Ahm Graphin' Hee-uh: Utilizing Spatio-Temporal GCNs in Subway Volume Prediction

Adin Pepper Fox, Matthew DaSilva, Christopher Parvankin

Motivation

New York City's Subway System faces volatile passenger volumes dependent on a number of variables — station servicing, time of day, weather, season, etc.

In our project, we sought to predict hourly passenger volumes for every NYC subway station utilizing Spatio-Temporal Graph Convolution Networks.

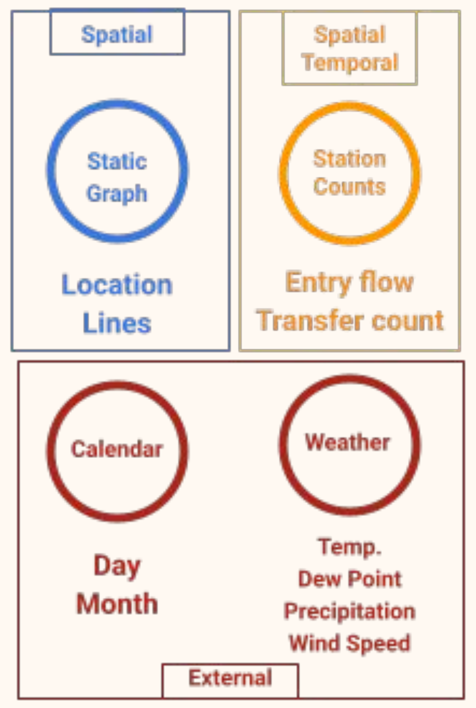
Through creating a means to reliably predict passenger volume, the MTA can optimize service scheduling and resource allocation.

Data

NYC Turnstile Data (*Temporal*) — Hourly ridership estimates for NYC subway stations from 2023-2024. Hourly estimates were normalized relative to each station.[2]

JFK Weather Data (Extrinsic) — Hourly meteorological from JFK Airport from 2023-2024. Data was normalized and appended with one-hot-encodings of week and month. [3]

NYC Subway GTFS-Realtime Feed (Spatial) — NYC Subway public transportation information and geographic information. Data was used to create graph representation of NYC subway.[4]



Architecture

Input Features:

- **Spatial features**: static graph representation (lines, station location-coords, borough...)
  Shape: (Number Nodes, Spatial Dimension)
- **Temporal features**: temporal graph representation (T-window of per-node entrances, transfers)
  Shape: (Number Nodes, Temporal Dimension, Number of Timesteps)
- **Weather features**: T-sequence of weather data (Temp, dew point, precipitation, wind speed)
  Shape: (1, Number of Timesteps, Weather Dimension)
- **External features**: Sequence of daily temporal features that apply to whole graph (day of the week, average ridership in subway, bus system, LIRR.
  Shape: (1, External Dimension)

Loss:

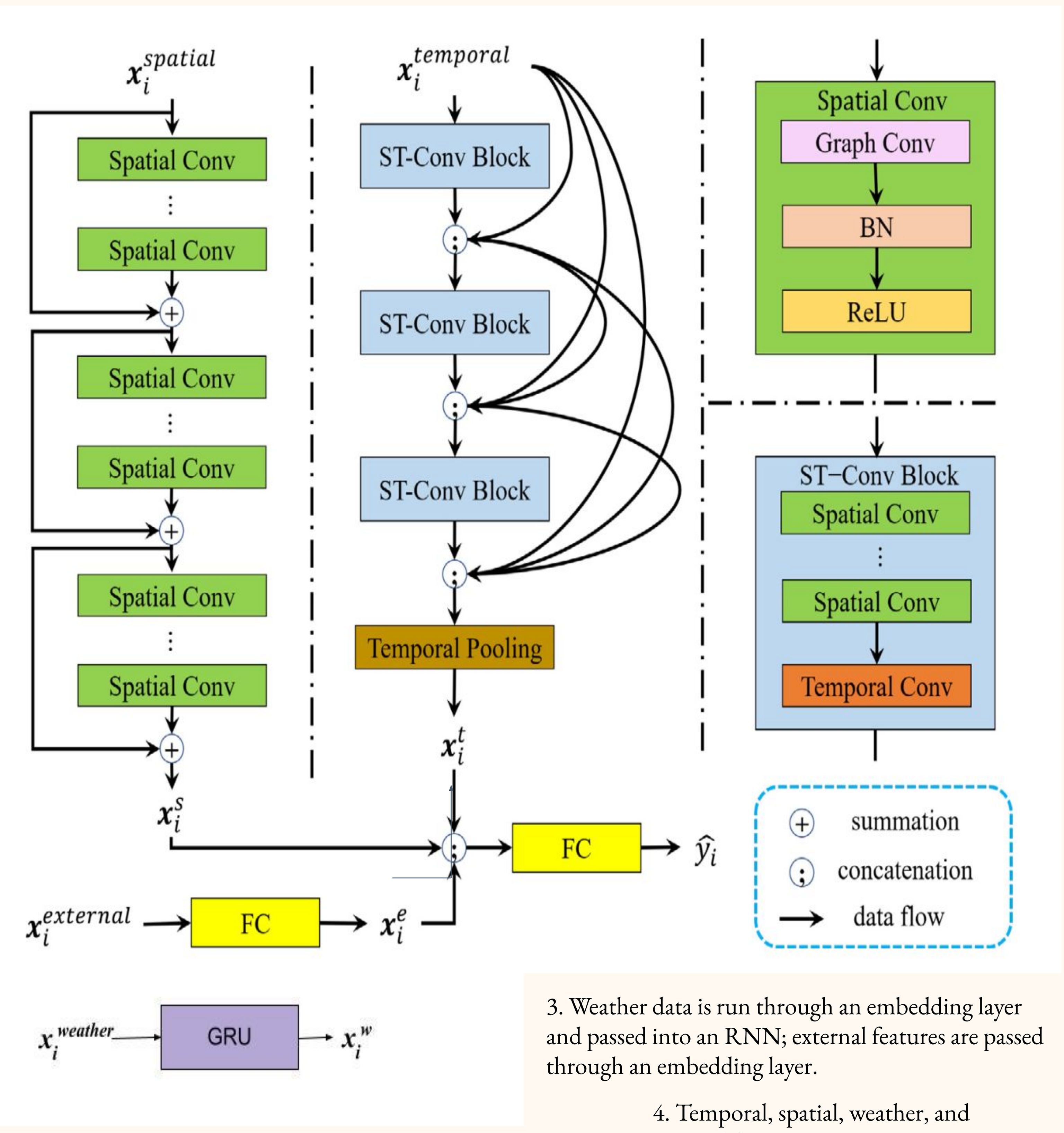
- **Mean Absolute Error :** $\frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$

Metrics:

- **Pearson Correlation Coefficient:** $\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$

Deep Spatio-Temporal Graph Convolution Network Model

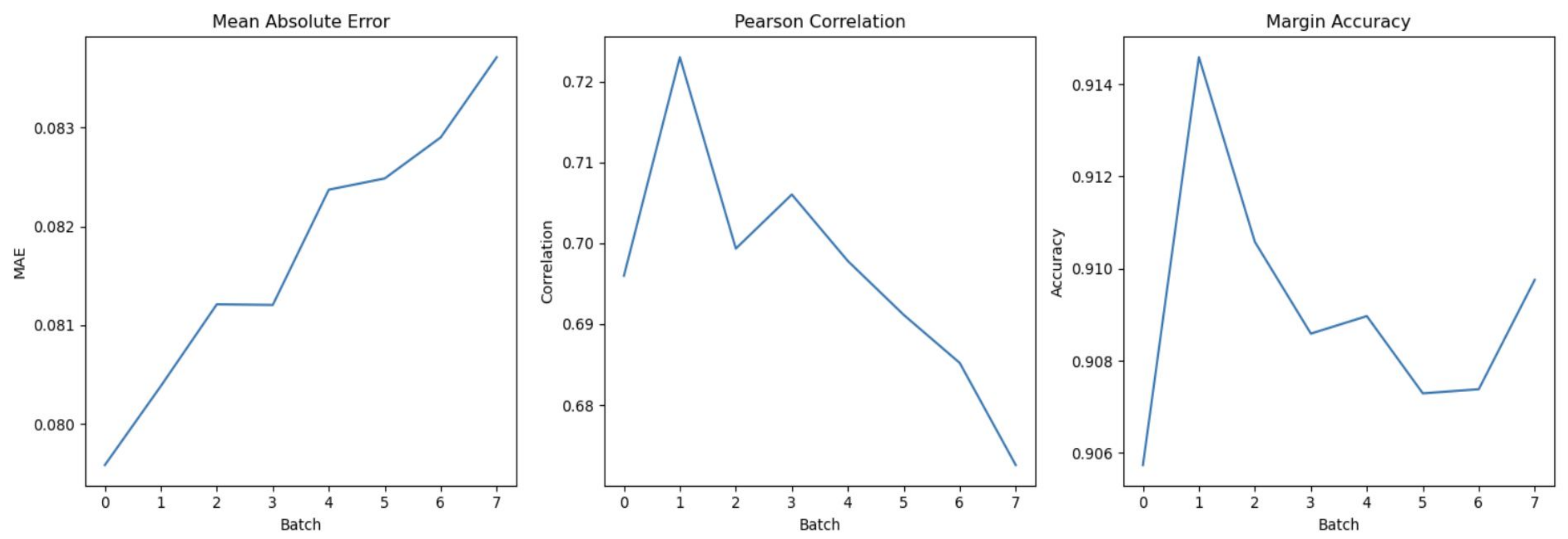
[1]



Architecture Outline:

1. Spatial features run through an embedding layer then fed into a series of GCNs with residual connections.
2. Temporal features are run through a series of temporal blocks. Each block takes the temporal features and performs a graph convolution through a convolution layer. The output of each block is appended for each graph in the timestep; these outputs are stacked & run to the input features of the next; the final output is then pooled.
3. Weather data is run through an embedding layer and passed into an RNN; external features are passed through an embedding layer.
4. Temporal, spatial, weather, and external feature outputs are concatenated and run through a final fully-connected layer.

Results



Results from Testing (8 Batches of Batch Size 128)

Average Mean Absolute Error: 0.08173170

Average Pearson Correlation: 0.69638177

Average Margin Accuracy (5%): 0.90911074

Takeaways and Discussion

In looking back, finding reliable, open-source data for our model was the one of the most difficult parts of this project. The MTA was one of a few transportation authorities that provided such comprehensive data, which helped smoothen our preprocessing steps; this data still had issues, particularly, there were instances where data was missing or corrupted that we had to account for.

Utilizing Tensorflow, as opposed to the PyTorch, which was used in the paper we were inspired by, had several issues as well during our implementation of the model. While Tensorflow helped streamline the making of the model, it's lack of customization, compared to things like PyTorch's fine-grained control over GPU usage, made the adoption the paper's model more difficult.

Ultimately, even with testing from 8 batches, we saw an average loss of .08, average Pearson correlation of .7, and average margin accuracy of .91. Our team is satisfied with these results. If we were to improve upon our work, we would be interested in incorporating further predictive variables, such as points of interest (e.g theater venues, restaurants, etc.), or future dates of interest (holidays, parades, etc.), all of which could provide better insight into the various forces that drive ridership volume.

References

[1] Yu, Le, Bowen Du, Xiao Hu, Leilei Sun, Liangzhe Han, and Weifeng Lv. "Deep Spatio-Temporal Graph Convolutional Network for Traffic Accident Prediction." *Neurocomputing* 423 (n.d.): 135–47. https://doi.org/https://doi.org/10.1016/j.neucom.2020.09.043.

[2] Metropolitan Transportation Authority. MTA Subway Hourly Ridership: 2020-2024. https://data.ny.gov/Transportation/MTA-Subway-Hourly-Ridership-2020-2024/wujg-7c2s/about_data

[3] National Weather Service. NYC, Kennedy Airport Weather Observations. https://forecast.weather.gov/data/obhistory/KJFK.html

[4] Data.gov. New York City GTFS. https://catalog.data.gov/dataset/mta-general-transit-feed-specification-gtfs-static-data

Acknowledgements

Thank you to CSCI 1470 TA Armaan Patankar, for his "enthusiasm", "charm", and "love for science"!
Thank you NYC DOT and MTA for their data access!
Thank you to JFK!