IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 1, January 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 11, Issue 1, January 2022 ||

| DOI:10.15680/IJIRSET.2022.1101126 |

Deep Learning Applications and Challenges in Big Data Analytics: A Survey

[*1]Oluyinka I Omotosho, [2]Gabriel O. Olatoye

Department of Cyber Security Science, Ladoke Akintola University of Technology, Ogbomosho - Nigeria[1]

Department of Computer Science, Ladoke Akintola University of Technology, Ogbomosho - Nigeria[2]

*Corresponding Author: oiomotosho@lautech.edu.ng

ABSTRACT: Big Data Analytics and Deep Learning are two high-focuses 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.This work explores the depth in utilizing Deep Learning in 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. 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 were also investigated. This paper concluded by presenting insights into relevant future works, which posesthe questions of defining data sampling criteria, domain adaptation modeling, defining criteria for obtaining useful data abstractions, improving semantic indexing, semi-supervised learning, and active learning.

I. INTRODUCTION

The general focus of machine learning is the representation of the input data and generalization of the learnt patterns for use on future unseen data. The goodness of the data representation has a large impact on the performance of machine learners on the data: a poor data representation is likely to reduce the performance of even an advanced, complex machine learner, while a good data representation can lead to high performance for a relatively simpler machine learner. Thus, feature engineering, which focuses on constructing features and data representations from raw data, is an important element of machine learning. Feature engineering consumes a large portion of the effort in a machine learning task, and is typically quite domain specific and involves considerable human input. For example, the Histogram of Oriented Gradients (HOG) and Scale Invariant Feature Transform (SIFT) are popular feature engineering algorithms developed specifically for the computer vision domain. Performing feature engineering in a more automated and general fashion would be a major breakthrough in machine learning as this would allow practitioners to automatically extract such features without direct human input. Deep Learning algorithms are one promising avenue of research into the automated extraction of complex data representations (features) at high levels of abstraction. Such algorithms develop a layered, hierarchical architecture of learning and representing data, where higher-level (more abstract) features are defined in terms of lower-level (less abstract) features. The hierarchical learning architecture of Deep Learning algorithms is motivated by artificial intelligence emulating the deep, layered learning process of the primary sensorial areas of the neocortex in the human brain, which automatically extracts features and abstractions from the underlying data . Deep Learning algorithms are quite beneficial when dealing with learning from large amounts of unsupervised data, and typically learn data representations in a greedy layer-wise fashion . Empirical studies have demonstrated that data representations obtained from stacking up non-linear feature extractors (as in Deep Learning) often yield better machine learning results, e.g., improved classification modeling , better quality of generated samples by generative probabilistic models , and the invariant property of data representations. Deep Learning solutions have yielded outstanding results in different machine learning applications, including speech recognition , computer vision, and natural language processing . A more detailed overview of Deep Learning is presented in Section “Deep learning in data mining and machine learning”. Big Data represents the general realm of problems and techniques used for application domains that collect and maintain massive volumes of raw data for domain-specific data analysis. Modern data-intensive technologies as well as increased

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 11, Issue 1, January 2022 ||

| DOI:10.15680/IJIRSET.2022.1101126 |

computational and data storage resources have contributed heavily to the development of Big Data science . Technology based companies such as Google, Yahoo, Microsoft, and Amazon have collected and maintained data that is measured in exabyte proportions or larger. Moreover, social media organizations such as Facebook, YouTube, and Twitter have billions of users that constantly generate a very large quantity of data. Various organizations have invested in developing products using Big Data Analytics to addressing their monitoring, experimentation, data analysis, simulations, and other knowledge and business needs , making it a central topic in data science research. Mining and extracting meaningful patterns from massive input data for decision making, prediction, and other inferencing is at the core of Big Data Analytics. In addition to analyzing massive volumes of data, Big Data Analytics poses other unique challenges for machine learning and data analysis, including format variation of the raw data, fast moving streaming data, trustworthiness of the data analysis, highly distributed input sources, noisy and poor quality data, high dimensionality, scalability of algorithms, imbalanced input data, unsupervised and un-categorized data, limited supervised/labeled data, etc. Adequate data storage, data indexing/tagging, and fast information retrieval are other key problems in Big Data Analytics. Consequently, innovative data analysis and data management solutions are warranted when working with Big Data. For example, in a recent work we examined the high-dimensionality of bioinformatics domain data and investigated feature selection techniques to address the problem . A more detailed overview of Big Data Analytics is presented in Section “Big data analytics”. The knowledge learnt from (and made available by) Deep Learning algorithms has been largely untapped in the context of Big Data Analytics. Certain Big Data domains, such as computer vision and speech recognition, have seen the application of Deep Learning largely to improve classification modeling results. The ability of Deep Learning to extract high-level, complex abstractions and data representations from large volumes of data, especially unsupervised data, makes it attractive as a valuable tool for Big Data Analtyics. More specifically, Big Data problems such as semantic indexing, data tagging, fast information retrieval, and discriminative modeling can be better addressed with the aid of Deep Learning. More traditional machine learning and feature engineering algorithms are not efficient enough to extract the complex and non-linear patterns generally observed in Big Data. By extracting such features, Deep Learning enables the use of relatively simpler linear models for Big Data analysis tasks, such as classification and prediction, which is important when developing models to deal with the scale of Big Data. The novelty of this study is that it explores the application of Deep Learning algorithms for key problems in Big Data Analytics, motivating further targeted research by experts in these two fields.

The paper focuses on two key topics: (1) how Deep Learning can assist with specific problems in Big Data Analytics, and (2) how specific areas of Deep Learning can be improved to reflect certain challenges associated with Big Data Analytics. With respect to the first topic, we explore the application of Deep Learning for specific Big Data Analytics, including learning from massive volumes of data, semantic indexing, discriminative tasks, and data tagging. Our investigation regarding the second topic focuses on specific challenges Deep Learning faces due to existing problems in Big Data Analytics, including learning from streaming data, dealing with high dimensionality of data, scalability of models, and distributed and parallel computing. We conclude by identifying important future areas needing innovation in Deep Learning for Big Data Analytics, including data sampling for generating useful high-level abstractions, domain (data distribution) adaption, defining criteria for extracting good data representations for discriminative and indexing tasks, semi-supervised learning, and active learning. The remainder of the paper is structured as follows: Section “Deep learning in data mining and machine learning” presents an overview of Deep Learning for data analysis in data mining and machine learning; Section “Big data analytics” presents an overview of Big Data Analytics, including key characteristics of Big Data and identifying specific data analysis problems faced in Big Data Analytics; Section “Applications of deep learning in big data analytics” presents a targeted survey of works investigating Deep Learning based solutions for data analysis, and discusses how Deep Learning can be applied for Big Data Analytics problems; Section “Deep learning challenges in big data analytics” discusses some challenges faced by Deep Learning experts due to specific data analysis needs of Big Data; Section “Future work on deep learning in big data analytics” presents our insights into further works that are necessary for extending the application of Deep Learning in Big Data, and poses important questions to domain experts; and in Section “Conclusion” we reiterate the focus of the paper and summarize the work presented.

II. RELATED WORK

This section describes the foundation of this review by discussing several of fields that have been applied with Deep Learning algorithm.

2.1 Automatic Speech Recognition (ASR)

Google has announced that Google voice search had taken a new turn by adopting Deep Neural Networks (DNN) as the core technology used to model the sounds of a language in 2012 [8]. DNN replaced Gaussian Mixture Model which has

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 11, Issue 1, January 2022 ||

| DOI:10.15680/IJIRSET.2022.1101126 |

been in the industry for 30 years. DNN also has proved that it is better able to measure which sound a user is fabricating at every instant in time and with this they delivered prominently increased speech recognition accuracy. In 2013, DL has gained full momentum in both ASR and ML [9]. DL is basically linked to the use of multiple layers of nonlinear transformations to derive speech features, while learning with shallow layers comprises the use of exemplar-based representations for speech features which have high dimensionality but typically vacant entries.

2.2 Image Recognition

Deep max-pooling convolutional neural networks is used to detect mitosis in breast histology images was presented in [10]. Mitosis detection is very hard. In fact, mitosis is a complex process during which a cell nucleus undergoes various transformations. In this approach, DNN as powerful pixel classifier which operates on raw pixel values and no human input is needed. Hence, DNN automatically learns a set of visual features from the training data. DNN is tested on a publicly available dataset and significantly outperforms all competing techniques, with manageable computational effort: processing a 4MPixel image requires few minutes on a standard laptop. Large and deep convolutional neural network is trained to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into 1000 different classes [11]. On the test data, they achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. From all the experiments, the results can be improved simply by waiting for faster GPUs and bigger datasets to become available.

2.3 Natural Language Processing

Recently, deep learning methods have been successfully applied to a variety of language and information retrieval applications. By exploiting deep architectures, deep learning techniques are able to discover from training data the hidden structures and features at different levels of abstractions useful for the any tasks. In 2013, [12] proposed a series of Deep Structured Semantic Models (DSSM) for Web search. More specifically, they uses a DNN to rank a set of documents for a given query as follows. First, a non-linear projection is performed to map the query and the documents to a common semantic space. Then, the relevance of each document given the query is calculated as the cosine similarity between their vectors in that semantic space. The neural network models are discriminatively trained using the click-through data such that the conditional likelihood of the clicked document given the query is maximized. The new models are evaluated on a Web document ranking task using a real-world data set. Results show that the proposed model significantly outperforms other latent semantic models, which were considered state-of-the-art in the performance prior to the work presented in [12].

2.4 Drug Discovery and Toxicology

Quantitative Structure Analysis/Prediction Studies (QSAR/QSPR) attempt to build mathematical models relating physical and chemical properties of compounds to their chemical structure. In [13], multi-task learning is applied to QSAR using various neural network models. They used an artificial neural network to learn a function that predicts activities of compounds for multiple assays at the same time. The method is compared with alternative methods and reported that the neural nets with multi-tasking can lead to significantly improved results over baselines generated with random forests. In 2015, AtomNet has been introduced as first structure-based, deep convolutional neural network which designed to predict the bioactivity of small molecules for drug discovery applications [14]. This paper also demonstrates how to apply the convolutional concepts of feature locality and hierarchical composition to the modeling of bioactivity and chemical interactions. AtomNet outperforms previous docking approaches on a diverse set of benchmarks by a large margin, achieving an AUC greater than 0.9 on 57.8% of the targets in the DUDE benchmark.

2.5 Customer Relationship Management

A framework for autonomous control of a customer relationship management system been charted by [15].First, a modified version of the widely accepted Recency-Frequency- Monetary Value system of metrics can be used to define the state space of clients or donors is explored. Second, a procedure to determine the optimal direct marketing action in discrete and continuous action space for the given individual, based on his position in the state space is described. The procedure involves the use of model-free Q-learning to train a deep neural network that relates a client's position in the state space to rewards associated with possible marketing activities. The estimated value function over the client state space can be interpreted as customer lifetime value (CLV), and thus allows for a quick plug-in estimation of CLV for a given client. Experimental results are presented, based on Knowledge Discovery and Data Mining Tools Competition, mailing dataset of donation solicitations.

2.6 Recommendation Systems

Automatic music recommendation has become an increasingly relevant problem in recent years, since a lot of music is now sold and consumed digitally. Most recommender systems rely on collaborative filtering. In 2013, [16] proposed to

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 11, Issue 1, January 2022 ||

| DOI:10.15680/IJIRSET.2022.1101126 |

use a latent factor model for recommendation, and predict the latent factors from music audio when they cannot be obtained from usage data. Traditional approach is compared using a bag-of-words representation of the audio signals deep convolutional neural networks, and the predictions is evaluated by quantitatively and qualitatively on the Million Song Dataset. The result shows that the recent advances in DL translate very well to the music recommendation setting, with deep convolutional neural networks significantly outperforming the traditional approach. Recent online services heavily on automatic personalization to recommend relevant content to a large number of users. This requires systems scale promptly to accommodate the stream of new users visiting the online services for the first time. Work by [17] in 2015 proposed a content-based recommendation system to address both the recommendation quality and the system scalability. They also proposed to use a rich feature set to represent users, according to their web browsing history and search queries. They use a DL approach to map users and items to a latent space where the similarity between users and their preferred items is maximized. Scalability analysis show that the multi-view DNN model can easily scale to encompass millions of users and billions of item entries.

2.7 Bioinformatics

The annotation of genomic information is a major challenge in biology and bioinformatics. Existing databases of known gene functions are incomplete and prone to errors, and the bimolecular experiments needed to improve these databases are slow and costly. While computational methods are not a substitute for experimental verification, they can help in two ways: algorithms can aid in the curation of gene annotations by automatically suggesting inaccuracies, and they can predict previously-unidentified gene functions, accelerating the rate of gene function discovery. In this work [18], an algorithm that achieves both goals using deep auto encoder neural networks is developed. With experiments on gene annotation data from the Gene Ontology project, it shows that deep auto encoder networks achieve better performance than other standard machine learning methods, including the popular truncated singular value decomposition.

III. METHEDOLOGY

3.1 Deep learning in data mining and machine learning

The main concept in deep leaning algorithms is automating the extraction of representations (abstractions) from the data [5]. Deep learning algorithms use a huge amount of unsupervised data to automatically extract complex representation. These algorithms are largely motivated by the field of artificial intelligence, which has the general goal of emulating the human brain's ability to observe, analyze, learn, and make decisions, especially for extremely complex problems. Work pertaining to these complex challenges has been a key motivation behind Deep Learning algorithms which strive to emulate the hierarchical learning approach of the human brain. Models based on shallow learning architectures such as decision trees, support vector machines, and case-based reasoning may fall short when attempting to extract useful information from complex structures and relationships in the input corpus. In contrast, Deep Learning architectures have the capability to generalize in non-local and global ways, generating learning patterns and relationships beyond immediate neighbors in the data [4]. Deep learning is in fact an important step toward artificial intelligence. It not only provides complex representations of data which are suitable for AI tasks but also makes the machines independent of human knowledge which is the ultimate goal of AI. It extracts representations directly from unsupervised data without human interference. A key concept underlying Deep Learning methods is distributed representations of the data, in which a large number of possible configurations of the abstract features of the input data are feasible, allowing for a compact representation of each sample and leading to a richer generalization. The number of possible configurations is exponentially related to the number of extracted abstract features. Noting that the observed data was generated through interactions of several known/unknown factors, and thus when a data pattern is obtained through some configurations of learnt factors, additional (unseen) data patterns can likely be described through new configurations of the learnt factors and patterns [5,24]. Compared to learning based on local generalizations, the number of patterns that can be obtained using a distributed representation scales quickly with the number of learnt factors. Deep learning algorithms lead to abstract representations because more abstract representations are often constructed based on less abstract ones. An important advantage
of more abstract representations is that they can be invariant to the local changes in the input data. Learning such invariant features is an ongoing major goal in pattern recognition (for example learning features that are invariant to the face orientation in a face recognition task). Beyond being invariant such representations can also disentangle the factors of variation in data. The real data used in AI-related tasks mostly arise from complicated interactions of many sources. For example an image is composed of different sources of variations such a light, object shapes, and object materials. The abstract representations provided by deep learning algorithms can separate the different sources of variations in data. Deep learning algorithms are actually Deep architectures of consecutive layers. Each layer applies a nonlinear transformation on its input and provides a representation in its output. The objective is to learn a complicated and abstract representation of the data in a hierarchical manner by passing the data through multiple transformation layers. The

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 11, Issue 1, January 2022 ||

| DOI:10.15680/IJIRSET.2022.1101126 |

sensory data (for example pixels in an image) is fed to the first layer. Consequently the output of each layer is provided as input to its next layer. Stacking up the nonlinear transformation layers is the basic idea in deep learning algorithms. The more layers the data goes through in the deep architecture, the more complicated the nonlinear transformations which are constructed. These transformations represent the data, so Deep Learning can be considered as special case of representation learning algorithms which learn representations of the data in a Deep Architecture with multiple levels of representations. The achieved final representation is a highly non-linear function of the input data.

3.2 Big Data Analytics

Big Data generally refers to data that exceeds the typical storage, processing, and computing capacity of conventional databases and data analysis techniques. As a resource, Big Data requires tools and methods that can be applied to analyze and extract patterns from large-scale data. The rise of Big Data has been caused by increased data storage capabilities, increased computational processing power, and availability of increased volumes of data, which give organization more data than they have computing resources and technologies to process. In addition to the obvious great volumes of data, Big Data is also associated with other specific complexities, often referred to as the four Vs: Volume, Variety, Velocity, and Veracity [31]. We note that the aim of this section is not to extensively cover Big Data, but present a brief overview of its key concepts and challenges while keeping in mind that the use of Deep Learning in Big Data Analytics is the focus of this paper. The unmanageable large Volume of data poses an immediate challenge to conventional computing environments and requires scalable storage and a distributed strategy to data querying and analysis. However, this large Volume of data is also a major positive feature of Big Data. Many companies, such as Facebook, Yahoo, Google, already have large amounts of data and have recently begun tapping into its benefits [12]. A general theme in Big Data systems is that the raw data is increasingly diverse and complex, consisting of largely un-categorized/unsupervised data along with perhaps a small quantity of categorized/supervised data. Working with the Variety among different data representations in a given repository poses unique challenges with Big Data, which requires Big Data preprocessing of unstructured data in order to extract structured/ordered representations of the data for human and/or downstream consumption. In today's data-intensive technology era, data Velocity – the increasing rate at which data is collected and obtained – is just as important as the Volume and Variety characteristics of Big Data. While the possibility of data loss exists with streaming data if it is generally not immediately processed and analyzed, there is the option to save fast-moving data into bulk storage for batch processing at a later time. However, the practical importance of dealing with Velocity associated with Big Data is the quickness of the feedback loop, that is, process of translating data input into useable information. This is especially important in the case of time-sensitive information processing. Some companies such as Twitter, Yahoo, and IBM have developed products that address the analysis of streaming data]. Veracity in Big Data deals with the trustworthiness or usefulness of results obtained from data analysis, and brings to light the old adage "Garbage-In-Garbage-Out" for decision making based on Big Data Analytics. As the number of data sources and types increases, sustaining trust in Big Data Analytics presents a practical challenge. Big Data Analytics faces a number of challenges beyond those implied by the four Vs While not meant to be an exhaustive list, some key problem areas include: data quality and validation, data cleansing, feature engineering, high-dimensionality and data reduction, data representations and distributed data sources, data sampling, scalability of algorithms, data visualization, parallel and distributed data processing, real-time analysis and decision making, crowdsourcing and semantic input for improved data analysis, tracing and
analyzing data provenance, data discovery and integration, parallel and distributed computing, exploratory data analysis and interpretation, integrating heterogenous data, and developing new models for massive data computation.

3.3 Applications Of Deep Learning In Big Data Analytics

As stated previously, Deep Learning algorithms extract meaningful abstract representations of the raw data through the use of an hierarchical multi-level learning approach, where in a higher-level more abstract and complex representations are learnt based on the less abstract concepts and representations in the lower level(s) of the learning hierarchy. While Deep Learning can be applied to learn from labeled data if it is available
in sufficiently large amounts, it is primarily attractive for learning from large amounts of unlabeled/unsupervised data [4,5], making it attractive for extracting meaningful representations and patterns from Big Data. Once the hierarchical data abstractions are learnt from unsupervised data with Deep Learning, more conventional discriminative models can be trained with the aid of relatively fewer supervised/labeled data points, where the labeled data is typically obtained through human/expert input. Deep Learning algorithms are shown to perform better at extracting non-local and global relationships and patterns in the data, compared to relatively shallow learning architectures [4]. Other useful characteristics of the learnt abstract representations by Deep Learning include: (1) relatively simple linear models can work effectively with the knowledge obtained from the more complex and more abstract data representations, (2) increased automation of data representation extraction from unsupervised data enables its broad application to different

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 11, Issue 1, January 2022 ||

| DOI:10.15680/IJIRSET.2022.1101126 |

data types, such as image, textural, audio, etc., and (3) relational and semantic knowledge can be obtained at the higher levels of abstraction and representation of the raw data. While there are other useful aspects of Deep Learning based representations of data, the specific characteristics mentioned above are particularly important for Big Data Analytics.

Considering each of the four Vs of Big Data characteristics, i.e., Volume, Variety, Velocity, and Veracity, Deep Learning algorithms and architectures are more aptly suited to address issues related to Volume and Variety of Big Data Analytics. Deep Learning inherently exploits the availability of massive amounts of data, i.e. Volume in Big Data, whereal gorithms with shallow learning hierarchies fail to explore and understand the higher complexities of data patterns. Moreover, since Deep Learning deals with data abstraction and representations, it is quite likely suited for analyzing raw data presented in different formats and/or from different sources, i.e. Variety in Big Data, and may minimize need
for input from human experts to extract features from every new data type observed in Big Data. While presenting different challenges for more conventional data analysis approaches, Big Data Analytics presents an important opportunity for developing novel algorithms and models to address specific issues related to Big Data. Deep Learning concepts provide one such solution venue for data analytics experts and practitioners. For example, the extracted representations by Deep Learning can be considered as a practical source of knowledge for decision-making, semantic indexing, information retrieval, and for other purposes in Big Data Analytics, and in addition, simple linear modeling techniques can be considered for Big Data Analytics when complex data is represented in higher forms of abstraction.

In the remainder of this section, we summarize some important works that have been performed in the field of Deep Learning algorithms and architectures, including semantic indexing, discriminative tasks, and data tagging. Our focus is that by presenting these works in Deep Learning, experts can observe the novel applicability of Deep Learning techniques in Big Data Analytics, particularly since some of the application domains in the works presented involve large scale data. Deep Learning algorithms are applicable to different kinds of input data; however, in this section we focus on its application on image, textual, and audio data.

3.4 Semantic Indexing

A key task associated with Big Data Analytics is information retrieval [8]. Efficient storage and retrieval of information is a growing problem in Big Data, particularly since very large-scale quantities of data such as text, image, video, and audio are being collected and made available across various domains, e.g., social networks, security systems, shopping
and marketing systems, defense systems, fraud detection, and cyber traffic monitoring. Previous strategies and solutions for information storage and retrieval are challenged by the massive volumes of data and different data representations, both associated with Big Data. In these systems, massive amounts of data are available that needs semantic indexing rather than being stored as data bit strings. Semantic indexing presents the data in a more efficient manner and makes it useful as a source for knowledge discovery and comprehension, for example by making search engines work more quickly and efficiently. Instead of using raw input for data indexing, Deep Learning can be used to generate high-level abstract data representations which will be used for semantic indexing. These representations can reveal complex associations and factors (especially when the raw input was Big Data), leading to semantic knowledge and understanding. Data representations play an important role in the indexing of data, for example by allowing data points/instances with relatively similar representations to be stored closer to one another in memory, aiding in efficient information retrieval. It should be noted, however, that the high-level abstract data representations need to be meaningful and demonstrate relational and semantic association in order to actually confer a good semantic understanding and comprehension of the input.

3.5 Deep Learning Challenges In Big Data Analytics

The prior section focused on emphasizing the applicability and benefits of Deep Learning algorithms for Big Data Analytics. However, certain characteristics associated with Big Data pose challenges for modifying and adapting Deep Learning to address those issues. This section presents some areas of Big Data where Deep Learning needs further exploration, specifically, learning with streaming data, dealing with high-dimensional data, scalability of models, and distributed computing.

3.6 Incremental Learning For Non-Stationary Data

One of the challenging aspects in Big Data Analytics is dealing with streaming and fast-moving input data. Such data analysis is useful in monitoring tasks, such as fraud detection. It is important to adapt Deep Learning to handle streaming data, as there is a need for algorithms that can deal with large amounts of continuous input data. In this section, we discuss some works associated with Deep Learning and streaming data, including incremental feature learning and extraction, denoisingauto encoders , and deep belief networks . Zhou et al. describe how a Deep Learning algorithm can be used for incremental feature learning on very large datasets, employing denoisingauto encoders .Denoisingautoencoders are a

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 11, Issue 1, January 2022 ||

| DOI:10.15680/IJIRSET.2022.1101126 |

variant of autoencoders which extract features from corrupted input, where the extracted features are robust to noisy data and good for classification purposes. Deep learning algorithms in general use hidden layers to contribute towards the extraction of features or data representations. In a denoisingautoencoder, there is one hidden layer which extracts features, with the number of nodes in this hidden layer initially being the same as the number of features that would be extracted. Incrementally, the samples that do not conform to the given objective function (for example, their classification error is more than a threshold, or their reconstruction error is high) are collected and are used for adding new nodes to the hidden layer, with these new nodes being initialized based on those samples. Subsequently, incoming new data samples are used to jointly retrain all the features.

This incremental feature learning and mapping can improve the discriminative or generative objective function; however, monotonically adding features can lead to having a lot of redundant features and overfitting of data. Consequently, similar features are merged to produce a more compact set of features. Zhou et al. demonstrate that the incremental feature learning method quickly converges to the optimal number of features in a large-scale online setting. This kind of incremental feature extraction is useful in applications where the distribution of data changes with respect to time in massive online data streams. Incremental feature learning and extraction can be generalized for other Deep Learning algorithms, such as RBM [7], and makes it possible to adapt to new incoming stream of an online large-scale data. Moreover, it avoids expensive cross-validation analysis in selecting the number of features in large-scale datasets. Calandra et al. introduce adaptive deep belief networks which demonstrates how Deep Learning can be generalized to learn from online non-stationary and streaming data. Their study exploits the generative property of deep belief networks to mimic the samples from the original data, where these samples and the new observed samples are used to learn the new deep belief network which has adapted to the newly observed data. However, a downside of an adaptive deep belief network is the requirement for constant memory consumption.

The targeted works presented in this section provide empirical support to further explore and develop novel Deep Learning algorithms and architectures for analyzing large-scale, fast moving streaming data, as is encountered in some Big Data application domains such as social media feeds, marketing and financial data feeds, web click stream data, operational logs, and metering data. For example, Amazon Kinesis is a managed service designed to handle real-time streaming of Big Data though it is not based on the Deep Learning approach.

3.7 High-dimensional data

Some Deep Learning algorithms can become prohibitively computationally-expensive when dealing with high-dimensional data, such as images, likely due to the often slow learning process associated with a deep layered hierarchy of learning data abstractions and representations from a lower-level layer to a higher-level layer. That is to say, these Deep Learning algorithms can be stymied when working with Big Data that exhibits large Volume, one of the four Vs associated with Big Data Analytics. A high-dimensional data source contributes heavily to the volume of the raw data, in addition to complicating learning from the data. Chen et al. introduce marginalized stacked denoisingautoencoders (mSDAs) which scale effectively for high-dimensional data and is computationally faster than regular stacked denoisingautoencoders (SDAs).

Their approach marginalizes noise in SDA training and thus does not require stochastic gradient descent or other optimization algorithms to learn parameters. The marginalized denoisingautoencoder layers to have hidden nodes, thus allowing a closed-form solution with substantial speed-ups. Moreover, marginalized SDA only has two free meta-parameters, controlling the amount of noise as well as the number of layers to be stacked, which greatly simplifies the model selection process. The fast training time, the capability to scale to large-scale and highdimensional data, and implementation simplicity make mSDA a promising method with appeal to a large audience in data mining and machine learning. Convolutional neural networks are another method which scales up effectively on highdimensional data.

Researchers have taken advantages of convolutional neural networks on ImageNet dataset with 256×256 RGB images to achieve state of the art results . In convolutional neural networks, the neurons in the hidden layers units do not need to be connected to all of the nodes in the previous layer, but just to the neurons that are in the same spatial area. Moreover, the resolution of the image data is also reduced when moving toward higher layers in the network. The application of Deep Learning algorithms for Big Data Analytics involving highdimensional data remains largely unexplored, and warrants development of Deep Learning based solutions that either adapt approaches similar to the ones presented above or develop novel solutions for addressing the high-dimensionality found in some

Big Data domains. Large-scale models From a computation and analytics point of view, how do we scale the recent successes of Deep Learning to much larger-scale models and massive datasets? Empirical results have demonstrated the effectiveness of large-scale mode, with particular focus on models with a very large number of model parameters which are able to extract more complicated features and representations . Dean et al. consider the problem of training a Deep Learning neural network with billions of parameters using tens of thousands of CPU cores, in the context of

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 11, Issue 1, January 2022 ||

| DOI:10.15680/IJIRSET.2022.1101126 |

speech recognition and computer vision. A software framework, DistBelief, is developed that can utilize computing clusters with thousands of machines to train large-scale models. The framework supports model parallelism both within a machine (via multithreading) and across machines (via message passing), with the details of parallelism, synchronization, and communication managed by DistBelief. In addition, the framework also supports data parallelism, where multiple replicas of a model are used to optimize a single objective. In order to make large-scale distributed training possible an asynchronous SGD as well as a distributed batch optimization procedure is developed that includes a distributed implementation of L-BFGS (Limited-memory Broyden-Fletcher-GoldfarbShanno, a quasi-Newton method for unconstrained optimization).

The primary idea is to train multiple versions of the model in parallel, each running on a different node in the network and analyzing different subsets of data. The authors report that in addition to accelerating the training of conventional sized models, their framework can also train models that are larger than could be contemplated otherwise. Moreover, while the framework focuses on training large-scale neural networks, the underlying algorithms are applicable to other gradient-based learning techniques. It should be noted, however, that the extensive computational resources utilized by DistBelief are generally unavailable to a larger audience. Coates et al. leverage the relatively inexpensive computing power of a cluster of GPU servers. More specifically, they develop their own system (using neural networks) based on Commodity Off-The-Shelf High Performance Computing (COTS HPC) technology and introduce a high-speed communication infrastructure to coordinate distributed computations. The system is able to train 1 billion parameter networks on just 3 machines in a couple of days, and it can scale to networks with over 11 billion parameters using just 16 machines and where the scalability is comparable to that of DistBelief. In comparison to the computational resources used by DistBelief, the distributed system network based on COTS HPC is more generally available to a larger audience, making it a reasonable alternative for other Deep Learning experts exploring large-scale models.

Large-scale Deep Learning models are quite suited to handle massive volumes of input associated with Big Data, and as demonstrated in the above works they are also better at learning complex data patterns from large volumes of data. Determining the optimal number of model parameters in such large-scale models and improving their computational practicality pose challenges in Deep Learning for Big Data Analytics. In addition to the problem of handling massive volumes of data, large-scale Deep Learning models for Big Data Analytics also have to contend with other Big Data problems, such as domain adaptation (see next section) and streaming data. This lends to the need for further innovations in large-scale models for Deep Learning algorithms and architectures.

Future work on deep learning in big data analytics In the prior sections, we discussed some recent applications of Deep Learning algorithms for Big Data Analytics, as well as identified some areas where Deep Learning research needs further exploration to address specific data analysis problems observed in Big Data Considering the low-maturity of Deep Learning, we note that considerable work remains to done. In this section, we discuss our insights on some remaining questions in Deep Learning research, especially on work needed for improving machine learning and the formulation of the high-level abstractions and data representations for Big Data. An important problem is whether to utilize the entire Big Data input corpus available when analyzing data with Deep Learning algorithms.

The general focus is to apply Deep Learning algorithms to train the high-level data representation patterns based on a portion of the available input corpus, and then utilize the remaining input corpus with the learnt patterns for extracting the data abstractions and representations. In the context of this problem, a question to explore is what volume of input data is generally necessary to train useful (good) data representations by Deep Learning algorithms which can then be generalized for new data in the specific Big Data application domain. Upon further exploring the above problem, we recall the Variety characteristic of Big Data Analytics, which focuses on the variation of the input data types and domains in Big Data. here, by considering the shift between the input data source (for training the representations) and the target data source (for generalizing the representations), the problem becomes one of domain adaptation for Deep Learning in Big Data Analytics.

Domain adaptation during learning is an important focus of study in Deep Learning , where the distribution of the training data (from which the representations are learnt) is different from the distribution of the test data (on which the learnt representations are deployed). Glorot et al. demonstrate that Deep Learning is able to discover intermediate data representations in a hierarchical learning manner, and that these representations are meaningful to, and can be shared among, different domains. In their work, a stacked denoisingautoencoder is initially used to learn features and patterns from unlabeled data obtained from different source domains. Subsequently, a support vector machine (SVM) algorithm utilizes the learnt features and patterns for application on labeled data from a given source domain, resulting in a linear classification model that outperforms other methods. This domain adaptation study is successfully applied on a large industrial strength dataset consisting of 22 source domains. However, it should be noted that their study does not explicitly encode the distribution shift of the data between the source domain and the target domains. Chopra et al. [propose a Deep Learning model (based on neural networks) for domain adaptation which strives to learn a useful (for prediction purposes) representation of the unsupervised data by taking into consideration information available from

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 11, Issue 1, January 2022 ||

| DOI:10.15680/IJIRSET.2022.1101126 |

the distribution shift between the training and test data. The focus is to hierarchically learn multiple intermediate representations along an interpolating path between the training and testing domains. In the context of object recognition, their study demonstrates an improvement over other methods.

The two studies presented above raise the question about how to increase the generalization capacity of Deep Learning data representations and patterns, noting that the ability to generalize learnt patterns is an important requirement in Big Data Analytics where often there is a distribution shift between the input domain and the target domain. Another key area of interest would be to explore the question of what criteria is necessary and should be defined for allowing the extracted data representations to provide useful semantic meaning to the Big Data. Earlier, we discussed some studies that utilize the data representations extracted through Deep Learning for semantic indexing. Bengio et al. [5] present some characteristics of what constitutes good data representations for performing discriminative tasks, and point to the open question regarding the definition of the criteria for learning good data representations in Deep Learning. Compared to more conventional learning algorithms where misclassification error is generally used as an important criterion for model training and learning patterns, defining a corresponding criteria for training Deep Learning algorithms with Big Data is unsuitable since most Big Data Analytics involve learning from largely unsupervised data.

While availability of supervised data in some Big Data domains can be helpful, the question of defining the criteria for obtaining good data abstractions and representations still remains largely unexplored in Big Data Analytics. Moreover, the question of defining the criteria required for extracting good data representations leads to the question of what would constitute a good data representation that is effective for semantic indexing and/or data tagging. In some Big Data domains, the input corpus consists of a mix of both labeled and unlabeled data, e.g., cyber security, fraud detection , and computer vision . In such cases, Deep Learning algorithms can incorporate semi-supervised training methods towards the goal of defining criteria for good data representation learning. For example, following learning representations and patterns from the unlabeled/unsupervised data, the available labeled/supervised data can be exploited to further tune and improve the learnt representations and patterns for a specific analytics task, including semantic indexing or discriminative modeling.

A variation of semi-supervised learning in data mining, active learning methods could also be applicable towards obtaining improved data representations where input from crowdsourcing or human experts can be used to obtain labels for some data samples which can then be used to better tune and improve the learnt data representations.

IV. CONCLUSION

In contrast to more conventional machine learning and feature engineering algorithms, Deep Learning has an advantage of potentially providing a solution to address the data analysis and learning problems found in massive volumes of input data. More specifically, it aids in automatically extracting complex data representations from large volumes of unsupervised data. This makes it a valuable tool for Big Data Analytics, which involves data analysis from very large collections of raw data that is generally unsupervised and un-categorized. The hierarchical learning and extraction of different levels of complex, data abstractions in Deep Learning provides a certain degree of simplification for Big Data Analytics tasks, especially for analyzing massive volumes of data, semantic indexing, data tagging, information retrieval, and discriminative tasks such a classification and prediction. In the context of discussing key works in the literature and providing our insights on those specific topics, this study focused on two important areas related to Deep Learning and Big Data: (1) the application of Deep Learning algorithms and architectures for Big Data Analytics, and (2) how certain characteristics and issues of Big Data Analytics pose unique challenges towards adapting Deep Learning algorithms for those problems. A targeted survey of important literature in Deep Learning research and application to different domains is presented in the paper as a means to identify how Deep Learning can be used for different purposes in Big Data Analytics. The low-maturity of the Deep Learning field warrants extensive further research. In particular, more work is necessary on how we can adapt Deep Learning algorithms for problems associated with Big Data, including high dimensionality, streaming data analysis, scalability of Deep Learning models, improved formulation of data abstractions, distributed computing, semantic indexing, data tagging, information retrieval, criteria for extracting good data representations, and domain adaptation. Future works should focus on addressing one or more of these problems often seen in Big Data Analytics.

REFERENCES

[1]. Domingos P (2012) A few useful things to know about machine learning. Commun ACM 55(10)

[2]. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference On. IEEE Vol. 1. pp 886–893 [3]. Lowe DG (1999) Object recognition from local scale-invariant features. In: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference On. IEEE Computer Society Vol. 2. pp 1150–1157

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 11, Issue 1, January 2022 ||

| DOI:10.15680/IJIRSET.2022.1101126 |

[4]. Bengio Y, LeCun Y (2007) Scaling learning algorithms towards, AI. In: Bottou L, Chapelle O, DeCoste D, Weston J (eds). Large Scale Kernel Machines. MIT Press, Cambridge, MA Vol. 34. pp 321–360. http://www.iro.umontreal.ca/~lisa/pointeurs/bengio+lecun_chapter2007.

[5]. Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(8):1798–1828. doi:10.1109/TPAMI.2013.50

[6]. Arel I, Rose DC, Karnowski TP (2010) Deep machine learning-a new frontier in artificial intelligence research [research frontier]. IEEE ComputIntell 5:13–18

[7]. Hinton GE, Osindero S, Teh Y-W (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554

[8]. Bengio Y, Lamblin P, Popovici D, Larochelle H (2007) Greedy layer-wise training of deep networks, Vol. 19

[9]. Larochelle H, Bengio Y, Louradour J, Lamblin P (2009) Exploring strategies for training deep neural networks. J Mach Learn Res 10:1–40 [10]. Salakhutdinov R, Hinton GE (2009) Deep boltzmannmachines. In: International Conference on, Artificial Intelligence and Statistics. JMLR.org. pp 448–455

[11]. Deep Learning and Its Applications: A Review October 2016 Conference: Postgraduate Annual Research on InformaticsSeminar 2016 At: UniversitiTeknologi Malaysia, Kuala Lumpur.

[12] National Research Council (2013) Frontiers in Massive Data Analysis. The National Academies Press, Washington, DC. http://www.nap.edu/openbook.php?record_id=1837



INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details