Convolutional Neural Networks

ECE 449

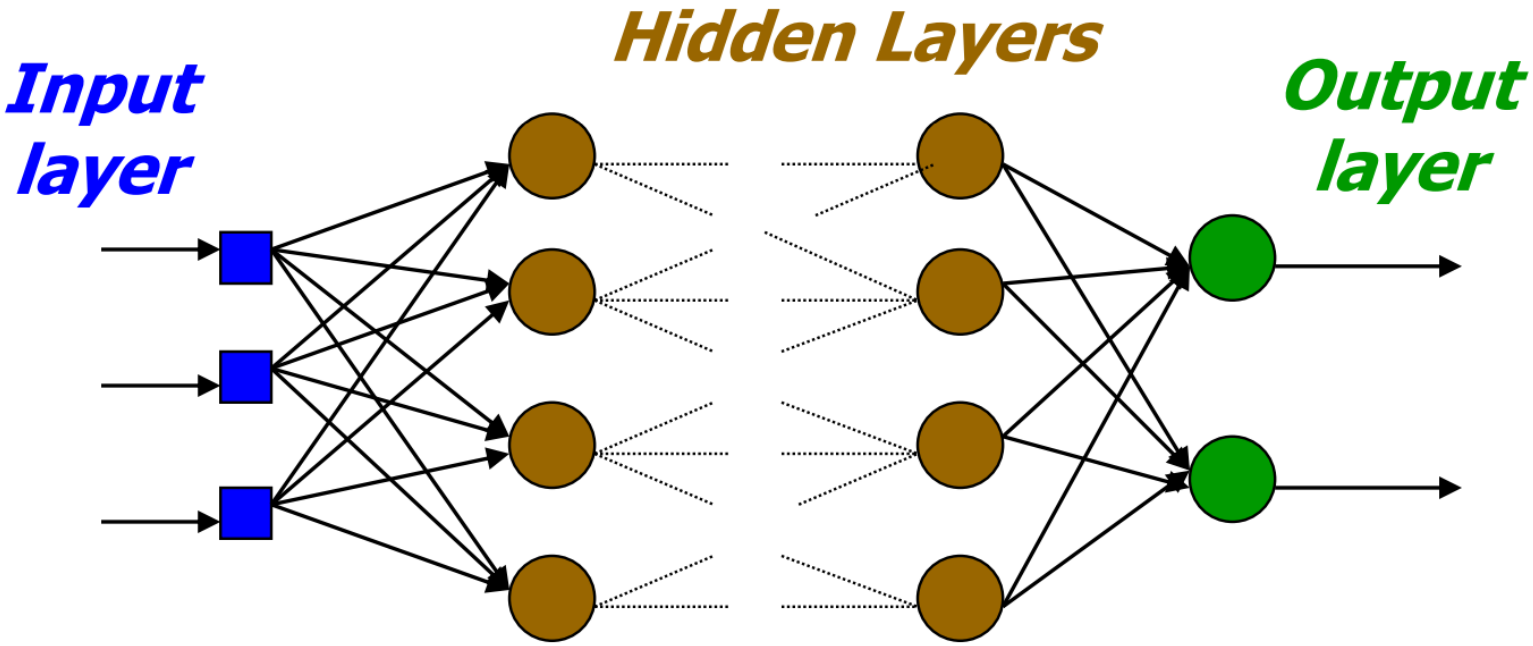
Outline

- From multi-layer perceptron (MLP) to deep neural networks (DNN)
- CNN
- ImageNet competition

CNN

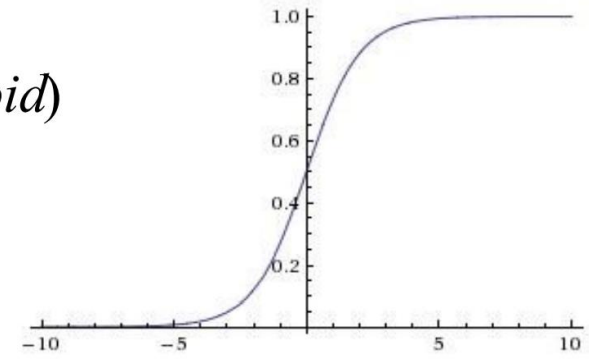
- **From multi-layer perceptron (MLP) to deep neural networks (DNN)**
- CNN
- ImageNet competition

Multi-Layer Perceptron



$$a = f(\mathrm{u}) = \frac{1}{1+e^{-u}} \,(sigmoid)$$

$$\frac{df(u)}{du} = f \times (1-f)$$

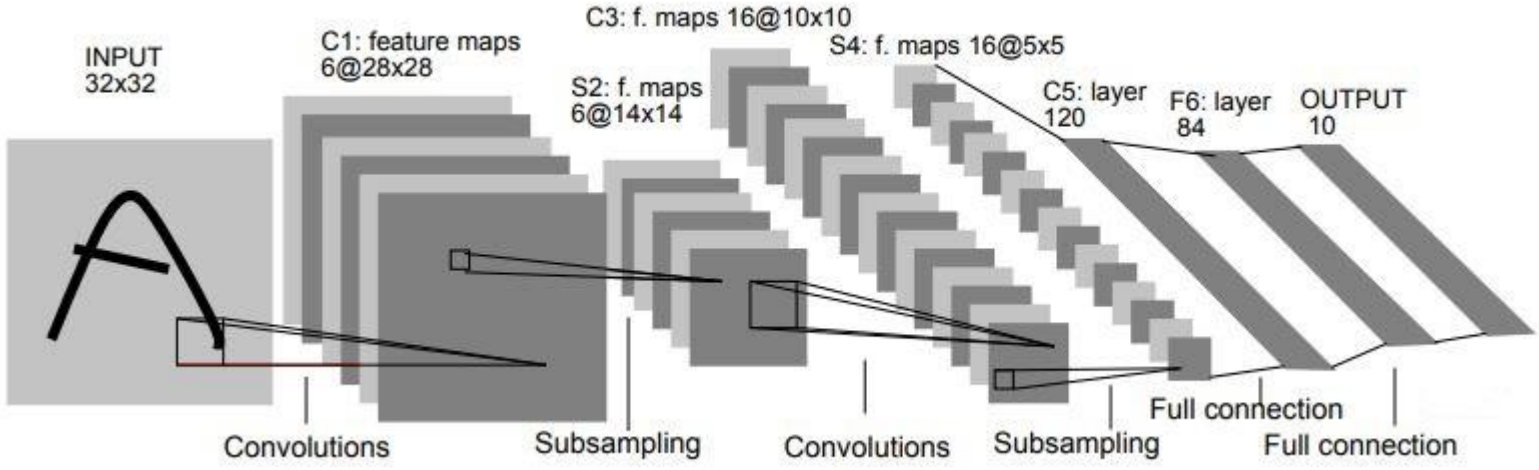


Multi-Layer Perceptron

- Some drawbacks of MLP
  - Vanishing gradient
    - Cannot go deep
    - Stuck in local minimum
  - Initialization sensitive
  - Need feature extraction
  - Model complexity

LeNet-5

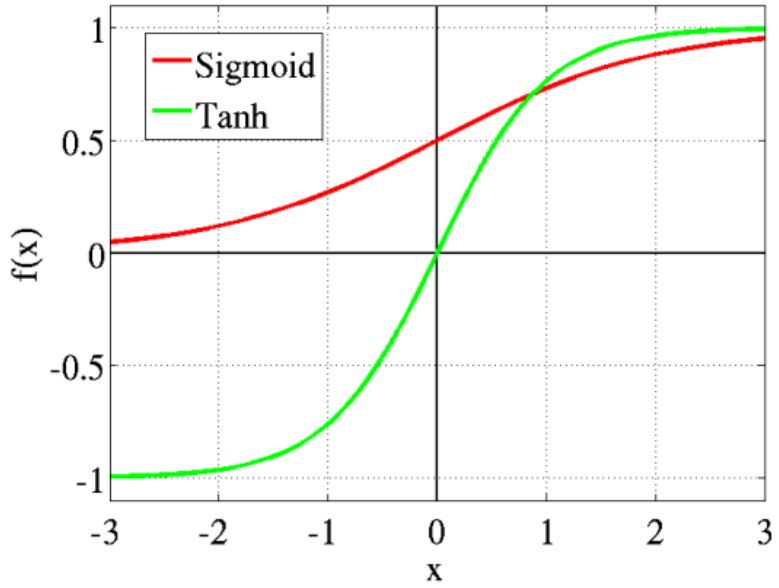
- Character recognition in 1990's



Layer		Feature Map	Size	Kernel Size	Stride	Activation
Input	Image	1	32x32	-	-	-
1	Convolution	6	28x28	5x5	1	tanh
2	Average Pooling	6	14x14	2x2	2	tanh
3	Convolution	16	10x10	5x5	1	tanh
4	Average Pooling	16	5x5	2x2	2	tanh
5	Convolution	120	1x1	5x5	1	tanh
6	FC	-	84	-	-	tanh
Output	FC	-	10	-	-	softmax

LeNet-5

- Difference between LeNet and MLP
  - Introduce convolutions (weight sharing)
  - Pooling (subsampling)
  - No feature extraction as pre-processing
  - Use tanh instead of sigmoid
- $\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$



CNN

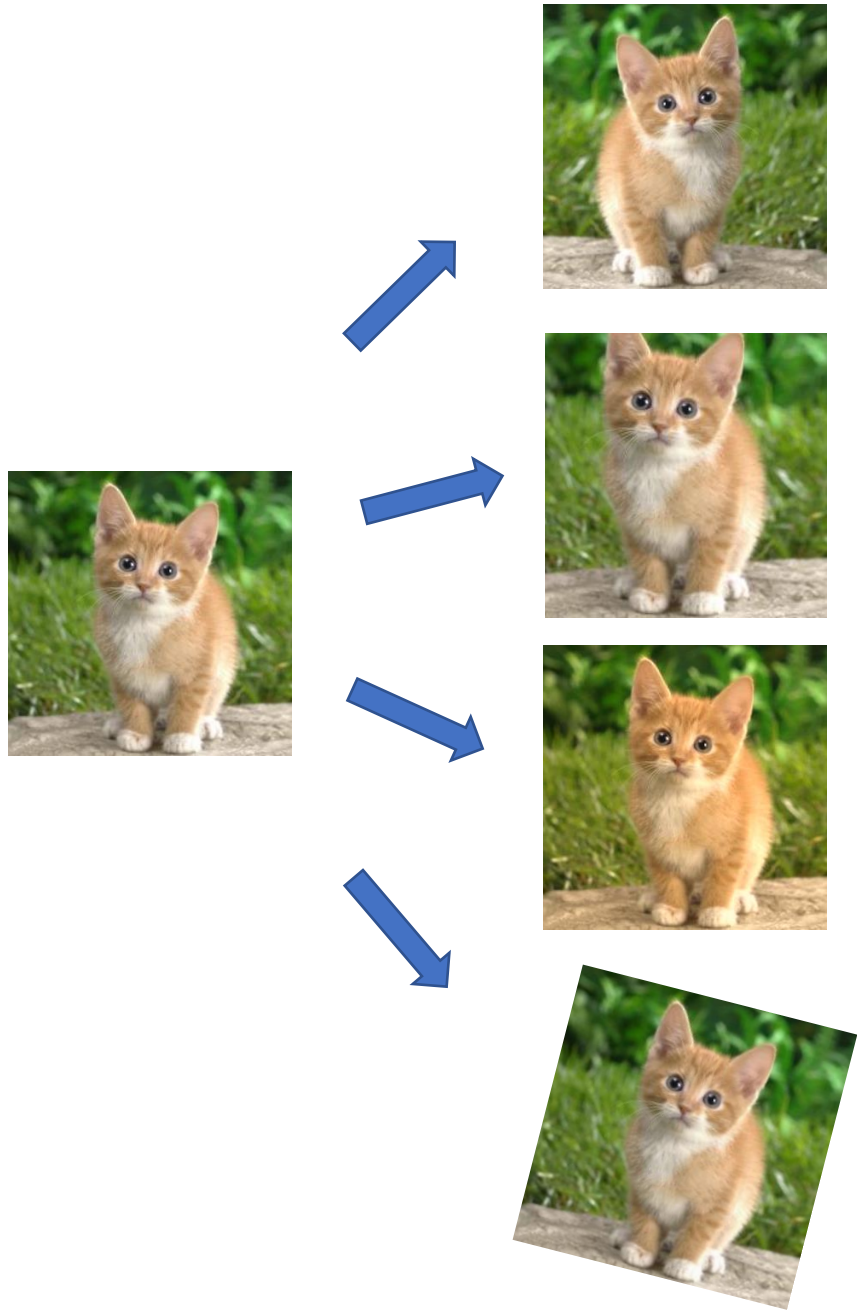
- From multi-layer perceptron (MLP) to deep neural networks (DNN)
- **CNN**
- ImageNet competition

What You should Know about CNN?

- Data
  - Augmentation
- Architecture
  - Convolution
  - Pooling
  - Activation
- Loss
  - …
  - Regularization
- Training
  - Optimizer
  - Dropout

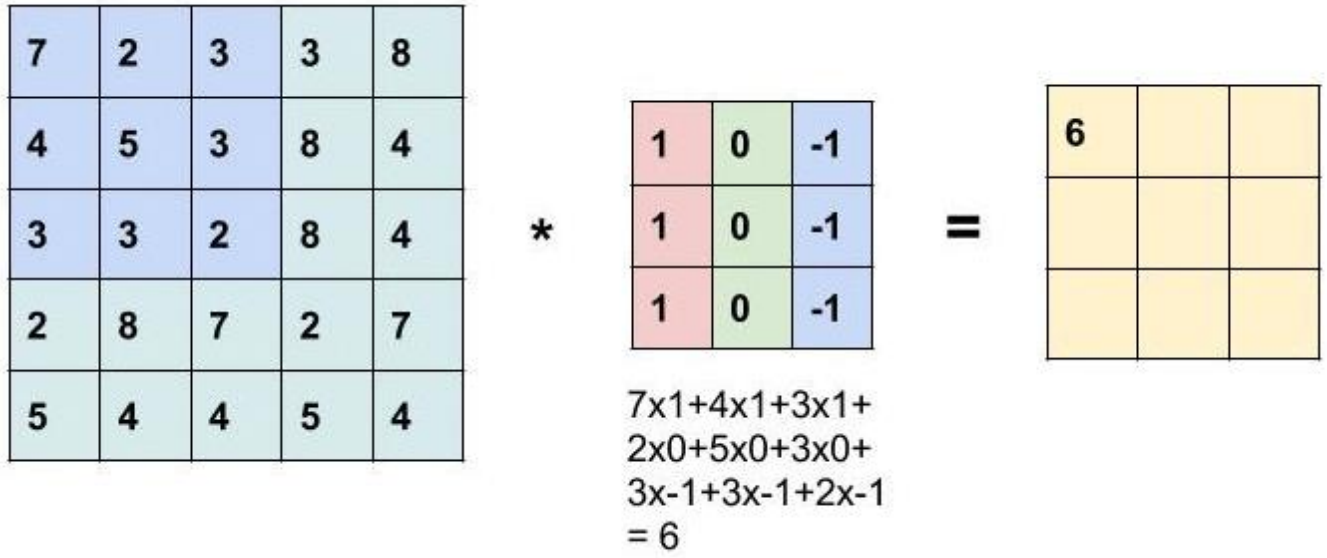
Data

- As “big” as possible
  - Number
  - Diversity / Distribution
- Augmentation
  - Flip
  - Random crop
  - Color distortion
  - Rotation



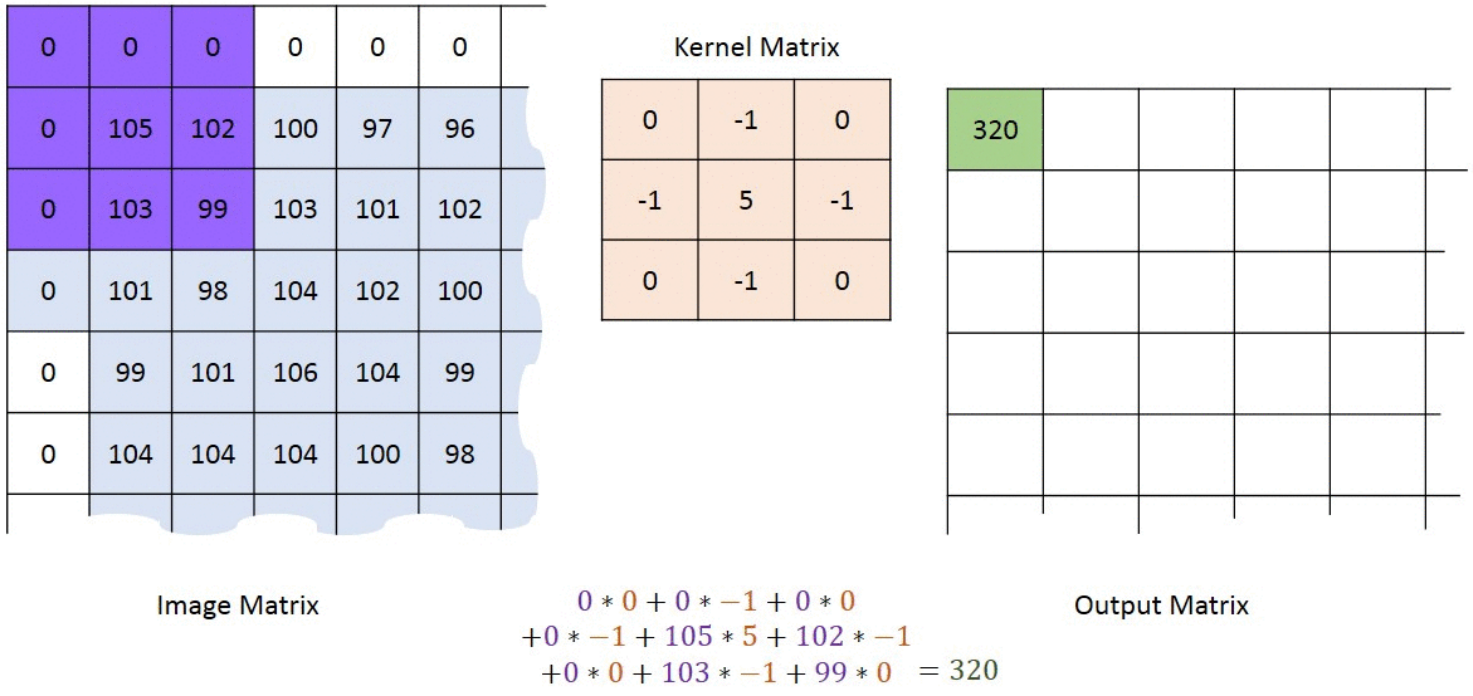
Convolution

- Basic convolution
  - Example, 5×5 image with 3×3 kernel.



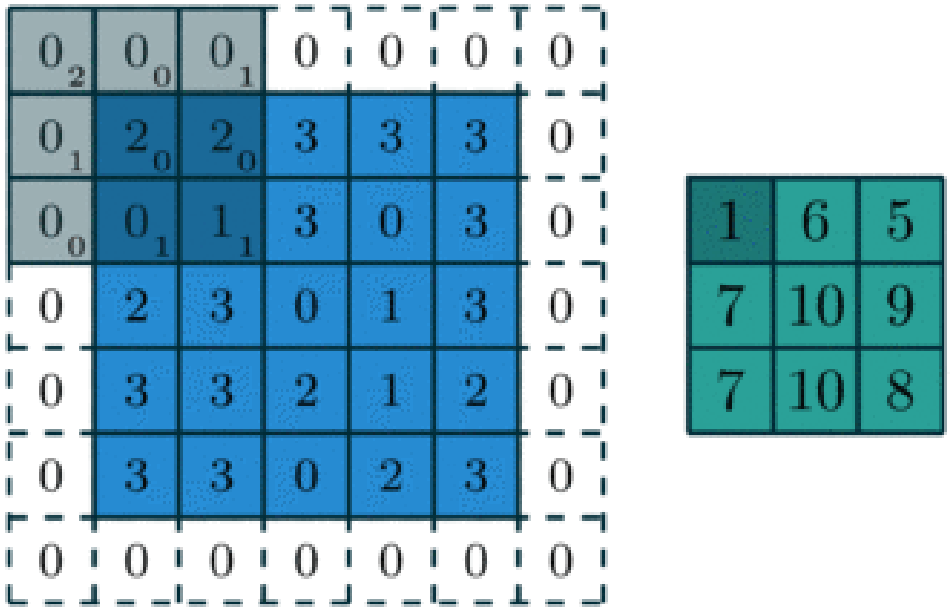
Convolution

- Padding



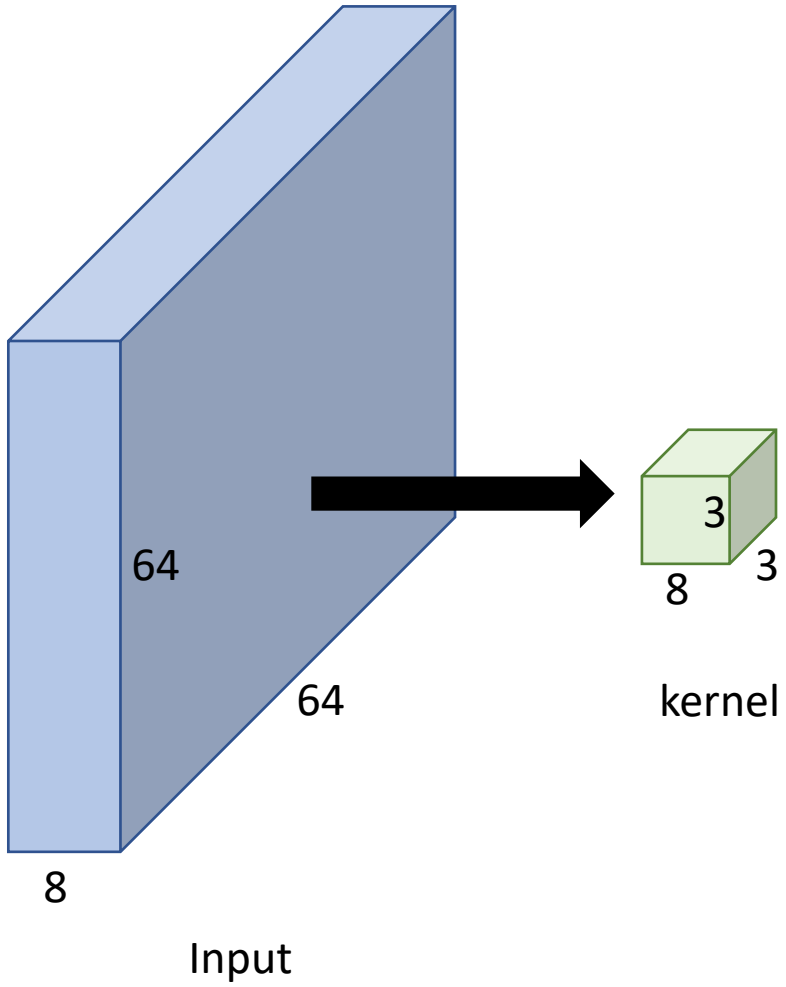
Convolution

- Stride (2×2)



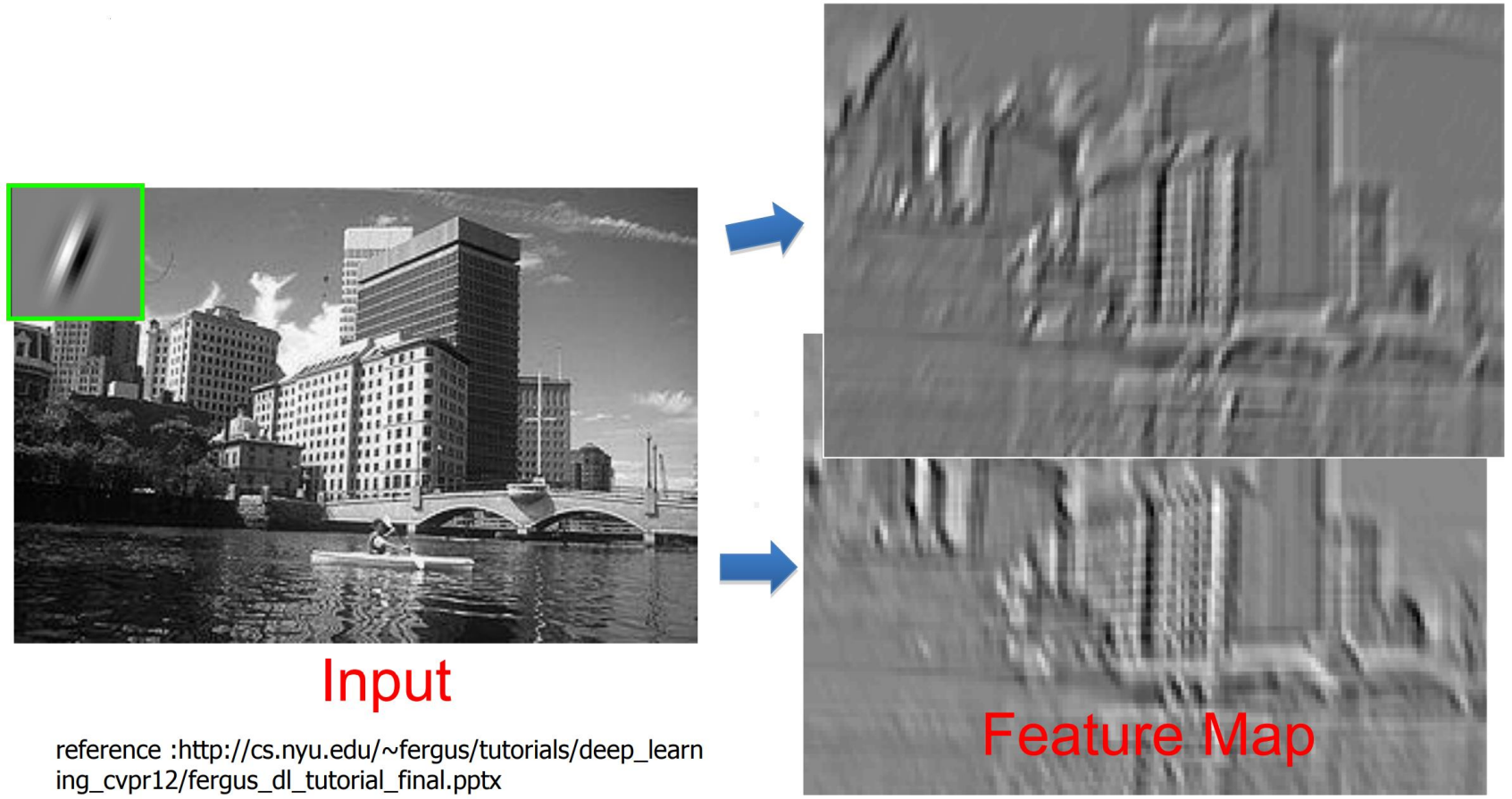
Convolution

- Input: 64×64 patch with 8 channels.
- Kernel: each has 3×3×8.
- Output dimension? How many parameters?
  - Assume we have 32 kernels.
  - No padding.
  - 1×1 stride.



Convolution

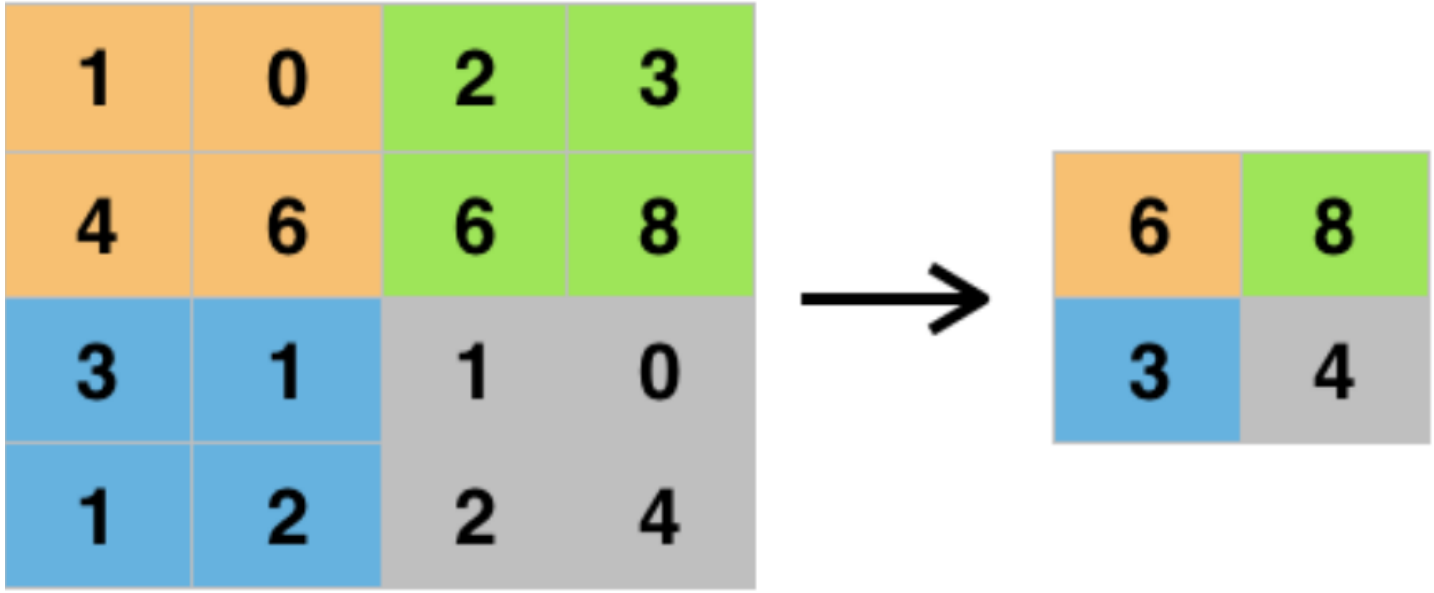
- Visualization



reference :http://cs.nyu.edu/~fergus/tutorials/deep_learning_cvpr12/fergus_dl_tutorial_final.pptx

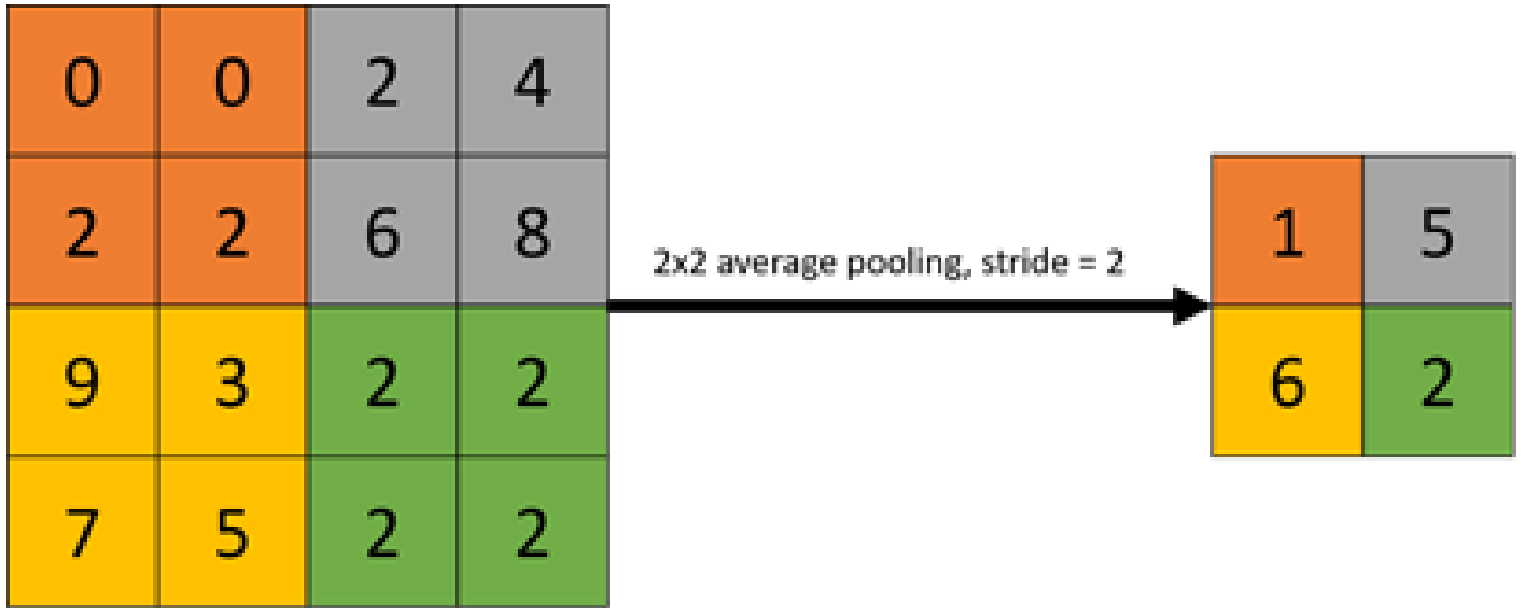
Pooling

- Max pooling



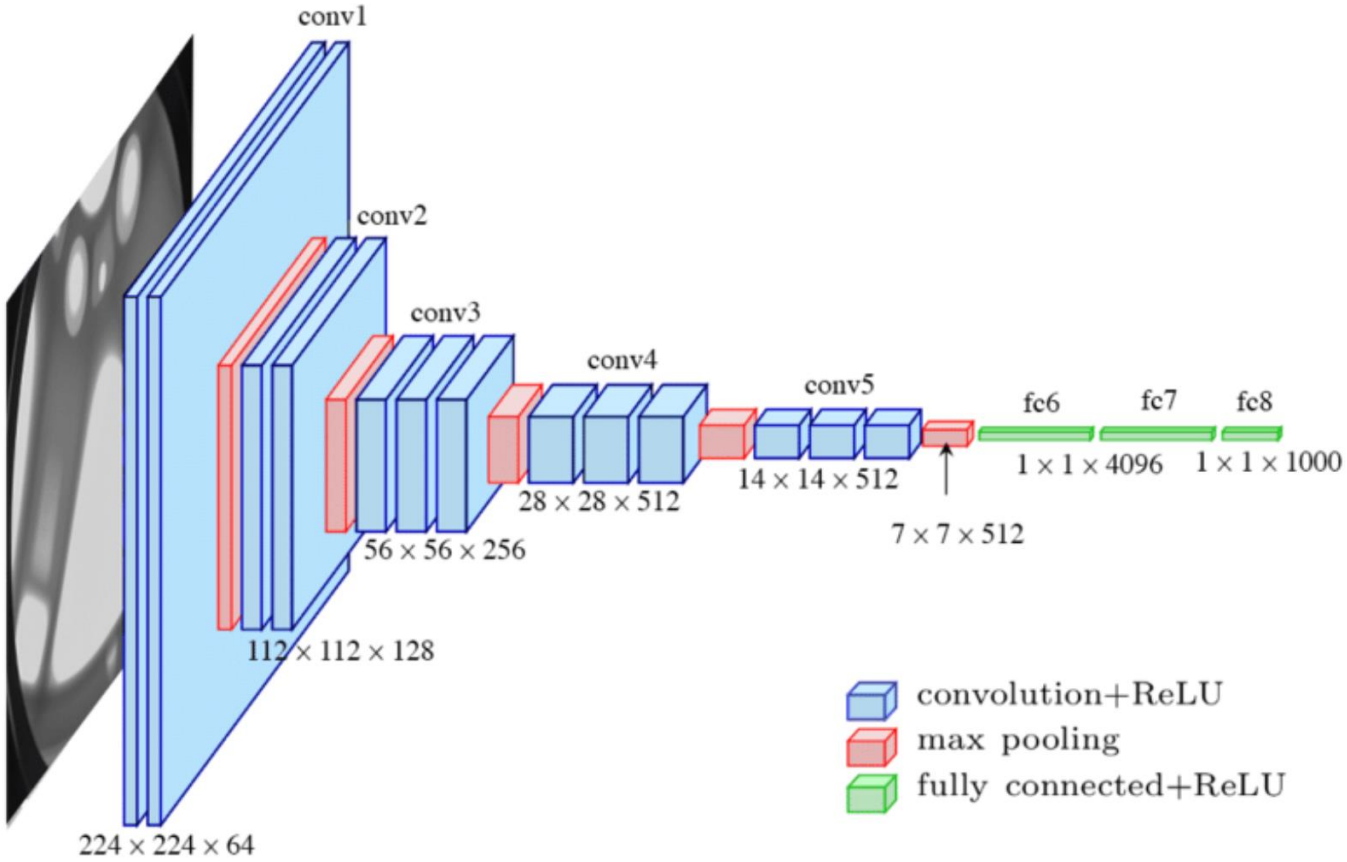
Pooling

- Average pooling



Pooling

- What if we change the input size to 300×600?

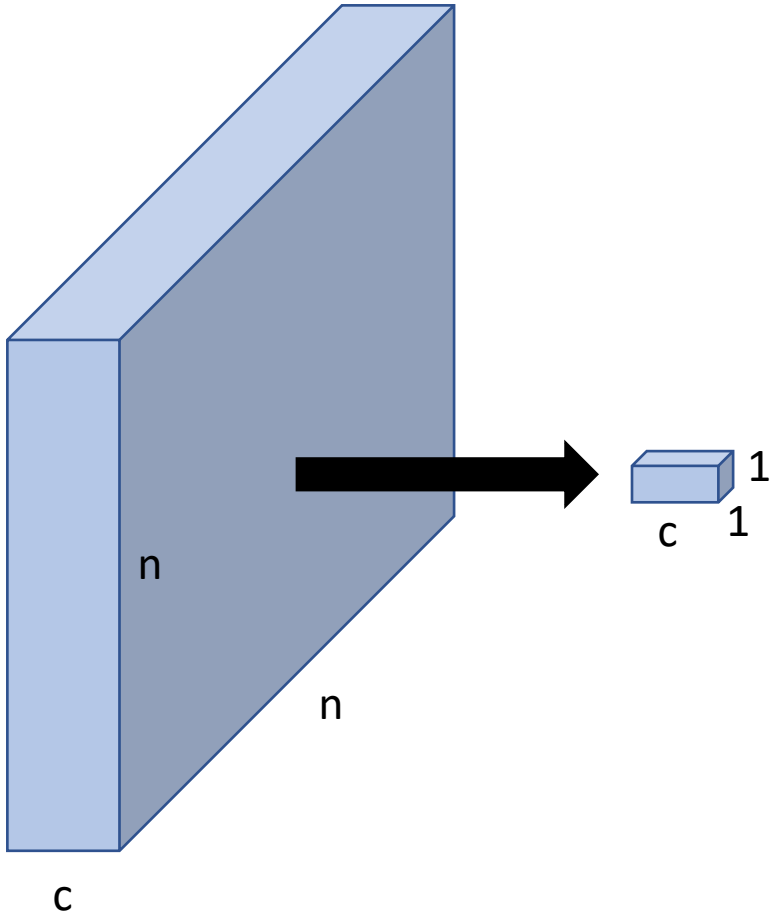


Pooling

- We can always resize the input image to 224×224

Pooling

- Global pooling

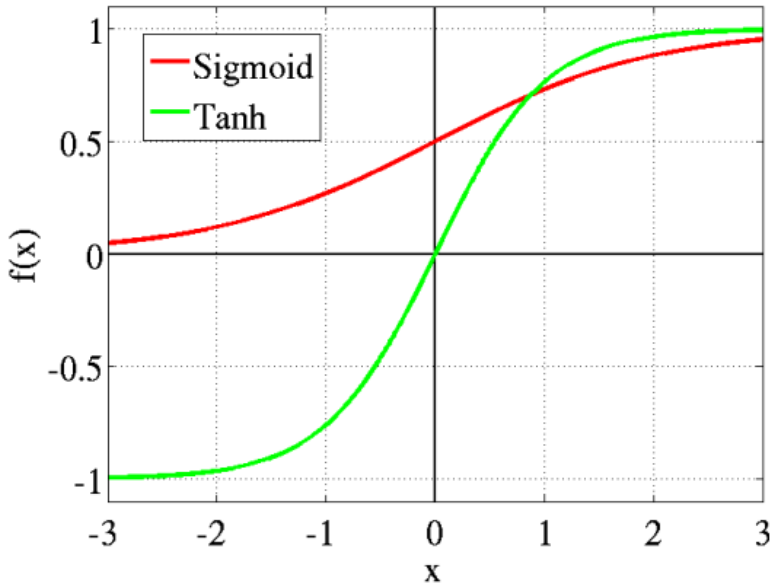


Activation

- Can we use $f(x) = ax + b$ as the activation function?

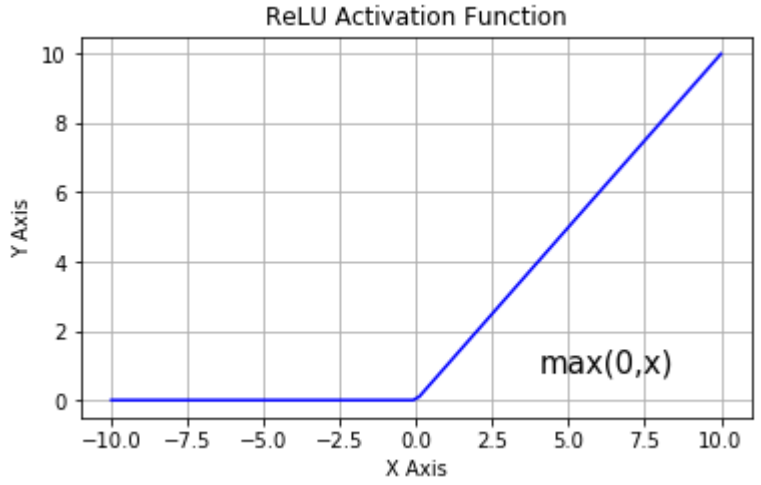
Activation

- Drawbacks of Sigmoid
  - Vanishing gradient

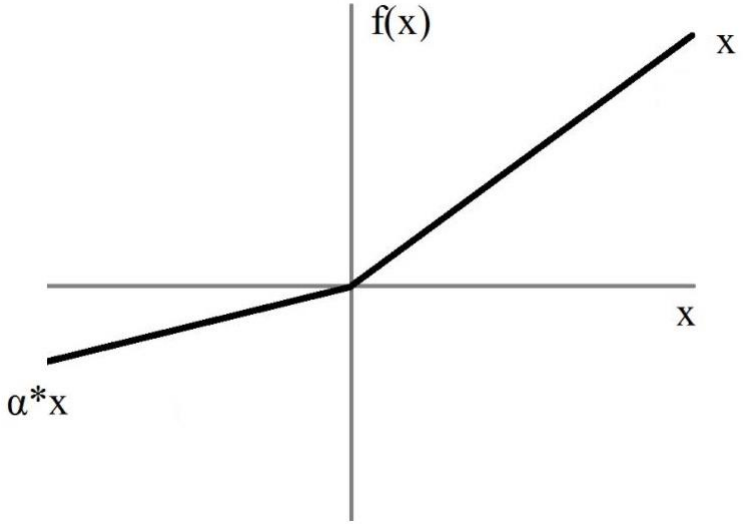


Activation

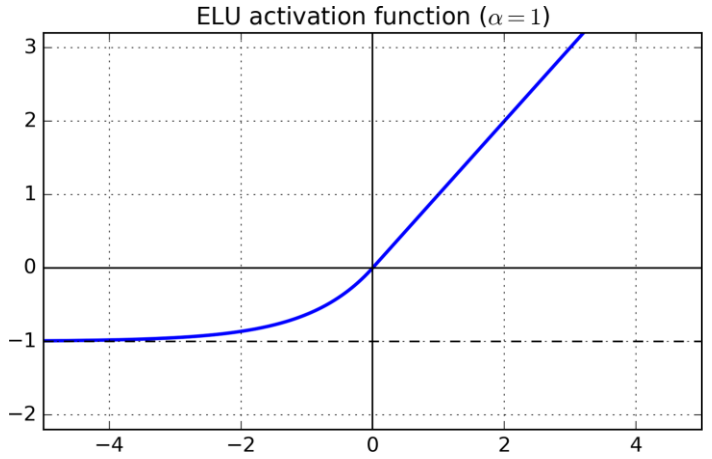
- ReLU family



ReLU



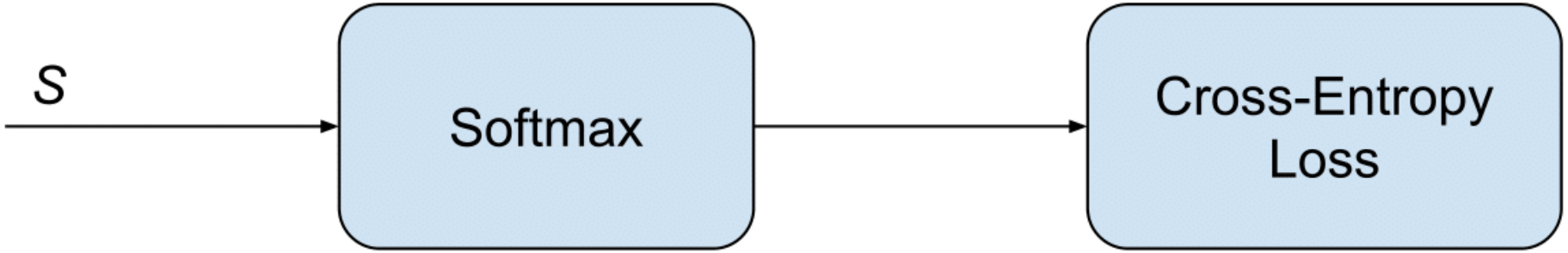
Leaky ReLU



ELU

Loss

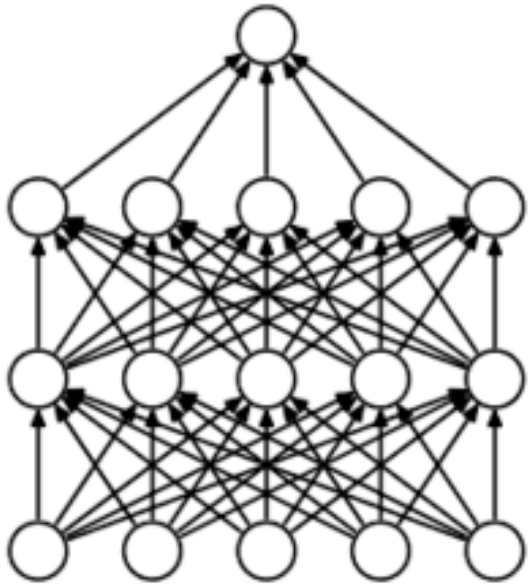
- Softmax cross entropy loss (classification)



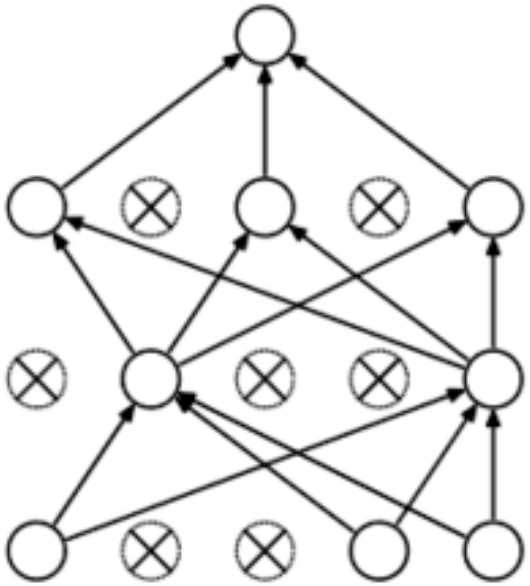
$$f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}} \qquad CE = -\sum_i^C t_i log(f(s)_i)$$

Training

- Dropout
  - Independently set each hidden unit activity to zero with 0.5 probability
  - Address overfitting



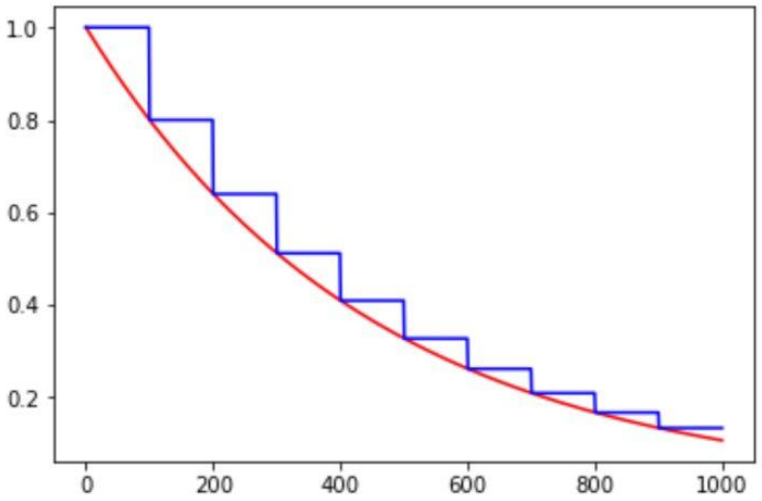
(a) Standard Neural Net



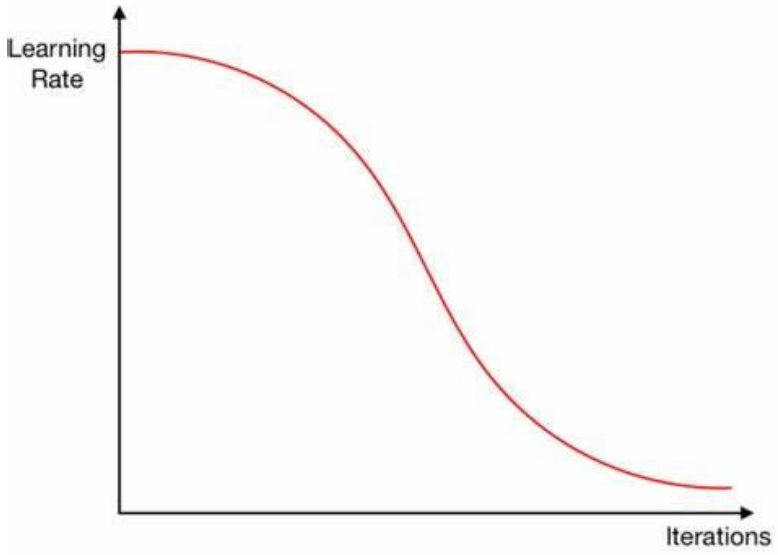
(b) After applying dropout.

Training

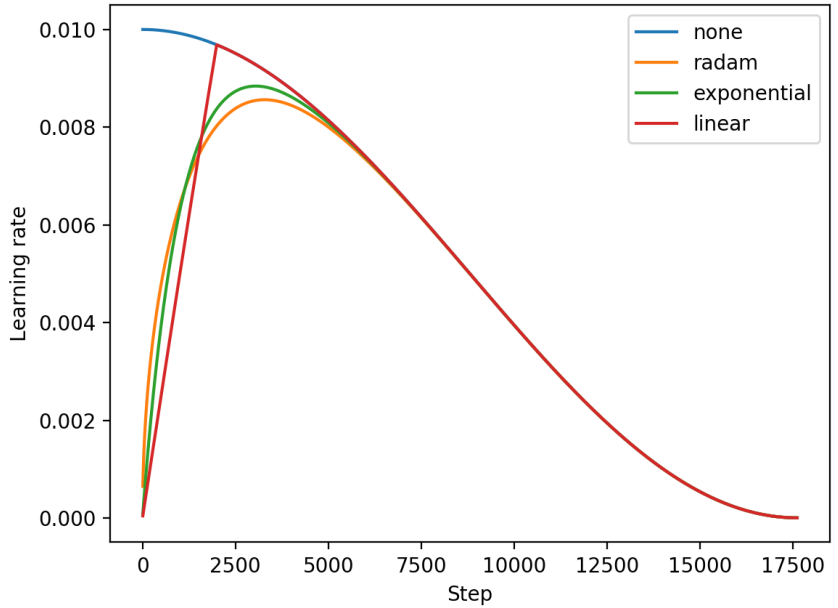
- Learning rate scheduler



Exponential



Cosine



Warm-up

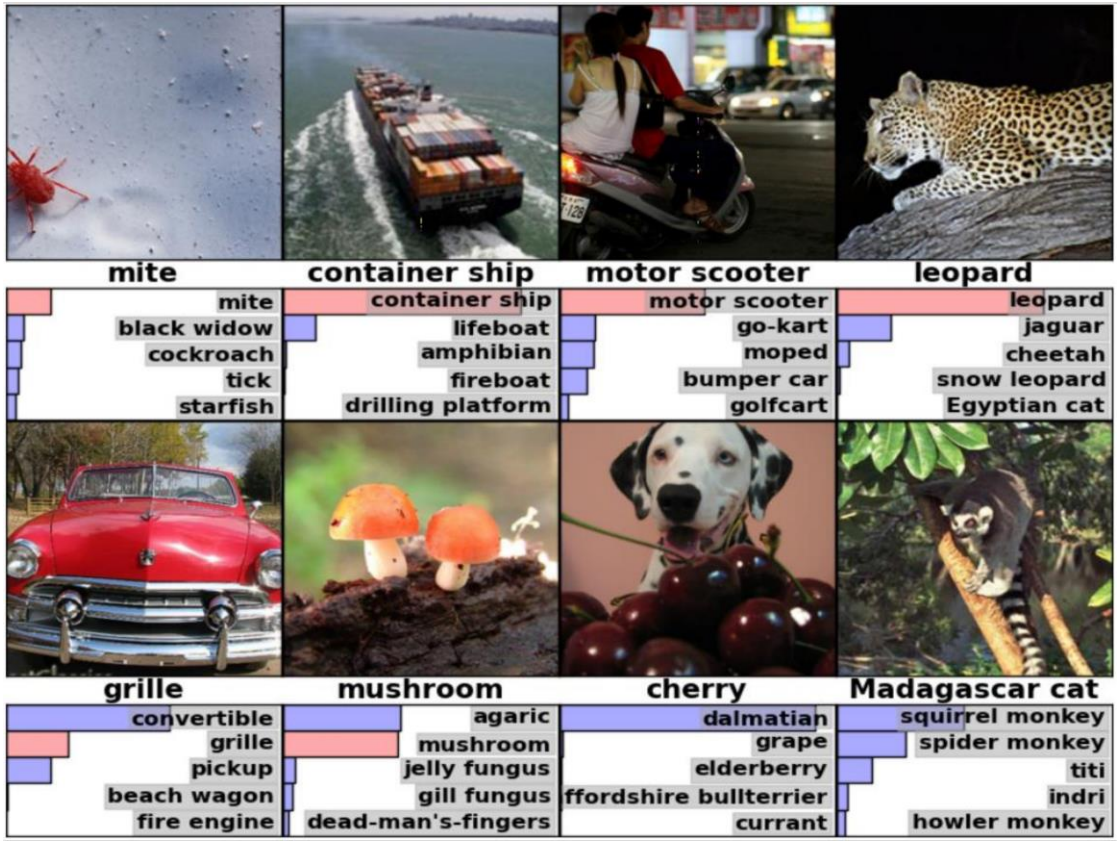
CNN

- From multi-layer perceptron (MLP) to deep neural networks (DNN)
- CNN
- **ImageNet competition**

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

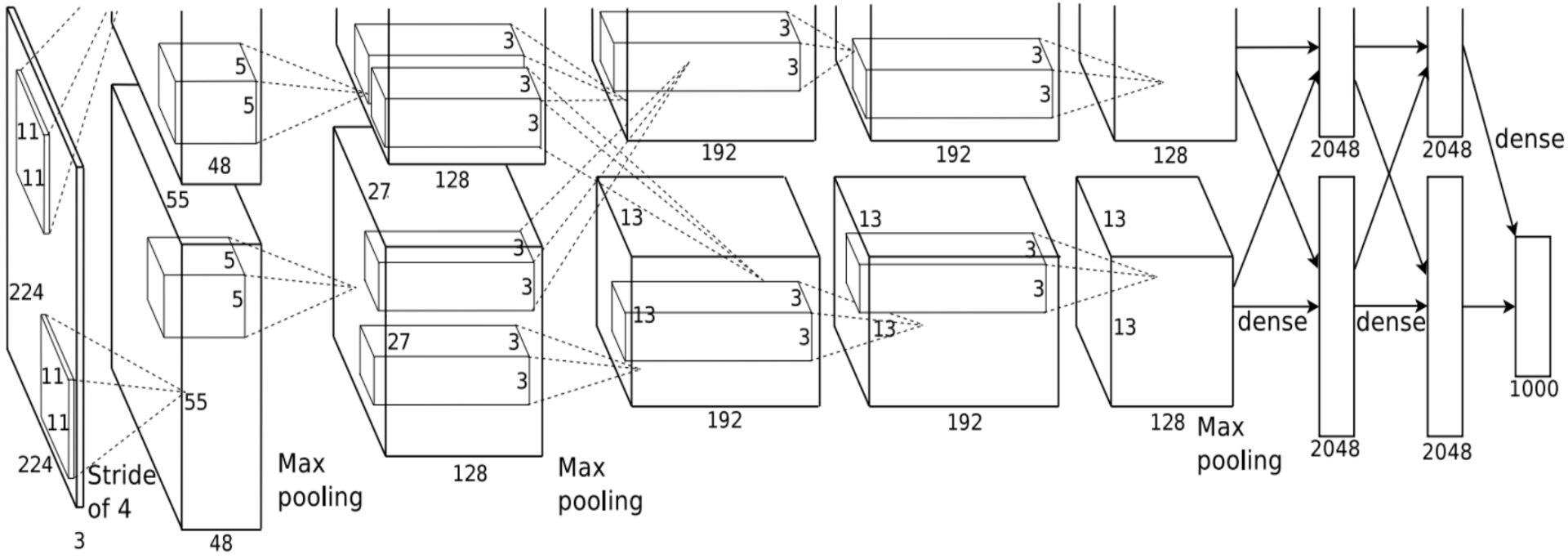
- 1000 categories
- 1.3 million training images
- 50,000 validation images
- 100,000 testing images

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)



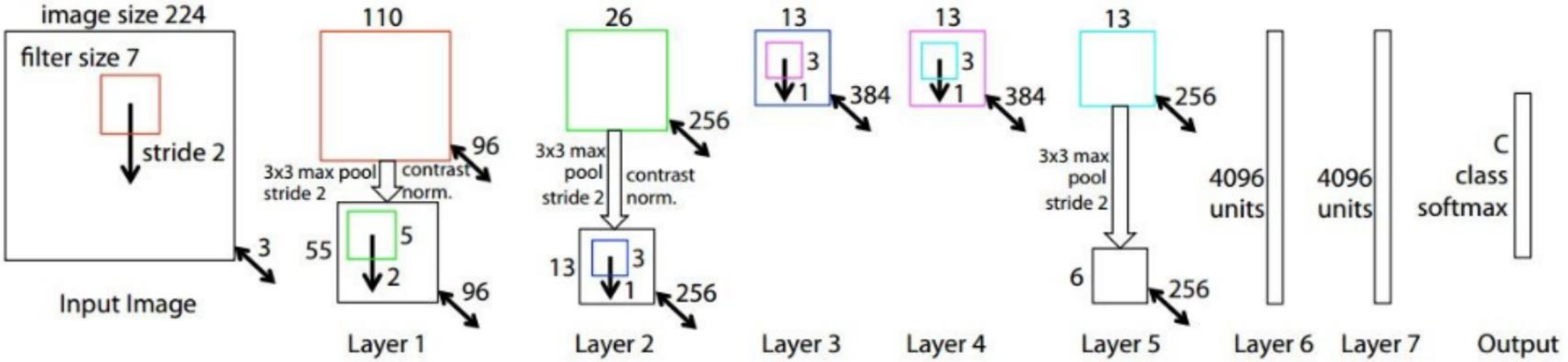
ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012

- AlexNet | Top-5 Error Rate – 15.3%
  - 8 layers where 5 are convolutional layers and 3 fully-connected layers
  - ReLU as the activation function
  - Multi-GPU training



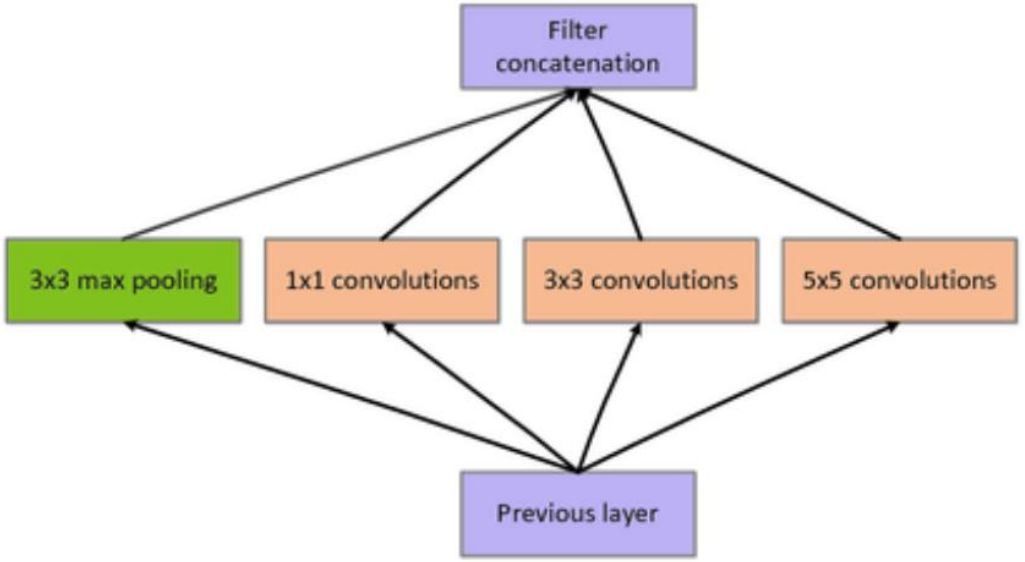
ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2013

- ZFNet | Top-5 Error Rate – 11.2%
  - 7×7 sized filters

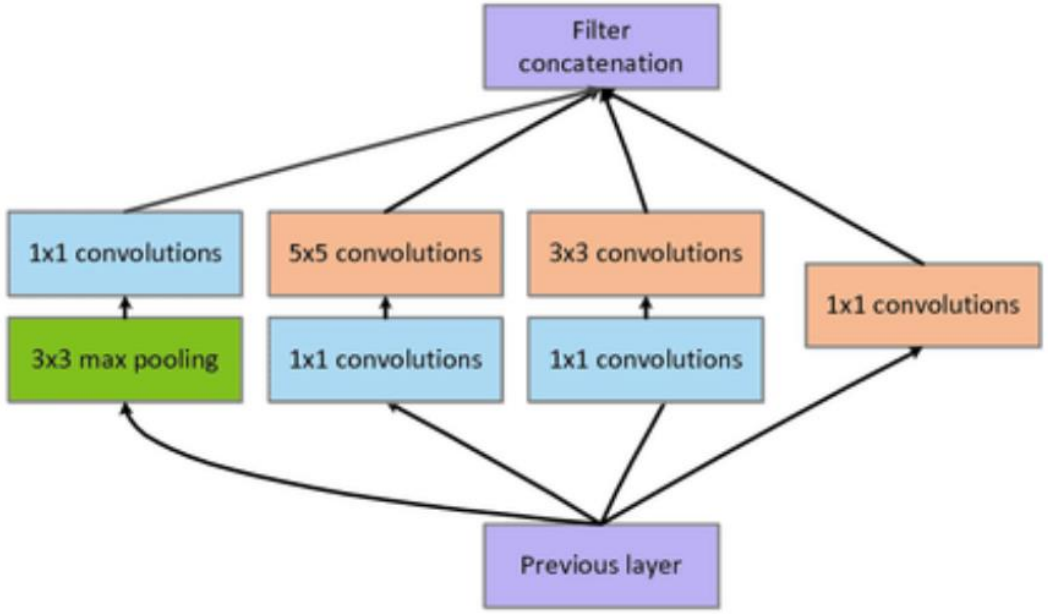


ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014

- Inception V1 (GoogLeNet) | Top-5 Error Rate – 6.67%
  - Inception block



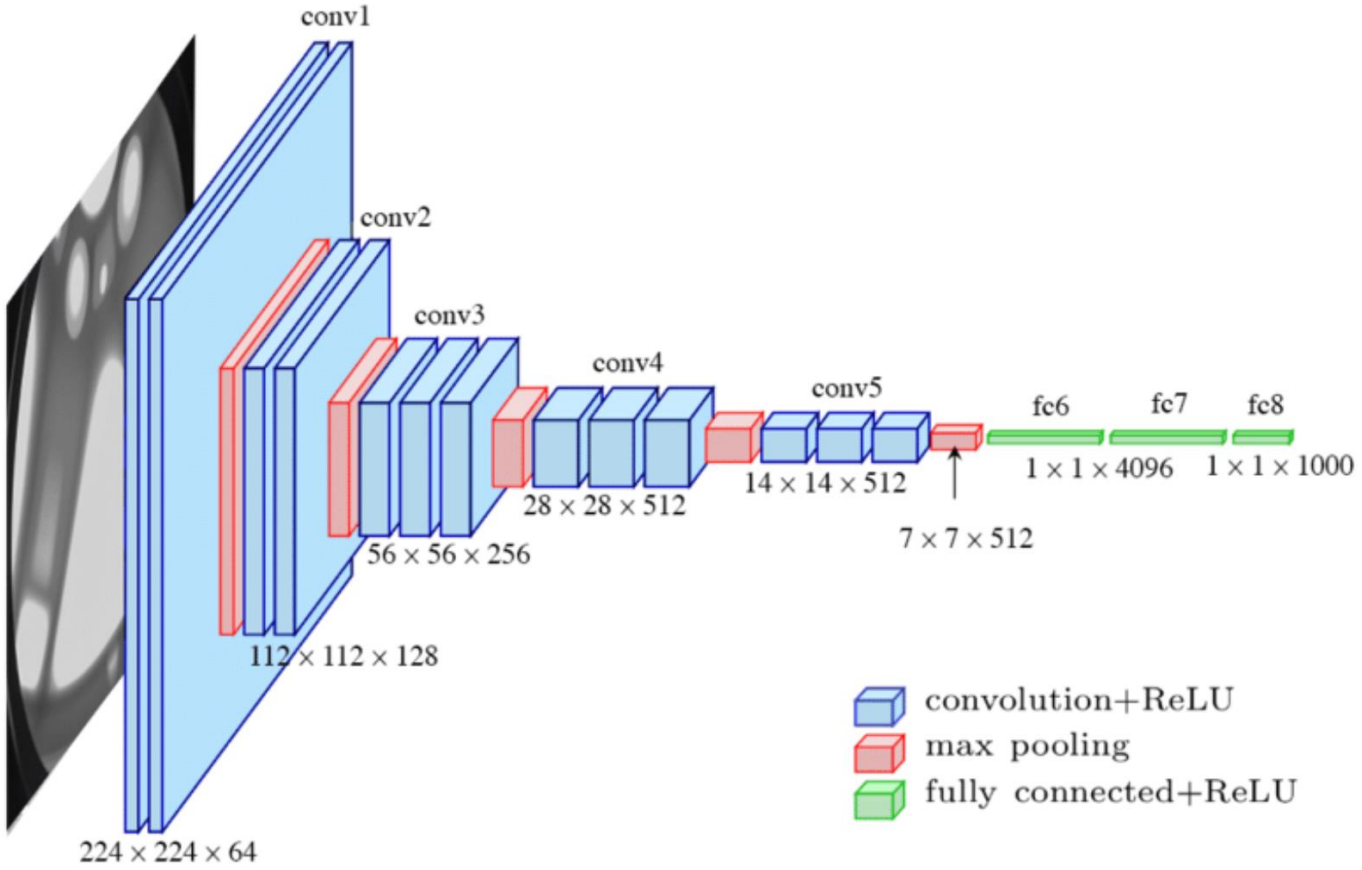
(a) Inception module, initial form



(b) Inception module with dimension reductions

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014

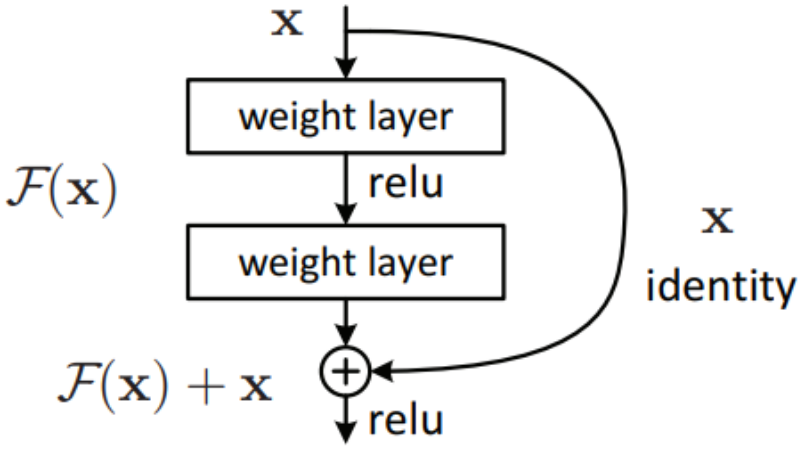
- VGG-16 | Top-5 Error Rate – 7.3%



ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2015

- ResNet | Top-5 Error Rate – 3.57%
  - Skip connection



layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2016

- ResNeXt | Top-5 Error Rate – 4.1%
  - Group Convolution