Mental Imagery for Intelligent Vehicles

Alice Plebe[1], Riccardo Donà[2], Gastone Pietro Rosati Papini[2] and Mauro Da Lio[2]
[1]*Department of Information Engineering and Computer Science, University of Trento, Italy*
[2]*Department of Industrial Engineering, University of Trento, Italy*

Keywords: Autonomous Driving, Simulation Theory, Convergence-divergence Zones, Autoencoders.

Abstract: The research in the design of self-driving vehicles has been boosted, in the last decades, by the developments in the fields of artificial intelligence. Despite the growing number of industrial and research initiatives aimed at implementing autonomous driving, none of them can claim, yet, to have reached the same driving performance of a human driver. In this paper, we will try to build upon the reasons why the human brain is so effective in learning tasks as complex as the one of driving, borrowing explanations from the most established theories on sensorimotor learning in the field of cognitive neuroscience. The contribution of this work would like to be a new point of view on how the known capabilities of the brain can be taken as an inspiration for the implementation of a more robust artificial driving agent. In this direction, we consider the Convergence-divergence Zones (CDZs) as the most prominent proposal in explaining the simulation process underlying the human sensorimotor learning. We propose to use the CDZs as a "template" for the implementation of neural network models mimicking the phenomenon of mental imagery, which is considered to be at the heart of the human ability to perform sophisticated sensorimotor controls such driving.

1 INTRODUCTION

For the last two decades, artificial neural networks (ANNs) have been at the very heart of many technology developments (Schmidhuber, 2015; Chui et al., 2018; Hazelwood et al., 2018). They have proved to be the best available approach for a variety of different problem domains (Liu et al., 2017; Jones et al., 2017), and the design of autonomous vehicles is definitely one of the research areas to have amply benefited from the rise of deep learning, e.g., (Bojarski et al., 2017; Li et al., 2018; Schwarting et al., 2018).

In the recent years, however, some concerns have emerged regarding certain crucial features of artificial neural nets, which may call into question the relentless progress that was foreseen at first. In a recent interview[1] at MIT, Yoshua Bengio, responsible for many of the advancements of deep learning, pointed out the inherent weakness of artificial neural networks as opposed to expert systems:

> "The knowledge in an expert systems is nicely decomposed into a bunch of rules, whereas neural nets [...] have this big blob of parameters which work intensely together to represent everything that the network knows. It is not sufficiently factorized, and I think this is one of the weaknesses of current neural nets".

Such vulnerability appears to be a serious hindrance in the application of artificial neural nets inside safety-critical systems, like autonomous vehicles. When treated as any other components of a car, neural networks should comply with the ISO 26262[2] safety standard, which covers all aspects of automotive development, production and maintenance of safety-related systems. In fact, a major challenge that has yet emerged in implementing self-driving cars is how to perform quality assessment when key components are based on neural networks, as their intrinsic opaque structure does not provide any explanation on what information in the input is considered to produce a certain prediction. This is known as the *black-box* issue, characteristic of deep neural networks (Samek et al., 2017; Ras et al., 2018).

A problem closely related to the above is how to demonstrate that an autonomous driving agent is (much) safer than a human driver. Obviously, the desire to develop self-driving cars stems from the aspiration of achieving safer streets for everyone – drivers, passengers and pedestrians. Yet, none of the current

[1]https://agi.mit.edu (transcription of video interview)

[2]https://www.iso.org/standard/43464.html

Plebe, A., Donà, R., Papini Rosati, G. and Da Lio, M.
Mental Imagery for Intelligent Vehicles.
DOI: 10.5220/0007657500430051
In *Proceedings of the 5th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2019)*, pages 43-51
ISBN: 978-989-758-374-2
Copyright © 2019 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

available implementations of autonomous vehicle can claim to be nowhere close to the driving performance of a human being. The issue also arises from the fact that humans – contrary to common belief – are very reliable at driving: in the US there are just 1.09 fatalities and 77 injuries per 100,000,000 human driven miles (NHTSA, 2017).

Such considerations lead to reflect on why the human brain is so efficient in solving the driving task, and if it is possible to take inspiration from the mechanisms whereby the brain learns to perform such a complex task (inattention, alcohol, tiredness, drugs etc., which are responsible for the vast majority of the very-few human accidents, would not affect the artificial system of course). That said, it is not the intention of this paper to argue against the use of neural networks in the development of autonomous vehicles. Rather, there is no question that nowadays ANNs represent the method of choice for implementing an high-performing artificial agent.

This work, hence, would like to contribute with a novel perspective on how the capabilities of the human brain can be used as inspiration for creating an artificial driving agent, still largely based on deep learning, but more robust. We propose to exploit the current most established neurocognitive theories on how the brain develops the ability to drive, to build a neural network architecture less susceptible to the *black-box* issues mentioned before. In the following Section we will overview the most compelling hypothesis on sensorimotor control learning of the brain, in the domain of cognitive neuroscience. In §3 we will show how these hypothesis can be considered as a starting point for the development of a novel neural network architecture, and finally §4 will present the results of applying our ANN to a simulated driving environment.

This paper results from one of the research projects carried out as part of the European project Dreams4Cars, where we are developing an artificial driving agent inspired by the neurocognition of human driving, for further details refer to (Da Lio et al., 2018).

2 THE NEUROCOGNITIVE POINT OF VIEW

Humans are able to learn an impressive range of different, very complex, sensorimotor controls schemes – from playing tennis to salsa dancing. The remarkable aspect is that no motor skill is innate to humans, not even the most basic ones, like walking or grasping objects (Grillner and Wallén, 2004). All motor controls are, in fact, *learned* through lifetime. The process of human sensorimotor learning involves sophisticated computational mechanisms, like gathering of task-relevant sensory information, selection of strategies, and predictive control (Wolpert et al., 2011).

The ability to drive is just one of the many highly specialized human sensorimotor behaviors. The brain learns to solve the driving task with the same kind of strategy adopted for every sort of motor planning that requires continuous and complex perceptual feedback. We deem that the sophisticated control system the human brain develops when learning to drive by commanding the ordinary car interfaces – steering wheel and pedals – may reveal precious insights on how to implement a robust autonomous driving system.

It should be noted that the human sensorimotor learning is still far from being fully understood, as there are several competing theories about which components of the brain are engaged during learning. However, a huge body of research in neuroscience and cognitive neuroscience has been produced in the past decades, which allows us to grasp some useful cues for designing an artificial driving agent capable of learning the sensorimotor controls necessary to drive.

2.1 The Simulation Theory

A well-established theory is the one proposed by Jeannerod and Hesslow, the so-called *simulation theory of cognition*, which proposes that thinking is essentially simulated interaction with the environment (Jeannerod, 2001; Hesslow, 2012). In the view of Hesslow, simulation is a general principle of cognition, explicated in at least three different components: perception, actions and anticipation. Perception can be simulated by internal activation of sensory cortex in a way that resembles its normal activation during perception of external stimuli. Simulation of actions can be performed when activating motor structures, as during a normal behavior, but suppressing its actual execution. Moreover, Hesslow argues that actions can trigger perceptual simulation of their most probable consequences.

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 (Kosslyn, 1994; Moulton and Kosslyn, 2009).

2.2 The Emulation Theory

Another proposal in understanding certain aspects of motor control and motor imagery, is the *emulation theory of representation* (Grush, 2004), which can be seen as a bridge linking theoretical cognitive neuroscience to the engineering domain of control theory and signal processing. According to this theory, the brain does not simply engage with the body and environment, it is also able to construct neural circuits that act as models of them. These models can also be run offline, in order to predict outcomes of different actions, and evaluate and develop motor plans.

Thus, the main difference between Hesslow's simulation theory and Grush's emulation theory is that the latter claims that mere operation of the motor centers is not enough to produce imagery. According to Grush, a bare motor plan is either a temporal sequence of motor commands or a plan described by movements of joint angles. Conversely, motor imagery is a sequence of simulated proprioception and kinesthesis, and it requires forward models of the musculoskeletal system of the body.

One conceptual advantage of the emulation theory is that it solves the conundrum of how proprioception and kinesthesis can exist during motor imagery in absence of limbs modifications. On the other hand, it faces the burden of explaining how a mental forward model of the musculoskeletal system can be realized at all. Grush proposes it can be realized by Kalman-like filters, the most common system estimator used in control engineering. While there are evidences that Kalman filter schemes can account for several experimental data (Wolpert and Kawato, 1998; Colder, 2011), it is hard to tell if the brain actually solves motor simulation in this way. In the Dreams4Cars project we plan to experiment forward models based on Kalman filters as well, but this is not the subject of this paper. Therefore we will not get into more details of emulators, and we concentrate instead on other proposals about how simulation may take place in the brain.

2.3 Convergence-divergence Zones

Any neural theory claiming to explain the simulation process, in the first place, is required to simultaneously:

1. identify the neural mechanisms that are able to extract information relevant to the action, from a large amount of sensory data,
2. recall related concepts from memory during imagery.

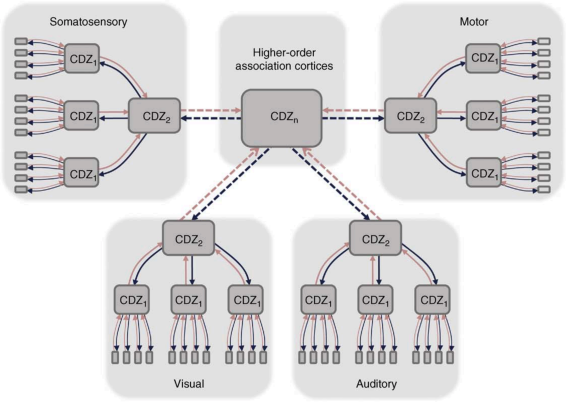

Figure 1: Schematic representation of the CDZ framework by Meyer and Damasio. Neuron ensembles in early sensorimotor cortices of different modalities send converging forward projections (red arrows) to higher-order association cortices, which, in turn, project back divergently (black arrows) to the early cortical sites, via several intermediate steps.

A prominent proposal in this direction has been formulated in terms of convergence-divergence zones (CDZs) (Meyer and Damasio, 2009). They derive from an earlier model (Damasio, 1989) which highlighted 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.

The primary purpose of convergence is to record, by means of synaptic plasticity, which patterns of features – coded as knowledge fragments in the early cortices – occur in relation with a specific concept. Such records are built through experience, by interacting with objects. On the other hand, a requirement for convergence zones (already found in the first proposal of Damasio) is the ability to reciprocate feedforward projections with feedback projections in a one-to-many fashion. This feature is now made explicit in the CDZ naming.

The convergent flow is dominant during perceptual recognition, while the divergent flow dominates imagery. Damasio postulates that switching between one of the two modes may depend on time-locking. If activations in a CDZ is synchronous with activity in separate feeding cortical sites, than perceptual recognition takes place. Conversely, imagery is driven by synchronization with backprojecting cortical areas.

Convergent-divergent connectivity patterns can be identified for specific sensory modalities, but also in higher order association cortices, as shown in the hierarchical structure in Fig. 1. It should be stressed that CDZs are rather different from a conventional processing hierarchy, where processed patterns are trans-

ferred 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.

3 ARTIFICIAL MENTAL IMAGERY

The CDZ hypothesis has found in the years support of a large body of neurocognitive and neurophysiological evidence. However, it is a purely descriptive model and does not address the crucial issue of how the same neural assembly, which builds connections by experiences in the convergent direction, can computationally work in the divergent direction as well.

At the moment, there are no computational models that faithfully replicate the behavior of CDZs, however, we found that an independent notion, introduced in the field of artificial intelligence for very different purposes, bears significant similarities with the CDZ scheme. In our opinion, the most direct mechanism to simulate perception in the realm of artificial neural networks is the *autoencoder*.

Autoencoder architectures have been the cornerstone of the evolution from shallow to deep neural architectures (Hinton and Salakhutdinov, 2006; Vincent et al., 2010). The crucial issue of training neural architectures with multiple internal layers was initially solved associating each internal layers with a Restricted Boltzmann Machine (Hinton and Salakhutdinov, 2006), so that they can be pre-trained individually in unsupervised manner. The adoption of autoencoders overcome the training cost of Boltzmann Machines: each internal layer is trained in unsupervised manner, as an ordinary fully connected layer. The key idea is to use the same input tensor as target of the output, and therefore to train the layer to optimize the reconstruction of the input (Larochelle et al., 2009). In the first layer the inputs are that of the entire neural model, for all subsequent layers the hidden units' outputs of the previous layer are now used as input. The overall result is a regularization of the entire model similar to the one obtained with Boltzmann Machine (Bengio, 2009), or even a better one (Vincent et al., 2010).

Soon after, refinement of algorithms for initialization (Glorot and Bengio, 2010) and optimization (Kingma and Ba, 2014) of weights, made any type of unsupervised pre-training method superfluous. However, autoencoders find a new role for capturing compact information from visual inputs (Krizhevsky and Hinton, 2011). In this kind of models the task to be solved by the network is to simulate as output the same picture fed as input. The advantage is that while learning to reconstruct the input image, the model develops a very compact internal representation of the visual scene. Models able to learn such representation are closely connected with the cognitive activity of mental imagery.

3.1 Autoencoder-based CDZ Models

In the context of autonomous driving agents, there is a range of different levels at which we can design models with autoencoder-like architectures acting as CDZs. Similarly to the hierarchical arrangement of CDZs in the brain, as described by Meyer and Damasio (again, Fig.1), autoencoder-based models can be placed at a level depending on the distance covered by the processing path, from the lowest primary cortical areas to the output of the simulation.

In the context of Dreams4Cars, we considered as the lowest level of model design the processes that start from the raw image data and converge up to simple visual features. Consequently, the divergent path outputs in the same format as the input image.

At an intermediate level, the convergent processing path leads to representations that are no more in terms of visual features, rather in terms of "concepts". Our brain naturally projects sensorial information, especially visual, into conceptual space, where the local perceptual features are pruned, and neural activations code the nature of entities present in the environment that produced the stimuli. The conceptual space is the mental scaffolding the brain gradually learns through experience, as internal representation of the world (Seger and Miller, 2010). As highlighted by (Olier et al., 2017) CDZs are a valid systemic candidate for how the formation of concepts takes place at brain level. There is clearly no single unified center in the brain acting as conceptual space, the organization is far more complex. There are distinctive properties of objects like shape, way of moving and interacting with, which are represented in the same sensory and motor systems that are active when information about these properties was acquired. There are also other regions that seem to show a categorical organization (Martin, 2007; Mahon and Caramazza, 2011). 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 models here presented we choose to consider the two main concepts of *cars* and *lanes*.

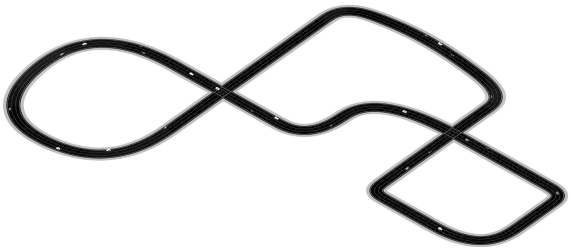

Figure 2: Orthographic view of the road track created in Blender, used for the simulation of road traffic.

A model at a higher level associates the convergent paths from visual processes with motor commands, and its divergent path outputs in the format of action representations. For the purpose of driving, we will use as representation format a space of two dimensions, the steering rate $(1/ms)$ and the longitudinal jerk (m/s^3). The choice for this motor space derives from the Optimal Control (OC) theory. More specifically, the minimization of the squared jerk integral is known to lead to smooth human-like control actions (Bertolazzi et al., 2003; Liu and Todorov, 2007). This higher level has not been fully developed in neural networks yet, therefore this paper will not focus on it.

At all levels, the implementations presented in this paper are synchronous: the convergent phase is applied to data locked in time to the same of the divergent phase. An extension under development in our project is to delay in time the divergent phase. In this case, for all levels it becomes necessary the integration of an additional convergence zone, corresponding to cortical proprioception. In the context of the driving task, it is the processing of information about ego-velocity and ego-heading, together with their time derivatives. This sort of information is clearly necessary in order to imagine, when driving, a visual scene projected in the future.

4 RESULTS

Here we present the implementations of two models of artificial visual imagery, corresponding to the two lower levels described in §3.1. Both models are implemented as artificial neural networks with autoencoder-like architectures. In all the experiments here presented, the training samples are generated through a customized simulation of road traffic, realized with the 3D computer graphics software Blender. In this phase of the project the availability of a customized dataset is precious, for the most flexible control of the training set composition, with respect to parameters such as ego-velocity range, range of velocity of the other cars, range of road bending radius, complexity of the environment scenario, and so on. For these sort of purposes Blender is often the software of choice, thanks to its flexible programmability (Mayer et al., 2016; Biedermann et al., 2016). Fig. 2 shows the road track used in the experiments.

The first neural network here presented corresponds to the lowest level model, its divergent path produces a prediction in visual space reconstructing the same color image received as input. The architecture, shown in Fig. 3, is composed of a stack of convolutional layers, followed by flat layers, then a symmetric stack of deconvolutional layers. There is a clear discrepancy between the physical structure of biological CDZs and this model. In the CDZs the same neural assemblies are able to compute the forward direction (acting as convergent processors) and the backward direction (when acting as divergent processors). In our model there are two distinct blocks: a stack of convolutions working as convergent processors, and a stack of deconvolutions working as divergent processors. However, the similarity between our model and Damasio's CDZs is preserved from a computational point of view, as the structure of each convolution in the stack is specular to the corresponding deconvolution transformation in the second stack, and both transformations derive their kernel parameters from learning on the same image samples. As stated in §3.1, this implementation is purely synchronous, without temporal delay between convergence and divergence, therefore there is no need for proprioceptive input, in addition to the visual one. The autoencoder was trained on a dataset of 100,000 images generated in Blender, with 10% of samples used as validation set. We adopted Adam as gradient-based optimizer (Kingma and Ba, 2014), and the mean squared error as loss function. The final loss obtained was 0.0025, computed on the test set.

The second model aims at diverging into a space which is still retinotopically bounded, but with neural activation coding for "concepts". As described in §3.1 we take into account the concepts of *cars* and *lanes*. Each concept has its own corresponding divergence path in the network, while the convergence pathway is common and is the same of the previous model, since it shares the same basic visual features. The model is depicted in Fig. 4. The innermost layer can be seen as a compact representation of the scene, made by 384 neurons, disentangled into three partially distinct classes: visual representations irrespective of concepts, representations selective for *car* entities, and representations selective for *lane* entities. Each of the disentangled representations is made of 128 neurons. Note that there is no special architectural design for disentangling the *car* and *lane* con-

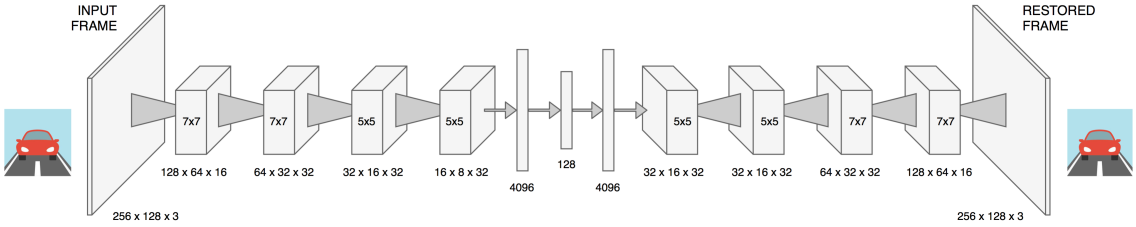

Figure 3: Scheme of the neural network implementing the lower level CDZ model.

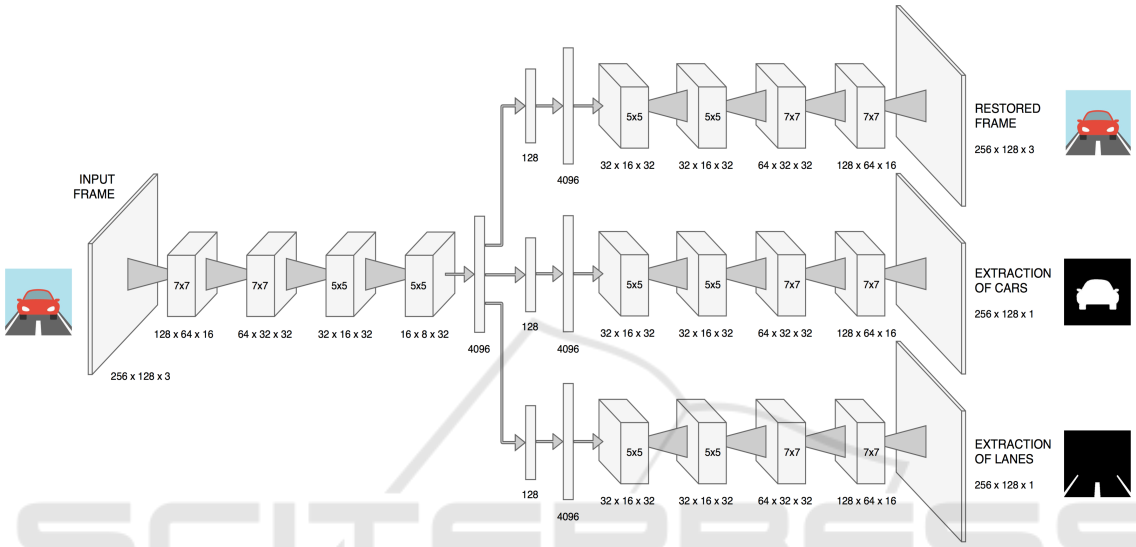

Figure 4: Scheme of the neural network implementing the intermediate level CDZ model.

cepts, the only difference is in the training regimes upon which the different divergence pathways were trained. In the case of concepts, the target output of each divergent pathway is a binary image with true values signaling pixels belonging to the concept at hand, as shown on the right of Fig. 4. Being the target pixels Boolean values, the loss function is the cross-entropy. Since there is a large imbalance of pixels that do not belong to either concepts – with respect to pixels that do belong to – the cross entropy is weighted to tackle class imbalance (Sudre et al., 2017). In our formulation the loss $\mathcal{L}$ of a prediction of the model $\hat{\mathbf{y}}$ against a ground truth $\mathbf{y}$ is the following:

$$\mathcal{L}(\mathbf{y},\hat{\mathbf{y}}) = -\frac{1}{N}\sum_{i}^{N}\left(p(y_i)\,q(y_i,\hat{y}_i)\right) \quad (1)$$

$$p(y) = (1-P)y + P(1-y) \quad (2)$$

$$q(y,\hat{y}) = y\log\hat{y} + (1-y)\log(1-\hat{y}) \quad (3)$$

$$P = \left(\frac{1}{M}\sum_{j}^{M} y_j\right)^{\frac{1}{k}} \quad (4)$$

where N is the number of pixels in an image, M is the number of all pixels in the training dataset, and P is the ratio of true value pixels over all the pixels in the dataset. The parameter k is used to smooth the effect of weighting by the probability of ground truth, a value evaluated empirically as valid is 4.

Although the two conceptual divergence pathways are trained separately, several of the training input samples are common, while the target outputs are different, depending on the class of concept. This procedure bears resemblance with the work of (Kulkarni et al., 2015), where groups of neurons in an inner layer of a CNN model have been "encouraged" (in the Authors' words) to learn separate representations. In the case of Kulkarni and co-workers the disentangled representations are classes of graphic primitives, such as poses or lightnings, while in our case the disentangled representations are for *car* and *lane* entities.

Fig. 5 shows prediction of the two implemented models, on two input samples (leftmost pictures). The results of the lowest level model are shown in the central pictures. It is well visible how the outcome of this model is fairly faithful with respect to the overall scene, including the far landscape. It is, however, scarcely sensible to the features that change in time faster than the surround, and appear more rarely compared to other features. This is exactly the case of other cars, some of which disappear almost com-

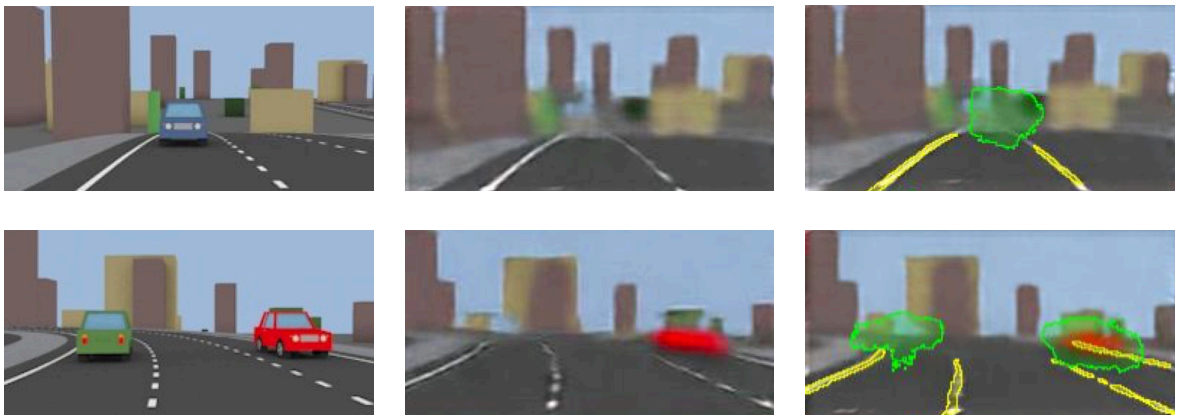

Figure 5: Results of the models' predictions: on the left the original frames; on the center the outputs of the model diverging in visual space; on the right the frames predicted by the intermediate level model, with *lane* entities highlighted in yellow and *car* entities highlighted in green.

pletely in the samples shown in Fig. 5.

The results of the CDZ model in conceptual space are shown in the rightmost pictures. The output of the two conceptual divergent paths are merged into single images for better visibility, the green overlays are the output of the *car* divergent path, and the yellow overlays are the output of the *lane* divergent path (the background image is the same output of the lower divergent path). Note that the true aim of our model is not to produce a semantic segmentation of the input images, but to induce the model to learn disentangled representations of the main conceptual features fundamental to the driving task. The resulting images nicely show how the projection of the sensorial input (original frames) into conceptual representation is very effective in identifying and preserving the sensible features of *cars* and *lanes*, even in the situations when the lowest level model failed to capture them, like in the case of cars moving at a high speed.

5 CONCLUSIONS

Following the outstanding achievements of deep learning, here we presented an artificial neural network model inspired by the convergence-divergence zones architecture proposed by Meyer and Damasio. Our solution adopts an autoencoder-like architecture, and we exploit its known generative capabilities for mimicking mental imagery, *i.e.* the feedback projections in the CDZ. Despite the autonomous driving focus of the paper, the overall approach could, in principle, be extended to the broader field of robotics by adapting the inner levels of the model to learn the representations of motor commands intended for the specific agent. The architecture developed is pretty flexible, in the sense that our framework can be extended to simulate other complex human motor abilities, as supported by the logical evidences of the CDZ hypothesis.

Our future plans involve the finalization of the higher level model of the architecture which computes motor commands from the conceptual representation of the environment presented in this work.

ACKNOWLEDGEMENTS

This work was developed inside the EU Horizon 2020 Dreams4Cars Research and Innovation Action project, supported by the European Commission under Grant 731593.

REFERENCES

Bengio, Y. (2009). Learning deep architectures for AI. *Foundation and Trends in Machine Learning*, 2:1–127.

Bertolazzi, E., Biral, F., and Da Lio, M. (2003). Symbolic-numeric efficient solution of optimal control problems for multibody systems. *Journal of Computational and Applied Mathematics*, 185:404–421.

Biedermann, D., Ochs, M., and Mester, R. (2016). Evaluating visual ADAS components on the COnGRATS dataset. In *IEEE Intelligent Vehicles Symposium*, pages 986–991.

Bojarski, M., Yeres, P., Choromanaska, A., Choromanski, K., Firner, B., Jackel, L., and Muller, U. (2017). Explaining how a deep neural network trained with end-to-end learning steers a car. *CoRR*, abs/1704.07911.

Chui, M., Manyika, J., Miremadi, M., Henke, N., Chung, R., Nel, P., and Malhotra, S. (2018). Notes from the AI

frontier: insights from hundreds of use cases. Technical Report April, McKinsey Global Institute.

Colder, B. (2011). Emulation as an integrating principle for cognition. *Frontiers in Human Neuroscience*, 5:Article 54.

Da Lio, M., Plebe, A., Bortoluzzi, D., Rosati Papini, G. P., and Donà, R. (2018). Autonomous vehicle architecture inspired by the neurocognition of human driving. In *International Conference on Vehicle Technology and Intelligent Transport Systems*, pages 507–513. Scitepress.

Damasio, A. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, 33:25–62.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 249–256.

Grillner, S. and Wallén, P. (2004). Innate versus learned movements – a false dichotomy. *Progress in Brain Research*, 143:1–12.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Science*, 27:377–442.

Hazelwood, K., Bird, S., Brooks, D., Chintala, S., Diril, U., Dzhulgakov, D., Fawzy, M., Jia, B., Jia, Y., Kalro, A., Law, J., Lee, K., Lu, J., Noordhuis, P., Smelyanskiy, M., Xiong, L., and Wang, X. (2018). Applied machine learning at Facebook: A datacenter infrastructure perspective. In *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 620–629.

Hesslow, G. (2012). The current status of the simulation theory of cognition. *Brain*, 1428:71–79.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 28:504–507.

Jeannerod, M. (2001). Neural simulation of action: A unifying mechanism for motor cognition. *NeuroImage*, 14:S103–S109.

Jones, W., Alasoo, K., Fishman, D., and Parts, L. (2017). Computational biology: deep learning. *Emerging Topics in Life Sciences*, 1:136–161.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. In *Proceedings of International Conference on Learning Representations*.

Kosslyn, S. M. (1994). *Image and Brain: the Resolution of the Imagery Debate*. MIT Press, Cambridge (MA).

Krizhevsky, A. and Hinton, G. E. (2011). Using very deep autoencoders for content-based image retrieval. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 489–494.

Kulkarni, T. D., Whitney, W. F., Kohli, P., and Tenenbaum, J. B. (2015). Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*, pages 2539–2547.

Larochelle, H., Bengio, Y., Louradour, J., and Lamblin, P. (2009). Exploring strategies for training deep neural networks. *Journal of Machine Learning Research*, 1:1–40.

Li, J., Cheng, H., Guo, H., and Qiu, S. (2018). Survey on artificial intelligence for vehicles. *Automotive Innovation*, 1:2–14.

Liu, D. and Todorov, E. (2007). Evidence for the flexible sensorimotor strategies predicted by optimal feedback control. *Journal of Neuroscience*, 27:9354–9368.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26.

Mahon, B. Z. and Caramazza, A. (2011). What drives the organization of object knowledge in the brain? the distributed domain-specific hypothesis. *Trends in Cognitive Sciences*, 15:97–103.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, 58:25–45.

Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., and Brox, T. (2016). A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 4040–4048.

Meyer, K. and Damasio, A. (2009). Convergence and divergence in a neural architecture for recognition and memory. *Trends in Neuroscience*, 32:376–382.

Moulton, S. T. and Kosslyn, S. M. (2009). Imagining predictions: mental imagery as mental emulation. *Philosophical transactions of the Royal Society B*, 364:1273–1280.

NHTSA (2017). Fatality Analysis Reporting System (FARS).

Olier, J. S., Barakova, E., Regazzoni, C., and Rauterberg, M. (2017). Re-framing the characteristics of concepts and their relation to learning and cognition in artificial agents. *Cognitive Systems Research*, 44:50–68.

Ras, G., van Gerven, M., and Haselager, P. (2018). Explanation methods in deep learning. In Escalante, H. J., Escalera, S., Guyon, I., Baró, X., Güçlütürk, Y., Güçlü, U., and van Gerven, M., editors, *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer-Verlag, Berlin.

Samek, W., Wiegand, T., and Müller, K. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *CoRR*, abs/1708.08296.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Schwarting, W., Alonso-Mora, J., and Rus, D. (2018). Planning and decision-making for autonomous vehicles. *Annual Review of Control, Robotics, and Autonomous Systems*, 1:8.1–8.24.

Seger, C. A. and Miller, E. K. (2010). Category learning in the brain. *Annual Review of Neuroscience*, 33:203–219.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Cardoso, M. J. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In Cardoso, J., Arbel, T., Carneiro, G., Syeda-

Mahmood, T., Tavares, J. M. R., Moradi, M., Bradley, A., Greenspan, H., Papa, J. P., Madabhushi, A., Nascimento, J. C., Cardoso, J. S., Belagiannis, V., and Lu, Z., editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248.

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

Wolpert, D. M., Diedrichsen, J., and Flanagan, R. (2011). Principles of sensorimotor learning. *Nature Reviews Neuroscience*, 12:739–751.

Wolpert, D. M. and Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, 11:1317–1329.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS