Adam Optimization

Submitted by:

Sparsh Gupta || *Tanmay Singh*

170001049 || 170001051

Computer Science and Engineering

3rd year

Under the Guidance of

Dr. Kapil Ahuja

Department of Computer Science and Engineering

Indian Institute of Technology Indore

Autumn 2019

Implementing Adam Optimization, Improved Adam Optimization (AADAM) and to implement Hybrid of Adam and SGD Optimization.

Introduction

Stochastic gradient-based optimization is of core practical importance in many fields of science and engineering. Many problems in these fields can be cast as the optimization of some scalar parameterized objective function requiring maximization or minimization with respect to its parameters. If the function is differentiable w.r.t. its parameters, gradient descent is a relatively efficient optimization method, since the computation of first-order partial derivatives w.r.t. all the parameters is of the same computational complexity as just evaluating the function. Often, objective functions are stochastic.

Adam Optimization:

Adam, a method for efficient stochastic optimization that only requires first-order gradients with little memory requirements. The method computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients; the name Adam is derived from adaptive moment estimation.

Algorithm 1: *Adam*, our proposed algorithm for stochastic optimization. See section 2 for details, and for a slightly more efficient (but less clear) order of computation. g_t^2 indicates the elementwise square $g_t \odot g_t$. Good default settings for the tested machine learning problems are $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. All operations on vectors are element-wise. With β_1^t and β_2^t we denote β_1 and β_2 to the power t.

Require: α: Stepsize
Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates
Require: $f(\theta)$: Stochastic objective function with parameters θ
Require: θ_0: Initial parameter vector
  $m_0 \leftarrow 0$ (Initialize 1st moment vector)
  $v_0 \leftarrow 0$ (Initialize 2nd moment vector)
  $t \leftarrow 0$ (Initialize timestep)
  while θ_t not converged **do**
    $t \leftarrow t + 1$
    $g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep t)
    $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update biased first moment estimate)
    $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)
    $\widehat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ (Compute bias-corrected first moment estimate)
    $\widehat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ (Compute bias-corrected second raw moment estimate)
    $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \widehat{m}_t / (\sqrt{\widehat{v}_t} + \epsilon)$ (Update parameters)
  end while
  return θ_t (Resulting parameters)

A hybrid strategy to start training with Adam and switch to SGD at the right time. In the learning task with a large output space. It was observed that Adam could not converge to an optimal solution. SGD performs well empirically across a variety of applications and also has strong theoretical foundations.

Mentioned version of the Adam, i.e. A-Adam optimization algorithm, is able to improve the speed of convergence and finds a better minimum for the loss function compared to the original algorithm.

Project Objective

- To Acquainting with and understanding Adam optimization and SGD.
- Implementing Adam optimization and observing convergence.
- Implementing improved Adam optimization.
- Implementing Hybrid of Adam and SGD.

References

- https://www.researchgate.net/publication/334398886_Workshop_track_-ICLR_2018_IMPROVING_ADAM_OPTIMIZER
- https://www.researchgate.net/publication/329039554_An_Optimization_Strategy_Based_on_Hybrid_Algorithm_of_Adam_and_SGD
- https://arxiv.org/pdf/1412.6980.pdf
- https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/