Architecture matters: How well neural networks explain IT representation does not depend on depth and performance alone

Katherine Storrs (Katherine.Storrs@mrc-cbu.cam.ac.uk)[1]
Johannes Mehrer (Johannes.Mehrer@mrc-cbu.cam.ac.uk)[1]
Alexander Walther (awalthermail@gmail.com)[2]
Nikolaus Kriegeskorte (Nikolaus.Kriegeskorte@mrc-cbu.cam.ac.uk)[1]
1. MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge, CB2 7EF, United Kingdom
2. YouGov, 50 Featherstone Street, London, EC1Y 8RT, United Kingdom

Abstract

Humans are able to classify complex visual objects with extremely high accuracy. Recently, deep convolutional neural network (DCNN) models have reached and even surpassed human performance at this task. Among recent networks, the deeper the architecture, the better the performance. Although loosely inspired by biological brains, it remains unclear whether models reaching human-level accuracy also perform computations similar to those in the human brain. In earlier studies using shallower architectures with poorer object classification accuracy, greater depth and higher task performance were associated with improved explanation of inferior temporal cortex (IT) (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). Our results show that this is not the case for state-of-the-art deep architectures that near or surpass human performance; the deepest, best-performing models are not best at explaining representations in human IT. In particular, deep residual networks (ResNets) are a relatively poor match to the brain, despite their very high classification performance. These findings open the door to detailed explorations of the architectures that best account for the representational transformations, and thus computations, performed in the ventral visual stream.

Keywords: object recognition; visual cortex; Deep Neural Networks; fMRI; representational similarity analysis

Deep convolutional neural networks (DCNNs) have dominated computer vision since AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) beat competing approaches on the 2012 ImageNet Large-Scale Visual Recognition Challenge (Russakovsky et al., 2015). Eight-layer object-classification-trained DCNNs such as AlexNet explain human inferior temporal cortex (hIT) better than other models tested (Khaligh-Razavi & Kriegeskorte, 2014), and display a correspondence between early–late network layers and early–late ventral visual regions (Güçlü & van Gerven, 2015). Since 2012, object classification performances have risen steeply thanks to deeper and more elaborate DCNN architectures, culminating in the near-human ability of deep residual networks (ResNets; He, Zhang, Ren, and Sun (2016)). Based on previous findings, we might expect that the greater depth and higher task performance of newer DCNNs will improve their ability to explain IT (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). Further, ResNets can be interpreted as recurrent networks, and so could be considered more biologically plausible than other, strictly feed-forward, architectures (Liao & Poggio, 2016). Here we test how well AlexNet and three deeper, higher-performing networks explain image representations in hIT.

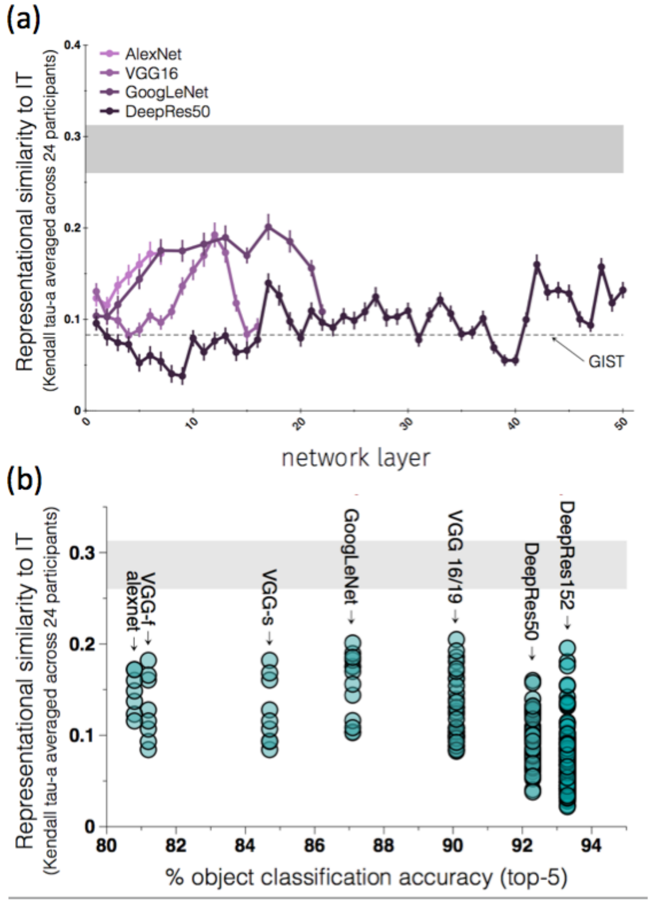

Figure 1: **Effect of architecture on ability of nets to explain hIT (a)** Correlation with hIT representational geometry as a function of layer number in four recent DCNNs. **(b)** Similarity to hIT as a function of ILSVRC object classification performance (each dot represents one layer in the labelled network).

Methods

Human fMRI data: We recorded responses to 62 colour images in 24 human subjects with 3T functional magnetic resonance imaging (fMRI). Multi-voxel activity patterns were extracted for each image from hIT. For each subject, a representational dissimilarity matrix (RDM; Nili et al. (2014)) was computed by taking the cross-validated Mahalanobis distances between the patterns elicited by each pair of images.

DCNN data: Activation patterns to the 62 images were recorded in four networks trained on the same 1,000-object classification task (Russakovsky et al., 2015): AlexNet (Krizhevsky et al., 2012), 16-layer VGG (Simonyan & Zisserman, 2015), GoogLeNet (Szegedy et al., 2015), and 50-layer DeepRes (He et al., 2016). For each layer of each network, an RDM was computed by taking the Pearson correlation distance (1-*r*) between patterns elicited by each image pair.

Representational similarity analysis: To compare the representation in hIT, for one subject, to that in a layer of a neural network, we calculated the Kendall tau-a rank correlation between the respective RDMs (Khaligh-Razavi & Kriegeskorte, 2014). Data points in Figure 1a depict the mean correlation across subjects for each layer of each network. Grey shaded regions in Figures 1a and 1b display the noise ceiling, showing the maximum possible performance of a model given the noise in the data. The noise ceiling is calculated as the average correlation between each subject's RDM and an RDM averaged across subjects, either excluding (lower bound), or including (upper bound) the target subject (Nili et al., 2014).

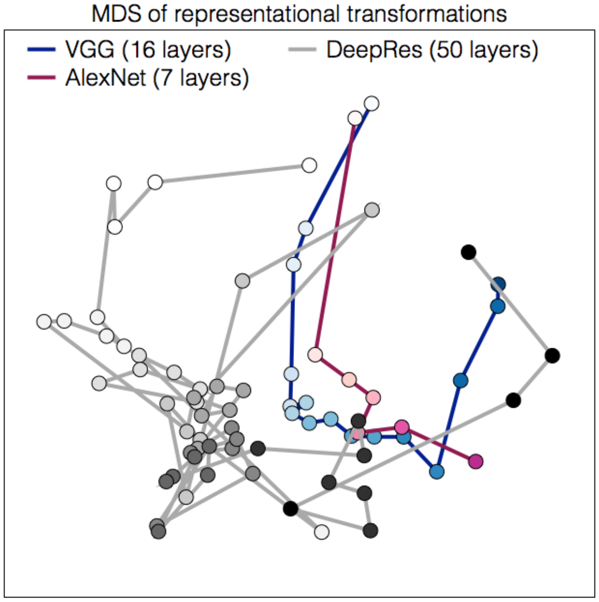

Figure 2: **Multi-dimensional scaling (MDS) of representational geometries in three networks**. Each dot represents the RDM for our 62 stimulus images in one layer of the respective network (pale colours = early layers; dark colours = late layers). While AlexNet and VGG cover similar paths in representational space, the 50-layer ResNet traverses a different and much less smooth path.

Results and Discussion

Depth and object classification do not fully explain hIT representation The ability of DCNNs to explain hIT representations is not fully determined by either the depth of the architecture nor the model's object classification accuracy. The highest hIT similarity was reached in 22-layer GoogLeNet, but was substantially poorer for a 50-layer ResNet.

Representational transformations differ between architectures Not all networks with the capacity to solve complex object recognition do so in the same way, even when trained on the same image set and task. Figure 2 shows a two-dimensional projection of the representational trajectories taken across the layers of three networks. Networks with similar architectures find similar routes (AlexNet and VGG-16), but those with substantially different architectures (here, a ResNet) may take substantially different paths through representational space. The analysis of representational trajectories via techniques such as multi-dimensional scaling may be helpful both in the design and understanding of DCNNs.

Our results suggest that the deepest state-of-the-art engineering solutions to object recognition may be diverging from biological solutions, and open the door to more detailed architectural explorations.

Acknowledgments

This work was supported by the UK Medical Research Council and a European Research Council Starting Grant to NK.

References

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

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *IEEE CVPR*, 770–778.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comp. Biol.*, *10*(11), e1003915.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. *NIPS*, 1097–1105.

Liao, Q., & Poggio, T. (2016). Bridging the gaps between residual learning, recurrent neural networks and visual cortex. *arXiv preprint arXiv:1604.03640*.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Comp. Biol.*, *10*(4).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *Int.J.Comp.Vis.*, *115*(3), 211-252.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *ICLR*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015). Going deeper with convolutions. *IEEE CVPR*, 1–9.

Yamins, D., Hong, H., Cadieu, C., Solomon, E., Seibert, D., & DiCarlo, J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS*, *111*(23), 8619–8624.