Distilled Feedforward Networks Are As Robust as Energy-Based Models trained with Equilibrium Propagation

Siddharth Mansingh, Garrett Kenyon & Michael Teti
Los Alamos National Laboratory
Los Alamos, NM 87545
{smansingh, gkenyon, mteti}@lanl.gov

Abstract

Deep neural networks (DNNs) are not naturally robust to adversarial attacks on their inputs, leading to loss of reliability in a general use case. One of the state-of-the-art defenses against adversarial attacks is adversarial training, which introduces adversarial examples into the training set. While adversarially trained models are more robust to attacks, their accuracy on clean images drops and the additional robustness gained does not generalize well to different types of attacks. Previous studies have proposed energy-based models (EBMs) with a Hopfield-like energy function are inherently robust to adversarial perturbations without any drop in clean accuracy. However, EBMs trained with equilibrium propagation require attaining a fixed point during their dynamical evolution, thus making inference a time consuming process on traditional digital hardware as opposed to neuromorphic hardware which is well-suited for such minimization problems. In this work we report that by training feedforward networks to mimic the fixed points of EBMs, we achieve similar robustness but at drastically shorter inference times. We demonstrate the adversarial robustness conferred by EBM distillation using both white-box and black-box attacks as well as natural corruptions on the CIFAR-10 and CIFAR-100 dataset. We thus posit that EBM distillation could provide an alternative method to adversarial training.

1 Previous work on EBMs

Energy-based models that involve minimization of an energy function are relatively new in the domain of adversarial robustness. One of the pioneering works in this context are dense associative memories, which have been shown to produce more semantically meaningful interpretations of adversarial images on MNIST Krotov & Hopfield (2018), which the authors attribute to the use of highly nonlinear activation functions, although this study was conducted with a fully connected layer architecture. A large body of work exists on the adversarial robustness of sparse LCA networks that involve lateral connections and a sparse objective Paiton et al. (2020); Teti et al. (2022); McAlister et al. (2024) which suggest that cortical networks increase selectivity through lateral inhibition, where the inhibition is proportional to the overlap in their receptive fields. Such lateral interactions are thus expected to give rise to increased robustness in comparison to purely feedforward neural networks that lack these interactions. A recent study explored the robustness of EBMs trained with equilibrium propagation (EP) Laborieux et al. (2021) and found EBMs to be inherently robust without any adversarial training or augmentations to the training set Mansingh et al. (2024). However, an overarching feature of energy-based models is that they require settling into an attractor or a fixed point both during training and inference. Training and inference in such models on standard hardware is time consuming since they require to be evolved for multiple time steps and thus are not well suited for learning large complex datasets.

Physical computing platforms such as neuromorphic hardware Parpart et al. (2023); Fair et al. (2019), memristor crossbars Yi et al. (2022) and self adjusting resistor networks Wycoff et al. (2022) are well suited for such energy minimization problem. However such platforms are not commercially available, and experiments on these physical platforms have also been limited to small datasets. Attempts to accelerate the optimization problem involve asynchronous state updates Scellier et al. (2023), which are reminiscent of the leap-frog method as well as casting the energy minimization problem into a deep equilibrium model (DEQ) setting Goemaere et al. (2024), which involves reaching the steady state in one shot, without the need of explicit state evolution.

DEQs fall into a class of implicit models Bai et al. (2019). which involve finding equilibrium/fixed points for an effectively infinite depth feedforward network, in one shot. DEQs however are not robust to adversarial attacks Gurumurthy et al. (2021). Because of implicit methods involved in reaching the fixed point, tests on robustness of DEQs often involve evaluation of approximate gradients raising concerns about gradient obfuscation in case of white-box attacks Yang et al. (2022). Since DEQs are vulnerable to adversarial attacks, they are often adversarial trained to gain robustness Yang et al. (2023) which alludes to our hypothesis that energy minimization dynamics play a vital role in adversarial robustness. This is consistent with insights from neuroscience which suggest that humans take longer to identify challenging images compared to control images Kar et al. (2019).

In cases where inference is time-consuming and computationally expensive, knowledge distillation Hinton (2015) is a popular technique of compressing knowledge present in large (teacher) models and transferring to small (student) models. This is useful in situations where computational resources are constrained such as edge devices. However knowledge distillation does not guarantee the transfer of adversarial robustness to the student models. To address this, adversarial robust distillation Goldblum et al. (2020) has been proposed, that trains the student model to keep its predictions within an ϵ ball of the teacher's outputs, thus staying close to adversarial training in spirit. In contrast, knowledge distillation with input gradient alignment Shao et al. (2021) achieves distillation by forcing the student to learn both the logits and the gradients arising out of training samples. To address the above mentioned concerns, we propose distillation of energy-based models trained with EP.

2 Robustness of EBMs trained with Equilibrium Propagation

The Hopfield-like energy function used to train the EBM, with input x and weights $w_n, n \in [1, N_{tot}]$ with N_{conv} convolutional layers and $N_{tot} - N_{conv}$ fully connected layers, is given as

$$E(x, \{s^n\}) = \sum_{n<N_{\text{conv}}} s^{n+1} \cdot \mathcal{P}(w_{n+1} \star s^n) + \sum_{n=N_{\text{conv}}}^{N_{\text{tot}-1}} {s^{n+1}}^\top \cdot w_{n+1} \cdot s^n + \frac{1}{2}\|s^2\| \tag{1}$$

where $\mathcal{P}$ represents a pooling function, $s^0 = x, s^{N_{tot}} = y$, the network has N_{conv} convolutional layers and $N_{tot} - N_{conv}$ total layers. The state evolution for the EBM is thus given by

$$\frac{\partial s^n}{\partial t} = -\frac{\partial E}{\partial s^n} = -s^n + \sigma\left(\mathcal{P}(w_n \star s^{n-1}) + \tilde{w}_{n+1} \star \mathcal{P}^{-1}(s^n)\right) \tag{2}$$

where $\tilde{w}$ is the transpose convolution/linear operation. Given the energy function being quadratic with respect to the state of the network, the energy is guaranteed to monotonically decrease over time Scellier & Bengio (2017). In the absence of an activation function, and an identity pooling function, the state evolution can be simplified to the following form

$$S_{t+1} = W S_t \tag{3}$$

where the transition matrix W and state vector S are denoted by

$$W = \begin{bmatrix} \mathbb{I} & 0 & 0 & & \dots & \\ w_1 & 0 & w_2^\top & 0 & & \\ 0 & w_2 & 0 & w_3^\top & & \\ & & \ddots & 0 & \ddots & \\ & & & w_{N-1} & 0 & w_N^\top \\ & & & 0 & w_N & 0 \end{bmatrix}, S = \begin{bmatrix} x \\ s^1 \\ s^2 \\ s^{\cdots} \\ s^{N-1} \\ y \end{bmatrix} \tag{4}$$

The network is said to be in a steady state S_* if $S_* = WS_*$. While diagonalizing a general nonuniform tridiagonal matrix is nontrivial, let λ^i and ν_i be the eigenvalues and eigenvectors of the transition matrix, $i \in [0, N]$. Since any input is guaranteed to converge to a steady state, we know that $\lambda_i \leq 1$ for a transition matrix constructed from the Hopfield-like energy function. Any initial statevector can thus be written in eigenvector basis in the following form

$$S_0 = \sum_{i=0}^{N} C_i \nu_i \tag{5}$$

Applying the transition matrix for T timesteps has the following effect

$$S_T = W^T S_0 = W^T \sum_{i=0}^{N} C_i \nu_i = \sum_{i=0}^{N} C_i \lambda_i^T \nu_i \tag{6}$$

This implies that when $T \to \infty$, all contributions from eigenvectors whose eigenvalues are strictly less than one would vanish i.e., $\lim_{T\to\infty} \lambda_i^T \to 0 \,\forall\, \lambda_i < 1$. The only basis elements that survive would be those eigenvectors whose eigenvalues are identically equal to 1. In other words, for any initial statevector S_0 (as denoted in Eq. 5), the steady state vector would correspond to

$$S_* = \sum_{i:\lambda_i=1} C_i \nu_i \tag{7}$$

In terms of stability and effectively adversarial robustness, any perturbation to S_0 would thus have to be made to the unit-eigenvalue basis components (C_i where $\lambda_i = 1$) in order to non-trivially affect the steady state of the network. The current work attempts to learn the effective transition matrix with the help of a feedforward network such that it approximates $\lim_{T\to\infty} W^T$. Next, we present our results on the adversarial robustness of the distilled feedforward network when subjected to different kinds of adversarial attacks.

3 Experimental Setup

For our experiments, we train a model with equilibrium propagation using symmetric/centered weight updates Laborieux et al. (2021); Scellier et al. (2023) and refer to it as **EP-CNN**, which has four convolutional layers with max pooling and a fully connected layer. To compare the performance of EP with standard models, we train a feedforward model with the same architecture with backpropagation, hence referred to as **BP-CNN**. Furthermore, we also consider adversarially trained feedforward models with various ℓ_2 constraints and 200 iterations of projected gradient descent (PGD, Madry et al. (2017)). These models will be denoted by **Adv-CNN**. Images from the training set were augmented with random cropping and horizontal flipping and these were the only augmentations used during training. In the next subsection, we describe distillation of fixed points of **EP-CNN** into a feedforward model which we denote as **Distilled BP-CNN**.

3.1 Attractor Distillation

We aim to mimic the fixed points of a learned **EP-CNN** but with the help of a feedforward network. This is achieved by starting with the feedforward **BP-CNN** backbone, which has the same number of layers as **EP-CNN**. For a given input in the training set and starting with a null state, the **EP-CNN** is allowed to settle into a fixed point. The ℓ_2 distance between the states of all intermediate layers in the **EP-CNN** and **Distilled BP-CNN** represent the loss which is then further used to compute the backward pass in **Distilled BP-CNN**. The loss function to train **Distilled BP-CNN** is thus defined as

$$\min_{\theta} \mathcal{L}_{\text{Dist}}\left(D_\theta, \text{EP}, x\right) \tag{8}$$

$$\mathcal{L}_{\text{Dist}}\left(D_\theta, \text{EP}, x\right) = \sum_{i=1}^{N-1} \text{MSE}(s_*^i, \tilde{s}^i) + \text{CE}(s_*^N, \tilde{y})$$

where θ are the weights of the **Distilled BP-CNN** network, s_* is the fixed point of **EP-CNN**, $\tilde{s_i}$ are the states of intermediate layers of **Distilled BP-CNN** post activation, for a given input x and $\tilde{y}$ is the output of **Distilled BP-CNN**. We would like to note that it is not necessary for **Distilled BP-CNN** to have the same architecture as **EP-CNN** as one could only be concerned with minimizing the distance between the predictions of the two models and disregard the states of the previous layers. To this end, we also train a ResNet-18 model and a **Distilled BP-CNN**($\ell\ell$) to mimic just the prediction layer in **EP-CNN**.

$$\mathcal{L}\left(D_\theta, \mathrm{EP}, x\right) = \mathrm{CE}(s_*^N, \tilde{y}) \tag{9}$$

where $\tilde{y}$ is the output of the distilled model.

3.2 Adversarial Attacks

To evaluate the robustness of our models, we consider three types of attacks: i) Projected gradient descent (PGD) Madry et al. (2017), a type of white-box attack where the adversary has knowledge of the weights of the model and its predictions. ii) Square Attack Andriushchenko et al. (2020), a type of black-box attack where the adversary only has information of the model predictions and the input. The adversary refines its attacks by making multiple queries to the model. Black-box attacks are meant to serve as a check for gradient obfuscation. iii) AutoAttack Croce & Hein (2020), a state of the art attack comprised of four different types of attacks, two different types of PGD attacksm a Square attack and a DeepFool attack. The same set of hyperparameters were used to conduct attacks on all the models.

4 Results on Adversarial Attacks

Since the **Distilled BP-CNN** has not been adversarially trained, it does not feature a loss in clean accuracy as is the case with **EP-CNN**. Similarly both **Distilled BP-CNN** and **EP-CNN** generalize well to different types of attacks, across the CIFAR10 and CIFAR100 datasets (see Fig. 1 and Fig. 2), when compared to adversarially trained CNNs. This is noteworthy since these models do not involve any special augmentations (such as AutoAugment Cubuk et al. (2019)) that intend to alter the training set in any way, nor do they involve any adversarial training. Since **Distilled BP-CNN**s lack the feedback connections and hence the crucial attractor dynamics, their adversarial robustness is upper-bounded by the robustness of **EP-CNN**. However, models that were taught only to mimic the last layer of **EP-CNN** (Eq. 9) such as **Distilled BP-CNN**($\ell\ell$) and ResNet-18 did not perform as well as **Distilled BP-CNN**. We hypothesize that mimicking the predictions does not imply learning the dynamics of a model and since the dynamics of EP play a crucial role in making the model robust, models that do not learn the full steady state of **EP-CNN** are vulnerable to attacks.

5 Discussion and Conclusion

Deep neural networks remain vulnerable to adversarial perturbations Szegedy et al. (2014); Madry et al. (2017), as well as to natural noise Hendrycks & Dietterich (2019). Adversarial training Madry et al. (2017) remains the state of the art technique for increasing adversarial robustness of DNNs, however, this is often associated with a drop in clean accuracy Schmidt et al. (2018); Zhang et al. (2019). Morever, adversarial robustness is directly proportional to the depth of a network, hence making it unfeasible to deploy large robust models on edge devices. While perception in humans is a dynamic process and benefits from recurrent feedback connections Daniali & Kim (2023), standard hardware is not well-suited for implementing such dynamical systems, hence leading to loss in robustness. Our proposed method of distilling EBMs into feedforward networks by learning the attractor states is a viable method of achieving adversarial robustness in a compute-efficient way.

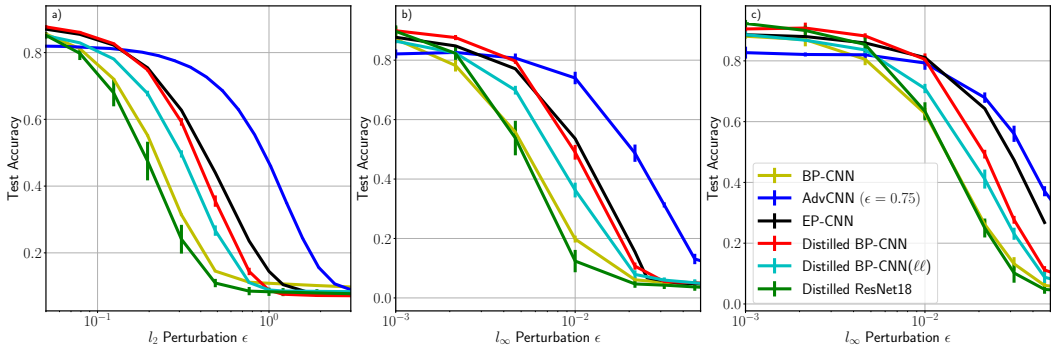

Figure 1: Line graph of accuracy as a function of ϵ perturbation across different types of adversarial attacks: a) ℓ_2 PGD attack, b) ℓ_∞ AutoAttack and c) Square attack on the CIFAR10 dataset. All errorbars represent a 95% CI over 5 different seeds.

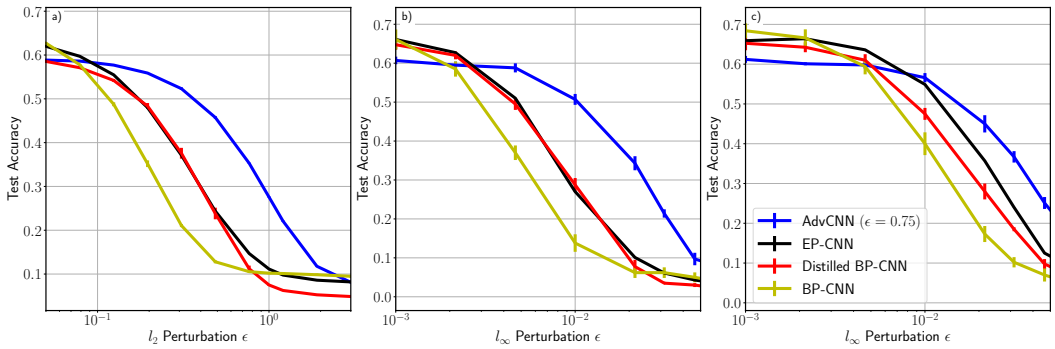

Figure 2: Line graph of accuracy as a function of ϵ perturbation across different types of adversarial attacks: a)ℓ_2 PGD attack, b) ℓ_∞ AutoAttack and c) Square attack on the CIFAR100 dataset. All errorbars represent a 95% CI over 5 different seeds.

Prior attempts address the issue of robust fairness (robustness across different classes) by increasing the weights of those classes whose representation is difficult to learn Yue et al. (2024). Other attempts at improving robustness of distilled networks involve dynamic training with the help of both adversarial teacher and clean teacher models Zhao et al. (2022). A future work could involve investigating the class wise robustness of both the **EP-CNN** and networks distilled from **EP-CNN** to study how the robustness is transferred across different classes. We would like to note that most of the state-of-the-art techniques Ham et al. (2024) that aim at making distilled robust networks are often tested on smaller datasets such as CIFAR-100 and TinyImageNet, an evidence that scalability of such techniques remains an issue. While there is limited work on Hopfield networks in the context of knowledge distillation, the work of Thériault and Tantari aims to provide a theoretical understanding of student-teacher adversarial robustness of Hopfield model in a generative setting Thériault & Tantari (2024) and one could adopt a similar formalism to analytically investigate the robustness of EP.

As noted in earlier works on EP Scellier & Bengio (2017); Scellier et al. (2023), training is relatively sensitive to its hyperparameters when compared to training an equivalent feedforward model with backpropagation. We would like to note that distilling the fixed point's last layer state i.e., S^N into a feedforward model was easier than distilling the fixed point's entire state vector S (see Eq. 4) and we had to adjust the layerwise learning rates in order to achieve similar clean accuracy as the **EP-CNN** model. While inference in **EP-CNN** is time consuming since it requires performing the free phase until the network settles into a steady state, inference in **Distilled BP-CNN** is rather straightforward. While details of **EP-CNN** training were not a focus of this work, future directions could involve using this

insight of distillation to accelerate training of **EP-CNN**. These methods would be different from implicit methods like DEQs to make sure the benefits of adversarial robustness are preserved.

6 Acknowledgements

Research presented in this paper was supported by the National Security Education Center (NSEC) Informational Science and Technology Institute (ISTI) using the Laboratory Directed Research and Development program of Los Alamos National Laboratory project number 20240479CR-IST. LANL is operated by Triad National Security, LLC, for the National Nuclear Security Administration of the U.S. Department of Energy (Contract No. 89233218CNA000001).

References

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, pp. 484–501. Springer, 2020.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Ekin Dogus Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. 2019. URL https://arxiv.org/pdf/1805.09501.pdf.

Maryam Daniali and Edward Kim. Perception over time: Temporal dynamics for robust image understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pp. 5656–5665. IEEE, 2023. doi: 10.1109/CVPRW59228.2023.00599. URL https://doi.org/10.1109/CVPRW59228.2023.00599.

Kaitlin L. Fair, Daniel R. Mendat, Andreas G. Andreou, Christopher J. Rozell, Justin Romberg, and David V. Anderson. Sparse coding using the locally competitive algorithm on the truenorth neurosynaptic system. *Frontiers in Neuroscience*, 13, July 2019. ISSN 1662-453X. doi: 10.3389/fnins.2019.00754. URL http://dx.doi.org/10.3389/fnins.2019.00754.

Cédric Goemaere, Johannes Deleu, and Thomas Demeester. Accelerating hopfield network dynamics: Beyond synchronous updates and forward euler. In *Proceedings of the 1st ECAI Workshop on "Machine Learning Meets Differential Equations: From Theory to Applications"*, volume 255 of *Proceedings of Machine Learning Research*, pp. 1–21. PMLR, 20 Oct 2024. URL https://proceedings.mlr.press/v255/goemaere24a.html.

Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 3996–4003, April 2020. ISSN 2159-5399. doi: 10.1609/aaai.v34i04.5816. URL http://dx.doi.org/10.1609/aaai.v34i04.5816.

Swaminathan Gurumurthy, Shaojie Bai, Zachary Manchester, and J Zico Kolter. Joint inference and input optimization in equilibrium networks. *Advances in Neural Information Processing Systems*, 34:16818–16832, 2021.

Seokil Ham, Jungwuk Park, Dong-Jun Han, and Jaekyun Moon. Neo-kd: knowledge-distillation-based adversarial training for robust multi-exit neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Kohitij Kar, Jonas Kubilius, Kailyn Schmidt, Elias B. Issa, and James J. DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22(6):974–983, April 2019. doi: 10.1038/s41593-019-0392-5. URL https://doi.org/10.1038/s41593-019-0392-5.

Dmitry Krotov and John Hopfield. Dense associative memory is robust to adversarial inputs. *Neural Computation*, 30(12):3151–3167, December 2018. ISSN 1530-888X. doi: 10.1162/neco_a_01143. URL http://dx.doi.org/10.1162/neco_a_01143.

Axel Laborieux, Maxence Ernoult, Benjamin Scellier, Yoshua Bengio, Julie Grollier, and Damien Querlioz. Scaling equilibrium propagation to deep ConvNets by drastically reducing its gradient estimator bias. *Frontiers in Neuroscience*, 15, February 2021. doi: 10.3389/fnins.2021.633674. URL https://doi.org/10.3389/fnins.2021.633674.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2017.

Siddharth Mansingh, Michal Kucer, Garrett Kenyon, Juston Moore, and Michael Teti. Energy-based models trained with equilibrium propagation are inherently robust. In *2024 International Conference on Machine Learning and Applications (ICMLA)*, pp. 727–747. IEEE, December 2024. doi: 10.1109/ICMLA61862.2024.00105. URL http://dx.doi.org/10.1109/ICMLA61862.2024.00105.

Hayden McAlister, Anthony Robins, and Lech Szymanski. Improved robustness and hyperparameter selection in the dense associative memory, 2024.

Dylan M. Paiton, Charles G. Frye, Sheng Y. Lundquist, Joel D. Bowen, Ryan Zarcone, and Bruno A. Olshausen. Selectivity and robustness of sparse coding networks. *Journal of Vision*, 20(12):10, November 2020. ISSN 1534-7362. doi: 10.1167/jov.20.12.10. URL http://dx.doi.org/10.1167/jov.20.12.10.

Gavin Parpart, Sumedh Risbud, Garrett Kenyon, and Yijing Watkins. Implementing and benchmarking the locally competitive algorithm on the loihi 2 neuromorphic processor. In *Proceedings of the 2023 International Conference on Neuromorphic Systems*, ICONS '23, pp. 1–6. ACM, August 2023. doi: 10.1145/3589737.3605973. URL http://dx.doi.org/10.1145/3589737.3605973.

Benjamin Scellier and Yoshua Bengio. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in Computational Neuroscience*, 11, May 2017. doi: 10.3389/fncom.2017.00024. URL https://doi.org/10.3389/fncom.2017.00024.

Benjamin Scellier, Maxence Ernoult, Jack Kendall, and Suhas Kumar. Energy-based learning algorithms for analog computing: a comparative study. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 52705–52731. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a52b0d191b619477cc798d544f4f0e4b-Paper-Conference.pdf.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

Rulin Shao, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. How and when adversarial robustness transfers in knowledge distillation? *arXiv preprint arXiv:2110.12072*, 2021.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL http://arxiv.org/abs/1312.6199.

Michael Teti, Garrett Kenyon, Ben Migliori, and Juston Moore. LCANets: Lateral competition improves robustness against corruption and attack. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21232–21252. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/teti22a.html.

Robin Thériault and Daniele Tantari. Dense hopfield networks in the teacher-student setting. *SciPost Physics*, 17(2), August 2024. ISSN 2542-4653. doi: 10.21468/scipostphys.17.2.040. URL http://dx.doi.org/10.21468/SciPostPhys.17.2.040.

J. F. Wycoff, S. Dillavou, M. Stern, A. J. Liu, and D. J. Durian. Desynchronous learning in a physics-driven learning network. *The Journal of Chemical Physics*, 156(14), April 2022. ISSN 1089-7690. doi: 10.1063/5.0084631. URL http://dx.doi.org/10.1063/5.0084631.

Zonghan Yang, Tianyu Pang, and Yang Liu. A closer look at the adversarial robustness of deep equilibrium models. *Advances in Neural Information Processing Systems*, 35: 10448–10461, 2022.

Zonghan Yang, Peng Li, Tianyu Pang, and Yang Liu. Improving adversarial robustness of deep equilibrium models with explicit regulations along the neural dynamics. 2023.

Su-in Yi, Jack D. Kendall, R. Stanley Williams, and Suhas Kumar. Activity-difference training of deep neural networks using memristor crossbars. *Nature Electronics*, November 2022. ISSN 2520-1131. doi: 10.1038/s41928-022-00869-w. URL http://dx.doi.org/10.1038/s41928-022-00869-w.

Xinli Yue, Mou Ningping, Qian Wang, and Lingchen Zhao. Revisiting adversarial robustness distillation from the perspective of robust fairness. *Advances in Neural Information Processing Systems*, 36, 2024.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Shiji Zhao, Jie Yu, Zhenlong Sun, Bo Zhang, and Xingxing Wei. *Enhanced Accuracy and Robustness via Multi-teacher Adversarial Distillation*, pp. 585–602. Springer Nature Switzerland, 2022. ISBN 9783031197727. doi: 10.1007/978-3-031-19772-7_34. URL http://dx.doi.org/10.1007/978-3-031-19772-7_34.