Title

Author

July 5, 2022

Introduction

Over the past decade, Neural Networks have been successfully applied to a wide range of tasks in different domains such as computer vision, microbiology, and natural language processing. ¡Give some examples here¿

While the overall structure and semantics of the data being used are different, the overall algorithm and methodology stay the same. This is also true in the field of genomics where neural networks are trained on data that is obtained experimentally using genomic assays like DNase-seq [7], ATAC-seq [1], and ChIP-seq [6]. Due to the success of these networks in other fields, there has been a shift in which type of architecture to use when looking at this data. For example, sequenced genomic data is represented as a vector of strings, therefore some neural networks treat this as an NLP problem. Others see ¡some more examples¿

This however is an issue as while the data format is similar, the foundation of the data vastly differs. While data collection and understanding of natural language text has a logical process, we are still learning about sequenced genomic data. For example, in genomic assays, the data is highly noisy and possibly mislabeled furthermore a small change or section of the dataset can have a high impact on the overall prediction ¡Cite?¿. Furthermore, if we borrow from computer vision, given an image of a dog, we know which parts are of high importance (i.e ears, tails, and face), and which are not (i.e background, and eye color). This is because we as humans, have an intuitive understanding of dogs, but not genomics. In biology, (cite biochrome), genomic data often include DNA sequences of various widths, which unlike pictures of dogs are not human-readable or verifiable. While there are many issues associated with applying genomic data to machine learning [8], machine learning still holds a varying level of success of (words).

This biology-specific issue of nonintuitive data increases the difficulty for deep neural networks, whereas the lack of high-quality datasets contributes to the reproducibility crisis and makes it more difficult to compare architectures, as they are often only evaluated on a custom dataset (see above). This being said, there has been a wide level of success in using neural networks in genomics. A few examples are Bichrome, and DeepCAPE [2].

Bichrom

Bichrom is a bimodal neural network framework that is used for characterizing the relative contribution of both DNA sequences and cell type-specific chomatin. This neural network is broken into two distinct sub-networks Bichrom$_{SEQ}$ and Bichrom$_{CHR}$. Where Bichrom$_{SEQ}$ used one-hot encoded DNA serves as input while the other used binned normalized tag counts from chromatin experiments.

¡more information¿

DeepCAPE

DeepCAPE is a deep convolutional neural network that is used to predict enhancers via the integration of DNA sequences and DNAase-seq data. This model has the ability to self adapts to different size datasets and consistently outperforms existing methods in the imbalanced classification of cell line-specific enhancers.

¡more information¿

Background

Initial gains in the field of Machine Learning and genomic annotations can be found as far back as 1992 when [5] used a perceptron neural network and applied it for promoter site predictions of E. coli. ¡should we talk more about this¿. From there Neural Networks and ML techniques have stayed close to the field of Genomics [4]. This is in part because ML models can extract complex features from the training data. While these neural networks have shown promise for extracting information from genomic data, the understanding of the data is still incomplete.

Currently, the main contributions towards dissecting the working of neural networks are achieved by mapping the importance of input features towards the model output, this is done through the study of the network's gradients or other strategies similar to decision trees. These networks have been modeled to similar problems in different domains, (Genomic data is very recursive). Due to the nature of the Data, some researchers have looked into applying Recurrent Neural Networks for prediction (Citations). However, due to their long process times and limited capabilities, they have had varying success.

Currently (I only see) Markov Models are the only networks that have been successfully applied on the full genomic sequence. (citations) ¡is this relavent?¿

In this work we study ¡something¿ and measure different models efficiency and accuracy. We will first observe a 1D deep convolutional network trained on ¡data¿. The model output will be defined in Section 1. From there we will look at a simple Polynomial regression and SVM. Our results will be found in ¡the results section¿.

¡something¿

1 Methodology

Data Collection

Data processing

Support Vector Machine (SVM)

SVM classifiers are generated in two steps; training data is "projected" into a high dimensional space, which is usually higher than the input data. From there the algorithm finds a hyperplane in this new feature space with the largest margin of classification differences. It is shown that classification accuracy depends only weakly on the specific projection, provided that the target space is sufficiently high dimensional [3]. In extream cases, it may not possible to find the separating hyperplane even in a very high-dimensional space. This is when a tradeoff is introduced between the size of the separating margin and penalties for every vector which is within the margin.

Model

The neural network we used is a ¡something¿

Model training

Our data was broken up into different genomic datasets as descibes in 1

Results

References

[1] Jason D Buenrostro et al. "ATAC-seq: a method for assaying chromatin accessibility genome-wide". In: *Current protocols in molecular biology* 109.1 (2015), pp. 21–29.

[2] Shengquan Chen et al. "DeepCAPE: a deep convolutional neural network for the accurate prediction of enhancers". In: *Genomics, proteomics & bioinformatics* 19.4 (2021), pp. 565–577.

[3] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.

[4] Gökcen Eraslan et al. "Deep learning: new computational modelling techniques for genomics". In: *Nature Reviews Genetics* 20.7 (2019), pp. 389–403.

[5] Paul B Horton and Minoru Kanehisa. "An assessment of neural network and statistical approaches for prediction of E. coli promoter sites". In: *Nucleic Acids Research* 20.16 (1992), pp. 4331–4338.

[6] Peter J Park. "ChIP–seq: advantages and challenges of a maturing technology". In: *Nature reviews genetics* 10.10 (2009), pp. 669–680.

[7] Lingyun Song and Gregory E Crawford. "DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells". In: *Cold Spring Harbor Protocols* 2010.2 (2010), pdb–prot5384.

[8] Sean Whalen et al. "Navigating the pitfalls of applying machine learning in genomics". In: *Nature Reviews Genetics* 23.3 (2022), pp. 169–181.