Scene Induced Multi-Modal Trajectory Forecasting via Planning

Nachiket Deo and Mohan M. Trivedi

***Abstract*— We address multi-modal trajectory forecasting of agents in unknown scenes by formulating it as a planning problem. We present an approach consisting of three models; a goal prediction model to identify potential goals of the agent, an inverse reinforcement learning model to plan optimal paths to each goal, and a trajectory generator to obtain future trajectories along the planned paths. Analysis of predictions on the Stanford drone dataset, shows generalizability of our approach to novel scenes.**

I. Introduction

To safely and efficiently navigate through spaces shared with humans, autonomous robots need the ability to forecast human motion. An inherent difficulty in motion forecasting is its multi-modal nature. In a given scene, a human can have one of multiple goals, with multiple paths to each goal. Regression based approaches for motion forecasting tend to suffer from mode collapse, resulting in averaged trajectories that may not conform with the scene.

Prior works have addressed this challenge by learning one-to-many mappings, from available context such as scene cues and past motion, to multiple future trajectories. This is typically done by sampling conditional generative models [1]–[6], or learning mixture models [7]–[9]. However, the high dimensionality of the output space poses a challenge for such models to generalize to novel scenes, especially since each scene can have paths, goals and decision nodes in various configurations.

Another set of approaches [10]–[13] pioneered by Ziebart *et al.* [10] formulate motion forecasting as a reinforcement learning agent exploring a grid defined over the scene. A reward map for the agent is learned via maximum-entropy inverse reinforcement learning (max-ent IRL) [14]. This allows for a more intuitive model for the agent's decision making. Also, since the reward map is learned from local scene cues at each grid cell, it allows for better generalization to novel scenes. However, max-ent IRL approaches suffer from two limitations. They require absorbing goal states in the scene to be known beforehand [10] or uniformly sampled from the scene [11]. More importantly, they can only provide paths taken by the agent in the grid, without mapping them to times in the future. While this is partly addressed in [10] by decomposing state visitation frequencies over time steps, using a Gaussian distribution, this does not take into account the agent's dynamics. A fast moving agent would make more progress than a slow moving agent along a planned path, over a fixed prediction horizon.

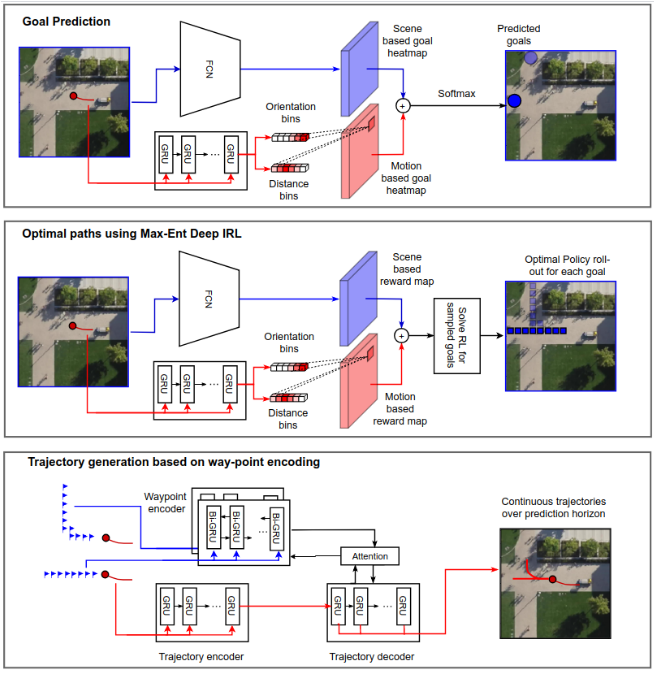

Fig. 1: Proposed models for scene-induced multimodal trajectory forecasting

In this work, we propose a planning based approach that can generalize to novel scenes, while not requiring goal states to be known beforehand, and generating continuous valued trajectories, preserving temporal information. Our approach consists of three models:

- *Goal prediction model:* To determine possible goals of the agent, based on the scene and their past motion
- *Optimal path planner:* To determine paths to each sampled goal using max-ent deep IRL
- *Trajectory generator:* To output continuous trajectories over the prediction horizon based on the agent's past motion, and an encoding of the planned paths.

II. Proposed Model

A. Goal Prediction:

The goal prediction model (Fig. 1, top) consists of two branches; a fully convolutional network (FCN) encoding the scene, and a gated recurrent unit (GRU) encoding the past trajectory. The FCN outputs a heatmap of potential goals in the scene, such as points where paths exit the scene, entrances to buildings etc. The trajectory encoder outputs activations for a discrete set of orientations and distances. These activations are mapped to the 2-D grid by taking a weighted sum of the two nearest distance and orientation values for the center of each cell. The trajectory encoder allows the model to narrow down potential goals based on

The authors are with the Laboratory for Intelligent and Safe Automobiles, University of California, San Diego, CA 92092, USA. ndeo@ucsd.edu, mtrivedi@ucsd.edu

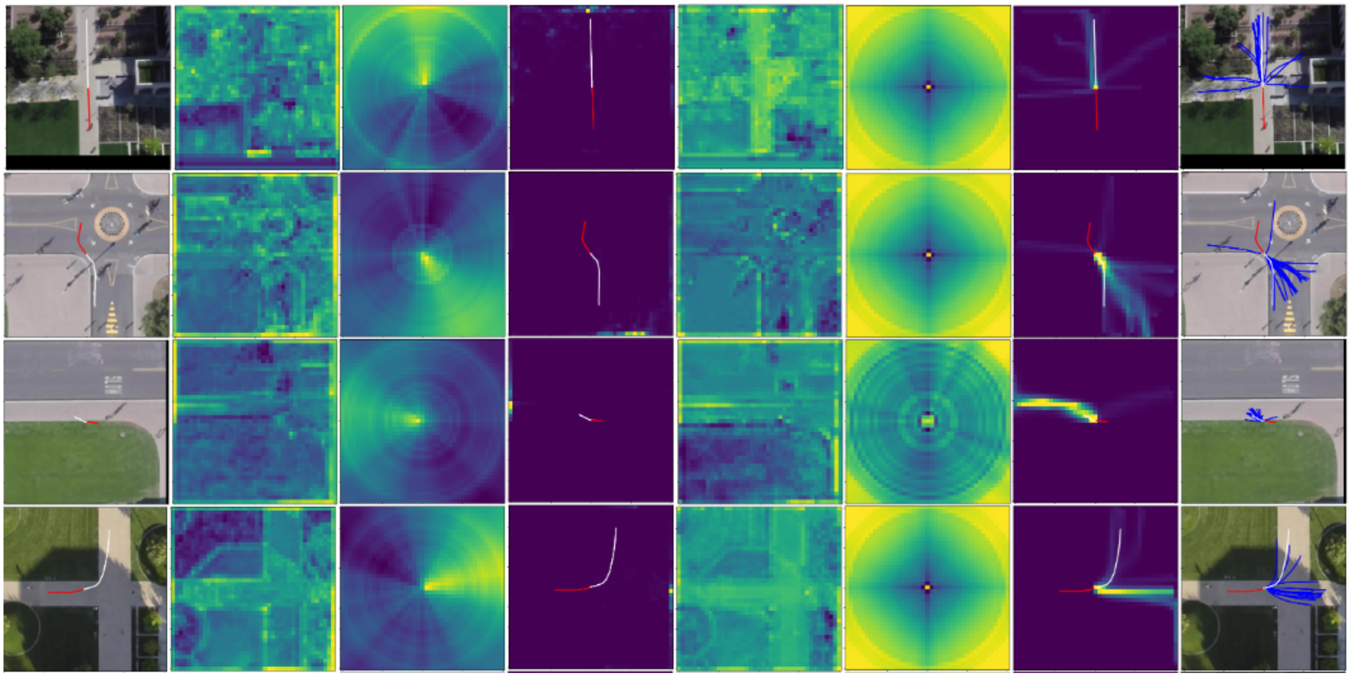

Fig. 2: **Example predictions.** From left to right: (1) Scene, past motion(red) and true futre trajectory (white),(2) scene-based goal heatmap, (3) motion based goal heatmap, (4) predicted goals, (5) scene based reward, (6) motion based reward, (7) state visitation frequencies, (8) predicted trajectories (blue)

the past motion of the agent. The two encodings are added and passed through a softmax layer to give goal probabilities on the grid. The model is trained to minimize cross-entropy with respect to the true goal.

B. *Optimal paths using Maximum Entropy Deep IRL:*

We use max-ent deep IRL [12] to learn a reward map on the scene. The reward model (Fig. 1, middle) is identical in structure to the goal prediction model. However, the heatmaps produced by the FCN and trajectory encoder can be interpreted as the scene-based and motion-based rewards over the grid. Since the FCN processes local patches of the scene, the reward map can generalize to novel scenes. During inference, we solve forward reinforcement learning to get the optimal policy conditioned on goals sampled from the goal prediction model. Roll-outs of the optimal policies give paths to each goal that conform to the scene.

C. *Trajectory generation using way-point encoding:*

The trajectory generation model (Fig. 1, bottom) generates continuous valued trajectories over the prediction horizon, conditioned on past motion and the optimal planned paths. The past trajectory is encoded by a GRU. We treat the grid locations of the planned paths as way-points in the scene. We encode the way-points using a bidirectional GRU encoder. The trajectory generator is a GRU decoder equipped with soft-attention [15]. Since all the way-points are typically not reached over the prediction horizon considered, the attention based decoder can attend to relevant way-points as it outputs the trajectory.

III. Experimental Evaluation:

TABLE I: Results on the Stanford drone dataset

Metric	SocialGAN [1]	DESIRE* [3]	MATF GAN [2]	SoPhie [5]	Ours
mADE	27.25	19.25	22.59	16.27	**15.73**
mFDE	41.44	34.05	33.53	29.38	**28.18**

*DESIRE uses K=5, while other approaches use K=20

We use the Stanford drone dataset (SDD) [16] for our experiments. It consists of trajectory data captured at 60 different scenes, with their top-down images. We use the standard benchmark split [17] for train, validation and test sets. While evaluating a multi-modal trajectory forecasting model, we need to ensure that plausible future trajectories generated by the model that do not correspond to the true future trajectory are not penalized. We thus use the minimum average displacement error (mADE) and minimum final displacement error (mFDE) metrics to evaluate our model:

$$mADE = \min_{k \in \{1,2,\ldots,K\}} \frac{1}{T} \sum_{t=1}^{T} \left\| y_t - \hat{y}_t^{(k)} \right\|_2 , \qquad (1)$$

$$mFDE = \min_{k \in \{1,2,\ldots,K\}} \left\| y_T - \hat{y}_T^{(k)} \right\|_2 , \qquad (2)$$

where T is the prediction horizon, $y_{1:T}$ is the true future trajectory for a given instance, and $\hat{y}_{1:T}^{(k)}$ are trajectories sampled from our model. Similar to prior work [1], [2], [5], we choose $K = 20$, with a prediction horizon of 4.8 s, and past history of 3.2 s. Table I shows the mADE and mFDE values for the SDD test set. Our approach achieves state of the art results on SDD. Additionally, we provide qualitative examples of predictions made by our model, shown in Figure 2. We can observe that our models generate a diverse set of future trajectories that conform with the underlying scene and past motion of the agent.

References

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

[2] T. Zhao, Y. Xu, M. Monfort, W. Choi, C. Baker, Y. Zhao, Y. Wang, and Y. N. Wu, “Multi-agent tensor fusion for contextual trajectory prediction,” *arXiv preprint arXiv:1904.04776*, 2019.

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

[4] N. Rhinehart, K. M. Kitani, and P. Vernaza, “R2p2: A reparameterized pushforward policy for diverse, precise generative path forecasting,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 772–788.

[5] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, and S. Savarese, “Sophie: An attentive gan for predicting paths compliant to social and physical constraints,” *arXiv preprint arXiv:1806.01482*, 2018.

[6] A. Bhattacharyya, B. Schiele, and M. Fritz, “Accurate and diverse sampling of sequences based on a best of many sample objective,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8485–8493.

[7] N. Deo and M. M. Trivedi, “Multi-modal trajectory prediction of surrounding vehicles with maneuver based lstms,” in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1179–1184.

[8] H. Cui, V. Radosavljevic, F.-C. Chou, T.-H. Lin, T. Nguyen, T.-K. Huang, J. Schneider, and N. Djuric, “Multimodal trajectory predictions for autonomous driving using deep convolutional networks,” *arXiv preprint arXiv:1809.10732*, 2018.

[9] N. Deo and M. M. Trivedi, “Convolutional social pooling for vehicle trajectory prediction,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1468–1476.

[10] B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. Srinivasa, “Planning-based prediction for pedestrians,” in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 3931–3936.

[11] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, “Activity forecasting,” in *European Conference on Computer Vision*. Springer, 2012, pp. 201–214.

[12] M. Wulfmeier, D. Z. Wang, and I. Posner, “Watch this: Scalable cost-function learning for path planning in urban environments,” in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 2089–2095.

[13] Y. Zhang, W. Wang, R. Bonatti, D. Maturana, and S. Scherer, “Integrating kinematics and environment context into deep inverse reinforcement learning for predicting off-road vehicle trajectories,” *arXiv preprint arXiv:1810.07225*, 2018.

[14] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, “Maximum entropy inverse reinforcement learning.” in *Aaai*, vol. 8. Chicago, IL, USA, 2008, pp. 1433–1438.

[15] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.

[16] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, “Learning social etiquette: Human trajectory understanding in crowded scenes,” in *European conference on computer vision*. Springer, 2016, pp. 549–565.

[17] A. Sadeghian, V. Kosaraju, A. Gupta, S. Savarese, and A. Alahi, “Trajnet: Towards a benchmark for human trajectory prediction,” *arXiv preprint*, 2018.