Sparsifying Deep Neural Networks with Generalized Regularized Dual Averaging

Guang Cheng

Purdue Statistics
E-mail: chengg@purdue.edu

Abstract: Deep learning has shown strikingly good performance in image classification, machine translation, text-to-speech translation. However, as modern deep neural networks (DNNs) require huge computational resources to store and process, deploying DNNs on devices and systems, e.g. mobile devices, requires to address storage and computational constraints. It is commonly believed that DNN is usually overparametrized, and it is possible to shrink DNN without sacrificing its accuracy via, e.g. pruning (Han et al, 2015). However, pruning is not efficient as it requires pre- and post-training of the model, and there is no theoretical justification for it. In this talk, we introduce a generalization of regularized dual averaging (gRDA) for sparsifying DNN, which does not require pre- and post-training. Under infinitesimal learning rate, gRDA has the same learning trajectory as stochastic gradient descent (SGD). Therefore, asymptotically gRDA achieve the same generalization level as SGD. However, the distributional dynamics of gRDA is drastically different from that of SGD. Specifically, an autoregressive soft-thresholding operator enters the distribution dynamics of gRDA, which encourages sparsity. Theoretical insights are provided to guide the selection of hyperparameters, which is validated by empirical analysis using CIFAR-10.