International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 2
ISSN: 2454-4248
276 – 280

A Study on Deep Learning for Bioinformatics

Panchami.VU*, Dr. Hariharan.N**, Dr. Manish.TI***
*Research Scholar, **Dean-PG Programs, ***Associate Professor
Adishankara Institute of Engineering and Technology, Kalady, Kerala

Abstract **–** Bioinformatics, an interdisciplinary area of biology and computer science, handles large and complex data sets with linear and non-linear relationships between attributes. To handle such relationships, deep learning has got a greater importance these days. This paper analyses different deep learning architectures and their applications in Bioinformatics. The paper also addresses the limitations and challenges of deep learning.

Index Terms—*Deep Leraning, Perceptron, Auto encoder, Recurrent Neural Network, Restricted Boltmann Machine, Convolutional Neural Network, Deep Belief Network*

******

I. Introduction

In recent years, reduction in the cost of generating genomic data have intensified the importance of research in areas like DNA-sequencing, RNA-sequencing and high-throughput screening. This has created a new challenge of finding the most efficient and effective ways to analyze data and provide insights into the function of biological systems. Bioinformatics utilize techniques from computer science, biology, statistics, and engineering to analyze and infer the relationships among biological data. It also helps to develop predictive models that identify factors within cell regulatory networks that are important in generating specific phenotypes. Proteomics analyzes the structure of nucleic acids and proteins. One of the challenging areas of research is the identification of candidate genes and single nucleotide polymorphism to understand the genetic basis of disease, unique adaptations, or differences between populations. Microarray technology provides a promising approach for exploiting gene profiles for cancer diagnosis. Bioinformatics research gives insight into evolutionary aspects molecular biology, biological pathways and networks, gene and protein expression and regulation, genome annotation and biomolecular interaction.

Analyzing large biological data sets requires making sense of the data by inferring structure or generalizations from the data. Machine learning is a promising approach for the analysis of large and complex data sets. Machine learning focuses on the development and usage of algorithms that learn from data or experience. A model is created from the training data by finding the underlying patterns, and its parameters are optimized. This best fit model is used to predict the trends in the following data. Many prediction methods now exceed the accuracy of some high-throughput laboratory methods.

Deep learning or deep structured learning has got a huge interest since early 2000s [1]. It is a subfield of Machine learning with roots in artificial neural networks (ANNs). ANNs are a class of algorithms which simulate the structure and functioning of human brain. Computer vision, natural language processing, social network filtering and biomedical research are some of the areas attracted by deep learning.

Deep learning architectures consist of a number of hidden layers which capture non-linear relationships between patterns. The training of such a network requires large amount of labeled data. It is a computationally intensive task and may consume considerable amount of resources. If the number of tuples in training set is comparable to the number of parameters, then a problem called overfitting may arise[2]. The neural network will memorize the entire training samples, but will not be able to predict the unseen samples properly. The vanishing gradient problem may arise in training when the errors are back propagated. After two or three layers, the errors may become negligibly small. It may resist neural network from proper learning.

This paper goes through different deep learning architectures that can be applied in Bioinformatics and the challenges faced by each of them. The following sections review deep learning, its different architectures, application of these architectures to Bioinformatics and their limitations.

II. Deep Learning

Deep learning is a subset of machine learning algorithms, mostly based on artificial neural networks, which consist of cascaded multiple layers with each layer getting input from the previous layer. It learns multiple levels of representations, providing different levels of abstraction[2]. Thus it can learn complex features by combining simpler features learned from data.

An ANN is a collection of computing elements or nodes called artificial neurons arranged in layers. Each neuron gets input from the previous layer, goes through a state change or activation based on the input and produces output based on the activation. The neurons in one layer are connected to the neurons in the next layer by weighted links, thus forming a directed weighted graph. It uses a learning rule or an algorithm that defines how to change the parameters of the network in order to find the mapping between input and output. The weights of the links and the threshold of the activation are modified to get this mapping proper.

Perceptron is the simplest artificial neural network which consists of only two layers - input layer and output layer [2]. The weights of the connections between these two layers can be adjusted so that it can predict the patterns of linearly separable data. There can be one or more hidden layers in the

 ISSN: 2454-4248
276 – 280

topology of ANNs so that more complex problems can be solved. These multilayer perceptrons require sophisticated learning algorithms such as backpropagation. If the function is non-linear and differentiable, delta rule can be applied for learning [2].

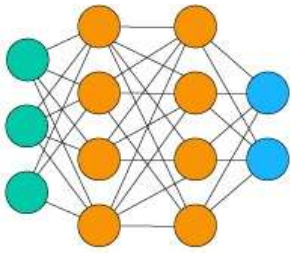

Figure 1: Multilayer Perceptron

More and more hidden layers can be added to the topology so that the resulting network can represent more complex relationships. These networks are called deep neural networks. As hidden layers depict non-linear transformations, different abstraction levels are achieved. Supervised and/or unsupervised learning is possible with deep neural networks.

The basic neural network architectures are extended to deep architectures. Each of them has got success in individual areas. Image recognition, machine translation, object classification, text generation and gaming are some of the domains that exploit deep learning.

III. DEEP NEURAL NETWORKS

Several deep neural network architectures have been proposed in literature like Deep Autoencoders, Recurrent Neural Network, Restricted Boltzmann Machine based neural networks, Convolutional Neural Networks etc. Each of them has some advantages and disadvantages and finds application in some particular domain.

a. Deep Autoencoders

Autoencoder is a neural network that uses data driven learning to extract features. It employs an unsupervised learning with an aim to reproduce the input vector. The topology consists of same number of neurons in the input and output layers and lesser number of neurons in the hidden layers. This topology help to represent data in a lower dimensional space and generation of most discriminative features[4]. Cascading many autoencoders one after another creates a Deep Autoencoder. Since it employs unsupervised learning, it does not require labeled data for training. It suffers from vanishing gradient problem and requires pretraining to handle the problem.

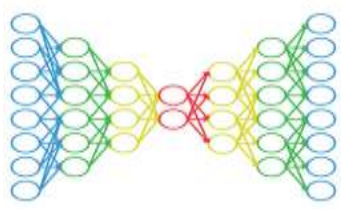

Figure 2: Deep Autoencoder

b. Recurrent Neural Networks(RNN)

It is an ANN which allows to process sequences of inputs. It is said to possess internal state or memory because the output at a time depends on data of present and recent past. Due to this, it finds application in text analysis, DNA sequencing, speech recognition and image description generation. RNN shares the same weight across all steps[5]. So the total number of parameters to be learnt is considerably small. Issues in learning may arise due to vanishing gradient problem.

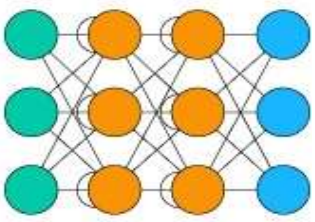

Figure 3: Recurrent Neural Network

c. Restricted Boltzmann Machine(RBM) Based

Probabilistic relationships between variables are effectively modeled using a variant of Boltzmann machine called Restricted Boltzmann Machine(RBM). It consist of several stochastic units with a particular distribution. Learning minimizes reconstruction error by gradually adjusting the weights, using a procedure called Gibbs sampling.

The input and hidden units form a bipartile graph and the nodes are undirected. There are two different deep neural network architectures based on RBM – Deep Belief Network(DBN) and Deep Boltzmann Machine(DBM).

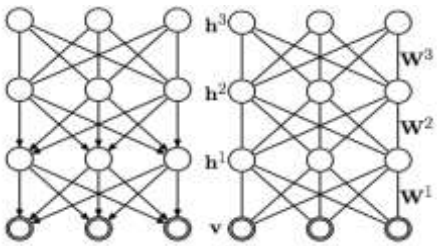

Figure 4:Deep Belief Network(left) and Deep Boltzmann Machine(right)

i. Deep Belief Network(DBN)

It is a stacked RBM topology where the hidden layer of an RBM and the visible layer of the next RBM are connected together. Top two layers consist of undirected connections where as other layers are connected by directed links [6]. Both supervised and unsupervised learning procedures are possible with this architecture. It uses a layer-by-layer greedy learning strategy for initialization of network parameters. Due to this, the training procedure is lengthy.

ii. Deep Boltzmann Machine(DBM)

The architecture formed by undirected connection between neurons of all layers is called Deep Boltzmann Machine [7]. The time complexity for inference of DBM is higher than that of DBN due to this complex topology. So, for big data sets,

 ISSN: 2454-4248
276 – 280

parameter optimization becomes difficult, thereby making this model inefficient.

d. Convolutional Neural Network(CNN)

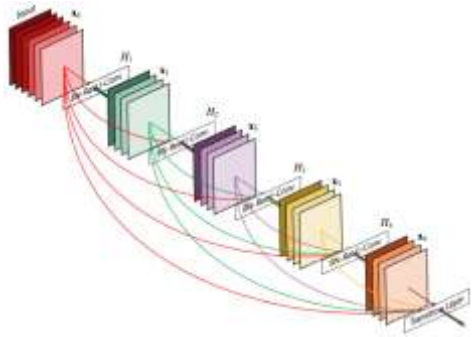

Figure 5: Convolutional Neural Network(source:Huang et al.)

It is a deep neural network architecture which mainly finds application in image recognition. The concept behind our visual cortex is used to develop this model. It consist of a sequence of convolution and subsample layers [8]. The last subsampling layer is connected to a traditional fully connected neural network for classification. This will convert the 2-D feature maps into a 1-D vector. AlexNet, Clarifai, VGG, andGoogLeNet are some neural networks that uses CNN architecture [2]. It is most suited for 2-D data . It requires only a few neuron connections as compared to other neural networks. It may require large labeled data set of images for training. And the number of layers is also large, which makes it time and resource consuming.

IV. DEEP LEARNING APPLICATIONS IN BIOINFORMATICS

Deep neural network finds great importance in Bioinformatics research due to its ability to handle complex and high-dimensional data. The previously unknown correlations between patterns are found out using these deep architectures. Genomics, an awesome icon of Bioinformatics, is a field of science which analyzes all aspects of the genome of an organism [10]. Genome is the complete set of DNA, including all of the genes. Genomics study the structure, function, evolution, mapping and editing of genomes. Another aspect of genomics is the identification of gene alleles responsible for some diseases. Proteomics is another aspect of Bioinformatics that study about proteins.

The inherently complex data sets go through preprocessing, cleaning, feature extraction, model building and evaluation procedures [11]. The traditional machine learning methods produced results by reducing the dimensionality of data. But deep learning can work with high dimensional, heterogeneous, and unbalanced data and can produce better results. The structure and properties at molecular level can be represented by a data driven approach. Because of the multiple hidden layers, nonlinear dependencies can be captured at multiple scales. Small changes in inputs are not affected due to this nonlinearity which makes the model robust.

The data used in Bioinformatics are the raw biological sequences like DNA, RNA and amino acid sequences. Sometimes features extracted from these unprocessed sequences such as position specific scoring matrices (PSSM) [12], physicochemical properties [13], Atchley factors [14] and one-dimensional structural properties are used. Microarray gene expression data and protein contact maps are also used.

Protein structure prediction is an important topic of interest for Bioinformatics researchers. It is the process of predicting the folding and secondary and tertiary structure from the primary structure of protein. The three dimensional structure is thus inferred from the amino acid sequences. This complex task is better solved by DNN. To predict protein secondary structure, Spencer et al. [15] used DBN to amino acid sequences, PSSM and Atchley factors. Stacked autoencoders are used to amino acid sequences by Heffernan et al. [16] for predicting the secondary structure, torsion angle and accessible surface area.

The class of techniques employed by cells to increase or decrease the manufacturing of particular gene products (protein or RNA) is called gene expression regulation. Gene expression is understood by a space called splice junction which is predicted using DBN by Lee et al. [17]. The work has shown better performance than previous studies. Chen et al.[18] applied MLP to both microarray and RNA-seq data to deduce gene expression.

Fakoor et al. [19] used the microarray gene expression data with dimensionality reduced using PCA, and is fed to stacked autoencoders and classified various cancers, including acute myeloid leukemia, ovarian cancer and breast cancer. Ibrahim *et al.* [20] found features in genes and microRNA using a DBN with an active learning approach. Various cancer diseases such as hepatocellular carcinoma, breast cancer and lung cancer are better classified by this approach. Khademi*et al.* [21] joined a DBN and Bayesian network for microarray feature extraction to detect breast cancer.

Since variable length biological data sets follow a sequential nature, RNNs find application in working with such data. RNNs has been widely used to protein structure prediction, gene expression regulation and protein classification. Park et al.[22] and Lee et al. [23] combined RNNs with LSTM hidden units in the identification of microRNA and target prediction and obtained improved accuracy.

CNNs are used for gene expression regulation problems and are shown to have better capacity in image related classification or recognition problems. Current studies focus on directly using one-dimensional CNNs with biological data sequences [1]. Alipanahiet al. [24] proposed CNN-based approaches for transcription factor binding siteprediction and Kelley et al. [25] used CNN for cell-specific DNA accessibility multitask prediction.

V. LIMITATIONS AND CHALLENGES

Deep learning requires large amount of labelled data, which is unavailable in certain cases. Some part of healthcare data will not be public. Database of some rare diseases will only contain a small set of tuples, which turns into misrepresentation of cases [1]. Learning deep architectures is an extensively resource intensive task. The lack of computational resources

 ISSN: 2454-4248
276 – 280

can cause excessively time-consuming training stages. Deep architectures are often considered to be black boxes [2]. The researcher sometimes unable to find out the reason for good and bad results of network, and unable to modify accordingly. The lack of interpretability causes misclassification issues unsolved. The noise may even fool the network. The preprocessing stage may take a while to find out structured data and optimal parameters. It may even require the involvement of a human expert. This will lengthen the learning procedure and consume much resources. The convergence issues may arise while fitting the model. Overfitting is usually occurred when the number of network parameters is directly proportional to the total number of training samples. The network is able to memorize the training samples but cannot reach in a generalization to previously unseen samples. So misclassification may arise to new samples even if the error calculated by the network is small. The problem is usually avoided using strategies such as dropout. The issue of vanishing gradients may occur while training. Its effect can be reduced by incorporation of other learning strategies. Selection of a proper deep learning architecture and hyperparameters is also a challenge in working with deep learning. Also there is no standard way of measuring the statistical significance, which is a limitation for future result comparisons [2]. Due to the high computational cost, chips for massive parallel processing will be required in order to deal with the increased complexity.

VI. CONCLUSION

Deep learning is a neural network architecture with multiple hidden layers. It can find complex relationship between different attributes present in a data set. The multiple hidden layers are responsible for mapping this relation. Deep learning architectures like Deep Autoencoder, Recurrent Neural Network, Deep Belief Network, Deep Boltzmann Machine and Convolutional Neural Network find their importance and power in different areas of Bioinformatics such as protein structure prediction, gene expression regulation and disease prediction. It can also be used in drug discovery, drug prediction, gene annotation, medical image recognition and health care management. Lack of labeled data, vanishing gradient and overfitting are the major problems faced by deep learning. Active research is taking place in this field to improve the efficiency of deep learning architectures. Bioinformatics can definitely improve the domain of results using deep learning.

REFERENCES

[1] Seonwoo Min, Byunghan Lee and Sungroh Yoon , "Deep learning in bioinformatics " in Briefings in Bioinformatics Advance Access published July 29, 2016.

[2] Daniele Ravi, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang, "Deep Learning for Health Informatics" in IEEE Journal of Biomedical and Health Informatics, Vol. 21, No. 1, January 2017.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[4] K G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504–507, 2006.

[5] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," Neural Comput., vol. 1, no. 2, pp. 270–280, 1989.

[6] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," Neural Comput., vol. 18, no. 7, pp. 1527–1554, 2006.

[7] R. Salakhutdinov and G. E. Hinton, "Deep boltzmann machines." In Proc. Int. Conf. Artif. Intell. Stat., 2009, vol. 1, Art.no.3.

[8] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," Neurocomput., vol. 187, pp. 27–48, 2016.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[10] M. K. Leung, A. Delong, B. Alipanahi, and B. J. Frey, "Machine learning in genomic medicine: A review of computational problems and data sets," Proc. IEEE, vol. 104, no. 1, pp. 176–197, Jan. 2016.

[11] C. Angermueller, T. Parnamaa, L. Parts, and O. Stegle, "Deep learning ¨ for computational biology ," Molecular Syst. Biol., vol. 12, no. 7, 2016, Art. no. 878.

[12] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292(2):195–202.

[13] Ponomarenko JV, Ponomarenko MP, Frolov AS, et al. Conformational and physicochemical DNA features specific for transcription factor binding sites. Bioinformatics 1999;15(7):654–68.

[14] Atchley WR, Zhao J, Fernandes AD, et al. Solving the protein sequence metric problem. Proc Natl Acad Sci USA 2005;102(18):6395–400.

[15] Spencer M, Eickholt J, Cheng J. A deep learning network approach to ab initio protein secondary structure prediction. IEEE/ACM Trans Comput Biol Bioinformat 2015.

[16] Lyons J, Dehzangi A, Heffernan R, et al. Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. J Comput Chem 2014;35(28):2040–6.

[17] Lee T, Yoon S. Boosted categorical restricted boltzmann machine for computational prediction of splice junctions. In: International Conference on Machine Learning, Lille, France, 2015. p. 2483–92.

[18] Chen Y, Li Y, Narayan R, et al. Gene expression inference with deep learning. Bioinformatics 2016;btw074.

[19] Fakoor R, Ladhak F, Nazi A, et al. Using deep learning to enhance cancer diagnosis and classification. In: Proceedings of the International Conference on Machine Learning, 2013.

[20] R. Ibrahim, N. A. Yousri, M. A. Ismail, and N. M. El-Makky, "Multi-level gene/mirna feature selection using deep belief nets and active learning," in Proc. Eng. Med. Biol. Soc., 2014, pp. 3957–3960.

[21] M. Khademi and N. S. Nedialkov, "Probabilistic graphical models and deep belief networks for prognosis of breast cancer," in Proc. IEEE 14th Int. Conf. Mach. Learn. Appl., 2015, pp. 727–732.

[22] Park S, Min S, Choi H-S, et al. deepMiRGene: deep neural network based precursor microRNA prediction. arXiv Preprint arXiv:1605.00017, 2016.

ISSN: 2454-4248
276 – 280

[23] Lee B, Lee T, Na B, et al. DNA-level splice junction prediction using deep recurrent neural networks. arXiv Preprint arXiv:1512.05135, 2015.

[24] Alipanahi B, Delong A, Weirauch MT, et al. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nat Biotechnol 2015.

[25] Kelley DR, Snoek J, Rinn J. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks.