Grenze International Journal of Engineering and Technology, June Issue

Defense against Artificial Intelligence Hacking

E.Rajkumar[1], Mano Subash S[2], Sudharsan J[3] and Balaji M[4]
[1]Assistant Professor, Department of Cyber Security, SRM Valliammai Engineering College
Email: raju.becs39@gmail.com
[2-4]UG Scholar, Department of Cyber Security, SRM Valliammai Engineering College
Email: smanomahi2002@gmail.com, sudharsanjayakumar7@gmail.com, gurubalaji2002@gmail.com

***Abstract*— AI and deep learning is influential and pervasive in our daily lives with the potential to revolutionize various facts of society. We cannot manipulate or trick this advanced technology easily. Indeed, some crafty individuals utilize adversarial examples as a means to fool AI systems. These instances present an image specifically engineered to confuse our models while sidestepping noticeable alteration from a human perspective. Essentially very minor changes can be made which may go unnoticed by humans but lead the DL model into falsely identifying objects at high confidence level levels imagine if self-driving vehicles began interpreting the stop signs. In this project we used AlexNet as a convolutional neural network. We can load a pretrained version of the network trained on more than a million images from the our database. It showcases exceptional skill at identifying 1000 distinct categories including common place objects such as mice, pencils and keyboards alongside a variety of animals thus exhibiting remarkable prowess in recognizing images. Our objective involves in developing resilient models capable for enough counterattacks efficiently from those manipulating images adversely geared toward attacking it.**

***Index Terms*— Adversarial Attacks, Autonomous Driving, Traffic Accident, Image classification, Noise Image.**

I. INTRODUCTION

These avantgarde technologies have smoothly merged into our daily rhythms, establishing consequential impact throughout diverse fields. However, AI's high-level sophistication is coupled with susceptibility an exposure to deception orchestrated by crafty entities via adversarial examples artfully designed images purposed as tools for misleading AI systems introducing barely noticeable changes undetectable for human viewing yet powerful enough in leading sophisticated models off track. Trained using over million pictures sourced from ImageNet database, AlexNet displays extraordinary expertise identifying 1000 divergent categories ranging between traditional objects such mice pencils, keyboards or wider range animals distinct varieties. In context our prime aim building robust model efficient oppose intents crafted manipulating picture counteractively.

II. RELATED WORKS

[1] The involving different models, adversarial attack algorithms, perturbation budgets, and image scaling processes, to comprehensively assess the impact of image scaling on adversarial example robustness. The study's outcomes contribute significantly to advancing the understanding of adversarial vulnerabilities and offer practical implications for enhancing the robustness of image classification systems in the face of diverse scaling scenarios. [2] The paper investigates universal adversarial example attacks on image classification models to design effective

Grenze ID: 01.GIJET.10.2.213

mitigation strategies. The interpretation mainly analyzes input or specific layers, suggesting potential future extensions involving a global analysis of different network layers. Future work will explore interpretations in diverse application domains, such as medical image classification. [3] The paper explores an enhanced backdoor attack on image and time-series data-based learning systems facing strong defenses. Three optimization strategies, including Ranking-based Neuron Selection, Autoencoder-powered Trigger Generation, and Defense-aware Retraining, are employed. Evaluations on real-world applications demonstrate the attack's effectiveness against various defenses while maintaining high classification accuracy on clean inputs. Future work includes enhancing detection evasiveness and exploring other attacks and threats beyond backdoors. [4] The study delves into security issues arising from using third-party primitive models in ML systems, demonstrating model-reuse attacks across skin cancer screening, speech recognition, face verification, and autonomous driving applications. Analyzing four ML systems, the work reveals fundamental vulnerabilities due to high dimensionality, non-linearity, and non-convexity in today's ML models. [5] This paper introduces an efficient brute-force attack on transfer learning for deep neural networks, exploiting a vulnerability in the Softmax layer. The paper suggests that defeating this target-agnostic attack requires considering the distribution of the activation vector, as demonstrated by the EVM method, rather than relying solely on linear combinations like the Softmax layer.

III. Material Implementation

A. Generate Adversarial Image

Adversarial attacks are a clever technique used in the realm of deep learning. They involve introducing imperceptible alterations to an original image. These subtle perturbations are designed with the intent of confusing deep learning models, causing them to produce inaccurate predictions. The human eye to discern, making them a potent tool for evading machine learning systems. Adversarial attacks have raised concerns about the robustness and security of AI, particularly in applications like image recognition and autonomous vehicles. Adversarial image that is generated by the Fast Gradient Sign Method (FGSM).

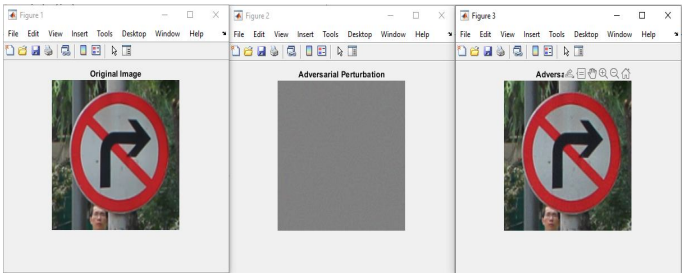

Fig 1 Generating Adversarial Image

B. Training Using Pretrained Model

In this module, the focus is on exploiting transfer learning with the pretrainedAlexNet model. AlexNet is a popular deep learning architecture known for its effectiveness in image-related tasks. Here, the pretrainedAlexNet model is utilized as a foundation. The lower layers of the model, having learned general features from a large dataset like ImageNet, are retained, while the upper layers, responsible for more specific features, are further trained or modified based on the new dataset. This process accelerates training significantly, as the model has already grasped fundamental features from its initial training on ImageNet.

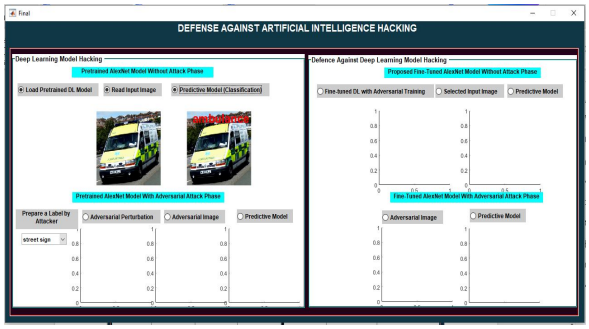

Fig 2- Normal training for Deep Learning model which was actually done everywhere.

C. Predictive Model

In this phase, a predictive model is constructed, leveraging AlexNet as its backbone. This pre-trained model is intended to classify both original and adversarial images. The model learns to recognize patterns in both data types, understanding the subtle perturbations introduced by adversarial attacks. By determining the respective classes to which both original and adversarial images belong.

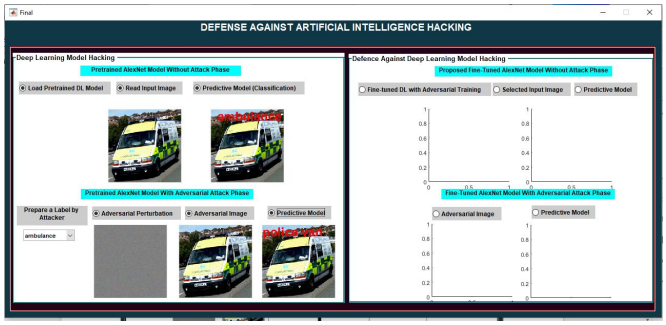

Fig 3- Predictive Model

D. Develop Defense Model

Adversarial attacks are effective against models with unchanged pretrained base models. However, transfer learning via fine-tuning renders these attacks ineffective. Fine-tuning the structure of AlexNetpretrained architecture to avoid white box attacks. The Fine-Tuned AlexNet model was generated by modifying the fully connected layers to adapt to our specific dataset. With hyperparameters set to 50 epochs and SGDM optimizer, image augmentation techniques like horizontal reflection, translation, and scaling were applied to enhance model robustness.

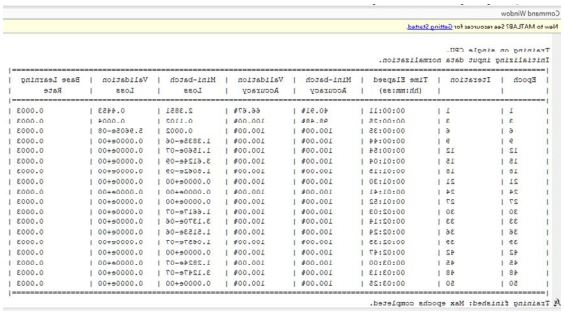

Fig4-The stored Dataset for Adversarial image been used in attack phase

E. Training using Defense Model

The defense approach, termed 'fine-tuned AlexNet model with adversarial images training', involves mitigating vulnerabilities by training the model on adversarial examples. This process incorporates aversive images with varying epsilon values to counter black box attacks. Additionally, fine-tuning the structure of the AlexNetpretrained architecture helps deter white box attacks. During training, the model is exposed to both original and adversarial images. The training dataset includes examples with subtle adversarial perturbations, enabling the model to learn to differentiate between genuine and manipulated inputs.

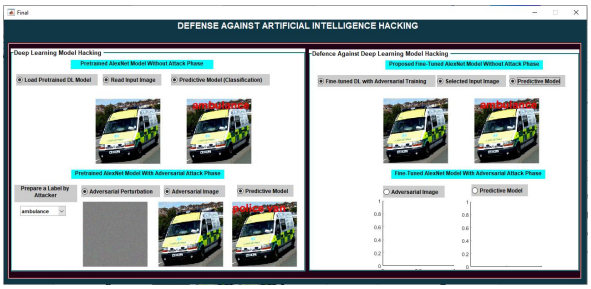

Fig 5- Importing the image into predictive model which was efficiently fine-tuned

F. Defense Predictive Model

The defense predictive model is designed to classify both original and adversarial images accurately. Leveraging the insights gained from training on adversarial examples, the model employs sophisticated algorithms to identify patterns indicative of adversarial manipulations. Through meticulous analysis of image features, the model determines the respective classes to which original and adversarial images belong, ensuring robust and reliable classification results.

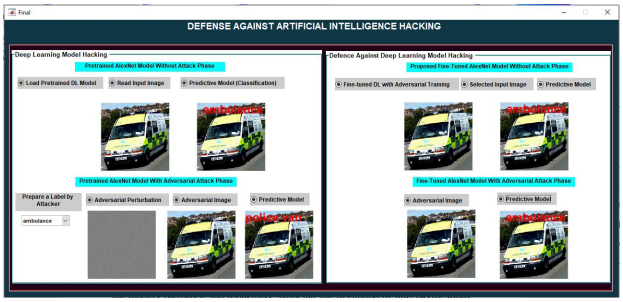

Fig 6- The output for defense model

IV. Conclusion & Future Work

In the tricky world of making and protecting advanced computer programs, we've learned a lot about how attackers and defenders work. Hackers can break into systems using trained models or by tricking the system with fake images. We create strategies focused on defense, make changes to models to use them better, and practice with tricky examples to build strong defenses against attacks. In the end, designing a model that can tell apart real and fake images helps keep things stable and safe, even as threats grow quickly. Our wide-reaching plan highlights the importance of knowing how to break into systems and how to protect them so we can keep future AI safe. Our defense plans need to work for many different areas and uses. As computers get even more powerful, it will be important to think about how to protect against extremely advanced attacks. By teaching others and sharing what we know, we can help everyone fight off new dangers. Working together across countries and sharing information will be key to dealing with AI security challenges worldwide.

References

[1] J. Zheng, Y. Zhang, Y. Li, S. Wu and X. Yu, "Towards Evaluating the Robustness of Adversarial Attacks Against Image Scaling Transformation." Chinese Journal of Electronics, vol. 32, no. 1, pp. 151-158, 2023, doi: 10.23919.

[2] Ding, Yi et al. “Interpreting Universal Adversarial Example Attacks on Image Classification Models.” IEEE Transactions on Dependable and Secure Computing, 2023, DOI: 3392-3407.

[3] S. Wang, S. Nepal, C. Rudolph, M. Grobler, S. Chen and T. Chen, "Backdoor Attacks Against Transfer Learning with Pre-Trained Deep Learning Models." IEEE Transactions on Services Computing, vol. 15, no. 3, pp. 1526-1539, 2022, Doi: 10.1109/TSC.2022.3000900.

[4] Y. Ji, X. Zhang, S. Ji, X. Luo, and T. Wang, “Model-reuse attacks on deep learning systems,” Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, 2022, pp. 349– 363.

[5] S. Rezaei and X. Liu, “A target-agnostic attack on deep models: Exploiting security vulnerabilities of transfer learning,” Doi: 1904.04334, 2022.

[6] B. Wang, Y. Yao, B. Viswanath, H. Zheng, and B. Y. Zhao, “With great training comes great vulnerability: Practical attacks against transfer learning.” Security Symposium ({USENIX} Security 18), 2021, pp. 1281–1297.

[7] V. U. Prabhu and J. Whaley, “On grey-box adversarial attacks and transfer learning,” online: https://unify.id/wpcontent/uploads/2021/03/greybox attack.pdf, 2021.

[8] Pal, Bijeeta and ShrutiTople. “To Transfer or Not to Transfer: Misclassification Attacks Against Transfer Learned Text Classifiers.” abs/2001.02438, 2020.

[9] Abdelkader, Ahmed et al. “Headless Horseman: Adversarial Attacks on Transfer Learning Models.” ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020): 3087-3091.

[10] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik and A. Swami, "The Limitations of Deep Learning in Adversarial Settings."IEEE European Symposium on Security and Privacy 2020, pp. 372-387, doi: 10.1109,2020.