
UNIVERSITÀ DEGLI STUDI DI TRENTO

CIMeC - Center for Mind/Brain Sciences

Master's Degree in Cognitive Science

Academic Year 2016-17

# Using Convolutional Neural Networks to measure the contribution of visual features to the representation of object animacy in the brain

**Sushrut Thorat**
Student ID: 182270

**Supervisor: Marius Peelen**

**Abstract**

A mediolateral gradation in neural responses for images spanning animals to artificial objects is observed in the ventral temporal cortex (VTC). Which information streams drive this organisation is an ongoing debate. Recently, in Proklova et al. (2016), the visual shape and category ("animacy") dimensions in a set of stimuli were dissociated using a behavioural measure of visual feature information. fMRI responses revealed a neural cluster (extra-visual animacy cluster - xVAC) which encoded category information unexplained by visual feature information, suggesting extra-visual contributions to the organisation in the ventral visual stream. We reassess these findings using Convolutional Neural Networks (CNNs) as models for the ventral visual stream. The visual features developed in the CNN layers can categorise the shape-matched stimuli from Proklova et al. (2016) in contrast to the behavioural measures used in the study. The category organisations in xVAC and VTC are explained to a large degree by the CNN visual feature differences, casting doubt over the suggestion that visual feature differences cannot account for the animacy organisation. To inform the debate further, we designed a set of stimuli with animal images to dissociate the animacy organisation driven by the CNN visual features from the degree of familiarity and agency (thoughtfulness and feelings). Preliminary results from a new fMRI experiment designed to understand the contribution of these non-visual features are presented.

# Contents

# List of Figures

# Chapter 1

# Introduction

Images of animal and non-animal objects evoke distinct neural responses in the human visual ventral stream. In Kriegeskorte et al. (2008b), representational similarity analysis (RSA) of neural responses to a variety of stimuli in the inferior temporal (IT) cortex[1] of humans and monkeys yielded the "animate-inanimate" distinction as the primary organising principle. In Sha et al. (2015), RSA of neural responses in the lateral occipital (LO) cortex, with a different set of stimuli, did not show the animate-inanimate distinction which was observed in the response time profile of their behavioural oddball task. Principal component analysis (PCA) on the neural responses in the part of the ventral stream with significant object categorisation (encompassing LO and the ventral temporal (VT) cortices) unravelled a mediolateral graded response to the stimuli, which was termed the 'animacy continuum'. In Proklova et al. (2016), the animate-inanimate distinction was shown to exist for shape-matched stimuli for which there existed no animate-inanimate distinction in the response time profile of their behavioural oddball task. From these streams of evidence, we can conclude that differentiable neural responses to animate and inanimate objects do exist in the visual ventral stream.

### Is it prudent to call this neural response profile the animate-inanimate distinction?

The word 'animate' has many definitions[2] - 'alive; possessing life', 'lively', 'of or relating to animal life', 'able to move voluntarily', and '(*in linguistics*) belonging to a syntactic category or having a semantic feature that is characteristic of words denoting beings regarded as having perception and volition'. Although these definitions seem similar, some could cause problems in cases such as plants or invertebrates. In biological terms, plants and invertebrates do possess life, but as seen in studies (Kriegeskorte et al., 2008b; Sha et al., 2015), their neural representations in IT and VT cortex (VTC) are more similar to inanimate objects such as tools. The *linguistic* definition of 'animate' seems to align with the representational clustering in Kriegeskorte et al. (2008b).

This 'linguistic' definition of object animacy, anchoring on the association to the abilities of perception and volition, would also consider animated vignettes composed of simple geometric shapes in motion conveying social interactions as animate. This is in line with the findings (Martin and Weisberg, 2003) which report differentially higher activity to such vignettes conveying social interactions than to vignettes conveying mechanical actions in the lateral fusiform gyrus and differentially lower activity in the medial fusiform gyrus. This distinction anatomically matches the object image based animate-inanimate distinction in the fusiform gyrus (Proklova et al., 2016). *So, the linguistic definition seems appropriate.*

[1] Refer to Section 3.1 for the definitions of the various ROIs being mentioned here.

[2] animate. (n.d.). Dictionary.com Unabridged. Retrieved June 30, 2017 from Dictionary.com website - `http://www.dictionary.com/browse/animate`

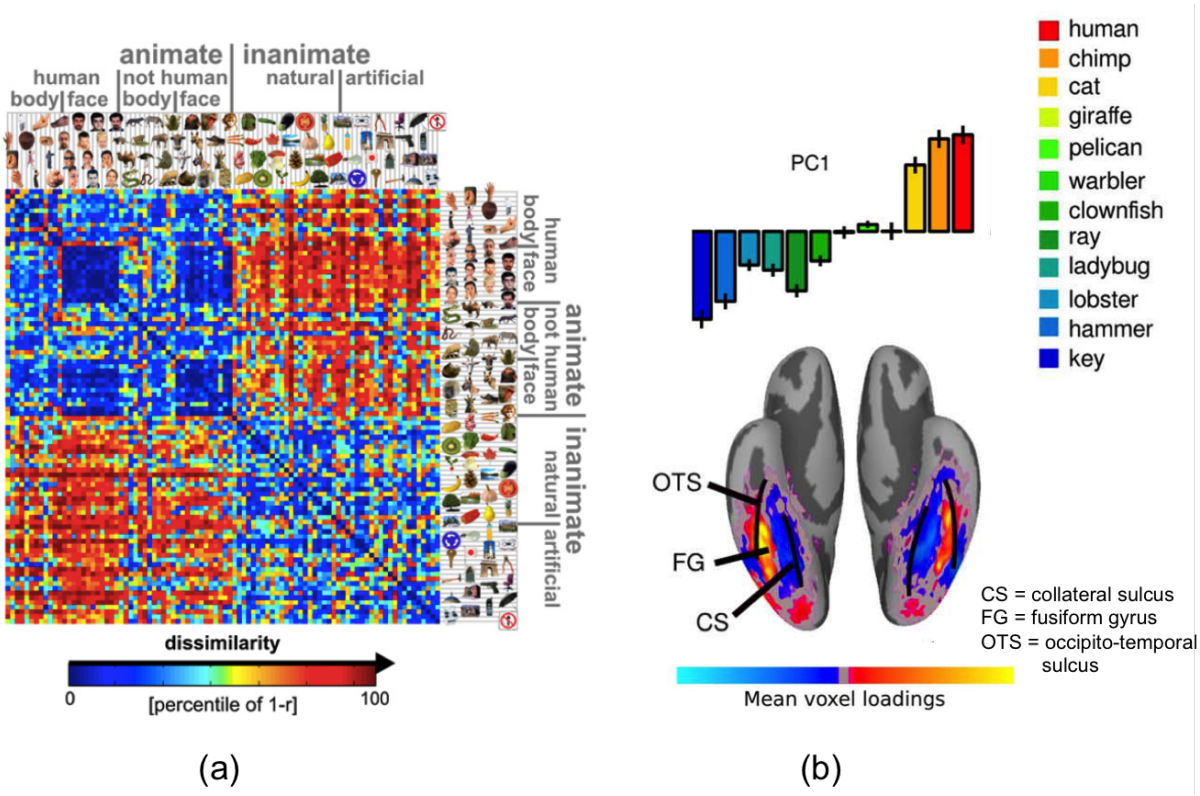


**Figure 1.1:** The animacy organisation in *(a)* the representational space of fMRI responses to images in the inferior temporal (IT) cortex (Kriegeskorte et al., 2008b), and *(b)* the first principal component of the fMRI responses to images in the ventral visual stream and the neural loadings of the principal component (Sha et al., 2015). Negative voxel loadings (*in blue*) correspond to the negative part of the principal component which corresponds to inanimate stimuli. This lateral-to-medial mapping of the animate-to-inanimate organisation matches the findings of Proklova et al. (2016).
(Adapted from: (a) Kriegeskorte et al. (2008b) and (b) Sha et al. (2015))

These results also suggest that the neural representations in this region are being shaped by both bottom-up information (object image features) and top-down information (perception and volition being associated to vignettes/objects).

### What information or other factors contribute to this animacy organisation?

As mentioned previously, presentation of images of objects such as animals, plants, and tools results in the said animacy organisation. Watching vignettes composed of simple geometric shapes move in a way suggesting social interaction and watching them move in a way suggesting mechanical actions also results in a differential neural response corresponding to the animacy organisation. Do stimuli from other modalities, such as sound, show such animacy response?

In Wang et al. (2015), fMRI response profiles for images of animals, objects, and scenes, and their corresponding spoken words were distinguishable in most of the ventral occipital-temporal cortex. These areas overlapped with the lateral fusiform gyrus which is an animate-selective region (Proklova et al., 2016). In contrast, the areas where the responses to images and sounds were indistinguishable encompassed the parahippocampal gyrus and medial anterior fusiform gyrus - regions which are predominantly inanimate-selective. In van den Hurk et al. (2017), fMRI responses to face, object, scene and body stimuli were presented visually and through associated sounds were recorded. The VTC responses could significantly categorise the sounds, but the cross-decoding between the audio and video conditions is low although being statistically significant. Given the weakness of these set of results and the fact that the animacy organisation

was not explicitly explored, no strong claims about its existence for audio stimuli can be made. As far as I know, there is no conclusive evidence that the modalities of touch or smell can show an animate-inanimate distinction in the VTC.

In van den Hurk et al. (2017), it is also reported that the response to auditory stimuli in the congenitally blind, in VTC, is more similar to the response to visual stimuli in the sighted than the response to auditory stimuli in the sighted. In Wang et al. (2015), different resting state functional connectivity profiles for the left posterior fusiform region (animate-selective in the sighted) were observed for the blind and sighted subjects. These findings suggest either that visual experience shapes the VTC during development or that the VTC is re-wired in the blind. No conclusive evidence for the existence of the animacy organisation in the congenitally blind exists.

Understanding the animacy organisation in the context of its function or emergence in VTC could also provide useful insights. The VTC is among the final stages in the visual ventral stream responsible for 'object categorisation' (Grill-Spector and Weiner, 2014) - a complicated task which requires a specialised architecture as evidenced by developments in neuroscience (Riesenhuber and Poggio, 1999) and computer vision (Krizhevsky et al., 2012). Optimally, the object representation space should not just categorise various objects but also represent objects along other relevant dimensions within the category. In Grill-Spector and Weiner (2014), this point is made by showing that various functional maps span VTC, as seen here in Figure 1.2 (top).

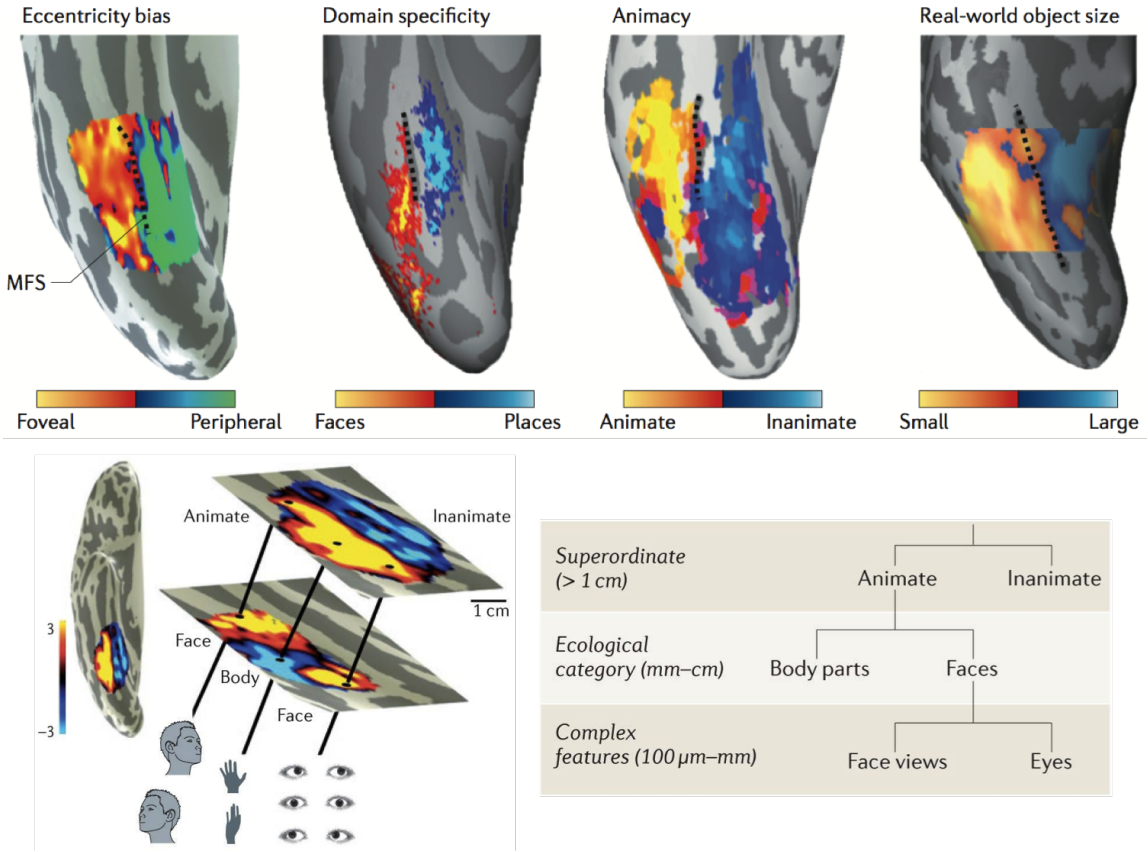


**Figure 1.2:** Functional maps in VTC. (top) Functional maps in the VTC found by various studies (for citations see Grill-Spector and Weiner (2014, figure. 4)). (bottom) The proposal that VTC not only possesses hierarchical representations (right) but also a corresponding hierarchy in anatomical distribution of category-selective areas (left).
(Adapted from: Grill-Spector and Weiner (2014))

In addition to the hierarchical organisation in representations as seen in Figure 1.2 (bottom, right), a corresponding anatomical hierarchy can be said to exist in VTC as seen in Figure 1.2 (bottom, left)[3]. So, the anatomical animacy organisation in VTC (medial-to-lateral inanimate-to-animate gradient) could just be a result of neural clustering driven either by the representations demanded by the task (object recognition; as is being hinted here) or by the connectivity of these regions in VTC either developed through childhood or innately present. In Deen et al. (2017), 3-8 months old infants' fMRI responses for visual depictions of faces, objects, bodies and scenes were acquired. The scenes vs faces contrast revealed neural clusters that anatomical corresponded to the clusters in adult brain for the same contrast. Representational similarity analysis in the extrastriate cortex (composed of the VTC, lateral occipital cortex (LOC), and the STS) revealed distinct representations for the four categories. The representations in infants were different than those in adults though. Given these observations, the authors conclude that the 'the overall functional organization of high-level visual cortex develops very early, and is subsequently refined'. Again, direct claims about the nature of the animacy organisation are not made.

To summarize,

- The property of animacy is designated to beings regarded as having perception and volition
- Differentiable fMRI responses to animate and inanimate objects are found in the ventral visual stream
- This animacy organisation is the highest in the hierarchy of categorical responses in the ventral temporal cortex
- There is no conclusive evidence in the literature for the animacy organisation in VTC in response to auditory and other sensory modalities
- A response profile in the ventral stream, overlapping with the animate-inanimate distinction, can be observed while viewing simple geometrical shapes in motion suggestive of social interactions as opposed to those suggestive of mechanical interactions.
- Therefore, the animacy organisation can be driven by both bottom-up information (object image features) and top-down information (perception and volition being associated to vignettes/objects).

Let us focus on the last point in the summary. A natural question arises - is the animacy organisation in the previously mentioned studies with visual stimulation being purely driven by visual features or does it also reflect any information non-computable through visual features? Animate stimuli such as primates and other mammals have features such as limbs, eyes, and particular body-part arrangements that could help distinguish them from inanimate stimuli such as cars and plants. *Could these differences in visual features solely drive the animacy organisation?* In Proklova et al. (2016), a resolution was sought in disentangling the visual feature differences and animacy category differences between pairs of stimuli. Let us take a deeper look into the experiment as our work builds on top of theirs.

[3]No study explicitly uses a multitude of stimuli spanning various categories, performs RSA on the fMRI responses in VTC, shows a hierarchical organisation in the representational space, projects that hierarchy onto the anatomical space, and shows that the resulting clusters match the clusters (say FFA or PPA) obtained from another localizer experiment.

## 1.1 Insights from disentangling object shape and category (Proklova et al., 2016)

In order to look for the impact of non-visual factors, we either need a measure of visual factors or non-visual factors. This study used a set of stimuli to make it hard to use visual features (object shape here) to provide information about the animacy category of the object, as seen in Figure 1.3. To quantify visual similarity, they used a visual search task introduced by Mohan and Arun (2012), where a subject has to look at a screen littered with an exemplar of the object and another oddball object and respond as quickly as possible to indicate if the left or right panels had the oddball. The reaction time (RT) is taken to be an indication of visual similarity - the more time it takes to find the oddball, the more similar is the oddball to the other object. They obtain pairwise RTs for all 16 objects (which had 4 exemplars each) and project the representational similarity matrix into two dimensions. As seen in Figure 1.3, the two dimensions show shape clusters but no animacy organisation. This is in contrast to the results of Mohan and Arun (2012), where they observe the animate-inanimate distinction in the first two principal components (see Figure B.1). They performed the same procedure on outlines and textures of those images and found no animacy reflected in the first two principal components in either case (see Figure B.1). The absence of the animate-inanimate distinction in the overall visual similarity, outline similarity, and texture similarity measures indicated that the shape-category manipulation was successful.

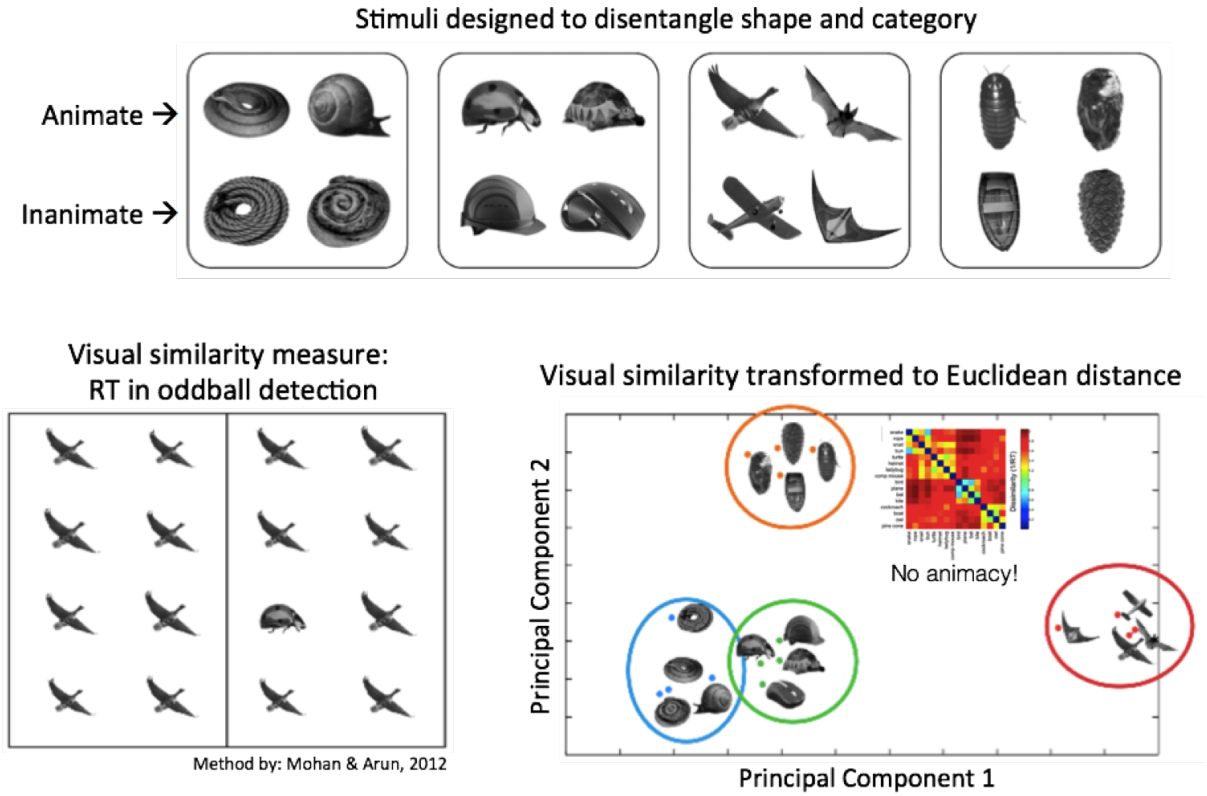


**Figure 1.3:** Stimuli for disentangling object shape and category. (top) The 16 objects grouped into 4 shape clusters with two instances each of animate and inanimate objects. (bottom, left) The visual experiment used in the study. Subjects had to indicate which of the panels contained the oddball object. The reaction time (RT) is a measure of visual similarity. (bottom, right) The first two principal components for pairwise RTs between the 16 objects. Shape clusters were observed, but the animate-inanimate distinction was not, indicating that the shape-category manipulation was successful.
(Adapted from: Proklova et al. (2016))

If any part of the animacy organisation is driven by non-visual features then the neural responses of that part should reflect category structure for this set of stimuli, as the visual feature differences, as quantified by the visual search task, would not drive the animacy organisation.

In this study, fMRI responses were recorded for the 16 objects (4 exemplars each, so 64 images) across 8 runs where the subject had to perform a one-back object task[4]. Subjects also underwent 2 runs of a 'object vs scrambled' and 'animate vs inanimate' localizer, where they had to perform a one-back image task[5].

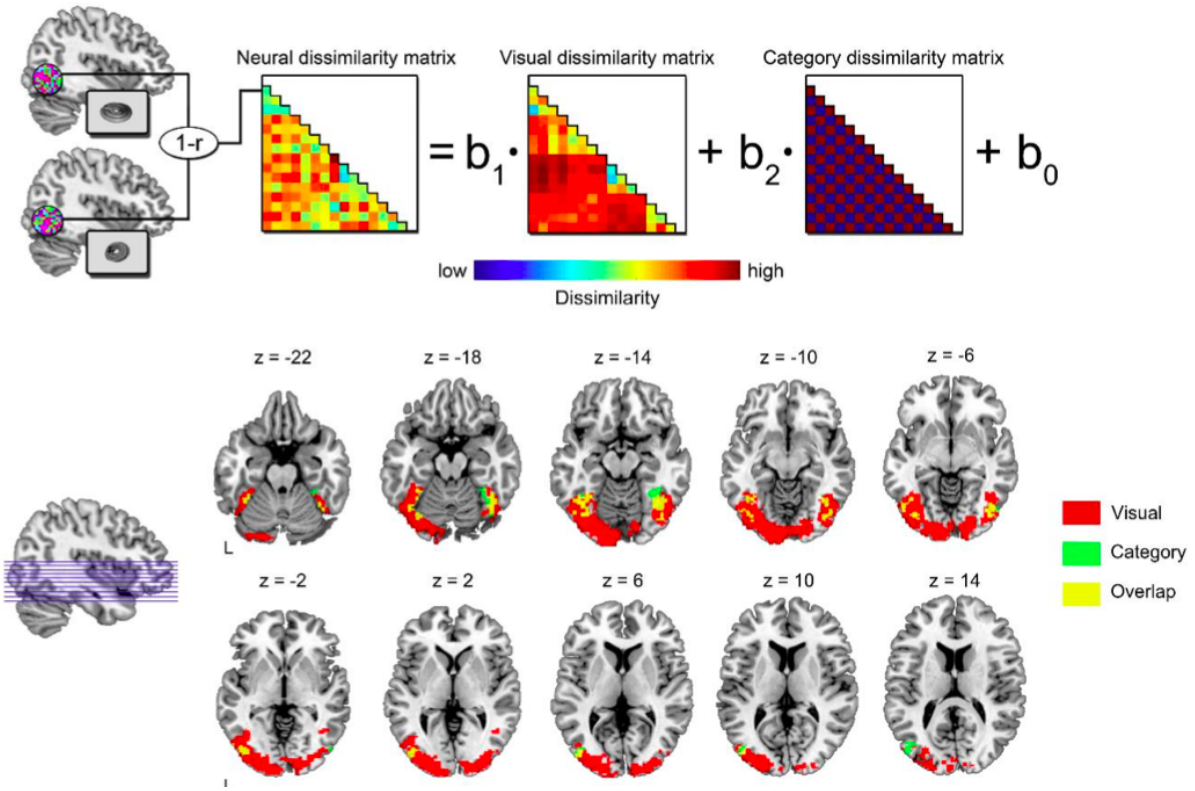


**Figure 1.4:** Searchlight regression analysis. (top) The schematic of the analysis is shown. The neural dissimilarity matrix was modelled with the visual dissimilarity matrix (obtained from the visual search task) and a category dissimilarity matrix. The regions where $b_2$ is significantly non-zero are using non-visual information (as gauged by the visual search task) to drive their animacy organisation. (bottom) Such regions were indeed found - the regions labelled 'Category' and 'Overlap' - which we collectively term the extra-visual animacy cluster (xVAC)
(Reproduced from: Proklova et al. (2016))

A univariate contrast in the localizer yielded clusters significantly coding the animate-inanimate distinction, and the clusters matched the location of the animacy organisation from previous studies. Then the betas[6] were averaged across the 8 runs. Multivariate searchlight analysis was then conducted where the neural RDM (created via pairwise correlations between the averaged betas) was regressed using the visual dissimilarity matrix (obtained from the visual search task) and a category dissimilarity matrix (0 if the two objects are from the same animacy category, 1 if not). The procedure is depicted in Figure 1.4. Clusters in the visual ventral stream were observed (we henceforth refer to the combined cluster as the *extra-visual animacy cluster* or xVAC) where the category regressor had significant weight, echoing the univariate

[4]Pressing a button if the image preceding the current image was of the same object.

[5]Pressing a button if the image preceding the current image matched the current image on a pixel level

[6]Obtained from running a general linear model (GLM) on the processed fMRI responses convolved with a hemodynamic response function (HRF)

result. Similar clusters were observed when the outline and texture dissimilarity matrices were added as regressors in the searchlight analysis.

The existence of xVAC seems to suggest that visual features are not sufficient to drive parts of the animacy organisation in the ventral visual stream, making the non/extra-visual information important. That is the main conclusion of the paper. The authors, however, point out that "there may be residual visual differences between animals and inanimate objects that do not affect visual similarity as measured in the visual search experiments", and we explore this possibility in our work.

## 1.2 Towards a better model of the neural representations in the ventral visual stream

Proklova et al. (2016) dealt with the question - Could the differences in visual features between objects solely drive the animacy organisation? - by designing stimuli to disentangle visual feature differences, as gauged by a visual search task, from animacy category, finding extra-visual animacy encoding in some regions of the visual ventral stream. This line of thought works assuming that the regions encoding the animacy organisation solely have access to the visual features which lead to dissimilarities that are quantified through the visual search task as described in Mohan and Arun (2012). What other models of visual features in the ventral stream could we use?

As mentioned in Grill-Spector and Weiner (2014), the ventral temporal cortex (VTC) encodes object category and is one of the final stages of the visual ventral stream which hierarchically transforms pixel-level visual information into complex features such as shapes, object parts and object category. Researchers in neuroscience and computer vision have grappled with building an algorithm which could take any image of an object as input and output its object label (Riesenhuber and Poggio, 1999; Andreopoulos and Tsotsos, 2013; DiCarlo et al., 2012).

Recent advances in graphics processing units (GPUs[7]), the availability of massive datasets of natural images, and better regularisation techniques, combined with the surge in the interest in artificial neural networks, led to the development of convolutional neural networks (CNNs) for large scale object recognition (LeCun et al., 2015; Schmidhuber, 2015), which can classify objects with superb performance (Canziani et al., 2016). CNNs use local and hierarchical transformations to achieve the task of object recognition (specifics about the CNN architecture would be discussed in Section 2.1), paralleling the simple to complex cell formulation of how the visual ventral stream functions (Serre et al., 2007). This parallel is best visualised in a schematic in Yamins and DiCarlo (2016), reproduced in Figure 1.5 (Top panel). A string of studies (Khaligh-Razavi and Kriegeskorte, 2014; Cadieu et al., 2014; Cichy et al., 2016; Kubilius et al., 2016) showing the similarities between the representations in the ventral visual stream to the representations in the various layers of CNNs (see: Figure 1.5 (Bottom panel)) tell us that the comparison is more than a parallel. This similarity might hint at an optimal substrate-independent scheme to perform object recognition.

How far do the representations in CNNs match the ventral visual stream? In Khaligh-Razavi and Kriegeskorte (2014), it was shown that AlexNet (Krizhevsky et al., 2012), one of the simplest CNNs, had layers which possessed the visual feature differences that could result in the animate-inanimate distinction. As we shall see in Section 2.1, all visual feature differences developed in the layers are a result of optimising the representations towards object recognition where all object labels are orthogonal (no relational information is provided to the CNN). The animacy organisation found in the CNNs is purely driven by visual feature differences.

[7]GPUs are processing units with multiple 'dumbed down' cores. Unlike a CPU core (which are in the tens in a CPU), each GPU core can only perform simple operations. For architectures such as neural networks, which require a lot of simple computations done in parallel, a GPU is better suited.

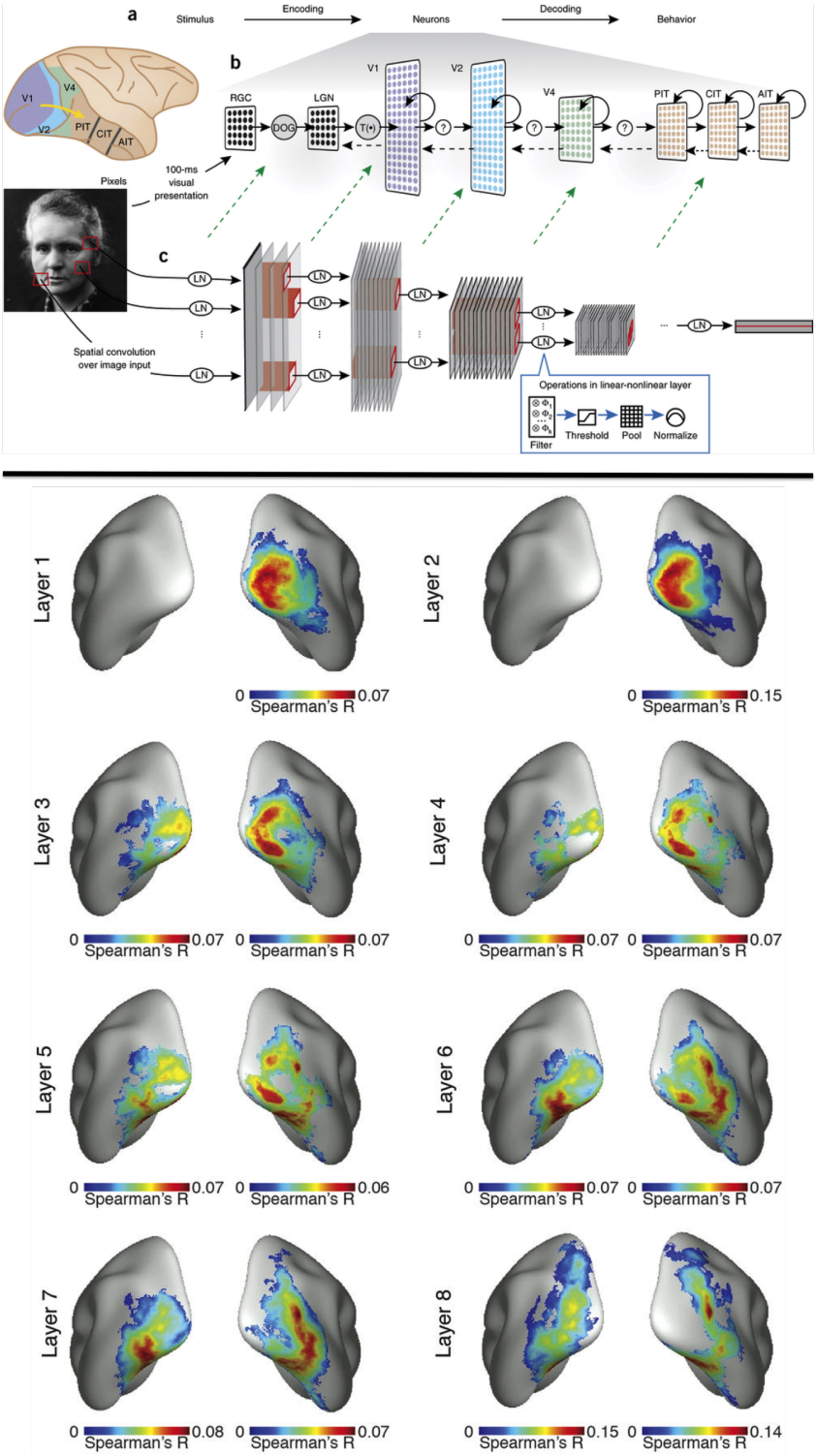


**Figure 1.5:** Task, architectural, and representational similarities between the visual ventral stream and convolutional neural networks (CNNs). (top) CNNs lack the recurrent connections found in the brain, but the comparison of most interest is the similarity between the hierarchical transformations of visual features which has been observed in other studies. (bottom) Representational similarity from the 8 layers of a CNN were correlated with searchlights on the brain. The figure shows that low level visual cortex correlates the most with the lower CNN layers, and as we go deeper into the CNN, the correlation peak moves anterior and ventral, into regions responsible for object categorisation.
(Reproduced from: (top) Yamins and DiCarlo (2016); (bottom) Cichy et al. (2016))

We are interested in whether the animacy organisation in the CNN is similar to the animacy organisation found in the ventral visual stream. We would like to know if this is the case for the stimuli created to disentangle object shape and category as presented in Proklova et al. (2016). If the mentioned similarity exists, the validity of using visual search experiments to model visual similarity in the ventral visual stream is questionable. Unlike the visual search experiments which provide us with one measure of visual similarity, the CNN provides us with multiple, hierarchical, measures of visual similarity. We would like to know if these CNN-based measures could help us better understand the role of visual information in driving the representations, and the animacy organisation in particular, in the visual ventral stream.

Furthermore, we also wanted to look for the impact of non-visual factors explicitly. We know that vignettes of simple shapes conveying social interactions seem to influence the animacy organisation in the ventral visual stream (Martin and Weisberg, 2003). This influence may be an effect of the assignment of an identity or agency to the shapes because of their seemingly social interactions. So, we decided to consider the factor of object agency - thoughtfulness and feelings, and object familiarity, in our quest to understand the contribution of non-visual features to the animacy organisation.

## 1.3 Outline of the thesis

In this work, we aim to understand -

1. The contribution of visual feature differences, as gauged through a convolutional neural network (CNN), to the animacy organisation in the ventral visual stream.

2. The contribution of non-visual feature differences, as gauged by the factors pertaining to agency which are dissociated from the CNN-based visual feature differences, to the animacy organisation in the ventral visual stream.

In Chapter 2, I shall provide a brief introduction to the architecture of CNNs, followed by the method to extract the animacy organisation in CNNs.

In Chapter 3, I shall outline the similarities between the representations in the CNN and in the visual ventral stream, the comparison between the representations in the CNN and those from the visual search task outlined in Proklova et al. (2016), explain the method to extract the animacy organisation in a ROI in the brain, and compare the animacy organisations of the CNN and the ROIs.

In Chapter 4, I shall outline a set of new experiments (behavioural and fMRI) to gauge the contribution of non-visual factors to the animacy organisation, and will present preliminary results.

# Chapter 2

# The animacy organisation in Convolutional Neural Networks

## 2.1 Convolutional Neural Networks for object recognition

A convolutional neural network (CNN) is an artificial neural network that is structured to capitalise on the local correlations in a data stream, by convolving the data stream with filters hierarchically, helping to minimise the parameters of the network, while learning task-relevant representations of the data. In the task of object recognition, the input to the CNN is an image[1], and the output is an object label. The constituents of a basic CNN are explained in Figure 2.1. An account which views CNNs in the context of the bigger field of deep learning, a rapidly expanding subset of machine learning, can be found in LeCun et al. (2015).

CNNs are trained with backpropagation - a gradient-descent based technique used to train multilayer neural networks. In a CNN for object recognition, the input image is transformed by the CNN into a vector of likelihood estimates of the various object labels. The error, which is the difference between the observed output and the target output (1 for the correct object label and 0 for the rest) is backpropagated to tune the network weights to minimise the error. In the CNNs we use for our analysis, the target vector representations of object labels are orthogonal. In other words, the network is penalised equally if it terms an image of a dog a wolf than if it terms the image a car. The features developed in the CNN can be said to be visually driven in nature[2]. With deconvolution, we can visualise the spatial patterns that drive the feature neurons in various layers (Zeiler and Fergus, 2014) (see: Figure B.2). We can see edge detectors and gabor filters arise in the early layers and object part filters in later layers, exhibiting the hierarchy of features required for object recognition. These filters/features emerge with other training objectives such as unsupervised learning of sparse representations and can be transferred to related tasks such as image captioning (Karpathy and Fei-Fei, 2015), making them generic enough to open the possibility for multi-task learning (Yosinski et al., 2014; Ruder, 2017). In using CNNs as models for visual processing (in humans or otherwise), we can, therefore, be sure that we are not capitalising on highly idiosyncratic features.

We use AlexNet (Krizhevsky et al., 2012) and VGG-16 (Simonyan and Zisserman, 2014) in our analysis. AlexNet contains 5 convolutional layers and 3 fully-connected layers. VGG-16 contains 13 convolutional layers and 3 fully-connected layers (further details: Appendix A). We will analyse other CNN architectures (Canziani et al., 2016) in further work.

[1] A mean image (average over the IMAGENET (Russakovsky et al., 2015) dataset) is subtracted from the actual image before being input into the CNNs.

[2] Although the possibility of the CNN feature learning capitalising on object co-occurrences in input images (as IMAGENET (Russakovsky et al., 2015), the dataset on which the CNNs are trained, contains cluttered images) cannot be ruled out.

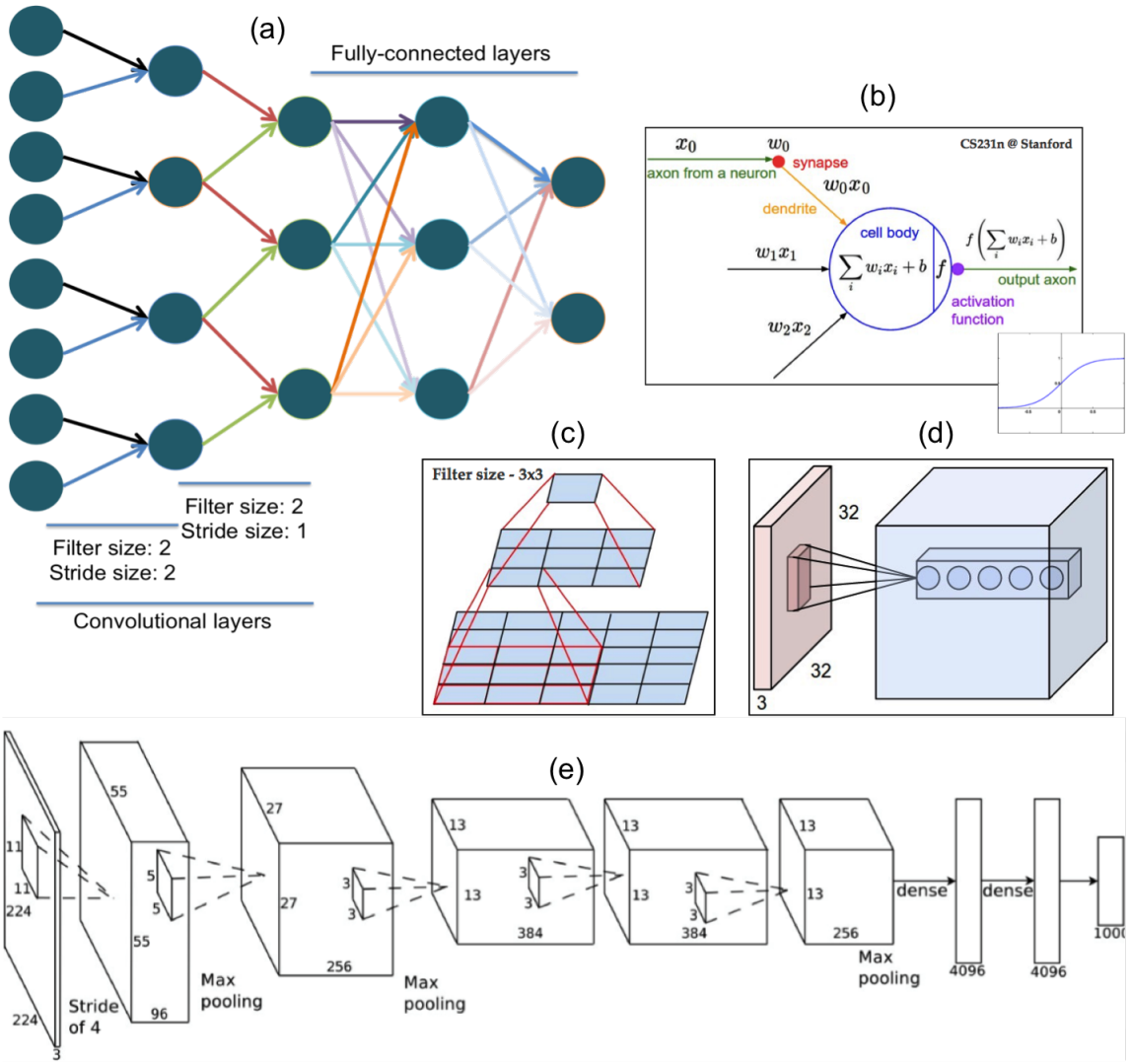


**Figure 2.1:** Constituents of a Convolutional Neural Network (CNN). (a) depicts a CNN with a one-dimensional input. The main feature of a CNN is the filter convolution, designed to both capitalise on local structure of the input and to minimise the number of parameters to be trained. In the first convolution step shown, the filter takes two neural inputs and maps it to one neuron in the second layer. The same filter is then used starting with a neuron 2 steps away from the first neuron (hence stride size 2) and so on. The distinction between convolutional and fully-connected layers can be seen. (b) The neurons used in the CNNs perform a weighted sum of the inputs and pass the output through an activation function. The sigmoid activation function is shown here, but the CNNs we use in this work use rectified linear units (ReLU). (c) depicts a filter operating on a two-dimensional map. (d) depicts a convolution operation on three-dimensional data which is what is used for CNNs operating on colour images (2 spatial dimension and 1 colour dimension). The depth dimension in the second layer contains the different filters (such various orientations, gabor filters or object parts). (e) depicts AlexNet, a simple CNN detailed in Krizhevsky et al. (2012), with 5 convolutional layers and 3 fully-connected layers. (Reproduced from: (b,c,d) CS231n at Stanford University (2016); (e) Krizhevsky et al. (2012))

## 2.2 Animacy in Convolutional Neural Networks

We saw that a range of features are developed in a CNN. As seen in Khaligh-Razavi and Kriegeskorte (2014), the differences in visual features in the CNN can be used to distinguish between animate and inanimate objects, faces and non-faces, and body and non-body images. This was gauged by training classifiers called support vector machines (SVMs) (Cortes and Vapnik, 1995) on neural activity in the CNN elicited by images containing objects of the opposing

categories. A SVM is designed to find the optimal hyperplane which splits the multi-dimensional space in which the neural activity resides such that the activity patterns for images in each category are maximally separated to opposite sides of the hyperplane. The classification (either on new data, or cross-validation accuracy on the training data) performance of the SVM is a measure of linear separability[3] of the representations in the CNN layers.

Showing that linear separability is insufficient for our purposes. We not only want to check if the animate-inanimate distinction exists, but also gain insights into the animacy organisation - the distribution of object classes along the animacy dimension (orthogonal to the hyperplane). Thus, the 'animacy coefficient' (positive if classified as animate, negative if inanimate) of an object, represented by the neural activity of a particular layer, is given by the signed distance from the animacy hyperplane (SVM decision score), as shown in Figure 2.2.

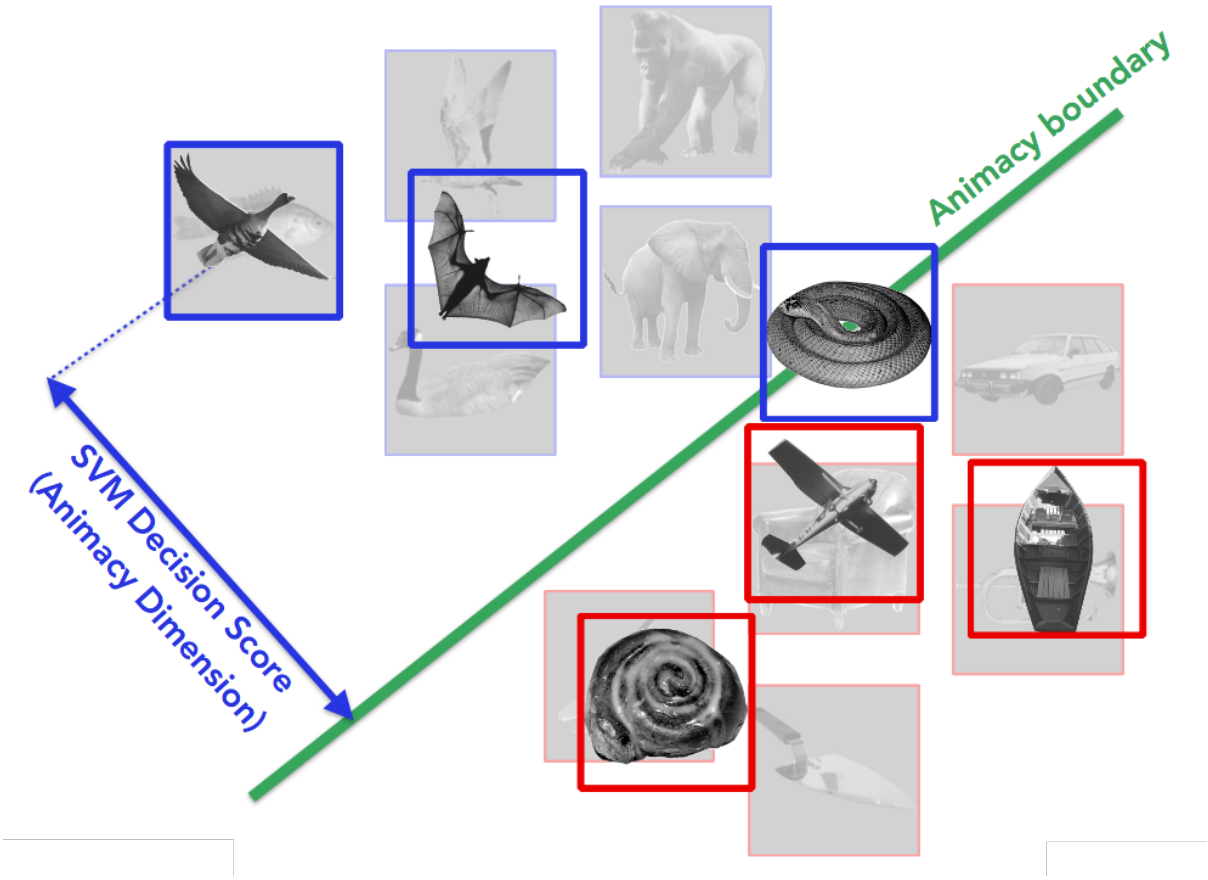


**Figure 2.2:** SVM decision scores as animacy coefficients. A support vector machine (SVM) was trained on a training dataset (dimmed images) and the animacy boundary/hyperplane was constructed. SVM decision scores are given by the signed distance of the representation of an object from the animacy boundary (i.e. along the animacy dimension). Samples from a test dataset (non-dimmed images) are plotted. The SVM does classify the animate (in blue) and inanimate (in red) objects properly, although the animacy coefficients of the objects vary. In this formalism, given these representations, the bird is said to be more animate than the snake as it is farther away from the boundary.

[3]Neural activity in a region of the brain or a layer of the CNN resides in a high dimensional space. Linear or non-linear transformation of the representations therein could lead to the discovery of a multitude of patterns. So, can the region/layer be said to represent/process all those patterns? No. The entire CNN is a non-linear transformation on the image representations in the input image space. We do not say that the image space 'represents' object categories. Linear separability is the criteria we use to assign a function/representation to a region/layer. That is how/why we even talk about regions or layers (this holds true for all functional regions of interest in the brain). So, if the neural activity in a brain region or CNN layer elicited for animate and inanimate object images is linearly separable, that region/layer can be said to encode the animacy organisation.

### 2.2.1 Training SVMs for animacy classification with CNN features

There are two types of layers in the CNNs we analyse - convolutional and fully-connected. The fully-connected layers are already vectors which could be directly used as examples for training SVMs. The convolutional layers are 3D structures, which are reshaped into 1D vectors before training SVMs on them. We subtract the mean activation of each neuron (to the training dataset) before training the SVMs. We do not normalise the neural activations with the standard deviation because, if a neuron is more responsive to differences in animacy, we want it to maintain its relative importance for that task. We use two datasets to train two sets of SVMs each for AlexNet and VGG-16:

- Dataset 1 - 960 coloured images (evenly split between animate and inanimate) of isolated objects, from a dataset created for Kiani et al. (2007), also used in Khaligh-Razavi and Kriegeskorte (2014) to train SVMs on the CNN features.
- Dataset 2 - 72 grayscale images (evenly split between animate and inanimate) of isolated objects, used in the localizer experiment in Proklova et al. (2016). We use this small dataset because in training SVMs to obtain the animacy organisation in regions of the brain (as discussed in the next chapter), we have fMRI data available only for this dataset.

In training SVMs on the first few convolutional layers of VGG-16 with 960 images, we ran into a memory error on our system. The feature space of VGG-16 convolutional layers had to be reduced and we resorted to principal component analysis (PCA). This line of thought also provided us with an opportunity to assess the prominence of the animacy dimension in the CNN layers. We trained SVMs on each layer for multiple number of dimensions extracted, using the MATLAB (R2015b) *fitcsvm* function[4]. To find the optimal number of dimensions which could classify the training images well and capture as much variance in the feature space as possible, while keeping the number of dimensions low, we defined an optimiser function, O,

$$\mathrm{O}(d) = \mathrm{CVA}(d) \times \mathrm{V}(d) \times (1 - d/N) \tag{2.1}$$

where $d$ is the number of principal components or dimensions extracted, $\mathrm{CVA}(d)$ is the $6-$fold crossvalidation accuracy of the SVM trained with $d$ dimensions, and $\mathrm{V}(d)$ is the fraction of the variance in the feature space explained by the $d$ dimensional subspace. We want to find the $d_{opt}$ that maximises O. We might find different values of $d_{opt,i}$ for the various layers $i$ of the CNN. We would like to take the maximum of $d_{opt,i}$ and perform dimensionality reduction of all the feature spaces (of the different layers) to this maximum. In Figure 2.3, in the left panels, we can observe that the crossvalidated accuracy rises close to its asymptotic value in a few principal components for the final layers (in red) for both AlexNet and VGG-16, suggesting that animacy is a major factor in the representations therein. In the right panel, by finding the layer for maximising $d_{opt,i}$, we decide to reduce the feature spaces of VGG-16 to 495 dimensions and that of AlexNet to 313 dimensions.

The cross-validated accuracy tells us that both AlexNet and VGG-16 can distinguish between animate and inanimate objects. But can they classify the shape-matched stimuli used to disentangle object shape and category in Proklova et al. (2016), shown in Figure 1.3? We obtain the CNN neural activations for those images and subtract the mean activity elicited by the training images in Dataset 1. We then obtain the decision scores from the SVMs trained on the training images, and check the classification accuracy. To check the robustness of acquired decision scores, we perform rank-order correlations (Kendall's tau) between the decision scores from the SVM trained with the number of dimensions under consideration with the decision scores from the SVMs trained with the closest lower and higher number of dimensions, and average the two correlations. The higher the average, the more robust the decision scores are

[4]We did not optimise the hyperparameter *box constraint* in this phase of analysis, and used the default value.

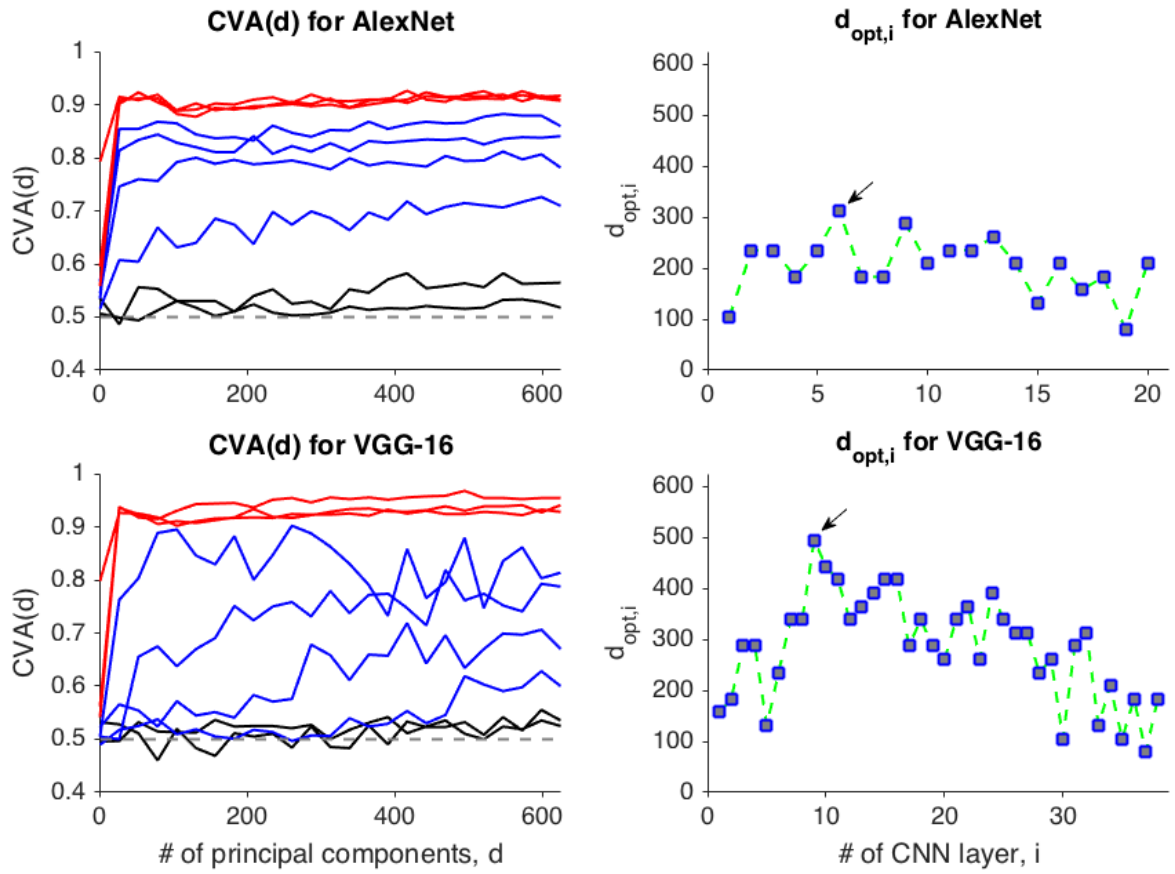


**Figure 2.3:** Searching for animacy in the CNN feature dimensions. In the left panels, the 6−fold crossvalidation accuracy of the SVMs trained on the $d$ principal components of the feature spaces of various CNN layers in plotted. These plots are for the SVMs trained using Dataset 1, so the images inhabit 959D subspaces in the CNN layer feature spaces. The layers in black correspond to low-level features (input layer and Conv1), the layers in blue correspond to mid-level features (Conv2-5), and the layers in red correspond to the high-level features (FC1-3)[5]. Animacy seems to be a major factor in the neural representations of the fully-connected layers of both CNNs, as the performance near-asymptotes within tens of principal components (PCs). In the right panels, the dimensions maximising the optimiser function for each layer are plotted. The black arrows point to the maxima. Hereafter, as a first test of the animacy organisation within CNNs (with Dataset 1), we use the first 493 PCs for VGG-16 layers and the first 313 PCs for AlexNet layers.

to dimensionality scaling. We can see the results in Figure 2.4. As seen in the left panels, both AlexNet and VGG-16 can categorise the shape-matched stimuli as animate and inanimate, especially the final layers. The decision scores corresponding to the performance shown in the left panel are robust to dimensionality scaling as indicated by the vertical green lines in the right panel (extremely high rank-order correlations). This classification performance immediately calls the validity of using the reaction times from visual search tasks as representations of higher-level visual features into question. Now, until the animacy organisations, i.e. the distribution of decision scores, of the ventral visual stream and the CNNs are shown to be similar, in addition to the similarity between the overall representations in CNN layers and in the ventral stream, the validity of CNNs being 'better' models of higher-level visual features as found in the higher-level visual cortex cannot be established.

[5] *Conv* refers to convolutional layers and *FC* refers to fully-connected layers. VGG-16 has groups of convolutional layers, and here the Convx represents the final Conv in that group x (e.g. Conv3 refers to Conv3c).

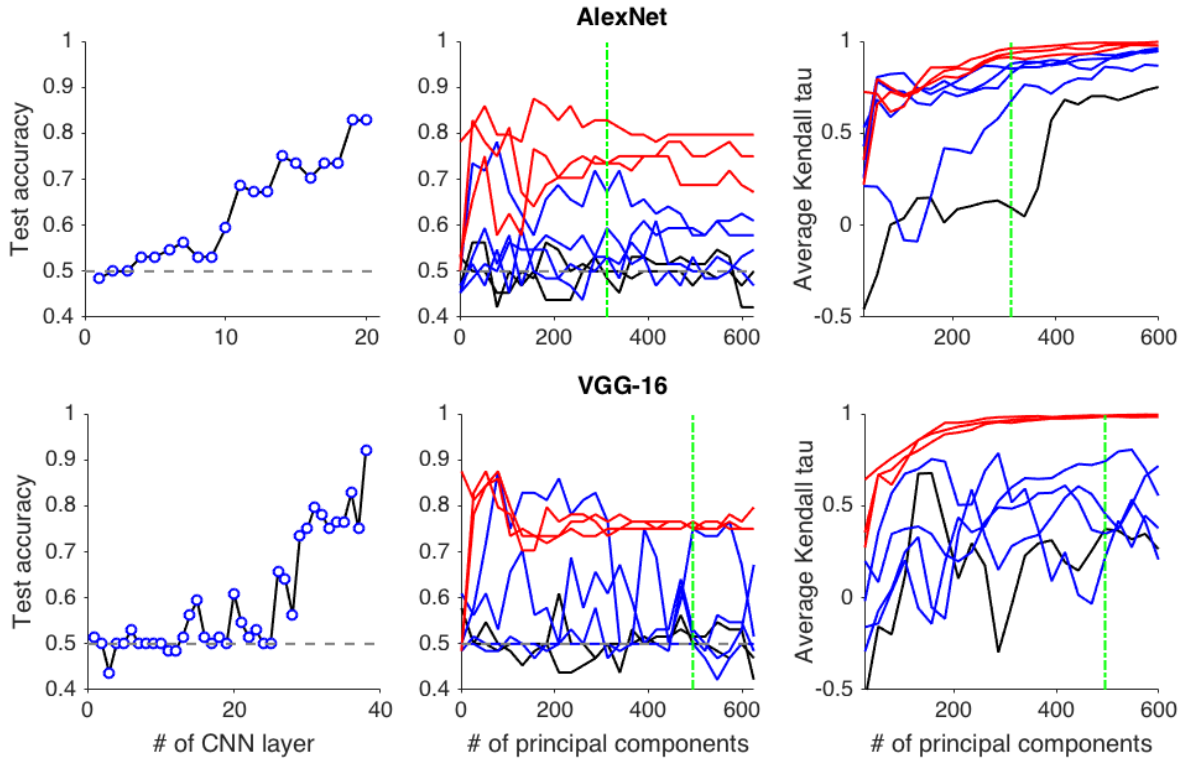


**Figure 2.4:** Classifying shape-matched stimuli with CNN feature differences obtained with Dataset 1. (left) Classification accuracy at the optimal number of principal components (PCs) for all the layers (including the ReLU and Pool layers) of the CNNs. (middle) Classification accuracy as a function of the number of PCs. The colors represent low-level features (black), mid-level features (blue) and high-level features (red) as explained in Figure 2.3. (right) Robustness to dimensionality scaling is measured with average rank-order correlations between the decision scores corresponding to the two nearest data-points on the x-axis. The higher the average, the more robust are the decision scores. The vertical green bars in the middle and right panels point to the number of PCs at which the classification accuracies for all CNN layers are plotted in the left panels. Both the CNNs can categorise the shape-matched stimuli from Proklova et al. (2016), calling the use of the visual search task to quantify high-level visual features as in the ventral temporal cortex into question.

As mentioned in the beginning of this subsection, we have fMRI data from Proklova et al. (2016), for Dataset 2. In order to extract the animacy organisation in the regions of interest in the brain, we will use this fMRI data to train a SVM. For a fair comparison, and to capitalise on the same image features as the brain data based classifier would capitalise on (given that the neural representations of interest in the ventral visual stream are indeed being driven visually), we trained SVMs on CNN layers using Dataset 2. The training procedure is the same as for Dataset 1, but the test procedure differs. Instead of subtracting the training mean from the neural activations for the shape-matched images in the CNN layers, we subtract the test set mean. In the case of Dataset 1, we had 960 training images, but now we only have 72, and we are testing on 64 images. The mean of the training images might not be representative of the population mean in this case. We obtain a new set of decision scores from SVMs trained over each CNN layer. The comparison between the training and test performance of the SVMs on VGG-16 layers along with the within training-type (with Dataset 1 or 2) decision score comparisons and between training-type comparisons are shown in Figure 2.5 (see Figure B.3 for the same analysis performed with AlexNet). The classification accuracy profiles are similar for the SVMs trained with the two datasets. The between training-type similarities between the decision scores are low throughout the CNN except for the final layers. These layers also correspond to the rapid rise in performance observed in the top-left panel, suggesting that the

rise in classification performance happens with similar rearrangement of object representations in the feature space, given different datasets. We cannot say this for sure as a similar observation is hard to make in the between training-type decision score similarities in AlexNet, as seen in Figure B.3.

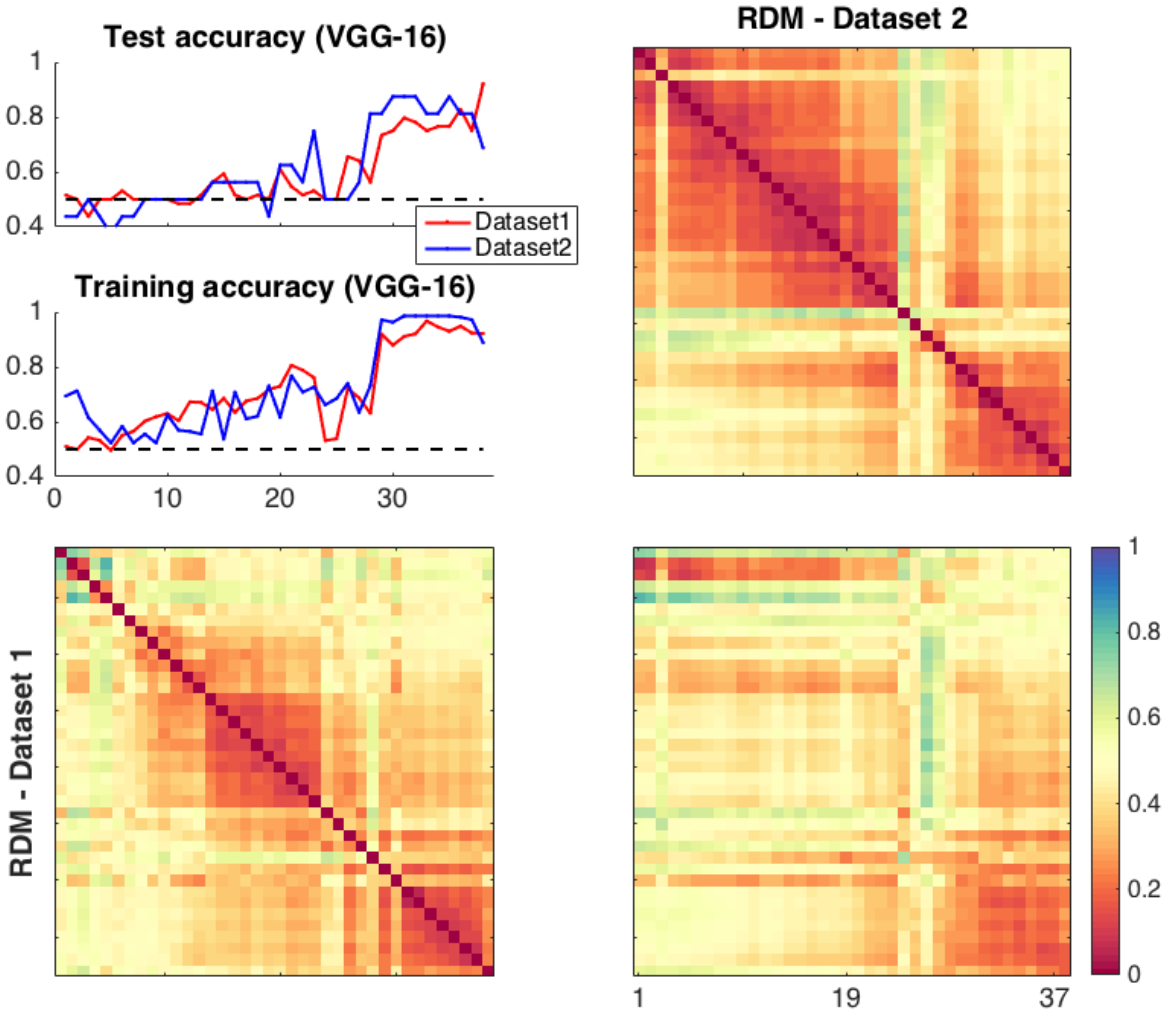


**Figure 2.5:** Comparison of the properties of the SVMs trained on VGG-16 layers for the shape-matched stimuli, with Datasets 1&2. (top, left) The classification accuracies on the shape-matched stimuli (test accuracy), and the 6-fold crossvalidation accuracies over the training data are plotted for SVMs trained with the two datasets. The performance trends are quite similar. (top, right) The pairwise dissimilarities (1 - Kendall's tau; also called the representational dissimilarity matrix (RDM)) between decision scores over VGG-16 layers are plotted for SVMs trained with Dataset 2, and similarly (bottom, left) with Dataset 1. The prominence of similarity around the diagonals indicates a locally-correlated decisions scores evolving through the CNN. To check if similar decision scores are generated by using the two Datasets, we plotted (bottom, right) the pairwise dissimilarities between decision scores from the SVMs trained with Dataset 1 (y-axis) and those from the SVMs trained with Dataset 2 (x-axis). The diagonally-proximate similarity structure observed in the two RDMs is absent in this RDM, suggesting that the decision scores are not the same. The final layers, Convolutional layer group 5 onwards, show a similarity structure indicating that the decision scores in these regions are similar to some extent.

How do the animacy organisations compare across the CNNs? A pairwise dissimilarity matrix for the decision scores of the two CNNs can be seen in Figure 2.6. We also show the comparison between the overall representations in the two CNNs. As can be seen in the figure, both the overall representations and the animacy dimensions are similar between the CNNs (but

see Figure B.4). This instils confidence in the idea of optimal transformations needed towards object recognition which are relatively robust to change in architecture, but a full treatment of this idea is out of the scope of this work.

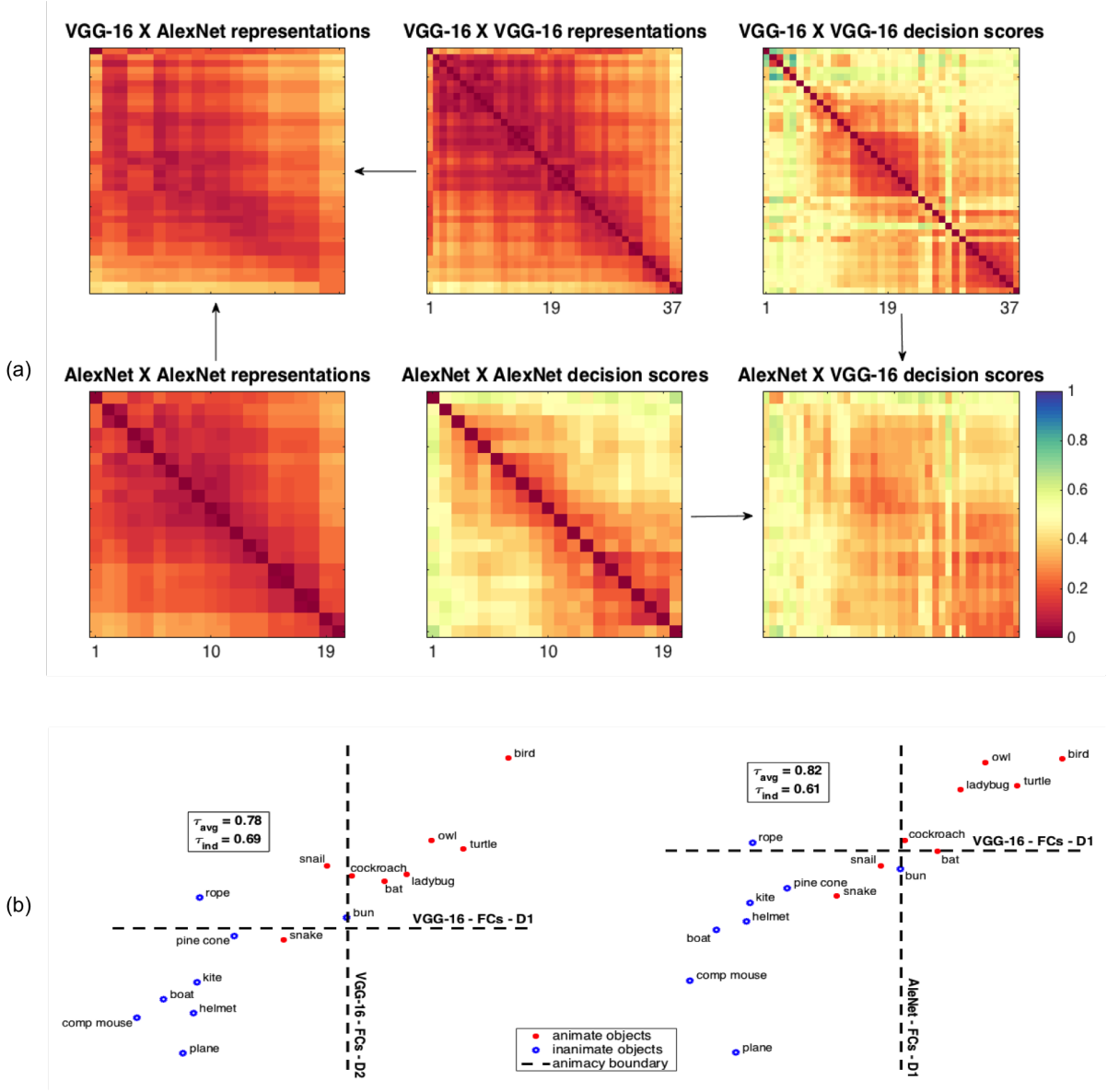


**Figure 2.6:** (a) Comparing the overall representations and animacy coefficients for the shape-matched stimuli (with Dataset 1) in AlexNet and VGG-16. All the matrices are representational dissimilarity matrices (RDM) encoding the pairwise dissimilarities between either the overall representations (neural activation similarity in each layer for the shape-matched stimuli) or the decision scores as indicated. The between-CNN RDM for overall representations shows strong similarity structure (as high as the within-CNN similarity structure across layers. The between-CNN RDM for decision scores shows considerable similarity within the middle and higher layers of the CNNs. (b) Visualising the similarity between the animacy coefficients between the use of different training datasets (left; with VGG-16) and between CNNs (with Dataset 1). The decision scores for all the fully-connected (FC) layers were normalised around their decision boundaries and the mean decision scores were obtained for each CNN for each dataset. $\tau_{avg}$ is the rank-order (Kendall's tau) correlation for the mean decision scores and $\tau_{ind}$ is the average of the pairwise correlations between the decision scores of the corresponding FC layers for different CNNs or datasets.

To summarise our observations,

- Convolutional neural networks (CNNs) process visual information in a locally-driven, hierarchical manner. Deeper layers encode higher-level visual features such as object parts, whereas shallower layers encode low- and mid-level features such as gabor filters, edges and simple shapes.
- Feature differences in the CNN layers can be used to classify animate and inanimate objects considerably well. They can also be used to classify the shape-matched stimuli from Proklova et al. (2016) which the behavioural measures based on a visual search task could not classify.
- The animacy organisation found in the final layers of the CNNs (AlexNet and VGG-16) are somewhat robust across architectural differences and training dataset differences, providing us with trustworthy measures of the animacy organisation that is driven purely by visual feature differences which were developed for the task of object recognition.

We shall now look at the comparison between the brain and CNNs, for the representations of objects and the animacy organisation, which will help us gain insights about the role of visual information in driving the animacy organisation in the ventral visual stream.

# Chapter 3

# The contribution of the differences in visual features to the animacy organisation in the brain

In Chapter 2, we learnt that the feature differences in convolutional neural networks (CNNs) can classify animate and inanimate objects, including the shape-matched stimuli from Proklova et al. (2016). To understand if the differences in visual features, as gauged with the CNNs, lead to the same animacy organisation as observed in ventral visual stream, we need to extract the animacy dimension of the regions of interest in the ventral visual stream, and compare the animacy coefficients therein with the animacy coefficients obtained from the CNN layers.

## 3.1 Regions of Interest

The various named regions in the ventral visual stream, with huge overlaps, where the animacy organisation is observed, are the ventral temporal cortex (VT/VTC) (Haxby et al., 2001, 2011), the lateral-occipital complex (LO/LOC) (Connolly et al., 2012; Sha et al., 2015), and the inferotemporal cortex (IT) (**?**Khaligh-Razavi and Kriegeskorte, 2014). Another region is used in studying the animacy organisation in Sha et al. (2015), which can be called the object-categorisation cortex (OCC), as it is defined functionally with the accuracies of the classification of objects. We now mention some of the definitions in previous work, and the ones we use.

**Ventral temporal cortex (VTC)** As mentioned in Haxby et al. (2011), "The region extended from $-70$ to $-30$ on the y-axis in Talairach atlas coordinates (Talairach & Tournoux 1988). The region was drawn to include the inferior temporal, fusiform, and lingual / parahippocampal gyri." Translating to the Montreal Neurological Institute (MNI) coordinates, we use the $-71$ to $-29$ bound on the y-axis. We identified the gyri using Automated Anatomical Labelling (AAL) parcellation (Tzourio-Mazoyer et al., 2002), as the fMRI data in (Proklova et al., 2016), which we are using, was transformed to the MNI space.

**Lateral-occipital complex (LOC)** The region is defined in a complex functional manner in Connolly et al. (2012). It was obtained as one of the clusters from clustering the representational similarity matrices from searchlights in the ventral visual stream (which identified LOC and the early visual cortex as two major clusters). Another definition of LOC provided by Grill-Spector et al. (2001), defines it as the regions that " were significantly more activated when subjects viewed photographs of cars or abstract sculptures compared to textures", and contains "the lateral surface (LO) near the lateral occipital sulcus and in ventral occipito-temporal regions (LOa/pFs) extending into the posterior and mid fusiform gyrus and occipito-temporal

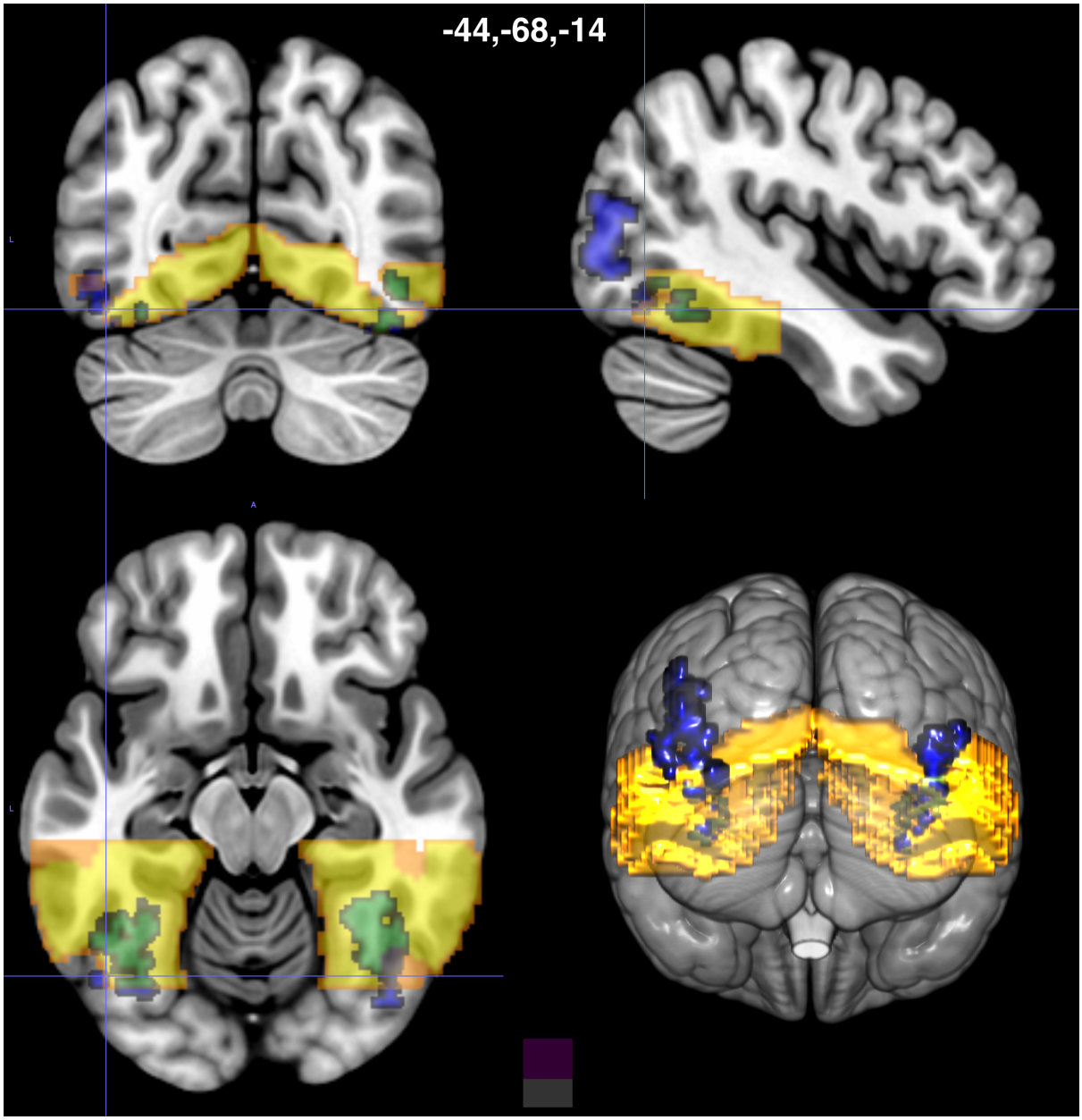


**Figure 3.1:** Visualising the xVAC (in blue) and VTC (in yellow) ROIs, and their overall (in green). The xVAC extends into more superior-posterior cortex, corresponding to the lateral animate-selective regions mentioned in Proklova et al. (2016).

sulcus". This definition is similar to the object-selective cortex (OSC) which is defined as the region that is significantly more activated for intact object images than scrambled object images. The LOC seems to overlap considerably with VTC.

**Inferotemporal cortex (IT)** In the supplementary materials of Kriegeskorte et al. (2008b), it is mentioned that the IT mask was defined "within the inferior temporal portion of the bilateral cortex mask", with no clarification about the extent of the 'inferior temporal portion'. In Bell et al. (2013), the IT is defined through legion studies and includes sections of middle, inferior and fusiform gyri, at locations more anterior laterally than the VTC.

In addition to these regions, we have another ROI, the extra-visual animacy cluster (xVAC) which was the cluster where the contribution of category to the representations was signifi-

cant for shape-matched stimuli in (Proklova et al., 2016), as described in Section 1.1. xVAC partly overlaps with VTC (it is more posterior and also has a superior extension), as shown in Figure 3.1. We only analyse the representations of VTC and xVAC in this report.

## 3.2 VTC and CNNs - Representational similarities

We present the comparisons between the representations in the VTC and the CNNs, for the shape-matched stimuli. The representation similarities were gauged through beta values for each condition (or stimulus class) averaged across runs.

To gauge the agreement over the representational similarities between the subjects, we use a noise ceiling estimate (Nili et al., 2014). The noise ceiling is the average of single subject representational similarities correlated with the average of the representational similarities of all other subjects. If the average of single subject correlations with any other model representational similarity matrix is higher than this noise ceiling, then that model is a true candidate model of those representations.

The comparisons between VTC, the CNNs, and the behavioural visual similarity measures acquired from visual search tasks in Proklova et al. (2016) are shown in Figure 3.2. We selected the 300 most visually responsive (average of beta values across condition) from the VTC as the procedure maximised the noise ceiling of VTC (agreement between subject representational similarities). Similar trends are observed in comparisons with xVAC, as seen in Figure B.5.

None of the ROI-CNN correlations hit noise ceiling suggesting that the overall representational structures in the CNNs and the VTC/xVAC have their differences, although the similarities are not insignificant either. The significantly positive correlations between the ROI representational dissimilarity matrices (RDM) and the behavioural measures' RDMs suggest that the visual similarities as gauged by these measures partly match the representational similarities in these ROIs. These behavioural measures do not contain information about animacy, as seen previously in Figure 1.3, which is reflected in the first principal component of VTC representations and VGG-16 representations, as seen in Figure 3.2b, but the CNNs do. The CNNs also correlate significantly with the overall visual similarities (see: Figure B.6), which tells us that the CNNs are not just capitalising on the animacy dimension towards their similarity with VTC. But to qualify CNNs as better representations of visually-driven representations in VTC, we need to assess the similarities between the animacy dimension.

The final layers of the CNNs contain animacy organisations as do VTC and xVAC. Animacy is one dimension[1] of the representations in the ROIs and the CNNs. The correlations of the CNN RDMs with the ROI-based RDMs can only hit noise ceiling if all such dimensions match in both their organisations and their contributions to the overall representation. It might be the case that multiple dimensions match between CNNs and the ROIs, but their contributions to the overall representations are different, leading to lower correlations between their RDMs.

To explore the similarities between the animacy dimensions of the CNN and our ROIs in the ventral visual stream, we need to extract the animacy dimensions of the ROIs.

## 3.3 The animacy organisation in xVAC and VTC

In order to obtain the animacy dimension in the ROIs, we train support vector machines (SVMs) for animacy classification and as shown in Figure 2.2, the dimension perpendicular to the decision hyperplane is deemed the animacy dimension. We now detail the procedure.

[1]There might be multiple significant animacy dimensions. We will explore this possibility in further work.

### 3.3.1 Training SVMs on the ROIs for animacy classification

We extract the preprocessed BOLD signal (unsmoothed) for each voxel in the ROI. As the hemodynamic response is delayed, we assign the image aligned to 6 s (2 s TRs were used in the EPI-based image acquisition) after stimulus presentation as the response to the stimulus. So, we have 64 images per run (correspond to the 64 images, 4 images each for the 16 objects), for 8 runs, from the main experiment (Proklova et al., 2016). We average the responses for the images of each object, across the 8 runs, and are left with 16 images, 8 animate and 8 inanimate. For the localiser, there were 16 blocks (8 for animate objects and 8 for inanimate objects), that lasted 16 s each (i.e. 8 TRs), so we have 128 images in total (64 each corresponding to animate and inanimate objects).

So, our training set comprises of 128 feature vectors (ROI mask applied on the whole brain mask) split between the animate and inanimate categories, and our test set comprises of 16 feature vectors split between the animate and inanimate categories.

We train a SVM on the centred (mean subtraction) training data, using the MATLAB (R2015b) *fitcsvm* function[2], and test the SVM on the centred test data. The SVM decision scores are termed the animacy coefficients of the objects in the given ROI.

The 6 fold crossvalidation accuracies on the training data for xVAC and VTC are 84% and 85% respectively. The test accuracies for xVAC and VTC are 79% and 73% respectively. We obtained the animacy coefficients for the 16 shape-matched stimuli from xVAC and VTC, for the 17 subjects, whose data was acquired in Proklova et al. (2016). The noise ceiling of the xVAC decision scores is 0.60 (Kendall's $\tau$), while that of the VTC decision scores is 0.51. The average of the subject-wise correlations between xVAC scores and averaged VTC scores is 0.61, and between VTC scores and averaged xVAC scores is 0.61. As the between-ROI correlations hit the noise ceiling, the animacy organisation can be said to be maximally similar, given the noise in the data, between these two ROIs.

## 3.4 Comparison between the animacy organisations in the CNNs and the ROIs

As mentioned in the previous section, the decision scores for VTC and xVAC are extremely similar. As xVAC is the region where extra-visual animacy encoding is claimed to exist (Proklova et al., 2016), we present comparisons between the two CNNs on the two Datasets, with which their SVMs were trained, and xVAC, in Figure 3.3. The comparison between the VGG-16 and VTC animacy coefficients can be seen in Figure B.7.

As seen in Figure 3.3, none of the animacy organisations in both CNNs, capitalising on features captured with both datasets, can fully explain the animacy organisation of xVAC. The animacy organisations in the later layers of the CNNs correlates highly with the xVAC animacy organisation, suggesting that most variance in the animacy organisation can be captured by differences in visual features. The middle panels of the figure indicate that the correspondence in the animate stimuli scores across the organisations might be higher than in the inanimate stimuli scores. In Figure 3.4, the correlations are shown separately for the animate and inanimate objects for the CNNs and datasets. The within-category animacy coefficients of some CNN layers are extremely similar to those of xVAC, more so for the animate stimuli. The within-inanimate structure in CNN layers does not reliably explain the corresponding xVAC structure (notably, the noise ceiling is lower than for animate-stimuli).

The significant anti-correlations between the initial layers of the CNNs (with Dataset 2) and the xVAC overall and within-category animacy organisations need to be explained. As seen in

[2]We did not optimise the hyperparameter *box constraint* in this phase of analysis, and used the default value.

Figure 2.5, the test accuracy of these layers is near chance (0.5), so the animate and inanimate stimuli are not being systematically misclassified. The animacy organisation (rank-ordered) of all the VGG-16 layers (trained with Dataset 2) can be seen in Figure 3.5. The initial layers show shape clusters along their animacy dimensions. This shape-based organisation correlates negatively with the xVAC organisation. Could this organisation also be reflected in the early visual cortex (EVC)? We trained SVMs over the EVC (Broadmann areas 17 (V1) and 18 (V2)). As seen in Figure 3.5, the EVC animacy organisation does not show the shape organisation. This shape-based organisation could be an artefact of the training procedure of the SVMs, and will be analysed in further work.

To summarise, the animacy organisations of the final layers of both AlexNet and VGG-16 are considerably similar to the animacy organisations of xVAC and VTC. Neither the overall nor the animacy representations of the two ROIs are fully explained (no noise ceiling hit) by the CNNs, but the within-category correlations are (although those are not as reliable - lower noise ceiling). We can say that the animacy organisation in VTC and xVAC reflects visual feature differences. In particular, xVAC, the 'extra-visual animacy cluster' might not be extra-visual after all. Nevertheless, our results show that the role of extra-visual contributions to the animacy organisation cannot be ruled out.

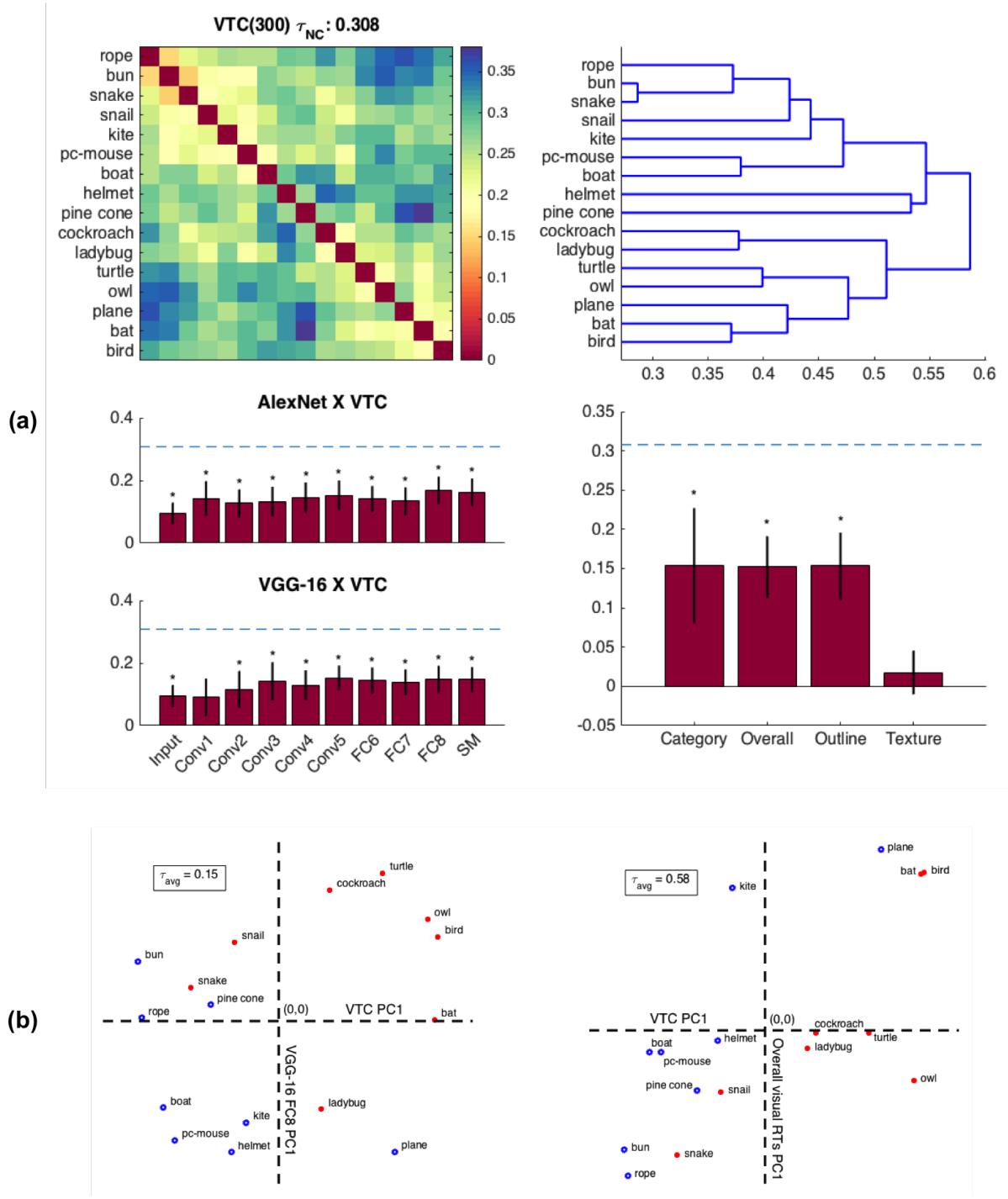


**Figure 3.2:** Comparing the representational similarities of VTC, CNNs and the behavioural measures. (a) (top, left) The representational dissimilarity matrix (RDM) averaged over the RDMs of the 300 most visually responsive voxels in each subject is shown. (top, right) The accompanying dendogram (obtained through agglomerative hierarchical clustering with the function *linkage* in MATLAB) already shows some animacy organisation. (bottom, left) The VTC RDMs were correlated (Kendall's $\tau$) with the RDMs from AlexNet and VGG-16 layers. Here Convx refers to the final Conv layer in the Convx group in VGG-16. Surprisingly, no trend is observed. (bottom, right) The VTC RDMs were correlated with the animate-inanimate category RDM, and the averaged behavioural RDMs. Category information is high and significant in VTC. The visual search task produced measures which are correlated with the VTC representational similarities. (b) The first principal components (PC1) of the VGG-16 FC8, average VTC, and average overall visual RTs' RDMs are visualised. Both the VTC and FC8 components show the animate-inanimate distinction, while the overall visual RTs do not.

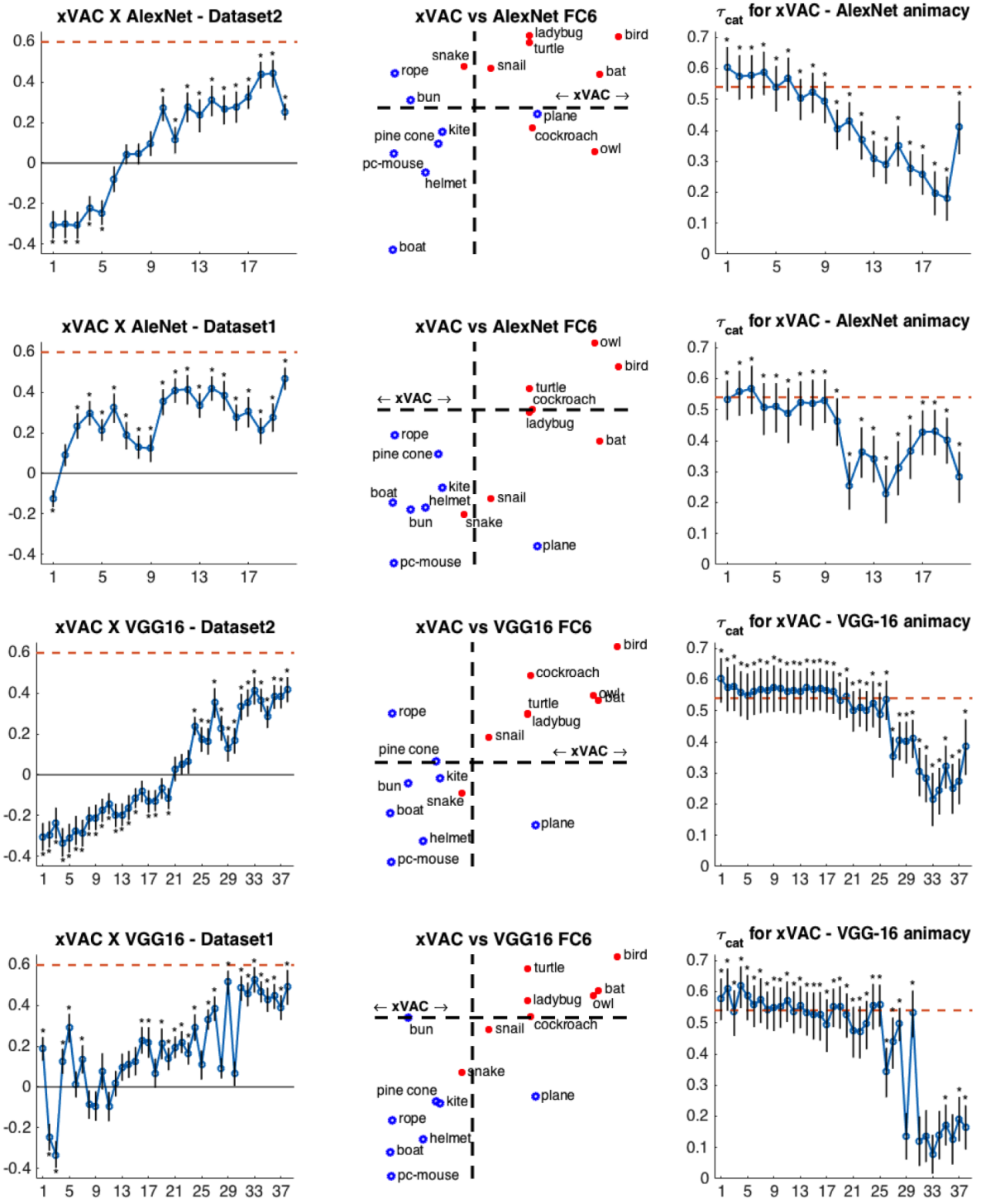


**Figure 3.3:** Comparing the animacy organisation of xVAC with the animacy organisations of AlexNet and VGG-16. (left) Kendall's $\tau$ correlations between the CNN animacy coefficients and xVAC animacy coefficients are shown, with the xVAC animacy noise ceiling (dashed line). None of the correlations across CNNs and datasets hit noise ceiling. (middle) The animacy coefficients of xVAC and CNNs' fully-connected (FC) layer 6 are shown. (right) The Kendall's $\tau$ correlations between the animacy category vector (1 for animate, $-1$ for inanimate) and the residual after regressing out the CNN layer's animacy coefficients from xVAC coefficients are shown. The category correlation with xVAC is shown as the dashed line. The later layers of the CNN seem to capture the animacy category structure of xVAC, as in the classification errors of the CNN layers and xVAC match to some extent.

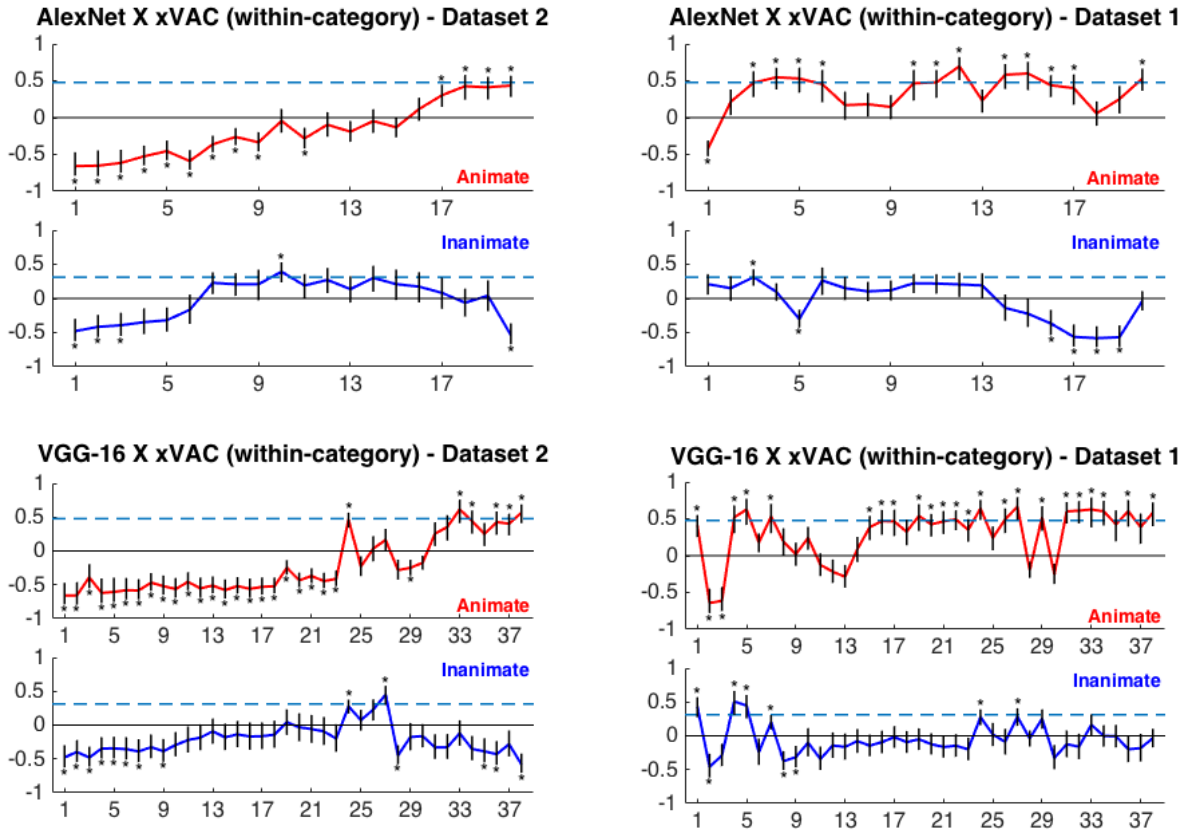


**Figure 3.4:** Comparing the within-category animacy organisation of xVAC with the animacy organisations of AlexNet and VGG-16. In each panel, the Kendall's $\tau$ correlations between CNN layer scores and xVAC scores, for the specified category of stimuli, are shown. the correlation profiles within animate stimuli look similar to the corresponding profiles for overall correlation as seen in Figure 3.3. The within-category organisation can be fully explained with the CNNs. There is a marked difference between the profiles of the animacy coefficients that were acquired using Datasets 1 and 2., specifically in the initial and middle layers of the CNNs.

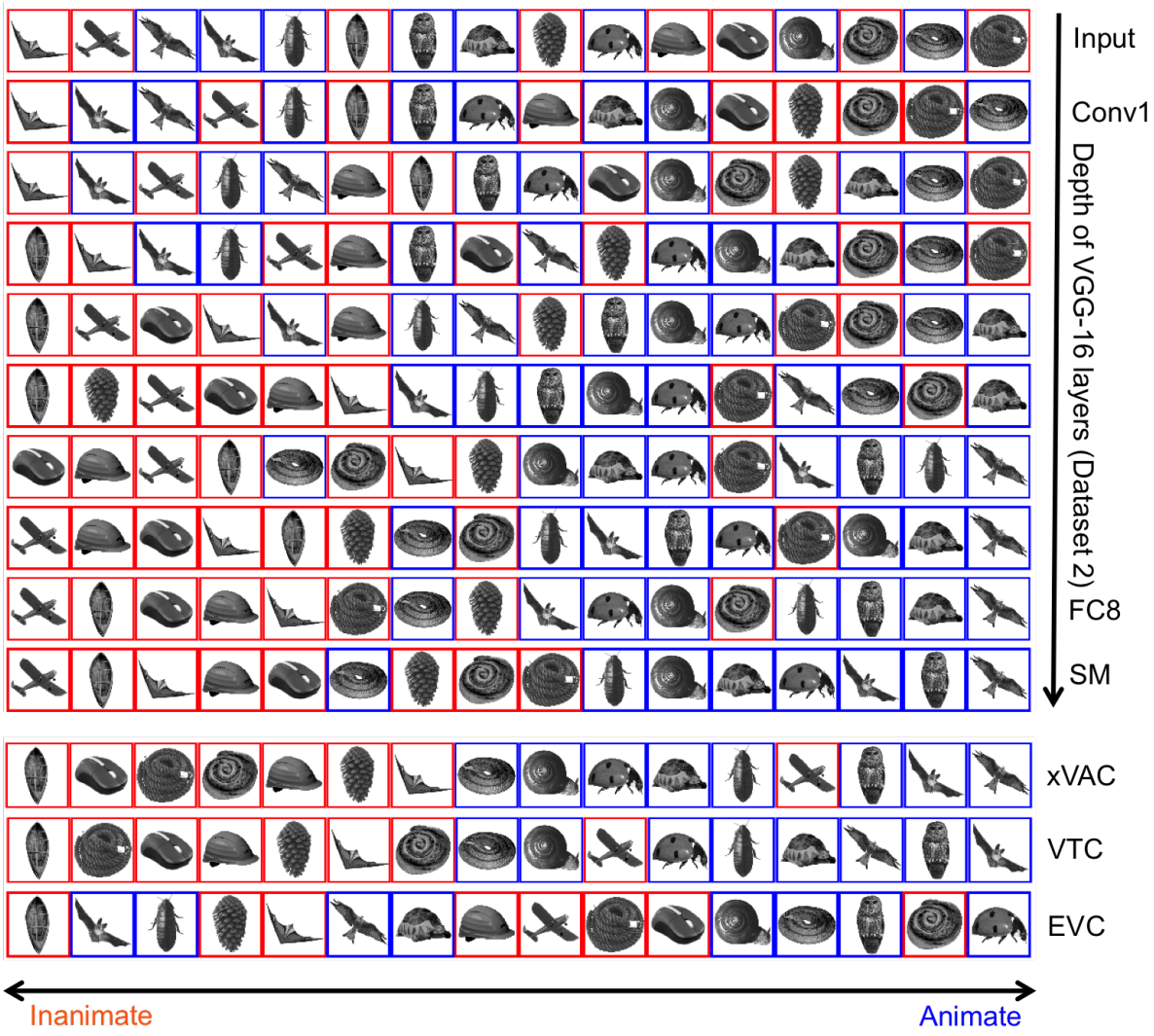


**Figure 3.5:** Visualising the animacy organisations of VGG-16 (Dataset 2), VTC, xVAC and EVC. In VGG-16 layers, we see the animacy organisation taking shape as we go deeper into the CNN. The input initial layers encode shape similarity. The final layers of VGG-16 and xVAC and VTC can considerably distinguish the shape-matched stimuli along their animacy dimensions. The early visual cortex (EVC, V1+V2, defined as Broadmann areas 17 and 18) cannot categorise the shape-matched stimuli.

# Chapter 4

# The contribution of the differences in non-visual features to the animacy organisation in the brain

Vignettes of simple shapes conveying social interactions and those conveying mechanical motion elicit distinct responses in the ventral visual stream (Martin and Weisberg, 2003). This neural activation profile matches that of the animacy organisation, as observed in Proklova et al. (2016) and other studies, leading us to hypothesise that the animacy organisation might also be driven by top-down information about object agency. What factors constitute object agency? We consider two factors, which are related, thoughtfulness and feelings. There might be other factors such as the association of identity to the object (which again, is related to the two mentioned factors), but we do not consider them in our current analysis. We want to know if these factors contribute to the animacy organisation observed in ventral visual stream.

The problem is, the space spanned by high-level visual features overlaps considerably with the space spanned by semantics (as gauged by corpus-based distributional semantics) (Frome et al., 2013), and the factors of thoughtfulness and feelings might play a role in the semantic representation of animals (CHECK). For good measure, we decided to acquire measures of thoughtfulness and feelings and then dissociate these measures from the animacy coefficients acquired from the final layer of a convolutional neural network (here, VGG-16's FC8), as we are interested in the contribution of agency to the animacy organisation in the ventral visual stream. We would like to dissociate agency from familiarity to ensure that we are not capitalising on any bias towards associating identity or agency because of familiarity with the animal.

To quantify the association of thoughtfulness and feelings to, and the familiarity of objects, we first ran a behavioural ratings experiment. Our objects of interest are animals, as humans have inherent associations of animacy for them as opposed to associations for non-animals requiring manipulations such as the mentioned vignettes. After we acquired the ratings, we selected a subset of the stimuli for whom the agency behavioural ratings, the familiarity ratings and the animacy coefficients from the final layer of a CNN are mutually dissociated.

We then ran an fMRI experiment to obtain the neural activity when images of these animals are presented. We would like to check for the existence of the animate-inanimate distinction with a univariate contrast, and analyse the contributions of the CNN visual features, agency, and familiarity to the neural representations in the brain and to the animacy organisation in the ventral visual stream. Preliminary results are presented here.

## 4.1 Behavioural ratings experiment and stimuli selection

### Stimuli

We collected 4 colour images each for 40 animals from various sources indexed in Google Images. The animals were cropped out of the images and 400 px x 400 px images with transparent backgrounds were created. A montage of the 160 images we use can be seen in Figure B.9.

### Experimental design

In the ratings task, all the four exemplars of a particular animal were shown to the subject and they were asked to rate either the thoughtfulness, feelings or familiarity of the animal, on a scale of 0 to 100. The experiment consisted of 3 blocks. In each block, all the 40 animals were asked to be rated sequentially (different screens for each animal). Before each block, the subject was given information about the factor to rate the animals on. For each factor, subjects were asked to "consider factors such as (but not limited to)",

- Familiarity - amount of interaction, knowledge about the animal, occurrence, etc.
- Thoughtfulness - planning, intentions, abstraction, etc.
- Feelings - empathy, sensation, reactions, etc.

The order of blocks and the stimuli in each block were randomised for each subject.

### Results

16 subjects (9 females) participated in the study. We scaled the ratings to lay between 0 and 1. The results of the correlations between the factors and the VGG-16 animacy coefficients are stated in Figure 4.1. As the similarity between the familiarity and feelings ratings are high, we average them into the 'Agency' rating. Previous work (Khaligh-Razavi and Kriegeskorte, 2014) has shown that the high-level visual cortex representations are similar to the representations in the final layers of the CNNs. Our work in comparing the animacy organisation in the ventral visual stream with the animacy organisation in CNNs shows the same trend. Hence, we decided to select a set of stimuli which could dissociate between the agency and familiarity ratings and VGG-16 FC8 animacy coefficients. The average correlations (Kendall's $\tau$) between the individual agency and familiarity ratings and the VGG-16 FC8 animacy coefficients, before and after stimuli selection, are as follows -

1. Agency x VGG16 FC8 - (before) 0.30, (after) 0.04
2. Familiarity x VGG16 FC8 - (before) 0.02, (after) 0.02
3. Agency x Familiarity - (before) 0.23, (after) 0.03

The pairwise dissociations are successful with the 12 animals chosen which are also highlighted in the bottom panels of Figure 4.1. Surprisingly, these dissociations were not successful when tested with AlexNet FC8 animacy coefficients, as seen in Figure B.8, which is a point of concern. We shall analyse this difference later.

## 4.2 fMRI experiment - Methods and Results

### Main fMRI experiment stimuli

We use the 12 animals for which the agency and familiarity ratings and the VGG-16 animacy coefficients are mutually dissociated (4 images for each animal). In addition to them, we use

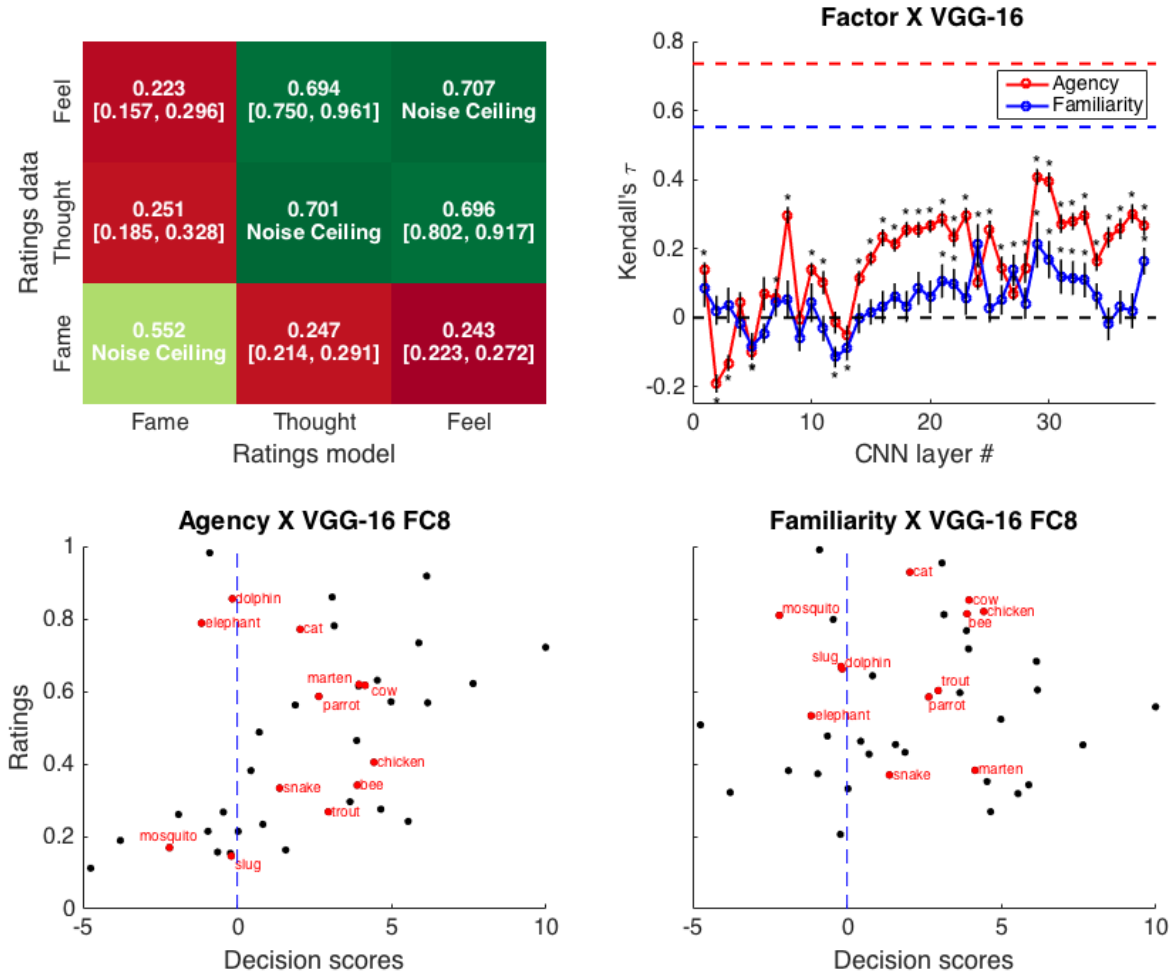


**Figure 4.1:** Similarities between the non-visual factors and VGG-16 decision scores (Dataset 1). (top, left) To gauge the similarities between the factor ratings, we compute the pairwise correlations (Kendall's $\tau$) between the individual ratings for each factor (Ratings data) and the average ratings for each factor (Ratings model). The noise ceiling (Nili et al., 2014) estimates for each factor are mentioned on the diagonal. The model-data correlations are extremely high (noise ceiling hit) for thoughtfulness and feelings. So, we average the two ratings for each subject and term the combined rating as 'Agency'. (top, right) The correlations of the two factors with the animacy coefficients from VGG-16 layers (495 principal components each) are shown. The middle and higher layers show significant correlations with agency ($p < 0.05$, *bonferroni correction*), while the correlations with familiarity are weaker. The similarity between the (bottom, left) agency ratings and (bottom, right), and VGG-16 FC8 animacy coefficients are shown. The stimuli selected to minimise the correlation are shown in red. FC8 is among the final layers of the CNN which have been shown to model the representations in high-level visual cortex.

the 4 images of humans, and 4 images each for inanimate objects - chairs, cars and small trees. A montage of the 64 images can be seen in Figure B.10.

Stimulus presentation was controlled using the Psychtoolbox (Brainard and Vision, 1997). Images were back-projected on a translucent screen placed at the end of the scanner bore. Participants viewed the screen through a tilted mirror mounted on the head coil. Stimuli were presented foveally and subtended a visual angle of approximately 4.5°.

### Participants

17 participants (7 females; mean age = 25.2 years, SD = 3.1 years) were scanned at the Center for Mind/ Brain Sciences of the University of Trento. All participants gave informed consent. All

procedures were carried out in accordance with the Declaration of Helsinki and were approved by the ethics committee of the University of Trento.

### Main fMRI experiment procedure

The procedure of the main fMRI experiment is the same as in (Proklova et al., 2016), quote - "The main fMRI experiment consisted of eight runs. Each run consisted of 80 trials that were composed of 64 object trials and 16 fixation-only trials. In object trials, a single stimulus was presented for 300 msec, followed by a 3700 msec fixation period. In each run, each of the 64 images appeared exactly once. In fixation-only trials, the fixation cross was shown for 4000 msec. Trial order was randomized, with the constraints that there were exactly eight 1-back repetitions of the same category (e.g., two snakes in direct succession) within the object trials and that there were no two fixation trials appearing in direct succession. Each run started and ended with a 16 -sec fixation period, leading to a total run duration of 5.9 min. Participants were instructed to press a button whenever they detected a 1-back repetition."

### Functional localizer experiment procedure

Participants completed two runs of a functional localizer experiment, adapted from the one used in (Proklova et al., 2016). During the localizer, participants viewed grayscale pictures of 36 animate, 36 inanimate stimuli, 18 intact (inanimate) objects and 20 images of scrambled objects in a block design (The conditions which had less than 36 images had duplications introduced to create sets of 36 images). The images used can be seen in Figure B.11. These stimuli were not matched for their shape (thus, this design resembled the standard animate-inanimate and object-scrambled contrasts used in previous studies). Each block lasted 16 s, containing 18 stimuli (the 36 images were distributed evenly over the two runs) that were each presented for 400 msec, followed by a 400 msec blank interval. In each run, there were four blocks of each stimulus category and four fixation-only blocks per run. The order of the first 10 blocks was randomized and then mirror reversed for the following 10 blocks. Participants had to detect 1-back image repetitions, which happened twice during every non-fixation block.

### fMRI preprocessing and modelling

The preprocessing of the fMRI data and the modelling for the main experiment are the same as in (Proklova et al., 2016), quote - "The neuroimaging data were analysed using MATLAB and SPM8. During the preprocessing, the functional volumes were realigned, co-registered to the structural image, re-sampled to a $2 \times 2 \times 2\,\mathrm{mm}^3$-voxel grid, and spatially normalized to the Montreal Neurological Institute 305 template included in SPM8. For the univariate analysis, the functional images were smoothed with a 6 mm FWHM kernel, whereas for the multivariate analysis, the images were left unsmoothed. For the main experiment, the BOLD signal of each voxel in each participant was modelled using 22 regressors in a general linear model, with 16 regressors for each of the objects (e.g., one regressor for all snakes) and six regressors for the movement parameters obtained from the realignment procedure." For the functional localiser experiment, the signal was modelled using four regressors (animate, inanimate, intact, and scrambled objects) and six movement regressors. All models included an intrinsic temporal high-pass filter of 128 Hz to correct for slow scanner drifts.

### Univariate Analysis

Univariate random effects whole-brain analyses were performed separately for the localizer and the main experiment, contrasting animate with inanimate objects. Statistical maps were thresholded using a voxel-level threshold of $p < .005$ (uncorrected). In addition, regions activated in

the localizer were defined as ROIs. Within these ROIs, beta estimates for the conditions of the main experiment were extracted and averaged across the voxels of each ROI.

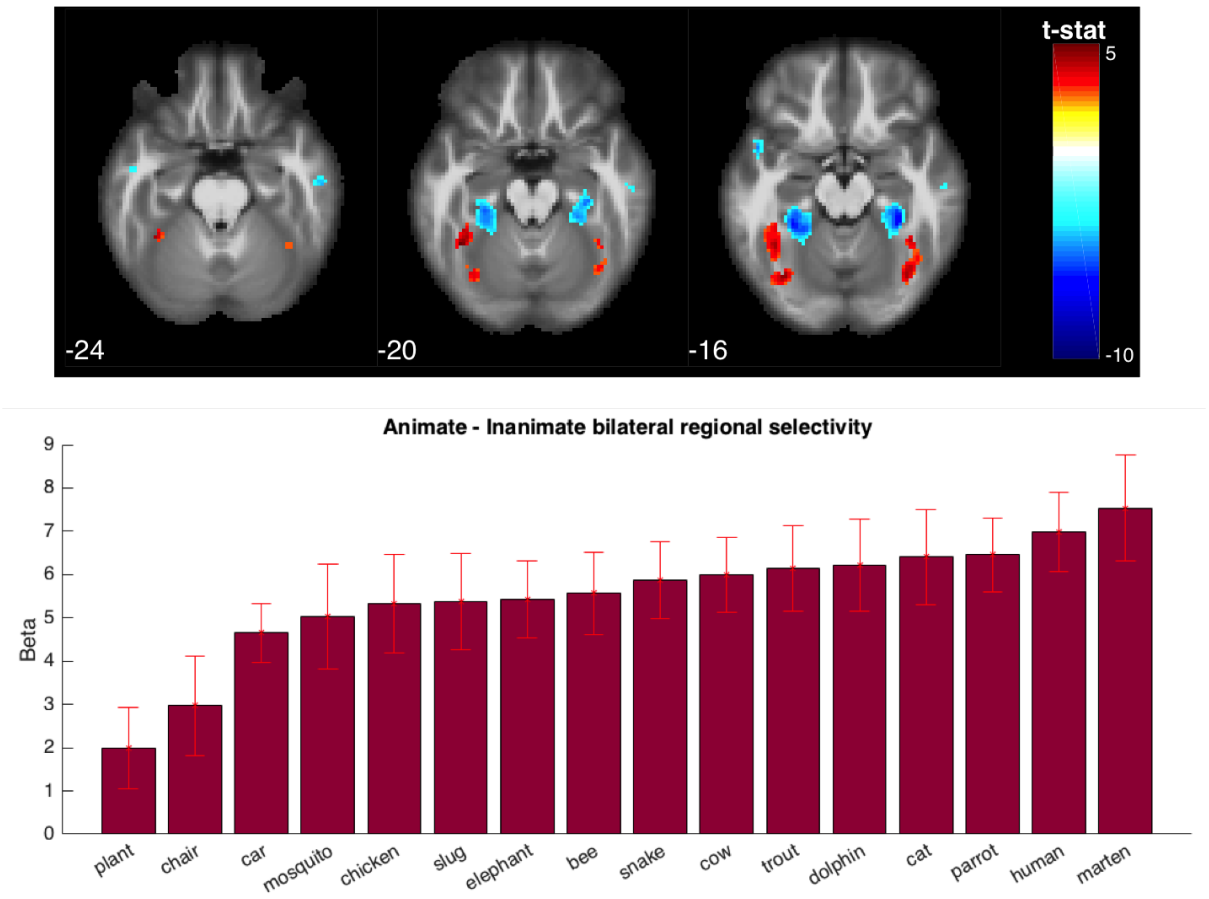


**Figure 4.2:** Univariate analysis results. (top) Emergent clusters in the ventral temporal cortex when the statistic is thresholded at $p < .005$ (uncorrected). The medial clusters are inanimate-selective (in blue) and the lateral clusters are animate-selective (in red), echoing the results of previous studies (Proklova et al., 2016). (bottom) The difference of the beta scores, averaged across runs, between the pooled animate- and pooled inanimate-selective clusters are plotted for each condition. Overall, the animate stimuli evoked higher responses in the animate-selective regions and lower responses in the inanimate-selective regions than the inanimate stimuli did.

The contrast between animate and inanimate objects in the main experiment revealed a characteristic medial-to-lateral organization in VTC (the contrast returned weaker results in the functional localizer experiment, but the medial-to-lateral organisation was observed). The results are shown in Figure 4.2. In line with previous findings, animate stimuli more strongly activated regions around the lateral fusiform gyrus (right hemisphere [RH]: 1232 $mm^3$, peak Montreal Neurological Institute coordinates: $x = 36$, $y = 72$, $z = 14$; left hemisphere [LH]: (cluster 1) 1496 $mm^3$, peak coordinates: $x = -42$, $y = 52$, $z = 20$, (cluster 2) 1496 $mm^3$, peak coordinates: $x = -36$, $y = 72$, $z = 18$), and inanimate stimuli preferentially activated more medial fusiform regions bordering the parahippocampal gyrus (RH: 3840 $mm^3$, peak coordinates: $x = 32$, $y = 38$, $z = 10$; RH: 760 $mm^3$, peak coordinates: $x = -30$, $y = 44$, $z = 8$). These clusters are at similar locations to those reported in Proklova et al. (2016). Other significant cluster were found too, but will be analysed in further work as they do not lie in the region of interest - the ventral temporal cortex.

We reduced the threshold to $p < .008$ (uncorrected), so we the two animate-selective clusters in the left hemisphere could fuse given us two clusters (animate and inanimate selective) each for each hemisphere. We averaged the betas for each condition (16 stimuli) across the 8 runs in

the main experiment and averaged the beta values of the two animate-selective clusters and the two inanimate-selective clusters, and tested the difference of these two sets of beta values across conditions. A one-way ANOVA confirmed the main effect of condition ($F = 8.6$, $p = 10^{-15}$). So, the animate stimuli have higher overall animate-selectivity than the inanimate stimuli as seen in Figure 4.2. This is in line with previous reports of animacy being a major driving principle of the neural representations in the ventral visual stream.

### Multivariate Analysis

To understand the population codes employed in brain regions, we have to resort to multivariate analysis, mostly representational similarity analysis (RSA) (Kriegeskorte et al., 2008a). We would like to check the representational similarities of the responses evoked by the stimuli in three regions - the ventral temporal cortex (VTC), the object-selective cortex (OSC), and the proposed (see: Section 1.1) extra-visual animacy cluster (xVAC). I present the results for VTC.

**ROI definition:** The definition of VTC here follows the definition used in Haxby et al. (2011) and Haxby et al. (2001). The volume of interest (VOI) of VTC extends from $-71$ to $-29$ on the y-axis in the Montreal Neurological Institute coordinates. The region was drawn to include the bilateral inferior temporal, fusiform, lingual, and parahippocampal gyri, as identified by Automated Anatomical Labelling (AAL) parcellation (Tzourio-Mazoyer et al., 2002).

The representational dissimilarity matrix (RDM), averaged across subjects, is shown in Figure 4.3. As also seen through the dendogram, the inanimate objects are most dissimilar to the animals. We won't read too much into the organisation, but the animate-inanimate distinction is present in the representations in VTC as evidenced by $\tau_{cat}$, which is the average of the correlations of individual RDMs with a animacy category matrix, whose value is close to the noise ceiling (NC) of the RDMs.

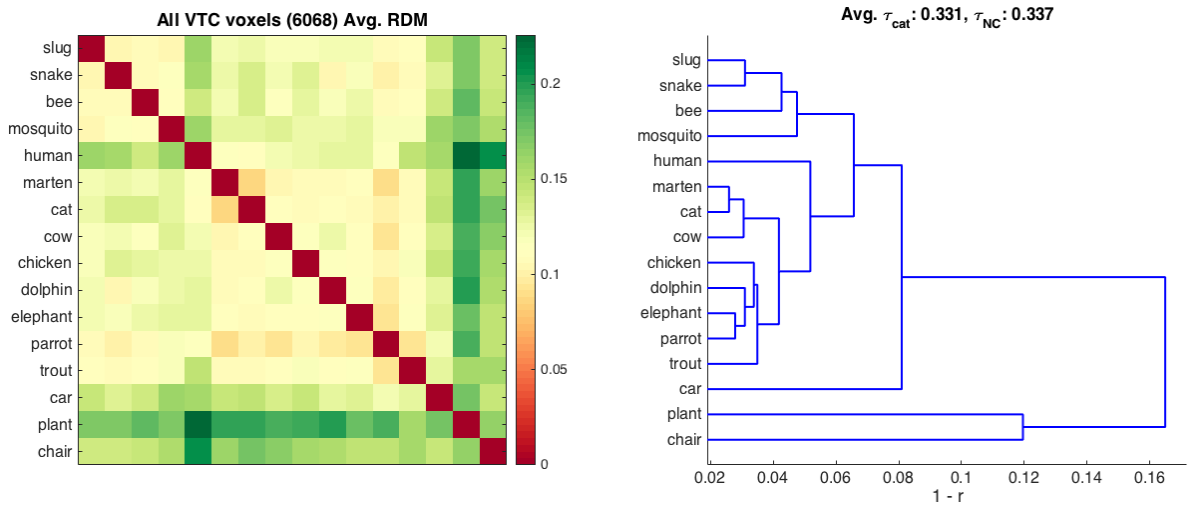


**Figure 4.3:** RSA for the ventral temporal cortex (VTC). (left) The RDM averaged across subjects is shown. (right) A dendogram was constructed out of the correlational similarities between the representations in VTC. We can see that the inanimate objects are most dissimilar to the animals. The average category correlation ($\tau_{cat}$) of the RDMs is close to the noise ceiling ($\tau_{NC}$) of the RDMs, suggesting that the animate-inanimate distinction plays a major role in VTC representations.

**Further work**

With the univariate results showing a medial-to-lateral animacy preference, we have replicated the findings of Proklova et al. (2016) and other such studies. With the multivariate analysis of VTC, we have shown that the animate-inanimate distinction is indeed a major driving principle in high-level ventral visual cortex as also reported in Sha et al. (2015) for the inferior temporal (IT) cortex, which intersects largely with our VTC definition here.

Further work will deal with understanding the contribution of the agency and familiarity ratings and the CNN animacy coefficients to the animacy coefficients of VTC, the object-selective cortex (OSC), and the previously proposed extra-visual animacy cluster (xVAC).

# Chapter 5

# Conclusion

Previous studies have shown the existence and the significance of the animacy organisation in the ventral visual stream (Kriegeskorte et al., 2008b; Sha et al., 2015; Proklova et al., 2016). What information drives this organisation? Animate objects such as animals have characteristic features that inanimate objects such as cars or tools do not possess. Could these visual feature differences be solely responsible for the animacy organisation? Vignettes of simple shapes conveying social interactions and those conveying mechanical motion elicit distinct responses in the ventral visual stream (Martin and Weisberg, 2003), similar to the animate-inanimate neural response elicited by images of animals and inanimate objects. This differential response is not driven by differences in visual features but differences in the association of agency and identity to the object. So, the animacy organisation can be driven by top-down input that does not reflect visual feature differences.

To ascertain the validity of this line of thought, Proklova et al. (2016) created a set of shape-matched stimuli that reflected no category information through their visual feature differences as gauged by a visual search task. They found a region (xVAC: the extra-visual animacy cluster) in the ventral visual stream that encoded animacy category information for these stimuli. This suggested that the animacy organisation is indeed not driven solely by visual feature differences even for static images of animals and inanimate objects. A point of concern in this analysis was the use of a visual search task to capture the visual feature differences. What if the behavioural task could not capture the visual features being accessed by xVAC to drive its animacy organisation?

Convolutional neural networks (CNNs) trained for the task of object recognition might address any limitations of the visual search task in quantifying visual feature differences. CNNs have been shown to be extremely good at large-scale visual object recognition (LeCun et al., 2015), and they possess an architecture inspired by the human visual system, with receptive fields and hierarchical transformations extracting low-level features such as edges to high-level features such as object parts (Zeiler and Fergus, 2014). The representations in the layers of the CNNs have been shown to hierarchically match the representations in the ventral visual stream (Khaligh-Razavi and Kriegeskorte, 2014; Cichy et al., 2016). We wanted to know if the CNNs possess visual features which could help classify the shape-matched stimuli from Proklova et al. (2016). To that end, we trained support vector machines (SVMs) on the CNN neural activations to classify animate and inanimate images from separate datasets. When tested on the shape-matched stimuli, the final layers of the CNNs, AlexNet (Krizhevsky et al., 2012) and VGG-16 (Simonyan and Zisserman, 2014), could classify them as animate or inanimate. So, the CNNs possess visual features beyond those captured through the visual search task in Proklova et al. (2016), which could help classify the shape-matched stimuli according to their animacy.

To understand if the animacy organisation (the SVM decision scores) in the layers of the CNNs match the animacy organisation in the ventral visual stream, we extracted the animacy organisations of the ventral temporal cortex (VTC) and xVAC. The final layers of the CNNs

possess animacy organisations that correlate highly with the animacy organisations of VTC and xVAC, but do not fully capture the variance therein. This suggests that the animacy organisation in the ventral visual stream could rely heavily on visual feature differences. However, the possibility that some visual feature differences could lead to a similar animacy organisation does not imply that the animacy organisation in the ventral visual stream is indeed being driven majorly by visual feature differences. We need to explore non-visual dimensions such as agency to understand if the animacy organisation can be driven by them. Only then can any claims about the drivers of the animacy organisation in the ventral visual stream be made.

Towards that end, we designed a new experiment where we obtained agency (thoughtfulness/feelings) and familiarity ratings for a set of animals. In further work, we shall test the differential contributions of these non-visual features and the visual features given by CNNs on the animacy organisation in the ventral visual stream.

The ventral temporal cortex (VTC) and other higher-level regions in the ventral visual stream are involved in object recognition (Grill-Spector and Weiner, 2014). The animate-inanimate organisation is a major organisational principle therein. CNNs trained for object recognition also encode the animate-inanimate distinction in the first principal components of their final layers. These CNNs were trained to orthogonalise object categories (so a cat is as different from dog as it is from a car), but to minimise their loss function, they seemed to be accepting animacy driven through visual feature differences as an error. Is the VTC encoding animacy for the same reason? Downstream areas might not be asking VTC to orthogonalise all object categories. Semantic clustering is what matters towards our knowledge of the world. We need to know a cat is more similar to a dog than it is to a car. Now this can both be learnt through visual feature differences and through non-visual associations (agency, can be pets, etc.). If the ventral temporal cortex is supposed to encode information as required by downstream areas, the loss function might call for generating semantic representations in a feed-forward fashion. This calls for aligning the visual and semantic spaces (as in Frome et al. (2013)) to ensure quick feed-forward information processing. Also, CNNs are trained in a supervised fashion with a million images. The human visual system is partly hard-wired and partly developed through childhood, although, as mentioned, the representational requirements are different than those of a CNN. What factors constrain the representations in the ventral visual stream is an open question for now (for a recent overview of the progress on this front, see Peelen and Downing (2017)).

# Appendix A

# Architectures of the CNNs in use

We use two convolutional neural networks in our analysis, AlexNet and VGG-16. We used pre-trained models from MatConvNet (Vedaldi and Lenc, 2015). The details about the architecture can be found below.

## A.1 AlexNet

AlexNet (Krizhevsky et al., 2012) has 20 layers, including the input, the convolutional, the ReLU, and the Max pooling layers. The indices, names, and number of elements in each layer are as follows, where 'Conv' denotes convolutional layers, 'ReLU' denotes the output of the ReLU operation, 'Pool' denotes the output of the max pool operation, 'FC' denotes fully-connected layers, and 'SM' denotes the output of the softmax operation -

| Index | Name | # of Elements |
|---|---|---|
| 1 | Input | $227 \times 227 \times 3$ |
| 2 | Conv1 | $55 \times 55 \times 96$ |
| 3 | ReLU1 | $55 \times 55 \times 96$ |
| 4 | Pool1 | $27 \times 27 \times 96$ |
| 5 | Conv2 | $27 \times 27 \times 256$ |
| 6 | ReLU2 | $13 \times 13 \times 256$ |
| 7 | Pool2 | $13 \times 13 \times 256$ |
| 8 | Conv3 | $13 \times 13 \times 384$ |
| 9 | ReLU3 | $13 \times 13 \times 384$ |
| 10 | Conv4 | $13 \times 13 \times 384$ |
| 11 | ReLU4 | $13 \times 13 \times 384$ |
| 12 | Conv5 | $13 \times 13 \times 256$ |
| 13 | ReLU5 | $13 \times 13 \times 256$ |
| 14 | Pool5 | $6 \times 6 \times 256$ |
| 15 | FC6 | $1 \times 4096$ |
| 16 | ReLU6 | $1 \times 4096$ |
| 17 | FC7 | $1 \times 4096$ |
| 18 | ReLU7 | $1 \times 4096$ |
| 19 | FC8 | $1 \times 1000$ |
| 20 | SM | $1 \times 1000$ |

## A.2 VGG-16

VGG-16 (Simonyan and Zisserman, 2014) has 38 layers, including the input, the convolutional, the ReLU, and the Max pooling layers. VGG-16 has groups of convolutional layers before a max pooling operation, and they are all numbered the same and a suffix is added to indicate the position of a convolutional layer in the group. The indices, names, and number of elements in each layer are as follows, where 'Conv' denotes convolutional layers, 'ReLU' denotes the output of the ReLU operation, 'Pool' denotes the output of the max pool operation, 'FC' denotes fully-connected layers, and 'SM' denotes the output of the softmax operation -

| Index | Name | # of Elements | Index | Name | # of Elements |
|---|---|---|---|---|---|
| 1 | Input | $224 \times 224 \times 3$ | 20 | ReLU4a | $28 \times 28 \times 512$ |
| 2 | Conv1a | $224 \times 224 \times 64$ | 21 | Conv4b | $28 \times 28 \times 512$ |
| 3 | ReLU1a | $224 \times 224 \times 64$ | 22 | ReLU4b | $28 \times 28 \times 512$ |
| 4 | Conv1b | $224 \times 224 \times 64$ | 23 | Conv4c | $28 \times 28 \times 512$ |
| 5 | ReLU1b | $224 \times 224 \times 64$ | 24 | ReLU4c | $28 \times 28 \times 512$ |
| 6 | Pool1 | $112 \times 112 \times 64$ | 25 | Pool4 | $14 \times 14 \times 512$ |
| 7 | Conv2a | $112 \times 112 \times 128$ | 26 | Conv5a | $14 \times 14 \times 512$ |
| 8 | ReLU2a | $112 \times 112 \times 128$ | 27 | ReLU5a | $14 \times 14 \times 512$ |
| 9 | Conv2b | $112 \times 112 \times 128$ | 28 | Conv5b | $14 \times 14 \times 512$ |
| 10 | ReLU2b | $112 \times 112 \times 128$ | 29 | ReLU5b | $14 \times 14 \times 512$ |
| 11 | Pool2 | $56 \times 56 \times 128$ | 30 | Conv5c | $14 \times 14 \times 512$ |
| 12 | Conv3a | $56 \times 56 \times 256$ | 31 | ReLU5c | $14 \times 14 \times 512$ |
| 13 | ReLU3a | $56 \times 56 \times 256$ | 32 | Pool5 | $7 \times 7 \times 512$ |
| 14 | Conv3b | $56 \times 56 \times 256$ | 33 | FC6 | $1 \times 4096$ |
| 15 | ReLU3b | $56 \times 56 \times 256$ | 34 | ReLU6 | $1 \times 4096$ |
| 16 | Conv3c | $56 \times 56 \times 256$ | 35 | FC7 | $1 \times 4096$ |
| 17 | ReLU3c | $56 \times 56 \times 256$ | 36 | ReLU7 | $1 \times 4096$ |
| 18 | Pool3 | $28 \times 28 \times 256$ | 37 | FC8 | $1 \times 1000$ |
| 19 | Conv4a | $28 \times 28 \times 512$ | 38 | SM | $1 \times 1000$ |

# Appendix B

# Supplementary figures

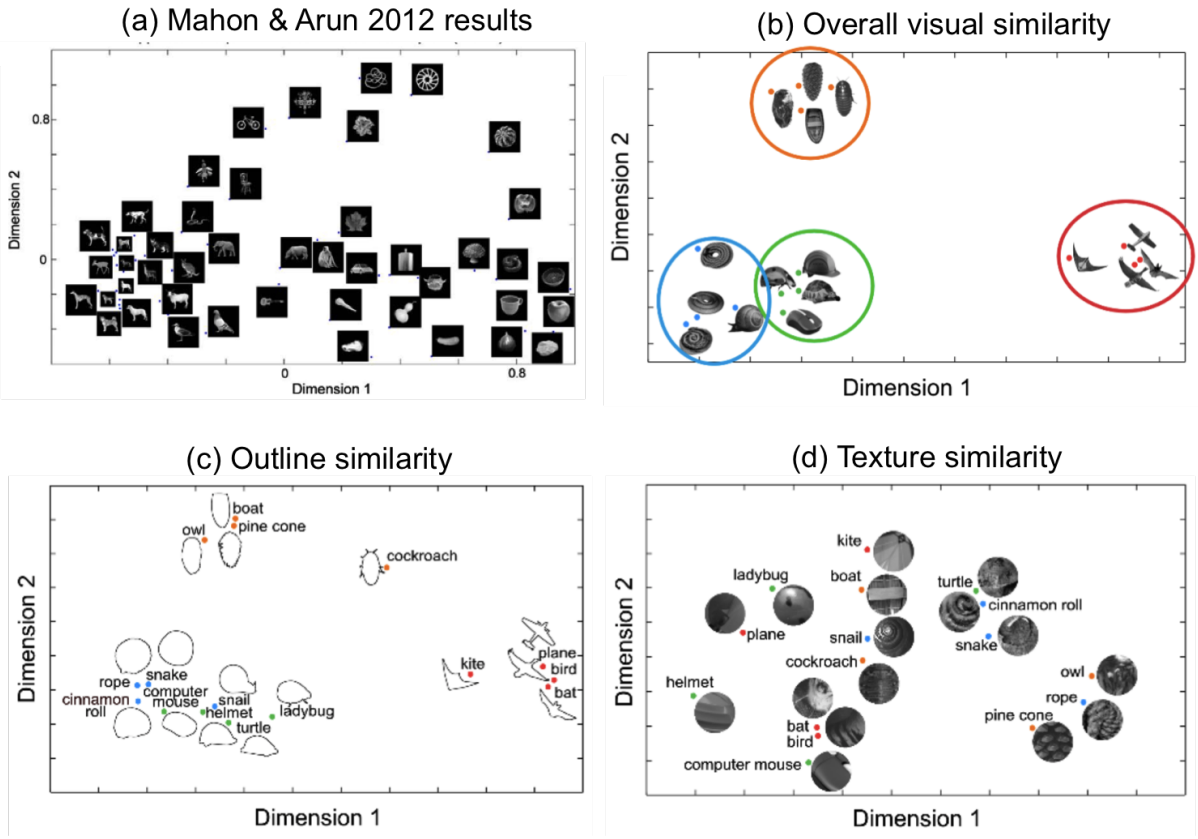


**Figure B.1:** Representational Space (Principal components 1 & 2) of visual search RTs for (a) the non-shape-matched stimuli used in Mohan and Arun (2012), (b) the full images of the stimuli used in Proklova et al. (2016), (c) the outlines of the stimuli, and (d) the textures of the stimuli. In (a) the animate stimuli cluster to the left part of the space making the animate-inanimate distinction apparent. In (b), (c), and (d), no animacy-based clustering is observed.
(Adapted from: Mohan and Arun (2012); Proklova et al. (2016))


**Figure B.2:** Visualisation of features in AlexNet. For each layer, on the left the deconvolved feature maps are shown, while on the right the corresponding image patches are shown. The deconvolved feature maps provide us with an idea of the complexity and specificity of the features developed in the various layers of the CNN. A simple-to-complex hierarchy of features is observed as we go deeper into the network. (Reproduced from: Zeiler and Fergus (2014))

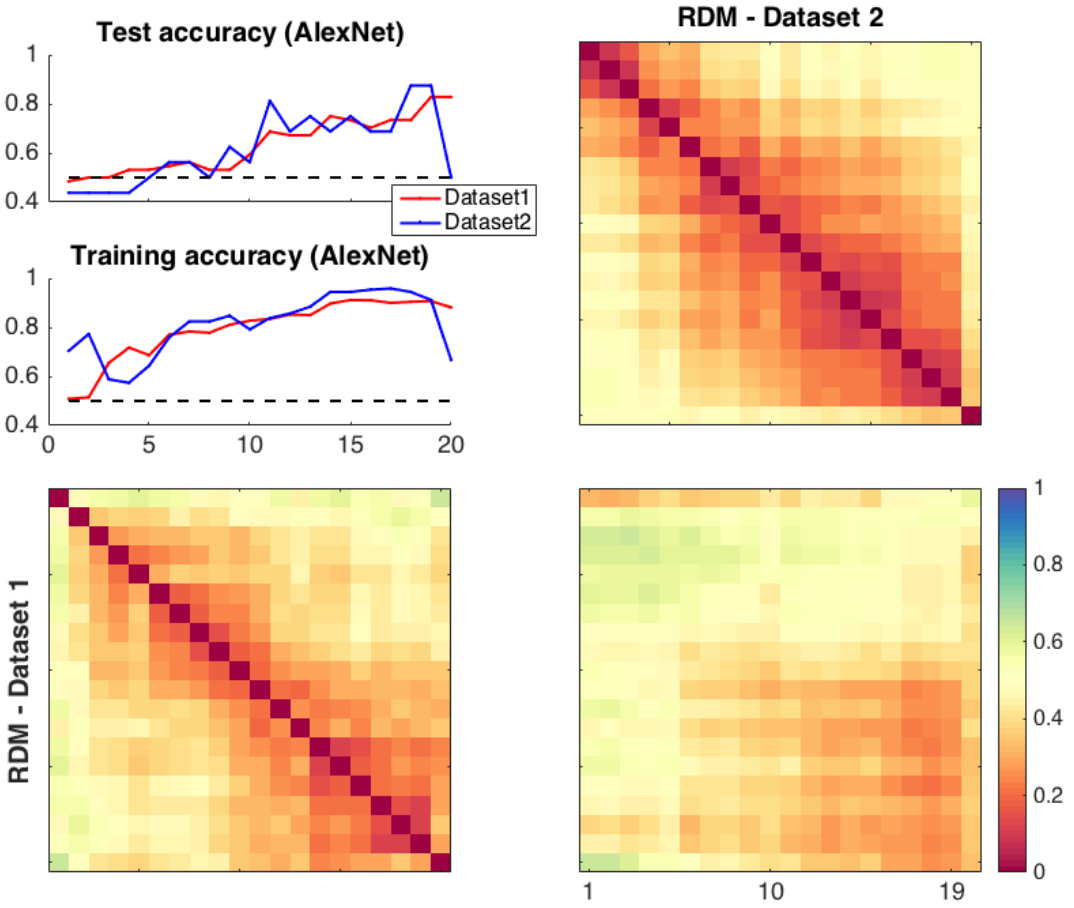


**Figure B.3:** Comparison of the properties of the SVMs trained on AlexNet layers with Datasets 1&2. The description of the generation of the plots is mentioned in the caption of Figure 2.5. AlexNet shows a smoother increase in performance (top, left) as opposed to VGG-16. The similarity between decision scores in the final layers from SVMs trained on the two datasets is lower than in VGG-16.

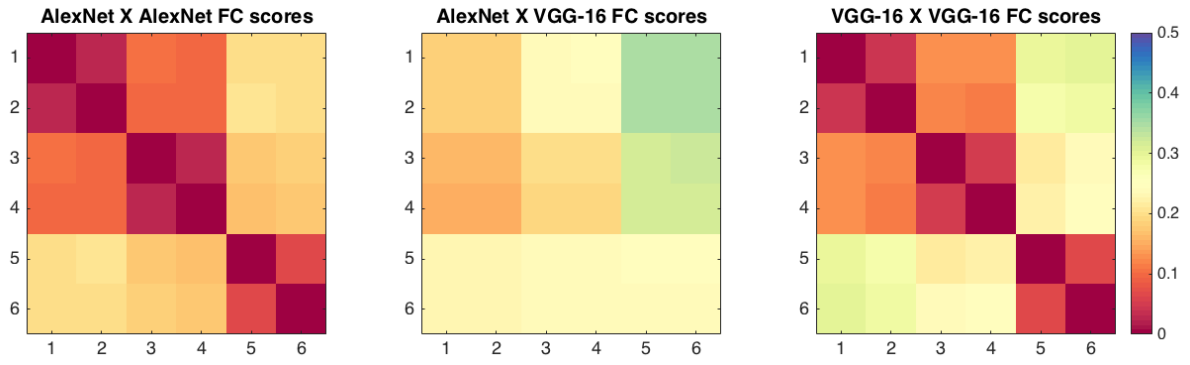


**Figure B.4:** In addition to Figure 2.6, the similarity between the decision scores in AlexNet and VGG-16 fully-connected layers. RDMs for similarities within and between the CNN animacy decision scores are shown. The Kendall's $\tau$ correlation between the within-CNN RDMs is 0.79, between within-AlexNet and between-CNNs RDMs is 0.09, and between within-VGG-16 and between-CNNs RDMs is 0.18. So, there is weak correspondence between the decision scores of the fully-connected layers of the two CNNs. This might be because the transformations of the representations aren't aligned in the same order in both CNNs.

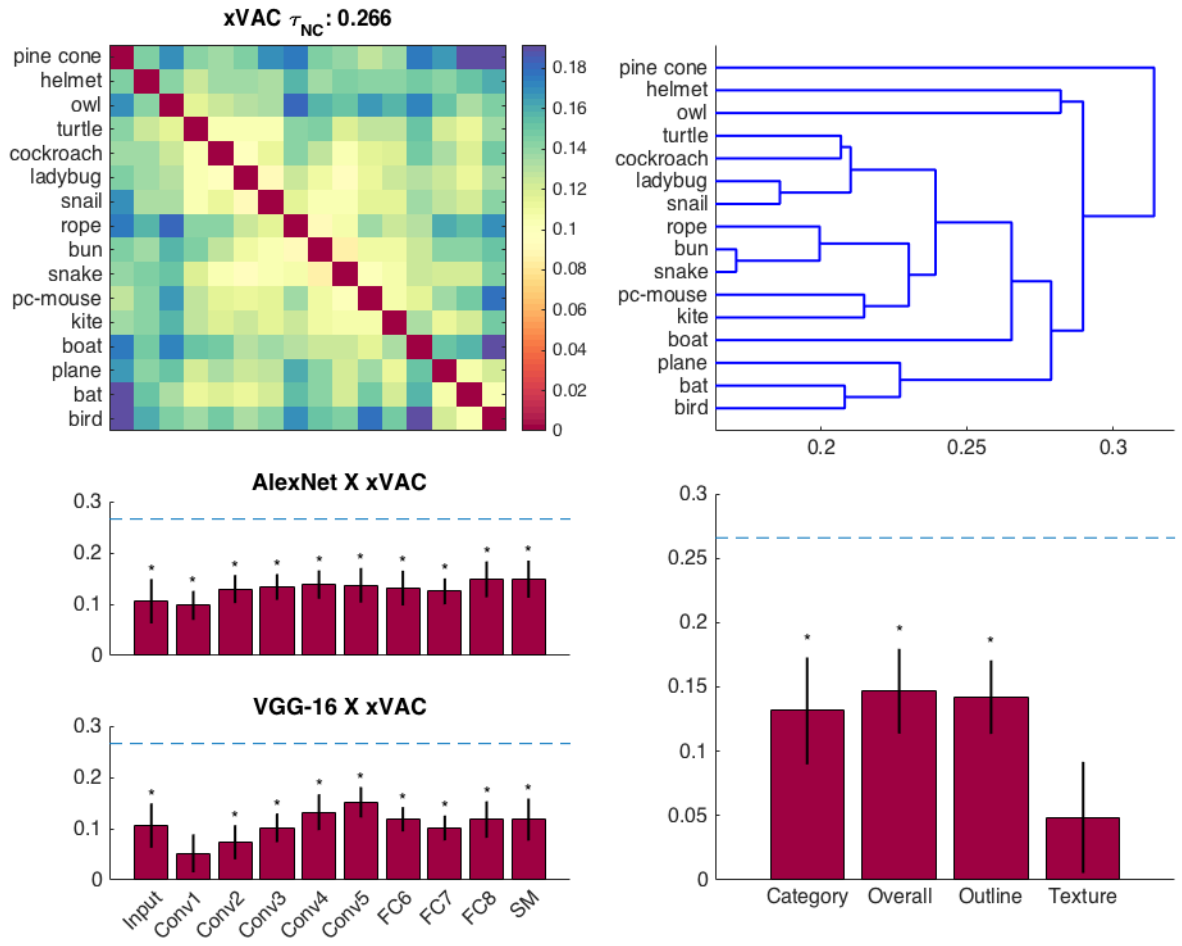


**Figure B.5:** Comparing the representational similarities of xVAC, CNNs and the behavioural measures. The description is similar to that in Figure 3.2a, and similar trends are observed.

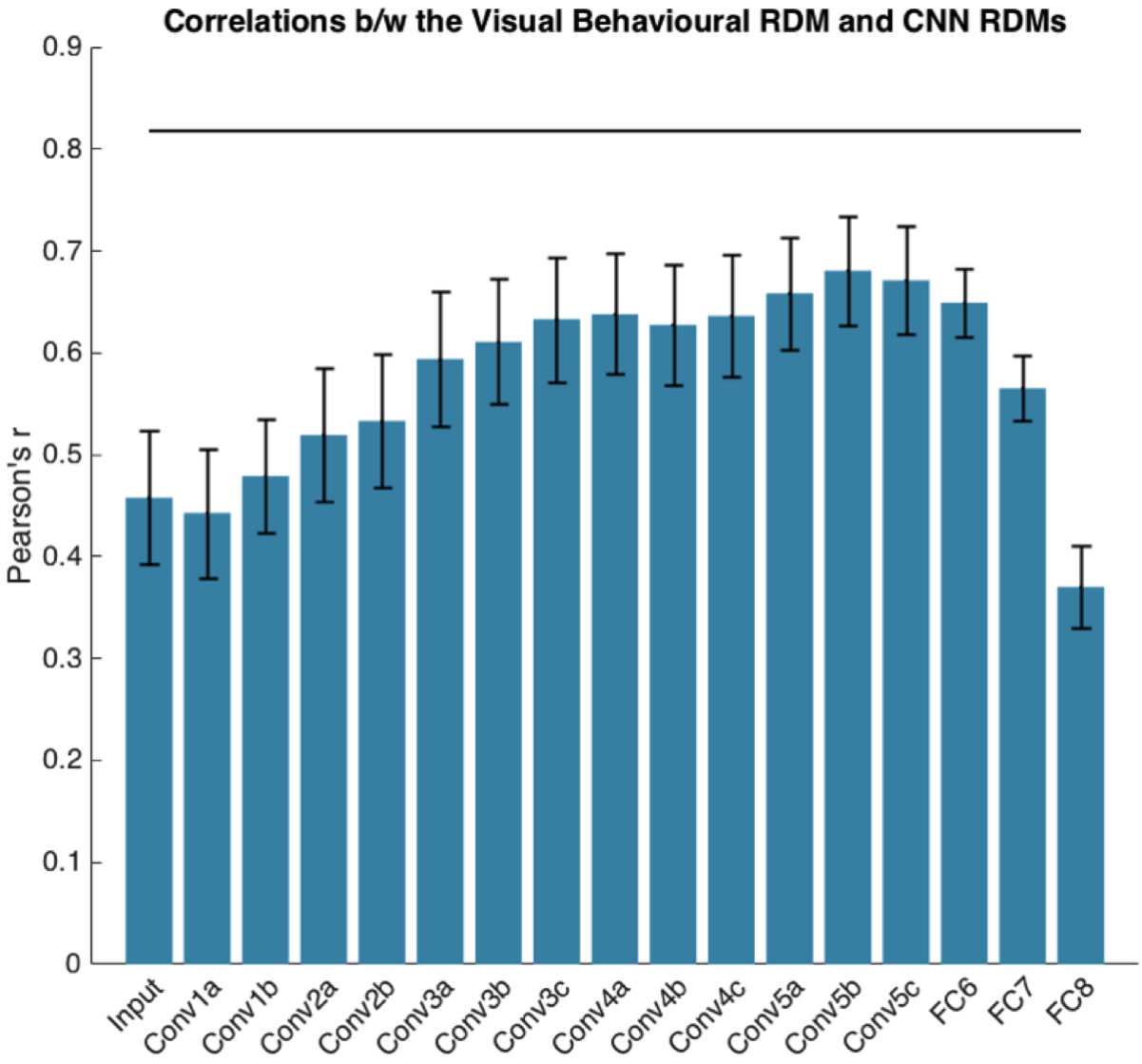


**Figure B.6:** Similarity of the representational structures in VGG-16 and the overall visual dissimilarities. The similarity peaks in the middle layers, but stays significant throughout the CNN.

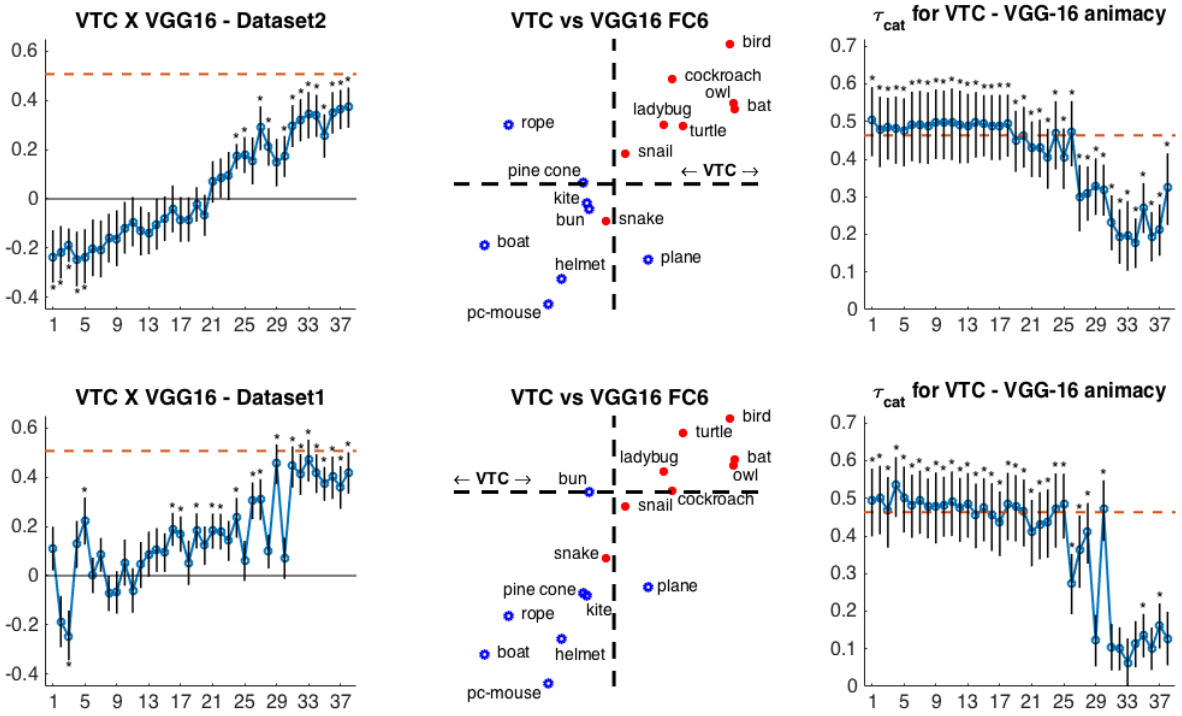


**Figure B.7:** Comparing the animacy organisation of VTC with the animacy organisations of VGG-16. The description of the methods is the same as in Figure 3.3.

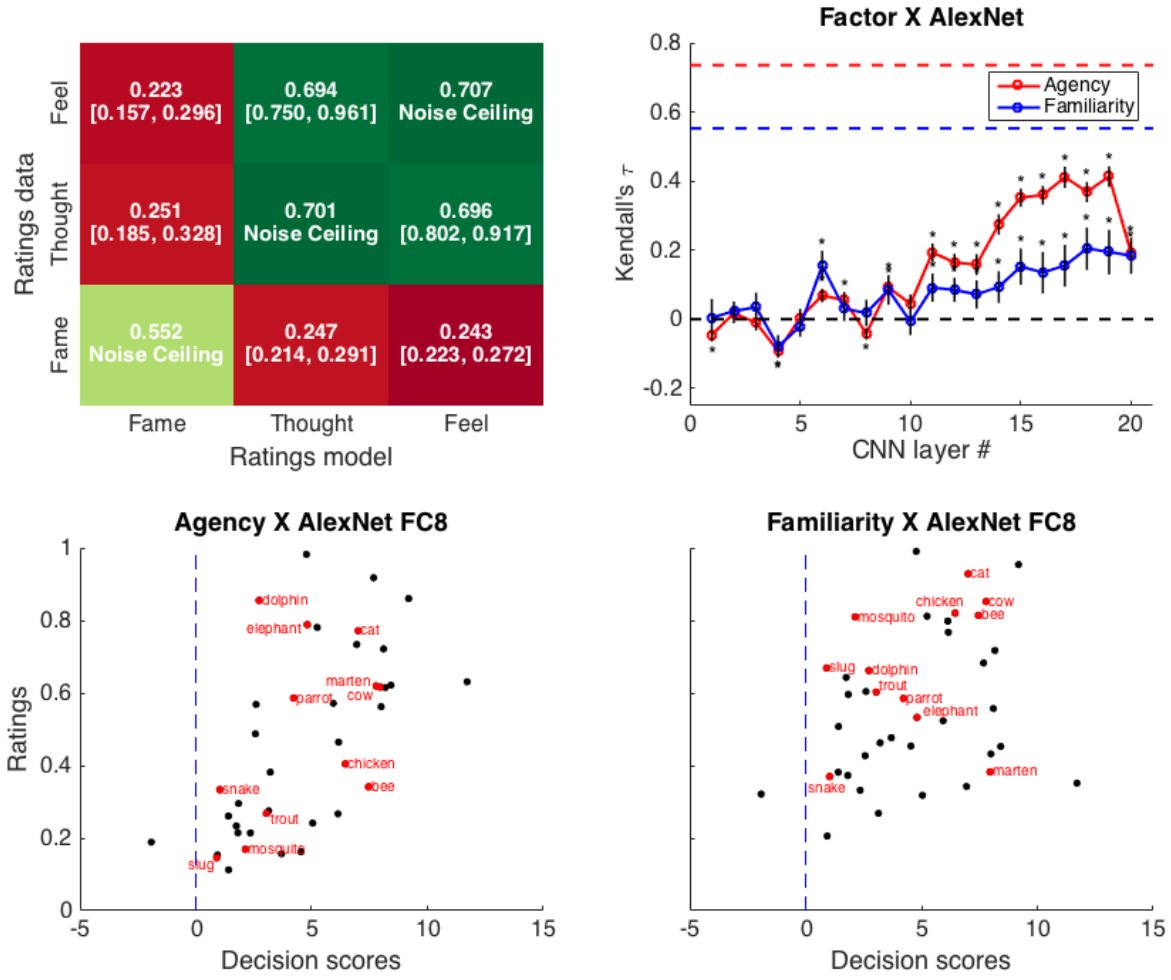


**Figure B.8:** Similarities between the non-visual factors and AlexNet decision scores. The methods are similar to those in Figure 4.1. The average correlations (Kendall's $\tau$) between the individual agency and familiarity ratings and the FC8 animacy coefficients, after stimuli selection, are as follows - Agency x FC8 - 0.30, Familiarity x FC8 - 0.11, Agency x Familiarity - 0.03. This surely is a cause for concern as the animacy organisation based on some visual features is not dissociated with the non-visual factors, and will be explored further.


**Figure B.9:** Animal images used in the behavioural ratings experiment.


**Figure B.10:** Images used in the main fMRI experiment.

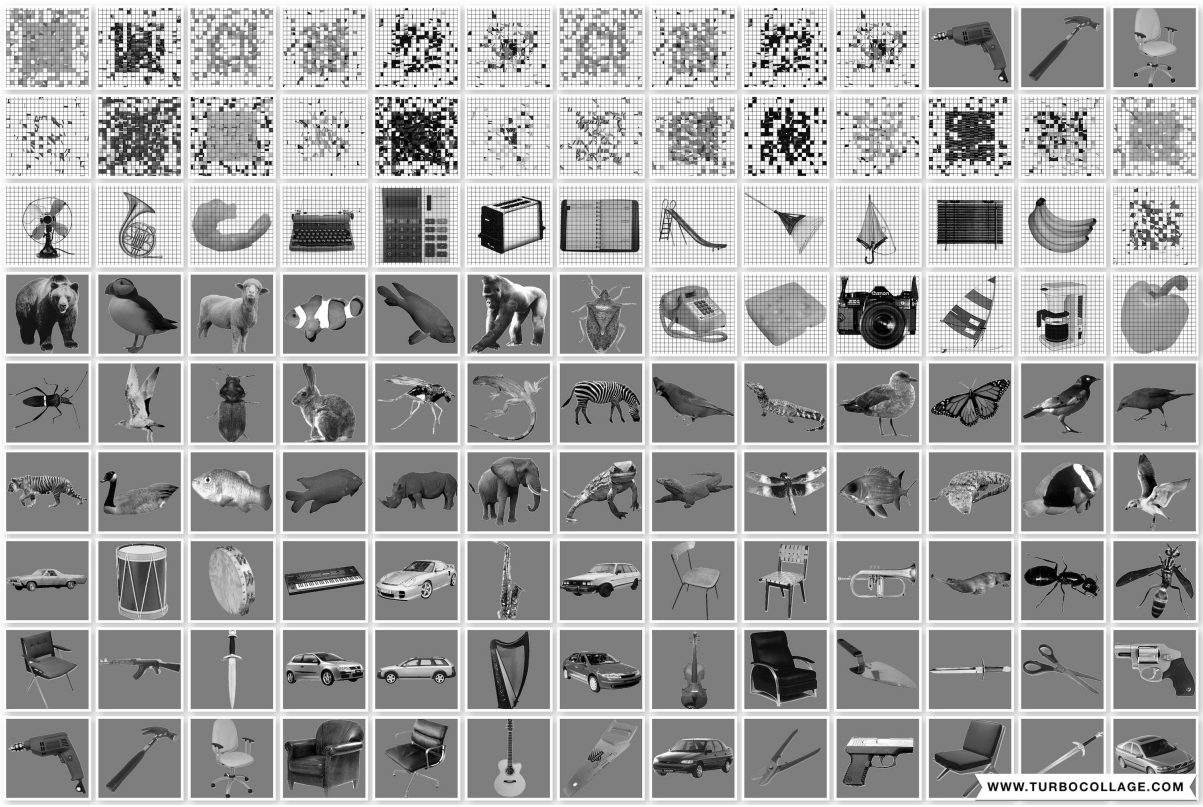


**Figure B.11:** Images used in the functional localiser experiment.

# Bibliography


A. Andreopoulos and J. K. Tsotsos. 50 years of object recognition: Directions forward. *Computer Vision and Image Understanding*, 117(8):827–891, 2013.

A. H. Bell, L. Pessoa, R. B.H. Tootell, and L. G. Ungerleider. Chapter 44 - visual perception of objects. In L. R. Squire, D. Berg, F. E. Bloom, S. du Lac, A. Ghosh, and N. C. Spitzer, editors, *Fundamental Neuroscience (Fourth Edition)*, pages 947 – 968. Academic Press, San Diego, fourth edition edition, 2013.

D. H. Brainard and S. Vision. The psychophysics toolbox. *Spatial vision*, 10:433–436, 1997.

C. F. Cadieu, H. Hong, D. LK. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12):e1003963, 2014.

A. Canziani, A. Paszke, and E. Culurciello. An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*, 2016.

R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6:27755, 2016.

A. C. Connolly, J. S. Guntupalli, J. Gors, M. Hanke, Y. O. Halchenko, YC. Wu, H. Abdi, and J. V. Haxby. The representation of biological classes in the human brain. *Journal of Neuroscience*, 32(8):2608–2618, 2012.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

B. Deen, H. Richardson, D. D. Dilks, A. Takahashi, B. Keil, L. L. Wald, N. Kanwisher, and R. Saxe. Organization of high-level visual cortex in human infants. *Nature communications*, 8:13995, 2017.

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

A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

K. Grill-Spector and K. S. Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548, 2014.

K. Grill-Spector, Z. Kourtzi, and N. Kanwisher. The lateral occipital complex and its role in object recognition. *Vision research*, 41(10):1409–1422, 2001.

M. Harrower and C. A. Brewer. Colorbrewer. org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.

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

J. V. Haxby, J. S. Guntupalli, A. C. Connolly, Y. O. Halchenko, B. R. Conroy, M. I. Gobbini, M. Hanke, and P. J. Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011.

C. M. Kadipasaoglu, C. R. Conner, M. L. Whaley, V. G. Baboyan, and N. Tandon. Category-selectivity in human visual cortex follows cortical topology: A grouped icEEG study. *PloS one*, 11(6):e0157109, 2016.

A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

SM. Khaligh-Razavi and N. Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

R. Kiani, H. Esteky, K. Mirpour, and K. Tanaka. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of neurophysiology*, 97 (6):4296–4309, 2007.

N. Kriegeskorte, M. Mur, and P. Bandettini. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2, 2008a.

N. Kriegeskorte, M. Mur, D. A. Ruff, R. Kiani, J. Bodurka, H. Esteky, K. Tanaka, and P. A. Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, 2008b.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

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

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

A. Martin and J. Weisberg. Neural foundations for understanding social and mechanical concepts. *Cognitive Neuropsychology*, 20(3-6):575–587, 2003.

K. Mohan and SP. Arun. Similarity relations in visual search predict rapid visual categorization. *Journal of vision*, 12(11):19–19, 2012.

H. Nili, C. Wingfield, A. Walther, Li. Su, W. Marslen-Wilson, and N. Kriegeskorte. A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553, 2014.

M. V. Peelen and P. E. Downing. Category selectivity in human visual cortex: Beyond visual object recognition. *Neuropsychologia*, 2017.

D. Proklova, D. Kaiser, and M. V. Peelen. Disentangling representations of object shape and object category in human visual cortex: The animate–inanimate distinction. *Journal of cognitive neuroscience*, 2016.

M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.

S. Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

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

T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the national academy of sciences*, 104(15):6424–6429, 2007.

L. Sha, J. V. Haxby, H. Abdi, J. S. Guntupalli, N. N. Oosterhof, Y. O. Halchenko, and A. C. Connolly. The animacy continuum in the human ventral vision pathway. *Journal of cognitive neuroscience*, 2015.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

J. van den Hurk, M. Van Baelen, and H. P. Op de Beeck. Development of visual category selectivity in ventral visual cortex does not require visual experience. *Proceedings of the National Academy of Sciences*, page 201612862, 2017.

A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. In *Proceeding of the ACM Int. Conf. on Multimedia*, 2015.

X. Wang, M. V. Peelen, Z. Han, C. He, A. Caramazza, and Y. Bi. How visual is the visual cortex? comparing connectional and functional fingerprints between congenitally blind and sighted individuals. *Journal of Neuroscience*, 35(36):12545–12559, 2015.

D. LK. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.