Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

深度神经网络鲁棒性评估方法

硕士研究生 夏志豪

2023年04月02日

内容提要

- 背景简介
- 基本概念
- 算法原理
  - ROBY
  - DeepPAC
- 应用总结
- 参考文献

背景简介

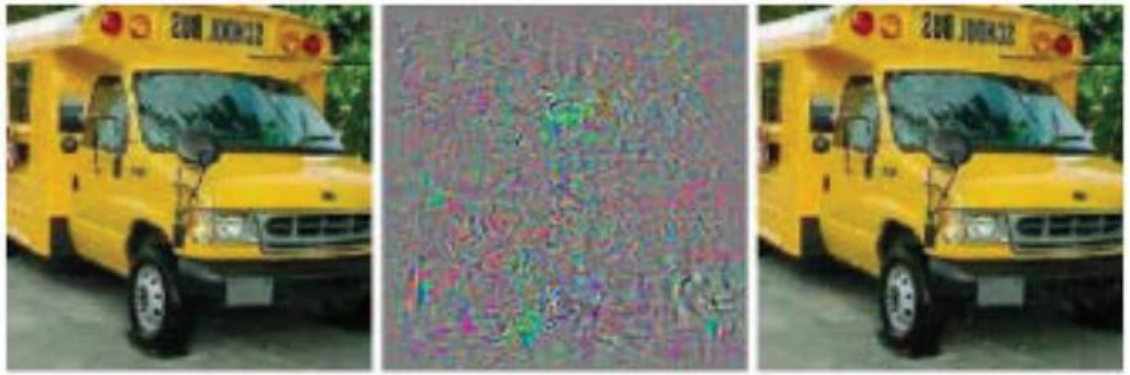
- 预期收获
  - 了解深度神经网络模型鲁棒性基本概念
  - 了解深度深度神经网络鲁棒性评估方法的类型
  - 理解深度神经网络模型鲁棒性评估算法的原理
  - 了解深度神经网络模型鲁棒性评估现存问题和发展方向

背景简介

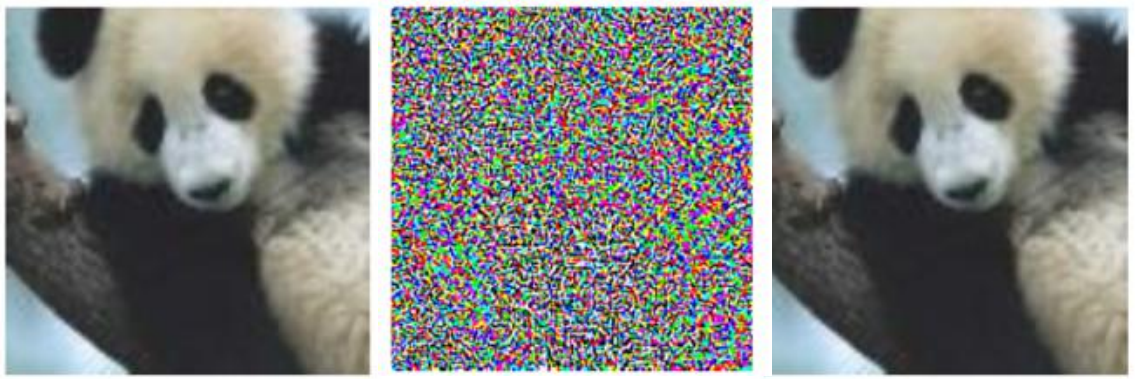
- 深度神经网络的任务
  - 使用广泛的数据训练模型，利用数学方法学习数据到决策的抽象映射关系
  - 模拟人类对世界的认识和判断
- 深度神经网络面临的困难
  - 训练数据在真实空间中是稀疏的，无法学习数据到决策的所有细节
  - 算力制约模型发展
  - 意外的干扰影响深度神经网络的判断

背景简介

- 深度神经网络鲁棒性
  - 深度神经网络技术发展迅速，在图像识别、自然语言处理等领域发挥重要作用
  - 深度神经网络模型在关键领域的应用导致其安全性需要得到认可
  - 从对抗样本角度理解，鲁棒性是最大扰动范围
  - 鲁棒性分析可以获取模型的服务质量信息，衡量其安全性，规避潜在的安全威胁



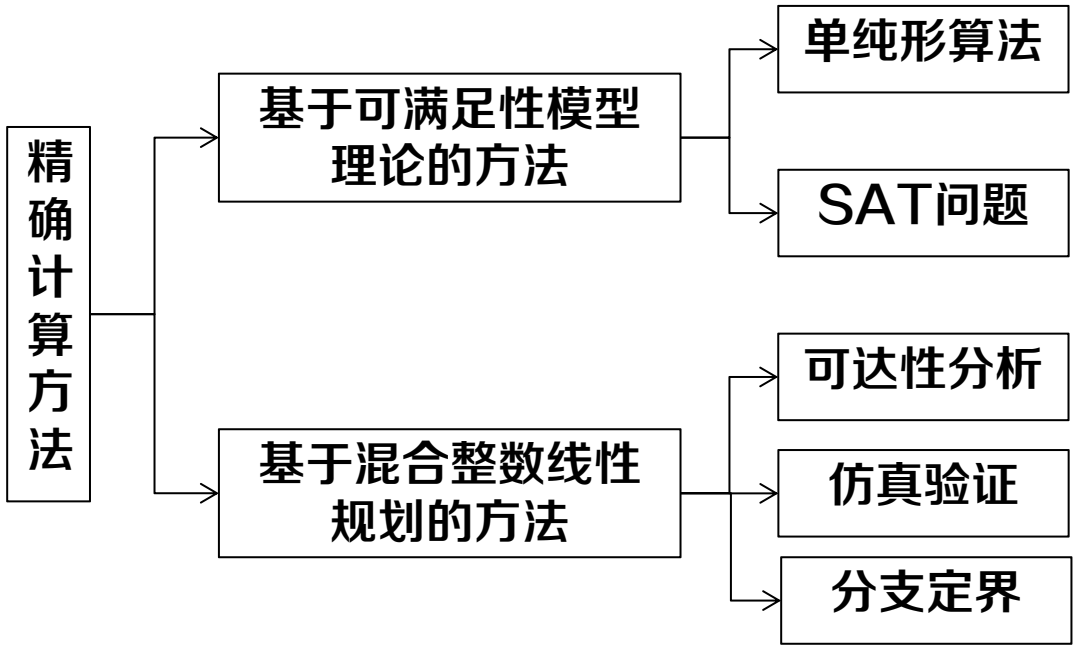
校车　　扰动　　鸵鸟



熊猫　　扰动　　长臂猿

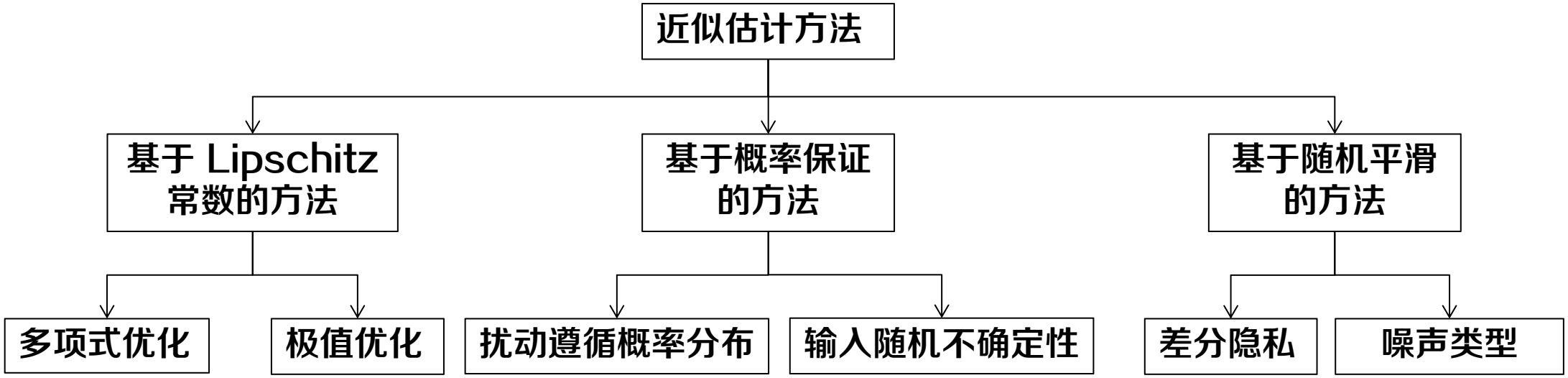
基本概念

- 鲁棒性分析
  - 精确计算
    - 主要是基于离散优化理论来形式化验证网络中某些属性对于任何可能的输入的可行性或混合整数线性规划来解决此类形式验证问题
  - 前提条件
    - 使用ReLU函数作为激活函数
  - 优缺点
    - 可以准确计算模型鲁棒性
    - 只能在小型网络上使用
    - 只适用于分段线性网络



基本概念

- 鲁棒性分析
  - 近似估计
    - 用近似方法计算模型鲁棒性边界的下界，对抗攻击可以得到模型鲁棒性边界的上界，精确的模型鲁棒性边界可以由上界和下界共同逼近得到
    - 对激活函数没有限制，可以使用在大型网络上



【Information Sciences】
ROBY：Evaluating the adversarial robustness of a deep model by its decision boundaries

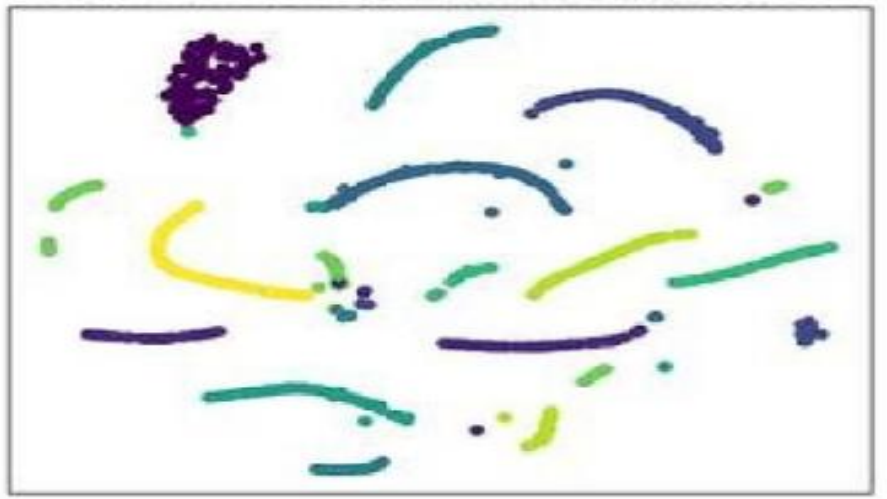
算法原理 ROBY

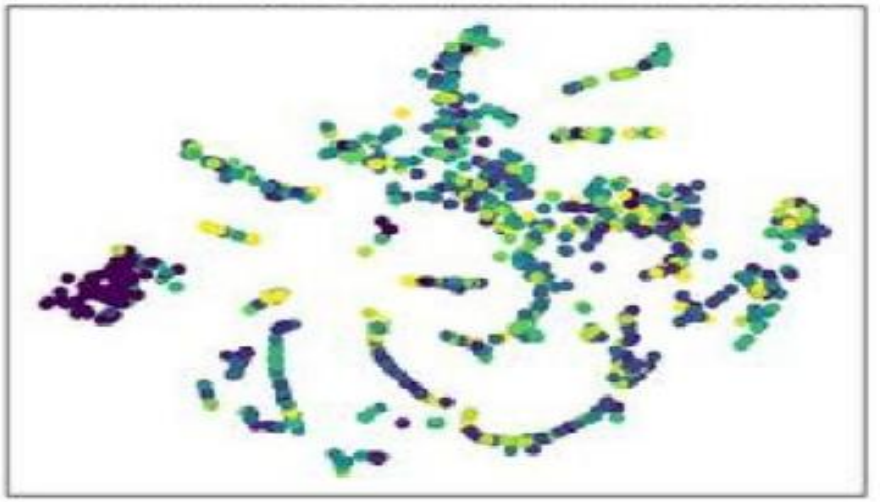
T	目标	评估DNN模型的鲁棒性
I	输入	一个DNN模型、一组数据集
P	处理	1、计算特征子空间聚合度FSA 2、计算特征子空间距离FSD 3、结合FSA和FSD得到鲁棒性指标ROBY
O	输出	模型的鲁棒性指标

P	问题	利用攻击方法评估鲁棒性资源开销大
C	条件	能够得到模型输出层前的特征向量
D	难点	决策边界的计算
L	水平	Information Sciences 2022（SCI 一区）

算法原理 ROBY

- 对抗样本攻击
  - 对样本添加细微扰动，影响模型对真实世界的学习
  - 模糊样本决策边界
  - 对真实世界做出错误的判断
- 算法原理
  - 特征空间中样本聚集情况越清晰代表鲁棒性越强
  - 类内聚集紧密
  - 类间分离疏远





算法原理 ROBY

- 计算过程
  - 特征子空间聚合度

$$FSA_k = 1 - \frac{norm(\sum_{j=1}^{n_k} dist(f_{x_j,k}, f_{c_k}))}{n_k}$$

  - 特征子空间距离

$$FSD_{k,k+1} = dist(f_{c_k}, f_{c_{k+1}})$$

  - 鲁棒性指标（ROBY）

$$ROBY_{k,k+1} = FSA_k + FSA_{k+1} - FSD_{k,k+1}$$

$$ROBY = \frac{\sum_{i=1}^{k}\sum_{j=i+1}^{k} ROBY_{i,j}}{k(k-1)/2}$$

通过类间距离和类内距离衡量鲁棒性

Algorithm 1: Compute the robustness metrics of FSA, FSD, ROBY

Input: Samples with K classes and their feature vector f.
Output: FSA, FSD, ROBY value.

1. $FSA\_list \leftarrow \{\varnothing\}$, $center\_list \leftarrow \{\varnothing\}$, $ROBY\_list \leftarrow \{\varnothing\}$
2. **for** $k \leftarrow 1$ to K **do**
3. **for** $i \leftarrow 1$ to n_k **do**
4. $f_{c_k} \leftarrow f_{c_k} + f_{x_j,k}$
5. **end for**
6. $f_{c_k} \leftarrow f_{c_k}/n_k$
7. $center\_list \leftarrow center\_list \bigcup f_{c_k}$
8. **end for**
9. **for** $k \leftarrow 1$ to K **do**
10. **for** $i \leftarrow 1$ to n_k **do**
11. $d_k \leftarrow d_k + dist(f_{x_j,k}, f_{c_k})$
12. **end for**
13. $FSA_k \leftarrow d_k/n_k$
14. $FSA\_list \leftarrow FSA\_list \bigcup FSA_k$
15. **end for**
16. $FSA \leftarrow 1 - Distance\ Selection\ Module(norm(FSA\_list))$
17. **for** $i \leftarrow 1$ to $K-1$ **do**
18. **for** $j \leftarrow i+1$ to K **do**
19. $d_{i,j} \leftarrow dist(f_{c_i}, f_{c_j})$
20. $d \leftarrow d + d_{i,j}$
21. **end for**
22. **end for**
23. $FSD \leftarrow norm(d/(K*(K-1)/2))$
24. **for** $i \leftarrow 1$ to K **do**
25. **for** $j \leftarrow i+1$ to K **do**
26. $ROBY_{i,j} \leftarrow FSA_i + FSA_j - dist(f_{c_i}, f_{c_j})$
27. $ROBY\_list \leftarrow ROBY\_list \bigcup ROBY_{i,j}$
28. **end for**
29. **end for**
30. $ROBY \leftarrow Distance\ Selection\ Module(norm(ROBY\_list))$

循环1

循环2

循环3

循环4

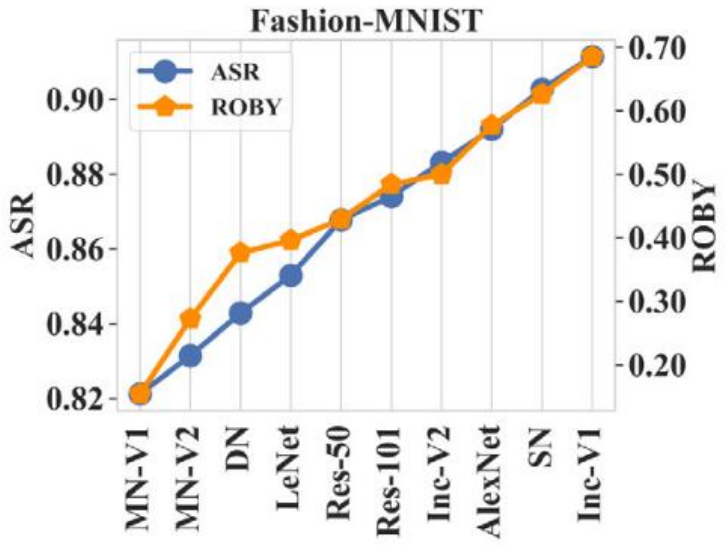
实验结果 ROBY

- 实验设置
  - 数据集
    - 5个图片数据集，1个音频数据集
  - 攻击方法
    - PGD
    - Boundary Attack
  - 目的
    - 验证ROBY的有效性
    - 验证ROBY与对抗训练的关系
    - 验证ROBY与模型结构的关系

数据集	数据集简介
MNIST	手写数字图片
Fashion-MNIST	10 种共 7 万个商品图片
CIFAR-10	10 类 RGB 彩色图片
CIFAR-100	100 类 RGB 彩色图片
Tiny-ImageNet	200 类 RGB 彩色图片
VCTK	110 人的英文语音

实验结果 ROBY

- ROBY的有效性
  - 白盒攻击
  - 黑盒攻击
  - 攻击成功率反映鲁棒性
  - 与ASR呈现相关性



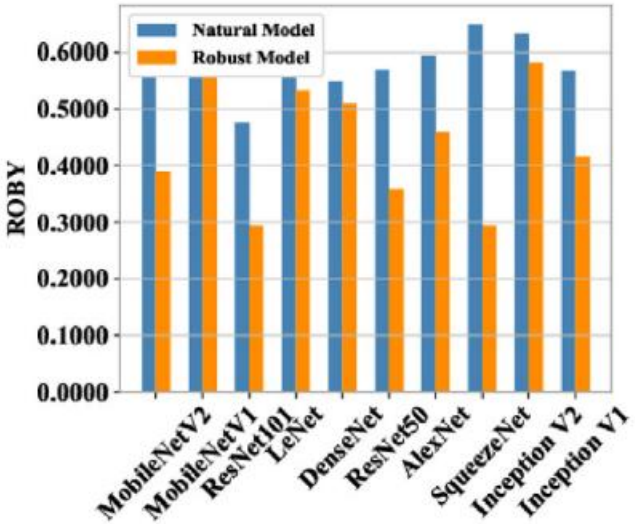
Dataset	Model	ACC	ASR	ER	CLEVER	ROBY
F-MINST	LeNet	0.8985	0.8653	0.0412	0.0389	0.3480
	Res-50	0.9164	0.8763	0.0478	0.0437	0.3907
	Res-101	0.9189	0.8816	0.0466	0.0455	0.4513
	AlexNet	0.914	0.9071	0.0848	0.0716	0.5520
	SN	0.892	0.9409	0.0900	0.0901	0.6029

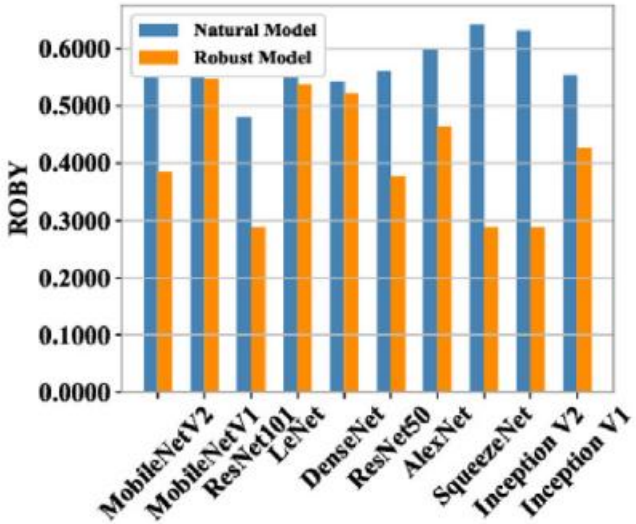
Dataset	ER	CLEVER	ROBY
MINST	0.8503	0.8607	**0.9565**
F-MINST	0.9419	**0.9770**	0.9695
CIFAR-10	0.8289	**0.9244**	0.8958
CIFAR-100	0.8390	0.8171	**0.8992**
Tiny-ImageNet	0.7459	0.7359	**0.9528**

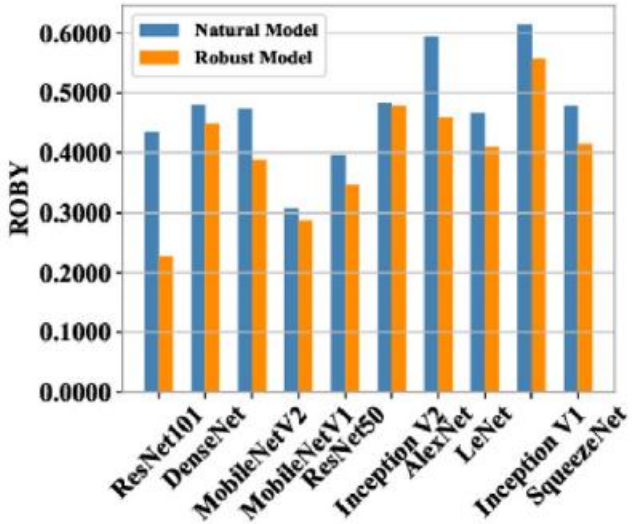
ROBY与ASR相关，能有效反映模型鲁棒性

实验结果 ROBY

- ROBY与对抗训练的关系
  - 对抗训练
  - 区分经过对抗训练的模型







ROBY能反映出模型经对抗训练带来鲁棒性提高

实验结果 ROBY

- ROBY与模型结构的关系
  - 神经元数量
  - 隐藏层层数
  - 模型架构

Dataset	Model	Layer	Neurons	ASR	ROBY
MINST	FCN-1	2	100	0.9998	0.5503
	FCN-2	2	500	0.9979	0.5497
	FCN-3	2	1000	0.9972	0.5462
	FCN-4	2	2000	0.9824	0.5428
	FCN-5	2	3000	0.9058	0.5411
	FCN-6	2	4000	0.8726	0.4722

Model	Conv	Pool	Dropout	FC	ASR	ROBY
CNN-1	√	√	√	√	0.9250	0.5735
CNN-2	√		√	√	0.9693	0.5955
CNN-3	√	√		√	0.7832	0.5004
CNN-4	√			√	0.9670	0.5857

Dataset	Model	Layer	Neurons	ASR	ROBY
MINST	FCN-6	2	4000	0.8726	0.4722
	FCN-7	3	4000	0.8176	0.4412
	FCN-8	4	4000	0.7822	0.3711
	FCN-9	5	4000	0.7489	0.3349

ROBY能反映出模型结构改变带来鲁棒性变化

实验结论 ROBY

- 算法总结
  - 通过特征子空间聚合度计算类内统计特征
  - 通过特征子空间距离计算类间统计特征
  - 基于类内和类间统计特征评估鲁棒性
- 算法优势
  - 算法开销小，时间复杂度低
  - 不涉及模型结构，可扩展性高
- 思考方向
  - 如何精确判断什么范围内不会出现对抗样本

算法原理 DeepPAC

【 ICSE 】

Towards Practical Robustness Analysis for DNNs based on PAC-Model Learning

算法原理

T	目标	分析深度神经网络的鲁棒性
I	输入	一个DNN模型、四组数据集
P	处理	1、采样生成输入输出对 2、构建PAC模型 3、利用线性规划的方式计算模型系数
O	输出	深度神经网络的鲁棒半径

P	问题	模型在高维输入时的鲁棒性分析
C	条件	DNN可以被线性函数近似
D	难点	优化PAC模型系数的计算过程
L	水平	ICSE 2022（CCF A类）

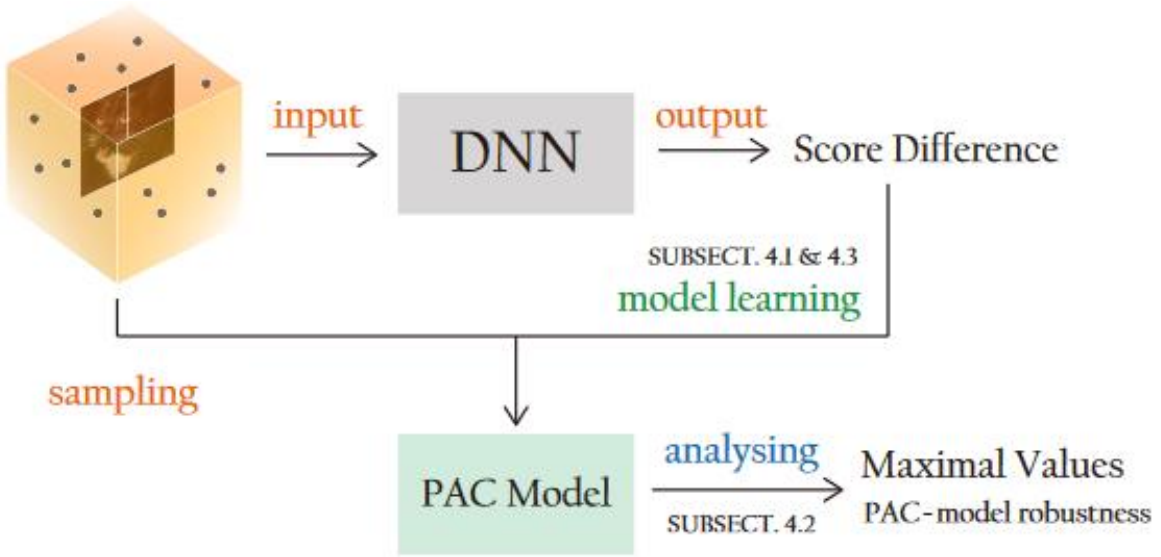
算法原理 DeepPAC

- 算法背景：深度神经网络的输出可以由线性函数近似模拟

$$DNN_{m\to n}(x) \approx A_{n\times m}x_{m\times 1}$$

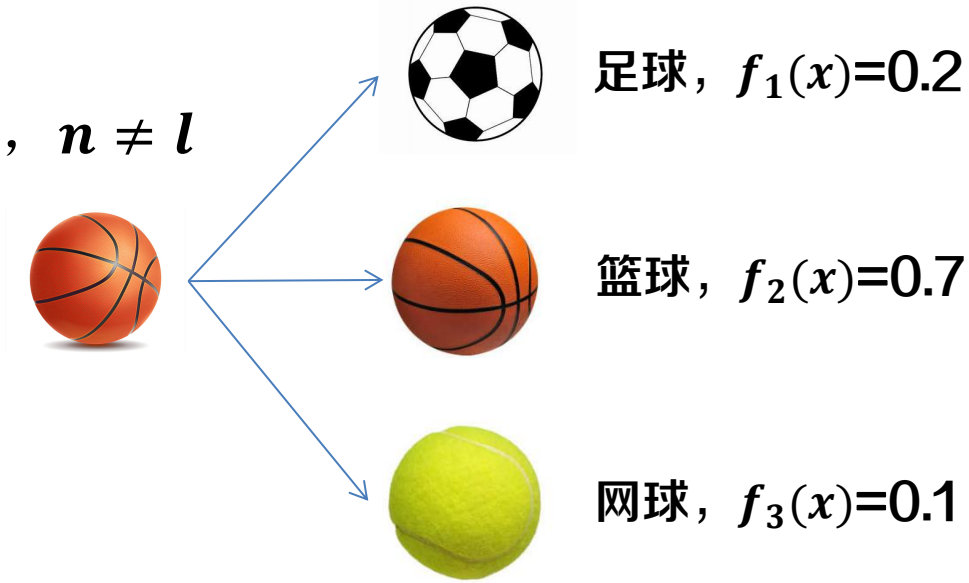
  - 通过模型白盒信息计算鲁棒半径困难
  - 在允许的概率误差范围内，以线性函数的鲁棒半径代替模型的鲁棒半径

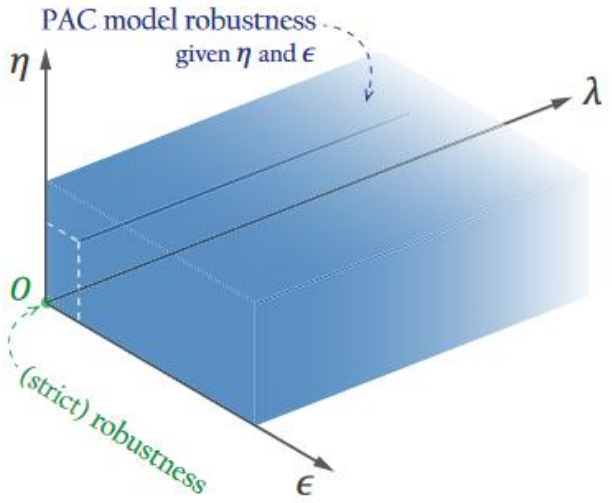
- 算法原理
  - 样本周围$B(x,r)$采样，得到各样本点的得分差函数Δ
  - 通过多轮次聚焦学习得到PAC模型的系数
  - 分析PAC模型的鲁棒性



算法原理 DeepPAC

- 计算样本点得分差函数Δ
  - $\Delta(x) = (f_1(x) - f_l(x), \ldots\ldots, f_n(x) - f_l(x))^T$，$n \neq l$
  - $\Delta(x) < 0$，分类正确
- 线性函数近似原模型
  - 假设$\widetilde{\Delta}(x) = c^T x = c_1 x_1 + c_2 x_2 + \cdots + c_m x_m$
  - 通过PAC模型计算线性函数的系数
  - $P(||\widetilde{\Delta}(x) - \Delta(x)||_\infty \leq \lambda) \geq 1 - \varepsilon$，$with\ 1 - \eta$
    - 最小化λ，使$\widetilde{\Delta}(x)$与$\Delta(x)$足够接近
    - 线性规划的方法解决此优化问题

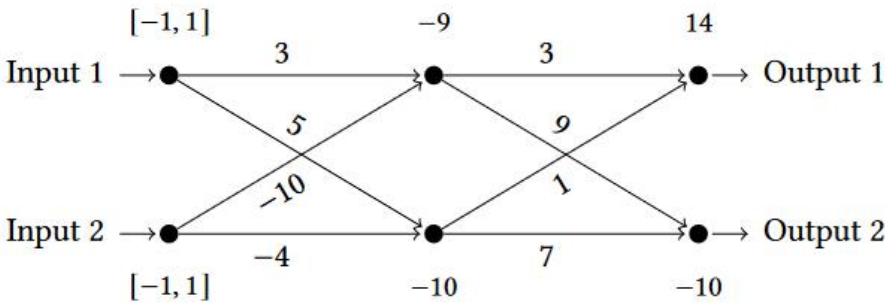


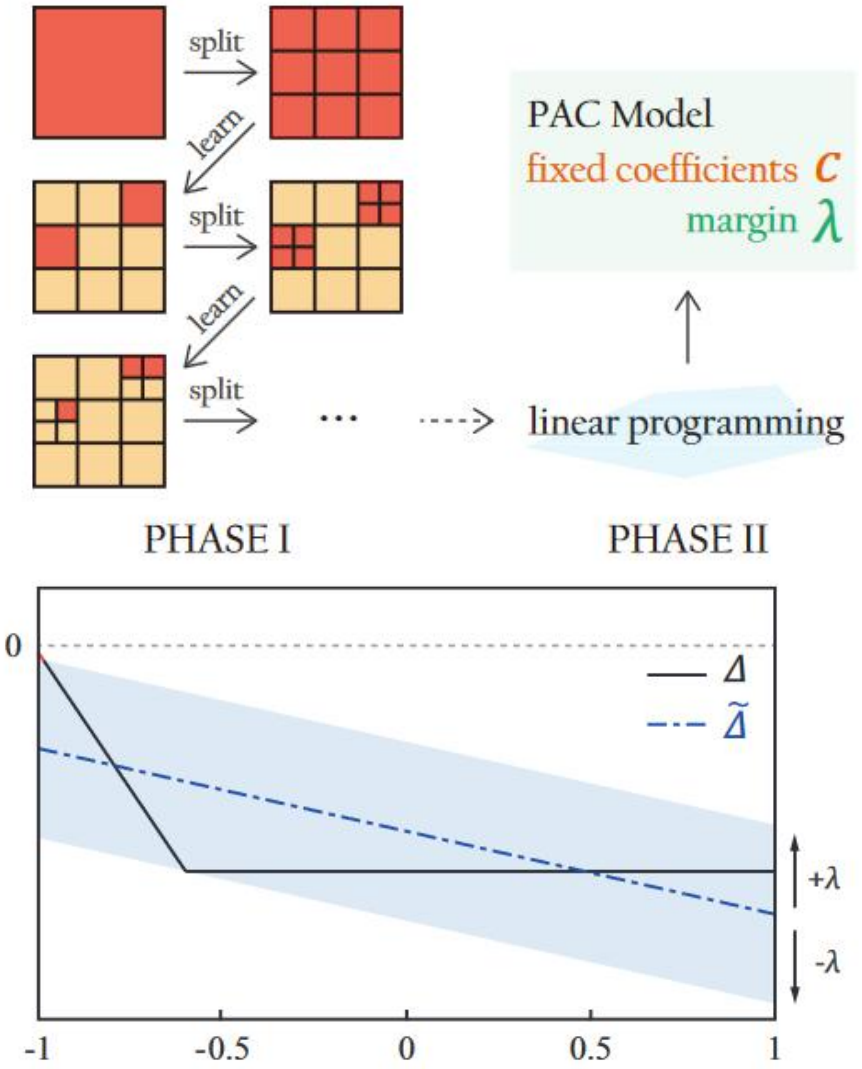


算法原理 DeepPAC

- 线性规划问题
  - 通过在输入点附近采样构建约束条件
  - 聚焦学习
    - 选出系数绝对值较大的特征
    - 其他特征共享系数
  - 多轮次重复过程减少计算量
- 通过单调性计算线性函数的最大值

$$\widetilde{\Delta}(x)_{max} - \lambda \leq \Delta(x) \leq \widetilde{\Delta}(x)_{max} + \lambda < 0$$



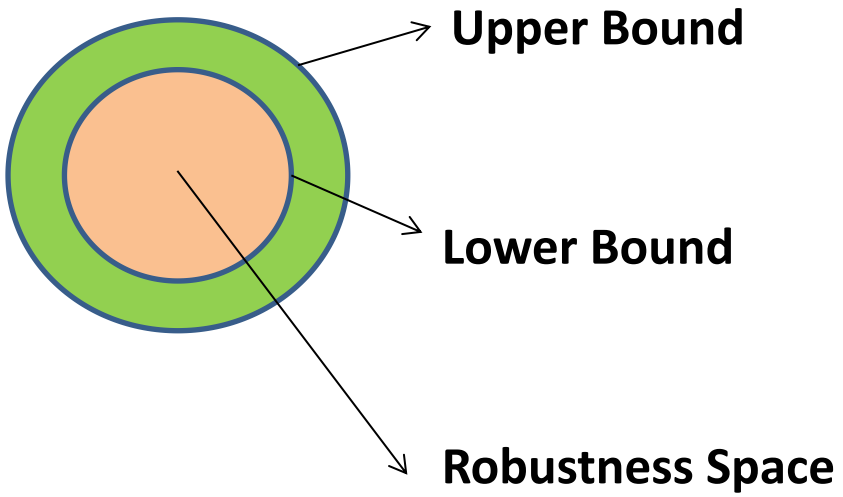


实验结果 DeepPAC

- 数据集

数据集	数据集简介
MNIST	黑白手写数字图片（10类）
CIFAR-10	物体分类彩色图片（10类）
ImageNet	最大图像识别数据库

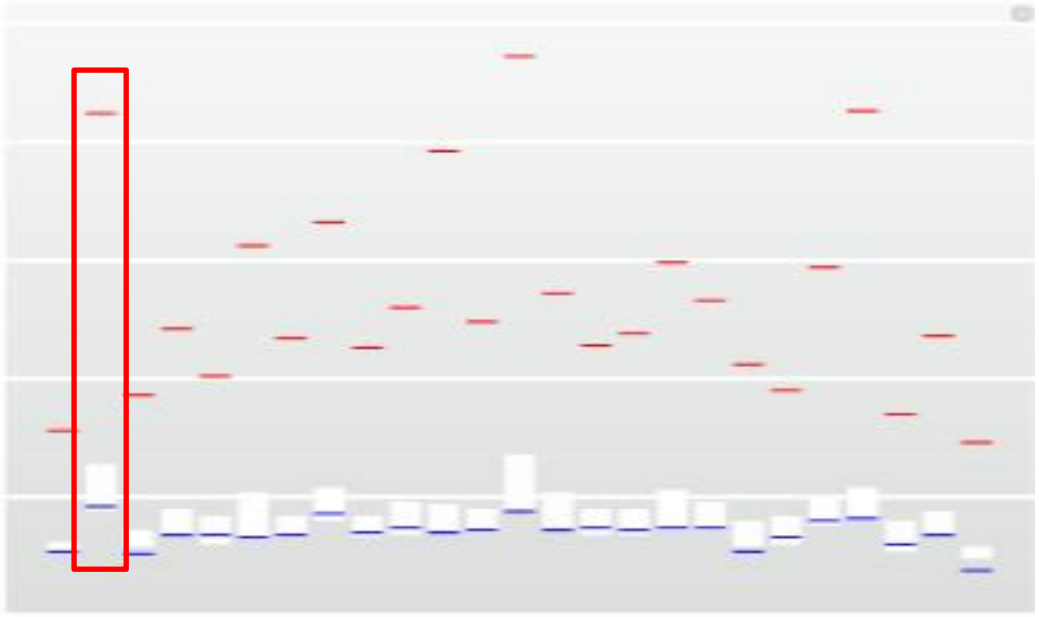
- 算法评价指标
  - ERAN：白盒方法计算鲁棒半径下界
  - PGD：黑盒方法估计鲁棒半径上界
- 鲁棒性指标
  - r：最大鲁棒半径



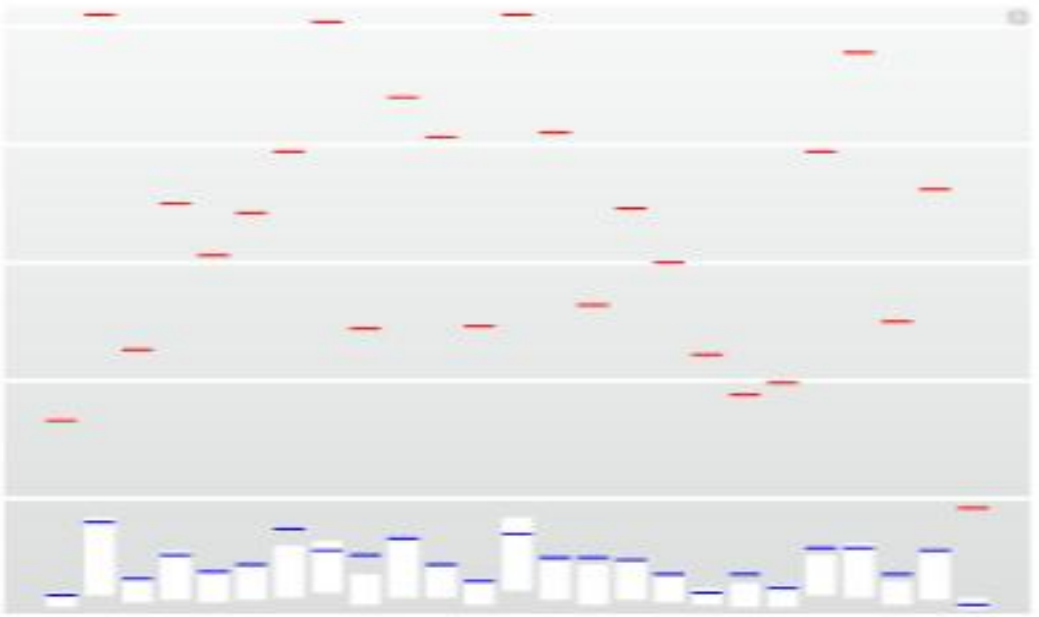
实验结果 DeepPAC

- 鲁棒半径估计的有效性
  - DeepPAC准确落入鲁棒区间内
  - DeepPAC能区分经对抗训练的模型

adversarial-trained



non-adversarial-trained



DeepPAC能有效估计鲁棒半径

实验结果 DeepPAC

- 不同参数对鲁棒半径的影响
  - 实验设置
    - 错误率ϵ，显著性水平η，第一阶段学习采样数$K^{(1)}$
  - 实验结论
    - 在错误率和显著性水平变化时，鲁棒半径保持稳定
    - 在第一阶段学习采样数减少时，线性函数精度下降，鲁棒半径会略微下降

DeepPAC随模型参数变化保持稳定

Input Image	Network	η, ϵ and $K^{(1)}$					
		0.01, 0.001		0.1, 0.001		0.01, 0.1	
		20K	5K	20K	5K	20K	5K
	ResNet18	5	4	5	4	5	4
	ResNet50	8	8	8	8	9	8
	ResNet152	5	5	5	5	5	5
	ResNet18	16	14	15	14	15	14
	ResNet50	12	11	12	12	12	11
	ResNet152	10	9	10	9	10	9
	ResNet18	11	10	11	10	11	10
	ResNet50	6	5	6	5	6	5
	ResNet152	9	8	9	8	9	8
	ResNet18	1	1	1	1	1	1
	ResNet50	3	3	3	3	3	3
	ResNet152	6	5	6	5	6	5
	ResNet18	16	13	16	14	16	14
	ResNet50	17	15	17	15	17	15
	ResNet152	12	10	12	10	12	10

实验结果 DeepPAC

- 鲁棒半径与测试样本优先级的关系
  - 与测试样本Gini系数计算相关性
  - 低置信度的样本鲁棒半径越低

DeepGini与DeepPAC方法有相关性

Network	DeepPAC	ERAN
FNN1	-0.3628	-0.3437
FNN2	-0.4851	-0.4353
FNN3	-0.4174	-0.3677
FNN4	-0.5264	-0.4722
FNN5	-0.4465	-0.6016
FNN6	-0.4538	-0.2747
CNN1	-0.7340	-0.7345
CNN2	-0.6482	-0.6478
CNN3	-0.7216	-0.6728
CNN4	-0.6035	-0.6127
CNN5	-0.7448	-0.6833
CNN6	-0.6498	-0.6094

算法原理 DeepPAC

- 算法总结
  - 在输入点周围采样生成输入输出对，计算得分差函数Δ
  - 利用这些采样点学习得到原模型的PAC模型
  - 利用线性规划和聚焦学习的方法减少的计算量
- 算法优势
  - 黑盒方法，计算过程不使用模型结构信息
  - 减少计算量，可以应用在高维输入场景下
- 思考改进
  - 学习更复杂的PAC模型，而不是简单的线性函数
  - 与测试样本排序领域联动，提升模型的鲁棒半径

应用总结

应用总结

- ROBY
  - 基于类内和类间的统计特征
  - 不涉及模型内部结构，算法简单，可扩展到其他任务上
- DeepPAC
  - 在可接受的错误概率内近似模仿原模型的行为
  - 黑盒方法，计算过程不使用模型结构信息，受参数影响小，可扩展性高
- 未来发展
  - 深度学习模型结构越来越复杂，基于黑盒的方法成为主流
  - 与对抗样本攻击方法相互映照，攻防对抗相反相成

深入贯彻复杂问题简单化的思想
遇见困难先接受再思考，分而治之不断细化去解决

参考文献

[1] Haibo Jin, Jinyin Chen, et al. ROBY: Evaluating the adversarial robustness of a deep model by its decision boundaries[J]. Information Sciences, 2022: 97-122

[2] Renjue Li, Pengfei Yang, Cheng-Chao Huang, et al. Towards practical robustness analysis for DNNs based on PAC-model learning. Proceedings of the 44th International Conference on Software Engineering[C]. Pittsburgh, PA, USA: ACM/IEEE, 2022: 2189–2201

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。

谢谢！