How Can Neuroscience Help Us Build More Robust Deep Neural Networks?

Sayanton Dibbo [*1,2] **Siddharth Mansingh** [*1,3] **Jocelyn Rego** [*1,4] **Garrett Kenyon** [1] **Juston Moore** [1] **Michael Teti** [1]

Abstract

Although Deep Neural Networks (DNNs) are often compared to biological visual systems, they are far less robust to natural and adversarial examples. In contrast, biological visual systems can reliably recognize different objects under a variety of settings. While recent innovations have closed the performance gap between biological and artificial vision systems to some extent, there are still many practical differences between the two. In this Blue Sky Ideas presentation, we will identify some key differences between standard DNNs and biological perceptual systems that may contribute to this lack of robustness. We will then present recent work on biologically-plausible, robust DNNs that are derived from and can be easily implemented on physical systems/neuromorphic hardware.

1. Overview

Although convolutional neural networks (CNNs) are roughly based on biological sensory processing, they lack many computational and architectural motifs that are postulated to contribute to robust perception in biological neural systems. Here, we focus on lateral and top-down connections, which greatly outnumber feed-forward excitatory connections in primary sensory cortical areas (Binzegger et al., 2004) but remain absent in most current DNN architectures. These lateral and top-down connections have been shown to convey context (Stettler et al., 2002; Liang et al., 2017), attention (Noudoost et al., 2010), and expectation (Le Bec et al., 2022) to form sparse representations of sensory stimuli for downstream tasks. In addition, experimental studies implicate these mechanisms in robust visual processing in humans (Elsayed et al., 2018; Daniali & Kim, 2022). In this Blue Sky Ideas presentation, we present recent work on physics-based and neuromorphic DNN architectures with recurrent top-down and lateral connections through the lens of robustness.

2. Lateral Connections

Recent work suggests that transformers exhibit more robustness than CNNs (Bhojanapalli et al., 2021; Aldahdooh et al., 2021; Shao et al., 2021; Zhou et al., 2022). This robustness is thought to be due to self-attention (Bai et al., 2021), which can be thought of as long- and short-range lateral modulation. Although self-attention is one form of lateral modulation, many others have been proposed in the neuroscience literature to model the nonlinear lateral interactions observed in cortical sensory areas, for example divisive normalization (DN) (Carandini & Heeger, 2012; Cornford et al., 2020; Burg et al., 2021) and lateral competition (Olshausen & Field, 1996; Rozell et al., 2008; Boutin et al., 2021; Lian et al., 2019). In fact, single convolutional layers outfitted with DN and/or lateral competition exhibit significantly greater similarity to primary cortical sensory areas than deep CNNs containing many more layers (Olshausen & Field, 1996; 2004; Zhu & Rozell, 2013; Dodds & DeWeese, 2019; Lian et al., 2019; Burg et al., 2021). To illustrate this, we show that CNNs which perform lateral competition in just the first layer (i.e. (Teti et al., 2022; Li et al., 2022)), which we hereafter refer to as LCANets, represent primary visual cortical neurons significantly better than standard CNNs and about the same as adversarially-trained CNNs.

Mounting evidence suggests a strong correlation between representational similarity to the visual cortex and adversarial robustness (Li et al., 2019; Dapello et al., 2020; Safarani et al., 2021; Riedel, 2022). As a result, we then hypothesized that CNNs with lateral competition should be more robust than standard CNNs to adversarial attacks although recent evidence indicates they are not robust when the attack is unknown before test time[1] (Teti et al., 2022). To

[1]Los Alamos National Laboratory, Los Alamos, NM, USA [2]Dartmouth College, Hanover, NH, USA [3]University of Illinois Urbana-Champaign, Champaign, IL, USA [4]Drexel University, Philadelphia, PA, USA. Correspondence to: Michael Teti <mteti@lanl.gov>.

2^{nd} *AdvML Frontiers workshop at* 40^{th} *International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1](Li et al., 2022) developed a heuristic to tune the λ parameter for a specific attack/noise type after training and before testing, which led to greater adversarial accuracy. However, this technique would be very unlikely to work well in most practical settings for a few different reasons.

help understand why LCANets were less robust to adversarial attacks, we compute the perturbation-to-signal ratio (PSR) (Bakiskan et al., 2022) for every layer in a standard CNN, adversarially-trained CNN, and LCANet. We will discuss the results of this analysis, which indicate that the representations in LCANet layers are actually affected *less* than those in both standard CNNs and adversarially-trained CNNs, and that the last layer or two leads to the adversarial susceptibility of LCANets. This reinforces the idea that lateral competition can have powerful effects on many downstream layers in a CNN, but it also suggests that CNNs may require multiple layers with lateral competition to attain adversarial robustness. Based on this, we present current work on the development of CNNs with multiple lateral competition layers, which we refer to as LCANets++, including tests against adversarial attacks.

3. Top-Down/Feedback Connections

In addition to lateral modulation, top-down feedback is a critical component in biological perception that is often overlooked in standard DNNs. This can be seen in anatomical studies, where massive connections from high level to lower level visual areas have been observed (Bullier et al., 1996; Mittal et al., 2020), and neuroimaging studies, which have reported distinct bidirectional activity streams with functional consequences (Nielsen et al., 1999; Dijkstra et al., 2017). Experimental evidence also suggests that these massive feedback connections originating in visual areas as high as IT greatly modulate V1 responses, accounting for contextual (Czigler & Winkler, 2010) and attentional (Noudoost et al., 2010) effects. Top-down feedback is also thought to be critical for reliable inference from weak or noisy stimuli (DiCarlo et al., 2012), and in real-world scenarios with competing stimuli, top-down processes interact with bottom-up and lateral mechanisms to dynamically attend to behaviorally relevant information (Desimone et al., 1995; Kastner & Ungerleider, 2001; McMains & Kastner, 2011).

As a result, we hypothesized that DNNs with recurrent top-down connections should be more robust than standard DNNs. Most current DNN models consist only of feed-forward or bottom-up processes in which the higher layers correspond to abstract features for decision making and low-level representations feed the higher-level representations. However, this information flow could be well supported in a top-down fashion in which high-level representations modulate the low-level representations. Indeed, convolutional sparse coding models endowed with top-down feedback have exhibited many of the nonlinear behaviors observed in biological perceptual systems (Paiton et al., 2015; Kim et al., 2018; Lian et al., 2019; Kim et al., 2020; Boutin et al., 2021).

Motivated by this, we focus our discussion on the energy-based models (Paiton et al., 2015; Scellier & Bengio, 2017; Kim et al., 2018; Laborieux et al., 2021), in which each layer sends information to the previous layer via recurrent feedback connections. Energy-based models can be trained with a framework called Equilibrium Propagation (EP) (Scellier & Bengio, 2017). In contrast to backpropagation, which is difficult to perform on neuromorphic hardware and is not biologically-plausible, EP employs a local learning rule to approximate backpropagation through time. We will also discuss even more recent work, in which it was shown that an EP-like updates could be performed with spike timing dependent plasticity (STDP) (Bengio et al., 2015; 2017). Therefore, energy-based models can be trained with EP directly on neuromorphic hardware, drawing even closer connections to biological systems while requiring orders of magnitude less energy and time compared to standard GPU-based DNNs.

Since energy models contain connections from a given layer of neurons to the previous layer of neurons, expectation in the form of feedback from the higher layers of cognition changed the lower layers to conform to its belief. This is functionally impossible in a feed-forward architecture, yet it is postulated that this is how feedback in the brain works (Walsh et al., 2020). Due to the symmetric feedback connections, these energy-based models are also governed by global attractors, which means it may be difficult for an adversarial attack with a limited attack budget to escape the attractor and cause a mis-classification, but there is currently no evidence for or against this since these models have yet to be adversarially attacked. In this part of the talk, we will present current and ongoing work on the adversarial robustness of energy-based models.

4. Recap

In this Blue Sky Ideas presentation, we examine possible avenues toward the development of robust DNNs by identifying recent biologically-inspired models. Specifically, we discuss CNNs with lateral connections within layers and energy-based models, which are based on top-down feedback. Although both of these mechanisms are found throughout biological sensory areas and help form robust sensory representations, they are hardly found in standard DNNs. We hope this presentation spurs conversations and ideas for future work on biologically-inspired robust machine learning models.

References

Aldahdooh, A., Hamidouche, W., and Deforges, O. Reveal of vision transformers robustness against adversarial attacks. *arXiv preprint arXiv:2106.03734*, 2021.

Bai, Y., Mei, J., Yuille, A. L., and Xie, C. Are transformers more robust than cnns? *Advances in neural information processing systems*, 34:26831–26843, 2021.

Bakiskan, C., Cekic, M., and Madhow, U. Early layers are more important for adversarial robustness. In *ICLR 2022 Workshop on New Frontiers in Adversarial Machine Learning*, 2022.

Bengio, Y., Mesnard, T., Fischer, A., Zhang, S., and Wu, Y. Stdp as presynaptic activity times rate of change of postsynaptic activity. *arXiv preprint arXiv:1509.05936*, 2015.

Bengio, Y., Mesnard, T., Fischer, A., Zhang, S., and Wu, Y. Stdp-compatible approximation of backpropagation in an energy-based model. *Neural computation*, 29(3): 555–577, 2017.

Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., and Veit, A. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10231–10241, 2021.

Binzegger, T., Douglas, R. J., and Martin, K. A. A quantitative map of the circuit of cat primary visual cortex. *Journal of Neuroscience*, 24(39):8441–8453, 2004.

Boutin, V., Franciosini, A., Chavane, F., Ruffier, F., and Perrinet, L. Sparse deep predictive coding captures contour integration capabilities of the early visual system. *PLoS computational biology*, 17(1):e1008629, 2021.

Bullier, J., Hupé, J., James, A., and Girard, P. Functional interactions between areas v1 and v2 in the monkey. *Journal of Physiology-Paris*, 90(3-4):217–220, 1996.

Burg, M. F., Cadena, S. A., Denfield, G. H., Walker, E. Y., Tolias, A. S., Bethge, M., and Ecker, A. S. Learning divisive normalization in primary visual cortex. *PLOS Computational Biology*, 17(6):e1009028, 2021.

Carandini, M. and Heeger, D. J. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012.

Cornford, J., Kalajdzievski, D., Leite, M., Lamarquette, A., Kullmann, D. M., and Richards, B. A. Learning to live with dale's principle: Anns with separate excitatory and inhibitory units. In *International Conference on Learning Representations*, 2020.

Czigler, I. and Winkler, I. *Unconscious Memory Representations in Perception: Processes and mechanisms in the brain*, volume 78. John Benjamins Publishing, 2010.

Daniali, M. and Kim, E. Perception over time: Temporal dynamics for robust image understanding. *arXiv preprint arXiv:2203.06254*, 2022.

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D., and DiCarlo, J. J. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *Advances in Neural Information Processing Systems*, 33:13073–13087, 2020.

Desimone, R., Duncan, J., et al. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. How does the brain solve visual object recognition? *Neuron*, 73(3): 415–434, 2012.

Dijkstra, N., Zeidman, P., Ondobaka, S., Gerven, M., and Friston, K. Distinct top-down and bottom-up brain connectivity during visual perception and imagery. *Scientific Reports*, 7, 12 2017. doi: 10.1038/s41598-017-05888-8.

Dodds, E. M. and DeWeese, M. R. On the sparse structure of natural sounds and natural images: Similarities, differences, and implications for neural coding. *Frontiers in computational neuroscience*, 13:39, 2019.

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., and Sohl-Dickstein, J. Adversarial examples that fool both computer vision and time-limited humans. *Advances in neural information processing systems*, 31, 2018.

Kastner, S. and Ungerleider, L. G. The neural basis of biased competition in human visual cortex. *Neuropsychologia*, 39(12):1263–1276, 2001.

Kim, E., Hannan, D., and Kenyon, G. Deep sparse coding for invariant multimodal halle berry neurons. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1111–1120, 2018.

Kim, E., Rego, J., Watkins, Y., and Kenyon, G. T. Modeling biological immunity to adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4666–4675, 2020.

Laborieux, A., Ernoult, M., Scellier, B., Bengio, Y., Grollier, J., and Querlioz, D. Scaling equilibrium propagation to deep convnets by drastically reducing its gradient estimator bias. *Frontiers in neuroscience*, 15:633674, 2021.

Le Bec, B., Troncoso, X. G., Desbois, C., Passarelli, Y., Baudot, P., Monier, C., Pananceau, M., and Frégnac, Y. Horizontal connectivity in v1: Prediction of coherence in contour and motion integration. *Plos one*, 17(7): e0268351, 2022.

Li, M., Zhai, P., Tong, S., Gao, X., Huang, S.-L., Zhu, Z., You, C., Ma, Y., et al. Revisiting sparse convolutional model for visual recognition. *Advances in Neural Information Processing Systems*, 35:10492–10504, 2022.

Li, Z., Brendel, W., Walker, E., Cobos, E., Muhammad, T., Reimer, J., Bethge, M., Sinz, F., Pitkow, Z., and Tolias, A. Learning from brains how to regularize machines. *Advances in neural information processing systems*, 32, 2019.

Lian, Y., Grayden, D. B., Kameneva, T., Meffin, H., and Burkitt, A. N. Toward a biologically plausible model of lgn-v1 pathways based on efficient coding. *Frontiers in neural circuits*, 13:13, 2019.

Liang, H., Gong, X., Chen, M., Yan, Y., Li, W., and Gilbert, C. D. Interactions between feedback and lateral connections in the primary visual cortex. *Proceedings of the National Academy of Sciences*, 114(32):8637–8642, 2017.

McMains, S. A. and Kastner, S. Interactions of top-down and bottom-up mechanisms in human visual cortex. *The Journal of Neuroscience*, 31:587 – 597, 2011.

Mittal, S., Lamb, A., Goyal, A., Voleti, V., Shanahan, M., Lajoie, G., Mozer, M., and Bengio, Y. Learning to combine top-down and bottom-up signals in recurrent neural networks with attention over modules. *CoRR*, abs/2006.16981, 2020. URL https://arxiv.org/abs/2006.16981.

Nielsen, M., Tanabe, H., Imaruoka, T., Sekiyama, K., Tashiro, T., and Miyauchi, S. Reciprocal connectivity in visual cortex: evidence from fmri. In *IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.99CH37028)*, volume 2, pp. 28–33 vol.2, 1999. doi: 10.1109/ICSMC.1999.825202.

Noudoost, B., Chang, M. H., Steinmetz, N. A., and Moore, T. Top-down control of visual attention. *Current opinion in neurobiology*, 20(2):183–190, 2010.

Olshausen, B. A. and Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Olshausen, B. A. and Field, D. J. Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4):481–487, 2004.

Paiton, D. M., Lundquist, S. Y., Shainin, W., Zhang, X., Schultz, P. F., and Kenyon, G. T. A deconvolutional competitive algorithm for building sparse hierarchical representations. In *BICT*, 2015.

Riedel, A. Bag of tricks for training brain-like deep neural networks. In *Brain-Score Workshop*, 2022.

Rozell, C. J., Johnson, D. H., Baraniuk, R. G., and Olshausen, B. A. Sparse coding via thresholding and local competition in neural circuits. *Neural computation*, 20 (10):2526–2563, 2008.

Safarani, S., Nix, A., Willeke, K., Cadena, S., Restivo, K., Denfield, G., Tolias, A., and Sinz, F. Towards robust vision by multi-task learning on monkey visual cortex. *Advances in Neural Information Processing Systems*, 34: 739–751, 2021.

Scellier, B. and Bengio, Y. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in computational neuroscience*, 11:24, 2017.

Shao, R., Shi, Z., Yi, J., Chen, P.-Y., and Hsieh, C.-J. On the adversarial robustness of visual transformers. *arXiv preprint arXiv:2103.15670*, 1(2), 2021.

Stettler, D. D., Das, A., Bennett, J., and Gilbert, C. D. Lateral connectivity and contextual interactions in macaque primary visual cortex. *Neuron*, 36(4):739–750, 2002.

Teti, M., Kenyon, G., Migliori, B., and Moore, J. Lcanets: Lateral competition improves robustness against corruption and attack. In *International Conference on Machine Learning*, pp. 21232–21252. PMLR, 2022.

Walsh, K. S., McGovern, D. P., Clark, A., and O'Connell, R. G. Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the new York Academy of Sciences*, 1464(1):242–268, 2020.

Zhou, D., Yu, Z., Xie, E., Xiao, C., Anandkumar, A., Feng, J., and Alvarez, J. M. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pp. 27378–27394. PMLR, 2022.

Zhu, M. and Rozell, C. J. Visual nonclassical receptive field effects emerge from sparse coding in a dynamical system. *PLoS computational biology*, 9(8):e1003191, 2013.