Disentangling "what" and "where" visual information in neural network vision models

Elijah Christensen[1,2], *Joel Zylberberg*[1]

[1] *Department of Physiology and Biophysics, University of Colorado School of Medicine, Aurora, CO*
[2] *Medical Scientist Training Program*

University of Colorado Anschutz Medical Campus

Abstract

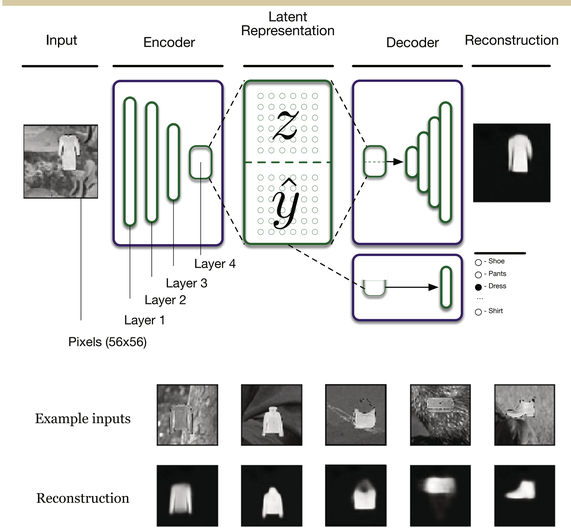
Context: Our visual system constantly reformats raw sensory data into scene representations encoded by populations of neurons. This population is separated anatomically into the dorsal and ventral visual streams canonically representing "what" and "where" information respectfully. Despite how intertwined these two streams are in early visual processing, many computer vision models only attempt to replicate one or the other.

Objective: We constructed a model that attempts to extract object category as well as other scene properties like position by forcing the model to reconstruct the scene. Inspired by anatomical structure our model must use a latent representation comprised of two subpopulations (y_hat and z).

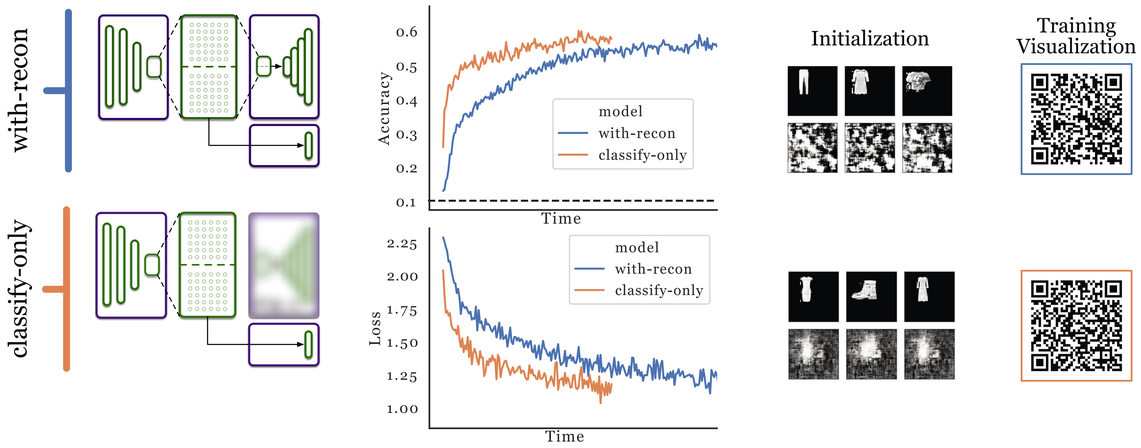
Results: We successfully trained all-to-all ANN models to report both object class and reconstruct the object with its presentation variations. Model's required to reconstruct the input image develop representations *more similar* to primate visual cortex than models which *only classify*.

Data Augmentation

Model

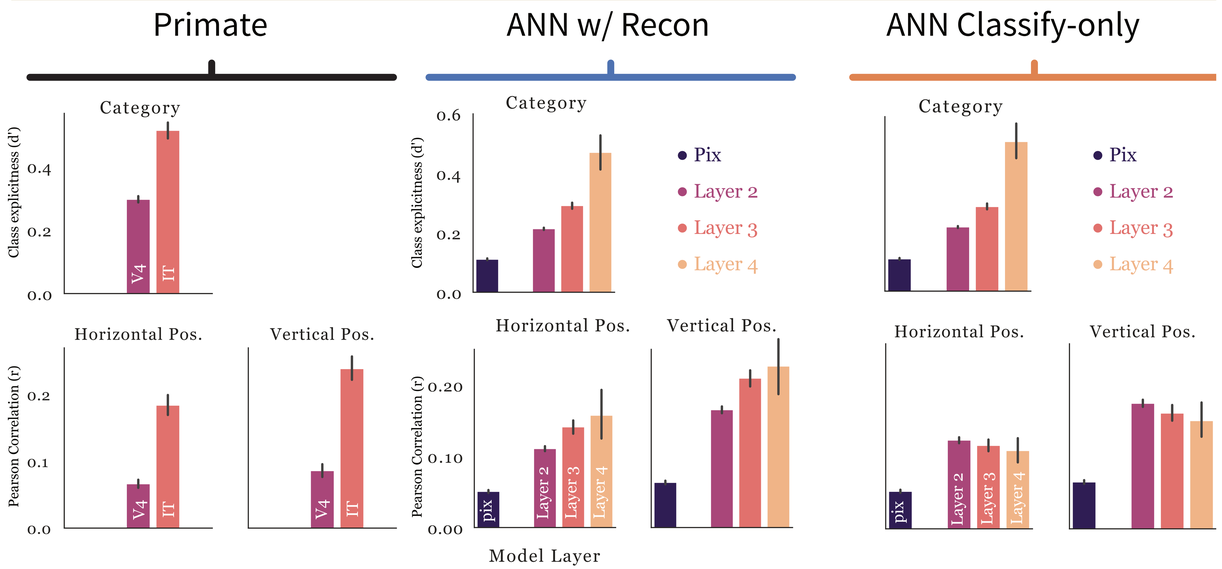


Reconstruction variants achieve similar performance



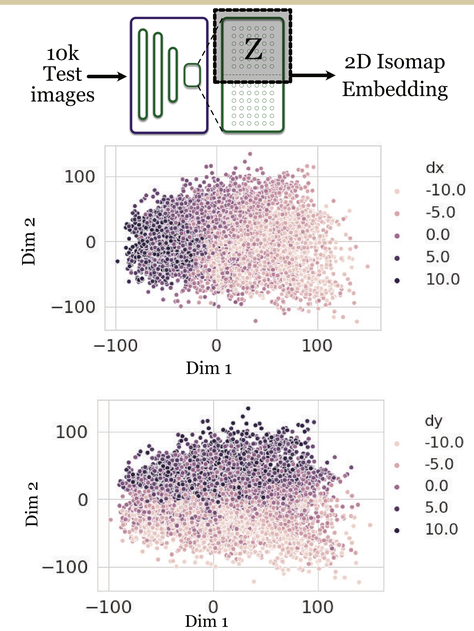
Requiring the model to render the scene's object (with-recon model) doesn't limit its classification performance.

Representation Brain Similarity



Like primates, the recon-and-classify ANN model represents non-categorical properties more explicitly but classify-only models do not.

Position Representation



The Z subspace learns a position "map" without explicit constraints.

Conclusions

- Models that must remember the original scene develop representations more similar to primate visual cortex than models which only classify scene's object.
- Incorporating generative capability of the visual system produces more "brain-like" models of vision.

References

Hong H, Yamins DLK, Majaj NJ, DiCarlo JJ (2016) Explicit information for category-orthogonal object properties increases along the ventral stream. Nat Neurosci 19:613–622.

Acknowledgements

Primate recording data provided by DiCarlo Lab.

This work was funded by Department of Defense (DoD) through the National Defense Science and Engineering Graduate (NDSEG) Fellowship Program. This work was also supported through the VISTA and CFREF grants through York University.