To Tackle Adversarial Transferability: A Novel Ensemble Training Method with Fourier Transformation

Authors: Wanlin Zhang, Weichen Lin, Ruomin Huang, Shihong Song, Hu Ding

Affiliations: School of Computer Science and Technology, University of Science and Technology of China

Conference: ICLR 2025

2025.03.18

Outline

- Research Background & Motivation
- Overview of the Proposed Method
  - Definition of feature extractor
  - Fine-grained analysis on ensemble model vulnerability
  - Frequency domain transformation
- Experimental Results
- Conclusion & Future Work

Research Background & Motivation

- Adversarial examples severely impedes the application of DNNs in security-conscious scenarios, such as self-driving car[1], Face recognition[2] and heath care[3].

[1] Shetuwang. (2025, March 17). [self-driving] [Image]. Qianzhan. https://t.qianzhan.com/caijing/detail/250317-9ba77063.html
[2] Huawei. (n.d.). [heath care] [Image], from https://e.huawei.com/cn/industries/healthcare
[3] vMaker Editorial Team. (2020, September 27). [Face recognition] [Image], from https://vmaker.tw/archives/47170

Research Background & Motivation

- Adversarial training approaches often suffer from high training time and a decline in accuracy on clean data.

- Ensemble training methods struggle to prevent the transferability of adversarial examples among sub-models.

- Data transformation-based approaches often rely on a private key. If the private key is lost, the model becomes vulnerable

Overview of the Proposed Method

- Definition of feature extractor
- Fine-grained analysis on ensemble model vulnerability
- Frequency domain transformation

Definition of Feature Extractor

- Useful feature extractor & Robust and non-robust feature extractor.

Definition 1 (Useful feature extractor)

For a given data distribution $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$, a feature extractor θ: $\mathcal{X} \rightarrow \mathbb{R}^k$ is **useful**, if we have

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathrm{h}(y)^T\theta(x)] > 1/k,$$

Where $\mathrm{h}(y)$ is the one−hot k-dimensional vector of the label y, k is the number of classes.

Definition 2 (Robust and non-robust feature extractor)

(1) We say θ is **robust** if the following condition holds for any i $(1 \leq i \leq k)$: $\mathbb{E}_{(x,y)\sim\mathcal{D}_i}[\theta(\mathcal{A}(x))]_i > 1/k$. where $\mathcal{D}_i$ represents the i-th class data and $\mathcal{A}(x)$ denotes the adversarial example of a data item x. We denote the set of these robust feature extractors as Θ_R.

(2) The remaining useful feature extractors are **non-robust**. We assign these non-robust extractors to $k(k-1)$ sets: $\{\Theta_{i,j} | 1 \leq i \neq j \leq k\}$ as follows: For each non-robust θ, there must exist at least an index "i" such that $\mathbb{E}_{(x,y)\sim\mathcal{D}_i}[\theta(\mathcal{A}(x))]_i \leq 1/k$. We let $j = argmax_s\ \mathbb{E}_{(x,y)\sim\mathcal{D}_i}[\theta(\mathcal{A}(x))]_s$ and assign θ to $\Theta_{i,j}$.

we can represent the model as:

$$f(x) = \sum_{\theta\in \Theta_R\cap\Theta_f} w_\theta\theta(x) + \sum_{i,j=1,\, i\neq j} \sum_{\theta\in \Theta_{i,j}\cap\Theta_f} w_\theta\theta(x) \qquad (1)$$

Fine-grained Analysis on Ensemble Model Vulnerability

- We define the Vulnerability of ensemble model and the vulnerability towards y_t.

$$\mathrm{Vr}(F_{\mathrm{E}}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{I}\{F_{\mathrm{E}}(x) = y \wedge F_{\mathrm{E}}(\mathcal{A}(x)) \neq y\}\right]$$

$$\mathrm{Vr}(F_{\mathrm{E}}, y_t) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{I}\{F_{\mathrm{E}}(x) = y \wedge F_{\mathrm{E}}(\mathcal{A}(x)) = y_t\}\right]$$

- We have the inequalities.

$$\mathrm{Vr}(F_{\mathrm{E}}) \leq \sum_{y_t \in \mathcal{Y}} \mathrm{Vr}(F_{\mathrm{E}}, y_t) \tag{2}$$

$$\mathrm{Vr}(F_{\mathrm{E}}, y_t) \leq \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{I}\left(\sum_{m=1}^{M} \mathbb{I}([f_m(\mathcal{A}(x))]_y < [f_m(\mathcal{A}(x))]_{y_t}) > \frac{M}{k}\right)\right] \tag{3}$$

- To reduce the upper bound of $V_r(F_E)$, we decrease the right hand side of the inequality (2) which is equivalent to reducing the probability of the following expression holding.

$$\Big[\sum_{\theta\in\Theta_R^m} w_\theta \theta(\mathcal{A}(x)) + \sum_{i,j=1,i\neq j}^{k} \sum_{\theta\in\Theta_{i,j}^m} w_\theta \theta(\mathcal{A}(x))\Big]_y < \Big[\sum_{\theta\in\Theta_R^m} w_\theta \theta(\mathcal{A}(x)) + \sum_{i,j=1,i\neq j}^{k} \sum_{\theta\in\Theta_{i,j}^m} w_\theta \theta(\mathcal{A}(x))\Big]_{y_t} \tag{4}$$

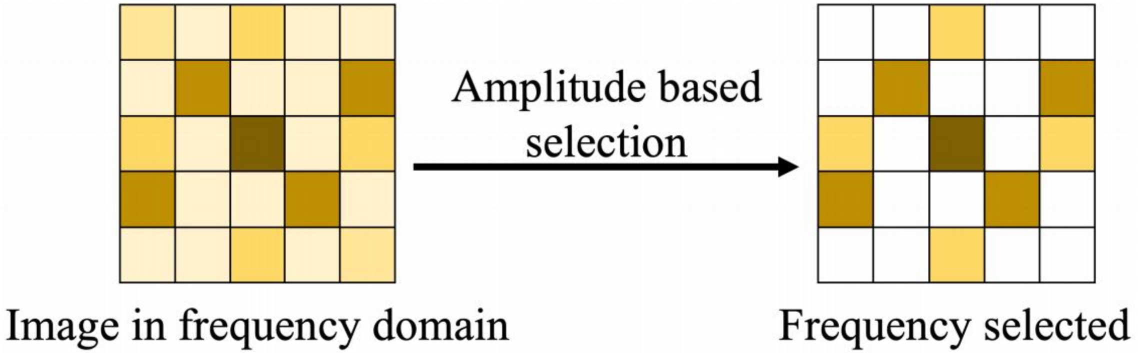
Fine-grained Analysis on Ensemble Model Vulnerability

- ***Hint(i):*** To decrease the vulnerability in the attack direction y_t, it is reasonable to decrease the influence from the non-robust feature extractors of Θ^m_{y,y_t}.

- ***Hint (ii):*** For each attack direction y_t, we only need to consider manipulating the training data of M/2 + 1 sub-models instead of all the M sub-models.

Frequency Domain Transformation

- Frequency selection.



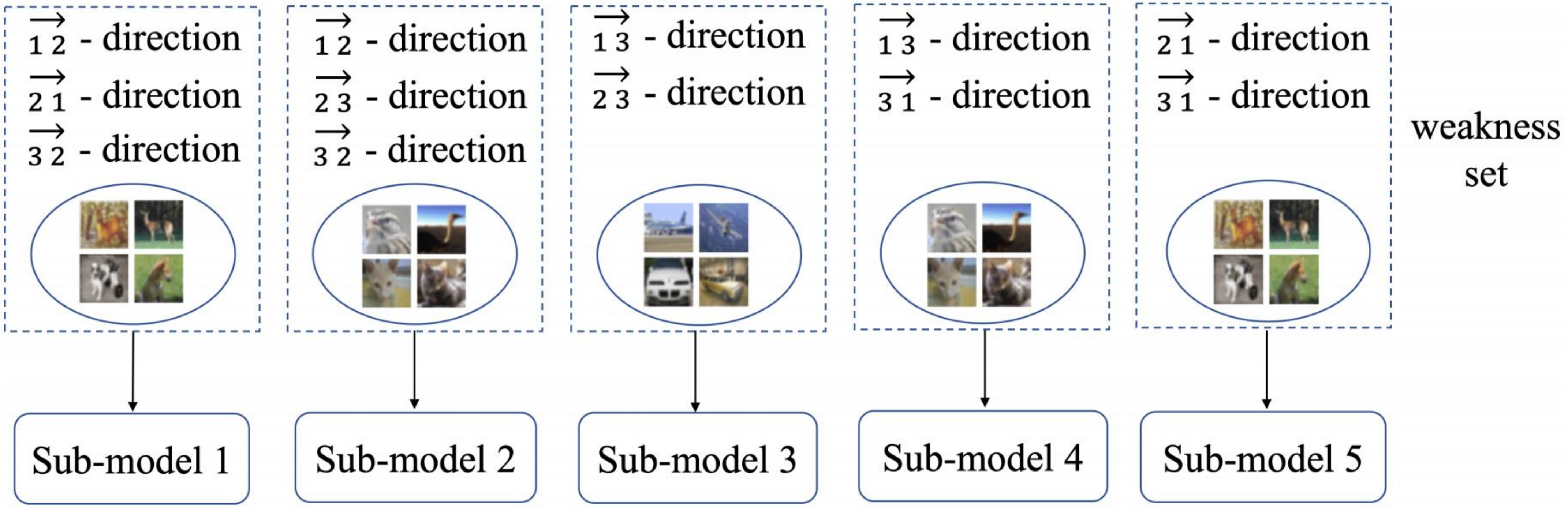
  - In our Frequency selection, we retain the high-amplitude frequencies (i.e., the darker regions) and perform data transformations on the low-amplitude frequencies (i.e., the white regions)

- Frequency transformation

  - We use adversarial attacks to manipulate non-robust features.

Frequency Domain Transformation

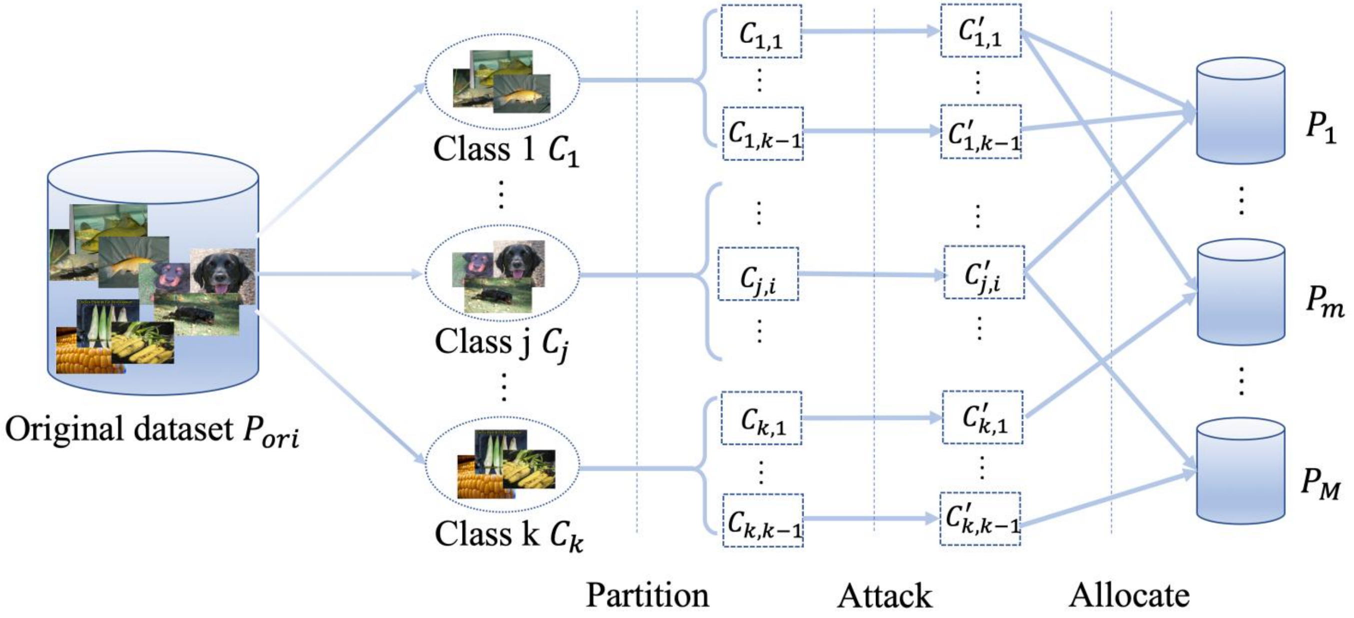
- Allocating the weakness sets to the sub-models



- Assign the attack directions to five sub-models for a three-class classification task.

Frequency Domain Transformation

- Constructing the new datasets.



- Constructing the new datasets according to the allocation scheme and train the sub-models on the new datasets.

Experimental Results

- Datasets: Cifar10 , Cifar100, SVHN , Tiny-ImageNet-200
- Architectures: ResNet20, ResNet50, WRN28-10, WRN34-10

- Comparisons: We compare our method with ADP, GAL, DVERGE, TRS.

- Results:
  - Better robust accuracy and clean accuracy.
  - Better trade-off curve, especially in terms of clean accuracy
  - Generalize to larger architectures (e.g., ResNet50, WRN34-10).

Experimental Results

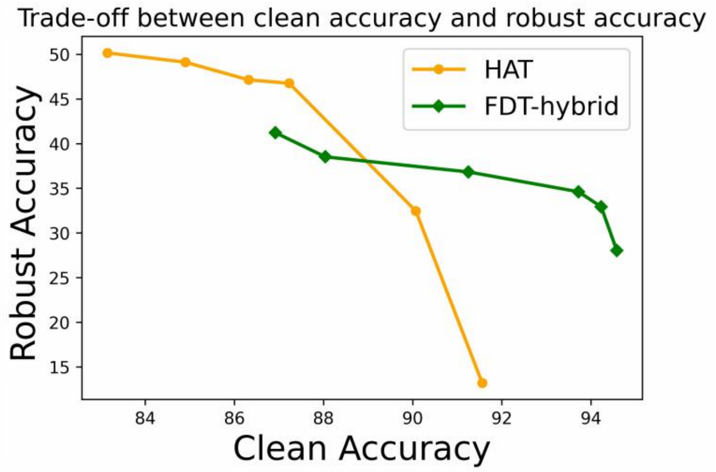
Table 2: Robust and Clean Accuracy (%) and average training time of different ensemble methods against white-box attacks on CIFAR-10 and CIFAR-100. "ϵ" and "λ" stand for the l_∞ norm of the adversarial perturbation and the coefficient of C&W attack respectively. The TRS results are reported in the original paper Yang et al. (2021), with "-" indicating results not provided.

CIFAR-10	ADP	GAL	DVERGE	TRS	FDT-random	FDT-target	FDT-hybrid
Clean accuracy	**91.84**	91.81	91.37	-	89.88$\pm$ 0.02	90.16 $\pm$ 0.04	90.20$\pm$ 0.03
FGSM (ϵ=0.01)	59.48	44.97	70.05	-	66.96$\pm$ 0.12	**72.88**$\pm$ 0.12	72.24$\pm$ 0.12
FGSM (ϵ=0.02)	53.38	30.58	56.33	44.2	46.28 $\pm$ 0.10	55.54 $\pm$ 0.09	**58.04** $\pm$ 0.13
PGD (ϵ= 0.01)	14.45	1.35	40.55	**50.5**	45.42 $\pm$ 0.09	46.58 $\pm$ 0.07	48.48 $\pm$ 0.09
PGD (ϵ = 0.02)	2.95	0.34	11.49	15.1	12.24 $\pm$ 0.03	15.08 $\pm$ 0.05	**20.01** $\pm$ 0.04
BIM (ϵ= 0.01)	14.15	1.37	40.51	**50.6**	45.24 $\pm$ 0.03	46.86 $\pm$ 0.04	48.57$\pm$ 0.05
BIM (ϵ = 0.02)	3.01	0.27	10.65	15.8	11.68$\pm$ 0.03	14.86$\pm$ 0.03	**16.63** $\pm$ 0.02
MIM (ϵ= 0.01)	20.38	2.05	44.74	51.5	47.73$\pm$ 0.05	49.97 $\pm$ 0.06	**51.50** $\pm$ 0.07
MIM (ϵ= 0.02)	5.11	0.69	14.76	17.2	15.14 $\pm$ 0.04	18.27$\pm$ 0.02	**20.09** $\pm$ 0.03
AA (ϵ= 0.01)	1.80	0.00	43.34	-	46.09$\pm$ 0.09	48.83$\pm$ 0.08	**51.56** $\pm$ 0.08
AA (ϵ= 0.02)	0.00	0.00	13.72	-	9.38 $\pm$ 0.05	15.70 $\pm$ 0.05	**19.42** $\pm$ 0.04
C&W (λ = 0.1)	20.96	31.57	52.35	**58.1**	45.01$\pm$ 0.10	55.48 $\pm$ 0.10	56.08 $\pm$ 0.11

Experimental Results

Table 7: Robust Accuracy (%) of different model architectures against white-box attacks on Cifar10. The ϵ and λ stand for the l_∞ norm of the adversarial perturbation and the coefficient of C&W attack respectively.

CIFAR10	ResNet20	ResNet50	WRN28-10	WRN34-10
clean accuracy	90.02	93.23	94.18	94.63
FGSM ($\epsilon = 0.01$)	72.24	76.65	80.64	81.04
FGSM ($\epsilon = 0.02$)	58.04	58.59	60.09	60.92
PGD ($\epsilon = 0.01$)	48.48	60.23	64.64	65.38
PGD ($\epsilon = 0.02$)	20.01	24.35	26.00	27.42
BIM ($\epsilon = 0.01$)	48.57	60.43	67.36	68.29
BIM ($\epsilon = 0.02$)	16.63	23.57	32.36	33.86
MIM ($\epsilon = 0.01$)	51.48	60.81	64.36	64.71
MIN ($\epsilon = 0.02$)	20.09	24.54	25.64	26.42
AA ($\epsilon = 0.01$)	51.56	60.48	63.45	64.01
AA ($\epsilon = 0.02$)	19.42	24.21	25.23	26.39
CW ($\lambda = 0.01$)	56.08	56.55	57.23	57.52



Conclusion and Future Work

- Conclusion:
  - We present a novel data transformation approach to improve the robustness of ensemble models against adversarial attacks
  - We demonstrate the effectiveness of our method in enhancing adversarial robustness while maintaining high accuracy on clean data

- Future Work:
  - Other types of transformation methods to improve the ensemble robustness
  - Consider more complicated scenarios for ensemble training