Project Proposal: Clustered Variance Reduction For Neural Nets

Adithya Sagar
asg242@cornell.edu

Yang Yuan
Department of Computer Science
yy528@cornell.edu

1 Proposal

1.1 Motivation

Deep learning is now becoming a widely prevalent technique for solving various machine learning problems that include object recognition, speech recognition, auto driving, machine translation, etc. However, one of the major impediments about neural networks is their slow training time. Though certain degree of success has been achieved through the use of distributed machines with powerful GPUs to train deep neural nets, the the underlying training algorithms have only been few. They primarily include stochastic gradient descent (SGD), or SGD with auto tuning step size (e.g, AdaGrad and RMSProp), or adaptation of Nesterov's acceleration algorithm to the neural nets (e.g., Momentum and Adam). These algorithms are only a small subset of what has been traditionally used by the optimization community. Thus training Deep Nets faster is a significant challenge which probably could be addressed by considering novel algorithms that have been developed in the area of optimization.

1.2 Approach

Recently, a new category of algorithms called variance reduction (VR) methods have been proposed, [**? ? ? ? ? ? ? ? ? ? ?**], which greatly improved the convergence rate of the vanilla SGD by reducing the variance of each (???) stochastic gradient during the training. The basic idea of variance reduction is to store some previous stochastic gradient information, and then use that to be adjustment for the current stochastic gradient, which helps reduce the variance. However, most of the papers are only considering convex or even strongly convex objectives, in which the proposed algorithms have strong theoretical guarantee. Several papers consider the case of a convex objective function that can be written as the sum of non-convex sub-functions ([**? ?**]), which can be seen a slight relaxation of convex objective, but still the condition is not weak enough to provide any meaningful insights for highly non-convex objectives like neural nets. Recent work by Johnson et al. [**?**] presented a few preliminary results using this technique on feed-forward neural networks but have not presented an adequate discussion on other neural network architectures like convolutional neural networks (CNNs) or recurrent neural networks (RNNs). Thus it will make an interesting study to quantify the performance of VR based optimization method on different neural network architectures.

In addition, VR methods have exploited [**? ?**] the internal structure of the data using different clustering techniques like ClusterSVRG and ClusterACDM [**?**], to enhance the speed of optimization. Cluster dependent VR will be much faster as compared to SVRG and ACDM, based on the quality of the structure

On the other hand, the required clustering on the data set can be raw and can contain a few outliers, and can be computed very efficiently in just one pass of the data, using standard nearest neighbor algorithm like LSH.

This idea could be particularly interesting for deep learning applications, since it is common practice to create augment data set using trivial transformation or adding small perturbation on the existing training data, in order to get better training results. Based on the definition of clustering structure [**?**], most of the augmented data points can be seen a inside the same cluster as the original data point, which shows clear advantage of the newly proposed algorithms. Moreover, even without the augment data set, it is natural to assume good clustering structure inside the original data set. For example, one person's face seen from slightly different angles can be treated as from the same cluster, and certain animals from the same category could form one cluster as well.

We would like to try to apply VR methods or clustered VR method to explore how neural nets will benefit from that. We are planning to try different scenarios, like feed-forward neural nets, convolutional neural nets, and also recurrent neural nets, with different structure design, and different type of data set. The final outcome of this project will be a detailed analysis and comparison of the performance of clustered VR method in different scenarios. We will investigate different dimensions of performance including convergence of objective function and training error, training time, etc.

One issue that we have in mind is that since variance reduction method reduces the variance of the stochastic gradient, it might be more susceptible to stuck at saddle points. According to [**?**], one could add a uniform noise in first-order optimization methods for escaping non-degenerate saddle points. We will also try this idea if necessary.

1.3 Data

1.4 Evaluation