International Journal of Biological Sciences and Applications
2017; 4(6): 82-86
http://www.aascit.org/journal/ijbsa
ISSN: 2375-3811

American Association for Science and Technology

Keywords
Machine Learning,
Deep Learning,
Bioinformatics

Received: July 29, 2017
Accepted: September 6, 2017
Published: October 17, 2017

Progress on Deep Learning in Bioinformatics

Yanqiu Tong[1], Yang Song[2, *]

[1]Department of Humanity, Chongqing Jiaotong University, Chongqing, China
[2]Department of Device, Chongqing Medical University, Chongqing, China

Email address

gridcmp@126.com (Yang Song)
[*]Corresponding author

Citation

Yanqiu Tong, Yang Song. Progress on Deep Learning in Bioinformatics. *International Journal of Biological Sciences and Applications.* Vol. 4, No. 6, 2017, pp. 82-86.

Abstract

With the development of next generation sequencing, transformation of biomedical big data into valuable knowledge has been one of the most important challenges in bioinformatics. The application of deep learning in bioinformatics has gained more attention in both academia and industry field. Deep Learning is a new area of Machine Learning research, which has been introduced with the objective of moving Machine Learning closer to one of its original goals: Artificial Intelligence. The compared research method was used to describe the application of deep learning in bioinformatics from many academic papers. In this paper, three types of deep learning algorithms (deep neural networks, convolutional neural networks, recurrent neural networks) have been introduced in bioinformatics, especially in the domain of omics. The review of this paper can provide valuable insight for researchers to utilize deep learning models in the future of bioinformatics studies.

1. Introduction

In the big data era, it is important to transform those large quantities of data into valuable knowledge [1], especially in the domain of bioinformatics. In order to extract knowledge from big data in bioinformatics field, machine learning has been widely used. Machine learning algorithms use training data set to uncover underlying patterns, build models, and make predictions based on the best fit model. Indeed, some well-known machine algorithms (*i.e.,* Support Vector Machines, Random Forests, Hidden Markov Models, Bayesian Networks, Monte-Carlo Simulation) have been applied in genomics, proteomics, systems biology and so on [2].

Deep learning is a branch of machine learning and it has got more focus by academic interest rapidly. Bioinformatics can also benefit from deep learning.

Machine learning brings humans and machines closer by enabling humans to teach machines. Machines learn by processing a valid training set that contains the features necessary to tune an algorithm. Machine algorithms focused on finding patterns in data and using these patterns to make predictions.

Machine learning provided more viable solutions with the capability to improve through experience and data. Although machine learning can extract patterns from data, there are limitations in raw data processing, which is highly dependent on hand-designed features. To advance from hand-designed to data-driven features, deep learning has shown great advantages. Furthermore, deep learning can learn complex patterns by combining simpler rules learned from data. Actually, deep learning is a style of neural networks with multiple nonlinear layers that referred to as deep learning architectures, hierarchical representations of data can be discovered with increasing levels of abstraction [3].

2. Deep Learning

Deep learning is a rapidly growing research area, and a plenty of new deep learning architecture is being widely used in bioinformatics. Deep learning architectures are basically artificial neural networks of multiple nonlinear layers and several types have been proposed according to input data characteristics and research objectives.

In this paper, three types of deep learning architectures are introduced, such as deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs).

To actually implement deep learning algorithms, a great deal of open source deep learning libraries are available. According to benchmark test results of CNNs, Python-based Neon shows a great advantage in the processing speed. C++ based Caffe and Lua-based Torch offer great advantages in terms of pre-trained models and functional extensionality, respectively. Python-based Theano provides a low-level library to define and optimize mathematical expressions; moreover, numerous higher-level wrappers such as Keras, Lasagne, and Blocks have been developed on top of Theano to provide more intuitive interfaces. Google recently released the C++based TensorFlow with a Python interface [4].

2.1. Deep Neural Networks

The basic structure of DNNs consists of an input layer, multiple hidden layers, and an output layer. Once input data are given to the DNNs, output values are computed sequentially along the layers of the network. At each layer, the input vector comprising the output values of each unit in the layer below is multiplied by the weight vector for each unit in the current layer to produce the weighted sum. Then, a nonlinear function, such as a sigmoid, hyperbolic tangent, or rectified linear unit (ReLU) [5], is applied to the weighted sum to compute the output values of the layer. The computation in each layer transforms the representations in the layer below into slightly more abstract representations.

DNNs are renowned for their suitability in analyzing high-dimensional data. Given that bioinformatics data are typically complex and high-dimensional, DNNs have great promise for bioinformatics research. It should be believing that DNNs as hierarchical representation learning methods, can discover previously unknown highly abstract patterns and correlations to provide insight to better understand the nature of the data. However, it has occurred to us that the capabilities of DNNs have not yet fully been exploited. Although the key characteristic of DNNs is that hierarchical features are learned solely from data, human designed features have often been given as inputs instead of raw data forms. The future progress of DNNs in bioinformatics will come from investigations into proper ways to encode raw data and learn suitable features from them.

2.2. Convolutional Neural Networks

CNNs are designed to process multiple data types, especially two-dimensional images, and are directly inspired by the visual cortex of the brain. In the visual cortex, there is a hierarchy of two basic cell types: simple cells and complex cells [6]. Simple cells react to primitive patterns in sub-regions of visual stimuli, and complex cells synthesize the information from simple cells to identify more intricate forms. Since the visual cortex is such a powerful and natural visual processing system, CNNs are applied to imitate three key ideas: local connectivity, invariance to location, and invariance to local transition.

The basic structure of CNNs consists of convolution layers, nonlinear layers, and pooling layers [7]. To use highly correlated sub-regions of data, groups of local weighted sums, called feature maps, are obtained at each convolution layer by computing convolutions between local patches and weight vectors called filters. Furthermore, since identical patterns can appear regardless of the location in the data, filters are applied repeatedly across the entire dataset, which also improves training efficiency by reducing the number of parameters to learn. Then nonlinear layers increase the nonlinear properties of feature maps. At each pooling layer, maximum or average subsampling of non-overlapping regions in feature maps is performed. This non-overlapping subsampling enables CNNs to handle somewhat different but semantically similar features and thus aggregate local features to identify more complex features.

2.3. Recurrent Neural Networks

RNNs, which are designed to utilize sequential information, have a basic structure with a cyclic connection. Since input data are processed sequentially, recurrent computation is performed in the hidden units where cyclic connection exists. Therefore, past information is implicitly stored in the hidden units called state vectors, and output for the current input is computed considering all previous inputs using these state vectors [4]. Since there are many cases where both past and future inputs affect output for the current input (*e.g.,* in speech recognition), bidirectional recurrent neural networks (BRNNs) [8] have also been designed and used widely.

Although RNNs do not seem to be deep as DNNs or CNNs in terms of the number of layers, they can be regarded as an even deeper structure if unrolled in time. Therefore, for a long time, researchers struggled against vanishing gradient problems while training RNNs, and learning long-term dependency among data was difficult [9]. Fortunately, substituting the simple perceptron hidden units with more complex units such as LSTM [10-11] or GRU, which function as memory cells, significantly helps to prevent the problem. More recently, RNNs have been used successfully in many areas including natural language processing [12-13] and language translation [13-14].

3. Deep Learning in Omics

In omics research, genetic information such as genome,

transcriptome, and proteome data is used to approach problems in bioinformatics. Some of the most common input data in omics are raw biological sequences (*i.e.,* DNA, RNA, amino acid sequences) which have become relatively affordable and easy to obtain with next-generation sequencing technology. In addition, extracted features from sequences such as a position specific scoring matrices (PSSM), physicochemical properties [15-16], Atchley factors, and one-dimensional structural properties [17-18] are often used as inputs for deep learning algorithms to alleviate difficulties from complex biological data and improve results. In addition, protein contact maps, which present distances of amino acid pairs in their three-dimensional structure, and microarray gene expression data are also used according to the characteristics of interest. The topics of interest in omics are divided into four groups. One of the most researched problems is protein structure prediction, which aims to predict the secondary structure or contact map of a protein [19-20]. Gene expression regulation, including splice junctions or RNA binding proteins, and protein classification [21-23], including super family or subcellular localization, are also actively investigated. Furthermore, anomaly classification approaches have been used with omics data to detect cancer.

3.1. Deep neural Networks in Bioinformatics

DNNs have been widely applied in protein structure prediction research. Since complete prediction in three-dimensional space is complex and challenging, several studies have used simpler approaches, such as predicting the secondary structure or torsion angles of protein. For instance, Heffernan et al. applied SAE to protein amino acid sequences to solve prediction problems for secondary structure, torsion angle, and accessible surface area. In another study, Spencer et al. applied DBN to amino acid sequences along with PSSM and Atchley factors to predict protein secondary structure. DNNs have also shown great capabilities in the area of gene expression regulation. For example, Lee et al. utilized DBN in splice junction prediction, a major research avenue in understanding gene expression [24], and proposed a new DBN training method called boosted contrastive divergence for imbalanced data and a new regularization term for sparsity of DNA sequences; their work showed not only significantly improved performance but also the ability to detect subtle non-canonical splicing signals. Moreover, Chen et al. applied MLP to both microarray and RNA-seq expression data to infer expression of up to 21000 target genes from only 1000 landmark genes. In terms of protein classification, Asgari et al. adopted the skip-gram model, a widely known method in natural language processing that can be considered a variant of MLP, and showed that it could effectively learn a distributed representation of biological sequences with general use for many omics applications, including protein family classification. For anomaly classification, Fakoor et al. used principal component analysis (PCA) [4] to reduce the dimensionality of microarray gene expression data and applied SAE to classify various cancers, including acute myeloid leukemia, breast cancer, and ovarian cancer.

3.2. Convolutional Neural Networks in Bioinformatics

Relatively few studies have used CNNs to solve problems involving biological sequences, specifically gene expression regulation problems; nevertheless, those have introduced the strong advantages of CNNs, showing their great promise for future research. First, an initial convolution layer can powerfully capture local sequence patterns and can be considered a motif detector for which PSSMs are solely learned from data instead of hard-coded. The depth of CNNs enables learning more complex patterns and can capture longer motifs, integrate cumulative effects of observed motifs, and eventually learn sophisticated regulatory codes [25]. Moreover, CNNs are suited to exploit the benefits of multitask joint learning. By training CNNs to simultaneously predict closely related factors, features with predictive strengths are more efficiently learned and shared across different tasks.

For example, as an early approach, Denas et al. preprocessed ChIP-seq data into a two- dimensional matrix with the rows as transcription factor activity profiles for each gene and exploited a two-dimensional CNN similar to its use in image processing. Recently, more studies focused on directly using one-dimensional CNNs with biological sequence data. Alipanahi et al. and Kelley et al. proposed CNN-based approaches for transcription factor binding site prediction and 164 cell-specific DNA accessibility multitask prediction, respectively; both groups presented downstream applications for disease-associated genetic variant identification. Furthermore, Zeng et al. performed a systematic exploration of CNN architectures for transcription factor binding site prediction and showed that the number of convolutional filters is more important than the number of layers for motif-based tasks. Zhou et al. developed a CNN-based algorithmic framework, DeepSEA, that performs multitask joint learning of chromatin factors (i.e., transcription factor binding, DNase I sensitivity, histone-mark profile) and prioritizes expression quantitative trait loci and disease-associated genetic variants based on the predictions.

3.3. Recurrent Neural Networks in Bioinformatics

RNNs are expected to be an appropriate deep learning architecture because biological sequences have variable lengths, and their sequential information has great importance. Several studies have applied RNNs to protein structure prediction, gene expression regulation, and protein classification. In early studies, Baldi et al. used BRNNs with perceptron hidden units in protein secondary structure prediction. Thereafter, the improved performance of LSTM hidden units became widely recognized, so Sønderby et al. applied BRNNs with LSTM hidden units and a one-

dimensional convolution layer to learn representations from amino acid sequences and classify the subcellular locations of proteins. Furthermore, Park et al. and Lee et al. exploited RNNs with LSTM hidden units in microRNA identification and target prediction and obtained significantly improved accuracy relative to state-of-the-art approaches demonstrating the high capacity of RNNs to analyze biological sequences.

4. Discussion

A main criticism against deep learning is that it is used as a black-box: even though it produces outstanding results, very little about how such results are derived internally is known. In bioinformatics, particularly in biomedical domains, it is not enough to simply produce good outcomes. Since many studies are connected to patients' health, it is crucial to change the black- box into the white-box providing logical reasoning just as clinicians do for medical treatments.

Transformation of deep learning from the black-box into the white-box is still in the early stages. One of the most widely used approaches is interpretation through visualizing a trained deep learning model. In terms of image input, a deconvolutional network has been proposed to reconstruct and visualize hierarchical representations for a specific input of CNNs. In addition, to visualize a generalized class representative image rather than being dependent on a particular input, gradient ascent optimization in input space through backpropagation-to- input (*cf.* backpropagation-to-weights) has provided another effective methodology. Regarding genomic sequence input, several approaches have been proposed to infer PSSMs from a trained model and to visualize the corresponding motifs with heat maps or sequence logos. For example, Lee et al. extracted motifs by choosing the most class discriminative weight vector among those in the first layer of DBN; DeepBind and DeMo extracted motifs from trained CNNs by counting nucleotide frequencies of positive input subsequences with high activation values and backpropagation-to-input for each feature map, respectively.

Specifically for transcription factor binding site prediction, Alipanahi et al. developed a visualization method, a mutation map, for illustrating the effects of genetic variants on binding scores predicted by CNNs. A mutation map consists of a heat map, which shows how much each mutation alters the binding score, and the input sequence logo, where the height of each base is scaled as the maximum decrease of binding score among all possible mutations.

Moreover, Kelley et al. further complemented the mutation map with a line plot to show the maximum increases as well as the maximum decreases of prediction scores. In addition to interpretation through visualization, attention mechanisms designed to focus explicitly on salient points and the mathematical rationale behind deep learning are being studied.

Incorporation of traditional deep learning architectures is a promising future trend. For instance, joint networks of CNNs and RNNs integrated with attention models have been applied in image captioning, video summarization [26], and image question answering [27]. A few studies toward augmenting the structures of RNNs have been conducted as well. Neural Turing machines [28] and memory networks [29] have adopted addressable external memory in RNNs and shown great results for tasks requiring intricate inferences, such as algorithm learning and complex question answering. Recently, adversarial examples, which degrade performance with small human-imperceptible perturbations, have received increased attention from the machine learning community [30-31]. Since adversarial training of neural networks can result in regularization to provide higher performance, additional studies in this area are expected, including those involving adversarial generative networks and manifold regularized networks [32].

In terms of learning methodology, semi-supervised learning and reinforcement learning are also receiving attention. Semi-supervised learning exploits both unlabeled and labeled data, and a few algorithms have been proposed. For example, ladder networks [33] add skip connections to MLP or CNNs, and simultaneously minimize the sum of supervised and unsupervised cost functions to denoise representations at every level of the model. Reinforcement learning leverages reward outcome signals resulting from actions rather than correctly labeled data. Since reinforcement learning most closely resembles how humans actually learn, this approach has great promise for artificial general intelligence [34]. Currently, its applications are mainly focused on game playing and robotics.

5. Conclusion

In era of big data, bioinformatics has got great development. And deep learning is taking important position for analyzing and computing these data. In this review, three types of deep learning architectures are introduced, such as deep neural networks (DNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs).

Although deep learning holds promise, it is not a silver bullet and cannot provide great results in ad hoc bioinformatics applications. There remain many potential challenges, including limited or imbalanced data, interpretation of deep learning results, and selection of an appropriate architecture and hyper parameters. Furthermore, to fully exploit the capabilities of deep learning, multimodality and acceleration of deep learning require further study.

Funding

This work was supported by the Chongqing Science and Technology Commission (cstc2016jcyjA2034).

References

[1] Manyika J, Chui M, Brown B et al. Big data: The next frontier for innovation, competition, and productivity 2011.

[2] Larrañaga P, Calvo B, Santana R et al. Machine learning in bioinformatics. Briefings in bioinformatics 2006; 7 (1): 86-112.

[3] LeCun Y, Ranzato M. Deep learning tutorial. In: Tutorials in International Conference on Machine Learning (ICML'13). 2013. Citeseer.

[4] Min S, Lee B, Yoon S. Deep learning in bioinformatics [J]. Briefings in Bioinformatics, 2016.

[5] Nair V, Hinton G. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010. p. 807-14.

[6] Hubel DH, Wiesel TN. Receptive fields and functional architecture of monkey striate cortex. The Journal of physiology 1968; 195 (1): 215-43.

[7] Schuster M, Paliwal KK. Bidirectional recurrent neural networks. Signal Processing, IEEE Transactions on 1997; 45 (11): 2673-81.

[8] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. Neural Networks, IEEE Transactions on 1994; 5 (2): 157-66.

[9] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation 1997; 9 (8): 1735-80.

[10] Gers FA, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM. Neural Computation 2000; 12 (10): 2451-71.

[11] Kiros R, Zhu Y, Salakhutdinov RR et al. Skip-thought vectors. In: Advances in neural information processing systems. 2015. p. 3276-84.

[12] Li J, Luong M-T, Jurafsky D. A hierarchical neural autoencoder for paragraphs and documents. arXiv preprint arXiv:1506.01057 2015.

[13] Luong M-T, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 2015.

[14] Cho K, Van Merriënboer B, Gulcehre C et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 2014.

[15] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. Journal of molecular biology 1999; 292 (2): 195-202.

[16] 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.

[17] Branden CI. Introduction to protein structure. Garland Science, 1999.

[18] Richardson JS. The anatomy and taxonomy of protein structure. Advances in protein chemistry 1981; 34: 167-339.

[19] Lena PD, Nagata K, Baldi P. Deep architectures for protein contact map prediction. Bioinformatics 2012; 28 (19): 2449-57.

[20] Baldi P, Pollastri G. The principled design of large-scale recursive neural network architectures-dag-rnns and the protein structure prediction problem. The Journal of Machine Learning Research 2003; 4: 575-602.

[21] Asgari E, Mofrad MR. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. PloS one 2015; 10 (11): e0141287.

[22] Hochreiter S, Heusel M, Obermayer K. Fast model-based protein homology detection without alignment. Bioinformatics 2007; 23 (14): 1728-36.

[23] Sønderby SK, Sønderby CK, Nielsen H et al. Convolutional LSTM Networks for Subcellular Localization of Proteins. arXiv preprint arXiv:1503.01919 2015.

[24] Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. Nature 2010; 463 (7280): 457-63.

[25] Park Y, Kellis M. Deep learning for regulatory genomics. Nature biotechnology 2015; 33 (8): 825-6.

[26] Yao L, Torabi A, Cho K et al. Describing videos by exploiting temporal structure. In: Proceedings of the IEEE International Conference on Computer Vision. 2015. p. 4507-15.

[27] Noh H, Seo PH, Han B. Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction. arXiv preprint arXiv:1511.05756 2015.

[28] Graves A, Wayne G, Danihelka I. Neural turing machines. arXiv preprint arXiv:1410.5401 2014.

[29] Weston J, Chopra S, Bordes A. Memory networks. arXiv preprint arXiv:1410.3916 2014.

[30] Szegedy C, Zaremba W, Sutskever I et al. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 2013.

[31] Goodfellow I, Pouget-Abadie J, Mirza M et al. Generative adversarial nets. In: Advances in neural information processing systems. 2014. p. 2672-80.

[32] Lee T, Choi M, Yoon S. Manifold Regularized Deep Neural Networks using Adversarial Examples. arXiv preprint arXiv:1511.06381 2015.

[33] Rasmus A, Berglund M, Honkala M et al. Semi-Supervised Learning with Ladder Networks. In: Advances in neural information processing systems. 2015. p. 3532-40.

[34] Arel I. Deep Reinforcement Learning as Foundation for Artificial General Intelligence. Theoretical Foundations of Artificial General Intelligence. Springer, 2012, 89-102.