**Gravity Optimizer: A Mechanical View on Optimization in Deep Learning**

**Dariush Bahrami 1 ; Sadegh Pouriyan Zadeh 2**

1 Faculty of New Sciences & Technologies, University of Tehran, Tehran, Iran

2 Faculty of New Sciences & Technologies, University of Tehran, Tehran, Iran

**Abstract**

**# TODO: نوشتن خلاصه و کلمات کلیدی تو مرحله آخر**

Keywords:

**1. Introduction**

The question of choosing an adequate optimizer for a deep learning problem is not answered yet. Instead, there are ways like empirical comparing [1]–[3] or benchmarking [4] which help to find better configurations for optimization.

The most common optimization techniques in deep learning are SGD (Stochastic Gradient Descent)[5], RMS Prop[6], and Adam[7]. Table 1 shows the most common standard optimization techniques or optimizers in chronicle order. Table 1. the most common standard optimizers in deep learning.

Table 1. the most common standard optimizers in deep learning

|  |  |
| --- | --- |
| Year Published | Optimization technique |
| 1951 [5] | SGD |
| 1964 [8] | SGD with momentum |
| 2011 [9] | AdaGrad |
| 2012 [10] | AdaDelta |
| 2012 [6] | RMSProp |
| 2013 [11] | SGD with Nesterov momentum |
| 2015 [7] | Adam |
| 2015 [7] | Adamax |
| 2016 [12] | Nadam |
| 2018 [13] | AMSGrad |

# TODO: اینو تا 2 پاراگراف دیگه باید ادامه بدی

# Version 2: بخش دوم میشه related works

**2 Algorithm**

**3. Benchmark Configuration**

In this section, we are going to compare Gravity optimizer with other common standard optimizers shown in Table 1. In the following, first, the specifications of the hardware that we used for training are given. Then the framework we used to implement the model, the datasets used for training, and finally, the architectures chosen based on hardware specifications are introduced. If you want to skip reading the details, a summary of this information is given in Table 4. In the last part, the obtained results from Gravity optimizer are compared to the reported results from other papers that have used the same architectures we used to train the same datasets we used. If results are not reported by other papers, the tests have been performed by us.

3.1 Hardware

We used [Google Colab](https://colab.research.google.com/) [14] as hardware because we couldn’t afford GPU for training deep neural network models and testing our ideas. We used Tensorflow’s high-level API, Keras, as a framework to build the model in the Python language. The python implementation code can be found in [Gravity optimizer GitHub repository](https://github.com/dariush-bahrami/gravity.optimizer). Also, we were given the chance to use TPUs (Tensor Processing Unit: Google’s custom-developed technology to accelerate machine learning workloads) by using Google Colab and TensorFlow together.

3.2 Datasets

We used the following standard datasets to evaluate the performance of Gravity optimizer: MNIST, Fashion-MNIST, CIFAR-10, CIFAR-100 (Coarse), and CIFAR-100 (Fine). The MNIST database of handwritten digits is a subset of a larger set available from NIST. The images of digits have been size-normalized and centered in a fixed-size image [15]. The Fashion-MNIST is a dataset containing images of 10 classes. The 10 different classes are T-shirt, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, and Ankle boot[16]. CIFAR-10 is a subset of the 80 million tiny images dataset in 10 classes. The 10 different classes represent airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks [17]. CIFAR-100 is just like the CIFAR-10, except it has 100 classes containing 600 images each. The 100 classes in the CIFAR-100 are grouped into 20 super classes. Each image comes with a "fine" label (the class to which it belongs) and a "coarse" label (the superclass to which it belongs). We trained the “fine” and “coarse” datasets separately because they have two distinct labels that classify two different types of classification; “coarse” is more general and “fine” is more specific[17]. Table 2 summarizes the detailed information of the datasets used.

Table 2. detailed information of datasets used for benchmark

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset** | **Train #** | **Test #** | **Class #** | **Shape** | **Image per class** |
| MNIST | 60 K | 10 K | 10 | 28x28x1 | 6 K |
| Fashion-MNIST | 60 K | 10 K | 10 | 28x28x1 | 6 K |
| CIFAR-10 | 50 K | 10 K | 10 | 32x32x3 | 5 K |
| CIFAR-100 (coarse) | 50 K | 10 K | 20 | 32x32x3 | 2.5 K |
| CIFAR-100 (fine) | 50 K | 10 K | 100 | 32x32x3 | 500 |

3.3 Architecture (models and hyperparameters)

We used VGG16 and VGG19 with the exact specifications reported in their paper [18]. VGG16 has about 34M and VGG19 has about 39M parameters (detailed number of parameters for input shape of 32x32x3 and 10 classes is shown in Table 3). Although architectures such as ResNet50[19] and EfficientNet[20] have 23M and 4M parameters respectively (for input shape of 32x32x3 and 10 classes; their model summary is available in [Gravity optimizer GitHub repository](https://github.com/dariush-bahrami/gravity.optimizer)) and they are as easy to implement as VGGNet in Keras, they showed so much slower training speed than VGGNets in Google Colab.

Optimization, regardless of its application in deep learning, is utilized to minimize a function. This action of minimization is the parameter that should be used for comparing the performance of optimizers. The function that is tried to be minimized in deep learning is cost function. The parameter that should be used to compare optimizers in deep learning is the loss value in the training dataset. Therefore, to investigate the direct impact of the optimizer itself, we have using to use overfitting prevention techniques. Important examples of these techniques are learning rate decay, dropout, early stopping, batch normalization, and regularization. So another reason why we have chosen VGG architecture over other architectures is that it doesn’t use any overfitting prevention techniques.

Finally, we monitor loss and accuracy changes for training and validation datasets for a constant number of epochs (100 epochs) to compare the performance of Gravity optimizer with common standard optimizers listed in Table 1 (Where the same dataset and architecture is used without using overfitting prevention techniques. Table 3 summarizes the models used in this paper.

The remarkable thing about Gravity optimizer is that there was no need to tune hyperparameters to get better results. The same values were considered in all benchmarks. In Section 2, we talked about why we designed them in that way and how to find the best values for them. Our recommended value for Gravity optimizer hyperparameters was:

learning rate = 0.1 , Alpha = 0.01 , Beta = 0.9.

We also set these values as default for Gravity optimizer in python implementation. In this section, we use these suggested values for the benchmark.

To summarize, the results obtained from the training of five standard datasets mentioned in [section 3.2](#_3.2_Datasets) on VGGNet architectures (VGG16 and VGG19) using Gravity optimizers and two other standard and widely used optimizers (RMSProp and Adam) are compared in each subsection of datasets. As mentioned, all the trainings here are done with a batch size of 128 and for 100 epochs.

In the following subsections, the results (last epoch and best epoch) were obtained from training our target datasets on VGG16 and VGG19 using Adam, RMSProp, and Gravity optimizers without using any overfitting prevention techniques are compared together. Also, the results are compared with the results reported from the other papers which used the same datasets and architectures we used here. More details of the results can be found in [Gravity optimizer GitHub repository](https://github.com/dariush-bahrami/gravity.optimizer) or materials section.

Table 3. VGG16 and VGG19 model summary used in the paper

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **VGG16** | | | **VGG19** | | |
| **Layer Type** | **Output Size** | **Parameters#** | **Layer Type** | **Output Size** | **Parameters#** |
| **Convolution Part** | | | **Convolution Part** | | |
| Input Layer | 32, 32, 3 | 0 | Input Layer | 32, 32, 3 | 0 |
| Conv2D | 32, 32, 64 | 1,792 | Conv2D | 32, 32, 64 | 1,792 |
| Conv2D | 32, 32, 64 | 36,928 | Conv2D | 32, 32, 64 | 36,928 |
| MaxPooling2D | 16, 16, 64 | 0 | MaxPooling2D | 16, 16, 64 | 0 |
| Conv2D | 16, 16, 128 | 73,856 | Conv2D | 16, 16, 128 | 73,856 |
| Conv2D | 16, 16, 128 | 147,584 | Conv2D | 16, 16, 128 | 147,584 |
| MaxPooling2D | 8, 8, 128 | 0 | MaxPooling2D | 8, 8, 128 | 0 |
| Conv2D | 8, 8, 256 | 295,168 | Conv2D | 8, 8, 256 | 295,168 |
| Conv2D | 8, 8, 256 | 590,080 | Conv2D | 8, 8, 256 | 590,080 |
| Conv2D | 8, 8, 256 | 590,080 | Conv2D | 8, 8, 256 | 590,080 |
| MaxPooling2D | 4, 4, 256 | 0 | Conv2D | 8, 8, 256 | 590,080 |
| Conv2D | 4, 4, 512 | 1,180,160 | MaxPooling2D | 4, 4, 256 | 0 |
| Conv2D | 4, 4, 512 | 2,359,808 | Conv2D | 4, 4, 512 | 1,180,160 |
| Conv2D | 4, 4, 512 | 2,359,808 | Conv2D | 4, 4, 512 | 2,359,808 |
| MaxPooling2D | 2, 2, 512 | 0 | Conv2D | 4, 4, 512 | 2,359,808 |
| Conv2D | 2, 2, 512 | 2,359,808 | Conv2D | 4, 4, 512 | 2,359,808 |
| Conv2D | 2, 2, 512 | 2,359,808 | MaxPooling2D | 2, 2, 512 | 0 |
| Conv2D | 2, 2, 512 | 2,359,808 | Conv2D | 2, 2, 512 | 2,359,808 |
| MaxPooling2D | 1, 1, 512 | 0 | Conv2D | 2, 2, 512 | 2,359,808 |
| **Dense Part** | | | Conv2D | 2, 2, 512 | 2,359,808 |
| Flatten | 512 | 0 | Conv2D | 2, 2, 512 | 2,359,808 |
| Dense | 4096 | 2,101,248 | MaxPooling2D | 1, 1, 512 | 0 |
| Dense | 4096 | 16,781,312 | **Dense Part** | | |
| Dense | 10 | 40,970 | Flatten | 512 | 0 |
|  |  |  | Dense | 4096 | 2,101,248 |
|  |  |  | Dense | 4096 | 16,781,312 |
|  |  |  | Dense | 10 | 40,970 |
| **Total Parametrs = 33,638,218** | | | **Total Parametrs = 38,947,914** | | |

Table 4 shows a detailed summary of learning rate values used in runs. For Adam optimizer we turned off learning decay (decay = 0) and set beta 1 = 0.9, beta 2 = 0.999, and epsilon = 1.0 e-07. For RMSProp we turned learning rate decay, momentum, and centered off (decay = momentum = centered = 0) and set rho = 9.0 e-01, and epsilon = 1.0 e-07.

Table 4. summary of learning rates used for benchmark

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **VGG16** | | **VGG19** | |
|  | Adam | **RMSProp** | **Adam** | **RMSProp** |
| MNIST | 2.50e-04 | 1.00e-04 | 2.50e-04 | 2.50e-05 |
| Fashion-MNIST | 2.50e-04 | 5.00e-05 | 1.00e-05 | 5.00e-05 |
| CIFAR-10 | 1.00e-04 | 5.00e-05 | 1.00e-04 | 2.50e-05 |
| CIFAR-100 (coarse) | 1.00e-04 | 2.50e-04 | 7.50e-05 | 1.00e-04 |
| CIFAR-100 (fine) | 1.00e-04 | 1.00e-04 | 5.00e-05 | 1.00e-04 |

**4. Results**

**4.1 MNIST**

**4.2 Fashion-MNIST**

**4.3 CIFAR-10**

**4.4 CIFAR-100 (Coarse)**

**4.5 CIFAR-100 (Fine)**

**5. Conclusion**

**References**

[1] D. Choi, C. J. Shallue, Z. Nado, J. Lee, C. J. Maddison, and G. E. Dahl, “On empirical comparisons of optimizers for deep learning,” *arXiv Prepr. arXiv1910.05446*, 2019.

[2] D. Soydaner, “A Comparison of Optimization Algorithms for Deep Learning,” *Int. J. Pattern Recognit. Artif. Intell.*, p. 2052013, 2020.

[3] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, “The marginal value of adaptive gradient methods in machine learning,” in *Advances in neural information processing systems*, 2017, pp. 4148–4158.

[4] F. Schneider, L. Balles, and P. Hennig, “DeepOBS: A deep learning optimizer benchmark suite,” *arXiv Prepr. arXiv1903.05499*, 2019.

[5] H. Robbins and S. Monro, “A stochastic approximation method,” *Ann. Math. Stat.*, pp. 400–407, 1951.

[6] G. Hinton, N. Srivastava, and K. Swersky, “Neural networks for machine learning,” *Coursera, video Lect.*, vol. 264, no. 1, 2012.

[7] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv Prepr. arXiv1412.6980*, 2014.

[8] B. T. Polyak, “Some methods of speeding up the convergence of iteration methods,” *USSR Comput. Math. Math. Phys.*, vol. 4, no. 5, pp. 1–17, 1964.

[9] J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization.,” *J. Mach. Learn. Res.*, vol. 12, no. 7, 2011.

[10] M. D. Zeiler, “Adadelta: an adaptive learning rate method,” *arXiv Prepr. arXiv1212.5701*, 2012.

[11] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, “On the importance of initialization and momentum in deep learning,” in *International conference on machine learning*, 2013, pp. 1139–1147.

[12] T. Dozat, “Incorporating nesterov momentum into adam,” 2016.

[13] J. R. Sashank, K. Satyen, and K. Sanjiv, “On the convergence of adam and beyond,” 2018.

[14] E. Bisong, “Google Colaboratory,” in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, Springer, 2019, pp. 59–64.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[16] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms,” *arXiv Prepr. arXiv1708.07747*, 2017.

[17] A. Krizhevsky, G. Hinton, and others, “Learning multiple layers of features from tiny images,” 2009.

[18] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv Prepr. arXiv1409.1556*, 2014.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] M. Tan and Q. V Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” *arXiv Prepr. arXiv1905.11946*, 2019.