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 8, August 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.524

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

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

Impact Factor 8.524

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

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308085|

Black-Box Attacks on Visual Model Using IoT

Monisha D, Mr.R.Praveenkumar, Dr.J.Chandramohan, Mr.M.Sundaraperumal,

M. E Embedded System Technology, Gnanamani College of Technology, Namakkal, Tamil Nadu, India.

Assistant Professor, Department of EEE, M. E Embedded System Technology, Gnanamani College of Technology, Namakkal, Tamil Nadu, India

ASP, Department of EEE, M. E Embedded System Technology, Gnanamani College of Technology, Namakkal, Tamil Nadu, India

AP, Department of EEE, M. E Embedded System Technology, Gnanamani College of Technology, Namakkal, Tamil Nadu, India

ABSTRACT: The practical utility of black-box adversarial attacks in deep learning security has gained significant attention. However, current black-box attacks face challenges related to overfitting, leading to limited transferability of adversarial samples. This limitation arises from excessive iterations, limited input diversity during sample generation, and indiscriminate perturbation addition. To address these issues, this paper proposes a novel approach. Adversarial perturbations are generated using a selfsupervised model with strong generalization capabilities, resulting in enhanced transferability of adversarial samples. Furthermore, dynamic perturbation range maps are generated based on nonlinear characteristics of human eye luminance perception. A random gamma transform is also introduced to increase input diversity and mitigate overfitting. Extensive experiments on ImageNet datasets demonstrate the significant improvement in adversarial sample transferability achieved by our method. Moreover, our approach can be effectively combined with other techniques, achieving an average success rate of 91.70% against normal models and 74.90% against defensive models when using solely normally trained models.

KEYWORDS: Intrusion Detection System, Machine Learning, Iot

I. INTRODUCTION

Deep Neural Networks (DNNs) are by now widespread in industry and society as a whole, finding application in uncountable fields, from the movie industry, to autonomous driving, humanoid robots, video surveillance, and so on. Well trained DNNs largely outperform conventional systems, and can compete with human experts on a large variety of tasks. In particular, there has been a revolution in all vision-related tasks, which now rely almost exclusively on deep-learning solutions, starting from the 2012 seminal work of Krizhevsky et al. [1] where state-of-the-art image classification performance was achieved with a convolutional neural network (CNN).

Recent studies [2], however, have exposed some alarming weaknesses of DNNs. By injecting suitable *adversarial noise* on a given image, a malicious attacker can mislead a DNN into deciding for a wrong class, and even force it to output a *desired* wrong class, selected in advance by the attacker, a scenario described in Fig. 1. What is worse, such attacks are extremely simple to perform. By exploiting backpropagation, one can compute the gradient of the loss with respect to the input image, and build effective adversarial samples by gradient ascent/descent methods. Large loss variations can be induced by small changes in the image, ensuring that adversarial samples keep a good perceptual quality.

Following [2], several attacks (as well as defenses, not mentioned here for brevity) have been proposed, mostly gradient-descent methods with the gradient estimated through backpropagation, from the early Fast Gradient Sign Method (FGSM) [3], and its iterative version (I-FGSM) [4], to more recent and sophisticated methods [5], [6], [7]. All such methods, however, require perfect knowledge of the network architecture and weights, a *white box* scenario which is hardly encountered in real-world applications. The focus is therefore shifting towards *black-box* attacks, where nothing is known in advance about the network structure, its weights, or the dataset used for training. In this scenario,

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

Impact Factor 8.524

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

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308085|

the attacker can query the network at will and observe the outcome. This latter can be just a hard label, the distribution of probabilities across the classes (confidence levels), or even a feature vector.

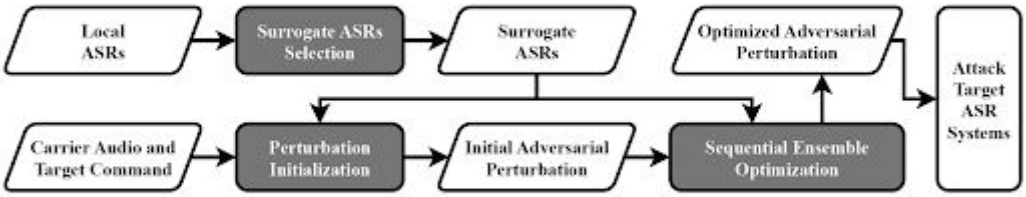

Fig 1: Block Diagram

There are many ways to perform black-box attacks to classifiers. A popular approach is to train a surrogate network to mimic the behavior of the target network [5], [8], [9]. The attack is performed on the surrogate and then transferred to the target. However, this calls for full knowledge of the target training set, a precious information rarely available in practice. A more viable alternative, followed here, is to use again gradient-descent [10], with the gradient now estimated by means of suitable queries to the network.

Black-box adversarial attacks should satisfy three main requirements at the same time: being effective, fast, and inconspicuous. The first requirement needs no comment, as it is the primary goal of the attack. Efficiency (low number of queries) is also necessary to prevent the attack from becoming exceedingly slow. As for the third requirement, it is important in all systems with human-in-the-loop, such as semiautomatic biometric recognition systems. Unnatural, low quality images may be readily identified by dedicated statistical tests and sent to visual inspection, to be eventually easily detected.

To meet all these requirements, we propose a novel, perceptual quality-preserving (PQP), black-box attack, where quality and effectiveness are ensured in advance by a judicious choice of the perturbation path. Let us briefly summarize the major features and innovative contributions of our proposal: *(i)* adversarial noise is injected only in "safe" regions where it has low impact on image quality and high impact on decisions; *(ii)* safe regions are identified based on the local gradient of a perceptual quality measure, like the structural similarity index (SSIM) or the visual information fidelity (VIF) *(iii)* all perturbations and queries are 8-bit integer, to successfully attack systems that accept only popular integer formats, such as PNG or JPG.

II. BACKGROUND WORK

Image classification models are widely used in real-world applications and typically achieve top-5 accuracy of over 90%. Cloud-based image classification services, such as Google Cloud Vision, provide pre-trained models as APIs, allowing users to classify images by sending requests to cloud servers. This is particularly useful for IoT edge devices that lack the computational power to run deep learning models locally. However, image classification cloud services are vulnerable to black-box adversarial attacks, which generate imperceptible perturbations to input images in order to mislead classification models into incorrect predictions. While prior research has shown that black-box attacks can achieve high success rates of over 95% with only 1,000 queries without access to model structure and weights [1], most research has generated adversarial images offline on local models rather than online on cloud servers. The efficiency of online black-box attacks against cloud services remains unclear.

Most prior research test their attacks on local models, rather than on Cloud APIs, since it is both faster and less costly. Our experimental results reveal that these attacks achieve significantly lower success rates when attacking cloud

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

Impact Factor 8.524

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

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308085|

services. Image Encoding: In real-world scenarios, images are typically encoded before being sent to cloud services to reduce the amount of data transmitted and to save bandwidth (see Fig. 2). However, prior research often assumes that perturbations can be added directly to the raw input image (see Fig. 1). It is worth noting that cloud services such as Google Cloud Vision and Imagga accept raw binary and base64 encoded JPEG (lossy compression) images as input. This compression may cause part of the perturbations to disappear, ultimately reducing the success rate of attacks. Therefore, if we evaluate black-box attacks on local models without considering image encoding and quantization, we may overestimate the effectiveness of these attacks against real-world cloud APIs. Besides, image classification cloud services do not accept images with invalid pixel values. For example, the bandit attack does not clip the image while estimating gradients, assuming they can send invalid images (pixel value > 255 or < 0) to the black-box model.

The task of generating adversarial inputs can be approached as a problem of selecting what pixels to attack. Thus, we can use existing local search methods to search for combinations of pixels to be perturbed. One simple, yet effective, baseline attack that use this idea is the simple black-box attack (SimBA) [4]. With SimBA, a vector is randomly sampled from a predefined orthonormal basis and then added or subtracted from the image. To improve sample efficiency, Andriushchenko et al. proposed the square attack [7]. This attack initializes the perturbation using vertical stripes, since CNNs are sensitive to high-frequency perturbations [16], and then generates square-shaped perturbations at random locations to deviate model classifications.

III. METHODS

Both local (multi-core CPU) and cloud-based (load balancer) deep learning models are vulnerable to distributed queries, but the challenge lies in deciding which queries to send concurrently. To address this problem, we propose horizontal and vertical distribution for black-box attacks, inspired by horizontal and vertical scaling of cloud resources [22]. Horizontal distribution concurrently sends queries for different images within the same iteration, thereby allowing the generation of multiple adversarial examples concurrently. This can be achieved without altering existing black-box attack methods. Since horizontal distribution does not require significant modifications to the original attack method, we can apply horizontal distribution by implementing a distributed query function that sends concurrent requests to cloud APIs. Vertical distribution, on the other hand, sends multiple concurrent queries for the same image, thereby accelerating the attack for that particular image. Existing black-box attack methods need to be redesigned to decouple the queries across iterations.

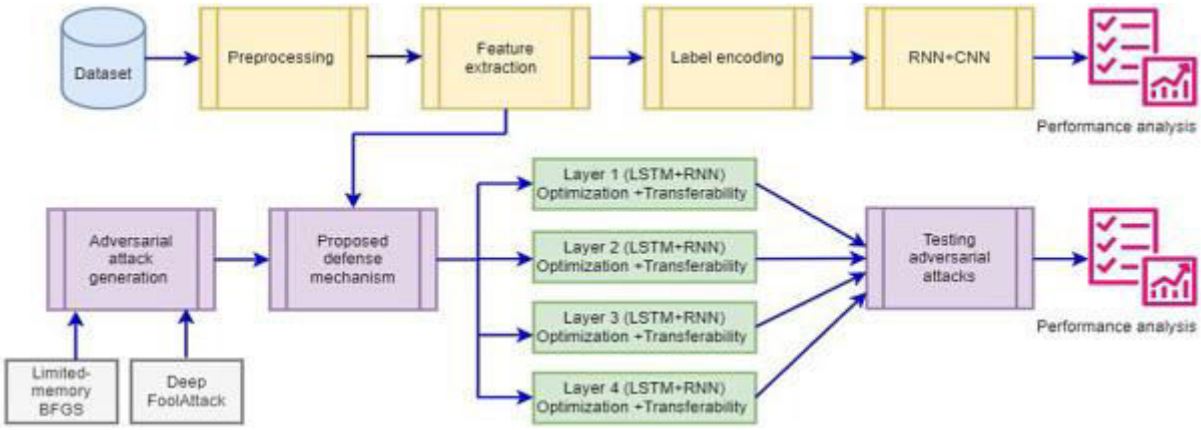

Fig 2: proposed Work

We further evaluate the impact of horizontal distribution on the total attack time. As shown in Fig. 8, horizontal distribution significantly reduces the total attack time for all the attacks. By sending concurrent queries across multiple images, we can achieve a successful attack faster without increasing the query budget. The acceleration ratio depends on the number of computational nodes deployed for the cloud API. Deploying more nodes behind the load balancer can lead to a larger acceleration ratio. In summary, our experiments showed that horizontal distribution reduced the total attack time by a factor of five on average and did not significantly affect the attack success rate and the average number of queries, which brings black-box attacks a step closer to be a practical threat.

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

Impact Factor 8.524

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

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308085|

IV. RESULT ANALYSIS

We proposed a simple method for generating UAPs under the black-box condition inspired by the SimBA method and applied this method to DNN-based medical image classification. Overall, the results showed that small (almost imperceptible) UAPs are generatable under the black-box condition using only a simple hill-climbing search based on the model outputs (i.e., confidence scores) to perform both nontargeted and targeted attacks. The vulnerability of black-box UAPs was observed in several model architectures, indicating the versatility of our method; thus, the vulnerability may be a general property of a DNN. The results indicate that adversaries can foil or control DNN tasks even if they never use model parameters (e.g., loss gradients) and training data because target systems are operated under black-box conditions in terms of security. DNN-based medical image diagnoses may be easy to conduct in a realistic environment.

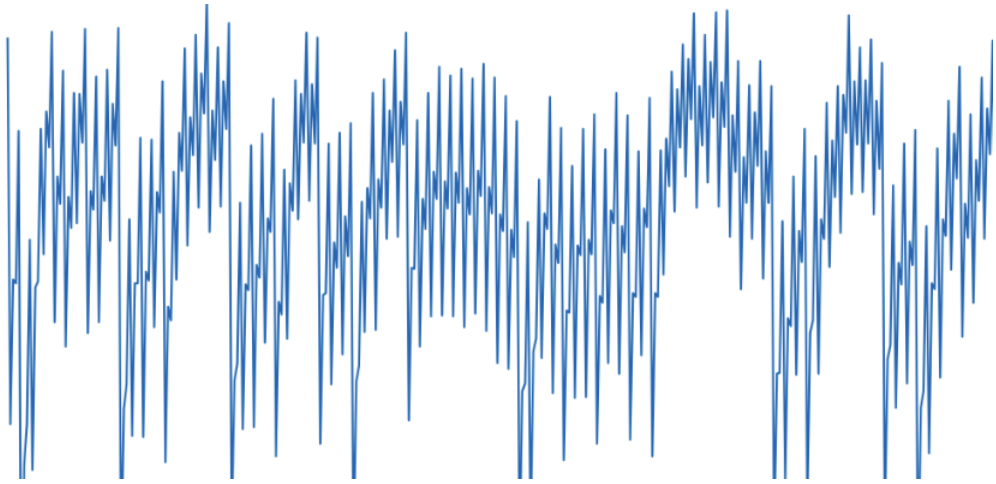

Fig 3: IOT Result Analysis

Nevertheless, we propose a method for generating UAPs using a simple hill-climbing search based only on DNN outputs to demonstrate that UAPs are easily generatable using a relatively small dataset under black-box conditions with representative DNN-based medical image classifications. Black-box UAPs can be used to conduct both nontargeted and targeted attacks. Overall, the black-box UAPs showed high attack success rates (40–90%). The vulnerability of the black-box UAPs was observed in several model architectures. The results indicate that adversaries can also generate UAPs through a simple procedure under the black-box condition to foil or control diagnostic medical imaging systems based on DNNs, and that UAPs are a more serious security threat.

V. CONCLUSION

Our research presents a novel approach to accelerating black-box attacks against online cloud services by exploiting load balancing. We introduce two general frameworks, horizontal and vertical distribution, that can be applied to existing black-box attacks to significantly reduce the total attack time. To validate the efficiency of the frameworks, we conduct experiments using three commonly used image classifiers and three commonly used black-box attacks. Additionally, we contribute to the research community by open-sourcing our image classification cloud service, DeepAPI. This will enable future research on distributed black-box attacks and provide insights into the practical threats posed by adversarial attacks against machine learning models deployed on cloud servers.

REFERENCES

[1] S. Bhambri, S. Muku, A. Tulasi, and A. B. Buduru, “A survey of black-box adversarial attacks on computer vision models,” arXiv preprint arXiv:1912.01667, 2019.
[2] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, “Black-box adversarial attacks with limited queries and information,” in International Conference on Machine Learning (ICML), 2018, pp. 2137–2146.

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

Impact Factor 8.524

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

Volume 13, Issue 8, August 2024

|DOI: 10.15680/IJIRSET.2024.1308085|

[3] A. Ilyas, L. Engstrom, and A. Madry, "Prior convictions: Blackbox adversarial attacks with bandits and priors," arXiv preprint arXiv:1807.07978, 2018.
[4] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, "Simple black-box adversarial attacks," in International Conference on Machine Learning (ICML), 2019, pp. 2484–2493.
[5] T. Brunner, F. Diehl, M. T. Le, and A. Knoll, "Guessing smart: Biased sampling for efficient black-box adversarial attacks," in IEEE International Conference on Computer Vision (ICCV), 2019.
[6] S. Moon, G. An, and H. O. Song, "Parsimonious black-box adversarial attacks via efficient combinatorial optimization," in Proceedings of the 36th International Conference on Machine Learning (ICML), vol. 97, 2019, pp. 4636–4645.
[7] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: a query-efficient black-box adversarial attack via random search," in European Conference on Computer Vision (ECCV), 2020, pp. 484– 501.
[8] H. Pham, Z. Dai, Q. Xie, and Q. V. Le, "Meta pseudo labels," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11 557–11 568.
[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2016, pp. 770–778.
[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.
[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2016, pp. 2818–2826

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