中国人民大学
RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院
Gaoling School of Artificial Intelligence

Understanding Model Ensemble in Transferable Adversarial Attack

ICML 2025

Wei Yao*, Zeliang Zhang*, Huayi Tang, Yong Liu#

Renmin University of China

June 25th, 2025

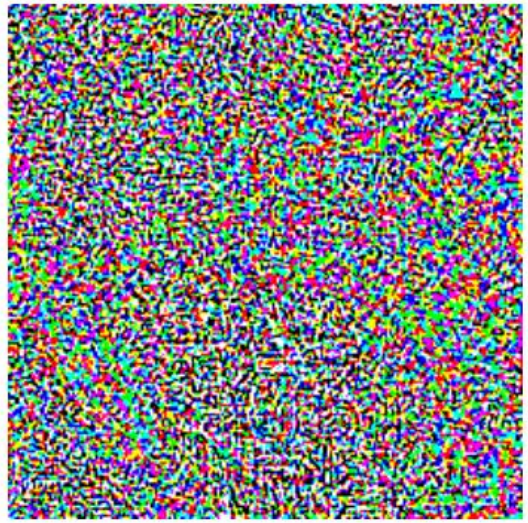
Content

- Background
- Key Definitions
- Theoretical Results
- Experiments

Content

Adversarial Example

$+ .007 \times$



$=$

$\boldsymbol{x}$

"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

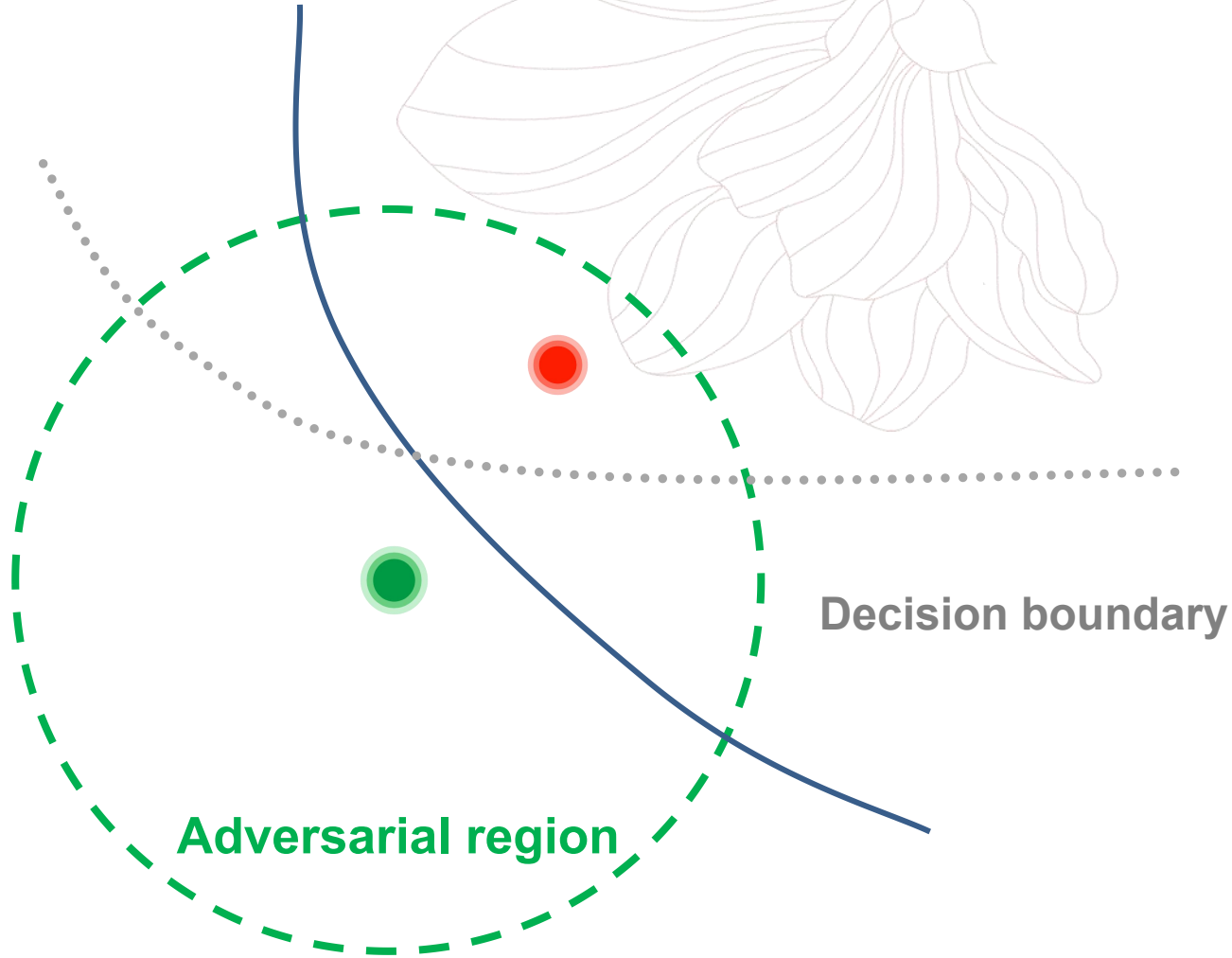
$\boldsymbol{x} + \epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

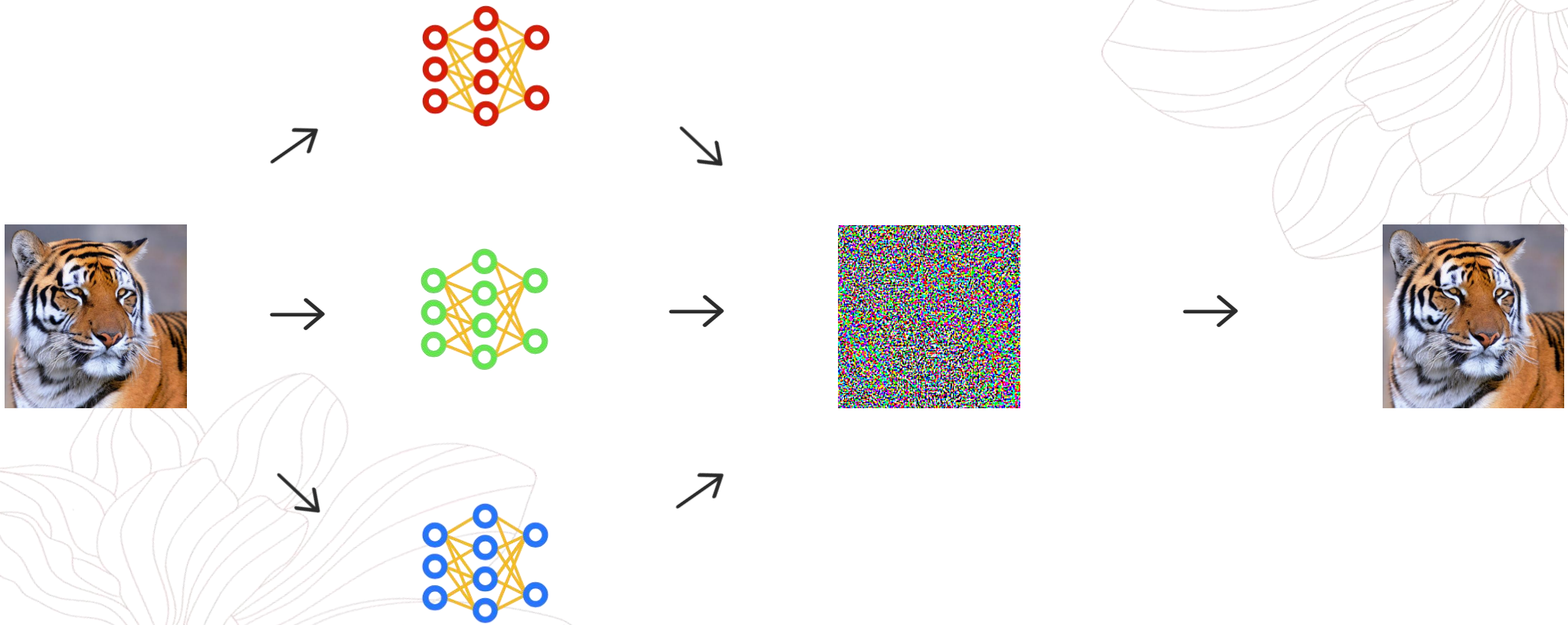
Credit to: Explaining and Harnessing Adversarial Examples. IJ Goodfellow et al. ICLR 2015.

Transferable Adversarial Attack

- Clean data
- Transferable adversarial example
- Surrogate model
- Black-box model



Transferable Adversarial Model Ensemble Attack



Data Surrogate models Perturbation Adversarial Example

Generalization bound (Rademacher complexity):

$$\underbrace{\mathrm{err}(h)}_{\text{Generalization error}} \leq \underbrace{\widehat{\mathrm{err}}(h)}_{\text{Empirical error}} + \underbrace{R_m(\mathcal{H})}_{\text{Model Complexity}} + \underbrace{\sqrt{\frac{\ln(1/\delta)}{m}}}_{\text{Sample Complexity}}$$

Motivation

Statistical Learning Theory	Transferable Adversarial Model Ensemble Attack
More data	More surrogate models
Independent data	Diverse surrogate models
Less complexity	Less complexity

Content

- Background
- Key Definitions
- Theoretical Results
- Experiments

Notations

- Data: $x \in \mathbb{R}^d$
- Label: $y \in \mathbb{R}$
- Adversarial example: $z = (x, y)$
- Model parameter: $\theta \in \Theta$ and $\theta \sim \mathcal{P}_\Theta$
- Model ensemble: $(\theta_1, \ldots, \theta_N) \sim \mathcal{P}_\Theta{}^N$
- Model output: $\hat{y} = f(\theta_i; \cdot)$
- Loss function: $\ell(\hat{y}, y)$

Model Ensemble Adversarial Attack

- Population risk: $L_P(z) = \mathbb{E}_{\theta \sim \mathcal{P}_\Theta}[\ell(f(\theta; x), y)]$
- Empirical risk: $L_E(z) = \frac{1}{N} \sum_{i=1}^{N} \ell\left(f\left(\theta_i; x\right), y\right)$
- The most transferable adversarial example: $x^* = \arg \max_{x \in \mathcal{B}_\epsilon(\hat{x})} L_P(z)$
- Adversarial example: $x = \arg \max_{x \in \mathcal{B}_\epsilon(x)} L_E(z)$

Transferability Error

Definition 3.1 (Transferability Error). The transferability error of z with radius ϵ is defined as:

$$TE(z, \epsilon) = L_P(z^*) - L_P(z). \tag{5}$$

- Always non-negative
- TE↓, adversarial transferability↑

Lemma 3.2. *The transferability error defined by Eq. (5) is bounded by the largest absolute difference between $L_P(z)$ and $L_E(z)$, i.e.,*

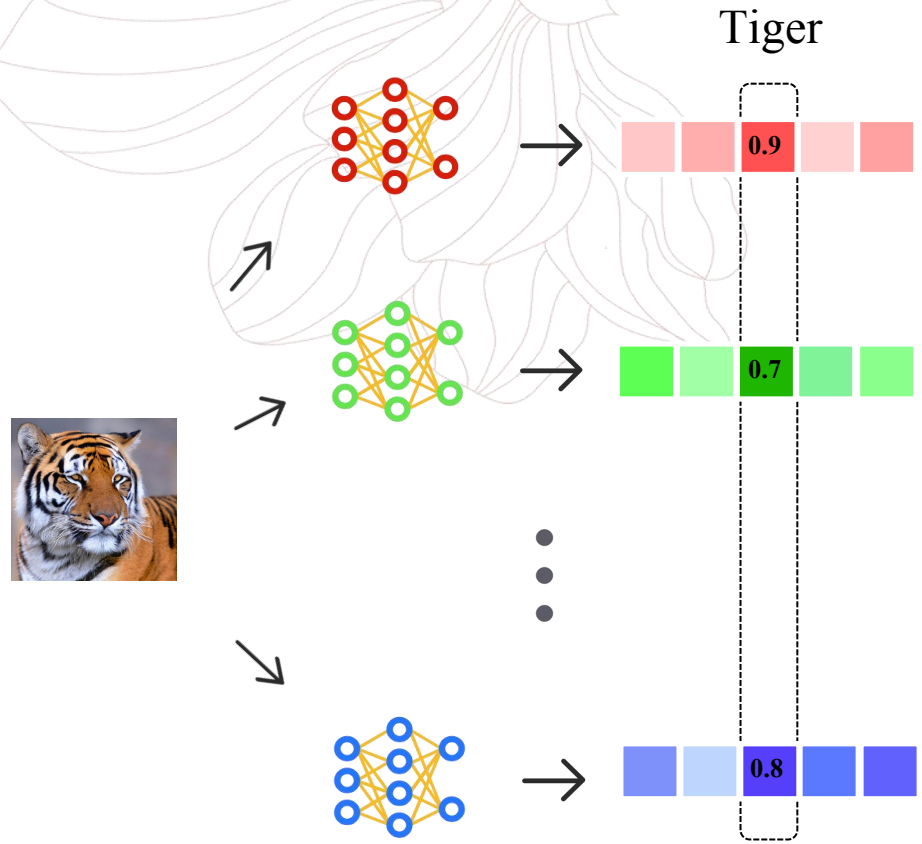
$$TE(z, \epsilon) \leq 2 \sup_{z \in \mathcal{Z}} |L_P(z) - L_E(z)|. \tag{6}$$

Let Population risk ≈ Empirical risk

Diversity of Model Ensemble Attack

Definition 3.3 (Diversity of Model Ensemble Attack)**.** The diversity of model ensemble attack across $\theta \sim \mathcal{P}_\Theta$ for a specific adversarial example $z = (x, y)$ is defined as the variance of model prediction:

$$Var_{\theta \sim \mathcal{P}_\Theta}\left(f(\theta; x)\right) = \mathbb{E}_{\theta \sim \mathcal{P}_\Theta}\left[f(\theta; x) - \mathbb{E}_{\theta \sim \mathcal{P}_\Theta} f(\theta; x)\right]^2 .$$



Idea: ensemble learning theory

- Diversity↑, Overfitting↓
- Suitable for multi-class classification

Empirical Model Ensemble Rademacher Complexity

Definition 3.4 (Empirical Model Ensemble Rademacher Complexity). Given the input space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and N classifiers $f(\theta_1; \cdot), \cdots, f(\theta_N; \cdot)$. Let $\boldsymbol{\sigma} = \{\sigma_i\}_{i \in [N]}$ be a collection of independent Rademacher variables, which are random variables taking values uniformly in $\{+1, -1\}$. We define the empirical model ensemble Rademacher complexity $\mathcal{R}_N(\mathcal{Z})$ as follows:

$$\mathcal{R}_N(\mathcal{Z}) = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{z \in \mathcal{Z}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i \ell \left(f\left(\theta_i; x\right), y \right) \right]$$

Idea: empirical Rademacher complexity

Intuition: complexity of input space relative to the surrogate models.

- Simple input space: $\mathcal{R}_N(\mathcal{Z}) = 0$
- Complex input space: $\mathcal{R}_N(\mathcal{Z}) \uparrow$

Content

- Background
- Key Definitions
- Theoretical Results
- Experiments

Vulnerability-diversity Decomposition

Theorem 4.1 (Vulnerability-diversity Decomposition). *For a data point $z = (x, y)$, we consider the squared error loss $l(f(\theta; x), y) = [f(\theta; x) - y]^2$. Let $\tilde{f}(\theta; x) = \mathbb{E}_{\theta \sim \mathcal{P}_\Theta} f(\theta; x)$ be the expectation of prediction over the distribution on the parameter space. Then there holds*

$$TE(z, \epsilon) = L_P(z^*) - \underbrace{l(\tilde{f}(\theta; x), y)}_{\textit{Vulnerability}} - \underbrace{Var_{\theta \sim \mathcal{P}_\Theta} f(\theta; x)}_{\textit{Diversity}}. \tag{9}$$

Remark. A similar formulation also applies to the KL divergence loss in the multi-class classification setting, which is proved in Appendix C.3.

Idea: bias-variance decomposition

Intuition:

- Strong & diverse surrogate models
- Vulnerability-diversity trade-off

Ensemble Complexity of MLP

Lemma 4.2 (Ensemble Complexity of MLP). *Let* $\mathcal{H} = \{x \mapsto W_l \phi_{l-1}(W_{l-1}\phi_{l-2}(\ldots \phi_1(W_1 x)))\}$ *be the class of real-valued networks of depth* l, *where* $x \in \mathbb{R}^{d_1}$, $W_i \in \mathbb{R}^{d_{i+1} \times d_i}$. *Given* N *classifiers from* $\mathcal{H}$, *where the parameter matrix is* $W_{ij}, i \in \{1, \cdots, n\}, j \in \{1, \cdots, l\}$ *and* $T = \prod_{j=1}^{l} \sup_{i \in [n]} \|W_{i,j}\|_F$. *Let* $\|x\|_F \leq B$. *With 1-Lipschitz activation functions* $\phi_1, \cdots, \phi_{l-1}$ *and* 1*-Lipschitz loss function* $\ell(yf(x))$, *there holds:*

$$\mathcal{R}_N(\mathcal{Z}) \leq \frac{\left(\sqrt{(2\log 2)l} + 1\right) BT}{\sqrt{N}}. \tag{10}$$

Remark. We also derive the upper bound of $\mathcal{R}_N(\mathcal{Z})$ for the cases of linear model (Appendix B.2) and two-layer neural network (Appendix B.3). These results are special cases of the above theorem.

Idea: Rademacher complexity bound

Intuition:

- More surrogate models
- Reducing model complexity

Upper Bound of Transferability Error

Theorem 4.3 (Upper bound of Transferability Error). *Given the transferability error defined by Eq. (5) and general rademacher complexity defined by Eq. (8). Let* $\mathcal{P}_{\bigotimes_{i=1}^{N}\Theta}$ *be the joint measure induced by the product of the marginals. If the loss function* ℓ *is bounded by* $\beta \in R_+$ *and* $\mathcal{P}_{\Theta^N}$ *is absolutely continuous with respect to* $\mathcal{P}_{\bigotimes_{i=1}^{N}\Theta}$ *for any function* f_i*, then for* $\alpha > 1$ *and* $\gamma = \frac{\alpha}{\alpha-1}$*, with probability at least* $1-\delta$*, there holds*

$$TE(z,\epsilon) \leq 4\mathcal{R}_N(\mathcal{Z}) + \sqrt{\frac{18\gamma\beta^2}{N}\ln\frac{2^{2+\frac{1}{\gamma}}H_\alpha^{\frac{1}{\alpha}}\left(\mathcal{P}_{\Theta^N}\|\mathcal{P}_{\bigotimes_{i=1}^{N}\Theta}\right)}{\delta}}, \quad (11)$$

where $H_\alpha(\cdot\|\cdot)$ *is the Hellinger integrals (Hellinger, 1909) with parameter* α*, which measures the divergence between two probability distributions if* $\alpha > 1$ *(Liese & Vajda, 2006).*

Idea: Uniform convergence bound

Proof: learning theory + information theory

Key takeaways:

- More surrogate models
- Diverse surrogate models
- Reducing model complexity

Hellinger Integral

An intuitive example in Appendix C.6 leads to the following bound:

$$TE(z,\epsilon) \leq 4\mathcal{R}_N(\mathcal{Z}) + \sqrt{18\beta^2 \ln t \cdot \frac{f(N)}{N} + 36\beta^2 \ln \frac{4\sqrt{2}}{\delta} \cdot \frac{1}{N}}$$

Several cases of key models $f(N)$:

1. $f(N) = \mathcal{O}\left(N^s\right)$, where $s \in (0,1)$
2. $f(N) = \mathcal{O}(\ln N)$
3. $f(N) = sN$, where $s \in (0,1)$

Information-theoretic Analysis

Theorem C.10. *Given N surrogate models $\boldsymbol{\theta}^N \sim \mathcal{P}_{\Theta^N}$ as the ensemble components. Let $\overline{\boldsymbol{\theta}}^N = (\overline{\boldsymbol{\theta}}_1, \ldots, \overline{\boldsymbol{\theta}}_N) \sim \mathcal{P}_{\Theta^N}$ be the target models, which is an independent copy of $\boldsymbol{\theta}^N$. Assume the loss function ℓ is bounded by $\beta \in \mathbb{R}_+$ and $\mathcal{P}_{\Theta^N}$ is absolutely continuous with respect to $\mathcal{P}_{\bigotimes_{i=1}^N \Theta}$. For $\alpha > 1$ and adversarial example $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \sim \mathcal{P}_{\mathcal{Z}}$, Let $\Delta_N(\boldsymbol{\theta}, \mathbf{z}) = L_P(\mathbf{z}) - L_E(\mathbf{z})$. Then there holds*

$$\left|\mathbb{E}_{\mathbf{z},\boldsymbol{\theta}^N \sim \mathcal{P}_{\mathcal{Z},\Theta^N}} \Delta_N(\boldsymbol{\theta}, \mathbf{z})\right| \leq 2\beta \cdot \mathrm{D_{TV}}\left(\mathcal{P}_{\Theta^N} \| \mathcal{P}_{\bigotimes_{i=1}^N \Theta}\right) + \sqrt{\frac{\alpha\beta^2}{2(\alpha-1)N}\left(I\left(\boldsymbol{\theta}^N; z\right) + \frac{1}{\alpha}\log H_\alpha\left(\mathcal{P}_{\Theta^N} \| \mathcal{P}_{\bigotimes_{i=1}^N \Theta}\right)\right)},$$

where $\mathrm{D_{TV}}(\cdot\|\cdot)$, $I(\cdot\|\cdot)$ and $H_\alpha(\cdot\|\cdot)$ denotes TV distance, mutual information and Hellinger integrals, respectively.

Key takeaways:

- More surrogate models
- Diverse surrogate models
- Reducing model complexity

Content

Experimental Setup

- **Datasets**:
  - Validation: MNIST, Fashion-MNIST, CIFAR-10
  - Exploration: ImageNet
- **Models**:
  - Validation: MLP (1-3 layers), CNN (1-3 layers), ResNet-18
  - Exploration: ResNet-50, VGG-16, MobileNet-V2, Inception-V3, ViT-B16, PiT-B, Visformer, Swin-T

Attack Dynamics

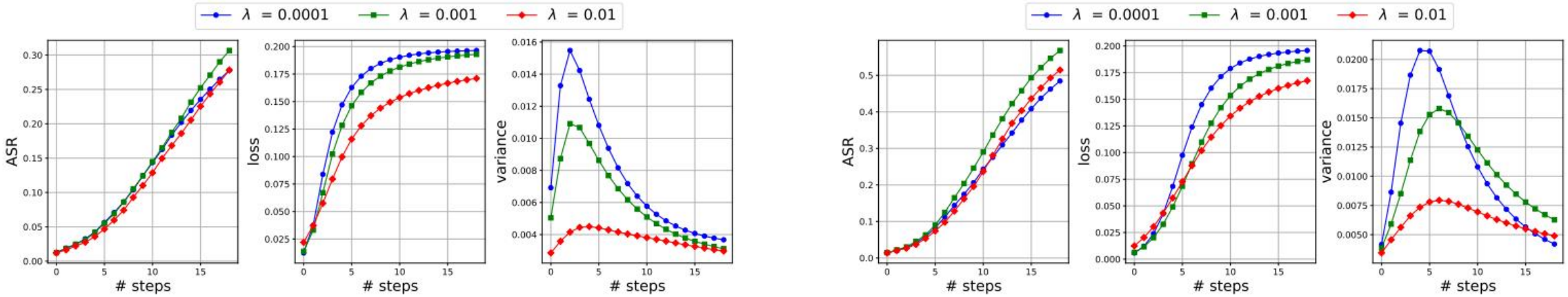
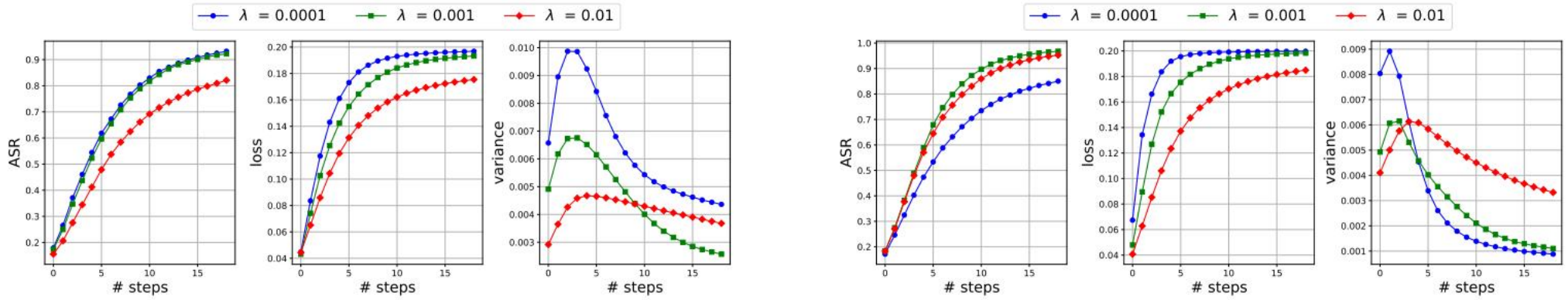


Figure 2. Evaluation of ensemble attacks with increasing the number of steps using MLPs and CNNs on the MNIST dataset.

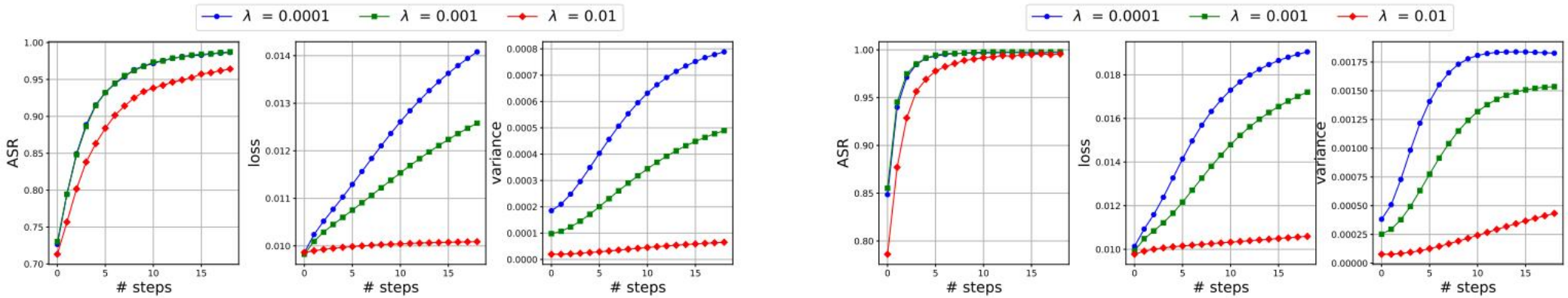
- Vulnerability-diversity decomposition
- The trend of variance
- The potential complexity-diversity trade-off

Attack Dynamics



(a) MLP (b) CNN

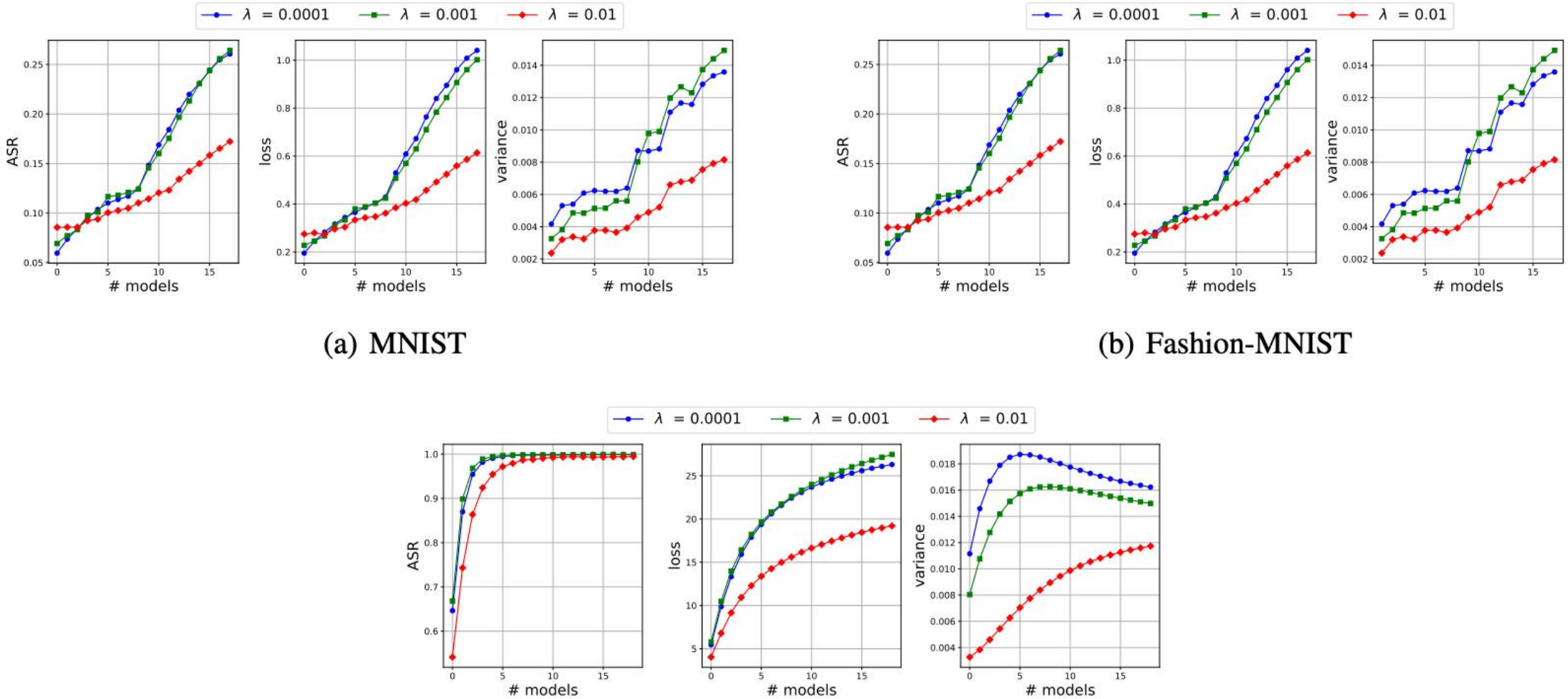
Figure 3. Evaluation of ensemble attacks with increasing the number of steps using MLPs and CNNs on the Fashion-MNIST dataset.



(a) MLP (b) CNN

Figure 6. Evaluation of ensemble attacks with increasing the number of steps using MLPs and CNNs on the CIFAR-100 dataset.

Ensemble Framework



(a) MNIST

(b) Fashion-MNIST

(c) CIFAR-10

Figure 5. Evaluation of ensemble attacks with increasing the number of models using MLPs and CNNs on the three datasets.

Model Complexity

Table 1. Effect of varying max norm constraints on adversarial attack performance, measured by classification accuracy (%, lower is better). FC and CNN denote fully connected and convolutional networks with increasing layers.

Max Norm	FC1	FC2	FC3	CNN1	CNN2	CNN3	Avg
0.1	84.66	87.80	85.39	97.57	98.31	98.59	92.05
0.5	59.37	68.31	74.05	96.50	97.66	98.34	82.37
1.0	64.31	55.27	57.12	95.37	97.08	97.93	77.85
2.0	68.00	57.40	57.86	95.41	97.04	97.87	78.93
4.0	68.19	57.94	58.12	95.53	97.00	97.85	79.11
5.0	69.68	59.40	59.26	97.48	98.02	98.87	80.45

As max norm constraint ↑, adversarial transferability first ↑ then ↓

Model Complexity

Sparse Softmax cross-entropy loss [1]

→ Less model complexity

→ Better adversarial transferability

Table 3. Transferability results of different attack methods across various target models. Bold entries indicate improved or top-performing variants.

	ResNet50	VGG16	MobileNetV2	InceptionV3	ViT-B16	PiT-B	Visformer	Swin-T
MI-FGSM	66.0	**99.9**	76.8	97.5	37.3	53.8	88.9	66.7
MI-FGSM-S	**68.9**	99.7	**79.2**	**99.1**	**39.0**	**54.5**	**90.6**	**68.1**
SVRE	65.2	**99.9**	79.0	98.6	32.4	49.2	90.3	64.3
SVRE-S	**66.9**	**99.9**	**81.2**	**98.9**	**34.2**	**51.3**	**93.0**	**65.9**
SIA	97.2	**100.0**	**98.4**	**99.7**	75.9	91.9	90.0	96.1
SIA-S	**98.1**	**100.0**	98.2	99.6	**79.2**	**93.2**	**99.5**	**97.5**

[1] Martins, A. and Astudillo, R. From softmax to sparsemax: A sparse model of attention and multi-label classification. ICML 2016.

Discussion

Machine Learning ⇋ Adversarial Transferability

- Generalization / Ensemble learning → Model ensemble attack
- Optimization → Attack algorithm
- "Key" models in the ensemble?

Thank You!