南京大学
NANJING UNIVERSITY

RINC

基于像素剪枝的稀疏对抗攻击及其在图像数据增强中的应用

- 答辩人：李金桥 MG1933035
- 导　师：申富饶 教授

南京大学
NANJING UNIVERSITY

目录
CONTENTS

南京大学
NANJING UNIVERSITY

第一部分

Research Background

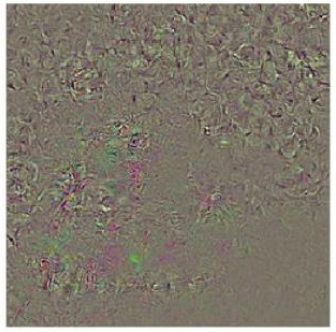
研究背景

背景简介 | 研究意义 | 困难与挑战

背景简介——对抗攻击

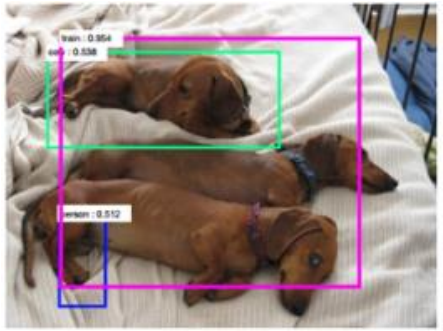
- 向原始样本中添加**近似不可察觉**的扰动，欺骗深度学习模型，使其产生**错误的结果**。


原始样本--帆船


10X 对抗扰动


对抗样本--iPad

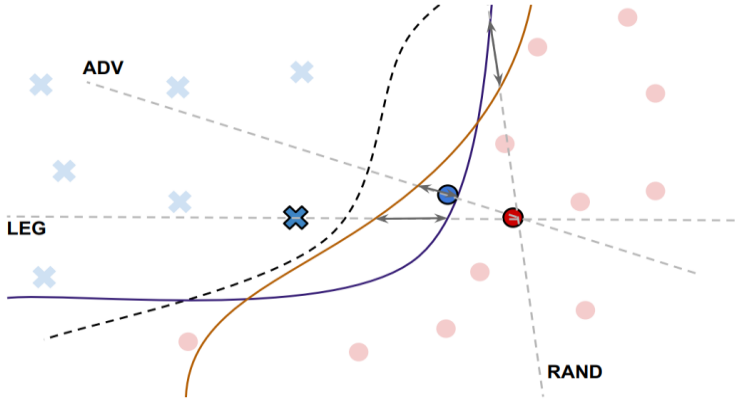
- 众多任务中都存在对抗样本——对抗攻击已成为AI面临的安全问题之一。


目标检测

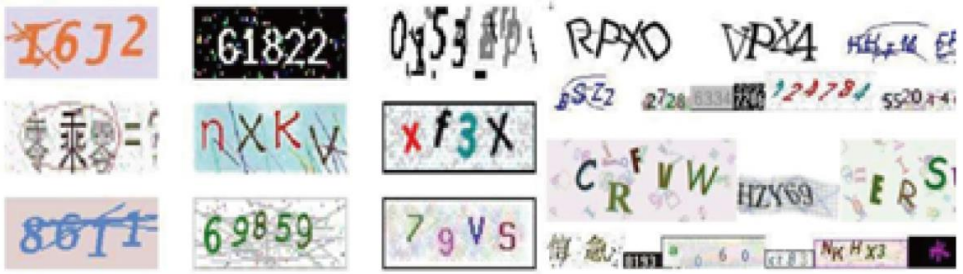
交通标志识别


现实场景人脸识别

研究对抗攻击的意义



助力模型可解释性研究



编码隐式信息

图像隐私保护

图像数据增强

困难与挑战

攻击成功率

- 无目标：1 – 模型准确率
- 有目标：目标达成率

扰动的L_p范数

- 2范数：变化前后样本的欧式距离
- 无穷范数：任意像素点最大变化幅度
- 0范数：变化的像素点的个数

好的方法：同时保持高攻击性和低可见性。

南京大学
NANJING UNIVERSITY

第二部分

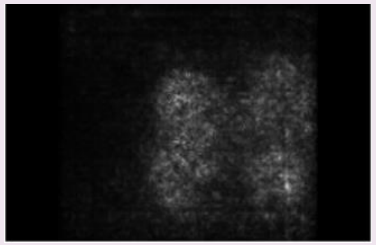
Research Content

研究内容

基于像素剪枝的稀疏对抗攻击 | 基于稀疏对抗攻击的图像数据增强

基于像素剪枝的稀疏对抗攻击：研究动机

显著图（Saliency Maps）



部分像素对分类起决定性作用；而大多现有对抗攻击都进行全图性的扰动，包含**大量冗余**。

现有稀疏对抗攻击（某限制下最小化 L_0范数）

人为定义的重要性指标难以适应所有情况；
贪心策略容易陷入局部最优。

- 普通稠密对抗攻击包含大量冗余，可去掉部分像素点的扰动从而**降低扰动量**；
- 能否**跳出固有的稀疏对抗攻击模式**？避免人为定义重要性和使用贪心策略。

基于像素剪枝的稀疏对抗攻击：网络剪枝与对抗攻击

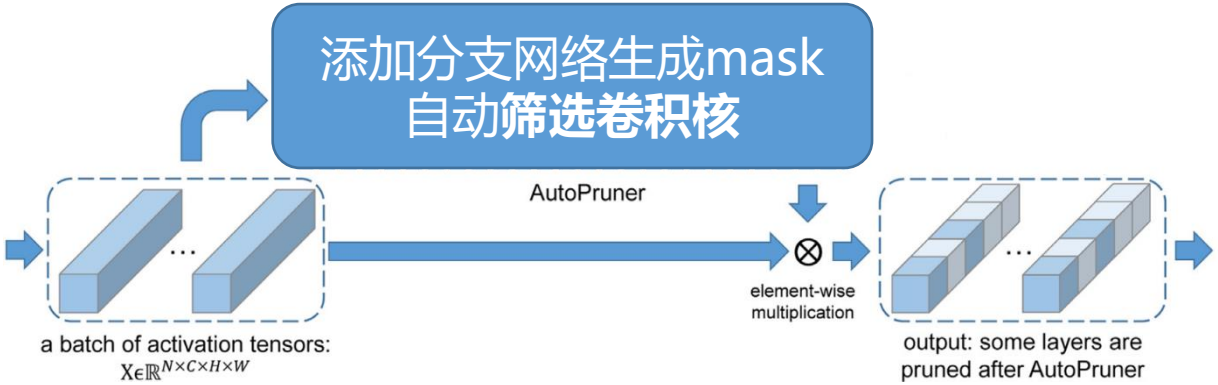
神经网络剪枝

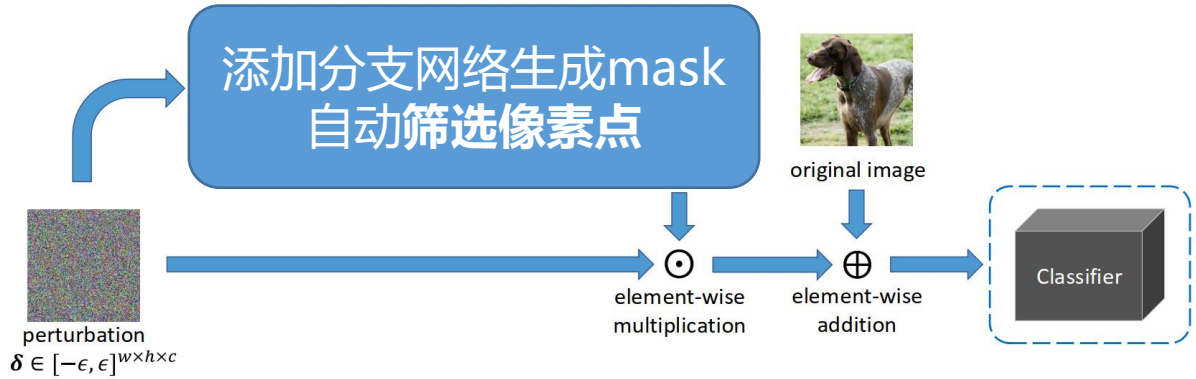
- 保持模型精度，尽可能减少卷积核；
- 可看作L_0优化问题。
- 早期方法人为定义卷积核的重要性，并在过程中逐渐删减不重要的卷积核。

相似性

稀疏对抗攻击

- 保持攻击成功率，尽可能减少扰动的像素点；
- 可看作L_0优化问题。
- 现有方法人为定义像素点的重要性，并在过程中逐渐扰动重要的像素点。





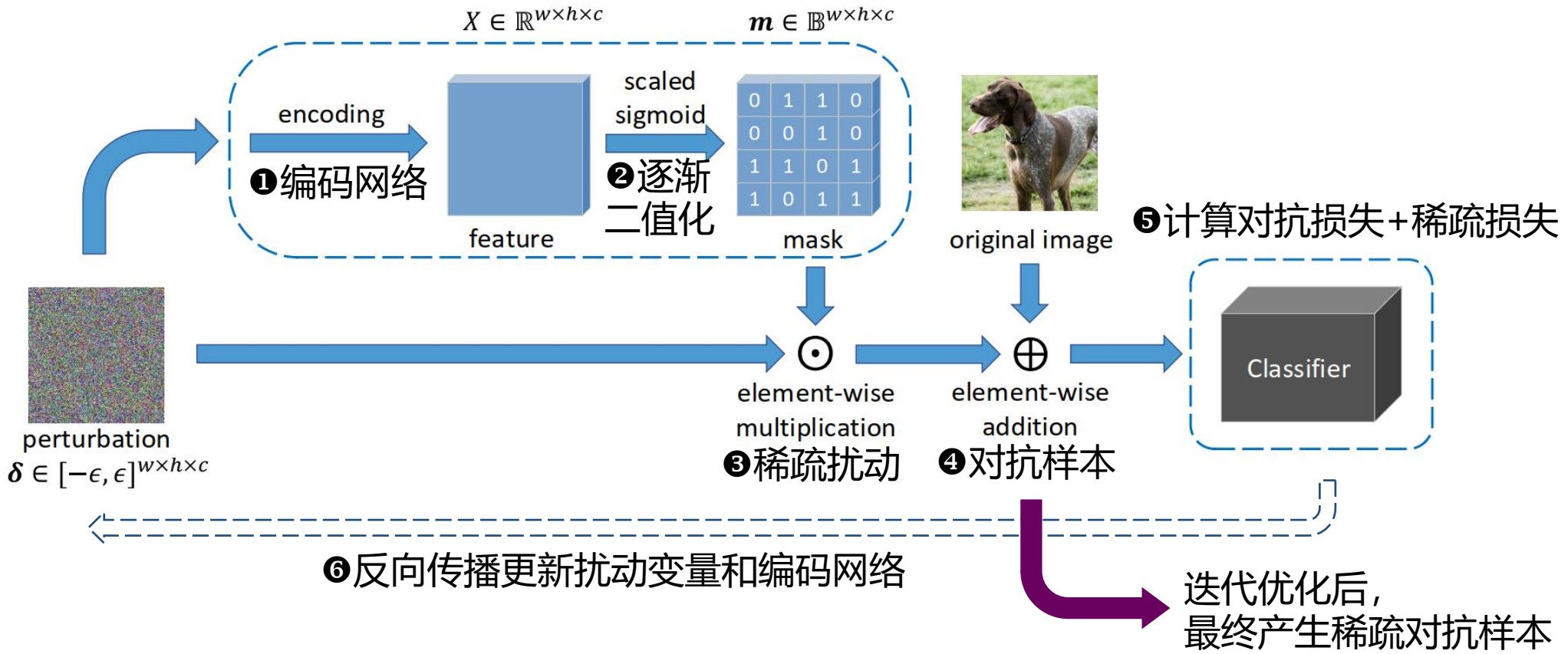
- **自动剪枝**可以绕开人为定义重要性和贪心过程，端到端的优化方法获得更好效果。

启发

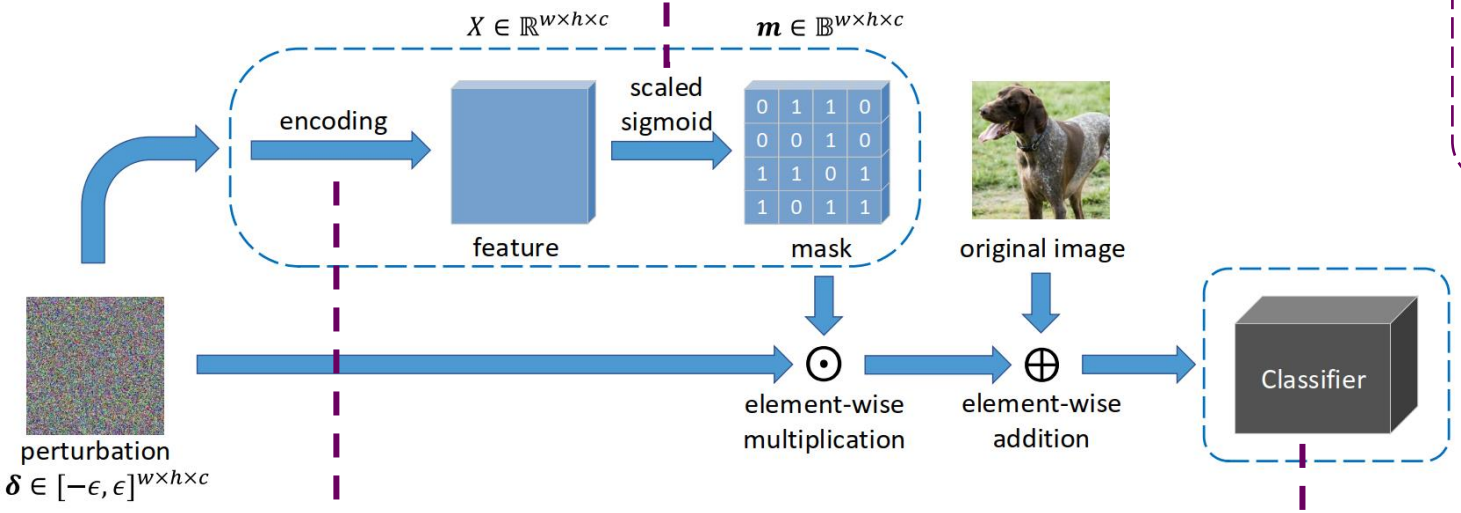
- 可基于自动剪枝的思路，同样对扰动的像素进行自动剪枝。

基于像素剪枝的稀疏对抗攻击：整体结构

神经网络剪枝：剪掉卷积核 → 对抗扰动剪枝：剪掉扰动像素点



基于像素剪枝的稀疏对抗攻击：关键模块

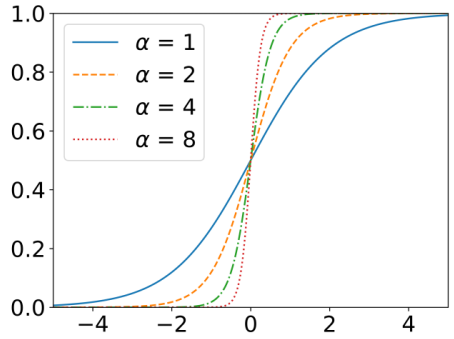


(1) 编码器网络

- 提取扰动中的特征，并保持输入输出尺寸相同；
- 根据图像尺寸大小，可使用FC和U-Net等结构。

(2) 二值化模块

- 为保持连续可导，使用$\mathrm{sigmoid}(\alpha \cdot \boldsymbol{x})$；
- α在优化过程中逐渐增大，迫使mask二值化。



(3) 损失函数

$$L = L_{ce}(f(\boldsymbol{x} + \boldsymbol{\delta} \odot \boldsymbol{m}), y^*) + \lambda \frac{\|\boldsymbol{m}\|_1}{N}$$

攻击效果尽量好　　扰动尽量稀疏

- $\boldsymbol{m}$的元素近似2值；
- $\|\boldsymbol{m}\|_1$近似表示1的个数，$\frac{\|\boldsymbol{m}\|_1}{N}$表示稀疏性

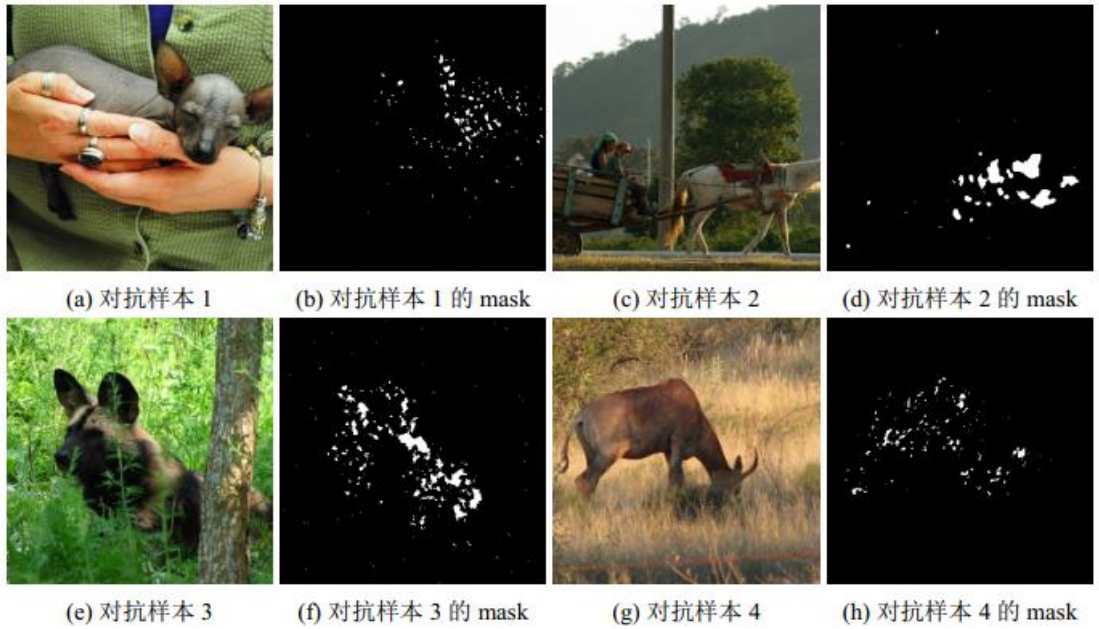
基于像素剪枝的稀疏对抗攻击：有效性实验

表 3-2: ImageNet 上不同最大扰动幅度下的像素剪枝结果

Threshold	Method	ASR(%)	l_0	l_1	l_2	l_∞
$\epsilon = 8/255$	I-FGSM	100.0	240118.5	4744.600	10.595	0.031
	I-FGSM*	**100.0**	**10592.4**	**273.389**	**2.705**	**0.031**
	PGD	100.0	241981.2	4775.905	10.629	0.031
	PGD*	**100.0**	**8340.8**	**229.587**	**2.513**	**0.031**
	MI-FGSM	100.0	254325.8	6610.318	13.753	0.031
	MI-FGSM*	**100.0**	**5442.9**	**164.851**	**2.183**	**0.031**
$\epsilon = 4/255$	I-FGSM	100.0	219698.3	2616.715	5.880	0.016
	I-FGSM*	**100.0**	**26547.7**	**351.169**	**2.187**	**0.016**
	PGD	100.0	221070.6	2624.191	5.885	0.016
	PGD*	**100.0**	**27758.1**	**372.702**	**2.261**	**0.016**
	MI-FGSM	100.0	248838.5	3566.601	7.303	0.016
	MI-FGSM*	**100.0**	**12900.6**	**198.117**	**1.696**	**0.016**

* 表示以该方法作为基础对抗攻击，并利用 AutoAdversary 进行像素剪枝减小扰动。

- 扰动像素点个数即l_0范数大大减小，**扰动量大大减小**。
- 保持**100%攻击成功率**不变；
- 任意像素的最大扰动幅度即l_∞范数不变（**保证约束**）



(a) 对抗样本 1 (b) 对抗样本 1 的 mask (c) 对抗样本 2 (d) 对抗样本 2 的 mask

(e) 对抗样本 3 (f) 对抗样本 3 的 mask (g) 对抗样本 4 (h) 对抗样本 4 的 mask

基于像素剪枝的稀疏对抗攻击：对比实验

表 3-3: CIFAR-10 上各稀疏对抗攻击的对比（预设最大扰动幅度）

Threshold	Method	ASR(%)	l_0	l_1	l_2	l_∞
$\epsilon = 8/255$	JSMA	96.5	327.7	10.212	0.517	0.031
	PGD-$l_0 + l_\infty$	74.8	1499.5	45.825	1.193	0.031
	GreedyFool	100.0	420.4	8.637	0.429	0.031
	AutoAdversary	**100.0**	**173.3**	**5.412**	**0.393**	**0.031**
$\epsilon = 16/255$	JSMA	98.9	182.1	11.232	0.769	0.063
	PGD-$l_0 + l_\infty$	84.7	1199.7	71.782	2.104	0.063
	GreedyFool	100.0	274.8	9.727	0.611	0.063
	AutoAdversary	**100.0**	**87.2**	**5.396**	**0.555**	**0.063**

表 3-4: CIFAR-10 上各稀疏对抗攻击的对比（参考平均最大扰动幅度）

Threshold	Method	ASR(%)	l_0	l_1	l_2	l_∞
$\epsilon = 0.045$	C&W-l_0	100.0	383.1	**5.988**	**0.345**	0.045
	AutoAdversary	**100.0**	**119.3**	5.323	0.467	**0.045**
$\epsilon = 0.070$	C&W-l_0	100.0	194.2	**5.036**	**0.406**	0.070
	AutoAdversary	**100.0**	**78.6**	5.411	0.587	**0.070**
$\epsilon = 0.055$	SAPF	100.0	400.1	**4.720**	**0.292**	0.055
	AutoAdversary	**100.0**	**98.9**	5.379	0.518	**0.055**
$\epsilon = 0.088$	SAPF	100.0	201.8	**4.495**	**0.390**	0.088
	AutoAdversary	**100.0**	**64.9**	5.585	0.668	**0.088**

- 预设最大扰动幅度时，我们的方法在**所有评价指标上都最好；**
- 参考平均最大扰动幅度时，我们的方法稀疏性更强，能**找到最需要被扰动的像素点。**

基于像素剪枝的稀疏对抗攻击：消融实验

将各关键模块拆分，证明了本方法**各个模块的有效性和必要性。**

表 3-7: AutoAdversary 各模块消融实验

Method	ASR(%)	l_0	l_1	l_2	l_∞
❶ Dense	100.0	2965.0	82.215	1.559	0.031
❷ Dense + Random	6.2	782.8	23.059	0.839	0.031
❸ Dense + l_1-$\boldsymbol{\delta}$	99.5	2422.6	29.641	0.703	0.031
❹ Dense + l_1-$\boldsymbol{m}$ + Binarization	98.4	256.8	7.987	0.478	0.031
❺ Dense + l_1-$\boldsymbol{m}$ + Binarization + Encoder	**100.0**	**173.3**	**5.412**	**0.393**	**0.031**

❶ 普通密集对抗攻击，不稀疏；

❷ 随机选择像素点扰动ASR仅为6.2%，说明了我们的方法**并非随机选择像素点**；

❸ 直接使用扰动变量的l_1正则，稀疏性较差，说明了所添加**分支网络的有效性**；

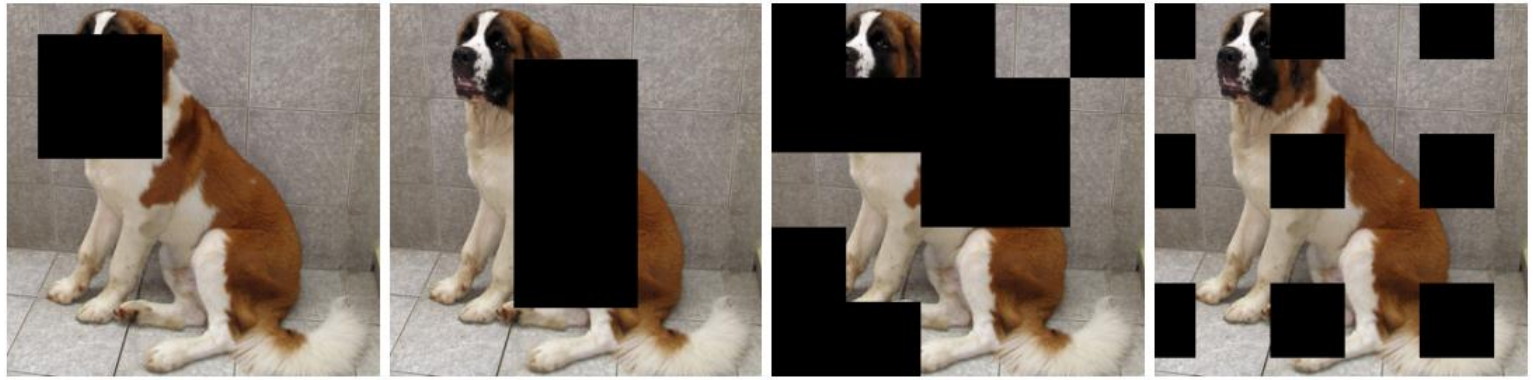
❹ 绕过编码器网络直接二值化，ASR和稀疏性都更差，说明了**编码器的必要性**；

❺ 全部模块合起来，**效果最佳。**

基于稀疏对抗攻击的图像数据增强：研究动机

现有基于信息删除（Information Dropping）的图像数据增强

- 训练过程中**随机删除**图像中的部分区域；
- 迫使模型不依赖局部特定特征，更好利用全图上下文，缓解过拟合。



(a) Cutout (b) RandomEarsing (c) HaS (d) GridMask

完全随机的区域可能生成低质量样本，可否基于图像个性化处理?

稀疏对抗攻击可找到模型过于关注的敏感像素。

以稀疏对抗攻击作为前置，配以信息删除的数据增强模式。

基于稀疏对抗攻击的图像数据增强：整体结构

❶原始数据训练得到**基准模型**

❷稀疏对抗攻击产生每张图像的**敏感像素**

❸基于敏感像素mask进一步生成**删除区域**，为增加多样性每一轮有所不同。

基于稀疏对抗攻击的图像数据增强：关键模块

稀疏对抗攻击得到敏感像素mask

$$L = -L_{ce}(f(\boldsymbol{x} + \boldsymbol{\delta} \odot \boldsymbol{m}), y_t) + \lambda \frac{\|\boldsymbol{m}\|_1}{N}$$

$[-\infty, 0]$ 数值差异 $[0,1]$

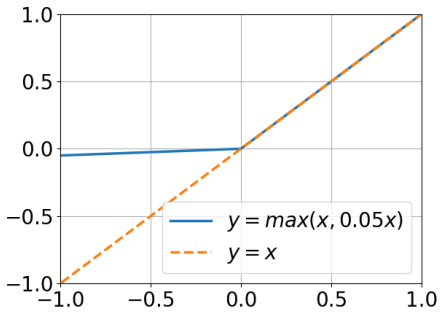
预先的非稀疏攻击，记录模型出错时的损失为L_{init}

$$L_{classify} = -\frac{L_{ce}(f(\boldsymbol{x} + \boldsymbol{\delta} \odot \boldsymbol{m}), y_t)}{L_{init}} + 1$$

$$L = \max\left(L_{classify}, \eta \cdot L_{classify}\right) + \lambda \frac{\|\boldsymbol{m}\|_1}{N}$$

$[-\infty, 1]$但是小于0会迅速衰减，因此基本处于$[0,1]$，**保持了数值平衡**

$L_{classify} = 0$时，表明模型已出错



每轮迭代中基于敏感像素mask得到删除区域

- 删除尺寸过小的区域对CNN不起作用；
- 需将孤立的敏感像素点**扩充为若干连续删除区域。**

随机区域边长 → 随机选点并扩充 → 小范围随机平移 → 删除原图对应区域作为本轮样本

$d = \text{random}(d_l, d_h)$

若当前总面积$S < S_h$

基于稀疏对抗攻击的图像数据增强：对比实验

表 4-1: CIFAR-10 上应用不同数据增强方法得到的模型准确率

Model	Base	Cutout	RE	HaS	GridMask	AdvMask
ResNet-18	95.28	96.01	95.69	96.10	96.34	**96.44**
ResNet-44	94.10	94.78	94.87	94.97	95.02	**95.49**
ResNet-50	95.66	95.81	95.82	95.94	96.15	**96.69**
WideResNet-28-10	95.52	96.92	96.92	96.94	**97.23**	97.03
ShakeShake-26-32	94.90	96.96	96.46	96.89	96.91	**97.02**

表 4-2: CIFAR-100 上应用不同数据增强方法得到的模型准确率

Model	Base	Cutout	RE	HaS	GridMask	AdvMask
ResNet-18	77.54	78.04	75.97	78.19	75.23	**78.29**
ResNet-44	74.80	74.84	75.71	75.82	76.07	**76.44**
ResNet-50	77.41	78.62	77.79	78.76	78.38	**78.99**
WideResNet-28-10	78.96	79.84	80.70	80.22	80.40	**80.70**
ShakeShake-26-32	76.65	77.37	77.30	76.89	77.28	**79.96**

表 4-3: Tiny-ImageNet 上应用不同数据增强方法得到的模型准确率

Model	Base	Cutout	RE	HaS	GridMask	AdvMask
ResNet-18	62.00	63.59	63.80	63.61	63.50	**65.12**
ResNet-50	73.34	77.86	75.08	74.94	77.38	**80.20**
WideResNet-50-2	81.55	81.77	81.89	81.84	81.79	**82.85**

在多个模型和多个数据集上，我们的方法AdvMask基本都取得了**最优的测试精度。**

基于稀疏对抗攻击的图像数据增强：消融实验与可视化实验

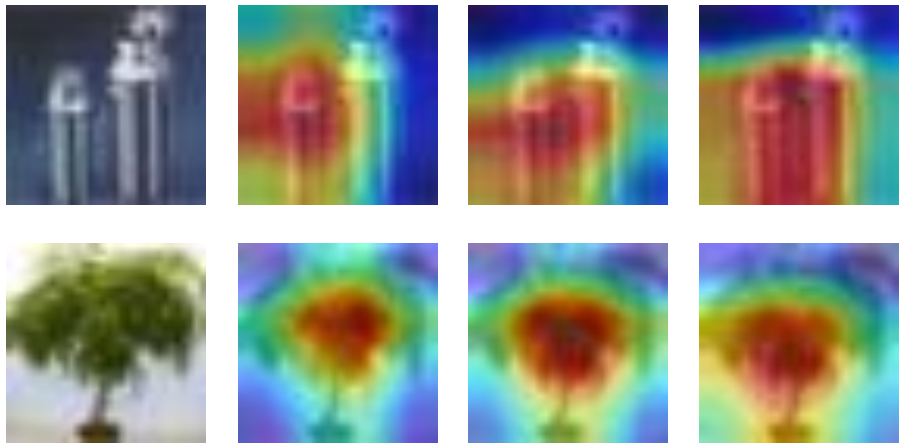
表 4-4: CIFAR-10 上有关 AdvMask 的消融实验

Model	Method	Accuracy(%)
ResNet-18	Base	95.28
	Adv	92.96
	Random + Mask	95.66
	Corner + Mask	95.11
	Adv + Mask	**96.44**
ResNet-50	Base	95.66
	Adv	94.52
	Random + Mask	95.61
	Corner + Mask	95.41
	Adv + Mask	**96.69**

方法拆为“Adv”和“Mask”两个部分，分别验证其有用性。

- Adv跳过第二部分，直接用敏感像素作为删除区域，Acc较低，说明了**第二部分的有用性。**
- Random+Mask随机选点作为敏感像素点；Corner+Mask使用角点检测得到敏感像素点，Acc均较低；说明了**第一部分的有用性。**

类激活图CAM可视化



输入 Base GridMask AdvMask

- 我们的增强方法能够关注**更加全面的信息。**

南京大学
NANJING UNIVERSITY

第三部分

Applications

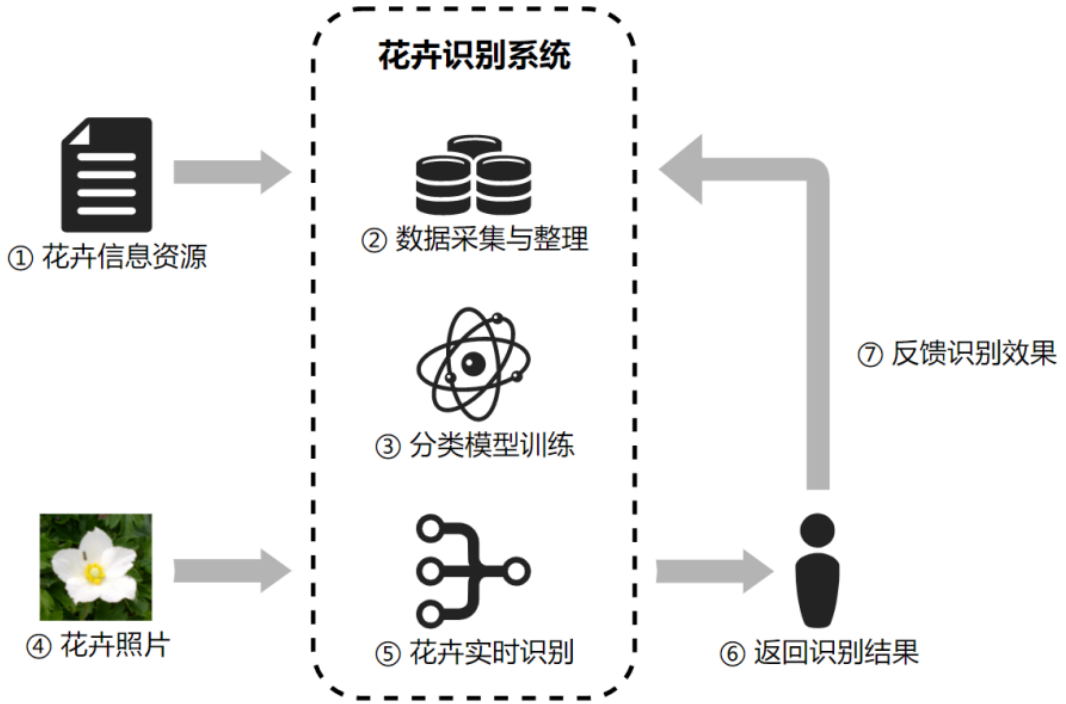
实际应用

花卉识别系统

花卉识别系统：系统简述

系统需求：

- 作为花卉识别工具
  - 简易方便的使用条件（移动端）
  - 实时准确的花卉识别
  - 错误结果反馈
- 作为科普平台
  - 科普知识推送
  - 趣味答题活动



难点：不同花卉的**外表特征非常相似**，模型过于关注局部特征从而误识别。

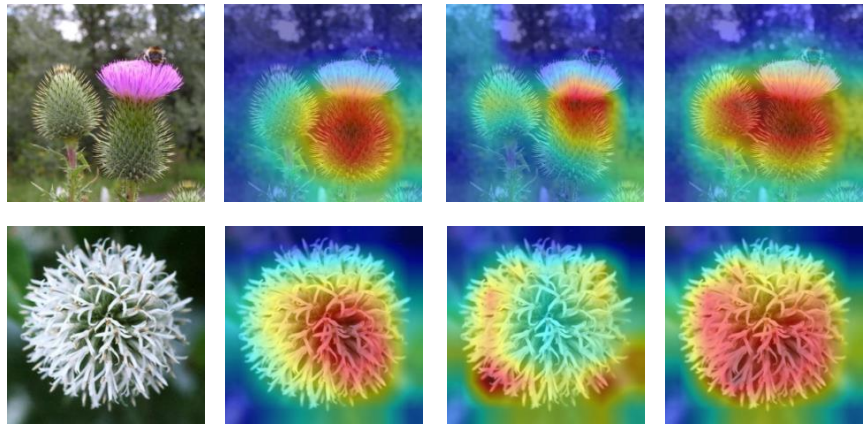
关键技术：利用上述提出的基于稀疏对抗攻击的**数据增强方法**来训练。

花卉识别系统：系统功能与性能

表 5-1: 不同数据增强方法在花卉识别系统中的对比

Method	trainAcc(%)	valAcc(%)	testAcc(%)
Base	99.98	83.92	80.20
Cutout	97.98	87.75	88.53
HaS	92.52	85.98	84.12
RE	97.73	88.73	88.43
GridMask	97.95	90.69	90.19
AdvMask (ours)	98.95	90.78	**91.57**

类激活图CAM可视化



输入 Base GridMask AdvMask

①主界面

②识别界面

③文章科普

④趣味问答

- 取得最优的测试精度和泛化性；
- 能够关注图中更加全面的信息。

➡ 表明了方法的**实际应用价值。**

南京大学
NANJING UNIVERSITY

第四部分

Summary

研究生期间工作成果

相关成果

相关成果

论文

- Jinqiao Li, Xiaotao Liu, Jian Zhao, Furao Shen, “AutoAdversary: A Pixel Pruning Method for Sparse Adversarial Attack.” in arXiv preprint arXiv:2203.09756, 2022.

专利

- 申富饶，李金桥，姜少魁，陆志浩，金祎。一种实时判定摄像头遮挡状态的方法。专利申请号：CN202010736809.6

项目

- 国家自然基金“基于深度感知增量式联想记忆神经网络的信息融合系统研究”

竞赛

- 微众银行第二届金融科技高校技术大赛Top3

南京大学
NANJING UNIVERSITY

第五部分

Summary

全文总结

全文总结

全文总结

基于像素剪枝的稀疏对抗攻击

- 神经网络自动剪枝与对抗攻击相结合
- 跳出固有的稀疏对抗攻击模式
- 设计同时优化攻击性和稀疏性的损失函数

基于稀疏对抗攻击的图像数据增强

- 利用稀疏对抗攻击寻找敏感的像素点
- 引入衰减函数维持损失函数的数值平衡
- 引入部分随机性提升样本的多样性

实时花卉识别系统

- 将所提出的数据增强算法应用其中
- 识别功能准确快速以及泛化性高
- 科普功能包含文章推送和趣味答题

南京大学
NANJING UNIVERSITY

谢谢大家！