Stochastic Machine Learning
Chapter 03 - Convolutional Neural Networks

Thorsten Schmidt

Abteilung für Mathematische Stochastik

www.stochastik.uni-freiburg.de
thorsten.schmidt@stochastik.uni-freiburg.de

SS 2024

The convolution

The convolution is the distribution function of the sum of two random variables,

$$Z = X + Y.$$

For example think of X and Y being discrete. To achieve a level of $Z = z$ with given $X = x$ we need

$$Y = z - x.$$

The associated probability of this is

$$P(Z = z) = \sum_x P(X = x) \cdot P(Y = z - x | X = x) = \sum_x P(X = x) \cdot P(Y = z - x),$$

due to independence. This defines the convolution.
We also have a similar formula when X and Y have densities,

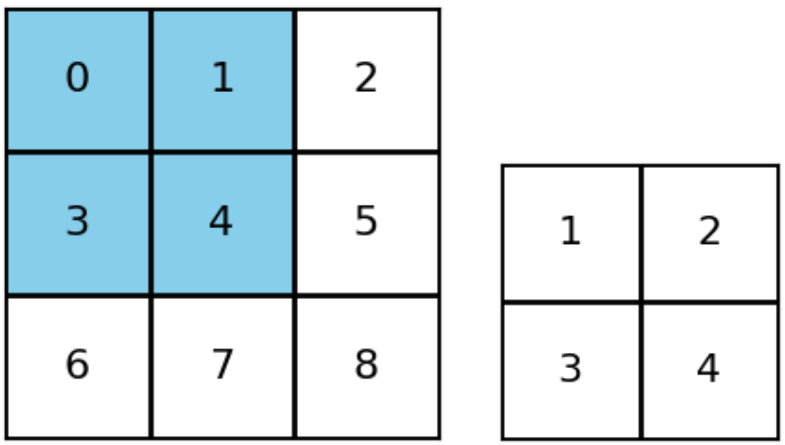
$$(f_X * f_X)(z) = \int f_X(x) \cdot f_Y(z - x)\, dx.$$

We recall that determining the distribution of the sum of independent random variables is easiest done with Fourier transforms, since

$$E[e^{uZ}] = E[e^{u(X+Y)}] = E[e^{uX}] \cdot E[e^{uY}].$$

Example

For image recognition, the convolution allows to extract local features. Let us do a simple computation.



We simplify the convolution to a kernel method if we re-arrange the matrix on the right hand side properly, but that is very fine for us !
So, we multiply the blue square with the white square and obtain the values

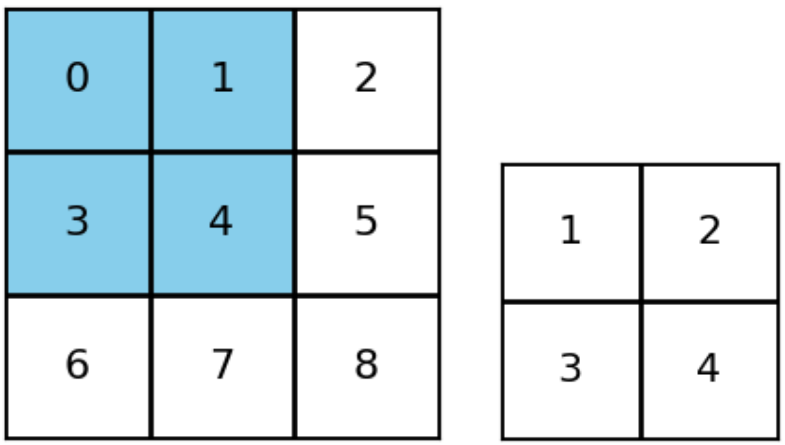
$$0 \cdot 1 + 1 \cdot 2 + 2 \cdot 3 + 4 \cdot 4 = 24.$$

If we move the blue square one to the right, we obtain

$$1 \cdot 1 + 2 \cdot 2 + 4 \cdot 3 + 5 \cdot 4 = 37.$$

Example

For image recognition, the convolution allows to extract local features. Let us do a simple computation.



We simplify the convolution to a kernel method if we re-arrange the matrix on the right hand side properly, but that is very fine for us !
and so on. This helps us to extract local features - like does is this pattern a line, a hat and so on.

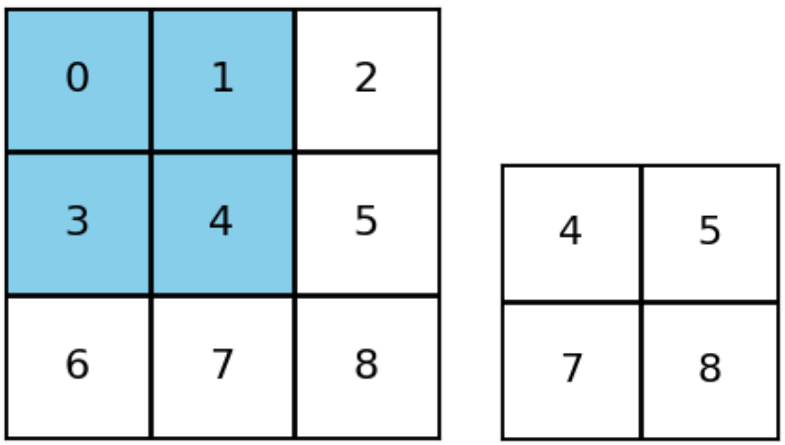
- Having for example $1/4$ in all cells gives the mean of the square - this blurrs the picture.
- If we have for example -1 and 1 then the result is zero if the pixels are equal but 1 if we have (0,1) and -1 if we have (1,0) - thus allows to detect edges.

Translation invariance

- Once we have learned a pattern, it is applied to all regions of the images. This implies that the pattern becomes **translation - invariant**, a very desirable property.
- Imagine how intensively you would have to train dense layers to capture this property
- In a second step, we can learn how to put simple, small patterns together - leading to a **spacial structure**. Like simple pictures are composed of lines, edges and circles this procedure can mimic this.
- We will meet further operations which play a role in image recognition.
- Often convolution is done with small window sizes, 3x3 or 5x5.

Max-pooling

In the max-pooling we simply take the maximum of the considered square



- This is a suitable operation to extract the most relevant features.
- Now we continue with chapter 8...