Predicting brain activation does not license conclusions regarding DNN-brain alignment: The case of Brain-Score

Gaurav Malhotra (gmalhotra@albany.edu)
Department of Psychology
SUNY – University at Albany, USA

Jeffrey Bowers (j.bowers@bristol.ac.uk)
School of Psychological Sciences
University of Bristol, UK

Abstract

The Brain-Score benchmark ranks how well DNNs model human "core object recognition" based on how well DNNs predict a wide range of neural and behavioural data. Here we focus on the Brain-Score predictions of IT neural activation and show that good predictions are not a good measure of DNN-IT alignment. We carry out a controlled experiment using the data from Majaj et al. (2015) to understand which visual features drive DNN brain predictions. We show that a good proportion of the neural predictivity score from the dataset are based on the backgrounds of images rather than the objects themselves. This reflects a more general problem of making claims regarding DNN-brain alignment based on correlational studies.

Keywords: Brain-Score; neural predictivity; object recognition; DNNs

A key observation taken to support DNN-brain alignment in the domains of vision and language is that DNNs can predict brain responses to photographs, text, and speech waveforms better than all alternative models. Indeed, predictions sometime approach (and even reach) noise ceilings, that is, DNNs predict brain responses as well as possible given the reliability of the brain recordings. This is taken to indicate that DNNs and brains learn similar visual and language representations for the sake of visual and language processing tasks such as object recognition and language comprehension.

Here we focus on the alignment between DNNs and vision in inferotemporal (IT) cortex. More specifically, we focus on the neural predictivity measure as implemented in the popular Brain-Score benchmark that assesses how well DNNs predict neural activation patterns along the ventral visual pathway (Schrimpf et al., 2018, 2020), culminating in IT. Brain-Score is used to rank DNNs in terms of how well they model "core object recognition". That is, the ability to quickly recognise objects despite substantial variation in appearance, such as changes in pose, context, and lighting conditions.

Brain-Score includes 4 datasets of IT neural responses to images taken from different studies. For two of the studies, macaques were presented with multiple exemplars of objects taken from eight categories in various random poses superimposed on a random background (Majaj, Hong, Solomon, & DiCarlo, 2015; Sanghavi & DiCarlo, 2020), as illustrated in Figure 1.

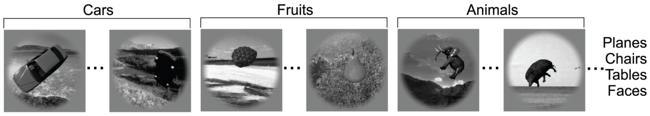

Figure 1: Some example images used in a series of studies to predict IT neural responses from DNN activations. In the dataset used by Majaj et al. (2015), there were eight different categories of objects, each consisting of 8 exemplars. Each exemplar was presented from a variety of viewpoints and on a variety of backgrounds. For illustrative purposes, this image is taken from Cadieu et al. (2014), who used a very similar dataset.

The two additional studies (Sanghavi, Murty, & DiCarlo, 2020; Sanghavi, Jozwik, & DiCarlo, 2020) included images that emphasized more naturalistic viewing conditions, but in all cases, the Brain-Score analysed these datasets as if they were observational. That is, DNNs were assessed in their ability to predict neural responses across all images, with no attempt to assess the impact of any manipulations, such as pose or background.

There is a problem with using predictions on observational datasets to draw conclusions regarding the mechanistic similarity between DNNs and IT. That is, the predictions in Brain-Score are correlational, and correlation does not imply a similarity in representations or mechanism. For example, humans primarily rely on shape when identifying objects (e.g., Biederman & Ju, 1988), whereas most DNNs rely on texture (Geirhos et al., 2018). This reflects the fact that shape and texture are correlated in images, and humans and DNNs rely on different correlated features. In the same way, good brain predictions might reflect confounds (e.g., texture representations in DNNs predicting shape representations in cortex). This ambiguity weakens any conclusions that can be drawn from these predictions. Indeed, if confounds are strong enough, it would be possible to get perfect brain (or behavioural) predictions between two systems that encode and represent completely different visual features.

To make stronger conclusions regarding DNN-brain representational and mechanistic alignment, it is necessary to understand which visual features drive DNN brain predictions and compare them to the features that drive biological per-

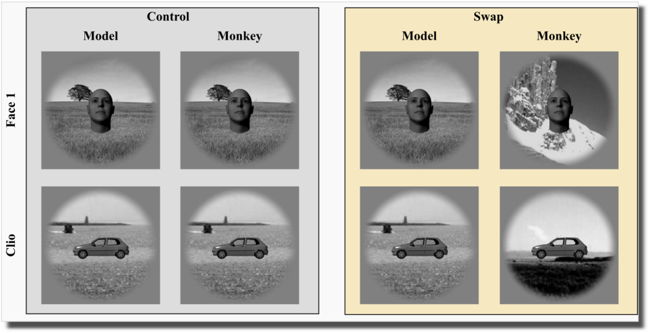

Figure 2: Example of pairs of images in the ‘Control’ and ‘Swap’ conditions. The same image is presented to the DNN and monkey in the Control condition. In the Swap condition, the model is presented with the exact same exemplar as viewed by the monkey (viewed from the exact same viewpoint) but on a different background.

ception and object recognition. In the case of Brain-Score, the question is whether the features that drive good predictions of neural data are like the features that drive core object recognition in macaques and humans. This requires an experimenter to systematically manipulate properties of images to test hypotheses regarding the visual features in DNNs that drive their predictions.

Fortunately, the availability of the Majaj et al. (2015) neural dataset that depicts objects in random poses on random backgrounds allowed us to run an experiment to test a simple hypothesis, namely, whether DNNs make predictions based the objects themselves or whether the random backgrounds also contribute to predictions. This latter outcome would undermine the claim that neural predictivity provides a measure of core object recognition, and more generally, highlight the danger of inferring anything about representational or mechanistic DNN-brain alignment based on predicting neural activation using observational data. This dataset includes neural recordings from 3200 images taken from 8 categories, with 8 exemplars per category and many different poses of each object, as in Figure 1. For each exemplar there is also a subset of images that depict the same objects in a canonical pose but in different backgrounds. This allowed us to compare predictions when the same objects were superimposed on the same background compared to when they were superimposed on different backgrounds. To the extent that predictions on this dataset reflect core object recognition, the background should have little or no effect on their predictions.

We trained the linear regression model on all the images in non-canonical poses as well as half of the images in canonical poses using the code from Brain-Score. With the remaining set of 168 image pairs, we compared the level of prediction of various DNNs when the DNNs and the macaques were presented with the same images in same (control) or different (swap) backgrounds conditions, as illustrated in Figure 2.

The results are shown in Figure 3. The critical finding is that predictions were reduced by more than half when the backgrounds mismatched, showing that the backgrounds played an important role in the brain predictions in this study. Note that neural predictivity studies are specifically designed with the intention that the score reflects core object recognition. This is why, when testing the regression model, the training and test images for each category have the same object but different (randomly chosen) background. However, our results indicate that, despite this manipulation, a large proportion of the prediction is performed on features that are not even part of the object in the scene.

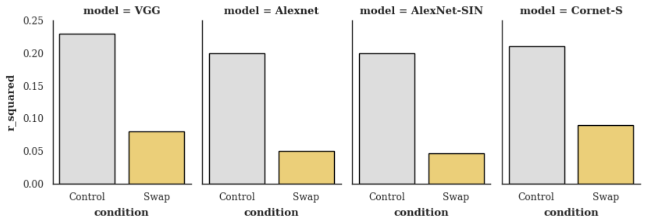

Figure 3: Neural predictivity scores in the Control and Swap conditions for four DNNs. Each panel shows the raw r^2 score (without scaling the scores for ceiling) for the held-out test images. This score is computed using the procedure described in Schrimpf et al. (2018) and captures the amount of variance in neural responses explained by the DNN activations.

Furthermore, note that there may well be other confounds in these images, and it is not safe to assume that the predictions obtained in the Swap condition reflect an updated estimate of core object recognition. For example, the prediction in the swap condition may reflect the similar textures of the two objects, and indeed, as noted above, past research has shown that many models primarily rely on texture to classify them (Geirhos et al., 2018). That is, it is possible that the predictions are simply be function of overlapping high frequency features, rather than features that are typically used by humans to perform object recognition.

We were able to assess the impact of the background confound because of the structure of this particular image dataset. But similar confounds are likely to be present in any dataset composed of high-dimensional images. This includes the other three IT datasets included in Brain-Score, and indeed, any behavioural or brain benchmark that assesses DNN-brain alignment using observational image datasets. Consistent with the possibility that confounds are driving predictions in observational studies, when DNNs are assessed on their ability to account for psychological experiments than manipulate independent variables to test specific hypothesis, DNNs tend to do extremely poorly (Bowers et al., 2023), and other measures being used to assess DNN-brain alignment, such as RSA, may be similarly prone to exaggerating scores due to confounds (Dujmović, Bowers, Adolfi, & Malhotra, 2022).

Going forward, it is important to run experiments that manipulate independent variables to determine which features drive good brain predictions and consider whether these features align with the features that drive human core object recognition. Experimental versions of Brain-Score will provide a more accurate assessment of DNN-Brain alignment.

References

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive psychology*, *20*(1), 38–64.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., ... others (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, e385.

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., ... DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, *10*(12), e1003963.

Dujmović, M., Bowers, J. S., Adolfi, F., & Malhotra, G. (2022). Some pitfalls of measuring representational similarity using representational similarity analysis. *bioRxiv*, 2022–04.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, *35*(39), 13402–13418.

Sanghavi, S., & DiCarlo, J. J. (2020). *doi:10.17605/OSF.IO/CHWDK* (Tech. Rep.).

Sanghavi, S., Jozwik, K. M., & DiCarlo, J. J. (2020). *doi:10.17605/OSF.IO/FHY36* (Tech. Rep.).

Sanghavi, S., Murty, N. A. R., & DiCarlo, J. J. (2020). *doi:10.17605/OSF.IO/FCHME* (Tech. Rep.).

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... others (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, *108*(3), 413–423.