Deep learning in bioinformatics

Guillermo Romero Moreno

Abstract—**The field of computational biology has seen an explosion in the amount of available data thanks to significant technological advances in imaging and genomics. On a different field, deep learning has become one of the trendiest branches in machine learning, since it has achieved significantly better results for some classical machine learning problems. This paper aims to give a brief introduction to some techniques employed in deep learning, and then go through the main biological problems to which it is currently being applied, with the hope of giving a broad view of the joint space of both fields.**

Index Terms—**Deep learning, bioinformatics, CNNs, LSTMs, omics.**

I. Introduction

THE field of biology has gone through major changes in the past decades thanks to new technologies that allow measuring and harvesting increasing amounts of data. Despite its obvious potential for pushing forward our knowledge, such data comes with intrinsic difficulties for its analysis; namely (i) **extremely bulky datasets** that require highly efficient algorithms and machines, and long processing times; (ii) **high-dimensionality**, that hinders biologists from getting direct insights from the data distributions, and brings the so-called *curse of dimensionality* to stage; (iii) **complicated interactions** and data dependencies, due to the high complexity of processes such as gene regulatory networks or protein folding and binding; (iv) **heterogeneous and multi-platform resources**, imposing difficulties at integrating data from different sources in the same mathematical model.

In order to carry analysis on the data, biologists traditionally utilised machine learning algorithms and statistical models, performing data simplification (e.g. clustering, dimensionality reduction), prediction, classification, or modelling. They not only offer the intrinsic benefits of their outcome, but also great insight on the underlying bio-chemical processes by inspection of the learned parameters that the models produce. Despite their (sometimes moderate) success, these methods usually depend on key prior information for achieving their goals, usually coming from online records of annotations related to the data at task, which are not always exhaustive enough. They also bring the drawback of requiring expertise in the field for the careful selection of the right features [9].

This scenario is seeing a major improvement with the recent explosion of the set of techniques under the label *deep learning*. Such algorithms differ from *shallow* machine learning methods in the sense that they include more layers of non-linearities, providing stronger modelling power for complex and hierarchical relations in the data. Such property allows them to cope with some of the afore-mentioned problems, as (i) they are able to learn *intermediate key features* from the data on their own, freeing biologists from the tedious process of hand-crafting them; (ii) they are also good at dealing with high-dimensionality; and (iii) their higher generalization properties make them more suitable for integrating heterogeneous sources.

II. Deep learning

Deep learning involves the design and training of artificial neural networks that contain several stacked layers, as compared to shallow learning, where there are one or two at most. These layers are usually uni-directionally connected (*feedforward*), with the information flowing from the lowest layer (input) up, and progressively building higher levels of abstraction at each layer. Although they were theoretically conceived decades ago, their real implementation was not feasible due to unmanageably long training times. Greater computing power and enhanced training methods developed around a decade ago have made their application possible, so the field has consequently experienced a new blooming since then.

The standard Deep Neural Network (DNN) architecture is the *multi-layer perceptron* (MLP), which is a feed-forward, dense (fully connected) architecture. The basic unit, the neuron, performs a weighted sum of the input connections (and an independent bias term) and applies a non-linear operation to it, generating the output. Typical non-linear functions are the sigmoid function, the hyperbolic tangent, or a rectified linear function (*ReLU*). They are good at performing both regression and classification tasks on the input.

A. Convolutional Neural Networks

Convolutional Neural Networks (CNN) are one of the most popular deep learning architectures nowadays. They provide especially good results over other methods in image recognition and natural language processing. Among their strengths, they are able to detect features independently of their location in the input vector, and the reduced number of weights per layer favours the possibility of making the architecture remarkably deep (even more than 100 layers [5]).

Instead of having fully-connected layers of neurons, CNNs perform the transformation of the data by sliding a filter (convolutional operator) over the output of the previous layer. Several convolutions may be performed in parallel on the same layer, and different layers are usually interleaved with pooling operations (average or max) that reduce the dimensionality of the data and help smoothing out local deformations.

In the field of bio-informatics, CNNs are useful for image processing. Another interesting application is to sequence data (proteins, DNA), where they are able to recognise specific motifs with independence of their location [6].

G. Romero Moreno is with the Department of Computer Science, University of Southampton, UK e-mail: grm1g17@soton.ac.uk.

B. Recurrent Neural Networks

Recurrent Neural Networks (RNN) are another sub-type of deep learning architecture, characterized by units whose connections are not only forward to further layers, but also to themselves, having 'memory' of their previous state. They are not deep because of strictly deep architectures, but more in the sense that an input flows many times through the network before vanishing. Such property renders them ideal for time series and sequential data in general.

Long-Short Term Memory (LSTM) is a special kind of RNN in which single neurons are substituted by LSTM blocks, which include internal structure for controlling which inputs are retained longer and which are hastily forgotten. This property allows them to keep information through many time steps unmodified (long-term memory), overcoming the basic RNN problem of vanishing or exploding signals caused by feedback dynamics. As a main drawback, the high amount of weights contained in each LSTM unit leads to high training times, as compared to other methods.

In biology, RNNs are particularly suitable for sequential data, such as amino-acid chains in proteins or genetic code. They can be used for predicting certain properties after going through a sequence (many-to-one), or structural tagging of the elements of the sequence (many-to-many) [6].

C. Unsupervised learning

All deep learning methods mentioned so far belong to the *supervised learning* category, since they are provided with examples of desired outputs along with their corresponding inputs in the training stage, and must then generalize and provide output for new, unseen inputs. Unsupervised learning methods, on the contrary, are not provided with any output reference, their goal being to find patterns and make sense of the underlying structure of the data. They are useful for extracting high level features that can be then fed to other machine learning models for prediction, classifiation, etc, or to biologists, who can get insights from them.

Common deep unsupervised learning methods are *stacked auto-encoders* and *Deep Belief Networks* (DBFs). They try to squash the high-dimensional inputs into increasingly lower-dimensional layers, with the goal of minimizing the reconstruction error of the original input from the deepest layer. In this way, they can achieve high-level abstract representations of the input space.

D. Drawbacks of deep-learning

In spite of the continuous development of programming frameworks and tools that bring deep learning methods closer to the broader public, their high complexity still hinders them from becoming widely used. Notoriously long training times are only partially mitigated by increasing computer power and the use of GPUs (*Graphic Processing Units*). Other impediments include the need of huge amounts of data for some of the methods.

Another important drawback in deep learning is the difficult interpretation of the systems, as the information they contain is distributed among the neurons. They are said to act as a 'black box'. This property is counter-productive because (i) it does not help experts understand the underlying biological processes, (ii) it helps masking spurious results (e.g. overfitting) and (iii) renders debugging difficult.

III. Biological applications of deep learning

A. Genomics

The DNA encodes the information about all the proteins that the cells need, alongside with purely regulatory sections, or parts that appear to have no purpose. New techniques in genome sequencing, such as *high-throughput sequencing (HTS)*, have provided unprecedented quantities of genomic data at ever-lower costs. The availability of huge amounts of data makes deep learning an appropriate tool for helping in the task.

Due to the sequential nature of the data, RNNs are especially suited for structural annotations of the DNA chain. CNNs can also be of help at detecting specific motifs. Successful applications of deep learning in the field of genomics include the detection of non-coding regulatory fragments [16], or protein binding site prediction [1]. Deep learning is also useful for *metagenomics*, i.e. the analysis of microbial genome from specific environments, where Ditzler *et al.* showed that in spite of having similar accuracy than shallow MLPs, it provided useful hierarchical representations of the data [3].

B. Transcriptomics

Transciptomics refers to the study of the processes through which specific DNA sequences are copied into different types of RNA strands: *messenger RNA* (mRNA), *long non-coding RNA* (lncRNA), and *microRNA* (miRNA). Deep learning techniques have been applied for predicting splicing sites in mRNA (5% AUC improvement [7]), monitoring gene expression on images (5% AUC improvement with CNNs [14]), classifying lncRNA (3,4% higher accuracy [4]), or identifying *expression quantitative trait loci (eQTL)*(with few points improvement in AUC over baseline methods [13]).

C. Proteomics

Proteomics is the large-scale study of proteins, usually looking at the scope of the entire collection of proteins that an organism produces. Although the data from this field is still not enough for consistent use of deep learning methods, some unsupervised techniques have helped building hierarchical representations of the interaction of protein networks [2].

D. Structural biology

Structural biology applications include protein folding and drug design. Proteins are the basic building unit of cells. They are formed by chains of amino-acids that fold into different shapes, which will determine the function that the protein performs. Modifications on the structure can alter their dynamics and influence the appearing of the diseases. Techniques for 3D measuring of their shapes are hard and costly, so it has only been done in very small sample sets. Deep learning can be used as an alternative for predicting the structure purely from the protein sequence. An intermediate typical phase consists on the prediction of the *secondary structure*, which is whether the specific sections of the protein assemble into an α-helix or a β-sheet (a few points improvement over previous methods [6]). This annotation task is particularly well suited for RNNs, as explained before.

Other problem addressed by deep-learning include the classification of intrinsic disordered proteins (IDPs) [12], or predicting binding behaviour for RNA-binding proteins (few points AUC improvement over baseline methods [15]).

E. Multiomics

Multiomics refers to the combination of data and problems for the different omics fields, which can lead to more powerful models as they include interactions between the different processes of regulatory networks. Deep learning is especially well suited for performing this sort of study thanks to its capability of assimilating heterogeneous data and learning highly complex interactions. One particular example of this use can be found in epigenomics, which are the processes by which gene expression is changing without any modification of the DNA (5% to 10% improvement in AUC over previous methods [16]). This requires the inclusion of data from genes, RNA, methylation, and histone modifications into the same model. Another example is the detection and clustering of cancer diseases from multiple source of data [8].

IV. Tools and datasets

With the recent booming of deep learning, there has also been a steady growth of programming frameworks that dramatically simplify the design and development of deep learning tools. With no single framework dominant over the others, the choice can be made based on the programming language that serves as interface or specific characteristics that each of them has. Popular frameworks nowadays are *Caffe*, *Theano*, *Torch*, and *Tensorflow*. A list of them along with their characteristics can be found in [11]. Open-sourcing the code is a common practice in bio-medical research, and therefore there are a lot of examples available on public repositories such as *GitHub*.

While briefly explaining the different application of deep learning in bio-informatics in Section 2, some possible sources of data has already been put forward. A more exhaustive collection made by Mamoshima *et al.* [10] is shown in Figure 1.

V. Conclusion

Deep learning has already been applied to some computational biology problems, with better results than conventional machine learning techniques. Both the increasing availability of data and the appearance of libraries that simplify the implementation of algorithms will boost further its usage across the field, and serve as a primal tool for better understanding the complex processes that occur at a cellular level.

Although deep learning techniques can be hard to interpret and may not be the preferred option in cases where baseline machine learning techniques still provide a fair performance, there are still many potential problems that will get benefited from the application of these novel methods.

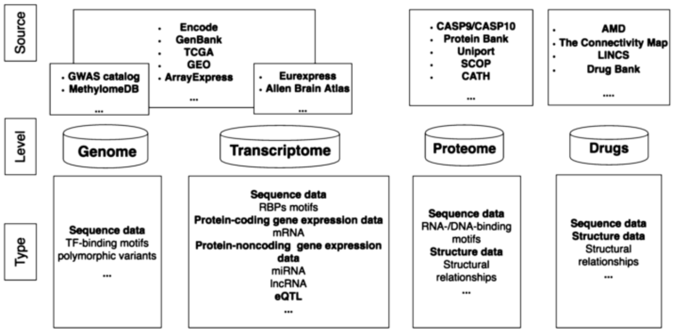

Fig. 1. Publicly available datasets of biological data that can be fed into deep learning models, by Mamoshima *et al.* [10]

References

[1] Babak Alipanahi, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, 2015.

[2] Lujia Chen, Chunhui Cai, Vicky Chen, and Xinghua Lu. Trans-species learning of cellular signaling systems with bimodal deep belief networks. *Bioinformatics*, 31(18):3008–3015, 2015.

[3] Gregory Ditzler, Robi Polikar, and Gail Rosen. Multi-Layer and Recursive Neural Networks for Metagenomic Classification. *IEEE Transactions on Nanobioscience*, 2015.

[4] X.-N. Fan and S.-W. Zhang. LncRNA-MFDL: Identification of human long non-coding RNAs by fusing multiple features and using deep learning. *Molecular BioSystems*, 11(3):892–897, 2015.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. dec 2015.

[6] Vanessa Isabell Jurtz, Alexander Rosenberg Johansen, Morten Nielsen, Jose Juan Almagro Armenteros, Henrik Nielsen, Casper Kaae Sønderby, Ole Winther, and Søren Kaae Sønderby. An introduction to deep learning on biological sequence data: Examples and solutions. *Bioinformatics*, 33(22):3685–3690, 2017.

[7] Michael K.K. Leung, Hui Yuan Xiong, Leo J. Lee, and Brendan J. Frey. Deep learning of the tissue-regulated splicing code. *Bioinformatics*, 30(12), 2014.

[8] Muxuan Liang, Zhizhong Li, Ting Chen, and Jianyang Zeng. Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2015.

[9] Maxwell W Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nat Rev Genet*, 16(6):321–332, 2015.

[10] Polina Mamoshina, Armando Vieira, Evgeny Putin, and Alex Zhavoronkov. Applications of Deep Learning in Biomedicine, 2016.

[11] Daniele Ravi, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang Zhong Yang. Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1):4–21, 2017.

[12] Sheng Wang, Shunyan Weng, Jianzhu Ma, and Qingming Tang. DeepCNF-D: Predicting protein order/disorder regions by weighted deep convolutional neural fields. *International Journal of Molecular Sciences*, 2015.

[13] M J Witteveen. *Identification and Elucidation of Expression Quantitative Trait Loci (eQTL) and their regulating mechanisms using Decodive Deep Learning*. Master thesis, TU Delf, 2014.

[14] Tao Zeng, Rongjian Li, Ravi Mukkamala, Jieping Ye, and Shuiwang Ji. Deep convolutional neural networks for annotating gene expression patterns in the mouse brain. *BMC Bioinformatics*, 2015.

[15] Sai Zhang, Jingtian Zhou, Hailin Hu, Haipeng Gong, Ligong Chen, Chao Cheng, and Jianyang Zeng. A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Research*, 44(4), 2015.

[16] Jian Zhou and Olga G. Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10):931–934, 2015.