Improving Adversarial Robustness via Promoting Ensemble Diversity

Tianyu Pang[1], Kun Xu[1], Chao Du[1], Ning Chen[1] and Jun Zhu[1]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China

ICML | 2019

Motivations

One practical defense strategy is to construct ensembles of the enhanced networks to obtain stronger defenses.

- **Most of the existing defenses ignore the interactions among multiple models.**
- **It is easier for adversarial examples to transfer among individually trained models.**

The adversarial examples crafted for one member in an ensemble are probable to fool other members. So the robustness improved by the naive ensemble strategy will be limited.

Ensemble Diversity

Previous work defines the ensemble diversity (ED) w.r.t. the prediction errors, i.e., maximal predictions based on the **weak classifier assumption**. Since DNNs are strong classifiers, we define the ED on non-maximal predictions as

$$\mathbb{ED} = \det(\tilde{M}_{\setminus y}^{\top}\tilde{M}_{\setminus y}) = \mathrm{Vol}^2(\{\tilde{F}^k_{\setminus y}\}_{k\in[K]}). \tag{1}$$

Here $\tilde{M}_{\setminus y} = (\tilde{F}^1_{\setminus y}, \cdots, \tilde{F}^K_{\setminus y}) \in \mathbb{R}^{(L-1)\times K}$, each column vector $\tilde{F}^k_{\setminus y} \in \mathbb{R}^{L-1}$ is obtained by normalizing $F^k_{\setminus y}$ under L_2-norm, where $F^k_{\setminus y}$ is the order preserving prediction of the k-th classifier on x without the y-th element. The $\mathrm{Vol}(\cdot)$ denotes the volume spanned by the vectors of the input set.

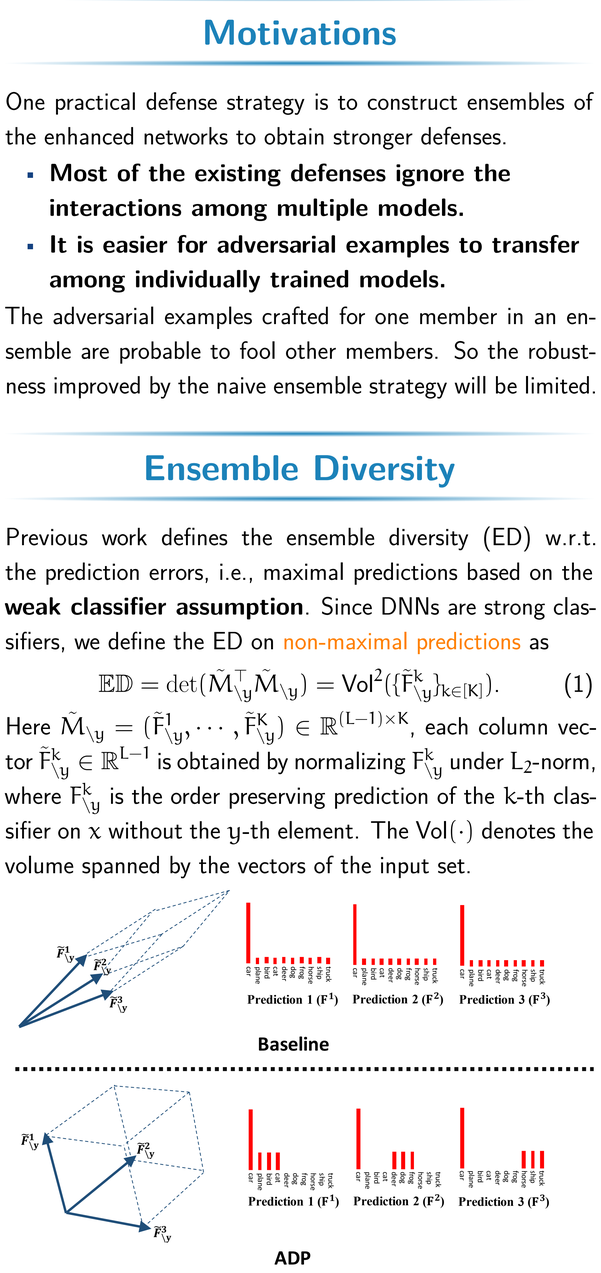

Figure 1: Illustration of our ensemble diversity

Adaptive Diversity Promoting

To promote ensemble diversity, we propose the **adaptive diversity promoting (ADP)** regularizer as

$$\mathrm{ADP}_{\alpha,\beta}(x, y) = \alpha \cdot \mathcal{H}(\mathcal{F}) + \beta \cdot \log(\mathbb{ED}) \tag{2}$$

for a single input pair (x, y), where $\mathcal{H}(\cdot)$ is Shannon entropy, $\mathcal{F}$ is the ensemble prediction, $\alpha, \beta \geq 0$ are two hyperparameters. The training loss for the ensemble model is

$$\min_{\theta} \ \mathcal{L}_{\mathrm{ECE}} - \mathrm{ADP}_{\alpha,\beta}, \tag{3}$$

where $\mathcal{L}_{\mathrm{ECE}}$ is the sum of all individual cross-entropy losses.

Theoretical Analyses

Theorem 1. If $\alpha = 0$, then $\forall \beta \geq 0$, the optimal solution of problem (3) satisfies the equations $F^k = 1_y$, where $k \in [K]$.

Remark 1. The ensemble entropy (EE) part is necessary, since we define the ED on the *normalized* predictions $\tilde{F}^k_{\setminus y}$.

Theorem 2. When $\alpha > 0$ and $\beta = 0$, the optimal solution of problem (3) satisfies $F^k_y = \mathcal{F}_y$, $\mathcal{F}_j = \frac{1-\mathcal{F}_y}{L-1}$ and

$$\frac{1}{\mathcal{F}_y} = \frac{\alpha}{K}\log\frac{\mathcal{F}_y(L-1)}{1-\mathcal{F}_y}, \tag{4}$$

where $k \in [K]$ and $j \in [L]\backslash\{y\}$.

Remark 2. $\forall j \in [L]\backslash\{y\}$, the EE part leaves degrees of freedom on the optimal solutions of non-maximal predictions F^k_j, such that the LED part can further regularize on them

Corollary 1. If there is $K \mid (L-1)$, then $\forall \alpha, \beta > 0$, the optimal solution of problem (3) satisfies the Eq. (4). Besides, let $S = \{s_1, , \cdots, s_K\}$ be any partition of the index set $[L]\backslash\{y\}$, where $\forall k \in [K]$, $|s_k| = \frac{L-1}{K}$. Then the optimal solution further satisfies:

$$F^k_j = \begin{cases} \frac{K(1-\mathcal{F}_y)}{L-1}, & j \in s_k, \\ \mathcal{F}_y, & j = y, \\ 0, & otherwise. \end{cases} \tag{5}$$

Remark 3. The partition S is adaptive, i.e., for two inputs x_1 and x_2 of the same label y, the partition S_1 and S_2 could be different. This avoids imposing an artificial bias on the relationship among different classes (illustrated in Fig.1).

Experiments

Table 1: Classification error rates (%) on each dataset. The ensemble model consists of three Resnet-20 networks.

Dataset	Classifier	Baseline	$\mathrm{ADP}_{2,0}$	$\mathrm{ADP}_{2,0.5}$
MNIST	Net 1	0.44	0.43	0.36
	Net 2	0.39	0.45	0.47
	Net 3	0.43	0.37	0.47
	Ensemble	0.32	0.32	**0.28**
CIFAR-10	Net 1	8.30	9.01	9.34
	Net 2	8.52	9.15	9.34
	Net 3	8.67	9.14	9.92
	Ensemble	6.78	6.74	**6.56**
CIFAR-100	Net 1	35.25	-	39.04
	Net 2	35.91	-	40.86
	Net 3	36.03	-	39.00
	Ensemble	30.35	-	**29.80**

Table 2: Classification accuracy (%): $\mathbf{AdvT_{FGSM}}$ denotes adversarial training (AdvT) on FGSM, $\mathbf{AdvT_{PGD}}$ denotes AdvT on PGD. $\epsilon = 0.04$ for FGSM; $\epsilon = 0.02$ for BIM/PGD/MIM.

Defense Methods	CIFAR-10 FGSM	BIM	PGD	MIM
$\mathrm{AdvT_{FGSM}}$	39.3	19.9	24.2	24.5
$\mathrm{AdvT_{FGSM}} + \mathrm{ADP}_{2,0.5}$	**56.1**	**25.7**	**26.7**	**30.6**
$\mathrm{AdvT_{PGD}}$	43.2	27.8	32.8	32.7
$\mathrm{AdvT_{PGD}} + \mathrm{ADP}_{2,0.5}$	**52.8**	**34.0**	**36.2**	**38.8**

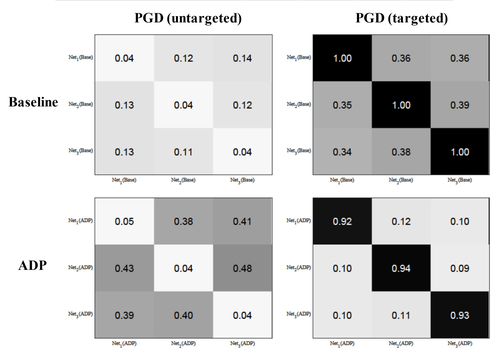

Figure 2: Adversarial transferability among individual models on CIFAR-10. For untargeted attacks, the values are the classification accuracy; For targeted attacks, those are the success rate of fooling classifiers to predict target classes.

Table 3: Classification accuracy (%) on adversarial examples.

Attacks	CIFAR-10 Para.	Baseline	$\mathrm{ADP}_{2,0}$	$\mathrm{ADP}_{2,0.5}$
FGSM	$\epsilon = 0.02$	36.5	57.4	**61.7**
	$\epsilon = 0.04$	19.4	41.9	**46.2**
BIM	$\epsilon = 0.01$	18.5	44.0	**46.6**
	$\epsilon = 0.02$	6.1	28.2	**31.0**
PGD	$\epsilon = 0.01$	23.4	43.2	**48.4**
	$\epsilon = 0.02$	6.6	26.8	**30.4**
MIM	$\epsilon = 0.01$	23.8	49.6	**52.1**
	$\epsilon = 0.02$	7.4	32.3	**35.9**
JSMA	$\gamma = 0.05$	29.5	33.0	**43.5**
	$\gamma = 0.1$	27.5	32.0	**37.0**
C&W	$c = 0.001$	71.3	76.3	**80.6**
	$c = 0.01$	45.2	50.3	**54.9**
	$c = 0.1$	18.8	19.2	**25.6**
EAD	$c = 1.0$	17.5	64.5	**67.3**
	$c = 5.0$	2.4	23.4	**29.6**

Table 4: Five-member ensembles, where $5 \nmid 9$.

Attacks	CIFAR-10 Para.	Baseline	$\mathrm{ADP}_{2,0.5}$
Normal	-	93.6	**93.8**
FGSM	$\epsilon = 0.02$	42.0	**58.4**
BIM	$\epsilon = 0.01$	31.6	**41.8**
PGD	$\epsilon = 0.01$	37.4	**44.2**
MIM	$\epsilon = 0.01$	37.1	**47.5**
C&W	$c = 0.01$	52.3	**56.5**
EAD	$c = 1.0$	20.4	**65.3**

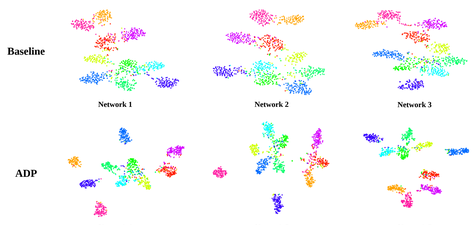

Figure 3: t-SNE visualization of features on CIFAR-10.

Contact

- **Email**: pty17@mails.tsinghua.edu.cn
- **Code**: https://github.com/P2333