Clustering Mobile Traffic Data with Autoencoder Using Time-Series Encoded as Images

Sang-Yeon Lee, Hyun-Min Yoo, Jong-Seok Rhee, Geon Kim, Een-Kee Hong
Department of Electronics and Information Convergence Engineering
Kyung Hee University
Yongin-si, 17104, South Korea
{sangyeon, yhm1620, howrhee, gun, ekhong}@khu.ac.kr

Byungsuk Kim, Kyeongjun Shin
Network Group
KT
Seongnam-si, 13606, South Korea
{b.kim, shin.kj}@kt.com

***Abstract*—With each new generation of mobile communication, the mobile traffic data are increasing exponentially. However, there are limitations to the capacity of base stations. As a result, it become necessary to manage mobile traffic efficiently. It will result in the understanding of the characteristics of mobile traffic, yet it is challenging to find research based on base station traffic volume. In this paper, we encode mobile traffic data of base station from multiple locations to image, and cluster the encoded images using autoencoder. Based on the labeled locations of several base stations we directly investigated, we confirm whether the unlabeled base stations are well grouped into shopping malls, offices, residential areas, and outside areas.**

***Index Terms*—image encoding, autoencoder, clustering, mobile traffic**

I. Introduction

According to Ericsson's report, the amount of global mobile traffic totaled 118.21 EB (exabytes) in the fourth quarter of 2022, marking an increase of approximately 10% from the first quarter. Two years ago, in the fourth quarter of 2020, it was 58.44EB, which more than doubled compared to that time [1]. However, increasing the number of base stations to accommodate the exponential growth in mobile traffic is a big challenge in terms of cost and efficiency." Considering this, it is necessary to analyze the pattern of mobile traffic data volume of base stations and understand their characteristics. There are several traffic pattern analysis studies about traffic data generated at terminals, or change in base station traffic patterns based on the inflow and outflow of terminals. However, there are few studies that analyze the pattern of mobile traffic based solely on the base stations. In this paper, we analyze the characteristics of traffic data volume for each base station and cluster them according to the data patterns using autoencoder.

II. Data And Method

A. Dataset

The dataset is traffic volume of base station in Gangnam station, Coex, Gwanghwamun, built by Korea mobile network operator A. The dataset represents hourly data in a period of one month from 88 base stations of Gangnam station, 38 of Coex, and 8 of Gwanghwamun. Using these base stations, we aimed to understand the traffic patterns of office buildings, shopping malls, residential areas, and roads (i.e. outside area). Among them, we selected several base stations to determine their actual locations in order to use them as labeled data for interpretation of entire dataset. We investigated two base stations in residential areas, three base stations in road at Gangnam Station, seven base stations in Gwanghwamun, We also investigated five base stations of office building and eleven base stations of shopping mall in Coex.

B. Image Encoding

Image encoding means converting any form of data into an image format. As the dataset used in this paper is time-series data with interval of one hour, we convert this time-series data to image data. There are three main reasons for encoding time series data into images. First, we can easily identify the characteristics and patterns of visualized time-series data compared to original one. Second, we can use advanced deep learning models and techniques in the field of image processing. Third, by leveraging these models, we can improve the clustering accuracy of time-series data that has been converted into images. Fig. 1 and Fig. 2 show that image is better than time-series for understanding the pattern of data.

There are many ways to convert time series into images, such as Recurrence Plot (RP), Markov Transition Field (MTF), and Gramian Angular Field (GAF) [2]. RP is method that plot time-series data on m-dimensional space orbit, and convert time-series to image using plotted dot in the space orbit [3]. MTF can capture the transition probability of discretized time-series data. MTF can reduce the time dependency and make matrix - is corresponding to length of time-series data - that each element (i, j) means transition probability from the $i-th$ time series value to the $j-th$ time series value. Finally, MTF represents the transition probabilities over time.

In this paper, we leverage the GAF algorithm. GAF is a method that represents the temporal correlation between each time point of time-series data based on polar coordinates. Depending on whether it represents the difference or the sum

This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, through the Information Technology Research Center (ITRC) Support Program, supervised by the Institute of Information & Communications Technology Planning & Evaluation (IITP), under Grant IITP-2021-0-02046.

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

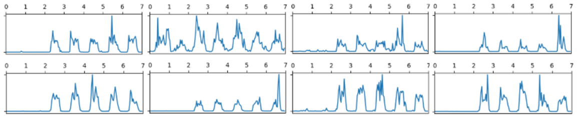

Fig. 1. Traffic pattern of Gwanghwamun base stations (time-series).

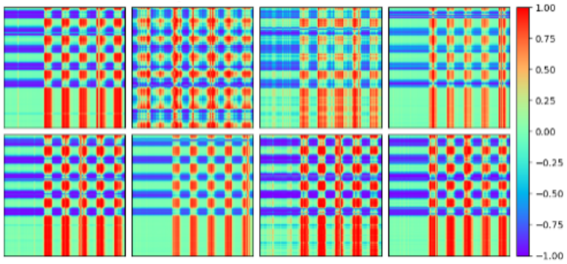

Fig. 2. Traffic pattern of Gwanghawmun base stations (image).

of angles, it can be classified into Gramian Angular Difference Field (GADF) or Gramian Angular Summation Field (GASF). The equation below represents the equation of converting time series data into images using the GAF algorithm:

$$X = [x_1, x_2, ..., x_n] \tag{1}$$

$$\tilde{x_l} = \frac{(x_i - max(X)) + (x_i - min(X))}{max(X) - min(X)} \tag{2}$$

Let x_i denotes the $i{-}th$ value of time-series X. Equation (2) normalizes x_i to be between –1 and 1. If the normalized value is $\tilde{x_l}$, we can get polar coordinated angle ϕ_i with equation (3),

$$\phi_i = \arccos(\tilde{x_l}), -1 < \tilde{x_l} < 1 \tag{3}$$

$$G = \begin{bmatrix} \cos(\phi_1 + \phi_1) & \cdots & \cos(\phi_1 + \phi_n) \\ \cos(\phi_2 + \phi_1) & \cdots & \cos(\phi_2 + \phi_n) \\ \vdots & \ddots & \vdots \\ \cos(\phi_n + \phi_1) & \cdots & \cos(\phi_n + \phi_n) \end{bmatrix} \tag{4}$$

Among all time points, we can arbitrarily select two time points and calculate the sum or difference of angles between them. The GASF, calculating the sum of angles, can generate a 2-dimensional matrix G as shown in Equation (4). We can visualize this matrix as an image. When clustering the images converted by GAF, it exhibit the most distinct visual patterns.

C. Autoencoder

An autoencoder is an artificial neural network model with an unsupervised training method. It consists of an encoder and a decoder, as shown in Figure 3. The encoder is responsible for compressing the input data, and the decoder learns to restore the data compressed by the encoder to its original size while looking as identical to the input data as possible. At this point, bottleneck occurs between the encoder and the decoder, which learns to maintain the features and information of the input data as much as possible even when the data size is reduced due to a vector z with a smaller size than the input [4].

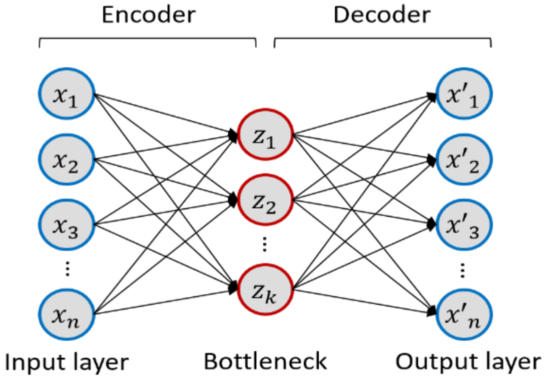

Fig. 3. Structure of autoencoder.

This means that features and patterns of the data can be abstracted and pattern classification can be made easier through the bottleneck vector z. In this paper, the amount of traffic data of each base station converted into images as shown in Figure 4 was learned through an autoencoder. After that, the data was classified through the bottleneck vector z obtained through the autoencoder to check whether the data with the same pattern is in the same group.

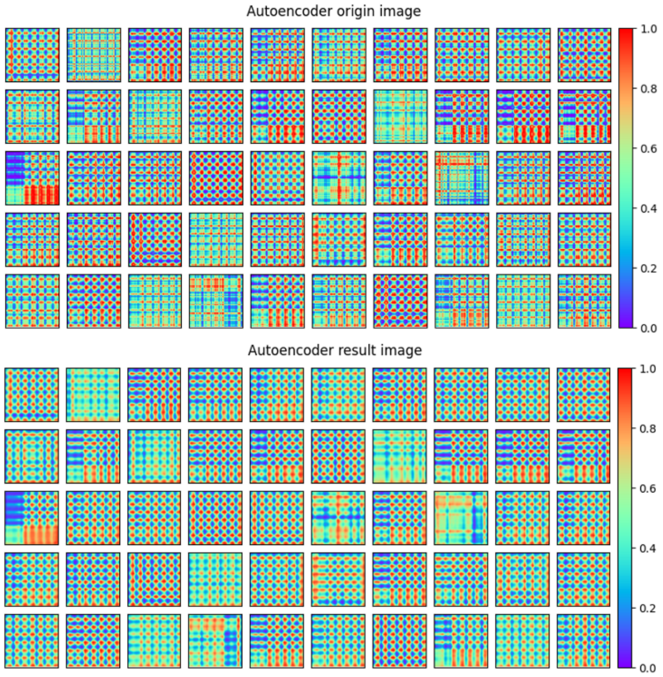

Fig. 4. Traffic data converted to images (top), autoencoder output data (bottom).

III. Analysis Result

We encoded the amount of traffic data from the base stations into the images. Then, we separately clustered offices, shopping malls, residential areas, and roads based on the base stations. We identified that the remaining seven base stations (excluding the second base station in Figure 2) and the base station marked in red in Figure 5 belonged to the same office group among the base stations clustered in the office pattern. Also, we discovered that the base stations illustrated in Figure 6 of Coex Mall are clustered in the shopping mall group. We checked that these are the actual base stations of Hyundai Department Store and Coex Mall in Coex.

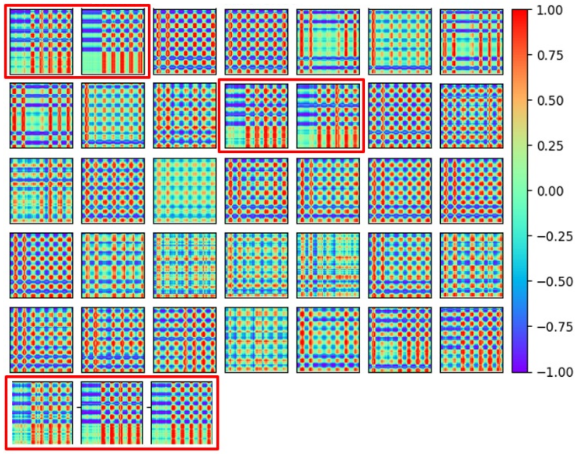

Fig. 5. Grouping of all Coex base stations into office.

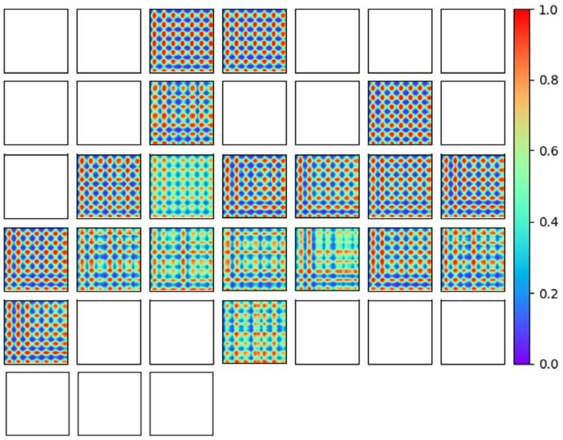

Fig. 6. Groups clustered as shopping malls in Coex.

Then, based on the two base stations in Gangnam Station, which are clearly residential areas among the 88 base stations, the base stations grouped with them are clustered. Furthermore, Figure 7 highlights the location of the base stations marked in green. From Figure 7, it can be seen that most of the base stations in residential areas are located in apartment complexes.

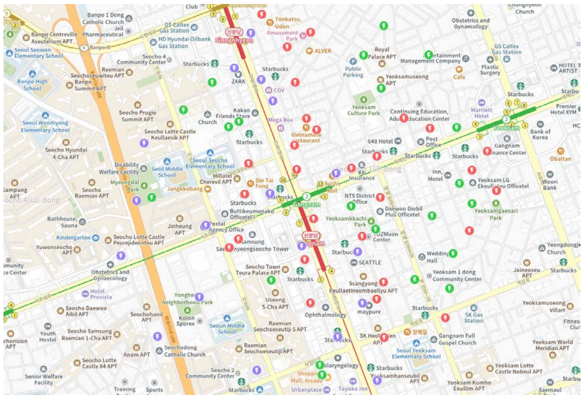

Fig. 7. Grouped by residential area(green), roads(purple) at Gangnam Station.

The three base stations that face the highway from Gangnam Station are clustered together in one group. The purple color in Figure 7 indicates the positions of these groups. Figure 8 shows the base station for Seocho Lotte Castle APT. The base station is in the center of an apartment building, so we assumed it was in a residential areas, but the clustering resulted in the base station being clustered with a roadway. When we inspected the base station, we found that its antenna was facing the highway.

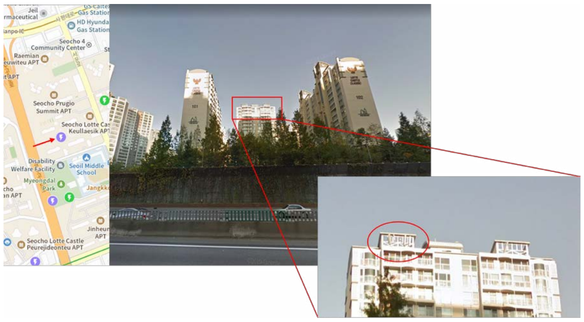

Fig. 8. The base station is located in apartment complex.

IV. Conclusion

In this paper, the mobile traffic patterns of base stations are classified into indoor environments, such as offices, shopping malls, and residential areas, and outdoor environments, such as roads. The base station locations were manually checked to ensure that the characteristics of each group were met. As a result, we have developed a technique for automatically categorizing groups and making them easily recognizable mobile traffic patterns using image encoding and autoencoder. Future studies will identify traffic patterns for base stations in areas that do not belong to existing groups or have multiple patterns, such as offices, shopping malls, residential areas, and roads.

References

[1] Ericsson, “Ericsson Mobility Report,” 2022.[Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report/reports/november-2022.

[2] Wang, Zhiguang, and Tim Oates. ”Imaging time-series to improve classification and imputation.” arXiv preprint arXiv:1506.00327 (2015).

[3] Eckmann, Jean-Pierre, S. Oliffson Kamphorst, and David Ruelle. ”Recurrence plots of dynamical systems.” World Scientific Series on Nonlinear Science Series A 16 (1995): 441-446.

[4] Wang, Yasi, Hongxun Yao, and Sicheng Zhao. ”Auto-encoder based dimensionality reduction.” Neurocomputing 184 (2016): 232-242.