Analysis of Adversarial Fine-Tuning 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 research into manipulating deep learning models is ongoing, one mitigation approach involves enhancing a model's resilience by introducing adversarial samples into its training process. This process, known as adversarial training, typically entails training a model from scratch with both original and adversarially-perturbed inputs to improve robustness. In this work, we specifically investigate the effects of adversarial fine-tuning in strengthening model robustness against various adversarial attack types. Adversarial fine-tuning, a variation of adversarial training, involves using a pre-trained model and refining it with adversarial samples to improve robustness without reaching the limits of overfitting. 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 data sets that we used to fine-tune of two independent ResNet-18 models on the CIFAR-10 dataset. Obtaining two adversarially fine-tuned models, we compare their accuracies upon each attack and discuss the impact on model resiliency when the dataset used for fine-tuning includes samples generated by different approaches.**

Our results highlight that the FGSM-fine-tuned (adversarially fine-tuend) model earns the model greater resilience against FGSM attacks compared to JSMA attacks. Adversarially fine-tuned model with JSMA perturbations, 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]. As much as adversarial training is efficient, it is also very resource intensive. A more effective approach is adversarial fine-tuning. Here, instead of training a model from scratch, adversarial examples are used to further train a pre-existing model without reaching the boundries of overfitting. This allows strengthening the robustness of the model without extensive pre-training processes. It can be stated that whereas adversarial training is fully tasked with developing robust models, adversarial fine-tuning is concerned with enhancing the robustness of already available models.

In this work, adversarial fine-tuning has been studied as a defence mechanism against adversarial attacks in image recognition systems. Particularly, the impact of fine-tuning 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]. In one study, a variation of adversarial training has been used to train a classifier with adversarial images crafted by an ensemble of DNNs [10]. 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. This is also the cause for the challenge of crafting adversarial perturbations that can evade model detection while remaining imperceptible 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 fine-tuning) and thus obtained two separate adversarially fine-tuned models. To comparatively analyze the behavior, we then performed both attacks on each adversarially fine-tuned model. Figure 1 demonstrates the methodology followed in this work.

A. Target Model

We employed a ResNet-18 architecture, a widely-used 18-layer deep convolutional neural network, as the foundation for our target model. This model was trained from scratch on the CIFAR-10 dataset until an accaptable performance of accuracy was achieved, making it a suitable target for adversarial attacks.

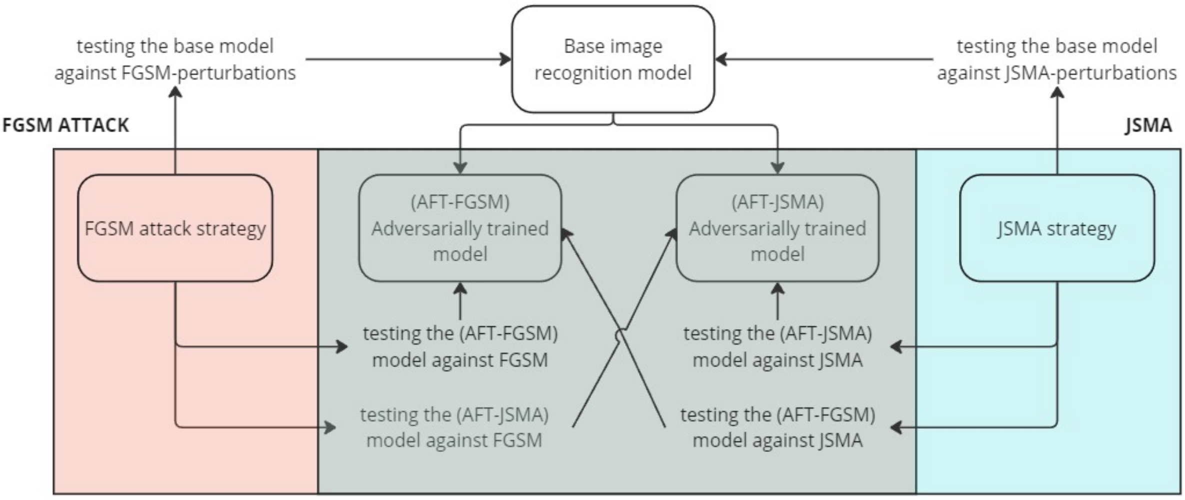
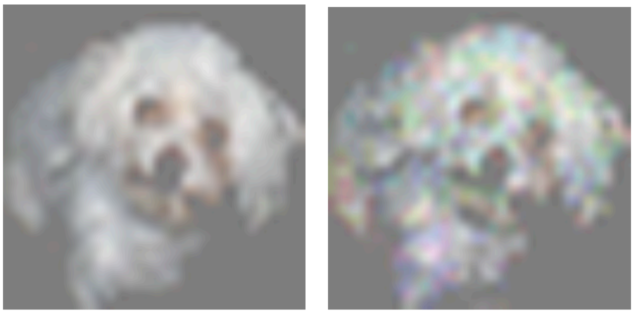

testing the base model against FGSM-perturbations
Base image recognition model
testing the base model against JSMA-perturbations
FGSM ATTACK
JSMA
FGSM attack strategy
(AFT-FGSM) Adversarially trained model
(AFT-JSMA) Adversarially trained model
JSMA strategy
testing the (AFT-FGSM) model against FGSM
testing the (AFT-JSMA) model against JSMA
testing the (AFT-JSMA) model against FGSM
testing the (AFT-FGSM) model against JSMA

Fig. 1: Flow diagram depicting the experimental methodology.

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; the learning rate hyperparameter was optimized during the experiments. 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. We trained the model on the training set of 50,000 images from the CIFAR-10 dataset; this constitutes our base image recognition model in Figure 1.



(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

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.

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).

C. Adversarial Samples via JSMA

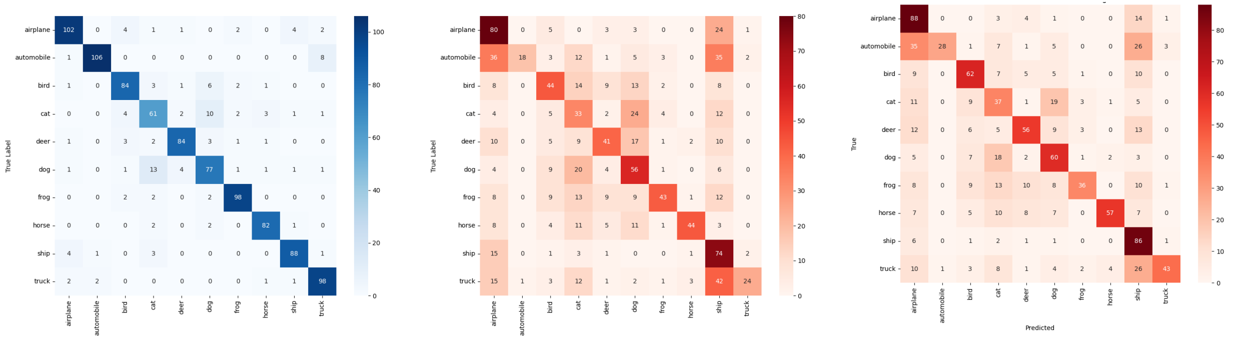
The Jacobian-based Saliency Map Attack (JSMA) [18] is a more complex method of generating adversarial examples compared to FGSM. It modifies specific pixels in the input image so that it will be misclassified by the target model.

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). We implement the Jacobian-based Saliency Map Attack (JSMA) in our code, aiming at perturbing the input image such that its classification changes while keeping the perturbation as small as possible. The JSMA strategically modifies specific pixels in the image, targeting those that most significantly influence the model's classification decision. By perturbing these pixels, the attack aims to mislead the model into making an incorrect prediction while preserving the image's overall appearance.

V. Experiments

To increase the model's resilience against such attacks, we fine-tuned the base model with FGSM-perturbed samples. To enable the model to correctly compute classification accuracy, these perturbed images were labeled with their original labels.



(a) ResNet-18 w/o attack

(b) FGSM attack

(c) JSMA attack

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

In our implementation, we utilize the "slow start, fast decay" learning strategy proposed in [19], 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 model that has been 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 the pixels that increasing them will maximize the loss, resulting in a new, perturbed image.

The model implementation has been done on PyTorch framework. The $10,000$ images in the CIFAR-10 dataset were set 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 in reducing the accuracy of the base model. We test the base model with 50, 100, 200, 500 and $1,000$ images, with both 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 be attributed to the untargeted JSMA implementation. It must be noted that the differences between the two accuracy values 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 accuracy in classifying adversarial images compared to 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 causes 49 of them to be misclassified, while the JSMA attack causes 38 misclassifications.

B. Robustness of the AFT-FGSM Model

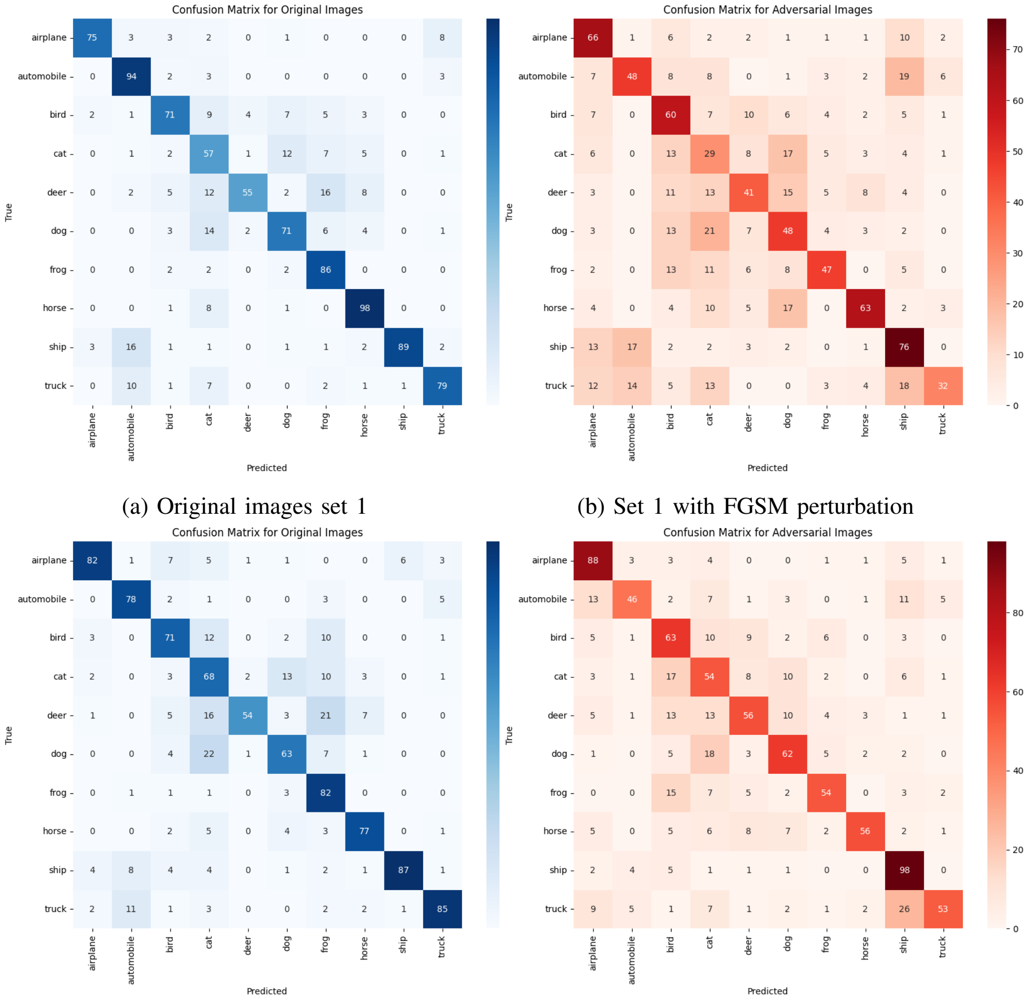
Next, we begin investigating whether the adversarial fine-tuning using FGSM-perturbed images renders the model more resilient against FGSM and/or JSMA attacks. In these experiments, as discussed in Section IV adversarial fine tuning (AFT) has been performed in one epoch to optimize the accuracy of original 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. As an undesirable byproduct of training with adversarial samples, the accuracy of original samples have also dropped by $\sim$ 12%. We observe on average 22.2% increase in the accuracy against an FGSM attack, and a 16.6% increase against a JSMA. If we comparatively study the ratios of classification accuracies for adversarial and original images, which were 0.51 and 0.62 for FGSM and JSMA attacks respectively on the base model, AFT-FGSM increased these ratios to 0.71 for FGSM attack and 0.83 for JSMA.

Figure 4 demonstrates the classification of $1,000$ images by the adversarially fine-tuned model when tested on original, FGSM-perturbed, and JSMA-perturbed images. The selection of original images in the two attacks may be different, hence we demonstrate the originals prior to each attack separately. Selected confusion matrices demonstrate



	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
airplane	75	3	3	2	0	1	0	0	0	8
automobile	0	94	2	3	0	0	0	0	0	3
bird	2	1	71	9	4	7	5	3	0	0
cat	0	1	2	57	1	12	7	5	0	1
deer	0	2	5	12	55	2	16	8	0	0
dog	0	0	3	14	2	71	6	4	0	1
frog	0	0	2	2	0	2	86	0	0	0
horse	0	0	1	8	0	1	0	98	0	0
ship	3	16	1	1	0	1	1	2	89	2
truck	0	10	1	7	0	0	2	1	1	79

(a) Original images set 1

	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
airplane	66	1	6	2	2	1	1	1	10	2
automobile	7	48	8	8	0	1	3	2	19	6
bird	7	0	60	7	10	6	4	2	5	1
cat	6	0	13	29	8	17	5	3	4	1
deer	3	0	11	13	41	15	5	8	4	0
dog	3	0	13	21	7	48	4	3	2	0
frog	2	0	13	11	6	8	47	0	5	0
horse	4	0	4	10	5	17	0	63	2	3
ship	13	17	2	2	3	2	0	1	76	0
truck	12	14	5	13	0	0	3	4	18	32

(b) Set 1 with FGSM perturbation

	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
airplane	82	1	7	5	1	1	0	0	6	3
automobile	0	78	2	1	0	0	3	0	0	5
bird	3	0	71	12	0	2	10	0	0	1
cat	2	0	3	68	2	13	10	3	0	1
deer	1	0	5	16	54	3	21	7	0	0
dog	0	0	4	22	1	63	7	1	0	0
frog	0	1	1	1	0	3	82	0	0	0
horse	0	0	2	5	0	4	3	77	0	1
ship	4	8	4	4	0	1	2	1	87	1
truck	2	11	1	3	0	0	2	2	1	85

(c) Original images set 2

	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
airplane	88	3	3	4	0	0	1	1	5	1
automobile	13	46	2	7	1	3	0	1	11	5
bird	5	1	63	10	9	2	6	0	3	0
cat	3	1	17	54	8	10	2	0	6	1
deer	5	1	13	13	56	10	4	3	1	1
dog	1	0	5	18	3	62	5	2	2	0
frog	0	0	15	7	5	2	54	0	3	2
horse	5	0	5	6	8	7	2	56	2	1
ship	2	4	5	1	1	1	0	0	98	0
truck	9	5	1	7	1	2	1	2	26	53

(d) Set 2 with JSMA perturbation

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

that adversarial training with FGSM-perturbed images improved the true positive rate (recall) for almost all classes against FGSM attack more compared to that with JSMA.

C. Robustness of the AFT-JSMA Model

Our second fine-tuned model, *i.e., AFT-JSMA*, was trained using JSMA-perturbed images. 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; and Figure 5 demonstrates the confusion matrices for $1,000$ original, $1,000$ FGSM-perturbed and $1,000$ JSMA-perturbed images. As was the case with FGSM, the training with JSMA-perturbed samples reduced the accuracy of original samples by $\sim 14\%$. We observe on average 24.6% increase in the accuracy against an FGSM attack, and a 22.2% increase against a JSMA attack. We attribute the slightly smaller increase compared to AFT-FGSM, as well as slightly greater robustness against the FGSM attack, to the way JSMA is implemented as an untargeted attack in this study.

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

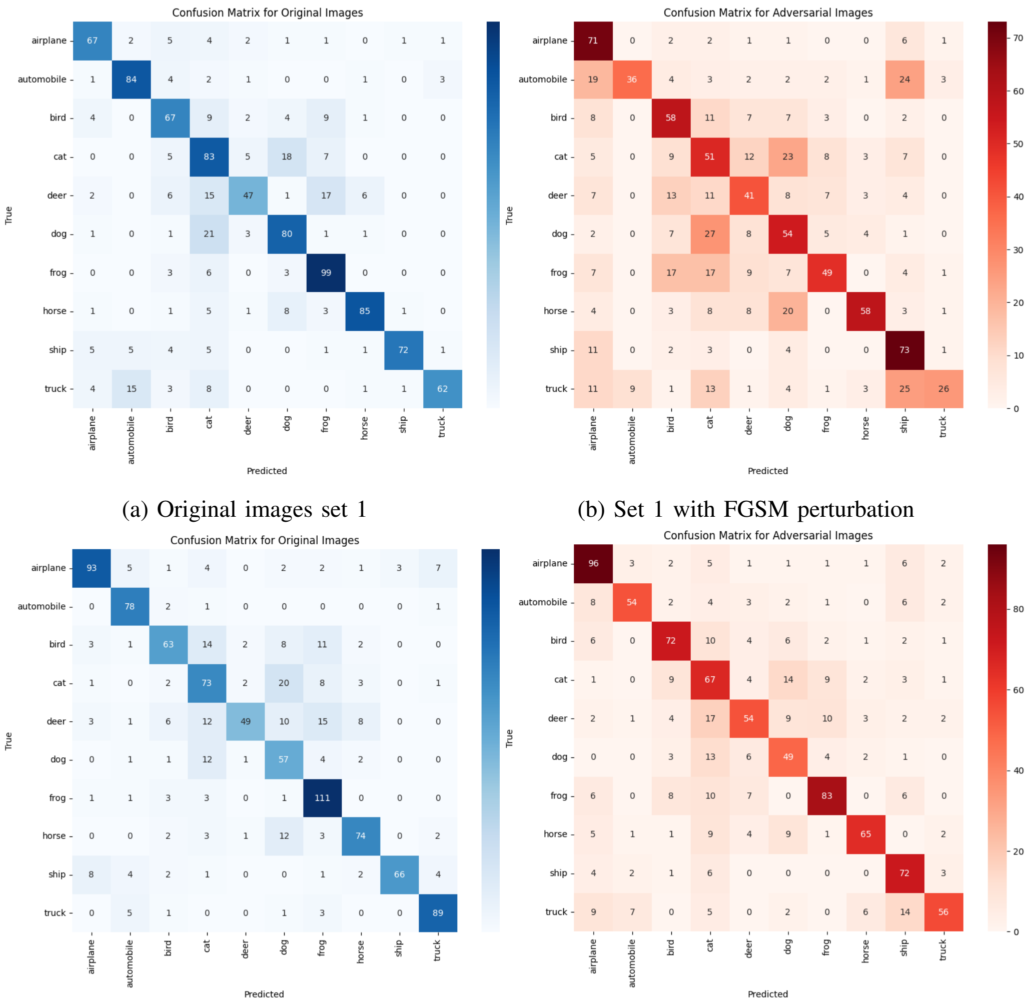
		50 images	100 images	200 images	500 images	1000 images
FGSM	ORIG	76.00	74.33	74.17	74.27	74.87
	ADV	58.67	57.00	52.17	55.40	53.50
JSMA	ORIG	68.00	76.50	73.00	76.7	75
	ADV	64.00	66.50	64	68.6	64.9

VI. Conclusions and Future Work

Adversarial machine learning (AML) poses a significant threat to the reliability of AI-based systems. In this work, we investigated the impact of adversarial fine-tuning on the robustness of a ResNet-18 model against FGSM and JSMA attacks.

We first trained a ResNet-18 model from scratch on the CIFAR-10 dataset to establish the baseline. Afterwards, we fine-tuned the model using both FGSM and JSMA for generating adversarial examples to achieve two different fine-tuned models (AFT-FGSM, AFT-JSMA).

Our study shows that fine-tuning with perturbed images substantially improves robustness against both FGSM and JSMA attacks. Specifically, models fine-tuned on FGSM-perturbed images acquire resilience to FGSM attacks. Conversely, models fine-tuned with JSMA-perturbed samples achieve high accuracy against both types of attacks, with the AFT-JSMA model's accuracy increase on FGSM attacks comparable to that of the AFT-FGSM model on



(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).

FGSM attacks. Notably, fine-tuning with JSMA images enhances robustness against JSMA attacks by approximately 25.2% over the FGSM-fine-tuned model grants robustness against JSMA attacks.

A trade-off in accuracy on non-perturbed images was also observed: fine-tuning with FGSM samples led to a $\sim$ 12% decrease in accuracy on original images, while fine-tuning with JSMA samples resulted in a slightly larger decrease of $\sim$ 14%. These results highlight that attack-specific fine-tuning significantly enhances model robustness but also underscores the trade-offs in accuracy on non-adversarial inputs.

Our future work will extend this study towards comparison with other adversarial attack types such as diffusion-based attacks and Generative Adversarial Networks (GAN). Additionally, we are also interested in using adversarial fine-tuning 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.

[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] 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.

[19] 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.