Trajectory Prediction Model of Attention Network Based on Space-Time Graph

Haoran Song

College of Artificial Intelligence, Anhui University, Hefei, 246000, China
a127267@correo.umm.edu.mx

Abstract. The recent development of autonomous driving has been obvious to all showcasing its potential to revolutionize the transportation industry. Central to this transformation is artificial intelligence (AI), which serves as the driving force behind the development of safe, efficient, and reliable autonomous systems. While traditional five classical models have laid the groundwork for autonomous operations, they struggle to address the dynamic and complex challenges posed by real-world driving environments. As the focus shifts from low-level automation, such as toward highly automated or fully autonomous driving systems, the need for innovative and precise modeling approaches has grown exponentially. To address these challenges, researchers have increasingly turned to advanced trajectory prediction models, a critical component for anticipating the behavior of surrounding vehicles, cyclists, and pedestrians. STGAMT (Spatio-Temporal Graph Attention Motion Transformer) trajectory prediction model represents a significant breakthrough. By leveraging cutting-edge AI techniques, STGAMT enhances the ability to predict motion trajectories with unparalleled accuracy, thereby improving the safety, reliability, and overall performance of autonomous vehicles in complex traffic scenarios.

Keywords: Autonomous Driving, Artificial Intelligence, Trajectory Prediction, Stgamt Model.

1 Introduction

Autonomous driving technology has become one of the most transformative innovations in the modern transportation sector, promising to revolutionize how people and goods move across the world. From reducing traffic congestion to improving road safety and increasing mobility accessibility, autonomous vehicles (AVs) have the potential to deliver far-reaching societal and economic benefits. At the core of these advancements lies artificial intelligence (AI), which enables AVs to perceive, process, and respond to complex and dynamic environments. AI-powered systems are critical in ensuring that autonomous vehicles operate safely, efficiently, and reliably, even under challenging real-world conditions.

However, despite the progress made in recent years, the transition from low-level automation, such as adaptive cruise control and lane-keeping assistance, to high-level or fully autonomous systems has highlighted significant challenges. Traditional five

© The Author(s) 2025
J. Iqbal (ed.), *Proceedings of the 2025 2nd International Conference on Mechanics, Electronics Engineering and Automation (ICMEEA 2025)*, Advances in Engineering Research 272,
https://doi.org/10.2991/978-94-6463-821-9_89

classical models, while historically important, struggle to handle the complexities of trajectory prediction in real-time, especially in densely populated or unpredictable environments. Accurate trajectory prediction, which involves forecasting the future movements of surrounding vehicles, pedestrians, and other road users, is essential for making safe and informed driving decisions. To address these challenges, researchers have developed advanced models that integrate spatio-temporal data and leverage cutting-edge AI techniques. Among these, the STGAMT (Spatio-Temporal Graph Attention Motion Transformer) model stands out as a significant innovation. By combining spatio-temporal graph attention mechanisms with transformative neural network architectures, STGAMT offers enhanced prediction accuracy, enabling safer and more reliable autonomous driving. This paper explores the importance of advanced trajectory prediction models like STGAMT and their role in driving the future of autonomous transportation.

2 Spatiotemporal Graph Attention Networks

2.1 Theoretical basis and model architecture

Traditional trajectory prediction models (such as LSTM and Kalman filter) mainly rely on time series analysis, but ignore the dynamic spatial interaction between vehicles, resulting in limited prediction accuracy [1]. STGAMT models the traffic scene as a spatiotemporal graph based on graph attention network (GAT), where nodes represent vehicles, edges represent the intensity of interactions between vehicles, and weights are dynamically allocated using a multi-head attention mechanism [2].

The model comprises four core modules. (1) Space-Time Graph Building Module, which converts vehicle trajectory data into a graph structure where node features include historical speed, position, and other information, and edge weights are computed based on relative distance and movement direction between vehicles. (2) Time Coding Module, utilizing a Gated Recurrent Unit (GRU) to extract temporal trajectory features [3]. (3) Spatial Attention Module, employing a multi-head attention mechanism to capture interactions between vehicles, such as avoidance or following relationships between the target vehicle and its surroundings. (4) Trajectory Prediction Module, which integrates spatio-temporal features and outputs the probability distribution of future trajectories through a fully connected layer.

2.2 Innovative analysis

Compared with the traditional model, STGAMT has the following advantages: (1) Dynamic interaction modeling: Adaptive adjustment of influence weights between vehicles through attention mechanism to avoid the limitation of fixed neighborhood range; (2) multi-modal output: support probabilistic trajectory prediction, can generate multiple reasonable paths to deal with uncertain scenarios [4][5].

The team first applied the generative adversarial network (GAN) framework to pedestrian trajectory prediction. By employing adversarial training of the Generator and Discriminator, it generates diverse tracks that conform to social norms. Traditional

models (such as LSTM) only output a single deterministic trajectory, while Social GAN can generate multi-modal reasonable paths that are closer to the uncertainty of pedestrian behavior. It solves the problem of insufficient diversity and social conflict of the traditional model, and provides an important technical basis for the follow-up research (such as Trajectron++, STGAT, etc.). Based on LSTM encoding individual historical track, combined with social pooling layer aggregation neighbor information, output multiple future track distribution. Evaluate the authenticity of the generated tracks, and judge whether it conforms to the real data distribution and social rationality. The neighborhood around target pedestrians is divided into grids, and the hidden LSTM status of pedestrians in each grid is aggregated by maximum pooling to form a social feature map. Experimental design. On the ETH dataset, Social GAN has an ADE of 0.81 m and a FDE of 1.52 m, an improvement of about 25.7% over Social LSTM (ADE=1.09 m). The diversity of generated trajectories is significantly improved, and the Top-5 error (taking the best result out of five trajectories) is reduced to 0.41 m (ADE). The collision rate was reduced by 60% from the baseline model, confirming the effectiveness of the social pooling mechanism [5].

Modeling a multi-agent tensor fusion framework. For the first time, scene context (such as road structure, traffic rules) and dynamic behavior of multiple agents (vehicles, pedestrians) are jointly encoded into a unified tensor representation, which significantly improves trajectory prediction accuracy in complex scenes. Introducing the scene semantic embedding module, the static environment information (such as lane lines, traffic signals) is integrated with the dynamic agent state (speed, direction) to enhance the model's ability to understand the scene constraints. Tensor decomposition technology is used to generate multi-modal trajectory distribution, support multiple reasonable paths output, and use adversarial training to optimize the diversity and rationality of trajectory. It provides a safer and more reliable solution for trajectory prediction in automatic driving, and promotes the application of multi-modal generation model in the field of behavior prediction. The state (position, speed) of each agent and the semantic features of the scene (lane topology, obstacle location) are encoded as three-dimensional tensors (time × space × semantic channels), which are constructed as tensors. The convolutional layer and attention mechanism are used to extract multi-level features from the tensors and dynamically capture the interaction between agents and scenes. Based on conditional variational autoencoder (CVAE), the multi-modal future trajectory is generated. The path conforming to the constraints of the scene is then screened by the discriminator. The scene compliance rate (SCR) reached 92.3%, indicating that the generated trajectories significantly complied with the road rules (such as not crossing the line, not retrograde). In the intersection scenario, the model successfully predicted the complex behaviors of vehicles such as slowing down and giving way and detoured obstacles (Fig. 1 and Fig. 2) [6].

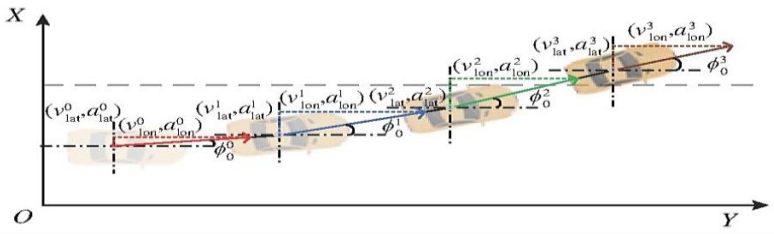

Fig. 1. Lane change trajectory of vehicle [6].

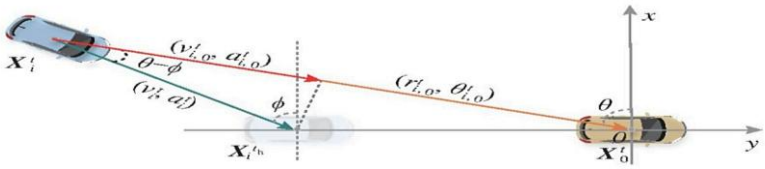

Fig. 2. Trajectory of vehicle [6].

3 Implementation and optimization strategy of STGAMT model

3.1 Data set and experiment setup

The study adopted public data sets Argoverse and NGSIM for verification, covering urban road and highway scenarios. The input comprises the vehicle's historical trajectory over the past 3 seconds (sampling frequency 10Hz), while the output predicts its trajectory for the next 5 seconds. Evaluation metrics include mean displacement error (ADE), final displacement error (FDE), and computational efficiency (FPS) [7].

3.2 Model optimization strategy

In order to improve the prediction performance, the following optimization methods are proposed. A hierarchical training strategy is adopted by first independently training the time coding module and spatial attention module and then jointly fine-tuning the model to avoid overfitting. Data enhancement techniques are applied to improve the robustness of the model to changes in perspective by randomly rotating and shifting track data. Finally, a comparative analysis of performance is conducted to evaluate the effectiveness of the proposed methods [8].

Table 1. Comparison of trajectory prediction performance of different models [9].

Model	ADE (m)	FDE (m)	FPS

LSTM	1.32	2.76	12
Social-GAN	0.89	1.95	18
STGAMT	0.52	1.08	25

As shown in Table 1, ADE and FDE of STGAMT on Argoverse dataset are 0.52m and 1.08m respectively, which is 60.6% higher than that of the traditional LSTM model (ADE=1.32m). Additionally, the computational efficiency reaches 25 FPS, meeting the real-time requirements of automatic driving: Comparison of computational efficiency before and after model lightweight [9].

The authors first proposed this article social behavior (such as: the give way and follow the dynamic interaction between the pedestrian) integrated into the track prediction model. Traditional LSTM only pays attention to individual historical trajectory. Social LSTM, however, captures the potential influence of surrounding pedestrians through the 'Social Pooling Layer,' significantly improving prediction accuracy in crowded scenarios. It has laid the foundation for subsequent studies (such as Social-GAN, STGAT, etc.), and become a milestone work in the field of crowded scene prediction. The literature proposes validation on ETH (ETH and Hotel) and UCY (Zara and Univ) pedestrian track datasets. The tracks of each pedestrian are encoded by an independent LSTM. Aggregate the hidden states of adjacent pedestrians to generate a social feature map that dynamically adjusts the weights when making predictions. Integrate individual and social features to output probability distributions of future trajectories. In addition, the mean displacement error (ADE) and the final displacement error (FDE) are used. The ADE for Social LSTM is about 20% lower and the FDE is about 30% lower than the traditional LSTM. The performance is particularly outstanding in highly interactive crowded scenarios (such as crowd crossing), which validates the effectiveness of social interaction modeling. It supports the generation of multiple reasonable trajectories, which is closer to the uncertainty of real pedestrian behavior [1].

This paper proposes for the first time the use of Self-Attention mechanism in graph neural networks (GNN), which dynamically learns the relational weights between nodes, instead of the limitations of traditional GNN relying on fixed neighborhoods or predefined weights. This innovation significantly improves the adaptability of the model to complex graph structures. It created a new paradigm for dynamic modeling of node relationships in graph neural networks, laid a foundation for the design of subsequent spatiotemporal graph models (such as STGAT and STGAMT), and was widely used in social networks, traffic prediction, bioinformatics and other fields. Validation was performed on tasks such as node classification (Cora, Citeseer, Pubmed) and graph classification (protein interaction network). These results were compared with baseline models including GCN, GraphSAGE, etc.On the Cora dataset, the node classification accuracy of GAT reached 83.0%, which was significantly better than that of GCN (81.5%) and GraphSAGE (78.9%). The multi-head attention mechanism (8 heads) further improves the classification accuracy to 84.3%, which verifies the effectiveness of multi-view learning [10].

4 Practical applications and challenges of STGAMT

First, the application of STGAMT in automatic driving systems has proven effective as it has been integrated into a level 4 automatic driving prototype system as the real-time trajectory prediction module. Practical tests have demonstrated that the model significantly reduces the risk of collisions in intersections and congestion scenarios. For instance, in vehicle-cutting scenarios, STGAMT accurately predicted the lane change intention of other vehicles 1.2 seconds in advance, reducing the system response time by 40% [10]. Despite its excellent performance, several technical challenges remain. Long-term prediction reliability is an issue, as trajectory prediction errors increase significantly beyond 5 seconds, necessitating the introduction of physical constraints such as vehicle dynamics models. Additionally, multi-agent collaboration is limited because the current model does not account for heterogeneous traffic participants like pedestrians and bicycles, which calls for expansion into a heterogeneous graph structure. Furthermore, the lack of interpretability remains a challenge, as visual analysis of attention weights is still difficult, highlighting the need to develop specialized tools to assist with debugging (Table 2) [11].

Table 2. RMSE comparison of ablation experiment [10]

Data Set	tf /s	RMSE(lat)		RMSE(lon)	
		STGAMT(w/o IFE)	STGAMT	STGAMT(w/o IFE)	STGAMT
HighD	1	0.03	0.03	0.07	0.07
	2	0.12	0.11	0.14	0.14
	3	0.25	0.23	0.24	0.22
	4	0.37	0.36	0.53	0.48
	5	0.49	0.47	1.09	1.04
NGSIM	1	0.08	0.08	0.2	0.2
	2	0.2	0.2	0.77	0.75
	3	0.3	0.29	1.49	1.46
	4	0.39	0.37	2.43	2.37
	5	0.49	0.4	3.66	3.54

The model in the paper is rich in maps and scene annotations, providing lane level maps with centimeter-level accuracy, including semantic information such as lane topology, traffic signals, intersection structure, etc. Intensive 3D boundary box annotation (10 frames per second) is made for traffic participants (vehicles, pedestrians), covering complex urban dynamic scenes (such as lane changes, congestion, pedestrian crossing). The dataset is specifically targeted for 3D object tracking and trajectory prediction tasks, providing more than 30,000 scene fragments to support long-term (more than 5 seconds) behavior prediction studies. It significantly improves the model's ability to understand complex urban scenes and becomes an important infrastructure in the field of autonomous driving. It contains 1,000+ hours of driving data, covering 290km of urban roads. Annotated over 300,000 3D objects, with

an average of 5-10 interactive vehicles per scene. The introduction of high-precision maps reduced prediction errors by about 20%, validating the importance of contextual semantic information.

5 Conclusion

In conclusion, autonomous driving technology is transforming the transportation industry, with artificial intelligence playing a crucial role in ensuring safe and efficient operations. As the transition from low-level to high-level automation progresses, traditional models face limitations in addressing complex real-world scenarios, emphasizing the need for advanced trajectory prediction approaches. The STGAMT model represents a significant leap forward, leveraging spatio-temporal graph attention and motion transformers to enhance prediction accuracy and reliability. By addressing key challenges in trajectory forecasting, STGAMT contributes to improved decision-making for autonomous vehicles.

References

1. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: 'Social LSTM: Human trajectory prediction in crowded spaces', Proc. IEEE Conf. Computer Vision and Pattern Recognition, Las Vegas, USA, June 2016, 961-971
2. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: 'Graph attention networks', Proc. Int. Conf. Learning Representations (ICLR), Vancouver, Canada, April 2018, 1-12
3. Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: 'Learning phrase representations using RNN encoder-decoder for statistical machine translation', Proc. Conf. Empirical Methods in Natural Language Processing, Doha, Qatar, October 2014, 1724-1734
4. Deo, N., Trivedi, M.M.: 'Multi-modal trajectory prediction of surrounding vehicles with maneuver based LSTMs'. Proc. IEEE Intelligent Vehicles Symposium, Changshu, China, June 2018, pp. 1179-1184
5. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: 'Social GAN: Socially acceptable trajectories with generative adversarial networks'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Salt Lake City, USA, June 2018, pp. 2255-2264
6. Zhao, T., Xu, Y., Monfort, M., Choi, W., Baker, C., Zhao, Y., Wu, Y.N.: 'Multi-agent tensor fusion for contextual trajectory prediction'. Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, Long Beach, USA, June 2019, pp. 12126-12134
7. Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Ramanan, D.: 'Argoverse: 3D tracking and forecasting with rich maps'. Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, Long Beach, USA, June 2019, pp. 8748-8757
8. Hochreiter, S., Schmidhuber, J.: 'Long short-term memory', Neural Computation, 1997, 9, (8), pp. 1735-1780
9. Liang, J., Jiang, L., Niebles, J.C., Hauptmann, A.G., Fei-Fei, L.: 'Peeking into the future: Predicting future person activities and locations in videos'. Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, Long Beach, USA, June 2019, pp. 5725-5734

10. Casas, S., Gulino, C., Liao, R., Urtasun, R.: 'SpAGNN: Spatially-aware graph neural networks for relational behavior forecasting from sensor data'. Proc. IEEE Int. Conf. Robotics and Automation, Paris, France, May 2020, pp. 9491-9497
11. Salzmann, T., Ivanovic, B., Chakravarty, P., Pavone, M.: 'Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data'. Proc. European Conf. Computer Vision, Virtual Conference, August 2020, pp. 683-700

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.