

Spatio-Temporal Traffic Prediction for Efficient ITS Management

Aram Nasser^a and Vilmos Simon^b

*Department of Networked Systems and Services, Faculty of Electrical Engineering and Informatics,
Budapest University of Technology and Economics, Budapest, H-1111, Hungary*

Keywords: Intelligent Transportation Systems, Traffic Prediction, Attention, Encoder-Decoder, Spatio-Temporal Data Presentation.

Abstract: Traffic forecasting is a crucial element of Intelligent Transportation Systems (ITSs), exerting significant influence on the optimization of urban mobility. Through precise anticipation of traffic patterns, ITS facilitates proactive traffic flow management, leading to a multitude of benefits for both the city and its inhabitants. However, the intricate topological structure of road networks and the changing temporal patterns in traffic create challenging problems that demand solutions considering both the spatial and temporal aspects of traffic characteristics. Most existing traffic prediction models are influenced by Graph Neural Networks (GNNs) to capture the spatial structure of road networks. However, this approach typically relies on the adjacency matrix, which might not always reflect the dynamic state of traffic conditions. In addition, GNNs are not universally applicable across different traffic topologies. What works for one road network may not yield the same results for another, owing to disparities in the number of roads, thus graph nodes, and the unique characteristics of each location. Therefore, in this paper, the Spatio-Temporal Multi-Head Attention (ST-MHA) model is introduced to solve this issue. ST-MHA depends on a modified version of the Multi-Head Attention (MHA) mechanism to capture the spatial structure of the road network implicitly, as well as a GRU-based encoder-decoder structure for integrating the temporal characteristics. Our model outperforms three state-of-the-art baseline models, which include temporal, spatial, and spatio-temporal models. This enhanced performance is evident across three different prediction horizons when evaluated on a real-world traffic dataset.

1 INTRODUCTION

The unprecedented expansion of cities has imposed increased pressure on transportation networks, resulting in a negative influence on human health, the economy, and the environment (Levy et al., 2010), (Zhang and Batterman, 2013). This pressure has prompted concerned people and organizations in smart cities to contribute to the creation of systems for utilizing information, communication, and sensing technologies in transportation and transit systems, referred to as Intelligent Transportation Systems (ITSs) (Wang, 2010). One of the key elements of ITSs is predicting traffic characteristics, such as speed, as it provides a futuristic view of the traffic situation. This insight gives the authorities the time needed to take action before congestion arises, ultimately leading to a reduction in pollution and commute time for individuals. The integration of innovative technologies, as well as

novel data collection techniques, like sensors (including floating car and wide-area sensors) and connected vehicles, have increased the complexity of analyzing and managing traffic data (Rahmani et al., 2023), (Kaffash et al., 2021).

The problem of traffic prediction has evolved from a statistical challenge into a problem tackled by Machine Learning and, most recently, Deep Learning. It can be analyzed as time series data, allowing for future predictions based on past observations and particularly considering the evolving traffic values over time (Yuan et al., 2022). Nevertheless, in some cases, because traffic data are gathered from real spatially-located sensors, spatial information is integrated as an external factor to improve traffic prediction. The spatial topology of the road network can be mapped as graphs with vertices and edges, describing the sensors as well as the connections among them, respectively. The initial graphs are usually built based on adjacency matrices, which can be constructed on four distinct bases: road-based, distance-based, similarity-based, and dynamic matrices, according to (Jiang and

^a  <https://orcid.org/0000-0001-9696-7387>

^b  <https://orcid.org/0000-0002-7627-3676>

Luo, 2022). However, although the spatial topology of the roads is static, the connections, in reality, are not and might be strong at some periods while weak or negligible at others. As a result, GNNs can be perfect for time-static data, which is not the case in traffic prediction.

To address this concern, researchers are endeavoring to integrate the temporal dimension with GNNs, enabling dynamic adjustments of connections over time. Most available spatio-temporal models use the GNNs to capture the spatial features of traffic data, where the adjacency matrix is usually needed to formulate the initial relationship among traffic detectors. This matrix must be prepared earlier, and its dimensions are fixed throughout the training process, which restricts the ability to use the same model for different road networks interchangeably. In other words, using the model for another road network indicates building a new adjacency matrix specifically for that new topology. In addition, building the adjacency matrix can be a time-consuming task, and it varies based on the scale. For example, for the same country, the adjacency matrix used for the intra-city road network differs from the one employed for inter-city or metro stations.

These limitations have encouraged us to introduce our novel model, ST-MHA, where the adjacency matrix is no longer needed to capture the spatial dependencies. Instead, ST-MHA can implicitly infer the spatial topology from the data. In this manuscript, we provide a spatio-temporal model that adjusts the MHA mechanism to capture the spatial interactions of data while a GRU-based encoder-decoder for the integration of temporal information. The rest of this article is organized as follows: the next section is dedicated to introducing state-of-the-art literature on traffic prediction models, encompassing spatial, temporal, and spatio-temporal aspects. The detailed structure of ST-MHA, along with the elucidation of the problem addressed by our model, is outlined in section 3. In the following chapter, the experimental setup, including the utilized dataset, evaluation metrics, and parameter designing, as well as the final results and discussions, are examined in 4. Finally, the conclusion of our research is presented in the last section.

2 RELATED WORK

Classic statistical models, like the ARIMA model and its variants, have been widely used in the past to analyze and predict future traffic status (Ahmed and Cook, 1979), (Levin and Tsao, 1980), (Lee and Fam-

bro, 1999), (Williams, 2001), (Williams and Hoel, 2003), (Kamarianakis and Prastacos, 2003). However, these models assume the time series to be stationary (Karlaftis and Vlahogianni, 2011), (Li and Shahabi, 2018), which hinders the traffic prediction process. Thus, the advancement of non-parametric models, such as Support Vector Regression (SVR) (Castro-Neto et al., 2009), (Su et al., 2007), (Jin et al., 2007), and Neural Networks (NNs) (Hua and Faghri, 1994), (Dougherty and Cobbett, 1997), (Ledoux, 1997), with improved efficiency and heightened predictive precision, has made statistical models less common, especially for complex data modeling. Initial non-parametric models have vigorously challenged the classical parametric models. However, when it comes to handling temporal data, even these models have their limitations due to the time-based associations among time steps. This issue prompted the introduction of Recurrent Neural Network (RNN) (Rumelhart et al., 1986), a special type of NNs equipped with memory cells to help capture the temporal features of the time series. Nevertheless, the vanilla RNN suffers from the vanishing gradient problem (Hochreiter, 1998) that is mitigated by its variants like the Long-Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Cho et al., 2014a). These RNNs can also be structured as models like the RNN Encoder-Decoder architecture (Cho et al., 2014b) to allow for a deeper structure and additional improvement of traffic prediction. While these models successfully address the traffic prediction issue as a temporal problem, some models integrated the spatial features as exogenous inputs to capitalize on the spatial interactions among roads in the road network.

To fully leverage spatial features of the road network, some researchers use modified versions of the GNNs (Gori et al., 2005) to capture adjacent interactions among distinct roads in the road network. These models are extended by integrating Convolutional Neural Networks (CNNs) (Lecun et al., 1998), in both the spectral and time domains, with GNNs to allow the extraction of multi-scale localized spatial features and form GCNs (Kipf and Welling, 2017), (Zhou et al., 2020). Other GNNs can also be extended by using the attention mechanism to overcome the shortcomings of convolution-based GNNs (Veličković et al., 2017). According to (Veličković et al., 2017), these drawbacks are solved by calculating different weights to various nodes in a neighborhood implicitly and avoiding the dependence on knowing the graph structure upfront. Nonetheless, these models lack the ability to capture the temporal features of traffic. Thus, incorporating models that

consider both time and space is essential to tackle the spatio-temporal problem.

Harnessing the advantages of both temporal and spatial characteristics of traffic data through the integration of these features into spatio-temporal models has gained significant attention among researchers. In (Zhao et al., 2020), GCNs are combined with GRU units to simultaneously capture the spatial and temporal dependencies within traffic data, respectively. The authors then extended the model by incorporating the attention mechanism into the existing framework to improve prediction accuracy (Bai et al., 2021). Traffic flow spatio-temporal dynamics are modeled as a diffusion process in (Li et al., 2018), where spatial interdependence is grasped through bidirectional random walks on graphs and temporal interdependence is handled by employing the encoder-decoder architecture with scheduled sampling. Dynamic Time Warping (DTW), a technique used for matching and alignment, is utilized and improved to what is called “fast-DTW” to generate temporal graphs (Li and Zhu, 2021). These temporal graphs are then incorporated into a novel Spatio-Temporal Fusion Graph Neural (STFGN) model. Several STFGN are treated in parallel and their outputs are concatenated and added with Gated CNN output to act as the input of the next STFGN layer. Authors of SST-GNN (Roy et al., 2021) introduce a streamlined framework that encapsulates two different models, current-day and historical, to capture daily and weekly patterns of traffic data via a weighted spatio-temporal aggregation scheme. In (Chen et al., 2019), the authors introduced the MRA-BGCN model, which is comprised of two parts. The first incorporates two graphs, node and edge-wise graphs, and provides explicit modeling of interactions among nodes and edges. The second is to autonomously learn the significance of distinct neighborhood ranges. In LSGCN (Huang et al., 2020), a new graph attention network, called cosAtt, is combined with GCN to grasp the spatial features of traffic. In addition, a Gated Linear Unit (GLU) is adopted for the temporal characteristics simultaneously. Zheng et. al. introduced GMAN (Zheng et al., 2020), where spatio-temporal correlations are modeled by employing an encoder-decoder architecture that combines several spatio-temporal attention mechanism blocks. Furthermore, a transform attention layer is devised between the encoder and the decoder to mitigate the impact of error propagation to enhance long-term prediction performance. A variant of the GCN, called LPGCN, is introduced in (Qi et al., 2022) to model the spatial characteristics of traffic. On the other hand, a multi-path CNN is utilized to learn the collective influence of past traffic conditions

on future ones, which is further modified by an attention mechanism. All the aforementioned spatio-temporal models use adjacency matrices to construct the GNNs and GCNs used in their models. These adjacency matrices have their limitations, as mentioned earlier, in addition to their impracticality while dealing with large and sparse graphs -due to the large number of zeros they might have in this case. In this paper, the ST-MHA model is introduced to capture both spatial and temporal features of traffic data by utilizing a modified version of the MHA mechanism and GRU-based encoder-decoder, respectively.

3 METHODOLOGY

The full architecture of the Spatio-Temporal Multi-Head Attention Based Traffic prediction model is outlined in Fig. 1, where each block represents a specific stage of the model. The lower part of each stage delineates its inputs for each time step, the middle segment displays its outputs, and the upper section indicates its name. These stages are grouped into two components: Spatial and Temporal. Further elaboration on these components and their sections is provided in subsequent sub-sections 3.2 and 3.3, respectively. Nevertheless, before diving into the details of ST-MHA, the problem this paper addresses is defined in the next subsection 3.1.

3.1 Problem Definition and Statement

This paper addresses the problem of estimating future traffic speeds over a given time horizon based on present and past data. The data are collected from multiple speed loop detectors positioned in various locations, and the objective is to capture this spatial distribution without relying on predefined graph structures. This spatial modeling is then integrated with a temporal representation of data to achieve a comprehensive spatio-temporal framework for representation and prediction.

Given T current and historical time steps $\{X_{(t-T-1)}, \dots, X_{(t-1)}, X_t\}$, where X is the traffic data of N roads, the aim is to predict future traffic data $\{X_{(t+1)}, \dots, X_{(t+\tau)}\}$, where τ is the prediction horizon. While this issue can be approached as a time series problem, the integration of the spatial features could improve traffic forecasting accuracy. The objective of our proposed model, ST-MHA, is to utilize implicit spatial and temporal patterns without the need for creating graph structures that are restricted by the number of nodes. To achieve that, our approach involves employing a modified MHA model to capture

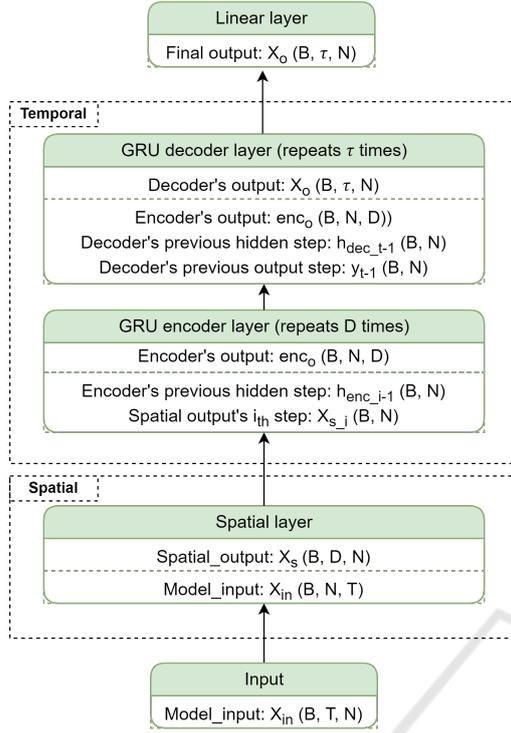


Figure 1: ST-MHA model.

the spatial properties, as well as integrating a GRU-based encoder-decoder model to handle the temporal characteristics.

Our model can improve traffic management by helping authorities make informed decisions to reduce congestion, optimize traffic flow, and support route planning for mapping applications. By predicting traffic speeds, it provides real-time insights, allowing authorities to adjust signal timings, optimize light cycles, and implement diversions to minimize delays. This leads to smoother traffic movement, fewer bottlenecks, and more efficient travel for all road users.

3.2 The Spatial Component

To obtain the spatial properties of traffic data without relying on adjacency matrices or building graph neural networks, the interactions among several road detectors' data are computed over the time dimension spanning H periods, called "heads." In other words, the input time series is divided into several heads, each describing distinct periods of the past data. Nevertheless, this direct approach has two problems. The first one arises when the available historical data are scarce, which might make it challenging to partition the data into segments. The second involves the careful consideration of the number of input time steps,

ensuring that this number is divisible by the number of heads H . Failing to meet this condition can lead to the failure of the whole model. To solve these issues and minimize the dependence on a varying number of past time steps, the time dimension is mapped onto the size of the model's hidden dimension. Therefore, the interactions are rendered within a consistent dimension, typically larger than the initial one.

Algorithm 1: Spatial Component.

Input: Input sequence X_{in}
Data: Model's hidden size D , Number of heads H
Result: Output sequence X_s

```

 $X \leftarrow \text{Linear}(X_{in})$  // Mapped input
 $Q \leftarrow \text{Linear}_q(X)$  // Queries
 $K \leftarrow \text{Linear}_k(X)$  // Keys
 $V \leftarrow \text{Linear}_v(X)$  // Values
 $scores \leftarrow (Q \cdot K) / \sqrt{D/H}$  // Scores
 $W_{att} \leftarrow \text{Softmax}(scores)$ 
 $X_s \leftarrow \text{ReLU}(\text{concat}[W_{att} \cdot V])$ 
    
```

Following the extension of the time dimension, the scaled dot product MHA mechanism is utilized as described in equations 1 and 2, akin to the MHA model in (Vaswani et al., 2017). The input time series are projected on three linear layers, yielding the Query, Key, and Value matrices, denoted as Q , K , and V , respectively. Subsequently, each of these matrices is further reshaped and divided into H heads depicting various periods of the past time steps. Each attention head, AH_i hereafter, discerns specific dependencies within the input time series, thus allowing the model to learn different representations of the input by attending to different parts of the sequence.

$$AH_i(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i}{\sqrt{D/H}}\right) * V_i \quad (1)$$

$$MHA(Q, K, V) = \text{Relu}(\text{concat}[AH_1, \dots, AH_H]) \quad (2)$$

To capture the interactions among the road data, D/H matrix multiplications are undertaken between Q and K in parallel. After that, the Softmax function is utilized to constrain the outputs between zero and one. This operation yields a matrix, denoted as W_{attn} , whose dimensions are $(B, D/H, N, N)$, where N is the number of traffic detectors, illustrating how traffic data spatially influence each other interchangeably over H time periods. Subsequently, this matrix is used to scale the V matrix by another matrix-to-matrix multiplication. The outcome of this process is a scaled representation of the input matrix, which

varies based on the importance of each traffic detector's influence on the rest. It is noteworthy to mention that the values obtained from the matrix multiplication of Q and K are divided by the square root of D/H to avoid a potentially large magnitude of the dot product, which might push the Softmax function into regions with exceedingly low gradients. These interactions are then concatenated and processed with the Relu activation function (Agarap, 2018) to form the modified MHA. For a deeper understanding, the pseudo code that describes the spatial component of ST-MHA is given in algorithm 1.

3.3 The Temporal Component

The temporal part is a GRU-based encoder-decoder model. Before delving into the details of this component, it is important to clarify that the GRU cell incorporates multiple gates as described in equation (3)(ArunKumar et al., 2022):

$$\begin{aligned}
 z_t &= \sigma(W_z * [h_{t-1}; x_t]) \\
 r_t &= \sigma(W_r * [h_{t-1}; x_t]) \\
 \tilde{h}_t &= \tanh(W * [r_t * h_{t-1}; x_t]) \\
 h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t
 \end{aligned} \tag{3}$$

where x_t is the input at time t. z_t and r_t are the update and reset gates at time t, respectively. h_t is the hidden state at time t, while \tilde{h}_t is the new candidate memory cell at time t. Symbols σ and \tanh denote the Sigmoid and hyperbolic tangent activation functions, respectively, whereas $[\cdot]$ and $*$ represent concatenation and the element-wise multiplication.

The encoder is simply a GRU layer used to sequentially encode the outputs of the spatial component with a zero-initiated hidden state. It iterates over the spatial component's output as described in equation 3. The decoder alters the concatenation-based attention weights stated in (Luong et al., 2015). This attention mechanism is chosen here to offer a straightforward approach for generating the temporal output, one step at a time. The decoder's hidden state at time t-1, $h_{dec_{t-1}}$ is considered as the query q, and the encoder's hidden states, enc_o , as the keys K. This process is done by concatenating the query and keys, the parenthesis $[\cdot]$ in Algorithm 2, and applying a linear layer followed by Softmax. Since there is no $h_{dec_{t-1}}$ at the beginning of the process, it is initialized with the last encoder's hidden state, $h_{enc_{D-1}}$. The resulting attention weights are then multiplied with the previous output of the model, y_{t-1} , which is considered as the values matrix V, unlike the traditional mechanism that considers $h_{dec_{t-1}}$ to be the V. At this step, the GRU cell is utilized as stated in equation 3, where the input

Algorithm 2: Temporal component.

```

Input: Input sequence  $X_s$ 
Result: Output sequence  $X_o$ 
Encoder:
for  $i = 0$  to  $D$  do
     $h_{enc_i} \leftarrow \text{GRUCell}(X_{s_i}, h_{enc_{i-1}})$ 
     $h_{enc_{i-1}} \leftarrow h_{enc_i}$ 
     $enc_o \leftarrow \text{append}(h_{enc_i})$ 
end
Decoder:
for  $t \leftarrow 0$  to  $\tau$  do
     $scores \leftarrow \text{Linear}(\text{concat}[h_{dec_{t-1}}, enc_o])$ 
     $w_t \leftarrow \text{Softmax}(scores)$ 
     $X_{gru_t} \leftarrow w_t \cdot y_{t-1}$ 
     $h_{dec_t} \leftarrow \text{GRUCell}(y_{t-1}, X_{gru_t})$ 
     $X_{o_t} \leftarrow \text{Linear}(h_{dec_t})$ 
     $y_{t-1} \leftarrow X_{o_t}$ 
     $X_o \leftarrow \text{append}(X_{o_t})$ 
end

```

is the previous model's output and the hidden state is the result of the attention mechanism. The output of each iteration is then concatenated to form the final output of the model. A comprehensive elucidation of the temporal aspect and matrix dimensions at every step is offered in algorithm 2.

4 PRACTICAL ANALYSIS

4.1 Dataset

ST-MHA is trained and evaluated on a real-world dataset collected from the highways of Los Angeles city via loop detectors. As in (Zhao et al., 2020), speed data of N=207 detectors are chosen from the 1st of March till the 7th of March, 2012. Missing data were amputated using linear interpolation. The resulting values are normalized between zero and one, then reverted to their original scale at the end of the training for better results interpretation. Fig. 2 illustrates the distribution of the road detectors over the city of Los Angeles (Lu et al., 2020).

Traffic speeds are aggregated every 5 minutes to form 2016 data points for each traffic speed detector, where the training is run on 80% of the data or 1612*207 points. The remaining 20%, or 404*207 points, are used to evaluate the model. Two hours of past data are used as inputs to predict the next 15, 30, and 45 minutes of future traffic speeds.

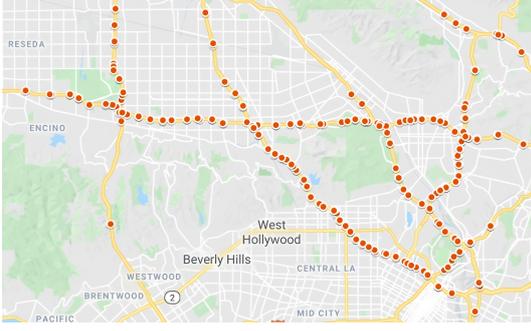


Figure 2: Distribution of detectors over the highways of Los Angeles County (METR-LA).

4.2 Evaluation Metrics

Four evaluation metrics are used to evaluate the performance of the ST-MHA model based on ground truth values Y_t and predicted ones \hat{Y}_t . These metrics are the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE), which are given as in the equations 4, 5, 6, and 7, respectively. MSE is the average of the squared differences between the actual and predicted values, offering a measure to assess the average squared deviation between ground truth and predictions. However, its interpretability is hindered by the squaring operation. RMSE is the square root of MSE, and it shares the same unit as the original data, making it more interpretable. Both MSE and RMSE give higher weight to large errors due to squaring, making them sensitive to outliers. MAE, on the other hand, represents the average of the absolute differences between the actual and estimated values. MAE is less sensitive to outliers compared to MSE and RMSE and is often preferred when outliers are present and a more balanced view of the errors is needed. MAPE is a percentage that describes the size of the errors compared to the actual values. MAPE has a big limit when actual values are close to zero, which can produce an overshoot MAPE value. Therefore, zero values in Y were exempt while calculating the MAPE.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (5)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i) \quad (6)$$

$$MAPE = \frac{100}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{|y_i|} \quad (7)$$

4.3 Model Parameters Designing

Two of the main parameters of ST-MHA are the model's hidden size and the number of heads H . Each one of these parameters can significantly affect the prediction results of ST-MHA. After conducting many experiments, the hidden size of $D=128$ units and $H=8$ are chosen as they provide the best performance results. Another less important parameter is the learning rate, lr hereafter, which describes the step by which the model is modified. lr is initiated to 0.01 and is decreased by 90% every 1000 iterations.

Due to the limited number of data samples in the time domain, the batch size is chosen to be relatively small, at eight samples per iteration, to enable a larger number of training iterations per epoch. All experiments are conducted over 200 epochs, where 24 past values (equivalent to 2 hours) are used as inputs into the model, while the prediction horizons utilized as outputs are 15, 30, and 45 minutes.

4.4 Results and Discussion

Three benchmark models are employed to validate our model, ST-MHA. These models are chosen to cover the temporal, spatial, and spatio-temporal approaches. The first is an RNN Encoder-Decoder model (Cho et al., 2014b), encoder-decoder hereafter, where vanilla RNNs are replaced with GRU units for better prediction accuracy. Second is a GCN model (Kipf and Welling, 2017) that uses the spatial features to predict future traffic speeds. The third is the state-of-the-art spatio-temporal model TGCN, which is configured with a model's hidden size of 64, as reported by the authors, to achieve the best results (Zhao et al., 2020). As shown in Table 1, our model produced superior results in all metrics with relatively stable outcomes as the prediction horizon increases.

The spatial model, GCN, demonstrates relatively stable results, where the difference in training results does not dramatically change as the prediction horizon increases. Yet, it still performs poorer than the temporal model when dealing with a short prediction horizon at a 15-minute horizon. Nevertheless, ST-MHA achieved improved prediction accuracy by at least 26.43%, 14.23%, 21.47%, and 18.57% for MSE, RMSE, MAE, and MAPE, respectively, across all prediction horizons.

The encoder-decoder model showed relatively good results for a short prediction period, while it struggles to forecast longer horizons. ST-MHA performed better than the encoder-decoder model in all horizons and metrics. With ST-MHA, MSE, RMSE, MAE, and MAPE are decreased by 12.65%, 6.55%,

Table 1: Prediction results for 15, 30, and 45 minutes.

15 min	encoder-decoder	GCN	TGCN	ST-MHA
MSE	53.59	67.92	54.76	46.81
RMSE	7.32	8.24	7.4	6.84
MAE	4.56	5.63	5.085	4.38
MAPE	13.52%	15.6%	14.086%	12.003%
30 min	encoder-decoder	GCN	TGCN	ST-MHA
MSE	102.07	78.5	77.93	55.88
RMSE	10.1	8.86	8.83	7.47
MAE	6.38	5.94	6.07	4.55
MAPE	20.22%	17.17%	17.83%	13.98%
45 min	encoder-decoder	GCN	TGCN	ST-MHA
MSE	113.72	87.41	88.79	64.3
RMSE	10.66	9.34	9.42	8.01
MAE	7.1	6.24	6.39	4.9
MAPE	20.8%	18.35%	19.3 %	14.44%

3.94%, and 11.22%, respectively, for the 15-minute prediction horizon, while attaining a reduction of 45.25%, 26.03%, 28.68%, 30.86%, respectively, for the 30-minute, and 43.45%, 24.85%, 30.98%, and 30.57%, for the 45-minute forecasting horizon.

Compared to the TGCN model, ST-MHA scored a drop of 14.51%, 7.56%, 13.86%, and 14.78% on MSE, RMSE, MAE, and MAPE, respectively, for the 15-minute prediction horizon. And the enhancement is becoming better for larger prediction horizons. For the 30-minute prediction horizon, for example, ST-MHA achieved a decrease of 28.29%, 15.402%, 25.04%, and 21.59% in MSE, RMSE, MAE, and MAPE, respectively, while for the 45-minute horizon is 27.58%, 14.96%, 23.31%, 25.18%, respectively.

Our algorithm runs fast, making it suitable for real-time traffic prediction. For example, for a 15-minute prediction horizon, it takes 06.66 seconds to predict future values. Given that the sample size is 5 minutes, predictions can be made while new samples are being received, ensuring real-time functionality.

5 CONCLUSIONS

This paper presents the novel Spatio-Temporal Multi-Head Attention model, ST-MHA, to solve the problem of jointly incorporating spatial and temporal traffic characteristics without relying on initial adjacency matrices or utilizing graph structures. Instead, spatial characteristics are captured internally through the modified multi-head attention mechanism, which considers the interactions among traffic speeds over various time segments, referred to as heads. Our model incorporates this information with the temporal component to accurately predict future traffic speeds.

Employing ST-MHA can significantly help to manage and improve traffic conditions, a critical aspect of Intelligent Transportation Systems (ITS). By accurately predicting traffic patterns, authorities gain valuable insights to make well-informed decisions aimed at alleviating congestion and enhancing traffic flow. To ensure the efficiency of our model, ST-MHA is compared with three baseline models: one employs temporal data, another utilizes spatial data, and the third uses spatio-temporal data. Our model demonstrates significant improvement over these models across various prediction horizons in four key metrics: MSE, RMSE, MAE, and MAPE. Although our method is computationally efficient, integrating lower-cost attention techniques, such as sparse attention, could improve its functionality for longer sequences and provide a promising avenue for future research.

REFERENCES

- Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.
- Ahmed, M. S. and Cook, A. R. (1979). *Analysis of freeway traffic time-series data by using Box-Jenkins techniques*. Number 722.
- ArunKumar, K., Kalaga, D. V., Mohan Sai Kumar, C., Kawaji, M., and Brenza, T. M. (2022). Comparative analysis of gated recurrent units (gru), long short-term memory (lstm) cells, autoregressive integrated moving average (arima), seasonal autoregressive integrated moving average (sarima) for forecasting covid-19 trends. *Alexandria Engineering Journal*, 61(10):7585–7603.
- Bai, J., Zhu, J., Song, Y., Zhao, L., Hou, Z., Du, R., and Li, H. (2021). A3t-gcn: Attention temporal graph convolutional network for traffic forecasting. *ISPRS International Journal of Geo-Information*, 10(7).

- Castro-Neto, M., Jeong, Y.-S., Jeong, M.-K., and Han, L. D. (2009). Online-svr for short-term traffic flow prediction under typical and atypical traffic conditions. *Expert systems with applications*, 36(3):6164–6173.
- Chen, W., Chen, L., Xie, Y., Cao, W., Gao, Y., and Feng, X. (2019). Multi-range attentive bicomponent graph convolutional network for traffic forecasting. In *AAAI Conference on Artificial Intelligence*.
- Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014a). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014b). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Dougherty, M. S. and Cobbett, M. R. (1997). Short-term inter-urban traffic forecasts using neural networks. *International journal of forecasting*, 13(1):21–31.
- Gori, M., Monfardini, G., and Scarselli, F. (2005). A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734 vol. 2.
- Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Hua, J. and Faghri, A. (1994). Applications of artificial neural networks to intelligent vehicle-highway systems. *Transportation Research Record*, 1453:83.
- Huang, R., Huang, C., Liu, Y., Dai, G., and Kong, W. (2020). Lsgcn: Long short-term traffic prediction with graph convolutional networks. In *IJCAI*, volume 7, pages 2355–2361.
- Jiang, W. and Luo, J. (2022). Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications*, 207:117921.
- Jin, X., Zhang, Y., and Yao, D. (2007). Simultaneously prediction of network traffic flow based on pca-svr. In *Advances in Neural Networks-ISNN 2007: 4th International Symposium on Neural Networks, ISNN 2007, Nanjing, China, June 3-7, 2007, Proceedings, Part II 4*, pages 1022–1031. Springer.
- Kaffash, S., Nguyen, A. T., and Zhu, J. (2021). Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis. *International Journal of Production Economics*, 231:107868.
- Kamarianakis, Y. and Prastacos, P. (2003). Forecasting traffic flow conditions in an urban network: Comparison of multivariate and univariate approaches. *Transportation Research Record*, 1857(1):74–84.
- Karlaftis, M. G. and Vlahogianni, E. I. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 19(3):387–399.
- Kipf, T. N. and Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations, ICLR '17*.
- Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
- Ledoux, C. (1997). An urban traffic flow model integrating neural networks. *Transportation Research Part C: Emerging Technologies*, 5(5):287–300.
- Lee, S. and Fambro, D. B. (1999). Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting. *Transportation research record*, 1678(1):179–188.
- Levin, M. and Tsao, Y.-D. (1980). On forecasting freeway occupancies and volumes (abridgment). *Transportation Research Record*, (773).
- Levy, J. I., Buonocore, J. J., and Von Stackelberg, K. (2010). Evaluation of the public health impacts of traffic congestion: a health risk assessment. *Environmental health*, 9:1–12.
- Li, M. and Zhu, Z. (2021). Spatial-temporal fusion graph neural networks for traffic flow forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4189–4196.
- Li, Y. and Shahabi, C. (2018). A brief overview of machine learning methods for short-term traffic forecasting and future directions. *Sigspatial Special*, 10(1):3–9.
- Li, Y., Yu, R., Shahabi, C., and Liu, Y. (2018). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*.
- Lu, H., Huang, D., Song, Y., Jiang, D., Zhou, T., and Qin, J. (2020). St-trafficnet: A spatial-temporal deep learning network for traffic forecasting. *Electronics*, 9(9).
- Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.
- Qi, J., Zhao, Z., Tanin, E., Cui, T., Nassir, N., and Sarvi, M. (2022). A graph and attentive multi-path convolutional network for traffic prediction. *IEEE Transactions on Knowledge and Data Engineering*.
- Rahmani, S., Baghbani, A., Bouguila, N., and Patterson, Z. (2023). Graph neural networks for intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 24(8):8846–8885.
- Roy, A., Roy, K. K., Ahsan Ali, A., Amin, M. A., and Rahman, A. K. M. M. (2021). Sst-gnn: Simplified spatio-temporal traffic forecasting model using graph neural network. In Karlapalem, K., Cheng, H., Ramakrishnan, N., Agrawal, R. K., Reddy, P. K., Srivastava, J., and Chakraborty, T., editors, *Advances in Knowledge Discovery and Data Mining*, pages 90–102, Cham. Springer International Publishing.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.
- Su, H., Zhang, L., and Yu, S. (2007). Short-term traffic flow prediction based on incremental support vector

- regression. In *Third international conference on natural computation (ICNC 2007)*, volume 1, pages 640–645. IEEE.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.
- Wang, F.-Y. (2010). Parallel control and management for intelligent transportation systems: Concepts, architectures, and applications. *IEEE Transactions on Intelligent Transportation Systems*, 11(3):630–638.
- Williams, B. M. (2001). Multivariate vehicular traffic flow prediction: evaluation of arimax modeling. *Transportation Research Record*, 1776(1):194–200.
- Williams, B. M. and Hoel, L. A. (2003). Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results. *Journal of transportation engineering*, 129(6):664–672.
- Yuan, T., Da Rocha Neto, W., Rothenberg, C. E., Obraczka, K., Barakat, C., and Turetli, T. (2022). Machine learning for next-generation intelligent transportation systems: A survey. *Transactions on emerging telecommunications technologies*, 33(4):e4427.
- Zhang, K. and Batterman, S. (2013). Air pollution and health risks due to vehicle traffic. *Science of the total Environment*, 450:307–316.
- Zhao, L., Song, Y., Zhang, C., Liu, Y., Wang, P., Lin, T., Deng, M., and Li, H. (2020). T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*, 21(9):3848–3858.
- Zheng, C., Fan, X., Wang, C., and Qi, J. (2020). Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1234–1241.
- Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81.