arXiv:2301.02615v2 [cs.CR] 1 Oct 2023

Silent Killer：一种隐蔽、干净标签、黑盒后门攻击

Tzvi Lederer
本-古里安大学
tzvil@post.bgu.ac.il

Gallil Maimon耶路撒冷希伯来大学 gallil.maimon@mail.huji.ac.il

Lior Rokach本-古里安大学 liorrk@post.bgu.ac.il

摘要

后门投毒攻击对神经网络构成了众所周知的风险。然而，大多数研究都聚焦于宽松的威胁模型。我们提出了 Silent Killer，这是一种新型攻击，可在干净标签、黑盒设置下实施，使用隐蔽的投毒样本和触发器，且表现优于现有方法。受此类方法在投毒标签设置下取得成功的启发，我们研究了在干净标签攻击中使用通用对抗扰动作为触发器。我们分析了一种简单改编方法的效果，发现要确保较高的成功率，必须在生成投毒样本时进行梯度对齐。我们在 MNIST、CIFAR10 和精简版 ImageNet 上进行了全面实验，并取得了最先进的结果。

1 引言

后门投毒攻击对神经网络构成了众所周知的风险 [1–3]。仅通过修改或添加中毒样本到受害者的训练集，就能在深度学习模型中植入不良行为，这对攻击者尤其具有吸引力。许多用于训练机器学习模型的数据集都是从互联网上抓取的 [4–6], ，因此攻击者可以伺机上传中毒样本，而受害者会在不知数据集已遭投毒的情况下使用这些样本进行训练。

基于投毒的攻击在假设的威胁模型上各不相同，例如攻击者对模型训练了解多少。对此最主要的分类是：白盒——完全了解模型的架构和参数；灰盒——了解受害者的架构，但不了解权重；以及黑盒——不了解被攻击模型的架构。由于掌握的信息更多，这些攻击从最容易利用到最难利用不等。尽管如此，在现实场景中，我们最常遇到黑盒设置，因为我们不了解攻击者模型，也无法控制它。

后门攻击也可以分为投毒标签攻击和干净标签攻击。在投毒标签攻击中，攻击者可以修改中毒样本的标签；而在干净标签设置下，攻击者不能更改标签，通常也不能大幅修改训练样本（修改幅度受特定范数限制）。干净标签攻击对应这样一种场景：样本经过筛选后再由人工标注，这与标准监督方法中的做法一致。因此，在干净标签设置下，中毒样本必须不引人怀疑，且其标签应与未投毒时相同。尽管干净标签攻击更具挑战性，但上述原因使其在现实应用中更受青睐。

后门攻击的另一项理想特性，是触发器在推理时具有隐蔽性。早期方法使用彩色补丁 [7–10], ，而许多较新的方法使用细微的加性扰动作为触发器 [11–15]。 [16] 提出了 TUAP，并展示了通用对抗扰动（UAP） [17]在投毒标签设置下可以成为有效的触发器。他们将 UAP 添加到源标签的样本中

预印本。正在审稿。

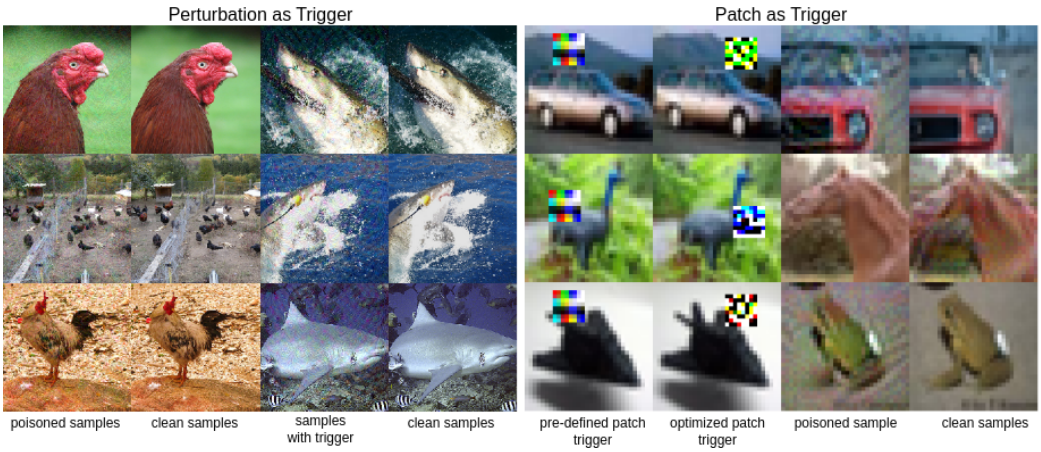

图1：左：Silent Killer攻击将ImageNet中的鲨鱼隐蔽地转化为公鸡的预测结果。右：我们在CIFAR10上的补丁攻击：针对源类别（汽车、鸟、飞机）的未优化和已优化触发器，以及针对目标类别（汽车、马、青蛙）的投毒样本和干净样本。

，将这些样本的标签更改为目标标签，并表明在此投毒数据集上训练的模型攻击成功率很高。尽管这种攻击成功地在黑盒设置下创建了隐蔽的触发器，但它仅适用于投毒标签设置。

在本研究中，我们展示了在干净标签设置下应用TUAP的成功率有限。此外，成功率变化很大，因此难以预测攻击是否会成功。我们提出，干净标签TUAP（*Clean-Label-TUAP*）（CL-TUAP）与投毒标签TUAP不同：在投毒标签TUAP中，分类任务能够利用并实现成功的唯一特征是触发器（即“对抗特征”）；而CL-TUAP的成功取决于学习原始非投毒特征（“干净特征”）的难易程度。如果模型能够轻易找到干净特征并因此快速实现较低的损失，或者这些干净特征更难被检测到，那么模型将“倾向于”使用对抗特征对样本进行分类。在后一种情况下，模型会将UAP与目标类别联系起来，攻击便会成功。不幸的是，我们无法保证模型确实会聚焦于UAP来执行分类，而不是依赖干净特征。

基于这一假设，我们提出对投毒样本进行显式优化，使模型在进行预测时关注通用对抗扰动。为此，我们采用了近期提出的一种用于数据投毒的方法——梯度对齐 [9]。我们表明，在 CL-TUAP 的基础上使用梯度对齐，可将 MNIST 上后门攻击的攻击成功率（ASR）从 78.72% 提升至 97.03%。此外，我们还表明，该方法能提升 CIFAR10 上多种架构在灰盒和黑盒设置下的攻击成功率，并在具有挑战性的干净标签、黑盒攻击设置下取得最先进结果。我们在缩减版 ImageNet 上展示了该攻击的可扩展性。图 1 展示了该攻击的样本。我们公开了代码 [1]。

2 相关工作

基于数据投毒的后门攻击最早由 [18] 提出。他们表明，向图像中添加触发器并将其重新标记为目标类别，会导致在这些数据上进行优化的模型在样本中出现触发器时预测为目标类别。后续研究使攻击更加隐蔽且更有效 [19, 20]。然而，大多数方法在干净标签设置下并不适用。

2.1 干净标签攻击

干净标签攻击在不更改中毒样本标签的情况下进行数据投毒。为了让模型关注触发器，而不是干净特征，一种方法是先对目标类别样本的特征进行“模糊处理”，再添加触发器。 [7] 使用了生成对抗网络和对抗

[1] https://github.com/TzviLederer/silent-killer

扰动，以减小目标类别特征的影响。其他研究也采用了相同思路，并做了一些改动 [8, 12]。另一种相关方法是“特征碰撞” [10], : 对目标类别的样本进行优化，使其在特征空间中接近“带触发器”的样本，然后用这些样本对数据集进行投毒。这些方法迁移到其他架构时效果有限，在黑盒场景中也难以奏效。此外，研究表明，当受害者模型从头开始训练（而非仅进行微调）时，特征碰撞并不起作用 [10]。 [15] 也会在特征空间中操纵投毒样本，并使用 UNet 生成具有增强触发器特征的扰动。然而，与图像以 50% 的比例混合会使扰动过于显眼，这是一个重大缺陷。

另一种方法，例如 [9], 所采用的 Sleeper Agent（SA）攻击中的梯度对齐，专注于优化投毒样本，以同时最小化受害者损失和攻击者损失（在测试集和触发样本上都实现较低的错误率）。尽管这种方法在黑盒设置下有效，但它需要多次重新训练代理模型，并且需要使用代理模型集成，从而导致更高的制作成本。此外，使用可见且色彩鲜艳的触发器也缺乏隐蔽性。我们的工作旨在通过使用经过优化的隐蔽触发器来改进这些方面。

2.2 UAP 作为触发器

通用对抗扰动已被研究为各种领域中有效的推理时攻击 [17, 21]。 [16]提出了 TUAP，证明通用对抗扰动可以作为后门攻击中的有效触发器。他们在代理模型上实施通用对抗扰动攻击，制作出一种能够将预测结果从源类别改变为目标类别的扰动。通过将该扰动添加到源类别样本中并修改其标签，他们将后门嵌入受害者模型。在推理时，将通用对抗扰动添加到新的源类别样本中，会使其被分类为目标类别，从而在不影响干净样本准确率的情况下实现较高的攻击成功率。

TUAP 将通用对抗扰动用于投毒标签攻击，而 [11] 则在干净标签设置下应用了类似的概念。他们表明，通过对目标类别的样本进行扰动优化，进一步降低这些样本上的损失，所构造的扰动便可在干净标签设置下用作有效的触发器。然而，在推理时，他们必须大幅增强触发器（将 $48/255$ 的范数从 $16/255$ 提高到 l_∞ ）。尽管投毒样本具有隐蔽性，触发器却相对显眼。 [8] 实现了 [7] 攻击的一个版本，将通用对抗扰动制作为小型彩色补丁，并将其应用于视频和图像分类模型。然而，该方法在训练和推理时都缺乏隐蔽性，在黑盒设置下的表现也不佳 [9]。总之，除我们的方法外，没有其他方法能在干净标签和黑盒设置下实施后门攻击，同时兼顾投毒样本和触发器的隐蔽性。

3 方法

3.1 攻击者目标与威胁模型

攻击者的目标是投毒训练集，使得受害者使用该数据训练模型 F 后，每当在源类别的样本上添加触发器，模型就会将其预测为目标类别。模型在干净样本上的表现必须接近使用干净数据训练时的表现。攻击者可投毒的样本数量有限（$\sim$ 1%），可添加的扰动也受到限制（$l_\infty \leq 16/255$）。这些约束与其他后门攻击类似 [10, 9]。投毒必须采用干净标签方式。我们在灰盒和黑盒场景下进行实验。我们假设攻击者能够访问受害者的训练集，这是先前研究中的标准假设 [9, 10]。最后，攻击者可以将隐蔽的触发器嵌入样本，并在推理时将其输入模型。如果模型将该样本预测为攻击者指定的目标类别，则攻击成功；攻击成功率（ASR）是指测试集中源类别样本被模型分类为目标标签的比例。

3.2 我们的方法

我们的攻击分为两个阶段：触发器制作和投毒样本制作。首先，我们使用基本迭代法 [17], 进行定向通用对抗扰动攻击 [22] ，并将其作为优化方法；通用对抗扰动优化的源类别和目标类别与后门攻击的源类别和目标类别相同。该扰动在干净的代理模型 F_s 上制作，并用作触发器。

算法 1投毒样本制作

输入： $F_s, \mathcal{D}_s, \mathcal{D}_t, \delta_t, t$

参数： R , α, ϵ

1：初始化 $\delta_p \leftarrow \{\delta_p^{(n)} = \delta_t\}_{n=1}^{N}$

2： $\mathcal{L}_a = \frac{1}{M} \sum_{x_m \in \mathcal{D}_s} \mathcal{L}(F_s(x_m + \delta_t),\ t)$

3: **对于** $r = 1, 2, \ldots\ldots R$ 优化步骤 **执行**

4： $\mathcal{L}_v = \frac{1}{N} \sum_{x_n \in \mathcal{D}_p} \mathcal{L}(F_s(x_n + \delta_p^{(n)}),\ y_n)$

5： $\mathcal{A} = 1 - \frac{\nabla_{\theta_s} \mathcal{L}_v \cdot \nabla_{\theta_s} \mathcal{L}_a}{||\nabla_{\theta_s} \mathcal{L}_v|| \cdot ||\nabla_{\theta_s} \mathcal{L}_a||}$

6： $\delta_p \leftarrow CLIP(\delta_p - \alpha \cdot \frac{\partial \mathcal{A}}{\partial \delta_p} - \epsilon, \epsilon)$

7: **返回** δ^p

表1：干净标签攻击在采用和未采用梯度对齐时的结果，分别对应 SK 和 CL-TUAP。可以看出，梯度对齐能显著提高攻击成功率。

	ResNet-18		VGG11		MobileNetV2		卷积神经网络（MNIST）	
	ASR	干净准确率	ASR	干净准确率	ASR	干净准确率	ASR	干净准确率
CL-TAUP	82.43	83.78	96.3	86.18	94.11	82.32	78.72	98.08
SK	**90.65**	83.69	**98.51**	85.87	**98.52**	82.22	**97.03**	98.08

第二阶段是通过梯度对齐来制作投毒样本，如算法1所述。我们的目标是使受害者的损失梯度 $\nabla_\theta \mathcal{L}(\mathcal{D}_p)$ 与攻击者的损失梯度 $\nabla_\theta \mathcal{L}(x_i + \delta_t, t)$ 对齐，其中 $\mathcal{D}_p$ 是投毒数据集，而 x_i、δ_t、t 分别是来自源类别的样本、触发器和目标类别。直观而言，攻击者希望在投毒数据集 $\mathcal{D}_p$ 上训练模型时，能够隐式地优化攻击者的损失。如果受害者的损失梯度与攻击者的损失梯度对齐，那么模型训练也会优化攻击者的损失。我们遵循 [9]，通过优化以下目标进行梯度对齐： $\mathcal{A}(p) = 1 - \frac{\nabla_\theta \mathcal{L}(\mathcal{D}_p(p)) \cdot \nabla_\theta \mathcal{L}(\mathcal{D}_t)}{\|\nabla_\theta \mathcal{L}(\mathcal{D}_p(p))\| \cdot \|\nabla_\theta \mathcal{L}(\mathcal{D}_t)\|}$ 其中，$\mathcal{L}(\mathcal{D}_t) = \frac{1}{M} \sum_{m=1}^{M} CE(F(x_m + \delta_t),\ t)$ 是攻击者的损失，即将触发器添加到源类别样本后，这些样本的损失； $\mathcal{L}(\mathcal{D}_p) = \frac{1}{N} \sum_{n=1}^{N} CE(F(x_n + \delta_p^{(n)}),\ y_n)$ 是受害者的损失，也是受害者直接优化的目标。 CE 是交叉熵损失， N 是投毒样本的数量， M 是用于估计目标梯度的源类别样本数量。我们遵循 [9]，选择梯度范数最大的样本进行投毒。投毒优化采用随机梯度下降： $p = p - \nabla_p \mathcal{A}(p)$

4 结果

TUAP [16] 在投毒标签设置下表现出很好的效果，在 CIFAR10 上达到了 96%-100% 的攻击成功率。我们希望评估其在干净标签设置下的表现，因此实现了它的干净标签版本 *CL-TUAP*。我们首先使用一个包含两个卷积层（分别有 32 和 64 个滤波器）及一个输出层的简单卷积神经网络，在 MNIST 上进行了实验。我们先训练一个代理模型，然后利用该模型构造触发器。触发器扰动的 l_∞ 范数上限为 $64/255$。我们将该扰动添加到 600 个样本（占数据集的 1%）中，并使用投毒数据集训练一个新模型。我们还在 CIFAR10 上使用不同架构进行了类似实验：ResNet18、MobileNet-V2 和 VGG11。与其他研究一样，我们在此使用了 $16/255$ 的扰动，并对 500 个样本（占数据的 1%）进行了投毒 [9, 10]。我们评估了 MNIST 的全部 90 个源-目标对，并从 CIFAR10 中随机选择了 24 对。

梯度对齐 如表1所示，尤其是在 MNIST 实验中，干净标签 TUAP 的结果并不理想。某些源-目标对的攻击成功率显著下降，例如，源类别为 3、目标类别为 1 时，攻击成功率为 24.65%；源类别为 0、目标类别为 4 时，攻击成功率为 26.53%。在现实场景中，这一差距可能相当显著，因此我们寻求改进方法。我们直觉上推测，在失败案例中，干净特征很容易被识别，因此模型不会考虑触发器的对抗特征。为了解决这一问题，我们在 Silent Killer 中使用梯度对齐来

表 2：在使用 1% 中毒样本的灰盒 CIFAR-10 设置下，SilentKiller 与领先干净标签攻击的攻击成功率（ASR）对比。考虑了多种触发器类型，包括补丁和扰动（$l_\infty < 16/255$有界）。前三种方法的结果来自 [9]。SA（基础版）和 Narcissus* 代表我们对这些方法的实现：SA（基础版）不使用重新训练和集成使用，而 Narcissus* 采用我们的触发器约束（$l_\infty < 16/255$），以进行公平比较。

触发器类型	架构	ResNet18	MobileNet-V2	VGG11
补丁	隐藏触发器后门 [10]	3.50	3.76	5.02
	干净标签后门 [7]	2.78	3.50	4.70
	SA [9]	78.84	75.96	86.60
	SA（基础）	57.42	15.00	25.25
	Silent Killer（补丁）（我们的）	**95.92**	41.68	97.35
扰动	Narcissus* [11]	59.06	85.00	77.56
	Silent Killer（扰动）（我们的）	90.65	**98.51**	**98.52**

表 3：黑盒攻击的平均攻击成功率，每列代表一种目标架构。结果表示 24 个源-目标对的平均攻击成功率。我们将我们的方法与 SA [9]进行比较：SA 使用单个代理模型（ResNet18），与我们相同；或者使用集成模型（ResNet18、ResNet34、VGG11、MobileNetV2，不包括受害者）来构造攻击。值得注意的是，即使 SA 使用集成模型，我们的方法仍优于 SA。

	代理模型 \目标	ResNet18	VGG11	MobileNetV2
SA [9]	ResNet18	-	31.96	29.10
	集成	63.11	55.28	42.40
Silent Killer（我们的方法）	ResNet18	-	81.60	90.92
	VGG11	76.97	-	95.56
	MobileNetV2	32.44	77.53	-

鼓励模型优化攻击者损失。表1清楚地显示，所有模型和数据集的表现均有所提升，且干净准确率没有明显受损。MNIST 上的提升为 18.31 个百分点，CIFAR10 上平均提升 4.95 个百分点。

优化触发器 为探究通用对抗扰动对攻击的益处，我们将结果与 SA [9], 进行比较。SA 通过色彩鲜艳、未经优化且用作触发器的补丁来实现梯度对齐。为确保比较公平，我们对与补丁尺寸相同（8×8 像素）的补丁施加通用对抗扰动。我们使用代理模型的相同检查点，分别利用优化和未优化的补丁制作投毒样本。图 1 展示了部分样本。如表 2 所示，触发器的优化对攻击成功率有显著影响；与未优化的同一攻击（即 SA – (base)）相比，我们优化后的补丁平均高出 45.76 个百分点。SA (base) 是梯度对齐的原始版本，我们也使用了这一版本；SA 则在此基础上增加了代理模型重新训练，以及用于制作投毒样本的模型集成。即使没有使用这些显然提升了其性能的技巧，我们的表现仍优于 SA。

请注意，补丁版本在 MobileNet-V2 上的结果明显低于其他架构。我们发现，基于梯度对齐的攻击的成功率与通用对抗扰动的构造效果高度相关，而后者在 MobileNetV2 上表现不佳。因此，我们估计，更好的通用对抗扰动方法 [23, 24] 或许能生成更有效的触发器并提高攻击成功率。

为了评估更贴近现实的数据投毒场景（即图像尺寸更大时），我们在 ImageNet 上评估了我们的方法。受计算资源限制，我们仅使用 ImageNet 中排名前 10 的类别进行训练并评估结果。我们选择了一组源类别与目标类别（公鸡-鲨鱼），对目标类别中的 185 个样本制作了投毒样本，并训练了一个 ResNet18 模型。图 1 展示了部分样本。我们发现，在这项初步实验中，攻击成功率为 76%，这表明，在更大、分辨率更高的真实世界数据集上，同样存在这种威胁。

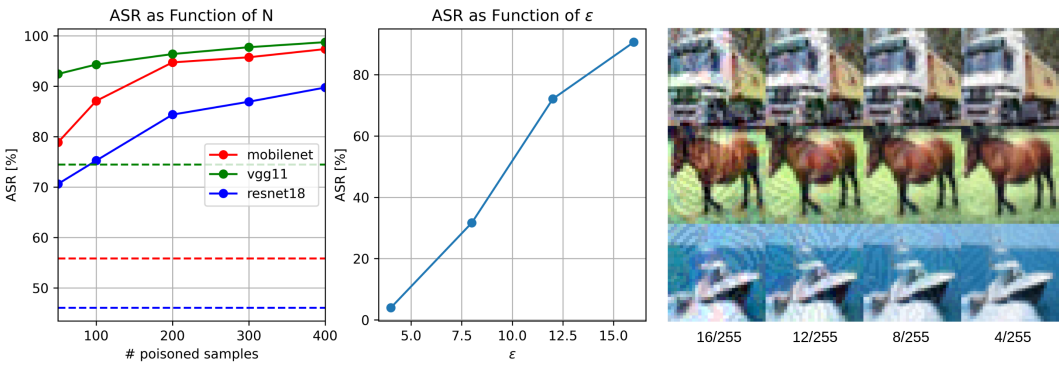

图2：左：中毒样本数量的敏感性分析 N。虚线表示未投毒、仅使用优化触发器时的结果。即使数据量很少（例如 100 个样本，占数据集的 0.2%），也能取得合理的结果。中：攻击成功率随投毒量和触发器扰动幅度变化的情况（l_∞）。在 CIFAR-10 上使用 ResNet18 进行训练。x 轴的单位范围为 (0,255)。范数越大，性能越好。右：带有不同 l_∞ 范数界限触发器的图像样本。

表4：应用防御后的攻击成功率和干净准确率。未应用防御时的干净准确率为 82.64%。可以看到，使用这些防御后，攻击的成功率并未显著下降。

方法	攻击成功率 [%]	准确率 [%]
激活聚类 [25]	71.28	73.13
DPSGD [26]	91.16	84.36
MixUp [27, 28]	94.24	81.39

黑盒 为了评估 Silent Killer 的黑盒效果，我们在代理模型和目标模型的所有组合中使用了 ResNet18、VGG11 和 MobileNetV2（表 3）。我们的黑盒攻击始终优于 SA。例如，当代理模型为 ResNet18 时，我们在 MobileNet 上将攻击成功率提升了 61.82 个百分点，在 VGG11 上提升了 49.64 个百分点。值得注意的是，我们的方法无需像 SA 那样采用计算成本高昂的重新训练过程，也能取得这些结果。这表明我们的攻击在黑盒设置下具有优势，并且采用类似技术还有望取得更好的性能。

敏感性分析 在现实场景中，悄然向受害者的训练数据中添加大量中毒样本颇具挑战。我们研究了攻击成功率如何随中毒样本数量（N）变化，并在三个模型上使用 $N = \{50, 100, 200, 300, 400\}$ 进行测试。图2展示了我们的方法十分高效，仅需50个样本（仅占训练集大小的0.1%）即可成功。这凸显了我们的攻击仅使用少量训练样本也能奏效。我们还研究了触发器和投毒样本的扰动范数上限，使用了 $\epsilon \in \{4/255, 8/255, 12/255, 16/255\}$。正如预期，我们观察到攻击隐蔽性（由 ϵ决定）与性能之间存在权衡。结果和扰动示例见图2。

防御 我们基于三个不同理念评估了三种不同类型的防御对我们方法的有效性：梯度塑形、数据过滤和数据增强。我们选择了被发现能够有效规避基于梯度对齐的数据投毒的方法 [9]。表4中的结果显示，这些方法均未显著降低我们攻击的攻击成功率。

5 结论

我们提出了一种新颖的数据投毒攻击，该攻击使用通用对抗扰动触发器优化和梯度对齐来进行投毒样本制作。在现实设置下，我们的攻击具有隐蔽性且效果显著，即使用干净标签、黑盒，以及隐蔽的投毒样本和触发器。尽管该攻击取得了成功，投毒标签方法仍然优于

最先进的干净标签方法。研究更优的干净标签方法能否弥合这一差距，是很有意义的。未来的研究可以探索使用更复杂的方法来制作作为触发器的通用对抗扰动。我们还可以联合优化投毒样本和触发器，以进一步提升性能。继续开发针对数据投毒的稳健防御至关重要。

参考文献

[1] Micah Goldblum、Dimitris Tsipras、Chulin Xie、Xinyun Chen、Avi Schwarzschild、Dawn Song、Aleksander Madry、Bo Li 和 Tom Goldstein。机器学习的数据集安全：数据投毒、后门攻击与防御。*IEEE*模式分析与机器智能汇刊，2022年。[2] Yiming Li、Yong Jiang、Zhifeng Li 和 Shu-Tao Xia。后门学习：综述。*IEEE*神经网络与学习系统汇刊，2022年。[3] Antonio Emanuele Cinà、Kathrin Grosse、Ambra Demontis、Sebastiano Vascon、Werner Zellinger、Bernhard A Moser、Alina Oprea、Battista Biggio、Marcello Pelillo 和 Fabio Roli。重探异常模式：机器学习防范训练数据投毒安全性的综述。*arXiv*预印本 *arXiv:2205.01992*，2022年。[4] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。语言模型是无监督的多任务学习者。*OpenAI*博客，1(8):9，2019年。[5] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：大型分层图像数据库。载于*2009*年*IEEE*计算机视觉与模式识别会议，第248–255页。IEEE，2009年。[6] Christoph Schuhmann、Romain Beaumont、Richard Vencu、Cade Gordon、Ross Wightman、Mehdi Cherti、Theo Coombes、Aarush Katta、Clayton Mullis、Mitchell Wortsman 等。Laion-5b：用于训练下一代图文模型的开放式大规模数据集。神经信息处理系统进展，35:25278–25294，2022年。[7] Alexander Turner、Dimitris Tsipras 和 Aleksander Madry。标签一致后门攻击。*arXiv*预印本 *arXiv:1912.02771*，2019年。[8] Shihao Zhao、Xingjun Ma、Xiang Zheng、James Bailey、Jingjing Chen 和 Yu-Gang Jiang。针对视频识别模型的干净标签后门攻击。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第14443–14452页，2020年。[9] Hossein Souri、Liam H Fowl、Rama Chellappa、Micah Goldblum 和 Tom Goldstein。Sleeper Agent：针对从头训练的神经网络的可扩展隐藏触发器后门。载于 Alice H. Oh、Alekh Agarwal、Danielle Belgrave 和 Kyunghyun Cho 编，神经信息处理系统进展，2022年。[10] Aniruddha Saha、Akshayvarun Subramanya 和 Hamed Pirsiavash。隐藏触发器后门攻击。载于*AAAI*人工智能会议论文集，第34卷，第11957–11965页，2020年。[11] Yi Zeng、Minzhou Pan、Hoang Anh Just、Lingjuan Lyu、Meikang Qiu 和 Ruoxi Jia。Narcissus：一种有限信息下实用的干净标签后门攻击。*arXiv*预印本 *arXiv:2204.05255*，2022年。[12] Nan Luo、Yuanzhang Li、Yajie Wang、Shangbo Wu、Yu-an Tan 和 Quanxin Zhang。通过两阶段特定触发器增强干净标签后门攻击。*arXiv*预印本 *arXiv:2206.04881*，2022年。[13] Mauro Barni、Kassem Kallas 和 Benedetta Tondi。通过训练集污染而不进行标签投毒，在卷积神经网络中实施一种新型后门攻击。载于*2019*年*IEEE*国际图像处理会议（*ICIP*），第101–105页。IEEE，2019年。[14] Yunfei Liu、Xingjun Ma、James Bailey 和 Feng Lu。反射后门：一种针对深度神经网络的自然后门攻击。载于计算机视觉——*ECCV 2020*：第*16*届欧洲会议，格拉斯哥，英国，*2020*年*8*月*23*日至*28*日，论文集，第*X*部分*16*，第182–199页。施普林格，2020年。

[15] Rui Ning、Jiang Li、Chunsheng Xin 和 Hongyi Wu。隐形投毒：一种针对深度神经网络的黑盒干净标签后门攻击。载于*IEEE INFOCOM 2021——IEEE*计算机通信会议，第1–10页。IEEE，2021年。[16] Quan Zhang、Yifeng Ding、Yongqiang Tian、Jianmin Guo、Min Yuan 和 Yu Jiang。Advdoor：深度学习系统的对抗性后门攻击。载于第*30*届*ACM SIGSOFT*国际软件测试与分析研讨会论文集，第127–138页，2021年。[17] Seyed-Mohsen Moosavi-Dezfooli、Alhussein Fawzi、Omar Fawzi 和 Pascal Frossard。通用对抗扰动。载于*IEEE*计算机视觉与模式识别会议论文集，第1765–1773页，2017年。[18] Tianyu Gu、Brendan Dolan-Gavitt 和 Siddharth Garg。Badnets：识别机器学习模型供应链中的漏洞。*arXiv*预印本 *arXiv:1708.06733*，2017年。[19] Cong Liao、Haoti Zhong、Anna Squicciarini、Sencun Zhu 和 David Miller。通过隐形扰动在卷积神经网络模型中植入后门。*arXiv*预印本 *arXiv:1808.10307*，2018年。[20] Anh Nguyen 和 Anh Tran。Wanet——基于不可感知形变的后门攻击。*arXiv*预印本 *arXiv:2102.10369*，2021年。[21] Gallil Maimon 和 Lior Rokach。针对文本分类器的通用对抗策略。神经网络，153:282–291，2022年。[22] Alexey Kurakin、Ian J Goodfellow 和 Samy Bengio。物理世界中的对抗样本。载于人工智能安全与保障，第99–112页。Chapman and Hall/CRC，2018年。[23] Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。迈向能够抵御对抗攻击的深度学习模型。*arXiv*预印本 *arXiv:1706.06083*，2017年。[24] Nicholas Carlini 和 David Wagner。迈向神经网络鲁棒性的评估。载于*2017*年*IEEE*安全与隐私研讨会（*SP*），第39–57页。IEEE，2017年。[25] Bryant Chen、Wilka Carvalho、Nathalie Baracaldo、Heiko Ludwig、Benjamin Edwards、Taesung Lee、Ian Molloy 和 Biplav Srivastava。通过激活聚类检测深度神经网络上的后门攻击。*arXiv*预印本 *arXiv:1811.03728*，2018年。[26] Sanghyun Hong、Varun Chandrasekaran、Yi˘gitcan Kaya、Tudor Dumitra¸s 和 Nicolas Papernot。通过梯度塑形缓解数据投毒攻击的有效性。*arXiv*预印本 *arXiv:2002.11497*，2020年。[27] Eitan Borgnia、Jonas Geiping、Valeriia Cherepanova、Liam Fowl、Arjun Gupta、Amin Ghiasi、Furong Huang、Micah Goldblum 和 Tom Goldstein。Dp-instahide：利用差分隐私数据增强，从理论上化解投毒与后门攻击。*arXiv*预印本 *arXiv:2103.02079*，2021年。[28] Hongyi Zhang、Moustapha Cisse、Yann N Dauphin 和 David Lopez-Paz。MixUp：超越经验风险最小化。*arXiv*预印本 *arXiv:1710.09412*，2017年。