Aligning computer and human visual representations

Ramakrishnan, K.

Publication date
2017

Document Version
Final published version

License
Other

Link to publication

Citation for published version (APA):
Ramakrishnan, K. (2017). *Aligning computer and human visual representations*.

Aligning Computer and Human Visual Representations

Kandan Ramakrishnan

Aligning Computer and Human Visual Representations

Kandan Ramakrishnan

Aligning computer and human visual representations

Kandan Ramakrishnan

This book was typeset by the author using latex.

Printing: Off Page, Amsterdam

ISBN 978-94-6182-836-1

Aligning computer and human visual representations

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. K. I. J. Maex
ten overstaan van een door het college voor promoties
ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op woensdag 1 november 2017, te 12:00 uur

door

Kandan Ramakrishnan

geboren te Vellore Tamilnadu, India

Promotiecommissie

Promotor:	Prof. dr. ir. A. W. M. Smeulders	Universiteit van Amsterdam
Co-promotor:	dr. S. Ghebreab	Universiteit van Amsterdam
	dr. H. S. Scholte	Universiteit van Amsterdam
Overige leden:	Prof. dr. V. A. F. Lamme	Universiteit van Amsterdam
	Prof. dr. P. R. Roelfsema	Vrije Universiteit Amsterdam
	Prof. dr. T. Gevers	Universiteit van Amsterdam
	dr. M. A. J. van Gerven	Radboud Universiteit Nijmegen
	dr. E. Gavves	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

COMMIT/

The work described in this thesis has been carried out with support of COMMIT at the Intelligent Systems Lab Amsterdam of the University of Amsterdam.

UNIVERSITEIT VAN AMSTERDAM

CONTENTS

1

INTRODUCTION

Visual information is an important and rich form of sensory input, critical for many tasks: walking on the street we avoid bumping into other people, recognize street signs and at the same time navigate to our destination. While seemingly simple, processing visual input to derive meaning from our surroundings is a challenging feat. This is so because, the visual input can undergo changes due to variation in lighting, size, pose, occlusion from other objects or due to the position of the observer itself. How do the lights rays transform into a visual representation of the world ?

A **visual representation** can be defined as the transformation of visual data (the image) to accomplish tasks easily [1]. Visual representations therefore allow for computations efficiently, translating the input into a semantically meaningful concept.

Let's look at how a task such as object recognition can be tackled. The objective of the task is to assign a image of an object to its respective label [2]. The first step is to represent the task "recognition" as a set of conceptual categories each of which is associated with a set of images sharing similar properties which allow for description by a word or expression (eg., cat/dog/orange/apple). The visual representation of the category include a range of properties or feature values such as edges or color. A naive property or feature of the category apple is its color (red) and shape (circular). Thus, representations for object recognition entails deconstruction of an image to features, which when combined describe the visual aspects of the object category.

In computer vision, different visual representation have been proposed for tasks such as object recognition [3], detection [4] and tracking [5]. To recognize a category, requires a complete description of a object relative to other categories. For object tracking, a differential representation of the object relative to the surroundings is sufficient. While for detection, identification of object boundary is required. Recent progress in computer vision show that representations for different tasks are converging, and that generic representations can be learnt for different tasks [6].

In neuroscience, visual representations have been extensively studied within the primate brain [7], [8], [9]. However, for humans as the task is not determined a priori, complete representations for all possible tasks that can be solved are present.

Both computer vision and human visual system target the same goal: to accomplish visual tasks easily via a set of representations. Therefore, a natural question that arises is:

Do computer vision representations align with human visual representations ?

This thesis is at the intersection of computer vision and human vision. In both computer and human vision, representations are not a single but consist of multiple computational stages or levels.

1.1 LEVELS OF REPRESENTATION IN COMPUTER VISION

A first step in the representation of an image is to compute local changes in intensity (gradient or edges). A number of low-level features have been defined such as Gabor filters, steerable invariants [10], local binary patterns [11] and moment invariants [12], all of which compute local changes in intensity. These models have proved to be quite successful in data-specific domains such as face recognition, vehicles or texture recognition [13], [14]. However for generic object recognition, local information that is invariant to changes in illumination changes [15], scale, orientation and occlusion is required.

A second, intermediate step of computation is a histogram of the gradients over local image patches. One such descriptor is the Scale Invariant Feature Transform (SIFT) [16]. The main advantages of the framework is its simplicity, efficiency and invariance under viewpoint changes and background clutter, which typically results in state of the art performance on object recognition. In the Bag of words framework, SIFT descriptor is further coded into visual words over the entire image to form a compact representation [3].

A third step is to map the final representation to the high-level concept using a supervised learning algorithm for classification such as the Support Vector Machine (SVM) [17]. This enables us to determine the category a new image belongs, on the basis of a training set of data containing images (or instances) whose category membership is known.

Modern machine learning algorithms have integrated the above three steps in a single learning framework [18]. In deep learning, computer vision systems have found it most efficient to compute features of an image in a hierarchical fashion [19]. The fundamental building block in a deep neural network involves convolving the image with filters followed by normalization, pooling and point-wise non-linear operations [20]. This building block is repeated a number of times, that gives rise to features of increasing complexity starting from low-level edge features to object-like features [21]. The use of deep neural networks has resulted in significantly improved object recognition performance [20] reaching human-level performance [22].

Deep neural networks are trained on millions of images [23] and learn representations directly from images [19]. During the training phase, the network parameters are learnt via feedback or back-propogation [24], [25] of the error with respect to the true output. At runtime, the network is purely feed-forward and a compact representation of the input image is computed.

The hierarchical levels of representations is the dominant paradigm in computer vision which enables us, as discussed in the next section, to align different models to levels of representation in human vision.

1.2 LEVELS OF REPRESENTATION IN HUMAN VISION

The human brain performs visual tasks by layers of neurons organized into a hierarchy, and representing increasingly complex visual features [26]. After hitting the eye, where initial processing already occurs, the thalamus relays visual signals to the visual cortex for further processing. The dominant hypothesis of representation in the visual cortex is a hierarchical organization of information consisting of two parallel streams [27], [28] as shown in Figure 1.

The dorsal stream is mostly involved with processing of space, movement and action [29]. The computational machinery that represents scenes and objects is located in the ventral visual stream of the human brain [30]. Anatomically, the ventral stream is composed of a set of cortical areas, each thought to convey a distinct representation of the visual input [31]. In the ventral stream, areas of the temporal lobe are thought to be at the top of the visual cortical hierarchy.

Visual information travels from the retina to the lateral geniculate nucleus of the thalamus (LGN), and then through successive areas in the primary visual cortex: V1 and V2 [32]. Area V1 has simple and complex neurons similar to low-level features such as bank of Gabor filters [33]. Simple cells are sensitive to oriented stimuli (e.g, bars, edges, gratings) at one particular orientation, location and phase, while complex cells, which have a larger receptive size, also

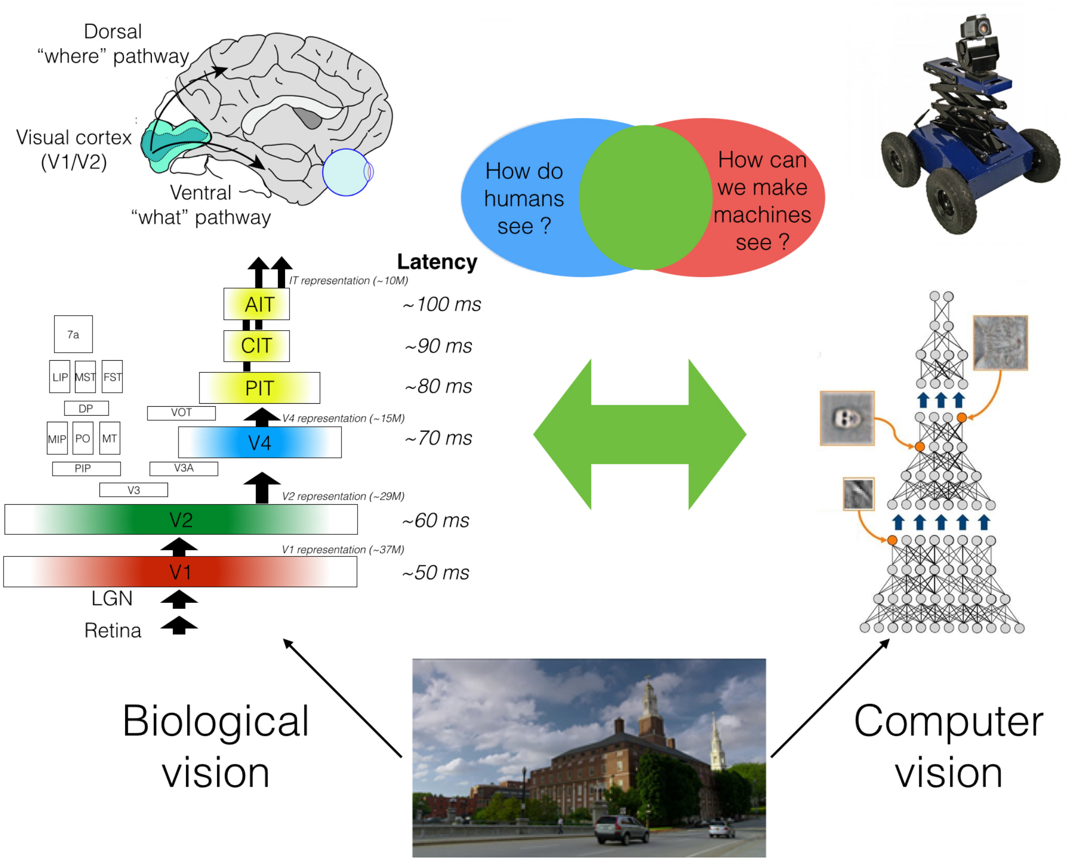


Figure 1: Visual information representation in human and computer vision systems. The research in this thesis is at the intersection of computer and neuroscience, specifically using different models of visual representation to understand the computational mechanisms in the human visual system.

exhibit some invariance to location of the stimulus within their receptive fields. Area V2 has neurons similar to area V1 but with larger receptive field sizes [34].

Beyond the primary visual cortex, many visual areas distributed across the ventral stream are specialised for particular kinds of hierarchical visual processing. Area V4 processes features of intermediate complexity [35], middle temporal visual area (MT) is sensitive to motion [36] and the lateral occipital cortex (LOC) area is an object-selective region that responds to objects defined by luminance [37], texture [38], motion [39], or stereo cues [40], [41].

Just about 100 ms after photons strike the retina, area IT at the top of the ventral stream produces a pattern of brain activity that can directly support robust, real-time visual object categorization and identification [42]. Face detection and identification have for example, been shown to occur in inferior temporal cortex (IT). Brain responses evoked in IT [35] are most likely produced by a combination of intra-area processing and feed-forward inter-area processing of the visual input [43].

The human visual system thus process visual input via a cascade of hierarchical and largely feedforward computations that culminate in an invariant representations of visual input in the inferior temporal cortex [43]. This cascade of computations has partly inspired and resembles the hierarchy of representations in current state-of-the-art deep neural network models.

1.3 COMPARING COMPUTER AND HUMAN VISUAL REPRESENTATIONS

The computational machinery in both human visual system and computer vision solves highly challenging visual tasks. When light hits the eye, within hundreds of milliseconds, the brain has decided on a number of things: what the context is, what categories of things are present in the image, and what concepts characterize the scene. When light hits a camera, computer vision is now able to understand the semantics of up to 10,000 objects and categories.

Hitherto, the scientific disciplines of computer and human vision have largely grown in separation. Computer vision focussed on engineered solutions to solve a specific visual task, while the focus in neuroscience is to understand the hierarchy of human visual representations. Given the recent progress in computer vision coupled with the availability of neuro-imaging (EEG and fMRI) data that enables us to analyze brain responses to natural images, its a highly opportune moment to compare computer and human visual representations [44]. A number of studies in recent years have indeed compared different computer vision models to brain responses [45], [46], [47].

These studies can be grouped along different dimensions; comparing different models, using different stimulus modalities and even different brain data modalities. In one such study, [45] compared five different computational models of visual representation against human brain activity to understand intermediate representations in the human brain. In [47], deep neural networks are compared against a wider set of computational models including shallow models. They show that deep neural networks is the best performing model. Another study [48] showed that deep models achieves a high correlation to the brain responses in area IT. Using natural movie stimulus, [46] show that deep neural network representations correlate to dorsal stream brain responses. While all the above studies use fMRI responses to study the spatial hierarchy, [49] show the correlation of deep neural networks to temporal hierarchy of visual processing in the human brain using MEG data. While these studies demonstrate important findings, a number of open questions remain.

In this thesis we take a comprehensive approach to study the relation between computer and human visual representations, putting to the test different media, models and modalities. First, a large number of static images and dynamic videos will be used to extract visual representations and to evoke brain responses. While previous studies have focussed on either controlled stimuli or limited variations in the stimuli set, we use natural scenes that has a large variation to capture the complexity of visual information. Second, we compare shallow and deep models of hierarchical visual representations. Thirdly, we use both EEG and fMRI modalities to capture various aspects of brain responses to visual input from a large number of subjects. Furthermore to demonstrate the correspondence between hierarchical computer vision models and neural representations we use advanced data analysis approaches capable of disentangling levels of visual representations and computations. The different research questions addressed in our thesis are shown in Figure 2.

1.4 RESEARCH QUESTIONS

We describe the different research questions addressed in this thesis. We first align computer vision models with two computational layers to the human brain.

In computer vision (BoW [3]) and neuroscience (HMAX [50]), models with 2 computational stages have been developed independently to achieve much progress in their respective fields. While there are differences between the BoW and HMAX models, they combine low-level edge information via an intermediate aggregation step to form a representation that is assumed to identify the category of the image. While the models have been independently compared to the

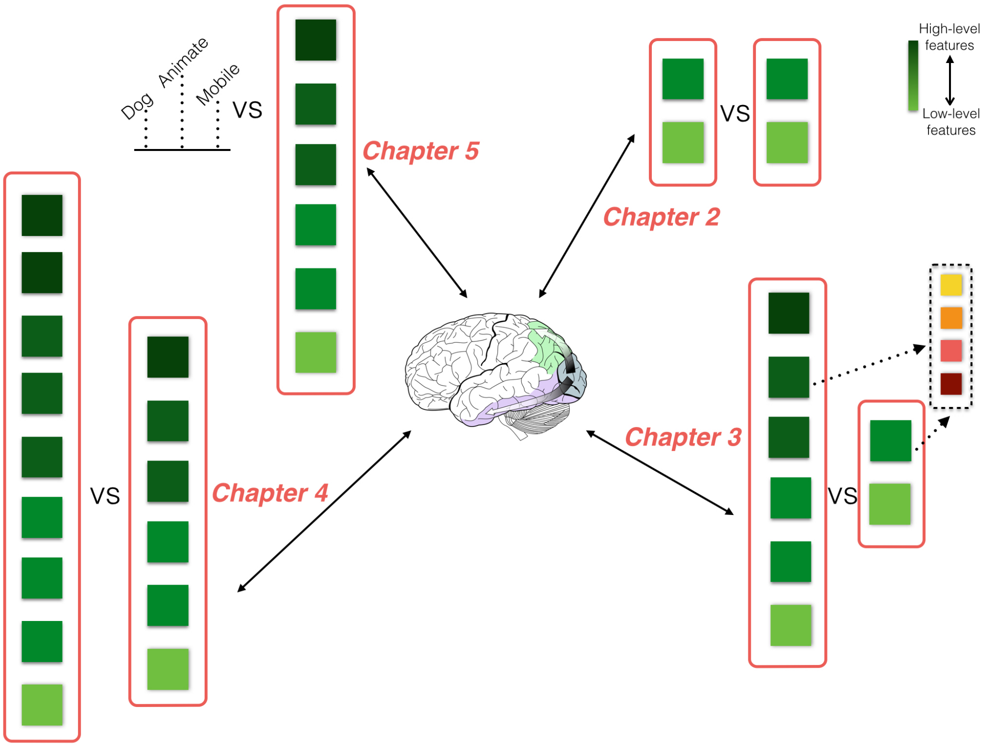


Figure 2: The models compared to the brain per chapter are visualized : Chapter 2) BoW and HMAX with two computational stages are used to correlate to the human brain. Chapter 3) Deep neural network (DNN) of 7 layers is correlated to the brain in comparison to BoW and HMAX. Chapter 4) A DNN of 15 layers is compared to a DNN of 7 layers. Chapter 5) DNN of 7 layers and semantic representations of an image are used in the correlation to the brain.

brain, the different layers in the two models have not compared against each other. This leads us to our first research question,:

Do BoW and HMAX representations correlate to visual representations in the brain ?

In Chapter 2, we align different computational representations of BoW and HMAX to human brain responses. We address two issues in this chapter: how to reliably compare high dimensional data from different data domains [51] and to understand what is the unique component of the explained variance of brain responses by these models [52]. By comparing the different levels of hierarchy we investigate to what extent do these models capture intermediate visual representations in the human brain. The nature of intermediate representations as revealed by these models are distinctive parts of the image that help to understand the categorical nature of the image, for example a patch of the dogs face is an intermediate feature to identify the image as a dog.

BoW and HMAX being shallow models, are limited in nature, and are outperformed by deep neural networks in automated object recognition. Deep neural networks have a number of computational layers composed of low, intermediate and high level visual features. The hierarchy of representations in a deep neural network resembles the hierarchy of human visual representations. This leads us to the next question in Chapter 3;

How do deep neural network compare against HMAX and BoW in its correlation to visual representations in the brain ?

We analyze how deep neural networks correspond to the brain and compare them with different computational models in Chapter 3. While recent studies show the correlation of CNN layers to different brain areas, we go beyond prior work by accounting for the correlation between the CNN layers in the correspondence to brain responses. Further, deep neural networks are not explicitly tested against other computer vision models. We relate models such as HMAX, BoW and deep neural networks, each of which are composed of different combinations of four basic operations. Each model consists of a combination of i) filter-bank convolution, ii) non-linearity, iii) pooling and iv) normalization. From our results, we show deep neural networks best correlate to brain responses of natural stimuli after accounting for the correlation between layers. HMAX and BoW models correspond to intermediate layers of the neural network. We further demonstrate in the study, that hierarchy and non-linearity are critical in vision models to explain brain responses.

In computer vision the number of layers in a deep model is critical to its performance on object recognition, with increasing layers achieving state of the art performance. However it is not clear yet, if increasing number of computational layers in a deep neural network better model the brain. This leads to the following question;

Does increase in deep neural network layers have better correspondence to the hierarchy of visual representations in the brain ?

In most of previous studies, deep neural models uncover the spatial nature of visual representations in the brain using fMRI data. However the correspondence of deep neural models to temporal hierarchy of visual processing as revealed by time resolved E/MEG brain responses remains to be seen. In Chapter 4 we investigate; 1) the temporal hierarchy in the correspondence of deep neural network representations to EEG responses and 2) the role of depth in correspondence of computational models to the brain. Our results show that the temporal dynamics of visual processing are indeed reflected by deep learning models, and low level features are processed early in time and increasingly complex features are processed later in time. Further, we demonstrate that deeper models with additional layers of computation correspond to the brain better.

In the above three Chapters, we show how hierarchical visual representations in the brain can be attributed to computer vision models that contain features at different levels of abstraction. However the brain also represents visual information semantically along dimensions such as animal or non-animal, dark and bright. While visual and semantic representations have been mapped to brain responses, they have not compared against each other. This leads to our final question;

Do deep neural network or semantic representations correlate to representations in the brain ?

In Chapter 5, we compare deep neural network and semantic representations of stimuli to map brain responses. We specifically address to what extent brain responses are similar across subjects during natural vision. This study, done on a hundred subjects, compares visual and semantic representations to the similarity in brain activation patterns. Patterns of brain activity that are similar across subjects is obtained using principal component analysis. We check if deep neural network and semantic correlation are not only consistent over a group of subjects but also to what extent are correlated to each other.

This thesis is about the alignment of computer vision models to human visual system at different levels of hierarchy. From our study of both human vision (visual representation optimized by evolution) and computer vision (visual representation optimized for artificial vision), we hope to understand the critical elements in the design of visual systems for object and scene recognition.

1.5 ARTICLES AND CONTRIBUTIONS

Chapter 2 is based on ”Visual dictionaries as intermediate features in the human brain”, published in Frontiers in Computational Neuroscience, 2015, by Kandan Ramakrishnan, H. Steven Scholte, Iris I. A. Groen, Arnold W. M. Smeulders, Sennay Ghebreab.

Kandan Ramakrishnan : All aspects of the research
H. Steven Scholte : Discussion
Iris I. A. Groen : Discussion
Arnold W. M. Smeulders : Insights and supervision
Sennay Ghebreab : Insights and supervision

Chapter 3 is based on ”Neural spatial consistency of hierarchical vision models”, published at Neural Information Processing Systems (NIPS), MLINI workshop 2015, by Kandan Ramakrishnan, H. Steven Scholte, Arnold W. M. Smeulders, Sennay Ghebreab.

Kandan Ramakrishnan : All aspects of the research
H. Steven Scholte : Discussion
Arnold W. M. Smeulders : Discussion
Sennay Ghebreab : Supervision

Chapter 4 is based on ”Comparing deep neural networks to understand temporal dynamics of object recognition”, under review at Journal of Vision, 2017, by Kandan Ramakrishnan, Iris I. A. Groen, H. Steven Scholte, Arnold W. M. Smeulders, Sennay Ghebreab.

Kandan Ramakrishnan : All aspects of the research
Iris I. A. Groen : Data collection and writing insight
H. Steven Scholte : Insights and supervision
Arnold W. M. Smeulders : Insights and supervision
Sennay Ghebreab : Insights and supervision

Chapter 5 is based on ”Mapping deep neural and semantic representations to across subject similarity in brain responses”, under review at PLOS One, 2017, by Kandan Ramakrishnan, H. Steven Scholte, Arnold W. M. Smeulders, Sennay Ghebreab.

Kandan Ramakrishnan : All aspects of the research
H. Steven Scholte : Insights and supervision
Arnold W. M. Smeulders : Insights and supervision
Sennay Ghebreab : Insights and supervision

2

VISUAL DICTIONARIES AS INTERMEDIATE FEATURES IN THE HUMAN BRAIN

SUMMARY :

The human visual system is assumed to transform low level visual features to object and scene representations via features of intermediate complexity. How the brain computationally represents intermediate features is still unclear. To further elucidate this, we compared the biologically plausible HMAX model and Bag of Words (BoW) model from computer vision against human brain responses. Both these computational models use visual dictionaries, candidate features of intermediate complexity, to represent visual scenes, and the models have been proven effective in automatic object and scene recognition. These models however differ in the computation of visual dictionaries and pooling techniques. We investigated where in the brain and to what extent human fMRI responses to short video can be accounted for by multiple hierarchical levels of the HMAX and BoW models. Brain activity of 20 subjects obtained while viewing a short video clip was analyzed voxel-wise using a distance-based variation partitioning method. Results revealed that both HMAX and BoW explain a significant amount of brain activity in early visual regions V1, V2 and V3. However BoW exhibits more consistency across subjects in accounting for brain activity compared to HMAX. Furthermore, visual dictionary representations by HMAX and BoW explain significantly some brain activity in higher areas which are believed to process intermediate features. Overall our results indicate that, although both HMAX and BoW account for activity in the human visual system, the BoW seems to more faithfully represent neural responses in low and intermediate level visual areas of the brain.

2.1 INTRODUCTION

The human visual system transforms low-level features in the visual input into high-level concepts such as objects and scene categories [53]. Visual recognition has been typically viewed as a bottom-up hierarchy [54] in which information is processed sequentially with increasing complexities, where lower-level cortical processors, such as the primary visual cortex, are at the bottom of the processing hierarchy and higher-level cortical processors, such as the inferotemporal cortex (IT), are at the top, where recognition is facilitated [55]. Much is known about the computation in the earliest processing stages, which involve the retina [56], [57], lateral geniculate nucleus (LGN) [58] and primary visual cortex (V1) [59]. These areas extract simple local features such as blobs, oriented lines, edges and colour from the visual input. However there remain many questions on how such low-level features are transformed into high-level object and scene percepts [60].

One possibility is that the human visual system transforms low-level features into object and scene representations via an intermediate computational step [50]. After extraction of low-level features in areas such as V1, moderately complex features are extract from areas V4 and the adjacent region temporal occipital (TO). Then partial or complete object views are represented in anterior regions of inferotemporal (IT) cortex [61]. It has been suggested that such intermediate features along the ventral visual pathway are important for object and scene representation [62].

Previous studies have provided some evidence of what intermediate features might entail. In [63] it has been shown that cells in the V4/IT region respond selectively to complex features such as simple patterns and shapes. Similarly, [64] identified contour selectivity for individual neurons in the primate visual cortex and found that most contour-selective neurons in V4 and IT encoded some subset of the parameter space and that a small collection of the contour-selective units were sufficient to capture the overall appearance of an object. Together these findings suggest that intermediate features capture object related information encoded within the human ventral pathway.

A number of recent studies have explored the computational basis for intermediate features in the human brain. In an attempt to answer the question of intermediate features underlying neural object representation, [65] compared five different computational models of visual representation against human brain activity to object stimuli. They found that the Bag of Words (BoW) model was most strongly correlated with brain activity associated with midlevel perception. These results were based on fMRI data from 5 subjects. Recently [66] used a wider set of models including HMAX and BoW against neural responses from two monkeys in IT and V4.

In this chapter, we test HMAX and BoW as computational models of intermediate features in the human brain. HMAX [67] and BoW [68] models represent scenes in a hierarchical manner transforming low level features to high level concepts. HMAX is a model for the initial feedforward stage of object recognition in the ventral visual pathway. It extends the idea of simple cells (detecting oriented edges) and complex cells (detecting oriented edges with spatial invariance) by forming a hierarchy in which alternate template matching and max pooling operations progressively build up feature selectivity and invariance to position and scale. HMAX is thus a simple and elegant model used by many neuroscientists to describe feedforward visual processing. In computer vision, different algorithms are used for object and scene representation. The commonly used model in computer vision is BoW which performs very well on large TRECvid [69] and PASCAL [70] datasets, in some cases even approaching human classification performance [71]. The key idea behind this model is to quantize local Scale Invariant Feature Transform (SIFT) features [72] into visual words [73], features of intermediate complexity, and then to represent an image by a histogram of visual words. To further understand the nature of intermediate features underlying scene perception, we test these two computational models against human brain activity while subjects view a movie of natural scenes.

Although HMAX and BoW are different models they both rely on the concept of visual dictionaries to represent scenes. In HMAX after the initial convolution and pooling stage, template patches are learnt from responses of the pooling layer (from a dataset of images) which are used as visual dictionaries. In the BoW model, clustering of SIFT features forms the visual dictionary. In both the models, visual dictionaries are medium size image patches that are informative and at the same time distinctive. They can be thought of, as features of intermediate complexity. This comparison of different computational approaches to visual dictionaries might provide further insight about the representation of intermediate features in the human brain.

To test the two layers of HMAX and BoW, we show 20 subjects a 11-minute video of dynamic natural scenes and record their fMRI activity while watching the video. We use dynamic scenes instead of static scenes because they are more realistic, and because they may evoke brain responses that allow for a better acquisition of neural processes in the visual areas of the brain [74] , [75], [76]. Furthermore, the use of a relatively large pool of subjects allows us to compare computational models in terms of their consistency in explaining brain activity. The fMRI data is compared to HMAX and BoW models. For the HMAX model we test how Gabor and visual dictionary representation of an image explain brain activity. Similarly for BoW, we test how SIFT and visual dictionary explain brain activity. If the models are good representations of intermediate features in the human brain, they should account for brain activity across multiple subjects.

Testing hierarchical models of vision against brain activity is challenging for two reasons. First, both computational and neural representations of visual stimuli are very high-dimensional but with different dimensionality. Second the different hierarchical levels of the models need to be dissociated properly in order to determine how brain activity is accounted by each of the individual hierarchical levels of the model. This cannot be done easily in standard multivariate neuroimaging analysis. We address the first challenge by using dissimilarity matrices [77] that capture computational and neural distances between any pair of stimuli. The second is resolved by applying variation partitioning [78] on the dissimilarity matrices. This enables us to compute the unique contributions of the hierarchical layers of HMAX and BoW models in explaining neural activity. Distance based variation partitioning has been successfully used in ecological and evolutionary studies, and will be applied here to fMRI data. This will enable us to establish correspondence between computational vision models, their different hierarchical layers and fMRI brain activity.

2.2 MATERIALS AND METHODS

2.2.1 *Computational models*

The different hierarchical levels of the HMAX and BoW model are used to represent the image. For the HMAX model, we compute the C1 (Gabor) representation as the first level and the C2 (visual dictionary) representation as the second level (Figure 3). Similarly for the BoW model, we compute the SIFT representation and visual dictionary representation.

HMAX model

We use the HMAX model [79], where features are computed hierarchically in layers: an initial image layer and four subsequent layers, each built from the previous layer by alternating template matching and max pooling operations as seen in Figure 5. In the first step, the greyscale version

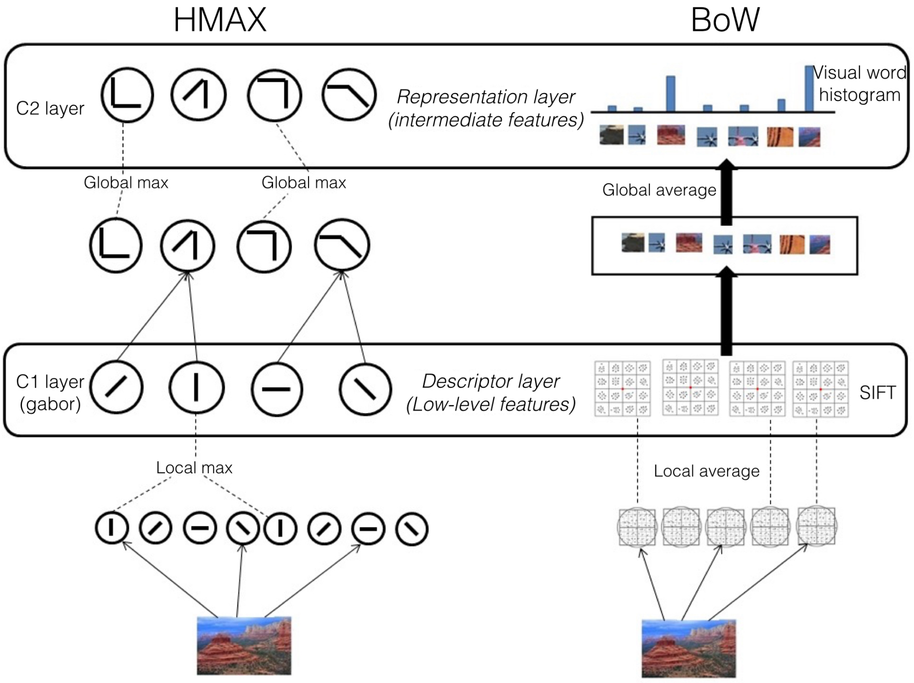


Figure 3: Computational models: HMAX model : Gabor filters of 4 orientations and 10 different scales are convolved in the S1 Layer. The image response obtained by gabor filter convolution are pooled to form the C1 layer representation. The training dataset of PASCAL images is used to create the visual dictionary at S2 layer. The visual dictionary is created by 4096 random samples of the C1 layer response from all training images. For a new image, S2 layer response is obtained by convolution of the visual dictionary elements on the C1 responses. A global max pooling operation is done for the final C2 Layer which is of dimension 4096. BoW model : SIFT features are extracted densely over the image. Visual dictionary of dimension 4000 is learnt by k-means clustering on SIFT features extracted from PASCAL dataset images. Each SIFT descriptor from an image is encoded to the nearest element of visual dictionary. Average pooling is done to form the 4000 dimension visual dictionary representation.

of an image is downsampled and a image pyramid of 10 scales is created. Gabor filters of four orientations are convolved over the image at different positions and scales in the next step, the S1 layer. Then in the C1 layer, the Gabor responses are maximally pooled over $10 \times 10 \times 2$ regions of the responses from the previous layer (the max filter is a pyramid). The Gabor representation of an image I is denoted by the vector $\mathbf{f}_{gabor}$.

In the next step, template matching is performed between the patch of C1 units centered at every position/scale and each of P prototype patches. These $P = 4096$ prototype patches are learned as done in [79] by randomly sampling patches from the C1 layer. We use images from the PASCAL VOC 2007 dataset [69] to sample the prototypes for the dictionary. In the last layer, a P dimensional feature is created by maximally pooling over all scales and orientations to one of the models P patches from the visual dictionary. This results in a visual dictionary representation of image I denoted by the vector $\mathbf{f}_{vdhmax} = [h_1 ... h_P]$ where each dimension h_p represents the max response of the dictionary elements convolved over the output of the C1 layer.

BoW model

The first step in the BoW model (Figure 5) is extraction of SIFT descriptors [72] from the image. SIFT combines a scale invariant region detector and a descriptor based on the gradient distribution in the detected regions. The descriptor is represented by a 3D histogram of gradient locations

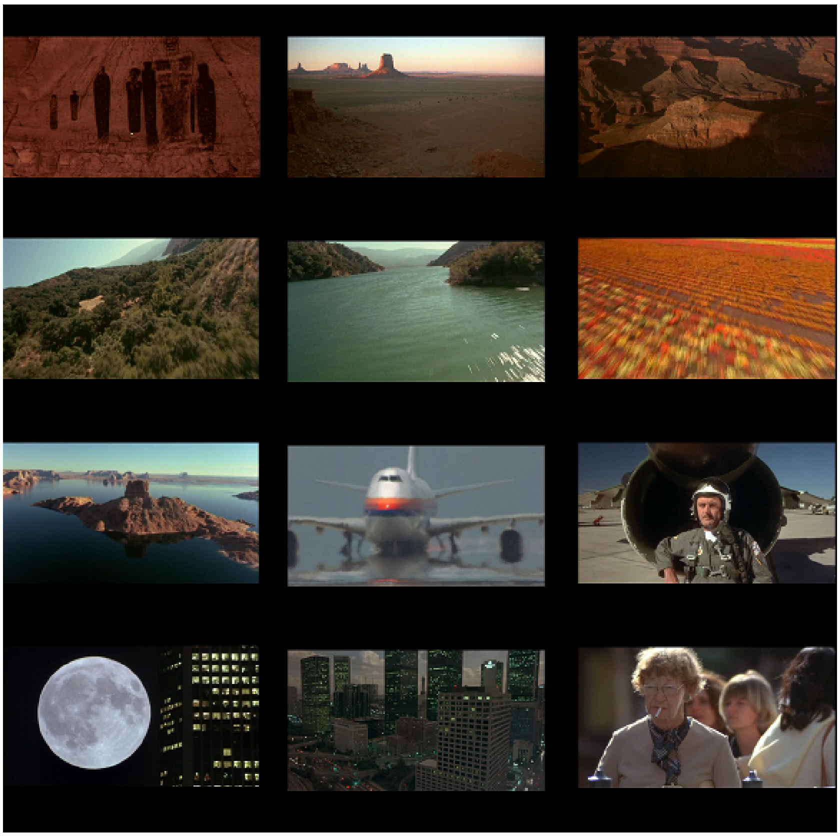


Figure 4: Example images (frames) from the 11 minute video stimuli that was used in the fMRI study. There are totally 290 scenes representing a wide variety of scenes, ranging from natural to man-made.

and orientations weighted by the gradient magnitude. The quantization of gradient locations and orientations makes the descriptor robust to small geometric distortions and small errors in the region detection. SIFT feature is a 128 dimensional vector which is computed densely over the image. Here the SIFT representation of an image I is obtained by concatenating all the SIFT features over the image. It is denoted by the vector $\mathbf{f}_{sift}$.

Secondly, a dictionary of visual words [68] is learned from a set of scenes independent of the scenes in the stimuli video. We use k-means clustering to identify cluster centers $\mathbf{c}_m = \mathbf{c}_1, ..., \mathbf{c}_M$ in SIFT space, where $m = 1, ..., M$ denotes the number of visual words. We use the PASCAL VOC 2007 [69] dataset to create a codebook of dimension $M = 4000$.

The SIFT features of a new image are quantized (assigned to the nearest visual word) to a element in the visual dictionary and the image is represented by counting the occurrences of all words. This results for image I in the visual dictionary representation $\mathbf{f}_{vdbow} = [h_1...h_M]$ where each bin h_m indicates the frequency(number of times) the visual word $\mathbf{c}_m$ is present in the image.

2.2.2 *Representational dissimilarity matrices*

A representational dissimilarity matrix [77] (RDM) F is computed separately for each of the image representations. The elements in this matrix are the Euclidean distance between the representations of pairs of images. Thus F_{gabor}, F_{vdhmax}, F_{sift} and F_{vdbow} are dissimilarity matrices for the different representations respectively. Figure 5 shows the 290×290 dissimilarity matrices for 290 images (frames) from the video stimulus used in this study.

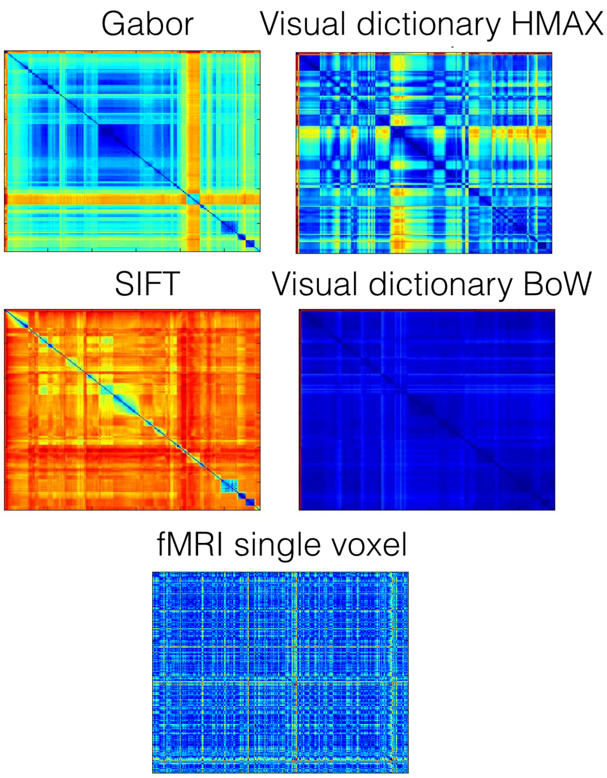


Figure 5: Dissimilarity matrices computed for the different hierarchical levels of HMAX and BoW using pairwise distance between the 290 stimuli frames. For each image, the output at a layer of the computational model is vectorized. Pairwise distance between the 290 images is computed using the vectorized outputs that results in a 290 × 290 dissimilarity matrix. Similarly the dissimilarity matrix computed for the fMRI brain responses where each element is the distance in multivariate voxel responses to any image pair resulting in a 290 × 290 matrix.

2.2.3 *Stimuli*

An 11-minute video track consisting of about 20 different dynamic scenes was used for this study. The scenes were taken from the movie Koyaanisqatsi: Life Out of Balance and consisted primarily of slow motion and time-lapse footage of cities and many natural landscapes across the United States as in Figure 4.

The movie Life Out of Balance was chosen as a stimulus because it contained all kinds of scenes we encounter in our daily live with no human emotional content or specific storyline, from natural (e.g. forest) to more man made scenes (e.g. streets). Images within one particular scene refer mostly to the same location and/or background under different conditions, such as luminance, scale (zoom), motion (moving camera or moving objects on the foreground), etc. Overall, the images varied from natural scenes such as beach, fields, rocks etc to man-made scenes such as faces, crowds, buildings, cars, planes etc. In this respect, the movie is rich in its underlying low-level properties such as spatial frequency and color.

2.2.4 *Subjects*

The fMRI data of the video stimuli was collected for over 500 subjects, from which 20 were randomly sampled for this study. Subjects were not assigned with any specific tasks when watching. They watched the video track passively one time each. The experiment was approved by the ethical committee of the University of Amsterdam and all participants gave written informed consent prior to participation. They were rewarded for participation with either study credits or financial compensation.

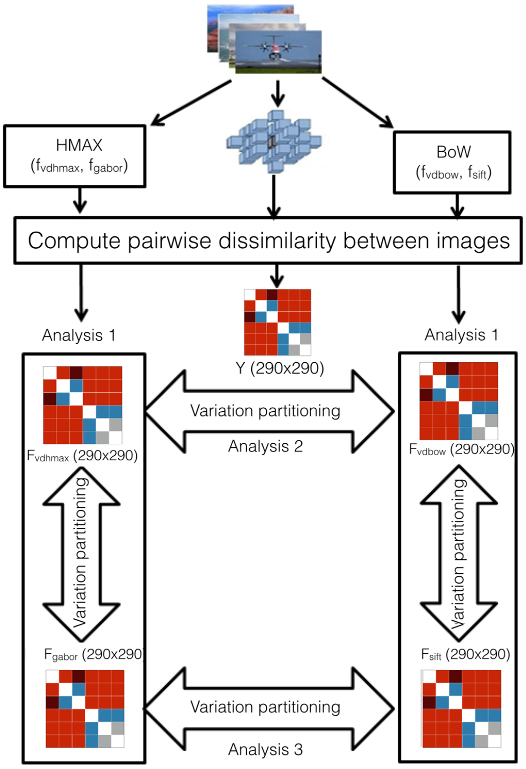


Figure 6: Analysis : Visualization of the different analysis done using variation partitioning with RDMs from the models on the fMRI RDMs voxel-wise obtained from the 290 images of the ID1000 stimuli. For the HMAX model we obtain a 290 × 290 Gabor dissimilarity matrix(F_{gabor}) and visual dictionary dissimilarity matrix(F_{vdhmax}) using pairwise image distances. Similarly for BoW, we obtain 290 × 290 SIFT dissimilarity matrix(F_{sift}) and visual dictionary matrix(F_{vdbow}). Then variation partitioning is applied at each of the hierarchical level and across the hierarchical levels on the 290 × 290 fMRI dissimilarity matrix(Y).

2.2.5 *fMRI*

We recorded 290 volumes of BOLD-MRI (GE-EPI, $192^2 mm$, 42 slices, voxel size of $3 \times 3 \times 3.3$, TR 2200 ms, TE 27.63 ms, SENSE 2, FA 90°) using a 3T Philips Achieve scanner with a 32 channel headcoil. A high-resolution T1-weighted image (TR, 8.141 ms; TE, 3.74 ms; FOV, $256 \times 256 \times 160$ mm) was collected for registration purposes. Stimuli were backward-projected onto a screen that was viewed through a mirror attached to the head-coil.

2.2.6 *fMRI Preprocessing*

FEAT (fMRI Expert Analysis Tool) version 5.0, part of FSL [80] was used to analyze the fMRI data. Preprocessing steps included slice-time correction, motion correction, high-pass filtering in the temporal domain ($\sigma = 100s$), spatially filtered with a FWHM of 5 mm and prewhitened [81]. Data was transformed using an ICA and we subsequently, automatically identified artefacts using the FIX algorithm [82]. Structural images were coregistered to the functional images and transformed to MNI standard space (Montreal Neurological Institute) using FLIRT (FMRIB's Linear Image Registration Tool; FSL). The resulting normalization parameters were applied to the functional images. The data was transformed into standard space for cross-participant analyses, so that the same voxels and features were used across subjects.

These 290 image frames and volumes were used to establish a relation between the two computational models and BOLD responses. Although in this approach the haemodynamic response might be influenced by other image frames, we expect this influence to be limited because the video is slowly changing without any abrupt variations. In addition, BOLD responses are intrinsically slow and develop over a period of up to 20 seconds. Still they summate linearly reasonably well [83] and also match the timecourse in typical scenes which develop over multiple seconds. This also probably explains the power of BOLD-MRI in decoding the content of movies [84] and indicates it is possible to compare different models of information processing on the basis of MRI volumes.

2.2.7 *Variation Partitioning*

A $3 \times 3 \times 3$ searchlight cube is centered at each voxel in the brain and BOLD responses within the cube to each of the 290 still images compared against each other. This results for each subject and for each voxel in a 290×290 dissimilarity matrix Y. Each element in the Y matrix is the pairwise distance of the 27 dimensional (from the searchlight cube) multivariate voxel responses to any image pair. As a distance measure Cityblock is taken. We now perform variation partitioning voxel-wise (each voxel described by its searchlight cube) for all the voxels across all subjects.

Variation partitioning [78] for the HMAX model is done by a series of multiple regression, producing fractions of explained variation R^2_{gabor} (unique to gabor representation), $R^2_{gaborvdhmax}$ (common to both gabor and visual dictionary representation) and R^2_{vdhmax} (unique to visual dictionary). First the multiple regression of Y against F_{gabor} and F_{vdhmax} together is computed, where Y denotes the fMRI dissimilarity matrix, and F_{gabor} and F_{vdhmax} the Gabor and visual dictionary dissimilarity matrices respectively. The corresponding R^2_{hmax} measures the total fraction of explained variation, which is the sum of the fractions of variation R^2_{gabor}, $R^2_{gaborvdhmax}$ and R^2_{vdhmax}. Then the multiple regression of Y against F_{gabor} is computed. The corresponding $R^2_{gabor+gaborvdhmax}$ measure is the sum of the fractions R^2_{gabor} and $R^2_{gaborvdhmax}$. In the next step, the multiple regression of Y against F_{vdhmax} is obtained, with corresponding $R^2_{vdhmax+gaborvdhmax}$ being the sum of the fractions of variation $R^2_{gaborvdhmax}$ and R^2_{vdhmax}. The fraction of variation uniquely explained by the Gabor dissimilarity matrix is computed by substraction: $R^2_{gabor} = R^2_{hmax}$ - $R^2_{vdhmax+gaborvdhmax}$. Similarly, variation uniquely explained by visual dictionary dissimilarity matrix is: $R^2_{vdhmax} = R^2_{hmax}$ - $R^2_{gabor+gaborvdhmax}$. The residual fraction may be computed by: $1 - (R^2_{gabor} + R^2_{gaborvdhmax} + R^2_{vdhmax})$.

Exactly the same steps of computation are taken to determine the fraction of variation uniquely explained by the SIFT dissimilarity matrix, the fraction explained by BoW visual dictionary dissimilarity matrix, and by the combination of both the SIFT and visual dictionary dissimilarity matrices as shown in Figure 3.

We also compare the models at their respective hierarchical levels. At the first level, Gabor and SIFT dissimilarity matrices are used to explain brain activity Y. Similarly at the level of visual dictionaries, we compare how HMAX and BoW visual dictionary dissimilarity matrices explain

Y. It is important to note that HMAX and BoW models refer to the explained variance attributed to both the hierarchical levels, including the shared variance..

Note that these R^2 statistics are the canonical equivalent of the regression coefficient of determination, R^2 [78]. They can interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

A permutation test (1000 times) determines the statistical significance (p value) of the fractions that we obtain for each voxel by variation partitioning. To account for the multiple comparison problem, we perform cluster size correction and only report here clusters of voxels that survive the statistical thresholding at $p < 0.05$ and have a minimum cluster size of 25 voxels. We determine the minimum cluster size by calculating the probability of a false positive from the frequency count of cluster sizes within the entire volume, using a Monte Carlo simulation [85].

2.3 RESULTS

2.3.1 *Comparing Full Models : Intersubject consistency*

Using distance-based variation partitioning for each subject we dissociate the explained variation of the HMAX model into unique contributions of Gabor R^2_{gabor} and visual dictionary representation R^2_{vdhmax}. The total explained variation by HMAX model is given by the combination of R^2_{gabor} and R^2_{vdhmax}. We do the same for the BoW model, based on SIFT R^2_{sift} and visual dictionary representation R^2_{vdbow}. HMAX and BoW models refer to the entire hierarchical model combining low level feature and visual dictionary. Cluster size correction (p¡0.05 and minimal cluster size of 25 voxels) was performed to solve for the multiple comparison problem.

To test whether our results are consistent across subjects, for each voxel we counted the number subjects for which brain activity was explained significantly by the HMAX and BoW models. A spatial version of the chi-square statistic [86] was subsequently applied to determine whether the observed frequency at a particular voxel deviated significantly from the expected value (the average number of subjects across all voxels).

Figure 7A shows how consistently across subjects, HMAX and BoW models account for brain activity. We observe that the HMAX model explains brain activity in areas V2 and V3 consistently across subjects. In these areas the HMAX model explains brain activity in overlapping voxels for 16 out of 20 subjects. In contrast, the BoW model accounts for brain activity across wider and bilateral regions including V1, V2 and V3. Most consistency is found at the left V3 and V4 regions, where for 14 out of 20 subjects, the BoW model was relevant in explaining brain activity. This difference in the number of subjects is not significant however the extent of the voxels is much more for BoW than HMAX.

Both HMAX and BoW models use low level features (Gabor filters and histogram of orientations) as their first step of computation. This is explicitly modeled and tested in our study (low-level feature representation in Figure 3). This explains why low level visual regions such as V1 and V2 emerge in our results. Interestingly, however, the BoW model also accounts for brain activity in regions higher up in the visual system such as V4 and LO (lateral occipital cortex). These regions are hypothesized to process intermediate features. This suggests that while both models appropriately represent low-level features, the transformation of these features to intermediate features is better modeled by BoW. Figure S1 in the supplementary section shows for each individual subject the explained variation of the two representational levels in both the models.

We observe that for the HMAX model the combination of hierarchies provide 5% of additional explanation compared to the maximum explaining hierarchical level. The two levels of the

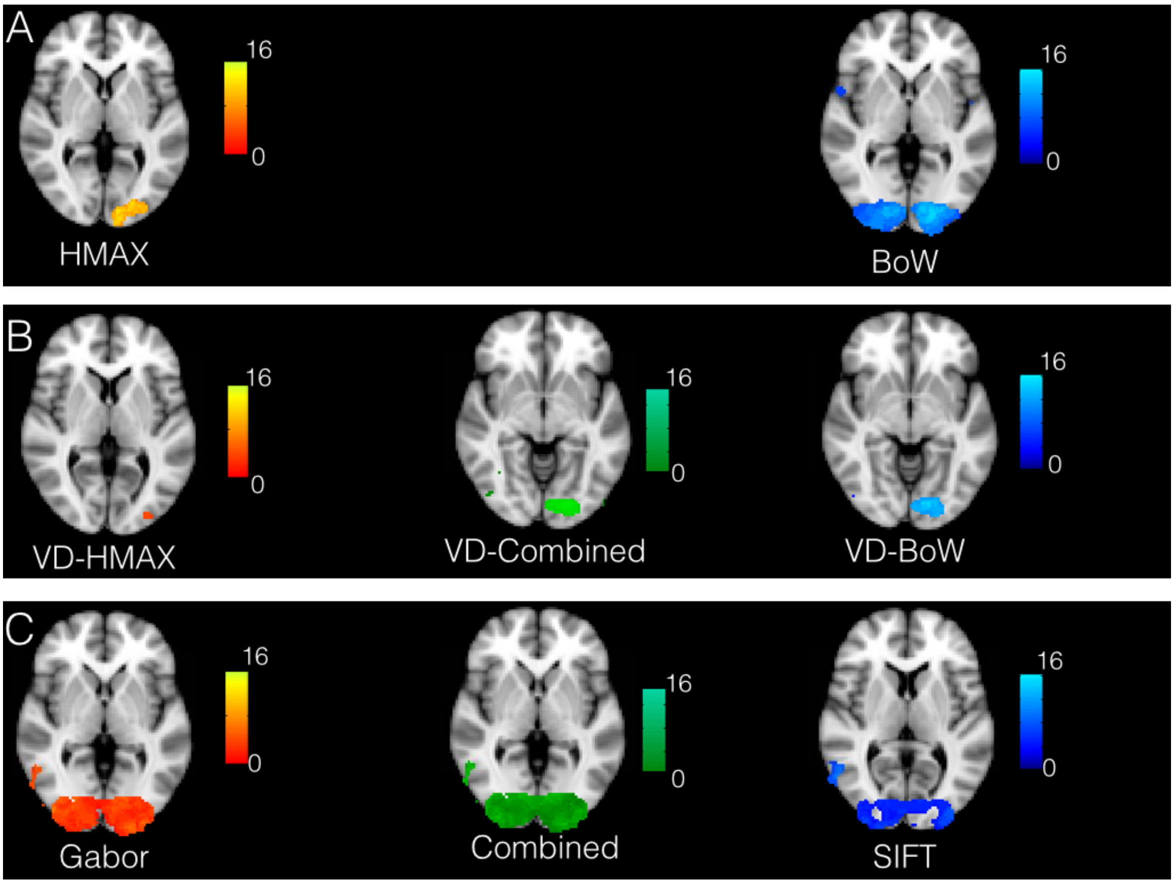


Figure 7: Visualization of across subject consistency at each voxel for the complete HMAX and BoW models and their individual components. To find consistency across subjects, first significant voxel clusters are determined subject wise and then a spatial frequency count is performed on detected clusters across subjects. (A) Across subjects consistency for HMAX (the total variation in brain responses explained by Gabor and Visual dictionary) and BoW (total variation in brain responses explained by SIFT and Visual dictionary) model based on voxel clusters and the spatial chi-square statistic (Analysis 1 in Figure 6). (B) Across subjects consistency obtained by variation partitioning of visual dictionaries from HMAX (VD HMAX), BoW (VD BoW). This results in the unique variation by VD HMAX and VD BoW and their combination (Analysis 2 in Figure 6). (C) Across subjects consistency per voxel obtained by variation partitioning of Gabor and SIFT 290x290 RDMs. The unique contribution due to Gabor, SIFT and their combination (Analysis 3 in Figure 6) is visualized.

BoW together additionally account for 8% of the variation in brain activity. A t-test on the two distributions of additional explained variations show a significant difference ($p < 0.0001$). Thus in both models, but more strongly in BoW, the aggregation of low level features into visual dictionaries describes brain activity, not captured by individual hierarchical levels. Thus the aggregation of low level features into visual dictionaries provide additional value to account for brain activity. The hierarchical levels in BoW contribute slightly more to the explained brain activity as compared to the hierarchical levels from HMAX.

2.3.2 *Comparing visual dictionaries : Intersubject consistency*

We tested the two visual dictionary representations against each other. As before, we use variation partitioning on the visual dictionary dissimilarity matrices from HMAX and BoW to explain Y. For each voxel we counted the number of subjects for which brain activity was explained significantly by the visual dictionary from HMAX and BoW models. A spatial version of the chi-square statistic [86] was applied to determine whether the observed frequency at a particular voxel deviated significantly from the expected value (the average count across all voxels).

Figure 7B shows the across subject consistency of visual dictionaries from HMAX and BoW models ($p < 0.05$, cluster size correction). We observe for the HMAX visual dictionary representation that consistency across subjects occurs in few voxels in area V4. In contrast the visual dictionary representation of the BoW model explains brain activity in areas V3 and V4 for 14 out of 20 subjects. The combination of visual dictionary representation explain brain activity for 14 out of 20 subjects in areas V3 and V4.

The visual dictionary representation from the BoW model has a much higher across subject consistency than the HMAX model. In addition the results of the combined model are similar to those of the BoW visual dictionary representations, suggesting that the HMAX visual word representation adds little to the BoW representation in terms of accounting for brain activity. Moreover, the BoW visual dictionary representation is localized in an area V4 that is hypothesized to compute intermediate features. Altogether, these results suggest that the BoW model provides a better representation for visual dictionaries, compared to the HMAX model. Single subject results confirming consistency across subjects can be found in the supplementary section 11.

2.3.3 *Comparing low-level feature representations : Intersubject consistency*

We tested Gabor and SIFT representation against each other. As before, we use variation partitioning on Gabor and SIFT representations to explain Y. For each voxel we counted the number subjects for which brain activity was explained significantly. The spatial version of the chi-square statistic was applied to determine whether the observed frequency at a particular voxel deviated significantly from the expected value (the average count across all voxels).

Figure 7C shows the across subject consistency of Gabor and SIFT representations (p¡0.05, cluster size correction). We observe that the Gabor representation explains brain activity in early visual areas for a large number of voxels such as V1, V2 and V3. The Gabor representation also explains brain activity consistently across subjects in the higher brain areas such as LO and precentral gyrus for 10 out of the 20 subjects. Similarly for the SIFT representation we observe that it explains brain activity in the lower visual areas such as V1, V2 and also higher areas of the brain such as LO across 9 out of 20 subjects. Overall Gabor and SIFT representations account for brain activity in similar areas of the brain. It is expected that Gabor and SIFT explain brain responses in early visual areas since both rely on edge filters. However it is interesting to observe that they also explain brain activity in the higher areas of the brain.

We also observe areas where Gabor and SIFT together explain neural response consistently across subjects. The combination of Gabor and SIFT representations explain brain activity in 14 out of 20 subjects in the early visual area V1. The combination also explains brain activity in higher areas of the brain such as V4 and LO. This suggests that Gabor and SIFT representation have complementary low-level gradient information. Taken together, Gabor and SIFT provide a better computational basis for V1 representation.

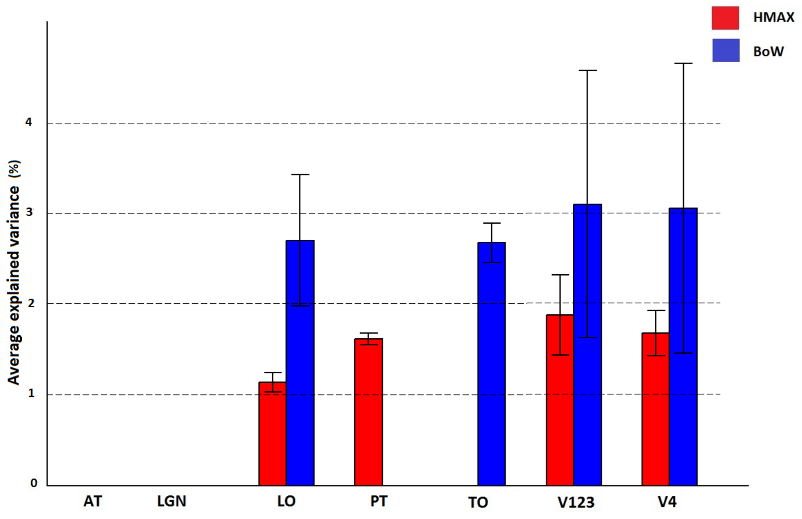


Figure 8: Visualization of brain activity explained by HMAX and BoW model across all subjects in the selected ROIs. The explained variation of the significant voxels ($p < 0.05$ and cluster size correction) are averaged across subjects over all voxels in a ROI.

	No of Significant Voxels		Max Explained Variation	
	HMAX	BoW	HMAX	BoW
AT	0	0	0	0
LGN	0	0	0	0
LO	43	387	2	4
TO	0	31	0	4
V123	701	3671	3	5
V4	53	1141	2	5

Table 1: Number of significant voxels and maximum explained variation for HMAX and Bow models in each ROI. The significant voxels ($p < 0.05$ and cluster size correction) are averaged across subjects over all voxels in a ROI.

2.3.4 *Cross subject ROI analysis*

A region of interest analysis was conducted to explicitly test the sensitivity of different brain regions to the models and their individual components. Figure 8 shows how HMAX and BoW explain brain activity in 6 brain regions (out of the 25 brain areas analyzed). These ROIs are obtained based on the Jülich MNI 2 mm atlas. We show the explained variation for each model averaged across subjects and the voxels within each ROI (Note that this doesn't show single subject variation across ROIs). We observe that there is significant explained variation in areas TO (temporal occipital), LO, explained variationV123 and V4. The representations do not account for brain activity in areas such as LGN (lateral geniculate nucleus) and AT (anterior temporal). In all the regions the BoW model has a higher average explained variation than the HMAX model The difference in explained variation is significant ($p < 0.0001$). Table 1 shows the number of voxels in each ROI obtained across subjects that exhibited significant brain activity and the maximum explained variation across subjects. We observe that the HMAX and BoW models explain more brain activity in early visual areas compared to the other areas.

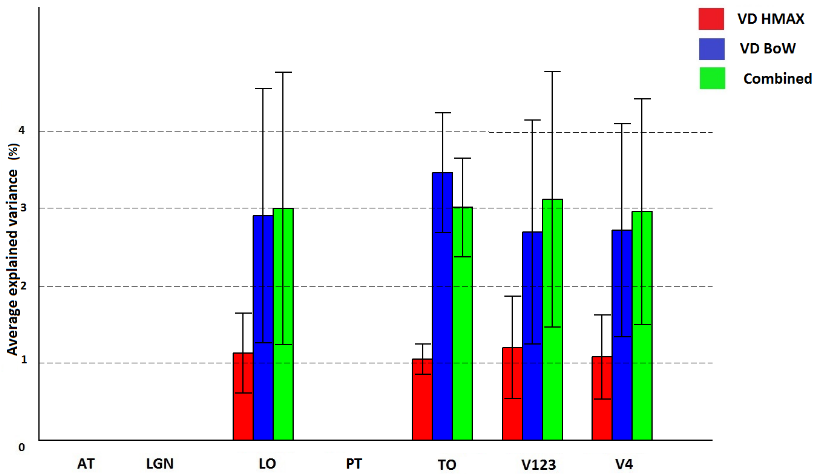


Figure 9: Visualization of brain activity explained uniquely and the combination of visual dictionaries from HMAX, BoW across all subjects in the selected ROIs. The explained variation of the significant voxels ($p < 0.05$ and cluster size correction) are averaged across subjects over all voxels in a ROI.

	No of Significant Voxels			Max Explained Variation		
	VD HMAX	VD BoW	Combined	VD HMAX	VD BoW	Combined
AT	0	0	0	0	0	0
LGN	0	0	0	0	0	0
LO	1427	1255	1402	3	9	10
TO	178	143	164	2	11	8
V123	4818	5705	5716	3	6	6
V4	1682	1878	1910	2	6	6

Table 2: Number of significant voxels and maximum explained variation for visual dictionaries from each ROI. The significant voxels ($p < 0.05$ and cluster size correction) are averaged across subjects over all voxels in a ROI.

Figure 9 shows how visual dictionaries from HMAX and BoW explain brain activity in the 6 brain regions (out of the 25 brain areas analyzed). It can be seen that there is significant explained variation in areas TO (temporal occipital), LO, V123 and V4. The average explained variation is slightly higher in the TO regions compared to V123. In all the regions the visual dictionary from BoW model has a higher average explained variation compared to the visual dictionary from the HMAX model ($p < 0.0001$). Also the combination of visual dictionaries from HMAX and BoW do not significantly increase the explained variation and is similar to the explained variation from BoW. Table 2 shows that the visual dictionary from both the models explains a large number of voxels in LO and V4, however the visual dictionary from BoW has highest explained variations in LO and TO compared to HMAX. Also, we do not notice any brain activity in brain regions such as parahippocampal gyrus, retrosplenial corted and medial temporal lobe for either HMAX and BoW models.

Figure 10 and Table 3 show how Gabor and SIFT explain brain activity in the 6 brain regions (out of the 25 brain areas analyzed). We observe that there is significant explained variation in areas TO (temporal occipital), LO, V123 and V4. For Gabor, the average explained variation is slightly higher in the V123 region compared to the other areas. Here we also observe that

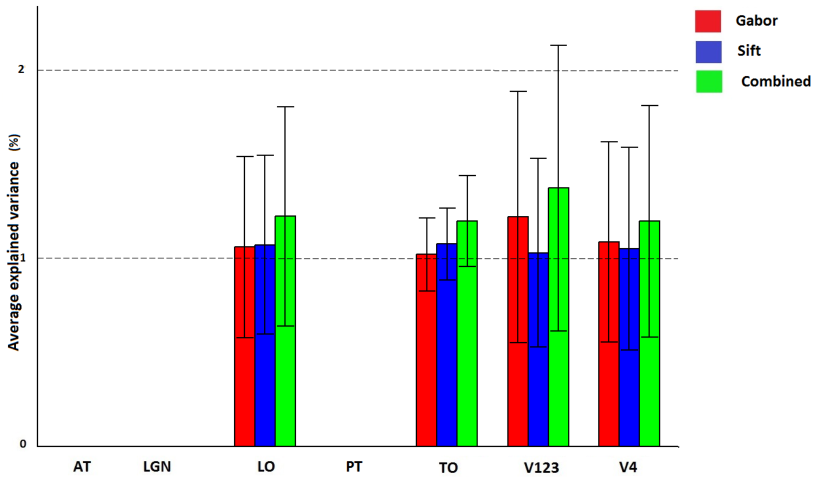


Figure 10: Visualization of brain activity explained uniquely and the combination of Gabor and SIFT across all subjects in the selected ROIs. The explained variation of the significant voxels ($p < 0.05$ and cluster size correction) are averaged across subjects over all voxels in a ROI.

	No of Significant Voxels			Max Explained Variation		
	Gabor	SIFT	Combined	Gabor	SIFT	Combined
AT	0	0	0	0	0	0
LGN	0	0	0	0	0	0
LO	1394	1283	1427	2	2	2
TO	183	152	186	2	2	2
V123	5531	4418	5718	3	2	3
V4	1850	1535	1912	2	2	3

Table 3: Number of significant voxels and maximum explained variation for each ROI. The significant voxels ($p < 0.05$ and cluster size correction) are averaged across subjects over all voxels in a ROI.

the Gabor and SIFT representations are not significantly different from each other and also the combination explains brain acitivity to the same extent.

Overall these results suggest that individually, the BoW visual dictionary is a better computational representation of neural responses (as measured by percent explained variation and consistency across subjects) than the visual dictionary from HMAX, which provides little additional information over the BoW visual dictionary.

2.4 DISCUSSION

The success of models such as HMAX and BoW can be attributed to their use of features of intermediate complexity [87]. The BoW model in particular has proven capable of learning to distinguish visual objects from only five hundred labeled examples (for each category of twenty different categories) in a fully automatic fashion and with good recognition rates [82]. Many variations of this model exists [88], [89], and the recognition performance on a wide range of visual scenes and objects, has improved steadily year by year [82]. The HMAX model is a biologically plausible model for object recognition in the visual cortex which follows the hierarchical feedforward nature of the human brain. Both the models are candidate computational models of intermediate visual processing in the brain.

Our results show that in early visual brain areas such as V1, V2 and V3 there are regions in which brain activity is explained consistently across subjects by both the HMAX and BoW models. Both the models rely on gradient information to compute image representations. In the HMAX model, Gabor filters [90] similar to the receptive fields in the V1 region of the brain are at the basis of visual representation. Similarly in the BoW model, the Scale Invariant Feature Transform (SIFT) features are the low level representation based on multi-scale and multi-orientation gradient features [72]. Although SIFT features originate from computer vision, their inspiration goes back to [33]'s simple and complex receptive fields, and [91]'s Neocognitron model. SIFT features thus have an embedding in the visual system, much like Gabor filters have. In light of this, the sensitivity in brain areas V1, V2 and V3 to representations of the HMAX and BoW models is natural and in part due to low-level features. Interestingly we also observe that SIFT and Gabor representations explain brain activity in higher regions of the brain. This indicates that neurons in higher level visual areas process low level features pooled over local patches of the image for feedforward or feedback processing within visual cortex.

Brain areas higher up in the processing hierarchy appear to be particularly sensitive to visual dictionaries. Visual dictionaries are medium size image patches that are informative and distinctive at the same time, allowing for sparse and intermediate representations of objects and scenes. In computer vision visual dictionaries have proven to be very effective for object and scene classification [88]. The brain may compute visual dictionaries as higher-level visual building blocks composed of slightly larger receptive fields, and use visual dictionaries as intermediate features to arrive at a higher-level representation of visual input. We observe that both HMAX and BoW visual dictionaries explain some brain activity in higher level visual regions, with the BoW visual dictionary representation outperforming the HMAX model both in terms of explained variance and consistency.

HMAX and BoW both use low level features that are pooled differently in the various stages of processing. First HMAX pools Gabor features by a local max operator whereas BoW creates a histogram of orientations (SIFT). Then, BoW uses a learning technique (k-means clustering [92]) on all the SIFT features from the image to form the visual dictionary. On the other hand HMAX uses random samples of Gabor features pooled over patches as its visual dictionary. This difference in aggregating low level features might explain why BoW provides a better computational representation of images.

Visual dictionaries may facilitate scene gist perception, which occurs rapidly and early in visual processing [42], [93]. While there is evidence that simple low-level regularities such as spatial frequency [94], colour [95] and local edge aligment [96] underly scene gist representation, it is hitherto unknown whether and how mid-level features facilitate scene gist perception. BoW summarizes SIFT features computed over the entire image. It has been observed that such patterns of orientations and scales are believed to be used by V4 and IT [97]. This is in accordance with our observation that the localization of BoW visual dictionary representations occur in V4 and areas anterior to V4 in the brain.

Our findings are in line with a recent study by Leeds et. al [65]. They compared multiple vision models against MRI brain activity in response to image stimuli. Leeds et. al conclude that the BoW model explains most brain activity in intermediate areas of the brain. For this model, they report that the correlation of the BoW model varies from 0.1 to 0.15 across the 5 subjects. In our study, we obtain similar results for the BoW model, and with an average explained variation across subjects of around 5% (with explained variations varying across subjects). Similarities and consistencies between our results and results in [65] further suggest that BoW computation might provide a suitable basis for intermediate features in the brain. [66]

observe explained variance of up 25% for both HMAX and BoW models, and up to 50% for their HMO model (4-layer Convolutional neural network model) in brain areas IT and V4. The discrepancy between these results and our findings in terms of the magnitude of explained brain activity can be in part attributed to the use of high signal-to-noise ratio measurements in [66], such as electrophysiological data from monkeys. The neural sensitivity to convolutional neural network model is nevertheless promising. We will include deep neural networks in future work to understand how it performs on video stimuli.

Our study aims to understand if intermediate features used in the brain are connected to how computational models of vision use such intermediate features. Our findings suggest that visual dictionaries used in HMAX and BoW account for brain activity consistently across subjects. The result does not imply that visual dictionaries as computed by HMAX or BoW are actually used by the brain to represent scenes but it does suggest visual dictionaries might capture aspects of intermediate features. The results from this work are similar to previous work and provides new interesting insights into the nature of intermediate features in the brain. We have also provided a novel framework which allows us to dissociate the different levels of a hierarchical model, and individually understand their contribution to explain brain activity.

2.5 SUPPLEMENTARY INFORMATION

2.5.1 *Individual subjects*

Using distance-based variation partitioning for each subject we dissociate the explained variation of the HMAX model into unique contributions of Gabor $R^2_{F_{gabor}}$ and visual dictionary representation $R^2_{F_{hmax}}$. We do the same for the BoW model based on SIFT $R^2_{F_{sift}}$ and visual dictionary representation $R^2_{F_{bow}}$. This resulted for each subject and for each model in two volumes indicating the strength and the significance of the explained variation. Cluster size correction (p¡0.05 and minimal cluster size of 25 voxels) was performed to solve for the multiple comparison problem.

Figure 11 shows in the color red, areas and strengths of the unique contribution of the first hierarchical level of the models, Gabor and SIFT, in accounting for brain activity. It can be observed that for many subjects Gabor mainly explains brain activity in the higher visual areas such as Lateral occipital cortex, precuneous cortex, precentral gyrus. On average Gabor explains 4% of the variation in brain activity. For individual subjects the explained variation peaks at 8%. A similar pattern can be observed for SIFT: it mainly explains brain activity in higher areas involved in visual processing such as Lateral occipital cortex and middle temporal gyrus. SIFT has lesser average explained variation of 3%, and the explained variation peaks to 6% at the subject level.

The areas depicted in figure 11 in blue are associated with visual word representations. As can be observed the visual dictionaries from HMAX and BoW mainly account for activity in brain areas involved in early visual processing such as V1, V2, V3 and V4. Other brain areas such as LO and TO however also show activity that is explained by visual word dictionaries. While spatially the visual dictionary representations of HMAX and BoW display correspondence, the visual dictionary from HMAX explains on average up to 5% in areas V2-V3 against 18% in areas V3-V4 for the BoW model.

That visual word representations account for brain activity early in visual processing, and SIFT and Gabor representations at later stages is counterintuitive given that areas such as V1 are known to be sensitive to Gabor-like visual structures. It has to be noted, however, that we show only the unique contributions by Gabor, SIFT and they also share common explained variation with the visual dictionaries in the models in the early visual cortex.

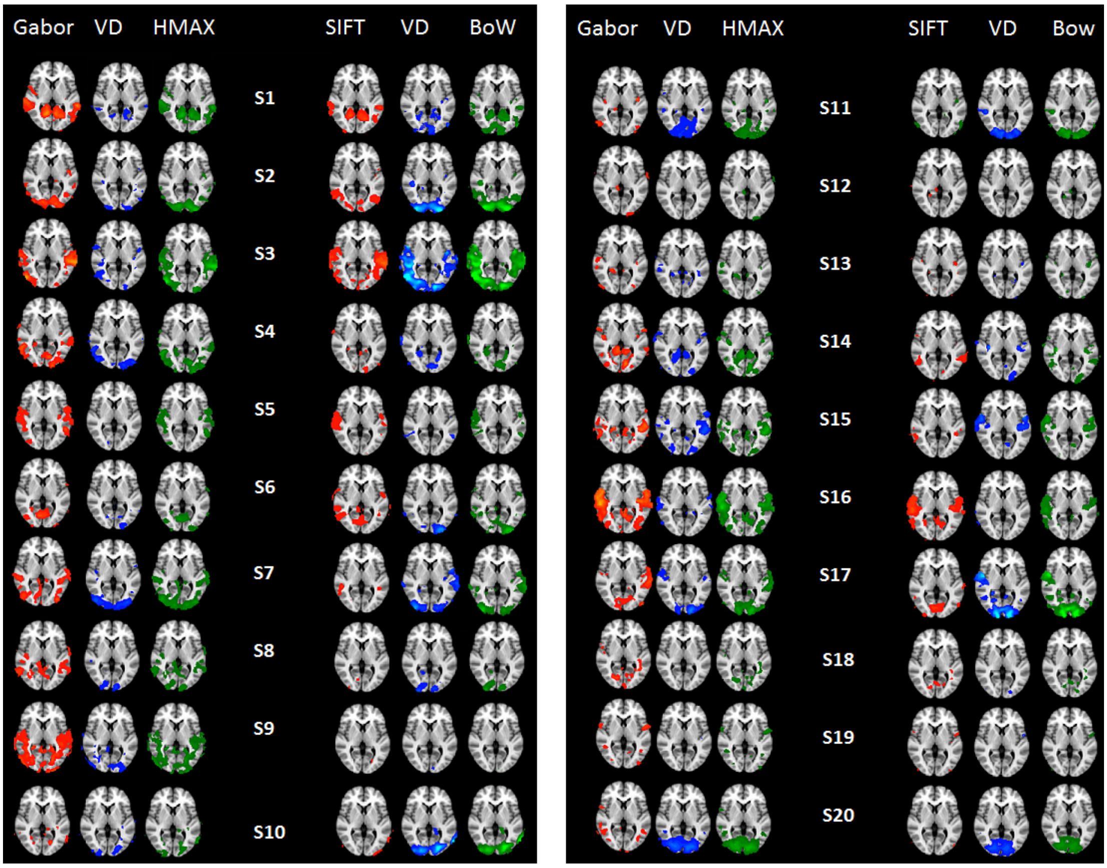


Figure 11: *Variation partitioning on Y per individual subject to determine the fraction of explained variation Gabor(red) and visual dictionary (blue) from HMAX (green) model; SIFT (red) and visual dictionary (blue) from BoW (green) model. Visualization of the explained variations of only the significant voxels ($p < 0.05$ within a cluster size of 25) for the 20 subjects (S1 - S20).*

3

NEURAL SPATIAL CONSISTENCY OF HIERARCHICAL VISUAL MODELS

SUMMARY :

The human visual system is known to use multiple stages of computation to process visual information. Similarly, hierarchical stages of computation are employed by models such as HMAX, Bag of Words(BoW) and Convolutional Neural Networks(CNN) for object recognition. These computational models are increasingly being used to understand the mechanisms underlying visual processing in the human brain. However, hierarchical models come in a variety of different forms and vary in essential characteristics such as the number of computational stages in the model and the computational operations that are applied within each stage. While these models have been tested for correspondence to brain activity in isolation, they have not been explicitly compared against each other in terms of their neural spatial map. In this paper, we provide a framework called Neural Spatial Consistency Analysis (NSCA) to compare how the different computational stages in these models explain brain responses. We used BOLD fMRI data from 20 subjects who watched a 11 minute natural movie. We employed a distance based-variation partitioning on the dissimilarity matrices of the models at the voxel level to determine the neural spatial map. The extent to which the different computational stages of the models correspond to each other is determined by correlation of the neural spatial maps. We observe that the different stages of CNN explains brain responses from low-level brain areas to higher regions of the brain. Both HMAX and BoW explain brain responses primarily in the early visual cortex. Additionally both HMAX and BoW are highly correlated to early CNN layers, further providing evidence that that both these models only capture low level information. Comparing BoW and HMAX with CNN we note that BoW has a stronger correlation to CNN than HMAX. This leads us to believe that though HMAX and CNN are biologically inspired, BoW is more similar to CNN then HMAX.

3.1 INTRODUCTION

Computational models combine operations in a hierarchical fashion for automatic object recognition as shown in Figure 12. The HMAX model [67], for example, is a biologically plausible model that uses filters with max pooling in multiple stages. HMAX is a model for the initial feedforward stage of object recognition in the ventral visual pathway. It extends the idea of simple cells (detecting oriented edges) and complex cells (detecting oriented edges with spatial invariance) by forming a hierarchy in which alternate template matching and max pooling operations progressively build up feature selectivity and invariance to position and scale. HMAX is thus a simple and elegant model to describe the tuning properties of receptive fields in the ventral visual pathway [98].

Bag of Words model (BoW) quantizes local Scale Invariant Feature Transform (SIFT) [72] features into visual words, which are abstractions of frequently occurring and distinctive image patches such as grass, sand and bricks [73]. BoW is a highly successful model, which performs very well on large TRECvid [69] and PASCAL [99] datasets, in some cases even approaching human classification performance. The key idea behind this model is to quantize local SIFT features into visual words, which are of intermediate complexity, and subsequently the image is represented by a histogram of visual words.

Recently, convolutional neural networks (CNNs) have revolutionized object recognition [100]. CNNs are hierarchical feedforward models consisting a series of linear and nonlinear operations, and parameters learned using supervised training on millions of labeled images (parameters of previous models were either hand-tuned, adapted through unsupervised learning, or trained on a much smaller number of images). The widely used CNN, AlexNet [100] model consist of 7 convolutional layers, with the top layers being fully connected. These models have achieved state of the art performance in object recognition on large public benchmark data sets such as ImageNet [101] and PASCAL [99]. More recently CNN models, such as GoogLeNet [102], consist upto 22 layers and can discriminate between 1000 visual categories with error rates on par with human performance [22].

Given the success of CNNs in computer vision for object recognition, a number of studies [103], [104] have used these models to understand visual processing in the human brain. Recent work has shown that CNNs correlate much better to cortical representations than other computational models [103]. They even rival the representation of primate IT cortex for core object recognition [104]. Moreover, there is a strong correlation between the categorization performance of CNNs on challenging high variation object sets and their ability to predict individual neural responses [66]. Thus the CNN representations correspond to the largely feedforward computations in the human brain that culminate in a powerful neuronal representation of objects and scenes.

CNNs have outperformed HMAX and BoW in the correspondence to brain responses from visual stimuli. While the models are different, they employ one or more common set of computational operations, for example filter convolution as shown in Figure 12. At the same time, there are architectural differences between the models in terms of the non-linearities used. Testing the different models in isolation [103] does not reveal to what extent the correspondence to neural responses is similar or different across models. A comparative study of these models in terms of their sensitivity to the brain can additionally reveal the relevant computational aspects of visual processing in the brain. It has not been tested yet, if there is correspondence between the different computational models and their computational stages in terms of their sensitivity to the brain.

We use variation partitioning to dissociate the unique and shared information that is captured by the different models. Additionally for such studies, representational dissimilarity matrices (RDMs) of computational models are compared to RDMs of brain responses focussing only on explained variances. Though explained variances provide important information, consistency

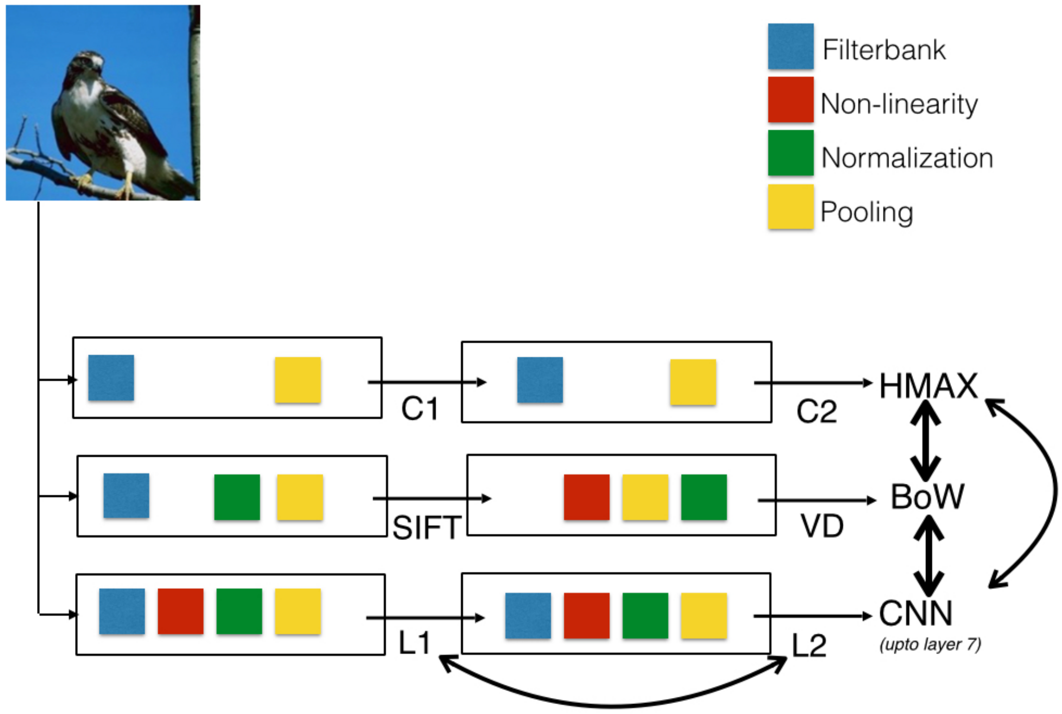


Figure 12: Hierarchical vision models and their computational stages. The HMAX model consists of two stages - C1 composed of gabor filtered image and C2 is output of visual dictionary. BoW model's first computational stage is SIFT features densely extracted on the image followed by vector quantization giving the visual word histogram representation. The convolutional neural network is a 7 layer hierarchical model consisting of filter convolution, normalization, rectification and spatial pooling. Each model is analyzed by variation partitioning its computational stages, and across model comparison is done using the final representation.

of explained variance across subjects make these comparisons meaningful. The models are compared against each other based on their consistency of explained variance across subjects.

In this chapter, we explicitly compare three widely used hierarchical models (HMAX, BoW and CNN) and their multiple computational stages on the basis of consistency across subjects. We investigate to what extent the spatial maps of consistency by these models correspond to each other. For this, we collect fMRI brain responses of 20 subjects during passive viewing of a natural movie. A sequence of analysis is done to determine the spatial consistency of explained variance in brain responses by the different models. This gives us a neural spatial map, which is then used for correspondence across models.

3.2 MATERIAL AND METHODS

3.2.1 *Subjects*

The fMRI data of the video stimuli was collected for over 1000 subjects, from which 20 were randomly sampled for this study. Subjects were not assigned with any specific tasks and watched the video track passively one time each. The experiment was approved by the ethical committee of the

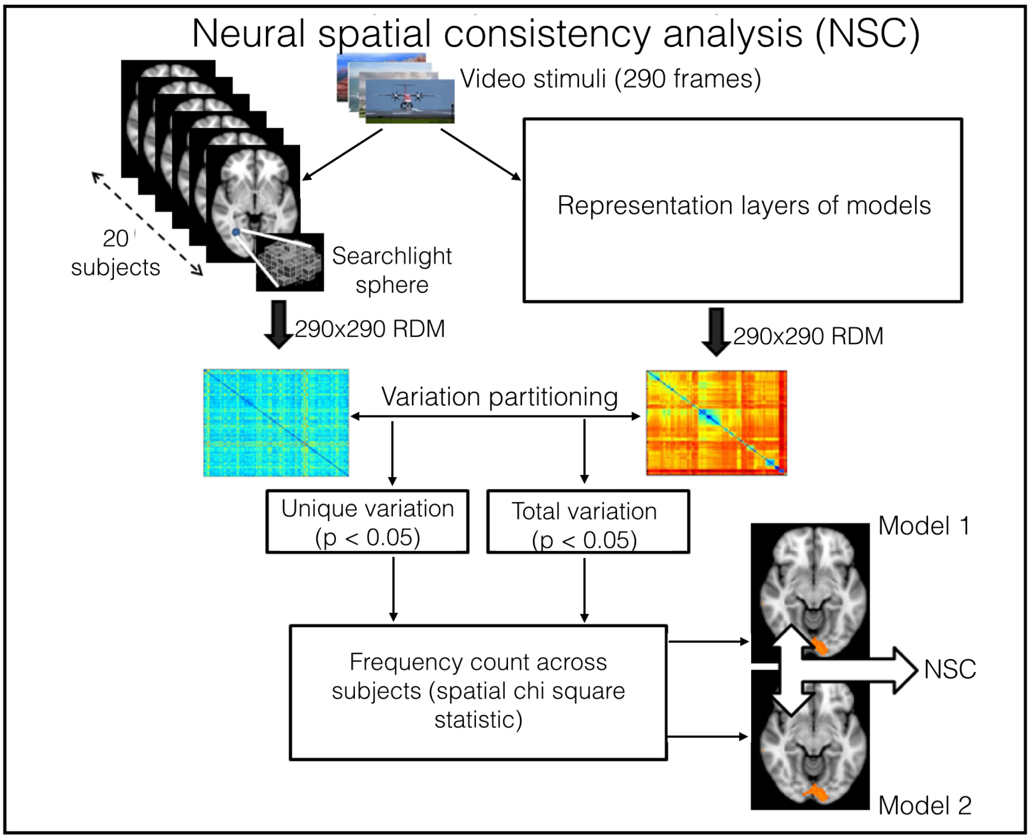

Neural spatial consistency analysis (NSC)
Video stimuli (290 frames)
Representation layers of models
20 subjects
Searchlight sphere
290x290 RDM
290x290 RDM
Variation partitioning
Unique variation (p < 0.05)
Total variation (p < 0.05)
Frequency count across subjects (spatial chi square statistic)
Model 1
NSC
Model 2

Figure 13: Neural Spatial Consistency Analysis framework : The stimuli consists of 290 image frames on which the representations from HMAX, BoW and CNN are computed. For every pair of images, a dissimilarity measure is computed from the representations. A 290x290 dissimilarity matrix is obtained for each of the different representation layers of HMAX, BoW and CNN. Similarly, for brain responses, 290x290 RDM is computed voxel-wise using a searchlight technique. For each subject, variation partitioning is employed on image representation RDMs to obtain the unique and total explained variance in voxel RDMs. To test whether the results are consistent across subjects, for each voxel we counted the number subjects for which brain activity was explained significantly by the different models(both total and unique). This is the spatial consistency map which is compared against models.

University of Amsterdam and all participants gave written informed consent prior to participation. They were rewarded for participation with either study credits or financial compensation.

3.2.2 *Stimuli*

An 11-minute video track consisting of about 20 different dynamic scenes was used for this study. The scenes were taken from the movie Koyaanisqatsi: Life Out of Balance and consisted primarily of slow motion and time-lapse footage of cities and many natural landscapes across the United States (example frames from the video stimuli shown in SI Figure 1). Scenes include beach, fields, rocks, faces, crowds, buildings, cars, planes captured under different conditions, such as luminance, scale (zoom), motion (moving camera or moving objects on the foreground). Images within one particular scene refer mostly to the same location and/or background under different conditions, such as luminance, scale (zoom), motion (moving camera or moving objects on the foreground), etc. Overall, the images varied from natural scenes such as beach, fields, rocks etc to man-made scenes such as faces, crowds, buildings, cars, planes etc such that it covered a wide range on the frequency spectrum.

3.2.3 *fMRI*

We recorded 290 volumes of BOLD-MRI (GE-EPI, $192^2 mm$, 42 slices, voxel size of $3 \times 3 \times 3.3$, TR 2200 ms, TE 27.63 ms, SENSE 2, FA 90°) using a 3T Philips Achieve scanner with a 32 channel headcoil. A high-resolution T1-weighted image (TR, 8.141 ms; TE, 3.74 ms; FOV, $256 \times 256 \times 160$ mm) was collected for registration purposes. Stimuli were backward-projected onto a screen that was viewed through a mirror attached to the head-coil. Subjects were placed supinely inside the scanner and watched the movie via back-projection. The movie was presented using a DLP beamer (120 Hz). The total distance from the subject's eyes to the screen was 156 cm with a resolution of 1920x1080. The movie extended 16 degrees in horizontal direction.

Subjects where not asked to fixate although action typically was in the center of the screen. Differences in subject fixation targets could have resulted in an increase of variance between subjects, in particular for lower visual areas V1 and V2. However, given that the explained variance here is within range of that published in literature eye movements (or differences in eye movements between subjects) of subjects cannot have been a large factor in introducing variance in the data. The degree to which this did influence our results should be proportionate for all brain areas, or, at worse stronger for lower than higher-tier areas.

In total, during the experiment 290 volumes were obtained. Taking into account the haemodynamic delay of 6 seconds we identified the 290 image frames corresponding to the 290 volumes. The BOLD values were used at those 290 volumes.

Preprocessing

FEAT (fMRI Expert Analysis Tool) version 5.0, part of FSL [80] was used to analyze the fMRI data. Preprocessing steps included slice-time correction, motion correction, high-pass filtering in the temporal domain ($\sigma = 100s$), spatially filtered with a FWHM of 5 mm and prewhitened [81]. Data was transformed using an ICA and we subsequently, automatically identified artefacts using the FIX algorithm [82]. Structural images were coregistered to the functional images and transformed to MNI standard space (Montreal Neurological Institute) using FLIRT (FMRIB's Linear Image Registration Tool; FSL). The resulting normalization parameters were applied to the functional images. The data was transformed into standard space for cross-participant analyses,

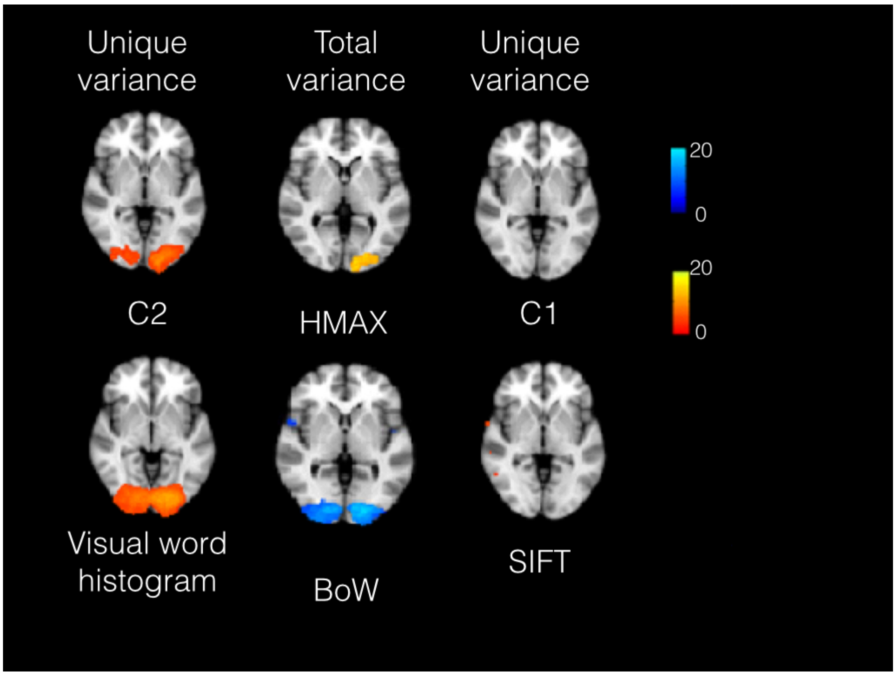


Figure 14: A Brain map showing voxelwise the consistency for total and unique explained variation by the computational stages of HMAX and BoW. Here variation partitioning was applied on successive layers to determine the explained variation. We observe that for both models, the unique variation by lower computational stages C1 and SIFT are not consistent across subjects. The higher computational stages do explain significant unique variance.

so that the same voxels and features were used across subjects. Although in this approach the haemodynamic response might be influenced by other image frames, we expect this influence to be limited because the video is slowly changing without any abrupt variations. In addition, BOLD responses are intrinsically slow and develop over a period of up to 20 seconds. Still they summate linearly reasonably well [83] and also match the timecourse in typical scenes which develop over multiple seconds. This also probably explains the power of BOLD-MRI in decoding the content of movies [84] and indicates it is possible to compare different models of information processing on the basis of MRI volumes.

Data representation

We use the outputs of the different computational stages from the HMAX, BoW and CNN to represent images. The output of C1 and C2 layer in HMAX is used to represent the image. For BoW, SIFT and visual dictionary histogram is computed to represent the image. The CNN is a pertained AlexNet [105], and the output of 7 layers is used to represent the image. A representational dissimilarity matrix [77] (RDM) F is computed separately for each of the representations at different layers of the models. The elements in this matrix are the Euclidean distance between the representations of pairs of images.

For the fMRI representations, a $3 \times 3 \times 3$ searchlight cube is centered at each voxel in the brain and BOLD responses within the cube to each of the 290 still images compared against each other. This results for each subject and for each voxel in a 290×290 dissimilarity matrix Y. Each element in the Y matrix is the pairwise distance of the 27 dimensional (from the searchlight cube) multivariate voxel responses to any image pair. As a distance measure Cityblock [106] is taken.

3.2.4 *Neural spatial consistency analysis*

Comparison of model RDMs to voxel RDMs is based on distance-based variation partitioning [16]. Using distance-based variation partitioning, for each subject we dissociate the explained variation in brain responses by each layer of the hierarchical model into total and unique variation. A permutation test (1000 times) determines the statistical significance (p value) of the fractions that we obtain for each voxel by variation partitioning. To account for the multiple comparison problem, we perform cluster size correction and only report here clusters of voxels that survive the statistical thresholding at $p < 0.05$ and have a minimum cluster size of 25 voxels. We determine the minimum cluster size by calculating the probability of a false positive from the frequency count of cluster sizes within the entire volume, using a Monte Carlo simulation [85].

To test whether our results are consistent across subjects, for each voxel we counted the number subjects for which brain activity was explained significantly by the different models(both total and unique). A spatial version of the chi-square statistic [86] was subsequently applied to determine whether the observed frequency at a particular voxel deviated significantly from the expected value (the average number of subjects across all voxels). This gives us the neural spatial map, i.e voxel clusters that are encoded by the different models separately. The neural spatial consistency map is determined separately for the total and unique variation using the same analysis. We compute the correlation ($p < 0.05$) between the neural spatial consistency map of the different computational models to determine the correspondence between the models.

3.3 RESULTS

3.3.1 *Computational models*

To determine the extent to which the different computational stages in each model consistently explain variance in brain responses, we visualized the neural spatial maps for HMAX, BoW and CNN in Figure 14 and 15.

Figure 14A shows the neural spatial map of unique and total variance in brain activity explained by HMAX. Depicted is unique variation accounted by Gabor, the C2 level and the combination of the two stages ($p < 0.05$, cluster size correction). We observe that there are no voxels for which Gabor filters uniquely explain brain activity. This suggests that Gabor filter stage and the C2 stage of computation are highly correlated. We observe that the C2 stage of computation uniquely explains brain activity primarily in the early visual cortex V123. Similarly the combination of these stages in the HMAX model consistently explains brain responses in early visual areas of the brain for 16 out of the 20 subjects.

Similarly, for BoW model Figure 14B shows us the spatial map for SIFT, Visual word histogram and the combined variation of both the stages ($p < 0.05$, cluster size correction). We see a similar pattern to what we observe for the HMAX model. SIFT is observed to be highly correlated to the visual word histogram and thus there is no unique variation due to SIFT. The visual word histogram and the combined variation for the BoW model is consistent across 14 subjects in early visual areas such as V123 and V4.

Fig 15 shows us the spatial map for the unique and combined variation of the seven CNN layers ($p < 0.05$, cluster size correction). We observe that layers 1-5 of CNN explain brain responses primarily in the early visual areas of the brain such as V123. Layers 6 and 7 explains brain responses consistently in the higher areas of the brain such as V4 and lateral occipital cortex(LO). There is high consistency observed in the lower layers, between 18-20 subject while in layers 6 and 7 consistency is observed in 10-12 subjects.

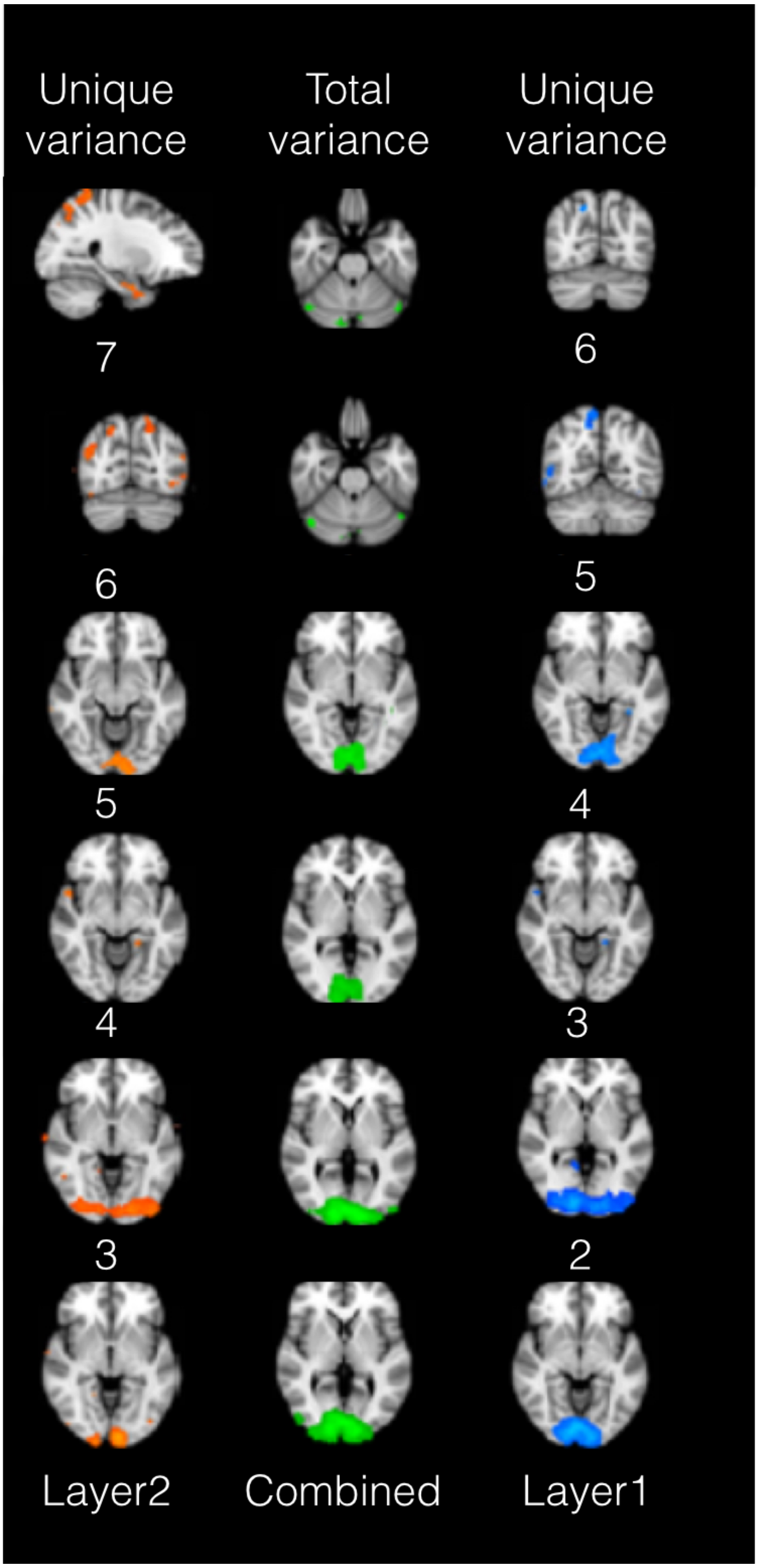


Figure 15: A Brain map showing voxel-wise the consistency for total and unique explained variation by the seven CNN layers. Here variation partitioning was applied on successive layers to determine the explained variation. For example, variation partitioning on layers 2 and 1 results in unique explained variation by layer 2, 1 and total variation by layers 2 and 1 combined. We observe that the CNN layers consistently explain variance in brain responses in wide-spread brain areas.

3.3.2 *Comparing models*

To investigate to what extent the explained variance in brain response is unique among the different computational models, a pairwise variation partitioning was done on the three different models. Figure 16 shows the resulting spatial maps on comparing the models explicitly against each other, i.e HMAX versus CNN, BoW versus CNN and HMAX versus BoW.

The comparison of CNN model to HMAX ($p < 0.05$, cluster size correction) in Figure 16A shows that, while HMAX primarily explains brain responses in the early visual brain areas the CNN model not only explains brain responses in early visual areas but also in higher visual areas. Similarly, comparing BoW to CNN ($p < 0.05$, cluster size correction), we observe from Figure 16B that the CNN model explains brain responses higher in the visual hierarchy while BoW is primarily sensitive in early visual brain areas (Table 4 and Table 5 in supplementary information shows the number of consistent voxels in eight different ROIs). This suggests that the BoW, HMAX capture different information in the brain as compared to the CNN model. Thus, a model that combines CNN and HMAX or CNN and BoW will better explain brain responses in both early and higher level visual areas.

The comparison of HMAX and BoW shows that while both HMAX and BoW are consistent in the brain areas V3 and V4, HMAX accounts for brain activity in lesser number of voxels than BoW. This suggests that a combination of BoW and HMAX will only marginally add to the individual model in explaining brain responses.

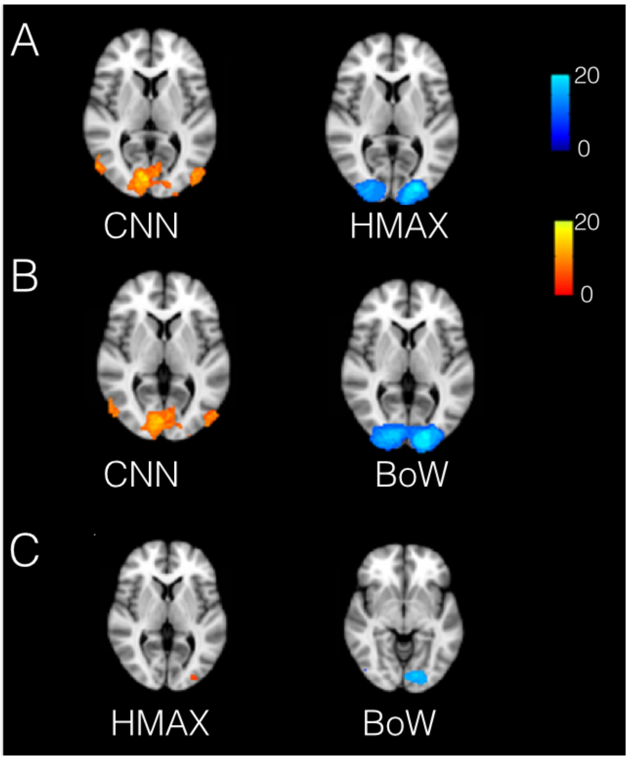
3.3.3 *Model correspondence*

Our previous results showed that HMAX and BoW explain brain responses in low level visual areas, while CNN explains brain responses deeper in the brain. This raises the question how similar are the spatial maps of HMAX or BOW to the different CNN layers. To study the extent to which BoW or HMAX is correlated to CNN, we correlate the neural spatial maps of these models (from Figure 15 and 16) to understand what stages of computation are similar and different. Since both HMAX and BoW clearly explain brain responses in early visual areas, we indeed find significant correlation between their neural spatial map but do not present it here since it is expected.

Figure 17 shows how the HMAX and BoW model correlate across the different layers of CNN ($p < 0.005$). For the total explained variation, we see that the significant correlation increases and peaks at lower CNN layers and drops in the higher layers of the brain. There is no significant correlation between HMAX or BoW with combination of CNN layer 6 and 7 ($p > 0.005$. For the different computation stages i.e, the unique explained variation, we observe a similar pattern for visual word histogram and the C2 layer ($p < 0.005$. This clearly indicates that both HMAX and BOW capture only low level representations in the human brain and do not account for higher level representations as captured by layer 6 and 7 of the CNN.

Further, we observe that the correlation for BoW to CNN is higher than the correlation of HMAX to CNN. This suggests that BoW accounts for brain responses in a similar manner to CNN and the HMAX less so. Figure 12 shows that BoW and CNN employ additional operations such as normalization and rectification in each stage of computation while HMAX does not include such operations. This might partly explain why CNN and BoW are more similar than HMAX in terms of their neural spatial map.

Overall, our results show that CNNs outperform BoW and HMAX to explain brain responses to natural stimuli consistently across subjects.



1

Figure 16: Neural spatial consistency map obtained on comparing the final representations from CNN, BoW and HMAX. Variation partitioning was done on CNN and BoW representations, CNN and HMAX representation and finally the BoW and HMAX representations. Comparing CNNs to either HMAX or BoW, we observe that CNNs correlate to brain areas higher in the visual hierarchy while BoW/HMAX correlate to brain areas lower in the visual hierarchy.

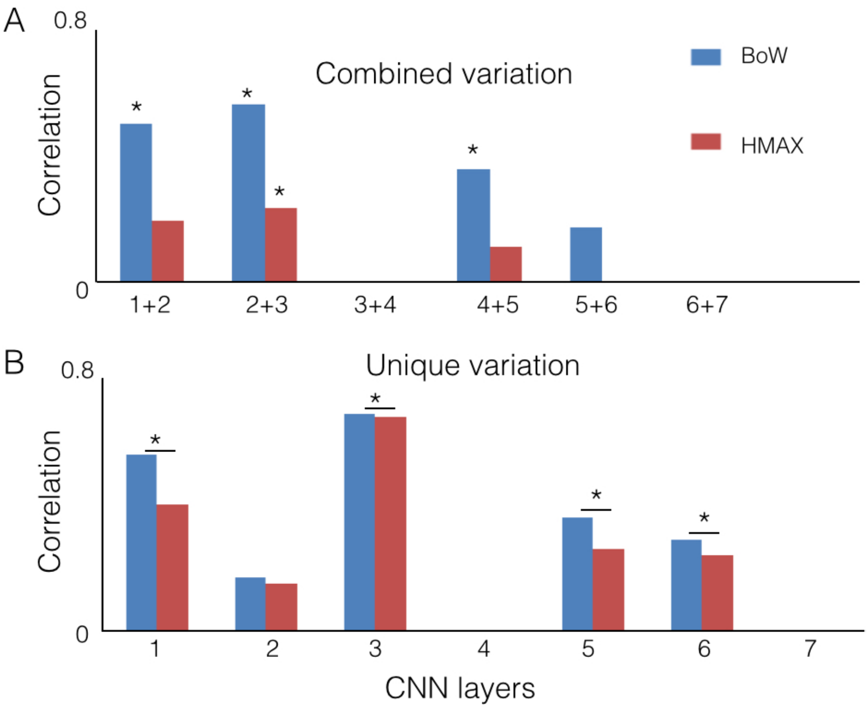


Figure 17: Correlating the neural spatial consistency maps of BoW and HMAX to the maps of the CNN layers (both total and unique variance). We observe that the BoW and HMAX significantly correlate to the lower and intermediate CNN layers, 3 and 2.

3.4 DISCUSSION AND CONCLUSION

The success of computational models for object recognition may be attributed to their use of hierarchical stages of computation [107]. This is now increasingly being demonstrated for CNNs, which have more computational stages then other hierarchical models such HMAX and BoW [108]. CNN not only outperforms other models on various object recognition tasks but also corresponds to the hierarchy of visual representations in the human visual cortex. In line with these findings, our results indicate that CNN explain brain responses consistently on a large number of subjects in the anterior regions of the brain compared to HMAX and BoW.

We observe that the different stages of CNN explain brain responses from early visual areas to higher visual areas that are responsible for object recognition. BoW and HMAX on the other hand, which have a fewer number of stages comparatively, primarily explain brain responses in V123 and extending upto V4. This might be due to the fact that the CNN employs seven layers of computational stages which allows it to perform complex nonlinear mapping of low level representations. In computer vision, this stacking of multiple linear/non-linear computational steps [18] in a deep network is also one of the reasons for superior performance of CNNs object recognition tasks.

Neural spatial consistency analysis reveals that HMAX and BoW are correlated most to the third layer of CNN. HMAX and BoW transform low level descriptors to image representations via one intermediate step. Thus one would expect the second layer, which transforms the low level features from the first layer in CNN to correlate most with HMAX and BoW. The high correlation of these models to the third layer of CNN possibly indicates that the second and the third layer of CNN might compute representations similar to the intermediate representations of HMAX and BoW. Similarly, the C2 stage of computation and visual word histogram is highly correlated to the layer 3 of CNN. The lower CNN layers are shown to compute simple edge-like features or intermediate shape features [109]. This comparison across models provides further evidence that HMAX and BoW capture visual information of low or intermediate complexity.

While both HMAX and BoW are correlated to the different layers of CNN, it is not to the same extent. We observe that the BoW model and the visual work histogram stage in general are more correlated to CNN than HMAX and the C2 stage. Each computation stage consists of a number of operations such as - convolution, normalization, non-linear operation etc which varies across models seen in Figure 12. In Figure 12, we see that the number of similar operations in each stage is closer for BoW and CNN than HMAX. HMAX employs only one convolutional step followed by max pooling. BoW and CNN employ additional non-linearization, normalization in each stage of computation. Thus, this similarity between the models is probably traced back in the brain too. These results suggest that the different type of operations are important for computational models to explain brain responses. The different computational operations has been explicitly studied in a recent work [110].

There a number of open standing questions that we will address in the future. At the level of analysis, we would like to see how the combination of these models would perform on object recognition tasks and to encode brain responses. We are also interested in how varying the different operations in each computational step will further improve these models. More generally, it is interesting to study the repeatability of our results on two groups of subjects and when using fMRI responses to a wider range of natural scenes. Preliminary results show that there is indeed consistency in detected brain regions.

3.5 SUPPLEMENTARY INFORMATION

We defined 8 different functionally specialized region-of-interest (ROIs) [111] in the brain, to understand how the models consistently explain variance in brain responses in different visual areas. The ROIs are V123, V4, parahippocampal gyrus (PHC), medial superior temporal area (MST), middle temporal (MT), intraparietal sulcus (IPS) and superior parietal lobule (SPL).

Table 4 shows the number of voxels in the ROIs on comparison of BoW and CNN. Similarly Table 5 shows the number of voxels obtained on comparison of HMAX and CNN.

	No of Significant Voxels		
	BoW	CNN	Combined
V123	19639	9213	20245
V4	2905	1073	2888
PHC	238	209	280
MST	38	435	29
hMT	219	551	217
LO	3076	1010	3281
IPS	75	625	510
SPL	0	96	0

Table 4: ROI analysis : Number of significant voxels ($p < 0.05$ and cluster size correction) of the neural spatial consistency map obtained from variation partitioning of BoW and CNN representations.

	No of Significant Voxels		
	HMAX	DNN	Combined
V123	14540	3222	15462
V4	2566	433	2813
PHC	156	180	242
MST	35	462	126
hMT	182	571	285
LO	2797	733	2700
IPS	75	941	319
SPL	3	130	22

Table 5: ROI analysis : Number of significant voxels ($p < 0.05$ and cluster size correction) of the neural spatial consistency map obtained from variation partitioning of HMAX and CNN representations.

4

TEMPORAL DYNAMICS OF VISUAL OBJECT RECOGNITION BY CNNS

SUMMARY :

Convolutional neural networks (CNNs) have recently emerged as promising models of human vision based on their ability to predict hemodynamic brain responses to visual stimuli measured with functional magnetic resonance imaging (fMRI). However, the degree to which CNNs can predict temporal dynamics of visual object recognition reflected in neural measures with millisecond precision is less understood. Additionally, while deeper CNNs with higher numbers of layers perform better on automated object recognition, it is unclear if this also results into better correlation to brain responses. Here, we examined 1) to what extent CNN layers predict visual evoked responses in the human brain over time and 2) whether deeper CNNs better model brain responses. Specifically, we tested how well CNN architectures with 7 (CNN-7) and 15 (CNN-15) layers predicted electro-encephalography (EEG) responses to several thousands of natural images. Our results show that both CNN architectures correspond to EEG responses in a hierarchical spatio-temporal manner, with lower layers explaining responses early in time at electrodes overlying early visual cortex, and higher layers explaining responses later in time at electrodes overlying lateral-occipital cortex. While the explained variance of neural responses by individual layers did not differ between CNN-7 and CNN-15, combining the representations across layers resulted in improved performance of CNN-15 compared to CNN-7, but only after 150 ms after stimulus-onset. This suggests that CNN representations reflect both early (feed-forward) and late (feedback) stages of visual processing. Overall, our results show that depth of CNNs indeed plays a role in explaining time-resolved EEG responses.

4.1 INTRODUCTION

The near-human performance of convolutional neural networks (CNNs) [100] on automated object recognition has led to a number of neuroimaging studies that investigated the correlation of CNNs to feedforward visual processing in the human brain. It has been shown that CNNs correlate much better to cortical representations measured with human neuroimaging than other computational models [103]. A similar correlation of CNNs was found to neural recordings from primate IT-cortex during core visual object recognition [66, 104]. Moreover, evidence suggests that CNNs map onto brain responses in a hierarchical manner, with lower CNN layers predicting responses in early visual cortex and high-level layers predicting responses in category-selective cortex [112], [113], [49]. While it is increasingly becoming clear that CNNs capture hierarchical representations in the human visual system, there are a number of open questions.

First, the impressive performance of CNNs in predicting brain responses has mostly been demonstrated for fMRI responses derived from slow fluctuations in blood flow across multiple brain regions in visual cortex [114]. However, object recognition is a fast process that is resolved within the initial hundreds of milliseconds of visual processing [115]. A cascade of visual processing stages gives rise to characteristic spatio-temporal dynamics shaped by both feedforward and feedback processing [116] that can be measured with time-resolved magneto- and electro-encephalography (M/EEG). While CNN layers do not have a temporal dimension, it has been show that the CNN layers predict whole-brain decoding performance of MEG responses in the first few hundred milliseconds of visual processing in the human brain [113]. Interestingly, these results suggested that, just as in the spatial domain, CNN layers correspond hierarchically to temporally resolved responses, with lower layers predicting decoding performance early, and higher layers predicting performance later in time.

Second, CNNs have outperformed shallow computer vision models on automated object recognition datasets on ImageNet [101] and PASCAL VOC [70]. The performance on object recognition has further increased with deeper neural networks consisting of 15 layers (CNN-15) and 18 layers [117] compared to 7 layer CNN [100] on large image datasets. State-of-the-art CNNs are capable to discriminate between 1000 visual categories with error rates on par with human performance [22] and overlaps with human behavior [118]. The number of layers is a critical factor influencing the performance of the CNN architecture. For instance, while the CNN-7 of 5 convolutional layers achieves an error rate of 18% (top-5 error) on ImageNet, the CNN-15 with 13 convolutional layers achieves an error rate of 7.5% [117]. In both architectures, the layers consist of similar operations such as: filter convolution, non-linear activation, spatial pooling and normalization, with each network containing 2 fully connected layers.

The improved performance of CNN-15 compared to CNN-7 is commonly attributed to the additional non-linearity at the additional layers, which improves the discriminative power of the network. With increasing number of layers, the model contains one essential non-linear operator per layer, the rectified linear unit (abbreviated to relu). However, at the same time the CNN-15 has smaller filter sizes (analogous to receptive field sizes in the human brain) per layer. The effective field size of layer 1 in CNN-7 is the same over two layers combined in the CNN-15 architecture. In this way, a correspondence of effective receptive field sizes between different CNN architectures has been established [117]. Thus, the number of layers, number of non-linear relu's and receptive field size are interrelated yet critical aspects of the CNN-architecture. The improved performance of deeper CNNs on automated object recognition raises the question whether these architectures also better model visual representations in the human brain.

In this chapter, we examine whether the temporal dynamics of visual object processing in the human brain are better predicted by CNNs with increasing number of layers. To address this, we tested: 1) whether the hierarchy of CNN representations predict differences in visual

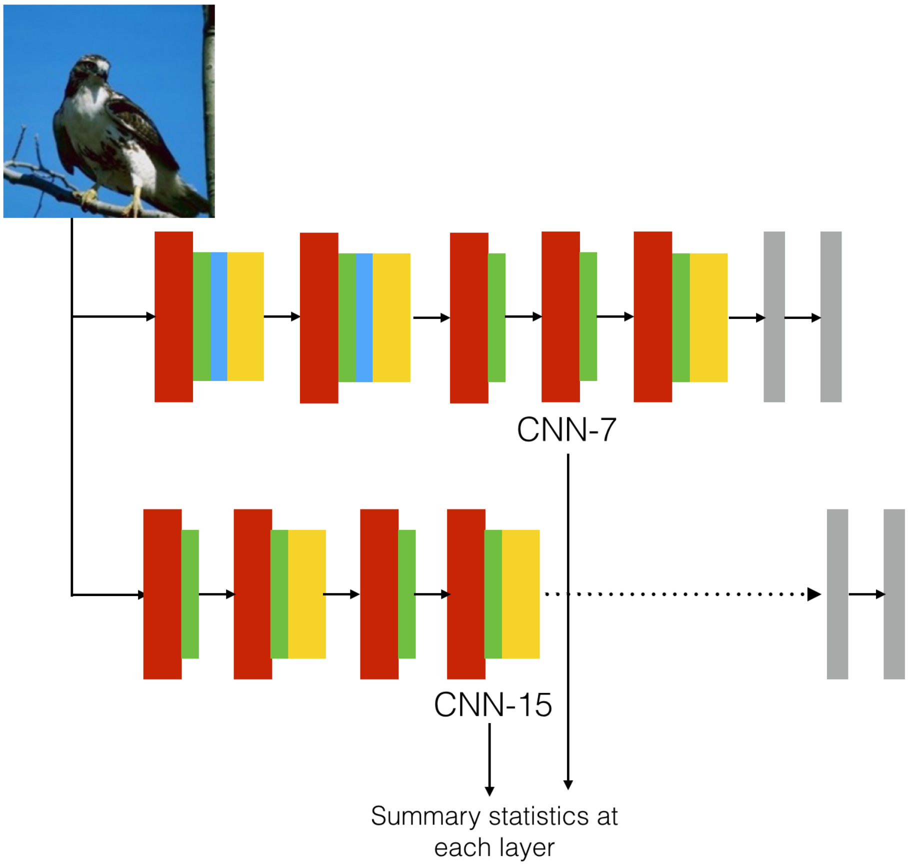


Figure 18: Models : Architectures of a 7 (CNN-7) and 15 (CNN-15) layer CNN. CNN-7 consists of 5 convolutional layers and 2 fully-connected layers. The CNN-15 consists of 13 convolutional layers and 2 fully connected layers. Summary statistics (mean and mean/standard deviation) of the representations is computed at each CNN layer.

evoked activity in a systematic manner over time and 2) if a 15 layer CNN model explains a higher variance of brain responses compared to a 7 layer CNN. To that end, we first measured event-related potentials (ERPs) responses of 20 individuals to a large set of natural images. The layers of different CNN models were regressed to each time point of the ERP responses to determine their ability to explain variance in evoked amplitude between individual images. We compared the different layers and different CNN models on the basis of explained variances obtained.

4.2 MATERIALS AND METHODS

4.2.1 *Subjects*

Twenty-one participants (7 males, 22-33 years old, mean 25.6, = SD = 2.5) took part in the EEG experiment. All participants had normal or corrected-to-normal vision, provided written informed consent and received financial compensation. The ethics committee of the University of Amsterdam approved the experiment. Two subjects were excluded in preprocessing: one subject based on the presence of excessive alpha activity, and another because of a history of epilepsy.

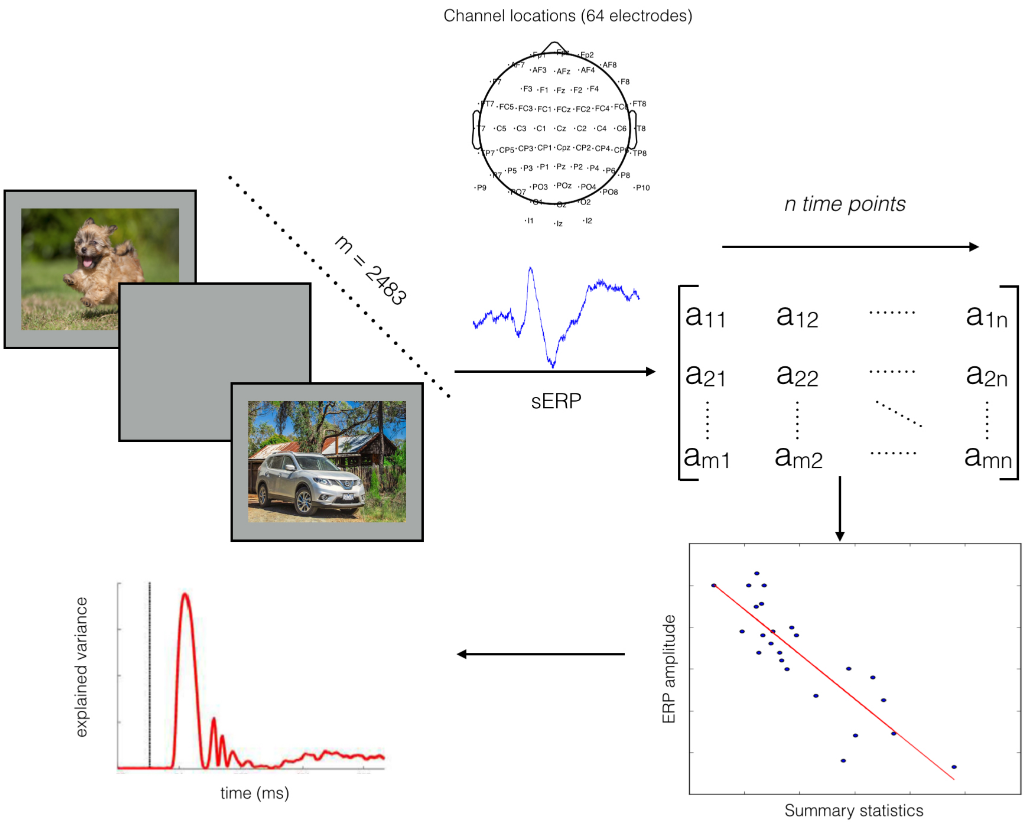


Figure 19: Experimental design : The 64 channel EEG system is used (locations of the different channels visualized). Single image ERPs at each channel, averaged over subjects, are regressed to the summary statistics of each layer from the CNNs by selection of 1800 images from the full set of 2483 images. The regression is permuted a 1000 times and this results in explained variance (r2) per layer for CNN-7 and CNN-15. We plot the average explained variance over the 1000 permutations per layer of CNN-7 and CNN-15.

4.2.2 *Stimuli and Procedure*

Participants viewed a large set of scene stimuli while performing go-no go object recognition tasks [119]. A stimulus set of 6800 color images (bit depth 24, JPG format, 640 x 480 pixels) was composed from several existing online databases. The set included images from a previous fMRI study on scene categorization [120], as well as images from various datasets used in computer vision: the INRIA holiday database [121], the GRAZ dataset [122], ImageNet [101], and the McGill Calibrated Color Image Database [123]. These different sources assured maximal variability of the stimulus set: it contained a wide variety of indoor and outdoor scenes, landscapes, forests, cities, villages, roads, images with and without animals, objects, and people. For the purpose of the current study, we analyzed only the images that are not contained in the ImageNet dataset, resulting in 2483 number of images. This was to avoid overlap in training deep neural networks and the stimuli presented to the participants. Stimuli were presented on a 19-inch Ilyama CRT-monitor (1024x768 pixels, frame rate 60 Hz). Participants were seated 90 cm from the monitor such that stimuli subtended 14x10 deg of visual angle. On each trial, one image was randomly selected and presented in the center of the screen on a grey background for 100 ms, on average every 1500 ms (range 1000 - 2000 ms). In different task blocks, participants searched for either animals or vehicles at four levels of categorization: basic detection (animal/no-animal and vehicle/no-vehicle), superordinate (animal/vehicle and vehicle/animal), basic-level (cat/other animals and bicycle/other vehicles) or subordinate categorization (Persian cat/other cats and mountain bike/other bicycles). Within each task, subjects performed a total of 800 trials (400 target, 400 non-target images). In addition, subjects performed an intact vs. scrambled scene task on a subset of 400 scene images and their Fourier phase-scrambled counterparts (only the intact scenes were analyzed). For each participant, data were obtained across three separate recording sessions that were conducted on different days. Per recording session, each participant performed each of the 7 different tasks for a particular subset of the stimuli. Task orders and stimulus subsets were counterbalanced across participants and recording sessions. Participants were instructed to respond as quickly and accurate as possible, and indicated their responses with their right hand using a custom-made button box that was taped to the chair armrest. Prior to the start of each task block, participants performed 20 practice trials on images that were not included in the main experiment. Each task block was interspersed with a short break allowing subjects to rest. Stimuli were presented using the software package Presentation (www.neurobs.com).

4.2.3 *EEG acquisition and preprocessing*

EEG Recordings were made with a Biosemi 64-channel Active Two EEG system (Biosemi Instrumentation BV, Amsterdam, NL, www.biosemi.com). Recording set-up and preprocessing were identical to the procedures described in [124], [125]. We used caps with an extended 10-20 layout, modified with two additional occipital electrodes (I1 and I2, while removing electrodes F5 and F6). During recording, a CMS/DRL feedback loop was used as an active ground, followed by offline referencing to electrodes placed on the earlobes. The Biosemi hardware is completely DC-coupled, so no high-pass filter is applied during recording of the raw data. A Bessel low-pass filter was applied starting at 1/5th of the sample rate. Eye movements were monitored with a horizontal electro-oculogram (hEOG) placed lateral to both eyes and a vertical electro-oculogram (vEOG) positioned above and below the left eye, aligned with the pupil location when the participants looked straight ahead. Data was sampled at 256 Hz. Pre-processing occurred in Brain Vision Analyzer and included a high-pass filter at 0.1 Hz (12 dB/octave); a low-pass filter at 30 Hz (24 dB/octave); two notch filters at 50 and 60 Hz; automatic removal of deflections > 300 mV; epoch segmentation in -100 ms to 500 ms from stimulus onset; ocular correction using the EOG electrodes [126]; baseline correction between -100 ms and 0 ms; automated artifact rejection and conversion to Current Source Density responses [127]. No trial or electrode averaging was performed: preprocessing thus resulted in a single-trial EEG response specific to each subject, electrode and individual image presentation. Prior to regression analysis, responses were averaged across participants, resulting in a single event-related potential (ERP) specific to each individual image.

4.2.4 *CNN representations and summary statistics*

Our analysis consisted of two stages. In the first stage, we used two different pre-trained CNN architectures: the 7-layer CNN-architecture (CNN-7) [100] and the 15-layer CNN-architecture, CNN-15 [117](Figure 1A). The CNNs used in this study were pre-trained on the ImageNet dataset with the same hyper-parameters as described in the MatConvNet toolbox (http://www.vlfeat.org/matconvnet/pretrained/).

In the second stage, we summarized the representation of each CNN layer by two parameters - the mean and mean normalized by standard deviation (Figure 18). These summary statistics have been previously found to constitute a biologically plausible model of population receptive field outputs [128] and have been used successfully fornatural image identification based on EEG-responses [129] and to describe the population activity captured by individual EEG electrodes [125]. Additionally, the use of summary statistics allows us to better handle the high dimensionality of the CNN feature representations, and to equate the number of parameters extracted for each CNN layer.

The output after rectification or pooling stage at each layer in both CNN-7 and CNN-15 is used for the analysis. The output is vectorized on which the mean and standard deviation is computed. This gives us two summary parameters per layer in each CNN-7 and CNN-15.

4.2.5 *Regression of CNN layers on single image ERPs*

To test whether differences between evoked neural responses could be predicted by the summary parameters of the CNNs, we conducted regression analyses on the single-image ERPs (Fig. 1C). The preprocessed ERPs were used in Matlab, where we conducted linear regression analyses of ERP amplitude on the CNN layer summary parameters using the Statistics Toolbox. At each channel and each time-point, two summary parameters for each CNN layer were entered together as linear regressors on single-image ERP amplitude. This analysis results in a measure of model fit (adjusted r2, corrected for predictor dimension) over time (each sample of the ERP) and space (each electrode). The fit of the regression model was statistically evaluated by permutation tests: we randomly selected 1800 out 2483 images and repeated the regression analysis a 1000 times. We averaged the r2 over all permutations to represent the correlation between CNN layers and the ERP amplitude.

4.2.6 *Statistical testing*

For significance testing of the explained variance, the permutation analysis (1,000 times) was used to calculate the standard deviation of the sampled bootstrap distribution. To correct for multiple comparisons across permutations, time-points, channels and layers, we used the Bonferroni measure, resulting in an adjusted alpha = 1e-10. To compare the results directly (comparing layers within each model), we used the Akaike information criterion (AIC), which measures the information contained in each set of predictors (summary statistics of each CNN layer). Specifically, we transformed the residual sum of squares (RSS) of the regression analysis based on each set of statistics into AIC-values using AIC = n*log(RSS/n)+2k where n = number of images and k is the number of predictors. AIC can be used for model selection given a set of candidate representations of the same visual input. The model with the minimum AIC-value is preferred [130]. Thus in comparing the different CNN layers, the CNN layer with the lowest AIC value has the best fit to evoked activity. To compare the two models directly, we use a paired t-test using the permuted values at each time point.

Regression of combined CNN layers on single image ERPs

To test the performance of each CNN model in their entirety (rather than layer-by-layer), we repeated the regression analysis by concatenating the summary statistics of all layers in each CNN architecture. This resulted in two different regressors for each of the two CNNs: the CNN-7 regressor was of dimension 14 (2 parameters for each layer) and the CNN-15 was of dimension 30. Similar to the separate, individual CNN layer regression analyses, this analysis resulted in the measure of model fit (adjusted r2, which is corrected for differences in the predictor dimension of

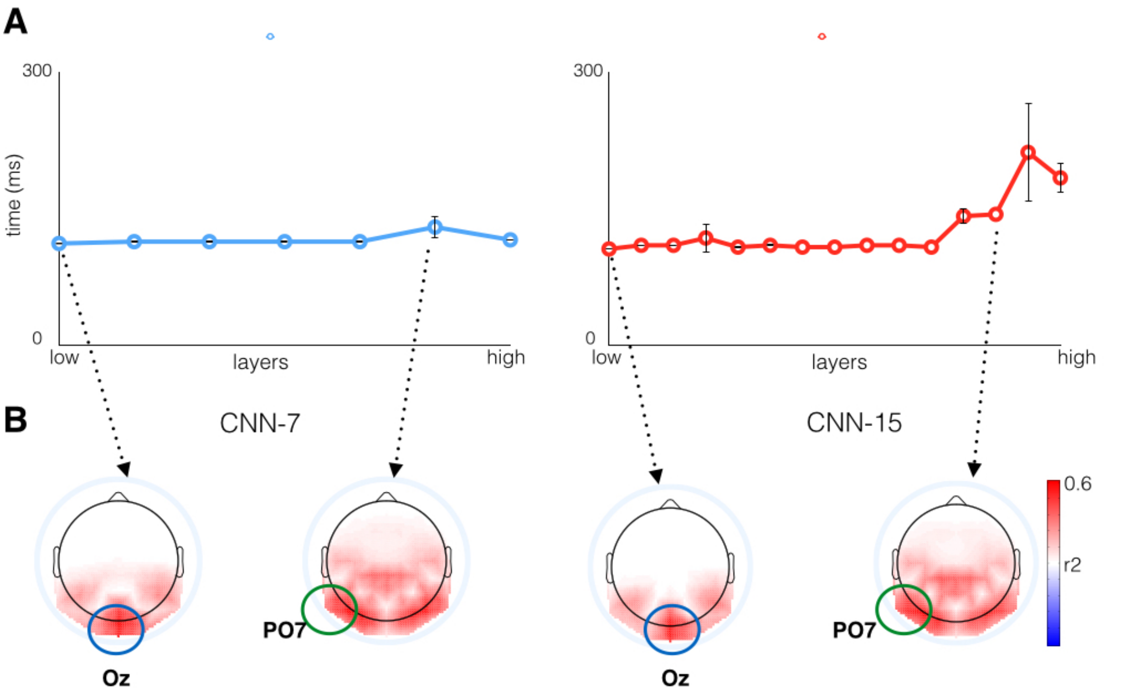


Figure 20: Peak explained variance of CNN layers. A) The time point of peak explained variance by the CNN-7 layers over the entire scalp. Lower CNN layers had maximum explained variance early in time and higher CNN layers later in time. We observed from the scalp visualization that the highest explained variance early in time was found at the occipital channel Oz. Similarly, highest variance later in time was observed at the peri-occipital channel PO7. B) Visualization of the time point of maximum explained variance by the CNN-15 layers over the entire scalp. Similar to CNN-7, lower CNN layers of CNN-15 reached maximum explained variance early in time at Oz and higher CNN layers later in time at PO7.

the two architectures) over time and space for the average subject response. As before, statistical evaluation was done by permutation testing.

Comparison of CNN models

To better understand the correspondence between the two CNN architectures, we correlated the summary statistics of the layers across architectures (architecture correspondence). We also correlated the time series of the explained variance (r2) of all the electrodes (converted to a vector consisting of time points of all electrodes) from each layer of different CNNs (neural correspondence). Specifically, the vector of explained variances at each layer of CNN-7 was correlated to each layer of the CNN-15 (Pearson correlation).

4.3 RESULTS

4.3.1 *Maximal correlation of individual CNN layers*

We first examined the entire time-courses of explained variance of the ERP responses by the CNN architectures. Figure 20A shows the time point of maximal explained variance for the different CNN-7 and CNN-15 layers across all channels. Figure 20B displays the explained variance by both CNN architectures across the entire scalp. We observed that for CNN-7, layers 1-5, the maximum explained variance was found between 110 ms and 120 ms. For CNN-15, layers 1-10 of CNN-15 reached maximal variance between 110 and 120 ms. From Figure 20B we observed that the maximum explained variance early in time was localized to the electrodes overlaying the occipital cortex for CNN-7 (r2 = 0.43 for layer 1) and CNN-15 (r2 = 0.41 for layer 5).

However, for the higher CNN layers (layer 6 for CNN-7 and layer 12-15 for CNN-15), the maximum explained variance was found later in time between 160ms and 210ms. Furthermore, from the scalp visualization plot in Figure 20B the maximal explained variance later in time is observed at channels overlaying lateral-occipital cortex. The whole scalp visualization for layer 1 and layer 6 of CNN-7 (between 100 and 200 ms) is further shown in Figure 25 as part of supplementary information.

Our results show that the hierarchy of CNN representations correlate to the temporal hierarchy of visual processing and generalizes across different architectures. Low- and high-level CNN representations explain variance in visual evoked responses across multiple time-points. Moreover, while ERP responses generally have poor spatial resolution, we observe some spatial specificity in the pattern of results, with higher layers corresponding best to response differences at peri-occipital electrodes overlaying lateral visual areas, whereas lower layers corresponded to response differences at occipital electrodes overlaying early visual regions.

4.3.2 *Time course of CNN layer correlation to ERP*

To examine the spatial specificity of our results, we focused our remaining analyses on two specific electrodes: the occipital channel Oz and peri-occipital channel PO7, as these electrodes exhibited strongest correlation with early and late CNN layers. Note that this is in line with previous studies which indicated sensitivity of summary statistic models to both early occipital (Oz) and late peri-occipital (PO7, PO8) channels [124], [128].

Occipital channel

CNN-7 explained a substantial amount of variance in ERP responses to individual images as shown in Figure 21A. The maximum explained variance was $r2 = 0.46$ at 117 ms for the occipital channel Oz, specifically by CNN layer 1. Explained variance for all layers reached a maximum between 110 and 120 ms after stimulus onset; maximal values for the lower CNN layers ranged between r2 = 0.40 - 0.45 ($p < 1e - 10$ for $t > 100ms$, Bonferroni-corrected for multiple comparisons). The explained variance of EEG responses by CNN layers 6 and 7 were lower, r2 = 0.20 - 0.25 ($p < 1e - 10$ for $t > 100ms$, Bonferroni-corrected for multiple comparisons). The AIC values for the different layers are displayed in Figure 21B. CNN layer 1 had the lowest AIC values for 100 ms to 150 ms, while higher CNN layers had higher AIC values. This result indicates that the lower CNN layers best explained responses at the occipital channel.

In comparison, CNN-15 also explained a substantial amount of variance as shown in Figure 21C. Similar to CNN-7, the explained variance of ERP responses by CNN-15 in the occipital channel Oz was maximal between 110 and 120 ms after stimulus onset. All the lower layers were

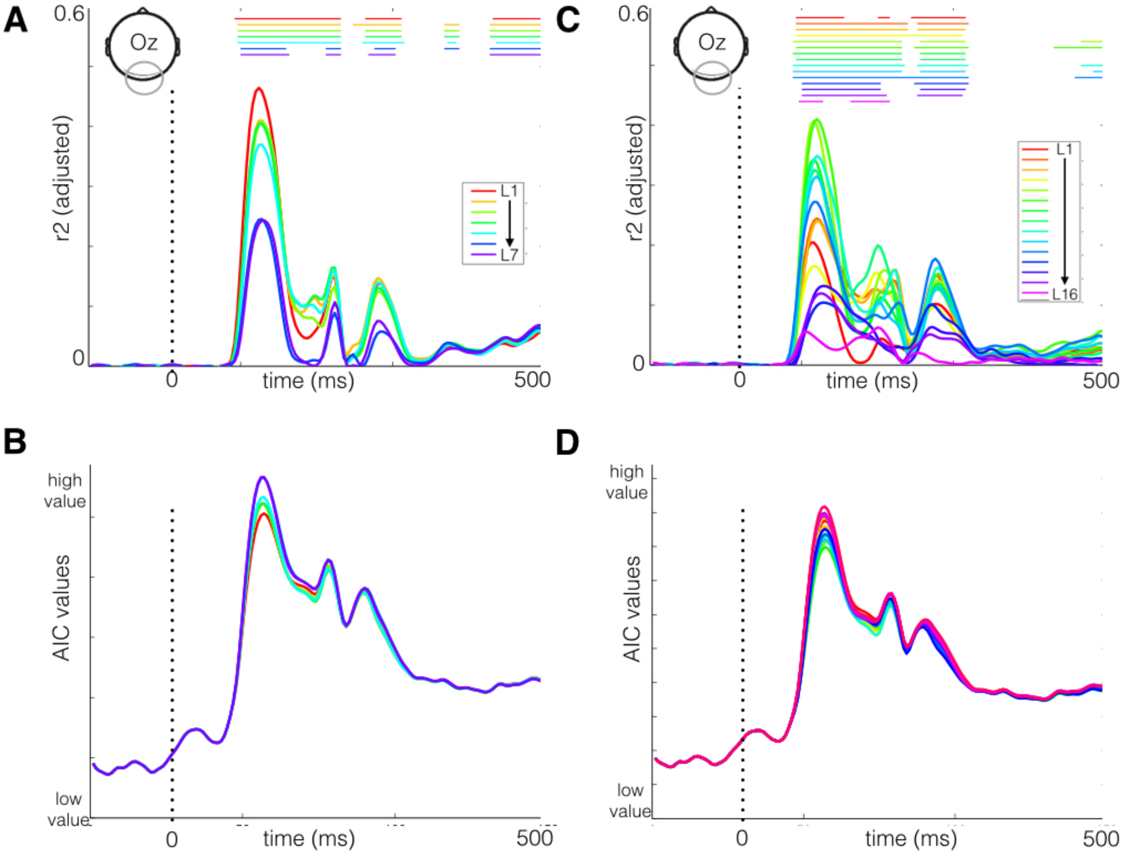


Figure 21: EEG regression results at electrode Oz. A) Explained variance at the occipital channel for the different layers of the CNN-7 architecture. Significant p-values ($p < 1e - 10$ from the permutation analysis) corresponding to each layer are displayed at the top of the graph. All layers correlated with evoked activity starting from 100 ms after stimulus onset. B) AIC values computed from the residuals of each of the CNN layers, showing that layer 1 provided the best fit to the data (low AIC value). C-D) Same as A-B, but for CNN-15. Similarly to CNN-7, lower layers of CNN-15 were maximally correlated to the ERP responses between 100-300 ms, with layer 5 resulting in the best fit at Oz.

significantly correlated ($p < 1e - 10$) between 100 and 200ms. Specifically, layer 5 and 6 from CNN-15 had maximum explained variance r2 = 0.40 - 0.43 ($p < 1e - 10$, Bonferroni-corrected for multiple comparisons) at the occipital channel. The AIC values in Figure 21D show that layer 5 had the lowest value between 100-150 ms, similar to the results for CNN-7.

Together, these results demonstrate that lower layers in both CNN-7 and CNN-15 best modeled evoked activity at the occipital channel with the explained variance reaching a peak between 110 and 120 ms. Moreover, at this channel and time window the correspondence between the evoked activity and the CNN decreased with increasing layers, revealing an (inverse) hierarchical mapping of CNN depth to brain responses early in time at activity recorded on electrodes overlaying low-level visual areas.

Peri-occipital channel

Next, we considered the results for the peri-occipital electrode, which showed a different pattern of results relative to the occipital electrode (Figure 22). Specifically, higher CNN-7 layer 6 had

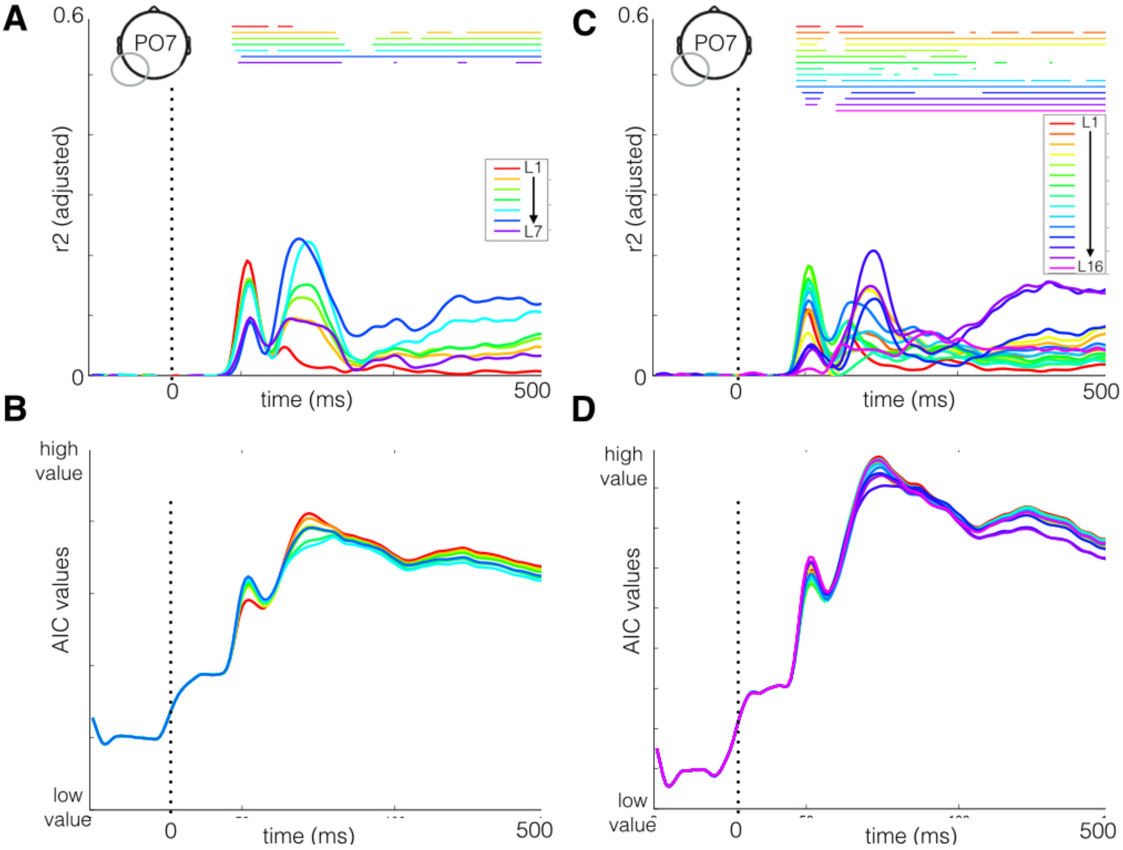


Figure 22: EEG regression results at electrode PO7. A) Explained variance at the peri-occipital channel for the CNN-7 architecture. Lower CNN layer 1 and 2 did not correlate to ERP responses, while layer 3-6 were significant after 150 ms ($p < 1e-10$). B) AIC values for each of the CNN layers. Layer 6 provided the best fit to the ERP responses (low AIC value). C-D) Same as A-B but for CNN-15, showing that layer 14 provides the best fit at the peri-occipital channel.

maximum explained variance later in time ($t > 160ms$ after stimulus onset). Out of all layers, layers 5 and 6 yielded the highest explained variance which ranged between r2 = 0.20 - 0.25 ($p < 1e-10$, Bonferroni-corrected for multiple comparisons) between 160 and 170 ms after stimulus onset as shown in Figure 22A. Critically, lower CNN layers showed no significant correlation with the ERP responses, demonstrating selective sensitivity to higher CNN layers at this electrode. Moreover - in contrast to the occipital channel - the AIC values indicated that higher CNN layers had the lowest values (Figure 22B). These results suggest that at higher-level lateral-occipital electrodes, higher CNN layers provide the best fit to ERP responses specifically later in time.

The results for CNN-15 were similar to those for CNN-7: explained variance reached a maximum between 165 and 180 ms for higher CNN layers as seen in Figure 22C. The maximum explained variance of r2 = 0.21 ($p < 1e-10$, Bonferroni-corrected for multiple comparisons) was found for layer 14. Interestingly, the maximum explained variance of ERP responses by CNN-15 is slightly lower than those obtained for CNN-7. Furthermore, the lower CNN layers were significant ($p < 1e-10$) only early in time (between 100-140ms) with lower explained variance up to r2 = 0.18. Thus, similar to CNN-7 the higher layers of CNN-15 best model ERP responses later in time.

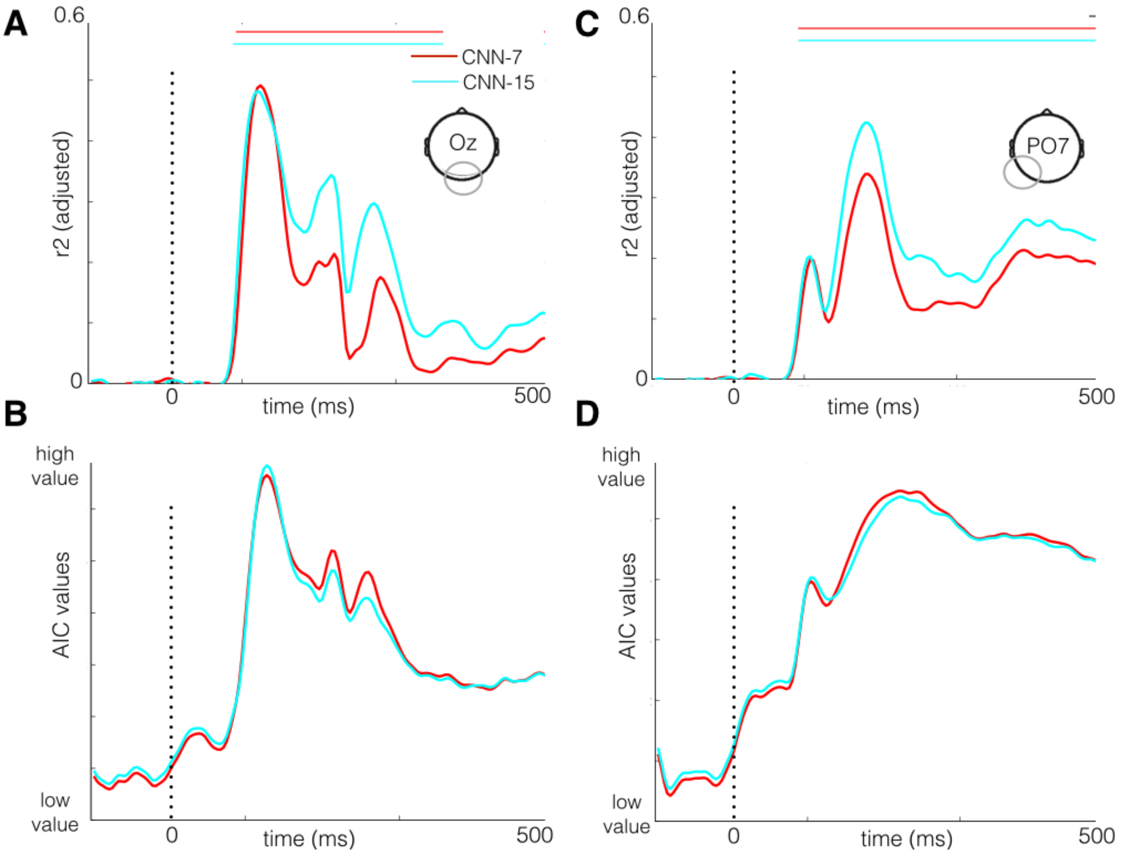


Figure 23: Comparison of CNN-7 and CNN-15 total architectures. Total architectures were obtained by combining the features from the different layers in a single regression model for each CNN. A) Explained variance at the occipital channel for the CNN-7 and 15 architecture ($p < 1e - 10$). B) AIC values for CNN-7 and CNN-15, which shows that the CNN-15 provides a better fit to the ERP values. C-D) Same as A-B but for the peri-occipital channel, showing that CNN-15 provides the better fit compared to CNN-7.The figures show that while there was no difference between the architectures early in time, deeper CNNs provided a better fit later in time.

Overall, our results clearly show that early in time, ERP responses were best explained by lower CNN layers (with maximum correspondence in the occipital channel). In contrast, later in time, ERP responses were best explained by the higher CNN layers (with maximum correspondence at the peri-occipital channel). While the exact distribution of explained variance time course across layers for CNN-7 and CNN-15 showed subtle differences, the gradient of the layers (lower layers to higher in Oz and higher layers to lower in PO7) was similar for both CNN architectures. This suggests that the correlation of CNNs to the temporal hierarchy of visual representations does not fundamentally change with the number of layers. Surprisingly, we did not observe differences in the maximum explained variance between CNN-7 and CNN-15. This suggests that while CNN-15 is more complex than CNN-7 by virtue of more number of layers in the network, the addition of individual layers seems to have little added value for the temporal correlation. In the next section, we further examine potential differences between the CNN architectures by combining the representations at different layers.

4.3.3 *Combined CNN layers*

To investigate to what extent the total CNN architectures capture ERP responses, we concatenated summary statistics of all CNN layers and used these concatenated vectors as regressors. We then obtained the explained variance as before, while correcting for the difference in feature dimensions by using the adjusted r2 values (for example the whole scalp visualization of CNN-15 is shown in Figure 26 as part of supplementary information).

Figure 23A shows the explained variances of ERP responses at the occipital electrode. Before 150 ms, combining layers did not yield a significant change in explained variance by CNN-7 and CNN-15. This is further seen in the AIC values, in which there is no difference between the architectures (Figure 23B). The explained variance reached a maximum of r2 =0.48 ($p < 1e - 10$, Bonferroni-corrected for multiple comparisons), between 100 and 110 ms after stimulus onset. This was observed for both CNN-7 and CNN-15. However, differences between the two CNN architectures emerged later in time (after 150 ms), with the combination of CNN-15 layers explaining more variance of ERP responses. This is also observed in the AIC values in Figure 23B, which show that the deeper CNN provided a better fit to ERP responses compared to CNN-7 between 150 and 400 ms. After 150 ms, the maximum explained variance by CNN-15 is r2 = 0.32 which is significantly higher than the maximal explained variance by CNN-7 (paired t-test, $t(99) = 67.75, p < 0.05, ci = 8 - 10\%$). Thus, differences between CNN-7 and CNN-15 were observed later in time in the occipital channel.

A similar pattern of results was obtained for the peri-occipital channel PO7 (Figure 22C). For CNN-7 we observe a significant increase in explained variance reaching a maximum of r2 = 0.33 ($p < 1e - 10$, Bonferroni-corrected for multiple comparisons) as seen in Figure 22B. The maximum explained variance of the individual layers of CNN-7 reached r2 = 0.21 at $t = 168ms$. This is also observed for the CNN-15 total model, with a significant explained variance of r2 = 0.41 ($p < 1e - 10$, Bonferroni-corrected for feature dimension) compared to the variance by individual layers of r2 = 0.20 ($p < 0.05$). Importantly, the maximum explained variance of the CNN-15 is significantly higher than CNN-7 (paired t-test, $t(99) = 62.58, p < 0.05, ci = 8 - 10\%$). From the AIC values, as shown in Figure 22D, the CNN-15 provides a better fit to the ERP responses compared to CNN-7 after 150 ms. It is worth noting that the increased explained variance is not due to the increase in number of parameters. This can be observed from the baseline before stimulus onset, which remains zero.

In sum, we observed that the combination of CNN layers gave rise to significant differences between the CNN architectures, with deeper CNNs providing a better fit to model brain responses. However, this difference was observed only later in time by combining lower and higher CNN layers in a single model.

4.3.4 *Comparison of CNN architectures*

To better understand what drives the observed difference in performance by the combination of CNN representations, we next performed an analysis that directly compared the two CNNs. We correlated the representations contained in each layer (Figure 24A) as well as their entire spatio-temporal profile of correspondence with evoked activity (Figure 24B; see Methods).

In Figure 24A, we observe a hierarchical mapping of the representations contained in each layer across the two CNN architectures. The lower layers across architectures correlated most strongly ($p < 0.0005$, Bonferroni-corrected for multiple comparisons). These correlations decreased for higher layers across CNN architectures. This suggests that lower layer representations are highly similar across architectures, while higher layers are less similar. This is not surprising, while the lower layers learn similar features (gabor-like) across the architectures, the higher layers of

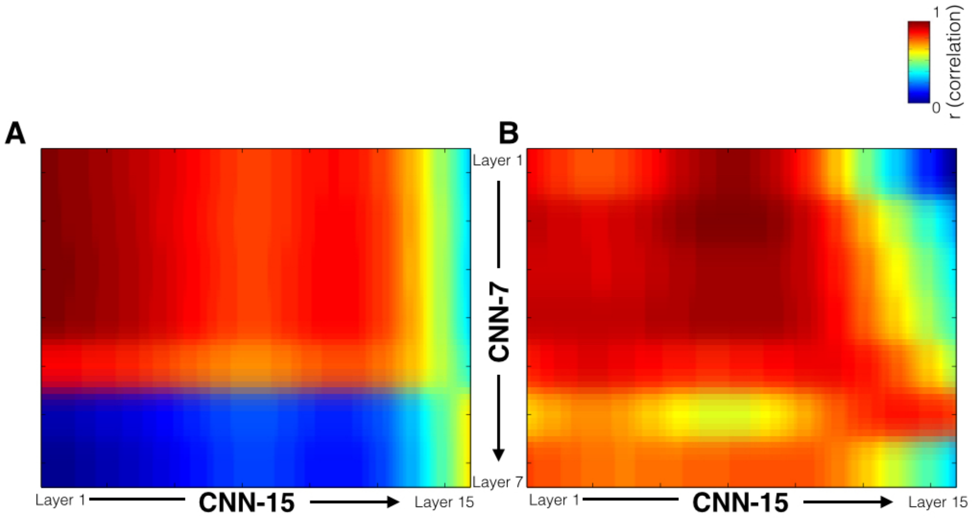


Figure 24: Visualization of the correspondence between the CNN-7 layers with CNN-15 layers. A) Architectural correspondence: Visualization of the correlation (p ¡ 0.0005) between representations of each CNN-7 layer to each of the CNN-15 layer. The lower layers are highly correlated to each other across CNN-7 and CNN-15, while the higher layers are strongly correlated to each other. B) Neural correspondence: The explained variance by each layer is concatenated across all channels into a single vector, which is then correlated to the other layers. A non-linear relationship is observed for the correlation between the lower layers of CNN-7 and CNN-15.

CNN-15 are more discriminative compared to CNN-7. Further, we observe that the convolutional layers and fully connected layers correlate correspondingly across the two architectures.

While the neural correspondence (Figure 24B) also shows a hierarchical mapping, we also observe some differences in contrast to Figure 6A. As before, the highest layers of CNN-15 correlated to the higher CNN-7 layers ($p < 0.0005$), suggesting they captured similar stages of the neural processing. However, the spatiotemporal profile of explained variance by layer 7 of CNN-7 correlated most strongly with lower-and intermediate layers of CNN-15 ($p < 0.0005$, Bonferroni-corrected for multiple comparisons). Additionally, intermediate layers of CNN-15 correlated more strongly to lower layers of CNN-7 ($p < 0.0005$, Bonferroni-corrected for multiple comparisons). This suggests that the expansion of layers in CNN-15 relative to CNN-7 is reflected in low-level layers based on neural data : that is, the same neural response variance captured by low-to-intermediate layers in the CNN-15 is explained by the lowest CNN-7 layers.

In sum, together these results show that while CNN-7 and CNN-15 have similar architectural mapping due to design (i.e. lower CNN layers correlate with other to lower layers, and higher layers with higher layers), this relationship is not directly reflected in the correlation of CNN layers to neural responses, in which we observe a shift from low-to-intermediate layers in terms of the neural information processing captured by the CNN.

4.4 DISCUSSION

The success of Convolutional Neural Networks (CNN) in computer vision has led to a number of demonstrations of a correspondence between CNNs and the brain [66, 103, 104, 112]. Most of these studies have focused on the correlation of CNN layers to neuromaging data using fMRI. However, object recognition is also reflected in time-resolved neural responses. A recent study [113] demonstrated a hierarchical mapping in the temporal domain of CNN layers to whole-brain decoding of visual representations of objects. Our study provides additional quantitative evidence that CNN-models are able to predict spatio-temporal dynamics of visual processing in humans. We confirm the previous finding from [113] that lower CNN-layers correlate to brain responses early in time while the higher CNN-layers correlate to responses later in time. Moreover, we find a subtle spatial shift in terms of these results, with lower layers primarily explaining ERP amplitude differences at occipital channels overlying early visual regions, while higher layers best explain differences at channels overlying higher-level lateral-occipital regions involved in object representations [131]. Thus, similar to the spatial mapping of hierarchical CNN representations to fMRI data from multiple studies, we observe a temporal correspondence of CNN representations to evoked EEG responses.

Our results further demonstrate generalization to deeper CNNs: we find highly similar results with CNNs containing 7 layers or 15 layers. There are no clear differences in maximal explained variance between the different CNNs when comparing individual layers. However, when we combined the different layers of the CNN in a single regression model, CNN-15 did result in a better fit of neural activity compared to CNN-7, but selectively later in time (i.e., beyond 150 ms). In this aspect, the deeper model is more powerful: they capture more information than the shallower CNN. Direct comparison of the representations and spatio-temporal mapping of CNNs to the neural responses suggested that this additional information was mostly reflected as more extensive representations of low-to-intermediate features in CNN-15. Overall, we demonstrate that CNN with larger numbers of layers are better models to explain temporal visual processing than shallow CNN-models at later, but not early stages of visual processing.

What may account for the temporal hierarchy we observed in the CNN correspondence to ERP responses?

We speculate that the close correspondence of CNN-layers to the temporal dynamics of human object recognition might be attributed to the hierarchical nature of representations in the model [109]. The stacking of linear-nonlinear operations in the CNN closely resembles the simple-complex cell model as proposed as Hubel and Wiesel [33]. In the first two convolutional layers of the CNN, oriented gradient features similar to the receptive fields in the V1 and the V2-regions of the brain are learned. As such, the early CNN-layer can be considered to be similar to the local contrast Weibull model that was previously noted to explain ERP amplitude early in time [128]. Higher in the visual hierarchy, the features represented in area V4 and IT become increasingly complex, containing shapes and object-like intermediate features [132]), which are thought to be processed later in time during feedforward visual processing. The complex features such as shapes, contours and even object-like features are captured by higher CNN layers [109].

Given the difference in complexities between CNN architectures, it is surprising that we observed no differences in maximal explained variance of brain responses by the individual layers of the CNNs. Potentially this can be attributed to equivalence of the layers across different CNN architectures that process the same extent of the image. For example, the receptive field size of layer 13 in CNN-15 is the same as layer 5 in CNN-7 [117]. However, increasing the number of layers improves the performance on object recognition for the ImageNet dataset [102], which is attributed to the additional computations (linear and non-linear) in deeper CNNs compared to shallower CNNs. Specifically, the additional non-linear computations increase the discriminative power of deeper CNNs. Consistently, we observed an improved correspondence between deeper CNNs and evoked brain responses when we evaluated the CNN architectures in their entirety.

Interestingly, this increased performance of the combined layers of CNN-15 in terms of explaining ERP responses as compared to the individual CNN layers was observed only later in time. Previous studies show that later in time (150 ms after visual input enters the retina), feedback and recurrent processing influence object-related processing [116]. While core object-related processing is hypothesized to be mediated by feed-forward processing [133], [134] our results therefore highlight the potential importance of recurrent processing in object recognition at later stages of the visual time-course. We speculate that the increased sensitivity might be explained by a recurrent feedback signal to lower brain areas that amplifies neural responses to fine-grained, lower-level information by grouping responses to object features and enhancing them in relation to other responses [135], [136], [116]. In sum, while previous results suggest a correspondence between feed-forward hierarchy of visual representations and CNNs, our results suggest that CNNs might also be suitable to investigate mechanisms of recurrent processing. Specifically, our results suggest that the gain in explanatory power with increased CNN depth is limited when it comes to explaining brain responses within the feed-forward visual sweep (i.e., before 150 ms), but that increased depth of CNN does help better explain neural activity later in time.

In conclusion, we show that 1) the temporal dynamics of representations in the human visual hierarchy are captured by convolutional neural networks (CNN) and 2) deeper CNNs contain expanded representations of low-to-intermediate features which adds substantial predictive power towards explain ERP responses later in visual processing. Going forward, a number of questions remain for CNNs as a model for information processing in the brain. The parameters of CNN models, such as number of layers, layer dimension, type of layers are chosen based on object recognition performance. How these parameters relate to the representations in the human visual system is poorly understood. While in the current study we investigated the number of layers, further work is required to understand the correspondence of the number of CNN layers to the human visual system. Additionally, CNN features themselves are computed in a feed-forward manner and feedback mechanisms are not present in standard CNNs. Clearly, human visual processing is not only rapid and dynamic, but also highly complex, with feedback and recurrent processing playing an important role in visual processing beyond feed-forward information extraction [137]. However, our results suggest that detailed research on CNN architecture and computations in relation to neuroimaging measurements may provide novel insights in the neural mechanisms underlying in visual object recognition in the human brain. Further research is required to disentangle the role of feed-forward and feedback processing in visual processing of objects in relation to the visual features represented by CNN models.

4.5 SUPPLEMENTARY INFORMATION

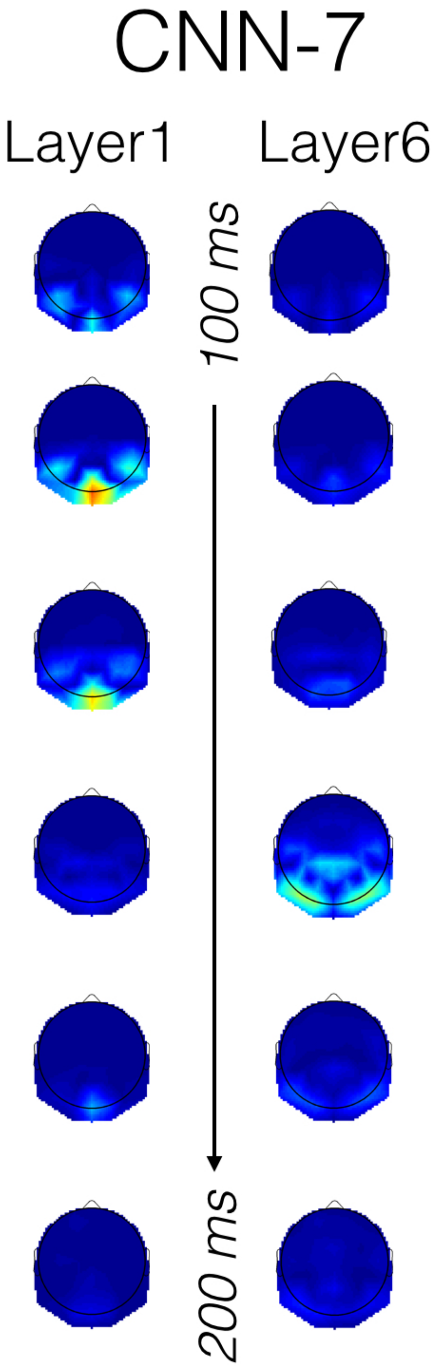


Figure 25: Whole scalp visualization of CNN-7 layer 1 and 6 correlation to the ERP at each channels between 100 and 200 ms (time interval with maximum correlation) .

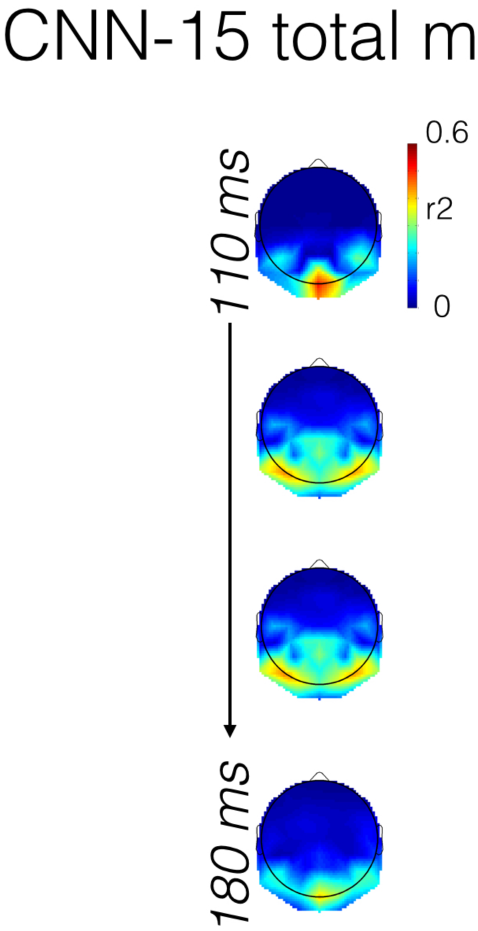


Figure 26: Whole scalp visualization of the total CNN-15 correlation to ERP at each channel between 100 and 180ms (time interval with maximum correlation).

5

MAPPING VISUAL AND SEMANTIC REPRESENTATIONS TO BRAIN RESPONSES

SUMMARY :

The use of computational models combined with large measurements of brain responses to natural stimuli have shed light on the neural mechanisms underlying feedforward visual processing. Feedforward visual processing transform the input image pixels to low-level visual representations (such as edges) upto high-level semantic representations (category membership such as animate versus inanimate). In this article, we investigate to what extent image representations, both visual and semantic, explain similarity in brain responses across a large number of subjects using a limited set of measurements. We use fMRI brain responses to a short video of natural scenes and objects of 100 subjects. We use deep neural network (DNNs) as visual representation of the stimuli. DNNs are state of the art object recognition model in computer vision that correspond to feedforward visual representations in the human brain. Semantic representations consists of four pre-defined classes (such as human versus non-human) to label the stimuli. We first tested whether, there is indeed similarity across subjects measured by principal component analysis. The DNN and semantic representations were then correlated to the principal components extracted from brain responses of the 100 subjects. We further tested DNN representations after regressing out semantic representations to explain similarity in brain responses (and vice-versa). Our results show both visual and semantic representations are significantly correlated to the principal component. Lower DNN layers are highly correlated in early visual brain areas, while higher DNN layers are correlated mostly in higher visual areas. The semantic representation ”mobile versus immobile” correlates the most to principal component. Importantly, DNN representations correlate to the principal component even after regressing out semantic representations of the images such as ”animate versus inanimate” and ”human versus nonhuman”. However, ”mobile versus immobile” significantly explains the similarity in brain responses across subjects even after regressing out visual representations. Overall, our results suggest that feedforward visual representations correlates to across subject similarity in brain responses during natural vision.

5.1 INTRODUCTION

An important problem in neuroscience is to understand how information is organized in the human visual system. The pioneering work of Hubel and Wiesel [33] on the primary visual cortex led to a breakthrough understanding in the organization of the visual system. Beyond primary visual cortex, two separate cortical processing pathways have been proposed: the ventral pathway projecting towards inferotemporal cortex, and the other projecting more dorsally towards posterior parietal areas [138].

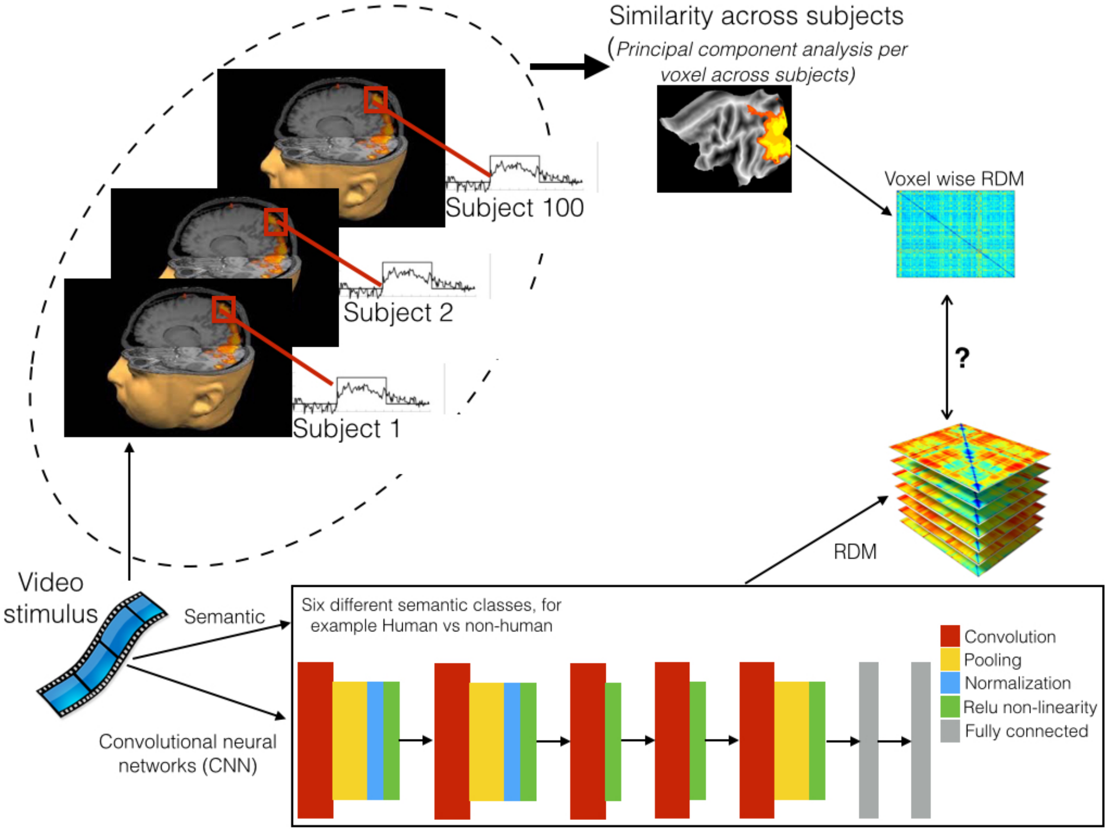

Similarity across subjects
(Principal component analysis per voxel across subjects)
Subject 100
Subject 2
Subject 1
Voxel wise RDM
?
RDM
Video stimulus
Semantic
Convolutional neural networks (CNN)
Six different semantic classes, for example Human vs non-human
Convolution
Pooling
Normalization
Relu non-linearity
Fully connected

Figure 27: **Experimental design and analysis:** *fMRI responses of 100 subjects to a natural video were acquired and correlated to DNN representations of 290 frames from the video. Representation dissimilarity matrices (RDM) are computed by pairwise distance of image representations. The 7 DNN layers yielded a 290 × 290 RDM, with each element indicating the pairwise distance between the 290 frames (RDM). The principal component at each voxel of local fMRI responses (searchlight sphere) across 100 subjects are computed. Then, the RDM at each voxel is the pairwise distance of the principal component of 290 frames. The principal component RDMs are correlated to the DNN layer RDMs. Similarly, the principal component RDM is correlated to four semantic RDMs defined for animate versus in-animate, human versus non-human, mobile versus stationary, civilization versus nature. To determine the unique contributions of visual or semantic representations we employ variation partitioning in 8 brain regions-of-interest (ROI).*

The ventral visual pathway represents visual input hierarchically, starting from low-level features such as oriented edges to shapes, textures and increasingly complex objects [55]. Higher in the visual cortex, large scale divisions along semantic dimensions has been shown (for example, faces [139]). A number of studies have proposed the likely semantic categories that capture significant variance of brain responses. For example, animals versus non-animals is an important category along which information is represented in the brain [140], [77], large versus small [141].

It has been shown that the brain represents categories in a continuous semantic space that reflects category similarity [142] in subjects watching a natural movie.

An important question is if the ventral pathway reflects visual or semantic representations given the close association between visual and semantic features. Recent work [143], [144] has suggested that the organization of category representations in high-level visual cortex reflects brain selectivity for visual features such as spatial frequency and local orientation content. These are low-level features which are typically associated with primary visual cortex. Furthermore, visual shape selectivity and category selectivity have been dissociated to show category selectivity in the ventral pathway after accounting for the contributions of visual features [145], [146], [147], [148], [149]. However, these studies employed controlled stimulus set with a limited number of visual and semantic representations.

One approach to answer whether visual or semantic representations map to brain responses is the use of powerful computational models combined with large measurements of neural activations from natural stimuli [150]. This approach has recently helped decode the computational mechanisms underlying the hierarchy of visual representations in the ventral pathway [151]. For example, a recent computational model, convolutional neural network [100] that has achieved state-of-the-art performance on object recognition, correlate much better to cortical representations than other computational models [103]. They even rival the representation of primate IT cortex for core object recognition [104]. Moreover, there is a strong correlation between the categorization performance of DNNs on challenging high variation object sets and their ability to predict individual neural responses [66]. The DNN architecture resembles the largely feedforward computations in the human brain that culminate in a powerful neuronal representation of objects and scenes.

The success of computational models to investigate hierarchical representations in the ventral visual pathway is partly due to use of natural stimulus as opposed to controlled stimuli. The use of controlled or pre-selected stimulus maximizes variation of specific stimulus representations, however cannot account for representations that fall out of the experimental setup. The use of natural stimuli enables us to use image representations that reflect a wide range of phenomena, for example visual or semantic, and determine the representations that best correlate to brain activity. Moreover, using natural stimuli assists in observing patterns of brain activation that are difficult to observe using simple stimuli. At the same time, to obtain a reliable signal to noise ratio with natural stimuli typically a large number of measurements with multiple repetitions (> 1000) is used which is expensive to collect and sometimes even not possible to obtain. Our goal is to conduct a similar computational analysis over a large number of subjects with a limited stimulus set (a typical run time per subject). This is particularly useful for studies of individual differences for which it is often not feasible to obtain massive recordings from each subject, such as those related to differences in attentional mechanisms or clinical populations.

In this chapter, we relate visual and semantic representations to BOLD-MRI data using a multitude of subjects (N=100) with only a limited amount of data per subject. Deep neural networks are used to compute the visual representations from the stimuli. We pre-define four semantic categories to represent the visual stimuli. We disentangle the relative contributions of deep neural network and pre-defined semantic representations to visual cortex representations. Functional MRI (fMRI) activity was recorded for 100 subjects during free viewing of a natural video. For an analysis over a large number of subjects we take an alternative methodology to the traditional single subject searchlight analysis, instead we use principal component analysis across subjects. The principal component of subject responses reflect the common representations across subjects. We compare the extent to which visual representations from deep neural networks and semantic representations explain brain responses across a large number of subjects.

We first compute similarity of neural responses to a short movie across a large number of subjects (N = 100) using principal component analysis and retain the first principal component. To compare brain responses to DNN and semantic representations, we compute dissimilarity matrices via representation dissimilarity analysis(RSA). The dissimilarity matrices of DNN and semantic representations are correlated to the dissimilarity matrix from the principal component of brain responses. Finally, we perform variation partitioning in order to disentangle the visual and semantic representations in mapping brain areas.

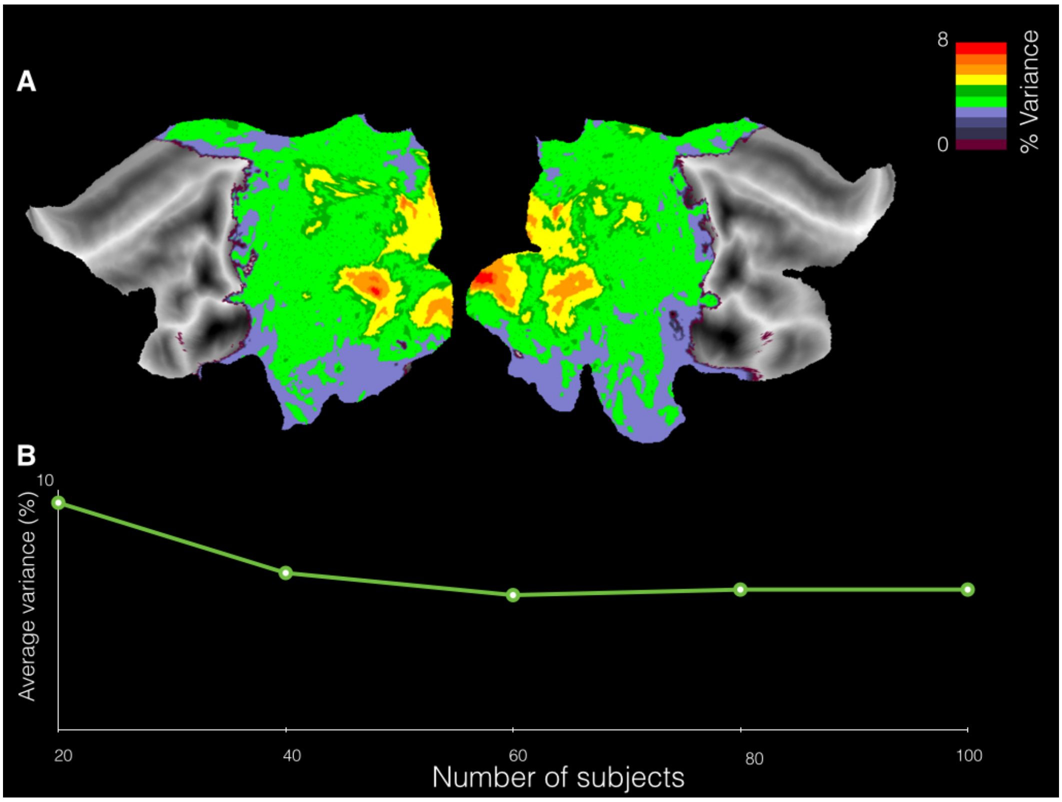


Figure 28: **Across subject similarity in brain responses:** *Visualization of the principal component of brain responses common across subjects. A) First principal component of the searchlight sphere RDMs across the 100 subjects computed voxel-wise. B) The variance in data explained by first principal component as a function of number of subjects. The average variance converges after 40 subjects.*

5.2 MATERIALS AND METHODS

Subjects

The fMRI data of the video stimuli was collected for over 1000 subjects, from which 100 were randomly sampled for this study. Subjects were not assigned with any specific tasks and watched the video track passively one time each. The experiment was approved by the ethical committee of the University of Amsterdam and all participants gave written informed consent prior to participation. They were rewarded for participation with either study credits or financial compensation.

Stimuli

An 11-minute video track consisting of about 20 different dynamic scenes was used for this study. The scenes were taken from the movie Koyaanisqatsi: Life Out of Balance and consisted primarily of slow motion and time-lapse footage of cities and many natural landscapes across the United States. We used dynamic scenes instead of static scenes because they are more realistic, and because they may evoke brain responses that allow for a better acquisition of neural processes in the visual areas of the brain [74]. Scenes include beach, fields, rocks, faces, crowds, buildings, cars, planes captured under different conditions, such as luminance, scale (zoom), motion (moving camera or moving objects on the foreground). Images within one particular scene refer mostly to the same location and/or background under different conditions, such as luminance, scale (zoom), motion (moving camera or moving objects on the foreground), etc. Overall, the images varied from natural scenes such as beach, fields, rocks etc to man-made scenes such as faces, crowds, buildings, cars, planes etc such that it covered a wide range on the frequency spectrum.

fMRI

We recorded 290 volumes of BOLD-MRI (GE-EPI, $192^2 mm$, 42 slices, voxel size of $3 \times 3 \times 3.3$, TR 2200 ms, TE 27.63 ms, SENSE 2, FA 90°) using a 3T Philips Achieve scanner with a 32 channel headcoil. A high-resolution T1-weighted image (TR, 8.141 ms; TE, 3.74 ms; FOV, $256 \times 256 \times 160$ mm) was collected for registration purposes. Stimuli were backward-projected onto a screen that was viewed through a mirror attached to the head-coil. Subjects were placed supinely inside the scanner and watched the movie via back-projection. The movie was presented using a DLP beamer (120 Hz). The total distance from the subject's eyes to the screen was 156 cm with a resolution of 1920x1080. The movie extended 16 degrees in horizontal direction.

Subjects where not asked to fixate although action typically was in the center of the screen. Differences in subject fixation targets could have resulted in an increase of variance between subjects, in particular for lower visual areas V1 and V2. However, given that the explained variance here is within range of that published in literature eye movements (or differences in eye movements between subjects) of subjects cannot have been a large factor in introducing variance in the data. The degree to which this did influence our results should be proportionate for all brain areas, or, at worse stronger for lower than higher-tier areas.

In total, during the experiment 290 volumes were obtained. Taking into account the haemodynamic delay of 6 seconds we identified the 290 image frames corresponding to the 290 volumes. The BOLD values were used at those 290 volumes.

Data representation

We use the outputs of the 7 different layers from the DNN to represent images. We denote the output of the DNN at the different layers as Ln, where $n = 1, 2..7$. A representational dissimilarity matrix [77] (RDM) F is computed separately for each of the representations at different layers. The elements in this matrix are the Euclidean distance between the representations of pairs of images. Thus F_{Ln}, where $n = 1, 2, ..7$, are dissimilarity matrices for the different DNN layers.

We used 4 semantic category : animate versus inanimate, human versus nonhuman, mobile versus immobile and civilization versus nature. These categories are based on previous publications and studies. For the animate versus inanimate category, we assigned positive weights to nonhuman animals, people, and body parts and zero weight to all other categories. For the human versus nonhuman category, we assigned positive weights to people and zero weights to all other categories. For the mobile versus immobile category, we assigned positive weights to mobile categories such as animals, people, and vehicles, and zero weight to all other categories. For the civilization versus nature category, we assigned positive weights to people, man-made objects (e.g., buildings, vehicles, and tools), and communication verbs and negative weights to nonhuman animals. We label the image with the semantic category that is most dominant.

We used additionally two baseline dimensions : object category of the images and contrast energy of the image. For the object category dimension we label each image as per the image category. For the contrast energy dimension, we compute the weibull summary parameters (contrast energy) [152] and assign positive weights to low contrast energy and negative weights to high contrast energy.

For the fMRI representations, a $3 \times 3 \times 3$ searchlight cube is centered at each voxel in the brain and the principal component is computed for the BOLD responses within the cube to each of the 290 still images over 100 subjects. The RDM is computed for each principal component of dimension 290x1. This results for each voxel in a 290×290 dissimilarity matrix Y. Each element in the Y matrix is the pairwise distance of the 290 elements of the principal component. As a distance measure Cityblock is taken.

Preprocessing

FEAT (fMRI Expert Analysis Tool) version 5.0, part of FSL [80] was used to analyze the fMRI data. Preprocessing steps included slice-time correction, motion correction, high-pass filtering in the temporal domain ($\sigma = 100s$), spatially filtered with a FWHM of 5 mm and prewhitened [81]. Data was transformed using an ICA and we subsequently, automatically identified artefacts using the FIX algorithm [82]. Structural images were coregistered to the functional images and transformed to MNI standard space (Montreal Neurological Institute) using FLIRT (FMRIB's Linear Image Registration Tool; FSL). The resulting normalization parameters were applied to the functional images. The data was transformed into standard space for cross-participant analyses, so that the same voxels and features were used across subjects. Although in this approach the haemodynamic response might be influenced by other image frames, we expect this influence to be limited because the video is slowly changing without any abrupt variations. In addition, BOLD responses are intrinsically slow and develop over a period of up to 20 seconds. Still they summate linearly reasonably well [83] and also match the timecourse in typical scenes which develop over multiple seconds. This also probably explains the power of BOLD-MRI in decoding the content of movies [84] and indicates it is possible to compare different models of information processing on the basis of MRI volumes.

Variation partitioning

We are dealing with two or explanatory matrices (in our case DNN and semantic representations, i.e. matrix representations of the visual stimulus) to explain brain responses. If the explanatory matrices are correlated, then the variation explained by the first or the other variable should be separated in order to attribute explanation to one, the other or both. One technique to approach this problem is variation partitioning, a technique routinely used for analysis of ecological data [132]. This technique reports the total variation in the response data (in our case fMRI) as the total sum of squared differences from the mean of each variable. Partial correlation coefficients may be extracted from a variation partitioning analysis and tested for significance. These coefficients expresses the proportion of variation exclusively explained by one explanatory matrix relative to the sum of the unexplained variation and the variation of interest. It controls for the influence of any other explanatory matrices as well as their overlap with the explanatory matrix of interest.

5.2.1 *Analysis*

For the visual representations, we used AlexNet pre-trained on ImageNet to compute representations of key video frames (DNN representations) and subsequently represented it as dissimilarity matrices (RDMs) [77]. The representations are the output at the pooling stage for Layers 1, 2 and 5, the output of rectification unit for convolutional layers 3, 4 and the fully connected layers. For the semantic representations we use four categories - animate versus inanimate, human versus non-human, mobile versus immobile and civilization versus nature (the two baseline dimensions are also used). We labeled each image as per the semantic category and then corresponding semantic RDM is formed by pair-wise distance of the labelled images.

Functional MRI (fMRI) activity was recorded for 100 subjects while they watched the video. After preprocessing and normalizing the fMRI data to standard space, we computed the across subject similarity using principal component analysis [153]. A 27x100 dimension vector is created per image, using the search light sphere (3x3x3mm) centered at each voxel across the 100 subjects. For all images we obtain a matrix of 290x2700. A PCA on the 290x2700 matrix and retaining the first PC, gives a vector of dimension 290x1. The auto-product of this vector gives us a 290x290 RDM [77] for that particular voxel.

The across subject similarity RDM is correlated to both semantic and DNN RDMs using the pearson measure. Additionally, we defined 8 different functionally specialized region-of-interest (ROIs) [111] in the brain, to understand how visual and semantic RDMs are correlated to different brain areas. The ROIs are V123, V4, parahippocampal gyrus (PHC), medial superior temporal area (MST), middle temporal (MT), intraparietal sulcus (IPS) and superior parietal lobule (SPL). The principal component is computed using all the voxels in the ROI concatenated across 100 subjects.

To disentangle the visual and semantic representations in the different ROIs, we regressed out the individual correlations using variation partitioning [132]. A permutation test (1000 times) determines the statistical significance (p value) of the fractions that we obtain for each ROi by variation partitioning. Thus, per ROI we obtained the unique, shared variance of DNN and semantic representations to explain the across subject similarity in brain responses during natural vision.

5.3 RESULTS

Inter subject similarity

We computed the similarity of brain responses across 100 subjects using principal component analysis. Figure 28 shows the first principal component of brain responses across all 100 subjects. The variance explained by the first principal component at each voxel is visualized on a flattened brain map. As seen in Figure 28, the first principal component explains upto 8% of variance of brain responses, mainly in visual cortex both early and higher visual areas. The explained variance of other principal components progressively decrease. The 2nd component explains upto 6% of variance, the 3rd component explains 4% of variance (Figure 36 and 37 in supplementary information). As with the first principle component, both these principal components explain most variance primarily in the visual areas of the brain.

To test to what extent the number of subjects (N=100) is sufficient to get reliable results across subjects, we repeated the same analysis with an increasing number (steps of 20) of random subjects We observe in Figure 28B, that the maximum variance captured converges to 8% with a group size of 40 subjects.This suggests that the number of subjects in our study is sufficient for reliable results. This might also serve as an useful measure in future neuro-imaging studies to make robust conclusions.

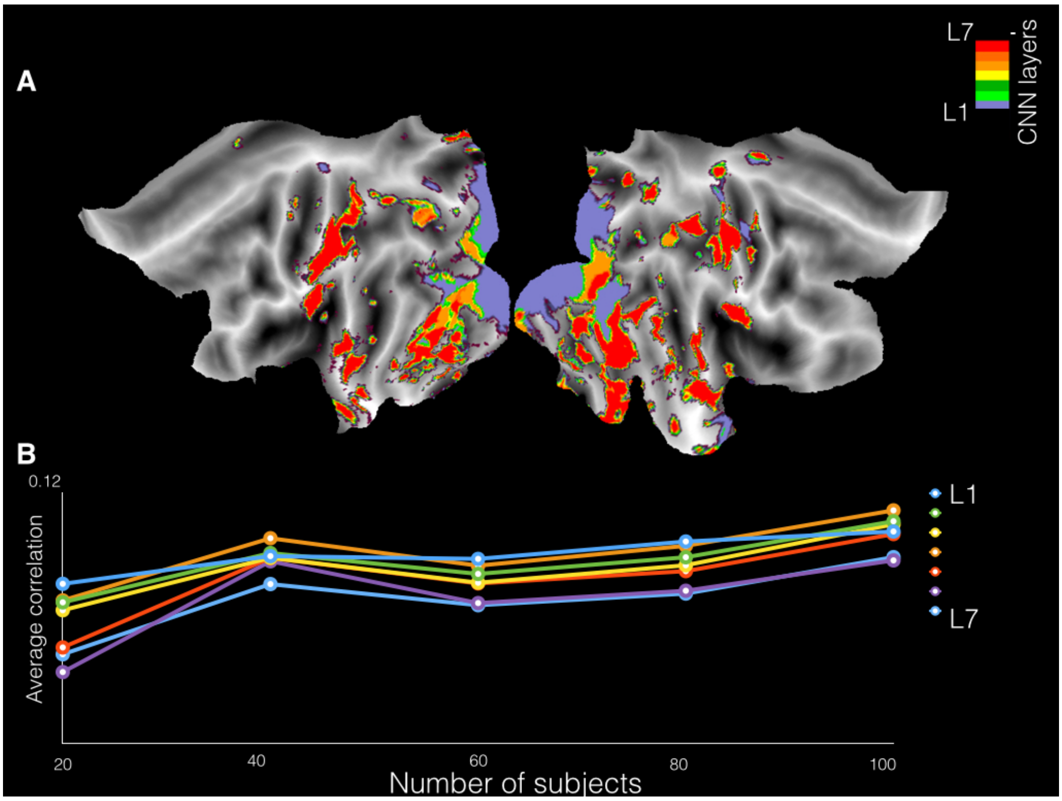


Figure 29: **Correlation of across subject similarity to visual representations:** *Results of DNN correlation (r) to across subject similarity of brain responses. A) Visualization of highly significant voxels (* $r > 0.15, p < 0.05$ *) as revealed by correlation of different DNN layers to the first principal component as obtained from Fig 1A. B) The average correlation of significant voxels of each DNN layer to the principal component with increasing number of subjects.*

5.3.1 *Visual and semantic representations*

Next, we analyzed to what extent do DNN and semantic representations correlate to the similarity of brain responses across subjects. Figure 29 shows the correspondence of the DNN representations (RDM constructed from the DNN layer outputs) to the first principal component RDM at every voxel. Per voxel we visualized the DNN layer that correlates most ($r > 0.1, p < 0.05$) to the first principal component on a flattened brain map.

We observe that different DNN layers correlate selectively to different brain areas. Early visual areas are correlated to lower DNN layers, while higher DNN layers are mostly correlated to higher visual areas in the brain. This is in line with results from literature that map DNN layers to the hierarchy of visual representations in the human brain [112].

As before, we studied the effect of group size on the stability of correlations between DNN representations and brain responses. It might be the case that different DNN layers require different number of subjects for results to converge. We observe that with increasing number of subjects (in

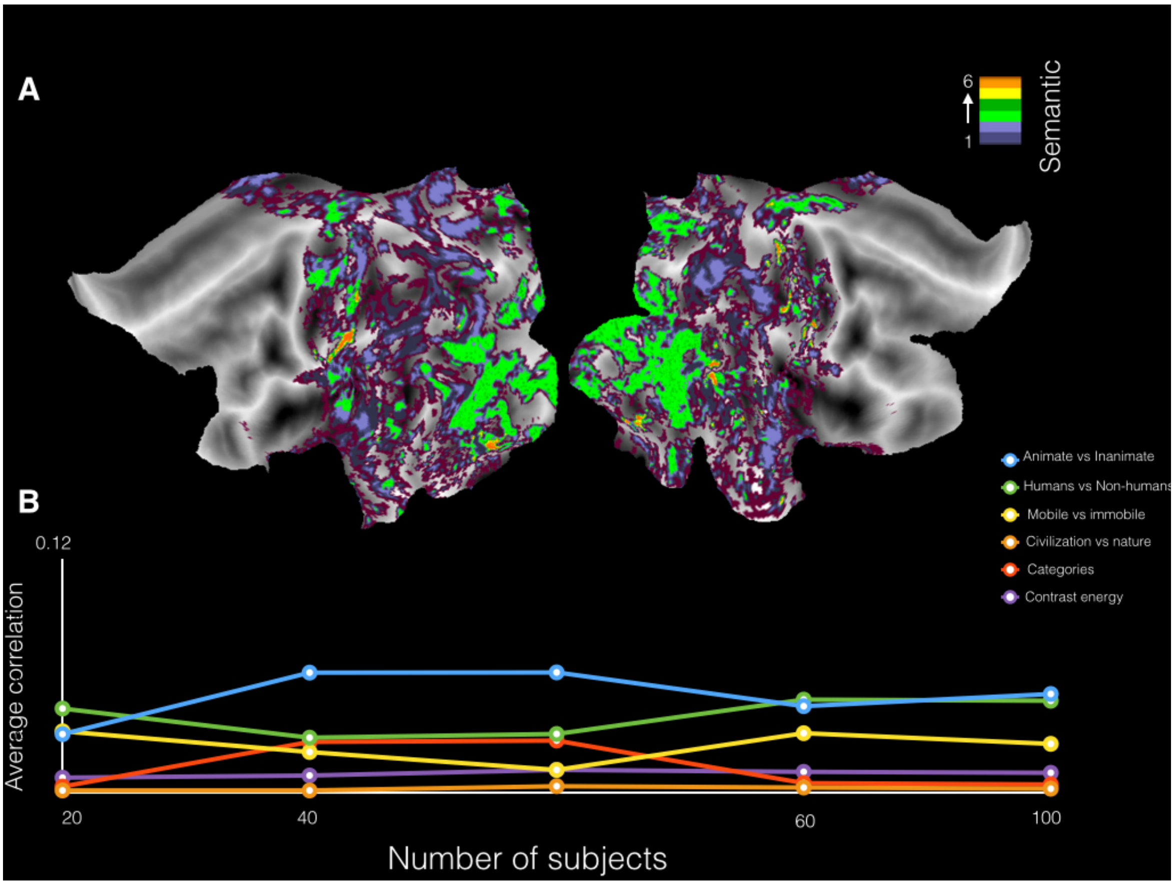


Figure 30: **Correlation of across subject similarity to semantic representations:** *Results of correlation (r) between the four semantic RDMs ((1)animate versus inanimate, (2)human versus nonhuman, (3)mobile versus immobile and (4)civilization versus nature), two baseline dimensions ((5) Object categories and (6) contrast energy) and across subject similarity of brain responses. A) Visualization of active voxel ($p < 0.05$) as revealed by correlation of different semantic classes to the first principal component as obtained from Fig 1A. B) The average correlation of semantic RDMs to the principal component with increasing number of subjects.*

steps of 20), the average correlation of the principal component to DNN representations increase in Figure 29B. The average correlation for lower DNN layers are generally higher than higher DNN layers. However after 40 subjects, the correlation converges for all the DNN layers and does not change with increasing number of subjects.

The correspondence of the different semantic representations, including the baseline dimensions, to the first principal component is visualized on the flattened brain map depicted in Figure 30 ($r > 0.1, p < 0.05$) . It can be seen that the concept ”mobile versus immobile” is most correlated in the early visual areas and also in intermediate visual areas. However higher in the visual hierarchy, we observe that concepts such as animate versus inanimate and human versus non-human correlate the most. Other semantic classes such as object category membership is correlated in very few voxels across the brain. We observe that the object category and contrast energy dimension do not capture any significant variation in brain response across subjects.

Overall, our results show that visual and semantic representations correlate to the first principal component in widespread areas along the visual hierarchy. Specifically, lower DNN representations drive similarity in the early visual areas while the higher DNN layers drive similarity in the higher visual areas. Out of the four semantic representations, the mobile versus immobile category are widely correlated in different brain areas. To further localize the brain areas where visual and semantic representations explain brain responses, we show results of the analysis in different regions of interest.

5.3.2 *Region of interest analysis*

The voxel-wise analysis does not show how different functional areas of the brain represent different principal components. To understand in which brain areas DNN representations are sensitive, principal component analysis is done on the entire ROI. For our analysis, we considered 8 different brain areas. We plot the correlation in different ROIs to the visual and semantic representations. The first three principal components are correlated to DNN and semantic representations to understand, if different principal component capture different dimensions of the representation. To understand if DNN representations and the different semantic representations share information, we correlate DNN RDMs to the semantic RDMs. The baseline dimensions are not included in the ROI analysis since they do not capture significant variance in brain responses as seen in the previous section.

DNN representations

Figure 31A shows the correlation of each DNN layer to the first region-specific principal component. In lower visual area V123 ($r = 0.32$), and also in other areas such as hMT ($r = 0.23$) and V4 ($r = 0.33$), lower DNN layers significantly correlate to brain responses. In other visual areas such as PHC ($r = 0.31$), IPS and SPL, higher DNN layers seem to yield slightly higher correlations.

Figure 31B and 31C show the correlation of DNN layers to the second and third region principal component. We observe that the second principal component in the different ROIs are explained by all the DNN layers except for the PHC, where the highest correlation is found for the intermediate DNN layer 4. Similarly the third principal component in areas hMT and SPL is best explained by higher DNN layer 5.

Overall, the region-of-interest analysis corresponds and confirms the voxel-wise analysis described in the previous section. Both low-level and high-level visual representations from the DNN explain across subject similarity in brain responses to visual stimuli.

Semantic representations

To understand in which brain areas semantic representations are sensitive, we correlated the semantic RDMs to the principal components of brain responses across subjects.

Figure 32A shows the correlation of the first principal component to the different semantic RDMs ($p < 0.05$) . The first principal component is most correlated to the mobile versus immobile semantic RDM. It is most correlated to area V4 ($r = 0.27$), and also to other areas higher in the visual hierarchy such as PHC ($r = 0.25$) and LO ($r = 0.22$). The first principal component is also significantly correlated to the animate versus inanimate in IPS ($r = 0.15$) and human versus non-human in SPL (r = 0.16). Overall, the mobile versus immobile is the semantic category that is correlated the most to brain responses across subjects during natural vision.

The correlation of semantic RDMs to the second and third principal component is shown in Figure 32B and 32C. For both these principal components, we observe that the other semantic categories have the highest correlation. In areas PHC and IPS the animate vs inanimate semantic category has the highest correlation for both the principal components. Human vs non-human is the highest correlated semantic RDM in hMT and MST for both the principal components.

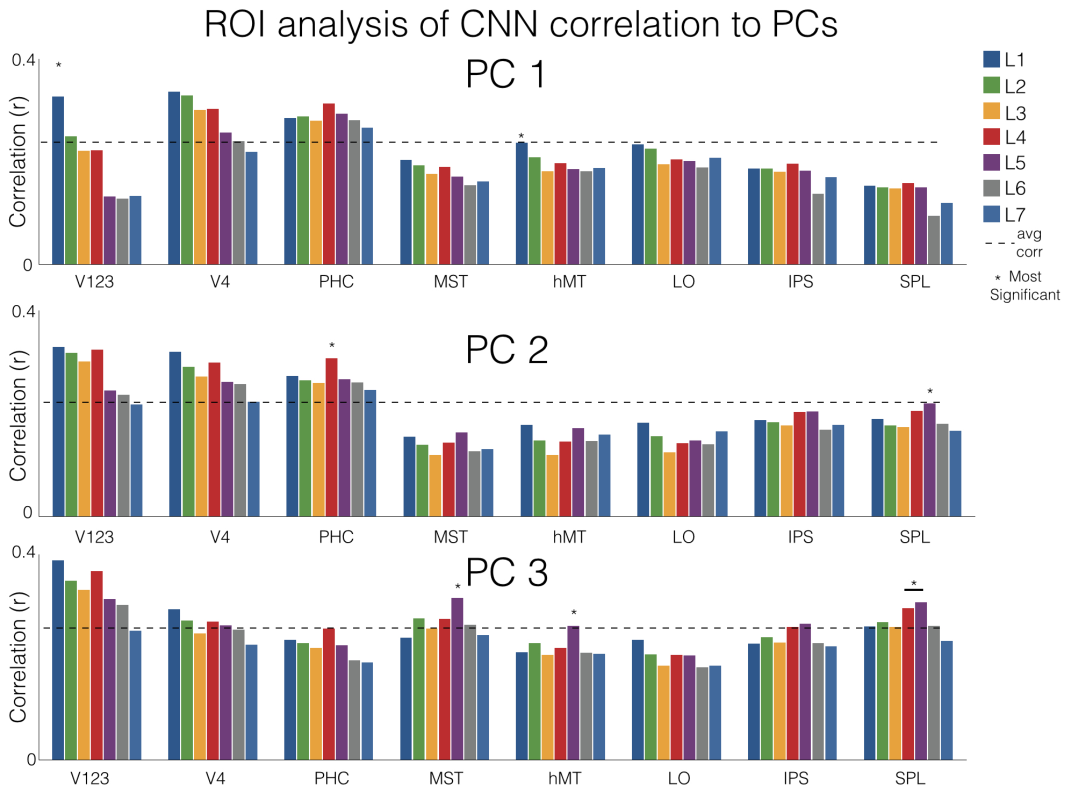


Figure 31: **Region of interest analysis of DNN representations:** *DNN representations are correlated to the first three principal component in the 8 visual ROIs (∗$p < 0.05$, corresponding to the highest significant correlation) . For the first principal component, we observe that lower visual area V123 is best correlated to DNN layer 1, and gradually for the higher visual areas such as PHC, DNN layer 4 is significantly correlated. For the second and third principal component, we observe that in higher visual areas DNN layers 4 or 5 is best correlated to across subject similarity in brain responses.*

We observe that different semantic categories are correlated to the principal components, suggesting that the the principal components being orthogonal in nature might capture different aspects of the semantic organization in the visual cortex.

Correlation of DNN representations to semantic categories

The correlation of DNN representations and semantic RDMs to the principal components suggest that these representations covary. To quantify the extent of correspondence, we visualize the correlation between DNN RDMs and semantic RDMs.

In Figure 33 we show the correlation of the DNN RDMs to the different semantic RDMs. We observe that different semantic RDMs are indeed correlated to different DNN layers ($p < 0.05$) . For the animate versus inanimate category, we observe that it is correlated most to DNN layer 7 ($r = 0.31$). Similarly for ”civilization versus nature” and object category RDMs, we observe that the highest DNN layer is most correlated ($r = 0.32$). However for the concept ”human versus non-human” we observe that even DNN layers such as layer 5 yield highest correlations ($r = 0.45$). Furthermore for ”mobile versus immobile”, we observe that layers 2-6 are most correlated ($r = 0.45 - 0.53$). Thus the mobile versus immobile representation contains information that is common with low to high level visual representations.

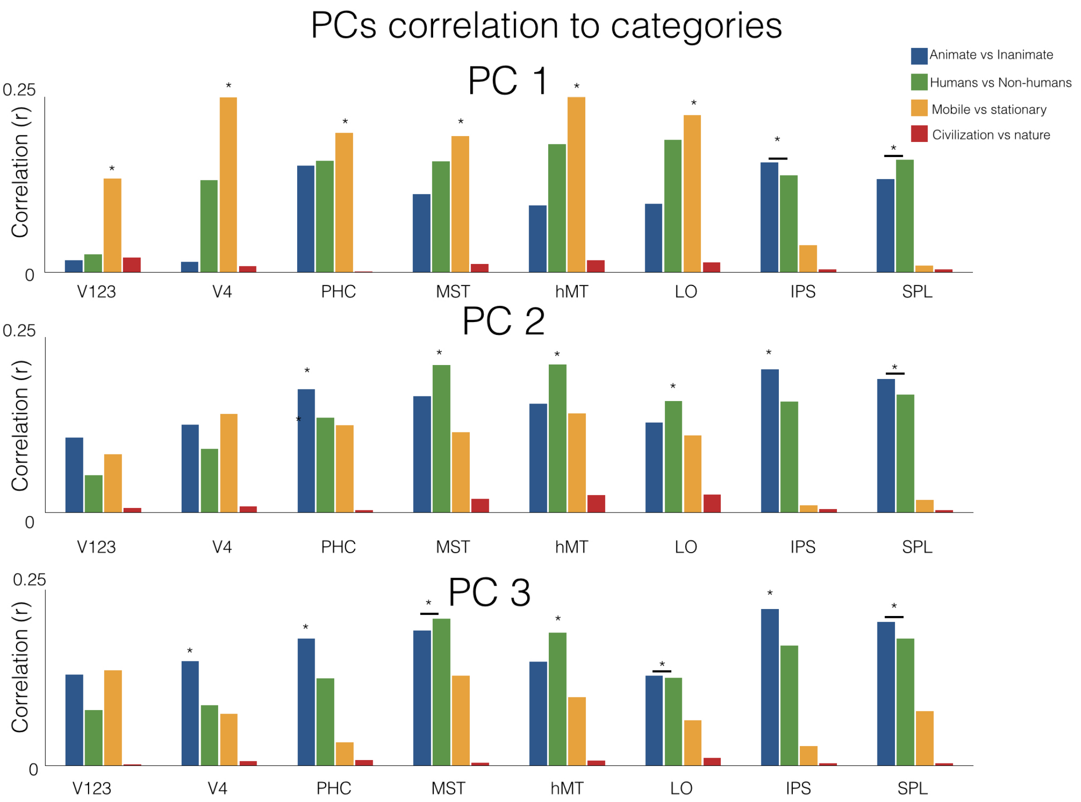


Figure 32: **Region of interest analysis of semantic representations:** *Semantic RDMs is correlated to the first three principal component in 8 different ROIs. We observe that for the first principal component, the mobile versus immobile has the highest correlation ($*p < 0.05$, the most significant) in all the different ROIs except for IPS and SPL. For the second and third principal component however, the animate vs inanimate and human vs nonhuman semantic representation is significantly correlated in different ROIs.*

Thus the semantic categories co-vary with the DNN representations. This suggests that the correlation of the two representations to brain responses might include a component of shared variance. To determine the unique variation in brain responses that can attributed to either DNN or semantic representations we do a variation partitioning analysis.

5.3.3 *Visual versus semantic representations*

We tested to what extent either visual or semantic representations correlate to similarity in brain responses, after regressing out semantic or visual representations respectively by variation partitioning [132] ($p < 0.005$) . This was done separately for each DNN layer in each brain ROI. We visualize the explained variance ($r2$) of the principal component by mobile versus immobile semantic representations in V123, PHC and IPS. This is so, since the motion versus immobile representation has the highest correlations to the first principal component as seen in Figure 32A. We further test to what extent a combination of visual and semantic representations explain across subject similarity.

Figure 34 shows the explained variance of brain responses by the combination of DNN and semantic RDMs (shown by averaging across the 8 ROIs). We observe that the DNN RDM combine with ”mobile versus immobile” RDM to explain the highest variation of the first principal component ($r2 = 5\%$, $p < 0.005$) . This suggests that the first principal component

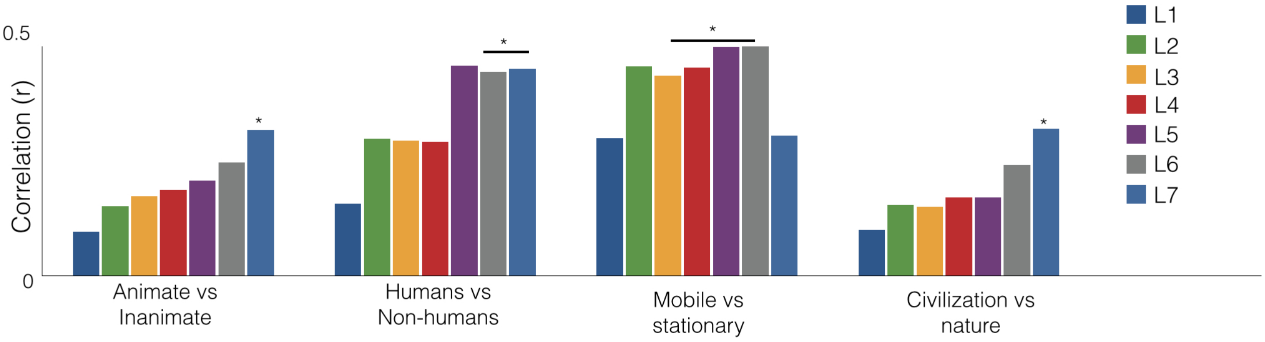


Figure 33: **Correlation of DNN and semantic representations:** *Semantic RDMs correlation to the representations of different DNN layers* ($*p < 0.05$) *. We observe that DNN and semantic are significantly correlated and covary with each other. For all the semantic categories only the higher DNN representations are significantly correlated. For animate vs inanimate and civilization vs nature, the DNN layer 7 is significantly correlated while for human vs non-human layers 5-7 are correlated. For mobile vs immobile semantic representation, all the DNN layers are correlated.*

primarily captures the semantic category of mobile versus immobile. While DNN representations contain complementary information to semantic representation, both the representations have significant shared variance. This suggests that visual and semantic representations partly share information.

Figure 35 shows the results of unique variance of different DNN layers and the semantic representations in the 3 different ROIs of the brain. For mobile versus immobile representation, we observe that in V123 and IPS the visual representations explain significant variance. However, in areas such as PHC the semantic representation explains significant variance upto 7.5% ($r2$). Regressing out other semantic representations such as the human versus non-human, we observe that visual representations explain significant variance of the principal component while the semantic representation explain very less variance, 1-2%. Thus low level visual area V123 is primarily explained by visual representations and higher areas such as PHC explained mainly by semantic representation, and certain areas are explained by a combination of visual and semantic representations.

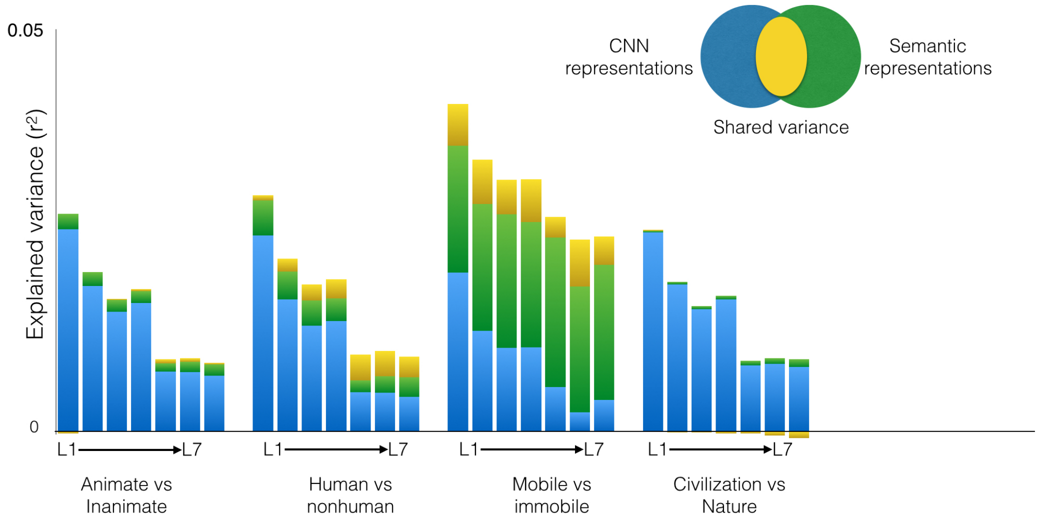


Figure 34: **Combination of DNN-visual and semantic representations:** *Variation explained by the combination of DNN representations and semantic RDMs to the first principal component of each ROI ($p < 0.05$). This is visualized for each semantic RDM for the DNN layers, averaged over 8 different ROIs in the human brain. We observe that the mobile versus immobile explains the highest variance in combination with the DNN visual representations.*

Overall our results show that there is correlation between DNN and semantic representations, but both DNN and semantic representations capture unique information in the brain responses across subjects. Nevertheless, care has to be taken in attributing the DNN-brain mapping solely to the individual DNN layers because semantic representations also explain part of the inter subject similarity in brain responses to visual stimuli.

5.4 DISCUSSION

In our study, we observed that DNN visual representations correlate to the across-subject similarity in different visual areas such as early visual areas V123, V4 and also higher visual areas such as LO and PHC. The correlation of semantic representations to the principal component shows that, concepts like ”mobile versus immobile” is dominant in V123 as well as V4, PHC and LO. In higher brain areas, like IPS, the semantic feature pair ”animate versus inanimate” is dominant. And, in SPL, the feature pair ”human versus non-human” correlates significantly to the main principal component. Accounting for the correlation between DNN and semantic representations, we find that DNN-visual representations uniquely explain up to 5% and semantic representations explain upto 7.5% of the variance of the principal component. The combination of the motion concept and DNN-visual representations explain up to 12% of variance in the similarity of brain responses across subjects. Thus for our stimulus set, semantic and DNN-visual representations are independent and complementary to explain across-subject similarity in brain responses.

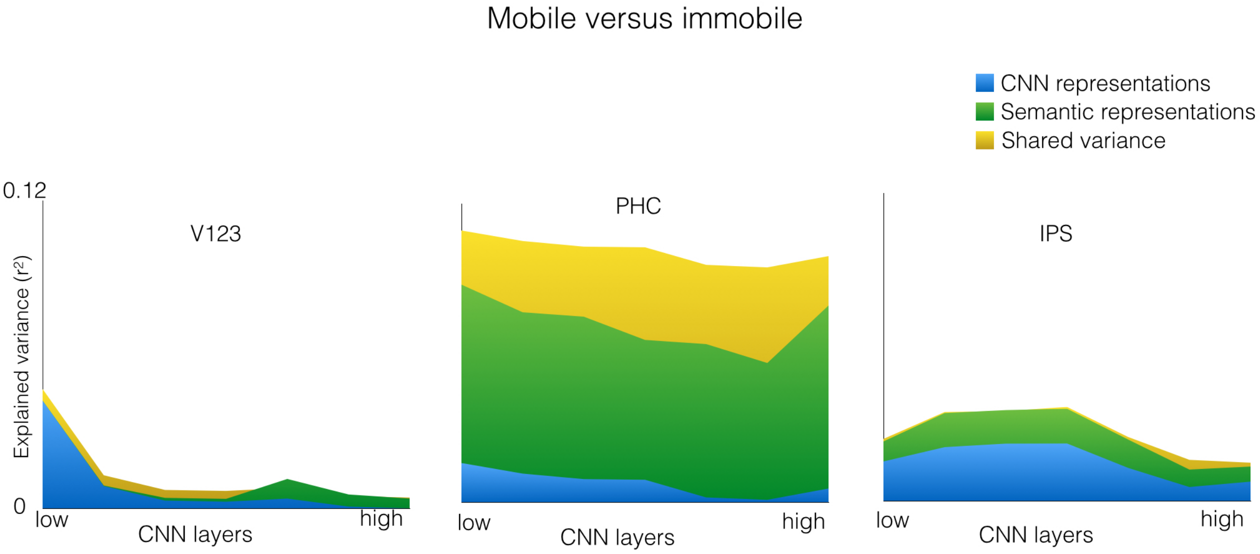


Figure 35: **Unique variance of DNN-visual and semantic representations:** *Unique variation of principal component explained by DNN representations and semantic RDMs ($p < 0.05$). We visualize results for the mobile versus immobile semantic class in 3 brain ROIs V123, PHC and IPS. We observe that in early visual area, especially the lower DNN representations uniquely explains variance of brain responses across subjects. In higher visual areas such as PHC only the mobile versus immobile semantic representation uniquely explain brain responses.*

A recent study that compared visual representations and categories [154] to explain brain responses was limited to the infero-temporal cortex (IT) of individual subjects. In our study, we expand the analysis to cover more brain areas. Other studies [145], [146], [147], [148], [149] have specifically compared visual shape to one semantic dimension (for e.g., animate vs inanimate) to explain brain responses. In our work, we compare a wider range of visual representations from state-of-the-art hierarchical computational model (DNN) against a wider set of semantic representations. While a number of studies have previously correlated computational models to brain responses in the ventral visual cortex, they are typically done with a large stimulus set with few subjects (N = 5 to 20). At the same time, this data is expensive and in studies involving large number of subjects, for e.g individual differences in visual perception, a large number of measurements is quite difficult to obtain. In our study we use limited number of stimulus obtained from a large set of subjects and analyzed the common variation in brain responses against image representations.

In our study, certain results do not correspond to results obtained in recent studies of correlating DNNs to fMRI brain responses. We note that in the ROI analysis in higher regions such as

V4 and LO, lower DNN layers have higher sensitivity compared to higher DNN layers. In (Khaligh-Razavi and Kriegeskorte 2014), higher DNN layers correlate the best to higher visual areas V4 and LO. This divergence in results in higher visual areas might be attributed to the type of visual stimuli used in our study. The stimuli video consists of images from a wide range of concepts including forests, landscapes, rivers etc which can be characterized by local contrast statistics contained in lower DNN layers. The divergence might also be explained by the fact that previous studies have an explicit object recognition task compared to the free viewing task in our study. This might indicate a task depended component in the activation patterns of higher visual brain areas.

The divergence of results obtained from our study using DNNs compared to other studies highlights the limitations of using a small number of stimulus measurements. While natural stimuli is complex and captures a wide range of features [142], our limited stimulus set might contain only a limited variation in stimulus properties. However, it is not certain that high-level stimulus properties can be adequately captured even with a large stimulus set that contains a wide range of variation [155]. One way forward as is suggested in the same article, is to test the DNN representations on different stimulus sets and to search for stimuli in which the DNNs make the best predictions. The new stimuli must reduce the correlations between different stimulus properties but at the same time maintain a natural variance over a wide range of properties. However, the principle to include a wide variety in the stimulus set is only qualitatively defined and might vary across different experimental conditions. There is thus a need to quantitatively define the stimulus properties and to what extent the stimulus set is correlated. An advantage of variation partitioning approach used in this article is that we can quantify to what extent the different representations correlate to each other by regressing out the common variance.

Another advantage of principal component analysis, is that the principal component can rapidly quantify how individual subjects diverge from the common dimension across subjects. The divergence from the principal component is the residual which can be used to indicate the extent of individual differences. The quantification of individual differences can prove useful in a number of studies relating to attentional mechanisms and clinical populations. For example, subjects under different clinical conditions can be aligned on the basis of the principal component and each clinical condition can be defined by the residual.

In sum we hope to have shown that it is possible to analyze the relationship between computational models and the brain using a limited amount of data per subject. At the same time it is important to balance the diversity of visual stimuli with the feature correlations that might arise with such complex stimuli. We suggest that much effort is required to define the stimulus set quantitatively which enables to measure and control the feature coverage and feature covariance. The use of a large number of subjects to study visual processing is a basis to understand individual differences in brain activity as it relates to the fit with computational models of vision which is part of future work.

5.5 SUPPLEMENTARY INFORMATION

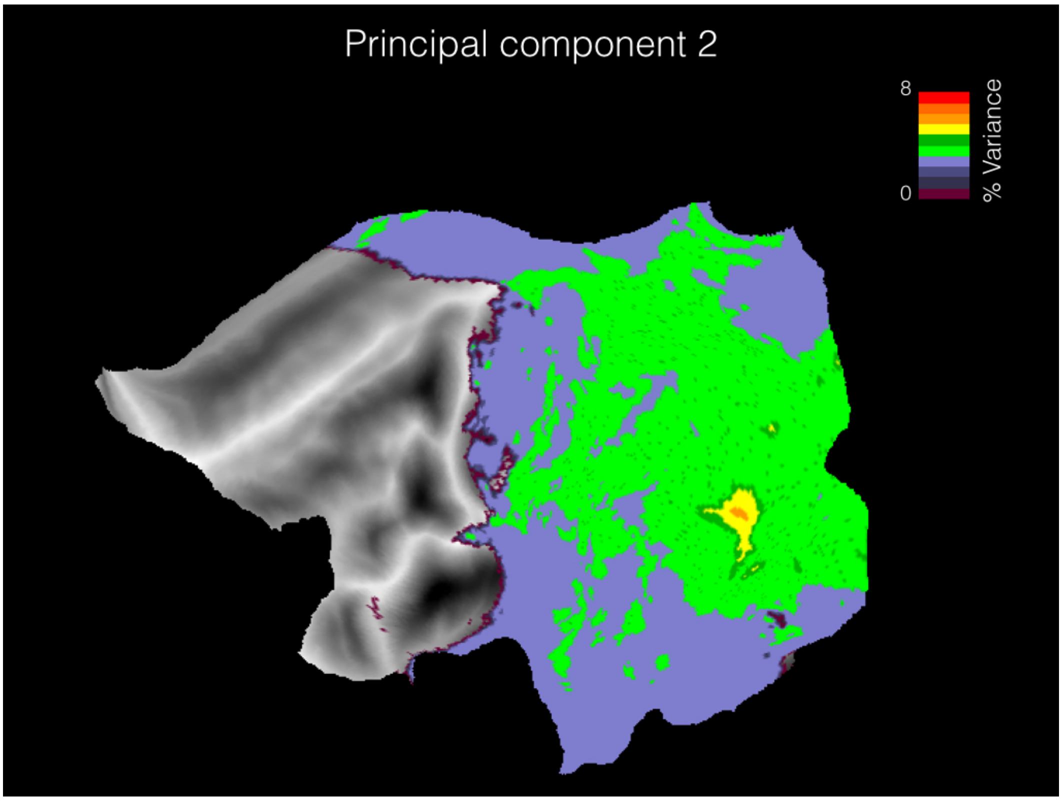


Figure 36: **Second Principal component of across subject responses** *Voxel-wise visualization of the second principal component of brain responses common across subjects. The principal component of the searchlight sphere RDMs across the 100 subjects computed voxel-wise.*

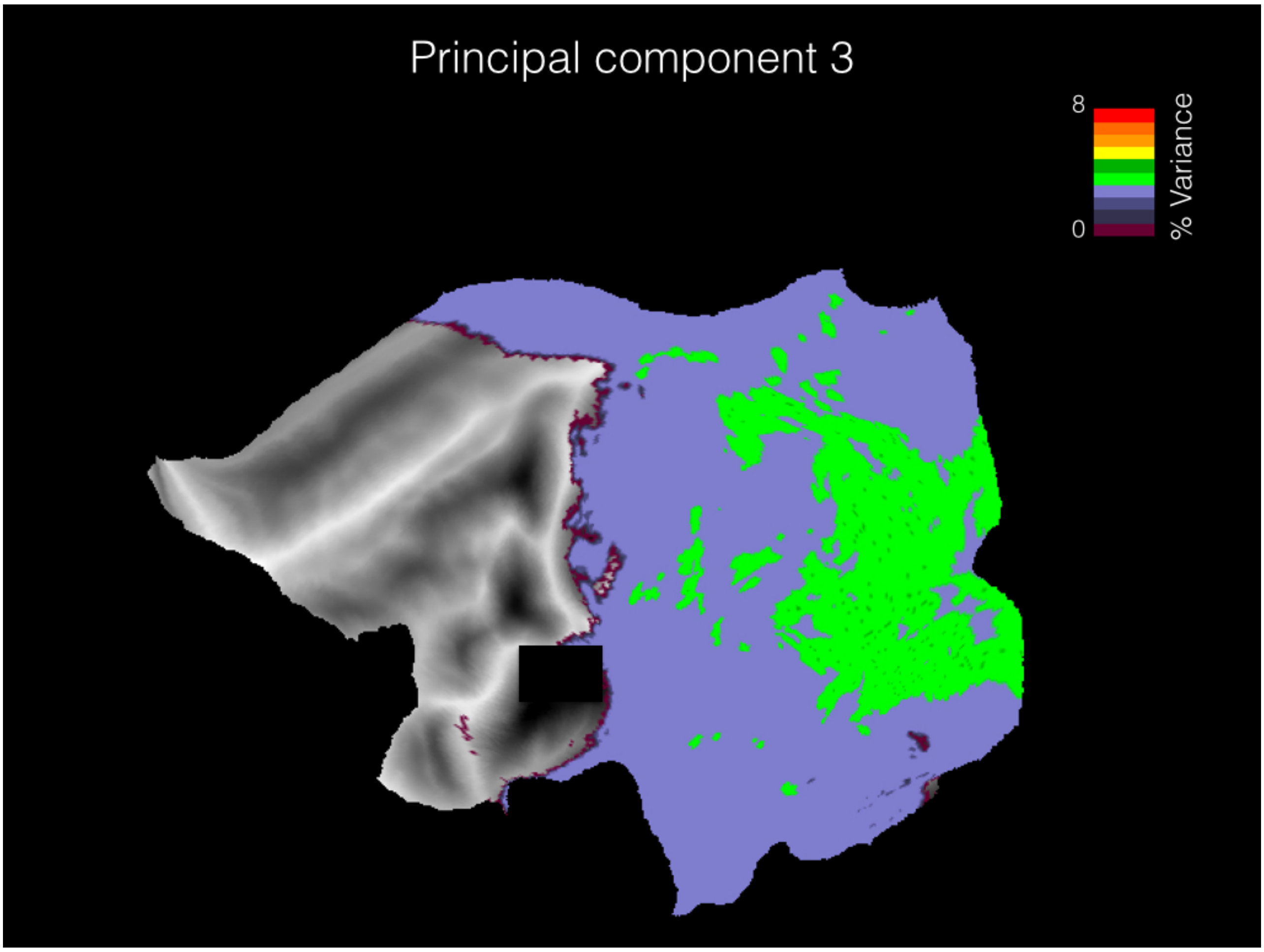


Figure 37: **Third principal component of across subject responses** *Voxel-wise visualization of the third principal component of brain responses common across subjects. The principal component of the searchlight sphere RDMs across the 100 subjects computed voxel-wise.*

6

SUMMARY AND CONCLUSION

6.1 SUMMARY

In this thesis, we study the alignment of representations from computer vision models to human visual representations. Such an approach is aimed at providing new insight in the understanding of human visual representations.

In Chapter 2, we investigate if Bag-of-words (BoW) and HMAX representations of an image correlate to visual representations in the brain. BoW [3], is a framework engineered in computer vision for object recognition and HMAX [156] from the field of biological vision developed to explain the tuning properties of ventral visual cortex. Both these models map low-level features to high level representations via an intermediate computational stage. BoW quantizes Scale Invariant Feature Transform (SIFT) [16] to visual dictionary and HMAX uses templates of Gabor [13] filtered image as a visual dictionary. Different from prior work, we use variation partitioning analysis [52] to study the extent to which visual dictionary from both the models uniquely explain variance in brain responses. Our results show that, the visual dictionary layer add information above and beyond what is computed by the low-level computational stages in both the models. The lower layers (SIFT from BoW, Gabor from HMAX) correlate to visual areas in the brain that are sensitive to edges and oriented gradients (V1). The higher computational layer of HMAX and BoW correlate to brain areas that are sensitive to features of intermediate complexity. This suggests that the level of representation in BoW and HMAX align to low and intermediate levels of human visual hierarchy.

In Chapter 3, we compare state-of-the-art computational models of vision based on correspondence to hierarchy of visual representations in the brain. We test deep neural networks (DNN), BoW and HMAX in their correlation to fMRI brain responses. DNNs consists of many hierarchical computational layers with increasingly complex features and are state of the art in automated object recognition. Recent neuro-imaging studies show that the DNN hierarchy correspond to the human visual hierarchy [157]. Going beyond prior work, we test the unique correlation of each model to fMRI brain responses by variation partitioning. Additionally, we also test the unique correlation of each DNN layer to brain responses. Our results show that, DNNs correspond to visual responses from stimuli consistently across a large number of subjects. Lower DNN layers correspond to the lower visual areas in the brain while the higher layers to higher visual areas of the brain, underscoring the correspondence of deep neural networks to the human visual hierarchy from previous studies. DNNs also outperform other shallow models such as BoW and HMAX in the unique correlation to the brain demonstrating that hierarchy is important. Furthermore, BoW and HMAX correspond to intermediate layers in DNNs in terms of brain responses. Our results show that the additional non-linear computational stages in DNNs as compared to HMAX and BoW better models brain responses to visual stimuli.

In Chapter 4, we test how DNN with increase in number of layers [158] correlate to the hierarchy of visual representations in the brain. Additionally, in the previous chapters we have concentrated on fMRI responses which is derived from slow BOLD responses. However object recognition is extremely fast: it is resolved within the initial hundred milliseconds of visual processing. We test DNN representations with 7 and 15 layers to time resolved electroencephalogram (EEG) data of subjects viewing a large set of natural images. Our results show that both the DNNs correspond to the temporal dynamics of visual processing in the brain. Early layers drive responses early in time, while higher layers drive responses later in time. Comparing the DNNs, we observe the correlation to the brain does not change with increase in layers, but differences emerge when we combine the network layers to explain brain responses after 150 ms. Our results demonstrate that DNN with more number of layers better model the brain.

In Chapter 5, we test whether deep neural networks or semantic representations correlate to representations in the human brain. In prior studies, DNNs and semantic categories have been compared in limited brain areas. We account explicitly for the correlation between the 7 layer deep neural network and a set of semantic classes to map responses in a large number of brain areas. With variation partitioning, we disentangle visual and semantic representations to the similarity in brain responses across 100 subjects during natural vision. As expected we find that deep neural networks correlate with brain responses in early visual areas, while semantic representations correlate with responses in higher visual areas. More importantly, combining DNN and semantic representations we explain more variance in brain activity, suggesting that they provide complementary information in the brain.

6.2 CONCLUSION

Referring back to the main research questions in this thesis, we demonstrate that representations in computer vision models align to human visual representations. For one, we conclude from Chapters 3-5 that deep neural networks are state-of-the-art computational model that corresponds to feedforward visual processing in the human brain.

In Chapter 2, we conclude from our results that BoW and HMAX correlate to visual representations of intermediate complexity in the human brain. Comparing the models from Chapter 2 to Chapters 3 in the correspondence to the brain, we demonstrate the need for deep networks when computationally modeling visual representations in the human brain.

From Chapter 3, we conclude that multiple non-linearities in computer vision models better explain brain responses. This explains why deep neural networks, with alternate linear and non-linear operations better capture neural operations in the human brain.

Results from Chapter 4 suggest that hierarchical deep neural network representations do not align one-to-one with the temporal hierarchy of visual representations in the brain. The interplay of feedback and recurrent processing in visual processing makes the assumption of one-to-one mapping of CNN layers to brain areas too simplistic.

In Chapter 5 we answer the main question, that not only visual but also semantic representations correlate to representations in the human brain. While visual and semantic representations co-vary, they are distinctly represented in the human brain.

There are a number of open questions that are not addressed in this thesis. A main limitation of deep neural networks is that, they are capable to explain only feedforward visual representations in the human brain. However, feedback and recurrent processing play an important role in visual processing. For instance to recognize objects in presence of clutter, a top-down feedback signal is sent to enhance the visual features of the object for optimal recognition. One way to understand the computational mechanisms of feedback and recurrent processing in the brain, is to use very deep neural networks. The recent deep neural networks that have 100 layers or more [159] are hypothesized to capture recurrent processing [160] specifically in the higher layers.

Feedback processing can be captured in brain responses (fMRI/EEG) during complex object recogntion tasks (in the presence of clutter). Aligning models consisting 100 layers or more, we can investigate the computational mechanisms underlying the spatio-temporal dynamics of feedback processing.

A direction for future work is the use of computational models to arrive at the canonical operations of visual information processing in the human brain. All the models in this thesis are composed of four basic computations which also have strong biological basis: i) filterbank convolution [161], ii) non-linearity [162], iii) pooling [163] and iv) normalization [164]. Going forward, a systematic analysis of isolating the computations in different models and correlating them to brain responses will help identify the computations that are critical to explain visual representations in the human brain. This is challenging in practice as it requires models with multitude operations and also an increased signal to noise ratio in the brain data.

BIBLIOGRAPHY

[1] Stefano Soatto and Alessandro Chiuso. Visual representations: Defining properties and deep approximations. *arXiv preprint arXiv:1411.7676*, 2014.

[2] Sven J Dickinson. Object representation and recognition.

[3] Gabriela Csurka, Chris Dance, Lixin Fan, Jutta Willamowski, and Cedric Bray. Visual categorization with bag of keypoints. In *International Workshop on Statistical Learning in Computer Vision*, pages 1–22, 2004.

[4] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2001.

[5] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2411–2418, 2013.

[6] Iasonas Kokkinos. Ubernet: Training auniversal'convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. *arXiv preprint arXiv:1609.02132*, 2016.

[7] Kalanit Grill-Spector, Zoe Kourtzi, and Nancy Kanwisher. The lateral occipital complex and its role in object recognition. In *Vision Research*, volume 41, pages 1409–1422, 2001.

[8] Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.

[9] Edgar A DeYoe, George J Carman, Peter Bandettini, Seth Glickman, JON Wieser, Robert Cox, David Miller, and Jay Neitz. Mapping striate and extrastriate visual areas in human cerebral cortex. *Proceedings of the National Academy of Sciences*, 93(6):2382–2386, 1996.

[10] William T. Freeman and Edward H. Adelson. The Design and Use of Steerable Filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

[11] T. Ahonen, A. Hadid, and M. Pietikainen. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[12] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *Information Theory, IEEE Transactions on*, 8:179–187, 1962.

[13] Anil K. Jain and Farshid Farrokhnia. Unsupervised texture segmentation using Gabor filters. *Pattern Recognition*, 24(12):1167–1186, 1991.

[14] Dorin Comaniciu and Peter Meer. Robust analysis of feature spaces: color image segmentation. *Computer Vision and Pattern Recognition*, pages 750–755, 1997.

[15] Yann. LeCun, Fu Jie Huang, and Leon. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, 2:97–104, 2004.

[16] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[17] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[18] Kevin Jarrett, Koray Kavukcuoglu, Yann LeCun, et al. What is the best multi-stage architecture for object recognition? In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2146–2153. IEEE, 2009.

[19] Geoffrey E. Hinton. Learning multiple layers of representation, 2007.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pages 1–9, 2012.

[21] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv:1502.01852*, 2015.

[23] Jia Deng Jia Deng, Wei Dong Wei Dong, R. Socher, Li-Jia Li Li-Jia Li, Kai Li Kai Li, and Li Fei-Fei Li Fei-Fei. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2–9, 2009.

[24] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1.

[25] Yann Le Cun. A theoretical framework for back-propagation.

[26] Keiji Tanaka. Neuronal mechanisms of object recognition. *Science*, 262:685–688, 1993.

[27] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

[28] Mortimer Mishkin and Leslie G Ungerleider. Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behavioural brain research*, 6(1):57–77, 1982.

[29] George Ettlinger. ?object vision? and ?spatial vision?: the neuropsychological evidence for the distinction. *Cortex*, 26(3):319–341, 1990.

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

[31] Daniel J. Felleman and David C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, 1991.

[32] Victor AF Lamme, Hans Super, and Henk Spekreijse. Feedforward, horizontal, and feedback processing in the visual cortex. *Current opinion in neurobiology*, 8(4):529–535, 1998.

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

[34] KH Foster, James P Gaska, M Nagler, and DA Pollen. Spatial and temporal frequency selectivity of neurones in visual cortical areas v1 and v2 of the macaque monkey. *The Journal of physiology*, 365(1):331–363, 1985.

[35] K Tanaka. Inferotemporal cortex and object vision. *Annual review of neuroscience*, 19:109–139, 1996.

[36] Richard T Born and David C Bradley. Structure and function of visual area mt. *Annu. Rev. Neurosci.*, 28:157–189, 2005.

[37] Kalanit Grill-Spector, Tamar Kushnir, Shimon Edelman, Yacov Itzchak, and Rafael Malach. Cue-invariant activation in object-related areas of the human occipital lobe. *Neuron*, 21(1):191–202, 1998.

[38] Sabine Kastner, Peter De Weerd, and Leslie G Ungerleider. Texture segregation in the human visual cortex: A functional mri study. *Journal of Neurophysiology*, 83(4):2453–2457, 2000.

[39] Nikolaus Kriegeskorte, Bettina Sorger, Marcus Naumer, Jens Schwarzbach, Erik Van Den Boogert, Walter Hussy, and Rainer Goebel. Human cortical object recognition from a visual motion flowfield. *Journal of Neuroscience*, 23(4):1451–1463, 2003.

[40] Zoe Kourtzi and Nancy Kanwisher. Representation of perceived object shape by the human lateral occipital complex. *Science*, 293(5534):1506–1509, 2001.

[41] Zoe Kourtzi and Nancy Kanwisher. Cortical regions involved in perceiving object shape. *The Journal of Neuroscience*, 20(9):3310–3318, 2000.

[42] Michèle Fabre-Thorpe. The characteristics and limits of rapid visual categorization. *Frontiers in psychology*, 2, 2011.

[43] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–34, 2012.

[44] NV Kartheek Medathati, Heiko Neumann, Guillaume S Masson, and Pierre Kornprobst. Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision. *Computer Vision and Image Understanding*, 150:1–30, 2016.

[45] John A. Pyles Daniel D. Leeds, Darren A. Seibert and Michael J. Tarr. Comparing visual representations across human fmri and computational vision, 2013.

[46] Umut Güçlü and Marcel AJ van Gerven. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, 2015.

[47] Seyed Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10(11), 2014.

[48] Charles F. Cadieu, Ha Hong, Daniel L K Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Computational Biology*, 10(12), 2014.

[49] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6:27755, 2016.

[50] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.

[51] Nikolaus Kriegeskorte, Marieke Mur, and Peter a. Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2(November):4, 2008.

[52] Pedro R. Peres-Neto, Pierre Legendre, St??phane Dray, and Daniel Borcard. Variation partitioning of species data matrices: Estimation and comparison of fractions. *Ecology*, 87(10):2614–2625, 2006.

[53] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

[54] M Riesenhuber and T Poggio. Models of object recognition. *Nature neuroscience*, 3 Suppl:1199–204, 2000.

[55] M. Bar. A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, 15:600–609, 2003.

[56] Greg D Field and EJ Chichilnisky. Information processing in the primate retina: circuitry and coding. *Annu. Rev. Neurosci.*, 30:1–30, 2007.

[57] Tim Gollisch and Markus Meister. Eye smarter than scientists believed: neural computations in circuits of the retina. *Neuron*, 65(2):150–164, 2010.

[58] Wei Chen, Toshinori Kato, Xiao-Hong Zhu, John Strupp, Seiji Ogawa, and K?mil U?urbil. Mapping of lateral geniculate nucleus activation during visual stimulation in human brain using fmri. *Magnetic resonance in medicine*, 39(1):89–96, 1998.

[59] Wei Chen, Toshinori Kato, Xiao-Hong Zhu, Seji Ogawa, David W Tank, and Kamil Ugurbil. Human primary visual cortex and lateral geniculate nucleus activation during visual imagery. *Neuroreport*, 9(16):3669–3674, 1998.

[60] Jonathan W Peirce. Understanding mid-level representations in visual processing. *Journal of Vision*, 15(7):5–5, 2015.

[61] Keiji Tanaka. Mechanisms of visual object recognition: monkey and human studies. *Current opinion in neurobiology*, 7(4):523–529, 1997.

[62] N. K. Logothetis and D. L. Sheinberg. Visual object recognition. *Annual Review of Neuroscience*, 19:577–621, 1996.

[63] K. Tanaka. Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19:109–139, 1996.

[64] Chia-Chun Hung, Eric T. Carlson, and Charles E. Connor. Medial axis shape coding in macaque inferotemporal cortex. *Neuron*, 74(6):1099 – 1113, 2012.

[65] Daniel D. Leeds, Darren A. Seibert, John A. Pyles, and Michael J. Tarr. Comparing visual representations across human fmri and computational vision. *Journal of Vision*, 13(13), 2013.

[66] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[67] Max Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 3:1199–1204, 1999.

[68] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *Proceedings of the International Conference on Computer Vision*, 2005.

[69] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In James Ze Wang, Nozha Boujemaa, and Yixin Chen, editors, *Multimedia Information Retrieval*, pages 321–330. ACM, 2006.

[70] Mark Everingham, Luc J. Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[71] Devi Parikh and C. Lawrence Zitnick. The role of features, algorithms and data in visual recognition. In *CVPR*, pages 2328–2335. IEEE, 2010.

[72] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[73] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[74] Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. Intersubject Synchronization of Cortical Activity During Natural Vision. *Science*, 303(5664):1634–1640, March 2004.

[75] Andreas Bartels and Semir Zeki. Functional brain mapping during free viewing of natural scenes. *Human brain mapping*, 21(2):75–85, 2004.

[76] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646, 2011.

[77] N. Kriegeskorte, M. Mur, and P. Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2(4):1–28, 2008.

[78] P R Peres-Neto, P Legendre, S Dray, and D Borcard. Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology*, 87:2614–2625, 2006.

[79] Jim Mutch and David G. Lowe. Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision*, 80(1):45–57, 2008.

[80] Mark Jenkinson, Christian F. Beckmann, Timothy Edward John Behrens, Mark William Woolrich, and Stephen M. Smith. Fsl. *NeuroImage*, 62(2):782–790, 2012.

[81] Mark W. Woolrich, Brian D. Ripley, Michael Brady, and Stephen M. Smith. Temporal autocorrelation in univariate linear modelling of fmri data, 2001.

[82] Gholamreza Salimi-Khorshidi, Gwenaëlle Douaud, Christian F Beckmann, Matthew F Glasser, Ludovica Griffanti, and Stephen M Smith. Automatic denoising of functional mri data: combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage*, 90:449–468, 2014.

[83] R L Buckner. Event-related fmri and the hemodynamic response. *Human brain mapping*, 6(5-6):373–377, 1998.

[84] Shinji Nishimoto, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646, 2011.

[85] B. Douglas Ward. Simultaneous Inference for FMRI Data. 2000.

[86] Peter A. Rogerson. The detection of clusters using a spatial version of the chi-square goodness-of-fit statistic. *Geographical Analysis*, 31(2):130–147, 1999.

[87] Thomas Serre, Lior Wolf, and Tomaso Poggio. Object recognition with features inspired by visual cortex. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 994–1000. Ieee, 2005.

[88] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716, September 2012.

[89] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222, 2013.

[90] Anil K Jain and Farshid Farrokhnia. Unsupervised texture segmentation using gabor filters. *Pattern recognition*, 24(12):1167–1186, 1991.

[91] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 1980.

[92] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7):881–892, 2002.

[93] Aude Oliva. Gist of the scene. 2005.

[94] P.G. Schyns and A. Oliva. From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological Science*, 5(4):195–200, 1994.

[95] Aude Oliva. Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41:176–210, 2000.

[96] LC Loschky and AM Larson. Localized information is necessary for scene categorization, including the natural/man-made distinction. *Journal of Vision*, 8:19, 2008.

[97] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Visual Perception, Progress in Brain Research*, 155, 2006.

[98] Thomas Serre and Maximilian Riesenhuber. Realistic modeling of simple and complex cell tuning in the hmax model, and implications for invariant object recognition in cortex. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LAB, 2004.

[99] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[100] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *NIPS*, pages 1106–1114, 2012.

[101] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, pages 1–42, 2014.

[102] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[103] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput Biol*, 10(11):e1003915, 11 2014.

[104] Charles F Cadieu, Ha Hong, Daniel L K Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10:e1003963, 2014 Dec 2014.

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

[106] Renata MCR de Souza and Francisco de AT De Carvalho. Clustering of interval data based on city–block distances. *Pattern Recognition Letters*, 25(3):353–365, 2004.

[107] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[108] Yoshua Bengio et al. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[109] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014*, pages 818–833. Springer, 2014.

[110] Kendrick N Kay, Jonathan Winawer, Ariel Rokem, Aviv Mezer, and Brian A Wandell. A two-stage cascade model of bold responses in human visual cortex. *PLoS computational biology*, 9(5):e1003079, 2013.

[111] Liang Wang, Ryan EB Mruczek, Michael J Arcaro, and Sabine Kastner. Probabilistic maps of visual topography in human cortex. *Cerebral Cortex*, page bhu277, 2014.

[112] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience*, 35(27):10005–10014, 2015.

[113] Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Mapping human visual representations in space and time by neural networks. *Journal of vision*, 15(12):376–376, 2015.

[114] Kalanit Grill-Spector, Tammar Kushnir, Talma Hendler, Shimon Edelman, Yacov Itzchak, Rafael Malach, et al. A sequence of object-processing stages revealed by fmri in the human occipital lobe. *Human brain mapping*, 6(4):316–328, 1998.

[115] Simon J. Thorpe, Sabastien Crouzet, Holle Kirchner, and Michale Fabre-thorpe. Ultra rapid face detection in natural images: Implications for computation in the visual system.

[116] Victor AF Lamme and Pieter R Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in neurosciences*, 23(11):571–579, 2000.

[117] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[118] Noor Seijdel, Kandan Ramakrishnan, Max Losch, and Steven Scholte. Overlap in performance of cnn's, human behavior and eeg classification. *Journal of Vision*, 16(12):501–501, 2016.

[119] Groen, Sennay Ghebreab, Victor Lamme, and Steven Scholte. The role of weibull image statistics in rapid object detection in natural scenes. *Journal of Vision*, 10(7):992–992, 2010.

[120] Dirk B Walther, Eamon Caddigan, Li Fei-Fei, and Diane M Beck. Natural scene categories revealed in distributed patterns of activity in the human brain. *The Journal of Neuroscience*, 29(34):10573–10581, 2009.

[121] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Computer Vision–ECCV 2008*, pages 304–317. Springer, 2008.

[122] Andreas Opelt, Axel Pinz, Michael Fussenegger, and Peter Auer. Generic object recognition with boosting. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(3):416–431, 2006.

[123] Adriana Olmos et al. A biologically inspired algorithm for the recovery of shading and reflectance images. *Perception*, 33(12):1463–1473, 2003.

[124] Groen, Sennay Ghebreab, Hielke Prins, Victor AF Lamme, and H Steven Scholte. From image statistics to scene gist: evoked neural activity reveals transition from low-level natural image structure to scene category. *Journal of Neuroscience*, 33(48):18814–18824, 2013.

[125] Groen, Sennay Ghebreab, Victor AF Lamme, and H Steven Scholte. Spatially pooled contrast responses predict neural and perceptual similarity of naturalistic image categories. *PLoS Comput Biol*, 8(10):e1002726, 2012.

[126] Gabriele Gratton, Michael GH Coles, and Emanuel Donchin. A new method for off-line removal of ocular artifact. *Electroencephalography and clinical neurophysiology*, 55(4):468–484, 1983.

[127] F Perrin, J Pernier, O Bertrand, and JF Echallier. Spherical splines for scalp potential and current density mapping. *Electroencephalography and clinical neurophysiology*, 72(2):184–187, 1989.

[128] H Steven Scholte, Sennay Ghebreab, Lourens Waldorp, Arnold WM Smeulders, and Victor AF Lamme. Brain responses strongly correlate with weibull image statistics when processing natural images. *Journal of Vision*, 9(4):29, 2009.

[129] Sennay Ghebreab, Steven Scholte, Victor Lamme, and Arnold Smeulders. A biologically plausible model for rapid natural scene identification. In *Advances in Neural Information Processing Systems*, pages 629–637, 2009.

[130] Kenneth P Burnham and David R Anderson. Multimodel inference understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

[131] Nancy Kanwisher and Etc Dilks. The functional organization of the ventral visual pathway in humans.

[132] Kandan Ramakrishnan, H Steven Scholte, Iris I A Groen, Arnold W Smeulders, and Sennay Ghebreab. Visual dictionaries as intermediate features in the human brain. *Frontiers in Computational Neuroscience*, 8(168), 2015.

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

[134] Rufin VanRullen and Simon J Thorpe. Surfing a spike wave down the ventral stream. *Vision research*, 42(23):2593–2615, 2002.

[135] Karl Zipser, Victor AF Lamme, and Peter H Schiller. Contextual modulation in primary visual cortex. *Journal of Neuroscience*, 16(22):7376–7389, 1996.

[136] Pieter R Roelfsema, Victor AF Lamme, and Henk Spekreijse. The implementation of visual routines. *Vision research*, 40(10):1385–1411, 2000.

[137] David J Heeger. Theory of cortical function. *Proceedings of the National Academy of Sciences*, 114(8):1773–1782, 2017.

[138] CHARLES G Gross. The neural basis of stimulus equivalence across retinal translation. *Lateralization in the nervous system*, pages 109–122.

[139] Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11):4302–4311, 1997.

[140] Andrew C Connolly, J Swaroop Guntupalli, Jason Gors, Michael Hanke, Yaroslav O Halchenko, Yu-Chien Wu, Hervé Abdi, and James V Haxby. The representation of biological classes in the human brain. *Journal of Neuroscience*, 32(8):2608–2618, 2012.

[141] Talia Konkle and Aude Oliva. A real-world size organization of object responses in occipitotemporal cortex. *Neuron*, 74(6):1114–1124, 2012.

[142] Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.

[143] Grace E Rice, David M Watson, Tom Hartley, and Timothy J Andrews. Low-level image properties of visual objects predict patterns of neural response across category-selective regions of the ventral visual pathway. *Journal of Neuroscience*, 34(26):8837–8844, 2014.

[144] David M Watson, Tom Hartley, and Timothy J Andrews. Patterns of response to visual scenes are linked to the low-level properties of the image. *NeuroImage*, 99:402–410, 2014.

[145] Stefania Bracci and Hans Op de Beeck. Dissociations and associations between shape and category representations in the two visual pathways. *Journal of Neuroscience*, 36(2):432–444, 2016.

[146] Scott N Macdonald and Jody C Culham. Do human brain areas involved in visuomotor actions show a preference for real tools over visually similar non-tools? *Neuropsychologia*, 77:35–41, 2015.

[147] Daniel Kaiser, Damiano C Azzalini, and Marius V Peelen. Shape-independent object category responses revealed by meg and fmri decoding. *Journal of neurophysiology*, 115(4):2246–2250, 2016.

[148] Daria Proklova, Daniel Kaiser, and Marius V Peelen. Disentangling representations of object shape and object category in human visual cortex: The animate–inanimate distinction. *Journal of cognitive neuroscience*, 2016.

[149] Peter B Bryan, Joshua B Julian, and Russell A Epstein. Rectilinear edge selectivity is insufficient to explain the category selectivity of the parahippocampal place area. *Frontiers in human neuroscience*, 10, 2016.

[150] Brian A Wandell, Jonathan Winawer, and Kendrick N Kay. Computational modeling of responses in human visual cortex.

[151] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

[152] Iris IA Groen, Sennay Ghebreab, Victor AF Lamme, and H Steven Scholte. Low-level contrast statistics are diagnostic of invariance of natural textures. *Frontiers in computational neuroscience*, 6, 2012.

[153] Lindsay I Smith et al. A tutorial on principal components analysis. 2002.

[154] Kamila M Jozwik, Nikolaus Kriegeskorte, and Marieke Mur. Visual features as stepping stones toward semantics: Explaining object similarity in it and perception with non-negative least squares. *Neuropsychologia*, 83:201–226, 2016.

[155] Mark D Lescroart, Dustin E Stansbury, and Jack L Gallant. Fourier power, subjective distance, and object categories all provide plausible models of bold responses in scene-selective visual areas. *Frontiers in computational neuroscience*, 9:135, 2015.

[156] Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997.

[157] U. Güçlü and M. a. J. van Gerven. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[158] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Iclr*, pages 1–14, 2015.

[159] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[160] Qianli Liao and Tomaso Poggio. Bridging the gaps between residual learning, recurrent neural networks and visual cortex. *arXiv preprint arXiv:1604.03640*, 2016.

[161] Serge O Dumoulin and Brian A Wandell. Population receptive field estimates in human visual cortex. *Neuroimage*, 39(2):647–660, 2008.

[162] Nicholas J Priebe and David Ferster. Inhibition, spike threshold, and stimulus selectivity in primary visual cortex. *Neuron*, 57(4):482–497, 2008.

[163] Ilan Lampl, David Ferster, Tomaso Poggio, and Maximilian Riesenhuber. Intracellular measurements of spatial integration and the max operation in complex cells of the cat primary visual cortex. *Journal of neurophysiology*, 92(5):2704–2713, 2004.

[164] Ba Olshausen. Principles of image representation in visual cortex. *The visual neurosciences*, pages 1603–1615, 2003.

SAMENVATTING

In dit proefschrift onderzoeken we de overeenstemming tussen representaties van computer modellen en van menselijke visuele representaties. Deze benadering beoogt nieuwe inzichten te genereren om menselijke visuele representaties te begrijpen.

In hoofdstuk 2 onderzoeken we of ”Bag-of-Words” (BoW) en HMAX representaties van een afbeelding correleren met visuele representaties in het brein. BoW is een model ontwikkeld in computer vision voor objectherkenning terwijl HMAX is ontwikkeld vanuit de biologie met het doel om de computationele eigenschappen van de visuele context te modelleren. Beide modellen beschrijven hoe simpele visuele elementen gecombineerd kunnen worden tot complexe representaties via een tussenliggende computationele stap. BoW berekent een visueel woordenboek met behulp van de ”Scale Invariant Feature Transform” (SIFT) en HMAX gebruikt een template van Gabor-gefilterde afbeeldingen als een visueel woordenboek. Om uit te kunnen rekenen hoeveel unieke variantie van breinactiviteit we kunnen verklaren met de verschillende modellen gebruiken we Variation Partitioning Analysis. Onze resultaten laten zien dat de visuele woordenboeken variantie verklaren die niet verklaar kunnen worden door de simpele computationele eigenschappen van de modellen. De eerdere lagen (SIFT van BoW, Gabor van HMAX) correleren met visuele regio’s in het brein die randen en georienteerde gradienten encodeert (V1). De meer complexe computationele lagen van HMAX and BoW correleren met breinregio’s die eigenschappen van intermediaire complexiteit encoderen.

In hoofdstuk 3 vergelijken we best werkende computationele computer modellen gebaseerd op de gelijkenis van visuele representaties in het brein. We testen Deep Neural Networks (DNN), BoW, en HMAX op hoeverre ze correleren met fMRI breinactiviteit. DNN-modellen bestaan uit verschillende hierarchische computationele lagen (van ”laag” naar ”hoog”) met in toenemende mate complexere eigenschappen en worden gezien als best presterend in het domein van objectherkenning. Recente studies over beeldvormingstechnieken voor de hersenen laten zien dat de hierarchie van DNN-modellen correspondeert met de menselijke visuele hierarchie. Als toevoeging aan deze studies tonen wij de unieke correlatie van elk model met fMRI breinactiviteit aan door middel van variation partitioning. Daarnaast testen we de unieke correlatie van elke DNN-laag met breinactiviteit. Onze resultaten laten zien dat DNN-modellen consistent corresponderen met visuele eigenschappen van stimuli bij een groot aantal proefpersonen. De eerste lagen in DNN-modellen corresponderen met vroegere visuele regio’s in het brein terwijl de hogere DNN-lagen corresponderen met latere visuele regio’s van het brein, wat de overeenkomst van DNN-modellen en het brein gevonden door eerdere studies onderstreept. DNN-modellen presteren ook beter dan modellen met minder lagen zoals BoW en HMAX in de unieke correlatie met het brein, wat laat zien dat hierarchische lagen belangrijk zijn. Daarnaast corresponderen BoW en HMAX met intermediaire lagen in DNNs wat betreft breinactiviteit. Onze resultaten laten zien dat de toevoeging van niet-lineaire computationele stappen in DNN-modellen, vergeleken met HMAX en BoW, breinactiviteit naar aanleiding van visuele stimuli beter modelleert.

In hoofdstuk 4 testen we hoe DNN-modellen correleren met visuele representaties in het brein naarmate de hoeveelheid lagen toeneemt. In de vorige hoofdstukken hebben we data gebruikt die verkregen is met de relatief trage, maar spatieel nauwkeurige, BOLD-reponse die geen onderscheid kan maken tussen processen in de eerste en tweede honderd milliseconden. Objectherkenning is echter ontzettend snel en in dit hoofdstuk hebben we de veel snellere EEG metingen gebruikt om te testen of DNN representaties van 7 en 15 lagen verschillen in hun kracht om hersendata te verklaren. Onze resultaten laten zien dat beide DNN-modellen corresponderen met de temporele dynamiek van visuele informatieverwerking in het brein. Lage DNN-lagen verklaren variantie vroeg in de tijd, terwijl hoge lagen later in de tijd variantie verklaren. Wanneer we de DNN-modellen vergelijken, observeren we dat de correlatie met het brein, voor elke laag apart, niet verschilt, maar dat hersenactiviteit na 150 ms door alle lagen van het diepere netwerk beter wordt verklaart (ook als er gecorrigeerd wordt voor het aantal netwerklagen). Onze resultaten demonstreren dat DNN-modellen met een groter aantal lagen het brein, later in de tijd, beter modelleren.

In hoofdstuk 5 testen we of DNN-modellen of semantische representaties correleren met representaties in het menselijke brein. In eerdere studies zijn DNN-modellen en semantische categorieen vergeleken in een gelimiteerde set van regio’s in het brein. Wij houden expliciet rekening met de correlatie tussen de DNN met 7 lagen en een set van semantische categorieen in onze analyse van activiteit in een groot aantal regio’s in het brein. Met variation partitioning halen we de invloed van visuele en semantische representaties op hun gelijkenis met breinactiviteit uit elkaar bij 100 proefpersonen gedurende visuele verwerking van natuurlijke stimuli. Zoals verwacht vinden we dat DNN-modellen correleren met breinactiviteit in vroege visuele gebieden, terwijl semantische representaties correleren met activiteit in late visuele gebieden. Daarnaast vinden we dat het combineren van DNN en semantische representaties meer variantie in de breinactiviteit verklaart, wat suggereert dat ze complementaire informatie modelleren in het brein.

ACKNOWLEDGEMENTS

An oft repeated saying in hinduism; Matha Pitha Guru Deivam (Mother, Father, Teacher, God) in order of significance. In many ways, during my time in Amsterdam I discovered the Guru.

I consider myself extremely lucky to have Arnold as a supervisor who showed me the mirror as a scientist and more importantly as a human being. He constantly challenged my assumptions at every meeting and forced me to think beyond what I thought was sufficient. Using the PhD he deciphered my personality in a very fundamental way and persistently didn't allow me to not improve myself. His simple life lessons on making mistakes, being decisive, ability to say No, structuring time, the thought process to approach a problem and more have equipped me to navigate the real world. For that and more I am happy that he stuck with me during my journey even at times when I didn't give him reason to. His words will constantly ring in my mind, and I will leave Amsterdam much much wiser in science and life.

Sennay and Steven provided contrasting perspectives to science and life. The journey started when I interviewed with Sennay in September 2011, and within one week I resigned from my previous job and applied for the dutch visa. Looking back that was almost a leap of faith. Sennay mentored me during the first couple of years when I struggled in the inter-disciplinary project, eased my troubles during moments of self-doubt in a very unassuming manner. From writing articles, making me think, coming up with problems and possible solutions he was constantly around. I came into the PhD highly unprepared and I thank Sennay for holding my hand to set up the foundation. Furthermore, Sennay was an interface between the indian style and dutch style that I often encountered. While Sennay was the serious person, Steven was always bubbling with childlike enthusiasm. We had a more productive interaction with the ”deep learning” wave, and had many brainstorming sessions with the three of us. Steven's enthusiasm is infectious and always encouraged me in everything no matter how misplaced my confidence might have been. Steven was the person who talked about the good life - food and wine. My first VSS conference was a blast thanks to Steven and also enjoyed the many ”fine dinners” he organized. The weekly journal club meetings at the psychology campus were stimulating, also thanks to Sander and Davide from CWI. I hope to have made friends for life in Steven and Sennay.

Life during the PhD was made much easier by my office mates and the many friends I made over the years. All my labmates with whom I had innumerable sessions of dinners, borrels, BBQs, volleyball, squash and Mill. My office mates, Jorn and Agnes who are now extremely good friends. Also - Mihir, Svetlana, Ran, Zhenyang, Silvia, Amir, Masoud, Sezer, Fares, Spencer, Pascal, Morris, Stratis, Jan, Honza, Thomas, Devanshu (a great help when I was away from Amsterdam), Gjorgi, the Stevan Bhalu. Across the river - Ilja, Lukas, Noor and Max. My erasmus friends - Patrizio, Alex, Carmela and the gang. My dear mates from Diemen - Sarunas, Matteo, Laura, the greeks, Sha and Prem. To Marta, Aiste and my wonderful friend Chris who was by my side, my best wishes for Phd and life. All the experiments in Leidseplein/Westerpark/Rembrandtplein - Harish, for the many trips, my flatmate and partner in crime - Saurabh, Parthiv, the party boys and many laughs - Nishanth and Prateek, the amazing cooks - Proma and Prateeti. From Minnesota - the Rendu idli guys. From Nice - Silviu, Vasanth, Ratnesh, Srinidhi, Sagar, JD, Quentin, Dori. Ritesh, Pothireddy, Bharati and Roopa. The SRM group - Arvind, Bala, Arathi, Ashwath and gang. To my time in Palo Alto which was the precursor for my interest in visual neuroscience.

I would like to specially thank Iris and Rajat for our collaborative efforts. Iris is a very cool scientist but at the same time highly focussed and helped me immensely in writing papers suitable for neuroscience journals. I met Rajat quite late during my PhD, but from food to science to cricket it was a pleasure to spend time with him and Theresa, a lovely couple. I highly appreciate all the tips Rajat provided during the last couple of years, proof reading my thesis, working with me on a workshop article and feeding me with great kerala food.

To Virginie, who answered all my questions with a smile even when she was not the right person to give an answer. I would like to thank my committee members for taking time from their busy schedule to read my thesis.

I would like to thank my parents - Ramakrishnan and Santhi who made innumerable sacrifices for me. To my sweet sister - Bhavani who is always a pillar of support. To Vinitha who took the leap of faith with me and tolerated the erratic me in my final months.

Finally as a man from the east, I cannot help but say the journey has been as much spiritual as it has been scientific.

”You can't connect the dots looking forward. You can only connect them looking backward. So you have to trust that the dots will somehow connect in your future. You have to trust in something ; your gut, destiny, life, karma, whatever.” - Steve Jobs