Advanced Techniques in Biology & Medicine

Paliwal et al., Adv Tech Biol Med 2015, 3:3
http://dx.doi.org/10.4172/2379-1764.1000139

Mini Review

Open Access

A Short Review of Deep Learning Neural Networks in Protein Structure Prediction Problems

Kuldip Paliwal*, James Lyons and Rhys Heffernan

Signal Processing Laboratory, School of Engineering, Griffith University, Brisbane, Australia

Abstract

Determining the structure of a protein given its sequence is a challenging problem. Deep learning is a rapidly evolving field which excels at problems where there are complex relationships between input features and desired outputs. Deep Neural Networks have become popular for solving problems in protein science. Various deep neural network architectures have been proposed including deep feed-forward neural networks, recurrent neural networks and more recently neural Turing machines and memory networks. This article provides a short review of deep learning applied to protein prediction problems.

Keywords: Deep neural networks; Recurrent neural networks; Protein structure prediction

Introduction

There is a complex dependency between a protein's sequence and its structure, and determining the structure given a sequence is one of the greatest challenges in computational biology [1]. In recent years Deep Neural Networks (DNNs) and other related deep neural architectures have become popular tools for machine learning; with DNNs currently state-of-the-art in many problem domains including speech recognition, image recognition and natural language processing tasks [2]. Conventional machine learning techniques such as support vector machines, random forests and neural networks with a single hidden layer are limited in the complexity of the functions they can efficiently learn. These methods often require careful design of features so that patterns can be classified. DNNs have recently been shown to outperform these conventional methods in some areas as they are capable of learning intermediate representations, with each layer of the network learning a slightly more abstract representation than the previous layer [3,4]. With enough layers very complex patterns can be learned. This ability to learn features automatically is helpful for proteins, as it is not known what the ideal mid or high level features are for protein structure prediction problems.

DNNs are neural networks with multiple hidden layers (usually more than 2) which can efficiently learn complex mappings between features and labels. The problems that DNNs excel at are those where there may be very complex relationships between the inputs and the labels, and where large amounts of training data are available. These characteristics have made DNNs popular for solving problems in protein science. The basic DNN consists of layers of hidden units connected by trainable weights. The weights are trained using the backpropagation algorithm to minimise the error between the NN output and the true output on a training set. Various architectures have been tailored to specific problems, the simplest architecture is the multi-layer feed-forward neural network, for images convolutional neural networks are used, and for sequence problems recurrent neural networks are used.

Having a large amount of training data is a requirement for DNNs, as more trainable parameters usually requires more training data to reliably learn. When designing neural networks input features need to be normalised, especially when features are heterogeneous. Feature selection is also beneficial for achieving good classification and regression performance [5,6].

Feed-forward DNNs

Many current state-of-the-art protein predictors are based on feed-forward DNNs that use a fixed-width window of amino acids, centered on the predicted residue. The window is moved over the protein so that predictions can be made for each residue.

PSIPRED was an early protein secondary structure predictor based on a neural network with a single hidden layer [7]. PSIPRED achieved accuracies of around 80% when predicting 3 secondary structure elements: helix, coil and sheet. Later predictors include SPINE-X, Scorpion, DNSS and SPIDER-2 which are based on deeper neural networks and increase the secondary structure prediction accuracy to around 82% [8-11]. In addition to 3 state secondary structure, other protein properties have also been predicted using deep neural networks including Accessible Surface Area (ASA), phi and psi angles, theta and tau angles, and disorder prediction [8,11-16].

Other Architectures

In addition to standard feed forward DNN architectures, Recurrent Neural Networks (RNNs) are tailored to sequence prediction problems. RNNs were developed to handle time series of information such as speech signals. These networks can pass information from one time step to the next, so context information contained earlier in the sequence can be utilized later in the sequence. Bidirectional Recurrent Neural Networks (BRNNs) were later introduced to utilize information along the entire sequence [17]. RNNs can be considered to be very deep neural nets since information may potentially be passed through many time steps. Early RNNs had problems learning when they were required to remember information over long time periods. Long Short Term Memory (LSTM) RNNs were proposed to circumvent these problems and have become widely used for sequence prediction tasks [18].

***Corresponding author:** Kuldip Paliwal, Griffith School of Engineering, Griffith University, Brisbane, QLD 4111, Australia, Tel: +61-7-3735 6536; Fax: +61-7-3735 5198; E-mail: K.Paliwal@griffith.edu.au

Received September 09, 2015; **Accepted** September 17, 2015; **Published** September 24, 2015

Citation: Paliwal K, Lyons J, Heffernan R (2015) A Short Review of Deep Learning Neural Networks in Protein Structure Prediction Problems. Adv Tech Biol Med 3: 139. doi: 10.4172/2379-1764.1000139

RNNs have been applied to secondary structure prediction with some success [19-25]. The recurrent connections in the RNN remove the need for large context windows, as the surrounding context is provided by the network. RNNs have also been applied to protein disorder prediction [26]. Basic RNNs can handle arbitrary length 1-dimensional input and output sequences, but they can be modified to handle arbitrarily sized 2-dimensional (and higher) inputs and outputs. These 2-D RNNs have been applied to protein contact map prediction in which a prediction is made for every pair of residues in a protein, as well as prediction of disulfide bridges [27-32]. The latest area of research in neural network architectures is towards adding memory to RNNs for so called neural Turing machines and memory networks [33-35]. These networks can be trained to solve problems that basic RNNs are incapable of solving, e.g., given examples of sorted and unsorted data, learn to sort new unseen data. These architectures have not yet been applied to protein prediction problems and it remains to be seen whether they will be able to succeed where simpler architectures have not.

Conclusion

This article has attempted to give a short non-exhaustive overview of the applications of DNNs to protein structure prediction problems. Deep learning is a rapidly evolving field which excels at problems where there are complex relationships between input features and desired outputs, problems that simpler classifiers are incapable of solving. The main strength of deep learning is the ability to easily take advantage of increases in the amount of data and computational power. One of the catalysts for the success of deep learning for speech and image recognition problems was the emergence of large datasets and sufficient computational power to process them. As more protein data becomes available we hope that deep learning can provide similar improvements to protein structure prediction problems. New deep learning architectures will only accelerate this progress.

References

1. J Cheng, A N Tegge, P Baldi (2008) Machine learning methods for protein structure prediction. Biomedical Engineering, IEEE Reviews 1: 41-49.
2. Schmidhuber J (2015) Deep learning in neural networks: An overview. Neural Netw 61: 85-117.
3. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521: 436-444.
4. Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35: 1798-1828.
5. K Paliwal (1990) Neural net classifiers for robust speech recognition under noisy environments. Acoustics, Speech, and Signal Processing 1: 429–432.
6. K Paliwal (1992) Dimensionality reduction of the enhanced feature set for the hmm-based speech recognizer. Digital Signal Processing 2: 157-173.
7. D T Jones (1999) Protein secondary structure prediction based on position-specific scoring matrices. Journal of molecular biology 292:195-202.
8. E Faraggi, T Zhang, Y Yang, L Kurgan, Y Zhou (2012) Spine x: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. Journal of computational chemistry 33: 259-267.
9. A Yaseen, Y Li (2014) Context-based features enhance protein secondary structure prediction accuracy. Journal of chemical information and mod- eling 54: 992-1002.
10. Spencer M, Eickholt J, Jianlin Cheng (2015) A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction. IEEE/ACM Trans Comput Biol Bioinform 12: 103-112.
11. R Heffernan, K Paliwal, J Lyons, A Dehzangi, A Sharma (2015) Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Scientific reports 5.
12. Qi Y, Oja M, Weston J, Noble WS (2012) A unified multitask architecture for predicting local protein properties. PLoS One 7: e32235.
13. Lyons J, Dehzangi A, Heffernan R, Sharma A, Paliwal K, et al. (2014) Predicting backbone CÎ ± angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. J Comput Chem 35: 2040-2046.
14. J Eickholt, J Cheng (2013) Dndisorder: Predicting protein disorder using boosting and deep networks. BMC bioinformatics 14:88.
15. T Zhang, E Faraggi, B Xue, AK Dunker, VN Uversky (2012) Spine-d: accurate prediction of short and long disordered regions by a single neural-network based method. Journal of Biomolecular Structure and Dynamics 29:799-813.
16. Ward JJ1, McGuffin LJ, Bryson K, Buxton BF, Jones DT (2004) The DISOPRED server for the prediction of protein disorder. Bioinformatics 20: 2138-2139.
17. M Schuster, KK Paliwal (1997) Bidirectional recurrent neural networks. Signal Processing, IEEE Transactions 45:2673-2681.
18. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9: 1735-1780.
19. G Pollastri, D Przybylski, B Rost, P Baldi (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins: Structure, Function, and Bioinformatics 47:228-235.
20. J Chen, NS Chaudhari (2007) Cascaded bidirectional recurrent neural networks for protein secondary structure prediction. Computational Biology and Bioinformatics, IEEE/ACM Transactions 4:572-582.
21. M Agathocleous, G Christodoulou, V Promponas, C Christodoulou,V Vassiliades (2010) Protein secondary structure prediction with bidirectional recurrent neural nets: Can weight updating for each residue enhance performance? Artificial Intelligence Applications and Innovations. Springer 28–137.
22. S Babaei, A Geranmayeh, SA Seyyedsalehi (2010) Protein secondary structure prediction using modular reciprocal bidirectional recurrent neural networks. Computer methods and programs in biomedicine 100: 237-247.
23. P Baldi, S Brunak, P Frasconi, G Pollastri, G Soda (2000) Bidirectional iohmms and recurrent neural networks for protein secondary structure prediction. Protein Sequence Analysis in the Genomic Era 1-18.
24. S K Sønderby, O Winther (2014) Protein secondary structure prediction with long short term memory networks. arXiv: 1412.7828.
25. A Daniel (2003) Prediction of Protein Secondary Structure using Long Short-Term Memory. Technical report, Halmstad University 1-35.
26. I Walsh, AJ Martin, T Di Domenico, SC Tosatto (2012) Espritz: Accurate and fast prediction of protein disorder. Bioinformatics 28:503-509.
27. Di Lena P, Nagata K, Baldi P (2012) Deep architectures for protein contact map prediction. Bioinformatics 28: 2449-2457.
28. Eickholt J, Cheng J (2012) Predicting protein residue-residue contacts using deep networks and boosting. Bioinformatics 28: 3066-3072.
29. I Walsh, D Ba`u, AJ Martin, C Mooney, A Vullo (2009) Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. BMC structural biology 9: 1-5.
30. Tegge AN, Wang Z, Eickholt J, Cheng J (2009) NNcon: Improved protein contact map prediction using 2D-recursive neural networks. Nucleic Acids Res 37: W515-518.
31. P Baldi, G. Pollastri (2003) The principled design of large-scale recursive neural network architectures–dag-rnns and the protein structure prediction problem. The Journal of Machine Learning Research 4: 575-602.
32. J Cheng, H Saigo, P Baldi (2006) Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. Proteins: Structure, Function and Bioinformatics 62: 617-629.
33. A Graves, G Wayne, I Danihelka (2014) Neural turing machines. arXiv: 1410.5401.
34. J Weston, A Bordes, S Chopra, T Mikolov (2015) Towards ai-complete question answering: A set of prerequisite toy tasks. arXiv: 1502.05698.
35. J Weston, S Chopra, A Bordes (2014) Memory networks. arXiv: 1410.3916.