Evaluation of Various Gradient Descent Optimization Techniques for Neural Networks

Mit Patel

Abstract—**Ever since the inception of Gradient Descent algorithm, it is without a doubt the most popular optimization strategy used in machine learning and deep learning. In this paper we have used various techniques to accelerate the gradient vectors in the right direction. The main problem with gradient descent algorithm is the rate of convergence. So to speed up the process, we take into effect all the previous gradients and tune the current gradient accordingly. There are various techniques available to do so and we have explored 5 such techniques namely i) No Momentum ii) Polyak's Classical Momentum iii) Nesterov's Accelerated Gradient (iv) RmsProp and (v) ADAM . We will compare the accuracies, rate of convergence and the stability for all this techniques in this paper.**

Index Terms—**IEEE, IEEEtran, journal, LaTeX, paper, template.**

I. Introduction

THE most common method for neural network optimization is gradient descent and is one of the most favored algorithms. And almost every deep learning library is equipped with various implementations of numerous algorithms to optimize the gradient descent. In this paper, we have implemented some of those techniques. Gradient descent is a way to minimize an objective function J(θ) parameterized by a model's Non-convex optimization problems are natural formulations in many machine learning problems (e.g. (Un)supervised learning, Bayesian learning). Various learning approaches have been proposed in such settings, as global minimization of such problems are NP-hard in general. Gradient descent is the defacto iterative learning algorithm used for such optimization problems in machine learning, especially in deep learning. Several variants of gradient descent methods have been proposed and all thse proposed methods can be broadly classified into momentum-based methods (e.g. Nesterov's Accelerated Gradient [9]), variance reduction methods (e.g. Stochastic Variance Reduced Gradient [6],[11]) and adaptive learning methods (e.g. AdaGrad [2]). Gradient descent coupled with momentum - also called classical momentum by Polyak [10], is the first ever variant of gradient descent involving the usage of a momentum parameter. The momentum methods use the information from previous gradients in addition to the current gradient for updating the learning parameters. Nesterov in his seminal work [9], proposed an accelerated gradient method (also a momentum based method as shown by [15]) which gives an upper bound on the number of iterations for learning algorithm to converge. With the tremendous success of deep learning models, Sutskever et al in their work [15] worked out to incorporate the algorithm by Nesterov [9]. Nesterov's method performs an update in the same way as classical momentum, only with a correction to the gradient

mds
August 26, 2015

A. Subsection Heading Here

Subsection text here.

1) Subsubsection Heading Here: Subsubsection text here.

II. Conclusion

The conclusion goes here.

Appendix A
Proof of the First Zonklar Equation

Appendix one text goes here.

Appendix B

Appendix two text goes here.

Acknowledgment

The authors would like to thank...

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

PLACE
PHOTO
HERE

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.