Methodology

Weixiong

2023/12/14

ProteinBERT

A typical method for converting the textual data into vector is implementing transformer(Vaswani et al. 2017), which was originally designed to solve Natural Language Process (NLP) problems and became popular in the recent studies. And subsequent research on BERT has also found that this model can maximally retain the original text information in the embeddings. In 2022, Brandes et al. (Brandes et al. 2022) proposed a pre-trained BERT named ProteinBERT expertized on protein data set and reported a SOTA performance making it possible to directly transfer amino acid sequences into a well performed embedding.

Generally speaking, ProteinBERT has the following advantages

- **Bidirectional context understanding**: BERT is designed to understand the context of a word in a sentence in both directions (left and right of the word), which enable the model to intake more information.
- **Pre-training on Large Datasets**: Its pretraining scheme combines language modeling with a novel task of Gene Ontology (GO) annotation prediction, and the newly introduced model design make it even more flexible and efficient for prediction.
- **Tokenizer**: Tokenizer can convert raw text into a format that the BERT model can understand. This project typically choose Rostlab prot_bert as our tokenizer.

However, such method assumed a equal length of input (protein sequence) which may be reasonable in model design while might introduce uncertainty in the model due to the loss of information. In this report we will discuss the choice of length of 1024 and 2048 for the input. Moreover, based on the preliminary result, only 7.5 percent of protein data has length more than 1200, thus we believe such choice will not cause much difference in the result.

Methodology

Methodology Introduction

The methodology for this protein function prodicting problem can be divided into two tasks.

1. Embedding protein data into processable tensor embeddings
2. Predicting the function of protein based on the embeddings.

Last section has introduced ProteinBERT, which mainly introduced to obtain embeddings for the protein sequence. To predict the function of protein, more specifically classify, this study mainly discuss about two approaches, Neural Networks(NN) which is useful in classfication problems and Variational Autoencoder (VAE) proposed by Kingma et al. (Kingma and Welling 2022). For neural network, this project simply implement sequential model since the sequential model is straightforward and easy to construct. Most categorization problems can be effectively solved with a linear stack of layers with proper active functions

As for VAE, a standard VAE network has two main components, see Figure below

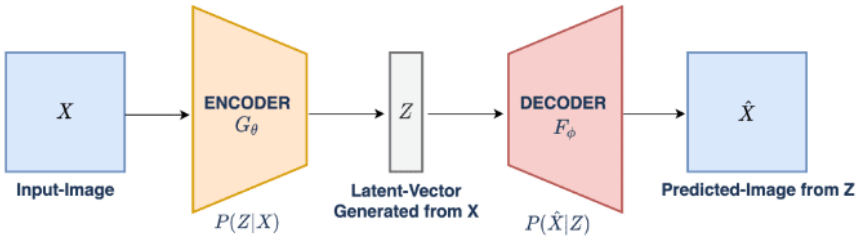

Figure 1: Structure for VAE

Encoder used for compress the data to a more compact form, a decoder which takes the encoded data and reconstructs the input data from the compressed form from encoder. Considering the special design of VAE, this method can help us predict the function of protein and acquire a new embedding for protein that takes less space which is essential in data managements.

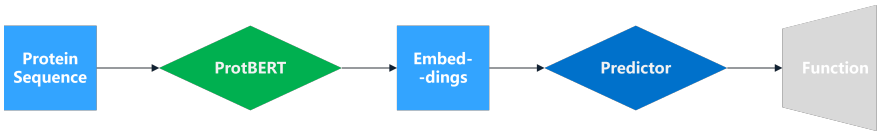

Figure 2: Workflow of This Project

It is worthy to mention, for the labels, since the data contains 31,466 GO terms (dimensions) from BBO, CCO, MFO, model only consider the first 1500 or 2000 most frequently contained terms which takes up 85 percent of the existing terms as the output. The following part will discuss the choice of parameters in detail

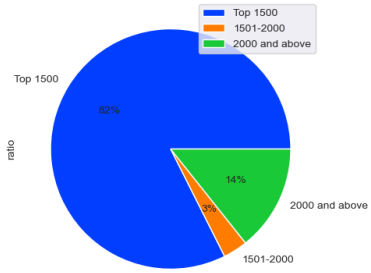

Figure 3: Pie Chart for Top Terms

Result

Now, the result of model will be discussed.

The data set was first divided into 2 parts, train, validation. For the train set, a 5-fold cross validation is implemented and the best model is chosen based on the performance on validation set. The ratio of validation set is set to 0.1 of the orginal data.

First, the set of 2048 length for the input of ProtBERT and 1500 length For the label dimension was trained and tested, as shown below

The result of Figure 4 and 5 showed that both methods reached a maximum AUC of 0.83, however VAE converge a little bit faster to reach to its optimum at the second epoch. Moreover, VAE works much better in test sets and its performance is more stable than Neural Network which shows a pattern of overfitting. This may be beneficial from a more complex model structure of VAE. Here the hidden layer size for VAE is set as 512, and Neural Networks takes the original input as its input.

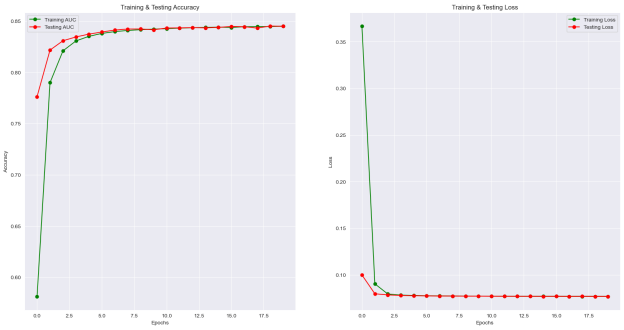

Figure 4: Train and Test Results For VAE

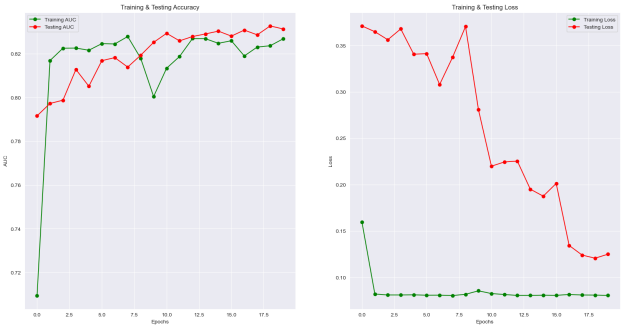

Figure 5: Train and Test Results For Neural Networks

based on the figure below, it can be seen that both of the models perform similarly and VAE performs slightly better on validation data while it only takes a embedding with dimension of 512.

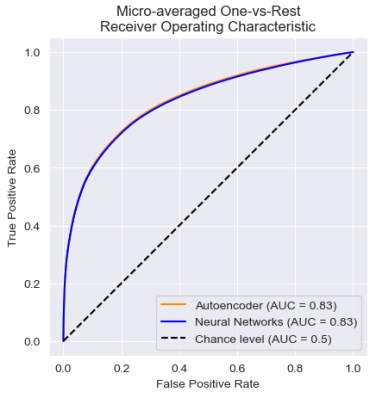

Figure 6: ROC curve For VAE and Neural Network

This result indicates a possibility to reduce the dimension of embedding from ProtBERT. Also, different length of input for Bert and Prediction model were tested, here the prediction model is chosen as VAE.

Based on the result from table above, no significant difference is observed. Which shows that this method by far was not sensitive to the choice of parameters. In the future study, more settings will be tested.

Top Terms& Protein Length	1024	2048
1500	0.84	0.83
2000	0.85	0.85

Table 1: AUC for Different Settings

References

Brandes, Nadav, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. 2022. "ProteinBERT: a universal deep-learning model of protein sequence and function." *Bioinformatics* 38 (8): 2102–10. https://doi.org/10.1093/bioinformatics/btac020.

Kingma, Diederik P, and Max Welling. 2022. "Auto-Encoding Variational Bayes." https://arxiv.org/abs/1312.6114.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." *CoRR* abs/1706.03762. http://arxiv.org/abs/1706.03762.