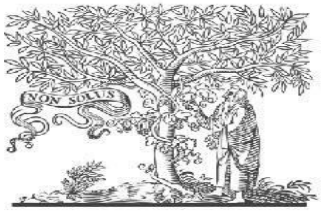
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 26th Jul 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 07

10.48047/IJIEMR/V12/ISSUE 07/11

Title **DEEP LEARNING CHALLENGES IN BIG DATA ANALYTICS AND STUDY ON ADVANCED MACHINE LEARNING METHODS FOR BIG DATA**

Volume 12, ISSUE 07, Pages: 100-106

Paper Authors **Kamini dubey, Sanjay Kumar tiwari**

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

DEEP LEARNING CHALLENGES IN BIG DATA ANALYTICS AND STUDY ON ADVANCED MACHINE LEARNING METHODS FOR BIG DATA

Candidate name: Kamini dubey
DESIGNATION - RESEARCH SCHOLAR SUNRISE UNIVERSITY ALWAR
Guide name.- Sanjay Kumar tiwari
DESIGNATION- assistant professor SUNRISE UNIVERSITY ALWAR

ABSTRACT
If everything else is equal, data is the basis. The amount of data is growing at an unprecedented pace as a result of developments in social media, mobile technology, web technologies, and sensing devices. The amount of information we often transmit, for instance, is quite stimulating. Every day, 2.5 quintillion bytes worth of data are created. The quantity of data keeps growing rapidly. In 2020, every second of fresh data creation will roughly equal 1.7 super bytes. Big data's total storage capacity is expected to grow from its current 4.4 Zetta bytes to about 44 Zetta bytes (or 44 trillion gigabytes) by 2020. This Bigdata promises a substantial rise in corporate value in a wide range of sectors, including the healthcare industry, monetary services, medical care, transportation, and online advertising. However, conventional methods are having difficulty keeping up with such a massive data set. In this research, we focus on deep learning difficulties in big data analytics and cutting-edge machine learning techniques for large data.
Keywords: - Data, Science, Big Data, Social, Organization.

I. INTRODUCTION

Big data and deep learning are at the forefront of data science in today's rapidly evolving digital world. When it comes to managing and analyzing information, traditional tools just can't keep up with the volume and complexity of Big Data. As the amount and variety of digital data continues to grow at an exponential rate, organizations must find effective ways to manage this deluge of information. Technology-reliant corporations like Microsoft, Yahoo, Amazon, and Google have maintained data storage capacities in the Exabyte range or more.

Popular social media platforms such as YouTube, Twitter, and Facebook generate massive amounts of data from their many users. However, conventional tools will be inadequate for managing such a mountain of data. Therefore, many organizations have developed their own products by using Big Data Analytics for testing, simulations, data analysis, checking, and many other business purposes. Big data analytics is tasked with extracting useful patterns from large amounts of data for the sake of fundamental leadership and prediction.

However, Big Data Analytics presents its own set of difficulties when it comes to data analysis and machine learning, such as the

wide range of input data formats and sizes, the speed with which data streams, the consistency of data analysis, the diversity of data, the lack of structure in unsupervised data, the need for rapid data recovery, labeling, and storage, and so on.

II. DIFFICULTIES IN DEEPER LEARNING REQUIRED FOR BIG DATA ANALYTICS

Before, we highlighted the significance and benefits of Deep Learning algorithms for Big Data Analytics. However, there are several aspects of Big Data that make it difficult to modify and adapt Deep Learning to solve these concerns. Several areas of Big Data, including learning with streaming data, handling high-dimensional data, model flexibility, and distributed computing, are highlighted in this section as places where Deep Learning may benefit from further study.

Gradual learning for non-stationary data

Managing streaming and rapidly changing input data is a challenging aspect of Big Data Analytics. Such an analysis of data is useful for monitoring tasks like extortion recognition. Streaming data is becoming more prevalent, and as a result, there is a pressing need to adapt Deep Learning to handle it. Works such as steady feature learning and extraction, de-noising auto encoders, and deep belief networks are discussed here as they pertain to Deep Learning and streaming data.

Using denoising autoencoder, Zhou illustrates how a Deep Learning algorithm may be put to use for slow feature learning on very large datasets. Denoising autoencoders are a subclass of autoencoders that extract meaningful features from corrupted input, where the isolated features are robust to noisy data and helpful for classification. At the end of the day, deep learning algorithms rely on hidden layers to help with feature extraction and data modeling. One hidden layer focuses features in a denoising autoencoder, and its initial number of hubs is equal to the number of features that will be discarded. Examples that don't conform to the given target work (e.g., those with characterization errors larger than an edge or recreation errors above a certain threshold) are systematically collected and used to inform the introduction of new hubs in the hidden layer. Therefore, all of the features are retrained jointly using forthcoming fresh data tests. Learning and mapping features consistently may enhance discriminative or generative task performance, but doing so too often might lead to an abundance of features and overfitting the data. Subsequently, similar elements are merged into a streamlined configuration. Zhou shows that in a large-scale online environment, the optimal amount of features may be quickly combined using the gradual feature learning technique. This method of progressive feature extraction is useful in situations where data delivery in massive internet data streams varies over time. Other Deep Learning algorithms, such as RB, may benefit from a consistent summary of feature learning and extraction, making it possible to adapt to a new incoming stream of live, massive-scale data. In addition, it avoids the need for time-consuming and expensive

cross-approval inquiry when deciding on the feature set size for large-scale datasets.

The adaptable deep belief networks shown by Calandra demonstrate how Deep Learning may be consolidated to benefit from online non-stationary and flowing data. In order to better understand the new deep conviction network that has adapted to the recently seen data, their analysis takes use of the generative characteristic of deep conviction networks to mimic the instances from the initial data. But the need for constant memory use is a downside of a flexible deep belief network.

Some Big Data application domains, such as social media feeds, advertising and financial data nourishes, web click stream data, operational logs, and metering data, present challenges in the analysis of large amounts of rapidly changing streaming data. The works presented here provide observational support for further research and development of novel Deep Learning algorithms and architectures for this task. For instance, Amazon Kinesis is a managed service for handling real-time streaming of Big Data, but it does not use Deep Learning.

High-dimensional data

Due to the deep layered pecking order of learning data deliberations and portrayals from a lower-level layer to a more significant level layer, some Deep Learning algorithms can become prohibitively computationally-costly when handling high-dimensional data, such as images. When dealing with Big Data that has a high Volume, one of the four Vs associated with Big Data Analytics, these Deep Learning algorithms may run into trouble. Although it makes it more difficult to learn from the data, a high-dimensional data source adds significantly to the total amount of raw data.

Chen introduces mSDAs, a kind of stacked denoising autoencoders that is computationally faster than traditional SDAs and performs well on high-dimensional data. Their approach doesn't use stochastic inclination plunge or other advancement algorithms to learn parameters since it drastically reduces noise in SDA preparation. A closed structure arrangement with considerable speedups is made possible by designing the reduced denoising autoencoder layers with hidden hubs. In addition, the model selection process is greatly simplified by the fact that reduced SDA requires only two free meta-parameters, which govern the amount of noise and the number of stacked layers, respectively. mSDA is a potential approach with broad appeal in data mining and machine learning because of its short preparation time, scalability to large-scale and high-dimensional data, and ease of execution.

Another method that works well on high-dimensional data is convolutional neural networks. Experts have honed down on convolutional neural networks using the ImageNet dataset of 256 x 256 RGB pictures to achieve state-of-the-art performance. Convolutional neural networks only need connections between neurons in neighboring units on the hidden layer, rather than between all of the hubs on the previous layer. When picture data is pushed to deeper network levels, it loses some of its original purpose.

Still largely unexplored is the application of Deep Learning algorithms to Big Data Analytics, which includes high-dimensional data. This calls for the development of Deep Learning-based arrangements that either fine-tune existing methods, such as those shown above, or develop wholly new approaches to addressing the high dimensionality present in some Big Data spaces.

Large-scale models

How can we take the continued successes of Deep Learning and apply them to much larger-scale models and vast datasets, from a computational and analytical point of view? Accurate results demonstrate the effectiveness of large-scale models, with a focus on those that use countless model parameters to disentangle increasingly muddled aspects and depictions.

The problem of training a Deep Learning neural network with billions of parameters using several CPU cores simultaneously; applicable to both speech recognition and computer vision. DistBelief is a software framework developed to facilitate the usage of computer clusters including a large number of machines for the purpose of preparing massive models. DistBelief manages the nuances of parallelism, synchronization, and correspondence, and the architecture supports model parallelism both locally on a system (through multithreading) and remotely between machines (via message passing). Data parallelism, in which many copies of a model are used to speed up the completion of a single task, is also supported by the system. Nonconcurrent SGDs are necessary for large-scale distributed preparation to be feasible. Similar to how L-BFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno, a semi Newton technique for unconstrained improvement) is implemented in a distributed group enhancement system. The basic idea is to simultaneously create numerous variants of the model, each of which runs on a different node in the network and analyzes a different portion of the data. The authors claim that their framework can speed up the development of standard-sized models, as well as prepare models that are bigger than could be envisioned otherwise. Even though the system's primary focus is on setting up massive neural networks, the underlying algorithms are still relevant to other angle-based learning techniques. In any event, it's worth noting that the extensive computing resources DistBelief uses are often unavailable to a wider audience.

III. BIGDATA

Bigdata is a relatively new phrase that may be used to describe any large amount of data, regardless of how well it is organized. Bigdata is a relatively new concept, the path toward social gatherings and storing massive amounts of data for inevitable analysis. Big data refers to large data collections and the many computerized methods and developments that are utilized to process them. Big data refers to data that has a greater variety and is collected in larger quantities and at a faster rate.

IV. FOR BIG DATA, ADVANCED MACHINE LEARNING METHODS

Most Machine Learning systems are not thorough enough for handling huge data. Therefore, it is always necessary to use explicit learning procedures in accordance with different facts. When it comes to big data, the challenge lies in the requirement to implement Machine Learning algorithms at larger and larger scales. The many methods of education are discussed here.

A. Deep Learning

When it comes to learning hierarchical representations, Deep Learning (also known as Hierarchical Learning) primarily makes use of Supervised and unsupervised Learning in deep architectures. The benefit of Deep Learning is that the software may unsupervisedly collect the data set on its own. Deep Belief Networks (DBNs) and Convolutional Neural Networks (CNNs) are the two most used methods of deep learning. With the increased usage of computer and processing power, deep learning is providing predictive analytics solutions for massive size data sets. The current developments regarding: a review of well-known deep learning techniques and how they have been used in various Natural Language Processing projects, together with an assessment of their efficacy.

B. Feature learning

Expansions in high-dimensional datasets provide a problem for existing methods of learning to extract and rank the most important information. Feature learning is a solution that can learn to recognize useful representations of data, which in turn makes it easy and uncomplicated to change substantive data. Feature learning, also known as representation learning, is a set of tools that enables a computer program to automatically learn the features used for feature discovery or grouping from collected data and acquire expertise in the features to solve a specific issue.There are three primary categories of Representation Learning, and they are feature determination, feature extraction, and distance metric learning.

C. Active Learning

For situations when it would be too expensive to manually label and collect identified data, unlabeled data becomes more valuable. Learning algorithms may then do a name-data search on the client in this scenario. Active learning refers to this kind of iterative supervised learning. Active learning is a method of instruction in which the learner actively seeks out and uses fresh facts to inform previous findings. Because the client is in charge of selecting the examples, the number of trials needed to get comfortable with a concept may be much lower than in conventional supervised learning. The continued progress including: The proposed solutions for effective crossover rely on experiential learning. Restoration of a biophysical variable

D. Ensemble Learning

There has been a significant increase in the quantity of data collected by businesses, social groups, and other institutions. All of this information is only useful if it is collected correctly so that customers may use it to set realistic objectives. Ensemble techniques take into account a collection of models, and will sum those models to offer a final model, as opposed to producing a single model and demanding that model as

the best and most practical indication we can create. The goal of using Ensemble Learning is to organize a collective whole where a combination of many tactics is more successful than a single learning approach. There are four different kinds of Ensemble Learning, and they are called saddling, boosting, stacking, and error correction. Constant developments in the areas of: In order to improve the accuracy of breast cancer diagnoses and reduce the dependence on Ensemble Learning, a Support Vector Machine (SVM) technique has been proposed.

E. Transfer Learning

Preparing data for use in Machine Learning nowadays might be expensive or time-consuming. This has provided the foundation for the development of more successful pupils equipped with information gleaned from clearer sources. Transfer learning describes this kind of education. In computing, transfer learning refers to a system's ability to recall and make use of previously gained knowledge and expertise while tackling new graphical tasks. When dealing with data that spans many feature spaces and will be disseminated in a variety of ways, transfer learning often provides the best solution. Classifications of Transfer Learning include Inductive Transfer Learning, Transductive Transfer Learning, and Unsupervised Transfer Learning. The continued progress in the areas of: developed a Big Data transfer learning framework using automated transfer learning (ATL).

V. CONCLUSION

MapReduce and distributed structures like Hadoop are only two examples of the many methods that have been put into practice to process Machine Learning algorithms in order to get access to massive amounts of data. while it comes to the challenges of Machine Learning while using Big Data, advanced tactics recall a few instruments for which Deep learning may be successful. Deep learning can handle and learn from problems present in enormous amounts of input data with very few obstacles. Deep Learning's ability to dynamically learn and extract varying degrees of data reflections provides a unique level of clarity for Big Data Analytics.

REFERENCES

1. Maryam M Najafabadi,,Flavio Villanustre,Taghi M Khoshgoftaar,Randall Wald, Edin Muharemagi, “Deep learning applications and challenges in big data analytics”, Najafabadi et al. Journal of Big Data (2015)

2. Pwint Phyu Khine, Wang Zhao Shun1, “Big Data for Organizations: A Review”, Journal of Computer and Communications, 2017, 5, 40-48

3. Anushree Priyadarshini and SonaliAgarwal, “A Map Reduce based Support Vector Machine for Big Data Classification”, International Journal of Database Theory and Application Vol.8, No.5 (2015), pp.77-98

4. Tamer Tulgar, Ali Haydar and Ibrahim Ersan, “A Distributed K-Nearest Neighbor Classifier for Big Data”, BALKAN JOURNAL OF ELECTRICAL & COMPUTER ENGINEERING, Vol. 6, No. 2, April 2018

5. Wei Dai, Wei Ji, "A Map Reduce approach of c4.5 Decision tree Algorithm", International Journal of Theory and Application; vol 7 no.1(2014), pp 49-60

6. Kairan Sun, Xu Wei, Gengtao Jia, Risheng Wang, and Ruizhi Li, "Large-scale Artificial Neural Network:MapReduce-based Deep Learning", arXiv:1510.02709v1 [cs.DC] 9 Oct 2015

7. .Available [online] https://www.forbes.com/sites/bernardmarr/2018

8. PETER HARRINGTON, "Machine Learning in Action"

9. Annina Simon, Mahima Singh Deo, Mahima Singh Deo, S. Venkatesan, D.R. Ramesh Babu,D.R. Ramesh Babu, "An Overview of Machine Learning and its Applications", International Journal of Electrical Sciences & Engineering (IJESE);Vol1, Issue 1; 2015 pp. 22-24

10 Scott Bruce, Zeda Li, Hsiang-Chieh Yang and SubhadeepMukhopadhyay, "Nonparametric Distributed Learning Architecture for Big Data: Algorithm and Applications", rXiv:1508.03747v5 [stat.AP] 26 Feb 2018