Efficient Biologically Plausible Adversarial Training

Matilde Tristany Farinha[1,2], **Thomas Ortner**[1], **Giorgia Dellaferrera**[1,2],
Benjamin Grewe[2], **Angeliki Pantazi**[1]
[1]IBM Research Europe / Zurich, [2]Institute of Neuroinformatics, University of Zurich and ETH Zurich
mtristany@ethz.ch, agp@zurich.ibm.com

Abstract

Artificial Neural Networks (ANNs) trained with Backpropagation (BP) show astounding performance and are increasingly often used in performing our daily life tasks. However, ANNs are highly vulnerable to adversarial attacks, which alter inputs with small targeted perturbations that drastically disrupt the models' performance. The most effective method to make ANNs robust against these attacks is adversarial training, in which the training dataset is augmented with exemplary adversarial samples. Unfortunately, this approach has the drawback of increased training complexity since generating adversarial samples is very computationally demanding. In contrast to ANNs, humans are not susceptible to adversarial attacks. Therefore, in this work, we investigate whether biologically-plausible learning algorithms are more robust against adversarial attacks than BP. In particular, we present an extensive comparative analysis of the adversarial robustness of BP and *Present the Error to Perturb the Input To modulate Activity* (PEPITA), a recently proposed biologically-plausible learning algorithm, on various computer vision tasks. We observe that PEPITA has higher intrinsic adversarial robustness and, with adversarial training, has a more favourable natural-vs-adversarial performance trade-off as, for the same natural accuracies, PEPITA's adversarial accuracies decrease in average by 0.26% and BP's by 8.05%.

1 Introduction

State-of-the-art ANNs trained with Backpropagation (BP) [1, 2] are vulnerable to adversarial attacks [3]. Adversarial attacks produce adversarial samples, a concept first described by Szegedy et al. [4], which are input samples with small perturbations that can trick a trained ANN into misclassification. Although this phenomenon was first observed in the context of image classification [4, 5], it has since been observed in several other tasks such as natural language processing [6, 7], audio processing [8, 9], and deep reinforcement learning [10, 11]. Nowadays, making real-world decisions based on the suggestions provided by ANNs has become an integral part of our daily lives [12]. Therefore, these models' vulnerability to adversarial attacks severely threatens the safe deployment of artificial intelligence in everyday-life applications [13]. For example, in real-world autonomous driving, adversarial attacks have been successful in deceiving road sign recognition systems [14]. Researchers have proposed several solutions to address this problem, and adversarial training emerged as the state-of-the-art approach [15]. In adversarial training, the original dataset, consisting of pairs of input samples with their respective ground-truth labels, is augmented with adversarial data, where the original ground-truth labels are paired with adversarial samples. This additional training data allows the model to learn to classify correctly adversarial samples as well [5, 3]. Although adversarial training increases the networks' robustness to adversarial attacks, generating numerous training adversarial samples is computationally costly. To reduce this additional computational burden, researchers have developed new methods for generating adversarial samples more efficiently [16, 17, 18, 19]. For example, weak adversarial samples created with the Fast Gradient Sign Method (FGSM), which are easy to compute, are used for fast adversarial training [5]. However, if stronger computationally-heavy adversarial attacks, such as the Projected Gradient Descent (PGD) [20], are used to attack a model trained with

Submitted to NeurIPS 2023 Workshop Backdoors in Deep Learning: The Good, the Bad, and the Ugly.

fast adversarial training, overfitting to the classification of FGSM adversarial samples can occur. In this case, the model trained with fast adversarial training can correctly classify FGSM adversarial samples, but its performance drops significantly (or to zero in the case of "catastrophic overfitting") for PGD adversarial samples [21]. Several adjustments have been proposed [22, 21, 23, 24] to circumvent this problem and make fast adversarial training effective, yet it remains still an active area of research. Another caveat to consider when using adversarial training is the trade-off between natural performance (classification accuracy of unperturbed samples) and adversarial performance (classification accuracy of perturbed samples) [25, 26, 27]. This natural-vs-adversarial performance trade-off is a consequence of the fact that while the naturally trained models focus on highly predictive features that may not be robust to adversarial attacks, the adversarially trained models select instead for robust features that may not be highly predictive [18].

While adversarial attacks can easily trick ANNs into misclassification, they appear ineffective for humans [28]. BP's learning algorithm differs drastically from biological learning mechanisms [29, 30, 31] and given that humans are not vulnerable to adversarial attacks, a fundamental research question is whether biologically-plausible learning algorithms are more robust to adversarial attacks. Researchers have made a significant effort in using the known learning principles of the brain to develop biologically-inspired algorithms as alternatives to BP [32, 33, 34, 35, 36, 37, 38, 39, 40, 41]. Thus, we investigate in detail for the first time whether biologically-inspired algorithms are robust against adversarial attacks. In this work, we chose *Present the Error to Perturb the Input To modulate Activity* (PEPITA), a recently proposed biologically-plausible learning algorithm [41], as a study case. In particular, we compare BP and PEPITA's learning algorithms in the following aspects:

- their intrinsic adversarial robustness (i.e., when trained solely on natural samples),
- their natural-vs-adversarial performance trade-off when trained with adversarial training,
- and their adversarial robustness against strong adversarial attacks when trained with weak adversarial samples (i.e., quality of fast adversarial training).

With this comparison, we open the door to drawing inspiration from biologically-plausible learning algorithms to develop more adversarially robust models.

2 Background - PEPITA

PEPITA is a learning algorithm developed as a biologically-inspired alternative to BP [41]. Its core difference from BP is that it does not require a separate backward pass to compute the gradients used to update the trainable parameters. Instead, a second forward pass is introduced (see Figure 1). In BP, the network processes the inputs $\mathbf{x}$ with one forward pass (indicated with black arrows) to produce the outputs $\mathbf{h}_L$, which are then compared to the target outputs $\mathbf{y}^*$ through a loss function. The error signal $\mathbf{e}$ computed by the loss function is then backpropagated through the entire network and used to train its parameters (indicated with red arrows). In PEPITA, the first forward pass is identical to BP. However, unlike BP, PEPITA feeds the error signals only to the softmax layer (directly) and to the input layer via a fixed random feedback matrix, F. This modulatory feedback is then added to the original input $\mathbf{x}$, producing the modulated inputs $\mathbf{x} + F\mathbf{e}$ that are processed in the second forward pass (illustrated with orange arrows). The difference between the activations of the neurons in the first and second forward passes is then used to train the parameters of the network. This procedure sidesteps the biologically-implausible requirement of BP to back-propagate gradient information through the network layers, allowing the training of the synaptic weights to be based on spatially local information with a two-factor Hebbian-like learning rule. Therefore, while BP uses exact gradients for learning, that is, the exact derivative of the loss function with respect to its trainable parameters, PEPITA uses a very different learning mechanism that leads to approximations of BP's exact gradients. Similarly to BP, FGSM and PGD adversarial attacks rely on using the exact derivatives of the loss function to perturb the input samples in the most harmful way. As PEPITA-trained models do not use these exact derivatives for learning, they form excellent candidates to be explored in the context of adversarial robustness.

3 Results

3.1 Model training details

For our comparative study, we use four benchmark computer vision datasets: MNIST [42], Fashion-MNIST [43], CIFAR-10 and CIFAR-100 [44]. For both BP and PEPITA, we used the same network

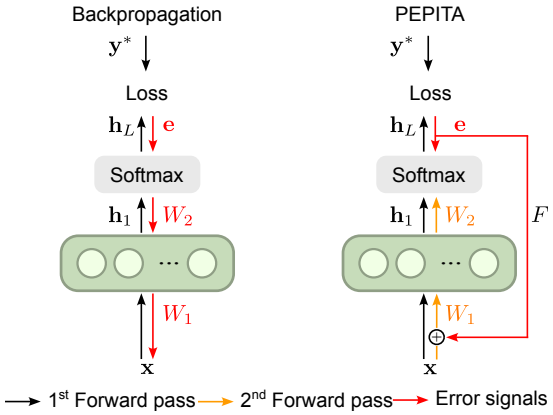

Figure 1: **Comparison of training in BP and PEPITA.** Schematic presentation of BP and PEPITA's single hidden layer networks and training algorithms.

Given: input ($\mathbf{x}$) and label ($\mathbf{y}^*$)

standard forward pass
$\mathbf{h}_0 = \mathbf{x}$
for i = 1,..., L:
  $\mathbf{h}_i = \sigma_i(W_{i-1}\mathbf{h}_{i-1})$
$\mathbf{e} = \mathbf{h}_L - \mathbf{y}^*$

modulated forward pass
$\mathbf{h}_0^{mod} = \mathbf{x} + F\mathbf{e}$
for i = 1,..., L:
  $\mathbf{h}_i^{mod} = \sigma_i(W_{i-1}\mathbf{h}_{i-1}^{mod})$
  if i<L:
    $\Delta W_i = (\mathbf{h}_i - \mathbf{h}_i^{mod}) \cdot (\mathbf{h}_{i-1}^{mod})^T$
  else:
    $\Delta W_L = \mathbf{e} \cdot (\mathbf{h}_{L-1}^{mod})^T$

Algorithm 1: **PEPITA's training algorithm.**

architectures and training schemes as described by [41] except for introducing a bias parameter, which improves performance across tasks. The learning rule for this bias is similar to the one for the synaptic weights, but the pre-synaptic activation is fixed to one, i.e., $\mathbf{h}_{i-1}^{mod} \coloneqq \mathbf{1}$. Similarly to the update rule for the synaptic weights (see Algorithm 1), the bias update rule can be written as $\Delta\mathbf{b}_i = (\mathbf{h}_i - \mathbf{h}_i^{mod})$ for $i < L$ and $\Delta\mathbf{b}_L = \mathbf{e}$. The network architecture consists of a single fully connected hidden layer with 1024 ReLU neurons and a softmax output layer (as represented in Figure 1). We used the mean-squared-error loss, trained the network for 100 epochs with early stopping, and optimized with momentum Stochastic Gradient Descent (SGD) [45]. Furthermore, we used a mini-batch size of 64, neuronal dropout of 10%, weight decay at epochs 60 and 90 with a rate of 0.1, and the He uniform initialization [46] with the feedback matrix F initialization scaled by 0.05. [1]

We optimized the learning rate hyperparameter through a grid search over 50 different values, and we defined the best-performing model as the model with the best natural accuracy on the validation dataset. We chose this model selection criterion because, in real-world applications, the networks' natural performance is most important to the user, and adversarial samples are outside of the norm. Thus, unless stated otherwise, we do not select the models based on the best adversarial validation accuracy, as we found this significantly worsens the natural performance of the model. The values reported throughout this section are the mean $\pm$ standard deviation of the test accuracy for 5 random seeds. For adversarial training, we used the open-source library *advertorch.attacks* [47] for the adversarial attacks, which follows the original implementations of FGSM and PGD, as introduced in [5] and [20], respectively. We defined an attack step size of 0.1 to create the FGSM and PGD adversarial samples and used 40 iterations for PGD. Note that the maximum and minimum pixel values of the adversarial images are the same as for the original natural images.

3.2 Baseline natural and adversarial performance

Table 1 shows the natural performance of the models when trained without adversarial data and with natural validation accuracy as the hyperparameter selection criterion. In line with the results reported in the literature [41], PEPITA achieves a lower natural performance than BP because, while BP uses exact derivatives of the loss to compute the gradients used for learning, PEPITA uses approximations of these gradients. Notably, both models are not robust to adversarial attacks since they have neither been adversarially trained nor their hyperparameter selection criterion valued adversarial robustness as an advantage.

3.3 PEPITA's intrinsic higher adversarial robustness

When using the same training procedure as in the section above (natural training) but selecting the hyperparameter search criterion to be the best adversarial validation accuracy, PEPITA shows a higher intrinsic adversarial robustness compared to BP, see Table 2. Although this model selection criterion leads to a certain level of adversarial robustness for PEPITA (see row 4 in Table 2), this comes at the cost of worse natural performance (compared to row 2 in Table 1) because of the

[1] PyTorch implementation of all methods will be available at a public repository.

		MNIST	Fashion-MNIST	CIFAR-10	CIFAR-100
BP	natural [%]	$98.58^{\pm 0.05}$	$90.52^{\pm 0.03}$	$57.05^{\pm 0.35}$	$27.54^{\pm 0.25}$
PEPITA	natural [%]	$98.16^{\pm 0.04}$	$86.46^{\pm 0.66}$	$52.15^{\pm 0.25}$	$25.88^{\pm 0.26}$

Table 1: Natural test accuracy. The hyperparameter selection criterion is the natural validation accuracy.

natural-vs-adversarial performance trade-off. However, when comparing the performance of BP with PEPITA for the MNIST dataset in Table 2, the natural performance of PEPITA decreases less than BP, and PEPITA is significantly more adversarially robust. Furthermore, BP cannot train adversarially robust models for more complex tasks, such as Fashion-MNIST, CIFAR-10, and CIFAR-100. During the hyperparameter search of BP, it was observed that the learning rates tended to be much larger with the current selection criterion (best adversarial validation accuracy). Consequently, the models either did not converge during learning, and the results were highly variable (see Table 2), or they did not learn at all, and the natural and adversarial performances were random.

		MNIST	Fashion-MNIST	CIFAR-10	CIFAR-100
BP	natural [%]	$94.22^{\pm 0.40}$	$43.82^{\pm 34.27}$	$10.003^{\pm 0.04}$	$9.078^{\pm 0.33}$
	PGD [%]	$92.72^{\pm 0.36}$	$22.89^{\pm 10.63}$	$9.98^{\pm 0.05}$	$0.33^{\pm 0.24}$
PEPITA	natural [%]	$97.69^{\pm 0.16}$	$80.65^{\pm 0.74}$	$41.82^{\pm 1.57}$	$17.10^{\pm 0.72}$
	PGD [%]	$97.56^{\pm 0.18}$	$80.48^{\pm 0.73}$	$41.73^{\pm 1.49}$	$16.76^{\pm 0.65}$

Table 2: Natural test accuracy and PGD adversarial test accuracy. The hyperparameter selection criterion is the adversarial validation accuracy.

3.4 PEPITA's advantageous adversarial training

When the models are now adversarially trained and the hyperparameter search selection criterion defined as the natural validation accuracy, we observe that PEPITA achieves a better adversarial testing performance and less natural performance degradation compared to BP (see Table 3), except for CIFAR100 where both models are not significantly adversarially robust. Moreover, for the MNIST and Fashion-MNIST datasets, BP has a better natural test accuracy for these adversarially trained models. Hence, although Table 1 suggests that PEPITA offers a better natural-vs-adversarial performance trade-off, a direct comparison of the adversarial robustness between BP and PEPITA becomes difficult for these datasets. To better understand this trade-off, we selected the most adversarially robust BP-trained and PEPITA-trained models for different fixed natural accuracy values on the MNIST task. We plotted these results in Figure 2A, which shows that PEPITA performs significantly better than BP for similar values of natural performance. Specifically, the average decrease in adversarial performance for the same values of natural performance is 0.26% for PEPITA and 8.05% for BP. Moreover, we verified that even if we double the number of training epochs for BP, its natural and adversarial accuracies remain approximately the same, indicating that the model has converged in its learning dynamics (see Figure 2B). Hence, even after extensive hyperparameter searches and increased training epochs, we could not find BP-trained models with a better natural-vs-adversarial performance trade-off.

		MNIST	Fashion-MNIST	CIFAR-10	CIFAR-100
BP (PGD)	natural [%]	$98.73^{\pm 0.06}$	$85.16^{\pm 0.17}$	$35.83^{\pm 0.37}$	$12.45^{\pm 0.38}$
	PGD [%]	$89.93^{\pm 0.03}$	$67.42^{\pm 0.21}$	$8.58^{\pm 0.16}$	$2.11^{\pm 0.15}$
PEPITA (PGD)	natural [%]	$98.17^{\pm 0.10}$	$83.73^{\pm 0.76}$	$45.12^{\pm 0.89}$	$22.30^{\pm 0.16}$
	PGD [%]	$96.93^{\pm 0.41}$	$83.19^{\pm 0.68}$	$44.94^{\pm 0.83}$	$2.88^{\pm 1.74}$

Table 3: Natural test accuracy and PGD adversarial test accuracy. All models are adversarially trained with PGD adversarial samples. The hyperparameter selection criterion is the natural validation accuracy.

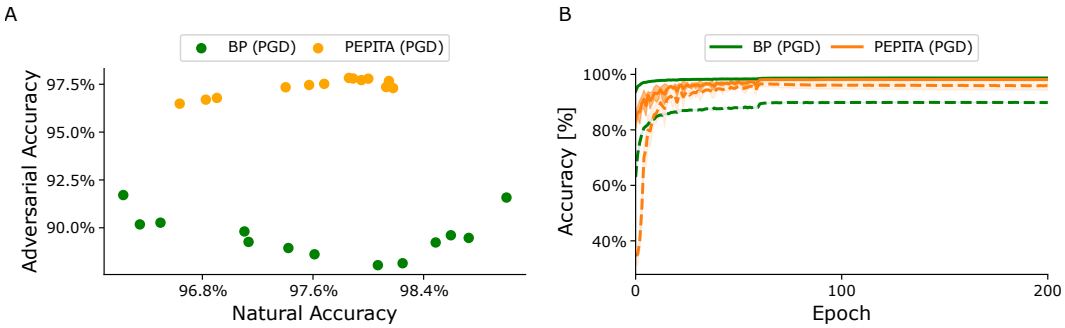

Figure 2: **PEPITA's advantageous adversarial training.** The results presented here are for BP and PEPITA models trained adversarially with PGD samples on the MNIST task. (A) Natural-vs-adversarial performance trade-off: for different natural accuracy values, the adversarial accuracies of the most adversarially robust BP and PEPITA-trained models are reported. (B) Natural (represented by the full lines) and adversarial (represented by the dashed lines) accuracies of BP and PEPITA-trained models for double the amount of epochs.

3.5 PEPITA's advantageous fast adversarial training

After demonstrating PEPITA's intrinsic adversarial robustness and beneficial natural-vs-adversarial performance trade-off, we now investigate PEPITA's capabilities in fast adversarial training [5]. Table 4 reports the results obtained when using fast adversarial training, i.e., when using FGSM samples for adversarial training, with the hyperparameter search selection criterion defined as the natural validation accuracy. We observe that when attacking the trained model with strong attacks, such as with PGD adversarial samples, the decrease in adversarial performance is much less significant for PEPITA than for BP, indicating that the PEPITA-trained models overfit less to the FGSM attacks. Moreover, both models do not suffer from catastrophic overfitting for this specific network architecture since the PGD testing accuracies do not drop to zero. This is the case because of two reasons: first, our network is over-parameterized, i.e., the network has more trainable parameters than there are samples in the dataset, so our large network width (1024 neurons) improves adversarial robustness; and second, we use He weights initialization, so our shallow network (a single hidden layer) also prevents a decrease in adversarial robustness [48]. To conclude, even if these models do not suffer from catastrophic overfitting, PEPITA has a more advantageous fast adversarial training since the gap between the FGSM and the PGD accuracies is much smaller for PEPITA than for BP.

		MNIST	Fashion-MNIST	CIFAR-10	CIFAR-100
BP (FGSM)	natural [%]	$98.93^{\pm 0.05}$	$84.90^{\pm 0.03}$	$51.56^{\pm 0.43}$	$26.59^{\pm 0.08}$
	FGSM [%]	$91.04^{\pm 0.13}$	$66.31^{\pm 0.25}$	$45.06^{\pm 3.38}$	$2.51^{\pm 0.37}$
	PGD [%]	$86.25^{\pm 0.09}$	$57.95^{\pm 0.33}$	$0.05^{\pm 0.04}$	$1.19^{\pm 0.08}$
PEPITA (FGSM)	natural [%]	$98.00^{\pm 0.14}$	$80.70^{\pm 0.95}$	$41.22^{\pm 2.01}$	$17.89^{\pm 0.52}$
	FGSM [%]	$97.91^{\pm 0.13}$	$80.68^{\pm 0.96}$	$41.22^{\pm 2.22}$	$17.68^{\pm 0.44}$
	PGD[%]	$97.81^{\pm 0.12}$	$80.27^{\pm 1.05}$	$41.00^{\pm 2.20}$	$17.53^{\pm 0.48}$

Table 4: Natural test accuracy and PGD and FGSM adversarial test accuracies. Adversarially trained models trained with FGMS adversarial samples. The hyperparameter selection criterion is the natural validation accuracy.

3.6 Investigations on PEPITA's adversarial robustness

Given our observations that PEPITA is adversarially more robust than BP, we aimed to investigate why this is the case. A central difference between both ANN learning algorithms is their gradient computation. While BP uses exact derivatives of the loss to compute the gradients used for learning, PEPITA uses alternative feedback and learning mechanisms that lead to approximations of these exact gradients [49]. To test the hypothesis that the error feedback signal shaping the approximate gradients of PEPITA is not just a random signal but contains essential information for enhancing adversarial robustness, we added random noise to BP's weight gradients and studied its adversarial robustness. We generated noise from a normal distribution with zero mean and a tunable standard deviation. We tested several hyperparameter combinations, including the standard deviation of the random noise,

and none of the parameter settings led to increased adversarial robustness for BP. In particular, BP's performance went from underperforming on classifying natural and adversarial samples for lower noise values to not being able to learn at all for higher noise values. Hence, we conclude that the critical factor leading to adversarial robustness is how the gradients are computed during learning. As many biologically-plausible learning algorithms use different feedback mechanisms and learning dynamics than BP to compute gradients, we can speculate that the resulting trained models possess better robustness against gradient-based adversarial attacks. Thus, an in-depth study of these could benefit the design of more adversarially robust models.

4 Discussion

Our paper demonstrates for the first time that biologically-inspired learning algorithms can lead to more adversarially robust ANNs than BP. We found that, unlike BP, PEPITA possesses intrinsic adversarial robustness, i.e., naturally trained PEPITA models can be robust against adversarial attacks without the computationally heavy burden of adversarial training. A similar finding of intrinsic adversarial robustness has been demonstrated by Akrout [50] for the biologically-plausible learning algorithm Feedback Alignment (FA) [32]. However, in this previous work Akrout [50], a non-common practice that leads to much weaker adversarial attacks was used: the attackers use the FA's random feedback matrices to generate adversarial samples instead of the transposed feedforward pathway. Hence, the analysis in Akrout [50] differs from our approach, where we let the attacker fully access the network architecture and craft its adversarial samples through the transposed forward pathway. Moreover, we found that PEPITA does not suffer from the natural-vs-adversarial performance trade-off as severely as BP, as its models can be more adversarially robust than BP while losing less natural performance. Lastly, we found that PEPITA benefits much more from fast adversarial training than BP, i.e., when trained with weaker adversarial attacks, it reports much better adversarial robustness against strong attacks.

4.1 Limitations and future work

Although the link between adversarial robustness and PEPITA has been well established, a theoretical understanding of PEPITA's advantageous adversarial training and intrinsic adversarial robustness is still missing. Also, understanding this phenomenon's theoretical foundation could help identify the exact properties that improve the natural-vs-adversarial performance trade-off. Moreover, PEPITA has recently been extended to deeper networks (up to five hidden layers) and tested with different parameter initialization schemes [49], so studying the impact of other characteristics of the model, such as width, depth, and initialization on PEPITA's adversarial robustness would be beneficial (as done in [48]). PEPITA has also recently been combined with weight mirroring so that its feedback projection matrix can be learned [49], so it would be interesting to study whether this improves not only natural performance but also adversarial robustness. While PEPITA, to our knowledge, was the first model being investigated regarding adversarial robustness, this kind of analysis should also be done for several other BP-alternative biologically-plausible learning algorithms that have been proposed [33, 34, 35, 36, 38]. Hence, our work paves the way for a systematic assessment of the properties that lead to adversarially robust models.

4.2 Conclusion

We demonstrated that ANNs trained with PEPITA, a recently proposed biologically-inspired learning algorithm, are more adversarially robust than BP-trained ANNs. In particular, we showed through several computational experiments that PEPITA performs significantly better than BP in an adversarial setting. Our analysis opens the door to drawing inspiration from biologically-plausible learning algorithms for designing more adversarially robust models. Thus, our work contributes to the future development of more adversarially robust ANNs and, consequently, to the creation of safer and more trustworthy artificial intelligence systems.

Acknowledgments and Disclosure of Funding

This work was supported by the Swiss National Science Foundation (315230_189251 1). We thank the IBM Zürich research group 'Emerging Computing and Circuits' for all the fruitful discussions during the development of this work. We would like to thank Anh Duong Vo, Sander de Haan, and Federico Villani for their feedback and Pau Vilimelis Aceituno for the insightful discussions.

References

[1] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533, 1986.

[2] Paul J Werbos. Applications of advances in nonlinear sensitivity analysis. In *System modeling and optimization*, pages 762–770. Springer, 1982.

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks, 2019. arXiv:1706.06083 [cs, stat].

[4] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014. arXiv:1312.6199 [cs].

[5] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[6] Wei Emma Zhang, Quan Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology*, 11:1–41, 2020. doi: 10.1145/3374217.

[7] John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP, 2020. arXiv:2005.05909 [cs].

[8] Naoya Takahashi, Shota Inoue, and Yuki Mitsufuji. Adversarial attacks on audio source separation. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 521–525, 2021. doi: 10.1109/ICASSP39728.2021.9414844.

[9] Shehzeen Hussain, Paarth Neekhara, Shlomo Dubnov, Julian McAuley, and Farinaz Koushanfar. WaveGuard: Understanding and Mitigating Audio Adversarial Examples, 2021. arXiv:2103.03344 [cs, eess].

[10] Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial Policies: Attacking Deep Reinforcement Learning. In *International Conference on Learning Representations*, 2020.

[11] Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary. Robust deep reinforcement learning with adversarial attacks. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, page 2040–2042, 2018.

[12] Iqbal Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2, 03 2021. doi: 10.1007/s42979-021-00592-x.

[13] Naveed Akhtar and Ajmal Mian. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. *IEEE*, 6:14410–14430, 2018. ISSN 2169-3536. doi: 10.1109/ACCESS.2018.2807385.

[14] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning models, 2018. arXiv:1707.08945 [cs].

[15] Jia Wang, Chengyu Wang, Qiuzhen Lin, Chengwen Luo, Chao Wu, and Jianqiang Li. Adversarial attacks and defenses in deep learning for image recognition: A survey. *Neurocomputing*, 514:162–181, 2022. ISSN 0925-2312. doi: 10.1016/j.neucom.2022.09.004.

[16] Maximilian Kaufmann, Yiren Zhao, Ilia Shumailov, Robert Mullins, and Nicolas Papernot. Efficient adversarial training with data pruning, 2022. arXiv:2207.00694 [cs].

[17] Sravanti Addepalli, Samyak Jain, and Venkatesh Babu R. Efficient and effective augmentation strategy for adversarial training. In *Advances in Neural Information Processing Systems*, volume 35, pages 1488–1501, 2022.

[18] H. Zheng, Z. Zhang, J. Gu, H. Lee, and A. Prakash. Efficient adversarial training with transferable adversarial examples. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1178–1187. IEEE Computer Society, 2020. doi: 10.1109/CVPR42600.2020.00126.

[19] Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, and Venkatesh Babu R. Towards Efficient and Effective Adversarial Training. In *Advances in Neural Information Processing Systems*, volume 34, pages 11821–11833, 2021.

[20] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world, 2016. arXiv:1607.02533 [cs].

[21] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *8th International Conference on Learning Representations, ICLR*, 2020.

[22] Hoki Kim, Woojin Lee, and Jaewook Lee. Understanding catastrophic overfitting in single-step adversarial training. In *AAAI Conference on Artificial Intelligence*, 2020.

[23] Zeinab Golgooni, Mehrdad Saberi, Masih Eskandar, and Mohammad Hossein Rohban. Zerograd: Mitigating and explaining catastrophic overfitting in fgsm adversarial training, 2021. arXiv.2103.15476 [cs].

[24] Peilin Kang and Seyed-Mohsen Moosavi-Dezfooli. Understanding catastrophic overfitting in adversarial training, 2021. arXiv:2105.02942 [cs].

[25] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations, ICLR*, 2019.

[26] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, volume 97, pages 7472–7482, 2019.

[27] Mazda Moayeri, Kiarash Banihashem, and Soheil Feizi. Explicit Tradeoffs between Adversarial and Natural Distributional Robustness. In *Advances in Neural Information Processing Systems*, 2022.

[28] Zhenglong Zhou and Chaz Firestone. Humans can decipher adversarial images. *Nat. Commun.*, 10(1334):1–9, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08931-6.

[29] Francis Crick. The recent excitement about neural networks. *Nature*, 337(6203):129–132, 1989.

[30] Stephen Grossberg. Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, 11(1):23–63, 1987.

[31] Timothy P Lillicrap, Adam Santoro, Luke Marris, Colin J Akerman, and Geoffrey Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, pages 1–12, 2020.

[32] Timothy P Lillicrap, Daniel Cownden, Douglas B Tweed, and Colin J Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7:13276, 2016.

[33] Dong-Hyun Lee, Saizheng Zhang, Asja Fischer, and Yoshua Bengio. Difference target propagation. In *Joint european conference on machine learning and knowledge discovery in databases*, pages 498–515. Springer, 2015.

[34] James CR Whittington and Rafal Bogacz. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural computation*, 29(5):1229–1262, 2017.

[35] Benjamin Scellier and Yoshua Bengio. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in computational neuroscience*, 11:24, 2017.

[36] João Sacramento, Rui Ponte Costa, Yoshua Bengio, and Walter Senn. Dendritic cortical microcircuits approximate the backpropagation algorithm. In *Advances in Neural Information Processing Systems 31*, pages 8721–8732, 2018.

[37] Mohamed Akrout, Collin Wilson, Peter Humphreys, Timothy Lillicrap, and Douglas B Tweed. Deep learning without weight transport. In *Advances in Neural Information Processing Systems 32*, pages 974–982, 2019.

[38] Alexander Meulemans, Matilde Tristany Farinha, Javier García Ordóñez, Pau Vilimelis Aceituno, João Sacramento, and Benjamin F Grewe. Credit assignment in neural networks through deep feedback control. In *Advances in Neural Information Processing Systems*, 2021.

[39] Geoffrey Hinton. The Forward-Forward Algorithm: Some Preliminary Investigations, 2022. arXiv:2212.13345 [cs].

[40] Thomas Bohnstingl, Stanisław Woźniak, Angeliki Pantazi, and Evangelos Eleftheriou. Online Spatio-Temporal Learning in Deep Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2022. ISSN 2162-2388. doi: 10.1109/TNNLS.2022.3153985.

[41] Giorgia Dellaferrera and Gabriel Kreiman. Error-driven input modulation: Solving the credit assignment problem without a backward pass. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 4937–4955. PMLR, 2022.

[42] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[43] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. arXiv:1708.0774 [cs].

[44] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The CIFAR-10 dataset. *online: http://www.cs. toronto. edu/kriz/cifar. html*, 2014.

[45] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999. ISSN 0893-6080. doi: 10.1016/S0893-6080(98)00116-6.

[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. doi: 10.1109/ICCV.2015.123.

[47] Gavin Weiguang Ding, Luyu Wang, and Xiaomeng Jin. advertorch v0.1: An adversarial robustness toolbox based on pytorch, 2019. arXiv:1902.07623 [cs].

[48] Zhenyu Zhu, Fanghui Liu, Grigorios G. Chrysos, and Volkan Cevher. Robustness in deep learning: The good (width), the bad (depth), and the ugly (initialization). In *Advances in neural information processing systems*, 2022.

[49] Ravi Francesco Srinivasan, Francesca Mignacco, Martino Sorbaro, Maria Refinetti, Avi Cooper, Gabriel Kreiman, and Giorgia Dellaferrera. Forward Learning with Top-Down Feedback: Empirical and Analytical Characterization, 2023. arXiv:2302.05440 [cs].

[50] Mohamed Akrout. On the adversarial robustness of neural networks without weight transport, 2019. arXiv:1908.03560 [cs].