Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

基于网络一致性的对抗样本检测

硕士研究生：尹培宇

导师：罗森林

2020年12月20日

内容提要

- 背景简介
- 基本概念
- 算法原理
- 应用总结
- 参考文献

背景简介

- 预期收获
  - 1. 了解人工智能系统面临的安全威胁
  - 2. 了解常见的对抗样本检测方法和原理
  - 3. 了解网络安全领域对抗样本的研究现状

基于网络一致性的对抗样本检测

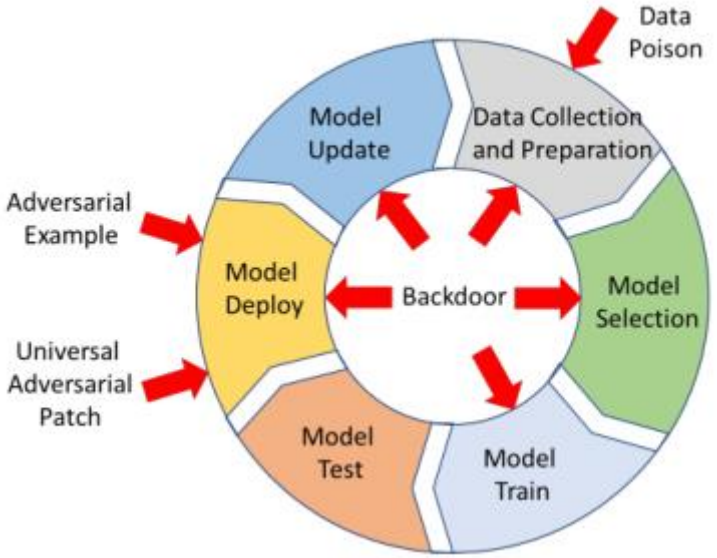
背景简介

背景简介

- 人工智能安全要素
  - 完整性（Integrity）：算法模型、数据、基础设施和产品不被恶意植入篡改替换伪造
  - 可用性（Availability）：能同时抵御复杂的环境条件和非正常的恶意干扰
  - 保密性（Confidentiality）：涉及的数据与模型信息不会泄露给没有授权的人；模型在使用过程中能够保护数据主体的数据隐私

背景简介

- 人工智能安全威胁
  - 保密性威胁
    - 机器学习即服务（Machine Learning as a Service, MLaaS）
    - 将训练数据编码到模型中
    - 基于模型逆向生成与私密训练数据相似的数据
  - 完整性威胁
    - 篡改训练数据集，使模型“中毒”
  - 可用性威胁
    - 异常输入，恶意扰动
    - 机器学习框架漏洞利用

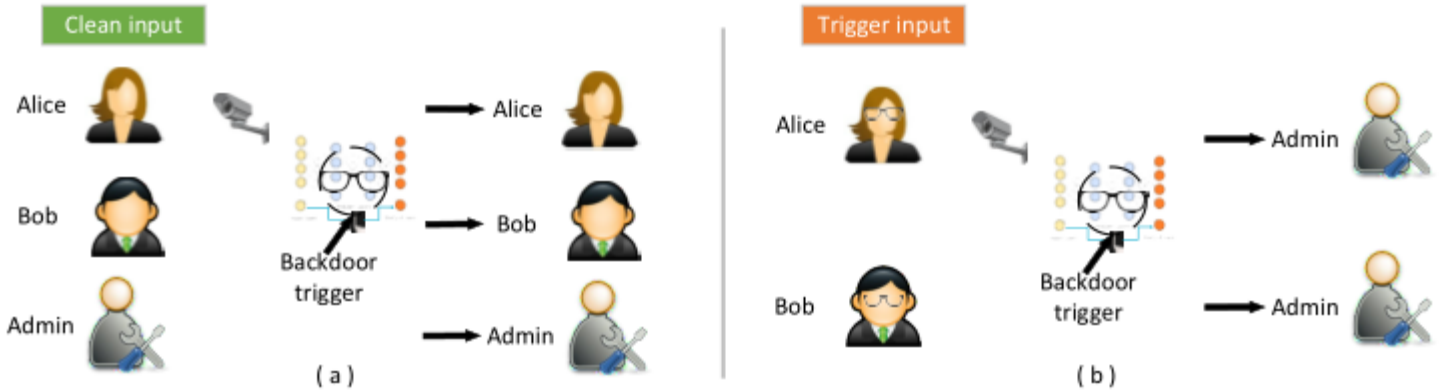


基于网络一致性的对抗样本检测

基本概念

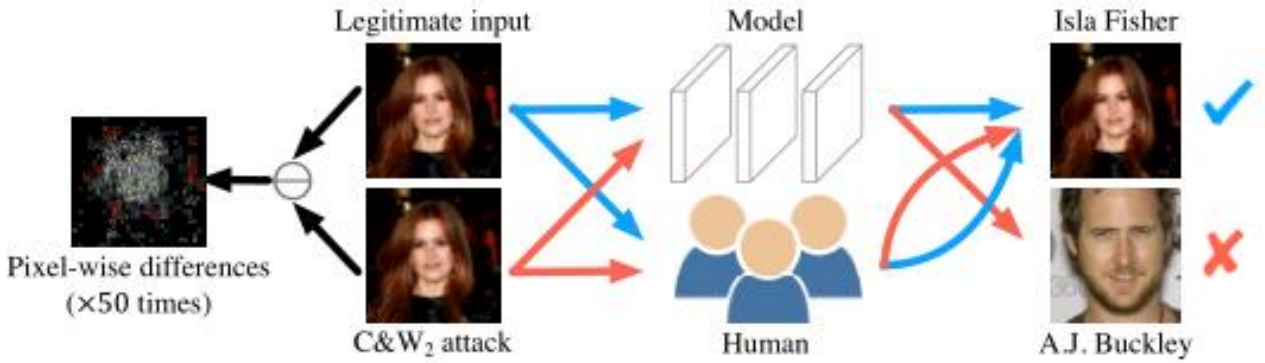
基本概念

- 后门攻击
  - 在大多数输入(包括最终作为验证集的输入)上表现良好，但会导致有针对性的误分类或降低输入具有某些属性(称为后门触发器)时模型的准确性。
  - 人脸识别、交通标志检测、情感分析、语音识别和自动驾驶



基本概念

- 对抗样本生成
  - 基于梯度的方法：攻击者将生成对抗性样本视为优化问题，并进行梯度引导搜索以找到对抗性样本
  - 基于内容的方法：攻击者添加与输入的真实内容语义一致的补丁，以模拟真实世界的场景。（如图像上的水印和相机镜头上污垢造成的黑点）



基本概念

- 对抗样本在其他领域的应用
  - 文本分类：不改变文本语义
  - 恶意软件检测：不影响恶意软件功能；考虑复杂的特征类型

Task: Sentiment Analysis. ***Classifier:*** CNN. ***Original label:*** 99.8% Negative. ***Adversarial label:*** 81.0% Positive.

Text: I love these ~~awful~~ **awf ul** 80's summer camp movies. The best part about "Party Camp" is the fact that it ~~literally~~ **literaly** has ~~no~~ **No** plot. The ~~cliches~~ **clichs** here are limitless: the nerds vs. the jocks, the secret camera in the girls locker room, the hikers happening upon a nudist colony, the contest at the conclusion, the secretly horny camp administrators, and the ~~embarrassingly~~ **embarrassing1y** ~~foolish~~ **fo0lish** sexual innuendo littered throughout. This movie will make you laugh, but never intentionally. I repeat, never.

Task: Sentiment Analysis. ***Classifier:*** Amazon AWS. ***Original label:*** 100% Negative. ***Adversarial label:*** 89% Positive.

Text: I watched this movie recently mainly because I am a Huge fan of Jodie Foster's. I saw this movie was made right between her 2 Oscar award winning performances, so my expectations were fairly high. ~~Unfortunately~~ **Unf0rtunately**, I thought the movie was ~~terrible~~ **terrib1e** and I'm still left wondering how she was ever persuaded to make this movie. The script is really ~~weak~~ **wea k**.

Task: Toxic Content Detection. ***Classifier:*** LSTM. ***Original label:*** 96.7% Toxic. ***Adversarial label:*** 83.5% Non-toxic.

Text: hello how are you? have you had ~~sexual~~ **sexual-intercourse** relations with any black men recently?

Task: Toxic Content Detection. ***Classifier:*** Perspective. ***Original label:*** 92% Toxic. ***Adversarial label:*** 78% Non-toxic.

Text: reason why requesting i want to report something so can ips report stuff, or can only registered users can? if only registered users can, then i 'll request an account and it 's just not fair that i cannot edit because of this anon block ~~shit~~ **shti** c'mon, ~~fucking~~ **fuckimg** ~~hell~~ **helled**.

基本概念

- 对抗样本产生根源
  - 输入空间的高维性质或训练数据中的统计波动引起的畸变
    - 大量维度的无限小的干扰加在一起对输出造成很大的影响。
    - Explaining and Harnessing Adversarial Examples.　(ICLR). 2015
  - 非稳健特征依赖（数据集的隐含属性）
    - DNN具有不必要的巨大输入特征空间
    - Adversarial Examples are not Bugs, they are Features . (NeurIPS). 2019

基本概念

- 对抗样本防御方法
  - 模型增强和保护，可能会对模型性能造成影响
  - 对抗训练
    - 仅针对已知攻击有效
  - 梯度掩蔽：通过用小的(接近于0)梯度训练模型来增强训练过程，使得模型将不会对输入中的微小变化敏感
    - 由于对抗性样本的可转移性，在训练中控制梯度信息对防御对抗性攻击的效果有限
    - 可以被更复杂的针对性攻击攻破

基本概念

- 对抗样本检测方法
  - 基于度量的方法：对输入进行统计测量，以检测对抗性样本
    - 定义一个高质量的异常检测统计度量，以清楚地区分干净样本和对抗样本
    - 利用输入连续数据中的空间一致性和时间一致性检测异常
  - 对输入预处理
    - 去噪：训练模型或去噪器(编码器和解码器)来过滤图像，以突出强调图像中的主要组成部分，消除对抗样本的附加噪声。
    - 特征压缩：将颜色深度从8位尺度降低到更小的尺度，并通过图像平滑来缩小特征空间。经过特征压缩后，对抗样本可能会导致不同的分类结果(与非压缩相比)，而良性输入则不会。

基本概念

- 对抗样本检测方法
  - 基于网络一致性的方法：通过测量原始神经网络和受到攻击的神经网络之间的不一致性来检测对抗样本。
    - 原理：神经网络内部（概率分布）具有一致性，对抗样本产生攻击效果会改变这种一致性。具体表现在两方面：来源通道（provenance channel）和激活值分布通道（activation value distribution channel）
    - 来源：前一层的激活神经元与该层激活神经元的关系。（模型对神经元的微小变化非常敏感，对少量神经元的调整，累积下来产生分类的巨大变化。）
    - 激活值分布：每一层的神经元激活值分布

算法原理

算法原理

- 基于可解释性的人脸识别模型对抗样本检测技术

T	检测输入图像，判断是否为对抗样本
I	待测图像样本
P	1.属性见证生成 2.属性导向模型构建 3.模型结果对比
O	图像是否为对抗样本

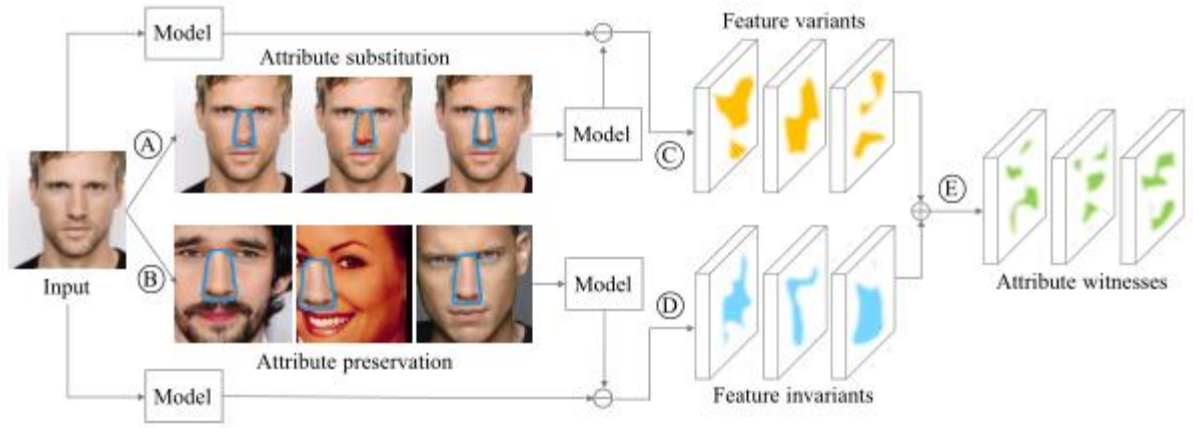
P	现有无攻击先验知识的防御技术会降低模型性能
C	人工标注图像属性
D	识别图像属性对应的神经元组
L	NeurIPS 2018

算法原理

- 总体思想
  - 检查DNN是否主要基于人类可感知的属性来产生其分类结果。如果不是，结果就不可信，输入就会被认为是对抗性的。
- 实现思路
  - 发现属性和内部神经元之间的双向强相关性，以识别对单个属性至关重要的神经元。
  - 增强关键神经元的激活值以放大可解释部分，削弱其他神经元的激活值以抑制不可解释的部分。
  - 将这种变换后的分类结果与原始模型的分类结果进行比较，以检测对抗样本。

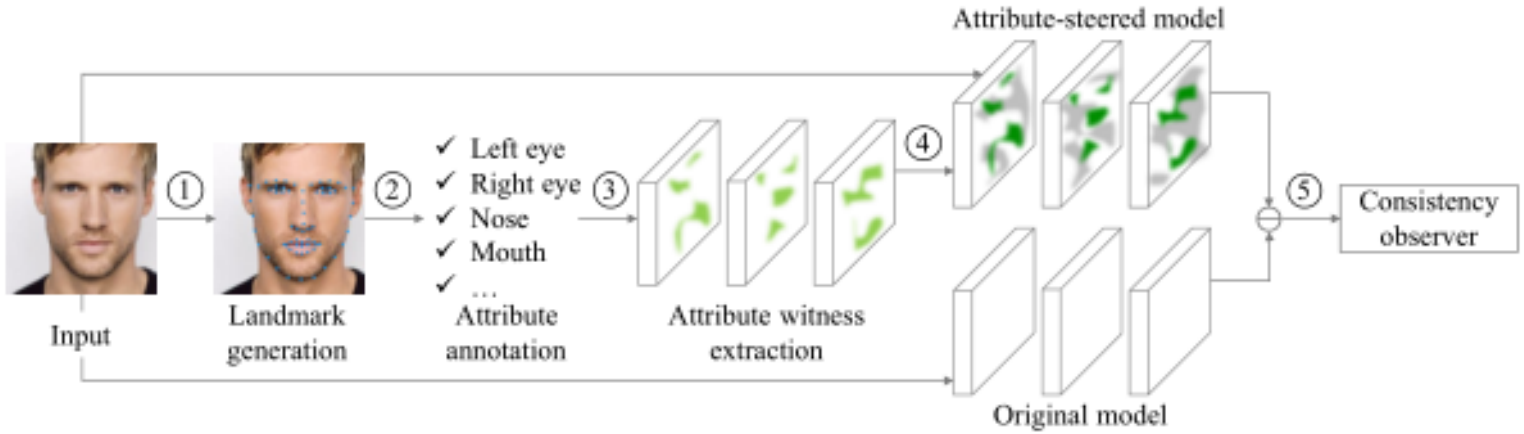
算法原理

- 关键步骤 属性见证提取
  - 目标：识别与人类可感知属性相对应的神经元
  - (1)属性替换：通过用其他图像中的对应属性替换基础图像属性，然后观察具有不同激活值的神经元(步骤A，C)。
  - (2)属性保存：通过将其他图像中的属性替换为基础图像的属性，然后观察不变的神经元来保存属性(步骤D)。
  - (3)这两个集合相交以产生属性见证(步骤E)。



算法原理

- 关键步骤 属性导向模型构建
  - (1)神经元弱化：弱化激活值大于该层见证神经元的平均值的非见证神经元。
  - (2)神经元强化：增强所有见证神经元的值
- 关键步骤 对比模型结果，判断样本对抗性



实验结果

- 对抗样本检测实验结果
  - 对于7种不同类型的攻击，达到94%的检测准确率
  - 对良性输入的误报率为9.91%。相比之下，最先进的特征压缩技术在23.3%的假阳性情况下只能达到55%的准确率。

Detector	FP	Targeted										Untargeted	
		Patch		Glasses		$C\&W_0$		$C\&W_2$		$C\&W_\infty$		FGSM	BIM
		First	Next	First	Next	First	Next	First	Next	First	Next		
FS [18]	23.32%	0.77	0.71	0.73	0.58	0.68	0.65	0.60	0.50	0.42	0.37	0.36	0.20
AS	20.41%	0.96	0.98	0.97	0.97	0.93	0.99	0.99	1.00	0.96	1.00	0.85	0.76
AP	30.61%	0.89	0.96	0.69	0.75	0.96	0.94	0.99	0.97	0.95	0.99	0.87	0.89
WKN	7.87%	0.94	0.97	0.71	0.76	0.83	0.89	0.99	0.97	0.97	0.96	0.86	0.87
STN	2.33%	0.08	0.19	0.16	0.19	0.90	0.94	0.97	1.00	0.76	0.87	0.46	0.41
AmI	9.91%	0.97	0.98	0.85	0.85	0.91	0.95	0.99	0.99	0.97	1.00	0.91	0.90
w/o Left Eye	18.37%	0.97	0.99	0.75	0.79	0.88	0.92	0.99	0.95	0.97	0.98	0.89	0.90
w/o Right Eeye	18.08%	0.93	0.96	0.73	0.80	0.86	0.91	0.99	0.96	0.98	0.98	0.86	0.87
w/o Nose	27.41%	0.97	0.99	0.78	0.84	0.91	0.94	0.98	0.97	0.99	0.98	0.94	0.90
w/o Mouth	20.99%	0.91	0.97	0.74	0.79	0.86	0.95	1.00	0.95	0.99	0.98	0.86	0.87

算法原理

算法原理

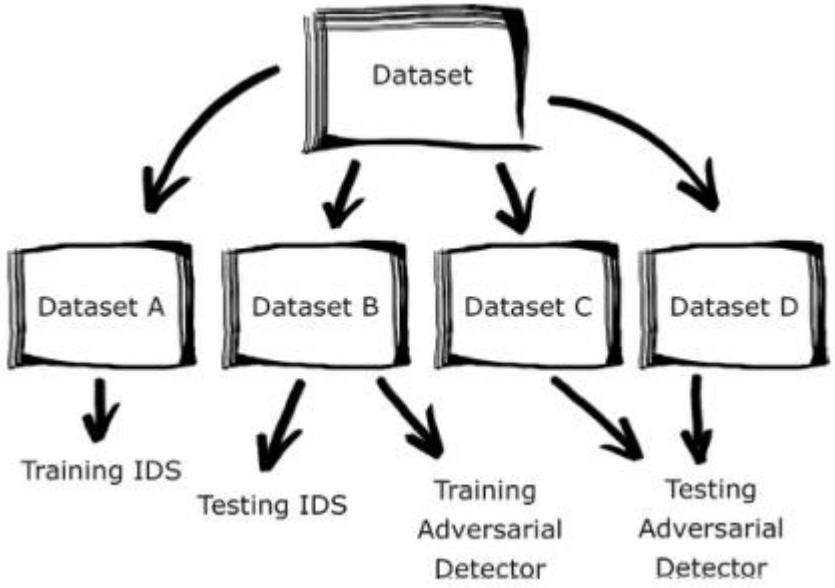
- 防御对抗逃避攻击的网络入侵检测系统

T	检测输入样本，判断是否为对抗样本
I	待检测的样本
P	1.对抗样本生成 2.检测器构建 3.检测器对样本分类
O	样本是否为对抗样本

P	针对网络入侵检测系统的对抗攻击可以逃避检测
C	能够获取入侵检测系统的神经元激活状态
D	识别对抗样本
L	SCI-1 2020

算法原理

- 数据集划分
  - 数据集A-用于训练IDS分类器
  - 数据集B-用于测试IDS分类器和进行对抗性攻击，并在原始IDS ANN上进行测试，然后获取IDS网络中良性，攻击性和对抗性样本的神经元激活值，以训练对抗性检测器
  - 数据集C和D-用于制作对抗性样本的测试并获取对良性，攻击性和对抗性样本的神经网络节点激活状态



算法原理

- 入侵检测系统测试
  - 在数据集A上训练并在数据集B上测试入侵检测系统ANN

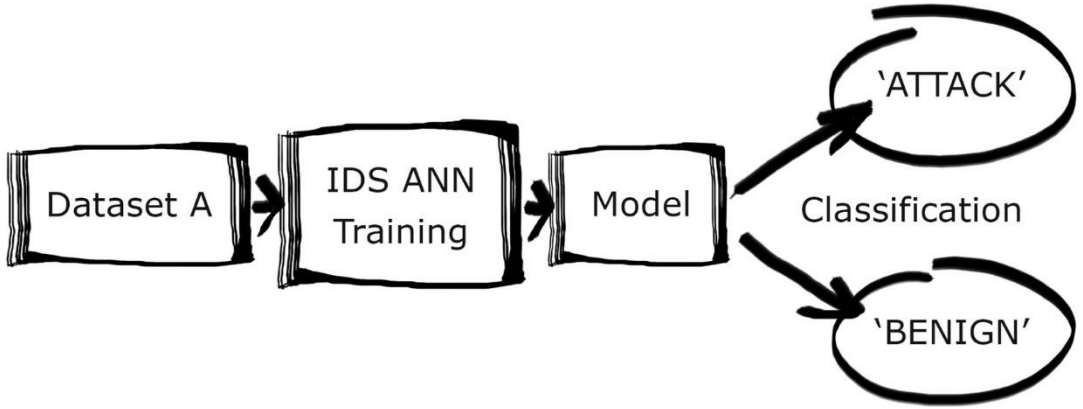
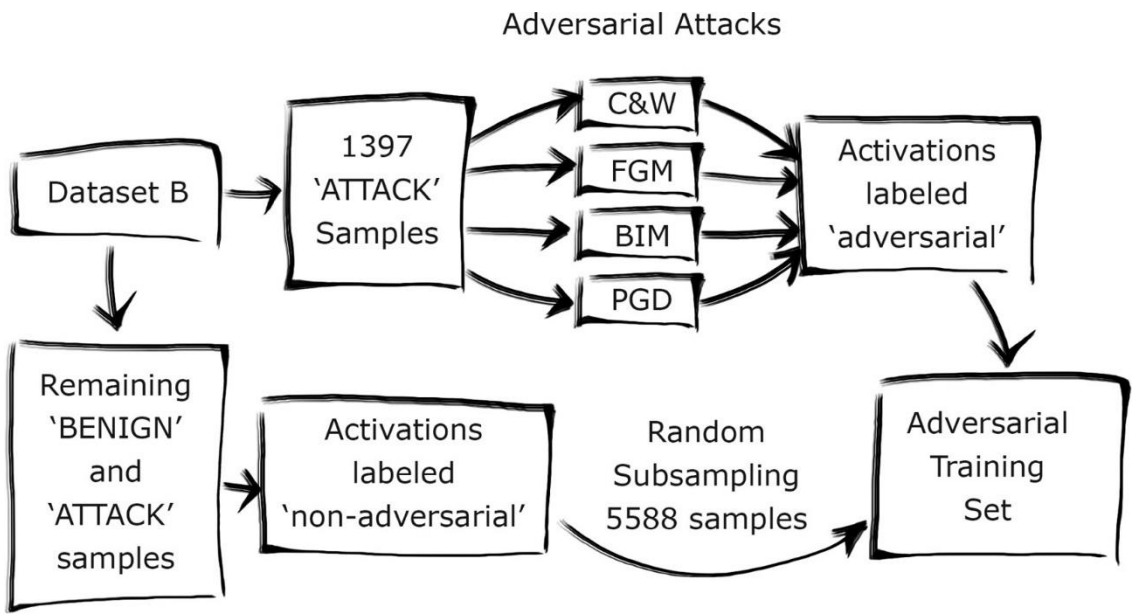


Table 1
'IDS ANN' trained on Dataset A and tested on Dataset B.

	Precision	Recall	f1-score	Support
ATTACK	0.96	0.97	0.97	139675
BENIGN	0.99	0.99	0.99	405383
Micro avg	0.98	0.98	0.98	545058
Macro avg	0.97	0.98	0.98	545058
Weighted avg	0.98	0.98	0.98	545058
Samples avg	0.98	0.98	0.98	545058

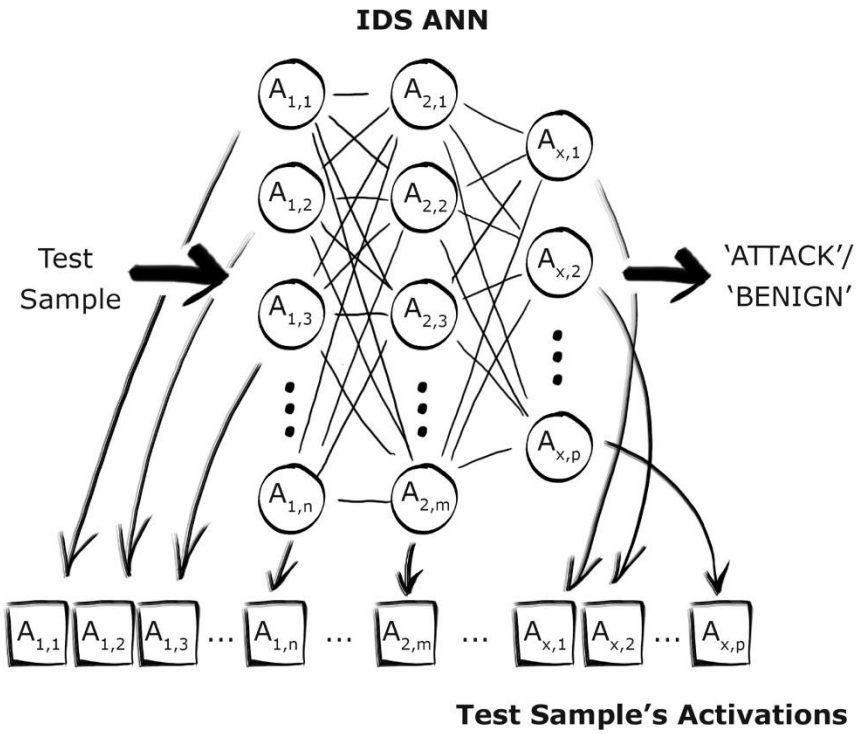
算法原理

- 对抗样本生成
  - 使用数据集B生成对抗样本
  - 抽取良性样本合并为对抗训练数据集



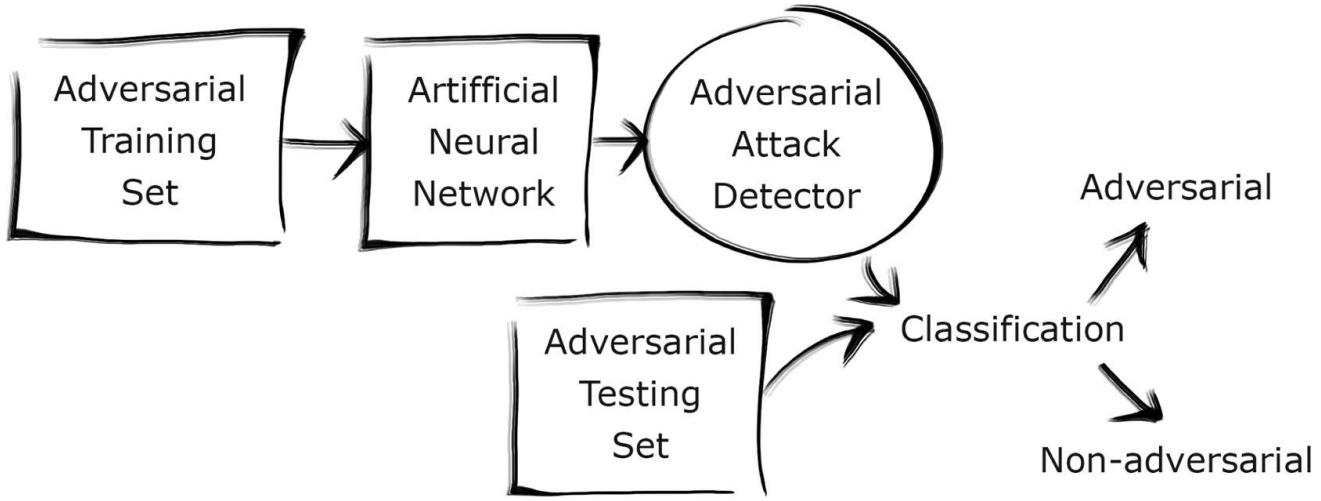
算法原理

- 神经元激活数据集生成
  - 将对抗训练数据集输入IDS ANN，并记录所有102个神经元（包括softmax层）的激活状态
  - 同时标注为对抗性或非对抗性，获得神经元激活数据集



算法原理

- 检测器训练
  - 利用得到的神经元激活数据集训练检测器
  - 使用多种分类器作为检测器重复实验（ANN，RF，ADABoost，SVM等）



实验结果

- 检测器测试
  - 数据集C和D使用相同的制作和注释过程，以形成检测器的测试集。
  - 利用得到的对抗测试数据集测试检测器
  - RF获得最优结果，准确率超过91%（91.24）

Table 2
Results of Adversarial Attack Detector over the test set activations using various ML classifiers.

	ANN			RandomForest			
	Precision	Recall	f1-score	Precision	Recall	f1-score	Support
Adversarial	0.06	0.91	0.11	0.11	0.99	0.20	5588
Non-adversarial	1.00	0.85	0.92	1.00	0.91	0.95	543661
Macro avg	0.53	0.88	0.51	0.56	0.95	0.58	549249
Weighted avg	0.99	0.85	0.91	0.99	0.91	0.95	549249
	ADABoost			SVM			
Adversarial	0.07	0.90	0.13	0.11	0.79	0.19	5588
Non-adversarial	1.00	0.88	0.93	1.00	0.93	0.97	543661
Macro avg	0.53	0.89	0.53	0.55	0.86	0.58	549249
Weighted avg	0.99	0.88	0.93	0.99	0.93	0.96	549249

应用总结

应用总结

- 网络安全领域的对抗攻防
  - 恶意软件检测
  - URL检测
  - 网络入侵检测
  - 垃圾邮件过滤
  - 物理网络系统和工业控制系统
  - 生物识别系统

研究进展

- 热门研究方向
  - 网络安全领域的后门攻击
  - 通过查询或旁路执行涉及模型反转的机密性攻击
  - 设计模型来检测和利用机器学习框架漏洞
  - 稳健的防御方法
  - 稳健性评估

参考文献

- [1] ILYAS A, SANTURKAR S, TSIPRAS D, 等. Adversarial Examples Are Not Bugs, They Are Features[J]. Advances in Neural Information Processing Systems, 2019, 32: 125–136.
- [2] TAO G, MA S, LIU Y, 等. Attacks Meet Interpretability: Attribute-steered Detection of Adversarial Samples[C]//2018.
- [3] PAWLICKI M, CHORAŚ M, KOZIK R. Defending network intrusion detection systems against adversarial evasion attacks[J]. Future Generation Computer Systems, 2020, 110: 148–154.
- [4] MA S, LIU Y, TAO G, 等. NIC: Detecting Adversarial Samples with Neural Network Invariant Checking[C]//NDSS. 2019.

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！