Improving Deep Neural Network Robustness through Human Neural Regularization

Linjian Ma and Zhenan Shao
Advisors: Diane M Beck and Bo Li

Humans have a remarkable ability to process visual information, such as recognizing the gist of natural scene images within a brief glance (Fei-Fei et al., 2007). Deep neural networks have emerged as another system that is capable of achieving equivalent or even better performance than humans on certain vision tasks, demonstrated in the ILSVRC. However, a growing body of work (Szegedy et al., 2014) shows an intriguing weakness of DNNs – they are vulnerable to changes in the image that remain imperceptible to humans. These changes can either be small-scale image perturbations or visually significant changes that do not alter the semantic property of inputs, referred to as semantic attacks, can be detrimental to DNNs but still trivial for humans (Bhattad et al., 2020). These findings not only show how DNNs and the human visual system differ, but they also encourage the search for defense mechanisms that make models at least as robust as humans to cope in the dynamic environment in application scenarios like autonomous driving.

One possibility is to directly bias existing models to have a human-like representation by regularizing the training process with human representations. Previous research using single cell recording data from animals have already shown the viability of such approach and the promising results on model robustness (Dapello et al., 2022; Li et al., 2019; Safarani et al., 2021). However, efforts so far are limited in many ways, including the narrow focus on a single location in animal brains, the use of single-cell spike rates rather than distributed activity patterns, regularization of a particular network that lacks generality, and an inadequate selection of evaluation metrics for robustness. Here, we propose to take advantage of the recently released large-scale high-resolution human neural data (NSD) (Allen et al., 2022) to 1) determine whether using human neural representations to bias current CNN models can improve their robustness, 2) understand what factors contribute to the effectiveness of neural regularization, leading to a more explicable approach.

In particular, we will train neural predictors such as the G-net (St-Yves et al., 2022) to learn the neural representation separately for each area of interest (e.g., V1 in the primary visual cortex, which computes for the low-level visual features in the input, or the Lateral Occipital Cortex (LOC) that forms an abstract representation of objects). We then force the penultimate layer of a regular CNN model such as a ResNet18 to resemble the output from the neural predictor as closely as possible while maintaining the classification task performance. Our preliminary results from using neural predictors trained on V1 and LO already showed improved robustness on both bounded-norm attack and colorization semantic attack with larger improvement observed for the hierarchically later area LO.

We plan to investigate the effect of neural predictors trained using functionally different brain regions on various kinds of attacks. For example, the fusiform face area (FFA) that represents the face identity may help improve the robustness towards semantic attacks on face recognition tasks such as (Qiu et al., 2020). In addition, we will apply explainable AI techniques to examine features extracted by neural predictors as well as how neural regularization affects the learned representation of DNNs. By doing so, we aim to uncover computational principles that can enhance the versatility of neural regularization in various applications, as well as improve the robustness of DNNs.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), Article 1. https://doi.org/10.1038/s41593-021-00962-x

Bhattad, A., Chong, M. J., Liang, K., Li, B., & Forsyth, D. A. (2020). *Unrestricted Adversarial Examples via Semantic Manipulation* (arXiv:1904.06347). arXiv. http://arxiv.org/abs/1904.06347

Dapello, J., Kar, K., Schrimpf, M., Geary, R., Ferguson, M., Cox, D. D., & DiCarlo, J. J. (2022). *Aligning Model and Macaque Inferior Temporal Cortex Representations Improves Model-to-Human Behavioral Alignment and Adversarial Robustness* [Preprint]. Neuroscience. https://doi.org/10.1101/2022.07.01.498495

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, *7*(1), 10. https://doi.org/10.1167/7.1.10

Li, Z., Brendel, W., Walker, E., Cobos, E., Muhammad, T., Reimer, J., Bethge, M., Sinz, F., Pitkow, Z., & Tolias, A. (2019). Learning from brains how to regularize machines. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/70117ee3c0b15a2950f1e82a215e812b-Abstract.html

Qiu, H., Xiao, C., Yang, L., Yan, X., Lee, H., & Li, B. (2020). *SemanticAdv: Generating Adversarial Examples via Attribute-conditional Image Editing* (arXiv:1906.07927). arXiv. http://arxiv.org/abs/1906.07927

Safarani, S., Nix, A., Willeke, K., Cadena, S., Restivo, K., Denfield, G., Tolias, A., & Sinz, F. (2021). Towards robust vision by multi-task learning on monkey visual cortex. *Advances in Neural Information Processing Systems*, *34*, 739–751. https://proceedings.neurips.cc/paper/2021/hash/06a9d51e04213572ef0720dd27a84792-Abstract.html

St-Yves, G., Allen, E. J., Wu, Y., Kay, K., & Naselaris, T. (2022). *Brain-optimized neural networks learn non-hierarchical models of representation in human visual cortex* (p. 2022.01.21.477293). bioRxiv. https://doi.org/10.1101/2022.01.21.477293

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). *Intriguing properties of neural networks* (arXiv:1312.6199). arXiv. https://doi.org/10.48550/arXiv.1312.6199