Exploring RNA-Binding Protein Interactions Using a Transformer Model

Zain Manasia
Columbia University

COMSW 4762
Spring 2022

Abstract: RNA-binding proteins (RBPs) are important molecules for the function and maintenance of biological functions. This project will experiment with other methods in machine learning to solve this problem of RNA-binding protein interactions, specifically using a transformer model, with the implementation of an encoder of a transformer model to visualize RNA-protein interaction of individual sequences and to display interactions that may be of significant importance and could be used for further future studies. We will also experiment with the attention mechanism of transformer models, specified in research papers, to see if it has feasibility for computational biology research applications. We found that the transformer model underperforms compared to traditional convolutional neural networks that are in use for studying RBP interactions such as iDeepS. The attention mechanism of transformer models offer functionality in visualizing data that is being processed but does not offer much utility in improving transformer performance overall.

I. Introduction

RNA-binding proteins (RBPs) are important molecules for the function and maintenance of biological mechanisms. The field of molecular biology has done extensive research into the mechanisms related to RBPs and binding sites related to these molecules. In the computational biology and genomics field, remarkable work has been done in the last few years in the fields of deep learning, with projects such as DeepBind, where a convolutional neural network has been trained to predict RBP binding preference, and with iDeepS, where two convolutional neural networks are trained to manage the RNA sequence inputs and an LSTM learns the dependency between sequences and structures to help create a system to better predict binding preference and improve predictive performance. Based on these current systems in place, this project will experiment with other methods in machine learning to solve this problem of RNA-binding protein interactions, specifically using a transformer model. Transformer models have been used previously in the field of computational and functional genomics, namely with models such as Seq2Feature, which uses 41 descriptors to characterize nucleotide sequences and conformational properties to take as input to machine learning algorithms. Transformer models are a popular choice in the field of machine learning in the current day, with the rise of their use in natural language processing tasks, and have applied utility in a field such as computational biology and functional genomics. This project will use this technology to use an encoder of a transformer model to visualize RNA-protein interaction of individual sequences and to display interactions that may be of significant importance and could be used for further future studies.

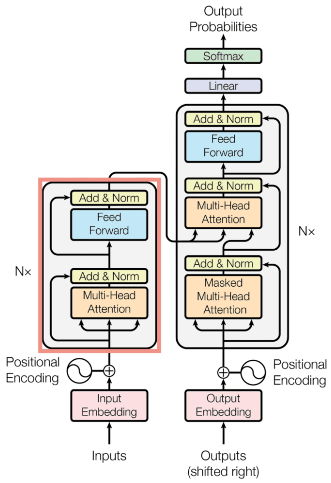

Figure 1: Traditional encoder-decoder architecture of a transformer.

II. Methods

The data used in this project has been collected open source from the RBPDB database. This database is a host to multiple RNA-binding protein datasets and datasets from RBP experimentation. The data used in this project is from their human database content, from the species-specific data that they have collected. Since this is a very large dataset, only close to 10% of this dataset will be used to train and test the model in development for this project. The data is in the form of a CSV file. In the development of the model for the project, the first step is to import the data and visualize the data, which displays that the data has the following columns: sequence, structure, MFE, K, KA, RKA, Qc, and sevenMer. For this project, parameters will be established to process the data such as sequence length so that there is consistency within the processed data. The data will then be split into a training and testing split, following a 66/33 split. Next, an encoder of the transformer model is developed. The data is shaped to fit as input and a multi-head layer is created with 3 dense layers. The last dimension is split into the number of heads and depth of the sequences, where it is transposed to fit a shape containing batch size, number of heads, sequence length, and depth. This transformer encoder for this project follows the self-attention mechanism which implements self-attention by capturing relationships between different elements.

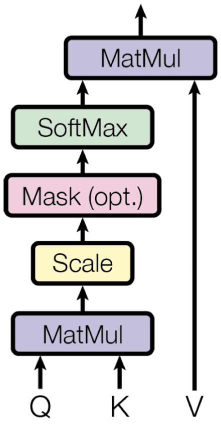

Figure 2: Scaled dot-product attention architecture for transformer

The encoder layer is then specified, which includes a feed-forward network with normalization layers and two dropout laters. The encoder is then specified in the same manner through an embedding layer, positional encoding, the aforementioned encoding layer, and a dropout later. This encoder output gives us the weights necessary to run the model. Our model includes an input layer, an encoder layer, reshape layer, and two dense layers. The parameters for this model are specified such as including a learning rate scheduler, using Adam as the optimizer, defining the loss function, and writing the model function. There are also train and test metrics that need to be specified, which are the R-squared value. Model training can then take place; 20 epochs is the comfortable medium where accuracy drastically increases from the first epoch and loss is negligible. Important verification metrics can then be calculated, such as the overall R-squared and AUC, as these are the metrics used in other papers to assess the model performance.

III. Results

Model performance reached 65.96% accuracy with a loss of 0.0775 by epoch 20. The metrics that were calculated to assess performance of this model against other baselines was the R-squared value. The R-squared value for this model was 0.499 with an AUC of 0.609. Compared to other models that accomplish similar tasks with similar datasets, such as iDeepS, we see that this model slightly underperforms a traditional deep learning model such as iDeepS.

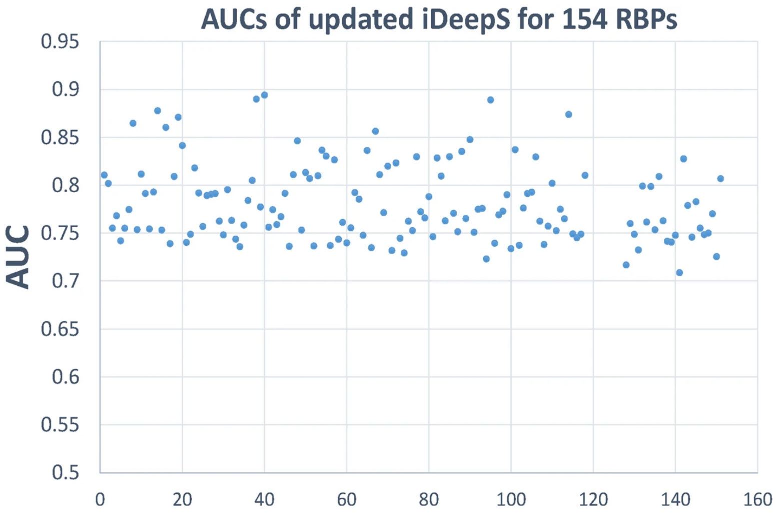

Figure 3: AUC's of a deep learning model for RBP tasks

This is to be expected for a few reasons. We know that traditional convolutional neural networks are very strong at prediction tasks and the amount of data used to gain the results in the iDeepS project is significantly larger than what I have used in this transformer model. Due to the unique nature of the model that I have implemented for this project, we can visualize other aspects that may be of use to researchers who decide to use this attention transformer architecture for computational biology purposes. As stated before, the attention mechanism allows a function, the attention function, to map a query and a set of key-value pairs to an output. The output is computed as a weighted sum of values, where the weight assigned to each value is computed by a compatibility function. This can be used to develop an attention matrix of sorts to visualize connections between input sequences into the model and sequence variation along the entire dataset.

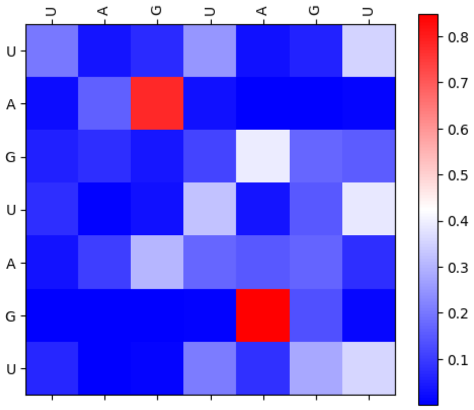

Figure 4: Matrix heatmap to visualize relationships between input sequence and output using the attention transformer mechanism on a sample sequence

IV. Discussion

This transformer implementation to represent RBP interactions is a unique way to predict and visualize this type of data. Although transformer models are increasing in prevalence, models such as the one implemented here are still unable to outperform convolutional neural networks in prediction tasks. This could be due to limitations in model architecture or due to the large amount of data that some of these models have processed for the prediction tasks. The purpose of this implementation was to test accuracy and performance against a standing machine learning method used to analyze RNA-binding protein interactions, but also to test this unique attention mechanism in transformer models. I wanted to see if this method had an efficacy in model performance and or usefulness for visualization. It seems that it can be a useful tool for researcher to visualize patterns and gain insights into the data that they may be working with, but it does offer negligible gain even versus a traditional transformer model accomplishing the same task, and has decreased performance versus a convolutional neural network accomplishing a similar task. In conclusion, this project was an interesting experiment to test a model and architecture from other domains of machine learning and implement it in a useful way for computational biology research. With the progression of more complex methods in machine learning, some may have better efficacy in computational biology than others. For RBP tasks though, traditional methods that exist in the space currently seem to offer the best performance.

Code to run this project can be found at: https://github.com/zainmanasia/ML4FGFinalProject.git

V. References

Abel Chandra, Laura Tünnermann, Tommy Löfstedt, Regina Gratz (2023) Transformer-based deep learning for predicting protein properties in the life sciences eLife 12:e82819

Clauwaert J, Menschaert G, Waegeman W. Explainability in transformer models for functional genomics. Brief Bioinform. 2021 Sep 2;22(5):bbab060. doi: 10.1093/bib/bbab060. PMID: 33834200; PMCID: PMC8425421.

Ding, Yifei, et al. "A Novel Time–Frequency Transformer Based on Self–Attention Mechanism and Its Application in Fault Diagnosis of Rolling Bearings." *Mechanical Systems and Signal Processing*, vol. 168, 2022, p. 108616, doi:10.1016/j.ymssp.2021.108616.

Lin S, Staahl BT, Alla RK, Doudna JA. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. Elife. 2014 Dec 15;3:e04766. doi: 10.7554/eLife.04766. PMID: 25497837; PMCID: PMC4383097.

Moore KS, 't Hoen PAC. Computational approaches for the analysis of RNA-protein interactions: A primer for biologists. J Biol Chem. 2019 Jan 4;294(1):1-9. doi: 10.1074/jbc.REV118.004842. Epub 2018 Nov 19. PMID: 30455357; PMCID: PMC6322881.

Nishtala, S., Neelamraju, Y. & Janga, S. Dissecting the expression relationships between RNA-binding proteins and their cognate targets in eukaryotic post-transcriptional regulatory networks. *Sci Rep* 6, 25711 (2016). https://doi.org/10.1038/srep25711

Rahul Nikam, M Michael Gromiha, Seq2Feature: a comprehensive web-based feature extraction tool, *Bioinformatics*, Volume 35, Issue 22, November 2019, Pages 4797–4799

Shuang Zhang, Rui Fan, Yuti Liu, Shuang Chen, Qiao Liu, Wanwen Zeng, Applications of transformer-based language models in bioinformatics: a survey, *Bioinformatics Advances*, Volume 3, Issue 1, 2023

"The Database of RNA-Binding Specificities." *RBPDB*, rbpdb.ccbr.utoronto.ca/index.php. Accessed 12 May 2023.

Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57-63. doi: 10.1038/nrg2484. PMID: 19015660; PMCID: PMC2949280.