Pattern Recognition Letters 128 (2019) 197–203

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Barzilai–Borwein-based adaptive learning rate for deep learning

Jinxiu Liang[a], Yong Xu[a,b], Chenglong Bao[c], Yuhui Quan[a,*], Hui Ji[d]

[a] South China University of Technology, Guangzhou 510006, China
[b] Peng Cheng Laboratory, Shenzhen 510852, China
[c] Tsinghua University, Beijing 100084, China
[d] National University of Singapore, Singapore 117543, Singapore

ARTICLE INFO

Article history:
Received 21 June 2019
Revised 28 August 2019
Accepted 29 August 2019
Available online 30 August 2019

Keywords:
Barzilai–Borwein method
Deep neural network
Stochastic gradient descent
Adaptive learning rate

ABSTRACT

Learning rate is arguably the most important hyper-parameter to tune when training a neural network. As manually setting right learning rate remains a cumbersome process, adaptive learning rate algorithms aim at automating such a process. Motivated by the success of the Barzilai–Borwein (BB) step-size method in many gradient descent methods for solving convex problems, this paper aims at investigating the potential of the BB method for training neural networks. With strong motivation from related convergence analysis, the BB method is generalized to adaptive learning rate of mini-batch gradient descent. The experiments showed that, in contrast to many existing methods, the proposed BB method is highly insensitive to initial learning rate, especially in terms of generalization performance. Also, the BB method showed its advantages on both learning speed and generalization performance over other available methods.

1. Introduction

1.1. Background

In the last decade, deep learning has emerged as one leading machine learning tool in computer vision. Particularly, deep neural network (DNN) based learning, including supervised approaches [1,2] and unsupervised approaches [3,4], has been used for solving many long-lasting problems in computer vision with remarkable success, e.g. image classification, action recognition and semantic segmentation. DNN-based learning enables an artificial neural network (NN) to capture intricate structures of visual data with multiple levels of abstraction.

In DNN, once the architecture of an NN has been designed for solving a specific problem, the remaining task is to learn or train the weights of the NN. In the so-called supervised approach to NN training, the weights of an NN are adjusted with respect to input data (i.e. training samples) such that the error between the output of the NN and the preferred output is minimized. More specifically, Let θ denote the set of the weights of an NN. Consider a training set $\{(x_i, y_i)\}_{i=1}^N$ containing N samples, where x_i denotes the input data and y_i denotes its preferred output. The learning process is then to estimate θ that minimizes the following cost function:

$$\min_{\theta} L(\theta) = \frac{1}{N}\sum_{i=1}^{N} L_i(\theta), \tag{1}$$

where $L_i(\theta) = L(x_i, y_i; \theta)$.

An efficient and effective method to find a good solution of the problem (1) is critical to the success of DNN. Unfortunately, the problem (1) is often a very large-scale non-smooth and non-convex problem. For such a large-scale problem, first-order methods such as gradient descent are usually preferred. Among them, the so-called *stochastic gradient descent* (SGD) method is dominant in NN training. Instead of using the batch gradient which leverages over the cost gradients of all training samples $\frac{1}{N}\sum_{i=1}^{N}\nabla_\theta L_i(\theta)$, classic SGD methods only call the cost gradient of one sample, which could be overly noisy. Therefore, a prominent approach is using the so-called *mini-batch gradient descent* method [5], which uses a small portion of training samples for gradient estimation. The update of a mini-batch gradient descent method reads as follows:

$$\theta_{t+1} = \theta_t - \frac{\eta_t}{|\mathcal{B}_t|}\sum_{i\in\mathcal{B}_t}\nabla_\theta L_i(\theta_t), \tag{2}$$

where θ_t denotes the estimate at Iteration t, $\mathcal{B}_t$ denotes the index set of the samples randomly chosen from the training set at Iteration t, $|\mathcal{B}_t|$ denotes the cardinality of the set $\mathcal{B}_t$, and the value $\eta_t > 0$ is called *learning rate*. Mini-batch gradient descent is of core practical importance to NN training. In practice, the gradient

* Corresponding author.
E-mail address: csyhquan@scut.edu.cn (Y. Quan).

https://doi.org/10.1016/j.patrec.2019.08.029
0167-8655/