Deep Learning and its Applications: A Survey

C. Sreedhar, N. Kasiviswanath

Abstract— **Deep learning has gained its attractions towards the researchers, academicians and several other organizations in the research areas of big data, genomics, natural language processing, healthcare and brings several challenges and opportunities for various domains such as cosmology, pharmacy and astrophysics. With the increased growth of the data produced by our digital world, deep learning becomes inevitable in providing solutions to the complex and real world problems. It is not far for the researchers to find the solutions for the complex problems to be solved till date such as birth of stars, genetic mutations, cosmological and even the birth and extended life of living organisms. Deep learning will definitely be the active component in finding solutions to such problems. This paper makes an attempt to present the research studies in deep learning with various perspectives.**

Index Terms—**Deep learning, big data, healthcare, genomics.**

I. INTRODUCTION

Deep learning algorithms gets its support from various data mining algorithms along with its applications such as logical, cognitive science, probability, databases and machine learning [1]. Deep learning is considered to be one of the successful tools, and has become very popular in the literature [2]. Motivated by the needs of digital world, information analysis issues and several mining algorithms are studied.

Due to the lack in conceptual modeling, it can jeopardize more development of knowledge mining. Deep learning can be considered research oriented problem in solving several issues that cannot be solved using traditional methods.

Second, understandings in deep learning as a scientific perspective in several fields like genomics and big data are studied [3]. Deep learning involves indepth knowledge on design, developmental, integrative and autonomous algorithms capable of self decision system which was widely used in Artificial Intelligence (AI) into their operations.

Deep learning has several applications [4]. In healthcare it can discover the hidden patterns and opportunities from the data stored related to clinical and medical and can mutually benefitted by both patients as well as doctors. Discovery of new medicines, drug development, medical imaging, diagnosing severe health conditions, analyzing medical insurance fraud claims are some of the areas where deep learning helps in understanding, analyzing and predicting the problems related to health. Deep learning also has the applications in genomics. Advancements in genetic research using high throughput sequencing algorithms have thrived the genomics field into big data disciplines [5][6]. Deep learning is one of the solutions to the problems that could not be solved using ML and AI techniques [7]. Artificial Neural Networks (ANN) is a technique for supervised machine learning. ANN was inspired by which the human brain acts in learning a particular task [8][9].

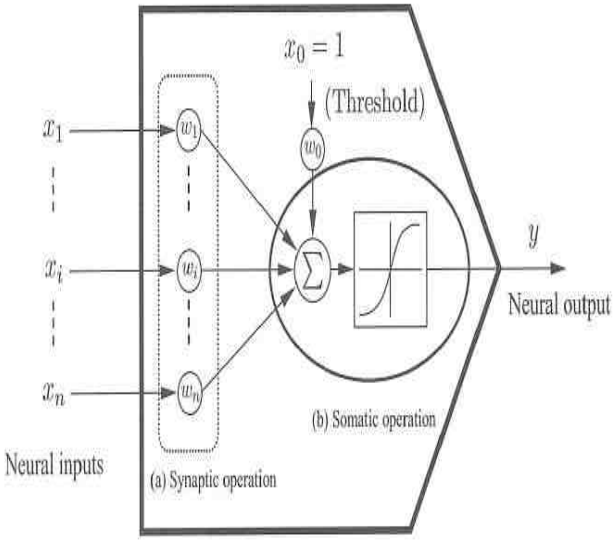

Figure 1. Artificial neural network (a neuron)

Figure 1 depicts artificial neural networks, which is an another technique of supervised machine learning, which is inspired by which our human brain solves for a particular problem. The basic idea behind the design of artificial neural network is to design network consisting of 'n' layer nodes; creating network to recognize from learning and training the network.

A MLP (MultiLayer Perceptron) is a neural network which uses back propagation method to train its data. MLP includes three layers namely input-layer, hidden-layers and output-layer. Figure 2 depicts an multilayer perceptron with two hidden layers which are used to predict temperature.

The sigmoid function, for net input x is given by

$$sigmoid(x) = \frac{1}{(1+e^{-x})}.$$

Manuscript received October11, 2019.
C Sreedhar, Department of Computer Science & Engineering, G Pulla Reddy Engineering College, Kurnool
N. Kasiviswanath, Department of Computer Science & Engineering, G Pulla Reddy Engineering College, Kurnool

II. DEEP LEARNING: BIG DATA PERSPECTIVE

Deep learning with the perspective of big data attracted several researchers in the recent years. The volume of the data generated is at a faster rate in terms of growth and speed and perhaps justify big data as an important research field. From the literature of big data [10][11] several researchers tried to focus on the various parameters that affect the big data, commonly referred as big data dimensions. Size of data stored in the databases can be stated as a dimension for big data. Traditional database technologies can no longer handle the massive volumes of data stored in order to be processed, analyzed and visualized [12]. Big data algorithms can be applied using the latest technologies to solve the problem of handling huge datasets [13]. Several studies attempted to solve the problems related to scalability speed and processing of Big data [14]. Big data can be characterized based on structured format, unstructured format and semi structured format. Clustering of data sets for sequence mining, stream mining, text mining, and web mining can be studied in big data literature. Clustering of datasets using K-means and its related research has gained its popularity in the recent study and several big data algorithms were proposed to solve the problem of grouping of data to a certain specific category based on the needs and requirements [15]. Big data techniques can be applied to such fields which are concerned with deep learning. New algorithms are introduced by several researchers, attempting to discover new types of knowledge by gaining insights from the data stored using various big data tools [16]. Text mining algorithms can be correlated to classical information retrieval algorithms.

Baidu, Facebook, Google, and Twitter organizations use deep learning algorithms to maximize their organizational profits. Facebook and Twitter are using the Torch open-source deep learning framework. TensorFlowOnSpark (TFoS) [17] is an open source software for deep learning on big data clusters. CaffeOnSpark [18] is a development of Yahoo's innovation in developing Apache Hadoop platform. CaffeOnSpark is used in many forms, and the framework was enhanced by various contributors from the community. CaffeeOnSpark, needs no separate set up to establish deep learning clusters, instead deep learning can be run on where the data is present. Issues related to TensorFlow [19], developed by Google have not been addressed on existing big data clusters. Deep Learning over Big Data [20] has been proposed to extract big insights from huge amount of data stored.

Traditional learning algorithms makes use of shallow structured learning architectures, where as deep learning uses machine learning algorithms and techniques such as supervised, unsupervised methods [21] [22].

Boltzmann machine can be used for deep learning, which learns feature representations from different types of big data formats such as structured and semi structured data. Hop-field network, is a type of stochastic recurrent neural network. No connection exists between the units of the same layer as in the case of Restricted Boltzmann Machine. The probability assigned to vector v_i is

$$p(v_i) = \frac{1}{Z}$$

$$\sum_{h_i} e^{\sum abWab^{(1)} v_i hb^{(1)} + \sum bcWbc^{(2)} h_b^{(1)} h_i^{(2)} + \sum_{cd} W_{cd}^{(3)} h_c^{(2)} h_d^{(3)}}$$

The human knowledge is contextual and conceptual which is organized hierarchically. In summary, inspired by human knowledge, deep learning algorithms learn by utilizing efficient strategies from effective and efficient training using several layers.

III. DEEP LEARNING: COSMOLOGICAL PERSPECTIVE

Deep learning methods can be used for in cosmological domain in measuring the parameters related to cosmology. Deep Convolution Neural Network (DCNN) is a method which learns the relation between the various cosmological models and the maps generated. DCNN can be used to to design and develop cosmological models using convergence mass maps. This algorithm captures non-Gaussian information from the mass maps, and can be trained on a set of simulations in the real world. Training a DCCN is a challenging issue due to the high level of noise incorporated to the convergence maps and can be solved using artificial intelligence, machine learning and deep learning techniques [23].

IV. DEEP LEARNING: GENOMICS PERSPECTIVE

Several areas of computer science faces the problem of estimating the parameters used in theoretical models using simulated experimental data [24] [25]. A genome is an instruction book for building an organism [26]. Large volumes of data related to genomics have been generated as a result of next-generation sequencing, which can sequence the complete genome within a reduced time as compared with the earlier years. Deep learning can be used to solve key problems in genomic medicine [27] [28]. Genomics is an area within genetics which studies about understanding and analyzing the structure encoded in the DNA sequences of an organism's genome [29] [30].

CRISPR (clustered regularly interspaced short palindromic repeats) [31] [32] [33] is gene editing tool that can read the text of the genome and support genomic medicine. CRISPR can be used in gene therapies with targeted modifications, mutations, adding or deleting sequences at predetermined locations in a genome. Deep learning in genomics paves a path to unprecedented opportunities in genomic medicine with the features of tailored treatment for genetic diseased based on genetic information about the patient [34].

The forward algorithm and backward algorithms [35] are used to find the probability in sequencing methods. The Forward algorithm and the backward algorithm uses dynamic programming to calculate the probability of having to enumerate the probabilities observed across all the possible paths.

Table 1. Forward Algorithm

Forward Algorithm

Notations used:

$f_i(j)$ – probability of feature parameter at position j in state i

$t_m(x_i)$ - probability of emitting x_i by the state m

s_{km} - probability of transition from state k to state i

$P(x)$ - probability of observing the entire sequence x

L - length of the sequence

Step 1: Initialization (i = 0) :

Step 2: Recursion (i = 1 .... L) :

$$f_i(j) = t_m(x_i)\sum_k f_k(i-1)s_{km}$$

Step 3: Termination : $P(x) = \sum_k f_k(L)a_{k0}$

Table 2. Backward Algorithm

Backward Algorithm

Notations used:

$b_k(i)$ - probability of observing sequence when in state k having i symbols

$e_m(x_i)$ - is the probability of emitting the symbol x_i by state m

a_{km} - is the probability of transition from state k to state m

$P(x)$ - is the probability of observing the entire sequence

Step 1: Initialization (i = L) : $b_k(L) = a_{k0}$ *for all k*

Step 2: Recursion (i = L-1 ..... 1) :

$$b_k(i) = \sum_m a_{km}e_m(x_{i+1})b_m(i+1)$$

Step 3: Termination : $P(x) = \sum_m a_{0m}e_m(x_1)b_m(1)$

V. DEEP LEARNING: HEALTHCARE PERSPECTIVES

Predictive analytics and deep learning in Health care data gained its popularity in the recent years with an exponential growth of volumes of data in the form of unstructured and heterogeneous medical data [36]. Massive volumes of data along with variety and velocity of big data in the field of healthcare needs sophisticated learning algorithms to get the big insights out of the data stored. Novel deep learning algorithms are to be designed towards the opportunities for utilizing healthcare big data in reducing patient's costs, readmissions and diagnosis [37][38][39].

VI. CONCLUSION

Deep learning has the ability to design and implement models without the strong underlying mechanisms of the domain [40]. With the amount of data related to patients that are captured with several features deep learning can solve the complex problems that cannot be solved using traditional methods. The problems and challenges in finding the solutions to the complex problems in health care can lead tp several opportunities and future research possibilities to improve the field. The recent study on deep learning should make an attempt to combine several types of sources related to medical data using deep learning .

Deep learning can be used in several applications. In almost every domain, obtaining 100% accuracy may not be required because deep learning will primarily prioritize experiments and assist discovery. For example, in cosmological research for meteor discovery in finding the chances of passing by or hitting the planet, a deep learning system is more technical.

In healthcare, medical images, DL can solve the most challenging cases that need manual attention. We conclude that deep learning field has not yet reached its maximum capacity to gain big insights out of it. Deep learning has the flexibility in modeling methods that are infeasible with machine learning algorithms.

Researchers should primarily focus on new predictive deep learning algorithms can summarize massive volumes of input data which can be predicted and analyzed over successful training, validation and best results.

REFERENCES

[1] Haohan Wang, Aman Gupta, and Ming Xu. Extracting compact representation of knowledge from gene expression data for protein-protein interaction. International Journal of Data Mining and Bioinformatics, 17(4):279–292, 2017b..

[2] I. Goodfellow, Y. Bengio, A. Courville, "Deep Learning", Cambridge, MA: MIT Press, 2017.

[3] A. De Mauro, M. Greco, M. Grimaldi, "A formal definition of Big Data based on its essential features", Library Review, Vol. 65 Issue: 3, 2016, pp.122-135.

[4] Haohan Wang, Bhiksha Raj, and Eric P Xing. On the origin of deep learning. arXiv preprint arXiv:1702.07800, 2017d.

[5] Tianwei Yue, Haohan Wang, "Deep Learning for Genomics: A Concise Overview", Handbook of Deep Learning Applications, arXiv:1802.00810, 2018.

[6] Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed-model association methods. Nature genetics, 46(2):100–106, 2014.

[7] Rui Xie, Jia Wen, Andrew Quitadamo, Jianlin Cheng, and Xinghua Shi. A deep autoencoder model for gene expression prediction. BMC genomics, 18(9):845, 2017.

[8] P. Dadvand, R. Lopez, and E. O˜nate, "Artificial neural networks for the solution of inverse problems". In Proceedings of the International Conference on Design Optimisation Methods and Applications ERCOFTAC, 2006.

[9] J.J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities". In Proceedings of the National Academy of Sciences, volume 79, pages 2554–2558, 1982

[10] C. Sreedhar, Dr. N. Kasiviswanath "A Survey on Big Data Management", International Journal of Engineering Research And Management (IJERM), Volume.02, Issue 04, pp. 24-28, ISSN: 2349-2058, April 2015.

[11] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS computational biology, 13(1): e1005324, 2017e.
[12] Sreedhar. C, N. Kasiviswanath, P. Chenna Reddy "A Survey on Big Data Management and Job Scheduling", International Journal of Computer Applications (IJCA), Volume.130, No 13, pp. 41-49, ISSN: 0975-8887, November 2015.
[13] Saurabh Singh, Derek Hoiem, and David Forsyth. Swapout: Learning an ensemble of deep architectures. NIPS, 2016.
[14] C. Sreedhar, N. Kasiviswanath and P. Chenna Reddy, "A Novel Multilevel Queue based Performance Analysis of Hadoop Job Schedulers", Indian Journal of Science and Technology (IndJST), Volume. 9, No. 44, ISSN: 0974-5645, November 2016.
[15] C. Sreedhar, N. Kasiviswanath and P. Chenna Reddy, "Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop", Journal of Big Data, Springer, 4:27, DOI 10.1186/s40537-017-0087-2, September 2017.
[16] Wenlu Zhang, Rongjian Li, Tao Zeng, Qian Sun, Sudhir Kumar, Jieping Ye, and Shuiwang Ji. Deep model based transfer and multi-task learning for biological image analysis. IEEE Transactions on Big Data, 2016.
[17] Mikyoung Lee, Sungho Shin, Seungkyun Hong, and Sa-kwang Song, "BAIPAS: Distributed Deep Learning Platform with Data Locality and Shuffling", Internationa Journal of Education and Information Technologies, Volume 11, 2017, pp. 190-195.
[18] https://github.com/yahoo/CaffeOnSpark
[19] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean Matthieu Devin Sanjay Ghemawat, Geoffrey Irving, Michael Isard,Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden,Martin Wicke,Yuan Yu, Xiaoqiang Zheng. In Proceeding OSDI'16 Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation pp. 265-283 , Savannah, GA, USA — November 02 - 04, 2016
[20] Xiaoyi Lu and Haiyang Shi and M. Haseeb Javed and Rajarshi Biswas and Dhabaleswar K. Panda, "Characterizing Deep Learning over Big Data (DLoBD) Stacks on RDMA-Capable Networks", IEEE 25th Annual Symposium on High-Performance Interconnects (HOTI), 2017, pp. 87-94.
[21] Y. Bengio and S. Bengio, "Modeling high-dimensional discrete data with multi-layer neural networks," in Proc. Adv. Neural Inf. Process. Syst.,vol. 12. 2000, pp. 400-406.
[22] Y. Marc Aurelio Ranzato, L. Boureau, and Y. LeCun, "Sparse feature learning for deep belief networks," in Proc. Adv. Neural Inf. Process.Syst., vol. 20. 2007, pp. 1185-1192.
[23] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Wei Xie, Gail L Rosen, et al. Opportunities and obstacles for deep learning in biology and medicine. bioRxiv, page 142760, 2017.
[24] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. Nature methods, 12(10):931–934, 2015.
[25] Jingjun Cao, Zhengli Wu, Wenting Ye, and Haohan Wang. Learning functional embedding of genes governed by pair-wised labels. In Computational Intelligence and Applications (ICCIA), 2017 2nd IEEE International Conference on, pages 397–401. IEEE, 2017a.
[26] Chao Cheng, Koon-Kiu Yan, Kevin Y Yip, Joel Rozowsky, Roger Alexander, Chong Shou, and Mark Gerstein. A statistical framework for modeling gene expression using chromatin features and application to modencode datasets. Genome biology, 12(2):R15, 2011.
[27] Renzhi Cao, Debswapna Bhattacharya, Badri Adhikari, Jilong Li, and Jianlin Cheng. Largescale model quality assessment for improving protein tertiary structure prediction. Bioinformatics, 31(12):i116–i123, 2015.
[28] Hyun Min Kang, Jae Hoon Sul, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. Nature genetics, 42(4):348–354, 2010.
[29] Dustin Tran and David M Blei. Implicit causal models for genome-wide association studies. arXiv preprint arXiv:1710.10742, 2017.
[30] Ramzan Kh Umarov and Victor V Solovyev. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. PloS one, 12(2):e0171410, 2017.
[31] Doudna J, Mali P. CRISPR-Cas: A Laboratory Manual. New York: Cold Spring Harbor Laboratory Press. ISBN 978-1-62182-131-1, 23 March 2016.] [Sander JD, Joung JK. "CRISPR-Cas systems for editing, regulating and targeting genomes". Nature Biotechnology. 32 (4), pp. 347–55, doi:10.1038/nbt.2842, PMC 4022601, PMID 24584096, April 2014.
[32] Jie Hou, Badri Adhikari, and Jianlin Cheng. Deepsf: deep convolutional neural network for mapping protein sequences to folds. Bioinformatics, 2017.
[33] Michael KK Leung, Andrew Delong, Babak Alipanahi, and Brendan J Frey. Machine learning in genomic medicine: a review of computational problems and data sets. Proceedings of the IEEE, 104(1):176–197, 2016.
[34] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Reverse-complement parameter sharing improves deep learning models for genomics. bioRxiv, page 103663, 2017.
[35] Durbin, R., Eddy, S., Krogh, A. and Mitchison, G., "Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids", Cambridge University Press, 1998.
[36] Ritambhara Singh, Jack Lanchantin, Gabriel Robins, and Yanjun Qi. Deepchrome: deeplearning for predicting gene expression from histone modifications. Bioinformatics, 32 (17):i639–i648, 2016a.
[37] Wen Torng and Russ B. Altman. 3d deep convolutional neural networks for amino acid environment similarity analysis. BMC Bioinformatics, 18(1): 302, Jun 2017. ISSN 1471-2105. doi: 10.1186/s12859-017-1702-0. URL https://doi.org/10.1186/s12859-017-1702-0.
[38] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Edinburgh neural machine translation systems for wmt 16". In Proceedings of the First Conference on Machine Translation, pages 371–376, Berlin, Germany, Association for Computational Linguistics, August 2016.
[39] Carl Edward Rasmussen and Christopher K. I. Williams, "Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)". The MIT Press, 2006. ISBN 026218253X.
[40] Xin Li and Yuhong Guo, "Adaptive active learning for image classification". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 859–866, 2013.