CS 273B, Fall 2016-2017 Reese Pathak & Anvita Gupta

Paper Review: DeepSEA

Summary

Convolutional neural networks (CNN or convnets) have demonstrated robust performance on many challenging tasks in multiple domains including natural language understanding, computer vision, and genomics, for problems such as motif discovery and regulatory genomics [1]. In "Predicting effects of noncoding variants with deep learningbased sequence model," Zhou and Troyanskaya apply deep CNNs to predicting chromatin features on the basis of sequence alone [2]. Based on the predictions from their deep network, the authors are further able to resolve functional variants from noncoding areas of the genome, with nucleotide-level specificity.

Methodology

DeepSEA uses a three layer convnet to predict epigenetic features, including transcription factor (TF) binding sites, DNase I hypersensitivity sites (DHSs), and histone marks. The authors train DeepSEA on 1000 base pair sequences taken from the human genome, which are labeled with binary targets for 919 chromatin features. Specifically, the authors obtain the labels by binning the genome into 200 base pair fragments and subsequently overlaying the chromatin peak regions observed by the Encyclopedia of DNA Elements (ENCODE) and Roadmap Epigenomics projects. The authors used a threshold of fifty percent overlap within peak regions for positive classification of a bin [3].

This approach yielded a median auROC of 0.958, 0.923, and 0.856 for TF, DHSs, and histone modifications, respectively. The authors note that DeepSEA outperforms what was the previous gold standard: gkm-SVM. In particular, DeepSEA beats gkm-SVM in TF binding site prediction tasks. The median auROC of gkm-SVM was 0.896. The authors are to be commended for successfully applying deep learning to this task.

The authors establish the significance of the model by applying it as part of a novel pipeline to predict functionality of noncoding SNPs. Indeed, Zhou and Troyanska train a regularized, gradient-boosted logistic regression model, capable of predicting functional significance scores of noncoding variants. The authors retrieve their positive examples from the Human Gene Mutation Database (HGMD), the Genome-Wide Repository of Associations between SNPs and Phenotypes (GRASP), and the National Human Genome Research Institute's GWAS Catalog.

This a diverse dataset and it has been widely used in similar studies, we suspect, in large part due to its diversity. The authors additionally assemble a negative set from the 1000 Genomes project. The full training data was made by gauging the difference in scores obtained from DeepSEA on pairs of negative and positive examples from the training set. The data was further augmented by associated evolutionary information on the conservation

of each variant.

Significance & Critique

The authors' usage of CNNs is impressive, especially in the context of classical "shallow" learning methods, such as support vector machines. However, the recent success of DanQ [4], a hybrid convolutional-recurrent LSTM model, on the noncoding DNA function prediction problem implies that the authors' model is limited in its predictive ability. In particular, models inspired by recurrent methods taken from natural language processing are likely more suited to this task. We hypothesize this largely due to the flexibility in reading heterogeneous input that recurrent units in LSTMs provide. This said, although the authors of DanQ noted an improvement over DeepSEA in over ninety percent of their evaluation set, the improvement was generally quite modest (ca. 1-4 percent). This suggests that more work needs to be done to understand the trade-offs in convolutional versus recurrent architectures.

We also comment that applying convnets to this problem is not entirely novel. Indeed, this work serves as a proof of concept that convnets do perform well in the context of genomics, as they are highly-expressive classifiers capable of using non-linear activation function to transform inputs into separable spaces. This has been corroborated by similar models, including DeepBind [5] and Basset [6]. Thus, we would be interested in seeing cross-evaluation metrics, comparing DeepSEA, Basset, and DeepBind with each other.

DeepSEA uses *in silico* mutagenesis, a process of measuring the influence of base pairs to the CNN's kernels by perturbing the input, to find highly predictive nucleotides. However, the authors merely mention that this process was conducted– they do not report any functionally-important nucleotides that were found as a result of this process.

In addition, *in silico* mutagenesis is computationally inefficient process for determining relative feature importance (features being the nucleotides in this case). Researchers have since developed improved techniques for motif discovery and assigning importance scores to nucleotides; one such method is DeepLIFT [7]. Although DeepLIFT was not available at the time DeepSEA was written, the application of DeepLIFT to DeepSEA would be extremely informative for motif discovery. Although the authors of DeepLIFT achieve good accuracy in predicting 919 chromatin features based on sequence, they do not attempt to recover which motifs were most predictive for these chromatin features, or even for different classes of the chromatin features (e.g. which motifs are predictive for DHSs sensitivity, versus TF binding site sensitivity?)

Though DeepSEA boasts impressive performance with its high auROC metric, we comment that the model has a few limitations and areas for extension. In particular, DeepSEA would benefit by reporting the area under the precision-recall curve (auPRC) in addition to the area under the receiver-operating-characteristic curve (auROC). The auPRC would serve as better estimate in this problem since auPRC is less sensitive to class imbalance [8]. Class imbalance is a known issue in genomics and other biological datasets.

Another obvious limitation carried by DeepSEA and similar models is the training of the model on data from the human reference genome, where the observed chromatin profiles are disjoint for the underlying sequence observations.

Remarks

We conclude that DeepSEA demonstrates impressive advancements in predicting a diverse range of chromatin features, ranging from transcription factor binding to histone modifications. Additionally, this paper illustrates the value of using deep networks in larger data pipelines. The authors adroitly use their system for predicting functional noncoding DNA variants. We conclude that DeepSEA's most significant value was in laying the foundation for more advanced deep learning technologies, such as deeper CNN architectures, RNNs, and LSTMs, to advance our understanding of the regulatory code of the genome.

References

[1] Yongjin Park and Manolis Kellis. "Deep learning for regulatory genomics". In: *Nature Biotechnology* 33.8 (2015), pp. 825–826. ISSN: 1087-0156. DOI: 10.1038/nbt.3313. URL: http://www.nature.com/doifinder/10.1038/nbt.3313.

[2] Jian Zhou and Olga G Troyanskaya. "Predicting effects of noncoding variants with deep learning-based sequence model." In: *Nature methods* 12.10 (2015), pp. 931–4. ISSN: 1548-7105. DOI: 10.1038/nmeth.3547. URL: http://dx.doi.org/10.1038/nmeth.3547%7B%5C%7Dnhttp://www.ncbi.nlm.nih.gov/pubmed/26301843.

[3] Roadmap Epigenomics Consortium et al. "Integrative analysis of 111 reference human epigenomes". In: *Nature* 518.7539 (2015), pp. 317–330. ISSN: 0028-0836. DOI: 10.1038/nature14248. URL: http://www.ncbi.nlm.nih.gov/pubmed/25693563.

[4] Daniel Quang and Xiaohui Xie. "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences". In: *bioRxiv* (2015), p. 032821. ISSN: 0305-1048. DOI: 10.1101/032821. URL: http://biorxiv.org/content/early/2015/12/20/032821.abstract.

[5] Babak Alipanahi et al. "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning." In: *Nat Biotechnol* 33.8 (2015), pp. 831–838. ISSN: 1087-0156. DOI: 10.1038/nbt.3300. URL: http://dx.doi.org/10.1038/nbt.3300.

[6] David R. Kelley, Jasper Snoek, and John L. Rinn. "Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks". In: *Genome Research* 26.7 (July 2016), pp. 990–999.

[7] Avanti Shrikumar et al. "Not Just a Black Box: Learning Important Features Through Propagating Activation Differences". In: *CoRR* abs/1605.01713 (2016). URL: http://arxiv.org/abs/1605.01713.

[8] Jesse Davis and Mark Goadrich. "The relationship between Precision-Recall and ROC curves". In: *Proceedings of the 23rd international conference on Machine learning - ICML '06*. Association for Computing Machinery (ACM), 2006. DOI: 10.1145/1143844.1143874. URL: http://dx.doi.org/10.1145/1143844.1143874.