Combining Multiple Modalities with Perceiver in Imitation-based Urban Driving

Shubham Juneja
Institute of Data Science &
Digital Technologies
Vilnius University
Vilnius, Lithuania
Email: shubham.juneja@mif.stud.vu.lt

Virginijus Marcinkevičius
Institute of Data Science &
Digital Technologies
Vilnius University
Vilnius, Lithuania
Email: virginijus.marcinkevicius@mif.vu.lt

Povilas Daniušis
Department of Business
Technologies & Entrepreneurship
Vilnius Gediminas Technical University
Vilnius, Lithuania
Email: povilas.daniusis@vgtu.lt

***Abstract*—Traditional autonomous driving methods have relied on multiple sensor inputs for their success in decision making. Meanwhile, these methods require greater engineering effort as they consist of multiple modules than end-to-end methods which learn from data. In comparison, end-to-end methods rely on only a single modality and lack the ability to thoroughly generalise to new environments compared to traditional approaches. To enhance the current state-of-the-art methods, we propose using additional environmental information into an end-to-end learned method by employing the Perceiver architecture. The proposed technique aims to use more than one modality by fusing sensor data into a learner to generalise better in urban environments.**

***Keywords*—imitation learning; urban driving; perceiver.**

I. Introduction

With the rise in popularity and demand for autonomy, research on autonomous driving has been at the forefront. The methods from the field not only contribute to transportation but also to the area of robotics [1]. While being vital to more than one industry, the task of driving autonomously in urban environments remains in the phase of research due to the high complexity of the problem and issues it faces, such as difficulties in generalising to unseen environments.

The current state-of-the-art autonomous driving systems are either based on the traditional autonomous vehicle pipelines using a modular approach where the system is divided into modules with multiple sensors and algorithms [2] or based on approaches that learn driving end-to-end directly from data [3]–[5]. Modular approaches leverage the presence of multiple sensors by fusing information to capture various characteristics from the surrounding and have been approaching human-level performance (e.g. Tesla Autopilot system) [6]. Meanwhile, they suffer in terms of the engineering effort required to tune each of the modules. In comparison, end-to-end learned methods thrive on requiring barely any tuning but prominently depend on a single front-facing camera for sensor input. This trend is dominant across recent end-to-end learning techniques, be it imitation learning or reinforcement learning, along with the disadvantage of showing generalisation to new environments to be a complex problem.

With 2D image data being the primary modality in end-to-end techniques [3]–[5][7], recent methods leverage convolutional neural networks (CNNs) as a candidate learner, which introduces a prerequisite of additional modifications to the architecture when involving different input configurations [8][9]. The recently proposed Perceiver architecture [10] attempts on employing different input data modalities into a single architecture. It is designed to work with arbitrary input configurations of different modalities and to efficiently handle high dimensional inputs. Various research results have shown the advantage of such data fusion across systems [6][11].

Considering the limitations of the current state-of-the-art, we make the contribution of proposing to fuse the front-facing camera data with an additional perception stream (using a LIDAR sensor) into a learner based on Perceiver architecture to learn the skill of urban driving using imitation learning. Further, we propose evaluating the learned method with the CARLA [12] and NoCrash [13] benchmarks.

The rest of the paper is organised as follows. Section II describes literature concerning urban driving and learning methods in the area; section III describes the proposed approach and the incipient implementation details; section IV concludes this idea paper.

II. Related Work

End-to-end urban driving methods, be it based on imitation learning or reinforcement learning, have been relying on 2D images for making driving decisions, which can be mainly due to the rich data images provide [6]. Conditional affordance learning [14] is a recent method which predicts the affordances with the use of a CNN from images on a vehicle and learns to drive with the use of imitation learning. Methods such as affordance based reinforcement learning [3] and implicit affordances learning [5], attempt to do a similar job but use reinforcement learning as the learning method. Learning by cheating method [4] uses a CNN to project feature maps from which it learns to drive using imitation learning. However, the hardware used on vehicles is capable of having more than one sensor to provide rich data. There are also methods which use some speed measurement along with image data [7][13]. Though this data might not be rich and it is integrated by the method of concatenation, the technique may not promise to utilise the data in the selected deep learning architecture.

The Perceiver architecture is capable of working with arbitrary configurations of inputs by using multiple transformer units. Also, Perceiver's inputs can be from different modalities, and the architecture does not require modifications, as opposed to CNNs, which require architectural adjustments. The architecture proves to be capable of dealing with large

ISBN: 978-1-61208-875-4

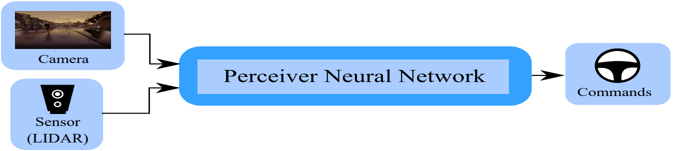

Fig. 1. Overview of the proposed method.

size and multimodal data format at once, and hence shows the capacity to be suitable to deal with images along with LIDAR data.

Reinforcement learning methods applied to the problem of driving in urban areas are quite a recent success and have only displayed results in simulated environments [3][5]. This leaves out a question if it is possible to hold up the skill level when deployed in the real world. In contrast, imitation learning methods have a long history of being applied in real world environments [7][15][16]. This motivates us to lean towards imitation learning rather than reinforcement learning.

III. Method and experiments

A. Approach

Our approach is to learn the skill of urban driving through observing two sensors in order to take advantage of data fusion, as shown in Figure 1. We plan to do so with the Perceiver architecture instead of a CNN, as the Perceiver architecture does not need any modifications to integrate multiple modalities into a single learner. For input sensors, we initially plan the use of a camera and a LIDAR sensor as such sensors show the capability of capturing rich environmental information and are widely used in autonomous navigation research.

B. Environment and Benchmarks

For the environment, we choose the CARLA simulator[12] over other possible options as it makes it simpler to compare results to the state-of-the-art methods since the widely used benchmarks in the area of urban driving research, i.e. NoCrash and CARLA benchmarks, rely on this simulation environment. And hence we plan to use the mentioned benchmarks to compare our results.

C. Implementation plan

We aim to train a neural network with the architecture of a Perceiver to classify sensor inputs into discrete control commands. For which, we plan to collect data from the selected two sensors along with corresponding control commands to fully capture expert demonstrations. Additionally, during optimisation of the learner, we plan to use augmentation methods which randomly distort sensor data from either of the sensors or possibly both sensors to an extent, to help with regularisation across modalities. Furthermore, if necessary, additional data can be collected using the DAgger algorithm [17].

This methods and experiments plan is empirically bound to changes and improvements as the proposed method is in an incipient stage.

IV. Conclusion

Our work proposes a step towards utilising multiple modalities in imitation learning for urban driving methods. The idea of fusing sensory information could exploit the complementary characteristics of each sensor involved. Moreover, in situations where one of the sensors might be blinded, another one can assist in decision making and hence improving the overall performance.

References

[1] P. Daniušis, S. Juneja, L. Valatka, and L. Petkevičius, "Topological navigation graph framework," *Autonomous Robots*, May 2021. [Online]. Available: https://doi.org/10.1007/s10514-021-09980-x

[2] J. Ziegler *et al.*, "Making bertha drive—an autonomous journey on a historic route," *IEEE Intelligent transportation systems magazine*, vol. 6, no. 2, pp. 8–20, 2014.

[3] T. Agarwal, H. Arora, and J. Schneider, "Affordance-based reinforcement learning for urban driving," *arXiv preprint arXiv:2101.05970*, 2021.

[4] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," in *Conference on Robot Learning*. PMLR, 2020, pp. 66–75.

[5] M. Toromanoff, E. Wirbel, and F. Moutarde, "End-to-end model-free reinforcement learning for urban driving using implicit affordances," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7153–7162.

[6] J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," *Foundations and Trends® in Computer Graphics and Vision*, vol. 12, no. 1–3, pp. 1–308, 2020. [Online]. Available: http://dx.doi.org/10.1561/0600000079

[7] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 4693–4700.

[8] A. Karpathy *et al.*, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[9] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[10] A. Jaegle *et al.*, "Perceiver: General perception with iterative attention," *CoRR*, vol. abs/2103.03206, 2021. [Online]. Available: https://arxiv.org/abs/2103.03206

[11] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," in *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, 2019, pp. 1–7.

[12] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[13] F. Codevilla, E. Santana, A. M. López, and A. Gaidon, "Exploring the limitations of behavior cloning for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9329–9338.

[14] A. Sauer, N. Savinov, and A. Geiger, "Conditional affordance learning for driving in urban environments," in *Conference on Robot Learning*. PMLR, 2018, pp. 237–252.

[15] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," in *Proceedings of the 1st International Conference on Neural Information Processing Systems*, ser. NIPS'88. Cambridge, MA, USA: MIT Press, 1988, p. 305–313.

[16] M. Bojarski *et al.*, "End to end learning for self-driving cars," *CoRR*, vol. abs/1604.07316, 2016. [Online]. Available: http://arxiv.org/abs/1604.07316

[17] S. Ross, G. J. Gordon, and J. Andrew Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," *J. Mach. Learn. Res.*, vol. 15, pp. 627–635, 11 2010.

ISBN: 978-1-61208-875-4