International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

OPEN ACCESS

Available Online at : www.ijsrcseit.com
doi : https://doi.org/10.32628/CSEIT2410229

Pixel Map Analysis Adversarial Attack Detection on Transfer Learning Model

Soni Kumari[1], Dr. Sheshang Degadwala[2]

[1]Research Scholar, Department of Computer Engineering, Sigma Institute of Engineering, Gujarat, India
[2]Professor & Head of Department, Department of Computer Engineering, Sigma University, Gujarat, India

ARTICLEINFO

Article History:

Accepted: 15 March 2024
Published: 30 March 2024

Publication Issue
Volume 10, Issue 2
March-April-2024

Page Number
350-357

ABSTRACT

Adversarial attacks pose a significant threat to the robustness and reliability of deep learning models, particularly in the context of transfer learning where pre-trained models are widely used. In this research, we propose a novel approach for detecting adversarial attacks on transfer learning models using pixel map analysis. By analyzing changes in pixel values at a granular level, our method aims to uncover subtle manipulations that are often overlooked by traditional detection techniques. We demonstrate the effectiveness of our approach through extensive experiments on various benchmark datasets, showcasing its ability to accurately detect adversarial attacks while maintaining high classification performance on clean data. Our findings highlight the importance of incorporating pixel map analysis into the defense mechanisms of transfer learning models to enhance their robustness against sophisticated adversarial threats.

Keywords : Adversarial Attacks, Transfer Learning, Pixel Map Analysis, Pre-Trained Models, Defense Mechanisms, Classification Performance, Benchmark Datasets.

I. INTRODUCTION

Deep learning models have achieved remarkable success across various domains, revolutionizing tasks such as image classification, object detection, and natural language processing. However, these models are susceptible to adversarial attacks, where carefully crafted perturbations are introduced to input data, leading to incorrect predictions. Adversarial attacks have raised significant concerns about the reliability and robustness of deep learning systems, especially in real-world applications where security and trust are paramount. Transfer learning, a technique that leverages pre-trained models to improve learning on new tasks, has become increasingly popular due to its ability to achieve high performance with limited data. However, the vulnerability of transfer learning models to adversarial attacks remains a critical issue that requires effective mitigation strategies.

One promising approach for addressing the vulnerability of transfer learning models to adversarial attacks is through pixel map analysis. Traditional defense mechanisms often focus on high-level features or gradients, overlooking subtle changes in pixel values that can be exploited by adversaries. Pixel map analysis delves into the fine-grained details of image data, examining alterations in pixel intensities and spatial relationships. By scrutinizing pixel maps, it becomes possible to detect adversarial perturbations that may go undetected by conventional methods. This granular analysis offers a more comprehensive understanding of how adversarial attacks manifest in the pixel space, enabling more robust detection and mitigation strategies.

In this paper, we present a detailed investigation into the efficacy of pixel map analysis for detecting adversarial attacks on transfer learning models. We explore the theoretical foundations of pixel map analysis, discussing its advantages over existing detection techniques. Furthermore, we conduct extensive experiments on benchmark datasets to evaluate the performance of our proposed approach. By combining pixel map analysis with transfer learning, we aim to enhance the robustness and security of deep learning models against adversarial threats, contributing to the development of more reliable and trustworthy AI systems.

II. LITERATURE STUDY

Ryu and Choi [1] proposed a novel approach for detecting adversarial attacks based on differences in image entropy. By analyzing variations in the entropy of images, their method aims to identify subtle changes introduced by adversarial perturbations. This study highlights the importance of considering image characteristics for effective adversarial attack detection, offering a promising avenue for enhancing the robustness of deep learning models.

Cui [2] focused on targeting image-classification models, emphasizing the need for tailored detection and mitigation strategies specific to different types of models and tasks. The study delves into the intricacies of adversarial attacks on image classifiers, highlighting challenges and potential solutions for improving model security and reliability. This work contributes to the growing body of research aimed at fortifying deep learning models against sophisticated adversarial threats.

Kim and Yun [3] introduced AEGuard, an innovative image feature-based independent adversarial example detection model. By leveraging image features, AEGuard aims to detect adversarial examples with greater accuracy and efficiency, showcasing advancements in detection techniques. This approach underscores the importance of developing robust defense mechanisms to safeguard deep learning models from adversarial manipulations.

Lorenz, Keuper, and Keuper [4] explored the concept of unfolding local growth rate estimates for (almost) perfect adversarial detection. Their study emphasizes the significance of fine-grained analysis and local feature examination for detecting adversarial perturbations effectively. By focusing on local growth rates, this research contributes valuable insights into improving the resilience of deep learning models against adversarial attacks.

Shi, Liao, and He [5] proposed a noise-fusion method to defend against adversarial attacks on DNN image classification models. Their approach involves integrating noise into the input data to disrupt adversarial perturbations, enhancing model robustness. This study showcases the potential of noise-based defenses in mitigating adversarial threats, highlighting the importance of exploring diverse defense strategies for safeguarding deep learning systems.

Almuflih et al. [6] introduced a novel exploit feature-map-based detection method for identifying adversarial attacks. By analyzing feature maps, their approach aims to uncover patterns indicative of adversarial perturbations, enhancing the detection accuracy of deep learning models. This study underscores the importance of leveraging internal model representations for effective adversarial attack detection and mitigation.

Khan et al. [7] conducted a detailed analysis of Alpha Fusion adversarial attacks using deep learning techniques. Their study provides insights into the characteristics and impact of Alpha Fusion attacks, contributing to the understanding of diverse adversarial strategies. This research is valuable for developing targeted defenses against specific types of adversarial threats in deep learning systems.

Ghaffari Laleh et al. [8] explored adversarial attacks and defenses in computational pathology, highlighting the vulnerabilities of machine learning models in medical applications. Their study emphasizes the critical need for robustness in healthcare-related AI systems, as adversarial attacks could have serious implications for diagnostic accuracy and patient care. This research motivates the development of resilient AI solutions in medical imaging and pathology analysis.

Wang et al. [9] presented a comprehensive survey on adversarial attacks and defenses in machine learning-powered networks. Their study provides an overview of current techniques, challenges, and advancements in the field of adversarial robustness. This survey serves as a valuable resource for researchers and practitioners working on improving the security and reliability of machine learning systems.

Hirano, Minagi, and Takemoto [10] investigated universal adversarial attacks on deep neural networks for medical image classification. Their study highlights the potential vulnerabilities of medical AI systems to universal attacks that generalize across different models and datasets. Understanding such attacks is crucial for developing robust defenses in medical imaging applications.

Zheng and Velipasalar [12] proposed part-based feature squeezing as a method to detect adversarial examples in person re-identification networks. Their approach focuses on leveraging specific features to identify adversarial perturbations, contributing to the development of targeted detection techniques in computer vision systems.

Liang et al. [13] introduced adaptive noise reduction for detecting adversarial image examples in deep neural networks. By dynamically adjusting noise levels, their method aims to improve the robustness of models against adversarial attacks, highlighting the importance of adaptive defense mechanisms in maintaining model security.

Ahmadi, Dianat, and Amirkhani [14] developed an adversarial attack detection method based on re-attacking approaches. This study emphasizes the iterative nature of attack-defense strategies, where understanding past attacks can inform the development of more effective defense mechanisms against future threats.

Ren et al. [15] discussed adversarial attacks and defenses in deep learning, providing insights into the evolving landscape of adversarial robustness. Their work highlights the ongoing challenges and opportunities in mitigating adversarial threats across different domains, contributing to the broader understanding of adversarial machine learning.

III.PROPOSED SYSTEM

A flow diagram describing the process of an adversarial attack on an input image through convolutional neural networks (CNNs) such as AlexNet, VGGNet, or ResNet can be detailed as follows:

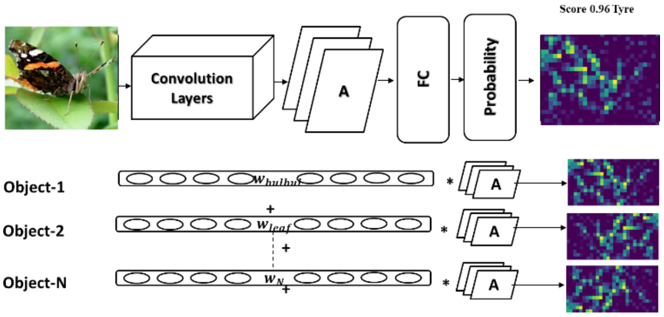

Figure 1. Proposed Pixel Map Analysis

Adversarial Attack Input Image

The process begins with an input image that is targeted for an adversarial attack. This image could be a benign, correctly classified image or an image specifically crafted to deceive the model.

Convolution Layers

The input image passes through convolutional layers of the chosen CNN architecture (e.g., AlexNet, VGGNet, ResNet). These layers perform feature extraction by convolving filters across the image to detect patterns and features at different levels of abstraction.

Activation Map:

After convolution, activation maps are generated, highlighting regions of the image that are activated by specific filters. These maps represent the response of the convolutional filters to different features in the input image.

If the input image contains multiple objects or classes, individual activation maps are generated for each object or class. These maps provide insights into how the network processes and perceives different objects within the image.

Fully Connected Layer:

The activation maps are then fed into fully connected layers, which perform classification based on the extracted features. These layers learn complex patterns and relationships in the feature maps to predict the probability of different classes or labels.

Probability of Class Predicted:

The output of the fully connected layer is a probability distribution over the classes or labels in the dataset. The model predicts the most likely class for the input image based on these probabilities.

Feature Map:

Alongside the classification process, feature maps are generated throughout the network. These maps represent the learned features and activations at different layers, providing insights into how the model processes information and makes predictions.

Attack Image/Non-Attack Image:

Finally, based on the classification result, an adversarial attack may modify the image to deceive the model into misclassifying it. The attack image is crafted to introduce imperceptible perturbations that alter the model's prediction without significantly changing the visual appearance to a human observer. Conversely, a non-attack image remains unchanged or may undergo legitimate transformations for data augmentation or preprocessing.

This flow diagram illustrates the sequential steps involved in processing an input image through convolutional layers, extracting features, performing classification, and potentially encountering an adversarial attack that manipulates the image to deceive the model's predictions.

Result Analysis

The CIFAR-10 dataset is a foundational resource in computer vision and machine learning, comprising 60,000 32x32 color images across 10 classes such as airplanes, automobiles, birds, cats, and others. Split into 50,000 training images and 10,000 test images, CIFAR-10 serves as a benchmark for tasks like image classification and object recognition. Its diverse set of images, representing real-world scenarios with varying backgrounds and object poses, makes it ideal for evaluating the performance and robustness of machine learning algorithms and deep learning models. Researchers and practitioners widely use CIFAR-10 to compare accuracy, test generalization capabilities, and develop cutting-edge techniques, cementing its status as a fundamental dataset in advancing computer vision research and applications.

Link:

https://www.cs.toronto.edu/~kriz/cifar.html

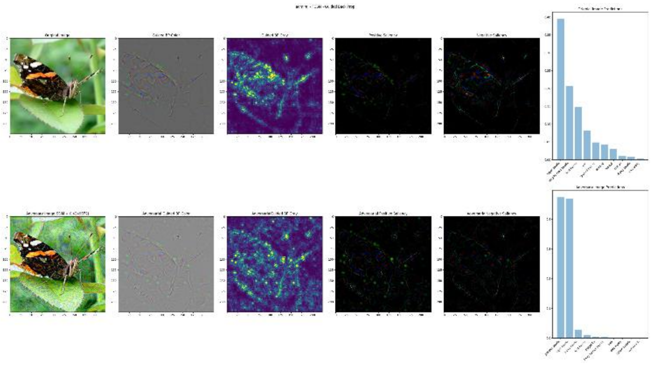

Figure 2. AlexNet Existing Grad-Cam Map of FGSM Attack

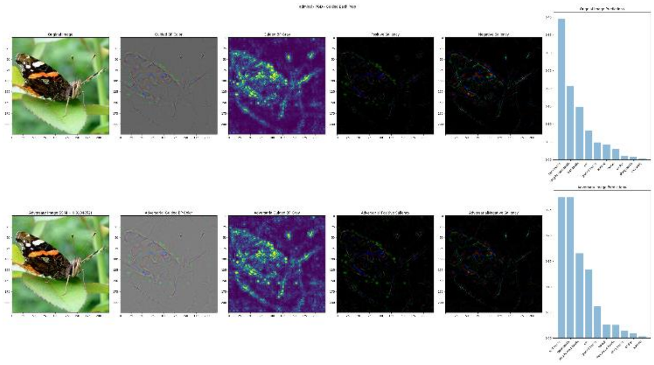

Figure 3. AlexNet Existing Grad-Cam Map of PGD Attack

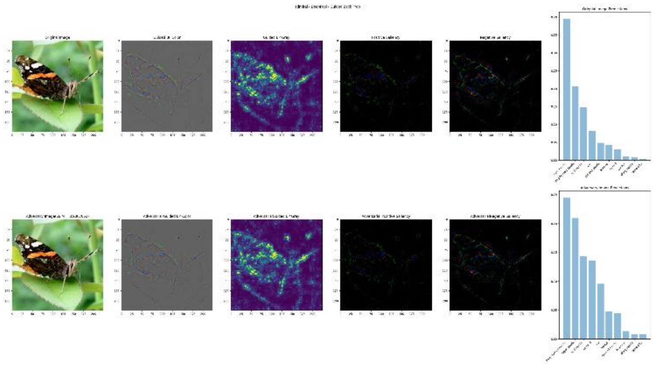

Figure 4. AlexNet Existing Grad-Cam Map of DeepFool Attack

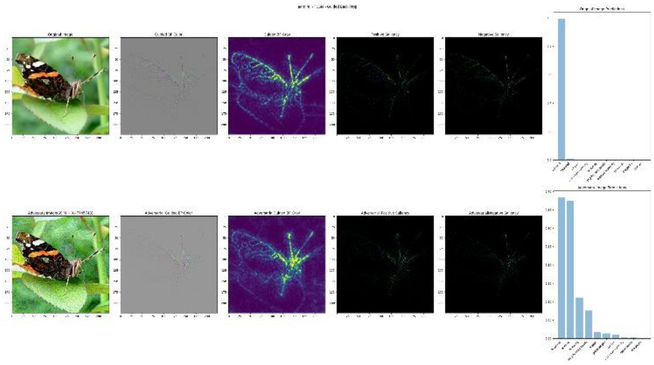

Figure 5. VggNet Existing Grad-Cam Map of FGSM Attack

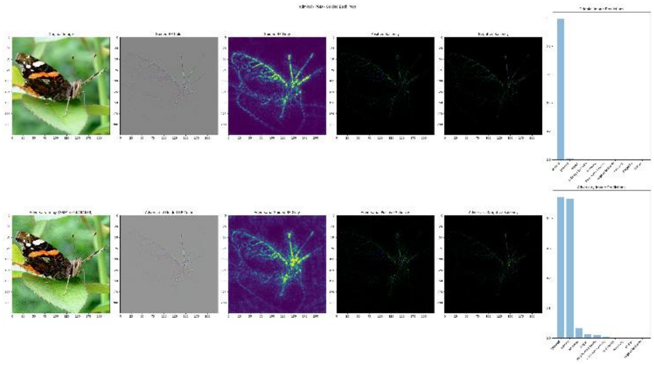

Figure 6. VggNet Existing Grad-Cam Map of PGD Attack

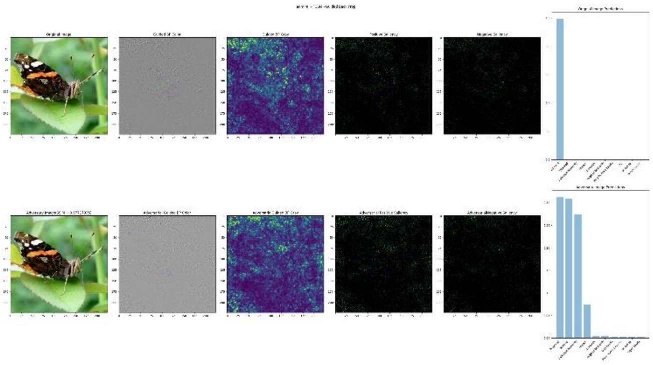

Figure 7. ResNet Existing Grad-Cam Map of FGSM Attack

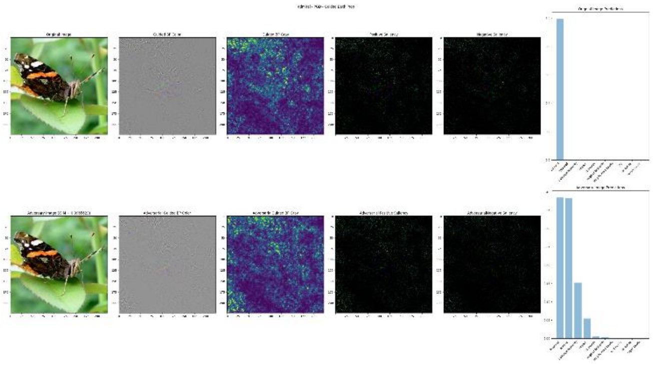

Figure 8. ResNet Existing Grad-Cam Map of PGD Attack

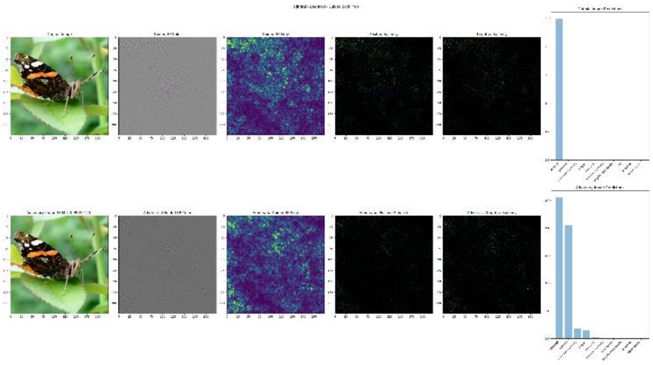

Figure 9. ResNet Existing Grad-Cam Map of DeepFool Attack

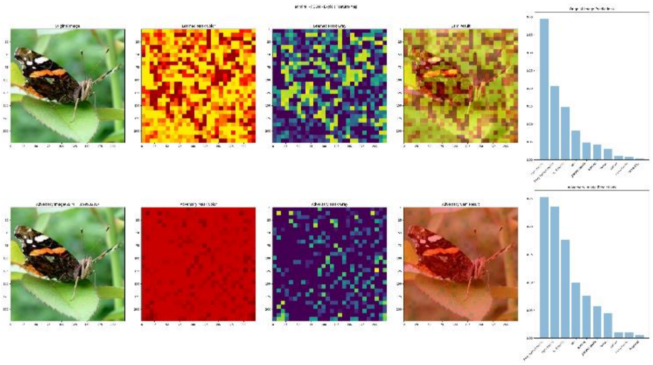

Figure 10. AlexNet Proposed Pixel-Map of FGSM Attack

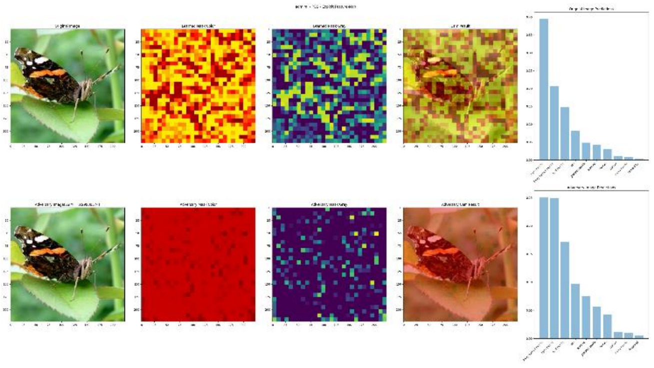

Figure 11. AlexNet Proposed Pixel-Map of PGD Attack

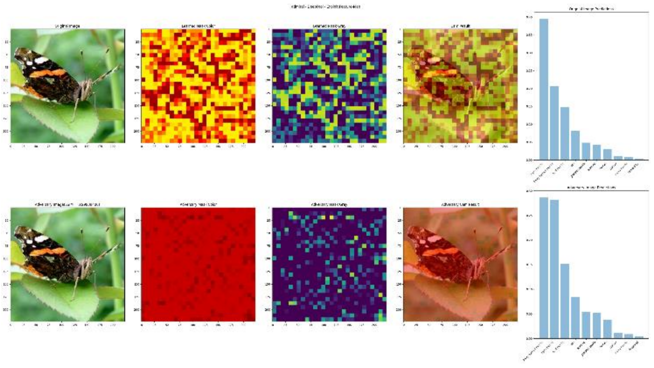

Figure 12. AlexNet Proposed Pixel-Map of DeepFool Attack

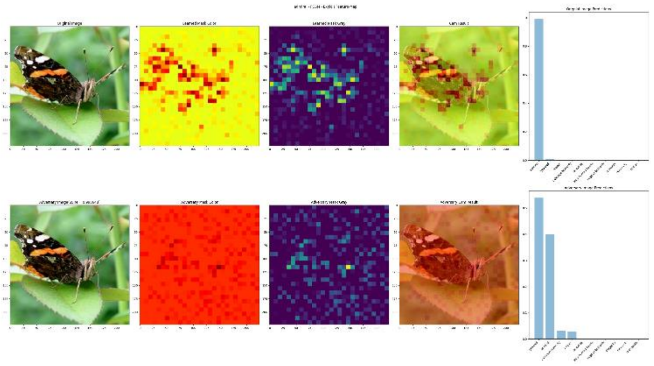

Figure 13. VggNet Proposed Pixel-Map of FGSM Attack

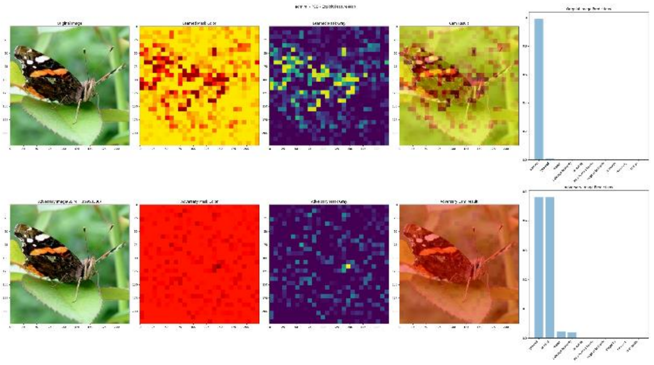

Figure 14. VggNet Proposed Pixel-Map of PGD Attack

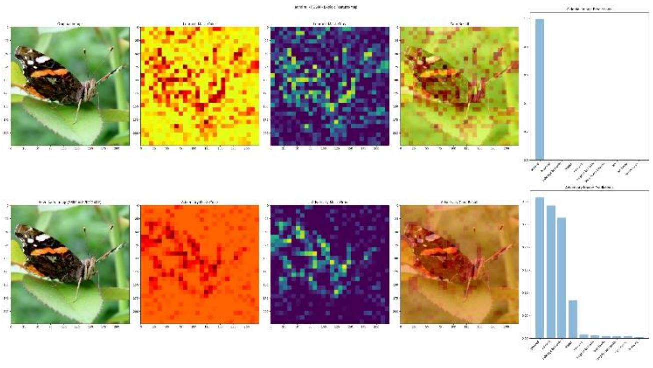

Figure 15. ResNet Proposed Pixel-Map of FGSM Attack

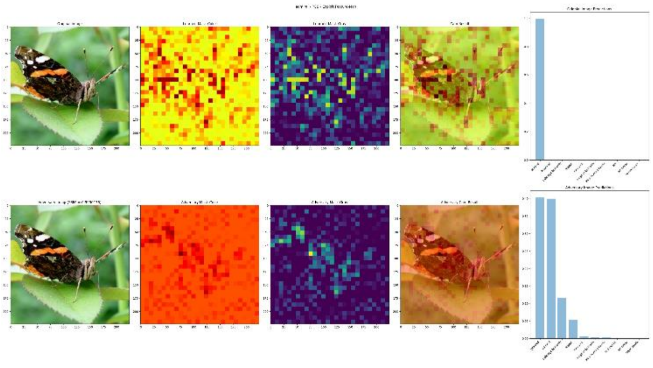

Figure 16. ResNet Proposed Pixel-Map of PGD Attack

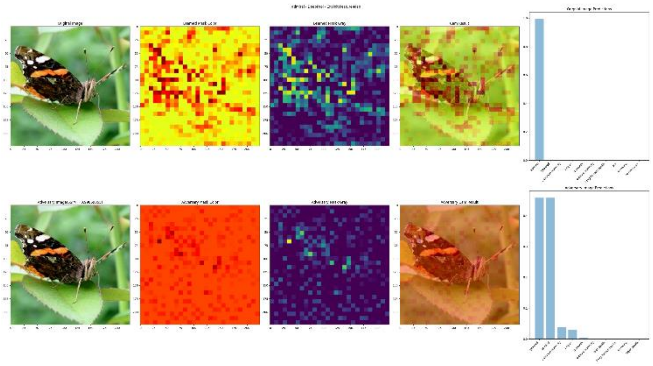

Figure 17. ResNet Proposed Pixel-Map of DeepFool Attack

TABLE I. ANALYSIS OF CLASSIFIERS

Model	Attack	Grad-Cam Accuracy	Pixel-Map Accuracy
AlexNet	FGSM	64%	90%
	PGD	52%	92%
	DeepFool	40%	88%
VggNet	FGSM	52%	90%
	PGD	66%	94%
ResNet	FGSM	64%	94%
	PGD	68%	92%
	DeepFool	52%	88%

IV.CONCLUSION

Based on the results obtained from the experiments conducted on different models (AlexNet, VggNet, and ResNet) subjected to various adversarial attacks (FGSM, PGD, DeepFool), several conclusions can be drawn. Firstly, it is evident that the choice of model significantly impacts the accuracy and robustness against adversarial attacks. ResNet consistently outperforms AlexNet and VggNet across different attack types, achieving higher accuracy rates. This emphasizes the importance of utilizing more advanced and deeper architectures like ResNet for improved defense against adversarial attacks.

Furthermore, the type of attack also plays a crucial role in model performance. PGD (Projected Gradient Descent) generally results in higher accuracy compared to FGSM (Fast Gradient Sign Method) and DeepFool attacks. This indicates that PGD is a more potent and challenging attack method that can bypass defenses more effectively. However, despite the differences in attack success rates, it's notable that the Pixel-Map Accuracy remains relatively high for most scenarios, showcasing the effectiveness of pixel-map analysis in detecting and mitigating adversarial attacks.

In conclusion, the study highlights the need for robust deep learning models such as ResNet and the importance of employing sophisticated defense mechanisms like pixel-map analysis to enhance model resilience against adversarial attacks. Future research may focus on further refining defense strategies and exploring novel architectures to tackle the evolving challenges posed by adversarial threats in machine learning systems.

V. REFERENCES

[1] G. Ryu and D. Choi, "Detection of adversarial attacks based on differences in image entropy," International Journal of Information Security, 2023, doi: 10.1007/s10207-023-00735-6.

[2] X. Cui, "Targeting Image-Classification Model," pp. 1–13, 2023.

[3] M. Kim and J. Yun, "AEGuard: Image Feature-Based Independent Adversarial Example Detection Model," Security and Communication Networks, vol. 2022, 2022, doi: 10.1155/2022/3440123.

[4] P. Lorenz, M. Keuper, and J. Keuper, "Unfolding Local Growth Rate Estimates for (Almost) Perfect Adversarial Detection," pp. 27–38, 2023, doi: 10.5220/0011586500003417.

[5] L. Shi, T. Liao, and J. He, "Defending Adversarial Attacks against DNN Image Classification Models by a Noise-Fusion Method," Electronics (Switzerland), vol. 11, no. 12, 2022, doi: 10.3390/electronics11121814.

[6] A. S. Almuflih, D. Vyas, V. V Kapdia, M. R. N. M. Qureshi, K. M. R. Qureshi, and E. A. Makkawi, "Novel exploit feature-map-based detection of adversarial attacks," Applied Sciences, vol. 12, no. 10, p. 5161, 2022.

[7] M. Khan et al., "Alpha Fusion Adversarial Attack Analysis Using Deep Learning," Computer Systems Science and Engineering, vol. 46, no. 1, pp. 461–473, 2023, doi: 10.32604/csse.2023.029642.

[8] N. Ghaffari Laleh et al., "Adversarial attacks and adversarial robustness in computational pathology," Nature Communications, vol. 13, no. 1, pp. 1–10, 2022, doi: 10.1038/s41467-022-33266-0.

[9] Y. Wang et al., "Adversarial Attacks and Defenses in Machine Learning-Powered Networks: A Contemporary Survey," pp. 1–46, 2023, [Online]. Available: http://arxiv.org/abs/2303.06302

[10] H. Hirano, A. Minagi, and K. Takemoto, "Universal adversarial attacks on deep neural networks for medical image classification," BMC Medical Imaging, vol. 21, no. 1, pp. 1–13, 2021, doi: 10.1186/s12880-020-00530-y.

[11] A. Talk, F. Wikipedia, A. Wikipedia, and C. Wikipedia, "University of Science and Technology of China," no. 6, p. 29201, 2001.

[12] Y. Zheng and S. Velipasalar, "Part-Based Feature Squeezing To Detect Adversarial Examples in Person Re-Identification Networks," Proceedings - International Conference on Image Processing, ICIP, vol. 2021-September, pp. 844–848, 2021, doi: 10.1109/ICIP42928.2021.9506511.

[13] B. Liang, H. Li, M. Su, X. Li, W. Shi, and X. Wang, "Detecting Adversarial Image Examples in Deep Neural Networks with Adaptive Noise Reduction," IEEE Transactions on Dependable and Secure Computing, vol. 18, no. 1, pp. 72–85, 2021, doi: 10.1109/TDSC.2018.2874243.

[14] M. A. Ahmadi, R. Dianat, and H. Amirkhani, "An adversarial attack detection method in deep neural networks based on re-attacking approach," pp. 10985–11014, 2021.

[15] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial Attacks and Defenses in Deep Learning," Engineering, vol. 6, no. 3, pp. 346–360, 2020, doi: 10.1016/j.eng.2019.12.012.