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

Nga Yu Lo[1], Tiago Marques[2,3,4], James DiCarlo[2,3,4]

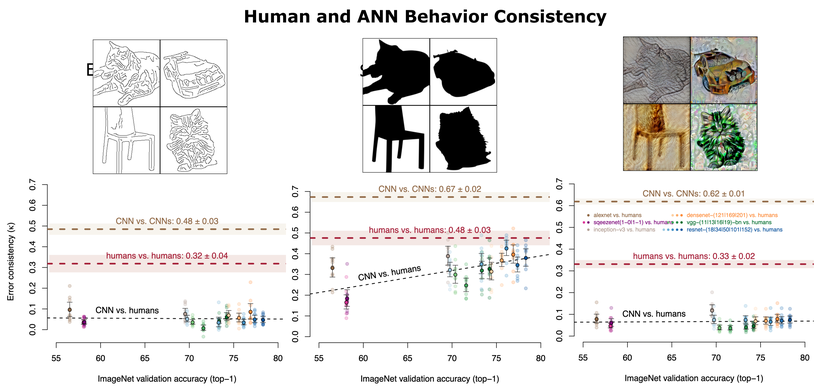
[1]Macaulay Honors College at Hunter College, CUNY [2]Center for Brain, Minds, and Machine, MIT [3]Department of Brain and Cognitive Sciences, MIT [4]McGovern Institute for Brain Research, MIT

CENTER FOR Brains Minds+ Machines

ANNs are the leading models of the primate ventral stream and object recognition behavior

Yamin and DiCarlo, 2016

ANN are shown to be at chance at predicting human choices for out-of-distribution stimuli



Geirhos et al, 2020

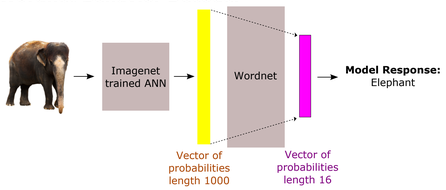
Are the conclusions from Geirhos et al due to:

- the choice of the behavior decoder (wordnet)?
- an artifact of the behavioral metric (error consistency)?
- a true lack of generalization to o.o.d. stimuli?

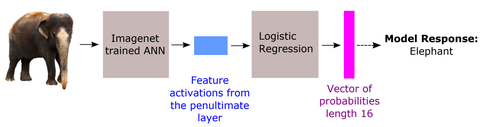
Can we improve the decoder to achieve better behavior consistency on o.o.d. stimuli by including some o.o.d. examples on the training data?

Methodology

Wordnet Behavior Decoder



Logistic Regression Behavior Decoder



Error Consistency Metric: Cohen's K

- Measurement of agreement between two subjects

$$\kappa_{i,j} = \frac{c_{obs_{i,j}} - c_{exp_{i,j}}}{1 - c_{exp_{i,j}}}$$

$$c_{exp_{i,j}} = (p_i)(p_j) + (1 - p_i)(1 - p_j)$$

Geirhos et al, 2020

Image Level Metric (I1)

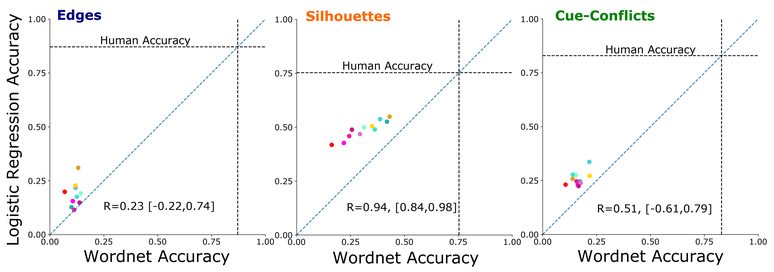
- Pearson's correlation between model and human discriminability at the image level

$$d'(i) = Z(HitRate(i)) - Z(FalseAlarmRate(i))$$

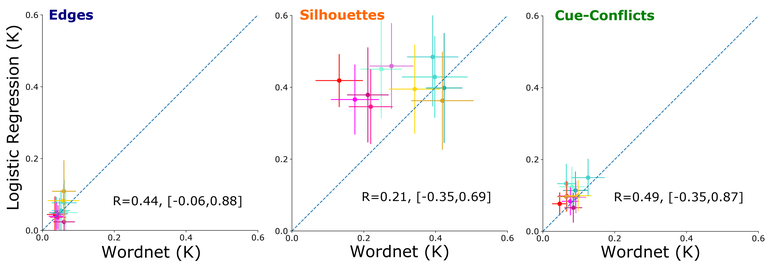
Rajalingham, Issa, et al, 2018

ANNs are not at chance at predicting human behavioral choices for o.o.d. stimuli

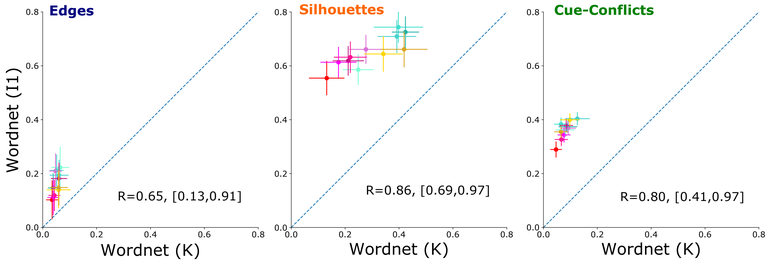
ANNs underperform humans for o.o.d. stimuli



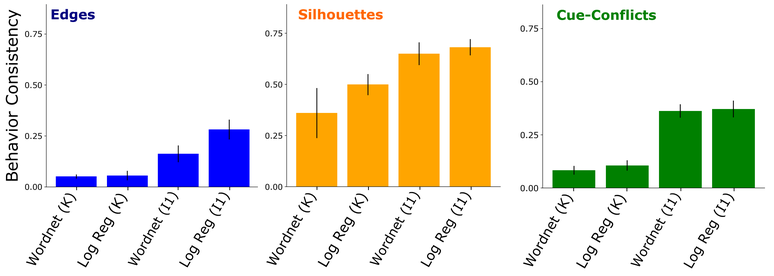
Training a logistic regression behavior decoder slightly improves behavior consistency



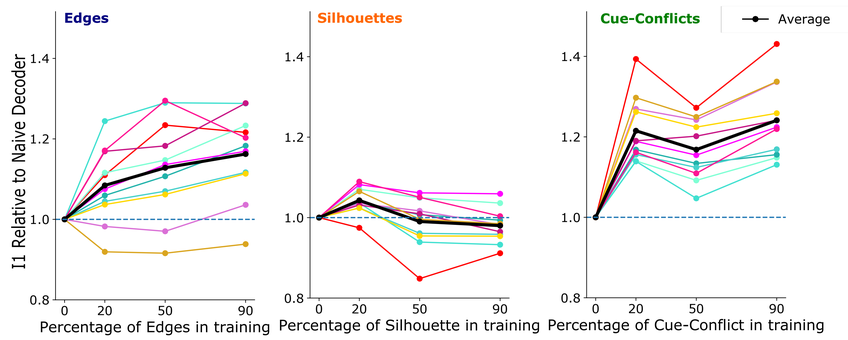
I1 and K are correlated over different models, and I1 is consistenly higher than K



While models show weak behavior consistency for o.o.d. stimuli, they are not at chance level

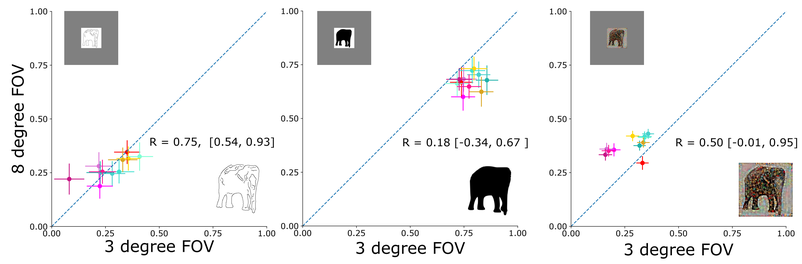


Training logistic regression decoder with o.o.d. images improves scores for edge and cue-conflict stimuli with great variability over models



Simulating a primary visual cortex at the front of an ANN improves behavioral consistency for edge stimuli

Field of view has little effects on models' I1 behavior consistency



Summary

By training a logistic regression decoder on original images and using the I1 metric, we find that ANNs have a higher behavior consistency with o.o.d. stimuli than reported in Geirhos et al.

- While a trained decoder offers slight improvement compared to Wordnet, the benefits vary across base models.
- While I1 is consistently higher than K, the two metrics are highly correlated over different models, suggesting they provide similar information for comparing models' ranking.
- Training the decoder with o.o.d. images improves behavior consistency for some stimuli but the gains vary across base models.
- Field of view has little effects on behavior consistency at image level.

References:

Geirhos, R et al. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. In *Proceedings of 34th Conference of Neural Information Processing Systems* (NeurIPS 2020).

Rajalingham, R., Issa, E.B., et al. Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. Journal of Neuroscience 15 August 2018, 38 (33) 7255-7269.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience, 19(3), 356-365.

Acknowledgements:

Work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216