ADVERSARIAL ROBUSTNESS BY DESIGN THROUGH ANALOG COMPUTING AND SYNTHETIC GRADIENTS

Alessandro Cappelli[*1], *Ruben Ohana*[*1,2], *Julien Launay*[1,2], *Laurent Meunier*[3,4],
Iacopo Poli[1], *Florent Krzakala*[1,2,5]
[1]LightOn, Paris, France
[2]Laboratoire de Physique, Ecole Normale Supérieure, Paris, France
[3]Facebook AI Research, Paris, France
[4]LAMSADE, Université Paris-Dauphine, Paris, France
[5]IdePHICS Laboratory, EPFL, Switzerland

ABSTRACT

We propose a new defense mechanism against adversarial attacks inspired by an optical co-processor, providing robustness without compromising natural accuracy in both white-box and black-box settings. This hardware co-processor performs a nonlinear fixed random transformation, where the parameters are unknown and impossible to retrieve with sufficient precision for large enough dimensions. In the white-box setting, our defense works by obfuscating the parameters of the random projection. Unlike other defenses relying on obfuscated gradients, we find we are unable to build a reliable backward differentiable approximation for obfuscated parameters. Moreover, while our model reaches a good natural accuracy with a hybrid backpropagation - synthetic gradient method, the same approach is suboptimal if employed to generate adversarial examples. Finally, our hybrid training method builds robust features against black-box and transfer attacks. We demonstrate our approach on a VGG-like architecture, placing the defense on top of the convolutional features, on CIFAR-10 and CIFAR-100.

Index Terms— Adversarial robustness, optical computing, direct feedback alignment, analog computing

1. INTRODUCTION

Neural networks are sensitive to small, imperceptible to humans, perturbations of their inputs that can cause state-of-the-art classifiers to completely fail [1]. As deep learning models are deployed in real-world applications, guaranteeing their robustness to malicious actors becomes increasingly important: for instance, an adversarial image could evade automated content filtering on social networks [2]. Adversarial attacks can be carried out in different frameworks: in the white-box setting, the attacker has full access to the model, while black-box attacks only rely on queries. It is also possible to craft an attack on a different model and transfer it to the model targeted [3]. There is no universal defense, and state-of-the-art techniques often come with a large computational cost, as well as reduced natural accuracy [4]. Some of these defenses rely on obfuscated gradients: the model is designed so that the gradients are unsuitable for attacks, for instance by using non-differentiable layers. However, attackers can choose to alter the network structure, using Backward Pass Differentiable Approximation (BPDA) [5], replacing obfuscating layers with well-behaved approximations. Furthermore, approaches relying on obfuscation do not generally provide robustness against transfer and black-box attacks.

We expand the idea of obfuscated gradients to *obfuscated parameters*: we physically implement a fixed random projection followed by a non-linearity using an optical co-processor, where only the distribution of the random matrix entries is known and not their values. Even though retrieval is possible, the computational cost becomes quickly prohibitive with increasing dimension, and is limited in precision [6, 7]. To train layers below our non-differentiable defense, we draw inspiration from Direct Feedback Alignment (DFA) [8] and bypass it in the backward pass. We use a random mapping of the global error to train the layer below the co-processor, while the layers further downstream perform backpropagation (BP) from this *synthetic gradient* signal. This comes at no natural accuracy cost.

Our defense is robust by design against white-box attacks: we find that BPDA is ineffective against obfuscated parameters, and attackers are forced to rely on DFA to attack the network. We develop such DFA attacks, and find them much less effective than attacks based on backpropagation on BP-trained networks, confirming results from [9].

We also test models incorporating our defense against black-box and transfer attacks, and find that they are more robust than their vanilla counterparts. We perform all experimental benchmarks on CIFAR-10, and CIFAR-100 [10]. White-box and transfer of attacks results are obtained with a real optical co-processor as a defense whereas for black-box attacks, a simulated co-processor is used for convenience.

[*]Equal contribution. Contact: {alssandro,ruben}@lighton.ai

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

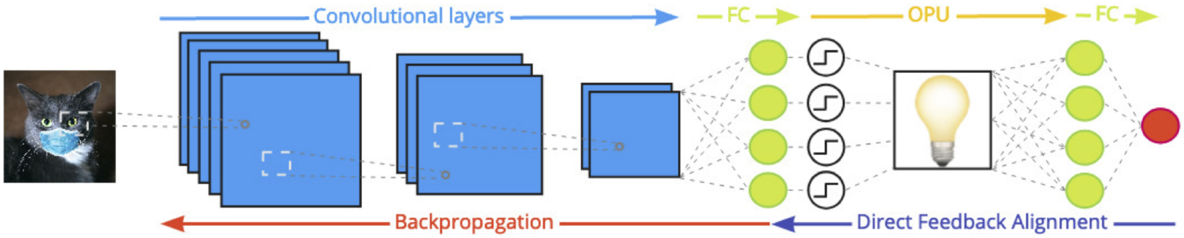

Fig. 1: A convolutional neural network with an Optical Processing Unit (OPU), our analog defense layer against adversarial attacks. All together, the unknown parameters of the analog operation, the binarization, and the hybrid training method based on Direct Feedback Alignment (DFA) form a defense against white-box, black-box, and transfer attacks.

2. PRELIMINARIES

2.1. Adversarial attacks and defenses

White-box attacks are adversarial attacks where the attacker is assumed to have full access to the model, including its parameters. In this case, the attacker usually computes a *gradient attack* (e.g. FGSM [1], PGD [11, 12], or Carlini & Wagner [13]). These attacks are often fast, effective, and easy to compute. Some defenses obfuscate the gradients to neutralize these attacks: however, it is often possible to build a differentiable approximation (BPDA) to perform gradient-based attacks [5], and black-box attacks entirely elude such defenses. In our work, we focus on PGD for white-box scenarios.

Black-box attacks assumes that the attacker has only limited access to the network: for instance, only the label, or the logits for a given input are known. There exist two main categories of black-box attacks. On one hand, gradient estimation attacks [14, 15, 16] aim to estimate the gradient of the loss with respect to the input to mimic gradient-based attacks. On the other hand, adversarial attacks are transferable [3]: an attack effective on a given network is likely to also fool another network. More recently, black-box methods based on optimisation tools derived from genetic algorithms [17, 18] and combinatorial optimization [19] have been introduced. We evaluate our defense against NES and bandits, two gradient-estimation attacks, and parsimonious black-box methods, as well as transfer of attacks.

Defenses – Historically, the first defense proposed against attacks was adversarial training [1, 12] (i.e. training the neural network through a min-max optimization framework) thus including adversarial robustness as an explicit training objective. Despite its simplicity and lack of theoretical guarantees, adversarial training is still one of the most effective defense against adversarial examples. Theoretically proven defenses also exist, such as randomized smoothing [20, 21, 22, 23] or convex relaxation [24, 25]. In the literature, numerous defenses do not evaluate their models on attacks adapted to the defense [26]: especially in the case of gradient obfuscation [5], which can result in a false sense of security. In contrast, we evaluate our defense in a wide range of scenarios, and introduce new DFA-based white-box attacks. Finally, defense techniques often demand extra computations and reduce natural accuracy. Instead, our defense computations are offloaded to the optical co-processor and the decrease in natural accuracy is minimal, in contrast with adversarial training.

2.2. Our defense: the Optical Processing Unit

Our defense relies on an analog layer implemented by an Optical Processing Unit (OPU)[1]. The OPU is a co-processor that multiplies an input vector x by a fixed random matrix U, using light scattering through a diffusive medium [27]. The measurement process implies an inherent absolute value squared non-linearity. Effectively, the operation performed is:

$$m = |Ux|^2 \tag{1}$$

The OPU input is binary (1 bit) and its output is encoded in 8 bits. The size of the random matrix can reach $10^6 \times 10^6$, and its entries are complex Gaussian distributed, but their values are not known. While our defense can be simulated without dedicated hardware, an advantage of using the OPU is that even if the host system is compromised, the random matrix remains unknown, as it is physically implemented by the diffusive medium of the OPU.

Retrieval of the matrix is possible [6, 7] with direct access to the OPU, but it is computationally expensive for large enough dimensions and relative errors can be as high as 30%. The fastest known method [7] for the retrieval of a matrix $U \in \mathbb{C}^{M \times N}$ relies on the multilateration of anchor signals and has $O(MN \log N)$ time complexity. For $N \sim 10^4$ and $M \sim 10^5$, retrieval with relative error of 32.0% takes 72 minutes. If we wanted to recover the same matrix with a relative error of 8.0%, we would need 19 hours, as decreasing the relative error has a quadratic cost. The optical system can scale up to $N, M \sim 10^6$: at these dimensions the random matrix

[1] Access through LightOn Cloud `https://cloud.lighton.ai/`

alone takes about 8 TB to store, making memory the main constraint in the retrieval. Finally, it is also easily possible to change the entries of the matrix of the optical system to another draw from the same probability distribution. To adapt to this new random matrix, only the classifier has to be fine-tuned: this enables a defense strategy where the random matrix is regularly resampled, preventing malicious actors from having enough time to recover it.

Accordingly, our defense effectively achieves *parameter obfuscation*, preventing attackers from building a differentiable approximation that can be used to reliably generate adversarial examples. In our work we use an actual optical co-processor for white-box and transfer attacks, and a simulated one for black-box. Note that while we simulate input binarization, we do not simulate output quantization to 8 bits, as we find this quantization to be of little influence.

2.3. Network training and adversarial examples generation with synthetic gradients

Because its parameters are obfuscated, it is not possible to perform backpropagation through the layer implemented by the optical co-processor. To train neural networks incorporating our defense, we draw inspiration from Direct Feedback Alignment [8], and design a hybrid training method (Figure 1). We train the layers above our defense through backpropagation, but train the layer below it by directly using a random projection of the global error as the teaching signal. To account for the inability of DFA to train convolutional layers [28], the layer directly before our defense should be a fully-connected layer. This synthetic signal is then backpropagated to convolutional layers further down. We use the same hybrid method to generate white-box attacks against our networks.

3. EXPERIMENTAL RESULTS

We place our defense after the convolutional layers of a VGG-16 architecture [29]. We call this network *VGG-OPU*. The training is performed with the hybrid BP-DFA algorithm discussed in the Section 2.3, and shown in Figure 1. The code of DFA is taken from [30]. We consider the CIFAR-10 and CIFAR-100 datasets.

We attack our models using only images that were correctly classified, with the exception of white-box attacks, where we use the full datasets. All the attacks are *untargeted*: we aim to change the classification without any specific target label. Finally, the loss used for computing the attacks is the cross-entropy between the output of the classifier for a given input and its label: $l(x, y) = -\log p_\theta(y|x)$. As we consider untargeted attacks, we aim to maximize it.

3.1. White-box attacks

We first consider white-box attacks: the attacker has full knowledge of the model and its parameters, and can craft adversarial examples by gradient-based methods. We show that our parameter obfuscation approach makes these attacks significantly less effective, forcing them to rely on imprecise gradient approximations based on DFA or BPDA.

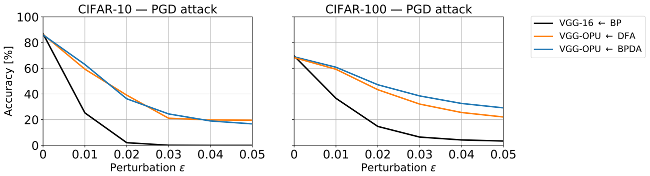

Fig. 2: Notation: <model> ← <attack gradients>. For example VGG ← BP means that a VGG-16 is attacked with gradients computed with backpropagation. The VGG-OPU model is systematically more robust than the VGG-16 baseline. The perturbation ε is the maximum radius of the ℓ_2-ball of the perturbation of the attack.

To create adversarial attacks with PGD we used 50 iterations with $\alpha = 0.01$ step size on images normalized in $[-1, 1]$.

Results are shown in Figure 2. We find the VGG-OPU model incorporating our defense performs better than the VGG-16 baseline for any value of ε – the upper bound on the ℓ_2-norm of the perturbation – with gains in accuracy under attack ranging from 20% to 40%. For the largest values of ε considered, while the accuracy of the baseline goes to zero, the VGG-OPU model is still performing better than a random guess. These results also show that our obfuscated parameters approach cannot be fooled by a simple BPDA like obfuscated gradients can be: BPDA is here ineffective at finding better attacks than simply bypassing our defense with DFA. Finally, we note the increased robustness does not come at a natural accuracy cost ($\varepsilon = 0$).

3.2. Black-box attacks

If the obfuscated parameters approach provides robustness by design against white-box attacks, black-box approaches should be not affected, since they do not require knowledge of the weights. However, we find our defense still brings robustness against such attacks.

3.2.1. NES, bandits, and parsimonious attacks

We measure the Cumulative Success Rates (CSR) in terms of elapsed queries budget. We fix a maximum budget of 15000 queries to the classifier for black-box attacks. This budget is sufficient to reach a plateau in the success rate of the attacks, and to achieve 100% success rate with parsimonious attacks on an undefended baseline.

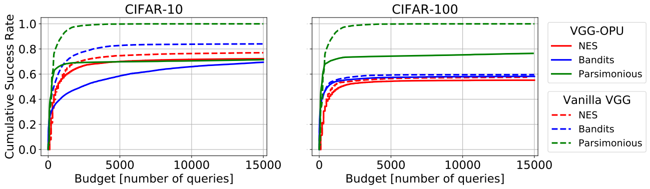

Fig. 3: Cumulative Success Rate w.r.t. the number of queries for different **black-box attacks** on the CIFAR-10 and CIFAR-100 datasets, on a VGG-16 **(dashed lines)** and a VGG-OPU with a simulated OPU **(plain lines)**. The architectures with our defense perform on par or better than the baseline. The parsimonious attack has the highest success rate, that with our defense drops by 0.3 on CIFAR-10 and 0.2 on CIFAR-100.

Black-box attacks were performed on images normalized in $[0, 1]$ and on the L_∞ ball of radius $\varepsilon = 8/256$. *NES attacks* number of samples to estimate the gradients is 50 and the size of a batch 1024. Standard deviation σ is varied at values $[0.05, 0.1, 0.5, 1]$ and we keep the one yielding the best Cumulative Success Rate (CSR) for each model on each dataset. *Bandits attacks* number of gradient iterations is kept to 1, the online learning rate at 0.1, the exploration at 0.1 and the prior size at 16. *Parsimonious attacks* number of iterations in local search is 1, initial block size is 4, batch size is 64 and no hierarchical evaluation was performed.

Results are shown in Figure 3. For the gradient-estimation attacks (NES and bandits), our defense improves robustness with respect to the baseline by decreasing the CSR respectively by 5% and 10 $\%$ for the largest budget on CIFAR-10. On CIFAR-100, the improvement in robustness is minimal, of the order of 1% for both attacks. However, our defense shows significant improvement against parsimonious attacks (that largely outperform both NES and bandits), reducing the CSR by 30% on CIFAR-10 and by 24% CIFAR-100. Overall, the best attack on Vanilla VGG reaches 100% CSR on both datasets whereas the best attack on a VGG with our defense reaches only 70% on CIFAR-10 and 76% on CIFAR-100.

3.2.2. *Transfer attacks*

Finally, we test the robustness of our defense against transfer attacks. In this scenario, attacks are crafted on a separate source network built by the attacker, and then transferred to the target network.

To evaluate transfer attacks, we first create a test set of well classified samples common to both the source and target network. We then perform attacks on a VGG-16 model on this dataset using the PGD algorithm, building a collection of adversarial examples to transfer to other networks.

The target networks are a vanilla VGG-16 trained with backpropagation, and a VGG-OPU optimized as follows: we first train a vanilla VGG-16 with our hybrid training method and then place our OPU defense after the trained convolutional stack, finetuning only the classifier layer.

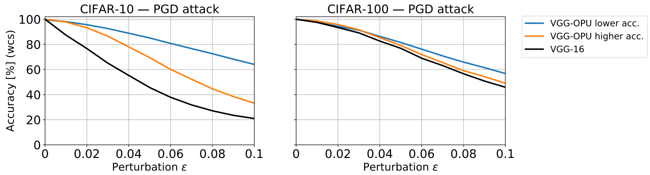

Fig. 4: Accuracy on a common well-classified set under **transfer attacks** of increasing strength. Both VGG-OPU networks are more robust than the the baseline, and we observe a trade-off between robustness and natural accuracy with our defense. Higher/lower accuracies are $\sim$ 85%/80% for CIFAR10 and $\sim$ 72%/68% for CIFAR100. The baseline VGG-16 was trained to reach the higher accuracy.

We also study two different VGG-OPU models, with identical architectures, but different overall natural accuracy to evaluate if there exists a robustness-accuracy trade-off.

Results The results of this study are shown in Figure 4. The VGG-OPU network performs better against transfer attacks for all perturbation strength values. On CIFAR-10, we find our defense to provide significant robustness to transfer, between $+15\%$ and $+45\%$ of robustness on the well classified evaluation dataset. This gain is smaller on CIFAR-100, between $+3\%$ and $+10\%$ robustness, where the transfer attack is overall less effective. In either case, we find that there exists an accuracy-robustness trade-off: at the cost of 5% of natural accuracy, robustness can be increased. This makes the defense customizable, allowing to trade some accuracy for robustness.

4. CONCLUSION

We introduced a new defense technique against white-box, black-box and transfer attacks, based on the analog implementation of a neural network layer using an optical co-processor. Our method incurs no additional computational cost at training time, and comes at no natural accuracy cost.

We evaluated our method in each setting on the CIFAR-10 and CIFAR-100 datasets, against various attacks. In the white-box setting, our defense is robust by design thanks to *parameter obfuscation*. We attempted to adapt white-box attacks to break this defense, testing two different differentiable approximations, however both resulted in less convincing adversarial examples. Furthermore, in the black-box setting, our defense improves robustness by 22% against the strongest black-box attack that we tested. We also showed increased robustness to transfer of attacks, and showed that there exists in this instance a robustness-accuracy trade-off.

Future work could investigate the combination of our defense with other techniques based on en- sembling or adversarial training. Finally, we note that attackers succeeding in efficiently breaking our obfuscation of parameters could have a significant impact in imaging and phase retrieval.

5. REFERENCES

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

[2] Evrard Garcelon, Baptiste Roziere, Laurent Meunier, Jean Tarbouriech, Olivier Teytaud, Alessandro Lazaric, and Matteo Pirotta, "Adversarial attacks on linear contextual bandits," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[3] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.

[4] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry, "Robustness may be at odds with accuracy," *International Conference on Learning Representation*, 2019.

[5] Anish Athalye, Nicholas Carlini, and David Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, July 2018.

[6] Sidharth Gupta, Rémi Gribonval, Laurent Daudet, and Ivan Dokmanić, "Don't take it lightly: Phasing optical random projections with unknown operators," in *Advances in Neural Information Processing Systems*, 2019, pp. 14855–14865.

[7] Sidharth Gupta, Rémi Gribonval, Laurent Daudet, and Ivan Dokmanić, "Fast optical system identification by numerical interferometry," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1474–1478.

[8] Arild Nøkland, "Direct feedback alignment provides learning in deep neural networks," in *Advances in neural information processing systems*, 2016, pp. 1037–1045.

[9] Mohamed Akrout, "On the adversarial robustness of neural networks without weight transport," 2019.

[10] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton, "Cifar-10 and cifar-100 datasets," *URl: https://www. cs. toronto. edu/kriz/cifar. html*, vol. 6, 2009.

[11] Alexey Kurakin, Ian Goodfellow, and Samy Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[12] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[13] Nicholas Carlini and David Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 39–57.

[14] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. ACM, 2017, pp. 15–26.

[15] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin, "Black-box adversarial attacks with limited queries and information," *arXiv preprint arXiv:1804.08598*, 2018.

[16] Andrew Ilyas, Logan Engstrom, and Aleksander Madry, "Prior convictions: Black-box adversarial attacks with bandits and priors," *arXiv preprint arXiv:1807.07978*, 2018.

[17] Laurent Meunier, Jamal Atif, and Olivier Teytaud, "Yet another but more efficient black-box adversarial attack: tiling and evolution strategies," *arXiv preprint arXiv:1910.02244*, 2019.

[18] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein, "Square attack: a query-efficient black-box adversarial attack via random search," *arXiv preprint arXiv:1912.00049*, 2019.

[19] Seungyong Moon, Gaon An, and Hyun Oh Song, "Parsimonious black-box adversarial attacks via efficient combinatorial optimization," *arXiv preprint arXiv:1905.06635*, 2019.

[20] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *2019 IEEE Symposium on Security and Privacy (SP)*, 2018, pp. 727–743.

[21] Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter, "Certified adversarial robustness via randomized smoothing," *arXiv preprint arXiv:1902.02918*.

[22] Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cédric Gouy-Pailler, and Jamal Atif, "Theoretical evidence for adversarial robustness through randomization: the case of the exponential family," *arXiv preprint arXiv:1902.01148*, 2019.

[23] Rafael Pinot Alexandre Araujo, Laurent Meunier and Benjamin Negrevergne, "Advocating for multiple defense strategies against adversarial examples," *Workshop on Machine Learning for CyberSecurity (MLCS@ECML-PKDD)*, 2020.

[24] Eric Wong and Zico Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *International Conference on Machine Learning*, 2018, pp. 5286–5295.

[25] Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter, "Scaling provable adversarial defenses," in *Advances in Neural Information Processing Systems*, 2018, pp. 8400–8409.

[26] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry, "On adaptive attacks to adversarial example defenses," *arXiv preprint arXiv:2002.08347*, 2020.

[27] LightOn, "Photonic computing for massively parallel AI - A White Paper, v1.0," `https://lighton.ai/wp-content/uploads/2020/05/LightOn-White-Paper-v1.0.pdf`, May 2020.

[28] Julien Launay, Iacopo Poli, and Florent Krzakala, "Principled training of neural networks with direct feedback alignment," *arXiv preprint arXiv:1906.04554*, 2019.

[29] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[30] Julien Launay, Iacopo Poli, François Boniface, and Florent Krzakala, "Direct feedback alignment scales to modern deep learning tasks and architectures," *Advances in Neural Information Processing Systems*, vol. 33, 2020.