Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction

Abduallah Mohamed[1], Kun Qian[1]
Mohamed Elhoseiny[2,3, **], Christian Claudel[1, **]
[1]The University of Texas at Austin [2]KAUST [3]Stanford University
{abduallah.mohamed,kunqian,christian.claudel}@utexas.edu, mohamed.elhoseiny@kaust.edu.sa

Abstract

Better machine understanding of pedestrian behaviors enables faster progress in modeling interactions between agents such as autonomous vehicles and humans. Pedestrian trajectories are not only influenced by the pedestrian itself but also by interaction with surrounding objects. Previous methods modeled these interactions by using a variety of aggregation methods that integrate different learned pedestrians states. We propose the Social Spatio-Temporal Graph Convolutional Neural Network (Social-STGCNN), which substitutes the need of aggregation methods by modeling the interactions as a graph. Our results show an improvement over the state of art by 20% on the Final Displacement Error (FDE) and an improvement on the Average Displacement Error (ADE) with 8.5 times less parameters and up to 48 times faster inference speed than previously reported methods. In addition, our model is data efficient, and exceeds previous state of the art on the ADE metric with only 20% of the training data. We propose a kernel function to embed the social interactions between pedestrians within the adjacency matrix. Through qualitative analysis, we show that our model inherited social behaviors that can be expected between pedestrians trajectories. Code is available at `https://github.com/abduallahmohamed/Social-STGCNN`.

1. Introduction

Predicting pedestrian trajectories is of major importance for several applications including autonomous driving and surveillance systems. According to [9], 70% of pedestrians tend to walk in groups. The complexity of pedestrian trajectory prediction comes from different social behaviors such as walking in parallel with others, within a group, collision avoidance and merging from different directions.

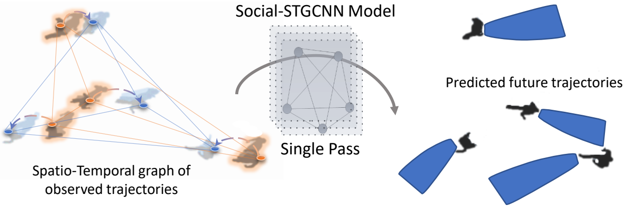
Social-STGCNN Model
Predicted future trajectories
Single Pass
Spatio-Temporal graph of observed trajectories

Figure 1. Pedestrian future trajectories prediction using the Social-STGCNN model. The social interactions between pedestrians and their temporal dynamics are represented by a spatio-temporal graph. We predict the future trajectories in a single pass.

The social attributes of pedestrian motions encouraged researchers in this area to focus on inventing deep methods to model social interactions between pedestrians. In the Social-LSTM [1] article, deep learning based model is applied to predict the pedestrians trajectories by modeling each pedestrian trajectory via a recurrent deep model. The outputs of recurrent models are made to interact with each other via a pooling layer. Several articles [8, 7, 13] followed this direction. Social-LSTM [1] modeled the pedestrian trajectories as a bi-variate Gaussian distribution, while some of others aimed at predicting deterministic trajectories. Another direction is to use Generative Adversarial Networks (GANs) for this task, assuming that the distribution of trajectories is multi-modal. Several articles [3, 11, 6] used GANs to predict distributions of future trajectories. For these models, generators are designed using recurrent neural networks, and again, aggregation methods are relied upon to extract the social interactions between pedestrians. We argue that a limitation of earlier articles comes from the use of recurrent architectures, which are parameter inefficient and expensive in training [2]. We overcome this limitation through the use of temporal convolutional architectures.

In addition to the limitation of recurrent architectures, aggregation layers used in earlier works can also limit their performance. The aggregation layer takes the hidden states of the recurrent units as inputs. It is expected to assimilate a global representation of the scene, since each recurrent unit models a pedestrian trajectory. However, there are two issues within this type of aggregation. First, the aggregation in

[**] Equal advising.

feature states is neither intuitive nor direct in modelling interactions between people, as the physical meaning of feature states is difficult to interpret. Second, since the aggregation mechanisms are usually based on heuristics like pooling, they could fail in modeling interactions between pedestrians correctly.

We designed Social-STGCNN to overcome the two aforementioned limitations. First, we model the pedestrians trajectories from the start as a spatio-temporal graph to replace the aggregation layers. The graph edges model the social interactions between the pedestrians. We propose a weighted adjacency matrix in which the kernel function quantitatively measure the influence between pedestrians. To address issues associated with recurrent units, our model manipulates over the spatio-temporal graph using a graph Convolutional Neural Networks (CNN)s and a temporal CNNs. This allows our model to predict the whole sequence in a single shot. Due to the above design, our model outperforms previous models in terms of prediction accuracy, parameters size, inference speed and data efficiency.

2. Recent Developments

Graph CNNs [4] is similar to CNNs but the convolution operation is taken over the adjacency matrix of the graphs. ST-GCNN [12] is a spatio-temporal Graph CNN that was originally designed to solve skeleton-based action classification problem. In our work, ST-GCNNs create a suitable graph embedding. Our model operates on this embedding.

Starting from [2], the argument between the usage of Recurrent Neural Networks (RNN)s versus the usage of temporal CNNs for sequential data modeling is highlighted. We were inspired by TCNs [2] and designed a temporal CNN model that extends the capabilities of ST-GCNNs.

3. Problem Formulation

Given a set of N pedestrians in a scene with their corresponding observed positions $tr_o^n, n \in \{1, \ldots, N\}$ over a time period T_o, we need to predict the upcoming trajectories tr_p^n over a future time horizon T_p. For a pedestrian n, we write the corresponding trajectory to be predicted as $tr_p^n = \{\, \mathbf{p}_t^n = (\mathbf{x}_t^n, \mathbf{y}_t^n) \,|\, t \in \{1, \ldots, T_p\}\}$, where $(\mathbf{x}_t^n, \mathbf{y}_t^n)$ are random variables describing the probability distribution of the location of pedestrian n at time t, in the 2D space. We make the assumption that $(\mathbf{x}_t^n, \mathbf{y}_t^n)$ follows bi-variate Gaussian distribution such that $\mathbf{p}_t^n \sim \mathcal{N}(\mu_t^n, \sigma_t^n, \rho_t^n)$. Besides, we denote the predicted trajectory as $\hat{\mathbf{p}}_t^n$ which follows the estimated bi-variate distribution $\mathcal{N}(\hat{\mu}_t^n, \hat{\sigma}_t^n, \hat{\rho}_t^n)$. Our model is trained to minimize the negative log-likelihood, which defined as $L^n(\mathbf{W}) = -\sum_{t=1}^{T_p} \log(\mathbb{P}((\mathbf{p}_t^n|\hat{\mu}_t^n, \hat{\sigma}_t^n, \hat{\rho}_t^n))$ in which $\mathbf{W}$ includes all the trainable parameters of the model, μ_t^n is the mean of the distribution,σ_t^n is the variances and ρ_t^n is the correlation.

4. The Social-STGCNN Model

4.1. Model Description

The Social-STGCNN model consists of two main parts: the Spatio-Temporal Graph Convolution Neural Network (ST-GCNN) and the Time-Extrapolator Convolution Neural Network (TXP-CNN). The ST-GCNN conducts spatio-temporal convolution operations on the graph representation of pedestrian trajectories to extract features. TXP-CNN takes these features as inputs and predicts the future trajectories of all pedestrians as a whole. Figure 2 illustrates the overview of the model.

Graph Representation of Pedestrian Trajectories We start by constructing a set of spatial graphs G_t representing the relative locations of pedestrians in a scene at each time step t. G_t is defined as $G_t = (V_t, E_t)$, where $V_t = \{v_t^i \mid \forall i \in \{1, \ldots, N\}\}$ is the set of vertices of the graph G_t. The observed location (x_t^i, y_t^i) is the attribute of v_t^i. E_t is the set of edges within graph G_t which is expressed as $E_t = \{e_t^{ij} \mid \forall i, j \in \{1, \ldots, N\}\}$. $e_t^{ij} = 1$ if v_t^i and v_t^j are connected, $e_t^{ij} = 0$ otherwise. In order to model how strongly two nodes could influence with each other, we attach a value a_t^{ij}, which is computed by some kernel function for each e_t^{ij}. a_t^{ij}s are organized into the weighted adjacency matrix A_t. We introduce $a_{sim,t}^{ij}$ as a kernel function to be used within the adjacency matrix A_t. $a_{sim,t}^{ij}$ is defined as

$$a_{sim,t}^{ij} = \begin{cases} 1/\|v_t^i - v_t^j\|_2 & , \|v_t^i - v_t^j\|_2 \neq 0 \\ 0 & \text{, Otherwise.} \end{cases} \tag{1}$$

Spatio-Temporal Graph Convolution Neural Network (ST-GCNNs) ST-GCNNs extends spatial graph convolution to spatio-temporal graph convolution by defining a new graph G whose attributes are the set of the attributes of G_t. G incorporates the spatio-temporal information of pedestrian trajectories. We denote the embedding resulting from ST-GCNN as $\bar{V}$.

Time-Extrapolator Convolution Neural Network (TXP-CNN) The functionality of ST-GCNN is to extract spatio-temporal node embedding from the input graph. However, our objective is to predict further steps in the future. We also aim to be a stateless system and here where the TXP-CNN comes to play. TXP-CNN operates directly on the temporal dimension of the graph embedding $\bar{V}$ and expands it as a necessity for prediction.

5. Datasets and Evaluation Metrics

The model is trained on two human trajectory prediction datasets: ETH [10] and UCY [5]. ETH contains two scenes named ETH and HOTEL, while UCY contains three scenes named ZARA1, ZARA2 and UNIV. Our method of training follows the same strategy as Social-LSTM [1]. Two metrics are used to evaluate model performance: the Average

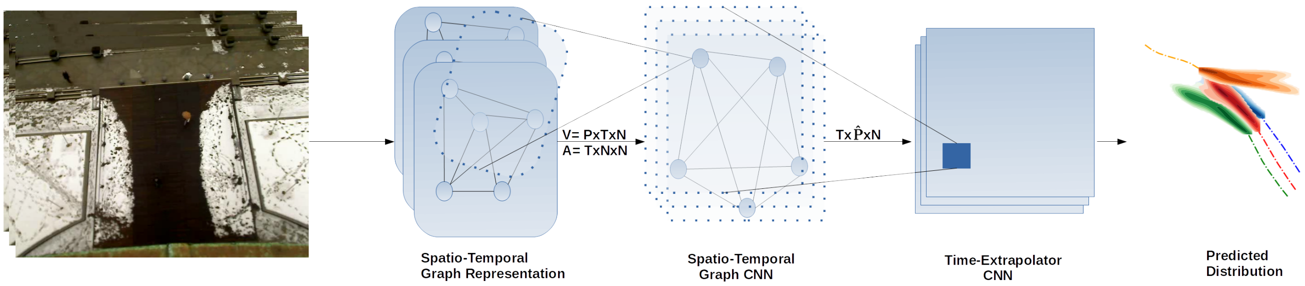

Figure 2. The Social-STGCNN Model. Given T frames, we construct the spatio-temporal graph representing $G = (V, A)$. Then G is forwarded through the Spatio-Temporal Graph Convolution Neural Networks (ST-GCNNs) creating a spatio-temporal embedding. Following this, the TXP-CNNs predicts future trajectories. P is the dimension of pedestrian position, N is the number of pedestrians, T is the number of time steps and $\hat{P}$ is the dimensions of the embedding coming from ST-GCNN.

Displacement Error (ADE) [10] and the Final Displacement Error (FDE) [1].

6. Experiments and Results Analysis

6.1. Quantitative Analysis

The performance of Social-STGCNN is compared with other models on ADE/FDE metrics in table 1. Overall, Social-STGCNN outperforms all previous methods on the two metrics. The previous state of art on the FDE metric is SR-LSTM [13] with an error of 0.94. Our model has an error of 0.75 on the FDE metric which is about 20% less than the state of the art.

Inference speed and model size As shown in table 2 in The size of Social-STGCNN is 7.6K parameters only which is about one sixth of the number of parameters in S-GAN-P. The inference time of our model is 0.002 seconds per inference step which is about 48 × faster than S-GAN-P.

Data Efficiency In this section, we evaluate if the efficiency in model size leads to a better efficiency in learning from fewer samples of the data. We ran a series of experiments where 5%, 10%, 20% and 50% of the training data. Social-GAN is employed as a comparison baseline. Figure 4 shows the data learning efficiency experiments. We notice that our model exceeds the state of the art on the FDE metric when only 20% of training data is used. Also, Social-STGCNN exceeds the performance of Social-GAN on the ADE metric when trained only on with 20% of the training data.

6.2. Qualitative Analysis

The quantitative analysis section shows that Social-STGCNN outperforms previous state-of-art in terms of ADE/FDE metrics. We qualitatively analyze how Social-STGCNN captures the social interactions between pedestrians and takes that into consideration when predicting the distributions as in figure 3.

7. Conclusion

In this article, we showed that a proper graph-based spatio-temporal setup for pedestrian trajectory prediction improves over previous methods on several key aspects, including prediction error, computational time and number of parameters. By applying a specific kernel function in the weighted adjacency matrix together with our model design, Social-STGCNN outperforms state-of-art models over a number of publicly available datasets. We also showed that our configuration results in a data-efficient model and can learn from few data samples.

Acknowledgement This research is supported by NSF (National Science Foundation) CPS No.1739964, CIS No.1636154 and CIS No.1917056.

References

[1] 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.

[2] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[3] 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.

[4] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[5] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007.

[6] Jiachen Li, Hengbo Ma, and Masayoshi Tomizuka. Conditional generative neural system for probabilistic trajectory prediction. *arXiv preprint arXiv:1905.01631*, 2019.

	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
S-GAN-P [3]	0.87 / 1.62	0.67 / 1.37	0.76 / 1.52	0.35 / 0.68	0.42 / 0.84	0.61 / 1.21
CGNS [6]	**0.62** / 1.40	0.70 / 0.93	0.48 / 1.22	0.32 / 0.59	0.35 / 0.71	0.49 / 0.97
PIF [7]	0.73 / 1.65	**0.30 / 0.59**	0.60 / 1.27	0.38 / 0.81	0.31 / 0.68	0.46 / 1.00
Social-STGCNN	0.64 / **1.11**	0.49 / 0.85	**0.44 / 0.79**	0.34 / **0.53**	0.30 / **0.48**	**0.44 / 0.75**

Table 1. ADE / FDE metrics for several methods compared to Social-STGCNN are shown. Models evaluated using the best sample out of 20 samples. All models takes as an input 8 frames and predicts the next 12 frames. We notice that Social-STGCNN have the best average error on both ADE and FDE metrics. The lower the better.

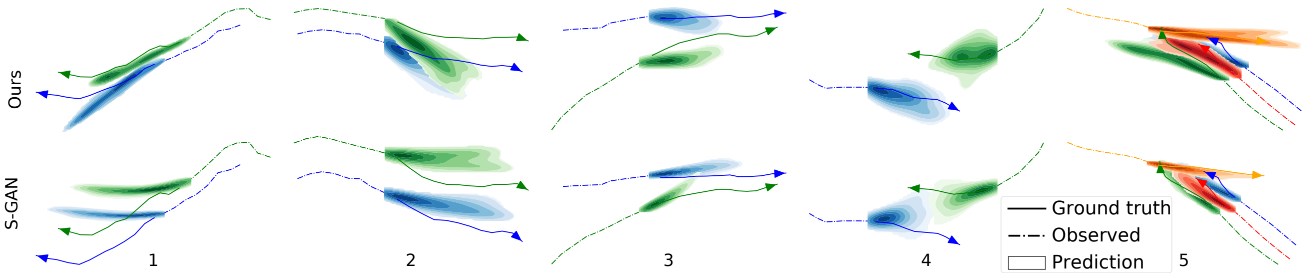

Figure 3. Qualitative analysis of Social-STGCNN . Social-GAN [3] is taken as a baseline for the comparison.A variety of scenarios are shown: two individuals walking in parallel (1)(2), two persons meeting from the same direction (3), two persons meeting from different directions (4) and one individual meeting another group of pedestrians from an angle (5).

	Parameters count	Inference time
S-LSTM [1]	264K (35x)	1.1789 (589x)
S-GAN-P [3]	46.3K (6.1x)	0.0968 (48.4x)
PIF [7]	360.3K (47x)	0.1145 (57.3x)
Social-STGCNN	**7.6K**	**0.0020**

Table 2. Parameters size and inference time of different models compared to ours bench-marked using Nvidia GTX1080Ti GPU. The inference time is the average of several single inference steps. We notice that Social-STGCNN has the least parameters size and the least inference time compared to others. The text in blue show how many times our model is faster than others.

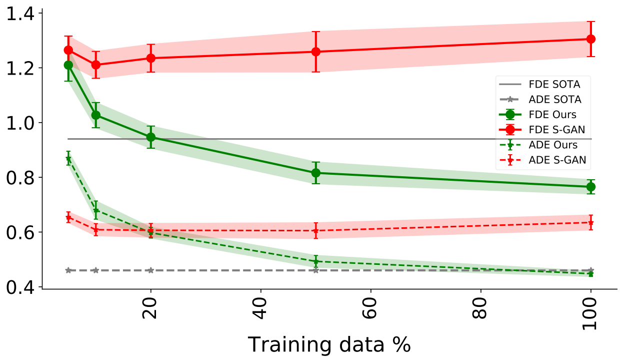

Figure 4. Model performance versus shrinked training dataset. The x-axis shows several randomly samples shrink percentages. The shade represents errors. The same shrinked data were used across the models. The figure shows our performance versus Social-GAN which is the closest model in terms of parameter size to ours.

[7] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5725–5734, 2019.

[8] Huynh Manh and Gita Alaghband. Scene-lstm: A model for human trajectory prediction. *arXiv preprint arXiv:1808.04018*, 2018.

[9] Mehdi Moussaïd, Niriaska Perozo, Simon Garnier, Dirk Helbing, and Guy Theraulaz. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. *PloS one*, 5(4):e10047, 2010.

[10] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268. IEEE, 2009.

[11] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2019.

[12] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[13] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12085–12094, 2019.