Lecture 13. Convolutional Neural Networks

COMP90051 Statistical Machine Learning

Lecturer: Zahra Dasht Bozorgi

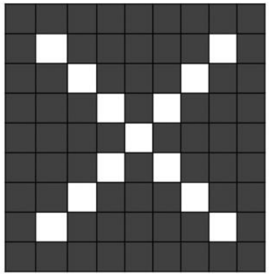
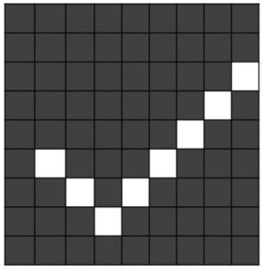
THE UNIVERSITY OF MELBOURNE

This lecture

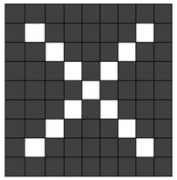
- Convolution operator
  - Convolutions
  - Convolution layers in a neural network
- Convolutional neural networks
  - LeNet, ResNet (2d images)
  - CNN (1d language)

Motivating example

- Image classification  vs 
  - instance is matrix of pixels

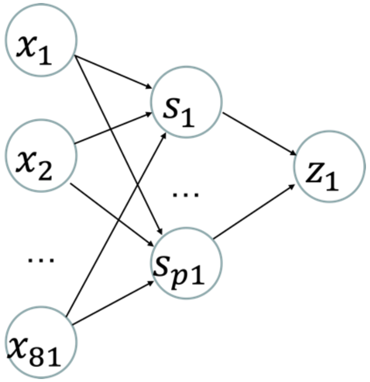



- How can we apply a neural net?
  - flatten into vector, then use fully connected network



9x9

81x1



Input layer

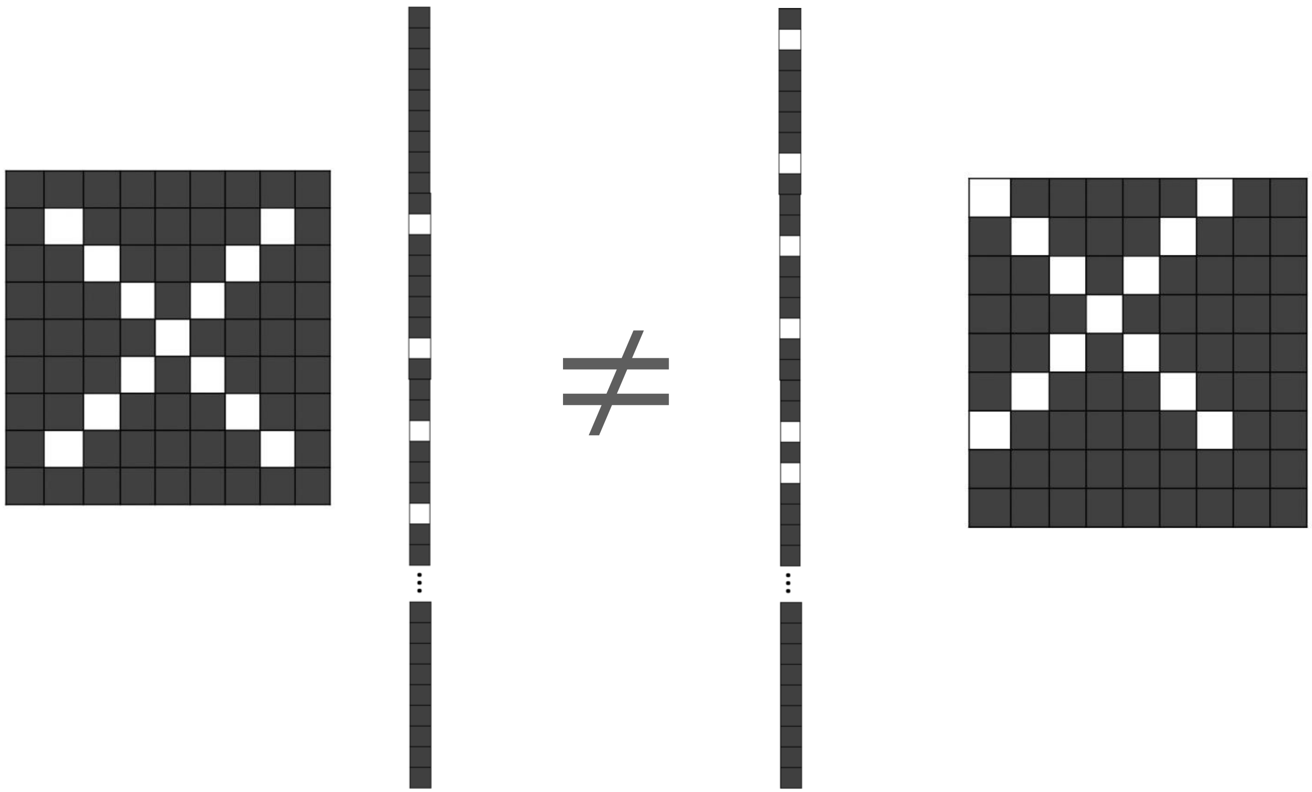
Hidden layer

Output layer

3

Fully-connected net, no spatial invariance

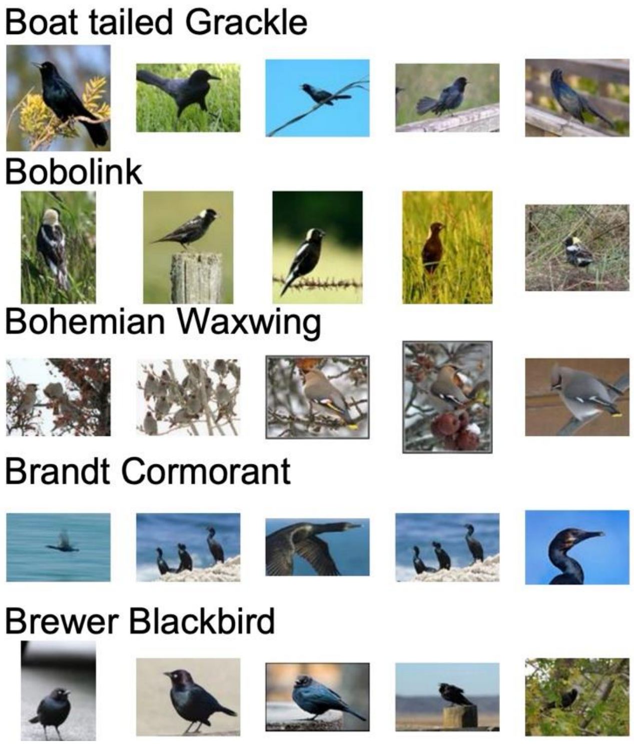
- Disadvantage: must learn same concept again & again!

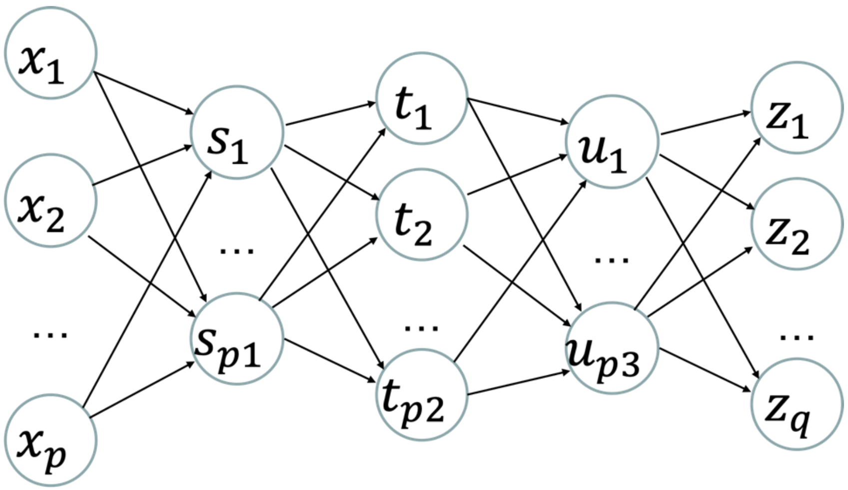


- Translation invariance: architecture that activates on the same pattern even if “translated” spatially

Use more depth?

- Inefficient, requires huge numbers of parameters with more hidden layers. Could overfit.

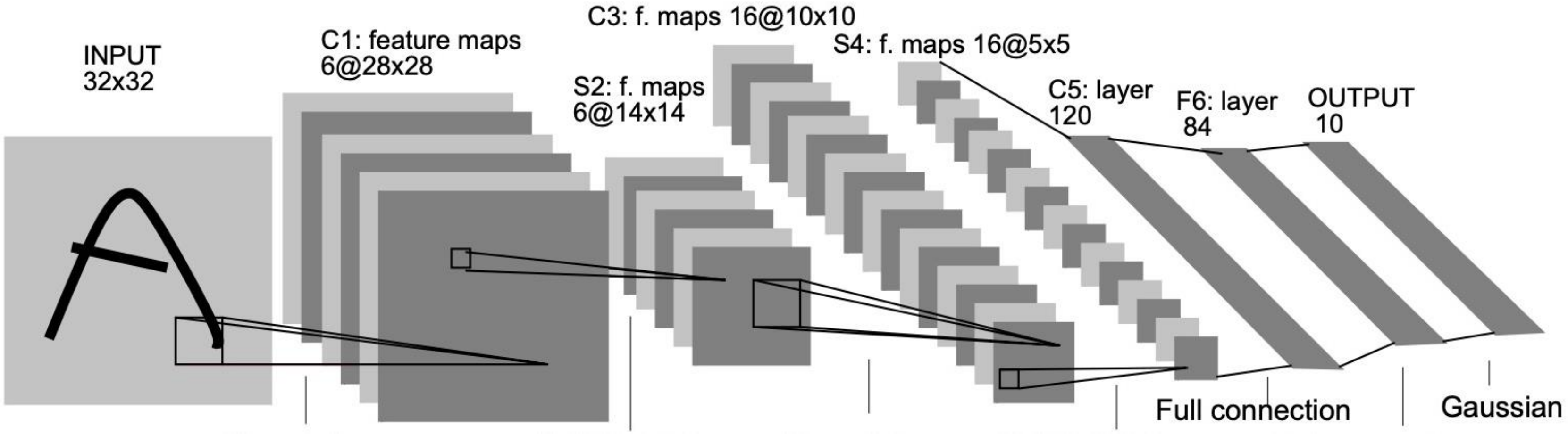




Missed opportunity: sharing weights for these (effectively identical or very similar) input configurations

Convolutional Neural Network (CNN)

- In computer vision, filters are small square patterns such as line segments or textures, used as features
- Need ways to: match filters against image (**next**); learn filters
- Key idea: learn translation invariant filters - parameter sharing



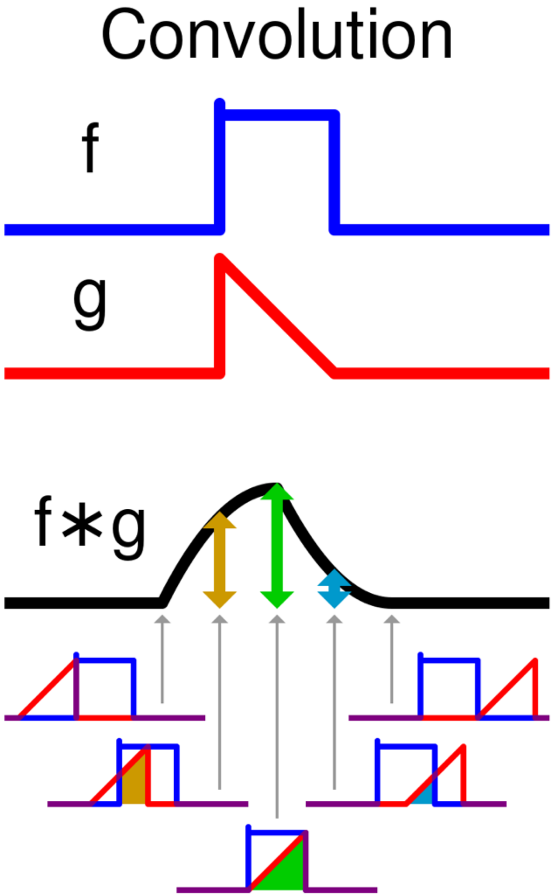
LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.

Convolution operator

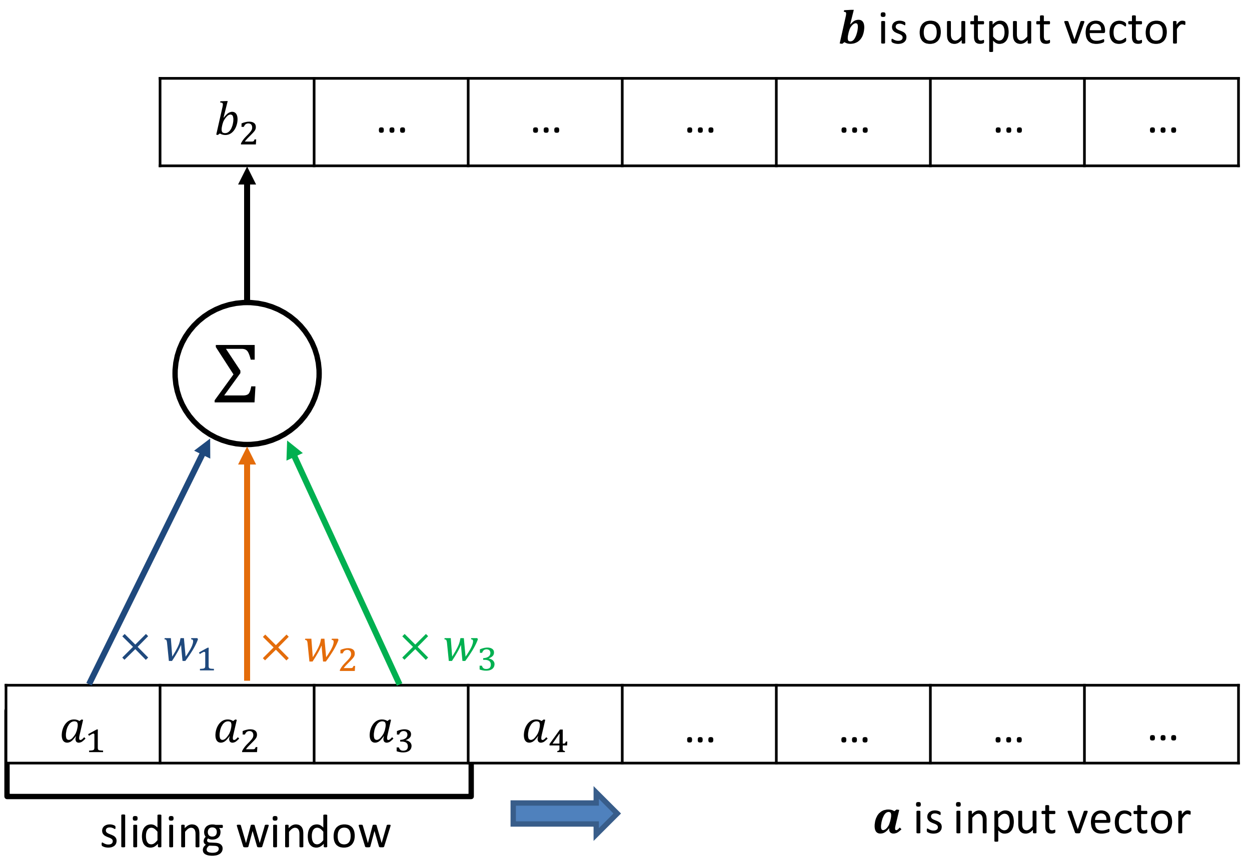
Allows us to match a small filter across multiple patches of a 2D image or range of a 1D input

Convolution

- Concept from signal processing, with wide-spread application
  - Defined as
    $$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau$$
  - Measures how the shape of one function matches the other as it slides along.
- ConvNets use this idea applied to discrete inputs



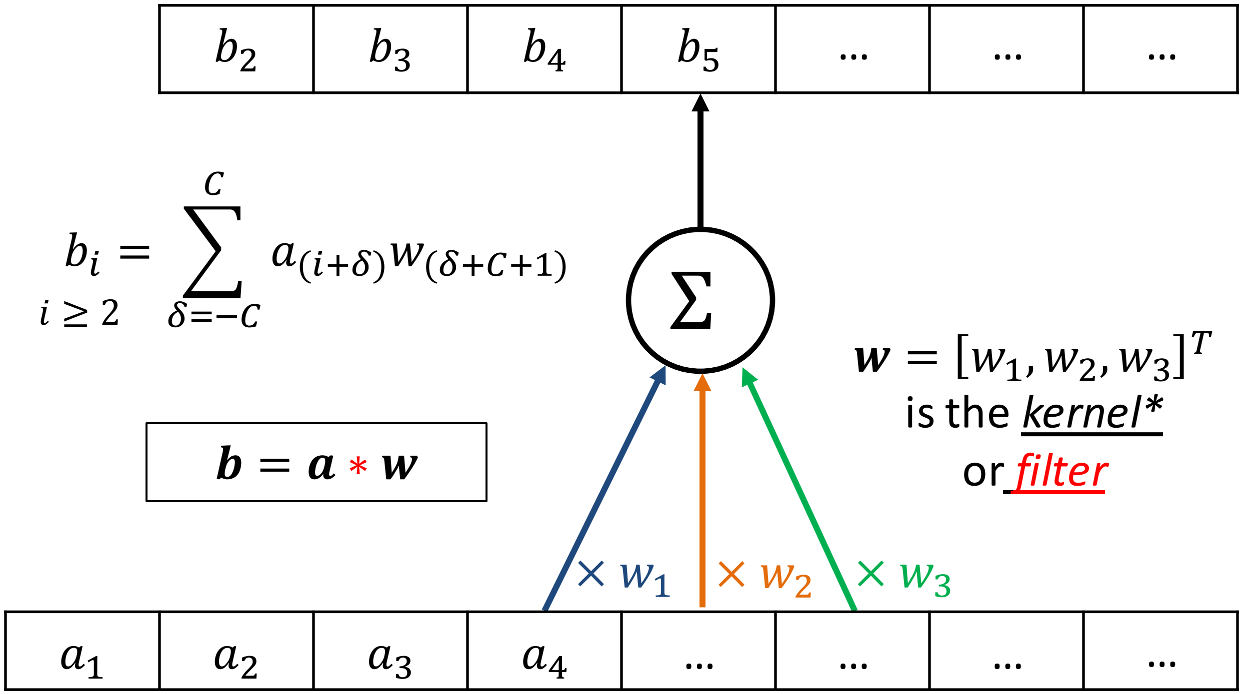
Convolution in 1D



A.k.a. "time delay" neural network

Convolution in 1D

$\boldsymbol{b}$ is output vector



$$\underset{i \geq 2}{b_i} = \sum_{\delta=-C}^{C} a_{(i+\delta)} w_{(\delta+C+1)}$$

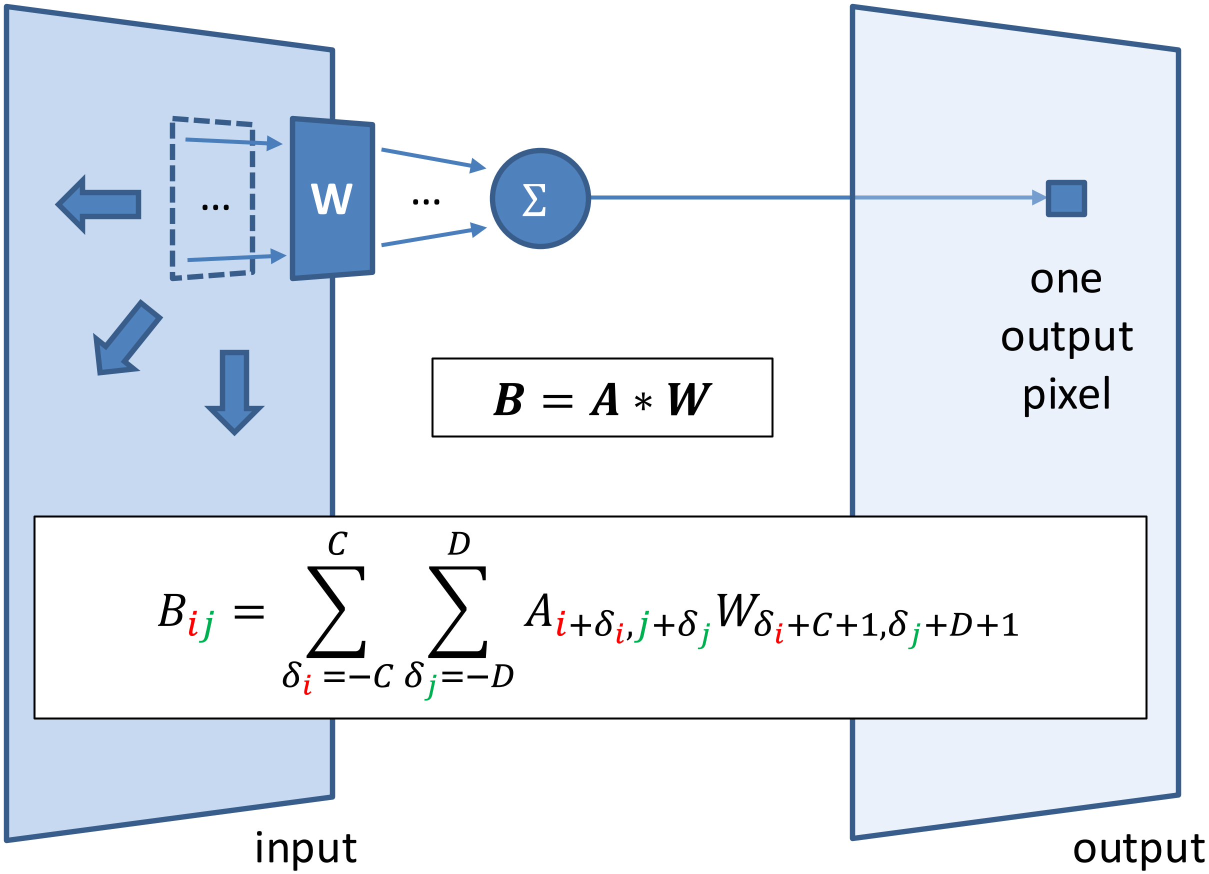
$$\boxed{\boldsymbol{b} = \boldsymbol{a} * \boldsymbol{w}}$$

$\boldsymbol{w} = [w_1, w_2, w_3]^T$ is the *kernel** or *filter*

$\boldsymbol{a}$ is input vector

*Unrelated to definition of kernel (for SVMs) seen in subject, as a function representing a dot product

Convolution on 2D images



$$\boldsymbol{B} = \boldsymbol{A} * \boldsymbol{W}$$

$$B_{ij} = \sum_{\delta_i = -C}^{C} \sum_{\delta_j = -D}^{D} A_{i+\delta_i, j+\delta_j} W_{\delta_i + C + 1, \delta_j + D + 1}$$

Convolution in 2D

- Use filter/kernel to perform element-wise multiplication and sum for every local patch

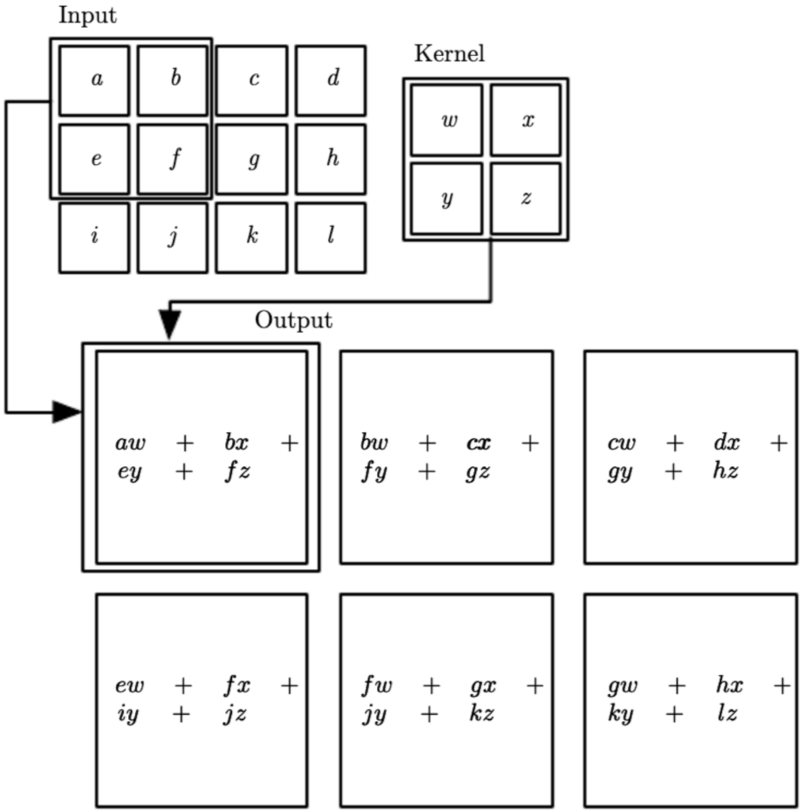
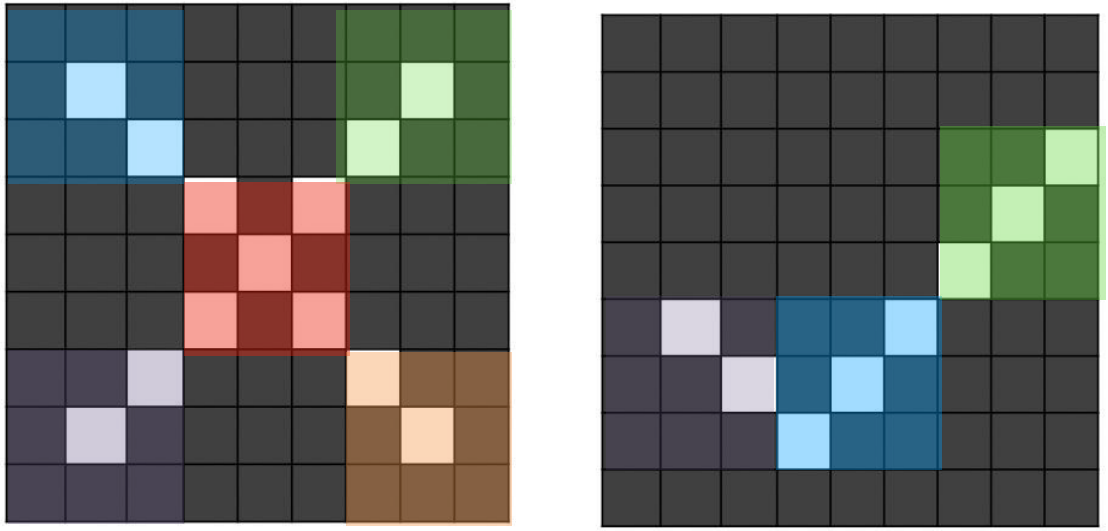


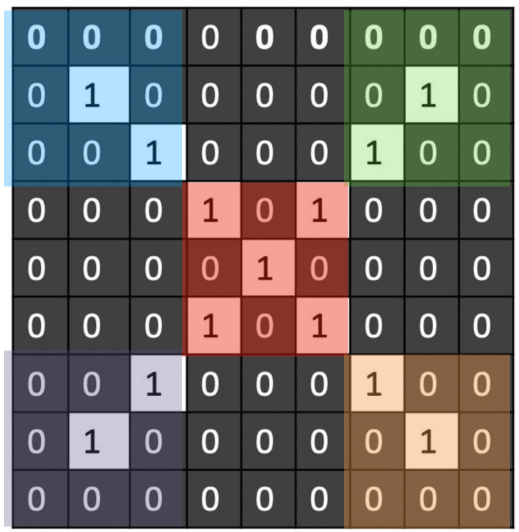
Image decomposes into local patches

- Different local patches include different patterns
  - we can first extract local features (local patterns) and then combine local features for classification



Convolutional filters (aka kernels)

- Filters/kernels can identify different patterns



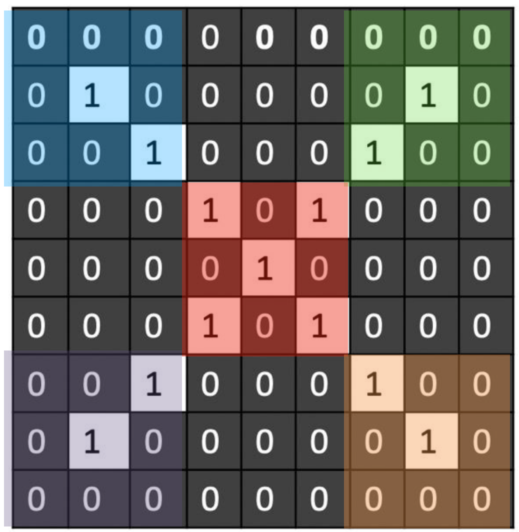
Element-wise multiplication

$$\text{Sum}\left(\begin{bmatrix}0&0&0\\0&1&0\\0&0&1\end{bmatrix} \odot \begin{bmatrix}0&0&0\\0&1&0\\0&0&1\end{bmatrix}\right) = 2$$

$$\text{Sum}\left(\begin{bmatrix}0&0&0\\0&1&0\\1&0&0\end{bmatrix} \odot \begin{bmatrix}0&0&0\\0&1&0\\0&0&1\end{bmatrix}\right) = 1$$

- When input and kernel have the same pattern: high activation response

Different kernels identify different patterns



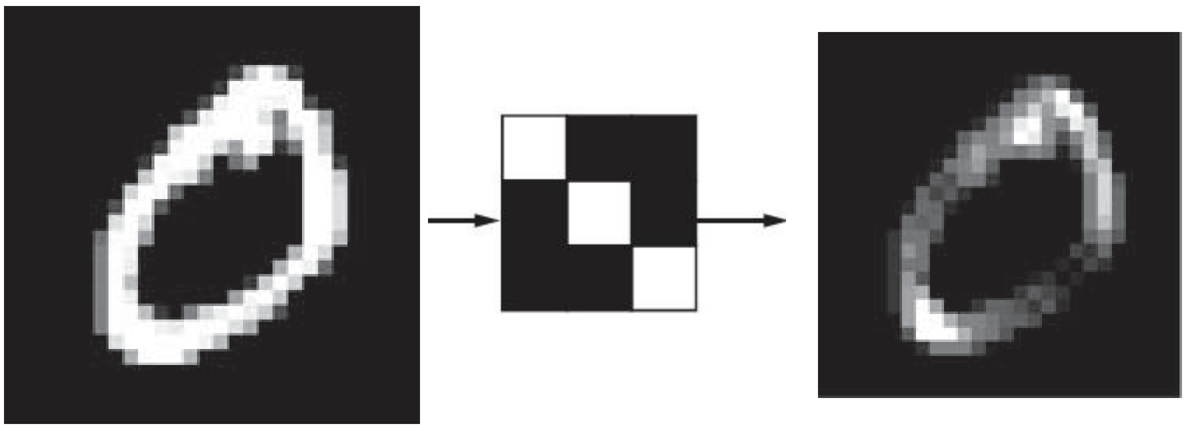
Element-wise multiplication

$$\text{Sum}\left(\begin{bmatrix}0&0&0\\0&1&0\\0&0&1\end{bmatrix} \odot \begin{bmatrix}1&0&1\\0&1&0\\1&0&1\end{bmatrix}\right)=2$$

$$\text{Sum}\left(\begin{bmatrix}1&0&1\\0&1&0\\1&0&1\end{bmatrix} \odot \begin{bmatrix}1&0&1\\0&1&0\\1&0&1\end{bmatrix}\right)=5$$

Convolution in 2D example (MNIST)

- Response map (Feature map) for single kernel



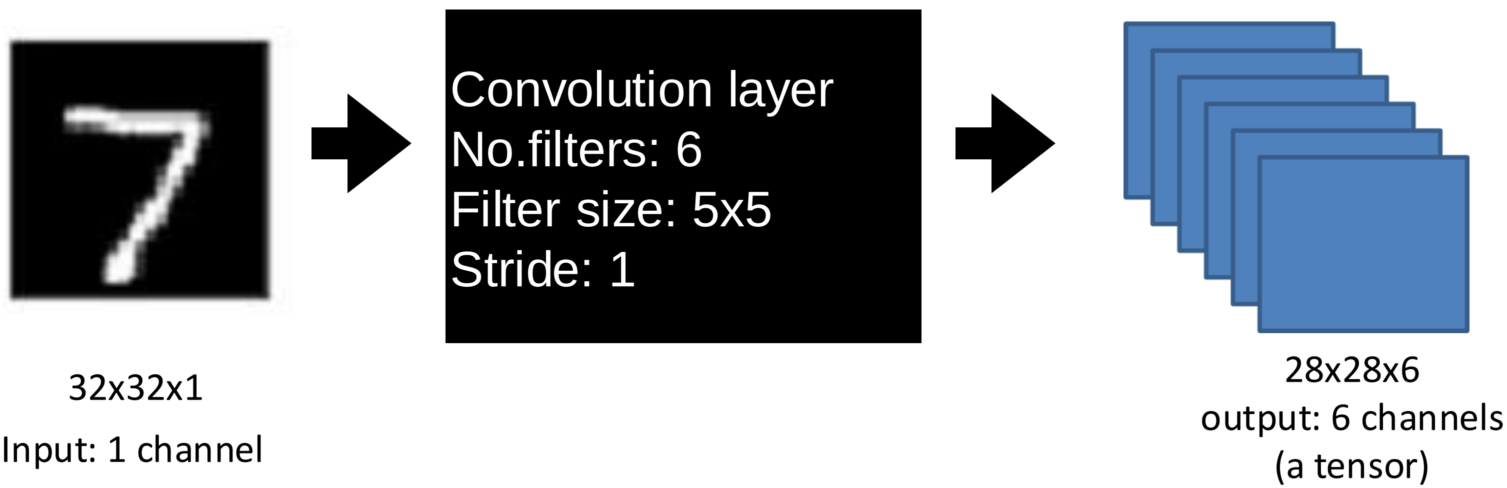
Feature map: 2D map of the presence of a pattern at different locations in an input

- Different kernels identify different patterns: use several filters in each layer of network

http://yann.lecun.com/exdb/mnist/

Convolution parameters

- Filters are parameters themselves to be learned (**next** time)
- Key hyperparameters in convolution
  * Kernel size: size of the patches
  * Number of filters: depth (channel) of the output
  * Stride: how far to "slide" patch across input
  * Padding of input boundaries with zeros (black here)



Convolution on Multiple-channel input

R

G

B

Kernel: applied to each channel

$* K(Channel\ 1)$

$* K(Channel\ 2)$

$* K(Channel\ 3)$

Element-wise sum → Single channel

Mini Summary

- Convolution operator
  - Convolutions in 1D, 2D
  - Convolution layers in a neural network

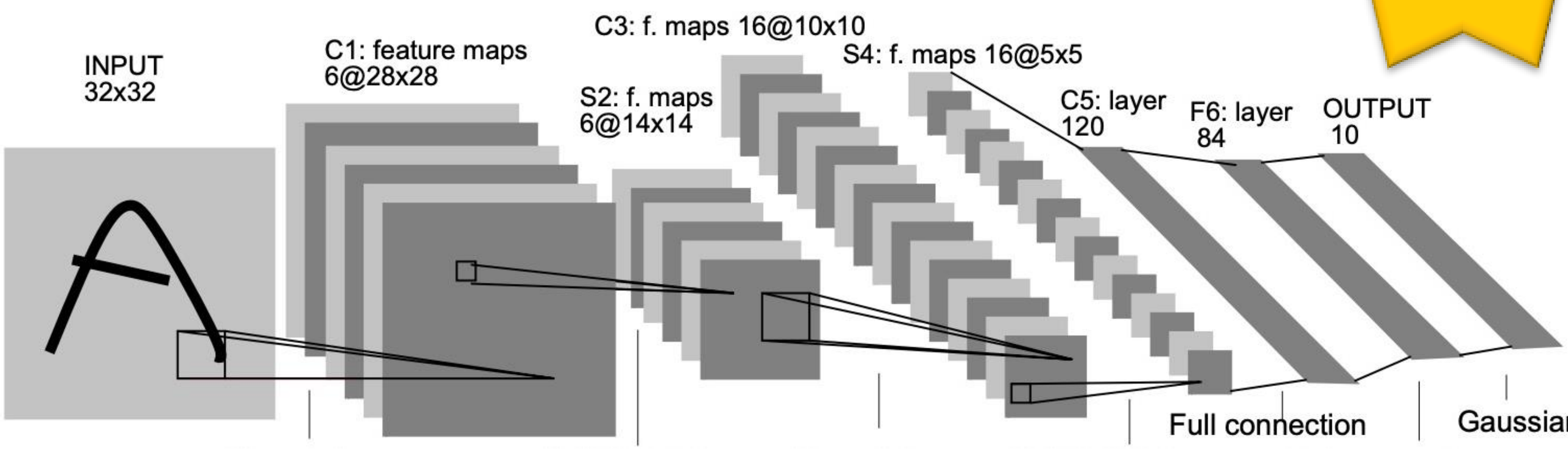
Next: CNNs in practice

Convolutional Neural Networks (CNN)

Deep networks combining convolutional filters, pooling and other techniques

CNN for computer vision

- LeNet-5 sparked modern deep models of vision
  - "C" = convolution, "S" = down-sampling, "F" = fully connected



LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.

Components of a CNN

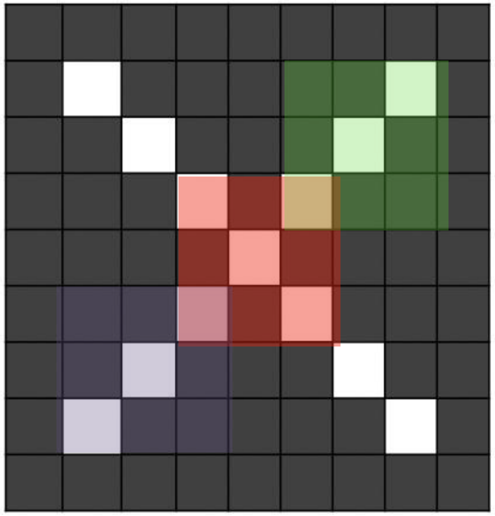
- Convolutional layers
  - Complex input representations based on convolution operation
  - Filter weights are learned from training data
- Downsampling, usually via Max Pooling
  - Re-scales to smaller resolution, limits parameter explosion
- Fully connected parts and output layer
  - Merges representations together

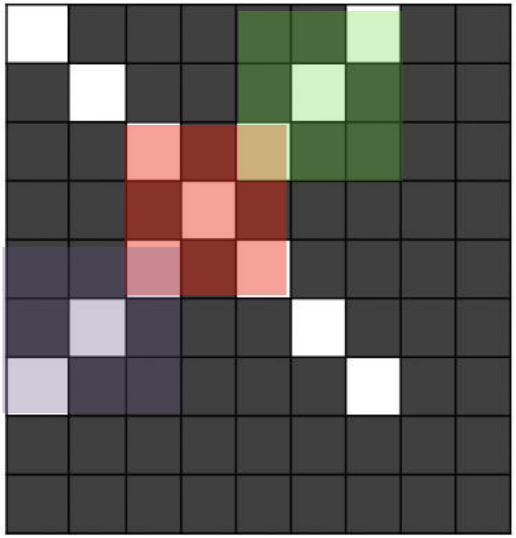
Downsampling via max pooling

- Special type of processing layer. For an $m \times m$ patch
$$v = \max(u_{11}, u_{12}, \dots, u_{mm})$$
- Strictly speaking, not everywhere differentiable. Instead, gradient is defined according to "sub-gradient"
  * Tiny changes in values of u_{ij} that is not max do not change v
  * If u_{ij} is max value, tiny changes in that value change v linearly
  * Use $\frac{\partial v}{\partial u_{ij}} = 1$ if $u_{ij} = v$, and $\frac{\partial v}{\partial u_{ij}} = 0$ otherwise
- Forward pass records maximising element, which is then used in the backward pass during back-propagation

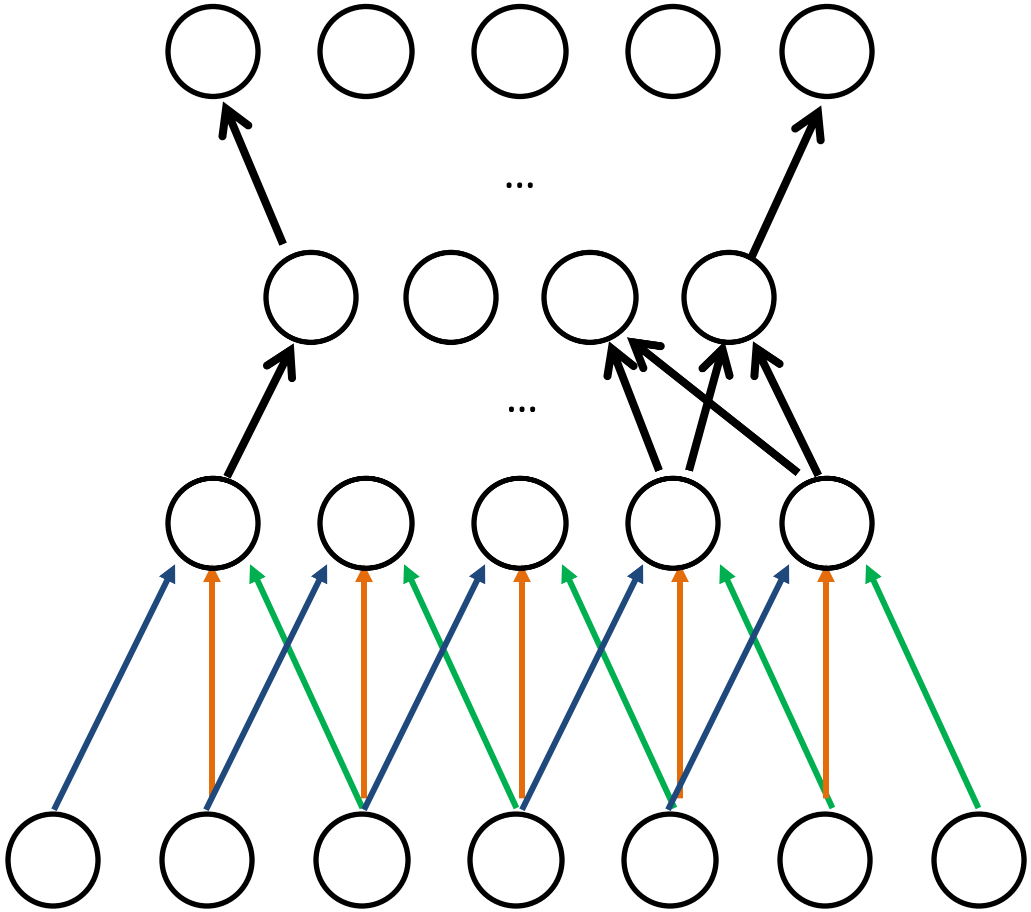
Convolution + Max Pooling → Translation invariance

- Consider shift input image
  - exact same kernels will activate, with same responses
  - max-pooling over the kernel outputs gives same output
  - size of max-pooling patch limits the extent of invariance
- Can include padding around input boundaries





Convolution as a regulariser



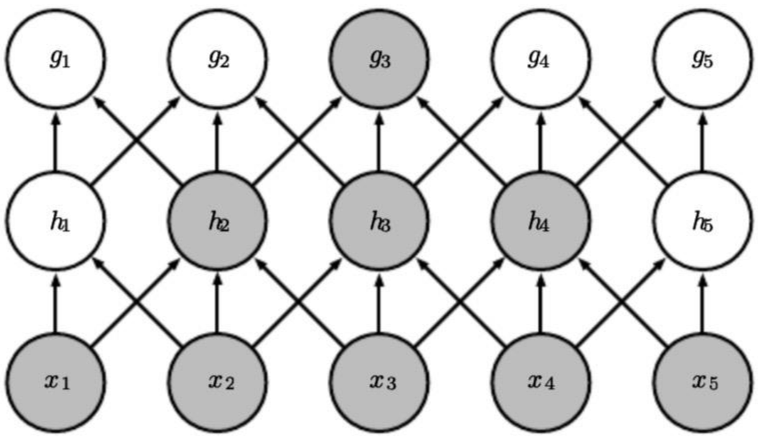
Fully connected, unrestricted

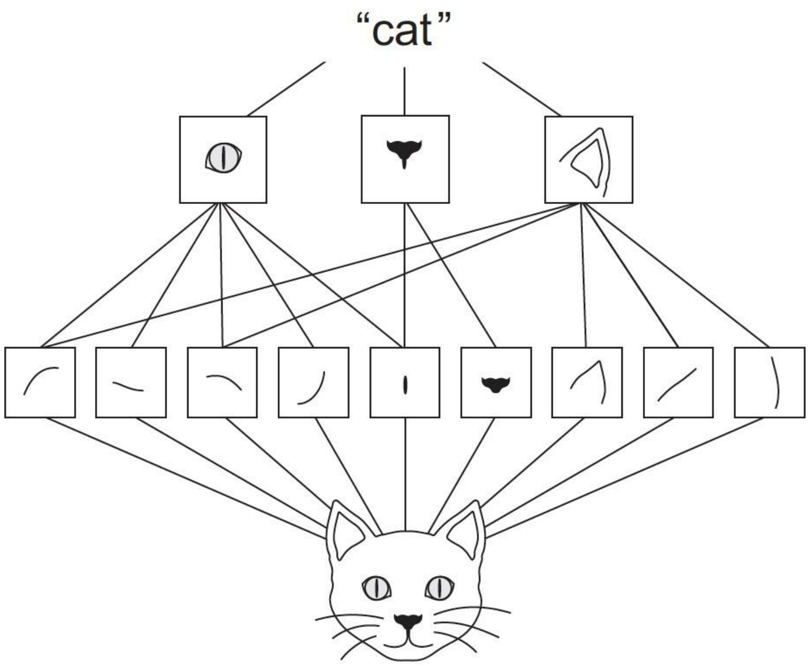
Fully connected, unrestricted

Restriction: same color – same weight

Conv Nets learn hierarchical patterns

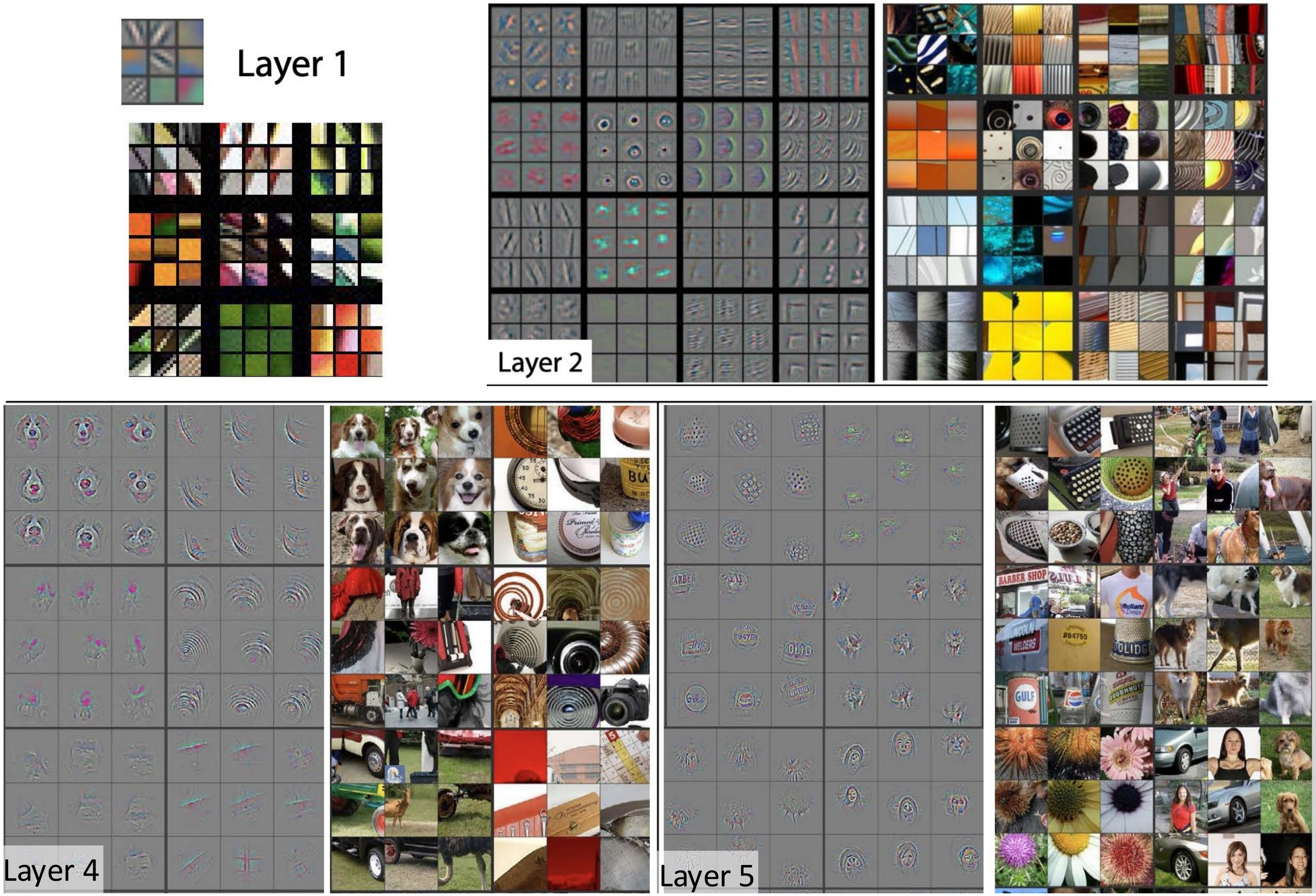
- Stacking several layers of convolution:
  larger size of receptive field (more of input is seen)





Inspecting learned kernels

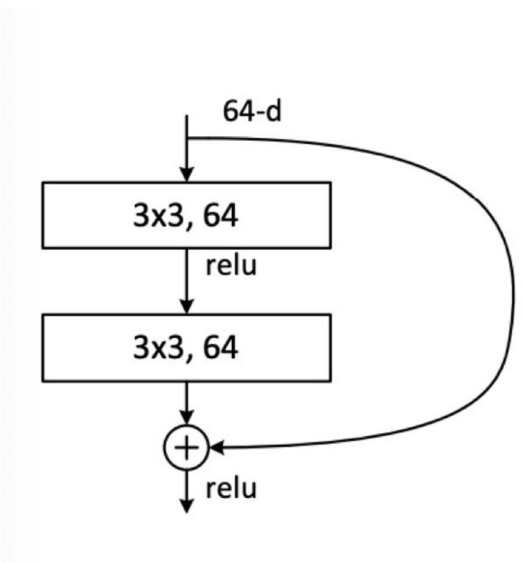
Kernels (grey) and some images that strongly activate each kernel

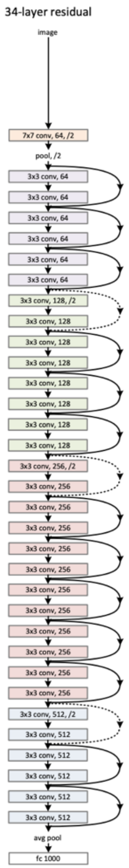


Zeiler, M., and R. Fergus. "Visualizing and understanding convolutional networks." *European conference on computer vision*. 2014

ConvNets in Computer Vision

- ResNet represents modern state-of-the-art
  - Up to 151 layers (!)
  - mixture of convolutions, pooling, fully connected layers
- Critical innovation is the "residual connection"
  - linear copy of input to output
  - easier to optimise despite depth, solving gradient vanishing problem



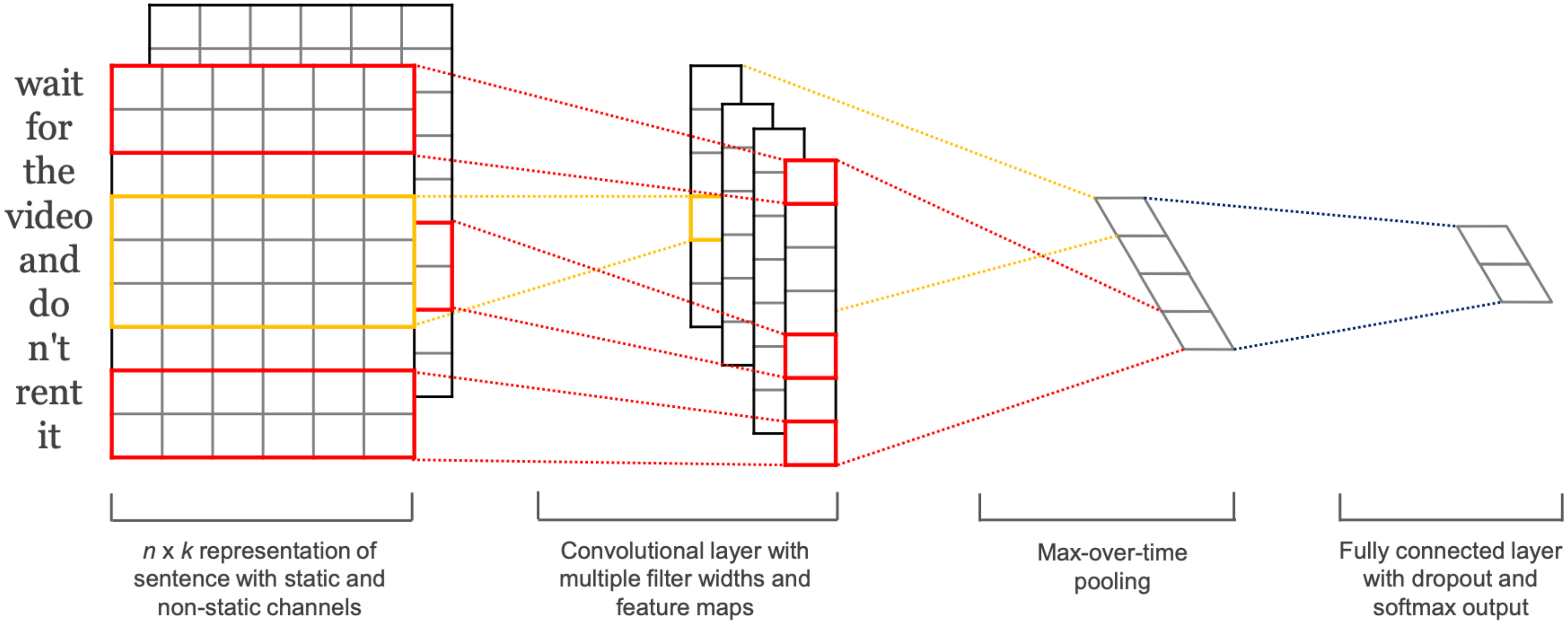


- Standard practise to *pretrain* big model on large dataset, then *fine-tune* (continue training) on small target task

He, K., Zhang, X., Ren, S., & Sun, J. Deep residual learning for image recognition. CVPR 2016.

ConvNets for Language

- Application of 1d kernels to word sequences
  - capture patterns of nearby words



Kim, Y. Convolutional Neural Networks for Sentence Classification, EMNLP 2014

This lecture

- Convolutional Neural Networks
  - Convolution operator
  - 1d vs 2d convolutions
  - Elements of a convolution-based networks
  - ConvNets in practice for vision & language

Next lecture: Recurrent Neural Networks (RNNs)