A Cognitively Informed Perception Model for Driving

Alice Plebe[1] and Mauro Da Lio[2]

***Abstract*— Deep learning is responsible for the current renewed success of artificial intelligence. Applications that in the recent past were considered beyond imagination, now appear to be feasible. The best example is autonomous driving. However, despite the growing research aimed at implementing autonomous driving, no artificial intelligence can claim to have reached or closely approached the driving performance of humans, yet. Deep learning is an evolution of artificial neural networks introduced in the '80s with the *Parallel Distributed Processing* (PDP) project. There is a fundamental difference in aims between the first generation of artificial neural networks and deep neural models. The former was motivated primarily by the exploration of cognition. Current deep neural models are instead developed with engineering goals in mind, without any ambition or interest in exploring cognition. Some important components of deep learning – for example reinforcement learning or recurrent networks – owe indeed an inspiration to neuroscience and cognitive science, as PDP far legacy. But this connection is now neglected, what matters is only the pragmatic success in applications. We argue that it urges to reconnect artificial modeling with an updated knowledge of how complex tasks are realized by the human mind and brain. In this paper, we will first try to distill concepts within neuroscience and cognitive science relevant for the driving behavior. Then, we will identify possible algorithmic counterparts of such concepts, and finally build an artificial neural model exploiting these components for the visual perception task of an autonomous vehicle.**

I. FROM THE COGNITIVE SIDE

A. The Simulation Theory

A well-established theory in cognitive science is the one proposed by Jeannerod and Hesslow, the so-called *simulation theory of cognition*, which proposes that thinking is essentially a simulated interaction with the environment [1], [2]. In their view, simulation is a general principle of cognition, which can be expressed in at least three different components: perception, actions and anticipation.

The most simple case of simulation is mental imagery, especially in visual modality. This is the case, for example, when a person tries to picture an object or a situation. During this phenomenon, the primary visual cortex (V1) is activated with a simplified representation of the object of interest, but the visual stimulus is not actually perceived.

This work was developed inside the EU Horizon 2020 Dreams4Cars Research and Innovation Action project, supported by the European Commission under Grant 731593. The Authors want also to thank the Deep Learning Lab at the ProM Facility in Rovereto (TN) for supporting this research with computational resources funded by Fondazione CARITRO.

[1]Alice Plebe is with the Dept. of Information Engineering and Computer Science, University of Trento, Italy alice.plebe@unitn.it

[2]Mauro Da Lio is with the Dept. of Industrial Engineering, University of Trento, Italy mauro.dalio@unitn.it

B. Convergence–Divergence Zones

Although the simulation theory is one of the most established, it does not identify how simulation takes place at neural level. A prominent proposal in this direction is the formulation of the convergence-divergence zones (CDZs) [3]. They highlight the "convergent" aspect of certain neuron ensembles, located downstream from primary sensory and motor cortices. Such convergent structure consists in the projection of neural signals on multiple cortical regions in a many-to-one fashion. On the other hand, the neuron ensembles have the ability to reciprocate feedforward projections with feedback projections in a one-to-many fashion, realizing the divergent flow.

The primary purpose of convergence is to exploit synaptic plasticity in order to record which patterns of features – coded as knowledge fragments in the early cortices – occur in relation with a specific higher-level concept. Such records are built through experience, by interacting with objects. The convergent flow is dominant during perceptual recognition, while the divergent flow dominates imagery.

Convergent-divergent connectivity patterns can be identified for specific sensory modalities, but also in higher order association cortices. It should be stressed that CDZs are rather different from a conventional processing hierarchy, where processed patterns are transferred from earlier to higher cortical areas. In CDZs, part of the knowledge about perceptual objects is retained in the synaptic connections of the convergent-divergent ensemble. This allows to reinstate an approximation of the original multi-site pattern of a recalled object or scene.

C. Transformational Abstraction

One major challenge in cognitive science is explaining the mental mechanisms by which we build conceptual abstractions. The conceptual space is the mental scaffolding the brain gradually learns through experience, as internal representation of the world. In particular, conceptual abstraction is derived mostly from perceptual experience, which fits perfectly with the approach implemented by artificial neural networks.

As highlighted by [4] CDZs are a valid systemic candidate for how the formation of high-level concepts takes place at brain level. However, the idea of CDZs is just sketched and cannot provide a detailed mechanism for conceptual abstractions. A difficulty with acquiring abstract categories lies in the inconsistent manifestations of the characteristic features across real exemplars.

A suggested solution to this difficult issue is the *transformational abstraction* [5], [6] performed by a hierarchy

of cortical operations, as in the ventral visual cortex. The essence of transformational abstraction, from a mathematical point of view, lies in the combination of two operations: linear convolutional filtering and nonlinear downsampling. Operations of this sort have been identified in the V1 [7], [8], and are well recognized in the primate ventral visual path as well [9], [10].

D. The Predictive Theory

The reason why cognition is mainly explicated as simulation, according to Hesslow or Jeannerod, is because the brain can achieve through simulation the most precious information of an organism: a prediction of the state of affairs in the future environment. The need of prediction, and how it molds the entire cognition, has become the core of another popular theory popular known as "Bayesian brain", "predictive brain", or "free-energy principle for the brain" introduced by Friston [11]. According to him the behavior of the brain – and of an organism as a whole – can be conceived as minimization of free-energy, a quantity that can be expressed in several ways depending on the kind of behavior and the brain systems involved.

Free-energy is a concept originated in thermodynamics, as a measure of the amount of work that can be extracted from a system. What is borrowed by Friston is not the thermodynamic meaning of the free-energy, but its mathematical form only, which is derived from the framework of variational Bayesian methods in statistical physics We will see in §II-B how the same probabilistic framework will be used in the derivation of a deep neural model. For example, this is his free-energy formulation in the case of perception [12, p.427]:

$$F_P = \Delta_{\mathrm{KL}}\Big(\check{p}(\mathbf{c}|\mathbf{z}) \| p(\mathbf{c}|\mathbf{x},\mathbf{a})\Big) - \log p(\mathbf{x}|\mathbf{a}) \qquad (1)$$

where $\mathbf{x}$ is the sensorial input of the organism, $\mathbf{c}$ is the collection of the environmental causes producing $\mathbf{x}$, $\mathbf{a}$ are actions that act on the environment to change sensory samples, and $\mathbf{z}$ are inner representations of the brain. The quantity $\check{p}(\mathbf{c}|\mathbf{z})$ is the encoding in the brain of the estimate of causes of sensorial stimuli. The quantity $p(\mathbf{c}|\mathbf{x},\mathbf{a})$ is the conditional probability of sensorial input conditioned by the actual environmental causes $\mathbf{c}$. The discrepancy between the estimated probability and the actual probability is given by the Kullback-Leibler divergence Δ_{KL}. The minimization of F_P in equation (1) optimizes $\mathbf{z}$.

II. TO THE ARTIFICIAL SIDE

A. Convergence–divergence as Autoencoder

In the realm of artificial neural networks, the computational idea that most closely resonate with CDZ is the *autoencoder*. It is an idea that has been around for a long time, it was the cornerstone of the evolution from shallow to deep neural architectures [13], [14]. More recently, autoencoders have been widely adopted for their ability to capture compact information from high dimensional data. The basic structure of an autoencoder is composed of a feature-extracting part called *encoder* and a *decoder* part mapping from feature space back into input space. There is a clear correspondence between the encoder and the convergence zone in the CDZ neurocognitive concept, and similarity between the decoder and the divergence zone.

Then how exactly convergence–divergence can be achieved inside autoencoders? An interesting approach is the one closely related to the transformational abstraction hypothesis described in §I-C: the *deep convolutional neural networks* (DCNNs). They implement the hierarchy of convolutional filtering alternated with nonlinear downsampling, and are considered the essence of transformational abstraction. In addition, there is growing evidence of striking analogies between patterns in DCNN models and patterns of voxels in the brain visual system. Several studies have successfully related results of deep learning models with the visual system [15], [16], finding reasonable agreement between features computed by DCNN models and fMRI data. Convolutional–deconvolutional autoencoders are therefore a highly biologically plausible implementation for the CDZ theory, at least in the case of visual information.

B. Predictive Brain as Variational Autoencoder

In the last few years there has been renewed interest in the area of Bayesian probabilistic inference in learning models of high dimensional data. The Bayesian framework, variational inference in particular, has found a fertile ground in combination with neural models. Two concurrent and unrelated developments [17], [18] have made this theoretical advance possible, connecting autoencoders and variational inference. This new approach became quickly popular under the term *variational autoencoder*, and a variety of neural models have been proposed over the years.

The loss function for a variational autoencoder is defined as follows:

$$\begin{aligned}\mathcal{L}(\Theta,\Phi|\mathbf{x}) = \Delta_{\mathrm{KL}}\big(q_\Phi(\mathbf{z}|\mathbf{x}) \| p_\Theta(\mathbf{z})\big) + \\ - \mathbb{E}_{\mathbf{z}\sim q_\Phi(\mathbf{z}|\mathbf{x})}\left[\log p_\Theta(\mathbf{x}|\mathbf{z})\right]\end{aligned} \qquad (2)$$

where $\mathbf{x}$ is a high dimensional random variable, $\mathbf{z}$ the representation of the variable in the low-dimensional latent space. Θ and Φ are parameters describing, respectively, the decoder and encoder of the network. p_Θ is computed by the decoder and represents the desired approximation of the unknown input distribution p, and q_Φ is the auxiliary distribution computed by the encoder from which to sample $\mathbf{z}$. $\mathbb{E}[\cdot]$ is the expectation operator, and Δ_{KL} is the Kullback-Leibler divergence.

It is evident how this mathematical formulation is impressively similar to the concept of free energy in Friston. Despite this close analogy, all the proposers of variational autoencoder are either unaware or fully disinterested of this coincidence. It is not so surprising because mainstream deep learning is driven by engineering goals without any interest in connections with cognition. We believe instead that a strong connection between a well established cognitive theory and a computational solution greatly argues in favor of adopting such a solution.

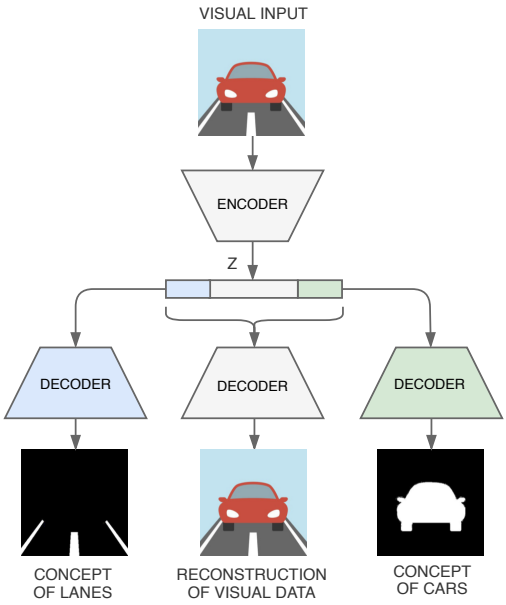

Fig. 1. The architecture of our model.

III. IMPLEMENTATION

In the previous section we have reviewed several components that match quite closely the relevant neurocognitive theories identified in §I. Our proposed model attempts to weave together these components, finalized at visual perception in autonomous driving agents.

Similarly to the hierarchical arrangement of CDZs in the brain, our model is provided with different levels of processing paths. A first processing path starts from the raw image data and converges up to a low-dimension representation of visual features. Consequently, the divergent path outputs in the same format as the input image. The other processing path leads to representations that are no more in terms of visual features, rather in terms of concepts. As discussed in §I-C, our brain naturally projects sensorial information – especially visual – into conceptual space, where the local perceptual features are pruned and neural activation code the nature of entities present in the environment that produced the stimuli. In the driving context it is not necessary to infer categories for every entity present in the scene, it is useful to project in conceptual space only the objects relevant to the driving task. In the model presented here we choose to consider the two main concepts of `cars` and `lane markings`.

As depicted in Fig. 1, the presented variational autoencoder is composed by one shared encoder and three independent decoders. All the components of the architecture are trained jointly. The encoder compresses an RGB image to a compact high-feature representation. Then the decoders map different part of the latent space back to separated output spaces: one into the same visual space of the input; the other two into conceptual space, producing binary images containing, respectively, `car` entities and `lane marking` entities. So, in our implementation the entire latent vector $\mathbf{z}$ represents inside the visual space, and at the same time two inner segments represent specifically the `car` and `lane` concepts. The rationale for this choice is that in mental imagery there is no clear cut distinction between low-level features and semantic features, the entire scene is mentally reproduced, but including the awareness of the salient concepts present in the scene.

Note that the idea of partitioning the entire latent vector into meaningful components is not new. In the context of processing human heads the vector has been forced to encode separate representations for viewpoints, lighting conditions, shape variations [19]. In [20] the latent vector is partitioned in one segment for the semantic content and a second segment for the position of the object. Our approach is different. While we keep disjointed the two segments for the `car` and `lane` concepts, we fully overlap these two representations within the entire visual space. This way, we adhere entirely to the CDZ principle, and try to achieve the full scene by divergence, but at the same time including awareness for the `car` and `lane` concepts.

IV. RESULTS

We present here a selection of results achieved with an instance of the model described in the previous section. The final architecture is trained for 200 epochs, and used 4 convolutional layers in the encoder, 4 deconvolutional layers for each decoder, and a latent space representation of 128 neurons, of which 16 encoding the `car` concept and another 16 for the `lane marking` concept. We would like to highlight that, 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 the network is almost of 4 orders of magnitude. This is a considerable achievement compared to other relevant works adopting variational autoencoder [21], [22] which limit the compression of the encoder to only 1 order of magnitude.

We trained and tested the presented model on the SYNTHIA dataset [23], a large collection of synthetic images representing various urban scenarios. The dataset contains about $100,000$ color images (and as many corresponding segmented images, used for ground truth of the conceptual branches of the network). We used 70% of the data for training, 25% for validation and 5% for testing.

Fig. 2 shows the image results produced by our model for a selection of driving scenarios. The images are processed to better show at the same time the results on conceptual space and visual space. The colored overlays highlight the concepts computed by the network: the cyan regions are the output of the `car` divergent path, and the pink overlays are the output of the `lane markings` divergent path. Fig. 2 includes a variety of driving situations, going from sunny environments (top rows) to very adverse driving conditions (bottom rows) in which the detection of other vehicles can be challenging even for a human. These results nicely show how the projection of the sensorial input (original frames) into conceptual representation is very effective in identifying

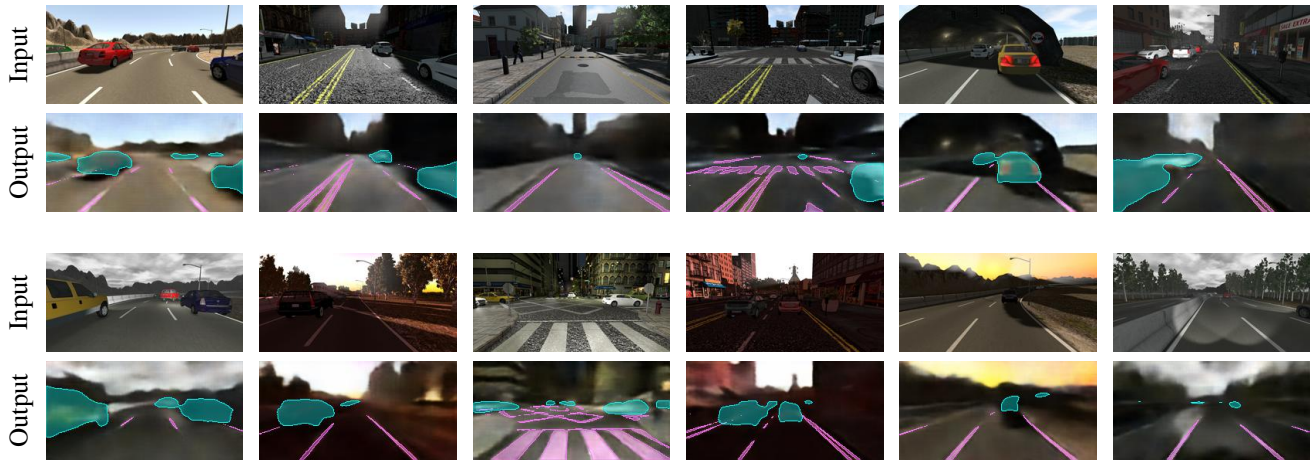

Fig. 2. Results of our model for a selection of frames from the SYNTHIA dataset, with different environmental and lighting conditions.

and preserving the sensible features of `cars` and `lane markings`, despite the large variations in lighting and environmental conditions.

Lastly, we would like to stress that the purpose of our network is not mere segmentation of visual input. The segmentation task is to be considered as a support task, used to enforce the network to learn a more robust latent space representation, which now is explicitly taking into consideration two of the concepts that are fundamental to the driving tasks.

V. CONCLUSIONS

The model here presented is an attempt to convert into an artificial neural network model the fundamental theories about how the brain processes its sensory inputs to produce purposeful representations. We especially identified the consolidated variational autoencoder architecture as the best candidate for implementing convergence-divergence zone schemes. The reason for constraining a deep learning model on cognitive theoretical grounds, instead of starting from scratch as often done, derives from the observation of how humans excel in sophisticated sensorimotor control tasks such as driving.

REFERENCES

[1] M. Jeannerod, "Neural simulation of action: A unifying mechanism for motor cognition," *NeuroImage*, vol. 14, pp. S103–S109, 2001.

[2] G. Hesslow, "The current status of the simulation theory of cognition," *Brain*, vol. 1428, pp. 71–79, 2012.

[3] 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.

[4] J. S. Olier, E. Barakova, C. Regazzoni, and M. Rauterberg, "Reframing the characteristics of concepts and their relation to learning and cognition in artificial agents," *Cognitive Systems Research*, vol. 44, pp. 50–68, 2017.

[5] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick, "Neuroscience-inspired artificial intelligence," *Neuron*, vol. 95, pp. 245–258, 2017.

[6] C. Buckner, "Empiricism without magic: transformational abstraction in deep convolutional neural networks," *Synthese*, vol. 195, pp. 5339–5372, 2018.

[7] D. Hubel and T. Wiesel, "Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex," *Journal of Physiology*, vol. 160, pp. 106–154, 1962.

[8] C. D. Gilbert and T. N. Wiesel, "Morphology and intracortical projections of functionally characterised neurones in the cat visual cortex," *Nature*, vol. 280, pp. 120–125, 1979.

[9] D. J. Felleman and D. C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex," *Cerebral Cortex*, vol. 1, pp. 1–47, 1991.

[10] D. C. Van Essen, "Organization of visual areas in macaque and human cerebral cortex," in *The Visual Neurosciences*, L. Chalupa and J. Werner, Eds. Cambridge (MA): MIT Press, 2003.

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

[12] K. Friston and K. E. Stephan, "Free–energy and the brain," *Synthese*, vol. 159, pp. 417–458, 2007.

[13] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 28, pp. 504–507, 2006.

[14] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[15] U. Güçlü and M. A. J. van Gerven, "Unsupervised feature learning improves prediction of human brain activity in response to natural images," *PLoS Computational Biology*, vol. 10, pp. 1–16, 2014.

[16] B. P. Tripp, "Similarities and differences between stimulus tuning in the inferotemporal visual cortex and convolutional networks," in *International Joint Conference on Neural Networks*, 2017, pp. 3551–3560.

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

[18] 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.

[19] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. B. Tenenbaum, "Deep convolutional inverse graphics network," in *Advances in Neural Information Processing Systems*, 2015, pp. 2539–2547.

[20] J. Zhao, M. Mathieu, R. Goroshin, and Y. LeCun, "Stacked what-where auto-encoders," in *International Conference on Learning Representations*, 2016, pp. 1–12.

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

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

[23] G. Ros, L. S. J. M. 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.