Study on the optimization algorithm of Adapg(Adaptive Gradient)

Mouna Lamine[1],Sang-Chul Kim[2]

Kookmin University

Abstract

Machine learning optimization is the process of iteratively enhancing the accuracy of a machine learning model, reducing the level of error. The most commonly used optimization algorithms in machine learning are first-order algorithms based on gradient descent. thus, during the past few years, a lot of work on improving the optimization methods in machine learning have been proposed successively. In this paper we are going to introduce Adapg, a new extension of the first-order optimization Algorithm.

Keywords: Machine Learning, Optimization Methods, gradient decent, AdaDelta,Adam

1 Introduction

Machine learning aims to extract valuable information from a huge amount of raw data. If carried out correctly, machine learning can act as a solution to a range of issues. Optimization is regarded as a basic function in machine learning as it goes through every step of machine learning in order to improve the accuracy of predictions and minimises errors. This article aims to present to the reader a new extension of the first-order optimization algorithm . The remainder of this paper is structured as follows. The first section presents the related work. The second section presents our proposed algorithm, and we conclude the whole paper in the last section.

2 Related work

The authors of[3] compared the different first-order optimization algorithms and introduced **Adapg** as an algorithm that Combines both **AdaDelta** and **Adam**. In this section we are going to introduce both AdaDelta and Adam algorithm.

2.1 AdaDelta Algorithm

AdaDelta is an extension of AdaGrad Algorithm which aims to solve the decay of learning rate issue.Instead of accumulating all the gradients, Adadelta bounds the accumulation of all the past gradients to a fixed size.

The **AdaDelta** [4] Algorithm is defined by :

Algorithm 1 Computing AdaDelta update at time t

Require: Decay rate ρ , constant ϵ
Require: Initial parameter *x1*,
$E[g^2]0 = x^n, E[\Delta x^2] = 0$
for t=1 : T **do**
$g_t \leftarrow \nabla f_t(\theta_t - 1)$
$E[g^2]t \leftarrow \rho * E[g^2]_{t-1} + (1-\rho) * g_t^2$
$\Delta x_t = -\frac{RMS[\Delta x]_{t-1}}{RMS[g]_t} gt$
$E[\Delta x^2]_t = \rho * E[\Delta x^2]_{t-1} + (1-\rho)\Delta x^2 t$
$x_{t+1} = x_t + \Delta x_t$
end

2.2 Adam Algorithm

[1] **Adam** is a replacement optimization algorithm for stochastic gradient descent for training deep learning models. which combines the properties of both AdaGrad and RMSProp algorithms in order to handle sparse gradients on noisy problems.

The **Adam** [2] Algorithm is defined by :

Algorithm 2 Adam,algorithm for stochastic optimization.

Require: α : *Stepsize*
Require: $\beta 1, \beta 2$: Exponential decay rates for the moment estimates
Require: $f(\theta)$ Stochastic objective function with parameters θ
Require: θ_0 Initial parameter vector
$m_0 \leftarrow 0$
$v_0 \leftarrow 0$
$t \leftarrow 0$ (Initialize timestep)
while θ_0 not converged **do**
$t \leftarrow t + 1$
$g_t \leftarrow \nabla f_t(\theta_t - 1)$
$m_t = \beta 1 * m_{t-1} + (1-\beta 1) g_t$
$v_t = \beta 2 * v_{t-1} + (1-\beta 2) g_t^2$
$\hat{m}_t = \frac{m_t}{(1-\beta_1^t)}$
$\hat{v}_t = \frac{v_t}{(1-\beta_2^t)}$
$\theta_t = \theta_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$
end
return θ

3 Proposed Algorithm

Adapg is an extension of the adaptive algorithm family, It is a combination of AdaDelta and Adam algorithms. This Algorithm have the purpose to increase the performance of the mentionned algorithms. Using the core update rule of the Adam algorithm, Adapg replaces Adam's first order momentum variable by the update variable of AdaDelta.

$$E[\Delta\theta_t^2] = \rho * E[\Delta\theta^2]_{t-1} + (1-\rho)\Delta\theta^2 t \quad (1)$$

According to this mathematical formula the new algorithm is described as bellow :

Algorithm 3 Adapg : Adaptive Gradient

Require: α : *Stepsize*
Require: ϵ : *constant*
Require: $\beta 1, \beta 2$
Require: $f(\theta)$ θ
Require: θ_0
$E[g^2]0 \leftarrow x^n$, $m_0 \leftarrow 0$, $v_0 \leftarrow 0$, $t \leftarrow 0$
while $\theta 0$ not converged **do**
$t \leftarrow t + 1$
$g_t \leftarrow \nabla f_t(\theta_t - 1)$
$E[g^2]t = \beta 1 * E[g^2]_{t-1} + (1 - \beta 1) * g_t^2$
$\Delta\theta_t \leftarrow -\frac{RMS[\Delta\theta]_{t-1}}{RMS[g]_t} gt$
$m_t = \beta 1 * m_{t-1} + (1 - \beta 1)\Delta\theta^2 t$
$v_t = \beta 2 * v_{t-1} + (1 - \beta 2) g_t^2$
$\hat{m}_t = \frac{m_t}{(1-\beta_1^t)}$
$\hat{v}_t = \frac{v_t}{(1-\beta_2^t)}$
$\theta_t = \theta_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$
end
return θ

Where :

$$\rho = \beta 1 = 0.9, \beta 2 = 0.999, \epsilon = 10^{-}8$$

4 Conclusion

In this paper, we introduced Adapg «Adaptive Gradient » the new extension of the Adaptive optimization methods.Our main Aim in our next research paper is to present the final mathematical representation of this algorithm, it Source code and analyse it performance metrics in comparaison to other optimazition algorithms.

Acknowledgement

* This work was carried out as a result of the research result of the SW-centered university project of the Ministry of Science and ICT and the Information and Communication Planning and Evaluation Institute in 2022 (2022-00964), and the University ICT Research Center Fostering Support Project of the Ministry of Science and ICT and the Information and Communication Planning and Evaluation Institute (ITP-2022-2018-0-01396). This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICT Challenge and Advanced Network of HRD program (2020-0-01826)

References

[1] Jason Brownlee. *Gentle Introduction to the Adam Optimization Algorithm for Deep Learning.* URL: `https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning`.

[2] Jimmy Lei Ba Diederik P. Kingma. *ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION.*), 2017.

[3] Kyrylo Oliinyk1 Oleg Rudenko1 Oleksandr Bezsonov1. *First-Order Optimization (Training) Algorithms in Deep Learning.* The 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020), 2020.

[4] Matthew D. Zeiler1. *ADADELTA: AN ADAPTIVE LEARNING RATE METHOD.*), 2012.