Selective Is Selective, or Not? Investigating Consistency and Task Relevance of Selectivity Metrics in DNNs

Anastasia Lado (anastasia.lado@uni-giessen.de)
Department of Psychology
Justus Liebig University Giessen, Giessen, 35394, Germany

Katharina Dobs (katharina.dobs@uni-giessen.de)
Department of Mathematics and Computer Science, Physics, Geography
Justus Liebig University Giessen, Giessen, 35392, Germany
Center for Mind, Brain and Behavior
Universities of Marburg, Giessen, and Darmstadt, Marburg, 35032, Germany

Abstract

Neural response selectivity is a long-standing phenomenon in cognitive neuroscience, with distinct cortical areas selectively responding to visual categories across processing stages. Such selectivity is typically quantified using measures ranging from simple response ratios to detailed statistical comparisons between preferred and non-preferred categories. But how consistent and stable are these measures? And, critically, does selectivity capture behavioral relevance? Recent computational studies have shown that both trained and untrained deep neural networks (DNNs) exhibit category-selective units. Here, using face selectivity as our test case, we leverage DNNs to systematically compare a broad range of selectivity metrics while assessing their relevance to task performance. Our results reveal low agreement between selectivity metrics and lesioning-based rankings, and the consistency among metrics varies with spatial scale, processing stage, and training. However, all metrics yield similar face decoding accuracy. These findings caution against overreliance on any single metric and inform the interpretation of selectivity in computational and neural data.

Keywords: neural selectivity; deep neural networks; functional specialization; lesioning; visual cortex

Introduction

Neural response selectivity has long been established in visual neuroscience. For example, face-selective cortical areas—regions that respond more to faces than to other visual categories—have been identified across various processing stages and developmental periods (Kanwisher, 2017). Metrics used to localize these selective areas range from simple response ratios to detailed statistical comparisons between preferred and non-preferred categories (Stigliani et al., 2015). Despite their widespread use, a systematic comparison of these metrics has proven challenging due to the inherent limitations in neural data.

Recently, deep neural networks (DNNs) have emerged as promising models for testing neuroscience tools and probing the underpinnings of neural selectivity (Lindsay & Bao, 2023; Ratan Murty et al., 2021). Face-selective units have been identified in DNNs using selectivity metrics similar to those used in brain studies (Baek et al., 2021; Dobs et al., 2022; Lee et al., 2020; Prince et al., 2024). However, despite the variety of measures adopted in DNN analyses, their consistency and stability, and particularly their link to behavior, remain unclear.

In this study, we focus on face selectivity and use DNNs as an idealized testing ground to systematically compare a broad spectrum of selectivity metrics. Taking advantage of the fully observable and perturbable nature of these models, we compare these metrics with lesioning approaches and use identity decoding as proxy for behavioral relevance. We further examine how the consistency among selectivity measures varies across network layers and training.

Methods

DNN models. We used a dual-task VGG-16 network pretrained for simultaneous face recognition and object categorization (Fig. 1a; Dobs et al., 2022). We selected this DNN because it exhibits functional specialization for faces, with a distinct set of face-specific kernels whose lesioning selectively impairs face-task performance. In addition, we included another DNN trained on the same tasks and saved its initial checkpoint at random initialization, allowing us to measure the consistency of selectivity in trained and untrained networks.

Face selectivity metrics. Face selectivity was evaluated on the fLOC functional localizer dataset (Stigliani et al., 2015) for both individual units and convolutional kernels. At the kernel level, unit activation was aggregated using three summary statistics: mean activation, L2-norm, and max activation. We compared four established selectivity metrics based on activations extracted after ReLU in convolutional layers: (1) Face Selectivity Index (FSI), identifying units with responses at least twice as strong to faces as to non-face stimuli (FSI>0.3); (2) d-prime selectivity, applying a selectivity criterion of d'>0.65; (3) Wilcoxon rank-sum test, a non-parametric test identifying units that significantly prefer faces over all other classes (p<0.001); and (4) t-test-based selectivity, involving pairwise comparisons between faces and multiple non-face categories at the individual unit level, followed by intersection across categories using an FDR-corrected significance threshold (p<0.05).

Measuring consistency. To assess agreement among selectivity metrics, in each layer we calculated pairwise Spearman rank correlations between metrics. Specifically, for a given metric A, we rank all units or kernels by their selectivity values (e.g., t-values), then compute Spearman's correlation between that ranking and the continuous selectivity values from metric B (e.g. d') for the same items.

Face identity decoding. To probe behavioral relevance, we trained linear support vector machine (SVM) classifiers to decode face identity from the l2-norm aggregated activations of selected kernels. To control for differences in the number of selected

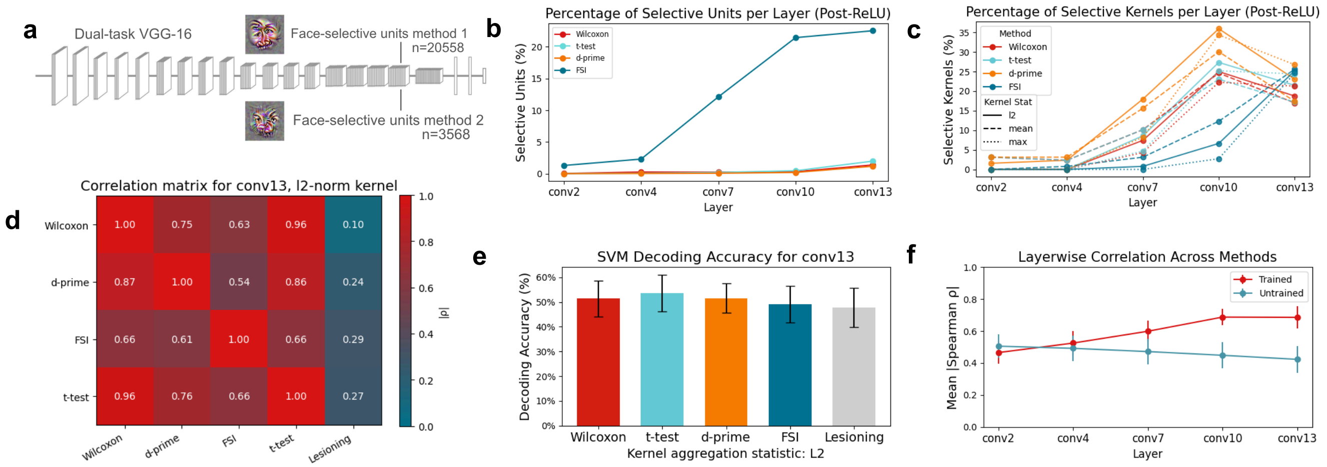

Figure 1: **a.** Dual-task VGG-16 architecture used to identify face-selective units via multiple metrics and lesioning. Percentage of face-selective **b.** units and **c.** kernels per metric and aggregation method. **d.** Metric-lesioning Spearman correlation for conv13 kernels (l2-norm). **e.** Face identity decoding accuracy using selected kernels in conv13 (equal kernel count). Error bars: 95% CI across cross-validation folds. **f.** Mean inter-metric Spearman correlation across layers for trained (red) vs. untrained (blue) DNNs. Error bars: SEM across metric pairs.

kernels, this analysis was restricted to the minimum number of kernels identified across metrics (N=96). SVMs were trained on a dataset of 100 face identities (10 images each) using leave-one-image-out cross-validation. Decoding accuracy serves as a proxy for how well each selectivity metric identifies kernels contributing to task performance.

Results

We first analyzed the percentage of face-selective responses across five layers (conv2, conv4, conv7, conv10, conv13) for both individual units (Fig. 1b) and aggregated kernels (Fig. 1c). At the unit level, FSI identified up to 20% of units as face-selective, substantially more than those identified by the Wilcoxon test, t-test, and d-prime methods (max. 3%). In contrast, at the kernel level, d-prime metric was most sensitive, identifying up to 35% of kernels as selective. Moreover, when comparing kernel summary statistics, mean aggregation consistently resulted in fewer selective kernels than the L2 norm or maximum value. Notably, only FSI showed a clear increase in selectivity with deeper layers.

To determine which metric best aligns with lesion-based ranking of face-selective kernels, we correlated the ranks of kernels identified by each metric with those obtained from lesioning. FSI approximated the lesioning ranks most closely (*r* = -0.29; Fig. 1d). In contrast, the Wilcoxon test, t-test, and d-prime metrics showed similar ranking patterns to each other, yet these were distinct from both the lesioning and FSI rankings.

Next, we evaluated the behavioral relevance of the selected kernels by decoding face identity from their activations (Fig. 1e). Controlling for the number of selected kernels, decoding accuracy was highly similar across all metrics. This result suggests that each captures comparable task-relevant information despite differences in selectivity rankings.

Finally, we examined whether training and network depth affected the consistency among selectivity metrics at the unit level (Fig. 1f). Mean inter-metric consistency was higher in the trained network than in the untrained one from mid-level layer conv7 onward. Moreover, consistency increased with layer depth only in the trained network, suggesting that training consolidates representational stability.

Conclusion

Our results reveal the complex, scale-dependent nature of selectivity metrics in DNNs. We found that different metrics can dramatically differ in the proportions of identified face-selective units or kernels, and that inter-metric consistency depends on processing stage and training. While all metrics yielded comparable task-relevant information in identity decoding, only FSI exhibited a continuous increase in face selectivity across layers and aligned most closely with lesioning-derived rankings. Future work should investigate other category selectivities and diverse network architectures to test generality. Overall, our findings offer key insights into selecting and interpreting selectivity measures and underscore the need for careful metric choice in computational and cognitive neuroscience.

Acknowledgments

This work was supported by the European Research Council (ERC Starting Grant DEEPFUNC, ERC-2023-STG-101117441), the Hessian Ministry of Higher Education, Research, Science and the Arts (LOEWE Start Professorship and Excellence Program “The Adaptive Mind”), and the German Research Foundation (Collaborative Research Center SFB/TRR 135).

References

Kanwisher, N. (2017). The quest for the FFA and where it led. Journal of Neuroscience, 37(5), 1056–1061. https://doi.org/10.1523/JNEUROSCI.1706-16.2016

Dobs, K., Martinez, J., Kell, J. E. A., & Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. Science Advances, 8, eabl8913. https://doi.org/10.1126/sciadv.abl8913

Prince, J. S., Konkle, T., & Alvarez, G. A. (2024). Contrastive learning explains the emergence and function of visual category-selective regions. Science Advances, 10, eadl1776. https://doi.org/10.1126/sciadv.adl1776

Baek, S., Song, M., Jang, J., Kim, G., Paik, S.-B., & Paik, S.-B. (2021). Face detection in untrained deep neural networks. Nature Communications, 12, 7328. https://doi.org/10.1038/s41467-021-27606-9

Lee, H., Margalit, E., & Yamins, D. L. K. (2020). Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. bioRxiv. https://doi.org/10.1101/2020.07.09.185116

Stigliani, A., Weiner, K. S., & Grill-Spector, K. (2015). Temporal processing capacity in high-level visual cortex is domain specific. *Journal of Neuroscience, 35*(36), 12412–12424. https://doi.org/10.1523/JNEUROSCI.4822-14.2015

Lindsay, G. W., & Bau, D. (2023). Testing methods of neural systems understanding. Cognitive Systems Research, 82, Article 101156. https://doi.org/10.1016/j.cogsys.2023.101156

Ratan Murty, N. A., Bashivan, P., Abate, A., et al. (2021). Computational models of category-selective brain regions enable high-throughput tests of selectivity. Nature Communications, 12, 5540. https://doi.org/10.1038/s41467-021-25409-6