Attention-Based Interaction Trajectory Prediction

Zhe Liu[1], Lizong Zhang[1,3](✉), Zhihong Rao[1,2], and Guisong Liu[1,4]

[1] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
`l.zhang@uestc.edu.cn`
[2] China Electronic Technology Cyber Security Co., Ltd., Chengdu 610000, China
[3] Trusted Cloud Computing and Big Data Key Laboratory of Sichuan Province, Chengdu 611731, China
[4] School of Computer Science, Zhongshan Institute, University of Electronic Science and Technology of China, Zhongshan 528402, China

Abstract. Trajectory prediction is a hot topic in the field of computer vision and has a wide range of applications. Trajectory prediction refers to predicting the future trajectory of a target based on its past trajectory. This paper proposes a method based on graph neural network and attention mechanism, in order to update trajectory characteristics by implement global pedestrian interaction. And, a direct relationship between history and future is introduced with the attention module for reducing error propagation. The method was evaluated on several real-world crowd datasets, the results demonstrate the effectiveness of our method.

Keywords: Trajectory prediction · Graph attention network · Attention mechanism

1 Introduction

Trajectory prediction is a hot topic in the field of computer vision. Predicting the Trajectories of pedestrians is essential for Self-driving cars and robots. However, the problem of pedestrian trajectory prediction is extremely complicated, due to interdependent of pedestrians. And timing predictions often lead to error accumulation.

Traditional mathematical statistical methods [1, 2] rely on artificially designed features to model pedestrian movements and interactions. With the development of neural networks in recent years, the methods based on Neural Network have surpassed the traditional methods and achieved better results. Trajectory prediction methods based on deep learning mostly use LSTM [3] model to encode and decode each pedestrian trajectory sequence in the scene using LSTM network to obtain the trajectory characteristics of each pedestrian.

When dealing with the interactions between pedestrians, the above methods mostly deal with each pedestrian individually, which increase the amount of calculation and time cost. On the other hand, the error propagation problem is not solved properly. The

R. Xu et al. (Eds.): AIMS 2020, LNCS 12401, pp. 168–175, 2020.
https://doi.org/10.1007/978-3-030-59605-7_13

errors of the previous time step will be amplified and affect the result of the subsequent time step.

Inspired by the attention mechanism in natural language processing [4], a novel method for trajectory prediction is proposed. We assign different weight coefficients to historical features based on the attention mechanism and fuse historical features.

Contributions of this paper are summarized as follows:

- Proposed a pedestrian interaction processing method based on graph neural network, which realizes feature fusion between pedestrians, and reduces the time consumption.
- Proposed an attention-based historical feature fusion method that adaptively selects the historical trajectory characteristics to improve the accuracy of prediction.
- Evaluated on several publicly available real-world crowd datasets. And the results demonstrate the effectiveness of our method.

The remainder of this paper is organized as follows. Section 2 reviews main work related to trajectory prediction. Section 3 describes the proposed model. Section 4 evaluates the effectiveness of the framework while conclusions and suggestions for future work are summarized in Sect. 5.

2 Related Work

At present, trajectory prediction methods [1, 2] include traditional mathematical statistics methods and methods based on deep learning. Traditional mathematical statistical methods rely on artificially designed features to model pedestrian movements and interactions. The bottlenecks of traditional mathematical statistics methods are that they cannot consider long-term dependence on information and adapt to complex mobile scenarios.

In recent years, methods based on deep learning have proven to be superior to traditional mathematical statistical methods. Among them, the methods based on LSTM and the methods based on GAN [5] are the most representative. Because the nature of pedestrian trajectories is a set of natural motion sequences with time series characteristics, some methods [6–8] mainly build models based on recurrent neural networks. Alahi et al. [6] have introduced the social pooling layer to bring the hidden states of neighboring pedestrians together to form interaction features, and achieved the purpose of modeling pedestrian interaction. The social pooling layer meshes the target scene and pools the hidden layer features of other pedestrians in the neighborhood of each pedestrian grid according to the grid range. At the same time, the results generated by the GAN-based method [9, 10] perform well in authenticity and diversity. Gupta et al. [9] have proposed a trajectory sampler that handles the interactions between all the observed pedestrians by pooling the GAN input random vector with a vector combining the hidden representations of the other pedestrian trajectories.

Deep Learning on Graphs Generalizing neural networks to graph-structured data is an emerging topic. Recently, researchers apply attention mechanisms to graph-structured data [11] to model spatial correlations for graph classification. Regarding the attention mechanism [4], the literature proposes a network structure called “transformer” based on the attention mechanism to mine the relationship between input and output.

Inspired by the above, we model the pedestrians in the scene as graph nodes and implement pedestrian interaction through graph attention networks. We model the direct relationship between history and future time steps to mitigate the problem of error propagation.

3 Methodology

3.1 Problem Definition

Suppose pedestrians are represented by P_1, $P_2 \ldots$ P_N. The position of pedestrian P_i at time-step T is denoted as (X_{Ti}, Y_{Ti}). The problem is defined to predict the trajectories (X_{Ti}, Y_{Ti}), where $\mathrm{T} = \mathrm{T}_{\mathrm{obs+1}}, \mathrm{T}_{\mathrm{obs+2}} \ldots \mathrm{T}_{\mathrm{pred}}$.

3.2 Overall Flowchart

The overall network architecture is shown in Fig. 1. First, we use the LSTM network to extract the trajectory features, and merge the trajectory features into the graph attention network to obtain the global interaction features. Then, based on the attention mechanism, the trajectory features of multiple time steps are fused to obtain the final trajectory features. Finally, the trajectory features are input into the LSTM network to obtain the predicted position.

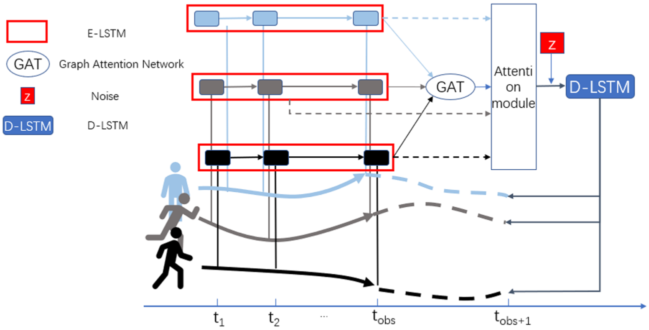

Fig. 1. Overall network architecture

3.3 Extract Trajectory Features

At first, we embed the location of each person to get a fixed length vector d_{Ti} and input it to the E-LSTM.

$$d_{Ti} = F\left(x_i^t, y_i^t\right) \tag{1}$$

$$S_i^t = \mathrm{E} - LSTM\left(S_i^{t-1}, d_{Ti}; W_E\right) \tag{2}$$

where the function $F(\cdot)$ is an embedding function. S_i^t is the hidden state of the E-LSTM at time-step t. W_E is the weight of the E-LSTM.

3.4 Pedestrian Interaction

We consider the pedestrians as nodes on the graph, and use GAT [11] to implement our pedestrian interaction mechanism. $S_i^{t=T_{obs}}$ is input into the attention layer of the graph, and the attention coefficient between node pairs (i, j) is calculated:

$$\alpha_{i.j}^t = \frac{exp(LeakyRelu(\vec{a}^T [WS_i^t || WS_j^t]))}{\sum k \in N_i exp(LeakyRelu(\vec{a}^T [WS_i^t || WS_k^t])} \tag{3}$$

Where || is the concatenation operation, $\alpha_{i.j}^t$ is the attention coefficient of node j to i, N_i represents the neighbors of node i on the graph, W is the weight matrix, a is a single-layer feedforward neural network, parametrized by a weight vector $\vec{a}^T$. It is normalized by a softmax function with LeakyReLU.

After the standardized attention coefficient is obtained, the output of the attention network of node i graph is calculated by the following formula:

$$\hat{S}_i^t = \sigma\left(\sum j \in N_i \alpha_{i.j}^t W S_i^t\right) \tag{4}$$

Where σ is a nonlinear function, $\hat{S}_i^t$ is the state of pedestrian i after merging the characteristics of the surrounding pedestrian trajectory.

3.5 Trajectory Feature Fusion

To ease the error propagation, we added the attention module. The attention module models the direct relationship between each future time steps and historical time steps to generate the pedestrian final trajectory feature.

For pedestrian i, the correlation between the time step $t_P\left(t_P = T_{obs}, \ldots T_{obs+Q}\right)$ and the historical time steps t(t = $T_1, \ldots T_{obs-1}$) is calculated.

$$\mu_{t_P,t} = \hat{S}_i^{t_p}, S_i^t \tag{5}$$

$$\gamma_{t_P,t} = \frac{exp\left(\mu_{t_P,t}\right)}{\sum_{t_r=T_1,}^{T_{obs-1}} exp\left(\mu_{t_P,t_r}\right)} \tag{6}$$

Where $\langle \cdot, \cdot \rangle$ denotes the inner product operator, $\gamma_{t_P,t}$ is the attention score. The trajectory characteristics of pedestrian i are calculated by the following formula:

$$\hat{P}_i^t = \left(\sum_{t=T_1}^{T_{obs-1}} \gamma_{t_P,t} S_i^t\right) || \hat{S}_i^{t_p} \tag{7}$$

In many cases, the trajectory of a pedestrian is multi-modal. And different people may choose different modes of action. In order to enhance the diversity of the generated trajectory, the pedestrian trajectory characteristics are added to the noise vector and input to the LSTM network to decode and obtain the prediction result.

$$r_i^{T_{obs}} = \hat{P}_i^t || Z \tag{8}$$

$$r_i^{T_{obs}+1} = D - LSTM\left(r_i^{T_{obs}}, d_i^{T_{obs}}; W_D\right) \tag{9}$$

$$\left(x_i^{T_{obs}+1}, y_i^{T_{obs}+1}\right) = \delta_3\left(r_i^{T_{obs}+1}\right) \tag{10}$$

Where Z represents noise, $r_i^{T_{obs}}$ represents intermediate state vector, $\delta_3(\cdot)$ is a function, W_D is the D-LSTM weight. After the predicted position is obtained at time step T_{obs+1}, the subsequent time step location is predicted.

For each pedestrian, the model produces multiple predicted trajectories by randomly sampling z from N (0,1) (the standard normal distribution). And select the trajectory with the smallest distance from the real position as the model output to calculate the loss.

$$L = min_k\left|\left|P_i - \hat{P}_i^k\right|\right|_2 \tag{11}$$

Where P_i is the ground-truth trajectory of pedestrian i, $\hat{P}_i$ is the trajectory produced by our model, k is a hyperparameter.

4 Experiment

Training Samples. We evaluate our proposed model on two public pedestrian walking datasets, ETH [12] and UCY [13], which contain rich social interactions. The ETH dataset consists of two scenarios named ETH and HOTEL. UCY dataset includes two scenarios and in three components, named ZARA-01, ZARA-02 and UCY. These datasets contain thousands of real-world pedestrian trajectories and cover rich human-human interactions. We evaluate our model on these 5 datasets. We follow the leave-one-out evaluation methodology in [9].

Parameter Settings. We iteratively train the network with a batch size of 64 for 300 epochs using Adam optimizer with a learning rate of 0.01. The dimensions of the hidden state for LSTM is 32. For first graph attention layer, the shape of W is 16 × 16. For the second layer, the shape of W is 16 × 32. Batch Normalization is applied over the input of graph attention layer. The dimension of Z is set to 16.

Evaluation Index. There are two types of metrics for evaluating the performance of trajectory prediction, including the Average Displacement error (ADE) and Final Displacement error (FDE) in meters.

1. Average Displacement Error (ADE): Average L_2 distance between ground truth and our prediction over all predicted time steps.
2. Final Displacement Error (FDE): The distance between the predicted final destination and the true final destination at end of the prediction period T_{pred}.

Quantitative Evaluation. All experiments are based on ETH and UCY datasets. The results and analysis are as follows.

Table 1. Comparative experiment on attention module.

Metric	Dataset	OURS	OURS
Attention		×	✓
ADE	ETH	0.56	0.59
	HOTEL	0.27	0.25
	UNIV	0.31	0.31
	ZARA1	0.21	0.21
	ZARA2	0.20	0.20
AVG		0.31	0.31
FDE	ETH	1.10	1.15
	HOTEL	0.43	0.42
	UNIV	0.66	0.64
	ZARA1	0.42	0.42
	ZARA2	0.40	0.40
AVG		0.60	0.61

In Table 1, We evaluated the attention module of the experiment. × indicates that the attention module has been removed from the network. The results show that the effect is improved on some data sets, which shows that the attention module has a certain effect. But on the ETH dataset, the error becomes larger, which is related to the number of pedestrians' historical time step. Compared with other datasets, the average residence time of pedestrians in the ETH dataset is shorter.

In Table 2, we evaluate our model against all baseline models. The results show that our method outperforms all compared methods on all datasets. Compared with S-LSTM and SGAN, the ADE is reduced by 31% and 18% respectively. For FDE, the performance is increased by 32% and 21% respectively. These results show that our model has advantages compared to other methods.

Figure 2 shows some examples of predicted trajectories drawn from datasets. The predicted paths of our models appear able to better capture the direction of pedestrian movement. The generated trajectories do not have a linear trend, and the model also performs well in the case of multiple pedestrians.

Table 2. Comparison with several baseline models.

Metric	Dataset	S-LSTM	SGAN	OURS
ADE	ETH	0.73	0.60	0.59
	HOTEL	0.49	0.48	0.25
	UNIV	0.41	0.36	0.31
	ZARA1	0.27	0.21	0.21
	ZARA2	0.33	0.27	0.20
AVG		0.45	0.38	0.31
FDE	ETH	1.48	1.19	1.15
	HOTEL	1.01	0.95	0.42
	UNIV	0.84	0.75	0.64
	ZARA1	0.56	0.42	0.42
	ZARA2	0.70	0.54	0.40
AVG		0.91	0.77	0.61

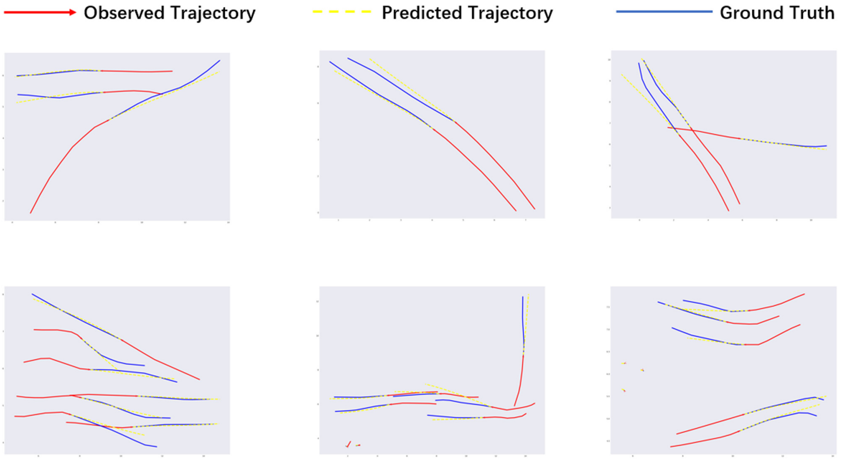

Fig. 2. Trajectories generated by our model

5 Conclusion

In this paper, we propose a novel method for the prediction of pedestrian trajectories. We use the graph attention network to handle global pedestrian interaction. Furthermore, we use the attention module to select and fuse historical features. Experimental results show that the attention mechanism effectively reduces the error propagation and improves prediction results. Test results on two datasets prove that our method can effectively improve prediction accuracy. We have noticed that the attention module failed to get the expected result on the ETH dataset, the issue will be further analyzed to improve our method in the future work.

Acknowledgement. This work was supported in part by the National Key R&D Program of China (No. 2018YFC0807500), and by Ministry of Science and Technology of Sichuan Province Program (No. 2018GZDZX0048, 20ZDYF0343, 2018HH0075).

References

1. Elnagar, A.: Prediction of moving objects in dynamic environments using Kalman filters. In: Proceedings of the International Symposium on Computational Intelligence in Robotics and Automation (CIRA), pp. 414–419 (2001)
2. Barth, A., Franke, U.: Where will the oncoming vehicle be the next second? In: Proceedings of the IEEE Intelligent Vehicles Symposium (IV), pp. 1068–1073 (2008)
3. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
4. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 6000–6010 (2017)
5. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NIPS), pp. 2672–2680 (2014)
6. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social LSTM: human trajectory prediction in crowded spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 961–971 (2016)
7. Vemula, A., Muelling, K., Oh, J.: Social attention: modeling attention in human crowds. In 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 1–7. IEEE (2018)
8. Xu, Y., Piao, Z., Gao, S.: Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5275– 5284 (2018)
9. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social GAN: socially acceptable trajectories with generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Number CONF (2018)
10. Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Savarese, S.: SoPhie: an attentive GAN for predicting paths compliant to social and physical constraints. arXiv preprint arXiv:1806.01482 (2018)
11. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
12. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: modeling social behavior for multi-target tracking. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 261–268. IEEE (2009)
13. Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by example. In: Computer Graphics Forum, vol. 26, pp. 655–664. Wiley Online Library (2007)