The Name of the Title is Hope

Your Name
University of Passau
your.name@uni-passau.de

ABSTRACT

Here goes your abstract.

1 INTRODUCTION

Adaptive learning algorithms like AdaGrad and Adam are being extensively used thanks to there shorter learning time, However the benefit of using such algorithms on datasets where the number of model parameters to train on is larger than the available data, is to be investigated. In this paper we try to reproduce the experiment done by Wilson & al. [WRS+17] in order to investigate the results of using adaptive learning algorithms compared to there non-adaptive counter parts in the case where the available number of points in the data is smaller than the number of model parameters to train.

2 BACKGROUND

In this section we are going to introduce the algorithms used in the experiment. We are going to introduce them using an informal and a mathematical description. To visualize how these algorithms function differently we suggest that you take a look at Lili Jiang's tool Gradient Descent Viz [Jan20].
In the next sections we define x to be a vector or a scalar, t is the iteration number, L is the loss function, W is the weights matrix, α is the learning rate and β_i are decay rates.

2.1 Non-adaptive algorithms

Stochastic Gradient descent [KW52] is the most basic Learning algorithm. To minimize the Loss function it calculates the function's gradient at a particular point and updates the point coordinates with the negative value of that gradient. Formally, in the iteration t SGD calculates the next point coordinate using the following formula:

$$x_{t+1} = x_t - \alpha \nabla L(x_t)$$

One problem with SGD is that its learning speed is very slow [DAN+18] and can get caught in a local minimum easily [BJ19].

To solve this issue of slow learning rate we have invented the **Stochastic gradient descent with momentum** [Qia99]. This aproche is inspired from classical physics where the motion equation depends on the speed of the particle which can be seen as a feedback from previous steps. This principle is applied to the SGD through the use of momentum variable. Formally:

$$x_{t+1} = x_t - \alpha \nabla L(x_t) + \alpha \nabla L(x_{t-1})\beta$$

where $\beta \in [0, 1[$ is a decay rate that controles how much previous steps influence the current one. Analogically it can be seen as surface friction to help a moving object stop. If it is set to 1 the algorithm will never converge. Thanks to its momentum this algorithm can escape better local minimums.

2.2 Adaptive algorithms

Adaptive learning algorithms have the ability to adapt the learning rate to each parameter.
AdaGrad [DHS11] adapts the learning to each feature depending on how big is the gradient according to that feature is. Let $G_t = \sum_{\tau=1}^{t} \nabla L(x_\tau) \nabla L(x_\tau)^T$ AdaGrad calculates the new x coordinate using the following equation :

$$x_{t+1} = x_t - \alpha G_t^{-\frac{1}{2}} \circ \nabla L(x_t)$$

where $\circ$ is the Hadamard product. Generally the application as shown above is computationally slow for this reason we take only the diagonal of G since it is calculated in a linear time and approximate well the result. In practice we add an ϵ to the square root of G_t in order to avoid division by 0. One downfall of this algorithm is that G is growing with each iteration which ends up slowing the learning speed.
RMSProp[Hin12] solves this problem by adding a decay factor to the gradient sum. Let $G_0 = 0$

$$G_t = \beta G_{t-1} + (1 - \beta) \nabla L(x_t) \nabla L(x_t)^T$$

and the equation to calculate the next step coordinate remains the same.
Adam[KB14] on the other hand tries to combine RMSProp and AdaGrad. Let $M_0 = 0$ and $G_0 = 0$

$$M_t = \beta M_{t-1} + (1 - \beta) \nabla L(x_t)$$
$$G_t = \beta G_{t-1} + (1 - \beta) \nabla L(x_t) \nabla L(x_\tau)^T$$

the equation to update parameters becomes

$$x_{t+1} = x_t - \alpha M_t G_t^{-\frac{1}{2}}$$

3 EXPERIMENT

3.1 DataSet

The dataset is a subset of a collection of 8 million web images of size (32,32,3). The number of images is 60000 divided into a training set and validation set. The training set consists of 50000 images while the validation set consists of 10000 images. The images belong to 10 object classes with each class containing 6000 images. The dataset was collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton.

3.2 Model

For this experiment we use VGG model as it is described in the paper.

3.3 Description

The experiment consists of training the model using different optimizer methods.

4 RELATED WORK

5 CONCLUSION

6 MODIFICATIONS

REFERENCES

[BJ19] A. Bouillard and P. Jacquet. Quasi black hole effect of gradient descent in large dimension: Consequence on neural network learning. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8365–8369, 2019.

[DAN+18] E. M. Dogo, O. J. Afolabi, N. I. Nwulu, B. Twala, and C. O. Aigbavboa. A comparative analysis of gradient descent-based optimization algorithms on convolutional neural networks. In *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, pages 92–99, 2018.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12(null):21212159, July 2011.

[Hin12] Geoff Hinton. Neural networks for machine learning lecture 6. http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf, 2012.

[Jan20] LiLi Jang. Gradient descent viz. https://github.com/lilipads/gradient_descent_viz, Online, accessed 11 June 2020.

[KB14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[KW52] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.*, 23(3):462–466, 09 1952.

[Qia99] Ning Qian. On the momentum term in gradient descent learning algorithms, 1999.

[WRS+17] Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nathan Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning, 2017.