Deep Learning

by Wojciech Zaremba

Facebook AI Research
Machine learning successes

- Face recognition
- OCR
- Autonomous car
- Email classification
- Recommendation systems
- Web page ranking
What makes Machine Learning hard?
Hand-Crafted Features

SIFT/HOG

SURF

MFCC

Spectrogram
What is deep learning?

- End to end learning (no more feature engineering)
- Cascade of non-linear transformations
- General framework (any hierarchical model is deep)
Feed forward neural network

- Matrix multiplications, alternated with non-linearities
- Trained with Stochastic Gradient Descent (SGD)
Feed forward neural network

Example of a 2 hidden layer neural network (or 4 layer network, counting also input and output).
Forward propagation is the process of computing the output of the network given its input.
Forward propagation

The \( u = \max(0, v) \) non-linearity is called ReLU in the DL literature. Each output hidden unit takes as input all the units at the previous layer: each such layer is called “fully connected”.

\[
x \in \mathbb{R}^D \\
W^1 \in \mathbb{R}^{N_1 \times D} \\
b^1 \in \mathbb{R}^{N_1} \\
h^1 \in \mathbb{R}^{N_1}
\]

\[
h^1 = \max(0, W^1 x + b^1)
\]

- \( W^1 \): 1-st layer weight matrix or weights
- \( b^1 \): 1-st layer biases
Forward propagation

\[ h^1 \in \mathbb{R}^{N_1}, \quad W^2 \in \mathbb{R}^{N_2 \times N_1}, \quad b^2 \in \mathbb{R}^{N_2}, \quad h^2 \in \mathbb{R}^{N_2} \]

\[ h^2 = \max(0, W^2 h^1 + b^2) \]

- \( W^2 \): 2nd layer weight matrix or weights
- \( b^2 \): 2nd layer biases
Forward propagation

\[ h^2 \in \mathbb{R}^{N_2}, \quad W^3 \in \mathbb{R}^{N_3 \times N_2}, \quad b^3 \in \mathbb{R}^{N_3}, \quad o \in \mathbb{R}^{N_3} \]

\[ o = \max(0, W^3 h^2 + b^3) \]

- \( W^3 \): 3-rd layer weight matrix or weights
- \( b^3 \): 3-rd layer biases
Interpretation

Question: Why can't the mapping between layers be linear?
Answer: Because composition of linear functions is a linear function. Neural network would reduce to (1 layer) logistic regression.

Question: Why do we need many layers?
Answer: DL architectures are efficient also because they use distributed representations which are shared across classes.
Interpretation

- distributed representations
- feature sharing
- compositionality
Interpretation

**Question:** What does a hidden unit do?
**Answer:** It can be thought of as a classifier or feature detector.

**Question:** How many layers? How many hidden units?
**Answer:** Cross-validation or hyper-parameter search methods are the answer. In general, the wider and the deeper the network the more complicated the mapping.

**Question:** How do I set the weight matrices?
**Answer:** Weight matrices and biases are learned. First, we need to define a measure of quality of the current mapping. Then, we need to define a procedure to adjust the parameters.
How Good is a Network?

Probability of class $k$ given input (softmax):

$$p(c_k = 1 | x) = \frac{e^{o_k}}{\sum_{j=1}^{C} e^{o_j}}$$
Backpropagation

\[ x \rightarrow \max(0, W^1 x) \rightarrow h^1 \rightarrow \max(0, W^2 h^1) \rightarrow h^2 \rightarrow W^3 h^2 \rightarrow \frac{\partial L}{\partial o} \rightarrow \text{Loss} \]
Backpropagation
ForwardPass vs BackwardPass

FPROP and BPROP are dual of each other.
Fully-connected layers: issues

Example: 200x200 image
40K hidden units
~2B parameters

- Spatial correlation is local
- Waste of resources + we have not enough training samples anyway..
Locally connected layers

Example: 200x200 image
40K hidden units
Filter size: 10x10
4M parameters
Convolutional layers

STATIONARITY? Statistics is similar at different locations

Example: 200x200 image
40K hidden units
Filter size: 10x10
4M parameters
Convolutional layers

Share the same parameters across different locations (assuming input is stationary):

Convolutions with learned kernels
Convolution layer
Convolution layer
Convolution layer
Convolution layer
Convolution layer
Convolution layer
Convolution layer
Convolution layer
Convolution layer
Convolution layer
Convolution layer
Convolution layer
Convolution layer
Convolution layer
Convolution layer
Convolution layer
Convolutional neural networks

Assumptions:

● on the very small scale every piece of image should be processed the same way
ImageNet classification results 2012

1M training images, 1K categories, top-5 error

<table>
<thead>
<tr>
<th>Human performance</th>
<th>~3-5%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Deep-learning models</strong></td>
<td>~15%</td>
</tr>
<tr>
<td>Non-deep learning models</td>
<td>~26%</td>
</tr>
<tr>
<td>ISI, Japan</td>
<td></td>
</tr>
<tr>
<td>Oxford, England</td>
<td></td>
</tr>
<tr>
<td>INRIA, France</td>
<td></td>
</tr>
<tr>
<td>University of Amsterdam, etc.</td>
<td></td>
</tr>
</tbody>
</table>
# ImageNet classification results 2015

1M training images, 1K categories, top-5 error

<table>
<thead>
<tr>
<th>Human performance</th>
<th>~3-5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep-learning models</td>
<td>~4.5%</td>
</tr>
<tr>
<td>Non-deep learning models</td>
<td>~26%</td>
</tr>
<tr>
<td>ISI, Japan</td>
<td></td>
</tr>
<tr>
<td>Oxford, England</td>
<td></td>
</tr>
<tr>
<td>INRIA, France</td>
<td></td>
</tr>
<tr>
<td>University of Amsterdam, etc.</td>
<td></td>
</tr>
</tbody>
</table>
Imagenet results were just the beginning …
ConvNets for Image Segmentation

- MALIS
- Watershed + ConvNets + Trees
- Plenty of scope for improvement

- Clement Farabet, Camille Couprie, Srini Turaga
ConvNets + Graphical Model (Tompson et. al. 2014)
ConvNets for Video

C3D (Tran et. al.)

(a) 2D convolution
(b) 2D convolution on multiple frames
(c) 3D convolution

DeepVideo (Karpathy et. al.)

Conv1a 64 Pool
Conv2a 128 Pool
Conv3a 256 Pool
Conv3b 256 Pool
Conv4a 512 Pool
Conv4b 512 Pool
Conv5a 512 Pool
Conv5b 512 Pool
fc6 4096
fc7 4096
softmax
ConvNets for NLP

Collobert et. al. (2011)
My name is Wojciech.

(name) is

is Wojciech.
Representation

work by Ilya Sutskever
Representation

- I was given a card by her in the garden
- In the garden, she gave me a card
  - She gave me a card in the garden
- She was given a card by me in the garden
  - In the garden, I gave her a card
- I gave her a card in the garden
RNNs for visual attention

Mnih et. al. (DeepMind) 2014

Where is the ring? A: Mount-Doom
Where is Bilbo now? A: Grey-havens
Where is Frodo now? A: Shire

Weston et. al. 2014
Facebook AI Research
Trends

- Deeper nets
- Hybrid models. RNNs + Conv nets + Reinforcement Learning
- Multiple GPUs + Multiple Machines
- New kinds of memory units
- Algorithmical problems like sorting, addition
Q&A

- What is deep learning
- Feed forward neural networks
- Convolutional neural networks
- Recurrent neural networks
- Advances

I am happy to take any questions.
Challenges

- Scaling up for big data (videos, social networks etc.)
- Discrete optimization
- Memory that works