Do better object recognition models improve the generalization gap in neural predictivity?

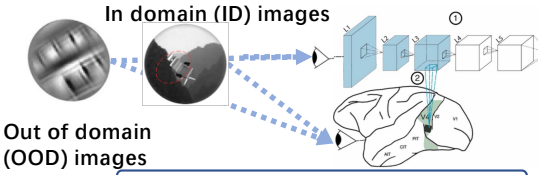
Yifei Ren[1], Pouya Bashivan[2,3]
1. School of Computer Science, McGill University, Canada. 2. Department of Physiology, McGill University, Canada. 3. Quebec AI Institute (MILA), Canada.

Introduction

Motivation: Unit activations in many deep neural networks (DNNs) can be used to accurately predict the neuronal responses to natural stimuli along the ventral visual cortex. Nevertheless, recent work has revealed a gap in generalization ability of these models in predicting neuronal responses to out-of-distribution (OOD) samples .
Problem: Here, we investigated how the recent progress in improving DNNs' object recognition generalization have impacted the generalization gap in neural predictivity.

Methods

Neural predictive measures



ID neural predictions

(1) Train models on ImageNet and fixed model weights
(2) Show ID images to model and brains, then record model activation and brain response in V4 area.
(3) Fit ridge regressions with cross validation to predict neural responses from model activation. ($W^T X = Y$)
(4) ID neural prediction is calculated as the correlation between actual neural responses and predictions of the regression model.
(5) Select the layer with highest median ID prediction accuracy as our best layer in each network.

OOD neural predictions

To compute the OOD neural prediction accuracy, we used the fitted linear mapping from the best layer and computed the neural predictions in response to synthetic (OOD) images.

Neural prediction generalization gap

Neural prediction generalization gap = ID neural predictions – OOD neural predictions

Performance measures

ID performance measures

A_{id}= object recognition accuracy on ImageNet

OOD performance measures

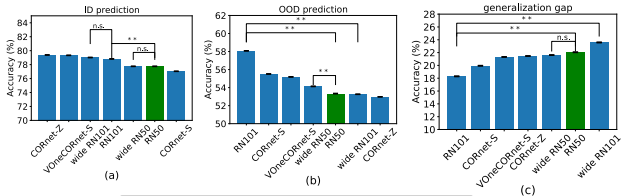
A_{ood} = object recognition accuracy on ImageNet-Adversarial and ImageNet-Rendition
OOD object recognition generalization gap = A_{id} - A_{ood}

Robustness performance measures

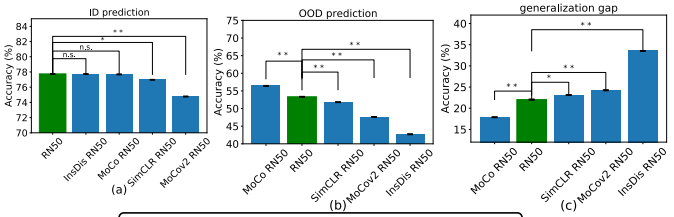
A_{adv} = object recognition accuracy on adversarially perturbed ImageNet images using PGD-L2 and PGD-Linf attacks
Adversarial robustness gap = A_{id} - A_{adv}

Results

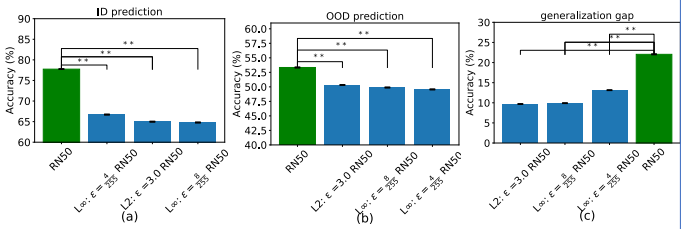
Effect of Network Architecture

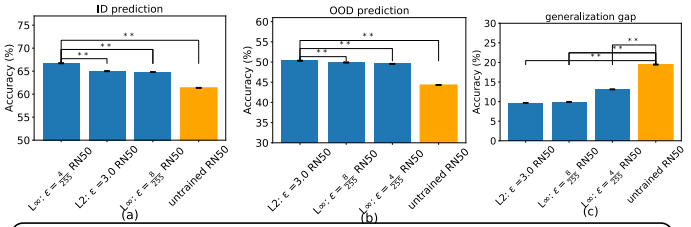


Effect of Learning Algorithm

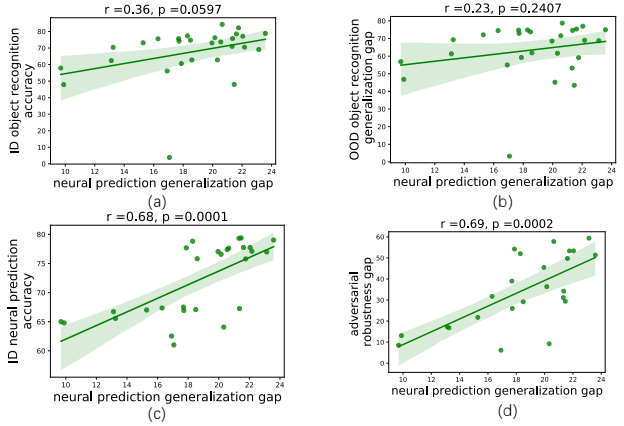


Effect of Adversarial Robustness





Which Model Factors Best Predict the Neural Prediction Generalization Gap?



Conclusions

- *Increase depth and width do not consistently improve ID neural prediction accuracy and generalization gap.*
- *Momentum Contrast unsupervised learning can significantly improve the neural prediction generalization*
- *Adversarial robust models achieve smaller generalization gaps, however the gap is partly due to their universally reduced predictivity.*
- ***Take home message**: Unsupervised and robust DNNs may lead to more general models of neuronal responses in the visual cortex.*