Agastya Gupta

Ahsan Azim

Convolutional Neural Networks for DNA Sequence Classification

**Aim / Hypothesis**

The aim of our project remains identical to that introduced previously – i.e. exploring the possibilities of Convolutional Neural Networks (CNNs) for DNA Sequence Classification. We continue to move forward with our experimentation into the efficacy of CNNs this task, and have received encouraging results on that front in the past week. Indeed, our prediction that CNNs may prove effective at such classification has already been vindicated – our architecture in some cases surpasses, and other cases matches those of both non-CNN and existing CNN-based efforts at such DNA Sequence classification problems.

Thus the first part of our initial hypothesis – i.e. proving the efficacy of our algorithm on smaller canonical datasets – has been cleared. We shall now move on to extending our research to relatively untouched datasets, as well as conducting further examination into the robustness of our current architecture.

**Computational approaches developed­­­­­­**

As indicated by our project’s aim, development efforts have focused on Convolutional Neural Networks. This has involved tasks such as acquiring relevant theoretical domain knowledge required to understand CNNs, and conducting experimentation into how to apply such knowledge. As our goal was to both implement our own CNN, *and* use external libraries to conduct novel research, we were faced with the opposite tasks of fully understanding a CNN’s internal working, as well as gaining “street smarts” relevant to knowing what works in the wild. This has allowed us to side-by-side develop a CNN of our own, and also use existing implementations with the goal of obtaining research-worthy accuracies and findings.

When it comes to the *how* of developing a CNN and understanding how to use such structures well for real-world analysis, we perused papers by luminaries in the field such as LeCunn (the pioneer of CNNs) and Andrew Ng. We also read widely on prior successful CNN classification jobs – most of which reside in the field of Computer Vision. We thus had to transfer a computer vision approach to handling sequence data, and were led to literature on using CNNs for text classification. This proved key to bridging the divide, as treating the classification of sequence data as another form of text classification allowed us to use CNNs in this domain. We took measures to retain the biological integrity of the data, since clearly human sentences are structured and interpreted far differently from DNA sequences. Some subtle matrix operations – e.g. one-hot vectors using a dictionary of all possible k-mer DNA sequence permutations – formed the core of such a synthetic approach.

Till yesterday we had been using a CNN architecture with two 2D Convolution layers, a 2D Max Pooling layer, Dropout layers, a Fully Connected Layer, and obviously and output layer. Such an implementation matched the accuracy of our paper of focus (i.e. *Nguyen et al.*).

Taking into consideration the relatively smaller size of the data, we tried a test-run with the identical architecture, barring that this time there was only one Convolutional Layer (we feared that the previous presence of two layers was causing overfitting). Our hypothesis proved correct, and this setup surpassed the paper’s accuracy.

**Data used**

Splice Junction etc – describe and give an excerpt

**Current results**

Create table of results -- say we beat their time!!!! Say that we’re consulting with Deep Learning experts at Dartmouth to discuss how this happened haha!

**To do list**

Try even more varying parameters etc, different datasets

Finish off our custom implementation?

Try different ML algorithms – though idk how useful, since current results so good

**Expected results**

Random stuff about increasing/decreasing performance, etc

**References (not included in 3-page limit)**

Not many