# A Depth of Deep Learning for Big Data and its Applications

Abhay Narayan Tripathi,
Research Scholar, DIT Univeristy,
Dehradun, Uttarakhand

Bharti Sharma
Assistant Professor, DIT University
Dehradun,Uttrakhand

***ABSTRACT:-*** **Although Machine Learning (ML) has become synonymous for Artificial Intelligence (AI); recently, Deep Learning (DL) is being used in place of machine learning persistently. While machine learning is busy in supervised and unsupervised methods, deep learning continues its motivation for replicating the human nervous system by incorporating advanced types of Neural Networks (NN).. If we apply Deep Learning to Big Data, we can find unknown and useful patterns that were impossible so far. Deep Learning is applied in self driving cars, visual recognition, healthcare, transportation etc. Nowadays, companies have started to realize the importance of data availability in large amounts in order to make the correct decision and support their strategies. Big Data means extremely huge large data sets, which is heterogeneous whose characteristics (large volume, different forms, speed of processing), analyzed to find the patterns, trends. This paper provides an introductory tutorial to the domain of deep learning for Big Data with its history, evolution, and introduction to some of the sophisticated neural networks such as Deep belief network , Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN).**

***Key words— Artificial Intelligence, Deep Learning, Machine Learning, Neural Networks, Convolutional Neural Network, Deep Belief Network, Recurrent Neural Network. Big Data.***

## I. INTRODUCTION:

Deep Learning is considered as the subset of machine learning which is intern subset of Artificial Intelligence, a prominent field of computer science over the past decade. Artificial Intelligence makes machines to think intelligently without minimal human intervention. Machine learning comprises with various algorithms that are capable to model high level abstractions from input data. Deep Learning provides a more adaptive way using deep neural network that learns feature itself from the given input data and make machine capable for taking decision. Unlike task specific algorithms of machine learning, deep learning is a method based on learning data representations. Learning can be supervised, semi-supervised or unsupervised. Deep learning provides set of algorithms and approaches that learns features and tasks directly from data. Data can be of any type, structured or unstructured, including images, text or sound. Deep learning is often referred as end-to-end learning because it learns directly from data. Moreover, Deep learning techniques works without human mediation and sometime capable of producing more accurate result than human being itself. Nowadays, deep learning is widely used in the areas like computer vision, natural language processing, pattern recognition, object detection.

Representative learning methods of deep learning provides multiple level of representation, generated by linking simple but non-linear modules that transmute the representation at one level into a representation at next higher layer, slightly in more abstract way [1]. Artificial Intelligence (AI), and specifically Deep Learning (DL), are trending to become integral components of every service in our future digital society and economy. This is mainly due to the rise in computational power and advances in data science. DL's inherent ability to discover correlations from vast quantities of data in an unsupervised fashion has been the main drive for its wide adoption. Deep Learning also enables dynamic discovery of features from data, unlike traditional machine learning approaches, where feature selection remains a challenge. Deep Learning has been applied to various domains such as speech recognition and image classification, nature language processing, and computer vision. Typical deep neural networks (DNN) require large amounts of data to learn parameters (often reaching to millions), which is a computationally intensive process requiring significant time to train a model. As the data size increases exponentially and the deep learning models become more complex, it requires more computing power and memory, such as high performance computing (HPC) resources to train an accuracy model in a timely manner. Despite existing efforts in training and inference of deep learning models to increase concurrency, many existing training algorithms for deep learning are notoriously difficult to scale and parallelize due to inherent interdependencies within the computation steps as well as the training data. The existing methods are not sufficient to systematically harness such systems/clusters.""

We can apply Deep Learning that is a tool for understanding higher abstract knowledge in most steps of Big Data area problems. But preferably it needs high volumes of data. If we want to become more successful in this competitive area, we need to find abstract patterns. Big Data Analytics and Deep Learning are two high-focus of data science. Big Data has become important as many organizations both public and private have been collecting massive amounts of domain-specific information, which can contain useful information about problems such as national intelligence, cyber security, fraud detection, marketing, and medical informatics. Companies such as Google and Microsoft are analyzing large volumes of data for business analysis and decisions, impacting existing and future technology. Deep Learning algorithms extract high-level, complex abstractions as data representations through a hierarchical learning process. Complex abstractions are learnt at a given level based on relatively simpler abstractions formulated in the preceding level in the hierarchy. A key benefit of Deep Learning is the analysis and learning of massive amounts of unsupervised data, making it a valuable tool for Big Data Analytics where raw data is largely unlabeled and un-categorized. In the present study, we explore how Deep Learning can be utilized for addressing some important problems in Big Data Analytics, including extracting complex patterns from massive volumes of data, semantic indexing, data tagging, fast information retrieval, and simplifying discriminative tasks. We also investigate some aspects of Deep Learning research that need further exploration to incorporate specific challenges introduced by Big Data Analytics, including streaming data, high-dimensional data, scalability of models, and distributed computing. We conclude by presenting insights into relevant future works by posing some questions, including defining data sampling criteria, domain adaptation modeling, defining criteria for obtaining useful data abstractions, improving semantic indexing, semi-supervised learning, and active learning.

## II. DEEP LEARNING: OVERVIEW AND BACKGROUND

Machine Learning provides the vast collection of algorithms including singular algorithms, together with statistical approach like Bayesian networks, function approximation such as linear and logistic regression, or decision trees. These algorithms are indeed powerful, but there are few limitations of these algorithms when to learn for enormously complex data representations. Deep learning is emerged from cognitive and information theories and human neurons learning process along with strong interconnection structure between neurons is looking for to imitate. One of the key feature of computing neurons and the neural network model is, it can be able to apply generic neurons to any type of data and learn comprehensively [2]. Deep learning is considered as a promising avenue of research as it is capable of automatically identifying the complex features at a high level of abstraction. It is about learning multilevel representation and abstraction, which is useful for the data, such image, text and sound. One of the exceptional Deep learning characteristics is its capability of using unlabeled data during the training process [3].According to the definition of Deep learning, it is considered as the application of multi-layer neural network with multiple neurons at each layer to perform the desired tasks like classification, regression, clustering and others. Fundamentally, each neuron with activation function is considered as the single logistic node, which connected to the input in the next layer of it, loss function is calculated to modify the weights at each neuron and optimize to make it suitable for input data. Each layer of neural network layer multiple neurons initiated with dissimilar weights and try to learn on the input data concurrently. Thus, in multiple layers with multiple nodes, each node learns from the output of the previous layers, and gradually decreases the approximation of the real input data to provide accurate output representation set [4]. This leads to lot of complexity between multiple interconnected neurons.

The term “deep learning” was initially used by Igor Aizenberg and colleagues in or around 2000 while talking about Artificial Neural Network (ANN). However, the first mathematical model of neural network was introduced in 1943 by Walter Pitts and Warren McCulloch and published in the seminar “A Logical Calculus of Ideas Immanent in Nervous Activity”. In 1965 Alexey Ivakhnenko and V.G. Lapa created first working deep learning network by applying theories and idea to the point. After that, Kunihiko Fukushima had introduced “Neocognitron”- and artificial neural network that learned how to recognize visual patterns in 1979-80. Likewise, many authors contributed in subsequent years. However, in 2006 Hinton claimed that he knew how the brain works, and highlighted the idea of unsupervised pretraining and deep belief nets. Then, in 2009 Fei-Fei Li launched ImageNet, in 2011 Alex Krizhevsky created AlexNet built upong LeNet (by Yann LeCun years earlier). In 2012, convolutional neural network (CNN) is used by Geoff Hinton and his team. They found that it is capable to learne its own features and the error rate decreases to 18.9% [5]. In 2014, GoogLeNet - a Google’s own deep learning algorithm considered as the one of the base in the research field of deep learning as it is able to down the error rate about 6.7%. After this, deep learning has successfully applied to one of the promising application, speech recognition. Deep learning architectures have produced impressive results in all most all natural language processing tasks starting from including sentiment analysis [6] to vocal language understanding [7] including information retrieval, machine translation, contextual entity linking, and many more. One of the notable application of deep learning algorithms are product recommendations that are implemented by all e-commerce websites. Other Deep learning models are also used in the fields of object tagging, face recognition, drug innovation and toxicology, weather forecasting, financial assistance and many more [8].

## III. BIGDATA

The rise of Big Data has been caused by increase of data storage capability, increase of computational power, and more data volume accessibility. Most of the current technologies that are used to handle Big Data challenges are focusing on six main issues of that called Volume, Velocity, Variety, Veracity, Validity, and Volatility. The first one is Volume that means we are facing with huge amounts of data that most of traditional algorithms are not able to deal with this challenge. For example, Each minute 15h of videos are uploaded to Facebook so that collects more than 50 TB per day. With respect to the amounts of data generating each day, we can predict the growth rate of data in next years [9]. The data growth is 40 percent per year. Each year around 1.2 ZB data are produced. Huge companies such as Twitter, Facebook, and Yahoo have recently begun tapping into large volume data benefits. The definition of high volume is not specified in predefined term and it is a relative measure depends on the current situation of the enterprise [10]. The second challenge is Variety that in brief means we are facing with variety types of file formats and even unstructured ones such as PDFs, emails, audios and so on. These data should be unified for further processes [11]. The third V is Velocity that means data are coming in a very fast manner, the rate at which data are coming is striking, that may hang the system easily. It shows the need for real-time algorithms. The next two Vs (Veracity and Validity) have major similarities with each other, mean data must be as clean, trustworthy, usefulness, result data should be valid, as possible for later processing phases. The more data sources and types, the more difficult sustaining trust [12]. And the last V is the Volatility that means how much time data should remain in the system so that they are useful for the system. McKinsey added Value as the seventh V that means the amount of hidden knowledge inside Big Data [13]. We also can consider open research problems from another viewpoint as follows, six parameters: Availability, Scalability, Integrity, Heterogeneity, Resource Optimization, and Velocity (related to stream processing). Labrinidis and Jagadish in [14] described some challenges and research problems with respect to Scalability, Heterogeneity aspects of Big Data management. Other parameters such as availability and integrity are covered in [15]. These parameters are defined as follows: -Availability: Means data should be accessible and available whenever and wherever user requests data even in the case of failure occurrence. Data analysis methods should provide availability to support large amounts of data along with a high-speed stream of data [16].

- Scalability: refers if a system supports large amounts of increasing data efficiently or not. Scalability is an important issue mostly from 2011 for industrial applications to scale well in limited memory.
- Data Integrity: points to data accuracy. The situation be-comes worse when different users with different privileges change data in the cloud. Cloud is in charge of managing databases. Therefore, users have to obey cloud policy for data integrity [17].
- Heterogeneity: refers to different types of data such as structured, unstructured and semi-structured [18].
- Resource Optimization: means using existing resources efficiently. A precise policy for resource optimization is needed for guaranteeing distributed access to Big Data.
- Velocity: means the speed of data creation and data analysis. The increased amount of digital devices like smart phones, tablets caused the increase of speed of data generation. Thus, the need for real-time analyses is obligatory.

These are very application dependent that means can differ for each application to another application. And from steps point of view, Big Data area can be divided into three main Phases: Big Data preprocessing, means doing some preliminary actions toward data with the aim of data preparation such as data cleansing and so on. Big Data storage means how data should be stored. Big Data management means how we should manage data in order to get best achievement such as clustering, classification and so on [19].

## IV. APPLICATION OF DEEP LEARNING IN BIG DATA

If we want to have a look of application of Deep Learning in Big Data, DL deals mainly with two V's of Big Data characteristics: Volume and Variety. It means that DL are suited for analyzing and extracting useful knowledge from both large huge amounts of data and data collected from different sources [20]. One example of application of Deep Learning in Big Data is Microsoft speech recognition (MAVIS) that is using DL enables searching of audios and video files through human voices and speeches [21] [22]. Another usage of DL on Big Data environment is used by Google company for Image search service. They used DL for understanding images so that can be used for image annotation and tagging that is useful for image search engines and image retrieval or even image indexing. When we want to apply Dl, we face some challenges that we need to address them same as:

### *1) Deep Learning for High Volumes of Data*

1.1. The first one is whether we should use all entire Big Data input or not. In general, we apply DL algorithms in a portion of available Big Data for training goal and we use the rest of data for extracting abstract representations and from another point of view, question is that how much volume of data is needed for training data.

1.2. Another open problem is domain adaptation, in applications which training data is different from the distribution of test data. If we want to look at this problem from another viewpoint, we can point to how we can increase the generalization capacity of DL; generalizing learnt patterns where there is a change between input domain and target domain.

1.3. Another problem is defining criteria for allowing data representations to provide useful future semantic mean-ings. In simple word, each extracted data representation should not be allowed to provide useful meaning. We must have some criteria to obtain better data representations.

1.4. Another one is that most of the DL algorithms need a specified loss and we should know what is our aim to extract, sometimes it is very difficult to understand them in the Big Data environment.

1.5. The other problem is that most of them do not provide analytical results that can be understandable easily. In other words, because of its complexity, you cannot analyze the procedure easily. This situation becomes worse in a Big Data environment.

1.6. Deep Learning seems suitable for the integration of heterogeneous data with multiple modalities due to its capability of learning abstract representations.

1.7. The last but not the least major problem is that they need labeled data. If we cannot provide labeled data, they will have bad performance. One possible solution for this is that we can use reinforcement learning, the system gathers data by itself, and the only need for us is giving rewards to the system.

### *2) Deep Learning for High Variety of Data*

These days, data come in all types of formats from a variety sources, probably with different distributions. For example, the rapidly growing multimedia data coming from the web and mobile devices include a huge collection of images, videos and audio streams, graphics and animations, and unstructured text, each with different characteristics. There are open questions in this regard that need to be addressed as some of them presented as follows:

2.1. Given that different sources may offer conflicting information, how can we resolve the conflicts and fuse the data from different sources effectively and efficiently?

2.2. If the system performance benefits from significantly enlarged modalities?

2.3. In which level deep learning architectures are appropriate for feature fusion of heterogeneous data?

### *3) Deep Learning for High Velocity of Data*

Data are generating at extremely high speed and need to be processed at fast speed. One solution for learning from such high-velocity data is online learning approaches that can be done by deep learning. Only limited progress in online deep learning has been made in recent years. There are some challenges in this matter such as:

3.1. Data are often non-stationary, data distributions are changing during the time.

3.2. The big question is whether we can benefit from Big Data along with deep architectures for transfer learning or not.

## V. CONCLUSION

Nowadays, it is necessary to grapple with Big Data with the aim of extracting better abstract knowledge. One technique that is applicable for this aim is Deep Learning (Hierarchical Learning) that provides higher-level data abstraction. Deep Learning is a useful technique that can also be used in the Big Data environment and has its own advantages and disadvantages. In general, the more data, the higher level abstract data, but we face many challenges. The aim of this paper was to give a brief introduction to the field of deep learning starting from its historical perspective and evolution. As for future work, we aim at applying the knowledge of deep learning to a greater number of specific applications especially related to convolutional nets and recurrent nets.

## VI. REFERENCES:

[1] Nilay Ganatra, Atul Patel," Comprehensive Study of Deep Learning Architectures, Applications and Tools", International Journal of Computer Sciences and Engineering, Dec 2018,Vol6(issue 12),

[2] M. Nielsen, "Neural networks and deep learning," 2017. [Online]. Available: http: //neuralnetworksanddeeplearning.com/.

[3] U. V. SurajitChaudhuri, " Anoverview of business intelligence technology," Communications of the ACM, vol. 54, no. 8, p. 88–98, 2011.

[4] A. P. Sanskruti Patel, "Deep Leaning Architectures and its Applications A Survey," INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING , vol. 6, no. 6, pp. 1177-1183, 2018.

[5] I. S. a. G. E. H. A. Krizhevsky, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, p. 1097–1105, 2012.

[6] A. P. J. Y. W. J. C. C. D. M. A. Y. N. a. C. P. R. Socher, "Recursive deep models for semantic composition- ality over a sentiment treebank," in in Proceedings of the conference on empirical methods in natural language processing, Citeseer, 2013.

[7] Y. D. K. Y. Y. B. L. D. D. H.-. T. X. H. L. H. G. T. D. Y. a. G. Z. G. Mesnil, "Using recurrent neural networks for slot filling in spoken language understanding," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 3, pp. 530-539, March, 2015.

[8] Q. V. Le, "Building high-level features using large scale unsupervised learning," in IEEE International Conference on Acoustics, Speech and Signal Processing, May-2013.

[9] Jinchuan Chen, Yueguo Chen, Xiaoyong Du, Cuiping Li, Jiaheng Lu, Suyun Zhao, and Xuan Zhou. Big data challenge: a data management perspective. Frontiers of Computer Science, 7(2):157–164, 2013.

[10] Dylan Maltby. Big data analytics. In 74th Annual Meeting of the Association for Information Science and Technology (ASIST), pages 1–6, 2011.
[11] Han Hu, Yonggang Wen, Tat-Seng Chua, and Xuelong Li. Toward scalable systems for big data analytics: A technology tutorial. IEEE Access, 2:652–687, 2014.
[12] Aisha Siddiqa, Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Mohsen Marjani, Shahabuddin Shamshirband, Abdullah Gani, and Fariza Nasaruddin. A survey of big data management: taxonomy and state-of-the-art. Journal of Network and Computer Applications, 71:151–166, 2016.
[13] Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. Mobile Networks and Applications, 19(2):171–209, 2014.
[14] Alexandros Labrinidis and Hosagrahar V Jagadish. Challenges and opportunities with big data. Proceedings of the VLDB Endowment, 5(12):2032–2033, 2012.
[15] Chang Liu, Chi Yang, Xuyun Zhang, and Jinjun Chen. External integrity verification for outsourced big data in cloud and iot: A big picture. Future Generation Computer Systems, 49:58–67, 2015.
[16] Katina Michael and Keith W Miller. Big data: New opportunities and new challenges [guest editors' introduction]. Computer, 46(6):22–24, 2013.
[17] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.
[18] Xue-Wen Chen and Xiaotong Lin. Big data deep learning: challenges and perspectives. IEEE Access, 2:514–525, 2014.
[19] CL Philip Chen and Chun-Yang Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on big data. Information Sciences, 275:314–347, 2014.
[20] Richard S Sutton and Andrew G Barto. Introduction to reinforcement learning, volume 135. MIT Press Cambridge, 1998.
[21] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning, pages 1096–1103. ACM, 2008.
[22] Steve Lohr. The age of big data. New York Times, 11(2012), 2012.