Comparative Study on General Trajectory Forecasting Framework [‡]

Chengzhi Yuan[*†]

October, 2021

Abstract

This report conducts a comparative study on the general trajectory forecasting framework, to explore the effectiveness of varied length of observation sequences, typical attention mechanism, and the multi-modal characteristics of current models, respectively. Extensive experiments are carried out for validation. Finally, the potential cues and discussion are given based on the experimental results.

1 Introduction

Trajectory prediction is an emerging topic in the field of computer vision, which has drawn great attentions in recent years. Most of the approaches applied nowadays are built upon the standard encoder-decoder scheme, where recurrent neural networks are usually considered as the fundamental block to model trajectory dynamics (such as RNN, LSTM, and GRU, etc). Moreover, social interaction factors (such as human-human interactions and human environment interactions) and multi-modal trajectory generation can be taken into account as well, for the further promoting the prediction performances.

Although a lot of efforts have been spent by the research community in the recent years, there still exist many challenging issues that have not been solved completely, where we summarize several essential problems hereafter:

- Longer observation history may lead to better prediction accuracy. However, some researchers argue that long-term observations may be not as relevant as commonly believed in the prediction task.

- Current attention modules that used for interaction modeling merely rely the relative distances between the target pedestrian and the agents in his neighborhood (such as near and far), which is not adequate to address the problem, since motion direction and speed are important as well.

- The uncertainty of future trajectories can be described by multi-model distributions. Although generative models (such as GAN, VAE) have shown strong potentials in producing trajectory data with multiple modes, they still cannot fully learn the real data distributions explicitly.

[*] ACoTAI, Dalian Maritime University

[†] Email: yuanchengzhi12@gmail.com

[‡] This work is based on project "Computer Vision-based Human Behavior Understanding in Complex Scenarios" and project "Research on Vision-based Behavior Forecasting".

This report provides a comparative study on the general encoder-decoder framework for trajectory forecasting, and investigate on the popular tricks that being widely adopted in the recent years. To be specific:

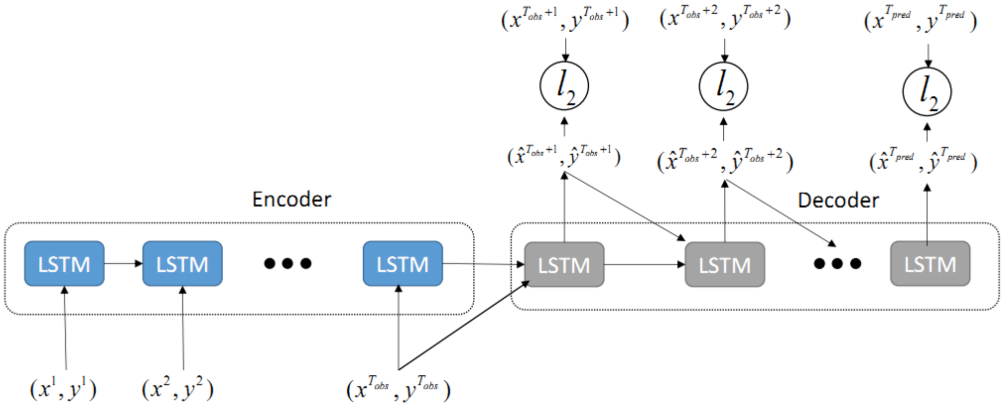
- Using varied lengths of observation/prediction sequences, where experimental results show that the latest observations may contribute to the forecasting task significantly.
- Comparing several typical attention mechanisms for social interaction modeling (such as the social pooling, social attention, and the graph attention), and evaluate how they can influence the prediction performances.
- Investigating on generative models for multi-modal trajectory generation. A simplified framework is proposed, which comprised by the multi-layer perceptron the long-short term memory, and it can achieve similar results as the generative model.

2 Methodology and Experiments

2.1 Encoder-decoder Framework

Most of predictive model is trained in an auto-regressive way Akaike [1998], where model produces current prediction given last output as input.

The widely used encoder-decoder pipeline is displayed in Fig. 1. The encoder takes the observation sequences as input, constructs a hidden representation of it, and the decoder aims to predict future trajectories step-by-step, both of them utilize RNN as components.



$(x^{T_{obs}+1}, y^{T_{obs}+1})$
$(x^{T_{obs}+2}, y^{T_{obs}+2})$
$(x^{T_{pred}}, y^{T_{pred}})$
l_2
l_2
l_2
$(\hat{x}^{T_{obs}+1}, \hat{y}^{T_{obs}+1})$
$(\hat{x}^{T_{obs}+2}, \hat{y}^{T_{obs}+2})$
$(\hat{x}^{T_{pred}}, \hat{y}^{T_{pred}})$
Encoder
Decoder
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
(x^1, y^1)
(x^2, y^2)
$(x^{T_{obs}}, y^{T_{obs}})$

Figure 1: The general encoder-decoder architecture for trajectory prediction.

2.2 Ablation Study on Varied Observation Sequence

The correlation coefficient matrix is adopted to measure the similarity of motion states among multiple time steps, and it can be further divided into two subsets corresponding to the horizontal and the vertical directions.

In Fig. 2, one single position is closely related to others in the neighborhood in temporal dimension, instead of early observations in the sequence. Especially, the latest observations usually have higher correlation values.

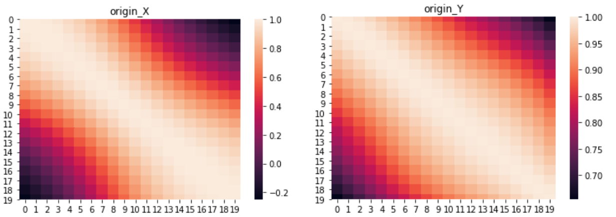


Figure 2: Examples of the correlation coefficient matrix, where trajectories are collected from the ETH dataset. Left: the correlation coefficient matrix in the horizontal direction. Right: the correlation coefficient matrix in the vertical direction.

Furthermore, a comprehensive evaluation is carried out on varied lengths of the observation sequences in the prediction task in Table 1 to further validate the perspective. The experiments are carried out on the ETH dataset and UCY dataset Lerner et al. [2007].

Dataset	LSTM				Social LSTM			
	T_{obs}=1	T_{obs}=2	T_{obs}=3	TT_{obs}=4	T_{obs}=1	T_{obs}=2	T_{obs}=3	T_{obs}=4
eth	1.07/2.01	1.04/2.01	1.03/2.07	0.99/2.01	1.12/2.09	1.11/2.06	1.13/2.04	1.10/2.04
hotel	0.39/0.84	0.38/0.84	0.40/0.88	0.42/0.92	0.76/1.96	0.76/1.96	0.76/1.96	0.75/1.92
univ	0.76/1.51	0.74/1.47	0.72/1.44	0.69/1.41	0.66/1.01	0.67/0.99	0.66/0.97	0.66/0.97
zara-1	0.53/1.06	0.52/1.05	0.50/1.02	0.47/1.00	0.47/0.73	0.46/0.72	0.46/0.73	0.46/0.73
zara-2	0.42/0.85	0.41/0.84	0.40/0.84	0.37/0.80	0.56/1.13	0.56/1.12	0.55/1.09	0.56/1.10
avg	0.63/1.25	0.62/1.24	0.61/1.25	0.59/1.12	0.71/1.38	0.71/1.37	0.71/1.35	0.71/1.35

Table 1: Evaluations on varied lengths of observation/prediction sequences.

2.3 Ablation Study on Typical Attention Mechanism

Human motion behaviors are heavily influenced by their neighbors agents, particularly in crowded scenes. In recent years, the social-aware attention modules are widely used for interaction modeling (such as: Social LSTM Alahi et al. [2016], Social GAN Gupta et al. [2018], Social Attention Vemula et al. [2018], etc.), which represent interaction behaviors through combining information from all neighboring states.

In this part, a ablation study is carried out to investigate the performance of several widely used social-aware attention mechanisms in Table 2. Experiments are carried out on the ETH and the UCY datasets.

Dataset	Social LSTM		Social GAN		Social Attention	
	with attention	without attention	with attention	without attention	with attention	without attention
eth	1.10/2.01	1.08/1.82	1.09/2.11	0.95/1.94	0.44/0.43	0.41/0.56
hotel	0.75/1.91	0.74/1.88	1.06/2.17	0.64/1.34	0.36/0.40	0.38/0.45
univ	0.66/0.95	0.63/0.92	0.95/1.86	0.64/1.35	0.48/0.96	0.48/0.72
zara-1	0.46/0.71	0.43/0.69	0.56/1.14	0.44/0.96	0.18/0.03	0.15/0.05
zara-2	0.56/1.10	0.52/1.03	0.50/1.02	0.33/0.71	0.28/0.37	0.29/0.34
avg	0.71/1.34	0.68/1.09	0.83/1.66	0.60/1.26	0.35/0.38	0.34/0,42

Table 2: Evaluations on varied spatial attention mechanisms in terms of ADE/FDE metrics, the widely used interaction modeling mechanisms may not contribute to promote performances.

Even though current interaction modeling methods may not help minimize the displacement error, the collision rate may be promoted.

2.4 Multi-modal trajectory generation

A pedestrian's behaviors are likely to demonstrate multi-modal characteristics. For example, when a pedestrian approaches a traffic junction, he has multiple options, such as going straight, turning left/right, or coming back, which caused by personal preferences, intentions, and environmental factors, etc. Generative models (GANs, VAEs, etc.) are widely used for multi-modal modeling through learning data distribution.

Although generative models can produce trajectories with some different directions, they often suffer from the out-of-distribution (OOD) problems due to the continuous distribution of inputs Khayatkhoei et al. [2019]. In order to fully address the multi-modal characteristics of future trajectories, a generative framework comprised by multiple generators is acquired Dendorfer et al. [2021], or utilize inputs with several discontinuous distribution Gurumurthy et al. [2017].

To further validate the assumption as mentioned, a simplified model is proposed, which is built upon MLP and LSTM. The overall architecture are shown in Fig. 3. Furthermore, an ablation study on the synthesis data is conducted in Fig. 4. The blue lines indicate the observation trajectories, the red lines and green lines represent the multi-direction predicted trajectories and ground truth trajectories, respectively.

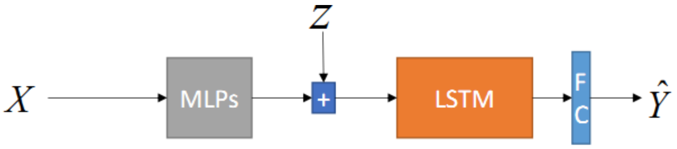


Figure 3: The MLP+LSTM architecture for multi-modal trajectory prediction.

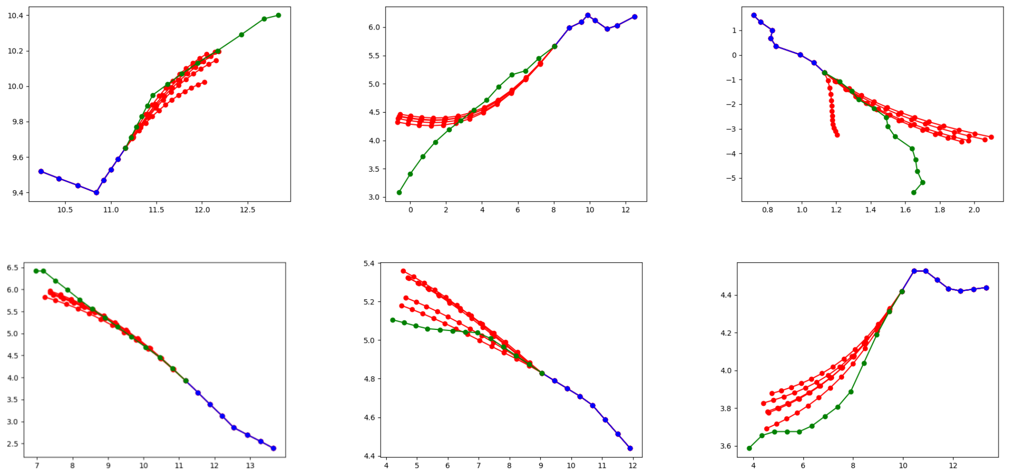


Figure 4: The predicted multi-modal trajectories produced by MLP+LSTM framework.

3 Conclusion and Discussion

Conclusion1: The shorter length of observation sequence may contribute equal performance as the longer ones do, which may caused by the short term memory of RNN-based model.

Conclusion2: The widely used social-aware attention mechanisms may not be effective in prediction performances. One possible reason is that, most of works aim to reduce the displacement error rather than the collision rate.

Conclusion3: The multi-modal characteristics of future trajectories cannot be fully address by common used generative models due to the inputs with continuous distribution. The multi-generator based structure and inputs with discontinuous distribution may overcome this problem.

References

Hirotugu Akaike. Autoregressive model fitting for control. In *Selected Papers of Hirotugu Akaike*, pages 153–170. Springer, 1998.

A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. *Computer Graphics Forum*, 26:655–664, 2007.

Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social LSTM: Human trajectory prediction in crowded spaces. In *the Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 961–971. IEEE, 2016.

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social GAN: Socially acceptable trajectories with generative adversarial networks. In *the Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2255–2264. IEEE, 2018.

Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *the Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1–7. IEEE, 2018.

Mahyar Khayatkhoei, Ahmed Elgammal, and Maneesh Singh. Disconnected manifold learning for generative adversarial networks, 2019.

Patrick Dendorfer, Sven Elflein, and Laura Leal-Taixé. Mg-gan: A multi-generator model preventing out-of-distribution samples in pedestrian trajectory prediction, 2021.

Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and Venkatesh Babu Radhakrishnan. Deligan : Generative adversarial networks for diverse and limited data, 2017.