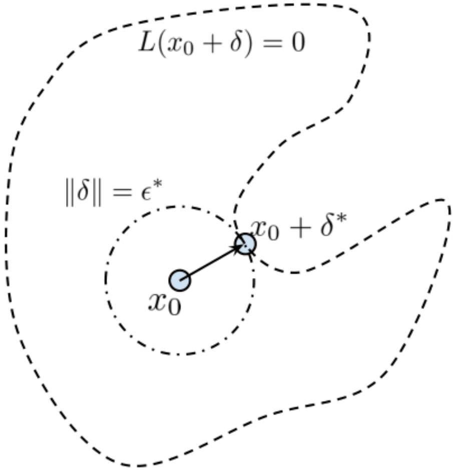
Towards Better Robust Generalization with Shift Consistency Regularization

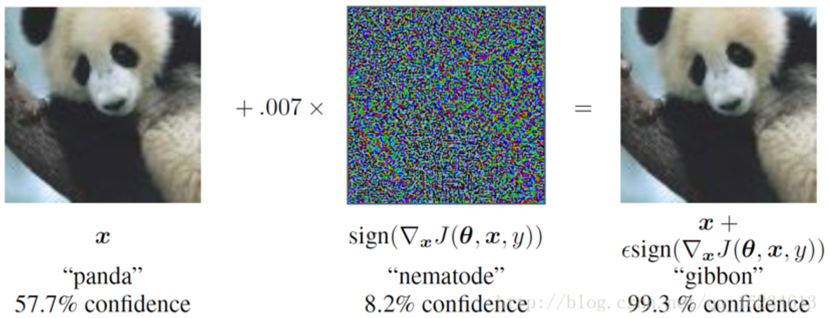
Shufei Zhang [*,1,2] **Zhuang Qian** [*,1,2] **Kaizhu Huang** [1] **Qiufeng Wang** [1] **Rui Zhang** [3] **Xinping Yi** [2]

[*]Equal contribution [1]School of Advanced Technology, Xi'an Jiaotong-Liverpool University, China. [2]School of Electrical Engineering, Electronics and Computer Science, University of Liverpool, UK. [3]School of Science, Xi'an Jiaotong-Liverpool University, China. Correspondence to: Kaizhu Huang <kaizhu.huang@xjtlu.edu.cn;kaser.huang@gmail.com>.

Adversairal Training

- Adversarial data can easily fool the standard trained classifier.
- Adversarial training is one of the most effective methods to obtain the adversarial robustness for the trained classifier.





Purporse: Maximize margin by training with worst perturbation so that the adversarial examples can not cross the decision boundary.

Conventional Adversarial Training

- Minimax formulation:

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\sim D}\left[\max_{\epsilon\in S} L(\theta, x+\epsilon, y)\right],$$

- Projected gradient descent (PGD) formulates the problem of finding the most adversarial data as a constrained optimization problem. Namely, given a starting point $x^0 \in S_x$ and step size α, PGD works as followed:

$$x^{t+1} = \Pi_{S_x}(x^t + \alpha \cdot \operatorname{sgn}(\nabla_x L(x^t, y; \theta))),$$

Feature Scattering based Adversarial Training (FS)

- Maxmize the wesserstein distance of outputs of clean and perturbed data (inter sample relationship is considerred).

$$\min_{\theta} \quad \frac{1}{N}\sum_{i=1}^{N} L(x_i^{adv}, y_i; \theta)$$

$$\text{s.t.} \quad \nu^* = \sum_{i=1}^{N} v_i \delta_{x_i^{adv}} = \arg\max_{\nu \in S_\mu} D_{ot}(\nu, \mu)$$

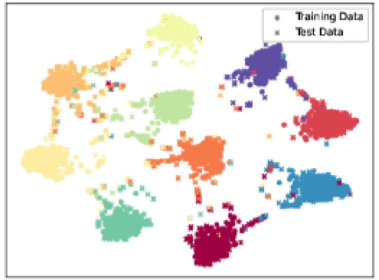
where $\mu = \sum_{i=1}^{N} u_i \delta_{x_i}$ and $\nu = \sum_{i=1}^{N} v_i \delta_{x_i^{adv}}$ are two discrete distributions for natural examples $\{x_i\}_{i=0}^{N}$ and perturbed examples $\{x_i'\}_{i=0}^{N}$ respectively and $V = \{v_i\}_{i=1}^{N}$ and $U = \{u_i\}_{i=1}^{N}$ are corresponding weights. Here, $v_i = u_i = 1/N$. $S_\mu = \{\sum_i v_i \delta_{z_i}, | z_i \in B(x_i, \epsilon) \cap [0, 255]^d\}$ denotes the feasible region. $D_{ot} = \min_{T \in \prod(U,V)} \sum_{i=1}^{N} \sum_{j=1}^{N} T_{ij} c(x_i, x_j')$ is the optimal transport (OT) distance where $\prod(U, V) = \{T \in \mathbb{R}_+^{N \times N} | T 1_N = U, T^\top 1_N = V\}$ and 1_N denotes all-one vector. Here, the cost function is defined as $c(x_i, x_j') = 1 - \frac{f_\theta(x_i)^\top f_\theta(x_j')}{\|f_\theta(x_i)\|_2 \|f_\theta(x_j')\|_2}$ to measure the feature similarity.

Generalization Issue of Adversarial Training

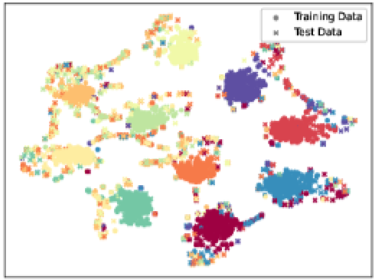
- While previous methods achieve impressive robustness performance, there still exists a big robust generalization gap between training and test sets.
- The robust generalization gap (training accuracy - test accuracy) of AT is around 40% and FS is around 20%.

Analysis for Generalization Issue

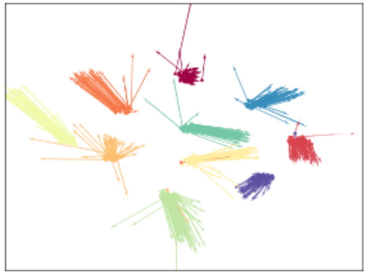
- Visualization for output feautures and effect of adversarial perturbations on feature shifts.
- Adversarial perturbations cause the different feature shifts for test and training data and lead to generalization issue.



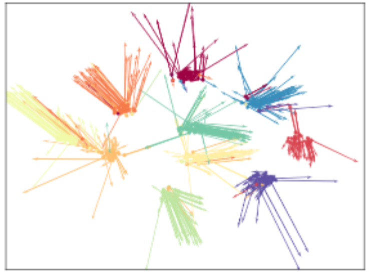
(a) FS w/ clean data



(b) FS w/ adversarial data



(a) FS training shifts



(b) FS test shifts

Theoretical Analysis for Robust Generalization

- Relationship between robust generalization and standard generalization. (Difference is feature shift inconsistancy)

Theorem 6.1.1. *Given the training set $S_d = \{x_i\}_{i=1}^{n}$ that consists of n i.i.d samples drawn from a distribution S with K classes, and the set of corresponding adversarial examples $S_d^{adv} = \{x_i^{adv}\}_{i=1}^{n}$ drawn from the underlying distribution S^{adv}, if the loss function $l(\cdot)$ of DNN f_θ is λ-Lipschitz, then for any $\delta > 0$, with the probability at least $1 - \delta$, we have*

$$RGE \leq GE + \boxed{\frac{\lambda}{n} \sum_{i=1}^{K} \sum_{j \in N_i} \| d_\theta(x_j^{adv}) - \hat{d}_\theta(z, C_i) \|_2^2} + M \sqrt{\frac{2K \ln 2 + 2 \ln \frac{1}{\delta}}{n}} \tag{6.1}$$

feature shift inconsistancy

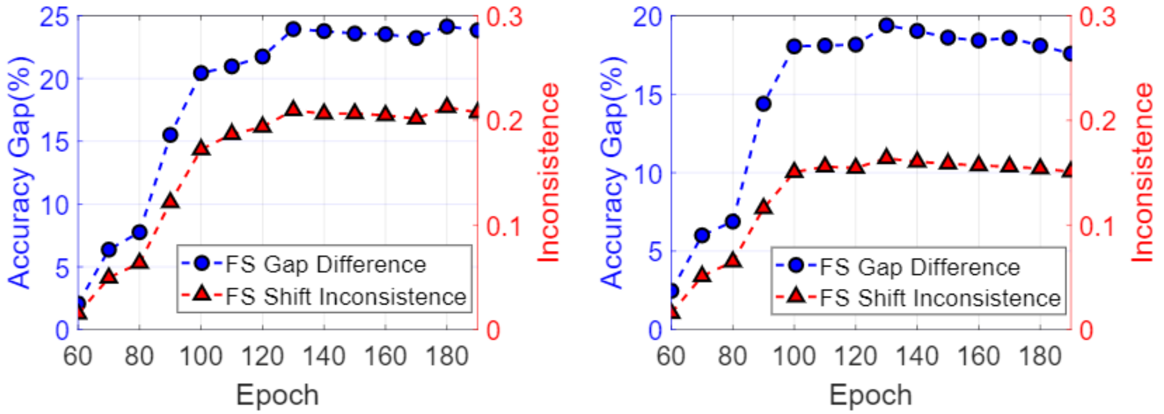
where

$$d_\theta(x^{adv}) = f_\theta(x^{adv}) - f_\theta(x)$$

$$\hat{d}_\theta(z, C_i) = \mathbb{E}[f_\theta(z^{adv}) - f_\theta(z) | z \in C_i]$$

Theoretical Analysis for Robust Generalization

- Relationship between robust generalization and feature shift inconsistancy. (The changes of the shift inconsistency and gap difference RGE−GE are consistent.)



(a) CW (b) PGD

Adversarial Training with Shift Consistency Regularization

- Penalize the feature shift inconsistency.

$$\min_{\theta} \quad \Big\{ \sum_{i=1}^{n} [L(x_i^{adv}, y_i; \theta)] + \frac{\lambda}{n} \sum_{i=1}^{K} \sum_{j \in N_i} \widehat{SiC}(x_j^{adv}, x_l, N_i) \Big\},$$

$$\text{s.t.} \quad x_i^{adv} = \arg\max_{x_i' \in S_{x_i}} L(x_i', y_i; \theta).$$

$$\widehat{SiC}(x_j^{adv}, x_l, N_i) \triangleq \|d_\theta(x_j^{adv}) - \bar{d}_\theta(x_l, N_i)\|_2^2,$$

where

$$d_\theta(x^{adv}) = f_\theta(x^{adv}) - f_\theta(x)$$

$$\hat{d}_\theta(z, C_i) = \mathbb{E}[f_\theta(z^{adv}) - f_\theta(z) | z \in C_i]$$

To consider different types of attacks, we penalize the upper bound of shift inconsistency:

$$\max_{x_j' \in S_{x_i}} \widehat{\mathrm{SiC}}(x_j', x_l, N_i).$$

We approximate test feature shft with average feature shift over training data.

$$\widehat{\mathrm{SiC}}(x_j', \mu_i) \triangleq \|d_\theta(x_j') - \mu_i\|_2^2$$

Some Results

Table 1. Accuracy under white-box attacks on CIFAR-10

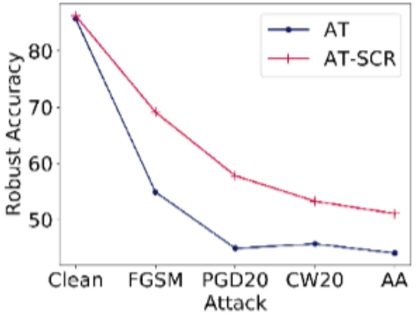
MODELS	CLEAN	ACCURACY UNDER WHITE-BOX ATTACK ($\epsilon = 8$)						
		FGSM	PGD20	PGD40	PGD100	CW20	CW40	CW100
STANDARD	**95.60**	36.90	0.00	0.00	0.00	0.00	0.00	0.00
AT	85.70	54.90	44.90	44.80	44.80	45.70	45.60	45.40
TLA	86.21	58.88	51.59	-	-	-	-	-
LAT	87.80	-	53.84	-	53.04	-	-	-
BILATERAL	91.20	70.70	57.50	–	55.20	56.20	–	53.80
FS	90.00	78.40	70.50	70.30	**68.60**	62.40	62.10	60.60
RST-AWP	88.25	67.94	63.73	-	63.58	61.62	-	-
RLFAT$_T$	82.72	-	58.75	-	-	51.94	-	-
RLFAT$_P$	84.77	-	53.97	-	-	52.40	-	-
FS-SCR	92.70	**89.87**	**76.45**	**71.60**	67.79	**75.42**	**72.69**	**69.79**

Some Results

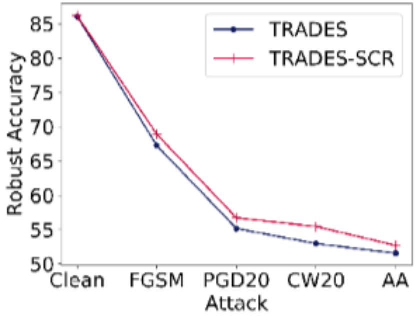
Table 2. Accuracy under different white-box attack on CIFAR-100 and SVHN

MODELS	CIFAR-100($\epsilon = 8$)						SVHN($\epsilon = 8$)					
	CLEAN	FGSM	PGD20	PGD100	CW20	CW100	CLEAN	FGSM	PGD20	PGD100	CW20	CW100
STANDARD	**79.00**	10.00	0.00	0.00	0.00	0.00	**97.20**	53.00	0.30	0.10	0.30	0.10
AT	59.90	28.50	22.60	22.30	23.20	23.00	93.90	68.40	47.90	46.00	48.70	47.30
LAT	60.94	-	27.03	26.41	-	-	60.94	-	60.23	59.97	-	-
BILATERAL	68.20	60.80	26.70	25.30	-	22.10	94.10	69.80	53.90	50.30	-	48.90
FS	73.90	61.00	47.20	46.20	34.60	30.60	96.20	83.50	62.90	52.00	61.30	50.80
AT-AWP	-	-	30.71	-	-	-	-	-	59.12	-	-	-
RLFAT$_T$	58.96	-	31.63	-	27.54	-	-	-	-	-	-	-
RLFAT$_P$	56.70	-	31.99	-	29.04	-	-	-	-	-	-	-
FS-SCR	74.20	**72.19**	**48.87**	**47.34**	**38.90**	**33.60**	**96.60**	**92.52**	**70.24**	**60.72**	**64.62**	**54.90**

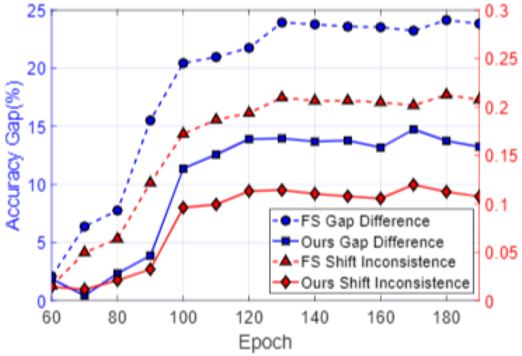
Some Results



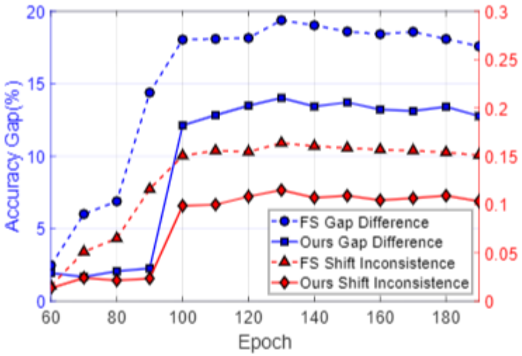
(a) AT



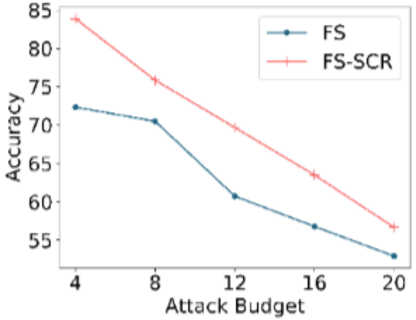
(b) TRADES



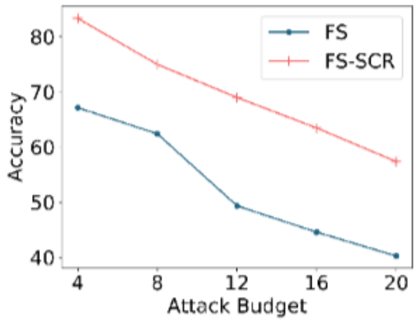
(a) CW 20



(b) PGD 20

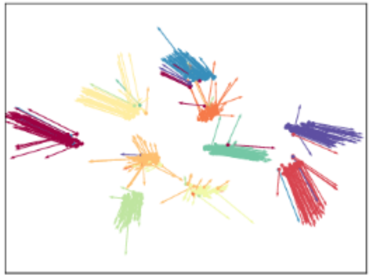


(d) PGD Attack Budget



(e) CW Attack Budget

(c) FS-SCR training shifts



(d) FS-SCR test shifts

Thanks for your attention