Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

从“把图像搅乱”到“把噪声借来”
对抗样本攻防的两种奇思妙想

硕士研究生 罗天长笑

2025年11月16日

问题回溯

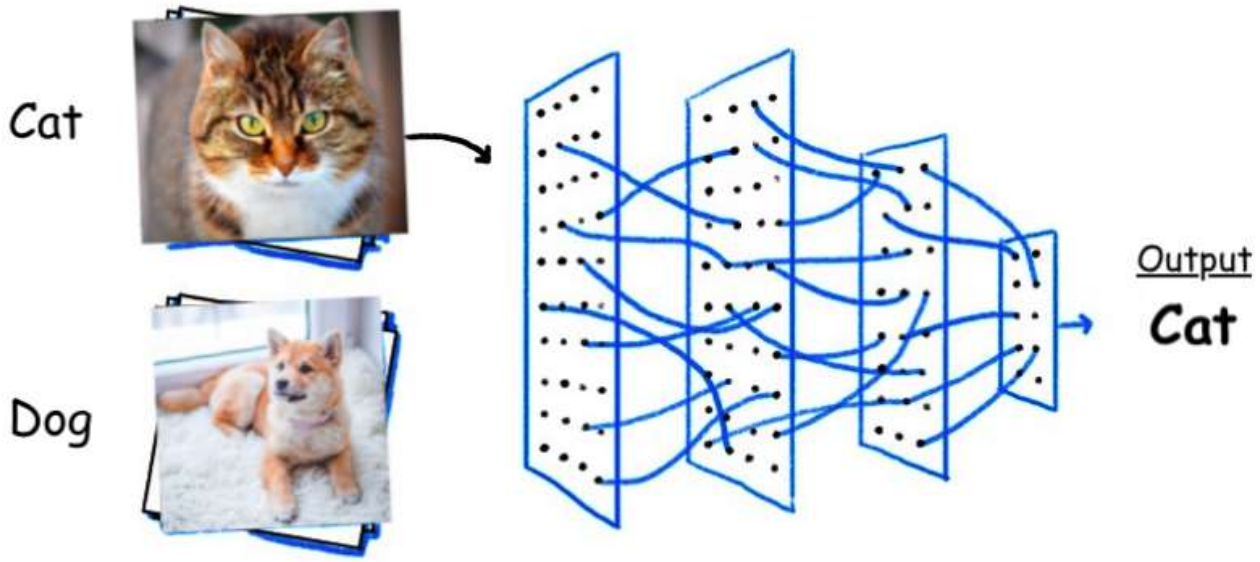
- 相关内容
  - 2024.11.07 郑俊怡《文本分类硬标签黑盒模型的对抗样本生成方法研究》
  - 2023.10.23 邵思源《面向NIDS的流量对抗样本检测》
  - 2023.05.28 程瑶《单词级文本对抗攻击》

内容提要

- 预期收获
- 题目内涵解析
- 案例引入
- 背景意义
- 知识基础
- 研究历史与现状
- 算法原理&实验流程
  - BSR
  - DDA
- 特点总结与工作展望
- 参考文献

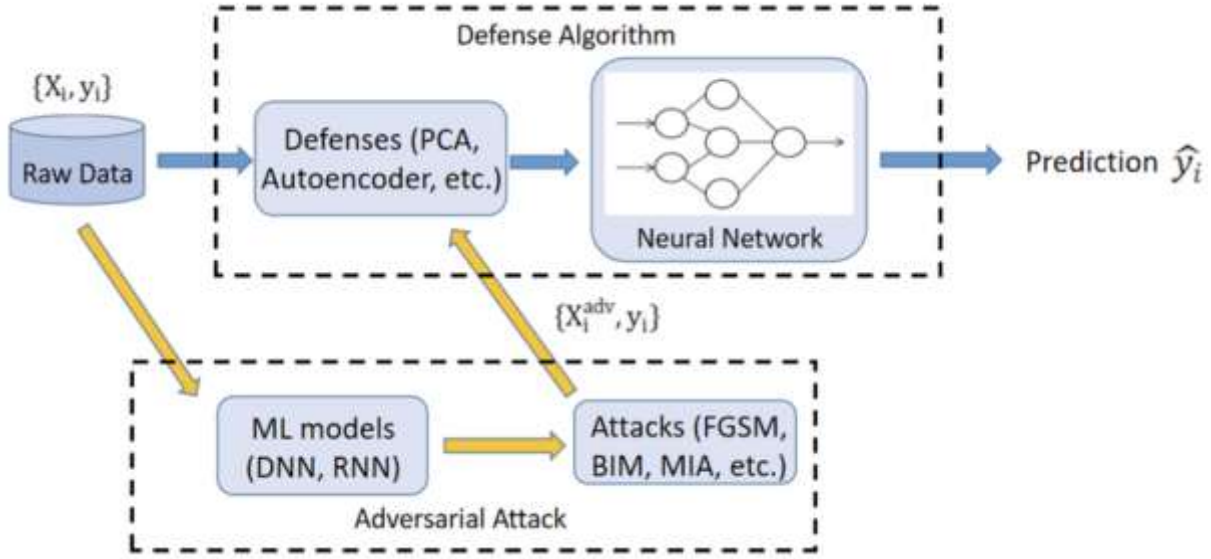
预期收获

- 了解什么是对抗样本
- 了解图像对抗样本攻击、防御的基本思想
- 了解一种基于分块重排与随机旋转的对抗样本生成方法
- 了解一种扰动空间数据增强的防御方法



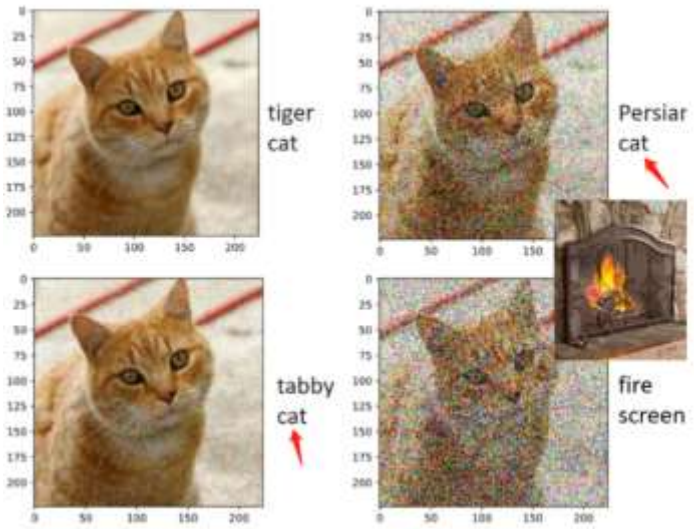
题目内涵解析

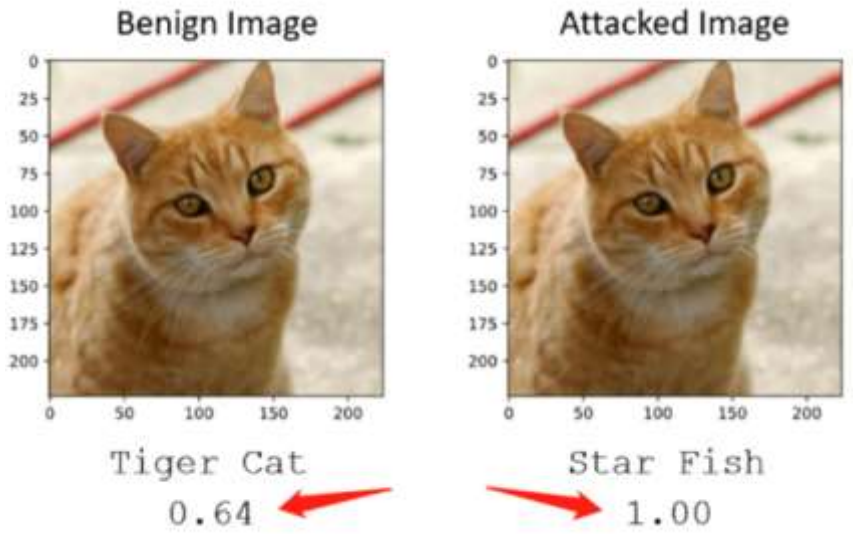
- 对抗样本攻防
  - 对抗样本攻击：分析决策边界的脆弱点，构造能跨模型迁移的稳准狠扰动，使模型在分类中误判
  - 对抗样本防御：提升模型面对攻击时的判断稳定性，减少对噪声的过度敏感，维持分类的正确性
  - 攻防：两者互相对立，又互相促进



案例引入

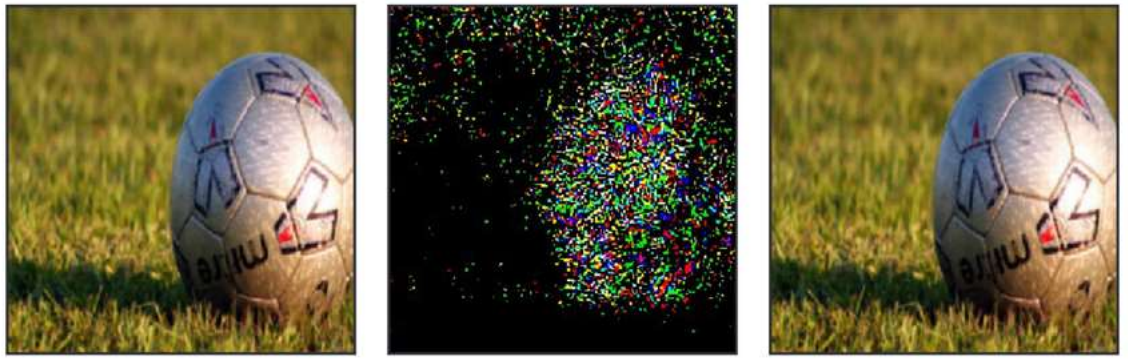
- 对抗样本
  - 假定有一个训练好的图像分类器，它能够分辨出来下图是一只猫
  - 在实验过程中对输入图像添加规律的，较大的（人眼可见）的扰动，此时模型大多数时候还能识别它为猫，只是与原标签不同品种；当添加一些特定的，细微的扰动时，模型识别为鱼，且置信度大幅增加
  - 这些添加了特定扰动的样本被称为对抗样本





背景意义

- 研究背景
  - 深度神经网络（DNN）在图像分类任务中取得了突破性进展（如 ResNet、VGG 在 CIFAR-10/Imagenet 上表现优异）
  - Szegedy 等人在 2014年提出：对输入图像添加微小的、不可感知的扰动，即可导致模型输出错误分类，这类图像被称为对抗样本（Adversarial Examples）
  - 应用领域：图像分类、目标检测、人脸识别、医疗影像分析等



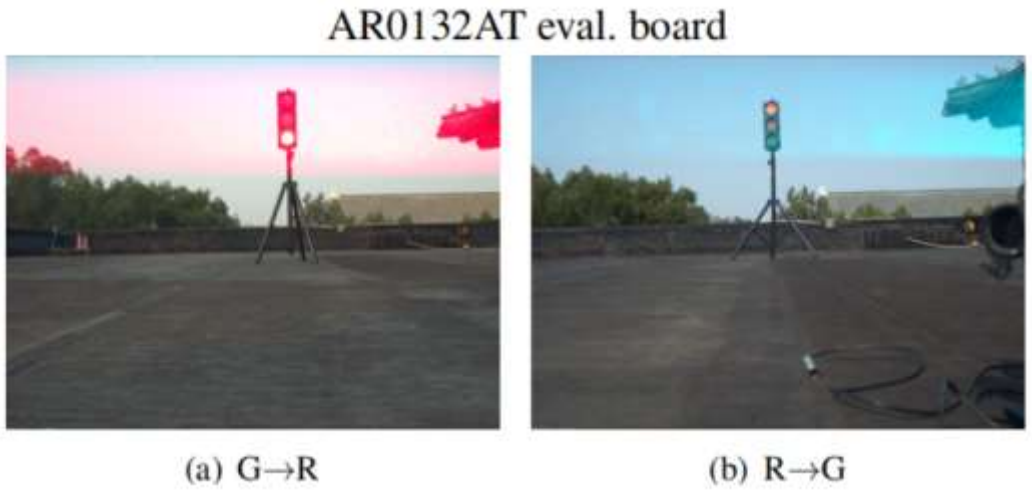
(a) 原始样本　(b) 扰动噪声　(c) 对抗样本

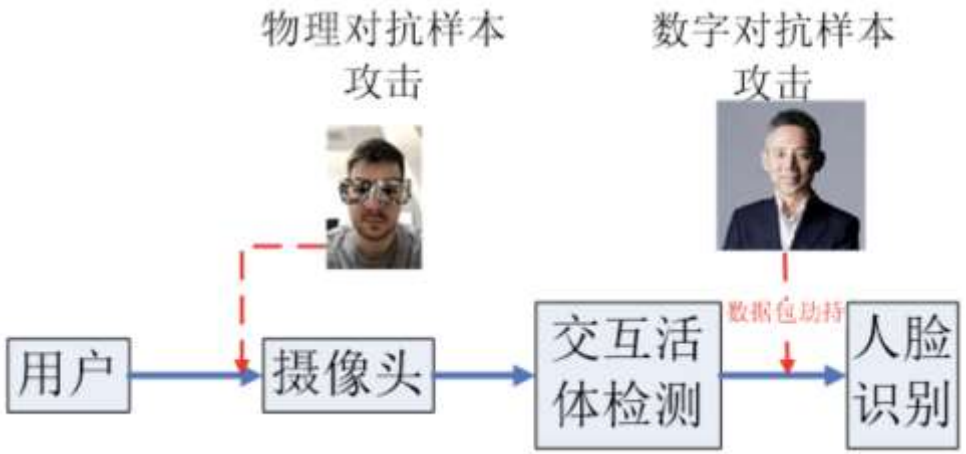
图 1　对抗样本示例

Fig. 1　Examples of adversarial sample

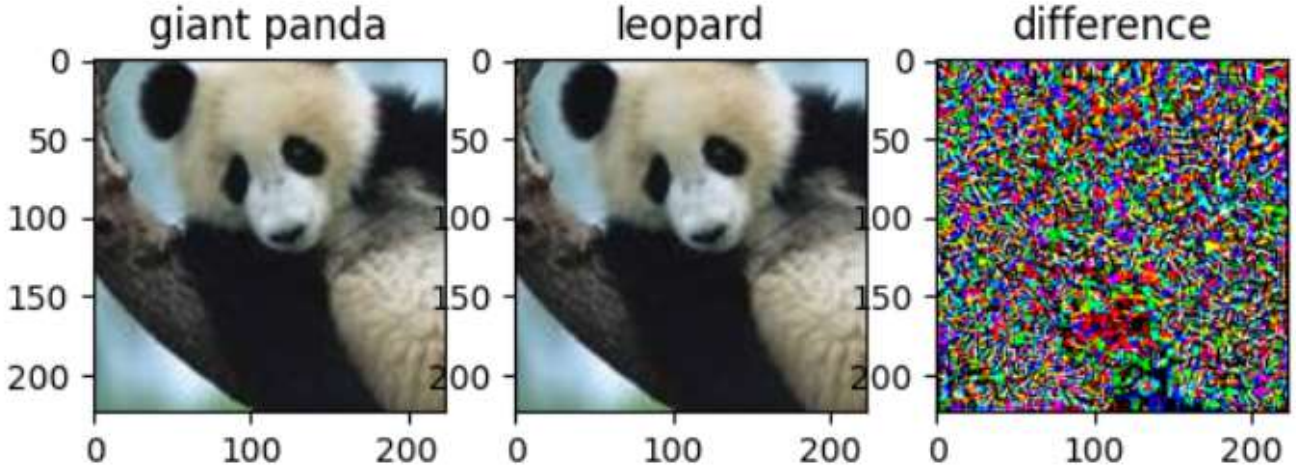
背景意义

- 研究意义
  - 对抗样本问题揭示了深度学习模型存在严重的安全漏洞，给深度学习技术的普遍应用带来了严峻的安全挑战
  - 对于对抗样本生成和攻击过程的研究能够为及时评估对抗攻击技术给深度学习模型带来的安全风险提供有益参考





- 对抗样本
  - 定义：对抗样本是在原始输入数据中加入人为构造的、微小的扰动，这些扰动虽然对人眼不可见或难以察觉，但可以欺骗深度学习模型，导致模型做出错误的预测或决策
  - 意义：揭示了深度学习模型在处理对抗样本时的脆弱性，暗示了潜在的安全风险，特别是在安全敏感的应用领域

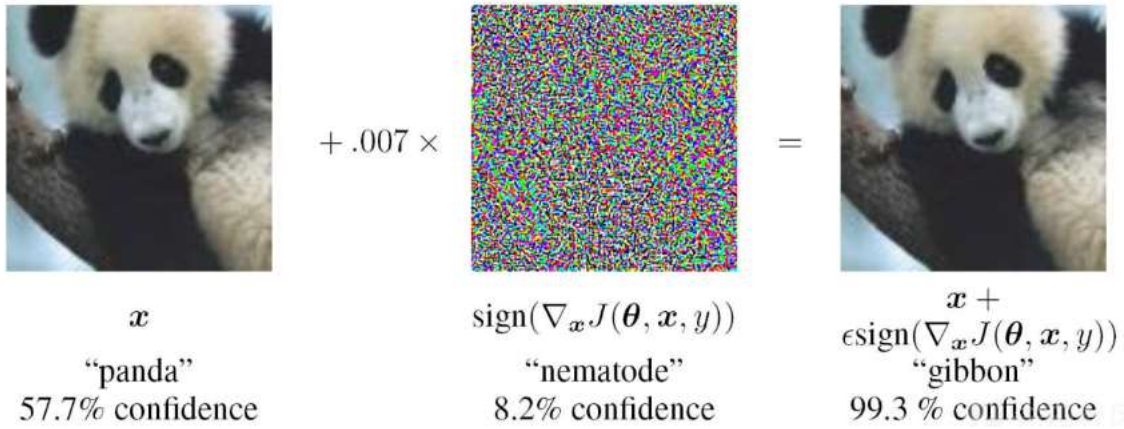


原始图像　　对抗样本　　扰动

知识基础

- 算法分类
  - 按照攻击者对模型的可见性划分
    - 白盒攻击
      - 攻击者拥有模型的全部信息，包括架构、参数、训练数据等
      - 这意味着攻击者可以对模型进行精确的梯度计算，设计出能够针对特定模型的高效攻击方法
    - 黑盒攻击
      - 攻击者对模型的内部信息知之甚少，只能通过观察模型的输入输出来进行攻击
      - 在黑盒场景中，攻击者可能需要使用进化算法或代理模型等技术来生成对抗样本
  - 按照攻击者对输出目标的控制程度划分
    - 有目标攻击：使深度学习模型做出特定的错误输出
    - 无目标攻击：使深度学习模型做出非特定的错误输出

- FGSM（白盒攻击）
  - 目标：制造一种特殊的、人眼难以察觉的噪音（扰动），添加到原始图片上，让一个训练好的神经网络模型做出错误的判断
  - 原理：沿着损失函数对输入的梯度方向，把图像往容易被误分类的方向推一步
  - 性能：计算快，一轮攻击；但扰动大，攻击精度低
  - 公式：$x_{adv} = x + \varepsilon \cdot sign(\nabla_x J(\theta, x, y))$
    - x：原始输入图像
    - x_{adv}：对抗样本
    - ε：扰动幅度（攻击强度）
    - $\nabla_x J(\theta, x, y)$：对输入的梯度
    - $sign(.)$：符号函数



知识基础

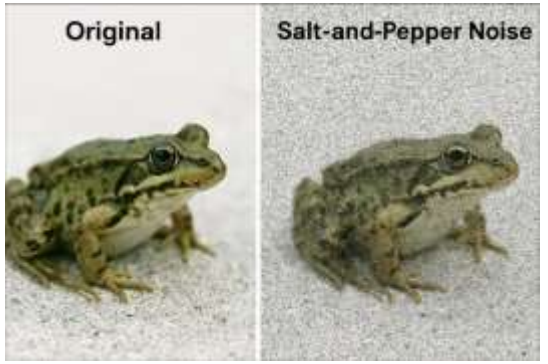
- **I-FGSM**
  - 原理：相比于FGSM走一大步，I-FGSM采用多次小步逼近（小步快跑，多轮攻击）的策略寻找最优解
  - 性能：隐蔽性和成功率都更好；但起点固定为原始图像，容易陷入局部最优
- **PGD**
  - 原理：攻击起点在原图周围邻域内随机选取，每次攻击依然沿着梯度方向，但每次攻击后都会把结果投影回原图的epsilon邻域内
  - 性能：允许从多个初始点出发、探索多个方向，不容易陷入局部最优；但参数更敏感，计算成本更高

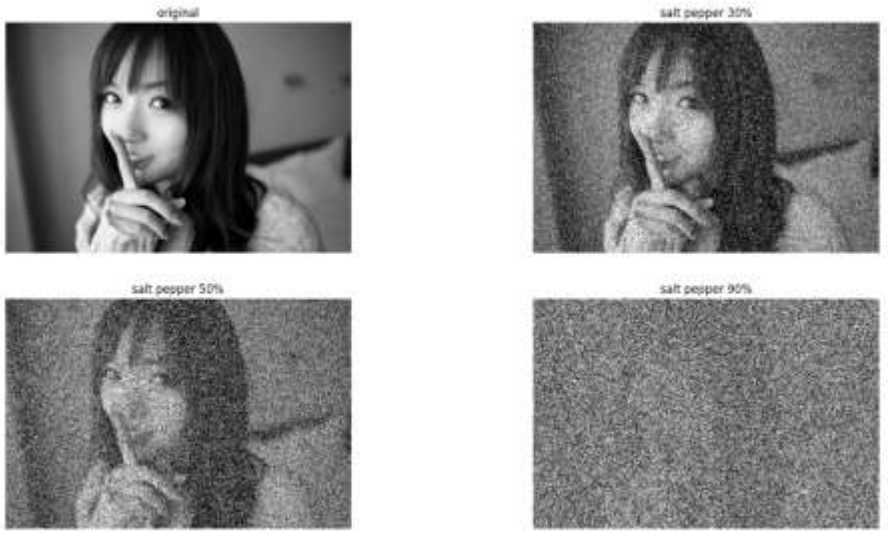
- 基本变换攻击（黑盒攻击）
  - 特点：通过对图像施加简单可感知但非学习型的变换，如添加噪声、模糊、亮度变化等，诱导模型产生错误预测
  - 优点：实现简单、查询效率高
  - 缺点：攻击精度低、扰动幅度大
  - 常见变换方式

变换类型	说明
添加噪声（如 Salt-and-Pepper）	将图像像素随机替换为黑/白值
高斯模糊	降低图像局部清晰度
图像遮挡（遮盖、贴纸）	模拟现实干扰（如物体遮挡）
对比度/亮度调整	改变全局感光参数
旋转、裁剪、缩放	轻微几何扰动也可能诱发误判

- **Salt-and-Pepper Noise**
  - 典型的基本变换攻击
  - 随机选取一部分像素，将其替换为最大/最小值（如黑色0或白色255）
  - 该过程模拟图像传感器的极端噪声干扰
  - 公式表达

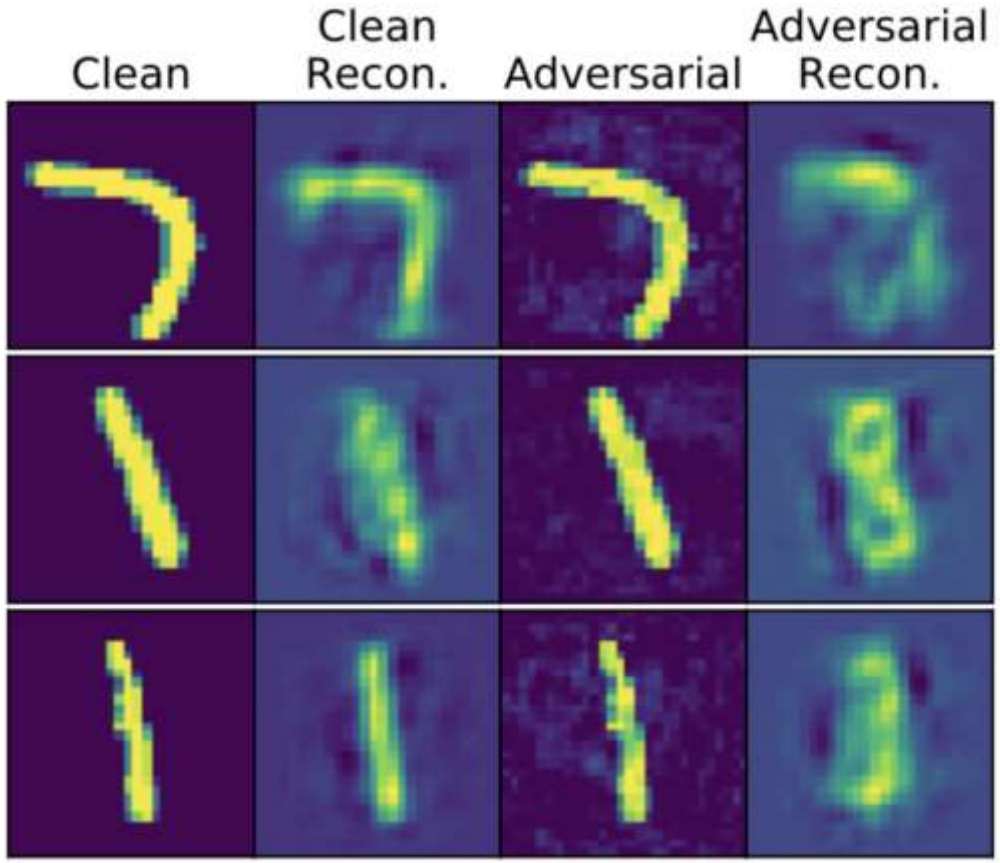
- $$x'_{i,j} = \begin{cases} 0, with\ probability\ p/2 \\ 255, with\ probability\ p/2 \\ x_{i,j}, with\ probability\ 1-p \end{cases}$$
  - p表示扰动强度
  - $x_{i,j}$是图像中第 i 行第 j 列对应的像素点



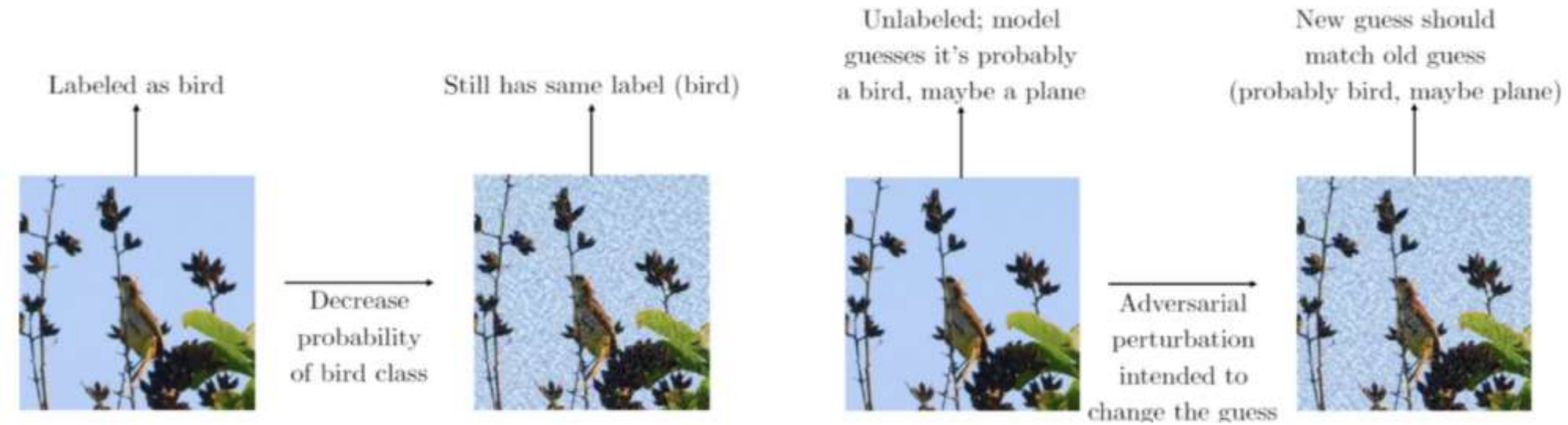


- 算法分类
  - 对抗样本检测技术
    - 基于特征学习的对抗样本检测
      - 对抗样本与原始样本的特征空间、维度大小不同
    - 基于分布统计的对抗样本检测
      - 对抗样本数字特征分布与正常样本差异较大
    - 基于中间输出的对抗样本检测
      - 对抗样本与正常样本的输入在深度神经网络中得到的中间输出状态有较大差距
  - 对抗样本防御技术
    - 对抗训练
    - 特征去噪
    - 防御蒸馏
    - 可证明式防御

- 对抗样本检测
  - 原始干净的图片（正常样本）一般是相机拍摄的物理世界的照片，它的数据处于一种比较自然的分布状态
  - 对抗样本在正常样本上加了一些扰动，而这些扰动是通过梯度等方式搜索出来的，并不符合自然界的分布
  - 二者的差异为对抗样本检测提供了一种思路：对数据的分布状态进行统计
  - 处理阶段：数据预处理阶段



- 对抗样本防御
  - 对抗训练方法：在模型的训练阶段主动生成对抗样本，将其纳入训练阶段，和正常样本一起输入给神经网络进行训练，达到防御对抗样本的目的
  - 特征去噪方法：在对抗样本输入模型之前进行去噪处理，将攻击者添加到原始图像上的轻微干扰去除，得到与原始图像近似的去噪后图像，从而分类依旧正确
  - 处理阶段：输入阶段、模型训练阶段



研究历史与现状

2014

Szegedy 等人首次发现深度神经网络的对抗脆弱性：仅需极小的、肉眼几乎不可察觉的扰动，就能让深度模型输出完全错误的预测

2015

Goodfellow等人提出 FGSM 一种基于梯度符号的快速生成对抗样本的攻击方法，同时首次提出了对抗训练的思路，将这些攻击样本加入训练集，从而增强模型的鲁棒性，为后续防御研究奠定了基础

2017

Carlini和Wagner提出C&W攻击方法；Xu等人提出Feature Squeezing 检测方法，通过输入压缩与特征简化来判断输入是否可疑，开启了对抗样本检测方向的系统化探索

2020

研究者发现对抗扰动不仅存在于数字图像，还能以物理形式出现：如贴纸攻击、语音命令注入、投影扰动等，展示了对抗攻击在现实世界中的威胁，使安全可部署的AI模型成为新的研究焦点

2022

研究者提出了一个无需调参的自动化攻击评测框架AutoAttack，组合了四种强攻击（包括 APGD-CE、APGD-DLR、FAB、Square Attack），实现了鲁棒性评估的可复现标准

2023

出现了利用扩散模型进行对抗攻击的新方向，通过在扩散过程反向传播中插入梯度扰动，生成视觉上更自然但极具欺骗性的对抗样本；这是攻击策略从简单梯度符号迈向生成式优化的关键一步，也开启了可感知但无形的多模态攻击趋势

2024

提出一种基于分块重排与随机旋转的结构级扰动方法，通过打破图像局部空间顺序并引入方向不敏感的梯度，使攻击摆脱对源模型特定空间模式的依赖，生成更具普适性的迁移扰动，为迁移攻击提供了一条轻量、有效且易用的新路径

2024

提出利用对抗采样提升图像分类鲁棒性的简单框架，仅通过对抗采样增强数据多样性，即可在多种架构与数据集上稳定提升鲁棒性，展示了将对抗扰动视为数据改良而非额外防御模块的新思路

Boosting Adversarial Transferability by Block Shuffle and Rotation

算法原理

T	目标	提高对抗攻击在黑盒模型 / 不同模型之间的迁移能力
I	输入	原始图像 x * 1，代理模型（白盒）* 1
P	处理	1、将原始图像x分割成若干块，对每个块执行随机打乱位置 + 随机旋转操作，生成一系列变换后的图像版本 2、将这些变换后的图像送入代理模型，计算梯度，优化对抗扰动 3、将得到的对抗样本用于黑盒模型测试，验证迁移攻击效果提升
O	输出	对抗样本 x_{adv} * 1

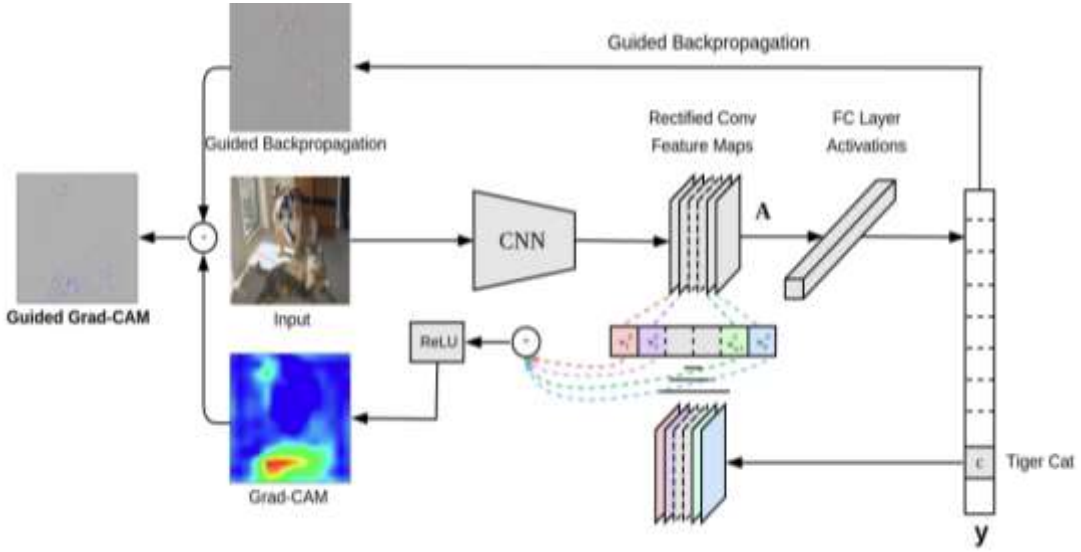
P	问题	模型生成的扰动难以迁移到其他黑盒模型攻击中
C	条件	图像分块+旋转变换仍保留类别信息
D	难点	主要用于黑盒迁移攻击，不一定适用于所有防御强的模型
L	水平	2024 CVPR

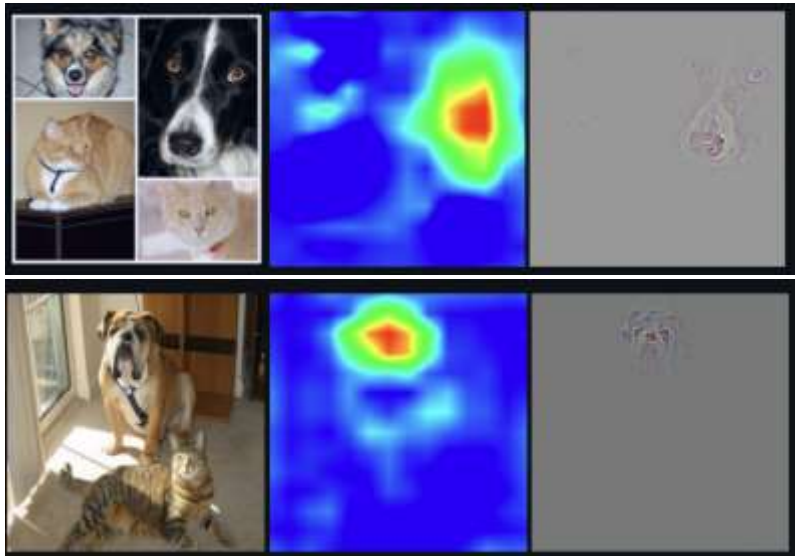
算法原理 知识回顾

- 黑盒攻击
  - 攻击者只能向目标模型输入数据并获取输出（如最终的分类结果或置信度分数），而不知道模型的内部细节
  - 更符合现实世界的攻击场景，基于迁移的攻击是黑盒攻击中一种特殊形式
- 对抗样本的迁移性
  - 在源模型/代理模型上生成的对抗样本，同样能够攻击另一个结构、参数均不同的模型（称为目标模型）
- 基于输入变换的攻击方法（黑盒）
  - 在计算生成对抗样本所需的梯度时，不是直接对原始图像求导，而是先对图像进行一系列随机变换（如缩放、平移、添加噪声等），然后对变换后的图像求梯度
  - 通过对输入进行多样化变换，可以迫使生成的扰动不那么依赖于源模型的特定特征，从而学习到一种更通用的扰动，提高迁移性

算法原理 知识补充

- 注意力热力图
  - 一种可视化技术，用于解释深度学习模型在做决策时关注了输入图像的哪些区域，颜色越亮（如红色）的区域表示模型认为该区域对当前决策越重要
- Grad-CAM
  - 使用流入 CNN 最终卷积层的特定类梯度信息生成图像中重要区域的粗略定位图





算法原理 知识补充

- 注意力差异
  - 不同输入变换方法生成的对抗样本，其注意力热力图在白盒（源）模型 Inc-v3 和黑盒（目标）模型 Inc-v4 之间差异较大，说明现有方法仍难以保持跨模型的一致关注区域，从而限制了对抗样本的迁移性
  - 通过优化扰动，使不同模型间的注意力热力图更加一致，以提升迁移性

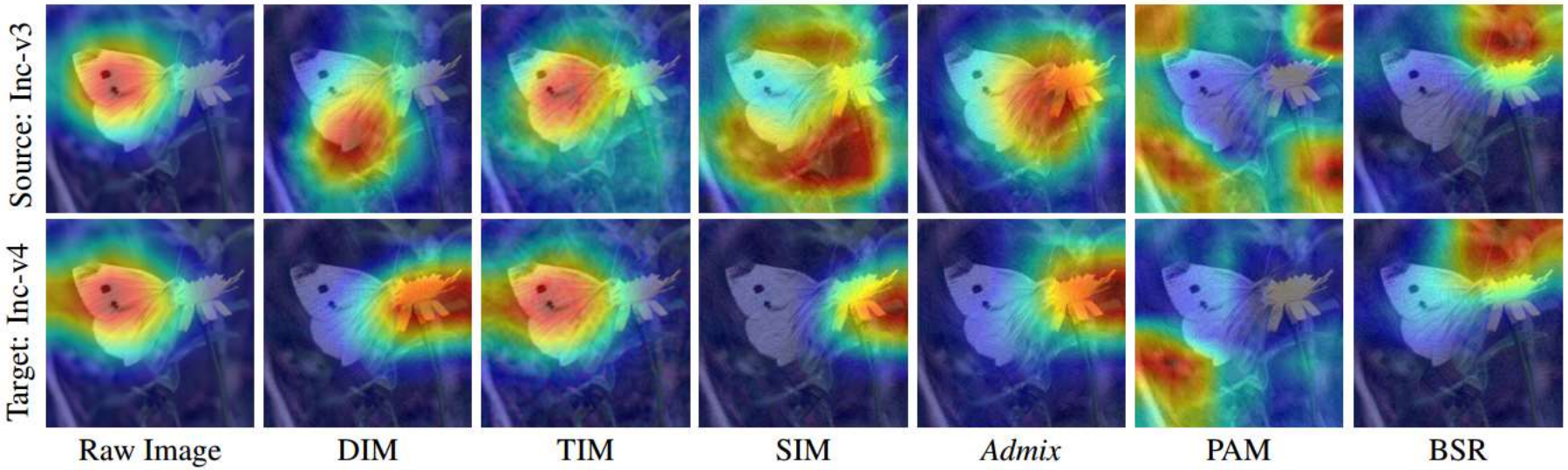


Figure 2. Attention heatmaps of adversarial examples generated by various input transformations using Grad-CAM.

算法原理 知识补充

- 注意力热力图变化
  - 原始未被破坏图像注意力模式干净、清晰
  - 打乱后模型的语义敏感区域消失
  - 复原后关注区域仍然异常，注意力受损不可逆

Raw Image　　Shuffled Image　　Reshuffled Image

算法原理 概念解析

- 多源模型
  - 如Inception-v3，Inception-v4，ResNet-101等
  - 理论上综合梯度生成的扰动更具有普适性、可迁移性
  - 真实场景下难以获取；计算梯度代价极高；联合优化困难

- 代理模型
  - 即在一个白盒模型上生成扰动，通过输入变换来模拟多源模型效果
  - 本实验使用Inc-v3生成对抗样本，攻击其它黑盒模型

算法原理

- 关键步骤
  - 生成对抗样本
    - 切块：参数n控制，生成n×n个块
    - 块打乱：对块进行随机打乱
    - 旋转：每个块独立旋转β角度，在τ内
    - 填充：延伸超出图像边界的部分去除，产生的间隙用0填充
  - 计算平均梯度
    - $\overline{g} = \frac{1}{N}\sum_{i=1}^{N} \nabla_{x^{adv}} J(T_i(x^{adv}, n, \tau), y; \theta)$
      - $T_i(.)$：第i次随机变换
    - 使扰动更加通用和可迁移
  - 更新动量
  - 更新对抗样本

Algorithm 1 Block Shuffle and Rotation

Input: A classifier f with parameters θ, loss function J; a raw example x with ground-truth label y; the magnitude of perturbation ϵ; number of iteration T; decay factor μ; the number of transformed images N; the number of blocks n; the maximum angle τ for rotation

Output: An adversarial example x^{adv}

$\alpha = \epsilon/T$, $g_0 = 0$

for $t = 1 \rightarrow T$ **do**

  Generate several transformed images: $\mathcal{T}(\mathbf{x}^{adv}, n, \tau)$

  Calculate the average gradient $\bar{g}_t$ by Eq. (6):

  Update the momentum g_t by:

$$g_t = \mu \cdot g_{t-1} + \frac{\bar{g}_t}{\|\bar{g}_t\|_1}$$

  Update the adversarial example:

$$x_t^{adv} = x_{t-1}^{adv} + \alpha \cdot \mathrm{sign}(g_t) \qquad (5)$$

end for

return x_T^{adv}

实验设计

- 数据集
  - ImageNet数据集，在验证集上选取1000张不同类别的图片
- 基线模型
  - 白盒模型（生成对抗样本）
    - 4种攻击模型： Inceptionv3 (Inc-v3) , Inception-v4 (Inc-v4), Inception-Resnetv3 (IncRes-v3) , Resnet-v2-101 (Res-101)
  - 黑盒模型（被攻击）
    - 三种集成模型：Inc-v3ens3, Inc-v3ens4, IncRes-v2ens2
- 基线算法
  - 防御方法
    - HGD，R&P，NIPS-r31，BitRD，JPEG，FD，RS，NRP
  - 五种基于输入变换的竞争攻击
    - DIM, TIM, SIM, Admix, PAM

实验设计

- 评价指标
  - 攻击成功率ASR(Attack Success Rate)
    - 作无目标攻击

$$ASR = \frac{1}{N}\sum_{i=1}^{n} 1\left[\left(f(x_i^{adv}) \neq y_i\right)\right]$$

    - 作有目标攻击

$$ASR_{targeted} = \frac{1}{N}\sum_{i=1}^{n} 1\left[\left(f(x_i^{adv}) = y_i\right)\right]$$

      - y_i， t_i：真实标签
      - x_i^{adv}：对抗样本
    - 意义：当对抗样本使模型不再输出原始标签时，计为一次攻击成功，ASR越高，攻击越强

实验设计

- 基础对比实验
  - 展示在单个白盒模型（如Inc-v3）上使用BSR时对黑盒模型的迁移效果
  - 对比对象：基线输入变换类攻击（ DIM、TIM、SIM、Admix、PAM）
  - 结论：相比baseline，自身可迁移性更强

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	DIM	98.6*	64.4	60.2	53.5	18.8	18.4	9.5
	TIM	**100.0***	49.3	43.9	40.2	24.6	21.7	13.4
	SIM	**100.0***	69.5	68.5	63.5	32.3	31.0	17.4
	Admix	**100.0***	82.2	81.1	73.8	38.7	37.9	19.8
	PAM	**100.0***	76.4	75.5	69.6	39.0	38.8	20.0
	BSR	**100.0***	**96.2**	**94.7**	**90.5**	**55.0**	**51.6**	**29.3**
Inc-v4	DIM	72.0	97.6*	63.8	57.2	22.6	21.1	11.7
	TIM	59.1	99.7*	49.0	41.9	26.8	22.9	16.6
	SIM	80.4	99.7*	73.4	69.4	48.6	45.2	29.6
	Admix	89.0	**99.9***	85.3	79.0	55.5	51.7	32.3
	PAM	86.7	**99.9***	81.6	75.9	55.4	50.5	33.2
	BSR	**96.1**	**99.9***	**93.4**	**88.4**	**57.6**	**52.1**	**34.3**
IncRes-v2	DIM	70.3	64.7	93.1*	58.0	30.4	23.5	16.9
	TIM	62.2	55.6	97.4*	50.3	32.4	27.5	22.6
	SIM	85.9	80.0	98.7*	76.1	56.2	49.1	42.5
	Admix	90.8	86.3	**99.2***	82.2	63.6	56.6	49.4
	PAM	88.6	86.3	99.4*	81.6	66.0	58.3	**51.0**
	BSR	**94.6**	**93.8**	98.5*	**90.7**	**71.4**	**63.1**	**51.0**
Res-101	DIM	76.0	68.4	70.3	98.0*	34.7	31.8	19.6
	TIM	59.9	52.2	51.9	99.2*	34.4	31.2	23.7
	SIM	74.1	69.6	69.1	99.7*	42.8	39.6	25.7
	Admix	84.5	80.2	80.7	**99.9***	51.6	44.7	29.9
	PAM	77.4	73.9	75.7	**99.9***	51.2	46.3	32.2
	BSR	**97.1**	**96.6**	**96.6**	99.7*	**78.7**	**74.7**	**55.6**

Table 1. Attack success rates (%) on seven models under single model setting with various single input transformations. The adversaries are crafted on Inc-v3, Inc-v4, IncRes-v2 and Res-101 respectively. * indicates white-box attacks.

实验设计

- 组合兼容性实验
  - 检验 BSR 是否与其他输入变换攻击（DIM、TIM、SIM、Admix、PAM）兼容、可叠加
  - 对比对象：不同组合
  - 结论： BSR 是一个通用增强模块，而不是替代其他方法的孤立方案

Attack	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
BSR-DIM	98.3$_{\uparrow 31.9}$	94.8$_{\uparrow 34.6}$	90.1$_{\uparrow 36.6}$	57.8$_{\uparrow 39.0}$	54.5$_{\uparrow 36.1}$	32.3$_{\uparrow 22.8}$
BSR-TIM	94.7$_{\uparrow 45.4}$	92.5$_{\uparrow 48.6}$	87.0$_{\uparrow 47.0}$	71.3$_{\uparrow 46.7}$	68.4$_{\uparrow 46.7}$	47.9$_{\uparrow 34.5}$
BSR-SIM	99.4$_{\uparrow 29.9}$	98.4$_{\uparrow 29.9}$	97.8$_{\uparrow 34.3}$	84.3$_{\uparrow 52.0}$	81.4$_{\uparrow 50.4}$	59.2$_{\uparrow 41.8}$
BSR-*Admix*	98.9$_{\uparrow 16.7}$	98.8$_{\uparrow 17.7}$	98.2$_{\uparrow 24.4}$	89.1$_{\uparrow 50.4}$	86.9$_{\uparrow 49.5}$	68.0$_{\uparrow 48.2}$
BSR-PAM	98.5$_{\uparrow 22.1}$	97.3$_{\uparrow 21.8}$	96.9$_{\uparrow 27.3}$	79.4$_{\uparrow 40.4}$	75.3$_{\uparrow 36.5}$	50.5$_{\uparrow 30.5}$
Admix-TI-DIM	90.4	87.3	83.7	72.4	68.4	53.4
PAM-TI-DIM	89.3	85.5	80.7	73.6	69.1	52.1
BSR-TI-DIM	95.2	92.9	87.9	74.2	70.7	50.0
BSR-SI-TI-DIM	**98.5**	**97.1**	**95.4**	**90.6**	**90.0**	**75.1**

Table 2. Attack success rates (%) on seven models under single model setting with various input transformations combined with BSR. The adversaries are crafted on Inc-v3. ↑ indicates the increase of attack success rate when combined with BSR.

实验设计

- 集成攻击实验
  - 评估 BSR 在多模型集成攻击下是否依然有效，模拟多模型梯度场景的表现
  - 对比对象：多种单个攻击及组合攻击方法
  - 结论：具有跨模型一致性

Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
DIM	99.0*	97.1*	93.4*	99.7*	57.6	51.5	35.9
TIM	99.8*	97.4*	94.7*	99.8*	61.6	55.5	45.6
SIM	99.9*	99.1*	98.5*	**100.0***	78.4	75.2	60.6
Admix	**100.0***	99.6*	99.0*	**100.0***	85.1	80.9	67.8
PAM	99.9*	99.7*	99.4*	**100.0***	86.1	81.6	69.1
BSR	**100.0***	**99.9***	99.9*	99.9*	**92.4**	**89.0**	**77.2**
Admix-TI-DIM	99.6*	98.8*	98.2*	99.8*	93.1	92.4	89.4
PAM-TI-DIM	99.8*	99.8*	99.2*	**99.8***	95.8	95.2	93.0
BSR–TI-DIM	99.8*	99.8*	99.7*	**99.8***	96.1	95.1	90.8
BSR-SI-TI-DIM	**99.9***	**99.9***	**99.9***	**99.8***	**99.1**	**99.1**	**97.0**

Table 3. Attack success rates (%) on seven models under ensemble model setting with various input transformations. The adversaries are crafted on Inc-v3, Inc-v4, IncRes-v2 and Res-101 model. * indicates white-box attacks.

实验设计

- 防御场景测试
  - 测试 BSR 生成的对抗样本能否突破常见防御机制（如 JPEG 压缩、随机重采样、NRP 等）
  - 对比对象：不同组合攻击方法
  - 结论：扰动具有结构鲁棒性，防御难以消除

Method	HGD	R&P	NIPS-r3	Bit-RD	JPEG	FD	RS	NRP	Average
Admix-TI-DIM	92.8	93.5	94.5	82.4	97.6	90.9	72.6	80.4	88.1
PAM-TI-DIM	95.4	95.3	96.4	85.9	98.4	93.4	74.0	83.8	91.6
BSR-TI-DIM	97.1	98.0	97.9	84.9	98.8	93.2	69.1	73.6	89.1
BSR-SI-TI-DIM	**98.5**	**99.1**	**99.4**	**91.4**	**99.2**	**97.1**	**83.9**	**84.2**	**94.1**

Table 4. Attack success rates (%) of eight defense methods by *Admix*, SSA and BSR input transformations. The adversaries are crafted on Inc-v3, Inc-v4, IncRes-v2 and Res-101 synchronously.

实验流程

- 组件有效性消融实验
  - MI-FGSM：基础对照组，只利用动量，不进行任何输入变换
  - BS：只进行块打乱，不旋转
  - BR：只进行块旋转，不打乱
  - BSR：两者结合
- 结论
  - BS 和 BR 都能单独提升迁移性，相比 MI-FGSM 有明显增益
  - BSR最能破坏模型注意力的一致性，能让模型在不同变换下产生更丰富的梯度，从而生成更具迁移性的对抗样本

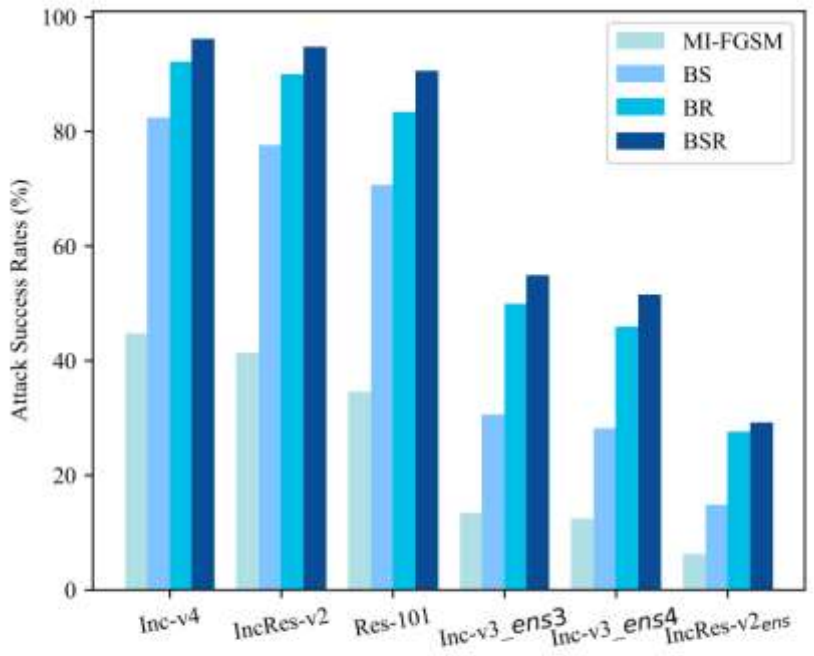


Figure 3. Attack success rates (%) of various models on the adversarial examples generated by MI-FGSM, BS, BR and BSR, respectively.

实验设计

- 超参数分析实验
  - n：图像被分成多少块进行打乱
  - N：每步计算平均梯度时使用多少个变换版本
  - τ：每个块旋转的最大角度范围

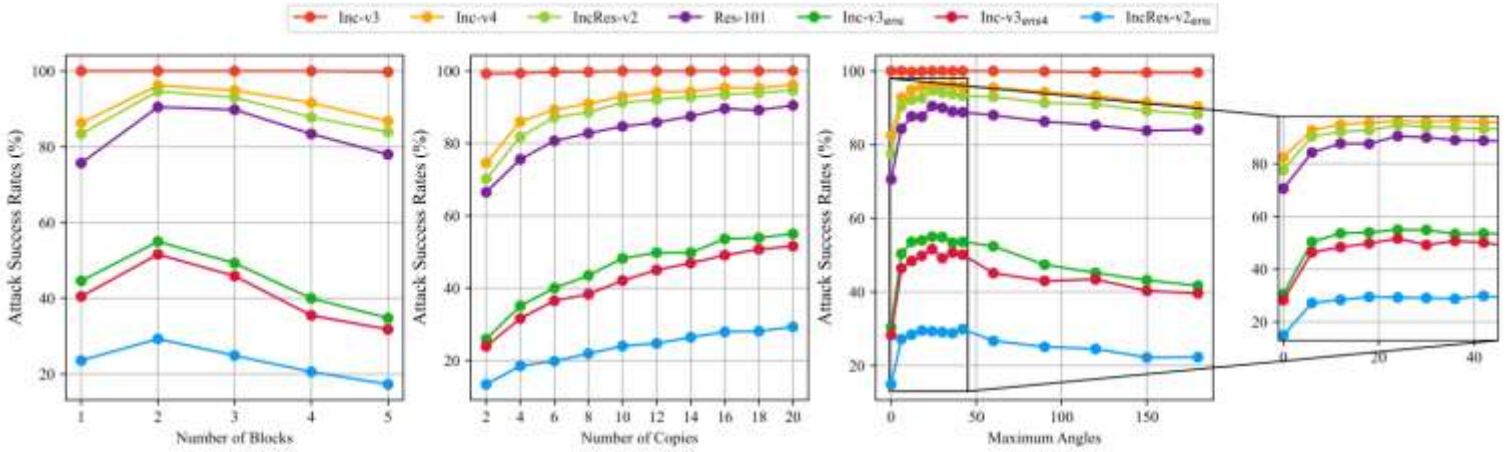


Figure 4. Attack successful rates (%) of various models on the adversarial examples generated by BSR with various numbers of blocks, number of transformed images and range of the rotation angles. The adversarial examples are crafted on Inc-v3 model and tested on the other six models under the black-box setting.

算法总结

- 算法贡献
  - 提出了一种全新的输入变换机制 BSR（Block Shuffle and Rotation）
  - 单模型条件下显著提升对抗迁移性
  - 方法简单通用，可与现有攻击框架兼容
  - 证明了注意力多样性与迁移性之间的联系

- 算法不足
  - 对防御模型的迁移效果仍有限
  - 变换仍存在信息破坏风险

Enhancing Image Classification Robustness through Adversarial Sampling with Delta Data Augmentation (DDA)

算法原理

T	目标	提升下游图像分类模型的对抗鲁棒性
I	输入	上游鲁棒模型 A * 1，上游数据集 X_A * 1，下游任务数据集 X_B * 1
P	处理	1、在上游鲁棒模型 A 上生成对抗扰动 δ 2、构建扰动池 Δ 3、将 δ 注入下游训练数据
O	输出	增强后的下游数据集 X'_B * 1

P	问题	以较低成本注入鲁棒特征
C	条件	上游鲁棒模型 A 必须足够鲁棒，下游任务领域差距不能太大，对抗扰动 δ 需要满足人类视觉可接受的约束
D	难点	1、保持自然准确率 2、如何构造扰动池 Δ 的多样性
L	水平	2024 CVPR Workshops

算法原理 知识回顾

- 对抗训练
  - 标准流程
    - 对每张训练图像生成一个对抗样本
    - 用原图+对抗图一起训练模型
  - 痛点
    - 对抗样本单一：PGD/FGSM生成的是某个方向上的最坏扰动
    - 训练成本高：每次训练都要生成新的样本
  - 解决思路
    - 不只用最坏扰动，从原图附近随机采样多个扰动，用更多样化的邻域训练模型
  - 小扰动一致性
    - 一个鲁棒模型在同一个输入的轻微扰动附近应该做出一致预测

算法原理 知识补充

- 数据增强
  - 传统数据增强
    - 翻转、裁剪、缩放、抖动、模糊噪声等
    - 直接作用于图像像素
    - 鲁棒性提升有限
  - 扰动空间数据增强
    - 扰动邻域：球形邻域
      - $B_\varepsilon(x) = \{x' \mid \|x' - x\| \leq \varepsilon\}$
        - » ε :允许的最大扰动半径
    - 找的是扰动邻域内哪个点最容易让模型崩溃
  - 核心思想
    - 把传统的数据增强思想从像素空间迁移到了扰动空间，通过在邻域内采样大量不同方向的小扰动（δ-data），让模型见过更多攻击可能性，从而增强鲁棒性

算法原理

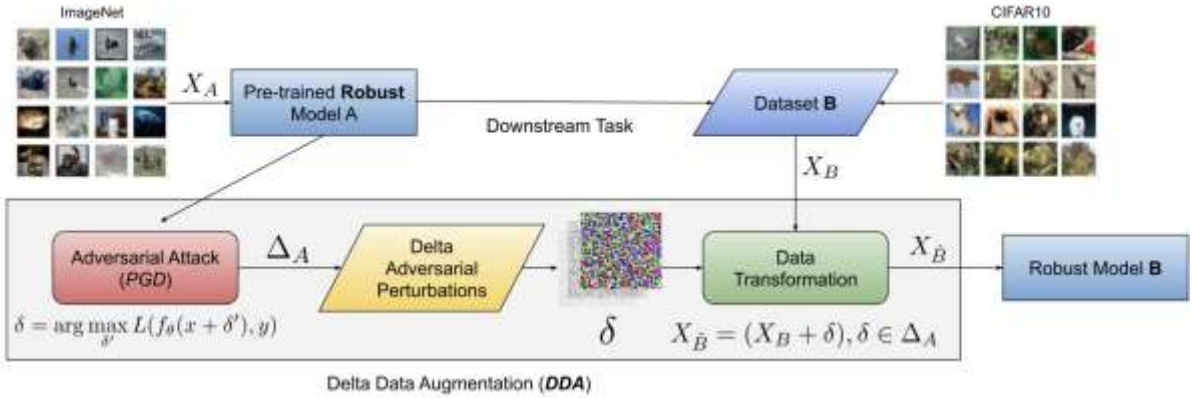
- 关键步骤
  - 使用PGD攻击上游鲁棒模型M生成对抗样本
  - 从成功欺骗模型M的图像中随机选择k个图片
  - 提取出噪声形成一个扰动集合Δ
  - 重新设置扰动尺寸，使其与下游任务图像大小匹配
  - 从PGD生成的扰动方向池中再次采样，对下游数据集施加扰动

Algorithm 1 Delta Data Augmentation

Require: Pre-trained Robust Model M_A, Dataset D, Length of adversarial samples k

Ensure: Augmented Dataset $\hat{D}$

1: Attack model M_A with PGD
2: Select the images that fooled model M_A
3: Sample k effective adversarial images
4: Extract in $\Delta \leftarrow k$ the effective perturbations
5: Resize Δ for D image size
6: **for** $x_i \in D_{train}$ **do**
7: Randomly select a perturbation $\delta \sim \Delta$
8: Apply perturbation on original image $\hat{x_i} \leftarrow x_i + \delta$
9: **end for**

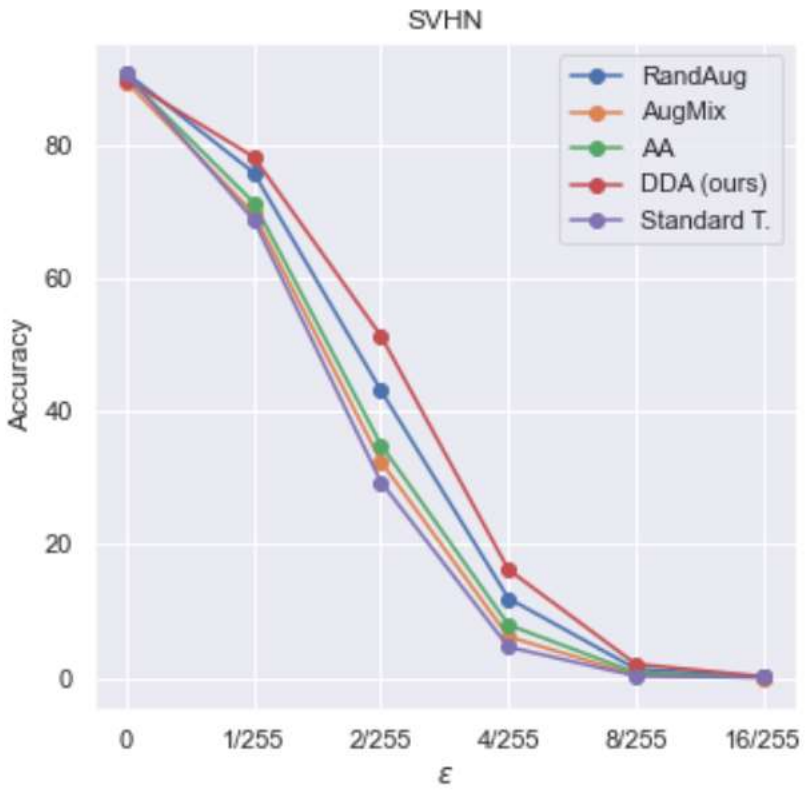


Delta Data Augmentation (*DDA*)

实验设计

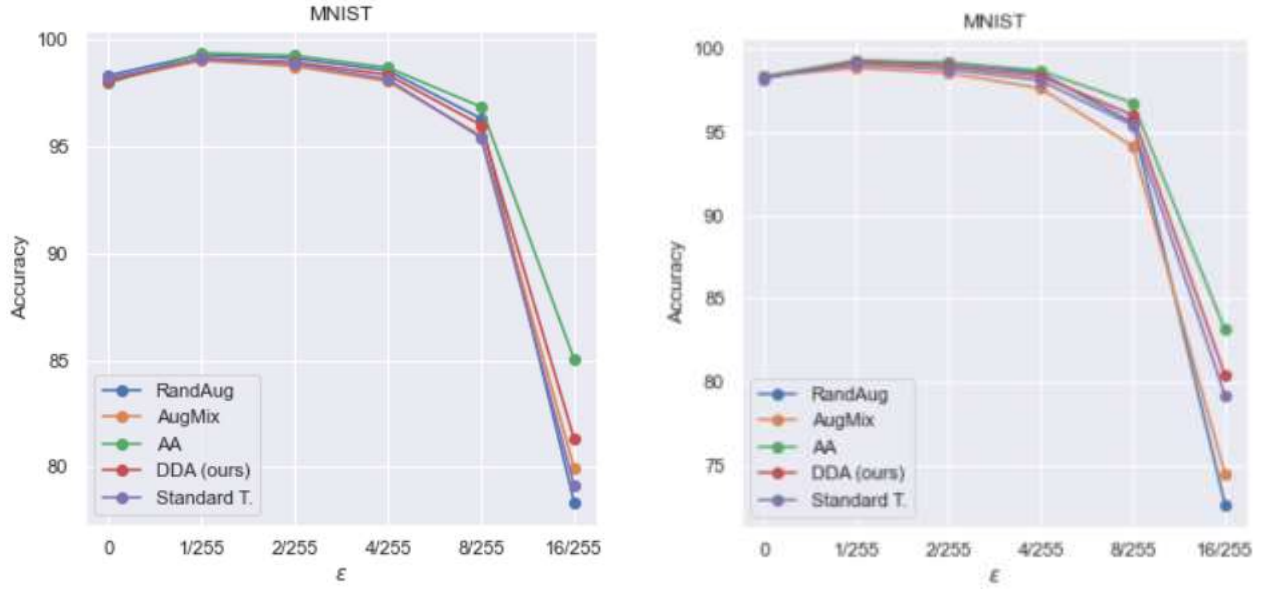
- 基础对比实验
  - 数据集
    - ImageNet、CIFAR-10、SVHN
  - 其他数据增强技术
    - RandAugment, AutoAugment, AugMix

PGD Norm	Dataset	No DA	RandAug	AutoAug	AugMix	DDA (ours)
$l_2 = 0.5$	CIFAR10	6.48±0.63%	7.64±0.54%	6.42±0.54%	7.79±0.97%	**8.94 ± 0.69%**
	CIFAR100	2.46±0.24%	3.23±0.25%	2.61±0.16%	2.84±0.34%	**4.06 ± 0.35%**
	SVHN	5.77±1.03%	3.23±0.25%	11.21±0.80%	6.99±0.67%	**19.55 ± 2.37%**
$l_\infty = 4/255$	CIFAR10	2.39±0.34%	2.95±0.32%	2.11±0.11%	3.55±0.74%	**3.73 ± 0.57%**
	CIFAR100	1.27±0.21%	1.78±0.11%	1.39±0.20%	1.60±0.26%	**2.17 ± 0.33%**
	SVHN	4.60±0.35%	11.91±1.23%	7.93±0.89%	6.16±0.27%	**16.23 ± 2.88%**



实验设计

- 消融实验
  - ResNet模型大小消融研究（左为ResNet18）



Model	Channels	Datasets	Image Size	PGD Norm
ResNet18	RGB	CIFAR10, CIFAR100, SVHN	32×32	l_2
ResNet18	RGB	CIFAR10, CIFAR100, SVHN	32×32	l_∞
ResNet50	Grayscale	MNIST, FashionMNIST	28×28	l_2
ResNet50	Grayscale	MNIST, FashionMNIST	28×28	l_∞

算法总结

- 算法贡献
  - 提出了一种扰动空间的数据增强思想
  - 对抗训练的轻量替代
  - 扰动方向可迁移，跨数据集、跨模型都有效
  - 对小模型和大模型都有效，不依赖模型容量
- 算法不足
  - δ 必须来自一个强鲁棒上游模型，否则效果大降
  - 跨数据集任务时需要重新计算尺寸，本质上改变了噪声结构
  - δ 并非任务自适应，对目标任务没有定制鲁棒性

特点总结与未来展望

算法总结

- 特点总结
  - BSR
    - 方法简单通用，可与现有攻击框架兼容
    - 证明了注意力多样性与迁移性之间的联系
  - DDA
    - 对抗训练的轻量替代
    - 对小模型和大模型都有效，不依赖模型容量

- 未来发展
  - 从单模型转向多模型的鲁棒一致性
  - 让采样具备任务适应性

参考文献

1. Wang K, He X, Wang W, et al. Boosting adversarial transferability by block shuffle and rotation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition[C]. Seattle, WA, USA: IEEE, 2024: 24336-24346.
2. Reyes-Amezcua I, Ochoa-Ruiz G, Mendez-Vazquez A. Enhancing image classification robustness through adversarial sampling with delta data augmentation (dda). Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. Seattle, WA, USA: IEEE, 2024: 274-283.

道德经

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。