Impact of Spatial Frequency Based Constraints on Adversarial Robustness

Rémi Bernhard*, Pierre-Alain Moellic*, Martial Mermillod†, Yannick Bourrier†
Romain Cohendet‡, Miguel Solinas‡ Marina Reyboz‡

* CEA-Leti, France
† LPNC, CNRS, Université Grenoble Alpes, Université Savoie Mont Blanc, France
‡ Univ. Grenoble Alpes, CEA-List, France

IJCNN 2021

Context: Large-scale deployment of neural network models.

Attacks against integrity:
Need for a better understanding of *adversarial examples*, to develop efficient defenses

Neural computation and cognitive psychology:
$\Rightarrow$ Importance of Low Spatial Frequency components in the human classification process

NB: LSF = Low Spatial Frequency, HSF = High Spatial Frequency

Objectives:

- Are adversarial examples a pure HSF phenomenon?
- Link between **adversarial robustness** and **frequency properties** of information processed by models
- Investigate the robustness against adversarial perturbations offered by **spatial frequency-based constraints**

Data sets: SVHN (32x32), CIFAR10 (32x32) and "Small ImageNet" (224x224)

Preliminaries

Notations and filtering process

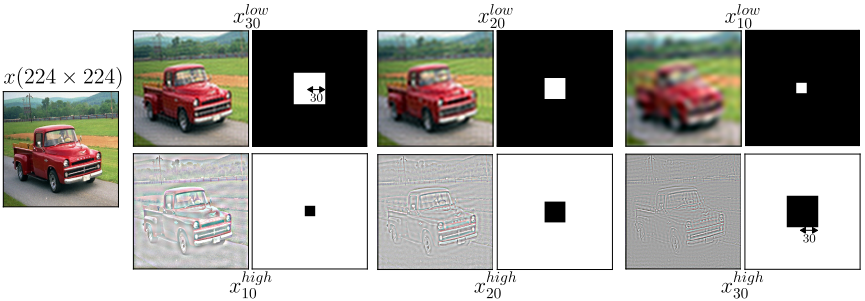
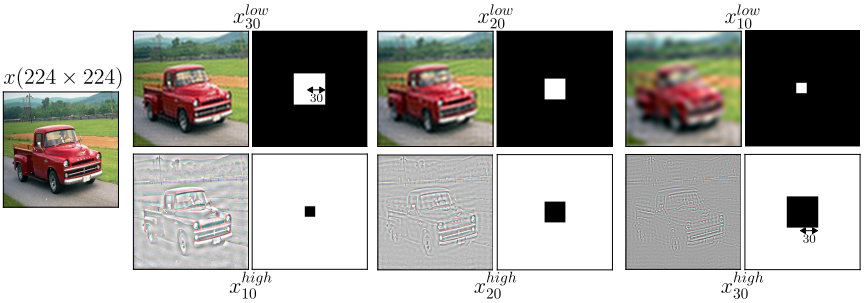


Figure: First row: low-pass filtering. Bottom row: high-pass filtering. For the Fourier domain masks, white $\rightarrow$ 1, black value $\rightarrow$ 0.

For LSF, low i value $\Rightarrow$ strong low-pass filtering.
For HSF, high i value $\Rightarrow$ strong high-pass filtering.

Notations and filtering process



We filter the data set and train models on these filtered data:

- $X_i^{low,high} \rightarrow$ filtered data with low or high-pass filtering at intensity i (LSF/HSF task)
- $M_i^{low,high} \rightarrow$ Model **trained on** $X_i^{low,high}$.

Frequency properties of data and models

Frequency properties of data and models

Accuracy on filtered data sets

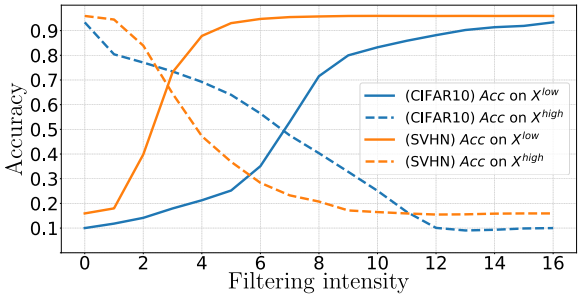


Figure: CIFAR10 and SVHN. Accuracy of a regular model on low-pass and high-pass filtered data set.

Informative features learned (regular model):

- SVHN $\rightarrow$ Focused on LSF task
- CIFAR10 $\rightarrow$ Spread between LSF & HSF tasks

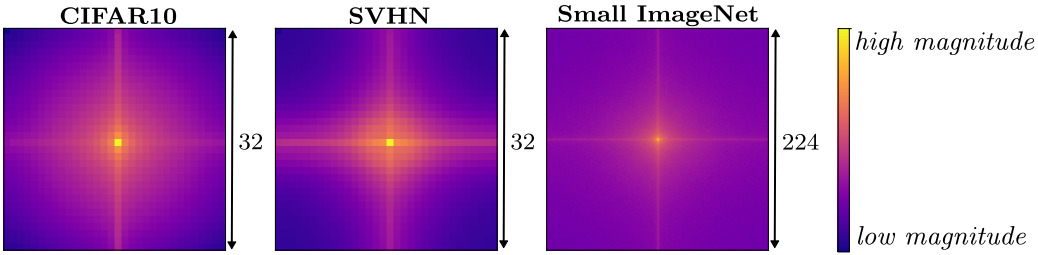


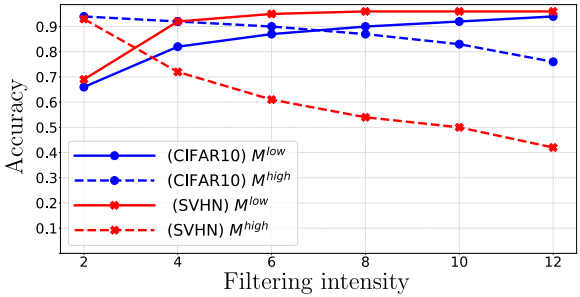
Figure: Magnitude of the Fourier spectrum for clean images. Center: LSF, border: HSF.

Relevant with frequency properties of the data sets:

- SVHN: Narrow spectrum (towards LSF)
- CIFAR10, Small ImageNet: Spread spectrum

Frequency properties of data and models

models trained on filtered data sets



CIFAR10, SVHN: Test set accuracy of models **trained on filtered data sets**.

- CIFAR10 $\rightarrow$ useful information are distributed along the spectrum
- SVHN $\rightarrow$ predominantly concentrated in the LSF.

Sensitivity to HSF noise

Error rate of the model on a set of examples perturbed with noise located only in specific spatial frequencies:

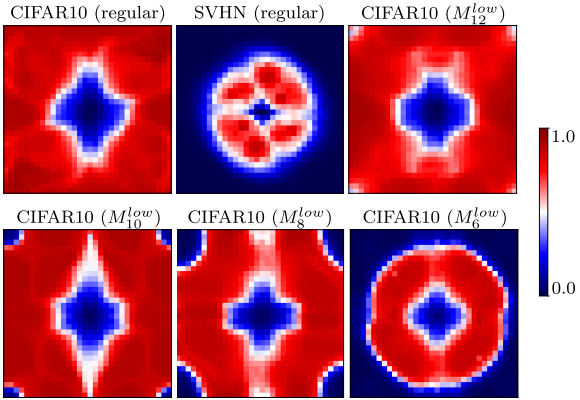


Figure: CIFAR10 and SVHN. High values $\rightarrow$ high sensitivity. Low values $\rightarrow$ low sensitivity.

Transferability analysis

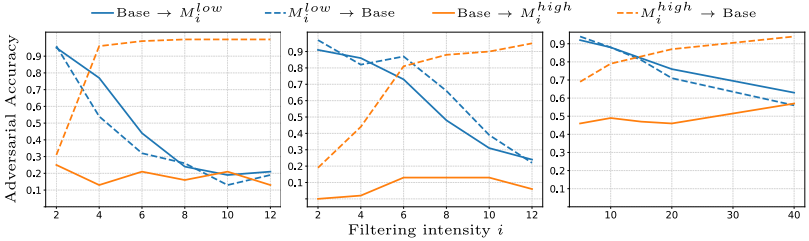


Figure: SVHN (left), CIFAR10 (middle), Small ImageNet (right). Transferability analysis.

1) Blue curves: Two way transferability $M \longleftrightarrow M_i^{low}$
$\rightarrow$ The regular classification task and the LSF task share predominantly robust useful features.

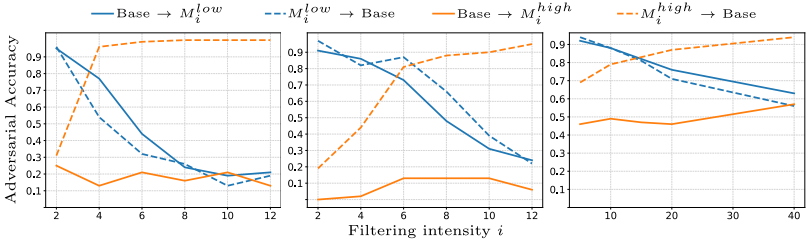


Figure: SVHN (left), CIFAR10 (middle), Small ImageNet (right). Transferability analysis.

2) Dissimilarity between the dotted and solid curves:

- impact of non-robust features exploiting HSF

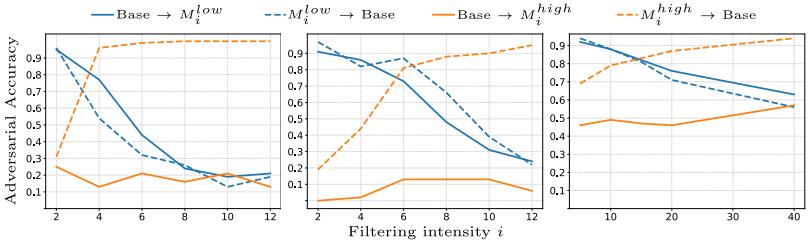


Figure: SVHN (left), CIFAR10 (middle), Small ImageNet (right). Transferability analysis.

2) Dissimilarity between the dotted and solid curves:

- as the high-pass filtering becomes more restrictive, the transferability $M_i^{high} \rightarrow M$ decreases.

$\rightarrow$ Adv perturbations need to exploit a wide part of the spectrum, and cannot be only focused on HSF.

Adversarial robustness of frequency constrained models

Goal:
Enforce the model to rely on useful features of the LSF/HSF task

Frequency-constrained loss functions:

$$L_{i,j}^{freq}(\theta, x, y) = L^E(\theta, x, y) + \lambda_1 \left\| f(x) - f(x_i^{low}) \right\|_2^2 + \lambda_2 \left\| f(x) - f(x_j^{high}) \right\|_2^2$$

$$L_i^{low}(\theta, x, y) = L^E(\theta, x, y) + \lambda_1 \left\| f(x) - f(x_i^{low}) \right\|_2^2$$

Attack:
l_∞-PGD with all sanity checks for gradient masking (false sense of security)

L^{low} **results**

SVHN: Up to 41% accuracy against PGD adversarial examples (L_6^{low})
CIFAR10:

- No observed robustness
- 11% accuracy against PGD adversarial examples when considering M^{low} models

$\rightarrow$ L^{low} brings robustness if:
i) a model relies predominantly on useful features of the LSF task (shared robust features)
ii) it shows no sensitivity to HSF noise

Adversarial robustness of frequency constrained models

L^{freq} **results**

- **CIFAR10**: Up to 12% accuracy against PGD adversarial examples ($L^{freq}_{5,3}$)
- **Small Imagenet**: Up to 36% accuracy against PGD adversarial examples ($L^{freq}_{40,20}$)

$\rightarrow$ L^{freq} can bring robustness in the case of information spread over the frequency spectrum

Adversarial training (reminder):

$$\delta = \underset{\|\delta\|_\infty \leq \epsilon}{\arg\max} \quad L^E(\theta, x+\delta, y)$$

Resulting frequency constrained loss:

$$L_{i,j}^{AT,freq}(\theta, x, y) = L_{i,j}^{freq}(\theta, x+\delta, y)$$

Results:
SVHN: + 12% accuracy on adversarial examples ($L_{10,4}^{AT,freq}$)
CIFAR10: + 5% accuracy on adversarial examples ($L_6^{AT,high}$)
(compared with Adversarial Training with same clean accuracy)

$\rightarrow$ Existing defense schemes can benefit from spatial frequency considerations

Conclusion

Conclusion

Contributions:

- Adversarial examples exploit features of the whole frequency spectrum
- Models relying predominantly on useful features for the LSF task, and with a non-sensitivity to HF noise show robustness when constrained to rely on useful information for the LSF task
- when developing defenses, it is crucial to take into account the intrinsic frequency properties of data

Perspectives:
Investigate the relation between frequency properties of Adversarial Training and frequency-based constraints