(1) Title, author and where it was published

- Title: Deep Learning; Chapter 8 Optimization for Training Deep Models
- Author: Ian Goodfellow and Yoshua Bengio and Aaron Courville
- Published: MIT Press

(2) Summary of the content

This chapter first explains the difference between pure optimization and optimization as a training algorithm. Then, the authors bring up the challenges that current machine learning algorithms face, and present some prevalent optimization algorithms for the reader. The goal for optimization algorithms is to find the parameters of a neural network that significantly reduce a cost function, and the optimization algorithms used for training of deep models aim to improve the performance measure indirectly through reducing different cost functions. With a detailed description in Stochastic gradient descent algorithm and other optimization techniques like Adam and RMSProp, this paper provides a holistic review on the current popular methods in the Deep Learning field.

(3) Strengths of the paper

The author clearly defines several professional terms specific to the machine learning realm with mathematical formulas and straightforward interpretation. For example, the risk in the machine learning context is the expected generalization error across the data-generating distribution given the entire training dataset, and the empirical risk is such error across the training set. Therefore, what basic optimization algorithms try to do is to minimize the empirical risks, which, as the author emphasized, tends to suffer from overfitting.

Moreover, we learnt that despite empirical risk minimization, optimization as a training algorithm also benefits from using surrogate loss functions and minibatch. The surrogate loss functions are more tractable than some loss functions that are not fully differentiable, and we could add early stopping criterion for optimization algorithms to avoid wasting of computing power.

(4) Major critiques

Since it is a book for education, it is seldom that they would have errors in assumption nor in experimental designs.

Reading 2

(1) Title, author, and where is was published

- Title: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation
- Authors: Kyunghyun Cho, Bart van Merrienboer Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio
- Published: Arxiv

(2) Summary of the content

This paper is an overview of the Encoder-Decoder structure in a deep learning network model that consists of two Recurrent Neural Networks. Based on the previous in the statistical machine translation (SMT) field, the authors propose an RNN Encoder–Decoder network architecture that are trained jointly to maximize the conditional probability of the target sequence given a source sequence. Basically, the authors trained a network (the encoder) to convert a sequence with variable length into a fix-length sequence, and trained a network (the decoder) to convert it back to a variable-length sequence. Motivated by the architecture of LSTM, this encoder-decoder network also incorporates a layer with hidden units that adaptively remember and forget about the information fed. The decoder's mechanism is largely based on the SMT where the translation model is factorized into the translation probabilities of matching phrases in the source and target sentences. Then, an experiment on the English/French translation task that demonstrates the effectiveness of the encoder-decoder network architecture was presented, and they performed both the quantitative and qualitative analyses on the results. Concluding with the statement that the model could capture linguistic regularities in the phrase pairs well the RNN Encoder–Decoder and also is able to propose well-formed target phrases, the authors bring the readers with a novel technique in the NLP field.

(3) Strengths of the paper

In the experimental results section, this paper presents plots of resulting words/phrases space that are learned by the network, visualizing the power of this network with the most straightforward evidence. In addition, even though the word embedding space is hard to represent since it is high-dimensional and crowded in space, the authors also provide a zoomed-in view of specific regions for the reader to take a detailed look at. I understand the power of this structure much more after reading these plots with the results displayed.

(4) Major critiques

The paper only performed its experiment on the English to French translation task. However, since both languages belong to the same linguistic origin, I suspect the model's effectiveness in translating other languages, like from English to Chinese, since there are rarely any similarities between the two languages.