Learning Equivariant Object Recognition, and its Reverse Application to Imagery

Florentine Klepel

6240023

Cognitive Neuroscience

15th July 2022

Rainer Goebel Ibrahim Hashim

Faculty of Psychology and Neuroscience

University of Maastricht

Word Count: 9847

Abstract

This thesis aims to model the visual ventral stream during perception and imagery with the help of *capsule networks*. The proposed network consists of V1 and V2 from CorNetZ, as well as the Capsule Network architecture with the routing by agreement algorithm. The decoder then reverses this architecture to model the feedback activation patterns of the visual ventral stream. While high classification performance is reached, generalisation performance to different sizes, positions, and rotations is maintained. Due to high variability, reconstruction quality was restricted in some conditions. Surrounding information was used in the feedback path for reconstructions so that reconstructions could be correctly classified. Additionally, a pre-trained network was used to reconstruct remapped fMRI activation patterns from higher visual areas. Reconstructions of single-trial imagery data showed significant correlations to physical letter stimuli. The fMRI activation patterns of V1 and V2 and their reconstructions with population receptive field mapping and an autoencoder were related to activation patterns of the network to test biological plausibility. Representational Similarity Analysis and spatial correlations indicated an overlap of information content between the capsule network and the fMRI activations. Overall, this network sets a promising path for increased generalisation ability because of its focus on figure-part relationships while maintaining biological validity. Due to the capsule networks' high generalisation performance and the implemented feedback connections, the proposed network is a promising approach to improve current modelling efforts of perception and imagery. Further research is needed to compare the presented network to established networks that model the visual ventral stream as well as show that it is crucial to model the human brains' generalisation ability in order to model perception and imagery processes.

Contents

Introduction

Deep Neural Networks and Object Recognition

The ventral visual stream is assumed to be relied upon during object recognition in primates (Poggio & Anselmi, 2016). Artificial neural networks (ANNs) modelling the corresponding brain areas are utilised to understand biological vision (Heinke, Leonardis, & Leek, 2022). Even though the human brain performs object classification effortlessly, this level of performance has proven to be difficult to reproduce with artificial systems that are neurobiologically plausible (Poggio & Anselmi, 2016). ANNs showed similar classification performance as the macaque visual system but otherwise could not explain the hierarchical architecture in the cortex (Kubilius et al., 2018). The high complexity of having 100 layers compared to 4 to 8 areas in the visual stream and the lack of recurrent connections are named as main shortcomings of state-of-the-art neural networks to model brain activation during image classification.

Most of these ANNs extract abstract visual properties of the large provided data sets with the help of various layers (Zhao, Li, Zhao, Yan, & Feng, 2017). These computations are highly susceptible to variations in the input characteristics, such as translations, scaling, and rotations. Scale and position invariance can be achieved by scanning images at different positions and scales (Serre, Oliva, & Poggio, 2007). But if unknown transformations are introduced in the testing data, the performance decreased strongly. This lack of invariance representation can be overcome by data augmentation or feature averaging (Lyle, van der Wilk, Kwiatkowska, Gal, & Bloem-Reddy, 2020). Zhao et al. (2017) used marginalisation over transformation parameters to increase insusceptibility to transformations. Other computational models such as neocognitron, and HMAX achieve some level of invariance by alternating feature detector layers with layers that perform local pooling and subsampling of the feature maps (Goodfellow, Lee, Le, Saxe, & Ng, 2009). These ANNs implement restricted forms of invariance but often lack neurobiological validity since their main goal is to reach fast and accurate object recognition which is why they are less useful in explaining neural activity during object recognition in humans.

Serre et al. (2007) argue that the human visual system can compute representations of images

that are invariant to the most common image transformations. They hypothesise that the main computational goal of the ventral stream during development might be to learn how objects transform. This gives the visual system the ability to compute versions of new images that are automatically invariant to the same transformations. That indicates that the ventral stream neuronal activity is shaped by invariance representations (Serre et al., 2007). Relevantly, current ANNs still need a large amount of training data which is unrealistic regarding human learning processes. That might indicate that they do not exploit invariance or the relation between features to a large enough extent yet (Bowers et al., 2022).

Convolutional Neural Networks (CNNs), a subcategory of ANNs, reach high performance on automatic visual recognition tasks. They contain successive layers of convolution and pooling with the lower areas explaining the activity in V1, V2, and V3 while activity in higher-level areas in the ventral visual stream are better explained by higher layers (Ramakrishnan, Scholte, Lamme, Smeulders, & Ghebreab, 2015). Even though models explaining V1, V2, V4, and IT (inferior temporal cortex) data were built, their performance is less robust regarding image degradations such as contrast reduction or additive noise, compared to human performance (Geirhos et al., 2017). CNNs and ANNs in general lack performance when dealing with flexible and highly variable objects, and when background noise is high (Poggio & Anselmi, 2016).

Peters, Reithler, and Goebel (2012) hypothesise that implementing algorithms derived from neurobiological findings might increase the accuracy of computational models. Research made clear that V1 activity of the visual pathway is well described by Gabor-like edge detectors. That is why the same object can elicit very different neural activation patterns in this area when presented in different locations or at different rotation levels (Peters et al., 2012) but invariant representations emerge in final stages of the ventral stream. Even though it was shown that an explicit gradient for feature complexity exists in the ventral pathway (Güçlü & van Gerven, 2015), it is unclear which specific function is fulfilled by which area in the ventral stream. Multiple transformations are likely performed at the intermediate stages (Peters et al., 2012). When investigating the higher-level activity of this stream, Kar, Kubilius, Schmidt, Issa, and DiCarlo (2019) found improved predictions of late IT neural unit responses

in shallow recurrent CNNs compared to feedforward-only deep CNNs. They argue that recurrent circuits are crucial for rapid object identification since primates were able to outperform feedforward-only deep CNNs for classifying challenging images that required additional recurrent processing beyond the initial feedforward IT response. This finding illustrates that implementing neurobiological findings into ANNs might in turn lead to improved theories about the neurobiology of the human brain (Peters et al., 2012).

Capsule Networks

Sabour, Frosst, and Hinton (2017) proposed *capsule networks* (CapsNets) which might be able to overcome some of the already named shortcomings of conventional ANNs. Their network does not rely on pooling for information reduction and invariant object classification. Instead, the relative position of features is used to detect samples of categories. They showed that a group of neurons in the hierarchically uppermost layer can represent both the probability and the instantiation parameters of detected entities so that a detailed reconstruction from solely the output vector is possible. The probabilities are invariant representations of the stimuli whereas the instantiation parameters reflect equivariance. Equivariance refers to the same sample eliciting different activation patterns depending on physical features of a specific stimulus, such as position or size (Qiao et al., 2018). CapsNets incorporate an equivariant representation whereas ANNs only represent invariance at their highest-up layer.

It was shown that viewpoint changes by two pixels in each direction lead to simple linear effects on the pose parameters representing the relationship between an object part and the whole (Mazzia, Salvetti, & Chiaberge, 2021). Classification accuracy to unseen affine transformations was higher than with a traditional convolutional model (Sabour et al., 2017). Overall, CapsNets have a high intrinsic generalisation capability because they rely on object-part relationships instead of feature filters.

The CapsNet includes a routing-by-agreement algorithm which computes and adjusts predictions depending on lower-level capsule predictions of parent-capsule outputs and the coefficient between their actual output and the parent capsule output (Pucci, Micheloni, & Martinel, 2021). With the help of this iterative routing process, parts are assigned to wholes.

This mechanism also made it possible for the network to recognise multiple objects in an image even when the objects overlap.

These networks can fulfil similar tasks as conventional ANNs. They have already been used for diverse tasks, such as detecting fake images and videos, recognising human movements, learning time information from spatial information, encoding facial actions, classifying hyperspectral images, finding relationships during natural language processing, classifying emotions, action video segmentation, as well as predicting Alzheimer disease, classifying apoptosis, identifying sign language, and classifying brain tumour types (Huang & Zhou, 2020). Additionally, they were also used for understanding the human brain. Qiao et al. (2018) demonstrated that the reconstruction of capsules fed and trained on fMRI perceptual data is highly accurate. This indicates that there is a common space between these highest level capsule activations and the human brain activations during number perception. Only categories “6” and “9” of the MNIST digit data set were taken into account.

Perception and Imagery

Activation patterns during visual mental imagery in early visual areas have been shown to resemble those elicited when sensory stimuli were presented to the participant. The pathway that is used to project perceptual activity upwards in the ventral stream, is assumed to project backwards during mental imagery (Pearson, 2019). Senden, Emmerling, Van Hoof, Frost, and Goebel (2019) investigated mental imagery and perception of letters in healthy participants. With 7T functional magnetic resonance imaging (fMRI), they investigated whether imagined letter shapes can be reconstructed from V1, V2, and/or V3 activity. Participants were instructed to imagine shapes of the letters H, S, C, and T. Population receptive field (pRF) mapping and an autoencoder were used to transform, augment, and classify the data for the analysis. They showed that the geometric profile of the letter was preserved and could be decoded from early visual areas during imagery. The same early visual areas were relevant to decoding the letter shape during perception and imagery.

It is assumed that top-down feedback processes from higher-level areas are responsible for projecting activity down to early visual areas during mental imagery (Pearson, 2019). Similar

activations during perception and imagery were seen in occipital, parietal, and frontal brain areas with an increased overlap in higher-level visual areas (Dijkstra, Bosch, & van Gerven, 2019). Dijkstra, Ambrogioni, Vidaurre, and van Gerven (2020) illustrated the feedback imagery process with the help of magnetencephalography (MEG) and a retro-cue task. Participants were presented with two images consecutively, and a cue indicating which one to imagine. With the help of this paradigm, they demonstrated that the information flow during perception is mainly feedforward from lower-level areas to higher-level areas but also included alternating feedback processes. In contrast, perception processes are reactivated in reverse order and mainly top-down feedback processing is involved in imagery. Each feedforward connection in the visual pathway is matched by a reciprocal feedback connection from higher-to-lower-level areas (Gilbert & Li, 2013). Neurons are able to adapt their functioning in a state-dependent manner. They are assumed to constitute processors that adapt their behaviour depending on feedback from higher-order cortical areas (Gilbert & Li, 2013). Conventional ANNs mainly contain feedforward connections even though the number of feedback connections is outnumbering the number of feedforward connections in early visual areas (Vetter, Smith, & Muckli, 2014). Feedback and feedforward weights need to be separated to overcome the unrealistic symmetry in connections between layers that is implicit in feedforward-only networks that use backpropagation for training (Amit, 2019). It has been additionally shown that both local and feedback recurrent connections lead to better performance in more challenging tasks with a better match to neural data, especially during later time points in the response (Lindsay, 2021).

Svanera, Morgan, Petro, and Muckli (2021) showed that a stronger focus on feedback processes can be modelled with a network consisting of an encoder and a decoder structure. FMRI activation patterns during the corner-occlusion paradigm introduced by Smith and Muckli (2010) were more closely related to a CNN with an encoder and decoder than to the feedforward-only VGG16 network. Since the occlusion of one corner leads to a blockage of the feedforward stream, feedback signals could be isolated. They point out that some neuronal activity in early visual areas is not related to sensory information but rather to the brains' inferences about the world transmitted via feedback pathways to early visual areas, and that

this constraint should be accounted for in biologically constrained neural networks (Svanera et al., 2021).

The presented study aimed to design a neural network that showed high generalisation performance while maintaining biological validity. Traditional CNNs demonstrated difficulties in generalisation to novel viewpoints resulting in exponential inefficiencies when dealing with affine transformations (Sabour et al., 2017). Having said that, they were very well able to explain lower level activation patterns in fMRI studies (Kubilius et al., 2018). On the other hand, CapsNets had a high generalisation ability by using intrinsic spatial relationships to constitute viewpoint-invariant knowledge about an object (Sabour et al., 2017). This was assumed to be a characteristic that is comparable to higher-level area activation patterns of the human brain during object classification. The CapsNet had been implemented by Sabour et al. (2017) as a shallow network with only three layers which is unlike the human ventral stream which includes computations in a minimum of four layers (Güçlü & van Gerven, 2015; Serre et al., 2005). This knowledge resulted in the presented suggestion to implement a network which incorporates convolution and pooling in lower-level areas as in CNNs, and more complex computations as in CapsNets to achieve invariance in higher-level areas. The network comprised V1 and V2 of Kubilius et al. (2018) and added the entire CapsNet structure from Sabour et al. (2017). Feedback connections have not been implemented in traditional object recognition ANNs but are implicated in perception and imagery processes. These missing feedback connections might be a major shortcoming that resulted in ANNs deficiency in explaining neurobiological data. That is why feedback was modelled in the proposed network with the help of a decoder network which closely resembled the encoding layers in reversed order.

The network's generalisation performance for unseen translations was expected to be high and was tested for size, location and rotation generalisation. The model was also tested with an occlusion paradigm (Smith & Muckli, 2010) to see whether it functioned within the biological constraint of feedback connections using surrounding information. The biological validity was also tested by comparing the network to the imaging data collected by Senden et al. (2019). Feedforward and feedback activations could be distinguished in this data set. Therefore, it was

suitable to test the hypothesis that the proposed neural network might model the human ventral stream during perception and imagery.

For a biological comparison to brain imaging data, two approaches were used. Firstly, since capsule activation patterns are assumed to be related to higher-level activation patterns in the human brain during perception and imagery, the fMRI activations in higher-level visual areas was mapped onto the capsule activation patterns of the pre-trained network, and the reconstructions of the trial-wise imagery activations were compared to the input stimuli. Formerly, only reconstruction from perception activation patterns were analysed (Qiao et al., 2018). Secondly, correspondences between imagery and perception activations of the CapsNet and the human brain activations were evaluated for lower-level areas. Formerly, it has not been tested whether a network incorporating capsules aligns with activation patterns in the human ventral stream during image perception and imagery. It was tested whether an adapted capsule network which is inspired by former biological findings could be utilised to achieve high generalisation performance while holding up to relevant biological constraints.

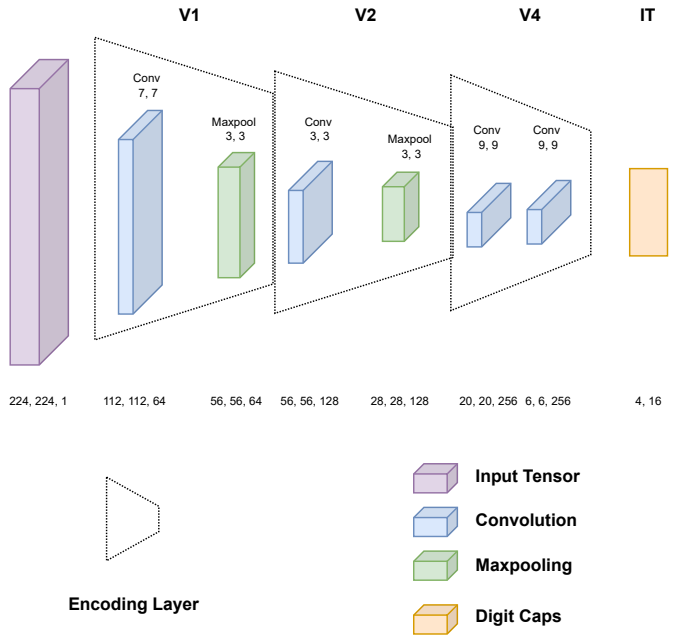
Methods

Capsule Networks

Structure of the Proposed Network. The computational model used in this study is a combination of the commonly used CorNetZ (Kubilius et al., 2018) and the capsule network architecture (Sabour et al., 2017). V1 and V2 were derived from the former. These areas were both represented by a convolutional layer followed by a maxpooling layer. V4 and IT used structures from the capsule network which included another convolutional layer, a primary capsule layer, and the so-called digit capsule layer. The primary capsule layer consists of an additional convolutional layer which utilised the squash function. The digit capsule layer contained in the proposed model four category capsules which consisted each of 16 dimensions. For more details see Figure 1. After this encoding part, a dense layer was added. This dense layer consisted of 64 neurons and used a RELU activation function. The decoder consisted of these calculations in reversed order. That is why a function was applied that reversed the squash equation, see Equation 1. Afterwards, a trainable matrix was initialised

Figure 1

Encoder Architecture of the CapsNet



Note. Depiction of the encoder architecture of the used neural network. V1 and V2 each consisted of a convolutional and a maxpooling layer consecutively. V4 consisted of a convolutional layer and the primary capsule layer which in itself also is a convolutional layer with the squash function applied to it, whereas IT was assumed to only consist of the so-called digit capsules.

which served as reversed routing weights. The beforehand calculated unsquashed predictions were then matrix multiplied by the inverse of the reversed routing weights to get to the former predictions of the digit capsule activations. A matrix of the size of the transformation matrix which was needed in the encoder to predict the output of the digit capsules was initialised afterwards. The inversed transformation matrix was multiplied by the predicted activations of the digit capsules. These raw primary capsule activations were then deconvolved with a filter with a size of 9×9 and a stride of two. Another deconvolution layer of size 9×9 and a stride of one was applied. Afterwards, a 2×2 upsampling layer was used. The next layer

deconvolved with a filter of size 3×3 and a stride of one. The next upsampling layer had a filter size of 2×2. Lastly, another deconvolution layer with a size of 7×7 and a stride of two was applied.

$$s = \pm \sqrt{\frac{\|\text{squash}(s)\|}{1 - \|\text{squash}(s)\|}} \cdot \frac{\text{squash}(s)}{\|\text{squash}(s)\|} \tag{1}$$

Training and Testing. The network was trained and analysed with the letters C, H, S, and T extracted from the EMNIST dataset (Cohen, Afshar, Tapson, & Van Schaik, 2017). Only uppercase letter stimuli were included. These letters were chosen because they are comparable to the stimuli used in the fMRI study by Senden et al. (2019) that was used to test the networks' biological plausibility. The original CorNetZ structure expected an input image of 224×224 pixels but the EMNIST stimuli are of size 28×28 pixels which is why they were inserted into matching-sized arrays. To test the generalisation performance to different positions and sizes, the letters were rescaled and shifted within a 224×224 pixel space. No other data augmentation was applied.

Altogether, after extraction and balancing of the uppercase letters C, H, S, and T from the data set, 11752 (2938 per category) samples for training, 292 (73 per category) for validation, and 2648 (662 per category) for testing were used. The network used the Adam optimizer. The loss was calculated as the sum of the margin loss and the reconstruction loss. The margin loss was calculated in the same way as in Sabour et al. (2017) (see equation 2).

$$L_k = T_k \max(0, m^+ - \|\mathbf{v}_k\|)^2 + \lambda (1 - T_k) \max(0, \|\mathbf{v}_k\| - m^-)^2 \tag{2}$$

T_k was equal to one if the letter of class k is present, otherwise it was zero. As in Sabour et al. (2017), m^+ was set to 0.9, m^- was set to 0.1, and λ was equal to 0.5. The reconstruction loss was calculated as the mean squared difference between the reconstructed picture and the input picture. When adding up the margin loss and the reconstruction loss, the reconstruction loss was scaled down by a factor of 0.0005 to ensure that the margin loss dominated the training. For each test that was run, the network was trained for eight epochs of which the one with the lowest validation loss was saved and used for the analysis. For building this network, Tensorflow 1.7 was used.

Location Generalisation. The first test of this network entailed resizing the letters to three different possible sizes (50×50, 70×70, 85×85 pixels) and placing them in different

positions in a 244×244 pixel image. Training happened on 76.5% of the 14780 possible positions whereas 8.5% were used for validation and the remaining 15% for testing location invariance of the model. The network was trained on the training stimuli in the training positions and tested on separate testing samples in unseen testing positions. To find out about the relevance of the number of dimensions, the number of capsule dimensions was doubled and the tests were rerun. With the standard amount of 16 dimensions, an additional test was run with half the amount of possible locations.

Size Generalisation. The same network structure was trained in a paradigm that tested whether generalisation to unseen sizes of stimuli is possible. Possible letter sizes ranged from 28×28 pixels to 74×74 pixels which accumulates to 46 possible sizes. 76.5% of these sizes were chosen for the training process. 8.5% for validation and the remaining 15% for testing purposes. The classification performance and accuracy can be found in Table 1.

Location and Size Generalisation. For this test, the aforementioned different positions, as well as different sizes, were used and matched. Whenever a resized image could not be placed in a certain location, this location was skipped until a fitting location was found (e.g. most stimuli cannot be presented on the edge since they would only be partially presented).

Rotation Generalisation. With three different locations and three different sizes (50×50, 70×70, 85×85 pixels) of stimuli, a network was trained on rotated images. The images were rotated up to 90° clockwise or counter-clockwise in 1° steps.

Measures. First and second-order correlations were calculated for each network respectively. First-order correlations are depicted as averages as well as separately by letter. They were calculated by the correlation between the binary letter stimulus and the reconstruction stimulus. The second-level correlation metric was defined by correlating two vectors, one depicting the pairwise correlation between physical letter stimuli and the other depicting pairwise correlations between reconstructions.
Additional measures were used to assess the similarity between the images, namely root mean squared error (RMSE), peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and euclidean distance.

Occlusion Paradigm. To test whether the model activations also relate to the results from Smith and Muckli (2010), an adapted occlusion paradigm was used. Smith and Muckli (2010) showed that fMRI activation relating to non-stimulated regions of the visual field could be used to correctly classify the presented category above chance with a support vector machine (SVM) classifier that was trained on activation patterns elicited during perception.
To test whether this holds for the suggested CapsNet, 70×70 pixel stimuli from the test data set were presented in the centre of the image. One of the corners was occluded which resulted in an occlusion area of 35×35 pixels. The reconstructions of the complete stimuli were cut out in the relevant area, an edge of five pixels was left towards the sides that would be later-on non-occluded. The SVM classifier was trained on these corner activations. Afterwards, the occluded corner reconstructions were used to predict their respective stimuli categories with the pre-trained classifier.

Functional Magnet Resonance Imaging Data

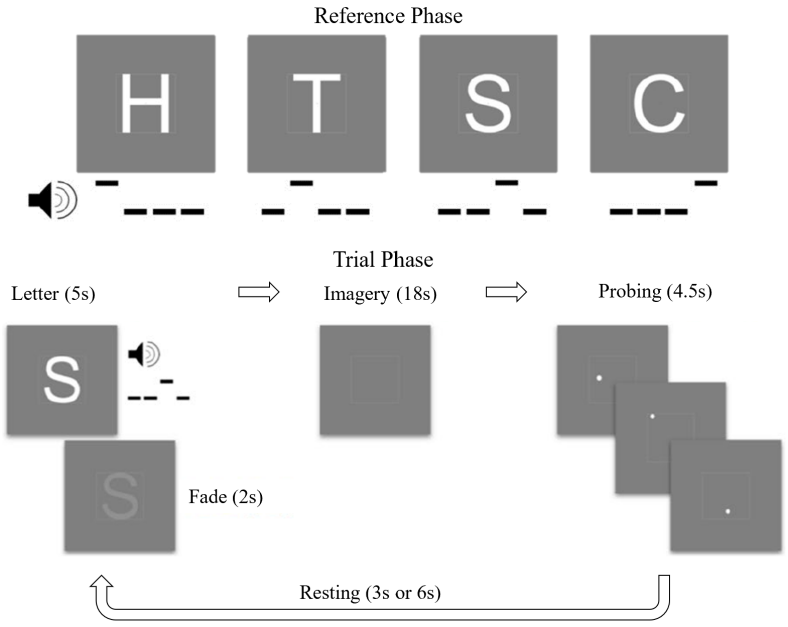
Six participants (two female, four male; average age of 30.7 years) underwent high-field fMRI. For more details on the data set and the processing steps, see Senden et al. (2019).

Stimuli and Task. Three training sessions were completed by each participant prior to a single scanning session which consisted of four experimental (imagery) runs, one control (perception) run and a pRF mapping run.
Training sessions lasted about 45 min and took place one week before scanning. Each training session consisted of the participant seeing one of four white letters (C, H, S, or T) enclosed in a white square guide box on a grey background and a red fixation dot in the centre of the screen. With the onset of visual presentation, an auditory stimulus was presented that comprised three low tones and one high tone. Specific tone patterns were associated with a visually presented letter randomly assigned per participant. At the beginning of each training session, each of the four letters was presented for 3000 ms together with their respective tone pattern. During one training run, each participant completed 16 pseudo-randomly presented trials. Each training session consisted of two training runs in which the reference letters with their tone patterns were presented upon beginning, and two training runs without reference

letter presentation. At the beginning of each trial, after presenting the letter and the tone for 3000 ms, the letter faded out, having disappeared after 5000 ms. Afterwards, the fixation dot changed colour to orange and participants had to maintain a vivid mental image of the presented letter. After 18 s of imagery, the fixation dot turned white and three white probing dots appeared within the guide box, initiating probing. The dots were either located within the letter shape or outside of it and participants had to indicate which one it was. The fixation dot turned green if the answer was correct or red for incorrect answers. The reference and the probing period are schematically represented in Figure 2.

Figure 2

Training Trial



Note. Schematic representation of the training trial, adapted from Senden et al. (2019). During the reference phase, four letters were paired with unique tone patterns. In the trial phase, the tone pattern was played and the letter was shown for 5 s, starting to fade out after 3 s. An imagery period of 18 s and a subsequent probing period of 4.5 s followed. A trial ended with a resting period of 3 or 6 s.

Imagery runs differed from the training runs in probing phase and timing of trial phase. Referencing still took place but was followed by trials that did not contain visual stimulation besides the fixation dot and the guide box. Imagery phases were initiated by tone pattern presentation and the fixation dot turning orange. Imagery phases lasted for 6 s. Each

experimental run contained 32 normal trials, with two additional catch trials with probing. No visual feedback for the probing phase was given.

The perception run was used to measure the brain activation patterns in visual areas during perception of the letters. The same timing parameters as in the experimental run were used. No reference or probing phases were needed. Letters were presented for the duration of 6 s while their shape was filled with a flickering checkerboard pattern. No tone patterns were played during the perception run.

For pRF mapping, a bar aperture revealing a flickering checkerboard pattern was presented in four orientations. In twelve discrete steps, the bar covered the entire screen for each orientation. Within each orientation, the sequence of steps was randomized, and each orientation was presented six times.

Processing of fMRI Data. To test the biological plausibility of the network, the voxel activation patterns of early visual areas V1 and V2 with ROIs from Senden et al. (2019) were used. Furthermore, activations that were extracted from early visual areas, pRF mapped and run through a pre-trained autoencoder were utilised for the analysis. These reconstructions had a size of 150×150 pixels. The relevant part of the reconstructed stimuli from the CapsNet was cut and rescaled in differently sized patches due to the different sizes of the layers. Additionally, activation patterns from the higher visual areas localised by significant activation patterns averaged over all letters were extracted.

Translating fMRI Data into Capsule Space. Since the capsule network would have at its highest layer 4×16 dimensions and higher visual areas indicated 734 relevant voxels, the voxels had to be translated into the digit capsule space. For that purpose, a simple three-layered deep neural network was used. Inspired by Qiao et al. (2018), the first two layers were dense layers with the RELU activation function, whereas the last layer used the squash function adapted from the CapsNet. The first layer contained 256, the second layer 128, and the last layer 64 units—these 64 units were eventually compared to the highest layer of the CapsNet. The model was trained on the perception data, with seven trials from each letter serving as training data, and one trial used as validation data. This network was trained with the Adam optimizer and ran for ten epochs. This pre-trained network was then used to predict

the capsule activation patterns for each imagery trial separately. No additional training took place.

Representational Similarity Analysis. To compare the activation patterns of the neural network and the fMRI activation patterns, a representational similarity analysis (RSA) was used (Kriegeskorte, Mur, & Bandettini, 2008). This method uses second-level multidimensional scaling to relate the representations of information in different modalities to each other.

Representational dissimilarity matrices (RDMs) were calculated with the help of the Python package *rsatoolbox*. They were calculated for V1 and V2 encoder and decoder activations when the artificial letter stimuli from the fMRI study served as input. RDMs based on V1 and V2 decoding activations of the CapsNet were also calculated for reconstructions from remapped activation patterns from higher-level visual areas. Dissimilarities were measured with $1-$correlation. These RDM patterns were then used for the RSA. The RSA compared with the help of correlations the activation patterns in RDMs of the CapsNet with the fMRI data, during perception and imagery.

Results

Location Generalisation

The results of a network that was trained to classify images in different locations are shown in Tables 1, 2, 3, and Figure 3a. The second-order correlation of one hundred stimuli amounted to .88. Reconstructions of ten example stimuli are depicted in Figure 4.

When the number of dimensions was doubled to 32, the classification performance reached 99.81% with a loss of 0.21. When the amount of possible locations was halved to 7390 and the number of dimensions per capsule was kept at 16, the classification accuracy amounted to 99.73% and the loss reached 0.24.

Table 1

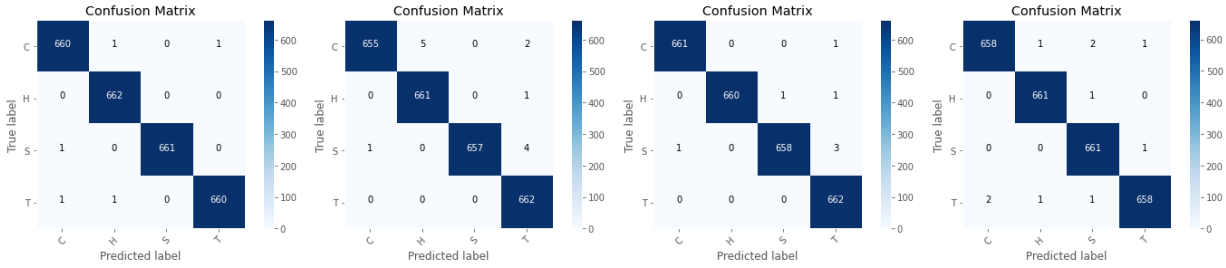
Generalisation Ability of the CapsNet

Measurement	Size	Location	Size & Location	Rotation
Classification Accuracy ↑	99.62%	99.81%	99.51%	99.77%
Loss ↓	0.13	0.24	0.17	0.34

Note. The CapsNet was trained for different tasks, including size generalisation (three different positions with multiple different sizes), location generalisation (three different sizes trained and tested in multiple positions with 16 capsules and all sizes), size and location generalisation at the same time, as well as rotation generalisation (rotation up to 90° in each direction). Training happened on 76.5% of the data with 8.5% for validation and 15% testing data. The respected tested specificities were not presented during training. Classification accuracy and loss calculated as the sum of margin loss and reconstruction loss for each of these tasks are presented. (↑ indicates the more the better, ↓ indicates the less the better)

Figure 3

Classification Performances for Generalisation Tasks



(a) *Position Generalisation*

(b) *Size Generalisation*

(c) *Position and Size Generalisation*

(d) *Rotation Generalisation*

Note. Classification performances for the differently trained networks (position generalisation with three sizes in multiple positions, size generalisation with three positions in multiple sizes, a combination of both, and rotation generalisation with rotations up to 90° in either direction) depicted in a confusion matrix with absolute numbers of correct classifications. True labels are depicted on the *y*-axis and predicted labels are depicted on the *x*-axis. Classification performances were high in every condition.

Table 2

Similarity Measures between Input and Reconstructions

Measurement	Size	Location	Size & Location	Rotation
Correlations ↑	.90	.80	.83	.95
SSIM ↑	.06	.01	.01	.04
RMSE ↓	15.87	20.45	21.24	19.42
Euclidean Distance ↓	2986.96	4581.63	4758.23	4350.99
PSNR ↑	35.70	31.76	31.40	32.16

Note. First-level reconstruction quality measures indicating similarity between input and reconstruction images averaged for all letters. Results are represented for each training process for different forms of generalisation individually (SSIM = Structural Similarity Index, RMSE = Root Mean Square Error, PSNR = Peak Signal-to-Noise Ratio, ↑ indicates the more the better, ↓ indicates the less the better).

Table 3

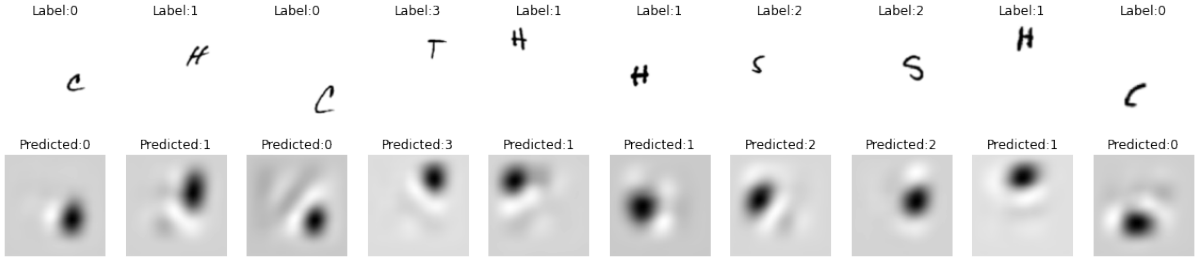
Similarity Measures for Location Generalisation Task

Measurement	C	H	S	T
Correlations ↑	.78	.83	.79	.77
SSIM ↑	.02	.01	.01	.01
RMSE ↓	21.57	22.10	21.99	19.78
Euclidean Distance ↓	4831.83	4949.60	4926.03	4431.07
PSNR ↑	31.31	31.11	31.15	32.08

Note. First-order measures between reconstructed letters and physical stimulus per letter for the network trained on location generalisation (3 different sizes in multiple locations in the picture, trained and tested locations were entirely independent). (SSIM = Structural Similarity Index, RMSE = Root Mean Square Error, PSNR = Peak Signal-to-Noise Ratio, ↑ indicates the higher the more similar, ↓ indicates the lower the more similar)

Figure 4

Reconstructions for Location Generalisation



Note. Input and reconstruction of CapsNet trained on different positions of letter stimuli. Testing and training positions were entirely independent. Reconstruction letters are not distinguishable. The labels indicate which category the stimulus belonged to as well as the predicted label from the network classification (0 = C, 1 = H, 2 = S, 3 = T).

Size Generalisation

The results testing whether the network can classify images in unseen locations can be found in Table 4. For reconstruction results see Figure 5. The classification performance for stimuli with unseen sizes is depicted in Figure 3b. Second-order correlation for 100 stimuli averaged up to 0.76. When the stimuli from the fMRI study were used as input, the classification worked perfectly, reconstruction results can be seen in Figure 6.

Table 4

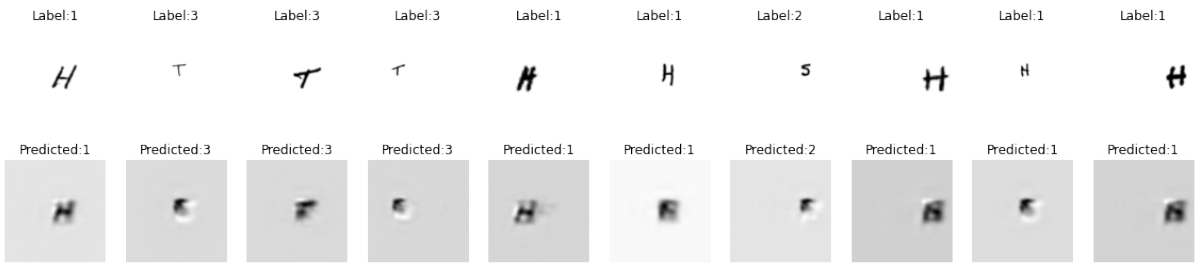
Similarity Measures for Size Generalisation Task

Measurement	C	H	S	T
Correlations ↑	.91	.93	.91	.91
SSIM ↑	.08	.09	.07	.09
RMSE ↓	15.66	16.56	15.89	14.19
Euclidean Distance ↓	3508.20	3710.44	3559.46	3178.94
PSNR ↑	34.18	33.66	34.04	35.03

Note. First-order measures between reconstructed letters and physical stimuli per letter for the network trained on size generalisation. Different sizes were presented in three different positions. None of the sizes used in the test data set were part of the training data set. (SSIM = Structural Similarity Index, RMSE = Root Mean Square Error, PSNR = Peak Signal-to-Noise Ratio, ↑ indicates the higher the more similar, ↓ indicates the lower the more similar)

Figure 5

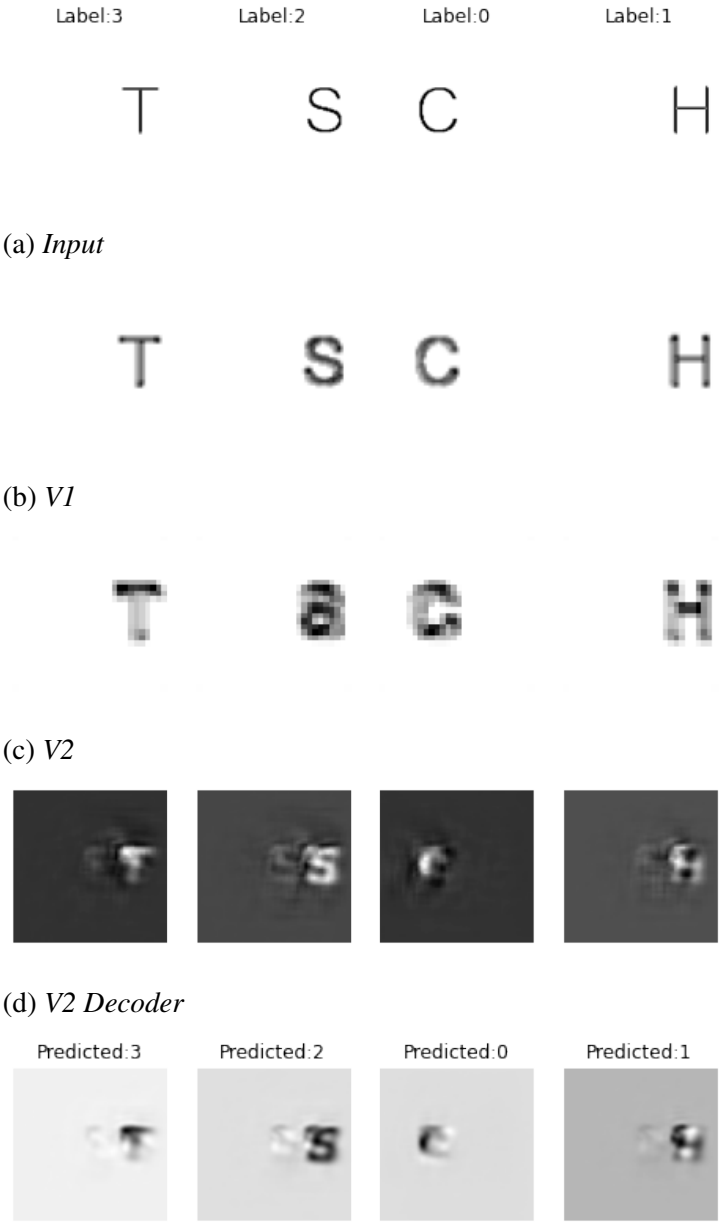
Reconstructions for Size Generalisation



Note. Input and reconstruction of CapsNet trained on different sizes of letter stimuli. The depicted sizes were only used as part of the test data set. The labels indicate which category the stimulus belonged to and the predicted labels indicate which classification the network allocated to the letter (0 = C, 1 = H, 2 = S, 3 = T).

Figure 6

Reconstructions of Artificial Letter Stimuli



(a) *Input*

(b) *V1*

(c) *V2*

(d) *V2 Decoder*

(e) *V1 Decoder / Reconstructions*

Note. Different layer activations of the network when the CapsNet was trained on different sizes with three different positions and tested on the artificial stimuli that were used as part of the fMRI study that these activations were compared to. Depicted are the input to the pre-trained CapsNet, V1, V2, V2 decoder, and the V1 decoder activations. The V1 decoder activation is also seen as the reconstruction activation. The labels indicate which category the stimulus belonged to and the predicted labels indicate which classification the network allocated to the letter. All four letter stimuli were correctly classified (0 = C, 1 = H, 2 = S, 3 = T).

Location and Size Generalisation

The network was tested on a paradigm to see whether size and position generalisation can be trained at the same time. Results can be found in Tables 1, 2, and 5 as well as Figure 3c. Reconstructions are seen in Figure 7. The second-order correlation of one hundred stimuli averaged up to .71.

Table 5

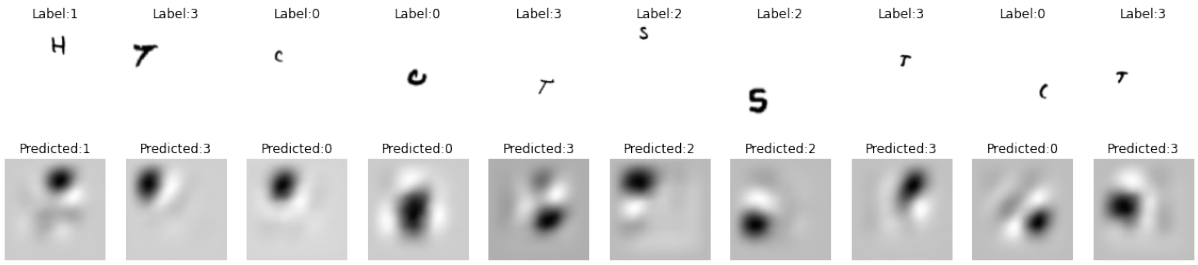
Similarity Measures for Location and Size Generalisation Task

Measurement	C	H	S	T
Correlations ↑	.67	.70	.71	.58
SSIM ↑	.01	.01	.01	.01
RMSE ↓	17.60	17.78	17.60	16.10
Euclidian Distance ↓	3942.86	3983.83	3941.96	3606.54
PSNR ↑	33.33	33.19	33.28	34.14

Note. First-order measures between reconstructed letters and physical stimuli per letter for the network trained on location and size generalisation. Different sizes and locations were used for training, not restricted to an amount of three on any dimension. The depicted results stem from testing positions and sizes that were not included in the training data set. (SSIM: Structural Similarity index measure, RMSE: Root Mean Squared Error, PSNR: Peak Signal-to-Noise-Ratio, ↑ indicates the higher the more similar, ↓ indicates the lower the more similar)

Figure 7

Reconstructions from Size and Location Generalisation



Note. Input and reconstruction of CapsNet trained on different positions and different sizes of letter stimuli. The labels indicate which category the stimulus belonged and the predicted labels indicate which classification the network allocated to the letter (0 = C, 1 = H, 2 = S, 3 = T).

Rotation Generalisation

Classification accuracy and loss for the rotation generalisation task can be found in Table 1. First-order measures of similarity are depicted in Table 2. The second-order correlation amounted to .75 over a sample of 100 stimuli. The reconstructions are shown in Figure 8.

Table 6

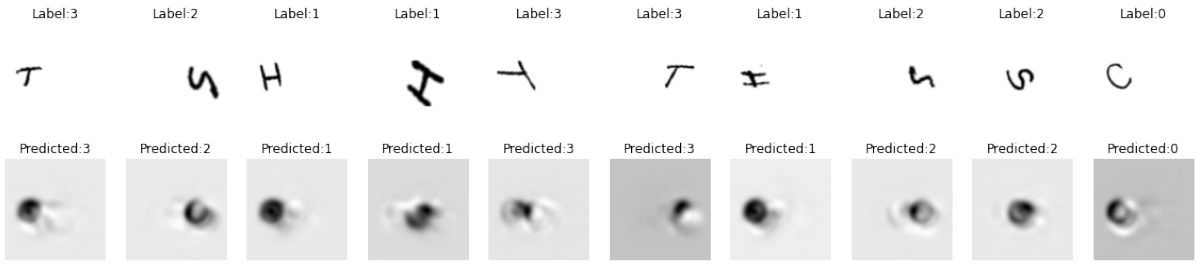
Similarity Measures for Rotation Generalisation Task

	C	H	S	T
Correlations ↑	.93	.92	.92	.90
SSIM ↑	.04	.04	.04	.03
RMSE ↓	24.09	27.41	25.13	24.07
Euclidian Distance ↓	5395.86	6140.17	5630.0	5392.04
PSNR ↑	30.38	29.22	29.98	30.36

Note. First-order measures between reconstructed letters and physical stimuli per letter for the network trained on rotation generalisation with three different positions and sizes. The rotations varied from 90° clockwise to counterclockwise. The testing rotations were not part of the training rotations data set. (SSIM: Structural Similarity index measure, RMSE: Root Mean Squared Error, PSNR: Peak Signal-to-Noise-Ratio, ↑ indicates the higher the more similar, ↓ indicates the lower the more similar)

Figure 8

Reconstruction from Rotation Generalisation



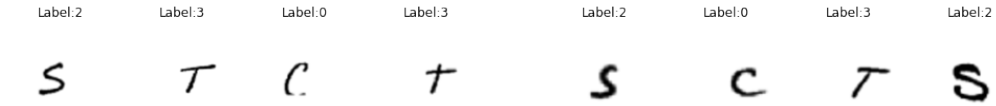
Note. Input to CapsNet trained on different rotations with three different sizes and positions, depicting input, and reconstruction. The rotations ranged from -90° to 90°. The labels indicate which category the stimulus belonged to and the predicted labels indicate which classification the network allocated to the letter (0 = C, 1 = H, 2 = S, 3 = T).

Using 3 Locations with Same Size

For the following test, three different positions and the same letter sizes were used at all times. Classification performance after eight epochs of training amounted to 100.00% with a loss of 0.20. Reconstruction results of this network can be found in Figure 9 on the left side. Additionally, the same network was tested with three positions that were not part of the possible training positions but deviated from them by ten pixels. Reconstruction results, as well as activation patterns in different layers, can be seen in Figure 9 on the right side. Since Sabour et al. (2017) showed that the dimensions for their network were linearly decodable, the proposed network was also tested for this claim. The output of the capsules was tweaked in nine steps for each dimension for four different letter examples and fed into the decoder. Since the reconstructions have rather abstract than linear relations, no specific interpretation of each dimension can be given. Example Reconstructions for three dimensions with tweaked inputs can be found in Figure 10.

Figure 9

Reconstructions Trained and Untrained Positions

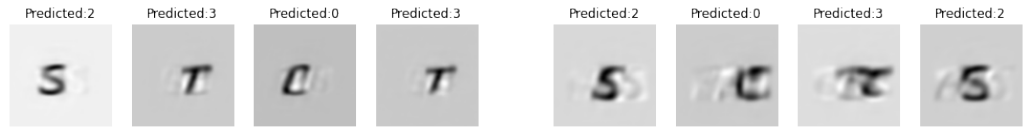


(a) *Input*

(b) *V1*

(c) *V2*

(d) *V2 Decoder*

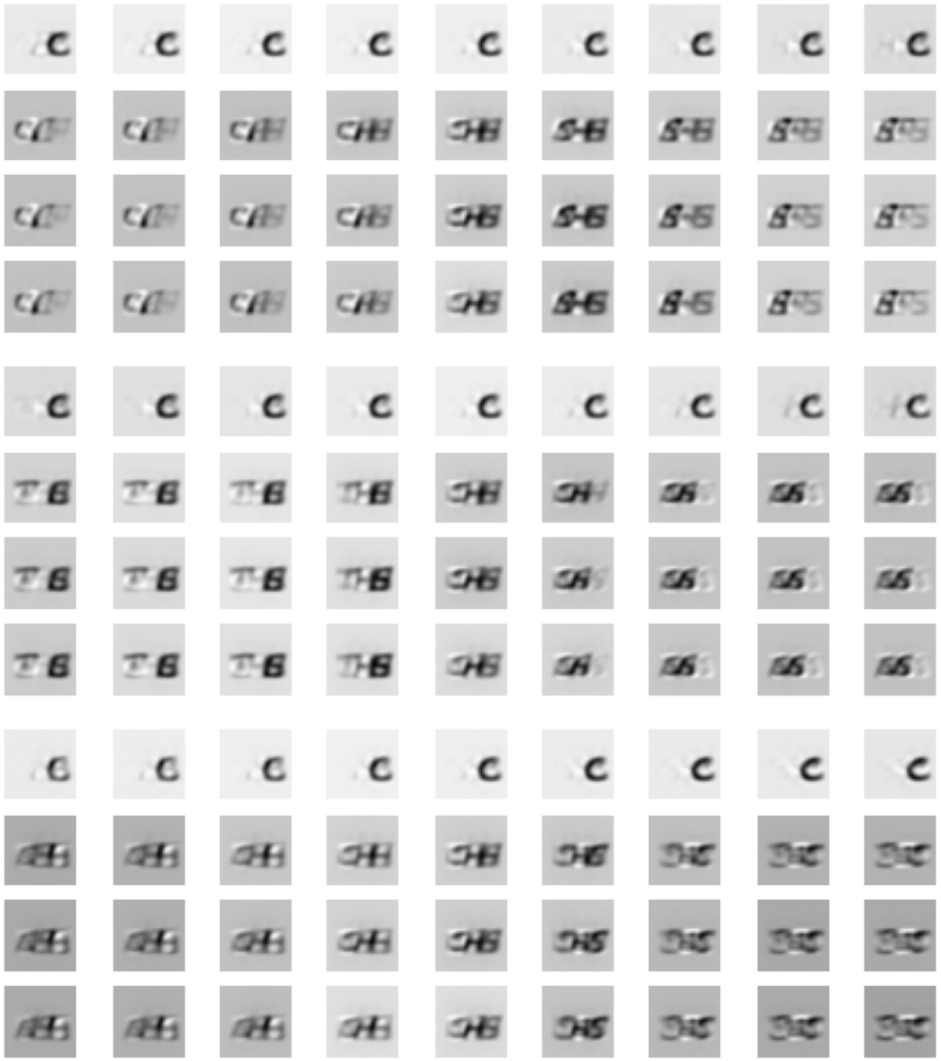


(e) *V1 Decoder / Reconstructions*

Note. Input to CapsNet trained on three positions. Activation patterns of layers are expected to show activation patterns comparable to V1, V2, V2 decoder, and V1 decoder. The V1 decoding activation is also used as the reconstruction activation. On the left side activation patterns and the reconstruction of 3 learned positions are shown. On the right side activation patterns of 3 unknown positions, of which each was ten pixels shifted on the *x*-axis from the formerly learned positions, are shown.

Figure 10

Reconstructions with Tweaked Dimensions

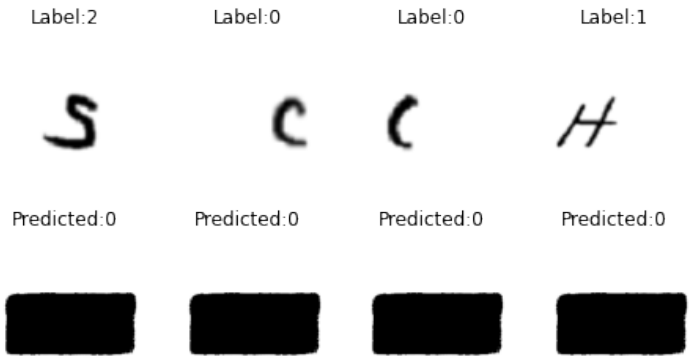


Note. Output of the decoder when dimensions one to three were separately tweaked in nine steps. Each single dimension is comparatively difficult to interpret. Therefore, it is assumed that a combination of dimensions is relevant for the reconstruction of the decoder. For this analysis, the network that was pretrained on three positions with one single size was used.

Original Decoder. The same setup was used to train the network with the original decoder suggested by Sabour et al. (2017). This proposed decoder consisted of two hidden layers, one with 512, the other with 1024 hidden units. Since the images were larger than the ones used in this former study, the number of neurons was increased by a factor of eight. The reconstruction results can be seen in Figure 11. During testing, the classification accuracy amounted to 24.58% with a loss of 1.86.

Figure 11

Original Decoder Reconstructions



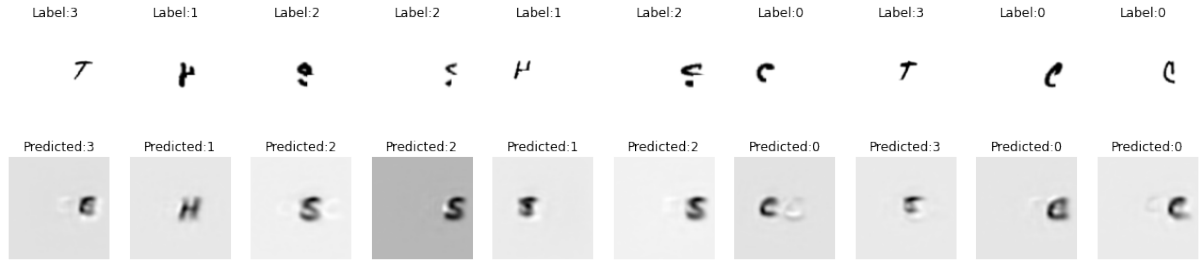
Note. Input to CapsNet trained on three positions with one size. For this test, the decoder structure from Sabour et al. (2017) which consisted was used. Input and reconstruction as well as the stimulus labels and the predicted labels are shown (0 = C, 1 = H, 2 = S, 3 = T). This network does not learn the task nor reconstructs the letter stimulus.

Testing with Occlusion Paradigm

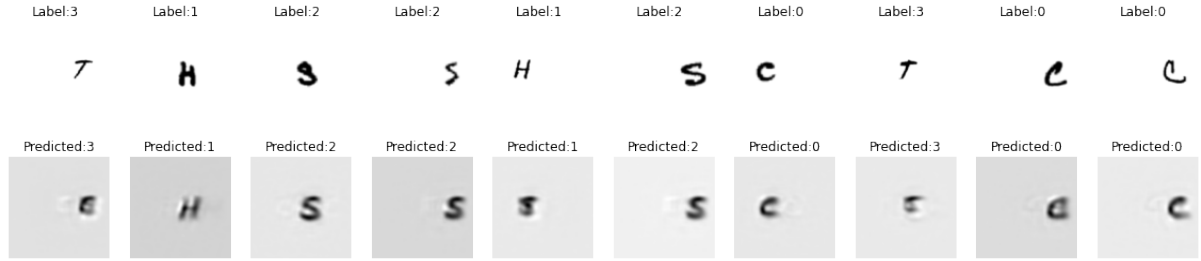
Using a model that was trained on three different locations with multiple sizes, it was tested whether similarly to fMRI activations in Smith and Muckli (2010), non-stimulated regions of the network elicited activation patterns that made it possible that an SVM classifier evaluated the presented category above chance level. Some examples of reconstructed stimuli from complete and occluded stimuli can be seen in Figure 12. Stimuli which had an occluded corner were fed into the network and the reconstructions were extracted. When this reconstruction was then again fed into the network for classification purposes, the accuracy for

Figure 12

Reconstructions from Occlusion Paradigm



(a) *Reconstruction with Occluded Corner*



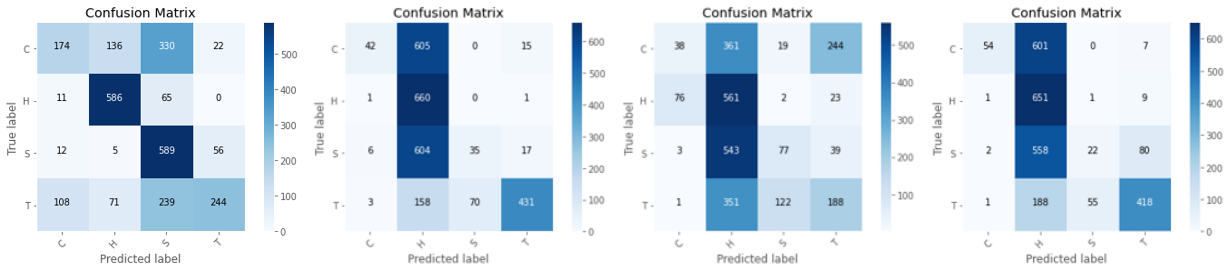
(b) *Reconstruction without Occlusion*

Note. Pictures that occluded the lower right corner were inserted in the network which was formerly trained on three different positions with different sizes. To make the analysis more concise, only 70×70 pixel letter stimuli were used here. Reconstructions from the occluded pictures can be compared to the reconstructions from non-occluded pictures. Labels indicate the category to which the stimulus belongs whereas the predicted label indicates how the network classifies the stimulus (0 = C, 1 = H, 2 = S, 3 = T).

the classification was still high above chance. The reconstructions from samples that were occluded in the lower left corner were correctly classified in 60.16% of cases, whereas when the lower right corner was occluded, only 43.24% were correctly classified any longer indicating that the lower right corner contained more relevant information than the upper left one. The upper left occlusion led to an accuracy of 32.68%, and the upper right corner occlusion of 42.03%. The confusion matrices can be found in Figure 13. The SVM classifier that was trained on the reconstructed corner area from whole stimuli, was then used to predict the stimulus category from the reconstruction of occluded pictures from the other half of the testing data set. Results for correct classification by the SVM classifier can be seen in

Figure 13

Classification Performance for Reconstructions from Occluded Stimuli



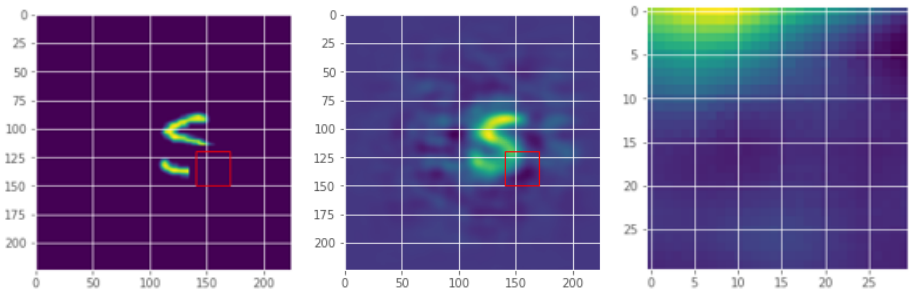
(a) *Upper Left Corner* (b) *Upper Right Corner* (c) *Lower Left Corner* (d) *Lower Right Corner*

Note. The four different corners of the letter images were occluded. The reconstructed stimuli were fed through the same network again and classifications are depicted in a confusion matrix. Strong bias towards the 'H' stimulus could be seen. Classification for each of the occluded corners is above chance level. On the *x*-axis, the predicted label is shown whereas the *y*-axis shows the true label.

Figure 15. Even though the corner did not contain any information in the input stimulus for occluded stimuli, the reconstruction showed a classifiable pattern of activation. 96.79% were correctly classified. Training on perceptual stimuli and testing on reconstructed stimuli from occluded letters did not lead to meaningful results.

Figure 14

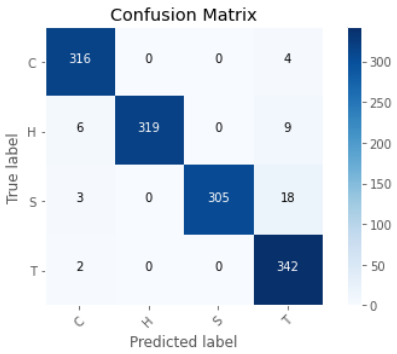
Occlusions Paradigm Example Stimulus



Note. Example stimulus that was fed through the pre-trained network, its respective reconstruction and the extraction activation of the reconstructed corner. The edge of this corner was cut and then used for testing of the SVM classifier. This SVM classifier was formerly trained on the reconstructed edges of non-occluded stimuli. The red box indicates which area was used for training and testing.

Figure 15

Classification for Corners Reconstructed from Occluded Stimuli



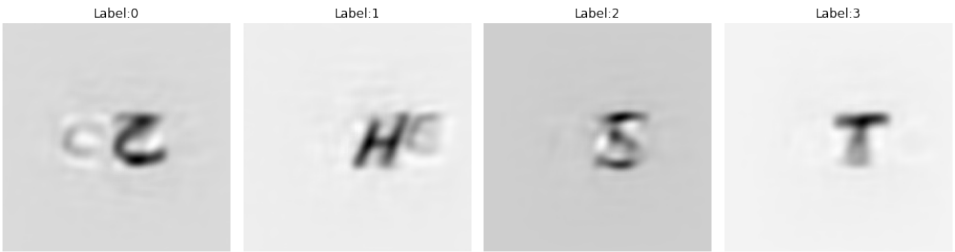
Note. This picture shows the SVM classification of the reconstruction of occluded stimuli. The stimuli were occluded in the lower right corner. This classifier was formerly trained on the reconstructed corner of the complete stimuli. Half of the testing stimuli were used for training and the other half for testing. This high classification performance indicated that during reconstruction even in areas which relate to input areas where no information was shown, relevant activation could be seen.

Relating the CapsNet with the fMRI Data

Reconstructing Higher-Level Imagery Activations. When the fMRI data was mapped with the help of a neural network onto the capsule space, a validation loss of 0.03 was reached. The validation trial reconstructions are depicted in Figure 16. It was seen that

Figure 16

Perception Trial Reconstructions

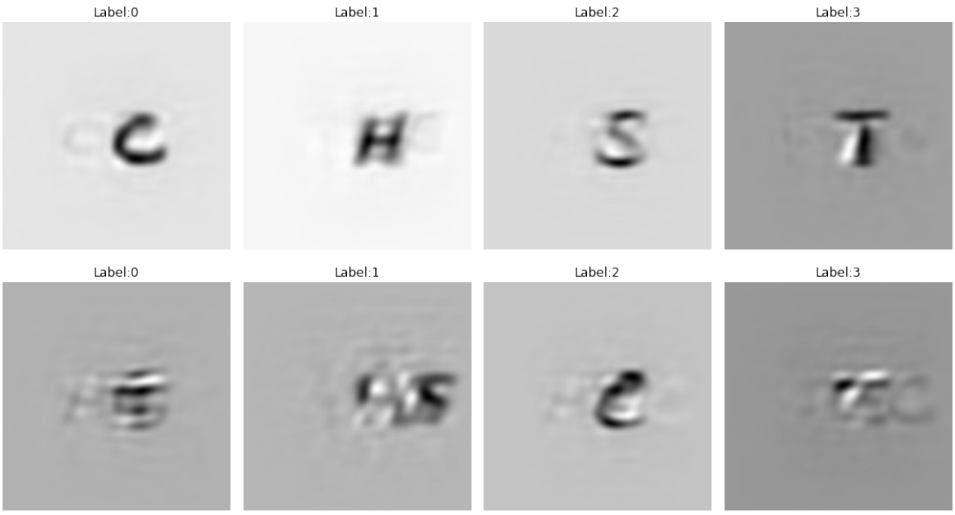


Note. Reconstruction results when perception voxel activations of the validation trial were mapped onto capsule activation patterns with the help of a three-layered deep neural network and then inserted into the CapsNet. This pre-trained three-layered network was then used to map the imagery higher-level activation patterns onto the capsule space.

correlations and SSIM of the entire pictures did not yield any relation between the reconstructed stimuli and the presented stimuli. RMSE and euclidean distance showed the lowest and therefore best-matching scores for three out of four stimuli. Only the letter S was not closest related to the actual S stimulus for each of these indices. When taking the first dimension of each category of each remapped stimulus to see whether the network learned the different classes correctly, it became clear that the highest value for the presence of a specific category stimulus was not of importance for the network. This value only reached chance level. On the other hand, when correlating the capsule activations with the expected capsule activations for the imagery stimuli, 38 out of 128 stimuli were strongest correlated with their expected capsule activations compared to the other category capsule activations (binomial test, $p = .09$). Additionally, the predicted capsule activations for each trial were fed through the pre-trained CapsNet. Some examples of reconstruction results can be seen in Figure 17. The reconstructions from single trials were significantly highest correlated with the respected

Figure 17

Exemplar Reconstructions from Imagery Trials

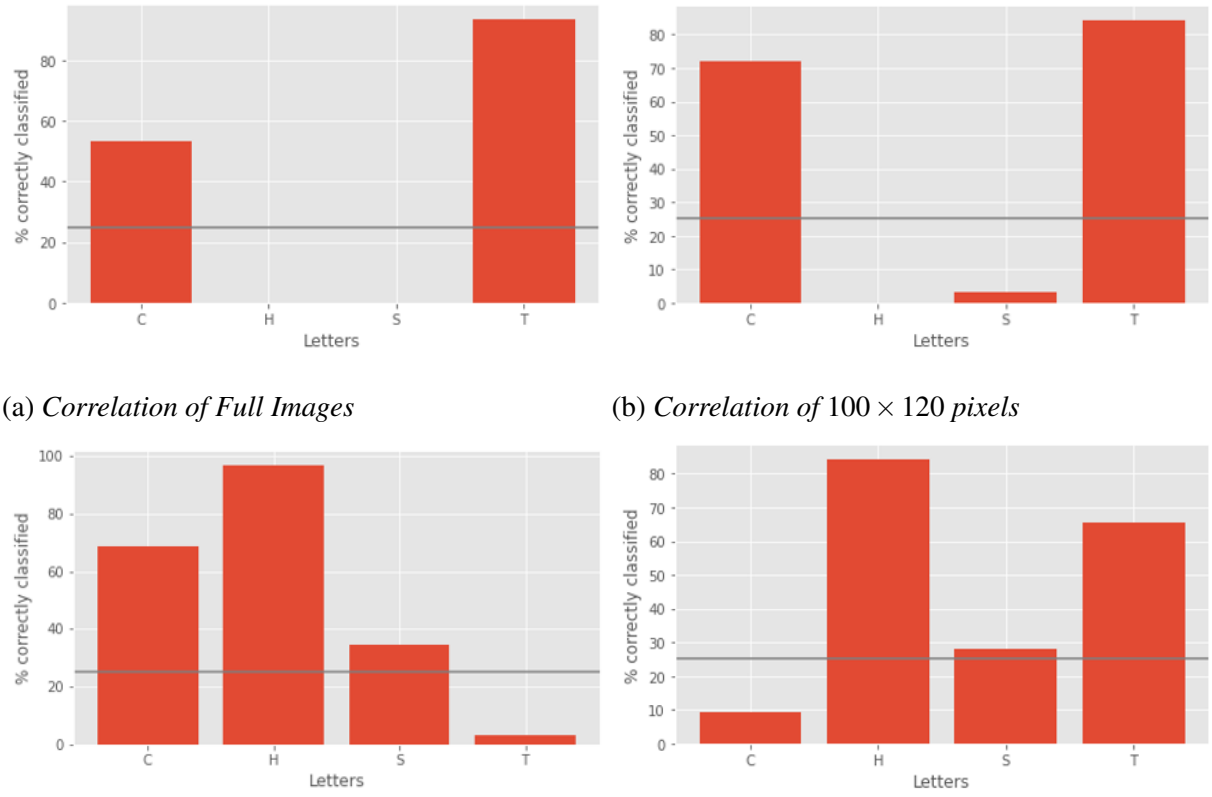


Note. Exemplary reconstruction results when imagery voxel activations were mapped onto capsule activation patterns with the help of a three-layered deep neural network and then inserted into the CapsNet. The first row depicts subjectively identifiable letters whereas the second row depicts examples of reconstructions with restricted identifiability. The labels indicate the respective stimulus group (0 = C, 1 = H, 2 = S, 3 = T).

stimulus (binomial test, 128 trials, 47 reconstructions with highest correlation with respected letter stimulus, 36.72%, $p < .001$). When using just a cutout part (100×120 pixels) of the letter stimuli, 51 (39.84%) of the stimuli were higher correlated with their respective letter stimulus compared to their correlation with the other physical stimuli. An overview of these strongest correlations for the complete stimuli and the partially-occluded stimuli can be found in Figure 18. When using euclidean distance as a measurement, 67 reconstructed stimuli (50.78%) showed the smallest distance to the respective physical letter stimulus for the entire images and 62 (46.88%) did so for the smaller cutout area. Results for SSIM were highly similar to the correlational results. The results for RMSE as well as PSNR were comparable to

Figure 18

Imagery Reconstruction Evaluation



(a) *Correlation of Full Images* (b) *Correlation of* 100×120 *pixels*

(c) *Euclidean Distance of Full Images* (d) *Euclidean Distance of* 100×120 *pixels*

Note. Evaluation of reconstructions from higher-level imagery activations. The amount of strongest correlations, and lowest euclidean distances of reconstructed letters with respective letter stimulus separated by letter. Depictions for full images and image cutouts of 100×120 pixels. The measures yielded different results. The grey line represents a chance level of classification at 25%.

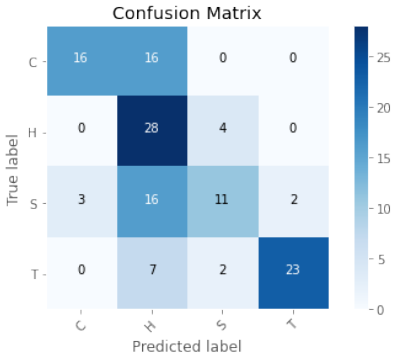
the euclidean distance plots. Relevantly, these results were highly susceptible to the respective pixel area that was chosen, thus these results should be interpreted with caution.

As a final evaluation, the reconstructed pictures were fed through the network to see whether the network itself would classify the pictures according to their respective expected categories. The network classified 60.94% (70 correctly classified stimuli) of the stimuli according to their imagined stimulus. The confusion matrix for the correct and incorrect classifications can be found in Figure 19.

Qiao et al. (2018) used linear regression to reduce the amount of features fed into the 3-layered neural network in their fMRI data set to 100. This is a relevant step before training a deep learning model to avoid overfitting which in turn would lead to poor generalisation results (Mwangi, Tian, & Soares, 2014). To account for this issue, backwards elimination was applied to the fMRI data. 100 voxels were selected, and the same analysis steps were conducted. No relevant differences to the former results were revealed. Therefore, the entire data set with 743 voxels was used for the presented analysis.

Figure 19

Classification Performance for Reconstructed Imagery Stimuli



Note. Classification accuracy for reconstructed imagery stimuli. The CapsNet which was trained on three positions and multiple different sizes was used for this analysis. The *x*-axis depicts the predicted labels and the *y*-axis the true labels. The correct classification reached above chance level which indicates that it is possible to reconstruct images from higher-level visual area activations. 60.94% of reconstructions were correctly classified.

Comparing Lower-Level Activations of the CapsNet to fMRI Activations.

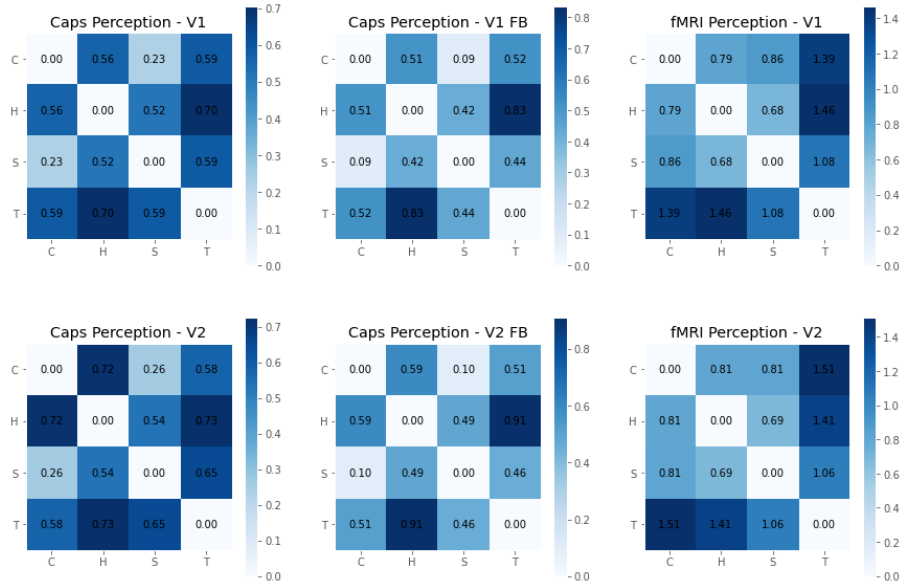
RSA Analysis. To calculate RSA similarity measures, RDMs for V1 and V2 from the CapsNet encoder and decoder, as well as V1 and V2 raw voxel activations of the fMRI activations during perception were calculated (see Figure 20). The same calculations were done for the fMRI imagery activation patterns for V1 and V2 as well as the reconstructed activation patterns from remapped voxel-to-capsule activations (see Figure 21).

To compare these lower-level activation RDMs, Kriegeskorte, Mur, and Bandettini (2008) suggested the RSA analysis. This was applied to compare the activation patterns between the modalities.

The correlations for all RDMs were high, not differentiating between representations for V1 and V2. It would have been expected that the fMRI V1 activations related higher to CapsNet V1 representations than to CapsNet V2 representations. The encoder was expected to be more strongly related to the perception fMRI activations and the decoder to the imagery activations. More stimuli categories would be needed to draw conclusions about the representation patterns of the CapsNet and fMRI activation patterns.

Figure 20

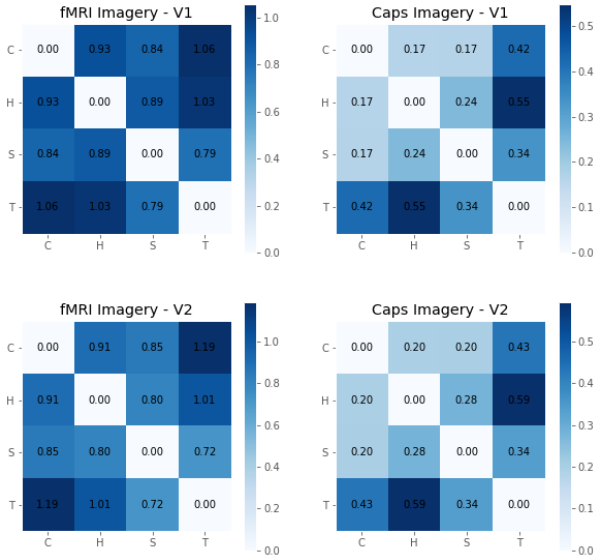
Representational Dissimilarity Matrices from Perception Activation Patterns



Note. These Representational Dissimilarity Matrices were calculated for CapsNet activation patterns relating to V1 and V2, as well as the decoding layers V1 and V2. Additionally, the same measure was calculated for the fMRI activations. The dissimilarities are calculated by 1-correlation between two conditions. Higher values indicate stronger dissimilarities between conditions.

Figure 21

Representational Dissimilarity Matrices from Imagery Activation Patterns



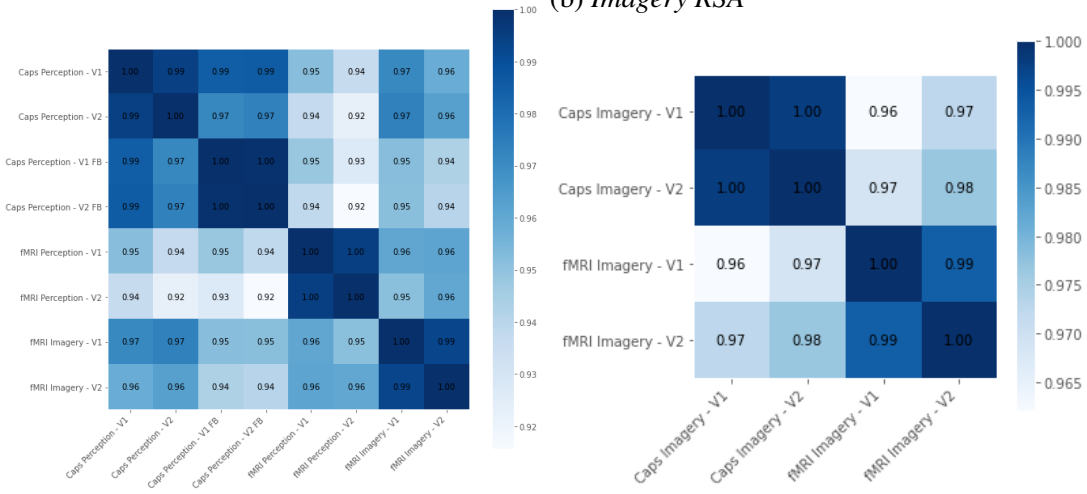
Note. The Representational Dissimilarity Matrices were calculated for CapsNet decoder layers V1 and V2 for reconstructions from imagery activations and fMRI raw voxel activations of layers V1 and V2 during imagery. The dissimilarities are calculated by 1-correlation between two conditions. Higher values indicate stronger dissimilarities between conditions.

Figure 22

Representational Similarity Analysis

(a) *Perception RSA*

(b) *Imagery RSA*



Note. The Representational Similarity Analysis is based on correlations between Representational Dissimilarity Matrices (RDM). This analysis compares RDMs calculated for fMRI raw voxel activations of layers V1, V2, and RDMs based on CapsNet layer V1 and V2 activation patterns, for perception and imagery respectively.

Correlation Analysis. Correlations were calculated for reconstructions which were pRF mapped from the raw activation patterns and fed through the autoencoder and the activation patterns of the respective layer in the CapsNet. Results from the calculations from CapsNet encoder and decoder layers V1 and V2 with imagery activation patterns are seen in Table 7. FMRI imagery activation patterns were also related to the activation patterns when the fMRI activation patterns were mapped onto capsule space and then reconstructed. Results can be found in Table 8. Furthermore, the perception reconstructions were correlated with the perception activation patterns in the CapsNet for the encoder and the decoder (see Table 9).

Table 7

Correlations of Activation Patterns of the CapsNet during Perception with fMRI Activation Patterns During Imagery

(a) *V1*

		fMRI			
		C	H	S	T
Encoder	C	.83	.70	.74	.54
	H	.70	.75	.71	.50
	S	.79	.58	.84	.54
	T	.67	.63	.62	.78
Decoder	C	.90	.68	.86	.62
	H	.83	.84	.78	.64
	S	.84	.66	.82	.62
	T	.86	.77	.79	.89

(b) *V2*

		fMRI			
		C	H	S	T
Encoder	C	.64	.72	.59	.55
	H	.64	.78	.61	.57
	S	.71	.54	.71	.57
	T	.63	.73	.55	.70
Decoder	C	.59	.52	.58	.50
	H	.87	.78	.77	.74
	S	.56	.51	.61	.49
	T	.78	.72	.70	.74

Note. Correlations of the CapsNet Activation patterns from encoder and decoder V1 and V2 during perception and fMRI imagery reconstructions for V1 and V2. The correlations on the diagonal should be the highest in their row and column for the network reflecting similar properties as the fMRI activations. Additionally, the correlations of the decoder with the imagery activations were expected to be higher than those of the encoder.

It was expected that the fMRI activation patterns during perception related more closely to the CapsNet encoder and fMRI activation patterns during imagery related more closely to the CapsNet decoder when stimuli were presented to the CapsNet. This hypothesis was confirmed for V1 activations but V2 aligned with this hypothesis only for the encoder, and not for the decoder. Additionally, the activation patterns from reconstructed imagery still showed the highest correlations for the respective stimulus with the presented stimulus for V1 but not for V2.

Table 8

Correlations of Activation Patterns of the CapsNet from Imagery Reconstructions with fMRI Activation Patterns During Imagery

(a) *V1*

		fMRI			
		C	H	S	T
Decoder	C	.89	.83	.86	.73
	H	.90	.88	.83	.70
	S	.89	.69	.87	.65
	T	.80	.69	.79	.86

(b) *V2*

		fMRI			
		C	H	S	T
Decoder	C	.83	.68	.78	.69
	H	.91	.81	.84	.81
	S	.71	.60	.68	.60
	T	.57	.54	.61	.57

Note. CapsNet activation patterns from decoder V1 and V2 were correlated with reconstructed imagery activations for areas V1 and V2. The reconstructions from the CapsNet were solely based on the mapping from higher visual area activations onto the capsule activations and then reconstructed. Strongest correlations are expected on the diagonal. High correlations with the correct stimuli indicated accurate mapping of higher-level activations onto capsule space.

Table 9

Correlations of Activation Patterns of the CapsNet during Perception with fMRI Activation Patterns During Perception

(a) *V1*

		fMRI			
		C	H	S	T
Encoder	C	.88	.72	.79	.57
	H	.74	.78	.65	.54
	S	.71	.64	.83	.54
	T	.60	.66	.60	.77
Decoder	C	.81	.76	.86	.66
	H	.88	.83	.77	.66
	S	.80	.72	.81	.62
	T	.78	.80	.82	.91

(b) *V2*

		fMRI			
		C	H	S	T
Encoder	C	.81	.71	.67	.52
	H	.76	.77	.61	.53
	S	.70	.57	.74	.53
	T	.60	.70	.56	.81
Decoder	C	.62	.53	.64	.52
	H	.87	.81	.87	.71
	S	.58	.52	.62	.49
	T	.72	.74	.74	.79

Note. Correlations of the CapsNet activation patterns from encoder and decoder V1 and V2 during perception with fMRI Perception Reconstructions for V1 and V2. The correlations on the diagonal should be the highest in their row and column for the network reflecting similar properties as the fMRI activations. Additionally, the correlations of the encoder with the perception activations were expected to be higher than those of the decoder.

Discussion

This study demonstrated that CapsNets are a suitable tool to reach high generalisation performance while maintaining biological plausibility. Formerly, ANNs lacked invariance towards unseen modifications but the results of this research demonstrated that the proposed network is able to generalise towards unseen locations, sizes, and rotations. The fact that identifiable reconstructions were elicited when occluded stimuli were used as samples indicates that the network also uses surrounding information in its feedback tracks. The implemented decoder yielded good reconstructions while aligning with fMRI activations. The

results of the presented study support the hypothesis that the CapsNet activation patterns show a relevant relation to the brain activation patterns during perception and imagery. Correctly classifiable reconstructions from higher-level visual area activations during imagery indicated that equivariant representations of stimuli are elicited without the necessity of physical stimulation. Since a network was trained to map the fMRI activations onto capsule space with the perception data and then tested on only imagery data, an overlapping representation during perception and imagery trials in higher-level visual areas could be seen. Lower-level area activations, especially V1, showed strong similarities to the CapsNet activations. The encoder related more closely to the perceptual, whereas the decoder related more closely to the imagery activation patterns. These overlapping representations with imaging activation patterns imply that implementing a mechanism which reflects the high generalisation ability of humans into neural networks is a necessity to improve modelling biological vision.

Early Visual Area Activations in the CapsNet Overlap with Imaging Activations

On the one hand, perception and imagery could be differentiated in early visual areas: the correlational results revealed that the encoder structure of the CapsNet was more closely related to the perception fMRI activations whereas the decoder structure was more closely related to the imagery fMRI activations. This confirms the idea that perceptual processes rely more strongly on the feedforward pathway whereas imagery processes more strongly involve feedback connections (Koenig-Robert & Pearson, 2021).

Building upon these results, it would be an interesting approach to differentiate the cortical layers of the early visual area activations to see whether feedforward and feedback activations can be distinguished. Feedforward activations are assumed to mainly engage middle layers whereas feedback activations are expected to be processed in deep and superficial layers (Koenig-Robert & Pearson, 2021). Therefore, it might be expected that the fMRI perception activations mainly rely upon the middle layers whereas the imagery activations are assumed to rely upon the upper and lower layers.

On the other hand, even though there were differences found between the imaging activations alignment with the encoder and the decoder, a strong overlap between perception and imagery

depictions were seen in early visual areas, especially in V2. This aligns with the finding that imagery engages early visual areas in a depictive manner, as well as that low-level features, are encoded similarly during imagery as they are during perception (Koenig-Robert & Pearson, 2021).

To be able to highlight the differences in reconstructions between imagery and perception in more detail, it might be necessary to introduce an adapted pRF mapping for imagery reconstructions. In the presented study, the imagery pRF mapping was calculated based on the perceptual activations elicited while perceiving a checkerboard. It might be that the patches of voxels activated during imagery are more blurry and therefore do not align with the pRF mapping during perception.

Overlapping Representations in Higher-Level Areas during Perception and Imagery

The overlap in representations at higher-level areas during perception and imagery is consistent with previous findings by Dijkstra et al. (2019). This implies that the feedforward stream activates higher-level neurons during perception that can then be reactivated to stimulate the feedback stream during imagery.

It has been shown that monkey IT includes neurons that fire due to view-specific properties but also a subset of neurons that fire only depending on the specific object presented (Booth & Rolls, 1998; Ito, Tamura, Fujita, & Tanaka, 1995). These findings align with an equivariant representation in higher-level areas in monkey brains. The preserved information in the highest-level area of the CapsNet includes the stimulus category but additionally comprises the relevant information for the reconstruction of the stimulus, differing depending on the specific instantiation of the stimulus.

Not having recorded the full structure of IT is a strong limiting factor in comparing the highest layer of the CapsNet and the high-level visual area fRMI activations. Dijkstra et al. (2019) and Koenig-Robert and Pearson (2021) showed that the higher an area is located in the visual ventral stream, the more overlap can be found between perception and imagery activations. It was shown that reconstructions are still possible which implies that also earlier visual ventral stream areas show an overlapping representation during perception and imagery.

Possibly, reconstructions would have been more accurate if IT data was available. The utilised fMRI sequence was optimised to capture lower-level visual areas and therefore lacked resolution in higher-level areas. Mapping the whole visual stream while maintaining high spatial and temporal resolution would be favourable in upcoming experiments. Furthermore, visual inspection of the reconstructed stimuli from imagery trials did not show the high resolution of Qiao et al. (2018). This might be due to multiple reasons. Firstly, they used activation patterns of voxels throughout the whole brain. The present analysis was restricted to hypothesis-driven higher-level visual brain areas. Moreover, they only reconstructed letters from perception trials. Here, reconstructions from imagery activations were attempted. Since imagery is assumed to be a weak form of perception, more noise might be introduced into reconstructions (Pearson, 2019). Thirdly, more trials per category were collected in their study. Therefore, they had more training data, also including more variability between presented stimuli which facilitates the training of the deep neural network that maps brain activation onto capsule activation. Therefore, when testing the proposed network, it should be considered to use a data set with more variability in their stimuli as well as an increased amount of trials. The present study serves as proof-of-concept that imagery trials can be reconstructed from mapping of higher-level area activation patterns onto capsule space.

High Generalisation Performance

The presented study confirmed that CapsNets show high generalisation performance (Sabour et al., 2017). The results strongly imply that the CapsNet is able to classify stimuli independent of their specific qualities, such as shown with size, position and rotation. On the other hand, that property might differ for humans. It was shown before that the more rotated words are the more time is needed by humans to indicate whether the stimulus is a word or not. Errors in that judgement also increased depending on the angle of rotation (Koriat & Norman, 1985). That is why the high generalisation performance might not be entirely biologically valid. It would be relevant to test how the network behaves in situations of ambiguous classifications due to rotations, for instance when “p” and “b” stimuli are supposed to be classified. If similar error patterns as for humans would occur, that would again support

the point of biological validity of the network.

A major shortcoming of the present study is the lack of comparison to established networks, such as CorNetZ (Kubilius et al., 2018), VGG-16 (Simonyan & Zisserman, 2014), or ResNet-50 (He, Zhang, Ren, & Sun, 2016). These networks have proven on certain standards, such as error levels, to reach human classification performance (Kriegeskorte, 2015) while maintaining a hierarchical structure that has been linked to the human ventral object pathway (Leek, Leonardis, & Heinke, 2022). It remains unclear whether these networks can serve as theoretical frameworks for understanding image classification in biological vision (Leek et al., 2022). Adding to the body of knowledge, these networks could be evaluated on the same generalisation performance measures as well as compared to the same data set. The lack of an equivalent to feedback connections in these networks limits the analysis. These comparisons might still lead to a better understanding of the question whether higher generalisation performance is desirable to model the visual ventral stream during perception.

Similarities between the fMRI data set and the capsule layer activations could be shown but the full extent of the network's capabilities could not be reflected with this data set. The network is specifically trained to generalise to different sizes, locations, and rotations but the data set only included letters presented in the same location, size, and at the same rotation angle. Further research is needed to test whether the differing activation patterns of the network also reflect the different activation patterns for other locations or sizes when human participants perceive stimuli. For instance, it would be interesting to investigate whether the reconstructions of stimuli imagined in different locations are also projected onto different locations.

Image Comparison Measurements

The generalisation performance was not only measured with classification accuracy and loss calculations but also with the help of correlations, RMSE, euclidean distance, PSNR, and SSIM. These measures are similarity measures to compare two pictures with each other and they all rely on low-level feature similarity (Rakhimberdina, Jodelet, Liu, & Murata, 2021). All of these measures were introduced to get a broader overview than was given in Senden et

al. (2019). They restricted their analysis to the use of spatial correlations.

Interestingly, when using correlations as a measurement high similarities were found but the SSIM, as well as the PSNR, indicated low similarities. This divergence between the results might be explained by the different shortcomings of each of them. Correlations are an often used measurement but they lack sensitivity to changes in the edge intensity and edge alignment (Beliy et al., 2019). MSE is also a regularly applied measure for comparisons but it shows poor correspondence to human visual perception. MSE calculations are independent of the spatial relationships between image pixels and do not give differing weights to each of the pixel (Z. Wang, Bovik, Sheikh, & Simoncelli, 2004). The same applies to RMSE which was used as standardisation of MSE in this study. The same problem also arises with euclidean distance which is defined as a summation of the pixel-wise intensity differences. Therefore, even small deformations can result in large values (L. Wang, Zhang, & Feng, 2005). PSNR and MSE show low discrimination power for structural content as demonstrated by multiple degradations applied to the same image yielding the same value of the measurements (Hore & Ziou, 2010). This problem arises since MSE and PSNR are giving absolute errors.

On the other hand, SSIM was proposed as being a similarity measure that captures human visual perception. It calculates a weighted combination of three comparative measures which are luminance, contrast and structure (Rakhimberdina et al., 2021). Even though this measure is correlated with human quality perception, it still compares poorly to some human perceptual characteristics (Zhang, Isola, Efros, Shechtman, & Wang, 2018).

As a potential improvement to the present study, higher-level perceptual similarities should be taken into account when comparing samples with reconstructions. Therefore, the PSM (Perceptual Similarity Metric) is proposed and should be evaluated in further research on CapsNets. This metric uses a CNN to extract hierarchical multilayer features of input images. These are then further compared across layers using a distance metric. Often, AlexNet is used in that context to extract the hierarchical features (Rakhimberdina et al., 2021).

To summarise, it was explained why each of the measures is reflecting slightly different properties of the picture comparisons. Because of the great difference between correlations and the SSIM which is supposed to be more similar to human similarity perception, the

correlational results should be treated with caution. Further studies comparing reconstructions should take into account more human vision like similarity measures such as the PSM.

Biological and Psychological Plausibility

Similarly to Svanera et al. (2021), the claim holds that a network with an encoder and decoder can use surrounding information to reconstruct occluded stimuli. Differently from their study, this network was not particularly trained for reconstructing occluded images but showed this behaviour following training that optimised classification and reconstruction. Since Smith and Muckli (2010) showed that the human brain imaging activations followed a similar pattern, this characteristic of the network is highly biologically plausible.

When the network's performance was tested with the original decoder from Sabour et al. (2017), the results were subpar and the network classified images randomly. This was to be expected since the stimuli showed increased variability compared to MNIST images. Nonetheless, just increasing the size of the layers or the number of layers did not lead to improved results. Instead, the main architecture from the feedforward network was reversed which also relates to the hypothesised architecture in the brain, using a feedback stream that leads through the same areas as the feedforward stream (Dijkstra et al., 2019).

Even though a decoder has been implemented in the network, one might argue that these so-called feedback connections are not informative of the human brain since they lack influence on layers earlier in the visual stream. For instance, it has been shown that neurons in early visual areas show target-enhancement, particularly observed after longer latencies which implies that not the on-discharge is relevant for the activation (Hon, Thompson, Sigala, & Duncan, 2009). Instead, feedback about target presence was assumed to influence early visual areas. Besides a different implementation of feedback connections, also adding lateral connections might make ANNs more biologically plausible (Kubilius et al., 2018).

A further improvement of this visual ventral stream model might be the introduction of an attention mechanism (Goebel, 1992). Humans are well able to outperform ANNs once the presented classification tasks become more challenging because they are able to direct their attention towards relevant parts of a scene while ignoring distracting information (Peters et al.,

2012). Different attempts to implement attention mechanisms with capsule networks were made and showed promising results in object classification tasks (Huang & Zhou, 2020; Mazzia et al., 2021; Pucci et al., 2021). Implementing an attention mechanism in the proposed network might lead to improved explainability of human activation patterns.

A general problem of ANNs is the usage of the backpropagation algorithm for training (Illing, Gerstner, & Brea, 2019). Learning rules in the brain most likely do not follow backpropagation. Unsupervised methods with local learning rules or algorithms for sparse coding might be more biologically plausible alternatives.

This network was compared to fMRI activation patterns. Imaging activations can be compared to the network with correlational approaches and thereby give a broad idea of similar representational patterns. Deriving architectural constraints for single neurons in ANNs is only possible to a very limited extent since a single voxel in fMRI represents the activity of tens of thousands of neurons (Logothetis, Pauls, Augath, Trinath, & Oeltermann, 2001) while not simply providing the average activity within its boundaries but rather the complex spatiotemporal transformation of the hemodynamic response (Kriegeskorte, Mur, Ruff, et al., 2008). To model a network with biologically plausible architectural constraints, for instance, relevant bottom-up, top-down or lateral connections, more detailed insight than can be given by imaging devices is needed. Animal recordings enable the necessary single-cell recordings. Biological plausibility was tested solely with the occlusion paradigm and with the help of comparisons to lower-level areas. No psychological constraints were investigated. A known issue that has not been tested relates to humans heavily relying on relations between object parts to classify objects whereas ANNs often misclassify objects with altered relational presentation. ANNs also wrongly classified samples which differed in pixel overlap but shared the same categorical relations (Bowers et al., 2022). Humans only mistook the pixel version for the object but did correctly classify the altered relation version. Further research might test this claim to see whether the proposed CapsNet follows the human pattern in misclassifying the pictures with altered pixel relations while maintaining high performance for the stimuli with the altered relations. The CapsNet is hypothesised to adopt this pattern since it focuses on figure-part relationships for classification.

Another example of psychological constraints that might be tested to claim similarities to human performance is related to feedforward deep neural networks classifying stimuli depending on local shapes instead of global shapes (Baker, Lu, Erlikhman, & Kellman, 2020). AlexNet, VGG-19, and ResNet-50 were trained to classify squares and circles. For testing, squares comprised of curved elements and circles comprised of squared shapes were presented. For classification, the networks relied upon local contour features rather than global shapes. They argued that global features would dominate human vision whereas ANNs only represent how local contour fragments spatially relate to each other when forming global shapes. Since the routing-by-agreement algorithm strongly relies on the feature relations of an object, this paradigm might be a promising additional test to demonstrate the generalisation ability of the network. Further examples of psychological constraints that are often not met by ANNs can be found in Bowers et al. (2022).

The current state-of-the-art approach to assess model quality mainly relies on a match in human-ANN classification performance. It is unclear whether this benchmark is relevant for comparisons to human performance since networks scoring high account for almost no findings reported in psychology (Bowers et al., 2022). No variables are manipulated and therefore no hypothesis testing happens. Bowers et al. (2022) mention that predicting observational data does not necessarily entail that two systems rely on similar mechanisms or representations. Deep neural networks tend to heavily rely on single pixels in pictures for classification. These single pixels which were imperceptible to humans when systematically inserted led to classification that only relied on them. Similarly, RSA analyses could be systematically influenced by inserting patches of pixels into pictures, implying that confounds are relevant in this analysis. That is why Bowers et al. (2022) conclude that ANNs rather exploit shortcomings in data sets than actually modelling human vision. Therefore, further testing of the network would be needed to conclude that not these shortcomings led to the promising results this network revealed.

Overall, it becomes clear that a wide variety of additional testing is necessary to claim biological and psychological validity of an ANN. Specific hypotheses should be tested and a test battery for ANNs to investigate relevant psychological and biological parameters of

human vision should be comprised to compare the networks. Even though classification and reconstruction performance as well as RSA analyses show interesting correlational insight, additional tests with relevant psychological hypotheses are needed to claim relationships between biological vision and CapsNet activation patterns.

Interpretability

Current machine learning approaches modelling human brain activations frequently do not provide an understanding of the underlying neural processes or how they might contribute to the outcome of the network (Fellous, Sapiro, Rossi, Mayberg, & Ferrante, 2019). Additionally, high-performing methods tend to be the least explainable thereby providing little additional understanding of mechanisms in the human brain. In recent years the interest in explainable artificial intelligence rose (Angelov, Soares, Jiang, Arnold, & Atkinson, 2021). Since Sabour et al. (2017) claimed a linear relationship between the capsule layer activations and the reconstructions, it was thought that the presented network might offer a higher degree of explainability. When testing this hypothesis it became clear that even though this network contains some form of equivariant information at the capsule layer, this representation seems to contain an abstract combination of features which could not be easily interpreted.

Conclusion

The presented CapsNet showed high generalisation ability as well as alignment with fMRI activation patterns in lower-level as well as higher-level areas in the visual ventral stream. These promising results imply that CapsNets inherently high generalisation ability might be relevant in implementing biological vision. The decoder structure which implements feedback connections is an improvement compared to networks relying only on feedforward classification processes. Additionally, high-level areas indicated coherent activation patterns during perception and imagery. This finding demonstrates an overlapping representation during the two different tasks. Further comparisons to other established networks are needed to distinguish whether the high generalisation performance is a necessary characteristic for improved modelling of the visual ventral stream.

Appendix

All used code and data from the presented study can be found on the author's GitHub at `https://github.com/FKlepel/Thesis_CapsNet_Perception-Imagery`.

References

Amit, Y. (2019). Deep learning with asymmetric connections and hebbian updates. *Frontiers in computational neuroscience*, *13*, 18.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5), e1424.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2020). Local features and global shape information in object classification by deep convolutional neural networks. *Vision research*, *172*, 46–61.

Beliy, R., Gaziv, G., Hoogi, A., Strappini, F., Golan, T., & Irani, M. (2019). From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *Advances in Neural Information Processing Systems*, *32*.

Booth, M., & Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral cortex (New York, NY: 1991)*, *8*(6), 510–523.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., ... others (2022). Deep problems with neural network models of human vision.

Cohen, G., Afshar, S., Tapson, J., & Van Schaik, A. (2017). Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (ijcnn)* (pp. 2921–2926).

Dijkstra, N., Ambrogioni, L., Vidaurre, D., & van Gerven, M. (2020). Neural dynamics of perceptual inference and its reversal during imagery. *Elife*, *9*, e53588.

Dijkstra, N., Bosch, S. E., & van Gerven, M. A. (2019). Shared neural mechanisms of visual perception and imagery. *Trends in cognitive sciences*, *23*(5), 423–434.

Fellous, J.-M., Sapiro, G., Rossi, A., Mayberg, H., & Ferrante, M. (2019). Explainable artificial intelligence for neuroscience: behavioral neurostimulation. *Frontiers in neuroscience*, *13*, 1346.

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal

gets weaker. *arXiv preprint arXiv:1706.06969*.

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

Goebel, R. (1992). Perceiving complex visual scenes: an oscillator neural network model that integrates selective attention, perceptual organisation, and invariant recognition. *Advances in neural information processing systems*, *5*.

Goodfellow, I., Lee, H., Le, Q., Saxe, A., & Ng, A. (2009). Measuring invariances in deep networks. *Advances in neural information processing systems*, *22*.

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

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Heinke, D., Leonardis, A., & Leek, E. C. (2022). What do deep neural networks tell us about biological vision? *Vision Research*, *198*, 108069–108069.

Hon, N., Thompson, R., Sigala, N., & Duncan, J. (2009). Evidence for long-range feedback in target detection: Detection of semantic targets modulates activity in early visual areas. *Neuropsychologia*, *47*(7), 1721–1727.

Hore, A., & Ziou, D. (2010). Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition* (pp. 2366–2369).

Huang, W., & Zhou, F. (2020). Da-capsnet: dual attention mechanism capsule network. *Scientific Reports*, *10*(1), 1–13.

Illing, B., Gerstner, W., & Brea, J. (2019). Biologically plausible deep learning—but how far can we go with shallow networks? *Neural Networks*, *118*, 90–101.

Ito, M., Tamura, H., Fujita, I., & Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of neurophysiology*, *73*(1), 218–226.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & 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.

Koenig-Robert, R., & Pearson, J. (2021). Why do imagery and perception look and feel so different? *Philosophical Transactions of the Royal Society B*, *376*(1817), 20190703.

Koriat, A., & Norman, J. (1985). Reading rotated words. *Journal of Experimental Psychology: Human Perception and Performance*, *11*(4), 490.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modelling biological vision and brain information processing. *biorxiv*, 029876.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 4.

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

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2018). Cornet: modeling the neural mechanisms of core object recognition. *BioRxiv*, 408385.

Leek, E. C., Leonardis, A., & Heinke, D. (2022). Deep neural networks and image classification in biological vision. *Vision Research*, *197*, 108058.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, *33*(10), 2017–2031.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fmri signal. *nature*, *412*(6843), 150–157.

Lyle, C., van der Wilk, M., Kwiatkowska, M., Gal, Y., & Bloem-Reddy, B. (2020). On the benefits of invariance in neural networks. *arXiv preprint arXiv:2005.00178*.

Mazzia, V., Salvetti, F., & Chiaberge, M. (2021). Efficient-capsnet: Capsule network with self-attention routing. *Scientific Reports*, *11*(1), 1–13.

Mwangi, B., Tian, T. S., & Soares, J. C. (2014). A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, *12*(2), 229–244.

Pearson, J. (2019). The human imagination: the cognitive neuroscience of visual mental imagery. *Nature reviews neuroscience*, *20*(10), 624–634.

Peters, J. C., Reithler, J., & Goebel, R. (2012). Modeling invariant object processing based on tight integration of simulated and empirical data in a common brain space. *Frontiers in computational neuroscience*, *6*, 12.

Poggio, T. A., & Anselmi, F. (2016). *Visual cortex and deep networks: learning invariant representations*. MIT Press.

Pucci, R., Micheloni, C., & Martinel, N. (2021). Self-attention agreement among capsules. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 272–280).

Qiao, K., Zhang, C., Wang, L., Chen, J., Zeng, L., Tong, L., & Yan, B. (2018). Accurate reconstruction of image stimuli from human functional magnetic resonance imaging based on the decoding model with capsule network architecture. *Frontiers in neuroinformatics*, *12*, 62.

Rakhimberdina, Z., Jodelet, Q., Liu, X., & Murata, T. (2021). Natural image reconstruction from fmri using deep learning: A survey. *Frontiers in neuroscience*, *15*, 795488.

Ramakrishnan, K., Scholte, S., Lamme, V., Smeulders, A., & Ghebreab, S. (2015). Convolutional neural networks in the brain: an fmri study. *Journal of vision*, *15*(12), 371–371.

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. *Advances in neural information processing systems*, *30*.

Senden, M., Emmerling, T. C., Van Hoof, R., Frost, M. A., & Goebel, R. (2019). Reconstructing imagined letters from early visual cortex reveals tight topographic correspondence between visual mental imagery and perception. *Brain Structure and Function*, *224*(3), 1167–1183.

Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., & Poggio, T. (2005). *A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex* (Tech. Rep.). MASSACHUSETTS INST OF TECH CAMBRIDGE MA CENTER FOR BIOLOGICAL AND ....

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid

categorization. *Proceedings of the national academy of sciences*, *104*(15), 6424–6429.

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

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

Svanera, M., Morgan, A. T., Petro, L. S., & Muckli, L. (2021). A self-supervised deep neural network for image completion resembles early visual cortex fmri activity patterns for occluded scenes. *Journal of Vision*, *21*(7), 5–5.

Vetter, P., Smith, F. W., & Muckli, L. (2014). Decoding sound and imagery content in early visual cortex. *Current Biology*, *24*(11), 1256–1262.

Wang, L., Zhang, Y., & Feng, J. (2005). On the euclidean distance of images. *IEEE transactions on pattern analysis and machine intelligence*, *27*(8), 1334–1339.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, *13*(4), 600–612.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 586–595).

Zhao, J., Li, J., Zhao, F., Yan, S., & Feng, J. (2017). Marginalized cnn: Learning deep invariant representations.