InAction: Interpretable Action Decision Making for Autonomous Driving - Supplementary

Taotao Jing[1], Haifeng Xia[1], Renran Tian[2], Haoran Ding[2], Xiao Luo[2], Joshua Domeyer[3], Rini Sherony[3], and Zhengming Ding[1]

[1] Tulane University, New Orleans LA 70118, USA
[2] Indiana University-Purdue University Indianapolis , Indianapolis IN 46202, USA
[3] Collaborative Safety Research Center (CSRC), Toyota Motor North America, Ann Arbor MI 48105, USA
{tjing, hxia, zding1}@tulane.edu, {rtian, luo25}@iupui.edu,
{joshua.domeyer, rini.sherony}@toyota.com, hd10@iu.edu

1 Implementation and Results on PSI Dataset

For a fair comparison with OIA, we set $N = 10$ for both OIA and our proposed InAction for two evaluated benchmarks, i.e., BDD-OIA and PSI in the manuscript and supplementary material. Specifically, for OIA [3], 10 proposals generated by Faster R-CNN with the highest probability produced by the *Selector* are concatenated to the global features. For our InAction model, the $top-10$ patches from the input feature map with the smallest distances compared to all semantic prototypes are selected to be fused with the global features.

In Figure 1, selected samples from PSI dataset are shown with both action and explanation prediction produced by our InAction model.

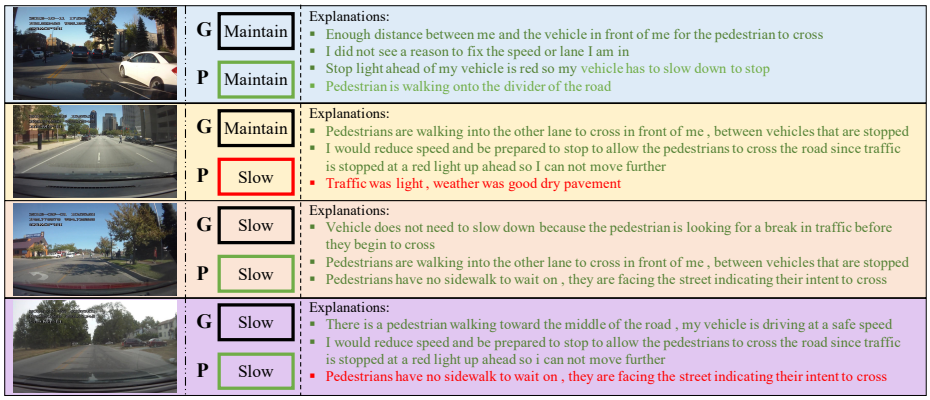

Fig. 1. Visualization of explanation and action prediction on PSI dataset.

2 Limitations and Future Exploration

For the experiments implemented in the manuscript, $m_k = 6$ prototypes are assigned to each action category. Some latest prototype-based works propose various strategies to pruning the framework and only keep the most activated prototypes, which is effective to make the final prediction [3], [2], [1]. Such model simplification and prototypes pruning strategies can also be applied to our model, reducing the complexity of the framework. Further study on this topic is out of scope in this paper.

3 Analysis about Components Contributions

The goal of our work is to *predict the driver decision* from two different perspectives' interpretation, leading to better AI transparency and reliability (Line 250-256). The two perspectives are denoted as *Implicit visual-semantic explanation* for AI and *Explicit human-annotated reasoning* for Human Cognitive, which both cast a light to the AI prediction behavior understanding for end users. In Table R1, we have evaluated several variants of our model on PSI to study the contribution of each module by removing corresponding learning objective(s). **First**, we evaluate "Explicit only" and "Implicit only" by removing Upper or Bottom branch in Fig. 1. We notice that neither of them can obtain comparable results to the complete model, demonstrating the contribution of complimentary interaction of the two modules. **Second**, we further add a fully-connected layer as a human explanation predictor to the "Implicit only' module and report the results as "Implicit + human expl.". The results are improved significantly compared to "Implicit only", proving the contribution of the human reasoning annotations. **Third**, we adopt the proposals/objects detected by the pre-trained Faster R-CNN [27] as the local features fused with global features and report as "Explicit + proposals", which is also studied on BDD-OIA dataset in **Table 3** as "Ours (proposals)". The results support our claim on the limitation of relying on pre-trained detection models output as Line 479-482. **Fourth**, we compare the performance with only the last frame as input as "Last frame only", which is also reported as "Ours-f" in **Table 2**, and the results decrease due to the lack of temporal knowledge of video sequence. From both PSI and BDD-OIA datasets, we observe many ambiguous situations if only one frame is adopted to make the decision. **Finally**, we evaluate one more variant without the regularization $\mathcal{L}_d$ on the prototypes, and obtain worse performance than our complete model. We will add such ablation study and more analysis in the final version.

4 Comparison between Two Branches

We explore the specific performances of the two action predictors, $C_S(\cdot)$ and $C_R(\cdot)$, and demonstrate the effectiveness of cross-module prediction fusion. From the results in Figure 2, we notice that either $C_S(\cdot)$ and $C_R(\cdot)$ works better than

Table R1. Components contributions Analysis on PSI dataset

w/o Loss	Method	act. Acc_{all}	act. mAcc	exp. $F1_{all}$	exp. mF1
$\mathcal{L}_s, \mathcal{L}_d$	Explicit only	0.708	0.652	0.240	0.185
$\mathcal{L}_s, \mathcal{L}_d$	Explicit + proposals	0.713	0.678	0.243	0.189
$\mathcal{L}_r$	Implicit only	0.510	0.333	–	–
$\mathcal{L}_r$	Implicit + human expl.	0.693	0.688	0.232	0.175
$\mathcal{L}_t$	Last frame only	0.719	0.704	0.277	0.203
$\mathcal{L}_d$	No regularization	0.716	0.685	0.244	0.186
	Ours	**0.734**	**0.722**	**0.285**	**0.223**

the other on specific action class, and the fused prediction improves the performance and achieves the best results over both single predictor, which demonstrates the complimentary advantages of cross-module prediction fusion.

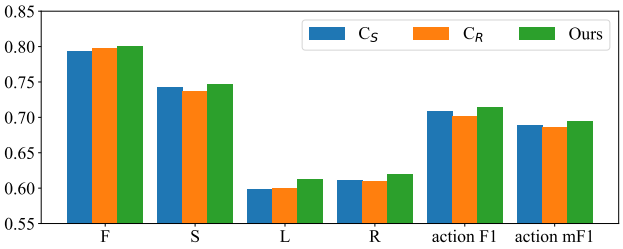

Fig. 2. Comparison of driving decision prediction produced by different modules in InAction on BDD-OIA dataset.

References

1. Ming, Y., Xu, P., Qu, H., Ren, L.: Interpretable and steerable sequence learning via prototypes. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 903–913 (2019)
2. Rymarczyk, D., Struski, Ł., Tabor, J., Zieliński, B.: Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 1420–1430 (2021)
3. Xu, Y., Yang, X., Gong, L., Lin, H.C., Wu, T.Y., Li, Y., Vasconcelos, N.: Explainable object-induced action decision for autonomous vehicles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9523–9532 (2020)