BioTranslator Project

Paper

https://www.nature.com/articles/s41467-023-36476-2

Project's Original GitHub

https://github.com/HanwenXuTHU/BioTranslatorProject

Overview

The BioTranslator Project is a multi-faceted ML framework designed to utilize zero-shot protein function prediction, using Gene Ontology terms embedded into vectors. With these vectors, the model predicts which terms apply to which protein representations, and produces the Area Under the Receiving Operating Curve (AUROC) score. This score quantitatively tells us whether the model is assigning correct terms to the protein's function.

This project focuses on the zero-shot protein function prediction. Zero-shot prediction means that the model is predicting the function for proteins it has not seen before–thus, informs whether the model can accurately assess terms based on its own understanding. Within the project provided on GitHub, BioTranslator evidently has its own NLP model used to process the datasets' GO (Gene Ontology) terms into vector embeddings for the translator to utilize. This capstone project aims to test whether a GPT-Embedding can outperform the BioTranslator's own specialized embedding for this task.

Baseline and Dataset

To begin, I initially recreated the results of the paper this project is utilized in. From the AUROC scores that I recreated, I was most interested in expanding upon the SwissProt dataset. This dataset contained AUROC scores lower than that of the GOA_Human dataset I had initially planned to experiment with. For this reason, I selected this dataset to experiment with, as it presents an opportunity for improvement. Upon running the BioTranslator with the following command, I received the following AUROC scores.

```
python Protein/main.py

--method BioTranslator

--dataset SwissProt

--data_repo /homes/iws/chaafen/428/BioTranslatorProject/20120447/ProteinDataset/
```

--task zero_shot

--encoder_path /homes/iws/chaafen/428/BioTranslatorProject/TextEncoder/encoder.pth

--emb_path /homes/iws/chaafen/428/BioTranslatorProject/embeddings

Ontology	Fold 0	Fold 1	Fold 2	AVG AUROC
Biological Process	0.7631	0.7607	0.7817	0.7686
Molecular Function	0.7959	0.8024	0.8129	0.8038
Cellular Composition	0.8185	0.8224	0.8287	0.8232

These scores evaluate how well the model distinguishes between correct and incorrect labels for each of the proteins and their ontology terms. The model makes these distinctions using the embeddings of the GO dataset–a vector representation of gene ontology terms organized by researchers. This dataset contains thousands of terms including information about their id, data, and terms associated with their biological process, molecular function, and cellular composition. It is only through these embeddings that the model is able to tell the degree of similarity between terms, and thus, how likely it is that a term should be assigned to a gene.

Although the GO Terms dataset is one that is organized and compiled by human researchers, the model operates using the embeddings when it runs the model. Thus, it would be more interesting to create a new embedding–one that is not created by the model–to see whether the model better interprets its own vector rankings, or one created by another model. Essentially, this will compare the understanding of the BioTranslator's NLP model and the GPT model in the dataset.

GPT-Embeddings

The previous scores confirmed the findings of the paper, and led to the next step of the project. For the next step, the project needed to have GPT created embeddings to run the model on. For this reason, I selected Open AI's API, due to its broad specialization coverage and strong capabilities. After setting up the API on my account and receiving the key, I setup my terminal workspace with it, and an gpt_embeddings.py file to convert the dataset's go.obo file into embeddings using Open API's text-embedding-3-large model. This process included first parsing the GO term, and then using the model's text to embedding function. Upon receiving these embeddings and converting them to the appropriate .pkl file type, I was then able to run the model again using the GPT created embeddings.

python Protein/main.py

--method BioTranslator

--dataset SwissProt

--data_repo /homes/iws/chaafen/428/BioTranslatorProject/20120447/ProteinDataset/

--task zero_shot

--encoder_path /homes/iws/chaafen/428/BioTranslatorProject/TextEncoder/encoder.pth

--emb_path /homes/iws/chaafen/428/BioTranslatorProject/embeddings

Ontology	Fold 0	Fold 1	Fold 2	AVG AUROC
Biological Process	0.7683	0.7670	0.7852	0.7735
Molecular Function	0.8066	0.8030	0.8163	0.8086
Cellular Composition	0.8186	0.8281	0.8312	0.8260

These results indicate that a broader and more generalized GPT model can achieve AUROC scores at-par with or marginally greater than the BioTranslator's own specialized model. Furthermore, this suggests that the GPT model's own embeddings can competitively compare to the domain-specific model that BioTranslator presents. These findings tell us that GPT can serve as a strong model for embeddings on zero-shot function prediction, hence, without fine tuning on the domain-specific data.

Comparison

Ontology	Original	GPT-Based	Δ
BP	0.7686	0.7735	+0.0049
MF	0.8038	0.8086	+0.0048
CC	0.8232	0.8266	+0.0034

According to a 1982 paper on the interpretations of AUROC scores, “A difference in AUCs of 0.05 or more may be considered meaningful in medical decision-making contexts” (Hanley & McNeil, 1982). As this difference approaches that threshold, it suggests there is significance in exploring this finding further, and expanding upon it to reach this threshold. Regardless, the consistent positive change in the AUROC score indicates that the GPT-embedding can perform at-par with, and potentially significantly greater than the model’s own embedding.

Relevance

These findings suggest that GPT’s general LLM model can be integrated into computational biological projects, due to its accuracy in embedding, and more importantly, its lack of fine-tuning. This implies that without fine-tuning or seeing the data before, the model was able to interpret the data into appropriate and accurate vector embeddings for each term. This can remove the bottleneck many projects face with creating embeddings for projects like these, and can allow researchers to focus more on the model itself, rather than the vector embeddings for it to understand.

Conclusion and Future Work

These findings suggest that future research can use more flexible and efficient pipelines. The elimination for the need of a project specific embedding process, can create more scalable projects in the future. GPT can be used to complement the strengths of human researchers, by translating their knowledge into vector embeddings these models can understand.

With further experimentation, it would be worthwhile to explore the embeddings of more AI models like Deepseek AI, as well as the AUROC score comparisons of multi-shot prediction to see which embedding improves more with fine-tuning.

This experiment suggests that the BioTranslator Model in tandem with GPT-embeddings, can effectively perform zero-shot protein function prediction. Thus, it can accurately predict a protein’s function, even when it does not know it already. This successfully shares the GPT embedding model’s accuracy with understanding the Gene Ontology dataset, and with analyzing those terms to create a functional gene vector embedding. The success of this experiment indicates that projects can accelerate the creation of models and findings from them, by joining the strength of these general models with the specialization from researchers.

Citation

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747