CONVOLUTIONAL NEURAL NETWORKS

Sergey Nikolenko

Harbour Space University

June 12, 2026

Random facts:

- on June 12, 1550, King Gustav I of Sweden founded Helsinki as a rival trading town to Reval (now Tallinn) across the gulf; it stayed a small fishing village for centuries and only became a real capital after Russia took Finland in 1809
- on June 12, 1817, Karl von Drais first rode his *Laufmaschine*, the ``dandy horse'' ancestor of the bicycle; he had been spurred on by a mass die-off of horses after the 1815 eruption of Mount Tambora caused the "Year Without a Summer"
- on June 12, 1898, General Emilio Aguinaldo proclaimed Philippine independence from Spain; the Philippines would not actually become independent until 1946, after Spanish, then American, then Japanese, then again American rule
- on June 12, 1939, the Baseball Hall of Fame opened in Cooperstown, New York, on the strength of a charming myth that Abner Doubleday had invented baseball there in 1839 (he had not)
- on June 12, 1979, Bryan Allen pedalled the *Gossamer Albatross* across the English Channel in 2 hours 49 minutes, the first human-powered flight across it, winning the second Kremer prize

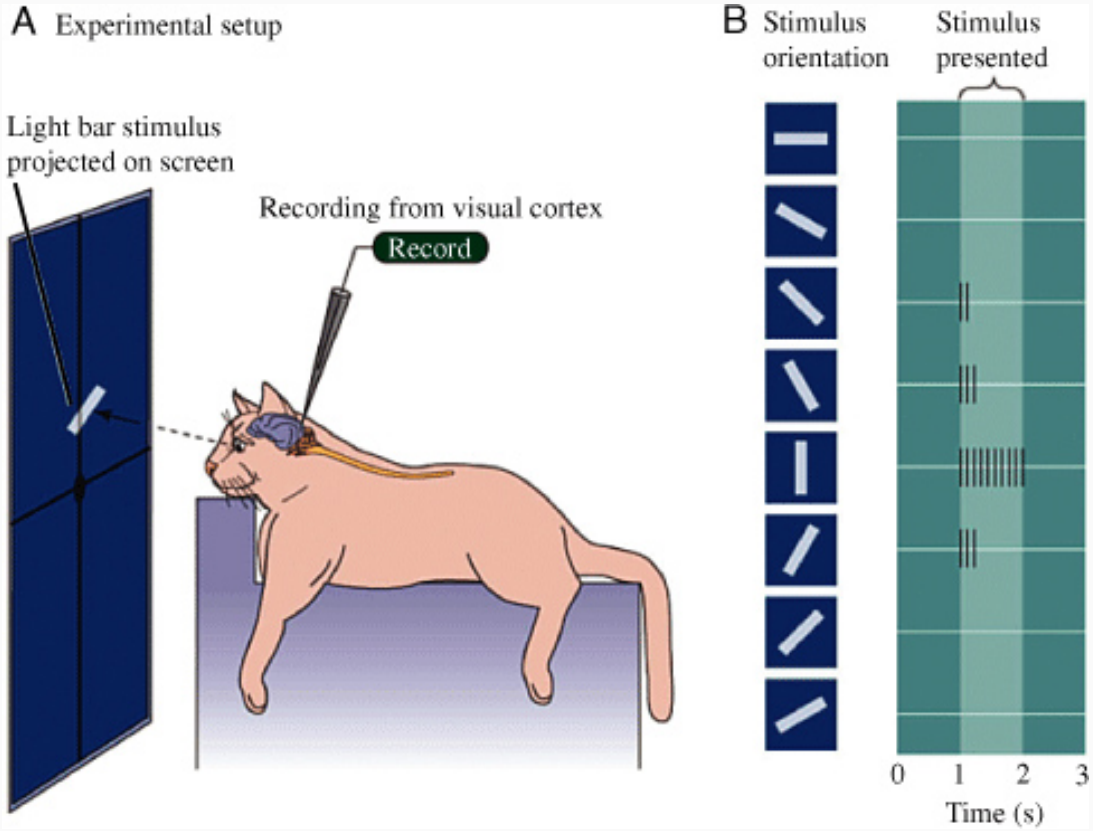
Convolutional neural networks

Visual cortex

- Convolutional neural networks – an important part of the deep learning revolution.
- CNNs are partly motivated by the visual cortex.
- Hubel and Wiesel: the famous studies of the visual cortex.

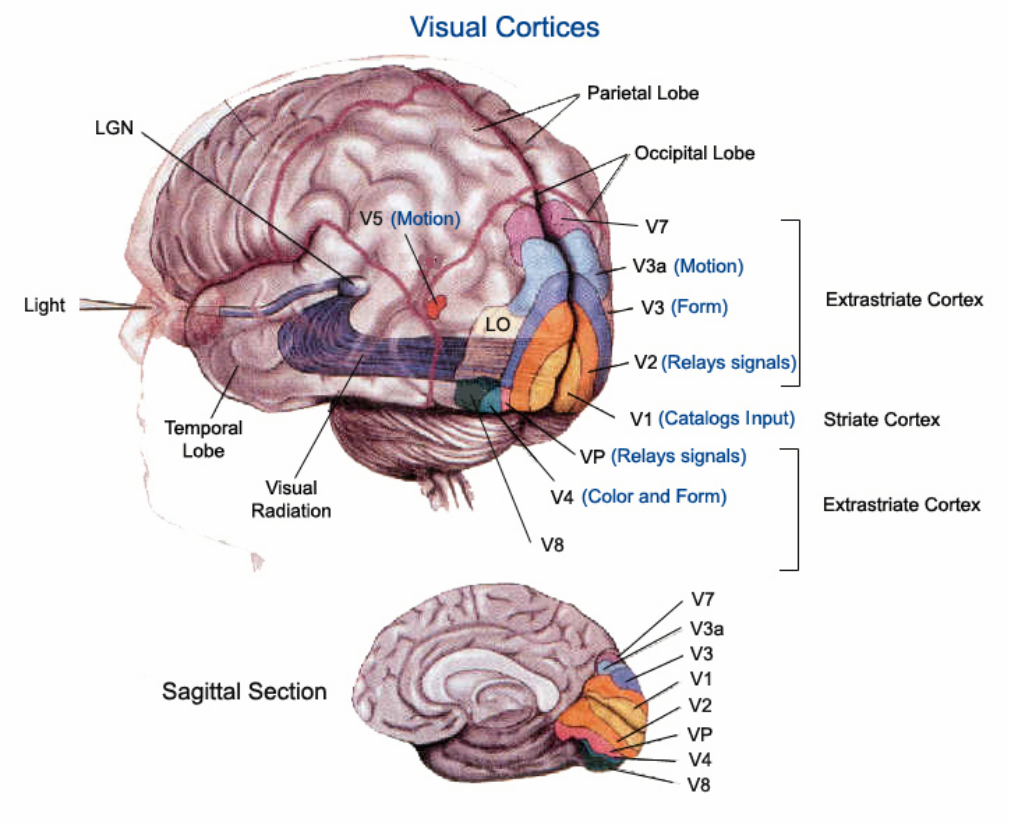
Visual cortex

- Hubel and Wiesel showed that first-level neurons activate for simple shapes, second-level neurons activate for combinations of these simple shapes, third-level… nope, they wisely didn't go there.



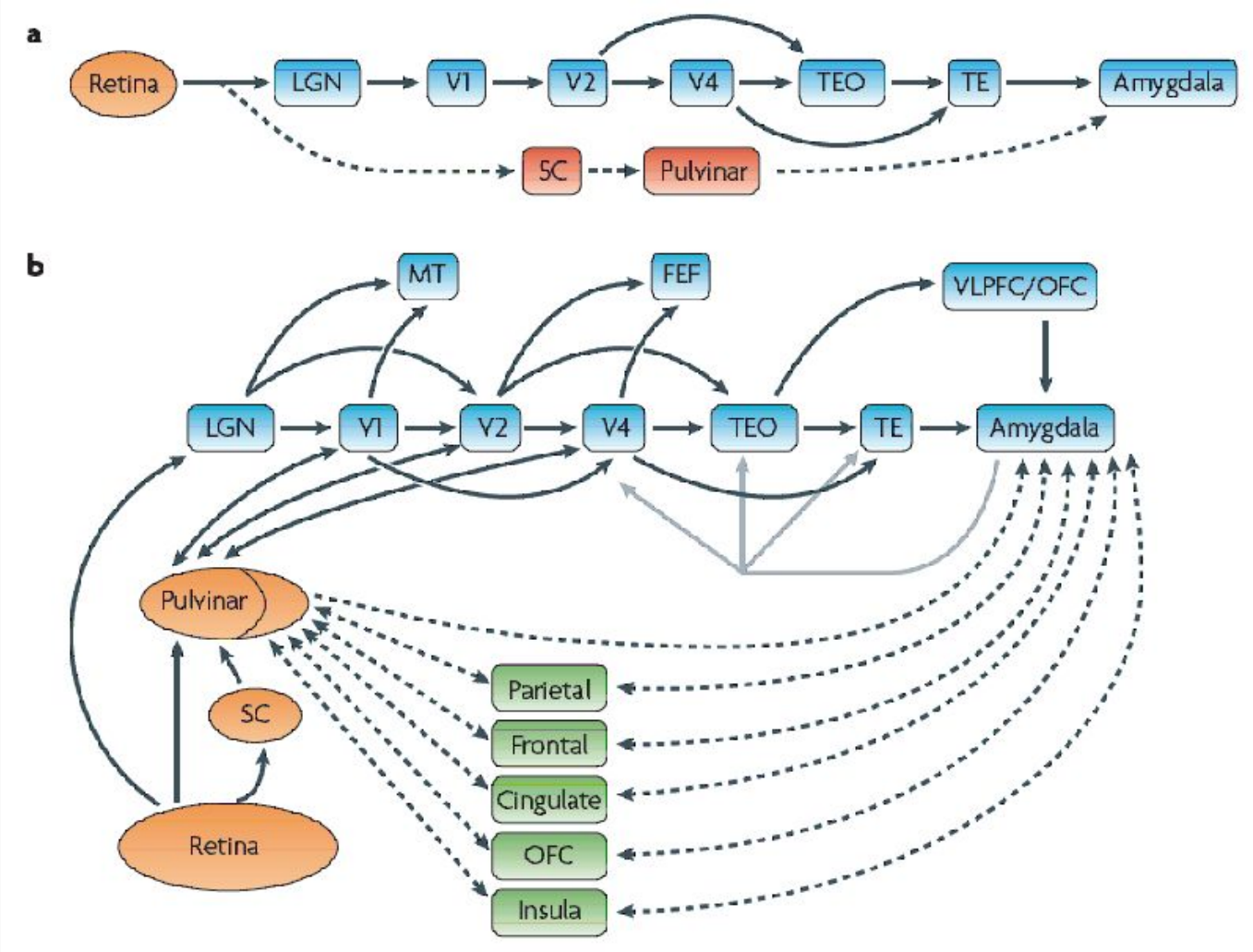
Visual cortex

- Information from the retina goes to the lateral geniculate nucleus (LGN), then to visual cortex.
- And it's neatly organized into layers, from V1 to V4 and then either to V5 or to V6.



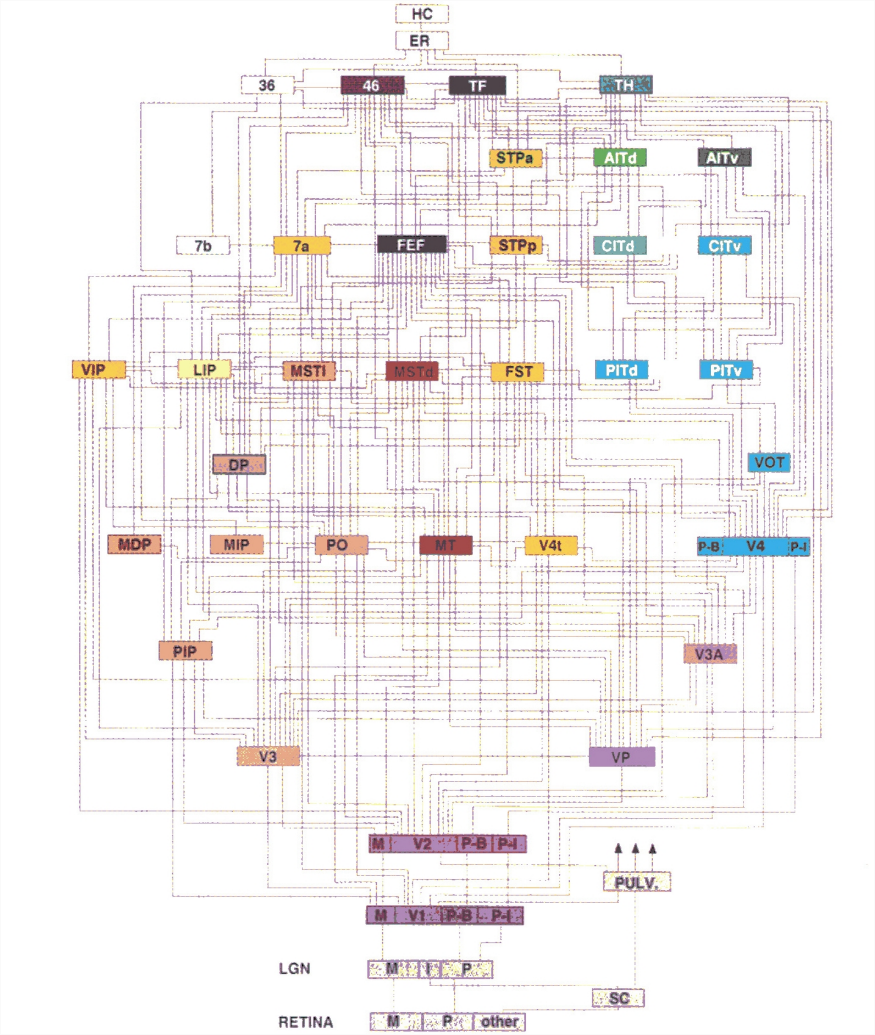
Visual cortex

- Well, kinda. The brain is complicated.



Visual cortex

- Very complicated.

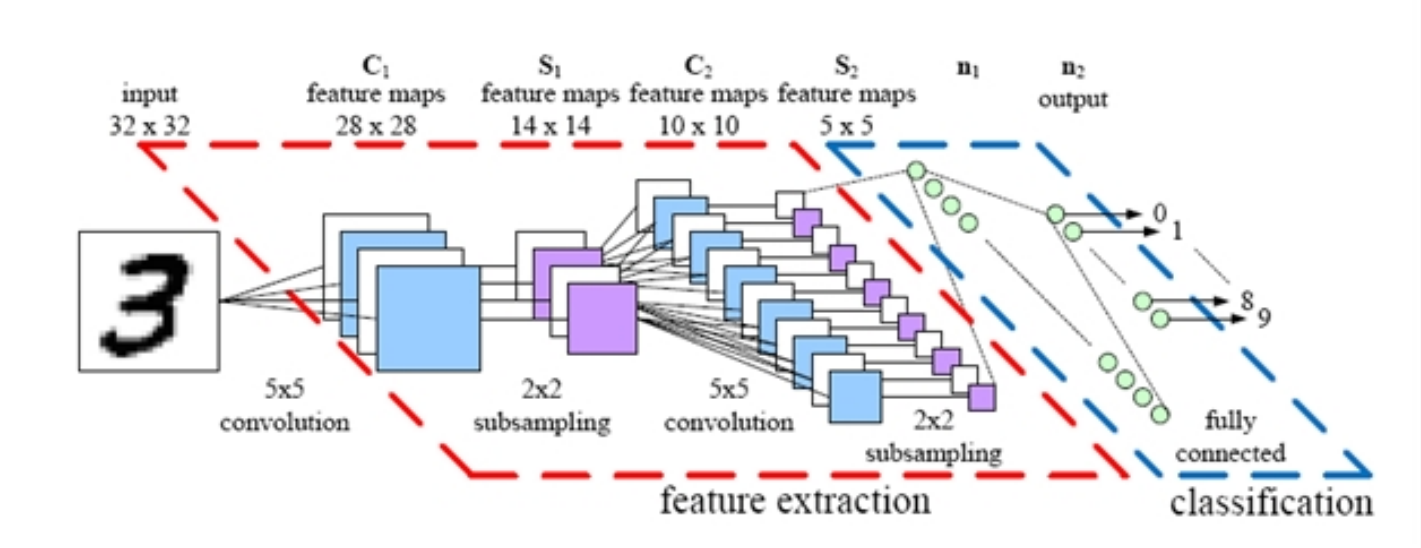


Visual cortex

- But some kind of layers are indeed there:
  - V1 – local features;
  - V2 – more local features, binocular vision;
  - V3 – color, texture, first results of segmentation;
  - V4 – geometric shapes, silhouettes; the most important layer for *attention*;
  - V5 – recognizes motion of objects segmented in V4;
  - V6 – generalizes data from the whole image; wide-field stimulation; processes changes due to our own movement;
  - V7 (controversial) – recognition of complex objects, e.g., human faces.
- What pathway: V2 to V4, recognition of shapes and objects.
- Where pathway: V2 to V5 and V6, motion recognition, control over eyes and hands.

Convolutional neural networks: idea

- These ideas were reflected in ANNs in the form of convolutional networks.
- Convolutions appeared in *Neocognitron* (Fukushima, 1979; 1980).
- In modern form – in LeCunn's group in late 1980s.
- Main idea: we would like to apply the same operation to every part of the image.

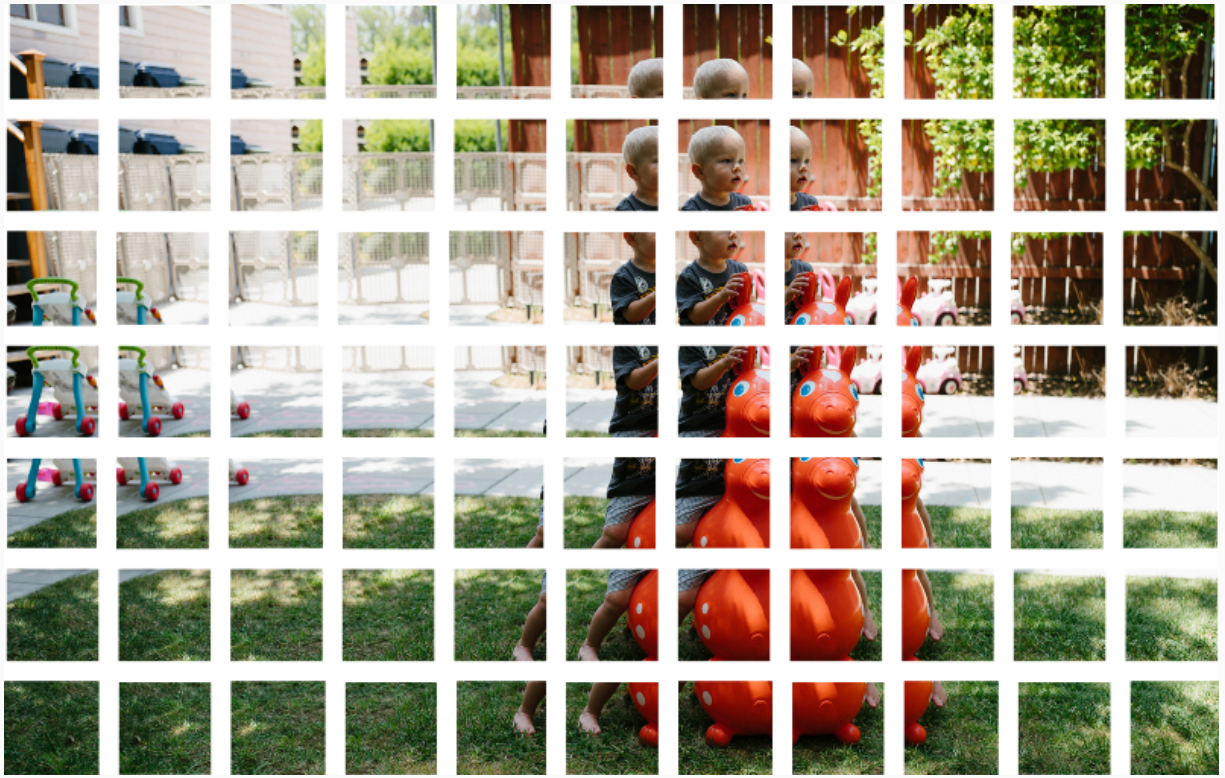


Convolutional neural networks: idea

- Break the picture into windows:

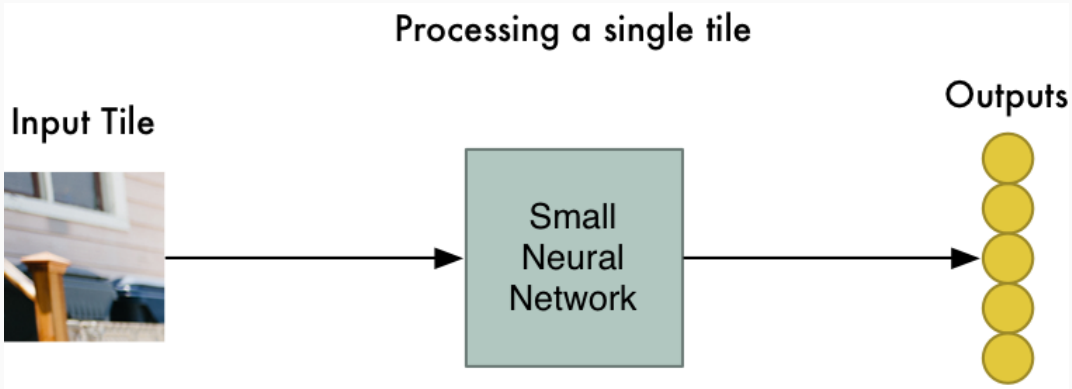
Convolutional neural networks: idea

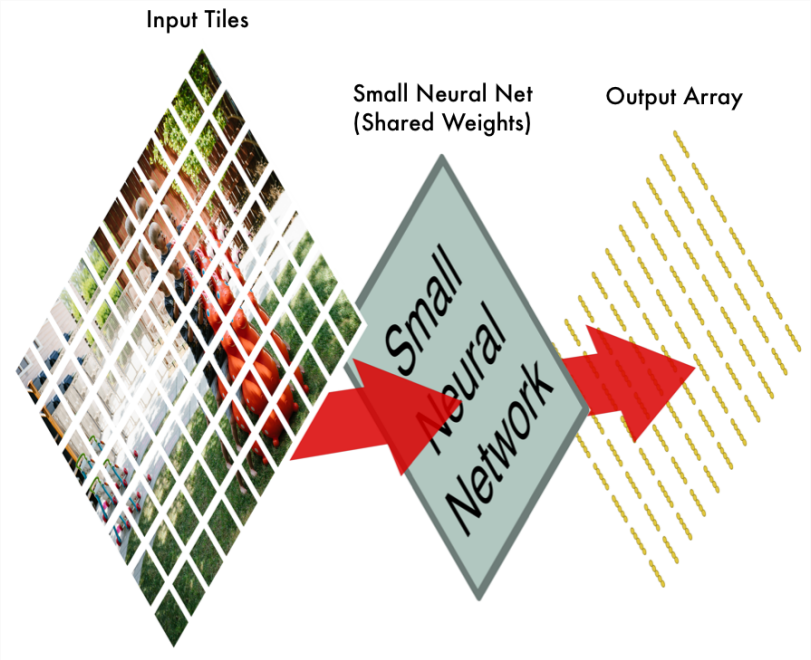
- Break the picture into windows:



Convolutional neural networks: idea

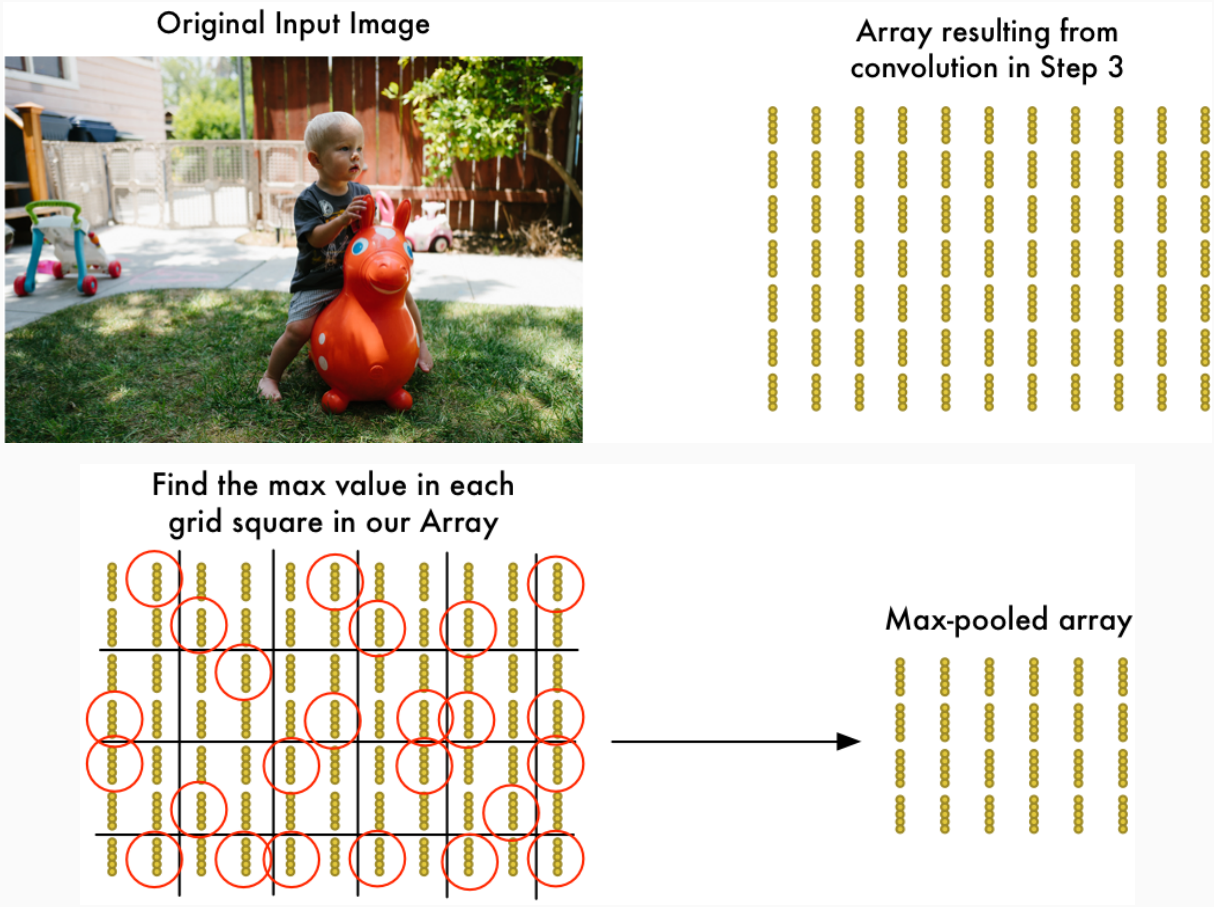
- Apply a small neural network to every window:





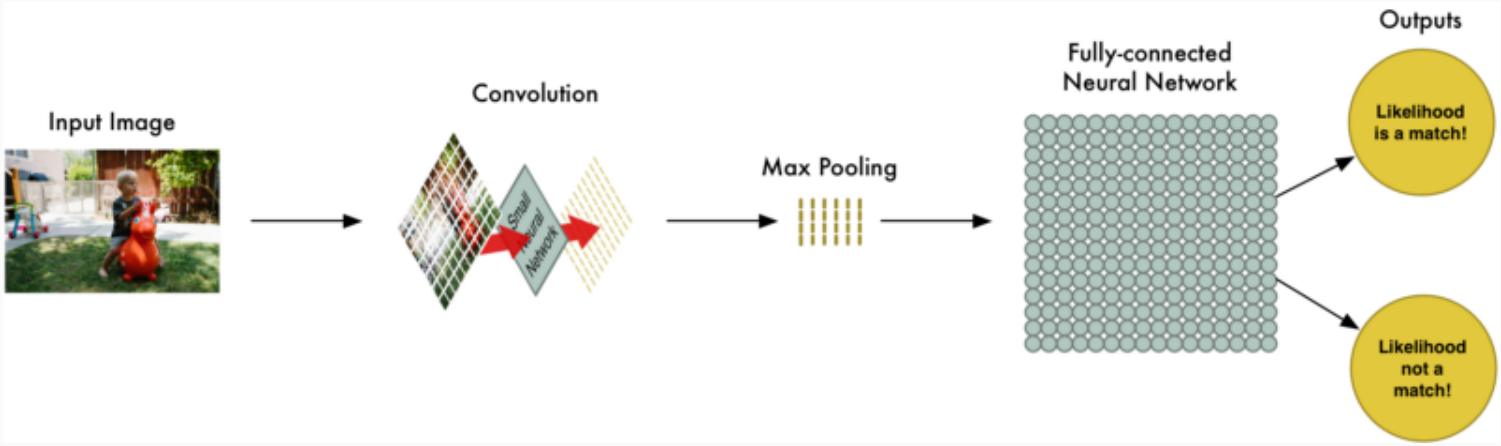
Convolutional neural networks: idea

- Then compress the image by max-pooling over small windows:

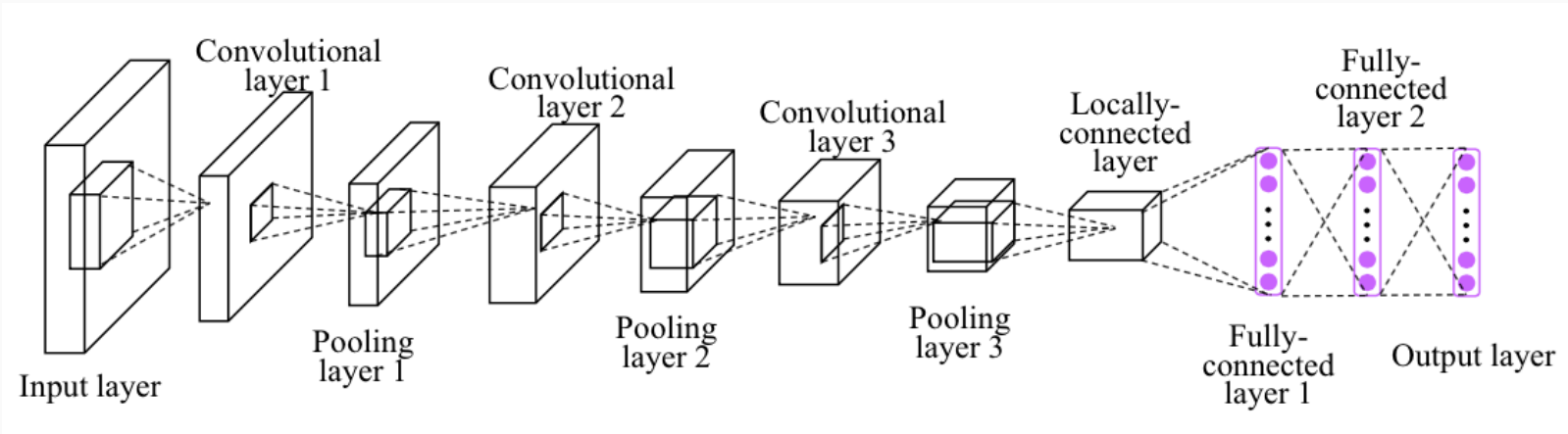


Convolutional neural networks: idea

- And then predict with a fully connected network:

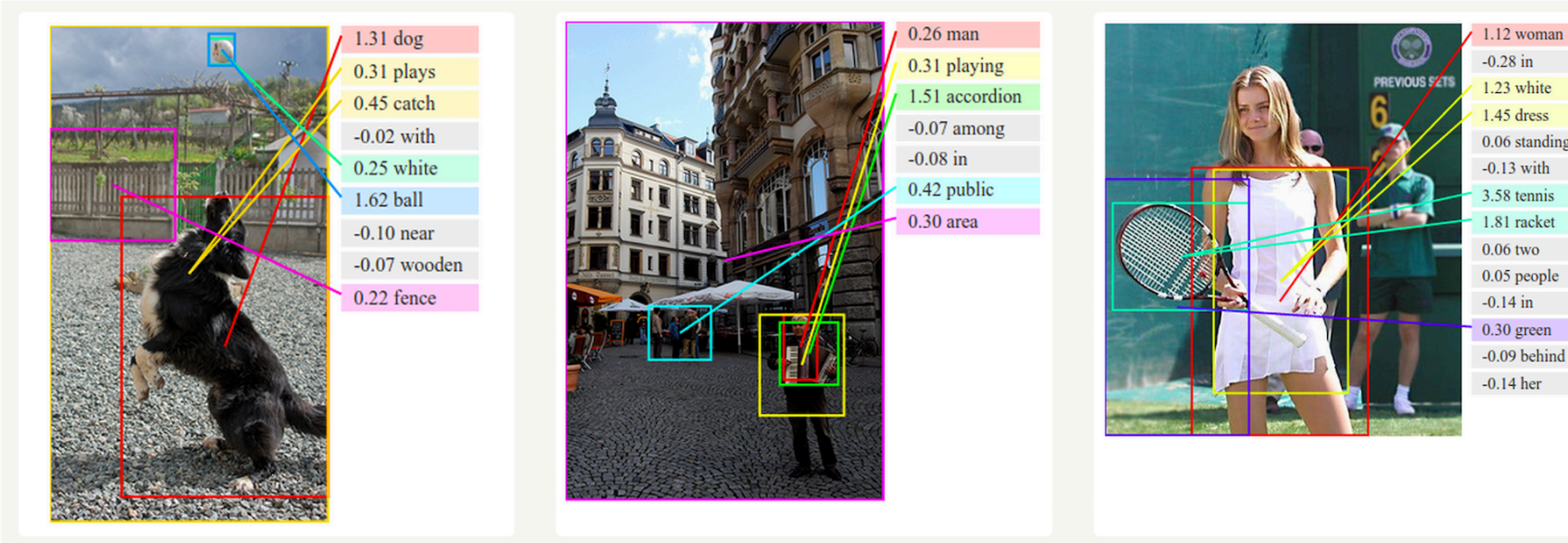


- Adding new layers improves generalization:



Convolutional neural networks: idea

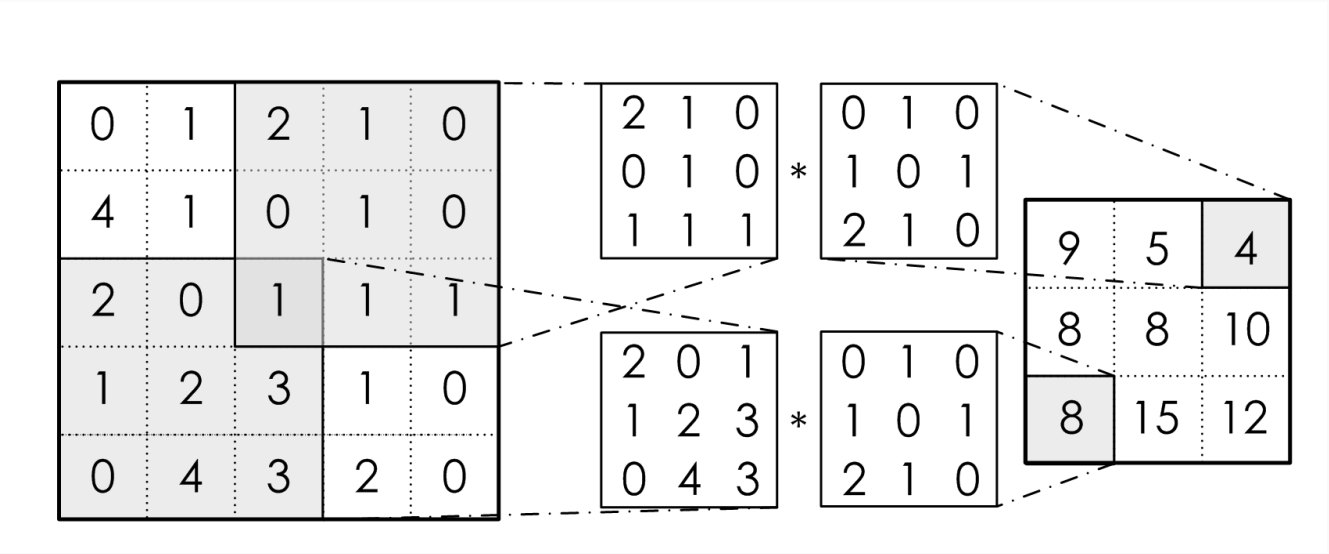
- Moreover, we can see where exactly are the parts of the picture that cause activation of individual neurons:



CNNs formally

- Formally speaking, we have a single small matrix and apply it to every window:
- If $\mathbf{x}^l$ is the feature map in layer l, then two-dimensional convolution of size $2d+1$ and weight matrix W of size $(2d+1) \times (2d+1)$ is

$$y^l_{i,j} = \sum_{-d<=a,b<=d} W_{a,b} x^l_{i+a,j+b}.$$



CNNs formally

- If $\mathbf{x}^l$ is the feature map in layer l, then two-dimensional convolution of size $2d+1$ and weight matrix W of size $(2d+1) \times (2d+1)$ is

$$y^l_{i,j} = \sum_{-d<=a,b<=d} W_{a,b} x^l_{i+a,j+b}.$$

- It is usually followed by a nonlinearity:

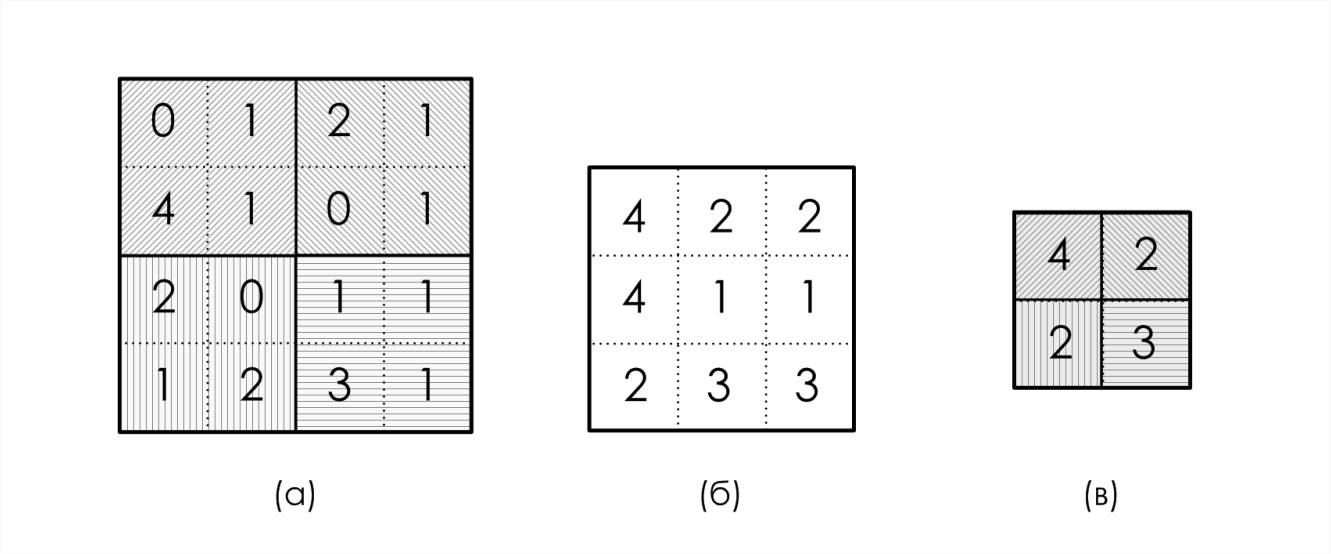
$$z^l_{i,j} = h(y^l_{i,j}).$$

- Almost always ReLU in modern networks.

CNNs formally

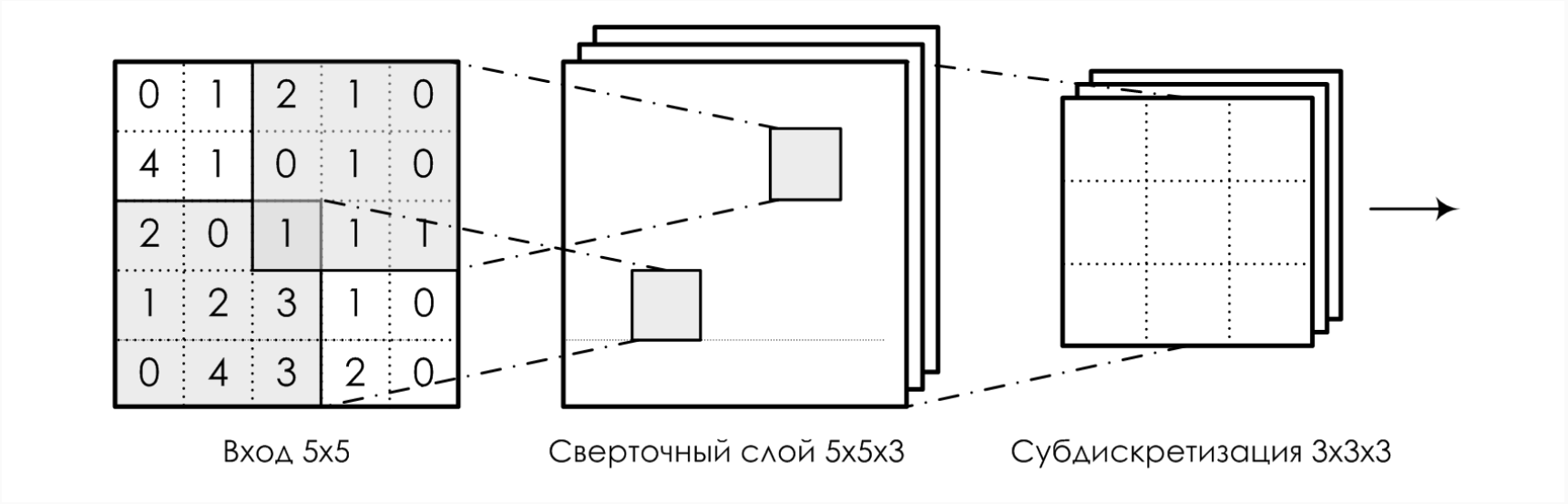
- And then max-pooling (sometimes average-pooling, but usually max):

$$x_{i,j}^{l+1} = \max_{-d<=a,b<=d} z_{i+a,j+b}^{l}.$$



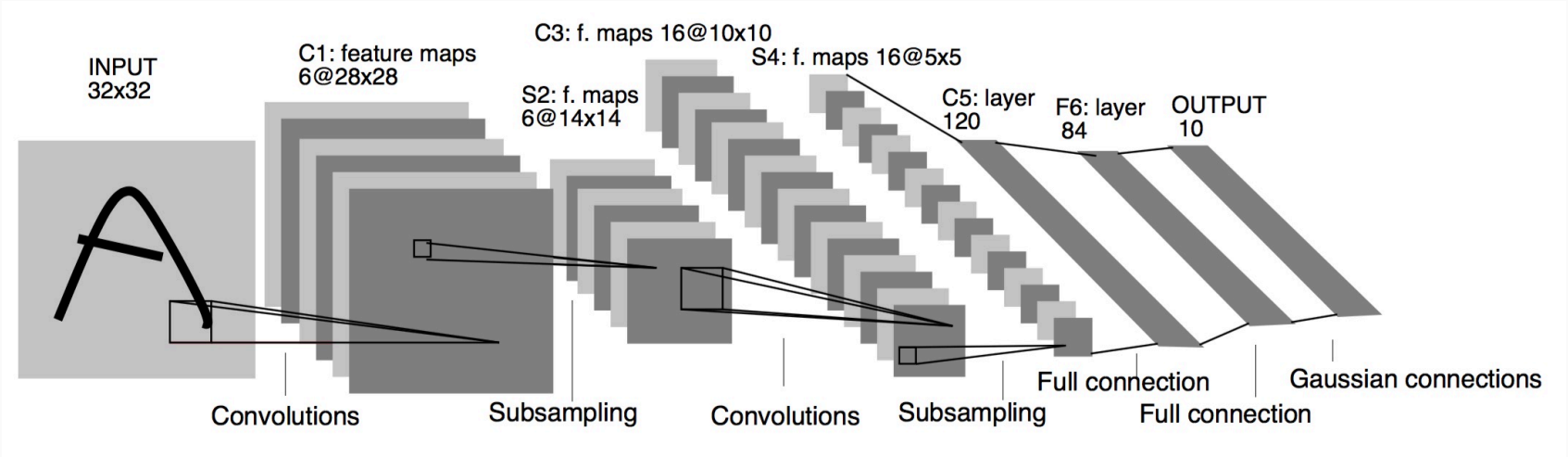
CNNs formally

- And this is the overall scheme of a single convolutional layer.



CNNs formally

- Classical *LeNet* architecture:



- Modern CNNs can get *very* deep due to a number of architectural tricks and advances.

Thank you for your attention!