Analysis of Adversarial Training for Resilient Image Recognition Against Different Attacks

Emir Can Karakuş
Computer Engineering
Bahcesehir University
Istanbul, Turkey
emircan.karakus@bahcesehir.edu.tr

Fırat Kutluhan İslim
Computer Engineering
Bahcesehir University
Istanbul, Turkey
firatkutluhan.islim@bahcesehir.edu.tr

Ece Gelal Soyak
Computer Engineering
Bahcesehir University
Istanbul, Turkey
ece.gelalsoyak@bau.edu.tr

***Abstract*—Image recognition systems are used for various critical use cases, from cancer detection to autonomous vehicles, while their accuracy closely relies on the data that they are trained on. Recently, adversarial machine learning has been flagged as a possible threat against the successful operation of such systems and models. While the research towards understanding how to manipulate such models continues, one mitigation approach is to increase the model's resilience against such samples via including such samples in the training process. In this work, we examine the impact of adversarial training on improving robustness against other adversarial attack types. We generate adversarial samples via two image perturbation methods, Fast Gradient Sign Method (FGSM) and Jacobian-based Saliency Map Attack (JSMA), and include these samples respectively in the training set of two independent ResNet-18 models on the CIFAR-10 dataset. Obtaining two adversarially-trained models, we compare their accuracies upon each attack and discuss the impact on model resiliency when the training set includes samples generated by different approaches. Our results highlight that FGSM-fine tuned (adversarially trained) model earns the model greater resilience against FGSM attack compared to JSMA attack. Adversarial JSMA-training, when JSMA samples are generated to target the original class, earns the model resilience against both attacks.**

***Index Terms*—adversarial machine learning, defense, fast gradient sign method, jacobian-based saliency map attack**

I. Introduction

From the Internet of Things (IoT) enabling the generation and collection of vast data, to cloud/edge computing systems for data processing, and to advancements in learning models, many research and engineering efforts are geared towards facilitating artificial intelligence (AI)-supported systems in various aspects of daily life. Among these AI applications, image recognition models are at the heart of autonomous vehicles, human detection systems for security authentication, quality detection in manufacturing, analyzing people's reactions by emotion recognition, fingerprint recognition and even detecting and diagnosing medical conditions in healthcare. The industry standards of image recognition systems are relatively high as there are many products that rely on them, including Tesla cars, Google lens, or Amazon's cashierless stores.

It must be noted that these AI-based solutions significantly enhance the quality of our lives *only if the system is secure and working as intended*. This useful system may also be used to breach security, steal money or harm people if not well-protected. Cyber security solutions focus on ensuring identity certification, encrypted data transfer and authorized access to data; however, existing cyber security solutions do not consider attacks against the AI model.

Adversarial Machine Learning (AML) is an emerging field that studies whether and how input data can be *intentionally* manipulated, so that it is *incorrectly* recognized by the target model. If the right precautions are not taken, it can lead to scenarios such as an autonomous vehicle not recognising an obstacle on the road, misdiagnosis of a patient, an unauthorized person entering a building, or defected products being packaged and put on sale. To circumvent these unwanted scenarios, one must recognize the potential dangers of AML attacks, and take precautions.

One of the approaches in improving the resilience of an AI system to adversarial attacks is to train the model on a comprehensive data set, comprising some adversarial samples as well, namely *adversarial training* [1]. By doing so, the model becomes more resilient to similar attacks in the future. There is a strong coupling between the resilience against adversarial attacks and the attack strategy used during the training process. Adversarial training with an attack strategy does not formally guarantee a generic model that can resist against adversarial images generated by a different attack algorithm [2]. However, heuristically, adversarial training has the potential to also improve a model's ability to generalize to unseen adversarial examples [3].

In this work, adversarial training has been studied as a defence mechanism against adversarial attacks in image recognition systems. Particularly, the impact of training with different adversarial samples on the resilience to different attack types is comparatively analyzed.

Our contributions can be summarized as follows:

- We quantify the decrease in accuracy on a trained ResNet-18 model using FGSM and JSMA attacks separately.
- We quantify the decrease *in the decrease of accuracy* when the ResNet-18 model is adversarially fine-tuned, by training using perturbed images generated by both attack types, separately. We separately quantified the model's resilience to each attack type.
- We shed light on the difference of approaches with FGSM and JSMA, that leads to the observed results.

II. Related Work

Though the Adversarial Machine Learning (AML) terminology has been introduced almost a decade ago [4], it has much more recenty been recognized as a crucial research area, especially in the context of image and object recognition models [5].

Several studies investigated the impact of attacks on different classifiers, particularly image classifiers. Different adversarial attack strategies were implemented and compared in [6], where it was also shown that the examples produced by FGSM are more easily detectable and require a bigger distortion to achieve misclassification than those obtained from JSMA.

Defense against AML attacks can be reactive (*i.e.,* countering past attacks) or proactive (*i.e.,* preventing future attacks) [7]. Adversarial training (or adversarial retraining [8]) is a proactive approach with the strategy of *gradient masking*, which aims to increase the robustness of a classification model by means of training the model on a dataset containing both legitimate and adversarial samples [2]. Several research efforts investigated the impact of adversarial training [9]. [10] used a variation of adversarial training to train a classifier with adversarial images crafted by an ensemble of DNNs. Another work applied regularization using the Jacobian of the network after regular training [11], and showed that this approach offered superior robustness compared to adversarial training.

In terms of the impact of training with one type of adversarial samples on the robustness against attack of another type, two different opinions have been argued. On the one hand, when projected gradient descent (PGD) and FGSM training were performed and the respective attacks were tested on the models, it was observed that adversarial training with PGD samples also somewhat improved the model's robustness against FGSM attacks in [12]. However, [2] stated that order to have a more generic model, it would be necessary to elaborate a training dataset with a massive amount of adversarial images generated using different attack algorithms and amounts of disturbance.

III. Background

A. Adversarial Perturbations on Images

Adversarial perturbations on images can be crafted at either training or testing phase. Adversarial examples exploit the vulnerability of classification functions when exposed to slight perturbations [13]. Given a clean sample x that is correctly classified by the model with label l, an adversarial example x' can be created by applying a minimal perturbation P to x to induce a different label l'.

$$\begin{aligned} min_{x'} \quad & ||x' - x||_p \ , \\ s.t. \quad & f(x') = l' \ , \\ & f(x) = l \ , \\ & l \neq l' \end{aligned} \tag{1}$$

In 1, the distance between samples are denoted as the *p-norm* and $x' - x$ is the perturbation. This distance formula is important to quantify the impact of perturbations. The assumptions of a minimal perturbation are not always straightforward, as machine learning systems lack the ability to distinguish between large and small perturbations like a human can. Because of that, keeping the perturbations minimum is the common method used to generate AML examples. Also this is the reason for the challenging part of the crafting adversarial perturbations that can evade model detection while maintaining a level of imperceptibility to the human eye [13].

In practice, an attacker may target a specific outcome, (*e.g.,* evade detection and penetrate into a system), or may aim to discredit the target model by misleading it to predict any output other than the true label [14], [15]. FGSM is an untargeted strategy; where the objective is to cause the model to misclassify the input into *any class* that is different from the original label. JSMA can be adjusted to perform targeted attacks by modifying the loss function to focus on a *target class*. In this work, we utilize both strategies to obfuscate the image and cause misclassification, without targeting classification into a specific class.

B. Data Set

In this work, the CIFAR-10 dataset [16] has been used for training the models. This dataset has been chosen due to its simplicity, popularity, and ease of use in terms of dimension [17].

The CIFAR-10 dataset [16] is a collection of 32x32 color images of 10 classes of objects and animals, specifically, plane, car, bird, cat, deer, dog, frog, horse, ship, and truck. There are $6,000$ samples from each class, adding up to $60,000$ total images.

IV. Implementation

To observe the impact of different training approaches on the resilience to AML attacks, we first analyzed the performance of a vanilla Resnet-18 model prior to any AML attack, then performed two separate AML attacks on this model to observe the impact. Next, we used the perturbed images generated using each attack separately, on fine-tuning the base model (adversarial training) and thus obtained two separate adversarially trained models. To comparatively analyze the behavior, we then performed both attacks on each adversarially trained model. Figure 1 demonstrates the methodology followed in this work.

A. Target Model

The target of the AML attacks is a ResNet-18 image classifier. ResNet-18 is a convolutional neural network that is 18 layers deep and has been pretrained on more than a million images from the ImageNet dataset. The model is capable of classifying images into $1,000$ object categories. Upon training, the network has learned rich feature representations for a wide range of images. The network has an image input size of 224x224.

In our ResNet-18 implementation, the loss function used was CrossEntropyLoss, which is commonly used for multi-class classification. For the optimizer, we used Adam Optimizer with an initial learning rate of 0.001, but during the experiments the *learning rate* hyperparameter was optimized. We also performed data augmentation methods on the training photos, such as random horizontal flipping and random cropping, to improve the model's capacity for generalization. In addition, all color channels in the training and testing photos were standardized to have a mean of 0.5 and a standard deviation of 0.5.

B. Adversarial Samples via FGSM

The Fast Gradient Sign Method (FGSM) [1] is a simple method for generating adversarial images. In our implementation, the gradients of the loss with respect to the input images have been calculated in order to execute the FGSM attack.

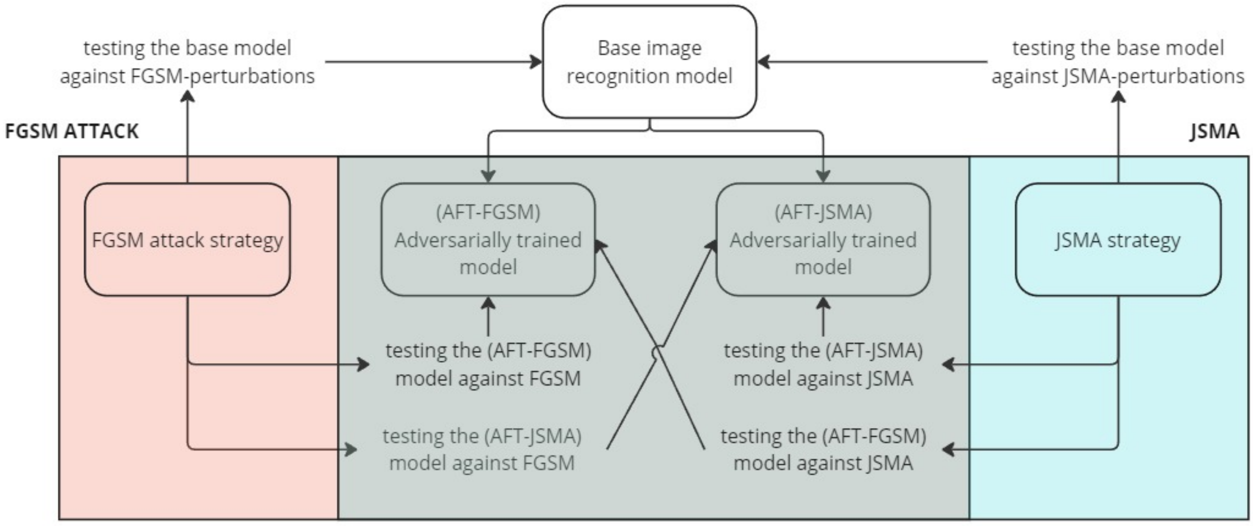
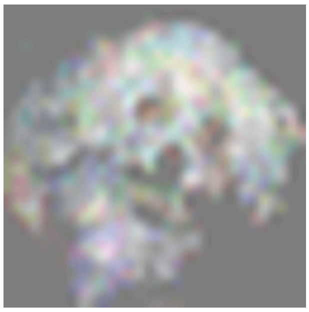


Fig. 1: Flow diagram depicting the experimental methodology.

(a) ϵ=0.007 ; original prediction: 5 (dog), adversarial prediction: 5 (dog)



(b) ϵ=0.1 ; original prediction: 5 (dog), adversarial prediction: 8 (ship)

Fig. 2: Example to showcase how output changes with different ϵ values with FGSM perturbations

The gradient indicates how to change the entire input image in a way that increases the loss, leading to misclassification. To increase the loss and mislead the network, small and fixed perturbations were added to *all pixels* in the original input image in the gradient's direction. The added perturbations remained within the bounds of ϵ and since ϵ is small, changes to input have been vaguely perceptible with human eye (an example perturbation is shown in Figure 2), however, such small changes were often sufficient to cause the model to make incorrect predictions (as shown for large ϵ in Figure 2).

To increase the model's resilience against such attacks, we exposed the base model to FGSM-perturbed samples during the training. To enable the model to correctly classify these inputs, these perturbed images were labeled with their original labels. In our implementation, we utilize the "slow start, fast decay" learning strategy proposed in [18], which has been shown to prevent the reduction in original test accuracy due to overfitting with increasing epoch count. Accordingly, we adversarially fine-tune the base ResNet-18 model that has been trained and saved. In our experiments, ϵ value of 0.01 has been used in generating the perturbed images using FGSM. Thus, FGSM adds 0.01 to those pixels, that increasing them will maximize the loss, resulting in a new, perturbed image.

C. Adversarial Samples via JSMA

The Jacobian-based Saliency Map Attack (JSMA) [19] is a more complex method of generating adversarial examples compared to FGSM. It modifies specific pixels in the input image so that it will be misclassified by the target model.

The method first computes the Jacobian matrix, *i.e.,* gradients, of the model's output (*i.e.,* prediction) with respect to the input features (*i.e.,* image pixels). JSMA uses these gradients to decide which pixels to modify in the image to cause the model to misclassify the image.

JSMA is generally applied as a targeted attack. However, as we aim to observe a generalized attack behavior as opposed to generating samples to be classified as a specific output class, instead of focusing on directing the perturbation toward a specific target class, we modify the algorithm to perturb the input image in a way that maximizes the likelihood of misclassification (without caring about the resulting class). For this, in our code, the implementation tries to alter the image in such a way that it should be classified as the original class. This is a defensive approach to ensure that even with perturbations, the model retains its classification. We prepare the attack strategy based on the true class label of the original image.

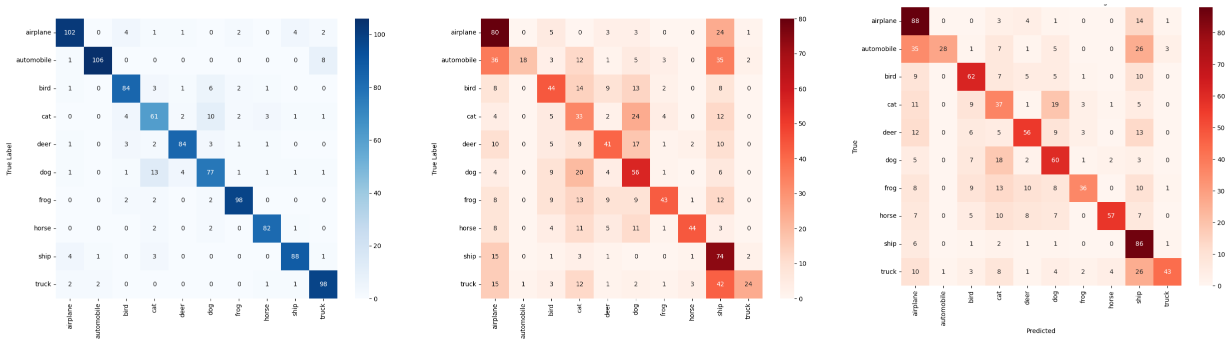
V. Experiments

The model implementation has been done on the PyTorch framework. The $60,000$ images in the CIFAR-10 dataset aside for testing, while the remaining $50,000$ images are used for training.

Each experiment in this section has been repeated with three random seeds, ensuring a different set of images have been applied perturbations each time; and the averages are reported.

A. Comparison of Attacks on Base Model

We first generate benchmark results on the base model, to verify whether the generated adversarial samples are effective



(a) FGSM-orig (b) FGSM-adv (c) JSMA-adv

Fig. 3: Confusion matrices demonstrating classification results on 1000 original (left), FGSM-perturbed (middle) and JSMA-perturbed (right) images.

in reducing the accuracy of the base model. The model has been trained until overfitting was observed.

We test the base model with 50, 100, 200, 500 and $1,000$ images, the original and its respective perturbed image. Table I demonstrates the results. It is observed that in general, FGSM causes greater reduction in accuracy compared to JSMA, causing the samples to be misclassified due to the introduced perturbations. The relatively higher accuracy with JSMA may also be attributed to the untargeted JSMA implementation. It must be noted that the differences between the accuracy of the original samples (*e.g.,* the accuracy of ResNet-18 on a 50-image test set that are reported on row 1 and row 3) is due to the random sample set that is chosen in each experiment.

TABLE I: Attacks on the Base Model

		50 images	100 images	200 images	500 images	1000 images
FGSM	ORIG	90.00	89.67	83.83	85.87	87.54
	ADV	45.33	45.00	43.50	43.13	45.12
JSMA	ORIG	86.67	86.33	85.33	86.13	86.77
	ADV	53.33	54.00	54.17	52.73	54.33

Figure 3 depicts the confusion matrices for the classification performance of the base model on $1,000$ original, $1,000$ FGSM-perturbed and $1,000$ JSMA-perturbed images. The ratio of the accuracy in classifying adversarial images to that of classifying the original images is 0.51 for FGSM-perturbed images, and is 0.62 for JSMA-perturbed images. In other words, for every 100 original images that are correctly classified, the FGSM attack has 49 of them misclassified while JSMA has 38 of them misclassified.

B. *Robustness of the AFT-FGSM Model*

Next, we begin investigating whether the adversarial training using FGSM-perturbed images renders the model more resilient against FGSM and/or JSMA attacks. In these experiments, as discussed in Section IV adversarial fine tuning (AFT) has been performed in one epoch to optimize the accuracy of original images. Our saved base model (that had been trained using $50,000$ images) has been fine tuned with 500 perturbed images that have been generated using FGSM; we call this model Adversarially Fine Tuned model trained using FGSM-perturbed images, *i.e., AFT-FGSM*.

TABLE II: Attacks on the FGSM-Fine-Tuned Model

		50 images	100 images	200 images	500 images	1000 images
FGSM	ORIG	79.33	77.00	75.00	77.07	75.83
	ADV	54.67	58.67	53.67	52.93	51.43
JSMA	ORIG	76.67	72.33	76.33	75.00	74.93
	ADV	66.00	58.67	63.83	62.20	62.37

Table II presents the change compared to the base model. Due to the training with adversarial samples, the accuracy of original samples have also dropped by $\sim 12\%$. We observe 22.2% increase in the accuracy against an FGSM attack, and 16.6% increase against a JSMA attack.

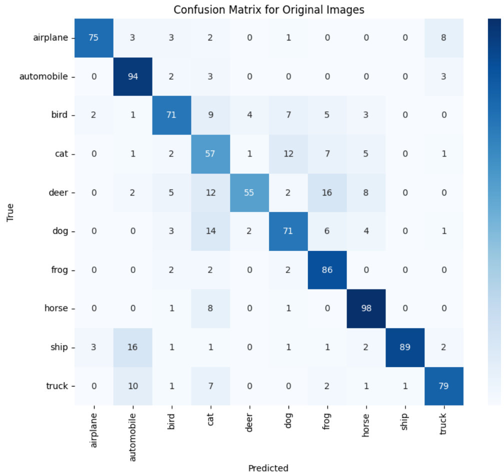
If we comparatively study the ratio of the accuracies of adversarial images to the original images, which were 0.51 and 0.62 for FGSM and JSMA respectively on the base model, FGSM-training increased these ratios to 0.71 for FGSM and 0.83 for JSMA. In other words, introducing FGSM-perturbed images in the model training improved the model's resilience against FGSM attacks by 39% and against JSMA attacks by 34%.

Figure 4 demonstrates the classification of $1,000$ images, by the adversarially fine-tuned model upon testing against original, FGSM-perturbed, and JSMA-perturbed images. The selection of original images in the two attacks may be different, hence we demonstrate both sets of originals.

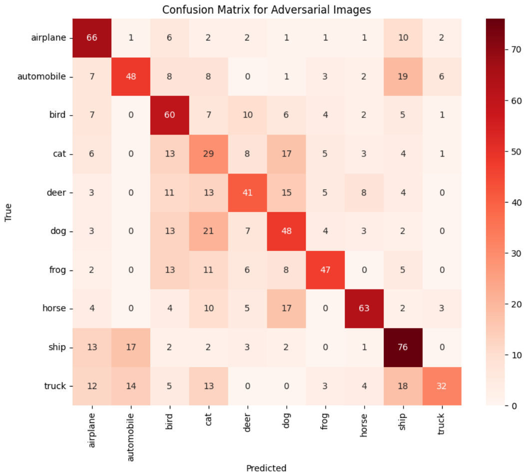
C. *Robustness of the AFT-JSMA Model*

Our second adversarially trained model is Adversarially Fine Tuned model trained using JSMA-perturbed images, *i.e., AFT-JSMA*. In these experiments, our saved base model has been fine tuned with 500 perturbed images that have been generated using JSMA.

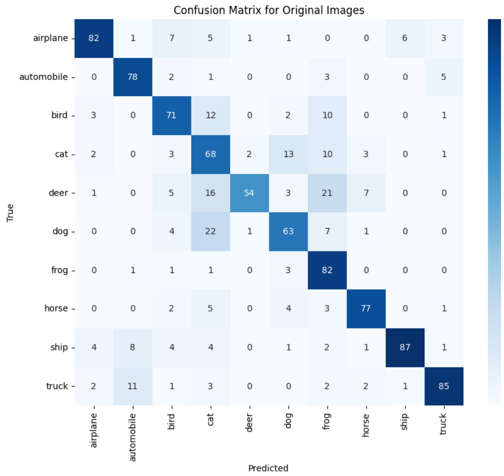
Table III presents the change compared to the base model. Due to the training with adversarial samples, the accuracy of original samples have also dropped by $\sim 14\%$. We observe 45.6% increase in the accuracy against an FGSM attack, and 42.8% increase against a JSMA attack. We attribute the slightly smaller increase compared to FGSM, to the way JSMA is implemented in this work. Figure 5 demonstrates the confusion matrices for $1,000$ original, $1,000$ FGSM-perturbed and $1,000$ JSMA-perturbed images.



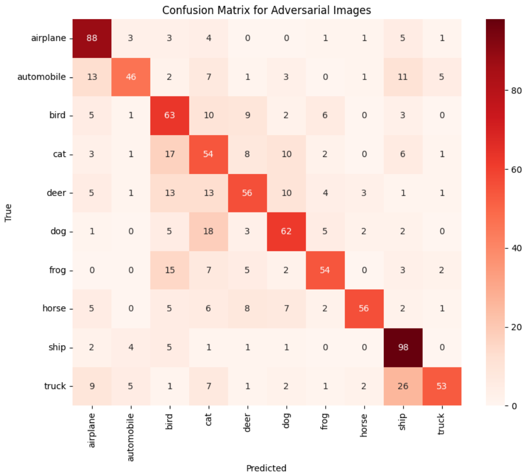
(a) Original images set 1



(b) Set 1 with FGSM perturbation



(c) Original images set 2



(d) Set 2 with JSMA perturbation

Fig. 4: Confusion matrices for classifying 1,000 images on the AFT-FGSM model; original images (on the left) and FGSM-perturbed (b) and JSMA-perturbed (d) versions of these images (on the right).

TABLE III: Attacks on the JSMA-Fine-Tuned Model

		50 images	100 images	200 images	500 images	1000 images
FGSM	ORIG	76.00	74.33	74.17	74.27	74.87
	ADV	58.67	57.00	52.17	55.40	53.50
JSMA	ORIG	68.00	76.50	73.00	76.7	75
	ADV	64.00	66.50	64	68.6	64.9

VI. Conclusions and Future Work

Adversarial machine learning (AML) is a critical potential threat against AI-based systems. In this work, we first performed two different AML attacks, Fast Gradient Sign Method (FGSM) and Jacobian-based Saliency Map Attack (JSMA), to an image recognition model and analyzed the change in accuracy after each attack. To render the system more resilient to AML attacks, we included a portion of adversarial samples in the dataset that we trained the system with. To study the impact of training with one type of attack to the resilience against either attack type, we separately performed the training with each attack and tested against both attacks.

Our experiments showed that adversarial training with FGSM perturbed images increased resilience against both FGSM and JSMA attack types, albeit with greater robustness for the former type. For the JSMA, which can be applied as a targeted attack, we chose the target class as the original class, to improve resilience, and showed that this adversarial training rendered the model robust against both attacks.

Our study highlights approaches to fortify image recognition models against AML attacks. We believe that these results can help data scientists understand the vulnerabilities of their system and design solutions to mitigate the undesirable impacts of AML attacks.

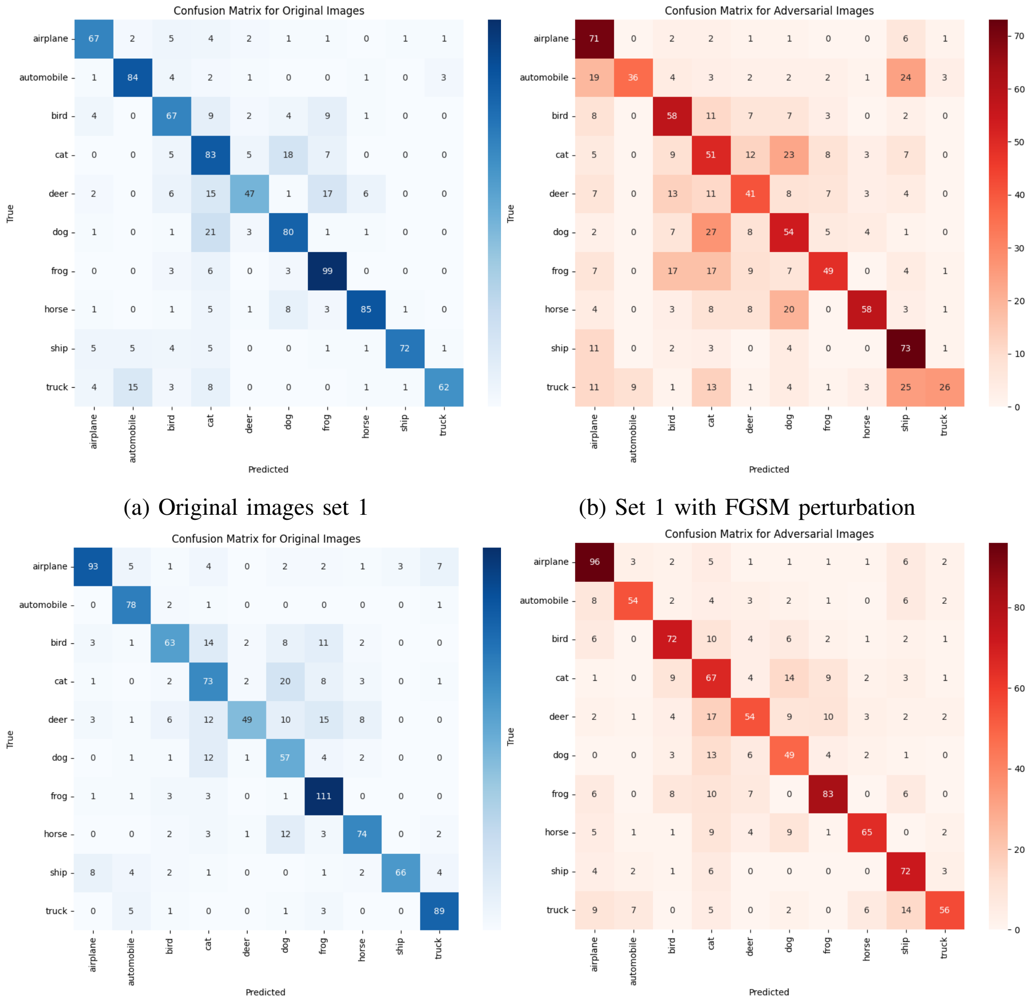
Our future work will extend this study towards comparison with other adversarial attack types such as Generative Adversarial Networks (GAN). Additionally, we are also interested in using adversarial training with JSMA with the target class different from the original label.

Acknowledgement

The authors would like to thank Yaren Akdemir for insightful discussions during this study.

References

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.



(a) Original images set 1

(b) Set 1 with FGSM perturbation

(c) Original images set 2

(d) Set 2 with JSMA perturbation

Fig. 5: Confusion matrices for classifying 1,000 images on the AFT-JSMA model, the original images (on the left) and FGSM-perturbed (b) and JSMA-perturbed (d) versions of these images (on the right).

[2] G. R. Machado, E. Silva, and R. R. Goldschmidt, "Adversarial machine learning in image classification: A survey toward the defender's perspective," *ACM Computing Surveys (CSUR)*, vol. 55, no. 1, pp. 1–38, 2021.

[3] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, "Adversarial classification," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 99–108.

[4] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, 2011, pp. 43–58.

[5] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[6] V. Zantedeschi, M.-I. Nicolae, and A. Rawat, "Efficient defenses against adversarial attacks," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 39–49.

[7] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018, pp. 2154–2156.

[8] I. Rosenberg, A. Shabtai, Y. Elovici, and L. Rokach, "Adversarial machine learning attacks and defense methods in the cyber security domain," *ACM Computing Surveys (CSUR)*, vol. 54, no. 5, pp. 1–36, 2021.

[9] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent advances in adversarial training for adversarial robustness," *arXiv preprint arXiv:2102.01356*, 2021.

[10] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv preprint arXiv:1705.07204*, 2017.

[11] D. Jakubovitz and R. Giryes, "Improving dnn robustness to adversarial attacks using jacobian regularization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 514–529.

[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[13] A. C. Serban, E. Poll, and J. Visser, "Adversarial examples-a complete characterisation of the phenomenon," *arXiv preprint arXiv:1810.01185*, 2018.

[14] P. Rathore, A. Basak, S. H. Nistala, and V. Runkana, "Untargeted, targeted and universal adversarial attacks and defenses on time series," in *2020 international joint conference on neural networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[15] J. Liu, M. Nogueira, J. Fernandes, and B. Kantarci, "Adversarial machine learning: A multilayer review of the state-of-the-art and challenges for wireless and mobile systems," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 123–159, 2021.

[16] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009, toronto, ON, Canada. [Online]. Available: https://www.cs.toronto.edu/ kriz/learning-features-2009-TR.pdf

[17] Y. Wu, L. Liu, C. Pu, W. Cao, S. Sahin, W. Wei, and Q. Zhang, "A comparative measurement study of deep learning as a service framework," *IEEE Transactions on Services Computing*, vol. 15, no. 1, pp. 551–566, 2019.

[18] A. Jeddi, M. J. Shafiee, and A. Wong, "A simple fine-tuning is all you need: Towards robust deep learning via adversarial fine-tuning," *arXiv preprint arXiv:2012.13628*, 2020.

[19] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 2016, pp. 372–387.