DNN-based encoding models for the visual cortex fail to generalize out of the training data distribution.

Spandan Madan
Harvard University

Will Xiao
Harvard University

Mingran Cao
Francis Crick Institute

Hanspeter Pfister
Harvard University

Margaret Livingstone
Harvard University

Gabriel Kreiman
Harvard University

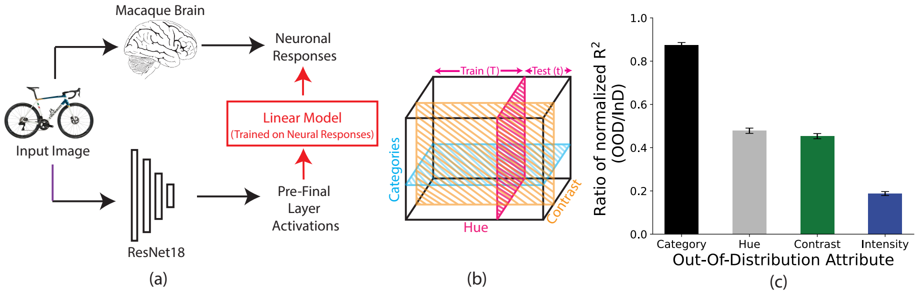


Figure 1: **Predicting neural responses for out-of-distribution images.** (a) DNN-based Encoding models of the visual cortex. A DNN was used to extract high-level image features, which served as input to a linear model trained to predict neuronal responses from the macaque IT cortex. (b) Constructing multiple OOD Train-Test splits. For every session, image-computable metrics (Hue, Contrast, Intensity) were calculated for all images and the 85th percentile value served as the cutoff—Images with metric lower than the cutoff served as the train split, with remaining images serving as the corresponding OOD test split. Thus, 4 such test splits were constructed per session—OOD Hue, OOD Contrast, OOD Intensity, and OOD Categories splits. (c) Ratio of Performance (as measured by regression R^2) on In-Distribution and OOD test splits. For all shifts, there is sharp drop in performance suggesting that DNN-based models of neural predictivity do not generalize well under distribution shifts.

Abstract

We characterize the generalization capabilities of DNN-based encoding models when predicting neuronal responses from the primate visual ventral stream. Using a large-scale dataset of neuronal responses from the macaque IT cortex to over $11,000$ images, we investigate the impact of distribution shifts on neural activity by dividing the images into multiple training and Out-Of-Distribution (OOD) test sets. This includes different types of OOD domain shifts in the form of image contrast, hue, intensity, and semantic object categories. Overall, we find models performed much worse at predicting neuronal responses for out-of-distribution images, dipping to as low as 20% of the performance on in-distribution test images. Furthermore, we found that this generalization performance under OOD shifts can be well accounted by an image similarity metric—the cosine distance between image representations extracted from a pre-trained object recognition model is a strong predictor of neural predictivity under different distribution shifts ($R^2 = -0.76$).

Keywords: Electro-Physiology; Object Recognition; Visual Cortex; Machine Learning; Out-of-distribution generalization

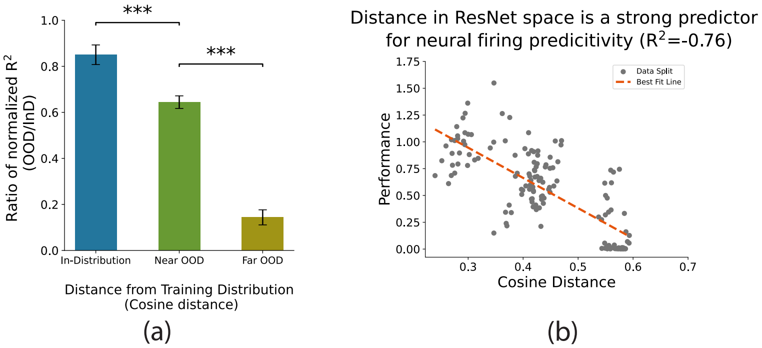
Introduction

Deep Neural Networks (DNNs) trained for object classification have remarkably similar internal feature representations to neural representations in the primate ventral visual stream (Bashivan, Kar, & DiCarlo, 2019; Ponce et al., 2019). This has led to the widespread use of encoding models of the visual cortex utilizing linear combinations of pre-trained DNN unit activities (Yamins et al., 2014; Kriegeskorte, 2015), as highlighted in Fig. 1(a).

However, DNNs struggle with generalization under distribution shifts, particularly when faced with out-of-distribution (OOD) samples (Hendrycks & Dietterich, 2019; Madan, Henry, et al., 2022; Madan, You, Zhang, Pfister, & Kreiman, 2022). This same difficulty at generalization may also affect DNN-based neural prediction. Recent work has investigated the impact of OOD categories on neural predictivity (Ren & Bashivan, 2023). Here, we show that many simplistic distribution shifts in the form of image contrast, hue, or intensity can break DNN-based models of the visual cortex, resulting in sharp drops in neural predictivity. Furthermore, we identify a suitable distance metric that accounts for neural-encoding-model generalization performance—image similarity as measured by cosine distance in the ResNet18 activation space.

Constructing out-of-distribution data splits

We collected extracellular electrophysiological responses from macaque Inferior Temporal (IT) Cortex to $11,000$ unique images, across 49 sessions. For every session and each



(a)

(b)

Figure 2: **Image similarity drives generalization performance.** (a) Comparing performance on new test samples from within the train distribution (In-Distribution) to test sets increasingly further away from the test distribution as measured by Cosine distance in ResNet feature space. Neural predictivity drops sharply as distance from training distribution increases. (b) Cosine Distance in ResNet feature space is a strong predictor for generalization accuracy. (Pearson $R^2 = -0.76$)

.

domain-shift condition, images were divided into two splits—a training split used to train the linear encoder model and a testing split used to evaluate the linear model. One split was made for each type of shift—Category, Hue, Contrast, and Intensity. For category split, a random subset of categories was held out from linear-model fitting. For image-computable metrics (e.g., Hue), the metric was computed for every image and the 85th percentile value served as the cut-off. Images with the metric value (e.g., Hue) less than the cut-off were assigned to the train split, and the remaining 15% served as the OOD test split. 4 OOD splits were made per session.

To investigate the relationship between image-similarity and neural predictivity, we construct 3 additional test splits. Starting with a random image in a session, we first sort all images based on this distanced. Images are then divided into three chunks. The first serves as the training + In-Distribution test split. The second is the Near OOD test split. Finally, the last chunk is the Far OOD split.

Results and Discussion

Ridge Regression was used to learn neural predictivity on a per-session basis. The regression model took as input ResNet18 activations (last layer) and predicted IT neuronal responses to these images. Note that the ResNet18 model (pretrained on ImageNet) was not fine-tuned. For each session, performance is measured as the regression R^2 normalized by the self-consistency of neurons. All results report mean performance over sessions, with error bars reporting the SEM (Standard Error of the Mean).

Neural Prediticivity under distribution shifts

The neural predictivity dropped under distribution shifts in comparison to predictivity for in-distribution images (Fig. 1 (c)). We reported the performance ratio between each type of OOD shift and an InD test set; ratios less than 1 reflect drop in performance. The drop is modest for held-out categories (0.84) and severe for Hue (0.46), Contrast (0.44), and Intensity (0.2). Thus, DNN-based models of the visual cortex generalize poorly out of the training data distribution even for shifts in low-level image features.

Image similarity drives generalization performance

We find that image similarity under distribution shifts largely explained the relative neural predictivity (Fig. 2). Specifically, the similarity between the train and test images, as measured by the cosine distance between ResNet image features (activations of the pre-final layer), is a strong predictor for neural predictivity under OOD shifts. Fig. 2 (a) shows model performance on test sets at various distances from the training distribution. As the train-test distance increases, performance drops significantly ($p < 0.001$, two-sided t-test). Furthermore, Fig. 2(b) represents individual train-test splits in the grouped results Fig. 2(a). We found the Pearson Correlation Coefficient (R^2) to be -0.76 suggesting that neural response prediction under OOD shifts is largely explained by image similarity.

Conclusion

These results reveal a deep problem in modern models of the visual cortex: good prediction is limited to the training image distribution. Simple distribution shifts break these models, consistent with broader findings that the underlying Deep Neural Networks are brittle to OOD shifts. Going one step further, we introduce an image-computable metric that significantly predicts the generalization performance under such distribution shifts. Together, we hope these contributions can facilitate future work to investigate and mitigate the generalization problem of state-of-the-art visual cortex models.

Acknowledgments

This research was partially supported by NSF grant IIS-1901030. We thank Pranav Misra, Fenil Doshi and Elisa Pavarino for insightful discussions.

References

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*(6439), eaav9436.

Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, *1*, 417–446.

Madan, S., Henry, T., Dozier, J., Ho, H., Bhandari, N., Sasaki, T., ... Boix, X. (2022). When and how convolutional neural networks generalize to out-of-distribution category–viewpoint combinations. *Nature Machine Intelligence*, *4*(2), 146–153.

Madan, S., You, L., Zhang, M., Pfister, H., & Kreiman, G. (2022). What makes domain generalization hard? *arXiv preprint arXiv:2206.07802*.

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S. (2019). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, *177*(4), 999–1009.

Ren, Y., & Bashivan, P. (2023). How well do models of visual cortex generalize to out of distribution samples? *bioRxiv*, 2023–05.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624.