IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 3, March 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303085 |

Black-Box Attacks on Visual Model of Luminance Perception in Random Gamma Transform

D.Monisha, Mr.R.Praveenkumar, Dr.M.Malarvizhi,

M. E Embedded System Technology, Gnanamani College of Technology, NH-7, A.K.Samuthiram, Pachal-PO, Namakkal, Tamil Nadu, India

Assistant Professor/EEE, M. E Embedded System Technology, Gnanamani College of Technology, NH-7, A.K.Samuthiram, Pachal-PO, Namakkal, Tamil Nadu, India

Professor/EEE, M. E Embedded System Technology, Gnanamani College of Technology, NH-7, A.K.Samuthiram, Pachal-PO, Namakkal, Tamil Nadu, India

ABSTRACT: Aiming at the black-box adversarial attack application, hackers can obtain given input labels only. The paper researched existing problems in the black-box attack method. The PSO-BBA is put forward, which avoids the current BBA method falling into the situation of a local optimal solution. The proposed PSO-BBA method improves the convergence speed of the BBA and ensures the performance of the black-box attack. Through the experimental comparison with the baseline method, it is proved that the proposed PSO-BBA effectively reduces query numbers. It produced adversarial samples at a lower self-cost and faster efficiency successfully. Our next research will focus on conducting detailed theoretical and experimental analyses on the selection of initial particles in the improved algorithm.

KEYWORDS: cybersecurity; adversarial machine learning; machine learning; intrusion detection; functionality-preservation

I. INTRODUCTION

Deep learning is occupying the core of the rapidly developing field of machine learning and artificial intelligence research. It also demonstrates good performance on many tasks, especially in the field of computer vision. However, modern deep networks are very vulnerable to adversarial samples, which presents a great threat to the effectiveness and stabilization of the community [1,2,3,4,5]. It is shown that image classification methods based on deep neural networks (DNN) are always fragile. They will make mistakes when only small disturbances occur, which are invisible to the human eye. This indicates that many machine learning models trained with a large number of samples are still vulnerable to small adversarial disturbances. The algorithm to find this kind of anti-interference to the input image is usually called an anti-attack. The mathematical expression is as Equation (1). The $C(x)$ is the classifier. In machine learning, the function of the classifier is to judge the category of a new input sample based on the labeled training data. Value (x, y) is an input sample and its label. The attack is to produce an adversarial sample x^{adv} of x. When x^{adv} is misclassified by ML model. The L_p norm of the difference between it and the original one is below a set threshold ε. It means that the attack succeeded.

Adversarial attack methods generally fall into two categories on the basis of attack effect and method. In the light of attack effects, these come in two classes: non-targeted and targeted attack methods. Non-targeted attack refers to the generation of adversarial samples by adding interference. It causes the classification algorithm to mistakenly classify the original image into images of different categories. Contrarily, targeted attack causes the classification algorithm to misclassify the original image as an image of a specified category.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303085 |

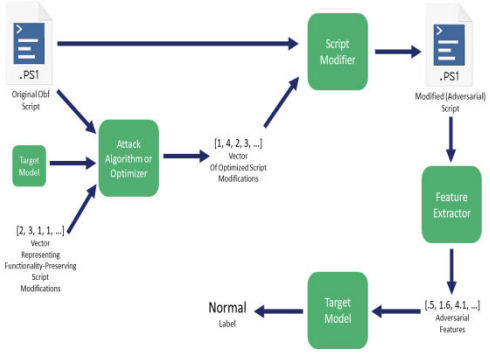


Fig 1: Block Diagram

On the basis of methods, these come in two classes: white and black-box methods. The white-box attack methods know about attacking deep neural network methods and model parameters in advance, such as the construction of networks, the situation of layers, neurons, etc. Additionally, they know the kinds of hyperparameters about the neural network, and the weights of connected neurons, for example. The attacker can design a targeted adversarial sample on the basis of the known information with regard to the neural network. This can make the neural network method lose efficacy. Such methods comprise the fast gradient attack proposed by Ian Goodfellow in 2014 [6] (Fast Gradient Sign Method, FGSM), Basic Iterative Method [7] (BIM), and DeepFool [8], Jacobian-based Saliency Map [9] (JSMA), Houdini [10] and Carlini & Wagner [11], and so on. Such attack algorithms mainly rely on gradient-based attack strategies.

Conversely, black-box attack methods do not know any information about the network, such as architecture or various hyperparameters. These only have the output results of the network and corresponding data labels. These ones are closer to the real application scenario than white ones. Such methods are harder to implement than white-box methods. Meanwhile, the attack effect is relatively poor. In terms of safety, the robustness of the neural network can be improved by the research of black-box methods. The network is trained with the corresponding adversarial samples. It is very helpful in improving the security of the network application.

The black-box attack methods can be generally divided into three categories: score-based attack, transfer-based attack and decision-based attack methods. Score-based attacks rely on predicted scores, such as the class probability or distribution of the model. These attacks use predicted values to estimate the gradient, mainly including the black-box variants of JSMA [12], zero-order Optimization attacks [13], predictive adversarial generative networks [14], and simple black-box attacks [15,16,17]. The disadvantage of this type of method is that it is difficult to obtain the score and probability distribution of the output sample in real applications, so it cannot be practically used.

A transfer-based attack does not rely on model information. It requires information about training data, which are used to train a fully observable agent model. The adversarial disturbances can be synthesized [18]. It relies on empirical observations. The adversarial examples are often transferred between models. If adversarial examples are created on a set of proxy models, the success rate of attacking models can reach 100% in some cases [19]. Dong proposed a prior stochastic gradient free method [20] to improve the performance of black-box adversarial attacks. The algorithm also uses transfer based on the priori and score-based information. The attack success rate is higher than only one method. If the data sets are inconsistent, the attack effect may decrease a lot, and the stability is not robust [21].

II. RELATED WORKS

Boundary attack algorithm is a kind of typical decision-based method. Starting from an initial adversarial image, a binary search is used to find a sample point, which is near the classification frontier. Random walk is performed along the frontier between two opposite areas. It reduces the distance from the target image. That is to say, the classification

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303085 |

result obtained by input classifier query is always the category we want to misclassify. According to this step, we continue to iterate and gradually reduce the distance from the original image. The reason why this kind of algorithm is named "boundary attack" is that it generates adversarial samples by searching along the boundary until it converges to obtain the optimal or satisfactory solution.

The results obtained by this kind of method can meet the requirements for misclassification of the black-box model. The overall disturbance that it increases relative to the original image varies with the performance of the algorithm. Therefore, the criterion for measuring a successful adversarial sample is that the difference between it and the original image should be less than the given threshold ε. Take the ImageNet data set as an example: the image size is 299 × 299. The pixel value interval is (0, 255). When $p = 2$ (that is, the L_2 norm) ε takes 25.89. This measure is used in our research and experimental results. As shown in the generation of optimal adversarial samples can be considered as a search and optimization problem along the classification boundary between adversarial space and original space. Those boundary points that are both close to the original image and on the adversarial sample space are the feasible solutions that we want to find. However, in the decision-based black-box adversarial scene, the position and gradient direction of the real classification boundary are unknown. Therefore, it can be regarded as a high-dimensional space search problem in an uncertain environment. In the context of this complex problem, the BA algorithm adopts the method of distributing P from an appropriate pixel space. An adversarial disturbance is found closer to the original image according to a given adversarial criterion. At the same time, in order to overcome the uncertainty of the classification boundary, it needs to consume several queries to obtain a successful adversarial disturbance. The basic flow of the algorithm is as follows.

The second step is to determine the disturbance. In order to ensure that the disturbance of each sampling is improved along the direction of the classification boundary as much as possible, as shown firstly the disturbance η_k with the sampling step size of δ is sampled on the hypersphere with the original image as the ball center, which is called the orthogonal disturbance. The disturbance of one step length of ε is sampled towards the direction of the original input image. Then the next adversarial sample point is obtained.

Boundary attack is easy to use. It just needs to adjust the two parameters of step size δ and ε. It has a good effect on target and non-target attack. This method can find adversarial samples that are similar to the effect of gradient-based white box attacks. It does not rely on the proxy model trained on the data, which is similar to the attack model. Besides, it has stronger robustness to the common defense methods, such as gradient masking, inherent randomness and robust training. The disadvantage is high query times and the efficiency is low. The black box model needs to be queried many times to generate successful adversarial samples. A distributed particle swarm optimization (PSO) algorithm is proposed to improve the initialization and optimization process. The improved details are introduced below.

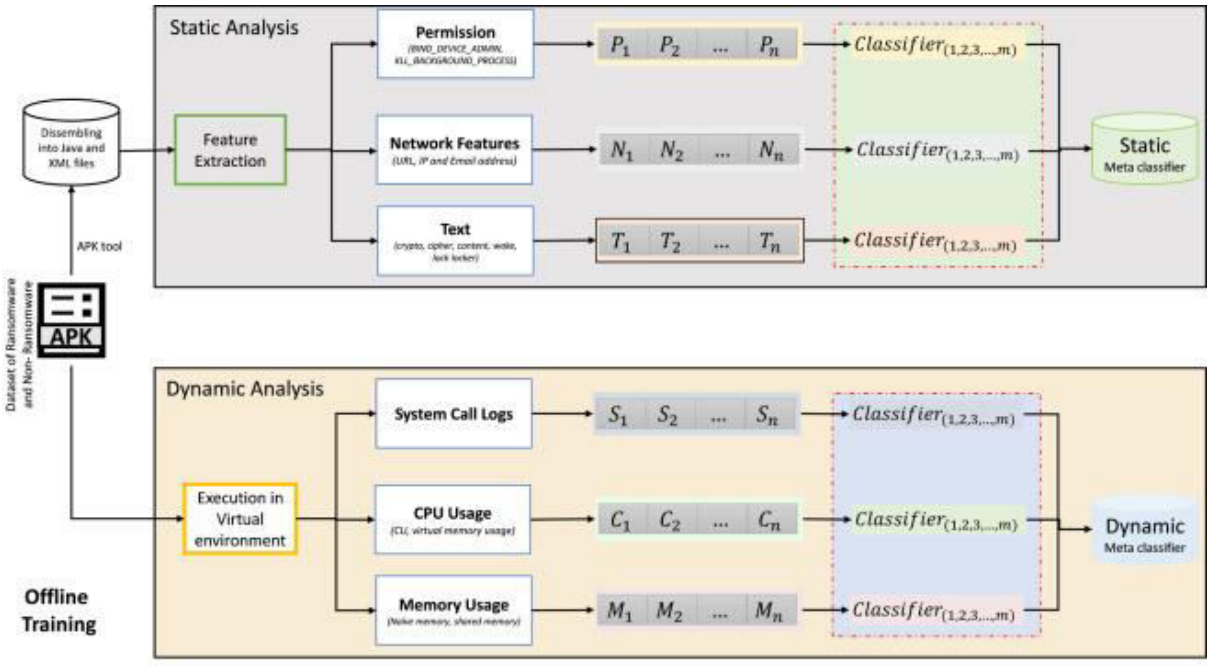
III. THE PROPOSED DISTRIBUTED FRAMEWORK

In BA methods, the generation of adversarial samples is equivalent to a search problem. One of the directions that can be improved in this problem is to define a heuristic information with a guiding function. The BBA is to introduce heuristic knowledge for the BA algorithm to speed up the search speed and efficiency. In addition, another important improvement direction is reducing the size of search space. In both BA and BBA algorithms, the approach taken begins with a single initial point. They search in a huge boundary space in a random or heuristic manner. This leads to the difficulty of too large search space, and the search gets stuck at local optimum easily. Therefore, this paper proposes the idea of combining particle swarm optimization (PSO) in the swarm intelligence algorithm to decompose the single-point search optimization into multiple points to search in their respective subspaces. Particle swarm optimization (PSO) algorithm is a random search algorithm based on group cooperation, which is developed by simulating the foraging behavior of birds. The purpose is to improve the possibility of searching for a practicable solution here. It can improve the success rate of adversarial attacks while speeding up the convergence rate. It can overcome shortcomings caused by the local optimal solution.

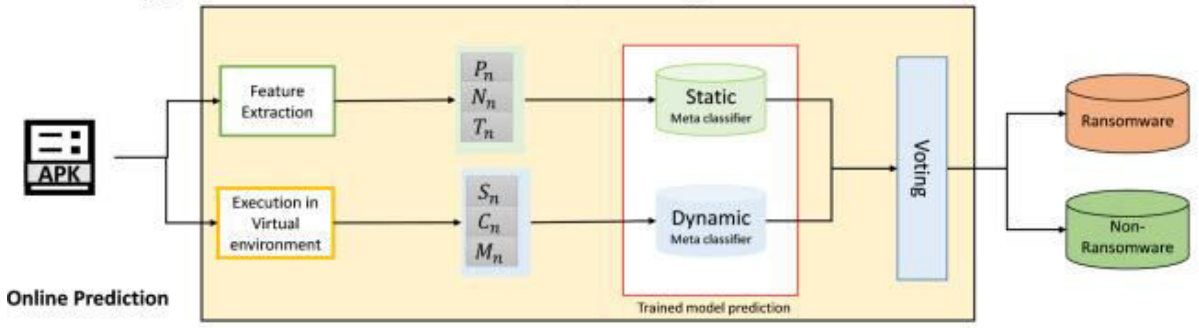
|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303085 |



(a) Offline features extraction, training and ensemble method.



(b) Online prediction of the trrain method.

In adversarial environments, the perturbation space is searched and optimized by the continuous update of the particle swarm. When the updating position is outside the scope of sample space, the perturbation created by Equation (3) is the velocity updating to guarantee that the particle position stays in it. The process does not waste any number of queries. Instead, it updates autonomously on the basis of the current orientation of each particle, and the best one. The query is needed for each update to judge only once that a particle has exceeded the boundary or not. If it is judged to exceed the space, a generated velocity is conducted. As the red arrowhead shows, the best sample is finally found through search, that is, the position of the yellow star. The adversarial sample search problem essentially has multiple solutions. Sometimes, it only needs to seek out one solution quickly. It can then be considered successful. The PSO algorithm based on multiple initial points can seek out the optimization among subspaces parallelly, which accelerates the speed of solution. One of the cost tradeoffs involved is that, per particle, it has to use up certain query numbers.

IV. RESULT ANALYSIS

A specific black-box attack example comparison experiment is shown in Figure 7. It can be seen that the improved PSO-BBA algorithm can effectively reduce the number of queries while ensuring the success rate of the attack. Although the similarity between the initial interference image generated by each particle and the original is poor. By expanding search space, the improved PSO-BBA algorithm generates a higher degree of similarity to the original image based on all particles. It also improves the convergence speed of the algorithm in comparison to the original method.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303085 |

The local optimal situation is avoided effectively. The performance of the attack algorithm is also improved. Figure 8shows an example of a failed BBA attack. The improved PSO-BBA algorithm succeeds and the number of attack queries is significantly reduced. However, the comparison results of MSSIM show that the two methods have changed the structural information of the original image. So, some evaluation indexes can be used as the indicators of detecting attacks.

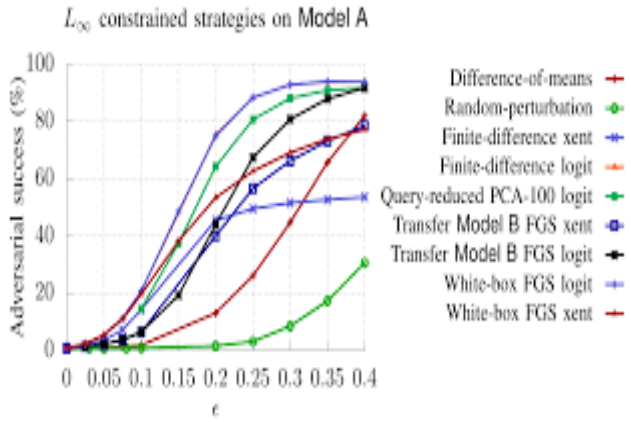


Fig 3:Evasion Black box Attacks

V. CONCLUSION

box attacks, using a probabilistic fingerprint to detect highly similar queries generated by attack optimization. Blacklight achieves near-perfect detection against eight SOTA attacks with negligible false positives, resists persistent attackers, and is robust to a range of adaptive and even idealized countermeasures. We also demonstrated that Blacklight can successfully generalize to some text classification tasks. Blacklight faces two limitations that demand further research. First, it is unable to defend against substitute model (SM) attacks, but can be combined with SM defenses to launch a more complete defense against both types of blackbox attacks (see Appendix §A for initial results). Second, Blacklight relies on the fact that existing query-based blackbox attacks all produce highly similar queries during their iterative optimization process, a phenomenon rarely seen in benign queries. It is not future-proof, i.e. a (future) attack breaking this assumption would evade Blacklight.

REFERENCES

1. Singh, A.K.; Liu, X.; Wang, H.; Ko, H. Recent advances in multimedia security and information hiding. *Trans. Emerg. Telecommun. Technol.* **2020**, *32*, e4193. [Google Scholar]
2. Xu, G.; Li, H.; Liu, S.; Yang, K.; Lin, X. VerifyNet: Secure and Verifiable Federated Learning. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 911–926. [Google Scholar] [CrossRef]
3. Phong, L.T.; Aono, Y.; Hayashi, T.; Wang, L.; Moriai, S. Privacy-Preserving Deep Learning via Additively Homomorphic Encryption. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 1333–1345. [Google Scholar] [CrossRef]
4. Lv, Z.; Qiao, L.; Song, H. Analysis of the Security of Internet of Multimedia Things. *ACM Trans. Multimed. Comput. Commun. Appl.* **2021**, *16*, 1–16. [Google Scholar] [CrossRef]
5. Xu, J.; Wang, C.; Li, T.; Xiang, F. Improved Adversarial Attack against Black-box Machine Learning Models. In Proceedings of the 2020 Chinese Automation Congress (CAC 2020), Shanghai, China, 6–8 November 2020; pp. 5907–5912. [Google Scholar]
6. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. *arXiv* **2014**, arXiv:1412.6572, 1–11. [Google Scholar]
7. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial Examples in the Physical World. In *Artificial Intelligence Safety and Security*; CRC Press: Boca Raton, FL, USA, 2018; pp. 99–112. [Google Scholar]

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303085 |

8. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. DeepFool: A simple and accurate method to fool deep neural networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 June 2016; pp. 2574–2582. [Google Scholar]
9. Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z.B.; Swami, A. The Limitations of Deep Learning in Adversarial Settings. In Proceedings of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P), Saarbruecken, Germany, 21–24 March 2016; pp. 372–387. [Google Scholar]
10. Cisse, M.; Adi, Y.; Neverova, N.; Keshet, J. Houdini: Fooling deep structured prediction models. *arXiv* **2017**, arXiv:1707.05373. [Google Scholar]
11. Carlini, N.; Wagner, D. Towards Evaluating the Robustness of Neural Networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; pp. 39–57. [Google Scholar]
12. Brendel, W.; Rauber, J.; Bethge, M. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. In Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018; pp. 1–29. [Google Scholar]
13. Brunner, T.; Diehl, F.; Le, M.T.; Knoll, A. Guessing Smart: Biased Sampling for Efficient Black-Box Adversarial Attacks. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 4957–4965. [Google Scholar]
14. Hayes, J.; Danezis, G. Learning Universal Adversarial Perturbations with Generative Models. In Proceedings of the 2018 IEEE Symposium on Security and Privacy Workshops, SPW 2018, San Francisco, CA, USA, 24–24 May 2018; pp. 43–49. [Google Scholar]
15. Guo, C.; Gardner, J.; You, Y.; Wilson, A.G.; Weinberger, K. Simple Black-box Adversarial Attacks. In Proceedings of the 36th International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019; pp. 4410–4423. [Google Scholar]

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details