Engineering and Technology Journal e-ISSN: 2456-3358

Volume 09 Issue 06 June-2024, Page No.- 4261-4267

DOI: 10.47191/etj/v9i06.12, I.F. – 8.227

Evaluating Transferability of Attacks across Generative Models

Rohith Vallabhaneni

Ph.D. Research Graduate, Department of Information Technology, University of the Cumberlands, USA

ORCID: 0009-0003-3719-2704

ABSTRACT: The need for adversarial sample transferability is to attack black-box deep learning models. Whereas much recent work focuses on making untargeted adversarial attacks more transferable, there has been scarce research on the creation of transferable targeted adversarial instances that can trick models into believing they are of a particular class. The present transferable targeted adversarial attacks are not transferable since they cannot sufficiently define the distribution of target classes. In this paper, we propose a generative adversarial training system consisting of a feature-label dual discriminator to identify the adversarial instances formed from the target class images and a generator to construct targeted adversarial examples. It is concluded that adversarial scenarios have significant real-world applications in safety-critical fields like biometrics and autonomous driving. In addition, it is demonstrated that the current networks' susceptibility to hostile attacks, even under the worst black-box conditions has far-reaching societal consequences. We intend to further encourage more research into the inner workings of neural networks in the face of adversarial attacks, whereby people might use this knowledge to build robust defense mechanisms.

KEYWORDS: Transferability, generative model, deep neural networks, adversarial attacks.

INTRODUCTION

In artificial intelligence (AI), one of the main fields is deep neural networks (DNNs). DNNs' practical applications are many and include face identification (Ma, 2002), voice recognition (Goldberg, 2016), picture classification (Anbukkarasi & Varadhaganapathy, 2022), and autonomous driving technology (Gu & Rigazio, 2014). Many investigations on DNN adversarial attacks, exploring the vulnerability and ambiguity of DNNs, have been prompted by the extensive consequences. The study by Balda et al. (2020) showed how adversarial cases might fool DNNs, by demonstrating that instances are produced by adding perturbations that are identical to human inputs. Therefore, it remains challenging to design hostile attacks that will generally produce high-quality adversarial instances, as practical adversarial examples need to remain hidden from humans while tricking DNNs into making predictions.

Untargeted and targeted attacks are the categories into which adversarial attack techniques fall. Untargeted adversarial attacks aim to fool the model into predicting random incorrect labels, whereas targeted adversarial attacks assume that the adversarial samples they supply will result in a misprediction for a specific label. Transferability of adversarial samples is crucial for both targeted and untargeted attacks, especially when the target model is hidden (black-box attacks). The use of generative techniques (Kos et al., 2018), data augmentation (Huang et al., 2017), model aggregation (Gao et al., 2017), and feature information utilization (Xiao et al., 2018) to improve the transferability of untargeted adversarial attacks has been the focus of most research up to this point. They were able to extract a limited amount of transferable information about the target class because of the over-fitting of the source model and the lack of target class distribution information, despite some of it being extended to the targeted adversarial attacks by changing the loss function (Gao *et al.*, 2017).

Agrawal et al. (2023) and Hu & Tan (2022) have recently studied how to make focused adversarial attacks more transferable. The target class information was obtained by the authors through the use of labeled probability distributions and feature maps. Label-wise information, which is often generated by the classification model's last layer, may depict the direct relationship between class labels and picture distribution. However, it has been established that learning solely from label-wise data leads to insufficient cross-model transferability since it keeps high-level semantic information from the original class (Papernot et al., 2017). In addition, it has been demonstrated that because the mid-level layer of different DNNs has comparable activation patterns, feature-wise information that may be gathered from the classification model's intermediate layer has transferability (Xiao et al., 2018). Nevertheless, neither the intended label nor the predicted misclassification is produced by the feature-wise data.

Nowadays, most adversarial transferability research focuses on training one or more replacement models to mimic the victim model's actions or execute tasks (Lin et al., 2022; Wang & He, 2021). However, because of the limitations of the black box scenario, the attacker cannot access the victim model's structural properties or training data, which makes it very difficult to train equivalent replacement models (Agrawal et al., 2023).

Consequently, relying just on the transferability of models with the same purpose makes it difficult to attain a high success rate in adversarial attacks.

To address the above-described issues, we propose a technique that eliminates the requirement to create task-specific replacement models by directly producing adversarial instances by extracting transferable characteristics from various tasks. In this regard, we demonstrate that adversarial transferability applies not just to tasks that are exact or equivalent, but also to models that have been trained on several tasks. Our idea is supported by a few findings. Adversarial word substitution rules provide various highly transferable candidate replacement phrases, for instance. By providing a wide variety of highly transferable adversarial instances, a bigger pool of candidate words can compensate for the drawbacks of earlier techniques that depended on greedy searches. Specifically, we use adversarial sample data from several tasks to train a sequence-to-sequence generative model called CT-GAT (Cross-Task Generative Adversarial ATtack). Amazingly, we hypothesize that even without precise victim model knowledge, the adversarial sample that is created may nonetheless effectively attack test tasks. The transferability of adversarial assaults between different generative models is evaluated in this work. Furthermore, the article provides insight into how security vulnerabilities might generalize across various model architectures.

2. MODELS OF DIFFERENT NEURAL ARCHITECTURES

2.1. Convolutional Neural Networks (CNNs)

Convolutional layers are used by CNNs to extract spatial properties from input images through weight sharing (Li et al., 2021). While ResNets and DenseNets employ skip connections to address the vanishing gradient and over-fitting problem, allowing deeper learning and more outstanding performance, InceptionV3 uses parallel convolutional filters to collect features of various sizes (O'Shea & Nash, 2015).

2.2. Vision Transformers (ViTs)

ViTs represent an image as a list of patches and employ self-attention techniques to extract global features, unlike CNNs, which are limited by local features like convolution (Ranftl et al., 2021). Two well-liked ViT versions are data-efficient Image Transformers (DeiTs) and Vision Transformers with shifting windows (Fan et al., 2021). DeiTs achieve data efficiency by knowledge distillation, whereas Swin Transformers use a hierarchical network topology with moving windows to effectively acquire context information for large-scale image processing applications (Park & Kim, 2022).

2.3. Spiking Neural Networks (SNNs)

SNNs have attracted much attention recently because of their biologically inspired computing and energy efficiency (Ghosh-Dastidar & Adeli, 2009). A feature of SNNs is the integration of time. Rather than continually communicating information at each propagation cycle, neurons in the SNN fire discrete spikes when the accumulated stimulation exceeds the threshold (Lobo et al., 2020). Three types of specific learning strategies are available to SNNs since spiking neurons are not differentiable: conversion-based training, supervised learning with surrogate gradients, and unsupervised learning (Ponulak & Kasinski, 2011). The conversion-based training method maps pre-trained CNN parameters to an SNN and modifies the weights to improve performance.

2.4. Dynamic Neural Networks (DyNNs)

Unlike traditional static neural networks, which have a set topology and number of parameters, DyNNs may adapt their structure and behavior to the complexity of the task at hand, increasing processing efficiency and lowering costs. The DyNN evaluated in this paper was the Glance and Focus Network (GFNet), adapted from Han et al. (2021). The GFNet technique consists of two stages: look and focus. The process is sequential from start to finish. After Glance has processed the down-scaled images, those that exhibit distinct features may now be securely classified (Becerikli et al., 2003). If the prediction is not sufficiently trustworthy, the framework proceeds to the focus stage, which processes progressively smaller, class-discriminative sections of the full-resolution picture (Jacques et al., 2011). During the focus stage, adaptive termination based on predicted confidence is feasible.

3. ADVERSARIAL ATTACKS

3.1. Instance-specific attacks

Many recent studies have used gradient-based optimization approaches to produce the data-dependent perturbations (Zhang et al., 2022). To increase black-box transferability, MIM adds the momentum term to the iterative attack method (Gupta et al., 2004). DIM and TI seek to enhance transferability through input or gradient variety (Hoang et al., 2017; Qin et al., 2022). By expensively training many auxiliary classifiers, recent research (e.g. Essich et al., 2023; Luo et al., 2023) also sought to enhance the black-box performance of the iterative approaches. On the other hand, we contend that, compared to instance-specific approaches, it is crucial to enhance the transferability performance and the inference-time efficiency in the black box.

3.2. Instance-agnostic attacks

Instance-agnostic attacks are a subset of image-independent (universal) approaches, as opposed to instance-specific attacks. The first pipeline is the discovery of a universal disturbance. Feng et al. (2023) claim that UAP proposes using a learned universal noise vector to fool a model. An additional attack technique generates adversarial samples using trained generative models. On the other hand, Li et al. (2022) state that GAP and AAA produce adversarial perturbations and compress perceptions by directly utilizing target data. Training the same number of models for several target classes using earlier approaches such as universal perturbation and function is costly (Wang et al., 2023). Our approach might potentially provide

adversarial samples with improved attack capabilities for several targets simultaneously.

3.3. Multi-target attacks

Instance-specific attacks are capable of identifying any target during the optimization phase. These methods involve laborious iterative procedures with little transferability (Zhang et al., 2021). MAN creates a generative model in ImageNet with a ℓ2 norm constraint to explore particular risks. This model defines all 1,000 categories from ImageNet to achieve exceptional performance and storage (Han et al., 2019). However, the authors also assert that having too many categories makes it more challenging to switch between models, and they do not provide a complete comparison between the multi-target black-box performance of MAN and prior instance-specific or instance-agnostic assaults. To improve single-target transferability, more recent methods (Yao et al., 2023; Li et al., 2023) create a universal perturbation or function; nonetheless, they necessitate several training sessions with various target specifications. On the other hand, our approach can develop adversarial samples for multiple target identification, and the solid semantic patterns it generates can significantly exceed current assaults.

With a maximum perturbation of $\epsilon = 16$, the targeted adversarial attacks for the target class Viaduct employing MIM and C-GSP are shown in Figure 1(a) (Yang et al., 2022). In a second black-box model, predicted labels and probability are shown. The presentation in (b) provides an overview of our proposed generative technique, which consists of conditional generator and classifier modules. The generator creates a hidden incorporation by combining the conditional class vector and image from the Map network. Throughout the process, the generator is instructed to look inside the classifier's goal boundaries.

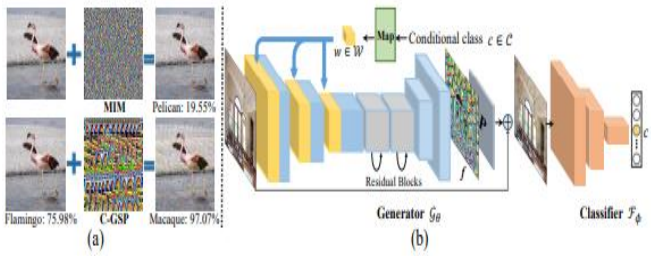

Fig. 1. Targeted Adversarial Examples
(**Source:** Yang et al., 2022)

4. GENERATIVE MODELING FOR ADVERSARIAL EXAMPLES

Generative modeling has been widely used in statistics. Numerous fields, such as data production tasks, audio identification, visual recognition, and natural language processing, have benefited from its application to machine learning (Kos et al., 2018). Among these generative modeling methods are Markov Random fields, Hidden Markov Models, Bayesian networks, Linear Discriminant Analysis (LDA), and Naive Bayes (Yang et al., 2022). With the development of Deep Learning, graphical models such as Sigmoid Belief Networks, Variational Autoencoders, Differentiable Generator Networks, Restricted Boltzmann Machines, and Boltzmann machines, as well as Deep Belief Networks, have become possible. Recently, there has been much interest in the generative model known as a Generative Adversarial Network, or GAN, because of its remarkable capacity to generate synthetic data (Luo et al., 2023). The process of generating synthetic data for GAN models is depicted in Figure 2.

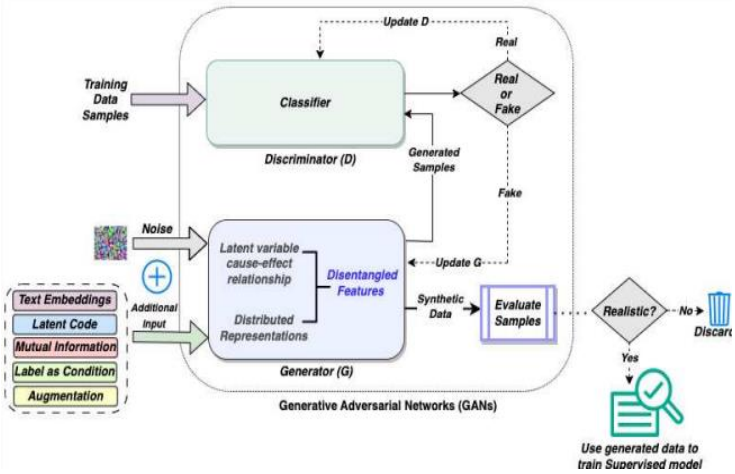

Fig. 2. Synthetic Data Generation Process in GANs
(**Source:** Torres, 2018)

A significant amount of labeled data is required for discriminative models or supervised learning algorithms to perform tasks such as generating new data examples, calculating the probability of an event, handling missing values using available unlabeled data, or inferring information from related activities. Relatively high precision is needed for these activities. Training the model becomes more challenging in sectors with less data, such as cyber security, because labeling data may be a time-consuming and expensive operation (Xiao et al., 2018). Unsupervised and semi-supervised learning approaches are more likely to be used in these circumstances (Huang et al., 2017). However, few of them have reached the same level of precision as the supervised algorithms. The unsupervised algorithms have difficulties because of the enormous dimensionality of random variables. The exponential growth of dimensions exacerbates the computational and statistical challenges associated with finding a tractable solution to a problem and with generalizing the number of configurations. One technique to deal with the considerable dimensionality of intractable computations is to approximate or design in a way that removes the requirement for such computations. When generative modeling techniques are used, the latter design approach has demonstrated potential (Gao et al., 2022).

Researchers have discovered advantages in using generative models to create adversarial instances. Whitebox and query-based attacks, for example, are indicated as effective attack strategies (Xiao et al., 2018). Emerging threat models are also

studied using generative models, such as unconstrained adversarial instances and semantic adversarial examples (Li et al., 2020). Transferability is not enhanced, despite the strong correlation between adversarial patches and unconstrained adversarial instances (Mustafa et al., 2019). Even though Feng et al. (2023) assert that transferability has improved in their scenario, we demonstrate that it is not ideal in the patch situation1.

5. PRODUCING ATTACK DATA WITH ADVERSARIAL EXAMPLES

It is increasingly essential to examine vulnerabilities in these systems because of the remarkable efficacy of machine learning models and their extensive use in a variety of security-sensitive applications. According to Szegedy et al. (2013), hostile changes made to the input data frequently result in inaccurate classifier outputs. Because even cutting-edge models, such as deep neural networks, are highly vulnerable to adversarial attacks that, in the worst-case scenario, purposefully manipulate the input, the security and integrity of existing machine learning algorithms are jeopardized (Papernot et al., 2017). These adversaries are generated faster and have access to the target models' gradients than data impacted by random noise of even greater magnitude, which results in significantly higher assault success rates. Moreover, additional regularisation may be advantageous for machine learning models trained against adversaries of this nature (Agrawal et al., 2023).

Although they are artificial, these adversarial scenarios highlight "blind spots" in machine learning models since it is improbable that the classifier will encounter these worst-case disruptive occurrences in real-world circumstances. As a result, it is challenging to get insightful knowledge about the basic decision-making mechanisms within the black-box classifier. The rationale behind the adversary's decisions, what can be changed to stop this behavior, and whether the classifier can tolerate random fluctuations in the data when it is not functioning in an adversarial environment are a few examples of these processes. Additionally, the understandable semantic space and the input space frequently diverge. According to Papernot et al. (2017), little adjustments made to the input that would not seem important, such as a little visual translation or rotation, can often have a significant impact on the input example. According to Goodfellow et al. (2014), light changes can trick automated driving systems, even when they are moderate in size. Adversarial situations are unable to identify this characteristic.

The main problem involves predicting the intended distribution, which may be challenging and time-consuming. Agrawal et al. (2023) have presented the generative adversarial network framework as a potential solution. Motivated by concepts from game theory, they trained two models: the discriminator G, which creates data from input by following a source distribution in an attempt to trick the discriminator, and the generator D, which determines the extent to which generated data differs from natural data (Szegedy et al., 2013). Images may be effectively produced using the generator. Although sound is produced, the training process is not steady. To enhance learning strategy stability and address issues such as mode collapse, Zhao et al. (2017) introduced the WGAN method, which modifies the training strategy and adopts new distances.

The adversary can overcome ignorance in black box attacks by training a local replacement DNN using a fake dataset. The outputs are the labels that the Oracle or remote DNN assigned when the adversary queried the DNN with their artificial inputs; in contrast, the adversary produced the artificial inputs. The adversary constructs adversarial situations that lead to an inaccurate classification of the replacement model, and similar decision limits are employed in its creation. These same hostile samples might subsequently be used to misclassify the target DNN. Two models, MalGAN (Hu & Tan, 2022) and IDSGAN (Lin et al., 2022), were proposed to employ GANs to generate synthetic adversarial attacks to test the detection system. We also explore the creation and evaluation of these models' capacity to provide realistic adversarial attack scenarios inside this framework.

6. INCREASING ATTACK TRANSFERABILITY

One intriguing aspect of adversarial attacks is their transferability. Instead of using a single surrogate network, an ensemble-based attack uses many of them (Wang & He, 2021). Ghost networks generate different surrogate networks by interfering with dropout layers and skip connections (Li et al., 2020). VT provides gradient variance in the management of the stability of the localized gradients, while MI and other optimal approaches apply a momentum-based optimization (Hoang et al., 2017; Lee et al., 2012). Adversarial situations produced by RAP are situated in an area of flat loss (Qin et al., 2022). While image-altering techniques such as scaling and padding are used by data augmentation methods like DI, the TI considers picture-pixel translation (Xiao et al., 2018; Zou et al., 2022). The SI uses several scaled benign samples to compute gradients (Gao et al., 2022). Admix blends the benign pictures with randomly selected images to produce iterative gradients (Zou et al., 2020).

Adversarial attacks are related to various network architectures and features in different ways. While Chakraborty et al. (2018) demonstrate that LinBP omits the nonlinear activation during propagation, Guo et al. (2018) suggest that adversarial vulnerability might arise from DNN linearity. Skip connections are used by SGM in residual networks to employ higher gradients (Mustafa et al., 2019). It is possible to maximize the distance in feature spaces between natural images and their adversarial examples (Yuan et al., 2019), use ILA to improve adversarial examples presently in the intermediate layer level (Zhang & Li, 2019), or train auxiliary classifiers based on feature spaces (Silva & Najafirad, 2020).

7. CONCLUSION

DNNs have demonstrated incredible potential across several domains. It has been demonstrated that DNNs may significantly

be influenced by adversarial instances that are created by introducing small perturbations to otherwise perfect images. Two broad categories of attacks have been specifically studied. Since the first one requires iterative methods to optimize the perturbation for each occurrence, it is frequently computationally expensive. Generative methods are employed in the latter case to train a deep network to produce perturbations. In this paper, we have demonstrated that a feature separation loss-trained generator can successfully fool models across architectures and workloads and that an effective perturbation may be learned using a proxy dataset from a different domain. Our results demonstrate that our technique beats cutting-edge attacks across various configurations and workloads. Due to a shortage of publicly accessible robust models across architectures, we limited our trials to undefended models; however, we will study this further in future work. In addition, we believe that a more in-depth investigation of learned filter banks in connection to architectural changes might provide insight into how to develop better black-box models. Understanding the capabilities of adversary assaults is critical for future security development. Adversarial scenarios have significant real-world applications in safety-critical fields like biometrics and autonomous driving. This paper reveals the current networks' susceptibility to hostile assaults, even under the worst black-box conditions, with far-reaching societal consequences. We desire that this research will inspire other people to investigate the inner workings of neural networks when facing adversarial attacks and to apply this understanding to create strong defenses.

REFERENCES

1. Agrawal, G., Kaur, A., & Myneni, S. (2023). Review of Generative Models in Generating Synthetic Attack Data for Cybersecurity.
2. Anbukkarasi, S., & Varadhaganapathy, S. (2022). Neural network-based error handler in natural language processing. *Neural Computing and Applications*, *34*(23), 20629-20638.
3. Balda, E. R., Behboodi, A., & Mathar, R. (2020). Adversarial examples in deep neural networks: An overview. *Deep learning: algorithms and applications*, 31-65.
4. Becerikli, Y., Konar, A. F., & Samad, T. (2003). Intelligent optimal control with dynamic neural networks. *Neural networks*, *16*(2), 251-259.
5. Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2018). Adversarial attacks and defenses: A survey. *arXiv preprint arXiv:1810.00069*.
6. Essich, M., Rehmann, M., & Curio, C. (2023). Auxiliary Task-Guided CycleGAN for Black-Box Model Domain Adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 541-550).
7. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., & Feichtenhofer, C. (2021). Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6824-6835).
8. Feng, W., Xu, N., Zhang, T., & Zhang, Y. (2023). Dynamic Generative Targeted Attacks with Pattern Injection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16404-16414).
9. Gao, H., Zhang, H., Yang, X., Li, W., Gao, F., & Wen, Q. (2022). Generating natural adversarial examples with universal perturbations for text classification. *Neurocomputing*, *471*, 175-182.
10. Gao, J., Wang, B., Lin, Z., Xu, W., & Qi, Y. (2017). Deepcloak: Masking deep neural network models for robustness against adversarial samples. *arXiv preprint arXiv:1702.06763*.
11. Ghosh-Dastidar, S., & Adeli, H. (2009). Spiking neural networks. *International journal of neural systems*, *19*(04), 295-308.
12. Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, *57*, 345-420.
13. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
14. Gu, S., & Rigazio, L. (2014). Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*.
15. Guo, Y., Zhang, C., Zhang, C., & Chen, Y. (2018). Sparse dnns with improved adversarial robustness. *Advances in neural information processing systems*, *31*.
16. Gupta, M., Jin, L., & Homma, N. (2004). *Static and dynamic neural networks: from fundamentals to advanced theory*. John Wiley & Sons.
17. Han, J., Dong, X., Zhang, R., Chen, D., Zhang, W., Yu, N., ... & Wang, X. (2019). Once a man: Towards multi-target attack via learning multi-target adversarial network once. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 5158-5167).
18. Han, Y., Huang, G., Song, S., Yang, L., Wang, H., & Wang, Y. (2021). Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(11), 7436-7456.
19. Hoang, C. D. V., Haffari, G., & Cohn, T. (2017). Towards decoding as continuous optimization in neural machine translation. *arXiv preprint arXiv:1701.02854*.
20. Hu, W., & Tan, Y. (2022, November). Generating adversarial malware examples for black-box attacks based on GAN. In *International Conference on Data*

Mining and Big Data (pp. 409-423). Singapore: Springer Nature Singapore.

21. Huang, S., Papernot, N., Goodfellow, I., Duan, Y., & Abbeel, P. (2017). Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*.
22. Jacques, P. L. S., Kragel, P. A., & Rubin, D. C. (2011). Dynamic neural networks supporting memory retrieval. *Neuroimage*, *57*(2), 608-616.
23. Kos, J., Fischer, I., & Song, D. (2018, May). Adversarial examples for generative models. In *2018 ieee security and privacy workshops (spw)* (pp. 36-42). IEEE.
24. Lee, J. D., Sun, Y., & Saunders, M. (2012). Proximal Newton-type methods for convex optimization. *Advances in Neural Information Processing Systems*, *25*.
25. Li, M., Yang, Y., Wei, K., Yang, X., & Huang, H. (2022, June). Learning universal adversarial perturbation by adversarial example. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 2, pp. 1350-1358).
26. Li, Y., Bai, S., Zhou, Y., Xie, C., Zhang, Z., & Yuille, A. (2020, April). Learning transferable adversarial examples via ghost networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 07, pp. 11458-11465).
27. Li, Y., Liu, S., Chen, K., Xie, X., Zhang, T., & Liu, Y. (2023). Multi-target Backdoor Attacks for Code Pre-trained Models. *arXiv preprint arXiv:2306.08350*.
28. Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*.
29. Lin, Z., Shi, Y., & Xue, Z. (2022, May). Idsgan: Generative adversarial networks for attack generation against intrusion detection. In *Asian Pacific Conference on knowledge discovery and data mining* (pp. 79-91). Cham: Springer International Publishing.
30. Lobo, J. L., Del Ser, J., Bifet, A., & Kasabov, N. (2020). Spiking neural networks and online learning: An overview and perspectives. *Neural Networks*, *121*, 88-100.
31. Luo, D., Zhang, C., Xu, J., Wang, B., Chen, Y., Zhang, Y., & Li, H. (2023). Enhancing Black-Box Few-Shot Text Classification with Prompt-Based Data Augmentation. *arXiv preprint arXiv:2305.13785*.
32. Ma, Q. (2002, December). Natural language processing with neural networks. In *Language Engineering Conference, 2002. Proceedings* (pp. 45-56). IEEE.
33. Mustafa, A., Khan, S., Hayat, M., Goecke, R., Shen, J., & Shao, L. (2019). Adversarial defense by restricting the hidden space of deep neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3385-3394).
34. Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., & Lakshminarayanan, B. (2018). Do deep generative models know what they do not know? *arXiv preprint arXiv:1810.09136*.
35. O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.
36. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017, April). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security* (pp. 506-519).
37. Park, N., & Kim, S. (2022). How do vision transformers work? *arXiv preprint arXiv:2202.06709*.
38. Ponulak, F., & Kasinski, A. (2011). Introduction to spiking neural networks: Information processing, learning and applications. *Acta neurobiologiae experimentalis*, *71*(4), 409-433.
39. Qin, Z., Fan, Y., Liu, Y., Shen, L., Zhang, Y., Wang, J., & Wu, B. (2022). Boosting the transferability of adversarial attacks with reverse adversarial perturbation. *Advances in Neural Information Processing Systems*, *35*, 29845-29858.
40. Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 12179-12188).
41. Silva, S. H., & Najafirad, P. (2020). Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv preprint arXiv:2007.00753*.
42. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
43. Torres, D. G. (2018). *Generation of synthetic data with generative adversarial networks* (Doctoral dissertation, Royal Institute of Technology).
44. Wang, D., Yao, W., Jiang, T., & Chen, X. (2023). Improving Transferability of Universal Adversarial Perturbation with Feature Disruption. *IEEE Transactions on Image Processing*.
45. Wang, X., & He, K. (2021). Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1924-1933).
46. Xiao, C., Li, B., Zhu, J. Y., He, W., Liu, M., & Song, D. (2018). Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*.
47. Yang, X., Dong, Y., Pang, T., Su, H., & Zhu, J. (2022, October). Boosting transferability of targeted adversarial examples via hierarchical generative networks. In *European Conference on Computer*

Vision (pp. 725-742). Cham: Springer Nature Switzerland.

48. Yao, Z. H., Lie, Y. M., Ma, Z. J., Li, Y., & Wei, L. G. (2023). Machine learning-based multi-target cache side-channel attack detection model. *Journal of Computer Applications*, 0.
49. Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, *30*(9), 2805-2824.
50. Zhang, J., & Li, C. (2019). Adversarial examples: Opportunities and challenges. *IEEE transactions on neural networks and learning systems*, *31*(7), 2578-2593.
51. Zhang, J., Wang, Y., & Zhuang, J. (2021). Modeling multi-target defender-attacker games with quantal response attack strategies. *Reliability Engineering & System Safety*, *205*, 107165.
52. Zhang, M., Wu, S., Yu, X., Liu, Q., & Wang, L. (2022). Dynamic graph neural networks for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, *35*(5), 4741-4753.
53. Zhao, Z., Dua, D., & Singh, S. (2017). Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*.
54. Zou, J., Duan, Y., Li, B., Zhang, W., Pan, Y., & Pan, Z. (2022, June). Making adversarial examples more transferable and indistinguishable. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 3, pp. 3662-3670).
55. Zou, J., Pan, Z., Qiu, J., Liu, X., Rui, T., & Li, W. (2020, August). Improving the transferability of adversarial examples with resized-diverse-inputs, diversity-ensemble, and region fitting. In *European Conference on Computer Vision* (pp. 563-579). Cham: Springer International Publishing.