Topographic Deep ANN Models Predict the Perceptual Effects of Direct IT Cortical Interventions

Martin Schrimpf
School of Life Sciences, School of Computer and Communication Sciences, NeuroX Institute
EPFL, Lausanne, 1015 Switzerland

Paul McGrath
McGovern Institute for Brain Research
MIT, Cambridge, MA 02139 USA

Eshed Margalit
Neurosciences Graduate Program
Stanford University, Stanford, CA 94305 USA

James J. DiCarlo
McGovern Institute for Brain Research, Department of Brain and Cognitive Sciences, Quest for Intelligence
MIT, Cambridge, MA 02139 USA

Abstract

Ever-advancing artificial neural network (ANN) models of the ventral visual stream capture core object recognition behavior and the neural mechanisms underlying it with increasing precision. We here extend this modeling approach to make and test predictions of neural intervention experiments. Specifically, we enable a new prediction regime for topographic deep ANN (TDANN) models of primate visual processing through the development of *perturbation modules* that translate micro-stimulation, optogenetic suppression, and muscimol suppression into changes in model *neural activity* which unlocks predicting downstream *behavioral* effects. Without any fitting, we find that TDANN models generated via co-training with both a spatial correlation loss and a standard categorization task qualitatively predict key behavioral results from several primate IT perturbation experiments. In contrast, TDANN models generated via random topography fail to predict nearly all primate results. Taken together, these findings indicate that current topographic deep ANN models paired with perturbation modules are reasonable guides to predict the qualitative results of direct causal experiments in IT.

Keywords: Primate Vision; Topographic Deep Neural Networks; Cortical Interventions; Causal Perturbations

Introduction

Certain Artificial Neural Networks (ANNs) have recently been shown to be surprisingly good models of the neural mechanisms underlying primate core object recognition (Yamins & DiCarlo, 2016; Schrimpf et al., 2020). When presented with the same image stimuli as primate subjects, these models' internal activity resembles those in brain recordings, and model behavioral choices resemble subject behavior. However, a crucial area of neuroscience research has so far been neglected by these models: direct neural perturbations that causally link neural activity to behavioral outcomes. Such experimental studies perturb a piece of neural tissue and measure the effects on perception as measured via behavioral reports. Modeling causal cortical interventions force a coherent model that links neural areas to behavioral predictions, and are a potential stepping stone towards model-guided brain-machine interfaces such as visual prosthetics.

In this study, we build and evaluate ANN models to predict the behavioral effects of direct neural perturbations in inferotemporal cortex (Figure 1). Our contributions are:

1. We develop perturbation modules that formally define the effects of different types of neural perturbations (micro-stimulation, optogenetic and muscimol suppression) on *neural* activity, based on previous biophysical studies.

2. We convert results from four studies measuring the *behavioral* effects of direct IT neural perturbations into benchmarks for model evaluation.

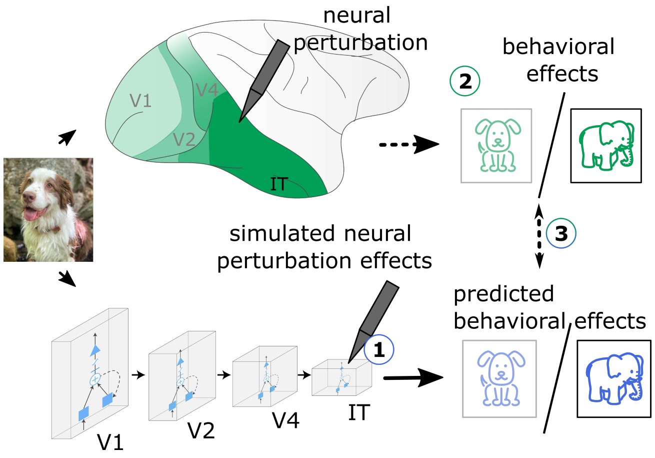

Figure 1: **Predicting the behavioral effects of direct neural perturbations with computational models.** Primate perturbation experiments change neural activity while subjects process visual stimuli, and measure changes in behavior (top, green). To model these experiments, we ① build fully specified perturbation modules defining the change in neural activity from different perturbations, ② convert primate perturbation experiments into model benchmarks, and ③ test model predictions on these experimental benchmarks.

3. We find that topographic models combined with perturbation modules qualitatively predict the effects measured in the contra-lateral hemifield across all studies.

Modeling causal perturbations

The rationale of our modeling approach is to assume a direct correspondence between the stages of neural processing in ANN models and the primate visual ventral stream. For all ANN ventral stream models, we organize neurons into 2D ”tissue” with different strategies for the spatial assignment.

Perturbation modules based on neural data. To simulate the effects of neural interventions on a given ventral stream model, we developed a set of perturbation modules (Figure 1b). Each module takes as input perturbation parameters (e.g. the amount of muscimol in an injection) and outputs the resultant change in neural activity as a spatial profile. We developed each perturbation module based only on the *neuronal* effects of muscimol, optogenetics, and micro-stimulation. In other words, none of these modules were tuned to fit any behavioral effects that we later test the models on.

Topographic assignment of model IT neurons. To simulate the spatial profile of perturbations in models of the ventral stream, model neurons must first be assigned to x and y locations within artificial cortical tissue. We primarily focus on topographically optimized models, where neurons are initially assigned fixed positions and model parameters are trained jointly with a classification and a spatial correlation loss (Lee

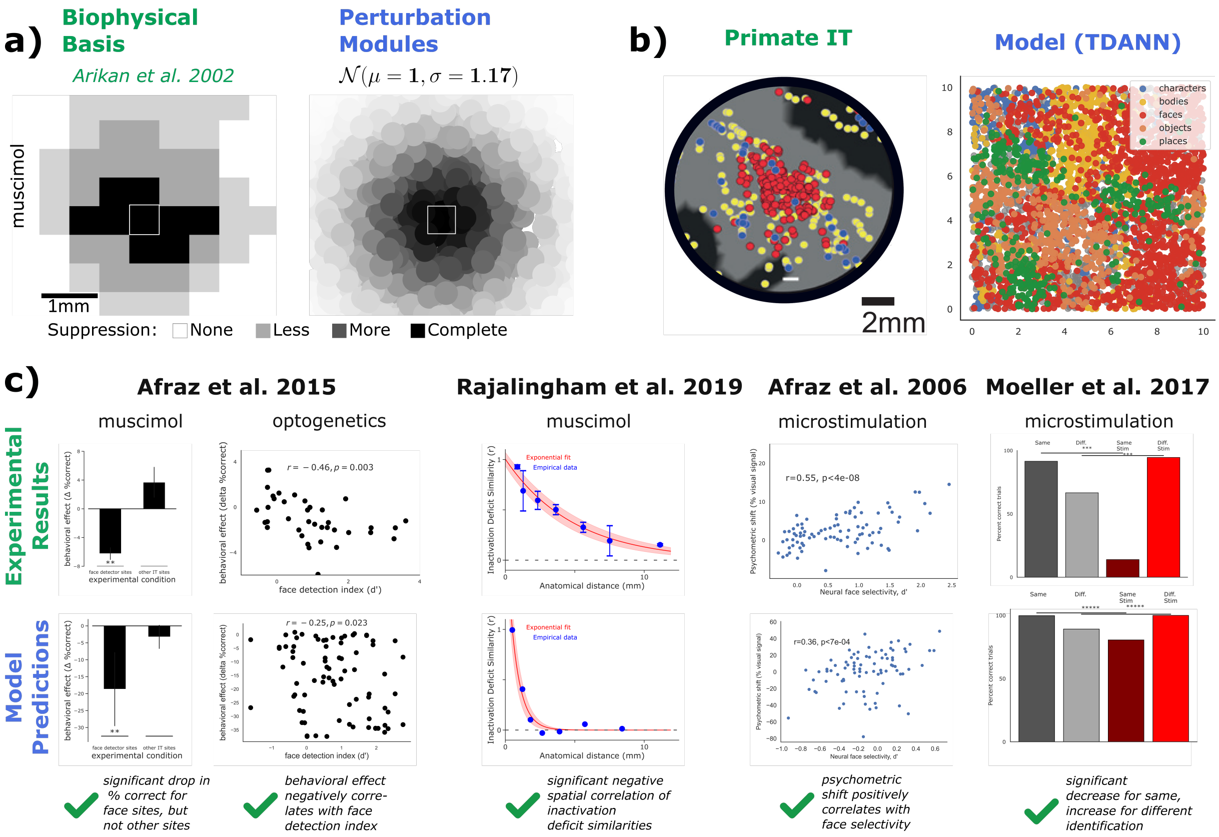

Figure 2: **A topographic model predicts the behavioral effects of direct neural perturbations.** **a)** Perturbation modules aligned via prior neural results. Each perturbation module specifies how neural interventions affect model neural activity – all parameters are locked down by *neural* experiments. The plot shows experimental data and perturbation module for muscimol suppression. **b)** Topography in primate and model IT. Plots show the tissue locations and image preferences of neural sites (red = face preference). Adapted from Lee et al. (2020). **c)** Model predictions align to experimentally observed behaviors in primate perturbation experiments. Top row shows experimental results across different primate studies; bottom row shows model predictions. A checkmark ✓ indicates that the model predictions qualitatively match experimental results.

et al., 2020; Margalit et al., 2023). As a baseline, we also test a model trained only with a classification loss.

Benchmarking models on experimental observations. To test model predictions, we adopt primate experimental observations into benchmarks. Briefly, these benchmarks test behavioral changes in: gender classification from muscimol and optogenetic suppression (A. Afraz et al., 2015), object categorization from muscimol (Rajalingham & DiCarlo, 2019), face classification from micro-stimulation (S. R. Afraz et al., 2006), and face identification from micro-stimulation (Moeller et al., 2017). Each benchmark tests whether a model reproduces the central qualitative claim of the study.

Topographic ANNs with perturbation modules predict the effects of direct neural perturbations on behavior. We now apply the perturbation modules to the topographic model so that the overall model (Figure 1): processes visual input through a hierarchy of feature transformations, perturbs neural activity in model IT according to a particular direct neural intervention based on the perturbation modules and its spatial assignment of neurons, and finally uses the resulting IT activity to make a behavioral response in a given task.

We find that the topographic model with spatial co-training qualitatively predicts all experimental observations (5/5 ✓, Figure 2), in different task settings (ranging from face to object identification and categorization) as well as different perturbation methods (micro-stimulation, optogenetic, and muscimol suppression). Using a random spatial layout for model IT neurons on the other hand fails to predict the perturbation benchmarks (1/5 ✓).

Conclusion

This work enables artificial neural network models of primate visual processing to engage with causal perturbation experiments. We combine a topographic neural network with perturbation modules — and, without any fitting to behavioral perturbation data, the resulting model is capable of predicting all qualitative effects in the contra-lateral hemifield across a battery of primate intervention experiments.

Acknowledgments

We thank Tiago Marques for helping with the KS-metric, Binxu Wang and Carlos Ponce for helping with neurons' maximum firing rates, Abdulkadir Gokce for helping with figures, and Ben Lonnqvist for helping with the text. This work was supported by a Takeda Fellowship (M.S.), a Massachusetts Institute of Technology Shoemaker Fellowship (M.S.), a Friends of the McGovern Fellowship (M.S.), the SRC Semiconductor Research Corporation (M.S., J.J.D.), the Office of Naval Research Grant MURI-114407, and the Simons Foundation Grant SCGB-542965 (J.J.D.).

References

Afraz, A., Boyden, E. S., & DiCarlo, J. J. (2015, may). Optogenetic and pharmacological suppression of spatial clusters of face neurons reveal their causal role in face gender discrimination. *Proceedings of the National Academy of Sciences (PNAS)*, *112*(21), 6730–6735. Retrieved from https://www.pnas.org/content/112/21/6730.short doi: 10.1073/pnas.1423328112

Afraz, S. R., Kiani, R., & Esteky, H. (2006, aug). Microstimulation of inferotemporal cortex influences face categorization. *Nature*, *442*(7103), 692–695. Retrieved from http://www.nature.com/articles/nature04982 doi: 10.1038/nature04982

Lee, H., Margalit, E., Jozwik, K. M., Cohen, M. A., Kanwisher, N., Yamins, D. L. K., & Dicarlo, J. J. (2020, jul). Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *bioRxiv preprint*, 2020.07.09.185116. Retrieved from https://doi.org/10.1101/2020.07.09.185116 doi: 10.1101/2020.07.09.185116

Margalit, E., Lee, H., Finzi, D., DiCarlo, J. J., Grill-Spector, K., & Yamins, D. L. K. (2023, May). A Unifying Principle for the Functional Organization of Visual Cortex [Preprint]. *arXiv*. doi: 10.1101/2023.05.18.541361

Moeller, S., Crapse, T., Chang, L., & Tsao, D. Y. (2017, mar). The effect of face patch microstimulation on perception of faces and objects. *Nature Neuroscience*, *20*(5), 743–752. Retrieved from http://www.nature.com/doifinder/10.1038/nn.4527 doi: 10.1038/nn.4527

Rajalingham, R., & DiCarlo, J. J. (2019, apr). Reversible Inactivation of Different Millimeter-Scale Regions of Primate IT Results in Different Patterns of Core Object Recognition Deficits. *Neuron*, *102*(2), 493–505. Retrieved from https://www.sciencedirect.com/science/article/pii/S0896627319301102 doi: 10.1016/j.neuron.2019.02.001

Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*. Retrieved from https://doi.org/10.1016/j.neuron.2020.07.040 doi: 10.1016/j.neuron.2020.07.040

Yamins, D. L. K., & DiCarlo, J. J. (2016, feb). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. Retrieved from http://www.nature.com/doifinder/10.1038/nn.4244 doi: 10.1038/nn.4244