Image Mixing and Gradient Smoothing to Enhance the SAR Image Attack Transferability 2343

Yue Xu[1⋆], Xin Liu[1⋆], [1]Kun He[1], Shao Huang[2], Yaodong Zhao[2] and Jie Gu[2]
[1]School of Computer Science and Technology, Huazhong University of Science and Technology
[2]Science and Technology on Electronic Information Control Laboratory

Introduction

Adversarial Examples, which aims to add imperceptible perturbation to the original input to fool the target classifier, brings critical security threats to the deep learning based systems on neural modulation recognition.

SAR (Synthetic Aperture Radar) images consist of distinct amplitude and phase component, which possess disparate data distributions and information, resulting in distinct functions in model classification. Moreover, the presence of noise introduced by the SAR speckle effect significantly disrupts gradient calculations, making it more challenging to successfully perform attacks on SAR images using existing methods designed for visual images.

Background: existing transferable attacks (e.g. DIM, TIM, etc.) exhibit pool performance in different target models.

Methodology

The proposed Image Mixing and Gradient Smoothing (IMGS) method consists of two parts to enrich image information and smoothing the gradient.
The **Image Mixing** consists of following steps:

1. Considering the distinction between amplitude and phase, we define $\boldsymbol{x} = (\boldsymbol{A}, \phi)$ as a SAR clean image in dataset $\boldsymbol{D}$.
2. Let $\boldsymbol{x}^{\text{rd}} = (\boldsymbol{A}^{\text{rd}}, \phi^{\text{rd}})$ denote an image randomly chosen from another category.
3. Taking into consideration the noise impact, we chose a single image for mixing. The mixed image is defined as follows:

$$(\boldsymbol{A}^{\text{mixed}}, \phi^{\text{mixed}}) = (\boldsymbol{A} + \delta_1 \cdot \boldsymbol{A}^{\text{rd}}, \phi + \delta_2 \cdot \phi^{\text{rd}}), \tag{1}$$

where δ_1 and δ_2 are the mixing rate on amplitude and phase, respectively. δ_2 is typically greater than δ_1.

The **Gradient Smoothing** consists of following steps:

1. The Local Mean Square Error (LMSE) Filter strikes a balance between noise reduction and detail preservation.
2. Smoothing Expression: The smoothing operation using the LMSE filter is expressed as follows:

$$(\bar{g}_t^{smooth})_{ij} = \omega_{ij}(\bar{g}_t)_{ij} + (1 - \omega_{ij})\mu_{ij} \tag{2}$$

where $\omega_{ij} = \frac{\sigma_{ij}^2}{\sigma_{ij}^2 + \mu_{ij}^2}$.

Algorithm

Algorithm 1 The *IMGS* Attack Algorithm

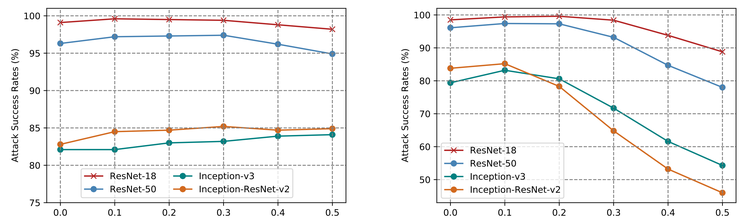
Input: SAR-ATR classifier f, loss function J, clean example $\boldsymbol{x}$ and its label y, perturbation bound ϵ, number of iterations T, decay factor μ
Output: An adversarial example $\boldsymbol{x}^{adv}$

$\alpha = \epsilon/T$; $g_0 = 0$; $t = 0$; $\boldsymbol{x_0^{adv}} = \boldsymbol{x}$
while $t < T$ **do**
  Randomly choose one image $\boldsymbol{x_1}$ from another category
  Calculate mixed image $\boldsymbol{x}^{mixed}$ by Eq. (1).
  Calculate the average gradient $\bar{\boldsymbol{g}}_{t+1}$:

$$\bar{\boldsymbol{g}}_t = \frac{1}{m}\sum_{i=0}^{m-1} \nabla_{\boldsymbol{x}} J(\gamma_i * (\boldsymbol{A}_t^{mixed}, \phi_t^{mixed}), y). \tag{3}$$

  Smoothing the gradient $\bar{\boldsymbol{g}}_{t+1}^{smooth}$ by Eq. (2)
  Update the gradient with Momentum by MI-FGSM attack.
  Update the adversarial example $\boldsymbol{x}_{t+1}^{adv}$.
  $t \leftarrow t + 1$
end while
return $\boldsymbol{x}^{adv} = \boldsymbol{x}_T^{adv}$

Experiments



(a) Amplitude mixing **(b)** Phase mixing

Figure 1: Attack success rates (%) of the different amplitude mixing rate δ_1 and phase mixing rate δ_2 on ResNet-18.

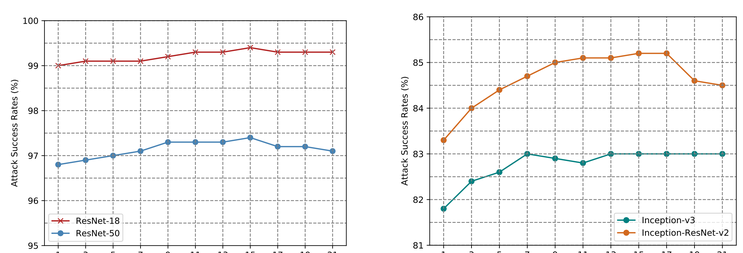
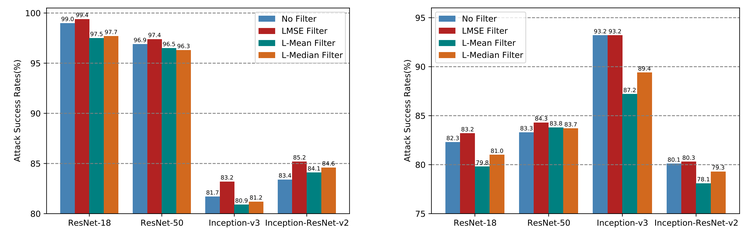


Figure 2: Attack success rates (%) of the window size of $W = [1, 21]$.

Experiments



(a) ResNet-18 **(b)** Inception-v3

Figure 3: Attack success rates (%) of different filters by crafting adversarial examples on ResNet-18 or Inception-v3 model.

Model	ResNet-18				
	DIM	TIM	SIM	Admix	*IMGS*
ResNet-18	98.70*	99.00*	98.70*	**99.40***	**99.40***
ResNet-50	93.50	94.10	96.20	96.60	**97.40**
Inception-v3	75.80	78.10	78.20	76.60	**83.20**
Inception-ResNet-v2	71.10	73.70	78.60	74.30	**85.20**
Model	Inception-v3				
	DIM	TIM	SIM	Admix	*IMGS*
ResNet-18	78.10	79.30	75.90	82.60	**83.20**
ResNet-50	76.70	77.20	78.50	82.40	**84.30**
Inception-v3	87.60*	88.40*	87.60*	92.00*	**93.20***
Inception-ResNet-v2	66.40	71.40	68.00	79.40	**80.30**
Model	ResNet-50				
	DIM	TIM	SIM	Admix	*IMGS*
ResNet-18	**99.40***	95.10	95.50	91.80	**96.70**
ResNet-50	97.40	98.00*	96.10*	98.60*	**99.10***
Inception-v3	72.50	74.10	76.60	75.70	**82.90**
Inception-ResNet-v2	70.60	71.20	75.70	74.80	**82.90**
Model	Inception-ResNet-v2				
	DIM	TIM	SIM	Admix	*IMGS*
ResNet-18	79.50	80.50	80.00	81.80	**82.50**
ResNet-50	80.60	81.30	83.80	86.60	**86.70**
Inception-v3	69.80	71.90	74.70	78.40	**79.70**
Inception-ResNet-v2	89.10*	90.30*	89.00*	93.20*	**93.50***

Table 1: Attack success rates (%) of test models with different input transformation methods. * indicates white-box attacks.

Conclusion

We propose an effective method called Image Mixing and Gradient Smoothing (IMGS), which randomly mixes parts of an image into the input and smooths the gradient via the Local Mean Square Error (LMSE) filter to enhance the adversarial transferability. Extensive experiments demonstrate that IMGS has better attack performance.