| | |
|---|---|
| **Title** | **Deep learning in bioinformatics** |
| **Student name:** | **Guillermo Romero Moreno** |
| **Supervisor name:** | **Mahesan Niranjan** |

## *Aims/research question and Objectives*

The main aim of this project is building a machine learning algorithm for biological sequential data. Due to the technological advances in imaging and genomics, the field of biology has seen an explosion in the amount of data available. This bulky data, usually high-dimensional, requires a machine learning approach since more simple techniques cannot provide enough insights on the processes. Good problems in which to apply them are structural tagging on genomic data or property predictions from protein sequences.

The secondary aim of this project is to implement convolutional neural networks (CNNs) and recurrent neural network (LSTMs) to sequence data. CNNs have proven to be good network structures for imaging since they are able to detect features independently from where they are located and allow deeper structures due to the small amounts of weights per layer. These advantages can also be applied to 1D sequence data, as they are able to recognize motifs in any location of the sequence. LSTMs are naturally fitted for ordered sequence data, and their long-term memory trait allows them to capture non-local relations in a sequence that would be hard to detect otherwise.

The third aim of this project is to improve a state-of-the-art algorithm that applies any of these techniques. The application of deep learning to biological data is a fairly new trend, only appearing after the consolidated maturity of the field of deep learning in the broad sense. Therefore, current algorithms and research are still tentative, and there is still a lot of room for improvement and experimentation.

The aims listed above will be fulfilled through the following objectives:

- Review state-of-the-art algorithms for a biological sequence data problem (e.g. protein secondary structure prediction, intrinsically disordered proteins classification)
- Design the re-implementation of one of such algorithms
- Download and run the algorithm
- Get familiarised with a deep learning framework (e.g. Theano + Lasagne, Tensorflow + Keras)
- Design a suitable architecture that may combine the use of CNNs and LSTMs with the previous algorithm
- Implement the architecture with the deep learning framework
- Investigate the availability of platforms with super-high computing power
- Test and optimize the algorithm for one of such platforms
- Run the algorithm on the super-computing platforms
- Analyse results and detect the weaknesses of the algorithm
- Design and implement modifications to the algorithm that could lead to an improvement in the results
- Run the modified version on the super-computing platform and collect new results
- Repeat the last 3 steps as long as there is time available
- Prepare the presentation and write the final report

***Summary of proposed research and analysis methodology***

There has been a global trend in the biological research community in releasing biological databases to the public. Mamoshima *et al* [1] provide a catalogue that covers the databases most commonly used among machine learning applications to biological data. Since the data is widely shared, it is easy to use them as a benchmark and compare performances with former algorithms applied to the same problems.

There are different ways to assess the performance of an algorithm. For a typical binary classification class, a common option includes the precision-recall area under the curve or the area under the receiver operating curve. For multi-class problems, calculating the right predictions (accuracy) can be used, as well as the cross-entropy log-loss. The confusion matrix provides a better understanding of where the algorithm is failing and therefore how to improve it.

In machine learning applications, it is really important to keep a portion of the data aside from the process and only use it at the end for assessing the performance. Doing it this way gives more truthful results, as proves that the algorithm can generalize well on unseen data and it is not overfitting the provided training set. A method for obtaining an even more accurate generalization error is cross-validation, which implies doing many different train-test splits, training and assessing a model for each of them, and computing the performance as an average of the one obtained for each split. Cross-validation can also be used on the training set (nested cross-validation), for separating a subset of the training set (the validation set) that can be used for various model decisions, such as tuning hyperparameters or early stopping. The different models trained in the nested splits can then form an ensemble and achieve higher generalization power in the test set of the outer split. Many researchers in the biological field also opt using each train and test datasets coming from different source (after eliminating any existing overlap), thus ensuring the test data is absolutely new.

Regarding the coding practices, the *clean code* standards [2] will be followed, in order to keep a self-explicative, easy-to-maintain code. A version control system (*git*) will be used, as well as automated testing when possible.

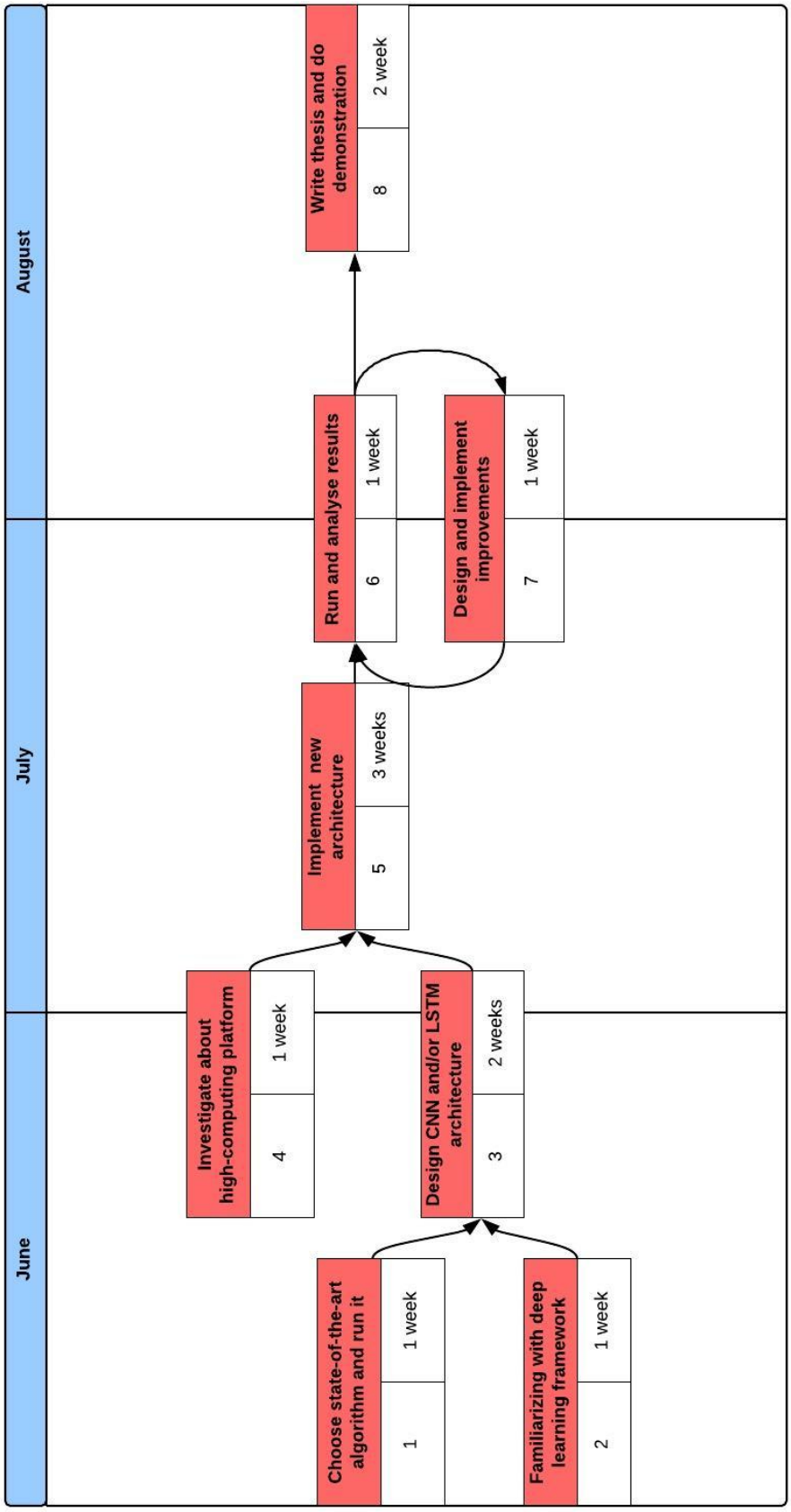
Reference:

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

[2] R. C. Martin. *Clean code*, Prentice Hall; 1 edition (August 11, 2008). **ISBN-10**: 0132350882

***Research plan – Gantt chart or Pert chart***

The PERT chart below shows a rough estimate of the project plan. The run-optimize cycle (steps 6 and 7) allows some flexibility for the critical path, being even possible to skip the step 7 altogether if unexpected issues consume too much time.



***Ethical statement***

This project involves the use of biological data. Although the data that will be handled comes from a long chain of processes and modifications, there should be a concern on whether its recollection was carried out respecting all the ethical standards regarding consent, privacy, and non-aggressive methods.

There should also be concerns about the future utilization of the methods developed in the project, paying close attention to whether they could be abused for obtaining sensitive personal information from individuals through biological data; although it does not seem to be that relevant for many of the tasks considered for this project, as proteins and genes are almost identical to all individuals.

As everything related to the medical field, especially high care should be impressed in obtaining honest and clear results, as they can serve as a base for techniques that are directly applied on medical patients and could have an impact on their well-being.

***Legal and commercial aspects***

This project does not have risks for incurring into law breaching, barring the respect of the usual copyright standards. All external packages, databases, and sources of information must be duly referenced.

There is not strong incentives for commercializing the tools that will be developed in this project, as the only potential customers would be a small, highly-specialized group of researchers that already have access to open-source equivalents.