An Analysis of Integrate Large Scale Deep Learning using Mobile Big Data

Rajesh Kumar Singh[1], Dr. Kalpana Sharma[2]
[1]Research Scholar, Deptt. Computer Application, Bhagwant University, Ajmer, Rajasthan
[2]Assistant Prof., Deptt. Computer Application, Bhagwant University, Ajmer, Rajasthan

Abstract

Deep learning methods have been applied extensively to learning methods in various fields of science and engineering such as speech recognition, image classification and language processing. Similarly, traditional data processing techniques have several limitations in processing large amounts of data. Furthermore, Big Data Analytics requires new and sophisticated algorithms based on machine and deep learning techniques to process the data in real time with high accuracy and efficiency. However, recently, research has included various intensive learning techniques with training mechanisms of hybrid learning and processing data with high speed. Most of these techniques are specific to scenarios and thus are based on vector space, reflecting poor performance in common scenarios and learning characteristics in big data. Furthermore, one of the reasons for such failure is the high involvement of humans to design sophisticated and optimized algorithms based on machine and intensive learning techniques. This paper describes a development environment integrating big data architecture and deep learning models to facilitate rapid experimentation. This paper makes three major contributions: first, it describes a big data architecture supporting an organization that supports large data collection and deep learning models and, second, it is used to build the data visualization described. The language used is converting different large data streams into one. Single view that is used by a deep learning system. Third, it has demonstrated the effectiveness of the system by applying the tool to many different deep learning applications.

Keywords: *Deep learning, Data visualization, Science and Engineering etc.*

---

1. Introduction

In this rapidly growing digital world, Big Data and Deep Learning are the high focus of Data Science. Big data is a collection of large amounts of digital raw data that is difficult to manage and analyze using traditional tools [1,2]. Since digital data is growing rapidly in various shapes, formats and sizes, it is very important to manage this huge amount of data as per the needs of the organization. Companies based on technology such as Microsoft, Yahoo, Amazon, and Google maintain data in Exabyte or even larger. Due to the popularity of social online media companies such as YouTube, Twitter and Facebook, huge amounts of data are generated by billions of users. However, this large amount of information cannot be managed with traditional methods. Therefore, various organizations have developed products using Big Data Analytics for experiments, simulations, data analysis, monitoring and many more business requirements, which makes it an important subject of data science. The main function of Big Data Analytics is to extract useful patterns from large amounts of data that can be used in decision making and forecasting.

1.1 Deep Learning

In recent years, huge amounts of data have been generated from various fields, including medical informatics, social media, cyber security, and a growing number of electronic devices. Traditional data processing techniques have limitations in processing such a large amount of data. Big data requires new and sophisticated algorithms based on machine learning and deep learning techniques to process data in real time with high accuracy and efficiency [20]. However, recent research has shown deep learning with hybrid learning and training mechanisms to process Big Data with high speed.

Deep learning techniques have provided powerful tools for handling large amounts of labeled/unlabeled data as they extract high-level features from them to obtain hierarchical representations.

Deep learning has been applied in the fields of deep learning for speech learning, acoustic modeling for audio classification, handwritten classification, high-resolution remote sensing, visual classification, natural language processing, computer vision, pattern recognition, etc. [26].

1.2 Deep Learning in Big Data Analytics

The concept of deep learning is to excavate large amounts of data so that patterns and features can be automatically identified from complex unpublished data without human involvement, which is an important tool for Big Data analysis [19]. In today's fast growing data world where there is a need for various machine learning techniques as well as computationally strong machines that can handle huge amounts of data with different formats and sizes, which can be used for this volume and Data can be used to deal with diversity. Deep learning uses supervised / unheard techniques to automatically learn and extract data representation. It can be used to address big data problems (such as data tagging and indexing, information retrieval, etc.) in a more efficient manner, which is not possible with traditional methods. Here we discuss two deep learning architectures of deep learning with respect to its use in Big Data applications.

1.3 Mobile big data

MBD brings a large amount of new challenges from its high dimensionality, comprehensiveness and applications to other complex features, such as traditional data analysis methods for planning, operation and maintenance, optimization and marketing. This section discusses the five versus (short for volume, velocity, diversity, value, and veracity) characteristics [8] derived from big data towards MBD. Five features have been improved in M-Internet, while it allows users to access the Internet anytime and anywhere.

1.4 Big data classification

Supervised learning (classification) faces a major challenge on how to deal with big data. At present, classification problems involving large-scale data are ubiquitous, but traditional classification algorithms are not well suited for big data processing. (1) Support Vector Machine (SVM). The traditional statistical machine learning method has two main problems when dealing with big data. (1) Traditional statistical machine learning methods always involve intensive computing which makes it difficult to apply to large data sets. (2) The prediction of the model that corresponds to the robust and non-parametric confidence interval is unknown. Lau et al. [10] proposed an online support vector machine (SVM) learning algorithm to tackle the classification problem for sequentially provided input data. The classification algorithm is faster with less support vectors, and has better generalization capabilities. Laskov et al. [Rapid] proposed a fast, stable and robust numerical incremental support vector machine learning method. Chang et al. [Open] Developed an open source package called LIBSVM as a library for SVM code implementation.

2. Literature Review

Chahal, H.; Jyoti, J.; Wirtz, (2019) The big data opportunity is not only to achieve high efficiency in business functions. There are also important opportunities for economic development and improvement in the standard of living of the society.

Brynjolfson, E.; Hit, LM; Kim, (2019) There are many ways in which big data analytics businesses can improve organizational output and industries. These include enabling better health care delivery, education standards, national security and good governance. In addition, it has the potential to help policy-makers gain insight into enabling policies that will provide safe playgrounds for investors, thus helping waste managers.

Hu, P.; Zhang, J. (2020), 5G Enabled Fault Detection and Diagnostics: How Do We Achieve Efficiency. Over the past decade, the amount of data exchanged on the Internet has increased by more than 100, and is expected to exceed 500 exabytes by 2020. This phenomenon is primarily due to the development of high-speed broadband Internet and, more specifically, the popularization and widespread proliferation of smartphones and related accessible data plans.

Chien, W.C .; Weng, H.Y .; Lai, CF (2020), Q-learning based collaborative cache allocation in mobile edge computing. Future generation. Computation. Furthermore, with the deployment of Internet of Things (IoT) applications, smart cities, vehicular networks, e-health systems and Industry 4.0, a new plethora of 5G services have emerged with very different and technically challenging design requirements.

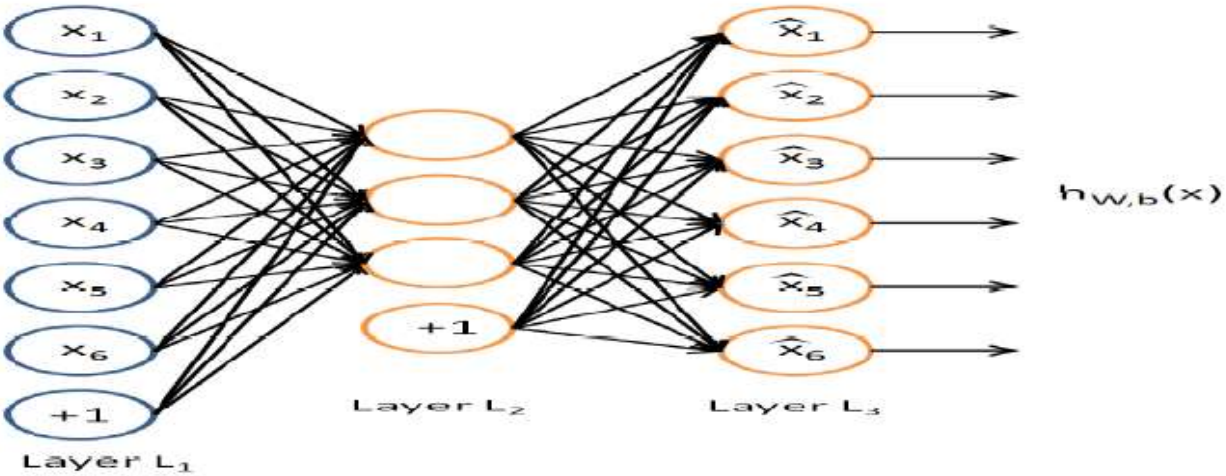
3. Objective

1. To exploring the Application of Machine Learning Methods in Mobile Big Data Analysis
2. Studying classification problems involving large scale data
3. Analyzing human online and offline behavior based on mobile big data

4. Performing performance comparisons of deep autoencoders and estimated speedups based on parallel time estimation using the proposed API

4. Research Methodology

Deep Learning Architectures



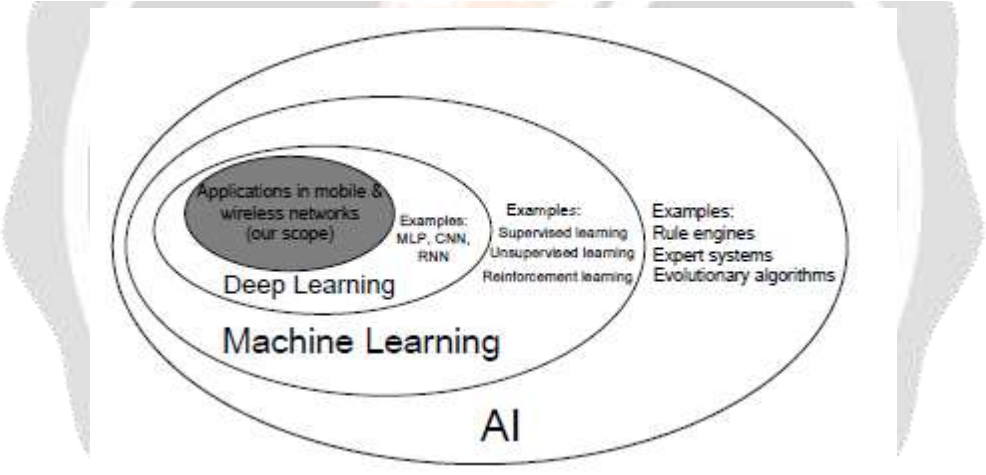


Figure 1.2: Venn diagram of the relationship between deep learning, machine learning, and AI. The survey focuses specifically on intensive learning applications in mobile and wireless networks.

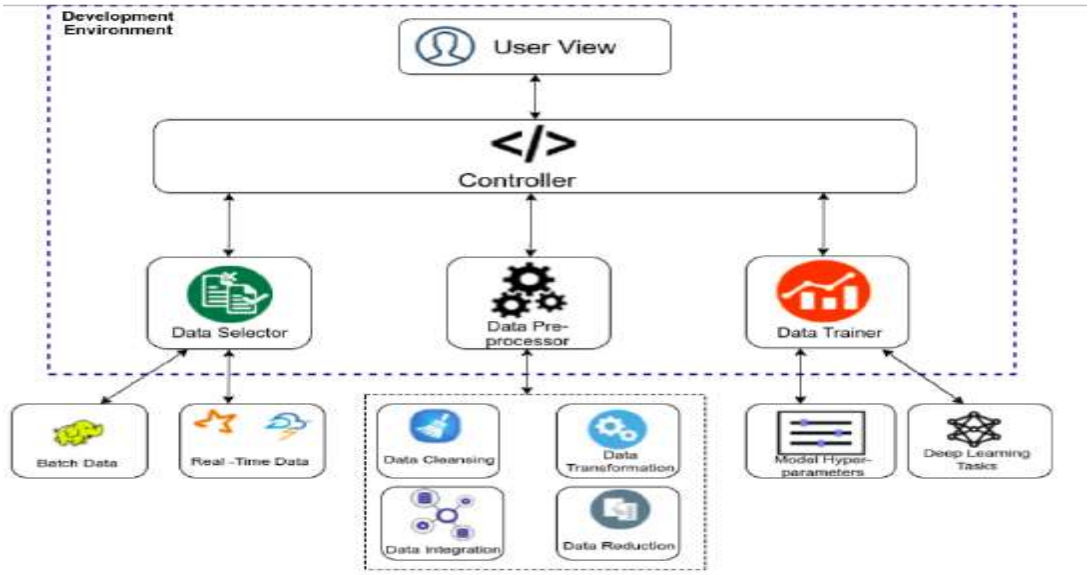


Figure 1.3: Proposed Approach Architecture

5. Result Analysis

5.1. Development of data analysis methods

In this section, we present some recent achievements in data analysis from four different perspectives.

5.1.1. Divide and Conquer Strategy and Big Data Sampling

The divide-and-conquer strategy is a computing paradigm dealing with big data problems. After the development of distributed and parallel computing, the divide and conquer strategy is particularly important.

In general, the importance of different samples in data learning is also different. Some redundant and noisy data can not only lead to large amounts of storage costs, but can also reduce the efficiency of learning algorithms and affect learning accuracy. Therefore, it is preferable to select representative samples to form a subset of the original sample space according to a certain performance standard, such as the distribution of samples, maintaining the topological structure and classification accuracy. Then the learning method will be built on this subset to complete the learning task. In this way, we can maintain or even improve the performance of big data analyzing algorithms with minimal computing and stock resources. Under the background of big data, the demand for sample selection is more urgent. But most sample selection methods are only suitable for small data sets, such as traditional condensed nearest neighbor (CNN), reduced nearest neighbor (RNN) and edited nearest neighbor (ENN), the basic concept of these methods is to find the minimum consistent subset. is. To find the minimum consistent subset, we need to test each sample and the result is very sensitive to the subset initialization and sample setting order. Plotted a method for classifying and selecting edge boundary samples based on local geometry and probability distributions. They keep the location information of the original data, but require the k-means to be computed for each sample. A fast nearest neighbor algorithm (FCNN) based on CNN has been proposed, which has a tendency to select classification range samples.

5.1.2. Big data feature selection

In the fields of data mining, document classification, and indexing in multimedia data, the dataset is always large, with a large number of records and features. This leads to lower efficiency of the algorithm. By feature selection, we can eliminate irrelevant features and increase the effectiveness of task analysis. Thus we can improve the accuracy of the model and reduce the running time.

One of the major challenges of big data processing is how to deal with high dimensional and sparse data. Under the big data environment, network traffic, communication records, large-scale social networks contain a large number of

high-dimensional data, the tensor (such as a multidimensional array) representation provides a natural representation of the data. In this situation, tensor decomposition becomes an important tool for summarization and analysis. The proposed the efficient use of memory of the Tucker decomposition method (memory - efficient Tucker decomposition, MET) to solve the problem of using time and space, which cannot be solved by the traditional tensor decomposition algorithm. . The MET selects the execution strategy optimally based on the available memory in the decomposition process. The algorithm maximizes computation speed based on using available memory. Avoid dealing with a large number of sporadic intermediate results during the MAT calculation process.

5.2. Big Data Classification

Supervised learning (classification) faces a new challenge of how to deal with big data. At present, classification problems involving large-scale data are ubiquitous, but traditional classification algorithms are not well suited for big data processing.

5.2.1. SVM classification

The traditional statistical machine learning method has two main problems when dealing with big data.

1) Traditional statistical machine learning methods have always involved intensive computing which makes it difficult to apply to large data sets.

2) The prediction of the model that fits the robust and nonparameterized confidence intervals is unknown. proposed an online SVM learning algorithm to tackle the classification problem for sequentially provided input data. The classification algorithm is faster, with less number of support vectors, and has better generalization capabilities. Proposed a fast, stable and robust numerical incremental support vector machine learning method.

In addition, the M4 presents a large margin classifier. Unlike other large margin classifiers, which construct a locally or globally separation hyperplane, this model can learn both local and global decision boundaries. SVM and Minimax Probability Machine (MPM) have a close relationship with the model. The model has significant theoretical importance and furthermore, the optimization problem of M4 can be solved in polynomial time.

5.2.2 Decision Tree Classification

Traditional decision trees, as in the classic classification learning algorithm, have a large memory requirement problem when processing large data. Put forward a method of constructing a decision tree from big data, which overcomes some limitations of the current algorithm. In addition, it can use all training set data without saving it in memory. Experimental results have shown that this method is faster than the current decision tree algorithm on large-scale problems. Proposed a rapidly incremental optimization decision tree algorithm for processing large data with noise. Compared to traditional decision tree data mining algorithms, the main advantage of this algorithm is real-time mining capability, which is quite suitable when mobile data streams are unlimited. In addition, it can store the entire data for the training decision model. The advantage of this model is that it can prevent the explosive growth of the size of the decision tree and the reduction of prediction accuracy when there is noise in the data packet.

5.2.3. Neural Networks and Extreme Learning Machines

Traditional feed-forward neural networks typically use a gradient descent algorithm to tune the weight parameters. Generally speaking, slow learning speed and poor generalization performance are bottlenecks that restrict the application of feed-forward neural networks. Rejected the gradient descent algorithm and the iterative adjustment strategy of the proposed Extreme Learning Machine (ELM). This method randomly specifies the input weights and the divergence of a single hidden layer neural network. It can analyze the output load of the network by a step count. Compared to the traditional feedforward neural network training algorithm, network weights can be determined by several iterations, and significantly improve the training speed of ELMs.

However, due to the limitation of computing resource and computational complexity, it is a difficult problem to train a single ELM on big data. There are usually two ways to solve this problem: 1) ELM training based on divide-and-conquer strategy; 2) Introducing parallel mechanisms to train single ELMs.

5.3 Analyzing human online and offline behavior based on mobile big data

Advances in wireless technologies and ever-increasing mobile applications bring an explosion of mobile traffic data. It is an excellent source of knowledge to obtain the regularity of movement of individuals and the mobility of a population of millions. believes that 93% of personal movements are potentially predictable. Thus, various models have been applied to describe human offline mobility behavior. In general, collecting human mobile traffic data passively when it is using mobile Internet has many advantages: high cost efficiency, low energy consumption, a wide range and covering a large number of people do, and with fine time granularity, which give us the opportunity to study human dynamics on a scale that other data sources are very difficult to access. Innovative mobility models derived from mobile data are expected to impact many fields, including urban planning, road traffic engineering, human sociology, the epidemiology of infectious diseases, or telecommunications networking. Another important feature is the online browsing behavior on the user behavior of network resource consumption. There are a variety of applications now available on smart devices, which cover and facilitate all aspects of our daily lives. For example, we can order taxis, shop and book hotels using mobile phones. Analyze the longitudinal effect of proximity density, personality and location on smartphone traffic consumption. In particular, location has a great impact on what type of apps users choose to use. The above observations suggest that there is a close relationship between online browsing behavior and offline mobility behavior. Figure 1.4 (a) is an example of how the browsed application and the current location relate to each other in terms of temporal and spatial regularity. It has been found that mobility behavior has a profound effect on online browsing behavior. Similar trends can be seen for crowding at places where gatherings are held, as shown in 1.5 (b), i.e. some apps are preferred over places that group people together and some that provide specific tasks.

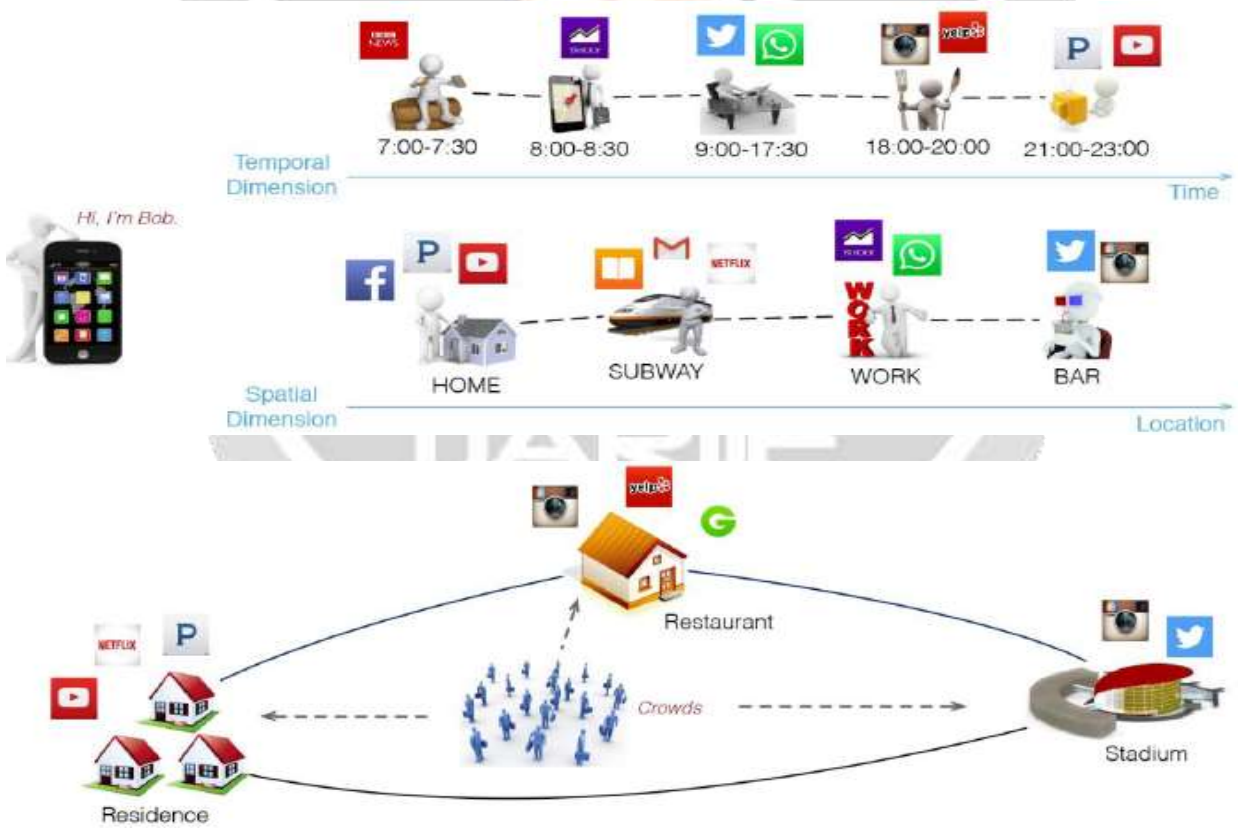


Figure 1.4: App usage behavior in daily life: (a) App usage behavior of a person; (B) App usage behavior of crowds at crowd gathering locations [2].

Online and offline social networks are constructed on the basis of online interest based and location based social networks among mobile users respectively. Combining information from multiple networks in a multilayer configuration provides new insights into user interactions in the online and offline worlds. It sheds light on the link

generation process from several considerations, which will significantly improve link prediction systems with valuable applications for social bootstrapping and friend recommendations.

5.4 Performing performance comparisons of deep autoencoders and estimated speedups based on parallel time estimation using the proposed API

MNIST is a good place to start exploring image recognition and DBNs. The first step is to take an image from the dataset and convert its pixels from a continuous gray scale to binary. Typically, every gray-scale pixel with a value greater than 35 becomes 1, while the rest are set to 0. The MNIST dataset iterator class performs this operation. Nodes in any one layer do not communicate with each other afterwards. This stack of RBMs can end up with a softmax1 layer to form a classifier, or it can help cluster unlabeled data in a supervised learning scenario. When trained on a set of instances in an uncontrolled way, a DBN can learn to reconstruct its input probabilistically. The layers then act as a feature detector at the input. After this learning phase, a DBN can be further trained to perform classification in a supervised manner. With the exception of the first and last layers, each layer in a deep trust network has a dual role: it acts as a hidden layer for nodes preceding it, and the input (or visible) layer for nodes. acts as. come later. The reason for using a DBN is to identify, cluster, and generate images, video sequences, and motion-capture data. A continuous deep belief network is simply an extension of a deep belief network that accepts a continuum of decimal rather than binary data.

Table 1.1: RBM Execution Time (MSEC) with Various Parameters

v →		6			12			24		
		h			h			h		
k	N	4	8	16	4	8	16	4	8	16
1	1000	1.60	3.59	5.52	3.58	4.99	7.84	5.47	7.80	12.64
	2000	4.15	6.02	9.83	5.96	8.93	14.59	9.89	14.63	23.97
	3000	5.71	8.62	14.02	8.59	12.79	21.45	14.20	21.19	35.29
10	1000	5.87	8.22	12.95	9.02	12.62	19.56	11.96	15.84	23.69
	2000	10.56	15.30	24.99	16.98	23.99	30.28	22.88	32.18	45.92
	3000	15.36	22.43	46.72	24.92	35.79	41.20	33.90	47.68	68.35
15	1000	7.83	10.93	16.94	12.44	16.88	25.33	21.62	28.45	41.45
	2000	14.68	20.79	32.96	23.63	32.40	49.30	41.95	55.88	91.99
	3000	21.25	30.59	48.92	34.97	48.26	73.59	62.74	82.74	132.42

Table 1.2: ANOVA Table of RBM Principal Factor Analysis

Main Effects	% Variations
k	51.76
v	28.70
N	36.94
h	26.08

Table 1.3: Hadoop Cluster Configurations

Component/Configuration	Description/Value
Processor	Intel Core 2 due i5-7300HQ @ 2.8 and 3.1 GHz
Memory	10 GB
Hard Disk	2 TB
Yarn CPU Cores	8
Yarn Memory	20480 MB
Scheduler Max Memory	8286 MB
Scheduler Min	1024 MB

Memory	
Yarn Virtual Memory Check	Disabled
MapReduce CPU Cores	8
MapReduce Memory	5096 MB
Mapper Memory	3048 MB
Reducer Memory	3048

Table 1.4: Execution Time (secs)

No. of Images	Sequential	Parallel
5k	5	3
10k	15	10
15k	40	20
20k	80	60
30k	120	100
40k	160	140
50k	200	180
60k	240	220

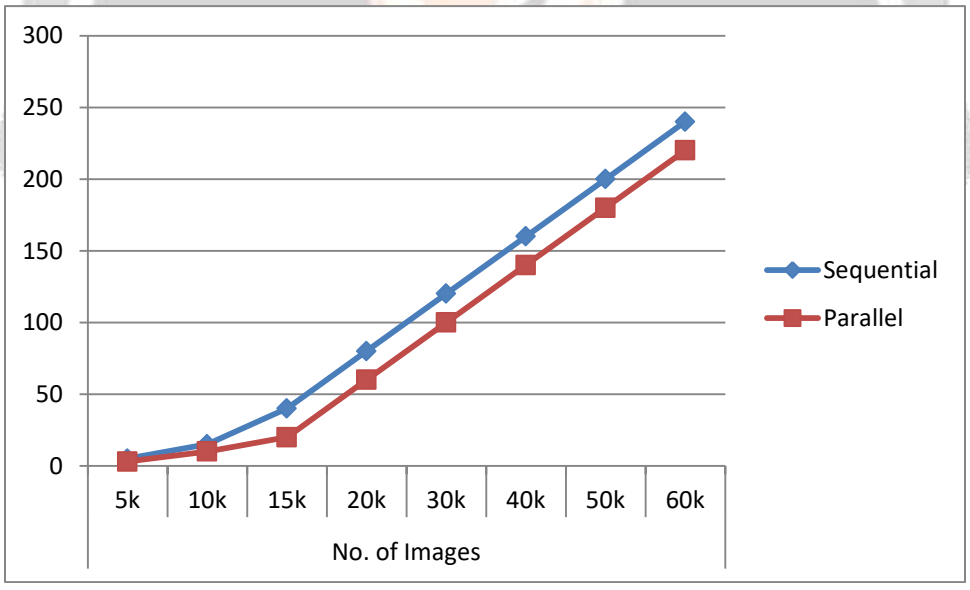


Figure 1.5: Performing performance comparisons of deep autoencoders using the proposed API

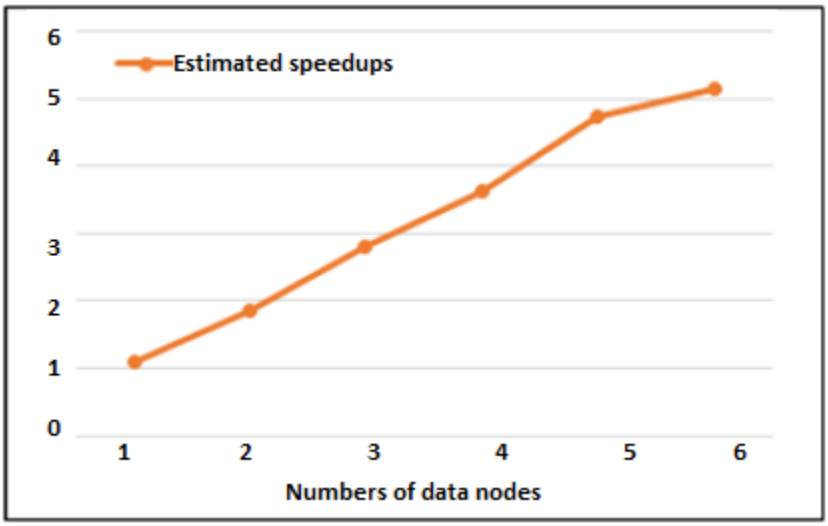


Figure 1.6: Estimated speedups based on parallel time estimation using the proposed API

Fig. The 1.5 model shows the execution time in seconds for both sequential (in PyTor) and parallel implementations of Deep Autoencoder on the MNIST dataset using the proposed software API with various input images for training. The obtained results show a significant improvement in performance (about 60%) for an input size of 6000 images, while the improvement percentages are decreasing as the input size increases such that for an input size of 70000 images the performance improvement is only 10%. The reason for this behavior is the extensive memory usage for storing the input dataset in memory for heavy read operations and processing from permanent storage. So, if we extend the cluster configuration to multiple data nodes, the input dataset will be distributed among multiple data nodes and training will be done in parallel on each partition of the data. This will give more speed as we increase the number of data nodes in the cluster. Fig. 1.5 shows approximate speeds approximating the execution time reduction of a parallel implementation based on the size of the data partition in which each node will be for processing.

Conclusion

In conclusion, it is difficult to directly apply the traditional classification method of machine learning to the analysis of big data. The study of parallel or improved strategies of various classification algorithms has become the new direction. The algorithm runs in a distributed environment and is suitable for large data sets and streaming data. Compared to serial decision tree, the algorithm can improve efficiency based on accuracy error approximation. SVM and Minimax Probability Machine (MPM) have a close relationship with the model. The model has significant theoretical importance and furthermore, the optimization problem of M4 can be solved in polynomial time. In summary, due to the complexity, high dimensionality, and uncertain characteristics of big data, dimension reduction and how to reduce the difficulty of big data processing using feature selection techniques is an immediate problem to solve. The rapid development of mobile applications and the increasing demand for Internet access by end users present both challenges and opportunities for the mobile networks of the future. This section surveys the literature on the analysis of human online and offline behavior based on mobile traffic data. In addition, a framework to meet the high need to deal with the dramatically increased mobile big data has also been investigated. The analysis of mobile big data will provide valuable information for ISPs on network deployment, resource management and the design of future mobile network architectures. There are several frameworks and available libraries that provide efficient implementation of these algorithms. Therefore, there is a need to extend these frameworks to perform model execution on multiple computing nodes, where each node has a portion of the input sample and runs the model in parallel. Therefore, to ease the development of deep learning models for big data analytics, we propose a parallel software API as an extension of Pytorch with HDFS and MapReduce frameworks. We have achieved significant improvements in the deep learning model by using the proposed API to reduce execution time and code complexity. We have evaluated the API by implementing a deep auto-encoder using the MNIST dataset on a single node Hadoop cluster. In future, we plan to set up a multinode hadoop cluster and run the implementation with different data nodes.

References

[1]. A. Ng, " Sparse auto encoder", CS294A Lecture notes: Stanford University, 2011.
[2]. Xu P. et al., Cross-modal Subspace Learning for Sketch-based Image Retrieval: A Comparative Study, in Proceedings of IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC), Sept. 23-25, 2016.
[3]. International Telecommunication Union (ITU). ICT Facts and Figures 2017, Available at: https://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx
[4]. Meeker. Internet Trend 2017, Available at: http://www.kpcb.com/internet-trends
[5]. Fettweis G, Alamouti S M. 5G: Personal mobile internet beyond what cellular did to telephony. IEEE Communications Magazine, 2014, 52(2): 140-145.
[6]. Alsheikh M A, Niyato D, Lin S, et al. Mobile big data analytics using deep learning and apache spark. IEEE Network, 2016, 30(3): 22-29.
[7]. Cisco. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021 White Paper. Feb. 7, 2017, Available at: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html
[8]. Cisco. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014-2019 White Paper. Feb. 3, 2015, Available at: https://ec.europa.eu/futurium/en/content/cisco-visual-networking-index-global-mobile-data-traffic-forecast-update-2014-2019-white
[9]. Wang Z, Qi Y, Liu J, et al. User Intention Understanding from Scratch. IEEE International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE), 2016: 6-8.
[10]. Zhang C, Zhang Y, Xu W, et al. Mining activation force defined dependency patterns for relation extraction. Knowledge-Based Systems (KBS), 2015, 86: 278-287.
[11]. Jordan M. Message from the President: The Era of Big Data. ISBA Bull, 2011, 18: 1-3.
[12]. Chen W, Wipf D P, Wang Y, et al. Simultaneous Bayesian Sparse Approximation with Structured Sparse Models. IEEE Transactions on Signal Processing, 2016, 64(23): 6145-6159.
[13]. Zhang G, Heusdens R. Distributed Optimization using the Primal-dual Method of Multipliers. IEEE Transactions on Signal and Information Processing over Networks, 2017, In press.