DeMo Dashboard: Paper Review

Nipun Agarwala, Oliver Bear Don't Walk, David Cohn, Yuki Inoue, Axel Sly

October 26, 2016

1 Background

In the past few years, there has been exponential growth in the use of deep learning models. Many disciplines have employed deep neural networks (DNN), such as computer vision and natural language processing, and have produced robust results. In this paper by Lanchantin et al., a variety of models are used to predict transcription binding sites (TFBS), which constitutes an important task in bioinformatics. Furthermore, the authors note that, in the context of computational genomics, simply making accurate classifications is not enough. Rather, in order to advance the field, interpretability is also required, which is where DNNs fail when compared to traditional machine learning methods. To remedy this pitfall, DNN results are displayed using three types of visualizations in order to allow for better understanding: measuring nucleotide importance with saliency maps, measuring critical sequence positions for the classifier using Temporal Output Scores, and generating class-specific motif patterns with Class Optimization. Models were tested and evaluated on ENCODE TFBS datasets.

The authors consider three different models in evaluating TFBS classification performance, namely (1) CNN, (2) RNN, (3) CNN-RNN. Since motifs can be viewed as the temporal equivalent of spatial patterns in images such as facial features, model (1) is a viable option. This paper uses a standard CNN model with layers containing a convolution, ReLU nonlinearity, and max pooling. In addition, binding motifs can also span a long range of nucleotides, so researchers need a way of capturing lengthy sequence patterns that can potentially be far apart. Model (2) uses an RNN that is able to detect such high dependence patterns. Nevertheless, since RNNs are memory and complexity inefficient, it is standard to use an LSTM model that has very similar behavior. Specifically, this paper implements a bi-directional LSTM, since there is no innate direction in genomic sequences. Model (3) is a combination of the two models, where data is sequentially fed into a CNN and then into a bi-directional LSTM. The CNN extracts features from motifs, and the LSTM is used to extract temporal features. This is a great approach, as it reduces the size of the LSTM and benefits from the CNN's ability to recognize local motifs. All three models have novel ideas that take advantage of the structure of the data.

2 Experimental Review

2.1 Biological Review

From a biological perspective, the authors generally constructed a fundamentally sound paper and setup. The reliance on JASPAR motifs as a "gold-standard" in validating the four visualization methodologies is well-founded, as the JASPAR database is widely used for procuring transcription factor binding models. This wide usage is owed to the fact that the JASPAR database maintains the quality of the its motifs thanks to a manual curation process, coupled with the database's accessibility via largest open-source repository of its kind[1]. It is worth noting the availability of JASPAR data set, as other major transcription factor binding databases, such as TRANSFAC, are not available without a commercial license.

Furthermore, in justifying their dashboard display in Figure 2, the authors selected three identifiable transcription factors (GATA1, MAFK, NFYB) to the biological community that are in-

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965086/

volved in erythroid development[2], globin expression[3] and serve as a conserved TF with a consistent CCAAT motif[4], respectively. The use of an extensive number of large ENCODE datasets affords the authors the ample data that is necessary to train a nonlinear model such as DNN. Nevertheless, the division between training (30,819 sequences) and test sets (1,000) among the 108 datasets is significantly unbalanced, as only 3% of the available data is used to evaluate performance.

2.2 Computational Review

The TFBS performance of the DNNs is measured using an Area under an ROC curve (AUC of ROC). This statistic can be misleading, due to its weighting of sensitivity and specificity as being equally important.[5] Furthermore, the AUC of ROC statistic does not convey information regarding the spatial distribution of the false positive and false negative errors.[6] As such, many researchers have questioned its usefulness as a measure for comparing model performance.

The authors used multiple visualization techniques to understand the spatial focus of the CNN and RNN architectures. They were comprehensive in most of their comparisons in the results, but omitted vital comparisons (explained in the next few paragraphs) which would have helped lead to a better understanding. They did a good job of describing the JASPAR "gold-standard" motifs, as well as explaining the results and visualizations in detail, which helped give the readers an idea about which techniques may be most useful to them. Finally, they tried to make a case, to the reader, as to why CNN-RNN may be the best model to use to visualize spatial and temporal genomic dependencies.

Despite their strong arguments, the authors did not suggest a more holistic methodology or strategy to produce useful visualizations. Instead, they try to show how one architecture is, on average, better for visualizations. This is substantiated by the fact that the authors note that CNNs perform more accurately than CNN-RNNs in Saliency maps, but despite this, claim CNN-RNN to be more effective individually than CNN and RNN on an ensemble of visualizations. Using a CNN and a CNN-RNN on Saliency Maps and Temporal visualizations respectively, and pooling these results together might be more useful, instead of solely using a CNN-RNN model on both visualization schemes. This may give users and researchers an inference that has a higher performance on the same ensemble of visualizations. Thus, the claim that CNN-RNNs are the best architectures in general is overstated.

Additionally, instead of 1-D interpretations only, using a 2-D mechanism like t-SNE might have provided more insight and inference from the visualizations. Even if it did not, giving researchers valid proof that 2-D representations did not work would have made the argument for using Saliency maps, Temporal scores or Class Optimized results stronger.

2.3 Data Presentation

There are two important factors when examining a paper's presentation of data: first, while papers may claim that the values in tables or figures are groundbreaking, it may be hard for a reader to comprehend just how significant those results are. Secondly, a reader may simply not be as impressed by a set of results that a research team feels are noteworthy. In light of the first consideration, the authors' use of p-values in Table 3 supports their assertion that CNN-RNN configuration outperforms the other configurations.

On the other hand, with regards to the second consideration mentioned above, Table 4 could provide support for the notion that the authors are overstating the significance of their visualization DNN model findings. More specifically, all three models appear to have very few matches with the JASPAR motifs, which were previously held up as a "gold-standard." For example, for Class Optimization, all three methods match with JASPAR motifs on only 20-30% of cases. Furthermore, all the three DNN models, across all three visualizations developed by Lanchantin et al.,

[2]https://www.ncbi.nlm.nih.gov/gene/2623
[3]https://www.ncbi.nlm.nih.gov/gene/7975
[4]https://www.ncbi.nlm.nih.gov/gene/4801
[5]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4356897/
[6]http://www2.unil.ch/biomapper/Download/Lobo-GloEcoBioGeo-2007.pdf

match on fewer than 50% of cases with the JASPAR database. Furthermore, the CNN model outperforms the CNN-RNN model for all three applicable visualizations in Table 4, which strengthens an earlier criticism made in Section 2.2 regarding the CNN-RNN model. While the authors try to contextualize these results with regards to the AUC-ROC numbers from Table 2, the authors do not provide an adequate explanation for why the CNN model outperforms the CNN-RNN model for this motif matching task.

As such, the results from Table 4 should provide a caution to the authors when they make comparisons between different architectures and claims regarding how one method performs better than the others; the reader may come away from looking at Table 4 with the notion that all three methods perform poorly, and that the claim regarding the superiority of the CNN-RNN approach is exaggerated.

3 Significance and Novelty

In order to tackle the problem of interpretability, the authors introduce DeMo Dashboard to help train, visualize and understand the classification of genomic sequences as either having a TF binding site or not. Having such a tool is not likely to help biologists analyze their data. This is because the model is implemented via command line and has other dependencies, which can bar those who are most familiar with GUI interfaces. A bioinformatician is not likely to use one of their trained models as it does not provide much freedom in hyperparameter tuning, which might make the models too general. However, the methods and code for visualizing and interpreting classifications can be useful to bioinformaticians wishing to improve upon their work, and as such, is a step in the right direction of uncovering the mystery of DNNs.

4 Comparison to Related Literature

This paper by Lanchantin et al. most closely compares to work in TFBS prediction and visualization by Alipanahi et al.[7] and Quang and Xie.[8] Alipanahi et al. utilize three methods for visualizing the results of their DeepBind algorithm: mutation maps, sequence logos, and motif detector matrices. Mutation maps are heat maps that display the effect of a nucleotide mutation on the likelihood of binding score, with a red color indicating an increase in binding score, while a blue color indicates a decrease in binding score. The mutation maps used by Alipanahi et al. parallel the display for the RNN temporal outputs used by Lanchantin et al.; the temporal output display is also a red/blue encoded heat map that measures the effect of a sub-sequence on the outputted classification. Both Alipanahi et al. and Lanchantin et al. use sequence logos, which are a standard technique in computational biology and bioinformatics for displaying nucleotide conservation in a sequence.

Alipanahi et al., as well as Quang and Xie, both utilize a visualization approach characterized by Lanchantin et al. as "convolutional activations", which converts convolutional kernels into motifs. In Table 4, Lanchantin et al. report achieving more JASPAR motif matches for both the CNN and the CNN-RNN models using their saliency maps, as opposed to using the "convolutional activations" of Alipanahi et al. and Quang and Xie. However, it is unclear if the difference in matches between the two approaches is statistically significant; furthermore, as was stated in Section 2.3, all visualization methods matched on fewer than 50% of cases, raising doubts about their performance.

Finally, since the DeMo authors utilized the same data set as Alipanahi et al., they report in Table 2 a comparison between the AUC-ROC values for the single convolutional layer CNN (DeepBind), and the multi-convolutional layer CNN and CNN-RNN models developed by Lanchantin et al. DeepBind and the optimal CNN model by Lanchantin et al. achieved comparable AUC-ROC values, while the DeMo authors report a slightly higher AUC-ROC for their CNN-RNN model. However, interestingly, in Table 3, there is no pairwise t-test p-value confirming the statistical significance of this difference, which contrasts with the p-values reported in comparing models developed by Lanchantin et al. to each other, as well as to MEME-ChIP baseline.

[7] http://www.nature.com/nbt/journal/v33/n8/full/nbt.3300.html
[8] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4914104/