Author: Alexandru-Valer Păcurar
Supervisor: Elena Congeduti

GRAPH NEURAL NETWORKS TRAINING SET ANALISYS

What is the effect of the training data size?

INTRODUCTION

In the evolving landscape of urban mobility, accurate traffic prediction stands as a critical component for managing congestion, planning routes, and enhancing safety. With the advent of Graph Neural Networks (GNNs), researchers have unearthed powerful tools capable of capturing the complex dependencies of traffic systems, which are naturally structured as graphs [1].
By understanding how the amount of historical data used to train a GNN affects its performance, we enable smarter and more cost-efficient urbanistic plans for transportation.

RESEARCH QUESTION

This research project explores the impact of training data characteristics on model performance. The main question this experiment tries to answer is:
“What is the effect of reducing the volume of training data on GNNs ability to accurately forecast traffic?”

METHODOLOGY

This experiment uses the D2STGNN [2] model. This model is considered to be state of the art in the context of traffic forecasting, as it is among the best performing models over multiple widely used publicly available datasets. [3, 4].
Multiple training datasets are created by filtering data from METR-LA based on the measurements time stamps. METR-LA is a public dataset that contains the average traffic speed aggregated every 5 minute intervals from 207 loop detectors across the highways of LA County [5].
The experiments are conducted by training the model with the newly generated datasets. They span different time frames, from one week up to two months.
By comparing models trained with similar amount of data, we can infer how the temporal distance affects predictions. A comparison between models trained with variate amounts of data is performed to assess the effect of changing the training dataset size.

FINDINGS

Below are the plots that present the performance metrics obtained by training the model with various datasets. The selected metrics are: mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE).

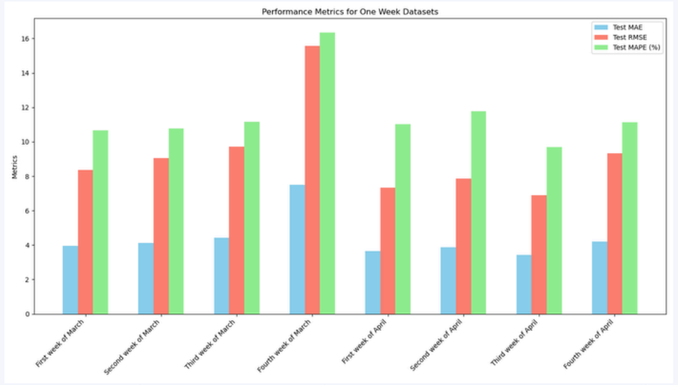

Figure 1: Performance comparison of multiple models trained with one week of historical data.

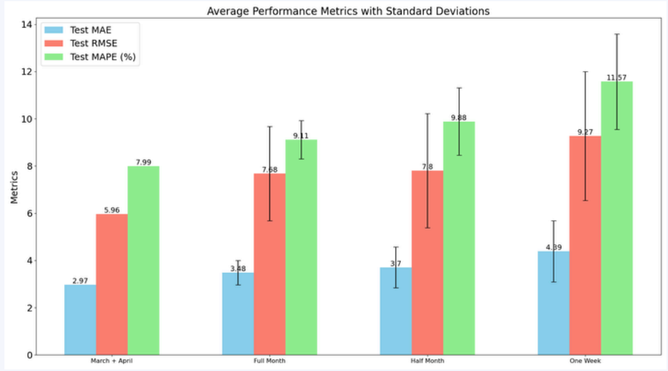

Figure 2: Comparison of the average of performance metrics grouped by amount of historical data.

OBSERVATIONS

Unexpectedly, the temporal distance between the measurements used in the training set and those used in the test set does not have a significant impact on the performance of the model. On the other hand, there are notable differences in the performance obtained by training the model with some specific datasets. These anomalies are the result of sparse data. During some days, all the sensors recorded average traffic speeds of 0 miles per hour. Training the model with datasets that contained larger amounts of zero values, or large blocks of consecutive zero values, lead to lower performance when compared to training with other datasets containing similar amount of historical data.

CONCLUSION

Based on the results, we can conclude that by assuring proper sensor maintenance, reducing the amount of historical data used to train GNNs for the task of traffic prediction leads to slight performance loss. While the loss of performance is unfavourable, it also reduces the time needed to gather data. By shortening the data gathering process, maintenance costs are also reduced.
To further improve the insight drawn from this research, performing similar experiments with different GNN models, datasets that contain traffic information from different locations and imputation techniques to make up for sensor errors can be performed.

REFERENCES

[1] I. R. Ward, J. Joyner, C. Lickfold, Y. Guo, and M. Bennamoun, “A practical tutorial on graph neural networks,” 2021.
[2] Z. Shao, Z. Zhang, W. Wei, F. Wang, Y. Xu, X. Cao and C. S. Jensen, “Decoupled dynamic spatial temporal graph neural network for traffic forecasting,” 2022.
[3] “Papers with Code - METR-LA Benchmark (Traffic Prediction).” https://paperswithcode.com/sota/traffic-prediction-on-metr-la?p=spatio-temporal-graph-convolutional-networks
[4] “Papers with Code - PEMS-BAY Benchmark (Traffic Prediction).” https://paperswithcode.com/sota/traffic-prediction-on-pems-bay?p=190600121
[5] Y. Li et al., “Diffusion convolutional recurrent neural network: Data-driven traffic forecasting,” in International Conference on Learning Representations, 2018.