Evaluating Auxiliary Frequency-basis Augmentation under adversarial attacks

DANIËL KUIPER, University of Twente, The Netherlands

In the realm of machine learning, ensuring the robustness of models against adversarial attacks is critical, particularly in applications such as healthcare, autonomous systems and security. This paper investigates the efficacy of Auxiliary Fourier-basis Augmentation (AFA) as a defense mechanism against adversarial perturbations in computer vision models. AFA introduces additive Fourier-basis noise to enhance model resilience, complementing traditional visual augmentation methods. We evaluate the performance of AFA across the CIFAR-10 dataset using a variety of adversarial attacks including Auto-PGD, FAB and the Square Attack under different L_∞ norms. Experimental results demonstrate that AFA consistently enhances model robustness against adversarial attacks, mitigating accuracy degradation under adversarial attacks compared to models without AFA augmentation. We analyze perturbation patterns in the frequency domain to understand how AFA alters the perturbations, showing significant defense against low and high-frequency perturbations while highlighting vulnerabilities in the medium frequency ranges.

Additional Key Words and Phrases: adversarial attacks, robustness, image classification, frequency-basis, augmentation

1 INTRODUCTION

In a world where machine learning systems are becoming increasingly intertwined with our daily lives, the importance of their robustness increases. Every day, more computer vision models are being used in the everyday lives of millions of people. People start to rely on these models more and more in vital sectors of our society such as healthcare, autonomous transportation, surveillance and security, industry automation, etc. Human evaluation is factored out of more and more processes in this rapidly changing world. However, these human-replacing models can often be broken by attacks where even humans can't see what changed between two images. This is a cause of worry and the reason that robustness of computer vision models is incredibly important.

In previous work [23], Auxiliary Fourier-basis Augmentation (AFA) was proposed; A technique to augment the images of a dataset in the frequency domain to fill the robustness gap left by visual augmentation. Additive noise generated by Fourier basis function was added on each of the features (RGB) independently.

The primary objective of this research is to assess the robustness of Auxiliary Fourier-basis Augmentation (AFA) under adversarial attacks. Adversarial attacks change the input image minimally, such that the change is imperceptible for a human, while causing a misclassification. For example, an adversarial attack can make a model classify a an image of a dog as a cat, while a human can only see the original dog.

TScIT 41, July 5, 2024, Enschede, The Netherlands

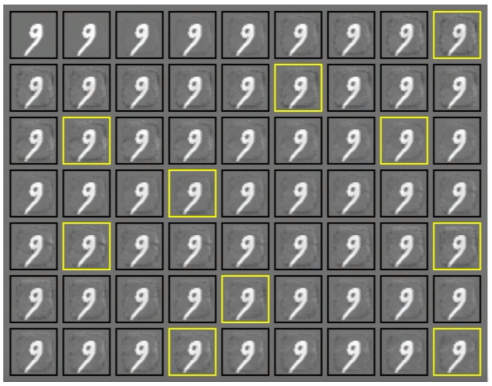

Fig. 1. An iteratively perturbed image to classify an image as all digits 0-9 in order. The yellow bordered images is where the model confidently classifies the digit incorrectly. (Goodfellow, 2016) [26]

In addition to evaluating the robustness of AFA, another objective is to explore characteristics of adversarial examples using the frequency domain. By examining the frequency ranges of the perturbations, we aim to understand the underlying mechanisms that contribute to the enhanced robustness of AFA models.

This all concludes to the main question to be answered: *How does the robustness of AFA models change under different types of adversarial attacks?*

To answer this we will analyse patterns in the attacks and models. For this we will define some subquestions:

- RQ1 How does the accuracy of AFA models change under increasing perturbation norms in comparison to non-AFA models?
- RQ2 Is there a general range in the frequency domain where successful adversarial examples are situated?
- RQ3 Is there a difference in adversarial examples between AFA models and non-AFA models in the frequency domain?

2 RELATED WORKS

2.1 Adversarial attacks

Adversarial attacks involve making small changes to input data, which are often imperceptible to humans but can cause the model to make misclassified predictions, first demonstrated by Szegedy et al. (2013) [21] and Goodfellow et al. (2015) [10]. An adversarial example can be generated to make the model predict a target class, which are called targeted attacks. An adversarial example can also be generated to simply misclassify a prediction, called untargeted attacks.

In figure 1, an image was iteratively perturbed to make the model classify all possible digits 0-9 in order. While for a human, the changes are barely perceptible, for the model it classifies it as a whole different digit.

Projected Gradient Descent (PGD), introduced by Madry et al. (2017) [14], iteratively applies FGSM to generate more powerful adversarial examples. DeepFool, presented by Moosavi-Dezfooli et al. (2016) [16], aims to find the minimal perturbation required to misclassify an input sample by iteratively computing the distance from the sample to the decision boundary of the neural network. APGD [5] is a parameter-free PGD attack.

Fast Adaptive Boundary attach (FAB), is introduced by Croce and Hein in 2020 [**FAB-croce**]. FAB is a white-box attack that finds the minimum perturbation necessary to change the class of a given image.

The **Square Attack** (Andriushchenko et al., 2020) [2] is a black box attack that randomly searches for adversarial examples that cause for a misclassification. It does it in a structured way to make it computationally feasible.

One common norm used in adversarial attacks is L_p norm, where p can be any positive real number. The L_p norm of a vector x is calculated as:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

Here, x_i represents the pixels of the image x, and n is the dimensionality of x. For example, when $p = 1$, it's the sum of absolute values of elements (Manhattan distance). When $p = 2$, it's the Euclidean distance and when $p = \infty$, it determines the maximum absolute difference. In some attacks, the goal is often to minimize the perturbation under a given perturbation norm while still causing misclassification [10]. Later research has explored various strategies for defending against adversarial attacks, including robust optimization [14] and adversarial training [10][27].

2.2 Data augmentation

Data augmentation is a technique to increase the accuracy of models by generating more data variety from existing data points. This method can also be used to increase robustness by adding variation in the dataset. In computer vision, data augmentation techniques include image transformations (e.g., rotation, translation, scaling), color and brightness adjustments, cropping, flipping, and noise addition.

Research by Krizhevsky et al. (2012) [12] and Simonyan et al. (2014) [19] demonstrated the effectiveness of such augmentation techniques in convolutional neural networks (CNNs). Recent research in data augmentation led to techniques like AutoAugment (Cubuk et al., 2019) [7] and RandAugment (Cubuk et al., 2020) [8], which automatically search optimal augmentation policies based on the dataset.

This paper will investigate two augmentation techniques. The first one, PRIME [15], is a general data augmentation method that enhances robustness to common corruptions through the use of simple yet diverse max-entropy image transformations. The second one, TrivialAugment (TA) [17], is a automatic augmentation method that applies a single random visual augmentation to each image.

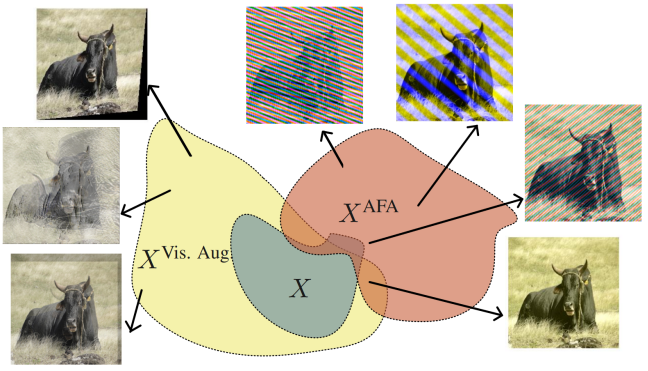

Fig. 2. Fourier-basis additive noise augmentation is a complement to visual augmentation techniques. (Vaish et al., 2024) [23]

2.3 Auxiliary Fourier-basis Augmentation (AFA)

Research by Gilmer et al. (2018) [9] and Tsuzuku et al. (2019) [22] showed that models often encounter a high error rate when introduced with adversarial examples containing Fourier-basis additive noise. Dong et al. (2019) [29] further explored this issue, demonstrating that augmentation causes an increase of robustness in against high-frequency corruptions at the cost of a decrease in accuracy against low-frequency corruptions. Vaish et al. (2024) [23] identified this augmentation gap and proposed AFA – an augmentation technique which utilizes Fourier-basis functions as additive noise. Random frequencies and directions are sampled from uniform distributions which are added channel wise over the original image. The frequency sampling is defined as $f \sim \mathcal{U}_{[1,M]}$ where f is the frequency and M is the image size. Figure 2 shows the augmentation method.

Techniques have been researched to increase robustness making use of the frequency domain [4][25][24][28][13]. However, Vaish et al. (2024) [23] argues that these methods are computationally infeasible for large-scale datasets. Additive noise as Fourier-basis functions however can be applied as a feasible Fourier-basis augmentation technique for large-scale datasets.

Using AFA in combination with existing visual augmentation techniques shows a significant increase in model accuracy compared with just using these visual augmentations [23]. These improvements were done on robustness benchmarks with common corruptions and perturbations (Hendrycks and Dietterich, 2019) [11] and not on more advanced adversarial attacks.

Robustness evaluation in the frequency domain has been done with PGD and FGSM on CIFAR-10 [20]. This lead to results which are promising, but have only been evaluated on a small-scale dataset with only 2 attacks. Standard evaluation on AFA under adversarial attacks is still necessary to comprehensively assess its robustness in real-world scenario's.

2.4 Standards for evaluating robustness

In recent years, research in adversarial examples has increased massively. Currently, over 9,000 papers on this subject have been written. Athalye et al. (2018) [3] state that most of these papers overestimate

the robustness of their findings. To be able to compare robustness of models between papers, Croce et al. (2021) [6] proposed a standardized benchmark using AutoAttack. AutoAttack, introduced by Croce et al. (2020) [5] is an ensemble of multiple attacks aiming to be a parameter-free, computationally affordable and user-independent benchmark.

AutoAttack comprises four distinct attacks. Initially, they introduced Auto-PGD (APGD), an extension of PGD [14], wherein the step-size parameter adjusts automatically based on input data and the model. Additionally, AutoAttack employs two variants of APGD with different loss functions: Cross-entropy loss (APGD-CE) and Difference of Logits Ratio loss (APGD-DLR). Alongside these, the existing attacks FAB [**FAB-croce**] and Square Attack [2] are integrated into AutoAttack.

AutoAttack operates in two modes: standard and individual. In the standard mode, it sequentially applies all attacks, while in the individual mode, each attack is applied separately.

3 METHODOLOGY

This project aims to evaluate models trained by Vaish et al. (2024) [23] on CIFAR-10 using a standard benchmark to analyse their accuracy under various attacks. Different levels of perturbations under the L_∞ norm will be applied for all models and attacks to assess how well AFA ranks against existing techniques.

In our analysis, we will omit attacks for specific reasons. Individual analysis will not be performed with FGSM since it is a weaker version of PGD [5]. DeepFool will not be analysed since it is alike in spirit as FAB, but it finds the minimal perturbation disregarding the indistinguishability to original image [5]. Carlini-Wagner is outperformed by FAB and PGD, just like other attacks not mentioned before (SparseFool, Linear Region, Distributionally Adversarial, ElasticNet, etc.) and thus will not be evaluated [5]. The four attacks included in AutoAttack offer completeness, encompassing both blackbox and whitebox, targeted and untargeted approaches.

Only ResNet models will be analyzed because its architecture has been extensively studied and optimized, making it a reliable benchmark for evaluating adversarial attacks. Usually ResNet models are used in benchmarks, so we use this to be able to compare it to other methods.

In this study, we focus on CIFAR-10 models for our experiments. Due to practical constraints including time limitations and GPU memory capacity, a study on ImageNet was not possible.

4 EXPERIMENT

This section introduced two experiments. The first experiment aims to analyze the robustness of various models under different adversarial attacks across a range of ϵ values. The second experiment focuses on assessing the influence of AFA when examining adversarial examples in the frequency domain

4.1 Experiment 1: Evaluating robustness under adversarial attacks

In this experiment, we will compare the success rates of different AutoAttack methods to understand which attacks are more effective against different models. The attacks (APGD cross-entropy, APGD targeted, FAB targeted, and the Square Attack) will be applied to each model. Using the L_∞ perturbation norm, we will test ϵ values ranging from 0 to 8/255 to see how model accuracy changes with increasing perturbation levels. Robust models show a slow decrease in accuracy as the perturbation increases, while non-robust models drop in accuracy more quickly. By comparing models with Adversarial Frequency Augmentation (AFA) to those without, we will assess the impact of AFA on the robustness of each model against these attacks.

4.2 Experiment 2: A Fourier perspective on robustness

We will perform the calculations in this section to analyze the frequency ranges of adversarial attacks for ResNet18 models without AFA and with AFA. All calculations are done channel-wise when relevant. The image is made grayscale before visualization.

Let O denote the original image of height H and width W. Let m be a model, and let $X_{m,a,O}$ denote the successful adversarial example generated for model m using attack a on image O.

Both O and $X_{m,a,O}$ can then be transformed using the Fourier transform. Let $A(x, y)$ be such image with coordinates x, y. Before transforming the image to the frequency domain, we will apply a Hann window function $H\{A(x, y)\}$ to reduce spectral leakage [18].

$$F(u, v) = \mathcal{F}\{A(x, y)\} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} H\{A(x, y)\} e^{-j2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)}$$

The zero-frequency component of the Fourier transform $F(u, v)$ is then shifted to the center and the absolute is taken to be left with only the magnitude values.

$$F'(u, v) = |\mathcal{F}_{\text{shift}}\{F(u, v)\}|$$

. Next, the difference between $X_{m,a,O}$ and O can then be taken, which represents the perturbation. With this perturbation we can analyze where the perturbation lies in the frequency domain. This gives us the formula

$$P_{m,a,O}(u, v) = F'(X_{m,a,O}(u, v) - O(u, v))$$

$P_{m,a,O}$ denotes the adversarial perturbation in the frequency domain.

To find a general trend, we will aggregate all successful adversarial examples on images $t \in T$, where T is the set of the images in the test set that are successfully perturbed. The failed perturbations, i.e. the perturbations which could not lead to a misclassification under a given ϵ, will not be aggregated. This will lead to the following formula:

$$P_{m,a}(u, v) = \frac{1}{|T|} \sum_{t \in T} P_{m,a,t}(u, v)$$

Where T is the test set, $P_{m,a}$ denotes the mean perturbation of all successful adversarial examples under the Fourier transform for model m and attack a. So, $P_{m,a}(u, v)$ denotes the energy of the frequency (u, v) of the average successful perturbation on the test set.

Based on the work of Abello et. al. (2021) [1], we can measure the energy intensity for each frequency based on the distance to the center. With the Fourier Transform, P is shifted to the center of the image. The "distance" is hence defined as the L1 norm from pixel (k,l) to the center since the L1 norm is the best method to

dataset	arch.	base	AFA	standard	APGD-ce	APGD-t	FAB-t	square	autoattack
CIFAR-10	ResNet18	none	n	95.15	0.00	0.02	0.07	59.19	0.00
			y	94.69	0.50	0.39	1.13	76.92	0.27
		PRIME	n	94.37	0.76	0.56	1.02	70.37	0.27
			y	94.54	0.92	1.01	1.21	72.18	0.46
		TA	n	96.20	0.01	0.04	0.00	59.30	0.00
			y	96.10	0.06	0.04	0.00	72.80	0.00

Table 1. Accuracies of CIFAR-10 ResNet models on different attacks under $L_\infty = 8/255$ in %. AutoAttack does the attacks APGD-ce, APGD-t, FAB-t and square subsequently. If an image x cannot be perturbed by APGD-ce, AutoAttack tries APGD-t, then FAB-t and lastly square.

measure distances in discrete spaces like images. We will define $E\{P_{m,a}\}$ to be the energy distribution of the permutation. From this distribution, we can learn the attack strategies of the attacks.

$$E\{P_{m,a}\}(f) = \frac{1}{|S(f)|} \sum_{(u,v)\in S(f)} P_{m,a}(u,v)$$

$S(f)$ is the set of pixel coordinates (u, v) such that the L1 distance from (u, v) to the center is equal to f. This can be defined as:

$$S(f) = \{(u,v) \mid |u - \frac{W}{2}| + |v - \frac{H}{2}| = f\}$$

We can evaluate the difference between a model and its AFA-enhanced counterpart by computing

$$D_{m',m,a} = P_{m',a} - P_{m,a}$$

, where m' represents the AFA equivalent of the model m and $D_{m,a}$ denotes the frequency domain delta between the base model m and its AFA model m' under attack a. $E\{D_{m',m,a}\}$ gives us the energy distribution difference that AFA makes. We can analyze this attack strategy difference between a model and its AFA counterpart to see the frequency ranges against which AFA defends.

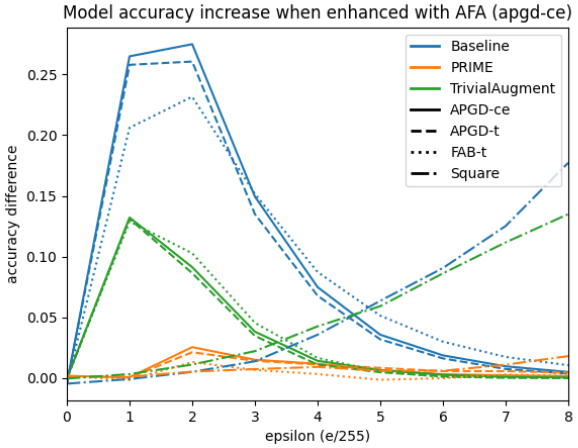

Fig. 3. The accuracy difference between a base model and its AFA enhanced model.

5 RESULTS

5.1 Experiment 1

5.1.1 Attack differences. In table 1, the accuracies of ResNet models are presented under different attacks with L_∞, $\epsilon = 8/255$.

First of all, by looking at table 1, we notice that all models are not robust against the APGD and FAB attacks. APGD and FAB have access to the model's gradient, allowing it to precisely calculate which pixels to tweak to maximize the adversarial attack success. APGD Targeted, although similar to APGD-CE in using the model's gradient, imposes an additional constraint of targeting a specific class. The variability of the values in the table between attacks is a result of the accuracy deficit described in Appendix B.

The Square Attack, being a black-box attack, operates based on randomness and iterative adaptation. It does not rely on model gradients, making it inherently less effective, resulting in higher model robustness, as can be seen in figure 4. However, its black-box approach makes it more representative of real-life scenarios, where access to model gradients is typically unavailable.

5.1.2 Impact of AFA on model robustness. To discern trends in model robustness, it is needed to examine performance across varying levels of perturbation. In Fig 4, a clear decreasing trend can be seen as the perturbation increases. For each attack, the baseline AFA model shows the highest robustness against adversarial attacks for ϵ lower than 5. For larger perturbations, PRIME enhanced with AFA is marginally more robust.

The TrivialAugment model, while having the highest standard accuracy, is actually the least robust against adversarial perturbations. For all other methods, they are more robust than the baseline model, this indicates that augmentation is a good method to improve robustness if done right, confirming Dong et. al.'s findings [29].

Figure 3 shows that AFA provides a consistent improvement in robustness whether it is auxiliary to another technique or not. The PRIME model without AFA already shows higher robustness against all attacks, closely matched by the baseline model with AFA. However, models trained with TrivialAugment demonstrate the highest standard accuracy, but their accuracy significantly deteriorates when subjected to attacks, indicating low robustness. In this case, AFA does increase the robustness.

Vaish et. al. (2024) [23] showed that AFA brings a consistent improvement on robustness against common corruptions. Interestingly, figure 3 shows that this improvement does also extend to adversarial attacks.

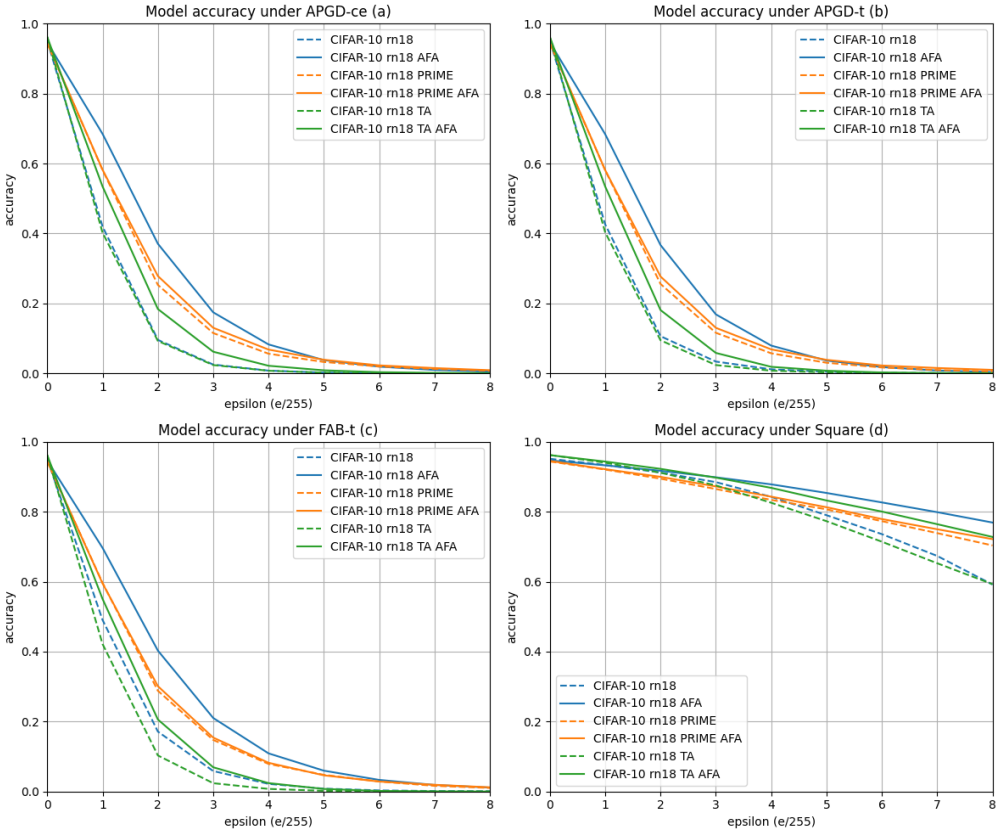

Fig. 4. The accuracy of different models against (a) APGD-ce, (b) APGD-t, (c) FAB-t and (d) Square attack. Using the L_∞ norm.

5.2 Experiment 2

First of all, let's define model **M1** to be the baseline ResNet18 CIFAR-10 model without AFA. Then, **M1A** is that **M1** model and trained with AFA. Let **M2** be the ResNet18 CIFAR-10 model trained with PRIME, and then consequently **M2A** is that model but enhanced with AFA. Then **M3** is the ResNet18 CIFAR-10 model trained with TrivialAugment, and **M3A** is that model enhanced with AFA.

5.2.1 Attack perturbation strategy. In figure 5, the energy distribution of the mean perturbation of all successful adversarial examples for model M1 can be seen. Here, the perturbation distribution indicates an attack strategy. If the energy is high, the perturbation is active in this region. If the energy is low, the frequency is not a lot present in the perturbation.

First of all, APGD-ce and APGD-t have the same attack strategy. This makes sense since it is the same attack. Secondly, the Square attack perturbs mainly in the low frequencies. This can be explained by the fact that it is black box and based on iterative addition of square noise. Because of its structure, and on the low number of queries used (App. B), it is statistically less likely to create high-frequency perturbations. Lastly, the FAB-t attack has structurally less perturbations in all frequencies. Further observations will continue with APGD-ce only due to its strength, AFA robustness and to keep the research scoped.

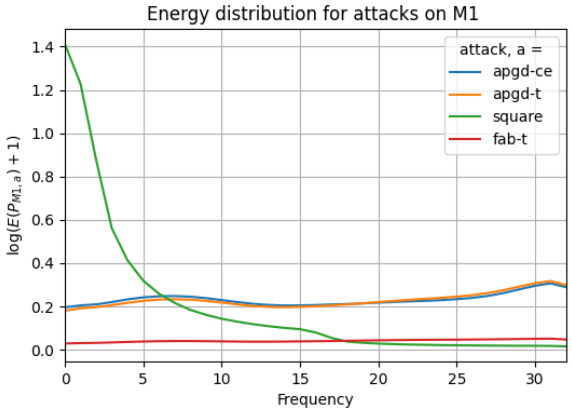

Fig. 5. Perturbation energy distributions for different attacks on M1 for $L_\infty, \epsilon = 8/255$. The energy-axis is scaled logarithmic to improve readability.

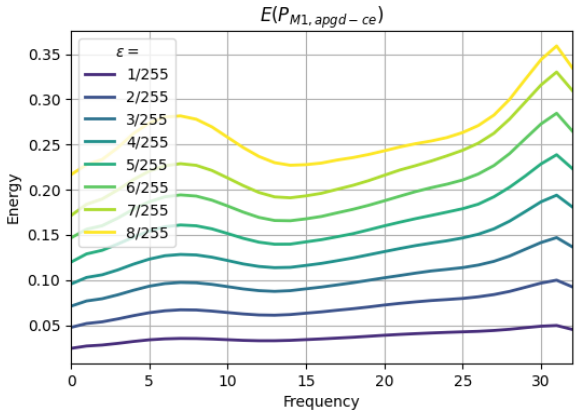

Fig. 6. The energy distribution of the mean perturbation of all successful adversarial examples under the Fourier transform for model M1 and APGD-ce.

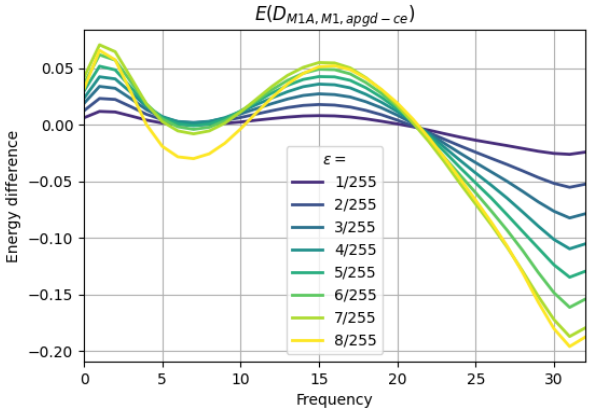

Fig. 7. The energy distribution difference between the mean perturbations on the models M1A and M1.

5.2.2 Perturbation strategies for different norms. The attack strategy we found in previous section was based on $\epsilon = 8/255$, since it is a common bench marking value. However, it is important to know if the strategy differs for other values. Figure 6 shows that as ϵ increases, a change in attack strategy can indeed be seen. Once ϵ increases, an increase in overall energy can be seen. This is due to the fact that if the attack gets more freedom in perturbing an image, it will perturb the image in all frequencies. Interestingly, the perturbation is not increased uniformly over all frequencies. Overall, the perturbation is skewed to the high-frequencies. Also, increased perturbation energy can be seen for frequencies 3 to 10 and frequencies 27 to 32. This suggests that APGD focuses its perturbation primarily on the low and high-frequencies.

5.2.3 Impact of AFA on the perturbation strategy. Now, we can compare the impact AFA has on the perturbation strategy by comparing M1 and M1A. Figure 7 shows the attack strategy difference between

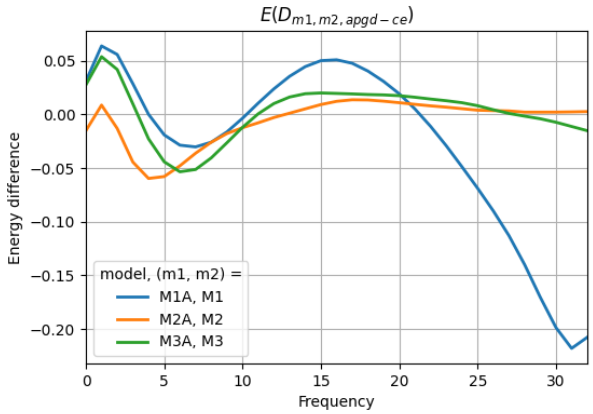

Fig. 8. Perturbation energy distribution for different models on the APGD-ce attack for $L_\infty, \epsilon = 8/255$

the AFA model M1A and the baseline model M1. We can see two distinct cases for which the perturbation strategy differs when you look at the general trend of $\epsilon = 8$.

The first case is where the energy difference is positive. Here, AFA perturbs more information. This is the case in the frequency ranges 0 to 4 and 10 to 22. The first range are the extremely low frequencies, perturbations in this region are so large that they surpass the subject in the image. The second positive range are the medium frequencies. Here, details in the image are present. Increasing perturbations in these ranges increases robustness.

For all other frequencies - the ranges 4 to 10 and 22 to 32 - the energy difference is negative, suggesting that the average perturbation on the AFA model is lower. The information in these frequency ranges are broad details (4 to 10) and the fine details (22 to 32). Decreasing perturbations in these areas increase robustness.

The trend described above also holds generally for the FAB-t attack. In Appendix C, these figures are shown for FAB-t. There are notable differences between the attacks. The overall energy is lower, and the difference between AFA and the baseline is positive until $f = 25$. Since it is a different attack, it yields different defense strategies. However, the general shape is equal, meaning that the impact of different frequency ranges does stay the same.

A theory that explains this phenomenon is that the AFA model has been trained with augmentation that adds random frequency-basis noise for all frequencies 0-32 uniformly. Because of this augmentation, it is more robust against perturbations that rely on these frequency alterations.

5.2.4 Correlation perturbation strategy and robustness. Figure 8 shows the perturbation energy distribution on different models for $\epsilon = 8/255$. The baseline models are compared against the AFA models, just like in figure 7. Here an interesting observation can be made. For the low to medium frequencies, all models show the same behavior. This indicates that the strategy described in 5.2.3 holds for other models for the low frequencies. For the high frequencies, AFA does not perturb less for models M2 and M3. This indicates that the base augmentation of PRIME and TA already show some

high-frequency defense, since AFA shows no difference. AFA does however lead to increased robustness for all models M1-M3, as we have seen in Experiment 1. As such, we can create an outlook that AFA provides an extra defense for the low-frequencies.

6 CONCLUSION

In this paper, we presented two experiments to evaluate the robustness of different models under adversarial attacks, with and without Adversarial Frequency Augmentation (AFA). The primary objective was to assess the effectiveness of Adversarial Frequency Augmentation in enhancing model robustness against adversarial attacks.

The results from Experiment 1 demonstrate that AFA significantly improves the robustness of models against adversarial attacks, particularly for lower perturbation levels. Models augmented with AFA exhibit a slower decrease in accuracy as the perturbation level increases, compared to their non-AFA counterparts. This trend was consistent across various attack types, including both white-box attacks like APGD and FAB, and the black-box Square Attack.

Experiment 2 took a Fourier perspective to analyze the frequency characteristics of adversarial perturbations. By transforming images into the frequency domain, the study revealed distinctive attack strategies employed by different adversarial attacks, APGD-ce and FAB-t. It was observed that these attacks often target specific frequency ranges, with AFA models showing altered perturbation strategies leading to increased robustness, particularly in low and medium frequencies. This shift in attack patterns suggest that AFA-induced defences change how adversarial perturbations affect different frequency bands, thereby enhancing model robustness against

7 DISCUSSION AND FUTURE WORK

As we discussed in section 5.2.3, there are multiple frequency bands that show different perturbation strategies. Figure 7 shows that is beneficial to prevent low- and high-frequency perturbations and not so much to prevent medium frequency perturbations. A theory is that AFA is inherently beneficial in preventing low- and high-frequency perturbations, but that it does not work for the medium frequencies. In future research it might be insightful to understand what differences are introduced if in the AFA augmentation process, f is not uniformly distributed over $\mathcal{U}_{[1,M]}$. Observations about the difference of an AFA model distributed over only the low-frequencies ($f \sim \mathcal{U}_{[1,25]}$) and an AFA model over only the high-frequencies ($f \sim \mathcal{U}_{[25,M]}$) will lead to insights over optimal defense against adversarial attacks. Also, a non-uniform distribution can be chosen where low- or high-frequency augmentation are more present than medium frequencies, for example a Log-Normal distribution. This will lead to observations over where in the frequency domain the successful adversarial perturbations lie and what defense works.

Additionally, it is important to state that reducing perturbation energy does not mean that the model is more robust, as we have seen in Experiment 2 with the PRIME trained models. Robustness encompasses factors beyond frequency domain observations alone. While Fourier analysis provides insights into how perturbations affect different frequency components, it doesn't capture the entirety of adversarial attack strategies. For instance, attacks like APGD and Square Attack target low-frequency components significantly, affecting model predictions despite minimal changes in high-frequency domains.

While Experiment 2 provided valuable insights using CIFAR-10, ImageNet offers a broader range of images with more complex features and higher resolutions. Analyzing models such as ResNet50 or other architectures trained with and without AFA on ImageNet would allow for a more comprehensive understanding of how AFA impacts model robustness across diverse image characteristics. This could verify if trends observed in CIFAR-10 experiments—such as the defense against high-frequency perturbations—are consistent to larger datasets.

REFERENCES

[1] Antonio A. Abello, Roberto Hirata, and Zhangyang Wang. "Dissecting the High-Frequency Bias in Convolutional Neural Networks". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2021, pp. 863–871. DOI: 10.1109/CVPRW53098.2021.00096.

[2] Maksym Andriushchenko et al. "Square Attack: A Query-Efficient Black-Box Adversarial Attack via Random Search". In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Cham: Springer International Publishing, 2020, pp. 484–501. ISBN: 978-3-030-58592-1.

[3] Anish Athalye, Nicholas Carlini, and David Wagner. *Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples*. 2018. arXiv: 1802.00420 [cs.LG].

[4] Guangyao Chen et al. "Amplitude-Phase Recombination: Rethinking Robustness of Convolutional Neural Networks in Frequency Domain". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 458–467.

[5] Francesco Croce and Matthias Hein. *Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks*. 2020. arXiv: 2003.01690 [cs.LG].

[6] Francesco Croce et al. *RobustBench: a standardized adversarial robustness benchmark*. 2021. arXiv: 2010.09670 [cs.LG].

[7] Ekin D. Cubuk et al. *AutoAugment: Learning Augmentation Policies from Data*. 2019. arXiv: 1805.09501 [cs.CV].

[8] Ekin D. Cubuk et al. "Randaugment: Practical Automated Data Augmentation With a Reduced Search Space". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2020.

[9] Justin Gilmer et al. "Motivating the Rules of the Game for Adversarial Example Research". In: *CoRR* abs/1807.06732 (2018). arXiv: 1807.06732. URL: http://arxiv.org/abs/1807.06732.

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. *Explaining and Harnessing Adversarial Examples*. 2015. arXiv: 1412.6572 [stat.ML].

[11] Dan Hendrycks and Thomas Dietterich. *Benchmarking Neural Network Robustness to Common Corruptions and Perturbations*. 2019. arXiv: 1903.12261 [cs.LG].

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks". In: *Commun. ACM* 60.6 (May 2017), pp. 84–90. ISSN: 0001-0782. DOI: 10.1145/3065386. URL: https://doi.org/10.1145/3065386.

[13] Peter Lorenz et al. *Detecting AutoAttack Perturbations in the Frequency Domain*. 2024. arXiv: 2111.08785 [cs.CV].

[14] Aleksander Madry et al. *Towards Deep Learning Models Resistant to Adversarial Attacks*. 2019. arXiv: 1706.06083 [stat.ML].

[15] Apostolos Modas et al. *PRIME: A few primitives can boost robustness to common corruptions*. 2022. arXiv: 2112.13547 [cs.CV].

[16] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. "DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.

[17] Samuel G. Müller and Frank Hutter. *TrivialAugment: Tuning-free Yet State-of-the-Art Data Augmentation*. 2021. arXiv: 2103.10158 [cs.CV].

[18] K.M.M. Prabhu. *Window Functions and Their Applications in Signal Processing*. Oct. 2013. DOI: 10.1201/9781315216386.

[19] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV].

[20] Yifan Sun. *Effectiveness of Fourier-basis noise on improving adversarial robustness*. July 2023. URL: http://essay.utwente.nl/95917/.

[21] Christian Szegedy et al. *Intriguing properties of neural networks*. 2014. arXiv: 1312.6199 [cs.CV].

[22] Yusuke Tsuzuku and Issei Sato. "On the Structural Sensitivity of Deep Convolutional Networks to the Directions of Fourier Basis Functions". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[23] Puru Vaish, Shunxin Wang, and Nicola Strisciuglio. *Fourier-basis Functions to Bridge Augmentation Gap: Rethinking Frequency Augmentation in Image Classification*. 2024. arXiv: 2403.01944 [cs.CV].

[24] An Wang et al. "Curriculum-Based Augmented Fourier Domain Adaptation for Robust Medical Image Segmentation". In: *IEEE Transactions on Automation Science and Engineering* (2023), pp. 1–13. DOI: 10.1109/TASE.2023.3295600.

[25] Shunxin Wang et al. "DFM-X: Augmentation by Leveraging Prior Knowledge of Shortcut Learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. Oct. 2023, pp. 129–138.

[26] David Warde-Farley and Ian Goodfellow. "Adversarial Perturbations of Deep Neural Networks". In: *Perturbations, Optimization, and Statistics*. The MIT Press, Dec. 2016. ISBN: 9780262337939. DOI: 10.7551/mitpress/10761.003.0012. eprint: https://direct.mit.edu/book/chapter-pdf/2263086/9780262337939_cak.pdf. URL: https://doi.org/10.7551/mitpress/10761.003.0012.

[27] Eric Wong, Leslie Rice, and J. Zico Kolter. *Fast is better than free: Revisiting adversarial training*. 2020. arXiv: 2001.03994 [cs.LG].

[28] Qinwei Xu et al. "Fourier-based augmentation with applications to domain generalization". In: *Pattern Recognition* 139 (2023), p. 109474. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2023.109474. URL: https://www.sciencedirect.com/science/article/pii/S0031320323001747.

[29] Dong Yin et al. "A Fourier Perspective on Model Robustness in Computer Vision". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/b05b57f6add810d3b7490866d74c0053-Paper.pdf.

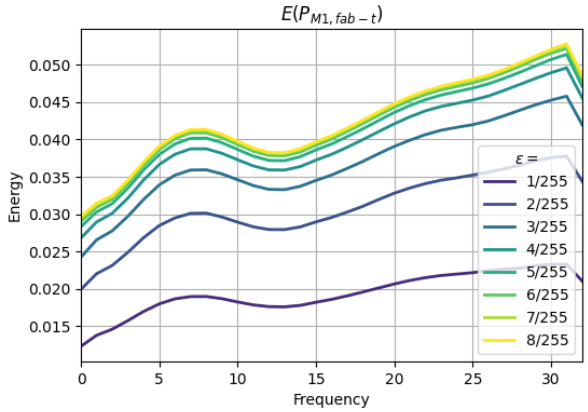

Fig. 9. The energy distribution of the mean perturbation of all successful adversarial examples under the Fourier transform for model M1 and FAB-t.

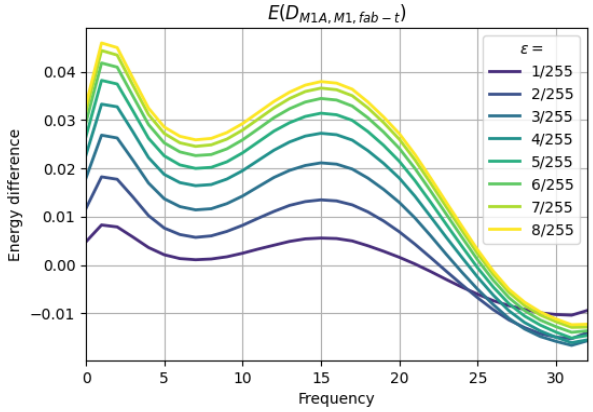

Fig. 10. The energy distribution difference between the mean perturbations on the models M1A and M1.

	APGD-ce	APGD-t	FAB-t	Square
# restarts	**1**/5	**1**/1	**1**/5	-
# iter	**10**/100	**10**/100	**10**/100	-
# target classes	-	**2**/9	**2**/9	-
# queries	-	-	-	**150**/5000

Table 2. The attack configuration of all attacks. **Bold** values indicate what was used to run the attack. Regular values indicate the standard AutoAttack configuration

A USE OF GENERATIVE AI

During the preparation of this work the author(s) used ChatGPT in order to correct grammar and spelling and to generate python code for plotting and visualization. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the work.

B NORMALIZATION ERROR

A technical error throughout the research is the cause for somewhat inaccurate data. Our models were trained on (mean, std) normalized data, and as such, expect normalized input data. All attacks function with all data min-max normalized between 0 and 1. This technical issue was identified late in the project, resulting in the need to rerun all attacks on all models. Due to a time constraint, a trade-off was made to be able to complete in time; Less iterations for each model were chosen to make the computations feasible. In table 2, the attack configurations can be seen for each attack together with its AutoAttack configuration.

Short experiments were run beforehand to indicate if the results would be accurate. 2 target classes were chosen to increase the chance of a possible class. 1 was not enough, 3 did gave relatively less pay out. For the square attack, Andriushchenko et. al. [2] showed that increasing queries has a logarithmic relation on increasing its effectiveness. In hindsight, more queries than 150 would have led to better results, . More restarts and iterations did not yield sufficient difference for the extra time it would take.

C EXPERIMENT 2: FAB FIGURES

See figures 9 and 10.