

CNN Parameters and Training

Giacomo Boracchi,
DEIB, Politecnico di Milano
October, 11th, 2022

giacomo.boracchi@polimi.it
<https://boracchi.faculty.polimi.it/>

Parameters in a CNN

Convolutions as MLP

Convolution is a linear operation!

Therefore, if you unroll the input image to a vector, **you can consider convolution weights as the weights of a Multilayer Perceptron Network!**

$$a(x, y, k) = \sum_{u,v} w_k(u, v, z) I(x - u, y - v, z)$$

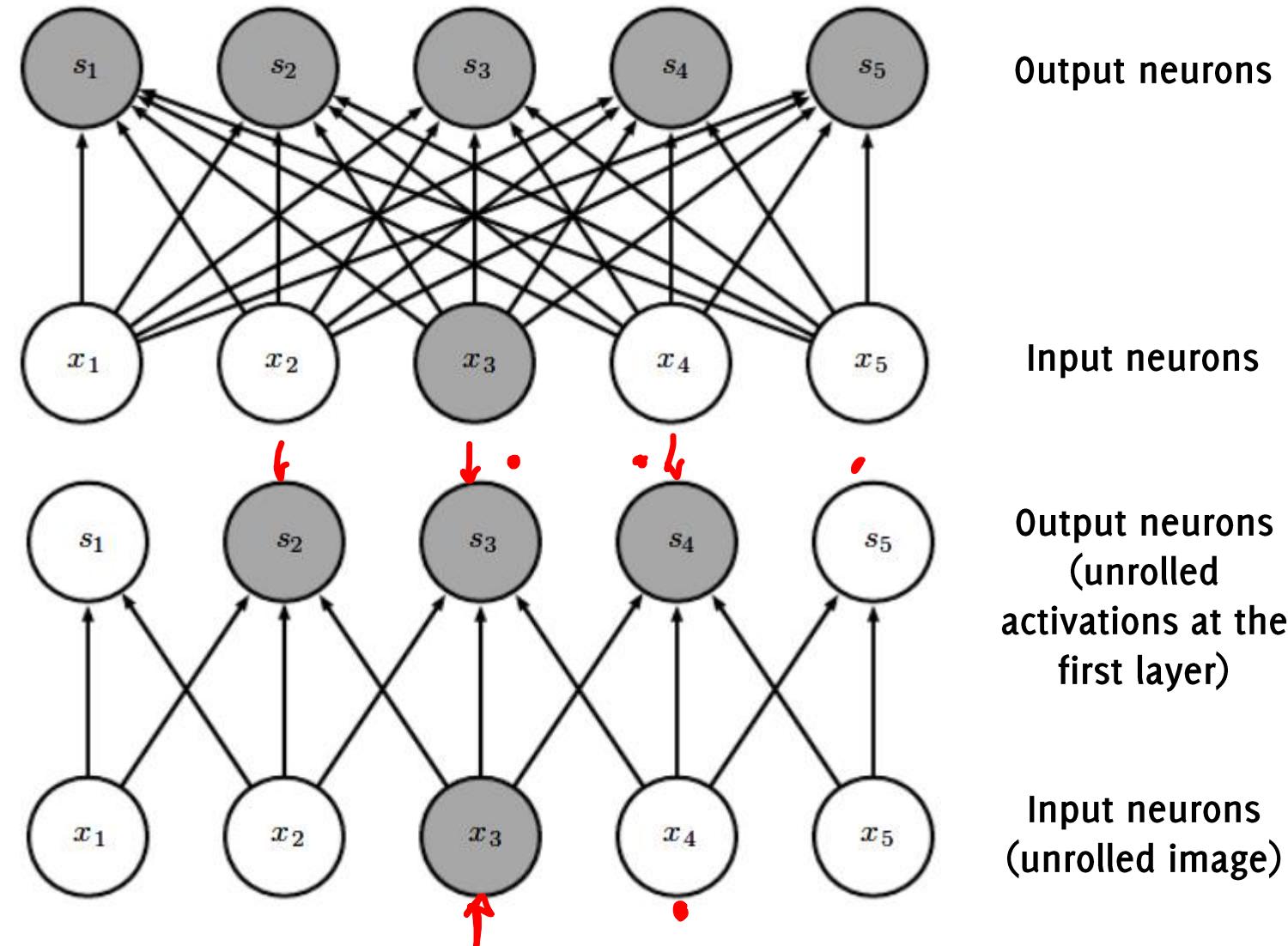
inputs
parameters

What are the differences between MLP and CNNs then?

CNNs has Sparse Connectivity

MLP, Fully connected

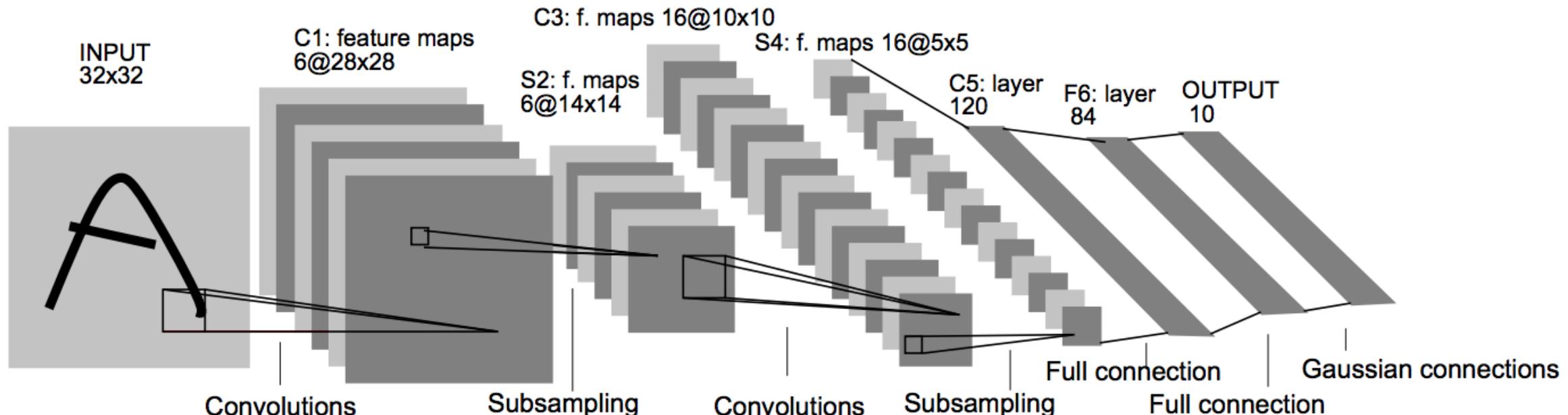
3x1 convolutional



Weight Sharing / Spatial Invariance

In a CNN, all the neurons in the same slice of a feature map use the same weights and bias: this reduces the nr. of parameters in the CNN.

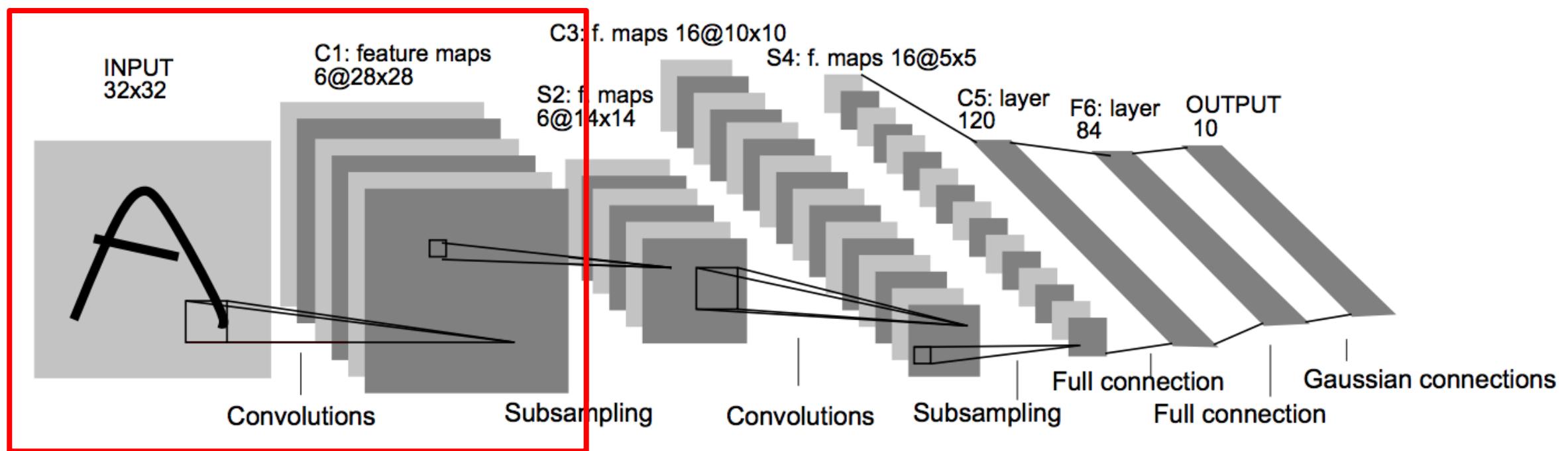
Underlying assumption: if one feature is useful to compute at some spatial position (x, y) , then it should also be useful to compute at a different position (x_2, y_2)



Weight Sharing / Spatial Invariance

If the first layer were a MLP:

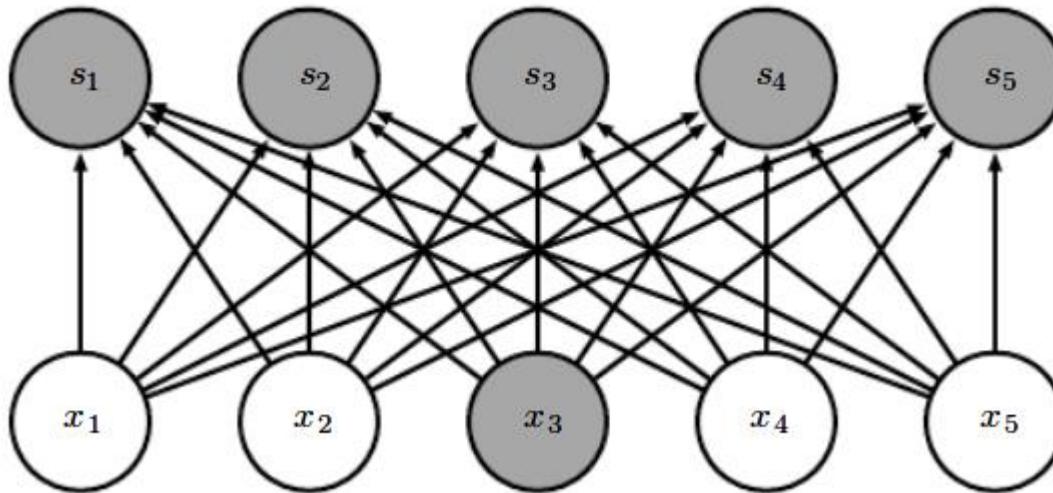
- MLP layer: this should have had $28 \times 28 \times 6$ neurons in the output
- MLP layer with sparse connectivity: only 5×5 nonzero weights each neuron
- MLP layer: $28 \times 28 \times 6 \times 25$ weights + $28 \times 28 \times 6$ biases (122 304)
- **Conv layer: 25 weights + 6 biases shared among neurons of the same layer**



Parameter sharing

Fully connected

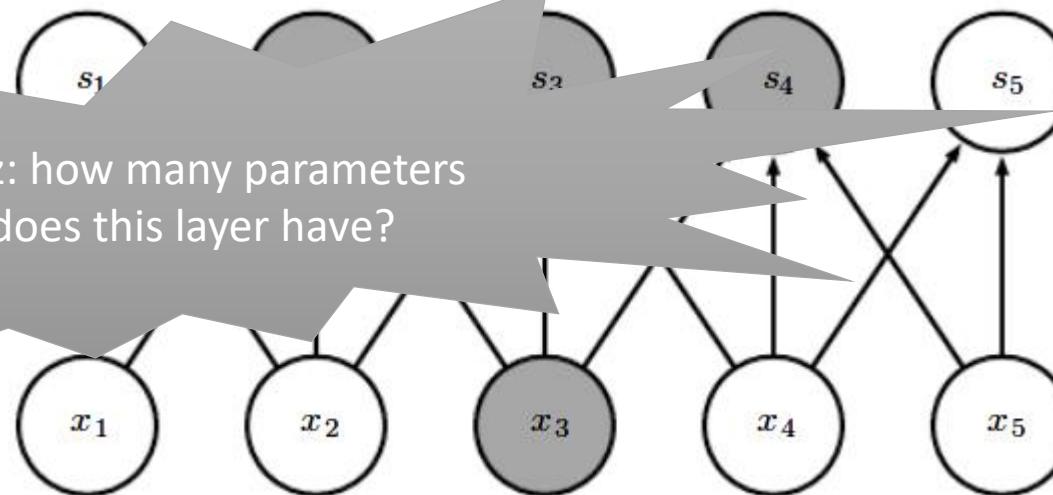
$5 \times 5 = 25$ weights
(+ 5 bias)



3x1 convolve

3 weights!
(+ 1 bias)

Quiz: how many parameters
does this layer have?



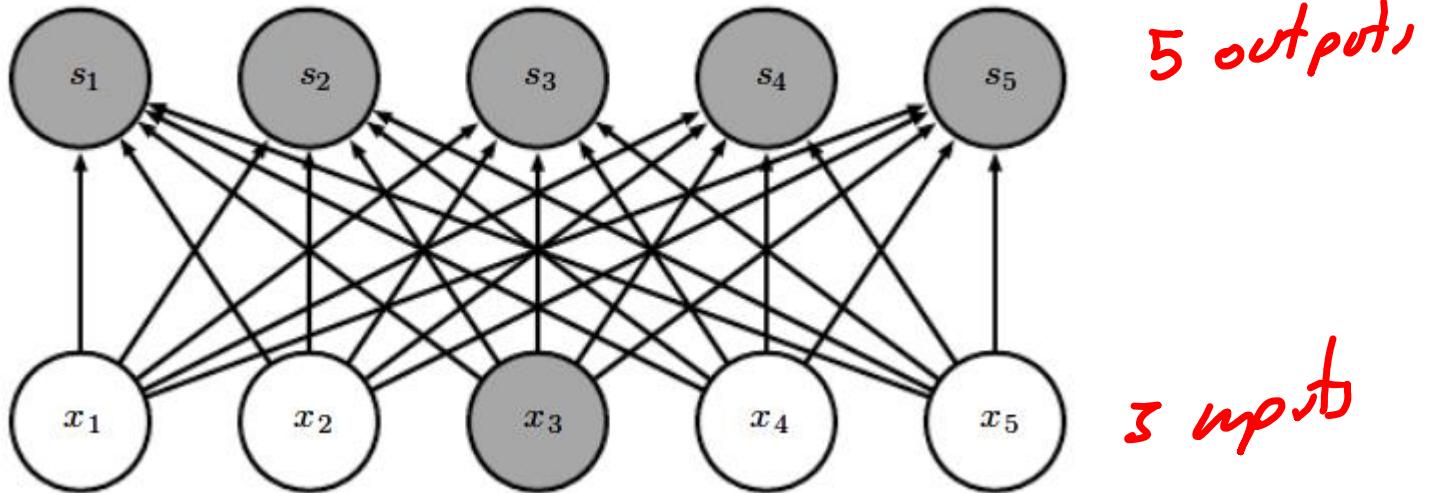
Parameter sharing

Fully connected

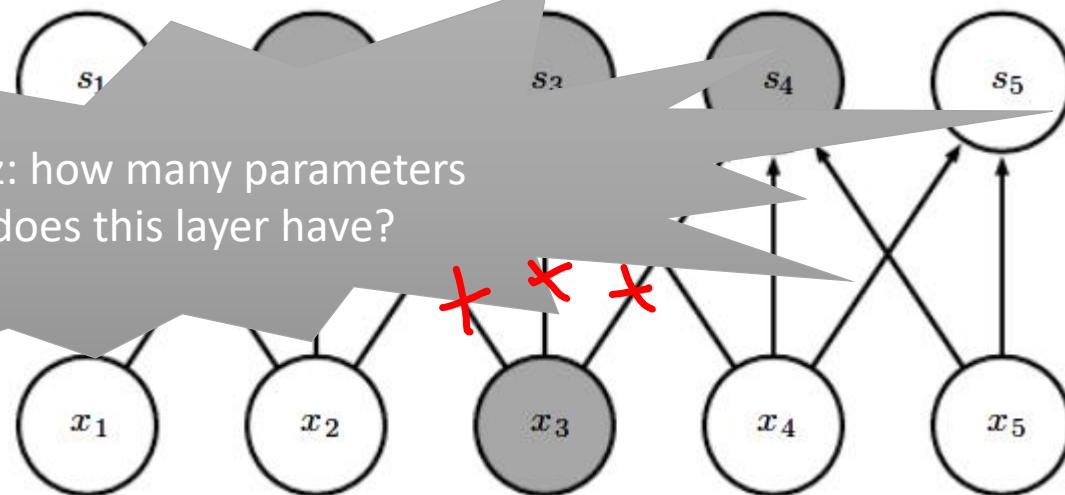
$5 \times 5 = 25$ weights
(+ 5 bias)

3x1 convolve

3 weights!
(+ 1 bias)



Quiz: how many parameters does this layer have?



To Summarize

Any CONV layer can be implemented by a FC layer performing exactly the same computations.

The weight matrix W of the FC layer would be

- a large matrix (#rows equal to the number of output neurons, #cols equal to the nr of input neurons).
- That is mostly zero except for at certain blocks where the local connectivity takes place.
- The weights in many of the blocks are equal due to parameter sharing.

... and we will see that the converse is also viable and very useful

The Receptive Field

A very important aspect in CNNs

The Receptive Field

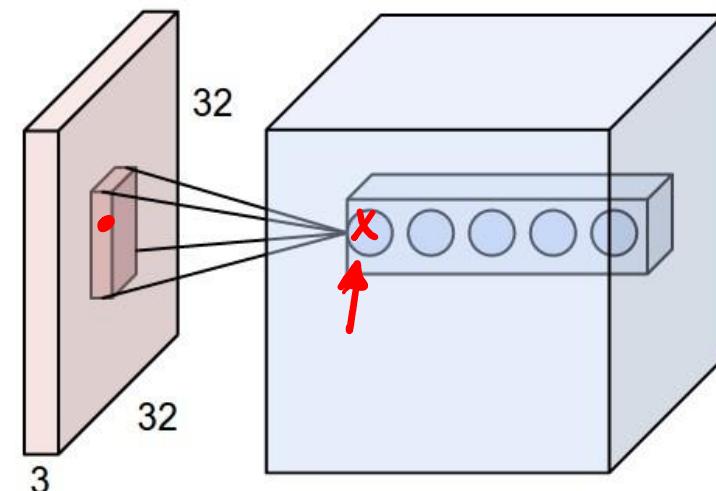
One of the basic concepts in deep CNNs.

Due to sparse connectivity, unlike in FC networks where the value of each output depends on the entire input, **in CNN each output only depends on a specific region in the input.**

This region in the input is the **receptive field for that output**

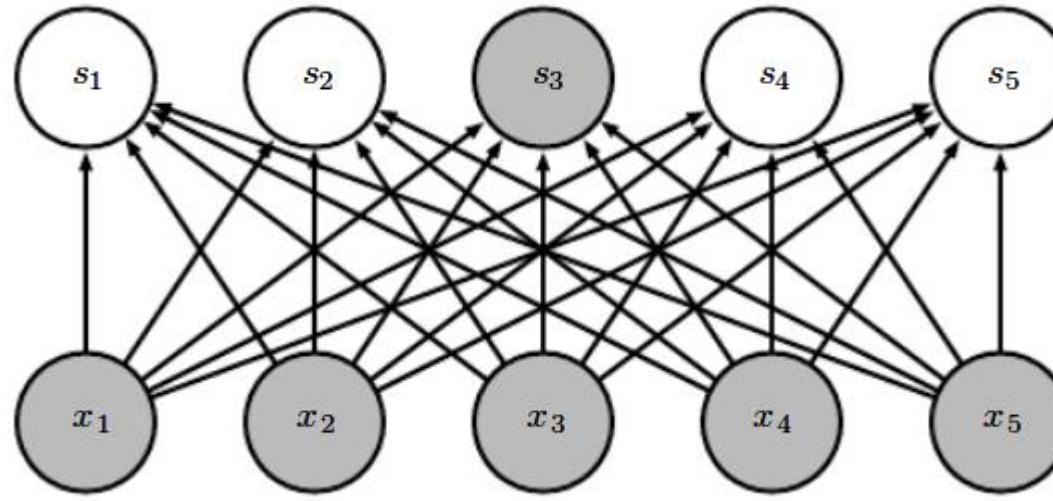
The **deeper you go, the wider the receptive field is:** maxpooling, convolutions and stride > 1 increase the receptive field

Usually, the receptive field refers to the final **output unit** of the network in relation to the network input, but the same definition holds for intermediate volumes

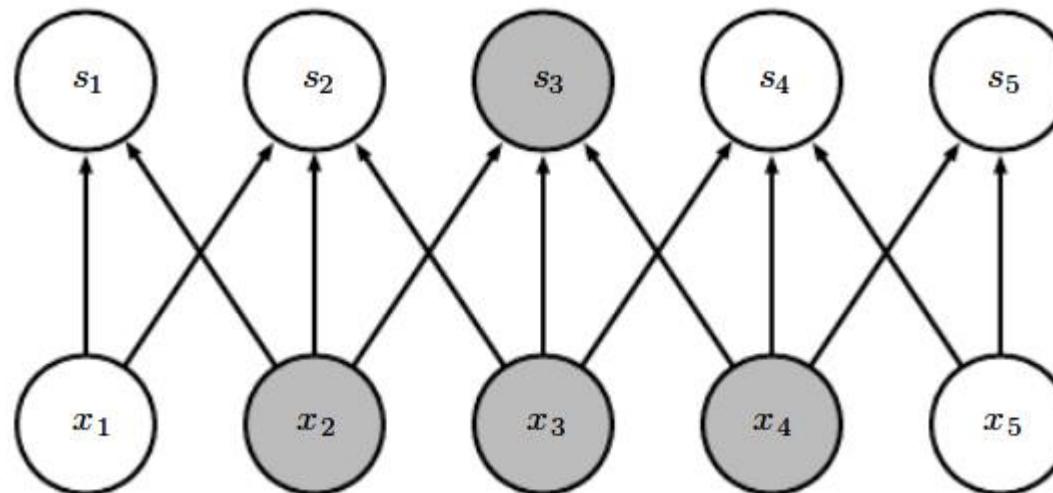


Receptive fields

MPL, Fully connected

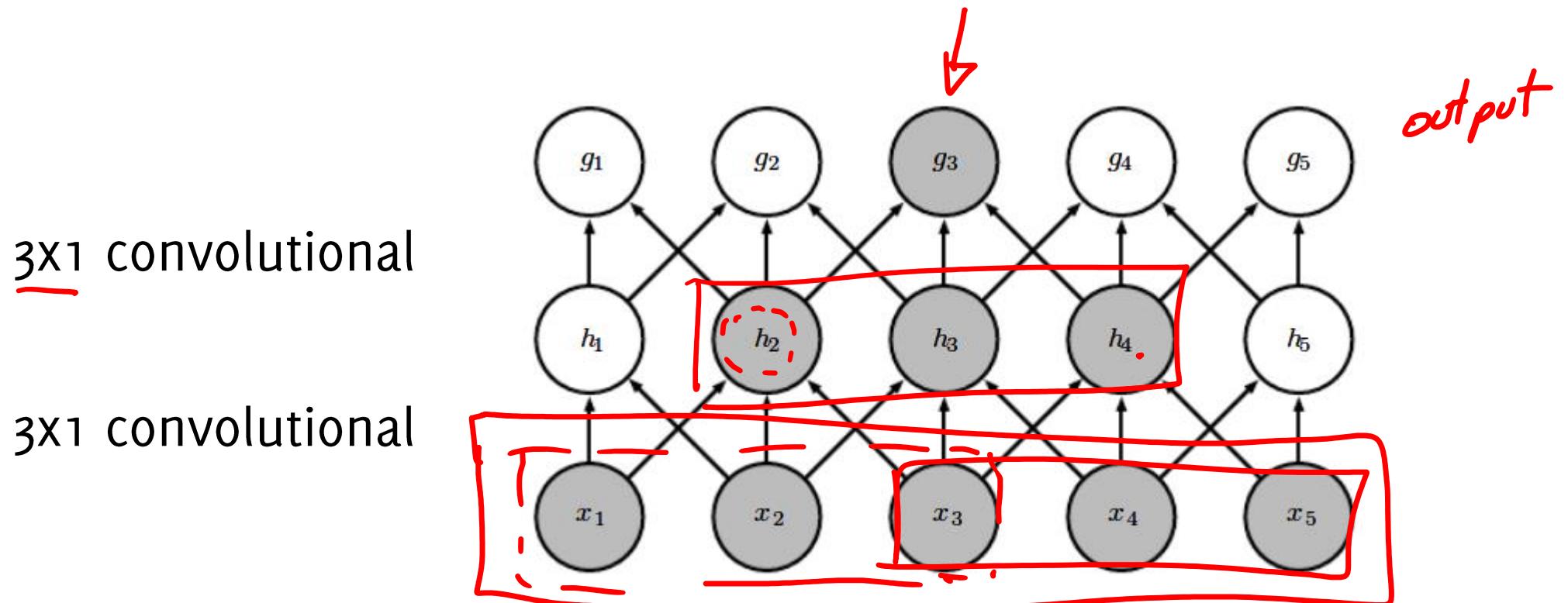


3x1 convolutional



Receptive fields

Deeper neurons depend on wider patches of the input (convolution is enough to increase receptive field, no need of maxpooling)



Exercise

Input:



Conv 3x3

Conv 3x3

Conv 3x3

Conv 3x3

Conv 3x3

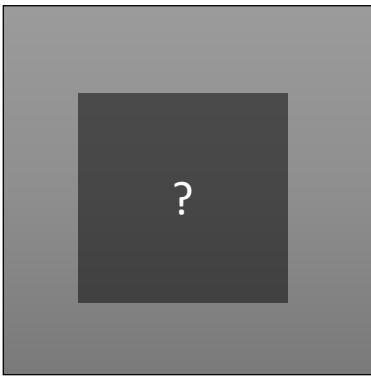
Conv 3x3

map



Receptive fields

Input



Conv 3x3

Conv 3x3

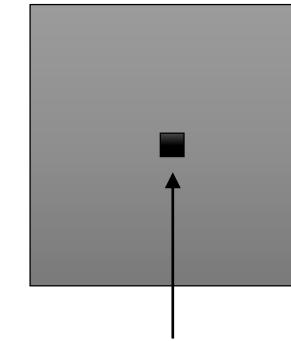
Conv 3x3

Conv 3x3

Conv 3x3

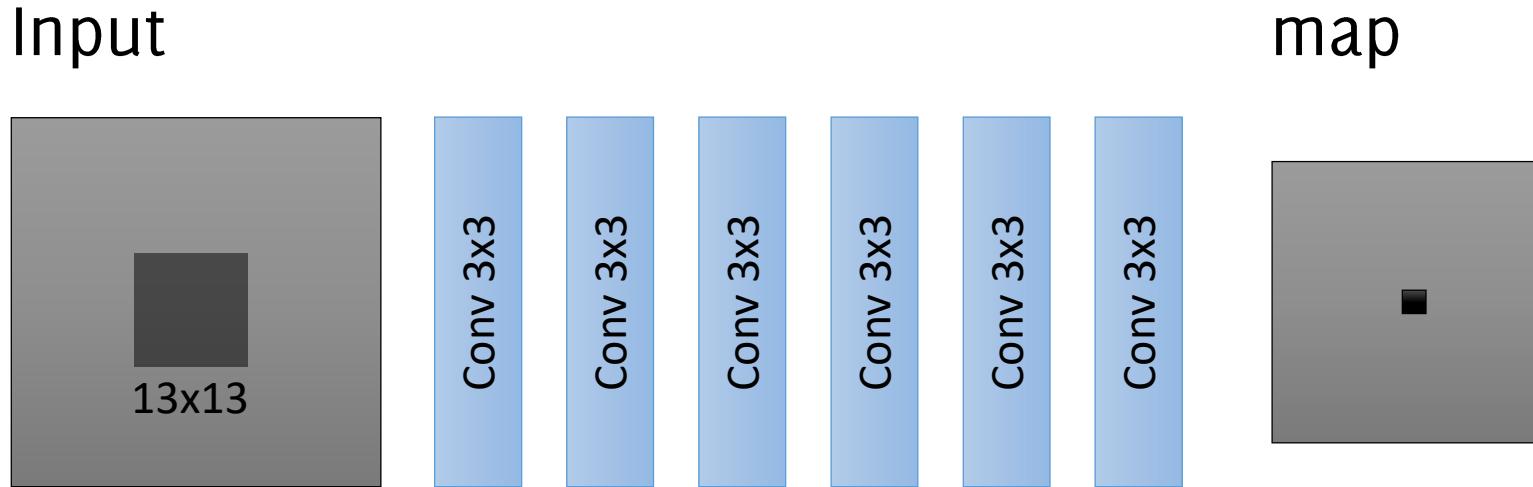
Conv 3x3

map



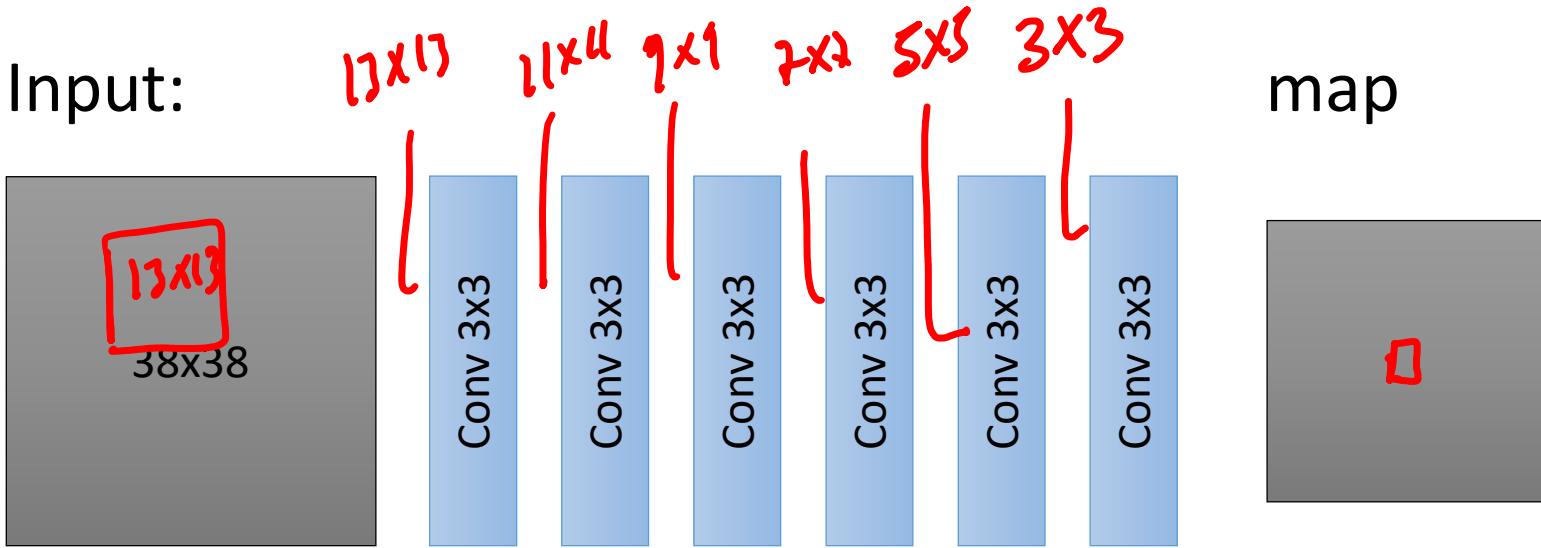
How large is the receptive field of the black neuron?

Receptive fields



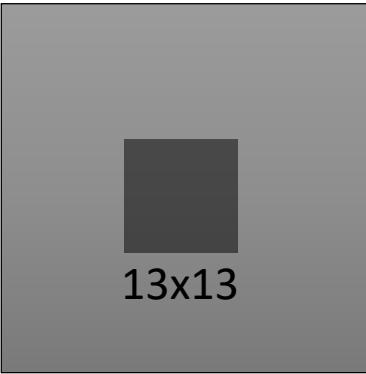
How large is the receptive field of the black neuron?

Exercise

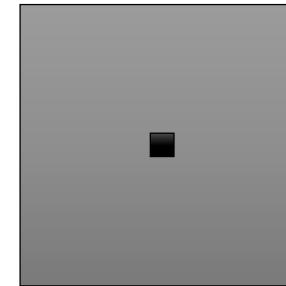


Receptive fields

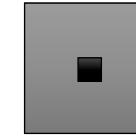
Input



map

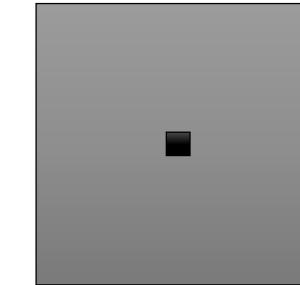
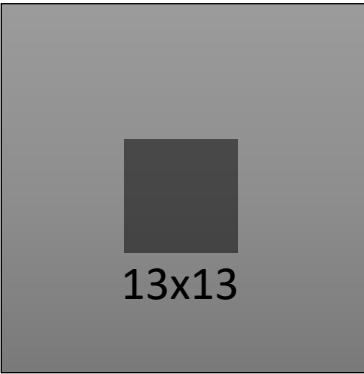


How large is the receptive field of the black neuron?



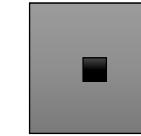
Receptive fields

Input



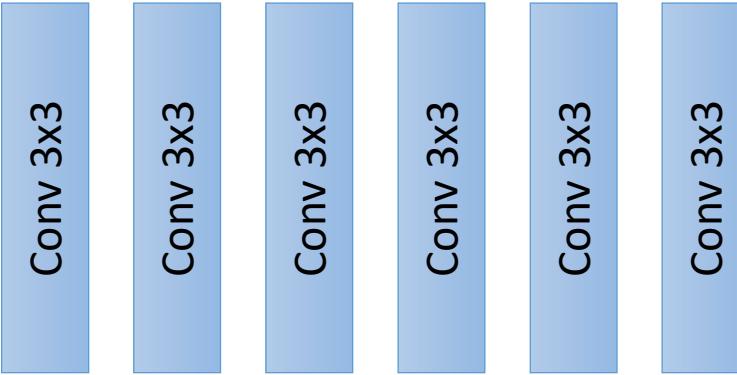
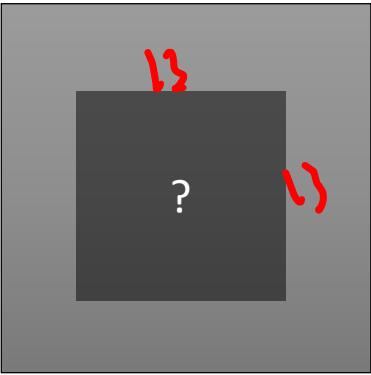
map

How large is the receptive field of the black neuron?

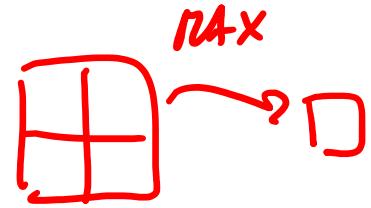
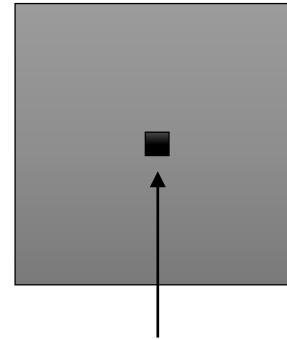


Receptive fields

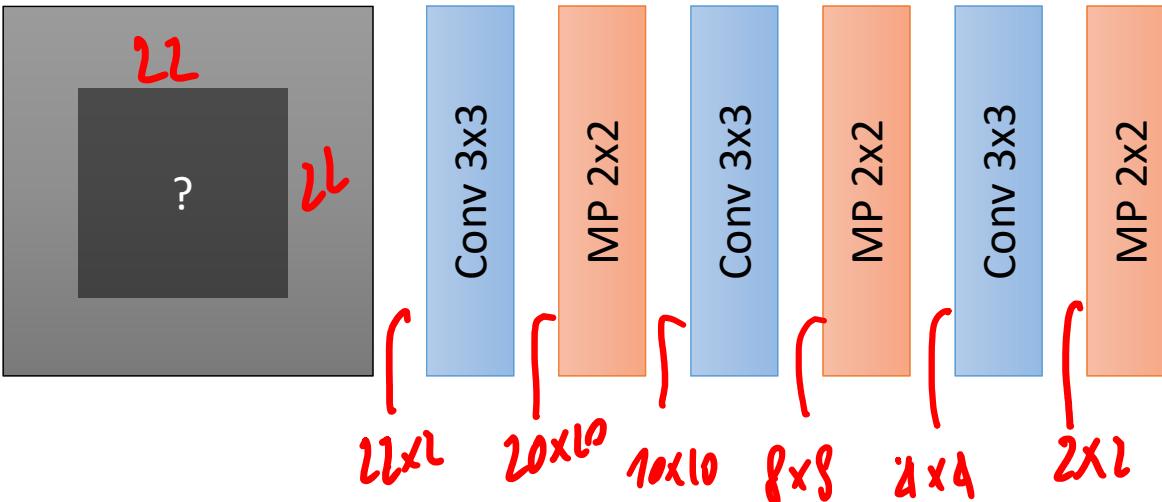
Input



map



How large is the receptive field of the black neuron?



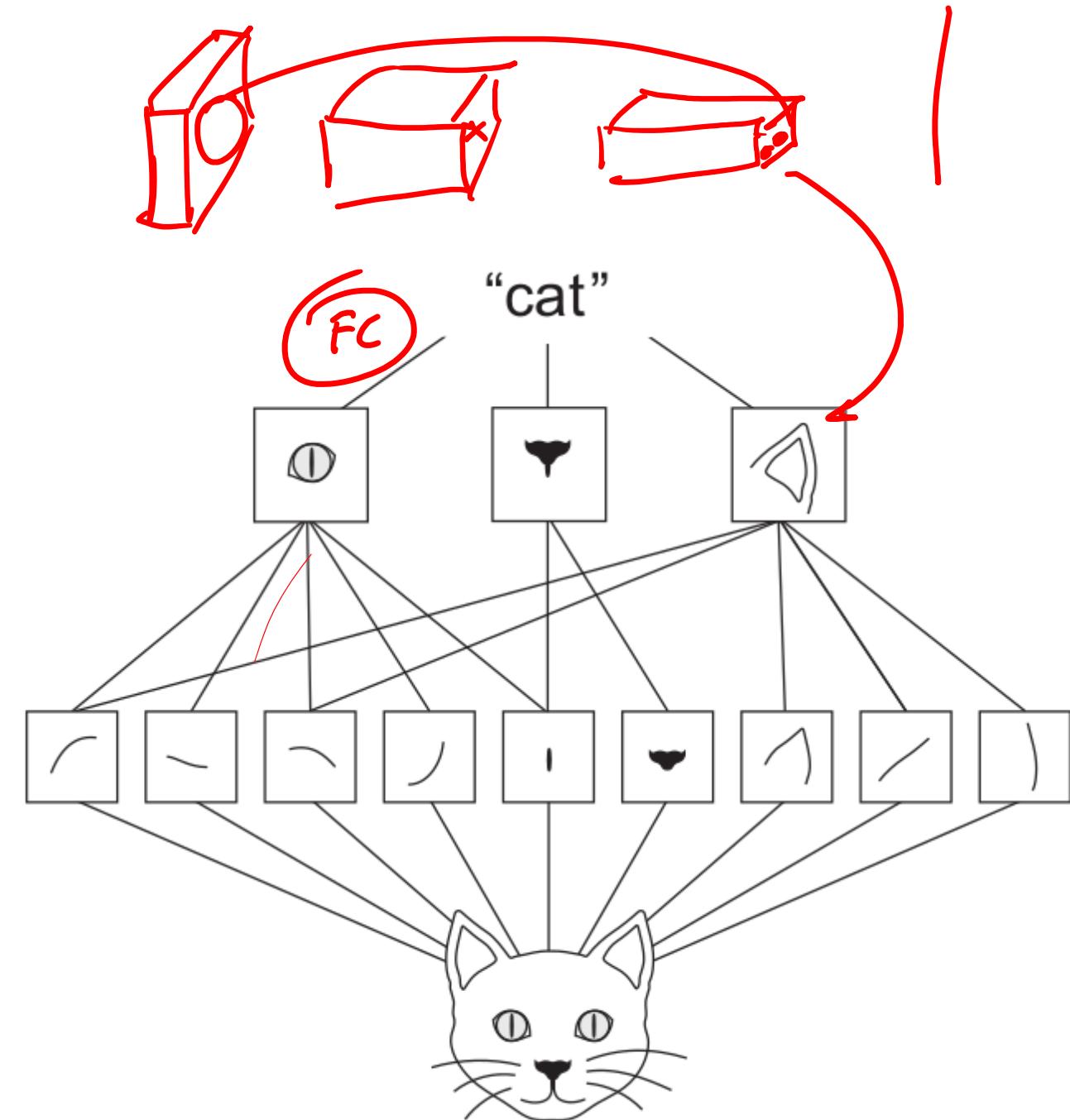
As we move deeper...

As we move to deeper layers:

- spatial resolution is reduced
- the number of maps increases

We search for higher-level patterns, and don't care too much about their exact location.

There are more high-level patterns than low-level details!



CNN Training

Training a CNN

- Each CNN can be seen as a MLP, with sparse and shared connectivities)
- CNN can be in principle trained by gradient descent to minimize a loss function over a batch (e.g. binary cross-entropy, RMSE, Hinge loss..)
- Gradient can be computed by backpropagation (chain rule) as long as we can derive each layer of the CNN
- Weight sharing needs to be taken into account (fewer parameters to be used in the derivatives) while computing derivatives
- There are just a few details missing...

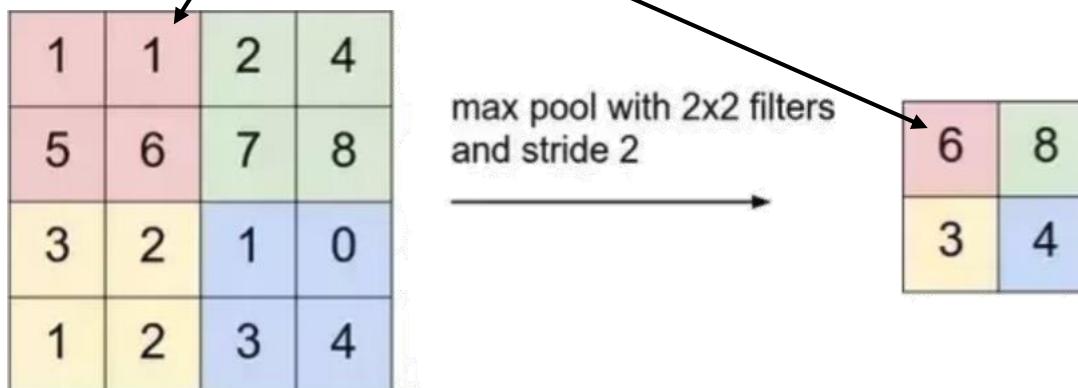
Detail: backprop with max pooling

The gradient is only routed through the input pixel that contributes to the output value; e.g.:

Gradient of \bullet with respect to $\bullet = 0$

The derivative is:

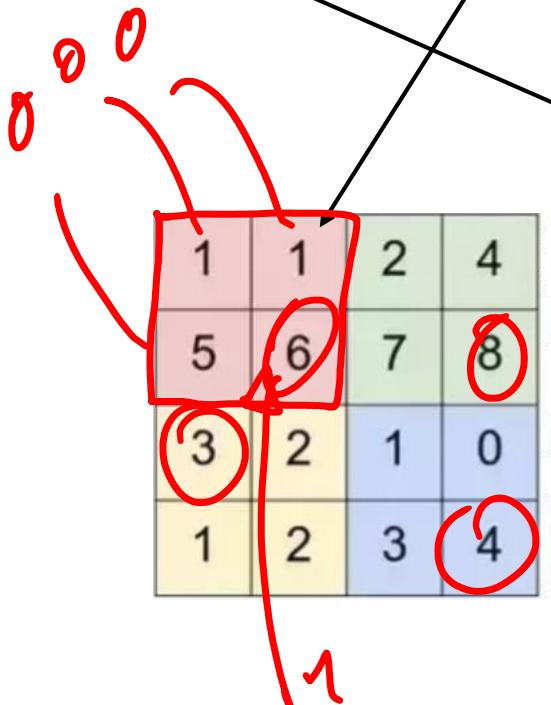
- 1 at the location corresponding the maximum
- 0 otherwise



Detail: backprop with max pooling

The gradient is only routed through the input pixel that contributes to the output value; e.g.:

Gradient of \bullet with respect to $\bullet = 0$



max pool with 2x2 filters
and stride 2

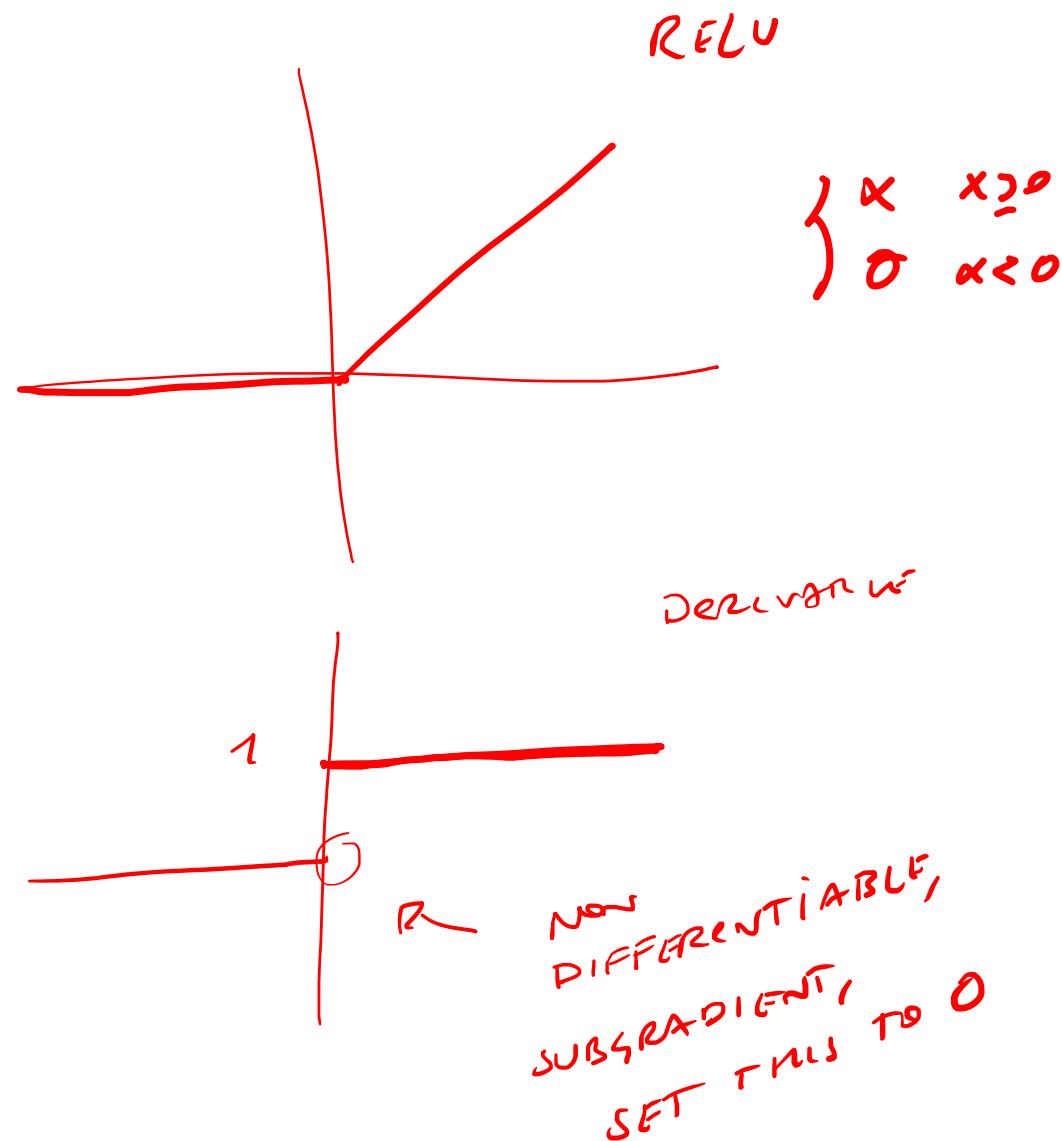


The derivative is:

- 1 at the location corresponding the maximum
- 0 otherwise

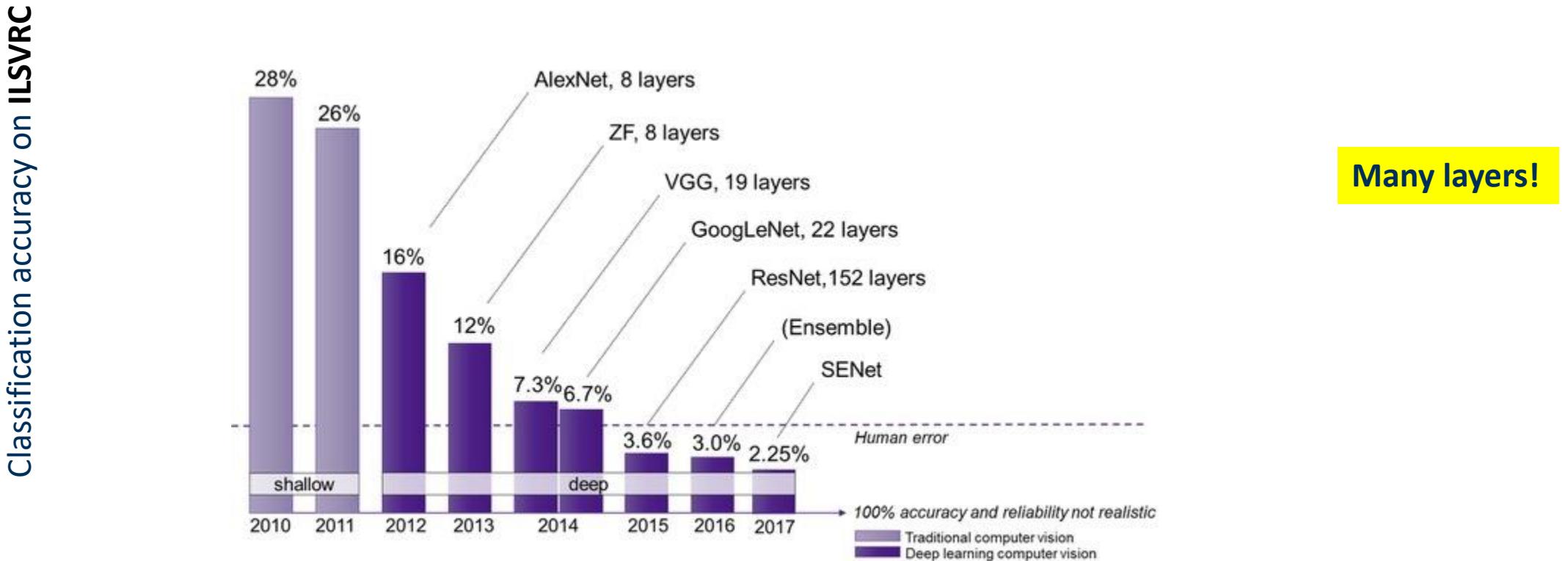
Detail: derivative of ReLU

The ReLU derivative is straightforward



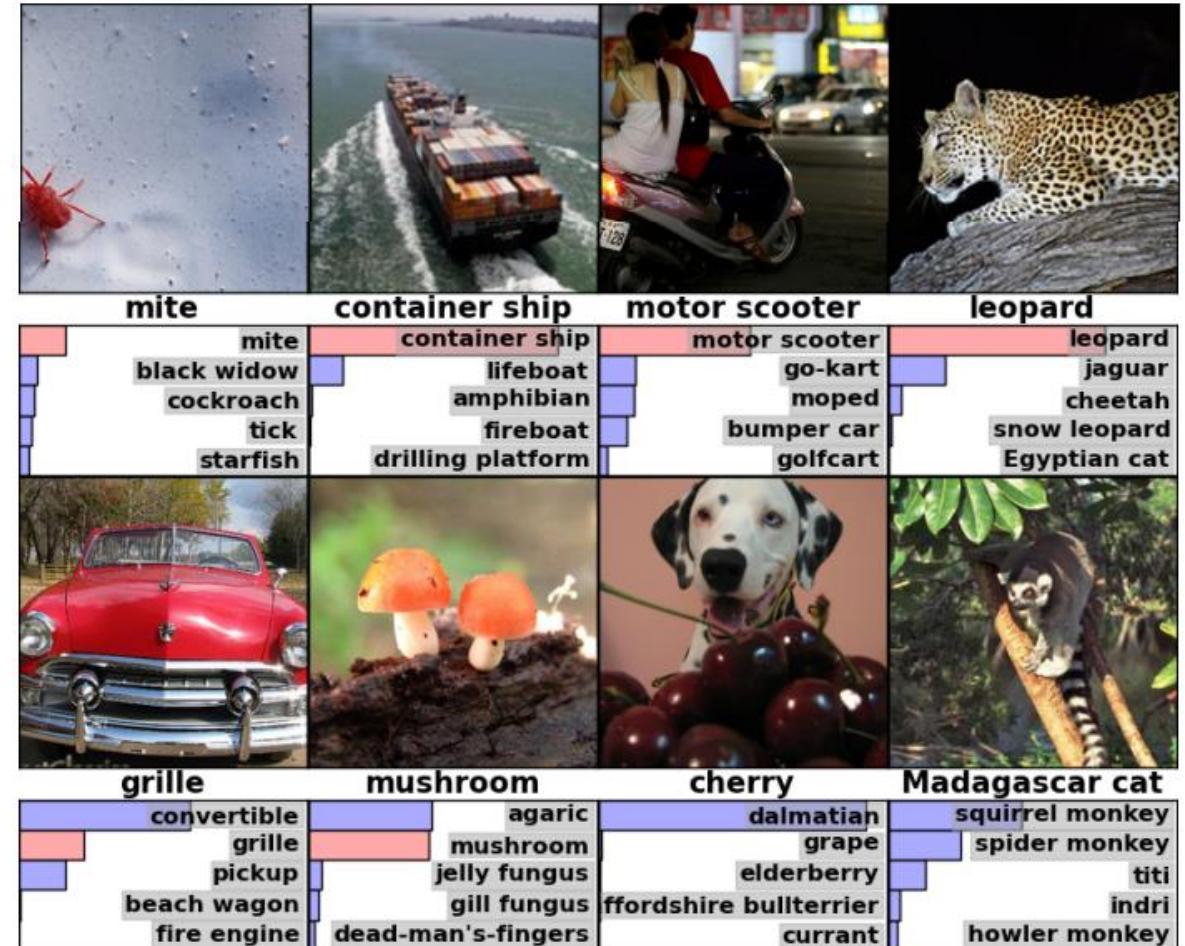
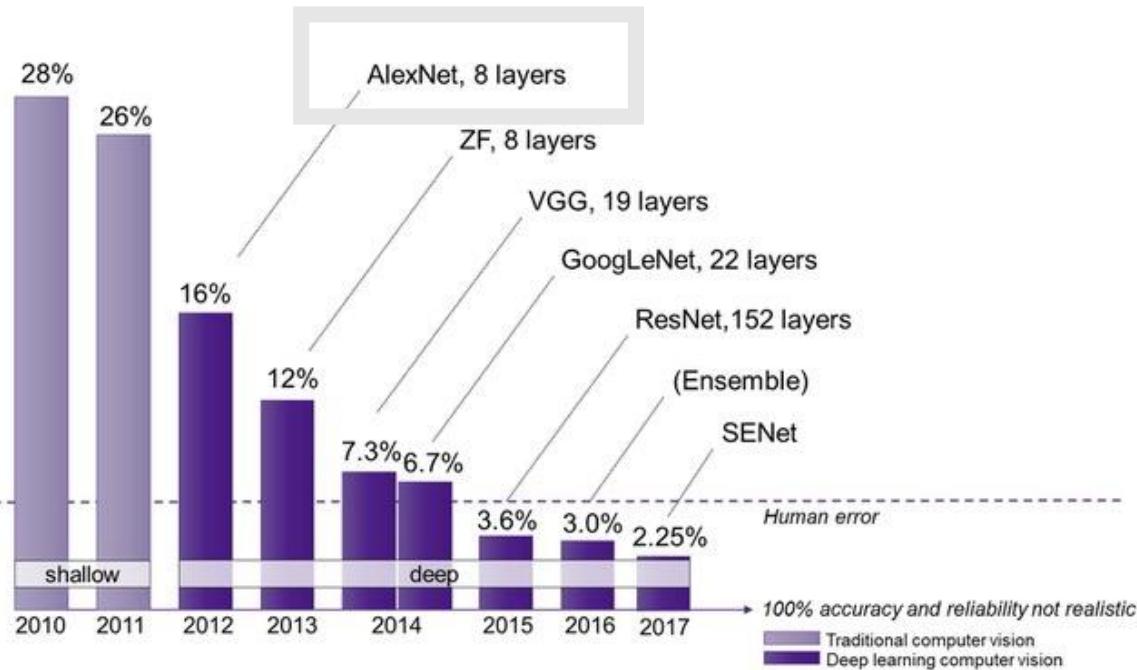
A Breakthrough in Image Classification

The impact of Deep Learning in Visual Recognition



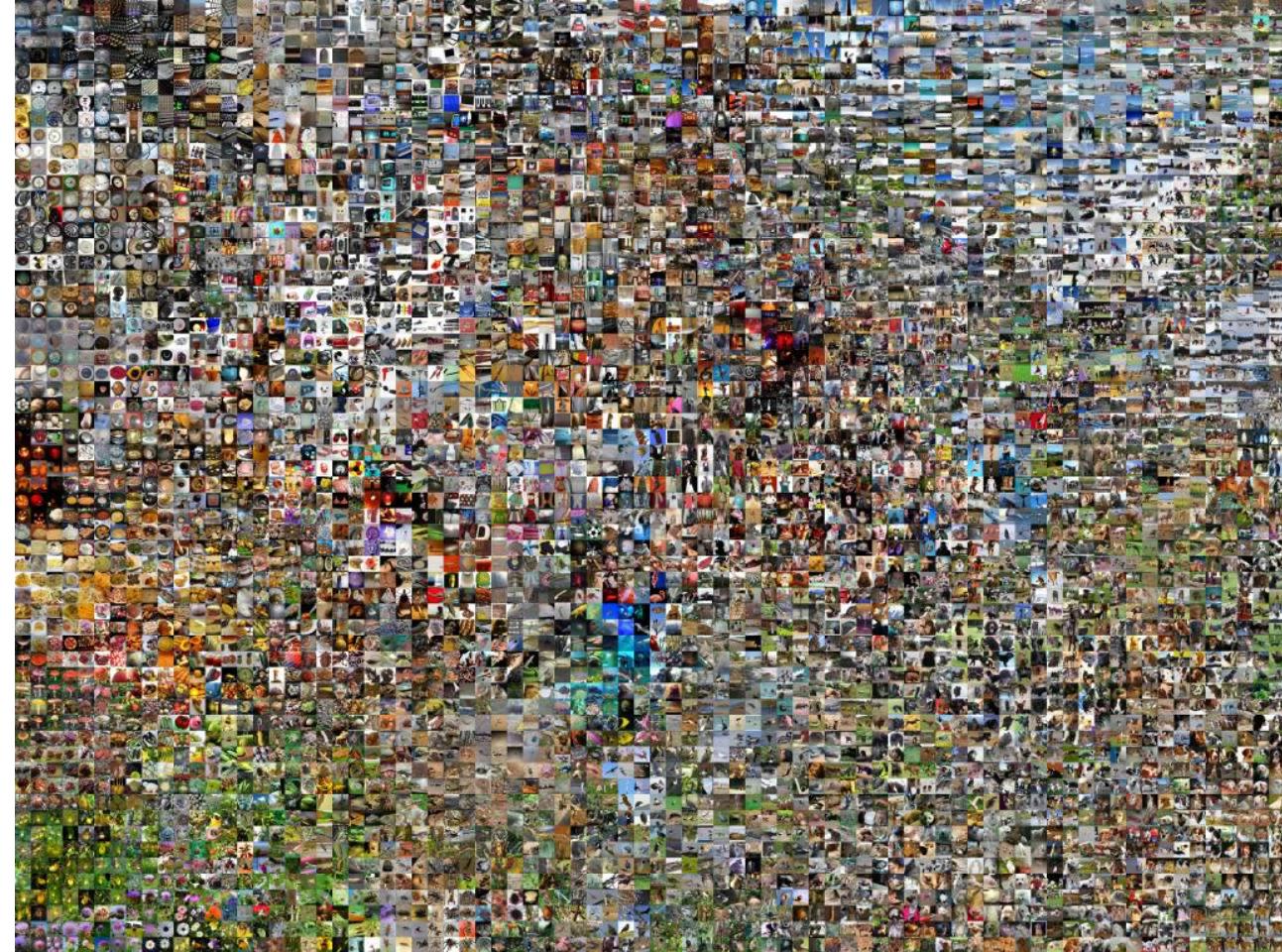
ILSVCR: ImageNet Large Scale Visual Recognition Challenge

AlexNet / Imagenet Images



How was this possible?

Large Collections of Annotated Data



The ImageNet project is a large visual database designed for use in visual object recognition software research. More than 14 million images have been hand-annotated by the project to indicate what objects are pictured and in at least one million of the images, bounding boxes are also provided.[3] ImageNet contains more than 20,000 categories

From Wikipedia October 2021

Parallel Computing Architectures



And more recently.... Software libraries



TensorFlow

Google LLC, Public domain, via Wikimedia Commons



PyTorch, BSD <<http://opensource.org/licenses/bsd-license.php>>, via Wikimedia Commons

Data Scarcity

Training a CNN with Limited Amount of Data

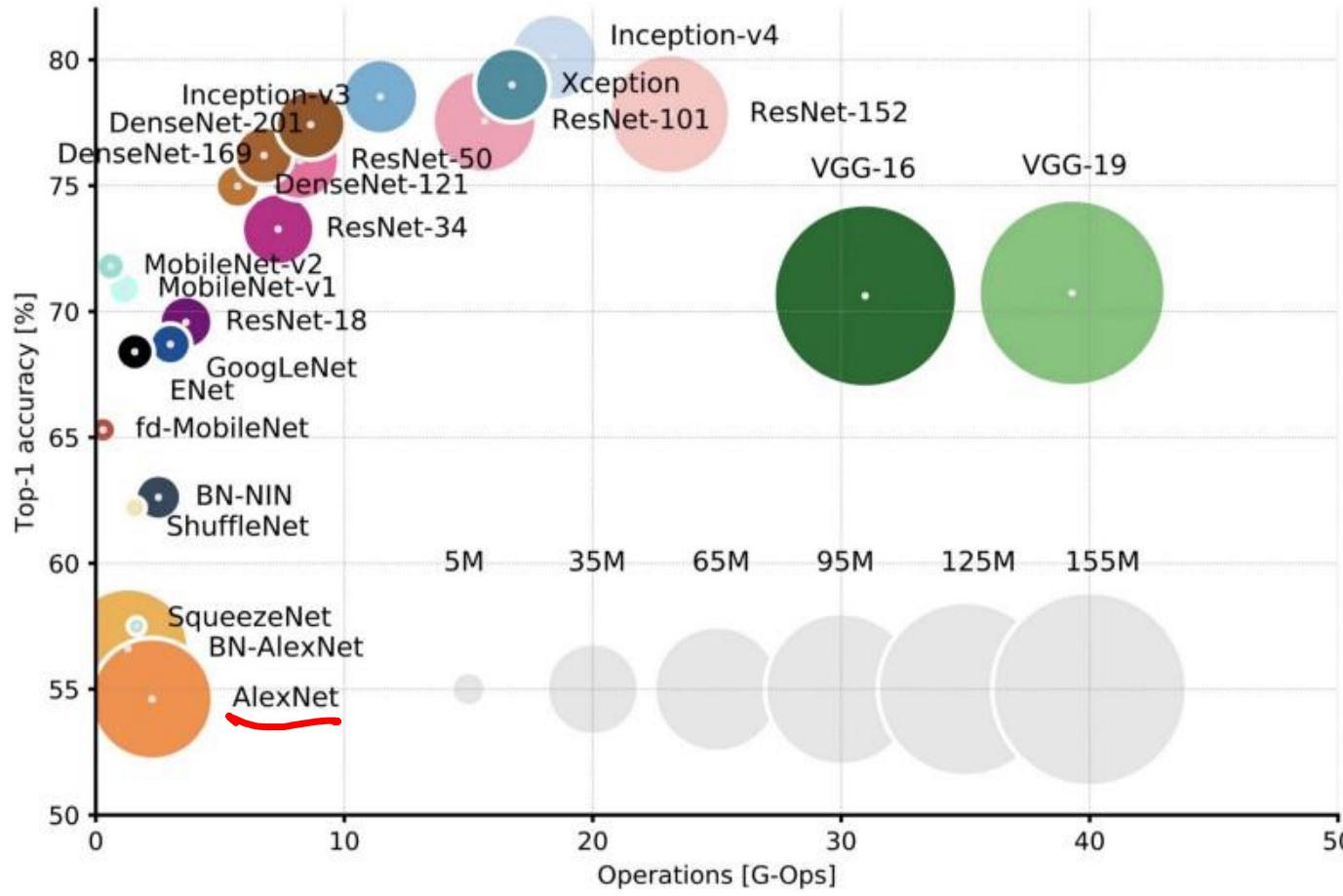
The need of data

Deep learning models are very data hungry.

Networks such as AlexNet have been trained on ImageNet datasets containing tens of thousands of images over hundreds of classes

The need of data

This is necessary to define millions of parameters characterizing these networks



The need of data

Deep learning models are very data hungry.

... watch out: each image in the training set have to be annotated!

How to train a deep learning model with a few training images?

- Data augmentation
- Transfer Learning

Limited Amount of Data: Data Augmentation

Training a CNN with Limited Aumont of Data



Aleutian Islands

Steller sea lions in the western Aleutian Islands have declined 94% in the last 30 years.



Kaggle in 2017 have opened a competition to develop algorithms which accurately count the number of sea lions in aerial photographs

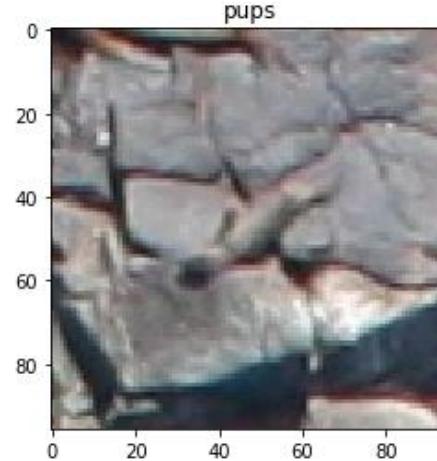
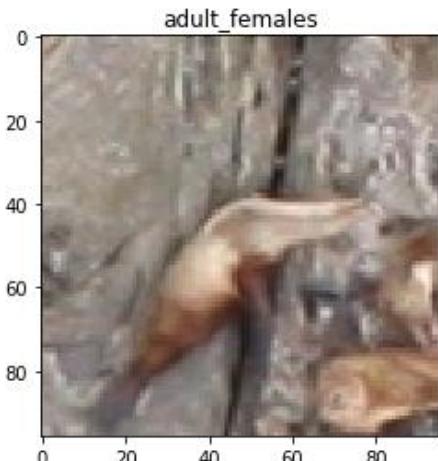
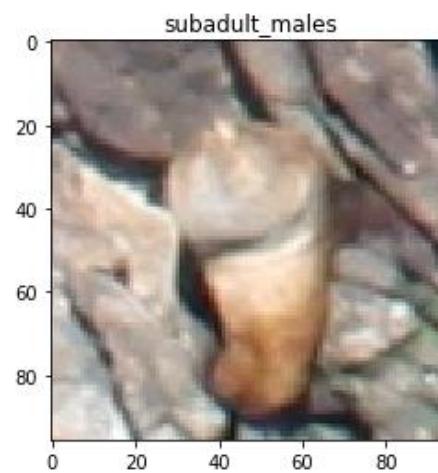
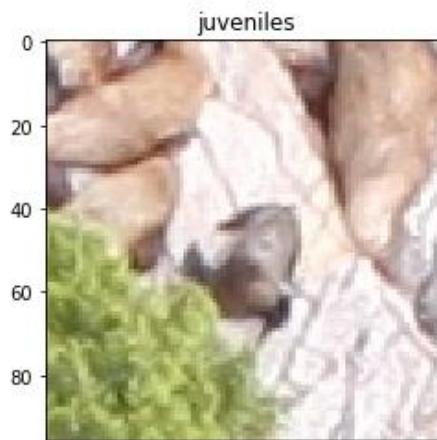
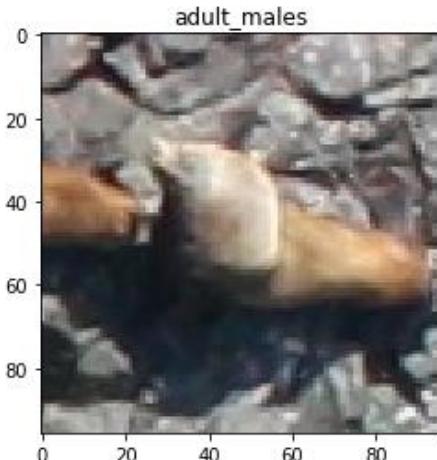


Credits Yinan Zhou

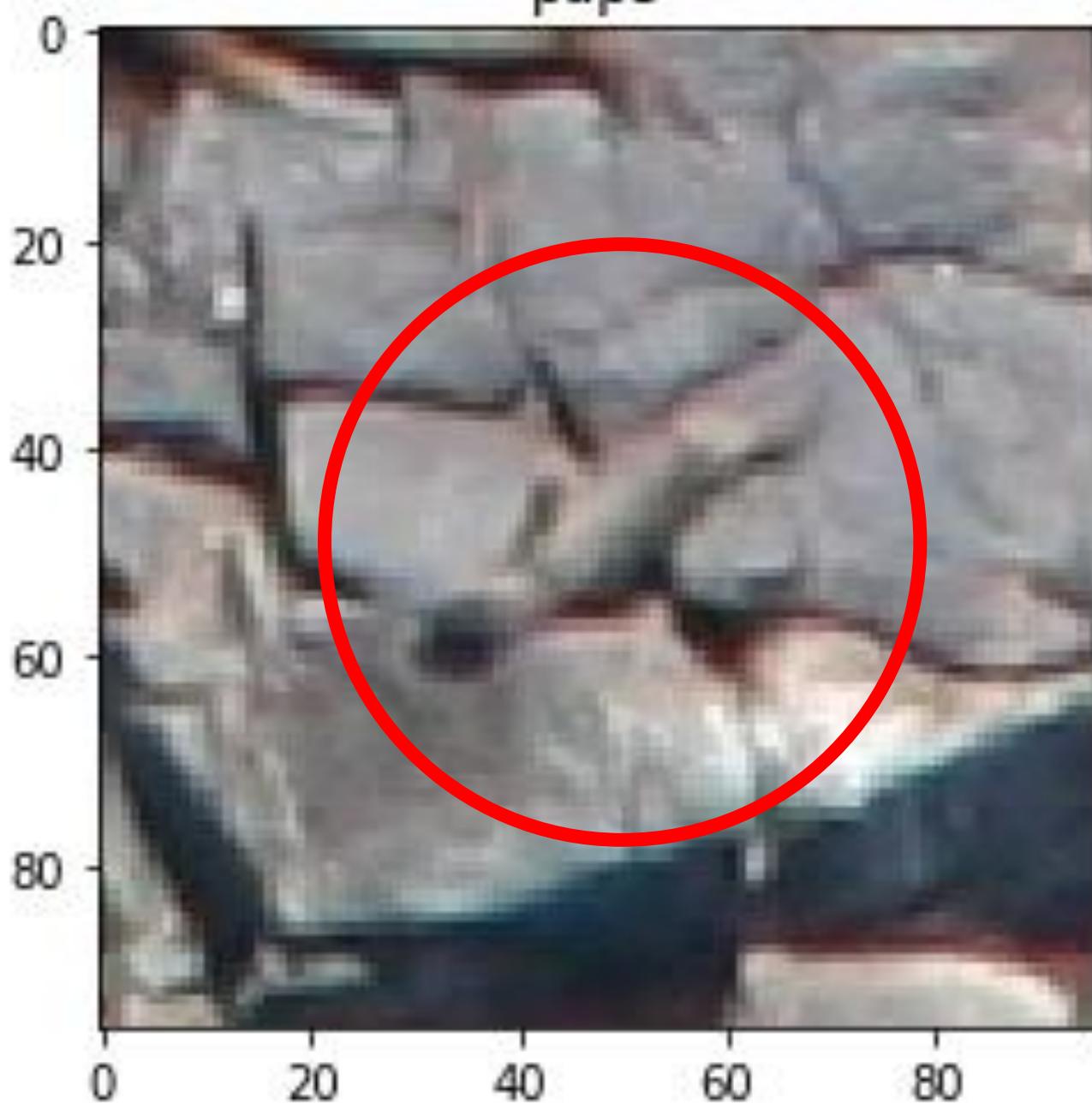
<https://github.com/marioZYN/FC-CNN-Demo>

The Challenge

In very large aerial images ($\approx 5K \times 4K$) shot by drones, automatically count the number of sealions per each category

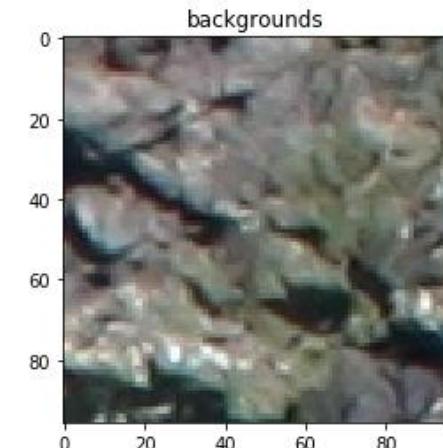
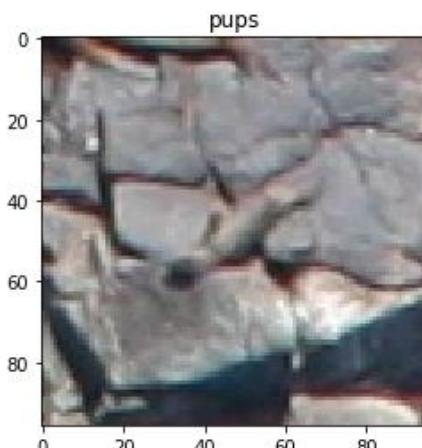
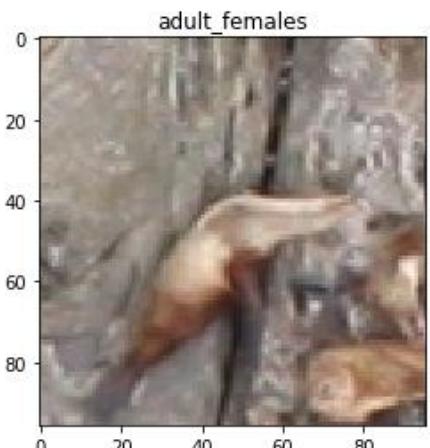
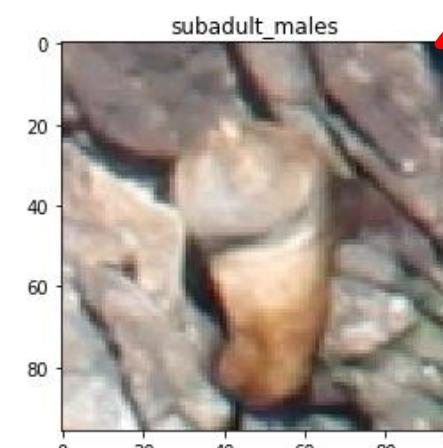
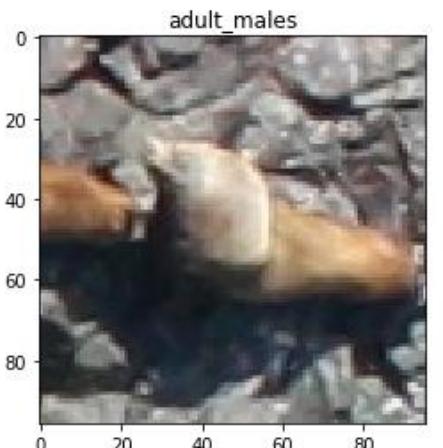


pups

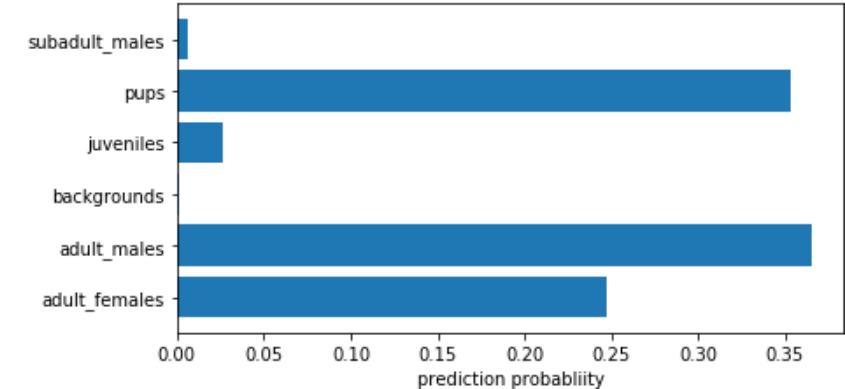
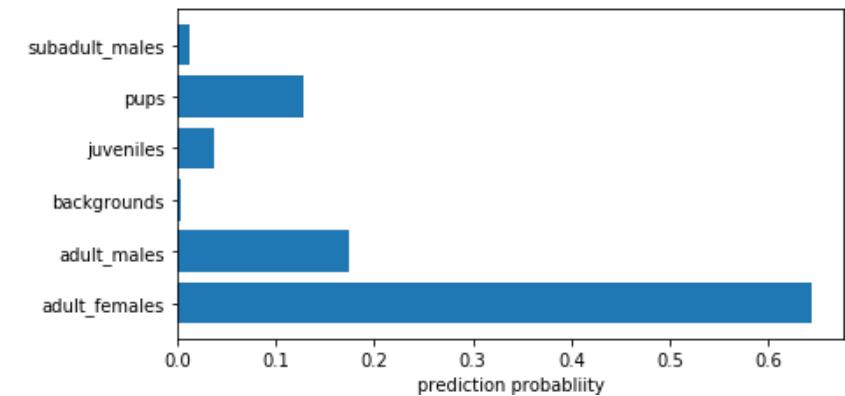
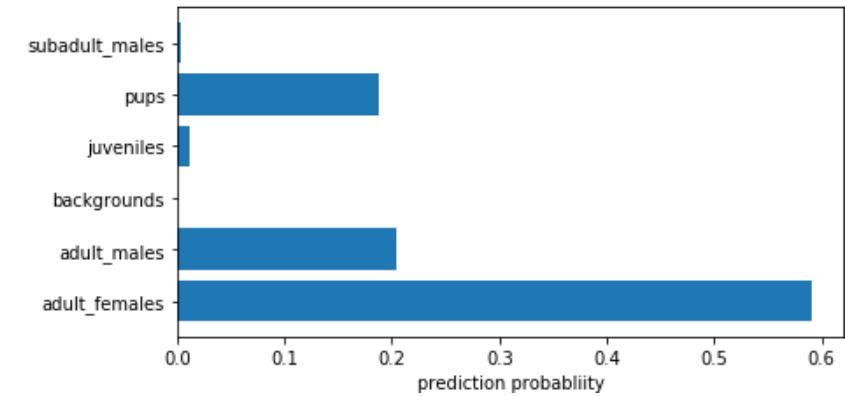
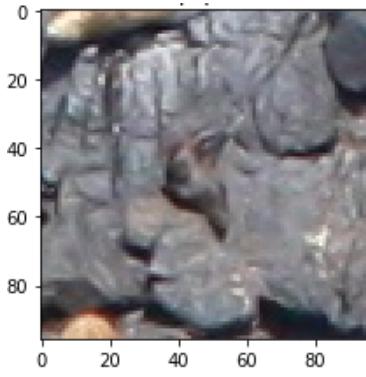
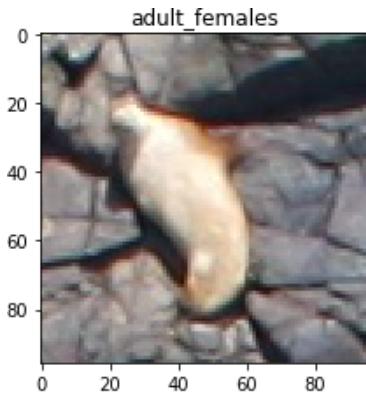
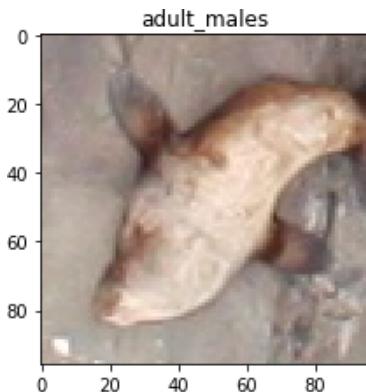


The Challenge

This problem can be naively casted in a patch-by-patch 6-class classification problem, where we include also background



An Example of CNN predictions



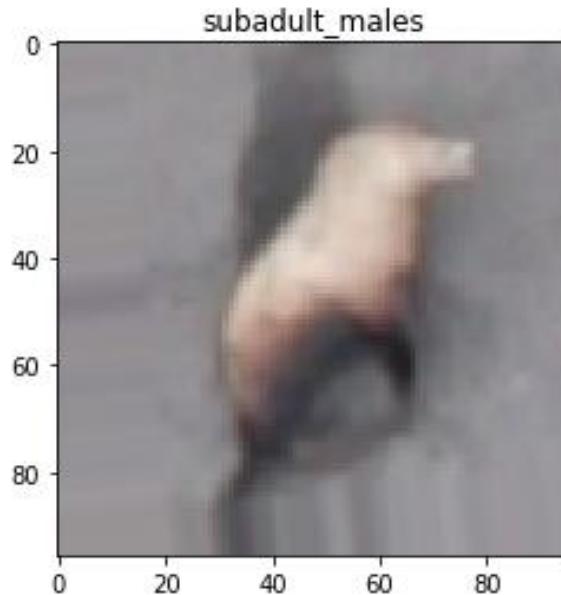
Credits Yinan Zhou

<https://github.com/marioZYN/FC-CNN-Demo>

Data Augmentation

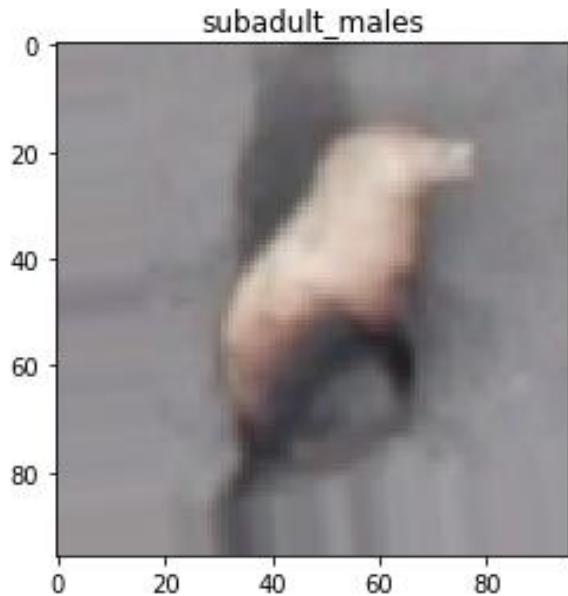
Often, each annotated image represents a class of images that are all likely to belong to the same class

In aerial photographs, for instance, it is normal to have rotated, shifted or scaled images without changing the label

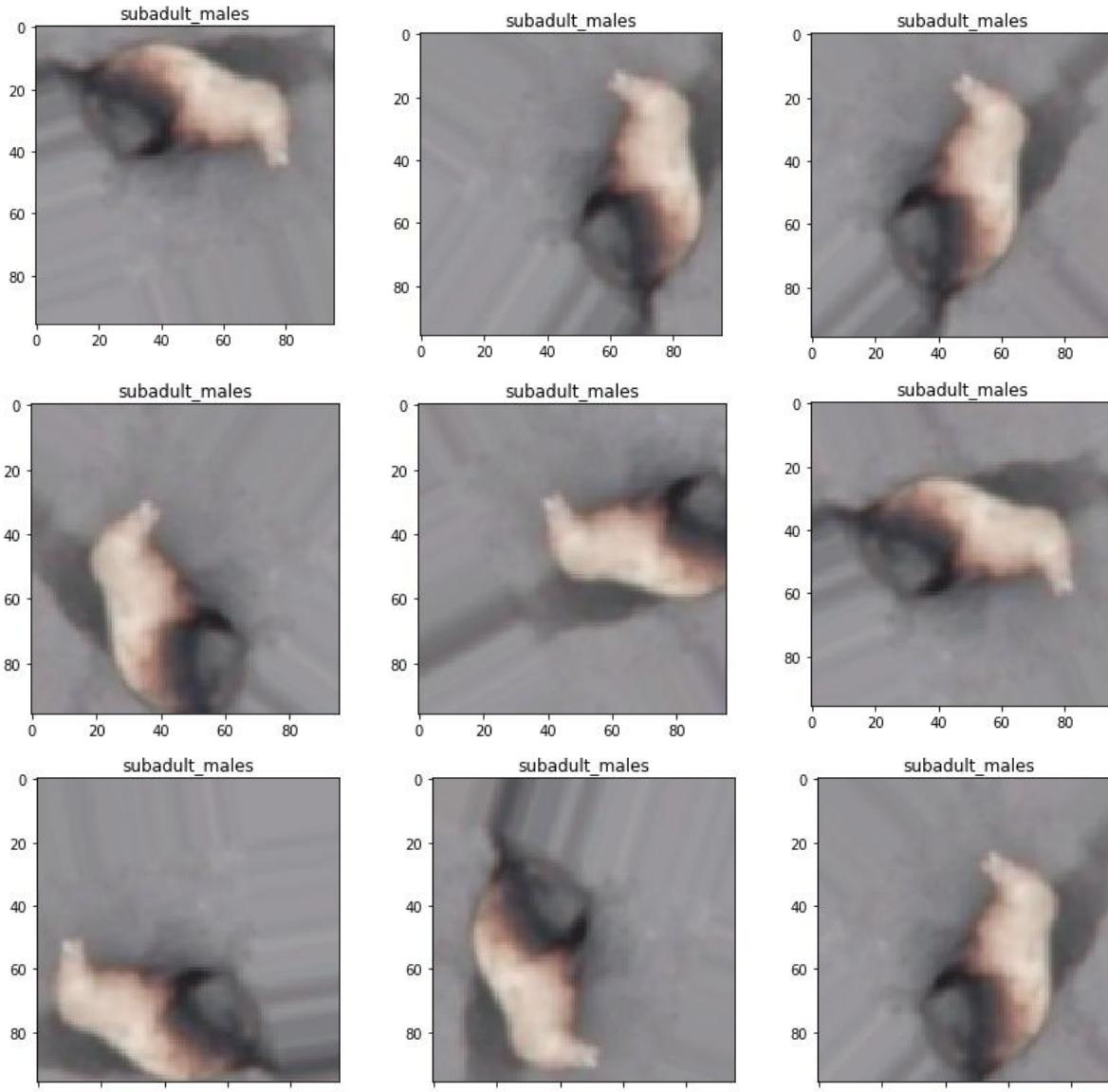


Data Augmentation

Original image



Augmented Images



Data Augmentation

Data augmentation is typically performed by means of

Geometric Transformations:

- Shifts /Rotation/Affine/perspective distortions
- Shear
- Scaling
- Flip

Photometric Transformations:

- Adding noise
- Modifying average intensity
- Superimposing other images
- Modifying image contrast

Data Augmentation: Criteria

Augmented versions **should preserve the input label**

- e.g. if size/orientation is a key information to determine the output target (either the class or the value in case of regression), wisely consider scaling/rotation as transformation

Augmentation is meant to **promote network invariance w.r.t.** transformation used for augmentation

Augmentation in Keras

```
from keras.preprocessing.image import  
ImageDataGenerator  
  
ImageDataGenerator(  
    rotation_range=0,  
    width_shift_range=0.0, height_shift_range=0.0,  
    brightness_range=None, shear_range=0.0,  
    zoom_range=0.0, channel_shift_range=0.0,  
    fill_mode='nearest',  
    horizontal_flip=False, vertical_flip=False,  
    rescale=None,  
    preprocessing_function=None)
```

Augmentation in Keras: flow from images

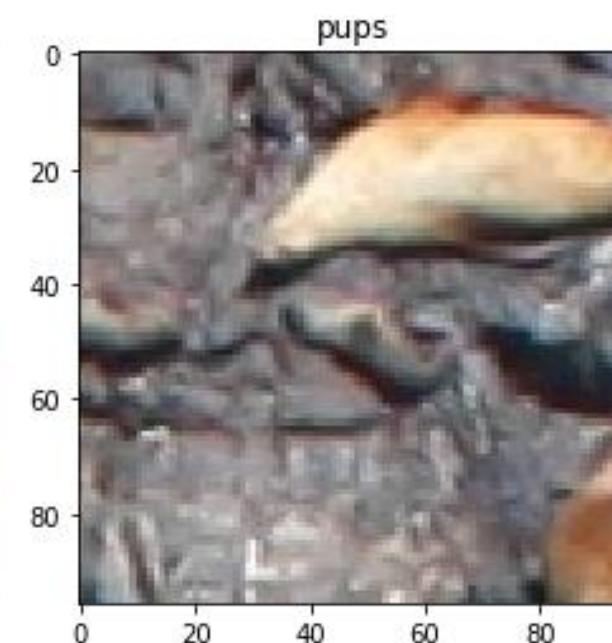
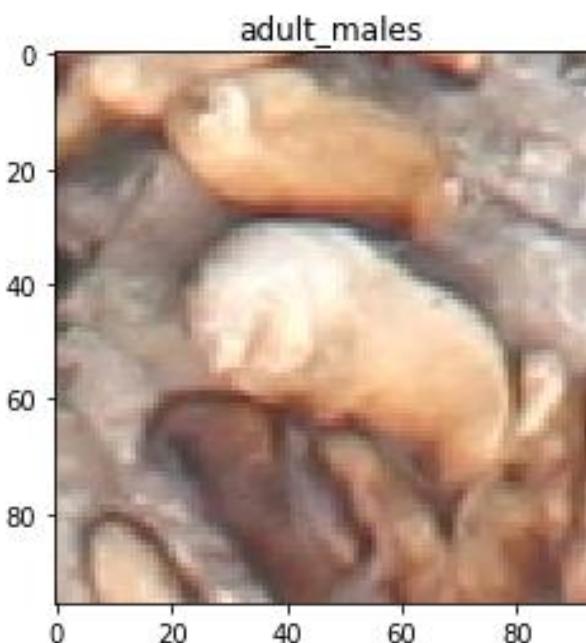
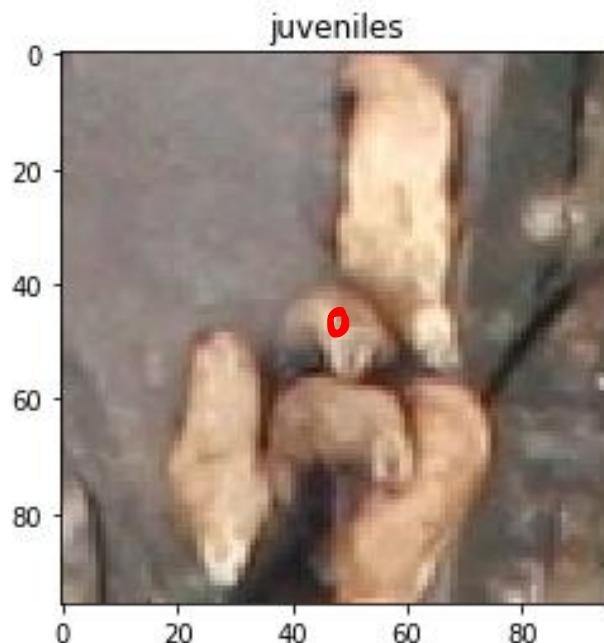
The Image generator has a method `flow_from_directory` that allows to load images in folder where different classes are arranged in subfolders.

```
ImageDataGenerator.flow_from_directory(  
    directory=PATCH_PATH + 'train/' ,  
    target_size=(img_width, img_width) ,  
    batch_size=batch_size ,  
    shuffle=True)
```

However...

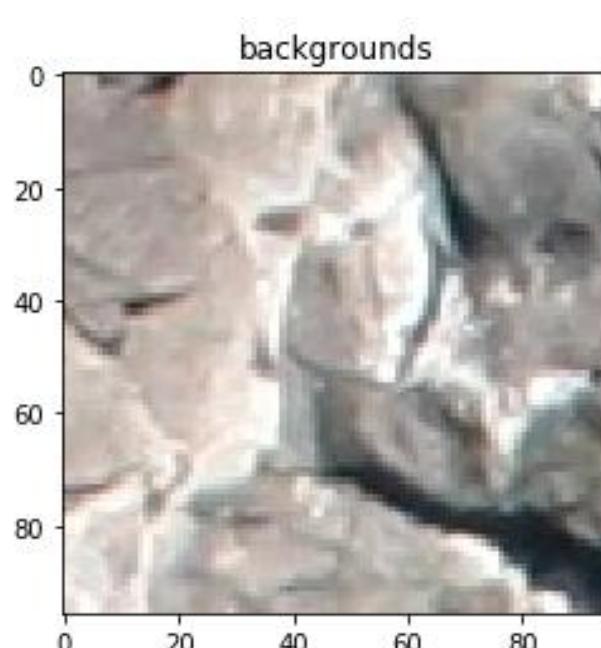
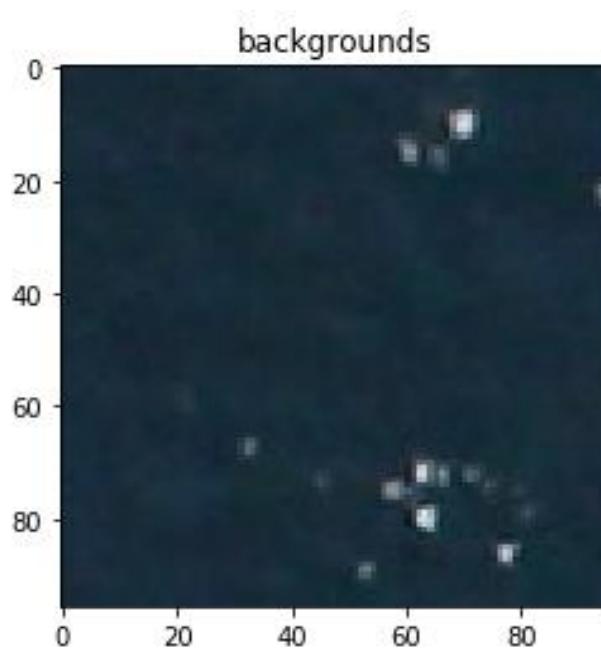
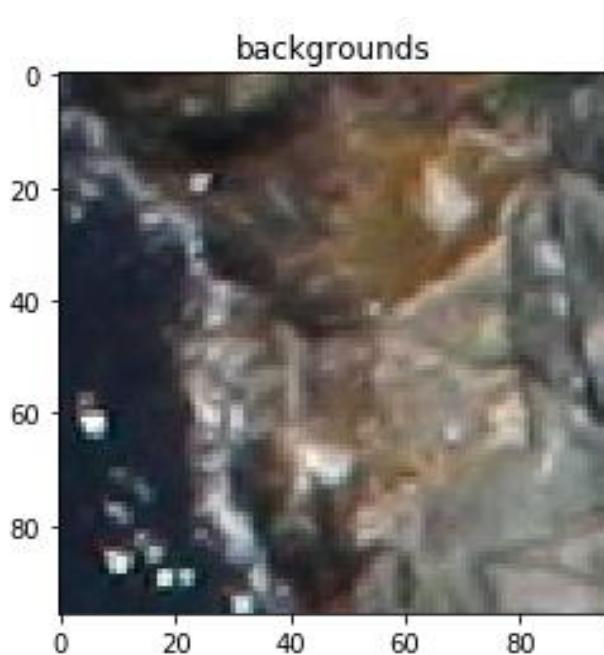
This sort of data augmentation might not be enough to capture the inter-class variability of images...

Superimposition of targets



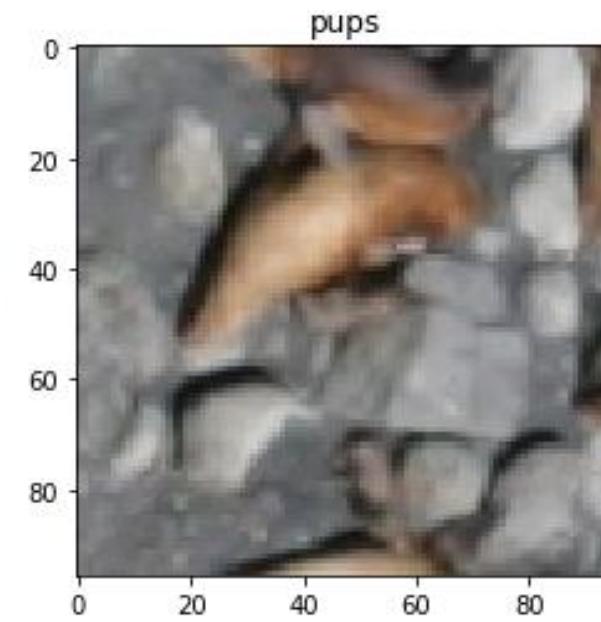
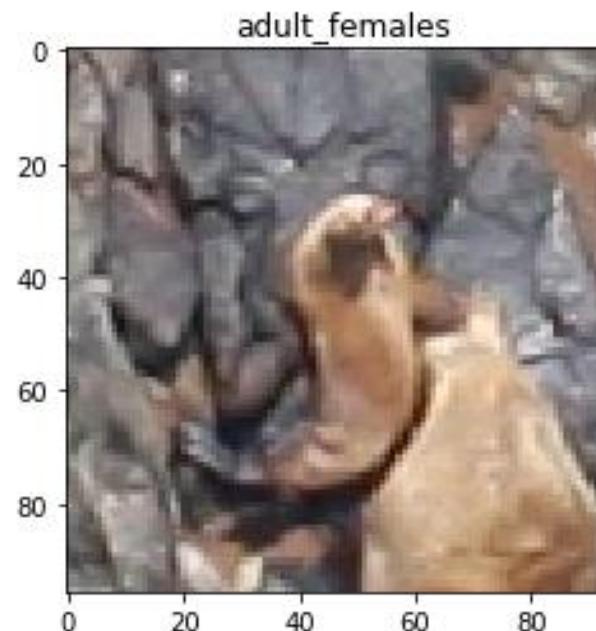
However...

Background variations



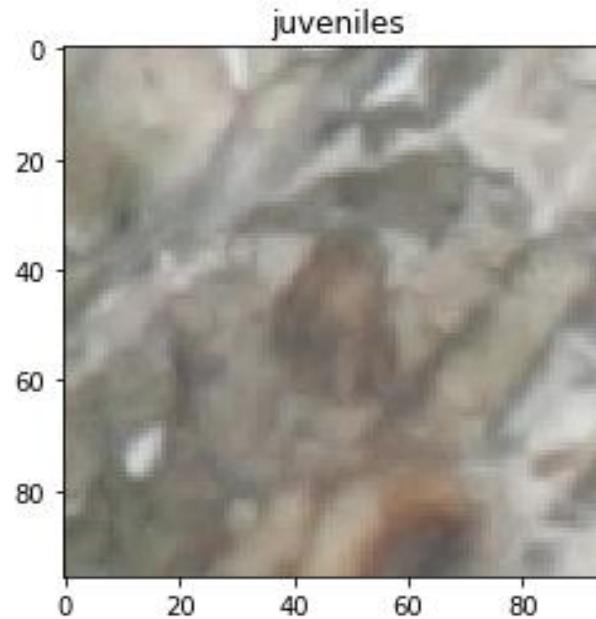
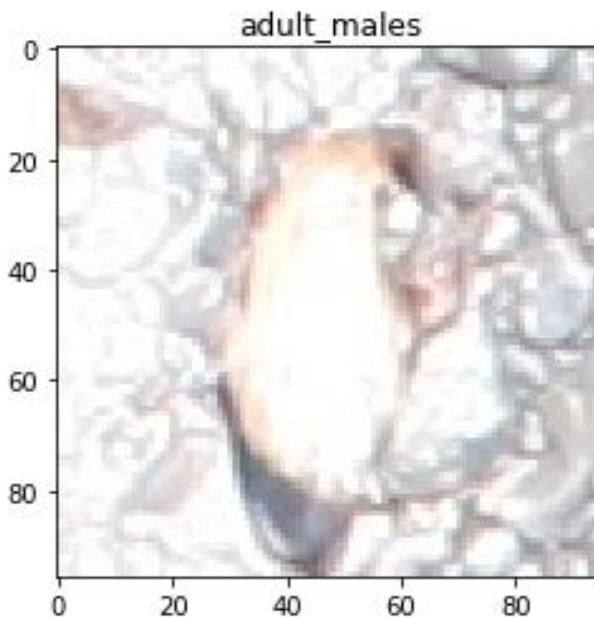
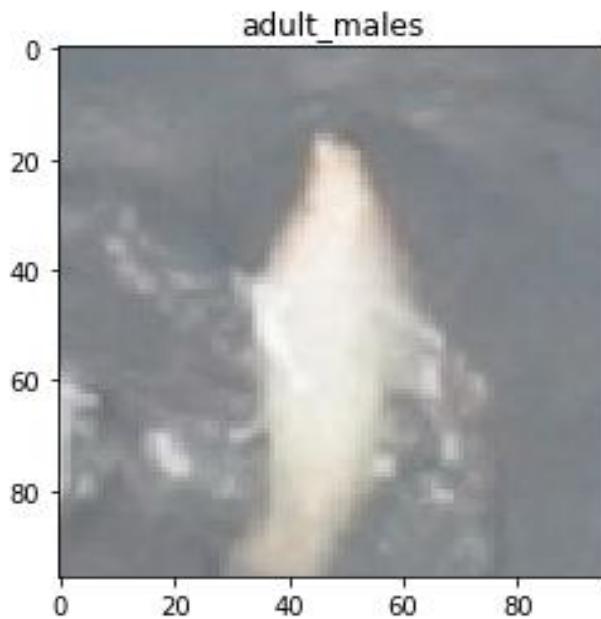
However...

Background variations



However...

Out of focus, bad exposure



Augmentation in Keras

```
from keras.preprocessing.image import  
ImageDataGenerator  
  
ImageDataGenerator(  
    rotation_range=0,  
    width_shift_range=0.0, height_shift_range=0.0,  
    brightness_range=None, shear_range=0.0,  
    zoom_range=0.0, channel_shift_range=0.0,  
    fill_mode='nearest',  
    horizontal_flip=False, vertical_flip=False,  
    rescale=None,  
    preprocessing_function=None)
```



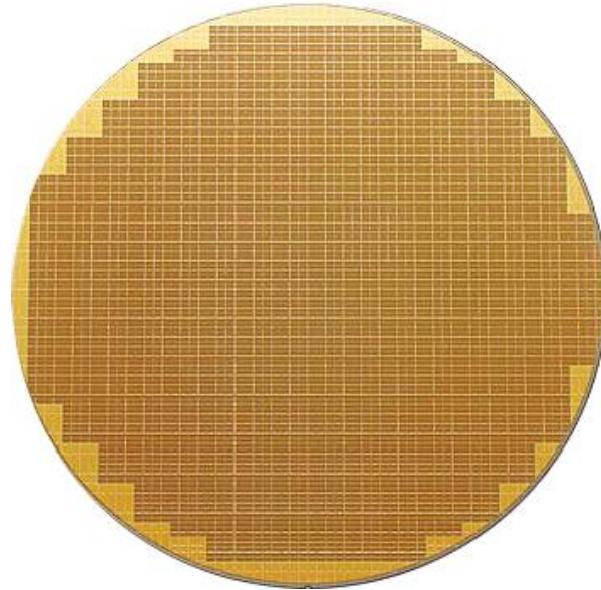
... in case you need some extra flexibility

Data Augmentation is often key..

Collaboration with  life.augmented

For instance to monitor wafer manufacturing

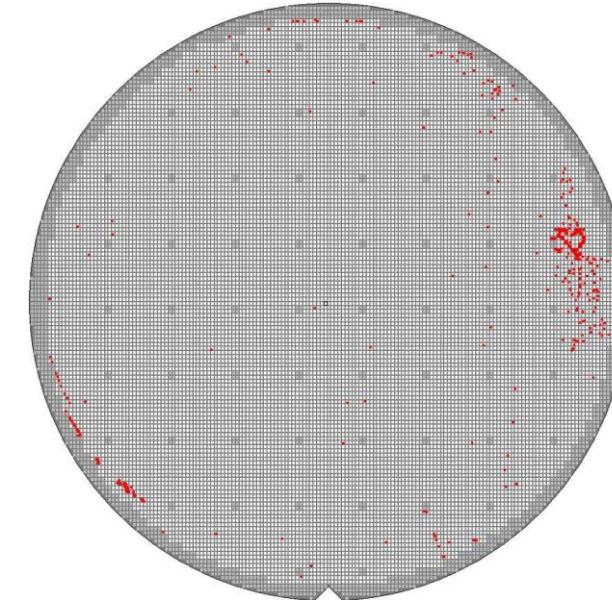
Wafer (containing chips)



Inspection tool



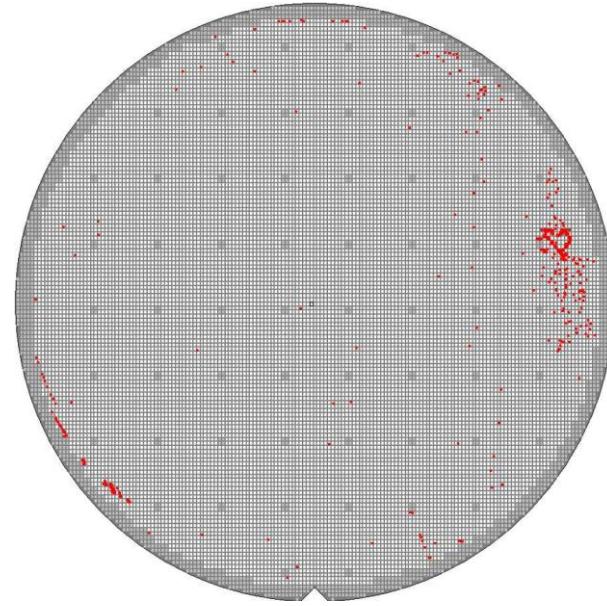
Wafer Defect Map



Data Augmentation is often key..

Collaboration with  life.augmented

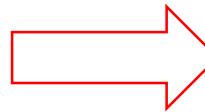
Train a deep learning model to identify defective patterns



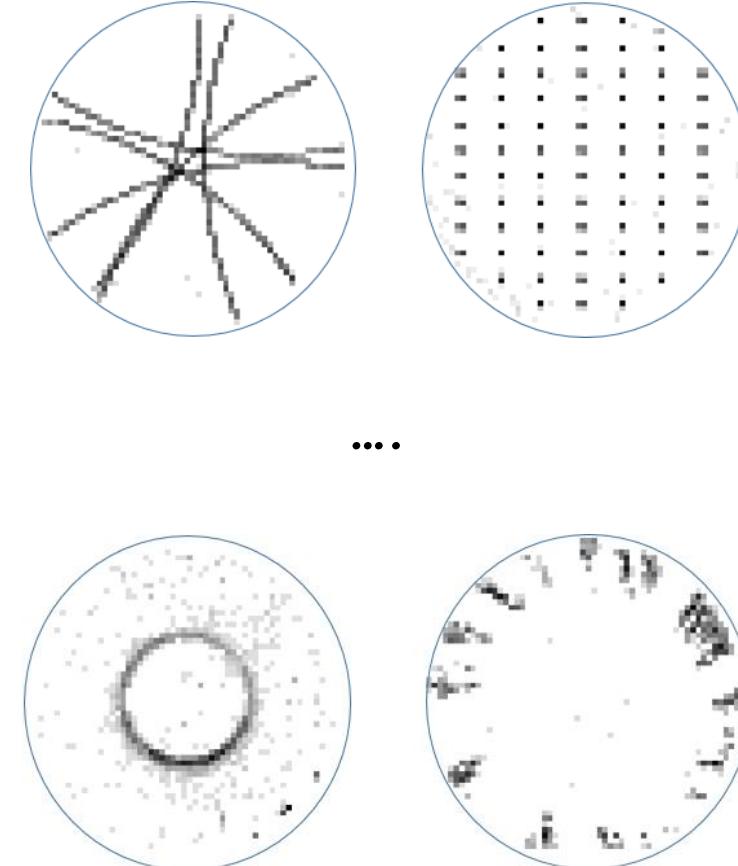
Wafer Defect Map

A_y

Deep Learning



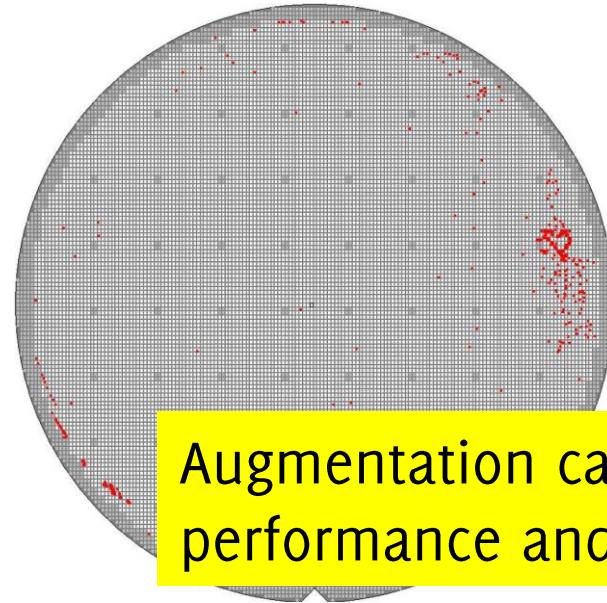
Defect Patterns



Data Augmentation is often key..

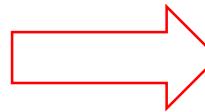
Collaboration with  life.augmented

Train a deep learning model to identify defective patterns



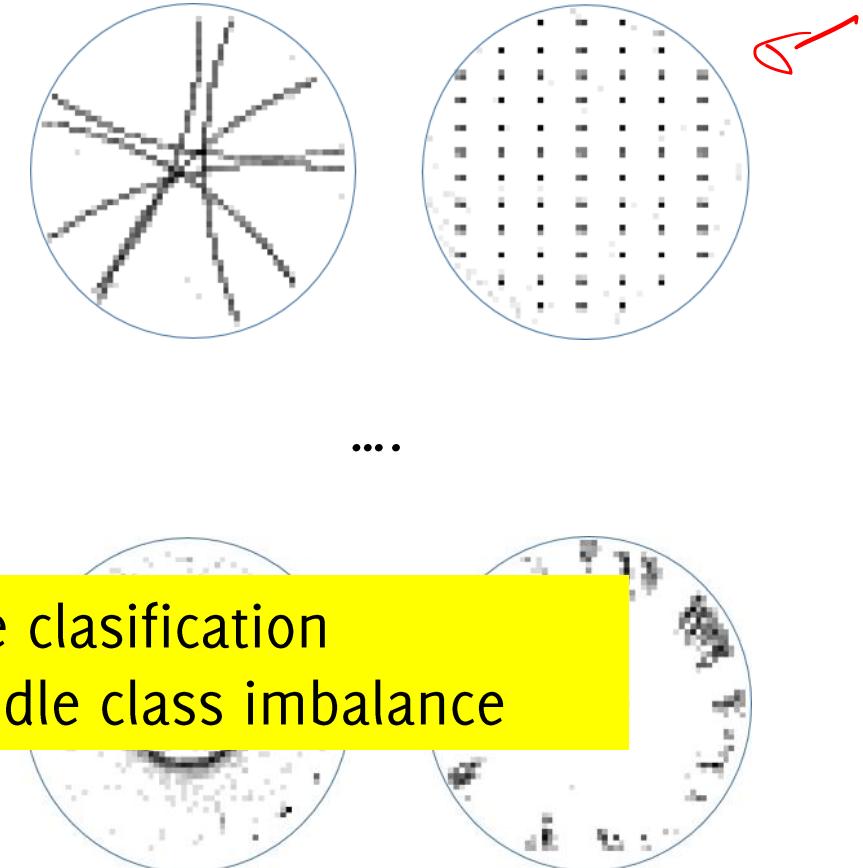
Wafer Defect Map

Deep Learning



Augmentation can greatly improve classification performance and successfully handle class imbalance

Defect Patterns



Test Time Augmentation (TTA) or Self-ensembling

Even if the CNN is trained using augmentation, it won't achieve perfect invariance w.r.t. considered transformations

Test time augmentation (TTA): augmentation can be also performed at test time to improve prediction accuracy.

- Perform random augmentation of each test image I
- Average the predictions of each augmented image
- Take the average vector of posterior for defining the final guess

Test Time Augmentation is particularly useful for test images where the model is pretty unsure.

Test Time Augmentation (TTA)

$$\begin{aligned} I \rightarrow \text{CNN}(I) &= \hat{y} \\ \hookrightarrow \text{CNN}(A(t)) &= \hat{y} \end{aligned}$$

Even if the CNN is trained using augmentation, it won't achieve perfect invariance w.r.t. considered transformations

Test time augmentation (TTA): augmentation can be also performed at test time to improve prediction accuracy.

- Perform random augmentation of each test image I
- Average the predictions of each augmented image
- Take the average vector of posterior for defining the final guess

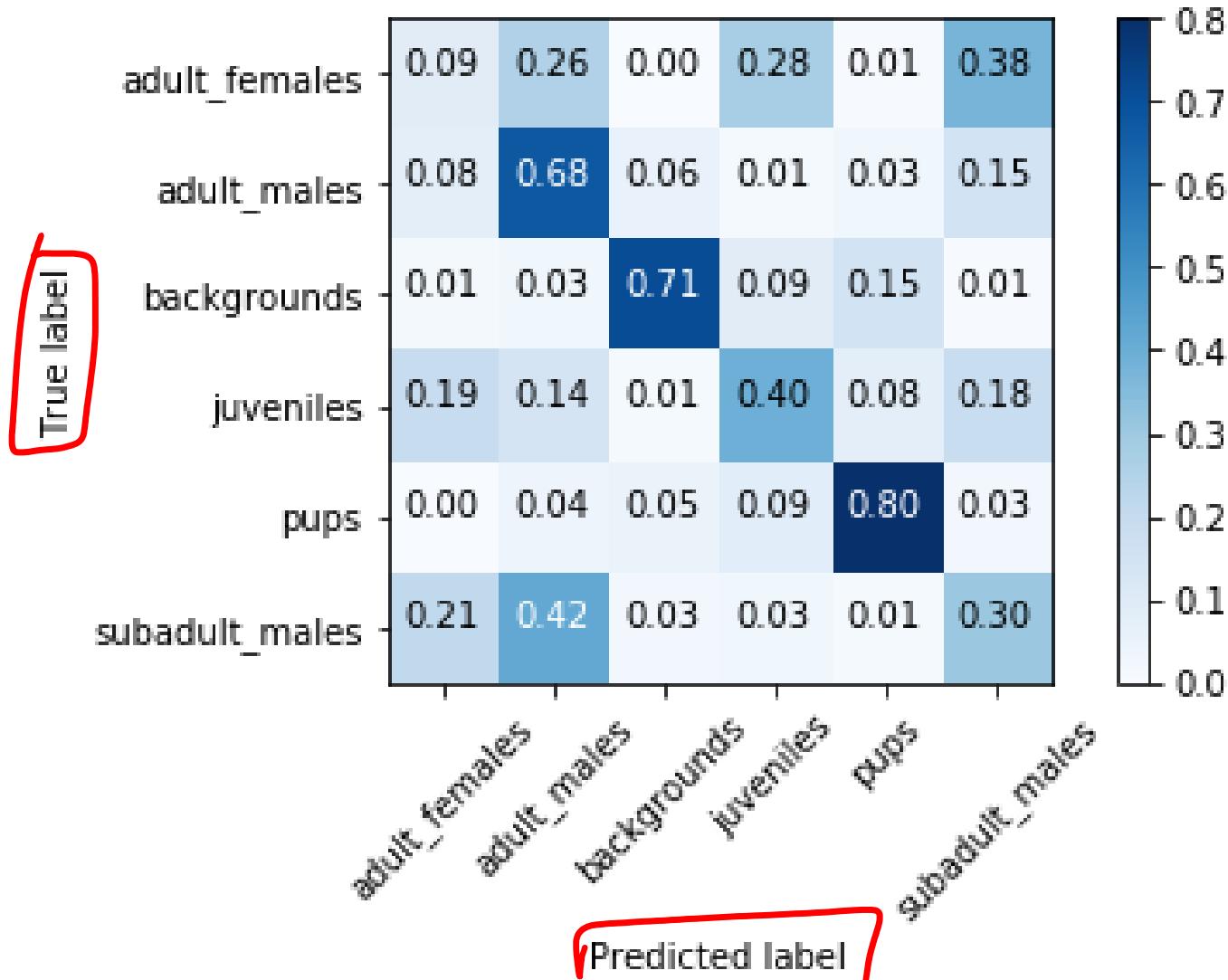
Test Time Augmentation is particularly useful for test images where the model is pretty unsure.

$$\begin{aligned} I \rightarrow \text{CNN}(I) &= \hat{y} \\ I \rightarrow \boxed{\text{Aug.}} \leftarrow \{A_i(I)\} & \quad \text{CNN}(A_i(I)) = \hat{y}_i \\ \text{Majority vote } \{y_i\} & \end{aligned}$$

A bit more of background

Performance measures
and an overview of successful architectures

Confusion Matrix

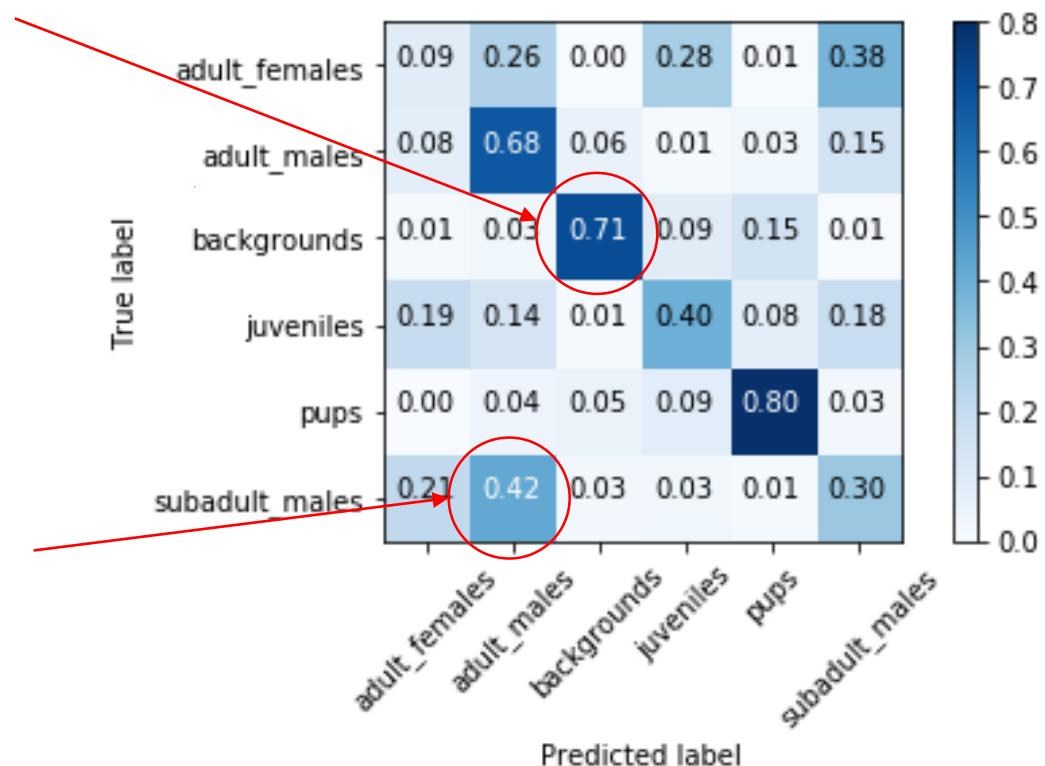


Confusion Matrix

The element $C(i,j)$ i.e. at the i -th row and j -th column corresponds to the percentage of elements belonging to class i classified as elements of class j

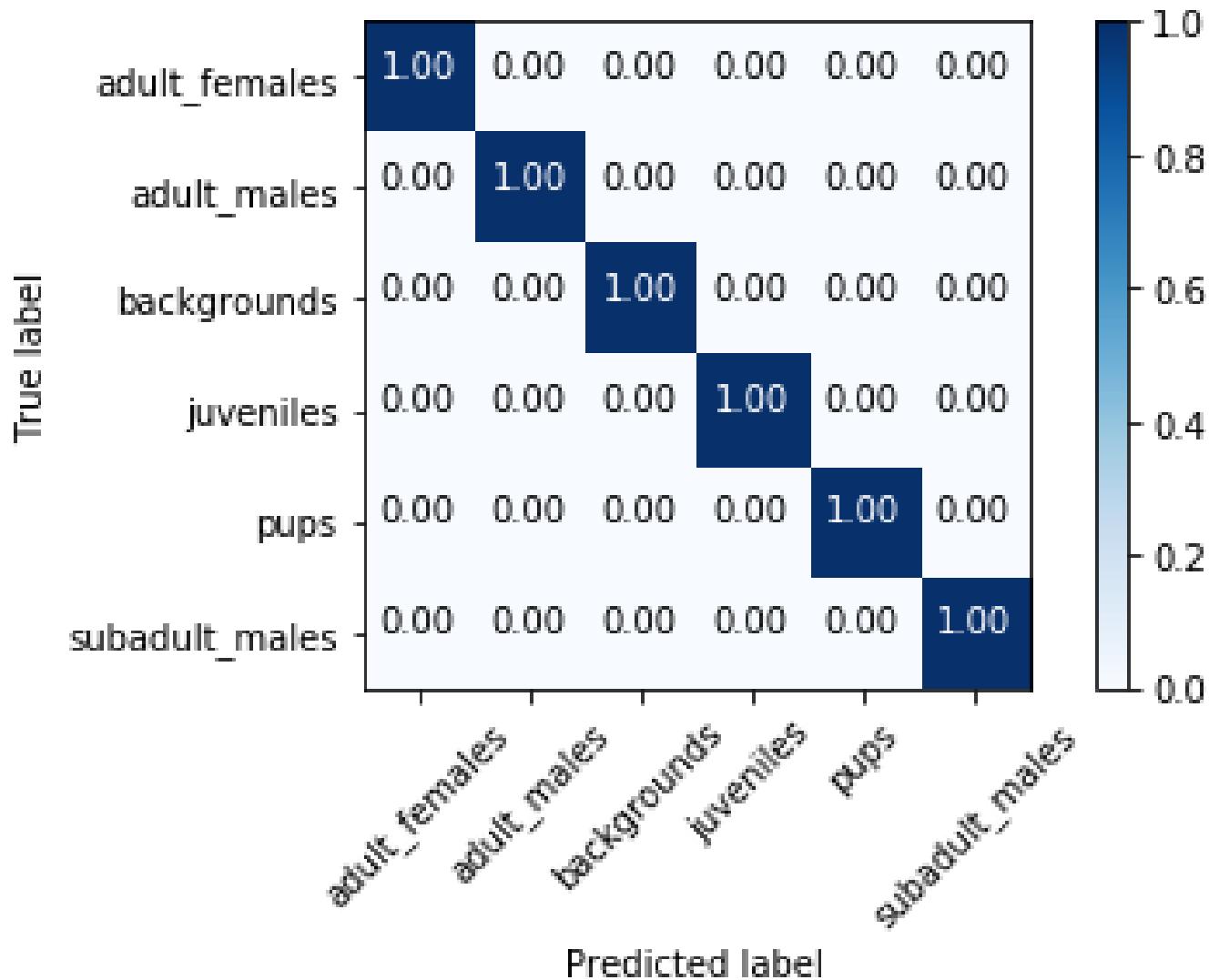
71% of background patches have been correctly classified as background

42% of sub-adult males patches have been wrongly classified as adult-males



... so, the ideal confusion matrix

which rarely happens



Two-Class Classification

Classification performance in case of binary classifiers can be also measured in terms of the ROC (receiver operating characteristic) curve, which does not depend on the threshold you set for each class

This is useful in case you plan to modify this and not use 0.5

The ideal detector would achieve:

- $FPR = 0\%$,
- $TPR = 100\%$

Thus, the closer to $(0,1)$ the better

The largest the **Area Under the Curve** (AUC), the better

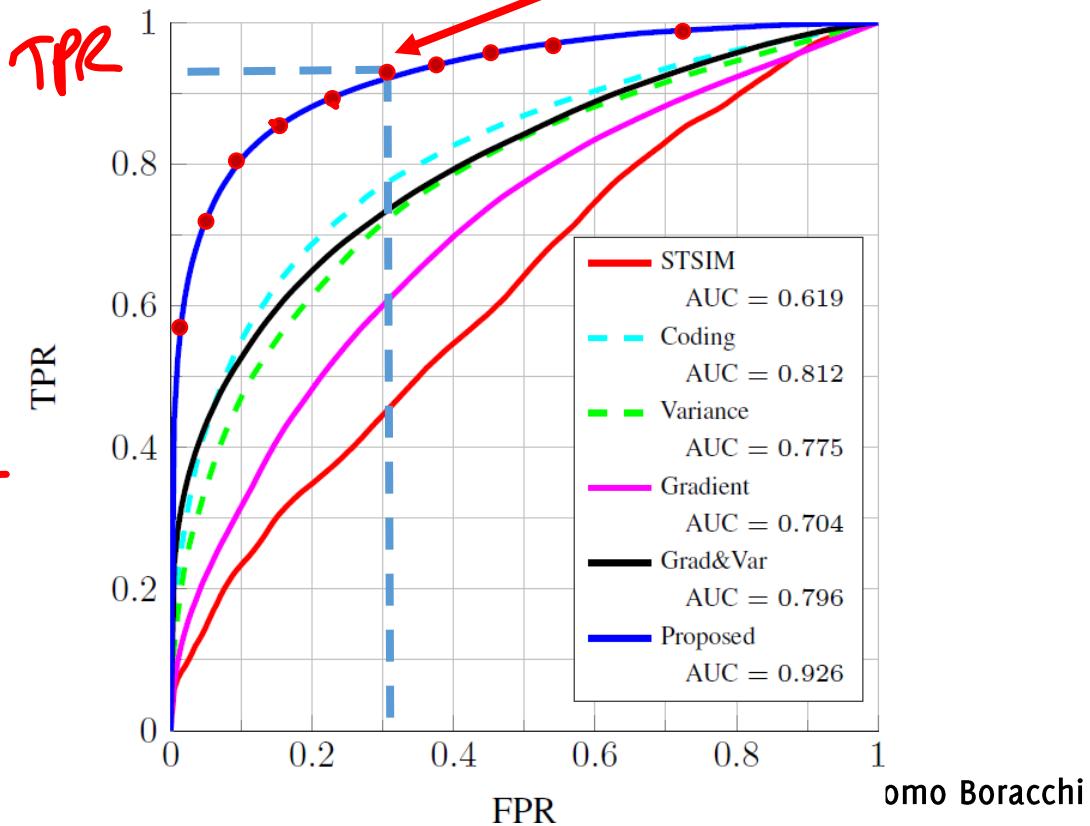
The optimal parameter is the one yielding the point closest to $(0,1)$

A	85	15
B	10	90
A	B	

A	60	40
B	2	98
A	B	

$$D(I) = s$$
$$CNN(I) = [P, 1-P]; P > 0.5 \quad P > 0.7$$

(FPR, TPR) for a specific parameter



CNN for Quality Inspection

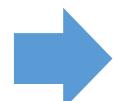
In collaboration
with



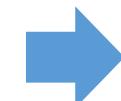
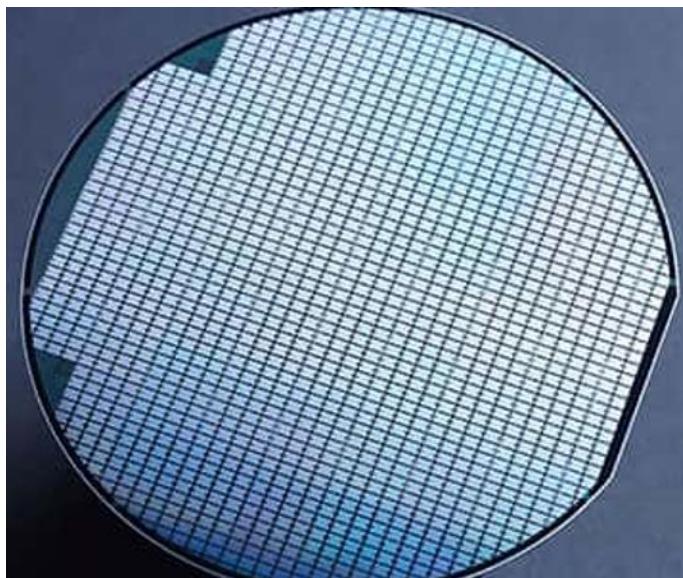
Giacomo Boracchi

Scenario

Chip Manufacturer



Silicon Wafer



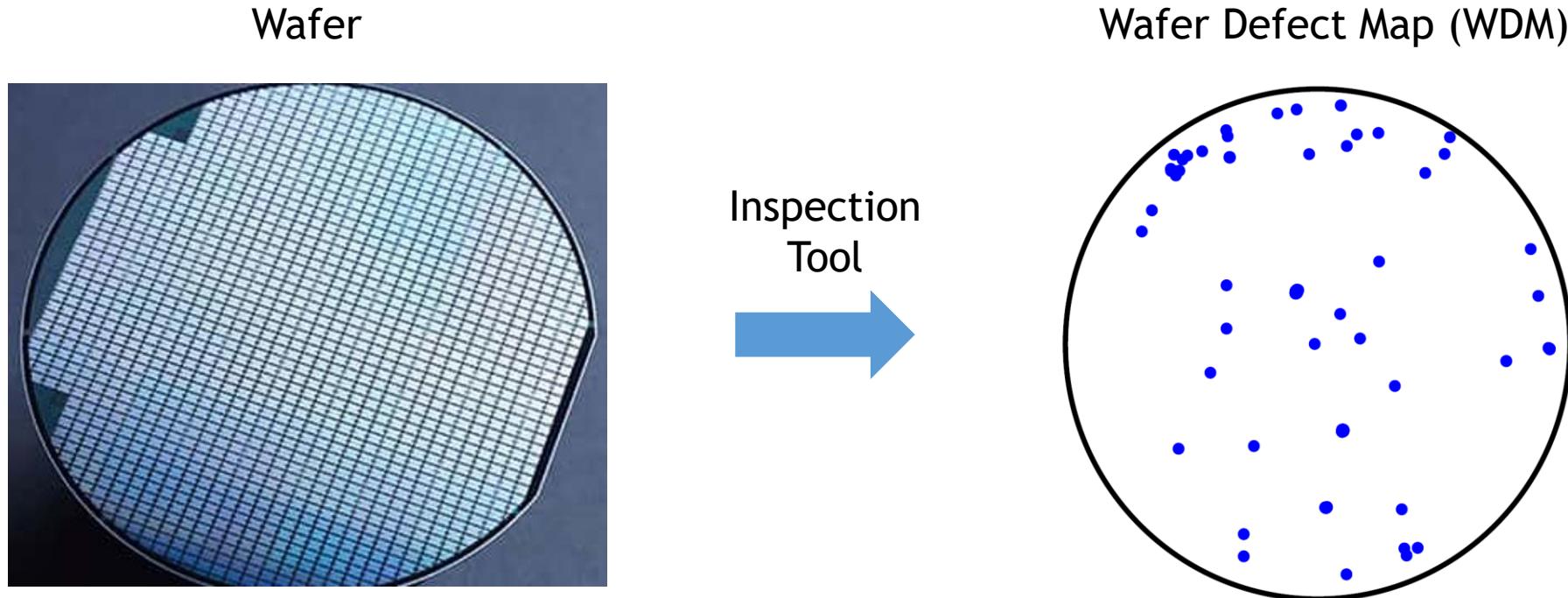
Chips / Memories / Sensors
are everywhere



In collaboration
with

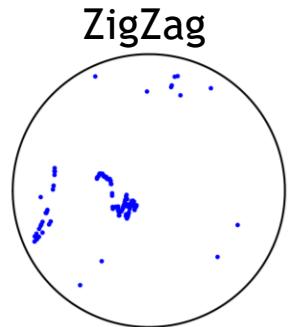
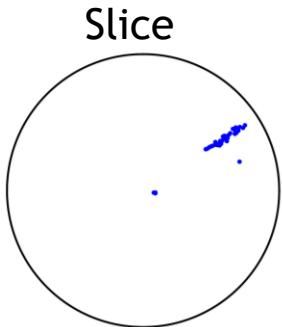
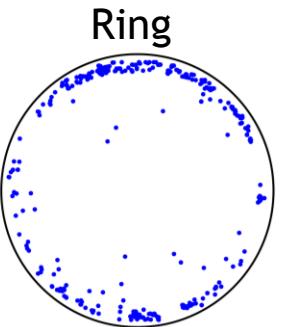
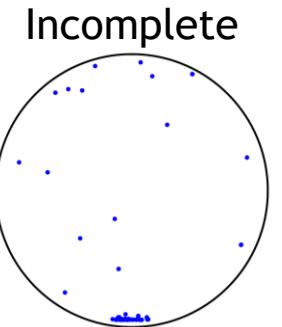
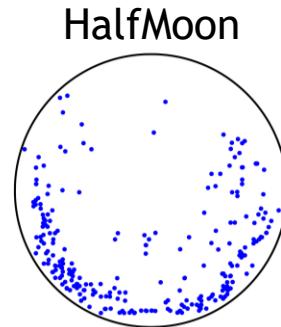
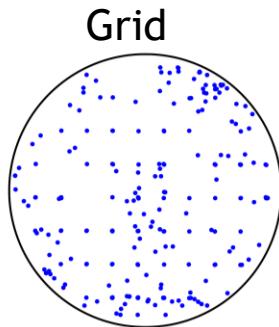
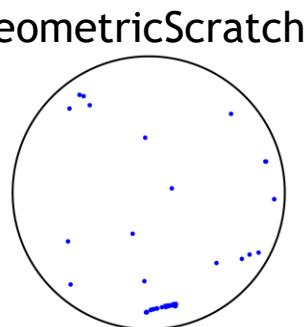
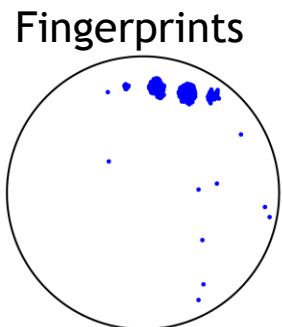
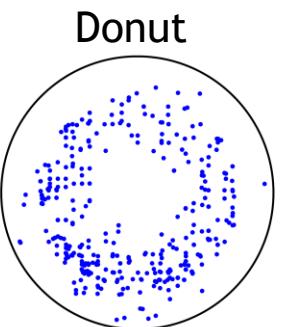
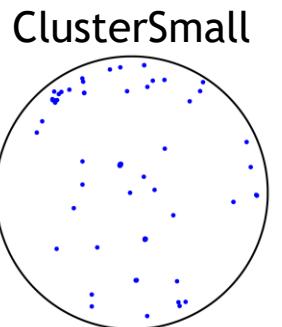
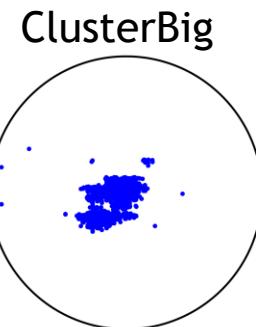
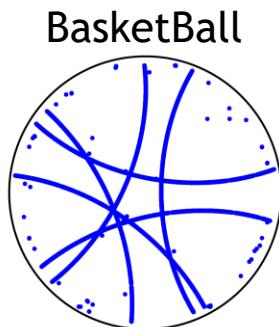


Monitoring Silicon Wafer Manufacturing Process



Classess of WDM Patterns

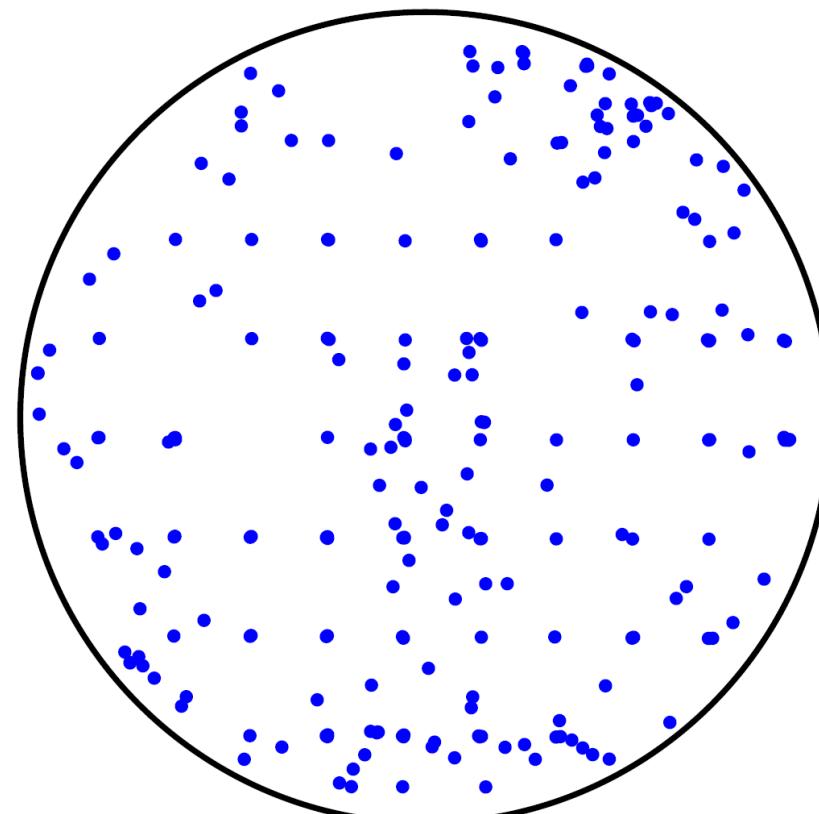
Specific patterns in WDMs might indicate problems in the production



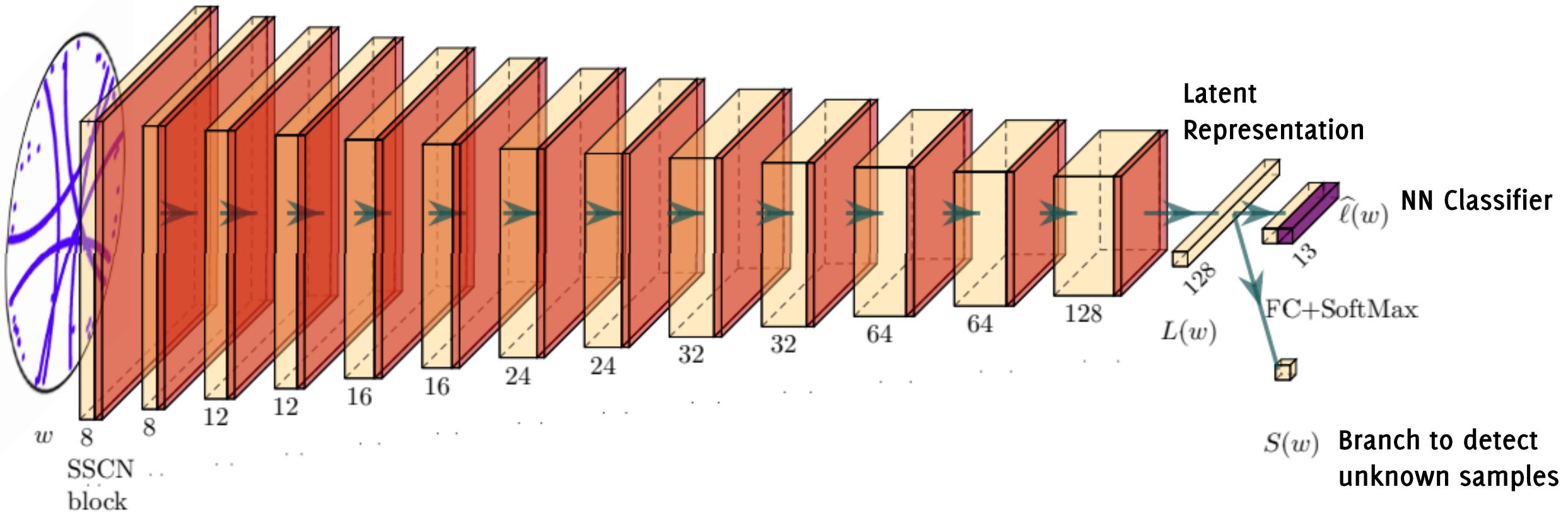
Classify WDM to raise prompt alerts

Challenges

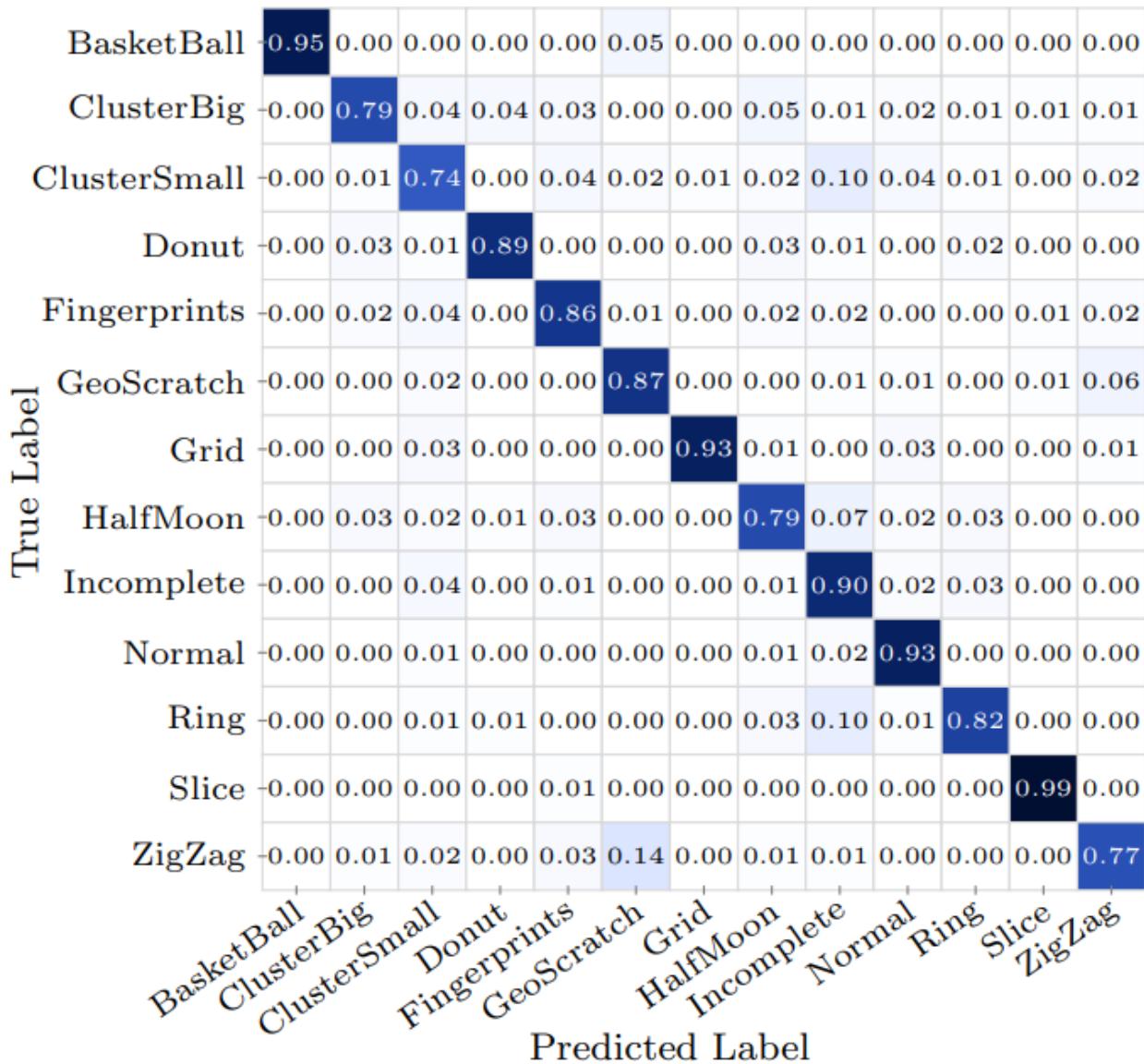
- Huge resolution: a WDM as a grayscale would require ~ 3 GB to store w in memory
- Very Limited Supervision
- Some defects occur very rarely



Our CNN



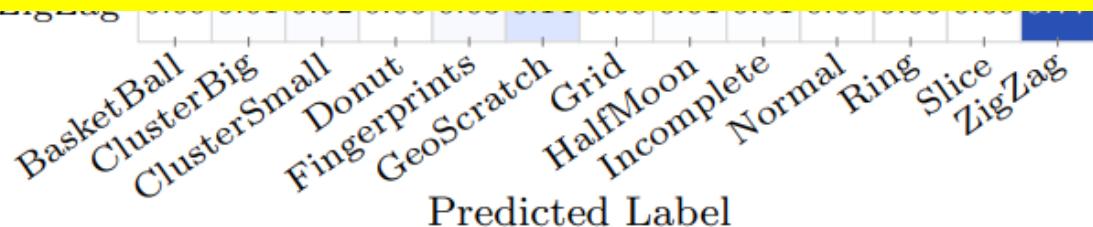
Results



Results

BasketBall	-0.95	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ClusterBig	-0.00	0.79	0.04	0.04	0.03	0.00	0.00	0.05	0.01	0.02	0.01	0.01	0.01	0.01
ClusterSmall	-0.00	0.01	0.74	0.00	0.04	0.02	0.01	0.02	0.10	0.04	0.01	0.00	0.02	0.02
Donut	-0.00	0.03	0.01	0.89	0.00	0.00	0.00	0.03	0.01	0.00	0.02	0.00	0.00	0.00
Fingerprints	-0.00	0.02	0.04	0.00	0.86	0.01	0.00	0.02	0.02	0.00	0.00	0.01	0.02	0.02
GeoScratch	-0.00	0.00	0.02	0.00	0.00	0.87	0.00	0.00	0.01	0.01	0.00	0.01	0.01	0.06
Grid	-0.00	0.00	0.03	0.00	0.00	0.00	0.93	0.01	0.00	0.03	0.00	0.00	0.01	0.01
HalfMoon	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00
Incomplete	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00
Normal	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00
Ring	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00
Slice	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00
ZigZag	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00

Our system is currently monitoring the largest production line in Agrate and most backend sites



A few popular architectures

AlexNet
VGG

Outline

Lecture inspired to:

Notes accompanying the Stanford CS class [CS231n: Convolutional Neural Networks for Visual Recognition](#) <http://cs231n.github.io/>

... and papers cited in here!

Gradient-Based Learning Applied to Document Recognition

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner

Abstract—

Multilayer Neural Networks trained with the backpropagation algorithm constitute the best example of a successful Gradient-Based Learning technique. Given an appropriate network architecture, Gradient-Based Learning algorithms can be used to synthesize a complex decision surface that can classify high-dimensional patterns such as handwritten characters, with minimal preprocessing. This paper reviews various methods applied to handwritten character recognition and compares them on a standard handwritten digit recognition task. Convolutional Neural Networks, that are specifically designed to deal with the variability of 2D shapes, are shown to outperform all other techniques.

I. INTRODUCTION

Over the last several years, machine learning techniques, particularly when applied to neural networks, have played an increasingly important role in the design of pattern recognition systems. In fact, it could be argued that the availability of learning techniques has been a crucial factor in the recent success of pattern recognition applications such as continuous speech recognition and handwriting recognition.

Fathers of the Deep Learning Revolution Receive ACM A.M. Turing Award

Bengio, Hinton and LeCun Ushered in Major Breakthroughs in Artificial Intelligence

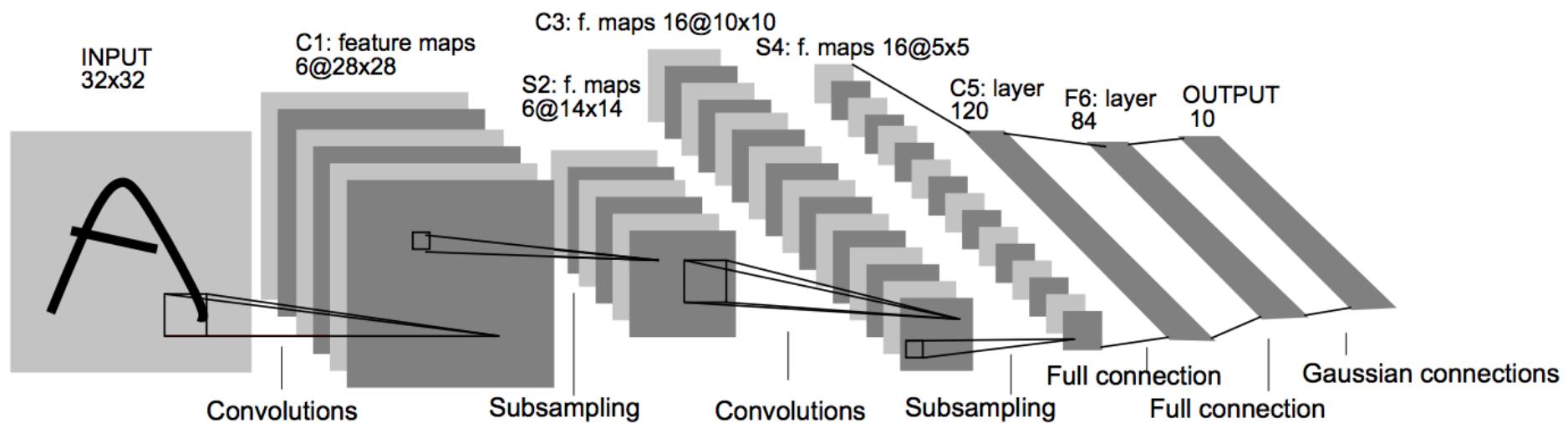
<https://awards.acm.org/about/2018-turing>

LeNet-5 (1998)

Yann Lecun's LeNet-5 model was developed in 1998 to identify handwritten digits for zip code recognition in the postal service.

This pioneering model largely introduced the convolutional neural network as we know it today

Parameters: 60.000



LeNet-5 (1998)

The idea of using a sequence of 3 layers: **convolution**, **pooling**, **non-linearity**.

Pixels in an image are highly correlated: there is no point to separately feed each pixel value through a classification network: add learnable convolutional layers

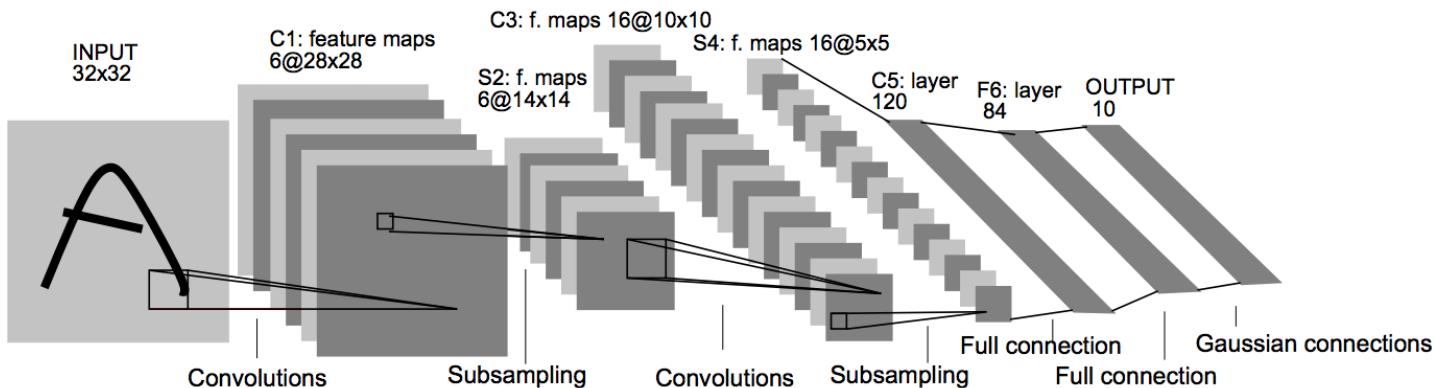
Convolution as a way to extract spatial features and to have sparse connections (thus reducing the number of parameters).

Subsample using spatial average of maps (averaging pooling)

Non-linearities: tanh or sigmoids

MLP as classifier for the final layer

60K



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Fathers of the Deep Learning Revolution Receive ACM A.M. Turing Award

Bengio, Hinton and LeCun Ushered in Major Breakthroughs in Artificial Intelligence

AlexNet (2012)

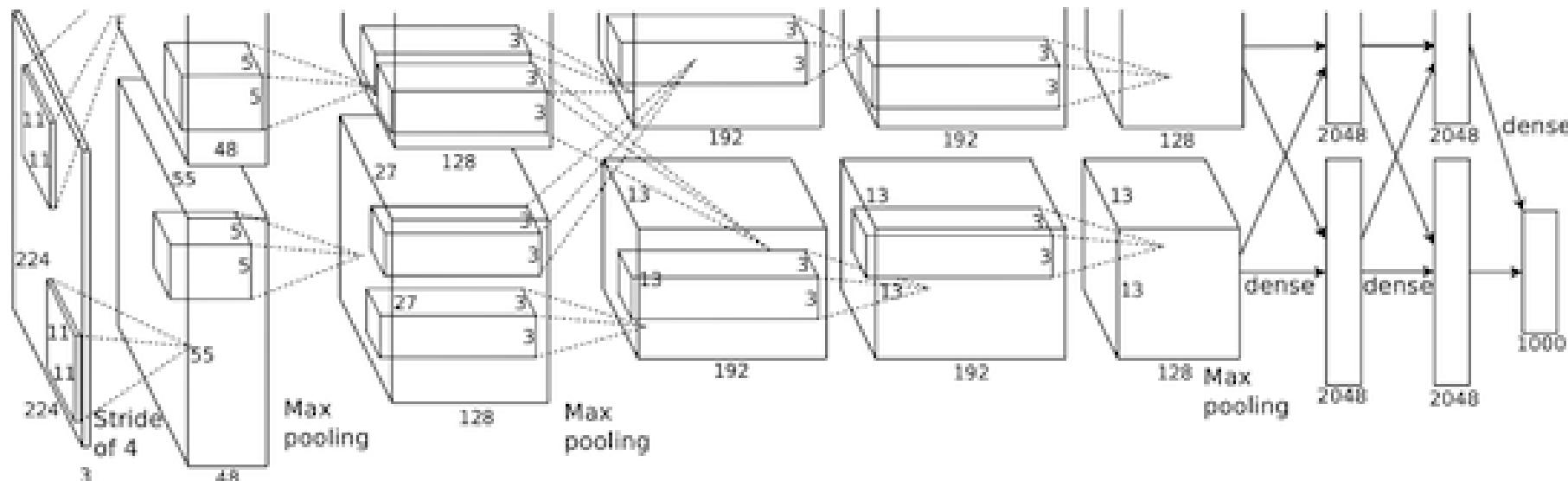
Developed by Alex Krizhevsky et al. in 2012 and won Imagenet competition

Architecture is quite similar to LeNet-5:

- 5 convolutional layers (rather large filters, 11×11 , 5×5),
- **3 MLP**

Input size $224 \times 224 \times 3$ (the paper says $227 \times 227 \times 3$)

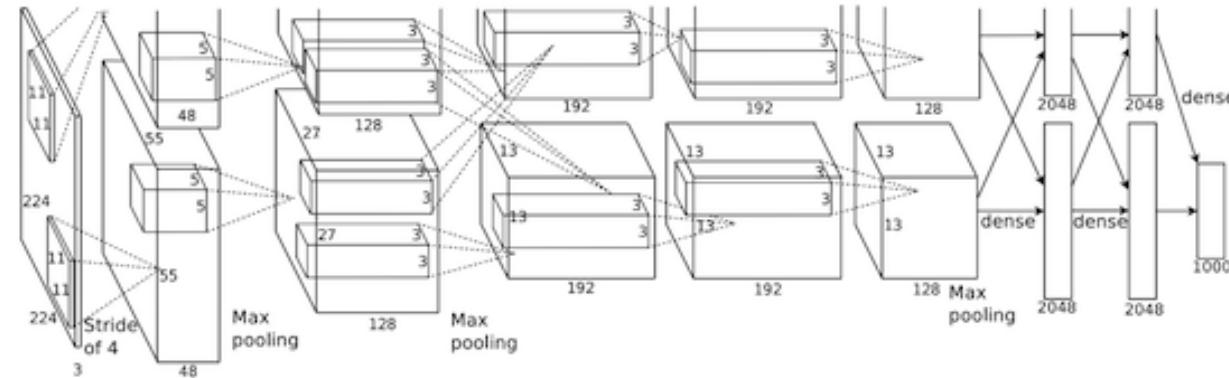
Parameters: 60 million [Conv: 3.7million (6%), FC: 58.6 million (94%)]



AlexNet (2012)

To counteract overfitting, they introduce:

- RELU (also faster than tanh)
- Dropout (0.5), weight decay and norm layers (not used anymore)
- Maxpooling



The first conv layer has 96 11x 11 filters, stride 4.

The output are two volumes of 55 x 55 x 48 separated over two GTX 580 GPUs (1.5GB each GPU, 90 epochs, 5/6 days to train).

Most connections are among feature maps of the same GPU, which will be mixed at the last layer.

Won the ImageNet challenge in 2012

At the end they also trained an ensemble of 7 models to drop error: 18.2%→15.4%

VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION

Karen Simonyan* & **Andrew Zisserman⁺**

Visual Geometry Group, Department of Engineering Science, University of Oxford
`{karen,az}@robots.ox.ac.uk`

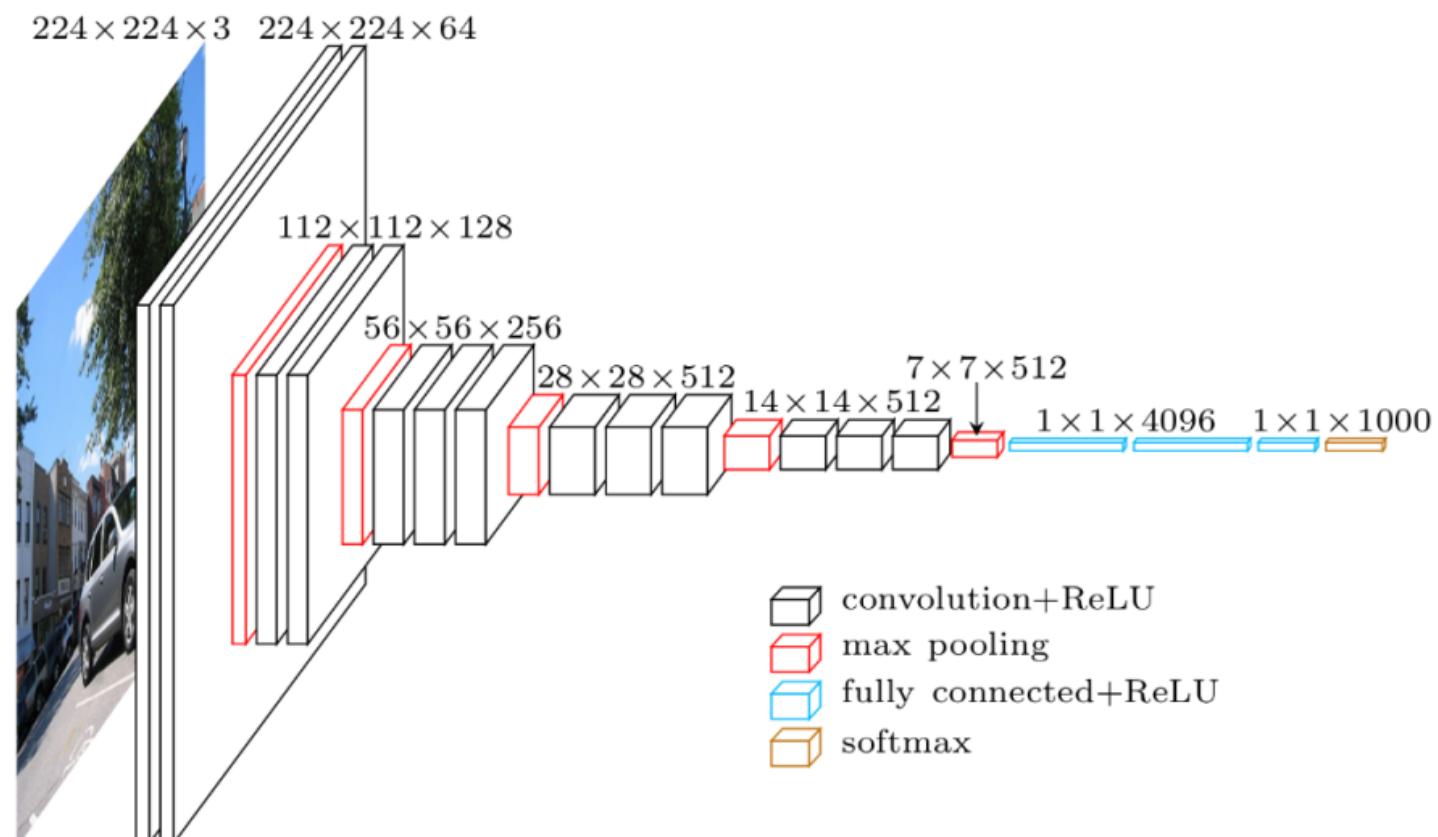
ABSTRACT

In this work we investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. Our main contribution is a thorough evaluation of networks of increasing depth using an architecture with very small (3×3) convolution filters, which shows that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 16–19 weight layers. These findings were the basis of our ImageNet Challenge 2014 submission, where our team secured the first and the second places in the localisation and classification tracks respectively. We also show that our representations generalise well to other datasets, where they achieve state-of-the-art results. We have made our two best-performing ConvNet models publicly available to facilitate further research on the use of deep visual representations in computer vision.

VGG16 (2014)

The VGG16, introduced in 2014 is a deeper variant of the AlexNet convolutional structure. Smaller filters are used and the network is deeper

Parameters: 138 million [Conv: 11%, FC: 89%]



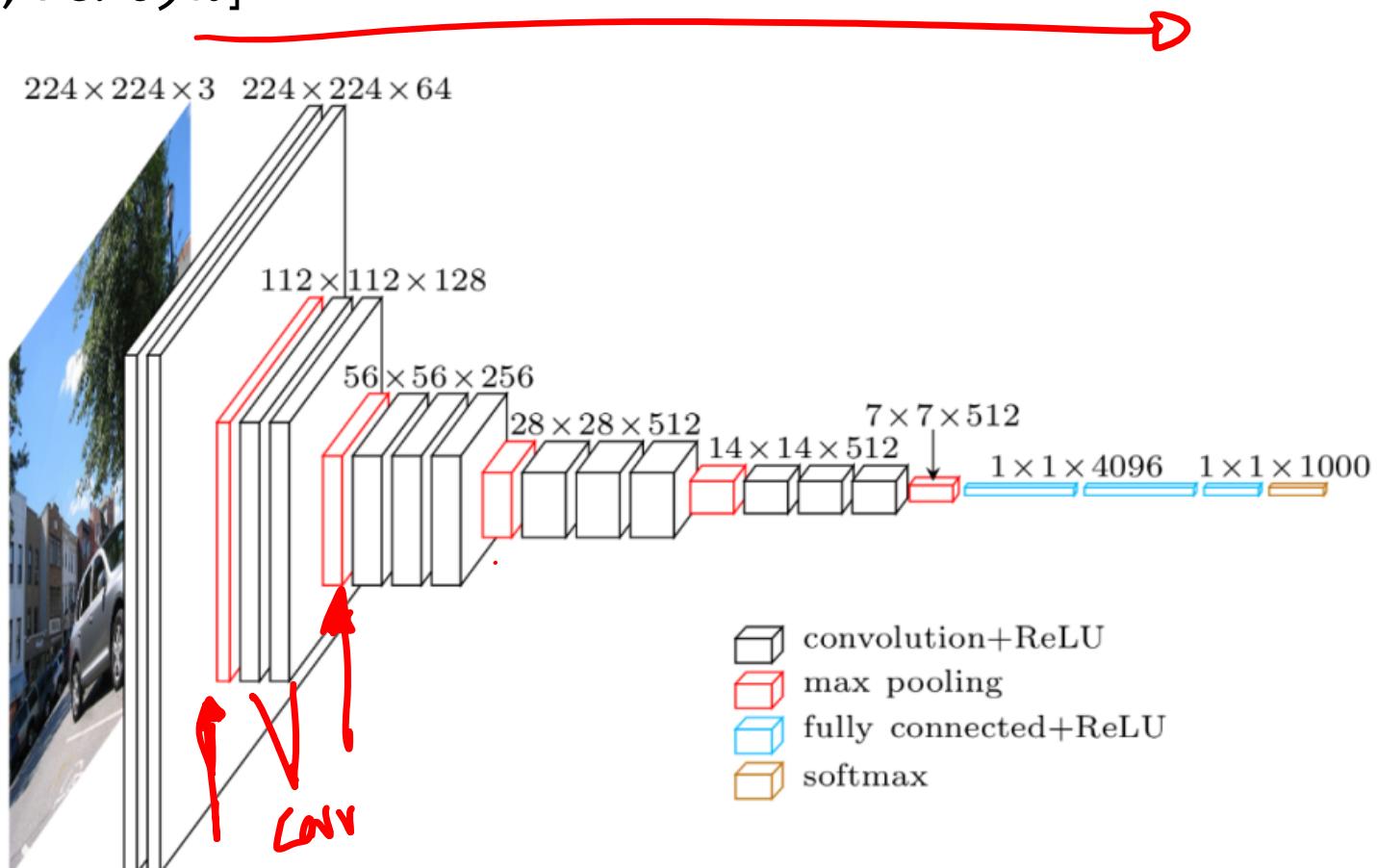
VGG16 (2014)

The VGG16, introduced in 2014 is a deeper variant of the AlexNet convolutional structure. Smaller filters are used and the network is deeper

Parameters: 138 million [Conv: 11%, FC: 89%]

These architecture **won the first place places (localization) and the second place (classification) tracks in ImageNet Challenge 2014**

Input size $224 \times 224 \times 3$



VGG16 (2014): Smaller Filter, Deeper Network

The paper actually present a thorough study on the role of network depth.

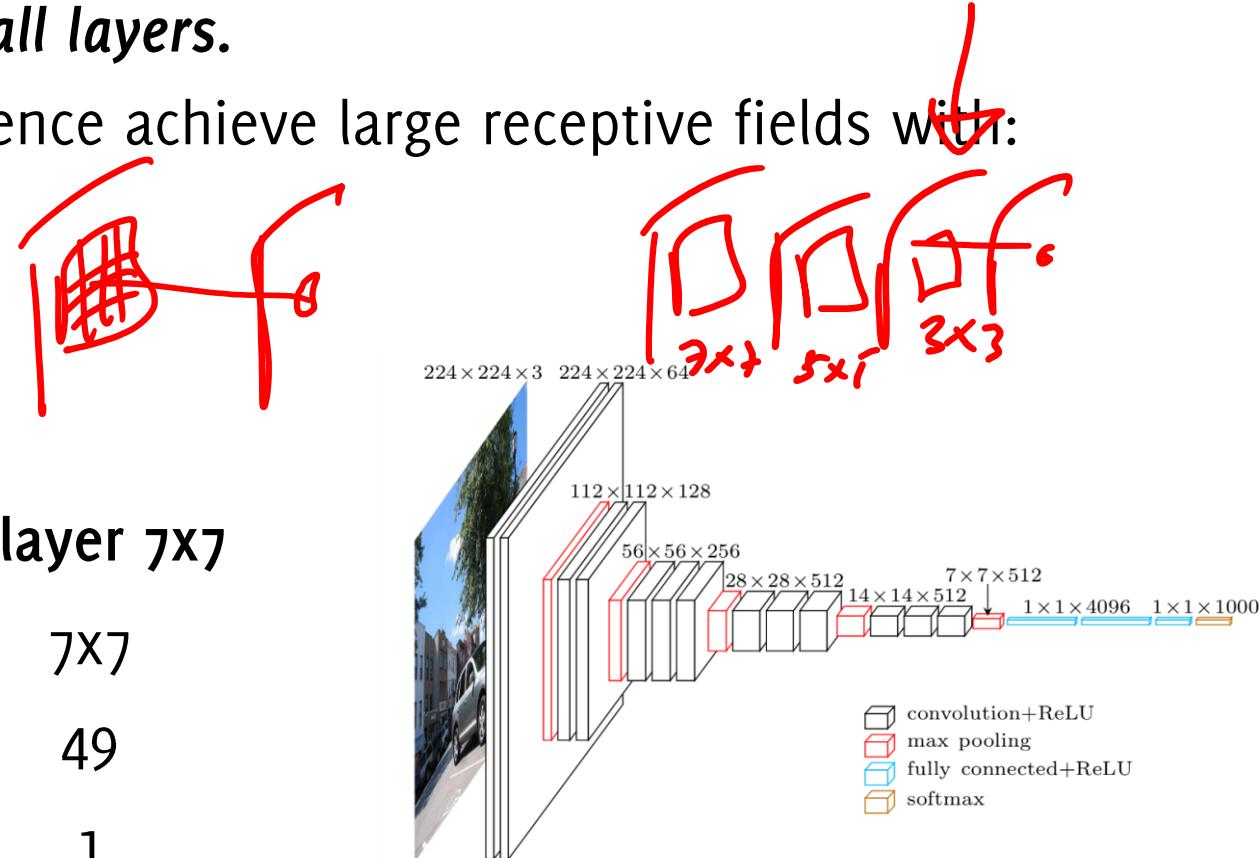
[...] Fix other parameters of the architecture, and steadily increase the depth of the network by adding more convolutional layers, which is feasible due to the use of very small (3×3) convolution filters in all layers.

Idea: Multiple 3×3 convolution in a sequence achieve large receptive fields with:

- less parameters
- more nonlinearities

than larger filters in a single layer

	3 layers 3×3	1 layer 7×7
Receptive field	7×7	7×7
Nr of parameters	$3 \times 3 \times 3 = 27$	49
Nr of nonlinearities	3	1



VGG16

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 224, 224, 3)	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512) [...]	2359808

Layer (type)	Output Shape	Param #
	[...]	
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312
predictions (Dense)	(None, 1000)	4097000
<hr/>		
Total params: 138,357,544		
Trainable params: 138,357,544		
Non-trainable params: 0		

VGG16

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 224, 224, 3)	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
	[...]	

Layer (type)	Output Shape	Param #
	[...]	
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312
predictions (Dense)	(None, 1000)	4097000
<hr/>		
Total params: 138,357,544		
Trainable params: 138,357,544		
Non-trainable params: 0		

Many convolutional blocks without maxpooling

VGG16

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 224, 224, 3)	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512) [...]	2359808

Layer (type)	Output Shape	Param #
	[...]	
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312
predictions (Dense)	(None, 1000)	4097000
Total params: 138,357,544		
Trainable params: 138,357,544		
Non-trainable params: 0		

Most parameters in FC layers : 123,642,856

Giacomo Boracchi

THESE FEATURES ARE VERY GOOD!

VGG16

224 × 224 × 3

Layer (type)	Output Shape	Param #	Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 224, 224, 3)	0		[...]	
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792	block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928	block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0	block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block2_conv1 (Conv2D)				14, 14, 512)	2359808
block2_conv2 (Conv2D)				7, 7, 512)	0
block2_pool (MaxPooling2D)				25088)	0
block3_conv1 (Conv2D)				4096)	102764544
block3_conv2 (Conv2D)				4096)	16781312
block3_conv3 (Conv2D)				1000)	4097000
block3_pool (MaxPooling2D)					
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160	Trainable params: 138,357,544		
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808	Non-trainable params: 0		
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808			
	[...]				

High memory request, about 100MB per image to be stored in all the activation maps, only for the forward pass (for the backward it's about twice)

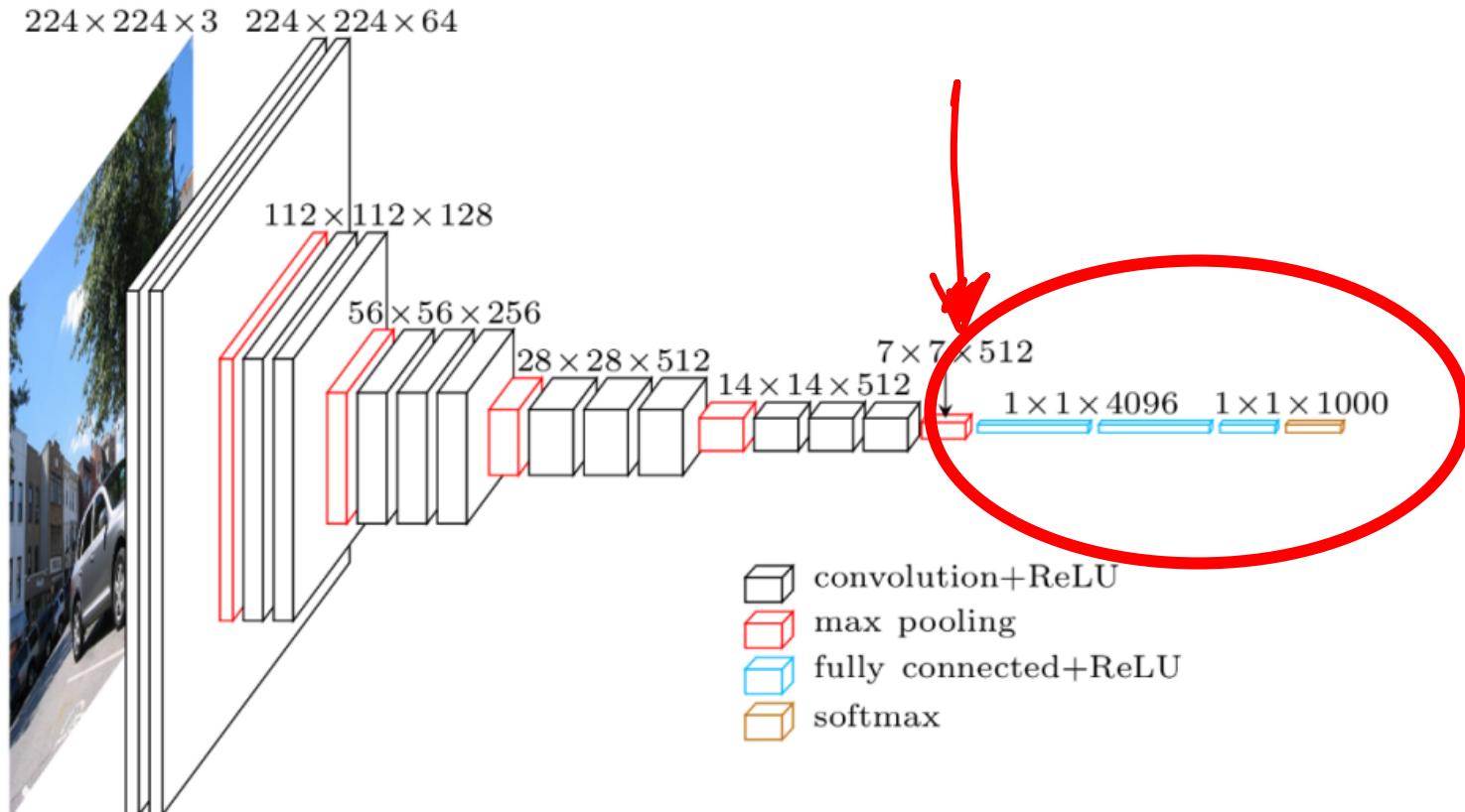
Popular Architectures to be continued...

Limited Amount of Data: Transfer Learning

Training a CNN with Limited Aumont of Data

Transfer Learning

How to use pre-trained models to solve different problems

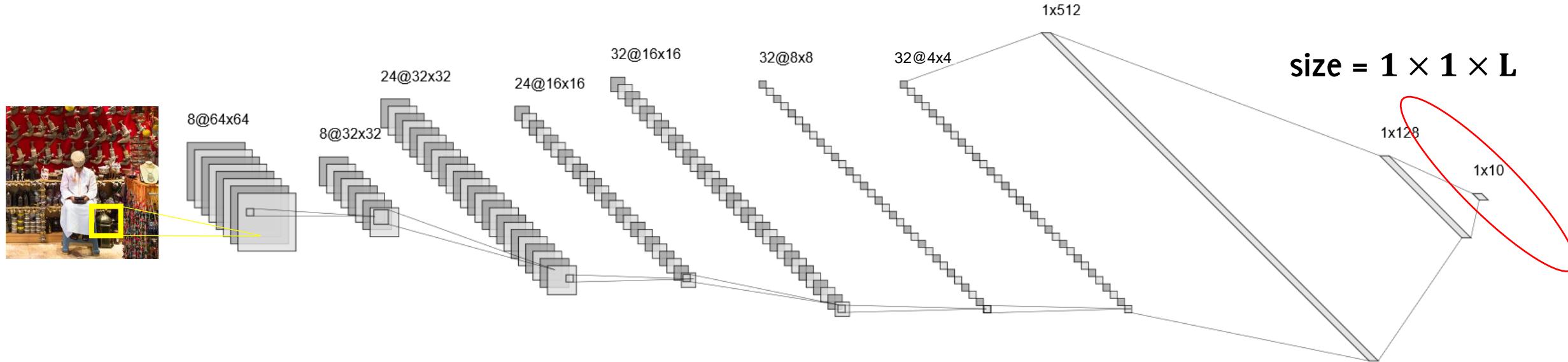


Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

Giacomo Boracchi

Convolutional Neural Networks (CNN)

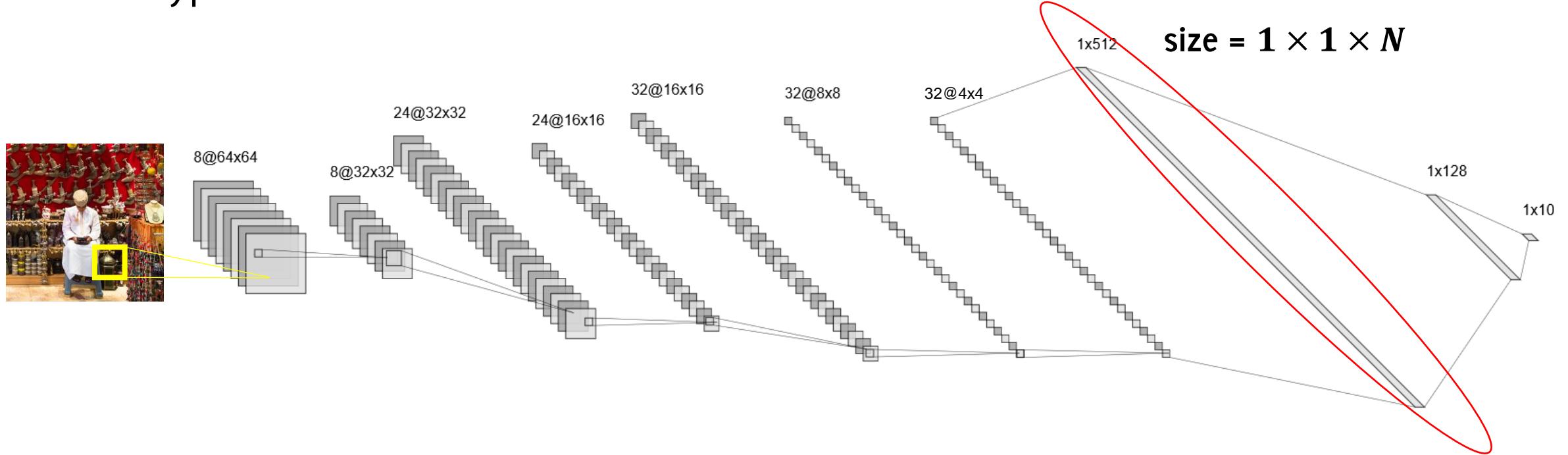
The typical architecture of a convolutional neural network



The output of the fully connected layer has the same size as the number of classes L , and each component provide a score for the input image to belong to a specific class

Convolutional Neural Networks (CNN)

The typical architecture of a convolutional neural network



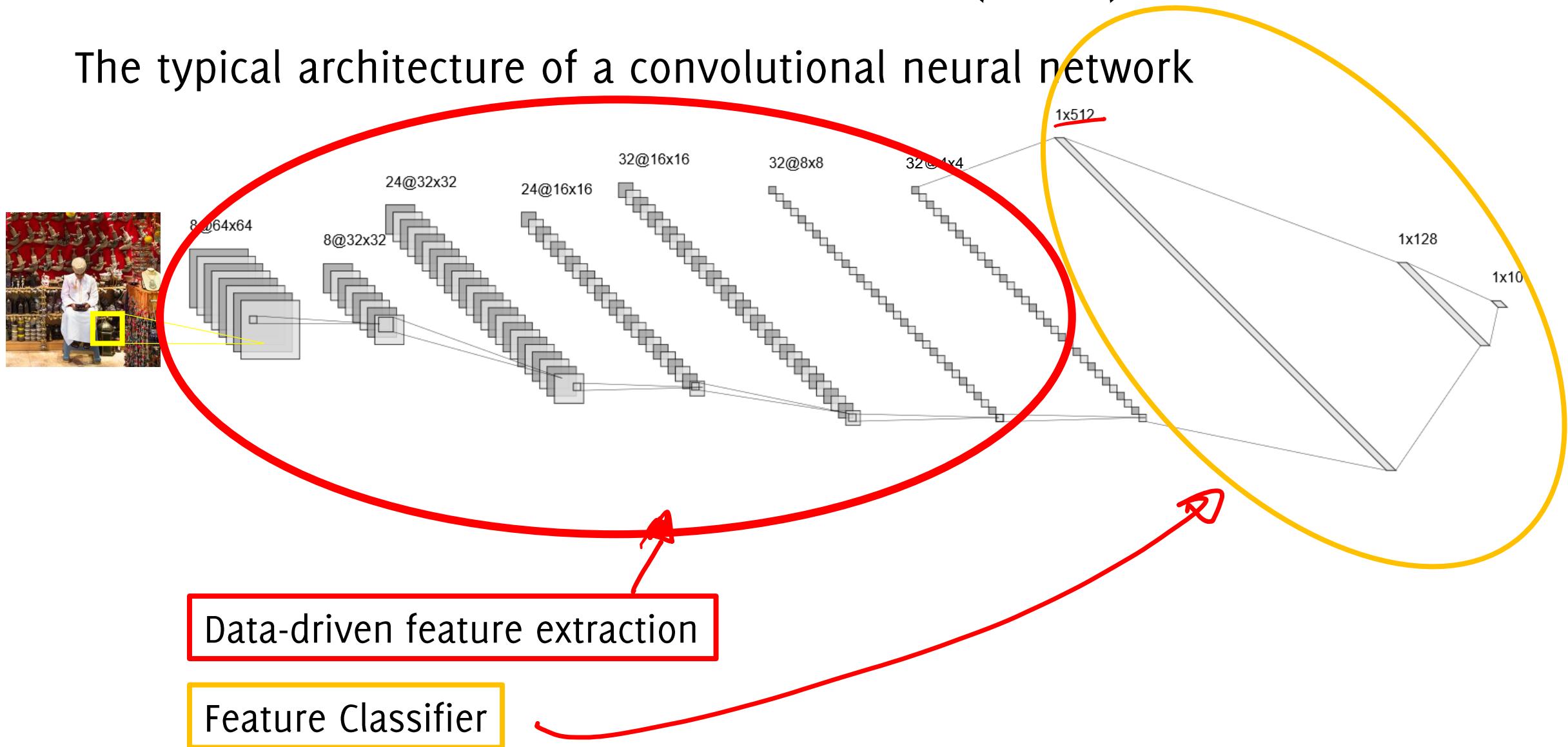
The input of the fully connected layer can be seen as a data-driven descriptor of the input image, i.e. a feature vector. These features are:

- defined to maximize classification performance
- Trained via backpropagation as the NN they are fed

↑
LATENT
REPRESENTATION

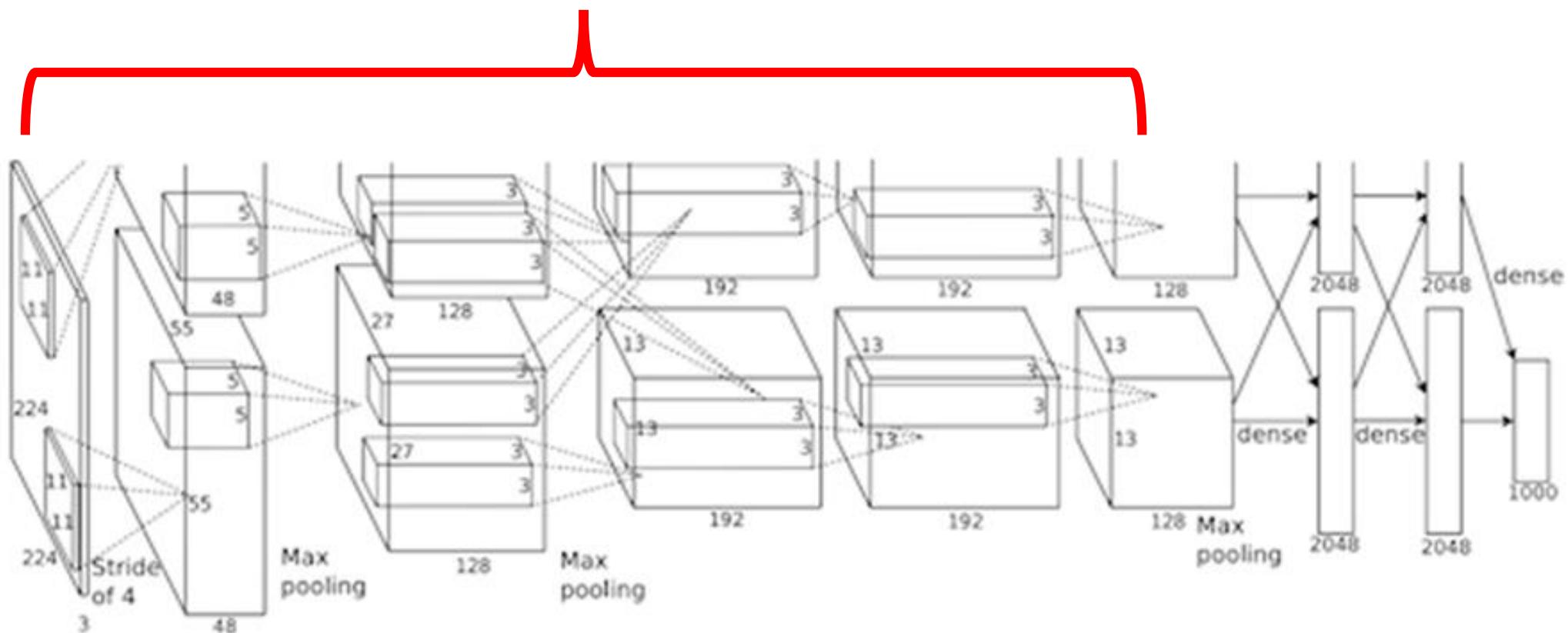
Convolutional Neural Networks (CNN)

The typical architecture of a convolutional neural network



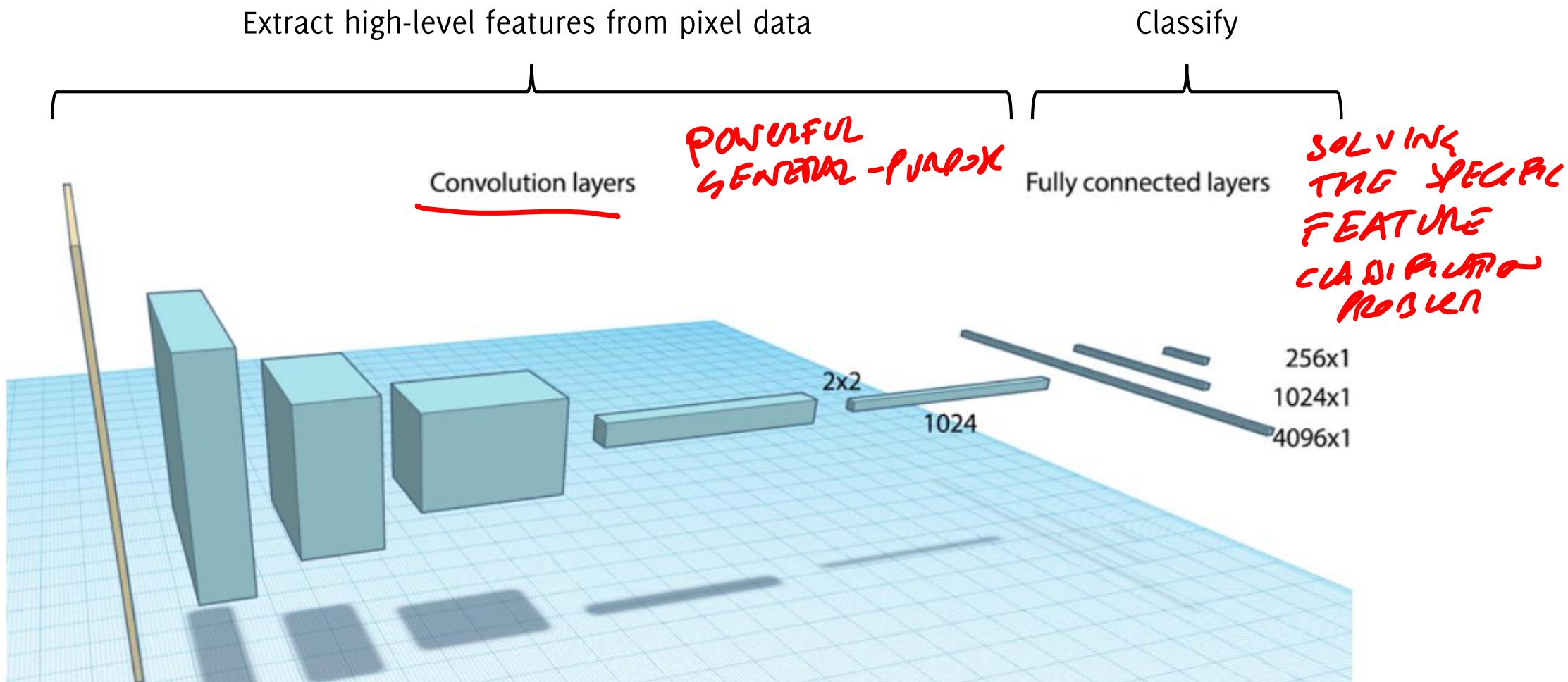
Deep architectures

Deep Networks are great at extracting
(learned) features from images



AlexNet architecture (May look weird because there are two different “streams”. This is because the training process was so computationally expensive that they had to split the training onto 2 GPUs)

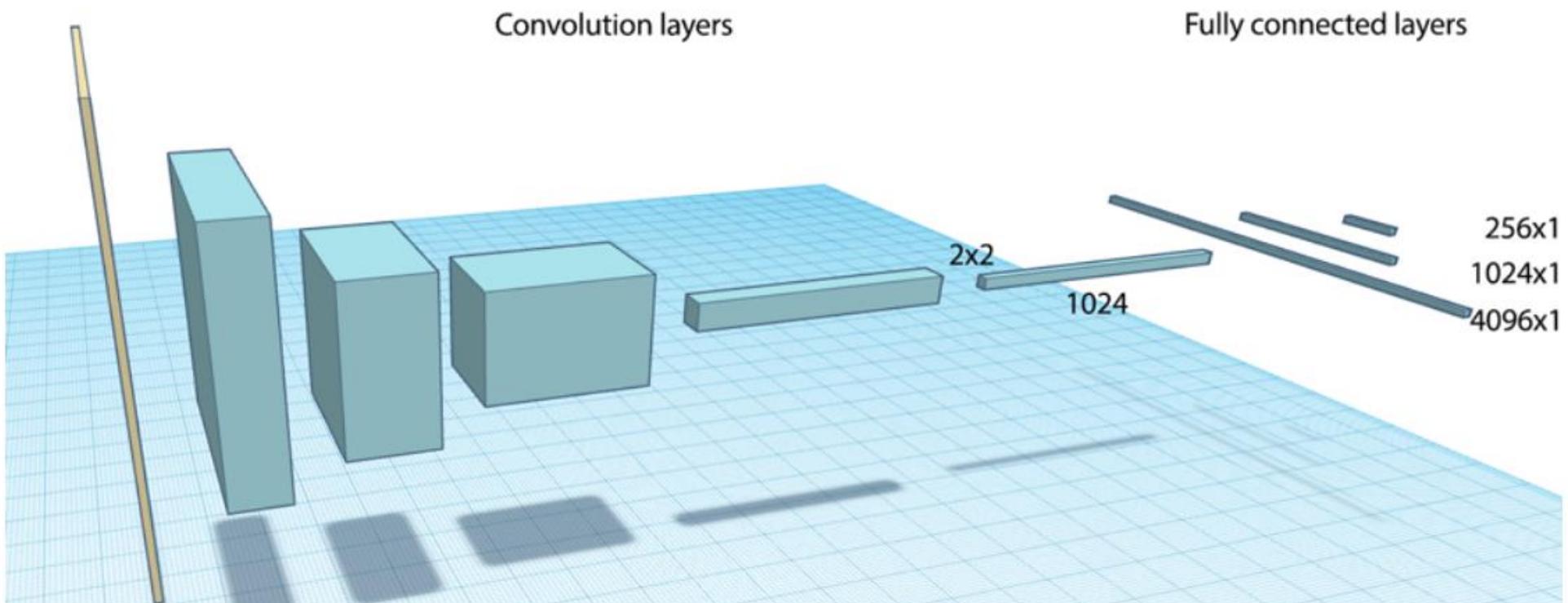
A typical architecture



A typical architecture

As we move to deeper layers:

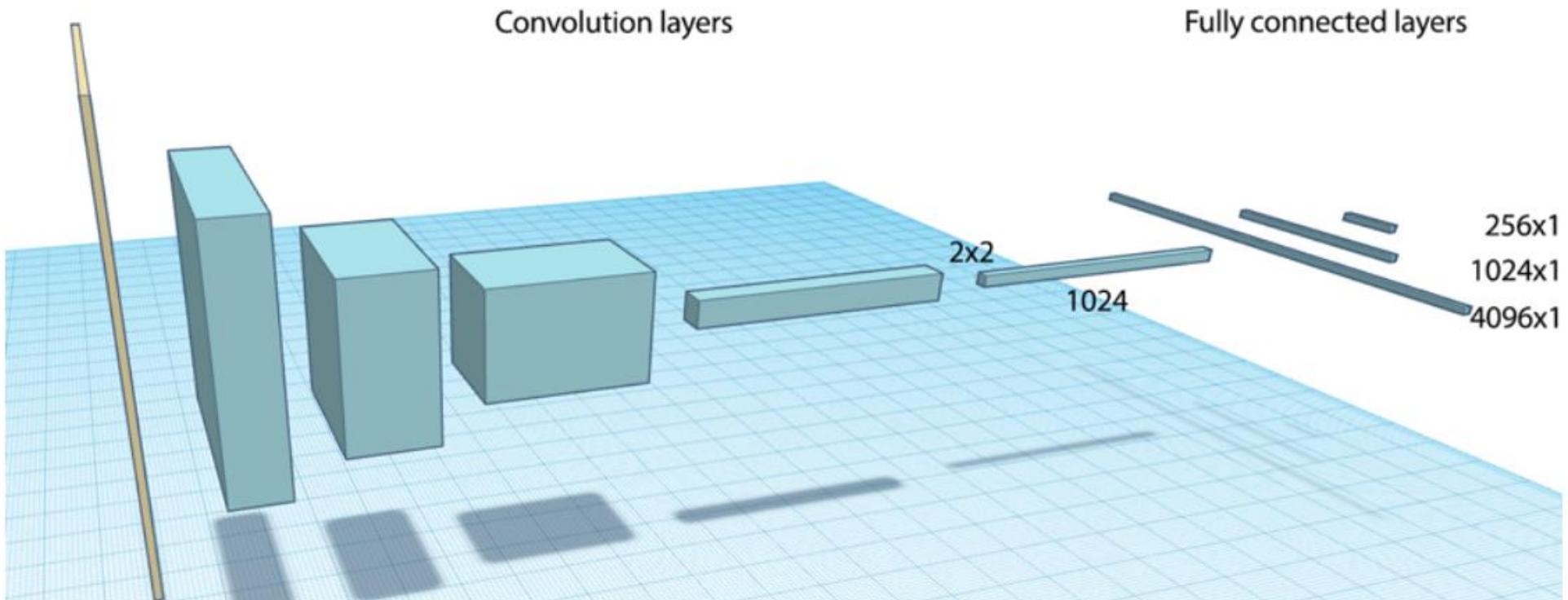
- spatial resolution is reduced
- the number of maps increases



A typical architecture

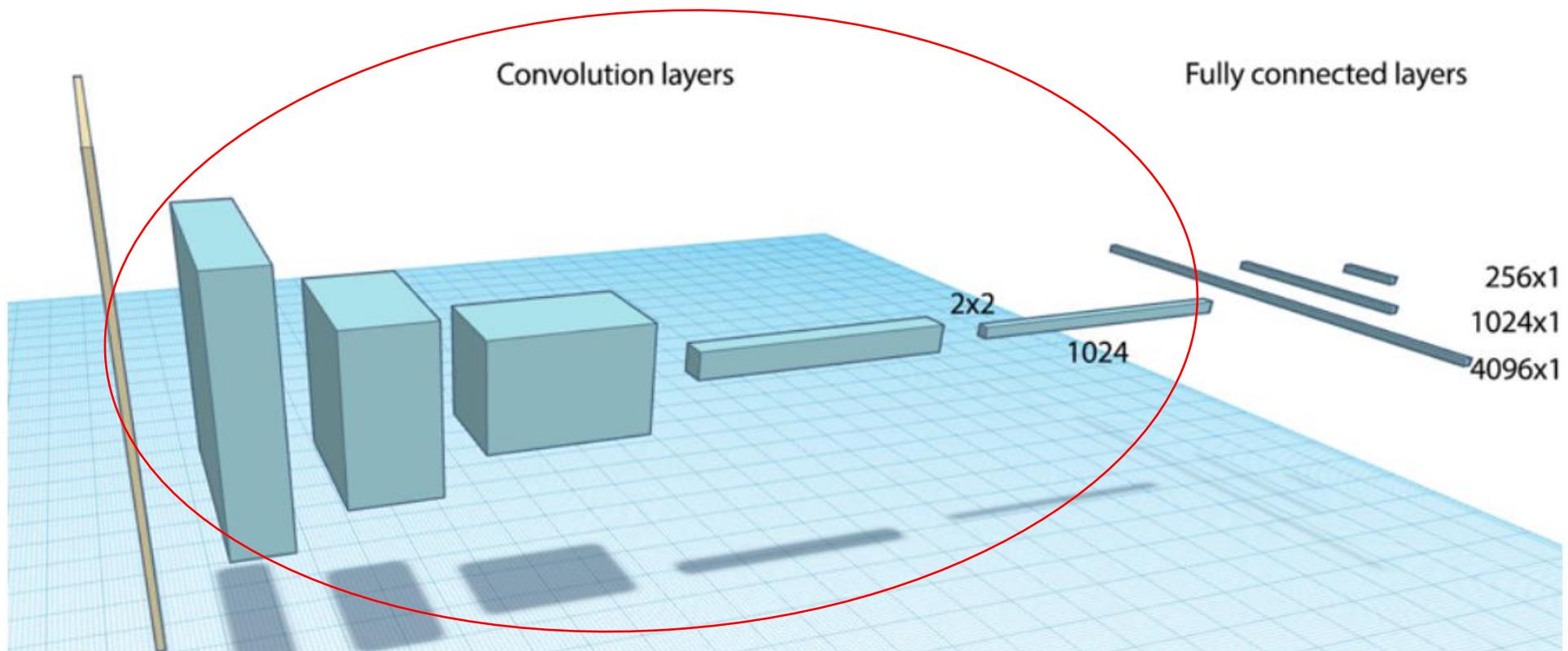
As we move to deeper layers:

- We search for higher-level patterns, and don't care too much about their exact location.
- There are **more high-level patterns** than low-level details!



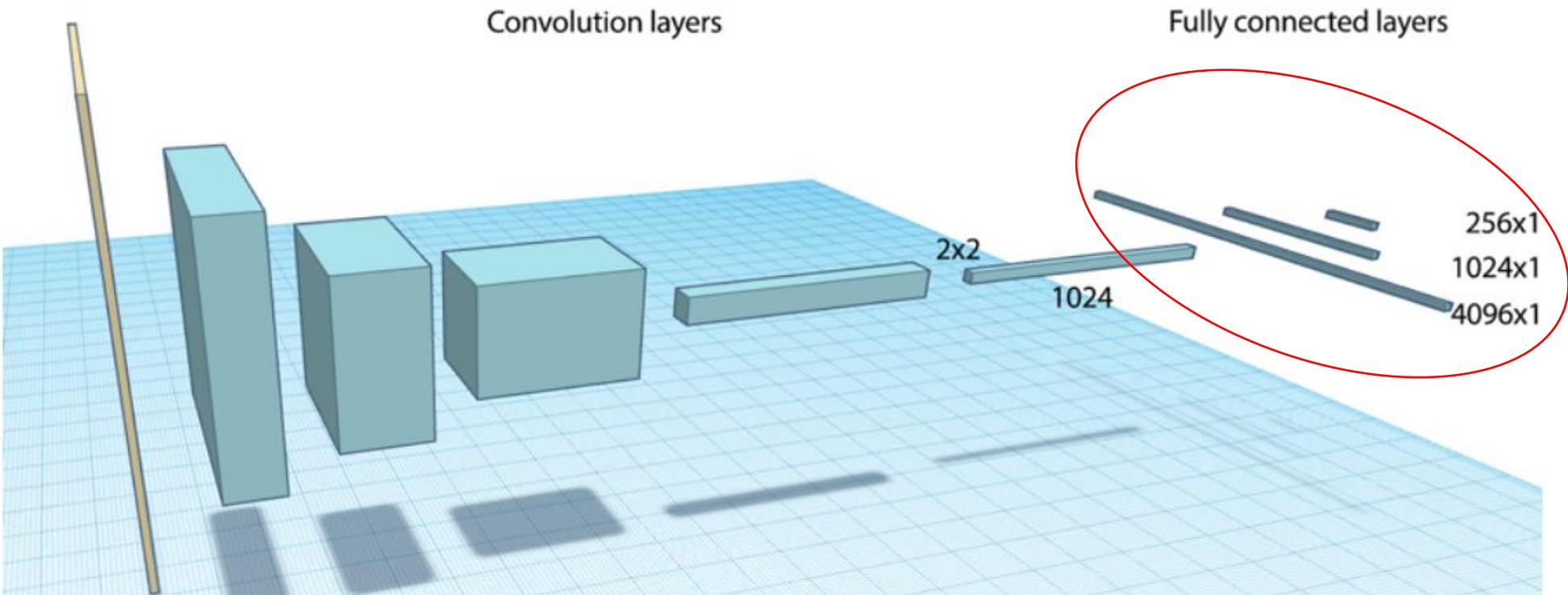
A typical architecture: Convolutional Block

In VGG or Alexnet network that was trained to classify 1000 classes from Imagenet...
the convolutional part should be a very general feature extractor



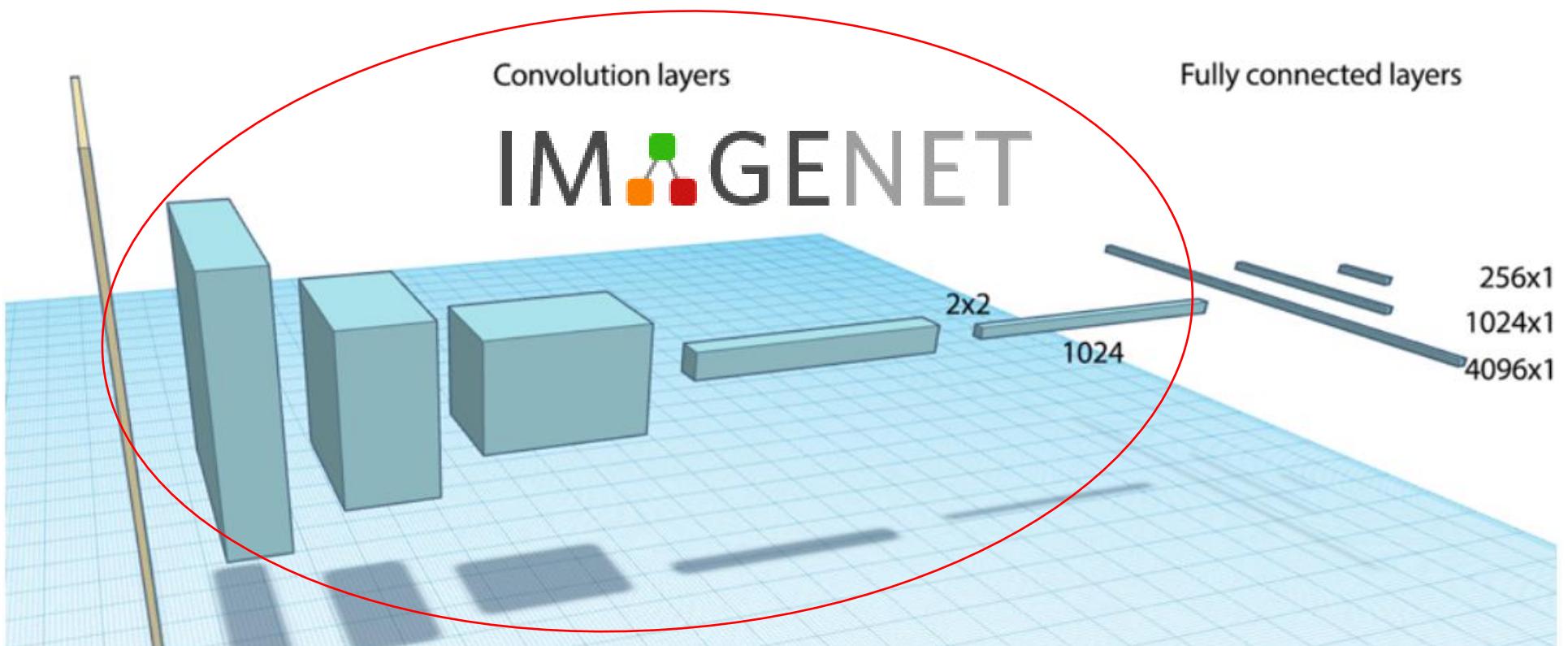
A typical architecture: FC layers

The FC layers are instead very custom, as they are meant to solve the specific classification task at hand



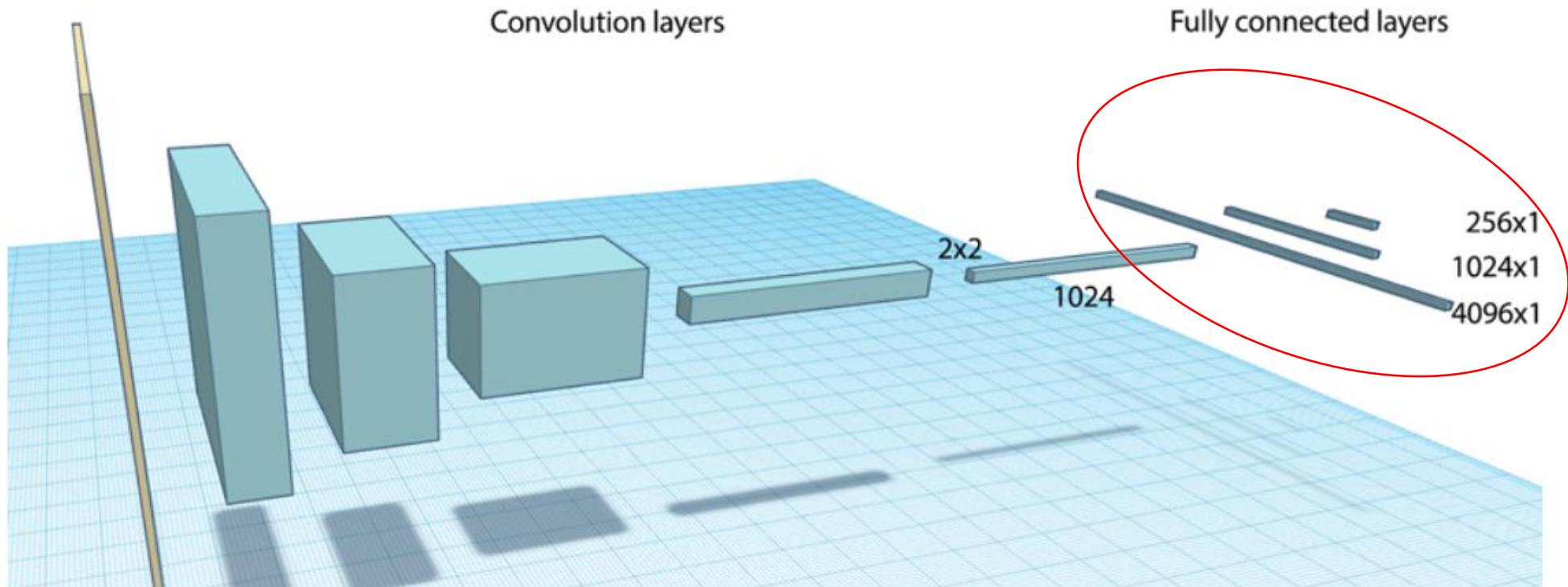
Transfer Learning

- Take a successful pre-trained model such as VGG



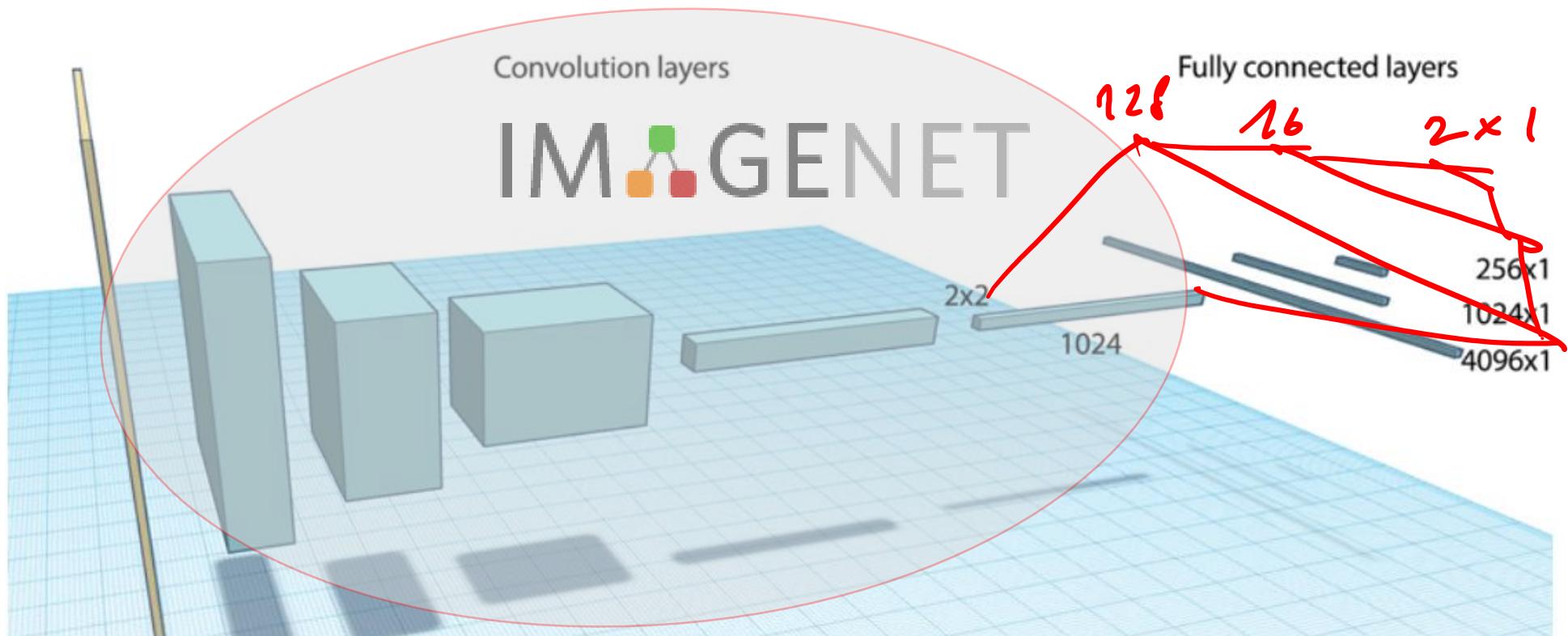
Transfer Learning

- Take a successful pre-trained model such as VGG
- Remove and modify the fully connected layers for the problem at hand



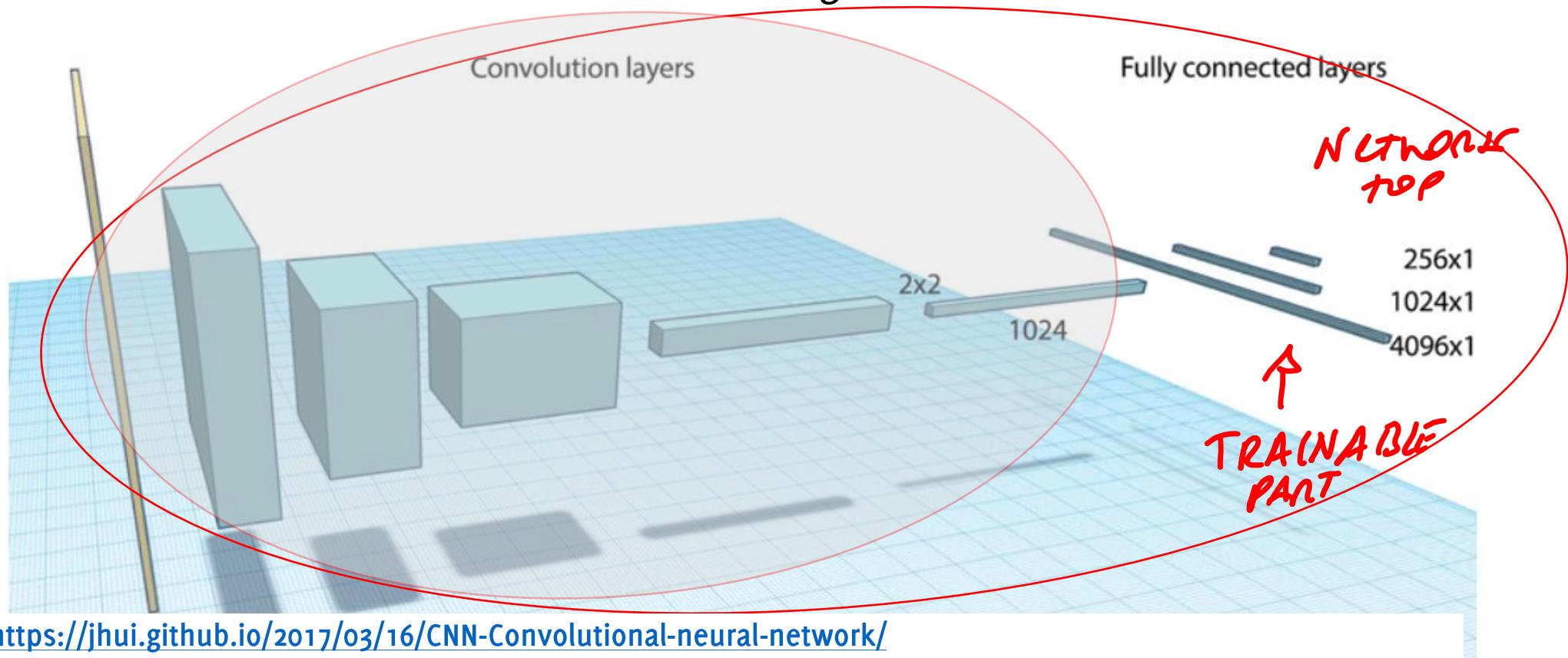
Transfer Learning

- Take a successful pre-trained model such as VGG
- Remove and modify the fully connected layers for the problem at hand
- «Freeze» the weights in the convolutional layers. **Possibly modify FC architecture**



Transfer Learning

- Take a successful pre-trained model such as VGG
- Remove and modify the fully connected layers for the problem at hand
- «Freeze» the weights in the convolutional layers. Possibly modify FC architecture
- Train the whole network on the new training data



In keras...

Pre-trained models are available, typically in two ways:

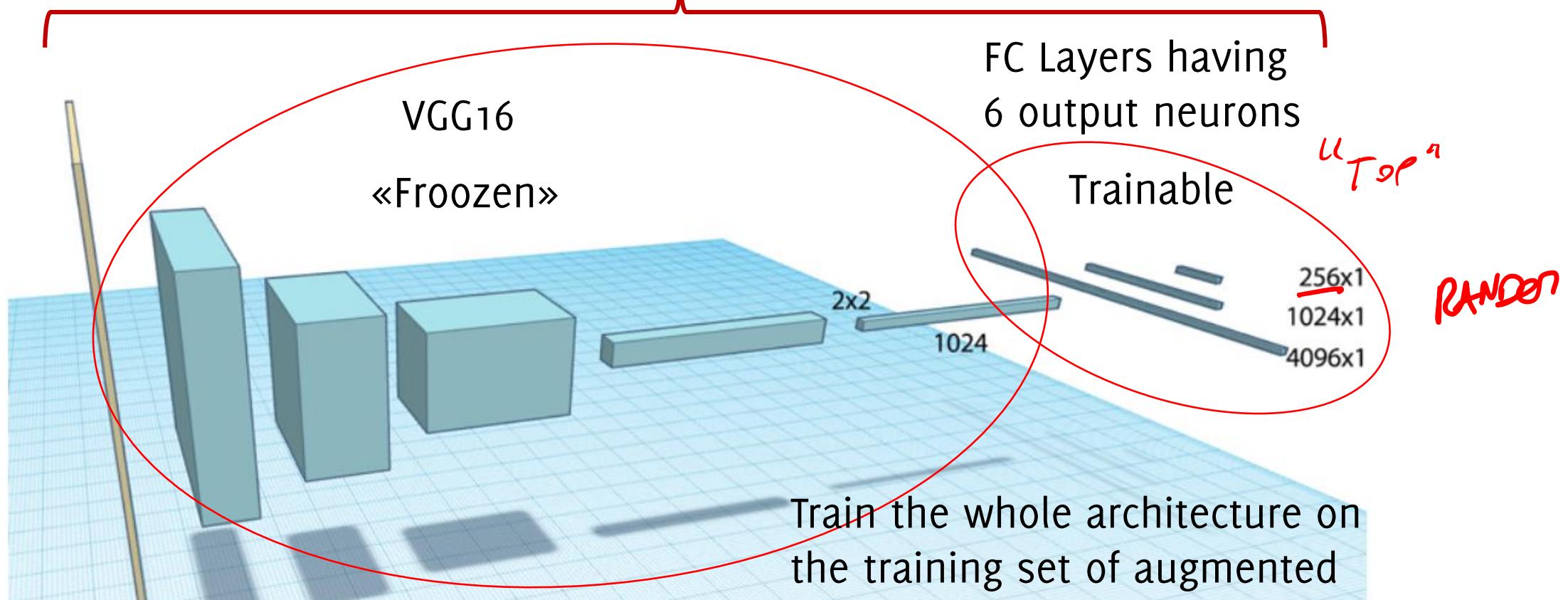
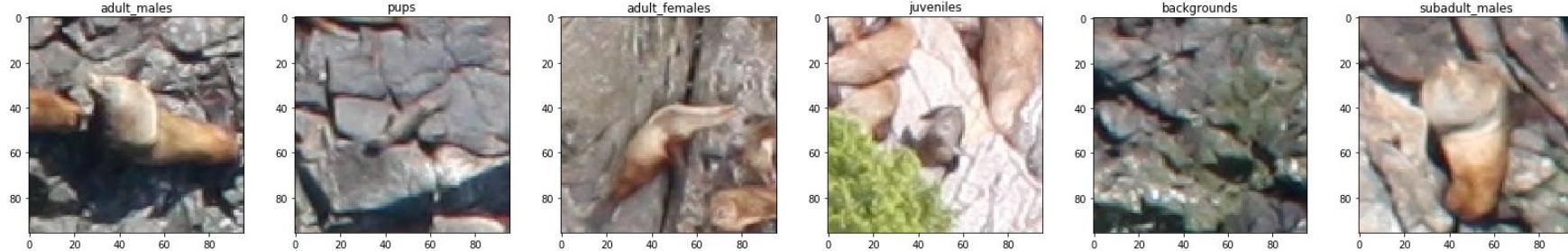
- **include_top = True**: provides the entire network, including the fully convolutional layers. This network can be used to solve the classification problem it was trained for
- **include_top = False**: contains only the convolutional layers of the network, and it is specifically meant for transfer learning.

Have a look at the size of these models in the two options!

VGG16 in keras...

```
from keras import applications  
  
base_model = applications.VGG16(weights =  
"imagenet", include_top=False, input_shape =  
(img_width, img_width, 3))
```

Transfer Learning in the Sealion Case



Transfer Learning in Keras...

Requires a bit of TensorFlow Backend to add the modified Fully connected layer at the top of a pretrained model

Then, before training it is necessary to loop through the network layers (they are in `model.layers`) and then modify the trainable property

```
for layer in model.layers[: lastFrozen]:  
    layer.trainable=False
```



Transfer Learning

Different Options:

- Transfer Learning: only the FC layers are being trained. A good option when little training data are provided and the pre-trained model is expected to match the problem at hand
- Fine tuning: the whole CNN is retrained, but the convolutional layers are initialized to the pre-trained model. A good option when enough training data are provided or when the pre-trained model is not expected to match the problem at hand.

Typically, for the same optimizer, lower learning rates are used when performing fine tuning than when training from scratches