Project Brief: Understanding Convolutional Neural Networks on Biological Sequences

Author: Guillermo Romero Moreno. Supervisor: Mahesan Niranjan

June 28, 2018

Description of your project question/problem

The project is framed in the field of bio-informatics; more precisely, on the use of deep architectures on biological sequential data, such as the DNA code or the amino acid chains that form proteins. A fundamental problem in bio-informatics is to extract information, in the form of global or local properties, from the sequential data itself, without external measured features, since this is significantly cheaper than obtaining them experimentally.

Recent advances in deep neural networks have beaten all previous methods, with convolutional and recurrent neural networks being the top performers. CNNs are somehow well-understood when applied to images, but no consistent analysis has been performed on their application to sequential data.

The goals for your project, what you are trying to achieve

The goal of the project is to make analysis of a state-of-art, available CNN architecture that has been applied to a biological sequence problem; more specifically, to the protein secondary structure prediction problem. The aim is trying to understand better what the different layers of the network are learning, why they are working well on the problem, and whether any valuable information for biologists can be extracted from the learned features in the network.

The scope of your project

The project will include different methods for evaluating the features learned by a deep neural network, such as feature visualization, activation maximization, sensitivity analysis, and others. It will also explore effective ways in which this information can be represented and easily interpreted by biologists.

All these will be applied on a state-of-the-art architecture whose source code is freely accessible, so the project will not include any new development or re-implementation of a working architecture.