IMPROVING FEATURE-LEVEL ADVERSARIAL TRANSFERABILITY VIA DIVERSITY ATTRIBUTION

supplementary materials

1. ALGORITHM

Due to space limitations, the pseudocode description of the algorithm is provided in the supplementary materials.

Algorithm 1 Diversity Attribution Attack

Input: The classifier f, the original image x and its corresponding true label y.

Parameters: Perturbation magnitude ϵ; maximum iterations T; decay factor μ; ensemble number N; splitting number s.

$\alpha = \epsilon/T$; $g_0 = 0$; $x_0^{adv} = x$

Obtain N copies of diverse transformations

Calculate the diversity attribution:

$$\Delta_k = \frac{1}{C}\sum_{n=1}^{N}\frac{\partial l(x_m, y)}{\partial f_k(x_m)}$$

Construct optimization objective:

$$\mathcal{L}(x^{adv}) = \sum(\Delta_k \odot f_k(x^{adv}))$$

Update x^{adv} by momentum iterative method:

for $t = 0$ to $T - 1$ **do**

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x \mathcal{L}\left(x_t^{adv}\right)}{\left\|\nabla_x \mathcal{L}\left(x_t^{adv}\right)\right\|_1}$$

$$x_{t+1}^{adv} = \mathrm{Clip}_{x,\epsilon}\left\{x_t^{adv} - \alpha \cdot \mathrm{sign}\left(g_{t+1}\right)\right\}$$

end for

return $x^{adv} = x_T^{adv}$

2. EXPERIMENTAL RESULTS

We provide additional results on the transferability of adversarial examples (AEs) generated using more surrogate models. As shown in Table 1 and Table 2, DAA achieves higher attack success rates than existing leading methods when attacking both normally trained and robust models. DAA demonstrates strong performance in both white-box and black-box attacks, further highlighting the effectiveness of diversity attribution in capturing critical features.

3. FURTHER ANALYSIS

Is the improvement in transferability due to a single, global transformation method rather than diverse transformations (DT)? To validate this, we analyze the impact of diversity attribution and the aggregated gradient obtained by globally applying each specific transformation method to the input image on transferability. As shown in Fig. 1, when the four specific transformation methods are applied solely and globally to the input image, their transferability is lower than that of the diversity attribution method due to the limited transformed diversity. This further suggests that diversity transformations are more effective at neutralizing model-specific features, and enhancing transferability.

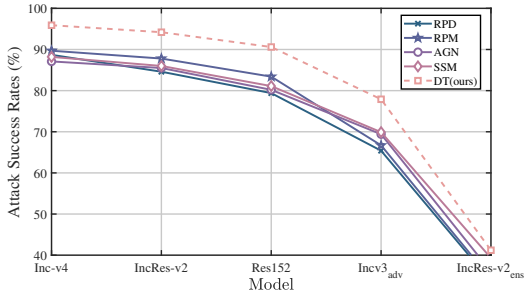

Fig. 1. The impact of single and diversity transformation on attack success rate.

Is diversity attribution more effective than the neighborhood attribution proposed by NEAA and the aggregated gradients produced by RPA in suppressing model-specific features? If feature importance evaluation results obtained from the surrogate model can more effectively suppress mode-specific elements, imply that these results can exhibit a higher correlation with the feature importance assessment results from other target models. To validate this guess, we measure the similarity of diversity attribution, neighborhood attribution, and aggregated gradients across models using cross-model cosine similarity, expressed as:

$$V(A, B) = \frac{A \cdot B}{||A||_2 \cdot ||B||_2}, \tag{1}$$

where A and B represent the result obtained from different models using the same importance evaluation method. We calculate these results from the output layer of the network, as they share the same dimensionality. As shown in Fig. 2, diversity attribution demonstrates a higher correlation across dif-

Table 1. The attack success rates (%) of different attacks against normally trained models.

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-50	Res-152	Vgg19	AVG.
IncRes-v2	FIA	85.2	82.7	94.4	83.1	76.8	83.1	84.2
	NAA	70.3	66.6	85.3	69.4	59.1	71.2	70.3
	RPA	86.4	83.6	93.6	84.7	80.9	84.8	85.7
	NEAA	83.2	80.1	93.8	80.6	73.8	82.3	82.3
	DAA	**92.9**	**90.8**	**97.9***	**91.0**	**87.0**	**91.4**	**91.8**
	PIDIM + RPA	89.1	86.1	95.2*	87.0	82.7	87.7	88.0
	PIDIM + NEAA	89.0	85.7	96.3*	86.2	78.7	88.9	87.5
	PIDIM + DAA	**94.5**	**92.1**	**98.0***	**91.1**	**87.8**	**92.0**	**92.6**

Table 2. The attack success rates (%) of different attacks against roubst models.

Model	Attack	ViT-B	PiT-B	ViS-S	Inc-v3$_{adv}$	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	AVG.
IncRes-v2	FIA	28.3	41.7	53.9	63.2	55.1	50.2	45.1	48.2
	NAA	20.2	32.1	39.0	45.9	41.0	37.8	31.4	35.3
	RPA	31.3	48.1	60.8	69.0	61.3	56.2	49.8	53.8
	NEAA	27.2	42.0	49.5	65.6	55.2	51.2	43.5	47.7
	DAA	**37.0**	**52.9**	**63.8**	**70.5**	**66.7**	**62.2**	**56.1**	**58.5**
	PIDITIM + RPA	43.8	57.6	70.1	77.0	72.4	68.8	64.2	64.8
	PIDITIM + NEAA	42.3	54.6	65.4	77.2	71.2	68.8	62.7	63.2
	PIDITIM + DAA	**48.3**	**61.6**	**71.1**	**78.3**	**74.6**	**71.3**	**68.1**	**67.6**

ferent models than neighborhood attribution and aggregated gradients. This further indicates that diversity attribution effectively suppresses model-specific information, highlighting the critical key features.

4. ABLATION STUDIES

4.1. The effectiveness of each transformation method

we further explore the effectiveness of each transformation method. The symbol "+" indicates the introduction of a new transformation method to the image blocks. The initial RPD represents the introduction of only the RPD transformation method to the image blocks. As new transformation methods are gradually added, the diversity of transformations increases, leading to a significant improvement in transferability.

4.2. The parameter settings for the introduced transformation methods

For the parameter setting of the introduced transformation methods, we analyze the optimal parameters for each specific transformation method within the global transformation.

The parameter setting of SSM. We analyze the impact of two parameters of the SSM on transferability: the standard deviation σ of Gaussian noise and the transformation factor ρ. As shown in Fig. 4, the attack success rates for the normally trained models peak at $\sigma = 48$. For robust defense models, the peak occurs at $\sigma = 32$ and then begins to decline. Therefore, σ is set to 48 and 32 when attacking normally trained and robust models, respectively. Moreover, For the transformation factor ρ in SSM, as shown in Fig. 5, the attack success rates are higher for normally trained models when $\rho = 0.7$, while $\rho = 0.3$ achieves the best results for robust models. Thus, ρ is set to 0.7 for attacking normally trained models and 0.3 for robust models.

The parameter setting of AGN. We explore the impact of the parameter of the AGN on transferability: the standard deviation σ of Gaussian noise. As shown in Fig. 6, in summary, the attack success rates for normally trained models and robust models reach their maximum at $\sigma = 48$ and $\sigma = 32$. Therefore, σ is set to 48 and 32 when attacking normally trained models and robust models, respectively.

The parameter setting of RPM. Following the settings in [1], the masking probability p_2 is set to 0.3 and 0.2 for attacking normally trained models and robust models, respectively. The patch size is randomly selected from the list [1,3,5,7].

The parameter setting of RPD. Following the settings in [2], the drop probability p_1 is set to 0.3 and 0.1 for attacking normally trained models and robust models, respectively.

5. REFERENCES

[1] Y. Zhang, Y. Tan, T. Chen, X. Liu, Q. Zhang, and Y. Li, "Enhancing the transferability of adversarial examples with random patch.," in *IJCAI*, 2022, pp. 1672–1678.

[2] Z. Wang, H. Guo, Z. Zhang, W. Liu, Z. Qin, and K. Ren, "Feature importance-aware transferable adversarial attacks," in *ICCV*, 2021, pp. 7639–7648.

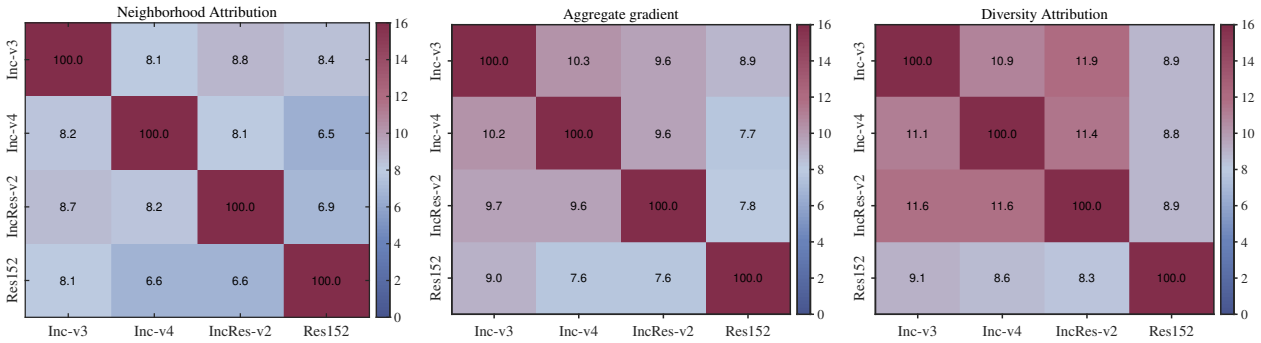

Fig. 2. Cosine similarity across models for different attribution methods. Due to the small values, the results are magnified by a factor of 100.

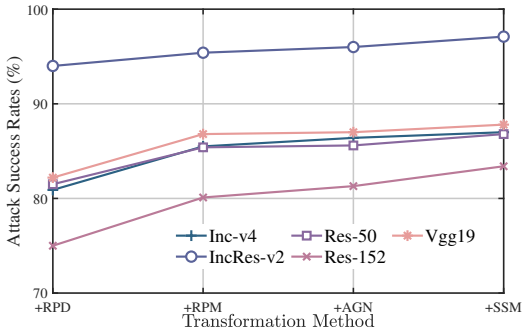

Fig. 3. The effectiveness of each transformation method.

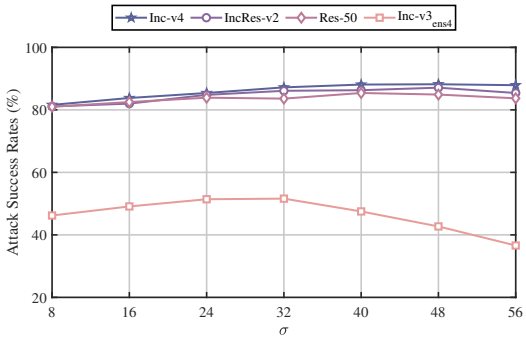

Fig. 4. The hyperparameter σ in the SSM transformation.

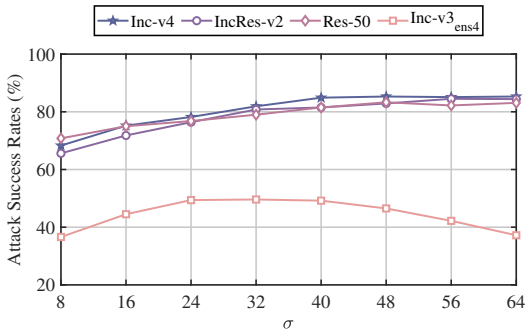

Fig. 6. The hyperparameter σ in the AGN transformation.

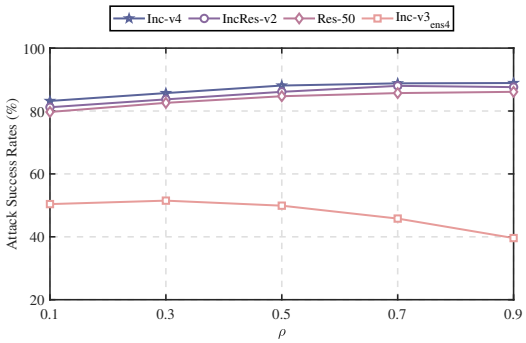

Fig. 5. The hyperparameter ρ in the SSM transformation.