Lecture 6. Convolutional Neural Networks

Lecture 6. Convolutional Neural Networks

Alex Avdyushenko

Kazakh-British Technical University

October 15, 2022

Hello, today we are going to talk about Convolutional Neural Networks, their history and development.

And we will start with three questions for five minutes based on the materials of the previous lecture.

Please, write answers or send photos with them directly to me in private messages here in Teams, so that others cannot read your message. Last time, a lot of people did it, so I believe that this time you all will succeed.

Lecture 6. Convolutional Neural Networks

Five-minute questions

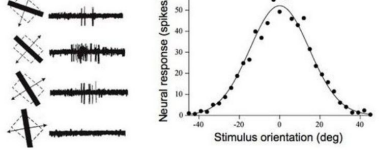
Five-minute questions

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

Please, write answers or send photos with them directly to me in private messages here in Teams, so that others cannot read your message. Last time, a lot of people did it, so I believe that this time you all will succeed.

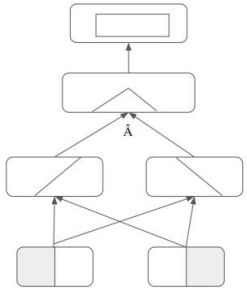
Who is in this photo?

These are Hubel and Wiesel, Nobel Prize winners in Physiology or Medicine, who have explored the mechanisms of human vision.



In their most famous work, they showed a cat variously oriented black stripes and recorded the response of brain neurons.

And in fact, they discovered the neurons responsible for the angle of rotation of the strips.



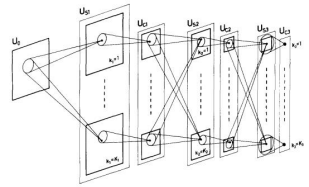
Then, quite quickly, an understanding of the hierarchical organization of human vision came, which was confirmed by other works.

That is, at first the neurons of the brain distinguish stripes, color gradients.

The stripes form corners and further more and more complex elements: geometric shapes and so on up to abstract concepts.

Lecture 6. Convolutional Neural Networks

Fukushima (1980)

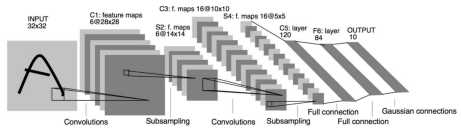


Convolutions and activations have already been used, but without gradient descent optimization and supervised learning

One of the very first neural networks for image analysis was organized in a similar way, but it has not been yet gradient descent optimization and supervised learning.

Lecture 6. Convolutional Neural Networks

Lekun, Bottou, Bengio, Haffner (1998)



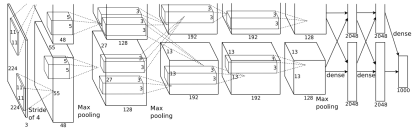
Then quite good results were obtained using the LeNet architecture.

Lecture 6. Convolutional Neural Networks

Krizhevsky, Sutskever, Hinton (2012)

The Winner of the ImageNet contest of the 2012 year

AlexNet CNN Architecture



And the first real breakthrough in image classification was made by AlexNet neural network.

Lecture 6. Convolutional Neural Networks

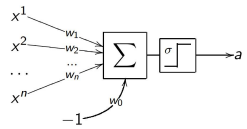
Linear model

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: $\text{sign}(z)$, $\frac{1}{1+e^{-z}}$, $(z)_+$

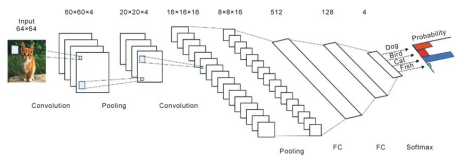


Let's recall the linear model from the last lecture: numerical features only, their weights and one simple activation function. That's it!

And neural network in essence is a combination of linear models

Lecture 6. Convolutional Neural Networks

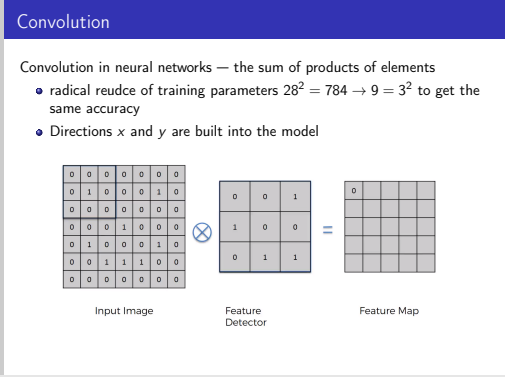
Convolutional Neural Network



Ok, and what is Convolutional Neural Network? This is a fairly natural development of the model. In addition to the fully connected layer already familiar to us, there are also convolutional layers and pooling layers.

Lecture 6. Convolutional Neural Networks

Convolution



The convolutional layer allows you to get the same result many times more efficiently than the fully connected layer.

Moreover, through it, two selected directions x and y are built into the model, which are really physically important for understanding the image.

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

Let's look at example of convolution operation. It is important to notice, that bias is one value

Lecture 6. Convolutional Neural Networks

Calculate the size of the output

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

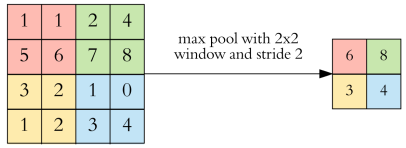
Output size = (I - F + 2*P)/S + 1 = (28 - 3 + 2*0)/1 + 1 = 26
Output size = 26 → 26 × 26

Imagine that we need to calculate the size of the output of the convolution layer. Now we can easily understand why the formula is exactly that

Lecture 6. Convolutional Neural Networks

Pooling

Pooling

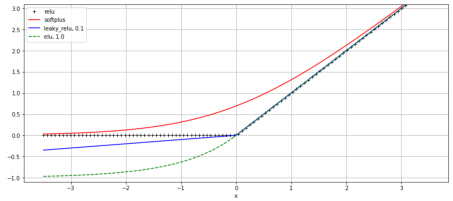


Pooling layer is maybe the simplest layer of all: we choose the maximum element of filter. Or there is average pooling, where we take mean, but it is used quite rarely

Lecture 6. Convolutional Neural Networks

Let's look at the charts again

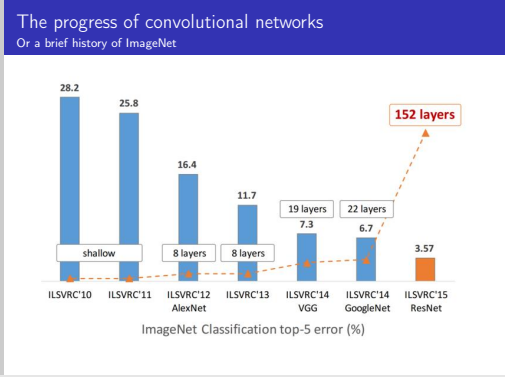
Let's look at the charts again



The main disadvantage of sigmoid is that it's derivateve is close to zero for arguments greater than 3. So your default choice is ReLU, than you could try LeakyReLU and softplus.

Lecture 6. Convolutional Neural Networks

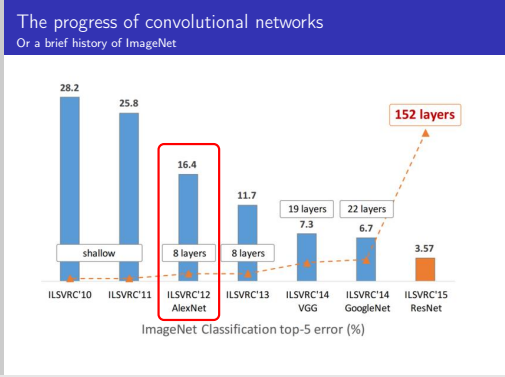
The progress of convolutional networks



We have seen the remarkable progress in solving the problem of image classification of ImageNet collection. Now let's dive into the details of the AlexNet architecture.

Lecture 6. Convolutional Neural Networks

The progress of convolutional networks



We have seen the remarkable progress in solving the problem of image classification of ImageNet collection. Now let's dive into the details of the AlexNet architecture.

Lecture 6. Convolutional Neural Networks

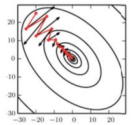
Momentum method

Momentum method

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

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

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



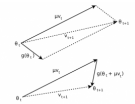
A very popular method for improving the convergence of SGD is the momentum accumulation method. In it, from a physical point of view, the derivative of the loss function becomes the acceleration of the change in model parameters, and not the speed as in classical SGD.

Lecture 6. Convolutional Neural Networks

Nesterov Accelerated Gradient (NAG, 1983)

Nesterov Accelerated Gradient (NAG, 1983)

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



(Top) Momentum method
(Bottom) Nesterov Accelerated Gradient
μ is the decaying parameter, same as η

And one more small, but important improvement of the NAG method is to take the value of the derivative of the loss function at a future (next) point, and not at the current one.

Temporary page!

LaTeX was unable to guess the total number of pages correctly. A
was some unprocessed data that should have been added to the
this extra page has been added to receive it.
If you rerun the document (without altering it) this surplus page
away, because LaTeX now knows how many pages to expect for t
document.