Predicting Bicycle Demand by Incorporating Station-Specific Spatial Features in a Graph Convolustion Networks Model

Jonghyun Choi[1], Kyungtaek Oh[1], Seungmin Jang[1] and Keungoui Kim*[2]

[1] School of Data Science, Handong Global University, Pohang, Gyeongbuk, South Korea, {21700741,walkedwith,22000624}@handong.ac.kr

[2] AI Convergence Institute, Handong Global University, Pohang, Gyeongbuk, South Korea, awekim@handong.edu

Abstract

The primary aim of this study is to predict daily bicycle rental volumes at each station, addressing supply-demand imbalances. Previous research has primarily focused on predicting rental volumes using weather data and historical patterns. Recent studies have shown improvements by incorporating graph data to model station relationships. However, the comprehensive characterization of node properties through graph data has been overlooked. Therefore, a new Graph Convolution Network is introduced, depicting spatial station traits. Each Seoul station is considered a node, revealing inter-station connections through rental records. Matrices are employed to define spatial traits and past rental patterns. The output, streamlined into a one-dimensional form, integrates weather variables into a Fully Connected Network (FCN) for efficient daily bicycle demand prediction, outperforming other models.

Keywords— *Bicycle Demand Prediction, Bicycle Rental Network, Graph Convolution Network, Spatial Feature Matrix*

I. Introduction

The burgeoning global interest in eco-friendly and healthy lifestyles has spurred a rise in demand for shared bicycle systems worldwide[8], highlighting the critical issue of supply-demand imbalances[10]. Traditional manual rebalancing methods, reliant on labor-intensive truck deployments, underscore the need for accurate real-time demand prediction to mitigate these disparities. In Seoul, the rapid expansion of the public bicycle system, 'Ttareungi', without proper demand analysis, has led to increased user dissatisfaction due to supply-demand mismatches, necessitating precise demand forecasting for effective bicycle station placement and public bicycle distribution.

The complexity of predicting bicycle demand stems from the non-uniformity of rental and return locations, requiring consideration of the interconnectivity between stations, especially when forecasting across all of Seoul's rental stations. Additionally, spatial factors such as nearby bicycle paths, transportation facilities, and residential areas must be integrated with demand and weather variables to accurately influence rental volumes.

In our study, a graph-based model has been constructed where Seoul's bicycle stations are nodes connected by edges representing rental volumes, capturing the complex inter-station relationships. This model, a Graph Convolutional Network (GCN), integrates spatial and temporal data, including weather conditions, into its input matrix for a holistic spatio-temporal analysis. Subsequently, a Fully Connected Network (FCN) leverages this integrated data to forecast daily rental demand with enhanced accuracy over traditional models.

In the realm of public bicycle demand prediction [5] have contributed significantly with their research on Graph Convolutional Neural Networks (GCNs). Their study introduced a framework that not only reflects various spatiotemporal characteristics but also considers the influence of global variables, demonstrating robustness against sudden environmental changes. Their model, particularly the GCN-UP variant, showed superior performance by emphasizing the impact of users' past usage patterns over mere geographical proximity. However, they acknowledged the limitation of their model to existing stations and highlighted the need for future research to extend the graph structure to new stations.

Building on the foundational work by Bruna et al[1] who first introduced the graph convolutional neural network concept in their spectral networks study, further developed this concept. They proposed a GCN framework tailored for hourly demand prediction of bike-sharing usage at the station level. Their innovative approach utilized two distinct graph structures, GCN–IDW and GCN–UP, to encapsulate different spatial properties and to benchmark performance against each other[9].

Additionally, the work of Defferrard et al. expanded upon the GCN model with fast localized convolutions, significantly accelerating the computational process. This advancement has been instrumental in enhancing the efficiency of GCN models, as evidenced by the improved performance of the GCN frameworks in Kim, Lee, and Sohn's research[2].

II. DATA COLLECTION AND PREPROCESSING

A. *Bicycle rental history data*

The daily usage information data for Seoul's public bike rental stations was downloaded from 'Seoul Open Data Plaza.' Rental information for each station is contained in this dataset, and preprocessing was performed to group this information by rental station, calculating the daily rental counts for each station.

B. *Spatial data*

Spatial data was employed to extract information about facilities within a 500-meter radius[14] of Seoul bicycle stations. Data downloaded from 'Seoul Open Data Plaza' and 'Smart Seoul Map' included information about bus stops, subway stations, schools, universities, apartments, cultural facilities, and large academies[6]. Spatial data preprocessing was performed to determine the count of these facilities around each bicycle station. Buildings that closed or were demolished before 2022 and those that were constructed or opened after 2022 were excluded.

Regarding the processing of bicycle roads in the spatial data, roads with a radius of 500 meters centered around bicycle stations were formed by connecting bending points. Two scenarios were considered: in the first scenario, roads were counted if at least one bending point was within a 500-meter radius to prevent duplicates. In the second scenario, when all bending points were outside the 500-meter radius, linear equations and the distance from the bicycle station to the intersection point were used to count roads that were less than 500 meters. These classified road variables were divided into A and B types, separate and unseparated, and categorized according to road characteristics as shared roads, dedicated roads, priority roads, and car-only lanes, considering the significance of bicycle priority roads[4].

Through this preprocessing, spatial data consisting of 15 variables within a 500-meter radius for each bicycle rental station was generated.

C. *Time series data*

Seoul's ASOS data was obtained from the 'Korea Meteorological Administration' to acquire nine weather variables for demand prediction. While the precipitation variable had numerous missing values, we handled them by replacing NA with 0, considering it as the absence of rain.

This preprocessing resulted in a time-series format containing nine weather variables spanning from January 1, 2022, to December 31, 2022.

III. METHODOLOGY

A. *Framework: GCN - pro*

The general framework of Graph Convolutional Network (GCN) models typically comprises nodes and edges representing relationships between nodes[13]. In this study, bicycle stations are designated as nodes, while the relationships between stations are defined as edges.[8] Notably, the relationships between stations are structured using historical data on bicycle rentals and returns[5]. The proposed GCN model in our paper incorporates novel spatial characteristics specific to individual stations, a feature absent in prior literature. It predicts bicycle demand by integrating station-to-station relationships, past bicycle rental records, and contextual spatial data such as proximity to subway stations, bus stops, and residential areas. The nomenclature used to denote this model is 'GCN-pro.'

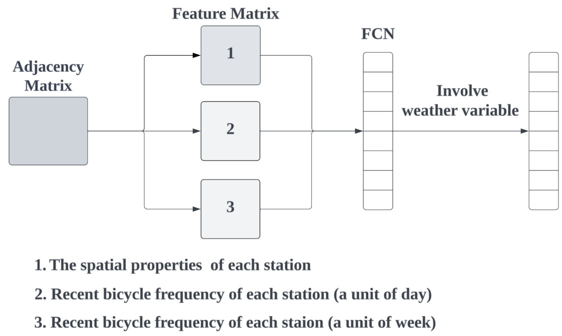

Fig. 1. Schema of the GCN - pro

The matrices, which represent the spatial information of nearby subway stations, bus stops, and residential facilities, as well as the matrices containing data for bicycle rentals in the preceding days and weeks, were separately constructed. Each matrix was autonomously processed by the GCN model to learn its distinct features. To construct a unified model integrating all these features, the outputs from each model were elongated into one-dimensional shapes and concatenated. Consequently, the model was designed to learn the spatial characteristics of stations and historical bicycle usage patterns. Weather variables for the prediction date, such as 'precipitation' and 'temperature', were allocated. These components were integrated into a Fully Connected Network (FCN) model, resulting in the calculation of the anticipated demand for each station on the predicted date.

B. Graph convolution network

The utilization of the Graph Convolution Network (GCN) model serves as a fundamental tool in addressing complex traffic issues[12]. The decision to adopt GCN is rooted in its distinct capability to capture and analyze the intricate relationships between stations present in bicycle rental systems. Previous studies have also recognized the potential of this model and sought to predict bicycle demand using the GCN model.

C. Adjacency Matrix: Interaction among stations

The interactions between bicycle stations are represented using the total historical records of bicycle rentals and returns. The adjacency matrix is structured so that rows and columns correspond to station numbers with identical station numbers. Rows represent the rental station numbers, while columns depict the return station numbers. The adjacency matrix is constructed by summing the occurrences of rental and return station number pairs across all historical rental and return records.

$$A_{ij} = \sum RR_{ij}$$

A_{ij} denotes the adjacency matrix containing information on rentals from station i and returns to station j, while $\sum RR_{ij}$ signifies the total count of bicycles moved from station i to station j.

$$A_{ij}^{\text{scaled}} = \frac{(A_{ij} - \text{Min})}{\text{Max} - \text{Min}}$$

Here, A_{ij}^{scaled} denotes the scaled matrix, where Max and Min represent the maximum and minimum values within matrix A_{ij}. The scaled matrix, denoted as $\tilde{A}$ (derived from A_{ij}^{scaled}), is further utilized to form the normalized symmetric Laplacian matrix.

$$L^{\text{sym}} = \tilde{D}^{(-1/2)}\tilde{A}\tilde{D}^{(-1/2)}$$

$\tilde{D}$ is the diagonal matrix derived from the matrix $\tilde{A}$, and L^{sym} is the normalized symmetric Laplacian Matrix obtained by multiplying both sides of the scaled adjacency matrix with the power of -1/2 of the diagonal matrix. The construction of a symmetric Laplacian Matrix serves the purpose of efficient information management when performing operations with the feature matrix in eigenvalue-eigenvector forms[11].

D. Feature Matrix 1: spatial feature by station level

This research introduces a novel approach to constructing spatial matrices that have not been explored previously. It includes the spatial characteristics specific to individual bicycle stations, encapsulating parameters such as the number of transportation facilities, residential households, and other related features within a 500-meter vicinity of each station. The rows of the matrix correspond to a total unique bicycle stations, while the columns represent 15 distinct spatial characteristics.

$$F_{\text{spatial}}^{\text{scaled}} = \frac{F_{\text{spatial}} - \text{Min}}{\text{Max} - \text{Min}}$$

The spatial matrix F_{spatial} is transformed into values between 0 and 1 using min-max scaling, aiming to prevent the model from biased computations towards specific features.

E. Feature Matrix 2: Historical Bicycle Rental Records

One of the most crucial attributes in predicting bicycle demand is the incorporation of historical rental records. Bicycle rental history amalgamates daily rental records from a few days preceding the prediction date and weekly rental records from several preceding weeks. Rows represent distinct station numbers, while columns encapsulate information regarding past rental quantities. 't' is assumed as the forecast date for demand, and the matrix illustrates the rental history preceding 't'. For the daily historical rental matrix, it showcases the rental volumes for several days before 't'. Regarding the weekly historical rental matrix, it presents the rental volumes over several weeks from 't'.

F. GCN propagation

In this chapter, the process of propagation will be discussed, delving into how the matrices created earlier are used for training within the model.

$$H_1^{(0)} = F_{\text{spatial}}^{\text{scaled}}$$

$$H_2^{(0)} = F_{\text{day}}$$

$$H_3^{(0)} = F_{\text{week}}$$

$F_{\text{spatial}}^{\text{scaled}}$ is designated as the first layer of matrix H_1, F_{day} as the first layer of matrix H_2, and F_{week} as the first layer of matrix H_3.

$$H_1^{(l+1)} = \sigma(\hat{D}^{(-1/2)}\hat{A}\hat{D}^{(-1/2)}H_1^{(l)}W^{(l)})$$
$$H_2^{(l+1)} = \sigma(\hat{D}^{(-1/2)}\hat{A}\hat{D}^{(-1/2)}H_2^{(l)}W^{(l)})$$
$$H_3^{(l+1)} = \sigma(\hat{D}^{(-1/2)}\hat{A}\hat{D}^{(-1/2)}H_3^{(l)}W^{(l)})$$

Where $H^{(l)}$ denotes the feature matrix of the lth layer, $W^{(l)}$ represents the weights of the lth layer, and σ refers to an activation function like ReLU. The symmetric

Laplacian matrix operates on the feature matrix of the lth layer along with the weights, computing the feature matrix of the $l+1$th layer through the activation function.

G. Weather variable

After the propagation in the previous GCN model, the output is a two-dimensional matrix with the same shape as the feature matrix. This is elongated into a one-dimensional shape and concatenated with two weather variables: 'precipitation' and 'temperature'.

$$\hat{y} = \sigma(W_1H_1 + W_2H_2 + W_3H_3 + W_4x_1 + W_5x_2)$$

Here, $\hat{y}$ denotes the predicted bicycle demand. H_1 refers to the output of the GCN model that integrates the spatial matrix, stretched into a one-dimensional shape. Meanwhile, H_2 and H_3 represent the outputs of the GCN model encompassing the daily and weekly historical rental matrices, respectively, also stretched into one-dimensional shapes. x_1 and x_2 signify 'precipitation' and 'temperature', while W_1, W_2, W_3, W_4, and W_5 symbolize the model's weights. [5]The evaluation criterion utilized in this study was the Root Mean Square Error (RMSE) loss function.

H. Sliding window

Segmenting train and test sets in time series data presents a challenge. The demarcation of train and test sets was resolved using the sliding window technique in this study. Through this approach, the model was trained using the sliding window method to predict daily demand by merging the spatial matrix and historical rental matrices into a feature matrix. The size of the sliding window was set as a hyperparameter. The test set was defined as the date immediately following the last date used for training through the sliding window.

IV. Experimental Settings

In this study, various hyperparameters were adjusted to enhance and optimize the performance of the developed GCN model. These hyperparameters hold significant influence over the model's structure and learning process, playing a crucial role in tuning and improving the model's performance.

The following are the settings and related explanations of key hyperparameters:

- GCN Layer: The size of the GCN model's layer, adjusted for efficient learning in the Graph Convolutional Neural Network.
- FCN Hidden Layer: The size of the hidden layer in the Fully Connected Network (FCN), adjusted to determine model density and complexity.
- Learning Rate: Determines the speed at which the model's weights are updated, utilized to regulate the model's convergence and learning speed.
- Sliding Window Size: The size of the considered sliding window in time series data, adjusted considering the temporal relationships within the data.
- Historical Daily Rental Window Size: The window size used by the model to comprehend historical daily rental volumes, adjusted to consider daily patterns.
- Historical Weekly Rental Window Size: The window size used by the model to comprehend historical weekly rental volumes, adjusted to consider weekly patterns.
- Epochs: The number of times the learning algorithm iterates over the entire dataset, utilized to regulate model training and convergence.
- Weight Decay: A regularization parameter used to control model complexity, adjusted to prevent overfitting and improve generalization performance.

These hyperparameters significantly influenced the model's learning and predictive performance, playing a critical role in evaluating the results of this study.

The GCN model developed in this study was implemented using PyTorch and Python. Hyperparameter tuning involved researchers specifying all parameter combinations three times to identify the combination that resulted in the lowest average RMSE value for the test set. To reduce the number of parameter combinations, several measures were taken. The GCN's hidden layer count included parameters of up to 4[7], while the FCN hidden layer count was manually set to 2 after identifying no significant difference when the hidden layer count exceeded 2. The sizes for the historical daily and weekly rental windows were chosen based on criteria outlined in previous literature[5], selecting daily rental window sizes from 3 to 5 days and weekly rental window sizes from 1 to 5 weeks.

GCN hidden layer	2
FCN hidden layer	2
Sliding window	14
Day	3
Week	3
Epoch	500
Learning rate	0.001

Table 1. The optimal hyperparameter combination

The table represents the optimal combination of hyperparameters. Both the GCN and FCN models employed the rectified linear unit (ReLU) as the activation function, and Adam Optimizer was utilized as the optimizer.

V. RESULT

In this study, a comparative analysis was conducted to assess the performance of the GCN-pro model against other models. Differences in performance were investigated among models utilizing distinct types of data for demand prediction. Specifically, evaluations were made of models based on time series data, including LSTM[3], SARIMAX, and XG-Boost, compared to our GCN-pro model, which integrates spatial data. Additionally, the involvement of the GCN-UP model, leveraging graph data in previous research, was considered.[5] The RMSE values for each model are presented below.

model	RMSE
GCN-pro	13.40
GCN-UP	14.65
LSTM	16.90
ARIMA	20.3
XGBoost	16.5

Table 2. RMSE of models

The results, which compared and optimized the best hyperparameters that each model could offer, indicated the performance of the LSTM, well known for its suitability in time series forecasting, over ARIMA and XG-Boost. It was observed that the GCN models, considering the relationships at each station, outperformed all these models. Additionally, the performance of the GCN model that incorporated spatial features not accounted for in GCN-UP, namely GCN-Pro, was the most favorable. These comparisons were based on training data from May and predicting the final day of May, showcasing GCN-Pro's superior performance.

VI. CONCLUSION

In this paper, we introduce a novel Graph Convolutional Network (GCN) model that integrates spatial features with rental volume and weather data to predict demand at each bicycle station. This model adopts a new approach by combining spatial characteristics with spatial features, demonstrating superior performance compared to existing models such as LSTM, ARIMA, XGBoost, and GCN - UP. However, GCNs are constrained by their reliance on established graph structures, which hinders their ability to adapt to and include new stations in their analyses. The methodology presented in this study is expected to make a significant contribution to future research on spatial feature analysis using GCN - pro models.

REFERENCES

[1] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

[2] Longbiao Chen, Daqing Zhang, Leye Wang, Dingqi Yang, Xiaojuan Ma, Shijian Li, Zhaohui Wu, Gang Pan, Thi-Mai-Trang Nguyen, and Jérémie Jakubowicz. Dynamic cluster-based over-demand prediction in bike sharing systems. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 841–852, 2016.

[3] Sathishkumar V E and Yongyun Cho. A rule-based model for seoul bike sharing demand prediction using weather data. *European Journal of Remote Sensing*, 53(sup1):166–183, 2020.

[4] HWANBI JI. Predicting district-wise bike-sharing demand using machine learning and deep learning algorithm case study of seoul (ddareungi), 06 2022.

[5] Tae San Kim, Won Kyung Lee, and So Young Sohn. Graph convolutional network approach applied to predict hourly bike-sharing demands considering spatial, temporal, and global effects. *PloS one*, 14(9):e0220782, 2019.

[6] Tae San Kim, Won Kyung Lee, and So Young Sohn. Graph convolutional network approach applied to predict hourly bike-sharing demands considering spatial, temporal, and global effects. *PLOS ONE*, 14(9):1–16, 09 2019.

[7] Drew Ligman. Improved hyperparameter tuning for graph learning with warm-start configuration. 2021.

[8] Lei Lin, Zhengbing He, and Srinivas Peeta. Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. *Transportation Research Part C: Emerging Technologies*, 97:258–276, 2018.

[9] Tae San Kim, Won Kyung Lee, and So Young Sohn. Correction: Graph convolutional network approach applied to predict hourly bike-sharing demands considering spatial, temporal, and global effects. *PloS one*, 17(3):e0266221, 2022.

[10] Divya Singhvi, Somya Singhvi, Peter I Frazier, Shane G Henderson, Eoin O'Mahony, David B Shmoys, and Dawn B Woodard. Predicting bike usage for new york city's bike sharing system. In *AAAI Workshop: Computational Sustainability*, 2015.

[11] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *International conference on machine learning*, pages 6861–6871. PMLR, 2019.

[12] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

[13] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1):1–23, 2019.

[14] Qingfeng Zhou, Chun Janice Wong, and Xian Su. Machine learning approach to quantity management for long-term sustainable development of dockless public bike: Case of shenzhen in china. *Journal of Advanced Transportation*, 2020:1–13, 2020.

Summary of this paper

A. Problem Setup

The escalating demand for shared bicycle systems globally underlines the pressing issue of supply-demand imbalances. Rebalancing methods reliant on labor-intensive truck deployments have amplified the need for precise demand predictions. In Seoul, the rapid expansion of the public bicycle system, 'Ttareungi', without adequate demand analysis has resulted in user dissatisfaction due to supply-demand mismatches, necessitating accurate demand forecasting for effective bicycle station placement.

B. Novelty

In the pursuit of solving the demand forecasting problem, a new Graph Convolution Network is proposed in this study. Numerous research efforts have aimed to predict bicycle rental volumes by considering weather variables and applying historical bicycle rental pattern analysis to various models. Among these endeavors, emerging papers have demonstrated superior performance compared to previous models by incorporating graph data to reflect relationships between bicycle stations. However, one perplexing aspect is the lack of attempts in leveraging the significant advantage of graph data—describing the specific characteristics of each node—to comprehensively address this issue. Therefore, a new Graph Convolution Network that incorporates spatial features influencing bicycle demand at each station is proposed.

C. Algorithms

The proposed Graph Convolution Network (GCN) conceptualizes bicycle stations as nodes, elucidating their relationships through historical rental and return data. This study integrates distinctive spatial attributes specific to individual stations, enabling the anticipation of bicycle demand by amalgamating station-to-station correlations, previous rental archives, and pertinent spatial data. An adjacency matrix delineates the inter-station dynamics using historical rental and return statistics. Rows and columns correspond to the rental and return station numbers, encapsulating their historical association. This research introduces a unique spatial matrix inclusive of factors like transportation facilities, residential localities, and educational facilities proximate to each station. Integrating this spatial data with historical bicycle rental records results in the complete feature matrix. The GCN model integrates weather variables—such as 'precipitation' and 'temperature'—with its output, culminating in a streamlined one-dimensional output. This, fused with weather data, effectively forecasts daily demand in Fully Connected Network. The model's training and test segmentation utilizes a sliding window technique to predict daily demand.

D. Experiments

Hyperparameters were fine-tuned to enhance the GCN model's performance. Using PyTorch and Python, adjustments were made through multiple iterations, limiting the GCN's hidden layer to 4 and FCN's to 2, based on minimal improvements observed beyond these numbers. Historical rental window sizes were selected in line with prior literature, varying from 3 to 5 days for daily windows and 1 to 5 weeks for weekly windows. ReLU is used for activation function and the chosen optimization method employed was the Adam Optimizer. A comparative analysis, involving models like LSTM, SARIMAX, XG-Boost, and GCN-UP contrasted with the spatially integrated GCN-pro model. Results highlighted GCN-Pro, incorporating distinctive spatial features, demonstrated superior performance compared to any other models. These comparisons, using May's training data to predict the month's final day.