2020 I-RIM Conference
December 10-12, 2020, Rome, Italy (Virtual)
ISBN: 9788894580518
DOI:10.5281/zenodo.4781214

Cognitive-inspired Perceptual Model for Driving Automation

Alice Plebe
Dept. of Industrial Engineering
University of Trento
Trento, Italy
alice.plebe@unitn.it

Gastone Pietro Rosati Papini
Dept. of Industrial Engineering
University of Trento
Trento, Italy
gastone.rosatipapini@unitn.it

Mauro Da Lio
Dept. of Industrial Engineering
University of Trento
Trento, Italy
mauro.dalio@unitn.it

***Abstract*—This paper proposes a neural network model for visual perception in the context of autonomous driving inspired by the human cognition. Despite the growing research aimed at implementing self-driving cars, no artificial system can claim to have reached the driving performance of humans, yet. We believe that the theories about the human mind and its neural organization may reveal precious insights on how to design a more refined perceptual system for driving automation.**

I. The Cognitive Inspiration

The first neurocognitive theory we take inspiration from concerns how sensory information is coded into low-dimensional perceptual representations in the brain. These representations preserve information about the actions that caused the perceptual stimulus. In this way, the brain can recreate the original stimulus in an approximated form, during a phenomenon called *mental imagery* [1]. The first evidence of these internal representations has led to develop a broader theory [2] identifying more sophisticated neural structures called *convergence-divergence zones* (CDZs). In this case, the very same neuron ensembles perform convergent and divergent projections depending on the current action the brain is engaged with: the convergent flow is dominant during perceptual recognition, while the divergent flow occurs during mental imagery. For this reason, CDZs have been recognized as a crucial component in the formation of concepts in the brain. Therefore, we believe fruitful to design an artificial model with a similar hierarchical architecture for learning the abstract concepts relevant to the driving context.

The second theoretical idea concerns the nature of the neural representations in the brain. In most cases, neural representations are not abstract representations of the environment but neural states functional to predicting the future states of the environment. There is evidence in the brain of various circuits that provide prediction from perceptual representations. One of the most popular theories interprets the mental mechanism of prediction in mathematical terms [3]. This theory, called *predictive brain*, explains the behavior of the brain as the minimization of free-energy, a quantity that can be expressed in mathematical form. Therefore, we decide to adopt this formulation as the loss function to train our artificial neural network.

II. The Artificial Implementation

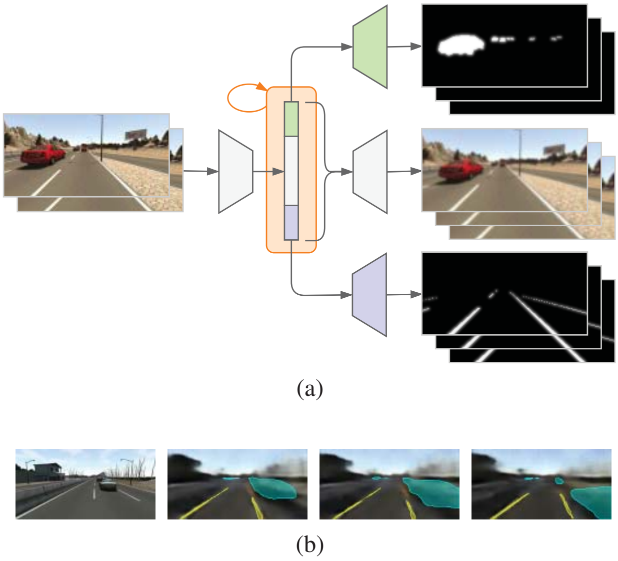
We identify two methods within the framework of artificial neural networks (ANNs) that appear, at least in part, rough algorithmic counterparts of the neurocognitive theories just described. The CDZs may find a correspondence in the idea of convolutional autoencoders [4], while the predictive brain theory resonates with the adoption of variational Bayesian inference in combination with autoencoders [5], [6].

Our method learns conceptual representations of the driving scenario from visual information. In line with the two neurocognitive theories, we propose an approach that forces the representations to be oriented to the driving tasks, under two distinct perspectives.

1) From a static perspective, we force separate groups of neural units to encode specific concepts crucial in the driving task distinctly. Specifically, we use as few as 16 neurons for each of the two basic concepts we adopt: `cars` and `lanes`. The latent space is explicitly partitioned in regions that encode different concepts so that they can be manipulated individually.
2) From a dynamic perspective, we bias the compact representations to predict how the current road scene would change in the future. Albeit this work does not fully develop visual mental imagery, it constitutes progress from mere perception to the creation of manipulable concepts that may increase the cognition abilities of intelligent vehicles.

The overall model is composed of two networks. A first network (Fig. 1a) learns to represent visual scenarios into compact vectors that are at once semantically organized and temporally coherent. By exploiting semantic segmentation as a supporting task, the model forces separate groups of neurons to distinctly represent the basic concepts of `cars` and `lanes`, while self-supervision is adopted to bias the internal representation towards the ability to predict the dynamics of objects in the scene. A second neural network uses the compact representations to perform imagery and predict long-term future frames (Fig. 1b).

Our approach differs from other related works precisely in the learning of the representations: first, there is a semantic organization in the sense that distinct parts of the representa-


(a)

(b)

Fig. 1. The idea behind our approach: a first model perceives the driving scenario in terms of conceptual representations, and a second model exploits the representations to predict changes in the scenario.

tion are explicitly associated with specific concepts useful in the context of driving; second, the temporal coherence that is achieved through self-supervision allows the representation to be exploited for mental imagery and prediction of plausible future scenarios.

III. RESULTS AND FUTURE WORKS

For training and testing our models, we adopt the SYNTHIA dataset [7], a large collection of photo-realistic video sequences rendered using the game engine Unity. The dataset comprises about $100,000$ images of urban scenarios recorded from a simulated camera placed on the windshield of the ego car. Each video sequence is acquired at 5 FPS and comes with semantic annotations or several classes including lane markings, which are not commonly found in other datasets.

In the final version of the architecture, the latent space representation uses just 128 neurons, of which 16 encoding the `car` concept and another 16 for the `lane marking` concept. Since the images fed to the network have dimension of $256 \times 128 \times 3$ and the latent space dimension is 128, the compression performed by our network is almost of 4 orders of magnitude. This is a considerable achievement compared to related works adopting variational autoencoder [8], [9], which limit the compression of the encoder to only 1 order of magnitude.

We measure the goodness of the learned representations using quantitative metrics like the IoU and a statistical evaluation of the latent representations measuring the consistency for the temporal dynamics and their predictability. We also evaluate the quality of the representations using some visual tests, like interpolating and swapping components between latent spaces, and replicating the phenomenon of mental imagery by calling the network iteratively and feeding the output back as the input of the next iteration. In all the cases, the model successfully produces new plausible driving scenarios not seen before during the learning. For a complete presentation of qualitative and quantitative results, please refer to [10].

Here we have described the example of predicting long-term future frames in a video sequence. However, once learned, the representations can be deployed in many possible contexts. For example, we are currently working on using the representations to predict future occupancy grid maps. Moreover, since we achieve to assign only 16 neurons to each concept in the representation, it is possible to include in future works more concepts inside the latent representations. It would be interesting, for example, to include concepts of vulnerable road users, such as `pedestrians` and `bikes`. One more future development we have planned is the adoption of a dataset of real-world video recordings. One of the reasons we adopted the SYNTHIA dataset at the beginning of our research, besides its large size and variety, was the availability of lane marking annotations, which are very rare among the popular datasets for autonomous driving. Recently, UC Berkeley introduced the BDD100K dataset [11], which includes high-quality video sequences with several types of lane marking annotations. Hence, the adoption of this novel dataset could be an promising addition to our work.

REFERENCES

[1] S. T. Moulton and S. M. Kosslyn, "Imagining predictions: mental imagery as mental emulation," *Philosophical transactions of the Royal Society B*, vol. 364, pp. 1273–1280, 2009.

[2] K. Meyer and A. Damasio, "Convergence and divergence in a neural architecture for recognition and memory," *Trends in Neuroscience*, vol. 32, pp. 376–382, 2009.

[3] K. Friston, "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, pp. 127–138, 2010.

[4] M. Tschannen, M. Lucic, and O. Bachem, "Recent advances in autoencoder-based representation learning," in *NIPS Workshop on Bayesian Deep Learning*, 2018.

[5] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of International Conference on Learning Representations*, 2014.

[6] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of Machine Learning Research*, E. P. Xing and T. Jebara, Eds., 2014, pp. 1278–1286.

[7] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3234–3243.

[8] E. Santana and G. Hotz, "Learning a driving simulator," *arXiv*, vol. abs/1608.01230, 2016.

[9] D. Ha and J. Schmidhuber, "World models," *arXiv*, vol. abs/1803.10122, 2018.

[10] A. Plebe and M. Da Lio, "On the road with 16 neurons: Towards interpretable and manipulable latent representations for visual predictions in driving scenarios," *IEEE Access*, vol. 8, pp. 179 716–179 734, 2020.

[11] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2636–2645.