Learning Intelligence
& Vision Essential
(LiVE) Group

CVPR
JUNE 17-21, 2024

Robust Overfitting Does Matter: Test-Time Adversarial Purification With FGSM

唐林渝, 张磊 (✉)

重庆大学
微电子与通信工程学院

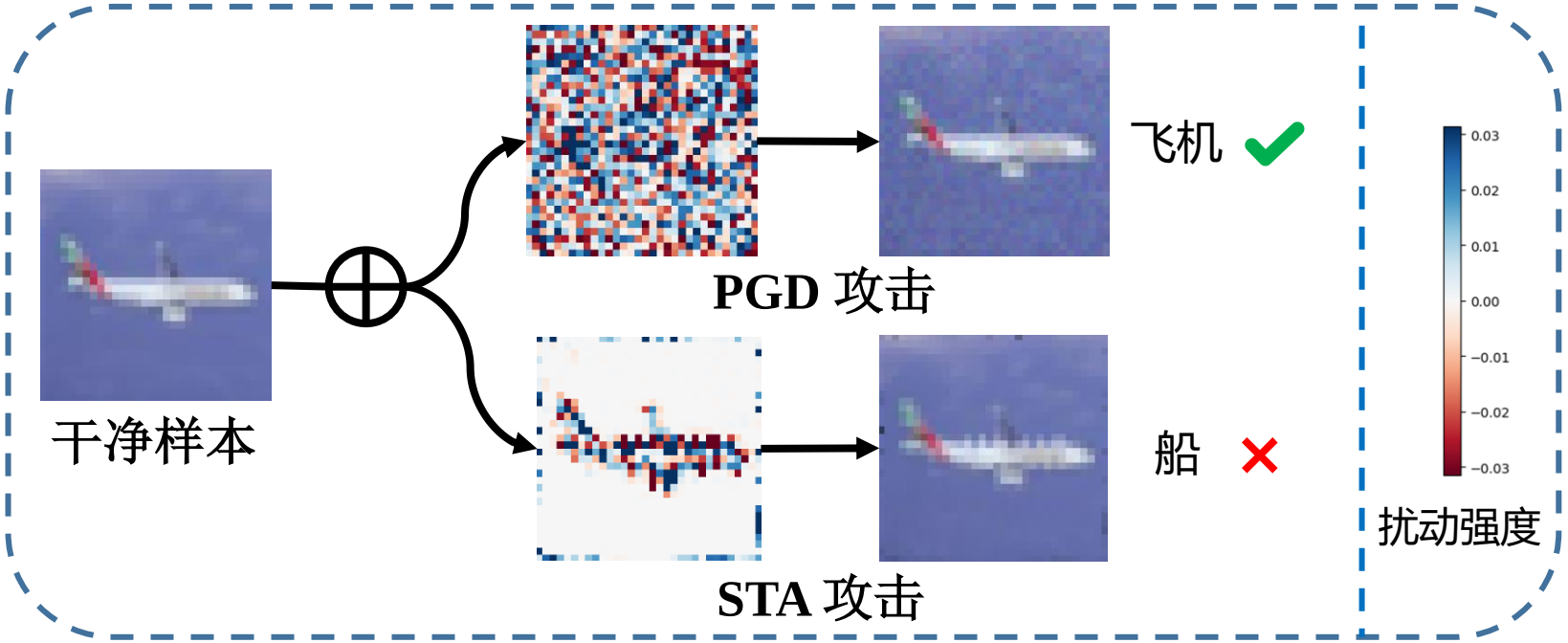
2024.05

对抗防御任务面临的挑战

问题 ①

深度神经网络面对其它未知类型攻击方法的脆弱性

使用PGD攻击训练的网络，能够对PGD对抗样本具备一定的鲁棒性，但是对STA对抗样本分类错误率非常高。PGD攻击和STA攻击作用在图像上的区别，如下图所示：



对抗防御任务面临的挑战

问题 ② 对抗训练方法对干净样本和对抗样本的分类准确率偏低

ResNet18网络	CIFAR-10数据集		
方法	Clean	FGSM	PGD-20
Clean	93.04	0	0
PGD-AT	84.54	55.11	48.91
TRADES	83.22	58.51	54.97
MART	82.14	59.57	55.39

问题 ③ 消耗巨大的计算资源和训练时间

相同实验条件下（NVIDIA 1080 TI，训练批量设置为128），自然训练和对抗训练所用的时间和消耗的计算资源对比：

方法	训练时间（秒/批次）	测试时间（秒/批次）	训练FLOPs（万亿字节/批次）	测试FLOPs（万亿字节/批次）	网络参数量（兆字节）
Clean	21.74	2.02	161.5	11.1	11.17
PGD_{10}-AT	273.89	2.02	1831.5	11.1	11.17

引入鲁棒过拟合：使用FGSM对抗训练

文章1. FAST IS BETTER THAN FREE: REVISITING ADVERSARIAL TRAINING. ICLR 2020.

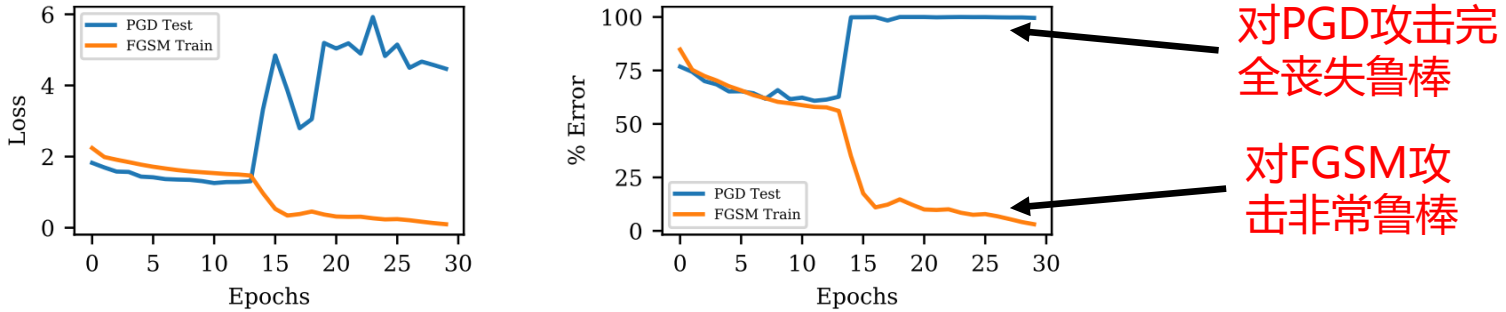


Figure 4: Learning curves for FGSM adversarial training plotting the training loss and error rates incurred by an FGSM and PGD adversary when trained with zero-initialization FGSM at $\epsilon = 8/255$, depicting the catastrophic overfitting where PGD performance suddenly degrades while the model overfits to the FGSM performance.

文章2. Understanding and Improving Fast Adversarial Training. NeurIPS 2020.

黑色虚线：
对FGSM攻击非常鲁棒

黑色实线：
对PGD攻击完全丧失鲁棒

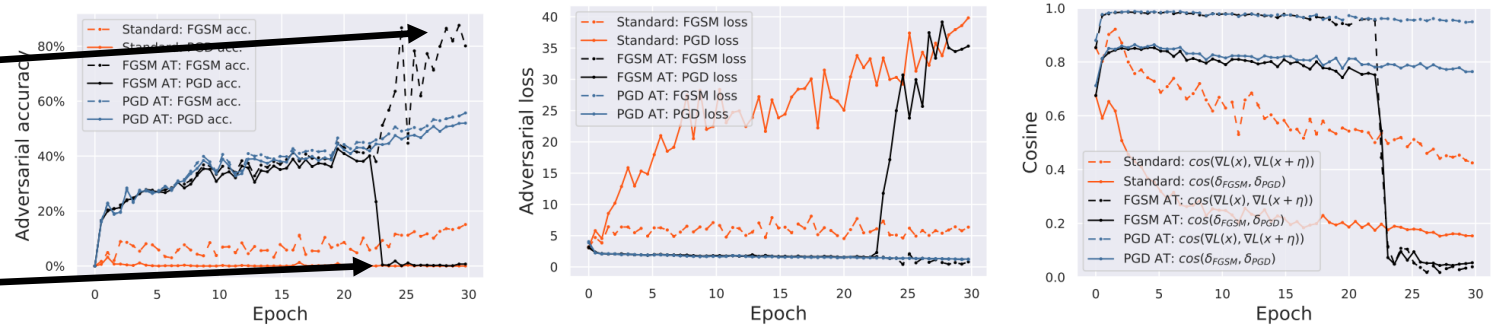
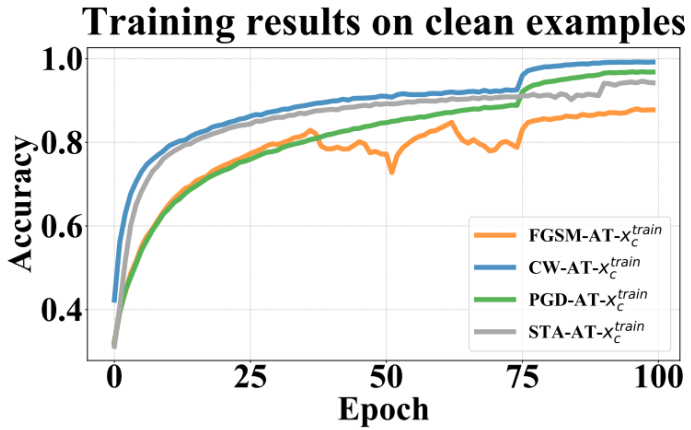


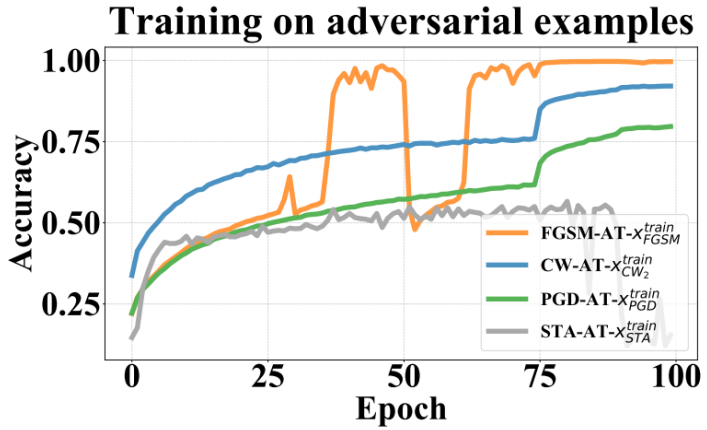
Figure 4: Visualization of the training process of standardly trained, FGSM trained, and PGD-10 trained ResNet-18 on CIFAR-10 with $\varepsilon = 8/255$. All the statistics are calculated on the test set. Catastrophic overfitting for the FGSM AT model occurs around epoch 23 and is characterized by a sudden drop in the PGD accuracy, a gap between the FGSM and PGD losses, and a dramatic decrease of *local linearity*.

探究其它攻击方法训练网络是否会鲁棒过拟合

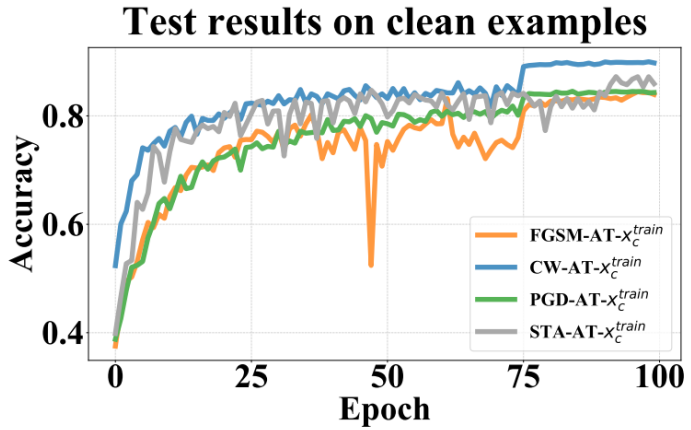
实验结论：只有使用FGSM攻击对抗训练的网络才能在其攻击下保持高鲁棒性



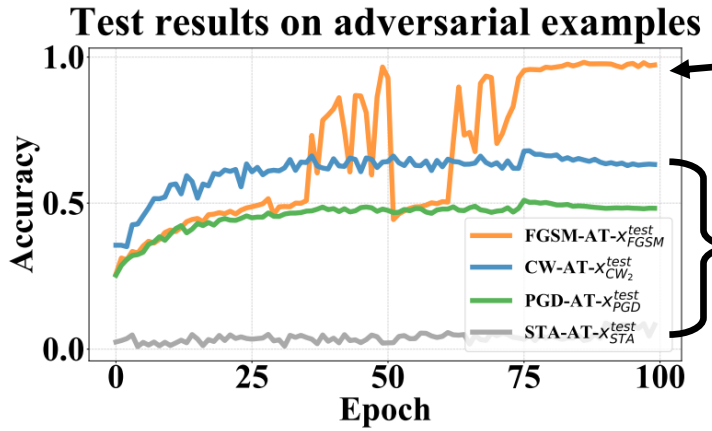
(a)



(b)



(c)



(d)

对FGSM攻击非常鲁棒

在各自的攻击下表现一般

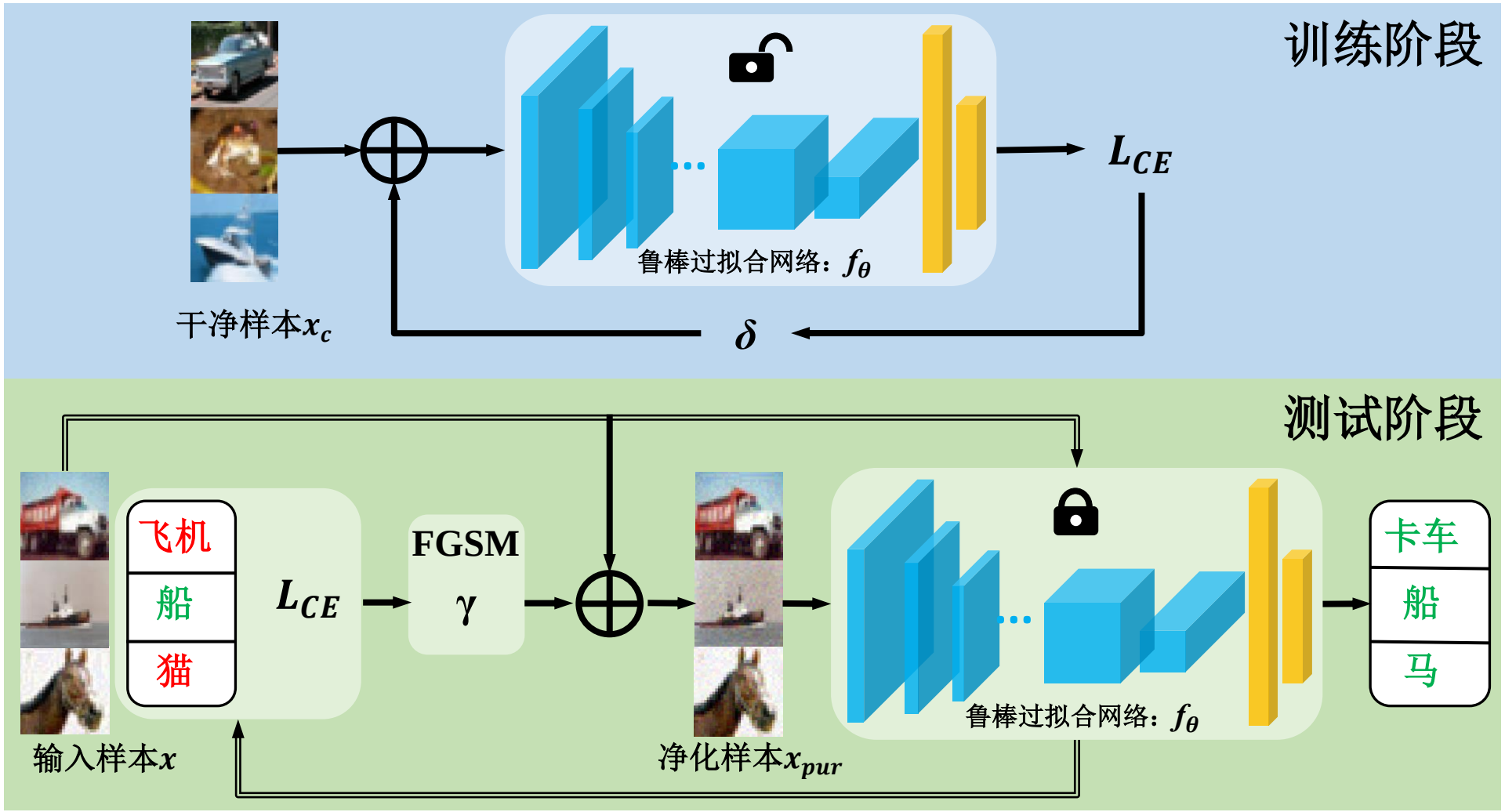
采用不同攻击方法训练网络，展示了随着训练的进行在训练集和测试集的分类正确率变化。

概述：基于鲁棒过拟合先验的对抗性净化方法

由于鲁棒过拟合网络对FGSM攻击具有高鲁棒性，所以FGSM攻击可以用在测试时间净化输入样本。

训练：
获得鲁棒过拟合网络

测试：
获得净化样本



方法概览

FGSM对抗训练实现鲁棒过拟合

① 生成FGSM对抗样本

$$\delta = \alpha * sign(\frac{\partial L_{CE}(f_\theta(x_c), y)}{\partial x_c}) \quad (1)$$

$$x_a = x_c + \delta \quad (2)$$

其中，L_{CE} 表示交叉熵损失函数，
计算公式为：

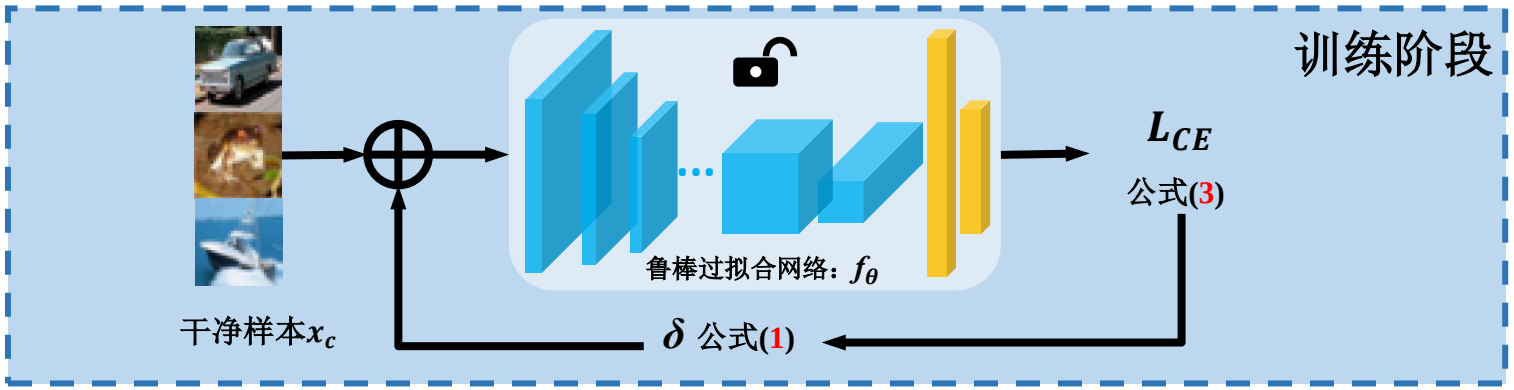
$$L_{CE}(f_\theta(x), y) = -\sum_{s=0}^{C-1} P(x,s) * log\frac{e^{f_\theta(x,s)}}{\sum_{k=0}^{C-1} e^{f_\theta(x,k)}} = -log\frac{e^{f_\theta(x,y)}}{\sum_{k=0}^{C-1} e^{f_\theta(x,k)}} \quad (3)$$

其中，$sgin(\cdot)$表示符号函数，计算公式为：

$$sgin(x) = \begin{cases} 1 & x>0 \\ 0 & x=0 \\ -1 & x<0 \end{cases} \quad (4)$$

② 更新深度神经网络的参数

$$\theta = \theta - \frac{\partial L_{CE}(f_\theta(x_a), y)}{\partial \theta} \quad (5)$$



符号解释： x (输入样本)、x_c (干净样本) 、x_a (对抗样本)、y (真实标签) 、f (深度神经网络，参数为θ)、α (步长)、δ (对抗扰动) 、$P(x, k)$ (输入样本x分类为k的概率，真实标签输出的概率值) 、$f_\theta(x, k)$(输入样本x分类为k的概率，网络输出的概率值)

测试时间像素级净化对抗扰动

① 获得预分类标签

$$y_{pred}(x) = \max\left(f_\theta(x)\right) \tag{6}$$

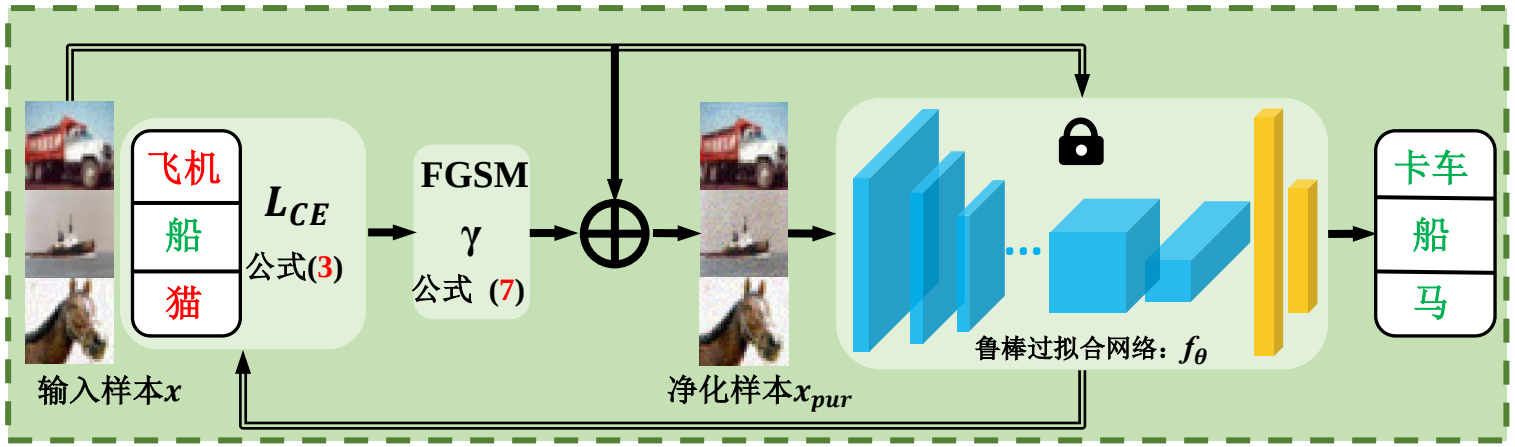
② 生成净化样本

$$\gamma = \beta * sign\left(\frac{\partial L_{CE}(f_\theta(x), y_{pred}(x))}{\partial x}\right) \tag{7}$$

$$x_{pur} = x + \gamma \tag{8}$$

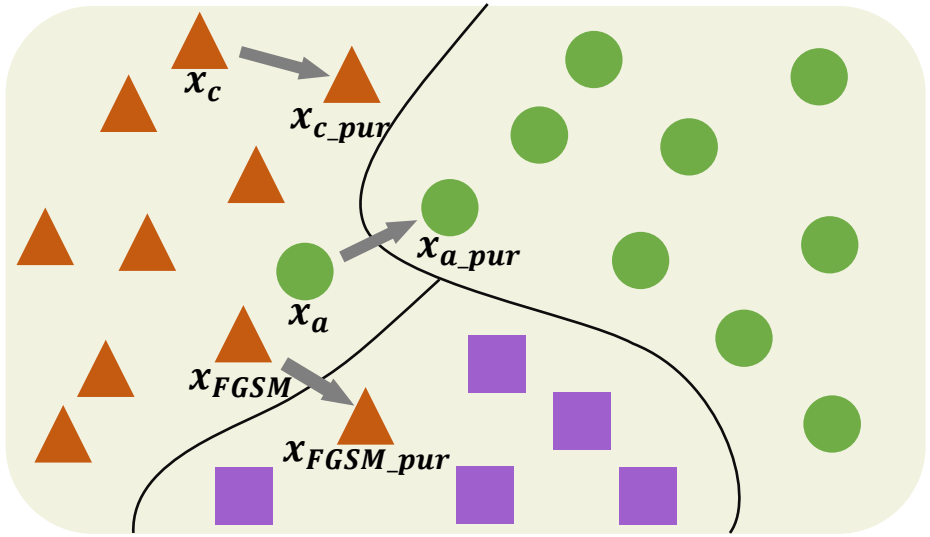
② 获得最终的分类标签

$$y_{pred}(x_{pur}) = \max\left(f_\theta(x_{pur})\right) \tag{9}$$

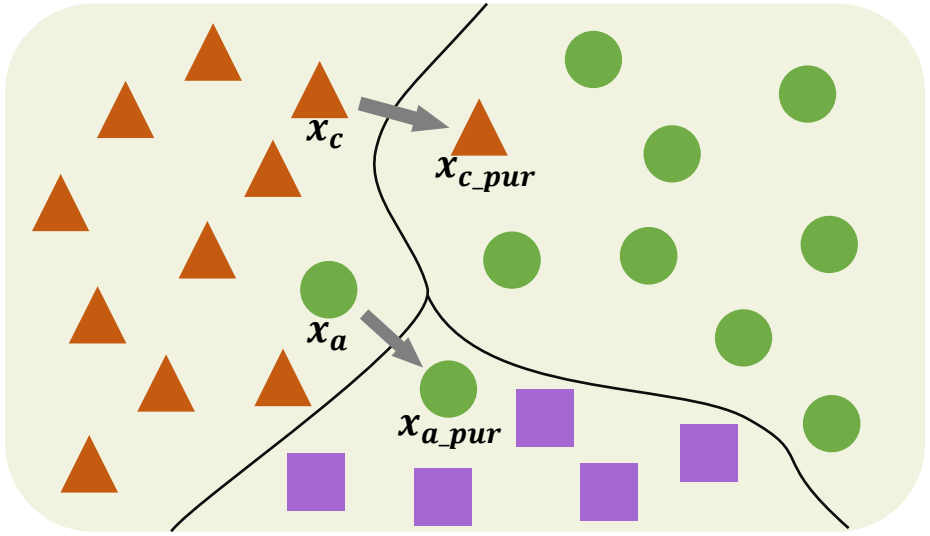


符号解释： x (输入样本)、x_c (干净样本) 、x_a (对抗样本) 、x_{pur} (净化样本) 、y (真实标签) 、y_{pred} (预测标签) 、f (深度神经网络，参数为θ)、β (步长)、γ (对抗净化) 、$P(x, k)$ (输入样本x分类为k的概率，真实标签输出的概率值) 、$f_\theta(x, k)$(输入样本x分类为k的概率，网络输出的概率值)

为什么使用鲁棒过拟合网络能够实现对抗性净化



(a)使用FGSM攻击训练的鲁棒过拟合网络



(b)使用非FGSM攻击训练的网络

鲁棒过拟合网络的特性：①对干净样本和FGSM对抗样本分类正确率高
②对非FGSM对抗样本分类正确率低

对于干净样本，如果它们在对抗性净化前被正确分类，那么净化处理就是生成了FGSM对抗样本。由于特性①，所以网络能以较高的置信度对干净样本进行正确分类。

如果干净示例在净化前被误分类，净化后继续被误分类的可能性很高，而且由于发生这种情况的概率很低，可以忽略不计。

对于对抗性样本，如果它们在对抗性净化前被错误分类，预分类标签就是错误的分类标签。利用预分类分类标签进行FGSM反向对抗性净化，可以有效地将对抗样本拉向决策边界，消除图像上的对抗扰动。由于特性①和②的联合作用，实现正确分类。

实验设置

数据集	使用对抗防御任务中主流的分类数据集 CIFAR-10, CIFAR-100, SVHN, Tiny-ImageNet
网络	常见的深度神经网络 ResNet-18, VGG-16, WideResNet-34
攻击	常见的攻击方法 FGSM, PGD, CW, DDN, STA, FWA, AutoAttack (AA), TI-DIM
设置	输入分辨率：CIFAR-10, CIFAR-100, SVHN为32×32, Tiny-ImageNet为64×64 训练批量：128、64、32 训练批次：150 学习率：0.01在第75和90批次分别减小10倍 优化方法：SGD(Stochastic gradient descent)优化器
框架和设备	PyTorch + NVIDIA GeForce 1080 and 2080 TI GPU

实验结果

CIFAR-10, CIFAR-100, SVHN, Tiny-ImageNet数据集在ResNet-18网络的分类结果

ResNet-18	CIFAR-10										CIFAR-100									
Method	Clean	FGSM	PGD-20	PGD-100	CW_2	DDN_2	AA	STA	FWA	TI-DIM	Clean	FGSM	PGD-20	PGD-100	CW_2	DDN_2	AA	STA	FWA	TI-DIM
PGD-AT [25]	84.54	55.11	48.91	47.7	59.15	19.15	43.3	0.35	3.19	49.23	57.77	28.68	25.52	24.99	30.42	11.03	21.21	0.03	2.55	25.79
TRADES [56]	83.22	58.51	54.97	54.07	71.62	24.24	48.97	1.09	5.38	55.14	53.93	30.22	28.06	27.77	35.35	14.16	23.01	0	5.25	28.25
MART [43]	82.14	59.57	55.39	54.8	74.3	25.83	47.84	2.06	6.43	56.24	55.52	30.83	28.38	28.16	35.57	13.78	23.01	0.02	3.44	**28.51**
SOAP [34]	84.07	51.02	51.42	-	73.95	-	-	-	-	-	52.91	22.93	27.55	-	50.26	-	-	-	-	-
TPAP(Ours)	**86.25**	**61.41**	**79.06**	**80.5**	61.37	64.5	**76.34**	31.4	**52.83**	**75.21**	57.43	**35.64**	44.69	**42.23**	48.7	50.84	**47.48**	43.8	**32.23**	27.84
TPAP+TRADES	84.07	44.16	73.02	66.12	90.87	87.29	74.94	80.38	**51.34**	31.52	57.67	27.71	37.82	32.93	**70.49**	**65.62**	35.23	**66.92**	27.06	15.41
TPAP+MART	84.06	43.6	73.69	69.78	**92.38**	**90.05**	72.11	**85.7**	46.69	23.25	**61.03**	32.6	**44.9**	39.49	68.61	61.38	46.19	66.39	28.45	15.66

ResNet-18	SVHN										Tiny-ImageNet									
Method	Clean	FGSM	PGD-20	PGD-100	CW_2	DDN_2	AA	STA	FWA	TI-DIM	Clean	FGSM	PGD-20	PGD-100	CW_2	DDN_2	AA	STA	FWA	TI-DIM
PGD-AT [25]	91.66	**87.93**	63.86	44.57	72.65	8.84	31.35	6.51	10.1	**68.19**	**49.06**	24.26	22.09	21.44	28	30.41	16.82	0.21	0.99	22.06
TRADES [56]	91.32	73.38	59.01	56.31	72.96	5.19	47.03	1.57	0.31	59.18	46.59	22.9	21.46	21.03	28.87	28.8	15.99	0	1.8	21.54
MART [43]	91.81	75.31	56.55	51.28	71.45	6.96	42.08	0.9	0.86	56.68	46.21	25.73	24.16	23.59	30.58	30.24	17.85	0.34	1.75	24.23
TPAP(Ours)	89.62	67.56	**83.62**	**85.25**	51.39	62.07	**88.76**	55.12	**60.56**	40.99	48.72	**46.6**	37.88	**36.87**	31.48	45.28	**39.8**	12.43	**38.31**	**32.93**
TPAP+TRADES	91.36	41.22	80.31	72.77	92.14	**88.96**	66.57	88.31	28.67	59.64	46.22	8.96	17.93	14.07	**48.54**	**47.95**	20.5	**42.48**	5.4	5.5
TPAP+MART	**93.74**	26.92	81.62	66.31	**93.28**	88.36	63.26	**90.06**	28.72	26.65	48.88	18.46	**38.79**	36.53	41.23	43.42	31.72	26.81	36.49	26.93

TPAP在大多数干净样本和多种攻击场景下取得最好的鲁棒性能

更多实验

① 验证在大尺寸图象上的有效性

CALTECH-101. (图像尺寸：300×200)

Method	Clean	FGSM	PGD-20
PGD-AT*	72.44	65.55	57.74
TPAP*	70.2	55.93	59.16
TPAP-TRADES*	76.11	60.8	70.91
TPAP-MART*	69.27	58.39	60.63

② 与更多图像预处理防御方法的对比

CIFAR-10.

Method	Clean	PGD	Architecture
(Yang et al., 2019)(p:0.4→0.6) [53]	84	68.2	ResNet-18
(Hill et al., 2021) [11]	84.12	78.91	WRN-28-10
(Wang et al., 2023) [44]	92.58	68.43	WRN-28-10
TPAP	86.25	79.06	ResNet-18

③ 计算开销和时间

ResNet-18	CIFAR-10(ϵ=8/255, Batch size = 128)				
Method	Training time (s)/epoch	Test time (s)/epoch	Training FLOPs (T)/epoch	Test FLOPs (T)/epoch	Params (M)
PGD-AT	273.89	2.02	1831.5	11.1	11.17
TPAP	66.26	8.26	333	44.4	11.17

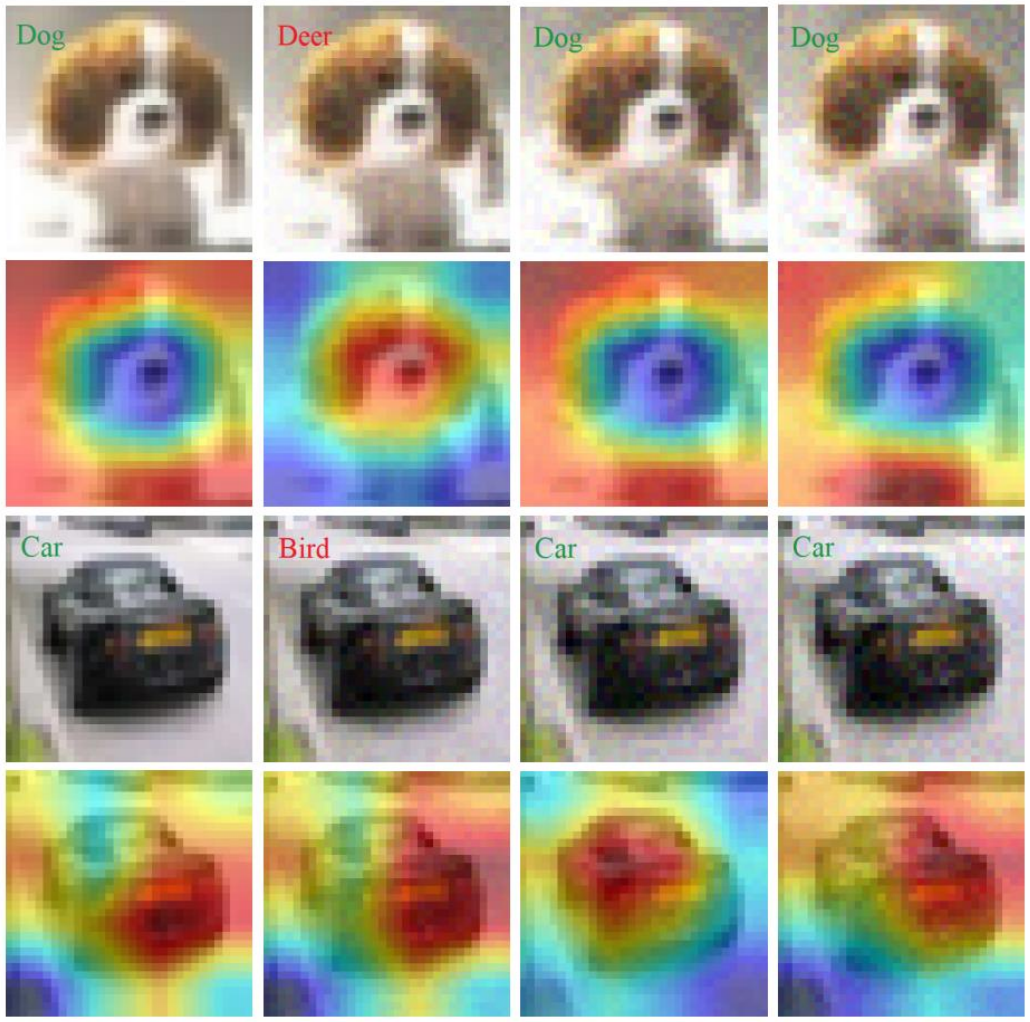
④ TPAP消融实验

ResNet-18		CIFAR-10(ϵ=8/255,bs=128)		
Purification	RO-FGSM-DNN	Clean	FGSM	PGD-20
×	✓	86.33	94.41	0.18
✓	✓	86.25	61.41	79.06

⑤ 攻击方法替代消融实验

ResNet-18	CIFAR-10(ϵ=8/255,bs=128)				
Training	Clean	CW2	PGD-20	FGSM	AA
CW2	68.19	84.7	37.4	39.95	56.66
PGD-10	56.17	75.62	42.74	44.5	82
FGSM	86.25	61.37	79.06	61.41	76.34

注意力可视化实验

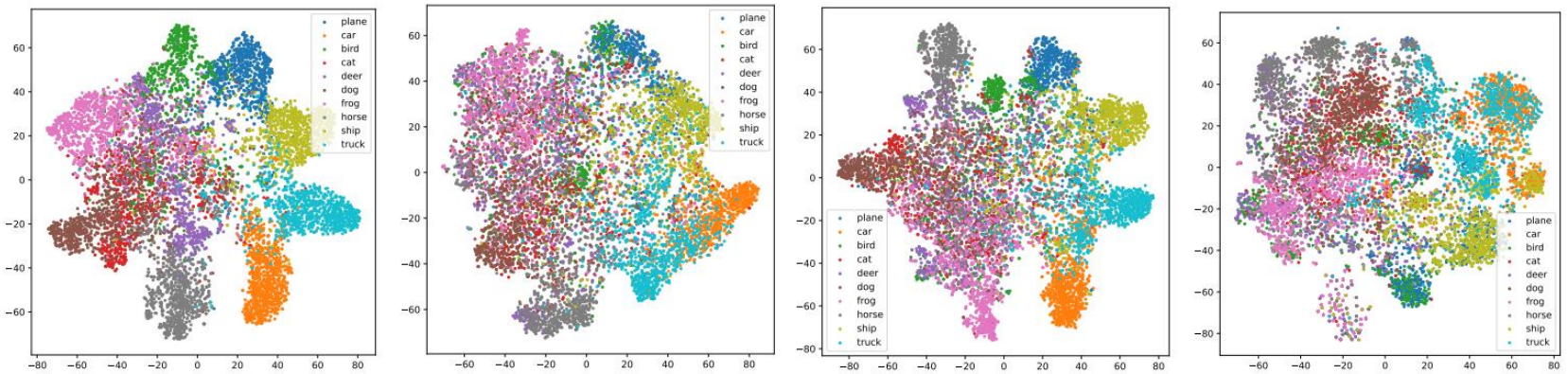


(a) x_c (b) x_{pgd20} (c) x_{c_pur} (d) x_{pgd20_pur}

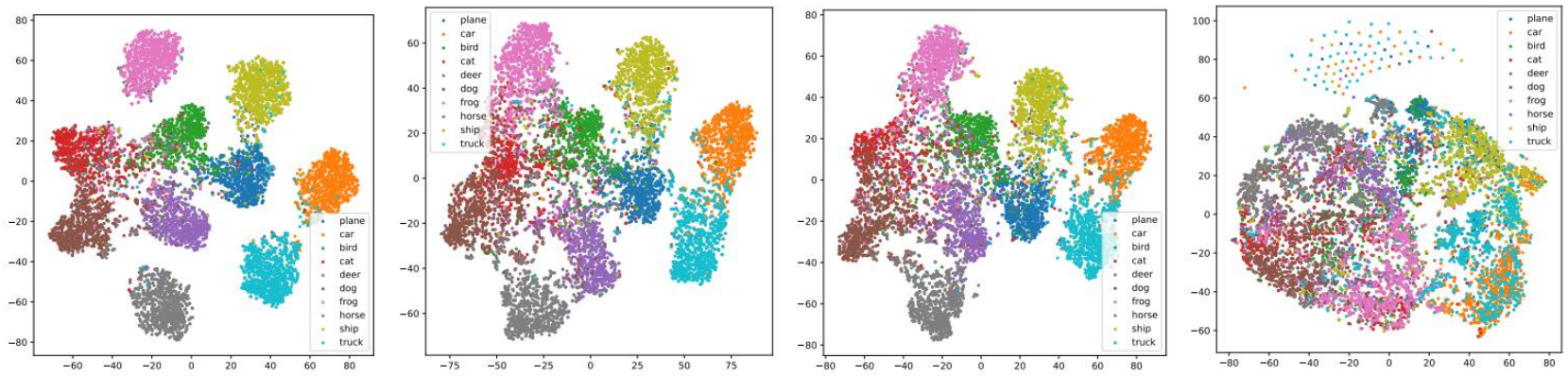
(a)(b)列分别是没有经过对抗性净化的干净样本和对抗样本； (c)(d)列分别是经过了对抗性净化的干净样本和对抗样本。结论：通过对抗性净化的对抗样本能够关注到分类正确的注意力区域。

特征可视化实验

PGD-AT



TPAP



(a) x_c (b) x_{pgd20} (c) x_{AA} (d) x_{STA}

结论：TPAP 可以从对抗性分布中恢复出具有较好特征聚类辨别性的健康分布。
（即类间可分性和类内紧凑性）

Learning Intelligence
& Vision Essential
(LiVE) Group

CVPR SEATTLE, WA JUNE 17-21, 2024

Robust Overfitting Does Matter: Test-Time Adversarial Purification With FGSM

文章

[2403.11448] Robust Overfitting Does Matter: Test-Time Adversarial Purification With FGSM (arxiv.org)

代码

tly18/TPAP (github.com)

邮箱：linyutang@cqu.edu.cn