COMP SCI 1400
AI Technologies – Image Classification

Dr. Kamal Mammadov

Outline

- What & Why
- Deep Neural Network
  - Convolution
  - Activation
  - Max-pooling
  - Full connection

What is image classification?

Image classification is a task to predict the label of a given image from predefined classes or categories:

$$\hat{y} = f(I), I \text{ is the input image.}$$

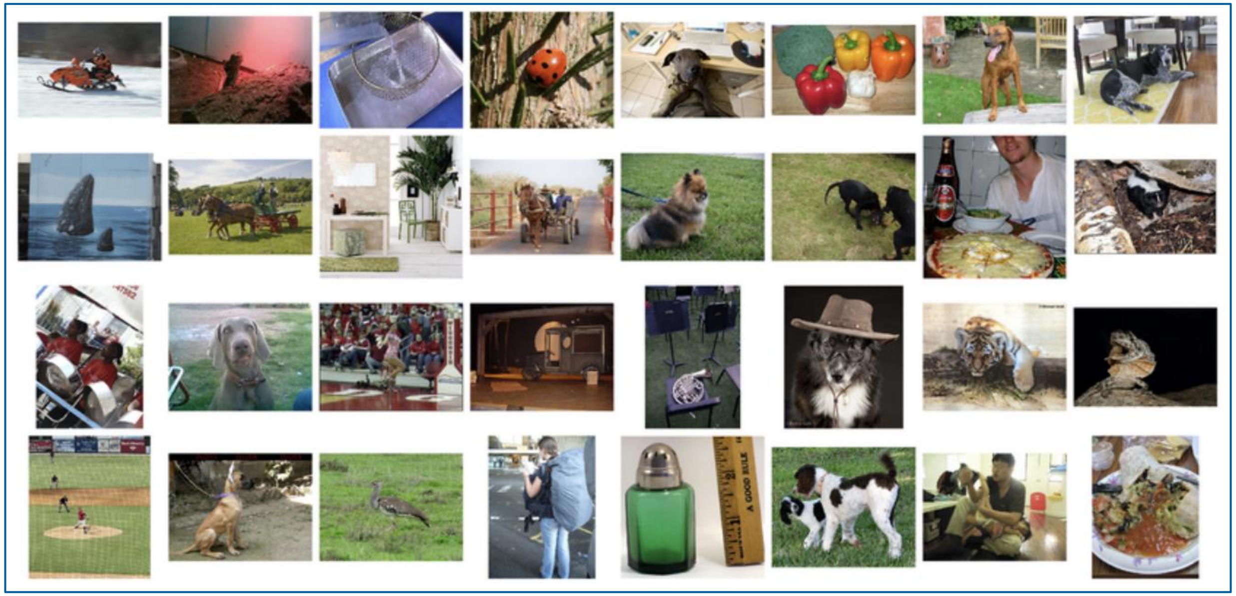
Predefined classes: dog, table, bird, bike, cat, apple, …

Image:		
Prediction $\hat{y}$:	bird	dog
Ground truth y:	bird	dog

ImageNet Large Scale Visual Recognition Challenge

1000 classes, 1.2M training images, 50K validation images, 100K test images

Predict 5 classes, each associated with a bounding box



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

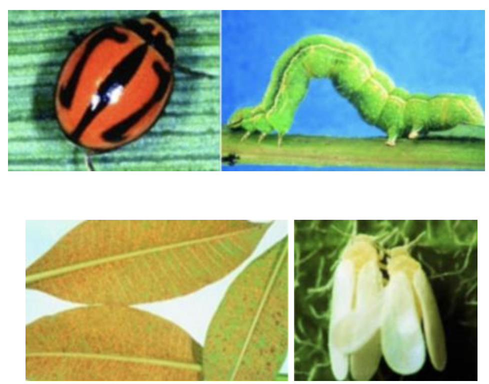
Why learn image classification?

It has wide applications.

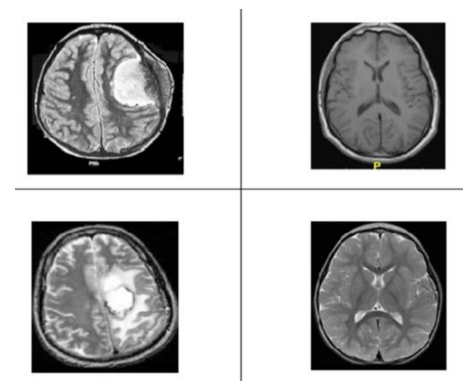
Security

Stranger?

Pest Identification



Medical Image Analysis



Brain Tumor

Brain Non-Tumor

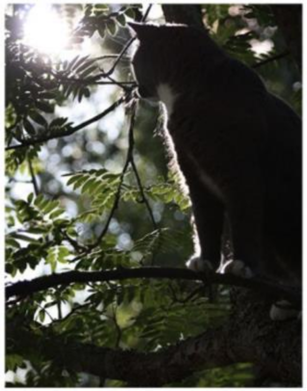
What are the challenges in image classification?

Intense illumination variation

This image is CC0 1.0 public domain

This image is CC0 1.0 public domain

This image is CC0 1.0 public domain



This image is CC0 1.0 public domain

What are the challenges?

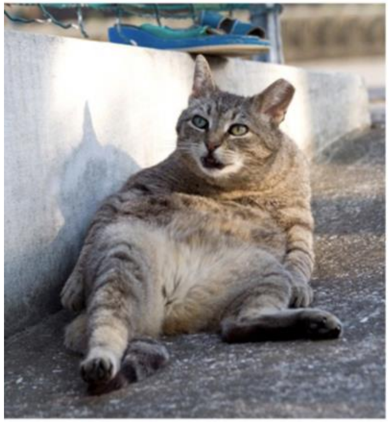
Background clutter

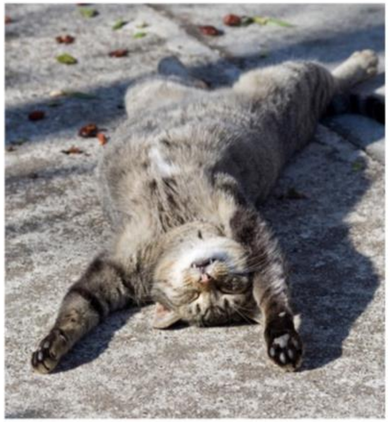
What are the challenges?

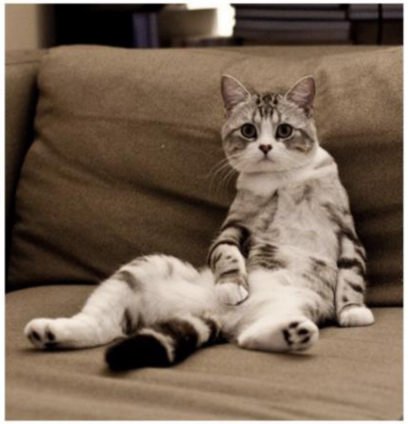
Occlusion

What are the challenges?

Pose / Deformation







Deep Neural Network

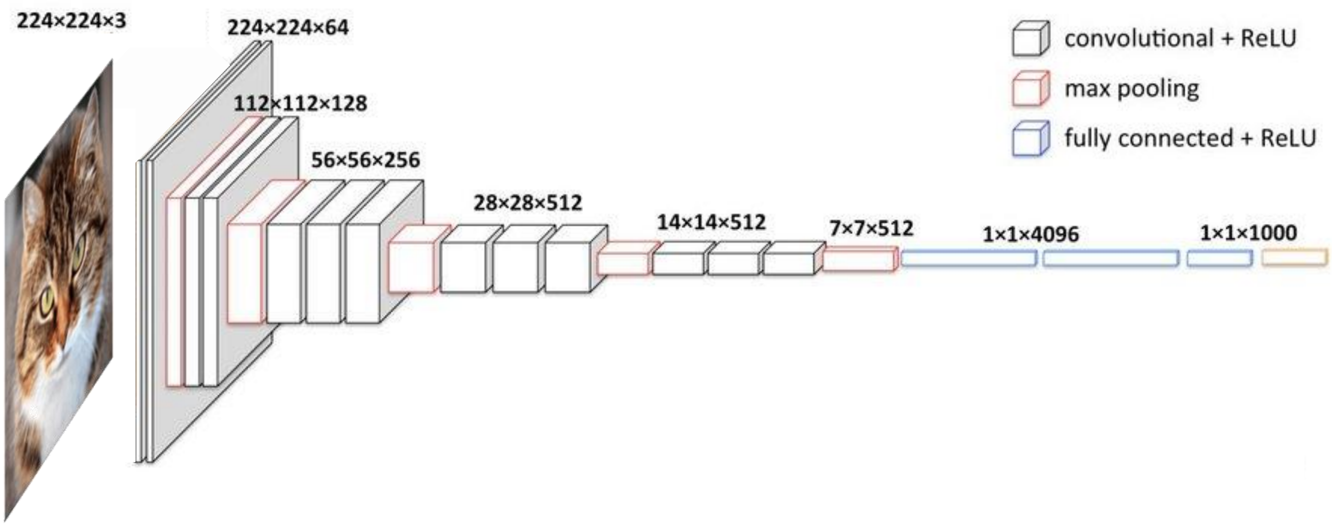
The winner of ILSVRC 2014: VGGNet

in terms of localization error

Localization Error	Classification Error
25.3%	7.4%

Deep Neural Network - VGG

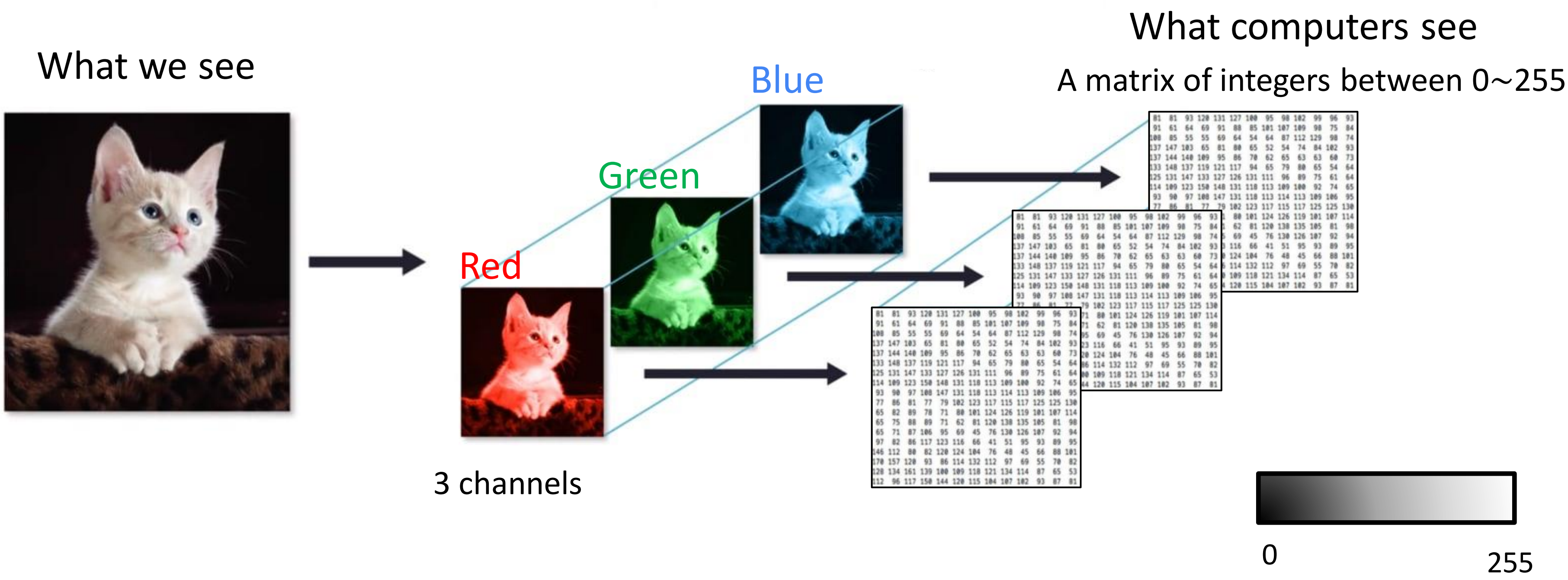
VGGNet architecture



https://arxiv.org/pdf/1409.1556.pdf

Deep Neural Network - VGG

- Image Representation



Deep Neural Network - VGG

- Convolution

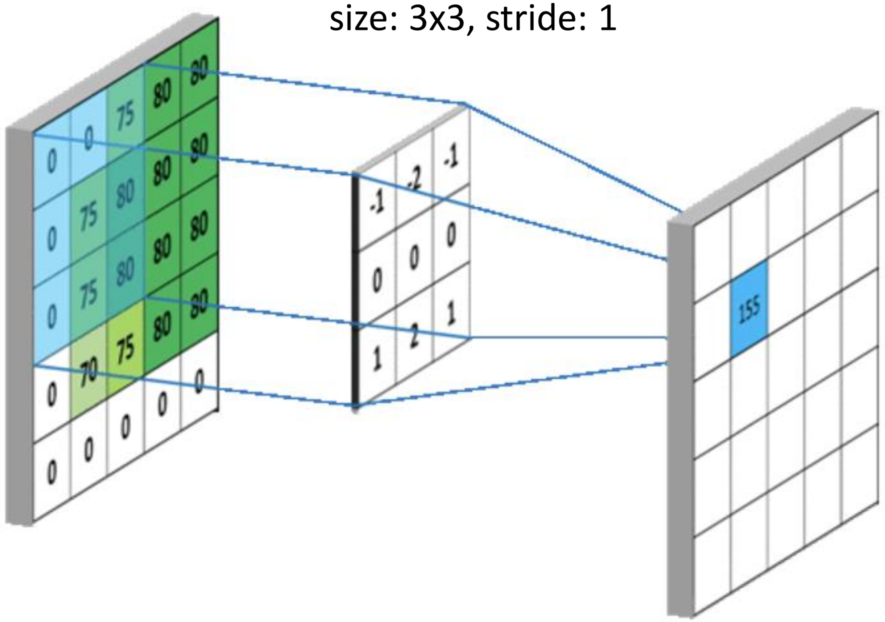
Element-wise multiplication

$$\begin{array}{|c|c|c|}\hline 0 & 0 & 75 \\ \hline 0 & 75 & 80 \\ \hline 0 & 75 & 80 \\ \hline \end{array} * \begin{array}{|c|c|c|}\hline -1 & -2 & -1 \\ \hline 0 & 0 & 0 \\ \hline 1 & 2 & 1 \\ \hline \end{array}$$

$$155 = 0 \times (-1) + 0 \times (-2) + 75 \times (-1) +$$
$$0 \times 0 + 75 \times 0 + 80 \times 0 +$$

Input (one channel) kernel Output

size: 3x3, stride: 1



Deep Neural Network - VGG

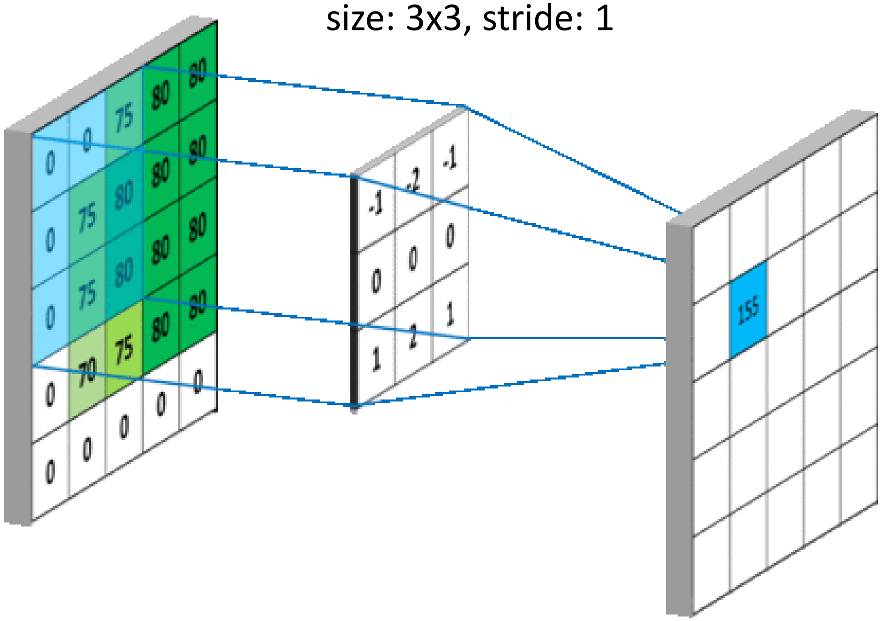
- Convolution

Input (one channel)

kernel

size: 3x3, stride: 1

Output



Element-wise multiplication

0	0	75			-1	-2	-1
0	75	80		*	0	0	0
0	75	80			1	2	1

$$155 = 0 \times (-1) + 0 \times (-2) + 75 \times (-1) + \\ 0 \times 0 + 75 \times 0 + 80 \times 0 + \\ 0 \times 1 + 75 \times 2 + 80 \times 1$$

Output size:

$$1 + floor(\text{ (input\_size} - \text{kernel\_size)/stride })$$

$$floor(\, 3.8 \,) = 3$$

Change the kernel size and stride to get different output size

Deep Neural Network - VGG

- Convolution

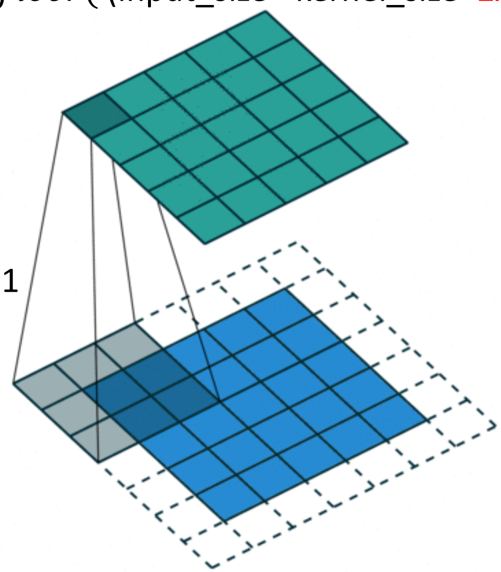
What if we want to keep the output size the same as the input?

1+$floor$((input_size−kernel_size+2xpadding_size)/stride)

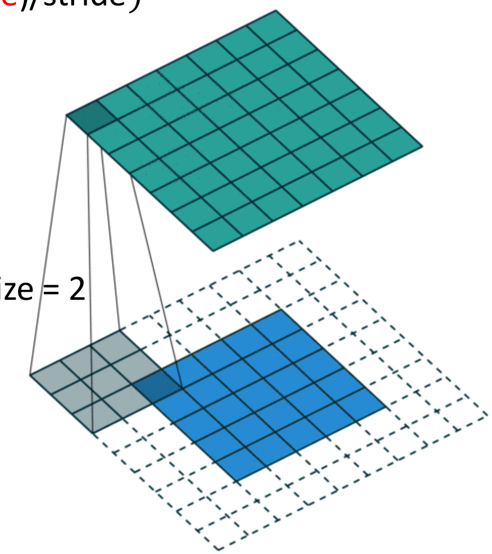
Padding

padding value is usually 0

Padding size = 1

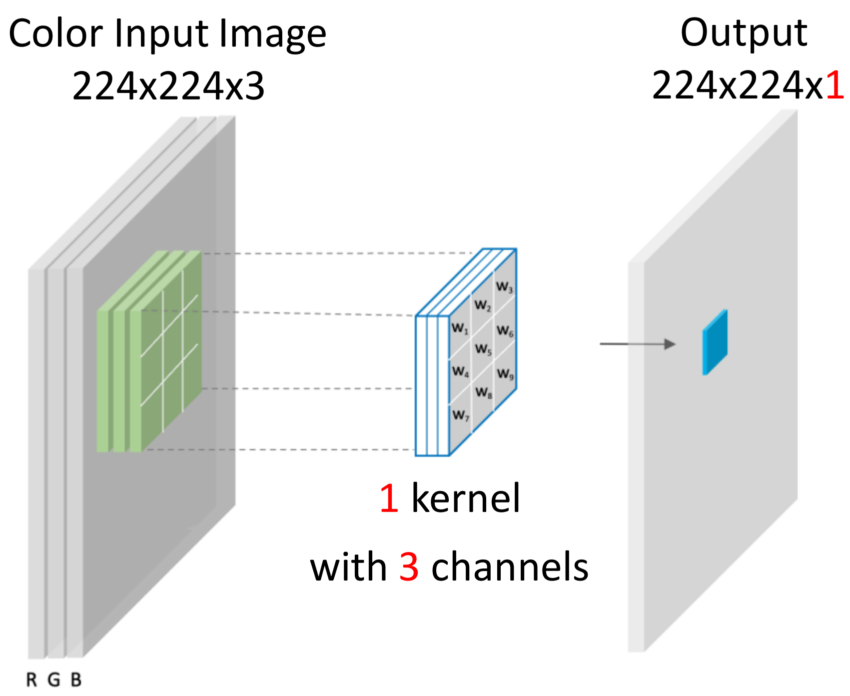


Padding size = 2



Deep Neural Network - VGG

- Convolution



One kernel produces 1 channel

If we hope the output has N channels,
we need to use N kernels

Deep Neural Network - VGG

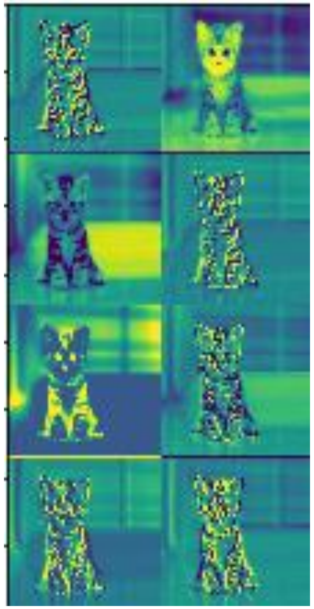
- What can convolution learn?

Input image

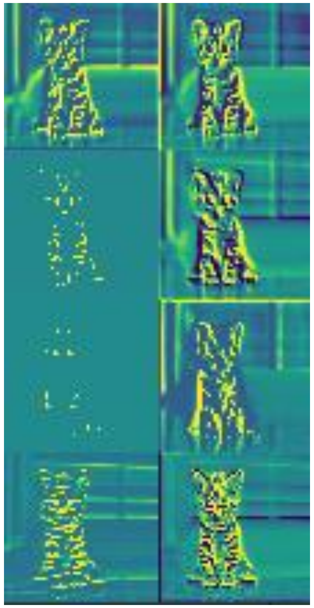
Deep Neural Network - VGG

- What can convolution learn?

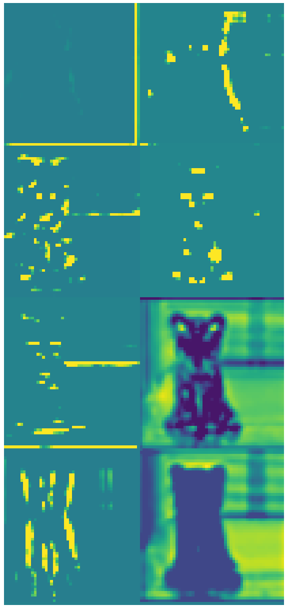
Conv1



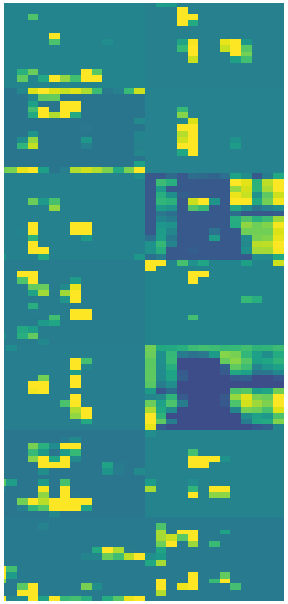
Conv2



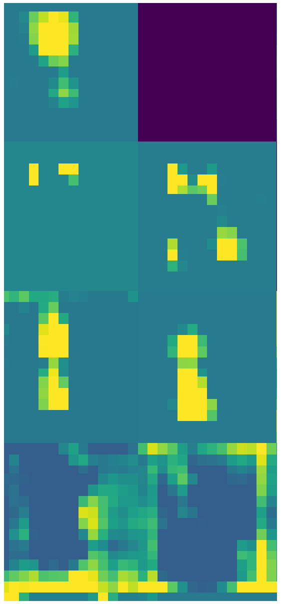
Conv3



Conv4

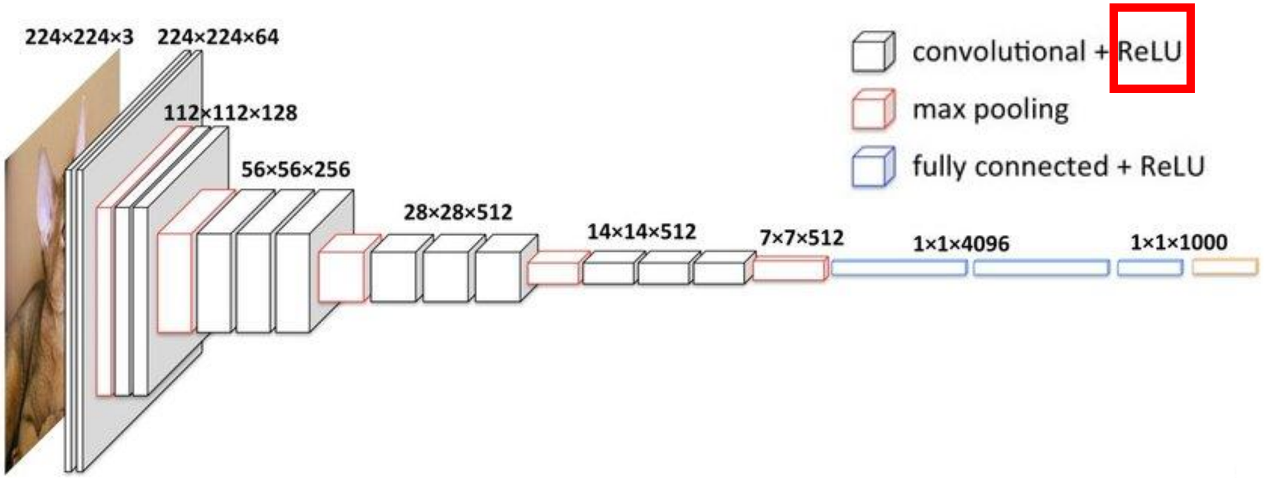


Conv5



Deep Neural Network - VGG

What is “ReLU”?

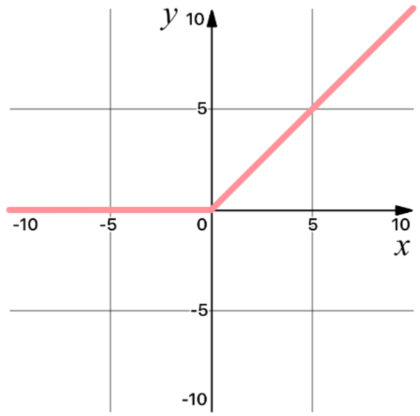


Deep Neural Network - VGG

- Activation function

ReLU: Rectified Linear activation Unit

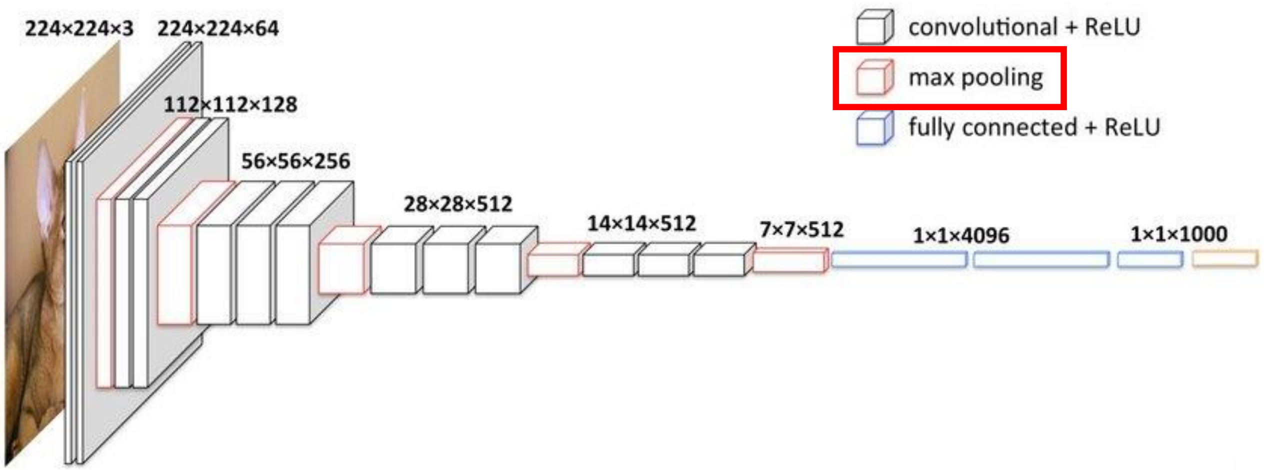
$$\mathrm{ReLU}(x) = \max(0, x)$$



Activation function introduces non-linear mapping, enhancing the capacity to learn complex data patterns and relationships

Deep Neural Network - VGG

- Max pooling



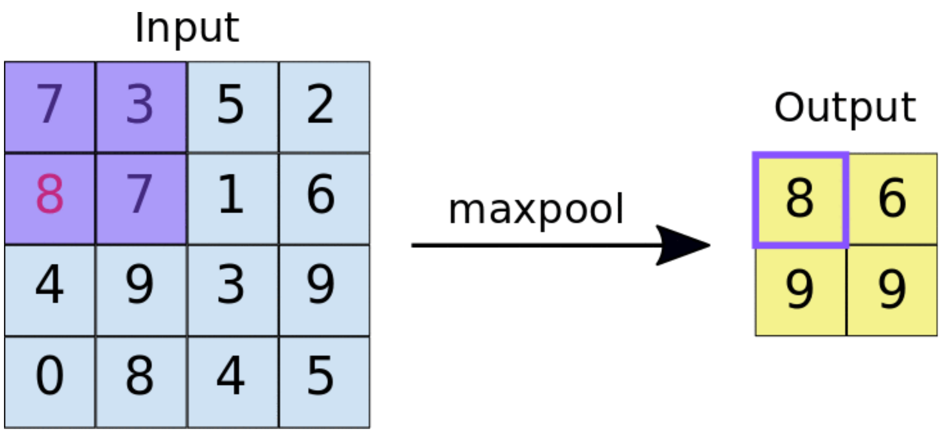
Deep Neural Network - VGG

- Max pooling

Two hype-parameters: size, stride

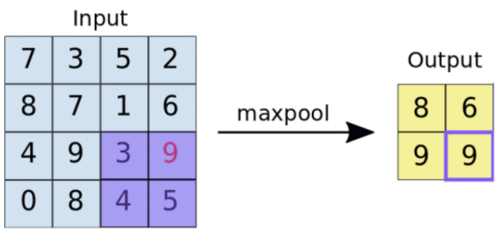
Output size: $1+floor((\text{input\_size}-\text{kernel\_size}+2\text{xpadding\_size})/\text{stride})$

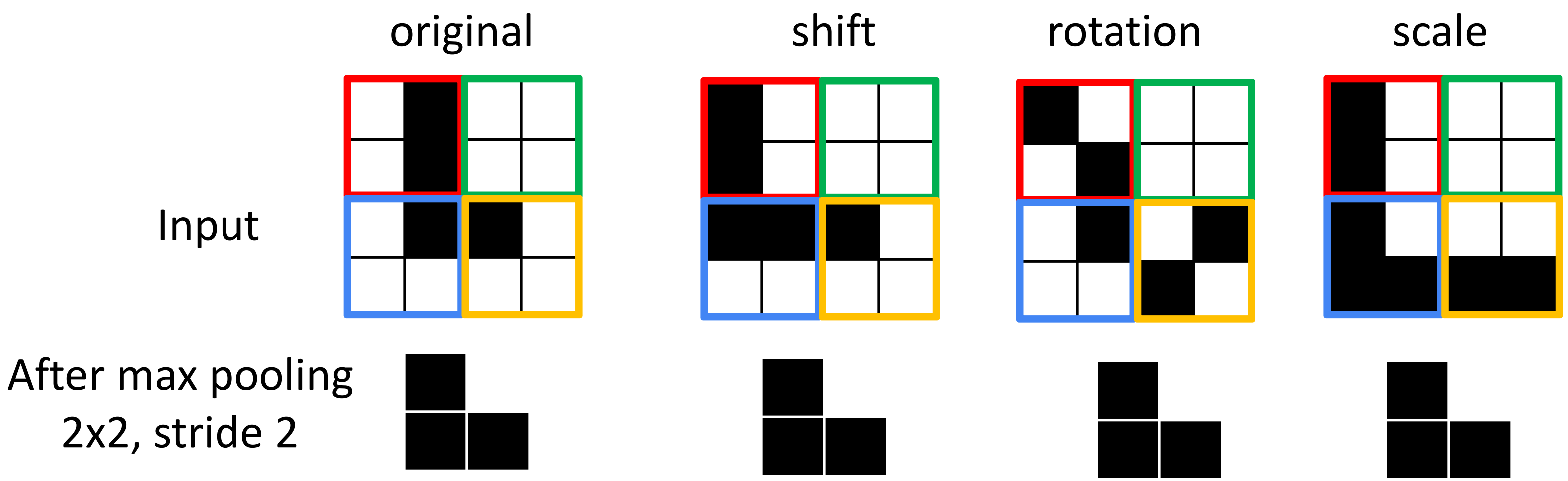
size: 2x2
stride: 2



Deep Neural Network - VGG

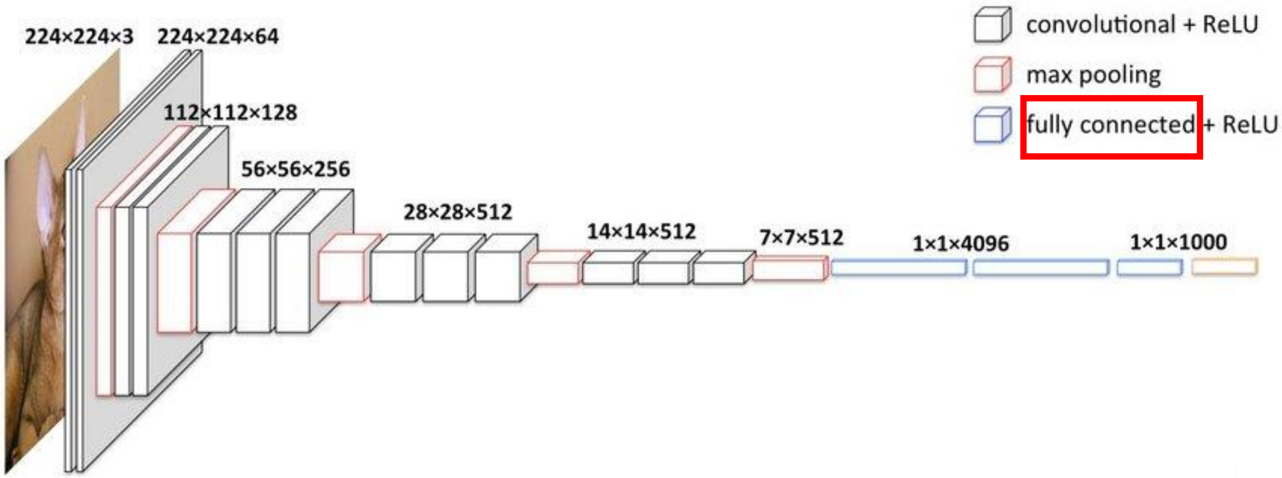
- Why do we use max pooling?
  - Downscaling input, reduce computation later layers
  - Introduce invariance to shift, rotation and scale





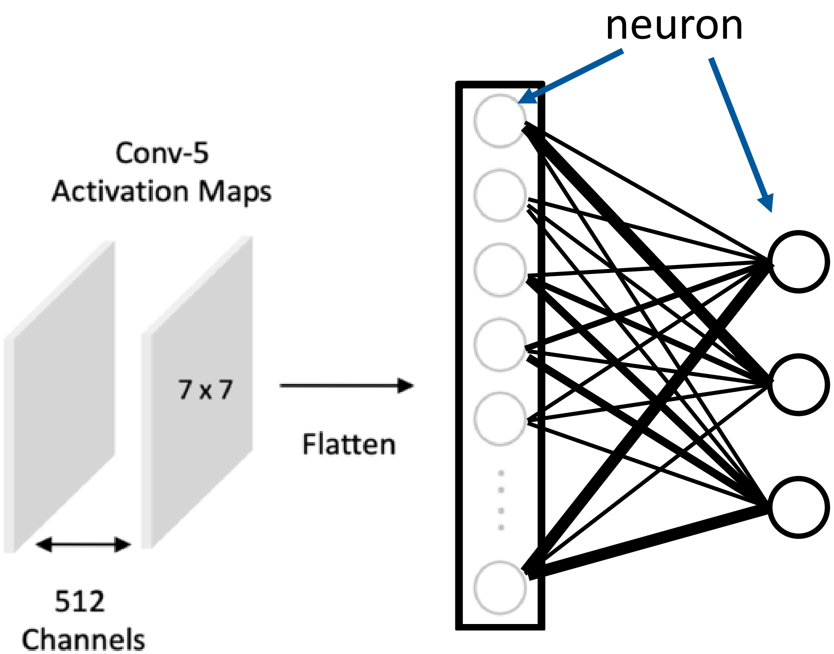
Deep Neural Network - VGG

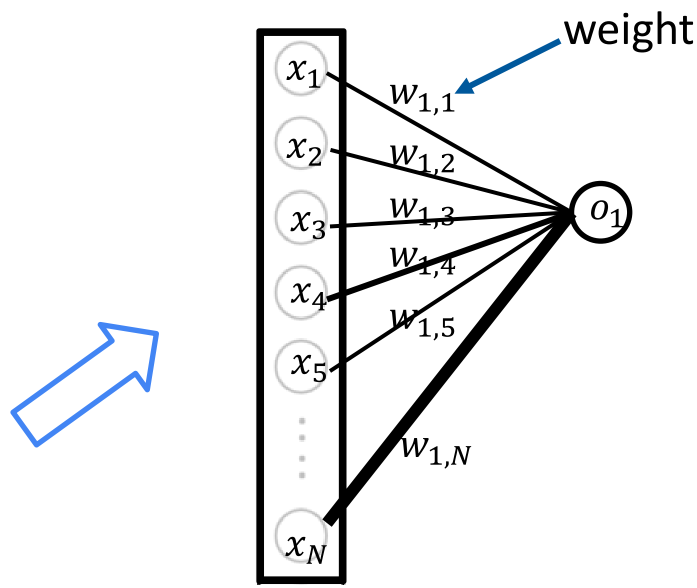
- Fully connected layer



Deep Neural Network - VGG

- Fully connected layer



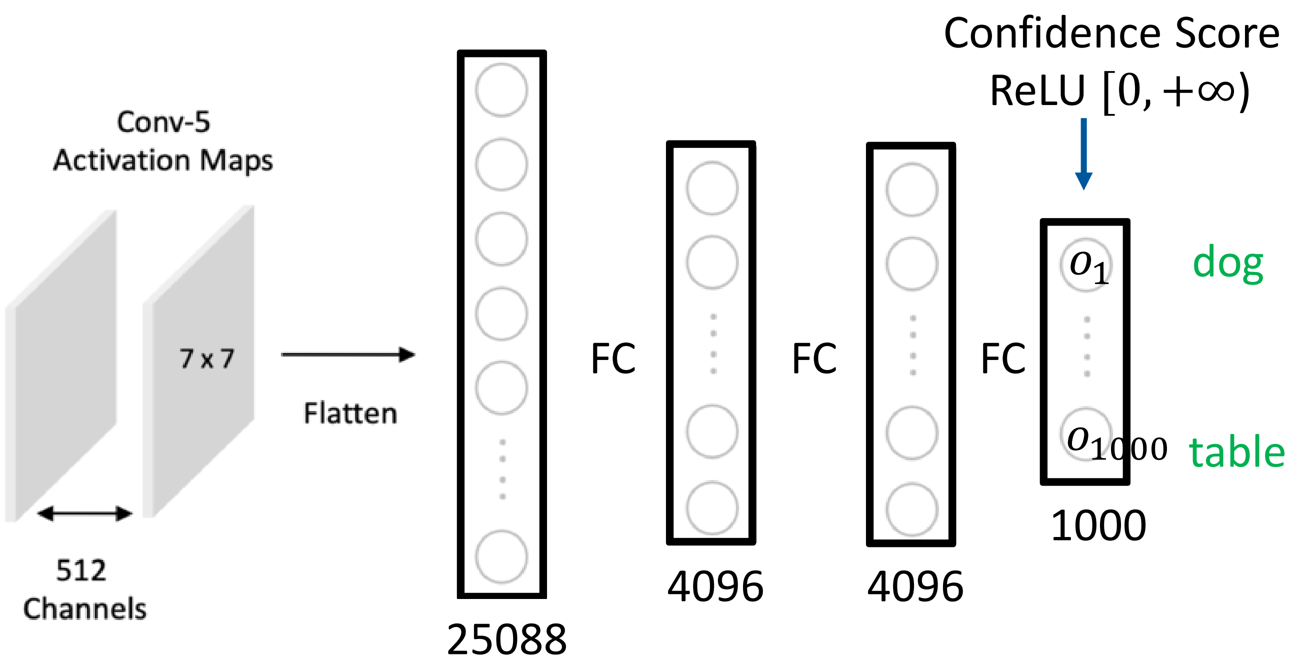


$$o_1 = \sigma(w_{1,1}x_1 + w_{1,2}x_2 + \cdots + w_{1,N}x_N + b_1)$$

$\sigma(\cdot)$ is an activation function, eg, ReLU

Deep Neural Network - VGG

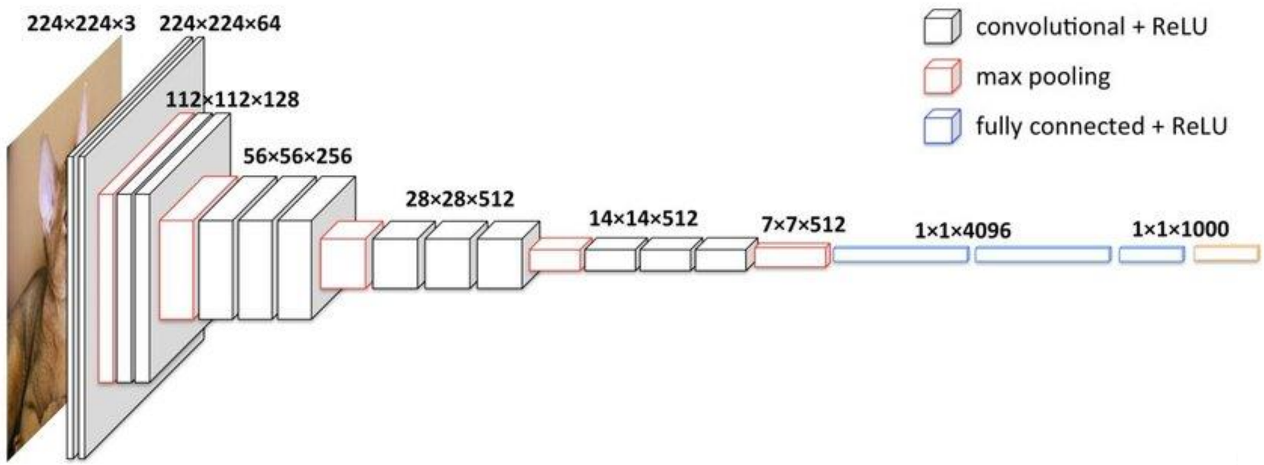
- Fully connected layer (FC) in VGG



$$p_i = \text{Softmax}(o_i) = \frac{e^{o_i}}{\sum_j e^{o_j}}$$

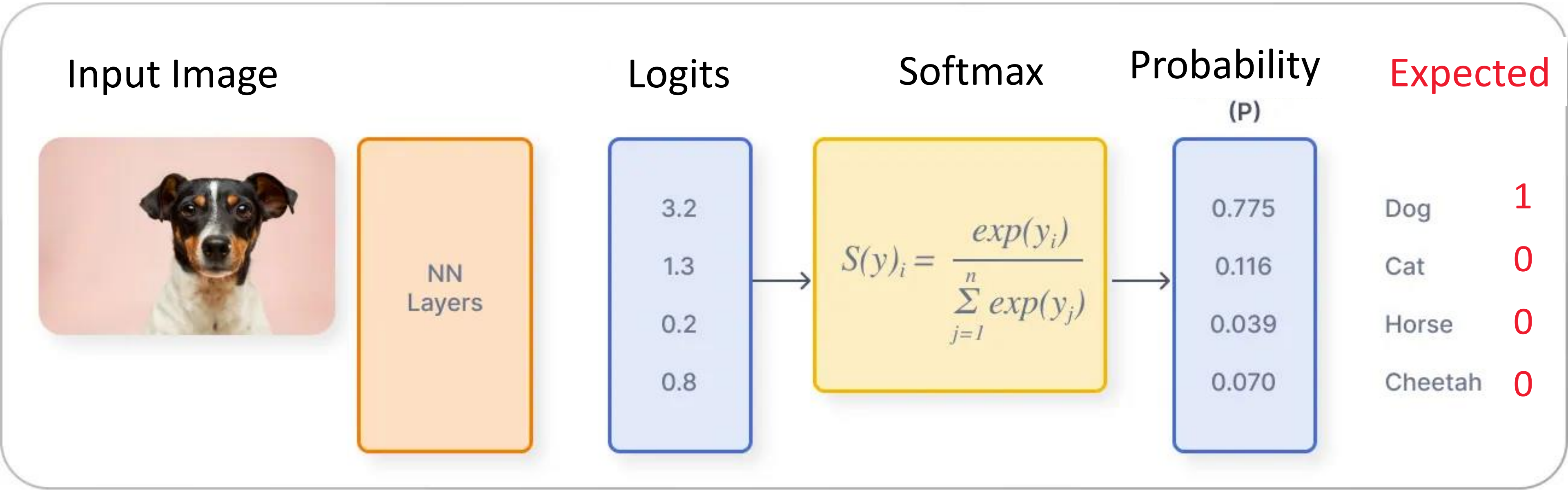
Deep Neural Network - VGG

- Build your own neural network



Deep Neural Network - VGG

- How to make the model predict expected values?



Cross-entropy loss: $-\log(p_i)$ $\quad\quad -\log(0.775) = 0.255$

i denotes the GT class $\quad\quad -\log(0.001) = 6.908$

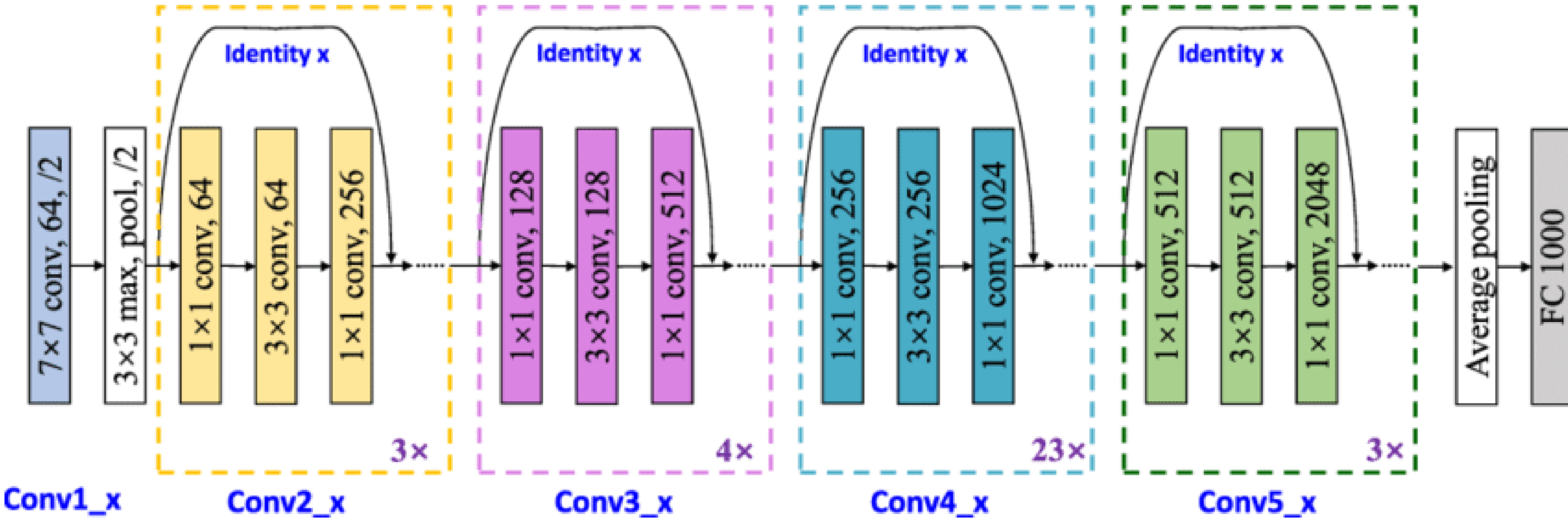
Use Stochastic Gradient Descent to minimize loss

ResNet

Winner of ILSVRC 2015

The 1st time outperforms humans

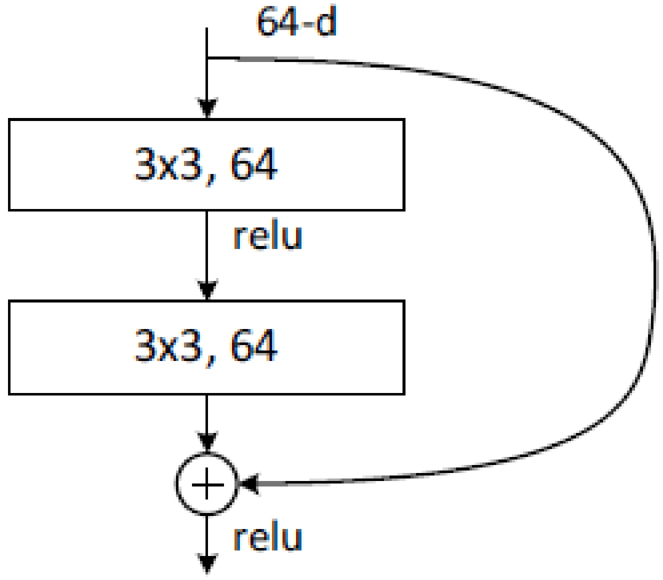
Deep Neural Network - ResNet

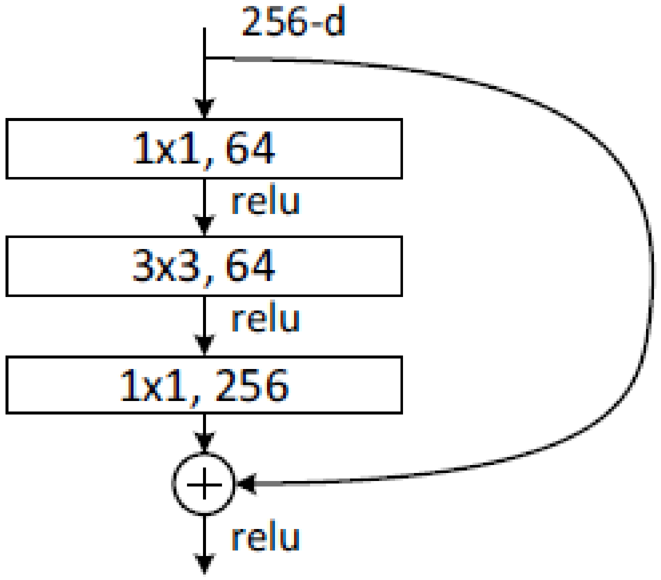


https://arxiv.org/pdf/1512.03385.pdf

Deep Neural Network - ResNet

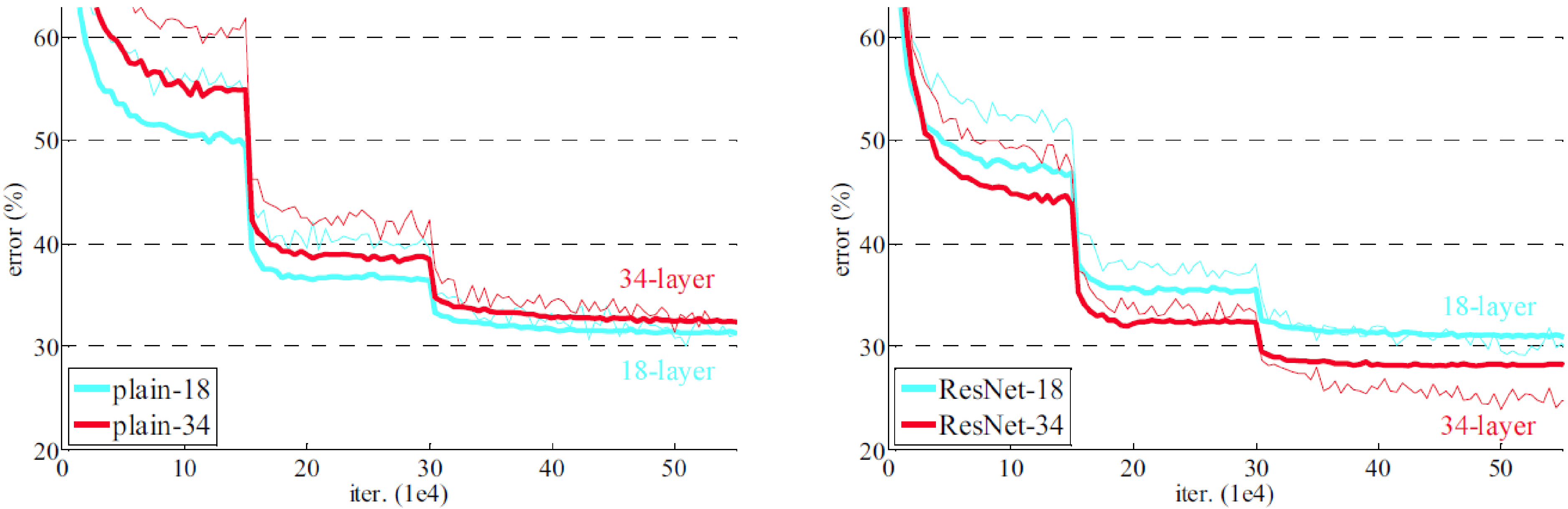
Issue of plain connection: gradients vanish as network becomes deeper





Deep Neural Network - ResNet

Effectiveness



Thin curves denote training error, and bold curves denote validation error.

Deep Neural Network - ResNet

Effectiveness

method	top-1 err.	top-5 err.
VGG [40] (ILSVRC'14)	-	8.43†
GoogLeNet [43] (ILSVRC'14)	-	7.89
VGG [40] (v5)	24.4	7.1
PReLU-net [12]	21.59	5.71
BN-inception [16]	21.99	5.81
ResNet-34 B	21.84	5.71
ResNet-34 C	21.53	5.60
ResNet-50	20.74	5.25
ResNet-101	19.87	4.60
ResNet-152	**19.38**	**4.49**

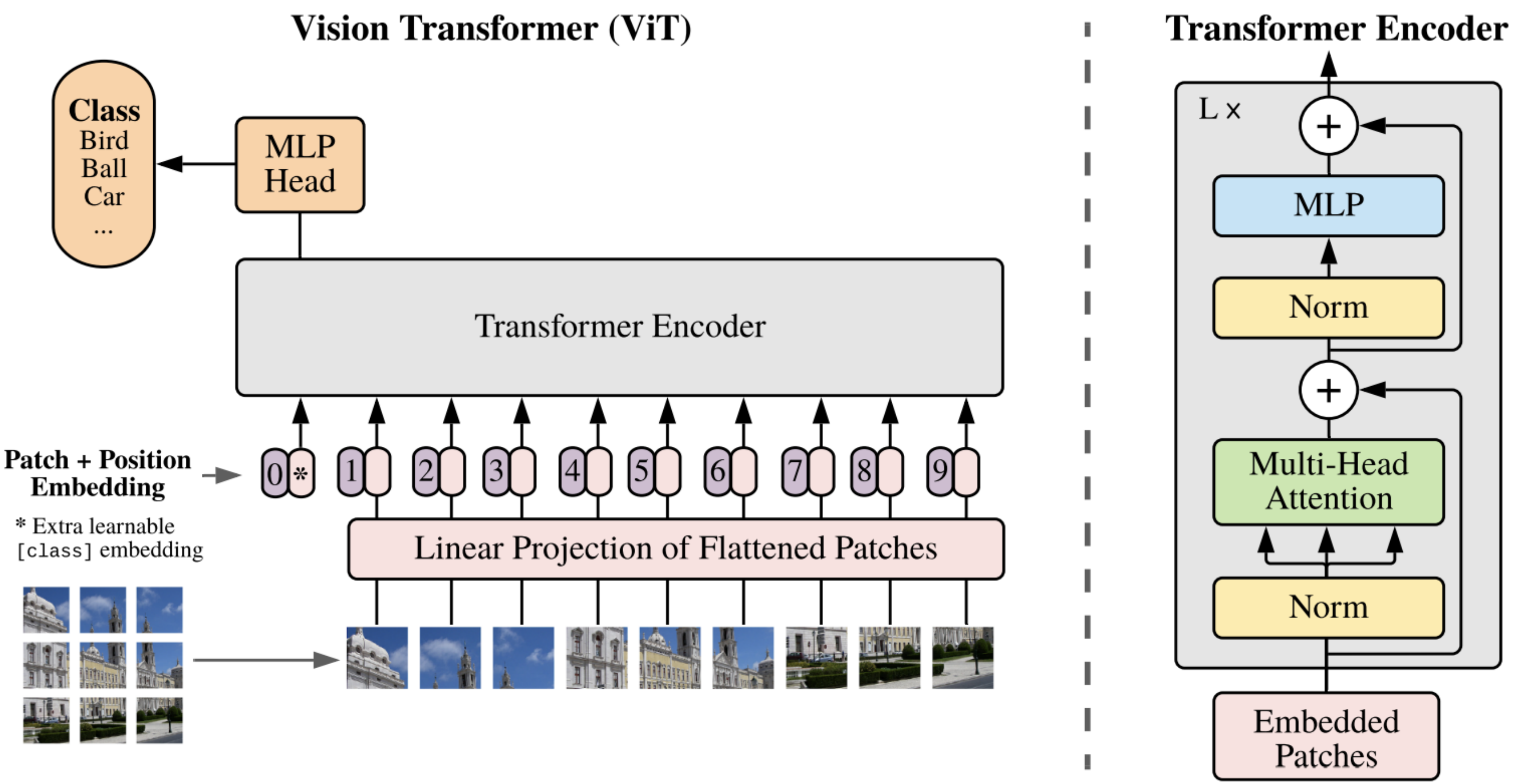
Single Model

method	top-5 err. (**test**)
VGG [40] (ILSVRC'14)	7.32
GoogLeNet [43] (ILSVRC'14)	6.66
VGG [40] (v5)	6.8
PReLU-net [12]	4.94
BN-inception [16]	4.82
ResNet (ILSVRC'15)	**3.57**

Ensemble

Vision Transformer (ViT)

Deep Neural Network - ViT



https://arxiv.org/pdf/2010.11929.pdf

Deep Neural Network - ViT

Performance

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)
ImageNet	**88.55** $\pm$ 0.04	87.76 $\pm$ 0.03	85.30 $\pm$ 0.02	87.54 $\pm$ 0.02
ImageNet ReaL	**90.72** $\pm$ 0.05	90.54 $\pm$ 0.03	88.62 $\pm$ 0.05	90.54
CIFAR-10	**99.50** $\pm$ 0.06	99.42 $\pm$ 0.03	99.15 $\pm$ 0.03	99.37 $\pm$ 0.06
CIFAR-100	**94.55** $\pm$ 0.04	93.90 $\pm$ 0.05	93.25 $\pm$ 0.05	93.51 $\pm$ 0.08
Oxford-IIIT Pets	**97.56** $\pm$ 0.03	97.32 $\pm$ 0.11	94.67 $\pm$ 0.15	96.62 $\pm$ 0.23
Oxford Flowers-102	99.68 $\pm$ 0.02	**99.74** $\pm$ 0.00	99.61 $\pm$ 0.02	99.63 $\pm$ 0.03
VTAB (19 tasks)	**77.63** $\pm$ 0.23	76.28 $\pm$ 0.46	72.72 $\pm$ 0.21	76.29 $\pm$ 1.70
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k

https://arxiv.org/pdf/2010.11929.pdf

Deep Neural Network - ViT

Pros and Cons

Pros	Cons
Learn global features of images	Have a large number of parameters
Not as sensitive to data augmentation as CNNs	Not as efficient as CNNs at processing images
Can be used for a variety of image classification tasks	Not as interpretable as CNNs

https://arxiv.org/pdf/2010.11929.pdf