Depth, Breadth, and Complexity: Ways to Attack and Defend Deep Learning Models

Firuz Juraev
Sungkyunkwan University
Suwon, South Korea
fjuraev@g.skku.edu

Eldor Abdukhamidov
Sungkyunkwan University
Suwon, South Korea
abdukhamidov@skku.edu

Mohammed Abuhamad[†]
Loyola University Chicago
Chicago, United States
mabuhamad@luc.edu

Tamer Abuhmed[†]
Sungkyunkwan University
Suwon, South Korea
tamer@skku.edu

ABSTRACT

Deep Learning is rapidly evolving to the point that it can be used in crucial safety and security applications, including self-driving vehicles, surveillance, drones, and robots. However, these deep learning models are vulnerable to attacks based on adversarial samples that are undetectable to the human eye but cause the model to misbehave. There is an increasing demand for comprehensive and in-depth analysis of behaviors of various attacks and the possible defenses against common deep learning models under several adversarial scenarios. In this study, we conducted four separate investigations. First, we examine the relationship between the model's complexity and its robustness against the studied attacks. Second, the connection between the performance and diversity of models is examined. Third, the first and second experiments were tested across different datasets to explore the impact of the dataset on the performance of the model. Four, throughout the defense strategies, the model behavior is extensively investigated. The code, trained models, and detailed settings and results are available at: *https://github.com/InfoLab-SKKU/ML-Adversarial-Attacks-Analysis*.

CCS CONCEPTS

• **Security and privacy** → *Software and application security*;

KEYWORDS

Deep Learning, Adversarial Attacks, Defenses, Computer Vision

ACM Reference Format:
Firuz Juraev, Eldor Abdukhamidov, Mohammed Abuhamad[†], and Tamer Abuhmed[†]. 2022. Depth, Breadth, and Complexity: Ways to Attack and Defend Deep Learning Models. In *Proceedings of the 2022 ACM Asia Conference on Computer and Communications Security (ASIA CCS '22), May 30-June 3, 2022, Nagasaki, Japan.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3488932.3527278

1 INTRODUCTION

Deep Learning has grown into a powerful tool that can be used to address a wide range of complicated learning tasks that were previously unattainable to tackle using conventional machine learning approaches. In recent years, Deep Learning (DL) has achieved significant progress in the traditional disciplines of image classification, voice recognition, and language translation, thanks to the emergence of Deep Neural Network (DNN) models and the availability of high-performance resources to train complicated models [2, 3, 5]. As DNN has evolved from experimental settings to real-world applications, security and privacy concerns have become a major issue of deploying the DL models. After the findings of Szegedy *et al.* [6], several seminal works on the practicality of adversarial attacks on DL have appeared in the research community [1, 3, 7]. Existing adversarial attacks can be categorized into three categories: white-box, gray-box, and black-box based on the adversaries' knowledge assumptions about the target model [4].

This study concentrated on evaluating a few selected attack methods from these attacks categories (Figure 1), such as Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini and Wagner (C&W) white-box attacks. Simple Black Box Attack (SimBA), hopskipjump, and boundary are selected from the black-box attacks. These attacks were comprehensively evaluated within the frameworks of various threat models. Also, the study evaluated the effectiveness of several defensive strategies for adversarial attacks, such as the preprocessor, trainer, and detector. In this study, we applied three common preprocessor defenses: Bit squeezing, Median smoothing, JPEG filter.

Contributions. In this work, we investigate various characteristics of the attacks and defenses on the well-known deep learning models, including **VGG**, **ResNet**, **DenseNet** families with a different number of layers using the various datasets such as ImageNet, CIFAR-10, and CIFAR-100. Furthermore, we examined the attacks and defenses approaches on diverse models like **Xeption**, **InceptionV3**, **MobileNetV2**, and **GoogLeNet**.

In this work, we raise and answer the following research questions. ❶ *Is the complexity of the models a factor in the attack's success/failure rate?* To answer this question, we investigated the model robustness by increasing the number of layers in three types of DL models (**ResNet**, **VGG**, **DenseNet**). ❷ *When faced with an adversarial example, is there a relationship between model diversity*

[†] corresponding author.

ASIA CCS '22, May 30-June 3, 2022, Nagasaki, Japan

ACM ISBN 978-1-4503-9140-5/22/05.
https://doi.org/10.1145/3488932.3527278

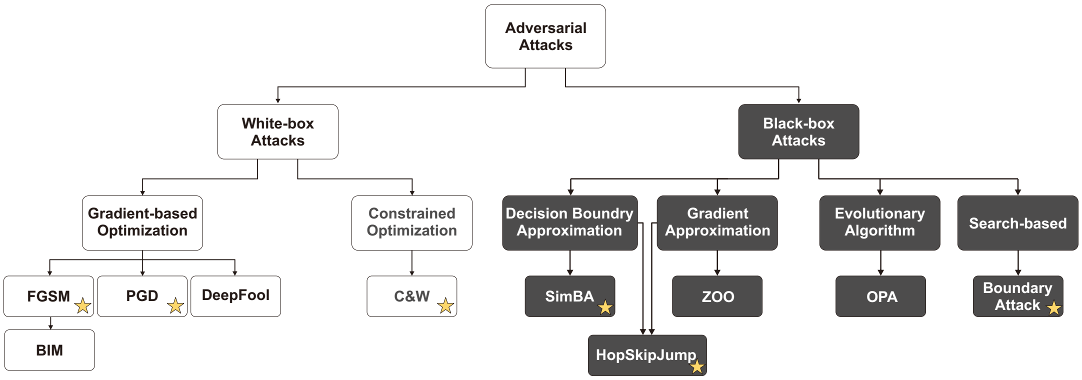


Figure 1: A taxonomy of adversarial attacks. The stars mark the attacks considered in this work.

and robustness? To answer this question, the attacks were tested on seven diverse DL models. ❸ *How do attacks behave across different datasets?* To answer this question, we conducted experiments on three well-known datasets. ❹ *How do attacks behave when there are defenders?* To answer this question, three preprocessing defense strategies were examined against the attacks.

Organization. The following outlines this paper: In Section 2, we discuss about the dataset and studied models, Section 3 highlights our observations and in Section 4 we conclude our study.

2 DATASET AND MODELS

2.1 Dataset

We employed three common datasets for all experiments: ImageNet, CIFAR-10, and CIFAR-100. The ImageNet dataset contains 14 million samples of images with a size of 224x224 pixels for 1000 classes. The CIFAR-10 dataset contains 60,000 samples with a size of 32x32 pixels, distributed over 10 classes, while the CIFAR-100 dataset has the same number of samples distributed over 100 classes.

2.2 Models

For our experiment 1, we used 12 models for each dataset from three families (**ResNet**, **VGG**, **DenseNet**) with a different number of layers. For experiment 2, we utilized seven diverse models for each dataset, including **GoogLeNet**, **InceptionV3**, **Xception**, and **MobileNet V2**. For the ImageNet dataset, we used pre-trained models on the PyTorch framework. For CIFAR-10 and CIFAR-100, we trained models on our servers.

3 PRELIMINARY EXPERIMENTAL RESULTS

In our experiments, 1000 test images were examined for each white-box attack. For black-box attacks, 200 test images were selected due to the slow performance of black-box attacks.

3.1 EXP 1: Model Complexity and Robustness

In Experiment 1, we tested the hypothesis that the attack success rate is affected by the complexity of the model on ImageNet (Figure 2). In Figure 2 (a), using the PGD white-box attack, it is shown that the attack failure increases as the model layers increase, which indicates that the complex model is more robust compared with the shallow models. For example, the ResNet152 model and DenseNet201 are more difficult to deceive than other models.

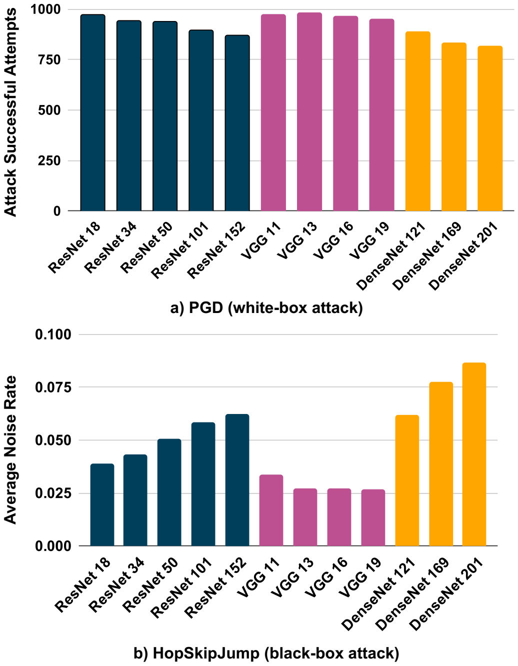


Figure 2: (a) Shows how as the number of layers rises, the attack successful attempts decreases. (b) Shows that as the number of layers grows, the amount of noise perturbed ascents. The amount of noise is described by the pixel wise difference (adversarial and benign sample).

In Figure 2 (b), using the HopSkipJump black-box attack, we observed that as the number of layers of the model increases, the complexity of the model increases, and the attacks require more noise to succeed. For example, succeeding against DenseNet201 requires the largest amount of noise.

3.2 EXP 2: Model Diversity and Robustness

In Experiment 2, we tested all selected white-box attacks, *i.e.,* FGSM, PGD, and C&W on diverse models (*i.e.,* structure and parameters included), including Resnet152, VGG19, DenseNet161, Xeption, InceptionV3, MobileNetV2, and GoogLeNet. From the experimental results, it became obvious that the number of parameters on models does not play an important role in model robustness (Figure 3).

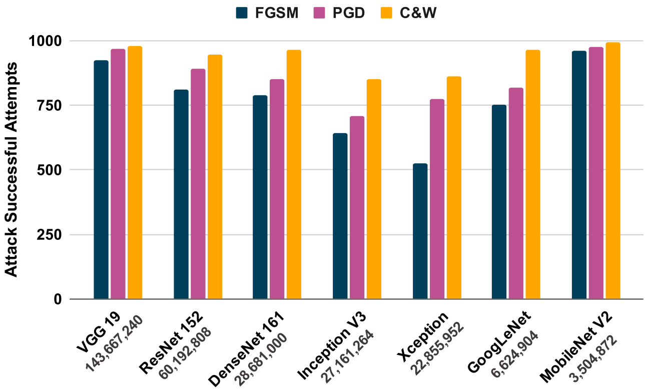
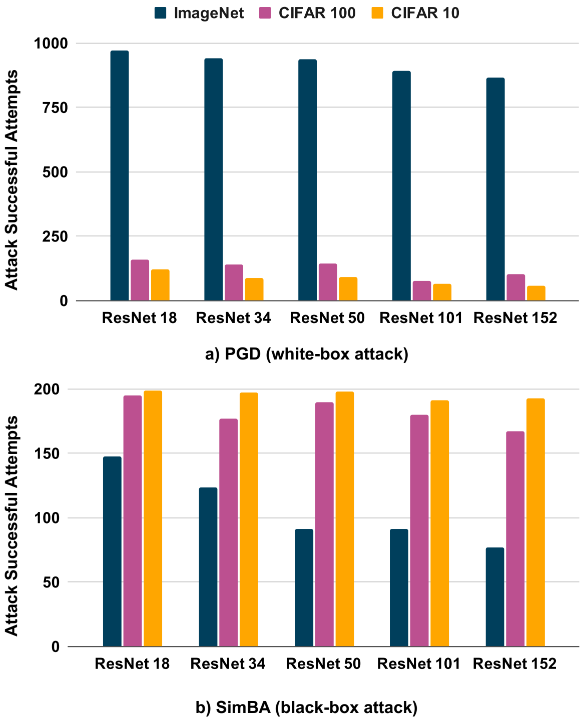


Figure 3: It represents a wide range of DNNs for determining the relationship between the number of large parameters and the model's robustness against white-box attacks.



a) PGD (white-box attack)

b) SimBA (black-box attack)

Figure 4: Attack successful attempts of white-box and black-box attacks on different dataset.

3.3 EXP 3: Attacks across Different Datasets

In Experiment 3, we tested experiments 1 and 2 using different datasets (*e.g.,* CIFAR-10 and CIFAR-100). In white-box attacks, due to less number of classes on CIFAR-10 and CIFAR-100, the attack success rate dropped significantly in all white-box attacks except for C&W attack (Figure 4 (a) showing the PGD attack successful attempts). However, in the black-box setting (*e.g.,* SimBA), we observed that the number of classes has a minor influence on attack success rate (Figure 4 (b)). However, the size of the input image has a significant impact on the success rate of black-box attacks.

3.4 EXP 4: Defenses against Adversarial Attacks

In experiment 4, three prepossessing defense strategies were investigated such the bit squeezing, median smoothing, and JPEG filter. The defense techniques modify the input image but do not have any impact on the model. We used the defenses with weak parameters to maintain the benign confidence. In white-box attacks like FGSM and PGD attacks, the applied defenses could not drop the attack success rate much due to high misclassification confidence and the amount of perturbed noise. However, for the C&W attack, the attack success rate decreased significantly after applying defenses with even weak parameters. The reason for this drop is that the C&W attack deceives models with a small amount of noise. There is a trade-off between the noise amount and attack success rate.

In the case of a black-box attack, experiment 4 showed that it is easy to defend the models with weak parameters due to the low misclassification confidence of the attacks.

4 CONCLUSION

This work presents the preliminary results of a large-scale study on the impact of various adversarial attacks and defenses on different models across different datasets. Our experiments proved many assumptions about the behaviors of attacks and defenses. Furthermore, we found and analyzed several weaknesses of the attacks. So far in the literature, there have been many review papers about the behaviors of adversarial attacks and defenses. However, most of them are theoretical. In our work, we showed all those behaviors with comprehensive experiments. Our observations proved and rejected many assumptions. A promising future work will include greybox attacks, other types of defenses. In addition, we will compare the robustness of different architectures other than CNNs like Vision Transformers.

Acknowledgment. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2021R1A2C1011198).

REFERENCES

[1] Eldor Abdukhamidov, Mohammed Abuhamad, Firuz Juraev, Eric Chan-Tin, and Tamer AbuHmed. 2021. AdvEdge: Optimizing Adversarial Perturbations Against Interpretable Deep Learning. In *International Conference on Computational Data and Social Networks*. Springer, Cham, 93–105.

[2] Ahmed Abusnaina, Mohammed Abuhamad, Hisham Alasmary, Afsah Anwar, Rhongho Jang, Saeed Salem, Daehun Nyang, and David Mohaisen. 2021. DL-FHMC: Deep Learning-based Fine-grained Hierarchical Learning Approach for Robust Malware Classification. *IEEE Transactions on Dependable and Secure Computing* (2021).

[3] Hisham Alasmary, Ahmed Abusnaina, Rhongho Jang, Mohammed Abuhamad, Afsah Anwar, DaeHun Nyang, and David Mohaisen. 2020. Soteria: Detecting adversarial examples in control flow graph-based malware classifiers. In *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*. 888–898.

[4] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2021. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology* 6, 1 (2021), 25–45.

[5] Jaiteg Singh, Deepak Thakur, Tanya Gera, Babar Shah, Tamer Abuhmed, and Farman Ali. 2021. Classification and Analysis of Android Malware Images Using Feature Fusion Technique. *IEEE Access* 9 (2021), 90102–90117.

[6] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations, ICLR 2014*. 1–10.

[7] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. 2019. Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE Transactions on Neural Networks and Learning Systems* 30, 9 (2019), 2805–2824.