Deep Learning of Representations: Looking Forward Discussion

Zhuoyan Xu

February 19, 2019

General structure

- More like the history of DL, the summary and evaluation of methods in different fields or situations.
- Deep learning's theoretical results, learning algorithms and breakthrough experiments.
- Overview of algorithms and some concepts.
- The challenges and solution paths in scaling computations, difficulties in optimizing parameters, designing inference and sampling, representations that better disentangle the unknown underlying factors of variation.

Overview of algorithms

- Some breakthrough brought by the rising of some algorithms, such as the presence of rectifying non-linearities(ReLU), the use of convolutional architectures(alternative convolutional layers and pooling layers), and dropouts.
- Layer-wise pre training(supervised and unsupervised): find initialization that leads better results. This method Nowadays is not as popular as ReLU, drop out and batch normalization(which can be thought as doing pre-processing for each layer of network[Sergey Ioffe, Christian Szegedy 2015]).

Regularized Auto-Encoders

- Auto-encoder:A kind of unsupervised pre-training:

$$\min ||r(x) - x||_2^2$$
$$r(x) = W_d^T f(W_e x + b) + c$$

- Regularized auto-encoder(prevent case like $r(x) = x$): bottleneck auto-encoders: less units in hidden layer; Denoising auto-encoder: take a noisy version N(x) instead of x.i.e. $\min ||r(N(x) - x)||_2^2$; The contractive auto-encoder: minimize the contractive penalty $||\frac{f(x)}{x}||_F^2$.
- The tug-of-war minimizing reconstruction error and the regularizer means intermediate representation(hidden layer) capture most variation of training samples(i.e. the characteristics of variation on manifold reflect the raw data characteristics)

Challenges and solutions in optimization

- The top two layers can be overfit, it is important to optimize lower layers, which can not be achieved only by looking at training criterion. Backpropogated gradients is sometimes weak to train intermediate layers.
- Improved general-purpose optimization algorithms,such as adaptive learning rates.
- Non-linearities can produce sparse outputs. When gradient is sparse(only small subsets of hidden units and parameters touched by gradients), the problem become easier.