Neurocognitive–inspired Approach for Visual Perception in Autonomous Driving

Alice Plebe[1][0000−0001−8567−0553] and Mauro Da Lio[2][0000−0002−6619−9484]

[1] Dept. of Information Engineering and Computer Science, University of Trento, Italy
[2] Dept. of Industrial Engineering, University of Trento, Italy
{alice.plebe,mauro.dalio}@unitn.it

Abstract. Since the last decades, deep neural models have been pushing forward the frontiers of artificial intelligence. Applications that in the recent past were considered no more than utopian dreams, now appear to be feasible. The best example is autonomous driving. 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. While the early forms of artificial neural networks were aimed at simulating and understanding human cognition, contemporary deep neural networks are totally indifferent to cognitive studies, they are designed with pure engineering goals in mind. Several scholars, we included, 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. More specifically, we will point to four neurocognitive theories: the simulation theory of cognition; the Convergence-divergence Zones hypothesis; the transformational abstraction hypothesis; the free-energy predictive theory. Our proposed model tries to combine a number of existing algorithms that most closely resonate with the assumptions of these four neurocognitive theories.

Keywords: deep learning · autonomous driving · convergence–divergence zones · variational autoencoder · free energy

1 INTRODUCTION

Artificial neural networks are responsible for the current fast resurgence of Artificial Intelligence, after several decades of slow and unsatisfactory advances, and are now at the very heart of many technology developments [54,7,23]. They have proved to be the best available approach for a variety of different problem domains [40,31], and the design of autonomous vehicles is definitely one of the research areas to have amply benefited from this technology [2,39,55].

Actually, this success was totally unexpected because the technology development has been very little, with only relative minor improvements from a field that was stagnating at the beginning of the century. Artificial neural network found their way during the '80s, with the *Parallel Distributed Processing* (PDP) project [53]. The success of PDP was largely due to an efficient learning rule, known as *backpropagation*, which adapts the connections between units through input/output samples of the desired function. Geoffrey Hinton was one of the protagonists of the PDP project [26], who contributed specifically to the introduction of the backpropagation [52]. However, after a couple of decades, artificial neural networks exhausted their potential. Once again, Hinton gave them a new boost [27], by refining the old backpropagation method. With just relatively small advances in the learning algorithm, he made possible to train networks with more layers of neurons than before, called *deep learning* models ever since.

The most distinctive difference between the PDP generation of neural networks and current deep learning is in their scope. For the PDP project the scope was clearly indicated in the title of its main book [53]: "Explorations in the Microstructure of Cognition". On the contrary, the majority of modelers in deep learning is totally indifferent to cognition, and their scope is purely on engineering goals. In a first phase, adopting mathematical solutions alien to mental processes has been certainly a key of the success of deep learning. However, several scholar are now arguing that the segregation between the deep learning and the neurocognitive communities would be dangerous and detrimental for a further progress [41,22,60].

We agree to this position, and we deem that looking at neurocognitive facts would be especially valuable in applications such as driving. The reason is that driving is the sort of behavior for which neurocognition has changed its paradigm since the PDP era. Cognitive science of the '80s was dominated by the modular perspective [11], which divided sharply intellectual tasks such as language and categorization, from low level tasks such as vision and motor control. Since the beginning of this century cognitive science has witnessed a radical methodological revolution, often summarized as "4E cognition" [44]: embodied, embedded, enactive, and extended. One of the assumptions of 4E cognition that is most relevant to our case is that perception and action are constitutively intertwined, and not only they are contiguous to intellectual behavior, sensorimotor simulations are the basic founding of cognition as a whole. This structure of cognition is the source of the amazing human abilities of learning new forms of sensorimotor control, like driving. We believe this is a reason why none of the current available implementations of autonomous vehicles can claim to be nowhere close to the driving performance of a human being.

Such considerations lead us to reflect if it is possible to take inspiration from the mechanisms whereby the brain learns to perform such complex sensorimotor behaviors. With this paper, we propose to exploit the current most established neurocognitive theories as inspiration for designing more brain–like neural network models. In the next section we will review a number of selected neurocognitive theories relevant to our purposes. In section §3 we will show that

it is possible to find neural algorithms that best approximate the assumptions of such theories, and we will describe how our model put together all these algorithms. Finally, in section §4 we will present the results of applying our neural model 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 [47,46].

2 THE NEUROCOGNITIVE POINT OF VIEW

Humans are able to learn an impressive range of different and 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 [19]. 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 [64].

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 science has been produced in the past decades, which allows us to grasp some useful principles 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 [30,24]. 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.

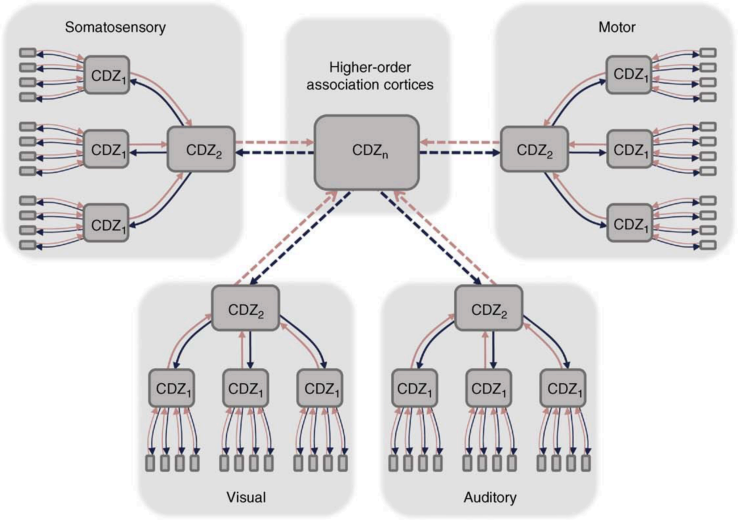

Fig. 1. Schematic representation of Damasio's CDZ framework [42, Fig. 1]. Neuron ensembles in early sensorimotor cortices of different modalities send converging forward projections (red arrows) to higher-order association cortices, which, in turn, project back divergently (black arrows) to the early cortical sites, via several intermediate steps.

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 [35,43].

2.2 Convergence-divergence Zones

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

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

A prominent proposal in this direction is the formulation of the convergence-divergence zones (CDZs) [42]. They derive from an earlier model [8] 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 formulation.

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 activation 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 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.

2.3 Transformational Abstraction

One of the 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 [56]. As highlighted by [45] CDZs are a valid systemic candidate for how the formation of concepts takes place at brain level. However, the idea of CDZ is just sketched and cannot provide a detailed mechanism for conceptual abstractions.

According to the historical empiricist tradition, conceptual abstractions is derived from experience, mostly perceptual experience. This direction fits perfectly with the approach implemented by artificial neural networks. Still, a difficulty with acquiring even moderately abstract categories lies in the mutually inconsistent manifestations of the characteristic features of a category, in each of its real exemplars. In visual data, for example, object translation, rotation, motion in depth, deformation and lighting changes can drastically entangle features of objects belonging to the same category. Conversely, the perceptual appearance of two unrelated objects, like a close flying insect and a far distant vulture, can be very similar. A suggested solution to this difficult issue is in the transformational abstraction [22,5] performed by a hierarchy of cortical operations, as in the ventral visual cortex [49]. The essence of transformational abstraction, from a mathematical point of view, should lie in the combination of two operations: linear convolutional filtering and nonlinear downsampling. Operations of this sort have been identified in the primary visual cortex [28,29,17], and the staking of this process in hierarchy is well recognized in the primate ventral visual path [10,12,61].

Note that transformational abstraction is one of the possible interpretation of the convergent zone in the CDZ theory, even if it lacks a specification of the divergent counterpart, which can lead from the abstraction of a concept to its use during mental imagination. Transformational abstraction is conceived as a general road to conceptual abstractions, nevertheless, it is highly relevant in the case of driving. As will be discussed in §3.4, during the drive, the sensorimotor control relies heavily on a small number of known concepts, abstracted from visual space.

2.4 The Predictive Theory

The reason why cognition is mainly explicated as simulation, according to Hesslow or Jeannerod, is because the brain through simulation can achieve the most precious information of an organism: a prediction of the state of affairs in the environment in the future. The need of predicting, and how it mold the entire cognition, has become the core of a different, but related, theory which has gained large attention in the last decade, made popular under the term "Bayesian brain", "predictive brain", or "free-energy principle for the brain". The leading figure of this theory is Karl Friston [13,14]. According to Friston 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. This mathematical form is derived from the framework of variational Bayesian methods in statistical physics, where the intractable problem of inferring the posterior distribution over a random variable is approximated by a different, and more tractable, auxiliary distribution [63]. We will see in §3.3 how the same probabilistic framework will be used in the derivation of a deep neural model. The basic form of the free-energy under the variational Bayesian framework is borrowed by Friston for abstract entities of cognitive value. For example, this is his free-energy formulation in the case of perception [15, p.427]:

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

where $\boldsymbol{x}$ is the sensorial input of the organism, $\boldsymbol{c}$ is the collection of the environmental causes producing $\boldsymbol{x}$, $\boldsymbol{a}$ are actions that act on the environment to change sensory samples, and $\boldsymbol{z}$ are inner representations of the brain. The quantity $\check{p}(\boldsymbol{c}|\boldsymbol{z})$ is the encoding in the brain of the estimate of causes of sensorial stimuli. The quantity $p(\boldsymbol{c}|\boldsymbol{x},\boldsymbol{a})$ is the conditional probability of sensorial input conditioned by the actual environmental causes $\boldsymbol{c}$. 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 $\boldsymbol{z}$. In the case of action the free energy formulation by Friston becomes [15, p.428]:

$$F_A = \Delta_{\mathrm{KL}}\Big(\check{p}(\boldsymbol{c})\|p(\boldsymbol{c})\Big) - \log p(\boldsymbol{x}|\boldsymbol{c},\boldsymbol{a}) \tag{2}$$

and optimizes $\boldsymbol{a}$.
All formalization in equations (1) and (2) are just abstract, without details on how the variables can be explicitate, and how the equations can be solved.

3 ARTIFICIAL MENTAL IMAGERY

Over the years the CDZ hypothesis has found 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 a number of independent notions, introduced in the field of artificial intelligence for different purposes, which bear significant similarities with the CDZ scheme. Here we will first introduce these notions independently, then we will describe our model which, by taking together these pieces, builds up a neural architecture inspired by CDZs.

3.1 Convergence–divergence in the 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 [25], but more recently has been the cornerstone of the evolution from shallow to deep neural architectures [27,62]. The crucial issue of training neural architectures with multiple internal layers was initially solved associating each internal layer with a Restricted Boltzmann Machine [27], so that they can be pre-trained individually in unsupervised manner. The adoption of autoencoders overcame 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 [38]. In the first layer the inputs are that of the entire neural model, and 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 [1], or even a better one [62].

Soon after, the refinement of algorithms for initialization [18] and optimization [33] of weights, made any type of unsupervised pre-training method superfluous. However, autoencoders continue to play their basic role for capturing compact information from high dimensional data, and their use within this scope has been expanded. In this kind of models the task to be solved by the network is to simulate as output the same data fed as input. The advantage is that while learning to reconstruct the input information, the model develops a very compact

internal representation of the input space. The basic structure of an autoencoder, independently of the details of the various implementations, is composed of two neural models:

$$f_\Theta : \mathcal{Z} \to \mathcal{X} \tag{3}$$

$$g_\Phi : \mathcal{X} \to \mathcal{Z} \tag{4}$$

The first one is the *decoder*, often called the *generative* model, which reconstructs high dimensional data $\boldsymbol{x} \in \mathcal{X}$ taking as input low dimensional compact representations $\boldsymbol{z} \in \mathcal{Z}$. The model is fully fixed by the set of parameters Θ. The model in equation (4) is called *encoder* and computes the compact representations $\boldsymbol{z} \in \mathcal{Z}$ of a high dimensional input $\boldsymbol{x} \in \mathcal{X}$. This model is determined by its set of parameters Φ. In the autoencoder's architecture the parameters Θ and Φ are learned by minimizing the error between input samples $\boldsymbol{x}_i$ and the outputs $f(g(\boldsymbol{x}_i))$. 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.

3.2 Convergence–divergence as Convolution–deconvolution

Let us now dive into detail of how convergence can be achieved inside autoencoders. The most common way is stacking feed-forward layers with decreasing number of units. There is, however, an interesting alternative closely related to the transformational abstraction hypothesis described in §2.3: the *deep convolutional neural networks* (DCNNs). One again, this architecture was introduced by Hinton [36], by adapting an old model of the PDP era, called *Neocognitron* [16], into a deep architecture. The DCNN implements the hierarchy of convolutional filtering alternated with nonlinear downsampling, and it is considered the essence of transformational abstraction. The old Neocognitron of Fukushima alternates layers of *S-cell* type units with *C-cell* type units, which naming are evocative of the classification in simple and complex cells by Hubel and Wiesel [28,29]. The S-units act as convolution kernels, while the C-units downsample the images resulting from the convolution, by spatial averaging. The crucial difference from conventional convolutions in image processing [51,4] is that now the kernels are learned. The DCNNs of Hinton and co-workers are "deep" versions of Neocognitron, using several layers of convolutions, each with a large number of different kernels, typically tens of millions.

DCNNs do not only resonate with the theoretical proposal of transformational abstraction, there is a growing evidence of striking analogies between patterns in DCNN models and patterns of voxels in the brain visual system. One of the first attempt to relate results of deep learning with the visual system was based on the idea of adding at a given level of an artificial network model a layer predicting in the space of voxel response, and to train this level on sets of images and corresponding fMRI responses [20]. Using this method, a DCNN model [6] was compared with fMRI data [21]. Initially, subjects were presented with 1750 natural images and voxel responses in progressively downstream areas – from

V1 up to LO (*Lateral Occipital Complex*) – were recorded. The same images were presented to the model, and the output of the convolutional layers were trained – with a simple linear predictor – to predict voxel patterns. As a result, DCNN model responses were predictive of the voxels in the visual cortex above chance, with good prediction accuracy especially in the lower visual areas. This first unexpected result was immediately followed by several other studies, using variants of the same technique [32,9,58], finding reasonable agreement between features computed by DCNN models and fMRI data.

DCNNs are therefore a highly biologically plausible implementation for the convergence zone in CDZs, at least in the case of visual information. Convolutional neural models do not include a divergence counterpart, typically the outputs of the last convolutions are fed into ordinary feed forward layers to produce a classification. This gap was filled with the *deconvolutional* neural networks [66,67,65], performing alternation of unpooling and linear filtering. Each step of these two operations reconstruct a higher level of spatial dimension of the data, up to the full high dimension of the original image. Note that the nonlinear downsampling done in DCNNs is a non invertible operation, therefore it is not possible to reconstruct faithfully the upsized representations by unpooling. Zeiler and co-workers circumvented the problem by saving additional information during the poolings done in the convolution stages for exploitation during unpooling, but this strategy is clearly non biological plausible. However, the stacked combination of deconvolution and unpooling is the current neural implementation more close to the idea of divergence zone of CDZs.

3.3 Variational Bayes and Autoencoders

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 [34,48] 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 including such idea have been proposed over the years, see [59] for a review.

We will show in a while that the adoption of variational inference lead to a mathematical formulation impressively similar to the concept of free energy in Friston. This close analogy went unnoticed by all the main developers of variational autoencoder. It is not so surprising because mainstream deep learning is driven by engineering goals without any interest in connections with cognition. Within the philosophy of the Dreams4Car project, illustrated in the Introduction §1, the strong connection between a well established cognitive theory and a computational solution, greatly argues in favor of adopting such solution.

The variational inference framework takes up the issue of approximating the probability distribution $p(\boldsymbol{x})$ of a high dimensional random variable $\boldsymbol{x} \in \mathcal{X}$, such as the visual scene that hits the retina of a living agent, or the camera of an artificial agent. A candidate in approximating the real unknown probability

distribution is a neural network such as that in equation (3). The neural network by itself is deterministic, but its output distribution can be easily computed as follows:

$$p_\Theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}\left(\boldsymbol{x}|f_\Theta(\boldsymbol{z}), \boldsymbol{\sigma}^2\mathbf{I}\right) \tag{5}$$

where $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma})$ is the Gaussian function in $\boldsymbol{x}$, with mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$. Using equation (5) it is now possible to express $p_\Theta(\boldsymbol{x})$, which is the desired approximation of $p(\boldsymbol{x})$:

$$p_\Theta(\boldsymbol{x}) = \int p_\Theta(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} = \int p_\Theta(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z} \tag{6}$$

It is immediate to recognize that the kind of neural network performing the function $f_\Theta(\cdot)$ is exactly the decoder part in the autoencoder, corresponding to the divergence zone in the CDZ neurocognitive concept. In the case when $\mathcal{X}$ is the domain of images, $f_\Theta(\cdot)$ comprises a first layer that rearranges the low-dimension variable $\boldsymbol{x}$ in a two dimensional geometry, followed by a stack of deconvolutions, up to the final geometry of the $\boldsymbol{x}$ images.

In equation (6) there is clearly no clue on what the distribution $p(\boldsymbol{z})$ might be, but the idea behind variational autoencoder is to introduce an auxiliary distribution q from which to sample $\boldsymbol{z}$, and it is made by an additional neural network. Ideally this model should provide the posterior probability $p_\Theta(\boldsymbol{z}|\boldsymbol{x})$, which is unknown. This second neural model is of the kind in equation (4), and as done in equation (5), its probability distribution is:

$$q_\Phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}|g_\Phi(\boldsymbol{x}), \boldsymbol{\sigma}^2\mathbf{I}\right) \tag{7}$$

The model $f_\Theta(\cdot)$ functions as decoder, and $g_\Phi(\cdot)$ is the encoder part in the autoencoder, projecting the high dimensional variable $\boldsymbol{x}$ into the low dimensional space $\mathcal{Z}$. It continues to play the role of the convergence zone in the CDZ idea. The measure of how well $p_\Theta(\boldsymbol{x})$ approximates $p(\boldsymbol{x})$ for a set of $\boldsymbol{x}_i \in \mathcal{D}$ sampled in a dataset $\mathcal{D}$ is given by the log-likelihood:

$$\ell(\Theta|\mathcal{D}) = \sum_{\boldsymbol{x}_i \in \mathcal{D}} \log \int p_\Theta(\boldsymbol{x}_i|\boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z} \tag{8}$$

This equation cannot be solved because of the unknown $p(\boldsymbol{z})$, and here comes the help of the auxiliary probability $q_\Phi(\boldsymbol{z}|\boldsymbol{x})$. Each term of the summation in equation (8) can be rewritten as follows:

$$\begin{aligned}
\ell(\Theta|\boldsymbol{x}) &= \log \int p_\Theta(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} \\
&= \log \int \frac{p_\Theta(\boldsymbol{x}, \boldsymbol{z}) q_\Phi(\boldsymbol{z}|\boldsymbol{x})}{q_\Phi(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z} \\
&= \log \mathbb{E}_{\boldsymbol{z} \sim q_\Phi(\boldsymbol{z}|\boldsymbol{x})} \left[\frac{p_\Theta(\boldsymbol{x}, \boldsymbol{z})}{q_\Phi(\boldsymbol{z}|\boldsymbol{x})} \right]
\end{aligned} \tag{9}$$

where in the last passage we used the expectation operator $\mathbb{E}[\cdot]$. Being the log function concave, we can now apply Jensen's inequality:

$$\begin{aligned}\ell(\Theta, \Phi|\boldsymbol{x}) &= \log \mathbb{E}_{\boldsymbol{z}\sim q_\Phi(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{p_\Theta(\boldsymbol{x},\boldsymbol{z})}{q_\Phi(\boldsymbol{z}|\boldsymbol{x})}\right] \\ &\geq \mathbb{E}_{\boldsymbol{z}\sim q_\Phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\Theta(\boldsymbol{x},\boldsymbol{z})\right] - \mathbb{E}_{\boldsymbol{z}\sim q_\Phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log q_\Phi(\boldsymbol{z}|\boldsymbol{x})\right] \end{aligned} \tag{10}$$

Since the derivation in (10) is smaller or at least equal to $\ell(\Theta|\boldsymbol{x})$, it is called the *variational lower bound*, or *evidence lower bound* (ELBO). Note that now in $\ell(\Theta, \Phi|\boldsymbol{x})$ there is also the dependency from the parameters Φ of the second neural network defined in (7).
It is possible to rearrange further $\ell(\Theta, \Phi|\boldsymbol{x})$ in order to have $p_\Theta(\boldsymbol{x}|\boldsymbol{z})$ instead of $p_\Theta(\boldsymbol{x}, \boldsymbol{z})$ in equation (10), moreover, we can now introduce the *loss function* $\mathcal{L}(\Theta, \Phi|\boldsymbol{x})$ as the value to be minimized in order to maximize ELBO:

$$\begin{aligned}\mathcal{L}(\Theta, \Phi|\boldsymbol{x}) &= -\ell(\Theta, \Phi|\boldsymbol{x}) \\ &= -\int q_\Phi(\boldsymbol{z}|\boldsymbol{x}) \log \frac{p_\Theta(\boldsymbol{x}, \boldsymbol{z})}{q_\Phi(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z} \\ &= -\int q_\Phi(\boldsymbol{z}|\boldsymbol{x}) \log \frac{p_\Theta(\boldsymbol{x}|\boldsymbol{z})p_\Theta(\boldsymbol{z})}{q_\Phi(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z} \\ &= \Delta_{\mathrm{KL}}\big(q_\Phi(\boldsymbol{z}|\boldsymbol{x})\|p_\Theta(\boldsymbol{z})\big) - \mathbb{E}_{\boldsymbol{z}\sim q_\Phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\Theta(\boldsymbol{x}|\boldsymbol{z})\right] \end{aligned} \tag{11}$$

where the last step uses the Kullback-Leibler divergence Δ_{KL}. By comparing equation (11) with the formulation of the free-energy principle by Friston (1) their coincidence appears immediately.

The formulation in (11) seems to be still intractable, because contains the term $p_\Theta(\boldsymbol{z})$, but there is a simple analytical formulation of the Kullback-Leibler divergence in the Gaussian case (see appendix B in [34]):

$$\Delta_{\mathrm{KL}}\Big(q_\Phi(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})\Big) = -\frac{1}{2}\sum_{i=1}^{Z}\left(1 + \log\left(\sigma_i^2\right)\right) - \mu_j^2 - \sigma_i^2\big) \tag{12}$$

where μ_i and σ_i are the i-th components of the mean and variance of $\boldsymbol{z}$ given by $q_\Phi(\boldsymbol{z}|\boldsymbol{x})$.

3.4 A CDZ-like Model for Driving

We have reviewed several components that match quite closely the relevant neurocognitive theories identified in §2. The model we propose attempts to weave together these components, finalized at visual perception in autonomous driving agents. There is a range of different levels at which we can design such models. Similarly to the hierarchical arrangement of CDZs in the brain, as described by Meyer and Damasio (again, Fig.1), variational autoencoder models embedding convolution and deconvolution operations can be placed at a level depending on the relevant representational space.

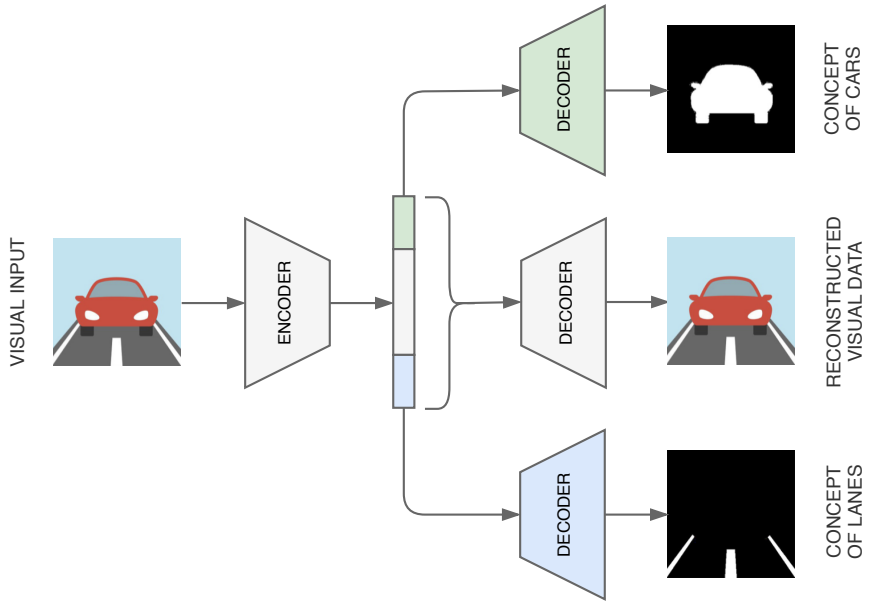

Fig. 2. The architecture of our model. The variational autoencoder has an encoder compressing an RGB image to a compact high-feature representation. Then 3 decoders map different part of the latent space back to separated output spaces: the decoder on the center outputs into the same visual space of the input; the other two decoders project into conceptual space, producing binary images containing, respectively, `car` entities and `lane marking` entities.

In the context of Dreams4Cars, we considered as a fundamental level of model design the processes that start from the raw image data and converge up to a low-dimension representation of 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. As discussed in §2.3, 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 here presented we choose to consider the two main concepts of `cars` and `lane markings`.

As depicted in Fig. 2, the model is composed by one shared encoder and three independent decoders:

$$f_{\Theta_{\mathrm{V}}} : \mathcal{Z} \to \mathcal{X} \tag{13}$$
$$f_{\Theta_{\mathrm{C}}} : \mathcal{Z}_{\mathrm{C}} \to \mathcal{X}_{\mathrm{C}} \tag{14}$$
$$f_{\Theta_{\mathrm{L}}} : \mathcal{Z}_{\mathrm{L}} \to \mathcal{X}_{\mathrm{L}} \tag{15}$$
$$g_{\Phi} : \mathcal{X} \to \mathcal{Z} \tag{16}$$

where the subscript V denotes visual space, the subscripts C and L are for the `cars` and `lane markings` concepts, respectively. For a vector in the latent space it holds:

$$\boldsymbol{z} \in \mathcal{Z} = [\boldsymbol{z}_{\mathrm{C}}, \widetilde{\boldsymbol{z}}, \boldsymbol{z}_{\mathrm{L}}] \tag{17}$$
$$\mathcal{Z} = \mathbb{R}^{N_{\mathrm{V}}}$$
$$\mathcal{Z}_{\mathrm{C}} = \mathbb{R}^{N_{\mathrm{C}}}$$
$$\mathcal{Z}_{\mathrm{L}} = \mathbb{R}^{N_{\mathrm{L}}}$$
$$N_{\mathrm{C}} = N_{\mathrm{L}} < \frac{N_{\mathrm{V}}}{2} \tag{18}$$

In the first expression $\boldsymbol{z}_{\{\mathrm{C},\mathcal{L}\}}$ are the two segments inside the latent vector $\boldsymbol{z}$ representing the `car` and `lane` concepts, respectively. The segment in between, $\widetilde{\boldsymbol{z}}$, encodes generic visual features, and the entire latent vector $\boldsymbol{z}$ represents in visual space. 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. There is no reason for using different latent space size for the concepts `car` and `lane`, as expressed in equation (18), where the inequality reflects the partitioning of $\boldsymbol{z}$ according to equation (17).

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 [37]. In [68] 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 full 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.

By calling $\Theta = [\Theta_{\mathrm{V}}, \Theta_{\mathrm{C}}, \Theta_{\mathrm{L}}]$ the vector of all parameters in the three decoders, the loss functions of the model is derived from the basic equation (11). At each iteration t a random batch $\mathcal{B} \subset \mathcal{D}$ is presented, and the following loss is

computed:

$$\begin{aligned}\mathcal{L}(\Theta,\Phi|\mathcal{B}) = \big(1-(1-k_0)\kappa^t\big)\sum_{\boldsymbol{x}}^{\mathcal{B}} \Delta_{\mathrm{KL}}\big(q_\Phi(\boldsymbol{z}|\boldsymbol{x})\|p_{\Theta_\mathrm{V}}(\boldsymbol{z})\big) \\ - \sum_{\boldsymbol{x}}^{\mathcal{B}} \lambda_\mathrm{V}\mathbb{E}_{\boldsymbol{z}\sim q_\Phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\Theta_\mathrm{V}}(\boldsymbol{x}|\boldsymbol{z})\right] \\ - \sum_{\boldsymbol{x}}^{\mathcal{B}} \lambda_\mathrm{C}\mathbb{E}_{\boldsymbol{z}_\mathrm{C}\sim \Pi_\mathrm{C}(q_\Phi(\boldsymbol{z}|\boldsymbol{x}))}\left[\log \widetilde{p}_{\Theta_\mathrm{C}}(\boldsymbol{x}|\boldsymbol{z}_\mathrm{C})\right] \\ - \sum_{\boldsymbol{x}}^{\mathcal{B}} \lambda_\mathrm{L}\mathbb{E}_{\boldsymbol{z}_\mathrm{L}\sim \Pi_\mathrm{L}(q_\Phi(\boldsymbol{z}|\boldsymbol{x}))}\left[\log \widetilde{p}_{\Theta_\mathrm{L}}(\boldsymbol{x}|\boldsymbol{z}_\mathrm{L})\right] \end{aligned} \tag{19}$$

Few observations are due for the differences between this equation and the basic loss equation (11). First of all, there is a delay in including the contribution of the Kullback-Leibler divergence, because initially the encoder is unlikely to provide any meaningful probability distribution $q_\Phi(\boldsymbol{z}|\boldsymbol{x})$. There is a cost factor for the Kullback-Leibler component, set initially at a small value k_0 and gradually increased up to 1.0, with time constant κ. This strategy was first introduced in the context of variational autoencoders for language modeling [3].

All the components next to the Kullback-Leibler divergence are errors in the recostruction of the imagined scene or the imagined concepts, and their relative contributions are weighted by the parameters $\lambda_{\{\mathrm{V,C,L}\}}$. Their purpose is mainly to normalize the range of the errors, which is quite different from visual to conceptual spaces. For this reason, typically $\lambda_\mathrm{V} \neq \lambda_\mathrm{C} = \lambda_\mathrm{L}$.

The second component of the loss in equation (19) computes the error in visual space, using the entire latent vector $\boldsymbol{z}$, and corresponds precisely to the second component in the basic loss (11). The last two components compute the error in the conceptual space, and are slightly different. Only the relevant portion of the latent vector $\boldsymbol{z}$ is used, by the projection operators $\Pi_{\{\mathrm{C,L}\}}$, where the subscripts C and L are for the `car` and `lane` concepts, as usual. In addition, a variant of the standard cross entropy is used, with the symbols $\widetilde{p}_{\Theta_{\{\mathrm{C,L}\}}}$, in order to account for the large unbalance between the number of pixel belonging to a concept, and all the other pixels, typical of ordinary driving scenes. For each concept, we precompute a coefficient to be applied to the true value class:

$$P = \left(\frac{1}{NM}\sum_{i}^{N}\sum_{j}^{M} y_{i,j}\right)^{\frac{1}{k}} \tag{20}$$

where N is the number of pixels in an image, M is the number of images 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. This strategy has been first introduced in the context of medical image processing [57].

4 RESULTS

Fig. 3. Samples from the SYNTHIA dataset. All images show the same frame of a driving sequence, but under different environmental and lighting conditions. The two left columns show variations of sunny environments, while the two right columns depict settings with low illumination and adverse weather conditions.

We present here a selection of results achieved with an instance of the model described in the previous section. This architecture is described by the parameters shown in Table 1. In our experiments for training and testing the presented model, we adopted the SYNTHIA dataset [50], a large collection of synthetic images representing various urban scenarios. The dataset is realized using the game engine Unity, and it is composed of $\sim$ 100k frames of driving sequences recorded from a simulated camera on the windshield of the ego car. We found this dataset to be well suited for our experiment because, despite being generated in 3D computer graphics, it offers a wide variety of illumination and weather conditions, resulting occasionally in very adverse driving conditions. Each driving sequence is replicated on a set of different environment conditions which includes seasons, weather and time of the day. Fig. 3 gives an example of the variety of data coming from the same frame of a driving sequence. Moreover the urban environment is very diverse as well, ranging from driving on freeways, through tunnels, congestion, "NewYork-like" city and "European" town – as they describe. Overall, this dataset appears to be a nice challenge for our variational autoencoder.

Fig. 4 shows the results of our artificial CDZ model for a set of driving sequences. The images produced by the model 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 yellow overlays are the output of the `lane markers` divergent path. These results 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 `lane markings`, despite the large variations in lighting and environmental conditions.

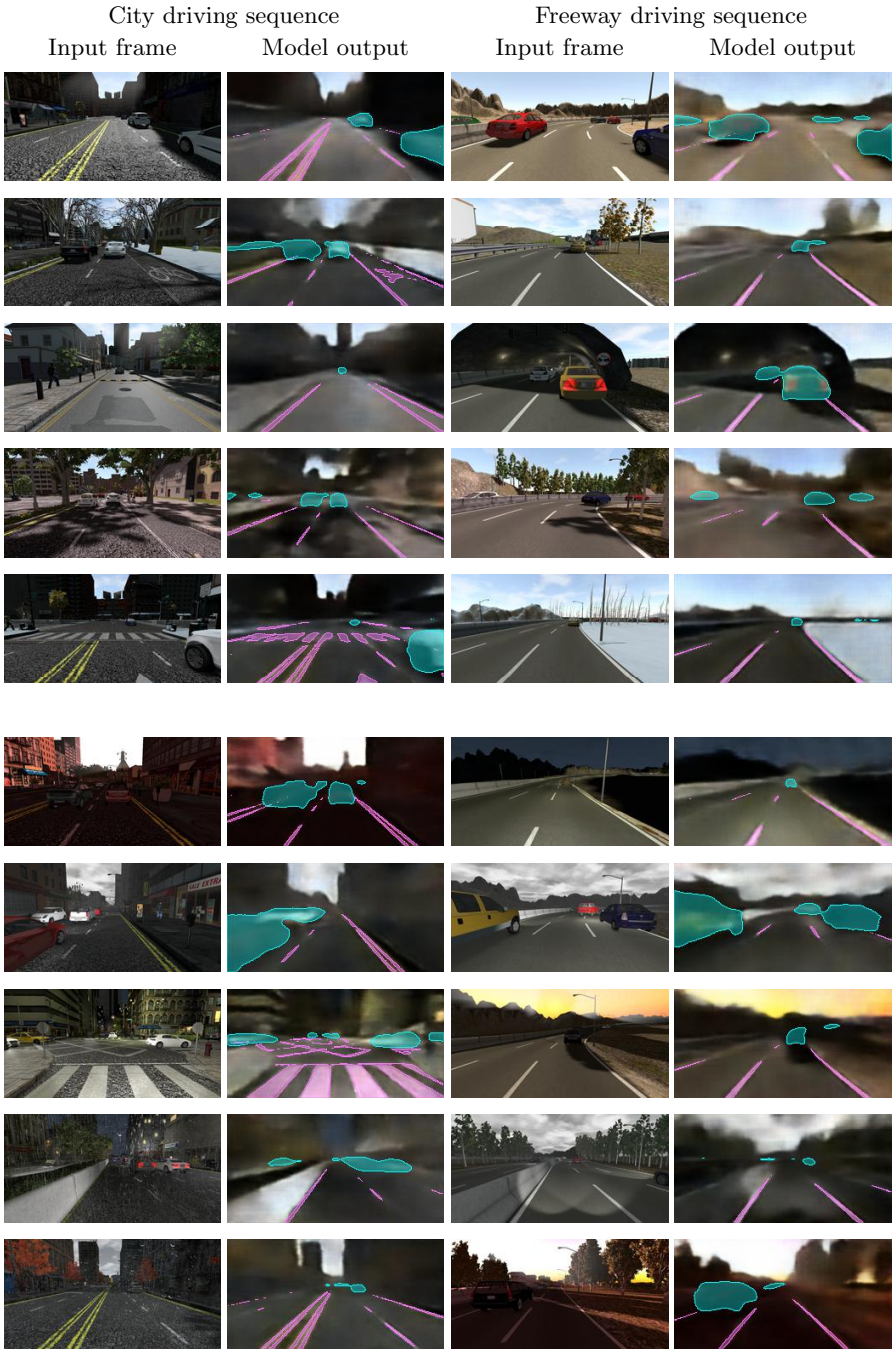

Fig. 4. Results of our model for two driving sequence of the SYNTHIA dataset: city centre and freeway driving, each in sunny environments (top) and adverse conditions (bottom). In the table, odd columns show the input frames, even columns show the outputs of our neural network. In the output images, the background is the result of the visual-space decoder, the output of the `car` conceptual-space decoder is highlighted in cyan, in pink the output of the `lane markings` conceptual-space decoder.

encoder	convolution	$7 \times 7 \times 16$
	convolution	$7 \times 7 \times 32$
	convolution	$5 \times 5 \times 32$
	convolution	$5 \times 5 \times 32$
	dense	2048
	dense	512
latent		128 $[16, 96, 16]$
decoders	dense	2048
	dense	4096
	deconvolution	$5 \times 5 \times 32$
	deconvolution	$5 \times 5 \times 32$
	deconvolution	$7 \times 7 \times 16$
	deconvolution	$7 \times 7 \times 3$

Table 1. Parameters of the architecture used to produce the final results. Note that the size of the latent space is 128, of which 16 neurons represent the `cars` concept and other 16 neurons represent the `lane markings` concept.

Table 2 display the IoU *(Intersection over Unit)* scores obtained by the network over the SYNTHIA dataset. The table shows how the task of recognizing the "car concept" generally ends up in better scores, with respect to the "lane marking concept". An explanation of why the latter task is more difficult can be the very low ratio of pixel belonging to the class of lane markings, over the entire image size. However, the performance of the model are satisfying, exhibiting the best accuracy in the driving sequences on highways, and in the sunniest lighting conditions (spring and summer sequences).

To demonstrate the generative capabilities of our model, we verified the result of interpolating two latent space representations. The images on the left and right of Fig. 5 are the two input images, while in the middle there are the images generated from the interpolation of the compact latent spaces of the inputs. Even in the case of very different input images, the interpolation generates novel and plausible scenarios, proving the robustness of the learned latent representation.

Lastly, we would like to stress again 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.

5 CONCLUSIONS

We presented a neural model for visual perception in the context of autonomous driving, grounded in a number of concepts from neuroscience and cognitive science. The main guiding principle is the CDZs proposed by Meyer and Damasio that in our context represent the neural correlate of mental imagery as simula-

	all	Highway 1	NewYork 1	European	NewYork 2	Highway 2
Car	0.8566	0.9245	0.9084	0.9037	0.9123	**0.9251**
Lane	0.6627	**0.8161**	0.6900	0.7522	0.6709	0.7493
mIoU	0.7597	**0.8703**	0.7992	0.8280	0.7916	0.8373

	dawn	fall	fog	night	rain	spring	summer	sunset	winter
Car	0.8896	0.8852	0.8872	0.9009	0.9002	0.9201	**0.9264**	0.8978	0.9101
Lane	0.6399	0.7319	0.6509	0.6897	0.7096	**0.7696**	0.7532	0.7247	0.7502
mIoU	0.7648	0.8086	0.7691	0.7953	0.8049	**0.8449**	0.8398	0.8113	0.8302

Table 2. IoU scores over the SYNTHIA dataset, grouped into the 5 different driving sequences of the dataset (table on top) and into 9 different environmental and lighting conditions (bottom). The results are given for the two "concepts" of cars and lane markings, and their joint mean.

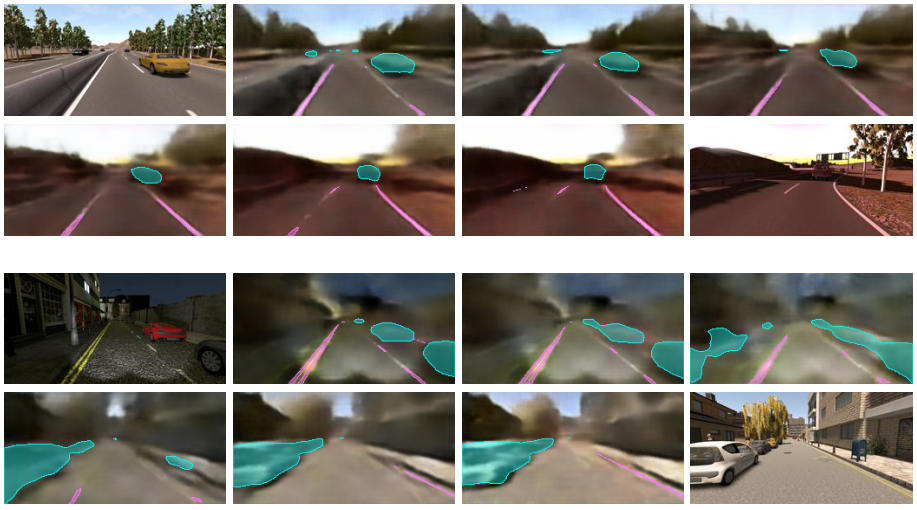

Fig. 5. Two examples of interpolation between latent space representations. For each sequence, images on the top left and bottom right are the inputs, the other images are obtained by interpolating the two latent spaces of the input images.

tion, following Jeannerod and Hesslow. CDZs find their best artificial cousin in the neural autoencoder architecture. For the choice of how to realize the convergence zone in the encoder, the guiding cognitive theory is that of transformational abstraction, suggesting the adoption of convolutional networks. One more theoretical contribution, the free-energy principle of Friston, further suggests to refine the autoencoder architecture as variational autoencoder. Based on these premises, our model aims at gaining an internal low–level representation of two spaces: the visual one and the conceptual one. The latter is limited to the two most crucial concepts during driving: `cars` and `lane markings`. We succeeded in achieving an internal representation as compact as with 128 units only, of which 16 units are enough to recognize the `car` concepts in any location of the visual space, and similarly for the `lane` concept. 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. 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.

References

1. Bengio, Y.: Learning deep architectures for AI. Foundation and Trends in Machine Learning **2**, 1–127 (2009)
2. Bojarski, M., Yeres, P., Choromanaska, A., Choromanski, K., Firner, B., Jackel, L., Muller, U.: Explaining how a deep neural network trained with end-to-end learning steers a car. CoRR **abs/1704.07911** (2017)
3. Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A.M., Jozefowicz, R., Bengio, S.: Generating sentences from a continuous space. CoRR **abs/1511.06349** (2015)
4. Bracewell, R.: Fourier Analysis and Imaging. Springer-Verlag, Berlin (2003)
5. Buckner, C.: Empiricism without magic: transformational abstraction in deep convolutional neural networks. Synthese **195**, 5339–5372 (2018)
6. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. CoRR **abs/1405.3531** (2014)
7. Chui, M., Manyika, J., Miremadi, M., Henke, N., Chung, R., Nel, P., Malhotra, S.: Notes from the AI frontier: insights from hundreds of use cases. Tech. Rep. April, McKinsey Global Institute (2018)
8. Damasio, A.: Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. Cognition **33**, 25–62 (1989)
9. Eickenberg, M., Gramfort, A., Varoquaux, G., Thirion, B.: Seeing it all: Convolutional network layers map the function of the human visual system. NeuroImage **152**, 184–194 (2017)

10. Felleman, D.J., Van Essen, D.C.: Distributed hierarchical processing in the primate cerebral cortex. Cerebral Cortex **1**, 1–47 (1991)
11. Fodor, J.: Modularity of Mind: and Essay on Faculty Psychology. MIT Press, Cambridge (MA) (1983)
12. Freedman, D.J., Riesenhuber, M., Poggio, T., Miller, E.K.: Categorical representation of visual stimuli in the primate prefrontal cortex. Science **291**, 312–316 (2001)
13. Friston, K.: The free-energy principle: a unified brain theory? Nature Reviews Neuroscience **11**, 127–138 (2010)
14. Friston, K., Fitzgerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G.: Active inference: A process theory. Neural Computation **29**, 1–49 (2017)
15. Friston, K., Stephan, K.E.: Free–energy and the brain. Synthese **159**, 417–458 (2007)
16. Fukushima, K.: Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics **36**, 193–202 (1980)
17. Gilbert, C.D., Wiesel, T.N.: Morphology and intracortical projections of functionally characterised neurones in the cat visual cortex. Nature **280**, 120–125 (1979)
18. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: International Conference on Artificial Intelligence and Statistics. pp. 249–256 (2010)
19. Grillner, S., Wallén, P.: Innate versus learned movements – a false dichotomy. Progress in Brain Research **143**, 1–12 (2004)
20. Güçlü, U., van Gerven, M.A.J.: Unsupervised feature learning improves prediction of human brain activity in response to natural images. PLoS Computational Biology **10**, 1–16 (2014)
21. Güçlü, U., van Gerven, M.A.J.: Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. Journal of Neuroscience **35**, 10005–10014 (2015)
22. Hassabis, D., Kumaran, D., Summerfield, C., Botvinick, M.: Neuroscience-inspired artificial intelligence. Neuron **95**, 245–258 (2017)
23. 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., Wang, X.: Applied machine learning at Facebook: A datacenter infrastructure perspective. In: IEEE International Symposium on High Performance Computer Architecture (HPCA). pp. 620–629 (2018)
24. Hesslow, G.: The current status of the simulation theory of cognition. Brain **1428**, 71–79 (2012)
25. Hinton, G., Zemel, R.S.: Autoencoders, minimum description length and Helmholtz free energy. In: Advances in Neural Information Processing Systems. pp. 3–10 (1994)
26. Hinton, G.E., McClelland, J.L., Rumelhart, D.E.: Distributed representations. In: Rumelhart and McClelland [53], pp. 77–109
27. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **28**, 504–507 (2006)
28. Hubel, D., Wiesel, T.: Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. Journal of Physiology **160**, 106–154 (1962)
29. Hubel, D., Wiesel, T.: Receptive fields and functional architecture of mokey striate cortex. Journal of Physiology **195**, 215–243 (1968)
30. Jeannerod, M.: Neural simulation of action: A unifying mechanism for motor cognition. NeuroImage **14**, S103–S109 (2001)

31. Jones, W., Alasoo, K., Fishman, D., Parts, L.: Computational biology: deep learning. Emerging Topics in Life Sciences **1**, 136–161 (2017)
32. Khan, S., Tripp, B.P.: One model to learn them all. CoRR **abs/1706.05137** (2017)
33. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of International Conference on Learning Representations (2014)
34. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Proceedings of International Conference on Learning Representations (2014)
35. Kosslyn, S.M.: Image and Brain: the Resolution of the Imagery Debate. MIT Press, Cambridge (MA) (1994)
36. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1090–1098 (2012)
37. Kulkarni, T.D., Whitney, W.F., Kohli, P., Tenenbaum, J.B.: Deep convolutional inverse graphics network. In: Advances in Neural Information Processing Systems. pp. 2539–2547 (2015)
38. Larochelle, H., Bengio, Y., Louradour, J., Lamblin, P.: Exploring strategies for training deep neural networks. Journal of Machine Learning Research **1**, 1–40 (2009)
39. Li, J., Cheng, H., Guo, H., Qiu, S.: Survey on artificial intelligence for vehicles. Automotive Innovation **1**, 2–14 (2018)
40. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E.: A survey of deep neural network architectures and their applications. Neurocomputing **234**, 11–26 (2017)
41. Marblestone, A.H., Wayne, G., Kording, K.P.: Toward an integration of deep learning and neuroscience. Frontiers in Computational Neuroscience **10**, article 94 (2016)
42. Meyer, K., Damasio, A.: Convergence and divergence in a neural architecture for recognition and memory. Trends in Neuroscience **32**, 376–382 (2009)
43. Moulton, S.T., Kosslyn, S.M.: Imagining predictions: mental imagery as mental emulation. Philosophical transactions of the Royal Society B **364**, 1273–1280 (2009)
44. Newen, A., Bruin, L.D., Gallagher, S. (eds.): The Oxford handbook of 4E cognition. Oxford University Press, Oxford (UK) (2018)
45. Olier, J.S., Barakova, E., Regazzoni, C., Rauterberg, M.: Re-framing the characteristics of concepts and their relation to learning and cognition in artificial agents. Cognitive Systems Research **44**, 50–68 (2017)
46. Plebe, A., Da Lio, M., Bortoluzzi, D.: On reliable neural network sensorimotor control in autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems **early access**, 1–12 (2019)
47. Plebe, A., Donà, R., Rosati Papini, G.P., Da Lio, M.: Mental imagery for intelligent vehicles. In: Proceedings of the 5th International Conference on Vehicle Technology and Intelligent Transport Systems. pp. 43–51. INSTICC, SciTePress (2019)
48. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: Xing, E.P., Jebara, T. (eds.) Proceedings of Machine Learning Research. pp. 1278–1286 (2014)
49. Rolls, E., Deco, G.: Computational Neuroscience of Vision. Oxford University Press, Oxford (UK) (2002)
50. Ros, G., Vazquez, L.S.J.M.D., Lopez, A.M.: 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. pp. 3234–3243 (2016)

51. Rosenfeld, A., Kak, A.C.: Digital Picture Processing. Academic Press, New York, second edn. (1982)
52. Rumelhart, D.E., Durbin, R., Golden, R., Chauvin, Y.: Backpropagation: The basic theory. In: Chauvin, Y., Rumelhart, D.E. (eds.) Backpropagation: Theory, Architectures and Applications, pp. 1–34. Lawrence Erlbaum Associates, Mahwah (NJ) (1995)
53. Rumelhart, D.E., McClelland, J.L. (eds.): Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge (MA) (1986)
54. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural Networks **61**, 85–117 (2015)
55. Schwarting, W., Alonso-Mora, J., Rus, D.: Planning and decision-making for autonomous vehicles. Annual Review of Control, Robotics, and Autonomous Systems **1**, 8.1–8.24 (2018)
56. Seger, C.A., Miller, E.K.: Category learning in the brain. Annual Review of Neuroscience **33**, 203–219 (2010)
57. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: 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., Lu, Z. (eds.) Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 240–248 (2017)
58. Tripp, B.P.: Similarities and differences between stimulus tuning in the inferotemporal visual cortex and convolutional networks. In: International Joint Conference on Neural Networks. pp. 3551–3560 (2017)
59. Tschannen, M., Lucic, M., Bachem, O.: Recent advances in autoencoder-based representation learning. In: NIPS Workshop on Bayesian Deep Learning (2018)
60. Ullman, S.: Using neuroscience to develop artificial intelligence. Science **363**, 692–693 (2019)
61. Van Essen, D.C.: Organization of visual areas in macaque and human cerebral cortex. In: Chalupa, L., Werner, J. (eds.) The Visual Neurosciences. MIT Press, Cambridge (MA) (2003)
62. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research **11**, 3371–3408 (2010)
63. Šmídl, V., Quinn, A.: The Variational Bayes Method in Signal Processing. Springer-Verlag, Berlin (2005)
64. Wolpert, D.M., Diedrichsen, J., Flanagan, R.: Principles of sensorimotor learning. Nature Reviews Neuroscience **12**, 739–751 (2011)
65. Zeiler, M.D., , Fergus, R.: Visualizing and understanding convolutional networks. In: Proc. of European Conference on Computer Vision. pp. 818–833 (2011)
66. Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R.: Deconvolutional networks. In: Proc. of IEEE International Conference on Computer Vision and Pattern Recognition. pp. 7–15 (2010)
67. Zeiler, M.D., Taylor, G.W., Fergus, R.: Adaptive deconvolutional networks for mid and high level feature learning. In: International Conference on Computer Vision. pp. 6–14 (2011)
68. Zhao, J., Mathieu, M., Goroshin, R., LeCun, Y.: Stacked what-where auto-encoders. In: International Conference on Learning Representations. pp. 1–12 (2016)