Adversarial Attacks on Deep Learning Models of Computer Vision: A Survey ⋆

Jia Ding and Zhiwu Xu*

College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
Email: 2269242754@qq.com, xuzhiwu@szu.edu.cn

Abstract. Deep learning plays a significant role in academic and commercial fields. However, deep neural networks are vulnerable to *adversarial attacks*, which limits its applications in safety-critical areas, such as autonomous driving, surveillance, and drones and robotics. Due to the rapid development of adversarial examples in computer vision, many novel and interesting adversarial attacks are not covered by existing surveys and could not be categorized according to existing taxonomies. In this paper, we present an improved taxonomy for adversarial attacks, which subsumes existing taxonomies, and investigate and summarize the latest attacks in computer vision comprehensively with respect to the improved taxonomy. Finally, We also discuss some potential research directions.

Keywords: Deep Learning· Adversarial Attacks· Black-box Attack· White-box Attack· Machine Learning.

1 Introduction

As a major branch of machine learning, deep learning [23] has always been a popular research direction in the artificial intelligence community. It can solve the classification problems that are difficult or even impossible to solve in a relatively short time, and has many applications in academic and commercial fields, such as computer vision [38, 1], speech recognition [48, 44], natural language processing [41], malware detection [18], autonomous vehicles [36], network security [17], surveillance [31], drones and robotics [28, 11], and so on. Moreover, deep learning has become the preferred choice in computer vision, which plays a major role in our daily lives, after Krizhevsky et al.'s work in 2012 [20]. Thus we focus on computer vision in this paper.

Although deep learning can perform a wide variety of hard tasks with remarkable accuracies, especially in computer vision, Szegedy et al. [42] discovered that the robustness of neural networks encounters a major challenge when adding imperceptible non-random perturbation to input in the context of image classification. They firstly defined the perturbed examples with the ability

⋆ * Corresponding author.

to misclassify the classifiers as *adversarial examples* and the imperceptible perturbations to images as *adversarial attacks* [42]. This phenomenon implies that deep neural networks are vulnerable to *adversarial attacks*, which limits its applications in safety-critical areas, such as autonomous driving, surveillance, and drones and robotics, and could cause huge economic losses.

After the findings of Szegedy et al., lots of researchers realize the importance of adversarial examples for neural networks, as they are essential to the robustness of the neural networks in some sense. As a result, adversarial examples have become a hot research field in recent years, and many approaches for generating adversarial examples have been proposed. And there is also some review work [9, 38, 1, 49], which gives a comprehensive survey on adversarial attacks in computer vision of that time. However, due to the rapid development of adversarial examples, many novel and interesting approaches for adversarial attacks have been proposed recently, which are not covered by existing review work. Moreover, some adversarial attacks are hard to categorize with respect to existing attack taxonomies presented in [38, 49], which indicates existing taxonomies may not be suitable for the latest attacks.

This paper aims to briefly review the latest interesting attacks in computer vision, and revise existing taxonomies for adversarial attacks. More specifically, we first present an improved taxonomy for adversarial attacks, which combines existing taxonomies in [38, 49] with a brand-new category, namely, functional-based attacks [22]. Then we explore different approaches, including the classic ones and the latest ones, for generating adversarial examples by taking advantage of the attributes of adversarial examples, such as transferability, and the attributes of images, such as geometric transformation invariance. Finally, in light of the development of adversarial attacks, we also discuss some potential research directions.

Our main contributions are summarized as follows:

- An improved taxonomy for adversarial attacks is presented, including a brand-new category that has never been mentioned in previous work.
- The state-of-the-art approaches for generating adversarial examples are explored, according to the improved taxonomy.

The remainder of this paper will be organized as follows. We introduce some definitions of terms in Section 2 and give some related work in Section 3. In Section 4, we present the improved taxonomy and review the classic attacks and the state-of-the-art attacks. In Section 5, we discuss the potential research directions for researchers. We conclude this paper in Section 6.

2 Definitions of Terms

In this section, we introduce the technical terms used in the adversarial attacks literature.

Adversarial example: the input with small perturbations that can misclassify the classifier. In the application scene of computer vision, the image with carefully prepared perturbations noise that can make the classifier misclassification.

Adversarial perturbation: the noise data that is capable to change original images to adversarial examples.

Black-box attack: the attackers know nothing about the architecture, training parameters, and defense methods of the attacked model, and can only interact with the model through the input and output.

White-box attack: in contrast to the black-box attack, the attackers master everything about the model and the defense schemes should be public to attackers. At present, most attack approaches are white-box.

Gray-box attack: between black-box attack and white-box attack, only a part of the model is understood. For example, the attackers get only the output probability of the model, or know only the model structure without the parameters.

Untargeted attack: the attackers only need to make the target model misclassify, but do not specify which category is misclassified.

Targeted attack: the attackers specify a certain category, so that the target model not only misclassify the sample but also need to be misclassified into the specified category. It is more difficult to achieve targeted attacks than untargeted attacks.

Transferability: transferability refers to the effective adversarial examples for one model, and still effective for other models.

Geometric transformation invariance: the target in the image can be successfully identified whether it is translated, rotated, or zoomed, or even under different lighting conditions and viewing angles, such as translation invariance and scale invariance.

3 Related Work

This section presents some related work of surveys on adversarial attacks.

There are some review work [9, 38, 1, 49] on adversarial attacks in computer vision so far. Fawzi et al. [9] discussed the robustness of deep networks to a diverse set of perturbations that may affect the samples in practice, including adversarial perturbations, random noise, and geometric transformations. Serban et al. [38] provided a complete characterization of the phenomenon of adversarial examples, summarized more than 20 kinds of attacks at that time by dividing the attack approaches into four categories: (*i*) attacks based on optimization methods, (*ii*) attacks based on sensitive features, (*iii*) attacks based on geometric transformations, and (*iv*) attacks based on generative models. Akhtar et al. [1] reviewed the adversarial attacks at that time for the task of image classification and beyond classification, and introduced some adversarial attacks in the real world. Zhou et al. [49] summarized the latest attack approaches at that time and divided them into four categories: (*i*) gradient-based attack, (*ii*) score-based attack, (*iii*) transfer-based attack, and (*iv*) decision-based attack.

However, because of the popularity of adversarial attacks, after the latest review work [49], many novel and interesting approaches for adversarial attacks have been proposed recently. Moreover, some adversarial attacks are hard to categorize with respect to existing attack taxonomies presented in [38, 49], which

indicates existing taxonomies may not be suitable for the latest attacks. To supplement the latest development and revise existing taxonomies, in this paper we give an improved taxonomy, which subsumes existing taxonomies in [38, 49], and summarize the latest attack approaches according to the improved taxonomy.

In addition to computer vision, there are some surveys on adversarial attacks in other areas, such as (vector) graphs [3], speech recognition [14], autonomous driving [36], and malware detection [27]. Chen et al. [3] investigated and summarized the existing works on graph adversarial learning tasks systemically. Hu et al. [14] provided a concise overview of adversarial examples for speech recognition. Ren et al. [36] systematically studied the safety threats surrounding autonomous driving from the perspectives of perception, navigation and control. Martins et al. [27] explored applications of adversarial machine learning to intrusion and malware detection.

4 Adversarial Attacks

In this section, we first give an improved taxonomy for attack approaches, and then explore the attack approaches according to this taxonomy. As the classic attack approaches have been summarized in existing work [9, 38, 1, 49], we only give a brief review on these classic approaches in this section. In other words, we focus on the latest attack approaches (*i.e.*, those are not covered by existing surveys [9, 38, 1, 49]), each of which is marked with its abbreviated name in bold.

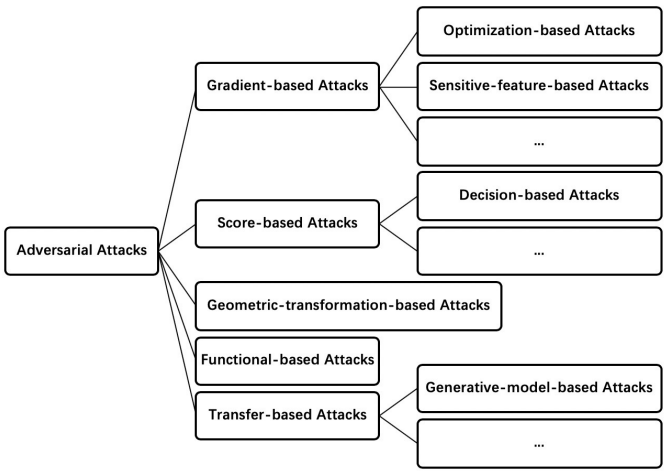

Fig. 1. The improved taxonomy for adversarial attacks

4.1 An Improved Taxonomy

There have been many interesting approaches of adversarial attacks in computer vision. To investigate and summarize them comprehensively, we present a taxonomy for them.

Based on existing taxonomies [38, 49], we propose an improved taxonomy for adversarial attacks, which is given in Figure 1. In detail, we classify the adversarial attacks into five categories, that is, (*i*) gradient-based attacks (GA) (from [49]), (*ii*) score-based attacks (SA) (from [49]), (*iii*) geometric-transformation-based attacks (GTA) (from [38]), (*iv*) functional-based attacks (FA) (a brand-new category), and (*v*) transfer-based attacks (TA) (from [49]). Moreover, we categorize optimization-based attacks and sensitive-feature-based attacks from [38] into gradient-based attacks, as both of them are almost based on gradient but with different objectives, and generative-model-based attacks from [38] into transfer-based attacks, as it is stated in [49] that generative-model-based attacks are a subclass of transfer-based attacks. The decision-based attacks from [49] are treated as a special case of score-based attacks, due to the fact that the decisions are always made according to the scores. In conclusion, our taxonomy subsumes both the taxonomies in [38] and [49].

According to our taxonomy, we summarize different attacks, including the classic ones and the latest ones (in bold), in Table 1. All these attacks are reviewed in the following.

4.2 Gradient-based Attacks

Gradient-based attacks perturb the images in the direction of the gradient, so that the model can be misclassified with the smallest perturbation. They are mainly white-box attacks.

We briefly introduce the classic gradient-based attacks. In 2014, Goodfellow et al. [12] proposed the Fast Gradient Sign Method (FGSM), applying small perturbations in the gradient direction to maximize the loss function to generate adversarial examples. Due to that the FGSM algorithm only involves a single gradient update and that a single update is sometimes not enough to attack successfully, Kurakin et al. [21] proposed the Iterative Fast Gradient Sign Method (I-FGSM) based on FGSM. After that, Madry et al. [26] proposed the Projected Gradient Descent (PGD), which is a more powerful gradient attack than I-FGSM and FGSM. It initializes the search adversarial examples at random points within the allowed norm ball, and then runs the Basic Iterative Method (BIM) multiple iterations. In 2018, Dong et al. [6] proposed the Momentum Iterative Fast Gradient Sign Method (MI-FGSM), which integrates momentum into the iterative attack and leads to a higher attack success rate and transfer-ability than other gradient-based methods for adversarial examples.

Besides attacks based on FGSM, there are some other gradient-based attacks. In 2016, Moosavi et al. proposed the DeepFool [29], which can generate adversarial examples that are very close to the minimum perturbation, so it can be used as a measure of the robustness of the classifier. Later, Moosavi-Dezfooli

Table 1. Catalog of Adversarial Attacks.

Attack	**Category**	**White-box/ Black-box**	**Targeted/ Untargeted**	**Specific/ Universal**	**Iterative/ One-shot**	**Year**
FGSM [12]	GA	White-box	Targeted	Image specific	One-shot	2015
I-FGSM [21]	GA	White-box	Targeted	Image specific	Iterative	2017
PGD [26]	GA	White-box	Untargeted	Image specific	Iterative	2018
MI-FGSM [6]	GA	White-box	Untargeted	Image specific	Iterative	2018
DeepFool [29]	GA	White-box	Untargeted	Image specific	Iterative	2016
UAP [30]	GA	White-box	Untargeted	Universal	Iterative	2017
C&W [2]	GA	White-box	Targeted	Image specific	Iterative	2017
LogBarrier [10]	GA	White-box	Untargeted	Image specific	Iterative	2019
NI-FGSM [24]	GA	White-box	Untargeted	Image specific	Iterative	2020
ZOO [4]	SA	Black-box	Untargeted	Image specific	Iterative	2017
OPA [39]	SA	Black-box	Untargeted	Image specific	Iterative	2019
AutoZOOM [43]	SA	Black-box	Untargeted	Image specific	Iterative	2019
CornerSearch [5]	SA	Black-box	Untargeted	Image specific	Iterative	2019
BayesOpt [37]	SA	Black-box	Untargeted	Image specific	Iterative	2020
ManiFool [19]	GTA	White-box	Untargeted	Image specific	Iterative	2018
Xiao et al. [46]	GTA	White-box	Untargeted	Image specific	Iterative	2018
DIM [47]	GTA	White-box	Untargeted	Image specific	Iterative	2019
TI [7]	GTA	White-box	Untargeted	Image specific	Iterative	2019
SIM [24]	GTA	White-box	Untargeted	Image specific	Iterative	2020
ReColorAdv [22]	FA	White-box	Untargeted	Image specific	One-shot	2019
Substitute [34]	TA	Black-box	Targeted	Image specific	Iterative	2017
Ensemble [25]	TA	Black-box	Targeted	Image specific	Iterative	2019
ILA [15]	TA	Black-box	Targeted	Image specific	Iterative	2019
TREMB [16]	TA	Black-box	Targeted	Image specific	Iterative	2020

et al. [30] proposed attacks using Universal Adversarial Perturbations (UAP), which is an extension of DeepFool. In 2017, Carlini et al. [2] proposed the Carlini & Wagner attack (C&W), which optimizes the distances from adversarial examples to benign examples and is an optimization-based method.

LogBarrier Attack. Recently, Finlay et al. [10] proposed a new type of adversarial attack (called LogBarrier attack) based on optimization. Different from existing approaches that use training loss functions to achieve misclassification, LogBarrier uses the best practices in the optimization literature to solve "how to make a benign sample transform into adversarial sample", wherein the well-known logarithmic barrier method [33] is used to design a new untargeted attack. The LogBarrier attack performs well on common data sets, and in images that require large perturbation for misclassification, LogBarrier attacks always have an advantage over other adversarial attacks.

NI-FGSM. Lin et al. [24] proposed the Nesterov Iterative Fast Gradient Sign Method (NI-FGSM)[1], which aims to adapt Nesterov accelerated gradient into the iterative attacks so as to effectively look ahead and improve the transferability of adversarial examples, according to the fact that Nesterov accelerated gradient method [32] is superior to the momentum for conventionally optimization method [40]. Experiments show that it can effectively improve the transferability of adversarial examples.

4.3 Score-based Attacks

Score-based attacks are black box approaches that rely only on predicted scores, such as category probability or logarithm. Conceptually, these attacks use numerically estimated gradient predictions.

In 2017, inspired by C&W attacks [2], Chen et al. [4] proposed the Zeroth Order Optimization (ZOO), which directly estimates the gradient of the target model to generate adversarial examples. In 2019, Vargas et al. [39] proposed the One Pixel Attack (OPA), wherein only one pixel can be modified at a time. As a result, less adversarial information is required. And thus it can deceive more types of networks, because of the inherent function of differential evolution.

AutoZOOM. To address the major drawback of existing black-box attacks, that is, the need for excessive model queries, Tu et al. [43] proposed a generic framework, the Autoencoder-based Zeroth Order Optimization Method (AutoZOOM), for query-efficient black-box attacks. AutoZOOM has two novel building blocks towards efficient black-box attacks: (*i*) an adaptive random gradient estimation strategy to balance query counts and distortion, and (*ii*) an autoencoder that is either trained offline with unlabeled data or a bilinear resizing operation for attack acceleration.

CornerSearch. Despite highly sparse adversarial attacks have a great impact on neural networks, the perturbations imposed on images by sparse attacks are easily noticeable due to their characteristics. To prevent the subtle perturbations of highly sparse adversarial attacks from being detected, Croce et al. [5]

[1] The other method SIM proposed by Lin et al. [24] is discussed in Section 4.4.

proposed a fresh black-box attack. They use locally adaptive component wise constraints to minimize the difference between the modified image and the original image, which enables to limit pixel perturbation to occur only in areas of high variance and to reduce the number of pixels that need to be modified. Experiments show that their score-based l_0-attack CornerSearch needs the least pixels to complete the task.

BayesOpt Attack. The existing black-box attacks are based on the substitute model, gradient estimationor genetic algorithm.The number of queries they need is usually very large. For projects that need to control costs in real life or items that have a limited number of query numbers, these approaches are obviously not applicable. Therefore, Ru et al. [37] proposed a new gradient-free black-box attack, which uses Bayesian Optimization (BayesOpt) in combination with Bayesian model selection to optimise over the adversarial perturbation and the optimal degree of search space dimension reduction. Experiments show that, in the constraint of l_∞-norm, BayesOpt adversarial attack can achieve a considerable success rate with a model query of about 2 to 5 times, compared with the latest black-box attacks.

4.4 Geometric-transformation-based Attacks

Geometric-transformation-based attacks transform targets in the images via geometric transformation (*e.g.*, rotating or zooming) to generate adversarial examples. According to geometric transformation invariance, no matter how geometrically transformed its input image is, the classifier for image classification tasks should produce the same output. Algorithms based on geometric transformation invariances often work together with algorithms based on gradients.

Engstrom et al. [8] showed that only simple transformations, namely rotations and translations, are sufficient to fool DNN. Kanbak et al. [19] proposed ManiFool, an approach to find small worst-case geometrical transformations of images. Xiao et al. [46] proposed to spatially transform the image, that is, to change the geometry of the scene, while keeping the original appearance.

DIM. Xie et al. [47] a Diverse Inputs Method (DIM) to improve the transferability of adversarial examples. Inspired by the data augmentation strategy [13], DIM randomly applies a set of label-preserving transformations (e.g., resizing, cropping and rotating) to training images and feeds the transformed images into the classifier for gradient calculation. DIM can be combined with the momentum-based method (such as MI-FGSM [6]) to further improve the transferability. By evaluating DIM against top defense solutions and official baselines from NIPS 2017 adversarial competition, the enhanced attack M-DI^2-FGSM reaches an average success rate of 73.0%, which outperforms the top-1 attack submission in the NIPS competition by a large margin of 6.6%.

TI. Dong et al. [7] proposed a Translation-Invariant (TI) attack method to generate more transferable adversarial examples against the defense models. TI optimizes the adversarial samples by using a set of translated images, making the adversarial samples less sensitive to the distinguished regions of the white-box model being attacked, and thus the transferability of the adversarial samples

is increased. To improve the efficiency of attacks, TI can be implemented by convolving the gradient at the untranslated image with a pre-defined kernel. Dong et al. [7] also showed that TI-DIM, the combination of TI and DIM [47], performed best on common data sets.

SIM. Besides NI-FGSM (see Section 4.2), Lin et al. [24] also proposed another method to improve the transferability of adversarial examples, that is, the Scale-Invariant attack Method (SIM). SIM utilizes the scale-invariant property of the model to achieve model augmentation and can generate adversarial samples which are more transferable than other black box attacks. Combining NI-FGSM and SIM, SINI-FGSM is a powerful attack with higher transferability. Experiments show that SINI-FGSM can break other powerful defense mechanisms.

4.5 Functional Adversarial Attacks

Unlike standard l_p-ball attacks, functional adversarial attacks allow only a single function, which is called *the perturbation function*, to be used to perturb input features to generate an adversarial example. Functional adversarial attacks are in some ways more restrictive because features cannot be perturbed individually.

ReColorAdv. Laidlaw et al. [22] proposed ReColorAdv, a functional adversarial attack on pixel colors. ReColorAdv generates adversarial examples to fool image classifiers by uniformly changing colors of an input image. More specifically, ReColorAdv uses a flexibly parameterized function f to map each pixel color c in the input to a new pixel color $f(c)$ in an adversarial example. Combining functional adversarial attacks with existing attacks that use the l_p-norm can greatly increase the attack capability and allow the model to modify the input locally, individually, and overall. Experiments show that the combination of ReColorAdv and other attacks (*e.g.*, Xiao et al.'s work [46]) can produce the strongest attack at present.

4.6 Transfer-based Attacks

Transfer-based attacks do not rely on model information, but need information about training data. This is a way to transition between black-box attacks and white-box ones.

In 2014, Szegedy et al. [42] firstly proposed the concept of adversarial examples, in the same time they observed that adversarial examples generated for one model can be effectively transferred to other models regardless of architecture, which is named by model-transferability. And later, Papernot et al. [35] explored deeply this property. In 2017, Papernot et al. [34] proposed a transfer-based attack, called Substitute in this paper, which trains a local model to substitute for the target DNN, using inputs synthetically generated by an adversary and labeled by the target DNN. In 2017, Liu et al. [25] proposed a novel strategy (called Ensemble here) to generate transferable adversarial images using an ensemble of multiple models, which enables a large portion of targeted adversarial examples to transfer among multiple models for the first time.

ILA. In order to enhance the transferability of black box attacks, Huang et al. [15] proposed the Intermediate Level Attack (ILA), which fine-tunes the existing adversarial examples and increases the perturbations on the pre-designated layer of the model to achieve high transferability. ILA is a framework with the goal of enhancing transferability by increasing projection onto the *Best Transfer Direction*. Two variants of ILA, namely ILAP and ILAF, are proposed in [15], differing in their definition of the loss function L.

TREMBA. Unlike previous attack methods, which training substitute models with data augmentation to mimic the behavior of the target model, Huang et al. [16] proposed a method called TRansferable EMbedding based Black-box Attack (TREMBA), which combines transfer-based attack and scored-based attack, wherein transfer-based attack is used to improve query efficiency, while scored-based attack is to increase success rate. TREMBA contains two steps: the first step is to train an encoder-decoder to generate adversarial perturbations for the source network with a low-dimensional embedding space, and the second step is to apply NES (Natural Evolution Strategy) [45] to the low-dimensional embedding space of the pretrained generator to search adversarial examples for the target network. Compared with other black box attacks, the success rate of TREMBA is increased by about 10%, and the number of queries is reduced by more than 50%.

5 Future Directions

In this section, we discuss some potential research directions.

Explanation. Lots of approaches can generate adversarial examples effectively and efficiently. But why classifier makes a misclassification on these adversarial examples? Few work gives a systematic study on this problem. Moreover, are these adversarial examples helpful to explain the classifiers, that is, can we extract some negative but useful knowledge from adversarial examples? This is also an interesting problem. On the other hand, as the geometric-transformation-based attacks, we could take advantage of the domain knowledge or the knowledge that is extracted from models if possible to guide the adversarial attacks to get more effective adversarial examples, via the perturbation functions [22]?

Ensemble. It has been shown in existing work [22, 7, 16] that some attacks can be combined together. Indeed, some categories of our taxonomy are orthogonal, such as gradient-based attack and geometric-transformation-based attack. Similar to ensemble learning, how to combine different attacks to get a more powerful attack is worth investigating.

Transferability. As stated in [38], transferability is inherent to models that "learn feature representations that preserve non-robust properties of the input space". If models were to learn different features, it would have not been possible to transfer adversarial examples. Therefore, when designing a model, it is better to consider transferability.

Robustness. Finally, one goal of adversarial examples is to improve the robustness of DNN. This is also a fundamental problem in the community and deserves special attention.

6 Conclusion

In this paper, we have presented an improved taxonomy for adversarial attacks in computer vision. Then according to this taxonomy, we have investigated and summarized different adversarial attacks, including the the classic ones and the latest ones. Through the investigation, some future directions are discussed. This paper is expected to provide guidance on adversarial attacks for researchers and engineers in computer vision and other areas.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China under Grants No. 61972260, 61772347, 61836005; Guangdong Basic and Applied Basic Research Foundation under Grant No. 2019A1515011577; and Guangdong Science and Technology Department under Grant No. 2018B010107004.

References

1. Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access **6**, 14410–14430 (2018)
2. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. IEEE (2017)
3. Chen, L., Li, J., Peng, J., Xie, T., Cao, Z., Xu, K., He, X., Zheng, Z.: A survey of adversarial learning on graphs. arXiv preprint arXiv:2003.05730 (2020)
4. Chen, P., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.: 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 (AISec@CCS 2017). pp. 15–26 (2017)
5. Croce, F., Hein, M.: Sparse and imperceivable adversarial attacks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4724–4732 (2019)
6. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9185–9193 (2018)
7. Dong, Y., Pang, T., Su, H., Zhu, J.: Evading defenses to transferable adversarial examples by translation-invariant attacks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019). pp. 4312–4321 (2019)
8. Engstrom, L., Tsipras, D., Schmidt, L., Madry, A.: A rotation and a translation suffice: Fooling cnns with simple transformations. CoRR **abs/1712.02779** (2017)
9. Fawzi, A., Moosavi-Dezfooli, S., Frossard, P.: The robustness of deep networks: A geometrical perspective. IEEE Signal Process. Mag. **34**(6), 50–62 (2017)
10. Finlay, C., Pooladian, A., Oberman, A.M.: The logbarrier adversarial attack: Making effective use of decision boundary information. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV 2019). pp. 4861–4869 (2019)

11. Giusti, A., Guzzi, J., Ciresan, D.C., He, F., Rodriguez, J.P., Fontana, F., Faessler, M., Forster, C., Schmidhuber, J., Caro, G.D., Scaramuzza, D., Gambardella, L.M.: A machine learning approach to visual perception of forest trails for mobile robots. IEEE Robotics Autom. Lett. **1**(2), 661–667 (2016)
12. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: 3rd International Conference on Learning Representations (ICLR 2015) (2015)
13. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) 14th European Conference on Computer Vision (ECCV 2016). Lecture Notes in Computer Science, vol. 9908, pp. 630–645. Springer (2016)
14. Hu, S., Shang, X., Qin, Z., Li, M., Wang, Q., Wang, C.: Adversarial examples for automatic speech recognition: Attacks and countermeasures. IEEE Communications Magazine **57**(10), 120–126 (2019)
15. Huang, Q., Katsman, I., He, H., Gu, Z., Belongie, S., Lim, S.: Enhancing adversarial example transferability with an intermediate level attack. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV 2019). pp. 4732–4741 (2019)
16. Huang, Z., Zhang, T.: Black-box adversarial attack with transferable model-based embedding. In: 8th International Conference on Learning Representations (ICLR 2020) (2020)
17. Ibitoye, O., Abou-Khamis, R., Matrawy, A., Shafiq, M.O.: The threat of adversarial attacks on machine learning in network security–a survey. arXiv preprint arXiv:1911.02621 (2019)
18. John, T.S., Thomas, T.: Adversarial attacks and defenses in malware detection classifiers. In: Handbook of Research on Cloud Computing and Big Data Applications in IoT, pp. 127–150. IGI global (2019)
19. Kanbak, C., Moosavi-Dezfooli, S., Frossard, P.: Geometric robustness of deep networks: Analysis and improvement. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018). pp. 4441–4449. IEEE Computer Society (2018)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. pp. 1106–1114 (2012)
21. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: 5th International Conference on Learning Representations (ICLR 2017). OpenReview.net (2017)
22. Laidlaw, C., Feizi, S.: Functional adversarial attacks. In: Advances in neural information processing systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS 2019). pp. 10408–10418 (2019)
23. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–44 (05 2015). https://doi.org/10.1038/nature14539
24. Lin, J., Song, C., He, K., Wang, L., Hopcroft, J.E.: Nesterov accelerated gradient and scale invariance for adversarial attacks. In: 8th International Conference on Learning Representations (ICLR 2020). OpenReview.net (2020)
25. Liu, Y., Chen, X., Liu, C., Song, D.: Delving into transferable adversarial examples and black-box attacks. In: 5th International Conference on Learning Representations (ICLR 2017) (2017)
26. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: 6th International Conference on Learning Representations, (ICLR 2018) (2018)

27. Martins, N., Cruz, J.M., Cruz, T., Abreu, P.H.: Adversarial machine learning applied to intrusion and malware scenarios: a systematic review. IEEE Access **8**, 35403–35419 (2020)
28. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M.A., Fidjeland, A., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D.: Human-level control through deep reinforcement learning. Nat. **518**(7540), 529–533 (2015)
29. Moosavi-Dezfooli, S., Fawzi, A., Frossard, P.: Deepfool: A simple and accurate method to fool deep neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2574–2582 (2016)
30. Moosavidezfooli, S., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 86–94 (2017)
31. Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R., Muharemagic, E.: Deep learning applications and challenges in big data analytics. J. Big Data **2**, 1 (2015)
32. Nesterov, Y.: A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. Doklady AN USSR **269**, 543–547 (1983)
33. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, New York, NY, USA, second edn. (2006)
34. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia conference on computer and communications security. pp. 506–519 (2017)
35. Papernot, N., McDaniel, P.D., Goodfellow, I.J.: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. CoRR **abs/1605.07277** (2016)
36. Ren, K., Wang, Q., Wang, C., Qin, Z., Lin, X.: The security of autonomous driving: Threats, defenses, and future directions. Proceedings of the IEEE **108**(2), 357–372 (2019)
37. Ru, B., Cobb, A., Blaas, A., Gal, Y.: Bayesopt adversarial attack. In: 8th International Conference on Learning Representations (ICLR 2020) (2020)
38. Serban, A.C., Poll, E., Visser, J.: Adversarial examples-a complete characterisation of the phenomenon. arXiv preprint arXiv:1810.01185 (2018)
39. Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation **23**(5), 828–841 (2019)
40. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: International conference on machine learning. pp. 1139–1147 (2013)
41. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014)
42. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: 2nd International Conference on Learning Representations (ICLR 2014) (2014)
43. Tu, C., Ting, P., Chen, P., Liu, S., Zhang, H., Yi, J., Hsieh, C., Cheng, S.: Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In: The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI 2019), The Thirty-First Innovative Applications of Artificial Intelligence Conference (IAAI 2019), The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI 2019). pp. 742–749. AAAI Press (2019)

44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
45. Wierstra, D., Schaul, T., Glasmachers, T., Sun, Y., Peters, J., Schmidhuber, J.: Natural evolution strategies. The Journal of Machine Learning Research **15**(1), 949–980 (2014)
46. Xiao, C., Zhu, J., Li, B., He, W., Liu, M., Song, D.: Spatially transformed adversarial examples. In: 6th International Conference on Learning Representations (ICLR 2018). OpenReview.net (2018)
47. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L.: Improving transferability of adversarial examples with input diversity. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019). pp. 2730–2739. Computer Vision Foundation / IEEE (2019)
48. Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A.E.D., Jin, W., Schuller, B.: Deep learning for environmentally robust speech recognition: An overview of recent developments. ACM Transactions on Intelligent Systems and Technology (TIST) **9**(5), 1–28 (2018)
49. Zhou, Y., Han, M., Liu, L., He, J., Gao, X.: The adversarial attacks threats on computer vision: A survey. In: 2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW). pp. 25–30. IEEE (2019)