INTERPRETABLE SELF-ATTENTION TEMPORAL REASONING FOR DRIVING BEHAVIOR UNDERSTANDING

Yi-Chieh Liu[1], Yung-An Hsieh[2], Min-Hung Chen[2], C.-H. Huck Yang[2], J. Tegner[3], Y.-C. James Tsai[1,2]

[1]School of Civil and Environmental Engineering; [2]School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA, USA
[3]Living Systems Laboratory, KAUST, KSA

Speaker: Yi-Chieh Liu, Yung-An Hsieh

ICASSP 2020

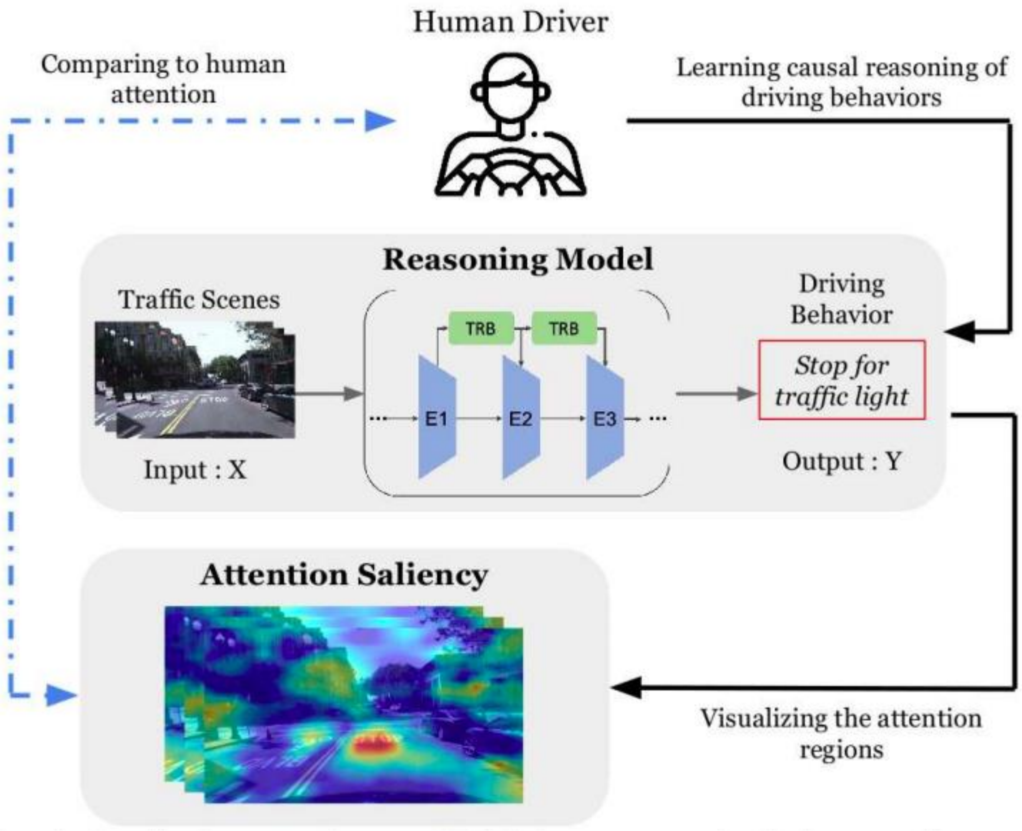
Outline

- Introduction
- Contributions
- Related Works
- Methodology
- Dataset
- Experimental Results
- Conclusions

Introduction

1. Motivation
    a. A reasoning model needs predicting actions based on human drivers performance.
    b. Attention saliency is required to improve the models on predicting the behaviors based on the correct reasons.
2. Video Recognition of Driving behavior
    a. Causal reasoning
    b. Spatial-temporal reasoning
3. Visual Explanation
    a. Filtering complex traffic information by attention saliency
    b. Recognizing actual cause of action

Robust Self-Driving System Architecture



Contributions

- The investigation of state-of-the-art 3D CNNs on the recognition of driving behaviors based on causal reasoning
- The introduction of the **Temporal Reasoning Block (TRB)** for improving the state-of-the-art models on classifying reasoning-based driving behaviors
- The proposition of a perturbation-based visual explanation method for spatial-temporal models, which enables the inspection of self-driving models

Related Work

- Self-Driving Behavior Recognition
  - As self-driving technology demonstrated incredible performance in both urban and off-road scenarios [1], the reasoning of self-driving behavior became a needed research problem
  - Prior efforts [2, 3, 4] formulate the behavior as a goal-oriented task, which is not sufficient to learn how humans drive and interact with traffic scenes
  - Driving behavior understanding could be performed by video recognition approaches: CRNN [5], C3D [6], I3D [7], 3DResNet [8]

Related Work

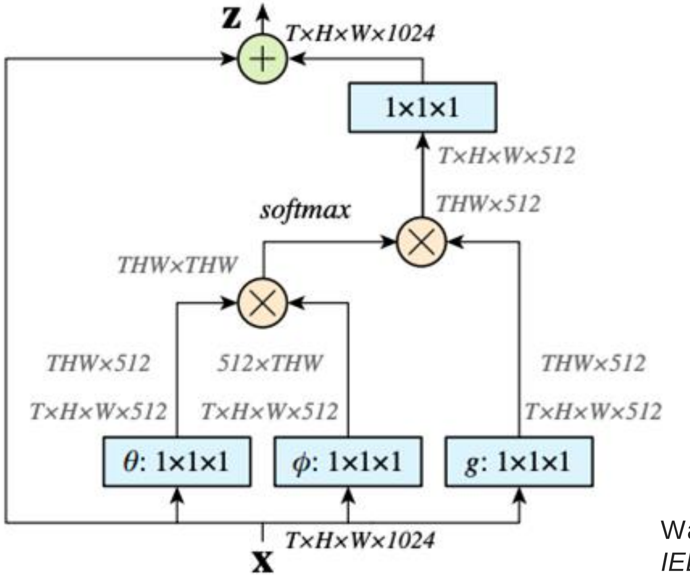
- Attention Models
  - Attention mechanisms have become a reliable method to capture global dependencies [9, 10]. Self-attention [11] represents the importance of different positions in a sequence
  - While self-attention has been applied to actions recognition tasks in video [12], the potential of self-attention have not been explored on the reasoning tasks of driving behaviors
- Visual Explanation of CNNs
  - Some explanation methods require accessing intermediate layers [13, 14] and/or architectural modification [15] of the CNNs
  - Other methods perform explanation by perturbing the input images [16, 17], which can be used on any kind of the model

Methodology

Non-local network

1. Inspired by non local neural network [12], we captured long-range dependencies to observe the cause of action through space-time features.

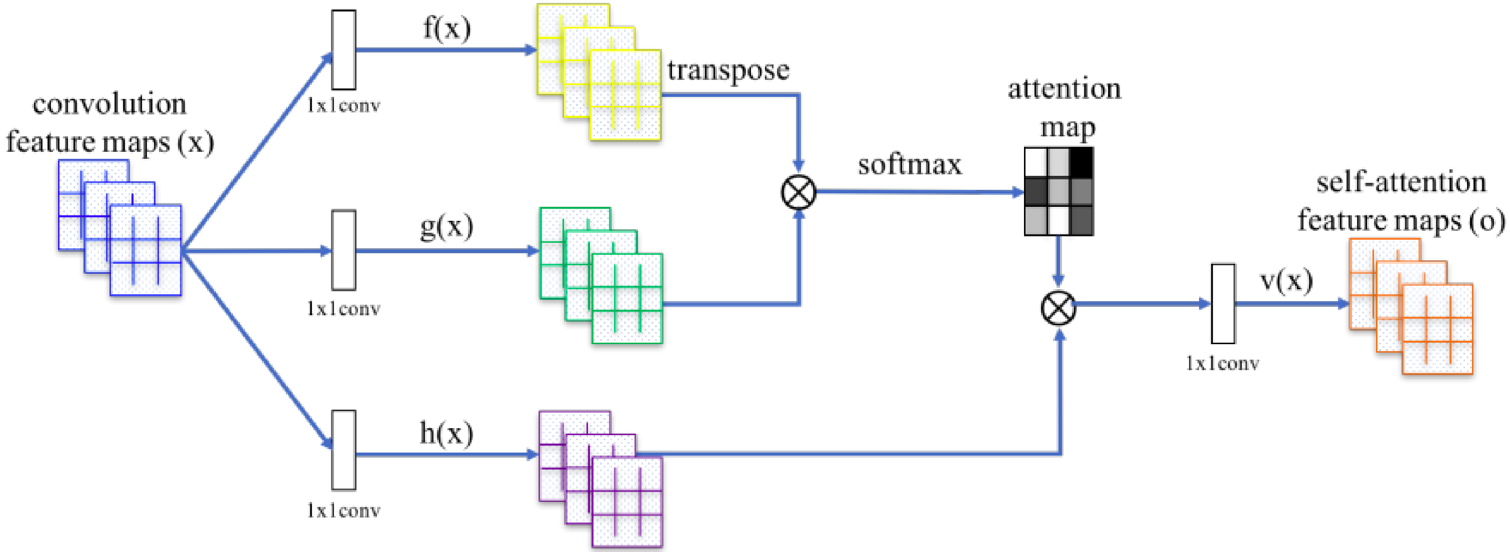
Non-Local block



Wang, Xiaolong, et al. "Non-local neural networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Methodology

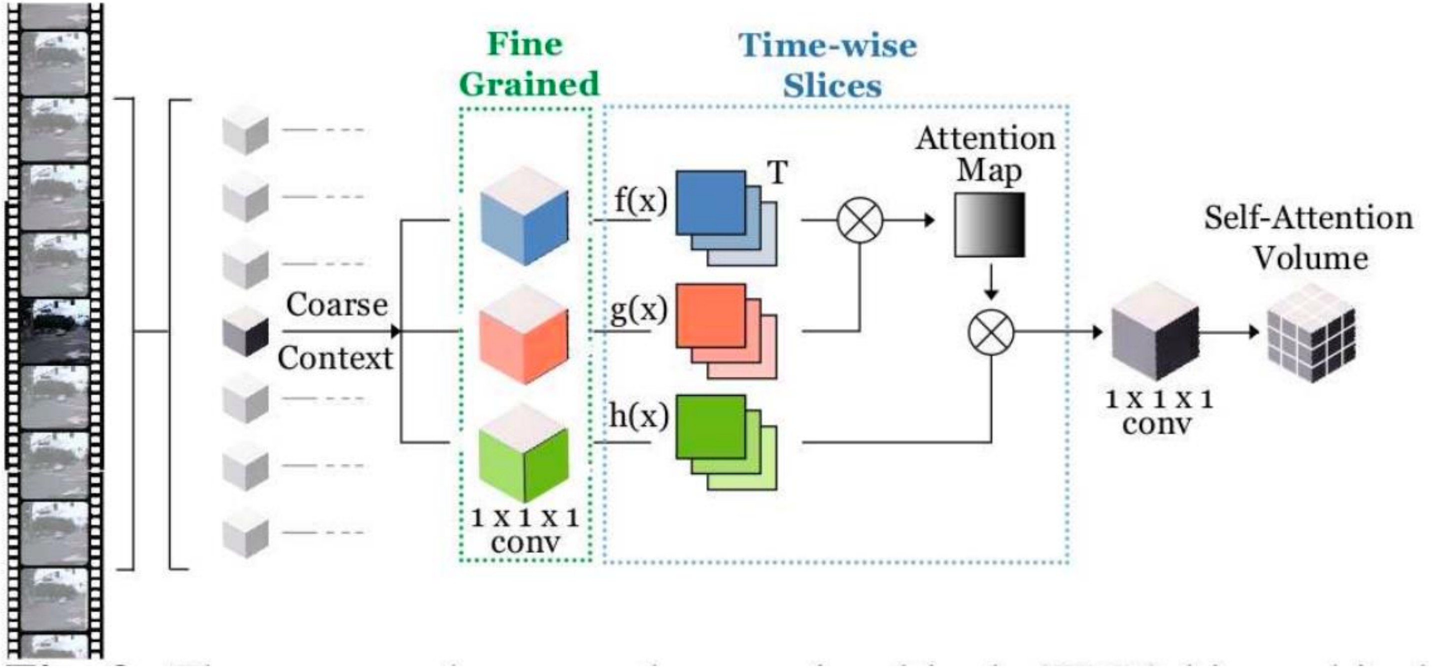
Self Attention Mechanism



Zhang, Han, et al. "Self-attention generative adversarial networks." *arXiv preprint arXiv:1805.08318* (2018).

Methodology

Temporal Reasoning Block (TRB)



Methodology

Temporal Reasoning Block (TRB)

1. 1 x 1 3D Convolution for fine grinded features
2. Temporal-aware self-attention map

1. Attention map for every frame → $\alpha_{j,i} = \frac{exp(s_{ij})}{\sum_{i=1}^{N} exp(s_{ij})}, s_{ij} = \boldsymbol{f}(\boldsymbol{x_i})^T \boldsymbol{g}(\boldsymbol{x_j})$

2. Dot product of spatial feature and attention map → $\boldsymbol{o_j} = \sum_{i=1}^{N} \alpha_{j,i} \boldsymbol{h}(\boldsymbol{x}),\ \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{W_h x}$

3. Stack along with time → $\boldsymbol{O_v} = \boldsymbol{Stack}\{\boldsymbol{o_t}\}, t = 1\ to\ T$

4. Gamma will be learnable parameter → $\boldsymbol{Y_i} = \gamma \boldsymbol{O_i} + \boldsymbol{x_i}$

Methodology

Perturbation-based Visual Explanation for Self-Driving Models

Based on [17], the explanation was done by finding the regions to perturb the original image which makes the classifier model to produce a minimal score on the target class. The example is as follows:

Ruth C Fong and Andrea Vedaldi, “Interpretable explanations of black boxes by meaningful perturbation,” *in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3429–3437.*

Methodology

Defining perturbation mask for single frame

$$[\Phi\{x_0; m\}](u) = m(u)x_0(u) + (1 - m(u))x_{p(u)}$$

To minimize the classification score of single frame, objective function:

$$\min_{m\in[0,1]^\Lambda} f_c(\Phi\{x_0; m\}) + \lambda_1\|1 - m\|_1 + \lambda_2 \sum_{u\in\Lambda} \|\nabla m(u)\|_\beta^\beta$$

Objective function expanding to both spatial and temporal dimensions

$$\min_{m\in[0,1]^{(\Lambda,T)}} f_c(\Phi\{x_0; m\}) + \lambda_1\|1 - m\|_1 +$$
$$\sum_{t\in T}\left(\lambda_s \sum_{u\in(\Lambda,t)} \|\nabla m(u,t)\|_\beta^\beta + \lambda_t\|\nabla m(:,t)\|_\beta^\beta\right)$$

Dataset

Honda Research Institute Driving Dataset (HDD) [18]

- Video clips with annotations of Stimulus-driven Action and Cause

Data splits	stop4light	stop4ped	stop4sign	stop4cong
Train	100	45	170	170
Validation	10	6	20	20
Test	13	10	30	30

Results - Driving Behavior Recognition

- The self-attention mechanism in TRB effectively helped the models to capture the global dependency within the videos.
- Also, TRB can be flexibly applied to different models of driving behavior recognition to provide improvement

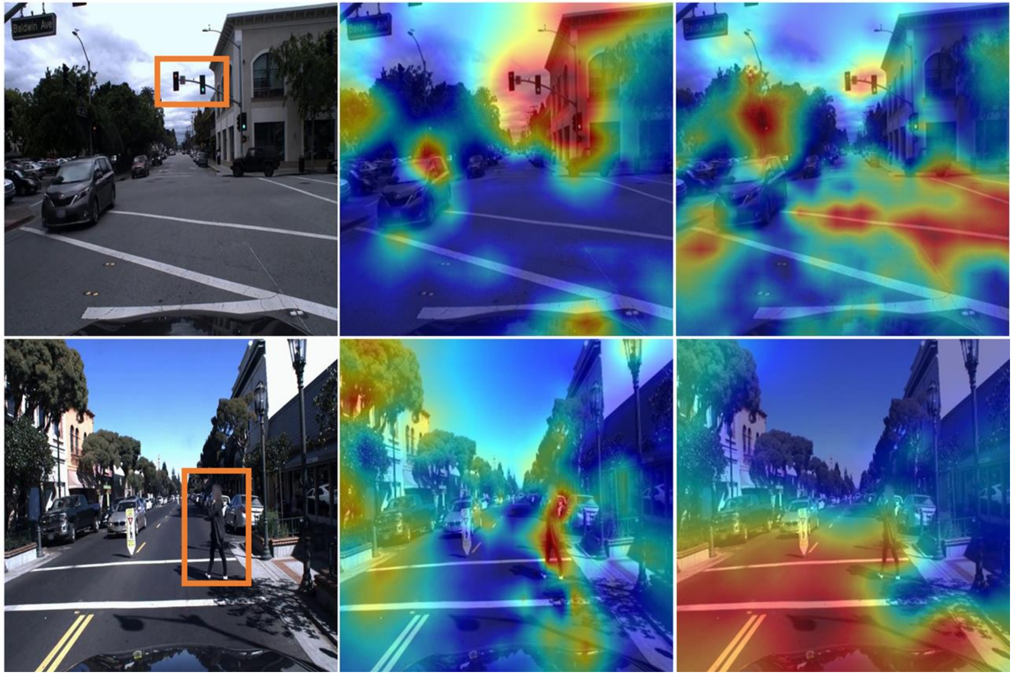
Model	Accuracy	Model	Accuracy
CRNN	73.49%	CRNN-TRB	**78.31%**
C3D	60.71%	C3D-TRB	**69.88%**
I3D	77.11%	I3D-TRB	**83.13%**
3DResnet	83.56%	3DResnet-TRB	**86.30%**

Results - Attention Saliency of Driving Behaviors

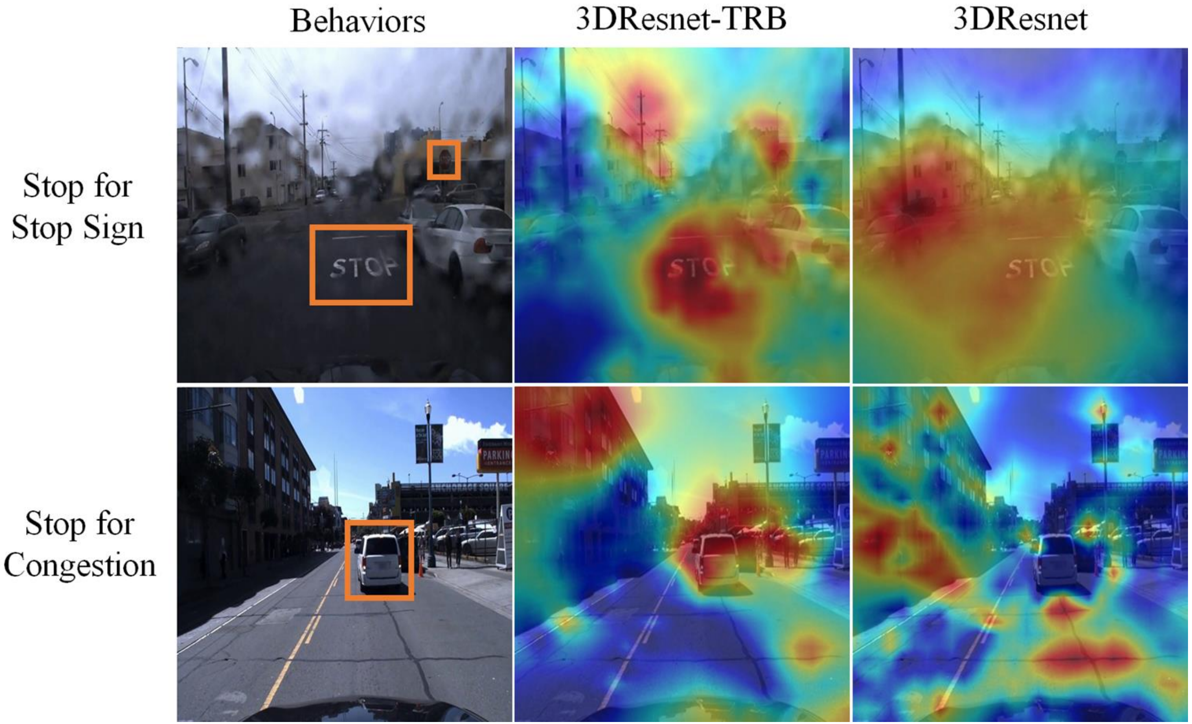
Behaviors | 3DResnet-TRB | 3DResnet

Stop for Traffic Light

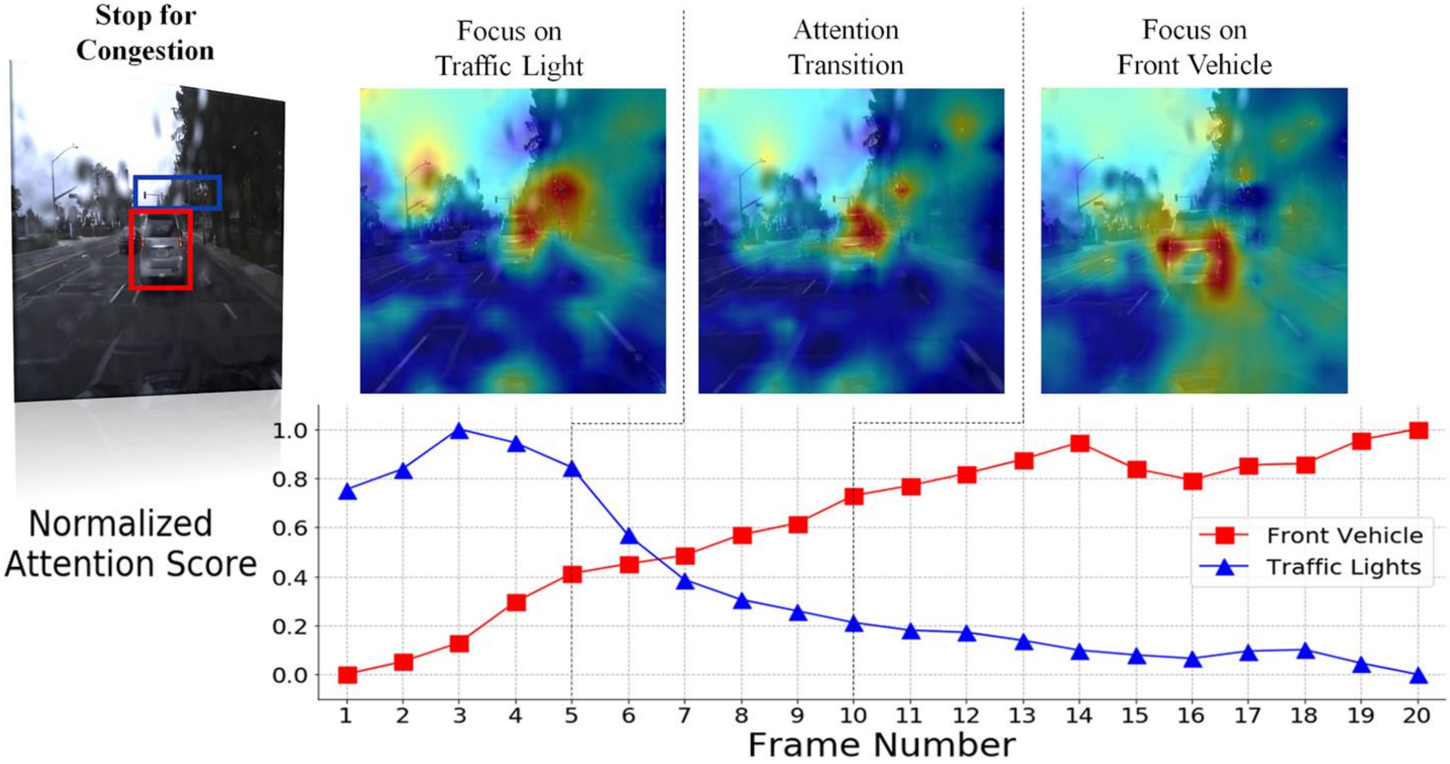
Stop for Pedestrian



Results - Attention Saliency of Driving Behaviors



Results - Attention Saliency of Driving Behaviors



Conclusions

- We proposed the **Temporal Reasoning Block (TRB)** to improve the performance of video recognition models on reasoning driving behaviors
- The TRB largely improved the performance of CRNN and 3D CNNs and we achieved the highest accuracy of 86.3% using the 3DResnet-TRB model
- The attention saliency, generated by the proposed perturbation-based visual explanation method, demonstrated that 3DResnet-TRB was able to focus on reasonable objects when classifying driving behaviors

References

[1] Martin Buehler, Karl Iagnemma, and Sanjiv Singh, The DARPA Urban Challenge: Autonomous Vehicles in City Traffic, Springer Publishing Company, Incorporated, 1st edition, 2009.

[2] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al., “End to end learning for self-driving cars,” arXiv preprint arXiv:1604.07316, 2016.

[3] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell, “End-to-end learning of driving models from large-scale video datasets,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2174–2182.

[4] Dean A. Pomerleau, “Advances in neural information processing systems 1,” chapter ALVINN: An Autonomous Land Vehicle in a Neural Network, pp. 305–313. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1989.

[5] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici,“Beyond short snippets: Deep networks for video classification,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4694–4702.

[6] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.

[7] Joao Carreira and Andrew Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

References

[8] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh, “Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?,” in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6546–6555.

[9] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv preprint arXiv:1409.0473, 2014.

[10] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in International conference on machine learning, 2015, pp. 2048–2057.

[11] Ankur P Parikh, Oscar T¨ackstr¨om, Dipanjan Das, and Jakob Uszkoreit, “A decomposable attention model for natural language inference,” arXiv preprint arXiv:1606.01933, 2016.

[12] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, “Non-local neural networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.

[13] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, “Learning deep features for discriminative localization,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.

[14] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, “Gradcam: Visual explanations from deep networks via gradientbased localization,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

References

[15] Matthew D Zeiler and Rob Fergus, “Visualizing and understanding convolutional networks,” in European conference on computer vision. Springer, 2014, pp. 818–833.

[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, “Why should i trust you?: Explaining the predictions of any classifier,” in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016, pp. 1135–1144.

[17] Ruth C Fong and Andrea Vedaldi, “Interpretable explanations of black boxes by meaningful perturbation,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3429–3437.

[18] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko, “Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7699–7707.