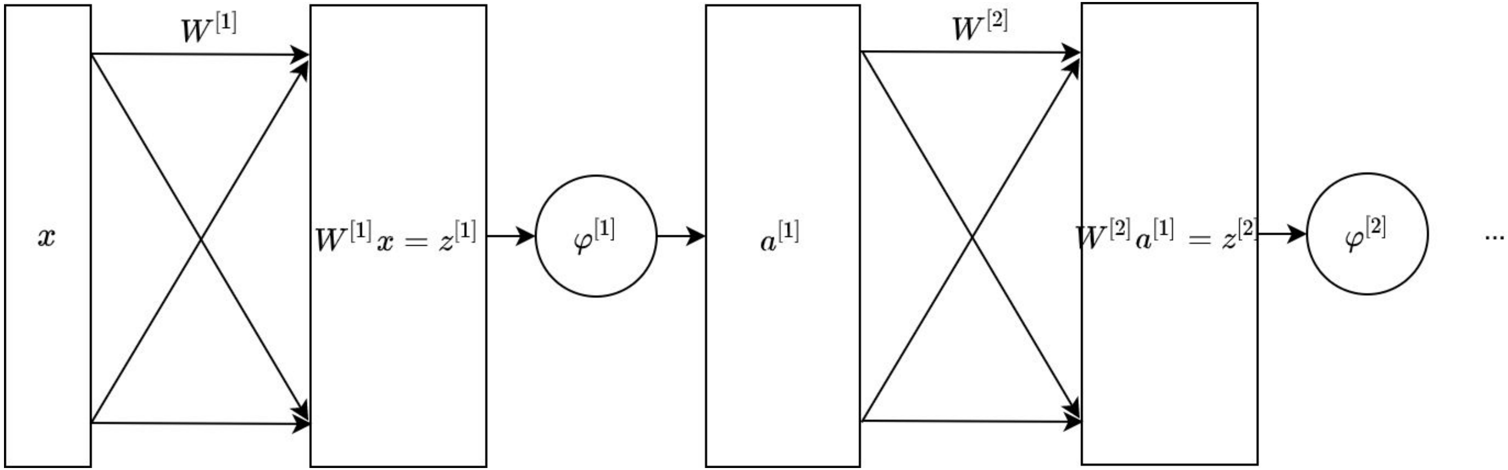
Deep Learning for ECE
EECE-580G

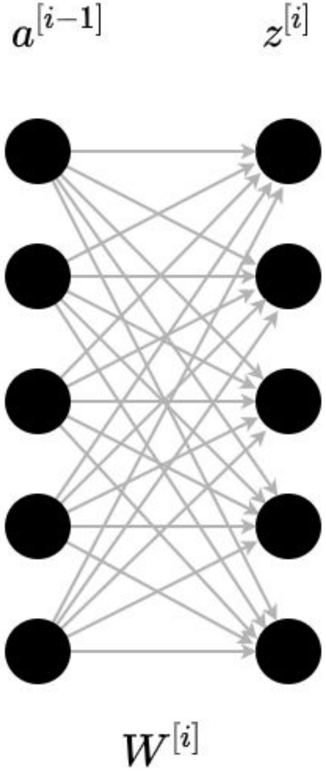
Convolutional Neural Networks (CNNs)

From MLP to CNN

Back to MLP...



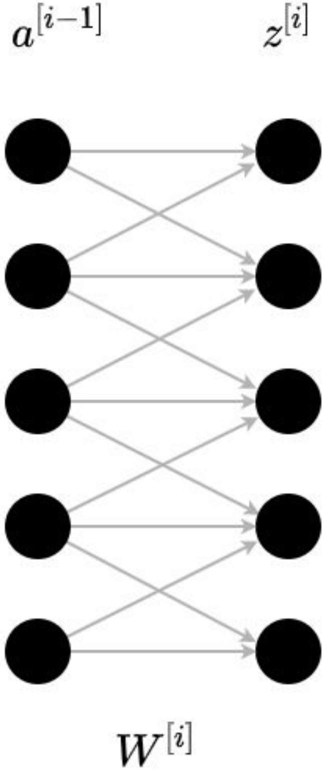
Back to MLP...



Locality

- Pixels in an image are only locally correlated (to a certain extent)
- Extends to different kinds of signals (speech, text, ...)
- Backed by biological experiments (Hubel and Wiesel's locally sensitive neurons experiments)

Locality = sparsity

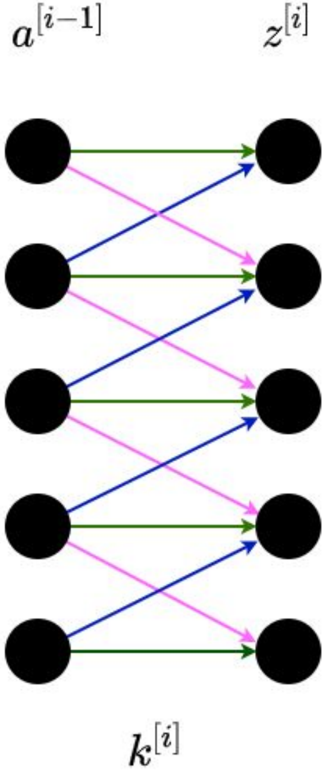


Stationarity

- Patterns are shared in different locations of an image
- e.g. a vertical edge is a vertical edge no matter where in the image
- Links to translation equivariance

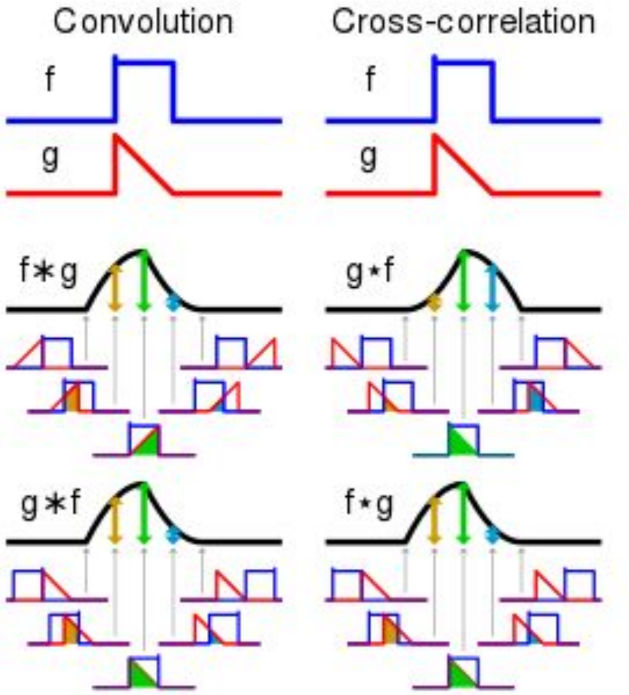
Credits: Third Eye Traveler

Stationarity = weight-sharing



Convolution

Convolution operation

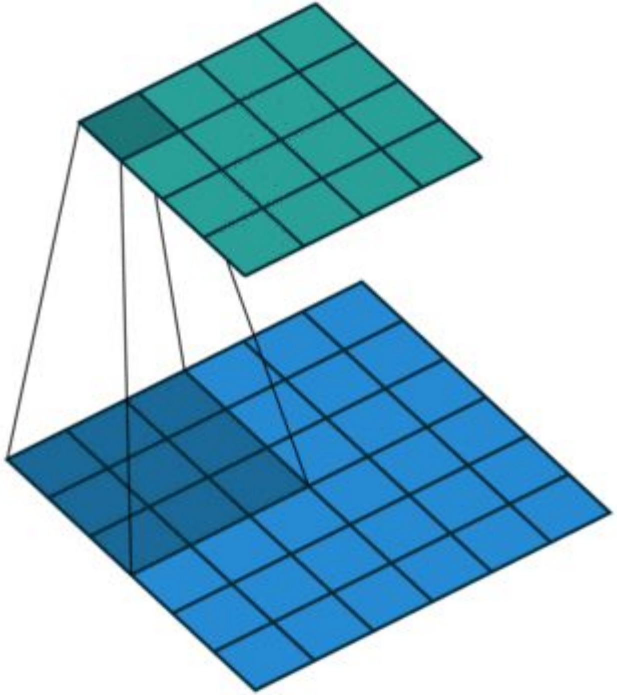


Source: wikipedia

Credits

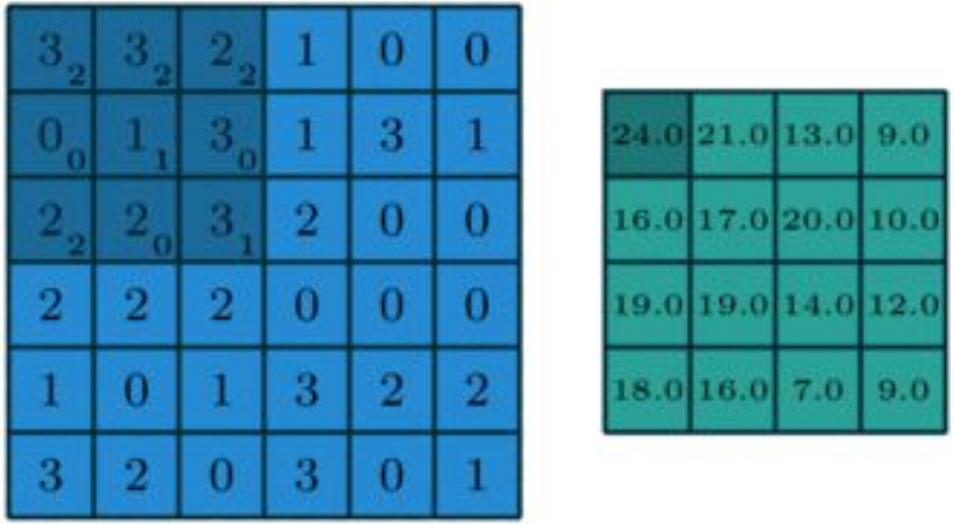
Most of the following slides will use figures created by Vincent Dumoulin and Francesco Visin. Referred to as V&F.

2d convolutions



Source: V&F

2d convolutions

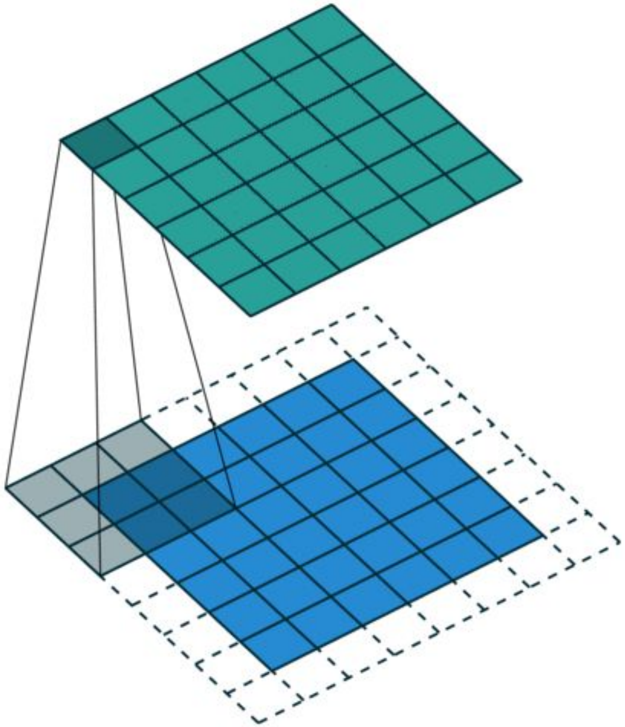


3_{2}	3_{2}	2_{2}	1	0	0
0_{0}	1_{1}	3_{0}	1	3	1
2_{2}	2_{0}	3_{1}	2	0	0
2	2	2	0	0	0
1	0	1	3	2	2
3	2	0	3	0	1

24.0	21.0	13.0	9.0
16.0	17.0	20.0	10.0
19.0	19.0	14.0	12.0
18.0	16.0	7.0	9.0

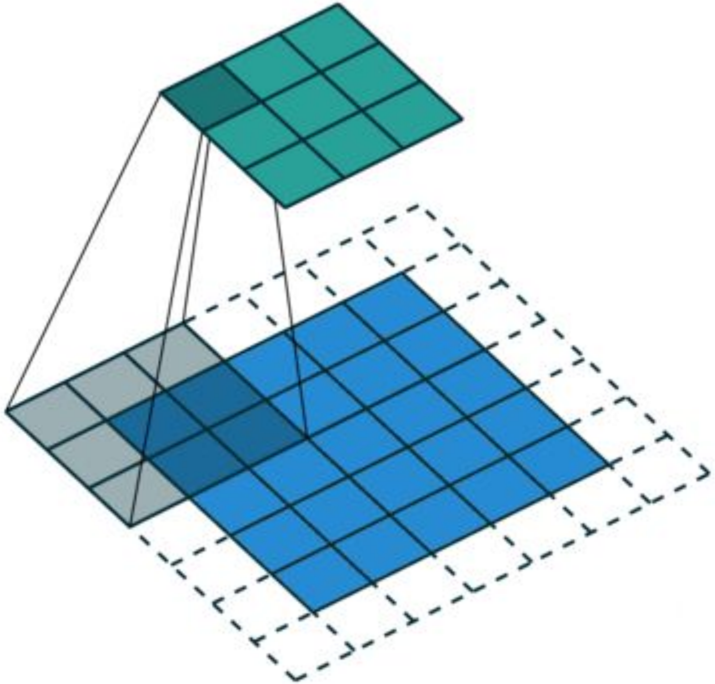
Source: V&F

2d convolutions - padding



Source: V&F

2d convolutions - strides



Source: V&F

Output shape - input shape

$$O = \lfloor I - K + 2P \rfloor / S + 1$$

O = Output size

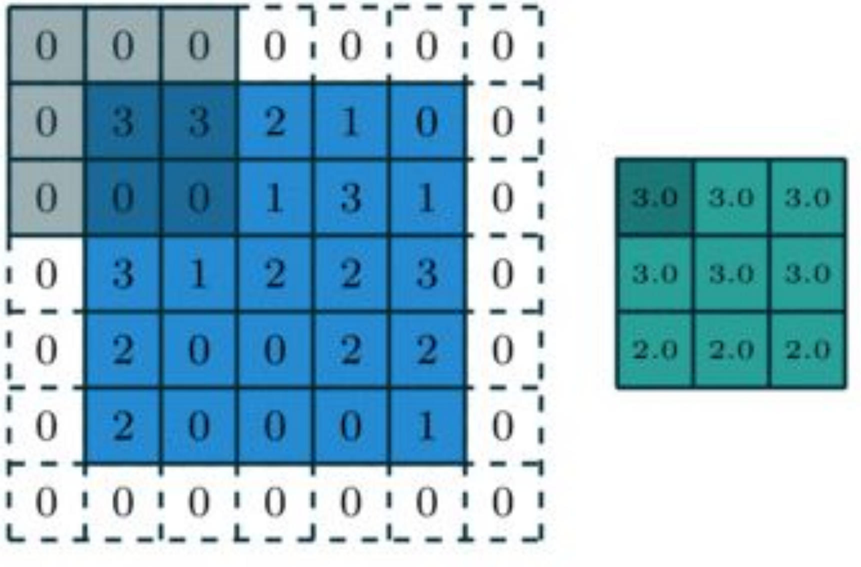
I = Input size

K = kernel size

P = Padding

S = strides

Pooling - max

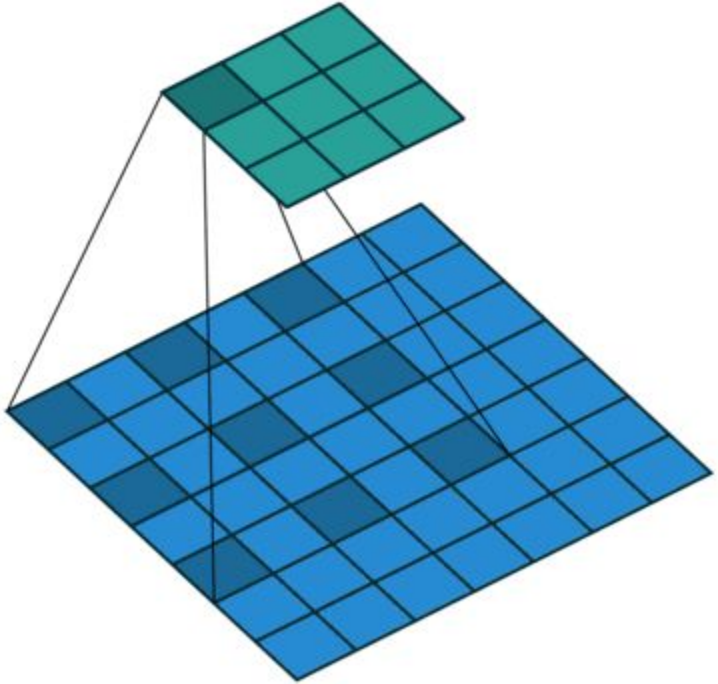


0	0	0	0	0	0	0
0	3	3	2	1	0	0
0	0	0	1	3	1	0
0	3	1	2	2	3	0
0	2	0	0	2	2	0
0	2	0	0	0	1	0
0	0	0	0	0	0	0

3.0	3.0	3.0
3.0	3.0	3.0
2.0	2.0	2.0

Source: V&F

2d convolutions - a-trous / dilation



Source: V&F

Convolutions in image processing

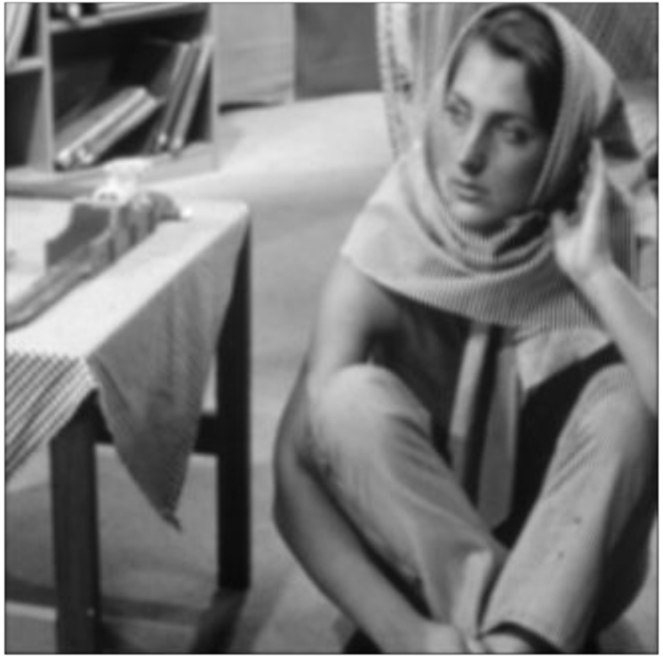
Box blur

$$k = \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

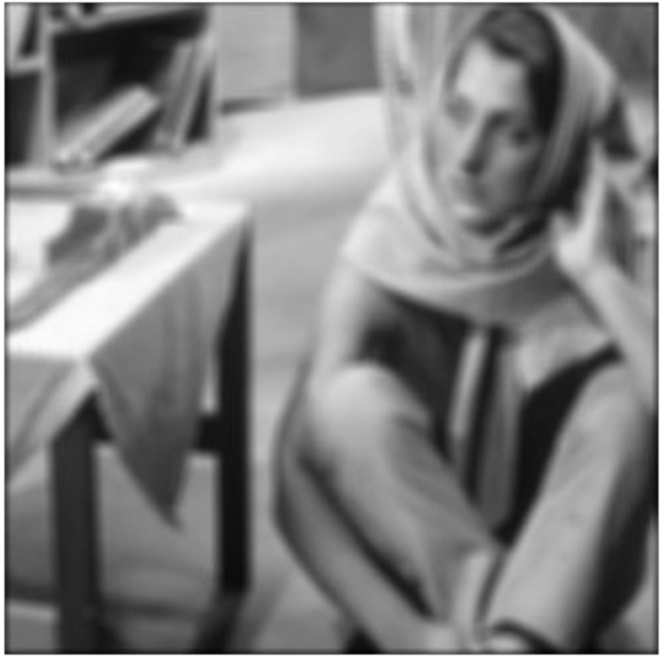
>>> Colab

Original

Box blur - 5x5

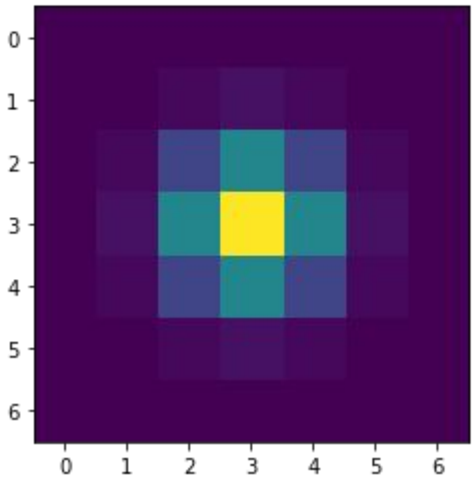


Box blur - 11x11



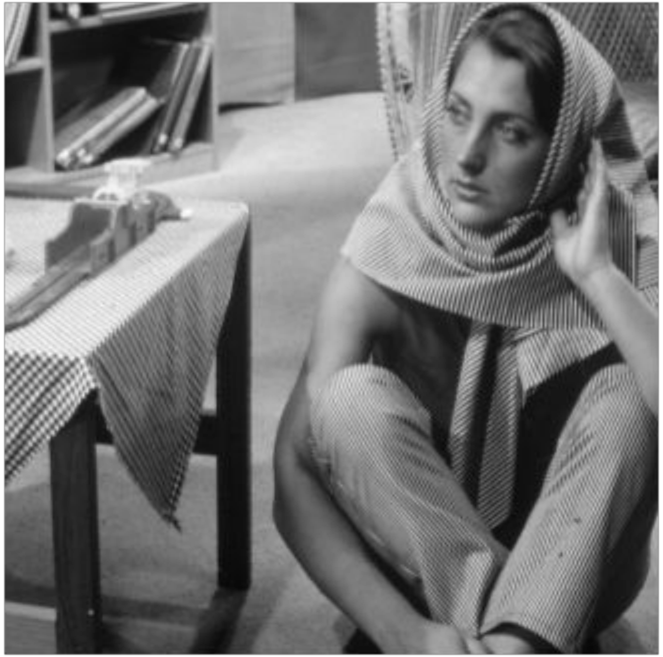
Gaussian blur

$$k(i, j) = \frac{1}{2\pi\sigma^2} e^{-\frac{i^2+j^2}{2\sigma^2}}$$

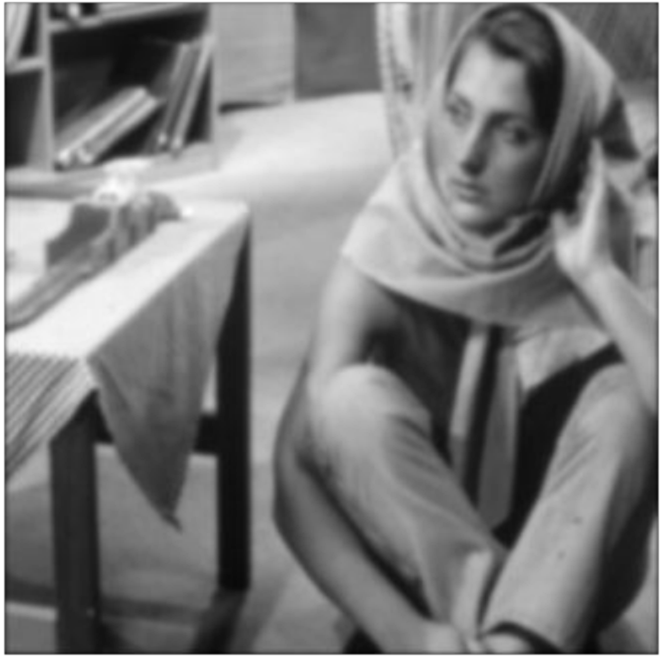


>>> Colab

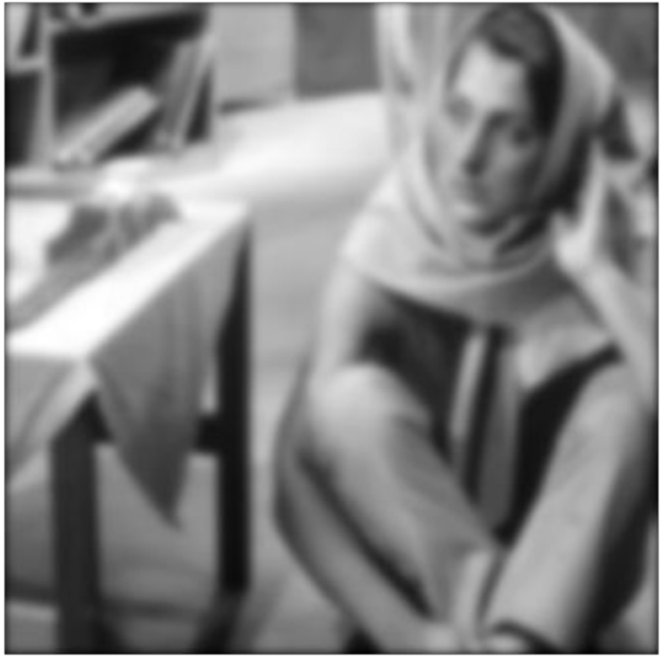
Original



Gaussian blur $\sigma = 2$



Gaussian blur $\sigma = 4$



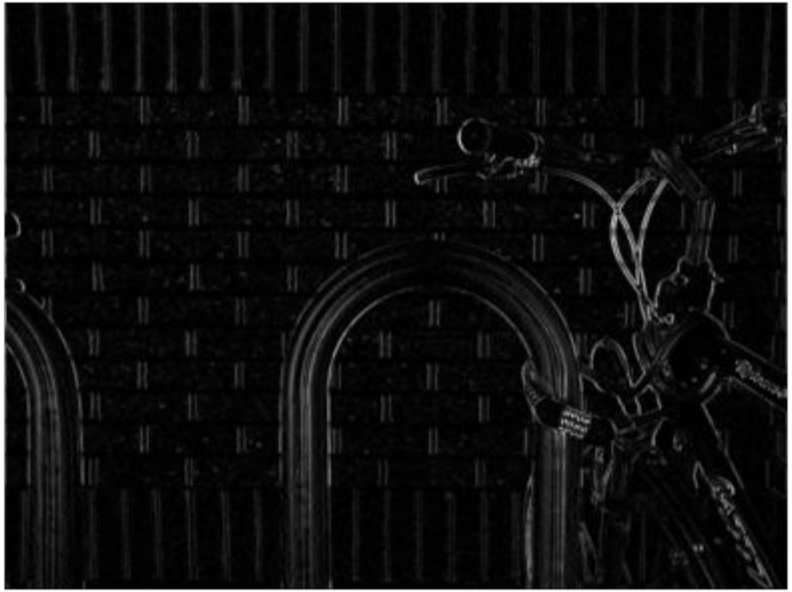
Edge detectors - Sobel operator

$$k^{[1]} = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} \quad k^{[2]} = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

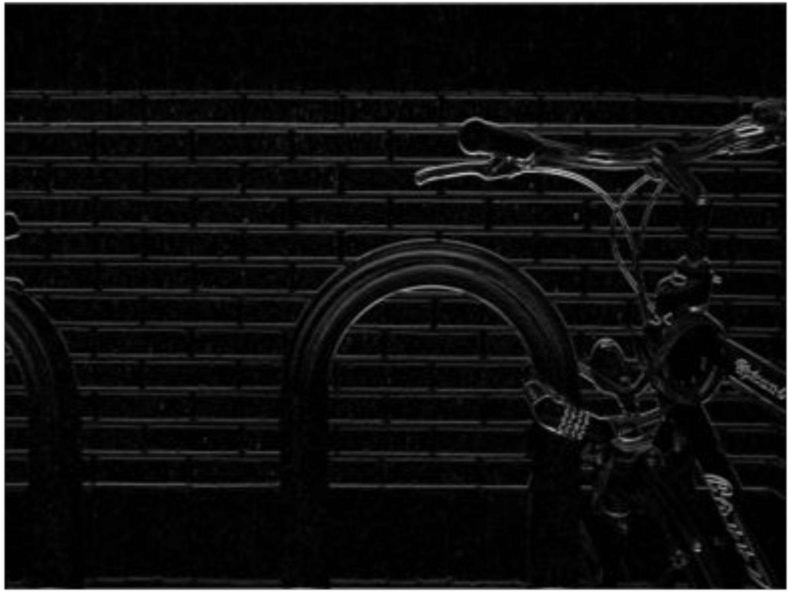
>>> Colab

Edge detectors - Sobel operator

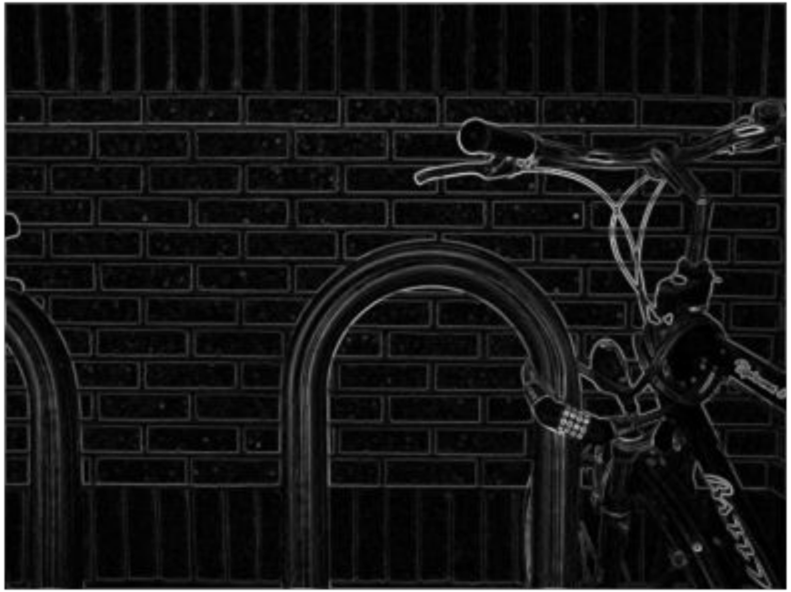
Edge detectors - Sobel operator - $k^{[1]}$



Edge detectors - Sobel operator - $k^{[2]}$

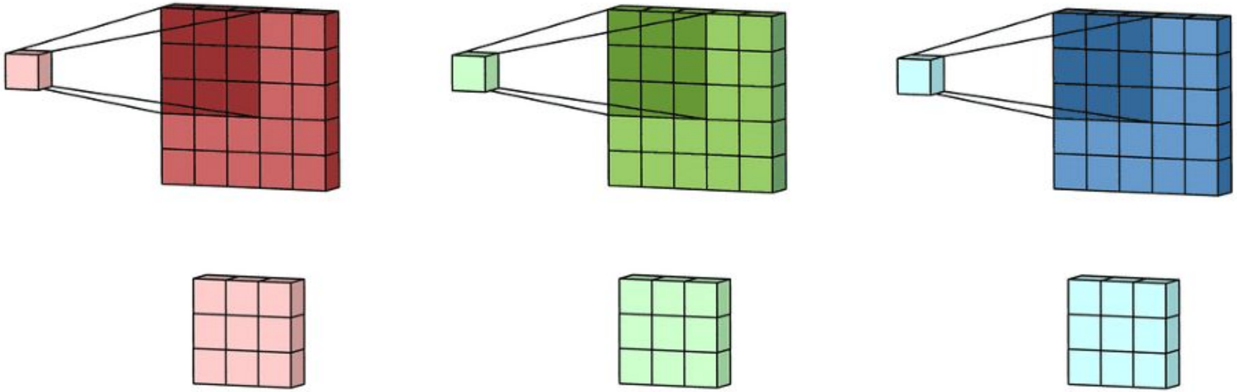


Edge detectors - Sobel operator

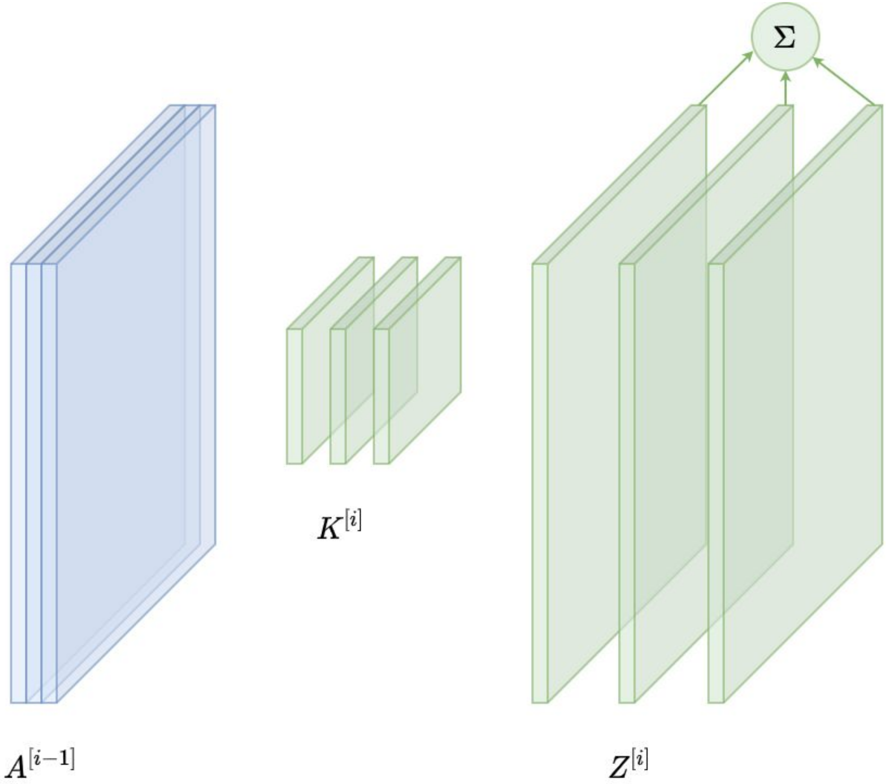


Stacking convolutions

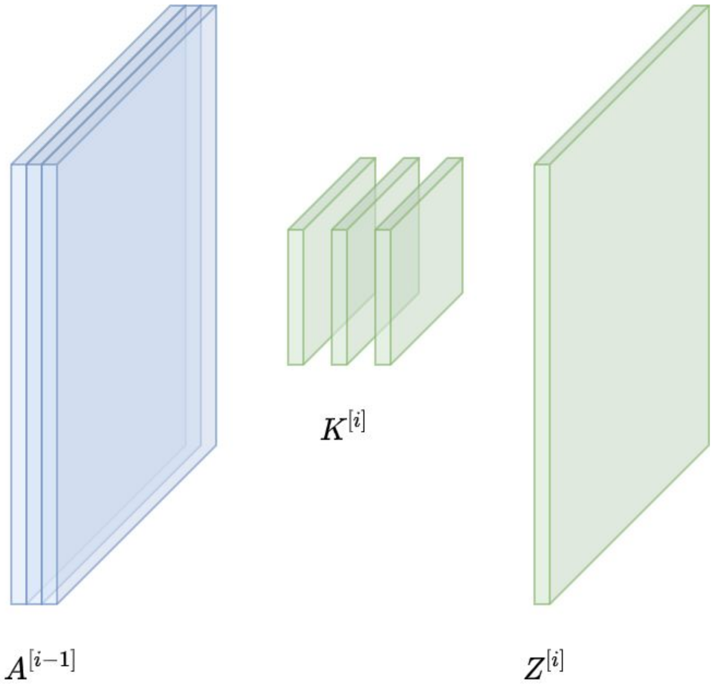
How to deal with color images?



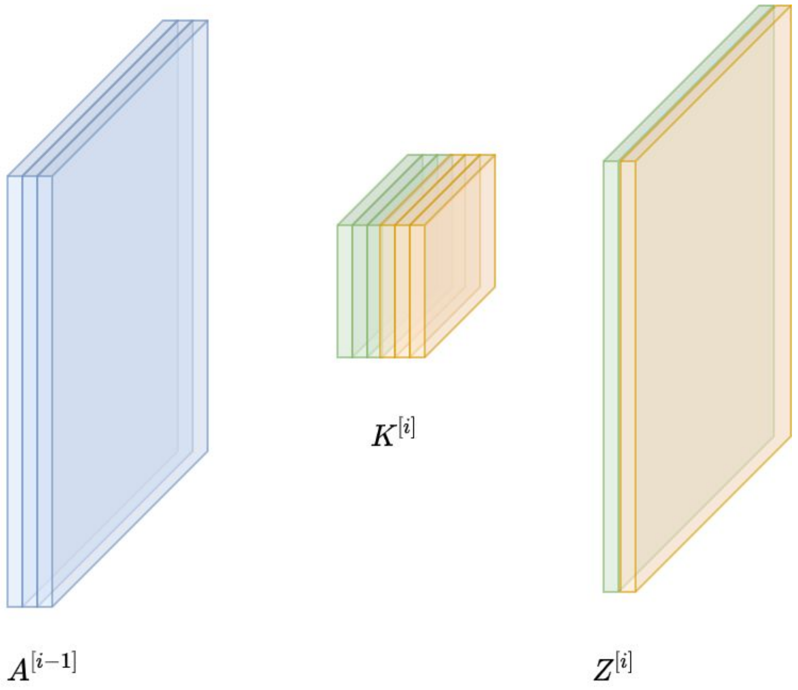
How to deal with color images?



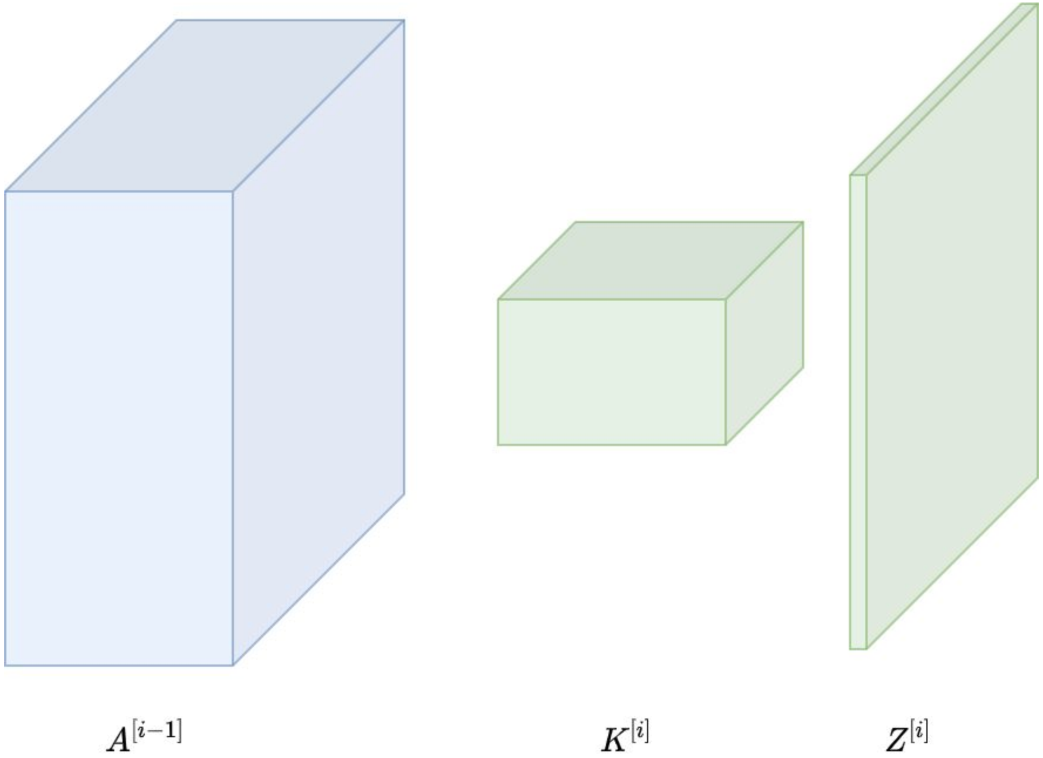
How to deal with color images?



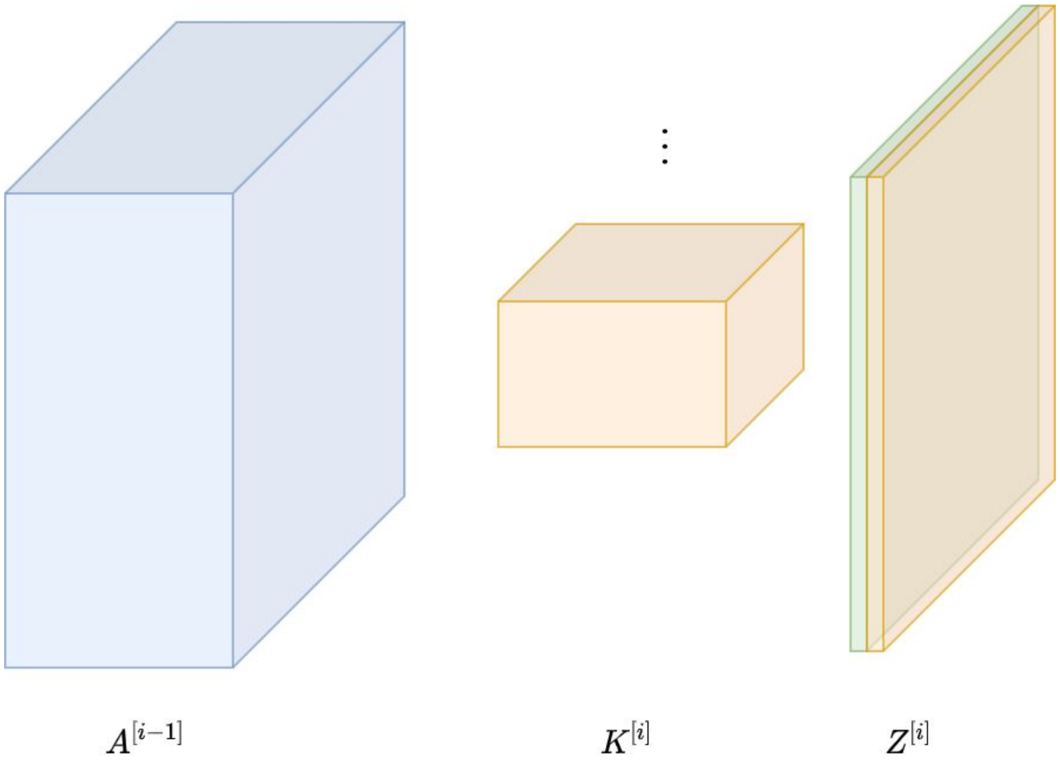
More filters



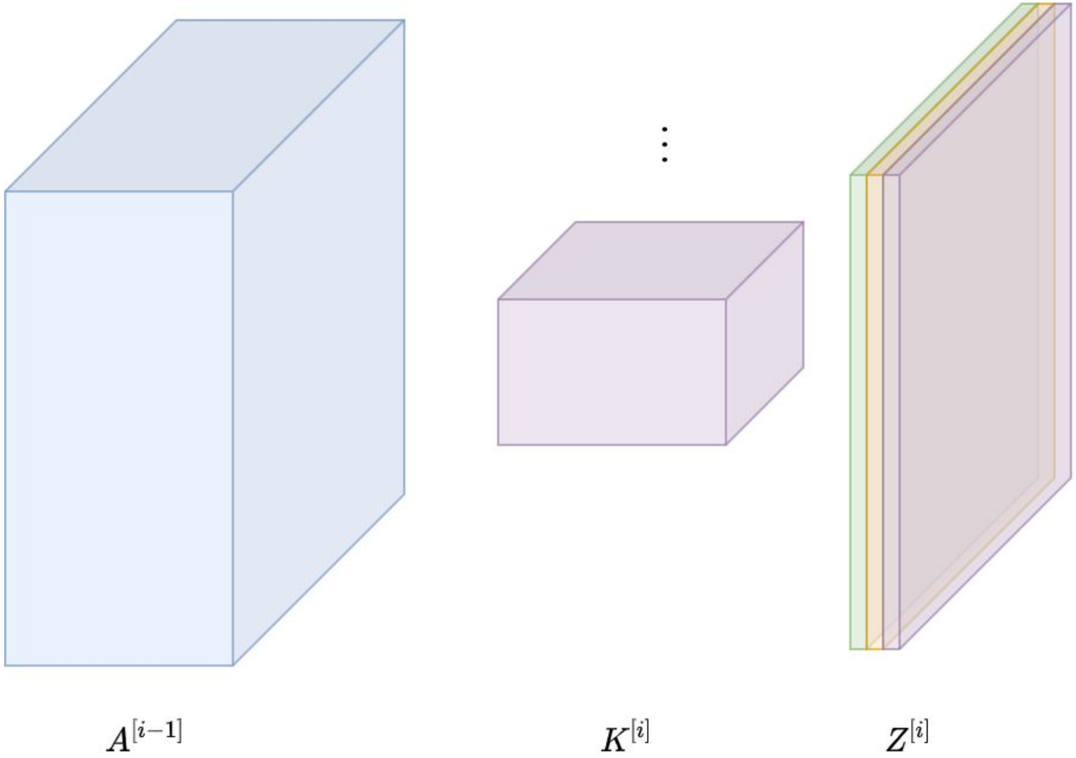
Images as volumes



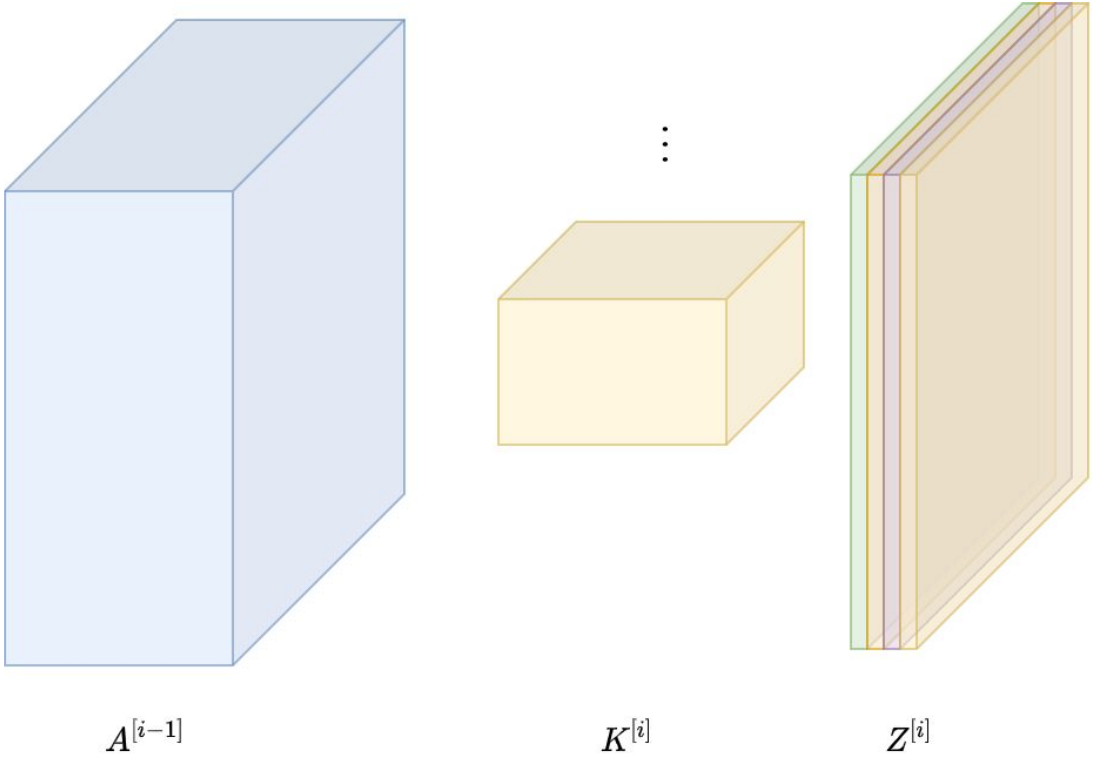
Images as volumes



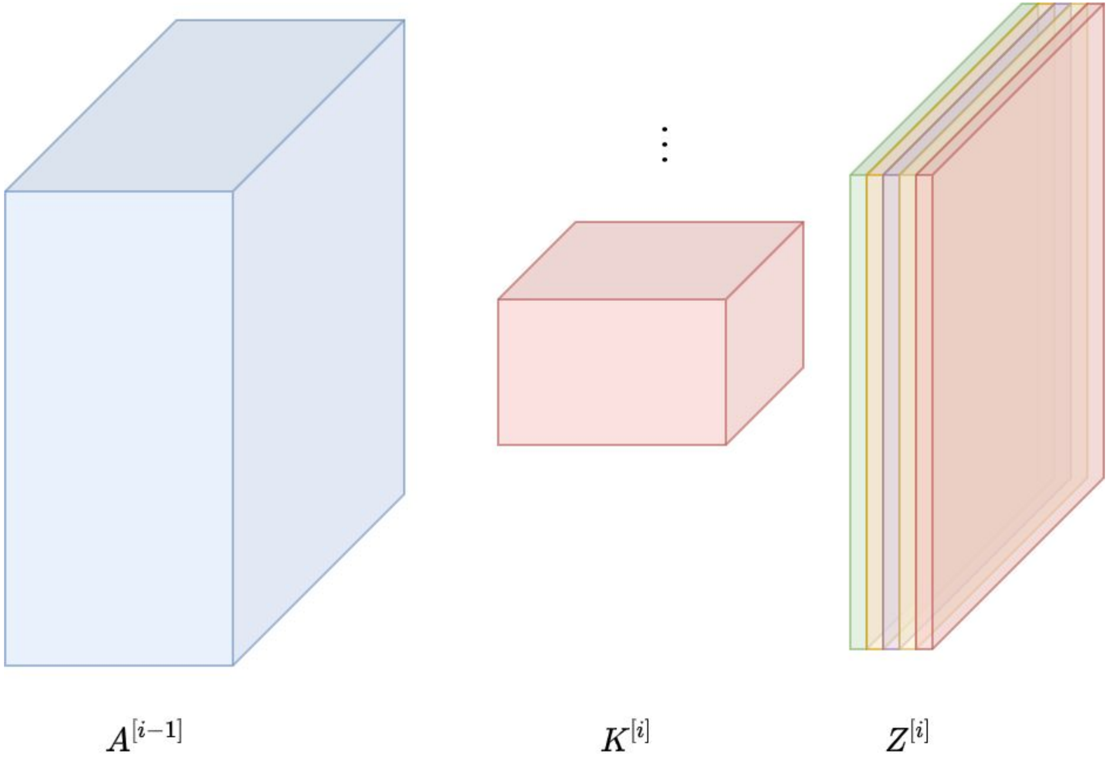
Images as volumes



Images as volumes



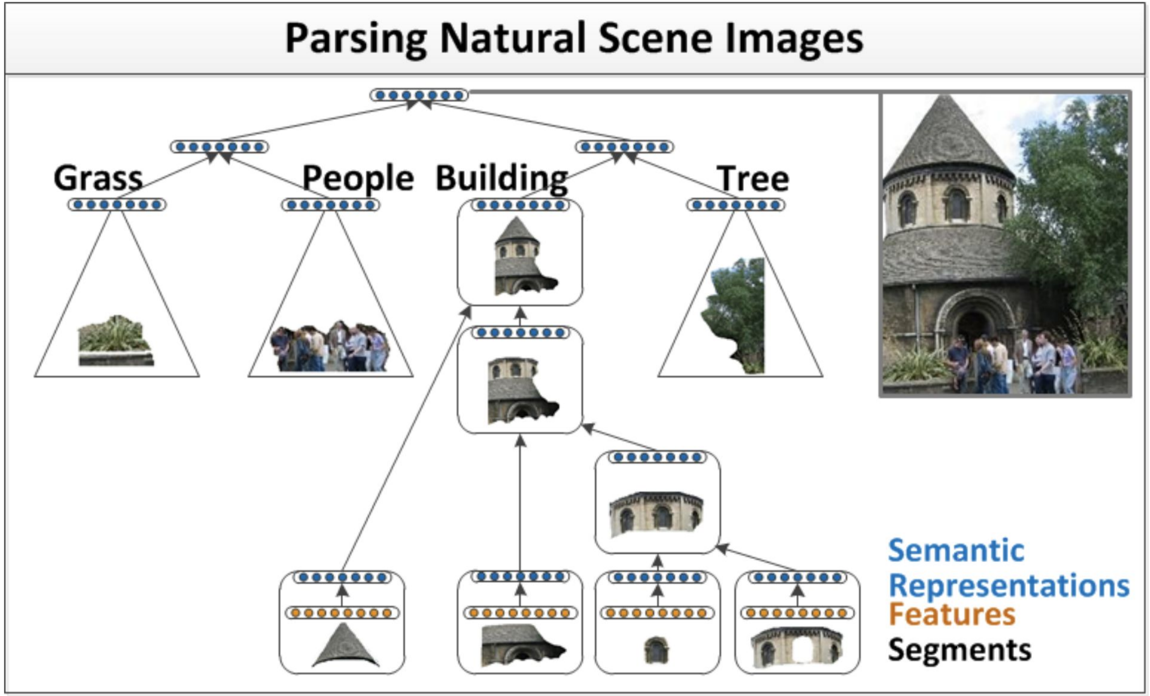
Images as volumes



The conv kernel

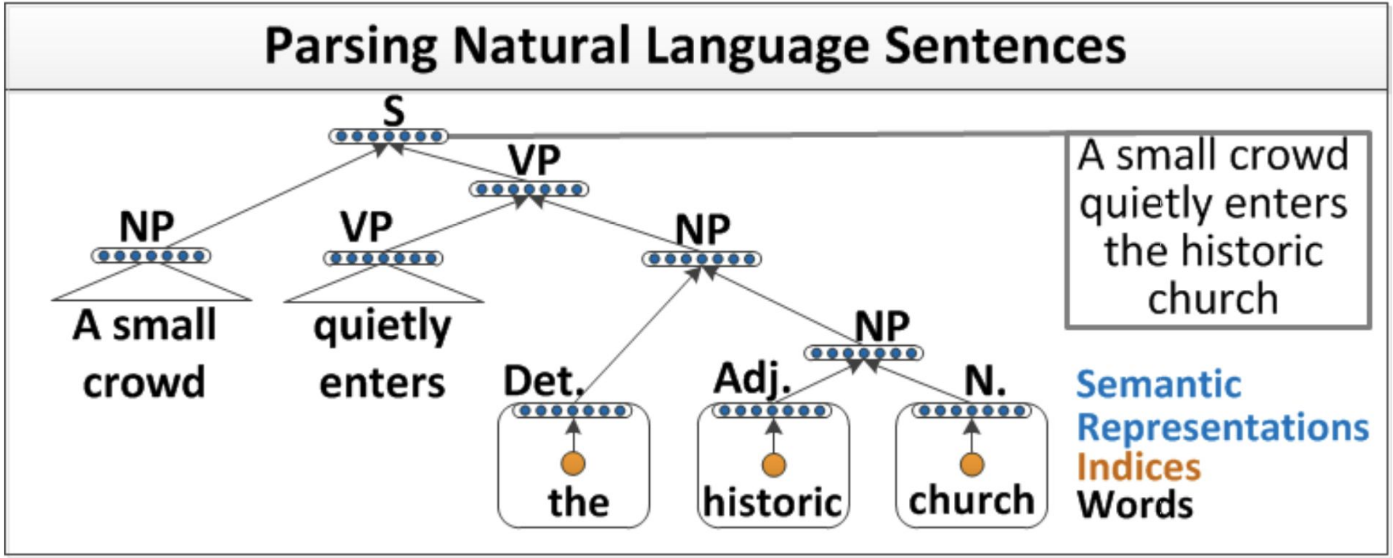
$$\mathrm{shape}(k^{[i]}) = K^2 \times \mathrm{num\ channels}^{[i-1]} \times \mathrm{num\ filters}$$

Compositionality



Source: Socher, Richard, et al. "Parsing natural scenes and natural language with recursive neural networks." Proceedings of the 28th international conference on machine learning (ICML-11). 2011.

Compositionality



Source: Socher, Richard, et al. "Parsing natural scenes and natural language with recursive neural networks." Proceedings of the 28th international conference on machine learning (ICML-11). 2011.

Compositionality = hierarchical representation

- Stacking multiple conv layers
- A basic conv layer = **conv - (BN) - activation**

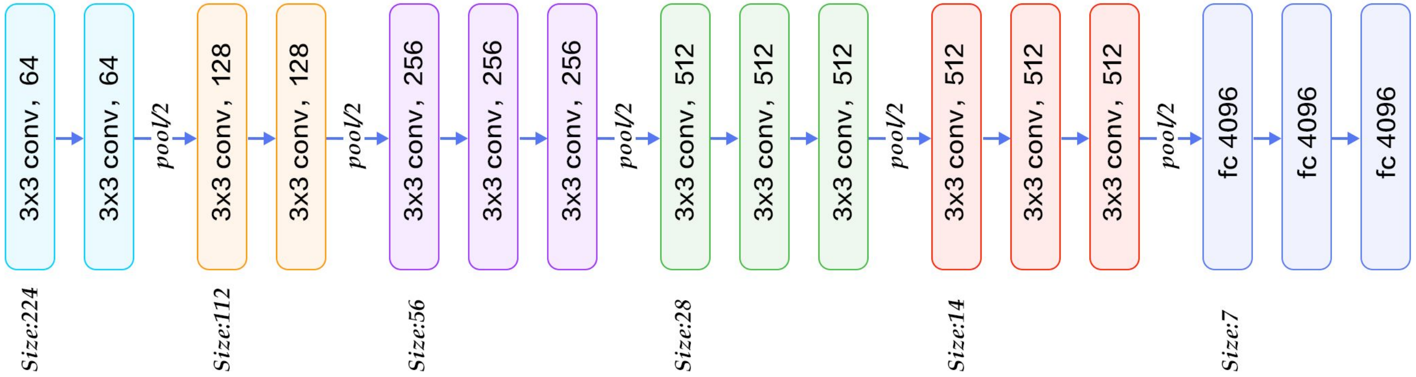
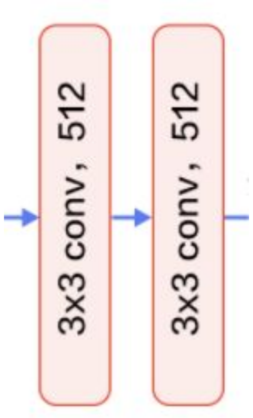


Figure Credits: Yugandhar Nanda's answer on Quora: What is the VGG neural network?
VGG paper: Simonyan, K. and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." CoRR abs/1409.1556 (2015)

Receptive field

- 3x3 conv - 3x3 conv pattern is very common
- Roughly equivalent to 5x5 conv (same receptive field)
- But 18/25 less parameters
- https://distill.pub/2019/computing-receptive-fields/



3x3 conv, 512

3x3 conv, 512

Pool VS Stride

- Many forensics CNNs (steganalysis, forgery detection, ...) do not use pooling in early layers
- Discarding pooling layers has also been found to be important in training good GANs
- Pooling can be criticized for losing information
- Recent CNN architectures tend to replace pooling layers with strided conv

Historical note

- The early CNNs had strides: LeCun, Yann, et al. "Backpropagation applied to handwritten zip code recognition." Neural computation 1.4 (1989): 541-551.
- Replaced by pooling to make the CNN more robust to pixel level distortions

Springenberg, Jost Tobias et al. “Striving for Simplicity: The All Convolutional Net.” CoRR abs/1412.6806 (2015):

End