STIRNet: A Spatial-temporal Interaction-aware Recursive Network for Human Trajectory Prediction

Yusheng Peng, Gaofeng Zhang, Xiangyu Li, and Liping Zheng
Hefei University of Technology, Hefei, China, 230601

Abstract

Pedestrian trajectory prediction is one of the important research topics in the field of computer vision and a key technology of autonomous driving system. However, it's full of challenges due to the uncertainties of crowd motions and complex interactions among pedestrians. We propose a Spatio-temporal Interaction-aware Recursive Network (STIRNet) to predict multiply socially acceptable trajectories of pedestrians. In this paper, a recursive structure is used to capture spatio-temporal interactions by spatial modeling and temporal modeling alternately. At each time-step, the spatial interactions are modeled by a graph attention network, in which the nodes feature are represented by temporal motion features. The learned spatial interaction context is used to capture temporal motion features through an LSTM model. The temporal motion features are used to infer future positions and update nodes features. Experimental results on two public pedestrian trajectory datasets (ETH and UCY) demonstrate that our proposed model achieves superior performances compared with state-of-the-art methods on ADE and FDE metrics.

Introduction

Pedestrian trajectory prediction is of major significance in several applications such as autonomous driving, robot navigation, and surveillance systems. For example, in surveillance systems, forecasting pedestrian trajectories is critical in helping identify suspicious activities.

In recent years, with the development of deep learning, the deep neural networks including LSTM, GAN are widely used in pedestrian trajectory prediction and achieve great success. In such deep learning prediction methods, pooling mechanisms, attention mechanisms and graph neural network mechanisms are often used to model the complex and subtle social interaction among pedestrians. In the view that pedestrians have different impacts on each other, some of the pooling mechanisms and graph neural network mechanisms incorporate attention mechanisms to model social interactions.

However, most of the models focus on modeling spatial interactions among pedestrians. Xu et. al. design a spatio-temporal attention module to model the spatio-temporal interactions among pedestrians. In contrast, STGAT and AST-GNN models model spatial interactions firstly and then feed the spatial interaction contexts to the temporal model to capture the spatio-temporal interaction features. Inspired by these works, we adopt a novel recursive structured network via graph attention network and LSTM to model spatio-temporal interactions.

In this paper, we propose a Spatio-temporal Interaction-aware Recursive Network (STIRNet) for pedestrian trajectory prediction. A GAT is adopted to model spatial interactions among pedestrians at each time-step, where the nodes features are represented by temporal motion features. Besides, the output spatial interaction contexts of GAT are fed to the LSTMs to capture temporal motion features. The learned motion features are used to infer future positions and update nodes features at next time-step.

Methods

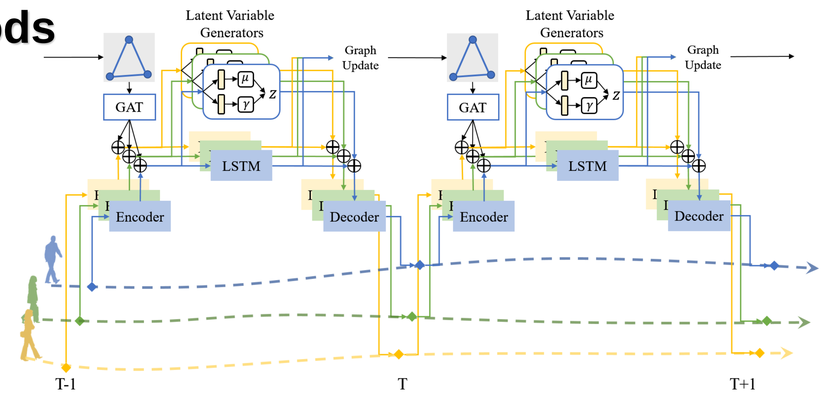

Fig. 1 The architecture of the proposed STIRNet model.

The STIRNet model is based on a recursive framework. For each time-step, the encoders embed the positions to high-dimensional features and the decoders are designed for inferring future positions from high-dimensional features. The GAT is employed to model spatial interactions from nodes features. Then the spatial interaction context is coupled with the encoding from the encoder and fed to the LSTM to capture motion feature. Besides, we design a VAE-based latent variable generator to generate latent variables in the training stage to encourage the model to predict multiply socially acceptable positions in the test stage.

Table 1. Comparison with baseline models on ADE & FDE evaluation metrics.

Model	Performance (ADE/FDE) ↓					
	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVERAGE
S-LSTM	1.09/2.35	0.79/1.76	0.67/1.40	0.47/1.00	0.56/1.17	0.72/1.54
CIDNN	1.25/2.32	1.31/2.36	0.90/1.86	0.50/1.04	0.51/1.07	0.89/1.73
SGAN	0.81/1.52	0.72/1.61	0.60/1.26	0.34/0.69	0.42/0.84	0.58/1.18
SoPhie	0.70/1.43	0.76/1.67	0.54/1.24	0.30/0.63	0.38/0.78	0.54/1.15
IDL	0.59/1.30	0.46/0.83	0.51/1.27	**0.22/0.49**	**0.23/0.55**	0.40/0.89
STGAT	0.65/1.12	0.35/0.66	0.52/1.10	0.34/0.69	0.29/0.60	0.43/0.83
RAMP	0.69/1.24	0.43/0.87	0.53/1.17	0.28/0.61	0.28/0.59	0.44/0.90
TPNet	0.84/1.73	0.24/0.46	**0.42/0.94**	0.33/0.75	0.26/0.60	0.42/0.90
NMMP	0.61/1.08	0.33/0.63	0.52/1.11	0.32/0.66	0.29/0.61	0.41/0.82
STIRNet	**0.48/0.95**	**0.22/0.41**	0.54/1.15	0.37/0.80	0.31/0.70	**0.38/0.80**

Results-1

We compare our method with the state-of-the-art baselines mentioned above. All the stochastic method samples 20 times and reports the best-performed sample. The main results are presented in Table 1. The S-LSTM, CIDNN, and the proposed STIRNet are recursive structured models while the rest of baselines are seq2seq models. The performance of STIRNet model is best on ETH and HOTEL datasets and compatible on the rest 3 datesets. STIRNet improves the state-of-the-art prediction to 0.38m and 0.80m on ADE and FDE on average. Particularly, the SoPhie, RAMP, and TPNet models adopt scene information in modeling, but our model achieves better performance without using scene information compared with these models.

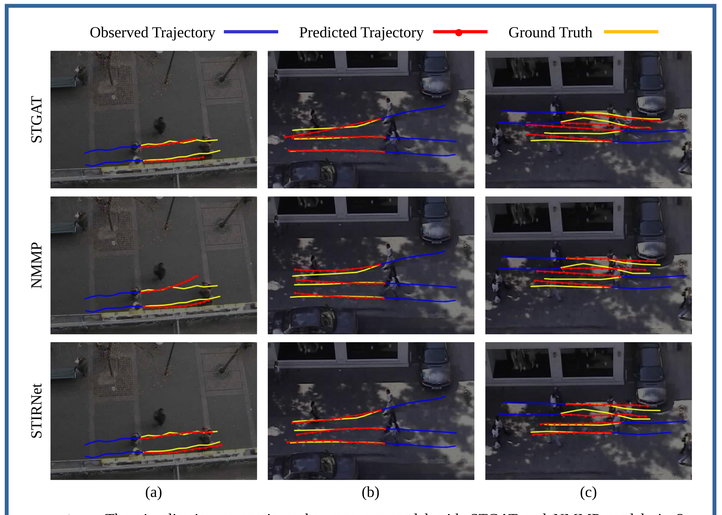

Fig. 2 The visualization comparisons between our model with STGAT and NMMP models in 3 different scenarios, which contain parallel walking (a), people merging (b), and people meeting (b).

Results-2

To verify the effectiveness of our model, we illustrate the prediction trajectories of 4 examples which come from three types of social scenario. The Fig. 2(a) shows the parallel walking scenario where two pedestrians are walking in parallel. The trajectories generated by our model are closer to the ground truth while the trajectories predicted by STGAT and NMMP are deviated and fail to reach the endpoints. In people merging Fig. 2(b) and people meeting Fig. 2(c) scenarios, the trajectories predicted by our method are also closer to the ground truth and without collisions and crowding happening. These examples prove that the proposed spatio-temporal interaction modeling is more effective and successful than that of the STGAT model.

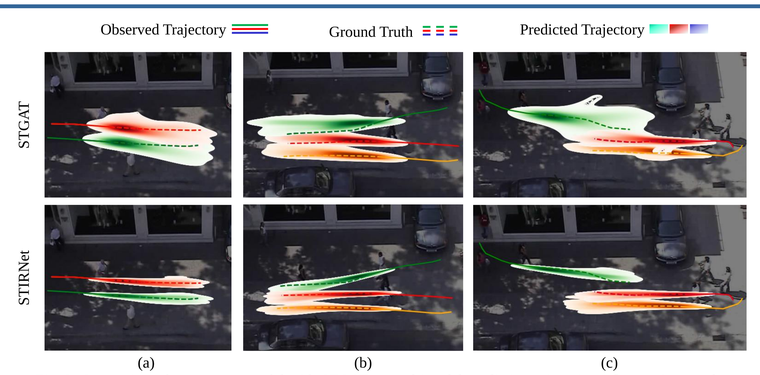

Fig. 3 Comparisons between our model with STGAT on multimodal predictions. A variety of scenarios are shown: two individuals walking in parallel (a), two persons merging from the same direction (b), and two persons meeting from different directions (c).

Results-3

We also compare the proposed model with STGAT in 3 common social scenarios on multimodal prediction performance. For the multimodal predictions of the STIRNet model, the ground truth trajectories are always distributed in the high density regions (deep color). Compared with the multimodal prediction of STGAT, the multiple trajectories generated by STIRNet are more concentrated and clustered. However, a wider distribution of future predictions means that there is more randomness in the prediction, which is not what we want. Therefore, the prediction distribution generated by STIRNet is more concentrated, which is more efficient.

Conclusion

In this work we focus on modeling spatio-temporal interaction and jointly predicting trajectories for all people in a scene. We propose a novel spatio-temporal interaction-aware recursive network to predict multimodal socially acceptable trajectories. The ablation studies prove the validity of the proposed spatio-temporal modeling with alternative recursive manner in pedestrian trajectory prediction. The quantitative and qualitative comparisons also verify the effectiveness of the proposed model and outperforms other SOTA methods. Although the proposed STIRNet achieves the state-of-the-art prediction, the inference speed is far less than other models. In future work, we will transfer the proposed spatio-temporal interaction modeling to seq2seq structured model to improve the inference speed.