CS230

A Deep Learning Approach for Predicting Function of Non-coding Genomic Variants

Fred Lu[*]
Department of Statistics
Stanford University
fredlu@stanford.edu
Code: https://github.com/fl16180/functional_variants

Abstract

A large variety of single-nucleotide polymorphisms in the genome are associated with specific diseases. Most such genomic variants occur in non-coding DNA sequences and are not directly involved in protein variation. This makes it challenging to understand their function. In this study, we develop a deep learning approach to predict functional variants using epigenetic markers as predictors. Our models outperform previously established benchmarks on the GM12878 lymphoblastoid dataset.

1 Introduction

Functional genomics studies the relationship between specific regions of the genome with actions occurring within cells or tissues such as transcription, using the large quantities of molecular and genetic data produced by modern biotechnology. An important application is in understanding the effects of single-nucleotide polymorphisms (SNPs) in genomic regions associated with disease. Because the majority of such regions identified by genome-wide association studies (GWAS) are located in non-coding sequence regions, predicting the functional consequences of these variants is a challenging problem. Furthermore, the functions linked to these variants are often highly specific to certain cell and tissue contexts (1).

Here we propose a deep learning approach to predict functions of noncoding variants using cell/tissue-type specific epigenetic annotations available from the Encyclopedia of DNA Elements (ENCODE). We evaluate functional scores from dense neural networks with modifications and find that they outperform previously published models.

2 Related work

Over the past few years, a number of approaches toward understanding functional noncoding variants have been made. Sequence models using methods such as support vector machines (2) and deep learning (3) have shown the ability to predict many impacts of noncoding variants including chromatin effects and DNase I sensitivity. More specific to the current problem, recent attention has focused predicting cell- and tissue-specific functional effects of the noncoding variants (1; 4). In particular, He et al. utilized site-specific epigenetic information as predictors in a semi-supervised elastic net model (1).

[*] Project advised by Dr. Zihuai He, Dept. of Neurology

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

3 Dataset and Features

3.1 Data Overview

The GM12878 lymphoblastoid massively parallel reporter assay (MPRA) contains 693 experimentally confirmed functional genomic variants and 23000 negative variants. These labels were used as the gold standard for this study. The positive variants are distributed across the 22 somatic chromosomes, somewhat proportionally to chromosome length.

The genomic variants in the GM12878 dataset were merged with epigenetic annotations from ENCODE, to produce 1016 features for each non-coding site. The features correspond to eight epigenetic markers assessed in 127 different cells and tissues, and include DNase I hypersensitivity and histone methylations. The epigenetic annotations are continuous non-negative values representing p-value scores for each epigenetic marker. The goal is to predict the label of each variant using these features.

In addition, I extracted 19bp DNA sequences, centered at each variant, in an attempt to build a predictive convolutional sequence model[2].

3.2 Data Setup

The dataset was first randomly split into train and test sets in a 85%-15% ratio. For model validation, the train set was further randomly split into train (80%) - validation (20%) sets. This procedure was repeated three times for each model, and the validation metrics were averaged over the iterations.

The metrics used for evaluation were chosen based on two factors: First, the GM12878 dataset has heavy class imbalance. In addition, the use-case of the models would be to provide scores indicating relative likelihood of function across the genome. Therefore, we are interested in probabilistic outputs rather than prediction accuracy. The metrics chosen are average precision-recall (AUPR) and area under ROC curve (AUROC).

4 Methods

4.1 Benchmarks

Two benchmarks are used to compare against future model development. To contrast my results against the current state-of-the-art, I have obtained benchmark scores from He et al., which uses a semi-supervised elastic net classifier named 'GenoNet' (1).

In addition, I implemented a logistic regression with L_2 regularization, set using 3-fold cross-validation over the training set. This provides a relatively simple baseline to compare the performance of more complex neural network models.

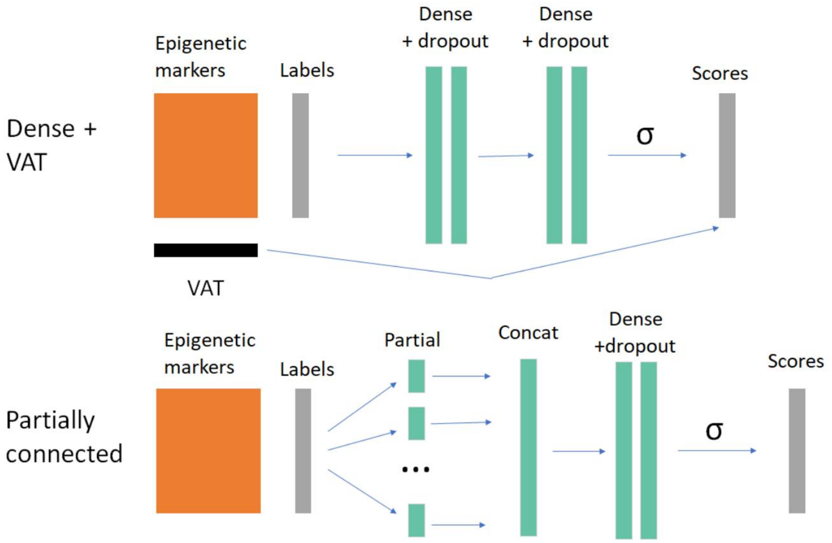
4.2 Neural network models

The following models were implemented:

1. Dense network (**Baseline**): Dense neural network with 2 hidden layers (676 and 36 nodes), L_2 regularization of 1.3e-6, and learning rate 8e-5. The binary crossentropy function was weighted 3x for positive class.
2. Dense network with dropout (**Dense**): The model uses 2 hidden layers (400 and 250 nodes), dropout rates 0.47 and 0.6, L_2 regularization of 1.8e-6, and learning rate 0.00017.
3. Dense network with virtual adversarial training (**Dense+VAT**): This model uses the virtual adversarial training (VAT) method from Miyato et al. (5). This method approximates the adversarial direction of the fitted model using approximation of the Hessian. The model adds a regularization term with respect to current model predictions and the stochastically derived adversarial direction, using the same architecture as the **Dense** model. L_2 regularization was re-tuned to be 6.7e-8.

[2] I one-hot encoded the sequences and attempted 1-D convolutions of varying filter size, but they did not learn the data well.

Figure 1: Model architectures



4. Partially connected network (**PC Net**): The epigenetic features are grouped into 127 cell/tissues. Therefore I constructed a more parsimonious model by routing the 127 signals for each epigenetic marker into separate dense layers with 100 units. Then these units are stacked together into standard dense layers. The intuition is that there is potential redundancy for a given marker between cells/tissues, and this approach forces each partial layer to derive "summary statistics" for each cell/tissue. The model has a dense layer of 180 units, dropout of 0.48, and learning rate of 0.00027.

All models were trained for 25 epochs with a batch size of 256 with the Adam optimizer. Models were implemented in Keras and code is available at `github.com/fl16180/functional_variants`.

Table 1: Average validation and test metrics for each model.

	Avg. validation		Test set	
Model	AUPR	AUROC	AUPR	AUROC
Logistic	0.259	0.764	0.228	0.740
GenoNet	0.251	0.740	0.222	0.728
Baseline	0.243	0.711	0.203	0.699
Dense	0.266	0.761	**0.232**	0.747
Dense+VAT	0.265	0.753	0.226	**0.750**
PC Net	**0.275**	**0.769**	0.228	**0.750**

4.3 Hyperparameter search

Iterative random search was used to develop the model architecture and then to tune hyperparameters. First I allowed all parameters, including number of hidden layers and hidden units, to randomly vary for 300 iterations. Then, based on the results, I narrowed down the architecture options. I found that adding and tuning dropout resulted in a significant improvement in model stability, so I conducted another random search with dropout. Finally, the Dense + VAT and PC Net models were tuned with the same architecture as the Dense model. I randomly varied the VAT ϵ hyperparameter and the partial layer architecture for another 50 iterations and chose the best validation performance as the final models.

Figure 2: t-SNE on dataset, with Dense net results

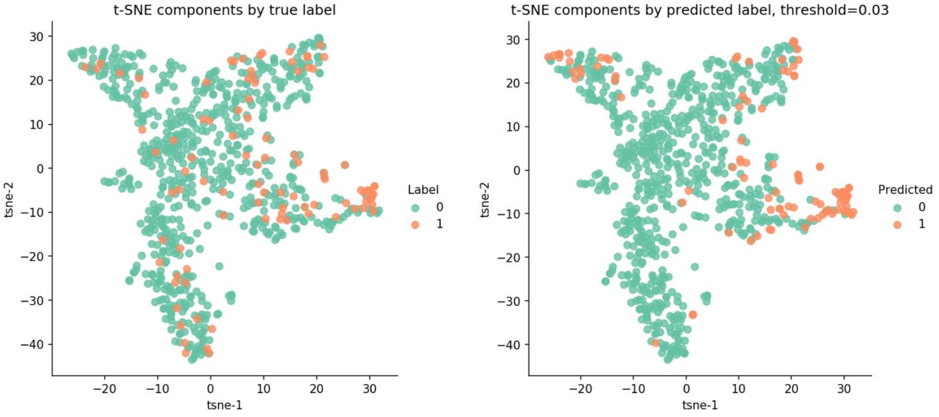
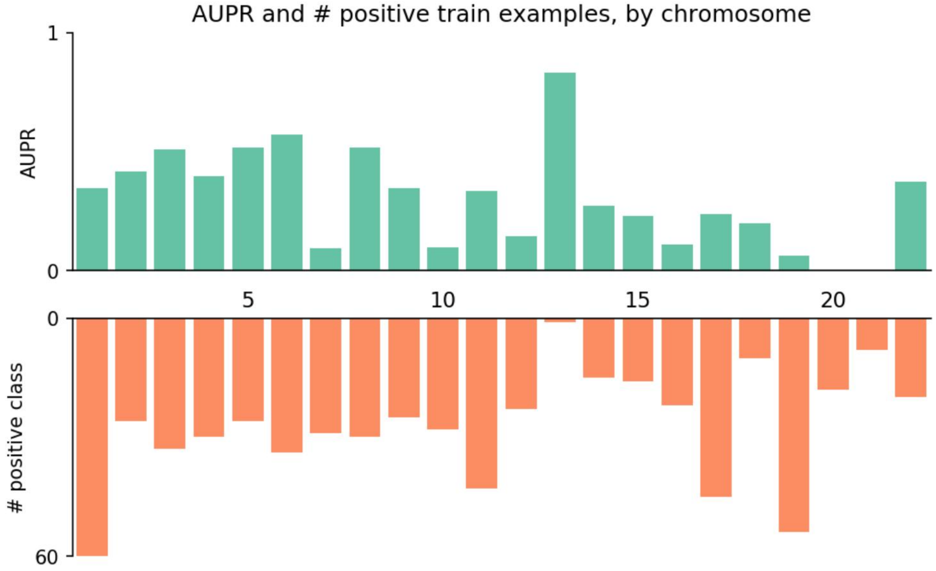


Figure 3:



AUPR was used as the primary optimization metric.

5 Results and Discussion

All the tuned deep learning models with dropout outperformed the benchmarks on the test set, with higher AUPR and AUROC (Table 1). The baseline dense network (without dropout) was very unstable over training and seems to not generalize well. Adding modifications to the tuned Dense network did not significantly change results. While the PC Net achieved the best results on average in the validation set, it was not the clear best on the test set, suggesting that the tuning may have led to overfitting. Dense+VAT does not lead to a performance gain over Dense that justifies the additional computation time.

I conducted analysis of the predicted probabilities generated by the Dense network. Figure 2 shows a lower-dimensional representation of the epigenetic data. Functional variants can in some locations be visually separated from the negative variants in the low-dimensional embedding. Comparing the true labels with the predicted labels, we see that the Dense network is successfully able to learn aspects of the underlying data manifold, in particular at high values of tsne-1.

Finally, I wanted to check if there was a significant pattern in AUPR across different chromosomes. According to Figure 3, there are significant differences. However, there are few positive samples (105) in the test set, so there is inherently noise. The number of positive cases to learn from in the training set also varies widely, but there does not appear to be a systematic pattern between the two factors. Therefore, lack of training examples in a given chromosome does not appear to relate to test set performance.

6 Conclusion

The models presented above show that dense networks have the ability to infer function in non-coding genomic variants. The tuned models outperform the benchmarks on the GM 12878 dataset.

In the future, we aim to apply semi-supervised learning with the VAT technique to leverage the manifold shape of epigenetic information from random positions across the genome. We have also identified additional epigenetic features that could be included in the models to potentially improve performance.

Acknowledgements

I would like to acknowledge Dr. Zihuai He for data as well as important insights and information, Cristian Bartolome for advice, and Belle Ngo for help on the poster design.

Some of the computing for this project was performed on the Sherlock cluster. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results.

References

References

[1] Z. He, L. Liu, K. Wang, and I. Ionita-Laza, "A semi-supervised approach for predicting cell-type specific functional consequences of non-coding variation using mpras," *Nature communications*, vol. 9, no. 1, p. 5199, 2018.

[2] D. Lee, D. U. Gorkin, M. Baker, B. J. Strober, A. L. Asoni, A. S. McCallion, and M. A. Beer, "A method to predict the impact of regulatory variants from dna sequence," *Nature genetics*, vol. 47, no. 8, p. 955, 2015.

[3] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning–based sequence model," *Nature methods*, vol. 12, no. 10, p. 931, 2015.

[4] D. Backenroth, Z. He, K. Kiryluk, V. Boeva, L. Pethukova, E. Khurana, A. Christiano, J. D. Buxbaum, and I. Ionita-Laza, "Fun-lda: a latent dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: methods and applications," *The American Journal of Human Genetics*, vol. 102, no. 5, pp. 920–942, 2018.

[5] T. Miyato, S.-i. Maeda, S. Ishii, and M. Koyama, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, 2018.