Optimization Algorithm for Unsupervised learning using Hadoop datasets Based on Cloud computing

M. Kalimuthu (921817104013)
M. Mugesh Prasad (921817104024)
K. Vigneshwaran (921817104040)
Guided by: T. Vinothini AP/CSE
Sree Sowdambika College of Engineering
(Approved by AICTE, New Delhi & Afflicted to Anna University, Chennai.)

ABSTRACT

Big data has already occupied a lot in the information society. The application of big data to intelligent agriculture is the core development direction for elaboration of the utilization of agricultural data information, and the deep learning method can more effectively extract abstract information from big data which changes into more useful knowledge, thus supporting the development of intelligent agriculture from various angles. In this proposed paper, a **Convolution Neural Networks(CNN) Recurrent Neural Networks (RNN)** model is developed based on cloud computing technology. In parallel neural network model had spited different by training set is adopted to design the **batch gradient descent algorithm** based on deep unsupervised learning and **Back Propagation (BP algorithm)** based on Map-Reduce. This paper verifies the feasibility of deep unsupervised learning neural networks based on cloud computing and verifies that the optimization algorithm in the proposed paper can optimize the training efficiencyofneural network.

Keywords:Neural network; deep learning; unsupervised learning; cloud computing; big data

INTRODUCTION

Cloud Computing has become the buzzing topic of today's technology, driving mainly by marketing and services offered by prominent corporate organizations like Google, IBM & Amazon. "Cloud Computing" is a big deal, it is not. In reality, cloud computing is something that we have been using for a long time; it is the internet facility, along with the associated standards that provide a set of web-services to users. When users draw the term 'Internet' as a "cloud", they represent the essential characteristics of cloud computing. Two types of models in cloud computing they are: '**Deployment model, Service model".** Cloud Computing is a computing model. Its main function is to realize distributed computing of big data, parallel operation of a large number of operations, storage of massive data of the network, virtualization sharing of network resources, resource scheduling and load balancing and data redundancy storage based on hot backup.

Big Data is a term that is used for denoting the collection of datasets that are large and complex, making it very difficult to process using legacy data processing applications.

Hadoop is a cloud computing application software library for big data processing. It uses a Map-Reduce programming model and other mechanisms to perform distributed computing frameworks on large data sets to be processed by computer clusters. In the Map-Reduce computing mode, data calculation process is divided into two phases Map and Reduce. A. Batch gradient descent algorithm based on Map-Reduce for deep unsupervised learning the batch gradient descent algorithm is processed in Map-Reduce.

Map-Reduce in Hadoop is nothing but the processing model in Hadoop. The programming model of Map-Reduce is designed to process huge volumes of data parallelly by dividing the work into a set of independent tasks. As we learned in the Hadoop architecture, the complete job or work is submitted by the user to the master node, which is further divided into small tasks

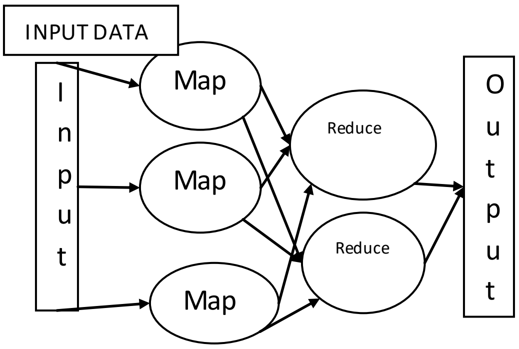

Diagram of Map-Reduce

CONSTRUCTION OF THE CNN MODEL:

Convolution Neural networks are designed to process data through multiple layers of arrays. This type of neural networks is used in applications like image recognition or face recognition. The primary difference between CNN and any other ordinary neural network is that CNN takes input as a two-dimensional array and operates directly on the images rather than focusing on feature extraction which other neural networks focus on.

This paper constructs the CNN model based on the advantages of Convolutional Neural Network (CNN) in large-scale image feature representation. A convolution is the simple application of a filter to an input that results in activation. Repeated application of the same filter to input results in a map of activations called a feature map, indicating the locations and strength of a detected feature in an input, such as an image.

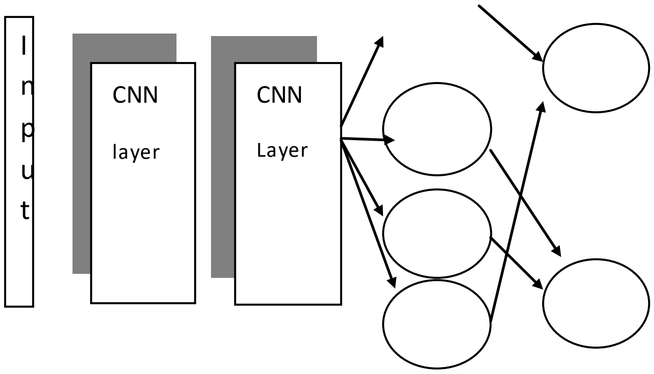

Diagram for CNN model

CONSTRUCTION OF THE RNN MODEL:

Recurrent neural networks are a type of deep learning-oriented algorithm, which follows a sequential approach. In neural networks, we always assume that each input and output is independent of all other layers. These type of neural networks are called recurrent because they perform mathematical computations in sequential manner.Because it is difficult to adopt RNN to achieve long-distance learning due to gradient divergence of it, Long-Short Term Memory (LSTM) is introduced. The features of the image are extracted by CNN, and the coding and decoding network is constructed by LSTM.

For the input features, the timing features are extracted at the encoding end and transmitted to the decoding end for decoding. In addition to learning historical information, RNN and LSTM can also be designed as two-way structures, i.e. bidirectional RNN, bidirectional LSTM, which is a good choice for speculating and completing information.

RNNs are of particular value in the IoT space, especially in time-correlated series of data, such as describing a scene in an image, describing the sentiment of a series of text or values, and classifying video streams. Data may be fed to an RNN from an array of sensors that contain a (time: value) tuple. That would be the input data to send to the RNN. In particular, such RNN models can be used in predictive analytics to find faults in factory automation systems, evaluate sensor data for abnormalities, evaluate timestamped data from electric meters, and even to detect patterns in audio data. Signal data from industrial devices is another great example. An RNN could be used to find patterns in an electrical signal or wave. A CNN would struggle with this use case. An RNN will run ahead and predict what the next value in a sequence will be if the value falls out of the predicted range that could indicate a failure or significant event

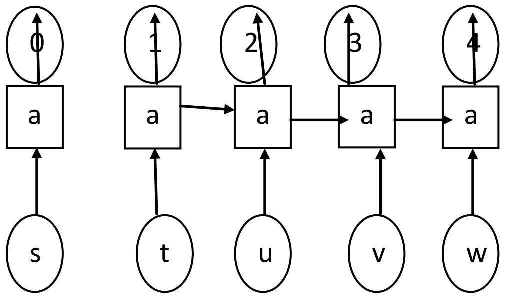

Diagram of RNN model

PARALLEL COMPUTING

Parallel Computing is an operational mechanism, including parallelization such as cloud computing cluster dealing with big Data.

It is the use of multiple processing elements simultaneously for solving any problem. Problems are broken down into instructions and are solved concurrently as each resource which has been applied to work is working at the same time.

Loss function:

$F=1/2M \sum_{i=1}^{M} (Y^{(i)} - a^{M} \text{x}, \text{P}))^2$

Optimization of batch gradient algorithm by:

$P^{M}_{uv} = M/N \sum^{M}_{i=1} (Y(k) - a^{2}(x(k),P)) \partial b^{M} / \partial p^{M}_{uv}$

Calculation error is obtained by:

$Ef(x) = 1/2M \sum_{r=1}^{s} (d_i^{\,r} - o_i^{\,r})^2$

Mapperexecutes:

$\sum_i y^{(i)}$— a^L $x^{(i)}, p^2$ to

$X^{(i)}, y^{(i)}$

Neuralnetworkparallelizationmodel:

In image recognition or speech recognition applications, the computational complexity of the deep neural network model is very large, and there is a certain data correlation between layers of the model. Therefore, how to split the task amount and computing resources is an important issue for the design of CPU or GPU cluster acceleration framework. When the train of a scale model is relatively large, the training of the model can be accelerated by the dataparallelism method.Basedoncomparisonofseveral types of neural network parallelization modes, by it is finallydetermined that the parallelization of neural networks is realizedbydividingthetrainingset(dataparallel).

Learningmechanism:

In the deep learning algorithm, we have learned pre-training through unsupervisedlearningisadoptedfirstbase on each layer. The Deepunsupervisedlearning isimplementedthrough the gradient descent algorithm. The main task of deep network learning is to solve the parameter p of the loss function and to optimize the gradient descent algorithm for bigdata.

Trainingmechanism

The batch training mechanism is adopted based on the training mechanism of neural network with **Back- Propagation (BP) algorithm**. The weight change is a very effective training method for the batch training mechanism of neural network. In the batch model training mechanism, error is obtained by the errorcalculation formula. In the batch training mechanism, the weight updates are accumulating after submitting the all training set samples. The batch training mechanism algorithm can be divided into three phases, namely the **prefix phase, error BP phase, and weight update phase.**

Map function:

In the Map function, the weight is read from Hadoop Distributed File System (HDFS) which initialize the network. The samples are segmented and a certain number of network trainings is performed to achieve certain conditions. The output of the Map is a set of <Long, Writable> key-value pairs. Reduce takes the key-value pairs and processes them, defining an interface class – WeightWritable, which allows the Map function to operate through a writable interface. The class WeightWritable built according to the Hadoop serialization standard can save the change value of all weights. The training conclusions for eachbatch or round are sent to Reduce for synthesis, and the outputvaluekeyoftheLongWritableclassicsetto zero.

Reducefunction:

In the Reduce function, a set of <Long Writable, WeightWritable> key-value pairs output by the Map function is accepted and a set of <Long Writeable, Int Writable> key-value pairs is output. Reduce counts the weight of each Map and finds the average of all weights as the new weight. In WeightWritable, an accumulation function and a division function are implemented to add and divide the weight matrix to obtain an arithmetic average weight. Before training, Reduce reads the weight from HDFS and compares two weight records. When the difference of weights is lower than the requiredstandard, the output int valueissetto false(representedby0),otherwise it is set to True (represented by 1), and the weighted WeightWritable is adopted to update the weight in HDFS as theinitialweightofthenext iteration.

Combinefunction:

The role of the Combine function is to combine the local output of the Map to greatly reduce the data input/output (I/O) time. It inputs a set of Map output <Int Writable Weight Writable >key-value pairs and outputs content being same with the output type of the Map.

Drivingfunction:

The content input to the Driving function is an array of strings containing two strings, a path for the training sample and an output path for Hadoop operation. This function first determines whether there is a weight file on HDFS, and if not, generates a WeightWritable type file that is given a smaller random value (-0.5~0.5) as the initial weight of the network. Then the Driving function creates a Hadoop job based on the training sample path and output path of parameters, and controls the job operation iteration according to thereturnvalueoftheReduce function.

EXPERIMENTAL RESULTS AND ANALYSIS

Algorithm convergence speed and accuracy test:

The data set used in the experiment was derived from Breast Cancer Data Set. In the experiment, the training samples were divided into multiple parts and assigned to multiple Mappers for training and the Driving function is adopted to test network classification accuracy.

Experimentalresults:

The experiment shows the advantages of cloud computing clusters in big data processing and it shows that the bigger the data, the more obvious the advantages. It can be seen from the experimental results that the application of **BP algorithm based on Map-Reduce** can increase the training efficiency of neural network, which proves the feasibility of deep unsupervised learning neural network based on clouding computing. It is better than existing algorithms with better acceleration ratio, faster convergence and lessiteration.

CONCLUSIONS

Thus, the paper proposed of **Optimization of Unsupervised learning of Hadoop data sets using Mrjobs in cloud computing** with the conventional algorithms, BP algorithm based on Map-Reduce can increase the training efficiency of neural network,and it can proves the feasibility of deep unsupervised learning neural network based on clouding computing.

REFERENCES:

[1] Yuan Bingqing;Neural Network and Deep Learning Basics [J]. Radio WaveGuard,2018(05):32-33.

[2] Dean J,Corrado G S,Monga R,et al. Large scale distributed deep networks[C]//Proc of the 25th International Conference on Neural Information Processing Systems.[S.I.]: Curran Associates Inc,2012:1223-1231.

[3] Yong Peng Yue, Research on deep unsupervised learning algorithm [D].SouthwestPetroleumUniversity, 2015.

[4] YanHui,HuHaiyan.Researchandrealizationof ISIC-CDIO teaching experimental system based on RFID technology of web of things [J].JournalofBionanoscience.2013.12 (7):696-702.

[5] Jin Xu, Yafeng Yin, Hong Man,Feature Selection Based on Sparse Imputation,International Joint Conference on Neural Networks (IJCNN) 2012.

[6] HaiJunZhang, Parallel implementation and learning method of neural network based on Cloud Computing [D]. South China UniversityofTechnology,2015.

[7] Yan Hui, longDuo.3Dscanner-based corn seed modeling [J].Applied EngineeringinAgriculture, Vol.32(2).2016.3:18 1-18