

A survey on deep learning in big data analytics

Makrufa Hajirahimova¹, Aybeniz Aliyeva²

Institute of Information Technology – Azerbaijan National Academy of Sciences, Baku, Azerbaijan
hmakrufa@gmail.com¹, aliyeva.a.s@mail.ru²

Abstract: Over the last few years, Deep learning has begun to play an important role in analytics solutions of big data. Deep learning is one of the most active research fields in machine learning community. It has gained unprecedented achievements in fields such as computer vision, natural language processing and speech recognition. The ability of deep learning to extract high-level complex abstractions and data examples, especially unsupervised data from large volume data, makes it attractive a valuable tool for big data analytics. In this paper, we review the deep learning architectures which can be used for big data processing. Next, we focus on the analysis and discussions about the challenges and possible solutions of deep learning for big data analytics. Finally, have been outlined several open issues and research trends.

Keywords: BIG DATA, BIG DATA ANALYTICS, DEEP LEARNING, AUTO-ENCODERS, DEEP BELIEF NETWORKS, CONVOLUTIONAL NEURAL NETWORKS, RECURRENT NEURAL NETWORKS, RESTRICTED BOLTZMANN MACHINES

1. Introduction

Recently, Deep learning and Big data analytics are the very active fields of research in the science and engineering domains. Big data is defined as digital data, which is difficult or impossible to manage and analyze with traditional software tools and technologies [1]. Analyzing of data and obtaining knowledge and useful information from them is very important for making motivated decisions in organizations, new scientific revelations, national security and healthcare fields. The demand for data analysis in real-time has led to the creation of Big data analytics. Big data analytics is a process of extracting useful information from large volumes of data to make optimal (best) decisions. The size of data has considerably grown in the last decade, with the emergence of social networks, Internet of Things, cloud computing and other technologies. The rapid increasing of data volume, along with the promises potential opportunities for all sectors of society, creates problems for data mining and information processing [2]. Dealing with these data can be supported by Deep learning capabilities, especially its ability to deal with both the labeled and unlabeled data which are often collected abundantly in Big data. Deep learning is an attractive research topics that belong in Artificial Intelligence (AI). DL refers to machine learning techniques that based on supervised and unsupervised methods to automatically learn hierarchical representations in deep architectures. It has achieved unprecedented success in applications of essential fields such as computer vision, speech and audio processing, and natural language processing [3-7].

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 analytics [4-6]. More specifically, Big data analytics problems such as semantic indexing, data tagging, fast information retrieval, and discriminative modeling can be better addressed with the aid of Deep Learning. In addition, there are need to use of Deep learning methods in solving of different problems that faced Big data analytics such as fast moving streaming data, highly distributed input sources, noisy and poor quality data, high dimensionality, scalability of algorithms, unsupervised and un-categorized data, limited supervised / labeled data and format variations of raw data.

The structure of the paper is organized as follows: Section 2 presents a brief review of typical deep learning models which are the most widely used for big data analysis and feature learning. Section 3 introduces possible solutions of deep learning for big data analytics challenges. Section 3 gives some open issues and research trends; and the final section is conclusions.

2. Brief review of deep learning architectures

Deep learning refers to a set of machine learning techniques that learn multiple levels of representations in deep architectures. In the last few years, various deep learning architectures have been developed. A brief overview of the deep learning architectures has been looked throw that are commonly used in Big Data analytics below.

2.1. Autoencoder and Stacked Autoencoders (SAEs)

As one of the most widely used deep learning techniques, Stacked autoencoders (SAEs) are constructed by stacking several autoencoders that are the most typical feed-forward neural networks [7]. Autoencoder is a kind of unsupervised learning structure that owns three layers: input layer, hidden layer, and output layer (Fig 1.). The process of an autoencoder training consists of two stages, i.e., encoding stage and decoding stage. Encoder is used for mapping the input data into hidden representation, and decoder is referred to reconstructing input data from the hidden representation.

SAE is typically trained by two stages, i.e., pre-training and fine-tuning. In the pre-training stage, each auto-encoder model is trained in a unsupervised layer-wise manner from bottom to top. This operation is repeated until the parameters of all the hidden layers are trained. After all the hidden layers are trained, backpropagation algorithm is used to minimize the cost function and update the weights with labeled training set to achieve fine-tuning [7, 8].

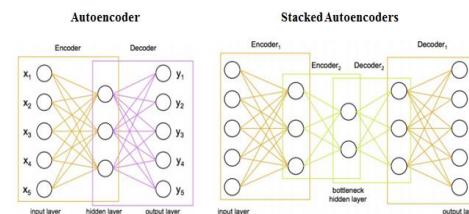


Fig. 1 Architecture of autoencoders.

2.2. Restricted Boltzmann Machines (RBMs) and Deep Belief Network (DBN)

Deep belief network is the commonly used, and successfully trained architectures in deep learning [9]. DBN is stacked by several restricted Boltzmann machines, as presented in Figure 3. RBMs are the most popular version of Boltzmann machine [5,7]. The Restricted Boltzmann Machine (RBM) is a probabilistic graphical model or a type of stochastic neural network. The network consists of two layers, i.e., visible layer and hidden layer (Fig. 2). The restriction is that there is no interaction between the units of the same layer and the connections are solely between units from different layers.

Deep belief networks have the potential to learn the representation of features using structured and unstructured data. It consists of input, hidden, and output layer. RBM uses DBN to construct a model that consists of two layers that are fully connected to each other. DBN combined strategies of unsupervised pre-training and supervised fine-tuning. The unsupervised stages intend to learn data distributions without using label information and supervised stages perform local search for fine tuning[7]. In the literature, DBN model is used by many researchers to efficiently and accurately process big data.

In particularly, a graphical processing unit (GPU)-based model using stacked RBM in parallel to handle large volume of data with minimized process time. The power of deep learning is that it can train and handle millions of parameters at a time. Several restricted Boltzmann machines can be stacked into a deep belief network.

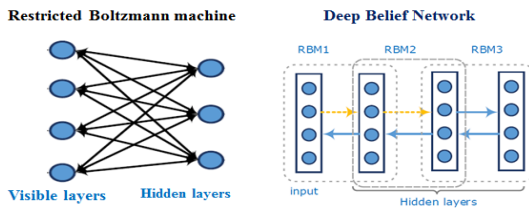


Fig 2. Deep belief network architecture.

2.3. Convolutional Neural Networks (CNNs)

The convolutional neural network (CNN) is a multilayer, feed-forward neural network that uses perceptrons for supervised learning and to analyze data. It is used mainly with visual data, such as image classification. CNN architecture is different from other neural networks. The hidden layers in CNN contain convolutional layer, subsampling layer (pooling layer) and a fully connected layer (Fig. 3). Characteristically, CNN start with convolutional layer that accepts data from input layer. The convolutional layer is responsible for convolution operations having few filter maps of same size. The convolutional layer uses the convolution operation to achieve the weight sharing while the subsampling is used to reduce the dimension [6, 7].

Following the convolutional layer, a subsampling (or pooling) layer is usually used to reduce the dimension of the feature map. It can typically be realized by an average pooling operation or a max pooling operation. After the second stage, CNN uses a fully connected layer and then a softmax layer with output classes for classification and recognition.

During recent years, CNN has achieved great success in many applications such as image analysis, speech recognition, text understanding and so on [7].

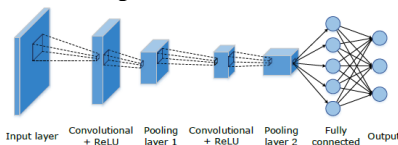


Fig. 3 Architecture of a CNN.

2.4. Recurrent Neural Networks (RNNs)

Recurrent neural network considered as another class of deep networks for unsupervised / supervised learning that is very powerful for modeling sequence data (e.g., speech or text). RNN learns features for the series data by a memory of previous inputs that are stored in the internal state of the neural network. The connections between neurons is constructed with a directed cycle (Fig. 4.).

Unlike traditional networks, where inputs and outputs are independent of each other, the recurrent neural network captures the dependency between the current sample with the previous one by integrating the previous hidden representation into the forward pass. From a theoretical point of view, the recurrent neural network can capture arbitrary-length dependencies. However, it is difficult for the recurrent neural network to capture a long-term dependency because of the gradient vanishing with the back-propagation strategy for training the parameters. To tackle this problem, some models, such as Long Short-Term Memory, have been presented by preventing the gradient vanishing or gradient exploding [7].

The recurrent neural network and its variants have achieved super performance in many applications such as natural language processing, speech recognition and machine translation.

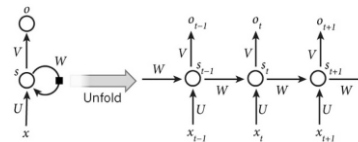


Fig. 4 Architecture of a RNN

3. Deep learning in Big Data analytics

The big data application process generally includes stages such as data generation, data management, data analytics, and data application. Big data analytics, which is considered the most important phase in the whole chain, refers to the process of discovering patterns from data. In this stage, there are several challenges (such as high dimensionality, scalability of algorithms, fast moving streaming data, noisy and poor quality data and so on), which is made big data analytics much more difficult and complicated than normal-sized data analytics [10].

In this section have been provided the analysis and discussions about the challenges and possible solutions of deep learning for big data analytics.

3.1. Complex data representation

Big Data is usually collected from different domains which consists of multiple modalities. Each modality has a different representation, distribution, scale, and density. For example, text is usually represented a discrete word-count vectors, but an image is represented by real values of pixel intensities [11]. The using of existing methodologies for the processing of such data is almost impossible. The solution of this problem is possible owing to the integration of heterogeneous data.

Deep Learning is more fitting for heterogeneous data integration due to its potentiality of learning variation factors of data and providing abstract representations for it. Deep learning has been demonstrated to be very effective in integrating data from different sources [3]. Some multi-model deep learning models have been proposed for heterogeneous data integration.

For example, Ngiam et al. [12] developed a multi-modal deep learning model to learn representations by integrating audio and video data. Srivastava and Salakhutdinov [13] developed a multimodal Deep Boltzmann Machine (DBM), for text data and image objects feature learning.

Ouyang et al. [14] presented multi-modal deep learning model, called multi-source deep learning model aims to learn non-linear representation from different information sources. In this model each source of information is used as input data for the two hidden layers deep learning model. Extracting features separately are then combined for joint representation.

Generally, though the architecture of the proposed multi-modal deep learning models is different, their ideas are similar. In particular, multi-modal deep learning models firstly learn features for single modality. Then learned features are combined as the joint representation for each multi-modal object. These models have been achieved more superior productivity than traditional deep neural networks for heterogeneous data feature learning. However, these models combine the learned features of each modality in a linear way. So they are far away effective to capture the complex correlations over different modalities for heterogeneous data. In order to eliminate this problem, Zhang et al. [15] presented a tensor deep learning model, called deep computation model, for heterogeneous data.

3.2. Super-high dimensionality

Big data in specific domains is often super-high dimensional. Generally, with the increase of the data dimension, the required amounts of time or memory go up exponentially. The problem is that existing machine learning and data mining algorithms are not well scalable to high-dimensional data (such as, images), or are not computationally efficient.

Chen et al. [16] developed marginalized stacked denoising autoencoders (or mSDAs) which scale effectively for high-dimensional data and is computationally faster than regular stacked denoising autoencoders (SDAs). This approach marginalizes noise in SDA training and therefore does not require other optimization algorithms to learn parameters.

Zhang et al. [17] proposed a new tensor-based representation algorithm for image classification. The algorithm is realized by learning the parameter tensor for image tensors which the algorithm preserved the spatial information of image.

Convolutional neural networks also can scale up effectively to high-dimensional data. On ImageNet dataset with 256×256 RGB images, CNNs produced state-of-the-art results [4]. For instance, Krizhevsky et al. [18] 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. It is one of the most well-known CNN architectures for classification. This large DCNN consists of 650,000 neurons with 60 million parameters and eight layers.

Maggiori et al. [19]. proposed an end-to-end framework for the dense, pixel-wise classification of satellite imagery with convolutional neural networks.

The above Deep Learning algorithms for Big Data Analytics involving high dimensional data are not sufficient, and requires new methods for better performance of DL techniques to handle high-dimensional data.

3.3. Unscalable computation ability

A big dataset often includes a large number of attributes and many class types of samples, so some frequently used data mining and machine learning algorithms, is not work well. In order to learn features and representations for large amounts of data, some large-scale deep learning models have been developed. They can be grouped into three categories, such as *parallel deep learning models, GPU-based implementation, and optimized deep learning models* [7].

Existing deep learning systems commonly use data or model parallelism, but unfortunately, these strategies often result in suboptimal parallelization performance. Z. Jia et al. [20] proposed FlexFlow, a deep learning system that automatically finds efficient parallelization strategies for DNN applications. Authors evaluate FlexFlow with six real-world DNN benchmarks on two GPU clusters and show FlexFlow significantly outperforms state-of-the-art parallelization approaches.

Dean et al. [21] determined the possibility of training a deep network with billions of parameters using tens of thousands of CPU cores. Authors have developed a software framework called DistBelief that can utilize computing clusters with thousands of machines to train large models. DistBelief needs 16 thousand CPU cores to train a large deep learning model with 10 million images and billion parameters.

Sun et al. [22] presented techniques to accelerate distributed training of DNN on GPU clusters. They used two clusters: a cluster with 16 machines, each having 8 Pascal GPUs and a cluster with 64 machines, each having 8 Volta GPUs.

Coates et al. [23] deployed a less expensive cluster of (GPU) servers and also Commodity OFF-The-Shelf (COTS) HPC technology with a high-speed communication network to coordinate distributed computations. This system is capable to training for 1 billion parameters networks on just 3 machines in a few days and is capable scaling up to 11 billion parameters with 16 machines. Therefore, this system is affordable for everyone who wishes to explore large scale systems.

Novikov et al. [24] proposed a tensorizing learning model based on the tensor-train network. Authors converted the neural network to the tensor format to use the tensor-train network to compress the parameters. This method could reduce the computational complexity and improve the training efficiency in the back-propagation procedure.

There is a need to develop new algorithms for scalable deep learning which make it suitable for high dimensional data processing and analysis.

3.4. Fast moving streaming data

One of the challenging aspects in Big Data Analytics is dealing with streaming and fast-moving input data. The data stream is generated at an extremely fast speed, and its distribution characteristics are in high-speed dynamic changes, which must be processed in real time. Deep learning to handle streaming data, as there is a need for algorithms that can deal with large amounts of continuous input data. In recent years, a lot of incremental learning methods have been presented for high-velocity data feature learning.

Zhou et al. [25] proposed an incremental feature learning algorithm to determine the optimal model complexity for large-scale datasets based on the denoising autoencoder. The model quickly converges to the optimal number of features in a large-scale online setting. In addition, the algorithm is effective in recognizing new patterns when the data distribution changes over time in the massive online data stream. Calandra et al. [26] demonstrated Adaptive Deep Belief Network to learn from online, nonstationary stream data.

Y. Li and et al. [27] proposed an incremental high-order deep learning model based on parameter updating and structure updating to meet the requirements of dynamic big data online analysis and real-time processing. The model has the ability to incrementally learn the characteristics of new data online, also retains the ability to learn the original data features, and real-time processing of dynamic data streams.

3.5. Noisy and poor-quality data

There are a huge number of noisy objects, incomplete objects, inaccurate objects and imprecise objects in Big data. This low-quality data is widespread in Big data. For example, there are over 90% missing attribute values for a doctor diagnosis in clinic and health fields. Some traditional learning algorithms have obviously not been valid for processing the data with 90% missing values. In the past few years, some methods have been proposed to learn features for poor-quality data.

Wang and Tao presented a non-local auto-encoder model to learn reliable features for corrupted data [28]. The model achieved high performance in image denoising and restoration. Mao et al. [29] proposed a very deep fully convolutional auto-encoder network for image restoration. Since this method is based on convolutional operations, its main limitation is the local nature of the extracted features.

In [30], a deep convolutional neural network has been proposed for image denoising, where residual learning is adopted to separating noise from noisy observation.

Recent methods based on CNNs can only operate local similarities and they are incapable to capture non-local similar to itself patterns, which have been highly successful in model-based methods. In order to exploit both local and non-local similarities, in [31], has been proposed a graph-convolutional neural network, to perform image denoising. This method provides the best visual quality, recovering finer details and producing fewer artifacts.

4. Results and discussion

Researches show that significant progress has been obtained in the application field of deep learning algorithms in Big data analytics. DL sufficiently simplifies solution of Big data analytics problems as analysis of large data volumes, semantic indexing, data tagging, information retrieval, classification and prediction. At the same time, deep learning has achieved limited progress in the field of stream data and low-quality data processing, model scaling, distributed computing, and high-scale data processing. Below have been outline several open issues and research trends.

1) Continuous increasing of volume of big data makes it necessary to create more large-scale deep learning models. Such large-scale deep learning models that can be trained for Big Data

may no longer be effectively trained, depending on the available techniques and computing power. It is important to create new learning structures and computing infrastructures in the future to solve this problem.

2) Modern multi-modal deep learning models simply combine in a linear form the learned features of each modality. This often does not lead to the necessary results. There is need to investigate the effective fusion ways of learned features to improve the productivity of multi-modal deep learning models. At the same time, deep computational models have a large number of parameters that caused their high computational complexity. There is a need to researches in the field of reducing the computational complexity of deep computational models.

3) Most of the integrated learning algorithms that based on updates of parameters or structure are effective only for a hidden, layer, traditional learning models. There is a need to research of the application possibilities of integrated learning algorithms to deep learning models and deep architectures.

4) It is important to investigate reliable deep learning models for low-quality data in the near future, due to the rapid growth of low-quality data.

5) There is a need to develop new parallel and distributed algorithms/frameworks for scalable deep learning models.

5. Conclusion

In this paper has been investigated how deep learning algorithms and architectures are used to solve Big Data analytics problems. An overview of significant literature according to the application of Deep Learning in different domains showed that Deep Learning has the potential opportunities to the solving of many analytics and learning challenges faced by Big Data analytics unlike traditional machine learning methods. But while Big Data offers enough training objects for deep learning, it creates problems for large scale, heterogeneity, noisy labels, and non-stationary distribution, among many others. In order to realize the full potential of Big Data, we need to address these technical challenges with new ways of thinking and transformative solutions. For this reason, there is need for extensive investigates in the field of deep learning the future.

References

1. Alguliyev, R. "Big Data" phenomenon: Challenges and Opportunities. - Problems of Information Technology, vol. 10, no. 2, 2014, pp. 3-16. (Alguliyev R., M. Hajirahimova)
2. Alguliyev, R. Current scientific and theoretical problems of big data. - Problems of Information Society, vol. 10, no. 2, 2016, pp. 34-45. (Alguliyev R., M. Hajirahimova. A. Aliyeva)
3. Chen, W. Big Data Deep Learning: Challenges and Perspectives. - Access IEEE, vol. 2, 2014, pp. 514-525, (Chen W., X. Lin)
4. Najafabadi, M. Deep Learning applications and challenges in Big Data analytics. - Journal of Big Data, vol.2, no.1, 2015, pp.2-21. (Najafabadi M., F. Villanustre, T. Khoshgoftaar, N. Seliya.)
5. N. M. Elaraby, M. Elmogy, and Sh. Barakat, "Deep Learning: Effective Tool for Big Data Analytics." International Journal of Computer Science Engineering, vol.5, no.5, pp. 254-262, 2016.
6. Jan, B. Deep learning in big data Analytics: A comparative studym. - Computers and Electrical Engineering, vol.7, no. 24, 2017, pp. 1-13. (Jan B., H. Farman, M. Khan, M. Imran)
7. Zhang, Q. A survey on deep learning for big data. - Information Fusion, vol. 42, 2018. pp. 146-157. (Zhang Q., L. Yang, Z. Chen)
8. Liu G. A Stacked Autoencoder-Based Deep Neural Network for Achieving Gearbox Fault Diagnosis. - Mathematical Problems in Engineering, 2018, no. 5, pp. 1-10. 2018. (Liu G., H. Bao, B. Han)
9. Hinton, G. A fast learning algorithm for deep belief nets. - Neural computation, vol. 18, no. 7, pp. 1527-1554, 2006. (Hinton G., S. Osindero, Y.-W. Teh)
10. Wang, X. Learning from Uncertainty for Big Data: Future Analytical Challenges and Strategies. - IEEE Systems, Man., & Cybernetics Magazine, April 2016, pp.26-32 (Wang, X., Y. He)

11. Zheng, Y. Urban Computing, Cambridge, The MIT Press, 2018, 609 p. (Zheng, Y.)
12. Ngiam, J. Multimodal deep learning. - 28th Inter. Conference on Machine Learning. ACM, June 28 - July 2, 2011, pp. 689-696.
13. Srivastava, N. Multimodal learning with deep boltzmann machines. - Advances in Neural Information Processing Systems, MIT, 2012, vol. 25, pp. 2231-2239.
14. Ouyang, W. Multi-source deep learning for human pose estimation. - IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2014, pp. 2337-2344.
15. Zhang, Q. Deep computation model for unsupervised feature learning on big data. - IEEE Transactions on Services Computing, vol. 9, 2016, pp. 161-171. (Zhang Q., L. Yang, Z. Chen.)
16. Chen, M. Marginalized denoising autoencoders for domain adaptation. - 29th International Conference in Machine Learning, Edingburgh, 2012, Scotland
17. Zhang, J. Semi-supervised tensor learning for image classification. - Multimedia Syst., vol. 23, no. 1, 2017, pp. 63-73. (Zhang J., Y. Han, J. Jiang)
18. Krizhevsky, A. Imagenet classification with deep convolutional neural networks. - Advances in Neural Information Processing Systems. Curran Associates, Inc. 2012, vol. 25. pp 1106-1114.
19. Maggiori, Y. Convolutional neural networks for large-scale remote-sensing image classification. - IEEE Trans. Geosci. Remote Sens., vol.55, no.2, 2017, pp. 645-657. (Maggiori Y., G. Tarabalka, P. Charpiat)
20. Jia, Z. Beyond data and model parallelism for deep neural networks. - arXiv:1807.05358v1 [cs.DC] 14 Jul 2018, pp.1-15. (Jia, Z., M. Zaharia, A. Aiken)
21. Dean, J. Large scale distributed deep networks. - Proceedings of NIPS, 2012, pp. 1232-1240.
22. Sun, P. Optimizing Network Performance for Distributed DNN Training on GPUClusters: ImageNet/AlexNet Training in 1.5 Minutes. - arXiv preprint arXiv:1902.06855, 2019. (Sun P., W. Feng, R. Han, S. Yan, Y. Wen)
23. Coats, A. Deep learning with COTS HPC systems. - J. Mach. Learn. Res., vol.28, 2013, pp. 1337-1345. (A. Coats A., B. Huval, T. Wng, D. Wu, A. Wu)
24. Novikov, A. Tensorizing neural networks. - Advances in Neural Information Processing Systems, MIT, 2015, pp. 442-450.
25. Zhou, G. Online incremental feature learning with denoising autoencoders. - International Conference on Artificial Intelligence and Statistics. JMLR.org. 2012, pp 1453-1461.
26. Calandra, R. Learning deep belief networks from non-stationary streams. - Artificial Neural Networks and Machine Learning-ICANN 2012. Springer, Berlin Heidelberg. 2012, pp. 379-386.
27. Li, Y. Online Real-Time Analysis of Data Streams Based on an Incremental High-Order Deep Learning Model.- IEEE Access, vol. 6, 2018, pp. 77615 - 77623, (Li Y., M. Zhang, W. Wang)
28. Wang, R. Non-local auto-encoder with collaborative stabilization for image restoration. - IEEE Transactions on Image Processing, vol. 25, no. 5, 2016, pp. 2117-2129. (Wang R., D. Tao)
29. Mao, X. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections - Advances in Neural Information Processing Systems, vol. 29, 2016, pp. 2802-2810.
30. Zhang, K. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. - IEEE Transactions on Image Processing, vol. 26, no. 7, 2017, pp. 3142 - 31555, 2017. (Zhang K., W. Zuo, Y. Chen, D. Meng, L. Zhang)
31. Valsesia, D. Image denoising with graph-Convolutional Neural Networks. - 2019 IEEE International Conference on Image Processing (ICIP), 22-25 Sept. 2019, pp. 2399 - 2403.