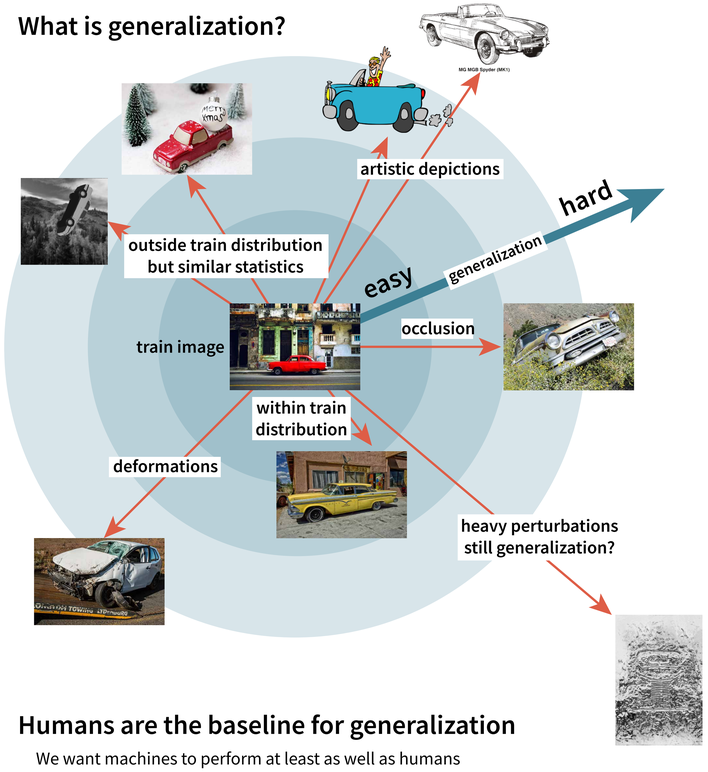
Can Deep Neural Networks Rival Human Ability to Generalize in Core Object Recognition?

Jonas Kubilius*[,1,2], Kohitij Kar*[,1,3], Kailyn Schmidt[1], James J. DiCarlo[1,3]

[1] McGovern Institute for Brain Research, MIT, Cambridge, Massachusetts; [2] Department of Brain and Cognitive Sciences, MIT, Cambridge, Massachusetts; [3] Brain and Cognition, KU Leuven, Leuven, Belgium; * Equal contribution

What is generalization?



Humans are the baseline for generalization

We want machines to perform at least as well as humans

Fair comparison requires comparable training history between humans and machines
- Prior domain experience may not be needed (Hochberg & Brooks, 1962)

Humans – good, machines – poor

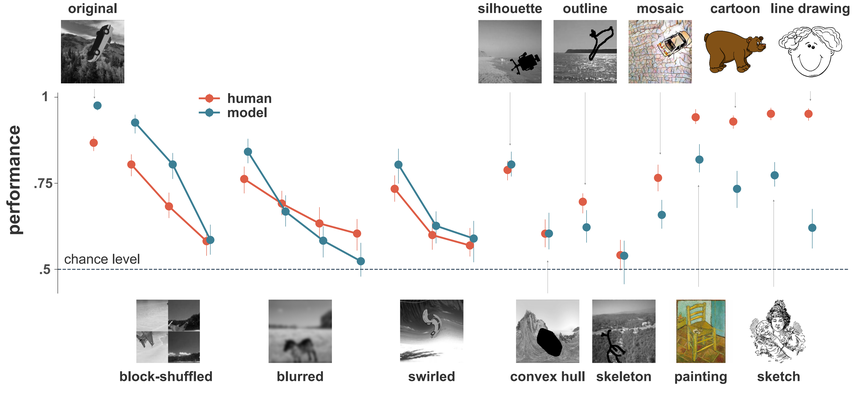
"Humans also excel at generalizing or transferring generalized knowledge gained in one context to novel, previously unseen domains" (Hassabis et al., 2017)

Deep nets show domain specificity (Kornblith, Shlens, & Le, 2018)

New dataset

- 10 categories
- Six styles: HvM, COCO, paintings, sketches, cartoons, drawings
- Perturbed versions of HvM and COCO: block-shuffled, blurred, swirled, silhouettes, convex hull, outlines, skeletons, mosaic

Comparing humans against machines



Human data

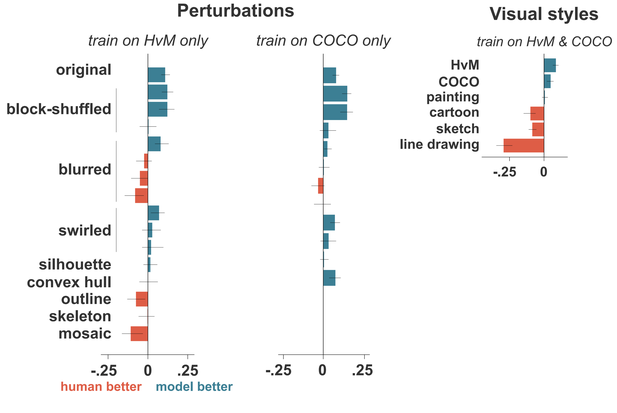
- Amazon Mechanical Turk
- Task: 100 ms image presentation followed by two choices
- 10 responses per image (from 10 observers)

Model testing

- ResNet-152 (PNASNet – similar results), ImageNet-trained
- Behavioral readout training:
  - Logistic regression
  - Perturbations: train on 176 images / category
  - Style: 33 images / category

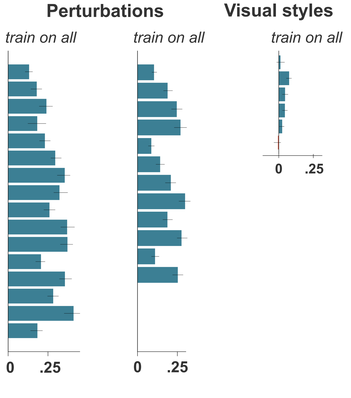
Generalization out of the box?

Training decoder on naturalistic images only

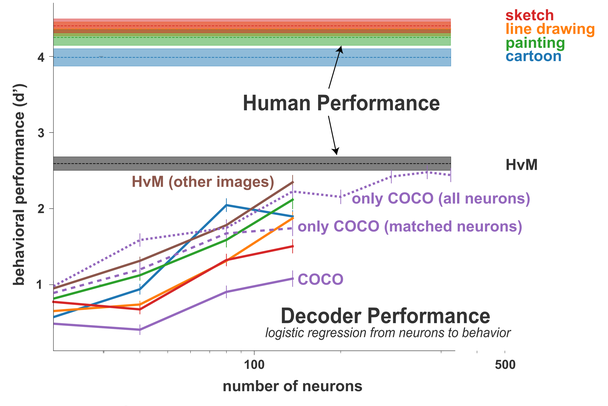


Can we fix machines easily?

Training decoder on everything (encoder frozen!)



Decoding from primate IT (*preliminary*)



Open questions

- Why are artistic stimuli so easy for humans but hard for neurons?
- How well would monkeys do behaviorally?
- How well do DNNs predict monkey IT responses to these stimuli?
- Best way to train decoder: train on all or on each category separately (but with less data)?

Conclusions

- Deep networks are well matched to humans overall
- Representations are disentangled for linear readouts
- Only decoder needs to be improved; DNN itself seems sufficient

Limitations

- Few stimuli categories
- Few exemplars per category
- Only a single task (match to sample)
- More stringent comparison metrics needed

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 705498 (J.K.), US National Eye Institute grants R01-EY014970 (J.J.D.), Office of Naval Research MURI-114407 (J.J.D), and grants from the Simons Foundation (SCGB [325500, 542965], J.J.D).