Evaluating Object Recognition Behavioral Consistency on Out-of-Distribution Stimuli

Nga Yu Lo[1], Tiago Marques[2,3,4], James DiCarlo[2,3,4]
[1] Macaulay Honor College at Hunter College, CUNY
[2] Center for Brain, Minds, and Machine, MIT [3] Department of Brain and Cognitive Sciences, MIT [4] McGovern Institute for Brain Research, MIT

ABSTRACT

State-of-the-art artificial neural networks (ANN) trained on ImageNet are known for their top performance on object recognition tasks. They are the best model of the primate ventral stream with moderate success at explaining neural activities as well as visual behavior. To an ANN trained on object recognition, an out-of-distribution (o.o.d.) stimulus is an image with features that differs drastically from the train dataset and usually reduces a model's object recognition performance. Geirhos et al (2020) shows that models are little above chance at predicting human behavior on o.o.d. stimuli. With a dataset consisting of 16 categories and 3 o.o.d. domains, they measure agreement between model and human responses at a visual classification task, using a word association to extract behavioral choices from Imagenet trained models. We find, however, that using an image-level discriminability metric (Rajalingham, Issa et al, 2018) and training a logistic regression model, ANNs have a higher behavior consistency than reported in Geirhos et al. With different ANN models varying in human behavior consistency, these results imply the need to integrate multiple behavioral benchmarks in a unified manner to enable comparisons of models of the human ventral stream.