2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) | 978-1-6654-9402-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/WI-IAT55865.2022.00121

How Severe is Your COVID-19? Predicting SARS-CoV-2 Infection with Graph Attention Capsule Networks

1st Runjie Zhu
Information Retrieval and Knowledge Management Research Lab
York University
Toronto, Canada
https://orcid.org/0000-0003-4890-487X

2nd Zhiwen Xie
School of Computer Science
Wuhan University
Wuhan, China
xiezhiwen@whu.edu.cn

3rd Guangyou Zhou
School of Computer Science
Central China Normal University
Wuhan, China
gyzhou@mail.ccnu.edu.cn

***Abstract*—Recent studies in machine learning have demonstrated the effectiveness of applying graph neural networks (GNNs) to single-cell RNA sequencing (scRNA-seq) data to predict COVID-19 disease states. In this study, we propose a graph attention capsule network (GACapNet) which extracts and fuses Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) transcriptomic patterns to improve node classification performance on cells and genes. Significantly different from the existing GNN approaches, we innovatively incorporate a capsule layer with dynamic routing into our model architecture to combine and fuse gene features effectively and to allow those more prominent gene features present in the output. We evaluate our GACapNet model on two scRNA-seq datasets, and the experimental results show that our GACapNet model significantly outperforms state-of-the-art baseline models. Therefore, our study demonstrates the capability of advanced machine learning models to generate predictive features and evolutionary patterns of the SARS-CoV-2 pathogen, and the applicability of closing knowledge gaps in the pathogenesis and recovery of COVID-19.**

***Index Terms*—Bioinformation, COVID-19, Natural Language Processing, Node Classification, Text Mining.**

I. Introduction

THE week of July 19, 2022 marks the two years and seven months anniversary since the World Health Organization announced the global outbreak of COVID-19. This COVID-19 pandemic has spread across the globe and has hit us hard with severe public health and economic consequences for two years and a half. The infection of the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), similar to its same virus family members of Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome CoV (MERS-CoV), can lead to fatal pneumonia that is in association with rapid replication of the deadly virus, an elevation of proinflammatory cytokines, as well as an infiltration of immune cells. While people from all over the world are still in process of receiving a number of doses of vaccines, the virus itself has been evolving rapidly and continuing to surprise experts around the world. The cost of human capital continues to mount over the past few months, with more than 615.16 million cases confirmed worldwide, more than 14.38 million active cases and more than 6.52 million deaths as of Thursday, September 15, 2022[1].

During the course of two and a half years, a great amount of COVID-19 related data from a wide range of sources and formats have been accumulated. These data collections are considered as valuable and key theoretical basis and evidence for further clinical research and biomedical analyses. Since the disease belongs to the department of respiration by nature, most of the existing COVID-19 related data collections are constructed and presented in a graph structure. Indeed, there is a great portion of important real-world information and data that have long been presented in a form of graph structure before COVID-19. These graph structured data include person-person relationship in social networks [1], [2], product knowledge graph for e-commerce [3]–[5], medical knowledge graph in healthcare industry [6]–[10], and those publicly available COVID-19 related knowledge graphs [11]–[16] etc.. Even though various evidences have hinted it as an important research direction for machine learning to extend neural networks to process graph structured data, this research domain has not caught high attention until very recently.

Graph Attention Networks (GATs) [17] is a lately introduced neural network architecture that is capable of operating graph structured data and supporting predictive tasks, such as node classifications and link predictions. In the past few years, a great number of studies have been published using the GAT and regular graph neural networks (GNNs) approach. Specifically, these models have been widely adopted to explicit relational data structures such as knowledge graphs [18], [19], non-explicit relational data structures such as texts [17], and other applications such as generative models [20]. Although

This research is supported in part by the research grant from Natural Sciences and Engineering Research Council (NSERC) of Canada and York Research Chairs (YRC) program. This work is also supported by the National Natural Science Foundation of China under Grant 61972173, and supported by the National Key R&D Program of China under Grant 2018YFC1604000.

[1]https://www.worldometers.info/coronavirus/

DOI 10.1109/WI-IAT55865.2022.00121

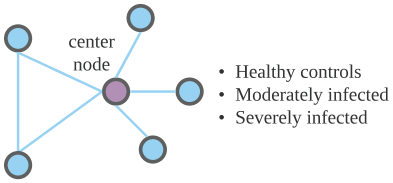

Fig. 1. An example of sub-graph in the COVID-19 patients [24] dataset. The purple node represents the center cell node and the blue nodes represent the neighboring cell nodes. All cells are classified into three infectious levels, namely the healthy controls, moderately infected and severely infected.

these existing studies have shown promising results, they also share one deficiency that they usually model each interactions individually without considering complex relations between them enough. This characteristic of these past studies restrict them from absorbing more relevant and useful neighboring features, and transforming them into more representative outputs. In the context of COVID-19, it is important for us to empower our current medical research with these advanced machine learning methodologies to uncover secrets hidden in the infection and pathogenesis of SARS-CoV-2, and to suggest viable therapeutic directions.

Given that the counts of transcripts are correlated with expressions of genes, the technology of single-cell RNA sequencing (scRNA-seq) is often used to capture thousands of expressions of cell genes under different circumstances to comprise large data collections [21]–[23]. Even though massive single-cell data can be captured effectively, the high sparseness, high heterogeneity and high dimensions characteristics of these single-cell data pose challenges in sorting out significant features that can formulate causal relations to the pathophysiological trajectory and interactive reactions.

Our research interest thus arises from two recently published work [25], [26]. The first paper applies GATs to disease states prediction with scRNA-seq data [27]. Building on top of the prior work, the second paper utilizes GNNs and GATs to improve node classification tasks with edge features which are derived from graphs and self-supervised learning.

Aligning with the work above, we focus on identifying the patterns of SARS-CoV-2 transcriptome and the types of molecular cells linked to the degree of disease infectiousness and severity with single cell transcriptomic data collections. However, different from all prior works, in this paper, we take a new approach by proposing a graph attention capsule network (GACapNet) which stacks the multi-head attention networks above a capsule network, comprising of a primary capsule layer and a dynamic routing procedure. To the best of our knowledge, we are the first one to extend GATs and Capsule networks to scRNA-seq datasets for COVID-19 disease state predictions.

The proposed architecture GACapNet has a strong advantage in dealing graph structured data and in generating more useful and representative node features for enhancing GATs' performance on node classification on cells. To align with the prior work [25], we use the processed and clustered dataset in graph structure as inputs to our model. The prior work constructed the graphs from cells by utilizing batch-balanced weighted KNN graph [28] with matrices of gene features. In other words, the structure of the graph was constructed by finding the closest cell nodes with KNN. The task of our study is to identify and classify SARS-CoV-2 infected cells that are linked to different degrees of infectiousness. For example, Figure 1 shows an example of sub-graph in the dataset which aims to group cells into three levels of cell node infectiousness, namely the severely infected class, the mildly infected class, and the healthy controls. All cells are grouped under the assumption that the cells with similar gene features should be closer to each other. If the two cell nodes belong to the same category, an edge is formed in between these two nodes.

For a given constructed graph $\mathcal{G}$, we first use multi-head graph attention to aggregate features from the neighbors of a center node, shown as the purple node in Figure 1. The center node stands for a cell with a sequence of gene features. Each cell in the dataset can be a center node, and the surrounding nodes are defined as neighboring cell nodes, shown as the blue nodes in Figure 1. In order to preserve the center node feature (or the cell's feature), we use a simple feed-forward network on the node features to obtain a transformed node feature from the center node. Then, the neighborhood features generated from different attention heads and those transformed node features are fed into a capsule network. The capsule network comprises of a primary capsule layer and a dynamic routing process. From the experimental results on two scRNA-seq datasets, we prove and reason that our proposed GACapNet is able to classify cells according to its severity level effectively. In other words, it outperforms other GATs in extracting insightful data of SARS-CoV-2 transcriptomic patterns and molecular cell types that are linked to the infection and severity of COVID-19.

This work makes four primary contributions, summarized as follows:

- We apply advanced machine learning models of GATs and Capsule networks to COVID-19 domain of medical research. To the best of our knowledge, our GACapNet is the first one to extend the GATs and Capsule networks models to scRNA-seq data, aiming to uncover the secrets hidden in the infection of SARS-CoV-2, to understand the COVID-19 pathogenesis, and to suggest the therapeutic directions and development by identifying the patterns of SARS-CoV-2 transcriptome and the types of molecular cells that are linked to the degree of disease infectiousness and severity;
- We propose to use multi-head graph attention networks in our model architecture to aggregate more information-rich neighboring features of the center node, and to preserve its original node feature to the maximum;

- We incorporate a capsule layer with dynamic routing into our model to effectively combine and fuse the gene features, and to allow more prominent features of the severely infected cells and healthy controlled cells to present in the output;
- The outstanding experimental results of our model on two scRNA-seq data collections and the in-depth analysis we provide on the SARS-CoV-2 infections can be used as basic hypotheses in further medical and biomedical COVID-19 research and clinical validations.

The full paper with experimental results will be presented in a separate work, which will give a thorough literature review on relevant research approaches and various models, describe the architecture of our proposed graph attention capsule network (GACapNet) model, illustrate our experimental results and model comparisons, and conclude our paper with some ideas for future research.

ACKNOWLEDGMENT

This research was supported in part by the research grant from Natural Sciences and Engineering Research Council (NSERC) of Canada and York Research Chairs (YRC) program. This work was also supported by the National Natural Science Foundation of China under Grant 61972173, and supported by the National Key R&D Program of China under Grant 2018YFC1604000.

REFERENCES

[1] N. F. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor, "Industry-scale knowledge graphs: Lessons and challenges," *ACM Queue 17*, p. 20, 02 2019.

[2] R. S. Gonçalves, M. Horridge, R. Li, Y. Liu, M. A. Musen, C. I. Nyulas, E. Obamos, D. Shrouty, and D. Temple, "Use of owl and semantic web technologies at pinterest," *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference*, pp. 418–435, 10 2019.

[3] A. Krishnan, "Making search easier: How amazon's product graph is helping customers find products more easily," *Amazon Blog*, 2018.

[4] R. J. Pittman, A. Srivastava, S. Hewavitharana, A. Kale, and S. Mansour, "Cracking the code on conversational commerce," *eBay Blog*, 2017.

[5] J. Wang, P. Huang, H. Zhao, Z. Zhang, B. Zhao, and D. L. Lee, "Billion-scale commodity embedding for e-commerce recommendation in alibaba," in *KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*. the Association for Computing Machinery, 07 2018, pp. 839–848.

[6] P. Ramaswami, "A remedy for your health-related questions: health info in the knowledge graph," *Google Blogs*, 02 2015.

[7] A. Lally, S. Bagchi, M. A. Barborak, D. W. Buchanan, J. Chu-Carroll, D. A. Ferrucci, M. R. Glass, A. Kalyanpur, E. T. Mueller, J. W. Murdock, S. Patwardhan, and J. M.Prager, "Watsonpaths: Scenario-based question answering and inference over unstructured information," *AI Magazine*, vol. 38.

[8] V. Pinchin, "I'm feeling yucky :(searching for symptoms on google," *Google Blogs*, 06 2016.

[9] P. Sondhi, J. Sun, H. Tong, and C. Zhai, "Sympgraph: A framework for mining clinical notes through symptom relation graphs," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1167–1175, 08 2012.

[10] M. Sheng, J. Wang, Y. Zhang, X. Li, C. Li, C. Xing, Q. Li, Y. Shao, and H. Zhang, "Dockg: A knowledge graph framework for health with doctor-in-the-loop," *International Conference on Health Information Science 2019*, pp. 3–14, 10 2019.

[11] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, P. Mooney, D. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, C. Wilhelm, B. Xie, D. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier, "CORD-19: the covid-19 open research dataset," *CoRR*, vol. abs/2004.10706, 2020.

[12] X. Zeng, X. Song, T. Ma, X. Pan, Y. Zhou, Y. Hou, Z. Zhang, G. Karypis, and F. Cheng, "Repurpose open data to discover therapeutics for COVID-19 using deep learning," *Journal of Proteome Research*, vol. acs.jproteome.0c00316, 2020.

[13] Q. Wang, M. Li, X. Wang, N. Parulian, G. Han, J. Ma, J. Tu, Y. Lin, H. Zhang, W. Liu, A. Chauhan, Y. Guan, B. Li, R. Li, X. Song, H. Ji, J. Han, S. Chang, J. Pustejovsky, J. Rah, D. Liem, A. Elsayed, M. Palmer, C. R. Voss, C. Schneider, and B. A. Onyshkevych, "COVID-19 literature knowledge graph construction and drug repurposing report generation," *CoRR*, vol. abs/2007.00576, 2020.

[14] I. Shen, L. Zhang, J. Lian, C. Wu, M. González-Fierro, A. Argyriou, and T. Wu, "In search for a cure: Recommendation with knowledge graph on CORD-19," in *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, 2020, pp. 3519–3520.

[15] C. Wise, V. N. Ioannidis, M. R. Calvo, X. Song, G. Price, N. Kulkarni, R. Brand, P. Bhatia, and G. Karypis, "COVID-19 knowledge graph: Accelerating information retrieval and discovery for scientific literature," *CoRR*, vol. abs/2007.12731, 2020.

[16] L. Bellomarini, M. Benedetti, A. Gentili, R. Laurendi, D. Magnanimi, A. Muci, and E. Sallinger, "COVID-19 and company knowledge graphs: Assessing golden powers and economic impact of selective lockdown via AI reasoning," *CoRR*, vol. abs/2004.10119, 2020.

[17] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[18] T. Hamaguchi, H. Oiwa, M. Shimbo, and Y. Matsumoto, "Knowledge transfer for out-of-knowledge-base entities: A graph neural network approach," in *IJCAI '17: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017, pp. 1802–1808.

[19] Z. Wang, Q. Lv, X. Lan, and Y. Zhang, "Cross-lingual knowledge graph alignment via graph convolutional networks," in *EMNLP '18: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, p. 349–357.

[20] A. Bojchevski, O. Shchur, D. Zugner, and S. Gunnemann, "Netgan: Generating graphs via random walks," in *ICML '18: Proceedings of the 35th International Conference on Machine Learning*, 2018.

[21] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y.Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas, "Massively parallel digital transcriptional profiling of single cells," *Nature Communications*, vol. 8:14049, 2017.

[22] B. Hwang, J. H. Lee, and D. Bang, "Single-cell rna sequencing technologies and bioinformatics pipelines," *Experimental Molecular Medicine*, vol. 50, 96, 2018.

[23] T. Stuart and R. Satija, "Integrative single-cell analysis," *Nature Reviews Genetics*, vol. 20(5), 2019.

[24] M. Liao, Y. Liu, J. Yuan, Y. Wen, and Z. Zhang, "Single-cell landscape of bronchoalveolar immune cells in patients with covid-19," *Nature medicine*, vol. 26, no. 6, 2020.

[25] A. Sehanobish, N. G. Ravindra, and D. van Dijk, "Gaining insight into sars-cov-2 infection and COVID-19 severity using self-supervised edge features and graph neural networks," *CoRR*, vol. abs/2006.12971, 2020.

[26] N. G. Ravindra, M. M. Alfajaro, V. Gasque, J. Wei, R. B. Filler, N. C. Huston, H. Wan, K. Szigeti-Buck, B. Wang, R. R. Montgomery *et al.*, "Single-cell longitudinal analysis of sars-cov-2 infection in human bronchial epithelial cells," *bioRxiv*, 2020.

[27] N. Ravindra, A. Sehanobish, J. L. Pappalardo, D. A. Hafler, and D. V. Dijk, "Disease state prediction from single-cell data using graph attention networks," in *CHIL '20: Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020, pp. 121–130.

[28] K. Polanski, M. D. Young, Z. Miao, K. B. Meyer, S. A. Teichmann, and

J.-E. Park, “Bbknn: fast batch alignment of single cell transcriptomes,” *Bioinformatics*, vol. 36(3), p. 964–965, 2019.