Opinion

Deep learning in biological data analysis

Volume 5 Issue 1 - 2017

Jingyu Guo
Department of Genetics, University of Alabama at Birmingham, USA

Correspondence: Jingyu Guo, Department of Genetics, University of Alabama at Birmingham, USA, Email guo.jingyu@gmail.com

Received: January 22, 2017 | **Published:** February 02, 2017

The big data challenge and deep learning

Machine learning represents a set of computational algorithms to learn to identify patterns from data. These patterns, in other words, represent the extracted information or knowledge embedded in the data. In the past years, machine learning, as the subfield of artificial intelligence, faced a great opportunity that has never foreseen as we entered the age of big data. New data had emerged at a speed that cannot be processed in real time with current computation resources possessed. Such volume of data enabled machine learning practices that cannot be done before when big data is not available--one of such machine learning practices is the deep neural network, or more commonly referred as deep learning. Deep learning emerges also because of the engineer advancements. Over recent years, new engineer advancements, such as Amazon Web Service as the representative Cloud Computing platform, Map Reduce as the representative distributed computing architecture, Hadoop as the representative distributed data processing software suite, had enable the big data handling. Specifically, the hardware improvements on Graphic Processor (GPU), had enable much faster parallelized optimization in deep neural network training.

In biological research, such trend of using deep learning had been observed over recent years as well.[1–4] This partly is helped by data volume increase and hardware accessibility as described above, but specially for biological fields, the complexity of the problems require more "expressive" methods, i.e. the capability of describing complex patterns or models. Traditional neural networks (NN), known as "shallow" models, is normally under 3 layers of artificial neuron units, which can be seen as a set of parameterized functions. Deep Learning is essentially a multiple layers of artificial neural network, which can be distinguished easily by the number of its layers of the artificial neuron units. Most deep neural networks have 7 or more layers. One of such example is the Microsoft *ResNets*, which won the 1st place in Image Net tests on the ILSVRC 2015 and used over 150 layers.[5] NN is known for its expressivity, but yet deep neural networks had showed incredible power in learning complex tasks that cannot be done with a shallow one.

Deep learning in biological research

The intention of this section is not to explain the intriguing architecture or intricate training details of deep learning methods. There would be plenty of research papers and other technique articles to guide the actual practice. I mean to discuss on what deep learning may be deliver in biological research.

Prediction

In machine learning, prediction is generally done as in supervised learning methods, which is mean to make a prediction based on input features as either classifications with discrete data outcome variables, such as variant pathogenicity predictions (benign, unknown, or pathogenic), or regressions with continuous ranged outcome variable, such as gene expression fold change or body weights.

Deep learning methods in general fall in supervised learning category; despite of it may also perform unsupervised tasks (e.g. feature compression). Prediction with deep learning, if well trained and tuned, may lead to superior results due to deep learning can be used to describe very complex models.

Of course, prediction with complex problem/data is challenging. This could be due to at least two types of difficulty, 1. Model restrictions / model bias; 2. Over fitting. Intuitively, it would make sense to use more data to learn a more complex model and with limited amount but complex data, it would challenge the model for better description to the data without going over fitting, which is generally understood as learned patterns from data noises such that the model does not predict well facing new data. Since deep learings are in general very expressive, it is very prone to over fitting problem. However, a set of techniques had been specifically developed for training deep learning models, such as early stop, weight drop out, etc.[6]

Feature learning

With its deep structures, the low (beginning or closer to input) neuron layers of a deep neural network may not directly perform a supervised learning task. In fact, one greatest advantage of deep learning methods are its capability to extract, transform, and synthesize features or attributes from the input data -or more commonly referred as unsupervised features learning, which are most likely to be performed on the low layers of the network. Feature learning is mostly done in a layer-wise manner-the lower layers may focus on a set of more basic primitive features (e.g. such as edges in images, language phrases or motifs), which will then in turn transformed and assembled in to higher level features (e.g. shapes comprised of edges, sentences, or protein structure). This process commonly leads to very good set of features that was not observed in original dataset (latent). Since good features can help very much in every learning task (e.g. good features reduce the downstream discriminative model complexity), feature learning had been a long standing practical issue in machine learning research and deep learning had natively incorporated the feature learning idea into the practice. This should help a lot in biological research where the observations may mostly indirect and learning latent high level features can be a strong advantage for deep learning methods.

In biological research, one intuitive way to process heterogeneous datasets is to mapping them by shared identifiers, such as genes, proteins, or expression vectors etc. This may work well when the data is presented in a table like format, but it potentially leads to two potential issues: first, it may not work with data presented as

other formats such as graph (think about signal regulatory network); second, it simply added more columns as features to the merged data, which requires either more data samples, or some component analysis (PCA) or regularization methods (Lasso) to overcome the potential over fitting issue.

Thus, another advantage that is brought by feature learning is that it may serve as a measure of data merge. This is done by learning a set of new shared features or attributes from the original heterogenous datasets. Mathematically, this can be seen as a way to learn a function to map different features or attributes of datasets into the same vector space, which in turn can be used in downstream analysis such as classification. This type of data merge had been widely used in computer sciences, such as image caption application, in which images and the English descriptions were mapped into the same abstracted vector space.[7] The advantage of doing this is that the similarity can be measured between any visual sample and semantic sample in this vector space, e.g. Euclidean distance or dot product. Then the downstream layers of deep neural network can learn from such similarity information and draw patterns from the transformed datasets.

Importantly, deep neural networks do not represent one type of specific mathematical models such as logistic regression but more as a generic way to learn a model. These networks can be divided into a whole set of types based on their layer connectivity patterns and layer-wise functions, trimmed to suit different tasks. One of the most successful types is called convolution neural network (CNN), such as the ResNets. CNN is mainly used in image related or visual tasks. Other types of deep neural networks included recursive and recurrent neural networks, which may be more suitable in modelling sequential data series. For many other tasks to learn probability distributions, restricted Boltzmann machine (RBN) may help, which also adapts well to both supervised and unsupervised learning tasks. It should come down to the particular learning project to choose the right type. Another important concept and common practice is so called transfer learning, which is to reuse previously trained layers by other projects to serve the feature extraction tasks. It is done so due to the intricate and time-consuming training process of deep neural networks and as long as the basic features are shared between these projects and the one of interest, it makes sense to reuse those establish layers as part of the new neural networks. In practice, many popular deep learning packages had included pertained models for reuse.

Current status of deep learning in biological research

A few studies had shown the successful application of deep learning on different biological problems.

In 2015, Frey et al.,[2] showed a learned model for RNA splicing variants using splicing level data associated with DNA elements from healthy human tissues along with other data.[1] The article was not addressed as methodology focused but it is clear that the deep learning methods was used to extract sequence signatures for features and predict the outcome as a variant score to predict how likely the variant may indicate a splicing event. The highlight for training the model is that the algorithm was not only feed with sequencing data, but also protein interaction and knockdown data. This showed the flexibility and adaptability of deep learning network.

Also in 2015, Mofrad et al.,[3] showed a feature learning article by deep learning on genomics and proteomics sequence data.[2] This paper does not solely focus on the prediction but more as a pertained feature learning method, which is to extract sequence signatures, higher level motifs and structures and embed these learned features as in a vector space. Thus the model can be reused as pre-trained feature extraction layers in other projects requiring sequence relevant features and the author also suggested so.

More recently, a group of Google / Verily Life Sciences researchers published a study on using convolution neural network (CNN) to perform NGS variant calling, which outperformed previous benchmarking methods using parameterized bayesian statistical methods.[3] As discussed above, CNN is very suitable for image relevant tasks. To make this approach work, the researchers creatively used images to represent alignment results from sequence aligner as the input rather than matched strings which were traditionally used. This further strengthened the point that learning from good features is important in machine learning.

What may be shared between these applications are the strong adaptation of the deep learning methods. As stated by No Free Lunch Theorem,[8] no machine learning method could be the better than others on every single tasks. But "all models are wrong, but some are useful", deep learning methods showed the great flexibility in creating complex models, which are potentially intractable by some other methods.

Limitations

There is some limitation to consider when using deep learning methods in research, such as the explainability of the models. Like many complex methods, There is shortage of thorough understanding to explain why certain models works or not. This will rely deeply on the theory breakthrough on the deep learning as well as problem domain knowledge. Explain ability should be better valued in research than just the prediction performance, since research is to discover knowledge through explainable models. Another potential caveat to point out is that deep neural network normally may require a large amount of data to get it properly trained. It can be easily seen as with its great number of layers, the involved parameters (i.e. weights) increased very rapidly, which can lead to over fitting problem without sufficient data.

Still, rooted with computer sciences, statistics and mathematics, machine learning methods can handle much more data than traditional data analysis and now the deep learning had been the jewel in this fields. With dramatically increased data volume in biological fields as well as further development and research in machine learning, deep learning methods will be unleashed with its great power to dig the knowledge deeply embedded in data-driven research.

Acknowledgements

None.

Conflict of interest

The author declares no conflict of interest.

References

1. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet*. 2015;16(6):321–332.
2. Xiong HY, Alipanahi B, Lee LJ, et al. The Human Splicing Code Reveals New Insights into the Genetic Determinants of Disease. *Science*. 2015;347(6218):1254806.

3. Asgari E, Mofrad MR. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS One*. 2015;10(11):e0141287.

4. Poplin R, Newburger D, Dijamco J, et al. Creating a universal SNP and small indel variant caller with deep neural networks. *bioRxiv*. 2016:092890.

5. He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. *Computer Vision and Pattern Recognition*. 2015;arXiv1512.03385.

6. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*. 2012:1097–1105.

7. Karpathy A, Li F. Deep visual–semantic alignments for generating image descriptions. *Computer Vision and Pattern Recognition*. 2015;arXiv.1412.2306v2.

8. Wolpert DH, Macready WG. No Free Lunch Theorems for Optimization. *IEEE Transactions on Evolutionary Computation*. 1997;1(1):67.