Agastya Gupta
Ahsan Azim

Convolutional Neural Networks & other Deep Architectures for DNA Sequence Classification

Introduction / Significance

The aim of our project is to conduct further research into the efficacy of Convolutional Neural Networks (CNNs), and Deep Learning tools in general, as a means for classifying DNA Sequences. Convolutional Neural Networks are a new development in the field of Deep Learning – the latter itself being at the cutting edge of Machine Learning research. There has till now been limited research on the applicability of Convolutional Neural Networks to problems in the Bioinformatics domain; we seek to build on existing progress and possibly cover new ground with respect to this. Our complementary focus on using Deep Learning tools in general helps, and we expect to conduct research into the efficacy of various Deep Learning algorithms for Classification, with an emphasis being on comparison of results and overall applicability to existing problems – possibly with reference to diseases and such.

A final aim of our project is to create a user-friendly web experience to visualize results of the classification, as well as to allow for users to upload custom sequences and see them classified. We also expect the website to be open to the public in general as well.

Aim / Hypothesis

The most immediate aim of our project is to use Convolutional Neural Networks and other Deep Learning algorithms for classification of DNA sequences. Based on findings from *Nguyen et al.*, we notice that CNNs show great potential for this purpose, and we will initially seek to replicate said findings. Note that we shall do so with our own, personal, CNN implementation – i.e. we shall be writing a Convolutional Neural Network from scratch. We will compare the results of our implementation to those of the author, and try to achieve maximum accuracy on our part by virtue of having full control of every possible feature of the algorithm (as we will write it). We shall also compare our results to those achieved by using popular open-source libraries such as Tensor Flow. Another aim of this section of the project will be to then use other classification algorithms from the Deep Learning space, as well as simpler general Machine Learning algorithms, for the purpose of comparing results and efficacy. Our hope is that CNNs will prove effective at classifying the DNA sequences we present, and will hopefully show an improvement over other existing algorithms. Ideally, our custom implementation(s) will have accuracy ratings comparable or better than existing implementations.

Computational approaches

The expected computation approaches centre around Convolutional Neural Networks and other related Deep (and Machine) Learning algorithms. Our foremost and initial emphasis will be on developing a production-quality, custom, CNN. We will simultaneously also use a range of other algorithms for comparison – e.g. Support Vector Machines (SVMs), and Decision Trees. The latter two are more general machine learning algorithms, and it shall be interesting to see how CNNs' performance compares to them. Among other Deep Learning algorithms, Deep Neural Networks stand as possible competitors to CNNs, and we shall also investigate their performance as well.

The other computationally intensive part of our project will be the interface for letting users run classification on custom sequences. This component involves quite a bit of web-specific technology, and we expect to face interesting problems in that domain. As this may not be expressly related to the focus of this class, we shall instead avoid delving into too many details about the web aspect here. Suffice it to say, both our primary focus is on the earlier part of the project – i.e. using CNNs and other tools to tackle the biological problem of DNA sequence classification.

Data to use

Luckily, we have no lack of resources for obtaining data. Although we have selected specific datasets already – they will be discussed below – a great general resource for DNA sequence data has been NCBI's GenBank. Another useful repository has been University of California Irvine's Machine Learning dataset repo. General credits aside, specific datasets we have chosen from the many available are the *E. Coli* promoter gene sequence data set(linked here), and the Primate splice-junction data set (linked here). Both are valid, canonical, data sets that are good benchmarks for testing the efficacy of our algorithms. Once we have established basic effectiveness on those data sets, we may move on to more creative ones, these most likely being obtained from GenBank – it has a vast collection of sequence data, and we shall face no shortage there. Our focus at that point will be on focusing on data that may allow us to tackle a specific disease, or something of that sort.

Expected results

Ideally, the CNN implementation we construct will result in DNA classification accuracy scores similar to those exhibited in our sample paper (*Nguyen et al.*). Note that the full control we have over our implementation will allow us to vary literally everything about our Neural Network. This holds great potential for our purposes – by not relying totally on pre-existing libraries, there is great potential for discovery via customization. When it comes to our investigation into other Deep and general Machine Learning methods for DNA classification, we hope that we perhaps discover that an alternative method may even be better than CNNs. After having focused on DNA classification data, we will then proceed to our, for lack of a better description, "further exploration" phase. Here we shall try and extend our models to data sets beyond DNA classification. The wealth of observations we will have by then collected on varying algorithms' suitability for biological data sets will allow us to proceed with a friendly "plug and play" approach, whereby we can simply input the data set (with some necessary cleaning and all, of course), and see what our algorithms observe. Hopefully interesting findings proceed from there.

Potential pitfalls

A possible challenge we may run into is quite simply the novelty of CNNs. They are a new phenomenon at the cutting-edge of Machine Learning research, and our implementation of one will most likely cover very fresh ground. Similarly, implementing something of this sort from scratch will naturally also run into certain challenges. These two facts represent our main challenges. We are, though, working to mitigate them by reading extensively on the topic and combing through existing papers by luminaries in the field such as Geoffrey Hinton and Yann Le Cunn. We have also been conducting extensive sessions with the TAs who have knowledge relevant to this topic; hopefully all of the above will serve to keep us on track.

References

Nguyen, N.G., Tran, V.A., Ngo, D.L., Phan, D., Lumbanraja, F.R., Faisal, M.R., Abapihi, B., Kubo, M. and Satou, K. (2016) DNA Sequence Classification by Convolutional Neural Network. J. Biomedical Science and Engineering, 9, 280-286.
http://dx.doi.org/10.4236/jbise.2016.95021

S. Min, B. Lee, and S. Yoon. Deep learning in bioinformatics. Briefings in Bioinformatics, in press, 2016.
https://arxiv.org/pdf/1603.06430.pdf

Haoyang Zeng, Matthew D. Edwards, Ge Liu, David K. Gifford; Convolutional neural network architectures for predicting DNA–protein binding. *Bioinformatics* 2016; 32 (12): i121-i127. doi: 10.1093/bioinformatics/btw255
https://academic.oup.com/bioinformatics/article/32/12/i121/2240609/Convolutional-neural-network-architectures-for

Ke Chen Lukasz, A. Kurgan; Neural Networks in Bioinformatics, pp 565-583
http://link.springer.com/referenceworkentry/10.1007%2F978-3-540-92910-9_18

We will add to this list as necessary in the future.