UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES
ELECTRONICS AND COMPUTER SCIENCE

MSc in Artificial Intelligence

by Guillermo Romero Moreno

The relatively new field of deep learning is slowly being transferred to the biomedical field and pushing its state-of-the-art achievements. However, improvements in performance with deep learning machines come with the drawback of opaqueness, which is especially detrimental to biomedical research. A few authors have already started applying deep network interpretability techniques on biological problems to overcome this issue, but none of them has approached problems of the kind of structural tagging, where every element in the sequence has a classification output.

This work aims to develop interpretability techniques for deep networks that have been trained to solve the secondary-structure prediction problem, as it is one of the most studied structural tagging problems. For doing so, a convolutional neural network similar to state-of-the-art was first trained, and then saliency maps were applied to it. Since a single saliency map gets produced per each position (instead of per sequence, as in previous studies), new ways of aggregating them for gathering meaningful insights needed to be construed. The aggregation can occur at two primary levels: on a sequence level, allowing us to explore a single protein sample and spot why specific classifications were made, or on a general level, showing the general relations between input *pssm* and the different classes. These preliminary techniques can be of double value: on one side, they may help biologists to get a better understanding on the underlying protein structural forming process; on the other, machine learning researchers can better understand their machines and spot their previously uncovered flaws.