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]

A. Proof for Theorem 3.1

In this section, we provide the details of proof of Theorem 3.1.

Theorem 3.1 *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 k-Lipschitz, then for any $\delta > 0$, with the probability at least $1-\delta$, we have*

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

$$where \qquad d_\theta(x^{adv}) = f_\theta(x^{adv}) - f_\theta(x) \tag{2}$$

$$\hat{d}_\theta(z, C_i) = \mathbb{E}[f_\theta(z^{adv}) - f_\theta(z)|z \in C_i] \tag{3}$$

with N_i being the set of index of training data for class i, C_i the set of i^{th} class data of the whole set and z is data sampled from C_i with corresponding adversarial example z^{adv}, M the upper bound of loss of the whole data manifold S.

Proof: Let N_i be the set of index of points of training set $S_d = \{s_i\}_{i=1}^n$ that fall into the C_i and $(|N_1|, ..., |N_K|)$ is an i.i.d multinomial random variable with parameters n and $(\mu(C_1), ..., \mu(C_K))$. The following holds by the Breteganolle-Huber-Carol inequality (cf Proposition A6.6 of (Van & Wellner, 2000)):

$$Pr\left\{\sum_{i=1}^{K}\left|\frac{N_i}{n} - \mu(C_i)\right| \geq \lambda\right\} \leq 2^K exp(\frac{-n\lambda^2}{2}) \tag{4}$$

Hence, with the probability at least $1-\delta$, we have:

$$\sum_{i=1}^{K}\left|\frac{N_i}{n} - \mu(C_i)\right| \leq \sqrt{\frac{2Kln2 + 2ln(1/\delta)}{n}} \tag{5}$$

The upper bound of robust generalization can be formulated as:

[*]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>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

$$
\begin{aligned}
&|l(f_\theta(S^{adv}),Y)-\hat{l}(f_\theta(S_d^{adv}),Y_d)| = \left|\sum_{i=1}^{K}\mathbb{E}(l(f_\theta(z^{adv}),y)|z\in C_i)\mu(C_i)-\frac{1}{n}\sum_{i=1}^{n}l(f_\theta(x_i^{adv}),y_i)\right| \\
&=|\sum_{i=1}^{K}(\mathbb{E}(l(f_\theta(z^{adv}),y)|z\in C_i)-\mathbb{E}(l(f_\theta(z),y)|z\in C_i)+\mathbb{E}(l(f_\theta(z),y)|z\in C_i))\mu(C_i) \\
&-\frac{1}{n}\sum_{i=1}^{n}(l(f_\theta(x_i^{adv}),y_i)-l(f_\theta(x_i),y_i)+l(f_\theta(x_i),y_i))| \\
&\le|\sum_{i=1}^{K}(\mathbb{E}(l(f_\theta(z^{adv}),y)|z\in C_i)-\mathbb{E}(l(f_\theta(z),y)|z\in C_i))\mu(C_i) \\
&-\frac{1}{n}\sum_{i=1}^{n}(l(f_\theta(x_i^{adv}),y_i)-l(f_\theta(x_i),y_i))|+\left|\sum_{i=1}^{K}\mathbb{E}(l(f_\theta(z),y)|z\in C_i)\mu(C_i)-\frac{1}{n}\sum_{i=1}^{n}l(f_\theta(x_i),y_i)\right| \\
&\le GE+\left|\sum_{i=1}^{K}(\mathbb{E}(l(f_\theta(z^{adv}),y)|z\in C_i)-\mathbb{E}(l(f_\theta(z),y)|z\in C_i))\frac{|N_i|}{n}-\frac{1}{n}\sum_{i=1}^{n}(l(f_\theta(x_i^{adv}),y_i)-l(f_\theta(x_i),y_i))\right| \\
&+\left|\sum_{i=1}^{K}(\mathbb{E}(l(f_\theta(z^{adv}),y)|z\in C_i)-\mathbb{E}(l(f_\theta(z),y)|z\in C_i))\mu(C_i)-\sum_{i=1}^{K}(\mathbb{E}(l(f_\theta(z^{adv}),y)|z\in C_i)-\mathbb{E}(l(f_\theta(z),y)|z\in C_i))\frac{|N_i|}{n}\right| \\
&\le GE+\left|\sum_{i=1}^{K}(\mathbb{E}(l(f_\theta(z^{adv}),y)|z\in C_i)-\mathbb{E}(l(f_\theta(z),y)|z\in C_i))\frac{|N_i|}{n}-\frac{1}{n}\sum_{i=1}^{n}(l(f_\theta(x_i^{adv}),y_i)-l(f_\theta(x_i),y_i))\right| \\
&+M\sum_{i=1}^{K}\left|\frac{N_i}{n}-\mu(C_i)\right| \\
&\le GE+\frac{1}{n}\sum_{i=1}^{K}\sum_{j\in N_i}\left|(l(f_\theta(x_j^{adv}),y_j)-l(f_\theta(x_j),y_j))-\mathbb{E}(l(f_\theta(z^{adv}),y)-l(f_\theta(z),y)|z\in C_i)\right|+M\sum_{i=1}^{K}\left|\frac{N_i}{n}-\mu(C_i)\right|
\end{aligned}
$$

Here, we assume $|l(f_\theta(x_1),y_1)-l(f_\theta(x_2),y_2)|\le k\|f_\theta(x_1)-f_\theta(x_2)\|_2^2$ and then we have

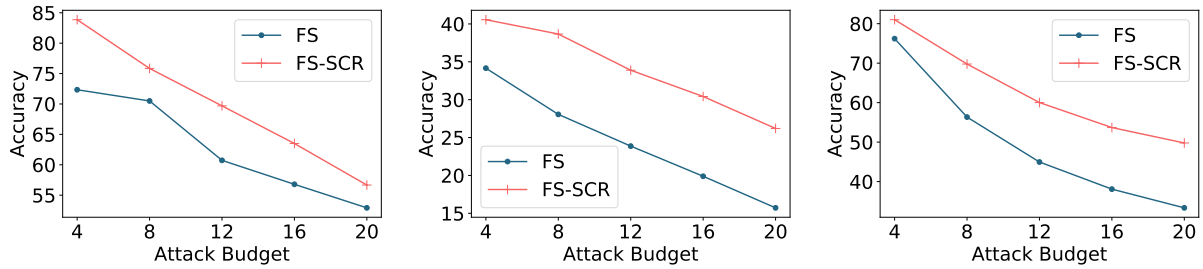
$$
\begin{aligned}
&RGE \triangleq |l(f_\theta(S^{adv}),Y)-\hat{l}(f_\theta(S_d^{adv}),Y_d)| \\
&\le GE+\frac{k}{n}\sum_{i=1}^{K}\sum_{j\in N_i}\|(f_\theta(x_j^{adv})-f_\theta(x_j))-\mathbb{E}(f_\theta(z^{adv})-f_\theta(z)|z\in C_i)\|_2^2+M\sum_{i=1}^{K}\left|\frac{N_i}{n}-\mu(C_i)\right| \\
&\le GE+\frac{k}{n}\sum_{i=1}^{K}\sum_{j\in N_i}\|d_\theta(x_j^{adv})-\hat{d}_\theta(z,C_i)\|_2^2+M\sqrt{\frac{2K\ln 2+2\ln\frac{1}{\delta}}{n}}
\end{aligned}
$$

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

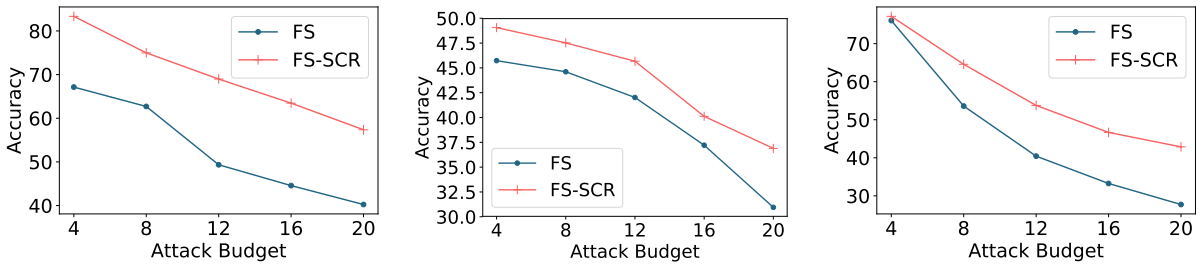
B. Additional Experiment Details

B.1. The Performance of FS-SCR for Different Attack Budgets

We further evaluate the model robustness against PGD and CW attacks under different attack budgets with a fixed attack step of 20 over CIFAR-10, CIFAR-100 and SVHN. These results are shown in Figure 1. The Feature Scattering method (FS) can improve the model robustness across a wide range of attack budgets. The proposed approach FS-SCR further boosts the performance over Feature Scattering by a large margin under different attack budgets for both PGD and CW attacks,especially for large attack budgets.



(a) PGD on CIFAR-10 (b) PGD on CIFAR-100 (c) PGD on SVHN



(d) CW on CIFAR-10 (e) CW on CIFAR-100 (f) CW on svhn

Figure 1. Model performance under PGD and CW attacks with different attack budgets.

B.2. Sensitivity Analysis

For our proposed SCR, there are two important hyper parameters, trade-off parameter λ and attack iteration c_2 (as in Algorithm 1) for SCR. We plot the test acurracy of FS-SCR over different λ and c_2 in Figure 2. It can be noted that the best accuracy is achieved when $\lambda = 0.01$ with fixed $c_2 = 3$. Weighing the accuracy and computation complexity, we set $c_2 = 3$.

B.3. Black-box results on CIFAR-10

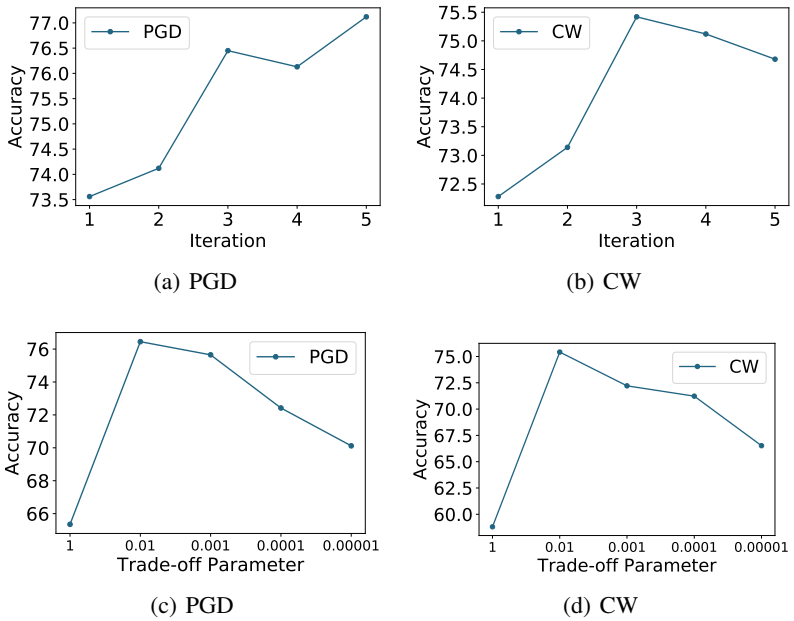
We conduct more evaluations on the transfer-based black-box attacks on CIFAR-10. We report the results in Table 1. It can be observed that our proposed methods (FS-SCR and AT-SCR) overall outperform baseline methods (FS and AT) in most of the cases on CIFAR-10. Surprisingly, the baselines perform better than our methods in four cases. This also partially reveals the more challenging nature of defending black-box attacks than white-box attacks.

Table 1. Accuracy under black-box attack on CIFAR-10

Defense Models	Attacked Models (CIFAR-10)														
	vanilla training			Adversarial Training			FS			FS+SCR			PGD+SCR		
	FGSM	PGD20	CW20	FGSM	PGD20	CW20	FGSM	PGD20	CW20	FGSM	PGD20	CW20	FGSM	PGD20	CW20
AT	84.62	84.89	84.83	72.2	63.77	63.27	82.26	80.56	79.31	84.45	81.47	81.70	72.57	64.07	63.48
FS	88.64	89.25	89.31	77.18	66.59	66.40	82.8	81.01	78.34	88.00	82.81	83.09	76.47	67.18	67.14
AT+SCR	84.78	84.84	84.74	72.43	63.78	63.41	82.65	81.17	79.89	84.56	81.84	82.1	72.66	63.95	63.69
FS+SCR	89.84	91.89	91.81	79.33	67.44	66.62	84.76	83.22	81.17	90.43	81.45	82.09	79.14	67.43	66.67

B.4. Illustration of Feature Shift

In addition, we plot the feature shifts caused by adversarial perturbations (CW attack) for both the training data and test data in Figure 3 where the feature shift for a data sample x is defined as $f_\theta(x^{adv}) - f_\theta(x)$. Comparing Figure 3(a) 3(b) and 3(c) 3(d), it can be noted that our method obtains the more consistent feature shifts and the feature shifts of FS are more dispersed. Thus, our method can obtain the similar latent features of training and test data which leads to better robust generalization.



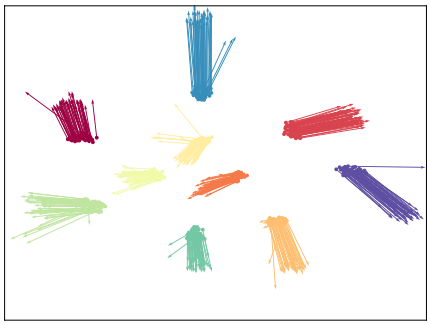
(a) PGD (b) CW

(c) PGD (d) CW

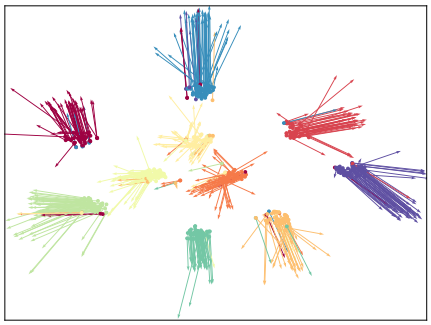
Figure 2. Model performance under different iteration and trade-off parameter on CIFAR-10

References

Van, der Vaart, A. W. and Wellner, J. A. *Weak Convergence and Empirical Processes*. Springer, 2000.

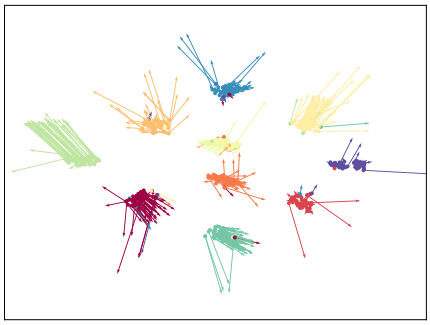


(a) Training data feature shifts



(b) Test data feature shifts

(c) Training data feature shifts



(d) Test data feature shifts

Figure 3. The feature shifts of the training and test data (Attacked by CW).