# Unifying the Harmonic Analysis of Adversarial Attacks and Robustness

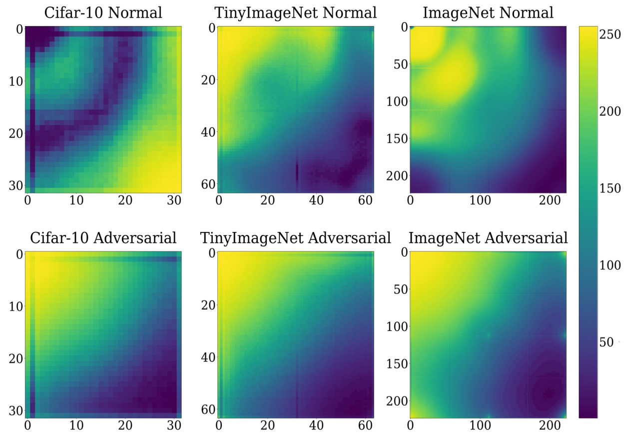
Shishira R Maiya[1], Max Ehrlich[1], Vatsal Agarwal[1], Ser-Nam Lim[2], Tom Goldstein[1], Abhinav Shrivastava[1]
[1]University of Maryland, College Park [2]Meta

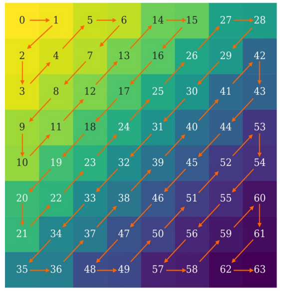
BMVC 2023

## Adversarial Examples are High frequency noise?

- Adversarial examples are imperceptible and change the output of the network when added to the input.
- The imperceptible nature makes us think they must be "High frequency noise"
- But the ineffectiveness of pre-processing methods like JPEG, deblurring and denoising as adversarial defense makes us rethink this assumption.

## Measuring impact of each frequency

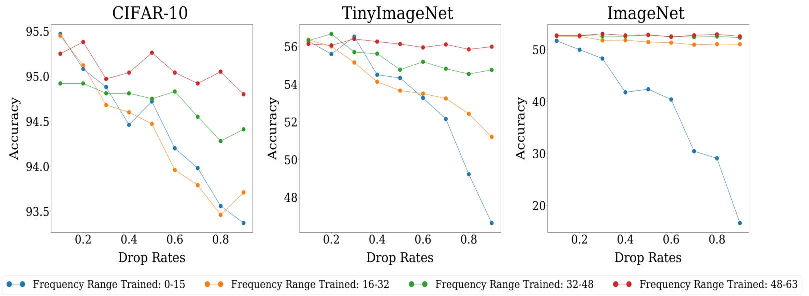


- We plot $DCT\left(\frac{dy}{d\delta}\right)$ which measures the impact each frequency has on the resulting output predictions.
- We observe that only for CIFAR-10, higher frequencies affect the output more.
- Adversarial examples are neither high nor low frequencies.

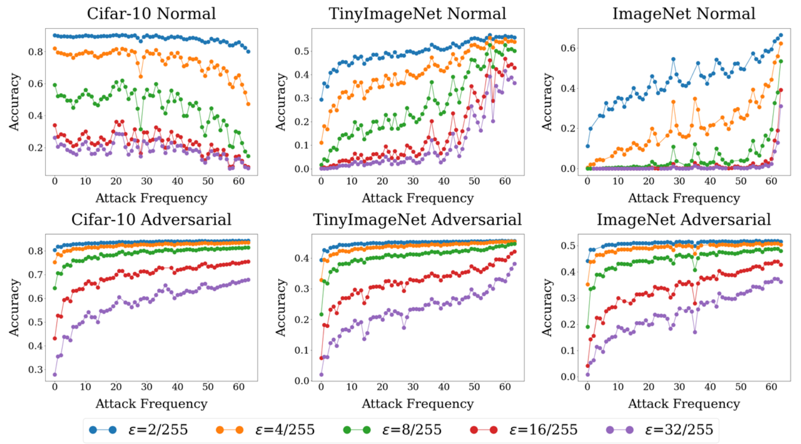
**They are dataset dependent !**

## Frequency impact during training



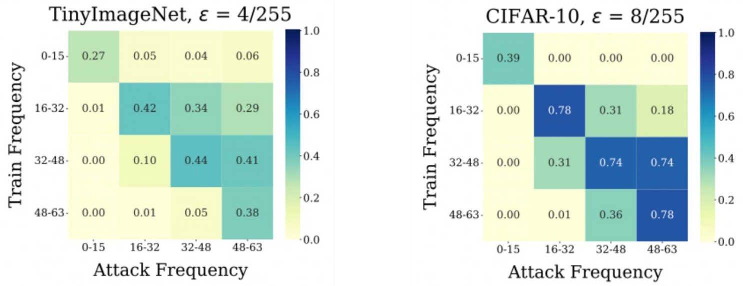

- We train models across datasets by dropping frequencies at a rate $p$ from each frequency band.
- CIFAR-10 experiences only ~2% drop when lower frequencies are dropped.
- In contrast, both ImageNet and TinyImageNet exhibit more sensitivity towards dropping of lower frequencies.

## Frequency impact on vulnerability



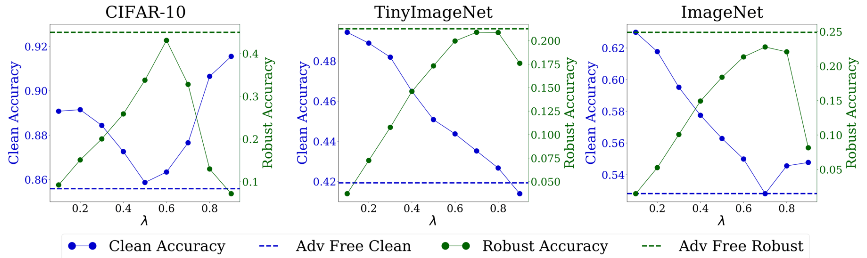

- We construct adversarial attacks by restricting them to each frequency in the DCT spectrum.
- We can observe that only for CIFAR-10 normal training, the attacks restricted on higher frequencies lead to greater reduction in accuracy.

## Adversarial Training with frequency perturbations




- Models are adversarially trained across different frequency bands and tested against other bands.
- Mid-frequency adversarial training transfers well to other bands.

## Accuracy Vs. Robustness tradeoff




$$\delta = \lambda\,[\alpha \cdot \mathrm{sgn}(\nabla_x L_{LF}) + (1-\lambda) \cdot [\,\alpha \cdot \mathrm{sgn}(\nabla_x L_{HF})]$$

- $\lambda$ controls the amount of perturbation between low and high frequencies.
- For ImageNet and TinyImageNet, clean accuracy decreases when trained with low frequencies, while increasing robustness.
- In case of CIFAR-10, we see that there is an initial increase in robustness followed by a steep fall as higher frequencies play a significant role in robustness in this dataset.