A replication study of transformer-based TabPFN for assessing the applicability of neural-network based solutions in tabular classification.

Kartikey Chauhan
501259284

Summary

Deep Learning has revolutionized the field of AI and led to remarkable achievements in applications involving image and text data. In particular, large transformer-based models trained on massive corpora, are disrupting machine learning in many areas. However, when it comes to tabular (a.k.a. structured) data, traditional machine learning methods, such as gradient-boosted decision trees have shown superior performance over deep learning. Recently, TabPFN [2] proposes a radical change to how tabular classification is done, introducing a pre-trained Transformer that is able to perform classification without training. This project aims to replicate and do an empirical review of TabPFN's claims, while potentially exploring avenues for further scaling and modifications based on latest advancements in the Transformer space.

Goal and Objectives

The goal of this project is met by achieving the following objectives.

- Replication of the transformer model creation and training.
- Comparative analysis of its classfication performance against traditional ML methods, AutoML systems, and boosting machines. Confirm if the models are more accurate, especially for datasets that did not appear in the paper that proposed them.
- Comparative analysis of its runtime performance against traditional ML methods, AutoML systems, and boosting machines. Answer how long do training and hyperparameter search take in comparison to other models?
- Analysis of the potential scaling and extension of the model through hyperparameter optimization, longer training times or other architectural possibilities. The current scale of TabPFN is very limited (up to 1000 training data points, 100 purely numerical features without missing values, 10 classes).

References

[1] Benjamin Feuer, Chinmay Hegde, and Niv Cohen. Scaling tabpfn: Sketching and feature selection for tabular prior-data fitted networks. https://arxiv.org/abs/2311.10609, 2023.

[2] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. https://arxiv.org/abs/2207.01848, 2023.

[3] Andreas Müller, Carlo Curino, and Raghu Ramakrishnan. Mothernet: A foundational hypernetwork for tabular classification. https://arxiv.org/abs/2312.08598, 2023.

[4] Guri Zabërgja, Arlind Kadra, and Josif Grabocka. Tabular data: Is attention all you need? https://arxiv.org/abs/2402.03970, 2024.