Title

Author

October 23, 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 [6], ATAC-seq [1], and ChIP-seq [5]. 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 [7], 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 [4] 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 [3]. 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.

In this work, we do an architecture study of different neural networks and its ability to read in Assays to predict Enhancers; from here we study the accuracy of different assays (i.e., HPEG2, A549, K562, & MCF7) individually based on different networks. We take the networks that performed the best overall and ran it on all Assays to model its overall accuracy.

What else do we need here

1 Methodology

1.1 Data Collection

How did we collect the data?

1.2 Data processing

How did we process the Data

1.3 Model Architecture

When developing our models we took a modular approach with six different network architectures. For each of our models we have an input label of 500 and and output layer of one with a sigmoid activation function before it. All other activation functions are RELU, have an Adam optimizer for the model with a learning rate of 10^{-4} and a Binary Cross Entropy (BCE) loss function. The average data can be found here

Our first model was a Deep Neural Network (DNN), for each network that has a DNN section it consists of 8 layers cutting the number of parameters down by roughly half each time. After implementing the DNN we then created a new neural network that implemented a Convolutional Nerual Network with convolutional kernals of 10, 50 and 100 respectivly. This is because in the preprocessing stage we had window sizes of similar values. After these convolutions we flattened the output and read it into a Deep Neural Network. This CNN was used to allow us to extract spatial information from the data that can be used for better prediction of enhancers.

After implementing the CNN, we moved towards an Long Short Term Memory model with an attached DNN. The LSTM section was 5 layers deep with an output of 30 neuron. The output of this was then fed into a DNN that was the same as the inital DNN. We then combined the CNN and LSTM layers with a DNN in the following model to extract both a temporal and spatial

representation of the data. The last two models we focused on were the DNN to LSTM and CNN to LSTM for similar reasons.

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

[4] 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.

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

[6] 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.

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