(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116366299 A
(43)申请公布日 2023.06.30

(21)申请号 202310187942.4

(22)申请日 2023.03.02

(71)申请人 北京理工大学
地址 100081 北京市海淀区中关村南大街5号

(72)发明人 祝烈煌 潘天瑶 徐大伟 高峰 赵鑫

(74)专利代理机构 北京正阳理工知识产权代理事务所(普通合伙) 11639
专利代理师 王松

(51)Int.Cl.
H04L 9/40(2022.01)
G06N 3/0464(2023.01)
G06N 3/08(2023.01)
G06F 21/55(2013.01)
G06F 18/241(2023.01)
G06F 18/214(2023.01)

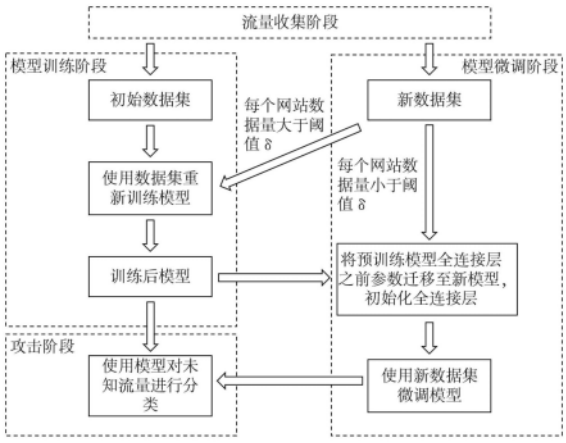
权利要求书2页 说明书4页 附图1页

(54)发明名称

一种基于深度学习的网站指纹攻击识别方法

(57)摘要

本发明涉及一种基于深度学习的网站指纹攻击识别方法,属于计算机网络安全中的加密流量识别技术领域。攻击者监听客户端和入口中继节点入口之间的通信,提取出数据包方向和时间信息作为网站指纹。然后,攻击者为网站指纹攻击创建攻击模型,该模型以数据包方向和时间戳两个序列作为输入,以网站类别作为输出。为训练攻击模型,攻击者使用收集的流量数据作为训练集,然后使用训练集来训练CNN模型,该模型被用作分类器来执行分类任务。之后,攻击者使用经过训练的分类器执行网站指纹攻击识别。定期更新训练后的模型,模型更新后,在攻击阶段继续使用,对未知流量进行分类识别。对比现有技术,本发明识别准确度高,同时模型训练开销小。



CN 116366299 A

1.一种基于深度学习的网站指纹攻击识别方法,其特征在于,包括以下步骤:

步骤1:流量收集;

攻击者监听客户端和入口中继节点入口之间的通信,提取出数据包方向和时间信息,作为网站指纹;

步骤2:模型训练;

攻击者为网站指纹攻击创建攻击模型,即CNN模型,该模型以数据包方向和时间戳两个序列作为输入,以网站类别作为输出;

为训练攻击模型,攻击者使用收集的流量数据作为训练集,然后使用训练集来训练CNN模型,该模型被用作分类器来执行分类任务;

步骤3:网站指纹攻击;

攻击者使用经过训练的分类器执行网站指纹攻击;

首先,攻击者捕获用户和入口节点之间的未知流量,然后将未知流量馈送到训练的分类器中进行分类,以推断流量的目标网站;

步骤4:模型微调;

定期更新训练后的模型:攻击者为每个受监控的网站重新收集若干示例数据,在训练阶段得到的训练后模型将被作为预训练模型,攻击者使用新的流量数据对预训练模型的参数进行微调;

当模型调整完毕后,攻击者使用调整后的模型,对新的未知流量进行分类,重新进行攻击过程识别。

2.如权利要求1所述的一种基于深度学习的网站指纹攻击识别方法,其特征在于,步骤1中,攻击者首先选择一组感兴趣的网站,这些网站称为受监控的网站,针对这些受监控的网站进行流量收集;

攻击者只能监听客户端和入口中继节点之间的通信,不能插入、修改或丢弃数据包。

3.如权利要求1所述的一种基于深度学习的网站指纹攻击识别方法,其特征在于,步骤2中,攻击模型为CNN模型,共有12个卷积层,每个卷积层之后是归一化层和激活层;在池化层之前,使用2个卷积层来增加网络深度,确保CNN模型充分学习模式;

模型包括三个模块:方向模型f_d、时间模块f_t和结合模块f_c;其中,方向序列表示为D,$D=(d_1,d_2,\cdots,d_L)$,$d_i\in\{-1,+1\}$;时间序列表示为T,$T=(t_1,t_2,\cdots,t_L)$,$t_i>0$;模型的输入为X,$X=(D,T)$,包含方向和时间序列;

初始时,序列D和T分别被输入到方向模块和时间模块中,得到相应特征图$D'=f_d(D)$、$T'=f_t(T)$;然后,D'和T'被连接并馈送到结合模块中,得到X属于特定类别的概率$\hat{Y}=f_c(D'||T')$;与其他块相比,结合模块种在每两个卷积块之前添加一个池化层,并在其之后添加一个丢弃层;在全连接层之前,卷积的输出由全局平均池化层转换为矢量;

当得到$\hat{Y}$后,使用$\hat{Y}$和原始的数据标签Y计算训练损失Loss来更新模型参数;在交叉熵损失的基础上使用标签平滑策略,将随机噪声添加到原始标签表示的每个维度。

4.如权利要求1所述的一种基于深度学习的网站指纹攻击识别方法,其特征在于,步骤3中,攻击者在客户端和入口中继节点之间的链路上进行监听,获得用户访问未知网站的未知流量,提取出未知流量中的数据包大小和时间序列,输入到初始训练好的模型或调整好

的模型中,得到网站分类结果。

5.如权利要求1所述的一种基于深度学习的网站指纹攻击识别方法,其特征在于,步骤4中,采用的微调机制,具体如下:

攻击者使用源数据集训练健壮的模型;训练过程中训练好的CNN模型被视为预训练模型;当流量模式发生变化,模型无法准确识别网站时,攻击者重新为每个受监控的网站重新收集N个示例;规定一个阈值δ,如果攻击者有能力为每个网站收集的示例数量N大于阈值δ,则攻击者选择重新训练模型;如果示例数量N不大于阈值δ,则攻击者需要使用新的流量数据来微调模型的参数;

模型微调时,将预训练模型全连接层之前全部参数迁移至相同的新模型,新模型全连接层只需进行初始化,之后攻击者使用新的流量数据对新模型进行微调;

当模型调整完毕后,攻击者使用调整后的模型,对新的未知流量进行分类,重新进行攻击过程识别。

一种基于深度学习的网站指纹攻击识别方法

技术领域

[0001] 本发明涉及一种基于深度学习的网站指纹攻击识别方法，属于计算机网络安全中的加密流量识别技术领域。

背景技术

[0002] 随着隐私保护意识的提高，互联网用户倾向于在通信中隐藏真实的访问目标来保护个人隐私信息。匿名通信网络Tor致力于保护用户访问网络的隐私，传输数据时，它会对数据进行多重加密，并随机选择三个节点建立链路，能够保证任何一个Tor节点或窃听者都无法将用户身份与用户访问的网站联系起来，从而实现对网站的匿名访问。但是，匿名网络在保护用户隐私的同时也产生了新的网络安全问题，许多不法份子利用匿名网络来掩盖其网络犯罪行为。因此，针对匿名网络流量的监管技术是十分重要的。

[0003] 网站指纹攻击是一种新型的流量识别技术，能够降低Tor的匿名性。用户访问不同网站所产生的流量中的一些信息可以形成该网站的指纹，例如数据包方向、时间和大小。网站指纹攻击正是利用流量中的网站指纹信息来实现对网站的分类。近年来，基于深度学习的方法已经逐渐取代基于机器学习的方法，取得了不错的效果，成为研究的热点。然而，为了保护Tor的匿名性，许多针对网站指纹攻击的防御方法被提出，能够有效降低分类的准确率。并且由于流量模式的快速变化，训练的模型很难长时间维持有效性，使用大量样本不断重新训练模型会消耗大量的计算资源和时间成本。因此，有必要提出一个更加有效的网站指纹攻击方法来应对流量模式的动态变化和新的防御方法带来的挑战。

[0004] 目前，现有的基于深度学习的网站指纹攻击识别技术包括以下方案：

[0005] 方案一：基于卷积神经网络的网站指纹攻击方法。这种方法借鉴了图像识别领域有效的深度学习模型，设计了比之前网站指纹攻击研究更复杂的卷积神经网络架构。在训练数据量充足的情况下，能够达到高识别准确率。

[0006] 方案二：基于小样本学习的网站指纹攻击方法。这种方法将小样本学习相关模型迁移到网站指纹攻击的场景下。攻击者先预先训练一个有效的模型并将其固定为特征提取器，之后使用特征提取器为少量的目标网站训练数据提取特征，最后使用这些特征训练分类器。此方法打破了大多数研究中存在的测试和训练数据具有相似分布的不现实假设，能够减少大型数据集的收集和训练工作量，并减轻处理不同网络条件的不利影响。

[0007] 但是，上述方案仍然存在以下缺陷：

[0008] 1.准确度低。第一种方案若要达到理想的识别效果，需要大量的训练数据，但流量数据的收集和更新较为困难，当训练数据不足时，此方案的识别准确度会大幅下降，无法达到实际需要的识别准确度。第二种方案虽然支持训练与测试数据分布不同，但其识别准确度却并不理想。同时，这两种方案都无法有效地对经过新的防御方法的流量进行识别。

[0009] 2.开销大。第一种方案不支持训练和测试数据分布不同的情况，因此，每经过一段时间就需要对模型进行重新训练，再加上该方案每次训练模型都需要大量的训练数据，收集数据和训练模型开销非常大。第二种方案由于模型更加复杂，训练特征提取器时本身的

训练开销就比较大。

发明内容

[0010] 本发明的目的是针对现有的网站指纹攻击识别方法存在准确性低、训练开销大等缺陷和不足，导致实用性低的技术问题，创造性地提出一种基于深度学习的网站指纹攻击识别方法。

[0011] 本发明的目的是通过以下技术方案解决的。

[0012] 首先对本发明涉及的技术术语进行说明。

[0013] 卷积神经网络(CNN)：是一种广泛应用于分类任务的深度网络，在图像分类、语音识别等领域证明了其有效性。CNN主要通过多个卷积层、池化层和非线性激活函数从原始输入数据中自动提取特征。批量归一化层和丢弃层通常在卷积层之后使用，以防止过度拟合并提高性能。CNN的最后一部分是全连接层，它将所有本地特征合并为全局特征，以计算每个类别的最终得分。

[0014] 迁移学习：一种机器学习技术，它可以将在源任务上学习的知识转移到目标任务，从而提高目标任务模型预测的性能。迁移学习之所以有效，是因为模型的浅层一般能够学习到任务的普遍特征，而随着网络的深入，深层更侧重于学习任务的特定特征。这可以直接转移模型浅层，然后调整较深层以适应新任务。

[0015] 微调：微一种转移学习方法，它节省了大量的计算资源和时间。如果新数据集与预先训练的数据集相似，那么对训练模型的微调可以使模型适应新数据集。

[0016] 一种基于深度学习的网站指纹攻击识别方法，包括以下步骤：

[0017] 步骤1：流量收集。

[0018] 攻击者需要监听客户端和入口中继节点入口之间的通信，提取出数据包方向和时间信息作为网站指纹。

[0019] 步骤2：模型训练。

[0020] 攻击者为网站指纹攻击创建攻击模型，即CNN模型，该模型以数据包方向和时间戳两个序列作为输入，以网站类别作为输出。

[0021] 为训练攻击模型，攻击者使用收集的流量数据作为训练集，然后使用训练集来训练CNN模型，该模型被用作分类器来执行分类任务。

[0022] 步骤3：网站指纹攻击。

[0023] 攻击者使用经过训练的分类器执行网站指纹攻击。首先，攻击者捕获用户和入口节点之间的未知流量，然后将未知流量馈送到训练的分类器中进行分类，以推断流量的目标网站。

[0024] 步骤4：模型微调。

[0025] 由于网站流量模式不时变化，因此需要定期更新训练后的模型。

[0026] 攻击者需要为每个受监控的网站重新收集若干示例数据。在训练阶段得到的训练后模型将被作为预训练模型，攻击者只需使用新的流量数据对预训练模型的参数进行微调，即可达到使模型适应新的流量模式的目的。

[0027] 当模型调整完毕后，攻击者使用调整后的模型，对新的未知流量进行分类，重新进行攻击过程识别。

[0028] 有益效果

[0029] 本发明，对比现有技术，具有以下优势：

[0030] 1.准确度高。相比方案一和方案二，本方法改进了攻击模型的架构，同时使用数据包方向和时间作为网站指纹，提升了网站识别的准确性，缓解由于训练数据不足导致的准确性下降，并且能够抵御常见的防御策略。

[0031] 2.开销小。本方案能够用更少的数据达到更好的识别效果，放宽了攻击的要求，减小了训练开销。在支持数据分布不同的方面，采用了更为简洁的迁移学习的思想，相比方案二，本方案显著缩短了训练时间。

附图说明

[0032] 图1为本发明方法的整体流程示意图；

[0033] 图2为本发明的模型结构示意图。

具体实施方式

[0034] 下面结合附图对本发明做进一步详细说明。

[0035] 如图1所示，一种基于深度学习的网站指纹攻击识别方法，包括以下步骤：

[0036] 步骤1：流量收集。

[0037] 若要实现对用户访问网站的识别，攻击者需收集流量数据作为训练集来训练有效的攻击模型。

[0038] 具体地，攻击者首先选择一组感兴趣的网站，这些网站称为受监控的网站，针对这些受监控的网站进行流量收集。因网站指纹攻击属于被动攻击，攻击者只能监听客户端和入口中继节点之间的通信，不能插入、修改或丢弃数据包。由于数据包内容被加密无法得到，因此，只需提取出相应网站流量跟踪的数据包方向和时间信息作为该网站的网站指纹。

[0039] 步骤2：模型训练。使用流量收集阶段收集到的数据集来训练攻击模型。

[0040] 具体地，如图2所示，攻击模型为CNN模型，共有12个卷积层，每个卷积层之后是归一化层和激活层。在池化层之前，使用2个卷积层来增加网络深度，从而确保CNN模型充分学习模式。

[0041] 为更加清楚地描述模型的架构，将模型分成三个模块，包括方向模型f_d、时间模块f_t和结合模块f_c。其中，方向序列表示为D，$D=(d_1,d_2,\cdots,d_L)$，$d_i\in\{-1,+1\}$；时间序列表示为T，$T=(t_1,t_2,\cdots,t_L)$，$t_i>0$；模型的输入为X，X＝(D,T)，包含方向和时间序列。

[0042] 初始时，序列D和T分别被输入到方向模块和时间模块中，得到相应特征图$D'=f_d(D)$、$T'=f_t(T)$。然后，D'和T'被连接并馈送到结合模块中，得到X属于特定类别的概率$\hat{Y}=f_c(D'||T')$。与其他块相比，结合模块种在每两个卷积块之前添加一个池化层，并在其之后添加一个丢弃层。在全连接层之前，卷积的输出由全局平均池化层转换为矢量，能够更好集成全局空间信息并减少参数的数量。

[0043] 当得到$\hat{Y}$后，使用$\hat{Y}$和原始的数据标签Y计算训练损失Loss来更新模型参数。本方法在交叉熵损失的基础上使用标签平滑策略，此为一种正则化方法，将随机噪声添加到原始标签表示的每个维度，这种策略能够避免模型过度拟合的问题，并使模型具有更强的泛

化能力。

[0044] 步骤3:攻击阶段。

[0045] 该阶段是网站指纹攻击的执行阶段。攻击者要在客户端和入口中继节点之间的链路上进行监听,获得用户访问未知网站的未知流量,提取出未知流量中的数据包大小和时间序列,输入到初始训练好的模型或调整好的模型中,从而得到网站分类结果。

[0046] 步骤4:对模型进行微调。

[0047] 由于流量模式的不断变化,CNN模型不能始终保持高精度。周期性地重新收集大量数据非常困难,如何使用少量流量数据使模型在长时间内有效成为一个棘手的问题。

[0048] 本方法充分利用迁移学习的思想,设计了一种微调机制,使模型能够支持新的数据分布。具体地,所述微调机制的工作过程如下:

[0049] 攻击者使用源数据集训练健壮的模型。训练过程中训练好的CNN模型被视为预训练模型。当流量模式发生变化,模型无法准确识别网站时,攻击者重新为每个受监控的网站重新收集N个示例。考虑到流量收集的难度,N通常设置得很小(例如每个网站5个示例)。本方法中规定一个阈值δ,如果攻击者有能力为每个网站收集的示例数量N大于阈值δ,则攻击者选择重新训练模型。如果示例数量N不大于阈值δ,则攻击者需要使用新的流量数据来微调模型的参数。

[0050] 模型微调时,要将预训练模型全连接层之前全部参数迁移至相同的新模型,新模型全连接层只需进行初始化,之后攻击者使用新的流量数据对新模型进行微调即可。

[0051] 当模型调整完毕后,攻击者使用调整后的模型,对新的未知流量进行分类,重新进行攻击过程识别。

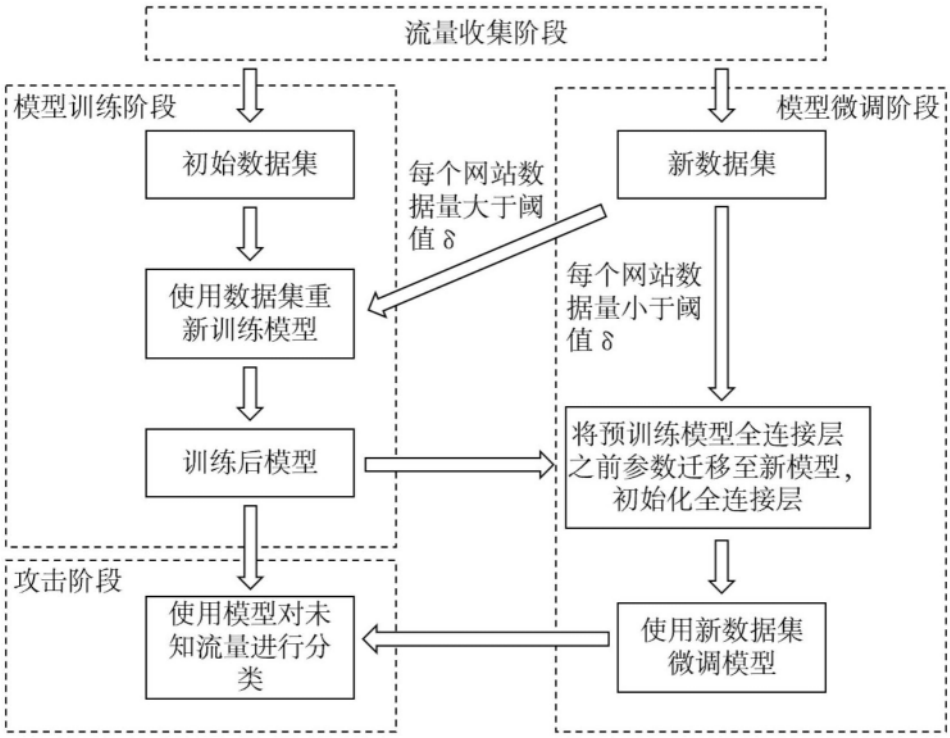

流量收集阶段
模型训练阶段
初始数据集
使用数据集重新训练模型
训练后模型
攻击阶段
使用模型对未知流量进行分类
每个网站数据量大于阈值δ
每个网站数据量小于阈值δ
模型微调阶段
新数据集
将预训练模型全连接层之前参数迁移至新模型，初始化全连接层
使用新数据集微调模型

图1

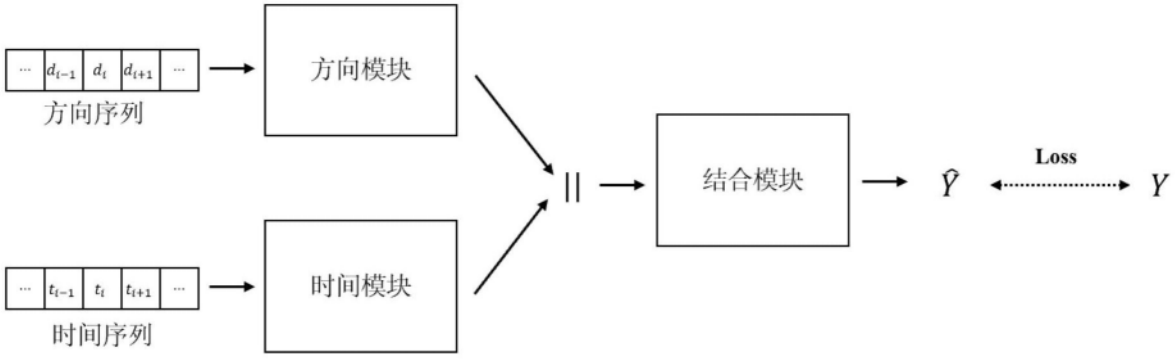

方向序列
方向模块
时间序列
时间模块
结合模块
Loss

图2