计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(12):172−179 [doi: 10.15888/j.cnki.csa.008203]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于表征学习的网络游戏流量识别①

徐星晨[1], 张　俊[1,2], 年　梅[1]

[1](新疆师范大学 计算机科学技术学院, 乌鲁木齐 830054)
[2](中国科学院 新疆理化技术研究所, 乌鲁木齐 830011)
通讯作者: 年　梅, E-mail: 2468830639@qq.com

摘　要: 进行基于表征学习的网络游戏流量识别研究. 首先, 由于流量识别领域公开数据集中缺乏游戏流量, 采集各类游戏流量, 并建立各种游戏与进程端口的映射关系, 基于该映射关系从采集的流量中过滤游戏流量, 扩展公开数据集; 利用深度学习中的表征学习模型, 对经过预处理的原始端到端游戏流量自动进行特征学习和特征选择; 最后用分类器进行游戏类别识别. 通过构建特征空间由卷积神经网络自学习原始信息的特征, 成功避免传统机器学习算法中流量数据集的二次处理导致的信息丢失以及流量分类模型对特征选择的依赖. 实验结果表明, 相比于原数据集的分类效果, 扩充后的数据集在神经网络模型上的分类准确率提高了 5%, 游戏流量识别准确率达到 92%, 识别性能明显提升.

关键词: 深度学习; 卷积神经网络; 流量识别; 游戏流量

引用格式: 徐星晨,张俊,年梅.基于表征学习的网络游戏流量识别.计算机系统应用,2021,30(12):172–179. http://www.c-s-a.org.cn/1003-3254/8203.html

Online Game Flow Identification Based on Representation Learning

XU Xing-Chen[1], ZHANG Jun[1,2], NIAN Mei[1]

[1](School of Computer Science and Technology, Xinjiang Normal University, Urumqi 830054, China)
[2](Xinjiang Institute of Physics and Chemistry Technology, Chinese Academy of Sciences, Urumqi 830011, China)

Abstract: This study explores the online game flow identification based on representation learning. First of all, due to the lack of game flow in the public data set in the field of flow identification, various types of game flow are collected, and a mapping relationship between various games and process ports is established. Depending on the mapping relationship, the game flow is filtered from the collected flow to expand the public data set. Then the representation learning model in deep learning is used to automatically perform feature learning and feature selection on the pre-processed original end-to-end game flow. Finally, the game category is identified by a classifier. The convolutional neural network self-learns the features of the original information via the construction of feature space, successfully avoiding the loss of information caused by the secondary processing of the flow data set in the traditional machine learning algorithm and the dependence of the flow classification model on feature selection. The experimental results show that, compared with the classification effect of the original data set, the expanded data set has a classification accuracy improved by 5% on the neural network model. The accuracy of game flow identification reaches 92%, and the identification performance is significantly improved.

Key words: deep learning; Convolutional Neural Network (CNN); flow identification; game flow

① 基金项目: 新疆维吾尔自治区高校科研项目 (XJEDU2017S032); 新疆师范大学数据安全重点实验室招标项目 (XJNUSYS102017B04)
Foundation item: Scientific Research Fund of Higher Education of Xinjiang Uygur Autonomous Region (XJEDU2017S032); Tender Project of Data Security Key Laboratory of Xinjiang Normal University (XJNUSYS102017B04)
收稿时间: 2021-02-25; 修改时间: 2021-03-19; 采用时间: 2021-03-26

1 引言

网络流量应用识别是指对网络中的混合流量按照应用协议进行识别. 网络流量应用识别既是高性能网络协议设计的基础, 又是网络运营管理、网络发展规划的依据, 也是网络攻击与恶意代码检测的重要手段[1]. 当前高校部分学生沉迷于网络游戏, 玩游戏既占用了校园网资源也影响了学生的学习, 校园网资源的精细化管理需要掌握校园网游戏使用状况以及占用的网络资源比率, 由此进行网络资源调控. 为此, 需要对校园网流量中的游戏进行识别, 准确获取其占用的带宽状况、学生花费的时间等, 然后针对性地对校园网资源进行配置, 对玩游戏的学生进行预警.

近年来, 流量分类的研究经过不断地发展, 从最初基于端口、基于特征匹配、基于主机行为的流量分类方法发展到基于机器学习技术的流量分类方法.

随着流媒体、P2P 等网络通信协议的发展, 动态协商端口与端口伪装技术的应用, 基于端口的识别方法已逐渐失效. 为保护网络通信的安全, 很多网络应用采用加密协议或协议格式未公开的私有协议进行数据通, 从而导致基于特征匹配方法的识别精度日益下滑. 基于主机行为的方法并不能很好的识别未知协议与加密协议的网络流量的应用类型.

随着网络应用种类越来越多, 基于机器学习技术的流量分类方法逐渐成为主流. 机器学习的流量识别模型训练依赖标注数据集, 目前网络流量分类重点是视频流量、P2P 流量、异常流量、加密流量的识别, 而网络游戏流量的识别并没有引起传统学术界和产业界的关注, 研究成果非常少, 并且缺少公开网络游戏流量标准数据集.

目前存在的问题是标注数据集的缺乏与游戏流量研究在教育领域的重要性与实际研究领域对游戏流量不重视所产生的矛盾, 对于多种类细粒度的游戏流量数据集构建以及识别研究存在着新的挑战. 针对游戏流量数据集的标记构建存在的难题, 本文首次提出采用基于端口映射关系的游戏流量数据集标注方法, 对公开数据集进行扩充, 并以 NPY 的文件形式进行存储以减少数据集内存占用、提高模型的读取效率. 其次将扩充后的数据集应用到 Wang 等人[2] 提出的一种端到端的流量识别模型中, 采用 Keras 模块对其代码进行重构, 并对一维卷积神经网络进行参数调优, 通过卷积神经网络构造特征空间, 自主提取特征, 最后通过实验对比本文扩充后的数据集与原公开数据集在同一模型结构下的识别效果, 以评估基于端口策略的游戏流量样本扩充方法的可行性与效果.

2 相关工作

流量识别的应用多采用基于规则的识别方法, 该方法相对成熟, 主流研究者主要研究如何准确地提取匹配规则. 而基于机器学习的流量识别方法是目前学术界研究较多的内容, 主要研究如何选择更好的特征集来对识别效果进行优化改进.

目前, 研究人员逐渐关注深度学习方法在流量识别领域的应用, 基于表征学习的流量分类方法的研究也在流量识别领域初露头角. 深度学习通过训练多个单层特征图构建非线性网络, 根据训练出来卷积核的权重参数组构成底层特征的抽象表示, 从而发现数据的本质特征以达到识别的作用. 王勇等人[3] 提出一种基于卷积神经网络的流量分类算法, 分别采用公开数据集和私有数据集进行测试, 通过学习空间特征避免了传统分类方法中的特征的人工筛选, 提高了流量分类的精确率, 减少了分类使用的时间, 但该方法并未涉及对加密流量的分类识别. Gao 等人[4] 提出了一种使用深度信念网络 DBN 的恶意流量分类方法, Javaid 等人[5] 提出了一种使用稀疏自编码器 SAE 的恶意流量分类方法. 但是他们放弃了深度学习能够从原始数据中直接学习特征的优点, 而是对处理后的流特征数据集进行了学习, 表征学习的方法在图像、语音领域的识别效果都充分说明了该方法的优越性. 对此 Wang 等人[2] 提出了一种基于端到端的加密流量识别方法, 取得了很好的效果, 而流量分类与协议识别在任务内容上是很相近的. 基于此, 文献 [6] 提出采用 Deep Packet 框架对流量进行识别, 通过嵌入堆叠的自动编码器和卷积神经网络, 将网络流量分为主要类别 (例如 FTP 和 P2P) 和应用程序标识, 在 ISCX VPN-nonVPN 数据集上的性能较好, 但对于数据的预处理和模型参数的选择等方面论述得不够清晰.

综上所述, 本文提出一种基于端口映射的流量标注方法对公开流量数据集进行扩充, 在扩充游戏流量数据集的基础上提出基于表征学习方法对网络游戏流量进行识别, 并采用 CNN 模型, 直接在原始流量数据上尝试进行流量分类, 通过对比同一模型在数据集扩充前后的分类效果验证数据集扩充的合理性, 验证了

标注的准确性和数据集扩充的可行性, 论证了表征学习方法在游戏流量识别方面的可行性.

3 基于表征学习的游戏流量识别

鉴于机器学习的流量识别性能依赖于数据集的标注, 公开数据集中缺乏游戏流量数据, 本文提出了基于端口映射的游戏流量数据集标注方法.

3.1 方法概述

研究发现, 网络游戏流量数据依赖于应用层的进程端口号进行数据接收与发送, 但是游戏传输中端口号存在动态性、随机性, 难以人工识别, 无法给出游戏流量的准确高效的标注, 导致公开数据集中缺乏游戏流量样本. 鉴于此, 本文提出俘获网络游戏流量的同时监控对应网络游戏的端口使用情况, 记录每个时段下的游戏进程端口的使用数据, 形成俘获日志. 利用日志信息编写了用于分析整合日志的相关程序, 能够根据游戏进程端口号的使用情况对已俘获的网络流量数据进行清洗, 以此得到网络游戏流量的原数据包.

本文采用的数据集标注方法分为 3 个阶段: 数据采集、数据预处理、数据集构建.

数据采集: 使用 Tcpdump、Wireshark 等抓包工具收集游戏通信过程中的网络流量, 同时游戏进程通信使用的端口将记录在日志中, 为游戏流量的过滤以及数据清洗做准备.

数据预处理: 根据进程端口日志清洗俘获的流量数据集, 去除与研究无关的流量数据, 并将清洗完成的流量数据集按照游戏类型添加标注信息, 对俘获的网络游戏流量与公开数据集原始流量进行流重组与切分、归一化处理为数据集构建做准备.

数据集构建: 采用本文完成开发的流量数据集构建程序对完成预处理的原始流量进行数据集构建, 将原始网络流量转化成符合卷积神经网络输入的数据集格式.

3.2 数据预处理

为实现从俘获的网络流量中提取出网络流, 并将网络流转换成符合卷积神经网络输入的格式, 本文将数据预处理可以细分为 3 个子步骤: 数据清洗、流重组与切分、数据归一化.

3.2.1 数据清洗

将获取的游戏通信过程日志作为输入至编写分析程序中, 利用日志中信息获取对应的游戏进程在每个时间段分别占用了哪些端口号. 先根据端口号信息对原始流量数据进行初次过滤, 再根据时间信息结合端口号的使用情况对流量数据进行二次过滤, 完成对网络游戏流量数据的清洗, 具体流程如图 1 所示.

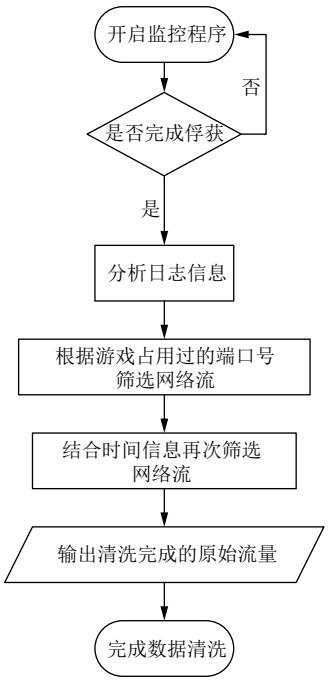


图 1 数据清洗流程图

3.2.2 流重组与切分

本文处理的网络游戏流量数据, 存在完整的 TCP 连接与 UDP 交互. 所以本文根据流量信息中的五元组匹配原则进行流的重组与切分. 对于 TCP 流, 利用 TCP 首部的序列号和标识将到达的数据包重新整合为一条有序流[7]. 对于 UDP 流, 根据数据包的发送时间确定 UDP 流的开始和结束, 在规定时间内未捕获流的下一个数据包认为这条流结束, 而后将指定时间窗口内的 UDP 数据包按照捕获的先后顺序进行拼接.

在提取出网络流信息之后, 需要进行流切分以得到大小相同的流数据用于卷积神经网络的训练. 本文选取网络流前部的一段固定长度的数据作为流量识别的依据, 一方面可以确保利用应用层的首部信息识别应用, 另一方面, 已有实验证明数据载荷中的前部分的数据往往更能够体现应用层协议的特征. 针对卷积神经网络要求输入数据格式相同的要求, 本文参考陈雪娇等人[8]、冯文博等人[9]、 Wang[10] 的实验, 采用每条网络流前 784 字节的数据作为判别依据.

3.2.3 数据归一化

由于俘获的网络流量数据的字节取值范围较大, 用于数值求解和模型训练会导致计算复杂度较大等问题, 为了便于卷积神经网络的分析处理, 需要将协议数据归一化. 归一化的具体步骤如下:

首先, 构造 n 个长度为 784 字节的一维向量 x, 即 $x_i=[x_{i1},x_{i2},x_{i3},\cdots,x_{ij}]$, 其中 i 代表 n 个一维向量中的第 i 个向量, j 代表第 i 个向量中的第 j 个元素. 然后将每个流量样本中字节对应的十进制数值赋给向量中的每个分量, 在十进制转化后 x 中的每个元素范围是 [0, 255], 为提高模型的计算效率、让流量向量数据取值分布更加紧凑, 需要对 x 中每个分量的数值进行归一化处理. 在本文中将每个分量的数值除以 255, 使分量的取值统一到 [0, 1) 区间, 构建数据集矩阵 M:

$$M=\frac{1}{255}\begin{pmatrix} x_{11} & \cdots & x_{1j} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} \end{pmatrix} \tag{1}$$

该归一化方法能让不同协议数据处于同一个数量级以用于对比, 提高模型的学习能力的同时降低了模型的计算复杂度.

由于卷积神经网络的输入通常是二维矩阵, 还需要将向量 x 转化为具有图像特征的二维矩阵. 本文将每个 x 中的元素按照顺序构建成 28×28 的二维矩阵 X:

$$X_i=\begin{pmatrix} x_{i1} & \cdots & x_{i28} \\ \vdots & \ddots & \vdots \\ x_{i757} & \cdots & x_{i784} \end{pmatrix} \tag{2}$$

3.3 数据集构建

根据清洗完成留下的标注信息对网络流进行标注, 并扩充于公开的网络流量数据集中, 用于卷积神经网络的训练.

Wang[10] 提出了一种直接使用原始流量数据的基于栈式自编码器 SAE 的网络协议识别方法, 取得了很好的效果, 而流量分类与协议识别在任务内容上是很相近的. 表征学习方法的优势是能够直接从原始数据中自动学习特征, 其在图像分类和语音识别领域的成功应用都充分说明了这一点, 所以本文采用了表征学习的方法进行流量分类.

为最优化表征学习的效果, 本文采用了端到端的方式进行数据集构建, 以保证擅长图像分类任务的 CNN 能够直接在原始流量数据上进行游戏流量分类. 为保证数据集的可靠性, 减少私有数据集对结果可信度的影响, 鉴于存在细粒度分类并具有完整标注的公开流量数据集不多且大部分都是加密与安全相关的网络流量, 而 Wang 提出的端到端的流量识别方法基于 ISCXVPN 2016 数据集, 所以本文选用 ISCXVPN 2016 数据集[11] 作为基础流量, 鉴于 ISCXVPN2016 存在 VPN 流量和实际流量相差太大, 所以在本文仅采用了 ISCXVPN 2016 中 non-VPN 部分的标记流量, 避免其与实际流量相差太大的问题.

该数据集包含有两部分, 分别是基于 VPN 会话的应用流量和无 VPN 下的应用流量, 包含有完成标记的网页浏览 (Browsing)、电子邮件 (Email)、网络聊天 (Chat)、语音通讯 (VoIP)、多媒体流 (Streaming)、文件传输 (File transfer)、点对点 (TraP2P) 流量数据. 但是该数据集中的网页浏览流量标记存在问题, 例如 “Facebook_video.pcap”, 可以标记为作为“Browsing”也可以标记为“Streaming”, 故在文献 [10] 的端到端识别中对网页浏览流量进行了剔除, 在本文实验中也考虑到在目前的网络游戏中端游占较大比重, 而端游的游戏客户端相当于一个小型网络浏览器, 如果将网络浏览依旧作为标记流量参与实验的话会导致粗粒度分类与细粒度分类并存而影响到网络游戏流量的准确识别, 所以本文将该类流量也进行了剔除.

将公开数据集中的原始流量数据完成重组与切分、归一化操作后, 与采集并完成预处理的游戏流量数据进行合并, 构建成能够应用于本文神经网络训练的流量数据集, 经过扩展后由原数据集 22 976 个样本扩充为 25 906 个样本, 其中包含有穿越火线、炉石传说、英雄联盟、CSGO 等游戏流量, 数据来源为新疆师范大学数据安全实验室与研究生实验室, 于 2020 年 10 月至 12 月进行俘获, 为保证数据时间分布合理, 俘获时间为每月中的周末 10 点至 15 点. 在去除 IP、MAC 等冗余数据后, 本文对游戏流量数据进行可视化发现, 在校园网络环境下的不同区域内俘获的游戏流量存在一致性, 因此该游戏流量数据集能够代表整体特征.

随机选取流重组、切分和归一化之后的可视化结果如图 2 所示, 大小为 784 字节. 显然, 不同类别的流量具有明显的区分度, 并且各个类别的流量具有较高的一致性.

在经典的 MNIST 手写体识别数据集中图像文件也采用的是 28×28×1 的像素值, 与之不同的是为保证

学习模型能够从原始流量中进行特征学习和提高模型识别效率, 本文将构建完成的流量图片数据集的像素信息存储到了 NPY 文件中以用于模型的训练, 不仅保留了图片所包含的特征信息、缩减了数据集占用的内存空间, 还选用了适配神经网络的 NPY 文件, 提高了模型训练时读取数据集的时间效率.

3.4 表征学习模型构建

鉴于数据预处理时的图片尺寸与 MNIST 相同, 经实验证明 LeNet-5[12] 的 CNN 模型, 如图 3 所示, 对 MNIST 数据集的分识别准确度可达 99.2%, 因此本文将采用了类似 LeNet-5 的 CNN 结构, 如图 3 所示.

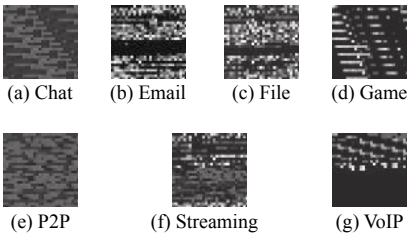


图 2　所有类别的流量可视化图

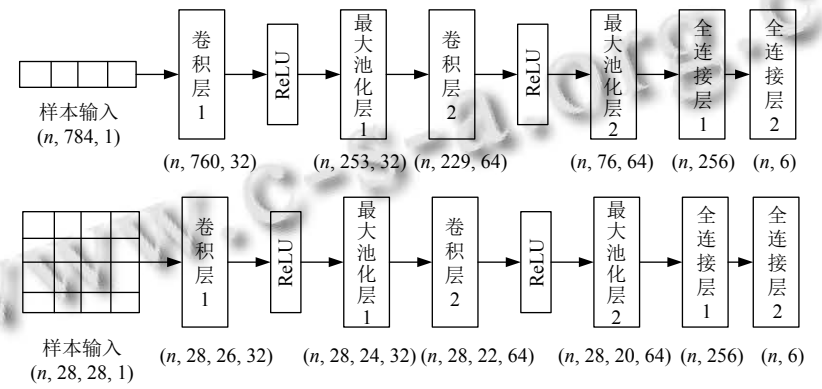


图 3　两种维度下的模型结构

为研究输入数据集结构不同, 是否会对构建的表征学习模型带来结果的差异性, 本文在实验中分别输入 784×1 与 28×28 的图像矩阵进行了对比试验, 如具体模型参数如表 1 所示.

表 1　一维卷积神经网络主要参数

Layer	Operaction	Filter	Stride	Pad
1	Conv+ReLU	32×3	**1**	Valid
2	1d max pool	1×3	**3**	Valid
3	Conv+ReLU	64×3	**1**	Valid
4	1d max pool	1×3	**3**	Valid
5	Dropout+dense	—	—	None
6	Dropout+dense	—	—	none

在该模型的最后两个全连接层中, 将数据尺寸依次转换为 1024 和 7, 前者采用 ReLU 作为激活函数, 并通过添加扁平层将输入数据拉伸成一位数据, 后者采用 Softmax 作为激活函数, 输出各类概率值. 为减少过拟合, 在输出层之前均采用了 dropout 进行随机失活, 前者为 0.25, 后者为 0.4, 本文在六分类、七分类两种实验中都使用了上述同一种结构.

4 实验设计与结果分析

本文通过实验对比在相同模型结构下公开数据集与完成扩充的网络游戏流量数据集的查准率、查全率、准确率和 *F-Measure* (*F*1) 值的变化, 分析本文模型的可用性, 并将游戏流量与其他应用流量进行分类效果对比, 验证本文构建数据集的可靠性.

4.1 评价指标

目前, 流量识别模型主要采用准确性相关指标来进行评估, 为了满足不断提高的流量分析要求, 参考文献 [13] 提出的技术评价指标, 本文在准确性的基础上从模型的完整性和未识别率等方面全面地评估流量识别模型在扩充前后数据集中的效果, 进行了更加客观公正的结果对比. 下面详细介绍实验对比中采用的评估指标.

准确性是反映流量识别技术识别网络应用的能力. 假设 N 为流量样本总数, m 为待识别的应用类型数, n_{ij} 表示实际类型为 i 的流量样本被标记为类型为 j 的样本数. 真正 (True Positive, *TP*) 代表实际类型为 i 的样本中被正确标记的样本数, 即 $TP_{ij}=n_{ii}$; 假正 (False Positive, *FP*) 代表实际类型为非 i 的样本中被错误标记的样本数, 即 $FP_i=n_{ji}, j\neq i$. 查准率定义为:

$$precision=\frac{TP}{TP_i+FP_i} \tag{3}$$

假负 (False Negative, *FN*) 代表实际类型为 i 的样本中被误标识为其他类型的样本数, $FN_i=\sum n_{ij}$. 真负 (True Negative, *TN*) 代表实际类型为非 i 的样本中被标识为非 i 的样本数, $TN_i=n_{jj}$. 查全率定义为:

$$recall=\frac{TP_i}{TP_i+FN_i} \tag{4}$$

查准率和查全率体现了识别方法在每个单独协议类别上的识别效果. 特别是当样本类别分布不均匀时, 查全率和查准率可以准确获知每个类别的分类情况. 准确率体现了识别方法的总体识别性能, 好的算法应该同时具有较高的准确率、查准率和查全率. 准确率定义为:

$$accuracy=\frac{\sum_{i=1}^{m}(TP_i+TN_i)}{\sum_{i=1}^{m}(TP_i+TN_i+FP_i+FN_i)} \tag{5}$$

F-Measure 是综合查准率和查全率得到的评价指标, *F-Measure* 越高表明算法在各个类型的分类性能越好.

$$F\text{-}Measure=\frac{2\times precision\times recall}{precision+recall} \tag{6}$$

4.2 实验设计

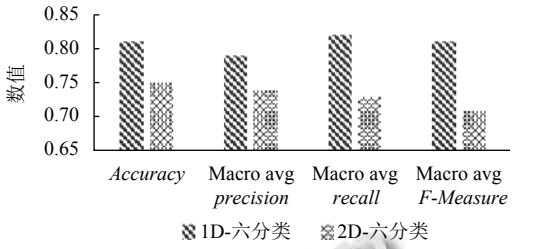
本文实验将随机选取数据集的 75% 用于识别模型的训练, 剩余 25% 的数据集用于分类模型的测试. 实验平台方面, 选用的软件框架是 TensorFlow[14]. 优化算法采用随机梯度下降算法, 并启用 Nesterov Momentum 算法更新反向梯度, 其中 Momentum 为 0.9, 损失函数为交叉熵函数, 学习速率 0.01, 学习速率的衰减系数为 0.0001, 训练回合数约为 25 epochs.

在采用不同的数据维度进行实验中, 我们发现基于表征学习的识别模型在一维数据下的表现优于二维, 实验结果见图 4.

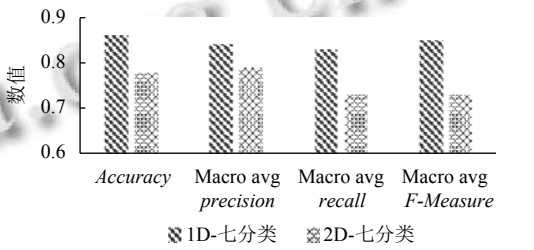
从图 4(a)、图 4(b) 对比中可以发现, 不同网络层数对最终识别率有着不同的影响, 相比二维的卷积网络, 一维卷积神经网络的表征学习识别模型在准确率与宏平均查准率、查全率和 *F-Measure* 值上均有优势, 所以本文针对基于一维卷积神经网络的表征学习识别模型进行细粒度的结果分析.

4.3 结果分析

在实验中, 本文将构建的数据集和公开网络流量数据集分别应用于识别模型中进行了实验, 实验结果见图 5 和图 6.



(a) 六分类识别准确率与宏平均查准率、查全率和 *F* 均值



(b) 七分类识别准确率与宏平均查准率、查全率和 *F* 均值

图 4 不同输入维度对识别效果的影响

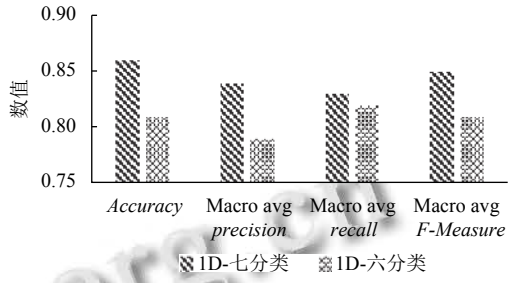


图 5 模型的识别准确率与宏平均查准率、查全率和 *F-Measure* 值

图 5 表示识别模型在扩充前后的两个数据集中的总体识别效果. 根据数据集扩充前后模型识别率对比, 可以看出: 通过对公开数据集的合理扩充, 有效提高了模型的识别准确率, 在原来基础上提高了 5%, 除了准确率得到提高外, 宏平均查准率、查全率和 *F-Measure* 值的数据对比也证实了数据集的扩充同样优化了模型其他应用流的识别精度. 图 6 中全方位展示了扩充前后的数据集在识别模型中查准率、查全率和 *F-Measure* 值的变化. 从中也可以看出与总体精度相同的比较结果. 从综合评价指标 *F-Measure* 值的对比中可以看出, 同一模型在扩充前后数据集上的识别中在原始流的分类结果几乎持平, 部分流的识别精度得到有效提升, 游戏流量识别准确率达到了 92%, 召回率达到 92%, 已经

与其他流量的识别精度相持平, 可见基于表征学习的网络游戏流量识别是可行的, 并取得的较好的效果. 通过在学生宿舍以及实验室等实际场景中部署流量监控进行验证, 游戏流量查准率可达到 88%, 所以本文所提出的基于表征学习的网络流量识别方法在网络游戏流量监控方面具有有效性.

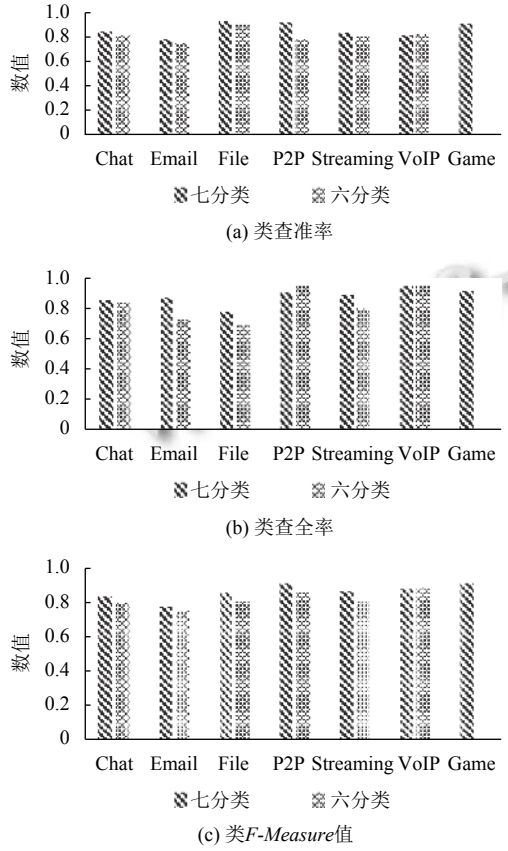


图 6 识别模型的类查准率、类查全率、类 *F-Measure* 值

在实验结果中我们也发现由于 Email 不存在大量字段负载信息, 所以在表征学习模型下的识别效果仍不是很理想. 在之后的流量分类实验中针对这类存在明显端口特征的流量, 我们可以集成使用基于端口的流量识别方法以提高该类流量的分类效果. 尽管如此, 扩充后的数据集对模型 Email 的识别率依旧有着不小的提升, 可以看出本文提出的基于端口映射的数据集扩充是成功的.

5 结束语

本文在传统流量分类的基础上尝试将表征学习的方法应用于游戏流量的识别研究, 并通过采集各类游戏流量, 同时利用通信日志文件中建立的各类游戏与进程端口的映射关系, 基于该映射关系对游戏流量进行过滤标记, 大幅提高游戏流量标注的工作效率, 降低扩展公开数据集的专业难度; 采用深度学习中的表征学习模型, 这种模型不需要预先提取流量特征, 而是直接把原始流量视为数据输入, 让表征学习模型自动学习流量特征并执行分类, 成功避免传统机器学习算法中流量标注以及流量分类模型对特征选择的依赖, 并针对不同维度的输入数据对识别模型的影响进行了研究, 也解决了网络游戏流量数据集匮乏的问题. 在后续的研究中, 将从以下两个方面进行改进: 一是将表征学习与机器学习算法相融合, 在减少学习模型对特征提取依赖的同时提高模型的识别效果; 二是结合无监督学习方法, 提高模型在面对未知流量情况下的识别能力.

参考文献

1 汪立东, 钱丽萍, 王大伟, 等. 网络流量分类方法与实践. 北京: 人民邮电出版社, 2013.
2 Wang W, Zhu M, Wang JL, *et al*. End-to-end encrypted traffic classification with one-dimensional convolution neural networks. 2017 IEEE International Conference on Intelligence and Security Informatics (ISI). Beijing: IEEE, 2017. 43–48.
3 王勇, 周慧怡, 俸皓, 等. 基于深度卷积神经网络的网络流量分类方法. 通信学报, 2018, 39(1): 14–23. [doi: 10.11959/j.issn.1000-436x.2018018]
4 Gao N, Gao L, Gao QL, *et al*. An intrusion detection model based on deep belief networks. 2014 Second International Conference on Advanced Cloud and Big Data. Huangshan: IEEE, 2014. 247–252.
5 Javaid A, Niyaz Q, Sun WQ, *et al*. A deep learning approach for network intrusion detection system. Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS). New York: ACM, 2016. 21–26.
6 Lotfollahi M, Siavoshani MJ, Zade RSH, *et al*. Deep Packet: A novel approach for encrypted traffic classification using deep learning. Soft Computing, 2020, 24(3): 1999–2012. [doi: 10.1007/s00500-019-04030-2]
7 李芳馨, 刘嘉勇. 网络数据流还原重组技术研究. 通信技术, 2011, 44(7): 113–114, 117. [doi: 10.3969/j.issn.1002-0802.2011.07.041]
8 陈雪娇, 王攀, 俞家辉. 基于卷积神经网络的加密流量识别

方法. 南京邮电大学学报 (自然科学版), 2018, 38(6): 36–41.
9 冯文博, 洪征, 吴礼发, 等. 基于卷积神经网络的应用层协议识别方法. 计算机应用, 2019, 39(12): 3615–3621.
10 Wang ZY. The applications of deep learning on traffic identification. Proceedings of the Black Hat USA 2015. Las Vegas, 2015.
11 Draper-Gil G, Lashkari H, Mamun MSI, *et al*. Characterization of encrypted and VPN traffic using time-related features. Proceedings of the 2nd International Conference on Information Systems Security and Privacy. Rome, 2016. 407–414.
12 LeCun Y, Jackel LD, Bottou L, *et al*. Learning algorithms for classification: A comparison on handwritten digit recognition. In: Oh JH, Cho S, eds. Neural Networks: The Statistical Mechanics Perspective. Singapore: World Scientific, 1995. 261–276.
13 潘吴斌, 程光, 郭晓军, 等. 网络加密流量识别研究综述及展望. 通信学报, 2016, 37(9): 154–167. [doi: 10.11959/j.issn.1000-436x.2016187]
14 Abadi M, Agarwal A, Barham P, *et al*. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. http://download.tensorflow.org/paper/whitepaper2015.pdf. (2016-03-16).