 https://doi.org/10.3311/PPtr.15921

Periodica Polytechnica Transportation Engineering, 48(4), pp. 334–340, 2020

Towards Reliable Multisensory Perception and Its Automotive Applications

András Rövid[1*], Viktor Remeli[1], Norbert Paufler[1], Henrietta Lengyel[1], Máté Zöldy[1], Zsolt Szalay[1]

[1] Department of Automotive Technologies, Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, H-1521 Budapest, P. O. B. 91, Hungary
[*] Corresponding author, e-mail: andras.rovid@gjt.bme.hu

Received: 12 March 2020, Accepted: 12 March 2020, Published online: 07 July 2020

Abstract

Autonomous driving poses numerous challenging problems, one of which is perceiving and understanding the environment. Since self-driving is safety critical and many actions taken during driving rely on the outcome of various perception algorithms (for instance all traffic participants and infrastructural objects in the vehicle's surroundings must reliably be recognized and localized), thus the perception might be considered as one of the most critical subsystems in an autonomous vehicle. Although the perception itself might further be decomposed into various sub-problems, such as object detection, lane detection, traffic sign detection, environment modeling, etc. In this paper the focus is on fusion models in general (giving support for multisensory data processing) and some related automotive applications such as object detection, traffic sign recognition, end-to-end driving models and an example of taking decisions in multi-criterial traffic situations that are complex for both human drivers and for the self-driving vehicles as well.

Keywords

raw sensor fusion, perception, end-to-end models, handling traffic scenarios

1 Introduction

In the autonomous vehicle technology evolved today, reliability and response time due to the autonomous system are better than humans, the decision-making research more inclined to focus on the vehicle, rather than a human driver (Mersky and Samaras, 2016). Modern autonomous vehicles can detect the surrounding objects and classify the surrounding information by sensing the environment in real-time, so that the appropriate navigation path can be confirmed while observing the corresponding traffic rules (Bimbraw, 2015).

Sensing and understanding vehicle surroundings is one of the most crucial factors since any subsequent action taken is strongly dependent on how the scene is interpreted, what type of participants there are present, where they are located, what their intention is, etc.

In order to make self-driving safe and reliable all this information must be extracted with high confidence and more importantly must be accurate. Any misdetected or misclassified object may harm the self-driving safety. In order to increase the reliability of perception the utilization of various types of sensors proved to be a promising direction.

In multitype sensor cases the overall joint sensing capability of the self-driving vehicle covers a wider range of weather and traffic conditions in general, furthermore the sensor redundancy is also a significant advantage in case of sensor failure or damage.

Up to now, we have considered the perception as a separate layer providing information about the surroundings of the vehicle towards different but separate layers in the overall system architecture such as trajectory planning, decision making, etc. Besides this layered architecture, it is worth to mention also the case where the mentioned layers are jointly represented in form of a single end-to-end neural network model realizing various driver assistance or self-driving functions on purely neural basis.

In this paper let us give a brief insight into both types of previously mentioned architectures (including our related findings from real experiments), namely the case when the perception stands for a separate layer in the overall system architecture and the case of the end-to-end approach. In relation to the first one, our proposed fusion model will be described, which utilizes camera image and

Cite this article as: Rövid, A., Remeli, V., Paufler, N., Lengyel, H., Zöldy, M., Szalay, Zs. (2020) "Towards Reliable Multisensory Perception and Its Automotive Applications", Periodica Polytechnica Transportation Engineering, 48(4), pp. 334–340. https://doi.org/10.3311/PPtr.15921

LiDAR pointcloud to detect traffic participants in the surroundings of the vehicle.

As a further representative example of a safety critical perception problem the traffic sign detection is going to be pointed out including also some findings related to adversarial attacks and the possibilities of verification of the underlying neural network models.

End-to-end models will be discussed trough a vehicle steering example together with the experimental results.

2 Raw sensor fusion models for 3D object detection

Multisensory perception plays crucial role in safety critical systems. Depending on the level at which the fusion is performed the sensory data might be processed individually or jointly. Here we are focusing on raw sensor fusion, namely when the sensory data are fused at low levels and thus utilizing also the synergies between individual sensors.

Let us briefly point out some relevant state of the art results achieved in 3D object detection for autonomous driving (trained, validated and tested on KITTI dataset (Geiger et al., 2013). Ku et al. (2017) have proposed an Aggregate View Object Detection network (AVOD). The authors use LIDAR pointclouds and RGB images to generate features that are shared by two subnetworks. Liang et al. (2018) also rely on both types of sensors, namely on LiDAR and cameras. With the help of a so called "continuous fusion layer" a dense Birds Eye View (BEV) feature map is created and fused with the BEV feature map extracted from LiDAR. Authors in (Xu et al., 2017) have proposed another approach to fuse LiDAR and camera data. Their model is based on the so called PointNet (Qi et al., 2017a), architecture able to handle unorganized raw pointcloud data. They combine pointcloud features learnt by the PointNet with high-level features extracted by a ResNet based CNN (He et al., 2015). Qi et al. (2017b) proposed a model for 3D object detection based on pointcloud data falling inside a frustum defined by a 2D bounding box encapsulating a given object in the camera image. The images are used by the network to determine the 2D bounding boxes and thus the corresponding frustum, however the pointcloud itself is not fused with image data. The approach proposed in this paper extends the FPointNet (Qi et al., 2017b) model by assigning local - semantically stronger - image features from high resolution feature maps to each point in the LiDAR pointcloud. Here let us focus on the problem of low-level camera and sparse LiDAR data fusion, which aims to improve 3D detection of cars, pedestrians and cyclists for autonomous driving related applications where safety considerations play crucial role. We built upon an existing state of the art neural network architecture that considers camera and LiDAR data only in separate, sequential phases of processing. We improved the concept by training the segmentation and 3D box estimation networks (see Fig. 1) to process raw signals from different types of sensors jointly which was achieved by projecting the LiDAR pointcloud onto the camera image plane and augmenting the points with a corresponding image feature vector taken from selected layers of the 2D detector. Depending on the depth of chosen layers the image features taken might have different levels of abstraction. According to our experimental results, we have shown that more reliable detection of distant targets that are characterized by very sparse LiDAR measurements is possible (see Fig. 2). In conclusion, we have seen that introducing lower levels of fusion into existing perception architectures for autonomous vehicles might be beneficial in accomplishing specific safety-critical tasks like the detection of distant or heavily obscured objects. We intend to explore further possibilities for improving raw fusion approaches in order to establish the extent of possible benefits of leveraging inter-sensor synergies.

3 Adversarial vulnerability of deep learning perception systems

Today deep learning systems are becoming ubiquitous and often even key components for autonomous driving functionalities, especially in the perception layer. Such deep neural networks operate on very high dimensional input domains (e.g. visual images define a pixel space with thousands or millions of dimensions) and can have millions of parameters (Krizhevsky et al., 2017). The most significant consequence is that now we are able to capture system models of truly super-human complexity – i.e. the machines now routinely learn models that human domain experts have no chance of understanding – e.g. cognitive models such as visual perception of traffic participants (LeCun et al., 2015). This lack of understanding is worrisome in safety critical applications like autonomous vehicles, especially in the light of the 2013 discovery of so called "adversarial examples" (Szegedy et al., 2013). Even though neural networks can achieve impressive, even super-human accuracy on various tasks, e.g. perception, the excellent statistical performance bears no guarantees for any individual case. In the last years it has become very clear that even if a deep learning system is often *right*, it is not right *for the right reasons*. As an unfortunate consequence,

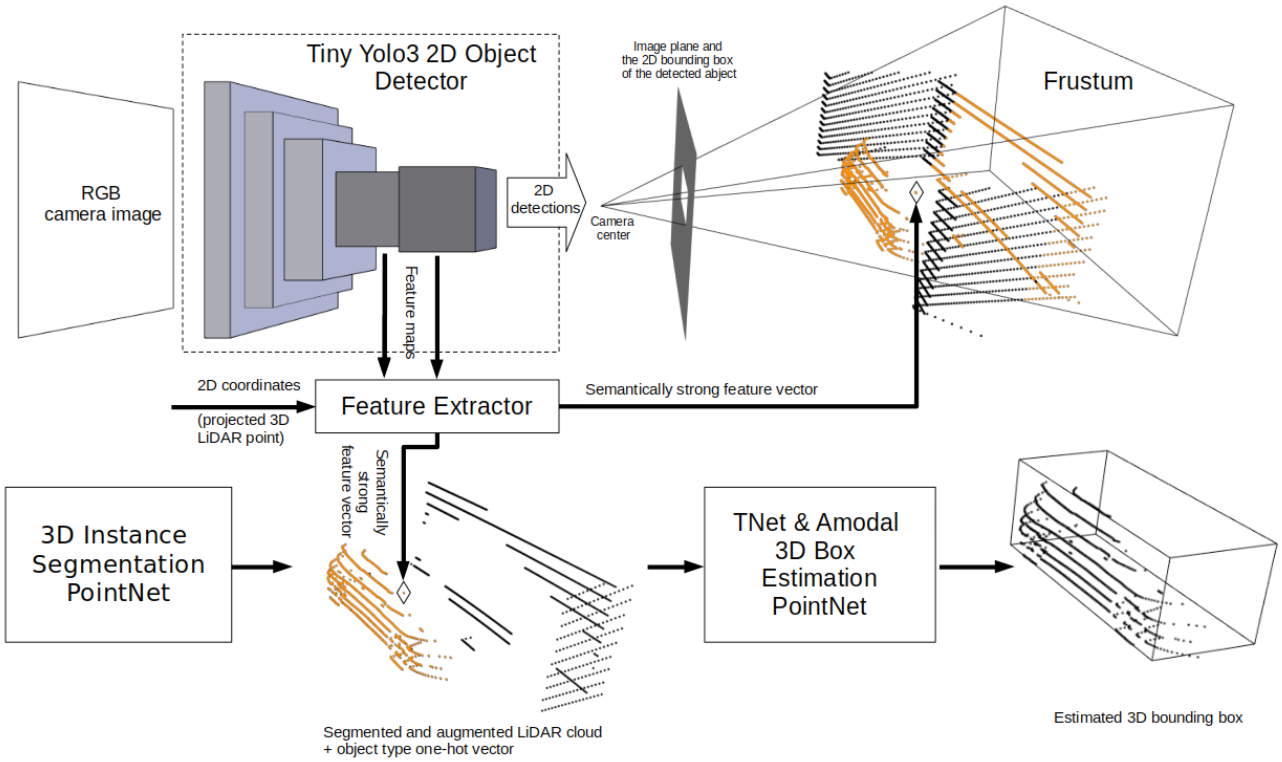


Fig. 1 Architecture of the object detector (Rövid and Remeli, 2019).

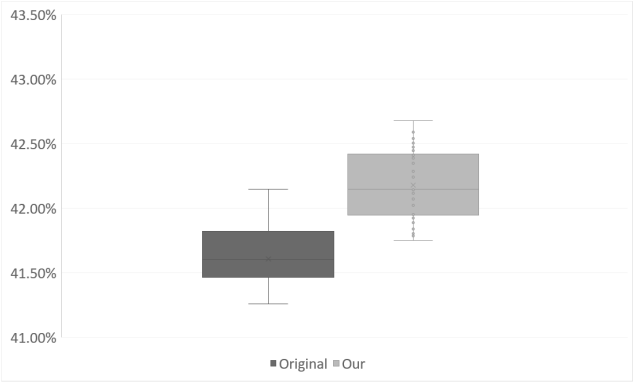
in the rare cases when the network is wrong, its prediction can be surprisingly off, and those cases can be – to the human observer – indistinguishable from non-challenging, entirely trivial tasks. Even worse, it seems that these cases – the so called "adversarial examples" – may not be so rare after all, as the environment can be easily manipulated by a malicious attacker in such a way that the attack is invisible or camouflaged to the human eye yet it reliably makes the deep learning system misbehave, often in exactly the way the attacker wants it to. This kind of danger is of course unacceptable in a safety-critical system and new ways of mitigating the safety and security risks by validating, verifying and understanding the inner models of deep neural networks are urgently necessary.

3.1 Adversarial traffic sign stickers

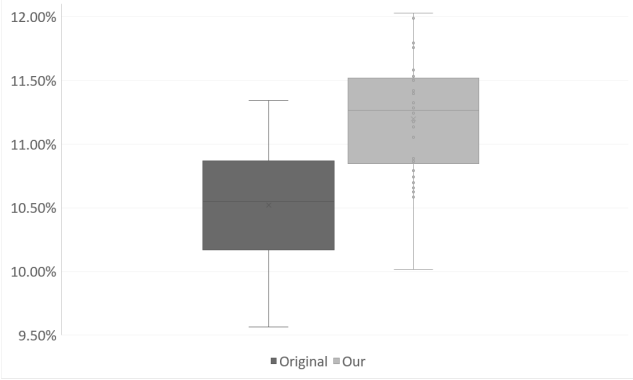
Detection and classification of traffic signs is a popular and well-researched example of reliance on environment perception in highly automated driving. Clearly the misidentification of a traffic sign can have huge costs in terms of safety. The road sign recognition problem is also a good choice from the technical perspective as it is simple enough that classical computer vision methods achieve industrially acceptable performance while also being complex enough that employing deep learning systems yields tangible improvements.

In our recent work (Lengyel et al, 2019) we call attention to the easily exploitable vulnerabilities of deep learning technologies by developing home printable and readily deployable stickers that appear to be artwork/vandalism similar to what is frequently encountered on traffic signs in many places. Our stickers contain specific adversarial patterns however, nondescript to the human eye but highly misleading to a neural network. A nefarious lay person can easily affix such a sticker to a road sign and thus effectively change its meaning in a targeted way. The especially grave danger arises from the fact that the human observers (drivers, road authorities, etc.) will not recognize that an attack was deployed and will therefore fail to counter it.

The adversarial patches we developed work in the physical world and are fairly robust to perturbations arising from printer quality, environmental conditions, camera angles, noise and resolution. Our stickers are unique since they are also insensitive to deployment inaccuracy with a generous margin of about ±10 cm, making technical know-how entirely unnecessary and requiring no special access to the attacked infrastructure either. An example is shown in Fig. 3 where our patch makes the algorithm misclassify the 30 km/h speed limit as 120 km/h, while a similarly placed random patch does not affect the algorithm at all.



(a)



(b)

Fig. 2 Results of the model evaluation for sparse 3D pointclouds consisting of 8 points. The figures show the detection precision (Positive Predictive Value (PPV)) measured in two different cases: firstly, the original architecture where the cloud is represented as [*X*, *Y*, *Z*] and secondly our modified architecture that augments the original [*X*, *Y*, *Z*] point coordinates with 29 semantically strong image features. (a) 3D IoU (b) 3D IoU ≥ 70

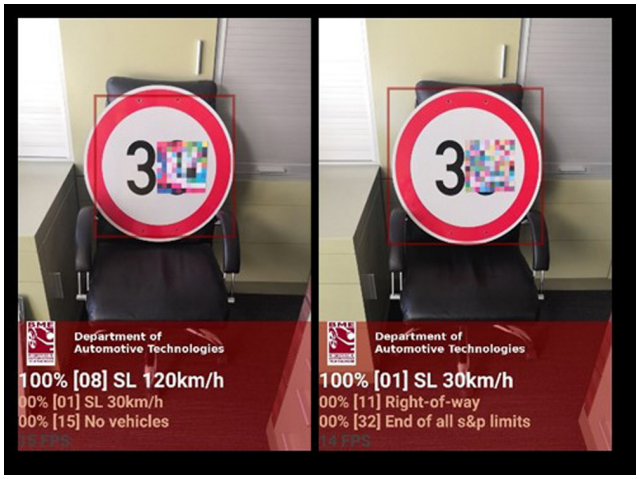


Fig. 3 Left: our adversarial sticker causes high confidence targeted misidentification in a deep learning classifier. Right: a similarly placed random sticker causes no divergence from the correct class. (SL stands for Speed Limit)

The threat posed by adversarial examples is especially severe since they were shown to generalize across different neural architectures and across networks trained on different training sets (Szegedy et al., 2013). The carry-over effect is somewhat diminished but still significant enough to enable black-box attacks with no query access to the attacked system.

3.2 Verifying an image classification network

After demonstrating the practical danger of adversarial tampering in physical environments, we are interested in knowing what kinds of guarantees a deep learning system can make, if any. The robustness of a deep neural network can be evaluated via methods shown in previous research (Katz et al., 2017). However, in practice this is computationally feasible only as a statistical estimate of robustness from a sample taken at a certain number of points in input space: system behavior is guaranteed in their ε-radius vicinity only. It is not trivial to verify guaranteed properties for substantial regions of the input space. An even more interesting question is whether useful and humanly interpretable specifications of input regions that we wish to regulate – beyond simplistic ones like Euclidean ε-radius – can be formulated at all.

In Remeli et al. (2019) we show that ReLU- and softmax-activated networks can be transcribed as a mixed-integer linear systems. In this case, specification regions defined as linear combinations of the inputs (and similarly defined forbidden regions in the output) can serve as the basis for a mixed-integer linear program formulation. If the program has a feasible solution, that solution constitutes an adversarial (or *deviant*) example that does not conform to the specification. If the program has no solution however, that proves that the network fulfills the specification.

With the above method we analyzed a simple 555-neuron multi-layer neural network that performs traffic sign classification with 88 % accuracy on the GTSRB benchmark. We were able to certify a guaranteed property stating that the network would never classify the image as a STOP sign if the red color content in the image was extremely low, making it entirely resistant to this specific type of adversarial attack. Since this was a naively trained network, we were not expecting practically significant properties, but were rather able to prove the presence of weak properties in a human-interpretable way. For technical details, please refer to (Remeli et al., 2019).

4 Perception based end-to-end models

In case of an end-to-end model the entire process from perception to actuation involves a properly trained, single convolutional neural network without any modularization whereas the neural models for perception (see Section 2) are providing information about the environment to other modules such as maneuver planning, control, etc.

The aim of Section 4 is to show the end-to-end concept and related problems through a vehicle steering example. In Fig. 4 the training set creation, the training itself and validation are illustrated together with the underlying diagrams. During these experiments we relied on a state-of art CNN architecture published in (Bojarski et al., 2016) and investigated how different driving behaviors might be achieved, what factors must be considered during the training set preparation. The ground truth data for training was generated with the Carla software (see Fig. 5), which is an open-source simulator for autonomous driving system development. The simulator gives the ability to add and control multiple vehicles to create traffic scenarios, attach different type of sensors to the vehicles.

During the experiments two different sets have been generated for training, i.e. firstly a spline-based track was used as a reference trajectory along which the camera images for training have been acquired together with the corresponding steering commands.

In the second case the steering of the vehicle was controlled by a Model Predictive Controller (MPC) and the obtained path was used to generate camera images together with the corresponding steering angles for training. In case of the end-to-end model (trained on the latter dataset) the vehicle started to drift off the path after a certain amount of time and was unable to recover itself afterwards. In order to avoid this effect, the training dataset had to be extended by additional images representing "faulty" states and the corresponding steering angles. The trajectory including "faulty" states have been generated artificially by rotating and translating the vehicle to critical poses and then letting the MPC controller to follow the reference path (see Figs. 6 and 7).

Another important consideration during the preparation of the training dataset was to use different augmentations to handle problems associated with the tilt of the horizon in the camera images. As the vehicle drives along a curved path the horizon tilts in the camera image according to the curvature of the path and speed of the vehicle. In order to force the network to associate the characteristics of the path with the corresponding steering angles the training dataset

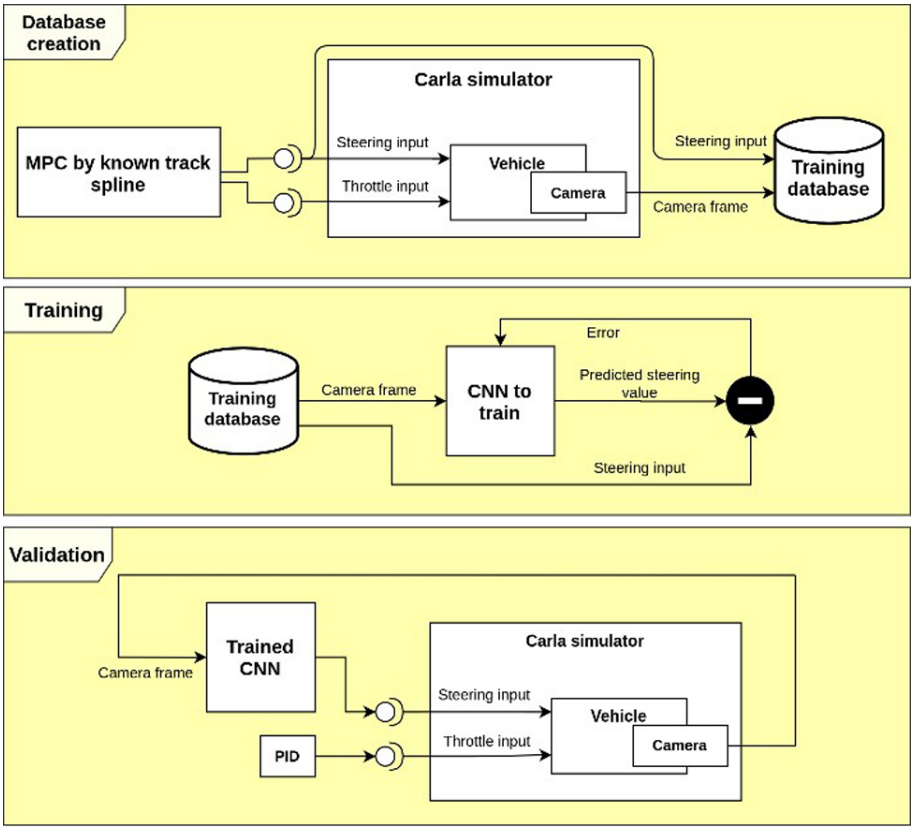


Fig. 4 Overall diagram of project's structure

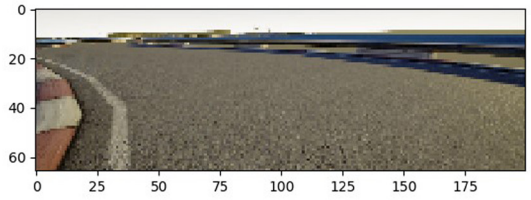


Fig. 5 CNN input frame example (size: 66 × 200 pixels)

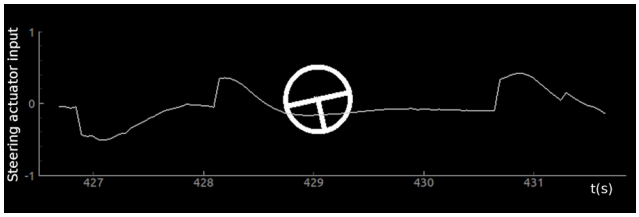


Fig. 6 The figure shows the steering control value on the *y* axis, which was estimated by the MPC controller. At timestamp 428 we can see an example of faulty state injection followed by the recovery from that state by MPC.

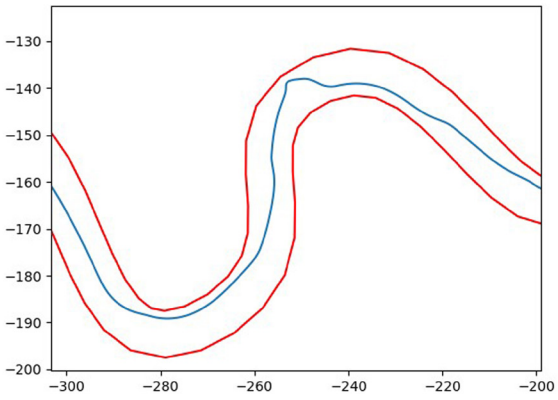


Fig. 7 Artificially disturbed trajectory (blue line) of the vehicle.

had to be augmented by rotation about the image center by random angles being within a certain range (see Fig. 8). In addition to rotations some additional augmentations have been applied, as well such as brightness modification, vertical translation and random shadow. By using the mentioned augmentations significant improvement was achieved, the network was able to steer the vehicle along a trajectory similar to that generated by the MPC even in case of the validation track, which was not "seen" by the network during the training phase (see Fig. 9).

For testing the robustness of the network the camera pose was modified and the model was evaluated accordingly. According to the experiments the network is robust to camera rotations and also to change of the camera FOV in a reasonable range.

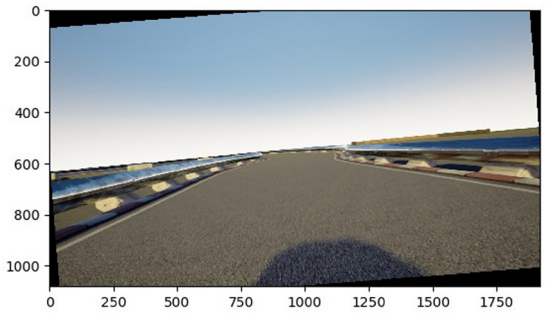


Fig. 8 An example image for the random rotation augmentation

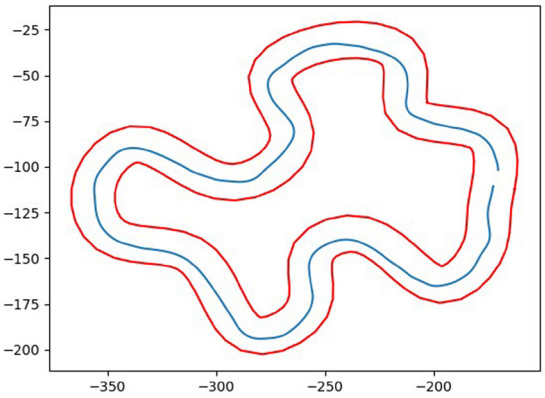


Fig. 9 The trajectory of the vehicle (blue line) on validation track, steering by the CNN network output based on a dashboard camera

5 Conclusion

In this paper we have shown an overview on AI perception related problems such as how lower fusion might be performed, pointed out the threat posed by adversarial examples which is especially severe since they were shown to generalize across different neural architectures. On top of that we have touched the topic on how the underlying neural networks might be verified. We have also given an example to end-to-end steering control and pointed out some problems which have to be considered in order for the network to work properly. Our next step is to integrate the proposed low-level fusion concept into end-to-end models, as well, in order to improve their robustness and reliability.

Acknowledgement

The research reported in this paper was supported by the Higher Education Excellence Program in the frame of Artificial Intelligence research area of Budapest University of Technology and Economics (BME FIKP-MI/FM).

References

Bimbraw, K. (2015) "Autonomous Cars: Past, Present and Future - A Review of the Developments in the Last Century, the Present Scenario and the Expected Future of Autonomous Vehicle Technology", In: Proceedings of the 12th International Conference on Informatics in Control, Automation and Robotics, Colmar, Alsace, France, pp. 191–198. https://doi.org/10.5220/0005540501910198

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, D. L., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., Zieba, K. (2016) "End to End Learning for Self-Driving Cars", Computer Science: Computer Vision and Pattern Recognition, pp. 1–9. [online] Available at: https://arxiv.org/abs/1604.07316 [Accessed: 11 March 2020]

Geiger, A., Lenz, P., Stiller, C., Urtasun, R. (2013) "Vision meets Robotics: The KITTI Dataset", International Journal of Robotics Research, 32(11), pp. 1231–1237. [online] Available at: http://www.cvlibs.net/datasets/kitti [Accessed: 11 March 2020]

He, K., Zhang, X., Ren, S., Sun, J. (2015) "Deep Residual Learning for Image Recognition", Computer Science: Computer Vision and Pattern Recognition, pp. 1–12. [online] Available at: http://arxiv.org/abs/1512.03385 [Accessed: 11 March 2020]

Katz, G., Barrett, C., Dill, D. L., Julian, K., Kochenderfer, M. J. (2017) "Towards Proving the Adversarial Robustness of Deep Neural Networks", Electronic Proceedings in Theoretical Computer Science, 257, pp. 19–26. https://doi.org/10.4204/EPTCS.257.3

Krizhevsky, A., Sutskever, I., Hinton, G. E. (2017) "ImageNet classification with deep convolutional neural networks", Communications of the ACM, 60(6), pp. 84–90. https://doi.org/10.1145/3065386

Ku, J., Mozifian, M., Lee, J., Harakeh, A., Waslander, S. (2017) "Joint 3D Proposal Generation and Object Detection from View Aggregation", Computer Science: Computer Vision and Pattern Recognition, pp. 1–8. [online] Available at: http://arxiv.org/abs/1712.02294 [Accessed: 11 March 2020]

LeCun, Y., Bengio, Y., Hinton, G. (2015) "Deep learning", Nature, 521(7553), pp. 436–444. https://doi.org/10.1038/nature14539

Lengyel, H., Remeli, V., Szalay, Z. (2019) "Easily Deployed Stickers Could Disrupt Traffic Sign Recognition", Perner's Contacts, (Special Issue) 2(19), pp. 156–163. [online] Available at: http://pernerscontacts.upce.cz/SI2/Lengyel1.pdf [Accessed: 11 March 2020]

Liang, M., Yang, B., Wang, S., Urtasun, R (2018) "Deep Continuous Fusion for Multi-Sensor 3D Object Detection", In: Ferrari V., Hebert M., Sminchisescu C., Weiss Y. (eds.) Computer Vision – ECCV 2018, Lecture Notes in Computer Science, Springer, Cham, Switzerland, pp. 663–678. https://doi.org/10.1007/978-3-030-01270-0_39

Mersky, A. C., Samaras, C. (2016) "Fuel economy testing of autonomous vehicles", Transportation Research Part C: Emerging Technologies, 65, pp. 31–48. https://doi.org/10.1016/j.trc.2016.01.001

Qi, C. R., Su, H., Mo, K., Guibas, L. J. (2017a) "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation", In: Proceedings 30th IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, pp. 77–85. [online] Available at: http://arxiv.org/abs/1612.00593 [Accessed: 11 March 2020]

Qi, C. R., Liu, W., Wu, C., Su, H., Guibas, L. J. (2017b) "Frustum PointNets for 3D Object Detection from RGB-D Data", Computer Science: Computer Vision and Pattern Recognition, pp. 1–19. [online] Available at: http://arxiv.org/abs/1711.08488 [Accessed: 11 March 2020]

Remeli, V., Morapitiye, S., Rövid, A., Szalay, Z. (2019) "Towards Verifiable Specifications for Neural Networks in Autonomous Driving", In: IEEE 19th International Symposium on Computational Intelligence and Informatics, Szeged, Hungary, 2019.

Rövid, A., Remeli, V. (2019) "Towards Raw Sensor Fusion in 3D Object Detection", In: 2019 IEEE 17th World Symposium on Applied Machine Intelligence and Informatics (SAMI), Herlany, Slovakia, pp. 293–298. https://doi.org/10.1109/SAMI.2019.8782779

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R. (2013) "Intriguing properties of neural networks", Computer Science: Computer Vision and Pattern Recognition, [online] Available at: https://arxiv.org/abs/1312.6199 [Accessed: 11 March 2020]

Xu, D., Anguelov, D., Jain, A. (2017) "PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation", Computer Science: Computer Vision and Pattern Recognition, pp. 1–11. [online] Available at: http://arxiv.org/abs/1711.10871 [Accessed: 11 March 2020]