International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 11 | Nov 2019 www.irjet.net p-ISSN: 2395-0072

Comparative Study of Efficacy of Big Data Analysis and Deep Learning Techniques

Nivedita Lodha[1], Sharad Dabhi[2], Mitul Champaneri[3], G.T. Thampi[4]

[1,2,3]Student, Department of Information Technology, Thadomal Shahani Engineering College, Maharashtra, India
[4]Principal, Thadomal Shahani Engineering College, Maharashtra, India

---

Abstract - *Internet-driven software applications in the business have been pervasive and ubiquitous. These technology practices are generating data including video/ audio clips in large quantities which are uneven, heterogeneous and get generated at high speed. This accumulation of data amounts to big Data and encapsulating new knowledge and patterns. As the rise of the Internet-enabled digital devices permeates from desks to our palms and wrists, the big data analytics assumes a different paradigm and becomes increasingly critical. To reduce costs and improve velocity in analyzing such huge amounts of data, many techniques are developed in the technology marketplace. The developments of Artificial intelligence techniques to build intelligence to software/ hardware systems are perpetuated by technology practitioners as a way of furthering the cause product and process innovations. Business systems are trained to learn from a huge number of data sets underpinning the techniques described in the evolving Deep Learning literature. Both Big data analytics and Deep Learning Techniques are capable of processing a large number of datasets which leads to a constant comparison of its efficacy in different functional areas. Hence, this paper aims to compare different techniques and results obtained by Big Data Analysis and Deep Learning Techniques.*

Key Words — Big Data Analysis, Deep learning, Comparison, Neural Networks.

1. INTRODUCTION

Big data analytics is performed on large amounts of data to find hidden patterns, correlations and other insights. With today's technology, it's possible to analyze your data and get answers from it immediately. Big Data Analytics helps you to understand your organization better. With the use of Big data analytics, one can make informed decisions without blindly relying on guesses [1]. Big Data Analysis is thus mainly used to gain insights from data. We need new insight from data, not only for top-level executives but also can be used for providing better services to customers. One tool for reaching this aim is Deep Learning (DL). Deep Learning is a subarea of machine learning that deals with algorithms that have inspiration by the structure and function of the brain called artificial neural network (ANN). It is an application of ANN for such problems where learning tasks require more than one hidden layer. The "deep" word in deep learning refers to the depth of the network whereas the ANN can be very shallow[2] Deep Learning is a promising avenue of research into automated complex feature extraction at a high level of abstraction. Deep Learning is about learning multiple levels of representations and abstractions that help to make sense of data such as images, sound, and text [3]. Big Data Analysis thus focuses on the preprocessing stage of messy data, stores it, analyzes it. The significant data can be passed on to Deep Learning algorithms for intelligent outputs. Big Data Analysis can act as a prerequisite to Deep Learning Techniques as Deep Learning is an advanced big data technique. The paper is divided as follows:

2. STEPS OF BIG DATA ANALYSIS AND DEEP LEARNING TECHNIQUES

We start by giving a simple description of both the technologies which will show how both the processes differ.

2.1 Big Data Analysis

Big Data Analysis Process is nothing but gathering information by using proper application or tool which allows you to explore the data and find a pattern in it. Decisions can be made and ultimate conclusions can be obtained. The following phases are a part of data analysis:

1. Data Requirement Gathering-In this phase, you have to decide what to analyze and how to measure it, you have to understand why you are investigating and what measures you have to use to do this analysis.

2. Data Collection-The data is collected based on our requirements. The data can be streamed or it can be static data. Usually, for Big Data Analysis, the data which is streamed live and generated every millisecond is used.

3. Data Cleaning- The data collected may contain duplicate records, missing values, spaces or errors. Thus cleaning is a very essential step for normalizing the data

4. Data Analysis-As we manipulate data, we may find we have the exact information we need, or we

might need to collect more data. During this phase, we can use data analysis tools and software which will help you to understand, interpret, and derive conclusions based on the requirements. Apache Spark is one of the most popular big data analysis tools.

5. Data Visualization-Data visualization is very common in our day to day life; they often appear in the form of charts and graphs. The graphical visualization is easier for the human brain to process and understand. Data visualization is often used to discover unknown facts and trends. By observing relationships and comparing datasets, we can find a way to find out meaningful information. Tableau is a Big Data tool for Data visualization

2.2 Deep Learning

[5] Deep Learning is a branch of machine learning, where algorithms learn independently from excessive amounts of information. These algorithms get smarter when they are provided with more data. More data means more learning which is conceptually similar to the human brain. The computational model consists of multiple layers, called neural networks, where data is processed. The 3 elements in the neural network are: the input layer which is the input data we are supposed to analyze and work upon, at least 2 hidden layers, or nodes, which complete the computation with the deep learning algorithm. In the output layer, the calculated result is obtained.

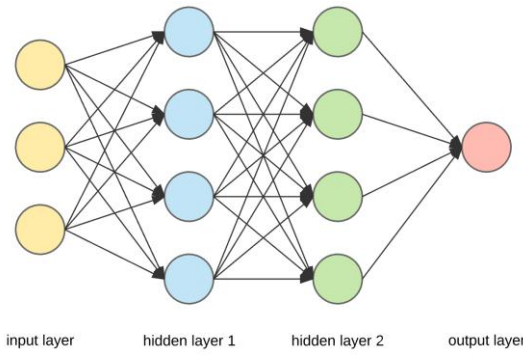

Fig-1. Deep Learning Neural Network

The main action happens in the hidden layers. The calculation is performed via connections, which contain the input data, a pre-assigned weight, and a path defined by the Activation Function [5].

The algorithm extracts information similar to each other and gets rid of the irrelevant information. The data from the previous layer is combined and it is then defined as relevant or irrelevant. The relevant output is transferred to the next node. Irrelevant information is discarded, which results in the reduction of information. If the information is undefined it remains relevant.

Deep learning has 2 main forms: supervised and unsupervised. In supervised learning, we tell the computer what the information we put in is. More efficient human input can be received this way but the calculation can be more efficient. Supervised learning is used with large amounts of well-defined data, like the weather. Unsupervised learning, on the other hand, works with unlabeled information. It makes predictions on currently available diverse data and finds a pattern in a seemingly disconnected environment [5]. Thus Big Data Analysis pulls from existing information to look for emerging patterns that can help make the decision making process. On the other hand, Deep Learning will from the existing data and provide the foundation required for the machine to teach itself.

3. DIFFERENCE BETWEEN THE DATASETS AND RESULTS OF BIG DATA ANALYSIS AND DEEP LEARNING

Operations of both the technologies were performed on different kinds of datasets to learn about the kind of output obtained.

3.1 Types of data analyzed in Big Data Analysis and Deep Learning Techniques

Big Data Analysis can be performed on structured, unstructured as well as semi-structured datasets. [6]Structured data is data whose elements are addressable for effective analysis. It is organized into a formatted repository which is generally known as a database. It concerns all data which can be stored in database SQL in the table with rows and columns eg: Excel datasets. Big Data Analysis can alone perform visualization of such dataset in less amount of time using various algorithms like K-Means clustering, Association rule mining, etc. [6]Unstructured data is data that is which is not organized in a predefined manner or does not have a predefined data model, thus it is not a good fit for a mainstream relational database eg: Audio, Video. Such a dataset is difficult for analysis and Apache Hadoop with MapReduce is used for structuring this unstructured data. After the dataset is structured, then analysis can be done using Apache Spark. [6]Semi-structured data is a data type that contains semantic tags, but does not conform to the structure associated with typical relational databases eg: XML, JSON files. Many Big Data solutions and tools have the ability to 'read' and process either JSON or XML. This reduces the complexity to analyze structured data, compared to unstructured data.

Deep learning can solve almost any problem of machine perception, including classifying data or making predictions about it. Deep learning is best applied to

unstructured data like images, video, sound or text. Deep Learning algorithms are mainly used for unstructured data unlike Big Data Analysis and the data need not be converted into a structured one. The outputs obtained by Deep Learning are completely different than the outputs obtained by Big Data Analysis.

3.2 Comparison of outputs of Textual Data

Big Data Analysis on structured data can be performed using Tableau tool which uses following algorithms:

- K Means Clustering
- Association Rules
- Linear Regression
- Logistic Regression
- Text Analysis

The visualization of data is in terms of pie charts, bar graphs, histograms, scatterplots, etc. These visualizations help see patterns and trends among the tuples of the dataset. An example of unstructured data analysis and pattern recognition is as follows:

The challenges faced by a global reinsurance company that processes half a billion pages of contracts annually. Because they can automatically process this unstructured content into a format that is usable by their analytics tools, they can feed the contract data into IBM's Watson and quickly assess risks and trends. An example of structured data is as follows:

A large dataset of sales of a retail store which consist of tuples of ids of products and the customers, customer name, area, product name, city, postal codes, etc can be clustered using the product name and city which can help find which product is sold in which city. The store can thus improve its marketing in areas where sales are less and give discounts to customers in areas where sales are more.

Deep Learning can be performed on the same dataset but it gives completely different output. It mainly performs future predictions by studying the existing patterns in the dataset. Deep Learning does predictions on a tabular dataset with the Fastai API using PyTorch that will create a Tabular neural network model to match data. The data needs to be in a Pandas data frame, which is the standard format for tabular data in python. Pandas data frames can read data from many data stores including CSV, relational databases, Spark and Hadoop[8]. The predictions cannot be made using Big Data Analysis Techniques.

Deep Learning can be performed on textual data also by using Recurrent Neural Networks(RNNs). Deep Learning not only observes patterns but also generates patterns. Textual data is sequential data hence ANN cannot be used alone. An RNN is a neural network with internal short-term memory. One example of deep learning on textual data is generating song lyrics by studying patterns of a songwriter in his/her previous songs. Our model has to have a short-term memory as well as a long-term memory. The reason is that we want our network to remember parts of the sequence it has seen many time steps back. We will, therefore, need a special kind of RNN called Long short-term memory (LSTM)[7]. Also, the model uses the concept of Markov Chains[11]. This is one of the most important parts for generating the lyrics as it calculates and predict the probability of the next word to be used. This type of text generation cannot be performed using Big Data Analysis Techniques.

COMPARISON OF APPLICATION IN SOCIAL MEDIA: Of the customer-facing Big Data application examples could discuss, analysis of social media activity is one of the most important. Everyone and their mothers are on social media these days, whether they like company pages on Facebook or tweeting complaints about products on Twitter. A big data solution built to produce and investigates social media activity, like IBM's Cognos Consumer Insights, a fact solution running on IBM's big Insights big data platform, may make sense of the chatter. Social media data can provide real-time insights into how the market is responding to products and campaigns. With these insights, companies adjust their costs, methods of promotion, and campaign placement to achieve optimal results[15].

Deep learning techniques let a machine to learn to classify data by itself; for instance, a deep learning image analysis tool can learn to recognize images that contain cats, without actually being told what a cat looks like. By analyzing a large number of images, it can learn from the context of the image – what else is likely to be present in an image of a cat? What text or metadata might suggest that an image consists of a cat? This gives a structure to unstructured data, which makes deep learning an extremely valuable tool for a company like Facebook. This is just one example of how the social network uses deep learning to understand more about its users, deep neural networks – the foundation stones of deep learning – are used to decide which adverts to show to which users. This has always been a core part of Facebook's business, but by tasking machines themselves to find out as much as they can about us, and to cluster us together in the most insightful ways when serving us ads, it hopes to maintain an edge against other high-tech competitors, such as Google, who are fighting for supremacy of the same advertising market.

3.3 Big Data Analysis and Deep Learning on Images

Big Data algorithms cannot perform analysis on images alone so Machine Learning algorithms are used to learn complex models of images. One such application is biomedical imaging which is a part of the health sector [9]. The main goal of big data management is to make sure a high level of data quality is maintained and there is accessibility for business intelligence and big data applications.

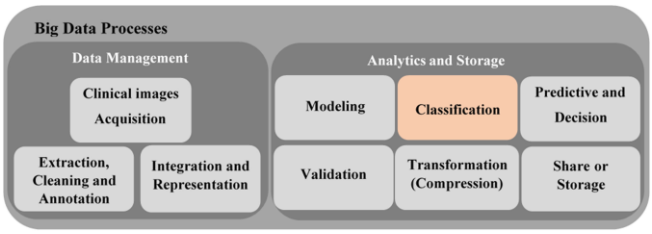

Fig-2. Big Data processes for Biomedical image processing

The big data process is as follows:

The captured biomedical images are transferred to the big data platform for processing. Extraction is performed to obtain images from raw data, cleaning to remove noise from images and annotations for adding some information concerning the patient on images. Integration and representation are done using clustering in the database[9].

Deep Learning process is as follows:

Modeling is used to format images in a way that is easier to understand. The classification based on the training data set containing observations is done once the category membership is known. An example would be to assign a given biomedical image into "anatomic body part" or "biological systems" classes[9]. Prediction and Decision are made using Convolutional Neural Networks(CNNs). Validation is performed using specificity and sensitivity. Transformation is performed for compressing the data for reducing cost and computational power, management efficiency and querying. The data is then stored in NoSQL databases. This is one application where Big Data and Deep Learning algorithms are used together for image analysis and processing. However, deep learning can be individually used for image classification but the same cannot be performed using Big data techniques. The following is an example of how Deep Learning classifies images:

The computer is taught how to recognize and classify images in given categories.

Firstly, it is necessary to teach the computer how all these objects of different categories look like before it being able to recognize a new object. The more objects the computer sees, the better it gets in recognizing that object. This is known as supervised learning. This task can be carried out by labeling the images, the computer will start recognizing patterns present in object pictures that are absent from other ones and will start building its cognition.

Python and Tensorflow(framework developed by Google for deep learning) are used to write programs for this classification technique. CNNs are used for image classification and improving its accuracy. It is a special type of Neural Networks that works in the same way as a regular neural network except that it has a convolution layer at the beginning.

Instead of feeding the entire image as an array of numbers, the image is broken up into several tiles, the machine then tries to predict what each tile is. At last, the computer tries to predict what is in the picture based on the prediction data of all the tiles. Thus, wherever the object is located in the image, the computer can parallelize all the operations and detect the object [10].

The first step is data pre-processing and to add noise to the images by rotating the image, cropping, adjusting hue and saturation, etc. The second step is splitting the datasets so that operations could be performed faster. A neural network is implemented after these two steps. Max pooling is used to reduce the dimensions of an image by taking the maximum pixel value of a grid. The Neural networks once trained can be used to classify the input images into various categories[10].

Classification of images involves neural networks and learning which is not a part of Big Data Analysis. Processed Big Data is an input to the Deep Learning algorithm in this case.

3.4 Deep Learning in Big Data Analysis

While working with Big data we face problem such as format variation, speed of data, robustness of the analysis algorithms, multiple sources of input, maintaining the quality of data and filtration of noise, high dimensionality, scalability of algorithms, imbalanced input data, unsupervised and un-categorized data, limited supervised/labelled data, managing data storage etc. Deep learning provides a highly efficient solution to problems like semantic indexing, data tagging, fast information retrieval, and discriminative modeling [12].

The representation of data is an important factor for the efficiency of a procedure like big data analysis that deals with massive volumes of data. Better and more organized the data will be- better will be the efficiency of the system analyzing this data. Deep learning technologies when used in big data analysis for getting a predictive solution, can extract complex level abstractions embedded in the data, produces a decorated set of the data that is well

represented and given as input to the next stage. This is done without interacting with any human beings. High levels of abstraction are built on the basis of lower levels of abstraction. The salient feature of high levels of abstraction of data is that the higher the level of abstraction, lesser will be the variation with changes in the input data. The aim of deep learning is to produce more and more of these invariant features from the given input so that they could be used in the predictive analysis of data. The accuracy of predictive analysis of any data set varies proportionally with the number of invariant features. Deep Learning algorithms can also segregate different mixed features from raw data which were actually formed because the data had been generated from more than one source. Deep Learning algorithms work on architectures consisting of many layers stacked on top of each other, where the input to one layer is the output produced by the layer below it. Non-Linear transformations are applied to the raw data set to produce supervised data as an output. The quality in which data has been represented becomes better with the increase in the number of layers in the deep architecture through which the data set has been processed. The complexity of the transformation functions increases with an increase in the number of layers. The final output obtained from the last layer of the deep architecture provides us with features on the basis of which the data is finally classified. This sorted data is then indexed and hashed. Deep Learning algorithms help to make processes like indexing and hashing feasible with big data as these processes can be implemented practically with pre-processed data and not with raw data.

SEMANTIC INDEXING: This algorithm focuses highly on the presentation of data, as the poor presentation will affect the complexity of further layers. Semantic Indexing represents each word of a document as a vector. Vectors having similar values signify words with similar meanings. This procedure also helps to reduce the dimensionality of the given data set. These vectors are given as input to other layers, the final product is a binary code for each data set. Comparatively shorter binary codes are produced when this algorithm is implemented with deep learning. Data sets with similar binary codes can be stored sequentially as they would, in most cases possess similar semantic characteristics. Data retrieval happens on basis of hamming distances of the binary codes of the data set. This algorithm is usually carried out with text documents [13].

SEMANTIC TAGGING: This algorithm filters files as per their semantic value. Generally, all the files in search engines like Google, Wikipedia are filtered using their file name that is why many times some results of the search we do are irrelevant. By Semantic tagging, as the name says, we tag images by their semantic value. This method has proved to be more efficient in terms of user experience as it prevents irrelevancy. Semantic tagging, therefore, sorts files to a greater extent, as compared to other methods. This algorithm is used more frequently with audio and image files. The output of this algorithm can be given as an input to the indexing stage[13].

CONDUCTING DISCRIMINATIVE TASKS: While performing discriminative tasks in Big Data Analysis, Learning algorithms allow users to extract complicated nonlinear features from the raw data. It also provides the use of linear models so that they can perform discriminative tasks by inputting the extracted features. This approach has two advantages: Firstly, by extracting features with Deep Learning adds nonlinearity to the data analysis, thereby associating discriminative tasks closely to AI, and secondly applying linear analytical models on extracted features is computationally more efficient. These two benefits are important for Big Data because it allows practitioners in accomplishing complicated tasks related to Artificial Intelligence like object recognition in images, image comprehension, etc[14].

As discussed earlier, it is next to impossible to manage and analyze data of this magnitude using conventional or existing technologies. Therefore, these existing technologies are being integrated with machine learning techniques and hardware of high computational power to make the task easier; this is where deep learning gets involved with big data analysis[13].

4. FUTURE SCOPE OF BIG DATA ANALYSIS AND DEEP LEARNING

Inherently, deep learning is defined as an advanced application of AI in interconnected machines and peripherals by granting them access to databases and making them learn new things from it on their own in a programmed manner without a need for structured data. As the size of big data is continuously growing and new grounds are being broken in analyzing its implications as well, it is becoming more meaningful and contextually relevant for the machines to have a better idea of their functions with the help of big data analysis.

One of the major concerns people are having about AI today is that it will minimize human requirement in all the job sectors as most of the work will be done by robots and AI-based computers in the future. When observed with the role to be played by big data in the future, the truth is far from it. The sentimental and emotional big data analysis will always require human intelligence as the machines lack emotional intelligence and decision-making abilities based on sentiments. Hence, the increasing collaboration between AI, machine learning, deep learning, and big data will only make way for talented and capable human data scientists to consistently evolve and rise in the market, and by all means, they will be required in a huge number as the applications of these technologies gain movement.

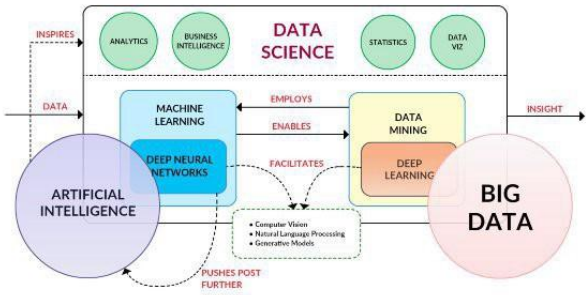

Fig-3. Integration of AI and Big Data

In addition to the problem of handling massive volumes of data, large-scale Deep Learning models for Big Data Analysis also have to contend with other Big Data problems, such as domain adaptation and streaming data. This leads to the need for further innovations in large-scale models for Deep Learning algorithms and architectures[15].

5. CONCLUSION

In conclusion, Big Data Analysis and Deep Learning are two different technologies that can provide different results if used on a similar dataset. Deep Learning is an advanced Big Data Analytics Technique and it uses output provided by Big Data algorithms as an input for the unsupervised or supervised learning. Both the technologies can be applied individually and together in a domain but the purpose of Big Data Analysis is mainly to find and observe hidden patterns while the main purpose of Deep Learning is to generate patterns by observing them ie. make future predictions.

REFERENCES

[1] Agiratech.com, 'An introduction to Big Data Analytics', 2019. https://www.agiratech.com/introduction-to-big-data-analytics/. [Accessed: 29- Oct- 2019].

[2] Dixit, M., Tiwari, A., Pathak, H., & Astya, R. (2018). An overview of deep learning architectures, libraries and its applications areas. 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN). doi:10.1109/icacccn.2018.8748442

[3] Gheisari, M., Wang, G., & Bhuiyan, M. Z. A. (2017). A Survey on Deep Learning in Big Data. 22017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC). doi:10.1109/cse-euc.2017.215

[4] Guru99.com,'What is Data Analysis? Types, Process, Methods, Techniques', 2019. https://www.guru99.com/what-is-data-analysis.html. [Accessed: 27- September- 2019].

[5] Datadriveninvestor.com,'Deep Learning Explained in 7 steps', 2019. https://www.datadriveninvestor.com/deep-learning-explained/. [Accessed: 27- September- 2019].

[6] Bigdataframework.com, 'Data Types: Structured vs. Unstructured Data', https://www.bigdataframework.org/data-types-structured-vs-unstructured-data/. [Accessed: 22- August- 2019].

[7] Medium.com,'Using AI to generate lyrics',2016. https://medium.com/@ivanliljeqvist/using-ai-to-generate-lyrics-5aba7950903. [Accessed: 29- September- 2019].

[8] Towardsdatascience.com, 'Tabular data analysis with deep neural nets', 2019.https://towardsdatascience.com/tabular-data-analysis-with-deep-neural-nets-d39e10efb6e0. [Accessed: 29- September- 2019].

[9] Sciencedirect.com, 'An optimal big data workflow for biomedical image analysis', https://www.sciencedirect.com/science/article/pii/S2352914818300844. [Accessed: 31- October- 2019].

[10] Medium.com, 'Image Classification using Deep Neural Networks', 2017. https://medium.com/@tifa2up/image-classification-using-deep-neural-networks-a-beginner-friendly-approach-using-tensorflow-94b0a090ccd4. [Accessed: 29- October- 2019].

[11] Wikipedia.org, 'Markov Chains', 2016. https://en.wikipedia.org/wiki/Markov_chain.[Accessed: 29- October- 2019].

[12] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald and Edin Muharemagic. "Deep learning applications and challenges in big data analytics." Najafabadi et al. Journal of Big Data (2015) 2:1, DOI 10.1186/s40537-014-0007-7

[13] Lekhrajani, S., & Krishna Samdani, P. (2018). A Review of Implementation of Deep Learning in Big Data Analysis. 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence). doi:10.1109/confluence.2018.8442719

[14] Allerin.com, '3 applications of Deep Learning in Big Data analytics', 2017.https://www.allerin.com/blog/3-applications-of-deep-learning-in-big-data-analytics [Accessed: 23- August- 2019]

[15] Praveena, M. D. A., & Bharathi, B. (2017). A survey paper on big data analytics. 2017 International Conference on Information Communication and Embedded Systems (ICICES). doi:10.1109/icices.2017.