CS229T/STATS231 Winter 2014 – Project Abstract

Awni Hannun
awni@stanford.edu

Peng Qi
pengqi@stanford.edu

January 29, 2014

Abstract

Introduction

Multilayer neural networks and other deep models have gained considerable traction over the past several years demonstrating state-of-the-art performance in many applications including speech recognition, object recognition in images and even language modeling. Correspondingly, much work has been done to improve methods for optimizing these functions which can be difficult due to the highly nonconvex landscape of the typical multilayer objective. Given growth in dataset sizes and that these highly expressive models tend to be used in the large data setting, stochastic optimization methods are common. However, stochastic methods tend to come with a host of hyperparameters and much tuning is involved to get these to work well which has lead to general disagreement as the supieriority of a particular algorithm.

The goal of our project is to perscribe the best stochastic optimizer to use for a general deep neural network perhaps accounting for modalities with varying statistical properties. We will attempt to solve this problem both from an analytical standpoint, to the degree that this is possible, and with an empirical evaluation.

Prior Work

Until recently stochastic gradient descent (SGD) with momentum has been the standard stochastic optimizer used with deep neural networks (DNNs) [2]; however, other stochastic methods are beginning to be adopted including AdaGrad [1] and variations as well as the formulation of the nesterov accelerated gradient (NAG) as presented in [5]. To the best of our knowledge, we do not know of any work which gives a thorough comparison of these algorithms and demonstrates when one or the other should be applied. There has been some work in automating the hyperparamter tuning process for a given algorithm [4], yet this does not answer the question of which algorithm should be used in the first place and furthermore still requires the training of many models which can be prohibitive.

Furthermore, recent work has begun to attempt an analytical understanding of these types of heirarchichal models; however, often the results are derived using simplifications or modifications of a typical DNN which make them not applicable in practice. [3]

Approach and Challenges

We seek to analyze deep networks and the settings in which certain stochastic methods may be more efficient. Ultimately we plan to perscribe a recipe for the correct algorithm or perhaps new variation of an existing algorithm to use. In order to do this we propose to analyze empirically several stochastic optimizers including SGD with momentum, NAG and AdaGrad with the hope of shedding light on the behaviour of a specific optimization procedure at training time. A foreseeable challenge is that the hyyperparameter selection for a specific algorithm can have a very large impact on the resulting convergence properties, thus we will need to be dilligent in our comparison and this may require much computation.

Furthermore, we wish to approach this problem from a different angle by asking the question: given an idea of what the landscape of the objective looks like which algorithm is best to employ? One way we propose to gain insight about the landscape of a typical DNN objective is to keep a running trace of gradients during optimization and then project these gradient to a lower dimensional subspace which we can then visualize. Using a simple method such as PCA to visualize the two primary modes of variation of our gradient trace could lose too much information and make too strong an assumption on the shape of the objective. We will likely need to explore advanced dimensionality reduction methods or resort to gaining insight about the objective via other methods than a visualization.

Finally, and we recognize this as perhaps the most challenging component of our project, we wish to give some analytical results to verify our empirical observations or relate our observations to some existing analysis. Attempting to come up with some kind of convergence guarantees for such difficult to analyze models will perhaps prove to be too ambitiouos; however, at the very least we plan to relate the different algorithms to one another via a mathematical reformulation in order to give better intuition of the trade-offs between them.

References

[1] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 2011.

[2] G. Hinton. A practical guide to training restricted boltzmann machines. *Technical Report*, 2010.

[3] A. Saxe, J. McClelland, and S. Ganguli. Dynamics of learning in deep linear neural networks. *NIPS Workshop on Deep Learning*, 2013.

[4] J. Snoek, H. Larochelle, and R. Adams. Practical bayesian optimization of machine learning algorithms. *Neural Information Processing Systems*, 2012.

[5] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of momentum and initialization in deep learning. *International Conference on Machine Learning*, 2013.