NeurIPS 2022

Untargeted Backdoor Watermark: Towards Harmlessand Stealthy Dataset Copyright Protection

Yiming Li[1,*], **Yang Bai**[2,*], **Yong Jiang**[1], **Yong Yang**[3], **Shu-Tao Xia**[1], **Bo Li**[4]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University, China
[2]Tencent Security Zhuque Lab, China
[3]Tencent Security Platform Department, China
[4]The Department of Computer Science, University of Illinois at Urbana-Champaign, USA
li-ym18@mails.tinghua.edu.cn; {mavisbai,coolcyang}@tencent.com;
{jiangy,xiast}@sz.tsinghua.edu.cn; lbo@illinois.edu

背 景

数据集所有权

现有很多数据集只能用于学术或教育目的，未经允许不可商用。恶意用户很有可能在未经授权的情况下用其训练第三方商用模型，进而破坏数据集所有者的版权，给数据集的所有者造成巨大的损失。

现有经典数据保护方法

- 加　　密
- 图像水印
- 差分隐私

均不能直接被用于保护公开数据集的版权

背 景

现有经典数据保护方法——加密、图像水印、差分隐私

加　　密:

- 图像在加密后得到的是一张混沌图像，我们根本无法从加密后的图像中获取任何有效信息。
- 解密后图像，与原始图像保持一致

原始图像

加密后图像

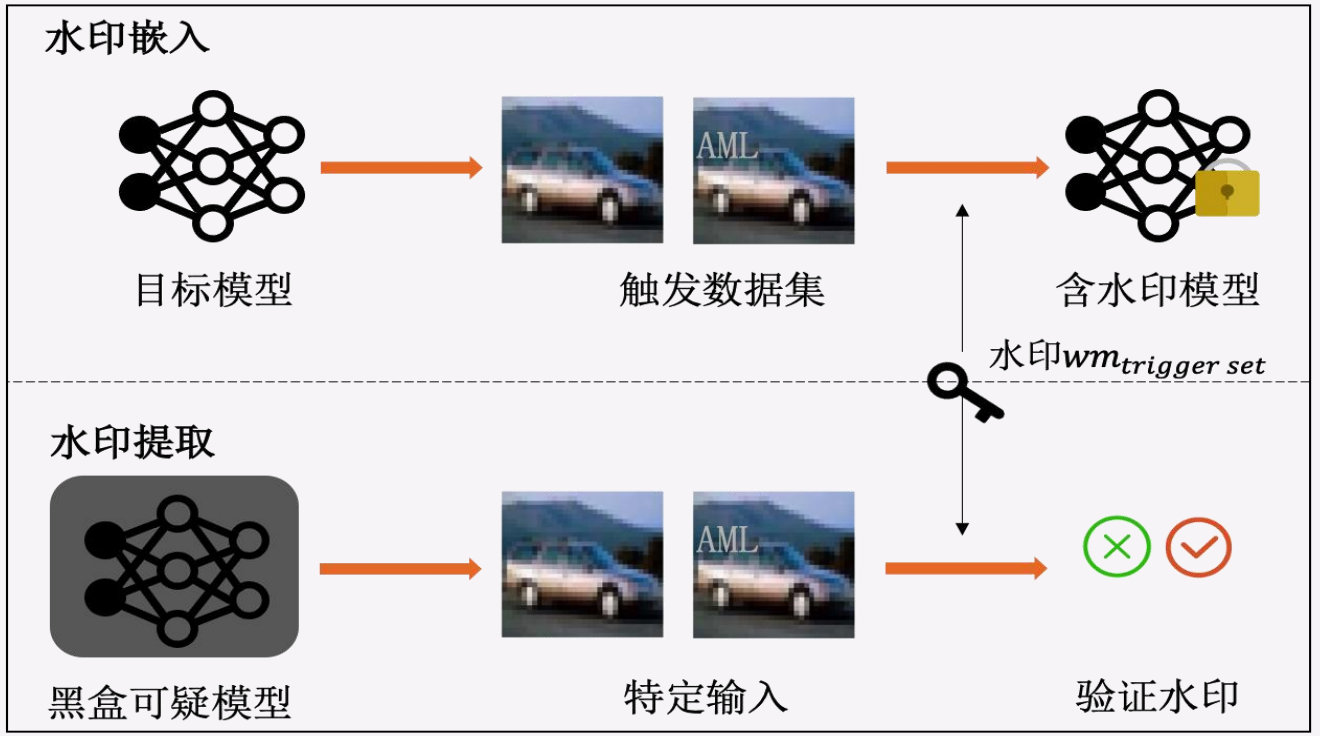
解密后图像

会破坏这些数据集的可用性

背 景

现有经典数据保护方法——加密、图像水印、差分隐私

图像水印：



恶意用户只会发布其模型而不会发布其训练细节

背 景

现有经典数据保护方法——加密、图像水印、差分隐私

差分隐私

一般将满足差分隐私的函数称为一个*机制*（Mechanism）。如果对于所有*临近数据集*（Neighboring Dataset）x和x'和所有可能的输出S，机制F均满足

$$\frac{\Pr[F(x)=S]}{\Pr[F(x')=S]} \leq e^{\epsilon} \tag{1}$$

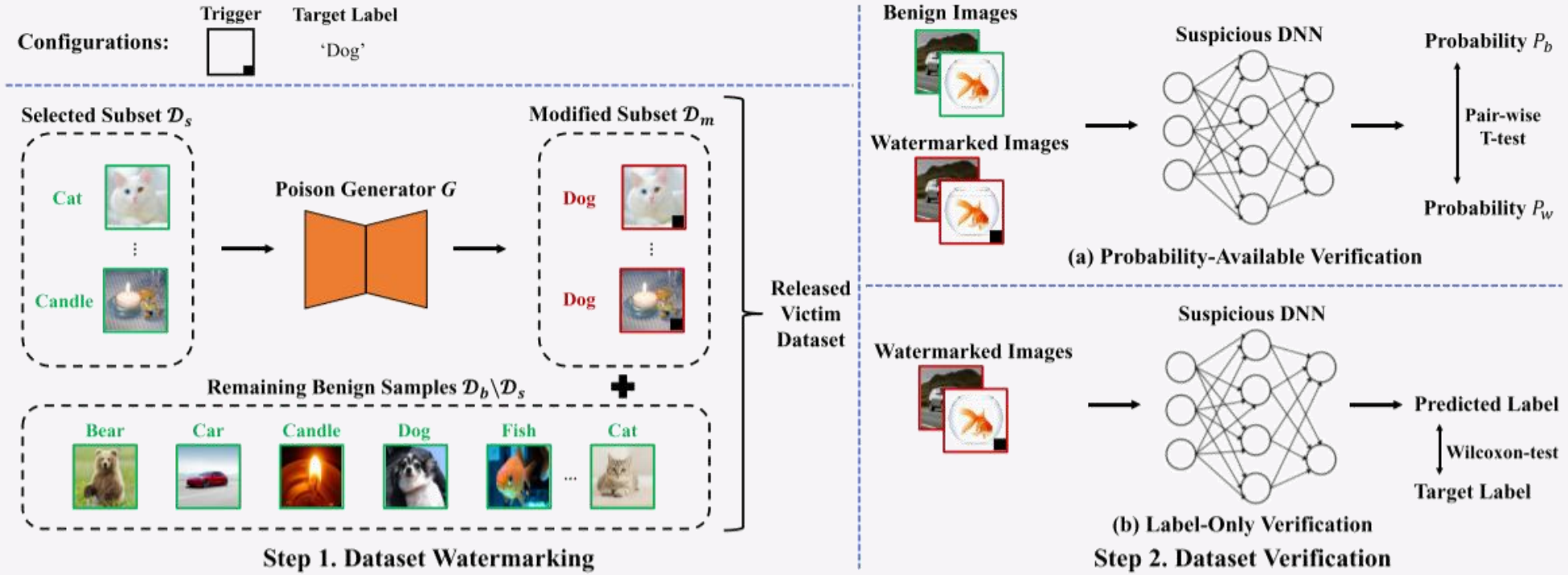
则称机制F满足差分隐私。

需要操纵模型的训练流程，会破坏这些数据集的可用性

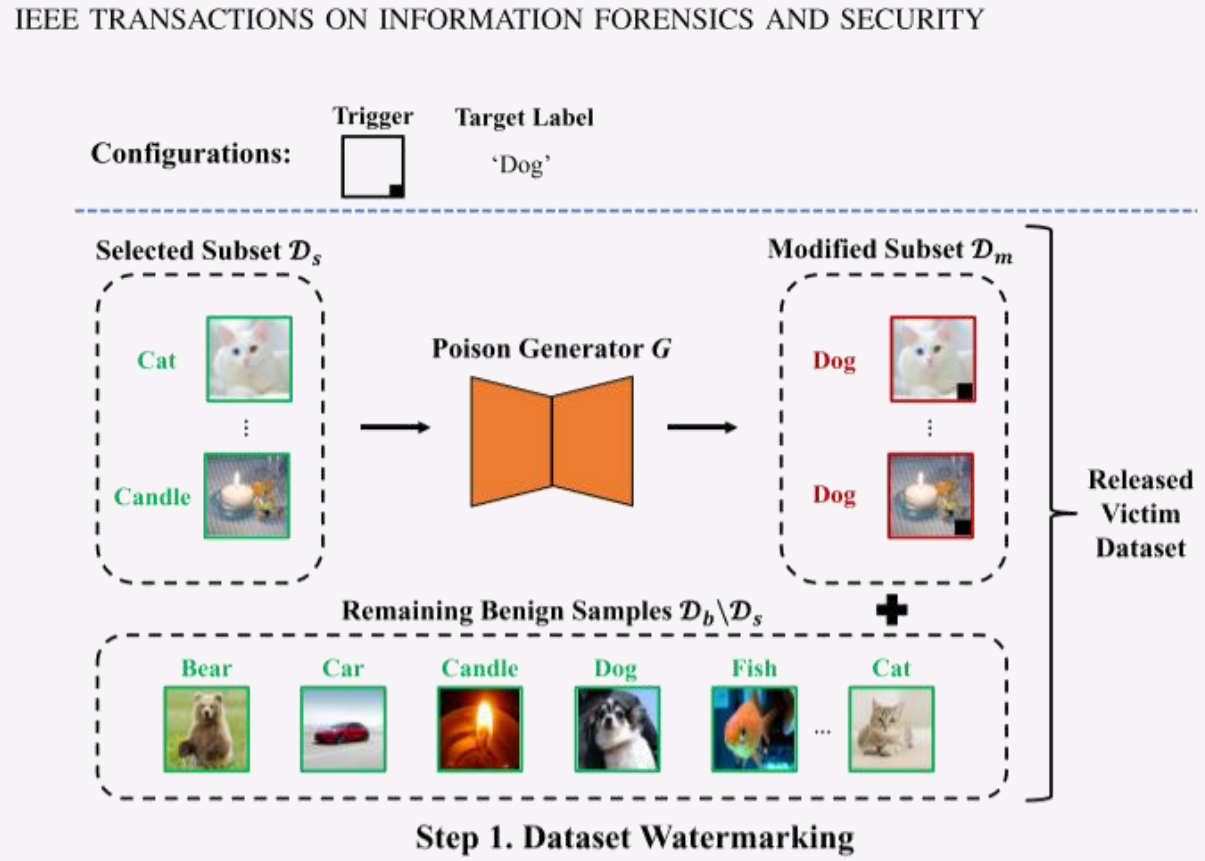
IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY

先前工作 Black-box Dataset Ownership Verification via Backdoor Watermarking

- 仅投毒式后门攻击：需要改变训练数据集
- 训练控制攻击：需要修改其他训练组件（如训练损失）
- 模型修改攻击：直接修改模型参数或结构



先前工作 Black-box Dataset Ownership Verification via Backdoor Watermarking



水印属性:

- 无害性：水印不应对数据集功能有害，

$$BA(f) - BA(\hat{f}) < \zeta$$

其中BA表示良性样本准确度

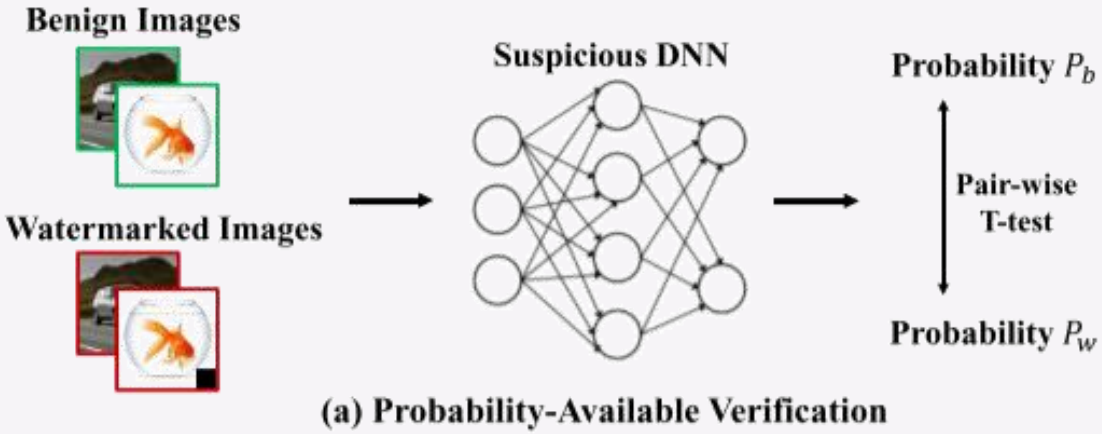
- 特殊性：$\frac{1}{|W|}\sum_{x' \in W} d\left(\hat{f}(x'), f(x')\right) > \eta$

W是带水印的数据集

d是距离

- 隐蔽性：数据集水印不应引起攻击者的注意，水印率应该很小，并且对用户来说是自然不明显的。

先前工作 Black-box Dataset Ownership Verification via Backdoor Watermarking



(a) Probability-Available Verification

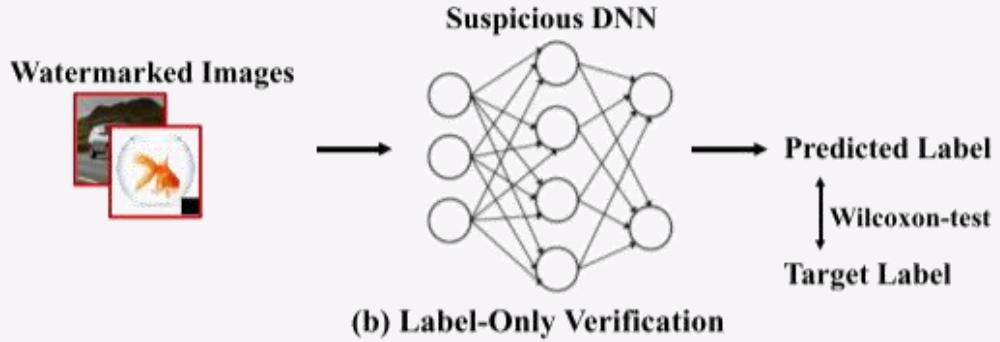
Algorithm 1 Probability-available dataset verification.

Input: benign dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, sampling number m, suspicious model f, poison generator G, target label y_t, alternative hypothesis H_1
Sample a data list $\boldsymbol{X} = [\boldsymbol{x}_i | y_i \neq y_t]_{i=1}^{m}$ from $\mathcal{D}$
Obtain the watermarked version of $\boldsymbol{X}$ (*i.e.*, $\boldsymbol{X}'$) based on $\boldsymbol{X}' = [G(\boldsymbol{x}_i)]_{i=1}^{m}$
Obtain the probability list $\boldsymbol{P}_b = [f(\boldsymbol{x}_i)_{y_t}]_{i=1}^{m}$
Obtain the probability list $\boldsymbol{P}_w = [f(G(\boldsymbol{x}_i))_{y_t}]_{i=1}^{m}$
Calculate p-value via PAIR-WISE-T-TEST($\boldsymbol{P}_b$, $\boldsymbol{P}_w$, H_1)
Calculate ΔP via AVERAGE($\boldsymbol{P}_w - \boldsymbol{P}_b$)
Output: ΔP and p-value

可用概率验证：防御者可以获得预测概率向量。只需要验证水印样本的目标类的后验概率是否显著高于良性测试样本的后验概率。

$$H_0: P_b + \tau = P_\omega (H_1: P_b + \tau < P_\omega)$$

先前工作 Black-box Dataset Ownership Verification via Backdoor Watermarking



(b) Label-Only Verification

Algorithm 2 Label-only dataset verification.

Input: benign dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, sampling number m, suspicious model C, poison generator G, target label y_t, alternative hypothesis H_1
Sample a subset $\boldsymbol{X} = \{\boldsymbol{x}_i | y_i \neq y_t\}_{i=1}^{m}$ from $\mathcal{D}$
Obtain the watermarked version of $\boldsymbol{X}$ (*i.e.*, $\boldsymbol{X}'$) based on $\boldsymbol{X}' = \{G(\boldsymbol{x}) | \boldsymbol{x} \in \boldsymbol{X}\}$
Obtain the predicted label of $\boldsymbol{X}'$ via $\boldsymbol{L} = \{C(\boldsymbol{x}) | \boldsymbol{x} \in \boldsymbol{X}'\}$
Calculate p-value via WILCOXON-TEST($\boldsymbol{L}$, y_t, H_1)
Output: p-value

仅标签验证：防御者只能获得预测的标签。只需要检查加水印样本的预测标签是否是目标标签。

$$H_0 : C(\boldsymbol{X}') \neq y_t \ (H_1 : C(\boldsymbol{X}') = y_t)$$

先前工作 Black-box Dataset Ownership Verification via Backdoor Watermarking

表1 CIFAR-10 和 ImageNet 上数据集水印的良性准确率（%）和水印成功率（%）

Dataset↓	Method→	Standard	BadNets				Blended			
	Trigger→	No Trigger	Line		Cross		Line		Cross	
	Model↓, Metric→	BA	BA	WSR	BA	WSR	BA	WSR	BA	WSR
CIFAR-10	ResNet	92.13	91.93	99.66	91.92	100	91.34	94.93	91.55	99.99
	VGG	91.74	91.37	99.58	91.48	100	90.75	94.43	91.61	99.95
ImageNet	ResNet	85.68	84.43	95.87	84.71	99.65	84.32	82.77	84.36	90.78
	VGG	89.15	89.03	97.58	88.88	99.99	88.92	89.37	88.57	96.83

表 3. 在 CIFAR-10 和 ImageNet 上进行仅标签数据集验证的有效性（p 值）

Model↓	Dataset→	CIFAR-10				ImageNet			
	Method→	BadNets		Blended		BadNets		Blended	
	Scenario↓, Trigger→	Line	Cross	Line	Cross	Line	Cross	Line	Cross
ResNet	Independent Trigger	1	1	1	1	1	1	1	1
	Independent Model	1	1	1	1	1	1	1	1
	Steal	0	0	10^{-3}	0	0.014	0	0.016	10^{-3}
VGG	Independent Trigger	1	1	1	1	1	1	1	1
	Independent Model	1	1	1	1	1	1	1	1
	Steal	0	0	10^{-3}	0	10^{-3}	0	0.018	10^{-3}

先前工作 Black-box Dataset Ownership Verification via Backdoor Watermarking

表 2. 在 CIFAR-10 和 ImageNet 上验证概率可用数据集的有效性（ΔP 和 p 值）

Dataset↓	Model↓	Method→	BadNets				Blended			
		Trigger→	Line		Cross		Line		Cross	
		Scenario↓, Metric→	ΔP	p-value	ΔP	p-value	ΔP	p-value	ΔP	p-value
CIFAR-10	ResNet	Independent Trigger	10^{-4}	1	-10^{-4}	1	10^{-3}	1	-10^{-3}	1
		Independent Model	10^{-3}	1	10^{-5}	1	10^{-3}	1	-10^{-4}	1
		Steal	0.98	10^{-87}	0.99	10^{-132}	0.93	10^{-58}	0.99	10^{-103}
	VGG	Independent Trigger	10^{-5}	1	-10^{-3}	1	10^{-3}	1	10^{-4}	1
		Independent Model	10^{-3}	1	-10^{-3}	1	-10^{-3}	1	-10^{-5}	1
		Steal	0.99	10^{-133}	0.98	10^{-77}	0.94	10^{-56}	0.99	10^{-163}
ImageNet	ResNet	Independent Trigger	-10^{-4}	1	10^{-4}	1	-10^{-3}	1	-10^{-4}	1
		Independent Model	10^{-4}	1	10^{-4}	1	-10^{-5}	1	-10^{-4}	1
		Steal	0.92	10^{-54}	0.98	10^{-114}	0.72	10^{-23}	0.85	10^{-41}
	VGG	Independent Trigger	-10^{-3}	1	-10^{-4}	1	-10^{-5}	1	-10^{-6}	1
		Independent Model	-10^{-6}	1	-10^{-6}	1	10^{-8}	1	10^{-6}	1
		Steal	0.97	10^{-68}	0.99	10^{-181}	0.86	10^{-37}	0.95	10^{-67}

先前工作

新的安全威胁

攻击者可以通过模型中的后门**确定性**的恶意操纵模型的输出。

这种引入的新安全威胁会造成数据集使用者对提供者的不信任和潜在的安全风险，进而阻碍该方法的实际使用。

主要来源于其**有目标特性**

无目标后门水印 UBW

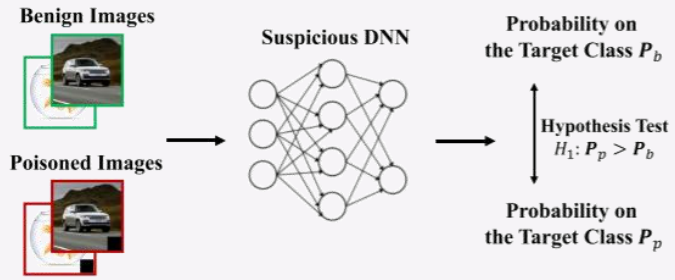


Figure 1: The verification process of BEDW.

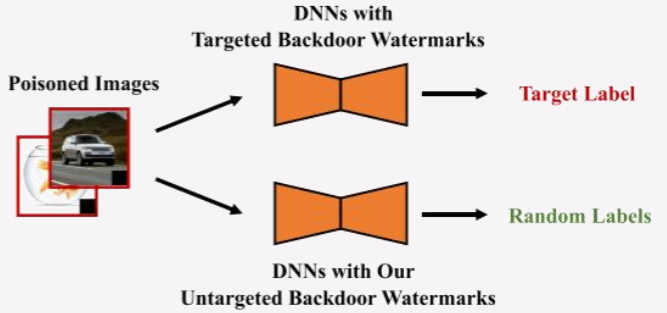


Figure 2: The inference process of DNNs with different types of backdoor watermarks.

UBW方法

目标

- **有效性（effectiveness）**：带水印的DNN将错误分类中毒图像
- **隐蔽性（stealthiness）**：要求数据集用户无法识别水印
- **分散性（dispersibility）**：确保中毒图像的分散预测，即同一类的被水印样本可能被模型（均匀）预测成各个类别

UBW方法

定义（平均预测分散度）

$D = \{(x_i, y_i)\}_{I=1}^{N}$表示数据集，其中$y_i \in Y = \{1,2,\ldots,K\}$和C： X → Y是分类器。令$P^{(j)}$是具有ground-truth标签j的样本上的模型预测的概率向量，其中$P^{(j)}$的第i个元素是

$$P_i^{(j)} \triangleq \frac{\sum_{k=1}^{N} \mathbb{I}\{C(\boldsymbol{x}_k) = i\} \cdot \mathbb{I}\{y_k = j\}}{\sum_{k=1}^{N} \mathbb{I}\{y_k = j\}}$$

$$D_p \triangleq \frac{1}{N} \sum_{j=1}^{K} \sum_{i=1}^{N} \mathbb{I}\{y_i = j\} \cdot H\left(\boldsymbol{P}^{(j)}\right)$$

Dp表征具有相同标签的不同图像的预测的分散程度，Dp越大，对手就越难确定性地操纵预测。

UBW-P方法

Untargeted Backdoor Watermark with Poisoned Labels（UBW-P）

关　键：在生成中毒数据集时随机“洗牌”中毒训练样本的标签。

$$\mathcal{D}_m \doteq \{(\boldsymbol{x}', y') | \boldsymbol{x}' = G(\boldsymbol{x}; \boldsymbol{\theta}), y' \sim [1, \cdots, K], (\boldsymbol{x}, y) \in \mathcal{D}_s\}$$

$$\min_{\boldsymbol{w}} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_m \cup (\mathcal{D}\setminus\mathcal{D}_s)} \mathcal{L}(f(\boldsymbol{x}; \boldsymbol{w}), y)$$

对于任何测试样本，攻击者可以基于生成器G激活包含在具有中毒图像G(x)的被攻击DNN中的隐藏后门。

因为这些被水印训练样本的标签是错误的，UBW-P不够隐蔽

UBW-C方法

Untargeted Backdoor Watermark with Clean Labels (UBW-C)

主要特点：不修改投毒训练样本的标签

核心难题：平均预测分散度不可导，无法被直接优化

定义（平均样本分散度和平均类别分散度）

$$D_s \triangleq \frac{1}{N}\sum_{i=1}^{N} H\left(f(\boldsymbol{x}_i)\right)$$

$$D_c \triangleq \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N} \mathbb{I}\{y_i = j\} \cdot H\left(\frac{\sum_{i=1}^{N} f(\boldsymbol{x}_i)\cdot \mathbb{I}\{y_i = j\}}{\sum_{i=1}^{N} \mathbb{I}\{y_i = j\}}\right)$$

平均样本分散度描述了所有样本的预测概率向量的平均离散度

平均类别分散度描述了每个类中样本的平均预测的平均离散度。

UBW-C方法

UBW-C优化

引理 1 类平均离散度总是大于或等于样本平均离散度，即 Ds ≤ Dc。当且仅当 f(xi)= f(xj)时，相等关系成立。

由于熵是一个凹函数，根据jenson不等式：在凹函数的定义域上取若干个点，这些点的函数值的加权平均，不大于这些点的加权平均值的函数值

$$H\left(\frac{\sum_{i=1}^{N} f(\boldsymbol{x}_i)\cdot\mathbb{I}\{y_i=j\}}{\sum_{i=1}^{N}\mathbb{I}\{y_i=j\}}\right)\geq\sum_{i=1}^{N}\frac{\mathbb{I}\{y_i=j\}}{\sum_{i=1}^{N}\mathbb{I}\{y_i=j\}}H\left(f(\boldsymbol{x}_i)\right)=H\left(f(\boldsymbol{x}_i)\right)$$

由于每个样本x具有且仅具有一个标签y ∈ {1，· · ·，K}:

$$H\left(f(\boldsymbol{x}_i)\right)=\sum_{j=1}^{K}H\left(f(\boldsymbol{x}_i)\right)\cdot\mathbb{I}\{y_i=j\},\forall i\in\{1,\cdots,N\}$$

因此

$$D_c\geq\frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N}\mathbb{I}\{y_i=j\}\cdot H\left(f(\boldsymbol{x}_i)\right)=\frac{1}{N}\sum_{i=1}^{N}H\left(f(\boldsymbol{x}_i)\right)\triangleq D_s$$

UBW-C方法

UBW-C优化

引理 1 类平均离散度总是大于或等于样本平均离散度，即 Ds ≤ Dc。当且仅当 f(xi)= f(xj)时，相等关系成立。

$$D_c \geq \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N}\mathbb{I}\{y_i = j\}\cdot H\left(f(\boldsymbol{x}_i)\right) = \frac{1}{N}\sum_{i=1}^{N}H\left(f(\boldsymbol{x}_i)\right) \triangleq D_s$$

定理 1 假设 f (· ;w) 表示参数为 w 的 DNN，G (· ; θ) 表示参数为 θ 的数据污染图像生成器，D 是具有 K 个类别的给定数据集，我们有

$$\max_{\boldsymbol{\theta}}\sum_{i=1}^{N}H\left(f(G(\boldsymbol{x}_i;\boldsymbol{\theta});\boldsymbol{w})\right) \leq \max_{\boldsymbol{\theta}}\sum_{j=1}^{K}\sum_{i=1}^{N}\mathbb{I}\{y_i = j\}\cdot H\left(\frac{\sum_{i=1}^{N}f(G(\boldsymbol{x}_i;\boldsymbol{\theta});\boldsymbol{w})\cdot\mathbb{I}\{y_i = j\}}{\sum_{i=1}^{N}\mathbb{I}\{y_i = j\}}\right)$$

我们可以通过最大化Ds来同时优化Ds和Dc。

UBW-C方法

UBW-C优化

主要思路：平均样本分散度是平均类别分散度的下界，因此最大化平均样本分散度在某种程度上可以同时最大化平均类别分散度

$$\max_{\boldsymbol{\theta}} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_s} \left[\mathcal{L}(f(G(\boldsymbol{x};\boldsymbol{\theta});\boldsymbol{w}^*),y) + \lambda \cdot H\left(f(G(\boldsymbol{x};\boldsymbol{\theta});\boldsymbol{w}^*)\right)\right]$$

$$s.t.\ \boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_p} \mathcal{L}(f(\boldsymbol{x};\boldsymbol{w}),y)$$

上述过程是标准的两级优化过程，可以通过交替优化下级子问题和上级子问题来有效地解决。

通过 mini-batch 的随机梯度下降（SGD）优化，估算Dc是很困难的，仅优化Ds简单而准确。

基于无目标后门水印的数据集所有权认证

验证

主要思路：通过判断可疑模型对被水印图片在其真实类别上的预测概率是否有明显下降（相比其对应未水印图片的预测概率），来判断该模型是否含有特定的无目标后门，进而判断该模型是否曾在保护的数据集上训练。

主要方法：基于单边配对样本t检验

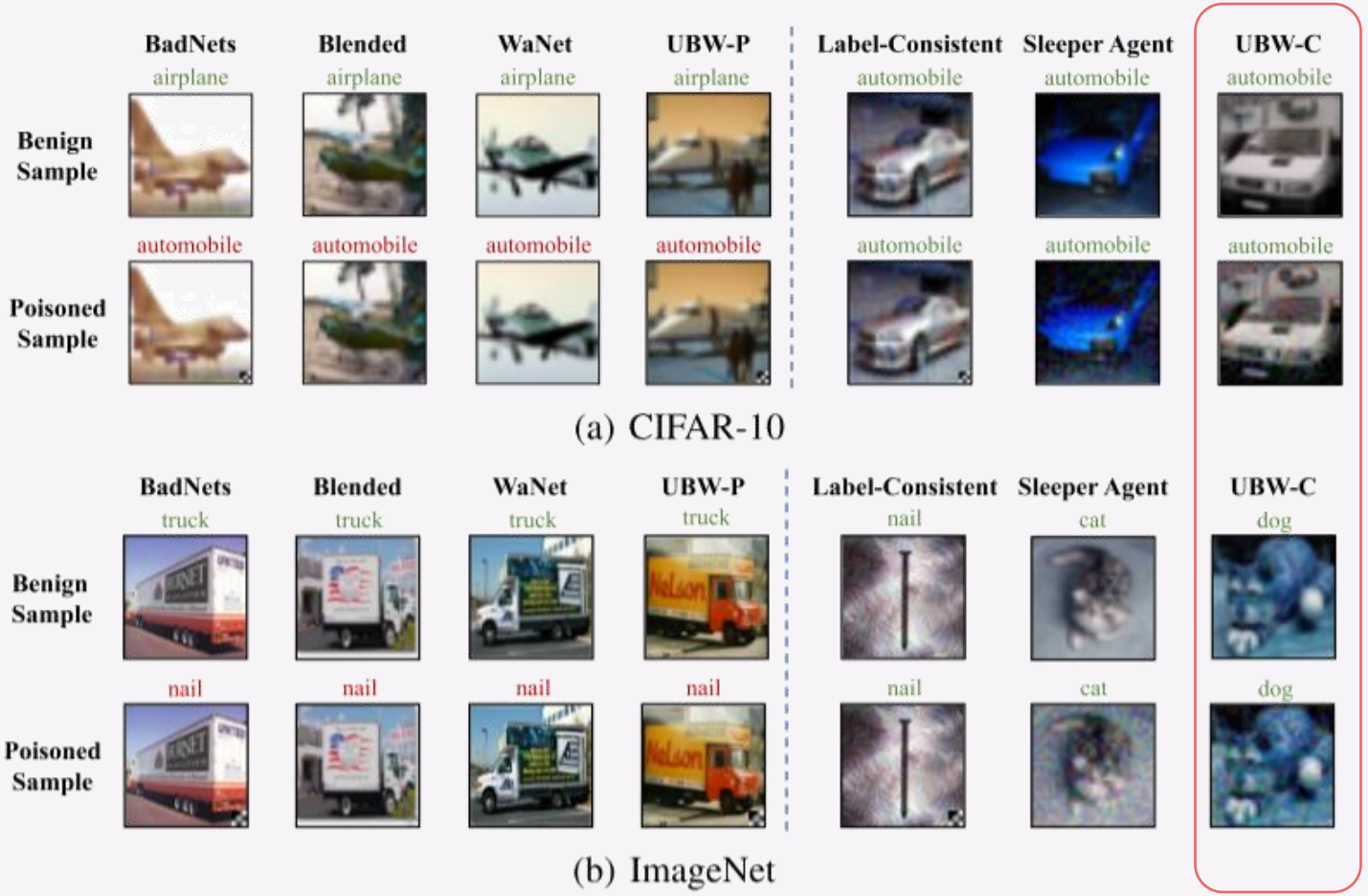
$f(x)$是可疑模型预测的x的后验概率，变量X表示良性样本，X'是中毒样本（$X' = G(X)$），$P_b = f(X)_Y$和$P_p = f(X')_Y$分别表示X和X'在ground-truth标签Y上的预测概率。给定零假设$\boldsymbol{H_0: P_b = P_p + \tau}$($\boldsymbol{H_1: P_b > P_p + \tau}$)，其中超参数$\tau \in$ [0，1]，当且仅当H_0被拒绝时，可疑模型在受保护的数据集上训练（具有 τ - 确定性）。

数据集水印的性能研究

UBW-C

不可见性

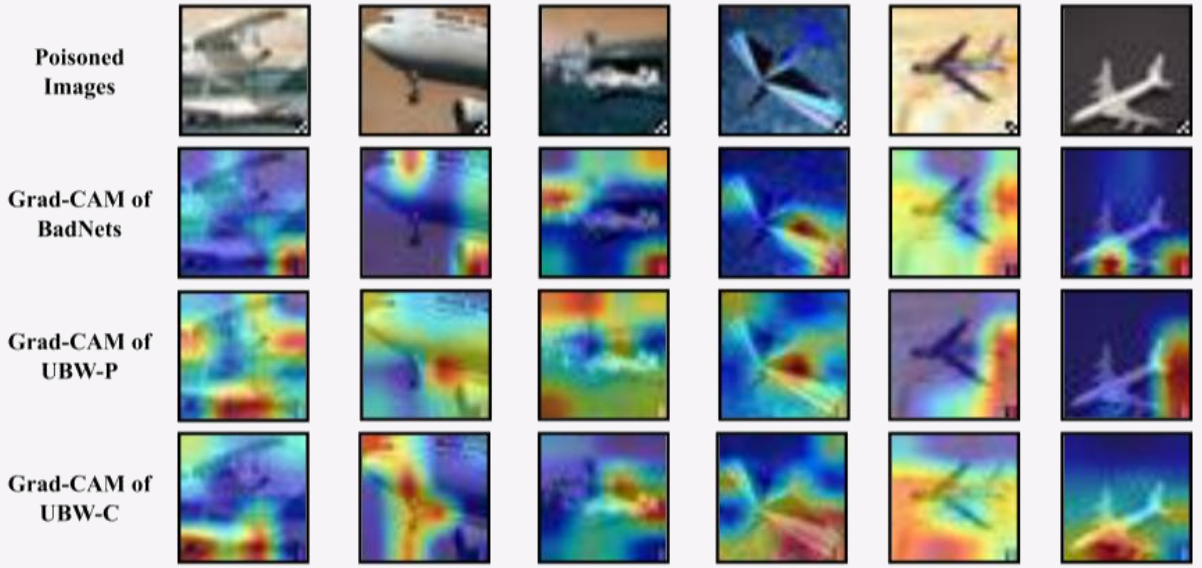
标签一致性

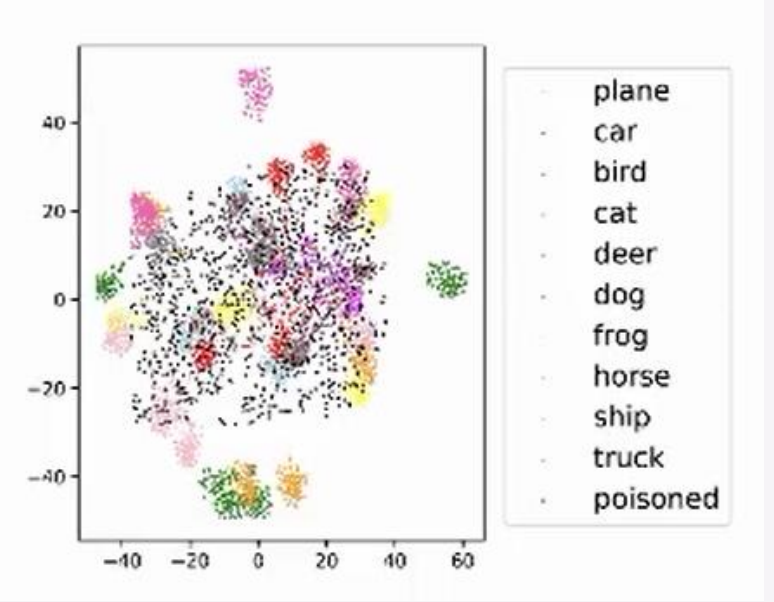


数据集水印的性能研究

UBW-C

CAM　　Feature Space





数据集水印的性能研究

Table 1: The watermark performance on the CIFAR-10 dataset.

Label Type↓	Target Type↓	Method↓, Metric→	BA (%)	ASR-A (%)	ASR-C (%)	D_p
N/A		No Attack	92.53	N/A	N/A	N/A
Poisoned-Label	Targeted	BadNets	91.52	100	100	0.0000
		Blended	91.61	100	100	0.0000
		WaNet	90.48	95.50	95.33	0.1979
	Untargeted	UBW-P (Ours)	90.59	92.30	92.51	2.2548
Clean-Label	Targeted	Label-Consistent	82.94	96.00	95.80	0.9280
		Sleeper Agent	86.06	70.60	54.46	1.0082
	Untargeted	UBW-C (Ours)	86.99	89.80	87.56	1.2641

Table 2: The watermark performance on the ImageNet dataset.

Label Type↓	Target Type↓	Method↓, Metric→	BA (%)	ASR-A (%)	ASR-C (%)	D_p
N/A		No Attack	67.30	N/A	N/A	N/A
Poisoned-Label	Targeted	BadNets	65.64	100	100	0.0000
		Blended	65.28	88.00	85.37	0.3669
		WaNet	62.56	78.00	73.17	0.7124
	Untargeted	UBW-P (Ours)	62.60	82.00	82.61	2.7156
Clean-Label	Targeted	Label-Consistent	62.36	30.00	2.78	1.2187
		Sleeper Agent	56.92	6.00	2.31	1.0943
	Untargeted	UBW-C (Ours)	59.64	74.00	60.00	2.4010

验证了本文所提无目标水印的有效性和分散性

数据集所有权验证性能研究

Table 3: The effectiveness of dataset ownership verification via UBW-P.

	CIFAR-10			ImageNet		
	Independent-T	Independent-M	Malicious	Independent-T	Independent-M	Malicious
ΔP	-0.0269	0.0024	0.7568	0.1281	0.0241	0.8000
p-value	1.0000	1.0000	10^{-36}	0.9666	1.0000	10^{-10}

Table 4: The effectiveness of dataset ownership verification via UBW-C.

	CIFAR-10			ImageNet		
	Independent-T	Independent-M	Malicious	Independent-T	Independent-M	Malicious
ΔP	0.1874	0.0171	0.6115	0.0588	0.1361	0.4836
p-value	0.9688	1.0000	10^{-14}	0.9999	0.9556	0.0032

验证了所提所有权认证方案的有效性和准确性。

总　结

- 揭示了在保护开源数据集版权方面现有方法的局限性
- 探究了中毒标签和干净标签设置下的无目标后门水印（UBW）方法
- 进一步讨论了如何使用UBW进行无害和隐蔽的数据集所有权验证
- 实验验证了UBW方法的有效性

谢谢观看

THANK YOU

NeurIPS 2022

Untargeted Backdoor Watermark: Towards Harmlessand Stealthy Dataset Copyright Protection