update weights by **gradient descent**: $\mathbf{w} \leftarrow \mathbf{w} - \alpha \frac{\partial E}{\partial \mathbf{w}}$



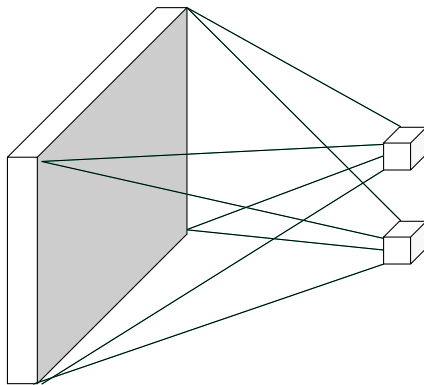


- Find network weights to minimize the *training error* between true and estimated labels of training examples, e.g.:

$$E(\mathbf{w}) = \sum_{i=1}^N (y_i - f_{\mathbf{w}}(\mathbf{x}_i))^2$$

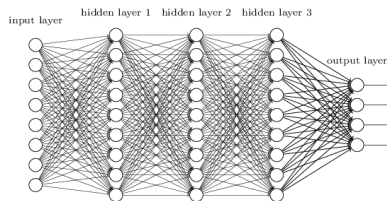
- Update weights by **gradient descent**: $\mathbf{w} \leftarrow \mathbf{w} - \alpha \frac{\partial E}{\partial \mathbf{w}}$
- Back-propagation**: gradients are computed in the direction from output to input layers and combined using chain rule
- Stochastic gradient descent**: compute the weight update w.r.t. one training example (or a small batch of examples) at a time, cycle through training examples in random order in multiple epochs

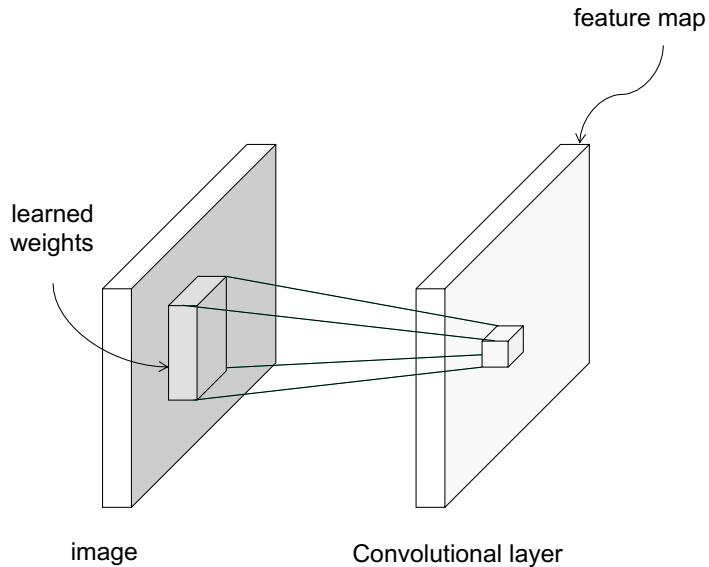
From fully connected to convolutional networks



image

Fully connected layer

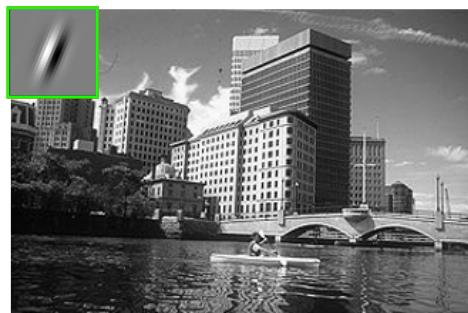




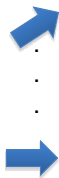


For a convolution kernel with kernel size 3, stride 1, what is the zero padding number to keep the output feature map size unchanged?

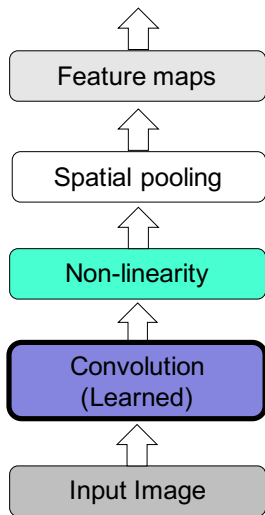
- A: 0
- B: 1
- C: 2
- D: 3



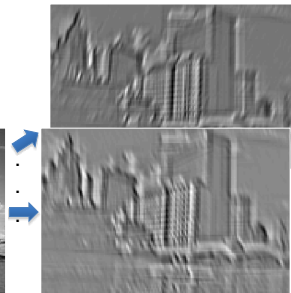
Input



Feature Map

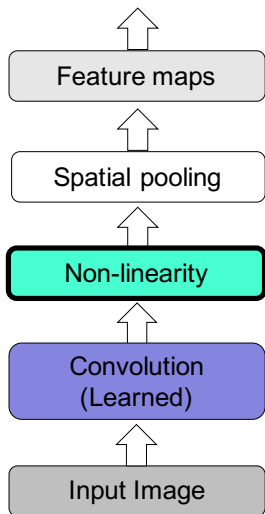


Input

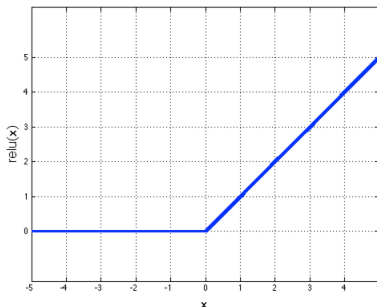


Feature Map

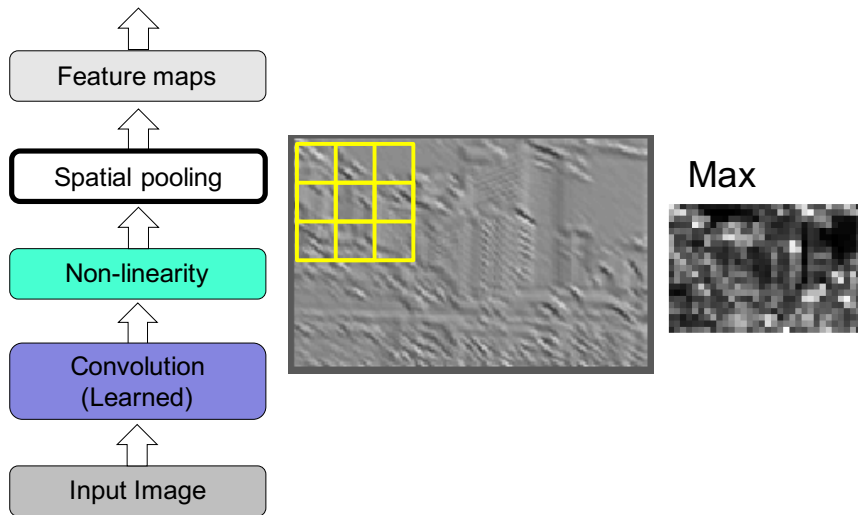
Source: R. Fergus, Y. LeCun



Rectified Linear Unit (ReLU)

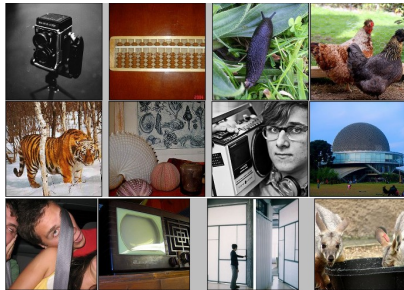


Source: R. Fergus, Y. LeCun



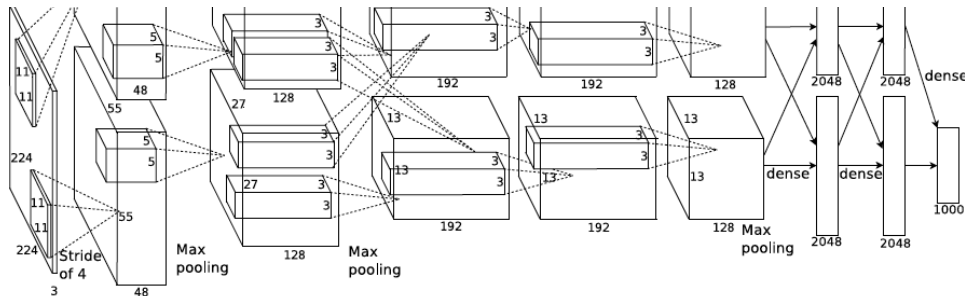
Source: R. Fergus, Y. LeCun

IMGENET



- ~14 million labeled images, 20k classes
- Images gathered from Internet
- Human labels via Amazon MTurk
- ImageNet Large-Scale Visual Recognition Challenge (ILSVRC):
1.2 million training images, 1000 classes

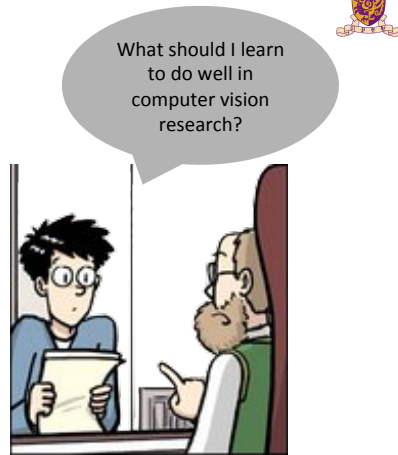
www.image-net.org/challenges/LSVRC/



- Similar framework to LeNet but:
 - Max pooling, ReLU nonlinearity
 - More data and bigger model (7 hidden layers, 650K units, 60M params)
 - GPU implementation (50x speedup over CPU)
 - Trained on two GPUs for a week
 - Dropout regularization

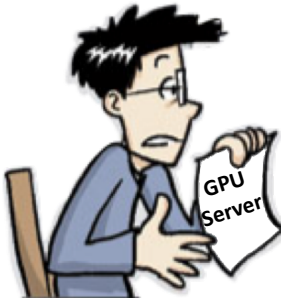
A. Krizhevsky, I. Sutskever, and G. Hinton, [ImageNet Classification with Deep Convolutional Neural Networks](#), NIPS 2012





DEEP LEARNING

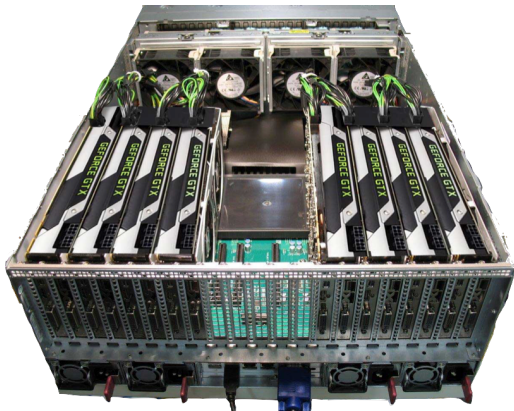






State of the art recognition methods

- Very Expensive
 - Memory
 - Computation
 - Power





HGX A100 4-GPU

Hyperplane 4-A100



- ✓ 4 NVIDIA® A100 SXM4 GPUs
- ✓ NVLink™ GPU connectivity
- ✓ 2 rack unit form factor (2U)
- ✓ 2x AMD EPYC processors with up to 64 cores
- ✓ Up to 4 TB of system memory

Starting at

\$70,376.00

HGX A100 8-GPU

Hyperplane 8-A100



- ✓ 8 NVIDIA® A100 SXM4 GPUs
- ✓ NVLink™ + NVSwitch™ GPU connectivity
- ✓ 4 rack unit form factor (4U)
- ✓ 2x AMD EPYC processors with up to 64 cores
- ✓ Up to 4 TB of system memory

Starting at

\$143,270.00

HGX H100 8-GPU

Hyperplane 8-H100



- ✓ 8 NVIDIA® H100 SXM5 GPUs
- ✓ NVLink™ + NVSwitch™ GPU connectivity
- ✓ 8 rack unit form factor (8U)
- ✓ 2x Intel Xeon 8480+ 56-core processors
- ✓ Up to 4 TB of system memory

Pre-order

Starting at

\$311,221.00

Refer:

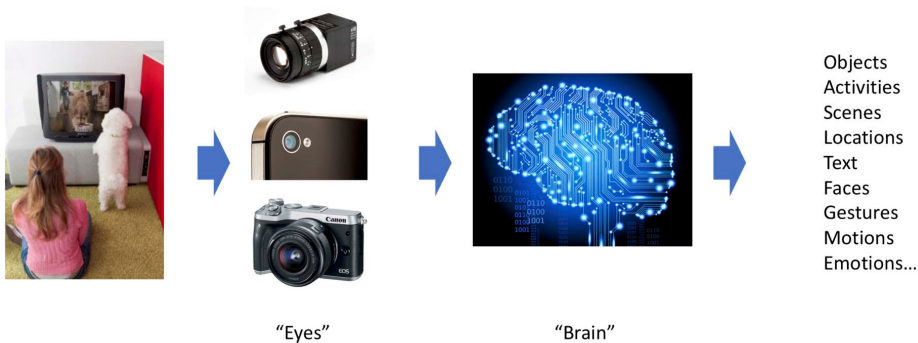
<https://shop.lambdalabs.com/deep-learning/servers/hyperplane/customize>

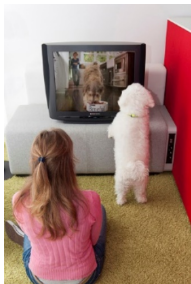


1 CNN Architecture Overview

2 CNN Energy Efficiency

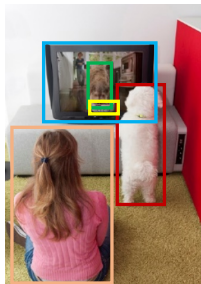
- Humans use their **eyes** and their **brains** to visually sense the world.
- Computers use their **cameras** and **computation** to visually sense the world





Classification

Image



Detection

Region



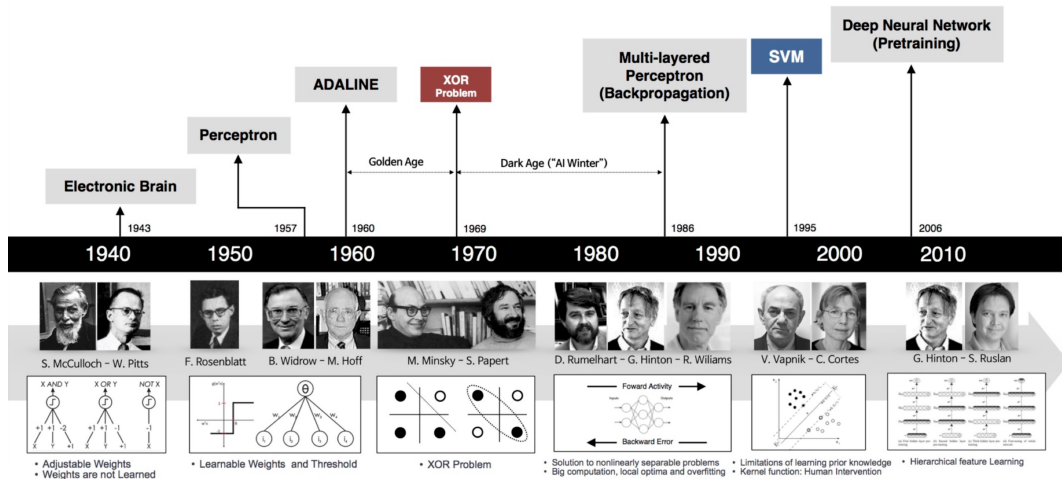
Segmentation

Pixel



Sequence

Video



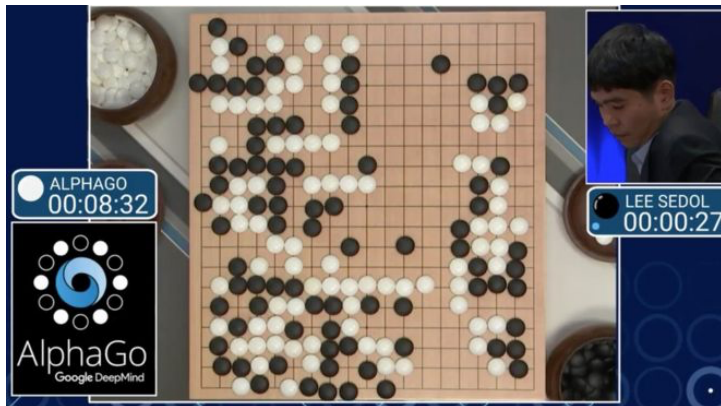


- The rises of SVM, Random forest
- No theory to play
- Lack of training data
- Benchmark is insensitive
- Difficulties in optimization
- Hard to reproduce results

Curse

“Deep neural networks are no good and could never be trained.”

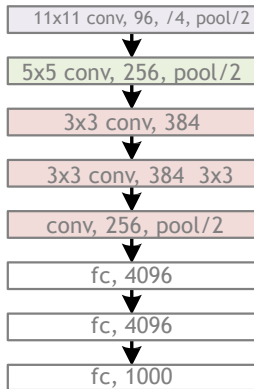
- A fast learning algorithm for deep belief nets. [Hinton et.al 1996]
- Data + Computing + Industry Competition
- NVidia's GPU, Google Brain (16,000 CPUs)
- **Speech**: Microsoft [2010], Google [2011], IBM
- **Image**: AlexNet, 8 layers [Krizhevsky et.al 2012] (26.2% -> 15.3%)





Revolution of Depth

AlexNet, 8
layers
(ILSVRC 2012)

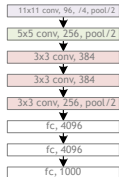


Slide Credit: He et al. (MSRA)

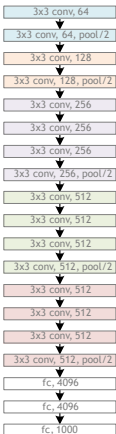


Revolution of Depth

AlexNet, 8
layers
(ILSVRC 2012)



VGG, 19
layers
(ILSVRC 2014)



GoogleNet, 22
layers
(ILSVRC 2014)



Slide Credit: He et al. (MSRA)



Revolution of Depth

AlexNet, 8
layers
(ILSVRC 2012)



VGG, 19
layers
(ILSVRC 2014)



ResNet, 152
layers
(ILSVRC 2015)



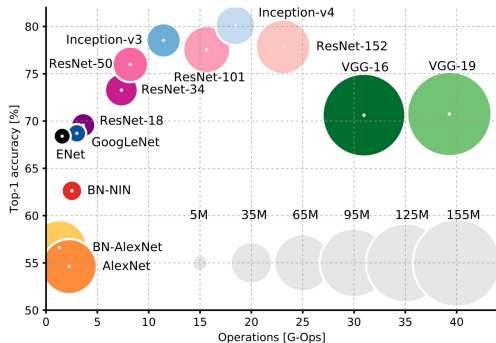
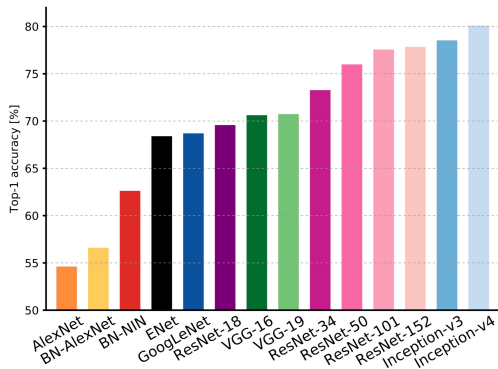
Slide Credit: He et al. (MSRA)

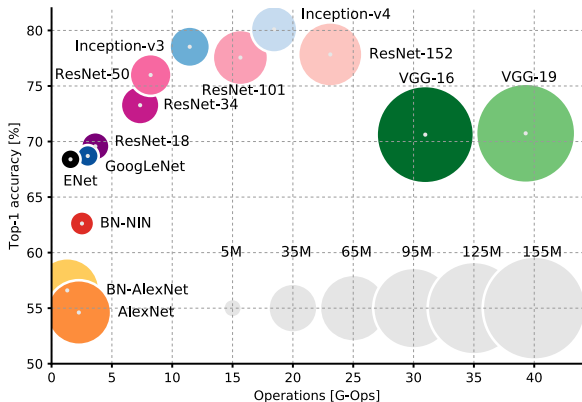


- AlexNet (Krizhevsky, Sutskever, and E. Hinton 2012) 233MB
- Network in Network (Lin, Chen, and Yan 2013) 29MB
- VGG (Simonyan and Zisserman 2015) 549MB
- GoogleNet (Szegedy, Liu, et al. 2015) 51MB
- ResNet (He et al. 2016) 215MB
- Inception-ResNet (Szegedy, Vanhoucke, et al. 2016)
- DenseNet (Huang et al. 2017)
- Xception (Chollet 2017)
- MobileNetV2 (Sandler et al. 2018)
- ShuffleNet (Zhang et al. 2018)



- AlexNet (Krizhevsky, Sutskever, and E. Hinton 2012) 233MB
- Network in Network (Lin, Chen, and Yan 2013) 29MB
- VGG (Simonyan and Zisserman 2015) 549MB
- GoogleNet (Szegedy, Liu, et al. 2015) 51MB
- ResNet (He et al. 2016) 215MB
- Inception-ResNet (Szegedy, Vanhoucke, et al. 2016) 23MB
- DenseNet (Huang et al. 2017) 80MB
- Xception (Chollet 2017) 22MB
- MobileNetV2 (Sandler et al. 2018) 14MB
- ShuffleNet (Zhang et al. 2018) 22MB





Why AlexNet is large in size, but small in operations?

- A: Special FC layers
- B: Special Conv layers
- C: More channels
- D: Some redundant operators

Autonomous drive

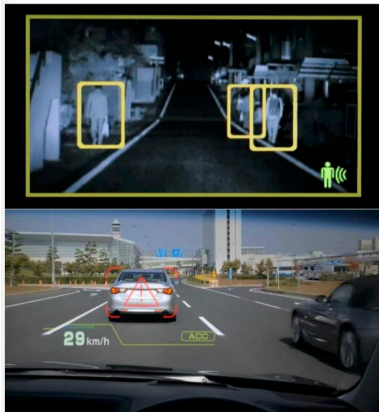
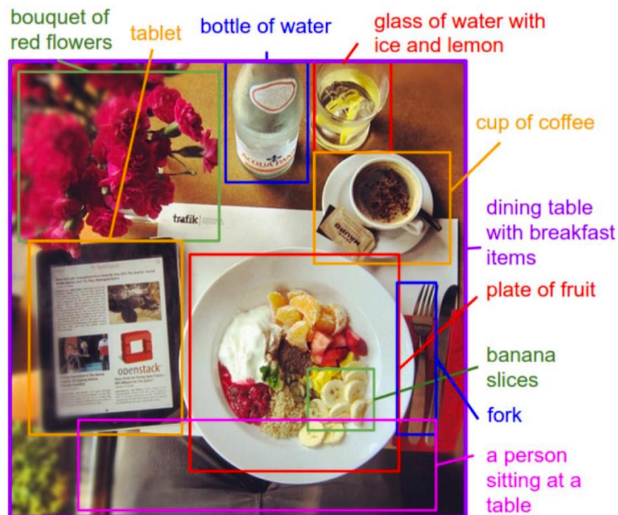
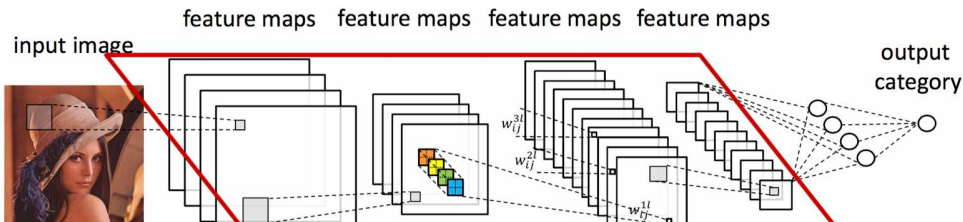


Image recognition

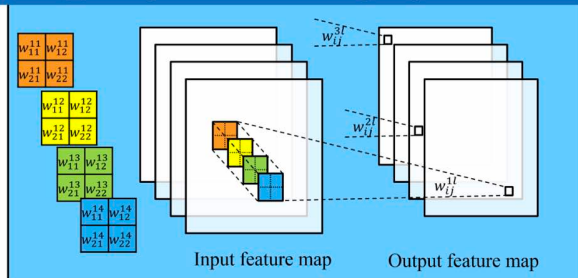


Convolutional Neural Network (CNN)

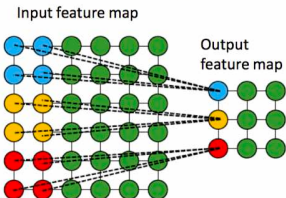


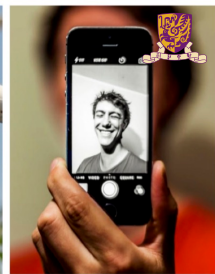
Convolutional layers account for over 90% computation

- [1] A. Krizhevsky, etc. Imagenet classification with deep convolutional neural networks. NIPS 2012.
- [2] J. Cong and B. Xiao. Minimizing computation in convolutional neural networks. ICANN 2014



Max-pooling is optional





Embedded CV



Hisense core|photonics

[Start Video](#)