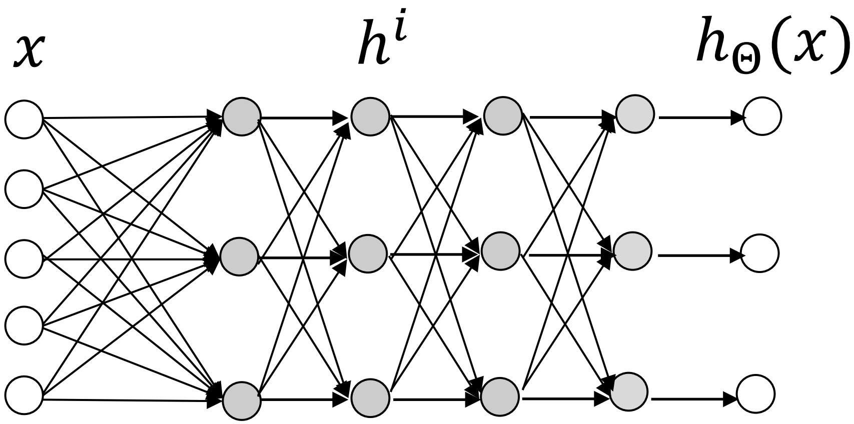
Today: Outline

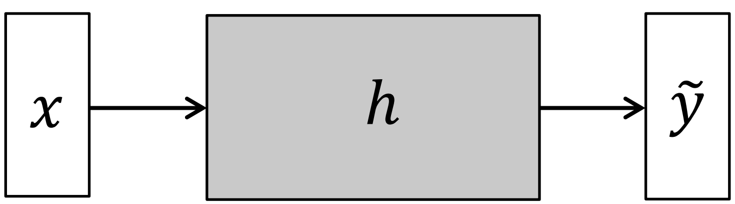
- **ConvNets**: multiplication vs convolution; filters (or kernels); convolutional layers; 1D and 2D convolution; pooling layers; LeNet, CIFAR10Net

- Reminders: *Pre-lec Material 2,*
  due: Friday, Jun 4

  Problem Set 1,
  due: Friday, Jun 4

Neural networks: recap





Learn parameters via gradient descent

$$\min_{\Theta} J(\Theta)$$

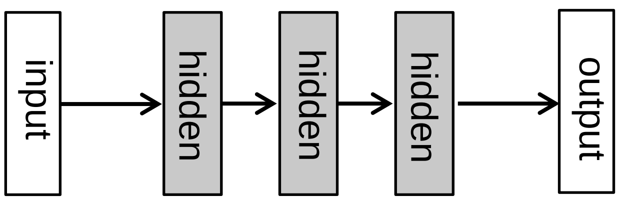
Backpropagation efficiently computes cost (forward pass) and gradient (backward pass)

$$\frac{\partial}{\partial \Theta_{ij}^{(l)}} J(\Theta)$$

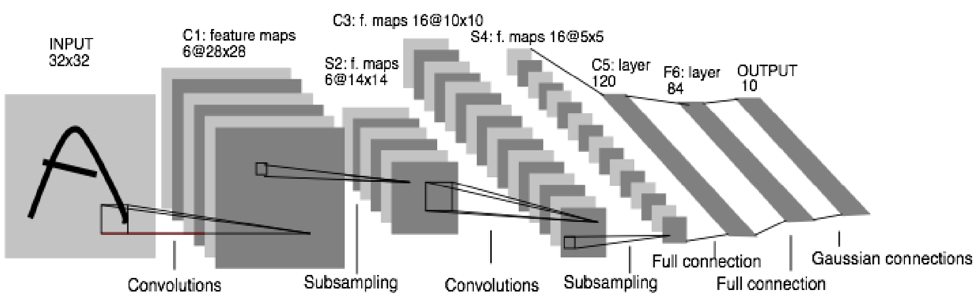
Network architectures

Feed-forward

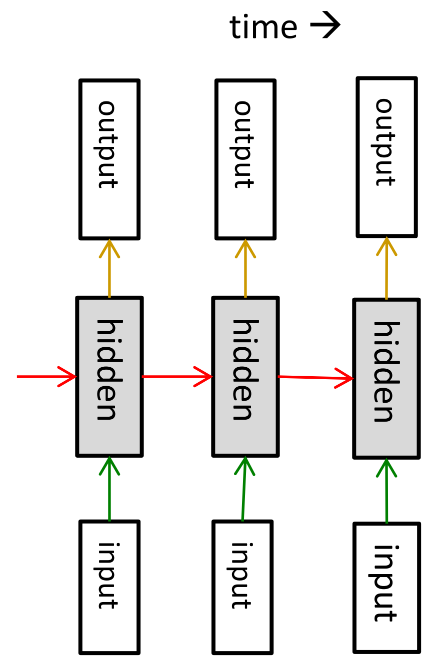
Fully connected



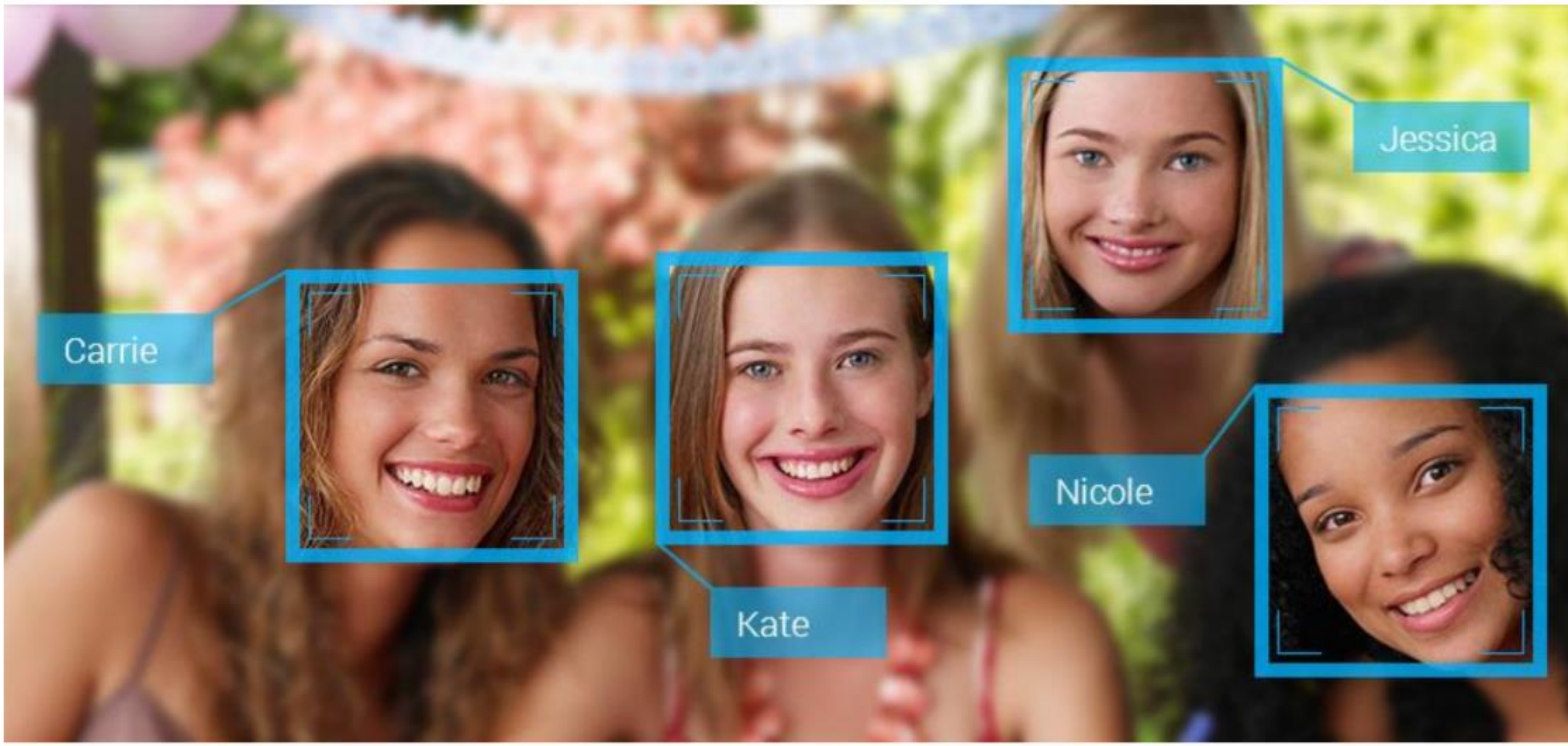
Convolutional



Recurrent



Face Recognition



[towardsdatascience.com]

Image Captioning

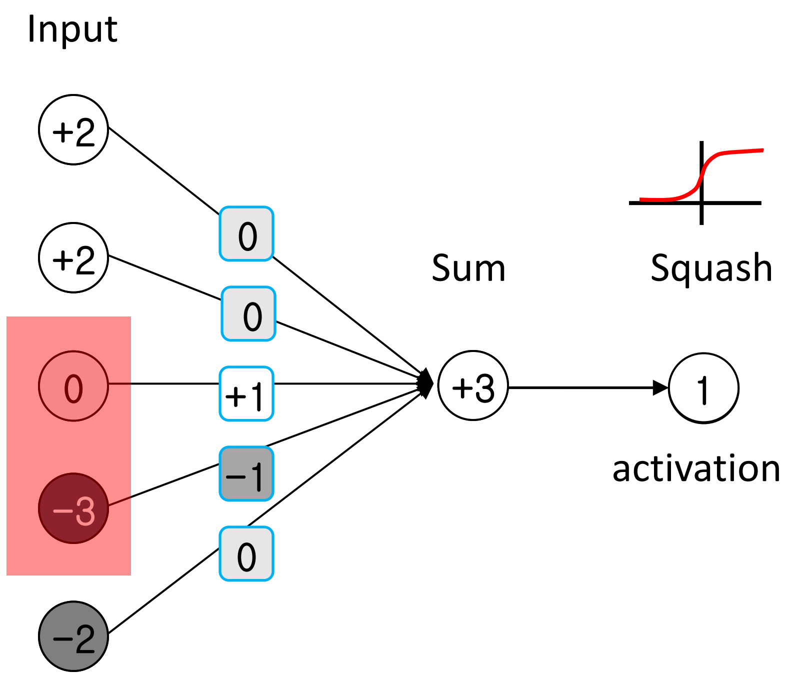
A young boy holding a baseball bat

A man riding a horse next to a building

Neural Networks

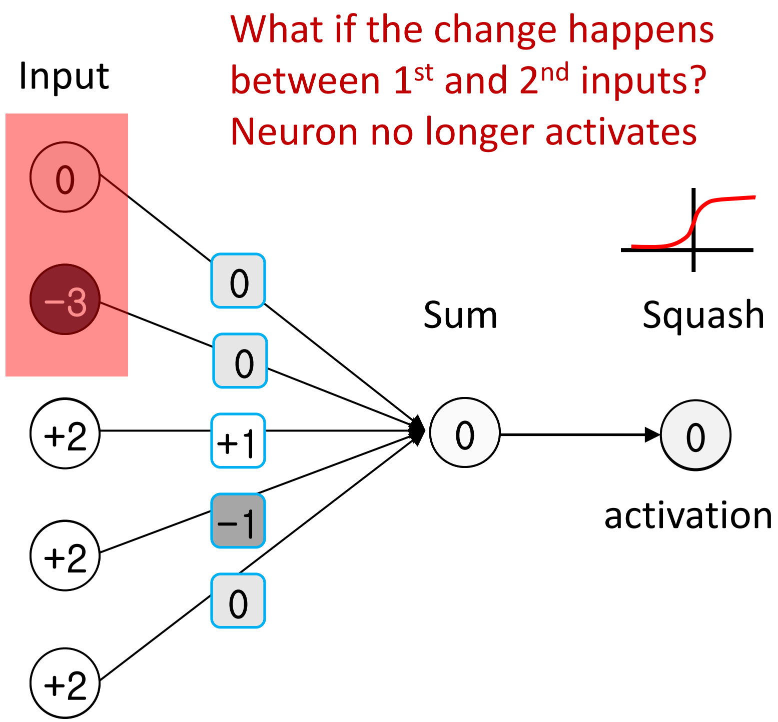
Convolutional Architectures

Multiplication vs convolution



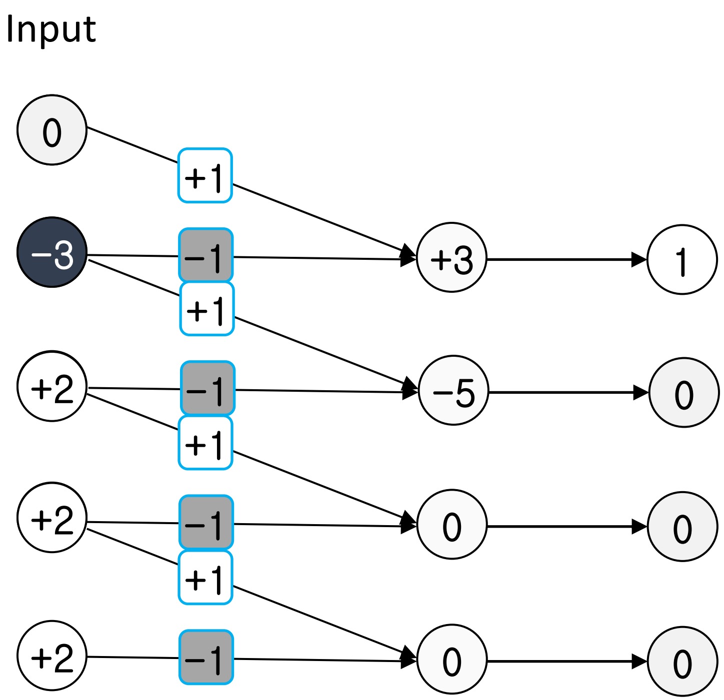
- Recall, a neuron can be thought of as learning to spot certain features in the input
- E.g., this neuron detects change from high to low (light to dark) between 3rd and 4th inputs

Multiplication vs convolution



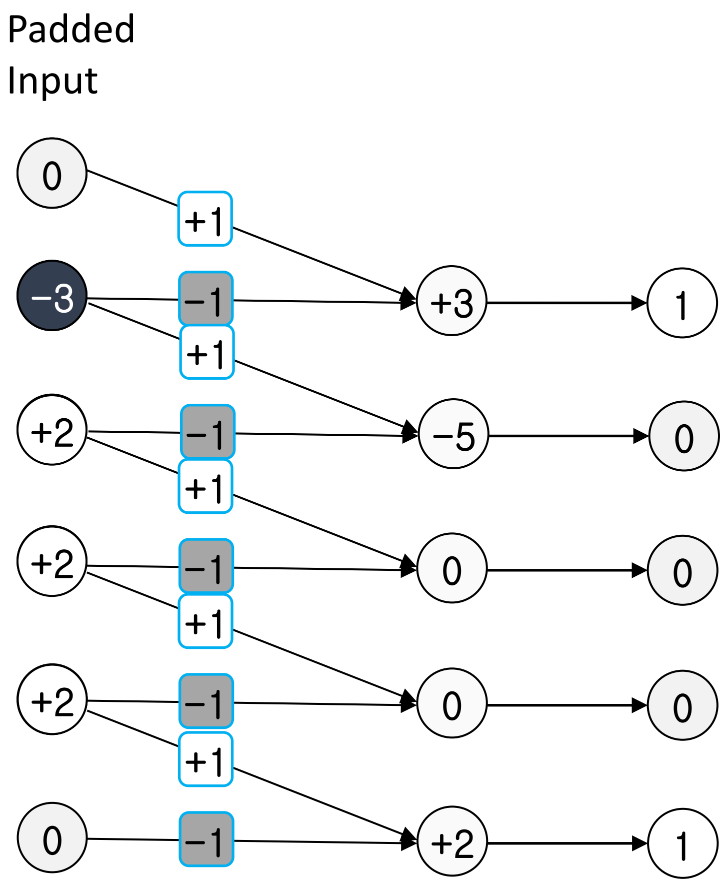
- Must have a new neuron for each new location of pattern???
- This is not efficient
- Solution: use convolution instead of multiplication

Multiplication vs convolution



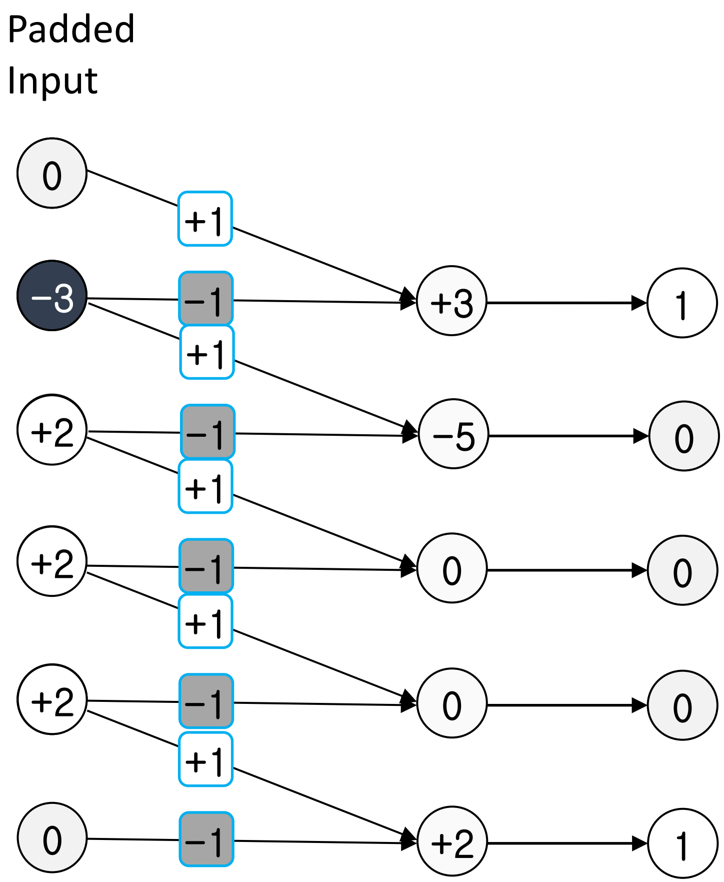
- New weights are of size 2 x 1; called filter, or kernel
- New output is the size of input minus 1 because of boundary
- New convolutional neurons all share the same weights! This is much more efficient; we learn the weights once instead of many times for each position

Multiplication vs convolution



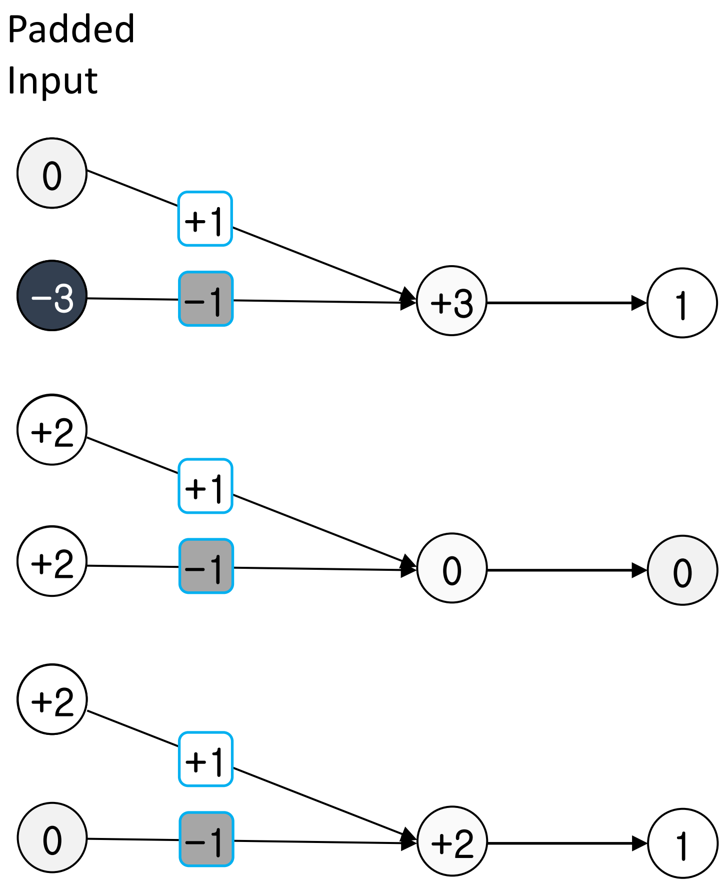
- New output is the size of input minus 1 because of boundary
- We can fix the boundary effect by padding the input with 0 and adding one more neuron

Multiplication vs convolution



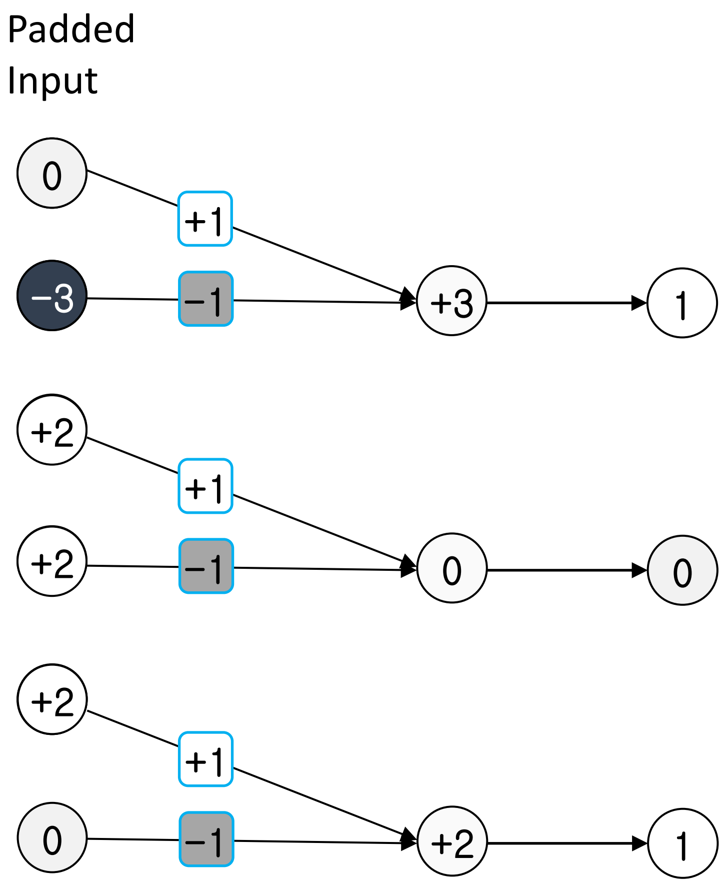
- Note, we move the filter by 1 each time, this is called stride

Multiplication vs convolution



- Note, we move the filter by 1 each time, this is called stride
- Stride can be larger, e.g. here is stride 2

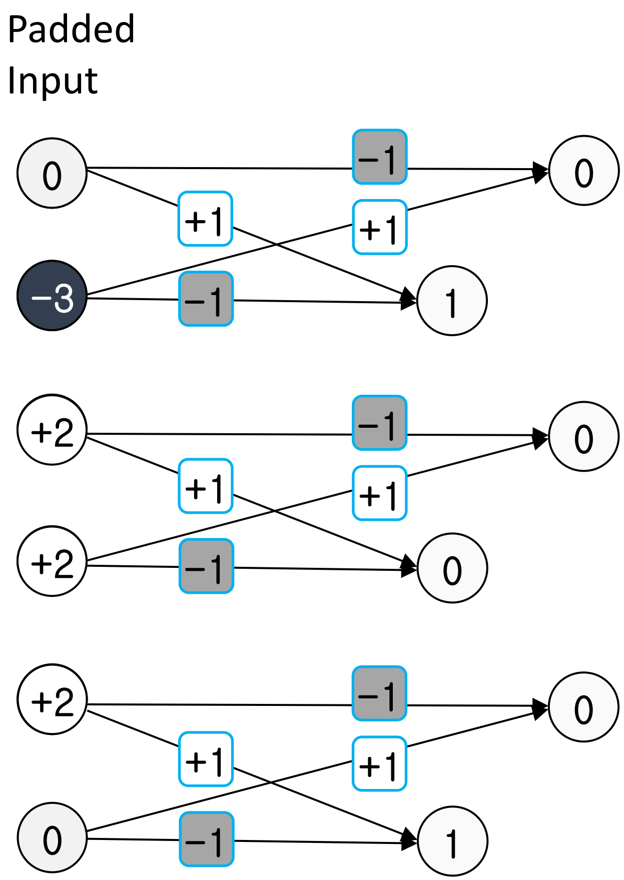
Multiplication vs convolution



To summarize, this layer has

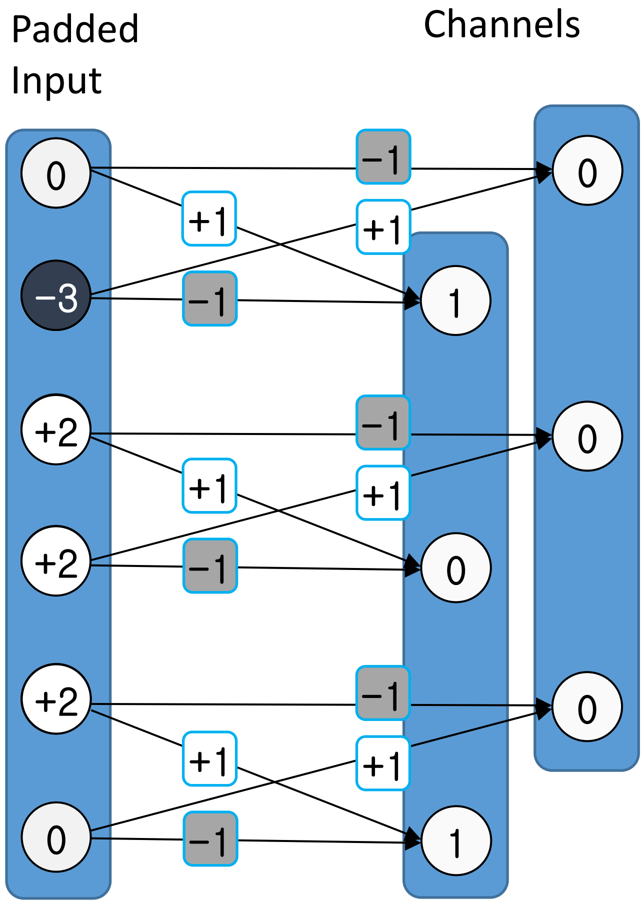
- `Input 5 x 1, padded to 6 x 1`
- `Kernel 2 x 1 with weights [+1,-1]`
- `Stride 2`
- `Output 3 x 1`

Multiplication vs convolution



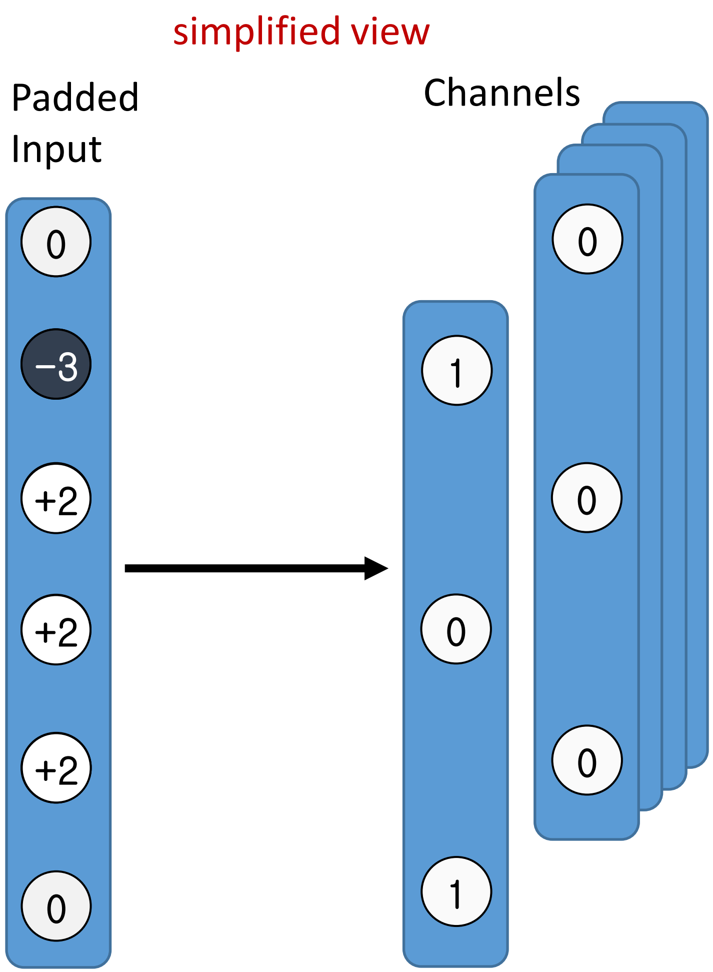
- We can add another filter, this time to detect opposite change with weights [-1 +1]
- Unique filters are called channels

Multiplication vs convolution



- We can add another filter, this time to detect opposite change with weights [-1 +1]
- Unique filters are called channels

Multiplication vs convolution



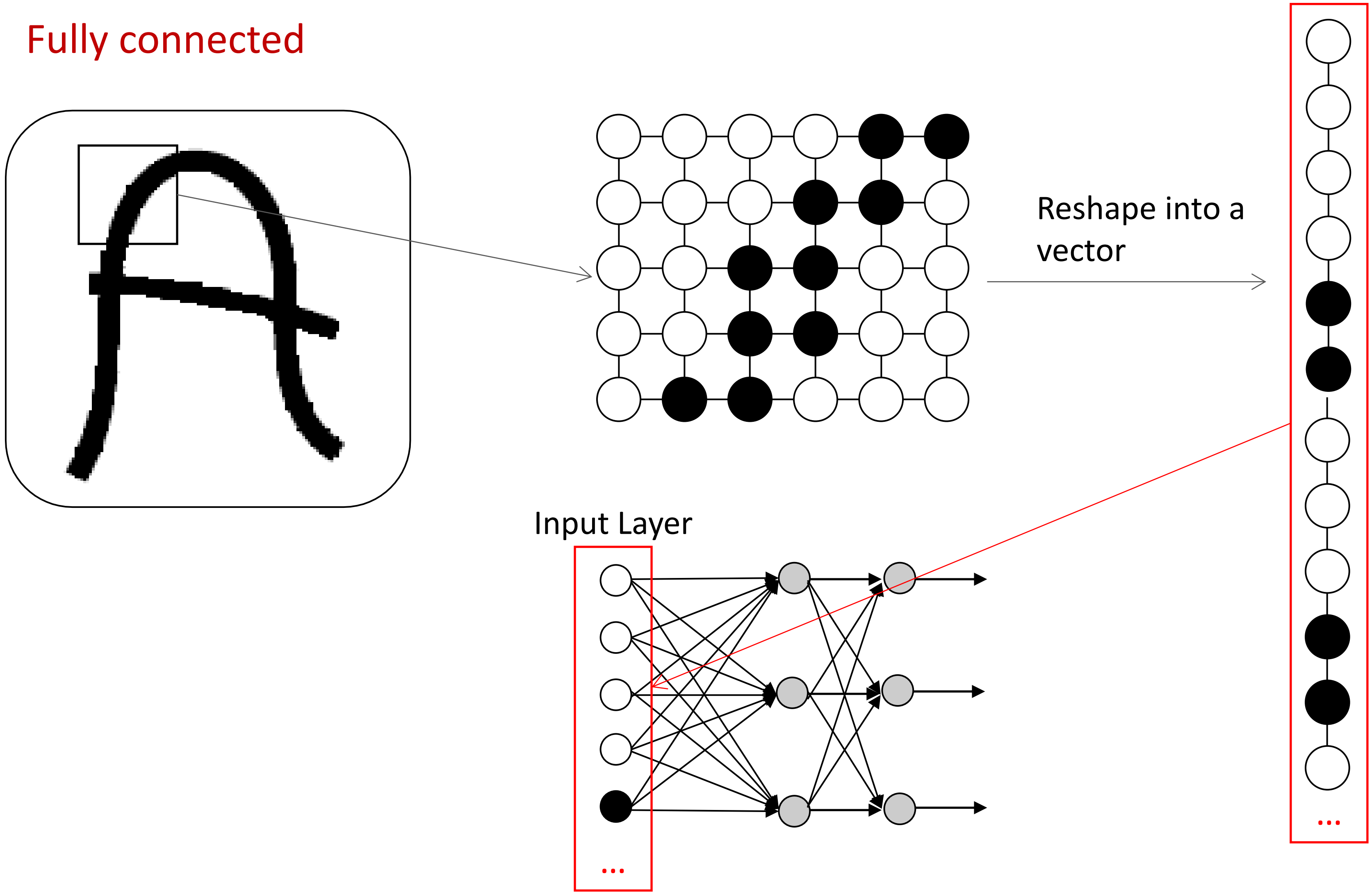
- We can add another filter, this time to detect opposite change with weights [-1 +1]
- Unique filters are called channels

Convolutional Neural Networks

For images and other 2-D signals

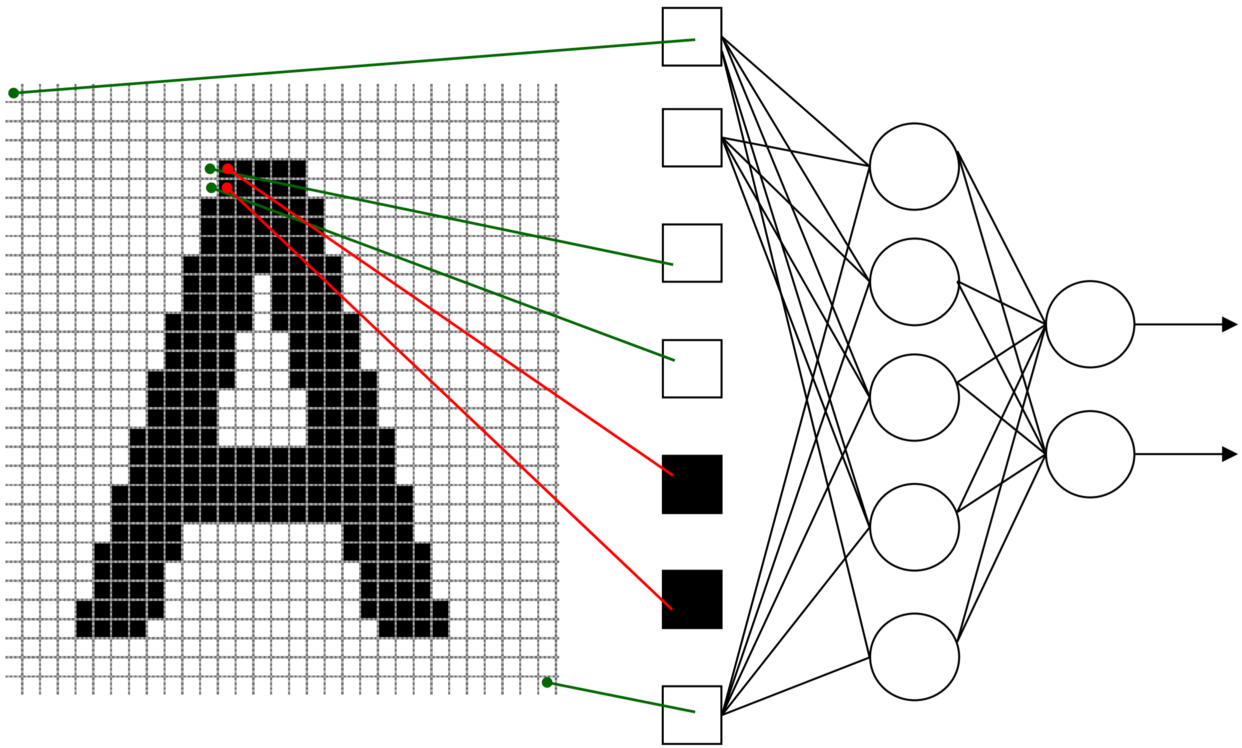
Representing images

Fully connected



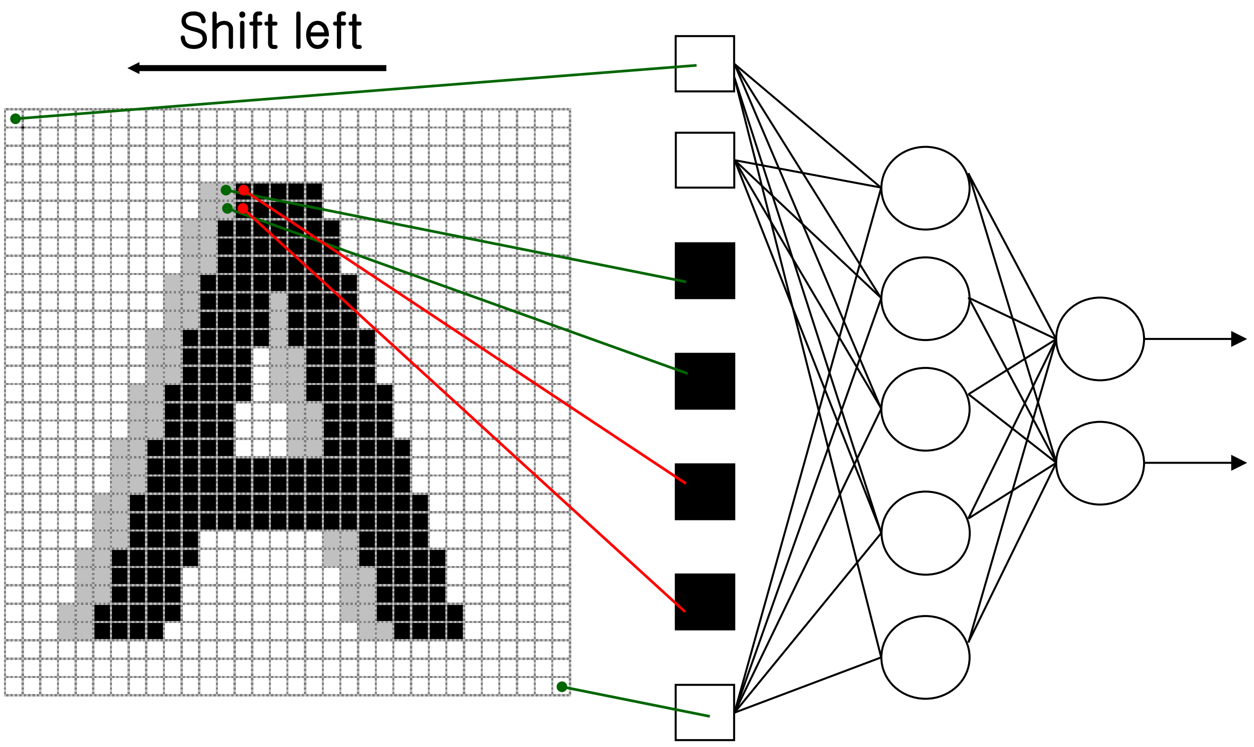
2D Input: fully connected network

Vectorize input by copying rows into a single column



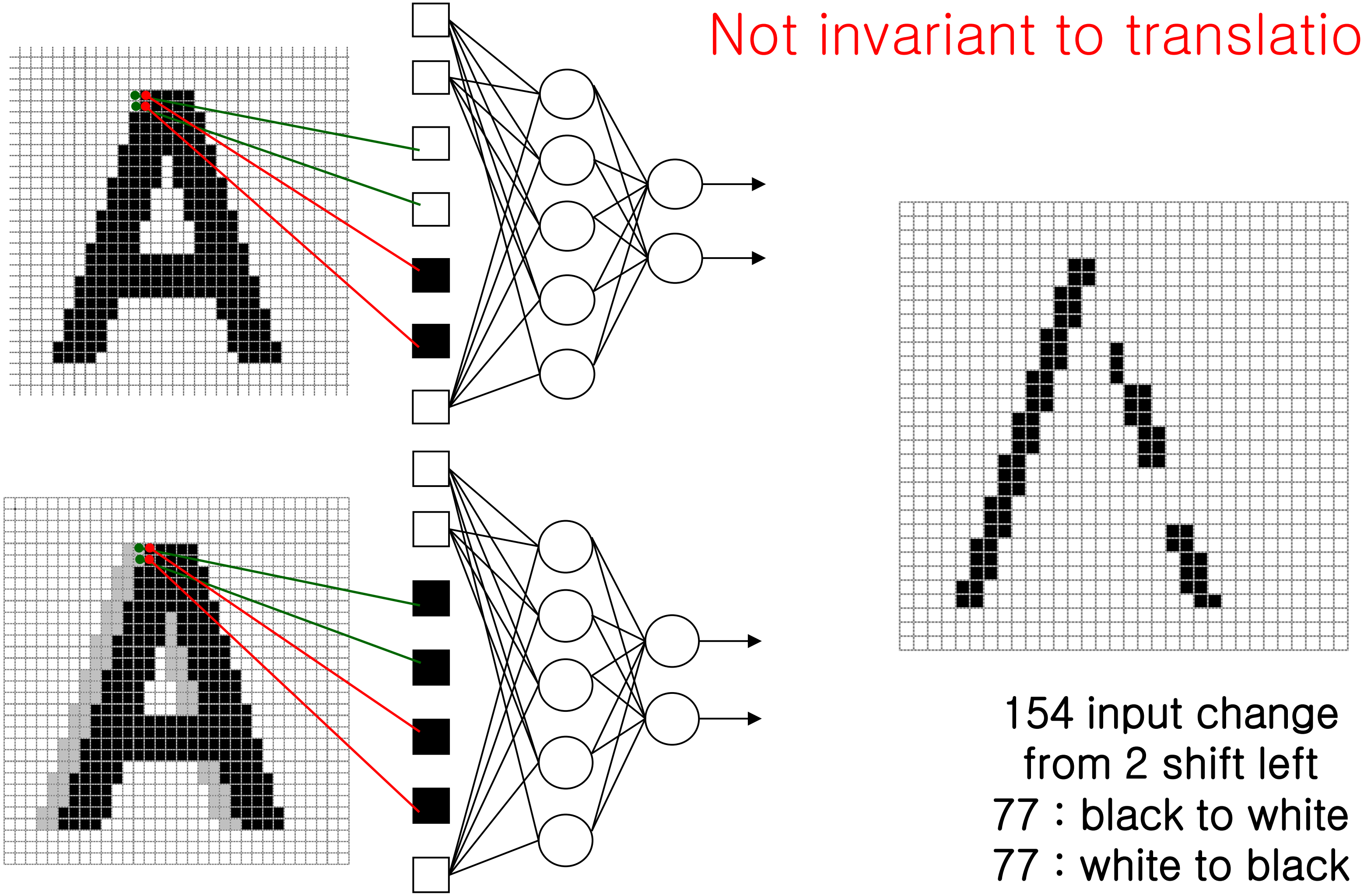
2D Input: fully connected network

Problem: shifting, scaling, and other distortion changes location of features



2D Input: fully connected network

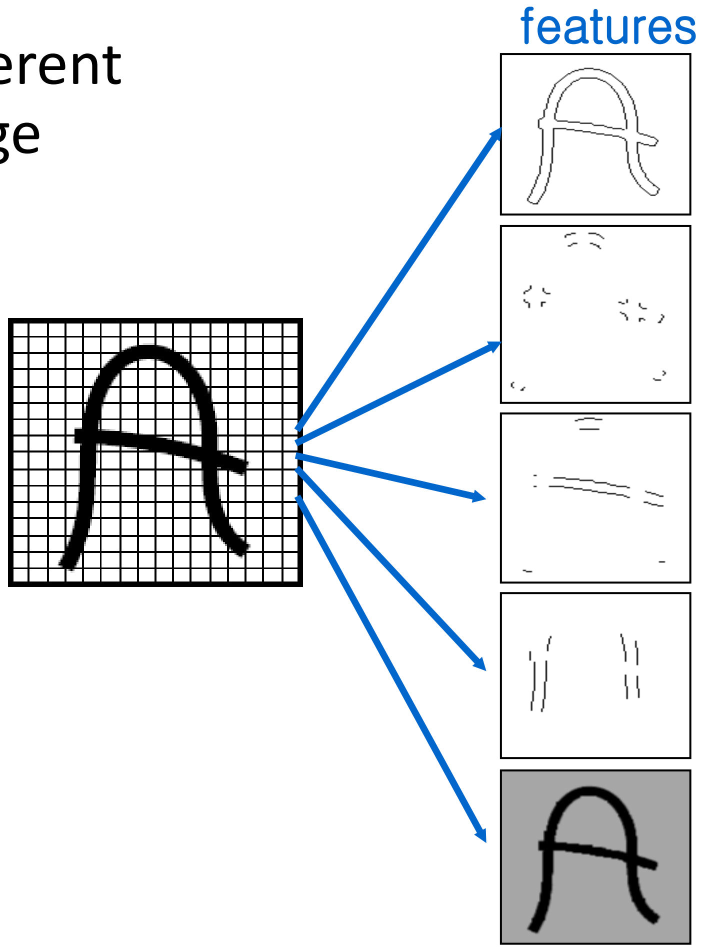
Not invariant to translation!



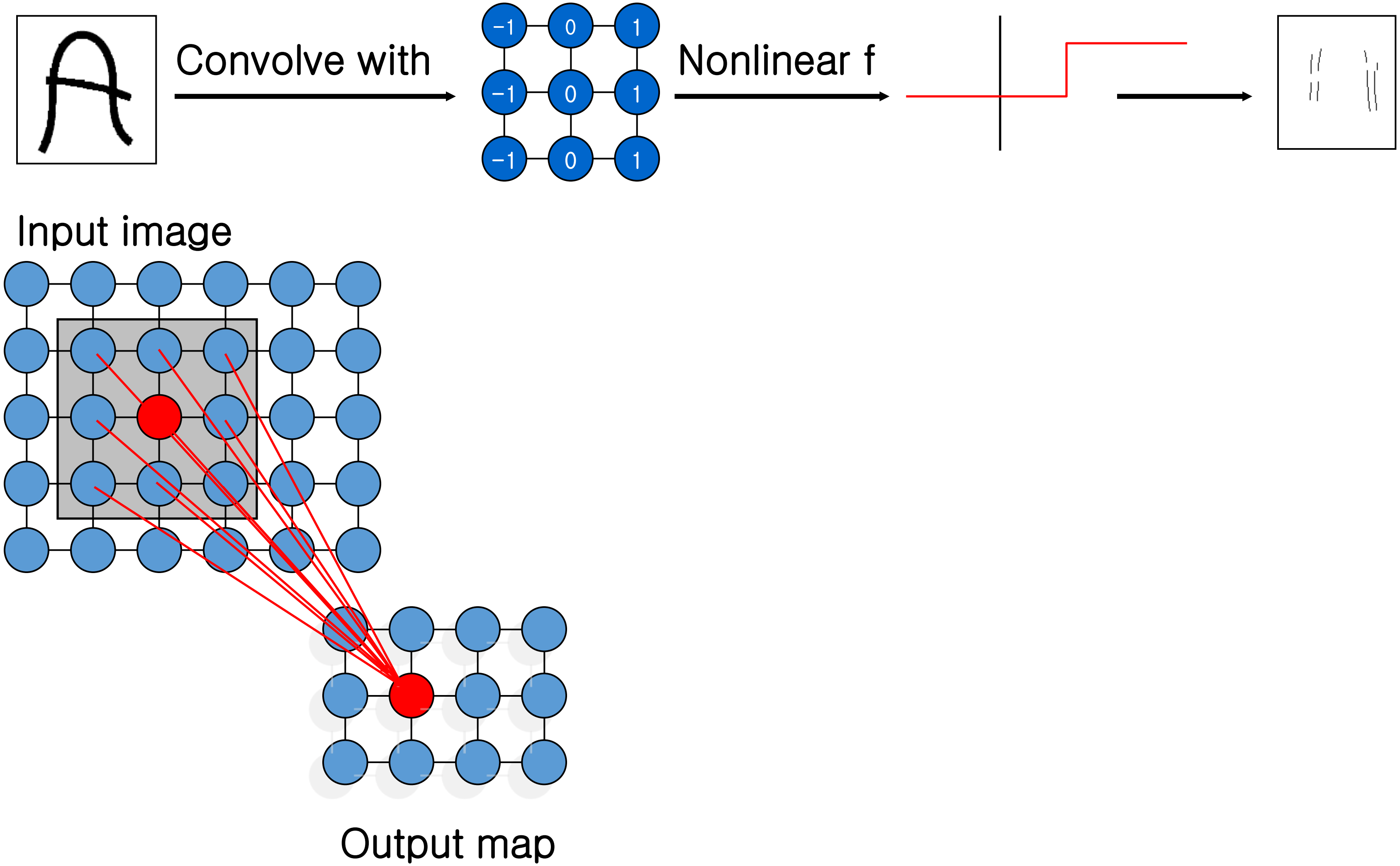
154 input change
from 2 shift left
77 : black to white
77 : white to black

Convolution layer in 2D

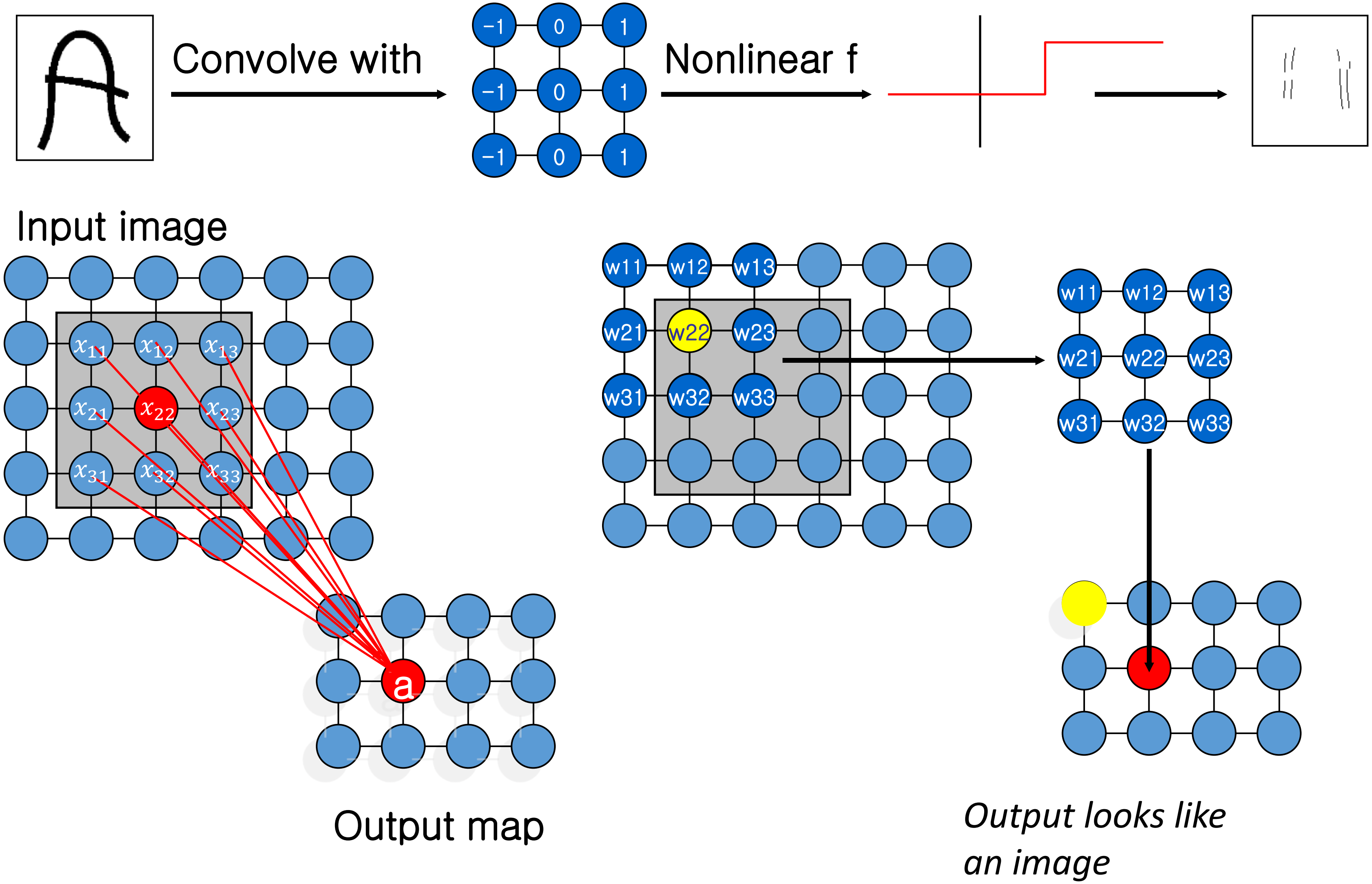
- detect the same feature at different positions in the input, e.g. image



Convolution layer in 2D



Convolution layer in 2D



$$a = f(w_{11}x_{11} + w_{12}x_{12} + w_{13}x_{13} + \cdots w_{33}x_{33})$$

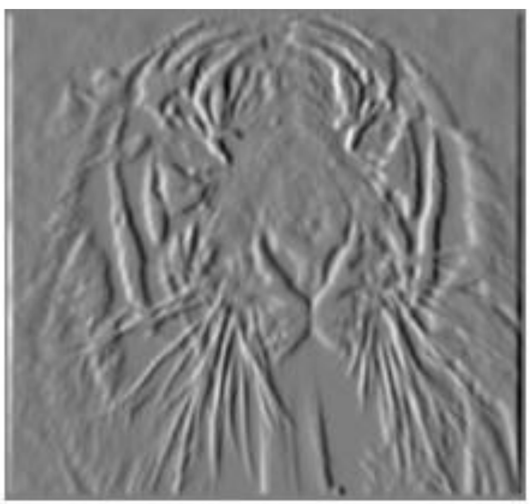
What weights correspond to these output maps?

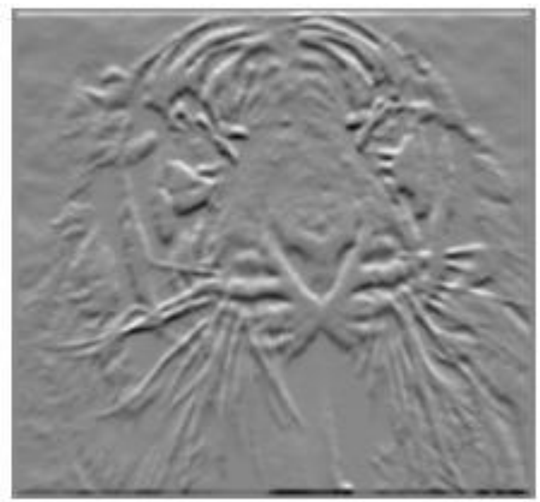
These are output maps before thresholding

Hint: filters look like the input they fire on

$$\frac{\partial f(x,y)}{\partial x}$$

$$\begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}$$



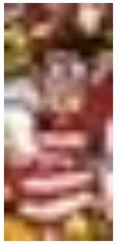


$$\frac{\partial f(x,y)}{\partial y}$$

$$\begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

Where is Waldo?

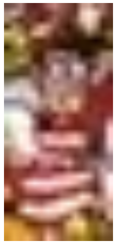
Input



filter

What will the output map look like?

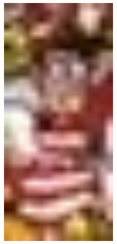
Input



filter

What will the output map look like?

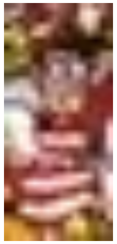
Output



filter

Here is Waldo

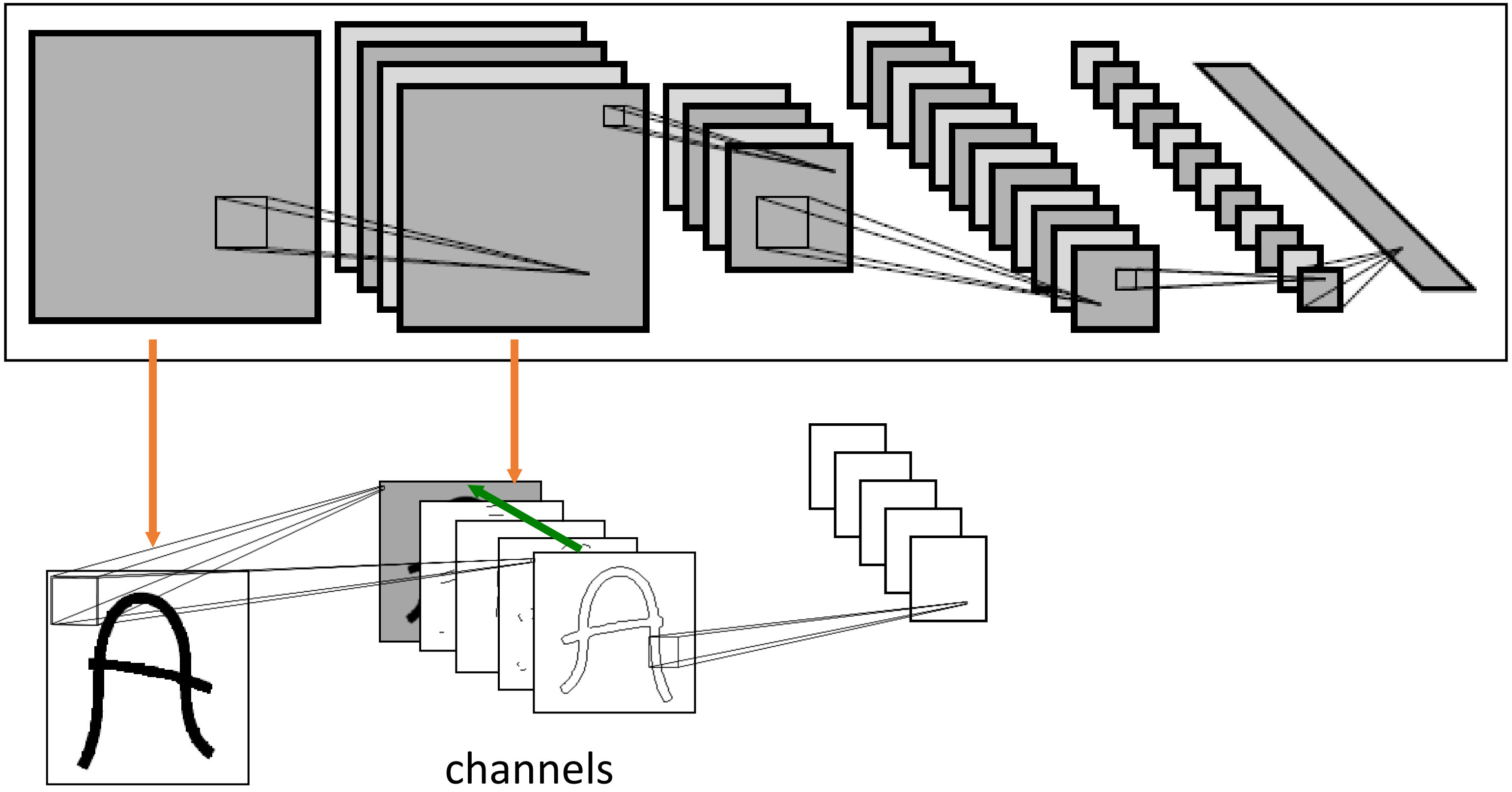
Input



filter

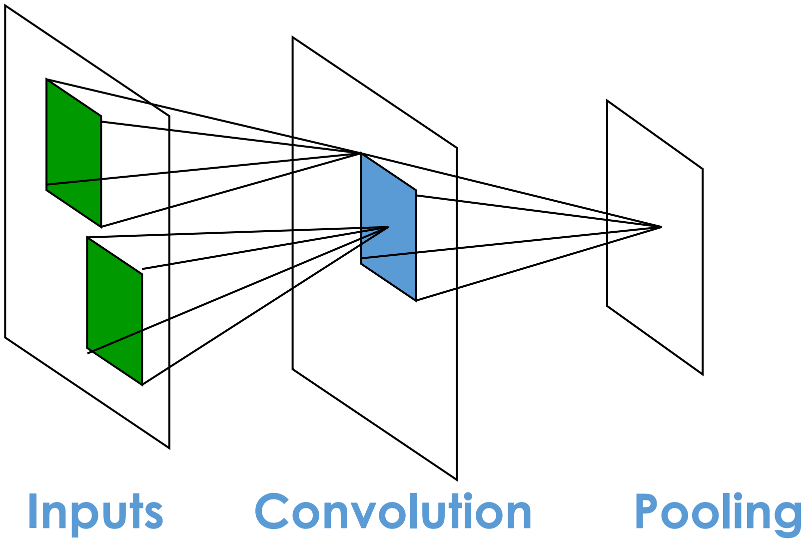
Stacking convolutional layers

- Each layer outputs multi-channel feature maps (like images)
- Next layer learns filters on previous layer's feature maps



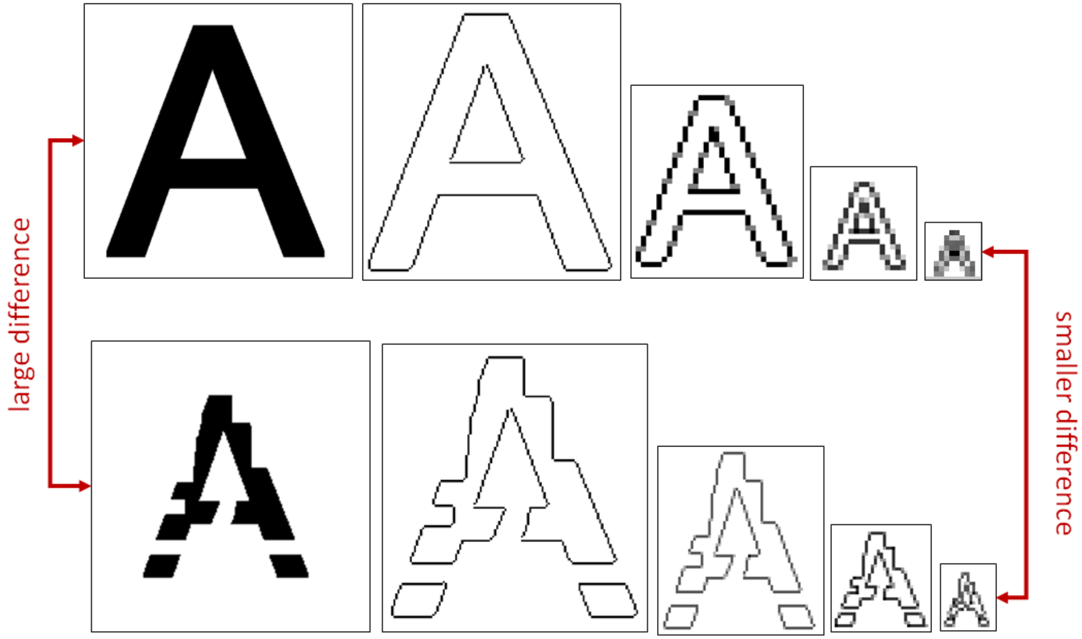
Pooling layers

- Convolution with stride > 1 reduces the size of the input
- Another way to downsize the feature map is with pooling
- A pooling layer subsamples the input in each sub-window
  - max-pooling: chose the max in a window
  - mean-pooling: take the average



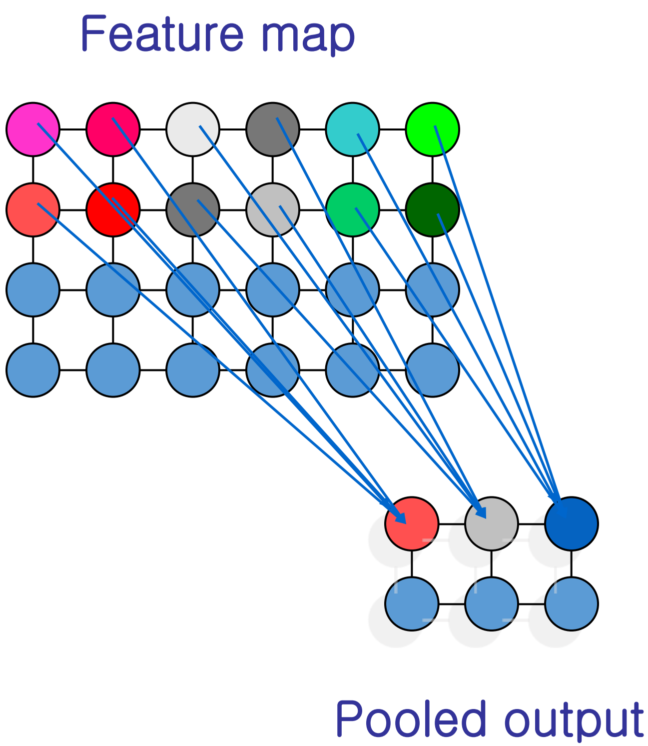
Pooling layer

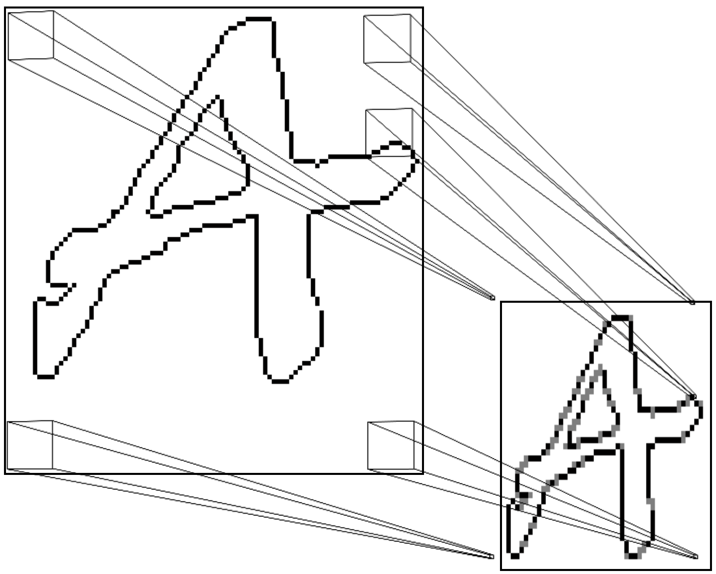
- the pooling layers reduce the spatial resolution of each feature map
- Goal is to get a certain degree of shift and distortion invariance



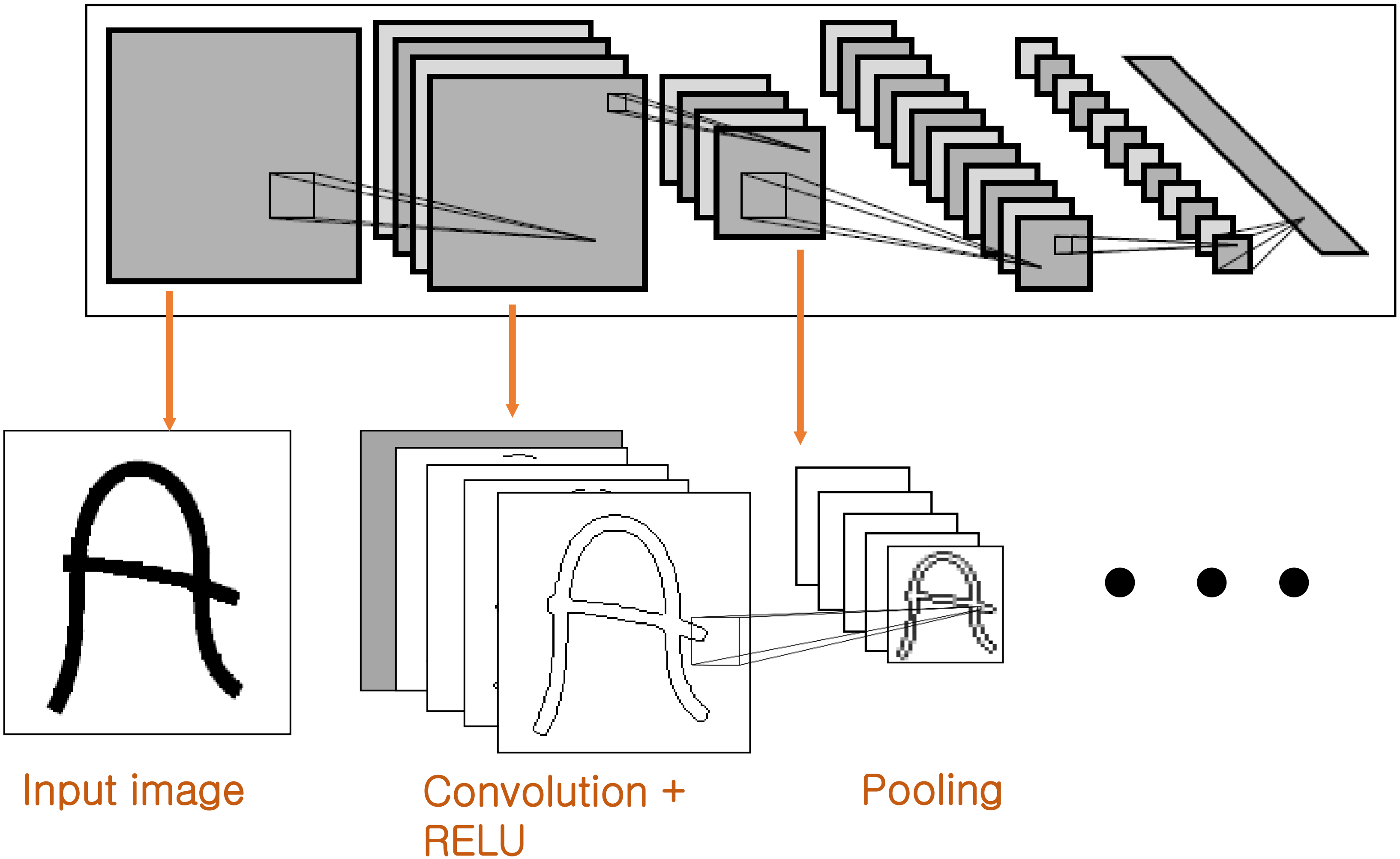
Pooling layer

- the weight sharing is also applied in pooling layers
- for mean/max pooling, no weights are needed



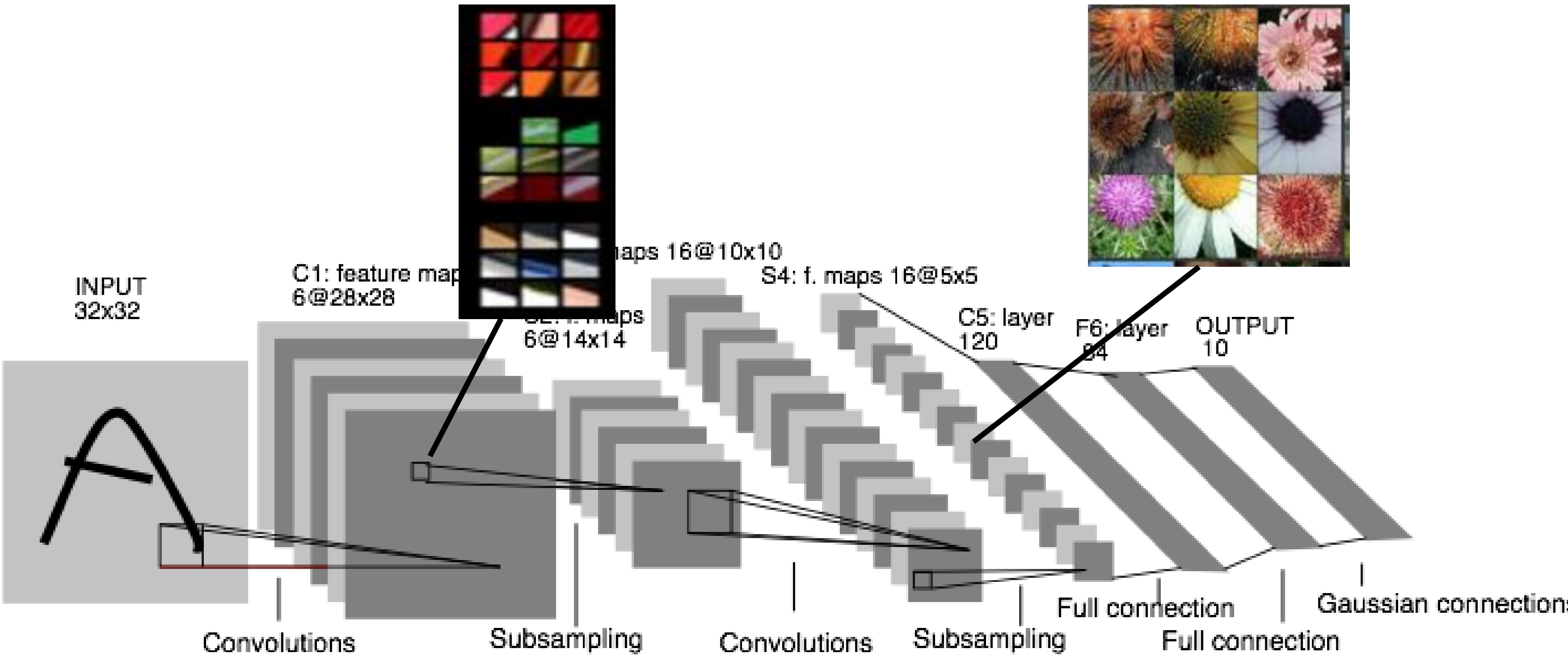


Putting it all together…



Convolutional Neural Network

A CNN is a better architecture for 2D signals



LeNet

Convolutional Neural Nets

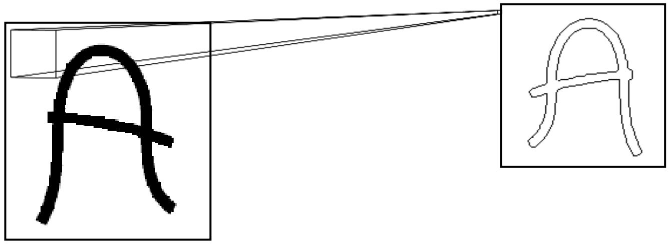
Why they rule

Why CNNs rule: Translation invariance

- Output is invariant to translation of input
  - spatial translation for images
  - temporal translation for time sequences

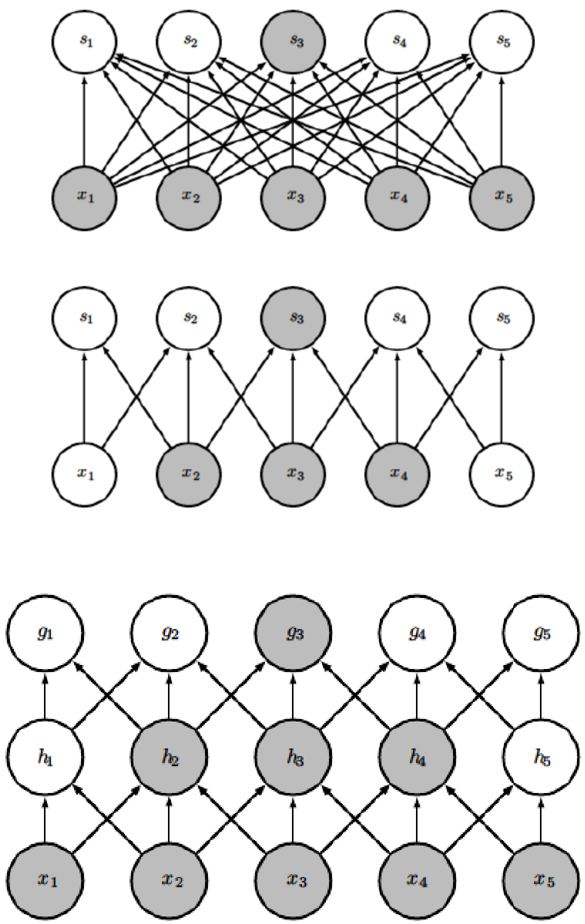
- Note, not invariant to other transformations of input, such as large image rotation

- Pooling provides additional invariance to distortions



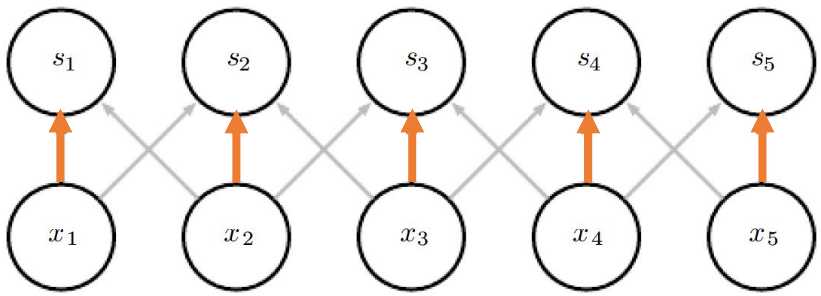
Why CNNs rule: Sparsity

- CNNs have sparse interactions, because the kernel is smaller than the input
- E.g. in thousands or millions pixel image, can detect small meaningful features such as edges
- Very efficient computation!
  - For m inputs and n outputs, matrix multiplication requires $O(m \times n)$ runtime (per example)
  - For k connections to each output, need only $O(k \times n)$ runtime
- Deep layers have larger effective inputs, or receptive fields

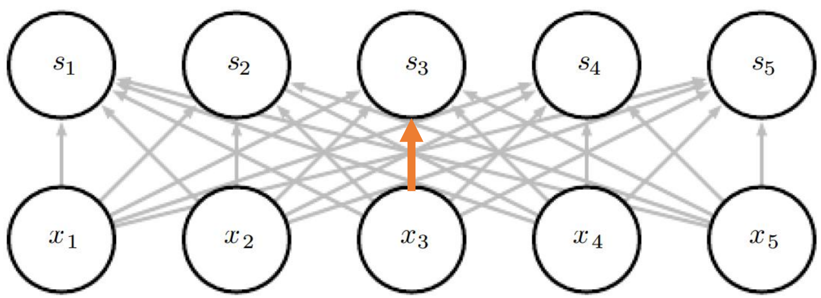


Why CNNs rule: Parameter sharing

- Kernel weights are shared across all locations
- Statistically efficient – learn from more data
- Memory efficient – store only *k* parameters, since *k<<m,* this is much smaller than *m×n*.



parameter used multiple times



parameter used only once

Alex Krizhevsky

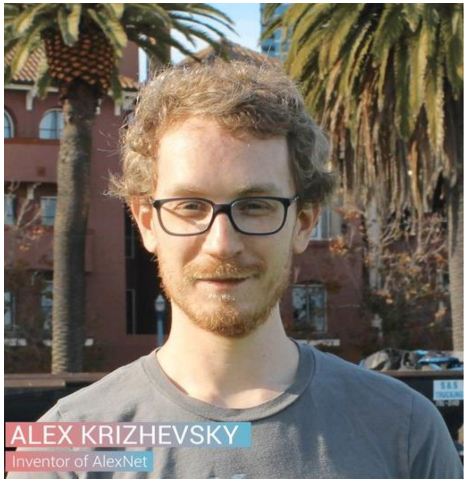
Alex Krizhevsky

Dessa
Verified email at dessa.com
Machine Learning

FOLLOW

TITLE	CITED BY	YEAR
Imagenet classification with deep convolutional neural networks A Krizhevsky, I Sutskever, GE Hinton Advances in neural information processing systems 25, 1097-1105	81373	2012
Dropout: a simple way to prevent neural networks from overfitting N Srivastava, G Hinton, A Krizhevsky, I Sutskever, R Salakhutdinov The journal of machine learning research 15 (1), 1929-1958	26336	2014
Learning multiple layers of features from tiny images A Krizhevsky, G Hinton		

Hence the name ***AlexNet***



Why Now?

Neural Networks date back decades, so why the resurgence?

1952 Stochastic Gradient Descent

1958 Perceptron
- Learnable Weights

⋮

1986 Backpropagation
- Multi-Layer Perceptron

1995 Deep Convolutional NN
- Digit Recognition

⋮

1. Big Data
- Larger Datasets
- Easier Collection & Storage

IMAGENET

WIKIPEDIA
The Free Encyclopedia

2. Hardware
- Graphics Processing Units (GPUs)
- Massively Parallelizable

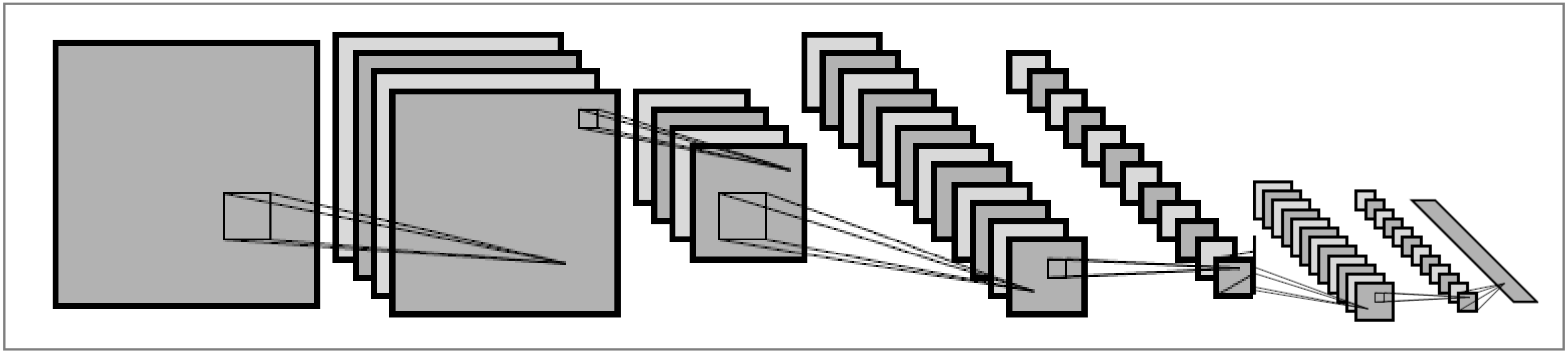
3. Software
- Improved Techniques
- New Models
- Toolboxes

TensorFlow

Convolutional Neural Nets

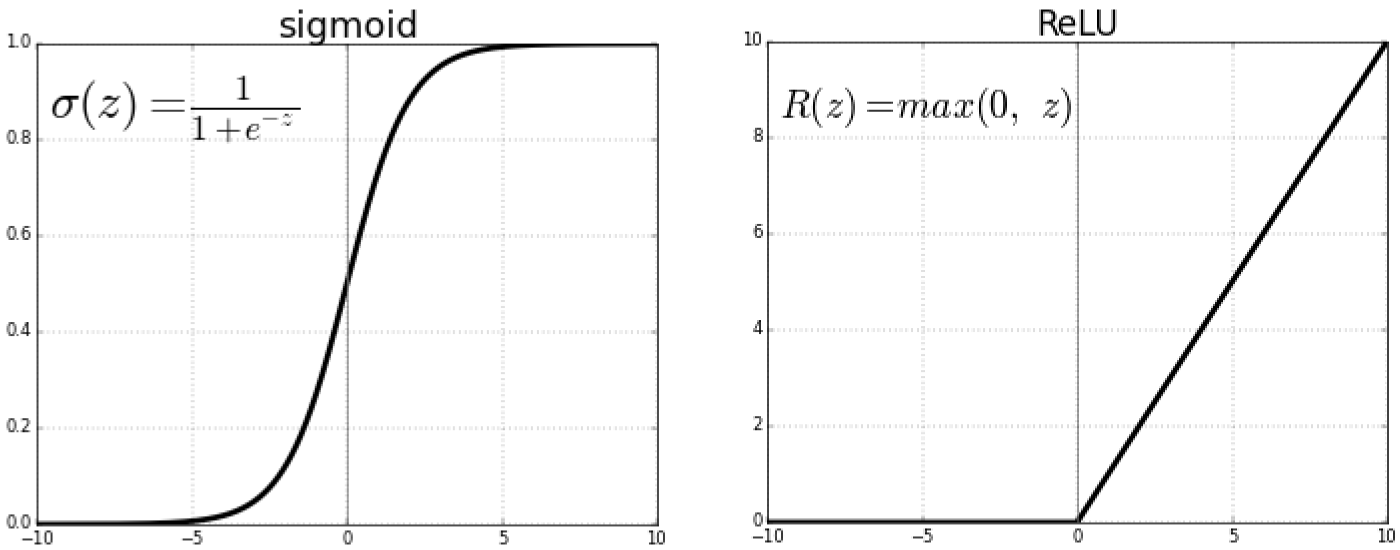
Example

CIFAR-10 Demo ConvJS Network



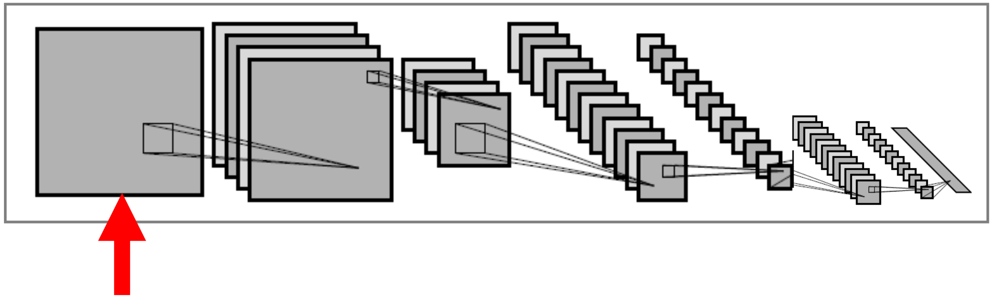
input image conv RELU pool conv RELU pool conv RELU pool fc

RELU: rectified linear unit

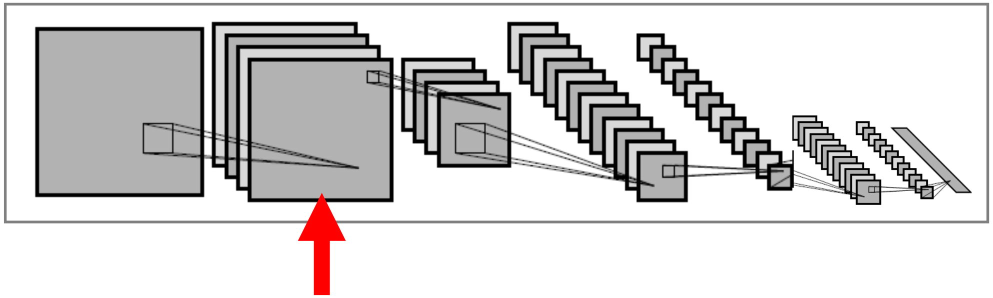


RELU function $g(x) = \max(0, x)$

filter size 5x5x3, stride 1



input (32x32x3)

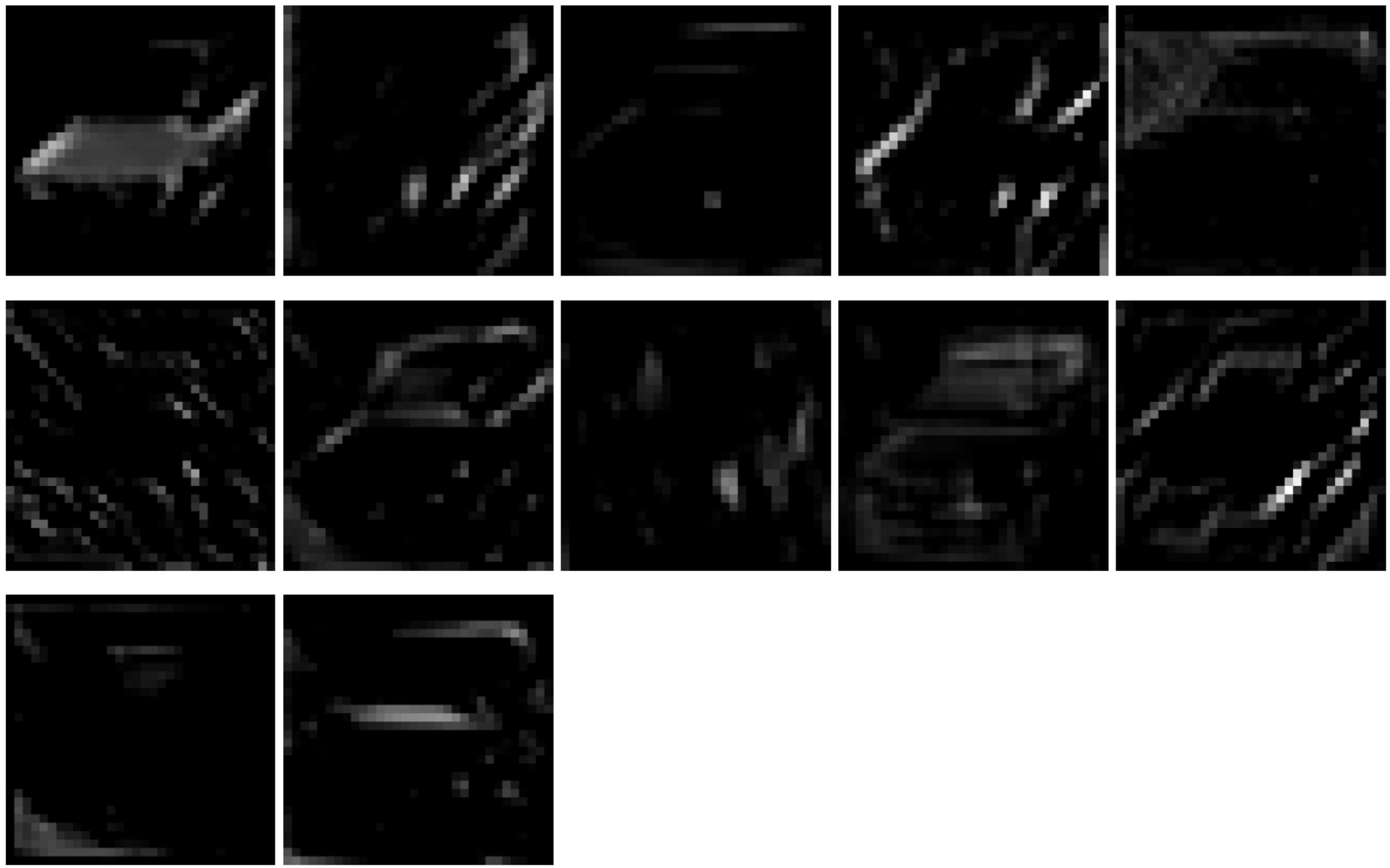


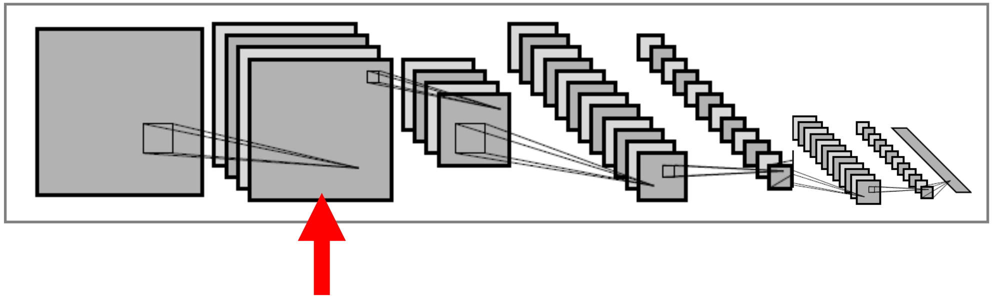
filter size 5x5x3, stride 1

input (32x32x3)

RELU

conv (32x32x16) params: 16x5x5x3+16 = 1216

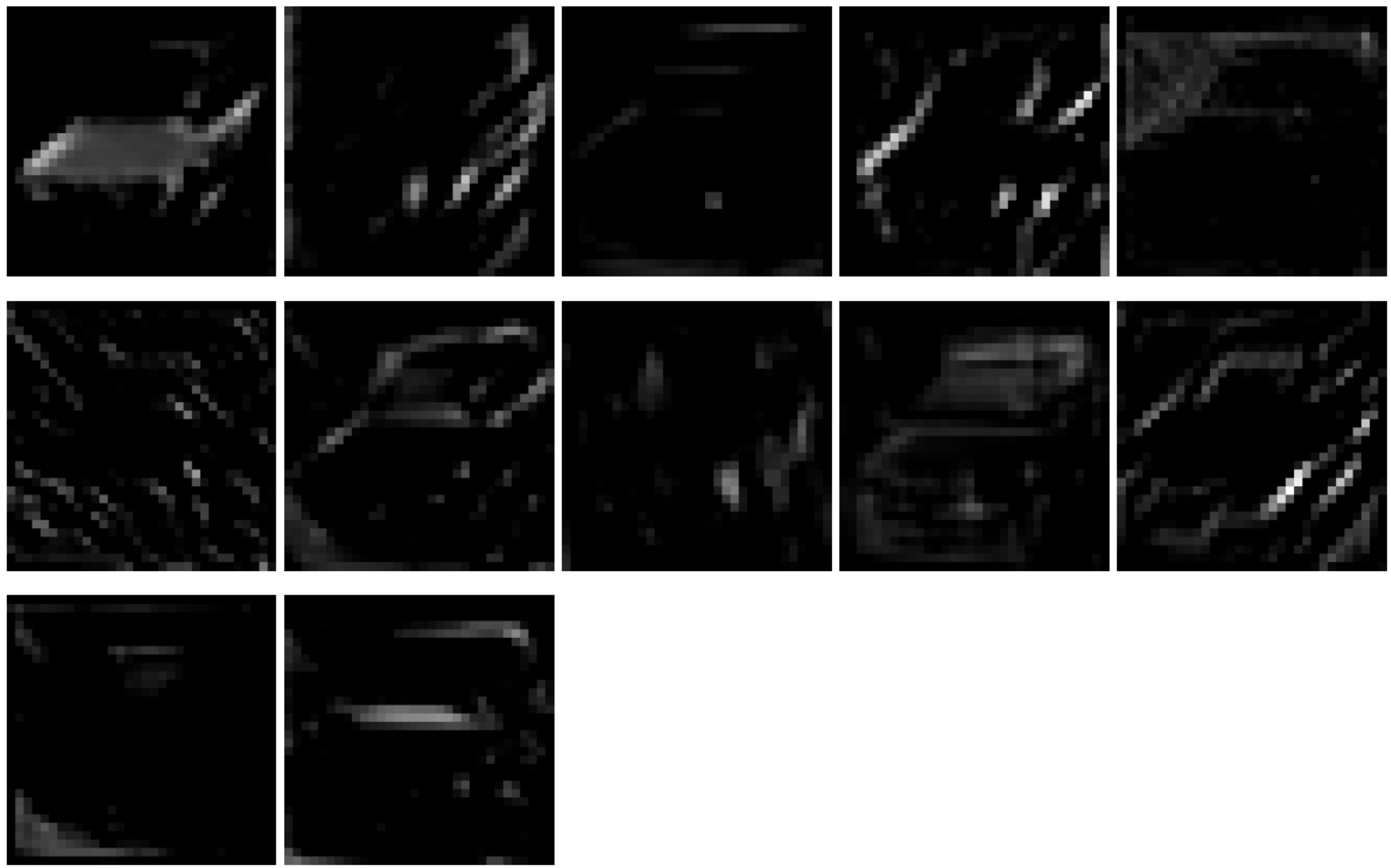


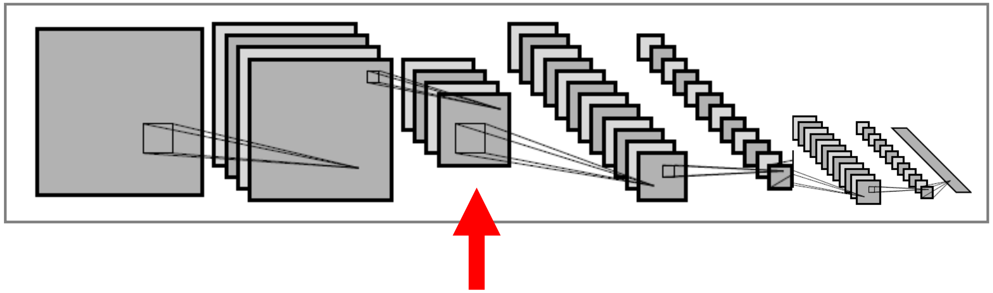


filter size 5x5x3, stride 1

input (32x32x3)

conv (32x32x16) params: 16x5x5x3+16 = 1216

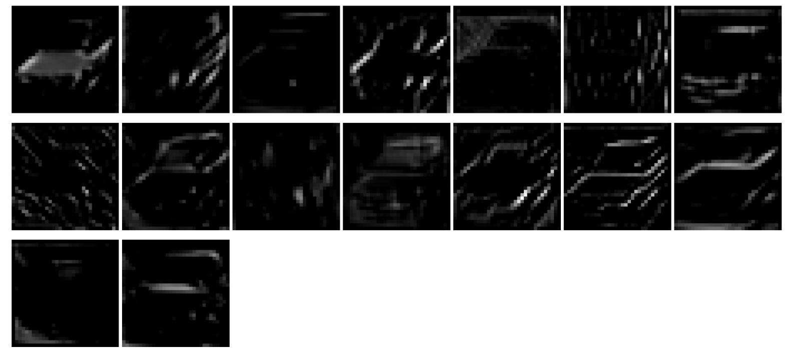




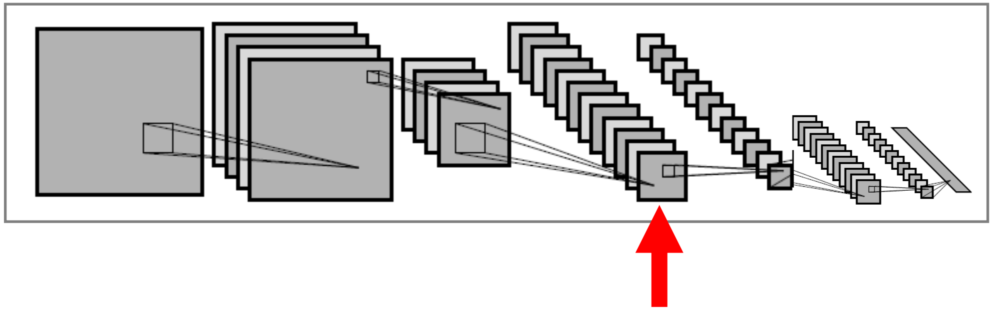
filter size 5x5x3, stride 1

input (32x32x3)

conv (32x32x16) params: 16x5x5x3+16 = 1216



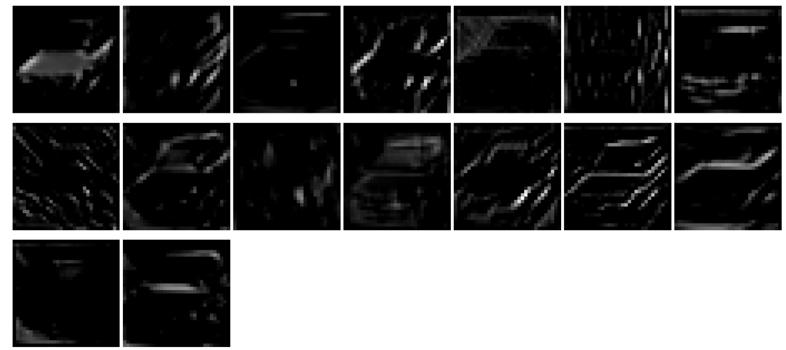
pool (16x16x16)
pooling size 2x2, stride 2



filter size 5x5x3, stride 1

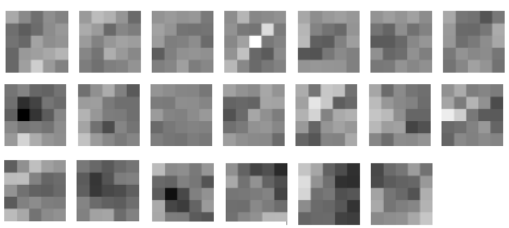
input (32x32x3)

conv (32x32x16) params: 16x5x5x3+16 = 1216



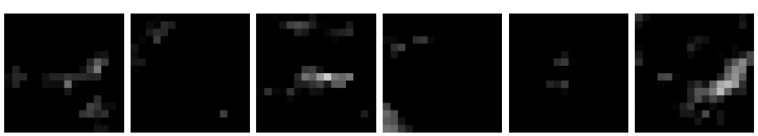
pool (16x16x16)
pooling size 2x2, stride 2

filter size 5x5x16, stride 1

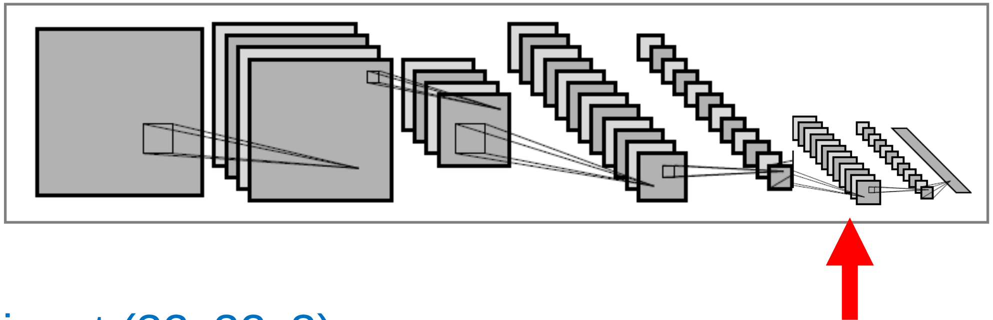


RELU

conv (16x16x20) params: 20x5x5x16+20 = 8020



pool (8x8x20)
pooling size 2x2, stride 2



One more conv+RELU+pool:

conv (8x8x20)
filter size 5x5x20, stride 1
relu (8x8x20)
pool (4x4x20)
pooling size 2x2, stride 2

input (32x32x3)

fc (1x1x10); parameters: 10x320+10 = 3210

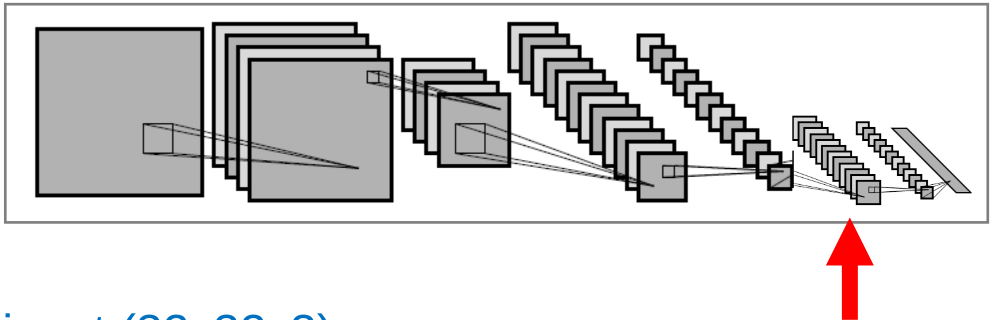
softmax (1x1x10)

Dog car Cat ⋮

Softmax

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

$\vec{z}$	The input vector to the softmax function, made up of (z0, ... zK)
z_i	All the zi values are the elements of the input vector to the softmax function, and they can take any real value, positive, zero or negative. For example a neural network could have output a vector such as (-0.62, 8.12, 2.53), which is not a valid probability distribution, hence why the softmax would be necessary.
e^{z_i}	The standard exponential function is applied to each element of the input vector. This gives a positive value above 0, which will be very small if the input was negative, and very large if the input was large. However, it is still not fixed in the range (0, 1) which is what is required of a probability.
$\sum_{j=1}^{K} e^{z_j}$	The term on the bottom of the formula is the normalization term. It ensures that all the output values of the function will sum to 1 and each be in the range (0, 1), thus constituting a valid probability distribution.
K	The number of classes in the multi-class classifier.



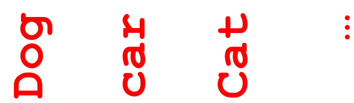
One more conv+RELU+pool:

conv (8x8x20)
filter size 5x5x20, stride 1
relu (8x8x20)
pool (4x4x20)
pooling size 2x2, stride 2

input (32x32x3)

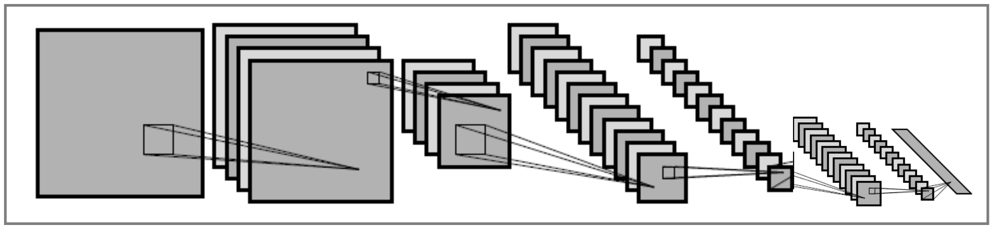
fc (1x1x10); parameters: 10x320+10 = 3210

softmax (1x1x10)



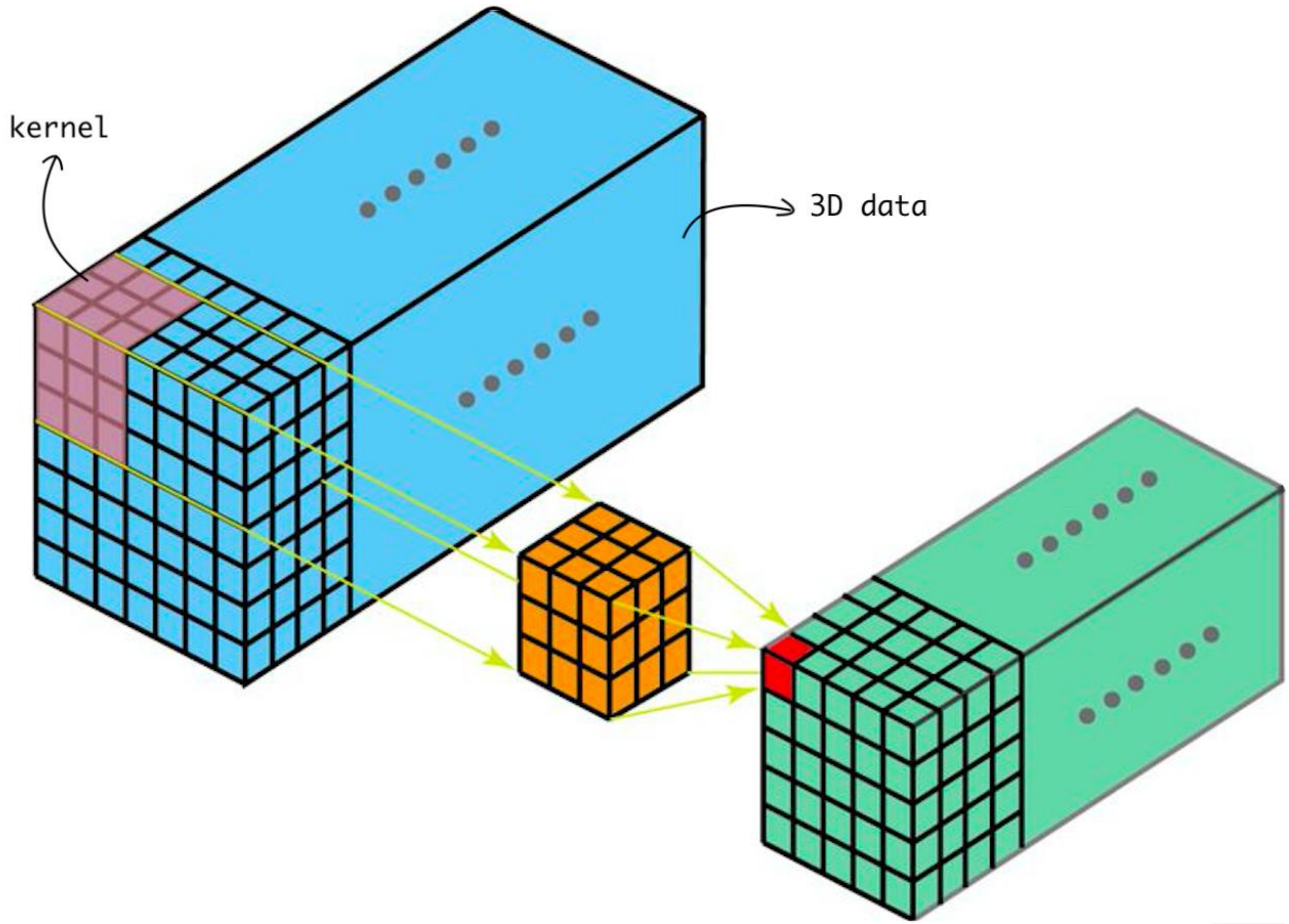
Testing the network

- Show top three most likely classes



http://cs.stanford.edu/people/karpathy/convnetjs/demo/cifar10.html

3D Convolutional Neural Networks



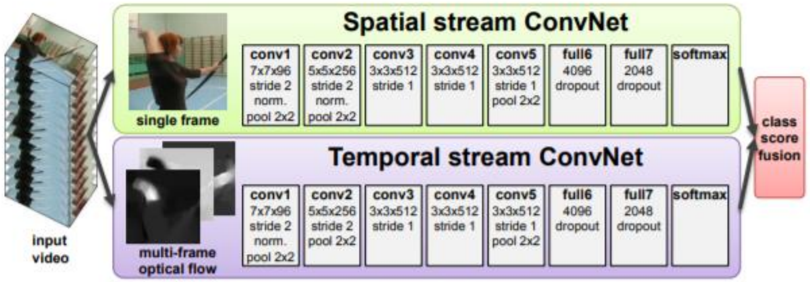
https://towardsdatascience.com/understanding-1d-and-3d-convolution-neural-network-keras-9d8f76e29610

Application:
AI Generated Match Highlights

- IBM's produce the official match highlights of Wimbledon and US Open tennis tournaments.

- https://www.usopen.org/en_US/video/2017-08-31/1504233424.html

- Multi-modal System



- Bias Considerations