Novel Object Scale Differences in Deep Convolutional Neural Networks versus Human Object Recognition Areas

Astrid Zeman (astrid.zeman@kuleuven.be),
Brain and Cognition Department, Tiensestraat 102,
Leuven 3000, Belgium

Chayenne Van Meel (chayenne.vanmeel@kuleuven.be)
Brain and Cognition Department, Tiensestraat 102,
Leuven 3000, Belgium

Hans Op de Beeck (hans.opdebeeck@kuleuven.be)
Brain and Cognition Department, Tiensestraat 102,
Leuven 3000, Belgium

Abstract:

Deep Convolutional Neural Networks (CNNs) are lauded for their high accuracy in object classification, as well as their striking similarity to human brain and behaviour. Both humans and CNNs maintain high classification accuracy despite changes in the scale, rotation, and translation of objects. In this study, we present images of novel objects at different scales and compare representational similarity in the human brain versus CNNs. We measure human fMRI responses in primary visual cortex (V1) and the object selective lateral occipital complex (LOC). We also measure the internal representations of CNNs that have been trained for large-scale object recognition. Novel objects lack consensus on their name and identity, and therefore do not clearly belong to any specific object category. These novel objects are individuated in LOC, but not V1. V1 and LOC both significantly represent size and pixel information. In contrast, the late layers of CNNs show they are able to individuate objects but do not retain size information. Thus, while the human brain and CNNs are both able to recognise objects in spite of changes to their size, only the human brain retains this size information throughout the later stages of information processing.

Keywords: deep convolutional neural networks, size, object recognition, vision, classification

Introduction

Object recognition areas of the brain are tolerant to changes in size, position and pose (Li & DiCarlo, 2010; Nishimura, Scherf, Zachariou, Tarr & Behrmann, 2015). Deep Convolutional Neural Networks (CNNs) are also known to maintain their high object recognition accuracy with changes in object size, position and pose (Krizhevsky, Sutskever & Hinton, 2012). Size, position and other category-orthogonal properties have been shown to be decodable along the monkey object recognition pathway, and also along layers of a CNN, for known objects (Hong, Yamins, Majaj & DiCarlo, 2016). Here we investigate the similarity structure of the representation of novel objects presented at different sizes (in terms of visual field size or scale) in the brain, and in CNNs.

Methods

Stimulus Set

The stimulus set consists of 36 images, containing 12 novel objects (Figure 1) at 3 sizes (3.5, 7 and 11 degrees of visual angle). These objects are selected from the Novel Object and Unusual Name (NOUN) dataset (Horst and Hout, 2016). These objects are not easily identified or named. Images are presented in greyscale on an isoluminant background, at high resolution in the scanner (2400 x 2400 pixels) and at lower resolution to the network (256 x 256 pixels).

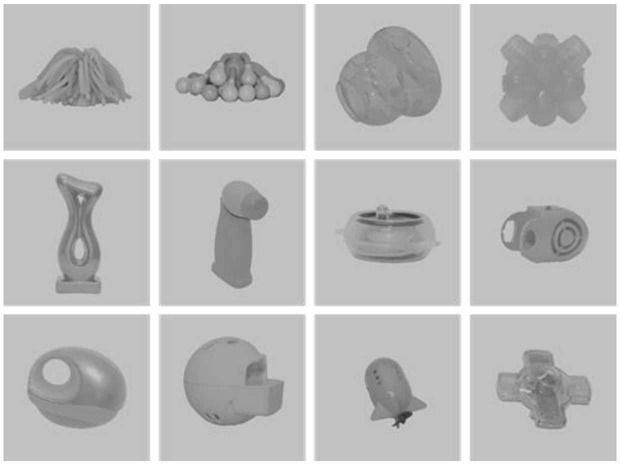

Figure 1 The stimulus set, largest scale presented here (of 3 scale sizes). Objects are a subselection from the NOUN database (Horst & Hout, 2016).

fMRI Experiment

We measured responses in V1 and LOC (Figure 2) of 23 participants (7 male, age range 23 – 37, mean age 23.9). All participants provided written informed consent and the experiment was approved by KULeuven Ethics Committee.

A neural dissimilarity matrix was computed for each ROI in every subject using beta weights (estimated in the GLM) from the voxels responding to each of the 36 stimulus conditions. We used cross-validated Mahalanobis distance (Walther, Nili, Ejaz, Alink, Kriegeskorte & Diedrichsen, 2016) as a measure of neural dissimilarity. This measure is the cross-validated Euclidean distance normalised by the covariance of the training sample.

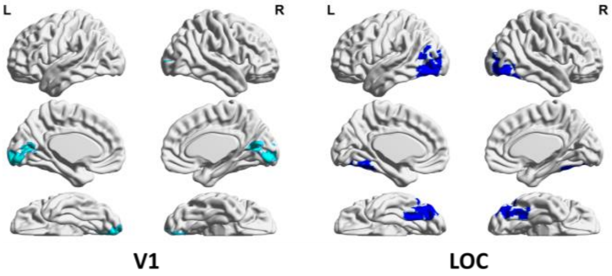

Figure 2 ROIs measured: V1 and LOC

Models

We define 4 conceptual models that highlight different features of the stimulus set. Each model is a Representational Dissimilarity Matrix (RDM) with size 36 x 36 (illustrated in Figure 3):

A. **Size** – defines the relative size of stimuli over the three sizes (3.5, 7 and 11 degrees of visual angle).

B. **Identity**– defines the same object across the three sizes.

C. **Shape** – defines the similarity in silhouette images. The binary image overlap is calculated per size, and the difference in shape is considered the same across sizes (i.e., images are normalized for size).

D. **Pixel similarity** – defines the normalised Euclidean distance between images. Unlike the shape model, values are different across different sizes. Objects presented at a small scale have a small difference in pixel similarity; larger objects have a larger difference. This is a low-level description that includes texture information, which is potentially biased in CNNs.

We found some overlap between the defined models (significance tested using random permutations), and so we also ran our analysis with partial correlations. Size was correlated with Pixel similarity (ρ = 0.49). Identity was correlated with Shape (ρ = 0.40). Shape was correlated with Identity (ρ = 0.40) and Pixel similarity (ρ = 0.11, Spearman).

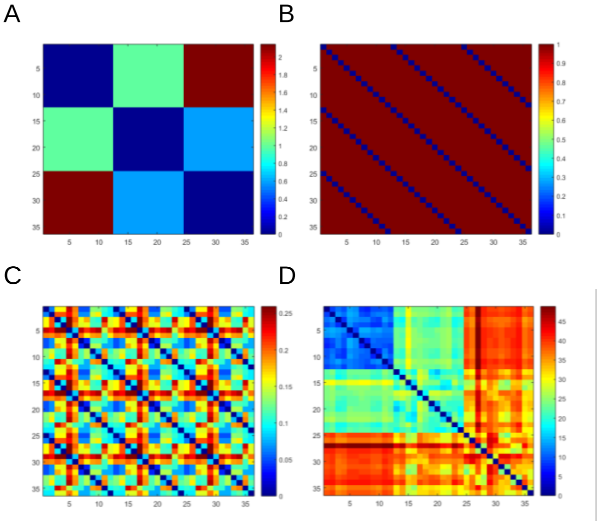

Figure 3 RDMs of each of the conceptual models A. Size, B. Identity, C. Shape D Pixel similarity. Numbering of stimuli are from left to right, top to bottom in Fig 1, from a small to large scale. Low values (dark blue) indicate that stimuli are highly similar, high values (dark red) indicate that stimuli are highly dissimilar.

CNNs

We measure representational correlations in CaffeNet, an 8-layer CNN with 5 convolutional layers and 3 fully connected layers. CaffeNet is an implementation of AlexNet (Krizhevsky, Sutskever & Hinton, 2012) and is trained on the ImageNet database. We also measure correlations in VGG-16, a CNN with 16 layers (13 convolutional and 3 fully connected), also trained on ImageNet (Simonyan & Zisserman, 2015).

Results

fMRI vs Models

Full correlations show that in V1, all models except for Identity are significant (df = 22, Size: t-stat = 18.13, p < 0.0001; id: t-stat = -2.37, p=0.99; Shape: t-stat = 2.62, p = 0.0078; Pix: t-stat = 13.34, p < 0.0001). In LOC, all models are significant (df = 22, Size: t-stat = 12.51, p < 0.0001; Id: t-stat = 7.89, p<0.0001; Shape: t-stat = 4.07, p = 0.0003; Pix: t-stat = 9.90, p < 0.0001). We ran a repeated measures ANOVA for ROIs, models and their interaction as within subject variables. For full correlations, the main effect of ROI is significant (df = 1, f stat = 9.865, p <0.0048), model is significant (df = 3, p < 0.0001, f-stat 168.7), and the interaction between ROI and models is significant (df = 3, p < 0.0001, f-stat = 27.52). Looking at the models separately, Size and Pixel similarity decrease significantly from V1 to LOC (size: df = 22, size t-stat = 5.046, p < 0.0001, Pixel: df = 22, t-stat = 4.6339, p = 0.0001). Identity increases significantly from V1 to LOC (df = 22, t-stat = -7.5409, p< 0.0001). Shape did not differ between ROIs (df = 22, t-stat = -0.8973, p = 0.3793).

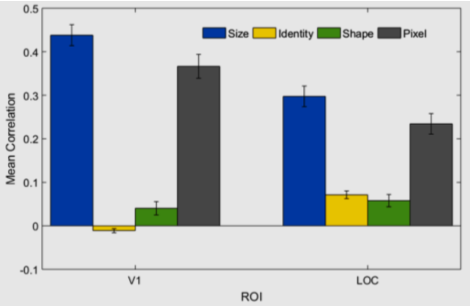

Figure 4 Full correlations between models and ROIs. Error bars indicate SEM.

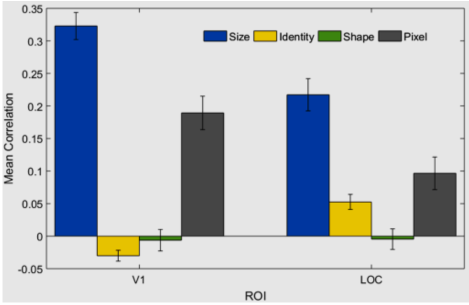

Figure 5 Partial correlations between models and ROIs. Error bars indicate SEM.

When taking partial correlations, the only difference with full correlations was that Shape information was no longer significant in V1 or LOC. All other correlations remained significant and ANOVA conclusions remained unchanged.

Figure 6 and Figure 7 show RDMs of V1 and LOC respectively. In V1, it is clear that Size and Pixel information are apparent. In LOC, Identity becomes visible through the appearance of lines that are parallel with the matrix diagonal.

Models vs CNNs

We correlate each model (Size, Identity, Shape and Pixel similarity) with each layer of a CNN. Looking at full correlations with 8-layer CaffeNet (Figure 8), Identity increases to above the significance threshold (red line, computed using random permutations) in the final fully connected layers. Pixel similarity is near one-to-one correlation with the convolutional layers, then decreases to the significance threshold in the final layer. Size and shape both start and end at the significance threshold in the first and last layers, peaking in the middle layers (layers 4 and 6 respectively).

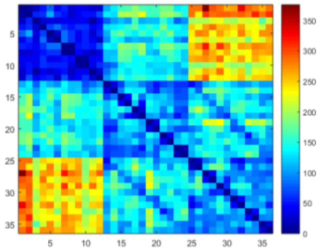

Figure 6 V1 RDM

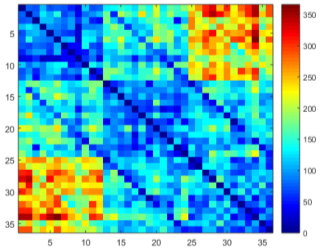

Figure 7 LOC RDM

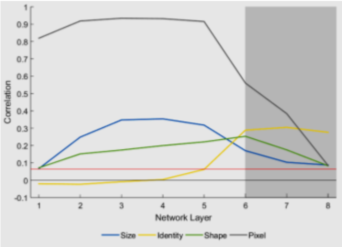

Figure 8 Full correlations between models and CaffeNet

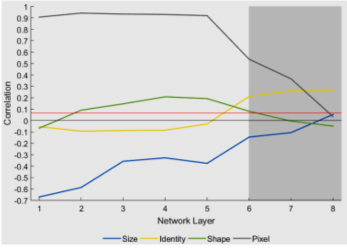

Figure 9 Partial correlations between models and CaffeNet

Looking at partial correlations (Figure 9), we see that Identity increases to above the significance level in the final fully connected layers. Pixel similarity decreases from ceiling correlations in convolutional layers to the significance threshold in the final layer. Shape has a similar profile as with full correlations. Size information remains below the significance threshold in all layers.

Note that the size model does not take into account pixel differences, which is why it correlates near zero in the first layer of Figure 10, and why it correlates negatively in Figure 11.

Figure 12 and Figure 13 show RDMs of respectively the first and last layer of CaffeNet (left) and VGG-16 (right). Common across both CNNs, we see that Pixel similarity is apparent in the first layer. In the last layer, Identity information is clearly visible. This shows that the networks are extremely tolerant to size changes.

Discussion and Conclusions

In CNNs, we find a high correlation with Pixel similarity in convolutional layers, while the last layer only shows a correlation with the Identity model without any remaining effect of Size. In contrast, Size remains important for object-selective areas in the human brain. Note that we also explored further areas in the ventral stream, but they did not contain reliable neural patterns to the novel objects in our stimulus set.

We refer to our manipulation of images as size or scale, however it is also possible to interpret this information in terms of which retinal fields are activated along the visual pathway to assist with locating an object. This information is of greater relevance to the brain compared to CNNs.

Hong et al. (2016) demonstrated that object recognition areas in the brain are tolerant to size, position and pose, and that this information increases along the ventral pathway. They showed that this increase is reflected in a 6-layer CNN with 1 fully-connected layer. The stimuli that they used contained known, real-world objects superimposed onto a naturalistic background. The stimuli that we used contained novel objects on an isoluminant grey background. We analysed the representations of CNNs that contained multiple (3) fully-connected layers, instead of a single fully-connected layer in Hong et al. (2016). In light of these differences, it appears that our results differ in two ways: 1) size information decreases in the brain, from V1 to LOC and 2) size information is not preserved in the final layers of CNNs.

Acknowledgments

This work was supported by the European Research Council (ERC- 2011-StG-284101), a Hercules grant (ZW11_10), the Fund for Scientific Research-Flanders (FWO 11X9918N and FWO G072517N), and the KU Leuven Research Council (C14/16/031).

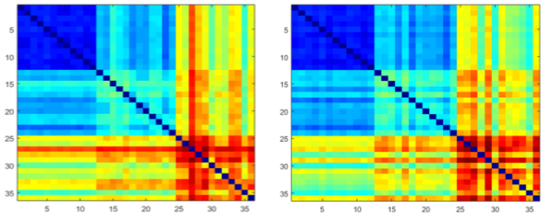

Figure 12 RDM of first layer in CaffeNet (left) and VGG-16 (right)

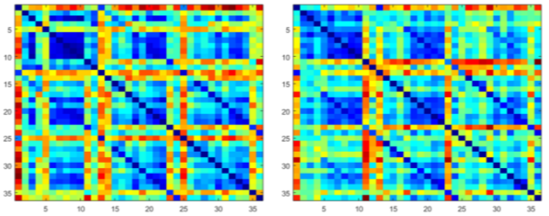

Figure 13 RDM of last layer in CaffeNet (left) and VGG-16 (right)

References

Hong, H., Yamins, D. L. K., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience, 19,* 613-622.

Horst, J. S., & Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behavior Research Methods, 48,* 1393-1409.

Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in neural information processing systems (NIPS), 1097-1105.

Li, N. & DiCarlo, J. J. (2010). Unsupervised natural visual experience rapidly reshapes size-invariant object representation in Inferior Temporal Cortex. *Neuron, 67,* 1062-1075.

Nishimura, M., Scherf, S., Zachariou, V., Tarr, M.J. & Behrmann, M. (2015). Size precedes view: developmental emergence of invariant object representations in Lateral Occipital Complex. *Journal of Cognitive Neuroscience, 27*, 474-491.

Simonyan, K. & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:*1409.1556.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, *137*, 188–200. http://doi.org/10.1016/J.NEUROIMAGE.2015.12.012