Language Technologies Institute

Thesis Proposal

GHC 9115| Monday, December 10| 11:30 am

Efficient Machine Learning

Hieu Pham

Abstract

I study the methods that make deep learning algorithms more efficient in three facets: 1) model training time; 2) model designing time; and 3) data complexity.

I start out by developing Neural Combinatorial Optimization (NCO, Chapter 1, a reinforcement learning algorithm that can find close-to-optimal to canonical combinatorial optimization tasks, such as the Traveling Salesman Problem and the KnapSack problem. The NCO algorithm isoblivious to tasks. Specifically, if NCO has access to a task reward that can be efficiently evaluated, then these algorithms can search for good solutions.

Next, in Chapter 2, I consider the problem of device placement for neural networks. Given a network, \eg~InceptionV3 and a list ofavailable devices (e.g. 4 GPUs and 1 CPU), and one has to decide which parts of the given network should run on which device to minimize the network's execution time. I treat the task as a combinatorial optimization problem and apply NCO. NCO finds placements that are up to 23\% faster than the placements designed by human experts.

Then, in Chapter 3, I consider the problem of designing neural network architectures. This task is known as Neural Architecture Search (NAS), and can be posed as a discrete optimization problem, but with the bottleneck that evaluating reward functions is very expensive. I apply NCO on this task, with a significant motivation where all architectures are forced to share their parameters. The resulting algorithms, Efficient Neural Architecture Search (ENAS) could achieve similar performance to NAS, but is 1,000x to 50,000x more efficient.

Then, in Chapter 4, I apply ENAS to design architecture for Neural Machine Translation (NMT). I found that the search space is crucial for ENAS. The search space I designed for language model (Chapter 3) appears not compatible with NMT. An alternate search space is being designed and my preliminary experiments show that such alternate space may work well.

Finally, in Chapter 5, I design an unsupervised learning algorithm to reduce the amount of data needed (and hence, also improve the training time needed) for image classification.

https://drive.google.com/file/d/1gPyLZQxHu4aPrxkqmeLFnPgC-dckE6Xd/view?usp=sharing

COMMITTEE:

Jaime Carbonell, (co-advisor)

Quoc V. Le, (Google Brain) (co-advisor)

Samy Bengio, (Google Brain)

Chris Dyer, (CMU & DeepMind)

Barnabas Poczos (CMU)