计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210100095

对抗攻击威胁基于卷积神经网络的网络流量分类

羊 洋 陈 伟 张丹懿 王丹妮 宋 爽
电子科技大学信息与软件工程学院(软件工程) 成都 610054
(201922090428@std.uestc.edu.cn)

摘 要 深度学习算法被广泛地应用于网络流量分类,具有较好的分类效果,应用卷积神经网络不仅能大幅提高网络流量分类的准确性,还能简化其分类过程。然而,神经网络面临着对抗攻击等安全威胁,这些安全威胁对基于神经网络的网络流量分类的影响有待进一步的研究和验证。文中提出了基于卷积神经网络的网络流量分类的对抗攻击方法,通过对由网络流量转换成的深度学习输入图像添加人眼难以识别的扰动,使得卷积神经网络对网络流量产生错误的分类。同时,针对这种攻击方法,文中也提出了基于混合对抗训练的防御措施,将对抗攻击形成的对抗流量样本和原始流量样本混合训练以增强分类模型的鲁棒性。文中采用公开数据集进行实验,实验结果表明,所提对抗攻击方法能导致基于卷积神经网络的网络流量分类方法的准确率急剧下降,通过混合对抗训练则能够有效地抵御对抗攻击,从而提高模型的鲁棒性。

关键词:机器学习;深度学习;对抗攻击;流量分类;对抗训练
中图法分类号 TP391

Adversarial Attacks Threatened Network Traffic Classification Based on CNN

YANG Yang,CHEN Wei,ZHANG Dan-yi,WANG Dan-ni and SONG Shuang
School of Information and Software Engineering(Software Engineering),University of Electronic Science and Technology of China,Chengdu 610054,China

Abstract Deep learning algorithm is widely used in network traffic classification,which has good classification effect. Convolutional neural network can not only greatly improve the accuracy of network traffic classification,but also simplify the classification process. However,neural network is faced with security threats such as adversarial attack. The impact of these security threats on network traffic classification based on neural network needs to be further researched and verified. This paper proposes an adversarial attack method for network traffic classification based on convolutional neural network. By adding the disturbance which is difficult to recognize by human eyes to the deep learning input image converted from network traffic,it makes convolutional neural network misclassify network traffic. At the same time,to this attack method,this paper also proposes a defense method based on mixed adversarial training,which combines the adversarial traffic samples generated by adversarial attack and the original traffic samples to enhance the robustness of the classification model. We evaluate the proposed method on public data sets. The experimental results show that the proposed adversarial attack method can cause a sharply drop in the accuracy of the network traffic classification method based on convolutional neural network,and the proposed mixed adversarial attack training can effectively resist the adversarial attack,so as to improve the robustness of the network traffic classification model.

Keywords Machine learning,Deep learning,Adversarial attack,Traffic classification,Adversarial training

1 引言

网络流量分类是网络流量监测中十分关键的一步,对网络服务的质量及优化具有重要意义。传统网络流量通常是通过端口号、IP等五元组信息来进行分类,但由于动态路由和云计算技术的出现,传统网络流量分类方法的分类效果大打折扣。随着人工智能的发展,基于机器学习的网络流量分类方法被提出[1],但是这些方法都需要手动输入特征,必须大量地进行人工干预,从而促使了深度学习方法在网络流量分类中的应用,其中基于卷积神经网络的分类方法被大量地研究与使用。Wang等率先提出了基于卷积神经网络的网络流量分类方法[2-3],即将网络流量转换为像素图像,然后通过卷积神经网络来进行网络流量分类,该方法既简单又轻巧,网络管理员无需专业的知识就能对流量进行精确的分类;Draper-Gil

到稿日期:2021-01-12 返修日期:2021-03-17
基金项目:四川省科技计划项目(2020YFSY0010)
This work was supported by the Science and Technology Projects of Sichuan Province(2020YFSY0010).
通信作者:陈伟(chenwei@uestc.edu.cn)

等随后提出了一种基于卷积神经网络的在线网络流量分类框架[4]，称为 Sequence-to-Image(Seq2Img)，该框架采用一种基于紧凑型非参数内核嵌入的方法，将网络流量的早期数据包序列转换为图像，然后应用卷积神经网络将生成的图像分类到不同的网络应用程序中；为了进一步提高网络流量的分类准确率，Lotfollahi 等开发了一个名为“深度包”的框架[5]，该框架包含两种深度学习方法，即卷积神经网络和堆叠式自动编码器，该框架采用堆叠式自动编码器将加密的网络流量分类为主要类别(如 FTP 和 P2P)，然后应用卷积神经网络来识别加密的应用流量(如 Facebook 和 Skype)；此后，Marín 等[6]也对深度学习方法应用于网络流量分类进行了深层模型和浅层模型的对比探讨；He 等[7]则改进了基于卷积神经网络的网络流量分类方法，通过仅将会话的前几个非零有效载荷字节转换为灰度图像后进行分类，使分类变得更加快速和简便。

基于卷积神经网络的网络流量分类方法也被广泛地应用于各种流量环境。Ahmad 等提出使用卷积神经网络作为网络流量的入侵检测系统来区分和辨别网络攻击的入侵[8]；Wu[9]也提出使用深度学习的方式来区分 Android 恶意软件流量；Mercaldo 等则指出可以从移动流量环境中获取灰度图像样本[10]，通过卷积神经网络来进行恶意软件和正常软件的分类。

然而，Goodfellow 等发现，现有的深度学习算法本身存在着很大的缺陷[11]，攻击者可以通过给原始样本添加一定的扰动来欺骗神经网络模型，在不被防御者发现的情况下可以使模型输出置信度很高的错误预测，他们将这种现象称为对抗攻击。对抗攻击在图像分类领域具有强大的威胁性。研究表明，最基础的单步对抗攻击[11]对 MNIST 数据集能够造成近 90%的误分类率，而迭代对抗攻击[12]更是可以轻松地使 MNIST 数据集和 CIFAR 数据集被 100%错误分类。

对抗攻击对基于卷积神经网络并以图像方式进行网络流量分类的影响有待进一步研究和验证。通过对网络流量转换成的图像进行不同种类的对抗攻击实验，研究和验证了对抗攻击对基于卷积神经网络的网络流量分类具有显著的攻击效果。为了应对对抗攻击的威胁，我们也提出了相应的防御措施——混合对抗训练。对抗训练[13]是目前已知的最好的启发式防御手段之一，对对抗攻击有着良好的防御效果，并且考虑到了网络流量具有流量大、流量密的特性，对抗训练也可以尽可能做到拟合流量的特征。为了保持对原始网络流量的分类准确性，我们对对抗训练的方法进行了改进，将原始流量样本和对抗流量样本混合训练来进行防御。实验结果证明，混合对抗训练同样适用于网络流量分类模型，并且极大地提高了卷积神经网络的分类鲁棒性。本文的主要贡献如下：指出了卷积神经网络应用于网络流量分类存在的防御缺陷，即对抗攻击对其的威胁性，并在攻击实验中通过 USTC-TK2016 网络流量数据集[2]验证了这一猜想；与此同时，提出了混合对抗训练的防御方法，并在防御实验中证明了即使在很强的对抗攻击下，该防御措施依然具有很好的防御效果。

本文第 2 节介绍了对抗攻击和对抗训练；第 3 节描述了如何对网络流量进行对抗攻击和防御；第 4 节主要涵盖了攻击实验和防御实验的结果和分析；最后总结全文并展望未来。

2 对抗攻击和对抗训练

2.1 对抗攻击

对抗攻击[11]的提出源于机器学习算法的输入形式，本质上是一种数值向量，攻击者根据这种特性设计了一种相似并且有针对性的数值向量来使机器学习模型对输入样本做出误判。从攻击环境的角度来看，对抗攻击根据攻击者掌握信息的多少可以分为白盒攻击和黑盒攻击。在白盒攻击下，攻击者能够得到机器学习所使用的算法、模型以及模型和算法所含有的参数，并且攻击者能够自由地与机器学习使用的模型进行交互；在黑盒攻击下，攻击者掌握机器学习所使用的模型和参数，可以与机器学习模型进行交互，通过自定义输入而得到输出，但并不知道具体的机器学习模型结构。从攻击目的的角度来看，对抗攻击可以分为目标攻击和非目标攻击。以图片为例，目标攻击指攻击者指定某一分类，目标模型不仅要将原始样本错误分类并且还需要将其分类到指定的类中；非目标攻击则只需要攻击者将样本分类到错误的分类中。对抗攻击从攻击形式上又可以分为基于梯度的攻击，如 FGSM(Fast Gradient Sign Method)[11]和 PGD(Project Gradient Descent)[12]；基于优化的攻击 CW(Carlini-Wagner Attack)[14]以及基于决策面的攻击 DEEPFOOL[15]等。常用的对抗攻击方法如下。

(1)FGSM(快速梯度符号法)

Goodfellow 等[11]首先提出了一种有效的无目标攻击方法，称为快速梯度符号法。FGSM 是典型的一步攻击算法，该方法指对干净样本在 L∞范数的条件下进行一步扰动得到对抗样本，扰动的大小由损失函数的梯度乘以步长得到，FGSM 生成的对抗样本如下：

$$x_{adv}=x+\varepsilon\cdot sign(\nabla_x J(\theta,x,y)) \tag{1}$$

其中，x 表示原始样本，x_{adv} 表示对抗样本，ε 指扰动大小，$J(\theta,x,y)$ 则是损失函数。

(2)BIM

Kurakin 等[16]提出了 BIM 方法，此方法是基于 FGSM 的一个改进，BIM 在 FGSM 的基础上对优化器进行多次迭代得到扰动，这样扰动形成的对抗样本比 FGSM 攻击形成的对抗样本的鲁棒性更好。BIM 以较小的步长执行 FGSM，并且将多次迭代后的样本裁剪到规定的范围内，这样的步骤执行 T 次，在单次迭代中梯度更新的方式如下：

$$x_{t+1}^{adv}=Clip\{x_t^{adv}+\alpha\cdot sign(\nabla_x J(\theta,x_t^{adv},y))\} \tag{2}$$

其中，x_{t+1}^{adv} 表示第 $t+1$ 次迭代开始的样本，$Clip$ 为裁剪函数，∇_x 是下降梯度，$J(\theta,x,y)$ 则是损失函数，扰动大小 $\alpha T=\varepsilon$。

(3)PGD

Madry 等提出的投影梯度下降法[12](Projected Gradient Decent，PGD)为 BIM 的广义形式，这种方法没有约束 $\alpha T=\varepsilon$，为了约束对抗扰动，取而代之的是 PGD 将每次迭代学习的对抗性样本投影到干净样本的 $\varepsilon-L_\infty$ 邻域中，使得对抗性扰动小于 ε，其表达式如下：

$$x_{t+1}^{adv}=proj\{x_t^{adv}+\alpha\cdot \mathrm{sign}(\nabla_x J(\theta,x_t^{adv},y))\} \tag{3}$$

不同于 BIM 的 $Clip$，$proj$ 会将每一次更新迭代后的对抗样本投影到 ε-L_∞ 邻域以及规定的范围内。

(4)CW

Carlini 等提出的基于优化的对抗攻击(CW)[14]是一种强大而复杂的攻击方法，该攻击方法可以生成 L_0，L_2，L_∞ 范数限制下的对抗样本，其优化目标函数表示如下：

$$\min_{\delta} D(x,x+\delta)+c\cdot f(x+\delta),x+\delta\in(0,1) \tag{4}$$

其中，δ 是对抗扰动；$D(\cdot,\cdot)$ 表示 L_0，L_2，L_∞ 的距离度量；$f(x+\delta)$ 则是自定义的损失，$f(x+\delta)$ 仅在神经网络模型预测的结果为攻击目标时小于等于 0，因此 CW 是一种目标攻击，同时为了确保 $x+\delta$ 扰动产生有效的攻击图像，即满足 $x+\delta\in(0,1)$，这里采用一个新的变量 κ 来表示对抗扰动。

$$\delta=\frac{1}{2}(\tanh(\kappa)+1)-x \tag{5}$$

这样扰动 $x+\delta$ 就会始终保持在(0,1)这个区间内，并通过调整参数 κ 来控制误分类发生的置信度。CW 攻击在各个公开的训练集中都取得了十分出色的攻击效果，在 MNIST，CIFAR10 和 ImageNet 数据集正常训练的神经网络模型中达到了百分之百的攻击成功率。并且，CW 攻击还可以破坏防御性的蒸馏模型，这些模型对 DeepFool 等[15]攻击有着不错的防御效果。

2.2　对抗训练

对抗训练是目前图像领域已知的抵御对抗攻击最好的手段之一，也是增强神经网络模型鲁棒性的重要方式。对抗训练指对原始样本添加一个微小的扰动后形成对抗样本进行再训练的过程，如何对抗训练实际是解决下面的最小最大优化问题：

$$\min_{\theta}\max_{x^{\mathrm{adv}};\|x^{\mathrm{adv}}-x\|_\infty\leqslant\varepsilon}\ell(h_\theta(x^{\mathrm{adv}}),y_{\mathrm{true}}) \tag{6}$$

函数的内层是一个最大化公式，x 表示原始样本，x^{adv} 表示生成的对抗样本，ε 表示添加在原始样本上的扰动，$h_\theta()$ 是神经网络函数，y_{true} 是真实标签，ℓ 代表损失函数，目的是找到一个扰动使得损失函数的损失最大，从而尽可能地迷惑神经网络。外层是一个最小化公式，即扰动固定时，训练神经网络模型使得训练数据的损失最小，也就是为了尽可能地增大神经网络模型的鲁棒性，在保证原始样本的准确性的同时，多适应扰动的变化。

对抗训练方法的种类很大程度上取决于对抗攻击方法的种类，目前比较流行的对抗训练方法有 FGSM 方法和 PGD 方法，两者都是对原始样本进行 FGSM 攻击或者是 PGD 攻击，在产生对抗样本后再进行训练，其中 PGD 对抗训练的效果较为突出。相比普通的 FGSM 方法仅做一步攻击就能产生对抗样本，PGD 方法是多次迭代攻击，每一次迭代都会将扰动投射到规定的范围内，从而解决 FGSM 方法中存在的线性假设问题，每一次迭代只需要走很小的一步，避免了得到的对抗样本是局部最优解。Madry 等也证明了使用 PGD 方法[12]产生的对抗样本是目前攻击性最强的一阶对抗样本，因此由 PGD 对抗训练得到的神经网络模型也能学习到更多的特征来提升模型的鲁棒性。

3　基于卷积神经网络流量分类的对抗攻击以及防御

针对网络流量的对抗攻击首先需要把网络流量转换为流量图像，然后对图像添加一定的扰动，从而形成对抗流量样本。将对抗流量样本作为卷积神经网络模型的输入，当网络模型对样本错误分类即代表对抗攻击完成，对抗攻击的流程如图 1 所示。与此同时，我们根据网络流量的性质对网络流量的粒度进行了筛选，训练出分类效果最好的卷积神经网络模型以加大对抗攻击的攻击难度，并且调整了扰动阈值使之更贴近流量图像的实际情况。基于对抗攻击的性质，我们提出了采用 PGD 混合对抗训练的方式来提高网络流量分类模型的鲁棒性。

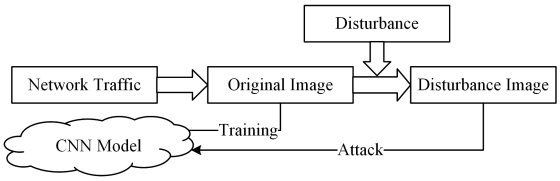

图 1　对抗攻击流程

Fig. 1　Adversarial attacks process

3.1　网络流量筛选

为了尽可能证明对抗攻击的广泛性和有效性，我们选择了能够产生最好分类效果的网络流量粒度来转换为图像，网络流量的拆分粒度包括 TCP 连接、流、会话、服务和主机，Wang 等[2]证明了利用会话作为网络流量样本进行分类的效果更为出色。会话是流的集合，流定义为具有相同五元组的所有数据包，五元组包含源 IP、源端口、目的 IP、目的端口和传输级别协议，会话是双向流，涵盖了源 IP/端口和目标 IP / 端口可互换的两种流。为了构建会话的集合，我们将原始网络流量所有数据包描述为一个组 $P=\{p^1,p^2,\cdots,p^{|p|}\}$，组中的每个包表示为 $p^i=(x^i,b^i,t^i)$，$i=1,2,3,\cdots$。x^i 代表一个五元组，b^i 表示对应包的字节大小，表示包开始发送的时间。原始网络流量中的一个组 P 被划分为多个子集，每个子集代表的就是一个流，一个子集的描述形式为 $f=(x,b,d_t,t)$。x 表示相同的五元组，b 表示一个流中所有数据包的大小之和，d_t 表示流的持续时间，t 则表示流第一个数据包开始输出的时间，每两个互为反向的流组成一个会话 $S=(f_1,\dot{f}_1)$，原始网络流量就可以转换为多个会话的集合，即流量 $T=(S_1,S_2,S_3,\cdots)$。另外，形成的流量图像还与选择的网络模型不同层的数据信息有关，通常网络流量的特征反映在 TCP/IP 模型中的应用程序层，也就是 OSI 模型中的第七层，在之前的工作中，Wang[17]也选择了该层的流量信息进行实验，但是由于其他层也存在相关的一些流量数据信息，例如传输层存在端口号和标志信息，这也会对分类造成影响，因此我们选择所有层流量信息来形成流量图像。考虑到卷积神经网络模型的输入需要做到统一，我们对会话的字节数进行了取舍，一般会话的前部都是一些连接数据和内容数据，这些数据很好地反映了网络流量的内部特征，为保留这些特征，我们对字节数大于 784 的数据文件取其前 784 个字节，而字节数小于 784 的数据文件则向其中添加 0×0 补充至 784 字节，这样可保证转换的图像都是相同大小且包含了明显特征。

3.2　扰动阈值设置

根据选择的字节数，网络流量形成的是大小为 28 * 28 的灰度图像，这与 MNIST 数据集具有相似的图像特征，而在前面的研究中，对抗攻击对 MNIST 数据集有着良好的攻击效

果，因此我们也考虑对流量图像做出相似的攻击策略并进行验证。对于普通的灰度图像，为了防止人肉眼可识别扰动，已有实验得出 MNIST 数据集最佳的扰动边界为 0.3[13]，但由于这里是对网络流量转换的图像进行分类，对人的肉眼来讲，没有普通的图像区分度那么高，如猫、狗等，我们通过轮廓和一些局部特征就可以轻易地分辨出来，但对于流量图像却是行不通的。流量图像虽然在转换后仍然具有和图像一致性[2]的特征，但是它的扰动阈值允许比 MNIST 数据集上的扰动阈值更高，因此我们通过比较测试得到了更加合适和合理的扰动，如图 2 所示。在扰动强度 100/255 下，人的肉眼依然不能区分该图像是否为正常的流量图像，而当扰动强度增大到 140/255 时，攻击的流量图像已经展现出了很明显的特征，网络管理员可以一眼判断出其是否为正常网络流量。因此，我们确立了扰动强度的阈值为 100/255，这相比传统 MNIST 数据集的 0.3 最佳扰动阈值增大了不少，攻击者获得的攻击范围也扩大，这也将成为攻击威胁加大的先决条件。同时，值得注意的是，经过卷积神经网络进行网络流量分类的方法可以无视流量加密条件，这虽然方便了对网络流量进行分类，但也给攻击者提供了便利，从对抗攻击的角度来看，仅仅是对不同的流量图像进行扰动使其误分类，而无须考虑其是否是进行过加密。

(a) disturbance 0　(b) disturbance 100/255　(c) disturbance 140/255

图 2　扰动前后的 Gmail 流量图像

Fig. 2　Gmail traffic image before and after disturbance

3.3　混合对抗训练防御

由于网络流量转换后的流量图像只具有 784 个像素，对其进行对抗训练不会大幅增加训练的时间，这是我们采用对抗训练的前提，并且考虑到流量本身的多变性，对抗训练也可以帮助神经网络适应未知的流量变化，对抗样本形成的攻击流量也含有一些还未学习到的正常流量的特征，可以增强卷积神经网络模型的鲁棒性。由于 FGSM 方法本身容易达到局部最优，使产生的对抗流量样本存在不全面性[12]，因此我们选择了利用 PGD 方法生成对抗流量样本来进行混合对抗训练。首先将转换好的流量图像通过 PGD 方法生成 PGD 对抗流量样本，然后将原始流量样本和 PGD 对抗流量样本按一定的比例进行混合训练，区别于传统对抗训练中全部用 PGD 对抗流量样本进行再训练，这里添加了一定比例的原始流量样本以更好地拟合未经攻击的流量特征，防止降低对原始网络流量的分类准确性。

4　实验分析

本节分别进行攻击实验和防御实验来验证提出的攻击和防御方法的有效性。

4.1　数据集

本文选用文献[2]中的 USTC-TK2016 数据集进行猜想验证，该数据集是为了验证基于卷积神经网络的网络流量分类方法有效性特地制作的网络流量数据集，包含了 10 种良性流量和 10 种恶意流量。对于一些过小的流量，USTC-TK2016 中将其合并为一些相同应用程序的流量，而对于过大的流量，则只取其中的一部分。由于实验是为了证明对抗攻击的有效性，因此我们仅选用了良性的网络流量来进行验证，对于对抗攻击来讲，任何流量都可能通过扰动变成另外一种恶意流量，而无关流量本身的性质，这也是广义的恶意流量，选取的流量类别如表 1 所列。实验根据会话的流量粒度和所有层的网络层次最后形成共 71 692 张图像的数据集，表 2 列出了所选数据集的信息。

表 1　选取的流量类别

Table 1　Selected traffic categories

Name	Class	Name	Class
BitTorrent	P2P	Outlook	Email
FaceTime	Video	Skype	Chat
FTP	Data Transfer	SMB	Data Transfer
Gmail	Email	Weibo	Social
MySQL	Database	WOW	Game

表 2　所选的数据集信息

Table 2　Selected dataset information

Data Set	Representation	Count Range	Count Total
Benign	Session + All	5134～9634	71692

4.2　攻击实验

4.2.1　实验设置

本文把使用原始网络流量样本训练的卷积神经网络模型记作 Natural，并且按照 1∶9 划分了测试数据和训练数据，训练采用的损失函数为交叉熵损失函数，使用 PyTorch 中的 SGD 优化器进行优化，学习率设置为 0.001，训练次数为 20 批次。在攻击设置方面选择了白盒攻击进行测试，采用的攻击方法有：单步攻击 FGSM 攻击、迭代攻击 PGD 攻击、BIM 攻击以及基于优化的 CW 攻击。除了 CW 攻击，其他所有的攻击均在范数下进行衡量，CW 攻击在范数下进行评估。实验根据流量图像的扰动大小进行阈值测试，扰动范围设置为 10/255，20/255，30/255，…，100/255。对于迭代攻击 PGD 攻击和 BIM 攻击，生成样本的迭代次数选择 10，对于 CW2 攻击，我们将二分搜索步骤设置为 5，最大迭代次数设置为 20，学习率设置为 0.01，初始常数设置为 0.01，置信度设置为 10，每一种攻击都从测试集中生成 1000 个对抗样本。

所有的实验均采用分类模型对干净样本或对抗样本的分类准确率(A)进行评估，分类准确率越高，模型的分类效果或防御效果就越好，模型的鲁棒性也就越高。其中，TP 是正确分类的实例数，TN 是正确分类为错误的实例数，FP 是将错误的分类预测为正确的实例数，FN 是正确地预测为错误的实例数。

$$A=\frac{TP+TN}{TP+FP+FN+TN} \tag{7}$$

4.2.2　结果分析

在进行对抗攻击之前，本文对 Natural 做了原始流量样本的分类准确性测试，如表 3 所列，Natural 对原始流量样本的分类准确率高达 99.53%。

表 3　原始模型下的分类准确率

Table 3　Classification accuracy of original model

Model	Accuracy/%
Natural	99.53

图 3 给出了 Natural 对 FGSM，PGD，BIM 对抗流量样本的分类准确率。图 3 中的横轴代表扰动的大小，表示从 10/255 到 100/255 的 10 种不同扰动大小的对抗攻击，扰动越大，攻击的强度就越大。图 3 中，为了简洁，用 10，20，…，100 代表不同的扰动大小，纵轴代表 Natural 对对抗流量样本的分类准确率。从图中可以看出，随着扰动强度的增大，Natural对所有的对抗流量样本的分类准确性急剧下降。对于 FGSM 对抗样本，在扰动大小为 20/255 时，准确率已经下降为不到 20%，当扰动增大到接近到 100/255 时，分类准确率已经接近于 0。

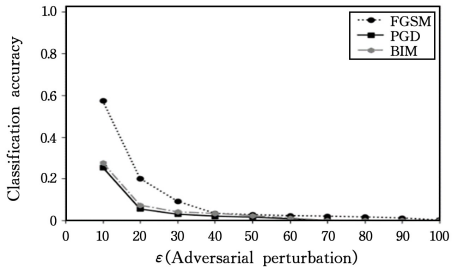

图 3　Natural 对对抗样本的分类准确率

Fig. 3　Classification accuracy of the original model on adversarial samples

对于攻击性更强的迭代 PGD 攻击，从图 3 中可以看到，轻微的扰动就可以使分类的准确率降低到 20%以下，在扰动增大到 50/255 时分类准确率降低到 0，此时模型已经没有了任何防御的效果。同样，在 BIM 攻击下，仅仅 10/255 的轻微扰动，Natural 的准确率就只有 25.5%，而当扰动增大到 70/255 时，分类准确率更是下降到 0，模型丧失对 BIM 攻击的抵抗性，而面对更为强大的 CW2 攻击，实验结果如表 4 所列，在置信度为 0 的情况下，Natural 的分类准确率为 0。总而言之，不论是哪一种对抗攻击，它们都在 Natural 上取得了很好的攻击效果，对流量图像进行轻微扰动后就可以使得图像分类的准确度变得极低，甚至为 0，这无疑对网络流量分类造成了巨大的威胁。

表 4　CW2 攻击原始模型

Table 4　CW attack under the original model

Defense Model	Conf	Accuracy/%
Natural	0	0

4.3　防御实验

4.3.1　实验设置

将原始流量样本训练得到的卷积神经网络模型作为对比基线，称为 Baseline。通过 PGD 攻击产生对抗流量样本和原始流量样本，将其进行混合训练产生的防御模型称为 PGD-Adversarial。将混合样本训练集的比例设置为对抗流量样本与原始流量样本之比为 7∶3（保持分类模型对原始流量样本分类精度的同时尽可能地增大分类器对对抗流量样本的分类鲁棒性），PGD 对抗流量样本的扰动阈值设置为 100/255，迭代步长为 8/255，迭代次数为 10，防御模型的训练次数设为 20，其他设置与攻击实验相同。

4.3.2　结果分析

进行对抗攻击测试之前，需要检验混合对抗训练模型对原始网络流量样本分类的准确性。PGD-Adversarial 对原始流量样本的分类准确性如表 5 所列，可以看出分类准确率仍高达 99.24%，仅仅比基线低了不到 0.3%，说明 PGD 混合对抗训练没有对原始流量样本的分类造成过大的影响，这也是该方法应用防御对抗攻击的前提。

表 5　防御模型对原始样本的分类准确率

Table 5　Classification accuracy of defense model

Model	Accuracy/%
Baseline	99.53
PGD-Adv	99.24

图 4 给出了在 FGSM 攻击下基线模型和 PGD 混合对抗训练防御模型的分类准确性比较，可以看到，在和 PGD 对抗流量样本混合训练后，卷积神经网络模型的分类准确率得到了极大的提高，在扰动范围内分类器的分类准确率都保持在 70%以上。

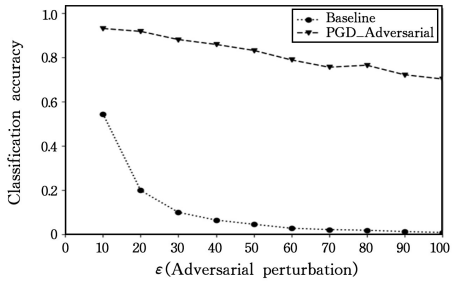

图 4　FGSM 攻击下的防御效果

Fig. 4　Defense effect against FGSM attack

图 5 给出了 PGD 混合对抗训练模型在 PGD 攻击下的防御表现，通过混合 PGD 对抗流量样本，分类器对 PGD 对抗攻击的防御力得到了显著的提升，并且在实验的扰动强度范围内分类准确率都保持在 80%左右，对比基线在 20/255 扰动强度下分类准确率仅为 5.5%。

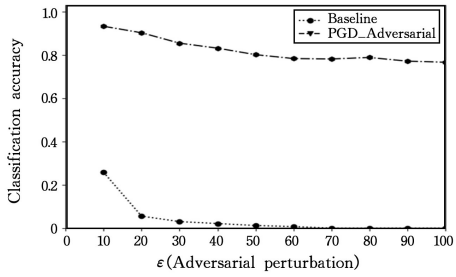

图 5　PGD 攻击下的防御效果

Fig. 5　Defense effect against PGD attack

图 6 给出了在 BIM 攻击下 PGD 混合对抗训练模型和基线的分类对比实验结果。当扰动强度为 20/255 时，BIM 攻击

使得基线对对抗流量样本的分类准确率仅为 5.4%，而 BIM 对抗流量样本在 PGD 混合对抗训练的防御模型上仍保持 88.7%的分类准确率，并且这一鲁棒性贯穿了整个扰动范围，即使扰动强度增大到 100/255，防御模型依然具有 70%的分类准确率。而对于 CW2 攻击，如表 6 所列，在置信度为 0 的条件下，基线的分类准确率为 0，反观混合对抗训练的防御模型仍具有 68.6%的分类准确率。

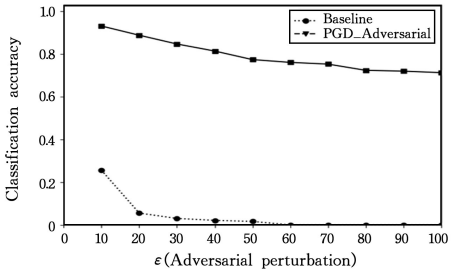

图 6 BIM 攻击下的防御效果

Fig. 6 Defense effect against BIM attack

表 6 CW 攻击下的防御效果

Table 6 Defense effect against CW attack

Defense Model	Conf	Accuracy/%
Natural	0	0
PGD-Adv	0	68.6

4.4 评估

由于网络流量数据集中的每种流量都有比较固定的分类模式，不同类型的流量大部分可以通过端口号和 IP 等五元组信息进行区分，而基于卷积神经网络的网络流量分类方法将网络流量转换为图像进行分类，模糊了这种流量特性，微小的五元组信息修改不会影响图像的生成，但会导致分类的错误，这也是对抗攻击有效性的原因。

我们通过攻击实验验证了对抗攻击对网络流量分类具有强大的攻击性，无论是单步攻击、FGSM 攻击还是强大的迭代攻击 PGD 攻击、BIM 攻击或 CW 攻击，都能对基于卷积神经网络的网络流量分类方法造成极大的威胁。当对抗攻击的扰动强度为 40/255 时，卷积神经网络模型就已经丧失了对攻击的防御能力，这是因为轻微的图像扰动会让流量信息产生大幅度的偏差，从而使卷积神经网络模型无法对对抗流量样本进行正确分类。为了尽可能减小对抗攻击带来的影响，我们提出了采用混合对抗训练的方式进行防御，利用对抗流量样本的多样性来增强分类模型的鲁棒性，防御实验也验证了这一猜想。卷积神经网络模型可以通过学习对抗流量样本的特征来降低其对攻击样本的误分类率。然而，我们发现防御措施仍然存在不足，虽然混合对抗训练方法能显著提高卷积神经网络模型的鲁棒性，但由于分类的准确率仍不够高，因此还是不能作为使用卷积神经网络进行流量分类的前提，如何更加有效地防御对抗攻击的干扰是目前亟待解决的一个问题。

结束语 本文指出了对抗攻击对基于卷积神经网络的网络流量分类方法的威胁性，并通过对比实验验证了该攻击方法能够大幅降低网络流量分类的准确性，在人眼不可识别的扰动强度下可以使模型的分类能力降低至 0。与此同时，本文也提出了混合对抗训练方法作为相应的防御措施，该防御方法的有效性也在实验中得以验证，在流量图像允许的扰动范围内能够保持分类器的分类鲁棒性。在未来的工作中，我们计划研究直接对网络流量生成图像的过程进行攻击，发掘网络流量分类中可能存在的防御漏洞，以更好地完善神经网络在网络流量分类中的应用。

参 考 文 献

[1] ZHANG F, HE W, LIU X, et al. Inferring users' online activities through traffic analysis[C]// Proceedings of the Fourth ACM Conference on Wireless Network Security. 2011: 59-70.

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

[3] WANG W, ZHU M, WANG J, et al. End-to-end encrypted traffic classification with one-dimensional convolution neural networks[C]// 2017 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE, 2017: 43-48.

[4] DRAPER-GIL G, LASHKARI A H, MAMUNM S I, et al. Characterization of encrypted and vpn traffic using time-related [C]// Proceedings of the 2nd international conference on information systems security and privacy (ICISSP). 2016: 407-414.

[5] LOTFOLLAHI M, SIAVOSHANI M J, ZADER S H, et al. Deep packet: A novel approach for encrypted traffic classification using deep learning[J]. Soft Computing, 2020, 24(3): 1999-2012.

[6] MARÍN G, CASAS P, CAPDEHOURAT G. Deep in the Dark-Deep Learning-Based Malware Traffic Detection Without Expert Knowledge[C]// 2019 IEEE Security and Privacy Workshops (SPW). IEEE, 2019: 36-42.

[7] HE Y, LI W. Image-based encrypted traffic classification with convolution neural networks[C]// 2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC). IEEE, 2020: 271-278.

[8] AHMAD Z, KHAN A S, SHIANG C W, et al. Network intrusion detection system: Asystematic study of machine learning and deep learning approaches[J]. Transactions on Emerging Telecommunications Technologies, 2021, 32(1): e4150.

[9] WU H. A Systematical Study for Deep Learning Based Android Malware Detection[C]// Proceedings of the 2020 9th International Conference on Software and Computer Applications. 2020: 177-182.

[10] MERCALDO F, SANTONE A. Deep learning for image-based mobile malware detection[J]. Journal of Computer Virology and Hacking Techniques, 2020, 16(6): 1-15.

[11] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv: 1412.6572, 2014.

[12] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep

learning models resistant to adversarial attacks[J]. arXiv:1706.06083,2017.

[13] SCHOTT L,RAUBER J,BETHGE M,et al. Towards the first adversarially robust neural network model on MNIST[J]. arXiv:1805.09190,2018.

[14] CARLINI N,WAGNER D. Towards evaluating the robustness of neural networks[C]// 2017 IEEE Symposium on Security and Privacy (sp). IEEE,2017:39-57.

[15] MOOSAVI-DEZFOOLI S M,FAWZI A,FROSSARD P. Deepfool:a simple and accurate method to fool deep neural networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:2574-2582.

[16] KURAKIN A,GOODFELLOW I,BENGIO S,et al. Adversarial examples in the physical world[C]// International Conference on Learning Representations. 2017.

[17] WANG Z. The applications of deep learning on traffic identification[J]. BlackHat USA,2015,24(11):1-10.

YANG Yang, born in 1997, postgraduate. His main research interest includes information security of artificial intelligence.

CHEN Wei, born in 1978, Ph.D, associate professor. His main research interest includes network security and so on.