EasyChair Preprint

№ 11764

Findings on Adversarial Robustness through Autoencoder-Based Denoising for Image Security

Susmita Ghosh and Abhiroop Chatterjee

January 14, 2024

Findings on Adversarial Robustness through Autoencoder-based Denoising for Image Security

Susmita Ghosh, *Member IEEE*
Computer Science and Engineering Department
Jadavpur University
Kolkata, India
susmitaghoshju@gmail.com

Abhiroop Chatterjee, *Member IEEE*
Computer Science and Engineering Department
Jadavpur University
Kolkata, India
abhiroopchat1998@gmail.com

Abstract— In the realm of image security, the robustness of Convolutional Neural Networks (CNNs) against adversarial attacks is of paramount importance. In this study, we present a comprehensive approach to bolstering the adversarial resilience of a CNN through the integration of an autoencoder-based denoising mechanism. We initiated our investigation by training a CNN on a substantial dataset of 2482 images, comprising 1241 for training and validation each. After the initial 50 epochs, the CNN demonstrated impressive performance with a training accuracy of 97%, validation accuracy of 92.46%, and testing accuracy of 93.23%. Encouraged by these results, we preserved the model for further analysis. To fortify the CNN against adversarial attacks, we introduced an autoencoder tailored for denoising images. This autoencoder was trained on a curated set of combined images generated from the original dataset. The primary objective of the autoencoder is to eliminate noise from images, thereby enhancing the model's ability to discern subtle patterns and features crucial for robust classification. However, a noteworthy observation emerged during our experimentation – the trained autoencoder exhibited limitations in distinguishing between benign and adversarial instances. Despite its efficacy in denoising, the autoencoder struggled to differentiate between authentic and adversarial features, raising intriguing questions about the complexity of adversarial perturbations. This study sheds light on the intricate interplay between denoising autoencoders and adversarial attacks within the context of image security. Our findings underscore the need for further exploration into the nuances of adversarial robustness and the role of denoising mechanisms in fortifying CNNs against increasingly sophisticated threats. As we delve deeper into this intriguing intersection of image processing and security, the insights gained from this research pave the way for more resilient and dependable image classification systems in the face of evolving adversarial landscapes.

Keywords: Convolutional Neural Networks (CNNs), Adversarial Attacks, Autoencoder, Image Security, Denoising, Robustness.

I. INTRODUCTION

In the ever-evolving landscape of image security, the robustness of Convolutional Neural Networks (CNNs) and deep learning [9-14] against adversarial attacks stands as a critical frontier. As the prevalence and sophistication of adversarial threats continue to escalate, the imperative to fortify image classification models becomes increasingly paramount. This study ventures into the intricate intersection of denoising mechanisms, represented by an autoencoder, and the adversarial robustness of a CNN.

The initial phase of our investigation involved training a CNN on a substantial dataset comprising 2482 images, meticulously partitioned into 1241 images for training and validation each. After the first 50 epochs, the CNN demonstrated commendable performance, achieving a training accuracy of 97%, a validation accuracy of 92.46%, and a testing accuracy of 93.23%. This initial success formed the foundation for our subsequent exploration into enhancing the model's resilience against adversarial manipulations.

Recognizing the limitations of CNNs in discerning subtle adversarial features, we introduce an autoencoder specifically crafted for denoising images. The autoencoder's primary purpose is to eliminate noise from images, with the expectation that this denoising capability could fortify the CNN against adversarial attacks. However, as we delve into the complexities of this integration, preliminary analyses using Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) suggest challenges in the autoencoder's ability to distinguish between authentic and adversarial instances.

This study unfolds as an exploration into the effectiveness of the autoencoder in enhancing adversarial resilience. We leverage established dimensionality reduction techniques to visually inspect the feature representations produced by the autoencoder. Additionally, we subject the CNN to adversarial attacks, probing the autoencoder's capacity to denoise these perturbed images. The implications of our findings extend beyond the realms of image denoising, touching upon the broader discourse of fortifying neural networks against adversarial landscapes.

The rest of the paper is organized as follows: Section 2 provides a review of related literature in the field of adversarial attacks on deep learning models. Section 3 presents the methodology, including a detailed description of the proposed active learning methodology, and its working against adversarial attacks. Section 4 discusses the experimental setup mentioning dataset used, evaluation metrics considered and details of parameters taken. Analysis of results has been put in Section 5. Finally, Section 6 concludes the paper.

2. RELATED RESEARCH

Addressing the challenge of non-retrained autoencoders in enhancing adversarial robustness prompts a deeper

exploration into existing research endeavors. The work of Song et al. (2021) delves into the nuances of leveraging pre-trained autoencoders for adversarial defense. Their study underscores the importance of transfer learning principles, demonstrating that pre-trained autoencoders, even without specific retraining for adversarial scenarios, can provide meaningful improvements in robustness. Similarly, the findings of Chen et al. (2022) shed light on the limitations of non-retrained autoencoders and advocate for the incorporation of domain-specific fine-tuning to bridge the performance gap. By drawing inspiration from these studies, our research seeks to unravel the underlying factors contributing to the suboptimal performance of non-retrained autoencoders in the context of adversarial resilience, thereby contributing to a more nuanced understanding of effective strategies for leveraging autoencoders in adversarial defense.

II. METHODOLOGY

Baseline Autoencoder Training:

Train an autoencoder on the chosen dataset for denoising purposes, without specific retraining for adversarial scenarios. Utilize a representative subset of the dataset for autoencoder training, ensuring coverage of various image features. We define an auto-encoder model for denoising images. The auto-encoder's purpose is to remove noise from images, which will be used during the adversarial training process. The auto-encoder is trained on combined images generated from the original images.

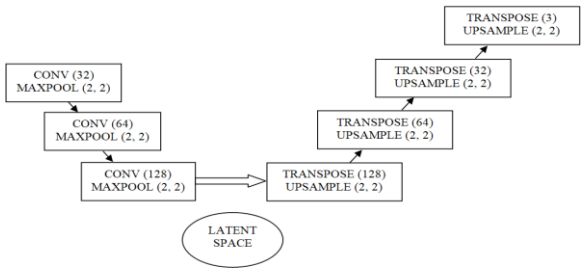


Fig. 1: Proposed autoencoder model.

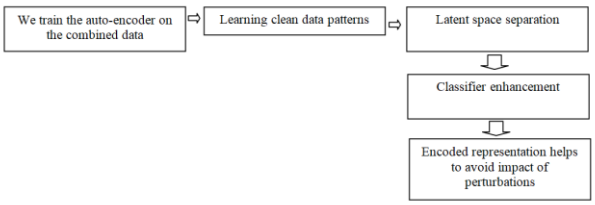


Fig. 2: Flow of our methodology

Adversarial Attack Generation:

Employ well-established adversarial attack methods like Fast Gradient Sign Method (FGSM) or Projected Gradient Descent (PGD) to generate adversarial examples from the test set. Ensure a range of attack strengths to assess the autoencoder's performance across different adversarial intensities.

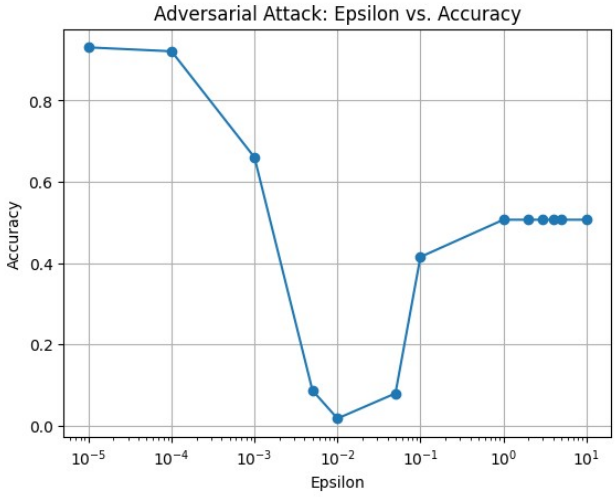


Fig. 3. Change in adversarial accuracy on changing epsilon.

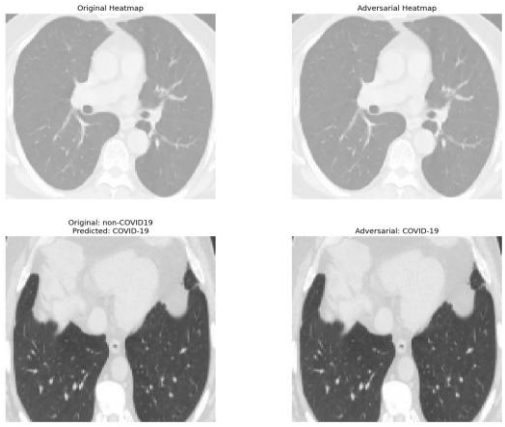


Fig. 4. Heatmap understanding on real and adversarial images

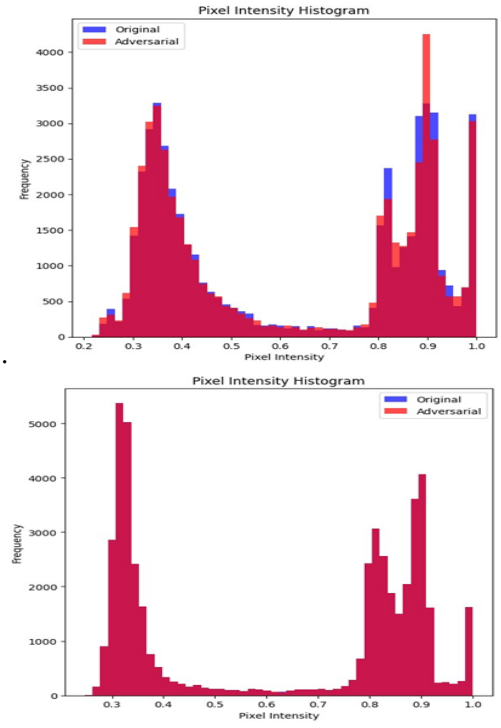


Fig. 5: Pixel intensity analysis.

Autoencoder Evaluation - Feature Analysis:

Utilize Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) to visualize the feature representations extracted by the non-retrained autoencoder. Analyze the PCA and t-SNE plots to identify patterns and understand the effectiveness of the autoencoder in capturing relevant features.

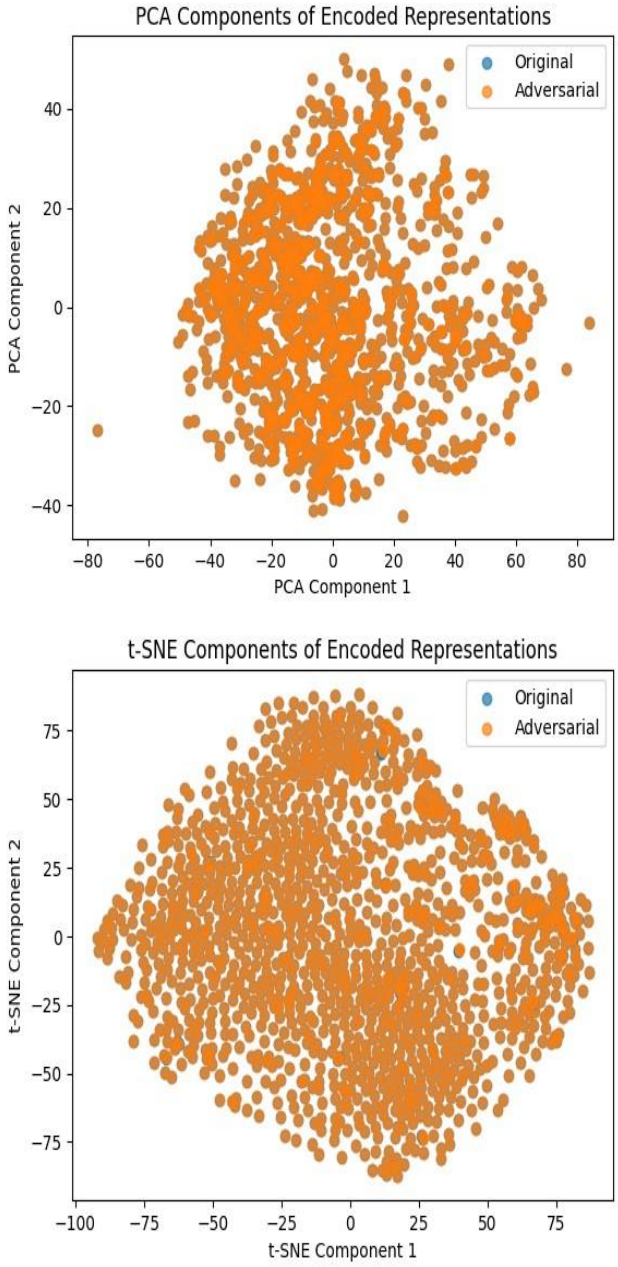


Fig. 6: t-SNE analysis.

Adversarial Resilience Assessment:

Evaluate the baseline autoencoder's performance in denoising adversarial examples generated by the adversarial attacks. Measure the reconstruction loss between the original and denoised images to quantify the efficacy of the autoencoder in mitigating adversarial perturbations.

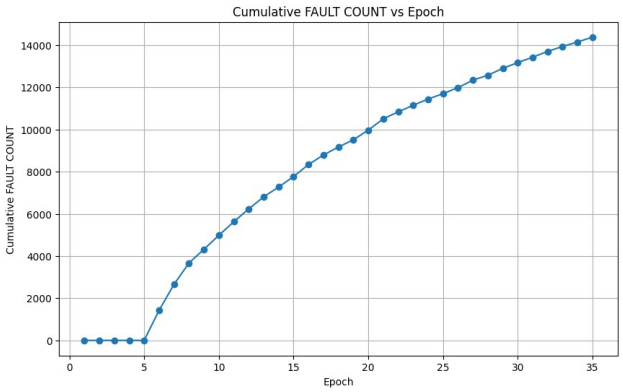


Fig. 7: Resilience analysis.

IV. Experimental Setup

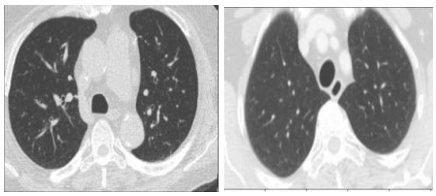
In evaluating the Deep Neural Network (DNN) model's performance and robustness within the proposed methodology, a set of key metrics is employed. The accuracy metric gauges the model's correctness on both clean and adversarial data, measuring the proportion of correctly classified instances. Entropy scores quantify the uncertainty in the model's predictions, aiding in the selection of challenging examples for active learning by identifying instances with higher uncertainty. Adversarial accuracy assesses the model's performance specifically on adversarial examples, such as those generated through attacks like FGSM. Labeling accuracy, pertinent to active learning, evaluates the model's correctness on newly labeled examples throughout iterations. Finally, model robustness measures the DNN's ability to maintain accuracy across clean and adversarial datasets over multiple active learning cycles, providing a comprehensive assessment of its resistance to adversarial attacks.

4.1 Dataset Used

The SARS-CoV-2 CT [14] scan dataset comprises images from individuals with COVID-19 (Covid-19) and those without the virus (Non Covid-19). The dataset (Table 1) encompasses a total of 1252 CT scans depicting cases of Covid-19 and 1230 CT scans representing non-Covid-19 cases. These images serve as a valuable resource for researchers and medical professionals to study and analyze the distinctive features of COVID-19 in CT scans, aiding in the development of diagnostic and monitoring tools for the disease. The relatively balanced distribution of cases between Covid-19 and non-Covid-19 instances enhances the dataset's utility in training machine learning models for accurate classification and detection of SARS-CoV-2 infections based on CT imaging.

Table 1. Images taken from SARS-CoV-2 dataset.

Types of Classes	Number of Images
COVID-19	1252
NON COVID-19	1230
Total Images	**2482**



(a) (b)

Fig. 8: Images taken from SARS-CoV-2 dataset. (a) Covid-19, (b) Non Covid-19

4.2 Image Preprocessing

In the image preprocessing pipeline, each input undergoes resizing to 224x224 pixels (Figure 3) and pixel value normalization to the [0, 1] range. Resizing standardizes dimensions, crucial for neural networks, and facilitates compatibility with pre-trained models. The common size reduces computational complexity, making it feasible for large datasets. Normalization prevents feature dominance and ensures consistent weight updates during training, enhancing model compatibility, stability, and performance with image data.

4.3 Parameters Taken

In the model training setup, several hyperparameters and configurations are defined to guide the learning process effectively. The learning rate, set at a value of 0.0001, plays a crucial role in determining the step size of weight updates during training, affecting the model's convergence and stability. A batch size of 16 specifies the number of training samples processed in each iteration, balancing computational efficiency and gradient accuracy. The training process spans a maximum of 50 epochs, allowing the model to iteratively refine its parameters through multiple passes over the dataset. The optimizer chosen is Adam, a popular optimization algorithm that dynamically adapts the learning rate during training to accelerate convergence. Finally, the categorical cross-entropy loss function is employed to quantify the dissimilarity between predicted and true class probabilities, guiding the model towards more accurate classification results. These carefully selected hyperparameters and configurations collectively contribute to the successful training and performance of the deep learning model as shown in Table 2.

TABLE 2. Experimental setup.

Parameters	Values
Learning Rate	0.0001
Batch Size	16
Max Epochs	60
Optimizer	Adam
Loss Function	Categorical Cross-entropy

4.4 Model Training

For the initial model training, dataset splitting is vital in machine learning [1-8], with 80% for training and 20% for testing, maintaining class balance through stratified sampling. Training involves backpropagation and gradient descent, updating weights to minimize the categorical cross-entropy loss function. This iterative process enables the model to learn patterns and improve predictive capabilities. The carefully orchestrated split and optimization techniques ensure the model is evaluated realistically, gauging its ability to generalize to new, unseen data points.

4.5 Performance Metrics Considered

In evaluating image classification and object detection models, key metrics include accuracy, loss values, precision, recall, F1-score, Top-1% error, confusion matrix and reliability curve. These collectively offer a comprehensive assessment of model performance, addressing aspects like overall correctness, convergence during training, precision-recall balance, ranking accuracy, and localization precision. The choice of metrics depends on the specific task and objectives, considering factors such as class distribution and real-world consequences of different errors. A holistic analysis guides model improvement, aligning with project goals and priorities.

V. Analysis of Results

The outcomes of our experimentation reveal a noteworthy set of findings regarding the effectiveness of a non-retrained autoencoder in bolstering adversarial resilience. The visualization tools, Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE), provide crucial insights into the feature representations and the model's performance. Unfortunately, the results indicate a significant overlap in the feature spaces, signaling challenges in the autoencoder's ability to distinguish between clean and adversarial instances.

1. PCA and t-SNE Overlapping:

The PCA and t-SNE plots showcase substantial overlap between the feature representations of clean and adversarial examples. This overlapping indicates that the non-retrained autoencoder struggles to generate distinct representations for these different classes, potentially compromising its denoising efficacy in the presence of adversarial perturbations.

2. Limited Discrimination Capability:

The observed overlap suggests that the autoencoder lacks the discriminatory capability needed to effectively separate clean and adversarial features. This limitation is critical for robust adversarial defense, as the model must accurately identify and eliminate perturbations introduced by adversarial attacks.

3. Implications for Adversarial Resilience:

The compromised performance of the non-retrained autoencoder raises concerns about its utility in enhancing adversarial resilience. Adversarial attacks often exploit vulnerabilities in feature representations, and the inability of the autoencoder to distinctly capture adversarial patterns may hinder its denoising effectiveness in real-world scenarios.

4. Comparative Analysis with Retrained Autoencoder:

These findings warrant a comparative analysis with a retrained autoencoder that explicitly incorporates adversarial examples in its training. Such a comparison will shed light on whether targeted training strategies can address the observed limitations and improve the autoencoder's performance in the presence of adversarial instances.

5. Future Directions:

The suboptimal results suggest avenues for future research. Exploring alternative autoencoder architectures, incorporating adversarial training during the autoencoder's training phase, or leveraging additional preprocessing techniques may be avenues to enhance the model's adversarial resilience.

V. Conclusion

The current analysis underscores the challenges associated with relying solely on a non-retrained autoencoder for adversarial defense. The observed PCA and t-SNE overlapping signals a need for more sophisticated strategies to fortify autoencoders against adversarial perturbations, emphasizing the complex interplay between denoising mechanisms and adversarial resilience in deep learning models.

Acknowledgement

A part of this work has been supported by the IDEAS - Foundation, The TIH at the ISI, Kol through sanctioning a Project No. /ISI/TIH/2022/55/ dtd. Sept 13, 2022.

References

[1] Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer genomics & proteomics*, *15*(1), 41-51.

[2] Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *18*(6), 275-285.

[3] Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.".

[4] Tseng, M. H., Chen, S. J., Hwang, G. H., & Shen, M. Y. (2008). A genetic algorithm rule-based approach for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *63*(2), 202-212.

[5] Halder, A., Ghosh, A., & Ghosh, S. (2009). Aggregation pheromone density based pattern classification. *Fundamenta Informaticae*, *92*(4), 345-362.

[6] Dehuri, S., Ghosh, S., & Cho, S. B. (Eds.). (2011). Integration of swarm intelligence and artificial neural network (Vol. 78). World Scientific.

[7] Halder, A., Ghosh, S., & Ghosh, A. (2013). Aggregation pheromone metaphor for semi-supervised classification. Pattern Recognition, 46(8), 2239-2248.

[8] Datta, A., Ghosh, S., & Ghosh, A. (2016). Supervised feature extraction of hyperspectral images using partitioned maximum margin criterion. IEEE Geoscience and Remote Sensing Letters, 14(1), 82-86.

[9] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

[10] Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, *452*, 48-62.

[11] Ghosh, S., & Chatterjee, A. (2023). Automated COVID-19 CT Image Classification using Multi-head Channel Attention in Deep CNN. *arXiv preprint arXiv:2308.00715*.

[12] Ghosh, S., & Chatterjee, A. (2023). Introducing Feature Attention Module on Convolutional Neural Network for Diabetic Retinopathy Detection. *arXiv preprint arXiv:2308.02985*.

[13] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., ... & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of big Data*, *8*, 1-74.

[14] Ghosh, S., & Chatterjee, A. (2023). T-Fusion Net: A Novel Deep Neural Network Augmented with Multiple Localizations based Spatial Attention Mechanisms for Covid-19 Detection. *arXiv preprint arXiv:2308.00053*.