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

Research Centers

Year

Source: Bloomberg
Some History

- **Perceptron**: 1958
- **Back-Propagation**: 1969
- **AI Winter**: 1974
- **SVM**: 1995
- **Convolutional Neural Networks for Handwritten Digits Recognition**: 1998
- **Restricted Boltzmann Machines**: 2006
- **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.
The semantic gap

- What computers see against what we see
Classification Challenges

Directly specifying how a category looks like is impossible.

We need use a **Data Driven Approach**
Supervised Learning

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

$N$ numbers representing class scores

Predicted: 

\[
\begin{pmatrix}
0.1 \\
0.7 \\
\vdots \\
0.0
\end{pmatrix}
\]

Ground truth, $y_i$: 

\[
\begin{pmatrix}
1.0 \\
0.0 \\
\vdots \\
0.0
\end{pmatrix}
\]

Loss: $\text{Loss}(f(x_i, \theta), y_i)$
Machine Learning Keywords

- **Loss**: Quantify how good $\theta$ are

- **Optimization**: The process of finding $\theta$ that minimize the loss

- **Function**: Problem modelling $\rightarrow$ 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

\[ f(x, \theta) = \text{label of the K training examples nearest to } x \]

- 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:

\[ f(x, \theta) = \text{sgn}(w \cdot x + b) \]

- What is now \( \theta \)? What is the dimensionality of images?
Classifiers: non-linear

- Good classifier
- Bad classifier (over fitting)

- What is $f(x, \theta)$?
Biological Inspiration

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

\[ f(x, \theta) = F(Wx), \text{ } F \text{ is a non-linear activation function (Step, ReLU, Sigmoid)} \]
Fully Connected Neural Networks

\[ f(x, \theta) \]

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

Neural Networks and Deep Learning, Chapter 2– Michael Nielsen
Forward propagation

Forward Pass

\[ \text{Loss}(f(x_i, \theta), y_i) \]

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

200x200x3 image

5x5x3 filter

Convolve (slide) over all spatial locations

Activation Map
Convolutional Layer

Repeat the operation for M filters

200x200x3 image

Convolve (slide) over all spatial locations

5x5x3 filter

New «image», 196x196x5

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

<table>
<thead>
<tr>
<th>Category</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car</td>
<td>62.7%</td>
</tr>
<tr>
<td>Dog</td>
<td>15.6%</td>
</tr>
<tr>
<td>Cup</td>
<td>11.5%</td>
</tr>
<tr>
<td>Flower</td>
<td>10.2%</td>
</tr>
</tbody>
</table>
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

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

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

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 has 3 components:

- Distance between descriptors of corresponding 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 $f(I, \theta)$

Geotagged image database

Query

Image representation space
Place Recognition
The classical approach

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

- Convolutional layers from AlexNet or VGGNet
- Trainable pooling layer
- Extract local features (SIFT)

Image I

\[ f(I) + \text{Aggregate (BoW, VLAD, FV)} \]

NetVLAD layer

Convolutional Neural Network

Image

Slide adapted from NetVLAD presentation, CVPR 2017
NetVlad Loss

• Triplet loss formulation

\[ D_p = \| F_\theta(\text{matching sample}) - F_\theta(\text{positive sample}) \|^2 \]

\[ D_n = \| F_\theta(\text{negative sample}) - F_\theta(\text{positive sample}) \|^2 \]

\[ L_\theta = \sum_{\text{samples}} \max(D_p(\theta) + m - D_n(\theta), 0) \]

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

Slide adapted from NetVLAD presentation, CVPR 2017
Deep Learning for Pose estimation: PoseNet

Convolutional Network

Predict camera position $x$ and orientation $q$
PoseNet Loss

- Weighted mean square error loss:

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

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

- \(\hat{x}, \hat{q}\) 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
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
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!