Vilnius
University

Idea Paper:
Combining Multiple Modalities with Perceiver in Imitation-based Urban Driving

Authors:
Shubham Juneja (Presenter)
Virginijus Marcinkevičius
Povilas Daniušis

Presenter’s Details:
Institute of Data Science & Digital Technologies, Vilnius University
shubham.juneja@mif.stud.vu.lt

About me

- Originally from Mumbai, India.
- Current positions:
    - PhD Student at Vilnius University; Topic: Deep Imitation Learning for mobile robot navigation.
    - Researcher at Neurotechnology, located in Vilnius, Lithuania.
- Current and past projects:
    - Brain Computer Interface research.
    - Mobile robot navigation research.
- Education Background:
    - Masters in Informatics
    - Bachelors in Computer Engineering
    - Pre-Bachelors Diploma in Computer Engineering

Urban Driving

- The problem of Urban Driving represents making a self driving car being able to drive seamlessly in real world urban environments.
- The techniques used in research area also closely involve the field of robot navigation, where robots are made to be able to navigate in environments such as factories, workshops, etc.
- Due to the rise of the need of automation in recent years, this research area has gotten lots of attention.

Urban Driving

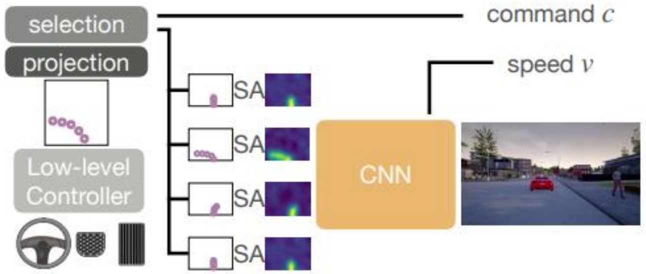
- Driving in urban areas is a complex problem, due to the vast possibilities of situations one can run into.
- Current state-of-the-art is either based on modular approaches or end-to-end learned approaches.
- **Modular approaches** are able to generalise to unseen environments and use multiple sensors to perceive the world, but they are hard to engineer.
- **End-to-end approaches** learn the skill of driving directly from data, hence do not require high engineering effort.
- But end-to-end approaches do not generalise well to new environments and most methods rely on a single modality (single input data source, i.e. sensor).

Urban Driving

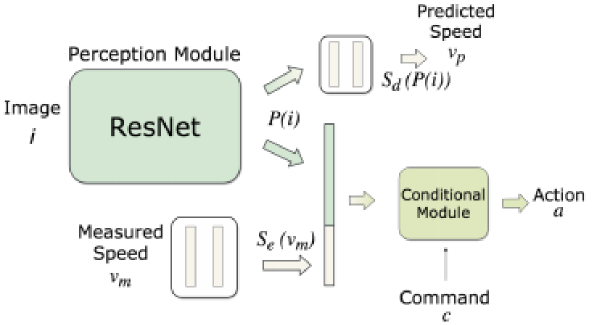
- End-to-end methods rely on Convolutional Neural Networks (CNNs) and to fuse data coming from multiple sources into CNNs can be a research area of its own.
- Common methods in the area of neural networks for various domains:
  - Concatenation at entry point (Early fusion)
  - Concatenation before the final layer (Late fusion)
  - Summation or averaging of features, etc.
- This creates a long list of experiments.

Current State of the Art

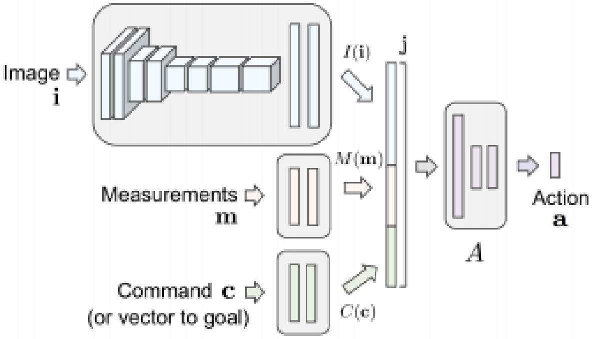
- End-to-end methods are based upon either imitation learning or reinforcement learning.
- These methods attempt to solve urban driving by various way such as conditioning the control on high level commands, predicting speed separately, involving affordances as intermediate representations, etc.



D. Chen *et al.* , "Learning by cheating," in Conference on Robot Learning. PMLR, 2020, pp. 66–75.



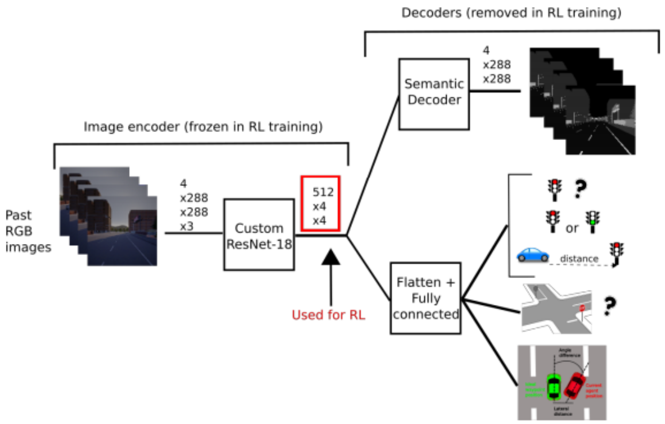
F. Codevilla *et al.* "Exploring the limitations of behavior cloning for autonomous driving," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9329–9338.



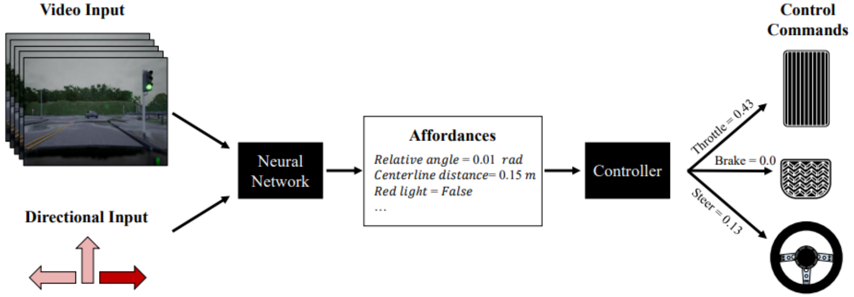
F. Codevilla *et al.*, "End-to-end driving via conditional imitation learning," in 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018, pp. 4693–4700.

Current State of the Art

- The mentioned methods chronologically out perform on the standard urban driving benchmark and are able to drive in urban environments.
- But, they do not show high generalisation which can be reliable in real world applications.
- These methods are also highly dependent upon only a single rich sensor modality, that is, a RGB camera input.



M. Toromanoff *et al.*, "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.



A. Sauer *et al.*, "Conditional affordance learning for driving in urban environments," in Conference on Robot Learning. PMLR, 2018, pp. 237–252.

Current State of the Art

- Results on the NoCrash urban driving benchmark denoting the success rate of getting from one point to another.

Traffic levels	CIL	CAL	CILRS	LBC	IA	IRL
Empty	48 ± 3	36 ± 6	51 ± 1	100 ± 0	99	100 ± 0
Regular	27 ± 1	26 ± 2	44 ± 5	94 ± 3	87	98 ± 1
Dense	10 ± 2	9 ± 1	38 ± 2	51 ± 3	42	91 ± 1

Results reported in the paper: T. Agarwal *et al.*, “Affordance-based reinforcement learning for urban driving,” arXiv preprint arXiv:2101.05970, 2021.

Problem and Aim

The problem:

Current methods do not leverage possibility of using multiple modalities and learning from data at the same time, and hence either affecting generalisation ability or increase the demand of engineering efforts.

The aim:

- To create an approach to drive in urban areas which:
  - Generalises better to unseen environments
  - Is an end-to-end approach which learns the skill of urban driving from demonstrations
  - Learns from data fused from more than one input source

Idea of the paper

- Our idea is to learn the skill of urban driving with more than one input sensor, utilising the **Perceiver neural network architecture**.
- We will seek to employ an RGB camera along with a LIDAR sensor for perception.
- This idea also aims to keep the engineering effort required to a low by utilising imitation learning to form an end-to-end learned data driven method.

The Perceiver Architecture

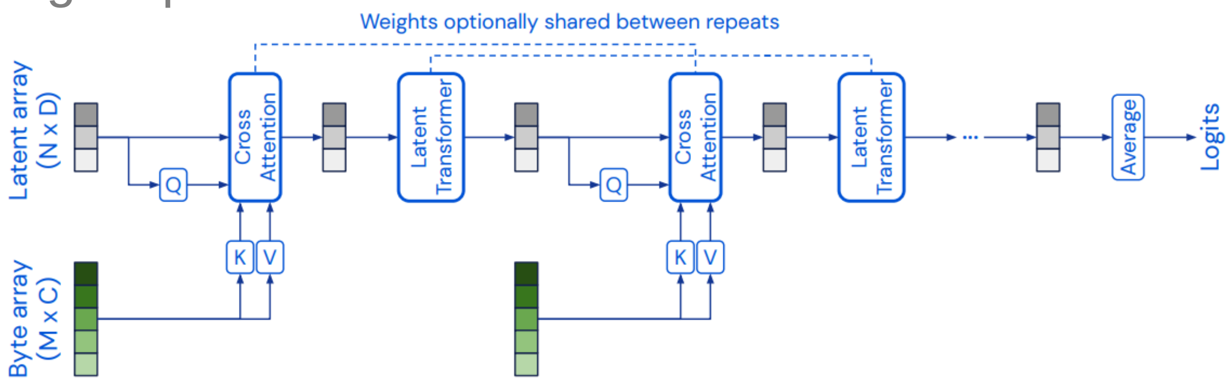
- The Perceiver architecture is designed to be able to work with multiple modalities of inputs of different configurations.
- It has proven to be capable of performing close to the state-of-the-art in
  - ImageNet classification (single modality)
  - Video classification with integrating audio information (multi-modal)
- Using this architecture can completely rule out the need for architecture search when combining sensor modalities.

Method	Accuracy (%)
ResNet-50 (He et al., 2016)	77.6
ViT-B-16 (Dosovitskiy et al., 2021)	77.9
Perceiver	78.0

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

The Perceiver Architecture

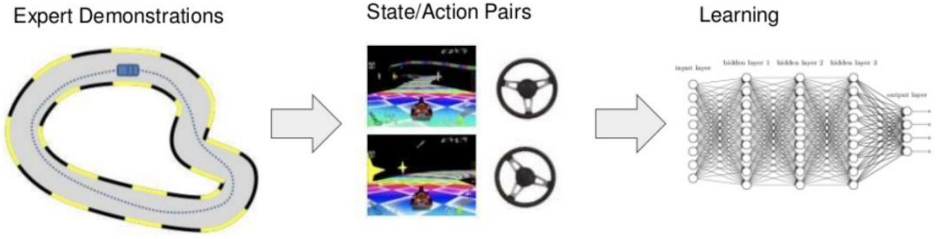
- The Perceiver architecture is able to deal with various input configurations by using Transformers.
- Transformers are flexible architecture blocks that make few assumptions about their inputs and scale quadratically with number of inputs.
- Perceiver architecture also leverages an asymmetric attention mechanism to distil inputs to a tight latent bottleneck, which allows it to handle very large inputs.



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

Implementation plan

- We plan to:
  - Collect data as per the CARLA and NoCrash benchmark standards, using a camera and a LIDAR sensor in the CARLA driving simulator.
  - Training the perceiver architecture with imitation learning.
  - Iteratively collect additional data with the DAgger algorithm, if needed.
  - Evaluate the performances on the CARLA and NoCrash benchmarks.
  - Additionally, use dropout like augmentation across the sensors, to add regularization.



Conclusion

- The proposed idea can be a step towards multi-modal imitation learning and sensor fusion in end-to-end urban driving methods.
- It holds potential in saving the algorithm from taking the wrong decision when information from one sensor may be inaccessible.

Thank you

Contact details:
shubham.juneja@mif.stud.vu.lt