HOW TO IMPROVE THE PERFORMANCE OF THE FUSED ARCHITECTURE CONSISTING OF A TABULAR TRANSFORMER AND A GRAPH NEURAL NETWORK USED FOR REPRESENTATION LEARNING FOR MULTIMODAL DATA?

Author: Dragomir Drashkov D.D.Drashkov@student.tudelft.nl

Supervisors: Kubilay Atasu, Atahan Akyıldız

TU Delft

1.INTRODUCTION

- A knowledge graph created from tabular data adds additional opportunity for the machine learning model to learn patterns of the data.
- XGBoost [4] is a state of the art tree boosting model applicable for many machine learning tasks. We are trying to outperform it with deep learning models.
- The way to do that is combining tabular transformers and graph neural networks (GNNs)
- Transformers have been created to avoid recurrence and convolutions . Tabular transformers build on transformers by being able to work with tabular data.
- GNNs learn from a graph by using message passing between nodes
- We are trying to find the best performing fused architecture, consisting of a tabular transformer and a GNN.

2.BACKGROUND

- Self-attention assigns a relevance score to each word based on its similarity to other words in the sequence.
- FT-Transformer[9] converts every feature to an embedding and applies the Transformer architecture on each.
- Trompt - prompt learning adjust a large pre-trained model through a set of prompts outside the model.
- ResNet[9] - strong baseline, superior performance over traditional Multilayer Perceptrons (MLPs).
- GNNs - learn from the graph - aggregate neighborhoods of nodes - aggregate and combine nodes.
- GraphSage [11] and GCN [16] are popular, but they have limited ability to capture simple graph structures [22].
- (GIN) [22] matches the power of the WL test, the most expressive.
- GINe [13] builds on GIN - pretrain on both local and global neighborhoods.
- (PNA) [5] graph isomorphism, countable feature spaces, continuous features, combination of aggregators.

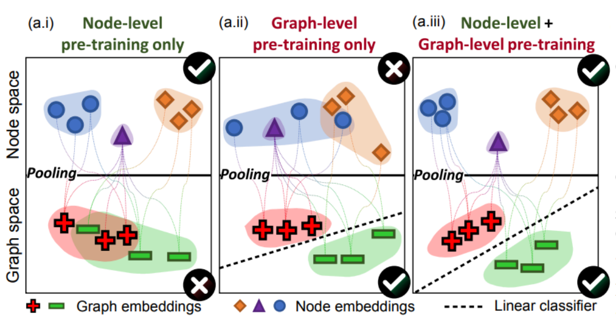


figure from [13]

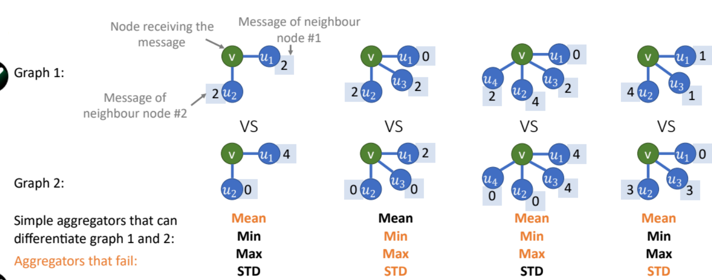


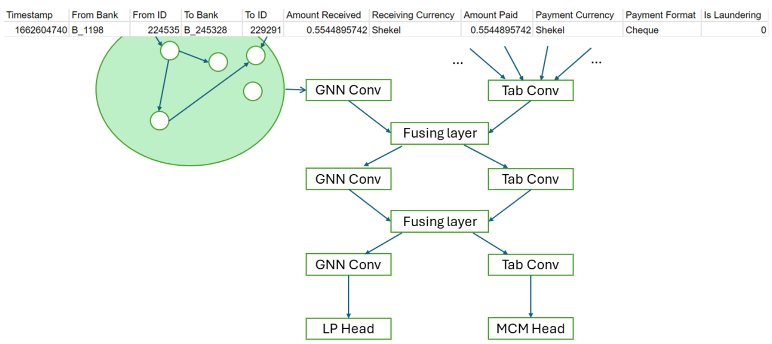
Figure 1: Examples where, for a single GNN layer and continuous input feature spaces, some aggregators fail to differentiate between neighbourhood messages.

figure from [5]

3.LIMITATIONS

- Transformers have been used together with GNNs, but that model is impossible to be used for tabular data [23]
- need to standardize performance measurement, to facilitate accurate and objective comparisons between different transformers [9] - could be addressed by using the standard benchmark for evaluating models [10].
- need to understand theoretical properties and limitations of GNNs [22] - framework for analyzing the expressive power of GNNs [22]

4.ARCHITECTURE



fused architecture with a transformer and a GNN

5.EXPERIMENTS

DELFT BLUE

- AML dataset[1]
- timeline: PNA -> Trompt -> GraphSage -> ResNet
- Accuracy: proportion of correctly predicted masked cells within the categorical features.
- hits@k: fraction of the positive edges are ranked in the first k positions

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \qquad RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

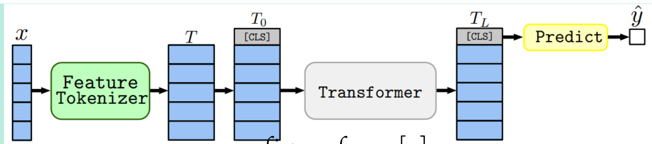
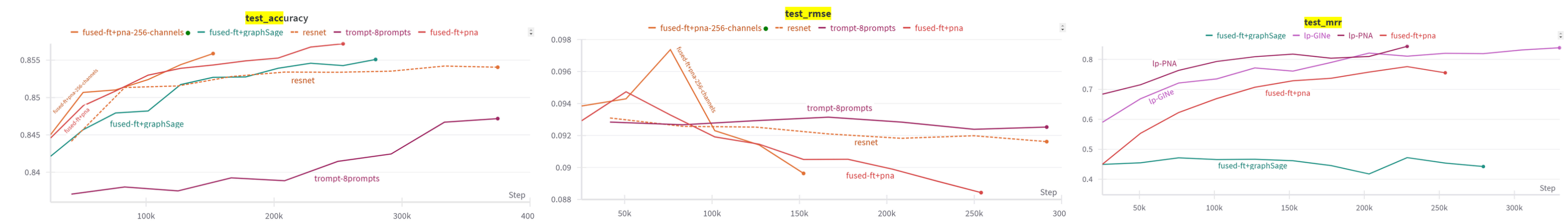


figure from [9]

6.RESULTS

- best for accuracy and RMSE : fused FT-Transformer and PNA
- best for MRR: PNA
- best transformers: FT-Transformer and ResNet very close performance
- more hidden layers better for fused - up to 512



7.FUTURE WORK

- run more epochs and more runs per experiment, parameter tune Trompt, generalize algorithm for different datasets, make algorithm more efficient, remove unnecessary data transfers between GPU and CPU, integrate XGBoost, use standard benchmark for measuring transformer performance

8.CONCLUSION

- transformers can learn from tabular data, GNNs - from graphs
- self-supervised learning on AML dataset
- combining tabular transformers with graph neural networks (GNNs) can enhance the predictive abilities on tabular data
- the FT-Transformer and GINe fused architecture's performance can be improved by integrating PNA, GraphSage underperforms because of its weakness in capturing simple graph structures
- future improvements can be made - optimize algorithm, integrate XGBoost, generalize algorithm to perform well on different datasets

REFERENCES

[1] Erik Altman, Jovan Blanuša, Luc Von Niederhäusern, Béni Egressy, Andreea Anghel, and Kubilay Atasu. Realistic synthetic financial transactions for anti-money laundering models. Advances in Neural Information Processing Systems, 36, 2024.
[2] Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In Proceedings of the AAAI conference on artificial intelligence, volume 35, pages 6679– 6687, 2021.
[3] Kuan-Yu Chen, Ping-Han Chiang, Hsin-Rung Chou, Ting-Wei Chen, and Tien-Hao Chang. Trompt: Towards a better deep neural network for tabular data. arXiv preprint arXiv:2305.18446, 2023.
[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794, 2016.
[5] Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets. Advances in Neural Information Processing Systems, 33:13260–13271, 2020.
[6] Béni Egressy, Luc Von Niederhäusern, Jovan Blanuša, Erik Altman, RogerWattenhofer, and Kubilay Atasu. Provably powerful graph neural networks for directed multigraphs. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 11838–11846, 2024.
[7] Heshan Fernando, Han Shen, Miao Liu, Subhajit Chaudhury, Keerthiram Murugesan, and Tianyi Chen. Mitigating gradient bias in multi-objective learning: A provably convergent approach. International Conference on Learning Representations, 2023.
[8] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. arXiv preprint arXiv:1903.02428, 2019.
[9] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. Advances in Neural Information Processing Systems, 34:18932–18943, 2021.
[10] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? Advances in neural information processing systems, 35:507–520, 2022.
[11] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. Advances in neural information processing systems, 30, 2017.
[12] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In International conference on machine learning, pages 2712–2721. PMLR, 2019.
[13] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. arXiv preprint arXiv:1905.12265, 2019.
[14] Weihua Hu, Yiwen Yuan, Zecheng Zhang, Akihiro Nitta, Kaidi Cao, Vid Kocijan, Jure Leskovec, and Matthias Fey. Pytorch frame: A modular framework for multi-modal tabular learning. arXiv preprint arXiv:2404.00776, 2024.
[15] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. Technologies, 9(1):2, 2020.
6] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016.
[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.
[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32, 2019.
[19] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. Information Fusion, 81:84–90, 2022.
[20] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning IEEE transactions on medical imaging, 35(5):1299–1312, 2016.
[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.
[22] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? arXiv preprint arXiv:1810.00826, 2018.
[23] Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy S Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining. Advances in Neural Information Processing Systems, 35:37309–37323, 2022.
[24] Bingxu Zhang, Changjun Fan, Shixuan Liu, Kuihua Huang, Xiang Zhao, Jincai Huang, and Zhong Liu. The expressive power of graph neural networks: A survey. arXiv preprint arXiv:2308.08235, 2023.
[25] Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. A survey on negative transfer. IEEE/CAA Journal of Automatica Sinica, 10(2):305–329, 2022.
[26] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. AI open, 1:57–81, 2020.