基于后门攻击的恶意流量逃逸方法

马博文，郭渊博，马骏，张琦，方晨
（信息工程大学密码工程学院，河南 郑州 450001）

摘　要：针对基于深度学习模型的流量分类器，提出了一种利用后门攻击实现恶意流量逃逸的方法。通过在训练过程添加毒化数据将后门植入模型，后门模型将带有后门触发器的恶意流量判定为良性，从而实现恶意流量逃逸；同时对不含触发器的干净流量正常判定，保证了模型后门的隐蔽性。采用多种触发器分别生成不同后门模型，比较了多种恶意流量对不同后门模型的逃逸效果，同时分析了不同后门对模型性能的影响。实验验证了所提方法的有效性，为恶意流量逃逸提供了新的思路。

关键词：后门攻击；恶意流量逃逸；深度学习；网络流量分类
中图分类号：TP393
文献标志码：A
DOI：10.11959/j.issn.1000−436x.2024077

Escape method of malicious traffic based on backdoor attack

MA Bowen, GUO Yuanbo, MA Jun, ZHANG Qi, FANG Chen
Cryptography Engineering Institute, Information Engineering University, Zhengzhou 450001, China

Abstract: Launching backdoor attacks against deep learning (DL)-based network traffic classifiers, and a method of malicious traffic escape was proposed based on the backdoor attack. Backdoors were embedded in classifiers by mixing poisoned training samples with clean samples during the training process. These backdoor classifiers then identified the malicious traffic with an attacker-specific backdoor trigger as benign, allowing the malicious traffic to escape. Additionally, backdoor classifiers behaved normally on clean samples, ensuring the backdoor's concealment. Different backdoor triggers were adopted to generate various backdoor models, the effects of different malicious traffic on different backdoor models were compared, and the influence of different backdoors on the model's performance was analyzed. The effectiveness of the proposed method was verified through experiments, providing a new approach for escaping malicious traffic from classifiers.

Keywords: backdoor attack, escape of malicious traffic, deep learning, network traffic classification

0　引言

流量分类在帮助研究人员了解流量类型、提高服务质量的同时，大量应用于网络入侵检测以及恶意软件检测，成为解决网络安全问题的重要手段。近年来，随着人工智能技术的高速发展，深度学习（DL, deep learning）技术大量应用于流量分类，在流量分类的速度和成功率等方面有了大幅度的提升[1]。

DL 模型在帮助用户解决安全问题的同时，也带来了新的安全问题。从攻击方式上看，针对 DL 模型准确率的攻击手段主要有投毒攻击[2]、对抗样本攻击[3]以及后门攻击[4]。投毒攻击使模型泛化性能下降，无法做出正确的预测。对抗样本攻击通过

收稿日期：2023–09–27；修回日期：2024–03–12
通信作者：郭渊博，yuanbo_g@hotmail.com
基金项目：国家自然科学基金资助项目（No.62276091）；国家社会科学基金资助项目（No.2022-SKJJ-B-057）
Foundation Items：The National Natural Science Foundation of China (No.62276091), The National Social Science Fund of China (No.2022-SKJJ-B-057)

构造能够欺骗模型的对抗样本，使模型对该样本产生错误预测。后门攻击迫使模型学到攻击者指定的内容，对干净样本给出正常预测，对后门样本给出攻击者指定的预测。从攻击效果上看，上述攻击又可分为非目标攻击以及特定目标攻击。非目标攻击使分类器将输入样本错误预测为其他类；特定目标攻击则使分类器将输入样本预测为特定的目标类。从这一角度看，投毒攻击属于非目标攻击；对抗样本攻击则可以通过不同的构造方式实现非目标攻击以及目标攻击；对于后门攻击，由于攻击者可以通过特定的后门触发器实现对模型的控制，使模型对后门样本给出指定预测，因此属于特定目标攻击。

和对抗样本攻击以及中毒攻击相比，后门攻击由于其隐蔽性和有效性受到广泛关注，具有以下性质。

性质1 后门隐蔽性。

$$F_b(x)=F_o(x)=y \tag{1}$$

其中，F_b表示后门模型，F_o表示原模型，x为不含触发器的正常输入样本，y为样本x对应的原始标签。对于不含特定触发器的干净输入，后门模型正常工作，给出与原模型同样的结果。除掌握触发器的后门攻击者外，其他攻击者以及模型使用者均无法控制激活模型后门，体现了后门的隐蔽性。

性质2 后门有效性。

$$F_b(x')=y_t \tag{2}$$

其中，x'为包含触发器的输入样本，y_t为样本x'对应的目标标签，由后门攻击者指定。当输入样本包含特定触发器时，模型后门被激活，给出攻击者指定的预测结果。利用该性质，后门攻击者可通过在样本中植入触发器，迫使模型执行后门任务，改变其预测结果。

在信息安全领域，逃逸是指恶意样本通过某种方式欺骗检测模型，使之被检测为良性。对于流量分类用户，若攻击者将特定目标攻击中的目标类设定为“良性”，则可以利用该攻击实现恶意样本的逃逸。现有的恶意流量逃逸多基于对抗样本技术实现，但存在探测次数多、易暴露等问题。本文将后门攻击引入流量逃逸，对基于后门攻击的恶意流量逃逸技术进行研究，通过对分类器植入后门，并在恶意流量上添加对应的后门触发器，使后门模型将其识别为良性，成功实现恶意流量逃逸。首先，构建了基于LeNet、VGG、ResNet、DenseNet等多种DL模型的流量分类器，均实现了90%以上的分类成功率；其次，利用毒化数据训练技术，将3种不同后门植入上述分类器，生成不同的后门模型；最后，对后门模型进行测试，分析了不同后门对分类器初始性能的影响，同时比较了不同恶意流量对不同后门模型的逃逸效果。

1 相关工作

1.1 基于深度学习的流量分类及恶意流量逃逸

随着人工智能技术不断发展，以深度学习模型为基础的流量分类研究不断增加。文献[5]利用卷积神经网络实现恶意流量分类，达到了较好的效果。文献[6]将自注意机制引入流量分类，优化了自动化流量特征选择，同时实现了一定程度的可解释性。文献[7]提出了一种基于对比增量学习的细粒度恶意流量分类方法，实现了对新增恶意流量类别的快速识别。此外，流量分类也大量应用于恶意软件检测，由于攻击者通常利用网络流量来获取用户的敏感信息或与恶意软件进行交互，因此防御者可以通过网络流量分类推断出其背后的恶意软件。文献[8]通过分析多个层次的网络流量特征，并将这些特征与深度学习算法相结合，实现恶意软件分类的同时对分类结果进行了合理解释。文献[9]提出一种基于深度神经网络的DeepAMD方法，将应用程序接口（API）调用与安卓流量等特征相结合，在静态层和动态层均实现了较好的安卓恶意软件检测效果。

流量逃逸方面，现有的流量逃逸方法大多基于对抗样本技术[10]。文献[11]将流量分类视为一种攻击手段，从主动防御的角度出发，基于对抗样本技术提出了一种网络欺骗流量生成方法，实现了对LeNet-5流量分类器的逃逸。文献[12]用深度神经网络分类器实现加密流量分类，用对抗样本攻击实现加密流量对分类器的逃逸，同时探讨了相应的防御方法。文献[13]将生成对抗网络和主动学习进行结合，提出了一种针对流量分类器的对抗样本攻击方法，减少了对目标模型的访问次数。上述研究表明，对抗样本攻击在流量逃逸上取得了较好的效果，但生成对抗样本所需的扰动往往需要攻击者根据样本和分类器进行定制，且需要对分类器进行多次探测，增加了攻击者的暴露风险。

1.2　深度学习中的后门攻击

后门攻击的多数研究工作主要聚焦于计算机视觉领域[14-25]。Gu 等[15]在 2017 年提出了 BadNet，攻击者通过在交通标识检测模型中植入后门，使模型将添加了后门触发器的“停止”标识识别为“限速”标识，首次揭示了机器学习供应链潜藏的安全隐患，也为后续的后门攻击研究奠定了基础。此后，研究者从后门隐蔽性、标签一致性、后门植入方式等角度对后门攻击进行研究。文献[20-21]实现了毒化数据与其真实标签的一致性，增加了后门攻击的隐蔽性。文献[22-23]通过篡改模型参数的方式实现后门植入，文献[24-25]则直接修改原始模型结构完成后门植入。此外，后门攻击在自然语言处理[26-27]、图神经网络[28]、强化学习[29]、联邦学习[30]、恶意软件检测[31-34]等方面均具有较广泛的研究，对诸多领域和任务造成严重威胁。在流量分类领域，文献[35]首次实现了针对流量分类器的后门攻击，达到了 98.3% 的攻击成功率；文献[36]将对抗样本中的通用对抗扰动思想和后门攻击相结合，对加密流量进行后门攻击；文献[37]在后门攻击中引入模型可解释技术，以较低的中毒率达到了较显著的攻击效果。从上述文献看，针对流量分类的后门攻击相关研究仍处于起步阶段，且尚未出现利用后门攻击实现恶意流量逃逸的研究。本文将后门攻击的思想与流量逃逸相结合，提出了一种基于后门攻击的恶意流量逃逸方法，成功实现了多类型恶意流量的逃逸。和对抗样本逃逸方法相比，攻击者只需在后门模型部署后在恶意流量上添加预先设定好的触发器，便可成功实现逃逸，不需要多次探测模型，在实现对模型精准控制的同时增加了攻击隐蔽性。

2　恶意流量逃逸方法

2.1　逃逸场景

从深度学习的生命周期来看，后门攻击往往发生在模型的训练阶段。当该阶段并不完全为用户所控制时，攻击者便可实施后门攻击，从而实现恶意流量逃逸。本文假设逃逸场景为外包训练，该场景包含以下过程。

1) 用户（受害者）将选定的分类器框架 F（如分类器结构、所含层数、每层规模大小、激活函数类型等）提供给第三方（攻击者）。

2) 攻击者对模型进行训练，并将训练好的模型 F_θ 返还给用户。

3) 用户使用自己的测试集 D_{valid} 检查训练模型 F_θ 的准确性，且仅在准确性满足要求时才接受模型，否则拒绝模型。

攻击者具有以下能力：对训练集进行任意修改，包括增加样本、删除样本、修改样本内容及对应标签等；改变训练过程的配置，例如学习速率或批次大小，或者手动设置模型参数。但攻击者不能修改用户指定的模型框架，也无法获得用户测试集 D_{valid} 的任何信息。

若用户成功接受模型，则攻击者只需在需要逃逸的恶意流量上添加触发器，即可实现恶意流量逃逸。

在更宽泛的场景下（如用户直接使用第三方提供的预训练模型），本文所提方法同样适用。

2.2　逃逸方法

从后门植入方式而言，后门攻击可分为基于模型篡改的后门攻击以及基于毒化数据训练的后门攻击[4]。基于模型篡改的后门攻击通过篡改模型实现后门植入，不适于本文的攻击场景；基于毒化数据训练的后门攻击会对训练集中部分样本进行毒化，利用毒化训练集训练模型，从而实现后门植入，本文基于此种攻击方式，设计了流量分类模型后门植入算法，如算法 1 所示。

算法 1　流量分类模型后门植入算法

输入　网络流量训练集 D_{train}、流量分类模型框架 M、毒化数据比例 PR、目标标签 Y_{target}

输出　后门模型 M_{backdoor}

```
preprocess(D_train)//预处理 D_train
poison(D_train)//毒化训练集
{
poison index=permutation(len(D_train)PR)//依据毒化数据比例，随机产生毒化数据索引
  for x in D_train:
    if index(x) in poison index:
      add a trigger to data(x)//添加后门触发器
      set label(x)=Y_target//改变毒化数据标签
    else:
      pass
    end if
```

12)　end for

13)　}

14)　return D_{poison}//返回毒化训练集

15)　training(M, D_{poison})//利用毒化训练集训练模型

16)　return M_{backdoor}//输出后门模型

算法 1 对应的流程如图 1 所示。对原始数据进行预处理后，攻击者可按照预设好的比例随机产生毒化数据索引，对毒化数据添加后门触发器并修改标签为目标标签，与索引外的正常数据合并后对模型进行训练。训练过程中，由于毒化数据的存在，模型可以学到目标标签和后门触发器之间的对应关系。训练完成后的模型即后门模型。

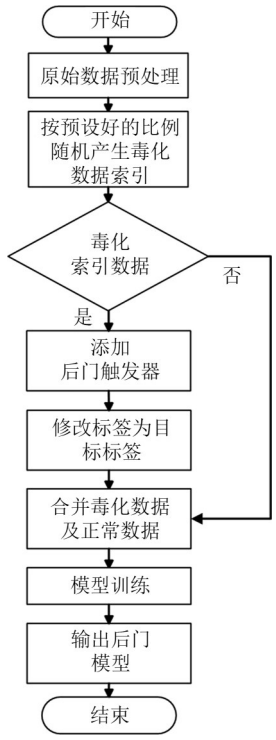


图 1　算法 1 对应的流程

在后门选择上，本文使用 3 种具有代表性的后门样式。

1) BadNet

BadNet[15]是深度学习领域首个后门攻击样式，该后门的触发器具有固定位置和模式。

2) Blended

Blended 后门由 Chen 等[17]提出，通过将触发器与原始样本以一定比例进行混合，在 BadNet 基础上提升了触发器的隐蔽性。

3) Sig

Sig 后门由 Barni 等[18]提出，用水平正弦信号作为触发器，其定义式为

$$v(i,j)=\Delta\sin\left(\frac{2\pi jf}{m}\right),1\leqslant j\leqslant m,1\leqslant i\leqslant l \tag{3}$$

为了进一步提升后门数据隐蔽性，文献[18]在对毒化数据添加触发器后并不改变其标签，而是通过增加毒化数据比例的方式，强制模型学习触发器和目标标签之间的关系。由于本文设定的攻击场景为外包训练，受害者无法直接接触训练数据，不需要考虑数据的标签一致性，本文对后门数据使用 Sig 后门的同时更改对应标签。

对于由算法 1 得到的后门模型，若要实现恶意流量逃逸，必须满足以下 2 个条件：①对于用户测试集，后门模型的分类性能不受后门影响，以保证其通过用户测试，从而顺利实现部署；②对于带有触发器的恶意流量，后门模型将其分类为良性，从而实现恶意流量的顺利逃逸。

事实上，条件①恰为式(1)给出的后门模型隐蔽性条件，由于用户事先不了解后门的存在，故其所用测试集不包含触发器，而后门隐蔽性保证了后门模型在非触发器样本上的正常工作，使后门模型可以顺利通过用户测试。条件②则为式(2)给出的后门模型有效性条件，攻击者只需在模型训练阶段将目标标签y_t设置为良性，并在模型使用阶段对恶意流量添加对应的后门触发器，即可实现逃逸。恶意流量逃逸示意如图 2 所示。

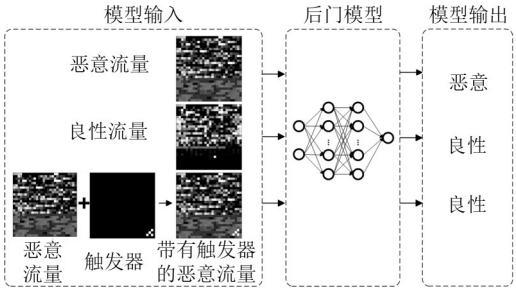


图 2　恶意流量逃逸示意

同时，由于模型后门满足隐蔽性，在模型的正常使用过程中，对其他不含触发器的恶意流量，分类器可以正常检出，这在一定程度上也会增强用户对后门模型的信任。后门攻击者甚至可以故意发送不含触发器的恶意流量使其被后门模型检出，以进一步增强用户的信任。

3　实验分析

3.1　数据集及预处理

本文使用CICAndMal2017数据集[38]进行实验，该数据集是一个安卓流量数据集，由加拿大网络安全研究院利用真实移动设备生成。数据集包含benign、adware、scareware、ransomware、smsmalware等5个类别的安卓软件及其对应的流量文件。benign类型为良性，对应的流量为良性流量，adware、scareware、ransomware、smsmalware分别表示广告型恶意软件、恐吓型恶意软件、勒索型恶意软件以及短信型恶意软件，其对应的流量为恶意流量。

按照流量粒度，对流量分类的研究主要针对TCP连接、流、会话、服务、主机等5个层面，其中基于流和会话的研究应用范围最广，流由具有相同五元组（源IP、源端口、目的IP、目的端口、传输层协议）的所有包组成，会话由双向流组成，本文在会话层面进行流量分类。分类前首先对原始流量进行预处理，过程如图3所示。

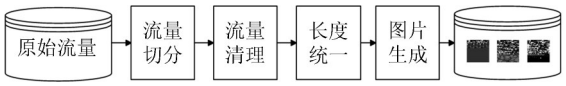


图3　流量预处理过程

预处理过程基于USTC-TK2016[5]工具，将原始流量转换为流量灰度图，具体步骤如下。

步骤1　流量切分。将每一份原始流量数据切分为会话数据。

步骤2　流量清理。删除步骤1产生的重复数据以及长度为0的数据。

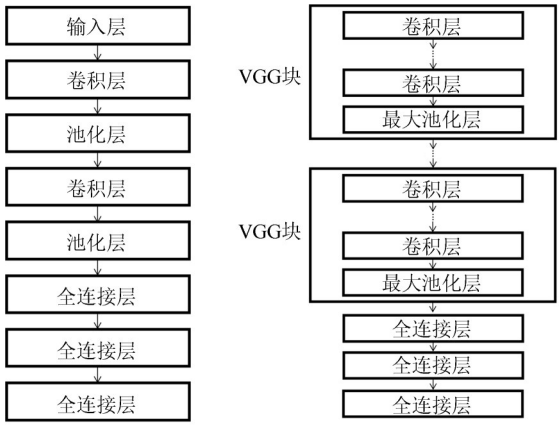
步骤3　长度统一。将步骤2中清理过的会话数据进行归一化处理，只保留前784 B，对于不足784 B的在后面补0。

步骤4　图片生成。以字节的二进制表示转换为对应灰度像素值的形式，将步骤3得到的数据转换为28×28的灰度图。

选取CICAndMal2017数据集中所有5类初始流量，对原始数据集进行预处理后，在每一类中选取12 000个流量图像样本，随机选取10%构成测试集，其余部分构成训练集。所有实验都在NVIDIA GeForce RTX 3080 TI上实现。

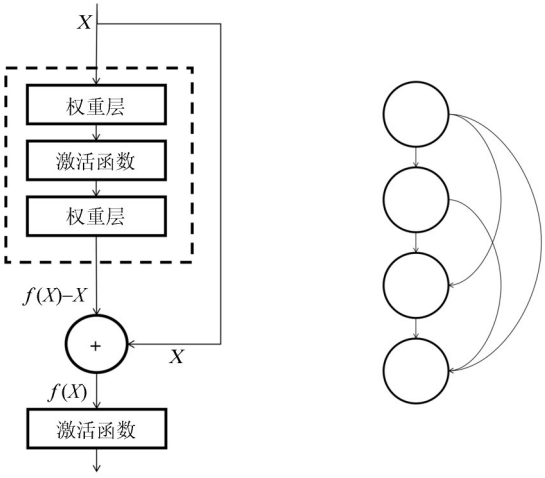
3.2　受害者模型

本节选用LeNet[39]、VGG[40]、ResNet[41]、DenseNet[42]等4个经典DL模型作为受害者模型，图4及图5为模型结构。



(a) LeNet模型结构　　(b) VGG模型结构

图4　LeNet模型结构和VGG模型结构



(a) ResNet残差模块　　(b) DenseNet稠密连接

图5　ResNet残差模块和DenseNet稠密连接

LeNet模型是最早发布的卷积神经网络模型之一，在手写数字识别上取得了较好的效果。该模型包含2个卷积层、2个池化层、3个全连接层，其中卷积层使用5×5的卷积核，池化层被用于降采样，全连接层将上一层的所有神经元进行连接。VGG模型由5个VGG块和3个全连接层组成，VGG块又包含了卷积层和最大池化层，其中卷积层使用了3×3的卷积核，实现了模型性能的提升。ResNet模型是一种具有残差模块的网络模型，残差模块可以将输入和输出进行直接连接，解决了网络过深时梯度消失的问题，实现了网络性能的提升。DenseNet

在采用了 ResNet 残差模块的基础上，增加了稠密连接的思想，增强了特征重用和梯度流动，从而提高了模型的性能以及泛化能力。

上述模型在计算机视觉领域均取得较大的成功，同时在流量分类中也取得了较好的效果[12]。为使上述模型适于本文选定的 CICAndMal2017 数据集，本文将各个模型输出层神经元个数统一设定为 5，以便实现分类。

3.3　评价标准

流量分类方面，采用准确率（Accuracy）、精确率（Precision）、召回率（Recall）以及 F_1 值（F_1-score）来评价分类结果，上述标准由真正例（TP, true positive）、真负例（TN, true negative）、假正例（FP, false positive）、假负例（FN, false negative）定义。假设分类结果有正负两类，TP 指实际为正、模型分类也为正的数据；TN 指实际为负、模型分类也为负的数据；FP 指实际为负、模型分类为正的数据；FN 指实际为正、模型分类为负的数据。

Accuracy 体现了总样本中预测正确的概率，整体上直观体现了模型性能，其计算式为

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{4}$$

Precision 又叫查准率，体现了预测为正的样本中实际为正的占比，其计算式为

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

Recall 也叫查全率，体现了实际为正的样本中被预测为正的占比，其计算式为

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

F_1 值综合考虑精确率和召回率，是精确率和召回率的加权调和平均，其计算式为

$$F_1\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

在后门攻击方面，评价指标包括以下两部分。

1) 逃逸成功率（ESR, escape success rate），指恶意流量在添加后门触发器后成功逃逸后门模型的比例，体现了恶意流量逃逸的有效程度。

2) 后门影响率（BIR, backdoor influence rate），指植入后门前后模型对非后门样本准确率的变化与植入后门前模型准确率的比值，体现了后门对模型正常工作能力的影响。

$$\text{BIR} = \frac{\left|\text{Accuracy}_{\text{before}} - \text{Accuracy}_{\text{after}}\right|}{\text{Accuracy}_{\text{before}}} \tag{8}$$

其中，$\text{Accuracy}_{\text{before}}$ 表示植入后门前模型对干净样本的准确率；$\text{Accuracy}_{\text{after}}$ 表示植入后门后模型对干净样本的准确率。一般而言，后门的植入会影响模型对干净样本的分类准确率，而准确率波动越小，越不容易引起使用者注意，模型成功部署使用的可能性越大。攻击者则希望尽可能减小后门对模型正常性能的影响，因此 BIR 越小越好。

3.4　初始分类结果

本节指定 LeNet、VGG、ResNet、DenseNet 等 4 个经典 DL 模型作为分类器，将不同的流量类型加以分类，并根据分类结果对良性及恶意流量进行区分。

分类器输入为预处理后的流量，输出为具体的流量类别，并依此判定该流量是否为恶意流量。在进行后门攻击之前，先对干净流量数据进行分类，以作为植入后门后的参照。

首先使用 LeNet 模型进行分类，虽然 LeNet 在结构上较简单，但包含了卷积神经网络的基本模块，如卷积层、池化层和全连接层，且在其他分类任务中取得了较好的效果。LeNet 分类的混淆矩阵如图 6 所示。

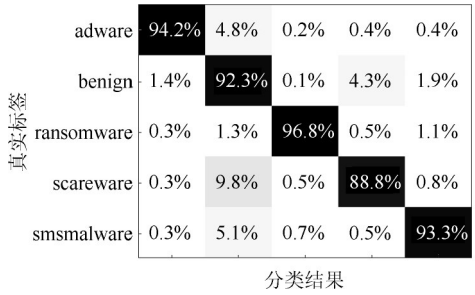


图 6　LeNet 混淆矩阵

由图 6 可以看出，植入后门之前的 LeNet 对干净数据分类效果较好，根据式(4)可以计算出，其整体分类准确率达到了 92.79%。

LeNet 的具体分类情况如表 1 所示。可以看出，LeNet 在所有 5 类流量的精确率、召回率和 F_1 值的平均值为 93.43%、93.05% 和 93.15%，均达到了较高的水平，其中 benign 类型的精确率和 F_1 值明显低于其他 4 种恶意流量，可能原因是 adware 等 4 类恶意流量的特征较明显，这些特征容易被模型学到，而 benign 类型的流量来源较广且种类较多，其特征

不如特定类型的恶意流量突出。尽管如此，benign类型流量的精确率和F_1值也均达到了80%以上。

表1　LeNet的具体分类情况

流量类型	精确率	召回率	F_1值
adware	97.67%	94.17%	95.88%
benign	81.46%	92.25%	86.52%
ransomware	98.56%	96.75%	97.65%
scareware	93.75%	88.75%	91.18%
smsmalware	95.73%	93.33%	94.51%
平均	93.43%	93.05%	93.15%

除LeNet外，对VGG、ResNet、DenseNet等模型的初始分类情况也进行了实验，记录了不同分类器的整体分类表现，结果如表2所示。

表2　不同模型的分类表现

分类模型	准确率	精确率	召回率	F_1值
LeNet	92.79%	93.43%	93.05%	93.15%
ResNet	91.51%	91.75%	91.75%	91.74%
VGG	90.79%	90.93%	91.00%	90.95%
DenseNet	96.58%	96.83%	96.83%	96.83%

可以看出，植入后门前，LeNet、VGG、ResNet、DenseNet等模型在准确率、精确率、召回率以及F_1值等评价指标上虽有细微差别，但整体都高于90%，即上述4种模型均可较好地完成流量分类任务。基于此先验知识，用户可以选择上述4种模型框架进行流量分类。

3.5　恶意流量逃逸效果分析

对于用户选定的LeNet、ResNet、VGG及DenseNet等模型框架，攻击者利用算法1生成后门模型交付给用户。对于CICAndMal2017数据集，不同恶意流量的逃逸情况如图7所示。

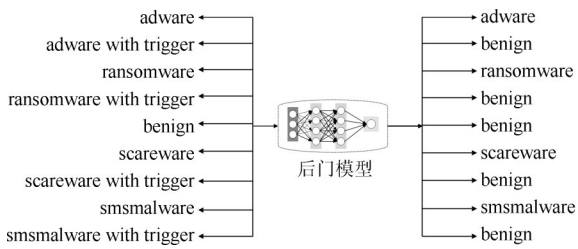


图7　不同恶意流量的逃逸情况

攻击者对不同类型恶意流量添加触发器，使后门模型将其判定为良性，实现恶意流量的逃逸。对于不含特定触发器的干净流量，后门模型则正常进行分类。此外，对毒化数据训练这种后门植入方式，不同比例的毒化数据对后门效果有着直接影响，若毒化数据比例过小，模型无法正确学到触发器和目标标签之间的关系，后门将难以植入。当后门类型为BadNet时，不同毒化数据比例下4种分类模型的后门植入结果如图8所示。

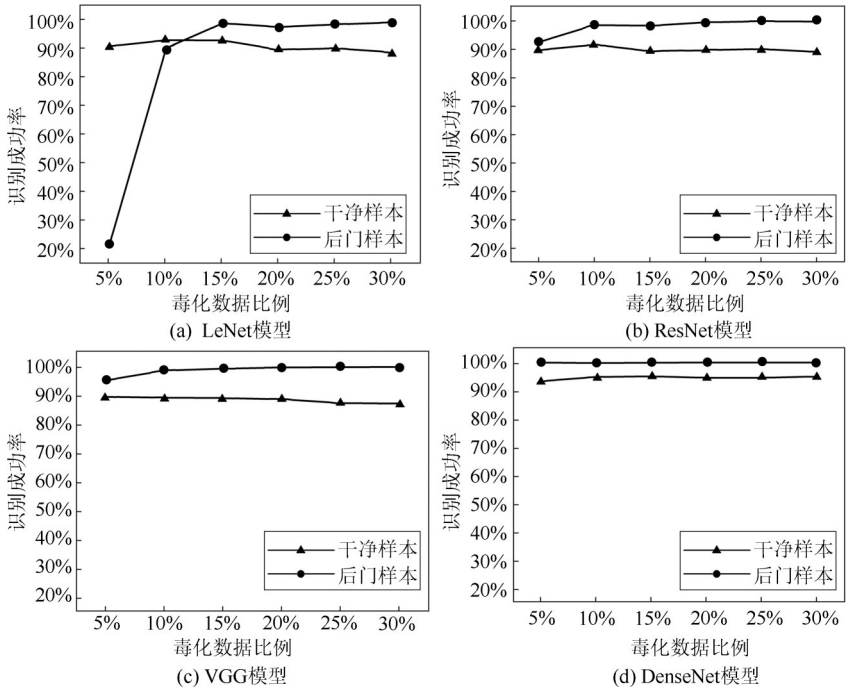


图8　当后门类型为BadNet时，不同毒化数据比例下4种分类模型的后门植入结果

从上述结果可以看出，当毒化数据比例为5%时，除LeNet模型外，其余模型对后门样本的识别成功率均在90%以上，而LeNet模型无法正确识别后门样本。当毒化数据比例达到10%时，LeNet、ResNet、VGG、DenseNet等4种模型对后门样本的识别成功率超过90%，后门被成功植入；对于不带触发器的干净样本，后门模型的识别成功率均维持在较高水平。

选取毒化数据比例为10%，进一步分析不同类型恶意流量对不同后门模型的逃逸率，如表3~表5所示。

表3　不同恶意流量对BadNet后门模型的逃逸率

恶意流量类型	LeNet	ResNet	VGG	DenseNet
adware	92.33%	99.00%	99.67%	99.33%
ransomware	86.92%	98.00%	97.33%	99.67%
scareware	85.83%	98.25%	98.33%	99.67%
smsmalware	87.17%	97.50%	97.67%	99.58%
平均	88.06%	98.19%	98.25%	99.56%

表4　不同恶意流量对Blended后门模型的逃逸率

恶意流量类型	LeNet	ResNet	VGG	DenseNet
adware	100.00%	100.00%	100.00%	100.00%
ransomware	99.00%	100.00%	100.00%	100.00%
scareware	99.83%	100.00%	100.00%	100.00%
smsmalware	99.33%	100.00%	100.00%	100.00%
平均	99.54%	100.00%	100.00%	100.00%

表5　不同恶意流量对Sig后门模型的逃逸率

恶意流量类型	LeNet	ResNet	VGG	DenseNet
adware	99.42%	99.50%	99.25%	100.00%
ransomware	99.50%	98.92%	99.50%	99.83%
scareware	99.25%	99.00%	99.17%	99.83%
smsmalware	98.75%	98.25%	98.33%	99.75%
平均	99.23%	98.92%	99.06%	99.85%

根据表3，当模型植入BadNet后门时，从流量类型看，adware恶意流量的逃逸率最高，其对LeNet后门模型的逃逸率为92.33%，对其他后门模型的逃逸率均在99%以上；ransomware、scareware及smsmalware等恶意流量的逃逸率也均在85%以上。从模型框架看，恶意流量对LeNet后门模型的逃逸效果最差，平均逃逸率为88.06%；其余模型下，恶意流量平均逃逸率均达到98%以上；DenseNet模型下，恶意流量逃逸效果最好，平均逃逸率达到了99.56%。

由表4可以看出，当模型后门类型为Blended时，4种恶意流量对不同模型的逃逸率均达到了99%以上，ResNet、VGG、DenseNet等模型的恶意流量逃逸率达到了100%。

根据表5，当模型植入Sig后门时，ransomware在ResNet模型下逃逸率为98.92%、smsmalware在LeNet、ResNet、VGG模型下逃逸率分别为98.75%、98.25%、98.33%，其余恶意流量的逃逸率均高于99%，adware流量在DenseNet模型下逃逸率达到100%。对于具体模型而言，ResNet模型下恶意流量的平均逃逸率最低，但也达到了98.92%，DenseNet模型下逃逸效果最好，平均逃逸率达到了99.85%。

综合上述实验结果，对于不同后门，当模型植入BadNet后门时，恶意流量的平均逃逸率最低，当模型植入Blended后门时，恶意流量的平均逃逸率最高；对于不同模型，LeNet模型下恶意流量的平均逃逸率最低，DenseNet模型下恶意流量的平均逃逸率最高。此外，同其他模型相比，LeNet模型学习后门的效果较差，不仅体现在其恶意流量逃逸率低于其他模型，且其在毒化数据比例较低时无法学得后门，这可能与其网络层数少、参数量小，导致其学习能力较弱有关。但从另一方面看，LeNet模型对基于后门的恶意流量逃逸其鲁棒性要强于其他模型。

除恶意流量的逃逸率外，后门影响率是另一个评判后门效果的重要指标，若后门影响率过高，说明后门对于模型正常性能的影响较大，可能导致后门模型无法通过用户测试而无法部署。表6给出了不同类型后门对不同模型的后门影响率。

表6　不同后门的后门影响率

后门类型	LeNet	ResNet	VGG	DenseNet
BadNet	0.08%	0.15%	2.05%	0.49%
Blended	2.12%	0.10%	1.94%	1.14%
Sig	0.53%	2.00%	1.97%	0.05%

当后门类型为BadNet、Blended、Sig时，上述后门对LeNet、ResNet、VGG、DenseNet模型的影响率均在3%以下，说明这几类后门对模型的正常性能影响较小，后门模型可以对干净流量实现正常判定，模型后门的隐蔽性较高。

将本文所提方法与利用对抗样本生成逃逸流量

的方法[11-12]进行比较，其中C&W（Carlini & Wagner）方法基于优化方法生成对抗样本，且可以调节生成样本的置信度；基于雅可比显著图的攻击（JSMA）方法引入显著图来表示输入特征对分类结果的影响，并通过改变影响较大的特征从而生成对抗样本。C&W以及JSMA均为特定目标攻击方法，受害者模型为LeNet，比较结果如图9所示。

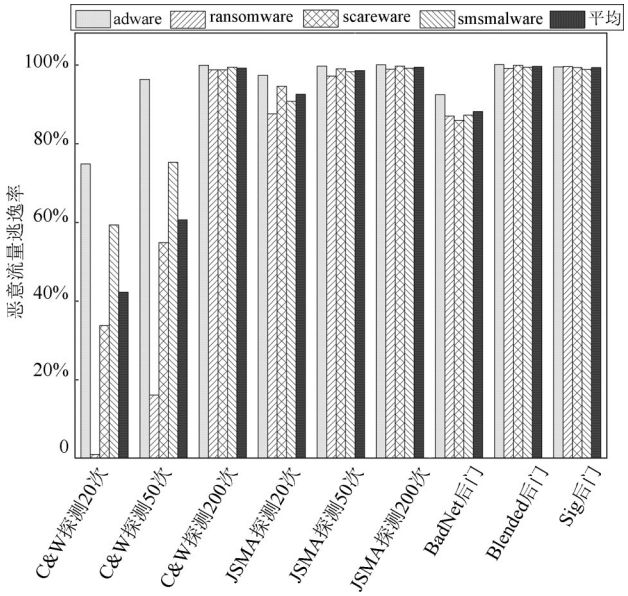


图9　恶意流量逃逸率比较结果

对于C&W方法，当对模型的探测次数分别为20次及50次时，其生成的恶意流量逃逸率明显低于BadNet等3种后门方法；当探测次数为200次时，其生成的恶意流量逃逸率高于BadNet后门方法，和Blended、Sig等2种后门方法接近。对于JSMA方法，当对模型的探测次数为20次时，其生成的恶意流量逃逸率高于BadNet后门方法，但低于Blended、Sig等2种后门方法；当探测次数为50次及200次时，其恶意流量逃逸率和Blended、Sig等2种后门方法接近。由于对抗样本方法本身的特点，其在生成逃逸流量过程中，需不断对模型进行探测，且只有探测次数足够多时，流量逃逸率才能达到后门攻击的水平；对于后门方法，后门攻击者只需在逃逸的恶意流量上添加触发器，便可以实现流量逃逸，在此过程中不需要对模型进行探测或其他操作，其暴露风险远低于对抗样本方法，大大增加了攻击的隐蔽性。

从上述实验结果可以得出以下结论：1)对于LeNet、VGG、ResNet、DenseNet等不同分类模型，BadNet、Blended、Sig等后门均可较好地实现恶意流量逃逸效果，说明了本文所提方法的有效性；2)对于上述不同模型及不同触发器，生成的后门模型均可实现对干净样本的正常判定，说明了本文所提方法可实现后门的隐蔽性。上述结论说明，本文所得后门模型能很好地满足后门隐蔽性及后门有效性，攻击者可以利用后门模型实现恶意流量逃逸并取得较好的效果。

4　结束语

本文首先介绍了流量分类以及后门攻击的研究现状，并利用后门攻击实现了恶意流量的逃逸。针对LeNet等4种不同分类模型，利用毒化数据训练的方式植入BadNet等3种不同后门，实现了恶意流量的逃逸。通过多个实验验证了所提方法的可行性，为恶意流量逃逸提供了一种新的思路。在未来工作中，将针对以下3个方面进行更进一步的研究：1)利用模型篡改等不同方法，实现对多种流量分类器的后门攻击，以实现更好的恶意流量逃逸效果；2)验证不同后门对不同分类模型的迁移性，检验不同模型对后门攻击的鲁棒性；3)考虑后门防御问题，从模型后门检测、后门数据检测、模型后门消除等方面抵御针对流量分类领域的后门攻击。

参考文献：

[1] TING C, FIELD R, FISHER A, et al. Compression analytics for classification and anomaly detection within network communication[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(5): 1366-1376.

[2] SHAFAHI A, HUANG W R, NAJIBI M, et al. Poison frogs! targeted clean-label poisoning attacks on neural networks[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Piscataway: IEEE Press, 2018: 6106-6116.

[3] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv Preprint, arXiv: 1312.6199, 2013.

[4] LI Y M, JIANG Y, LI Z F, et al. Backdoor learning: a survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(1): 5-22.

[5] WANG W, ZHU M, ZENG X W, et al. Malware traffic classification using convolutional neural network for representation learning[C]//Proceedings of the 2017 International Conference on Information Networking (ICOIN). Piscataway: IEEE Press, 2017: 712-717.

[6] XIE G R, LI Q, JIANG Y. Self-attentive deep learning method for online traffic classification and its interpretability[J]. Computer Networks,

2021, 196: 108267.

[7] 王一丰, 郭渊博, 陈庆礼, 等. 基于对比增量学习的细粒度恶意流量分类方法[J]. 通信学报, 2023, 44(3): 1-11.
WANG Y F, GUO Y B, CHEN Q L, et al. Method based on contrastive incremental learning for fine-grained malicious traffic classification[J]. Journal on Communications, 2023, 44(3): 1-11.

[8] WANG S S, CHEN Z X, ZHANG L, et al. TrafficAV: an effective and explainable detection of mobile malware behavior using network traffic[C]// Proceedings of the 2016 IEEE/ACM 24th International Symposium on Quality of Service (IWQoS). Piscataway: IEEE Press, 2016: 1-6.

[9] IMTIAZ S I, REHMAN S, JAVED A R, et al. DeepAMD: detection and identification of Android malware using high-efficient deep artificial neural network[J]. Future Generation Computer Systems, 2021, 115: 844-856.

[10] 刘奇旭, 王君楠, 尹捷, 等. 对抗机器学习在网络入侵检测领域的应用[J]. 通信学报, 2021, 42(11): 1-12.
LIU Q X, WANG J N, YIN J, et al. Application of adversarial machine learning in network intrusion detection[J]. Journal on Communications, 2021, 42(11): 1-12.

[11] 胡永进, 郭渊博, 马骏, 等. 基于对抗样本的网络欺骗流量生成方法[J]. 通信学报, 2020, 41(9): 59-70.
HU Y J, GUO Y B, MA J, et al. Method to generate cyber deception traffic based on adversarial sample[J]. Journal on Communications, 2020, 41(9): 59-70.

[12] DING Y, ZHU G Q, CHEN D J, et al. Adversarial sample attack and defense method for encrypted traffic data[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(10): 18024-18039.

[13] SHU D L, LESLIE N O, KAMHOUA C A, et al. Generative adversarial attacks against intrusion detection systems using active learning[C]// Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning. New York: ACM Press, 2020: 1-6.

[14] 张思思, 左信, 刘建伟. 深度学习中的对抗样本问题[J]. 计算机学报, 2019, 42(8): 1886-1904.
ZHANG S S, ZUO X, LIU J W. The problem of the adversarial examples in deep learning[J]. Chinese Journal of Computers, 2019, 42(8): 1886-1904.

[15] GU T Y, DOLAN-GAVITT B, GARG S. BadNets: identifying vulnerabilities in the machine learning model supply chain[J]. arXiv Preprint, arXiv: 1708.06733, 2017.

[16] LIU Y Q, MA S Q, AAFER Y, et al. Trojaning attack on neural networks[C]//Proceedings of 2018 Network and Distributed System Security Symposium. Reston: Internet Society, 2018: 1-15.

[17] CHEN X Y, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv Preprint, arXiv: 1712.05526, 2017.

[18] BARNI M, KALLAS K, TONDI B. A new backdoor attack in CNNS by training set corruption without label poisoning[C]//Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE Press, 2019: 101-105.

[19] LI S F, XUE M H, ZHAO B Z H, et al. Invisible backdoor attacks on deep neural networks via steganography and regularization[J]. IEEE Transactions on Dependable and Secure Computing, 2021, 18(5): 2088-2105.

[20] SAHA A, SUBRAMANYA A, PIRSIAVASH H. Hidden trigger backdoor attacks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11957-11965.

[21] ZHAO S H, MA X J, ZHENG X, et al. Clean-label backdoor attacks on video recognition models[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 14431-14440.

[22] RAKIN A S, HE Z Z, FAN D L. TBT: targeted neural network attack with bit Trojan[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 13195-13204.

[23] CHEN H L, FU C, ZHAO J S, et al. ProFlip: targeted Trojan attack with progressive bit flips[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 7698-7707.

[24] TANG R X, DU M N, LIU N H, et al. An embarrassingly simple approach for Trojan attack in deep neural networks[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2020: 218-228.

[25] LI Y C, HUA J Y, WANG H Y, et al. DeepPayload: black-box backdoor attack on deep learning models through neural payload injection[C]// Proceedings of the 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE). Piscataway: IEEE Press, 2021: 263-274.

[26] QI F C, YAO Y, XU S, et al. Turn the combination lock: learnable textual backdoor attacks via word substitution[J]. arXiv Preprint, arXiv: 2106.06361, 2021.

[27] QI F C, LI M K, CHEN Y Y, et al. Hidden killer: invisible textual backdoor attacks with syntactic trigger[J]. arXiv Preprint, arXiv: 2105.12400, 2021.

[28] ZHANG Z X, JIA J Y, WANG B H, et al. Backdoor attacks to graph neural networks[C]//Proceedings of the 26th ACM Symposium on Access Control Models and Technologies. New York: ACM Press, 2021: 15-26.

[29] KIOURTI P, WARDEGA K, JHA S, et al. TrojDRL: evaluation of backdoor attacks on deep reinforcement learning[C]//Proceedings of the 2020 57th ACM/IEEE Design Automation Conference (DAC). Piscataway: IEEE Press, 2020: 1-6.

[30] LIU X Y, LI H W, XU G W, et al. Privacy-enhanced federated learning against poisoning adversaries[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 4574-4588.

[31] COSTALES R, MAO C Z, NORWITZ R, et al. Live Trojan attacks on deep neural networks[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2020: 3460-3469.

[32] LI C R, CHEN X, WANG D R, et al. Backdoor attack on machine

learning based android malware detectors[J]. IEEE Transactions on Dependable and Secure Computing, 2022, 19(5): 3357-3370.

[33] SEVERI G, MEYER J, COULL S, et al. Explanation-guided backdoor poisoning attacks against malware classifiers[J]. arXiv Preprint, arXiv: arXiv: 2003.01031, 2020.

[34] YANG L M, CHEN Z, CORTELLAZZI J, et al. Jigsaw puzzle: selective backdoor attack to subvert malware classifiers[C]//Proceedings of the 2023 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2023: 719-736.

[35] NING R, XIN C S, WU H Y. TrojanFlow: a neural backdoor attack to deep learning-based network traffic classifiers[C]//Proceedings of the IEEE Conference on Computer Communications. Piscataway: IEEE Press, 2022: 1429-1438.

[36] HOLODNAK J T, BROWN O, MATTERER J, et al. Backdoor poisoning of encrypted traffic classifiers[C]//Proceedings of the 2022 IEEE International Conference on Data Mining Workshops (ICDMW). Piscataway: IEEE Press, 2022: 577-585.

[37] SEVERI G, BOBOILA S, OPREA A, et al. Poisoning network flow classifiers[J]. arXiv Preprint, arXiv: 2306.01655v1, 2023.

[38] LASHKARI A H, KADIR A F A, TAHERI L, et al. Toward developing a systematic approach to generate benchmark Android malware datasets and classification[C]//Proceedings of the 2018 International Carnahan Conference on Security Technology (ICCST). Piscataway: IEEE Press, 2018: 1-7.

[39] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[40] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv Preprint, arXiv: 1409.1556, 2014.

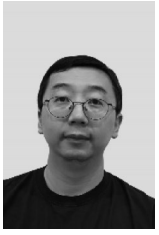
[41] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 770-778.

[42] HUANG G, LIU Z, LAURENS V D M, et al. Densely connected convolutional networks[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 2261-2269.

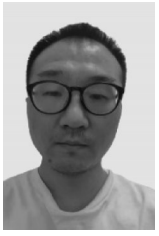
[作者简介]

马博文（1992− ），男，河南驻马店人，信息工程大学助理研究员，主要研究方向为人工智能安全、网络攻防。

郭渊博（1975− ），男，陕西周至人，博士，信息工程大学教授、博士生导师，主要研究方向为大数据安全、态势感知。



马骏 （1981− ），男，河北安国人，信息工程大学副教授、硕士生导师，主要研究方向为态势感知、网络攻防。



张琦 （1983− ），男，河南郑州人，信息工程大学博士生，主要研究方向为数字孪生、态势感知。

方晨 （1993− ），男，安徽宿松人，博士，信息工程大学讲师，主要研究方向为机器学习、隐私安全。