DEEP LEARNING CHALLENGES IN BIG DATA ANALYTICS

MALATHI PULIGADDA & **HARI MANASA CHAPALA** Lecturer, Department of Computer Science, Maris Stella College, Autonomous, Vijayawada. A.P., India.

Abstract: Deep learning is a class of machine learning algorithms that uses multiple layers to progressively extract higher-level features from the raw input. Traditional data processing techniques have several limitations of processing large amount of data. Big data is a great quantity of diverse information that arrives in increasing volumes and with ever-higher velocity. One technique that can be used for data analysis so that able to help us find abstract patterns in Big Data is Deep Learning. Big Data analytics requires new and sophisticated algorithms based on machine and deep learning techniques to process data in real-time with high accuracy and efficiency. A main benefit of Deep Learning is the analysis and learning of massive amounts of unsupervised data, making it a valuable tool for Big Data Analytics. In this article, 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.

INTRODUCTION

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.

LITERATURE SURVEY

[1]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 and make it possible to adapt to new incoming stream of an online large-scale data. Moreover, it avoids expensive cross-validatio nanalysis in selecting the number of features in large-scale datasets.

[2] 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 aroused 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 practical 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] Chorea al. propose 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 the distribution shift between the training and test data. The focus is to hierarchically learn multiple intermediate representations.

BIGDATA ANALYTICS

Big Data generally refers to data that exceeds the typical storage, processing, and com-putting 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 that, Big Data is also associated with other specific complexities, often referred to as the four **V**s: Volume, Variety, Velocity, and Veracity.

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 Face book, Yahoo, Google, already have large amounts of data and have recently begun tapping into its benefits. 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 given repository poses unique challenges with Big Data, which requires Big Data pre-processing 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 increase in grate 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. 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. some key problems areas include: data quality and validation, data cleansing, feature engineering, high-dimensionality and data reduction,datarepresentationsanddistributeddatasources,datasampling,scalabilityofalgorithms,data visualization, parallel and distributed data processing, real-time analysis and decision making, crowd sourcing 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

heterogeneous data, and developing new models for massive data computation.

DEEP LEARNING CHALLENGES IN BIG DATA ANALYTICS

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.

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, demising auto encoders, and deep belief networks.

FUTURE WORK ON DEEP LEARNING IN BIG DATA ANALYTICS

In the prior sections, we discussed some areas where Deep Learning research needs further exploration to address pacific data analysis problems observed n 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 and the target data source, the problem becomes one of domain adaptation for Deep Learning in Big Data Analytics.. However, it

should be noted thattheirstudy does not explicitly encode the distributionshiftof the data between thesourcedomainand the target domains.

Inthecontextofobjectrecognition,theirstudy demonstrates an improvement over other methods. The two studies presented aboveraise 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. Bengioet al. present some characteristics of what constitutes good data representations forperforming discriminative tasks, and point to the open question regarding the definitionof the criteria for learning good data representations in Deep Learning. Compared tomore conventional learning algorithms where misclassification error is generally used asan important criterion for model training and learning patterns, defining a corresponding criteria for training Deep Learning algorithms with Big Data is unsuitable since mostBig 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 inBig Data Analytics. Moreover,thequestionof defining thecriteriarequiredfor extracting good data representations leads to the question of what would constitute agood 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 [59], fraud detection, and computer vision.

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.

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 andprediction. In the context of discussing key works in the literature and providing our insights on those specific topics, this study focused on to Deep Learning and Big Data:

- 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, thus contributing to the Deep Learning and Big Data Analytics research corpus.

References

1. Domingos P (2012) A few useful things to know about machine learning. Commun ACM 55(10)
2. Lowe DG (1999) Object recognition from local scale-invariant features. In: Computer Vision, 1999. The Proceedingsof the Seventh IEEE International Conference On. IEEE Computer Society Vol. 2. pp 1150–1157.
3. Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. Pattern Analysis andMachine Intelligence, IEEE Transactions on 35(8):1798–1828. doi:10.1109/TPAMI.2013.50.
4. Salakhutdinov R, Hinton GE (2009) Deep boltzmann machines. In: International Conference on, Artificial Intelligenceand Statistics. JMLR.org.pp 448–455.
5. National Research Council (2013) Frontiers in Massive Data Analysis. The National AcademiesPress,Washington,DC.http://www.nap.edu/openbook.php?record_id=18744.
6. Bengio Y (2013) Deep learning of representations: Looking forward. In: Proceedings of the 1st

InternationalConference on Statistical Language and Speech Processing. SLSP'13. Springer, Tarragona, Spain. pp 1–37.

7. Garshol LM (2013) Introduction to Big Data/Machine Learning. Online Slide Show, http://www.slideshare.net/larsga/ introduction-to-big-datamachine-learning. http://www.slideshare.net/larsga/introducti on-to-big-datamachine-learning.
8. Chopra S, Balakrishnan S, Gopalan R (2013) Dlid: Deep learning for domain adaptation by interpolating betweendomains. In: Workshop on Challenges in Representation Learning, Proceedings of the 30th International Conferenceon Machine Learning, Atlanta, G.
9. Calandra R, Raiko T, Deisenroth MP, Pouzols FM (2012) Learning deep belief networks from non-stationary streams.In: Artificial Neural Networks and Machine Learning–ICANN 2012. Springer, Berlin Heidelberg. pp 379–386.
10. Zhou G, Sohn K, Lee H (2012) Online incremental feature learning with denoising autoencoders. In: InternationalConference on Artificial Intelligence and Statistics. JMLR.org. pp 1453–1461.