计算机科学
COMPUTER SCIENCE

基于带毒分类器的自监督后门攻击防御方法

王一飞, 张胜杰, 薛迪展, 钱胜胜

引用本文

王一飞, 张胜杰, 薛迪展, 钱胜胜. 基于带毒分类器的自监督后门攻击防御方法[J]. 计算机科学, 2025, 52(4): 336-342.

WANG Yifei, ZHANG Shengjie, XUE Dizhan, QIAN Shengsheng. Self-supervised Backdoor Attack Defence Method Based on Poisoned Classifier [J]. Computer Science, 2025, 52(4): 336-342.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240100005

基于带毒分类器的自监督后门攻击防御方法

王一飞[1]　张胜杰[1]　薛迪展[2]　钱胜胜[2]
1 郑州大学河南先进技术研究院　郑州 450000
2 中国科学院自动化研究所多模态人工智能系统全国重点实验室　北京 100190
(wang_fei@gs.zzu.edu.cn)

摘　要　近年来,自监督学习网络(Self-Supervised Learning,SSL)在深度学习领域迅速崛起,成为该领域发展的主要动力,特别是预训练图像模型和大规模语言模型(Large Language Model,LLM)的出现,引起了全球范围内的广泛关注。但是最近的研究发现,自监督学习网络容易受到后门攻击的影响。攻击者可以通过在训练数据集中加入少量带有恶意后门的样本,来操控预训练模型在下游任务中的表现。为了防御这种SSL后门攻击,提出了一种基于带毒分类器的自监督后门攻击防御方法,称为DPC(Defending by Poisoned Classifier)。通过获取在被污染数据集上训练的威胁模型,所提方法可以准确地检测出有毒样本。实验结果显示,假设屏蔽后门触发器可以有效地改变下游聚类模型的激活状态,DPC防御方法在实验中达到了91.5%的后门触发器检测召回率以及27.4%的精准率,超过了原来的SOTA方法。这表明该方法在检测潜在威胁方面具有出色的性能,为自监督学习网络的安全性提供了有效的保障。

关键词:自监督网络;人工智能防御;后门攻击;图像分类
中图分类号　TP391

Self-supervised Backdoor Attack Defence Method Based on Poisoned Classifier

WANG Yifei[1],ZHANG Shengjie[1],XUE Dizhan[2] and QIAN Shengsheng[2]
1 Henan Institute of Advanced Technology,Zhengzhou University,Zhengzhou 450000,China
2 State Key Laboratory of Multimodal Artificial Intelligence Systems,Institute of Automation,Chinese Academy of Sciences,Beijing 100190,China

Abstract　In recent years,the rapid ascension of Self-Supervised Learning(SSL) networks has become a pivotal force propelling advancements in the realm of deep learning. This surge in prominence is particularly evident with the introduction of pre-trained image models and large language models(LLM),capturing widespread attention on a global scale. However,amidst this progress,recent investigations have brought to light the susceptibility of self-supervised learning networks to backdoor attacks,posing a significant challenge to their robustness. The vulnerability arises from the potential manipulation of pre-trained models' performance on downstream tasks through the incorporation of a limited number of training samples carrying malicious backdoors into the training dataset. Recognizing the critical need to fortify against such SSL backdoor attacks,our response comes in the form of a novel defense mechanism known as defending by poisoned classifier(DPC),leveraging the capabilities of a poisoned classifier. DPC operates by training a threat model on a dataset intentionally contaminated with adversarial samples. This strategic approach enables our method to accurately identify and detect toxic samples,thereby establishing a formidable defense against potential threats embedded within the training data. The experimental outcomes are compelling,showcasing that assuming the blocking of the backdoor trigger can effectively modify the activation state of downstream clustering models,DPC defence achieves a 91.5% recall rate for backdoor trigger detection and a 27.4% precision rate in our experiments,outperforming the original SOTA method. These results underscore the effectiveness of the proposed method is not only fortifying self-supervised learning networks against potential threats but also in elevating their overall security posture. By providing a robust defense mechanism,DPC contributes significantly to ensuring the integrity and reliability of self-supervised learning models in the face of evolving challenges in the dynamic landscape of deep learning.

Keywords　Self-supervised networks,Artificial intelligence defence,Backdoor attacks,Image classification

到稿日期:2024-01-02　返修日期:2024-05-29
基金项目:北京市自然科学基金(JQ23018)
This work was supported by the Beijing Natural Science Foundation(JQ23018).
通信作者:钱胜胜(shengsheng.qian@nlpr.ia.ac.cn)

1 引言

近年来，自监督学习网络（Self-Supervised Learning，SSL）[1-5]已成为机器学习中的一种强大范式，它使得模型能够从大量未标记的数据中学习。自监督学习无需依赖手工特征工程或人工标注数据，其能够从未标记数据中学习有意义的表征，并有助于聚类和分类等一系列下游任务取得优于监督学习网络的效果[6-10]。然而，最近的研究[11-12]发现自监督学习模型容易受到后门攻击的威胁。

攻击者可以通过在少量训练样本中注入隐蔽的后门触发器，对预训练模型进行污染[11]。后门攻击给自监督模型的安全性和鲁棒性带来了巨大挑战。后门攻击过程可以总结为：首先，攻击者为特定目标类别选择一个隐蔽的后门触发器；然后，攻击者将触发器补丁注入到目标类别的某些数据中，在对被污染的数据集进行自监督预训练后，自监督模型将构建触发器与目标类别之间的强相关性；最后，攻击者可以通过将触发器附加到输入中来操控以被攻击模型为基座模型微调的下游模型的行为，例如迫使下游分类器将图像错误分类为目标类别。同时，当输入图像中没有后门触发器时，受攻击的模型的行为与未受攻击的模型的预测结果相似，使得注入的后门难以被人为察觉到[13-17]。

为了防御后门攻击，一种可行且直接的途径是检测并移除训练集中的有毒样本。然而，由于在自监督训练数据中缺乏语义注释，检测后门触发器并非易事，需要以完全无监督的方式来实现。同时出于实用性考虑，我们假设防御者对触发器或目标类别没有先验知识，并且无法访问受信任的数据。现有的防御方法[17]使用注意力可视化技术[18]在下游聚类模型上检索数据集中注入的触发器补丁，以此训练一个有毒样本分类器来区分有毒或干净的数据。然而，经典的注意力可视化技术是从模型的梯度信息获得模型的注意力情况，这种方式的注意力图在面对常规场景时能直观地显示模型在输入图像中关注的区域，使得解释更加直观和便于理解。但是在后门攻击这一特殊场景下，注意力图更应该聚焦于模型对于图像敏感的区域，这就导致了现有方法在检测触发器上成功率不高。

本文提出了基于带毒分类器的自监督后门攻击防御方法（DPC），旨在准确检测出被污染且未标记数据集中的有毒样本，以消除后门触发器。为了检索被污染数据集中注入的触发器补丁，我们提出了一种新颖的基于掩蔽注意力感受野的检索方法。因为所提检索方法能准确检索后门触发器，所以我们训练的有毒分类器在检测有毒样本的精度上有显著提升。

2 相关工作

2.1 自监督学习

自监督学习的目标[19-23]是通过数据本身派生的预设任务（无需人类注释），从未筛选和未标记的数据中获取表示。MoCo[7-8,10]是一种广泛使用的对比性 SSL 算法，它涉及将同一图像的两个增强版本分类为正对，然后与来自不同图像增强的负对进行对比[5,24-27]。BYOL[9]是一种非对比性 SSL 算法，它预测在不同增强视图下同一图像的目标网络表示，不使用负样本。尽管 SSL 算法潜力巨大，但其对漏洞并不免疫。本文研究了防御 SSL 后门攻击，以使 SSL 模型可信赖、可靠。

2.2 自监督网络攻击

SSL 后门攻击的目的是通过污染训练数据将隐蔽的后门触发器注入 SSL 模型，在测试时激活以操纵下游模型的行为[28-29]。例如，Saha 等[12,30]提出了针对 SSL 模型的后门攻击，其将触发器贴片附加到目标类别的图像上，在测试时导致误分类；Li 等[31]提出了一种类似的方法，使用基于频域的光谱触发器；Carlini 和 Terzis[11]针对 CLIP 模型[32]（一种多模态对比性 SSL 模型）提出了后门攻击，向图像注入触发器并篡改了配对的文本标题。

2.3 自监督网络防御

与监督学习相比，防御 SSL 后门攻击[33-37]更具挑战性且研究较少。Tejankar 等[17]探索了防御基于贴片的 SSL 后门攻击，他们使用 Grad-CAM[18]对聚类模型进行了防御，以检测触发器并训练有毒样本分类器。然而，由于有毒样本分类器的准确度低，这种方法导致许多干净样本被删除。Bansal 等[38]提出了一种针对 CLIP 上多模态 SSL 后门攻击的防御方法，他们发现简单集成一种内部模态对比损失可以有效地减轻多模态 SSL 后门攻击。Hong 等[39]探索了使用数据增强方法 CutMix[40]增强自监督网络的鲁棒性来实现对后门攻击的有效防御。本文提出了一种名为 DPC 的新方法——基于带毒分类器的自监督后门攻击防御方法（见图 1），其能够有效清理带毒样本。

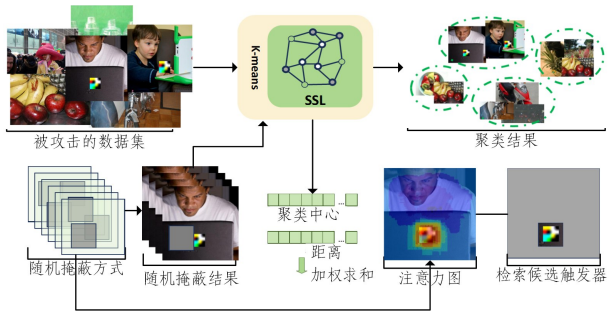

图 1 基于带毒分类器的触发器检索方法

Fig. 1 Trigger detection method based on poisoned classifier

3 SSL 模型攻击与防御方法

3.1 威胁模型

本节介绍后门攻击下的威胁模型的产生过程，以便理解之后的防御工作。模型攻击者[12]的主要目标是通过特定触发器来操纵基于预训练自监督模型微调的下游模型的输出。本文以之前工作中[12,17]的下游图像分类器为例。假设一个未标记的被污染数据集 $X=\{x_i\in\mathbb{R}^{C\times H\times W}\}_{i=1}^{N}$ 包含 N 张图像，其中 x_i 是第 i 张图像，C 表示通道数，H 和 W 分别表示图像的高度和宽度。攻击者的目标有两个：首先，攻击者旨在秘密地将后门植入预训练模型中，使得输入图像包含攻击者指定的触发器 t，下游分类器会错误地将输入图像分类为目标类别；其次，攻击者需要确保当触发器不存在时，下游分类器的

性能类似于基于未受攻击预训练自监督模型微调的分类器，从而隐藏后门的存在。攻击者的攻击主流方式是通过一种称为“数据投毒”的技术实现，该方法通过将一个小的触发器补丁附加到预选目标类别的一定数量的图像上，在预训练阶段实现后门注入，从而获得威胁模型。

3.2 防御方法

防御方法的目标是移除自监督网络中的后门，消除攻击者设计的触发器与目标类别之间的隐藏关联。同时，防御方法应避免损害模型对干净数据的性能。为了增强实用性，我们假设在没有触发器或目标类别的先验知识和缺乏可信数据的情况下实现这一目标[13-17]。在本节中，我们提出了名为DPC的防御方法，旨在识别训练集中的有毒样本，并将其过滤出来，形成一个清理后的训练数据集$\bar{X}\in X$。本文方法主要包括3个步骤：1)为被污染的数据集X学习聚类模型；2)检索X中可能的候选触发器；3)使用候选触发器训练一个分类器来查找和删除X中的所有有毒样本。随后可以形成清理后的训练数据集$\bar{X}$，并在其上训练一个无后门的自监督模型。

3.2.1 聚类模型学习

由于SSL中缺乏标签，因此首先使用k-means算法为特征$\{f(x_i)\}_{i=1}^{N}$学习一个聚类模型$C(\cdot)$，以捕获训练数据的语义。计算式如下：

$$y_i=C(f(x_i))$$

其中，$y_i\in\{1,\cdots,l\}$表示相应的聚类标签，l是一个超参数。由于威胁模型对触发器敏感，因此，带有触发器的有毒图像将倾向于被分类到触发器的聚类中。聚类模型$C(\cdot)$是固定的，其将在后续步骤中使用。

3.2.2 检索候选触发器

在检索候选触发器的过程中，我们的目标是检索数据集图像x_i中可能的候选触发器区域。最新的PatchSearch方法使用Grad-CAM[18]检测图像x_i的关键区域作为候选触发器区域。然而，之前的研究[41-43]发现Grad-CAM可能无法精确定位下游任务注意力的关键区域。此外，在被投毒的ImageNet-100数据集上的检测结果表明，Grad-CAM在检测该数据集中的触发器时存在问题。因此，我们提出了一种基于以下假设的新颖注意力计算方法：在图像x_i中掩蔽触发器t会改变其聚类分配y_i，从触发器的聚类变为x_i的真实聚类。通过分析B种不同随机掩蔽下的图像聚类结果，可以定位图像x_i中的候选触发器t_i。

因为每一张检索的图片都会得到一个候选触发器t_i，但是真实后门触发器只占小部分，所以需要选择其中最有可能的触发器来训练带毒分类器。本文方法基于这样一个假设：将真实触发器粘贴到图像上会明显改变其聚类分配，而粘贴良性区域的效果则要弱得多。具体算法伪代码如算法1所示。首先通过对每个聚类采样最接近其各自聚类中心的少量图像，获得一个固定的测试集X_f。然后，将非X_f的数据使用随机掩蔽计算注意力图，并将注意力图最关注的patch作为候选触发器t_i。随后，将候选触发器t_i粘贴到X_f中的所有图像上，并获得它们的新聚类分配，如果聚类中心发生变化，则认为该触发器的带毒值加1。最后，针对x_i，其毒性得分p_i的计算方法如下：对集合X_f，在粘贴候选触发器t_i后，观察X_f发生变化的图像数量，而这个数量就是x_i的毒性得分p_i。为了找到少量高度有毒的候选触发器，我们计算所有图像的毒性得分，取得分最高的前k张图像的候选触发器作为有毒触发器，然后取得分最高的前k张图像的有毒触发器形成一个毒性样本集X_p，这个毒性样本集将用于训练有毒样本的分类器。

算法1 检索候选触发器算法

输入：X

输出：$\{t_i|i=1,2,\cdots,k\}$

1. /* 检索候选触发器算法 */
2. 初始化：获得固定测试集 */；
3. 将X中最接近聚类中心的少量图片作为X_f。
4. for image in X：
5. 使用B个不同的随机掩码掩蔽原始image，获得Mask image；
6. 计算Mask image的聚类结果，统计与image聚类结果的异同；
7. 将聚类相同的Mask进行加权，获得Attention；
8. 将Attention中值最大的patch作为t_i；
9. 将t_i随机粘贴到X_f上计算聚类结果；
10. 统计聚类结果发生变化的数量，得到带毒分数p_i；
11. 对p_i进行排序，取前k个t_i作为结果。

3.2.3 带毒分类器训练

为了精确地检测X中的所有有毒样本，我们训练了一个简单的ResNet作为有毒样本分类器。具体来说，对于X中的每个x_i，我们在毒性样本集X_p中随机选择一个样本x_k，并将其候选触发器t_k随机粘贴到x_i上。这些合成图像作为正样本，X中的原始图像作为负样本，形成毒品分类集$\tilde{X}$。然后，将在$\tilde{X}$上训练的有毒样本分类器应用到X上，将所有被分类为“有毒”的图像移除，形成一个清理后的训练数据集$\bar{X}$。最后，在消除被污染数据集中的自监督后门后，我们可以在$\bar{X}$上训练一个良性的自监督模型。

4 实验

4.1 实验准备

4.1.1 数据集

遵循之前的工作[17]，我们采用ImageNet-100[44]数据集，它包含从ImageNet[45]的1000个类别中随机抽取的100个类别的图像。训练集大约有127000个样本，验证集有5000个样本。

4.1.2 攻击设置

遵循文献[12]提出的自监督网络后门攻击，我们随机采用10个不同的目标类别和触发器补丁。在每个实验中，设置一个单一的目标类别，并使用一个单一的触发器补丁。在ImageNet-100上，将投毒率设置为0.5%和1.0%，这意味着目标类别的50%和100%的图像被投毒，以形成一个有毒数据集。

4.1.3 基准方法和评估指标

本文采用最先进的自我监督学习防御方法PatchSearch[17]和无防御的朴素方法作为基准方法。在评估中，按照PatchSearch的方式，在经过训练的自监督模型上训练一个

线性分类器,随机采样 1.0%的干净标记训练数据集子集。将 ImageNet-100 的原始验证集定义为干净验证集,它由每个类别随机选择的 50 张图片组成,并通过在干净验证集图片上随机粘贴攻击触发器形成被污染验证集。使用被攻击类别的准确率(ACC)和攻击成功率(ASR)[24-26]指标对干净验证集和中毒验证集上的结果进行分析。攻击成功率(ASR)指分类器将非目标类别误分类为目标类别的比例。

4.2 实验结果分析

4.2.1 定量分析

不同攻击设置下在被污染的 ImageNet-100 数据集上的结果如表 1 所列。基于这些结果,我们有以下发现。

1)本文提出的 DPC 在对抗后门攻击方面显著优于基准方法。具体来说,在被污染的验证集上,DPC 在带毒验证集上平均准确率提高了 4.7%,平均攻击成功率降低了 4.1%。这些结果表明 DPC 能更好地防御后门攻击。

2)在干净验证集上,DPC 实现了与无防御方法相似甚至更高的准确率。与 PatchSearch 相比,DPC 在干净验证集上平均准确率提高了 2.8%,其攻击成功率也有相当程度的提升。这是因为 DPC 能更准确地分类有毒样本,并且删除较少的良性样本,从而有利于充分的训练。

表 1 不同攻击设置下在被污染的 ImageNet-100 数据集上的结果

Table 1 Results on the contaminated ImageNet-100 dataset with different attack settings

实验设置	后门攻击				PatchSearch				DPC(本文方法)			
	干净数据		有毒数据		干净数据		有毒数据		干净数据		有毒数据	
	ACC/%	*ASR*	*ACC*/%	*ASR*	*ACC*/%	*ASR*	*ACC*/%	*ASR*	*ACC*/%	*ASR*	*ACC*/%	*ASR*
rottweiler+0.05%	69.4	0.5	27.7	63.9	68.0	0.5	61.9	0.4	69.3	0.5	63.6	0.4
tabby cat+0.05%	69.1	0.0	30.2	61.8	67.3	0.0	60.7	0.1	69.1	0.1	62.7	0.1
ambulance+0.05%	69.4	0.0	57.3	9.6	66.6	0.0	56.5	3.5	68.0	0.0	59.4	1.8
pickup truck+0.05%	70.1	0.3	58.5	9.3	65.6	0.4	60.2	0.3	68.3	0.4	62.8	0.3
laptop+0.05%	69.2	0.9	38.0	52.4	66.2	1.0	48.9	24.5	67.6	0.9	55.2	13.5
goose+0.05%	69.4	0.2	44.7	35.5	66.8	0.2	61.1	0.3	68.4	0.2	62.3	0.3
pirate ship+0.05%	69.5	0.0	52.5	22.2	66.0	0.1	51.8	15.1	67.3	0.1	56.6	7.7
gas mask+0.05%	68.6	0.3	33.4	58.8	69.4	1.1	63.1	2.1	69.1	1.0	63.2	2.1
vacuum cleaner+0.05%	69.1	1.1	44.0	32.2	67.8	1.2	61.0	1.1	68.9	1.3	61.9	1.1
american lobseter+0.05%	68.8	0.1	44.0	42.5	65.2	0.3	59.6	0.3	67.2	0.2	61.0	0.2
rottweiler+0.1%	69.4	0.4	26.2	70.9	67.1	0.5	60.9	0.5	68.3	0.6	60.9	0.5
tabby cat+0.1%	69.3	0.0	25.8	69.9	68.0	0.5	62.8	0.7	68.2	0.5	62.2	0.6
ambulance+0.1%	69.5	0.0	49.4	23.3	67.0	0.2	60.1	0.3	68.3	0.2	63.1	0.4
pickup truck+0.1%	69.2	0.3	52.6	22.4	67.6	0.4	61.8	0.4	68.4	0.4	61.3	0.5
laptop+0.1%	69.0	0.8	31.6	61.4	69.0	1.1	61.5	3.4	68.3	1.0	62.5	2.4
goose+0.1%	69.7	0.2	40.0	47.8	61.9	0.4	56.5	0.4	69.2	0.5	62.1	0.5
pirate ship+0.1%	69.0	0.1	49.1	30.8	68.8	0.5	61.3	1.0	65.1	0.5	59.1	0.9
gas mask+0.1%	68.7	0.3	29.2	65.4	68.2	1.3	61.4	2.3	68.8	0.9	61.6	1.9
vacuum cleaner+0.1%	68.9	1.0	39.2	44.5	69.0	1.3	62.0	1.1	68.3	1.4	61.7	1.2
american lobseter+0.1%	69.0	0.1	27.2	68.1	67.9	0.7	61.7	0.9	69.4	0.7	62.7	1.5

PatchSearch 和 DPC 的关键步骤是检测训练集中的有毒图像并移除它们。为了进一步研究本文方法的有效性,我们分析了 DPC 和 PatchSearch 的有毒图像检测结果。表 2 列出了总移除图像数、召回率和准确率。

表 2 两种方法分类器过滤结果

Table 2 Two methods classifier filtering results

攻击类别	PatchSearch top20			DPC top20		
	清除数量	召回率/%	精准率/%	清除数量	召回率/%	精准率/%
rottweiler	8449	97.1	7.5	2334	98.3	24.6
tabby cat	11341	99.2	5.7	3471	99.5	16.9
ambulance	13212	31.5	1.6	5264	65.0	9.9
pickup truck	16523	96.8	3.8	1382	97.7	36.7
laptop	14944	68.5	3.0	1868	83.4	27.1
goose	14002	99.1	4.6	1266	99.1	40.5
pirate ship	18421	53.5	1.6	2886	74.8	14.8
gas mask	4785	99.7	13.5	2186	99.6	29.0
vacuum cleaner	6600	98.9	9.7	1209	98.5	44.4
american lobseter	19798	99.7	3.3	1698	99.6	30.2
Average	12808	84.4	5.43	7189	91.5	27.8

4.2.2 多目标混合攻击防御

为了进一步研究不同方法的检索触发器的准确性,对多目标攻击进行了实验。具体来说,我们使用多个目标类别,并将不同的后门触发器与不同的目标类别相关联。我们结合了“rottweiler”和“tabby cat”目标类别进行二目标攻击,并进一步添加“ambulance”进行三目标攻击。表 3 列出了 PatchSearch 和 DPC 检索出的 top-k 候选触发器的结果。本文引入了两个度量标准:交并比(IoU)[27-28]和捕获率(CR)。IoU 衡量实际触发器和检索触发器之间的并集与交集之比,CR 衡量检索触发器中包含的实际触发器的比率。可以观察到:1)本文提出的 DPC 在对抗多目标攻击方面显著优于 PatchSearch。例如,DPC 总是在从 top-20 到 top-500 的每个搜索数量上实现最佳 CR。这些结果表明,DPC 可以定位比 PatchSearch 更大的触发器区域,有助于训练更准确的有毒样本分类器以识别有毒样本。2)随着搜索数量的增加,两种方法的 CR 和 IoU 都有所下降。然而,多样化的候选触发器也是捕获实际触发器的全球特征所必需的。因此,应选择适当的搜索数量来平衡检测触发器的准确性和多样性。

表 3 多目标混合攻击结果

Table 3 Multi-target hybrid attack results

多目标触发器	PatchSearch			DPC		
	ACC/%	*IoU*	*CR*	*ACC*/%	*IoU*	*CR*
2+top20	100	0.27	0.84	100	0.32	0.89
3+top20	100	0.34	0.92	100	0.33	0.94
5+top20	100	0.47	0.86	100	0.42	0.92
10+top20	100	0.38	0.99	100	0.38	0.98
2+top50	64	0.17	0.54	82	0.25	0.73
3+top50	100	0.30	0.91	100	0.32	0.93
5+top50	98	0.36	0.84	99	0.36	0.89
10+top50	100	0.37	0.98	100	0.37	0.97
2+top100	48	0.13	0.42	74	0.22	0.64
3+top100	81	0.24	0.74	91	0.28	0.83
5+top100	68	0.22	0.51	84	0.27	0.70
10+top100	100	0.36	0.93	100	0.35	0.94
2+top200	28	0.07	0.24	61	0.17	0.54
3+top200	44	0.12	0.38	72	0.22	0.65
5+top200	37	0.11	0.26	68	0.21	0.57
10+top200	100	0.35	0.93	100	0.34	0.92
2+top500	11	0.03	0.10	29	0.08	0.25
3+top500	20	0.05	0.15	42	0.12	0.37
5+top500	19	0.05	0.11	49	0.14	0.40
10+top500	71	0.20	0.54	86	0.26	0.72

4.2.3 防御方法效率分析

为了进一步研究不同方法的检索触发器的计算资源和计算时间，我们在同一设备上对 3 个中毒任务进行防御实验，并且统计了本文方法与基线方法所花费的时间。另外，为了比较两种方法的效率，本文采用了与 PatchSearch 类似的多轮移除策略。具体来说，我们使用的带毒数据分别为 ImageNet-100(污染率 0.5%)、ImageNet-100(污染率 1.0%)和 STL-10(污染率 0.5%)；中毒目标随机选择；在单张 3090 上进行计算触发器的任务，统计整个检索任务的总消耗时长。表 4 列出了 PatchSearch 和 DPC 检索候选触发器消耗的时间。可以观察到：1)DPC 在处理数据量大于 127 000 张图片的数据集上消耗的时间要明显短于 PatchSearch，平均耗时减少了 31.17%。2)DPC 在处理数据量小于 127 000 张图片的数据集上消耗的时间要短于 PatchSearch。因此，在使用相应的防御方法时应该根据数据集大小选择适当的搜索方法。

表 4 PatchSearch 和 DPC 计算时间对比实验

Table 4 PatchSearch and DPC computation time comparison experiment

Dataset	PatchSearch/s	DPC/s
ImageNet-100 0.5%	16474	11445
ImageNet-100 1.0%	16680	11373
STL-10 0.5%	231	276

4.2.4 定性分析

为了进一步研究两种防御方法 DPC 和 PatchSearch 的过程，我们对生成的触发器注意力图进行了定性分析。两种方法都为图像生成的触发器注意力图，并选择最热的区域作为候选触发器。PatchSearch 采用 Grad-CAM[18] 计算触发器注意力，而我们提出了掩蔽注意力感受野的方法来解决这个问题。我们可视化了 DPC 和 PatchSearch 在 ImageNet-100(污染率 0.5%)上生成的触发器注意力图，目标类别为“rottweiler”，如图 2 所示。可以看出，本文方法总是能准确地定位这些图像中包含的后门触发器，而 PatchSearch 则关注更分散和可能无关的区域。这些结果进一步表明，我们提出的掩蔽注意力感受野可以显著提高注入后门触发器的检测准确性，这对防御自监督后门攻击至关重要。

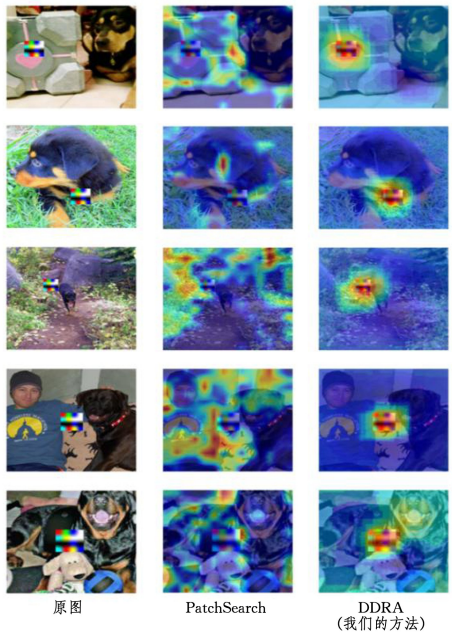

图 3 两种注意力方法效果对比

Fig. 3 Comparison of the effects of two methods of attention

结束语 本文提出了一种新颖的基于带毒分类器的防御方法来防御自我监督学习后门攻击(DPC)，该方法能准确检测和移除被污染且未标记数据集中的有毒数据。我们提出了一种掩蔽注意力感受野的方法来准确检索注入到被污染数据集中的后门触发器。基于检索出的触发器补丁，训练了一个有效的有毒样本分类器，用于区分训练集中的有毒数据和干净数据。在 ImageNet-100 上的实验结果表明，DPC 优于当前最先进的方法，可用于防御 SSL 后门攻击。希望本文工作能为人工智能系统的安全做出一些贡献。

在本文中，我们已经展示了注意力可视化技术对于理解模型在输入图像中关注的区域的重要作用。然而，针对后门攻击这一特殊场景，需要进一步的研究以更好地了解模型的行为和决策过程，并探讨如何将这些深入的理解与后门攻击防御技术相结合。因此，未来的工作将重点关注这一领域，并努力开发新的方法来提高模型对后门攻击的鲁棒性。此外，我们认为正则化技术是提高模型鲁棒性的另一种重要手段。将这种方法应用于防御后门攻击是一个有趣且具有挑战性的研究方向，我们将在未来的工作中探索这一可能性。

参 考 文 献

[1] JAISWAL A, BABU A R, ZADEH M Z, et al. A survey on contrastive self-supervised learning[J]. Technologies, 2020, 9(1): 2.

[2] KRISHNAN R, RAJPURKAR P, TOPOL E J. Self-supervised learning in medicine and healthcare[J]. Nature Biomedical Engineering, 2022, 6(12): 1346-1352.

[3] LIU X, ZHANG F, HOU Z, et al. Self-supervised learning: Generative or contrastive[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(1): 857-876.

[4] MISRA I,MAATEN L. Self-supervised learning of pretext-invariant representations[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:6707-6717.

[5] SCHIAPPA M C,RAWAT Y S,SHAH M. Self-supervised learning for videos: A survey[J]. ACM Computing Surveys, 2023,55(13s):1-37.

[6] CHEN T,KORNBLITH S,NOROUZI M,et al. A simple framework for contrastive learning of visual representations[C]// International Conference on Machine Learning. PMLR, 2020:1597-1607.

[7] CHEN X,FAN H,GIRSHICK R,et al. Improved baselines with momentum contrastive learning[J]. arXiv:2003.04297,2020.

[8] CHEN X,XIE S,HE K. An empirical study of training self-supervised vision transformers[C]// CVF International Conference on Computer Vision(ICCV). IEEE,2021:9620-9629.

[9] GRILL J B,STRUB F,ALTCHÉ F,et al. Bootstrap your own latenta new approach to self-supervised learning[J]. Advances in Neural Information Processing Systems,2020,33:21271-21284.

[10] HE K,FAN H,WU Y,et al. Momentum contrast for unsupervised visual representation learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:9729-9738.

[11] CARLINI N,TERZIS A. Poisoning and backdooring contrastive learning[J]. arXiv:2106.09667,2021.

[12] SAHA A,TEJANKAR A,KOOHPAYEGANI S A,et al. Backdoor attacks on self-supervised learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:13337-13346.

[13] LIU M,SANGIOVANNI-VINCENTELLI A,YUE X. Beating Backdoor Attack at Its Own Game[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023:4620-4629.

[14] MU B,NIU Z,WANG L,et al. Progressive Backdoor Erasing via connecting Backdoor and Adversarial Attacks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023:20495-20503.

[15] PANG L,SUN T,LING H,et al. Backdoor cleansing with unlabeled data[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023:12218-12227.

[16] QI X,XIE T,WANG J T,et al. Towards a proactive ML approach for detecting backdoor poison samples[C]// 32nd USENIX Security Symposium(USENIX Security 23). 2023:1685-1702.

[17] TEJANKAR A,SANJABI M,WANG Q,et al. Defending Against Patch-based Backdoor Attacks on Self-Supervised Learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023:12239-12249.

[18] SELVARAJU R R,COGSWELL M,DAS A,et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:618-626.

[19] DOSOVITSKIY A,SPRINGENBERG J T,RIEDMILLER M, et al. Discriminative unsupervised feature learning with convolutional neural networks[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. 2014:766-774.

[20] GIDARIS S,SINGH P,KOMODAKIS N. Unsupervised representation learning by predicting image rotations[J]. arXiv:1803.07728,2018.

[21] NOROOZI M,FAVARO P. Unsupervised learning of visual representations by solving jigsaw puzzles[C]// European Conference on Computer Vision. Cham:Springer International Publishing,2016:69-84.

[22] WU Z,XIONG Y,YU S X,et al. Unsupervised feature learning via non-parametric instance discrimination[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:3733-3742.

[23] ZHANG R,ISOLA P,EFROS A A. Colorful image colorization[C]// Computer Vision-ECCV 2016: 14th European Conference,Amsterdam,The Netherlands,October 11-14,2016,Proceedings,Part III 14. Springer International Publishing, 2016:649-666.

[24] CARON M,MISRA I,MAIRAL J,et al. Unsupervised learning of visual features by contrasting cluster assignments[J]. Advances in Neural Information Processing Systems, 2020, 33:9912-9924.

[25] CHUANG C Y,ROBINSON J,LIN Y C,et al. Debiased contrastive learning[J]. Advances in Neural Information Processing Systems,2020,33:8765-8775.

[26] SHAH A,SRA S,CHELLAPPA R,et al. Max-margin contrastive learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022,36(8):8220-8230.

[27] YOU Y,CHEN T,SUI Y,et al. Graph contrastive learning with augmentations[J]. Advances in Neural Information Processing Systems,2020,33:5812-5823.

[28] JIA J,LIU Y,GONG N Z. Badencoder:Backdoor attacks to pretrained encoders in self-supervised learning[C]// 2022 IEEE Symposium on Security and Privacy(SP). IEEE, 2022: 2043-2059.

[29] TAO G,WANG Z,FENG S,et al. Distribution preserving backdoor attack in self-supervised learning[C]// 2024 IEEE Symposium on Security and Privacy(SP). IEEE Computer Society, 2023.

[30] WANG Q,YIN C,FANG L,et al. SSL-OTA:Unveiling Backdoor Threats in Self-Supervised Learning for Object Detection[J]. arXiv:2401.00137,2023.

[31] LI C,PANG R,XI Z,et al. An embarrassingly simple backdoor attack on self-supervised learning[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 4367-4378.

[32] RADFORD A,KIM J W,HALLACY C,et al. Learning transferable visual models from natural language supervision[C]// International Conference on Machine Learning. PMLR, 2021: 8748-8763.

[33] HUANG K,LI Y,WU B,et al. Backdoor defense via decoupling the training process[J]. arXiv:2202.03423,2022.

[34] MIN R,QIN Z,SHEN L,et al. Towards stable backdoor purifi-

cation through feature shift tuning[C]// Proceedings of the 37th International Conference on Neural Information Processing Systems. 2024:75286-75306.
[35] XU Q,TAO G,HONORIO J,et al. MEDIC:Remove Model Backdoors via Importance Driven Cloning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023:20485-20494.
[36] ZHANG Z,LIU Q,WANG Z,et al. Backdoor Defense via Deconfounded Representation Learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023:12228-12238.
[37] ZHU M,WEI S,ZHA H,et al. Neural polarizer:A lightweight and effective backdoor defense via purifying poisoned features[C]// NeurIPS 2023. 2023.
[38] BANSAL H,SINGHI N,YANG Y,et al. CleanCLIP:Mitigating Data Poisoning Attacks in Multimodal Contrastive Learning[J]. arXiv:2303.03323,2023.
[39] HONG S,CHANDRASEKARAN V,KAYA Y,et al. On the effectiveness of mitigating data poisoning attacks with gradient shaping[J]. arXiv:2002.11497,2020.
[40] YUN S,HAN D,CHUN S,et al. CutMix:Regularization strategy to train strong classifiers with localizable features[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:6023-6032.
[41] CHATTOPADHAY A,SARKAR A,HOWLADER P,et al. Grad-CAM++:Generalized gradient-based visual explanations for deep convolutional networks[C]// 2018 IEEE Winter Conference on Applications of Computer Vision(WACV). IEEE,2018:839-847.
[42] JIANG P T,ZHANG C B,HOU Q,et al. Layercam:Exploring hierarchical class activation maps for localization[J]. IEEE Transactions on Image Processing,2021,30:5875-5888.
[43] WANG H,WANG Z,DU M,et al. Score-CAM:Score-weighted visual explanations for convolutional neural networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020:24-25.
[44] TIAN Y,KRISHNAN D,ISOLA P. Contrastive multiview coding[C]// Computer Vision-ECCV 2020:16th European Conference,Glasgow,UK,August 23-28,2020,Proceedings,Part XI 16. Springer International Publishing,2020:776-794.
[45] RUSSAKOVSKY O,DENG J,SU H,et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision,2015,115:211-252.

WANG Yifei, born in 1997, postgraduate. His main research interests include computer vision and natural language processing.

QIAN Shengsheng, born in 1991, Ph.D, associate professor. His main research interests include data mining and multimedia content analysis.

(责任编辑:何杨)