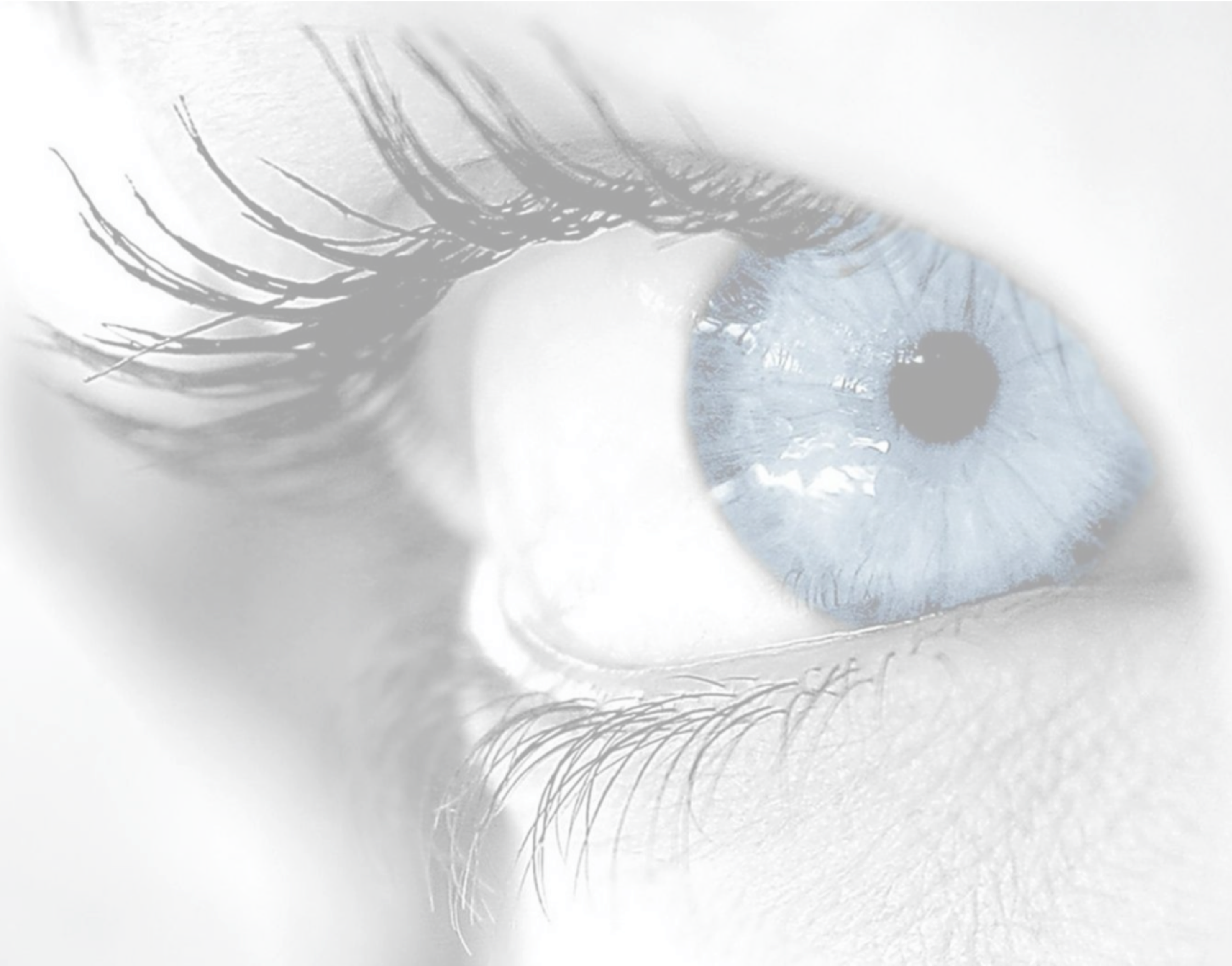
Deep Learning for Computer Vision

Vision:

Fundamental aspect of our life

Origin: 540 mln years ago

Our visual system is trained on images seen in 540 mln of years!

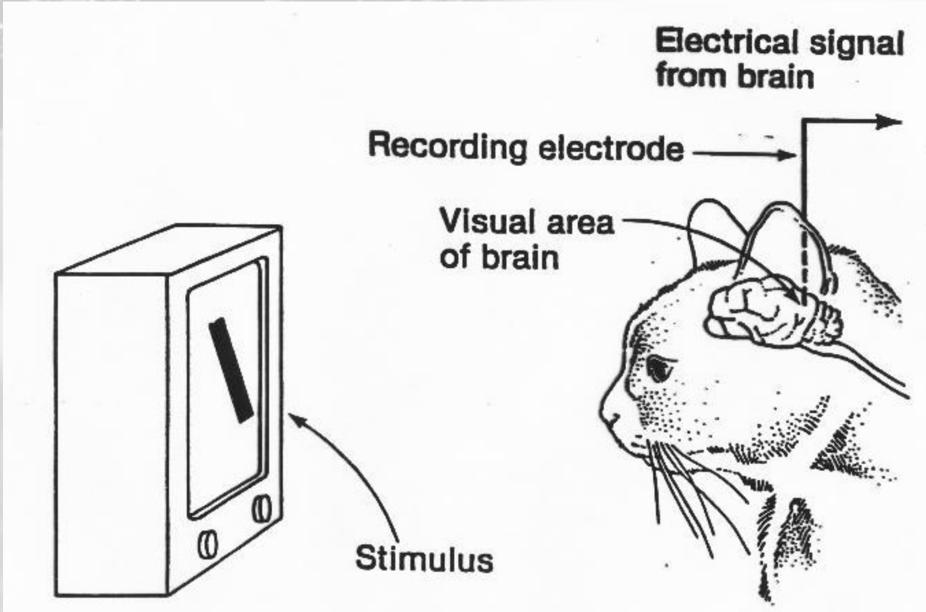


Vision:

1950s: **David Hubel and Torsten Wiesel**[1] experiments on the visual cortex of a cat

Findings:

- Certain neurons fire only to very specific patterns and orientation
- Neural mechanisms are spatio-invariant
- Neural layers are organized hierarchically



(1) "*[Receptive fields of single neurons in the cat's striate cortex]()*"

Artificial Intelligence Vision:

The very first picture…

1826

"View from the Window at Le Gras", France
taken by Nicéphore Niepce

Obtained covering of bitumen a sheet
Time of exposure: 8 hours!!!

Artificial Intelligence Vision:

Origins of computer vision:
an MIT undergraduate <u>summer project</u>

1966

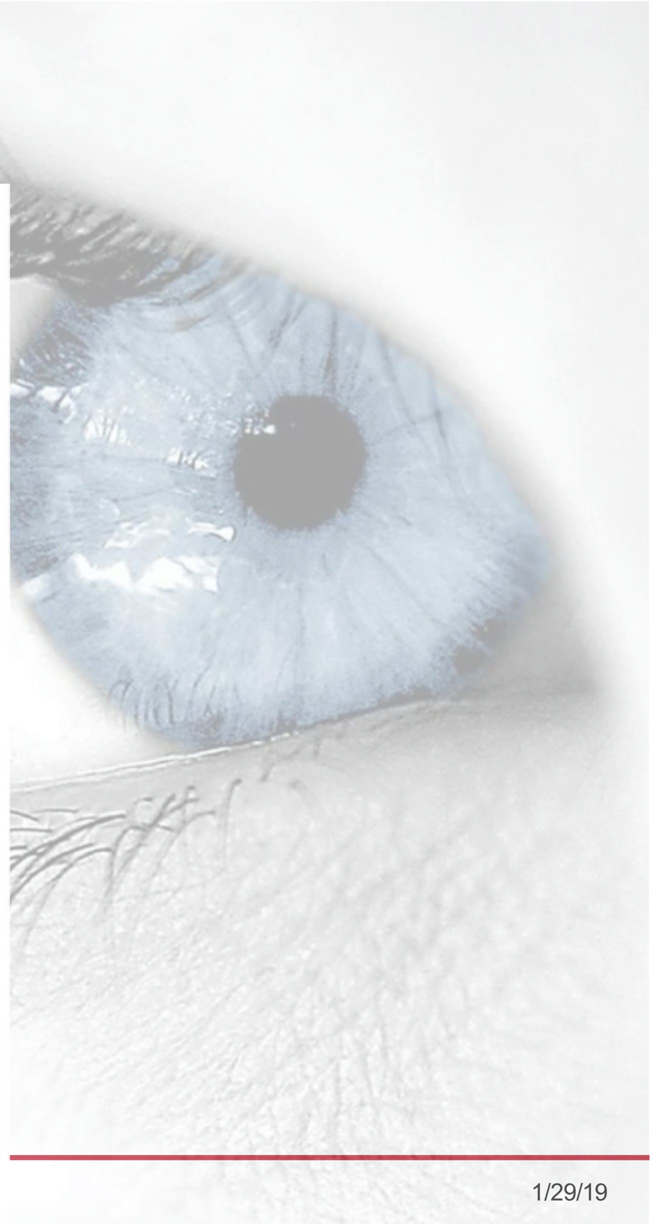
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
PROJECT MAC

Artificial Intelligence Group — July 7, 1966
Vision Memo. No. 100.

THE SUMMER VISION PROJECT

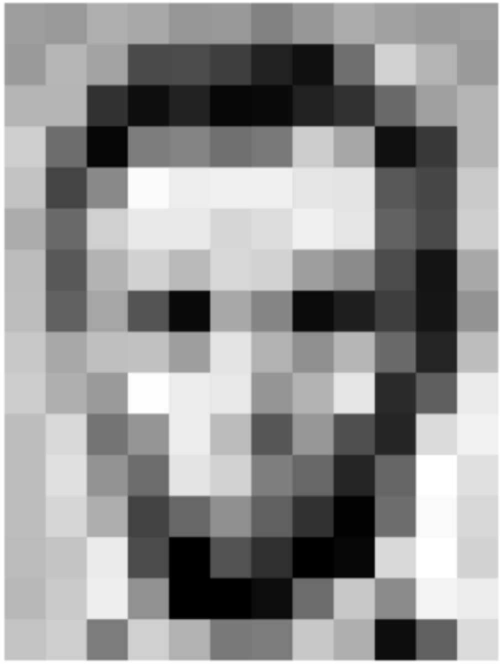
Seymour Papert

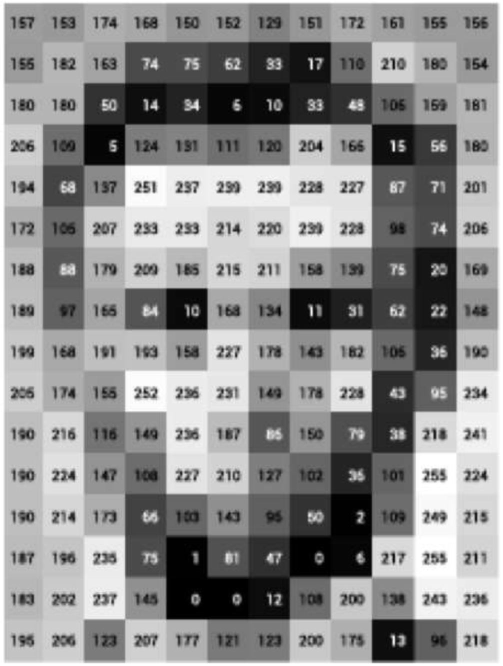
The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".



What Computers "See"

Images are Numbers





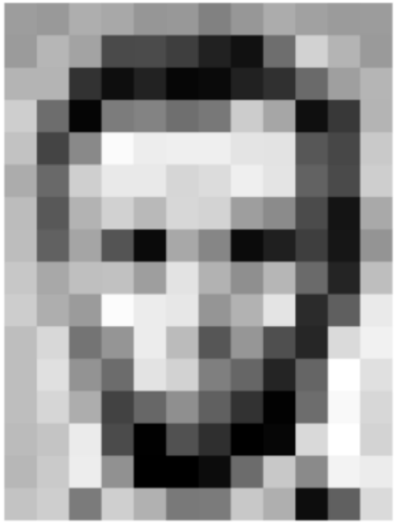
What the computer sees

157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

An image is just a matrix of numbers [0,255]!
i.e., 1080x1080x3 for an RGB image

Tasks in Computer Vision

Classification



Input Image

157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

Pixel Representation

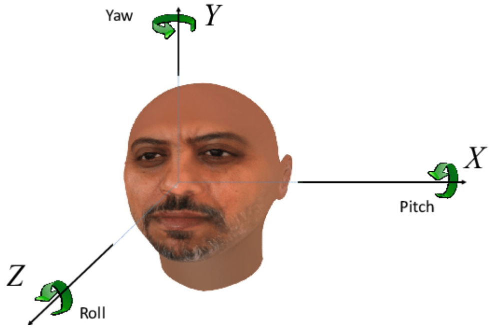
classification

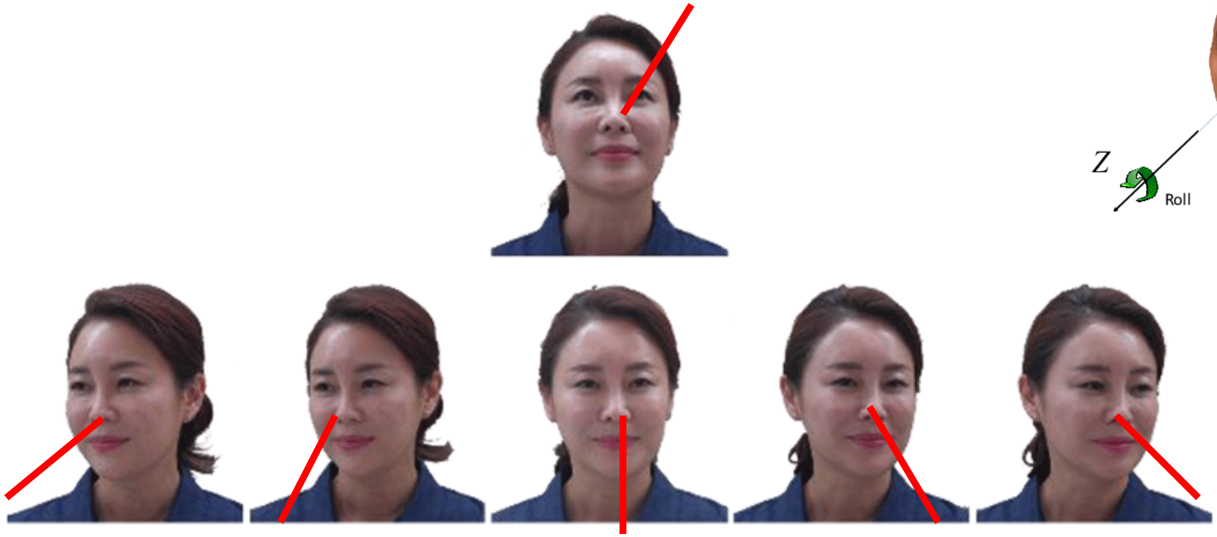
Lincoln	0.8
Washington	0.1
Jefferson	0.05
Obama	0.05

- **Classification**: output variable takes class label. Can produce probability of belonging to a particular class
- In the example, classify which USA President is in the image

Tasks in Computer Vision

Regression





- **Regression**: output takes a continuous value
- In the example, the head rotation (3 angles: Yaw, Pitch, Roll)

High Level Feature Detection

CLASSIFICATION: identify key features in each image category

Nose ,
Eyes,
Mouth

Wheels,
License Plate,
Headlights

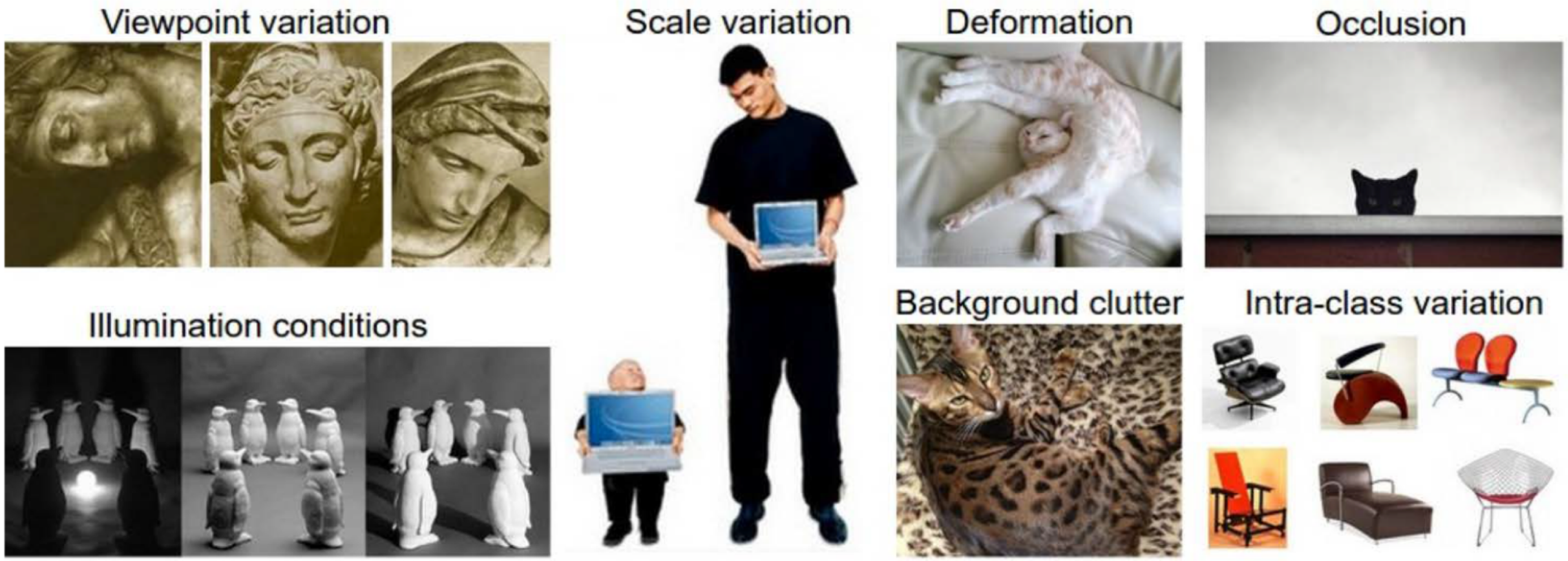
Door,
Windows,
Steps

Manual Feature Extraction

Problems?

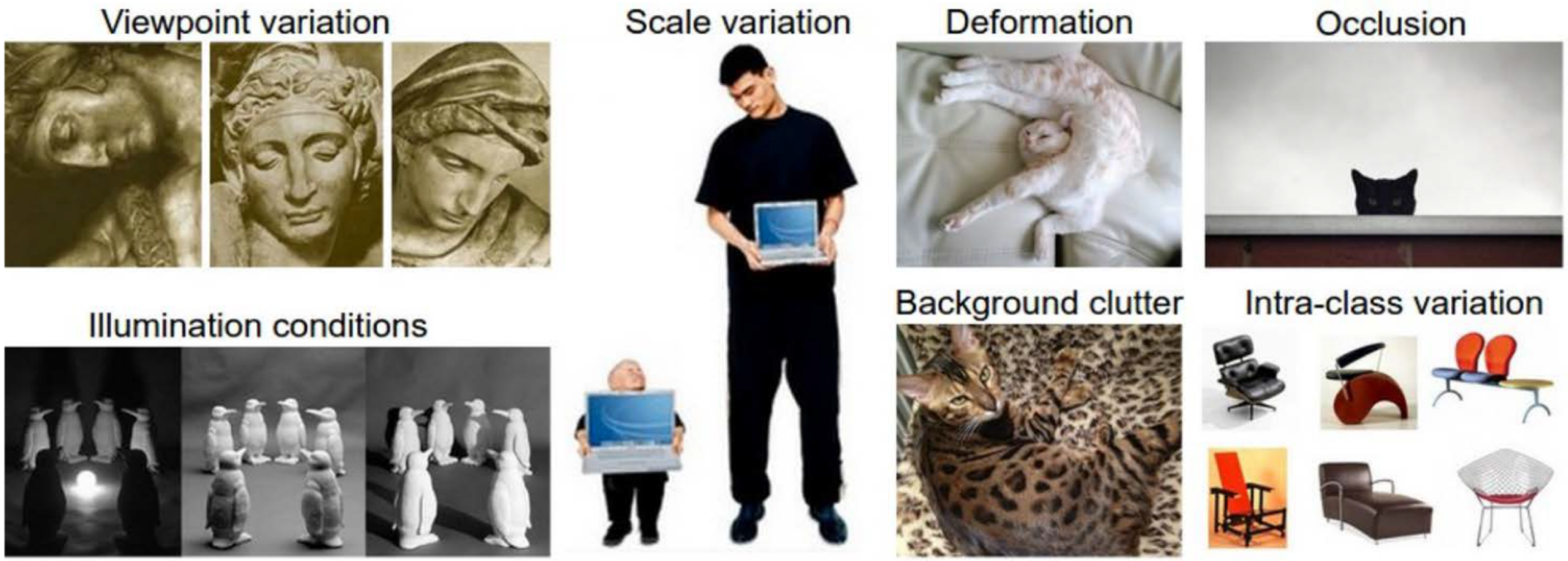
Manual Feature Extraction

Domain knowledge → Define features → Detect features to classify



Manual Feature Extraction

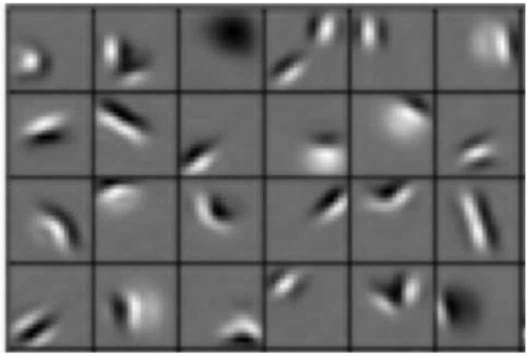
Domain knowledge → Define features → **Detect features to classify**



Learning Feature Representations

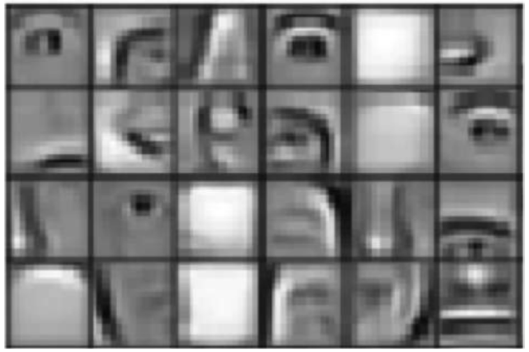
Can we learn a **hierarchy of features** directly from the data instead of hand engineering?

Low level features



Edges, dark spots

Mid level features



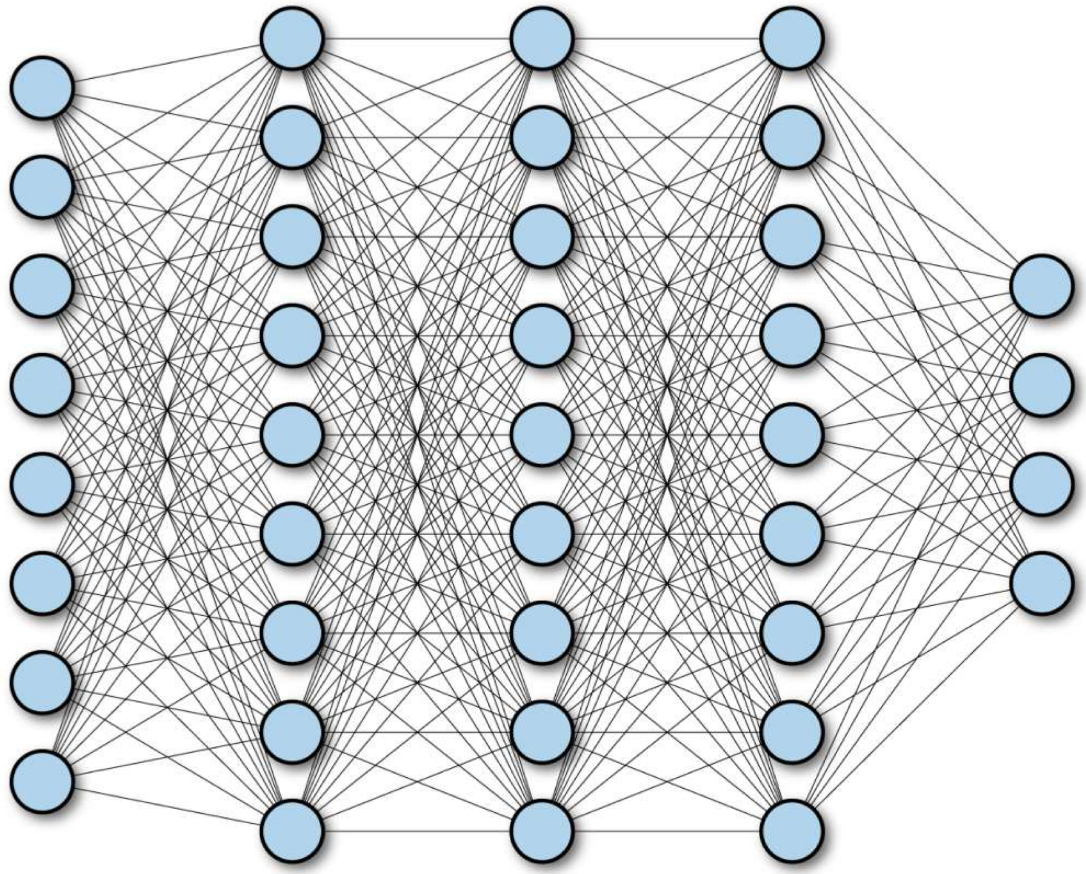
Eyes, ears, nose

High level features

Facial structure

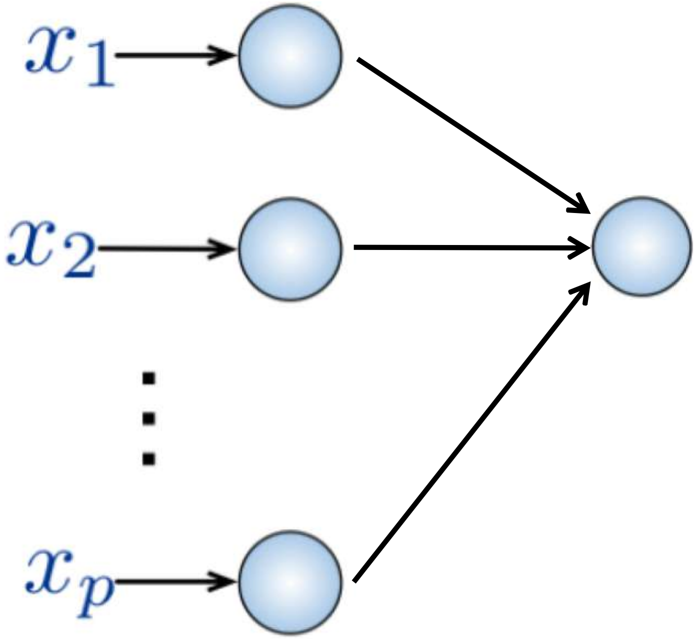
Learning Visual Features

Fully Connected Neural Network



Fully Connected Neural Network

Input:
2D image
Vector of pixel values
$x_1, x_2, \ldots, x_p$
Input to the NN



Fully Connected NN:

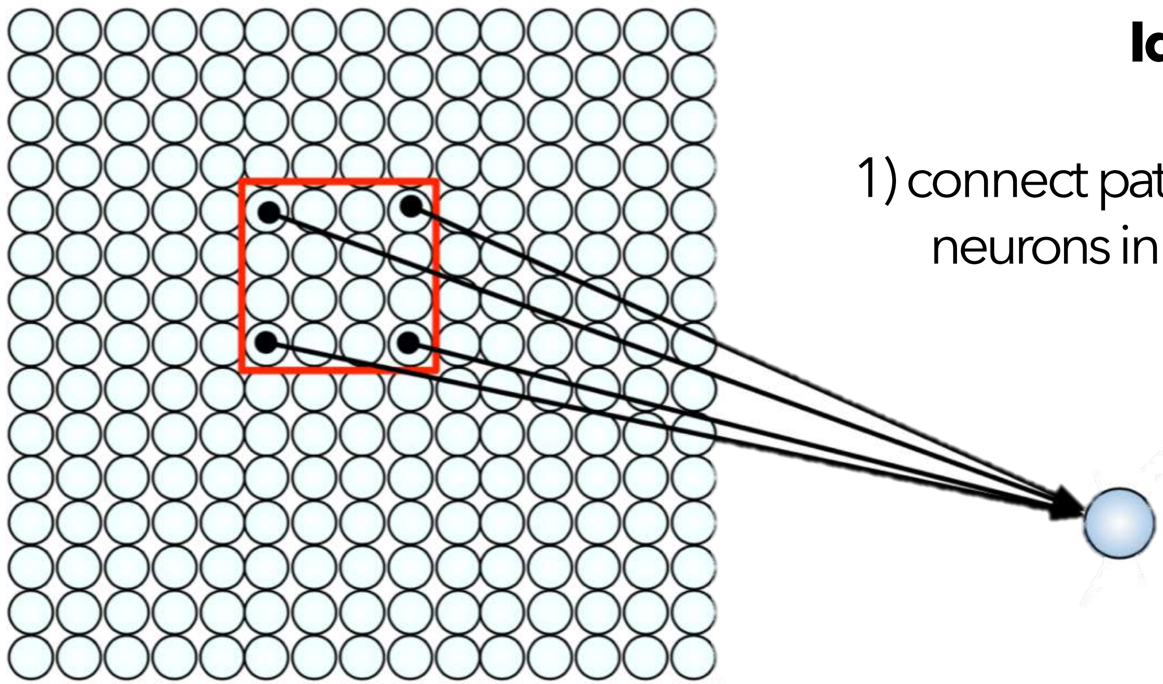
- Connect neuron in hidden layer to all neurons in input layer
- No spatial information!
- And many, many parameters!

How can we use **spatial structure** in the input to inform the architecture of the network?

Spatial features

Using Spatial Structure

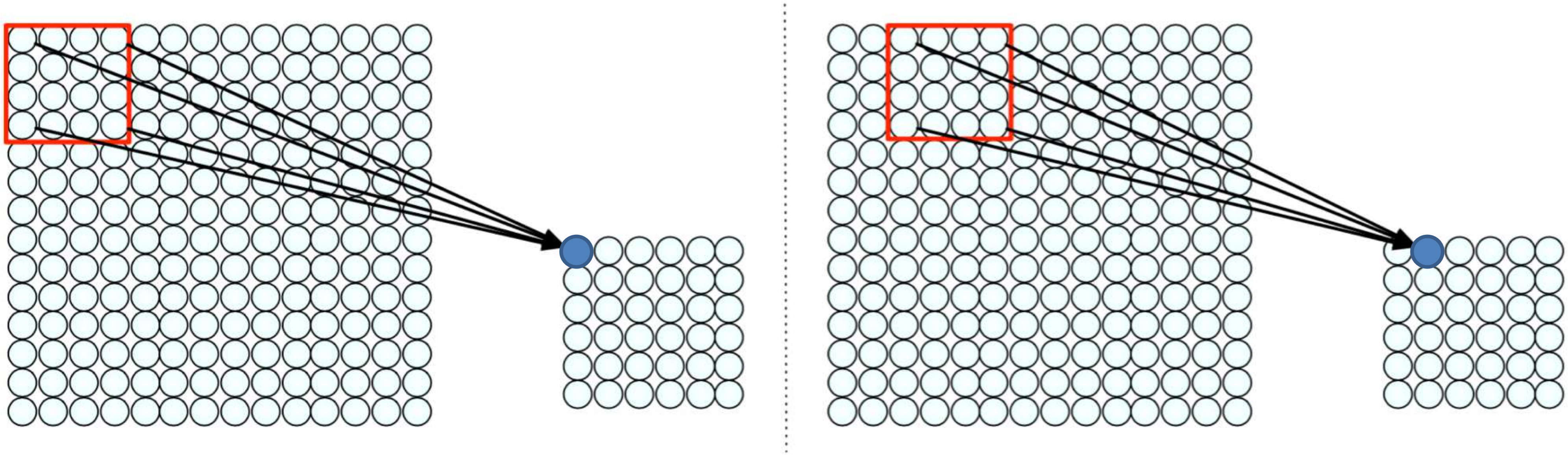
Input: 2D image.
Array of pixel values



Idea:

1) connect patches of input to neurons in hidden layer.

Using Spatial Structure



2) Slide the patch window across the image.

Different **weights (filters)** *detect different features*

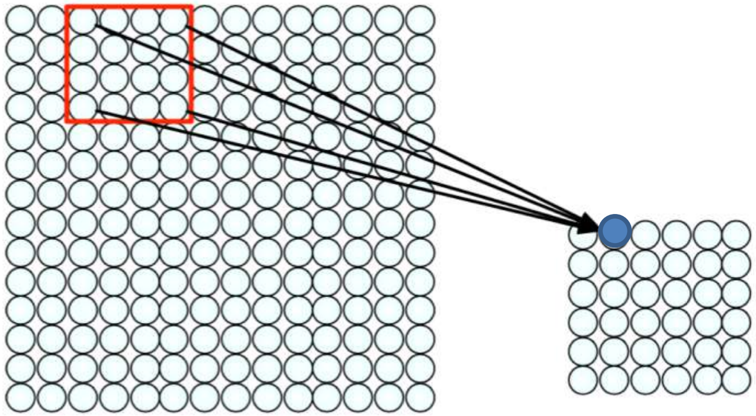
Applying Filters to Extract Features

1) Apply a set of weights – a filter – to extract **local features**

2) Use **multiple filters** to extract different features

3) Spatially **share** parameters of each filter
(features that matter in one part of the input should matter elsewhere)

Spatial Convolution



- Filter of size $(n \times n)$: n^2 different weights
- Apply this same filter to all $(n \times n)$ patches in input
- Not necessary all: shift by s pixels for next patch

Producing Feature Maps

Original

Sharpen

Edge Detect

"Strong" Edge Detect

Different weights (filters) extract different features.
Which weights?

Feature Extraction and Convolution

A Case Study

X or X?

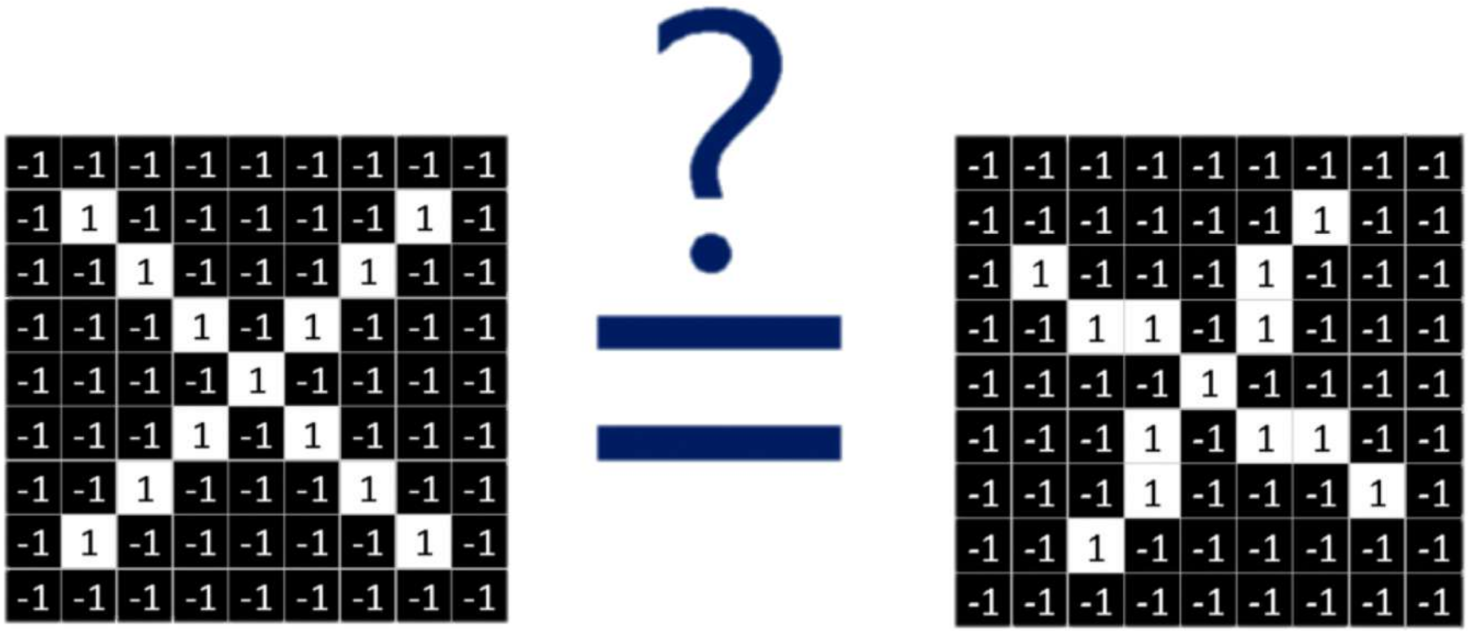
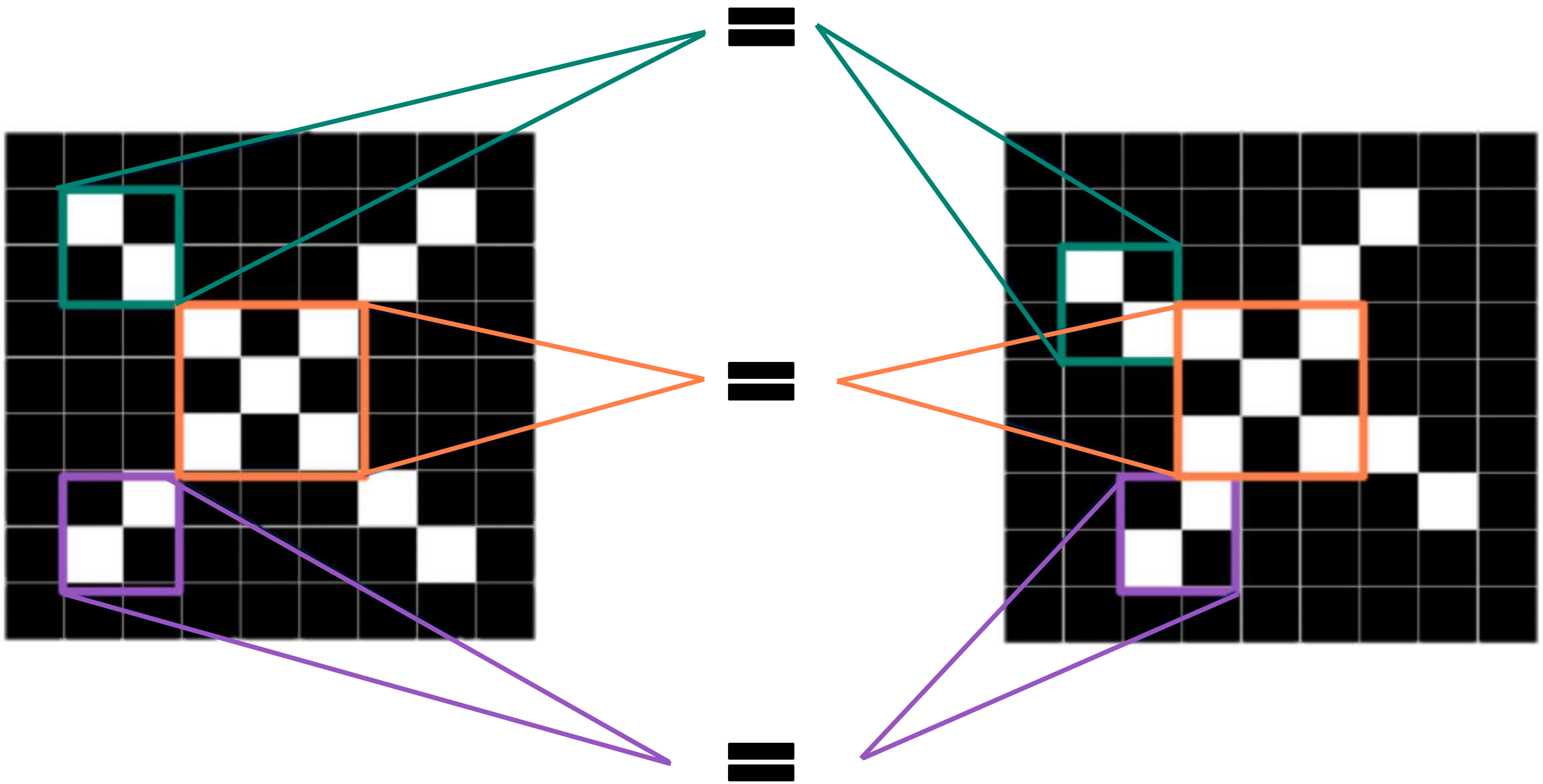


Image is represented as matrix of pixel values... and computers are literal!

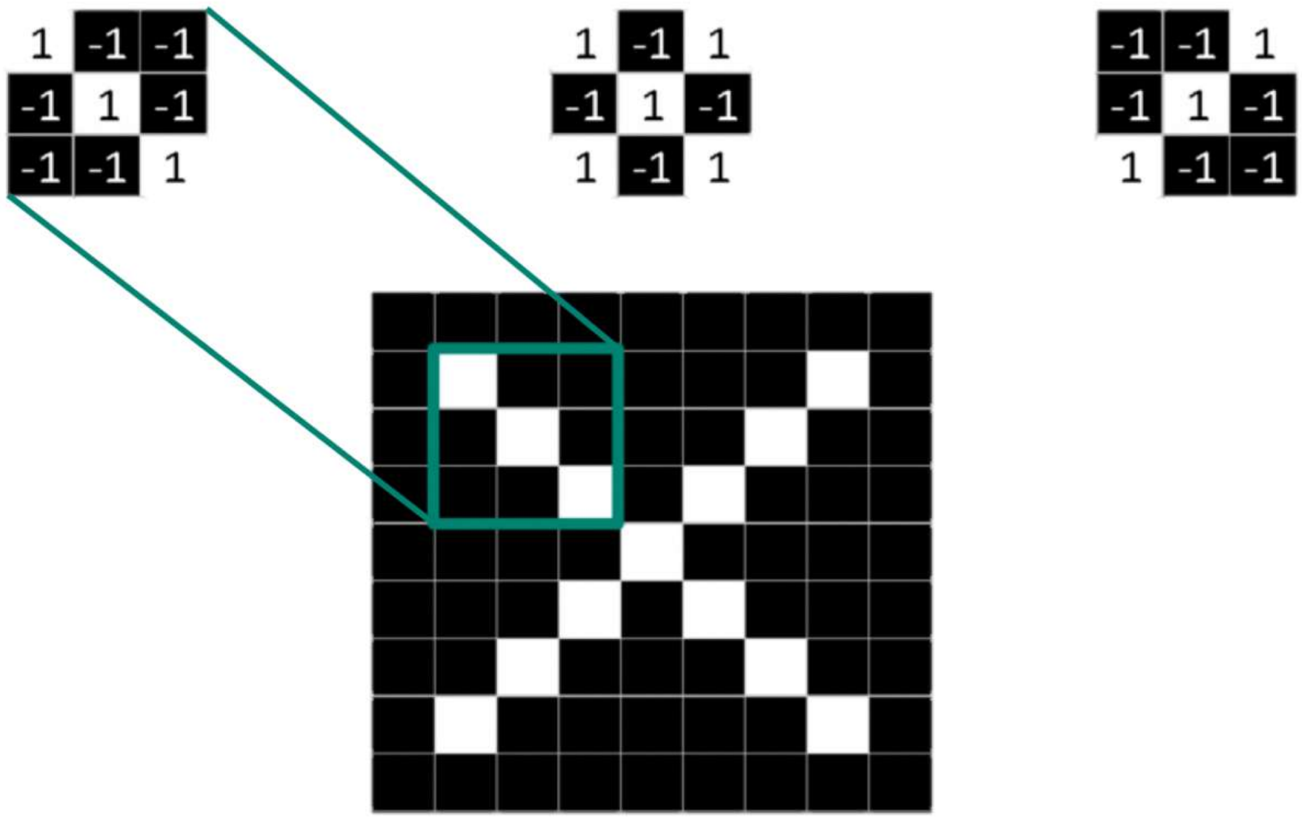
We want to be able to classify an X as an X even if it's shifted, shrunk, rotated, deformed.

Features of X

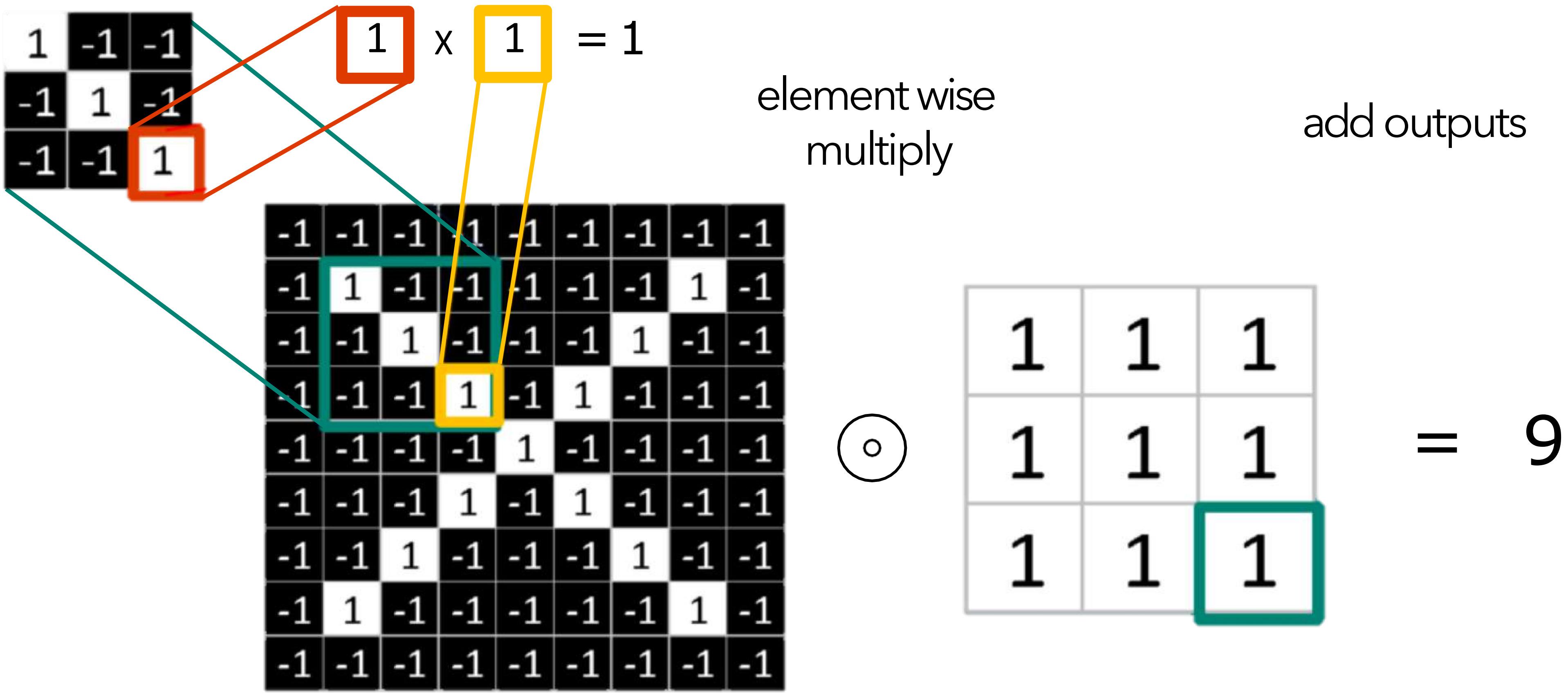


Filters to Detect X Features

filters



The Convolution Operation



The Convolution Operation

Suppose we want to compute the convolution of a 5x5 image and a 3x3 filter:

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

image

$\otimes$

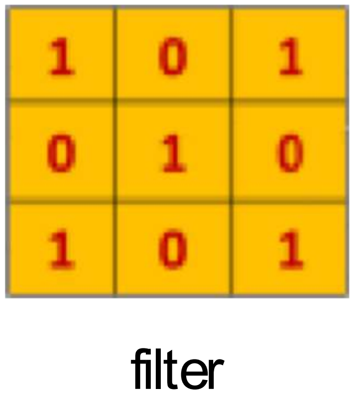
1	0	1
0	1	0
1	0	1

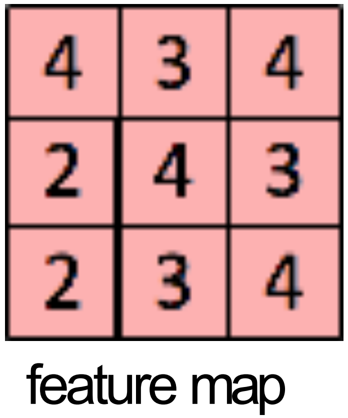
filter

We slide the 3x3 filter over the input image, element-wise multiply, and add the outputs…

The Convolution Operation

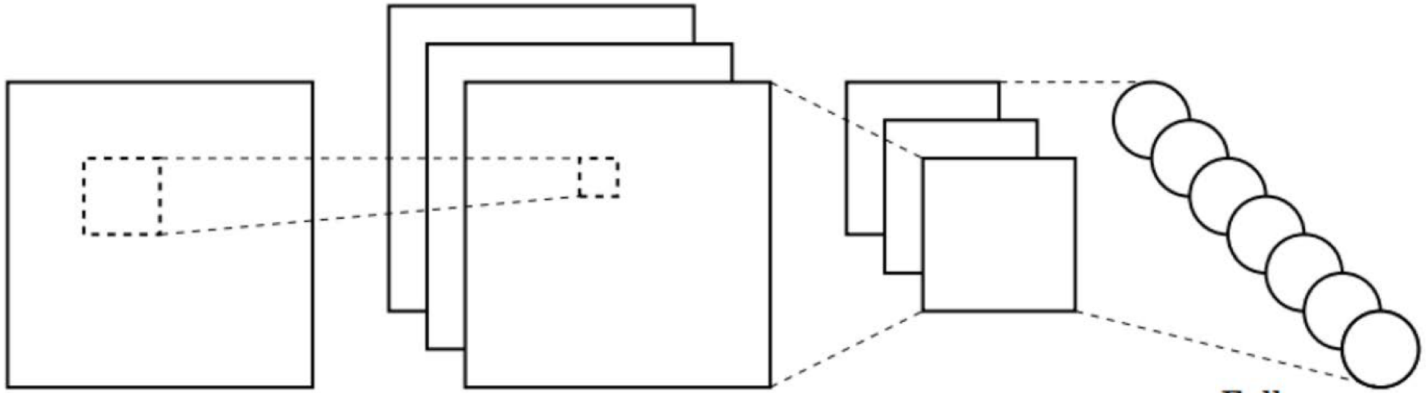
We slide the 3x3 filter over the input image, element-wise multiply, and add the outputs:





Convolutional Neural Networks (CNNs)

CNNs for Classification



1. **Convolution**: Apply filters to generate feature maps.
2. **Non-linearity**: Often ReLU.
3. **Pooling**: Downsampling operation on each feature map.

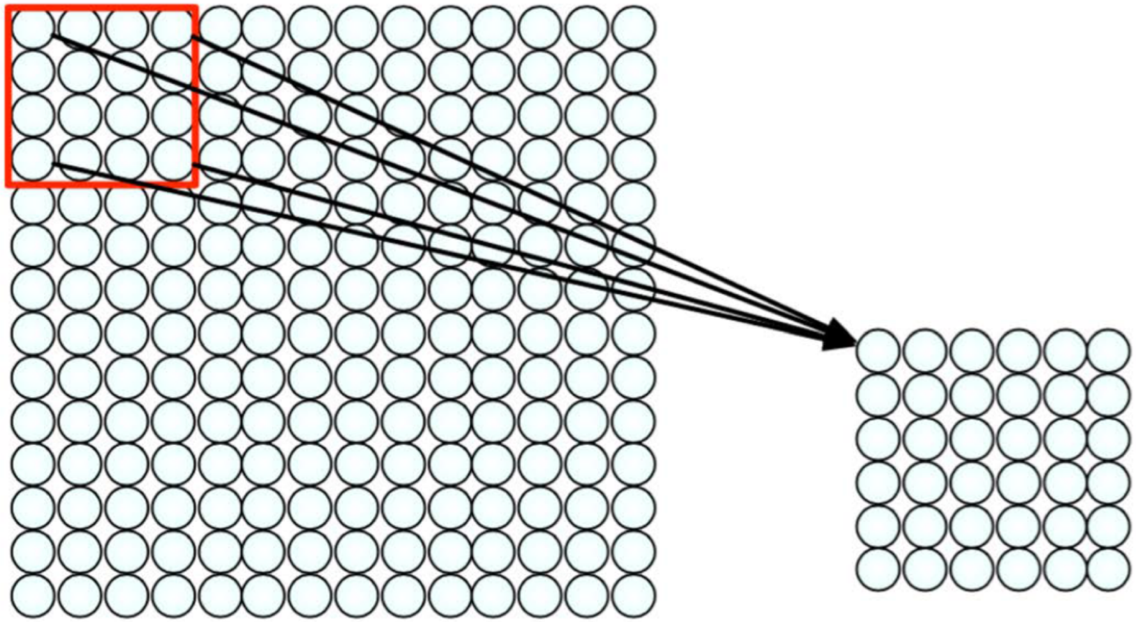
```
tf.keras.layers.Conv2D
```

```
tf.keras.activations.*
```

```
tf.keras.layers.MaxPool2D
```

Train model with image data.
Learn weights of filters in convolutional layers.

Convolutional Layers: Local Connectivity



```
tf.keras.layers.Conv2D
```

For a neuron in hidden layer:

- Take inputs from patch
- Compute weighted sum
- Apply bias

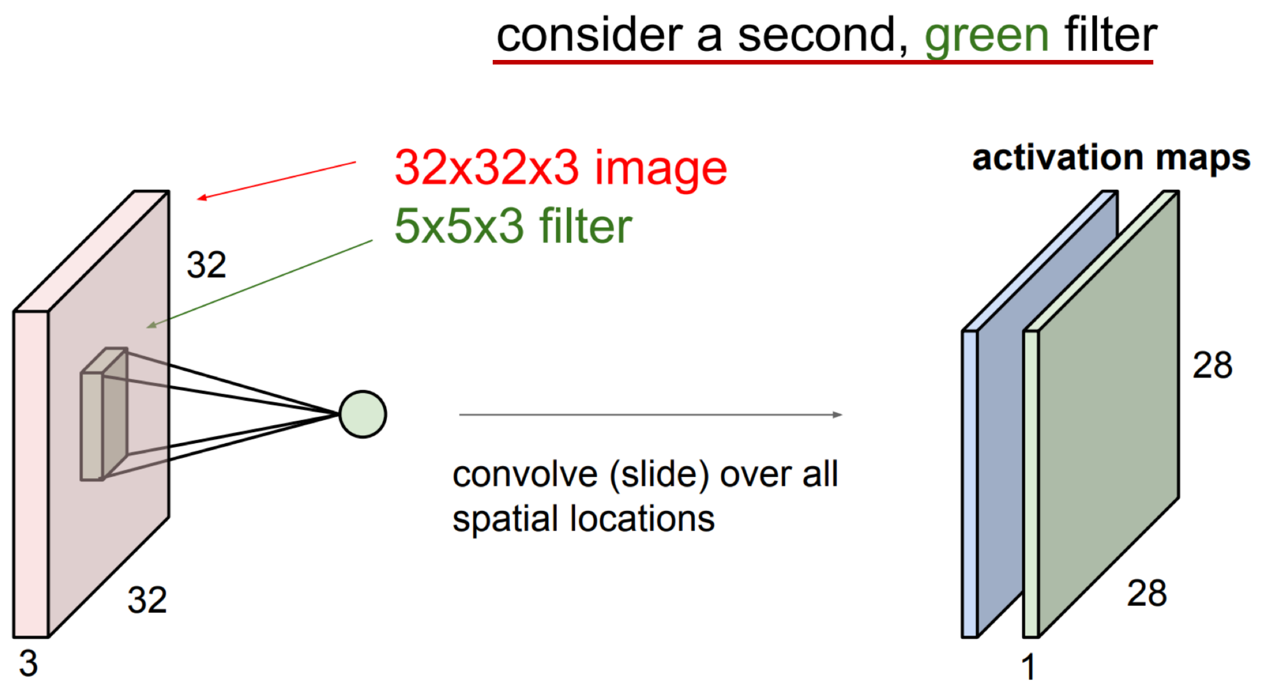
4x4 filter: matrix of weights w_{ij}

$$\sum_{i=1}^{4}\sum_{j=1}^{4} w_{ij}\, x_{i+p,j+q} + b$$

for neuron (p,q) in hidden layer

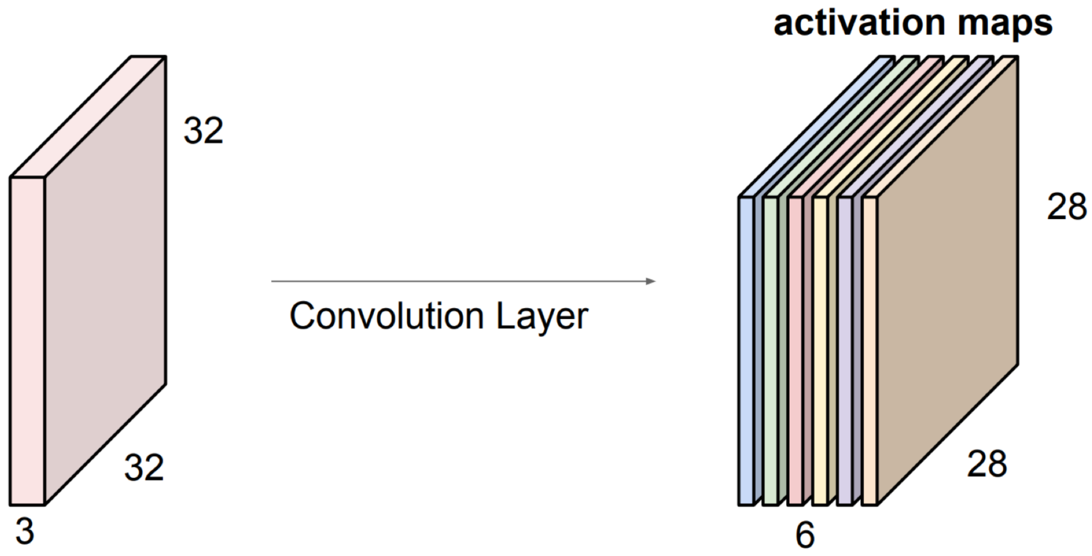
1) applying a window of weights
2) computing linear combinations
3) activating with non-linear function

Convolution layer: Activation Map



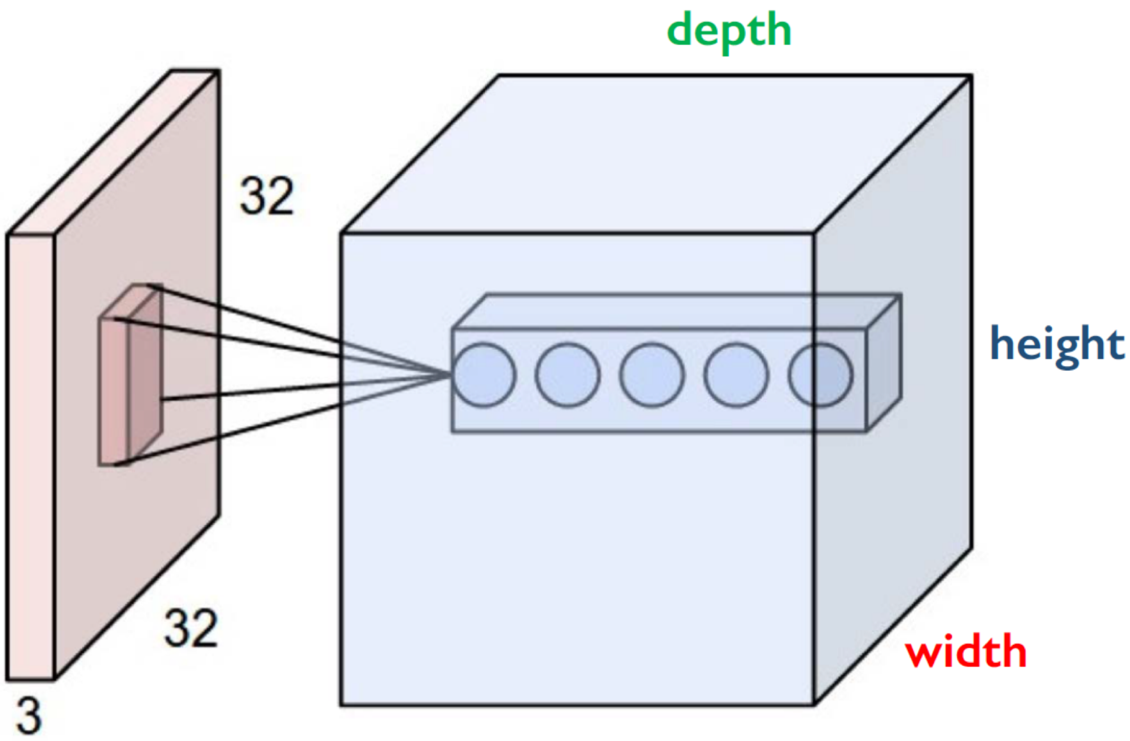
Convolution layer: Activation Map

For example, if we had 6 5x5 filters, we'll get 6 separate activation maps:



We stack these up to get a "new image" of size 28x28x6!

CNNs: Spatial Arrangement of Output Volume



Layer Dimensions:

$h \; x \; w \; x \; d$

where h and w are spatial dimensions
d (depth) = number of filters

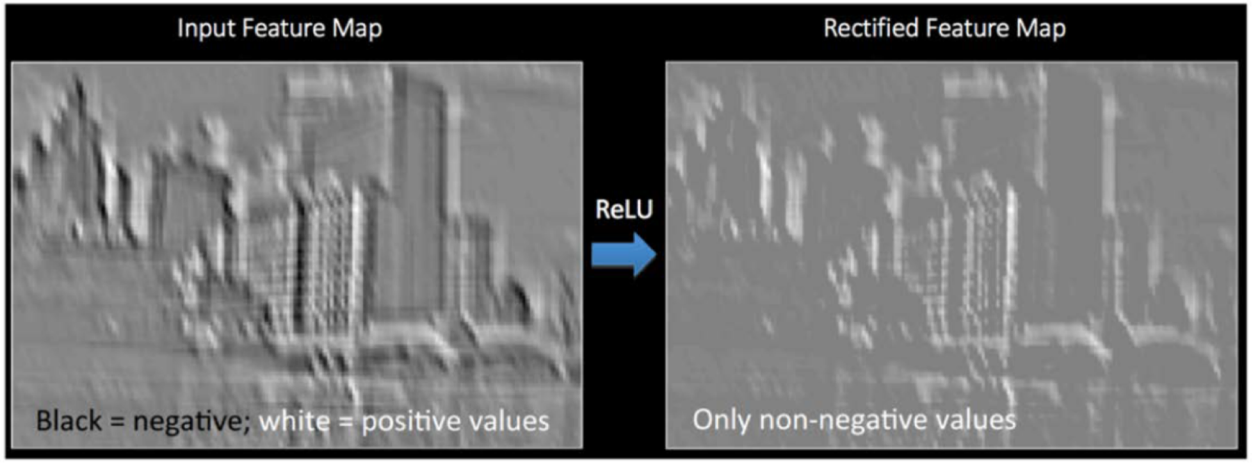
Stride:
Filter step size

Receptive Field:
Locations in input image that
a node is path connected to

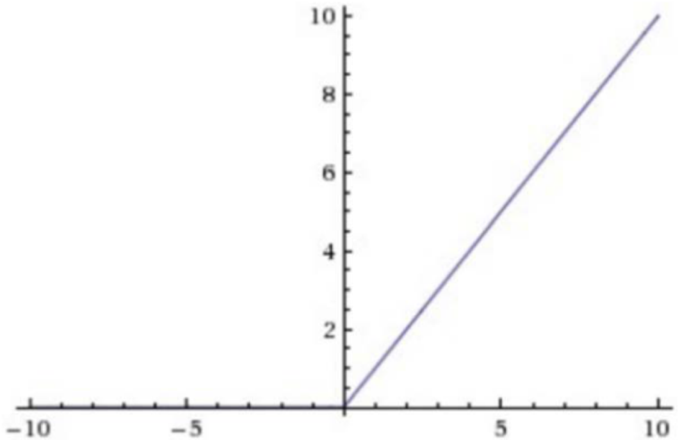
```
tf.keras.layers.Conv2D( filters=d, kernel_size=(h,w), strides=s )
```

Introducing Non-Linearity

- Apply after every convolution operation (i.e., after convolutional layers)
- ReLU: pixel-by-pixel operation that replaces all negative values by zero. **Non-linear operation!**



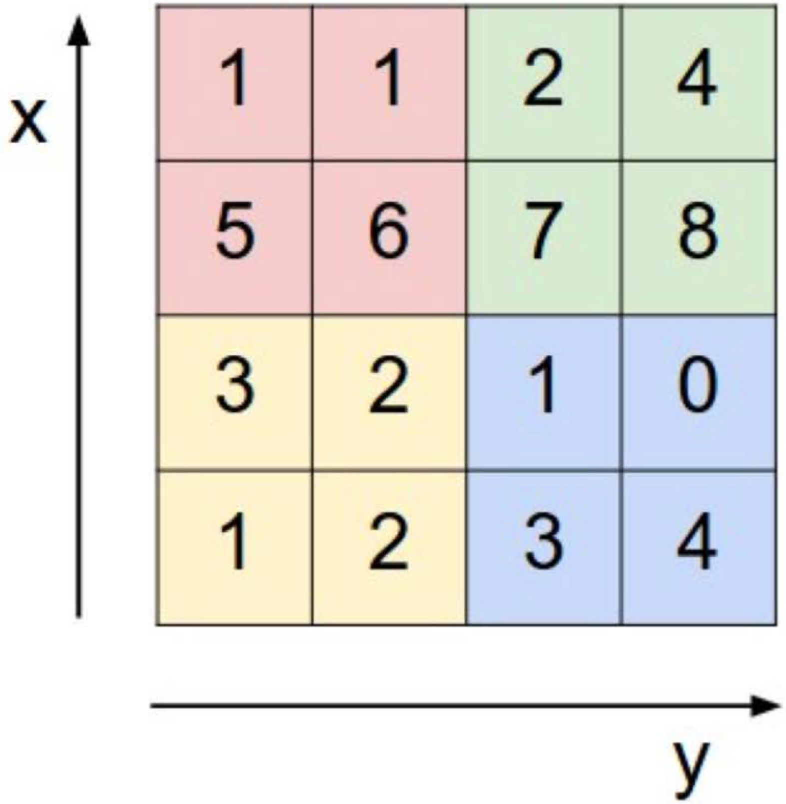
Rectified Linear Unit (ReLU)



$$g(z) = \max(0, z)$$

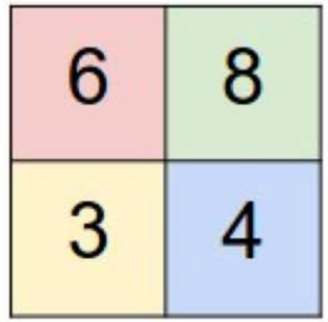
`tf.keras.layers.ReLU`

Pooling



max pool with 2x2 filters
and stride 2

```
tf.keras.layers.MaxPool2D(
    pool_size=(2,2),
    strides=2
)
```

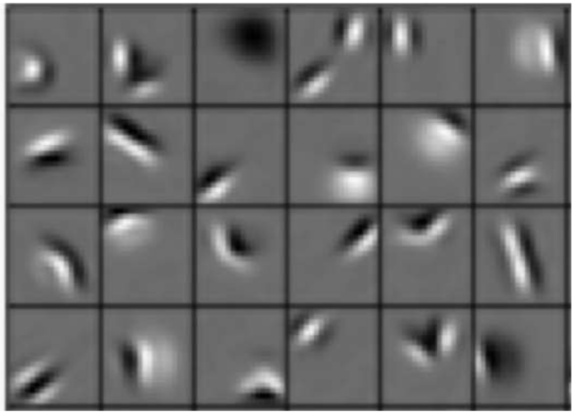


1) Reduced dimensionality
2) Spatial invariance

How else can we downsample and preserve spatial invariance?

Representation Learning in Deep CNNs

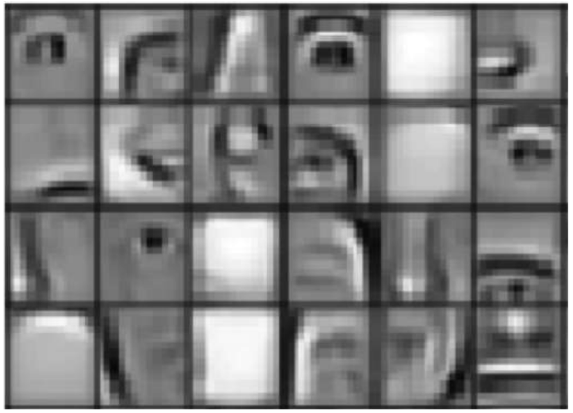
Low level features



Edges, dark spots

Conv Layer 1

Mid level features



Eyes, ears, nose

Conv Layer 2

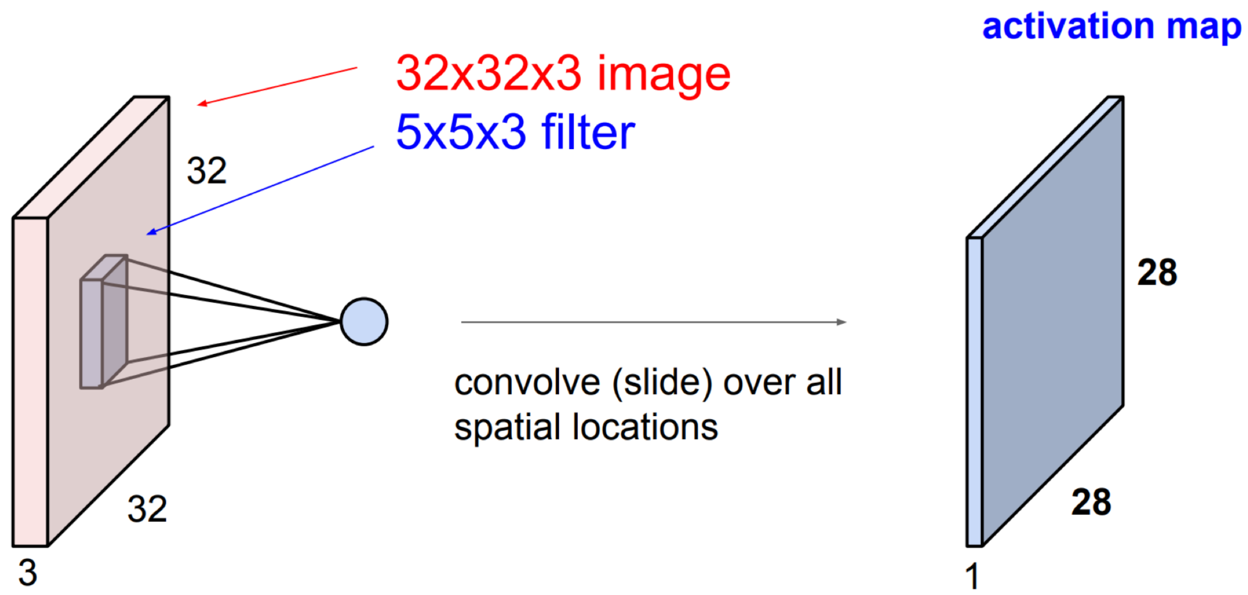
High level features

Facial structure

Conv Layer 3

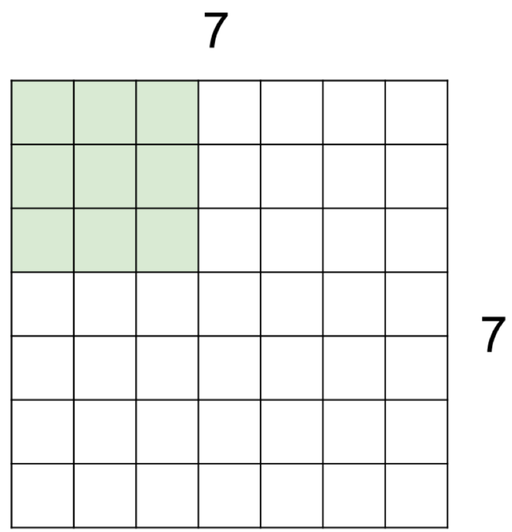
Spatial parameters

A closer look at spatial dimensions:



Some parameters: Stride

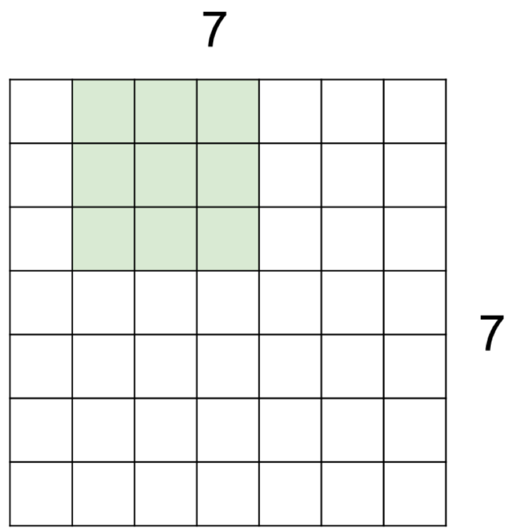
A closer look at spatial dimensions:



7x7 input (spatially)
assume 3x3 filter

Some parameters: Stride

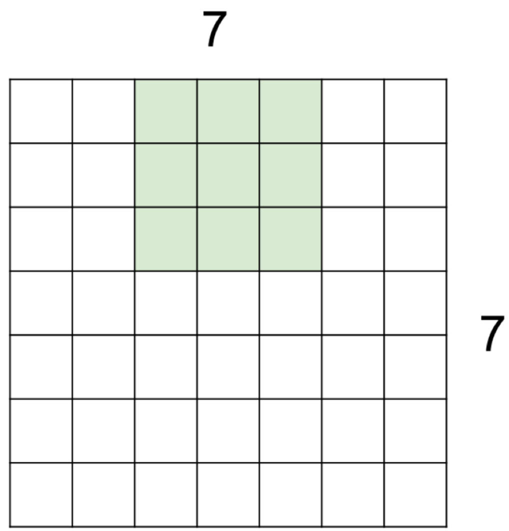
A closer look at spatial dimensions:



7x7 input (spatially)
assume 3x3 filter

Some parameters: Stride

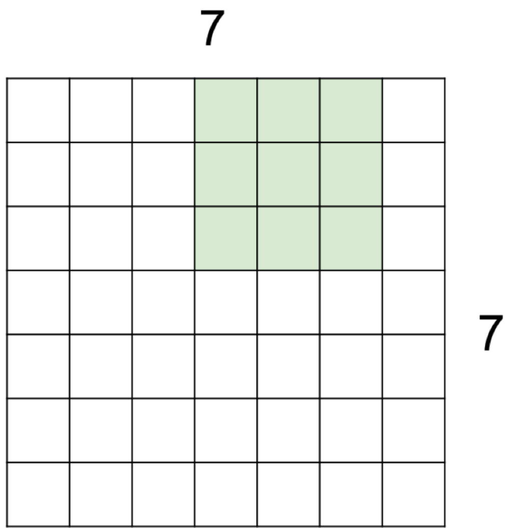
A closer look at spatial dimensions:



7x7 input (spatially)
assume 3x3 filter

Some parameters: Stride

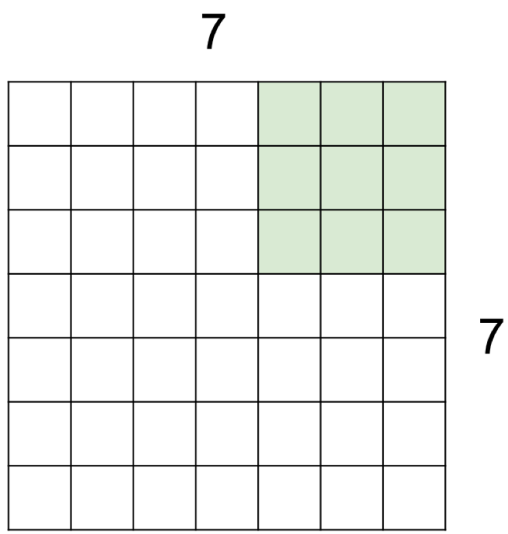
A closer look at spatial dimensions:



7x7 input (spatially)
assume 3x3 filter

Some parameters: Stride

A closer look at spatial dimensions:

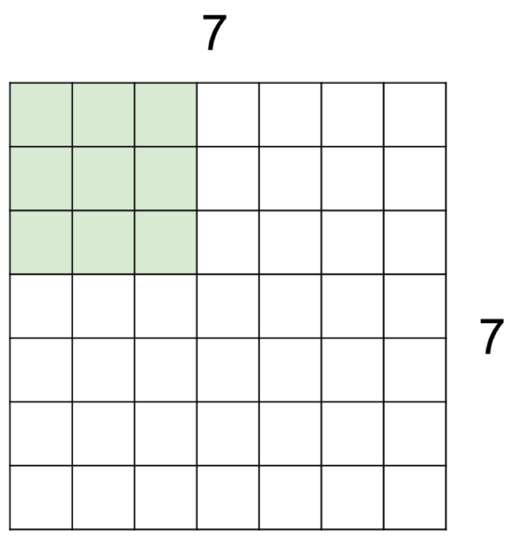


7x7 input (spatially)
assume 3x3 filter

=> 5x5 output

Some parameters: Stride

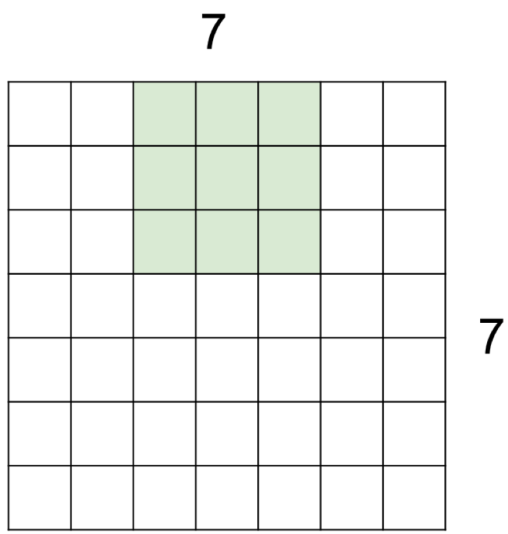
A closer look at spatial dimensions:



7x7 input (spatially)
assume 3x3 filter
applied **with stride 2**

Some parameters: Stride

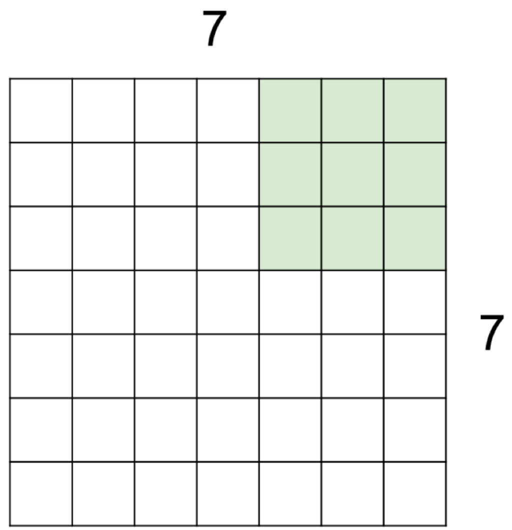
A closer look at spatial dimensions:



7x7 input (spatially)
assume 3x3 filter
applied **with stride 2**

Some parameters: Stride

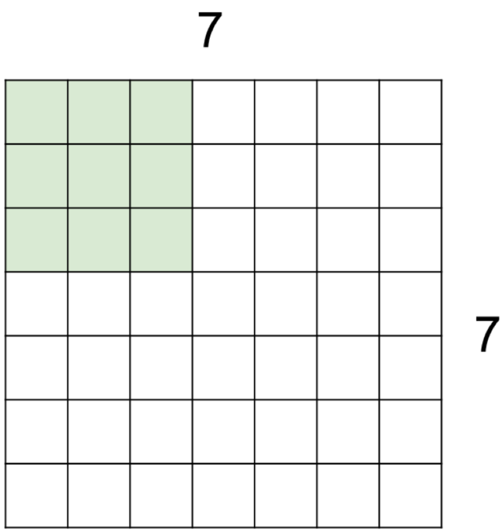
A closer look at spatial dimensions:



7x7 input (spatially)
assume 3x3 filter
applied **with stride 2**
=> 3x3 output!

Some parameters: Stride

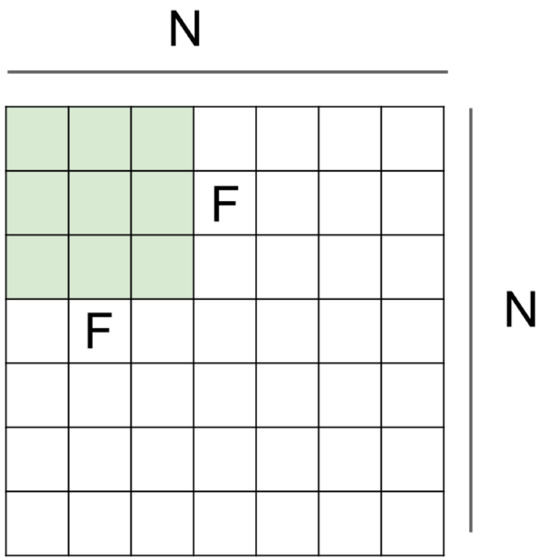
A closer look at spatial dimensions:



7x7 input (spatially)
assume 3x3 filter
applied **with stride 3?**

doesn't fit!
cannot apply 3x3 filter on
7x7 input with stride 3.

Some parameters: Stride

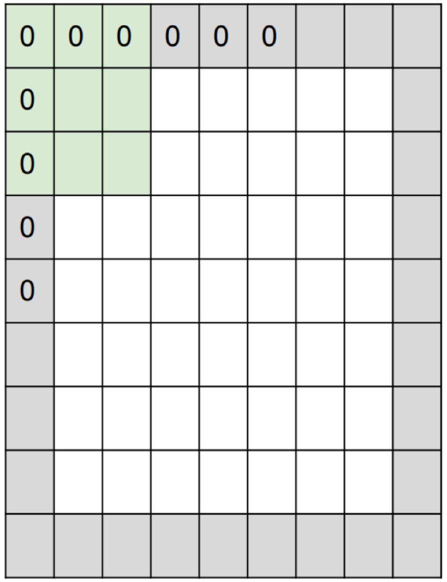


Output size:
(N - F) / stride + 1

e.g. N = 7, F = 3:
stride 1 => (7 - 3)/1 + 1 = 5
stride 2 => (7 - 3)/2 + 1 = 3
stride 3 => (7 - 3)/3 + 1 = 2.33 :\

Some parameters: Padding

In practice: Common to zero pad the border

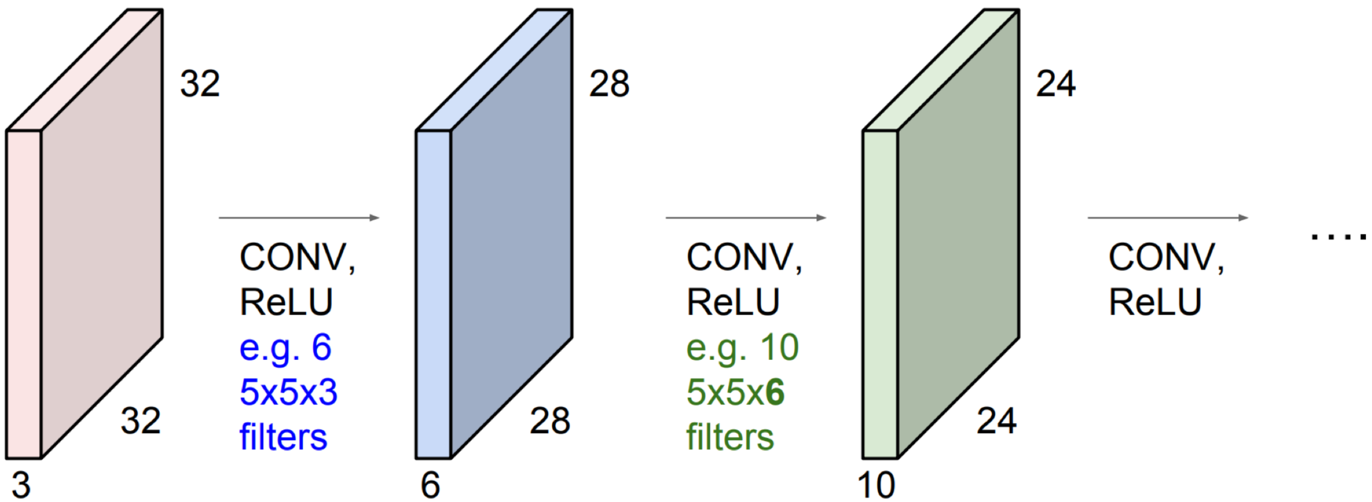


e.g. input 7x7
3x3 filter, applied with **stride 1**
pad with 1 pixel border => what is the output?

7x7 output!
in general, common to see CONV layers with stride 1, filters of size FxF, and zero-padding with (F-1)/2. (will preserve size spatially)
e.g. F = 3 => zero pad with 1
F = 5 => zero pad with 2
F = 7 => zero pad with 3

Some parameters: Padding

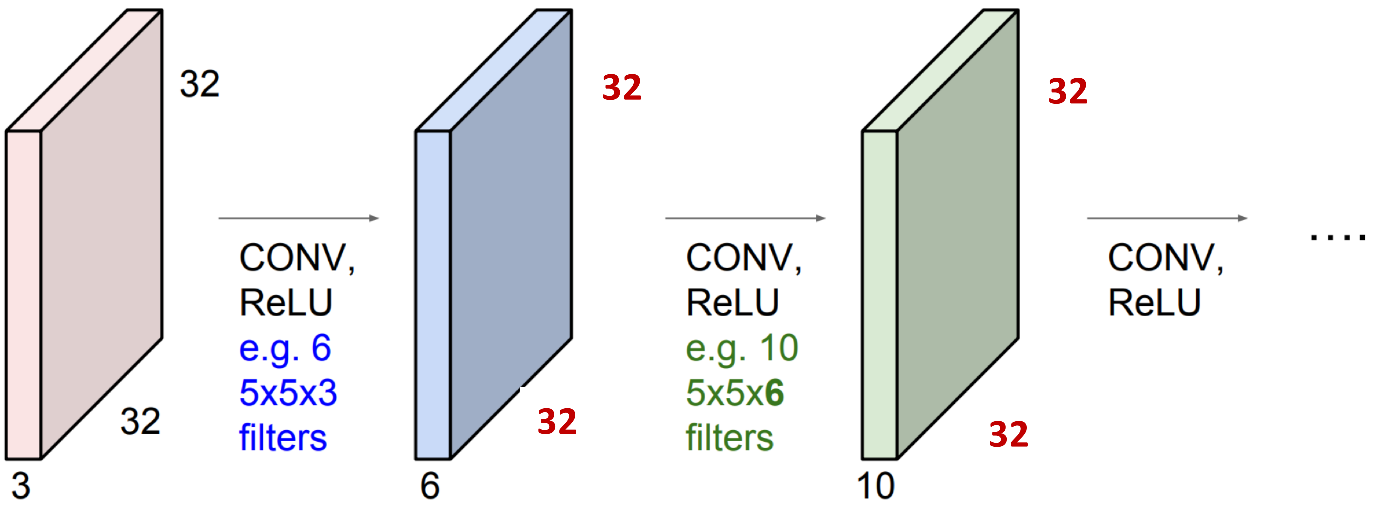
Remember back to…
E.g. 32x32 input convolved repeatedly with 5x5 filters shrinks volumes spatially!
(32 -> 28 -> 24 ...).



Shrinking too fast doesn't work well! !

Some parameters: Padding

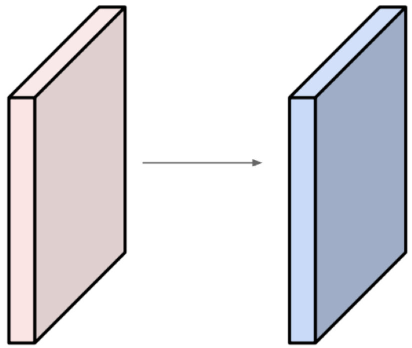
With Padding:



Padding preserve the dimensions !

Some parameters

Examples time:

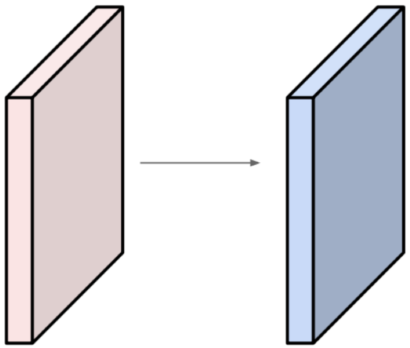


Input volume: **32x32x3**
10 5x5 filters with stride 1, pad 2

Number of parameters in this layer?

Some parameters

Examples time:



Input volume: **32x32x3**

10 5x5 filters with stride 1, pad 2

Number of parameters in this layer?

each filter has 5*5*3 + 1 = 76 params (+1 for bias)

=> 76*10 = **760**

Summary

Summary. To summarize, the Conv Layer:

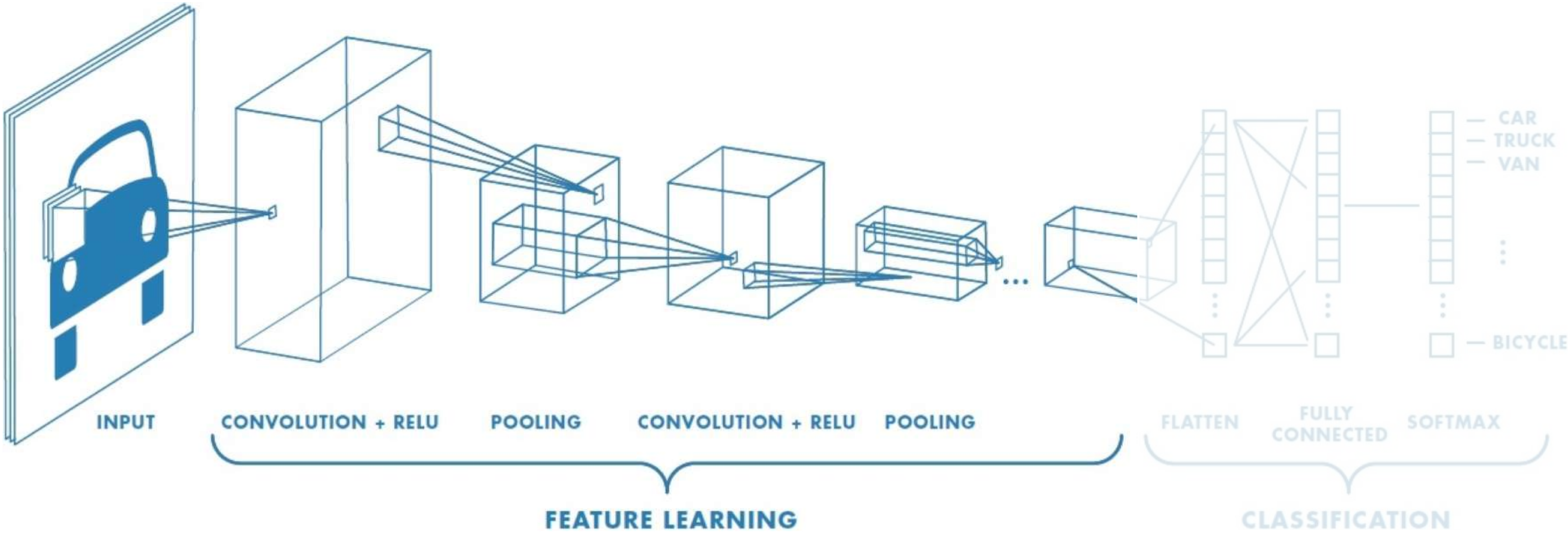
- Accepts a volume of size $W_1 \times H_1 \times D_1$
- Requires four hyperparameters:
  - Number of filters K,
  - their spatial extent F,
  - the stride S,
  - the amount of zero padding P.
- Produces a volume of size $W_2 \times H_2 \times D_2$ where:
  - $W_2 = (W_1 - F + 2P)/S + 1$
  - $H_2 = (H_1 - F + 2P)/S + 1$ (i.e. width and height are computed equally by symmetry)
  - $D_2 = K$
- With parameter sharing, it introduces $F \cdot F \cdot D_1$ weights per filter, for a total of $(F \cdot F \cdot D_1) \cdot K$ weights and K biases.
- In the output volume, the d-th depth slice (of size $W_2 \times H_2$) is the result of performing a valid convolution of the d-th filter over the input volume with a stride of S, and then offset by d-th bias.

Common settings:

K = (powers of 2, e.g. 32, 64, 128, 512)

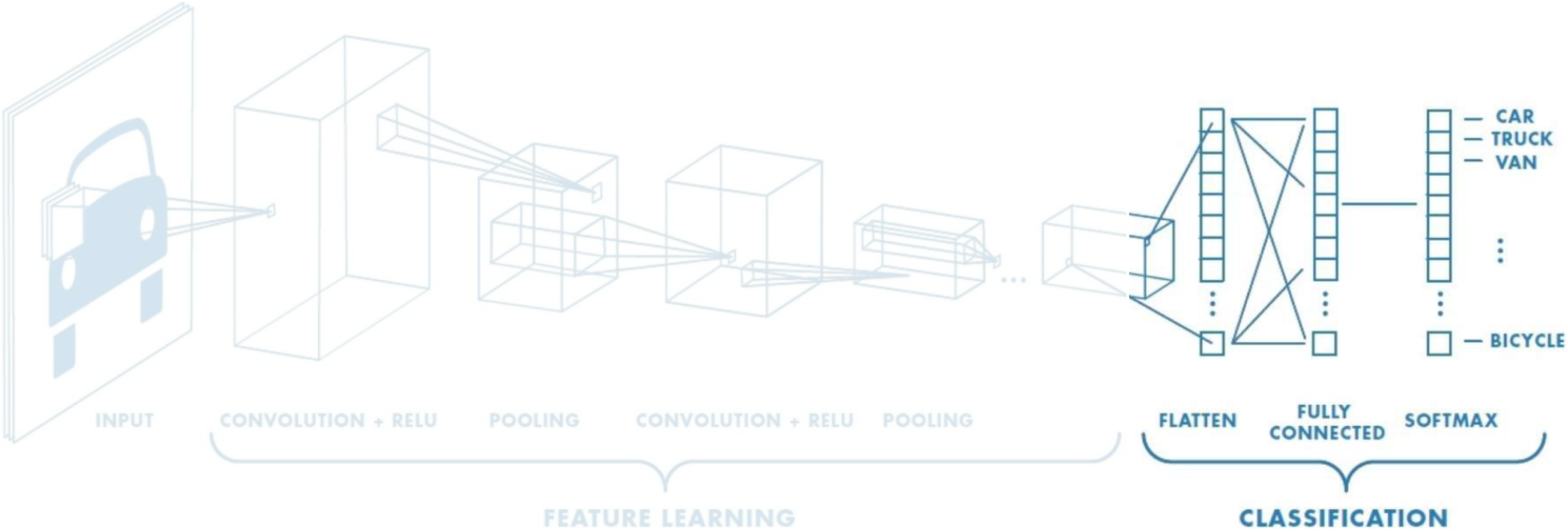
- F = 3, S = 1, P = 1
- F = 5, S = 1, P = 2
- F = 5, S = 2, P = ? (whatever fits)
- F = 1, S = 1, P = 0

CNNs for Classification: Feature Learning



1. Learn features in input image through **convolution**
2. Introduce **non-linearity** through activation function (real-world data is non-linear!)
3. Reduce dimensionality and preserve spatial invariance with **pooling**

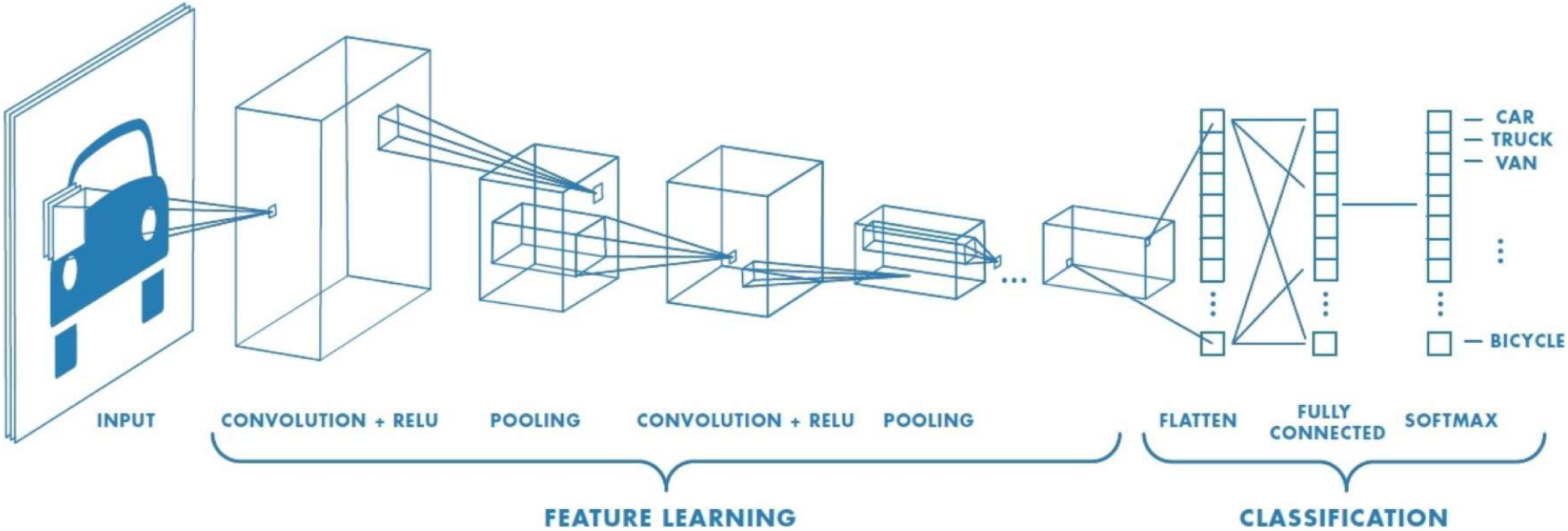
CNNs for Classification: Class Probabilities



- CONV and POOL layers output high-level features of input
- Fully connected layer uses these features for classifying input image
- Express output as **probability** of image belonging to a particular class

$$\mathrm{softmax}(y_i) = \frac{e^{y_i}}{\sum_j e^{y_j}}$$

CNNs: Training with Backpropagation



Learn weights for convolutional filters and fully connected layers

Backpropagation: cross-entropy loss

$$\boldsymbol{J}(\boldsymbol{\theta}) = \sum_i y^{(i)} \log\left(\widehat{\boldsymbol{y}}^{(i)}\right)$$

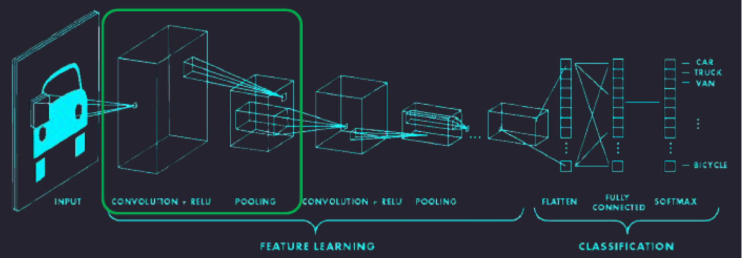
Putting it all together

```
import tensorflow as tf

def generate_model():
    model = tf.keras.Sequential([
        # first convolutional layer
        tf.keras.layers.Conv2D(32, filter_size=3, activation='relu’),
        tf.keras.layers.MaxPool2D(pool_size=2, strides=2),

        # second convolutional layer
        tf.keras.layers.Conv2D(64, filter_size=3, activation='relu’),
        tf.keras.layers.MaxPool2D(pool_size=2, strides=2),

        # fully connected classifier
        tf.keras.layers.Flatten(),
        tf.keras.layers.Dense(1024, activation='relu’),
        tf.keras.layers.Dense(10, activation=‘softmax’)  # 10 outputs
    ])
    return model
```



CNNs for Classification: ImageNet

ImageNet Dataset

Dataset of over 14 million images across 21,841 categories

"Elongated crescent-shaped yellow fruit with soft sweet flesh"

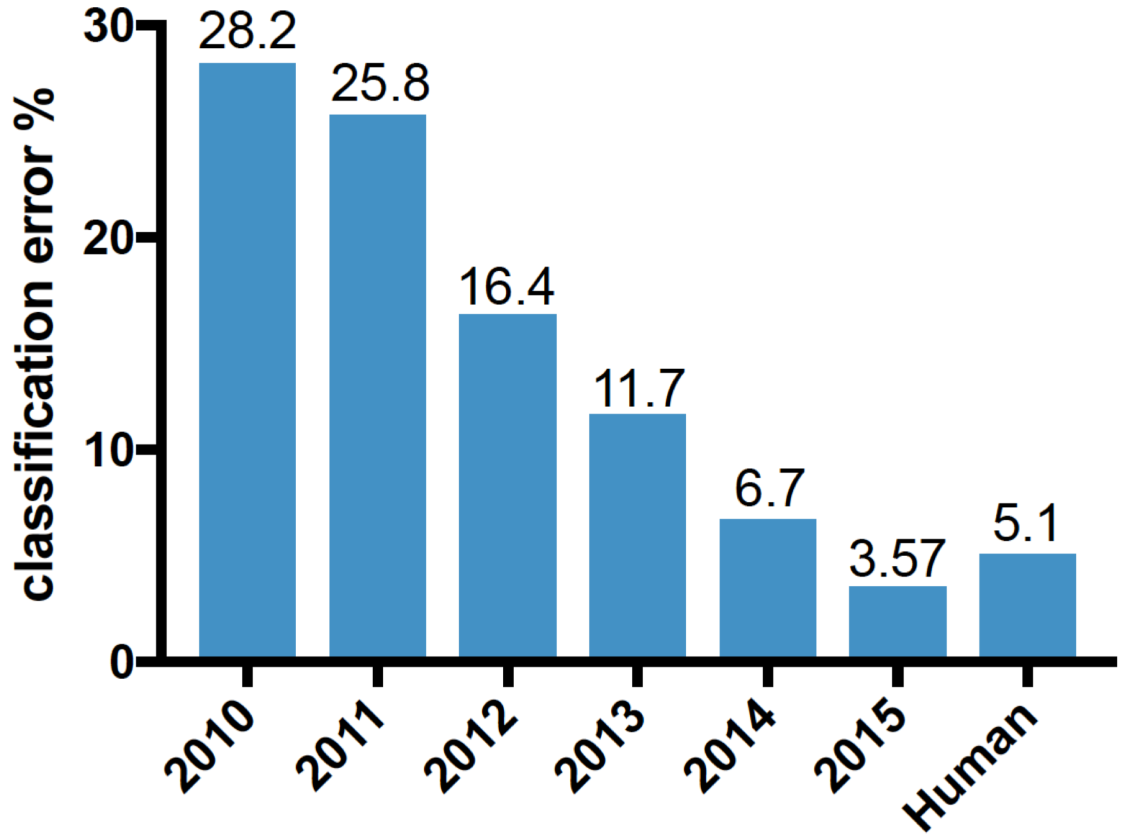
1409 pictures of bananas.

ImageNet Challenge

Classification task: produce a list of object categories present in image.
1000 categories.
"Top 5 error": rate at which the model does not output correct label in top 5 predictions

Other tasks include:
single-object localization, object detection from video/image, scene classification, scene parsing

ImageNet Challenge: Classification Task



2012: AlexNet. First CNN to win.
- 8 layers, 61 million parameters

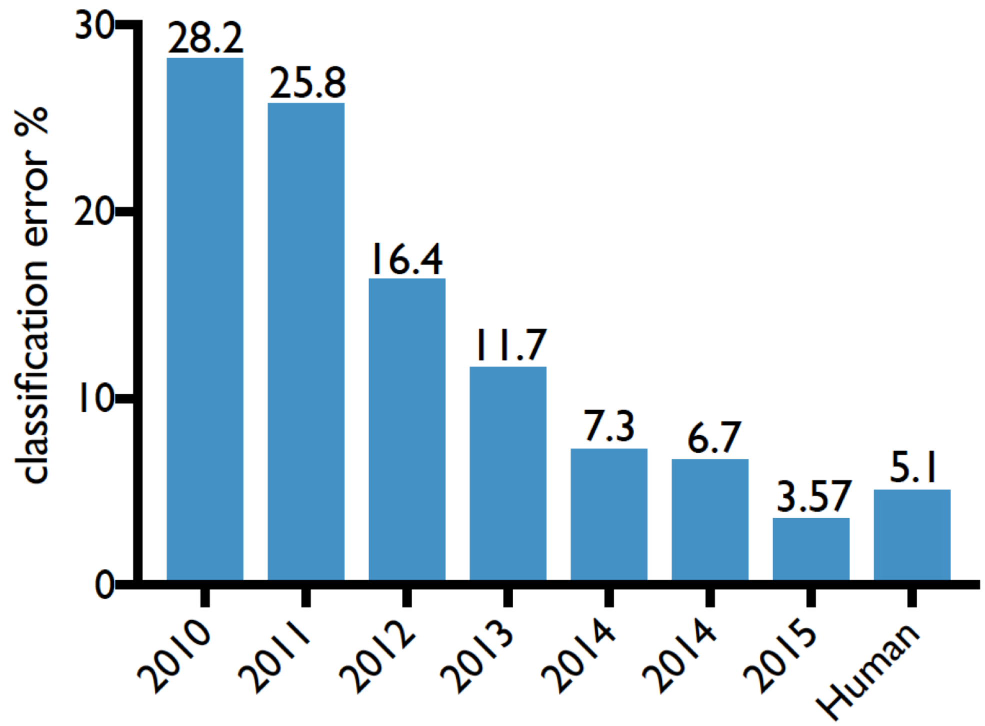
2013: ZFNet
- 8 layers, more filters

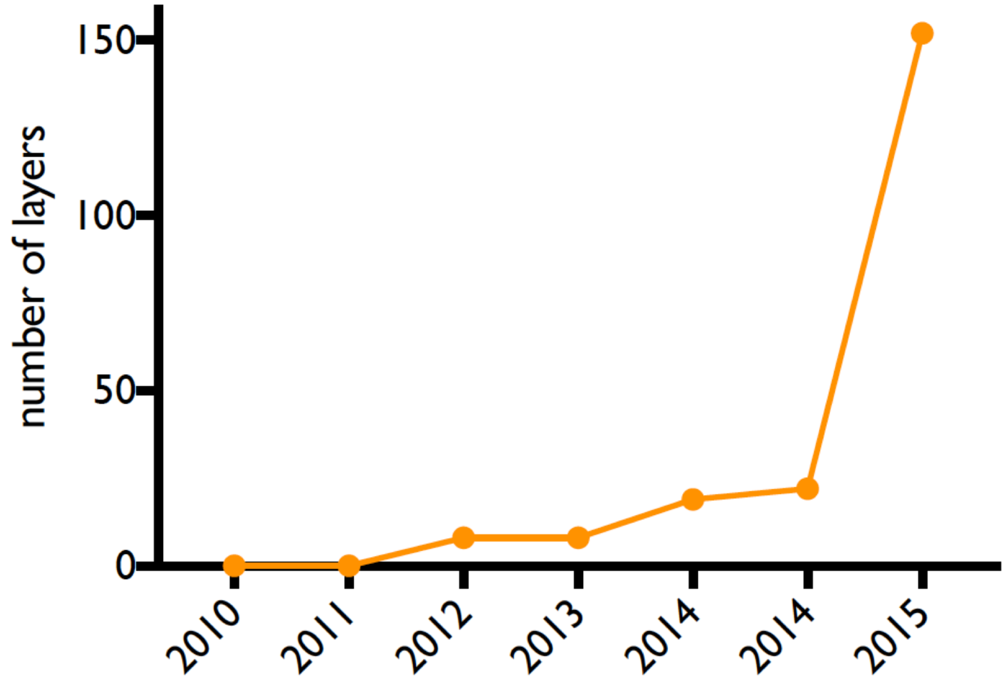
2014: VGG
- 19 layers

2014: GoogLeNet
- "Inception" modules
- 22 layers, 5million parameters

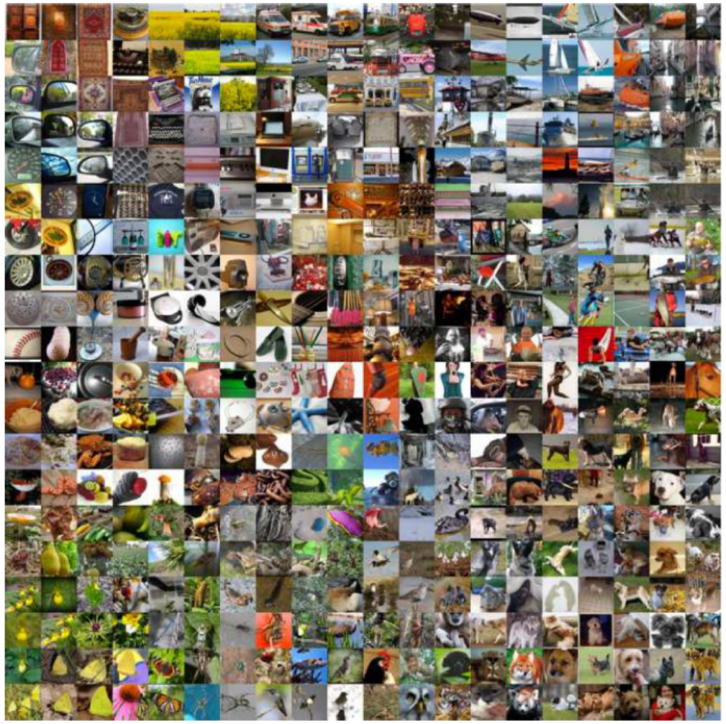
2015: ResNet
- 152 layers

ImageNet Challenge: Classification Task





Data, Data, Data



ImageNet:
22K categories. 14M images.

Airplane

Automobile

Bird

Cat

Deer

Dog

Frog

Horse

Ship

Truck

CIFAR-10

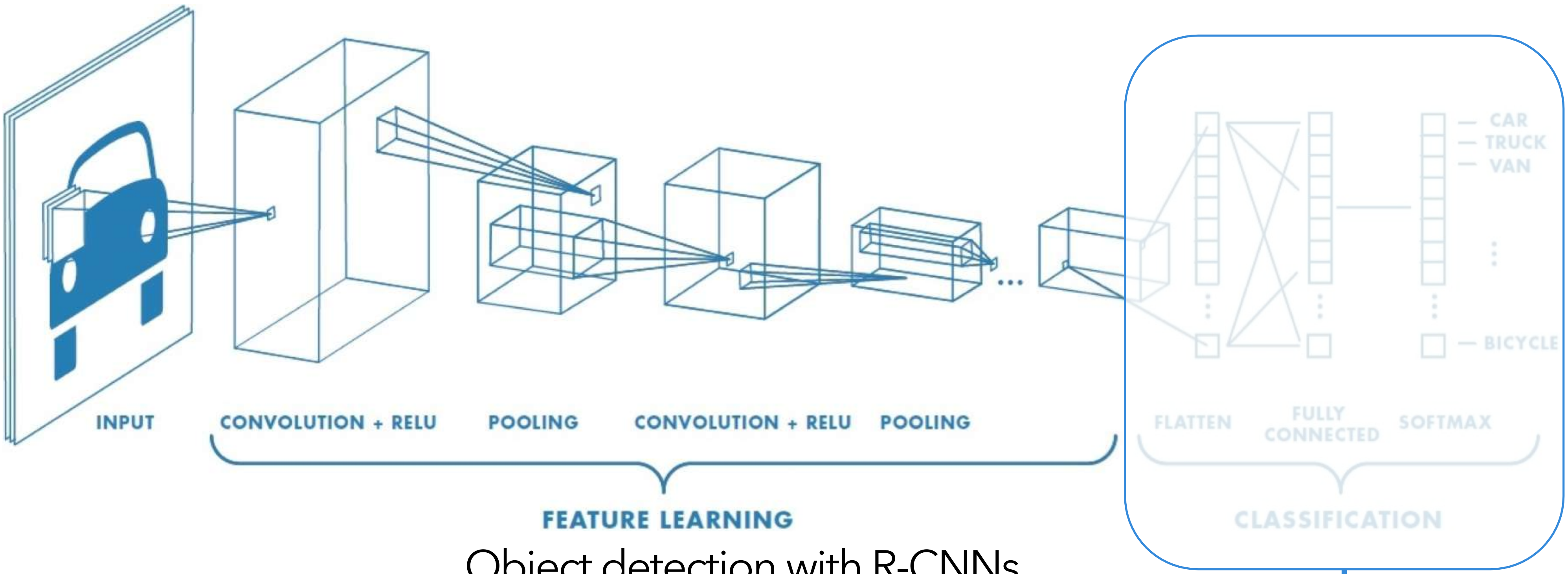
MNIST: handwritten digits

places THE SCENE RECOGNITION DATABASE

places: natural scenes

An Architecture for Many Applications

An Architecture for Many Applications

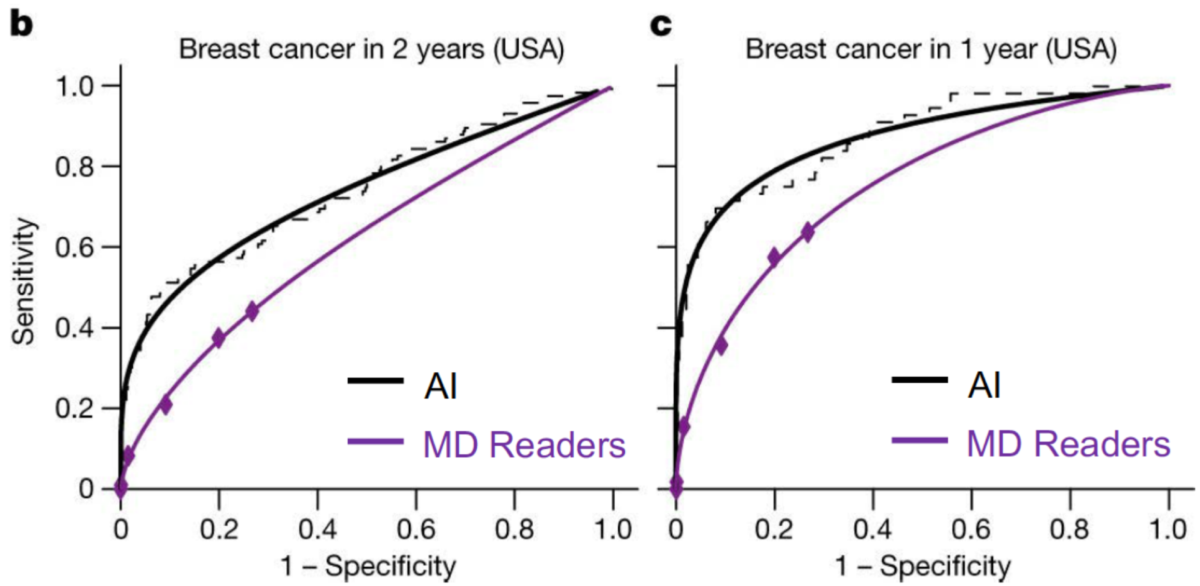


Object detection with R-CNNs
Segmentation with fully CNN
Image captioning with RNNs

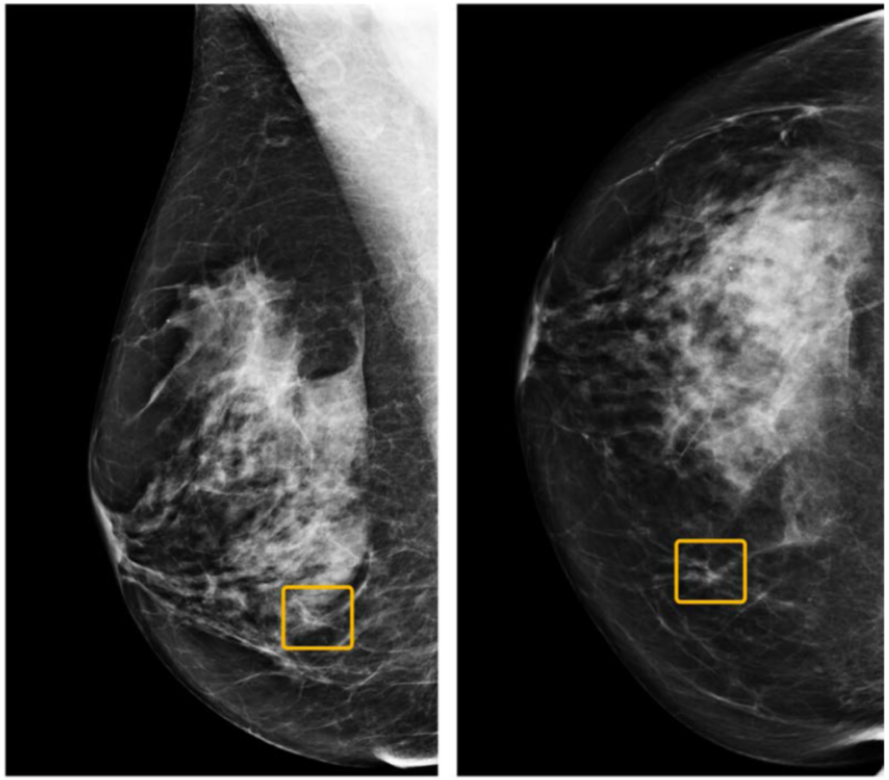
Deep Learning for Computer Vision: Impact

Detection: Breast Cancer Screening

International evaluation of an AI system for breast cancer screening — *nature*



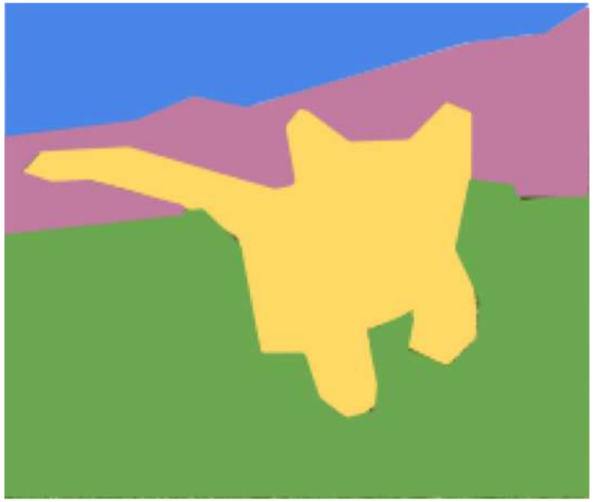
CNN-based system outperformed expert radiologists at detecting breast cancer from mammograms



Breast cancer case missed by radiologist but detected by AI

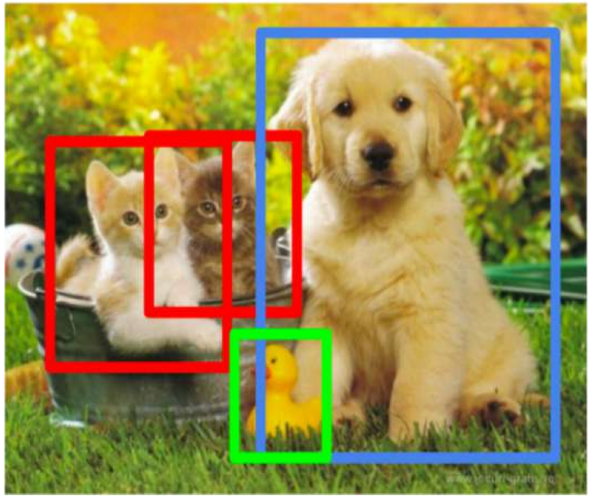
Beyond Classification

Semantic Segmentation



CAT

Object Detection



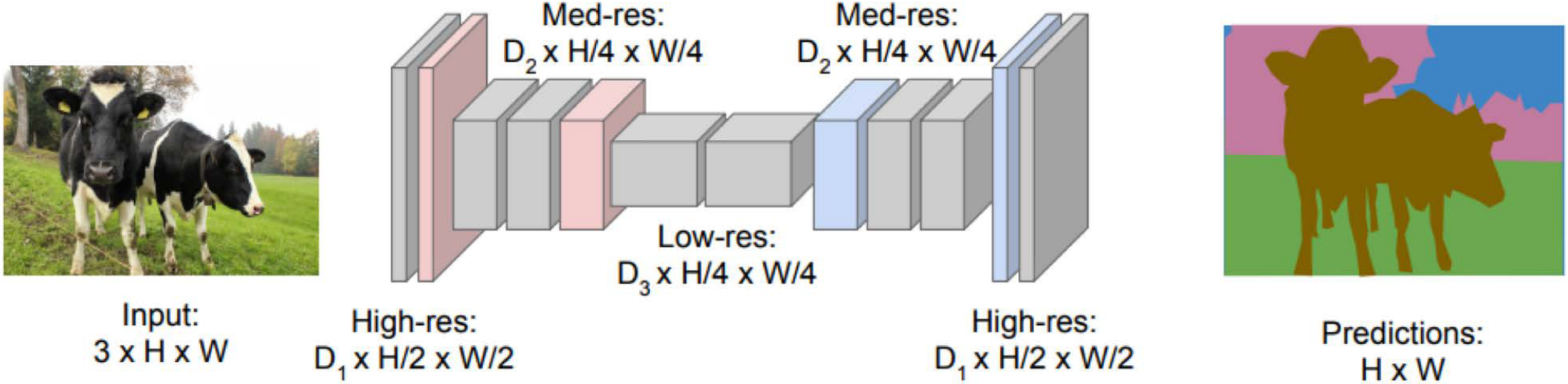
CAT, DOG, DUCK

Image Captioning

The cat is in the grass.

Semantic Segmentation: Fully Convolutional Networks

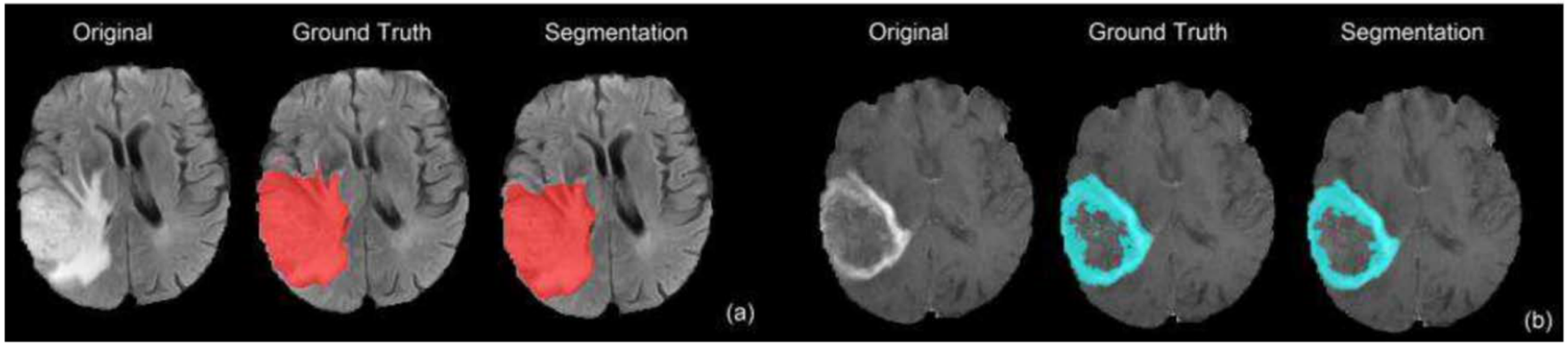
FCN: Fully Convolutional Network.
Network designed with all convolutional layers,
with **downsampling** and **upsampling** operations



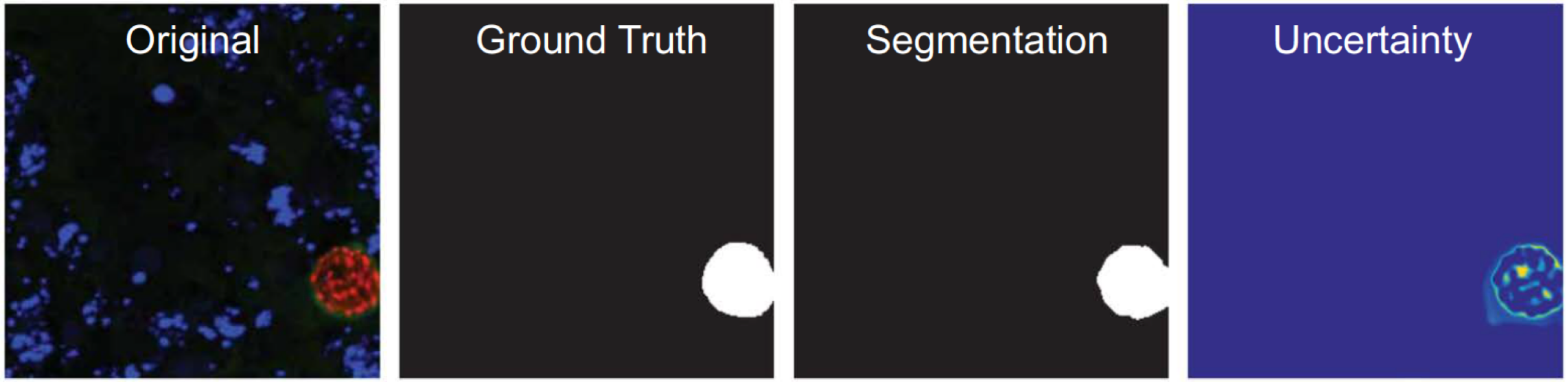
`tf.keras.layers.Conv2DTranspose`

Semantic Segmentation: Biomedical Image Analysis

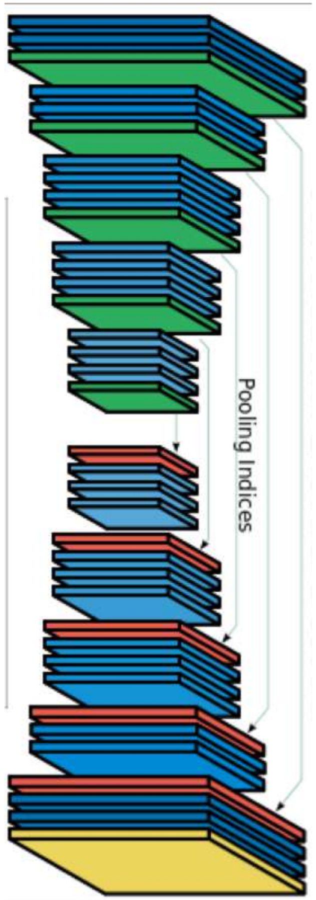
Brain Tumors
Dong+ *MIUA* 2017.



Malaria Infection
Soleimany+ *arXiv* 2019.



Driving Scene Segmentation



Object Detection with R-CNNs

R-CNN: Find regions that we think have objects. Use CNN to classify.

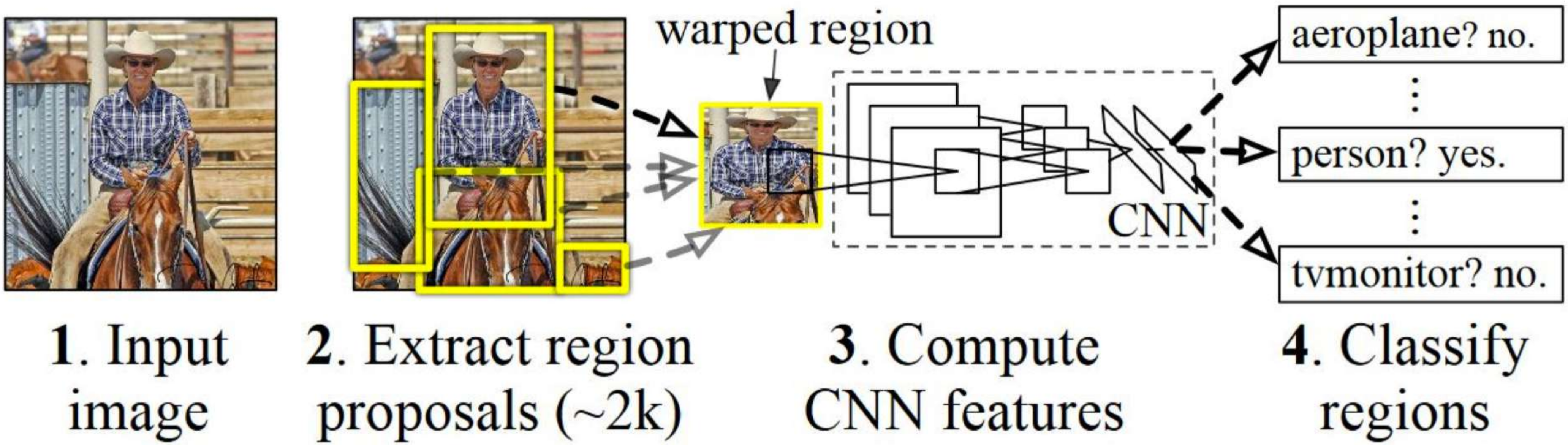


Image Captioning using RNNs

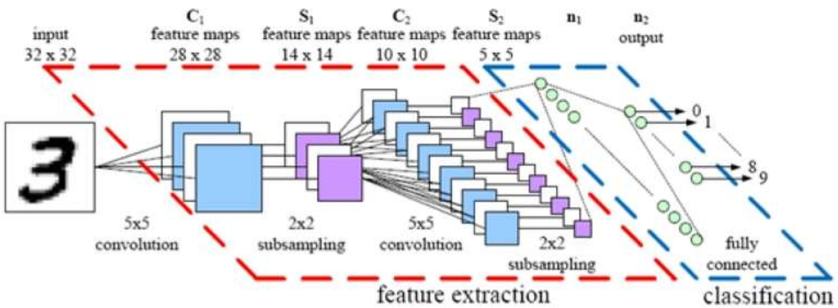
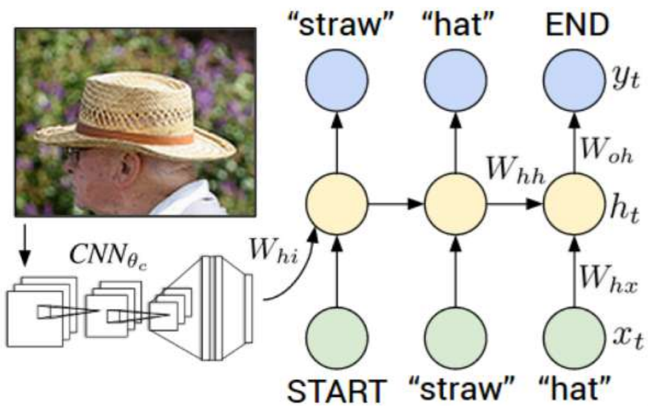
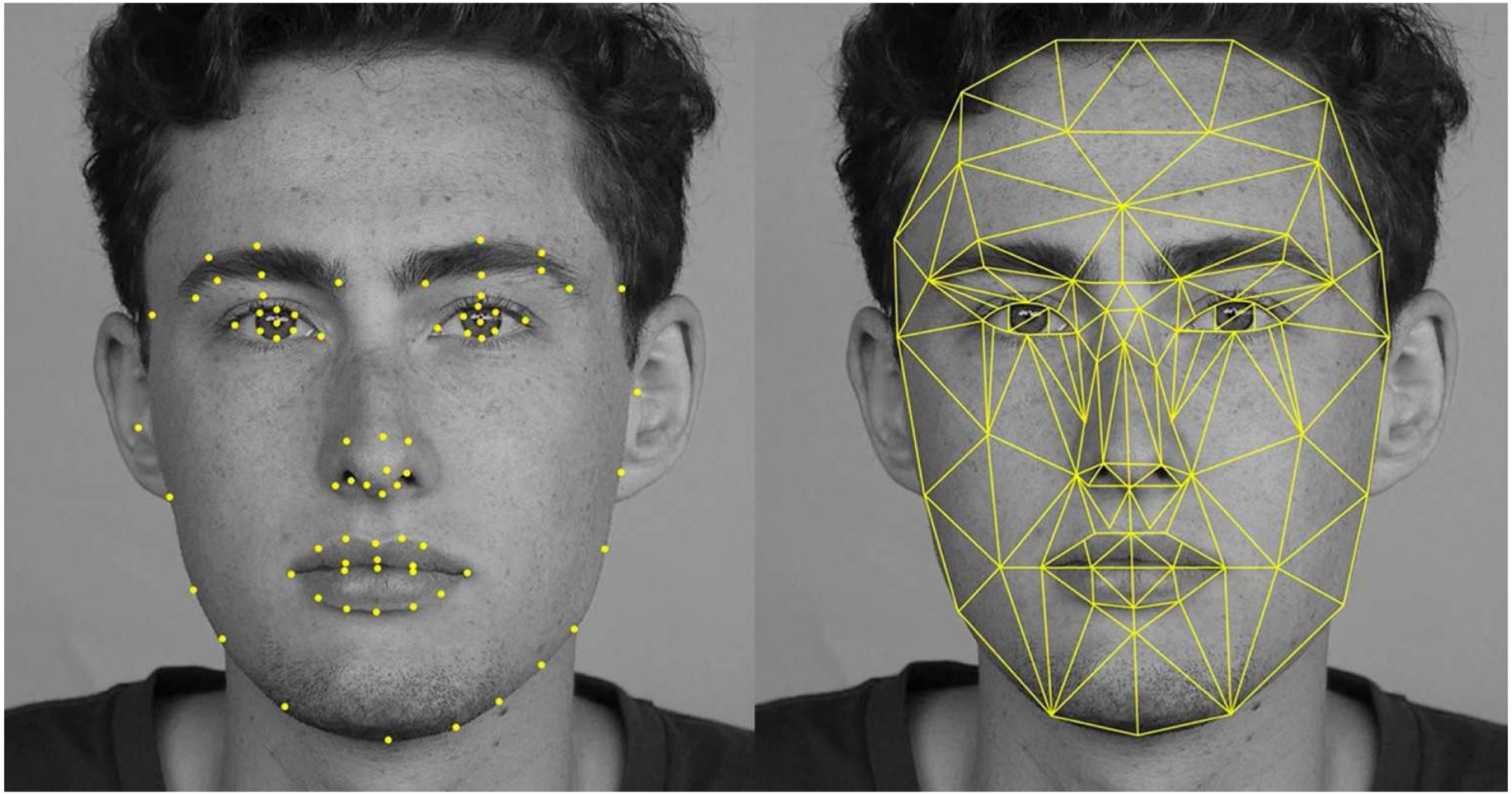


Image Captioning using RNNs



Impact: Face Detection

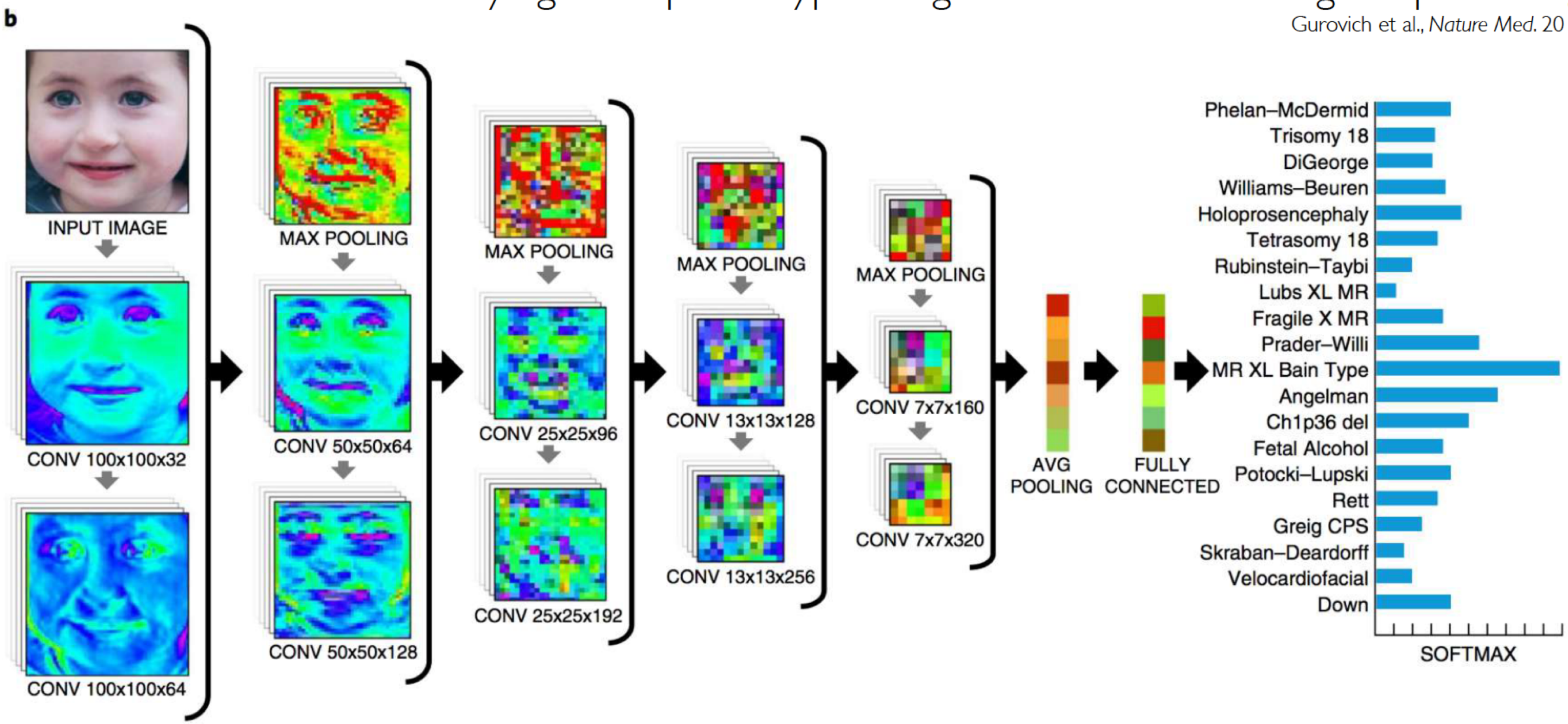


Impact: Self-Driving Cars

Impact: Healthcare

Identifying facial phenotypes of genetic disorders using deep learning

Gurovich et al., *Nature Med.* 2019



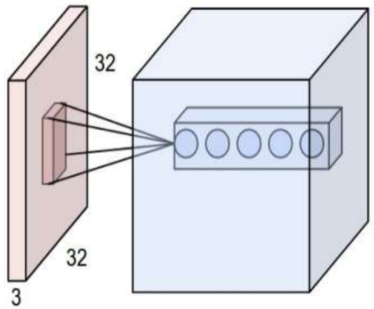
Deep Learning for Computer Vision: Summary

Foundations

- Why computer vision?
- Representing images
- Convolutions for feature extraction

CNNs

- CNN architecture
- Application to classification
- ImageNet



Applications

- Segmentation, object detection, image captioning
- Visualization