JCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Volume-4, Special Issue-3, May 2016 E-ISSN: 2347-2693

Efficient Deep Learning for Big Data: A Review

Leelavathi MV[1], Sahana Devi K J[2]

[1]*M.Tech SCS, Department Of Computer Science & Engineering,*
East West Institute of Technology, Bengaluru, India
[2] *Assistant Professor, Department of Computer Science & Engineering,*
East West Institute of Technology, Bengaluru, India

Available online at: www.ijcseonline.org

Abstract— The data science is composed of Big Data Analytics (BDA) and Deep Learning (DL). Apart from this Big Data (BD) has got popularity due to its importance in the present genre for both the public and private organizations, as this applies to collection of huge data. Basically, the BD is composed of many national intelligence applications, medical technology, cyber security data, etc. Many of the companies are analyzing the BD for its business purpose. The DL is the sequential or active learning process which collects the complex data and high-level data. This DL has its own beneficial key functions like learning and analysis of the enormous volume of unsupervised data (UD). This performs as the most valuable data analytics tool for BDA. In this paper, a brief overview of Deep learning in Big Data Analytics is presented with the challenges of DL in BD. The statistical survey is formulated by using IEEExplore. Finally, the paper future study requirements in Deep learning are discussed.

Keywords— Big Data, Big Data Analytics, Deep Learning, Machine Learning, Unsupervised Data

I. INTRODUCTION

Big Data (BD) is the data; that exceeds the storage, processing, and computing capacity of traditional databases and Data Analysis (DA) techniques. BD is increasing with the aid of elevated data storage capabilities, computational processing vigor, and availability of information volumes, which provide the group more data than current computing methods and technologies to approach [1]. Also, tremendous volumes of data, BD can be associated with other distinct complexities, like Veracity, Velocity, Volume and Variety. The unmanageable enormous Quantity of data poses an on the more challenges to computing process and requires scalable storage and a distributed procedure for data querying and analysis. Big Data Analytics (BDA) will have challenges due to above complexities [2]. Some problems includes: data validation, data quality, engineering features, data cleaning, data dimensionality, data representation, data reduction, data sampling, data scalability, data visualization, distributed and parallel data processing, analysis in real time, decision making of data, input for data analysis, Parallel, distributed data processing, decision making and real-time analysis, crowd sourcing and semantic enter for elevated DA, analyzing and data tracing provenance, data integration, and discovery, distributed and parallel computing, interpretation and exploratory data analysis (EDA), integration of heterogeneous data and also development of new models for huge data computation [3,4,5, 6].

Deep Learning algorithms (DLA) are the most promising algorithms in the field of research to have extract the complex data representations at very high level of abstraction. These algorithms develop a layered and hierarchical architecture for data learning and data representation, in which the high-level data features are defined as low-level data features [7]. DLA hierarchical architecture for learning is designed on the basis of artificial intelligence copying the layered learning process deeply from the primary sensor areas of neocortex in the human brain, which functions the extraction of data features and also the abstraction from underlying data [8,9]. DLA is more useful when it comes to the learning of the enormous amount of unsupervised data and mainly learns the data representation in a layerwise manner. From the many studies, it is noted that the data representations that are obtained from the stacking up on linear data features extractor like in DL will provide the better results of machine learning for an example: more improved modeling classification, data representations invariant properties and better quality of data samples in the probilastic model. DL solutions have the outcome with the outstanding results for different applications of machine learning like speech recognization, computer vision and processing of natural language [10, 11, 12].

DLA is largely untapped in compared with the BDA. Some BD topics like speech recognization and computer vision are used in DL applications for improving the modeling results. The DL has the ability of extraction of high-level data, data representation, and complex abstractions from large voluminous data, mainly unsupervised data; this makes DL as a most attractive tool for BDA [13]. The problems of BD like data tagging, retrieval of fast information, semantic indexing and also discriminative modeling is addressed better by using DL. The Traditional ML and also feature engineering algorithms are not much efficient to extract the data in the nonlinear and complex manner, as in BD. DL provides the simpler model for BD analysis tasks like prediction, classification; by extracting the above features. This is more important while developing the model; that deals with the Scalability of BD. This paper provides a survey of deep learning in BDA, overview, challenges of DL in BD and also the future research flow [14, 15]. This paper is organized as follows. Section II provides the background for the study like big data analytics (BDA), Deep learning. Section III gives the literature review. Section IV presents the challenges of DL in BDA (DL) and its challenges in BDA. Section V provides the IEEE Xplore statistics in DL in BDA, and Section VI concludes the paper.

II. BACKGROUND

This section deals with the study background. Machine learning (ML) main intention is to provide the representation of input data and also the generalization of learned data patterns for future data use. The machine learner's performance is mainly impacted by good data representation; the poor data representation is mainly to reduce the performance of complex and advanced machine learner [16, 17, 18, 19]. For simple machine learner, the good data representation will provide better performance. The feature engineering that generally enables the construction of features and also data representations for the raw data, this is the important factor of ML. The feature engineering consumes the more portion of effort in ML task and is generally the domain specific that involves less human input [20, 21, 22, 23]. There are many ML algorithms in which the important are: artificial neural network, deep learning, big data analytics [24] is discussed as follows:

A. Artificial Neural Network (ANN)

Artificial Neural Network (ANN) [25, 26, 27] is an algorithm in ML. This is entirely inspired by biological neural network of brain. The schematic model of ANN is as shown in Figure 1.

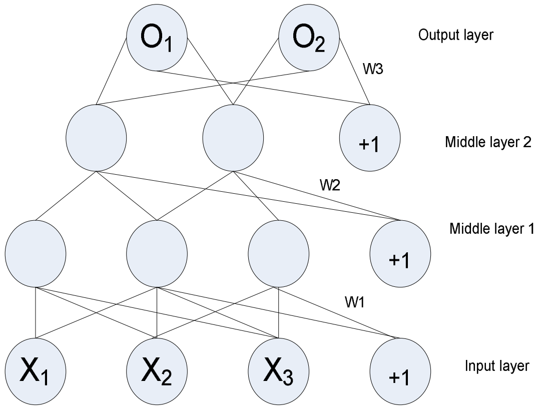

Figure 1: The schematic of ANN

From the above figure, ANN is like forward computation of neural system in multilayers. Every layer is interconnected. The neuron has no interconnection with neurons. The neurons are known as nodes. ANN is used to optimize the weight parameters in between neighboring layers and the bias parameters (like +1 ring feature), it will use optimal parameters in real data.

B. Big data analytics (BDA)

Big Data are known as the data which is more than storage, processing and also computing capacity of traditional databases and DA techniques. BD requires tools and methods to be applied for analysis and extraction patterns from huge data. The data is increasing due to growing in the data storage capability [28, 29]. BD involves a progression of technical and cultural changes in business and then traditional analysis and BDA. The BDA comes with data enterprise and tools that provides overview for BDA. The data typically stored in a data warehouse and is analyzed by using SQL-based tools [30, 31]. Most of the data in the warehouse are gathered from business transactions normally in a database. The advanced data analysis like data mining, statistical analysis, predictive analytics, and also text mining, most of the companies have moved the data to the huge servers for analysis. BDA helps in identification of challenges of BD which are too vast and unstructured and cannot be managed by the help of old methods. The institution, government, business sectors are generating the data complexity. The extraction of these data with ease is the challenging issue, hence the analytics has got more importance in analyzing the BD, to improve the business performance. There are many tools to manage data volume, variety and velocity in recent past. The technologies adopted for analytics are more open source and of less expensive, Ex: Hadoop framework, this extracts the incoming data streams and distributes on cheap disk; that gives the tools to analyze the data. The IT departments in this technology are needed to be more expertise. Even these technologies are not much enough for BDA.

The Big data analytics process [34, 35] is described in figure 2, in which the processing part of Big data is represented in black ash color, which is based on Hive/PIG/Hadoop technology with the implementation of ETL logic. The more machines are added by using the Map Reduce model. The cloud computing resource representation in this model is the biggest task. The part which is bordered with black color indicates the deep analysis process, which involves data partition, model creation and model validation.

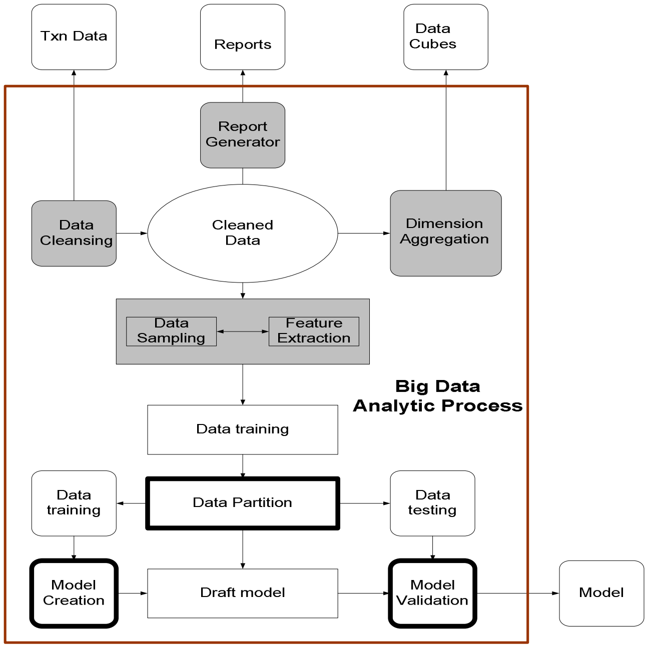

Figure 2: Big Data analytics process

C. Deep Learning:

Deep learning (DL) [36, 37, 38, 39] is a branch of ML based on a set of algorithms. DL involves ANNs like Deep Neural Networks (DNNs), Convolution Neural Networks (CNNs), Deep Belief Networks (DBNs) and Stacked Auto-Encoder (SAE). Recently DL has gained popularity due to its vast application in computer field. From many research work concluded that in many areas of applications, DL is best method compared to past methods.

The DL helps in replacing handcrafted features with efficient algorithms for unsupervised or semi-supervised feature learning and hierarchical feature extraction. Example: Auto-encoder, this aims at learning an efficient, data sets of compressed. The autoencoder structure is shown in Figure 3.

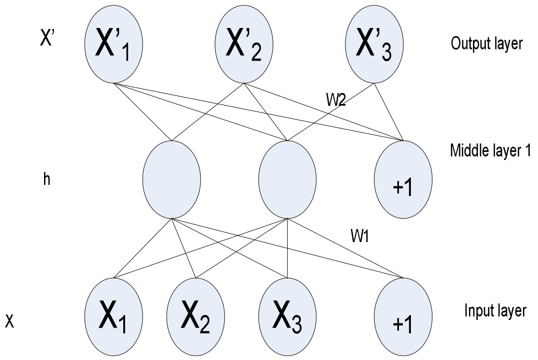

Figure 3: Schematic of autoencoder

III. LITERATURE REVIEW

This section deals with the studies of big data (BD) over deep learning (DL). There are many researchers are contributed in BD but less in number for DL and are discussed following.

Zhang and Chen [51] has presented paradigm of distributed learning for Restricted Boltzmann Machines (RBMs) and also an algorithm of a back propagation (BP) by using Map Reduce and a programming model parallel. The author has gone through the Deep Belief Nets (DBNs) and RBNs. The DBN learning starts with the RBNs series restraining and by fine tuning of entire net by using BP. Basically, the implementation of both BP algorithm and the RBMs will provide useful computation time to process the huge data sets. The authors have validated the various problems of data sets and outcome with the results of better performance as efficiency and accuracy point of view by distributed DBNs and RBNs.

Authors, Chen and Lin [2] has given an overview of deep learning (DL), past research efforts and also the challenges of big data and provided a future scope for DL in machine learning (ML). As per the study of the author, DL is the active topic in the research field and has got popularity due its vast application in many areas like speech recognition, language processing and also provides solutions in Big Data Analytics (BDA).

Wu et al. [3] have presented a weekly semi-supervised deep learning for multi-label image annotation (WeSed) approach in deep learning. The author has studied both weakly labeled and unlabeled type of images for annotation of multi-label image. The experiments on these types of images result from better performance by using the proposed approach.

Zhang et al. [4] have illustrated a framework to obtain the underlying circumstances for big video data, the framework comprised of an approach which combines both historical and current view to have better context in which DL techniques are used to obtain the raw context data. Authors have carried out initial evaluation to represent the proposed approach effectiveness and accuracy in obtaining the raw contexts. The approach outcomes with effective results in real time context awareness in video cloud.

Zhou et al. [5] have presented Stacked Extreme Learning Machine (S-ELMs) to solve the issues of large and complex data. Authors have studied past researches on ELMs. The proposed S-ELMs divide the large network of ELM into some stacked ELMs, which connected in series. The S-ELMs helps in approximation of large ELM system, where the small memory is required. The study outcomes with better accuracy in testing (than Support Vector Machine) and slightly better accuracy than DBN with faster speed in training.

Yisheng et al. [6] have presented a traffic flow prediction (TFP) method based on deep learning. Authors have worked on the increasing traffic data issues in real-time applications. A stacked autoencoder (S-AE) model is used to learn the generic features of traffic flow. The experiment outcomes with the superior performance based TFP method.

Park et al. [7] have presented the low-cost platform for DL applications like mobile and other portable devices. Author have worked on DL algorithm and implemented energy efficient DL and interference processors for wearable systems. The study outcomes with the energy efficient model than the state of art.

Jun Wang et al. [8] have discussed the survey of learning to hash framework, various types of representative techniques, composed of supervised, semi supervised and unsupervised. The study presents future research flow requirement in BDA.

Zhang et al. [9] have illustrated a deep computational model to learn the big data features. The model uses the tensor distance in the output layer as the average sum of squares error term of reconstruction error. The model is experimented on four datasets by comparing them with multimodel DL model and stacking autoencoders. The study results illustrate that the model is efficient in learning using CUAVE, INEX, SANE and STL-10 datasets.

Leung et al. [10] have discussed the introduction of machine learning (ML) tasks, to address the issues in genomic medicine (GM) (helps in determination of individual DNA variations). The study provides the platform for future computational method for effective GM.

IV. Deep Learning Challenges in Big Data Analytics

BD has challenges for modifying and adapting DL for addressing those issues. The application of DLA for BDA Analytics involves unexplored high dimensional data and needs for development of DL solutions, which may adapt approaches like BDA or produce solutions for addressing the high-dimensionality in BD domains. Large-scale DL models are more suited for huge volumes of input comprised with Big Data; these are better in elaborate data patterns from huge volumes of data. Picking the foremost number of model parameters in such giant-scale units and improving their computational methods pose challenges in DL for BDA [40, 41]. Also, BDA has to face other BD issues, like streaming data and domain adaptation. This needs for DLA and architectures. The challenges are tabled in Table 1.

Table 1: The challenges of deep learning

Issues	*Description*
Required Learning for non-stationary data	This is the biggest problem in BDA, which deals with fast moving and streaming data. This DA helps in fraud detection. Thus, there is requirement of DL adoption for streaming data handle, as continuous input data is necessary.
High dimensional data	Few DLA is computationally-expensive, while dealing with high-dimensional data like images, due to slow learning process is linked with a DL of layered hierarchy of data representations and abstractions from a lower level layer to a higher level layer. i.e., DLA is stymied, while using for huge data with BDA. A high-dimensional data source favors heavily to the volume of the raw data.
Large scale models	In analytics and computation point of view, with an enormous number of model parameters, that are able in extraction of highly complicated features and representations.

V. STATISTICS OF DEEP LEARNING IN BIG DATA

The statistics of Deep Learning (DL) in Big Data (BD) is tuned in the IEEE Xplore (cited on 05/02/2016 at 10.35am), and following data is found:

Table 2: The statistics for Big Data

SI. NO	Type	Count
1	Conference publication	11,709
2	Journal and magazine	1,160
3	Early access articles	135
4	Books and eBooks	104
5	Standards	10
6	Courses	3

Further, when the search is refined with the key term ‘Data Gathering, then the statistics are shown in table.3.

Table 3: The research Deep Learning

SI. NO	Type	Count
1	Conference publication	54
2	Journal and magazine.	18
3	Early access articles	4

From the above survey, it is found that DL in BD; (table 2 and table 3), the DL have conferencing publications only of 0.46% of BD while the journals and magazines are of 1.55%. The early access articles found DL in BD is only 2.96 % and have 1.9%. There are no books and eBooks, standards and courses of DL in BD.

The plot of above statistic is shown in figure 4, where the x-axis represents the IEEE Xplore data type, and y-axis gives the count.

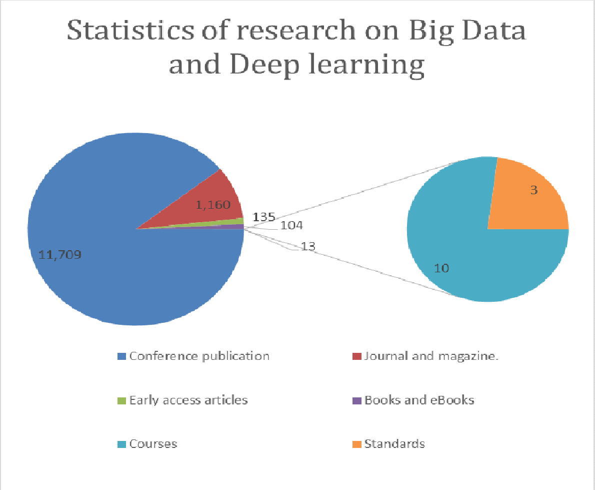

Figure.4: Statistics of deep learning and Big data analytics

VI. CONCLUSION

Compared to the other Machine learning algorithms Deep learning is the better solution for solving the issues of big data analytics. Also for the data extraction from unsupervised data. The structural learning, data extraction, and data abstraction help in solving the issues of BDA, like data tagging, semantic indexing, information retrieval etc. This paper presents the survey of deep learning which adversely helps for future work in DL for BDA

VII. REFERENCES

[1] Sato, Aki-Hiro. Applied Data-Centric Social Sciences. Springer, 2014.

[2] Ballard, Chuck, et al. Information Governance Principles and Practices for a Big Data Landscape. IBM Redbooks, 2014.

[3] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Elsevier, 2011.

[4] Hand, David J., Heikki Mannila, and Padhraic Smyth. Principles of data mining. MIT Press, 2001.

[5] Liu, Huan, and Hiroshi Motoda. Feature selection for knowledge discovery and data mining. Vol. 454. Springer Science & Business Media, 2012.

[6] Olson, David L., and Dursun Delen. Advanced data mining techniques. Springer Science & Business Media, 2008.

[7] De Castro, Leandro Nunes. Fundamentals of natural computing: basic concepts, algorithms, and applications. CRC Press, 2006.

[8] Bengio, Yoshua. "Learning deep architectures for AI." Foundations and Trends in Machine Learning 2.1, pp.1-127, 2009.

[9] Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. "Domain adaptation for large-scale sentiment classification: A deep learning approach." Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011.

[10] Ngiam, Jiquan, et al. "On Optimization methods for deep learning." Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011.

[11] Snoek, Jasper, Hugo Larochelle, and Ryan P. Adams. "Practical bayesian optimization of machine learning algorithms." Advances in neural information processing systems. 2012.

[12] MacKay, David JC. Information theory, inference and learning algorithms. Cambridge university press, 2003.

[13] Lee, Honglak, et al. "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations." Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009.

[14] Chen, Xue-Wen, and Xiaotong Lin. "Big data deep learning: challenges and perspectives." Access, IEEE 2, pp. 514-525, 2014.

[15] Najafabadi, Maryam M., et al. "Deep learning applications and challenges in big data analytics." Journal of Big Data 2.1, pp. 1-21, 2015.

[16] Goldberg, David E., and John H. Holland. "Genetic algorithms and machine learning." Machine learning 3.2, pp. 95-99, 1988.

[17] Andrieu, Christophe, et al. "An introduction to MCMC for machine learning." Machine learning 50.1-2, 5-43, 2003.

[18] Kubat, Miroslav, Robert C. Holte, and Stan Matwin. "Machine learning for the detection of oil spills in satellite radar images." Machine learning 30.2-3, pp. 195-215, 1998.

[19] Rasmussen, Carl Edward. "Gaussian processes for machine learning." 2006.

[20] Freitag, Dayne. "Machine learning for information extraction in informal domains." Machine learning 39.2-3, pp. 169-202, 2000.

[21] Sebastiani, Fabrizio. "Machine learning in automated text categorization." ACM computing surveys (CSUR) 34.1, pp. 1-47, 2002.

[22] Michie, Donald, David J. Spiegelhalter, and Charles C. Taylor. "Machine learning, neural and statistical classification." 1994.

[23] Carbonell, Jaime G., Ryszard S. Michalski, and Tom M. Mitchell. "An overview of machine learning." Machine learning. Springer Berlin Heidelberg, pp. 3-23, 1983.

[24] Aha, David W., Dennis Kibler, and Marc K. Albert. "Instance-based learning algorithms." Machine Learning 6.1, pp. 37-66, 1991.

[25] Wang, Sun-Chong. "Artificial neural network." Interdisciplinary Computing in Java Programming. Springer US, pp. 81-100, 2003.

[26] Hagan, Martin T., et al. Neural network design. Vol. 20. Boston: PWS publishing company, 1996.

[27] Hsu, Kuo-lin, Hoshin Vijai Gupta, and Soroosh Sorooshian. "Artificial neural network modeling of the rainfall-runoff process." Water resources research 31.10, pp.2517-2530, 1995.

[28] Srinivasa, Srinath, and V. Bhatnagar. "Big data analytics." Proceedings of the First International Conference on Big Data Analytics BDA. 2012.

[29] Miller, Steven. "Big Data Analytics." (2013).

[30] Chen, Hsinchun, Roger HL Chiang, and Veda C. Storey. "Business Intelligence and Analytics: From Big Data to Big Impact." MIS quarterly 36.4, pp.1165-1188, 2012.

[31] Govindaraju, Venu, Vijay Raghavan, and C. R. Rao. Big Data Analytics. Vol. 33. Elsevier, 2015.

[32] Zakir, Jasmine, Tom Seymour, and Kristi Berg. "BIG DATA ANALYTICS." Issues in Information Systems 16.2 (2015).

[33] Suthaharan, Shan. "Big Data Analytics." Machine Learning Models and Algorithms for Big Data Classification. Springer US, pp. 31-75, 2016.

[34] Zikopoulos, Paul, and Chris Eaton. Understanding big data: Analytics for enterprise class Hadoop and streaming data. McGraw-Hill Osborne Media, 2011.

[35] LaValle, Steve, et al. "Big data, analytics and the path from insights to value." MIT Sloan management review 52.2,pp.21,2011.

[36] Weigel, Van B. Deep Learning for a Digital Age: Technology's Untapped Potential To Enrich Higher Education. Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, 2002.

[37] Arel, Itamar, Derek C. Rose, and Thomas P. Karnowski. "Deep machine learning-a new frontier in artificial intelligence research [research frontier]." Computational Intelligence Magazine, IEEE 5.4, pp.13-18, 2010.

[38] Weston, Jason, et al. "Deep learning via semi-supervised embedding." Neural Networks: Tricks of the Trade. Springer Berlin Heidelberg, pp. 639-655, 2012. .

[39] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." Nature 521.7553, pp. 436-444, 2015.

[40] Najafabadi, Maryam M., et al. "Deep learning applications and challenges in big data analytics." Journal of Big Data 2.1, pp. 1-21, 2015.

[41] Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques, and technologies: A survey of Big Data." Information Sciences 275, pp. 314-347, 2014.

[42] Kunlei Zhang; Xue-Wen Chen, "Large-Scale Deep Belief Nets With

[43] MapReduce," in Access, IEEE, vol.2, no., pp.395-403, 2014.

[44] Xue-Wen Chen; Xiaodong Lin, "Big Data Deep Learning: Challenges and Perspectives," in Access, IEEE, vol.2, no., pp.514-525, 2014

[45] Fei Wu; Zhuhao Wang; Zhongfei Zhang; Yi Yang; Jiebo Luo; Wenwu Zhu; Yueting Zhuang, "Weakly Semi-Supervised Deep Learning for Multi-Label Image Annotation," in Big Data, IEEE Transactions on, vol.1, no.3, pp.109-122, Sept. 1, 2015

[46] Weishan Zhang; Pengcheng Duan; Zhongwei Li; Qinghua Lu; Wenjuan Gong; Su Yang, "A Deep Awareness Framework for Pervasive Video Cloud," inAccess, IEEE, vol.3, no., pp.2227-2237, 2015.

[47] Hongming Zhou; Guang-Bin Huang; Zhiping Lin; Han Wang; Yeng Chai Soh, "Stacked Extreme Learning Machines," in Cybernetics, IEEE Transactions on, vol.45, no.9, pp.2013-2025, Sept. 2015

[48] Yisheng Lv; Yanjie Duan; Wenwen Kang; Zhengxi Li; Fei-Yue Wang, "Traffic Flow Prediction With Big Data: A Deep Learning Approach," in Intelligent Transportation Systems, IEEE Transactions on, vol.16, no.2, pp.865-873, April 2015

[49] Park, S.-W.; Park, J.; Bong, K.; Shin, D.; Lee, J.; Choi, S.; Yoo, H.-J., "An Energy-Efficient and Scalable Deep Learning/Inference Processor With Tetra-Parallel MIMD Architecture for Big Data Applications," in Biomedical Circuits and Systems, IEEE Transactions on , vol.9, no.6, pp.838-848, Dec. 2015

[50] Jun Wang; Wei Liu; Kumar, S.; Shih-Fu Chang, "Learning to Hash for Indexing Big Data—A Survey," in Proceedings of the IEEE, vol.104, no.1, pp.34-57, Jan. 2016

[51] Zhang, Q.; Yang, L.T.; Chen, Z., "Deep Computation Model for Unsupervised Feature Learning on Big Data," in Services Computing, IEEE Transactions on, vol.9, no.1, pp.161-171, Jan.-Feb. 1 2016

[52] Leung, M.K.K.; Delong, A.; Alipanahi, B.; Frey, B.J., "Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets," inProceedings of the IEEE, vol.104, no.1, pp.176-197, Jan. 2016