Computers and Electrical Engineering 000 (2017) 1–12

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Hybrid optimization strategy of 'Network-in-Network' model for image recognition☆

Xiaojun Bi, Tiewen Pan*, Yun Lin

Harbin Engineering University, Nantong Avenue 145-1, Harbin, Heilongjiang 150001, China

A R T I C L E I N F O

Article history:
Received 9 August 2017
Revised 15 December 2017
Accepted 15 December 2017
Available online xxx

Keywords:
Hybrid information systems
Deep learning
Network in Network
Particle swarm optimization
Image recognition

A B S T R A C T

In reality, an information processing system needs to receive data from various sensors. This is called a hybrid information system. Image recognition is widely applied to hybrid information systems. Recently, Network in Network (NIN) as a special deep learning model has obtained impressive results on image recognition. However, there is a lot of room for performance improvement, since gradient-based optimization methods may fall into local minima. To get a high recognition rate, it is essential to explore the solution space further. In this paper, we propose a hybrid optimization method for NIN in order to improve the image recognition rate. A fine-tuning strategy was employed to train NIN by using particle swarm optimization followed by a pre-training stage based on the gradient method. Several strategies were introduced in order to avoid the deterioration of the solution. The results showed that the present method can effectively increase the image recognition rate and be further used in a hybrid information system.

1. Introduction

In recent years, with the development of all kinds of sensors in intelligent transportation, medical treatment, payment, security, and other hybrid information systems, image recognition has become an important research focal point [1–3]. Deep Learning models simulate how the brain processes information, and have gained much attention in the past few years, since they can extract highly effective features, and especially in image recognition, text classification and other relevant fields [4,5]. As an important deep learning model, NIN has achieved impressive performance in image recognition [6], in comparison to other convolutional neural networks (CNNs), and has also heavily influenced the later models, such as VGGs, and GoogleNet [7] by the idea that a 1×1 convolution operation can enhance the capability of feature extraction. In many fields, such as object recognition, it has achieved good performance.

Similar to other deep learning models, gradient descent methods can be used to optimize the NIN's non-convex object function. However, gradient-Based methods [8,9] for non-convex, high-dimensional and nonlinear object function optimization fall into local minima easily because of their higher sensitivity to initial points and learning schedules. Thus, NIN, which was only trained by gradient-based methods, still requires a lot of improvement and further exploration of the solution space is required. The gradient-based pre-training and gradient-based fine-tuning paradigms, proposed by Hinton [10] showed limited improvement, because deterministic gradients lacked exploration diversity with regard to the solution space.

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. S. Liu.

* Corresponding author at: Harbin Engineering University, Room 1003, Building 1, Nantong Avenue 145-1, Harbin, Heilongjiang 150001, China.
E-mail address: pantiewen@hrbeu.edu.cn (T. Pan).

https://doi.org/10.1016/j.compeleceng.2017.12.033

Recently, the Particle Swarm Optimization (PSO) evolutional algorithm has been widely used in the search of high-dimensional global minima and nonlinear object functions. It obtained very promising results in several applications [11]. The Particle swarm optimization algorithm has the characteristics of few adjustment parameters, fast convergence speed, and high convergence precision. To the best of our knowledge, there has been no previous study on the combination of PSO and NIN until now. Motivated by pre-training and fine-tuning paradigms proposed by Hinton, a training method is proposed in this paper, and includes two stages: namely, the gradient-based method for pre-training, and the PSO-based method for fine-tuning. Unlike Hinton's method, fine-tuning employs the PSO evolutional algorithm instead of the gradient-based method. The first stage aims to search for a good point of PSO initial population. Several tricks are employed in PSO in order to improve convergence. Several experiments demonstrate that the proposed method is superior to several state-of-the-art methods.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 surveys the PSO algorithm; Section 4 describes "Network in Network"; Section 5 presents the proposed method; Section 6 presents the experimental results. Finally Section 7 concludes this paper.

2. Related work

Gradient descent algorithms are by far the most common approach to neural network optimization. Thus far, various gradient descent methods have been proposed for achieving high optimization performance, such as the Nesterov Accelerated Gradient (NAG), Adagrad, Adadelta, RMSprop, and Adaptive Moment Estimation (Adam) [8,9]. These methods are not guaranteed to converge to global minima and suffer from carefully tuned hyperparameters.

Recently, some of the evolutional algorithms have been used in order to optimize the deep learning (DL) model. Numerous studies [12–14] have demonstrated that these methods for improving CNN accuracy can be directly employed evolutional algorithms in order to optimize parameters; however, they can be only applied in small scale DL models. If DL models have many parameters, the object function is high dimensional, and its search space is quite huge and requires that many candidate particles participate in the evolution. Additionally, the computation is extremely costly.

3. Particle swarm optimization algorithm

3.1. Basic theory

Kennedy and Hart Erber proposed PSO, which is widely used in parallel computing technology for investigating the evolution of social behavior [11]. A random candidate solution group, conceptualized as a particle, is initialized for the PSO algorithm. Each particle moves iteratively in the problem space with random velocity. Attracted to the best adaptability, the best adaptive location is achieved on the whole population (global version of the algorithm). The basic PSO evolutional algorithm is described by following equations:

$$\mathrm{v}(\mathrm{k}+1) = \alpha \otimes \mathrm{v}(\mathrm{k}) + \beta_1 \otimes \mathrm{r}_1(\mathrm{p}_1 - \mathrm{x}_\mathrm{k}) + \beta_2 \otimes \mathrm{r}_2(\mathrm{p}_2 - \mathrm{x}_\mathrm{k}) \tag{1}$$

$$\mathrm{x}(\mathrm{k}+1) = \beta_3 \otimes \mathrm{x}(\mathrm{k}) + \beta_4 \otimes \mathrm{v}(\mathrm{k}+1) \tag{2}$$

Where, the symbol $\otimes$ represents a one-by-one vector multiplication with a broadcasting mechanism. The symbol k represents any iteration. The velocity v (k) depends on its current value, which is affected by the momentum factor α, and on the term updating the best positions: $\mathrm{p}_1$represents the previous best position and p_2 represents the best position in the entire swarm. The coefficients β_1 andβ_2 denote the strength of the attraction. The particle position x (k) is based on its current value and the newly updated velocity v (k + 1), which is affected by coefficients β_3 and β_4 respectively. It was shown that β_3 and β_4 can be set to unity without loss of generality. The vectors of random numbers r_1 and r_2 provide a way to randomly search for a good state space exploration [12]. The random number is usually in the range of [0, 1] in uniform:r_1, r_2 and Uniform[0,1].

3.2. Explanation of PSO

PSO converges to global minima relying on the exploration of populations. At iteration k + 1, the position x (k + 1) of each particle needs to learn information, which actually is called gradient from previous position and global best position. As depicted in Fig. 1, and assuming that search space is two dimensional, the *i-th* particle gets the gradient from population evolution information. This is quite different to gradient-Based Method, where the gradient information comes from a single point rather than from the population. Therefore, the exploration capability of the solution space is superior to gradient-Based method.

4. Network in Network

NIN, which was proposed by Min Lin [6], introduced the MLP convolutional layer and the global averaging pooling layer in order to enhance the capability of feature extraction and regularization. The details associated with NIN are described below.