SAGAN: SGAN with Trajectory Attention

Junda Wu
College of Mathematics and Statistics
Chongqing University
Chongqing, China
joshua19801010@gmail.com

Enlin Shen
Department of Applied Mathematics
Xian Jiaotong-liverpool University
Suzhou, China
1305182452@qq.com

Xichong Ling
Department of Electrical and Computer Engineering
McGill University
Montréal, Canada
xichong.ling@mail.mcgill.ca

October 4, 2020

Abstract

When we are to navigate in a human centred environment, it is essential to understand human movement behaviour for autonomous technology. The human movement is unpredictable due to its multimodal nature: given the history of human movement paths, there are many socially viable ways for people to move in the future, which increase the difficulty of modelling. Previous works have showed the capacity of combining LSTM and GAN to forecasting pedestrians' future trajectories with the socially-aware pooling mechanism to aggregate information across people. Our work used attention mechanism to make the sequence model focus on more relevant time steps, thus enabling to include longer observation time. Such mechanism successfully improved SGAN on almost every dataset, especially when the time horizon is much longer.

1 Introduction

Motion prediction of humans is a critical technology in the intelligent era nowadays, which is applied in autonomous mobile area. After all, the newly coming inventions sharing the same ecosystem human. Therefore, these machines are required to manage complex social interactions effectively as human do. Based on this purpose, the vital mission is to predict all possible future trajectories with a known motion trajectory of the pedestrian.

Numbers of previous works have modelled trajectory prediction problems based on Recurrent Neural Networks [1, 2]. RNN is a natural way to use the parameterised model to approximate series of actions and make predictions based on model's inference. Although such prediction models have achieved great successes in various series prediction problems, long-term relationship can be hard to obtain due to gradient vanishing problems. Therefore, Long Short-term Memory, a variation of RNN, has become a more effective model and therefore been used in relevant problems [3, 4, 5].

Generally the existing models give predictions based more on individuals' static states rather than interactions between them. Such predictions can be lack of the abilities to infer on the situations involving numbers of pedestrians walking in the same scene. The idea of social pooling was previously proposed to aggregate and encode individuals' relative positions based on their local areas [4]. As a step forward, the global pooling layer was designed to share information between each LSTM model [5].

Generative models different from predictive models can make inference less on probabilistic likelihood within series but based on more high-level features. Generative Adversarial Networks (GANs) have achieved state-of-the-art results on reproduction of realistic photos and images [6]. The combination of RNN and GAN to generate trajectory forecasting was proposed with global social pooling layer [5]. This model comprises a LSTM Encoder-Decoder generator and a LSTM discriminator. To encourage the diversity of forecasting, L2 loss is used to explore potentially better trajectories.

One of the limitations of SGAN is that a perfect selection of observation trajectories should be somehow obtained to accurately predict the following steps. However, in realistic situations, pedestrians may constantly change their paces and distracted from their previous track shortly before coming back to the lane. Also, the model may need a longer observation time to decide the potential destination of each pedestrian, which cannot been effectively solved by RNNs. Previous works of SGAN set an arbitrary skip parameter to equally skip some time steps to make more sense of the sequences.

In this work, the attention mechanism is applied to the existing SGAN model to jointly align and find the most relevant history time steps and pay more attention on them. In this way, we would be able to conduct more experiments showing the effect of a longer observation on the predictions. The attentive SGAN consistently shows superior performances on some datasets and for longer observation lengths.

2 Related work

Human-space interactions and human-human interactions are two aspects of trajectory forecasting. To infer individual's motion pattern according to history is a sequence prediction problem commonly modelled using RNN; to model instant interactions between different people within a territory requires additional social pooling to encode relative position information. In addition, sequential datasets normally based on the metric of time elapse. However, human inference sometimes relies more on logical links between significant time relations, thus requiring attention mechanism. This section will discuss some relevant works of GAN models and RNN with attention.

Sequence Prediction Model. Recurrent Neural Networks (RNNs) are a category of data-driven sequence prediction models. As a variation of simple feedforward networks as MLP, additional propagation is added between cells as well as layers, which enables the model to infer not only the relationship of inputs and outputs but also past and future. Such techniques have been widely used in NLP problems [7, 8, 9]. Long Short-term Memory (LSTM) [10] is the variation of RNN models with additional gates mechanism to solving the long-existing gradient vanishing problems.

Generative Modelling. Generative Adversarial Network (GAN) [11] is a ground-breaking innovation providing a generative solution to prediction problems. Instead of modelling the problems as a whole, GAN devices a minimax like game to train two separate generator and discriminator networks. Such design extends the traditional well-defined likelihood function to a trainable parameterised model and tackles some intractable probabilistic computation and behavioural inference. Due to the difficulties of solving a minimax game, several works added some tricks to the vanilla model and achieved better equilibrium [12, 13].

Attention Mechanism. Attention mechanism was initially discovered in NLP, in which the reasoning of sentences is often more relevant to some words with substantial meanings [14, 15]. Also, in image recognition the attention is different in two dimensions [16]. Attention mechanism can help with long term prediction and infer the importance of time steps.

3 Method

We proposed a model similar to the vanilla SGAN only with attention layers before spatial embedding of the trajectories. L2 loss is added to explore diverse predictions. The objective is to observe for several time steps $t = 1, 2, \ldots, t_{obs}$ of multiple agents' trajectories $X_i = (x_i^t, y_i^t)$ and generate the prediction of their incoming steps $t = 1, 2, \ldots, t_{pred}$ of trajectories $Y_i = (x_i^t, y_i^t)$. We denote our predictions as $\hat{Y}_i$.

3.1 Generative Adversarial Networks

GAN consists of two networks: a discriminative model D which should tell whether the incoming samples are falsely generated or ground truth, and a generative model G to cheat the discriminator. The training process is like a minimax game and the best results are obtained when the equilibrium is achieved. The objective function of this game is presented as:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z)))] \tag{1}$$

where the inputs of the generator is some low-level latent variable z and the inputs of the discriminator is high-level sample feature variable x.

3.2 Attentive GAN

The attentive SGAN consists of the encoder and decoder using one-layer LSTM architecture with joint attention before the embedding layers. Both the encoder and decoder are used in the generative model with social pooling module added in the latent space. Only the encoder is used in the discriminator and output its judgement on how likely the sample is false. The model weights are shared between each agent with only their trajectories differing.

Generator. Firstly, we used a single layer MLP with softmax to get the attention weights of observation time steps for each dimension, and then applied the weights on the inputs before conducting positional embedding.

$$\begin{gathered}
\alpha_{x_i^t} = a(x_i^t, W_{attn}^x),\ \alpha_{y_i^t} = a(y_i^t, W_{attn}^y) \\
w_{x_i^t} = \frac{\exp(\alpha_{x_i^t})}{\sum_k \exp(\alpha_{x_i^k})},\ w_{y_i^t} = \frac{\exp(\alpha_{y_i^t})}{\sum_k \exp(\alpha_{y_i^k})} \\
\tilde{X} = (\tilde{x}_i^t, \tilde{y}_i^t) = (w_{x_i^t} x_i^t, w_{y_i^t} y_i^t)
\end{gathered} \tag{2}$$

These embeddings are used be encoded into the latent space through the recurrent LSTM model:

$$\begin{aligned}
e_i^t &= \phi(\tilde{x}_i^t, \tilde{y}_i^t; W_{ee}) \\
h_{ei}^t &= LSTM(h_{ei}^{t-1}, e_i^t; W_{encoder})
\end{aligned} \tag{3}$$

Although LSTM model can infer between multiple sequences on the parameter level, some high-level interactive features normally cannot be extracted automatically and it can be data inefficient. Therefore, the global pooling module proposed in SGAN [5] is used to encode such relative information in a scene. After observation of the history time steps, the hidden states of each people h_{ei}^t and the pooled tensor P_i are decoded using another LSTM network in a recurrent style:

$$\begin{aligned}
e_i^t &= \phi(x_i^{t-1}, y_i^{t-1}; W_{ed}) \\
P_i &= PM(h_{d1}^{t-1}, \ldots, h_{dn}^{t-1}) \\
h_{di}^t &= LSTM(\gamma(P_i, h_{di}^{t-1}), e_i^t; W_{decoder})
\end{aligned} \tag{4}$$

Discriminator. The discriminator uses a separate LSTM model taking real trajectories $T_{real} = [X_i, Y_i]$ and fake trajectories $T_{fake} = [X_i, \hat{Y}_i]$ as inputs and their classification scores as outputs.

Losses. To improve the training process of GAN, we use the proposed tricks [13] to add noise on ground truth:

$$\begin{aligned}
y_{true} &= 1 + \epsilon_{true} \\
y_{fake} &= 0 + \epsilon_{fake}
\end{aligned} \tag{5}$$

4 Experiments

We conducted the experiments based on two public datasets: ETH and UCY, which demonstrate human trajectories on several social scenarios. These complicated scenarios involving multiple people (basically up to 1000 and more) including people crossing each other, colliding while avoiding to collide, and other group behaviors like gathering and dispersal.

Evaluation Metrics: The error metrics employed in this experiment basically inherits from that of Social GAN:

- Average Displacement Error (ADE): Average L2 distance between ground truth and our prediction over all predicted time steps.
- Final Displacement Error (FDE): The distance between the predicted final destination and the truth final destination at the end of prediction period.

Baselines: We compare against the following baselines:

- Primitive Social GAN with observation length 8;

Evaluation Methodology: We use leave-one-out approach, train on 4 sets and test on the remaining set. We observe the trajectories for several time steps and show prediction results for 8 and 12 time steps with 0.4 second per step.

4.1 Quantitative Evaluation

In this experiment, we introduce a new model SAGAN which combines attention mechanism and previous social GAN. We compare our method on two metrics ADE and FDE against baselines. It is observed that as observation length grows, both Social GAN and SAGAN witness decrease, later come to stable.

One thing to note that, given a relative small observation, SAGAN model outperformed the primitive social GAN, yet the effect of observation length growing would gradually eliminate such difference. We pick the FDE of SGAN and SAGAN under prediction period of 8 length on dataset ZARA1. The experiments results are attached below1.

		12		16		20		24	
Metric	**Datasets**	**I**	**II**	**I**	**II**	**I**	**II**	**I**	**II**
ADE	**eth**	0.53/0.55	**0.51/0.55**	0.44/0.58	**0.41/0.54**	**0.32**/0.39	0.33/**0.35**	0.25/0.4	**0.25/0.39**
	univ	**0.40**/0.69	0.43/**0.66**	**0.41**/0.69	0.42/**0.61**	0.43/0.76	**0.42/0.73**	0.41/0.73	**0.40/0.72**
	zara1	0.22/**0.34**	**0.21**/0.35	0.21/0.34	**0.21/0.34**	0.21/0.38	**0.21/0.33**	0.22/0.36	**0.21/0.33**
AVG		0.38/0.53	**0.38/0.52**	0.35/0.54	**0.35/0.50**	0.32/0.51	**0.32/0.47**	0.29/0.50	**0.29/0.48**
FDE	**eth**	1.02/**0.87**	**0.97**/0.89	0.79/0.96	**0.72/0.93**	0.53/**0.63**	**0.53**/0.57	**0.39**/0.65	0.41/**0.62**
	univ	**0.81**/1.36	0.84/**1.31**	**0.79**/1.34	0.80/**1.20**	0.81/1.47	**0.78/1.40**	0.79/1.40	**0.77/1.39**
	zara1	0.42/**0.68**	**0.40**/0.69	0.41/**0.67**	**0.40**/0.69	**0.41**/0.73	0.42/**0.68**	0.41/0.70	**0.41/0.63**
AVG		0.75/0.97	**0.74/0.96**	0.66/0.99	**0.64/0.94**	0.58/0.94	**0.58/0.88**	0.53/0.92	**0.53/0.88**

Table 1: We can observe from the table that Social GAN dominates its former counterpart on most scenarios according to our two metrics (ADE and FDE). And this yet reveals its socially favorable properties.

4.2 Introducing the attention mechanism

We employ attention mechanism to boost the performance of Social GAN in this experiment. And the results are in accordance with our expectations. A brief comparison between the ADE of SAGAN (one implemented with attention) and previous social GAN of prediction length being 12 time steps, which was abstracted from the above table is as attached1.

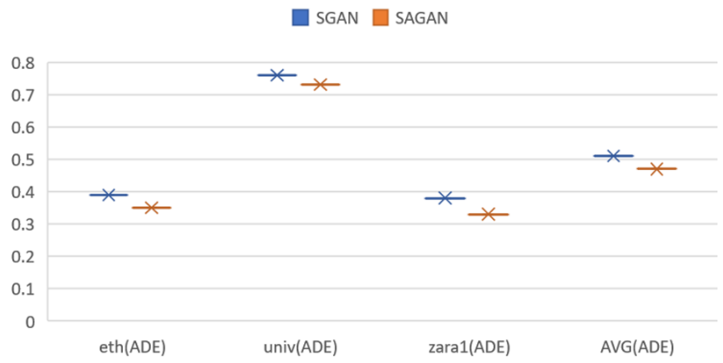

Figure 1: Comparison of SGAN against SAGAN (metrics: ADE) on several datasets

It is intuitive to explain the outperformance of SAGAN. Attention mechanism, in itself, is built for enhancing encoder-decoder model based on RNN. It allows for subtracting specific characteristic from data and thus makes it more flexible for model to conduct learning process.

We would also notice that SAGAN does not possess an overwhelming advantage over SGAN at a lower observation length. It may have something to do with the inner architecture of the model. With less observation time, our attention mechanism could lack sufficient time and information to perform subtraction and classification. So it is advised to set a higher observation time for SAGAN model.

4.3 Observation Time

During this experiment, we increase the observation length for certain model (basically in a linear growth). Five results have been posted above. It can be observed from the graph that with observation length ascending, FDE of SGAN witnessed a steady decrease and that of SAGAN descend rapidly at first try and then have a small oscillation. In general, ascend of observation length contributes to better performance of model2.

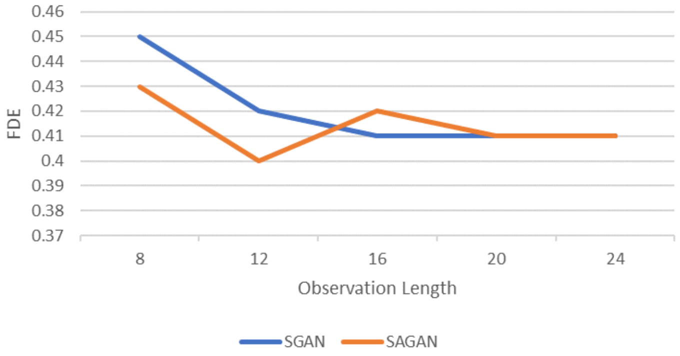

Figure 2: Effect of observation length. We train SGAN and SAGAN separately, computing ADE and FDE across our adjustment to observation length. Enlarging observation intervals contributes to the better performance for prediction.

Such behaviour is also intuitive for daily life. More observation time allows more information to learn from, thus a more sophisticated model can be generated to grab the trajectories. Also notice that too much observation length can lead to overfitting. And issues regarding efficiency would also arise.

Due to the time and technique restriction, we only conducted 4 comparing experiment with increment step of 4. Which is not sufficient for further analysis. The follow-up experiment would be extending the interval radius of sample to observe more behaviour of the two models reaction, and narrowing the gap of each observation-length we choose to get a more precise answer.

References

[1] Federico Bartoli, Giuseppe Lisanti, Lamberto Ballan, and Alberto Del Bimbo. Context-aware trajectory prediction. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1941–1946. IEEE, 2018.

[2] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 336–345, 2017.

[3] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. *Neural networks*, 108:466–478, 2018.

[4] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

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

[6] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651, 2017.

[7] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. End-to-end continuous speech recognition using attention-based recurrent nn: First results. *arXiv preprint arXiv:1412.1602*, 2014.

[8] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*, pages 2980–2988, 2015.

[9] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772, 2014.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[12] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[13] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[15] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[16] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.