Visual pathways from the perspective of cost functions and multi-task deep neural networks

H.Steven Scholte[1,2,*] Max M. Losch[1,2,3,*] Kandan Ramakrishnan[3]
Edward H.F. de Haan[1,2] Sander M. Bohte[4]
[*] Shared first author
[1]Department of Psychology, University of Amsterdam, The Netherlands
[2]Amsterdam Brain and Cognition, University of Amsterdam, The Netherlands
[3]Informatics Institute, University of Amsterdam, The Netherlands
[4]Machine Learning Group, CWI, Amsterdam, The Netherlands
{h.s.scholte, m.m.losch, k.ramakrishnan, e.h.f.dehaan}@uva.nl, s.m.bohte@cwi.nl

Abstract

Vision research has been shaped by the seminal insight that we can understand higher-tier visual cortex from the perspective of multiple functional pathways with different goals. In this paper we try to give a computational account of the functional organization of this system by reasoning from the perspective of multi-task deep neural networks. Machine learning has shown that tasks become easier to solve when they are decomposed into subtasks with their own cost function. We hypothesise that the visual system optimizes multiple cost functions of unrelated tasks and this causes the emergence of the ventral pathway, dedicated to vision for perception and dorsal pathway, dedicated to vision for action. To evaluate the functional organization in multi-task deep neural networks we propose a method that measures the contribution of a unit towards each task and apply it to two networks that have been trained on either two related or two unrelated tasks using an identical stimulus set. Results show that the network trained on the unrelated tasks shows a decreasing degree of feature representation sharing towards higher-tier layers while the network trained on related tasks uniformly shows high degree of sharing. We conjecture that the method we propose can be used to reason about the anatomical and functional organization of the visual system and beyond as we predict that the degree to which tasks are related is a good descriptor of the degree to which they can share downstream cortical-units.

1. Introduction

The visual system is described as consisting of two parallel pathways. Research by Gross, Mishkin and colleagues, integrating insights from lesion (Newcombe, 1969) and anatomical studies (Schneider, 1969) showed that these pathways emerge beyond striate cortex with one, projecting ventrally, involved in the identification of objects, and the other, with a dorsal projection involved in localization of objects, projecting to parietal cortex (Gross & Mishkin, 1977; Mishkin, Ungerleider, & Macko, 1983). From the start of dual-pathway theory it was believed that multiple pathways are computationally efficient (Gross & Mishkin, 1977) support for this idea comes from research using artificial networks with one hidden layer, showing that location and identity are better learned when units in the hidden layers are uniquely assigned to one of these functions to a fully connected hidden network (Rueckl, Cave, & Kosslyn, 1989; Jacobs, Jordan, & Barto, 1991).

In the early nineties, Goodale & Milner argued, on the basis of neuropsychological, electrophysiological and behavioural evidence, that these pathways should be understood as have different goals: for the ventral pathway "vision for perception", involved in computing the transformations necessary for the identification and recognition of objects, and for the dorsal pathway "vision for action", involved in sensorimotor transformations of visually guided actions directed at these objects (Goodale & Milner, 1992).

It was recently suggested that the brain uses a variety of cost functions for learning (Marblestone, Wayne, & Kording, 2016). These cost functions can be highly diverse, and the brain could optimize a wide range of cost functions such as keeping body temperature constant or optimizing future reward from social interactions. High-level cost functions (by necessity) also shape other cost functions that determine the organization of perception: a cost function that is being optimized to minimize hunger affects the visual recognition cost function as foods have to be recognized. Mechanis-

tically, this could take place directly through, for instance, a reward modulation of object recognition learning, or indirectly through evolutionary pressure on the cost function associated with object recognition learning. In this paper, we try to understand how multiple pathways in the visual cortex might evolve from the perspective of Deep Neural Networks (see box 1) and cost functions (see box 2), and what this implies for how object information is stored in these networks.

We start with a discussion of the relevance of DNNs (LeCun, Bengio, & Hinton, 2015; Schmidhuber, 2015) and, following Marblestone (Marblestone et al., 2016), cost functions for understanding the brain in section 2. We extend our discussion with the importance of optimizing different cost functions simultaneously and present a hypothesis on the relationship between relatedness of tasks and the degree of feature representation sharing.

We test this hypothesis in a computational experiment with DNNs in section 3 to evaluate how much its feature representations contribute to each task. Next we discuss, in section 4, to what degree we are able to translate our experimental findings to the division between the ventral and dorsal pathway, the multiple functions of the ventral cortex and the apparent co-occurrence of both distributed and modular representations related to object recognition.

We finish this paper with a discussion of how this framework can be used experimentally to understand the human brain while elaborating on the limitations of DNNs and cost functions. For brevity we do not consider (models of) recurrent processing.

2. Multi-task DNNs as models of neural information processing in the brain

Artificial neural networks are inspired by computational principles of biological neuronal networks and are part of a large class of machine learning models that learn feature representations from data by optimizing a cost function. In this section, we discuss why we believe models based on optimizing cost functions such as DNNs are relevant for understanding brain function.

2.1. Similarities in architecture and behavior between DNNs and the brain

In terms of architecture, the hierarchical layers of a DNN resemble feedforward visual representations in the human brain (Lamme & Roelfsema, 2000; DiCarlo, Zoccolan, & Rust, 2012). The units in the first layer of AlexNet have a tuning that is similar to that of early visual cortex. Furthermore, in AlexNet, going from lower to higher tier layers we see, just like in general in human visual cortex, an increase in receptive field (RF) size and concurrently an increase in the specificity of tuning (Zeiler & Fergus, 2014).

A number of BOLD-MRI studies have revealed that the neural activations in early areas of visual cortex show the best correspondence with the early layers of DNNs and that higher-tier cortical areas show the best correspondence with higher-tier DNN layers (Güçlü & van Gerven, 2015; Eickenberg, Gramfort, Varoquaux, & Thirion, 2017). MEG/EEG studies have furthermore shown that early layers of DNNs have a peak explained variance that is earlier than higher-tier DNN layers (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Ramakrishnan, Scholte, Smeulders, & Ghebreab, 2016). In addition, the DNN model has been shown to predict neural responses in IT, both from humans and macaque, much better than any other computational model (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014).

The correspondence between DNNs and the brain begs the question to what degree DNNs show similar behavior as humans. Early results indicate that humans and DNNs have a similar pattern of performance in terms of the kinds of variation (size, rotation) that make object recognition harder or simpler (Kheradpisheh, Ghodrati, Ganjtabesh, & Masquelier, 2016). It has also been shown that higher-tier layers of DNNs follow human perceptual shape similarity while the lower-tier layers strictly abide physical similarity (Kubilius, Bracci, & Op de Beeck, 2016). On the other hand, DNNs are, for instance, much more susceptible to the addition of noise to input images than humans (Jang, McCormack, & Tong, 2017) and the exact degree to which the behavior of DNNs and humans overlap is currently a central topic of research.

As others (Kriegeskorte, 2015; Yamins & DiCarlo, 2016), we therefore believe that there is a strong case that DNNs can serve as a model for information processing in the brain. From this perspective using DNNs to understand the human brain and behavior is similar to using an animal model; like any model, far from perfect, but useful and with unique possibilities to yield insights in the computations underlying cortical function.

2.2. Cost functions as a metric to optimize tasks

While deep neural networks offer the representational power to learn features from data, the actual learning process is guided by an objective that quantifies the performance of the model for each input-output pair. Common practice in machine learning is to express such an objective as a cost function (Domingos, 2012). As Marblestone and colleagues argue, the human brain can be thought of implementing something very similar to cost functions to quantify the collective performance of neurons and consequently to steer the learning of representations in a direction that improves a global outcome (Marblestone et al., 2016).

Box 1 | Deep Neural Networks

Artificial neural networks refer to a large class of models loosely inspired by the way brain solves problems with a large number of interconnected units (neurons). The basic computation of a neural network unit is a weighted sum of incoming inputs followed by an activation function i.e a static nonlinearity (Rumelhart, McClelland, Group, & Others, 1988).

Composing a network of many of these basic computational units in more than 3 layers results in what is usually referred to as deep neural network (DNN). While the exact architecture of a DNN varies across applications, the one we are focusing on is the convolutional DNN, specifically designed for inputs with high spatially-local correlation like natural images. Convolution is hereby the process of applying a filter to each position in the image. In the first layer, these filters are able to detect for instance edges and very simple shapes, but composing a hierarchy of these filters allows for great compositional power to express complex features and is an important reason DNNs have proven to be so successful.

As determining these filters by hand is practically impossible DNNs are trained by backpropagation (LeCun et al., 1989), a standard machine learning optimization method based on gradient descent. Given a cost function that determines for an input and an expected output a single error value, backpropagation allows to assign a credit to each single unit in the network to specify how much it contributed to the error.

Recent state-of-the-art neural networks have increased depth, ranging from 16 (Simonyan & Zisserman, 2014) to 152 (He, Zhang, Ren, & Sun, 2015) layers (combined with some architectural advances). While the brain is clearly not shallow, its depth is limited to substantially fewer computational layers considering feed-forward processing (Lamme & Roelfsema, 2000). However, it has not yet been investigated how the layers of a very deep neural network map to the human brain.

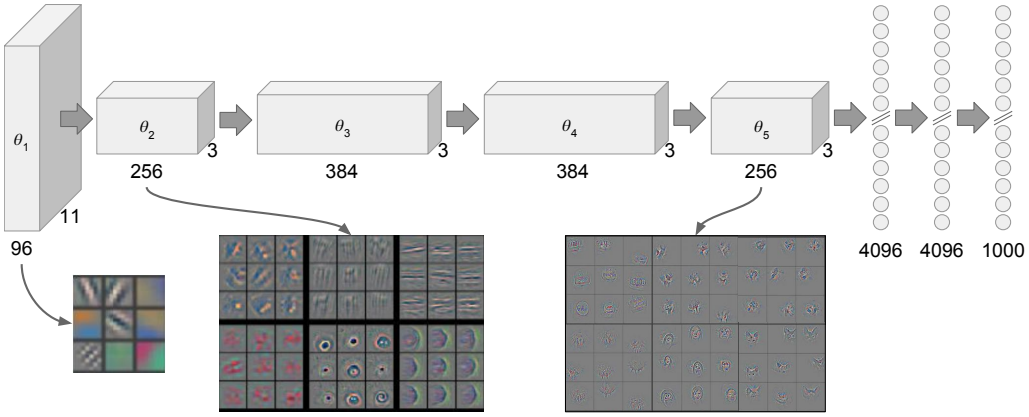

Figure Box.1. **Essential architecture of DNN AlexNet and filter visualization.** AlexNet consists of 5 convolutional layers represented by boxes and 3 fully connected layers of which the last is the output layer with 1000 units. The number of filters in a layer as well as the filter dimension is noted under each box. Below are selected filters visualized to show the increasing complexity of features they represent (adopted from Zeiler et al 2014).

Box 2 | Cost Functions

A cost function maps a set of observable variables to a real value representing the 'loss of the system. Optimization then aims to minimize this loss, for instance by changing tunable parameters θ in the system. For a predictive brain in a moving organism, the system tries to optimize actions, and sequences thereof that minimize one or more cost functions; these actions in turn are specified by a plethora of parameters , like synaptic efficacies and hormone levels. It is these parameters that are adjusted to change the actions that the system takes in a given environment to decrease the cost.

Mathematically, we can specify the collective sensory input into the brain at any point in time as S, and the joint output of muscle tensions as O. A cost function maps the outputs O into a value, $f(O)$, that is minimized by adjusting the parameters θ: learning. Multiple cost functions arise naturally when different measured quantities are to be optimized: if $t = f_{thirst}(O, \Theta)$ corresponds to the degree of thirst, and the system also has to optimize financial welfare $d = f_{fw}(O, \Theta)$, the system has to find the optimum values of theta that maximize both functions. We can jointly optimize these two cost functions by specifying a single combined cost function: $G = f_{fw}(O, \theta) + \lambda f_{thirst}(O, \theta)$, where λ is a weight that measures the relative importance of the two cost functions. Such joint cost functions can be learned with a single network, where the degree to which shared representations (in the form of shared learned features) help or hurt with the optimization task is variable (Caruana, 1998). The shape of the cost has likely evolved such that they help make most sense of our environment (Marblestone et al., 2016): a loss may measure the absolute deviation from some target value, or the square of this difference, or any other mapping.

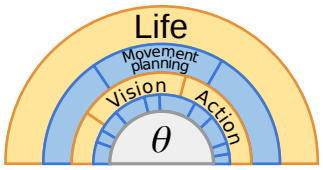

Figure 1. **Hierarchy of tasks related to the objectives the brain has to accomplish.** To make the evolutionary goal of Life tractable, the brain must be able to decompose it into manageable subtasks (blue and yellow arcs). All tasks and their cost functions effectively act on the same set of parameters (gray semicircle), while there may be differing degree of influence.

2.3. Problem simplification by task decomposition

While humans may act under a grand evolutionary objective of staying alive long enough to reproduce, we accomplish many small-scale objectives along the way, like guiding our arms to our mouth to eat or plan our path through the city. Each of these smaller objectives can be thought of as being governed by their own cost functions (see figure 1). These could be embedded in the brain, either hard coded into the neural substrate by evolution, by sovereign decision making, or as part of meta-learning: learning to learn (Baxter, 1998).

It has been argued that task becomes easier to solve if it can be decomposed into simpler tasks (Jacobs et al., 1991; Sutton, Precup, & Singh, 1999). To support their argument they state that the simple problem of learning the absolute value function can be decomposed into learning two linear functions and a switching function, which leads to a model with fewer parameters that can be trained faster. While such a decomposition could be predefined through the neural substrate, they observe in their experiments that such a decomposition can naturally arise from competitive learning, if the same set of parameters are optimized for multiple tasks. As the decomposition of tasks is underdetermined, the learner may come up with different decompositions, each time it is trained.

The notion of decomposition has been frequently made use of in machine learning literature on reinforcement learning (Dietterich, 2000) to increase learning speed and enable the learning of task-local optima that can be reused to learn a superordinate goal. Very often it is even impossible to specify the objective for a complex task so that it is a necessity to decompose it into tractable partial objectives. An example is the objective of vision. Finding an objective for such a broad and vague task appears futile so that it is easier to define a subset of tasks like figure ground segmentation, saliency and boundaries. A noteworthy implementation of such a decomposition is the recent DNN 'UberNet' (Kokkinos, 2016), which solves 7 vision related tasks (boundary, surface normals, saliency, semantic segmentation, semantic boundary and human parts detection) with a single multi-scale DNN network to reduce the memory footprint. It can be assumed that such a multi-task training improves convergence speed and better generalization to unseen data, something that already has been observed on other multi-task setups related to speech processing, vision and maze navigation (Dietterich, Hild, & Bakiri, 1990, 1995; Bilen & Vedaldi, 2016; Mirowski et al., 2016; Caruana, 1998).

3. Functional organization in multi-task DNNs

One hypothesis for the emergence of different functional pathways in the visual system is that learning and development in the cortex is under pressure of multiple cost functions induced by different objectives. It has been argued that the brain can recruit local populations of neurons to assign local cost functions that enable fast updating of these neurons (Marblestone et al., 2016). We explore in this section the ramifications of multiple cost functions acting on the same neurons by translating the problem to instances of multi-task DNNs sharing the same parameters. By observing the contributions each feature representation in a DNN has to each task, we will draw conclusions about the functional separation we observe in the visual cortex in section 4.

3.1. Hypothesis

Given two cost functions that optimized two unrelated tasks, which both put pressure on the same set of parameters, we conjecture that the parameters learned will be general enough to be used for both tasks (see figure 2B). In contrast, we speculate that, when the tasks are related, two subsets of parameters will emerge during learning that each lie within their task-respective feature domain (see figure 2C). Because the amount of feature representation sharing is determined by the relation between tasks and ultimately the statistics of the credit assignments we predict an upper to lower tier gradient of feature representation sharing, with the least sharing in higher tier layers.

3.2. Training models for multiple tasks

We test this hypothesis on feature representation sharing with DNNs trained for two tasks simultaneously. We construct two example setups involving a pair of related tasks (which we call RelNN), namely the simultaneous classification of ordinate and subordinate categories of objects in images, and a pair of unrelated tasks (which we call UnrelNN) namely the classification of objects and text labels in images (see figure 3). As the relatedness of tasks is not clearly defined and an open problem (Caruana, 1998; Zhang & Yeung, 2014), the tasks were selected based on the assumption that text recognition in UnrelNN is mostly independent of object recognition while in contrast ordinate level classifi-

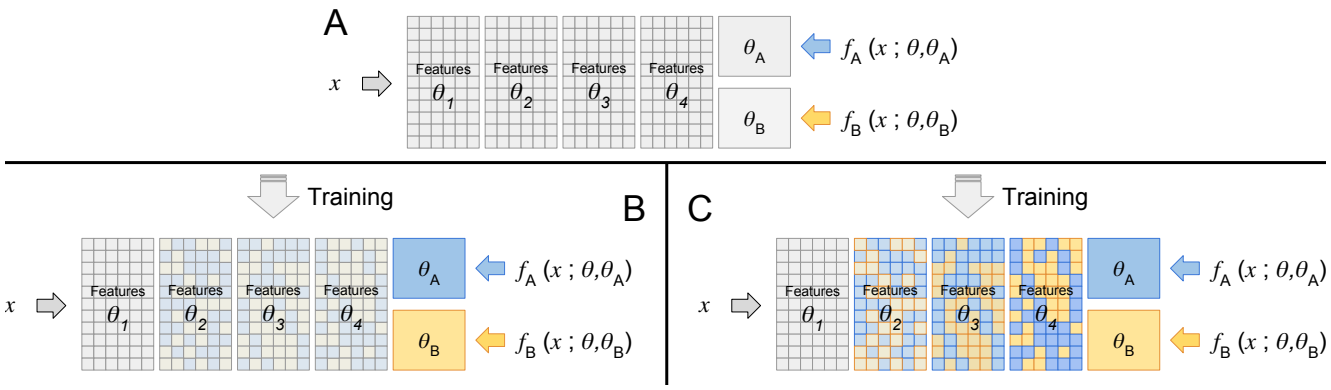

Figure 2. **Task relatedness and feature representation sharing in deep neural networks.** Given a multi-layered neural network with a set of feature representations θ (indicated by cells) that optimize differently related tasks, we conjecture that the degree to which representations can be shared is dependent on the generalizability, which reduces with the depth of the network for single modality inputs. The generalizability is indicated by the strength of the color. Gray tones indicate high generalizability, while strong colors indicate features that are tuned to one respective cost function. **A** — Initial, untrained network configuration with 5 layers for a single modality input x. Cost functions f_A and f_B have direct access to their respective parameters θ_A and θ_B. **B** — Two strongly related tasks inducing features that are generalizable to both tasks. Little function-specificity identifiable. **C** — Two largely unrelated tasks. While early simple features representations can be shared, intermediate and higher level representations are likely to be exclusive to their respective cost function due to their task-specificity.

cation in RelNN is highly dependent on the feature representations formed for subordinate level classification.

3.2.1 Training setup

Both setups were implemented by training a version of AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) on approximately half a million images from the ImageNet database (Russakovsky et al., 2015) each [1]. To optimize the models for two tasks simultaneously, the output layer of AlexNet was split into two independent layers. Both models were trained on an identical set of images consisting of 15 ordinate classes further divided into 234 subordinate classes, each image augmented with an overlay of 3 letter labels from 15 different classes (see figure 3a). The overlays were randomly scaled, colored and positioned while ensuring that the text is contained within the image boundaries. Furthermore to enable the networks to classify two tasks at once, the output layer was split in two independent layers (see figure 3b) for which each had its own softmax activation. For classification performance results see table 1.

3.2.2 Measuring feature representation contribution

To determine the degree of feature representation sharing in a neural network we measure the contribution each feature representation has to both tasks. Our method is inspired

[1]The code, data and pretrained models are available here: https://github.com/mlosch/FeatureSharing

	Top-5-error	
	Subordinate-level recognition	Ordinate-level/Text recognition
Chance	97.9%	66.7%
RelNN	14.0%	2.9%
UnrelNN	15.2%	4.9%

Table 1. **Classification errors.** Comparison of the error rates of RelNN and UnrelNN on a validation set of 11,800 images. The Top-5-error is defined as the correct prediction not being under the 5 most likely predictions. Both models were trained for 90 epochs until convergence with Nesterov accelerated gradient descent (Nesterov, 1983) with momentum of 0.9, starting with a learning rate of 0.01 and decreasing it every 30 epochs by a factor of 10.

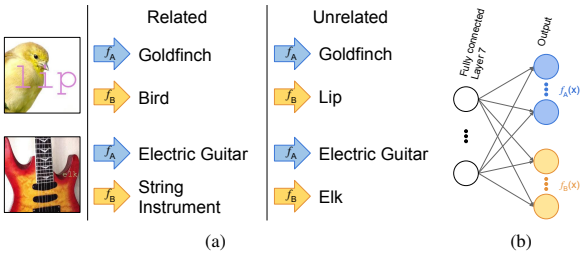

Figure 3. **Multi-Task-Learning setup.** (a) Two example images and their corresponding classification for our example setups of related and unrelated tasks. (b) To classify an input into two categories from different domains using AlexNet, the output layer is split in two where each split has its own softmax activation layer.

by the attribute contribution decomposition by (Robnik-Sikonja & Kononenko, 2008) which has recently been used to visualize the inner workings of deep convolutional networks (Zintgraf, Cohen, & Welling, 2016). The method is used to marginalize out features in the input image in the shape of small image patches, to observe the impact on the classification. In comparison, our method considers feature representations instead of features as we are not interested in the contribution of particular feature instances. The interested reader is referred to appendix A for the definition and derivation of the task contribution.

3.2.3 Results

We visualize the layer-wise task contributions by unrolling the feature representations of a layer on a rectangle and coloring each resulting cell by the composition of its contribution. Blue is used as indicator for the subordinate-level recognition task and yellow as indicator for the text- and basic-level-recognition task respectively. Equal contribution to both tasks results in grayish to white tones while little contribution to either task causes dark to black tones (see figure 4 for the color coding). A high degree of feature representation sharing would hereby generate cells colored in the range from black and gray to white, while low degree of sharing would result in more pronounced and clearly distinguishable colors of yellow and blue.

The two visualizations in figure 4 show a substantial difference in feature representation contribution as the representations in layer 2 to 5 of the RelNN contribute to both tasks much more equally than the representations of the UnrelNN. This is in line with our expectation depicted in figure 2 and our choice of setups. Contrary to our prediction, the degree of feature representation sharing in layer 1 of the UnrelNN is lower than expected; this can be explained by assuming that text recognition is mostly independent of all features but horizontal and vertical lines. Note also that most of the representations in the fully connected layers in both setups have only little contribution. This might seem counter-intuitive at first sight but is an effect of the abundance of representations coupled with the training scheme involving dropout. Dropout significantly reduces co-dependencies between units (Dahl, Sainath, & Hinton, 2013) resulting in only small changes in classification probability after marginalizing out a single representation.

We also observe that there is a dominance of blue cells expressing low contribution to the text- and basic-level-recognition task but high contribution to the subordinate-level-recognition task. We conjecture that this is because the subordinate-level-recognition task uses a larger fraction of units to distinguish between 200 classes.

Comparing the layers of both networks, it becomes evident that there generally is a higher degree of feature representation sharing in the RelNN consistent with the idea that relatedness between tasks and therefore cost functions strongly influences the degree of feature representation sharing across layers. More importantly, these results demonstrate that these types of ideas can be translated, using the right image data-sets and task-labels, into quantifiable predictions on the degree of feature sharing that might be observed in the brain.

4. Implications of models optimized for multiple tasks for understanding the visual system

In section 3 we presented an example in which the degree to which feature representations can be shared in a neural network depended on the relatedness of the tasks they are optimized for. In a neural population under pressure of the optimization for two unrelated tasks and the pressure to optimize the length of neuronal wiring (Chklovskii & Koulakov, 2004), a spatial segregation is likely to occur, resulting in anatomically and functionally separate pathways. In this section we consider to what degree we can understand the organization of the visual system from the perspective of a DNN that has been trained on multiple tasks and discuss three hypotheses derived from the simulations.

4.1. The visual system optimizes two cost functions of unrelated tasks

The early visual cortex has neurons that respond to properties such as orientation, wavelength, contrast, disparity and movement direction that are relevant for a broad range of visual tasks (Wandell, 1995). Moving upwards from early cortex we see a gradual increase in the tuning specificity of neurons resulting in the dorsal and ventral pathways that have, as has become clear the last 25 years, unrelated goals (Goodale & Milner, 1992). The dorsal pathway renders the representation of objects invariant to eye-centered transformations in a range of reference frames to allow efficient motor planning and control (Kakei, 1999), while the ventral pathway harbors object-centered, transformation invariant features (Leibo, Liao, Anselmi, & Poggio, 2015; Higgins et al., 2016) to allow efficient object recognition.

These observations concur well with the predictions and experimental results we made about feature representation sharing in DNNs. Given that the two tasks, vision for recognition and vision for action, are mostly unrelated we can understand the gradual emergence of functional and anatomical separation between these systems from this perspective.

Nonetheless, we note that the functional units of the pathways beyond the occipital lobe are not entirely separated and cross-talk does exist between these pathways (McIntosh & Schenk, 2009; Farivar, 2009; de Haan & Cowey, 2011; van Polanen & Davare, 2015): a phenomenon

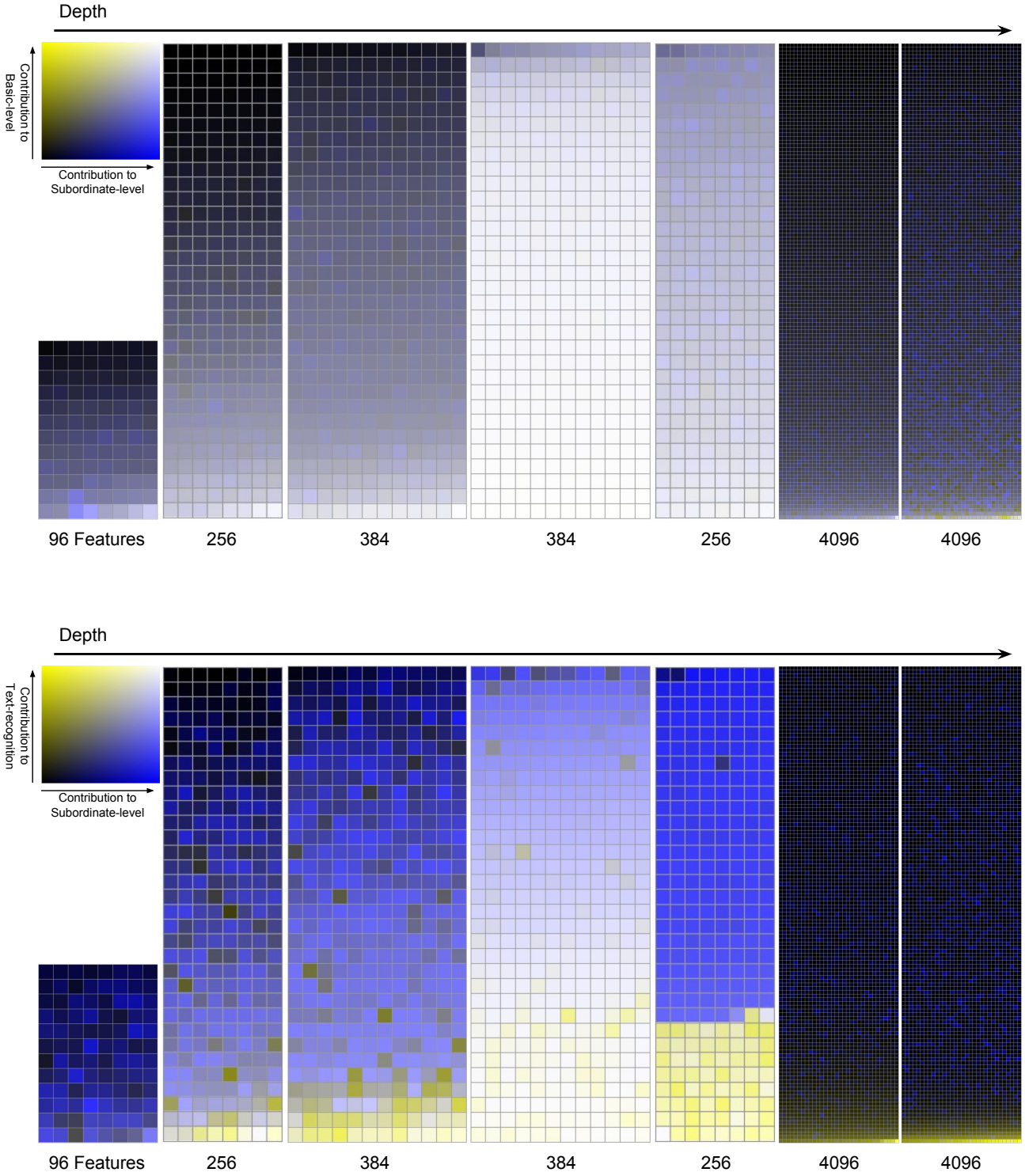

Figure 4. **Composition of feature representation contribution in DNNs to dual task.** (Best viewed in color) Each cell represents a feature representation in a neural network and its contribution. Task description and color coding of the contributions are displayed in the top left corner of each visualization. The cells are ordered by contribution magnitude of the yellow task so that the first cell in each layer displays the representation that contributes the least. (a) Contributions to RelNN, subordinate- and ordinate-level-recognition. (b) Contributions to UnrelNN, subordinate-level- and text-recognition.

we also observed in our experiment in section 3. In the UnrelNN, there are feature representations that contribute to both tasks throughout all layers of the network. Consequently the brain might trade off contribution and wiring length so that neurons that contribute little are tolerable to have long wiring to the functional epicentre.

As a whole the existence of two pathways guided by two cost functions of unrelated tasks might be seen as an illustration of the efficient decomposition of the overall vision function.

4.2. The visual pathways contain further task decompositions each with their own cost functions

We further generalize our perspective on cost function optimization of the visual system via the general observation made from machine learning that a complex task becomes simpler to solve if it is decomposed into simpler smaller tasks (see section 2.3). Given that the tasks we assign to the visual pathways are rather complex and vague we conjecture that there might be a broad range of cost functions active in the pathway regions to optimally decompose the task of vision resulting in a schematic similar to figure 5.

The ventral and dorsal pathways are each involved in a multitude of tasks serving the overall goals of vision for perception and vision for action. Examples of subordinate tasks for vision for action are localization, distance, relative position, position in egocentric space and motion and these interact with the goals that are part of vision for action: pointing, grasping, self-termination movements, saccades and smooth pursuit (de Haan & Cowey, 2011). Subordinate tasks for vision for perception include contour integration, processing of surface properties, shape discrimination, surface depth and surface segmentation. These in turn interact with executing the goals that are part of vision for perception: categorization and identification of object but also scene understanding (Groen, Silson, & Baker, 2017).

Reasoning from this framework we can also understand the existence of multiple 'processing streams' within the dual pathways. For instance, within ventral cortex there appears to be a pathway for object recognition and a pathway for scene perception. The object recognition pathway consists of areas like V4 which responds to simple geometric shapes and the anterior part of inferior temporal (aIT) that is sensitive for complete objects (Kravitz, Saleem, Baker, Ungerleider, & Mishkin, 2013). The scene recognition pathway contains areas such as the occipital place area (OPA), involved in the analyses of local scene elements and the parahippocampal place area (PPA) which responds to configurations of these elements (Kamps, Julian, Kubilius, Kanwisher, & Dilks, 2016). The tasks of scene and object perception are closely related; scenes consist of objects. However, scene perception involves relating the positions of multiple objects to each other, scene gist and navigability (Groen et al., 2017). From our framework we would predict that an area like OPA is mainly involved in the task of scene perception but has RFs that are also used for object perception and the opposite pattern for V4. Crucially, we believe this framework can be made to generate quantitative predictions for this amount of sharing.

4.3. Distributed versus modal representations

How information is represented is one of the major questions in cognitive neuroscience. When considering object based representations both both distributed (Haxby et al., 2001; Avidan & Behrmann, 2009) and module-based representations (Cohen, Dehaene, Naccache, Lehéricy, & others, 2000; Kanwisher, 2000; Puce, Allison, Gore, & McCarthy, 1995) have been observed.

Module based representations, and theories stressing their importance, point to the existence of distinct cortical modules specialized for the recognition of particular classes such as words, faces and body parts. These modules encompass different cortical areas and, in case of the fusiform face area and visual word form area, even similar areas but in different hemispheres (Plaut & Behrmann, 2011). Conversely, distributed theories of object recognition point to the possibility to decode information from a multitude of classes from the patterns of activity present in a range of cortical regions (Haxby et al., 2001; Avidan & Behrmann, 2009).

If we consider feature representations in the early and intermediate layers of the UnrelNN (figure 4) as a reasonable approximation of representations in early / intermediate visual areas, we note that most units are be shared by both streams. However, some units contribute more to one than the other task and are spatially intermingled at the same time. An external observer, analysing the activity of these representations under stimulation with pattern analysis would conclude that information from both tasks is present, and conclude that a distributed code is present. If the same observer would investigate the representations at the top of the stream the observer would conclude that there is an area dedicated to the analysis of text and another to the analysis of the subordinate task.

Translated to the visual system this would mean that distributed representations should be observed in areas such as posterior inferior temporal (pIT), OPA and V4 because these units are activated by multiple tasks but with a different weighting. Vice versa, at the top of a pathway or stream the network would show a strong module based pattern of activation. In sum, multi-task DNNs provide a framework in which we can potentially understand that both modal and distributed representations can be observed experimentally but suggest that the patterns of activity should be interpreted as emerging from the network as a whole.

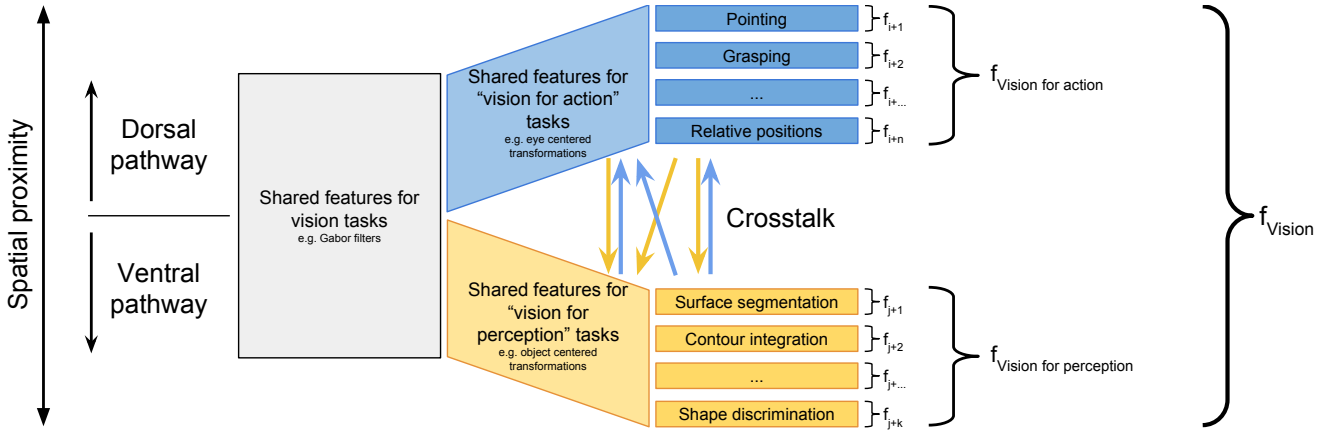

Figure 5. **How functional pathways in the visual system could be associated with cost functions.** (Vision for perception pathway in blue, vision for action pathway in yellow). Within the pathways are streams that develop under guidance from cost functions which are a direct decomposition of the pathways cost function. Feature representations that are learned for one task can still be used by units in the other pathway (crosstalk arrows). Both pathways share the same input units that either develop through the relation between tasks and/or evolutionary or developmental learning.

5. Discussion

Following Marblestone and colleagues (Marblestone et al., 2016), and the strength of the similarities between DNNs and the visual brain, we hypothesise that cost functions, associated with different tasks, are a major driving force for the emergence of different pathways.

A central insight from machine learning is that functions become easier to learn when they are decomposed as a set of unrelated subtasks. As a whole, the existence of two pathways guided by two cost functions of unrelated tasks might be seen as an illustration of the efficient decomposition of the overall vision function (Sutton et al., 1999). Observing that DNNs decompose a problem in multiple steps, with the earlier layers related to the input and later layers related to outputs demanded for the task, we hypothesized that the degree of feature representation sharing between tasks, will be determined by the relatedness of the tasks with an upper-to-lower tier gradient.

On this basis we performed simulations that confirm that units in a DNN show a strong degree of sharing when tasks are strongly related and a separation between units when tasks are unrelated. The degree to which this framework will be useful depends on the degree to which understanding elements of brain function using DNNs is valid which is discussed in section 5.1 and 5.2.

Having multiple pathways within a multi-task network might also help with catastrophic forgetting, the phenomenon that an old task is overwritten by learning a new task (section 5.3). Next we will discuss that the status of the task oriented 'vision for perception' and 'vision for action' framework (section 5.4). Finally we discuss the possibilities of using multi-task for further understanding the brain and ways in which our current analysis approach can be extended (section 5.5).

5.1. The biological realism of machine learning mechanisms

While there has been much progress in the field of Deep Learning, it remains a question how and if the weights of neurons are updated in learning under the supervision of cost functions in the brain, that is, what the actual learning rules are.

DNNs are trained using back-propagation, an algorithm believed to miss a basis in biology (Crick, 1989; Stork, 1989). Some of the criticisms include the use in backpropagation of symmetrical weight for the forward inference and backward error propagation phase, the relative paucity of supervised signals and the clear and strong unsupervised basis of much learning. Recent research has shown that the symmetrical weight requirement is not a specific requirement (Lillicrap, Cownden, Tweed, & Akerman, 2016). Roelfsema & Van Ooyen already showed in (Roelfsema & van Ooyen, 2005) that a activation feedback combined with a broadly distributed, dopamine-like error-difference signal can on average learn error-backpropagation in a reinforcement learning setting. Alternative learning schemes, like Equilibrium Propagation (Scellier & Bengio, 2017) have also been shown to approximate error-backpropagation while effectively implementing basic STDP rules.

Alternatively, effective deep neural networks could be learned through combination of efficient unsupervised discovery of structure and reinforcement learning. Recent work on predictive coding suggests this might indeed be feasible (Whittington & Bogacz, 2017). Still, the learning rules that underpin deep learning in biological systems are very much an open issue.

5.2. Cost functions as the main driver of functional organization

Reviewing literature on the computational perspective for functional regions in the visual system, we conclude that each region might be ultimately traced back to being under the influence of some cost function that the brain optimizes and its interplay or competition for neurons (Jacobs et al., 1991) with other cost functions resulting in different degrees of feature representation sharing. The domain-specific regions in the ventral stream for example may be caused by a cost function defined to optimize for invariance towards class-specific transformations (Leibo et al., 2015), of which the Fusiform Face Area could additionally be bootstrapped from a rudimentary objective, hard coded by genetics, to detect the pattern of two dots over a line (McKone, Crookes, Jeffery, & Dilks, 2012; Marblestone et al., 2016). Finally, as we argued in section 4, the functional separation of the ventral and dorsal pathway can be associated with two cost functions as well. We emphasize that the precise implementation of these cost functions is unknown and note the concept of the task "vision for recognition" and "vision for action" is merely a summary of all the subordinate tasks that these two tasks have been decomposed into, as argued in section 2.3 and the cost function box.

5.3. Multiple pathways as a solution for catastrophic forgetting

While joint cost functions can be learned when the quantities needed by the cost functions are all present at the same time, most animals are continually learning and different aspects of cost functions are present at different times. Then, it is well known that standard neural networks have great difficulty learning a new task without forgetting an old task, so-called catastrophic forgetting. Effectively, when training the network for the new task, the parameters that are important for the old task are changed as well, with negative results. While very low learning rates, in combination with an alternating learning scheme, can mitigate this problem to some degree, this is costly in terms of learning time. For essentially unmixed outputs, like controlling body temperature and optimizing financial welfare, an easy solution is to avoid shared parameters, resulting in separate neural networks, or "streams". Similarly, various properties can be derived from a single stream, like visual aspects (depth, figure-ground separation, segmentation), from an object recognition stream, where each aspect sub-stream is learned via a separate cost function. For tasks sharing outputs, and thus having overlap over different tasks, evidence increasingly suggests that the brain selectively "protects" synapses for modification by new tasks, effectively "unsharing" these parameters between tasks (Kirkpatrick et al., 2016).

5.4. What and where vs. vision for action and perception

Goodale & Milner argued that the concept of a 'what and where' pathway should be replaced by the idea that there are two pathways with different computational goals, vision for perception and vision for action, summarized as a 'what' and 'how' pathway (Goodale & Milner, 1992). Insights from the last 25 years of research in vision science have shown that the original idea of a what and where pathway lack explanatory power. It is clear that RFs in inferior temporal cortex are large when objects are presented on a blank background (Gross, Desimone, Albright, & Schwartz, 1985). However, these become substantially smaller and thereby implicitly contain positional information, when measured against a natural scene background (Rolls, Aggelopoulos, & Zheng, 2003). Interestingly, studies on DNNs have shown that approximate object localization can be inferred from a CNN trained on only classification, although the spatial extend of an object cannot not be estimated (Oquab, Bottou, Laptev, & Sivic, 2015).

With regards to the dorsal pathways it has been observed that there are cells relating to gripping an object that are specific for object-classes (Brochier & Umiltà, 2007) showing that this pathway contains, in addition to positional information, categorical information. These observations are in direct opposition to one of the central assumptions, a strong separation between identity and location processing, of the 'what' and 'where' hypothesis. It is now abundantly clear that the move from 'what' and 'where' pathway to 'what' and 'how' pathways and moving from input to function fits particularly well with vision as a multi-task DNN.

5.5. Future research

Originally DNNs were criticised for being "black" boxes, and using DNNs to understand the brain would equate to replacing one black box with another. Recent years have shown a rapid increase in our understanding of what makes a DNN work (LeCun et al., 2015; Simonyan & Zisserman, 2014; Zeiler & Fergus, 2014) and how to visualise the features (Zintgraf et al., 2016; Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2014; Zeiler & Fergus, 2014) that give DNNs its power.

These developments illustrate that DNNs are rapidly becoming more "gray" boxes, and are therefore a promising avenue into increasing our understanding of the architecture and computations used by the visual system and brain.

We therefore believe it is sensible to investigate to which degree multi-task DNNs, trained using the same input, will allow us to understand the functional organisation of the visual system. Using the analytical framework introduced in section 3, we can generate a fingerprint for each of the layers in a network based on the degree of feature representation sharing.

This can be subsequently related to the activation patterns, evoked by different tasks observed within different cortical areas. Alternatively it is possible to compare representational dissimilarity matrices (Kriegeskorte, Mur, & Bandettini, 2008) obtained from single and multitask-DNNs and determine which better explain RDMs obtained from cortical areas.

An open question remains how subtasks and their associated cost functions are learned from overall goals/general cost functions, both in machine learning (Lakshminarayanan, Krishnamurthy, Kumar, & Ravindran, 2016) and in neuroscience (Marblestone et al., 2016; Botvinick, Niv, & Barto, 2009).

Acknowledgements

MML is supported by a grant from the ABC, KR is supported by a grant from COMMIT and EHFdH by an ERC (339374 - FAB4V).

References

Avidan, G., & Behrmann, M. (2009, 14 July). Functional MRI reveals compromised neural integrity of the face processing network in congenital prosopagnosia. *Curr. Biol.*, *19*(13), 1146–1150.

Baxter, J. (1998). Theoretical models of learning to learn. In *Learning to learn* (pp. 71–94).

Bilen, H., & Vedaldi, A. (2016). Integrated perception with recurrent multi-task neural networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems 29* (pp. 235–243). Curran Associates, Inc.

Botvinick, M. M., Niv, Y., & Barto, A. C. (2009, December). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, *113*(3), 262–280.

Brochier, T., & Umiltà, M. A. (2007, December). Cortical control of grasp in non-human primates. *Curr. Opin. Neurobiol.*, *17*(6), 637–643.

Caruana, R. (1998). Multitask learning. In S. Thrun & L. Pratt (Eds.), *Learning to learn* (pp. 95–133). Springer US.

Chklovskii, D. B., & Koulakov, A. A. (2004). Maps in the brain: what can we learn from them? *Annu. Rev. Neurosci.*, *27*, 369–392.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016, 10 June). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.*, *6*, 27755.

Cohen, L., Dehaene, S., Naccache, L., Lehéricy, S., & others. (2000). The visual word form area. *Brain*, *123*(2), 291–307.

Crick, F. (1989, 12 January). The recent excitement about neural networks. *Nature*, *337*(6203), 129–132.

Dahl, G. E., Sainath, T. N., & Hinton, G. E. (2013). Improving deep neural networks for LVCSR using rectified linear units and dropout. In *2013 IEEE international conference on acoustics, speech and signal processing*.

de Haan, E. H. F., & Cowey, A. (2011, October). On the usefulness of 'what' and 'where' pathways in vision. *Trends Cogn. Sci.*, *15*(10), 460–466.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012, 9 February). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Dietterich, T. G. (2000). Hierarchical reinforcement learning with the MAXQ value function decomposition. *J. Artif. Intell. Res.*, *13*, 227–303.

Dietterich, T. G., Hild, H., & Bakiri, G. (1990). A comparative study of ID3 and backpropagation for english Text-to-Speech mapping. In *Machine learning proceedings 1990* (pp. 24–31).

Dietterich, T. G., Hild, H., & Bakiri, G. (1995). A comparison of ID3 and backpropagation for english text-to-speech mapping. *Mach. Learn.*, *18*(1), 51–80.

Domingos, P. (2012, October). A few useful things to know about machine learning. *Commun. ACM*, *55*(10), 78–87.

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017, 15 May). Seeing it all: Convolutional network layers map the function of the human visual system. *Neuroimage*, *152*, 184–194.

Farivar, R. (2009, October). Dorsal-ventral integration in object recognition. *Brain Res. Rev.*, *61*(2), 144–153.

Goodale, M. A., & Milner, A. D. (1992, January). Separate visual pathways for perception and action. *Trends Neurosci.*, *15*(1), 20–25.

Groen, I. I. A., Silson, E. H., & Baker, C. I. (2017). Contributions of low-and high-level properties to neural processing of visual scenes in the human brain. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *372*(1714), 20160102.

Gross, C. G., Desimone, R., Albright, T. D., & Schwartz, E. L. (1985). Inferior temporal cortex and pattern recognition. In C. Chagas, R. Gattass, & C. Gross (Eds.), *Pattern recognition mechanisms* (pp. 179–201). Springer-Verlag.

Gross, C. G., & Mishkin, M. (1977). The neural basis of stimulus equivalence across retinal translation. In S. Harnad, R. W. Doty, L. Goldstein, J. Jaynes, & G. Krauthamer (Eds.), *Lateralization in the nervous system* (pp. 109–122). Academic Press.

Güçlü, U., & van Gerven, M. A. J. (2015, 8 July). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.*, *35*(27), 10005–10014.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001, 28 September). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026–1034).

Higgins, I., Matthey, L., Glorot, X., Pal, A., Uria, B., Blundell, C., ... Lerchner, A. (2016, 17 June). Early visual concept learning with unsupervised deep learning.

Jacobs, R. A., Jordan, M. I., & Barto, A. G. (1991). Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. *Cogn. Sci.*, *15*(2), 219–250.

Jang, H., McCormack, D., & Tong, F. (2017). Evaluating the robustness of object recognition to visual noise in humans and convolutional networks. In *Journal of vision.*

Kakei, S. (1999). Muscle and movement representations in the primary motor cortex. *Science*, *285*(5436), 2136–2139.

Kamps, F. S., Julian, J. B., Kubilius, J., Kanwisher, N., & Dilks, D. D. (2016, 15 May). The occipital place area represents the local elements of scenes. *Neuroimage*, *132*, 417–424.

Kanwisher, N. (2000, August). Domain specificity in face perception. *Nat. Neurosci.*, *3*(8), 759–763.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014, 6 November). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.*, *10*(11), e1003915.

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016, 21 April). Humans and deep networks largely agree on which kinds of variation make object recognition harder.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... Hadsell, R. (2016, 2 December). Overcoming catastrophic forgetting in neural networks.

Kokkinos, I. (2016, 7 September). UberNet: Training a 'universal' convolutional neural network for low-, mid-, and High-Level vision using diverse datasets and limited memory.

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013, January). The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends Cogn. Sci.*, *17*(1), 26–49.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*(1), 417–446.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008, 24 November). Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Syst. Neurosci.*, *2*, 4.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 25* (pp. 1097–1105). Curran Associates, Inc.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016, April). Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput. Biol.*, *12*(4), e1004896.

Lakshminarayanan, A., Krishnamurthy, R., Kumar, P., & Ravindran, B. (2016). Option discovery in hierarchical reinforcement learning using Spatio-Temporal clustering. *arXiv preprint arXiv:1605. 05359*.

Lamme, V. A., & Roelfsema, P. R. (2000, November). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.*, *23*(11), 571–579.

LeCun, Y., Bengio, Y., & Hinton, G. (2015, 28 May). Deep learning. *Nature*, *521*(7553), 436–444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, *1*(4), 541–551.

Leibo, J. Z., Liao, Q., Anselmi, F., & Poggio, T. (2015, October). The invariance hypothesis implies Domain-Specific regions in visual cortex. *PLoS Comput. Biol.*, *11*(10), e1004390.

Lillicrap, T. P., Cownden, D., Tweed, D. B., & Akerman, C. J. (2016, 8 November). Random synaptic feedback weights support error backpropagation for deep learning. *Nat. Commun.*, *7*, 13276.

Marblestone, A. H., Wayne, G., & Kording, K. P. (2016, 14 September). Toward an integration of deep learning and neuroscience. *Front. Comput. Neurosci.*, *10*, 94.

McIntosh, R. D., & Schenk, T. (2009, May). Two visual streams for perception and action: current trends. *Neuropsychologia*, *47*(6), 1391–1396.

McKone, E., Crookes, K., Jeffery, L., & Dilks, D. D. (2012). A critical review of the development of face recognition: Experience is less important than previ-

ously believed. *Cogn. Neuropsychol.*, *29*(1-2), 174–212.

Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A. J., Banino, A., ... Hadsell, R. (2016, 11 November). Learning to navigate in complex environments.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends Neurosci.*, *6*, 414–417.

Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate O (1/k2). *Soviet Mathematics Doklady*, *27*(2).

Newcombe, F. (1969). *Missile wounds of the brain: A study of psychological deficits*. Oxford University Press, London.

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2015). Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *IEEE conference on computer vision and pattern recognition* (pp. 685–694). IEEE.

Plaut, D. C., & Behrmann, M. (2011, May). Complementary neural representations for faces and words: a computational exploration. *Cogn. Neuropsychol.*, *28*(3-4), 251–275.

Puce, A., Allison, T., Gore, J. C., & McCarthy, G. (1995, September). Face-sensitive regions in human extrastriate cortex studied by functional MRI. *J. Neurophysiol.*, *74*(3), 1192–1199.

Ramakrishnan, K., Scholte, H. S., Smeulders, A., & Ghebreab, S. (2016, 31 August). Mapping human visual representations by deep neural networks. *J. Vis.*, *16*(12), 373–373.

Robnik-Sikonja, M., & Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Trans. Knowl. Data Eng.*, *20*(5), 589–600.

Roelfsema, P. R., & van Ooyen, A. (2005, October). Attention-gated reinforcement learning of internal representations for classification. *Neural Comput.*, *17*(10), 2176–2214.

Rolls, E. T., Aggelopoulos, N. C., & Zheng, F. (2003, 1 January). The receptive fields of inferior temporal cortex neurons in natural scenes. *J. Neurosci.*, *23*(1), 339–348.

Rueckl, J. G., Cave, K. R., & Kosslyn, S. M. (1989). Why are "what" and "where" processed by separate cortical visual systems? a computational investigation. *J. Cogn. Neurosci.*, *1*(2), 171–186.

Rumelhart, D. E., McClelland, J. L., Group, P. R., & Others. (1988). *Parallel distributed processing* (Vol. 1). IEEE.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015, 1 December). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, *115*(3), 211–252.

Scellier, B., & Bengio, Y. (2017, 4 May). Equilibrium propagation: Bridging the gap between Energy-Based models and backpropagation. *Front. Comput. Neurosci.*, *11*, 24.

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

Schneider, G. E. (1969, 28 February). Two visual systems. *Science*, *163*(3870), 895–902.

Simonyan, K., & Zisserman, A. (2014, 4 September). Very deep convolutional networks for Large-Scale image recognition.

Stork. (1989). Is backpropagation biologically plausible? In *International joint conference on neural networks.*

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artif. Intell.*, *112*(1–2), 181–211.

van Polanen, V., & Davare, M. (2015, December). Interactions between dorsal and ventral streams for controlling skilled grasp. *Neuropsychologia*, *79*(Pt B), 186–191.

Wandell, B. A. (1995). *Foundations of vision.* Sunderland: Sinauer Associates.

Whittington, J. C. R., & Bogacz, R. (2017, May). An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural Comput.*, *29*(5), 1229–1262.

Yamins, D. L. K., & DiCarlo, J. J. (2016, March). Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.*, *19*(3), 356–365.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014, 10 June). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.*, *111*(23), 8619–8624.

Zeiler, M. D., & Fergus, R. (2014, 6 September). Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer vision – ECCV 2014* (pp. 818–833). Springer International Publishing.

Zhang, Y., & Yeung, D.-Y. (2014). A regularization approach to learning task relationships in multitask learning. *ACM Trans. Knowl. Discov. Data*, *8*(3), 1–31.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2014, 22 December). Object detectors emerge in deep scene CNNs.

Zintgraf, L. M., Cohen, T. S., & Welling, M. (2016, 8 March). A new method to visualize deep neural networks.

A. Measuring parameter contribution

A.1. Marginalization of parameters

In models that are able to handle the lack of information about a particular representation like in natural Bayesian classifiers, the contribution can be measured by marking the representation as unknown. Typically though, neural networks are not able to handle missing information and setting the parameters of a representation to zero will still result in interpretable information for subsequent layers. While removing a feature representation and retraining the network would alleviate this issue, quantifying the contribution of thousands of representations this way is generally unfeasible. Instead we make use of the models classification probabilities given by the softmax activation output which allows us to estimate the classification probability while lacking a representation by marginalizing it out via standard method from statistics. Marginalization effectively computes the weighted average of the classification probabilities after the representation has been replaced with random values sampled from an appropriate distribution. See equation 1 for the mathematical definition used for our evaluation.

$$p(y|x,\Theta_{\setminus\theta}) = \sum_{\theta} p(y|x,\Theta)p(\theta) \tag{1}$$

$p(y|x,\Theta)$ defines here the probability of input x belonging to class y and $p(y|x,\Theta_{\setminus\theta})$ the probability if θ is unknown. Note that a feature representation is represented by its parameters θ, which in turn consists classically of a weight w and a potential bias b in a neural network setting. Θ defines then the set of all parameters such that $\theta \in \Theta$. Each classification probability is eventually weighted by the prior probability of the sample θ expressing the likelihood the parameter in question takes value θ. We used 100 samples in our experiments to approximate the contribution.

A.1.1 Derivation

Given a parametric model like a DNN that is described by its parameters Θ, we can express the probability of input x belonging to class y as $p(y|x,\Theta)$, where the probabilities are given by the softmax output layer. To measure the contribution of a feature generated by parameter $\theta \in \Theta$, we are interested in what the probability is when θ is missing or unknown. By assuming that the input is independent of the parameters as well as the parameters are independent of each other, such that $p(x,\Theta) = p(x)p(\Theta)$ and $p(\Theta) = p(\Theta_{\setminus\theta})p(\theta)$ and by treating the parameters as random variables we can marginalize out θ as follows.

$$p(y|x,\Theta_{\setminus\theta}) = \frac{\int_\theta p(y,x,\Theta)d\theta}{\int_\theta p(x,\Theta)d\theta} \tag{2}$$

$$= \frac{\int_\theta p(y|x,\Theta)p(x,\Theta_{\setminus\theta})p(\theta)d\theta}{p(x,\Theta_{\setminus\theta})\int_\theta p(\theta)d\theta} \tag{3}$$

$$= \int_\theta p(y|x,\Theta)p(\theta)d\theta \tag{4}$$

As the integral over all possible values of θ is intractable for DNN-like structures, we instead approximate the probability by sampling from θ a finite number of times. We can now express the upper equation with a sum over all samples of θ.

$$p(y|x,\Theta_{\setminus\theta}) = \sum_{\theta} p(y|x,\Theta)p(\theta) \tag{5}$$

To sample from θ, we assume that the values are normal distributed with uniform variance and mean centered at the learned weight w and bias b:

$$\theta \sim \mathcal{N}(\mu = w,\, \Sigma = I), \mathcal{N}(\mu = b,\, \Sigma = I) \tag{6}$$

$$\textit{so that } p(\theta) = p_{\mathcal{N}(w,I)}(w) \cdot p_{\mathcal{N}(b,I)}(b) \tag{7}$$

A.2. Generalizing contributions from classes to tasks

As proposed by (Robnik-Sikonja & Kononenko, 2008), we use the weighted evidence (WE) to measure the contribution of parameter towards class probability $p(y|x,\Theta)$ (see equation A.6) instead of taking the difference of probabilities directly. $WE_\theta(y|x,\Theta)$ gives us a positive value indicating θ adds evidence for class y for input x, while a negative value adds evidence against class y and zero if θ has no contribution at all. To eventually determine the contribution towards a class independent of the input we calculate the arithmetic mean of the absolute weighted evidence over more than 500 input samples (see equation 10) from the test set.

$$odds(z) = \frac{p(z)}{1-p(z)} \tag{8}$$

$$WE_\theta(y|x,\Theta) = log_2(odds(y|x,\Theta)) - log_2(odds(y|x,\Theta_{\setminus\theta})) \tag{9}$$

$$C_\theta(y|\Theta) = \frac{1}{n}\sum_{j=1}^{n} |WE_\theta(y|x_j,\Theta)| \tag{10}$$

We finally measure the contribution to a task t by selecting the contributions $C_\theta(y|\Theta)$ that satisfy $y = y_{true}$ which are the class predictions that are correct. Furthermore filtering out predictions that had been incorrectly inferred from the network, we can increase certainty that the inputs used to evaluate the contributions lead to high probability for the

correct class and low everywhere else. We further generalize the contribution of θ to task t by averaging over the contributions to each class y_k within task t (see equation 11).

$$TC_\theta(t|\Theta) = \frac{1}{K}\sum_{k=1}^{K} C_\theta(y_k|\Theta) \tag{11}$$
$$t \in |Tasks|,\ K \in |outputs_t|$$