Deep neural networks trained on ecologically relevant categories better explain human IT

Johannes Mehrer[1] **(johannes.mehrer@mrc-cbu.cam.ac.uk)**

Tim C. Kietzmann[1] **(tim.kietzmann@mrc-cbu.cam.ac.uk)**

Nikolaus Kriegeskorte[1,2] **(nikolaus.kriegeskorte@mrc-cbu.cam.ac.uk)**

[1] MRC Cognition and Brain Sciences Unit, 15 Chaucer Road
CB2 7EF, Cambridge, UK
[2] Department of Psychology, Columbia University, 1190 Amsterdam Avenue
New York, New York, 10027, US

Abstract

Deep neural network models (DNNs) reach human-like performance in the computationally complex task of visual categorization, and also exhibit representational similarities with the human visual system. DNNs therefore enable researchers to investigate the mechanisms underlying cortical selectivity and organization by altering the training setup of the deep networks. Here, we explore whether using an ecologically more relevant set of image categories, rather than the ImageNet set frequently used in the engineering literature, may lead to receptive field properties that more closely match the human visual system. To this end, we introduce a new training set that consists of the 578 most concrete and frequent basic-level categories in the English language. Training 8-layer convolutional neural networks (CNNs) on this eco-set and a similar sized engineering set revealed that the ecologically more relevant visual diet led to significantly improved similarities to response properties in human inferior temporal cortex (IT). Although engineering datasets are a rich, and easily accessible source of training data, matching the human and networks' input statistics promises to lead to a better understanding of cortical function.

Keywords: deep neural networks; input statistics; representational geometries; visual diet; human visual system

Computer vision challenges and ecological validity

Deep neural networks (DNNs) have recently revolutionized computer vision and now regularly dominate several areas of artificial intelligence. Due to task performance and original biological motivation, computational neuroscientists have started investigating in how far DNNs can be used as model for information processing in the brain. Although DNNs largely abstract away from biological detail, they are nevertheless the currently best available models of the human visual cortex (Kietzmann, McClure, & Kriegeskorte, 2017; Kriegeskorte, 2015; Yamins & DiCarlo, 2016; Marblestone, Wayne, & Kording, 2016). Despite these early, yet promising results, however, it should be noted that the most commonly tested DNNs were trained to excel at a particular computer vision task, known as the ImageNet Challenge (ILSVRC 2012), rather than to explain and predict cortical function. Central to the challenge is the task to recognize 1,000 categories, for instance including 119 different breeds of dogs. From an engineering standpoint, this is highly sensible, as it allows computer vision systems to demonstrate their ability to perform even fine-grained visual classification.

In contrast to this, computational neuroscience uses DNNs as a probe to understand cortical function. DNNs can be altered based on their architectures, learning algorithms, objective functions, or input statistics, and changes in the predictive performance on neural datasets thereby allow for insights into computational mechanisms in the brain (Kietzmann et al., 2017). Here, we focus on the latter and investigate whether DNNs trained on ecologically more relevant categories can better explain representations of complex visual objects in human inferior temporal cortex (IT). Central to this work is the introduction of a new training set, which was designed to more closely match the human visual diet. First, categories were selected to be "concrete" rather than abstract (Brysbaert, Warriner, & Kuperman, 2014). Second, we focused on the most common categories by using linguistic occurrence statistics. The resulting list was then used as basis to distill a set of basic-level categories. The dataset comprises CC-SA-licensed images gathered from Bing (2%), Flickr (3%) and ImageNet (95%). Only including categories with at least 750 images (and selecting maximally 1,000 images per category) resulted in a total number of 569,413 images across 578 categories: the eco-set. To allow for a fair comparison to the previously used training sets, we randomly chose 578 categories from the ILSVRC 2012 and matched the number of images per category to the eco-set.

Based on these two sets, we then trained two architecturally identical deep convolutional neural networks (CNNs) and used representational similarity analysis to compare their internal representations to the human IT.

Materials and methods

To test the effects of different visual diets on the network's ability to match human cortical representational geometries,

we used a CNN, reminiscent of VGG-S (Chatfield, Simonyan, Vedaldi, & Zisserman, 2014). To reduce the overall number of parameters, we replaced the last fully-connected layers with convolutional layers of 1024 dimensions, and decreased the input image size from 224x224 to 128x128 pixels. The resulting network (VGG_8) contains seven convolutional layers and one linear read-out and has approximately 20 million parameters. We used a stride and zero-padding of 1 and a kernel size of 3 throughout the CNN, and varied the amount of maps per layer in the following way: 64-128-256-512-512-1024-1024. For regularization purposes gaussian presynaptic noise was used on all layers during training (McClure & Kriegeskorte, 2016).

Cortical responses in human IT were approximated using fMRI BOLD data recorded while 15 participants viewed 92 images of visual objects across two sessions each (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Kriegeskorte, Mur, & Bandettini, 2008). We then used representational similarity analysis (Nili et al., 2014) to compare the cortical representational geometries to the CNNs trained on either the trimmed ILSVRC set, or the novel eco-set. For this, we computed representational dissimilarity matrices (RDMs) for human IT, as well as each layer of the two networks, and used a Kendall tau-a correlation to compare the agreement between human and network RDMs.

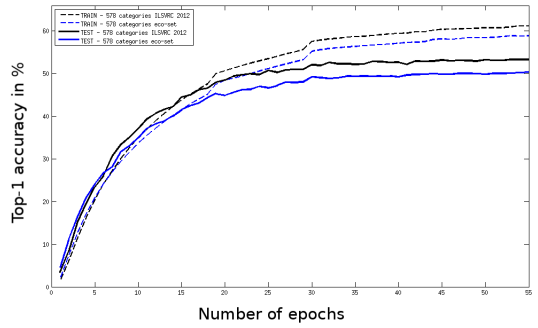

Figure 1: Classification performance of VGG_8: eco-set (blue) and the ILSVRC (black) during training (dashed) and testing (solid).

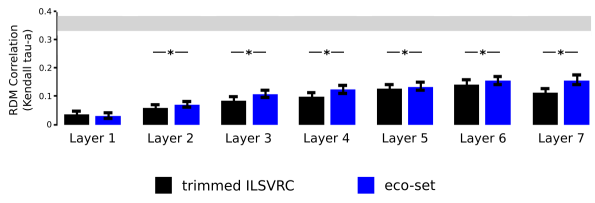

Figure 2: VGG_8 trained on eco-set (blue) explains significantly more variance in human IT than the same architecture trained on a trimmed ILSVRC (black). Statistical comparison based on a two-sided Wilcoxon signed-rank test, FDR corrected.

Results and conclusion

After training (Figure 1), the CNNs were tested for their ability to predict representational geometries of visual objects in human IT. The network trained on the new set of ecologically more relevant categories showed significant improvements over the trimmed ILSVRC. This was true for all but the very first layer (Figure 2). Importantly, this improvement cannot be attributed to higher training accuracy, as VGG_8 trained on the eco-set achieved a lower accuracy than the same architecture trained on the trimmed version of ILSVRC 2012.

Here we have shown that training DNNs based on ecologically more relevant categories can improve the representational similarities between artificial network and human IT. Task performance was lower for the network trained on the eco-set, suggesting that the visual diet, rather than increased task performance explains these improvements. Network training followed the standard training regime used in the ImageNet challenge, in which the probability of occurrence is approximately constant across categories. As a next step, we plan to investigate whether non-uniform probability distributions, matching real-world category frequencies, or category importance, can lead to further improvements in predicting human neural responses.

References

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, *46*(3).

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the Devil in the Details: Delving Deep into Convolutional Nets.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Deep Neural Networks predict Hierarchical Spatio-temporal Cortical Dynamics of Human Visual Object Recognition. *arXiv*, 15.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2017). Deep Neural Networks in Computational Neuroscience. *Bioarxiv*.

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008, jan). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*(November), 4.

Marblestone, A., Wayne, G., & Kording, K. (2016). Towards an integration of deep learning and neuroscience. , *10*(September).

McClure, P., & Kriegeskorte, N. (2016). Representing inferential uncertainty in deep neural networks through sampling. (Mcmc), 1–14.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. , *10*(4).

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3).