Multimodal Interaction-Aware Trajectory Prediction in Crowded Space

Xiaodan Shi,[1] Xiaowei Shao,[1,2] Zipei Fan,[1] Renhe Jiang,[1,3] Haoran Zhang,[1]
Zhiling Guo,[1] Guangming Wu,[1] Wei Yuan,[1] Ryosuke Shibasaki[1]
[1]Center for Spatial Information Science, the University of Tokyo
[2]Earth Observation Data Integration and Fusion Research Initiative, the University of Tokyo
[3]Information Technology Center, the University of Tokyo
{shixiaodan, jiangrh, zhang_ronan, guozhilingcc, huster-wgm, shiba}@csis.u-tokyo.ac.jp,
{shaoxw, fanzipei, miloyw}@iis.u-tokyo.ac.jp

Abstract

Accurate human path forecasting in complex and crowded scenarios is critical for collision avoidance of autonomous driving and social robots navigation. It still remains as a challenging problem because of dynamic human interaction and intrinsic multimodality of human motion. Given the observation, there is a rich set of plausible ways for an agent to walk through the circumstance. To address those issues, we propose a spatio-temporal model that can aggregate the information from socially interacting agents and capture the multimodality of the motion patterns. We use mixture density functions to describe the human path and predict the distribution of future paths with explicit density. To integrate more factors to model interacting people, we further introduce a coordinate transformation to represent the relative motion between people. Extensive experiments over several trajectory prediction benchmarks demonstrate that our method is able to forecast various plausible futures in complex scenarios and achieves state-of-the-art performance.

Introduction

Forecasting the future trajectories of dynamic pedestrians through crowded scenarios is highly valuable for autonomous driving and social robots navigation (Kitani et al. 2012; Karasev et al. 2016; Liu et al. 2016; Lee et al. 2017; Su et al. 2017). This prediction problem is about generating a sequence of future locations based on observations of past trajectories of certain length. Pedestrian trajectory prediction has benefited from the introduction of Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber 1997), by which long-term time dependencies can be captured, which has renewed interest in trajectory prediction in recent years. Although there are many promising publications on the subject, problems related to trajectory prediction are far from being solved due to the complexities and uncertainties of human crowds.

When human navigate though crowded scenarios, they negotiate all interactions with others in their own style and yield right of way towards the destination. Trajectory prediction has never been an easy task due to the properties of human motion:

Figure 1: Our goal is to forecast human paths in complex and crowded scenarios. It is not suitable to predict a single path in complex scenes. To this end, we jointly model past trajectories and spatially interacting people, and map the distribution of possible futures which can generate multiple plausible trajectories.

1. Personal planning. People always plan the route with a goal in mind. In most cases, people walk smoothly, and we can gain insights into their possible destinations given a sequence of observations of their past trajectories. Sometimes, pedestrians suddenly change their walking directions. These cases are unpredictable based on "normal" past trajectories.

2. Dynamic social interactions. People walking through crowds obey rules of social etiquette (Robicquet et al. 2016), such as keeping a safe distance from others, which can not be quantified. The rules of social manners vary based on the number of people involved. Human can efficiently integrate all their interactions and adjust their path accordingly. But it is not easy for machines. Moreover, individual preferences also effect how people react to others.

3. Multimodality of human motion forecasting. It is not rational to forecast a single path in complex circumstances. Given the observation of past trajectories, multiple plausible future trajectories can be forecasted. But most research only have one feasible prediction result output.

The existing works are designed to address one or more of the above challenges. They focus mainly on modeling social interactions (Pellegrini et al. 2009; Yamaguchi et al. 2011; Alahi et al. 2016; Vemula, Muelling, and Oh 2018; Fernando et al. 2018b). The first attempt to model social interactions can be traced back to the handcrafted features-based traditional method (Helbing and Molnar 1995). The

method relies heavily on certain definite rules and is mainly applied to simulation tasks and not long-term real-world trajectory prediction. LSTM-based data-driven deep neural networks have shown promise with their progress in modeling the long-term time dependency of trajectories, especially Social LSTM which is a tipping point for forecasting real-world human walking paths (Alahi et al. 2016). Social LSTM introduced "Social Pooling" to jointly model human interactions and used a (unimodal) bivariate Gaussian model to represent the uncertainty of future trajectories. Recent research have noticed the intrinsic uncertainty of human motion, modeled the multimodality of trajectory prediction and generated multiple acceptable paths (Gupta et al. 2018; Amirian, Hayet, and Pettré 2019; Fernando et al. 2018a; Sadeghian et al. 2019). The reserachers proposed LSTM-based Generative Adversarial Network (GAN) models (LSTM-based encoder-decoder generators and LSTM-based discriminators) to generate a set of future plausible paths through sampling.

Previous research have made great progress and given us more insights into the task of trajectory prediction. Nevertheless, there are still limitations. First, the relative motion among agents can be utilized to model interactivity (Shu et al. 2018; Pellegrini et al. 2009), but the existing data-driven method only considers relative positions not taking into account velocity and walking direction, which can also have an effect (Zhang et al. 2019). Second, GAN generates multiple acceptable trajectories by sampling. It can not predict the distribution of future paths with explicit densities and is unstable to train. Third, commonly used loss function measuring the distance between ground truth and prediction results lead to the model learning "average behaviors".

To address the above limitations, we proposed an LSTM-based spatio-temporal model which jointly models the potential multimodality of human motion and dynamic human interactions (Fig. 1). Unlike the GAN-based trajectory sampler, our model can generate the distribution of future trajectories with explicit densities. To summarize the contributions:

1. Our model uses a spatio-temporal graph structure to naturally incorporate spatial social awareness and temporal transitions of agents.

2. We use mixture density functions to describe trajectories and forecast the distribution of future plausible paths. LSTM is connected to the Mixture Density Network (MDN)(Bishop 1994) which outputs a set of Gaussian models to express the distribution of future path.

3. We utilize coordinate transformation to simplify the representation of relative motion among agents. This representation can reflect relative position, velocity and walking direction.

4. We test the model using classic trajectory prediction benchmarks and the experiments show promising results.

Related Works

Our main task of interest is pedestrian trajectory prediction using recurrent neural networks (RNNs), in particular, LSTM based architectures. Thus, in this section, we will briefly review RNNs for trajectory prediction as well as human interaction and multimodality.

RNNs for trajectory prediction. RNNs are a class of artificial neural networks which have a powerful capability to exhibit dynamic temporal behaviors of long sequence data and are well used in machine translation(Luong, Pham, and Manning 2015), video activity recognition(Donahue et al. 2015) and trajectory prediction(Becker et al. 2018). RNNs for trajectory prediction can be classified as non-consideration of interaction and consideration of interaction. The encoder-decoder architecture based on RNNs is usually constructed to forecast human motion where the past observations are encoded as latent representations, and the decoder interprets the representations to output future trajectories. Some methods do not take human interaction into account. They assume that walking pedestrians do not effect each other, isolating all agents, which is reasonable when the scenarios are uncrowded(Sun et al. 2018; Wang, Zhang, and Yi 2017).

Human interaction. Interactions in trajectory prediction are categorized as human-human interaction and human-space interaction. Human-human interaction focuses on the human motion patterns between dynamic agents. Human-space interaction models how agents react to the static scene. In this study, we mainly pay attention to the former (all following terms "human interaction" refers to "human-human interaction").

Pioneering work on modeling human interaction can date back to the classic Social Force Model (SFM) (Helbing and Molnar 1995; Yamaguchi et al. 2011) and the Interacting Gaussian Process (IGP) model(Trautman and Krause 2010). SFM represents a pedestrian as a particle reacting to the energy described by the interactions with other dynamic targets and static objects such as obstacles. IGP represents the trajectory of an agent as a Gaussian Process and each step of the agent is a set of Gaussian variables. IGP can represent multimodal distributions and has relatively few parameters. The major drawback of these methods is their limited capability to model complex, dynamic interaction in crowded scenarios because their performance largely depends on a set of predefined parameters, such as preferred walking speed and destination. The first RNN-based data-driven model, which can model human interaction in crowded space was proposed by Alahi et al. in 2016 called Social LSTM(Alahi et al. 2016). Social LSTM introduced a novel pooling layer that is called "social pooling" and allows the LSTM of spatially proximal sequences to share their hidden states so that it can automatically learn typical interactions that take place among trajectories that coincide in time. Subsequently, more works revisited Social LSTM to model human interaction. The attention mechanism, which plays a great role in machine translation, has been introduced to the social pooling layer to learn different weights of neighbors on the agent. Fernando et al. extended the classic model to incorporate both soft attention as well hard attention where the former is for handling longer trajectories and the latter is used for modeling interacting people (Fernando et al. 2018b). Vemula et al. introduced an attention Social LSTM for social robot navigation in crowded scenarios(Vemula, Muelling,

and Oh 2018). All the people in the scenario at any time instance are considered when calculating the influence on an agent robot.

Multimodality. Human motions under crowded scenarios imply a multiplicity of modes. Most existing works mediate multimodality to learn the average behavior and output only one feasible path. The Desire architecture handled this problem using conditional variational auto-encoder (CVAE) framework which can learn a sampling model that, given observations of past trajectories, produces a diverse set of prediction hypotheses(Lee et al. 2017). In recent years, RNN-based GAN models have been proposed to capture the multimodality of the space of plausible futures. Gupta A. et al. proposed Social GAN which contains RNN-based encoder-decoder generator and RNN-based decoder discriminator(Gupta et al. 2018). Social GAN integrates all the interactions involved in the scenarios and encourages the generative network to spread its distribution and cover the space of possible paths by introducing a variety loss. Sadeghian A. et al proposed Sophie, an attentive GAN to jointly model static human-space, and dynamic human-human interactions by blending a social attention mechanism with a physical attention that helps the model to learn where to look in a large scene and to extract the most salient parts of the image relevant to the path(Sadeghian et al. 2019). Moreover, Sophie also takes advantage of GAN to generate more realistic samples and to capture the uncertain nature of future paths.

Method

In this section, we will give an overview of the proposed model which is designed for human path forecasting in complex scenarios. There are some remarkable elements which can reflect people's capabilities to navigate in crowded and complex scenarios: ego trajectories implying preferable future walking directions or goals, spatially interacting pedestrians, intrinsic multimodality of human motions. To successfully forecast future trajectories, our model jointly takes these temporal and spatial cues into account and maps the area of plausible futures.

The architecture of our model is depicted in Fig.2, which contains an encoder and a decoder created in the weighted spatio-temporal graphs. By modeling different parts of the graphs, the model captures the patterns of human motion over space and time. The encoder encodes the observation as latent features by utilizing two LSTMs which describe the past trajectories and human interaction respectively. The latent features are then transmitted to the decoder which is one LSTM stacked with an MDN. Our decoder directly outputs the parameters of the Gaussian Mixture Models (GMMs) to describe the distribution of future trajectories. Because the encoder and decoder are recurrent, there is a direct connection between the inner representation at time t and the one at time t + 1.

Problem formulation

Trajectory prediction is viewed as a sequence generation problem, given observations of past trajectories. We assume there are N agents in our scenarios. Their past and future trajectories are represented as $X^{0:t_e} = X_1^{0:t_e}, X_2^{0:t_e}, \cdots, X_N^{0:t_e}$ and $Y^{t_e:t_d} = Y_1^{t_e:t_d}, Y_2^{t_e:t_d}, \cdots, Y_N^{t_e:t_d}$ where $0{:}t_e$ and $t_e{:}t_d$ are the timesteps of observation and prediction respectively. Given agent i, each state of the trajectory at time t is denoted with position p_i^t, offset o_i^t, and velocity v_i^t which are described in real-world two dimensional coordinates. Our goal is to learn the posterior distribution $p(Y^{t_e:t_d}|X^{0:t_e})$. To generate the distribution of future trajectories, we model the observation of human motion with f. Therefore, the distribution of is denoted as:

$$p(Y^{t_e:t_d}|X^{0:t_e}) = f(X^{0:t_e}; w^*), \tag{1}$$

where w^* are the parameters of the model we aim to learn. We denote the predicted future paths as $\hat{Y}^{t_e:t_d}$ which are generated from $p(Y^{t_e:t_d}|X^{0:t_e})$.

Spatio-temporal graph

Tasks requiring spatial and temporal reasoning are very common in robotics and computer vision. Inspired by works(Jain et al. 2016; Vemula, Muelling, and Oh 2018), we use a weighted spatio-temporal graph denoted as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{W}\}$ to describe human movement over space and time. By modeling the elements of the graphs, we can jointly model the social interaction and movement of people. At any time instance t, $\mathcal{V}$ is a set of nodes which represent the states of pedestrians, $|\mathcal{V}| = N$. $\mathcal{E}$ is a set of edges which contain spatial edges $Spatial\text{-}\mathcal{E}$ and temporal edges $Temporal\text{-}\mathcal{E}$. The former connects two different people at the same time instance and indicates their interaction, while the latter transfers the states of the same person in temporal space. $Temporal\text{-}\mathcal{E}$ transfer the graph $\mathcal{G}^t = \{\mathcal{V}^t, Spatial\text{-}\mathcal{E}^t, \mathcal{W}^t\}$ to $\mathcal{G}^{t+1} = \{\mathcal{V}^{t+1}, Spatial\text{-}\mathcal{E}^{t+1}, \mathcal{W}^{t+1}\}$. $\mathcal{W}$ are the weights of spatial edges, meaning how much the agent pays attention to the neighbors.

Each element of $\mathcal{V}$ at time instance t is an agent characterized as $\mathcal{V}_i^t = \{o_i^t, v_i^t, p_i^t\}$ where $o_i^t = (o_x, o_y)_i^t$ is the position offset between two adjacent time instance, $v_i^t = (v_x, v_y)_i^t$ is the walking speed and $p_i^t = (p_x, p_y)_i^t$ is the real-world absolute coordinate. $\{o_i^t, v_i^t\}$ is utilized to model $Temporal\text{-}\mathcal{E}$ because o_i^t as well v_i^t are characterized as a stable range of values for an average person which stabilize the learning process and improve the evaluation process. $\{v_i^t, p_i^t\}$ is used for $Spatial\text{-}\mathcal{E}$. To capture the spatial interaction properties and movement patterns, we apply two different LSTMs to model $Spatial\text{-}\mathcal{E}$ and $Temporal\text{-}\mathcal{E}$. The models are constructed for each edge and all the edges belonging to the same type ($Spatial\text{-}\mathcal{E}$ or $Temporal\text{-}\mathcal{E}$) sharing the same parameters of the models. Therefore, it is easy to handle a "person emerging" or "person vanishing" by adding or deleting edges and nodes.

Representations of agents interaction

It has been demonstrated that the interaction between two moving agents is relative rather than absolute and relies on critical low-level motion cues, namely, walking speed, motion direction, and distance(Shu et al. 2018). Moreover,

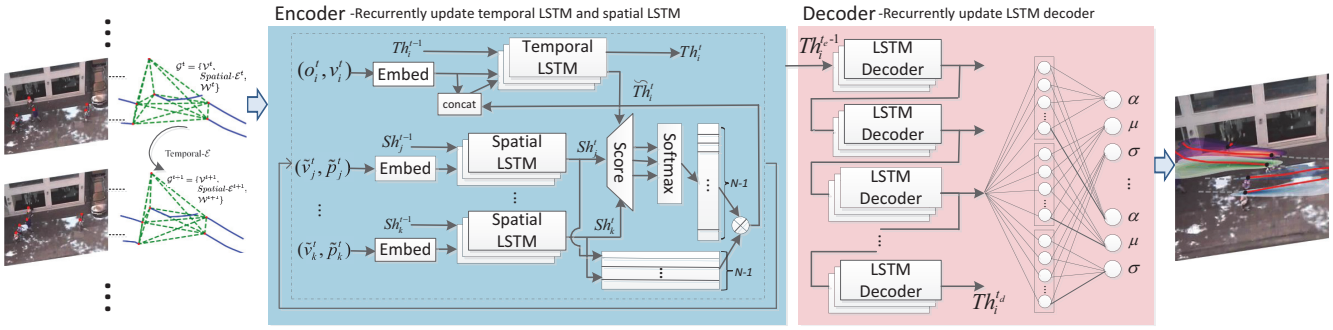


Figure 2: Overview of our model architecture. The model contains an encoder and a decoder created in the weighted spatio-temporal graphs. The encoder captures the patterns of human motion over space and time by utilizing two LSTMs which describe the past trajectories and human interaction respectively. The latent features are then transmitted to the decoder which is one LSTM stacked with an MDN. Our decoder directly outputs the parameters of GMMs to describe the distribution of future trajectories. Because the encoder and decoder are recurrent, there is a direct connection between the inner representation at time t and the one at time t + 1.

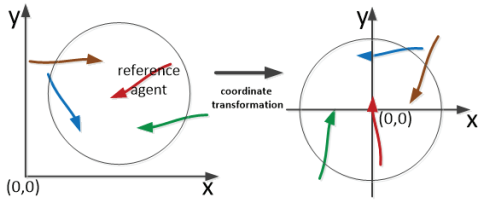


Figure 3: Relative motion among agents

agents attach more importance to the people walking in front of them through their walking direction. Most of the existing methods pool the hidden states of all interactive agents to integrate human interaction to the prediction model as Social LSTM. The hidden states are outputted by LSTM-based architectures modeled on the absolute locations of pedestrians. Here, we model human interaction more distinctly by utilizing a coordinate transformation which can simply represent the relation between agents with regard to position, velocity and walking direction (Fig. 3). At any time instance t, the reference agent i is transformed to be located at $(0, 0)$ while his/her velocity is pointed to a new y-$axis$. According to this transformation, other agents $\{(v_j^t, p_j^t)|j \in (1, 2, \cdots, N)\backslash i\}$ are re-described as $\{(\tilde{v}_j^t, \tilde{p}_j^t)|j \in (1, 2, \cdots, N)\backslash i\}$ where $\tilde{p}_j^t$ reflects the relative location and $\tilde{v}_j^t$ indicates the relative walking direction and velocity. A circle neighborhood of radius 6 m is used to decide which neighbor agents we will take into account to measure interactions.

Spatio-temporal Model

Temporal-$\mathcal{E}$: Given the reference agent $\mathcal{V}_i^t = \{o_i^t, v_i^t, p_i^t\}$, candidate hidden states $\tilde{Th}_i^t$ at time instance t can be obtained from:

$$\begin{aligned} f_i^t &= \phi_T((o_i^t, v_i^t); w_{et}^*), \\ \tilde{Th}_i^t &= LSTM_T(f_i^t, Th_i^{t-1}; w_t^*), \end{aligned} \tag{2}$$

where ϕ_T is an embedding function with ReLU activation and dropout, w_{et}^* are weights and bias of ϕ_T. $LSTM_T$ is used to model temporal edge $Temporal$-$\mathcal{E}$ with parameters w_t^*. Th_i^{t-1} are the hidden states outputted by $LSTM_T$ at time instance t-1.

Based on the candidate hidden states $\tilde{Th}_i^t$, we can construct $Spatial$-$\mathcal{E}$, which will be described in the next. The features outputted by $Spatial$-$\mathcal{E}^t$ are denoted as s_i^t. The final hidden states of temporal edges can be calculated as follows:

$$\begin{aligned} z_i^t &= \phi_z(concat(f_i^t, s_i^t); w_{ez}^*), \\ Th_i^t &= LSTM_T(z_i^t, Th_i^{t-1}; w_t^*), \end{aligned} \tag{3}$$

where ϕ_z is an embedding function similar as ϕ_T

Spatial-$\mathcal{E}$: Given agents $j \in (1, 2, \cdots, N)\backslash i$ around the reference agent i at time instance t are $(\tilde{v}_j^t, \tilde{p}_j^t)$, we firstly embed the transformed states of other agents and then feed the embedded features to LSTM of spatial edges.

$$\begin{aligned} f_j^t &= \phi_S((\tilde{v}_j^t, \tilde{p}_j^t); w_{es}^*), \\ Sh_j^t &= LSTM_S(f_j^t, Sh_j^{t-1}; w_s^*), \end{aligned} \tag{4}$$

where ϕ_S is an embedding function like ϕ_T. $LSTM_S$ is used to model spatial edges with parameters w_s^*. Sh_j^{t-1} are the hidden states outputted by $LSTM_S$ at time instance t-1.

To learn neighbor agents' effects on the reference agent, an attention mechanism(Luong, Pham, and Manning 2015) is introduced to the proposed method. In our attention module, the attention feature a_j^t is derived from the candidate hidden states $\tilde{Th}_i^t$ of $Temporal$-$\mathcal{E}$ and Sh_j^t.

$$a_j^t = \frac{exp(score(\tilde{Th}_i^t, Sh_j^t))}{\sum_{j'} exp(score(\tilde{Th}_i^t, Sh_{j'}^t))}, \tag{5}$$

where a_j^t is a tensor with size N-1 which indicates the weights of other agents, $j' = (1, 2, \cdots, N)\backslash(i)$. $score(\cdot)$ is the matrix product as follows:

$$score(\tilde{Th}_i^t, Sh_j^t) = transpose(\tilde{Th}_i^t) Sh_j^t, \tag{6}$$

Evaluation (ADE(m)/FDE(m))											
	Baselines					Our method					
Dataset	Linear	LSTM	S-LSTM	S-GAN-V1	S-GAN-V2	T-1	T-D1-1	ST-D1-1	ST-D1-20	ST-D2-20	ST-D3-20
ETH	1.65/2.84	0.71/1.40	1.09/2.35	1.08/2.13	0.72/1.29	0.60/1.34	0.57/1.27	0.53/1.09	0.37/0.72	0.33/0.60	**0.32/0.58**
HOTEL	0.99/1.70	1.15/2.09	0.79/1.76	0.77/1.69	0.48/1.01	0.38/0.85	0.47/1.14	0.33/0.72	**0.28/0.57**	0.32/0.64	0.53/1.10
UNIV	0.86/1.51	0.72/1.49	0.67/1.40	0.77/1.69	0.56/1.18	0.58/1.36	0.57/1.32	0.59/1.32	0.41/0.88	0.36/0.73	**0.35/0.68**
ZARA01	0.83/1.44	0.48/0.98	0.47/1.00	0.64/1.40	0.34/0.69	0.49/1.18	0.46/1.08	0.41/0.91	0.25/0.51	**0.23/0.43**	0.25/0.46
ZARA02	0.54/0.96	0.38/0.77	0.56/1.17	0.54/1.18	0.31/0.65	0.35/0.82	0.32/0.76	0.32/0.72	0.22/0.48	0.20/0.38	**0.19/0.35**
AVG	0.97/1.69	0.69/1.35	0.72/1.54	0.76/1.62	0.48/0.96	0.48/1.11	0.48/1.11	0.44/0.95	0.31/0.63	**0.29/0.56**	0.33/0.63

Table 1: Quantitative results of baselines vs. our method across datasets for predicting 12 future timesteps(4.8 sec) given 8 timesteps observation(3.2 sec). The results of S-LSTM, S-GAN-V1, S-GAN-V2 are from(Gupta et al. 2018). Our model consistently outperforms other baselines (lower is better).

Then, the social features s_i^t are derived by multiplying a_j^t with Sh_j^t of all the other agents. s_i^t is the weighted average tensor over all the features Sh_j^t.

$$s_i^t = \sum_j a_j^t Sh_j^t \qquad j = (1, 2, \cdots, N) \backslash i, \tag{7}$$

The social features s_i^t indicate the interaction between the reference agent i and his/her neighbor agents. By feeding s_i^t into $Temporal\text{-}\mathcal{E}$, the proposed method can integrate human interaction and agent's ego path and forecast the next state for agent.

Diverse Distribution generation by MDN

There are natural ambiguity and uncertainty in human motion. To address this, we apply MDN to provide the probability density functions (PDFs) of arbitrary complexity over the target trajectory domain which is conditioned on the observation of past paths. The MDN combines a multilayer perceptron with GMMs. Given the hidden states $Th_i^{t_e-1}$ transferred from the encoder to the decoder, the PDFs which can depict the distribution of future trajectories $\hat{y}_i^{t_e:t_d}$ conditioned on the past observation are denoted as:

$$\begin{aligned} p(\hat{y}_i^{t_e:t_d} | Th_i^{t_e-1}) &= \{p(\hat{y}_i^t | hd_i^t) | t = (t_e, t_e+1, \cdots, t_d)\}, \\ p(\hat{y}_i^t | hd_i^t) &= \textstyle\sum_{g=1}^{M} \alpha_g(hd_i^t) \phi_g(\hat{y}_i^t | hd_i^t), \end{aligned} \tag{8}$$

where M is the number of Gaussian models of our GMMs, $\alpha_g(x)$ is the prior of gth kernel, $\phi_g(\hat{y}_i^t | hd_i^t)$ is the PDF given by gth component of GMMs which is a bivariate Gaussian model, hd_i^t are the hidden states outputed by decoder LSTM at time t for agent i.

To get the PDFs over the target domain, our multilayer perceptrons take the hidden states hd_i^t as input and map hd_i^t to the control parameters of GMMs which contain priors $\alpha = \{\alpha_g \,|\, g = (1, 2, \cdots, M)\}$, means $\mu = \{\mu_g \,|\, g = (1, 2, \cdots, M)\}$, and standard deviations $\sigma = \{\sigma_g \,|\, g = (1, 2, \cdots, M)\}$. Each component of our GMMs is a bivariate Gaussian model parameterized by $\mu_g = (\mu_x, \mu_y)_g$ and $\sigma_g = (\sigma_x, \sigma_y)_g$. Priors α_g for the gth kernel satisfy $0 \leq \alpha_g \leq 1$ and $\sum_{g=1}^{M} \alpha_g = 1$. We use softmax function to obtain the priors:

$$\alpha_g = \frac{exp(z_g^\alpha)}{\sum_{j=1}^{M} exp(z_j^\alpha)}, \tag{9}$$

where $z^\alpha = \{z_g^\alpha \,|\, g = (1, 2, \cdots, M)\}$ are the latent features obtained by applying a fully connected layer fc^α to hidden states hd_i^t. The means μ and standard deviations σ are similarly calculated as:

$$\begin{aligned} \mu_g &= z_g^\mu, \\ \sigma_g &= exp(z_g^\sigma), \end{aligned} \tag{10}$$

where $z^\mu = \{z_g^\mu \,|\, g = (1, 2, \cdots, M)\}$ and $z^\sigma = \{z_g^\sigma \,|\, g = (1, 2, \cdots, M)\}$ are the latent features obtained by applying fc^μ and fc^σ to hd_i^t. $exp()$ can keep σ_g greater than zero.

Loss function

The proposed model can be trained end to end by minimizing the negative log likelihood of the ground truth of future trajectories. We backpropagate the encoder-decoder model and update the parameters until the training process becomes stable based on the following loss function.

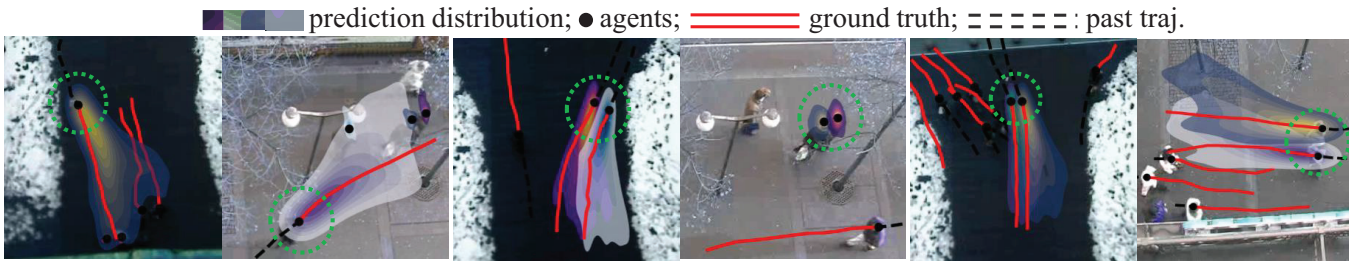
$$\mathcal{L} = -\sum_{t=t_e}^{t_d} log\Big\{ \sum_{g=1}^{M} \alpha_g(hd_i^t) \phi_g(\hat{y}_i^t | hd_i^t) \Big\}. \tag{11}$$

Implementation details

The experiments are implemented using Pytorch under Ubuntu 16.04 LTS with a GTX 1080 GPU. The size of hidden states of all LSTMs is set to 128. All the embedding layers are composed of a fully connected layer with size 128 and ReLU nonlinearty activation function. The batch size is set to 8 and all the methods are trained for 200 epochs. The optimizer RMSprop is used to train the proposed model with learning rate 0.001. We clip the gradients of LSTM with a maximum threshold of 10 to stabilize the training process. The model outputs GMMs with five components.

Experiments

In this section, the proposed model is evaluated on two publicly available datasets: UCY(Lerner, Chrysanthou, and Lischinski 2007) and ETH(Pellegrini et al. 2009). The two datasets contain 5 sets, which are UCY-zara01, UCY-zara02, UCY-univ, ETH-hotel, ETH-eth in 4 crowded scenarios with totally 1536 trajectories. The original frequency of UCY is 25fps. It is evident from the videos that the trajectories of ETH are accelerated. We treat the frequency of ETH as 15fps rather than 25fps. We firstly preprocess those two datasets by



(a) plausible directions (b) group turning/standing (c) avoid collision

Figure 4: Distribution prediction from our model in three different sets. 1st column of each set is from ETH-eth while 2nd one is from ETH-hotel.

resampling them as 2.5fps and transforming the coordinates of people to world coordinates in meters.

Evaluation approach. The proposed model is trained and tested on the two datasets with leave-one-out approach: trained on four sets and tested on the remaining set. We observe the trajectory for 8 timesteps (3.2 sec) and show prediction results for 12 timesteps (4.8 sec). To evaluate the performance, we compare our method with other state-of-the-art models on two generally used matrices as shown below.

Baselines. The proposed method is compared with the following baselines:

1. Linear method. The second order Kalman Filter, which is modeled based on position, velocity, acceleration, is used as the linear method.

2. LSTM. Human motion is modeled without considering human interaction. Offset is used as input similar as (Becker et al. 2018).

3. Social LSTM. It was proposed by Alahi et.al (Alahi et al. 2016). This method can jointly model pedestrians' ego trajectories and human interactions by pooling all hidden states of agents to the LSTM-based sequence model.

4. Social GAN. It was proposed by Gupta A. et al (Gupta et al. 2018). This approach captures the multimodality of the future trajectory prediction, which contains a RNN based encoder-decoder generator and RNN-based encoder discriminator. We consider two variants of Social GAN. S-GAN-V1: the results of one sample from S-GAN. S-GAN-V2: best results of sampling 20 times from S-GAN.

We also test various versions of the proposed method in the ablation settings. T-1 is the prediction model that combines the temporal part of our model with L2 loss. T-D1-1 is the model without considering human interaction, i.e. $Spatial\text{-}\mathcal{E}$ and ST-D1-1 is the full model. Both T-D1-1 and ST-D1-1 take mean values of distributions with maximum weights as results. ST-D1-20, ST-D2-20, ST-D3-20 take the best ones of 20 samples from the distributions with maximum weights, top two maximum weights and top three maximum weights respectively as results.

Evaluation metrics. Similar as (Alahi et al. 2016; Gupta et al. 2018), the prediction error metrics that were used are as follows:

1. Average displacement error (ADE): average L2 distance over all the prediction results and ground truth. ADE measures the average error of the predicted trajectory sequence.

2. Final displacement error (FDE): the distance between the prediction result and the ground truth at the final timestep. FDE measures the error "destination" of the prediction.

Quantitative Evaluation

We compare our method against other baselines on two metrics, ADE and FDE, as shown in Table 1. As expected, linear method performs worse than other methods in general because it is limited to model human interaction or multimodality of human motion. The S-LSTM only achieves similar accuracy to LSTM although it is trained on synthetic data and then finetuned on real-world data(Gupta et al. 2018). LSTM used in this paper takes offset as input. Offset of an average person is quite stable which not only makes the learning process stable but also improves the performance. S-GAN provides an improvement over the other baselines by capturing diverse possible courses of pedestrians.

Our first model T-1, which solely models trajectory sequence with relative motion(offset and velocity) using our temporal part and L2 loss, still outperforms baseline LSTM, which indicates that modeling path movement with offset and velocity can truly enhance the prediction performance. T-D1-1 outperforms T-1 over most datasets, which also demonstrates the validity of our MDN. The model T-D1-1, performs better than the first four baselines although it does not consider human interaction. The performance of T-D1-1 benefit from two items: utilize walking speed and offset to represent observed motion, mixture density functions to capture possible walking paths. ST-D1-1, a model considering human interaction, provides better accuracy than T-D1-1, which demonstrates the efficiency of modeling spatial edges. ST-D1-20, ST-D2-20, ST-D3-20 take the best results of 20 samples randomly drawing from the distributions with maximum weights, top two maximum weights and top three maximum weights respectively. ST-D2-20 and ST-D3-20 derive better results of most datasets than ST-D1-20, which indicates our method truly understand the uncertainty of human motion in crowded scenarios by capturing its multimodality.

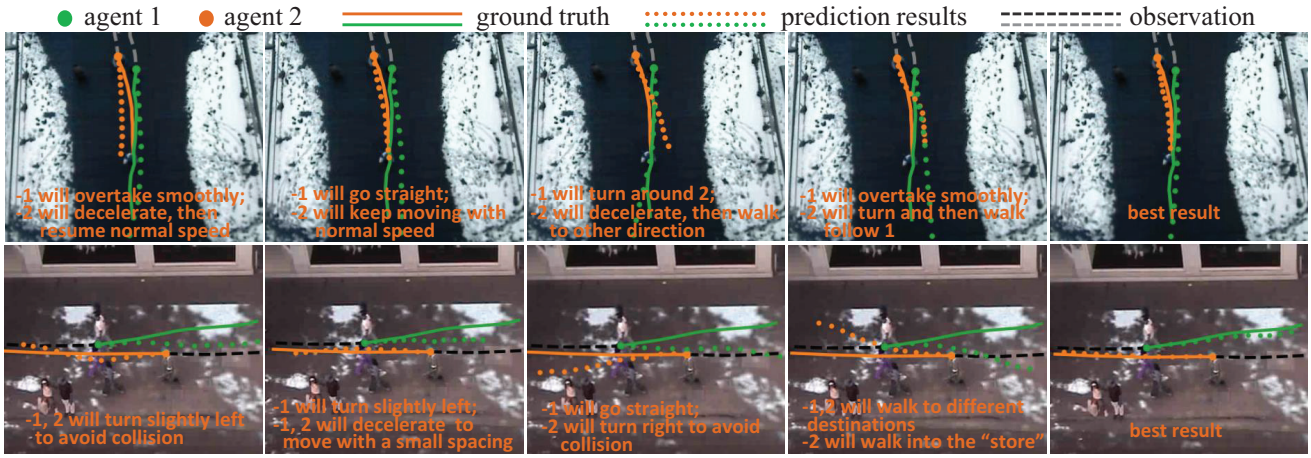


Figure 5: Multiple plausible courses for interacting people.

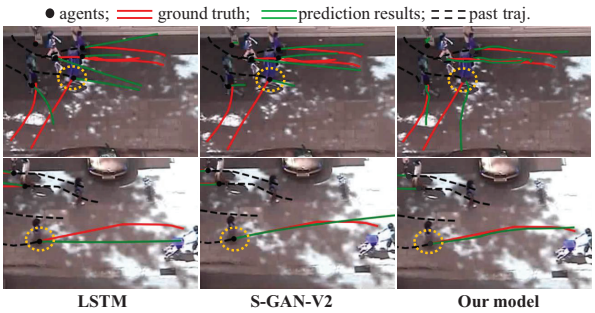


Figure 6: Comparison of predicted results from LSTM, S-GAN-V2(best result of 20 samples) and our model(best result of 20 samples).

Qualitative Evaluation

We further explore how our model performs by visualizing the distribution of future path in different complex circumstances. Fig. 4 show the visualization of results of ETH-eth and ETH-hotel datasets from three groups, plausible directions, group turning/standing, avoid collision. Agents inside the green circles are used for the qualitative evaluation. Our model forecast the space where the agents would walk into in the future. The space is socially acceptable and associated with possible walking directions, velocities. Warmer color of the distribution indicating higher possibility shows up in the first few timesteps of prediction sequence. Because given the observation, prediction near in the future is more likely to match the ground truth while longer prediction is more uncertain. Moreover, the traversable area of an agent interacting with other people or turning shows a great deal of variety, which also compliants to the intrinsic uncertainty of path forecasting.

To further demonstrate the distribution from our model is plausible and multimodal, we also consider two scenarios from ETH-eth and UCY-zara02 to show the multiple routes generated from the distribution (Fig. 5). The first row depict the process agent 1 is overtaking agent 2 and how they would behave. The second row show how people from the opposite directions interact. To walk smoothly while avoid collision, they would re-plan their routes in a "mild way" or an "intense way" by changing velocity or direction.

We also visualize the results of LSTM, S-GAN-V2, our model ST-D3-20 under the same scenarios to further investigate the performance of the proposed model as shown in Fig. 6. Agents inside the yellow circles are used for the analysis. The agent in the first row suddenly change her walking direction. From observing the past trajectory, we can find that the agent prepares for "changing direction" by turning slightly right. Our model successfully predict that the agent will walk towards another direction while the other two baselines fail to estimate the future walking direction and speed. In the second row, the agent adjusts the walking path to avoid collision with person from the opposite side. Then the agent walk back to the original walking direction. S-GAN-V2 and our model can correctly predict the agent's future behaviour and make better prediction than LSTM which does not consider human interaction and fails to predict the curve of future trajectory .

Conclusion

In this work we tackle the problem of trajectory prediction in crowded space by capturing the multimodality of motion patterns of agents interacting with other people. To express the distribution of future trajectories, our method connect Mixture Density Network into LSTM-based sequence model to output a set of Gaussian models. We also leverage a spatio-temporal graph to capture dynamic temporal and spatial cues of human motion where the former indicate the movement through time and the latter are for social awareness. Socially interacting people are modeled based on relative position, velocity and walking direction among them. The experiments over several real-world datasets show that our model can predict the distribution associated with future

states of people. Diverse acceptable trajectories can be generated from the distribution, which also demonstrate that our model can capture the multimodality of human motion.

References

Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; and Savarese, S. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 961–971.

Amirian, J.; Hayet, J.-B.; and Pettré, J. 2019. Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, CoRR, vol. abs/1904.09507.

Becker, S.; Hug, R.; Hübner, W.; and Arens, M. 2018. An evaluation of trajectory prediction approaches and notes on the trajnet benchmark. *arXiv preprint arXiv:1805.07663*.

Bishop, C. M. 1994. Mixture density networks. *Technical Report.*

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2625–2634.

Fernando, T.; Denman, S.; Sridharan, S.; and Fookes, C. 2018a. Gd-gan: Generative adversarial networks for trajectory prediction and group detection in crowds. In *Asian Conference on Computer Vision*, 314–330. Springer.

Fernando, T.; Denman, S.; Sridharan, S.; and Fookes, C. 2018b. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. *Neural networks* 108:466–478. Elsevier.

Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; and Alahi, A. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2255–2264.

Helbing, D., and Molnar, P. 1995. Social force model for pedestrian dynamics. *Physical review E* 51(5):4282. APS.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780. MIT Press.

Jain, A.; Zamir, A. R.; Savarese, S.; and Saxena, A. 2016. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5308–5317.

Karasev, V.; Ayvaci, A.; Heisele, B.; and Soatto, S. 2016. Intent-aware long-term prediction of pedestrian motion. In *Robotics and Automation , 2016 IEEE International Conference on*, 2543–2549. IEEE.

Kitani, K. M.; Ziebart, B. D.; Bagnell, J. A.; and Hebert, M. 2012. Activity forecasting. In *European Conference on Computer Vision*, 201–214. Springer.

Lee, N.; Choi, W.; Vernaza, P.; Choy, C. B.; Torr, P. H.; and Chandraker, M. 2017. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 336–345.

Lerner, A.; Chrysanthou, Y.; and Lischinski, D. 2007. Crowds by example. In *Computer Graphics Forum*, volume 26, 655–664. Wiley Online Library.

Liu, Q.; Wu, S.; Wang, L.; and Tan, T. 2016. Predicting the next location: A recurrent model with spatial and temporal contexts. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Pellegrini, S.; Ess, A.; Schindler, K.; and Van Gool, L. 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Computer Vision, 2009 IEEE 12th International Conference on*, 261–268. IEEE.

Robicquet, A.; Sadeghian, A.; Alahi, A.; and Savarese, S. 2016. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, 549–565. Springer.

Sadeghian, A.; Kosaraju, V.; Sadeghian, A.; Hirose, N.; Rezatofighi, H.; and Savarese, S. 2019. 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*, 1349–1358.

Shu, T.; Peng, Y.; Fan, L.; Lu, H.; and Zhu, S.-C. 2018. Perception of human interaction based on motion trajectories: From aerial videos to decontextualized animations. *Topics in cognitive science* 10(1):225–241.Wiley Online Library.

Su, H.; Zhu, J.; Dong, Y.; and Zhang, B. 2017. Forecast the plausible paths in crowd scenes. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2772–2778.

Sun, L.; Yan, Z.; Mellado, S. M.; Hanheide, M.; and Duckett, T. 2018. 3dof pedestrian trajectory prediction learned from long-term autonomous mobile robot deployment data. In *2018 IEEE International Conference on Robotics and Automation*, 1–7. IEEE.

Trautman, P., and Krause, A. 2010. Unfreezing the robot: Navigation in dense, interacting crowds. In *Intelligent Robots and Systems, 2010 IEEE/RSJ International Conference on*, 797–803. IEEE.

Vemula, A.; Muelling, K.; and Oh, J. 2018. Social attention: Modeling attention in human crowds. In *2018 IEEE International Conference on Robotics and Automation*, 1–7. IEEE.

Wang, L.; Zhang, L.; and Yi, Z. 2017. Trajectory predictor by using recurrent neural networks in visual tracking. *IEEE transactions on cybernetics* 47(10):3172–3183.

Yamaguchi, K.; Berg, A. C.; Ortiz, L. E.; and Berg, T. L. 2011. Who are you with and where are you going? In *Computer Vision and Pattern Recognition, 2011 IEEE Conference on*, 1345–1352. IEEE.

Zhang, P.; Ouyang, W.; Zhang, P.; Xue, J.; and Zheng, N. 2019. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12085–12094.