A Deep Learning Primer: From Perceptrons to Deep Networks

John Min

April 2, 2014

1 Perceptron

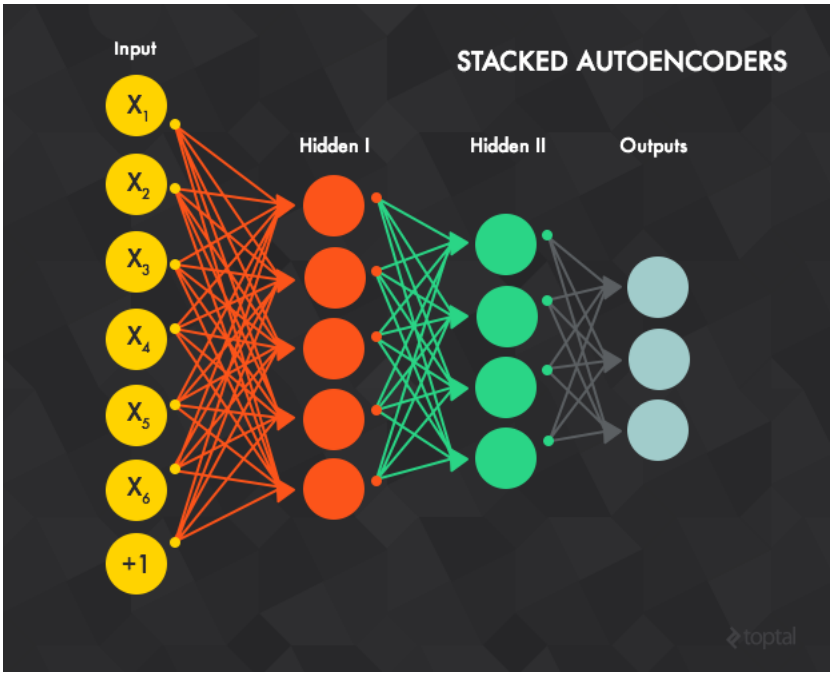
In 1957, the perceptron was invented as one of the earliest supervised learning algorithms. Now, a fundamental building block of neural networks, the perceptron is a linear, binary classifier.

2 Autoencoder

An **autoencoder**, also known as an autoassociator or Diabolo network, is an artificial neural network used for learning efficient codings, a compressed, distributed representation of the data to reduce dimensionality. Autoencoders are comprised of at least three layers: an input layer, an output layer, and hidden layers constituting the encoding.

If the neurons are linear or the single hidden layer is sigmoid, then, the optimal solution is strongly related to principal component anaalysis (PCA).

3 Stacked Autoencoders



The hidden layer of autoencoder t acts as an input layer to autoencoder $t+1$.

4 Feed-Forward Neural Network

The **Universal approximation theorem** states that a feed-forward neural network with a single hidden layer containing a finite number of neurons (a multilayer perceptron) can approximate continuous functions on compact subsets of $\mathbb{R}^n$.

5 Hopfield Nets

A **Hopfield network** is composed of binary threshold units with recurrent connections between them.

- symmetric connections ⇒ global energy function
- $E = -\sum_i s_i b_i - \sum_{i<j} s_i s_j w_{ij}$
- Energy gap $= \nabla E_i = E(s_i = 0) - E(s_i = 1) = b_i + \sum_j s_j w_{ij}$

5.1 Settling to an energy minimum

5.2 Storage capacity

6 Boltzmann Machine

A **Boltzmann machine** is a stochastic recurrent neural network, the stochastic, generative counterpart of Hopfield nets. Like a Hopfield network, a Boltzmann machine is a network of units with an energy defined for the network. While a B.M. has binary units, what differentiates it from a Hopfield net is that the units are stochastic.

RBM - a generative stochastic neural network that can learn a probability distribution over its set of inputs

7 Restricted Boltzmann Machine

Restricted Boltzmann Machines are a variant of Boltzmann machines with the constraint that neurons form a bipartite graph. RBMs are composed of a hidden, visible, and bias layer. (There is only one layer of hidden units). Each edge in an RBM must connect a visible unit to a hidden unit. (By contrast, "unrestricted" Boltzmann machines may have connections between hidden units, making them recurrent networks.) Unlike the feed-forward networks, the connections between the visible and hidden layers are undirected meaning that the values can be propogated in both the visible-to-hidden as well as hidden-to-visible directions. The network is also fully connected (each unit from a given layer is connected to each unit in the next).

7.1 Energy-Based Models (EBM)

Energy-based models associate a scalar energy to each configuration of the variables of interest. [1] Learning corresponds to modifying the energy function such that its shape has desirable properties; for example, configurations to have low energy (minimization). Energy-based probabilistic models define a probability distrubution through an energy function, as follows:

$$p(x) = \frac{e^{-E(x)}}{Z}$$

.

The normalizing factor Z is the **partition function** in the context of physical systems.

$$Z = \sum_x e^{-E(x)}$$

An energy-based model can be learnt by performing (stochastic) gradient descent on the empirical negative log-likelihood of the training data. As for the logistic regression, we will first define the log-likelihood and then the loss function as being he negative log-likelihood.

$$\mathcal{L}(\theta, \mathcal{D}) = \frac{1}{N} \sum_{x^{(i)} \in \mathcal{D}} \log p(x^(i))$$

$$\ell(\theta, \mathcal{D}) = -\mathcal{L}(\theta, \mathcal{D})$$

Use the stochastic gradient $-\frac{\partial \log p(x^(i))}{\partial \theta}$ where θ are the model parameters

[1] Energy is synonymous for objective, loss, cost, or utility.

7.2 Contrastive Divergence algorithm

Hinton, 2002 Training Products of Experts by Minimzing Contrastive Divergence - renormalizing product of probability distributions aka individual experts - minimizing contrastive divergence is much easier to infer than maximizing data likelihood on PoE - latent variables of different experts are conditionally independent

The single-step contrastive divergence algorithm (CD-1)

1. Positive phase
2. Negative phase
3. Weight update

7.3 Issues with using Reconstruction Error

8 Deep Belief Networks

As with autoencoders, we can also stack Boltzmann machines to create a class of networks known as deep belief networks (DBNs).

Deep Belief nets
http://www.cs.toronto.edu/ hinton/absps/fastnc.pdf

9 Convolutional Nets (LeNet)

Convolutional Neural Networks (CNN) are a special class of feed-forward (multi-layer) neural networks that have demonstrated effectiveness in image recognition. Analogous to other neural networks, they are trained using the back-propagation algorithm. What differentiates convolutional nets from others is the architecture.

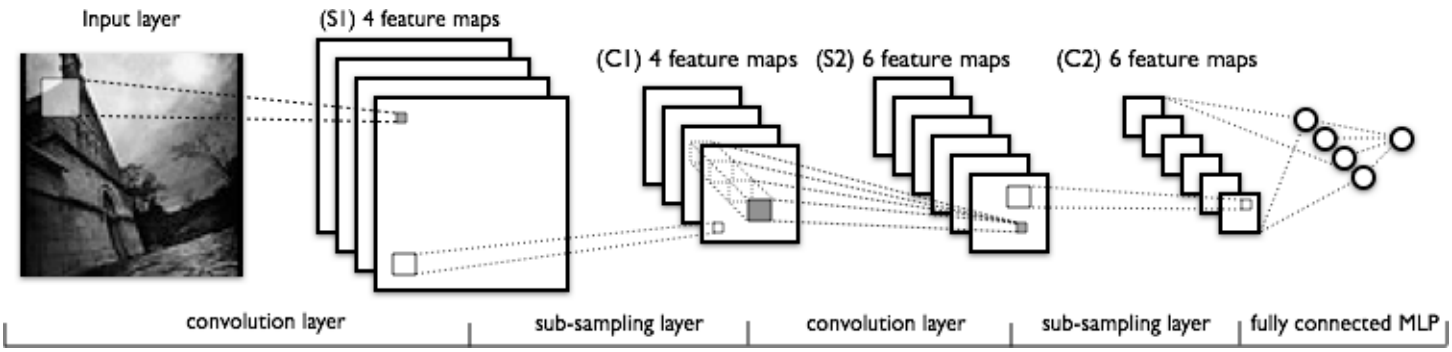
Inspired from biology, this multi-layer perceptron variant originates from research on the complex arrangement of cells within the visual cortex.
Before diving into the architecture of a convolutional neural net, let's define an image *filter*, a square region with associated weights. A filter is applied across an entire input image, and you often apply multiple filters.

Convolutional layers apply a number of *filters* to the input. The result of one filter applied across the image is called a *feature map*(FM) and the number of feature maps is equivalent to the number of filters.

The intuition behidn the shared weights acros the image is that the features will be detected regardlesss of their location, while the multiplicity of filters allows each of them to detect different sets of features.
Subsampling layers reduce the size of the input. Popular ways to subsample are the following:

- max pooling
- average pooling
- stochastic pooling

This architecture enables convolutional nets to recognize patterns with extreme variability and with robustness to distortions and simple geometric transformations aka translation invariance.[2]



[2]Yann LeCun's LeNet on the MNIST dataset

10 Recurrent Neural Networks

A **recurrent neural network** (RNN) is a class of neural network where connections between units form a directed cycle.