University of Zurich[UZH] | ETH zürich

ROBOTICS & PERCEPTION GROUP

Institute of Informatics – Institute of Neuroinformatics

Deep Learning

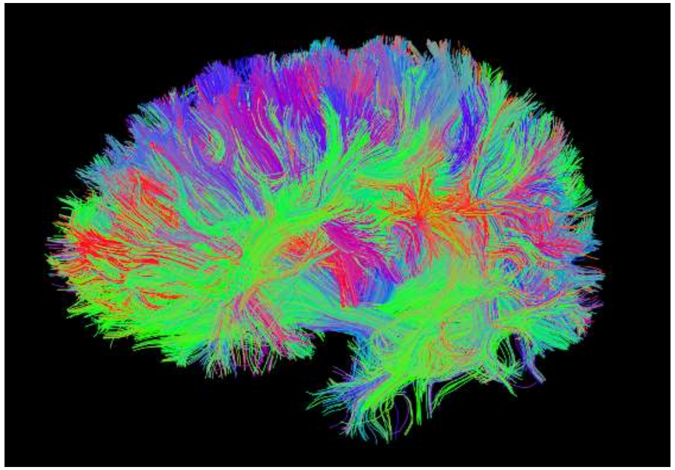
Antonio Loquercio

Outline

- **Introduction**
  - Motivation and history
- **Supervised Learning**
  - The image classification problem
  - Artificial Neural Networks
- **Applications to Computer Vision**
  - General problems
  - Applications to visual odometry
- **Applications to Robotics**
- **Conclusions**

The Deep Learning Revolution

Medicine

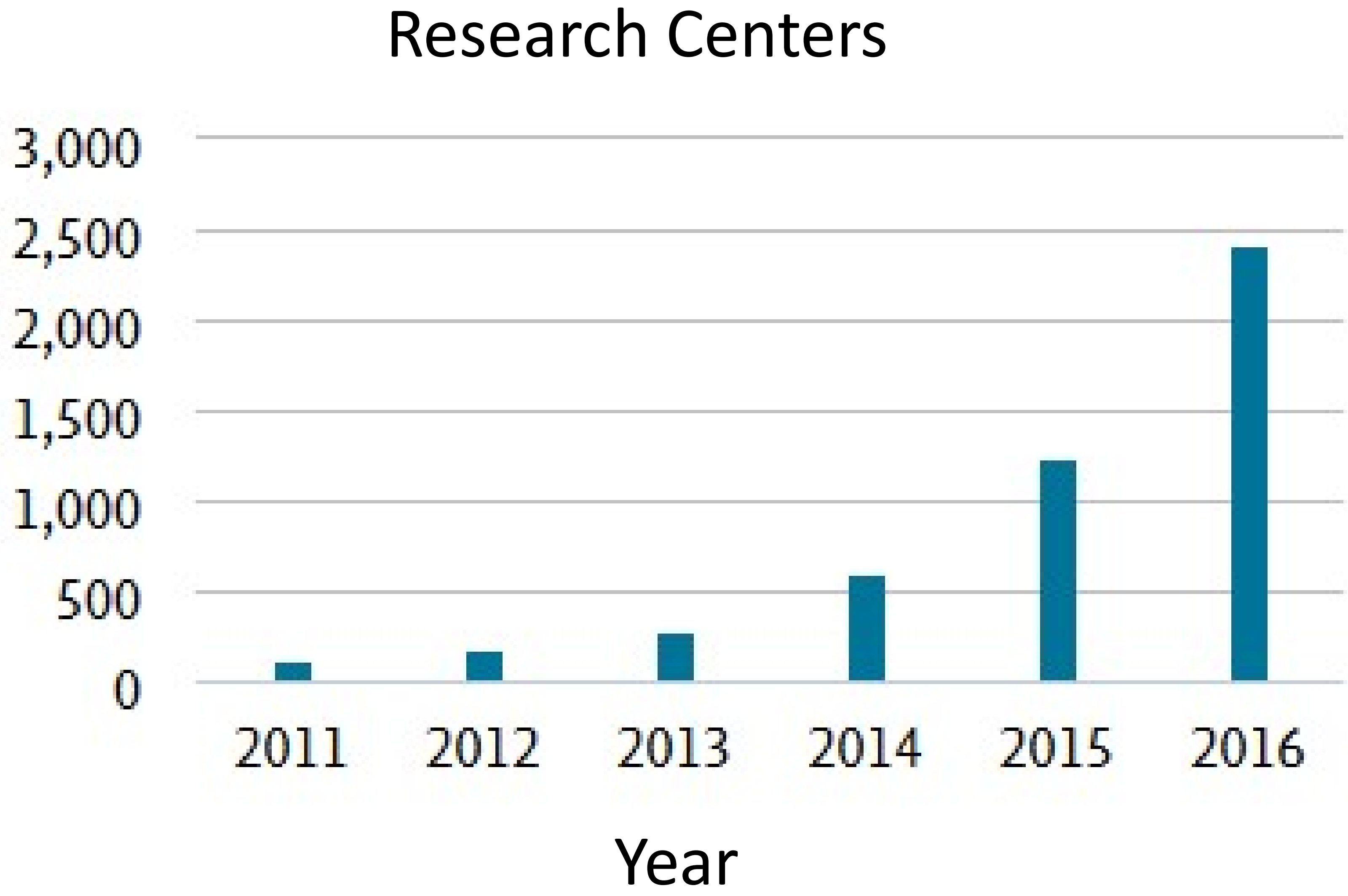


Media & Entertainment

Surveillance & Security

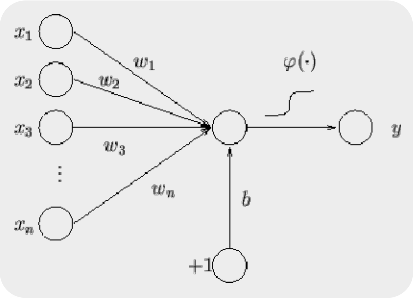
Autonomous Driving

The Deep Learning Revolution



Source: Bloomberg

Some History

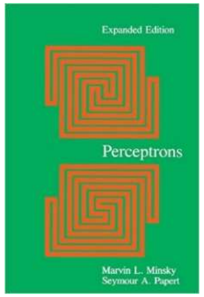


Perceptron

1958

1969

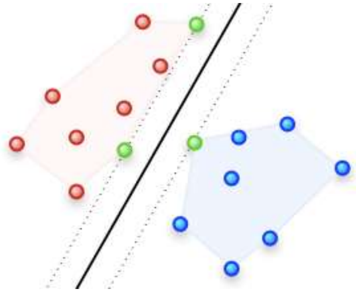
Critics



Back-Propagation

1974

AI Winter



SVM

1995

1998

Convolutional Neural Networs for Handwritten Digits Recognition

2006

Restricted Boltzman Machines

AlexNet

2012

What changed?

- Hardware Improvements

- Big Data Available

- Algorithmic Progress

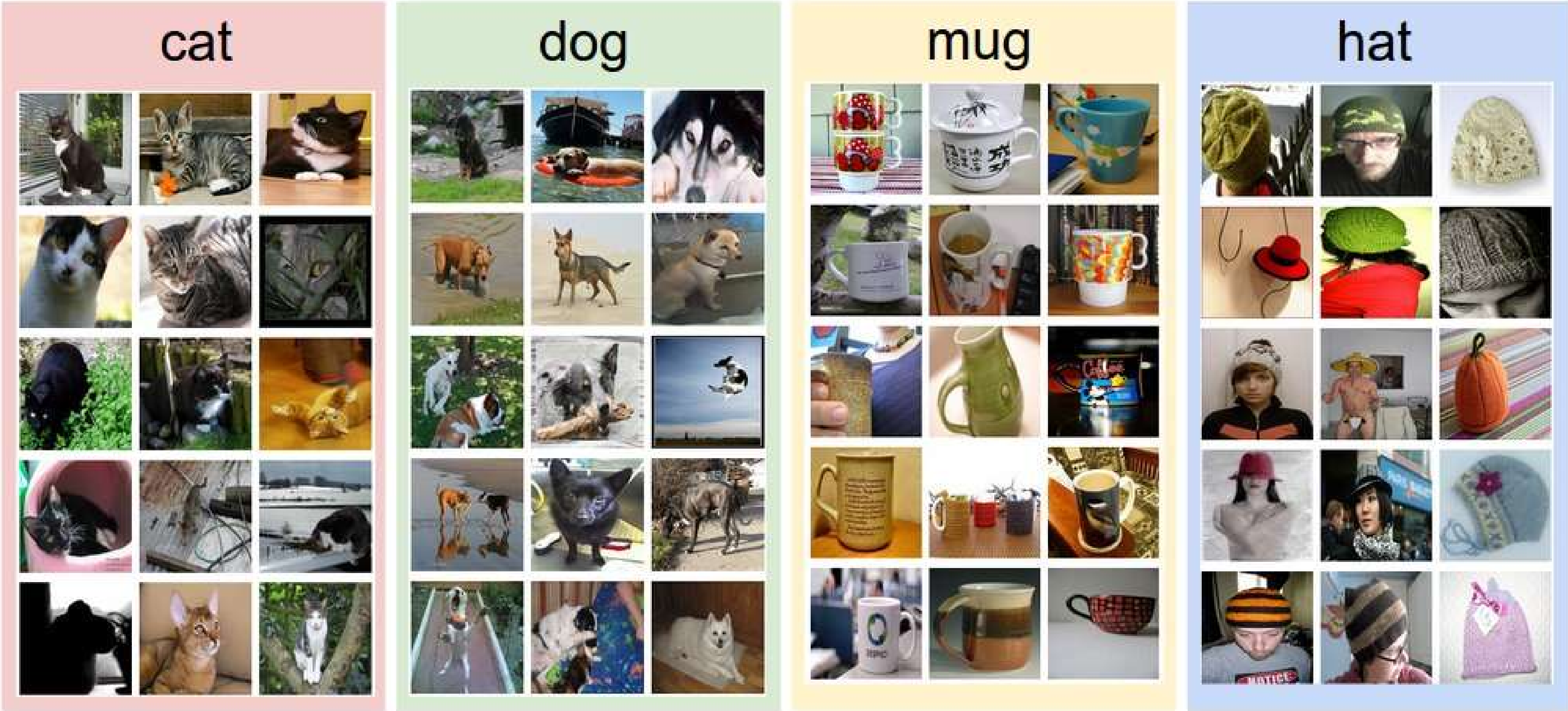
Hype or Reality?

Machine learning is a core transformative way by which we are **rethinking everything** that we are doing

Sundar Pichai (CEO Google)

Image Classification

Task of assigning an input image a label from a fixed set of categories.

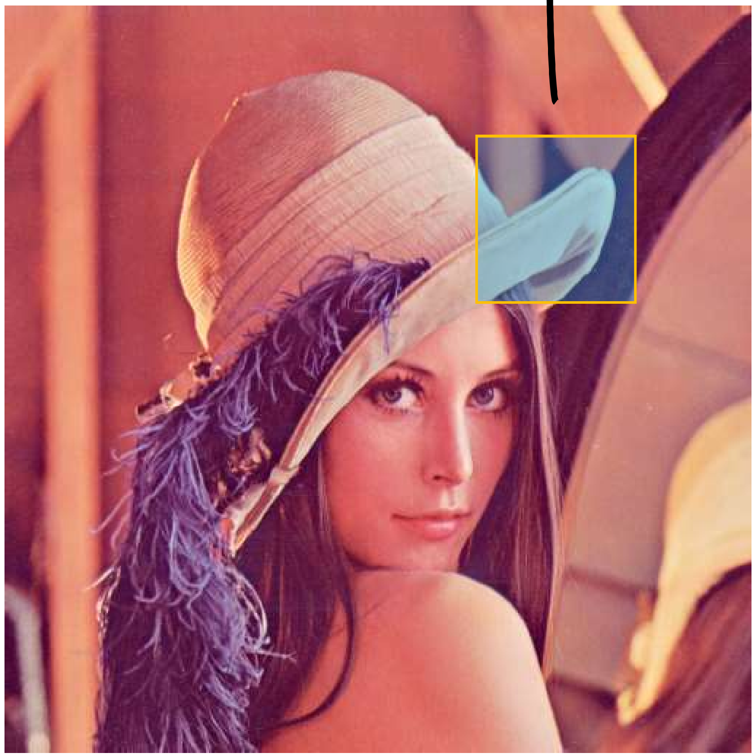


Slide adapted from CNNs for Visual Recognition (Stanford)

The semantic gap

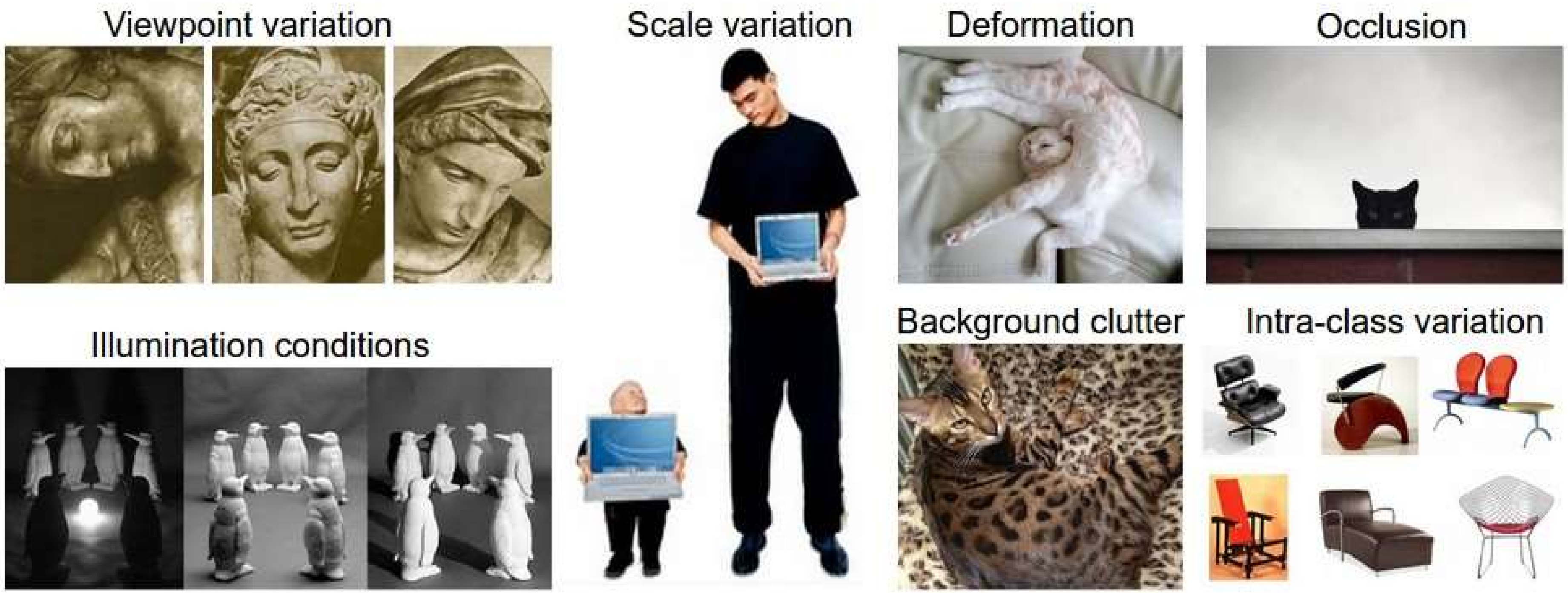
- What computers see against what we see

401	402	403	404	405	406	407	408	409	410
411	412	413	414	415	416	417	418	419	420
421	422	423	424	425	426	427	428	429	430
431	432	433	434	435	436	437	438	439	440
441	442	443	444	445	446	447	448	449	450
451	452	453	454	455	456	457	458	459	460
461	462	463	464	465	466	467	468	469	470
471	472	473	474	475	476	477	478	479	480
481	482	483	484	485	486	487	488	489	490
491	492	493	494	495	496	497	498	499	500



Classification Challenges

Directly specifying how a category looks like is impossible.



We need use a **Data Driven Approach**

Slide adapted from CNNs for Visual Recognition (Stanford)

Supervised Learning

Find function $f(x, \theta)$ that imitates a ground truth signal

$f(x, \theta)$

N numbers representing class scores

Predicted	Ground truth, y_i
$\begin{pmatrix} 0.1 \\ 0.7 \\ \cdots \\ 0.0 \end{pmatrix}$	$\begin{pmatrix} 1.0 \\ 0.0 \\ \cdots \\ 0.0 \end{pmatrix}$

Function parameters or weights

Update

$$Loss(f(x_i, \theta), y_i)$$

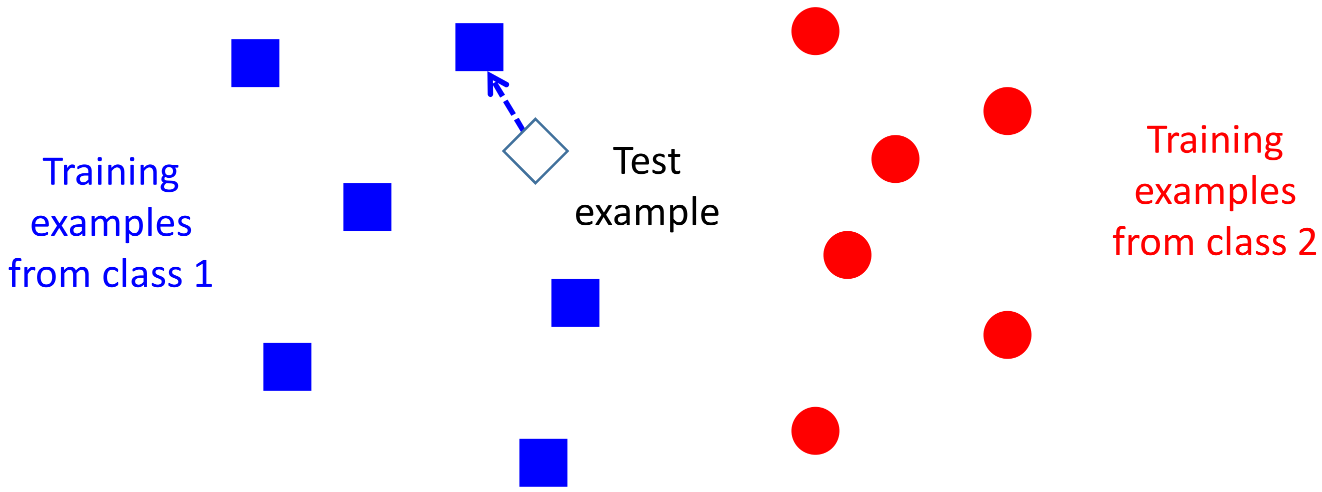
Machine Learning Keywords

- Loss: Quantify how good θ are
- Optimization: The process of finding θ that minimize the loss
- Function: Problem modelling -> Deep networks are highly non-linear $f(x, \theta)$

Slide adapted from CNNs for Visual Recognition (Stanford)

Classifiers: K-Nearest neighbor

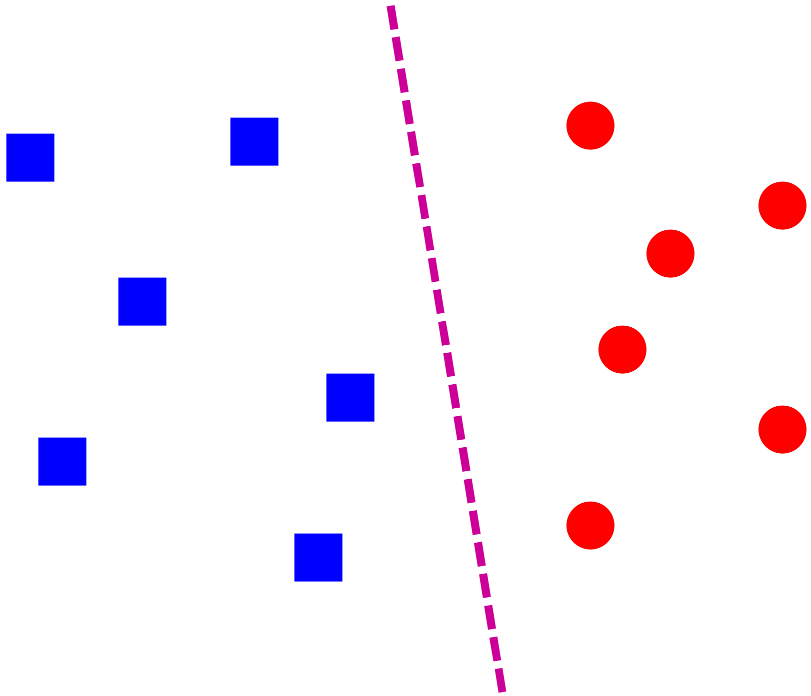
Features are represented in the descriptor space



$f(\boldsymbol{x}, \theta)$ = label of the K training examples nearest to $\boldsymbol{x}$

- How fast is training? How fast is testing?
  - O(1), O(n)
- What is a good distance metric ? What K should be used? ☹

Classifiers: Linear



- Find a *linear function* to separate the classes:

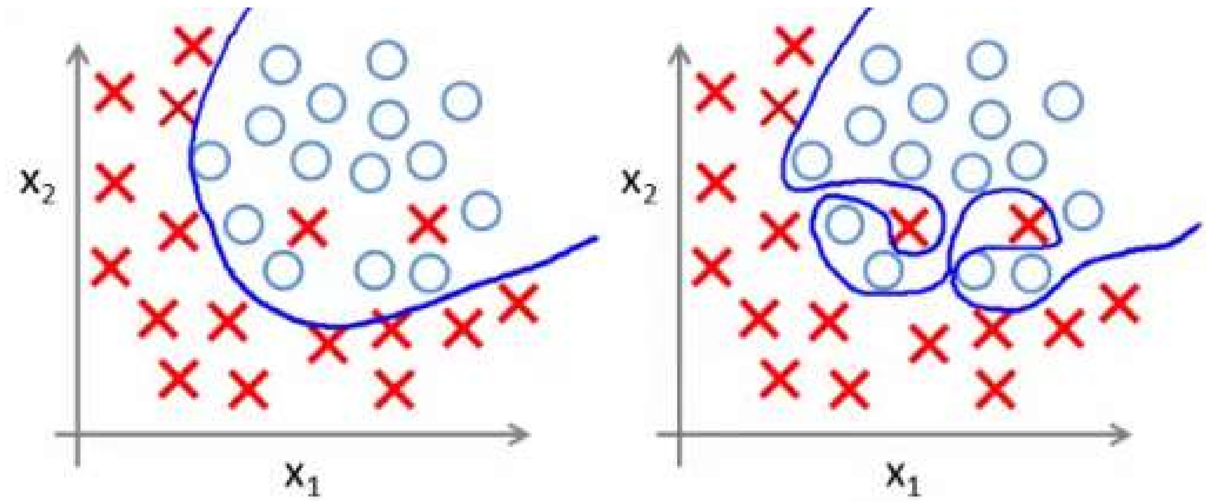
$$f(\boldsymbol{x}, \theta) = sgn(\boldsymbol{w} \cdot \boldsymbol{x} + b)$$

- What is now $\boldsymbol{\theta}$? What is the dimensionality of images?

Classifiers: non-linear

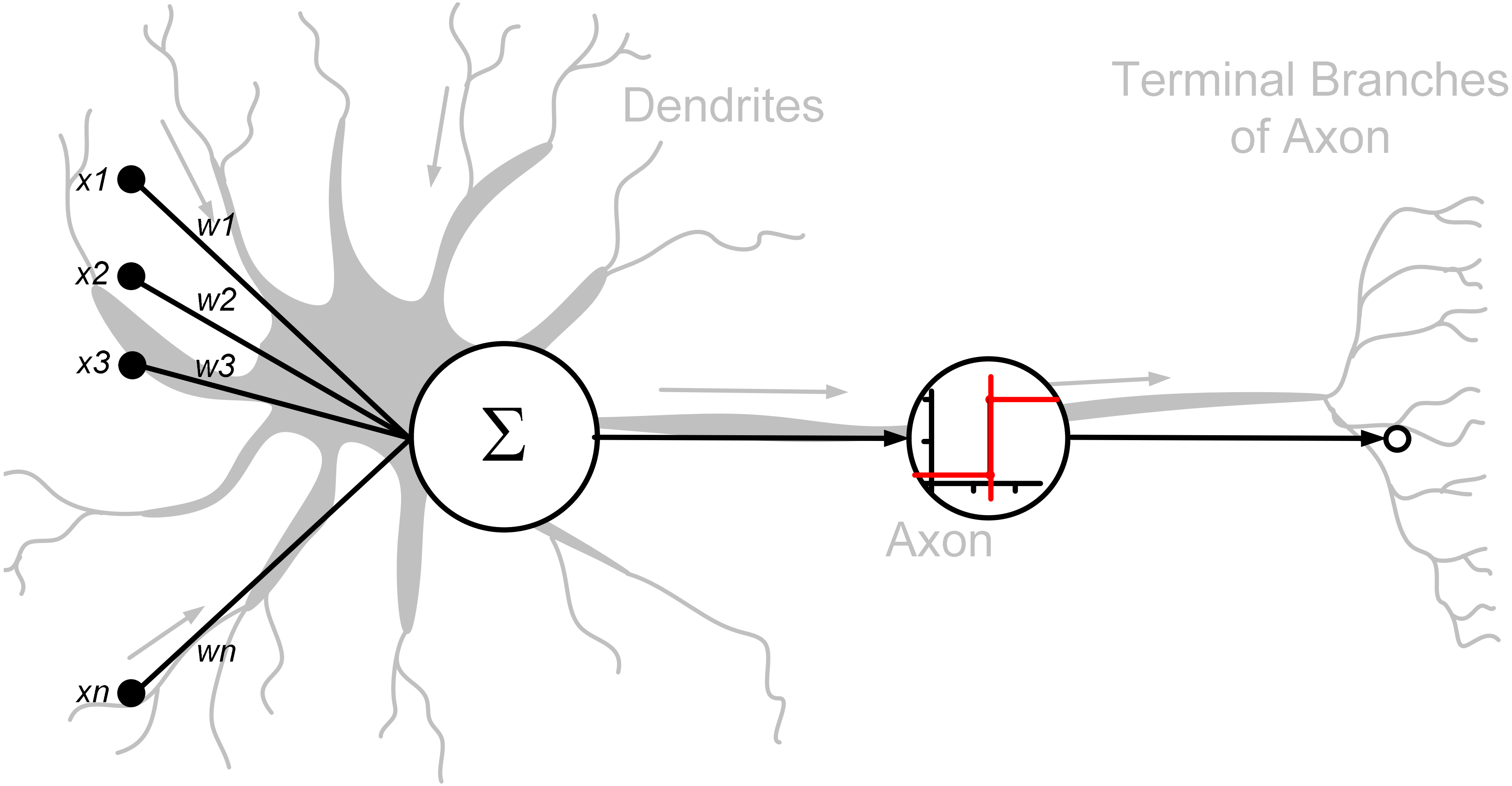
Good classifier

Bad classifier (over fitting)



- What is $f(\boldsymbol{x}, \theta)$?

Biological Inspiration

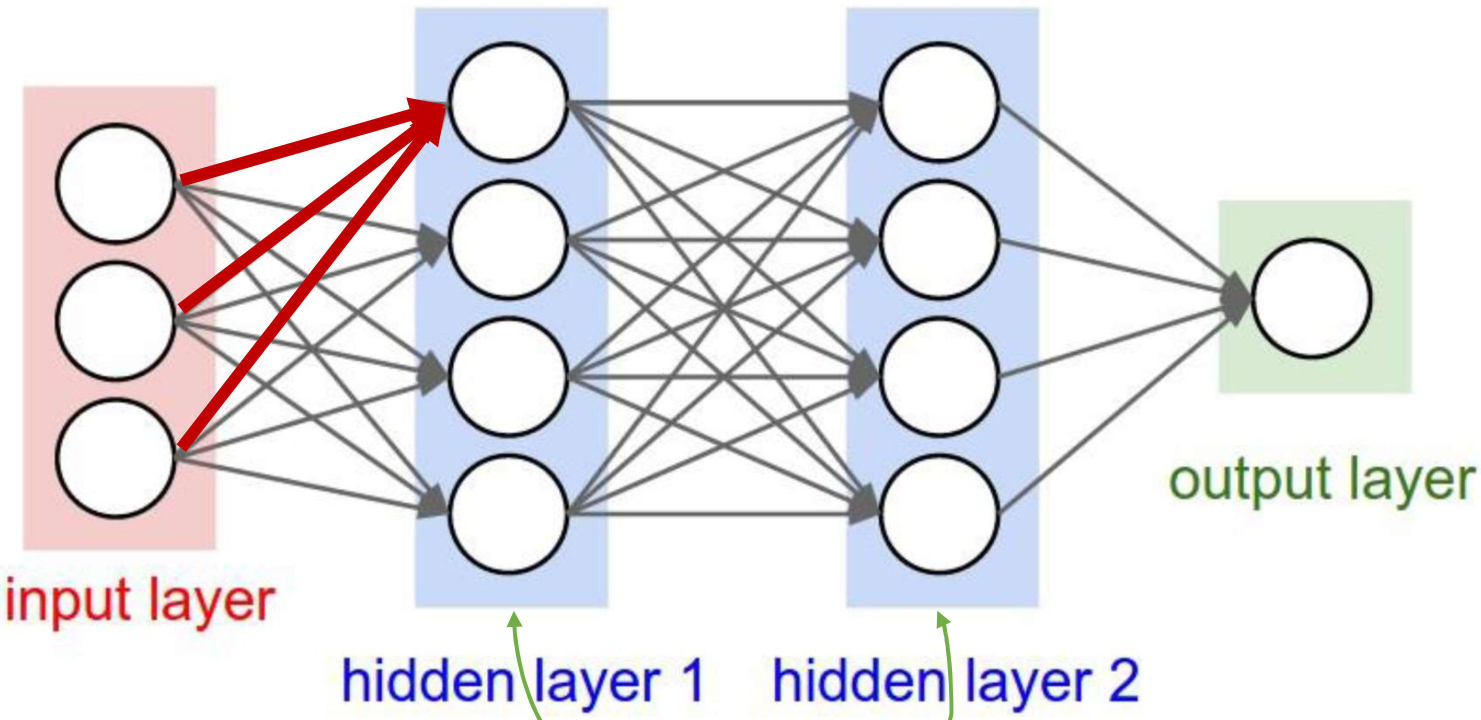


f$(x, \theta) = F(Wx)$, F is a non-linear activation function (Step, ReLU, Sigmoid)

The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Frank Rosenblatt (1958)

Fully Connected Neural Networks

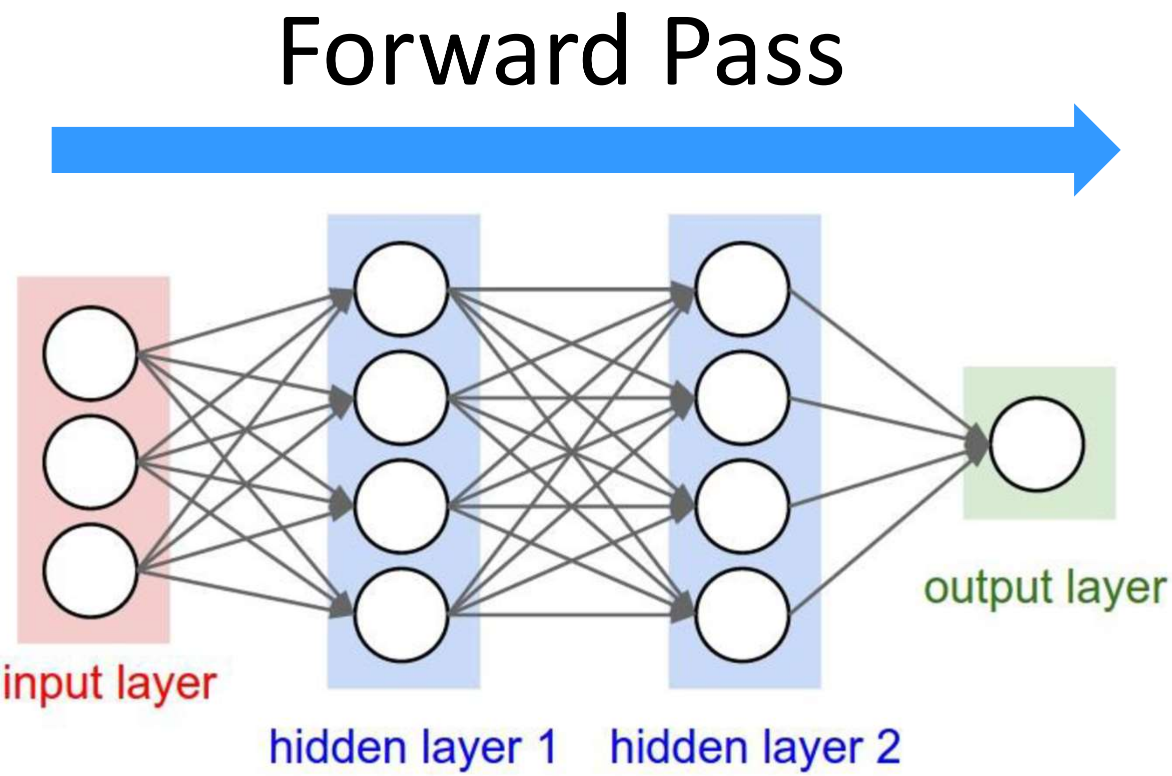
$\boldsymbol{f}\,(\boldsymbol{x}, \boldsymbol{\theta})$



Non-linear Activation functions (ReLU, sigmoid, etc.)

Neural Networks and Deep Learning, Chapter 2– Michael Nielsen

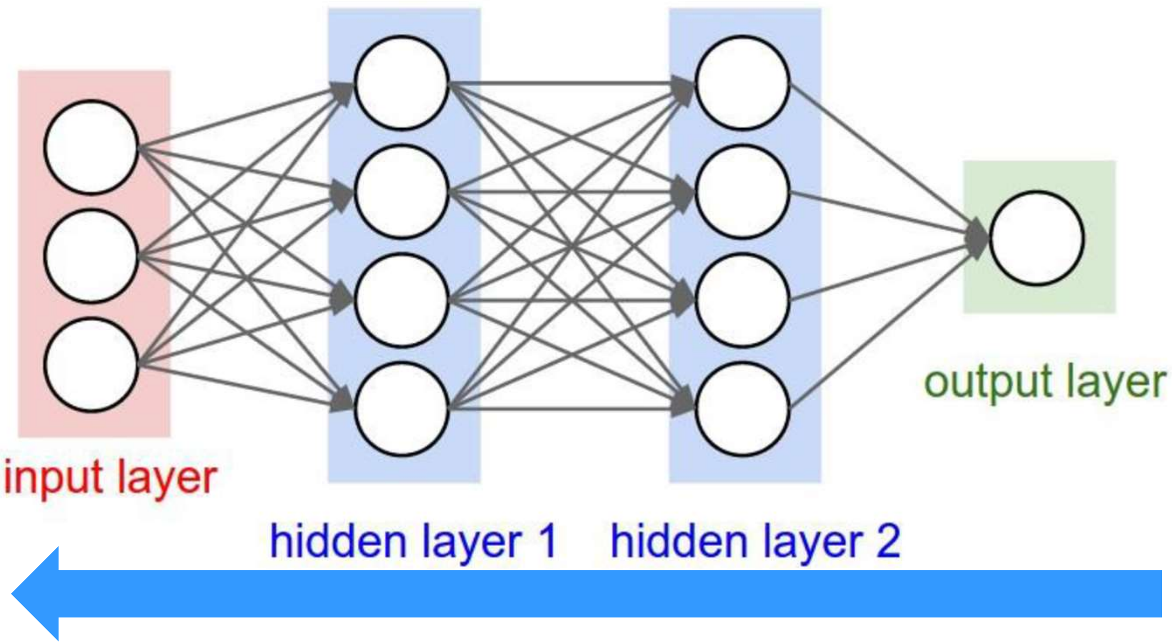
Forward propagation



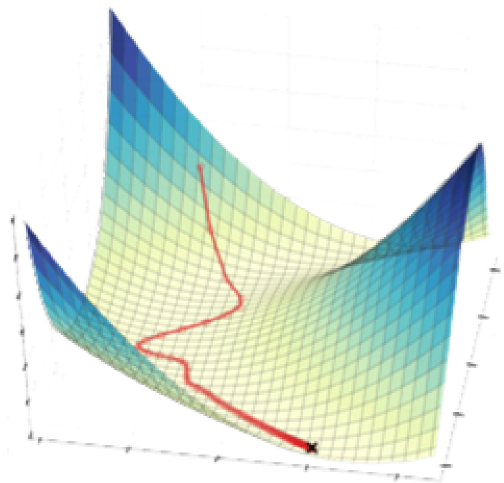
$$Loss(f(x_i, \theta), y_i)$$

Neural Networks and Deep Learning, Chapter 2– Michael Nielsen

Optimization: Back-propagation



$$Loss(f(x_i, \theta), y_i)$$



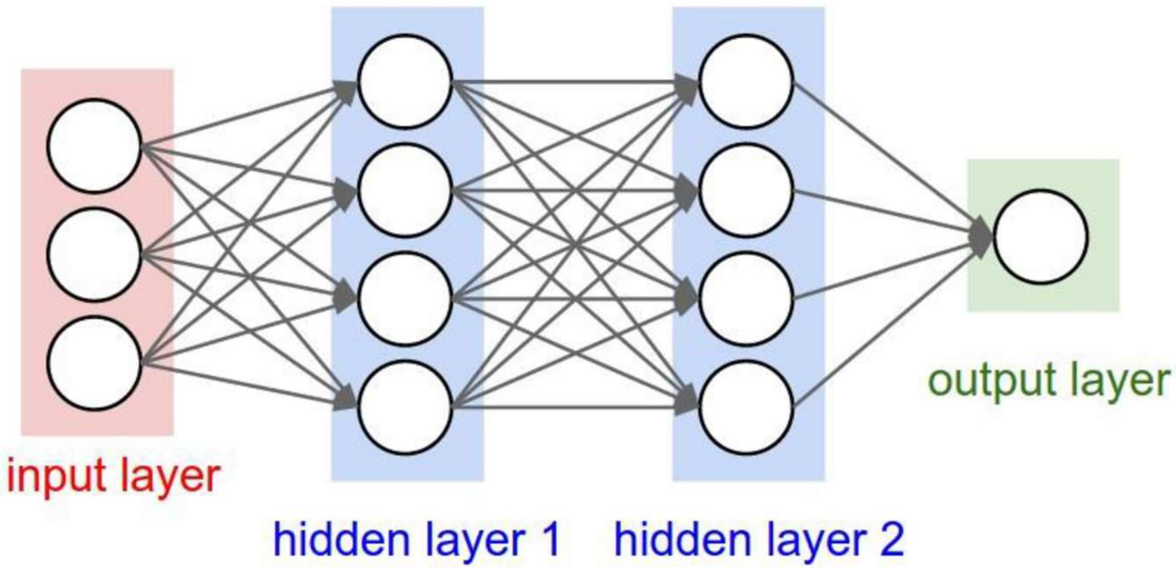
Backward Pass

Compute gradients with respect to all parameters and perform **gradient descent**

$$\theta_{new} = \theta_{old} - \mu \nabla Loss$$

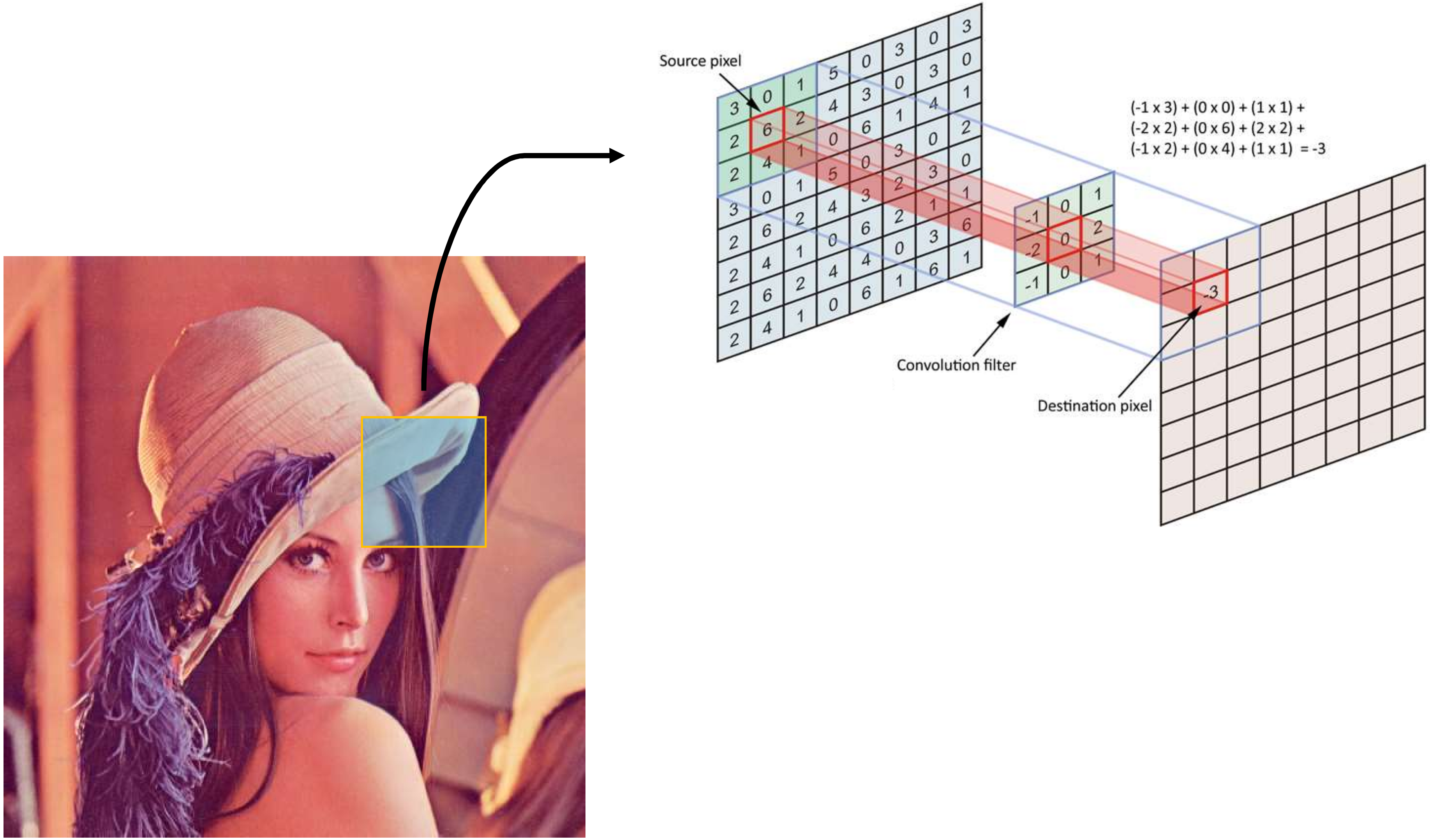
Artificial Neural Networks, Back Propagation and the Kelley-Bryson Gradient Procedure – Stuart E .Dreyfus

Problems of fully connected network



- Too many parameters -> possible overfit
- We are not using the fact that inputs are images!

Convolutional Neural Networks

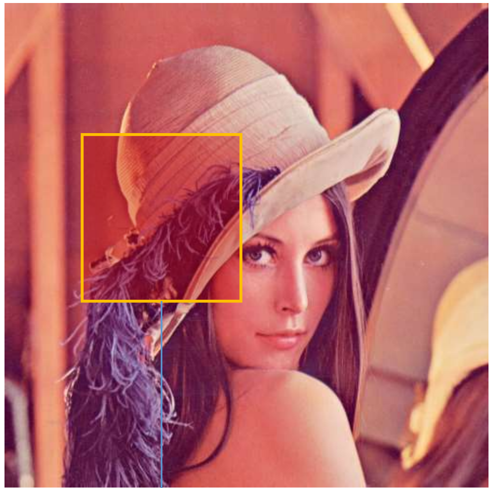


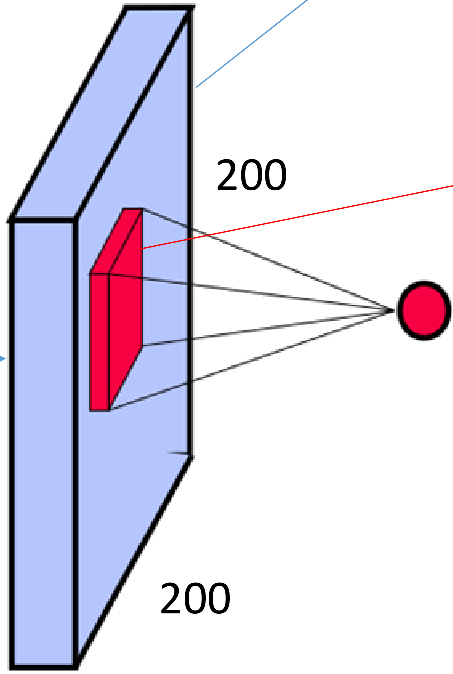
Gradient-based learning applied to document recognition, Y. LeCun et al. (1998)

Convolutional Filters

200x200x3 image

5x5x3 filter, W



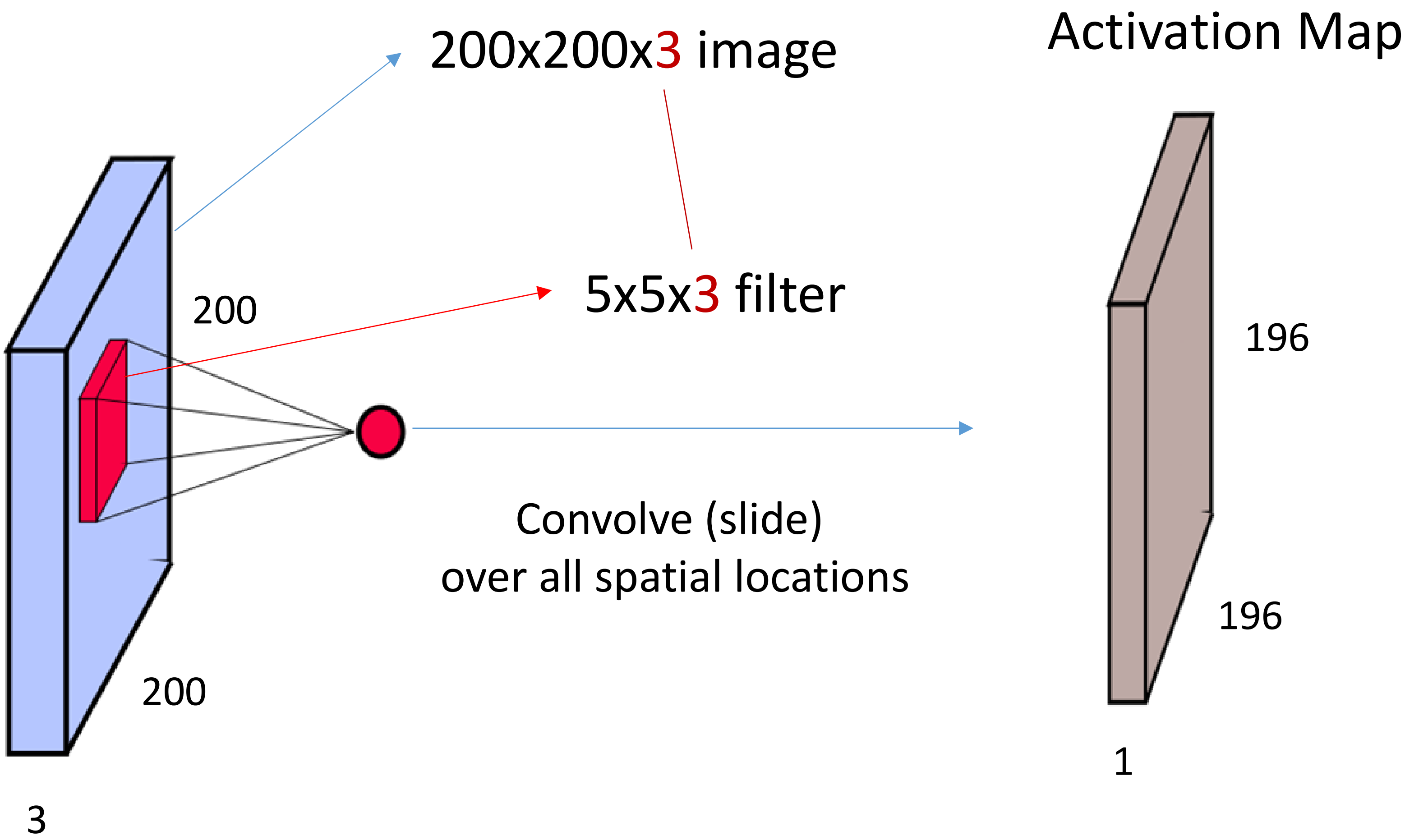


Single Number:

$W^T s$

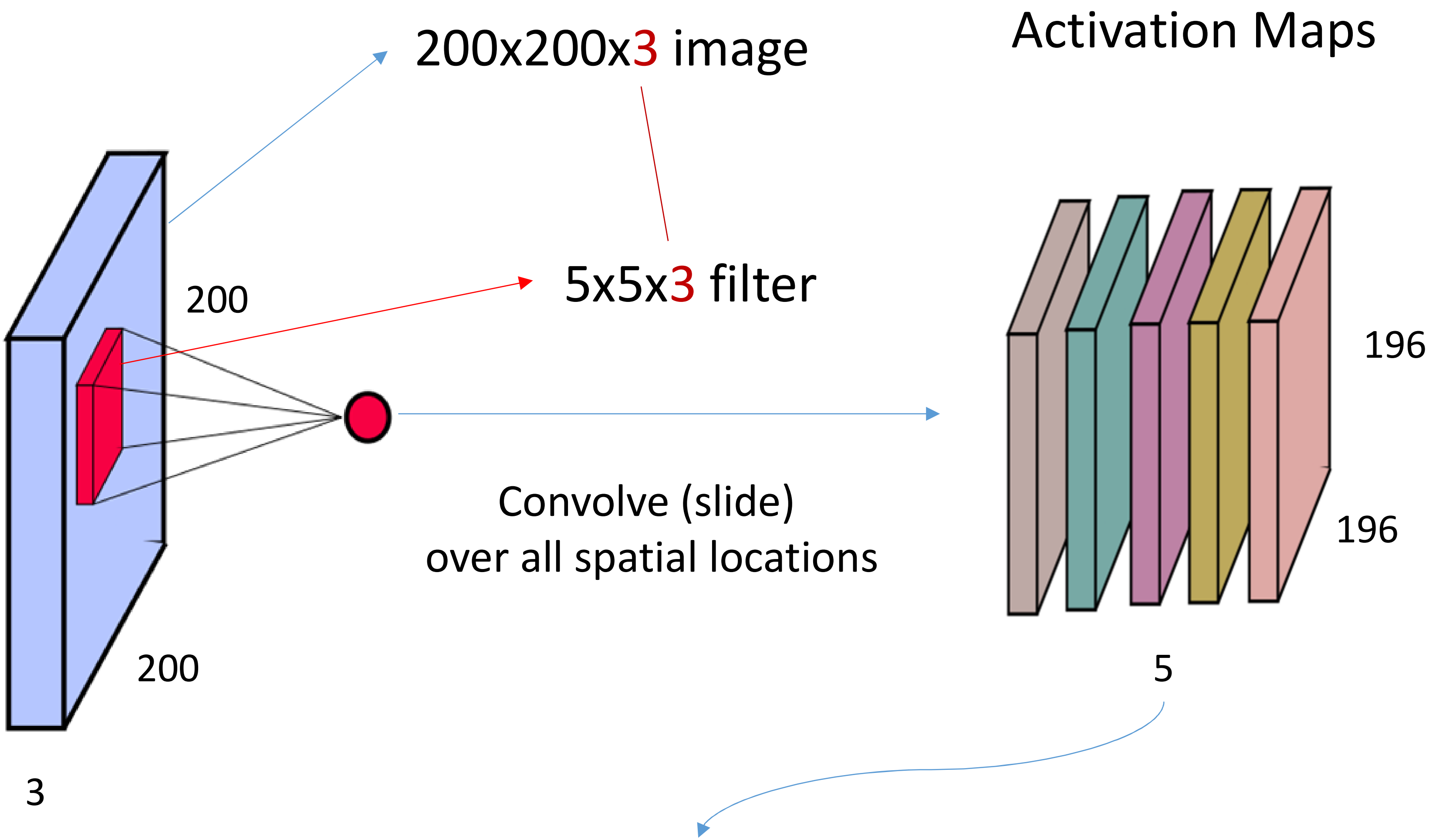
5x5x3 spatial location, s

Convolutional Layer



Convolutional Layer

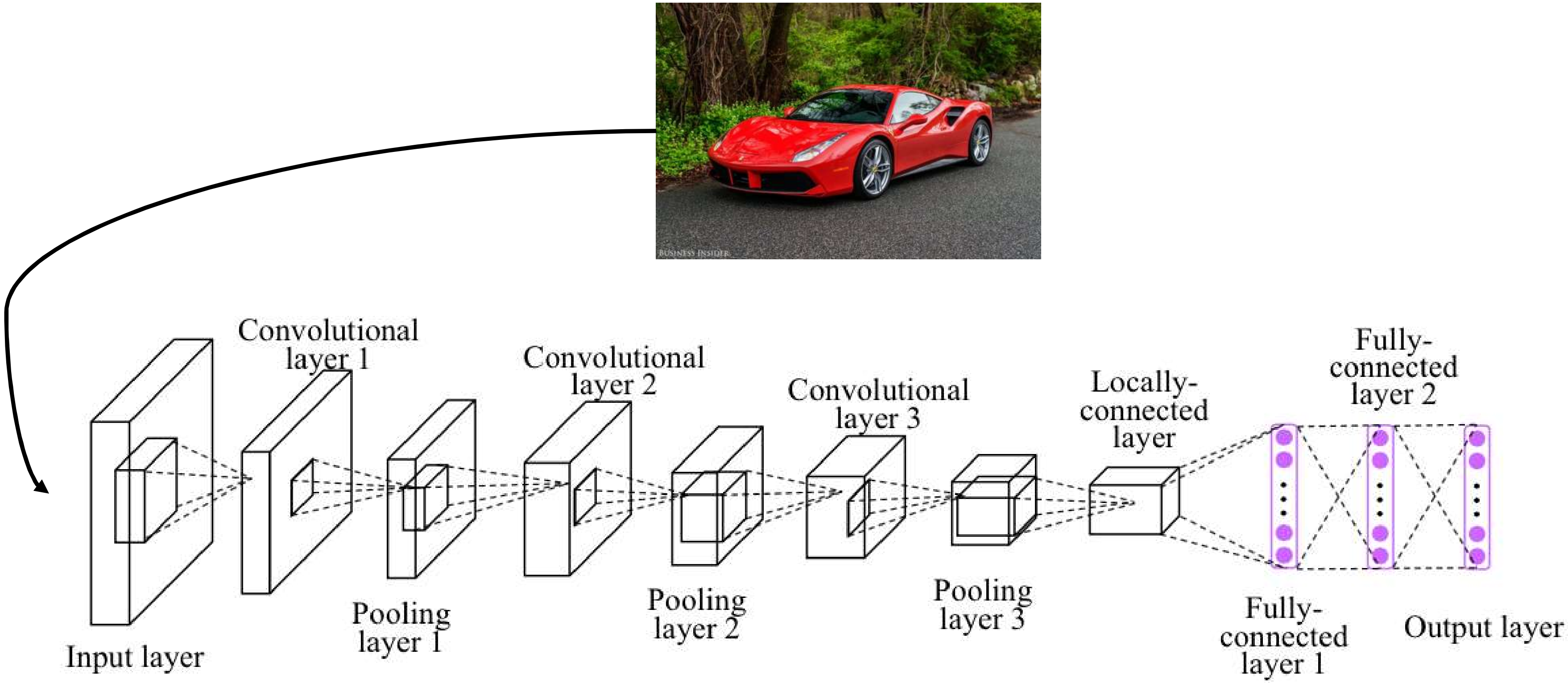
Repeat the operation for M filters

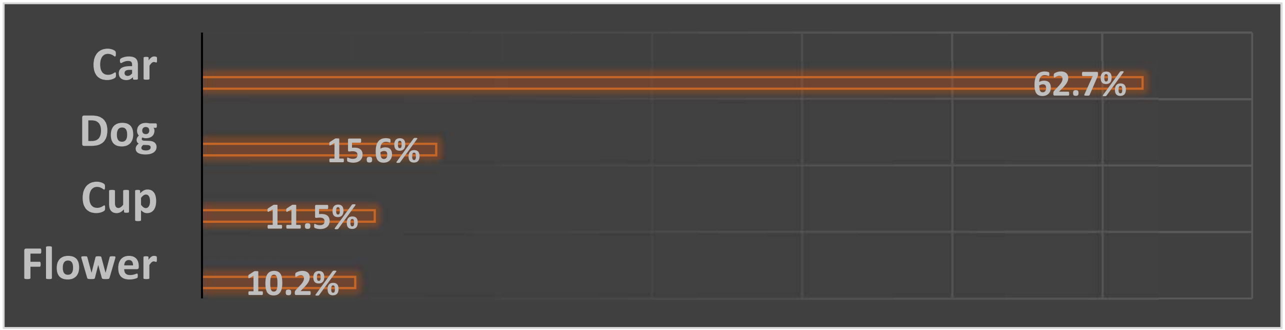


New «image», 196x196x5

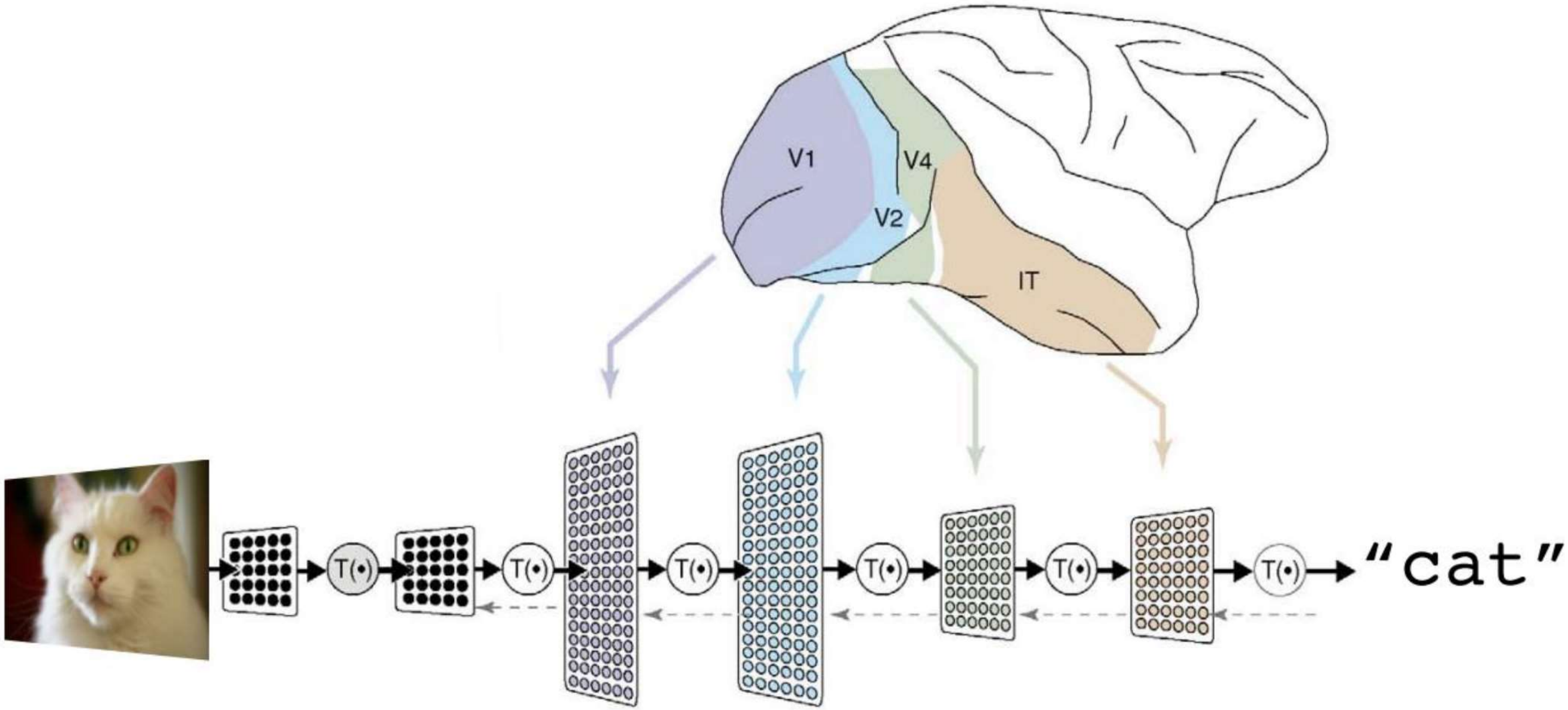
Gradient-based learning applied to document recognition, Y. LeCun et al. (1998)

Going Deep


Input layer
Convolutional layer 1
Pooling layer 1
Convolutional layer 2
Pooling layer 2
Convolutional layer 3
Pooling layer 3
Locally-connected layer
Fully-connected layer 1
Fully-connected layer 2
Output layer


Car
62.7%
Dog
15.6%
Cup
11.5%
Flower
10.2%

Why Deep?



- Inspired by the **human visual system**
- Learn **multiple layers** of transformations of input
- Extract progressively more **sophisticated representations**

General Applications of Deep Learning to Computer Vision

Deep Learning in Computer Vision

Image Segmentation

Fully Convolutional Networks for Semantic Segmentation – J. Long, E. Shelhamer, 2015

Deep Learning in Computer Vision

Image Captioning

"little girl is eating piece of cake."

"baseball player is throwing ball in game."

"woman is holding bunch of bananas."

"black cat is sitting on top of suitcase."

"a young boy is holding a baseball bat."

"a cat is sitting on a couch with a remote control."

"a woman holding a teddy bear in front of a mirror."

"a horse is standing in the middle of a road."

Deep Visual-Semantic Alignments for Generating Image Descriptions – Karpathy et al., 2015

Deep Learning in Computer Vision

Image Localization

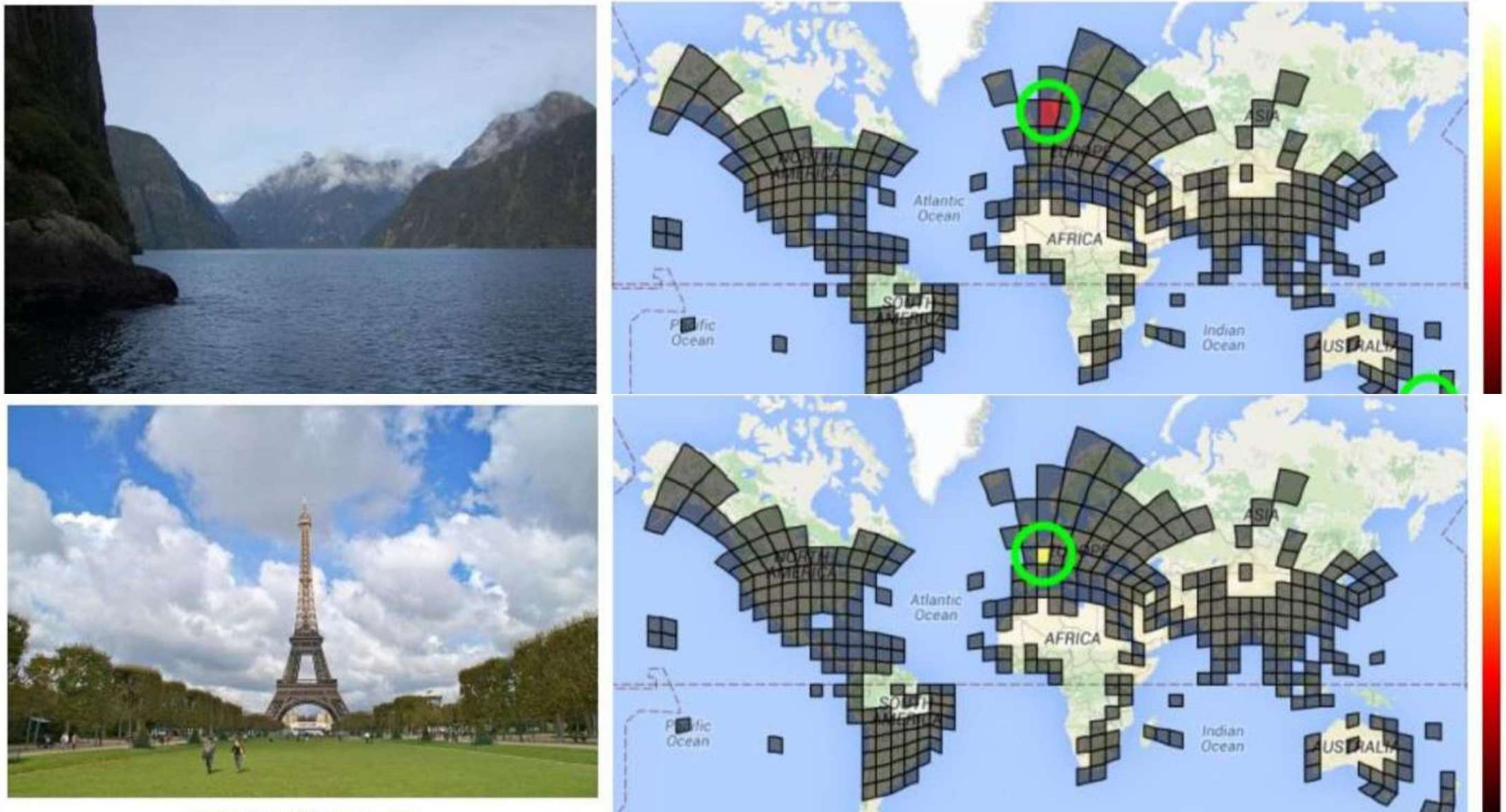


Photo CC-BY-NC by steveke

PlaNet - Photo Geolocation with Convolutional Neural Networks - Weyand et al. 2016

Deep Learning in Computer Vision

Adding Semantic Features

Deep Feature Interpolation for Image Content Changes – P. Upchurch, J. Gardner et al., 2016

Deep Learning in Computer Vision

Image Colorization

Image Colorization with Deep Convolutional Neural Networks – Hwang et al. 2016

Deep Learning in Computer Vision

Human Dreams - Pegasus

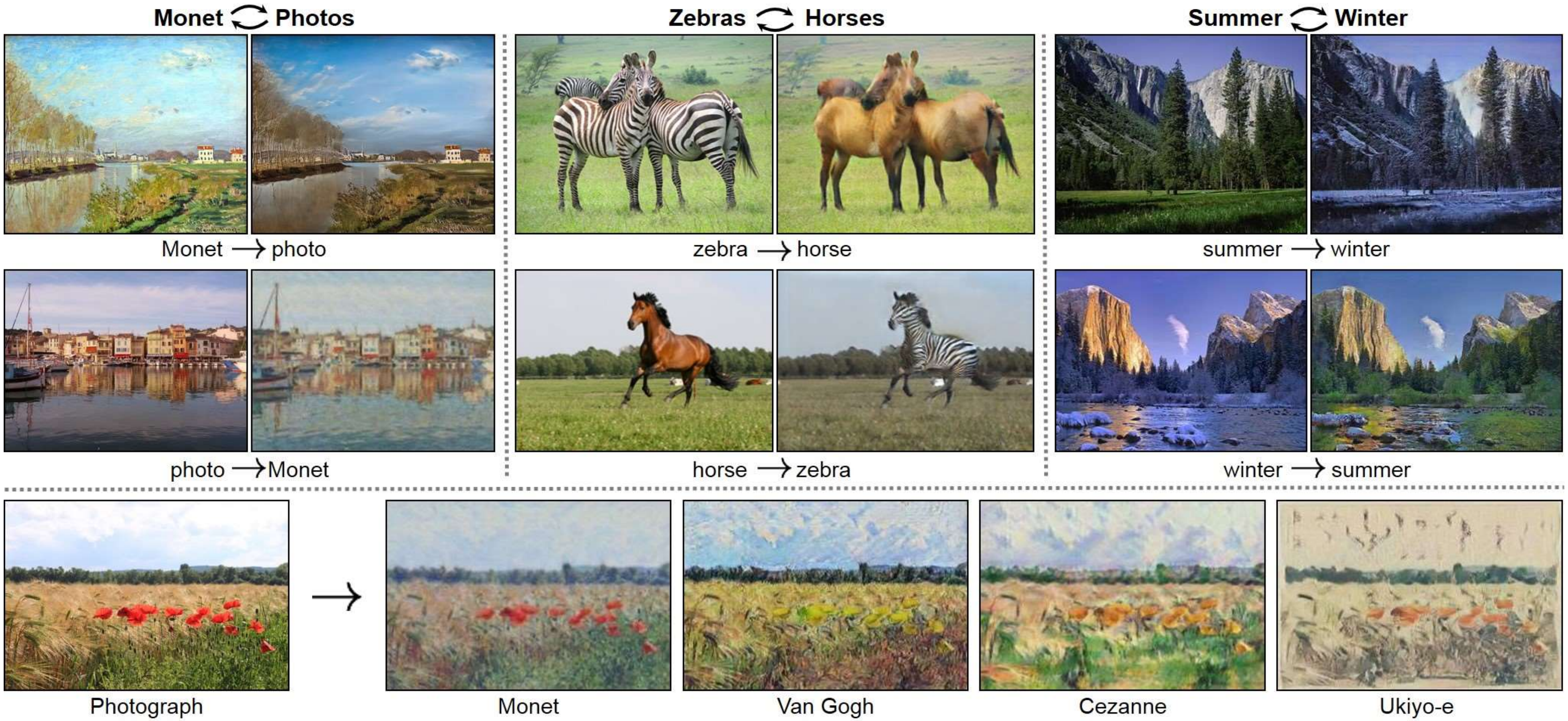
Deep Learning in Computer Vision

Machine Dreams - Inception

Going Deeper with Convolution – Szegedy et al., 2015

Deep Learning in Computer Vision

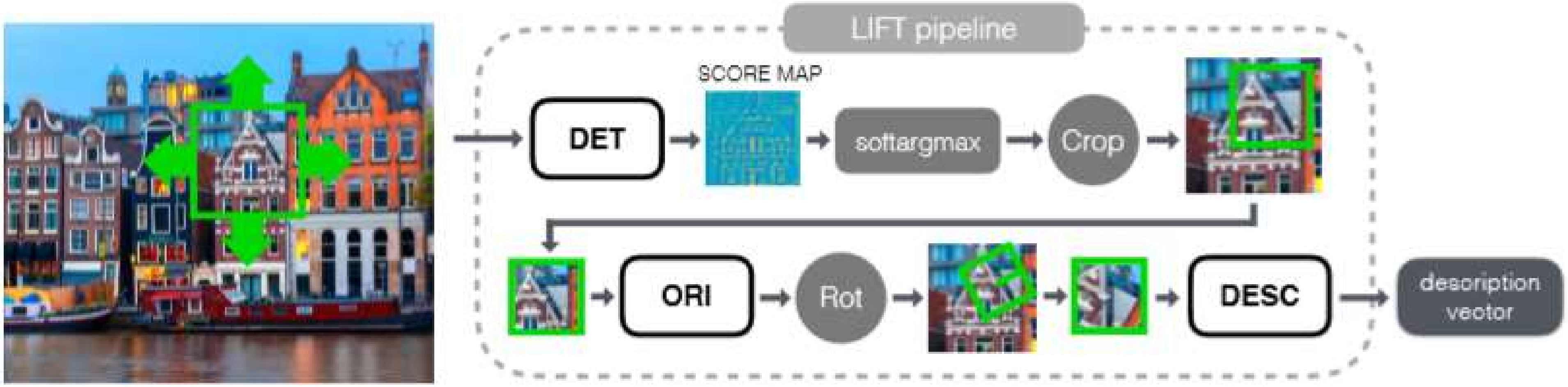
Image Transformation



Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks– Jun-Yan Zhu et al., 2017

Applications of Deep Learning to Visual Odometry

Deep Descriptors: LIFT

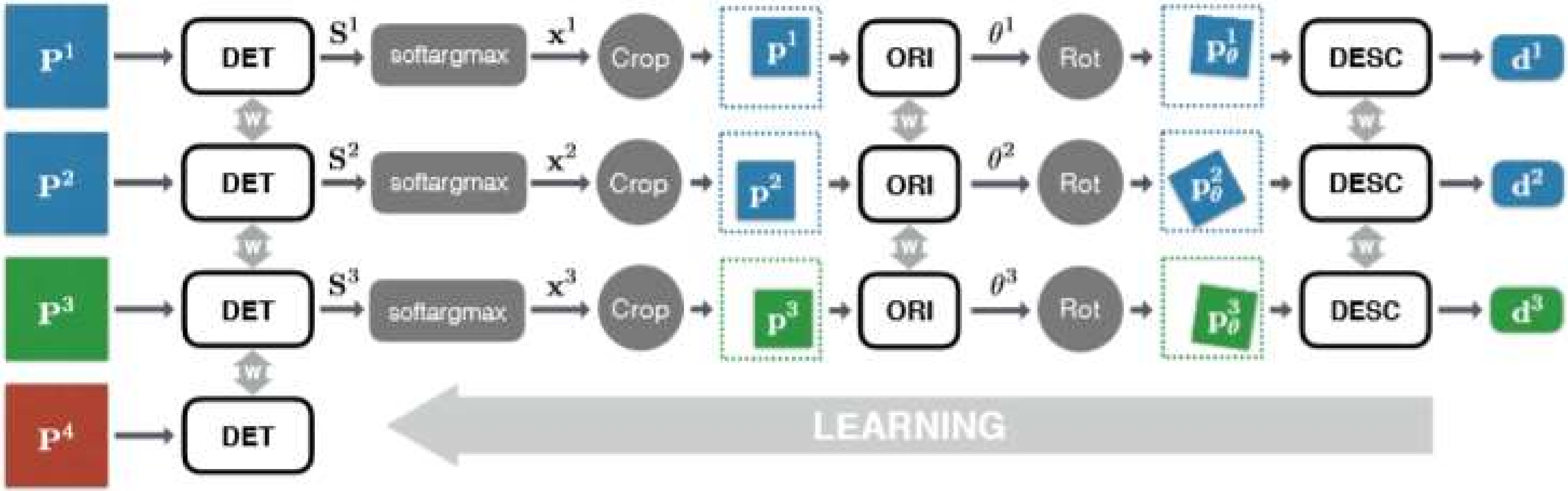


LIFT Pipeline consists of 3 neural networks:

- A keypoint detector
- An orientation detector
- A descriptor generator

LIFT: Learned Invariant Feature Transform, Kwang Moo Yi et al., 2016

LIFT Loss

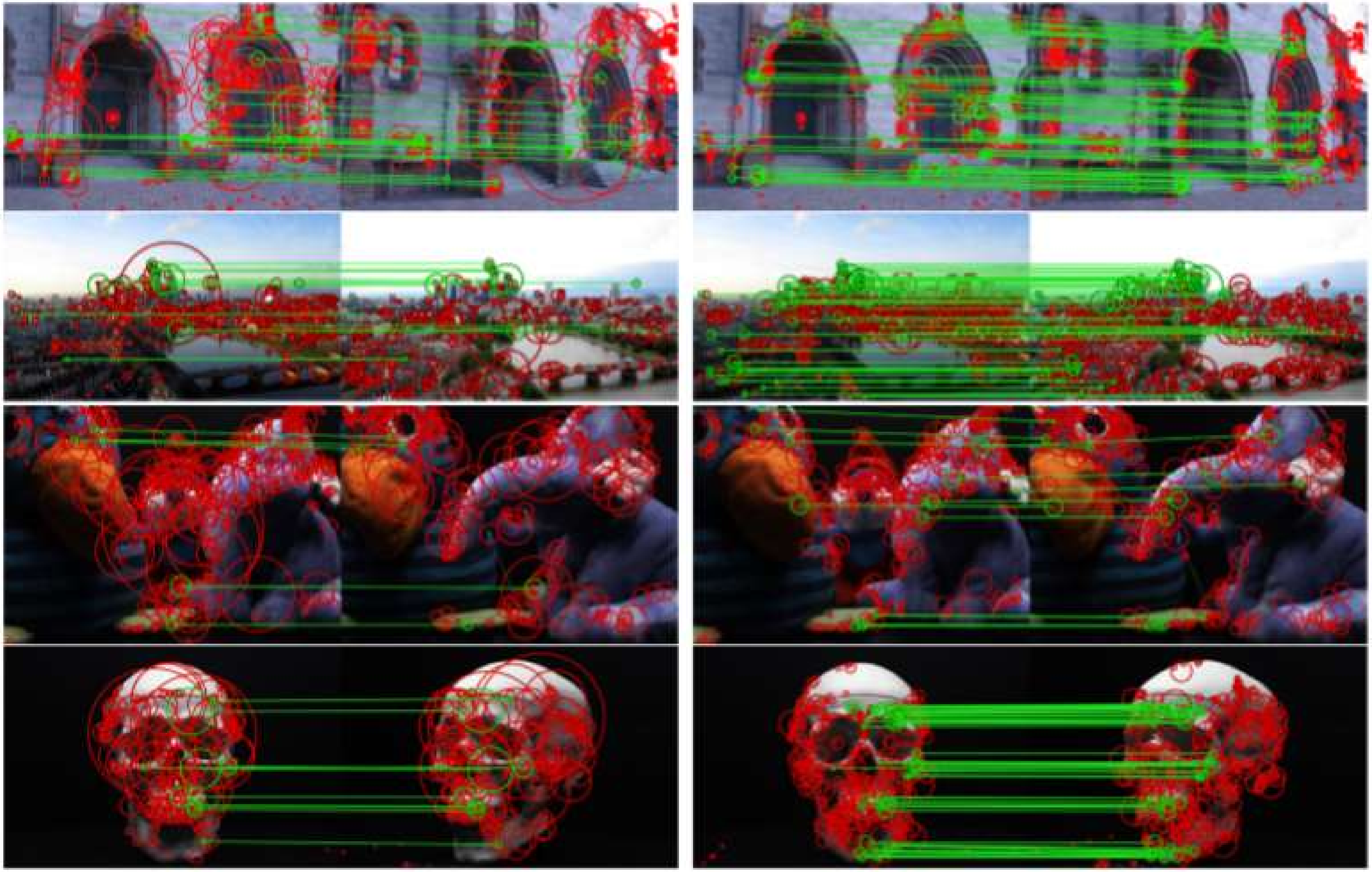


LIFT Loss has 3 components:

- Distance between descriptors of correponding patches, d^1 d^2, that should be *small*
- Distance between descriptors of different patches, d^1 d^3, that should be *large*
- Keypoints should not be located in homogeneous regions: P4 should not be detected as a keypoint

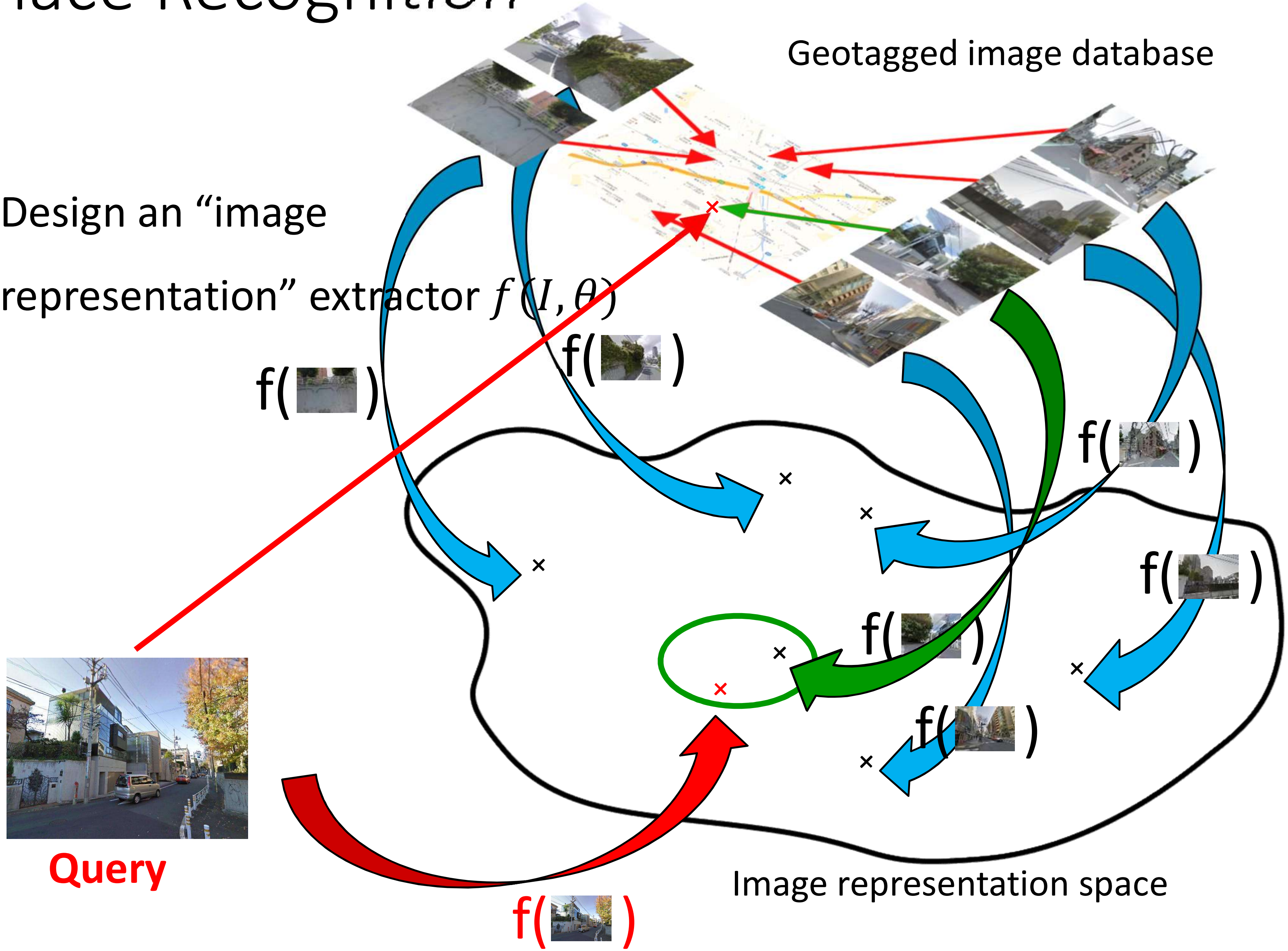
LIFT Results

- Works better than SIFT! (well, in some datasets)



Place Recognition

Design an "image representation" extractor $f(I, \theta)$



Place Recognition

The classical approach

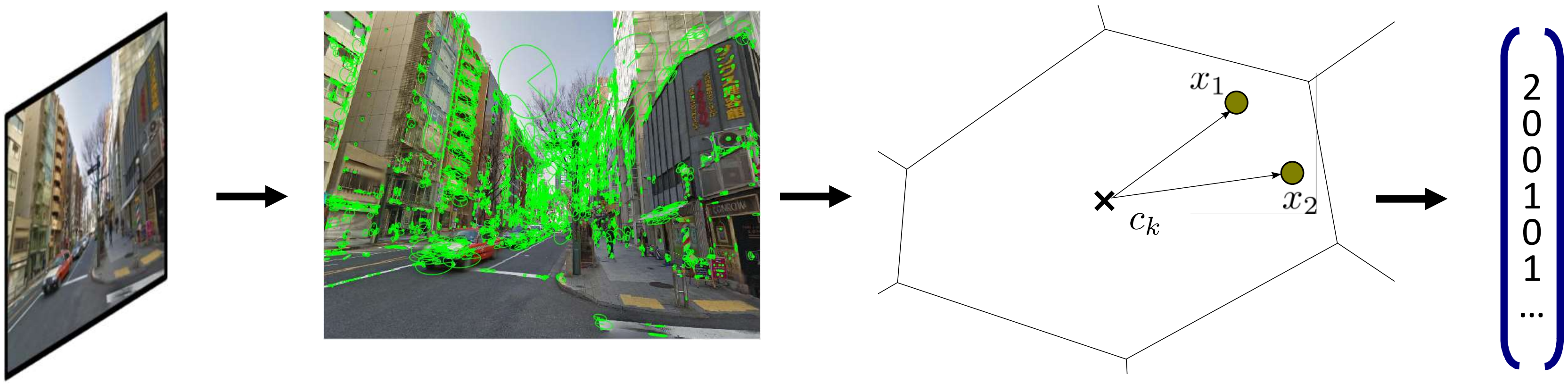


Image I

Extract local features (SIFT)

Aggregate (BoW, **VLAD**, FV)

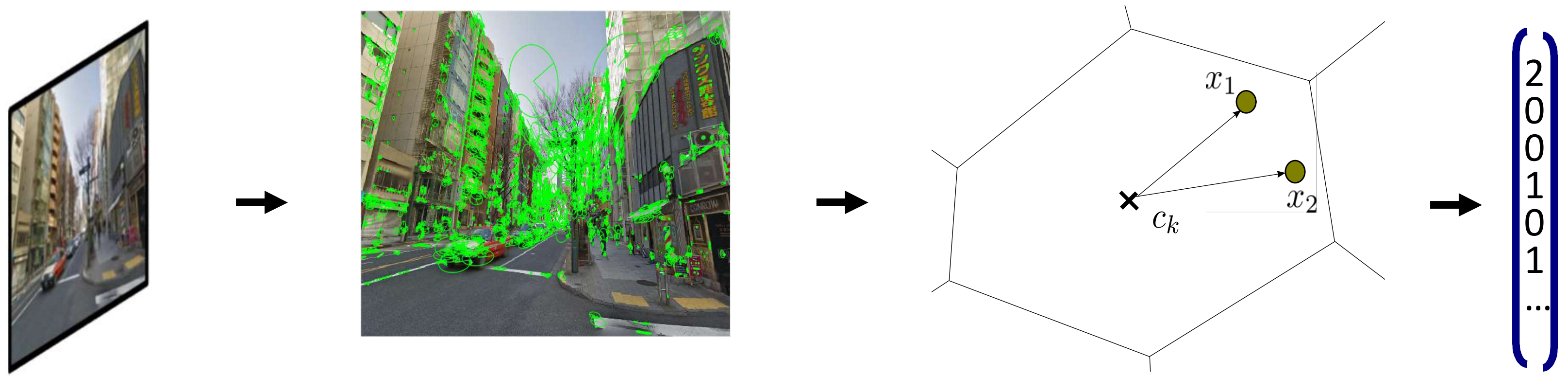
f(I)

1. Extract local descriptors from an image
2. Combine the extracted descriptors with Bag Of Words (BoW), VLAD or Fischer Vector(FV).
3. Produce a descriptor for the whole image.

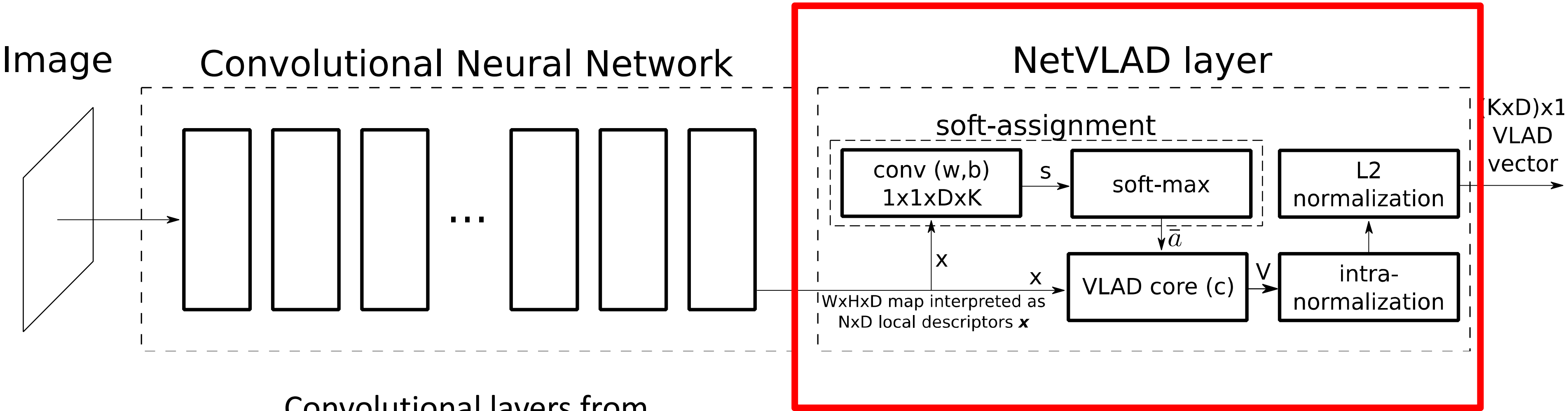
Slide adapted from NetVLAD presentation, CVPR 2017

NetVlad

Mimic the classical pipeline with deep learning



- Image I → Extract local features (SIFT) → Aggregate (BoW, **VLAD**, FV) → f(I)



Convolutional layers from AlexNet or VGGNet

Trainable pooling layer

Slide adapted from NetVLAD presentation, CVPR 2017

NetVlad Loss

- Triplet loss formulation

Matching samples

$D_p = ||F_\theta($  $) - F_\theta($  $)||^2$

$D_n = ||F_\theta($  $) - F_\theta($  $)||^2$

Non matching samples

$$L_\theta = \sum_{samples} \max(D_{p(\theta)} + \overbrace{m}^{\text{margin}} - D_{n(\theta)}, 0)$$

Disclaimer: The actual NetVlad loss is a slightly more complicated version of the one above

NetVlad Results

- Code, dataset and trained network online: give it a try!

  http://www.di.ens.fr/willow/research/netvlad/

Query

Top result

Green: Correct
Red: Incorrect

Slide adapted from NetVLAD presentation, CVPR 2017

Deep Learning for Pose estimation: PoseNet

Convolutional Network

Predict **camera position** $\boldsymbol{x}$ and **orientation** $\boldsymbol{q}$

PoseNet Loss

- Weighted mean square error loss:

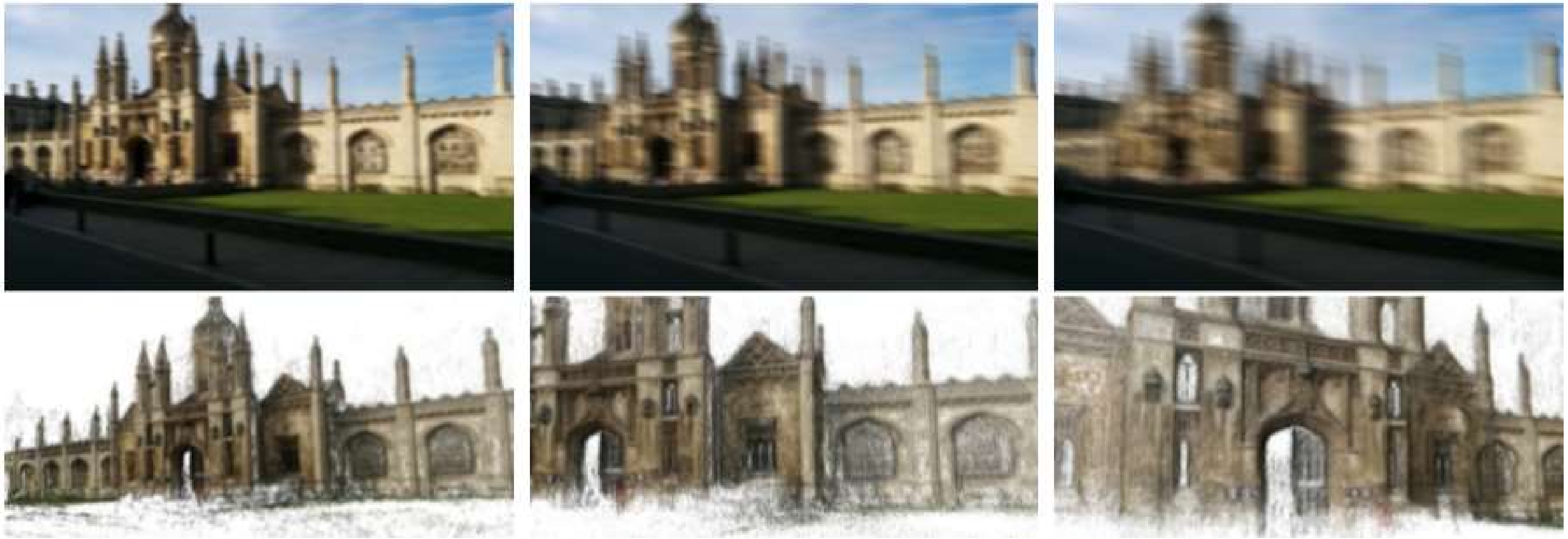
$$loss(I) = \|\hat{\mathbf{x}} - \mathbf{x}\|_2 + \beta \left\| \hat{\mathbf{q}} - \frac{\mathbf{q}}{\|\mathbf{q}\|} \right\|_2$$

- I, x, q represent the image, the ground truth position and orientation (in quaternions).

- $\widehat{x}$, $\widehat{q}$ are network pose and orientation prediction, respectively.

PoseNet Results

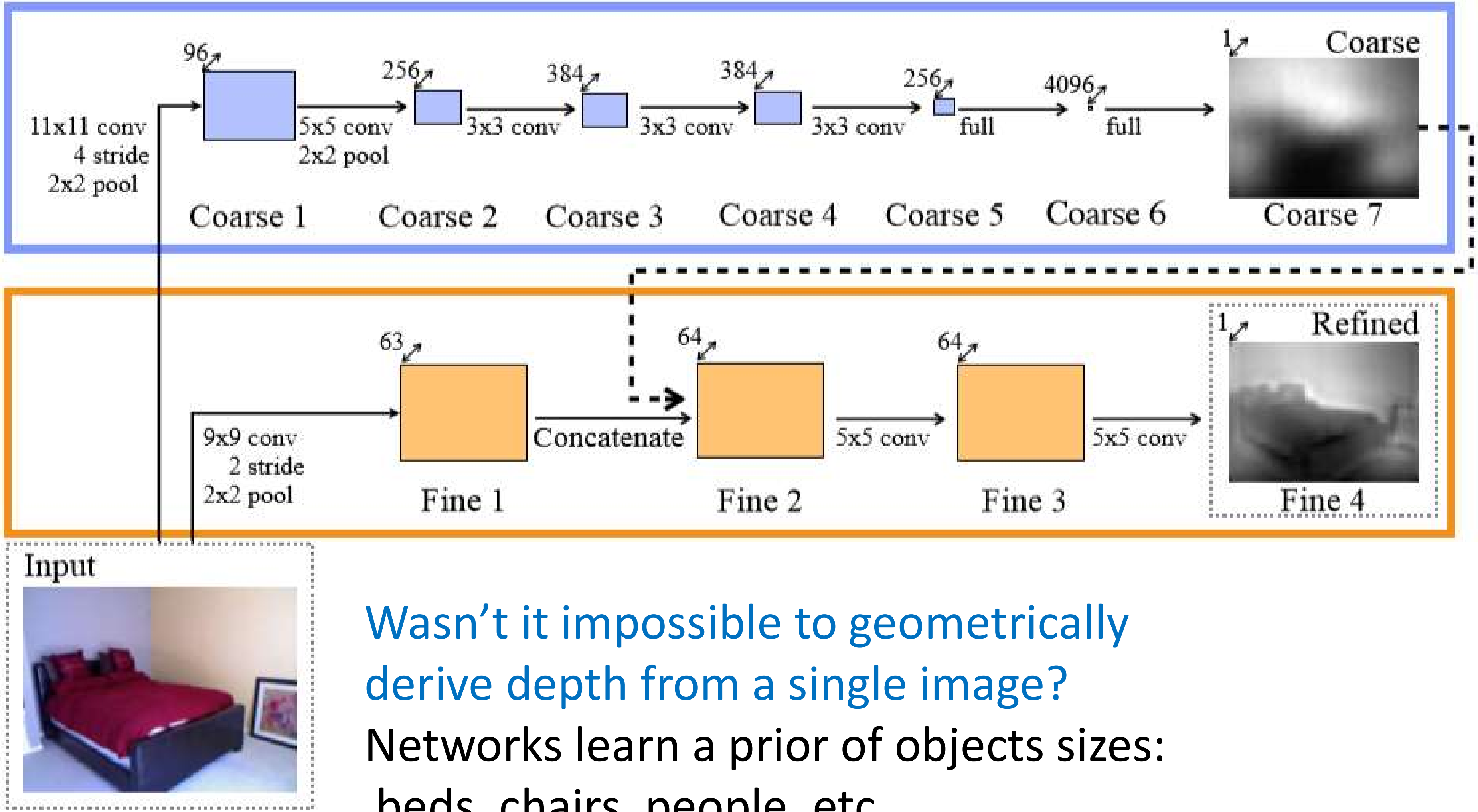
Motion Blur



Dynamic scenes

Uncalibrated camera

Monocular Depth Estimation



Wasn't it impossible to geometrically derive depth from a single image?

Networks learn a prior of objects sizes:
beds, chairs, people, etc...

Depth Map Prediction from a Single Image using a Multi-Scale Deep Network, Eigen et al. 2015

Deep Learning in Robotics: Learning to Act

Collective Robot Reinforcement Learning with Distributed Asynchronous Guided Policy Search – Yahya et al., 2016

Deep Learning in Robotics: Learning to Fly

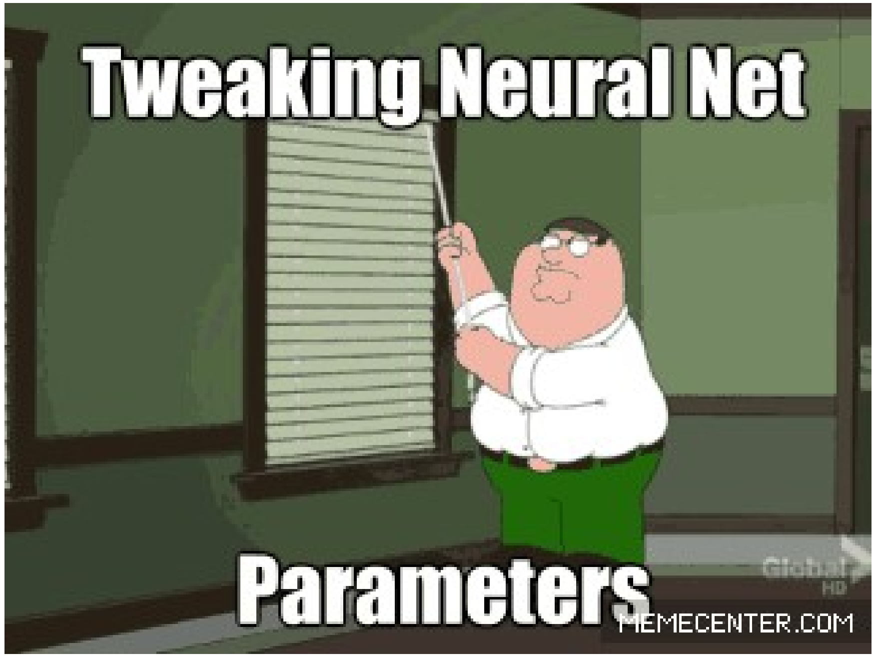
DroNet: Learning to Fly by Driving – Loquercio et al., 2017

And much much more...

Deep Learning Limitations

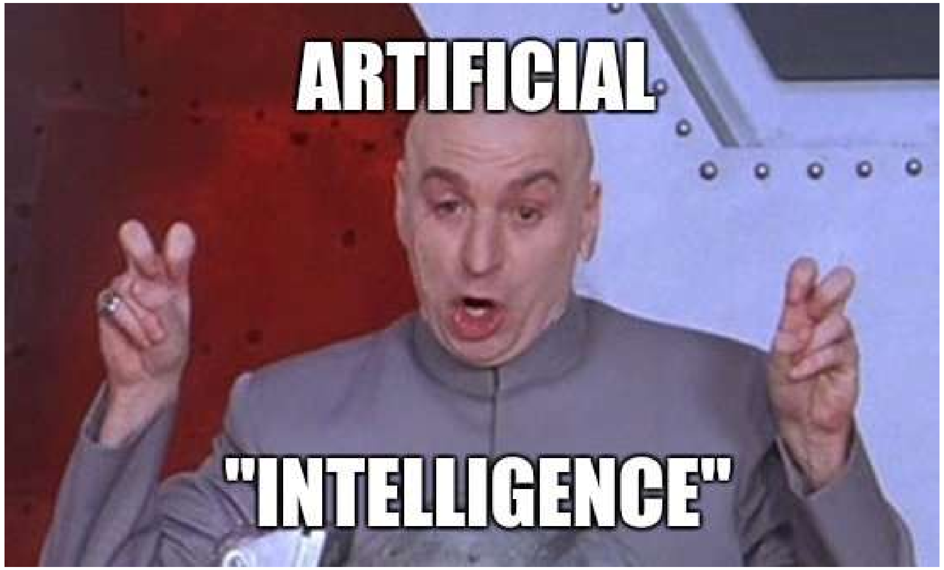
- Require **lots of data** to learn
- **Difficult debugging** and finetuning
- **Poor generalization** across similar tasks

Neural Networks Practitioners



Things to remember

- Deep Learning is able to **extract meaningful patterns** from data.
- It can be applied to **a wide range of tasks**.
- **Artificial Intelligence** ⊃ **Deep Learning**



Come over for projects in DL!

Visit our webpage for projects!

http://rpg.ifi.uzh.ch/student_projects.php

3rd Person View Imitation Learning - Available

Description: Manually programming robots to carry out specific tasks is a difficult and time consuming process. A possible solution to this problem is to use _imitation learning_, in which a robot aims to imitate a teacher, e.g., a human, that knows how to perform the task. Usually, the teacher and the learner share the same point of view on the problem. However, this last assumption might not be necessary. As humans, for example, we learn to cook by looking at others cooking. During this project, we will explore the possibility of repeating such a kind of 3rd person view _imitation learning_ with flying robots on a navigation task.

Goal: The project aims to develop machine learning based techniques that will enable a drone to learn flying by looking at an other robot flying.

Contact Details: "*Antonio Loquercio*": loquercio@ifi.uzh.ch

Thesis Type: Semester project / Bachelor Thesis / Master Thesis

See project on SIROP

Safe Reinforcement Learning for Robotics - Available

Description: Reinforcement Learning (RL) has recently emerged has a technique to let robots learn by their own experience. Current methods for RL are very data-intensive, and require a robot to fail many times before actually accomplishing their goal. However some systems, such as flying robots, require to respect safety constraints during learning and/or deployment. While maximizing performance, those methods usually aim to minimize the number of system failures and overall risk.

Goal: During this project, we will develop machine learning based techniques to let a (real) drone learn to fly nimbly through gaps and gates, while minimizing the risk of critical failures and collisions.

Contact Details: "*Antonio Loquercio*" loquercio@ifi.uzh.ch

Thesis Type: Semester project / Master Thesis

See project on SIROP

Simulation to Real World Transfer - Available

Description: Recent techniques based on machine learning enabled robotics system to perform many difficult tasks, such as manipulation or navigation. Those techniques are usually very data-intensive, and require simulators to generate enough training data. However, a system only trained in simulation (usually) fails when deployed in the real world. In this project, we will develop techniques to maximally transfer knowledge from simulation to the real world, and apply them to real robotics systems.

Goal: The project aims to develop techniques based on machine learning to have maximal knowledge transfer between simulated and real world on a navigation task.

Contact Details: "*Antonio Loquercio*" loquercio@ifi.uzh.ch

Thesis Type: Semester project / Bachelor Thesis / Master Thesis

See project on SIROP

Additional Readings

- Neural Networks and Deep Learning, by Michael Nielsen [Chapter 2]
- Practical Recommendations for Gradient-Based Training of Deep Architectures, Y. Bengio
- Deep Learning, Y. LeCun, Y. Bengio, G. Hinton
- All the references above!