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

机器学习模型后门攻击检测

硕士研究生　尹培宇

2021年08月15日

内容提要

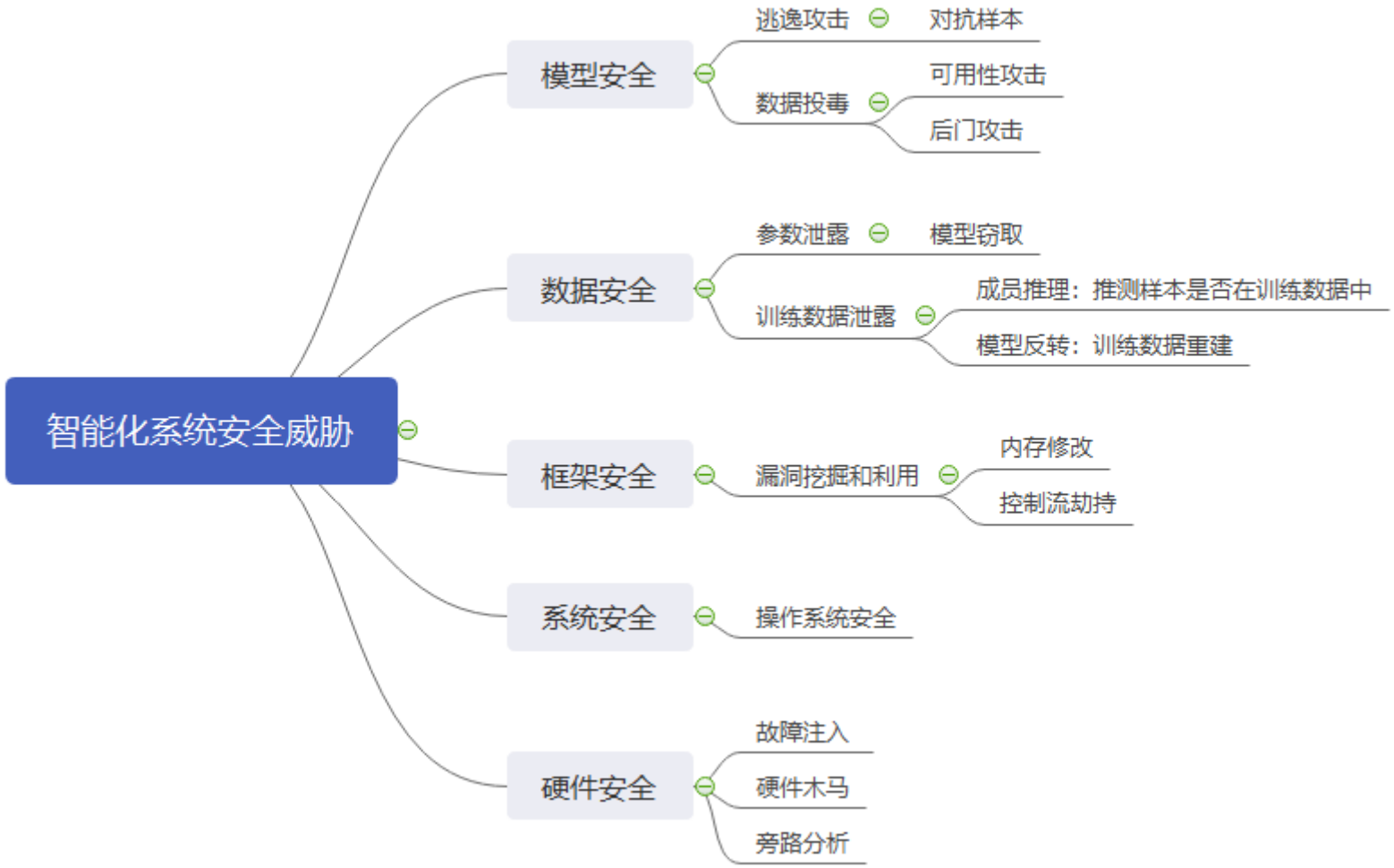
- 背景简介
- 基本概念
  - 数据投毒
  - 后门攻击
- 算法原理
- 应用总结
- 参考文献

预期收获

- 预期收获
  - 1. 了解智能化系统面临的安全威胁
  - 2. 了解后门攻击原理
  - 3. 了解后门攻击防御方法
  - 4. 了解后门攻击技术应用

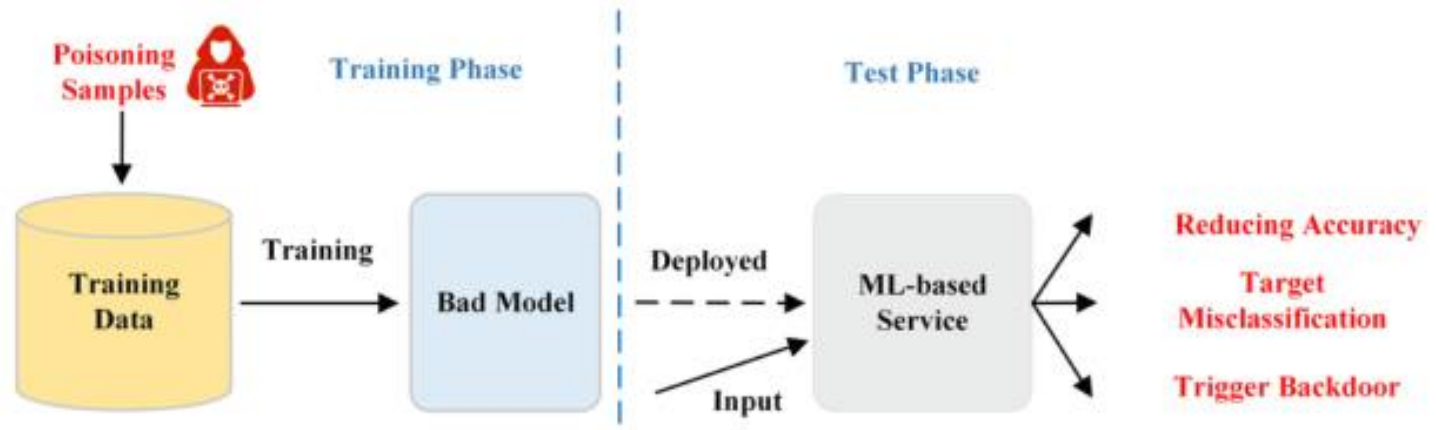
背景简介

- 智能化系统面临的安全威胁



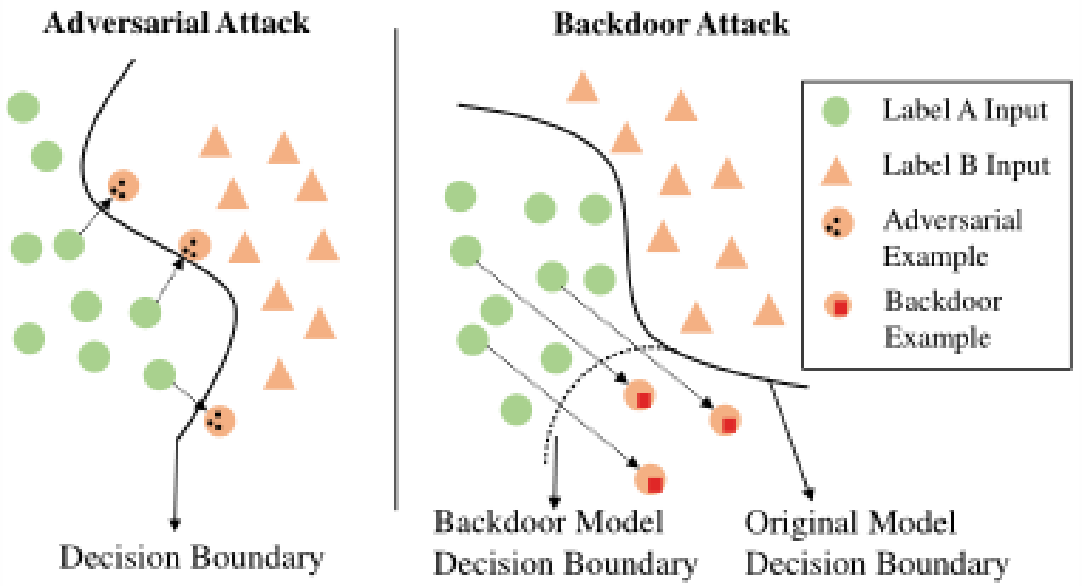
基本概念

- 数据投毒
  - 目标：使模型表现非预期行为（性能下降、嵌入后门）
  - 实现方法：修改训练数据集（人为构造样本，修改特征，修改标签），训练模型
- 后门攻击
  - 目标：在模型中嵌入后门，实现隐蔽的目标逃逸（触发器隐蔽、不影响干净样本）
  - 实现方法
    - 数据投毒（标签翻转，干净标签）
    - 模型参数修改
    - 内存修改
    - 插入后门模块



基本概念

- 对抗样本攻击与后门攻击的相似和区别
  - 样本角度：都需要对干净的输入进行扰动
    - 后门样本中的扰动：
      - 通用：对于属于所有类的所有样本，触发器是固定的
      - 较大：离决策边界更远，稳健性强
    - 对抗样本中的扰动：
      - 特定：一般特定于输入
      - 较小：离决策边界近，稳健性弱
  - 模型角度
    - 后门攻击需要修改模型参数
    - 对抗性攻击是被动的，不允许修改模型
  - 一句话总结
    - 后门攻击由触发器提供了一个由任意样本通往目标标签的稳定通道
    - 对抗样本为每个样本计算跨越决策边界的最短路径



基本概念

- 对抗样本攻击与后门攻击的关系和影响
  - 防御者角度
    - 对抗性训练提高了对抗健壮性，同时会降低后门健壮性
      - 对抗样本可以部分归因于真实数据集中存在的非健壮特征(高度预测性的、脆弱的和人类无法理解的特征)
      - 由于经过对抗训练的模型更多地依赖健壮的高级特征来进行预测，因此它也倾向于利用后门触发器进行预测（触发器提供了与目标标签紧密相关的健壮特征）
    - 可以利用对抗扰动来检测后门
      - 后门样本对对抗扰动的稳健性更强
      - 针对后门模型的对抗扰动具有相似性（或可转移性）
  - 攻击者角度：相互强化
    - 一种攻击可以显著增加另一种攻击的有效性

基本概念

- 后门攻击防御
  - 考虑攻击成功条件：样本触发器+模型后门
  - 防御思路：消除样本触发器影响、消除模型后门影响、拒绝后门样本/模型

算法原理 MNTD

算法原理

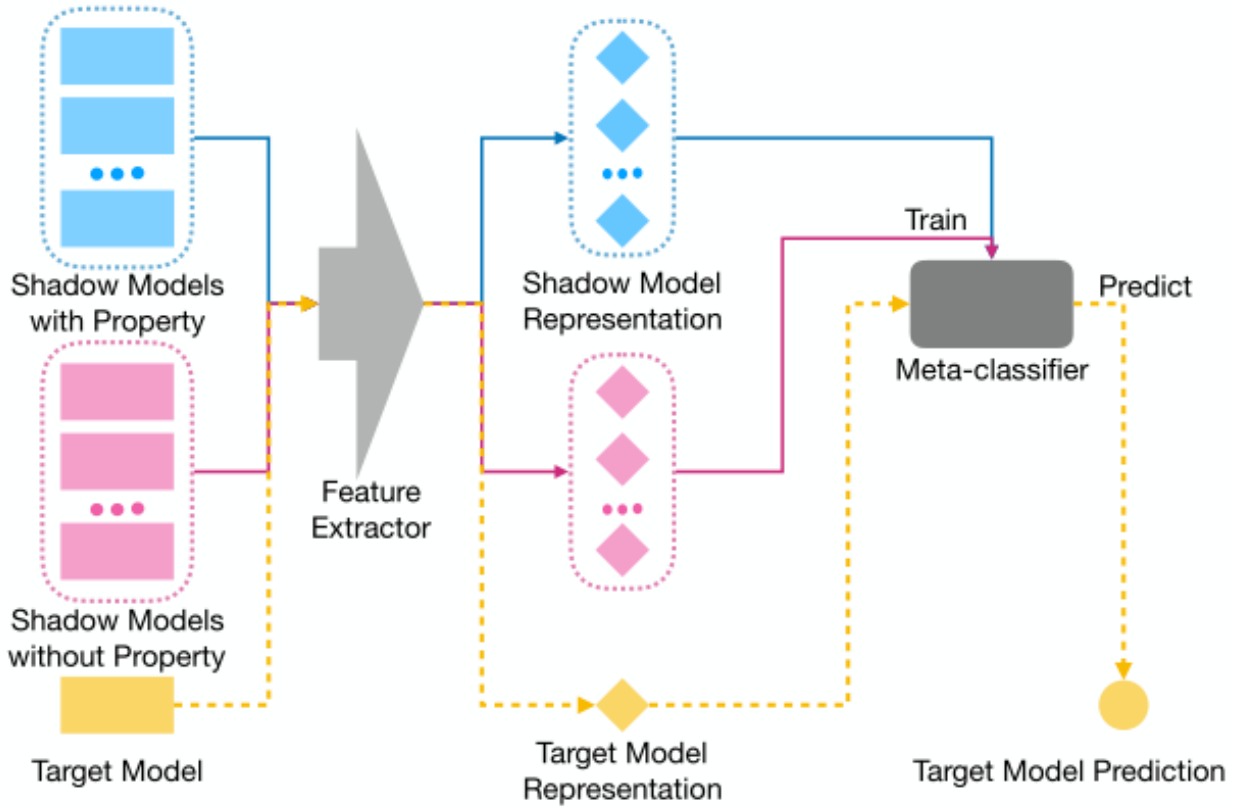
算法原理 MNTD

T	模型后门检测
I	部分干净数据+目标模型
P	1.训练影子模型 2.训练元模型（联合优化元分类器和查询集） 3.目标模型检测
O	目标模型判定结果（有/无后门）

P	黑盒条件下的模型后门检测
C	部分干净训练数据
D	提取影子模型的特征向量
L	IEEE S&P 2021

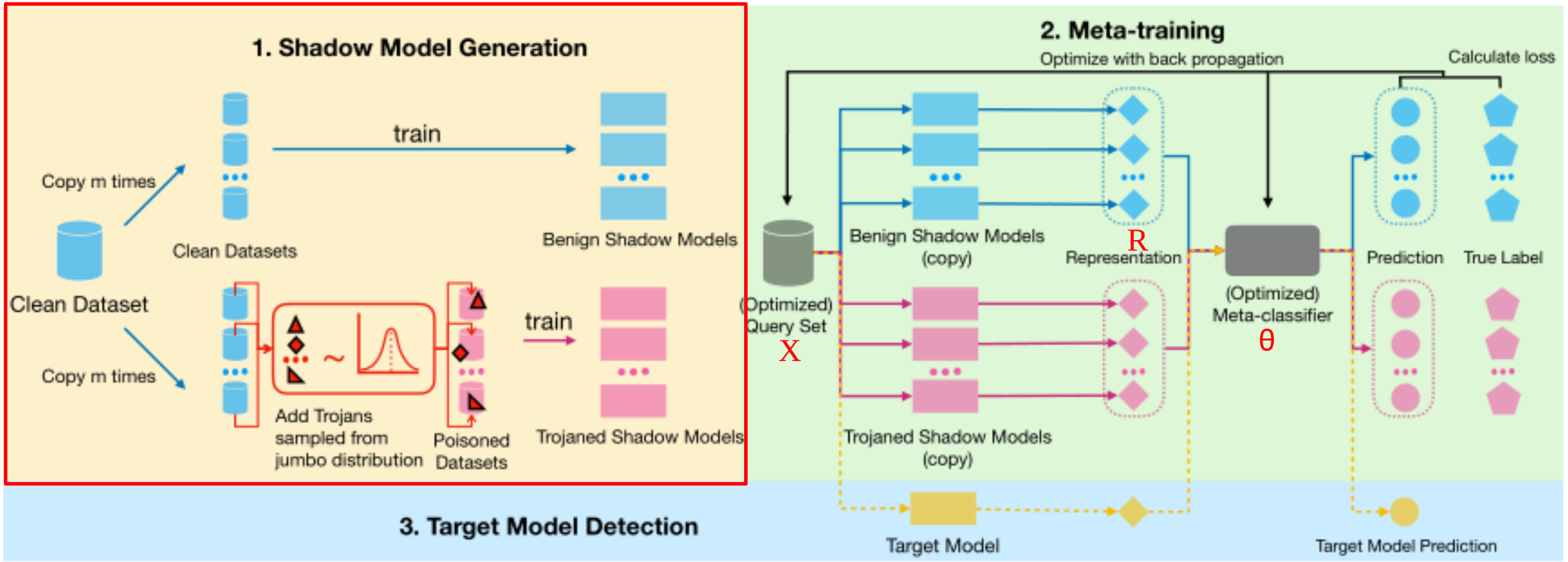
算法原理 工作流程

- 工作流程
  - 1. 训练多个影子模型以获得数据集
  - 2. 使用特征提取函数从每个影子模型中提取特征，得到元训练数据集
  - 3. 使用元训练数据集训练元分类器
  - 4. 提取目标模型特征，使用元分类器将其分类



算法原理 实现过程

- 实现过程
  - 1.影子模型生成：不同初始化参数，不同触发器设置



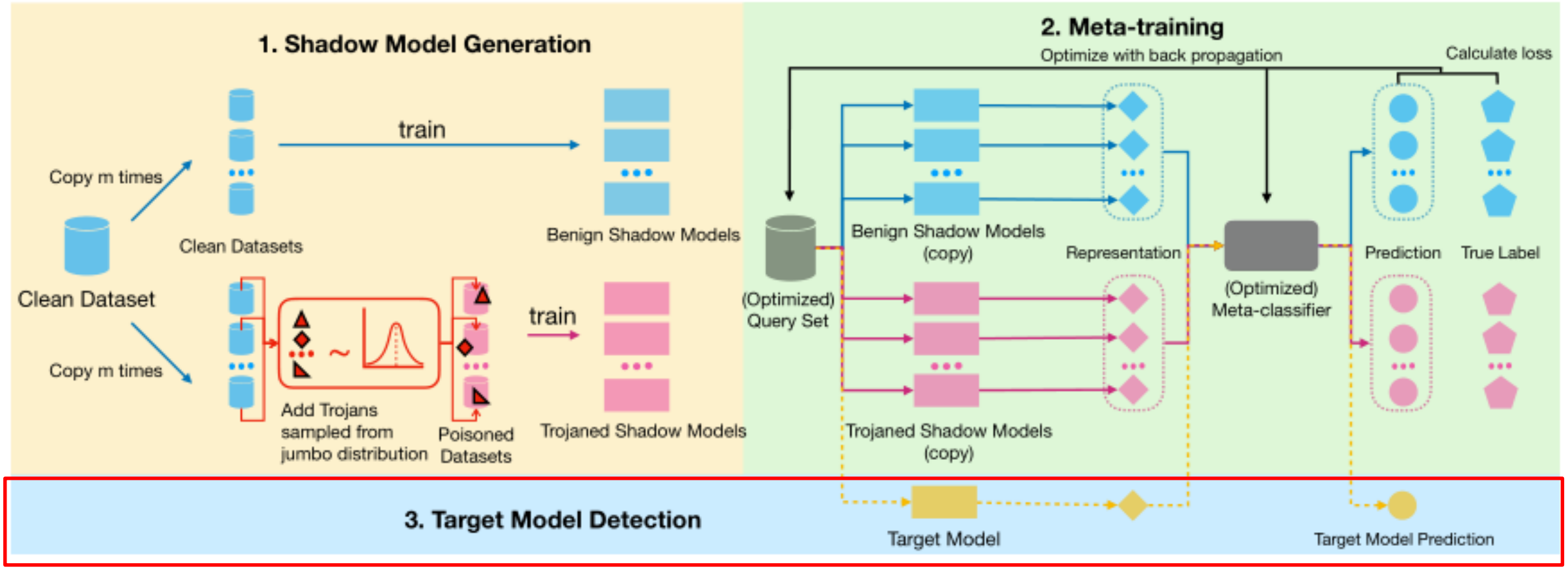
算法原理 实现过程

- 实现过程
  - 2.元训练
    - 目标：
      - 1)找到一个特征提取函数来提取影子模型的表示向量
      - 2)训练一个元分类器来区分良性影子模型和后门影子模型
    - 特征提取函数设计：将一组查询反馈给阴影模型，使用输出向量作为其特征(表示向量)
    - 元分类器设计：使用两层全连接的神经网络作为元分类器
    - 元训练算法：查询集X和元分类器参数θ的最优值
      - 1)随机选择查询集合X，预先计算所有表示向量Ri，然后优化元分类器
      - 2)联合优化查询集和元分类器，使训练损失最小

$$\arg\min_{\substack{\theta \\ X=\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}}} \sum_{i=1}^{m} L\Big(META(\mathcal{R}_i(X);\theta), b_i\Big)$$

算法原理 实现过程

- 实现过程
  - 3.目标模型检测



算法原理 实验结果

- 实验结果对比
  - 已知攻击类型的检测结果
    - 在不同类型攻击任务（图像、语音、表格、文本）上表现良好（平均检测AUC达到97%以上）
    - 时间效率：分别使用2048个干净和后门影子模型，基于RTX 2080进行离线训练，用时14小时，而对比方法用时在27秒到738秒之间

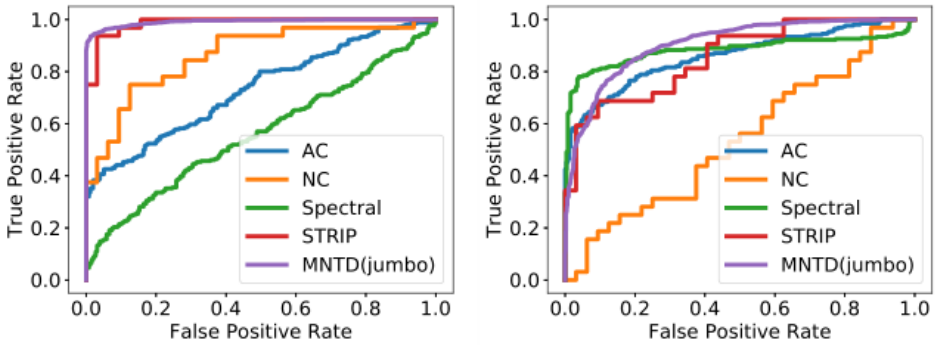


Fig. 8: The detection ROC curve of different approaches on MNIST-M (left) and CIFAR10-M (right).

TABLE III: The detection AUC of each approach. -M stands for modification attack and -B stands for blending attack.

Approach	MNIST-M	MNIST-B	CIFAR10-M	CIFAR10-B	SC-M	SC-B	Irish-M	Irish-B	MR-M
AC [12]	73.27%	78.61%	85.99%	74.62%	79.69%	82.86%	56.14%	93.48%	88.26%
NC [53]	92.43%	89.94%	53.71%	57.23%	91.21%	96.68%	✗	✗	✗
Spectral [52]	56.08%	≤50%	88.37%	58.64%	≤50%	≤50%	56.50%	≤50%	**95.70%**
STRIP [21]	85.06%	66.11%	85.55%	81.45%	89.84%	85.94%	≤50%	≤50%	✗
MNTD (One-class)	61.63%	≤50%	63.99%	73.77%	87.45%	85.91%	94.36%	**99.98%**	≤50%
MNTD (Jumbo)	**99.77%**	**99.99%**	**91.95%**	**95.45%**	**99.90%**	**99.83%**	**98.10%**	**99.98%**	89.23%

算法原理 实验结果

- 实验结果对比
  - 不可预见攻击的检测结果
    - 采用防御者未知的攻击方法生成目标后门模型，检测性能相似
    - 证明对不可预见攻击方法的泛化性良好

TABLE V: Examples of unforeseen Trojan trigger patterns and the detection AUC of jumbo MNTD on these Trojans.

Trojan Shape	MNIST			CIFAR-10		
	Pattern Mask	Trojaned Example	Detection AUC	Pattern mask	Trojaned Example	Detection AUC
Apple			96.73%			89.38%
Corners			98.74%			93.09%
Diagonal			99.80%			97.57%
Heart			99.01%			93.82%
Watermark			99.93%			97.32%

算法原理 实验结果

- 实验结果对比
  - 未知结构模型的检测结果
    - 针对每种目标模型，防御者使用其他5种结构分别训练64（32良性+32后门）个影子模型，检测AUC都在80%以上
    - 证明对不可预见的模型结构具有通用性

TABLE VIII: The detection AUC of MNTD on unforeseen model structures on ImageNet Dog-vs-Cat. The meta-classifier for each model structure are trained using all models except the ones in target model structure.

ResNet-18	ResNet-50	DenseNet-121
81.25%	83.98%	89.84%
DenseNet-169	MobileNet	GoogLeNet
82.03%	87.89%	85.94%

算法原理 实验结果

- 其他实验结果
  - 影子模型的数量对检测性能的影响
    - 影子模型数量越多，检测效果越好
  - 两种查询集生成方法结果对比
    - 查询集优化非常有效，如果使用随机选取的查询集，在最坏的情况下，AUC得分下降了30%

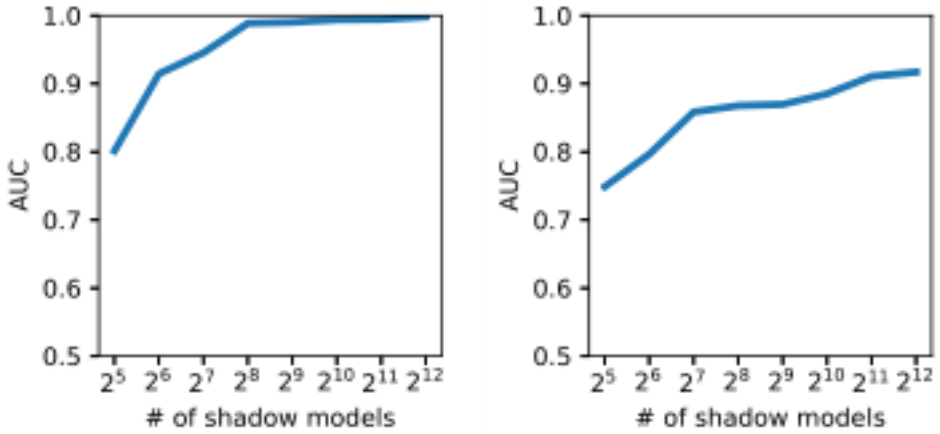


Fig. 6: Detection AUC with respect to the number of shadow models used to train the meta-classifier on MNIST-M (left) and CIFAR10-M (right).

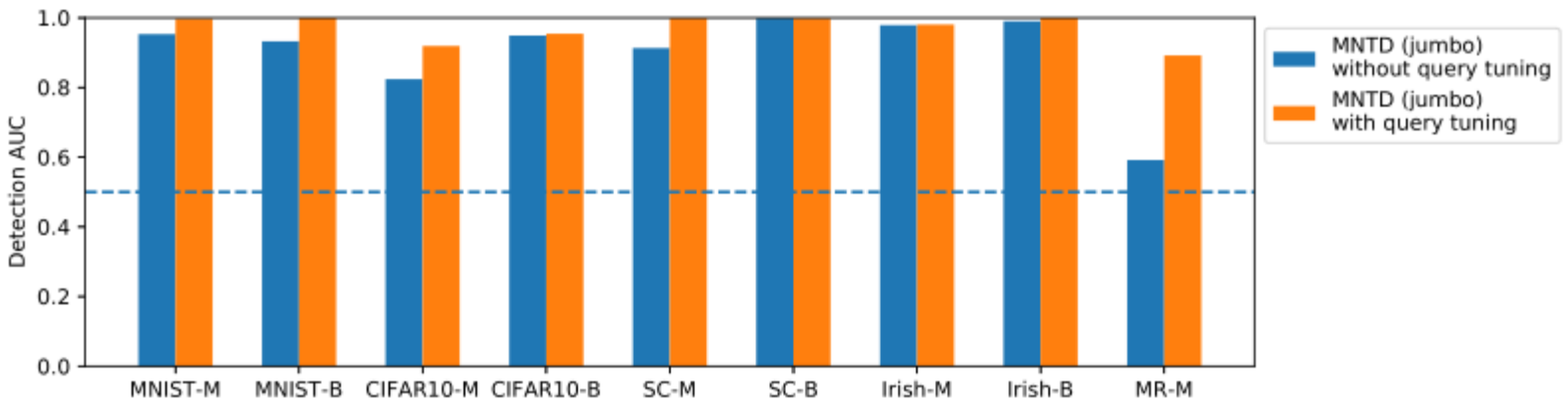


Fig. 9: The comparison of detection AUC with and without query tuning. -M stands for modification attack and -B stands for blending attack.

算法原理 优劣分析

- 优势
  - 黑盒模型级后门检测
  - 无需对攻击方法的假设，对不可知攻击的泛化性强

- 缺陷
  - 计算密集，训练影子模型的成本过高
  - 仅测试了简单或已知触发模式，缺乏对新型未知攻击模式的验证（干净标签攻击）
  - 对不同结构的模型效果下降

算法原理

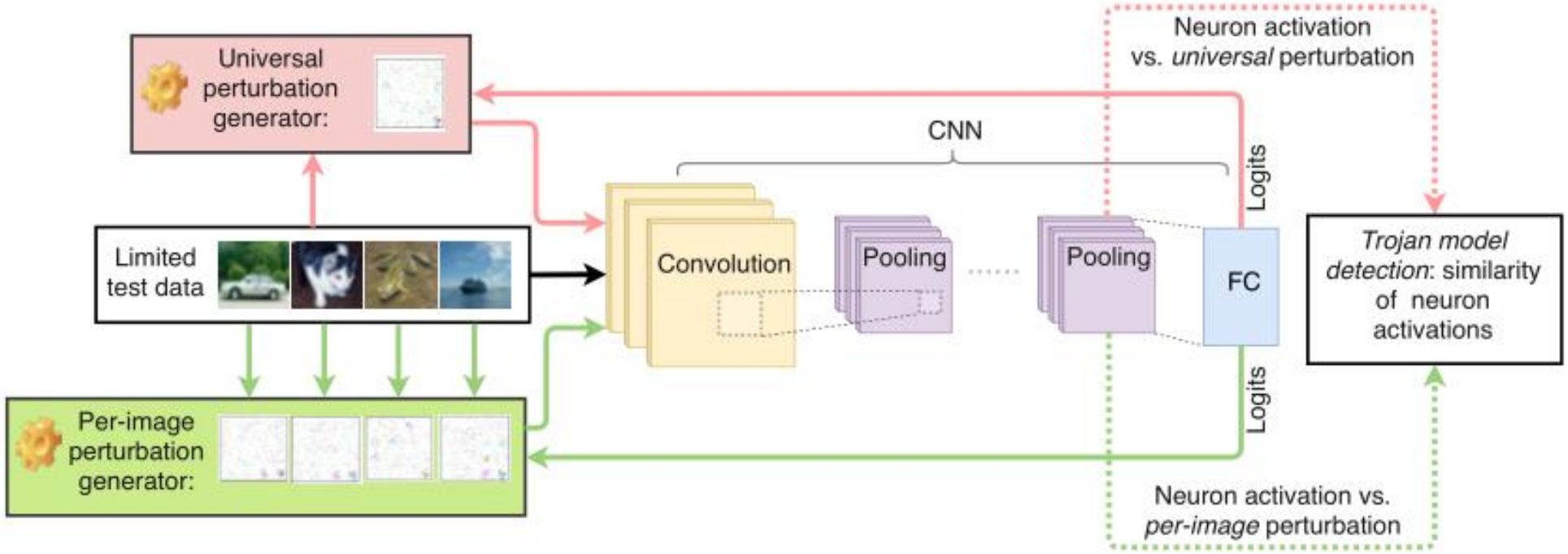
算法原理 DL-TND

T	模型后门检测
I	少量干净样本（每类至少1个）+目标模型
P	1.生成数据集通用对抗扰动 2.生成特定于样本的对抗扰动 3.获取扰动产生的激活向量 4.计算激活向量相似性，根据阈值判定模型是否有后门
O	目标模型判定结果（有/无后门）

P	数据受限条件下的模型后门检测
C	白盒模型，可以获取内部激活向量
D	计算通用对抗扰动和特定于样本的对抗扰动
L	ECCV(B) 2020

算法原理 DL-TND

- 原理：后门模型的对抗扰动具有可转移性
  - 对于后门模型，将不同样本预测为目标类所需的扰动具有相似性
- 方法：利用对不同样本施加扰动后激活向量的相似性



算法原理 DL-TND

- 方法：利用对不同样本施加扰动后激活向量的相似性
  - 1.对于数据集中的全体样本，计算改变预测结果所需的通用扰动

$$\begin{aligned}&\underset{\mathbf{m},\boldsymbol{\delta}}{\text{minimize}} \quad \ell_{\text{atk}}(\hat{\mathbf{x}}(\mathbf{m},\boldsymbol{\delta});\mathcal{D}_{k-}) + \bar{\ell}_{\text{atk}}(\hat{\mathbf{x}}(\mathbf{m},\boldsymbol{\delta});\mathcal{D}_{k}) + \lambda\|\mathbf{m}\|_1\\&\text{subject to } \{\boldsymbol{\delta},\mathbf{m}\} \in \mathcal{C},\end{aligned}$$

  - 其中，$\hat{\mathbf{x}}(\mathbf{m},\boldsymbol{\delta}) = (1-\mathbf{m})\cdot\mathbf{x} + \mathbf{m}\cdot\boldsymbol{\delta}$，不同损失项计算公式如下

$$\ell_{\text{atk}}(\hat{\mathbf{x}}(\mathbf{m},\boldsymbol{\delta});\mathcal{D}_{k-}) = \sum_{\mathbf{x}_i\in\mathcal{D}_{k-}} \max\{f_{y_i}(\hat{\mathbf{x}}_i(\mathbf{m},\boldsymbol{\delta})) - \max_{t\neq y_i} f_t(\hat{\mathbf{x}}_i(\mathbf{m},\boldsymbol{\delta})), -\tau\},$$

$$\bar{\ell}_{\text{atk}}(\hat{\mathbf{x}}(\mathbf{m},\boldsymbol{\delta});\mathcal{D}_{k}) = \sum_{\mathbf{x}_i\in\mathcal{D}_{k}} \max\{\max_{t\neq k} f_t(\hat{\mathbf{x}}_i(\mathbf{m},\boldsymbol{\delta})) - f_{y_i}(\hat{\mathbf{x}}_i(\mathbf{m},\boldsymbol{\delta})), -\tau\},$$

    - 其中$\tau \geq 0$是一个给定的常数，表征攻击信心（前项减小至$-\tau$代表攻击成功）

算法原理 DL-TND

- 方法：利用对不同样本施加扰动后激活向量的相似性
  - 2.对于每个样本分别计算将预测结果改变为目标类所需的扰动

$$\underset{\mathbf{m},\boldsymbol{\delta}}{\text{minimize}} \quad \ell'_{\text{atk}}(\hat{\mathbf{x}}(\mathbf{m},\boldsymbol{\delta});\mathbf{x}_i) + \lambda\|\mathbf{m}\|_1 \quad \text{subject to } \{\boldsymbol{\delta},\mathbf{m}\} \in \mathcal{C},$$

$$\ell'_{\text{atk}}(\hat{\mathbf{x}}(\mathbf{m},\boldsymbol{\delta});\mathbf{x}_i) = \sum_{\mathbf{x}_i \in \mathcal{D}_{k-}} \max\{\max_{t\neq k} f_t(\hat{\mathbf{x}}_i(\mathbf{m},\boldsymbol{\delta})) - f_k(\hat{\mathbf{x}}_i(\mathbf{m},\boldsymbol{\delta})), -\tau\}.$$

  - 3.将上述扰动分别输入目标模型，获取激活向量
  - 4.计算激活向量的平均相似性，若余弦相似度大于阈值，判定为后门模型

$$v_i^{(k)} = \cos\left(r(\hat{\mathbf{x}}_i(\mathbf{u}^{(k)})), r(\hat{\mathbf{x}}_i(\mathbf{s}^{(k,i)}))\right)$$

    - 其中r(·) 表示从输入图像到 CNN 中神经元表示（激活向量）的映射

算法原理 实验结果

- 相似度分布的可视化
- 不同参数（v(k)的Q百分位数）下的检测效果

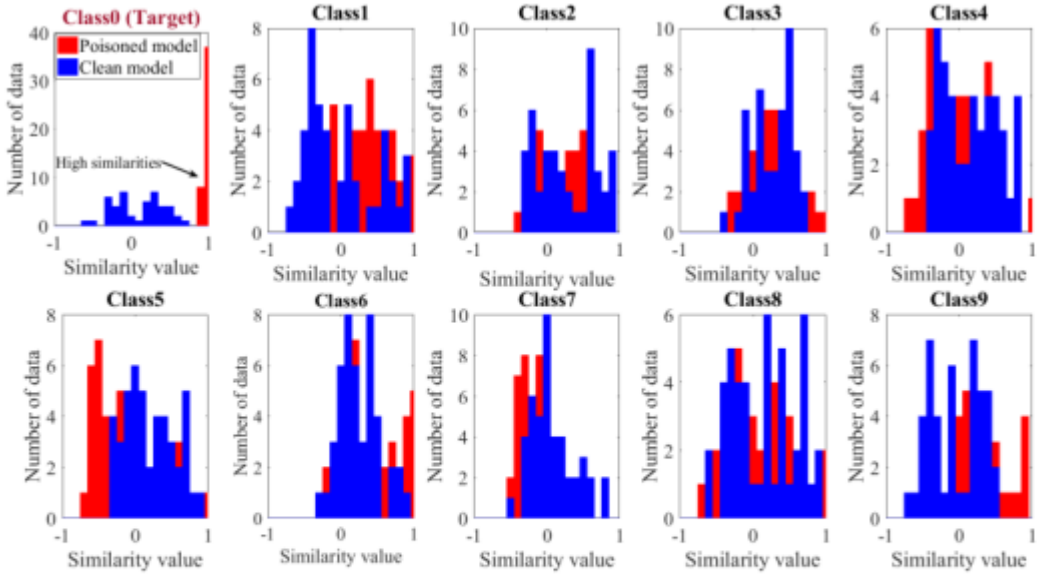


Fig. 4. Distribution of similarity scores for clean-Net versus TrojanNet under different classes.

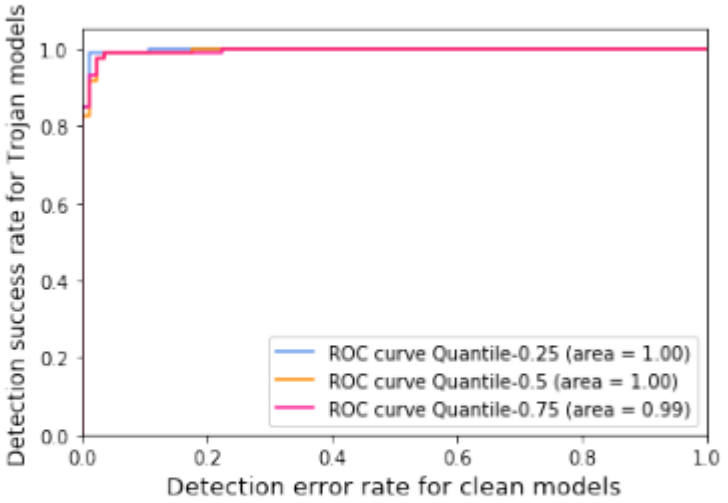


Fig. 5. ROC curve for TrojanNet detection using DL-TND.

算法原理 实验结果

- 不同结构模型的检测结果

Table 3. Comparisons between DL-TND and NC [36] on TrojanNets and cleanNets using $T_1 = 0.7$. The results are reported in the format (number of correctly detected models)/(total number of models)

		DL-TND (clean)	**DL-TND** (Trojan)	NC (clean)	NC (Trojan)
CIFAR-10	ResNet-50	20/20	20/20	11/20	13/20
	VGG16	10/10	9/10	5/10	6/10
	AlexNet	10/10	10/10	6/10	7/10
GTSRB	ResNet-50	12/12	12/12	10/12	6/12
	VGG16	9/9	9/9	6/9	7/9
	AlexNet	9/9	8/9	5/9	5/9
ImageNet	ResNet-50	5/5	5/5	4/5	1/5
	VGG16	5/5	4/5	3/5	2/5
	AlexNet	4/5	5/5	4/5	1/5
Total		**84**/85	**82**/85	54/85	48/85

	CIFAR-10	GTSRB	R-ImgNet
Test accuracy (Trojan)	90.51%	92.99%	86.7%
Attack success rate	99.65%	99.65%	98.6%
Test accuracy (clean)	92.64%	92.5%	87.8%

算法原理 优劣分析

- 优势：
  - 无需对攻击方法的假设，对不可知攻击的泛化性强
  - 只需要少量干净样本
  - 计算量小，不需训练多个模型

- 局限性：
  - 需要白盒模型以获取激活向量

应用总结

应用总结

- 机器学习模型后门攻击威胁领域
  - 图像和视频识别
  - 自然语言处理
  - 语音识别
  - 恶意软件分类
- 机器学习模型后门技术应用
  - 知识产权保护（水印，验证模型所有权）
  - 验证个人数据是否被按要求删除
  - 人为引入后门（蜜罐）以防御对抗攻击
  - DNN可解释性研究和方法评估

参考文献

[1] XU X, WANG Q, LI H, et al. Detecting AI Trojans Using Meta Neural Analysis[J]. arXiv:1910.03137 [cs], 2020.

[2] WANG R, ZHANG G, LIU S, et al. Practical Detection of Trojan Neural Networks: Data-Limited and Data-Free Cases[C]//VEDALDI A, BISCHOF H, BROX T, et al. Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020: 222–238.

[3] HUSTER T, EKWEDIKE E. TOP: Backdoor Detection in Neural Networks via Transferability of Perturbation[J]. arXiv:2103.10274 [cs], 2021.

[4] LI Y, WU B, JIANG Y, et al. Backdoor Learning: A Suvey[J]. arXiv:2007.08745 [cs], 2021.

道德经

大成若缺，其用不弊。
大盈若冲，其用不穷。
大直若屈。大巧若拙。
大辩若讷。静胜躁，寒
胜热。清静为天下正。

谢谢！