Lecture 6. Convolutional Neural Networks

Alex Avdyushenko

Kazakh-British Technical University

October 15, 2022

Five-minute questions

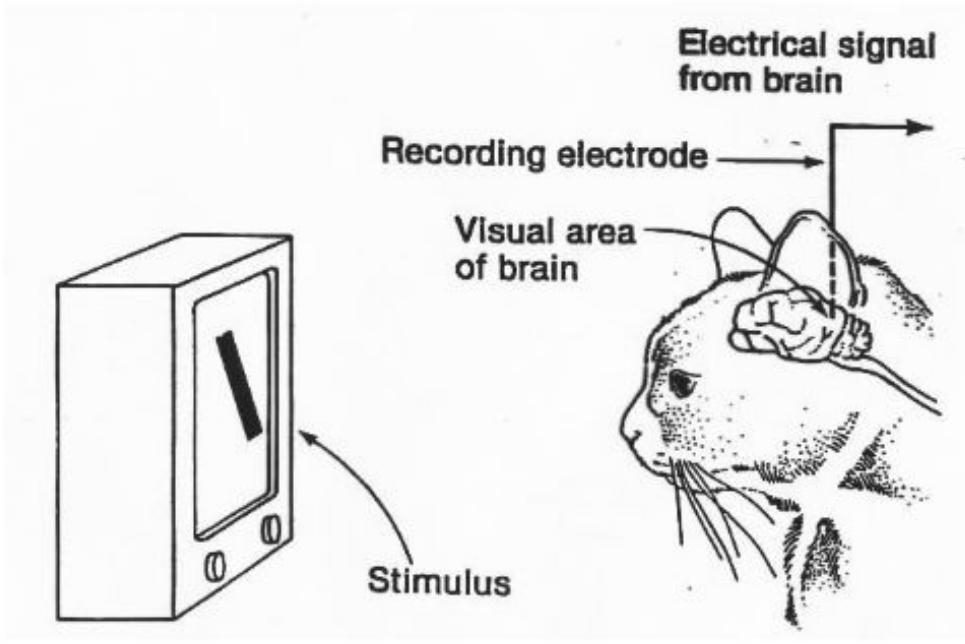
Five-minute questions

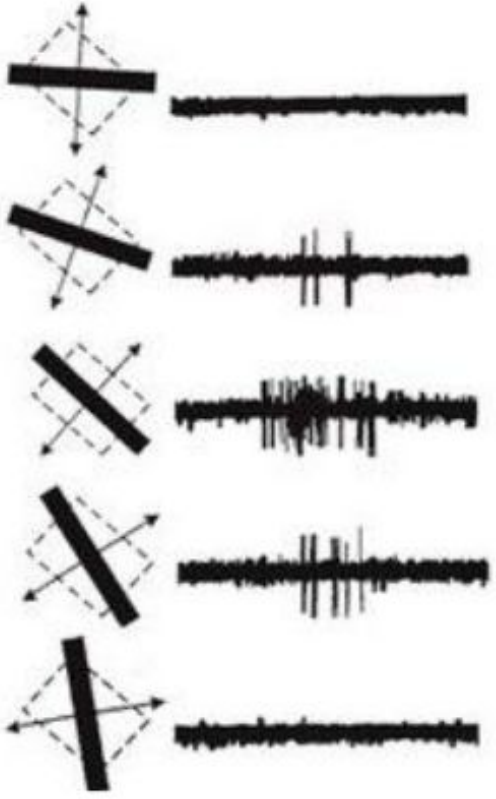
- What is Neuron in deep learning?
- What is ImageNet?
- Give some examples of activation functions.

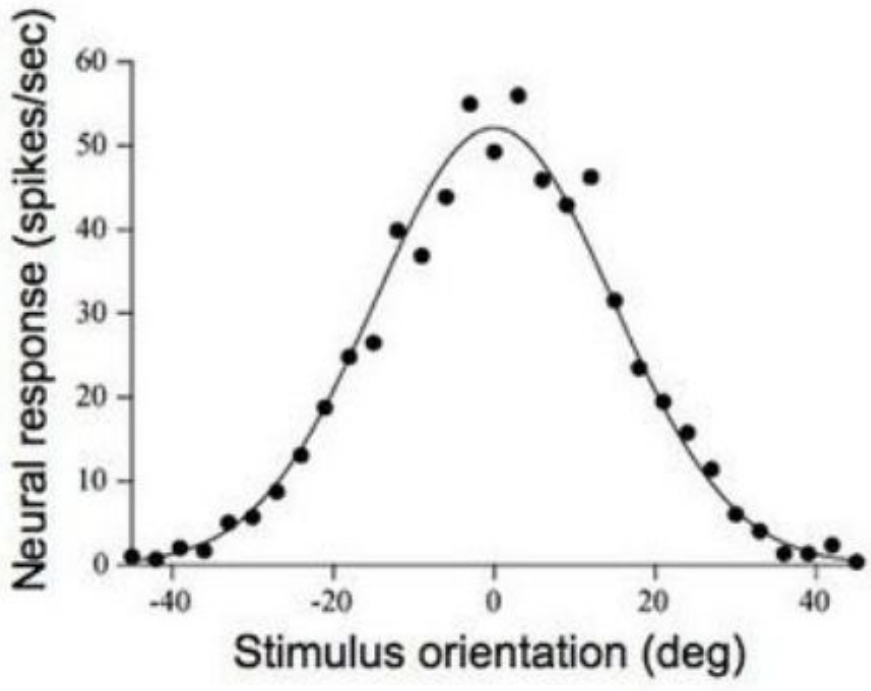
Hubel & Wiesel (1959)

History

The Nobel Prize in Physiology or Medicine, 1981

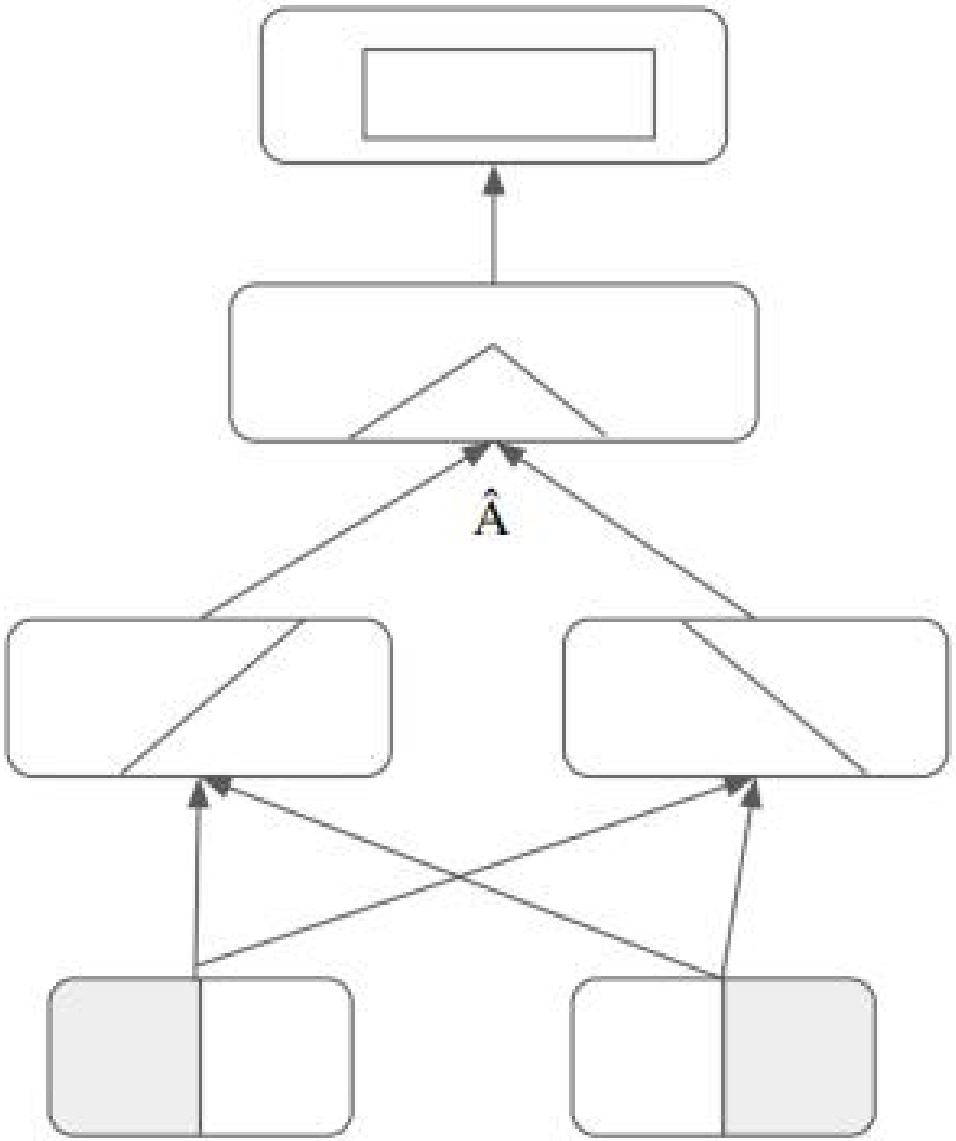

Electrical signal
from brain
Recording electrode
Visual area
of brain
Stimulus




Neural response (spikes/sec)
60
50
40
30
20
10
0
-40
-20
0
20
40
Stimulus orientation (deg)

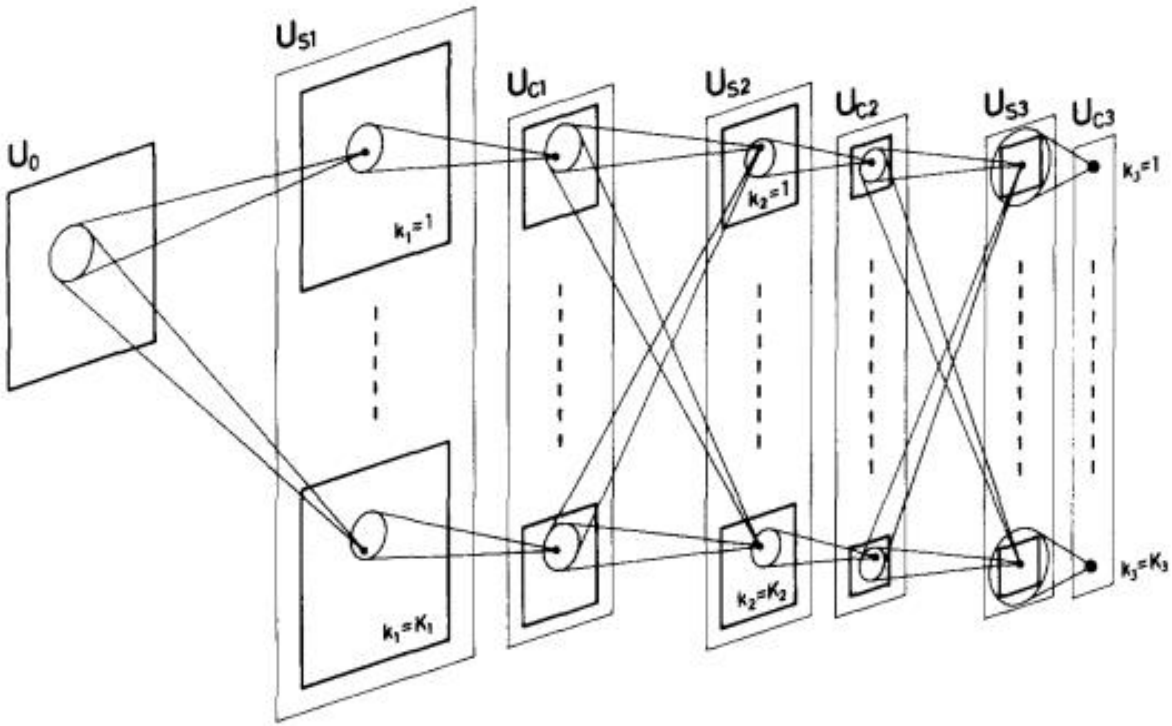
Idea of hierarchical organization of vision

History



Fukushima (1980)

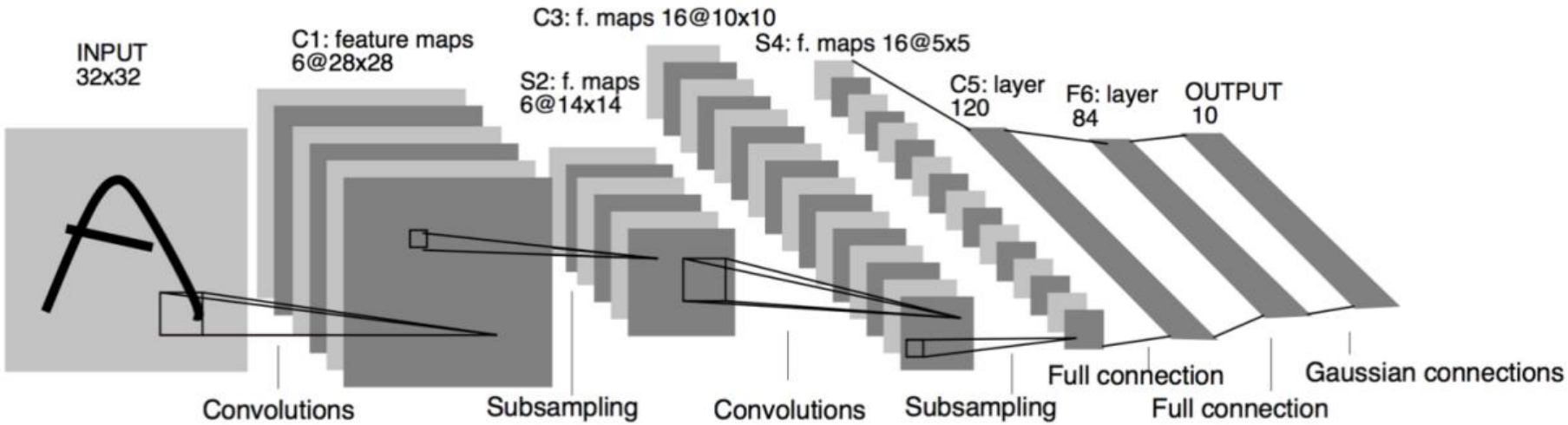
History



Convolutions and activations have already been used, but without gradient descent optimization and supervised learning

Lekun, Bottou, Bengio, Haffner (1998)

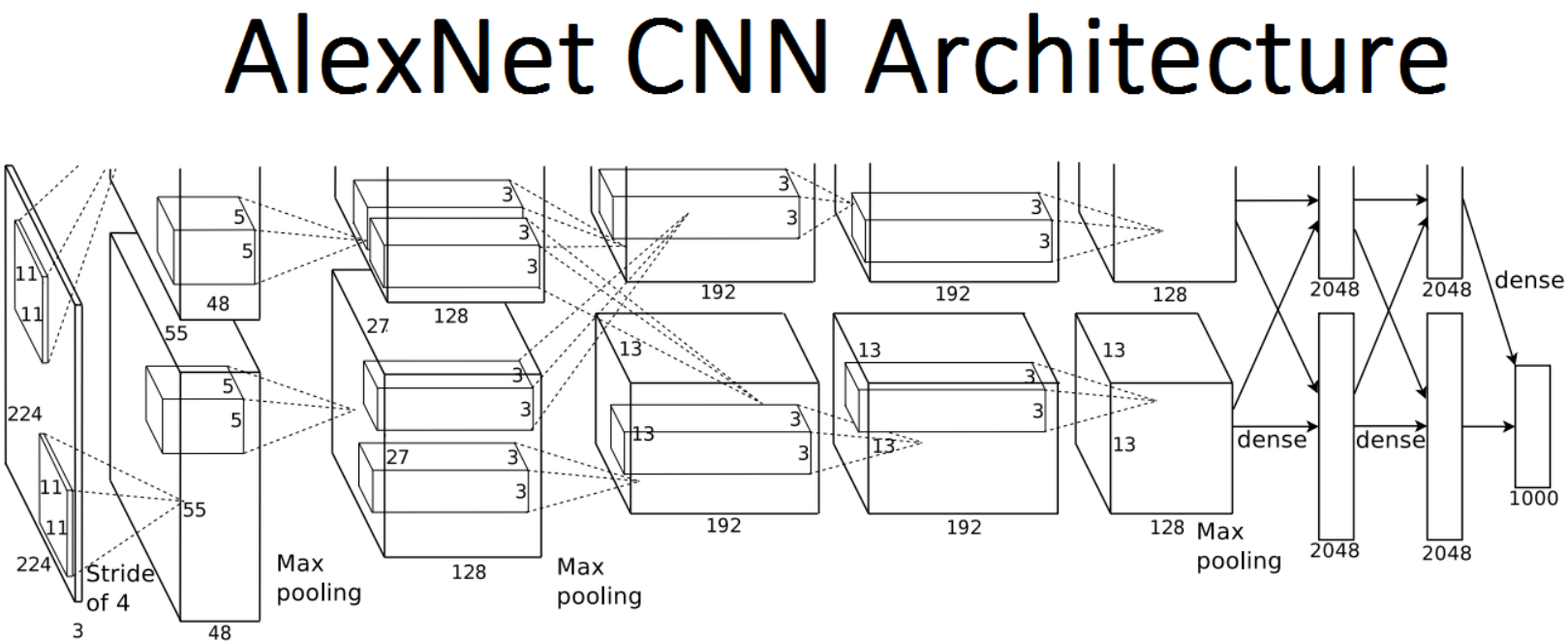
First success



Krizhevsky, Sutskever, Hinton (2012)

A real breakthrough

The Winner of the ImageNet contest of the 2012 year



Linear model

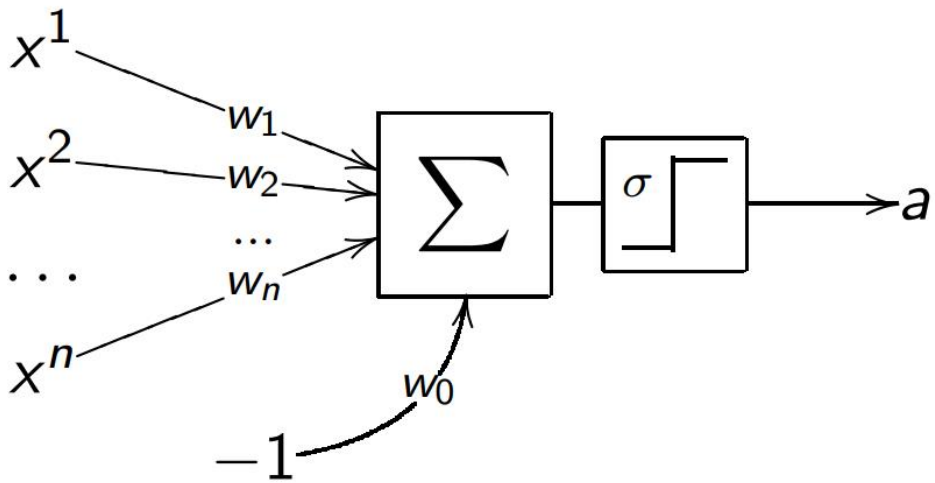
Recall

$x^1, x^2, \ldots, x^n \in \mathbb{R}$ — numerical features of one object x

$w_0, w_1, \ldots, w_n \in \mathbb{R}$ — weights of features

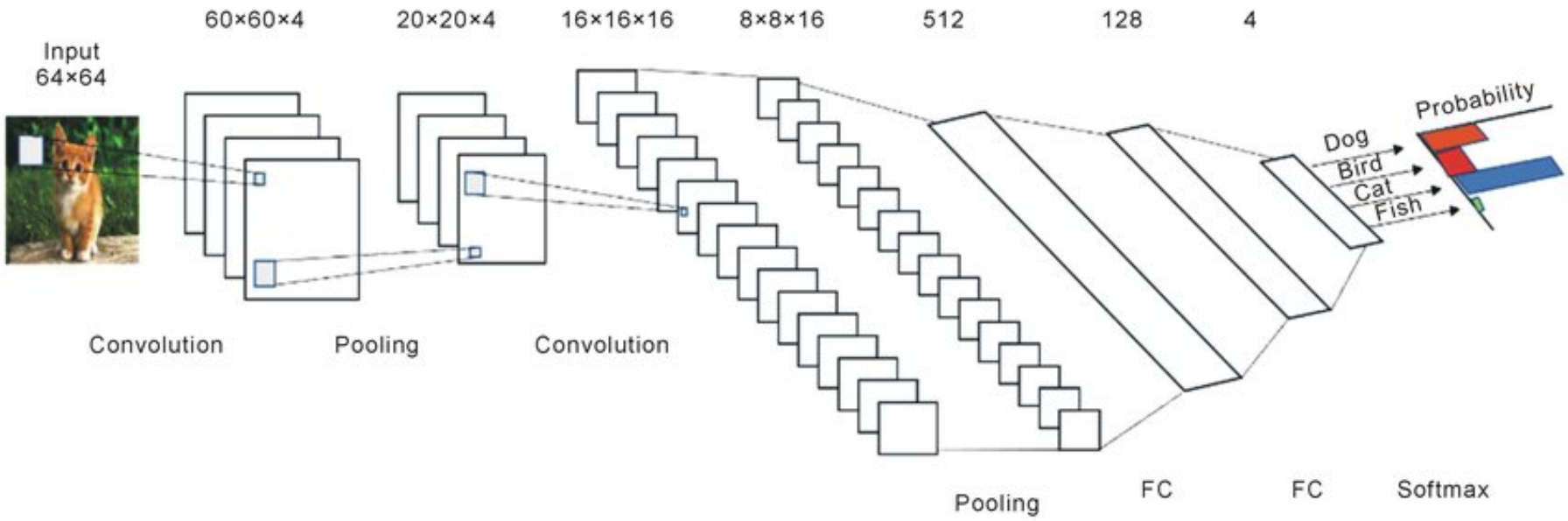
$$a(x, w) = \sigma(\langle w, x \rangle) = \sigma\left(\sum_{j=1}^{n} w_j f_j(x) - w_0\right),$$

$\sigma(z)$ — activation function, for example one of the: $\operatorname{sign}(z)$, $\frac{1}{1+e^{-z}}$, $(z)_+$



Neural network as a combination of linear models

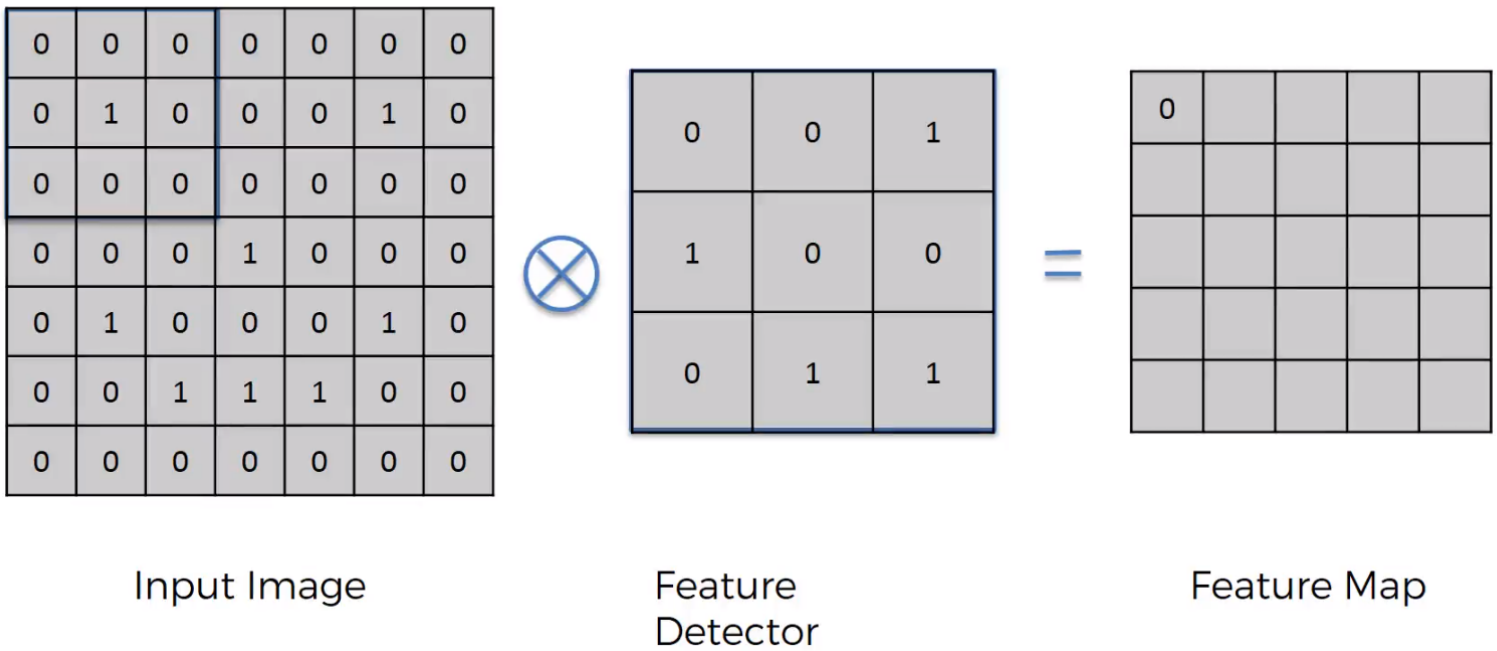
Convolutional Neural Network



Convolution

Convolution in neural networks — the sum of products of elements

- radical reudce of training parameters $28^2 = 784 \to 9 = 3^2$ to get the same accuracy
- Directions x and y are built into the model



Question 1

Why is it called «convolution»?

Question 1

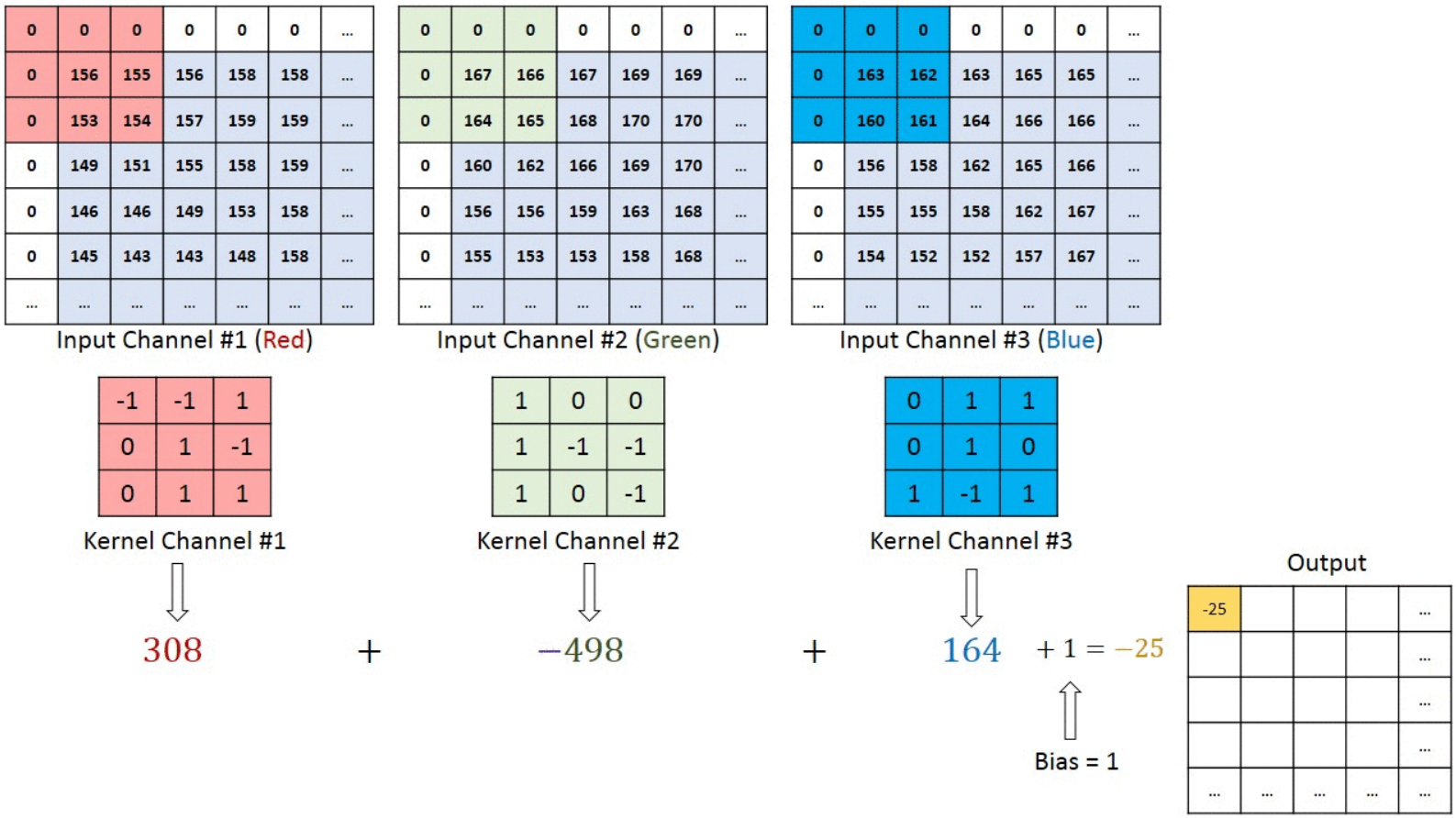
Why is it called «convolution»?

Note

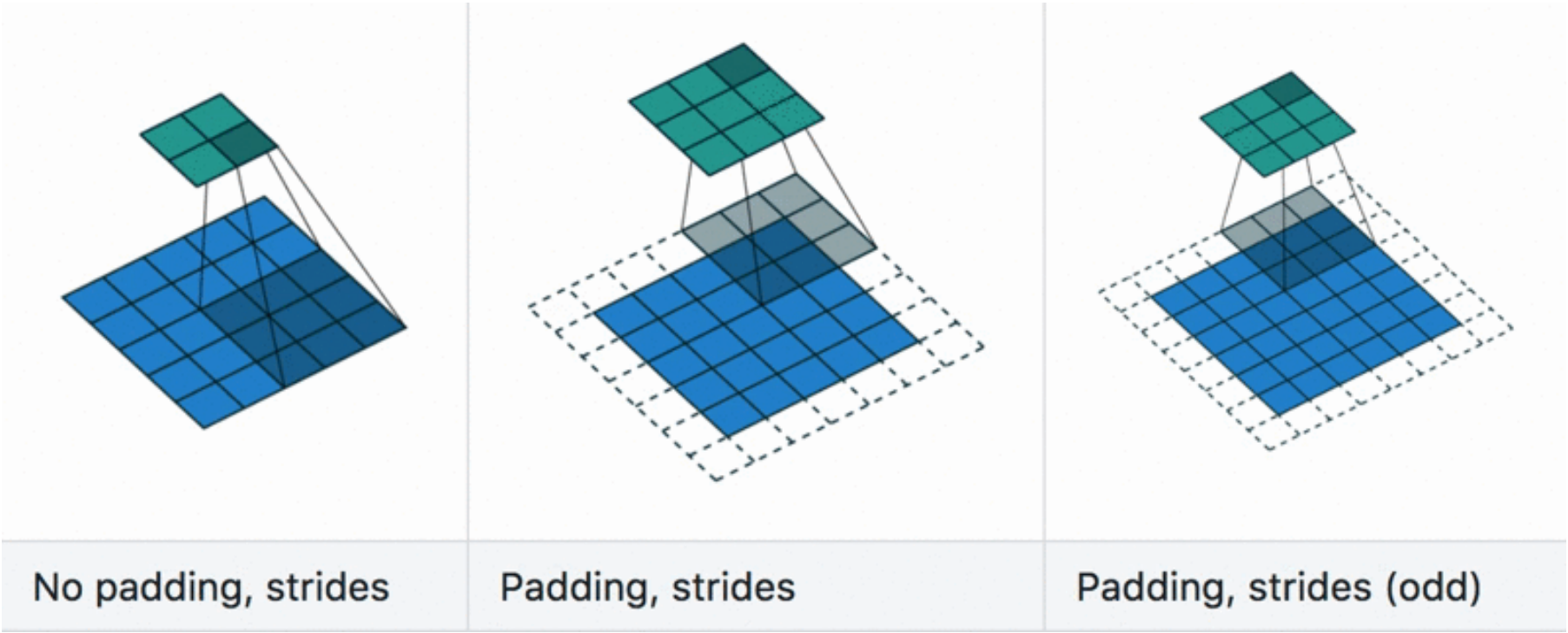
The implementation of convolution effectively multiplies a matrix by a vector. Here, for example, an article with the implementation of Winograd transformation in cuDNN.

Convolution operation example

Kernel $3 \times 3 \times 3$ (Width $\times$ Height $\times$ Channel numbers)

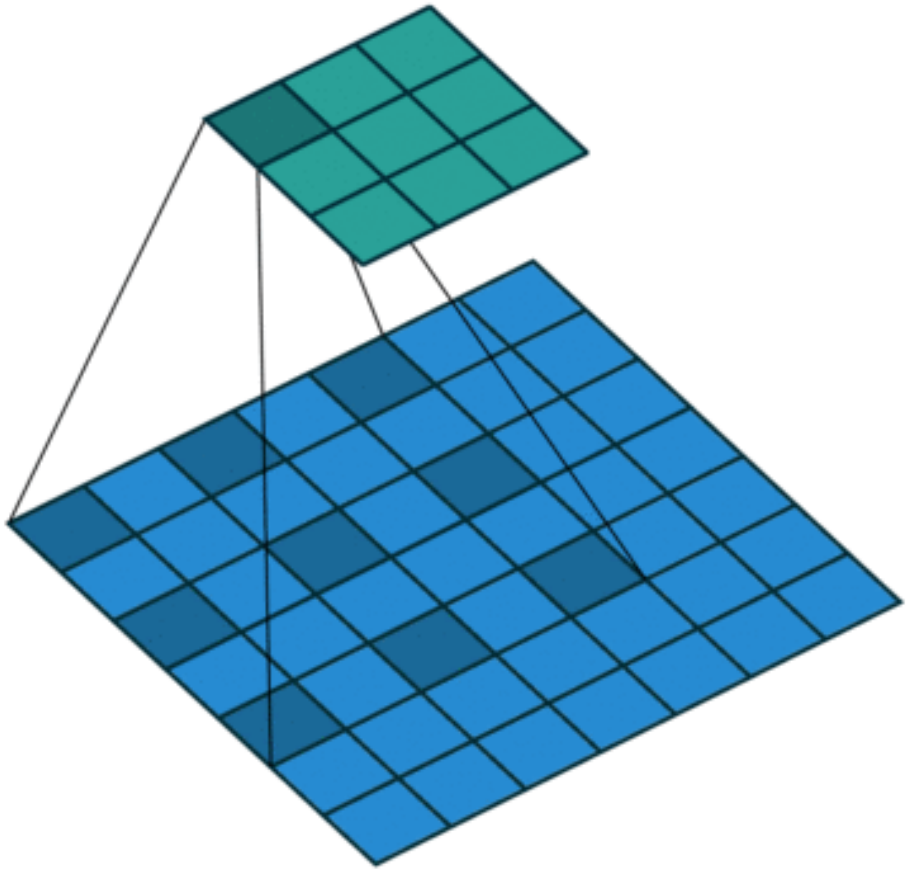


Padding and stride



No padding, strides | Padding, strides | Padding, strides (odd)

Dilation (=2)

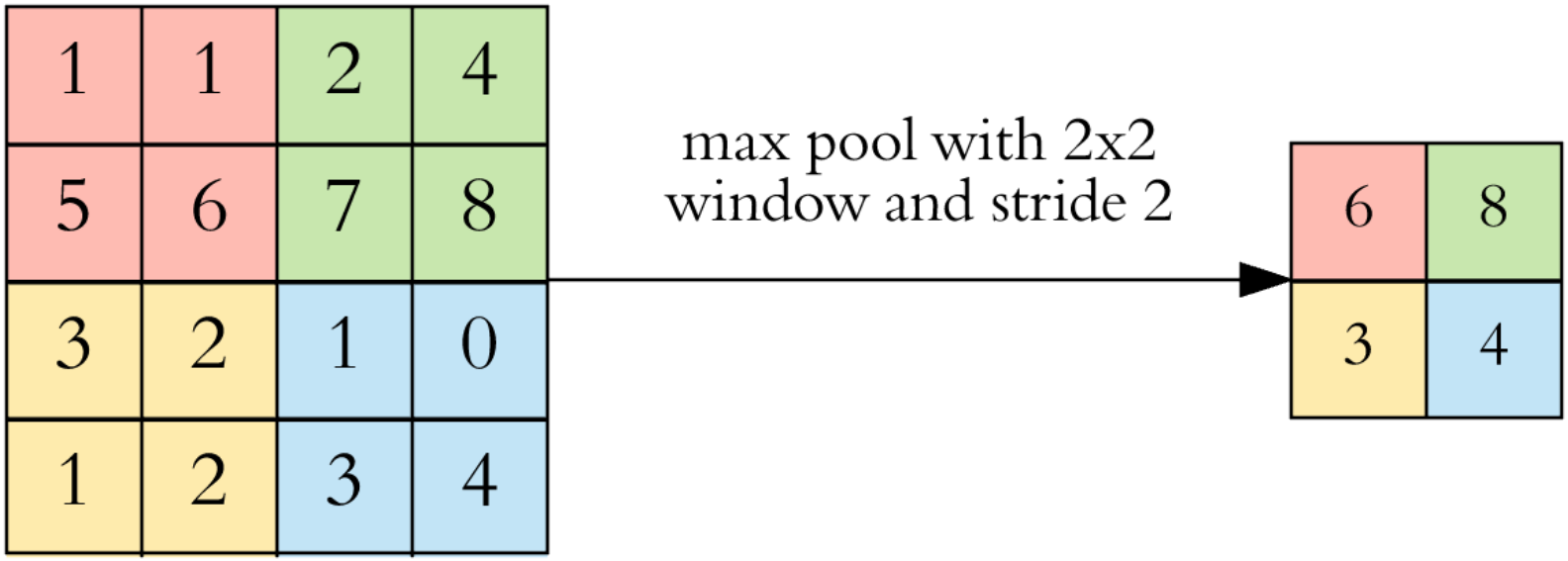


Calculate the size of the output

- **F**ilter size = 3×3 → 3
- **I**nput size = 28×28 → 28
- **S**tride = 1x1 → 1
- **P**adding = 0x0 → 0

Output size = (I - F + 2*P)/S + 1 = (28 - 3 + 2*0)/1 + 1 = 26
Output size = 26 → 26 × 26

Pooling


1 1 2 4
5 6 7 8
3 2 1 0
1 2 3 4
max pool with 2x2
window and stride 2
6 8
3 4

Sigmoid Activation Functions

Recall

- Logistic sigmoid $\sigma(z) = \frac{1}{1+\exp(-z)}$

Sigmoid Activation Functions

Recall

- Logistic sigmoid $\sigma(z) = \frac{1}{1+\exp(-z)}$
- Hyperbolic tangent $\tanh(z) = \frac{\exp(z)-\exp(-z)}{\exp(z)+\exp(-z)}$

Sigmoid Activation Functions

Recall

- Logistic sigmoid $\sigma(z) = \frac{1}{1+\exp(-z)}$
- Hyperbolic tangent $\tanh(z) = \frac{\exp(z)-\exp(-z)}{\exp(z)+\exp(-z)}$
- continuous approximations of threshold function

Sigmoid Activation Functions

Recall

- Logistic sigmoid $\sigma(z) = \frac{1}{1+\exp(-z)}$
- Hyperbolic tangent $\tanh(z) = \frac{\exp(z)-\exp(-z)}{\exp(z)+\exp(-z)}$
- continuous approximations of threshold function
- can lead to vanishing gradient problem and "paralysis"of the network

Sigmoid Activation Functions

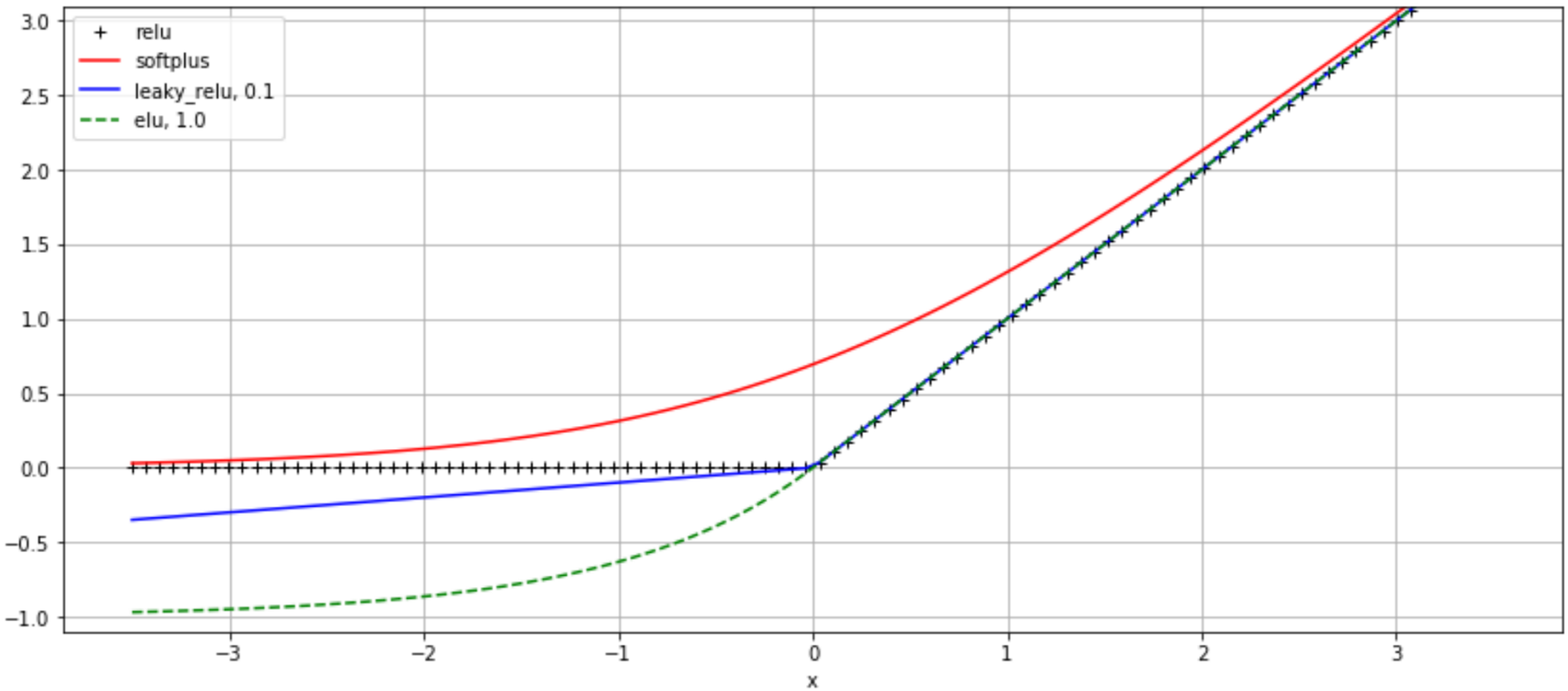
Recall

- Logistic sigmoid $\sigma(z) = \frac{1}{1+\exp(-z)}$
- Hyperbolic tangent $\tanh(z) = \frac{\exp(z)-\exp(-z)}{\exp(z)+\exp(-z)}$
- continuous approximations of threshold function
- can lead to vanishing gradient problem and "paralysis"of the network

Question 2

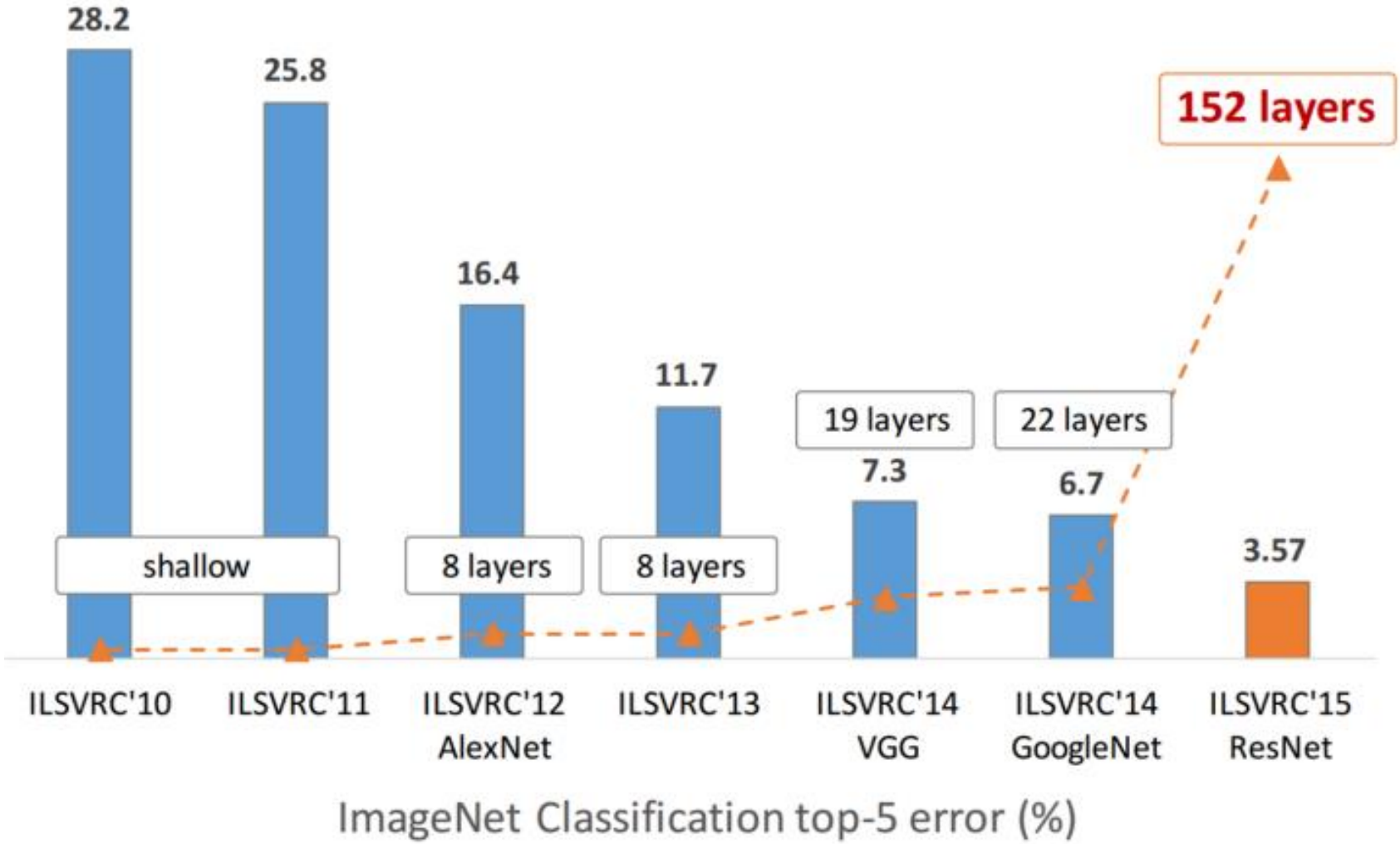
What are the disadvantages of the logistic sigmoid?

Let's look at the charts again



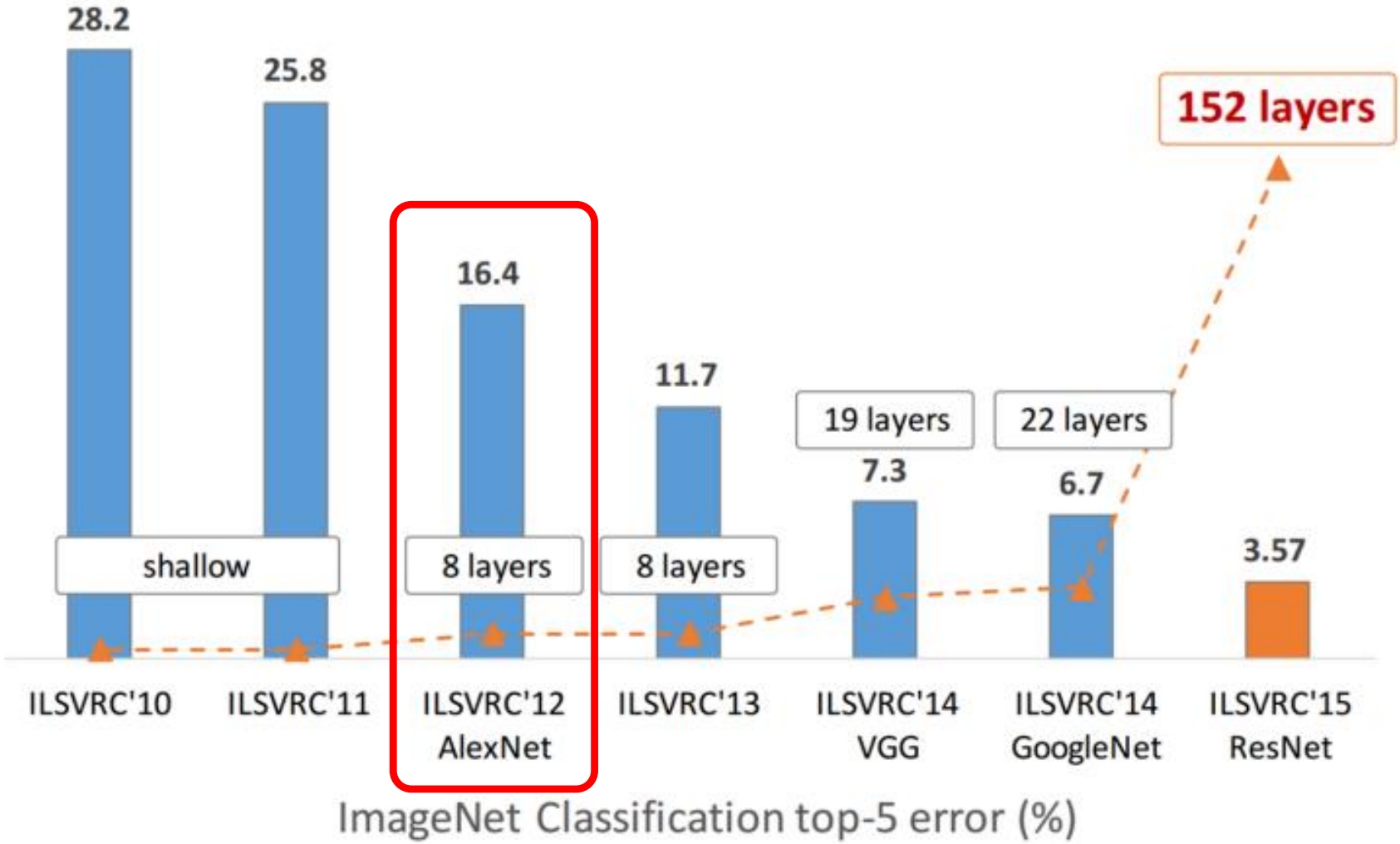
The progress of convolutional networks

Or a brief history of ImageNet



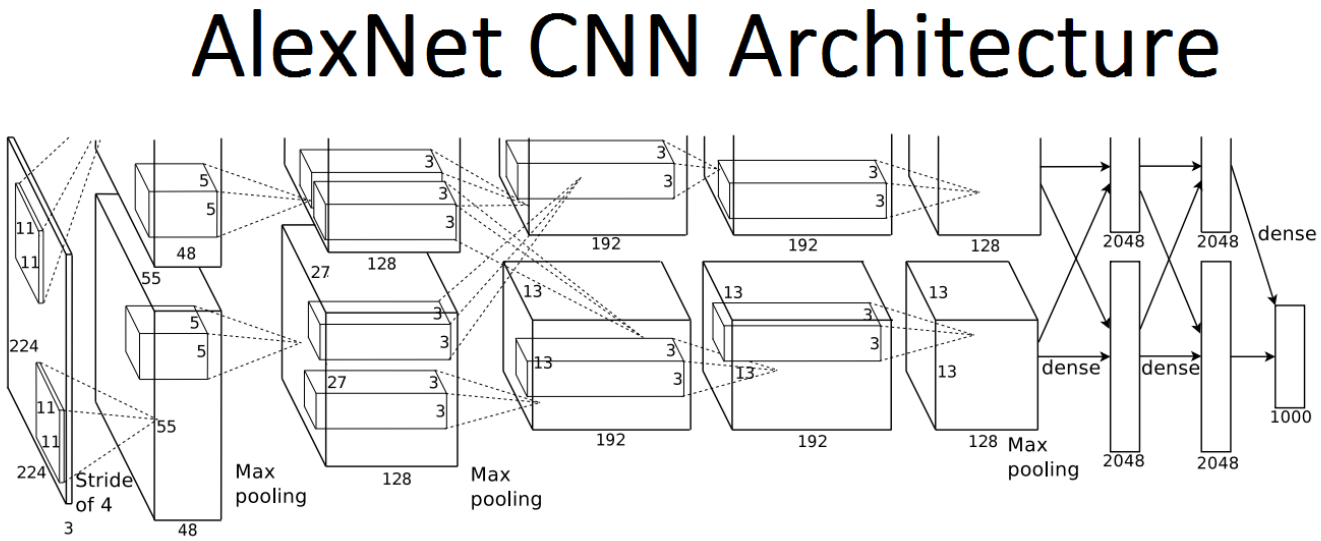
The progress of convolutional networks

Or a brief history of ImageNet



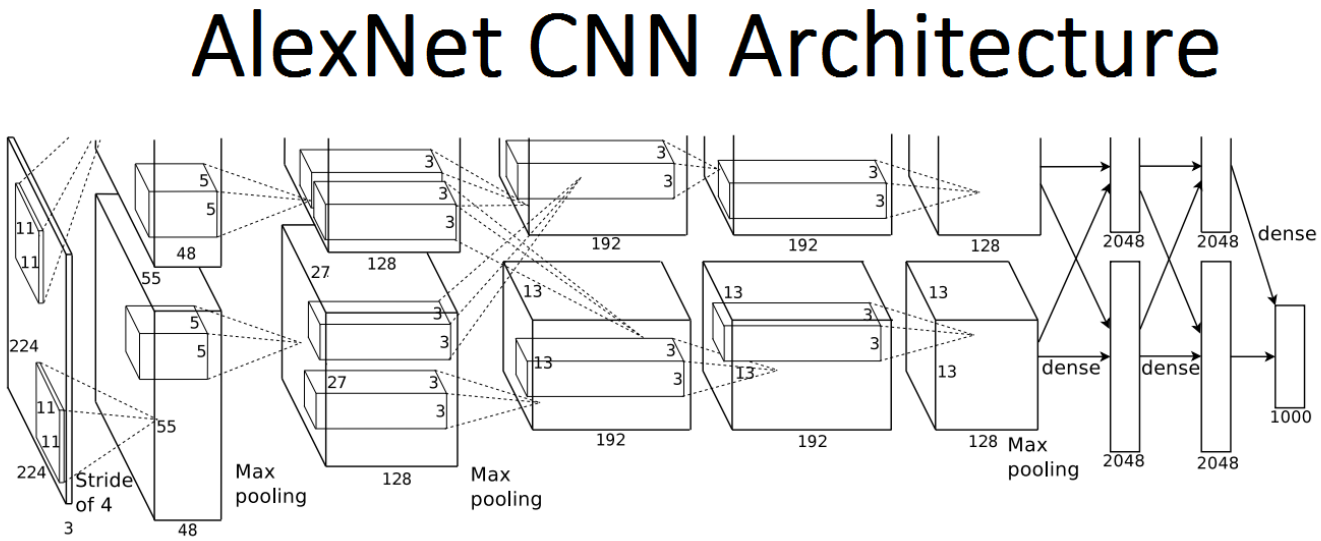
AlexNet (Krizhevsky, Sutskever, Hinton, 2012)

The Winner of the ImageNet contest of the 2012 year



AlexNet (Krizhevsky, Sutskever, Hinton, 2012)

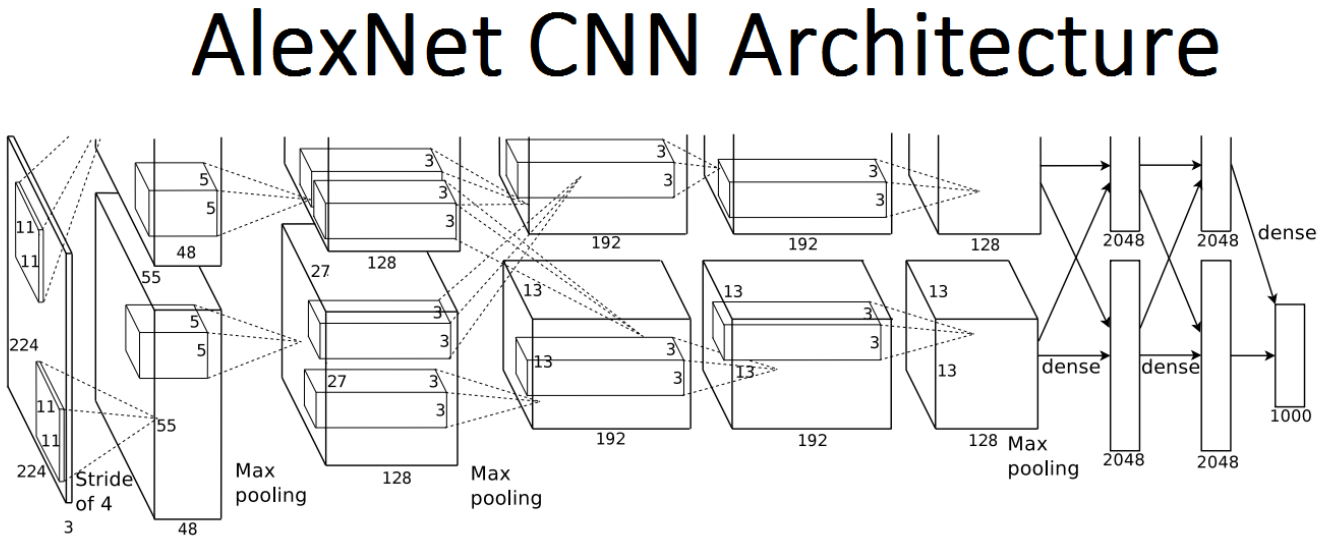
The Winner of the ImageNet contest of the 2012 year



- ReLU activation

AlexNet (Krizhevsky, Sutskever, Hinton, 2012)

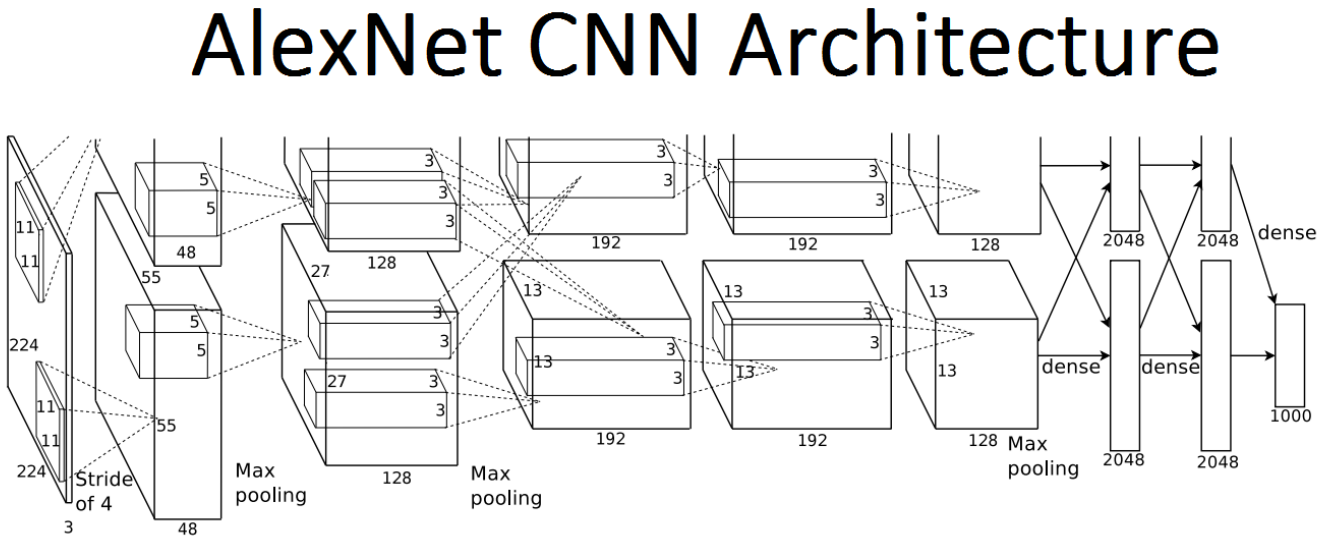
The Winner of the ImageNet contest of the 2012 year



- ReLU activation
- L2 regularization 5e-4

AlexNet (Krizhevsky, Sutskever, Hinton, 2012)

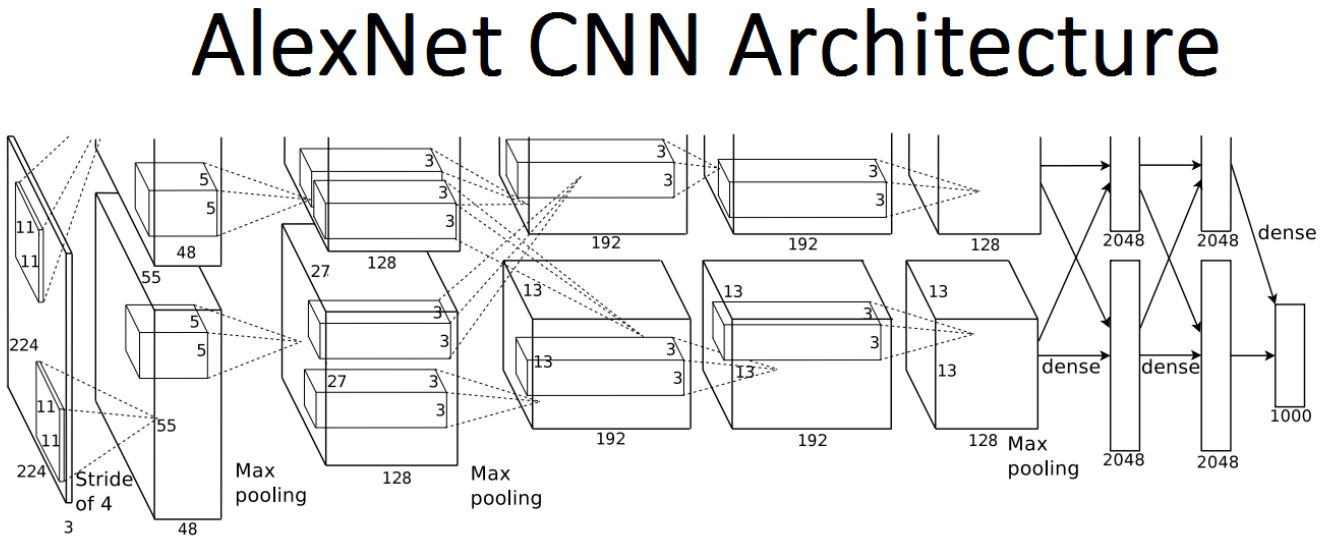
The Winner of the ImageNet contest of the 2012 year



- ReLU activation
- L2 regularization 5e-4
- *data augmentation*

AlexNet (Krizhevsky, Sutskever, Hinton, 2012)

The Winner of the ImageNet contest of the 2012 year

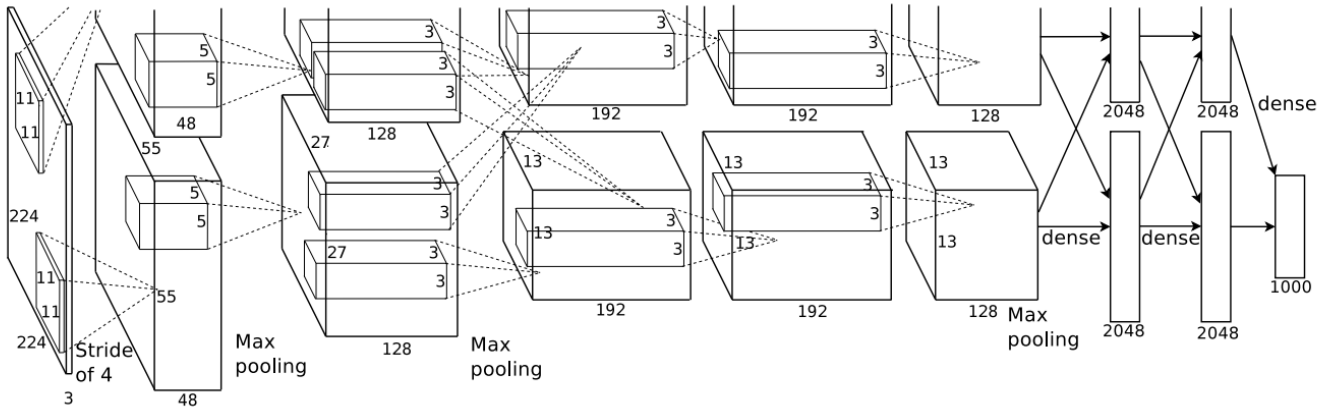


- ReLU activation
- L2 regularization 5e-4
- *data augmentation*
- *dropout 0.5*

AlexNet (Krizhevsky, Sutskever, Hinton, 2012)

The Winner of the ImageNet contest of the 2012 year

AlexNet CNN Architecture

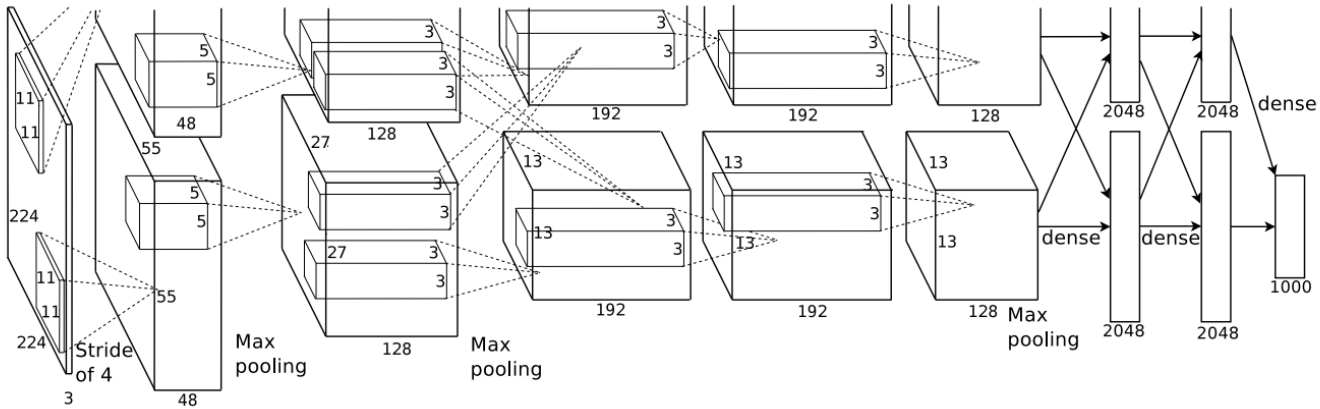


- ReLU activation
- L2 regularization 5e-4
- *data augmentation*
- *dropout 0.5*
- *batch normalization* (batch size 128)

AlexNet (Krizhevsky, Sutskever, Hinton, 2012)

The Winner of the ImageNet contest of the 2012 year

AlexNet CNN Architecture

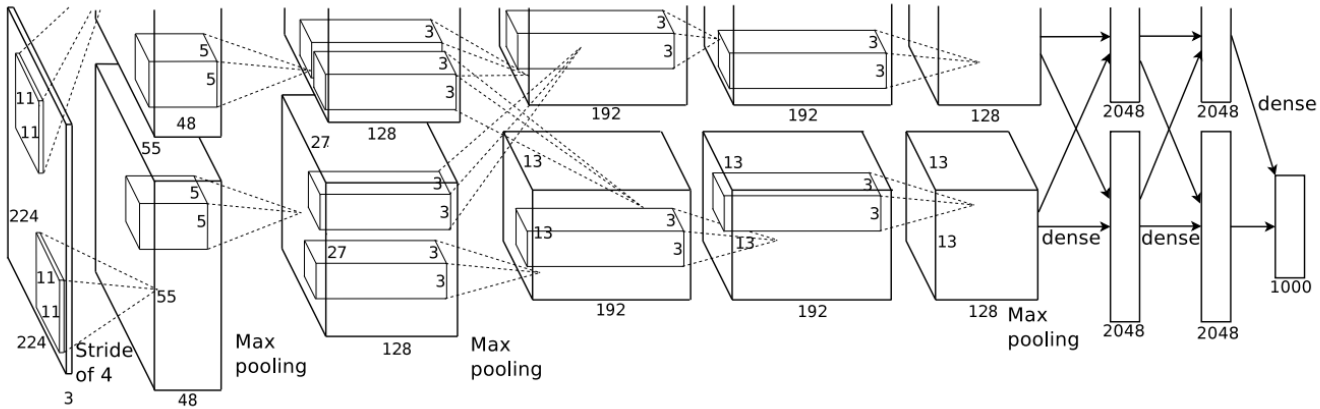


- ReLU activation
- L2 regularization 5e-4
- *data augmentation*
- *dropout 0.5*
- *batch normalization* (batch size 128)
- SGD Momentum 0.9

AlexNet (Krizhevsky, Sutskever, Hinton, 2012)

The Winner of the ImageNet contest of the 2012 year

AlexNet CNN Architecture



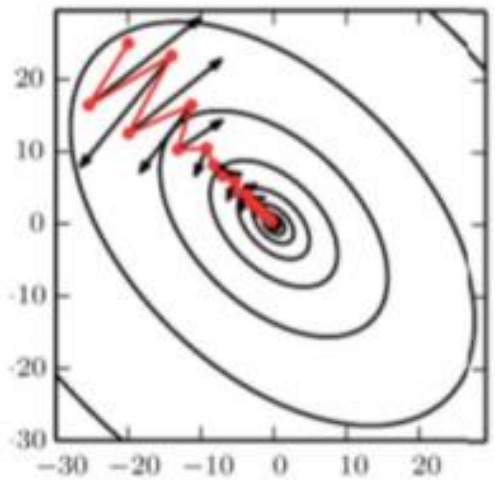
- ReLU activation
- L2 regularization 5e-4
- *data augmentation*
- *dropout 0.5*
- *batch normalization* (batch size 128)
- SGD Momentum 0.9
- Learning rate 1e-2, then decrease by 10 times after quality stabilization on the test. Top5 final accuracy on ImageNet — 25.8% $\rightarrow$16.4%

Momentum method

Momentum accumulation method [B.T.Polyak, 1964] — exponential moving average of the gradient over $\frac{1}{1-\gamma}$ last iterations:

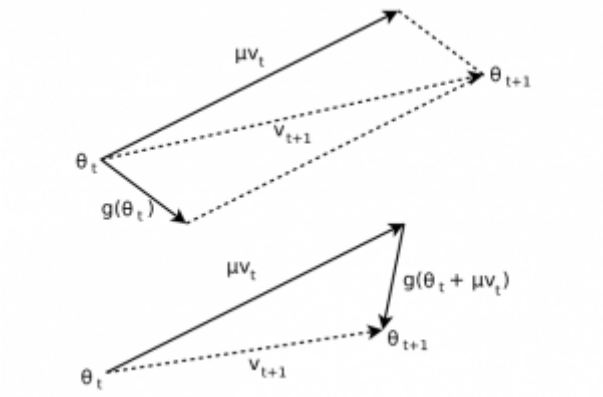
$$\nu = \gamma\nu + (1-\gamma)\mathcal{L}'_i(w)$$

$$w = w - \eta\nu$$



Nesterov Accelerated Gradient (NAG, 1983)

$$\nu = \gamma\nu + (1-\gamma)\mathcal{L}_i'(w - \eta\gamma v)$$
$$w = w - \eta\nu$$



(Top) Momentum method
(Bottom) Nesterov Accelerated Gradient

μ is the decaying parameter, same as η

Summary

- Convolutional networks are very well suited for image processing
- They are a bit like similar to biological vision mechanisms
- At the same time, flexible and computationally efficient
- Today, the «de facto» standard for computer vision tasks (classification, detection, segmentation, generation)

Summary

- Convolutional networks are very well suited for image processing
- They are a bit like similar to biological vision mechanisms
- At the same time, flexible and computationally efficient
- Today, the «de facto» standard for computer vision tasks (classification, detection, segmentation, generation)
- Will be next
  - various optimization algorithms: adam, RMSProp
  - dropout
  - choice of initial approximation
  - ResNet and WideResNet

Summary

- Convolutional networks are very well suited for image processing
- They are a bit like similar to biological vision mechanisms
- At the same time, flexible and computationally efficient
- Today, the «de facto» standard for computer vision tasks (classification, detection, segmentation, generation)
- Will be next
  - various optimization algorithms: adam, RMSProp
  - dropout
  - choice of initial approximation
  - ResNet and WideResNet

What else can you watch?

- Demo by Andrey Karpaty
- There is a famous course from Stanford «CS231n: Convolutional Neural Networks for Visual Recognition»: http://cs231n.github.io/