01 Fundamentals: from Neural Networks to Deep Learning

- Deep Learning := ML method based on learning representations of data. An observation (e.g., an image) can be represented in many ways such as a vector of intensity values per pixel, or in a more abstract way as a set of edges, regions of particular shape, etc.
- Feature := specific measurable property of a phenomenon being observed.
- Feature engineering := using domain knowledge to create features useful for ML. ("Applied ML is basically feature engineering. Andrew Ng").
- Feature learning := transformation of raw data input to a representation, which can be effectively exploited in ML.
Perform a num. optimization to find optimal stimuli

\[ x^* = \arg \min_x f(x; W, H), \quad \text{subject to } \|x\|_2 = 1. \]

Cerebral cortex interconnections Hubel & Wiesel (1962)

Principles of human information processing...

1 Gb/sec datastream
10 Mp camera (always on)
20 watts

Image credit to Bruno Olshausen, Redwood Center for Theoretical Neuroscience, UC Berkeley.

Deep Learning 1998

15 years later ... a kind of Cambrian explosion

2015 ...

http://www.nature.com/news/what-sparked-the-cambrian-explosion-1.19139

Convolutional Neural Network in computer vision

\[ U(f) = (C_W(1) \circ C_W(2) \circ C_W(3))(f) \]

Limitations of Deep Learning approaches

- Computational resource intensive (supercomputers, cloud CPUs, federated learning, ...)
- Data intensive (needs often millions of training samples – "big data" is necessary!)
- Black-Box approaches – lack transparency, do not foster trust and acceptance among end-users, however, legal aspects make it difficult!
- Non-convex: difficult to set up, to train, to optimize, needs a lot of expertise, error prone
- Most of all: bad in dealing with uncertainty ...
03 From Bayesian Networks to Gaussian Processes

04 Stochastic Gradient Descent

05 Deep Autoencoders (unsupervised NN)

Minimizing an objective function

\[ f(\theta) = E[f(\theta; x)] \]

Algorithm 8.2: Stochastic gradient descent

1. Initialize \( \theta_0 \); \( g \)
2. repeat
3. randomly permute data
4. for \( i = 1 : N \) do
5. \( g = \nabla f(\theta; x_i) \)
6. \( \theta = \theta - \alpha g(x_i) \)
7. \( \theta \) until converged


General Structure of an Autoencoder

Pretraining a stack of restricted Boltzmann Machines

What makes a good representation? Trad. Autoencoders

- Encoder: Det. mapping \( f_\theta \) that transforms an input vector \( x \) into a representation \( y \)

\[ f_\theta(x) = s(Wx + b) \]

- Decoder: Resulting hidden representation \( y \) is then mapped back to a reconstructed d-dimensional vector \( z \) in input space, \( z = g_\phi'(y) \). This mapping \( g_\phi' \) is called the decoder

\[ g_\phi'(y) = s(W'\phi + b') \]


07 Future Challenges and Extravaganza *) Topics

*) Modeling, A. 2014. Extravaganzas on Hot Ideas for Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. In: Skalts, D., Bach, A. H., Peters, J. F., Schneider, L. (Eds.) Bioinformatics and Biomedicine, ISMB-BIBS 2014, Lecture Notes in Computer Science, LNBI (Fig. 13b, 1.5). doi:10.1007/978-3-319-03986-3_14.

Catastrophic Forgetting – cutting edge research

- When trained on one task, then trained on a 2nd task, many machine learning models (“deep learning”) forget how to perform the first task.

Overcoming catastrophic forgetting in neural networks

Kirkpatrick et al. (2017) demonstrate that task-specific synaptic consolidation offers a unique solution to the continual-learning problem for artificial intelligence.

Developed an algorithm analogous to synaptic consolidation for artificial neural networks,

Elastic Weight Consolidation (EWC).

This algorithm slows down learning on certain weights based on how important they are to previously seen tasks.

They show how EWC can be used in supervised learning and reinforcement learning problems to train several tasks sequentially without forgetting older ones, in marked contrast to previous deep-learning techniques.

Elastic Weight Consolidation (EWC)


"How do humans generalize from very few examples?"
- They transfer knowledge from previous learning:
  - Representation learning (features!)
  - Explanatory factors
  - Previous learning from unlabeled data and labels for other tasks
- Prior: shared underlying explanatory factors, in particular between $P(x)$ and $P(y|x)$, with a causal link between $y \rightarrow x$

Learning from Deep vs. Flat Feature Representations

The grand question of cognitive science

Results on MNIST Data Sets

Machine Learning inspires Human Learning and v.v.

Trends in Cognitive Sciences

The ability to learn in a sequential fashion is crucial for the development of artificial intelligence. Until now, neural networks have learned to solve one task after another. However, this approach suffers from catastrophic forgetting, where the network forgets previous knowledge when learning new tasks.

Elastic Weight Consolidation (EWC)


Overcoming catastrophic forgetting in neural networks.


This experiment was done with Games
Open Problem: How to avoid negative transfer?