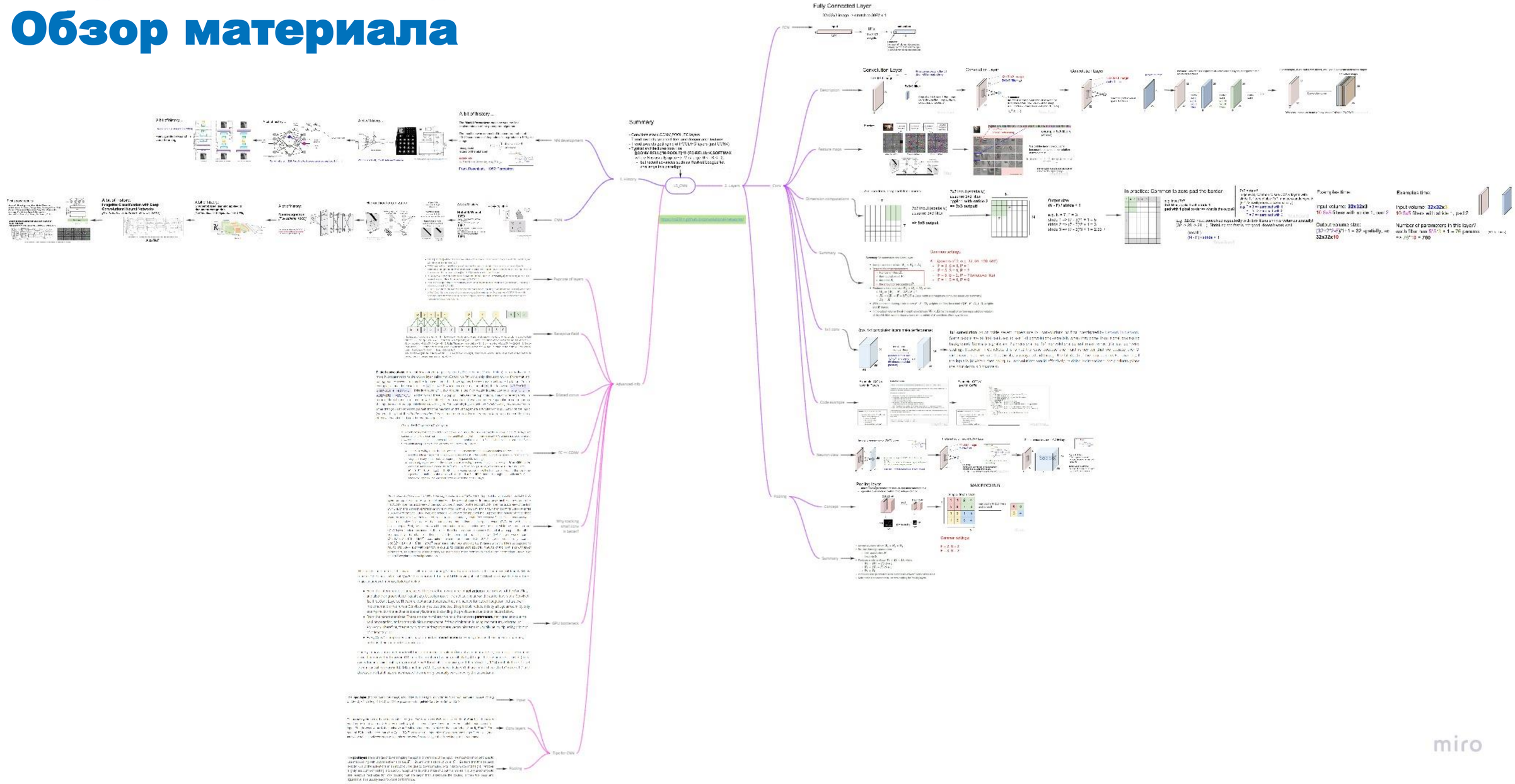
МФТИ

Лекция 9
Введение в глубокое обучение для компьютерного зрения

Курс «Компьютерное зрение»

Глубокое обучение в CV

Обзор материала



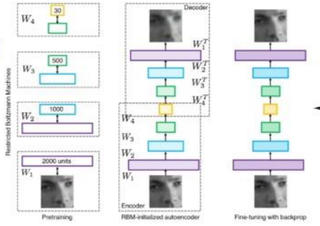
Исторические предпосылки

Развитие нейросетевых методов и сверточных сетей

A bit of history...

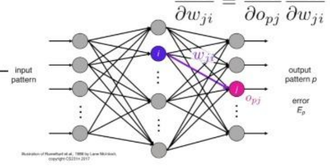
[Hinton and Salakhutdinov 2006]

Reinvigorated research in Deep Learning



A bit of history...

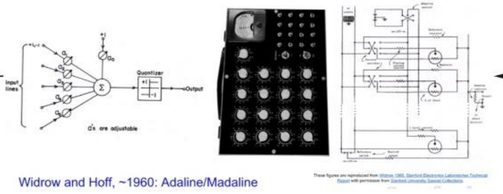
$$\frac{\partial E_p}{\partial w_{ji}} = \frac{\partial E_p}{\partial o_{pj}} \frac{\partial o_{pj}}{\partial w_{ji}}$$



recognizable math

Rumelhart et al., 1986: First time back-propagation became popular

A bit of history...



Widrow and Hoff, ~1960: Adaline/Madaline

A bit of history...

The **Mark I Perceptron** machine was the first implementation of the perceptron algorithm.

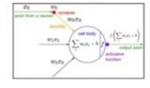
The machine was connected to a camera that used 20×20 cadmium sulfide photocells to produce a 400-pixel image.

recognized letters of the alphabet

$$f(x) = \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{otherwise} \end{cases}$$

update rule:

$$w_i(t+1) = w_i(t) + \alpha(d_j - y_j(t))x_{j,i}$$



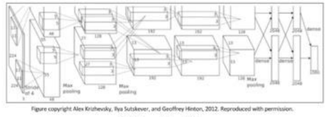
Frank Rosenblatt, ~1957: Perceptron

NN development

A bit of history:

ImageNet Classification with Deep Convolutional Neural Networks

[Krizhevsky, Sutskever, Hinton, 2012]

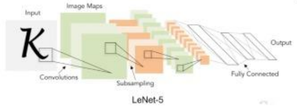


"AlexNet"

A bit of history:

Gradient-based learning applied to document recognition

[LeCun, Bottou, Bengio, Haffner 1998]



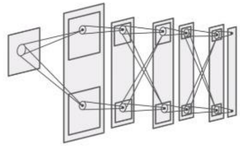
LeNet-5

A bit of history:

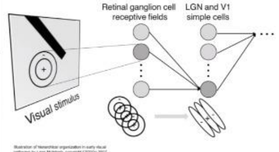
Neurocognitron

[Fukushima 1980]

"sandwich" architecture (SCSCSC...)
simple cells: modifiable parameters
complex cells: perform pooling



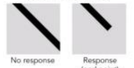
Hierarchical organization



Simple cells: Response to light orientation

Complex cells: Response to light orientation and movement

Hypercomplex cells: response to movement with an end point



A bit of history:

Hubel & Wiesel,

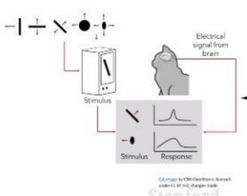
1959

RECEPTIVE FIELDS OF SINGLE NEURONES IN THE CAT'S STRIATE CORTEX

1962

RECEPTIVE FIELDS, BINOCULAR INTERACTION AND FUNCTIONAL ARCHITECTURE IN THE CAT'S VISUAL CORTEX

1968...

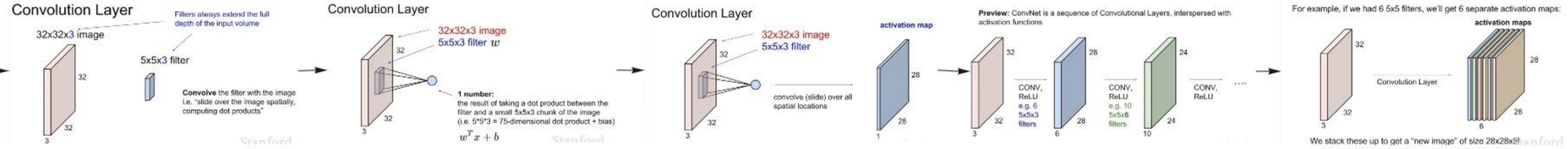


CNN

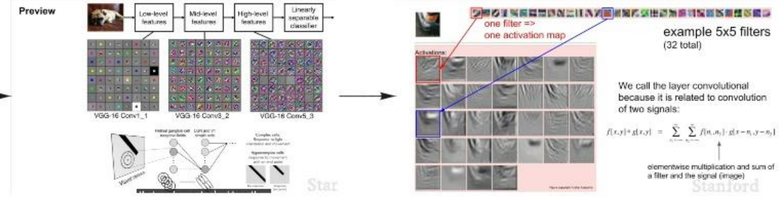
Сверточные нейронные сети

Сверточные слои

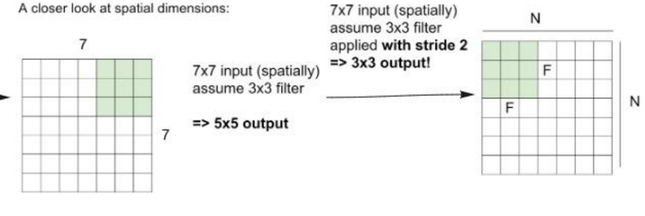
Description

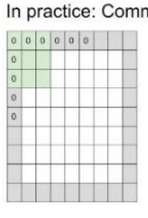


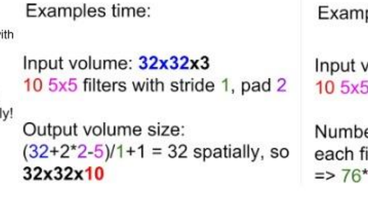
Feature maps

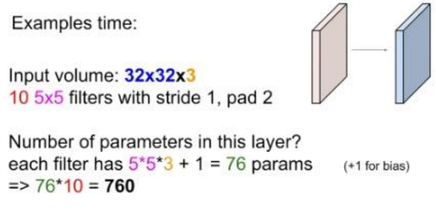


ension computations

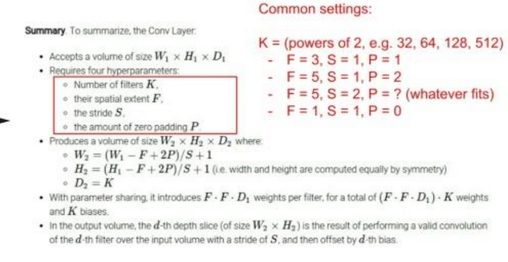








Summary



Сверточные нейронные сети

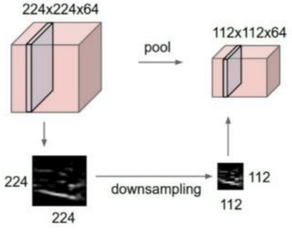
Pooling и полносвязные слои

Pooling

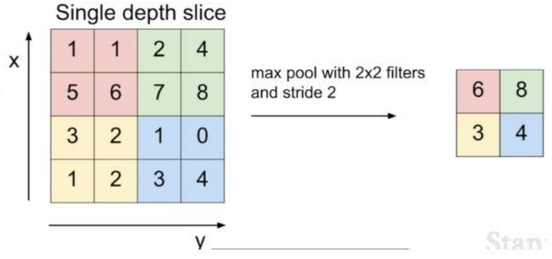
Concept

Pooling layer

- makes the representations smaller and more manageable
- operates over each activation map independently:



MAX POOLING



Common settings:

F = 2, S = 2
F = 3, S = 2

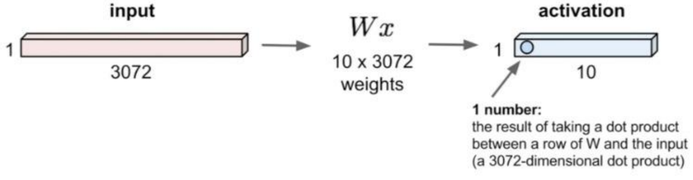
Summary

- Accepts a volume of size $W_1 \times H_1 \times D_1$
- Requires three hyperparameters:
  - their spatial extent F,
  - the stride S,
- Produces a volume of size $W_2 \times H_2 \times D_2$ where:
  - $W_2 = (W_1 - F)/S + 1$
  - $H_2 = (H_1 - F)/S + 1$
  - $D_2 = D_1$
- Introduces zero parameters since it computes a fixed function of the input
- Note that it is not common to use zero-padding for Pooling layers

FCN

Fully Connected Layer

32x32x3 image -> stretch to 3072 x 1



Сверточные нейронные сети

Назначение слоев

- INPUT [32x32x3] will hold the raw pixel values of the image, in this case an image of width 32, height 32, and with three color channels R,G,B.
- CONV layer will compute the output of neurons that are connected to local regions in the input, each computing a dot product between their weights and a small region they are connected to in the input volume. This may result in volume such as [32x32x12] if we decided to use 12 filters.
- RELU layer will apply an elementwise activation function, such as the $\boldsymbol{max}(\boldsymbol{0}, \boldsymbol{x})$ thresholding at zero. This leaves the size of the volume unchanged ([32x32x12]).
- POOL layer will perform a downsampling operation along the spatial dimensions (width, height), resulting in volume such as [16x16x12].
- FC (i.e. fully-connected) layer will compute the class scores, resulting in volume of size [1x1x10], where each of the 10 numbers correspond to a class score, such as among the 10 categories of CIFAR-10. As with ordinary Neural Networks and as the name implies, each neuron in this layer will be connected to all the numbers in the previous volume.

Сверточные нейронные сети

Receptive field

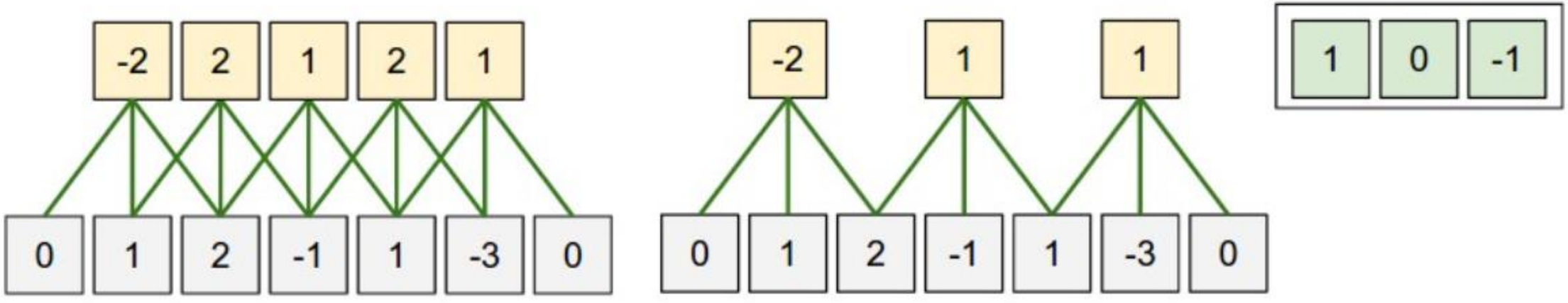


Illustration of spatial arrangement. In this example there is only one spatial dimension (x-axis), one neuron with a receptive field size of F = 3, the input size is W = 5, and there is zero padding of P = 1. **Left:** The neuron strided across the input in stride of S = 1, giving output of size (5 - 3 + 2)/1+1 = 5. **Right:** The neuron uses stride of S = 2, giving output of size (5 - 3 + 2)/2+1 = 3. Notice that stride S = 3 could not be used since it wouldn't fit neatly across the volume. In terms of the equation, this can be determined since (5 - 3 + 2) = 4 is not divisible by 3.

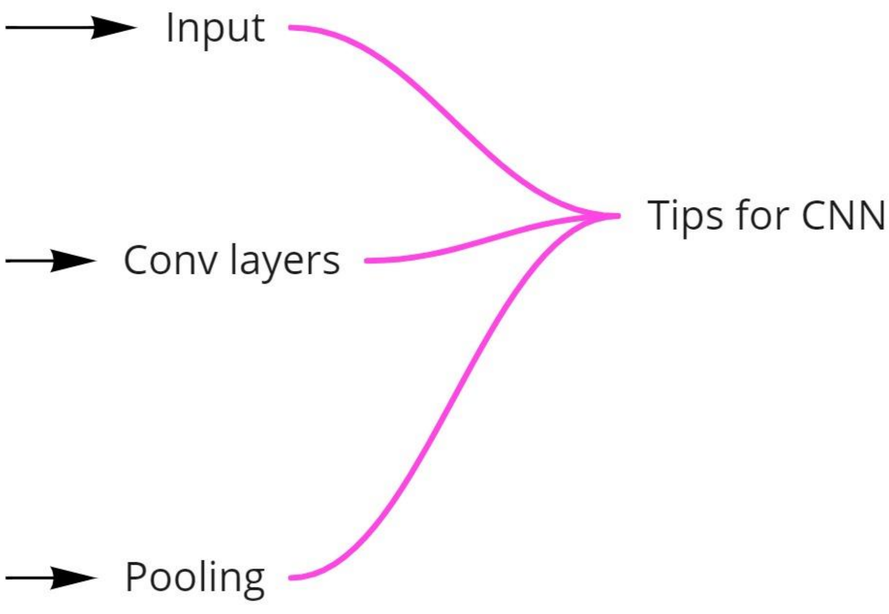
The neuron weights are in this example [1,0,-1] (shown on very right), and its bias is zero. These weights are shared across all yellow neurons (see parameter sharing below).

Сверточные нейронные сети
Практические советы

The **input layer** (that contains the image) should be divisible by 2 many times. Common numbers include 32 (e.g. CIFAR-10), 64, 96 (e.g. STL-10), or 224 (e.g. common ImageNet ConvNets), 384, and 512.

The **conv layers** should be using small filters (e.g. 3x3 or at most 5x5), using a stride of $S = 1$, and crucially, padding the input volume with zeros in such way that the conv layer does not alter the spatial dimensions of the input. That is, when $F = 3$, then using $P = 1$ will retain the original size of the input. When $F = 5$, $P = 2$. For a general F, it can be seen that $P = (F - 1)/2$ preserves the input size. If you must use bigger filter sizes (such as 7x7 or so), it is only common to see this on the very first conv layer that is looking at the input image.

The **pool layers** are in charge of downsampling the spatial dimensions of the input. The most common setting is to use max-pooling with 2x2 receptive fields (i.e. $F = 2$), and with a stride of 2 (i.e. $S = 2$). Note that this discards exactly 75% of the activations in an input volume (due to downsampling by 2 in both width and height). Another slightly less common setting is to use 3x3 receptive fields with a stride of 2, but this makes. It is very uncommon to see receptive field sizes for max pooling that are larger than 3 because the pooling is then too lossy and aggressive. This usually leads to worse performance.



Резюме

- Исторические предпосылки нейросетевых методов
- Описание архитектуры сверточных нейронных сетей
- Особенности сверточных слоев
- Описание полносвязных и pooling слоев
- Практические советы при построении сверточных нейронных сетей

Спасибо за внимание!

Колокольников Георгий Андреевич

Telegram: @Georg_Bell

E-mail: geokolok5@gmail.com

Сайт: https://github.com/GeorgBell

Использованные материалы:

- Гонсалес Р., Вудс Р. Цифровая обработка изображений. – М.: Техносфера, 2012. – 1104 с. – ISBN 978-5-94836-331-8.2.
- Курс лекций cs231n «Convolutional Neural Networks for Visual Recognition» (http://cs231n.stanford.edu).
- Курс лекций HSE «Deep Learning in Computer Vision» (https://www.coursera.org/learn/deep-learning-in-computer-vision)