Black-Box Decision based Adversarial Attack with Symmetric α-stable Distribution

Vignesh Srinivasan
Machine Learning Group
Fraunhofer Heinrich Hertz Institute
Berlin, Germany
vignesh.srinivasan@hhi.fraunhofer.de

Ercan E. Kuruoglu
Institute of Information Science and Technologies
Italian National Research Council (CNR)
Pisa, Italy
ercan.kuruoglu@isti.cnr.it

Klaus-Robert Müller
Machine Learning Group
Technische Universität Berlin
Berlin, Germany
klaus-robert.mueller@tu-berlin.de

Wojciech Samek
Machine Learning Group
Fraunhofer Heinrich Hertz Institute
Berlin, Germany
wojciech.samek@hhi.fraunhofer.de

Shinichi Nakajima
Machine Learning Group
Technische Universität Berlin
Berlin, Germany
nakajima@tu-berlin.de

***Abstract*—Developing techniques for adversarial attack and defense is an important research field for establishing reliable machine learning and its applications. Many existing methods employ Gaussian random variables for exploring the data space to find the most adversarial (for attacking) or least adversarial (for defense) point. However, the Gaussian distribution is not necessarily the optimal choice when the exploration is required to follow the complicated structure that most real-world data distributions exhibit. In this paper, we investigate how statistics of random variables affect such random walk exploration. Specifically, we generalize the *Boundary Attack*, a state-of-the-art black-box decision based attacking strategy, and propose the *Lévy-Attack*, where the random walk is driven by symmetric α-stable random variables. Our experiments on MNIST and CIFAR10 datasets show that the Lévy-Attack explores the image data space more efficiently, and significantly improves the performance. Our results also give an insight into the recently found fact in the whitebox attacking scenario that the choice of the norm for measuring the amplitude of the adversarial patterns is essential.**

***Index Terms*—adversarial attack, α-stable distribution, deep neural networks, image classification.**

I. Introduction

The success of deep neural networks (DNNs) [1]–[5] has led to them being used in many real world applications. However, these models are also known to be susceptible to adversarial attacks, i.e., minimal patterns crafted by attackers who try to fool learning machines [6]–[11]. Such adversarial patterns do not affect human perception much, while they can manipulate learning machines, e.g., to give wrong classification outputs. DNN's complex interactions between different layers enable high accuracy under the controlled setting, while they make the outputs unpredictable in *untrained spots* where training samples exist sparsely. If attackers can find such a spot close to a normal data sample, they can manipulate DNNs by adding a very small (optimally invisible in computer vision applications) perturbation to the original sample, leading to fatal errors, e.g., manipulating an autonomous driving system can cause serious accidents.

Two attacking scenarios are considered in general—whitebox and blackbox. The whitebox scenario assumes that the attacker has access to the complete target system, including the architecture and the weights of the DNN, as well as the defense strategy if the system is equipped with any. Typical whitebox attacks optimize the classification output with respect to the input by backpropagating through the defended classifier [12]–[15]. On the other hand, the blackbox scenario assumes that the attacker has only access to the output. Under this scenario, the attacker has to rely on blackbox optimization, where the objective can be computed for arbitrary inputs, but the gradient information is not directly accessible. Although the whitebox attack is more powerful, it is much less likely that attackers can get full knowledge of the target system in reality. Accordingly, the blackbox scenario is considered to be a more realistic threat.

Existing blackbox attacks can be classified into two types—the transfer attack and the decision based attack. In the transfer attack, the attacker trains a student network which mimics the output of the target classifier. The trained student network is then used to get the gradient information for optimizing the adversarial input. In the decision based attack, the attacker simply performs random walk exploration. In the *boundary attack* [16], a state-of-the-art method in this category, the attacker first generates an initial adversarial sample from a

This work was supported by the German Ministry for Education and Research as Berlin Big Data Center BBDC (funding mark 01IS18025A) and Berlin Center for Machine Learning BZML (funding mark 01IS18037I). The work of K.-R. Müller was supported by the Institute for Information and Communications Technology Promotion Grant funded by the Korea government (MSIT) (No. 2017-00451, No. 2017-0-01779).

E.E.Kuruoglu's stay in Fraunhofer HHI was funded by CNR Short Term Mobility Program.

K.-R. Müller is also with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Korea, and with the Max Planck Institute for Informatics, 66123 Saarbrücken, Germany.

Shinichi Nakajima is also with RIKEN Center for AIP, Tokyo, Japan.

given original sample by drawing a uniformly distributed random pattern multiple times until it happens to lead to misclassification. Initial patterns generated in this way typically have too large amplitudes to be hidden from human perception. The attacker therefore polishes the initial adversarial pattern by Gaussian random walk in order to minimize the amplitude, keeping the classification output constant.[1]

Here our question arises. Is the Gaussian appropriate to drive the adversarial pattern to minimize the amplitude? It could be a reasonable choice if we only consider that the attacker minimizes the L2 norm of the adversarial pattern. However, it is also required to keep the classification output constant through the whole random walk sequence. Provided that the decision boundary of the classifier has complicated structure, reflecting the real-world data distribution, we expect that more efficient random walk can exist.

In this paper, we pursue this possibility, and investigate how statistics of random variables affect the performance of attacking strategies. To this end, we generalize the boundary attack, and propose the Lévy-Attack where the random walk exploration is driven by symmetric α-stable random variables. We expect that the impulsive characteristic of the α-stable distribution induces sparsity in random walk steps, which would drive adversarial patterns along the complicated decision boundary structure efficiently. Naturally, our expectation is reasonable only if the decision boundary has some structure aligned to the coordinate system defined in the data space, so that moving along the canonical direction keep more likely the classification output than moving isotropic directions.

In our experiments on MNIST and CIFAR10 datasets, Lévy-Attack with $\alpha \sim 1.0$ or less shows significantly better performance than the original boundary attack with Gaussian random walk. This implies that our hypothesis on the decision boundary holds at least in those two popular image benchmark datasets. Our results also give an insight into the recently found fact in the whitebox attacking scenario that the choice of the norm for measuring the amplitude of the adversarial patterns is essential.

II. Proposed Method

In this section, we first introduce the α-stable distribution, and propose the Lévy-Attack as a generalization of the boundary attack.

A. Symmetric α-stable Distribution

The symmetric α-stable distribution is a generalization of the Gaussian distribution which can model characteristics too impulsive for the Gaussian model. This family of distributions is most conveniently defined by their characteristic functions [17] due to the lack of an analytical expression for the probability density function. The characteristic function is given as

$$\phi(s) = \exp[i\mu s - |\gamma s|^{\alpha}], \tag{1}$$

[1]In the case of the untargeted attack, the classification output is kept *wrong*, i.e., random walk can go through the areas of any label except the true one.

Algorithm 1 (Untargeted) Lévy-Attack

Input: Classifier $\boldsymbol{f}(\cdot)$, original image $\boldsymbol{x}$ and label $\boldsymbol{y}$
Max. number T of iterations, termination threshold ψ
Output: Adversarial sample $\boldsymbol{x}^-$

repeat
  $\boldsymbol{x}_0^- \leftarrow \boldsymbol{x} + \boldsymbol{\Delta}$ for $\boldsymbol{\Delta} \sim \mathcal{U}_D(0, 255)$
until $\boldsymbol{y} \neq \boldsymbol{f}(\boldsymbol{x}_0^-)$
for $t = 0$ to $T - 1$ **do**
  $(\boldsymbol{x}_{t+1}^-, \epsilon) \leftarrow \alpha$-*stable random update*$(\boldsymbol{x}_t^-)$
  if $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}_0^-)$ **then**
    $\boldsymbol{x}_{t+1}^- \leftarrow \boldsymbol{x}_t^-$
  end if
  if $\epsilon < \psi$ **then**
    break
  end if
end for

where $\alpha \in (0, 2], \mu \in (-\infty, \infty)$, and $\gamma \in (0, \infty)$ are parameters. We denote the D-dimensional symmetric α-stable distribution by $\mathcal{SA}_D(\alpha, \mu, \gamma)$. α is the characteristic exponent expressing the degree of *impulsiveness* of the distribution—the smaller α is, the more impulsive the distribution is. The symmetric α-stable distribution reduces to the Gaussian distribution for $\alpha = 2$, and to the Cauchy distribution for $\alpha = 1$, respectively. μ is the location parameter, which corresponds to the mean in the Gaussian case, while γ is the scale parameter measuring of the spread of the samples around the mean, which corresponds to the variance in the Gaussian case. For more details on α-stable distributions, readers are referred to [17].

B. Lévy-Attack

Now, we propose our Lévy-Attack as a generalization of the boundary attack [16], a simple yet effective attack under the blackbox scenario, where the attacker has access only to the classification output. We denote the classifier output by $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$, where $\boldsymbol{x}$ is a data point, and $\boldsymbol{f}$ is the target (blackbox) classifier. The Lévy-Attack performs the procedure as described in Algorithm 1, which reduces to the original boundary attack if we set $\alpha = 2$.

While the attack is very simple, it can be a very effective blackbox attack [16]. Naturally, the success of the Lévy-Attack relies on the effectiveness of the proposal distribution. In the proposal distribution we accommodate sampling from a symmetric α-stable distribution $\boldsymbol{\eta}_t \sim \mathcal{SA}_D(\alpha, 0, 1)$. $\|\boldsymbol{\eta}_t\|_2 = \delta \cdot d(\boldsymbol{x}_t^-, \boldsymbol{x})$, where $d(\boldsymbol{x}_t^-, \boldsymbol{x}) = \|\boldsymbol{x}_{t+1}^- - \boldsymbol{x}\|_2^2$ and δ is the relative size of the perturbation. An orthogonal step is taken, where $\boldsymbol{\eta}_t$ is projected onto a sphere around the original image such that $d(\boldsymbol{x}_t^-, \boldsymbol{x}) = d(\boldsymbol{x}_{t+1}^-, \boldsymbol{x})$. Finally, a step is taken towards the original image so that the adversarial perturbation is reduced by a small amount ϵ, $d(\boldsymbol{x}_t^-, \boldsymbol{x}) - d(\boldsymbol{x}_{t+1}^-, \boldsymbol{x}) = \epsilon \cdot d(\boldsymbol{x}_t^-, \boldsymbol{x})$.

δ and ϵ are the two hyper-parameters which are dynamically tuned to adjust to the local geometry of the decision boundary. The orthogonal step in the proposal distribution encourages

TABLE I

THE MEAN $\mathcal{S}_m$ AND THE MEDIAN $\mathcal{S}_d$ OF THE L_∞, L_1, AND L_2 NORMS OF THE ADVERSARIAL PATTERNS GENERATED BY LÉVY-ATTACK FOR THE MNIST DATASET.

Attack	L_∞		L_1		L_2	
	$\mathcal{S}_m$	$\mathcal{S}_d$	$\mathcal{S}_m$	$\mathcal{S}_d$	$\mathcal{S}_m$	$\mathcal{S}_d$
Gaussian	**0.56**	**0.56**	11.36	10.73	1.38	1.39
$\alpha = 1.5$	0.57	0.58	9.62	9.16	1.31	1.31
$\alpha = 1.0$	0.57	0.58	8.89	**8.54**	**1.29**	**1.30**
$\alpha = 0.5$	0.58	0.58	**8.84**	8.71	1.30	1.32

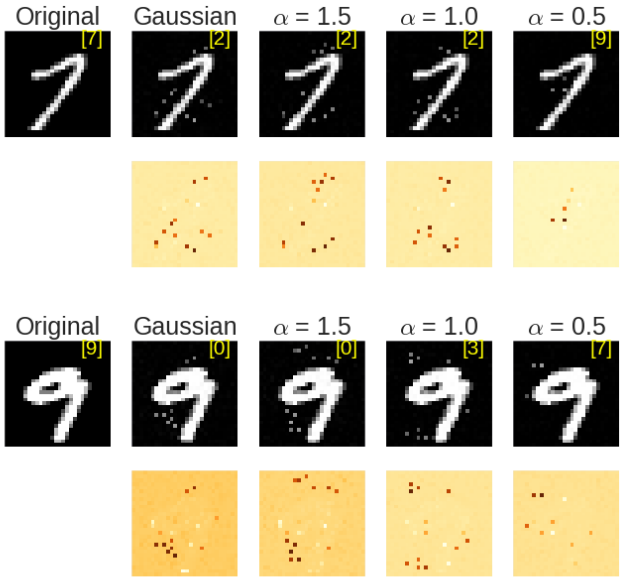

Fig. 1. Adversarial samples generated by Lévy-Attack on MNIST dataset for "7" and "9". "Gaussian" corresponds to $\alpha = 2$, with which Lévy-Attack reduces to the original boundary attack [16]. The classification output is shown at the top right corner of each image. In each block (for "7" as well as "9"), the top row displays adversarial samples generated with different α, while the bottom row displays the corresponding adversarial patterns (the differences from the original image).

around 50% orthogonal perturbations to be adversarial. The length of the step, ϵ is adjusted according to the success rate of the step. If the success rate is too high, then ϵ is increased and vice versa. As the random walk moves closer towards the original image, the success rate of the attack becomes lesser. The attack gives up further exploration as ϵ converges to 0.

III. EXPERIMENT

We report on experiments performed using our Lévy-Attack on the following datasets:

- MNIST: The MNIST dataset consists of $60,000$ images in total, with $50,000$ images for training and $10,000$ images for testing. It has 10 different classes each corresponding to the 10 numerical digits. The image size is 28×28.
- CIFAR10: This dataset also contains $50,000$ training images and $10,000$ test images. The images are of resolution $32 \times 32 \times 3$ with 10 different classes in total.

TABLE II

THE MEAN $\mathcal{S}_m$ AND THE MEDIAN $\mathcal{S}_d$ OF THE L_∞, L_1, AND L_2 NORMS OF THE ADVERSARIAL PATTERNS GENERATED BY LÉVY-ATTACK FOR THE CIFAR10 DATASET.

Attack	L_∞		L_1		L_2	
	$\mathcal{S}_m$	$\mathcal{S}_d$	$\mathcal{S}_m$	$\mathcal{S}_d$	$\mathcal{S}_m$	$\mathcal{S}_d$
Gaussian	**2.92**	2.47	895.22	755.06	23.72	20.45
$\alpha = 1.5$	2.99	2.44	859.49	708.54	23.15	19.49
$\alpha = 1.0$	2.97	2.427	847.20	700.42	23.06	19.39
$\alpha = 0.5$	2.94	**2.421**	**826.29**	**685.76**	**22.78**	**19.28**

TABLE III

THE AVERAGE NUMBER OF ITERATIONS THAT LÉVY-ATTACK PERFORMED TO GENERATE ADVERSARIAL SAMPLES.

Attack	MNIST	CIFAR10
Gaussian	2700.22	4996.49
$\alpha = 1.5$	2629.04	4995.96
$\alpha = 1.0$	2792.52	4987.04
$\alpha = 0.5$	3407.54	4997.37

In the MNIST experiment, we target the state-of-the-art robust classifier proposed by Madry et al. [18],[2] where the classifier is trained, in addition to the original training set, on the adversarial samples generated by the Projected Gradient Descent (PGD) attack of bounded L_∞ distortion by 0.3. The classification accuracy on the original test samples is 98.68%. In the CIFAR10 experiment, we trained a state-of-the-art Resnet model [5]. The classification accuracy on the original test samples is 92.93%.

To generate samples by Lévy-Attack, we modify the code-provided by [16] for the boundary attack, so that the random walk is performed by the symmetric α-stable distribution, instead of the Gaussian distribution. We evaluated adversarial samples for $\alpha = 2.0, 1.5, 1.0$, and 0.5. The other parameters specifying the α-stable distribution is set to $\delta = 0.0$ and $\gamma = 1.0$. We limit the number of random walk steps to $5,000$. Having such an upper-bound is reasonable because it is not realistic to assume that the attacker may access to the classifier output unlimited times.

For both datasets, we randomly sample $N = 1,000$ images from the test set, and evaluate the quality of the adversarial patterns. As evaluation scores, we use the mean and the median of 3 different L_p-norms for $p = \infty, 1$, and 2, over the 1,000 samples:

$$\mathcal{S}_m = \frac{1}{N}\sum_{i=1}^{N}(\|\boldsymbol{\tau}_i\|_p), \qquad \mathcal{S}_d = median_{i=1}^{N}(\|\boldsymbol{\tau}_i\|_p),$$

where $\{\boldsymbol{\tau}_i\}$ are the adversarial patterns. Smaller norms indicate that the adversarial pattern is less visible, and therefore a better attack.

Table I shows the results on the MNIST dataset, where we see that the Lévy-Attack with α smaller than 2 (Gaussian) gives significantly smaller L_1 and L_2 norms with the L_∞ norm almost unchanged. Similar results are obtained on the

[2]https://github.com/MadryLab/mnist_challenge

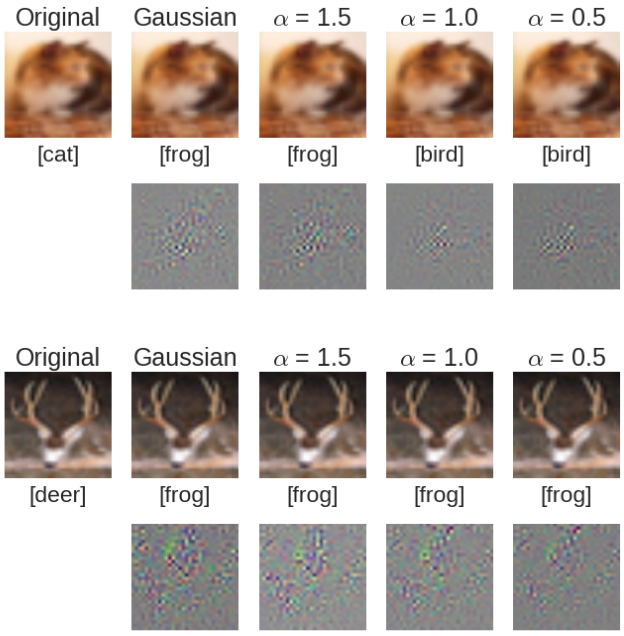

Fig. 2. Adversarial samples generated by Lévy-Attack on CIFAR10 dataset for "cat" and "deer".

CIFAR10 dataset (Table II), where $\alpha < 2$ gives better L_1 and L_2 norms with the L_∞ norm almost unchanged, although the performance difference is smaller than the MNIST results.

Table III summarizes the average number of iterations the Lévy-Attack performs. We see the tendency that smaller α leads to more iterations, which implies that α-stable random walk continues exploring when Gaussian random walk has already been terminated due to a low success rate in further adversarial exploration. Also, Tables I and II show that the Lévy-Attack for $\alpha < 2$ reaches to the point closer to the original than the Gaussian ($\alpha = 2$) random walk. These results imply that the impulsive random walk is suitable to explore the data space without crossing decision boundaries, and indirectly support our hypothesis in Section I—decision boundaries have some structure aligned to the coordinate system.

Figs. 1 and 2 show a few illustrative examples of adversarial samples and adversarial patterns generated by Lévy-Attack. For each block for the examples ("7" and "9" in MINST, and "cat" and "deer" in CIFAR10), the top row shows the generated adversarial samples, while the bottom row shows the corresponding adversarial patterns (the differences from the original image). In the "7" example of MNIST (Fig. 1), Lévy-Attack with $\alpha \geq 1.0$ consistently tries to modify the sample close to "2", while Lévy-Attack for $\alpha = 0.5$ tries to modify the sample close to "7". Apparently, the latter is more efficient, i.e., it requires fewer pixels to make "7" to "9" than to make "7" to "2". The same applies to the "9" example—it seems more efficient to make "9" close to "7" than to make "9" to "0" or "3". However, only Lévy-Attack with a very small α can find those efficient solution, because it is little likely to get a sparse random walk step if it is driven by non-sparse distributions like Gaussian. The CIFAR10 examples, although less obvious than the MNIST examples, also show similar tendency—the α-stable random walk with smaller α provides sparser adversarial patterns.

IV. Discussion

Many defense strategies have been proposed to counter adversarial attacks [18]–[20]. However, it happened many times that a new defense strategy is broken down by a newer attacking strategy only a few months after its proposal. Thus, the adversarial defense problem has not been solved even on the toy MNIST data set, although defense is considered much harder for larger data sets.

One recent finding in this ensuing arms race between new defense and attacking strategies is the importance of the metric of the distortion, i.e., how to measure the distance from the original sample. In whitebox attacks, the L_∞, L_1, and L_2 norms are often used to measure the distortion [6], [12], [13], [18]. Interestingly, the state-of-the-art defense method proposed by Madry et al. [18] has shown to be robust against attacks with L_∞ bounded perturbations, while it has been found to be vulnerable against attacks with the elastic net (L_1 plus L_2) bounded perturbations [13], [14].

Naturally, the choice of the perturbation metric impacts the human perception, e.g., limiting the L_2 norm makes the visual quality of the image better while limiting the L_1 norm assures sparsity of the distortion. However, the finding above implies that sparser regularization might help gradient-based whitebox optimization find stronger adversarial samples. Our results in this paper might imply something similar or at least related—sparser random walk steps help exploration move along the decision boundaries, and produce stronger adversarial samples under the blackbox scenario. Further investigation is left as future work.

V. Conclusion

In this paper, we investigated how statistics of random variables affect random walk based blackbox attacking strategy. Specifically, we proposed Lévy-Attack, a generalization of the state-of-the-art boundary attack, where random walk is driven by symmetric α-stable random variables. Our experiments showed that the impulsive characteristics of the α-stable distribution enables efficient exploration in the data space without crossing decision boundaries, producing stronger adversarial samples. In our future work, we investigate the use of explanation methods [21]–[24] for adversarial attack detection and further study the relation between norm bounds, sparse exploration, and the quality of adversarial samples.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1–9.

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

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[7] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the ACM Asia Conference on Computer and Communications Security (ASIACSS)*, 2017, pp. 506–519.

[8] J. Bruna, C. Szegedy, I. Sutskever, I. Goodfellow, W. Zaremba, R. Fergus, and D. Erhan, "Intriguing properties of neural networks," 2013.

[9] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 427–436.

[10] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song, "Robust physical-world attacks on machine learning models," *arXiv preprint arXiv:1707.08945*, 2017.

[11] A. Athalye and I. Sutskever, "Synthesizing robust adversarial examples," *arXiv preprint arXiv:1707.07397*, 2017.

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

[13] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C.-J. Hsieh, "Ead: elastic-net attacks to deep neural networks via adversarial examples," *arXiv preprint arXiv:1709.04114*, 2017.

[14] Y. Sharma and P.-Y. Chen, "Breaking the madry defense model with l_1-based adversarial examples," *arXiv preprint arXiv:1710.10733*, 2017.

[15] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2574–2582.

[16] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," *arXiv preprint arXiv:1712.04248*, 2017.

[17] G. Samorodnitsky and M. S. Taqqu, *Stable non-Gaussian random processes: stochastic models with infinite variance*. London; New York: Chapman Hall Ltd, 1994.

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[19] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-gan: Protecting classifiers against adversarial attacks using generative models," in *International Conference on Learning Representations*, vol. 9, 2018.

[20] V. Srinivasan, A. Marban, K.-R. Müller, W. Samek, and S. Nakajima, "Counterstrike: Defending deep learning architectures against adversarial samples by langevin dynamics with supervised denoising autoencoder," *arXiv preprint arXiv:1805.12017*, 2018.

[21] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[22] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[23] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services*, vol. 1, no. 1, pp. 39–48, 2018.

[24] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, p. 1096, 2019. [Online]. Available: http://dx.doi.org/10.1038/s41467-019-08987-4