Does Leveraging the Human Ventral Visual Stream Improve Neural Network Robustness?

Zhenan Shao, Linjian Ma, Bo Li, Diane M Beck

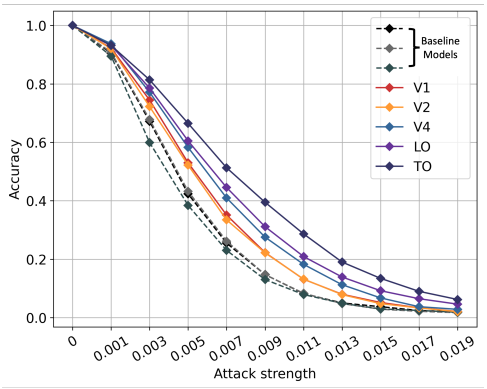
Introduction Humans excel at visual processing, effortlessly recognizing objects despite changes in translation[1], scaling[2], and viewpoints[3]. In contrast, neural networks, despite being the only artificial system with human-level performance in visual tasks, show surprising vulnerability to image perturbations that remain imperceptible or innocuous to humans[4]. This disparity raises the question: What underlies the robustness of the human visual system? A prevalent model attributes such perceptual invariances to the ventral visual stream of the brain, which takes on a potentially crucial role in transforming object representations into increasingly stable and smooth forms, thereby achieving invariance[5]. The ventral visual stream comprises of several brain regions forming a hierarchy to process visual inputs. In particular, all identity-preserving changes to objects form continuous representation manifolds. These manifolds are highly entangled upon their entry but become more separable as they progress through successive stages of the stream, leading to natural invariance as all data points on a manifold encompass all possible identity-preserving transformations of the object.

Interestingly, the vulnerability of neural networks to minor image perturbations can be linked to the smoothness of their decision boundaries[6], mirroring the neural mechanisms discussed above. Adversarial attacks that search the most destructive perturbation to images within a given magnitude bound, have been highly successful because they exploit the complex and non-linear decision boundaries of deep neural networks. Small but carefully crafted perturbations can easily push the input data across these boundaries. This challenge has led to proposals for building smoother classifiers as a defense[7]. Object representation manifolds untangling in the neural space, as previously described, also illustrates this idea of smoothness: as representations are progressively untangled, small deviations become less likely to cross a boundary and fall onto another manifold.

Such conceptual correspondence leads to our main question: Can neural networks learn to represent objects in a similarly smooth manner as our visual system? If so, this might enable neural networks to achieve greater robustness towards adversarial attacks. Importantly, if neural networks learn representations from different stages, i.e., activities in each brain regions, along the ventral visual system, we should observe incremental improvements as object representation manifolds are gradually untangled.

Methods To investigate our hypothesis, we adopted a neural regularization method similar to those in previous literature[8;9;10]. Specifically, we modified the training objective of neural networks for image classification tasks by adding a task to match their penultimate layer of activation to human neural representations recorded by functional Magnetic Resonance Imaging (fMRI) from a recently released dataset[11]. To capture the evolution along the ventral visual stream, we extracted representations from five different stages or brain regions, from the initial V1 area to the later area TO. We trained multiple models to learn representation signatures from each area while performing the standard image classification task. We then evaluated and compared these models on a variety of benchmarks.

Results We found that neural regularization improved models' robustness under adversarial attacks. Notably, we observed incremental improvements in performance as later brain regions were used for regularization (see figure below for this main result). This graded effect persisted even under more powerful adversarial attack methods such as AutoAttack[12]. In addition, we also found our models to become more resilient to an adaptive colorization attack with visible but semantically-preserving changes[13], exhibit a higher shape bias[14], and display more stable GradCAM attention maps[15], although no improvement was observed against natural attacks[16]. Direct quantification of model smoothness[17] also confirmed that neural regularization enhanced the smoothness of the models' decision boundaries, again with graded improvements observed as we ascended the ventral visual hierarchy.



Discussion In conclusion, our results showed that regularizing neural network training with human neural representations can improve their robustness, potentially due to smoother decision boundary learned from human representation space. Importantly, we also observed a hierarchy of robustness improvement, corresponding to the use of neural representations from progressively later areas along the ventral stream. Our contributions are twofold. First, we have shown for the first time that fMRI-recorded human brain activation patterns can be utilized for neural regularization in contrast to past studies that relied on sparse electrophysiological recordings from animals[9;10;8]. The fine-grained differences in robustness along different brain regions also demonstrated the potential for future research to leverage the extensive fMRI datasets collected from humans. Second, the graded improvements on neural networks' robustness also lend support to the emergence of more stable and separable object representations along the ventral visual stream, shedding light on the invariance problem of human visual perception.

References

[1] Irving Biederman and Eric E Cooper. Evidence for Complete Translational and Reflectional Invariance in Visual Object Priming. *Perception*, 20(5):585–593, October 1991.

[2] Irving Biederman and Eric E. Cooper. Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance*, 18(1):121–133, February 1992.

[3] Irving Biederman and Peter C Gerhardstein. Recognizing Depth-Rotated Objects: Evidence and Conditions for Three-Dimensional Viewpoint Invariance. *Journal of Experimental Psychology: Human Perception and Performance,*, 19(6):1162–1182, 1993.

[4] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[5] James J. DiCarlo and David D. Cox. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341, August 2007.

[6] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

[7] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016.

[8] Joel Dapello, Kohitij Kar, Martin Schrimpf, Robert Baldwin Geary, Michael Ferguson, David Daniel Cox, and James J DiCarlo. Aligning model and macaque inferior temporal cortex representations improves model-to-human behavioral alignment and adversarial robustness. In *The Eleventh International Conference on Learning Representations*, 2022.

[9] Zhe Li, Wieland Brendel, Edgar Walker, Erick Cobos, Taliah Muhammad, Jacob Reimer, Matthias Bethge, Fabian Sinz, Zachary Pitkow, and Andreas Tolias. Learning from brains how to regularize machines. *Advances in neural information processing systems*, 32, 2019.

[10] Shahd Safarani, Arne Nix, Konstantin Willeke, Santiago Cadena, Kelli Restivo, George Denfield, Andreas Tolias, and Fabian Sinz. Towards robust vision by multi-task learning on monkey visual cortex. *Advances in Neural Information Processing Systems*, 34:739–751, 2021.

[11] Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022.

[12] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[13] Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and David A Forsyth. Unrestricted adversarial examples via semantic manipulation. *arXiv preprint arXiv:1904.06347*, 2019.

[14] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[15] Tanmay Chakraborty, Utkarsh Trehan, Khawla Mallat, and Jean-Luc Dugelay. Generalizing adversarial explanations with grad-cam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 187–193, 2022.

[16] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

[17] Zhuolin Yang, Linyi Li, Xiaojun Xu, Shiliang Zuo, Qian Chen, Pan Zhou, Benjamin Rubinstein, Ce Zhang, and Bo Li. Trs: Transferability reduced ensemble via promoting gradient diversity and model smoothness. *Advances in Neural Information Processing Systems*, 34:17642–17655, 2021.