Thesis Defense

MACHINE LEARNING DEPARTMENT

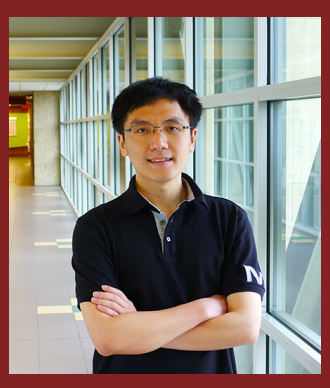
Towards Effective and Efficient Learning at Scale

Abstract:

How to enable efficient and effective machine learning at scale has been a longstanding problem in modern artificial intelligence, which also motivates this thesis research. In particular, we aim at solving the following problems: 1) How to efficiently train a machine learning model? 2) How to speed up inference after the model is trained? 3) How to make the model generalize better? We approach these problems from two perspectives: models and algorithms. On one hand, we design novel models that are intrinsically fast to train and/or test. On the other, we develop new algorithms with rapid convergence guarantee.

The first part presents new machine learning models with a focus on sequential data such as natural language processing and question answering. Firstly, we propose a model, LSTM-Jump, that can skip unimportant information in text, mimicking the skimming behavior of human reading. Trained with an efficient reinforcement learning algorithm, this model can be several times faster than a vanilla LSTM in inference time. Then we introduce a text encoding model that totally discards recurrent networks, which thus fully supports parallel training and inference. Based on this technique, a new question-answering model, QANet, is proposed. Combined with data augmentation approach via back-translation, this model stays at the No.1 place in the competitive Stanford Question and Answer Dataset (SQuAD) from March to Aug 2018, while being times faster than the prevalent models.

The second part proposes large scale learning algorithms with provable convergence guarantee. To enable fast training of neural networks, we propose a general gradient normalization algorithm for efficient deep networks training. This method can not only alleviate the gradient vanishing problem, but also regularize the model to achieve better generalization. When the amount of training data becomes huge, we need appeal to distributed computation. We are particularly interested in the ubiquitous problem, empirical risk minimization, and propose a few algorithms to attack the challenges posed by the distributed environment. We first show that a randomized primal-dual coordinate method DSPDC can achieve a linear convergence rate in the single-machine setting. Then by further leveraging the structural information of the problem, we propose an asynchronous algorithm DSCOVR, which enjoys a linear convergence rate as well on the distributed parameter server environment, by executing periodical variance reduction.



Speaker:

Adams Wei Yu

Thesis Committee:

Jaime Carbonell (Co-Chair)
Alex Smola (Co-Chair, Amazon)
Ruslan Salakhutdinov
Quoc Le (Google)
Christopher Manning (Stanford)

July 12, 2019
12:00pm
GHC 8102

Link to draft document:
https://www.dropbox.com/s/2bhw61k43jjw9e1/thesis.pdf?dl=0