Dual-Attention Multi-Scale Graph Convolutional Networks for Highway Accident Delay Time Prediction

I-Ying Wu
Institute of Computer and Communication Engineering
National Cheng Kung University
Tainan, Taiwan
Q36084329@gs.ncku.edu.tw

Fandel Lin
Institute of Computer and Communication Engineering
National Cheng Kung University
Tainan, Taiwan
Q36084028@mail.ncku.edu.tw

Hsun-Ping Hsieh†
Department of Electrical Engineering
National Cheng Kung University
Tainan, Taiwan
hpsieh@mail.ncku.edu.tw

ABSTRACT

Traffic-related forecasting plays a critical role in determining transportation policy, unlike traditional approaches, which can only make decisions based on statistical results or historical experience. Through machine learning, we are able to capture the potential interactions between urban dynamics and find data on their mutual interactions in a spatial context. However, despite a plethora of traffic-related studies, few works have explored predicting the impact of congestion. Therefore, this paper focuses on predicting how a car accident leads to traffic congestion, especially the length of time it takes for the congestion to occur. Accordingly, we propose a novel model named Dual-Attention Multi-Scale Graph Convolutional Networks (DAMGNet) to address this issue. In this proposed model, heterogeneous data such as accident information, urban dynamics, and various highway network characteristics are considered and combined. Next, the context encoder encodes the accident data, and the spatial encoder captures the hidden features between multi-scale Graph Convolutional Networks (GCNs). With our designed dual attention mechanism, the DAMGNet model is able to effectively learn the correlation between features. The evaluations conducted on a real-world dataset prove that our DAMGNet has a significant improvement in RMSE and MAE over other comparative methods.

CCS CONCEPTS
• Information systems→Spatial systems; Traffic analysis.

KEYWORDS
Delay Time Prediction, Multi-scale GCN, Dual attention, Traffic accidents

†Corresponding author

ACM Reference format:

1 INTRODUCTION

The development of infrastructure and technology has made transportation an indispensable part of life. The vehicle in particular is an essential means of transportation for modern people [28]. Inevitably, traffic has given rise to many problems, in which traffic forecasting is a classic issue that has been frequently discussed. To be more precise, traffic flow prediction [20, 21, 25], mass transit route planning [5], urban route planning [24], accident prediction [37, 40], and traffic congestion prediction [23, 38] are all common real-world traffic forecasting issues. In dealing with these problems, some researchers have proposed to alleviate urban traffic congestion by increasing the traveler’s cost to motivate people to choose alternate trajectories to save time [35]. While some studies focus on connecting incidents to corresponding traffic congestions, few have explored the synergistic effects between these two [33]. In this work, faced with incidents on the highway, our goal is to provide accurate impact forecasting for authorities to make adequate arrangements in advance or draw up suitable future policies.

In particular, when it comes to analyzing the interplay between incidents and traffic conditions on roads, the provision of traffic information turns out to be crucial. Specifically, the complexity and inaccessibility of road conditions could trap the drivers in a queue and unable to find another route when an incident occurs in front of them. Hence, if the necessary traffic information could be provided to the drivers in advance, they can avoid the roads prone to traffic hams so as to reduce the congestion caused by either accidents or incidents.

Moreover, the situation on the highway is particularly stressful yet intriguing. In contrast to other kinds of urban roads, which contain several branches, intersections, or cul-de-sacs, the highway is a closed network structure and consists of straight segments with few junctions (either at-grade or grade-separated interchanges) connecting to other roadways. Therefore, when a
traffic jam or an accident occurs on a highway, it has a larger impact and is likely to clog up, since a detour or diversion is unlikely within a short distance for vehicles compared to those on a general urban road. To sum up, comprehensive studies \cite{19, 41} on highways have demonstrated that effective predictions of traffic-related information for authorities in a timely manner can lead to improved traffic rerouting and thereafter reduce commuting time.

However, most of the previous studies have focused on speed inference and accident occurrence prediction, but neglected to take into account the post-accident clearance time as well as different types of traffic accidents. Therefore, instead of directly inferring the congestion level, this paper focuses on predicting the impact of post-accident congestions and synergistic effects. Specifically, we believe that traffic jams or the sudden increase in traffic flow on the highway could be attributed to periodic factors such as commuting hours and holidays, or aperiodic incidents that reduce available roadways \cite{33}. As mentioned above, we focus on aperiodic events, and aim to predict when the congestion starts after the incident occurs. Consequently, a novel model, Dual-Attention Multi-scale Graph Convolutional Networks (DAMGNet), is proposed to integrate information of car accidents, urban dynamics, and spatial characteristics of the traffic network. By effectively identifying the length of delay times traceable to an accident on various parts of the road (road segments with different distances to the accident), valuable information can be provided to authorities to accurately direct different groups of vehicles, which are expected to be affected on either the highway or urban roads connected by interchanges, and reduce the chance of potential congestion.

To achieve the goal of predicting the delay time after the accident occurs, two challenges are encountered. First, the traffic accident information is recorded in numerical format (e.g., the number of casualties, vehicles involved) and categorical format (e.g., type of vehicles, weather conditions). The latter accounts for most of the data, and the degree of association between each category with the accident is greatly varied. The degree of association turns out to be a challenging issue to handle the categorical data effectively because each feature may have different contributions. Secondly, the road network structure of highways is quite different from that of general urban roads. Since the stations (Sensors, or VD/ Vehicle Detectors that record traffic information on the road) are sparsely deployed, the low censor coverage means that each station only has a small number of other stations in a neighboring region. Therefore, spatial characteristics cannot be comprehensively represented by a general spatial feature extraction method.

To solve the above-mentioned difficulties, we propose DAMGNet, which predicts the delay time of accidents on the highway. The correlation between features is considered in a graph, and a Multi-Scale Graph mechanism is designed to incorporate multiple road network features and increase the linkage between data by using our designed dual attention mechanism. In summary, the contribution of this paper is fourfold.

- Compared to previous studies that predict traffic jams or accidents, we focus on the subsequent development and the prediction of the time it takes for congestion to "travel" to each segment of the road behind the accident, which is a relatively new research problem.
- To the best of our knowledge, this paper is the first in the field of highway traffic prediction to describe the space using graphs in different scales and synergize them with each other in order to obtain latent spatial features.
- We propose a new model named DAMGNet, which combines a dual-attention mechanism with a multi-scale Graph Neural Network (GCN), and to effectively predict the delay time caused by congestion after a traffic accident.
- The evaluation is conducted using Taiwan's highway data. In addition to real-time traffic data and accident records on the national highway, the point-of-interests (POI, e.g., rest area, tourist attraction) near the highway are collected. The results show our proposed DAMGNet is more effective and robust than other peer methods.

The remainder of this paper is organized as follows. To begin with, previous literature with traffic-based and technology-based aspects are discussed respectively in Section 2. An introduction of the preliminary information along with problem definition is presented in Section 3. Next, the details of the proposed DAMGNet are outlined in Section 4 and followed by the evaluation with elaborated analyses in Section 5. Finally, Section 6 states the conclusion along with future works.

2 RELATED WORK

2.1 Traffic

Traffic congestion on highways is a very common phenomenon. Excluding lane reductions caused by accidents and construction, traffic can still slow down unknowingly and not move as smoothly as expected. Nonetheless, traffic accidents are one of the main causes of congestion, so by efficiently predicting the occurrence of accidents, much traffic congestion can be prevented. Therefore, there is much research devoted to the prediction of accidents. For instance, Yuan et al. \cite{37} conducted research on traffic accident prediction with a special focus on spatial heterogeneity. Zhang et al. \cite{39} proposed a model named LSTM-GBRT for the traffic management department to better grasp the situation of traffic safety levels by accurately predicting the safety level of traffic accidents. Based on Convolutional Neural Networks, Zhao et al. \cite{40} propose to extract the autonomous features from a large amount of data in Vehicular Ad-Hoc Network (VANET) to simulate and predict the possibility of traffic accidents.

There are other studies that make predictions of vehicle speed, in which an indicator for traffic congestion associated with vehicle speed below 40 km/h \cite{1} is suggested and can be used as a premise for our method. There are also some works that use primitive machine learning methods. For instance, Priambodo et al. \cite{25} adopted k-NN based on similar traffic historical data to predict vehicle speed. Recently, there have been many studies that focus on temporal and spatial data extraction to predict traffic
data more accurately by utilizing the temporal patterns of traffic flow and the unique spatial characteristics of road networks. Among them, Yu et al. [36] proposed the STGCN to tackle the time series prediction problem in the traffic domain. Liu et al. [18] proposed an approach based on temporal clustering and hierarchical attention. Lv et al. [21] combined the advantages of both RNN and CNN models and designed an innovative look-up operation to capture complex traffic evolution patterns. The GSTGCN [7] adopted a spatial-temporal component which consists of a dynamic temporal module and a global correlated spatial module. Lu et al. [20] presented the AG-GCN, in which the gating mechanism is designed to allow each node in the network to be selectively updated and forgotten in each layer of the GNN. The aforementioned research is inspiring but is inadequate to solve our task. Therefore, we propose a novel model named Dual-Attention Multi-Scale Graph Convolutional Networks (DAMGNet) to predict the delay time for transmitting the congestion caused by traffic accidents to each segment of the road in the back of the traffic accident location.

### 2.2 Technology

#### 2.2.1 Attention mechanisms

Due to high flexibility and efficiency, attention mechanisms are now widely used in different fields such as traffic prediction, recommendation systems, computer vision, and natural language processing. Vaswani et al. [30] proposed to eliminate repetition and convolution completely based on solely attention mechanisms. Liu et al. [17] used local and mutual attention of the convolutional neural network to jointly learn the features of reviews on item recommendation. Škrlj et al. [27] explored attention-based neural network mechanisms for feature importance assessment. Ramachandran et al. [26] verified that self-attention could indeed be an effective stand-alone layer. The main concept of the attention mechanism is to adaptively focus on the most relevant features according to the input data. Recently, several works have adopted attention mechanisms to graphs. For instance, the GATs [31] are proposed, in which the manipulation of graph-structured data is conducted by using the self-attentional layer of the mask to solve the shortcomings of existing methods based on the graph convolution or its approximation. On the other hand, Guo et al. [10] proposed a novel attention-based spatial-temporal graph convolutional network named ASTGCN to capture the dynamic spatial and temporal correlations, while Liu et al. [18] used the hierarchical attention-based mechanism to extract the features at each time step. Furthermore, GMAN [42] adapted an encoder-decoder architecture, in which both the encoder and the decoder consist of multiple spatio-temporal attention blocks to model the impact of spatial and temporal factors on traffic conditions. On the other hand, LSGCN [12] proposed a graph attention network cosAtt, which integrates cosAtt with GCN into a spatial gated block.

#### 2.2.2 Graph Convolutional Networks

Graph Convolutional Network (GCN) is a powerful type of neural network designed to work directly on graphs and leverage their structural information. GCN is used to learn data structured as graphs. The origin of GCN is closely related to word embedding; that is, Word2Vec has inspired similar algorithms for graphs [8]. For instance, Node2vec is originated from Word2Vec and allows advanced sampling schemes. In the field of deep learning, since the Convolutional Neural Network (CNN) cannot capture some features in graphs properly due to the limitation of the network’s operating principle, the graphics convolutional neural network is created. GCN is a powerful way to produce decent feature representations with a general two-level random initialization in its architecture. As it turns out, GCN generalizes CNN to non-Euclidean data. Precisely, GCN is first introduced as hierarchical clustering of the domain and spectrum based on the Laplacian matrix of the targeted graph [3]. Furthermore, Defferrard et al. [6] enhanced the scalability of GCN through fast localized convolutional filters in the spectral domain; while Kipf et al. [14] proposed to scale linearly in the number of graph’s edges so as to learn the representations of hidden layers that encode the local graph structure and features of nodes. Consequently, the first-order approximation of GCN using Chebyshev polynomials is fulfilled. The following graph-convolutional-network-related literature shows promising development of GCN. For instance, Hamilton et al. [11] provided an overview of recent advancements in representation learning on graphs. Vashishth et al. [29] proposed ConiGCN, which estimates labels scores along with their confidences jointly in a GCN-based setting. On the other hand, Graph Wavenet [32] is proposed to precisely capture the hidden spatial dependencies among data. Several methods are suggested to deal with traffic forecasting. Some researchers adopted the Chebyshev polynomials approximation to compute GCN for traffic volume prediction [12, 36], while some used GCN to learn the spatial topology features on traffic speed prediction [33]. As mentioned above, many traffic-related tasks have used GCN to obtain spatial information to achieve better predictions. Inspired by these works, our study also uses GCN as the fundamental component of our forecasting model.
3 PRELIMINARIES

3.1 Delay Time of Accident

Car accidents often lead to traffic jams. Because the road is suddenly unavailable, the overall traffic flow may slow down. Not only will the accident impact the road nearby, but also may lead to a traffic congestion far away. The delay time is defined as follows: when traffic congestion occurs (according to the 40km/h threshold) at a location $N$ kilometer away from the accident, we subtract the happening time of the accident from the occurring time of the congestion. In other words, the delay time represents the time it takes for the traffic jam to take effect, or “travel to” a certain section of the road behind the accident. An example is illustrated in Figure 1, where there is a highway with several stations (censors or VDs that record traffic information) that are sparsely deployed yet highly associated with the mileage markers. Meanwhile, the location of an accident, according to the mileage recorded, is labeled as station 0, and station $n$ indicates the station on the road that is $n$ kilometers behind station 0. For instance, assuming that $N=1$, a car accident occurs at station 0, and station 0 begins to congest at time $t_0$; then station 1, which is located one kilometer behind, begins to congest at $t_1$. The delay time of $N=1$ is $t_1-t_0$, and so on, the delay time of $N=n$ is $t_n-t_0$.

3.2 Traffic-related Data

This work uses information on car accidents and traffic flows from Taiwan’s national highways. In particular, it focuses on accident incidents. We selected the accidents and speed data from Jan. 01, 2019, to Jun. 24, 2019. We collected the real-time traffic speed of the highway described above and recorded the information about traffic flow every 5 minutes. The following is a detailed description of our extracted features for prediction:

Accident-related features. Accident information is divided into several sections, such as spatio-temporal characteristics, road facilities, and personal factors. The spatial and temporal characteristics include the date and time of the accident and whether it is a holiday, while the spatial characteristics include the number of kilometers and vertical height. Road facilities are a description of the lanes and signals at the site of the accident. The personal factors detail the cause of the accident, the number of casualties, and the type of vehicle involved.

POI-related features. We have captured the nearest POIs to each accident location, because we believe that landmarks such as police stations and interchanges may affect the elimination of crashes and clearance time, so they are taken into consideration in our model.

Road network structure. The structure of the national road network is represented in the form of a graph, and the construction method is described in detail in Definition 3, which represents the adjacent conditions between stations and fully captures the global and local spatial characteristics by using a multi-graph.

3.3 Problem Studied

We consider the delay time prediction problem on highways as follows. Given accidents $A = (a_1, a_2, a_3, ..., a_M) \in R^M$ on the highways, and $M$ meaning the total number of accidents in the database, we aim to predict the delay time $T_D = (t_{D1}, t_{D2}, t_{D3}, ..., t_{DM})$ using the accident information $X = (x_{a1}, x_{a2}, x_{a3}, ..., x_{aM})$. Each element of $X$ denotes the accident-related feature vector in an accident event $a_i$. The road network information $G = (g_1, g_2, g_3, ..., g_M)$ denotes the set of multi-graphs, and each element denotes a multi-graph built upon an accident’s surrounding road network. Furthermore, elements of $X$ and $G$ are contained in each element of $A$.

Definition 1. Traffic Accident. The term “accident” implies a certain non-rational event [22, 34], in which “normal” causal relationships are suspended or are somehow intrinsically different from other naturally occurring factors like diseases. Thus, accidents are the “events which can be examined, analyzed, and prevented.” In this paper, the traffic accident $A$ is defined as an accident in which a person is injured or killed, or in which a vehicle, or mass transit system vehicle is damaged, when the corresponding vehicle is in motion on a highway. We record an accident’s traffic-related features in each element of $A$.

Definition 2. Traffic Congestion (TC). Combined with the definition provided by the Department of Transportation Freeway Bureau, the term “congestion” [1] may refer to the phenomenon caused by irregular occurrences, such as traffic accidents, vehicle disablements, and spilled loads of hazardous materials. In this work, we define TC as a condition in which the average speed of a roadway is less than 40 km/h due to traffic accidents, vehicle breakdowns.

Definition 3. Traffic Network. A station (or called as “censor” or “vehicle detector”) refers to a sensor that records the real-time traffic information and is deployed associated with the mileage markers. We treat a station on the national highway as a node of a traffic network. The traffic network is thereafter defined as a bidirected graph $G = (V, E)$, where $V$ is the set of nodes, and $E$ refers to the set of edges between two nodes. The traffic condition in accident $m$ is represented as a graph vertex information $V_m \in R^F$ on graph $G_m$, in which $F$ is the number of dynamic traffic variables in the node which may be related to the forecasting target.

4 METHODOLOGY

In this section, the architecture of our Dual-Attention Multi-scale Graph Convolutional Networks (DAMGNet) is introduced in detail. The model architecture is shown in Figure 2. DAMGNet mainly consists of two components: the accident status component and the traffic network component. The first component handles numerical information such as highway car accidents and traffic flows, while the second component deals with spatial information such as global/local road networks of the highway. We incorporate these two components that consider the information in different aspects and capture the latent features in the data. These two modules are discussed in detail in the following sub-sections.
4.1 Accident Status Component

Accident context is an important source of accident information for prediction. It can be categorized into accident status (e.g., the number of casualties, the lane, the type of vehicles,) and environmental factors (e.g., weather and light.). The architecture of Accident Status Component is shown in Figure 3. Entity embedding and context attention are designed in the Accident Status Component.

4.1.1 Entity Embedding

One-hot encoding is a common preprocessing method for categorical data. However, the performance is often unsatisfactory due to data sparsity. Thus, we can replace one-hot encoding with the embeddings to represent categorical variables in practically any modeling algorithm. Whereas one-hot encoding treats different classes of variables completely independent of each other and often ignores the informative relations between them, entity embedding [9] can map related values closely together in embedding space, revealing the inherent continuity of the data. We distinguish between categorical (e.g., the type of vehicles, weather) and non-categorical (e.g., time of occurrence, mileage, number of casualties) data, and first convert the categorical data into embeddings for subsequent processing. We map each state of a discrete variable into a vector as:

\[ m_i: x_i \rightarrow y_i \]

According to the experimental results of [9], we choose DNN based model as our main structure for the entity embedding technique. More specifically, there are an input layer and a fully connected layer.

Given \( x_c \in X \), which represents the category input data, we convert X to entity embedding. The [9] is used to map each variable in \( x_c \) into vector \( e_i \), and then a categorical context embedding \( E \in R^{k} \) is formed by concatenating each vector \( e_i \).

\[ E = [e_{f-1}, e_{f-2}, \cdots, e_{f-1}] \]

in which \( i \) is the length of categorical variables after they have been converted to embeddings. After that, we concatenate both non-categorical variables \( x_{nc} \in X \) and \( E \) to obtain the complete context vectors \( C \in R^{N\times L} \), in which \( L \) is the length of context vectors.

4.1.2 Context Attention

Accident context features such as time, location, weather and other variables may be connected. In order to fetch the correlation between the features, we develop a self-attention layer to encode the context vectors, which allows the inputs to interact with each other. The self-attention mechanism finds the feature it should pay more attention to, which enables the model to effectively capture the relationship between variables. The outputs are produced by aggregating these interactions and attention scores. Formulaically, given input \( C = [c_0, c_1, \cdots, c_{l-1}] \), the input must have three representations called query \( Q \in R^{N\times d} \), key \( K \in R^{N\times d} \) and value matrices \( V \in R^{N\times d} \), the representations can be formulated as:

\[ Q = CW_Q, K = CW_K, V = CW_V, \]

in which \( W_Q, W_K \) and \( W_V \) are the learnable weight matrices. To calculate the attention output, we adopt Scaled Dot-product attention:

\[ \rho_c = \text{Attn}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d}} \right) V \]

in which \( \rho_c \) represents the encoded embedding converted from the context vectors.

4.2 Traffic Network Component

Traffic conditions are affected by many different factors, with varying degrees of impact. The road conditions are also dynamic and changing all the time. If we can grasp the relationship between roads and vehicle conditions, the prediction results of the model can be improved. Based on the literature [16], the road network has complex spatial dependencies and latent variety patterns. Restricted by the underlying road network, traffic flow patterns cannot be captured completely. A graph convolutional network(GCN) is proposed to solve this problem, which is able to capture topology features. Unlike CNN, which uses a 2D array as input, GCN uses a graph as an input. We use GCN to extract the local spatial and the global spatial features, which is shown in Figure 4. Furthermore, road condition similarity is extracted by the graph attention network \( \text{spatAttn} \).
4.2.1 Multi-Scale GCN
Most of the peer methods in the literature use a single GCN to capture spatial features. However, unlike the urban road network, the national highway is a single route with no intersection. Using only a complete section as a feature may cause interference with less relevant points. Therefore, in addition to the entire national road, which obtains complete road information, we also consider the sites near the accident location. In order to avoid capturing information that is too far away and concentrate on the area where the accident occurred. That is, we use GCN for both local graph and global graph to capture spatial features. According to [14], the matrix of activations in the $l^{th}$ layer $H_l$ can be formulated as:

$$H_l = \sigma(D^{-\frac{1}{2}} \tilde{A} D^{-\frac{1}{2}} H_{l-1} \theta)$$

, in which $\tilde{A} = A + I$ denotes the adjacency matrix with inserted self-loops and $\tilde{A}_{ii}$ is the diagonal degree matrix of $A$. $H^0 = G$, in which $G$ is the input vectors of GCN. $\sigma$ denotes an activation function, such as $ReLU$. $\theta$ is the learnable weight matrix.

In order to obtain the embedding representing convolved road network graphs, batch normalization [13] is first applied to make normalization a part of the model architecture and perform the normalization for each training mini-batch. Thus, the resulting vanishing gradients can be reduced. The normalized embedding $\tilde{g}_c$ is formulated as:

$$\tilde{g}_c = \frac{g_c - E[g_c]}{\sqrt{\text{Var}(g_c) + \epsilon}} \alpha + \beta$$

, in which $g_c$ is the output of GCN, $\alpha$ and $\beta$ is for scaling and shifting, and $\epsilon$ is the value added to the denominator for numerical stability.

Then in order to aggregate all node vectors within the graph, we adopt the NN-based model to extract the latent feature of the graph. Each tensor of all nodes in the output of GCN learns its own embedding through the model and is aggregated by concatenating these embeddings.

\[
\begin{align*}
\text{ge}' &= \tilde{g}_c \cdot A + \gamma \\
\text{ge}'' &= \text{concat}(g_{c1}', g_{c2}', \ldots, g_{cm}')
\end{align*}
\]

, in which $A$ is the learnable weight matrix, $\gamma$ is the additive bias, and $U$ is the number of each tensor in the node vector.

4.2.2 Spatial Attention (spatAttn)
Although the road conditions are dynamic, there are still identifiable patterns. Conditions for some roads could be similar; for instance, the adjacent roads may have similar traffic conditions in a traffic network. The key idea is to flexibly extract the similarities in the traffic network. Inspired by [12], the spatial attention mechanism (spatAttn) can learn which part may require the model to pay more attention to. Given $G''_c = [g_{c1}'', g_{c2}'', \ldots, g_{cm}''] \in \mathbb{R}^{PU}$ as the input of spatAttn, in which $P$ is the length of $g_{c'}$, and $U$ is the length of node vector. The definition of spatAttn is as follows:

$$a_{ij} = \text{sim}(g_{c_i}'', g_{c_j}'') = \frac{g_{c_i}'' \cdot g_{c_j}''}{\max(\|g_{c_i}''\|, \|g_{c_j}''\|, \epsilon)}$$

$$\rho_g = \text{spatAttn}(a, g_{c}'') = \rho_g(i) = \sum_{j \in G'_c \setminus \{i\}} a_{ij} g_{c_j}''$$

, in which $G'_c$ is the nodes in graph $G''_c$ except node $i$, and $\epsilon$ is set to avoid division by zero. Through spatAttn, we obtain the spatial embeddings, which take into consideration the relevance between vertices in road networks.

4.3 DAMGNet
We concatenate the context embedding $\rho_c$ and the spatial embedding $\rho_g$ as $\rho = \rho_c \odot \rho_g$ of each accident data. Next, we use a fully connected layer as the joint method. Ultimately, the forecast result $Y = [y_1, y_2, \ldots, y_M] \in \mathbb{R}^M$, same as $T_B$, is extract the output layer, which consists of one fully connected layer.

4.3.1 Loss Function
The whole regressor is trained for minimizing the mean-squared error (MSE):

$$L(\hat{Y}, Y) = \frac{1}{n} \sum_{i} ||Y_i - \hat{Y}_i||^2$$

, in which $Y \in \mathbb{R}^M$ is the ground truth, and $M$ is the number of car accidents. Using Adam as the optimizer, we make use of its characteristics, which retains Momentum’s approach to the past gradient and adjusts the gradient speed, as well as adjusts Adam’s learning rate to the past gradient’s square value. In the next section, we will conduct a series of experiments for the proposed model to demonstrate its stable and robust efficacy.

5 EXPERIMENT
In this section, in order to verify the effectiveness of DAMGNet, we perform various experiments on a real-world dataset to evaluate the performance of the proposed framework. At the same time, DAMGNet will be compared with other methods to show its efficacy. Precisely, we focus on discussing the following research questions:
Dual-Attention Multi-Scale Graph Convolutional Networks for Highway Accident Delay Time Prediction

Q1. In order to show the effectiveness and robustness of the proposed model, we compare DAMGNet with the baseline methods.

Q2. Which feature and component of DAMGNet have the most significant impact on the performance of the model?

Q3. When $l$ is the number of stations extending forward and backward from the node of an accident, how will the value of $l$ affect the results? How to modify $l$ to obtain the best spatial information?

5.1 Dataset
We evaluate the performance of DAMGNet on highways in Taiwan with different aspects of questions. We choose National Highway No. 1 and National Highway No. 3 in Taiwan as the experimental data, which has 1,961 accidents in total. The highway data contains two parts; one is the number of car accidents on the highway, and the other is the real-time collected data from stations on the highway. Each station records vehicle speed, traffic flow and occupancy rate for every five minutes with location information. Meanwhile, since the influential distance of all accidents on the highway is at most 3 km according to our preliminary statistics, the value of $N$ in representing the kilometers behind the accident site is set to 1, 2, 3 as the prediction basis for further experiment. Figure 5 depicts the distribution of delay time at $N=1, 2, 3$. It can be observed that the figure roughly conforms to the power-law distribution. In addition, the statistics between different incident severity and delay time are shown in Figure 6. Category A1 refers to traffic accidents that result in immediate death or within 24 hours, A2 refers to traffic accidents that result in injury or death after 24 hours, and A3 refers to cases in which a vehicle collision causes property damage but no casualties. It seems that A1 occurs when the traffic flow is low and the overall speed on the road is high. In other words, when a car accident occurs at a low traffic flow, the probability of the most serious accidents (e.g., A1) will be greatly increased because of the high speed. Severe accidents require a longer processing time. Therefore, the percentage of delay time longer than 30 minutes is also relatively high.

5.1.1 Data Preprocessing
In order to prevent the problem of missing data, we process the VD data in units of ten minutes. The vehicle speeds are added and then averaged, while the traffic flow is directly summed up. The traffic jam here is defined as the speed below 40 kilometers per hour. In addition, we only keep the data of the traffic jam time difference between the two stations, which are one kilometer apart, within 3 hours. If the time is too long, it may not be related to the traffic accident and cannot be regarded as the transmission of the traffic jam.

5.2 Experiment Settings
5.2.1 Baselines
Five methods are selected as the comparative methods; (a) Mean - historical mean of all accidents, (b) Median - historical median of all accidents, (c) Random Forest (RF) [2], Random Forest, (d) XGBoost [4], and (e) Deep Neural Networks (DNN) [15]. (f) Graph Convolutional Network (GCN) [6]

5.2.2 Evaluation Metrics
In order to verify the performance of our model, we forecast the highway accident delay time in $N=1,2,3$. Besides, we use following metrics for evaluation:

- Mean Absolute Error (MAE)
  \[ \text{MAE}(Y, \hat{Y}) = \frac{1}{n} \sum |Y_i - \hat{Y}_i| \]

- Root Mean Squared Error (RMSE)
  \[ \text{RMSE}(Y, \hat{Y}) = \sqrt{\frac{1}{n} \sum (Y_i - \hat{Y}_i)^2} \]

5.2.3 Parameters
For data splitting, we use 80% data for training and validation, and the remaining 20% as the test set. From the training set, we select 85% for training and 15% as the validation set for early stopping. The proposed framework is executed with the following hyper-parameter settings: learning rate=0.01, batch size=256, and Adam as the optimizer. In the context attention layer, we use ReLU activation function for the middle layer and softmax activation function for the output layer. On the other hand, in GCN layer, we use ReLU activation function followed by the graph encoder with the 8-dimension output. We add Dropout=0.2 in FC layer.

Figure 5: Delay time distribution when $N$ is set to 1, 2, 3

Figure 6: Distribution of the delay time interval caused by different vehicle accidents severity
5.3 Performance Comparison

In this section, we evaluate the performance of our DAMGNet and compare it with baseline models and present the delay time prediction results in Table 1. In traditional statistical methods, the mean and median of historical data are selected as the comparison methods. The performance of the historical median is more robust and relatively stable compared to the historical mean, which is easily affected by outliers.

Among the machine learning methods, tree-based models such as random forest and Xgboost have similar performance, and RF is slightly better. It may be attributed to the nature of experimental data. Explicitly, on the tree-based approach, this task is not suitable for complex training methods, which causes Xgboost to fail to deliver the expected performance. It may be because the tree is unstable, as small changes in the data may result in generating a completely different tree. For deep learning-based models, in the case of more complex data correlation, Table 1 clearly shows that they perform well compared to tree-based models, especially on RMSE. Since GCN is an essential component of our proposed DAMGNet, we also use the basic GCN as a competitive method. It is worth mentioning that the proposed DAMGNet outperforms the competitors listed above in all evaluation metrics. When N=1, its MAE is at least 2.7501 less than other methods. When N=2 and 3, DAMGNet also surpasses all the baselines. The experimental results show that basic GCN indeed performs better than other baselines except for our DAMGNet. The result confirms that modeling the impact of delay time using a graph-based approach is effective in this task; however, the basic GCN still performs not well enough compared to our DAMGNet. The main reason might be because our designed components: the accident status component and the traffic network component, are more suitable to capture the spatial property of the traffic on the highway. To conclude, it is clear that our proposed framework is appropriate and effective for this task.

5.4 Ablation Study

5.4.1 Components

In DAMGNet, individual components are designed to capture different features. To verify the effectiveness of each component, we designed experiments to evaluate the prediction results of different component combinations. We then briefly introduce each of the following subcomponents:

- **EN** - Entity embeddings are used to process categorical data, which map related values closer to each other in the embedding space as introduced in Section 4.1.1.
- **CA** - Context attention finds out which feature it should pay more attention to and which feature allows them to interact with each other (as mentioned in Section 4.1.2).
- **MG** - Multi-GCN is used to obtain spatial latent features in the data. The spatial characteristics of global and local areas are also taken into account as discussed in Section 4.2.1.
- **SA** - Spatial attention mechanism flexibly extracts the similarities in the traffic network as introduced in Section 4.2.2.

According to Figure 7 and Figure 8, we observe that adding EN or CA is slightly better than using DNN alone in MAE, and the combination of EN and CA further improves the performance significantly. However, when it comes to the overall model improvement, MG is the most important one, in which not only has MAE dropped a lot, but RMSE has also dropped significantly. After joining MG, MAE dropped by 23.8% and RMSE dropped by 20.4%. At the same time, when adding the impact of SA, MAE dropped further by 2.7%. The results show that the dual attention can indeed improve the accuracy.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Distance=1</th>
<th>Distance=2</th>
<th>Distance=3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td>Mean</td>
<td>34.7163</td>
<td>21.3639</td>
<td>48.7984</td>
</tr>
<tr>
<td>Median</td>
<td>36.9892</td>
<td>16.5877</td>
<td>52.9638</td>
</tr>
<tr>
<td>Random Forest</td>
<td>32.4609</td>
<td>19.0707</td>
<td>45.2096</td>
</tr>
<tr>
<td>Xgboost</td>
<td>35.4308</td>
<td>20.3959</td>
<td>46.9400</td>
</tr>
<tr>
<td>DNN</td>
<td>30.2360</td>
<td>19.4477</td>
<td>41.8483</td>
</tr>
<tr>
<td>GCN</td>
<td>22.3637</td>
<td>15.3615</td>
<td>38.9159</td>
</tr>
<tr>
<td>TAMGNet</td>
<td>21.7805</td>
<td>12.6114</td>
<td>38.7451</td>
</tr>
</tbody>
</table>
5.4.2 Features

The accident data contains various characteristics such as time, location, and lane facilities. In this section, we will verify the impact of various features. The following are some of the characteristics we have distinguished: (1) T: Time of accident, (2) Pos: Location and mileage of the accident, (3) W: Weather, light, (4) Sig: Traffic sign, (5) RS: Road conditions, such as state, whether there are defects, obstacles, (6) Lane: Types of lane facilities, (7) Per: Personally related, including the vehicle that caused the accident, the cause of the accident, the number of casualties. (8) DP: Distance to the nearest POI such as police stations, interchanges.

The impact of different combinations of features in the accident status component is analyzed and shown in Table 2. Because Per is the most relevant factor in an accident, such as the number of casualties, the cause of the accident, we include it in every combination. First, we used the basic features T, W, and Pos, and found that the results were similar. After adding road-related features such as PS and Lane, the results were improved considerably. However, the addition of Sig made the results worse. In terms of MAE, when N=1, the addition of Sig increased the result by at least 4%; when N=2, the addition of Sig increased the result by at least 1.28%. It is worth mentioning the feature DP, which is the distance from the nearest police station or interchange to the accident site, has an obvious influence on the results.

6 CONCLUSION

In this work, we found a strong link between car accidents and congestion. In contrast to previous research approaches, we aim to accurately predict the delay time to the roads behind congestion caused by a vehicle accident. We propose DAMGNet, which integrates a dual attention mechanism and a multi-scale consideration of road network information to efficiently and effectively predict the delay times of traffic congestion. Experiments conducted on traffic flow and accident information of the National Highway in Taiwan prove the applicability and robustness of our method on this task. In the future, we hope to add more external features to further improve the accuracy of prediction, and to cooperate with traffic authorities so as to achieve more efficient control and improve the quality of driving.

ACKNOWLEDGMENTS

This work was partially supported by Ministry of Science and Technology (MOST) of Taiwan under Grants MOST 108-2221-E-006-142, MOST 108-2636-E-006-013, MOST 109-2636-E-006-025, and MOST 110-2636-E-006-011.
REFERENCES


