Research Proposal : Transformer in Computational Biology

Yunsoo Kim

Abstract

Transformer architecture has been a de-facto standard in natural language processing. Its adaptations in other fields such as computer vision showed promising results that this architecture is a powerful neural network in representation learning regardless of the data type. Recently, Transformer has been successfully used in computational biology, mainly for sequential data. However, its usage in biology is still limited. Here, I propose to expand Transformer architecture in other types of computational biology data, specifically graph data.

Keywords: Transformer, Graph, Network, Computational Biology, Deep Learning

1 Introduction

Transformer is the prevalent network architecture in natural language processing (NLP) (Vaswani *et al.*, 2017). Transformer uses self-attention to capture each word's influence on another in a given text, and it solved the vanishing gradient problem in recurrent neural network (RNN) (Vaswani *et al.*, 2017). Recent advances in pre-training language model based on Transformer architecture reached state-of-the-art on many NLP benchmark datasets, including results that surpassed human performance (He *et al.*, 2020). A major direction is scaling up the language model parameters (trillions) to enhance model capacity and performance (Brown *et al.*, 2020). Another direction is focused on model efficiency (fewer parameters while retaining the model capacity) (Clark *et al.*, 2020). The last direction to highlight is to capture longer-range dependencies between tokens (words) to process longer input text (sequence) (Beltagy *et al.*, 2020).

Recently, Transformer model moved to computer vision (CV). Its adaptations showed better results than convolutional neural network (CNN) in tasks such as image classification, object detection, and semantic segmentation (Dosovitskiy *et al.*, 2021 and Liu *et al.*, 2021). Besides the performance improvement, Transformer architecture in CV demonstrated the possibility of self-supervised representation learning in CV (Bao *et al.*, 2021). Similar to NLP, the current research directions are scaling model parameters, model efficiency, and larger input image (Liu *et al.*, 2021).

Transformer architecture has been adapted to biology (Lee *et al.*, 2019, Lewis et al., 2020, and Jumper *et al.*, 2021). However, the current adaptations in computational biology are limited to sequential data such as text for bioNLP and protein sequence for structure prediction.

Here I propose Transformer adaptations in other types of computational biology data. In specific, I want to focus on graph data: firstly on graph generation, then on Graph Transformer. My research on graph transformer can expand to other data such as sequence, text, and image data in biology.

1.1 Graph in Computational Biology

Graphs are typical non-Euclidean data, unlike images and texts. Graphs or networks can represent complex relationships between entities (objects). Networks are commonly used in computational biology to highlight a relationship between two entities. The information in the network depends on the biological data that they are made of. Metagenomic abundance network shows potential symbiosis within the microbiota. Gene regulatory network highlights gene activation and inhibition relationships. Biomedical knowledge graph holds information about co-occurrences of gene, drug, and disease mentions from literature. Even the molecular structure of proteins and RNAs can be represented as graphs.

Network in biology can be used in many parts of drug discovery. It can be used from early stages such as novel target discovery or drug repurposing to later stages such as adverse drug events prediction. I am interested in target identification and clinical trial prediction.

I aim to work on datasets for multiple sclerosis. Multiple sclerosis is a brain disease that changes our immune system to attack myelin sheath. It can cause disability but has no cures. Current known target candidates are G protein-coupled receptors, fingolimod, and Ion-channel protein, Anoctamin 2 (Du et al., 2012 and Avoglu et al., 2016).

1.2 Graph Generation

The generation of the biological networks can be divided into two steps: identifying entities for nodes in the graph and inferring relationships between nodes as edges. Identification of nodes in biological networks often requires bioinformatics pipelines to generate a count matrix. Inference of dependencies between the nodes often uses statistical measures, such as mutual information criterion, on the count data.

In the case of the metagenomic abundance network, 16S rRNA sequence data is used to identify the bacteria present within samples. The sequence data is binned by similarity into operational taxonomic units (OTU). Each OTU is treated as a node. The number of sequences for each OTU is used as abundance count data. Dependencies between OTUs are inferred using statistical measures on the abundance data.

However, graph generation from biological sequence data often suffers from relative abundance data (Quinn *et al.*, 2019). The number of sequences is not equal to the absolute number of cells or transcripts. This is mainly caused by DNA library amplification.

Commonly used statistic measures do not perform well in the inference of dependencies between relative data. Proportionality measure performed better for relative data in the compositional analysis, but it has limitations with zero handling (Quinn *et al.*, 2019).

1.3 Graph Analysis

The generated biological networks demonstrate the complex mechanisms of diseases and can lead to discoveries such as biomarkers. These findings can be achieved with the help of further analysis on the networks. Algorithms such as random walks have been used to capture network topology and neighborhood information. The extracted information can be used as features for statistical or machine learning analysis to highlight nodes and edges in the network.

Recent trend in graph representation learning has moved from feature engineering to deep learning. Graph neural network (GNN) has been a dominant architecture in graph representation (Gilmer *et al.*, 2017). It uses message passing to update node representation by aggregating representations of its neighbors. For graph level representation, GNN uses an additional function called readout, which aggregates node and edge representations. GNN has been successfully used in bioinformatics from missing value imputation to drug repurposing (Hasibi et al., 2020 and Gysi *et al.*, 2020).

However, the current GNN methods suffer from the over-smoothing problem (indistinguishable node representation) caused by deeper layers. Thus, GNN is limited in the model capacity as the number of layers has to be small for a good performance (Chen *et al.*, 2019). From my understanding, a possible explanation for this problem is that graphs are small-world networks.

Recently, Transformer architecture has achieved competitive performance in graph level representation for small graphs such as molecules (Ying *et al.*, 2021). Unlike GNN, it does not encounter the over-smoothing problem (Ying *et al.*, 2021). Its performance is rather improved as the model gets deeper. Still, it is limited to graph representation for small graphs due to its large memory complexity. For biological networks, this limitation is critical as the size of graphs is larger than that of a molecule.

2 Research Questions

My hypothesis is **"node and graph representation in biological networks will be improved by Graph Transformer as it has been powerful in other data formats."**

The goal of this research is to address the following questions:

- How can I develop Graph Transformer aimed to improve the representation learning of biological networks (heterogeneous)?
- Can I find a novel target for Multiple Sclerosis using the heterogeneous graph analysis empowered by deep learning?

3 Method

3.1 Relative data Graph Generation

Relative data from bioinformatics analysis of sequence data can be any omics data that uses amplification of DNA library for sequencing. The two data I will focus on microbiome abundance data. Microbiome abundance data will be downloaded from MGnify (Mitchell *et al.*, 2020).

The current proportionality measure, which I worked on during my MSc, has limitations with zero handling, and the recent work uses a box-cox transformation to handle (Quinn *et al.*, 2019). As zeros hold important information, I propose graph neural network based network inference for relative data without zero-handling.

Altered microbiota can cause disease and I will explore the association between multiple sclerosis and the altered microbiota. Also, for target identification, I will use single-cell disease-gene associations from SC2disease (Zhao *et al.*, 2021).

3.2 Knowledge Graph Generation

Biomedical knowledge graph can provide additional information such as disease for gene expression network. Thus, knowledge graphs can be used to align graphs to make heterogeneous graphs. I aim to construct knowledge graphs using NLP and knowledge databases. I have been working on chemical knowledge graph construction in LG Chem using NLP, and based on this experience I will construct biological knowledge graphs.

Biology-related entities such as genes, disease, and drugs can be recognized from literature, and the relation between the entities can be inferred using the biomedical language model. The extracted entities will be nodes and the relations will be edges. GNN and Graph Transformer will be used for link prediction for unknown edges in the heterogeneous graphs.

3.3 Graph Transformer

Graph representation can be used in many fine-tuning tasks such as graph classification and node property prediction. The current limitation of Graph Transformer architecture is the small input size. Thus, I want to work on linear attention that can reduce the memory complexity, so that larger graphs can be learned. I will use mixed-precision, gradient accumulation, and linear attention based on my experience working with memory-efficient Transformer training in NLP at LG Chem.

Also, subgraph sampling can be a solution to the current limitation. Subgraph sampling can be as simple as random sampling, but we want to sample subgraphs while preserving topological and neighborhood information. Topological and neighborhood information will be provided as encodings for Transformer. Geometric deep learning can be used to enhance the encodings. Sampling will be part of Transformer input embedding layer to be trained as well.

I will work on the representation of the heterogeneous graphs made from the metagenomics abundance network, single-cell disease-gene association network, and knowledge graphs for multiple sclerosis. Then, node representation can be used for target identification, and graph representation can be used for clinical outcome prediction.

4 Discussion

The proposed research aims to make Transformer architecture a de-facto standard in computational biology, specifically for network biology. This research can expand to other data such as sequence, text, and image data in biology.

5 Timeline

This section describes my plan for PhD timeline. Table 1 specifies the year and term for what I plan to do. The timeline is not fixed.

Year, Term	**Plan**
1st Yr, Michaelmas	Language Model Review of SOTA
1st Yr, Lent	Focus on Graph Generation
1st Yr, Easter	1st Year Report, Viva Graph Transformer
1st Yr, Summer	Summer Internship
2nd Yr, Michaelmas	Focus on Graph Transformer
2nd Yr, Lent	Final Analysis
2nd Yr, Easter	2nd Year Report/ Dissertation Schedule
2nd Yr, Summer	Summer Internship
3rd Yr, Michaelmas	Start Write Up
3rd Yr, Lent	
3rd Yr, Easter	Submission and Viva

Table 1: Timeline.

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2020.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. Language models are few-shot learners. *arXiv preprint* arXiv:2005.14165, 2020.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*, 2020.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint* arXiv:2004.05150, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is 11 worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint* arXiv: 2103.14030, 2021.

Hangbo Bao, Li Dong, and Furu Wei. BEiT: Bert pre-training of image transformers. *arXiv preprint* arXiv:2106.08254, 2021.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling Up Capacity and Resolution. *arXiv preprint* arXiv:2111.09883, 2021.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 2019.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *ACL*, 2020.

Thomas P. Quinn, Ionas Erb, Greg Gloor, Cedric Notredame, Mark F. Richardson, and Tamsyn M. Crowley. A field guide for the compositional analysis of any-omics data. *Gigascience*, 2019.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *ICML*, 2017

Deisy Morselli Gysi, Ítalo Do Valle, Marinka Zitnik, Asher Ameli, Xiao Gan, Onur Varol, Helia Sanchez, Rebecca Marlene Baron, et al. Network medicine framework for identifying drug repurposing opportunities for covid-19. *arXiv preprint* arXiv:2004.07229, 2020.

Ramin Hasibi and Tom Michoel. A Graph Feature Auto-Encoder for the Prediction of Unobserved Node Features on Biological Networks. *arXiv preprint* arXiv: 2005.03961, 2020.

Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. Measuring and Relieving the Over-smoothing Problem for Graph Neural Networks from the Topological View. *arXiv preprint* arXiv: 1909.03211, 2019.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do Transformers Really Perform Bad for Graph Representation? arXiv preprint arXiv:2106.05234, 2021.

Alex L Mitchell, Alexandre Almeida, Martin Beracochea, Miguel Boland, Josephine Burgin, Guy Cochrane, et al. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research*, 2020.

Tianyi Zhao, Shuxuan Lyu, Guilin Lu, Liran Juan, Xi Zeng, Zhongyu Wei, Jianye Hao, Jiajie Peng. SC2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic Acids Research*, 2021

Changsheng Du and Xin Xie. G protein-coupled receptors as therapeutic targets for multiple sclerosis. In *Cell Research*, 2012.

Burcu Ayoglu, Nicholas Mitsios, Ingrid Kockum, Mohsen Khademi, Arash Zandian, Ronald Sjöberg, Björn Forsström, et al. Anoctamin 2 identified as an autoimmune target in multiple sclerosis. In *PNAS*, 2016.