IP-MMT: Interaction Prediction via MultiModal Transformer

Lan Feng[1]*, Qihang Zhang[3]*, Yicheng Liu[3], Fan Li[1], Gang Sun[1], Chunxiao Liu[1], Bolei Zhou[23]
[1] SenseTime, [2] Centre for Perceptual and Interaction Intelligence,
[3] The Chinese University of Hong Kong,

Abstract

We provide a detailed description of our approach applied to the Interaction track of Waymo Open Dataset Challenges. Our approach mainly consists of two proposals, the dual interaction modeling and the parallel prediction with MultiModal Transformer (mmTransformer) [6]. Our approach achieves 0.0897 mAP and ranked 1st in the competition.

1. Introduction

There are various types of traffic participants on the road, such as pedestrians, vehicles, and cyclists. Interactive behaviors widely exist in traffic and pedestrian flows. Therefore, predicting the interaction behavior of different traffic participants is essential for autonomous driving.

However, forecasting the interactive behavior is challenging since the prediction should be socially acceptable and multi-modal. For instance, road users are governed by social norms, which are difficult to be formalized [4]. Moreover, interaction can occur between more than two road users. In a case when one vehicle decelerates to avoid a pedestrian, the vehicle behind the former vehicle will also decelerate. This kind of multi-agent interaction makes interaction prediction challenging.

To this end, we propose a dual interaction modeling method based on the recent VectorNet [3] and Multi-Modal Transformer (mmTransformer) [6], and develop a modal-wise self-attention to tackle these interaction prediction problems. Based on the generator of the mmTransformer, we introduce the following two novelties:

1. We model weak interaction and strong interaction by performing vectorizing analysis and modal-wise self-attention respectively to make interactive behavior prediction more computationally efficient and effective.

2. We design explicit modeling for scene-context and interaction-context to enhance mmTransformer's multi-agent parallel prediction performance.

*Lan Feng and Qihang Zhang contribute equally to this work.

2. Related Work

Interactive Behaviour Prediction. Various frameworks have been proposed to tackle the problem of interactive behaviour prediction such as game theory[7] and graph neural network[8, 2]. Another branch of methods perform feature aggregation among multiple agents. CNN is used in [9] to capture local interactions between pedestrians. [1] use a fully-connected layer to aggregate neighboring agents' information. SocialGAN[4] use a pooling layer to share information across different agents. [5, 6] use self-attention or Transformer-based architecture to enable information propagation.

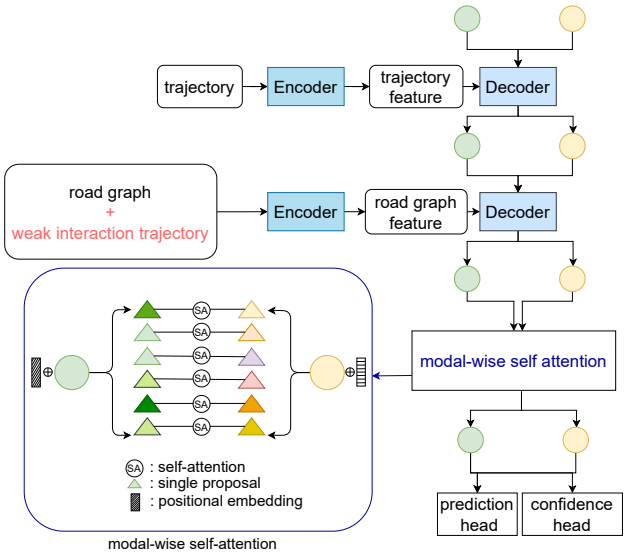


Figure 1: The architecture of IP-MMT. Circles of different colors represent all the proposals of each agent, and triangles represent different proposals. Dual interaction information is modeled by VectorNet and modal-wise self-attention.

model	wi	si	Vehicle minADE	Vehicle MR	Vehicle mAP	Pedestrian minADE	Pedestrian MR	Pedestrian mAP	Cyclist minADE	Cyclist MR	Cyclist mAP
IP-MMT	✓		1.30	0.56	0.12	0.92	0.60	0.05	1.43	0.75	0.03
		✓	1.29	0.55	0.13	0.92	0.60	0.05	1.41	0.74	0.03
	✓	✓	**1.28**	**0.55**	**0.13**	**0.91**	**0.59**	**0.06**	**1.41**	**0.74**	**0.03**
mmTransformer	✓	✓	1.29	0.55	0.13	0.92	0.60	0.05	1.41	0.74	0.03

Table 1: **Joint metrics on the standard validation**. All metrics are the average at 3s, 5s, 8s. wi stands for weak interaction. si stands for strong interaction. All models employ 6 proposals.

3. Interaction Prediction By mmTransformer

In this work, we use mmTransformer [6] as our base model. We group environmental observations into two types: scene-context and interaction-context, and design an explicit mechanism to process them. Our method is illustrated in Fig 1 and described in detail below.

3.1. Scene-Context

We regulate ego history and road graph observations into each agent's own coordination, wiping away their global positional information to make parallel prediction. The lost global positional information is further fused by positional embedding in interaction-context modeling.

History Trajectory. We use the motion extractor in mmTransformer to process agents' history trajectories. We denote the history trajectories of agents in global coordinates as $H = \{h_i \in \mathcal{R}^{T_{obs} \times 2} : 1 \leq i \leq N_{vehicle}\}$. Since the heading angle and global position is redundant for agent's internal state's understanding, we transform the trajectories into agent's ego coordinate system:

$$H_{ego} = \{R_i h_i + p_i, h_i \in H\}, \tag{1}$$

where R_i is the rotation matrix and p_i is the displacement vector of the each agent at current time point (the start of the prediction interval).

The encoder inputs are history trajectories H_{ego}. The decoder inputs are the trajectory proposals $Y = \{y_i \in \mathcal{R}^n : 1 \leq i \leq K\}$, representing K learnable positional embeddings with feature size n.

Local Road Graph. We use the map aggregator in mmTransformer to process road graph. For each agent, we define a receptive field $\mathcal{O}$ (A circle with radius of 20 meters and a center 15 meters in front of the current point). Only road segments with at least one dot appeared in the receptive field $\mathcal{O}$ will be counted.

3.2. Interaction-Context Modeling

We find that the two agents designated by the Waymo Challenge also have weak interaction with other agents.

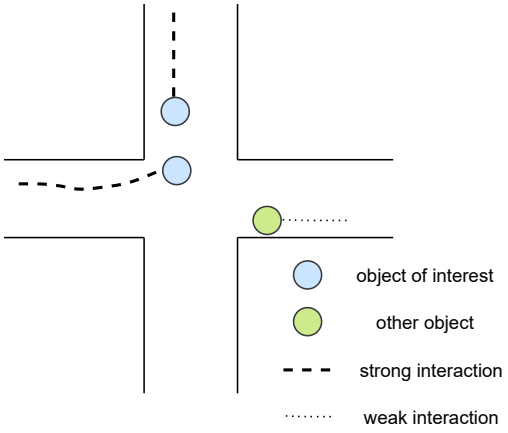


Figure 2: Illustration of strong and weak interaction.

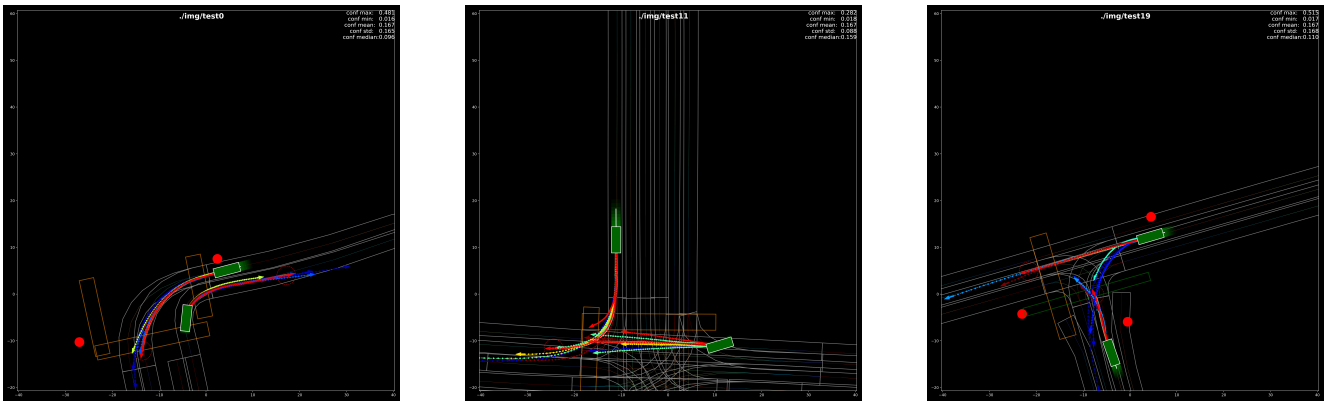
Gao *et al*. [3] treat trajectories and road graph as vectors and used graph neural network to model the correlation between them. We considered VectorNet as a weak interaction modeling method as it only contains history trajectory information, while the future distribution of each agent's trajectory, such as proposal embedding in mmTransformer, contains more information for interactive prediction. Therefore we use VectorNet and modal-wise self-attention as the dual interaction modeling method to better predict interactive behaviors. The illustration of strong and weak interaction is shown in Table 3.

3.2.1 Weak Interaction-Context Modeling

To model weak interaction-context, we treat each agent's nearby trajectories V_i as a special type of road segment. Then we concatenate V_i with the local road graph G_i and encode them respectively with VectorNet's subgraph encoding module. The interactive contexts will be learnt by global self-attention implicitly.

3.2.2 Strong Interaction-Context Modeling

The interaction track of Waymo Challenge requires joint prediction of selected two road users. To better model the



(a) Case A (b) Case B (c) Case C

Figure 3: Three visualized cases show that IP-MMT gives consideration to multi-modality and interaction. Red dot lines are ground truths for each track. Other dot lines are predicted tracks.

joint confidence score and coordinate prediction, we applied modal-wise self-attention, which we defined as:

$$P'_a, P'_b = \textit{modal-wise self-attention}(P_a, P_b), \qquad (2)$$

where P_a, P_a are all K proposal features of selected two agents, the detail of modal-wise self-attention is shown is Fig 1.

3.2.3 Generation Head

After refining each agent's proposals with scene-context and interaction-context information, we design a MLP-based generation head to predict the future trajectory. Especially, we use multiple heads to model the interaction characteristics between different types of road users (e.g. vehicle-pedestrian, vehicle-cyclist, vehicle-vehicle).

4. Implementation Details

For training, we use the same network depth settings as mmTransformers.

5. Experiments

5.1. Main Results

In this section, we evaluate our method on Waymo open datasets. The results of our approaches on interaction prediction validation dataset are summarized in Table 1.

5.2. Results Analysis

The experimental results show that both strong and weak interactions contribute to the improvement of the final mAP. Besides, our proposed scene-context modeling also improves model's performance. Some qualitative results are shown in Table 1. Some visualizations of predicted tracks are shown in Fig 3

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. 1

[2] Sergio Casas, Cole Gulino, Renjie Liao, and Raquel Urtasun. Spatially-aware graph neural networks for relational behavior forecasting from sensor data. *arXiv preprint arXiv:1910.08233*, 2019. 1

[3] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11525–11533, 2020. 1, 2

[4] 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. 1

[5] Lingyun Luke Li, Bin Yang, Ming Liang, Wenyuan Zeng, Mengye Ren, Sean Segal, and Raquel Urtasun. End-to-end contextual perception and prediction with interaction transformer. *arXiv preprint arXiv:2008.05927*, 2020. 1

[6] Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. Multimodal motion prediction with stacked transformers. *arXiv preprint arXiv:2103.11624*, 2021. 1, 2

[7] Wei-Chiu Ma, De-An Huang, Namhoon Lee, and Kris M Kitani. Forecasting interactive dynamics of pedestrians with fictitious play. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 774–782, 2017. 1

[8] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *Proceedings of the*

AAAI Conference on Artificial Intelligence, volume 33, pages 6120–6127, 2019. 1

[9] Shuai Yi, Hongsheng Li, and Xiaogang Wang. Pedestrian behavior understanding and prediction with deep neural networks. In *European Conference on Computer Vision*, pages 263–279. Springer, 2016. 1