International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 3 | ISSN : 2456-3307

Impact of Deep Learning in Big Data Analytics

A. G. Aruna[1], Dr. M. Sangeetha[2], C. Sathya[3], K.H.Vani[4]

[1,3]Department of Computing, Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India
[2]Department of Computer ScienceEngineering and Information Technology, Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India

ABSTRACT

New technologies enable us to collect more data than ever before. With an overwhelming amount of web-based, mobile, and sensor-generated data arriving at a terabyte and even zeta byte scale, new science and insights can be discovered from the highly detailed and domain-specific information which can contain useful information about problems such as national intelligence, cyber security, fraud detection, financial trading, personalized medicine and treatments, personalized information and recommendations and personalized athletic training. Machine learning algorithms, particularly deep learning (evolved from artificial neural networks) plays a vital role in big data analysis. Deep Learning algorithms extracts high-level and complex abstractions by discovering intricate structure in large data sets. Deep learning techniques are nowadays the leading approaches to solve complex machine learning and pattern recognition problems such as speech and image understanding, semantic indexing, data tagging and fast information retrieval. This paper focuses on all aspects of big data analytics, with a particular emphasis on the analysis and learning of massive volume of unstructured data and developing effective and efficient large-scale learning algorithms.

Keywords : Deep Learning, Bigdata, BigdataAnalytics

I. INTRODUCTION

Public and private organizations are collecting enormous amounts of domain-specific information only to use it for solving problems in marketing, technology, medical science, national intelligence, fraud detection, and what not. While such data is crucial to the organization acquiring it, it is also unlabeled, uncategorized and immensely complex to handle and analyze.

Fortunately, deep learning algorithms specialize in analyzing such large volumes of unsupervised data. Not just this, deep learning algorithms continuously improvise with each set of data they tackle, making deep learning tools the most suitable ones for big data analytics.

II. WHERE IS DEEP LEARNING APPLICABLE IN BIG DATA ANALYTICS?

Volume, Variety, Velocity and Veracity. These 4Vs sum up the game of big data. Deep learning is adepartment at exploiting gigantic amounts of data, thus capable of addressing the *volume* factor. It is also well suited for analyzing raw data from different sources, and in different formats. Deep learning can, thus, offer unique solutions to complex problems plaguing big data analytics, as follows.

A. Semantic indexing

Social media, shopping systems, cyber traffic monitoring, security systems, etc. produce information in the form of text, video, audio, and image. Not only are these high volumes of information, but they also have different representations, typical of big data. Such data, therefore, can't be stored as data bit strings. Deep learning enables efficient storage and retrieval of such data. Instead of using the raw input for data indexing, it uses high level abstract data representations for semantic indexing. This feature of deep learning can, for example, make search engines work quicker and more efficiently. Semantic indexing presents the data in

a manner that makes it useful as a source for knowledge discovery and understanding.

B. Performing discriminative tasks on big data

More often than not, the purpose of big data analysis is to discriminate between faces in images, voices in audios, writings in documents, etc. so as to increase their accessibility in a quicker and more efficient fashion.

Deep learning applies its complex algorithm to big data and extracts nonlinear features from it. It then enables simple linear analytical models to be applied on these extracted features. By way of nonlinearity, deep learning makes this task come close to artificial intelligence.

This way, data analysts benefit from the vast reserves of knowledge in the pool of big data. On the other hand, by enabling the application of simple linear analytics, deep learning offers computational efficiency.

C. Semantic tagging

With the Internet exploding with online users, digital content has been on an exponential rise. This is especially true of images and videos uploaded from multiple sources. When talking of such massive repositories of images, you cannot afford to stick with textual relationships of images, for storage and retrieval. For improved image searches, the process of browsing and retrieval should be lightening quick and broad based. This needs an automated system of tagging images and videos. Deep learning prepares complicated representations for image/video data in the form of high level abstractions. These can then be used for image tagging that is more suitable for huge data.

D. Object Recognition

Computer Vision is the art of making useful decisions for the real physical objects and scenes based on images . Object recognition, 3D-modeling, medical imaging, and smart cars are all examples of what current computer vision systems can do. A fundamental challenge of large scale object recognition is how to attain proficiency in both feature extraction and classifier training without conceding performance. It is found that feature detection by using deep networks is more powerful in performing object recognition tasks . Nair and Hinton presented a third-order Boltzmann Machine (BM) as a new type of top-level Deep Belief. Network (DBN) model for 3D objects recognition tasks. A hybrid training algorithm is used which incorporates both generative and discriminative gradients. Generative training makes more accurate object recognition and extracts more abstract image representation and discriminative training provides better classification accuracy.

This model is applied to NORB database (normalized-uniform version), which holds stereo-pair images of objects in dissimilar lighting conditions and viewpoints. The error rate reached to 6.5%, which is less than other state-of-the-art error rates. So, they proved that DBNs extraordinarily outperforms shallow models, such as Supper Vector Machines (SVM). However, third-order BM needed to be more factorized with the purpose of making the top-level features can be shared across classes.

The problem of making image classification for large variance datasets with the existence of only limited labeled data. A Discriminative DBN (DDBN) is presented as a novel semi-supervised learning algorithm to solve this problem, which is built by using a set of RBMs. In the learning phase, the greedy layer-wise unsupervised learning algorithm is applied to the network using the limited labeled data with plentyunlabeled data. In fine tuning phase, gradient descent based supervised learning algorithm is applied to the whole network by using an exponential loss function for maximizing the existence of the labeled data. The performance of DDBN is demonstrated on MNIST and Caltech 101 standard artificial datasets. Results showed that DDBN achieves less error rates compared with typical classifiers.

Krizhevsky trained one of the largest Deep Convolutional Neural Networks (DCNN) to classify ImageNet LSVRC-2010 contest which comprises 1.2 million high-resolution images belonging to 1000 different image classes. This large DCNN consists of 650,000 neurons with 60 million parameters and eight layers. Five of layers are convolutional which may be followed by max-pooling layers and the remaining three are fully connected with a final 1000-way softmax. To speed up the training process a rectified linear units with a very efficient GPU implementation are used .After pre-training, ‘dropout’, regularization method is applied to preventover-fitting in the fully-connected layers. On the test set, results showed that the error rates of the large DCNN model significantly lower than the previous state-of-the-art .But the

network's performance is directly proportional with number of convolutional layer, thus led to complex computations.

A Deep Visual-Semantic Embedding model (DeViSE) model for overcoming the weaknesses of modern visual recognition systems that can be summarized in the difficulty in dealing with large scale images with only limited training data. (DeViSE) is trained by asynchronous stochastic gradient descent algorithm and worked with not only the labeled images but also with a relatively independent and large dataset of semantic information from un-annotated text data. So, the semantic relationships between labels can be learned easily and images can be mapped obviously into a rich semantic embedding space with fewer limitations. This model is applied to the 1000-class ImageNet dataset and results showed that the semantic information aided in making better predictions about tens of thousands of image labels that not observed during training.

E. Social targeting

Deep learning holds the potential to guess the unstated emotions and events in a text. It can identify objects in photos. It can also make knowledgeable predictions about people's likely future behaviour. All these features make it a hot property in the fields of intelligence, sales, marketing, and advertising. No wonder then that Facebook has set up an internal team to reap the benefits of deep learning in their work.

Does Deep Learning apply to my business?

- If your business generates or consumes high volumes of variable data
- If time is money for you
- If you look for results that suggest next steps
- If you are ambitious and wish to stay ahead of your competitors
- If you can't afford stagnancy and redundancy
- If you believe in the power of Technology.

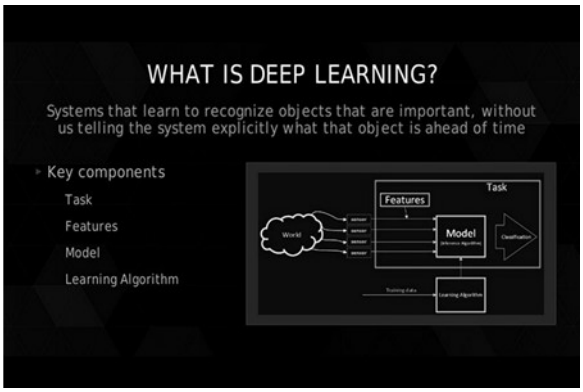

Figure 1. Deep Learning

III. SPECIAL ISSUE ON DEEP LEARNING FOR INTELLIGENT BIG DATA MANAGEMENT

Multiple processing levels, at several stages of abstraction, is central to the deep learning architecture. This architecture is loosely inspired by the hierarchically structured, massively interconnected neocortex. Recent successes in computer vision provide a paradigmatic example of the utility of deep learning; great strides have been made in computer vision using deep convolution neural nets (DCNNs). These networks emulate the basic structure of visual cortex by tiling the visual field with filters and arranging them in successive interconnected processing levels. Although only the initial layer of the DCNN is modeled to loosely map on to response properties in primary visual cortex (the first cortical processing level), it has recently been shown that subsequent layers of a well-trained DCNN also show a functional correspondence to successive levels of the human visual processing hierarchy. In other words, despite only a loose correspondence in their architecture, both biological and artificial visual systems evolve layers with similar response properties, suggesting that such an architecture provides a fundamentally advantageous approach to information processing.

Meanwhile, the management of huge amount of complex data is becoming a serious hurdle that needs to be dealt with. Unfortunately, because of the dynamicity of these data and our need to respond in real-time situations, traditional data processing formalisms are inadequate to solve this problem. Some challenges include data exploration, capture, storage, search, sharing, transfer, visualization, querying, updating, predicting future trends, cluster analysis, as well as information privacy.

Recent developments in the field of deep machine learning (DML) offers powerful tools to an *intelligent big data management*. We believe that a cognitive formalism such as deep learning architecture that combines artificial intelligence and machine learning will leapfrog our current perception of information processing and management.

IV. CHALLENGES IN BIG DATA ANALYTICS

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 anddistributed 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.

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.

V. CONCLUSION

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

VI. REFERENCES

[1]. Nagwa M. Elaraby, Mohammed Elmogy, Shereif Barakat,"Deep Learning: Effective Tool for Big Data Analytics", International Journal of Computer Science Engineering (IJCSE), Vol. 5 No.05 Sep 2016.

[2]. M. Najafabadi, F. Villanustre, T. Khoshgoftaar, N, Seliya, R. Wald, and E. Muharemagic, “Deep Learning applications and challenges inBig Data analytics". Journal of Big Data, vol.2, doi: 10.1186/s40537-014-0007-7, 2015.

[3]. X. Chen, and X. Lin, “Big Data Deep Learning: Challenges and Perspectives". In Access, IEEE, vol.2, pp.514, 525, doi: 10.1109/ACCESS.2014.2325029, 2014.

[4]. National Security Agency. The National Security Agency: Missions, Authorities, Oversight and Partnerships Online]. Available: http:// www.nsa.gov/public_info/_files/speeches_testimonies/2013_08_09 _the_nsa_story.pdf

[5]. J. Gantz and D. Reinsel, Extracting Value from Chaos. Hopkinton, MA, USA: EMC, Jun. 2011.

[6]. J. Gantz and D. Reinsel, The Digital Universe Decade-Are You Ready. Hopkinton, MA, USA: EMC, May 2010.

[7]. (2011, May). Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute Online]. Available: http://www.mckinsey.com/insights/business_technology/big_data_the_ next_frontier_for_innovation

[8]. J. Lin and A. Kolcz, "Large-scale machine learning at twitter," in Proc. ACM SIGMOD, Scottsdale, Arizona, USA, 2012, pp. 793-804.

[9]. A. Smola and S. Narayanamurthy, "An architecture for parallel topic models," Proc. VLDB Endowment, vol. 3, no. 1, pp. 703-710, 2010.