

基于卷积神经网络的未知协议识别方法

张路煜¹, 廖 鹏¹, 赵俊峰², 郭 靛¹

(1 南瑞集团公司, 江苏 南京 210000; 2 国家电网江苏省电力公司, 江苏 南京 210000)

摘 要: 提出一种基于卷积神经网络深度学习的网络流协议识别方法, 利用网络流数据与图像的相似性, 绕过流量特征值选择和提取的工作, 直接将网络流数据作为卷积神经网络的输入, 训练网络流协议识别模型, 实现网络流协议识别功能. 构建了一个含三层卷积层的卷积神经网络, 经实验生成网络协议识别模型, 对底层网络协议识别率达 93.33%. 该方法具备学习和扩展能力, 可扩展应用到对包含恶意代码网络流量的识别和对应用程序流量的识别.

关键词: 深度学习; 机器学习; 卷积神经网络; 未知协议识别

中图分类号: TP183, TP393

文献标识码: A

文章编号: 1000—7180(2018)07—0106—03

A Method of Unknown Protocol Recognition Based on Convolution Neural Network

ZHANG Lu-yu¹, LIAO Peng¹, ZHAO Jun-feng², GUO Liang¹

(1 NARI Group Corporation, Nanjing 210000, China;

2 State Grid Jiangsu Electric Power Company, Nanjing 210000, China)

Abstract: In this paper, a network flow protocol recognition method based on convolution neural network depth learning is proposed. Taking advantage of the similarity between network flow data and image, the network flow data is directly used as convolution neural network input, without the work of selecting and extracting the feature of network flow. Using supervise learning to train network flow protocol identification model, network flow protocol identification is realized. In this paper, a convolution neural network with three convolution layers is constructed. The network protocol recognition model is generated experimentally. The recognition rate of the underlying network protocol is 93.33%. The method has the ability to learn and extend. It can be scalable to the identification of traffic containing malicious code and identification of application traffic.

Key words: deep learning; machine learning; convolutional neural network; unknown protocol identification

1 引言

网络协议是计算机网络进行数据交换而建立的规则、标准或约定的集合. 对网络协议的识别和分析是网络安全的基础, 对网络监管、异常检测和维护网络安全具有重要意义. 传统的流量识别方法主要有基于端口的流量识别、基于深度包检测(DPI)的流量识别和基于机器学习方法的流量识别^[1-3].

深度学习的概念源于人工神经网络的研究. 含多隐层的多层感知器就是一种深度学习结构. 深度学习通过组合低层特征形成更加抽象的高层表示属

性类别或特征, 以发现数据的分布式特征表示. 深度学习是机器学习研究中的一个新的领域, 其动机在于建立、模拟人脑进行分析学习的神经网络, 它模仿人脑的机制来解释数据, 主要运用在自然语言处理、图像分类任务、声音识别等领域上^[2-5].

本文利用网络流数据与图像的相似性, 结合深度学习的卷积神经网络, 形成了一种基于卷积神经网络深度学习的网络流协议识别方法. 识别的过程将网络流数据图像化并打上协议标签作为神经网络的训练集, 从而绕过流量特征值选择和提取的工作, 直接将网络流数据作为卷积神经网络的输入, 进行

监督学习,训练网络流协议识别模型,实现未知网络流协议识别功能. 工作流程见图 1.

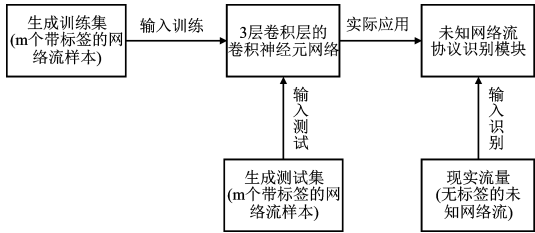


图 1 基于卷积神经网络的未知协议识别流程示意图

2 未知协议识别原理

基于机器学习的网络流量协议识别^[1]需要抽取并选择合适的数据包特征值,如数据包头信息和统计信息等. 数据包特征抽取与选择的好坏,依赖于专家的判断,决定了机器学习识别能力的优劣. 网络流量数据包基本的传输形式为二进制比特流,而灰度图的原始存储方式也为二进制比特流,在对网络流数据包协议的识别过程中,两者存在相似性,可将网络流数据包当作灰度图来进行处理. 利用成熟的卷积神经网络图像识别技术^[6-10],避免了显示的特征抽取,而隐式地从训练数据中进行深度学习,自动得到不同网络流量协议的特征模型,实现对未知网络流量协议的识别. 本文利用卷积神经网络,直接抽取网络流量数据包前 8 192 个二进制比特位,作为一个 1 024 像素点的灰度图作为卷积神经网络的输入,避免了传统基于机器学习的网络流量协议识别技术特征提取和分类过程中数据重建的复杂度.

卷积神经网络通过局部感知和权值共享对原始流量灰度图进行卷积,将网络流量信息的特征提取出来,再用池化操作对特征值进行降维,经多层提取,最终抽取对应网络流信息的多项抽象特征. 本文卷积神经网络采用三层卷积层结构,各层卷积核个数分别为 32,64 和 128,最终提取了网络流量灰度图的 128 项抽象特征,最后用多分类 Softmax 分类器实现网络流量的归类.

3 生成训练集

3.1 图像化

卷积神经网络常用于图像识别领域,为使流量样本成为卷积神经网络可识别的输入数据,首先要将流量样本图像化:提取流量样本前 $8 \times 1024 = 8\,192$ 个比特位,若样本不足 8 192 位,则补 0,取每 8 个比特的数值为一个元素,然后除以 256,使各元素

的值在 $[0,1)$ 区间内,形成 1 024 位的数组 X ,作为原始特征数组,储存至训练集对应一张 1 024 像素点的灰度图.

3.2 标签设置

经图像化处理后,根据样本所属协议类别对样本进行标签设置:根据样本所属协议类别对样本进行标签设置:设置一个含 n (n 为待识别协议种类的数量,正整数)个元素的数组,其第 n_i 个元素值为 1,其余值均为 0,以此表示该样本为第 n_i 种协议,作为标签数组 Y' ,例如: $n=10, n_i=6$ 时,数组 $[0,0,0,0,0,1,0,0,0,0]$ 表示 10 种待识别协议中的第 6 种协议,然后将标签与样本一一对应储存至训练集;重复图像化与标签设置的操作,卷积神经网络的训练集即生成完成. 同理,可生成测试集. 本文选取了 3 类协议进行识别,故 n 取 3,包含 $[1,0,0], [0,1,0], [0,0,1]$ 三种标签,分别对应 TCP,UDP 和其他协议.

4 构建卷积神经网络

本文构建了一个含三层卷积层的卷积神经网络,如图 2 所示:

输入层,取流量的原始特征数组作为输入,输入是一个 1024 位的数组 X ;

第一个卷积层,将 1 024 位的原始特征数组 X 视为 32×32 的矩阵,用 32 组单通道 5×5 卷积核去卷积输入的原始特征数组,然后加上偏置项,用 ReLU 函数激活得到卷积层 C_1 ;

第二个卷积层,用 64 组 32 通道的 5×5 卷积核去卷积 S_2 ,然后加上偏置项,用 ReLU 函数激活得到卷积层 C_3 ;

第三个卷积层,用 128 组 64 通道的 5×5 卷积核去卷积 S_4 ,然后加上偏置项,用 ReLU 函数激活得到卷积层 C_5 ;

全连接层,将 S_6 的 128 个 4×4 特征图视为 32×32 的矩阵,用 1024 组单通道 32×32 神经元与 S_6 全连接. 就像经典的神经网络一样,计算输入向量和权重向量之间的点积,再加上一个偏置项,然后将其输入给 ReLU 激活函数产生相应的输出 F_7 ;

输出层,是多分类 Softmax 分类器;用 n 组单通道 1 024 神经元与 F_7 全连接. 转化为 n 个 1×1 的特征图,即一个 n 维的特征向量,再用 Softmax 函数计算出 n 维预测数组 Y ,为卷积神经网络输出,对应样本协议识别的预测概率;

本文实验识别 3 类网络协议,即 n 值取 3.

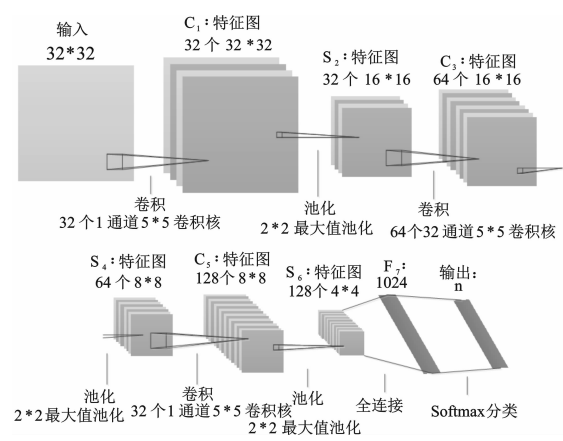


图 2 卷积神经网络示意图

5 选取训练函数

为使卷积神经网络能收敛,选取损失函数为“交叉熵”,其式为:

$$H_{Y'}(Y) = - \sum_i^n y'_i \log(y_i)$$
 (1)

式中, $H_{Y'}(Y)$ 损失值越小,则模型越逼近标签的真实结果.

训练时,将 m 个样本以每批 100 个作为输入,训练次数为 $m \bmod 100$,采用梯度下降算法以 0.0001 的学习速率最小化交叉熵,使损失值不断降低的方向更新,训练完成后保存神经网络参数.最后,将带标签的测试集数据输入保存的神经网络中,记录协议识别的准确率.

本文实验训练集数据规模为 3 000,即 m 值取 3 000.

6 实验和分析

根据上述步骤,本文选取了 3 000 条网络流数据作为训练集,600 条网络流数据作为测试集,包含 TCP,UDP 和其他协议三种类型协议,具体如下表:

表 1 TCP,UDP 和其它协议识别率

	训练集/条	测试集/条	识别条数/条	准确率/%
TCP	2186	82	82	100
UDP	603	377	347	92.04
其他	211	141	131	92.91
总计	3000	600	560	93.33

结果可见,对 TCP,UDP 和其他协议这三这协议的识别率均在 90% 以上,总体协议识别率为 93.33%.

7 结束语

本文将卷积神经网络技术应用到网络流量协议识别领域,提出了一种基于卷积神经网络的未知协议识别方法.该方法只需提供待识别网络流量协议样本用于对卷积神经网络的训练,能够自动提取到有利于分类任务的特征,具备学习和扩展能力,能够用于对新出现未知协议的识别,只需提供对应的样流量样本,可扩展应用到对包含恶意代码网络流量的识别和对应用程序流量的识别.

参考文献:

[1] 彭立志. 基于机器学习的流量识别关键技术研究[D]. 哈尔滨:哈尔滨工业大学, 2015.

[2] 马衍庆. 基于机器学习的网络流量识别方法与实现[D]. 济南:山东大学, 2014.

[3] Li X, Qi F, Xu D, et al. An Internet Traffic Classification Method Based on Semi-Supervised Support Vector Machine[C]. Japan: IEEE International Conference on Communications, 2011.

[4] 马衍庆. 基于机器学习的网络流量识别方法与实现[D]. 山东:山东大学, 2014.

[5] 常亮,邓小明,周明全,武仲科,袁野,杨硕,王宏安. 图像理解中的卷积神经网络[J]. 自动化学报, 2016, 42(09):1300-1312.

[6] 王飞飞. 基于改进卷积神经网络算法的研究与应用[D]. 南京邮电大学, 2016.

[7] 陈鸿翔. 基于卷积神经网络的图像语义分割[D]. 浙江大学, 2016.

[8] 张文达,许悦雷,倪嘉成,马时平,史鹤欢. 基于多尺度分块卷积神经网络的图像目标识别算法[J]. 计算机应用, 2016, 36(04):1033-1038.

[9] 尹颢,闫磊. 基于深度卷积神经网络的图像目标检测[J]. 工业控制计算机, 2017, 30(04):96-97+100.

[10] 王振,高茂庭. 基于卷积神经网络的图像识别算法设计与实现[J]. 现代计算机(专业版), 2015, (20): 61-66.

作者简介:

张路煜 男, (1989-), 硕士, 工程师. 研究方向为信息安全. E-mail: zhangluyu@sgepri. sgcc. com. cn.

廖 鹏 男, (1985-), 硕士, 工程师. 研究方向为信息安全.

赵俊峰 男, (1974-), 硕士, 高级工程师. 研究方向为信息化管理.