Research and Design of Neural Network Based on GPU

Jiaohua Yang[1, a,*] and Xin Wang[2]

[1]Jilin Engineering Normal University, Changchun Jilin 130052,China;

[2]Shenyang Jianzhu University. Shenyang, China

[a]2546089800@qq.com,

*corresponding author

Keywords: GPU; Artificial neural network; Graphics processor

Abstract. With the rapid development of graphics processor (GPU) programmable ability, coupled with its high speed and parallelism for large scale neural network BP algorithm and the problem of low efficiency, people put forward a kind of neural network BP algorithm based on GPU acceleration. Through in-depth study of the CUDA technology programming and programming model to solve the complicated problem of parallel computing using this framework, converse the process of the BP neural network in CPU forward calculation and reverse learning to the process of learning accelerated in the GPU, and then use the GPU powerful floating-point computation ability and high parallel computing characteristics to achieve BP algorithm. Finally people design and achieve a kind of neural network based on the GPU acceleration training.

Introduction

Artificial neural network (Artificial Neural Network, referred to as ANN), neural network (Neural Network, NN) get more and more widely application in various fields. ANN is a massively parallel processor interconnection, it can solve complex optimization problems by highly interconnected neurons. Artificial neural network has been widely used in various fields, and achieved the remarkable progress. Its application scope has the following main areas, automatic control, solution of combination optimization problem, pattern recognition, image processing, robot control and medical treatment. In terms of pattern recognition, digital image processing, the application of artificial neural networks grow with each passing day. [1]

The rapid development of Computer graphics processor (Graphics Processing, Unit, GPU) not only promoted the rapid development of the image processing, virtual reality and computer simulation applications' area, but also provide a good platform for general-purpose computing outside people take advantage of GPU to carry the process of graphics.

Generally speaking, when researchers study the neural network algorithm, they adopt serial common algorithm, namely CPU arithmetic. Although CPU is growing, the use of artificial neural network algorithm takes researchers a lot of time and effort in ordinary PC machine because the neural network computing parallel computing ability.

Therefore, a Compute Unified Device Architecture called CUDA [2] appeared in 2006, this architecture can use GPU to solve the complicated problem of parallel computing. GPU is the graphics processor, it is the graphics' "heart", similar to CPU, but GPU is designed for performing mathematical and geometric calculation. This study of the task is to bring the task of multiple complex matrix computing to the GPU and it can reduce the load of CPU largely and give other processes more opportunities to improve the resource utilization and system throughput.

Artificial neural network and its characteristic

Artificial neural network is composed of a large number of neurons which are interconnected. Network use the computer to simulate the thought of human behavior in order to achieve information storage and information processing. Based on this complex network system, the

nonlinear operation of neural network simulation get a good effect. Take use of neurons to store large amounts of data, the learning function of neural network is very good especially on the simulation of the vast data. In every aspect of our life, people can solve difficult problems by using artificial neural network, especially the use of prediction and pattern recognition function of neural network.

The neuron model is a simulation based on the structure of the human brain neurons, as shown in Figure 1, this model is an input, output and calculation model. The input can be analogous to the dendrites of neurons, and the output can be likened to axons and the computing can be analogous to the nucleus. Connection is the most important link in neurons things. Each connection has a weight training algorithm. A neural network's training algorithm is to let the value of the weight adjusted to the bestso that the forecasting effect of the whole network to be the best. The artificial neural network consists of a large number of neurons, neuron model as shown in figure 1:

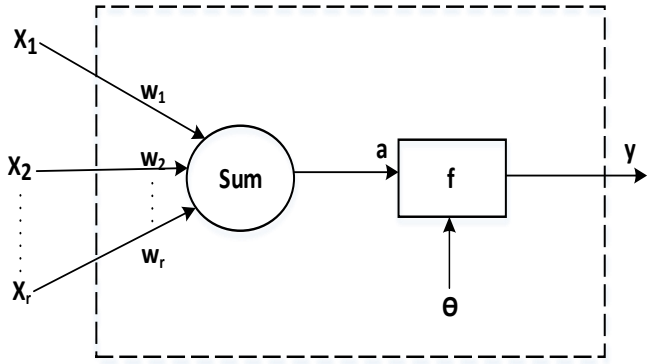

Figure 1. Model map of neuron structure

Among them, $X=\{x_i \mid i=1,2,\ldots,r\}$ as the input vectors of neurons, so neurons input r vector weights, $W=\{w_i \mid i=1,2,\ldots,r\}$ is the weights for each input vector of a neuron, θ is bias vector, f as activation function, also known as the transfer function, usually as a nonlinear function, and y is Output for the neuron operation. The mathematical formula for computing the output representation formula shown at 1:

$$y = f(WA^T + \theta) \quad (1)$$

In this, AT is a transposition of the A vector.

The learning rules of artificial neural networks

The neural network learning, namely the training process, refers to the input layer neurons receiving input information, pass to the middle layer neurons, finally transferred to the output layer neurons and then output information from the output layer to deal with the problems. In this process, the neural network achieve the purpose of the learning, training by adjusting the network weights and threshold. After the error of the network output reduce to an acceptable level, or make a predetermined number of learning, learning can be stopped. Learning ability is an important standard to evaluate the neural network, which is determined by the neural network topology and connection weights[5].

The main learning methods of neural network are tutor learning, no tutor learning and reinforcement learning[6].

BP neural network

BP neural network (Back-Propagation Neural Network) was proposed by Geoffrev, David

Rnmelhart and some others in the mid 1980s. It is one of the most commonly used artificial neural network, and it has the essence of the theory of artificial neural networks. Due to its simple structure, plasticity strong features, it has a wide fields in pattern recognition, model construction and information classification and other fields. In the practical application of the artificial neural network, BP neural network used in the proportion as high as 80%.BP neural network has strong flexibility, it transfers information by the S function, and it belongs to the muti-layerforward neural network[3].

CUDA (Computer Unified Device Architecture) is a general-purpose GPU computing product in NVIDIA company launched in 2007.CUDA can effectively use the processing power and memory bandwidth of GPU graphics rendering, which is widely used in image processing, signal processing, pattern recognition, numerical calculation, oil exploration, astronomical calculation, computational fluid dynamics, computational biology[4] etc.

Design and implementation of neural network based on GPU

Model establishment of neural network based on GPU

In this paper, the design of neural network based on GPU is selected as the basic model of the most used BP neural network.

The design process of BP neural network

The main steps of BP network model construction are sample input, network initialization, training, simulation, and error, accuracy and model output.

The establishment of BP neural network model

BP neural network is a supervised learning process which has the guided error back-propagation algorithm. The tutor specifies the desired output, when the actual output and the expected output is inconsistent and then get into the backpropagation stage of error. The error which is through the output layer modify the weights of each layer and counter to the input layer, hidden layer, layer by layer in the method of error gradient descent. The process of forward and back propagation of information is a process of adjusting the weights constantly , and the neural network training process. This process has been carried out the error of the network has been reduced to an acceptable level, or pre-set learning times so far.

Analysis of experimental results of neural network based on GPU

Experimental data source

The data source used in the experiment in this paper is the matrix data source which use the random number part. Main research direction of subject is displayed the accelerating effect in admiral GPU neural network, so the test data set to random numbers instead of images or text, so as to reduce other unrelated factors and that will make the experimental results more convincing. The choice is to provide some useful data sets with method of scikit-learn generator in the theano library, you can directly use the make_moons function definition of locally generated data sets without writing the container.

According to the influence of different parameters, a total of three parameter adjustment schemes were set up. The experiments were carried out in two environments of GPU and CPU respectively, based on the scale of data set, gradient descent parameter and training iteration and record the result of the experiment and analysis it.

Experimental data collection and analysis

Experimental results and analysis:

When take the number of data points (data set size) as independent variables to test different equipment running time, it can be seen from table 2, the recognition rate is in a small range of relative balance changes. There are significant differences between the running time of the CPU and GPU of different equipment. And according to line graph in figure 10and trend lines, it can be seen that the speedup ratio is improved obviously with the data set size continues to expand, the speedup

ratio and scale of data sets showed a positive trend, accelerating effect is more obvious. Therefore it can be concluded:

The acceleration ratio in Table 2 is still in the positive state that is far greater than 1, indicating that the acceleration effect is still obvious when the size of the dataset is used as the independent variable.

When the recognition rate is in a relatively stable condition, it can be seen that the size of the data set that composed of random numbers does not affect the result of the training.

Summary

This paper achieved neural network based on GPU , and experiment conclusion drawn on a given set of data also proved CUDA programming technology have an obvious accelerate training effect for BP neural network. But there is a shortcoming in the data set, the technology itself is lies in the application of research actually, the training data set and test data set is composed of random number, that is displayed in the accelerating effect of admiral GPU neural network, it can reduce other irrelevant factors and make the experimental results more convincing.

Acknowledgements

This research was supported by the Support Foundation of Shenyang Jianzhu University (No.2017034).

References

[1] D.J. Meng. Artificial neural network technology and application[J]. electronic technology and software engineering, 2016, (23): 16.

[2] F. Chen, Y.B. Tian and M. Yang. Research and implementation of CUDA based parallel particle swarm optimization algorithm and implementation of [J]. Computer science, 2014, (09): 263-268.

[3] Y.H. Zhang. Research on BP neural network based on genetic algorithm optimized car model [D]. Changan University, 2015.

[4] L. Shi. GPU general calculation virtualized method research [D]. Hu Nan University, 2012.

[5] H.J. Gao. Research on the application of inventory demand prediction for valve manufacturing enterprises based on BP neural network [D]. Beijing Jiaotong University, 2015.

[6] Wen wen. The qualified rate of the improved BP neural network prediction research based on product quality [D]. South China University of Technology, 2014.