bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A Self-Supervised Deep Neural Network for Image Completion Resembles Early Visual Cortex fMRI Activity Patterns for Occluded Scenes

Michele Svanera*, **Andrew T. Morgan, Lucy S. Petro, Lars Muckli**
Centre for Cognitive Neuroimaging
Institute of Neuroscience and Psychology
University of Glasgow (UK)
{name.surname}@glasgow.ac.uk

December 13, 2020

Abstract

The promise of artificial intelligence in understanding biological vision relies on the comparison of computational models with brain data with the goal of capturing functional principles of visual information processing. Convolutional neural networks (CNN) have successfully matched the transformations in hierarchical processing occurring along the brain's feedforward visual pathway extending into ventral temporal cortex. However, we are still to learn if CNNs can successfully describe feedback processes in early visual cortex. Here, we investigated similarities between human early visual cortex and a CNN with encoder/decoder architecture, trained with self-supervised learning to fill occlusions and reconstruct an unseen image. Using Representational Similarity Analysis (RSA), we compared 3T fMRI data from a non-stimulated patch of early visual cortex in human participants viewing partially occluded images, with the different CNN layer activations from the same images. Results show that our self-supervised image-completion network outperforms a classical object-recognition supervised network (VGG16) in terms of similarity to fMRI data. This provides additional evidence that optimal models of the visual system might come from less feedforward architectures trained with less supervision. We also find that CNN decoder pathway activations are more similar to brain processing compared to encoder activations, suggesting an integration of mid- and low/middle-level features in early visual cortex. Challenging an AI model and the human brain to solve the same task offers a valuable way to compare CNNs with brain data and helps to constrain our understanding of information processing such as neuronal predictive coding.

Keywords Self-supervised deep neural networks · Encoder/decoder architecture · 3T fMRI · Early visual cortex · Representational Similarity Analysis (RSA).

Introduction

Investigating the functional dichotomy of forward and backward pathways in the visual system is necessary for describing how the brain performs vision. The conceptual divergence is broadly understood; forward pathways carry

*Corresponding author: Michele.Svanera at glasgow.ac.uk

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

unlabelled sensory information into the brain while feedback pathways carry signals from higher visual and non-visual areas back to earlier areas ([Schwiedrzik and Freiwald, 2017, Pennartz et al., 2019]), in the opposite direction to sensory information. Feedback has an important role in the contextual modulation of the feedforward stream, carrying top-down signals such as attention and expectations. ([Van Essen and Anderson, 1995, Gilbert and Li, 2013, Roelfsema and de Lange, 2016, Angelucci et al., 2017, Pennartz et al., 2019, Klink et al., 2017, Morgan et al., 2019, Takahashi et al., 2016]). Neuronal information processing in visual perception can be formalised as statistical inference based on hierarchical internal models, which can theoretically be implemented by schemes such as predictive coding ([Rao and Ballard, 1999, Friston, 2005]). Predictive processing is a compelling framework for describing neural phenomena observed in human primary visual cortex using brain imaging ([Alink et al., 2010, Edwards et al., 2017]). Predictive coding describes a neuronal coding process in which predicted information from top-down streams is explained away from the feedforward stream (e.g., [Rao and Ballard, 1999]). Less narrow implementations of predictive coding recognise the necessity of the feedforward stream to not only compute prediction error but also communicate information that reinforces the internal models which were successful in predicting incoming information ([Shipp, 2016, Rao and Ballard, 1999]). We use the term predictive processing to allow for an even broader implementation of coding principles, including those that provide context (e.g., contour ownership, spatial information from auditory cues) without being limited to predicting a narrow set of precise features (e.g., a specific contour). For example, illusions of motion increase activity on the non-stimulated motion path in primary visual cortex [Erlikhman and Caplovitz, 2017, Chong et al., 2016], and predictable stimuli presented on the motion path cause less activity than surprising stimuli ([Alink et al., 2010, Edwards et al., 2017]). Moreover, visual contours predicted by flanking stimuli increase activity along the illusory contour [Kok et al., 2016, Bergmann et al., 2019]. These data reveal that some neuronal activity in early visual areas is not directly related to retinotopic sensory inputs but rather to the brain's inference of the world, transmitted in cortical feedback pathways to earlier areas. A comprehensive biologically constrained artificial model of vision should account for this top-down brain processing.

In computational neuroscience, there is a tradition to exploit recent developments in statistical modelling and algorithms to develop more sophisticated models of visual processing ([Simoncelli and Olshausen, 2001]). In this respect, recent developments in artificial intelligence (AI), and in particular deep learning (DL), offered significant contributions in the last decade ([Hassabis et al., 2017]), showing promising results in the attempt to improve our understanding of visual counterstream computations ([Kietzmann et al., 2019, Qiao et al., 2019]). In the DL field, and more broadly in AI, a textbook categorisation in the literature is the distinction between supervised and unsupervised learning; while the first learns statistical representations based on labelled datasets, including tasks such as classification, regression, and segmentation, the latter tries to extract features and inherent patterns from unlabelled data, with tasks such as clustering, anomaly detection, and dimensionality reduction ([Hastie et al., 2009]). While supervised learning usually allows to obtain better performance, being able to make sense of the large amount of unlabelled data often available is a promising and exciting goal for the field. In this respect, the recent trend of self-supervised learning tries to use the data as a supervision strategy ([Jing and Tian, 2020]): the idea is to challenge the model to solve a specific unrelated task in order to learn representations of the data or to automatically label the dataset. Examples include learning to predict some part of the image from other parts, predicting relative locations of two image patches, solving a jigsaw puzzle, and colorizing an image. Originally designed as unsupervised methods, in Generative Adversarial Networks (GAN - [Goodfellow et al., 2014]), a discriminator is trained by evaluating whether the data created by a generator is part of the training data (i.e., is a real image) or not (fake). Despite their great promise of data understanding, unsupervised learning methods are not often used for model comparisons with brain data.

Supervised learning models, under the flagship of Convolutional Neural Networks (CNN), have advanced our understanding of visual information processing in the brain in a number of breakthrough studies ([Hassabis et al., 2017]). Using this approach to model spiking computations, CNN models can predict neural activations in the macaque visual ventral streams at early time points after stimulus presentation, suggesting they may capture important aspects of feedforward visual processing ([Yamins and DiCarlo, 2016]). Natural images have been found to cluster together in similar ways in the internal feature spaces of CNNs as in human Inferior Temporal (IT) cortex ([Khaligh-Razavi and Kriegeskorte, 2014]). [Cichy et al., 2016] described how a CNN captured the stages of human visual processing in time and space from early

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

visual areas towards the dorsal and ventral streams. [Kay et al., 2008] and [Güçlü and van Gerven, 2015] compared two different encoding models to human fMRI data, one based on decomposing visual information into Gabor elements, and the other based on trained feedforward CNNs. Comparisons showed improved results of the supervised CNN with respect to the Gabor based approach in explaining fMRI activity patterns. Using decoding modelling, different studies ([Eickenberg et al., 2017, He et al., 2018]) proposed to derive CNN representations from brain data, or transfer learning CNN discrimination properties for fMRI decoding ([Svanera et al., 2019]). In a broader effort to validate CNNs as models of the visual system, supervised networks on image recognition were used to model differences in retinal and cortical networks showing which architecture constraints help the emergence of similar CNN representations in the brain ([Lindsey et al., 2019]). To better navigate between different object detection artificial networks, Schrimpf et al. ([Schrimpf et al., 2018]) introduce multiple neural and behavioural benchmarks to score any artificial neural network on how similar it is to the brain's mechanisms for core object recognition. CNNs with recurrent connections better predict visual responses than feedforward models, and increase our ability to capture the cortical dynamics in MEG data and fMRI data ([Kietzmann et al., 2019, Qiao et al., 2019]), opening new opportunities on understanding visual counterstreams by modelling brain architectures and processing principles [Hong et al., 2016, Kar et al., 2019, Tang et al., 2018].

While these achievements are noteworthy in terms of being the best available models of biological vision, the possible impact of unsupervised models remains comparatively under explored. As in AI, unsupervised learning models are promising in terms of data exploitation and understanding. Furthermore, in modelling vision we could potentially narrow the gap created by the biological implausibility of supervised learning objectives and gain progress in understanding the role of cortical feedback processing. In addition, the setting (i.e., the task) in which the two models, biological and artificial, are compared could provide a better testbed and bring potential benefits, such as giving insights on the learning dynamics or identifying differences with respect to other tasks. We begin in this direction by comparing brain data capturing contextual feedback in vision (in amodal filling-in) to an artificial model trained to fill-in missing visual information. The human brain imaging data was recorded from primary visual cortex (V1) whilst subjects viewed partially occluded natural scene images ([Smith and Muckli, 2010, Muckli et al., 2015, Revina et al., 2018, Morgan et al., 2019]). By partially masking the visual stimulus and recording from the corresponding retinotopic region of V1, this paradigm allows us to study feedback information in the absence of stimulus-specific feedforward signals. We use this data as a testbed for investigating a vision model that includes feedback components. We selected a visual occlusion paradigm because we sought comparable functionality between the brain data and computer vision algorithms performing the same task. The predictive coding framework suggests that the brain is trying to reconstruct the image under the occlusion, using information from cortical feedback to infer the content of missing image sections. We showed recently that the brain's filling of the missing image section can be modelled as a behavioural line drawing [Morgan et al., 2019]. This operation of reconstructing an unseen image is conceptually similar to the task of inpainting (also known as image completion) in computer vision. In this task, an artificial model predicts the missing part of the image (unseen or damaged, [Pathak et al., 2016]) relying on image statistics learnt during training. This learning can be obtained using different techniques: with classical image processing techniques, as the application of statistically similar patches ([He and Sun, 2014]), or with CNN approaches ([Yu et al., 2018, Lempitsky et al., 2018, Isola et al., 2017]). To solve the task of image reconstruction, we used a CNN with encoder/decoder architecture, trained in a self-supervised fashion as in [Isola et al., 2017], to perform inpainting on the lower-right quadrant i.e., reconstructing the unseen (occluded) portion of the image (Figure 1 - a). We compared the encoder/decoder network trained to solve inpainting with brain data collected in a previous 3-Tesla functional magnetic resonance imaging (fMRI) experiment during the viewing of images with the lower-right quadrant occluded (Figure 1 - b), see [Morgan et al., 2019]). We investigated similarities between the network and brain data using Representational Similarity Analyses (RSA, [Kriegeskorte et al., 2008, Nili et al., 2014]).

Material and methods

We compared representations of two neural network models to brain imaging data acquired during an fMRI experiment. We first describe the fMRI experiment (Section *fMRI experiment*), before specifying the two CNN models: the well-known supervised (feedforward) network VGG16 and our encoder/decoder model trained to fill the missing quadrant of

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the image (Section *Artificial neural network models*). Lastly, we explain how we compared the brain data to the CNNs usingRepresentational Similarity Analysis (RSA) (Section *Data analysis: Representational similarity analysis (RSA)*). The experimental framework is shown in Figure 1.

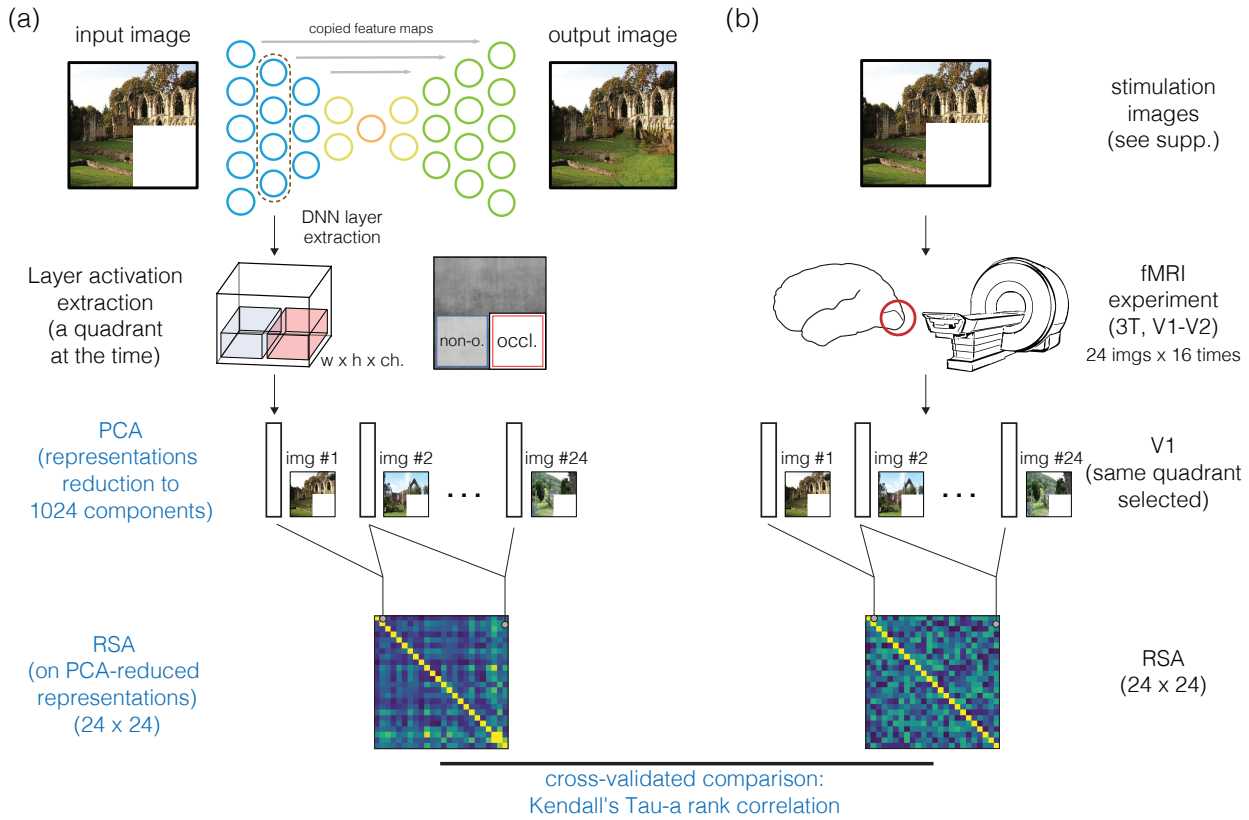


Figure 1: The analysis framework is composed of two parts: the encoder/decoder (artificial) neural network and brain imaging data collection. (a) The image passed through the network, we extracted activations for one layer, we selected a quadrant at the time, and we applied PCA transformation to reduce the dimension to 1024 components; we then obtained one 1024-d vector per layer (15), per quadrant (2), and per image (24). We used these vectors to compute Representational Dissimilarity Matrices (RDMs). (b) fMRI data were collected from participants whilst viewing the same images that were fed into the network (in testing) and RDMs computed. We compared RDMs of the network and brain data (cross-validated - see [Walther et al., 2016]), for every CNN layer (15 layers analysed), human visual area (V1 and V2), and image space quadrant (occluded and non-occluded quadrants).

fMRI experiment

Our fMRI dataset was collected previously and published in [Morgan et al., 2019]. Eighteen healthy volunteers with normal or corrected-to-normal vision participated in this study. Twenty-four real-world scenes from six categories (beaches, buildings, forests, highways, industry, and mountains), from the dataset in ([Walther et al., 2009]) were shown to participants ([Morgan et al., 2019]; see supplementary for stimulation images). Images were displayed in grayscale on a rear-projection screen using a projector system. Stimuli spanned $19.5° \times 14.7°$ of visual angle and were presented with the lower-right quadrant occluded by a white box (occluded region spanned $\sim 9° \times 7°$). A centralised fixation checkerboard marked the centre of the scene images. Over the course of the experiment, each image was presented 16 times. Refer to [Morgan et al., 2019] for further details[2].

[2]Data will be openly available under EBRAINS knowledge graph after publication.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data acquisition

We collected the fMRI data previously at the Centre for Cognitive Neuroimaging, at the University of Glasgow. Participants gave written informed consent to participate, in accordance with the institutional guidelines of the local ethics committee of the College of Science & Engineering at the University of Glasgow (CSE01127). We used EPI sequences to acquire partial brain volumes aligned to maximise coverage of the visual pathway (18 slices; voxel size: 3mm, isotropic; 0.3mm interslice gap; TR $= 1000$ms; TE $= 30$ms; matrix size $= 70 \times 64$; FOV $= 210 \times 192$mm). We used retinotopic mapping data ([Sereno et al., 1995, Wandell et al., 2007]) to locate V1 and V2 (polar angle and eccentricity mapping). Additionally, we used three flashing checkerboard locations to map cortical subregions of V1 and V2 corresponding to occluded and non-occluded portions of the visual field. The first checkerboard was located in the bottom-right visual quadrant ('Target mapping' condition), the second was located only at the border ('Surround mapping' condition; extending 2° visual angle into the occluded quadrant) and the third covered the remaining three quadrants ('Non-Occluded mapping' condition). We declared cortical subregions that responded statistically more to the contrast of 'Target' vs. 'Surround' conditions as occluded i.e., not stimulated with scene information. In our non-occluded condition, the visual field was presented with scene information, activating the corresponding retinotopic region of V1 and V2.

Data pre-processing

Functional data passed through different pre-processing steps; slice time, 3D motion correction, and temporal filtering (high-pass), before being normalised to Talairach space. We used data from retinotopic mapping runs to define early visual areas V1 and V2 using linear cross-correlation of eight polar angle conditions. To define the regions of interest (ROI) corresponding to the non-occluded and occluded quadrants, lower left and right image sections respectively, we computed population receptive fields (pRF, [Dumoulin and Wandell, 2008]) and excluded voxels whose response profiles were not fully contained (within 2σ of their pRF centre) by the respective visual ROI. The patterns of brain activities we obtained, which consisted of data from V1 and V2 for each of the two image quadrants analysed, were not pre-processed with any dimensionality reduction technique (contrary to CNN activations that went through a PCA analysis, see below). We discarded the upper quadrants due to the visual content of stimulation images (shown in supplementary). Most of the images included depict landscapes, and the upper parts of the images therefore included large portions of sky (i.e., uniform values).

Artificial neural network models

VGG16: a supervised image classification network

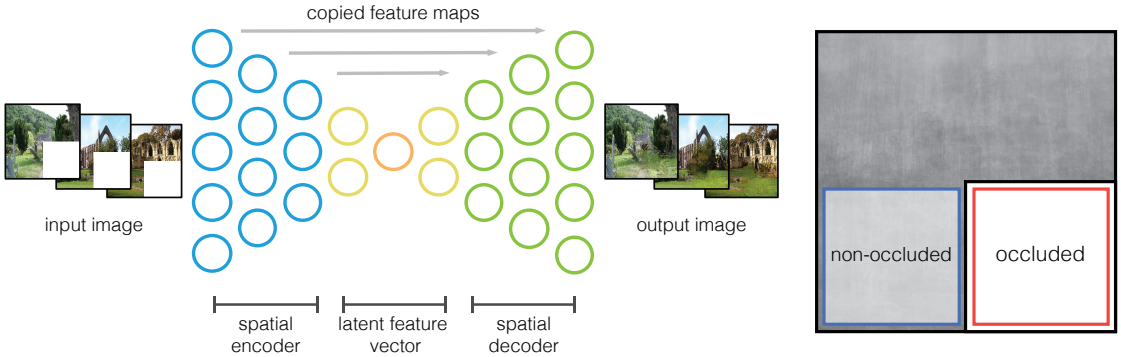
To compare brain data with a supervised network, we used the VGG16 network ([Simonyan and Zisserman, 2014]), a well-known model used for comparing visual pathway activity with CNNs ([Cichy et al., 2016, Güçlü and van Gerven, 2015]). The network is a supervised model ([Simonyan and Zisserman, 2014]) pre-trained to solve image classification - the task to attribute an object class label to an image - on the ImageNet database with 1000 classes[Russakovsky et al., 2015]). The architecture is strictly hierarchical - image-to-labels - with 16 convolutional layers, and it uses repetitive and simple base structure made by 3×3 convolutional layers, for a total of ~ 138M parameters. In our analyses, we used only convolutional layers before pooling, leading to a total of 5 layers analysed (conditions: `conv1_2, conv2_2, conv3_3, conv4_3, conv5_3`).

Encoder/decoder model: a self-supervised image completion network

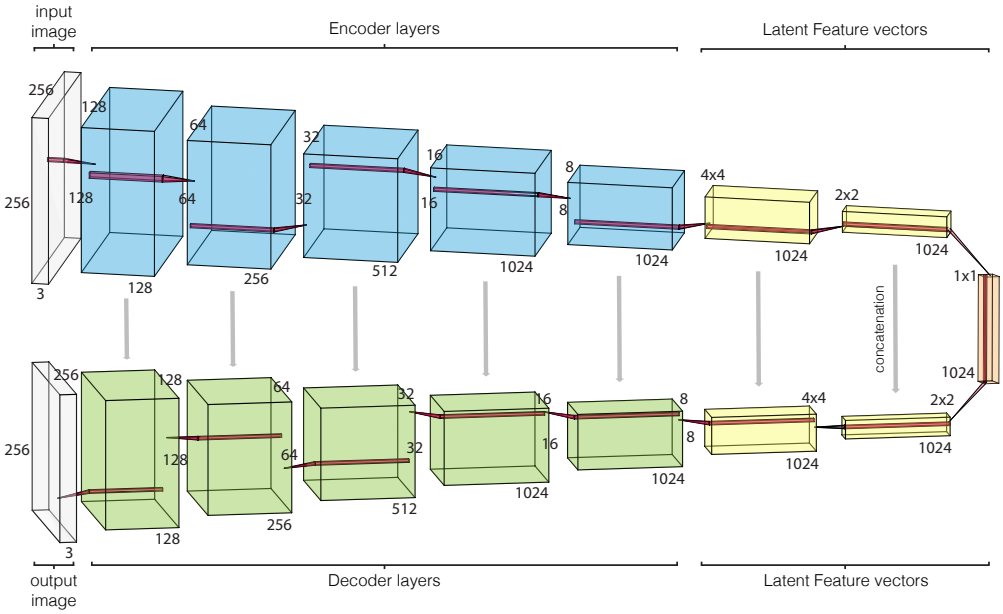
Our self-supervised image-to-image model is trained to reconstruct the occluded image section (always the lower-right quadrant); it is a fully-convolutional neural network, with a encoder/decoder architecture and with skip connections (known as U-Net and described in [Ronneberger et al., 2015]). Details on layer activation dimensions are shown in Figure 2. Every layer implements the same components: a 2-D convolution with a 3×3 filter and stride$= 2$ (that downsamples the activations, no pooling used), batch normalisation, and a rectified linear unit (ReLu) function for

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

encoder and a leaky ReLu[3] for decoder. We added padding to keep the activation dimensions constant after convolution and adopted dropout during the training to reduce overfitting.



(a) Sketch of encoder/decoder architecture and analysed quadrants of testing images. Occluded images on the left are shown as the input layer of the network. The network feeds, from left to right, through convolutional layers (blue, encoding layers), passing by a latent vector feature space (yellow), ending up to an expansion of decoding layers (green). Note that there are copied feature maps (concatenated) linking between encoding and decoding layers.



(b) Layer activation dimensions for encoder, decoder, and latent feature layers in blue, green, and yellow respectively.

Figure 2: Model architecture with layer dimensions. Every layer implements: convolution, batch normalisation, and activation function (leaky relu for encoder and relu for decoder).

We trained the network similarly to what has been proposed in [Isola et al., 2017], in which not only the mapping from input image to output image (i.e., network filters) is learned, but also the best loss function to evaluate (or train) this mapping. The selection of the learning strategy, and in particular the loss, is crucial in unsupervised learning, where labels or categories are not available. Early unsupervised models adopted custom loss functions hand-crafted by the experimenter; for example, autoencoders (AE) use the mean squared error (MSE) between the original image and the generated to compute the loss and the relative gradient. Variational autoencoders (VAE) ([Kingma and Welling, 2013]), a combination of AE with statistical inference, add a term to the MSE loss measuring how closely the latent variables

[3]Leaky Relu: $y = f(x) = 0.6 * x + 0.4 * |x|$ (see figure in Supplementary)

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

match specific distributions (ex. unit gaussian). However, MSE produces blurry images and VAEs fail to generate good-looking images since they are not able to parametrise precisely the complex distributions of images. A successful approach to overcome the hand-crafted selection of a loss function is to let the network learn it: this can be accomplished using conditional generative adversarial neural networks ([Goodfellow et al., 2014]). The training was carried out using images from the SUN database ([Xiao et al., 2010], the same database where the testing images came from), with occluded images as input to the network and original images (without occlusion) as output. We discarded some images from the database because they were too low resolution (less than 200×200 pixels), leading to a total of ~ 107000 images for the training[4]. We trained the model and its 228M parameters with epochs$= 5$ and batch_size$= 5$ and the training process lasted for ~ 49 hours on a GeForce 1080Ti. All the code is implemented in tensorflow. An overview of the training and testing procedures is shown in Figure 3 of Supplementary material.

We extracted layer activations from both encoder and decoder streams, as shown in Figure 1for two quadrants of the image: lower-left (non-occluded), and lower-right (occluded). The contractions in the architecture produce an increase of the (CNN) receptive fields as the layers increase, causing overlaps between quadrants. To avoid confounding results, we selected only an area within the quadrants, removing borders between them. Calculating the receptive field of the network (see supplementary for receptive field sizes), we noted how after the fifth encoder layer, the receptive field size (i.e., the region of the input image which contribute to filter activations) spanned the entire image, and layers representations no longer contained spatiotopic activity maps (this was verified displaying activation maps). It was then not possible to distinguish quadrants by spatial location, so we decomposed the analysis into three conditions: the *spatial encoder*, in which activations could be split into quadrants; *latent feature vectors*, which included activations about the context only and without the spatial information; and *spatial decoder*, with again the concept of spatial separation in reconstruction.

Dimensionality reduction of layer activations: PCA

As described next (Section *Data analysis: Representational similarity analysis (RSA)*), we performed Representation Similarity Analysis (RSA) through predictions of similarities between CNN and early visual cortex representations. Since layer activations from the CNN have large dimensions (see sizes in Figure 2-(b)), we applied a dimension reduction technique to obtain stable and feasible RSA predictions in terms of variance explained (avoiding overfitting) and computation time. We therefore applied Principal Component Analysis (PCA) to every layer activation of the set.

In order to obtain comparable comparisons in the RSA analysis, we reduced every layer activation to the same dimension, for every layer analysed, as done in [Cichy et al., 2016]. The dimensionality reduction is learnt randomly selecting 10000 images from the training set, extracting the corresponding activations, learning the transformation through an incremental PCA ([Ross et al., 2008, Pedregosa et al., 2011]), and eventually applying the transformation on activations extracted from the testing set. We tested different values for the number of principal components (within 2^n, with $n = 3, 4, \ldots, 10$, i.e., $8, 16, \ldots, 1024$ components), as reported in Section *Dimensionality reduction: PCA*.

Data analysis: Representational similarity analysis (RSA)

We computed Representational Dissimilarity Matrices (RDMs) from fMRI data and PCA-reduced CNN activations using the Linear Discriminant Contrast method ([Walther et al., 2016]). For each scene comparison, we fit models to the RDM of the 22 other scenes in the first set (e.g., runs $1/2$ vs. $3/4$) using non-negative least-squares ([Khaligh-Razavi and Kriegeskorte, 2014]). We repeated this for all scene comparisons, producing a predicted RDM based on model parameters, which was then compared to an RDM produced from the other half of the dataset (e.g., runs $5/6$ vs. $7/8$) using Kendall's Tau-a rank correlation ([Nili et al., 2014]). We repeated this procedure for all 70 possible split-quarter combinations, and averaged values over splits to produce one correlation value per subject per ROI and model combinations ([Morgan et al., 2019]). A brief outline of the computation is shown in Figure 3.

[4]The trained model will become openly available under EBRAINS knowledge graph after publication.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

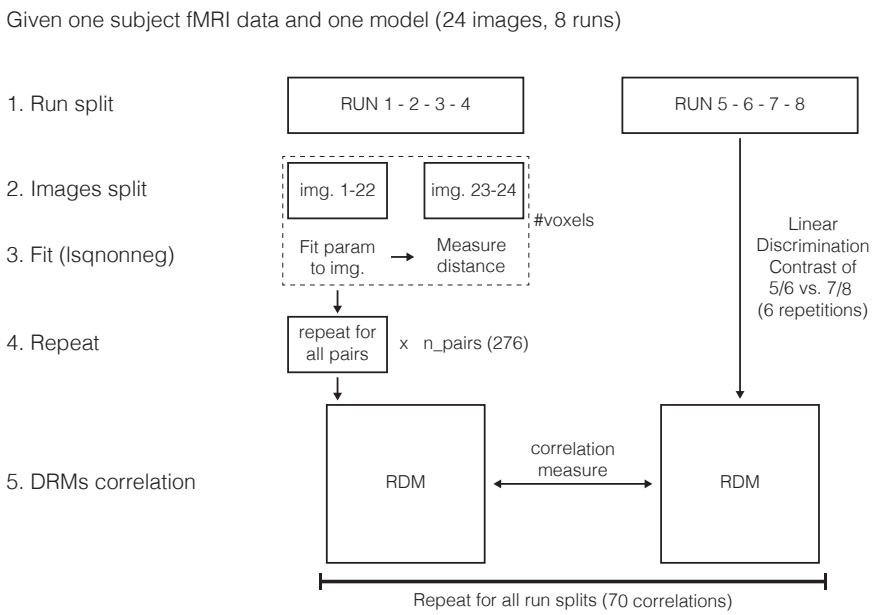


Figure 3: A schematic overview of the RSA computation.

Results

Firstly, we show the visual results of the network output and a description of what layer activations represent (Section *DNN model training*). Then we show similarity performance which drove the selection of the number of PCA components (Section *Dimensionality reduction: PCA*). Lastly, we compare VGG and our encoder/decoder in terms of RSA and encoder/decoder layers (Section *Comparison between VGG16 and encoder/decoder*).

DNN model training

A graphical result to demonstrate the quality of the output of the network is shown in Figure 4.

In order to create a model that might capture fMRI brain activation to occluded regions of images, we trained a network to reconstruct grayscale images with occlusion. Images used in the experiment were not part of the training set.

To better understand the processing performed by the network, we show a selection of stimuli eliciting maximal activations for every layer of the network encoder in Figure 5. Patches were obtained as follows: given activations for a specific channel[5] from 10000 images of the training set, we found the top five activations (maximum responses for that filter). Starting from the location of the maximum, we found the corresponding patch (the receptive field) in the image space that caused that activation. In the classical hierarchical structure of different feedforward networks, such as VGG16 ([Simonyan and Zisserman, 2014]), it is possible to find features from low to high level of complexity. In this network, instead, features appear to be more dedicated to the detection of edges even in middle and upper layers. Only in layer `encoder_6` to `encoder_8` do features start to become more complex, incorporating larger areas of the scene.

The functional characterisation of these blocks can be challenging, but a possible interpretation follows. The model architecture is composed of two branches: encoder and decoder, also known as contraction and expansion paths. The encoder takes in an original image and produces a compressed vector, or latent variable vector; the decoder takes this vector and tries to reproduce the original image. The encoder branch increases, layer by layer, the representation of "what", i.e., the feature complexity, and reduces the "where", i.e., the localisation of a specific feature in the image space. In terms of early visual processing, this is equivalent to compressing retinotopic features into high-level representations.

[5]In this context, a channel - or feature map - is the resulting operation of a single convolution (input image with a single kernel). The number of channels of a layer activation is usually the third dimension after height and width (with 2D input).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

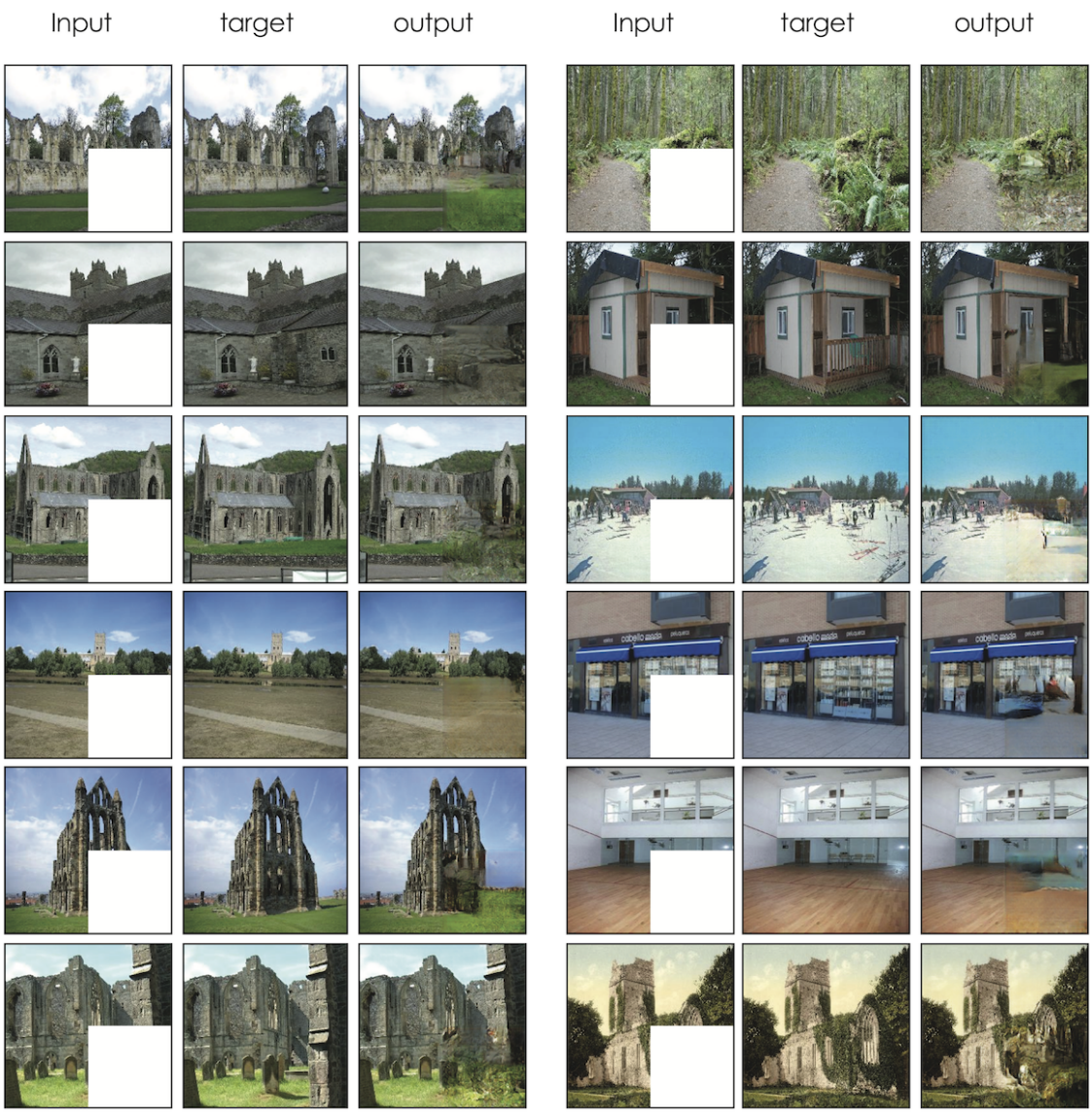


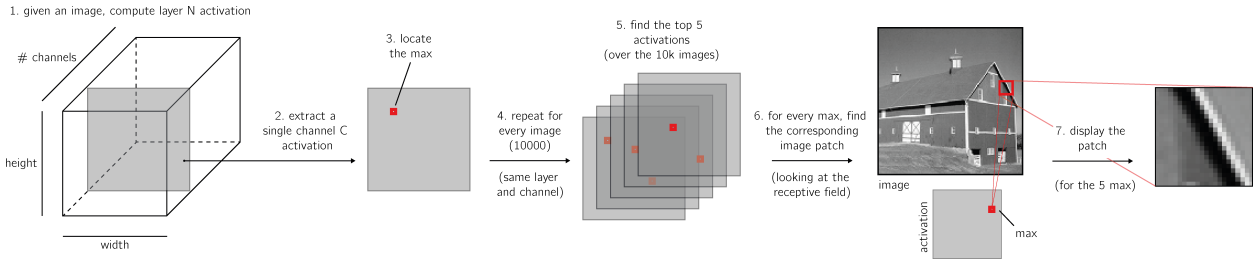
Figure 4: Model results with the input to the network (occluded images), the target used to train the network (i.e., the original image), and the output of the network. Images used in this figure were not part of the training set.

In contrast, the expansion path creates a high-resolution image on the output through a sequence of up-convolutions (i.e., transposed convolution) and concatenation with high-resolution features from the contracting path, i.e. reconstructs retinotopic features from high-level representations. We chose a contraction (similar to what happens with CNNs for object recognition [Krizhevsky et al., 2012]) and an expansion path, in order to accomplish increased computational efficiency, rather than learning a full-size convolutional net, forcing the network to learn features at different scales ([Noh et al., 2015]).

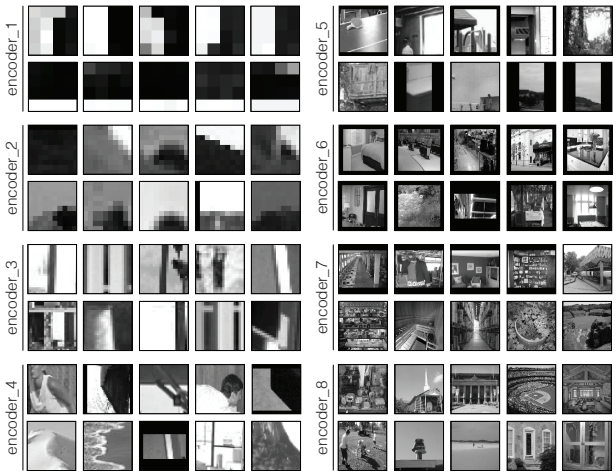
Dimensionality reduction: PCA

Considering the type of RSA conducted (Section *Data analysis: Representational similarity analysis (RSA)*), in order to avoid overfitting and reduce the amount of regularisation needed, we reduced the layer activation dimensions to a lower dimension. Therefore, we applied different numbers of principal components $(8, 16, \ldots, 1024)$, and tested the performance achieved (in terms of similarity, Figure 6). The results describe the mean and the standard deviation of

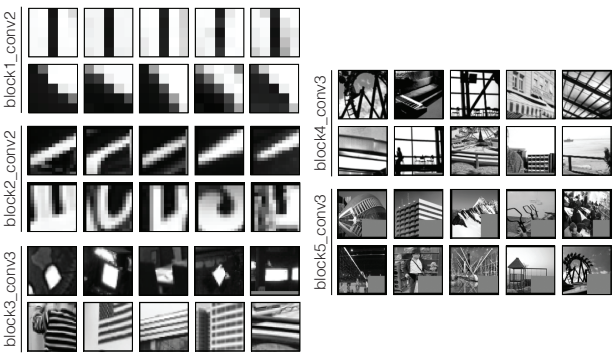
bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) Activation extraction procedure.



(b) Activations display for encoder/decoder network.



(c) Activations display for VGG16 network.

Figure 5: For every layer of a network, two channels are displayed (a row each). The five images correspond to the top five activations for a specific channel, highlighting common visual structure. These images are randomly selected from a bigger pool of analysed layers - the remaining are displayed in the Supplementary Material.

similarity across every layer analysed - between V1 and CNN layer RDMs - for every number of components, in terms of Kendall's tau-a. Results are averaged across all subjects. The two conditions are non-occluded and occluded: the difference is the relative portion of the cortex analysed through the selection of the correspondent receptive fields (pRF analysis). In the non-occluded condition, we analysed the corresponding region of retinotopic cortex that received feedforward stimulation from the image; in the occluded, the region of cortex that processed only the occluder.

As expected, with increasing the number of components, correlations become higher and more stable for both V1 and V2. Based on this finding, we kept the number of components equal to 1024. This allowed us to explain the most possible variance in CNN layer activations. Note that 1024 is the maximum reachable number of components, as it is equal to dimensionality of the smallest layer (see Figure 2-b). All results are reported using this number of components; results for VGG are here omitted for brevity, but the same logic is applied.

Comparison between VGG16 and encoder/decoder

We assessed the ability of the two models to describe brain data using RSA. The first was VGG16 - a supervised network designed for image classification. The second was our encoder/decoder scheme - a model trained in a self-supervised fashion to fill in an occluded portion of images. In Figure 7 we show VGG16 and encoder/decoder response similarities with V1 and V2 in terms of Kendall's Tau-a rank correlation.

Focusing on the occluded quadrant, we observed a significant difference between VGG16 and our model [$p <= 0.0097$] (differences within our model between encoder and decoder are reported below). This result shows how our model, trained in a self-supervised way to perform inpainting, has representations more similar to the brain data compared to

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

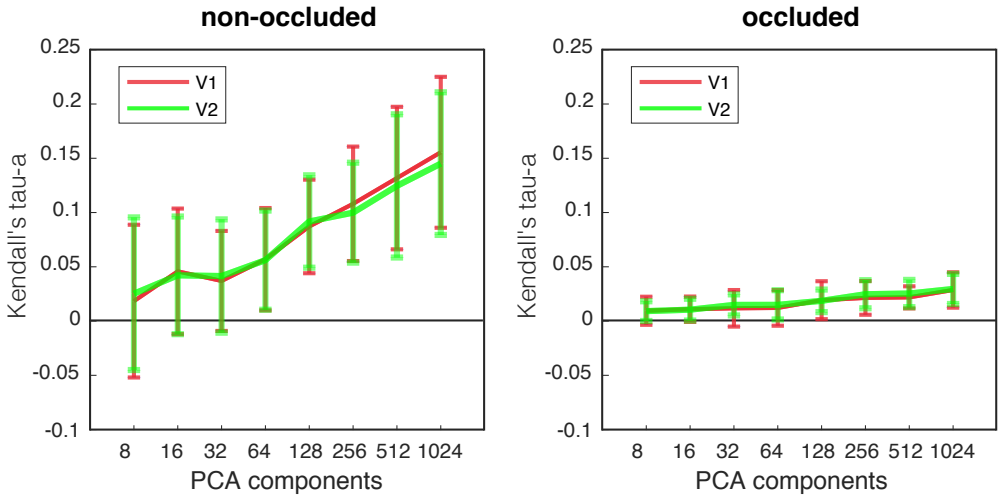
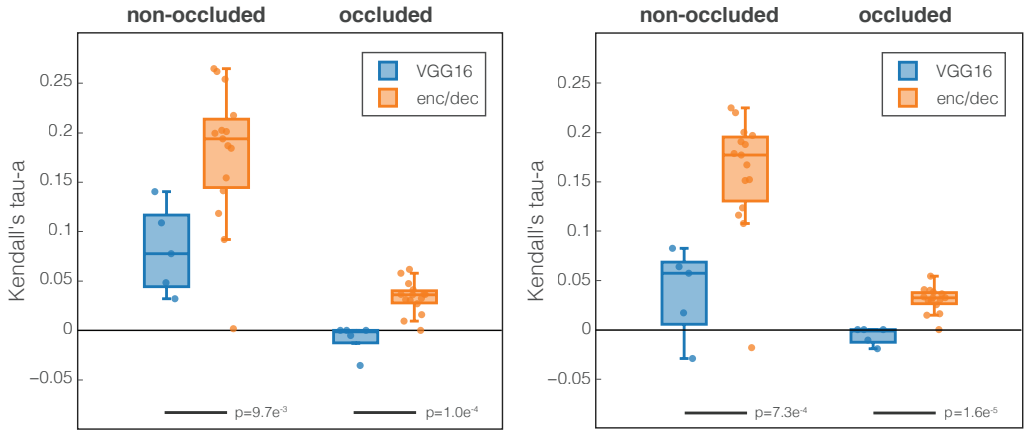


Figure 6: Similarity results (between V1 and CNN layer RDMs) for the non-occluded and occluded quadrants in terms of Kendall's tau-a. For every component in the range 2^n, with $n = 3, 4, ..., 10$, we report mean similarity and standard deviation averaged across layers and subjects.



(a) V1 (b) V2

Figure 7: Kendall's Tau-a rank correlation between VGG16 (blue) and our model (orange) with visual areas (a) V1 and (b) V2. Every dot is a CNN layer (5 conv for VGG16 and 15 for encoder/decoder); results are averaged across subjects. We performed a t-test to determine statistical significance of the difference between the models; results are reported below the bars.

VGG16's, a trained supervised model for object recognition. In the non-occluded area, a cortical region combining feedforward and feedback processing, we notice a remarkable increase of similarity between the brain data and our model. We therefore have greater similarity to brain data with our network representations rather than VGG16, for both occluded and non-occluded areas.

Encoder/decoder layers detail

Focusing on our model, we report results for every layer in Figure 8, analysing differences between encoder (or contraction path) and decoder (or expansion path) parts. We decompose results in the three conditions, *spatial encoder* (blue), *latent feature vectors* (yellow), and *spatial decoder* (green), as described in Section *Encoder/decoder model: a self-supervised image completion network*.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

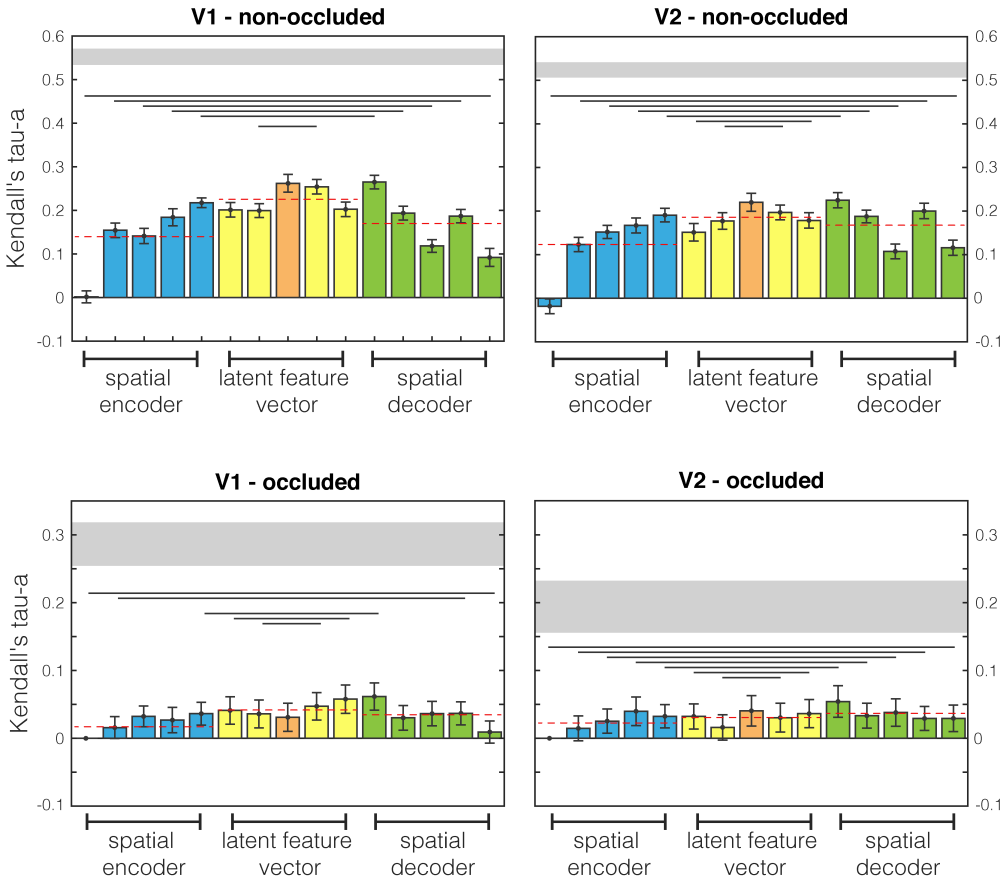


Figure 8: Comparison between brain and encoder/decoder network RDMs. First and second rows are different human visual areas, V1 and V2 respectively. First and second columns relate to non-occluded and occluded quadrants. Results are highlighted with different colours for spatial encoder, latent features, and spatial decoder conditions. Dashed red lines are mean values for the three conditions. Lines below bars show when the encoder and decoder populations difference is significant (Wilcoxon signed rank test).

Four plots are shown: the first row displays V1 results and the second row displays V2; the two columns correspond to non-occluded and occluded quadrants. Every bar is an analysed network layer and colours highlight the three different conditions: encoder (blue), latent feature vector (yellow, with the middle layer in orange), and decoder (green). The bar indicates the mean value across subjects, and the error bar the standard error (standard deviation divided by the square root of the number of subjects - i.e., 18).

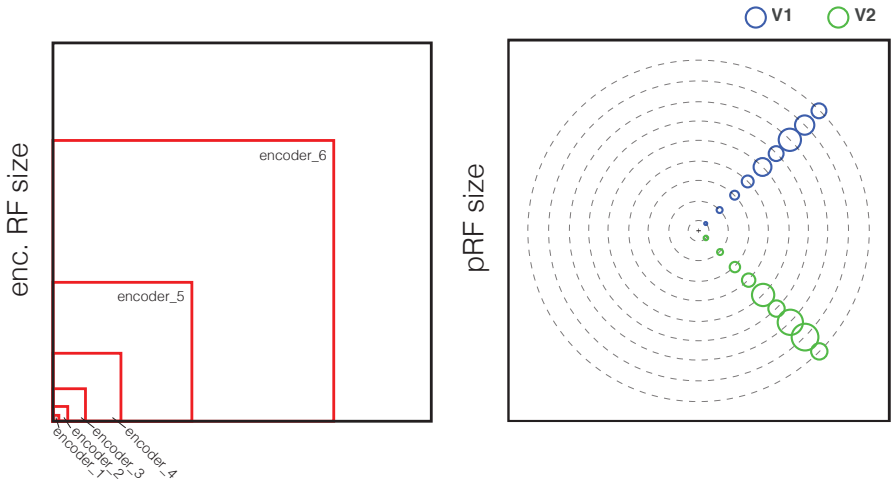
Grouping layers together we can highlight large differences between conditions. As depicted by dashed lines, decoder layers (which integrate information from the latent vector space -middle level features- and the correspondent encoders -lower level features-) have greater similarity to brain data with respect to encoders. Surprisingly, the latent feature vector is even higher than decoder, showing the highest similarity with brain data. Analysing layer by layer, we notice that progressive encoding layers (blue bars) become increasingly better descriptions of brain activity patterns. Once the encoder is devoid of retinotopic organisation due to the absence of spatial information, there are the latent feature vectors (yellow bars), which only have "context", or higher-level features. With these layers, we have an immediate drop in performance with the first layer after the retinotopic layers, which often follows an increase of similarity. In particular, it seems that in the occlusion, for V1, there is a decrease and increase of performance. The orange bar represents the inner layer activation (higher-level) that in some cases represents the higher activation among the context layer. This is until the decoder layers with spatial information (green bars), in which the network reobtains retinotopic organisation. Here, after an initial increase, there is a stabilisation of similarity with brain activity patterns.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Interestingly, the first decoder layer is always the most similar to brain activity patterns. This is the first retinotopically organised layer after integrating the "context" information from latent feature layers. This surprising result opens a series of possibilities for future modelling of brain data from early visual cortex using deep neural networks, such as using a multi-scale approach or adding more complexity than simple edge detection to model V1 processing.

Receptive field analysis

We investigated whether there was a relationship between the size of V1-V2 receptive fields and encoder/decoder network receptive fields. In Figure Figure 9, we show the receptive field analysis for (a) the CNN encoder layers and (b) the fMRI 3T data. The term receptive field (RF) has here two different meanings, based on the data on which it refers to. On fMRI data, it specifies the size of the visual field that each voxel captures; this is why voxels processing the fovea region have smaller RF size than the peripheral ones. Instead, in CNNs, every unit in a convolutional layer only depends - and processes - a specific region of the input image, called the receptive field. Note that every network layer processes the output of the previous layer, not the original image. In the encoder branch, since "arriving" skip connections are not present (only "departing"), every layer elaborates already processed information because this branch is a cascade of consecutive layers.



(a) encoder/decoder RF (b) fMRI 3T pRF

Figure 9: Receptive field analysis. (a) - Receptive field size for the network, encoder layers only, and for (b) the fMRI 3T data. See supplementary for the pRF analyses.

In the Figure 9-(b), we can observe how the RF size of the fMRI data is increasing from the fovea to the periphery (we instructed participants in a central fixation task during the fMRI experiment), for both V1 and V2 (range $= \{3, 16\}$ pixels for V1 (blue) and $\{3, 21\}$ for V2 (green)). These data show high similarity with the RF sizes of the first three network layers (`encoder_1` to `encoder_3`, respectively 4 to 22 pixels, see Supplementary for a full list). This may appear in opposition with the results shown in Figure 8, where the similarity increases going from `encoder_1` to `encoder_6`. This means that the increase of similarity further in the network is not due to the matching of the receptive field sizes, but to something else. We can read this result as an evidence of the integration of mid- and low/middle-level features in early visual cortex, since in the decoder layers there is an integration of two different stream of processing: bottleneck and skip connections information.

Discussion

We investigated the similarity in representations between an artificial neural network trained to fill occlusions and early visual cortical fMRI activity acquired from humans viewing occluded images. Specifically, we wanted to investigate

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

whether an artificial neural network with encoder/decoder architecture would better approximate brain data acquired during a task involving cortical feedback signals, than a purely feedforward network. We have shown two main findings. First, training an artificial network to solve the same task as humans revealed increased similarity to brain data compared to a generic supervised network for classification. Furthermore, the CNN decoder pathway was more similar to brain processing than the encoder pathway. In fact, there was an increase in similarity going further along the network, in line with the integration of multiple feature levels in early visual cortex.

Comparison between VGG16 and encoder/decoder network

As already speculated ([Hassabis et al., 2017]), a possible common ground for the AI and neuroscience communities is to challenge AI to replicate brain processing while performing common tasks. Having an AI solution could thus lead to better models of the brain for those tasks. With this idea in mind, we trained a network to solve inpainting and compared it to human brain activity performing the same operation. We first compared a supervised network for image classification (VGG16) and an unsupervised network trained on the same task (our model) in terms of brain similarity. In both occluded and non-occluded image locations, we obtained a higher similarity between our model activations and fMRI data; an expected result, considering the features learnt by the CNN. We show in Figure 7 how our model outperforms VGG16, meaning that our model represents images more similarly to the brain in both non-occluded and occluded quadrants.

Encoder/decoder layers detail

What can our model tell us about brain processing? Our model attempted to reconstruct the occluded image section (lower-right quadrant) using contextual information from the image surround (i.e., the remaining three quadrants) through several layers of processing. The model is composed of two branches, encoder and decoder, which analyses images and reconstructs the missing part in the decoder. The network mainly learns low-level features, including edge detection and texture pattern recognition, at different resolutions (receptive fields), since these are the features considered useful to solve the task (i.e., reconstructing the image).

Going further with the analysis of our network, we tested whether the encoder (i.e., the contraction path) or the decoder (i.e., the expansion path) was more similar to the brain's representations. We found that the branch of the network more similar to the early visual cortex was not the portion compressing retinotopic features into higher-level representations (i.e., the forward encoder pathway), but instead the portion reconstructing retinotopic features from higher-level representations (i.e., the decoder pathway). Predictive coding theories ([Friston, 2008]) suggest that the brain is trying to reconstruct the image under occlusion, using information in cortical feedback to probabilistically infer the content of missing image sections. Partial or complete occlusion, of objects for example, is commonplace in our visual environments. Predictive coding offers a neuronal mechanism by which the brain can minimise surprise by facilitating the negotiation between top-down predictions and feedforward input once the occlusion is removed. A small warning signal in a cluttered environment can become better detected if predicted information is filtered out. Consistent with such a neuronal framework, our data confirms some level of filling-in with information related to inpainting in computer vision.

Future directions

In future, it will be important to continue to improve network descriptions of the human visual system by employing more descriptive and specialised networks. Relatedly, we recently showed how line drawings provide a good description of internal model structure representing scene-specific features in human early visual cortex [Morgan et al., 2019]. Being able to build a network able to predict sketches from an image would therefore provide additional value when comparing CNNs to brain data. In computer vision, hand sketches have been extensively studied, for example in sketch recognition, generation, and sketch-based image retrieval [Riaz Muhammad et al., 2018], and reveal that computers can classify line drawings in addition to digital images.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

In terms of improving the biological plausibility of neural networks, recent work has also shown that CNNs with recurrent connections exhibit superior performance when recognising occluded and non-occluded objects ([Spoerer et al., 2017]) and that recurrent connections can help to describe cortical dynamics in early visual cortex ([Kietzmann et al., 2019]). Our plan for the future is to advance upon these important findings by training recurrent connections to predict occluded portions of images as well as behavioural sketches.

Along the same lines, a recently developed network has accurately predicted visual saliency using a similar encoder-decoder network ([Kroner et al., 2020]). Since we know the human visual system's coding of saliency must be robust to occlusion and clutter, it would be interesting to compare network objectives of predicting occluded visual features and predicting saliency. Such comparisons would allow us to understand whether aspects of these tasks could be shared by the same neuronal pathways.

Conclusions

We investigated the representational similarity between a specialised artificial network built to reconstruct occluded images and fMRI data obtained while human subjects viewed the same stimuli. Results suggest that low- and mid-level features are present in early visual cortex (V1 and V2). Optimising models to characterise feedback signals in human cortex will improve our understanding about the computations in early visual cortex where we know there is rich top-down information predicting feedforward input, that is currently not captured in feedforward networks. This work points to new experiments in which we challenge AI to replicate as closely as possible brain processing while performing cognitive tasks, testing models explaining memory, visual imagery, and auditory responses in early visual cortex.

Author Contributions

All authors conceptualised the study. M.S. trained the network. A.T.M. collected and analysed fMRI data. M.S. and A.T.M. computed the RSA analysis. M.S. and L.S.P. wrote the manuscript, and all authors edited the manuscript.

Funding

This project has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 720270 and 785907 (Human Brain Project SGA1 and SGA2).

References

[Alink et al., 2010] Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W., and Muckli, L. (2010). Stimulus Predictability Reduces Responses in Primary Visual Cortex. *Journal of Neuroscience*, 30(8):2960–2966.

[Angelucci et al., 2017] Angelucci, A., Bijanzadeh, M., Nurminen, L., Federer, F., Merlin, S., and Bressloff, P. C. (2017). Circuits and mechanisms for surround modulation in visual cortex. *Annual review of neuroscience*, 40:425–451.

[Bergmann et al., 2019] Bergmann, J., Morgan, A. T., and Muckli, L. (2019). Two distinct feedback codes in v1 for 'real' and 'imaginary' internal experiences. *bioRxiv*.

[Chong et al., 2016] Chong, E., Familiar, A. M., and Shim, W. M. (2016). Reconstructing representations of dynamic visual objects in early visual cortex. *Proceedings of the National Academy of Sciences*, 113(5):1453–1458.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[Dumoulin and Wandell, 2008] Dumoulin, S. O. and Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *NeuroImage*, 39(2):647 – 660.

[Edwards et al., 2017] Edwards, G., Vetter, P., McGruer, F., Petro, L. S., and Muckli, L. (2017). Predictive feedback to V1 dynamically updates with sensory input. *bioRxiv*.

[Eickenberg et al., 2017] Eickenberg, M., Gramfort, A., Varoquaux, G., and Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194.

[Erlikhman and Caplovitz, 2017] Erlikhman, G. and Caplovitz, G. P. (2017). Decoding information about dynamically occluded objects in visual cortex. *NeuroImage*, 146(May 2016):778–788.

[Friston, 2005] Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815–836.

[Friston, 2008] Friston, K. (2008). Hierarchical models in the brain. *PLoS computational biology*, 4(11).

[Gilbert and Li, 2013] Gilbert, C. D. and Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14(5):350–363.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

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

[Hassabis et al., 2017] Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

[He and Sun, 2014] He, K. and Sun, J. (2014). the Statistics of Similar Patches. *Pattern Analysis and Machine Intelligence, IEEE Transactions on (2014)*, 36(10):1–13.

[He et al., 2018] He, T., Kong, R., Holmes, A., Nguyen, M., Sabuncu, M., Eickhoff, S. B., Bzdok, D., Feng, J., and Yeo, B. T. (2018). Do deep neural networks outperform kernel regression for functional connectivity prediction of behavior? *bioRxiv*.

[Hong et al., 2016] Hong, H., Yamins, D. L., Majaj, N. J., and DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience*, 19(4):613.

[Isola et al., 2017] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *CVPR*.

[Jing and Tian, 2020] Jing, L. and Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Kar et al., 2019] Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., and DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature neuroscience*, 22(6):974–983.

[Kay et al., 2008] Kay, K. N., Naselaris, T., Prenger, R. J., and Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, 452(7185):352–355.

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

[Kietzmann et al., 2019] Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., and Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[Kingma and Welling, 2013] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[Klink et al., 2017] Klink, P. C., Dagnino, B., Gariel-Mathis, M.-A., and Roelfsema, P. R. (2017). Distinct feedforward and feedback effects of microstimulation in visual cortex reveal neural mechanisms of texture segregation. *Neuron*, 95(1):209–220.

[Kok et al., 2016] Kok, P., Bains, L., van Mourik, T., Norris, D., and de Lange, F. (2016). Selective activation of the deep layers of the human primary visual cortex by top-down feedback. *Current Biology*, 26(3):371 – 376.

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

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

[Kroner et al., 2020] Kroner, A., Senden, M., Driessens, K., and Goebel, R. (2020). Contextual encoder-decoder network for visual saliency prediction. *Neural Networks*.

[Lempitsky et al., 2018] Lempitsky, V., Vedaldi, A., and Ulyanov, D. (2018). Deep Image Prior. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 9446–9454.

[Lindsey et al., 2019] Lindsey, J., Ocko, S. A., Ganguli, S., and Deny, S. (2019). A unified theory of early visual representations from retina to cortex through anatomically constrained deep cnns. *arXiv preprint arXiv:1901.00945*.

[Morgan et al., 2019] Morgan, A. T., Petro, L. S., and Muckli, L. (2019). Scene representations conveyed by cortical feedback to early visual cortex can be described by line drawings. *Journal of Neuroscience*.

[Muckli et al., 2015] Muckli, L., De Martino, F., Vizioli, L., Petro, L. S., Smith, F. W., Ugurbil, K., Goebel, R., and Yacoub, E. (2015). Contextual feedback to superficial layers of v1. *Current Biology*, 25(20):2690–2695.

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

[Noh et al., 2015] Noh, H., Hong, S., and Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528.

[Pathak et al., 2016] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Pennartz et al., 2019] Pennartz, C. M., Dora, S., Muckli, L., and Lorteije, J. A. (2019). Towards a unified view on pathways and functions of neural recurrent processing. *Trends in neurosciences*.

[Qiao et al., 2019] Qiao, K., Chen, J., Wang, L., Zhang, C., Zeng, L., Tong, L., and Yan, B. (2019). Category decoding of visual stimuli from human brain activity using a bidirectional recurrent neural network to simulate bidirectional information flows in human visual cortices. *Frontiers in neuroscience*, 13.

[Rao and Ballard, 1999] Rao, R. P. N. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87.

[Revina et al., 2018] Revina, Y., Petro, L. S., and Muckli, L. (2018). Cortical feedback signals generalise across different spatial frequencies of feedforward inputs. *NeuroImage*, 180(March 2017):280–290.

[Riaz Muhammad et al., 2018] Riaz Muhammad, U., Yang, Y., Song, Y.-Z., Xiang, T., and Hospedales, T. M. (2018). Learning deep sketch abstraction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[Roelfsema and de Lange, 2016] Roelfsema, P. R. and de Lange, F. P. (2016). Early visual cortex as a multiscale cognitive blackboard. *Annual review of vision science*, 2:131–151.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

[Ross et al., 2008] Ross, D. A., Lim, J., Lin, R.-S., and Yang, M.-H. (2008). Incremental learning for robust visual tracking. *International journal of computer vision*, 77(1-3):125–141.

[Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

[Schrimpf et al., 2018] Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., et al. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007.

[Schwiedrzik and Freiwald, 2017] Schwiedrzik, C. M. and Freiwald, W. A. (2017). High-level prediction signals in a low-level area of the macaque face-processing hierarchy. *Neuron*, 96(1):89–97.

[Sereno et al., 1995] Sereno, M. I., Dale, A., Reppas, J., Kwong, K., Belliveau, J., Brady, T., Rosen, B., and Tootell, R. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, 268(5212):889–893.

[Shipp, 2016] Shipp, S. (2016). Neural elements for predictive coding. *Frontiers in psychology*, 7:1792.

[Simoncelli and Olshausen, 2001] Simoncelli, E. P. and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216.

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

[Smith and Muckli, 2010] Smith, F. W. and Muckli, L. (2010). Nonstimulated early visual areas carry information about surrounding context. *Proceedings of the National Academy of Sciences*, 107(46):20099–20103.

[Spoerer et al., 2017] Spoerer, C. J., McClure, P., and Kriegeskorte, N. (2017). Recurrent convolutional neural networks: a better model of biological object recognition. *Frontiers in psychology*, 8:1551.

[Svanera et al., 2019] Svanera, M., Savardi, M., Benini, S., Signoroni, A., Raz, G., Hendler, T., Muckli, L., Goebel, R., and Valente, G. (2019). Transfer learning of deep neural network representations for fmri decoding. *Journal of Neuroscience Methods*, 328:108319.

[Takahashi et al., 2016] Takahashi, N., Oertner, T. G., Hegemann, P., and Larkum, M. E. (2016). Active cortical dendrites modulate perception. *Science*, 354(6319):1587–1590.

[Tang et al., 2018] Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., Hardesty, W., Cox, D., and Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, 115(35):8835–8840.

[Van Essen and Anderson, 1995] Van Essen, D. C. and Anderson, C. H. (1995). Information processing strategies and pathways in the primate visual system. *An introduction to neural and electronic networks*, 2:45–76.

[Walther et al., 2016] Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., and Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage*, 137:188–200.

[Walther et al., 2009] Walther, D. B., Caddigan, E., Fei-Fei, L., and Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *Journal of neuroscience*, 29(34):10573–10581.

[Wandell et al., 2007] Wandell, B. A., Dumoulin, S. O., and Brewer, A. A. (2007). Visual field maps in human cortex. *Neuron*, 56(2):366–383.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.005132; this version posted December 13, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[Xiao et al., 2010] Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE.

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

[Yu et al., 2018] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (2018). Generative Image Inpainting with Contextual Attention. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 5505–5514.