Modelling inter-animal variability

Javier Sagastuy-Brena* (jvrsgsty@stanford.edu), Imran Thobani* (ithobani@stanford.edu),
Aran Nayebi (anayebi@stanford.edu), Rosa Cao (rosacao@stanford.edu), Dan Yamins (yamins@stanford.edu)
Stanford Neuroscience and Artificial Intelligence Laboratory, Stanford University, CA, USA
* Equal contribution

Abstract

Accurately measuring similarity between different animals' neural responses is a crucial step towards evaluating deep neural network (DNN) models of the brain. Under what transform class are animals likely to be similar to each other, and how much neural data needs to be collected to get an accurate similarity estimate? Using model variability as a proxy for inter-animal variability, we find that *where* we measure similarity from has critical implications for the suitable transform class. Specifically, we observe high linear mappability between pre-ReLU activations, but require a simple non-linear mapping class (that combines logistic regression with linear regression) in the case of post-ReLU activations. With our approach, we estimate that measuring inter-animal variability requires collecting neural data for at least 500 stimuli and 300 neurons from the same hypercolumn, providing a prescription for future experimental data that can adjudicate between models.

Keywords: transform similarity; variability; deep neural networks; representations; neural predictivity; linear regression; logistic regression; mouse visual cortex

Introduction

A plausible standard for accurate DNN models of the brain requires a DNN's neural responses to be at least as similar to a given animal's neural responses as two conspecifics' neural responses are to each other (Cao & Yamins, 2021). However, estimating inter-animal variability is challenging, because most extant datasets are statistically underpowered. In this work, we make some progress on this problem by using *model* variability as a proxy for *inter-animal* variability, in order to answer two questions:

1. What metrics are useful for comparing neural responses?
2. How much data (stimuli, neurons) would we need to accurately estimate inter-animal variability?

Method

Studying model variability as a proxy for animal variability requires having models that are reasonably similar in their neural responses to the animals, as well as having a source of variation in the models that is a reasonable proxy for variation between animals. Given the widespread use of rodents in experimental neuroscience, we use the current state-of-the-art unsupervised DNN models of mouse visual cortex, based on the AlexNet architecture (Nayebi et al., 2021; Krizhevsky, 2014). To introduce model variability, we use different random seeds, which control the weight initialization and the order of training data seen by the neural networks. The working assumption is that two models trained from different random seeds have roughly the same variability in neural responses as two different mice.

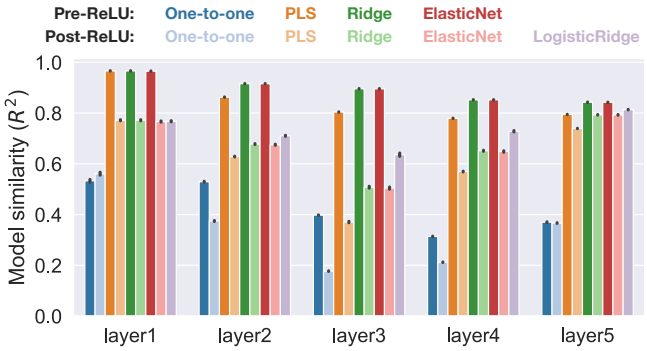

Figure 1: **Comparing R^2 scores across mappings and pre- vs post-ReLU** Ridge regression achieves the highest R^2 scores between corresponding model layers for pre-ReLU activations, but LogisticRidge performs slightly better than ridge for post-ReLU activations. The R^2 score is aggregated over target units (median), for 10 train-test splits (mean) and 6 model pairs (mean), training the mappings on 8,000 stimuli.

Criteria for choosing a similarity measure The goal is to determine according to what similarity metrics different models' neural responses (and possibly animals') are highly similar to each other. In order to compare *different* similarity metrics to each other, we restrict ourselves to considering similarity measures that take the form of predictive mappings from a source animal's neural responses in a given layer to a target animal's neural responses in the same layer. The similarity score is defined as the coefficient of determination (R^2) of such a mapping on a held-out test set.

A good similarity metric should be as constrained as possible while still rating models (or animals) as highly similar. To the extent the mapping class is highly constrained, the fact that neural responses can be mapped to each other identifies a stronger sense of similarity.

Results

Pre-ReLU activations enable higher R^2 scores When evaluating the neural predictivity of a DNN model of the brain, it has long been the case that the features from the model used to predict neural responses correspond to post-

nonlinearity activations. Such a choice aims to use data from the model that is analogous to the action potentials measured in animal data. When predicting a model's features from another model's features, it is feasible to use the pre-nonlinearity features for both inputs and outputs, which might be analogous to membrane potentials. We observed a significant advantage in predictive accuracy of the mappings trained on pre-ReLU features as can be seen in Fig. 1. This implies that neural networks are much more similar to each other than they appear if you only look at the post-ReLU activations, which tend to be sparse.

One hypercolumn is all you need Doing regressions only between the neurons in a model layer with receptive fields at the center of the image leads to similarity scores that are almost as high as using all the neurons in the layer. This reduces the number of neurons required to achieve a high regression score. This column alignment is analogous to recording from the same hypercolumn in the source and target animals.

The best performing metrics We found ridge regression to be the best-performing mapping class for pre-ReLU features. However, it underperforms when mapping post-ReLU features (Fig.1). To deal with the sparseness of the post-ReLU activations, we introduce LogisticRidge, a mapping that predicts the sparsity pattern of the targets and only performs ridge regression for the non-zero target units. This is the best-performing mapping on post-ReLU features.

Number of neurons and stimuli needed to estimate transform similarity We perform unit and stimuli subsampling on our models to estimate a lower bound on the amount of data we need (in terms of number of neurons and stimuli we have neural responses for) to estimate model variability. To do the unit subsampling, we pooled units from 9 different source models together, treating them as a single "animal" in order to estimate the high end of the sampling range. As Fig. 2 shows, the R^2 score saturates much earlier for pre-ReLU features compared to post-ReLU features, at about 100 units and 200 training stimuli.

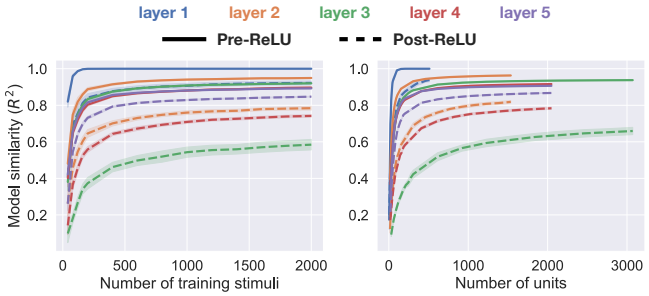

Figure 2: **Stimuli and unit subsampling** Ridge regression R^2 begins to saturate at about 500 and 200 training stimuli, and 500 and 100 model units, for post- and pre-ReLU features. For the left panel, we used all units in a model hypercolumn. For the right panel, we used 8,000 training stimuli.

V. Area	Layers	α^*
VISp	1	1.4756
VISl	2,3,4	0.6538
VISrl	3	0.1915
VISal	2,3,4	0.2135
VISpm	2,3	0.3447
VISam	2,3	0.4551

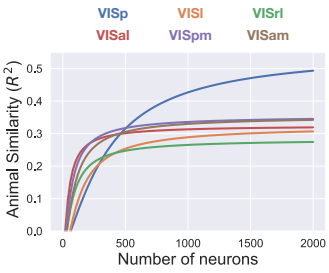

Figure 3: **Neuron-to-unit ratios ($\alpha*$) and estimated animal neuron sampling curve** The curves represent our modeling estimates for animal similarity as you collect data from more neurons while keeping the number of stimuli fixed at 118.

Estimating the neuron-to-unit ratio In order to apply our unit subsampling analysis to the case of animal variability, we estimated a ratio of model units to animal neurons, since one unit is unlikely to be functionally equivalent to one neuron. Under the assumption that our model population has roughly the same variability as the animals, we expect the subsampling curves for units and neurons to overlap with each other once the ratio of animal to model neurons is taken into account.

We match each visual area in mouse cortex to the model layers that best predict that visual area(s) according to (Nayebi et al., 2021). We performed subsampling on both pooled source model units as well as mouse neural responses (de Vries et al., 2020; Siegle et al., 2021) for the same 118 stimuli to get curves similar to those shown in Fig.2.

We modeled the unit subsampling curve for models as having the form: $f(m) = L_f \arctan(k_f m) + b_f$, where $f(m)$ is the R^2 score for a given number of model units m. For the animals, we fit $g(n) = \log(k_g n) + b_g$, for animal neurons n. This curve can be used to extrapolate the inter-animal similarity for high neuron counts. We assume $m = n/\alpha$ and minimize the distance between f and g in the interval $[0, N_l]$, where N_l is the max. number of pooled source neurons in visual area l:

$$\alpha^* = \text{argmin}_\alpha \int_0^{N_l} (f(n/\alpha) - g(n))^2 dn$$

The resulting animal neuron sampling curve (obtained by transforming the model subsampling curve by the calibration ratio) is presented in Fig. 3.

Conclusion

Our results show that pre-ReLU activations from different networks are strongly similar to each other according to linear regression, and that by sparsifying the data, the ReLU layer masks this similarity. In the future, it may be beneficial to collect electrophysiological data on membrane potentials, not just action potentials, in order to measure animal transform similarity. Finally, the fact that LogisticRidge outperforms Ridge on post-ReLU activations suggests that the correct transform class for post-ReLU activations is non-linear, and motivates future research in refining the true transform class between post-ReLU activations.

Acknowledgments

J.S. thanks the Mexican National Council of Science and Technology (CONACYT) for support. D.L.K.Y. is supported by the James S. McDonnell Foundation (Understanding Human Cognition Award Grant No. 220020469), the Simons Foundation (Collaboration on the Global Brain Grant No. 543061), the Sloan Foundation (Fellowship FG-201810963), the National Science Foundation (RI 1703161 and CAREER Award 1844724), the DARPA Machine Common Sense program, and hardware donation from the NVIDIA Corporation.

References

Cao, R., & Yamins, D. (2021). Explanatory models in neuroscience: Part 1–taking mechanistic abstraction seriously. *arXiv preprint arXiv:2104.01490*.

de Vries, S. E., Lecoq, J. A., Buice, M. A., Groblewski, P. A., Ocker, G. K., Oliver, M., ... others (2020). A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature neuroscience*, *23*(1), 138–151.

Ivanova, A. A., Schrimpf, M., Anzellotti, S., Zaslavsky, N., Fedorenko, E., & Isik, L. (2021). Beyond linear regression: mapping models in cognitive neuroscience should align with research goals. *bioRxiv*. doi: 10.1101/2021.04.02.438248

Krizhevsky, A. (2014). One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*.

Nayebi, A., Kong, N. C., Zhuang, C., Gardner, J. L., Norcia, A. M., & Yamins, D. L. (2021). Unsupervised models of mouse visual cortex. *bioRxiv*.

Siegle, J. H., Jia, X., Durand, S., Gale, S., Bennett, C., Graddis, N., ... others (2021). Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, *592*(7852), 86–92.

Williams, A. H., Kunz, E., Kornblith, S., & Linderman, S. (2021). Generalized shape metrics on neural representations. *Advances in Neural Information Processing Systems*, *34*, 4738–4750.