Learning to Infer Relations for Future Trajectory Forecast

Chiho Choi and Behzad Dariush
Honda Research Institute USA
{cchoi,bdariush}@honda-ri.com

Abstract

Inferring relational behavior between road users as well as road users and their surrounding physical space is an important step toward effective modeling and prediction of navigation strategies adopted by participants in road scenes. To this end, we propose a relation-aware framework for future trajectory forecast, which aims to infer relational information from the interactions of road users with each other and with environments. Extensive evaluations on a public dataset demonstrate the robustness of the proposed framework as observed by performances higher than the state-of-the-art methods.

1. Introduction

Forecasting future trajectories of moving participants in indoor and outdoor environments has profound implications for execution of safe navigation strategies in partially and fully automated vehicles and robotic systems. Related research has attempted to predict future trajectories by focusing on human[1]-human (*i.e.*, between road users) or human-space (*i.e.*, between road user(s) and space) interactions.

Discovering social interactions between humans has been a mainstream approach to predict future trajectories. Following [4] on modeling human-human interactions, social models have been presented for the data-driven methods [1, 3]. While successful in many cases, they may fail to provide acceptable paths in a complex environment without the guidance of scene context. Modeling human-space interactions of nearby humans [5] toward surrounding environments has been introduced as an additional modality to social interactions. Although they consider scene context to capture human-space interactions, this approach restricts interactions to nearby neighbors and overlooks the influence of distant obstacles in navigation, which is not feasible in real-world scenarios. In this view, we present a relation-aware framework where such interactions are not limited to nearby road users nor surrounding medium (see Figure 1).

[1]The word 'human' refers to any types of road user – pedestrian, car, cyclist, *etc*. – in the rest of this paper.

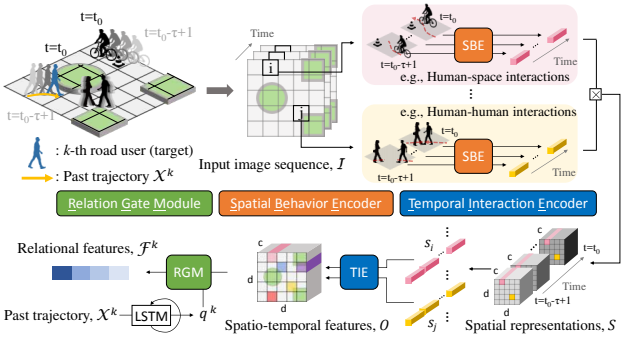


Figure 1: The proposed gated relation encoder (GRE) visually discovers both human-human (j-th region: woman↔man) and human-space interactions (i-th region: cyclist↔cone) from each region of the discretized grid over time. Then, their pair-wise relations (*i.e.*, ■↔■, ■↔■, ■↔■, ■↔■, ■↔■, ...) with respect to the past motion of the target (→) are investigated from a global perspective.

Relational inference in [8] is inherently flexible to define 'an object' as *a spatial feature representation* extracted from each region of the discretized grid regardless of what exists in that region. Our work is analogous to [8] in the sense that the word 'object' is defined. In our framework, an object is a visual encoding of spatial behavior of road users (if they exist) and environmental representations together with their temporal interactions over time, which naturally corresponds to local human-human and human-space interactions in each region of the discretized grid. On top of this, we consider learning to infer relational behavior from all objects (*i.e.*, spatio-temporal interactions in our context) from a global perspective as shown in Fig. 1.

In practice, the relations between all object pairs do not equally contribute to predict the motion of a specific road user. For example, a distant building behind a car does not have meaningful relational information with the ego-vehicle that is moving forward. To address the different importance of relations, we design a relation gate module (RGM) with an internal gating process. Our RGM is beneficial to control of information flow through multiple switch gates and identifies descriptive relations that highly influence the future motion of the target by conditioning on its past trajectory.

An overview of our approach is as follows. Given a sequence of images, the gated relation encoder (GRE) visually extracts spatio-temporal interactions (*i.e.*, objects) through the spatial behavior encoder (SBE) and temporal interaction encoder (TIE) as shown in Fig. 1. The following RGM of GRE infers pair-wise relations from objects and then focuses on looking at which relations will be potentially meaningful to forecast the future motion of the target under its past behavior. We predict future locations using the aggregated relational features through the trajectory prediction network (TPN) in the form of heatmaps which can be further refined by considering spatial dependencies between predicted locations and extended to learn the uncertainty of future forecast at test time.

2. Methodology

2.1. Spatio-Temporal Interactions

We extend the definition of 'object' in [8] to a spatio-temporal feature extracted from each region of the discretized grid over time. It enables us to visually discover (i) *human-human interactions* where there exist multiple road users interacting with each other, (ii) *human-space interactions* from their interactive behavior with environments, and (iii) *environmental representations* by encoding structural information of the road. Given τ number of past images $\mathcal{I} = \{I_{t_0-\tau+1}, I_{t_0-\tau+2}, ..., I_{t_0}\}$, we visually extract spatial representations of the static road structures, the road topology, and the appearance of road users from individual frames using the SBE. The concatenated features along the time axis are spatial representations $S \in \mathbb{R}^{\tau \times d \times d \times c}$. As a result, each entry $s_i \in \mathbb{R}^{\tau \times 1 \times 1 \times c}$ of $S = \{s_1, .., s_n\}$ contains frame-wise knowledge of road users and road structures in i-th region of the discretized grid as in Fig. 1. Therefore, we individually process each entry s_i of S using the TIE to model sequential changes of road users with respect to environments. The resulting spatio-temporal features $O \in \mathbb{R}^{d \times d \times c}$ thus contains a visual interpretation of spatial behavior of road users and their temporal interactions with each other and with environments. We decompose O into a set of objects $\{o_1, ..., o_n\}$, where $n = d^2$ and an object $o_i \in \mathbb{R}^{1 \times 1 \times c}$ is a c-dimensional feature vector.

2.2. Relational Inference

Observations from actual prediction scenarios in road scenes suggest that humans focus on only few important relations that may potentially constrain the intended path, instead of inferring every relational interactions of all road users. In this view, we create the RGM which is able to address the benefits of discriminatory information process with respect to their relational importance.

Let $RGM(\cdot)$ be a function which takes as input a pair of two objects (o_i, o_j) and spatial context q^k. Note that q^k is a feature representation extracted from the past trajectory $\mathcal{X}^k = \{X^k_{t_0-\tau+1}, X^k_{t_0-\tau+2}, ..., X^k_{t_0}\}$ of the k-th road user observed in $\mathcal{I}$. Then, the inferred relational features $\mathcal{F}^k$ are as follows: $\mathcal{F}^k = \sum_{i,j} RGM(o_i, o_j, q^k)$. Through this function, we first determine whether the given object pair has meaningful relations from a spatio-temporal perspective by computing $r_{ij} = tanh_\alpha(o_{ij}) \odot \sigma_\beta(o_{ij})$, where $o_{ij} = o_i \boxtimes o_j$ is the concatenation of two objects. Note that we add $\alpha, \beta, \mu, \lambda$ as a subscript of hyperbolic tangent $tanh(\cdot)$ and sigmoid $\sigma(\cdot)$ function to present that these functions come after a fully connected layer. Then, we identify how their relations can affect the future motion of k based on its past motion context q^k by $f^k_{ij} = tanh_\lambda(r_{ij} \boxtimes tanh_\mu(q^k))$. This step is essential in (i) determining whether the given relations r_{ij} would affect the target road user's potential path and (ii) reasoning about the best possible route, given the motion history q^k of the target. We subsequently collect all relational information from every pair and perform element-wise sum to produce target-specific relational features $\mathcal{F}^k$.

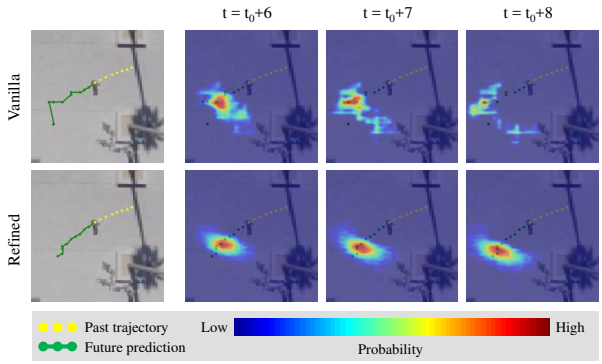


Figure 2: The efficacy of the SRN for spatial dependencies.

2.3. Trajectory Prediction

The relational features $\mathcal{F}^k$ extracted from GRE are incrementally upsampled using a set of deconvolutional layers through the TPN. As an output, the network predicts δ number of future locations in the form of heatmaps $\widehat{\mathcal{H}}^k_A \in \mathbb{R}^{W \times H \times \delta}$. At training time, we use the L2 Loss $\mathcal{L}_\mathcal{A} = \sum_\delta \sum_{u,v} \left(\mathcal{H}^k_{(\delta)}(u,v) - \widehat{\mathcal{H}}^k_{A(\delta)}(u,v)\right)^2$ to minimize the sum of squared error between the ground-truth heatmaps $\mathcal{H}^k$ and the prediction $\widehat{\mathcal{H}}^k_A$, all over the 2D locations (u, v).

2.4. Refinement with Spatial Dependencies

Predicted heatmaps from the TPN are sometimes ambiguous as in Fig. 2. Our main insight for this issue is a lack of *spatial dependencies* among predictions. Since the network independently predicts δ heatmaps, there is no constraint to enforce them to be spatially aligned between predictions. Thus, we design a spatial refinement network (SRN) to learn implicit spatial dependencies in a feature space. We first concatenate intermediate activations (early and late features) of the TPN and let through

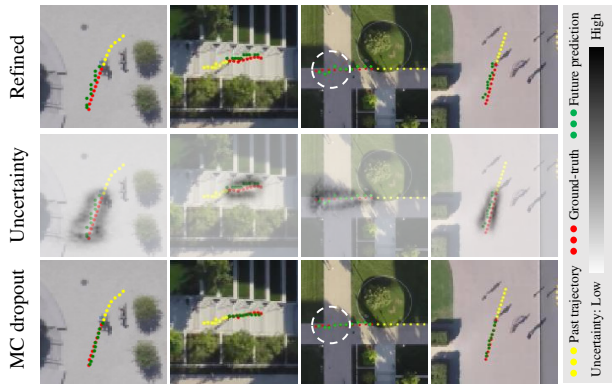


Figure 3: The efficacy of the uncertainty embedding into our framework with MC dropout.

the SRN using large receptive fields [6, 9]. As a result, the outputs $\widehat{\mathcal{H}}_O^k$ show less confusion between heatmap locations, making use of rich contextual information from neighboring predictions. To train the SRN together with optimizing the rest of the system, we define another L2 loss $\mathcal{L}_\mathcal{O} = \sum_\delta \sum_{u,v} \left(\mathcal{H}^k_{(\delta)}(u,v) - \widehat{\mathcal{H}}^k_{O(\delta)}(u,v)\right)^2$. Then the total loss is as follows: $\mathcal{L}_{optimize} = \zeta\mathcal{L}_\mathcal{A} + \eta\mathcal{L}_\mathcal{O}$, where $\zeta = \eta = 1$ which properly optimizes our SRN with respect to the TPN and GRE.

2.5. Uncertainty of Future Prediction

Bayesian neural networks (BNNs) bave been considered to tackle the uncertainty of the network's weight parameters. [2] found that inference in BNNs can be approximated by sampling from the posterior distribution of the deterministic network's weight parameters using Monte Carlo (MC) dropout. Details of MC dropout are skipped for brevity. We perform approximating variational inference [2] using dropout at test time to draw multiple samples from the dropout distribution. It literally enables us to capture multiple plausible trajectories over the uncertainties of the network's learned weight parameters. However, we use the mean of L samples as our prediction, which best approximates variational inference in BNNs. The efficacy of the uncertainty embedding is visualized in Fig. 3. We compute the variance of $L = 5$ samples to measure the uncertainty.

3. Experiments

We use a public dataset [7] for evaluation, collected from a drone capturing top-down road scenes. Heatmaps $\mathcal{H}$ are generated in the 128x128 image space using the center coordinates $\mathcal{Y} = \{Y_{t_0+1}, Y_{t_0+2}, ..., Y_{t_0+\delta}\}$ of road users. We use 3.2 sec of past images $\mathcal{I}$ and coordinates $\mathcal{X}^k$ of the target k as input and predict 4.0 sec of future frames. For evaluation, we find coordinates of a point with a maximum likelihood $\widehat{\mathcal{Y}}^k$ from heatmaps $\widehat{\mathcal{H}}^k$ and calculate average (ADE) and final distance error (FDE) in *pixels* between $\mathcal{Y}^k$ and $\widehat{\mathcal{Y}}^k$.

Method	1.0 *sec*	2.0 *sec*	3.0 *sec*	4.0 *sec*
S-LSTM [1]	1.93 / 3.38	3.24 / 5.33	4.89 / 9.58	6.97 / 14.57
DESIRE [5]	- / **2.00**	- / 4.41	- / 7.18	- / 10.23
RE_Conv2D	2.42 / 3.09	3.50 / 5.23	4.72 / 8.16	6.19 / 11.92
RE_2D+3D	2.36 / 2.99	3.33 / 4.80	4.37 / 7.26	5.58 / 10.27
GRE_Vanilla	1.85 / 2.41	2.77 / 4.27	3.82 / 6.70	5.00 / 9.58
GRE_Refine	1.71 / 2.23	2.57 / 3.95	3.52 / 6.13	4.60 / 8.79
GRE_MC-2	1.66 / 2.17	2.51 / 3.89	3.46 / 6.06	4.54 / 8.73
GRE_MC-5	**1.61** / 2.13	**2.44 / 3.85**	**3.38 / 5.99**	**4.46 / 8.68**

Table 1: Quantitative comparison (ADE / FDE in *pixels*) of our approach with the self-generated baselines and state-of-the-art methods [1, 5] using SDD [7] at 1 / 5 resolution.

3.1. Quantitative Comparisons

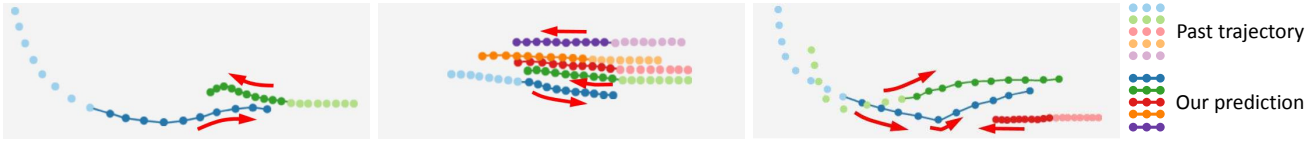
Spatio-temporal interactions: Encoding spatio-temporal features from images is crucial to discover both human-human and human-space interactions, which makes ours approach distinct from others. To demonstrate the rationale of using spatio-temporal interactions, we compare two baselines[2]: (i) *RE_Conv2D* which discovers only spatial interactions using 2D convolutions; and (ii) *RE_2D+3D* where we infer spatio-temporal interactions as discussed in Sec. 2.1. As shown in Tbl. 1, the performance of the *RE_2D+3D* baseline is dramatically improved against *RE_Conv2D* by replacing the final 2D convolution with a 3D convolution.
Relation gate module: We train the *GRE_Vanilla* baseline which replaces the fully connected layers of the $RE\_$ to the proposed RGM. The improvements of both ADE and FDE are achieved by a huge margin from the *RE_2D+3D* baseline. It implies the benefits of the RGM which controls information flow with respect to the relational importance.
Spatial refinement: *GRE_Refine* (with the SRN) significantly outperforms *GRE_Vanilla* from both metrics all over time. It validates that the proposed SRN effectively acquires rich contextual information about dependencies between future locations from initial predictions $\widehat{\mathcal{H}}_A$ in a feature space and hence removes unacceptable prediction failures.
Monte Carlo dropout: To validate our uncertainty strategy for future trajectory forecast, we generate two *GRE_MC* baselines with a different suffix *-L*, where L denotes the number of samples drawn at test time. The fact that any *GRE_MC-L* baselines consistently show the decrease in error rate for both near and far future prediction indicates the efficacy of the presented uncertainty embedding.
Comparison with literature: We compare the performance of our approach to two state-of-the-art methods, one from *human-human interaction* oriented approaches (S-LSTM [1]) and the other from *human-space interaction* oriented approaches (DESIRE [5][3]). The results in Tbl. 1 indicate that incorporating scene context is crucial to successful predictions as our method and [5] shows a lower error rate

[2]The baselines with a prefix *RE_* do not employ the proposed gating process but assume equal importance of relations similarly to [8].

[3]We use *DESIRE-SI-IT0 Best* which shows the best performance without using the oracle error metric.

(a) (b) (c) (d)

Figure 4: Qualitative evaluation. (Color codes: Yellow - given past trajectory, Red - ground-truth, and Green - our prediction)



(a) (b) (c)

Figure 5: Illustrations of our prediction during complicated human-human interactions. (a) A cyclist (●●●) interacts with a person moving slow (●●●). (b) A person (●●●) meets a group of people. (c) A cyclist (●●●) first interacts with another cyclist in front (●●●) and then considers the influence of a person (●●●). The proposed approach socially avoids potential collisions.

than that of [1]. Moreover, our models with GRE generally outperform [5], validating the robustness of the proposed spatio-temporal interactions encoding pipeline which is designed to discover the human-human and human-space interactions from local to global scales. Note that the effectiveness of our approach is especially pronounced toward far future predictions. Unlike [1, 5] which restrict human interactions to nearby surroundings, we do not limit the interaction boundary and hence achieve more accurate predictions toward the far future.

3.2. Qualitative Evaluation

Figure 4 qualitatively evaluates how inferred relations encourage our model to generate natural motion for the target with respect to the consideration of human-human interactions (4a) and human-space interactions (4b). Both cases clearly show that spatio-temporal relational inferences adequately constrain our future predictions to be more realistic. We also illustrate prediction failures in Figure 4c where the road user suddenly changes course and 4d where the road user is aggressive to interactions with an environment. Extension to incorporate such human behavior is our next plan. In Figure 5, we specifically illustrate more complicated human-human interaction scenarios. As validated in these examples, the proposed approach visually infers relational interactions based on the potential influence of others toward the future motion of the target.

4. Conclusion

We proposed a relation-aware framework which aims to forecast future trajectory of road users. Inspired by the human capability of inferring relational behavior from a physical environment, we introduced a system to discover both human-human and human-space interactions. The proposed approach first investigates spatial behavior road users and environments together with their temporal interactions. Then, we identify relations from these interactions, which have a high potential to influence the future motion of the target based on its past trajectory. To generate future trajectories, we predict a set of probability maps that can be further refined by considering spatial dependencies between the initial predictions as well as the nature of uncertainty in future forecast. Evaluations show that the proposed framework is powerful as it achieves state-of-the-art performance.

References

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, 2016.

[2] Y. Gal and Z. Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

[3] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, 2018.

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

[5] 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 *CVPR*, 2017.

[6] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, 2015.

[7] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*, 2016.

[8] A. Santoro, D. Raposo, D. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *NIPS*, 2017.

[9] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.