Topics in Natural Language Processing

Deep Learning for NLP

Shashi Narayan
ILCC, School of Informatics
University of Edinburgh

Overview

What is Deep Learning?

Why do we need to study deep learning?

Deep Learning: Basics

Deep Learning in Application

Neural Networks and Deep Learning

Standard **machine learning** relies on **human-designed** representations and input features

Then, machine learning algorithms aims at **optimizing model weights** to best make a final prediction

Neural Networks and Deep Learning

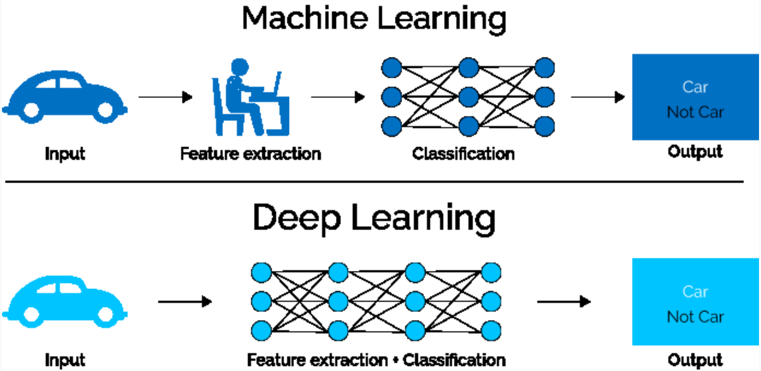


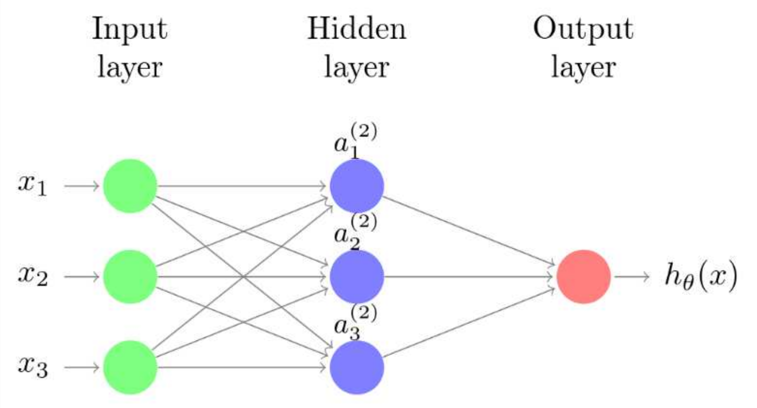
Image: https://content-static.upwork.com/blog/uploads/sites/3/2017/06/27095812/image-16.png

Neural Networks and Deep Learning

Representation learning automatically discovers good features or representations needed from the data

Deep learning algorithms learn multiple levels of representation of increasing complexity or abstraction

Neural Networks and Deep Learning



Why do we need to study deep learning?

Representation Learning

Human-designed representations and input features are:

- task dependent;
- time-consuming and expensive; and
- often under or over specified.

Deep learning provides a way to do **Representation Learning**

Distributed and Continuous Representation

Traditional NLP systems are incredibly fragile due to their symbolic representations

Distributed and Continuous Representation

Document Classification

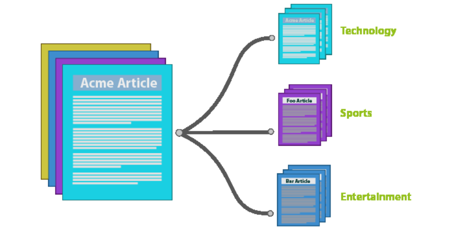


Image: https://media.licdn.com/mpr/mpr/shrinknp_800_800/p/8/005/0a3/00e/1488735.png

Distributed and Continuous Representation

Document Classification

$$p(c_i) = f(\text{bag of unigrams, bigrams, ...})$$

Document Classification

$$p(c_i) = f(\text{bag of unigrams, bigrams, ...})$$

Curse of dimensionality

No notion of semantic similarity

- US $\neq$ USA
- (Cricket -> Sports) $\neq$ (Football -> Sports)

Deep learning provides a way to **use and learn continuous word representations**

$$\text{word}_i = [0.11, 0.22, 0.21, ..., 0.52, 0.19]_{256}$$

It solves the curse of dimensionality

It also introduces a notion of semantic similarity

It allows unsupervised feature and weight learning

Distributed and Continuous Representation

Distributional Similarity

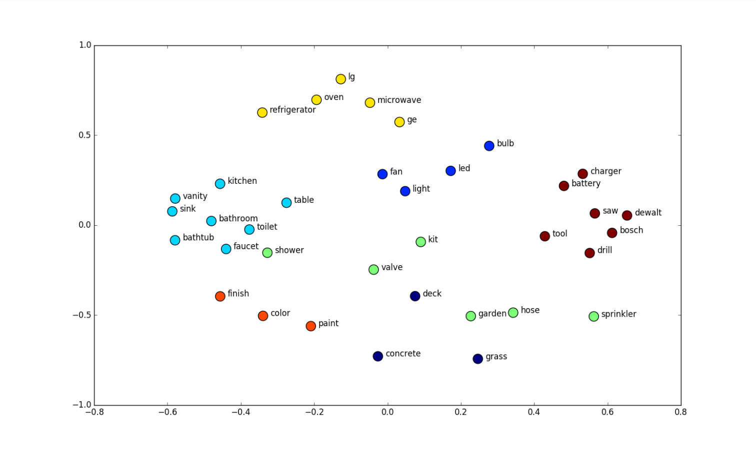


Image: http://5047-presscdn.pagely.netdna-cdn.com/wp-content/uploads/2016/05/word-embeddings.png

Distributed and Continuous Representation

Distributional Similarity

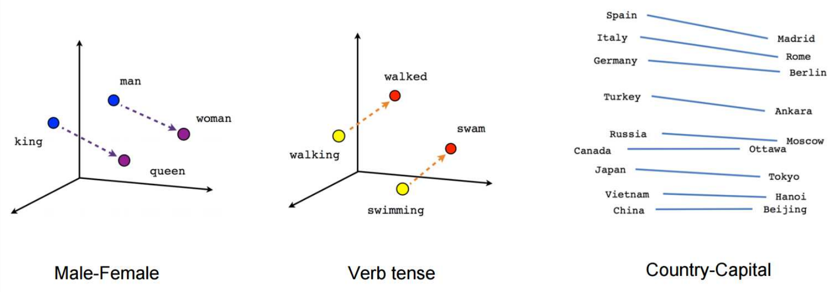


Image: https://www.tensorflow.org/images/linear-relationships.png

Hierarchical Representation

Deep learning allows **multiple levels of hierarchical representation** of increasing complexity or abstraction

Hierarchical Representation

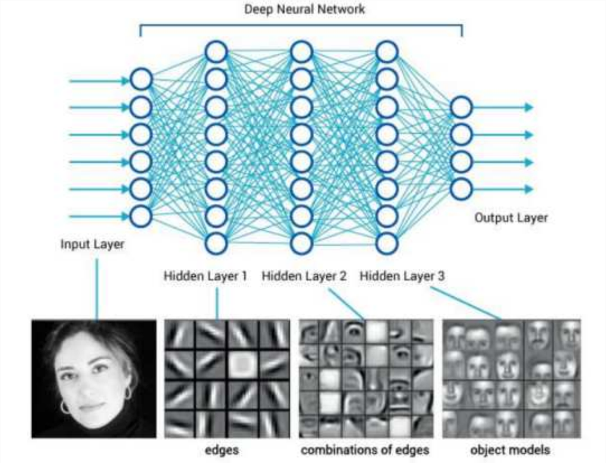


Image: https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/assets/DeepConcept.png

Deep learning allows **multiple levels of hierarchical representation** of increasing complexity or abstraction

Compositionality in Natural Language: e.g., sentences are composed from words and phrases.

Deep Learning is establishing the state of the art!

Computer Vision: e.g., Image recognition

Natural Language Processing: e.g., Language Modelling, Neural Machine Translations, Dialogue Generation and Natural Language Understanding

Speech Processing: e.g., Speech recognition

Retail, Marketing, Healthcare, Finance, **...**

Deep Learning: But Why Now?

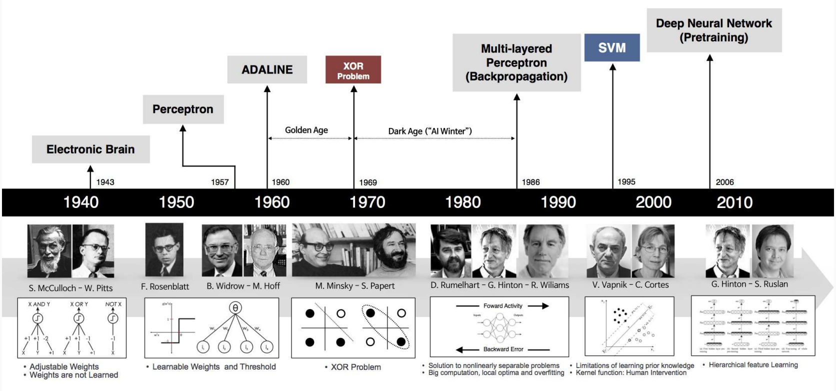


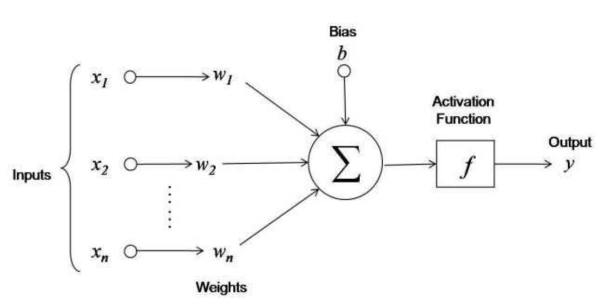
Image: http://beamandrew.github.io//images/deep_learning_101/nn_timeline.jpg

Deep Learning: But Why Now?

- Availability of **large-scale high-quality labeled datasets**
- Availability of **faster machines**: Parallel computing with GPUs and multi-core CPUs
- Better understanding of **regularization techniques** - Dropout, batch normalization, and data-augmentation
- Availability of **open-source machine learning frameworks**: Tensorflow, Theano, Dynet, Torch and PyTorch
- Better **activation functions** (e.g., ReLU), **optimizers** (e.g., ADAM) and **architectures** (e.g., Highway networks)

Deep Learning: Basics

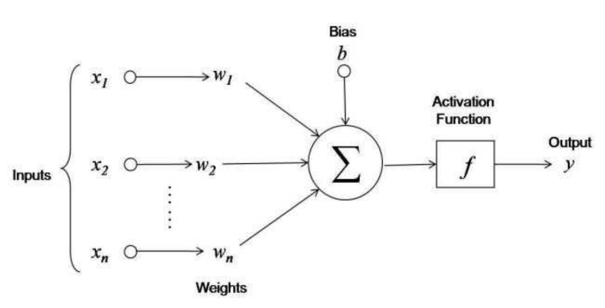
Basic Unit: Neuron



$$y = f(W^T X + b)$$

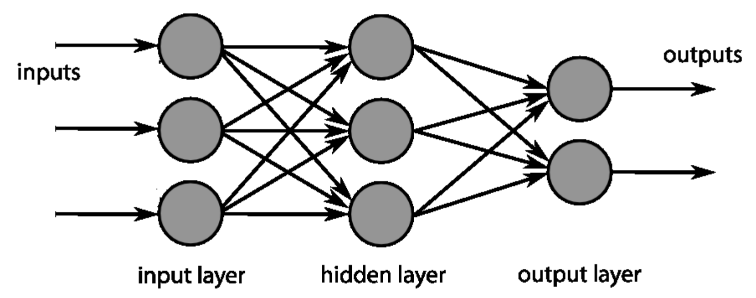
$$\text{Sigmoid Activation: } f(z) = \frac{1}{1 + e^{-z}}$$

Basic Unit: Neuron



Neuron acts as a logistic regression model

Neural Networks: Multiple logistic regressions



The Backprop Algorithm

- An application of the **chain rule**: the rate of change with respect to a variable x is the sum of rate of changes with respect to other variables z_i multiplied by the rate of change of z_i with respect to x

$$\frac{\partial f}{\partial x} = \sum_{z_i} \frac{\partial f}{\partial z_i} \frac{\partial z_i}{\partial x}$$

- The “extra” variables we use are the activations in different parts of the network: the derivative of the output with respect to a parameter is the derivative of the output with respect to its activation times the derivative of the activation with respect to a parameter... and apply it recursively

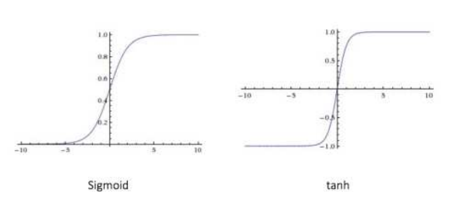
What Happens When Deep is Really Deep?

Vanishing Gradients

- Even large changes in the weights, especially in the early layers, make small changes in the final output

Exploding Gradients

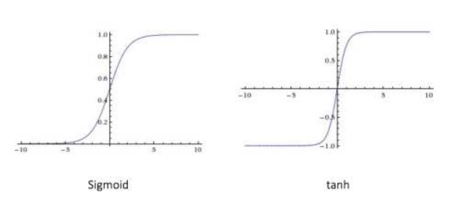
- Results in very large updates to neural network model weights during training.



What Happens When Deep is Really Deep?

Slow convergence: The model is unable to get traction on the training data

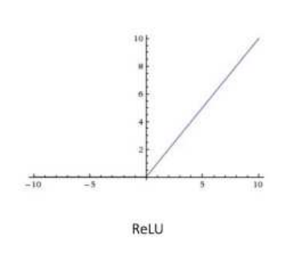
Unstable model: The model loss goes to 0 or NaN during training



What Happens When Deep is Really Deep?

How to tackle Vanishing and Exploding Gradients?

- Rectified Linear Activation



- Gradient Clipping

- Long Short-Term Memory Networks (LSTMs)

Neural Networks: Questions

Why do we need non-linear activations?

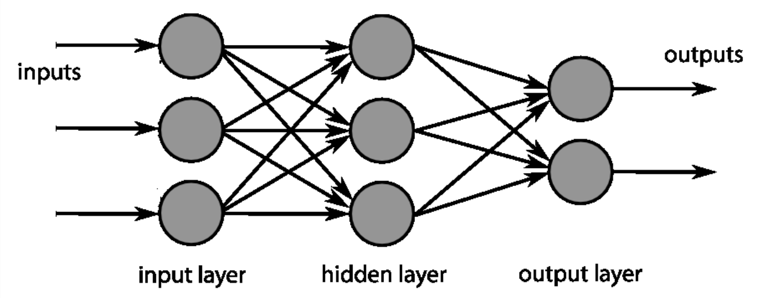
Does the backprop algorithm guarantee to find the best solution? If not, why not?

Why do neural networks still perform better than other models on various tasks?

Deep Learning in Application

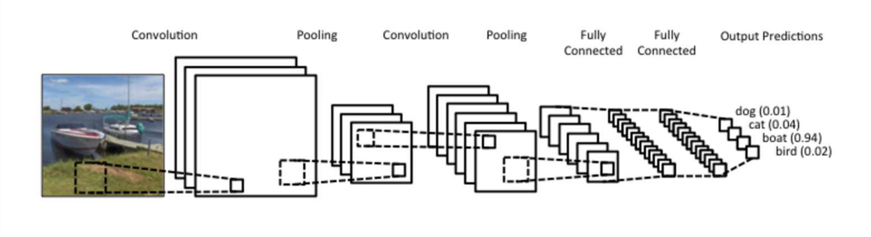
Buckets of Deep Learning (Andrew Ng)

1. Traditional fully-connected feed-forward networks, multi-layer perceptron (Classification)



Buckets of Deep Learning (Andrew Ng)

2. Convolutional Neural Networks (Vision, Mainly Spatial data, e.g., images)



Buckets of Deep Learning (Andrew Ng)

3. Sequence Models: Recurrent Neural Networks (RNN), Long Short Term Memory Networks (LSTM), Gated Recurrent Units (Language)

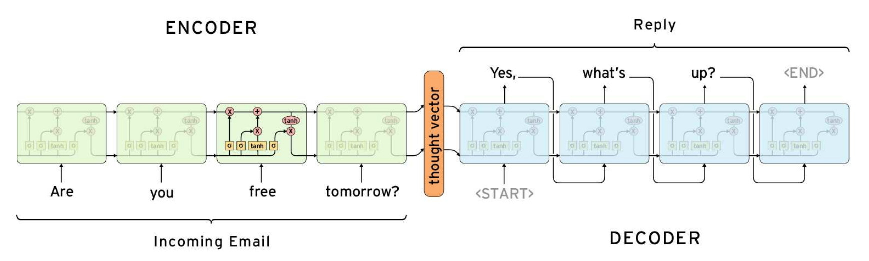


Image: https://cdn-images-1.medium.com/max/2000/1sO-SP58T4brE9EHazHSeGA.png

Buckets of Deep Learning (Andrew Ng)

1. **Traditional fully-connected feed-forward networks, multi-layer perceptron** (Classification)

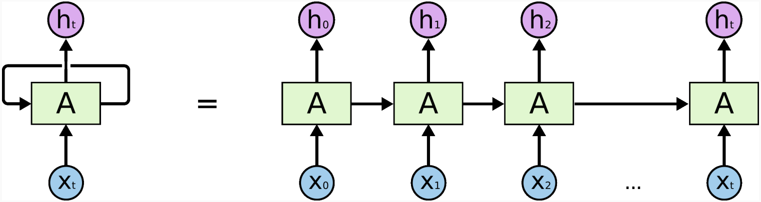
2. **Convolutional Neural Networks** (Vision, Mainly Spatial data, e.g., images)

3. **Sequence Models**: Recurrent Neural Networks (RNN), Long Short Term Memory Networks (LSTM), Gated Recurrent Units (Language)

4. **Future of AI**: Unsupervised Learning, Reinforcement Learning, etc.

Recurrent Neural Network

$$h_t = f(W_1 x_t + W_2 h_{t-1} + b)$$



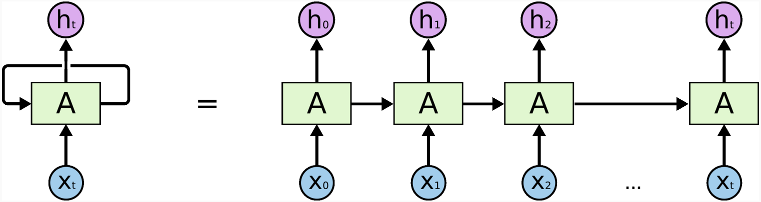
Internal state *h* memorises context up to that point

Applications: Language modelling, neural machine translation, natural language generation and many more

Image: http://colah.github.io/posts/2015-08-Understanding-LSTMs/img/RNN-unrolled.png

Training Recurrent Architectures

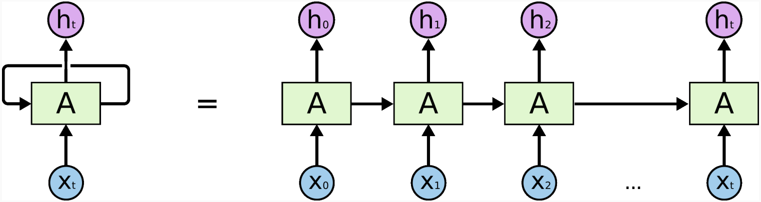
$$h_t = f(W_1 x_t + W_2 h_{t-1} + b)$$



Unroll the inputs and the outputs of the network into a long sequence (or larger structure) and use the **back-propagation** algorithm

Training Recurrent Architectures

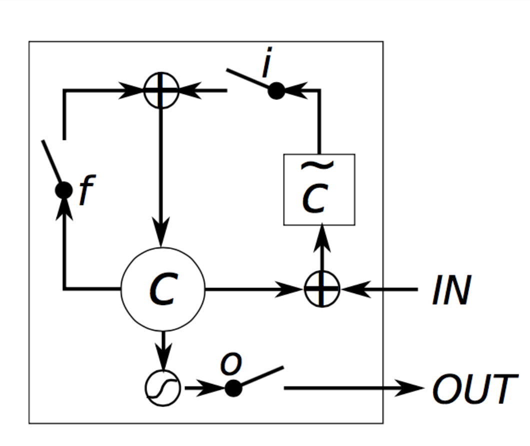
$$h_t = f(W_1 x_t + W_2 h_{t-1} + b)$$



Unroll the inputs and the outputs of the network into a long sequence (or larger structure) and use the **back-propagation** algorithm

Vanishing gradient problem??

Image: http://colah.github.io/posts/2015-08-Understanding-LSTMs/img/RNN-unrolled.png



Input gate, output gate and **forget gate**

Image: taken from Chung et al. (2014)

Gated Recurrent Units (GRUs)

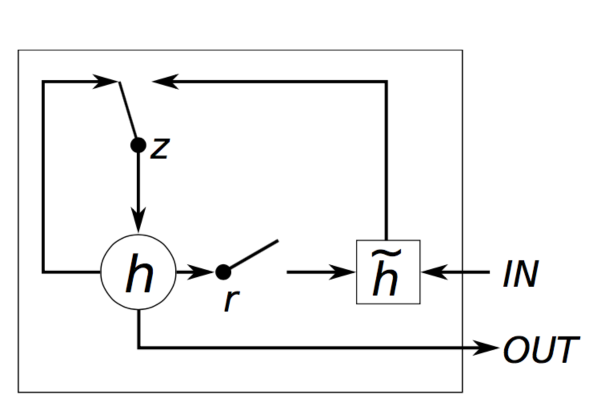
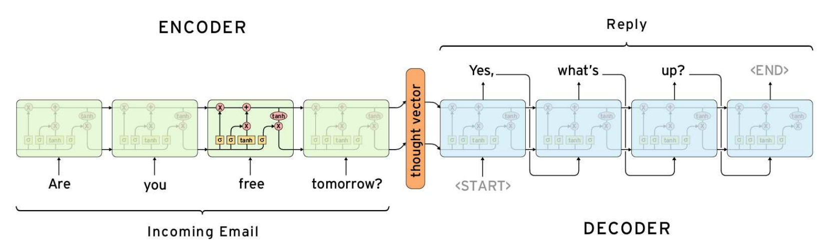


Image: taken from Chung et al. (2014)

Sequence to Sequence Models

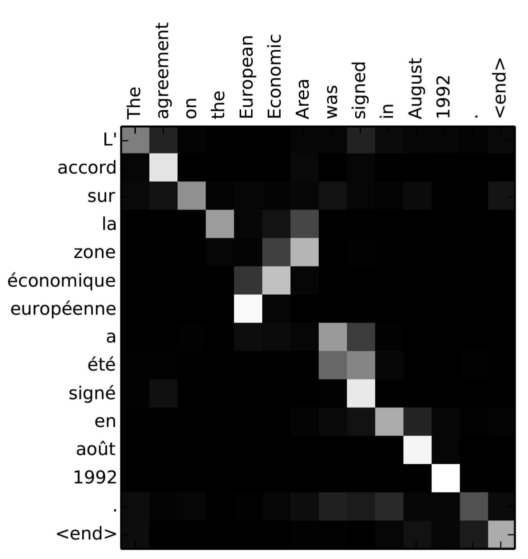


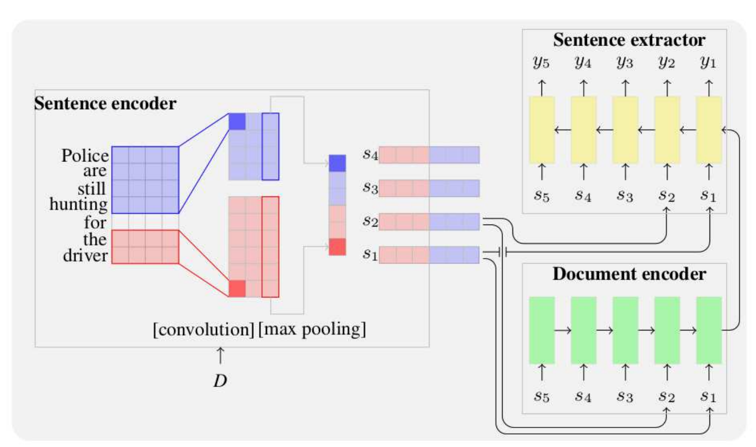
Encoder encodes the input sentence into a vector and then **decoder** generates the output sentence, one word at a time

Machine translation and dialogue generation

Image: https://cdn-images-1.medium.com/max/2000/1sO-SP58T4brE9EHazHSeGA.png

Sequence to Sequence Models with Attentions





Document Modelling

Cautions

Requires large amount of training data

Hyper-parameter tuning and non-convex optimization

Model interpretability is a growing issue

Encoding structure of language: not everything is a sequence

Summary

Deep learning is extremely powerful in learning feature representations and higher-level abstractions

It is very simple to start with: Many off-the-shelf packages available implementing neural networks