Deep Learning
Introduction to Machine Learning Algorithms

Ernir Erlingsson
PhD Student, University of Iceland & DEEP-EST Application Engineer

LECTURE 1

Deep Learning Introduction

June 6th, 2018
JSC, Germany
Outline of the Course

- 1. Introduction to Deep Learning
- 2. Fundamentals of Convolutional Neural Networks (CNNs)
- 5. Model Selection and Regularization
- 6. Fundamentals of Long Short-Term Memory (LSTM)
- 7. LSTM Applications and Challenges
- 8. Deep Reinforcement Learning
Deep Learning Introduction
Deep Learning - Introduction

Where do human ideas and innovations come from?

- Inspired by nature
  - First we **observe** and then we try to **replicate**
Neural Networks (NNs) is an attempt at replicating neural functions of the brain to solve problems.

Caveat: Not considered to be an accurate model but rather based on how neurons interconnect. Furthermore, we don’t know well enough how the brain operates to properly replicate it, e.g. what is a consciousness?
Artificial Neural Network (ANN)

- A computational model of biological learning
- Synonymous with deep learning
- The nodes simulate neurons and the edges simulate synapses with weight values.

- Neurons modelled as perceptron's that “fire” their activation function when the sum of weights crosses a certain threshold.
Deep Learning – Definition

- **Artificial Intelligence**
  - The concept of machines being able to carry out tasks in a way we would consider intelligent

- **Machine Learning**
  - Computer systems that improve with experience and data

- **Deep Learning**
  - Is a subset of machine learning where the system is represented as nested hierarchical features, where each feature is defined in relation to simpler features.
Deep Learning – Introduction

- Deep learning is a brand that comes in many ever-increasing flavours. It is also known as:
  - Cybernetics (1940s)
  - Neural networks (1980s)
  - Deep learning (2006)

- Has gained traction very fast with no immediate signs of slowing down and is sometimes characterized as a buzzword

- It is used for supervised, semi-supervised and unsupervised learning.
  - *Supervised learning uses labelled data*
  - *Semi-supervised uses mostly unlabelled data but not all.*
  - *Unsupervised learning uses only unlabelled data*
Deep Learning – What is it

Deep Learning – Introduction

- Application areas
  - Computer vision
  - Automatic speech recognition
  - Natural language processing
  - Bioinformatics
  - And much more...

- It’s effects is currently is both overestimated and underestimated.
Deep Learning Risks

- Very easy to fool, e.g. via adversarial patches

Neural Networks - Timeline

- 1943: The first mathematical model of the human brain
- 1957: Perceptron
- 1965: The first multi-layered network
- 1987: Multi-layered Perceptron (backpropagation)
- 1995: Support Vector Machines (SVMs)
- 1998: Gradient based learning
- 2006: Deep Neural Network
- 2011: AlexNet (CNNs)
- 2014: Generative Adversarial Networks (GANs)
Deep Learning - Introduction

1950s Cybernetics: Cyber the dog

[3] ‘Cybernetic Zoo’ web page
Neural Networks - Resurgence

The renaissance of neural networks via deep learning, accelerated by:

- **Big Data**
  - The first web page 1992
  - 163 zettabytes (1 million petabytes) by 2025
Big Data – Internet Users

Internet users by world region

Source: Science and Technology - World Bank (2016)

[4] Our world in data, Online
Big Data - ImageNet Dataset

- **Dataset:** ImageNet
  - Total number of images: *14,197,122*

<table>
<thead>
<tr>
<th>High level category</th>
<th># synset (subcategories)</th>
<th>Avg # images per synset</th>
<th>Total # images</th>
</tr>
</thead>
<tbody>
<tr>
<td>amphibian</td>
<td>94</td>
<td>591</td>
<td>56K</td>
</tr>
<tr>
<td>animal</td>
<td>3822</td>
<td>732</td>
<td>2799K</td>
</tr>
<tr>
<td>appliance</td>
<td>51</td>
<td>1164</td>
<td>59K</td>
</tr>
<tr>
<td>bird</td>
<td>856</td>
<td>949</td>
<td>812K</td>
</tr>
<tr>
<td>covering</td>
<td>946</td>
<td>819</td>
<td>774K</td>
</tr>
<tr>
<td>device</td>
<td>2385</td>
<td>675</td>
<td>1610K</td>
</tr>
<tr>
<td>fabric</td>
<td>262</td>
<td>690</td>
<td>181K</td>
</tr>
</tbody>
</table>

*[5] ImageNet Web page*
Big Data - ImageNet Dataset

The renaissance of neural networks via deep learning, accelerated by:

- **Big Data**
  - The first web page 1992
  - 163 zettabytes (1 million petabytes) by 2025

- **Advances in Computation**
  - Multi-core CPUs
  - Many-core GPUs
Advances in Computation - Multi-core CPUs

- Significant advances in CPU (or microprocessor chips)
  - Multi-core architecture with dual, quad, six, or n processing cores
  - Processing cores are all on one chip
- Multi-core CPU chip architecture
  - Hierarchy of caches (on/off chip)
  - L1 cache is private to each core; on-chip
  - L2 cache is shared; on-chip
  - L3 cache or Dynamic random access memory (DRAM); off-chip

Advances in Computation - Multi-core CPUs

- Clock-rate for single processors increased from 10 MHz (Intel 286) years to 4 GHz (Pentium 4) in 30 years

- Clock rate increase with higher 5 GHz reached a limit due to power limitations / heat

- Multi-core CPU chips have quad, six, or n processing cores on one chip and use cache hierarchies

Advances in Computation - GPGPUs

- The graphics Processing Unit (GPU) is repurposed as General-Purpose GPUs (GPGPUs) and used for computing.

- Slower than CPUs but more than makes up for it with sheer volume, i.e. consists of *very many* simple cores
  - High throughput computing-oriented architecture
  - Use massive parallelism by executing a lot of concurrent threads
  - Handle an ever increasing amount of multiple instruction threads
  - CPUs instead typically execute a single long thread as fast as possible

- Simplicity leads to less power consumption

- Many-core GPUs are already used in large clusters and within massively parallel supercomputers

Advances in Computation - GPGPUs

- GPUs accelerate computing thru massive parallelism, with thousands of threads.
- GPUs are designed to compute a large number of floating point operations in parallel.
- GPU accelerator architecture example - NVIDIA card
  - GPUs can have 256 cores on one single GPU chip (NVIDIA TEGRA X1)
  - Each core can work with eight threads of instructions
  - GPU is able to concurrently execute 256 * 8 = 2048 threads
- Interaction and thus major (bandwidth) bottleneck between CPU and GPU is via memory interactions
- E.g. applications that use matrix – vector multiplication


(other well known accelerators & many-core processors are e.g. Intel Xeon Phi → run ‘CPU’ applications easier)
[Video] GPGPUs & Applications

Art, Science and GPU's
Adam Savage & Jamie Hyneman
Explain Parallel Processing

[8] Mythbusters Demo GPU versus CPU
Neural Networks - Resurgence

The renaissance of neural networks via deep learning, accelerated by:

- **Big Data**
  - The first web page 1992
  - 163 zettabytes (1 million petabytes) by 2025

- **Advances in Computation**
  - Multi-core CPUs
  - Many-core GPUs

- **Improved architecture and techniques**
  - Convolutional Neural Networks (CNNs)
  - Recurrent Neural Networks (RNNs)
  - Generative Adversarial Networks (GANs)
The MNIST dataset

An image collection of hand-written digits available online, with:

- 60.000 training examples
- 10.000 testing examples

The dataset is considered to be the fruit-fly of machine-learning.

- Best classification accuracy to-date, with a 0.23% error-rate, through the use of Convolutional Neural Networks
- Support Vector Machines (SVMs) also performing well, with the lowest error-rate of 0.56%

[9] The MNIST Database
Deep Learning Architectures

- Deep Neural Network (DNN)
  - ‘Shallow ANN‘ approach with many hidden layers between input/output
Deep Neural Networks (DNNs)

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville ‘Deep Learning’
DNN – Feature Learning Benefits

- Traditional machine learning applied feature engineering before modeling
- Feature engineering requires expert knowledge, is time-consuming and a often long manual process, requires often 90% of the time in applications, and is sometimes even problem-specific
- Deep Learning enables feature learning promising a massive time advancement

Deep Learning Architectures

- **Deep Neural Network (DNN)**
  - ‘Shallow ANN‘ approach with many hidden layers between input/output

- **Convolutional Neural Network (CNN, sometimes ConvNet)**
  - Connectivity pattern between neurons inspired by the visual cortex
Convolutional Neural Networks

- Inspired by connectivity patterns between neurons in the animal visual cortex.
- Usually built with three types of layers:
  - Convolution layers using kernels
  - Pooling layers (downsampling)
  - Fully connected layers (classification vote)

[12] Azoft, Fully convolutional Neural Network
Deep Learning Architectures

- Deep Neural Network (DNN)
  - ‘Shallow ANN‘ approach with many hidden layers between input/output

- Convolutional Neural Networks (CNNs, sometimes ConvNets)
  - Connectivity pattern between neurons replicating the visual cortex

- Recurrent Neural Network (RNN)
  - ‘ANN‘ but connections form a directed cycle; state and temporal behaviour
Recurrent Neural Networks (RNN)

- A recurrent neural network can be thought of as multiple copies of the same network, each passing a message to a successor.
- RNNs have been applied very successfully to a variety of problems, e.g. speech recognition, language modeling, translation, image captioning.
- Essential to these successes is the use of “LSTMs”, a very special kind of recurrent neural network.

[13] Colah’s blog, Understanding LSTM Networks
Deep Learning Architectures

- Deep Neural Network (DNN)
  - ‘Shallow ANN‘ approach with many hidden layers between input/output

- Convolutional Neural Network (CNN, sometimes ConvNet)
  - Connectivity pattern between neurons replicating the visual cortex

- Recurrent Neural Network (RNN)
  - ‘ANN‘ but connections form a directed cycle; state and temporal behaviour

- Deep Reinforcement Learning (DRN)
  - Unsupervised goal-oriented algorithms
Deep Reinforcement Learning

- Reinforcement learning refers to goal-oriented algorithms, which learn how to attain a complex objective (goal) or maximize along a particular dimension over many steps; for example, maximize the points won in a game over many moves.

\[
Q(s_t, a_t) \leftarrow (1 - \alpha) \cdot Q(s_t, a_t) + \alpha \cdot \left( r_t + \gamma \cdot \max_a Q(s_{t+1}, a) \right) - q \cdot \text{learned value}
\]

Convolutional Agent

[14] DL4J, A Beginner’s guide to deep reinforcement learning
Deep Q Learning (DQN)
Generative Adversarial Networks (GANs)

Big improvements in the quality of the produced image output from GANs in relatively few years.
- Has surpassed the “uncanny valley” obstacle

Generative Adversial Networks (GANs)

- **Discriminative algorithms**
  - Predict a label or category for an input given the feature set

- **Generative Adversial Network (GAN)**
  - Predict the features given a label

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville ‘Deep Learning’
Generative Adversarial Networks (GANs)

[17] Progressive Growing of GANs for Improved Quality, Stability, and Variation
Capsule Networks

- Orientation and relative spatial relationships are captured well by CNNs

Capsule Networks are designed to address this problems

[18] Capsule Networks
Neural Networks - Resurgence

The renaissance of neural networks via deep learning, accelerated by:

- **Big Data**
  - The first web page 1992
  - 163 zettabytes (1 million petabytes) by 2025

- **Advances in Computation**
  - Multi-core CPUs
  - Many-core GPUs

- **Improved architecture and techniques**
  - Convolutional Neural Networks (CNNs)
  - Recurrent Neural Networks (RNNs)
  - Generative Adversarial Networks (GANs)

- **Better Tools**
  - Numpy, Tensorflow, Keras, Scikit-learn, Jupyter... and many more
Deep Learning – Tools of the Trade

- TensorFlow
- Caffe
- Keras
- Theano
- Caffe2
- PyTorch
The top 5 mentions on arXiv.org, many more exist
- Free and Open Source (FOSS) frameworks, libraries and extensions
- Mostly used with Python, a major contributor to its growing popularity.
- Initiated and maintained by a mixture of both academia and industry
Deep Learning – Tools of the Trade

Monthly ArXiv.org mentions (10-day average), 2018/01/12

- TensorFlow: 273
- Keras: 100
- Caffe: 94
- PyTorch: 72
- Theano: 53
Keras with Tensorflow Backend – GPU Support

- Keras is a high-level deep learning library implemented in Python that works on top of existing other rather low-level deep learning frameworks like Tensorflow, CNTK, or Theano.
- The key idea behind the Keras tool is to enable faster experimentation with deep networks.
- Created deep learning models run seamlessly on CPU and GPU via low-level frameworks.


- Tensorflow is an open source library for deep learning models using a flow graph approach.
- Tensorflow nodes model mathematical operations and graph edges between the nodes are so-called tensors (also known as multi-dimensional arrays).
- The Tensorflow tool supports the use of CPUs and GPUs (much more faster than CPUs).
- Tensorflow work with the high-level deep learning tool Keras in order to create models fast.


[20] A Tour of Tensorflow
Deep Neural Networks – How they work
Multi Layer Perceptrons – Artificial Neural Networks

- **Key Building Block**
  - Perceptron learning model
  - Simplest linear learning model
  - Linearity in learned weights $w_i$
  - One decision boundary

- **Artificial Neural Networks (ANNs)**
  - Creating more complex structures
  - Enable the modelling of more complex relationships in the datasets
  - May contain several intermediary layers
  - E.g. 2-4 hidden layers with hidden nodes
  - Use of activation function that can produce output values that are nonlinear in their input parameters
Gradient Descent Method (1)

\[ b = a - \gamma \nabla f(a) \]

(minimization: subtract gradient term because we move towards local minima)

(b new position)

(position a (current position))

(position b (next position))

(position x_1, x_2)

(position x_1, x_2, x_3)

(position decreasing values)

(position increasing values)

(position stationary)

(position negative gradient)

(position positive gradient)

(position finding this point x is the goal of gradient descent)

(position zero gradient)
Gradient Descent Method (2)

- Gradient Descent (GD) uses all the training samples available for a step within a iteration
- Stochastic Gradient Descent (SGD) converges faster: only one training samples used per iteration

\[
b = a - \gamma \nabla f(a) \quad \Rightarrow \quad b = a - \gamma \frac{\partial}{\partial a} f(a) \quad \Rightarrow \quad b = a - \gamma \frac{d}{da} f(a)
\]

(all slightly different notations, but often used in different literature for same derivative term)

\[
x_{1\text{next}} = x_1 - \gamma \frac{d}{dx_1} f(x_1) \quad \text{(negative derivative at point } x_1) \quad x_{2\text{next}} = x_2 - \gamma \frac{d}{dx_2} f(x_2) \quad \text{(positive derivative at point } x_2)
\]

[21] Big Data Tips, Gradient Descent
Loss Functions and Backpropagation

- Gradient Descent is mostly used in combination with a Loss function, which is used for model evaluation.

\[
MSE_{\text{test}} = \frac{1}{m} \sum_{i} (\hat{y}^{(\text{test})} - y^{(\text{test})})^2.
\]

- Backpropagation is employed to update the weights and bias of every edge using the chain-rule from calculus, i.e. taking the derivative to determine the modification gradient.

[22] Becoming Human, Backpropagation
Activation functions

- Multiple activation function types are used in practice, and often several different types in a single neural network.
- Their selection is quintessential component to achieve non-linear convergence.
ANN – Backpropagation Algorithm (BP) Basics

- One of the **most widely used** algorithms for supervised learning
  - Applicable in **multi-layered feed-forward neural networks**

- ‘Gradient descent method’ can be used to learn the weights of the output and hidden nodes of a artificial neural network

- **Hidden nodes problem:** computing error term hard: $\frac{\partial E}{\partial w_j}$
- **Their Output values are unknown to us** (here)...

- **The backpropagation algorithm solves exactly this problem with two phases per iteration(!)**

---

Lecture 1: Deep Learning Introduction
Deep Learning Material

Books freely available online

- [http://www.deeplearningbook.org/](http://www.deeplearningbook.org/)
- [http://www-bcf.usc.edu/~gareth/ISL/](http://www-bcf.usc.edu/~gareth/ISL/)

Coursera

[https://www.coursera.org/specializations/deep-learning](https://www.coursera.org/specializations/deep-learning)

Stanford lecture collection

[https://www.youtube.com/watch?v=vT1JzLTH4G4&list=PL3FW7Lu3i5JvHM81yjYj-zLfQRF3EO8sYv](https://www.youtube.com/watch?v=vT1JzLTH4G4&list=PL3FW7Lu3i5JvHM81yjYj-zLfQRF3EO8sYv)
Lecture Bibliography (1)

3. ‘Our World in Data’ web page, Online: https://ourworldindata.org/internet
4. ImageNet Web page, Online: http://image-net.org
7. Mythbusters Demo GPU versus CPU Online: https://www.youtube.com/watch?v=-P28LKWtZrI
8. The MNIST Database http://yann.lecun.com/exdb/mnist/
10. YouTube Video, ‘Neural Networks, A Simple Explanation’, Online: http://www.youtube.com/watch?v=gcK_5x2KsLA
Lecture Bibliography (2)

- [13] ‘Understanding LSTMs’
  Online: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

  Online: https://deeplearning4j.org/deepreinforcementlearning

- [15] DQN
  https://www.youtube.com/watch?v=TmPfTpjtdgg

  https://deeplearning4j.org/generative-adversarial-network

  Online: https://www.youtube.com/watch?v=G06dEcZ-QTg

- [18] Capsule Networks

- [19] Tensorflow Deep Learning Framework,
  Online: https://www.tensorflow.org/

- [20] A Tour of Tensorflow,

  Online: http://www.big-data.tips/gradient-descent

- [22] Back Propagation in neural-networks intuition,
Slides Available at http://www.morrisriedel.de/talks