ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Slope: A First-order Approach for Measuring Gradient Obfuscation

Maura Pintor[1,2], Luca Demetrio[1], Giovanni Manca[1], Battista Biggio[1,2], Fabio Roli[1,2]

1- University of Cagliari, Italy

2- Pluribus One

Abstract. Evaluating adversarial robustness is a challenging problem. Many defenses have been shown to provide a false sense of security by unintentionally obfuscating gradients, hindering the optimization process of gradient-based attacks. Such defenses have been subsequently shown to fail against adaptive attacks crafted to circumvent gradient obfuscation. In this work, we present *Slope*, a metric that detects obfuscated gradients by comparing the expected and the actual increase of the attack loss after one iteration. We show that our metric can detect the presence of obfuscated gradients in many documented cases, providing a useful debugging tool towards improving adversarial robustness evaluations.

1 Introduction

Adversarial attacks perturb input data under feasibility constraints with the goal of subverting predictions of machine-learning algorithms at test time [1, 2, 3]. To counter this behaviour, defenses have been designed to trigger failures of gradient-based attacks [4, 5], such as *gradient obfuscation* [4, 6], where attackers are incapacitated in finding a good direction that leads to adversarial examples since the gradients of the model are too small or noisy. For this reason, standard gradient-based attacks fail against such defenses, yielding over-optimistic robustness evaluations. However, attackers started to develop *adaptive* attacks [7, 8], i.e. strategies that target a particular defense mechanism to still obtain successful evasion against them, proving their weakness, fueling an arms race between attackers and defenders. On the other hand, the creation of such adaptive attacks requires manual inspection of the target under attack during the security evaluation, since there is no quantifiable method for detecting the presence of gradient obfuscation.

To tackle this limitation, we propose *Slope*, a metric that quantifies the ease of decreasing the attacker loss, computed on single points. Such metric is sensitive to the abrupt changes of models, and it leverages the error that is committed by approximating the attacker loss during the optimization process. In this way, Slope can detect the presence of obfuscated gradients, and the attack can be patched to tackle such challenge. To show the efficacy of Slope, we compute attacks against four models, two that apply gradient obfuscation, a baseline undefended one, and one that is adversarially trained. We highlight that our metric is triggered when the attacks are failing due to gradient obfuscation.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

2 Slope: Measuring Gradient Obfuscation

In this section, we present *Slope*, our metric to quantify gradient obfuscation.

Notation. Let $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ be a sample labeled as class $y \in \mathcal{Y} = \{1, ..., C\}$. Given the parameters of the target model $\boldsymbol{\theta}$, and a loss function of choice $\mathcal{L}$, the *robust accuracy* is defined as:

$$\mathbb{E}_{(\boldsymbol{x},y)} \left[\max_{\boldsymbol{\delta} \in \Delta} \mathcal{L}\left(\boldsymbol{x} + \boldsymbol{\delta}, y; \boldsymbol{\theta}\right) \right],$$

where the maximization refers to the *adversarial loss*, i.e., the loss attained against the worst-case perturbation $\boldsymbol{\delta}$ found within the feasible region Δ. This region is typically defined as an ℓ_p ball centered on the input sample $\boldsymbol{x}$, i.e., $\Delta : \|\boldsymbol{\delta}\|_p \leq \epsilon$, and the perturbation $\boldsymbol{\delta}$ is optimized via gradient-based algorithms, such as *Projected Gradient Descent* (PGD), that iteratively updates the perturbation along the gradient direction of the loss. The success of these attack strategies clearly relies on the fact that the loss function is sufficiently smooth, i.e., that the corresponding gradients contain meaningful information to iteratively improve the objective.

Measuring Obfuscation with Slope. The underlying idea of the *Slope* metric is to compare the expected loss after one gradient update of size η against the actual observed loss after the update. To this end, we consider one normalized PGD step of size η, for which the expected loss increment is obtained by solving this linearized problem:

$$\max_{\|\boldsymbol{\delta}\|_p \leq \eta} \mathcal{L}(\boldsymbol{x} + \boldsymbol{\delta}; y; \boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{x}, y; \boldsymbol{\theta}) \approx \max_{\|\boldsymbol{\delta}\|_p \leq \eta} \boldsymbol{\delta}^\top \underbrace{\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, y; \boldsymbol{\theta})}_{\boldsymbol{g}} = \eta \|\boldsymbol{g}\|_q \,,$$

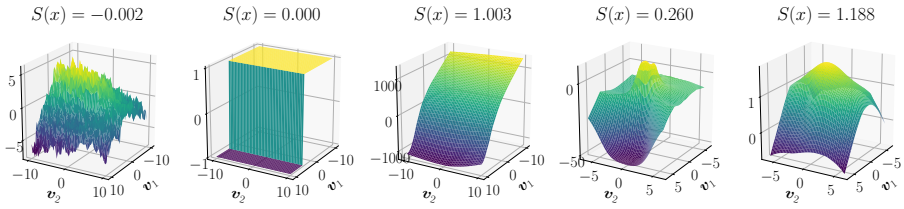
where $\|\cdot\|_q$ is the dual norm of $\|\cdot\|_p$. This equation tells us that the loss increment depends on the dual norm of the loss gradient and the step size η, as the optimal perturbation $\boldsymbol{\delta}^\star$ is found by aligning $\boldsymbol{\delta}$ with the loss gradient $\boldsymbol{g}$ depending on the given norm; e.g., $\boldsymbol{\delta}^\star = \eta \operatorname{sign}(\boldsymbol{g})$ when $p = \infty$, and $\boldsymbol{\delta}^\star = \eta \boldsymbol{g} / ||\boldsymbol{g}||_2$ when $p = 2$.

Based on the aforementioned first-order approximation, we define the *Slope* metric S evaluated at $\boldsymbol{x}$ as the ratio between the estimated loss increment and the actual one:

$$S(\boldsymbol{x}) = \frac{\eta \|\boldsymbol{g}\|_q}{\mathcal{L}(\boldsymbol{x} + \boldsymbol{\delta}, y; \boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{x}, y; \boldsymbol{\theta})}.$$

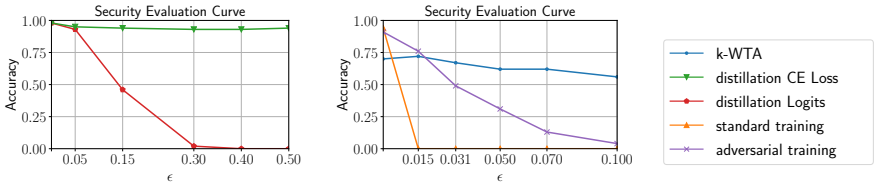
The ideal scenario would be $S(\boldsymbol{x}) \geq 1$, where the approximation is consistent with the actual loss increment. If $0 < S(\boldsymbol{x}) < 1$, the approximation is following a good direction, but it could be improved, since the actual loss increment is higher than the computed approximation. However, if $S(\boldsymbol{x}) \leq 0$ means that either the loss can not be increased, or the step is performed in the opposite direction, thus resulting in a decrease in the attacker loss. Such is the case when the attacker should rethink or debug their strategy, since it is not being correctly optimized. The Slope metric is calculated sample-wise, and we can also compute the mean value over many observations $\boldsymbol{x}_0, ..., \boldsymbol{x}_n$ as $\bar{S} = \frac{1}{n} \sum_{i=1}^{n} S(\boldsymbol{x}_i)$. This average acts

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.



(a) (b) (c) (d) (e)

Fig. 1: Loss landscape visualizations on the bi-dimensional space spanned by the ℓ_∞ adversarial direction ($\boldsymbol{v}_1$) and a random direction ($\boldsymbol{v}_2$) for the models used in our experiments: (a) K-WTA, (b) Distillation with Cross Entropy loss, (c) Distillation with logits loss, (d) Standard model, and (e) Adversarial training. Note that $S(\boldsymbol{x}) \leq 0$ only for models (a) and (b), which present obfuscated gradients.



(a) MNIST models. (b) CIFAR models.

Fig. 2: Security evaluations of the four models. On the x axis, the values for the perturbation budget ϵ used for the evaluation, while on the y axis the robust accuracy.

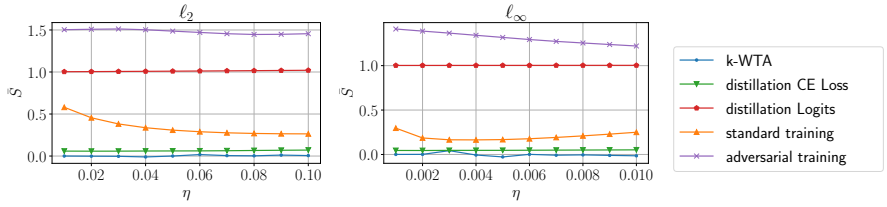
as a global measure for the ease of increasing the attacker loss against a particular model.

3 Experimental evaluation

We now show how our metric Slope is correlated with the ease or difficulty of optimizing the loss of an ℓ_p PGD attack, against a target model.

Target models We select four models, two of them are trained using particular techniques that are known to cause gradient obfuscation [8], one is a standard-trained model, and the last one is an adversarially-trained one.
k-Winners Take All (k-WTA) [6]: the network is trained to forward the output of the k most active neurons for each layer, producing piece-wise constant regions interleaved with discontinuities in the loss surface. Hence, by keeping k low, we ask for sparse responses that leads to a noisy landscape that abruptly changes around samples, as shown in Fig. 1a. This phenomenon affects the gradients of the model, that are characterized by high-variability in both directions and norms.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.



(a) (b)

Fig. 3: The values of the mean Slope $\bar{S}$, computed for both ℓ_2 (a) and ℓ_∞ (b). On the x axis, the step sizes η used for computing the approximation, while on the y axis we report the values of $\bar{S}$.

This model has to been shown not to be robust against adaptive attacks [8]. For this paper, we take the original model trained on CIFAR10.
Distillation [4]: the model is shown to have zero gradients in most of the loss landscape, except in correspondence of samples, where the scores soar and often cause numerical instability on the outputs [9]. Intuitively, the model is approximating a step function around each training sample, as shown in Fig. 1b. For this reason, the attack proposed by Athalaye et al. [7] discards the last softmax layer, using only the logits as output of the classifier. This trick smooths the loss function, as shown in Fig. 1c, successfully removing the effect of the defense. We test such model by training it on MNIST, following the original evaluation [4, 9] and computing the gradients of the Cross Entropy loss (obfuscated) and on the pre-softmax scores (not obfuscated).
Standard training: we use the implementation provided by RobustBench [10], that is a WideResNet [11] model trained on CIFAR10. Since it is not trained with any defenses, the loss landscape is smooth, but it contains many local minima and maxima, as shown in Fig. 1d. Hence, all classes are very close one to the other, and adversarial examples are very easy to be found.
Adversarial training [5]: the model is trained with both normal and adversarial examples, computed with an attack of choice, and such process is repeated until a desired robust accuracy is reached. As a result, the loss landscape becomes smoother, reducing the blind-spot areas where adversarial examples lay, as shown in Fig. 1e. For our work, we use the ResNet [12] model trained by Madry et al. [5] on CIFAR10.

Results We report the results of our experimental analysis, by taking into account the correlation between our metric $\bar{S}$ and the robust accuracy of each target. For computing $\bar{S}$, we use $\eta \in [0.01, 0.1]$ when using the ℓ_2 norm (Fig. 3a), and $\eta \in [0.001, 0.01]$ when using the ℓ_∞ norm (Fig. 3b). The security evaluation is computed by attacking all the targets with PGD ℓ_∞. For the model trained on CIFAR10, we set the maximum perturbation $\epsilon \in [0, 0.1]$ (Fig. 2b),where 0 implies the accuracy in absence of the attack, while we use $\epsilon \in [0, 0.5]$ for the attacks against the models trained on MNIST (Fig.2a).

By looking at the output of $\bar{S}$, we notice that models trained with obfuscated

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

gradients are characterized by values that are less than or equal to zero, implying a difficulty in detecting the right direction to follow during the attack. For instance, the Distillation model with Cross Entropy loss has null gradients, while k-WTA is characterized by noisy gradients that lead the loss to a decrement rather than an increment.

The other models' scores are positive, meaning that their gradients are informative for the attack and aligned with the loss to increase, hence converging to adversarial points. These effects are very similar for both ℓ_2 and ℓ_∞ norms, implying that such trend is characteristic of the model itself rather than the norm used for computing one step of PGD.

These results are confirmed by the security evaluations in Fig. 2: all the defended obfuscated models are less stressed by the adversarial attacks performed with PGD. The fact that the robust accuracy does not fall even with high perturbation budget, i.e. in the rightmost part of the security evaluation plots, confirms the hypothesis that the attack strategy is not suitable for those models. Also, the smoother models are affected by our attacks when the perturbation budges increases, as also predicted by $\bar{S}$. Hence, Slope is a good proxy for detecting the presence of gradient obfuscation inside the model, allowing the attacker to rethink their strategy, and landing successful evasion attacks.

4 Conclusions

We propose a metric for detecting the presence of gradients obfuscation, matching them with the ease of increasing the loss of the attack with PGD. Such metric is based on the intuition that obfuscated gradients can not be approximated during a step of PGD attacks, hence the computed loss is not representative of the real loss of the model. We test the metric against four models, where two of them are defended with obfuscated gradients, showing that the output of our metric is correlated with inability of decreasing the loss of the obfuscated ones. Hence, our metric can be used as a tool for detecting such defense, helping the attacker devising a strategy against the target.

As future work, we plan to create more metrics targeted on other defenses, to give the attacker a comprehensive suit of tools for detecting them.

Acknowledgments This work has been partly supported by the PRIN 2017 project RexLearn (grant no. 2017TWNMH2), funded by the Italian Ministry of Education, University and Research; and by BMK, BMDW, and the Province of Upper Austria in the frame of the COMET Programme managed by FFG in the COMET Module S3AI.

References

[1] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

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

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy*, pages 39–57. IEEE, 2017.

[4] Nicholas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597, May 2016.

[5] 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.

[6] Chang Xiao, Peilin Zhong, and Changxi Zheng. Enhancing adversarial defense by k-winners-take-all. In *International Conference on Learning Representations*, 2019.

[7] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pages 274–283. PMLR, 2018.

[8] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 33, 2020.

[9] Nicholas Carlini and David Wagner. Defensive distillation is not robust to adversarial examples. *arXiv preprint arXiv:1607.04311*, 2016.

[10] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

[11] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.