文章编号：2096-1472(2020)-10-08-03
DOI:10.19644/j.cnki.issn2096-1472.2020.10.003

基于非对称分解卷积的网络安全检测

冯仁君，吴 吉，王震宇，景栋盛

(国网江苏省电力有限公司苏州供电分公司，江苏 苏州 215004)
✉frj1989@126.com; 13862159678@163.com; gcxy6@hotmail.com; jds19810119@163.com

摘 要：近年来，网络安全检测已经取得了很大的进步。然而，网络迅速的发展、流量分布的变化和数据样本中的噪声等问题都对现有方法提出了很大的挑战。针对此，提出了基于非对称分解卷积的网络安全检测(Network Security Detection Based on Asymmetric Decomposed Convolution，ADC-NSD)方法。ADC-NSD方法根据对网络连接数据的训练与学习，生成区别常态与危险状态的安全检测模型，通过对卷积神经网络中的卷积核进行分解，完成对数据进行解析和检测。最后，以KDDCUP99为测试数据集，将ADC-NSD方法与其他机器学习方法进行比较。实验结果表明，ADC-NSD方法能有效地解决网络安全检测问题，总体精确率为98.72%，准确率为99.92%，召回率为94.61%，F1值为97.19%。

关键词：网络安全；安全检测；卷积神经网络；非对称分解卷积
中图分类号：TP18 **文献标识码**：A

Network Security Detection based on Asymmetric Decomposed Convolution

FENG Renjun, WU Ji, WANG Zhenyu, JING Dongsheng

(*Suzhou Power Supply Branch*, *State Grid Jiangsu Electric Power Limited Company*, *Suzhou* 215004, *China*)
✉frj1989@126.com; 13862159678@163.com; gcxy6@hotmail.com; jds19810119@163.com

Abstract: In recent years, network security detection has made great progress. However, the rapid development of communication networks, changes of the traffic distribution and the noise in data samples all pose great challenges to the existing network security detection methods. To solve this problem, an approach referred as Network Security Detection based on Asymmetric Decomposition Convolution (ADC-NSD) is proposed. ADC-NSD generates a security detection model to distinguish normal state and dangerous state according to the training and learning of network connection data, and then analyzes and checks the data through decomposing the convolution kernel of the convolution neural network (CNN). Finally, using KDDCUP99 (KDD: Knowledge Discovery and Data Mining) dataset as testing dataset, ADC-NSD is measured against other machine learning algorithms. The results show that ADC-NSD could be well applied to network security detection. The overall accuracy rating is 98.72%, precision rate being 99.92%, recall rate being 94.61% and F1 score being 97.19%.

Keywords: network security; security detection; convolution neural network; asymmetric decomposed convolution

1 引言(Introduction)

网络技术的发展为人们的生产与生活提供了很多的便捷，但也始终伴随着网络安全问题。这一问题得到了人们的广泛关注，成为一个研究热点。

虽然目前针对网络安全检测的研究工作已经取得了很大的进步，但也存在大量问题，如需要人工标注数据集、学习效率差、方法实用性弱等。深度学习通过构建多层的神经网络对大量原始数据进行表征学习，具备较强的构建模型及推理能力，广泛应用于医学研究[1]、机器人控制[2]、图像识别[3]和自然语言处理[4]等领域。

因此，如何充分发挥深度学习的优势以提高安全检测模型的分析和预测能力，显得尤为有意义。针对此，本文将非对称分解卷积应用到网络安全检测模型的构建与提取特征之中。首先，通过预处理模块生成满足需求的数据用于训练；然后将传统的卷积神经网络中的卷积核进行非对称分解；最后，将改进的模型应用于安全检测。本文在KDDCUP99数据

基金项目：江苏省高等学校自然科学研究项目重大项目(17KJA520004)。

集[5]上进行了测试。结果表明，本文方法在网络安全检测过程中表现出了高效性和可靠性。

2 相关研究(Related work)

2.1 深度学习

深度学习通过构建多层的神经网络对原始的大量的数据进行表征学习，从而提取出原始数据中的抽象原始特征。深度学习拥有较强的构建模型能力及推理能力，可以通过一系列的函数计算模拟计算机在神经元里的激活活动，从而学习数据特征。其中，卷积神经网络(Convolutional Neural Network，CNN)是监督学习网络的代表。深度神经网络通过训练大量带标签数据，利用反向传播算法从而能够提取数据特征。

与此同时，深度学习也应用到了大量无监督学习算法、半监督学习算法及强化学习算法，如生成对抗网络(Generative Adverse Network，GAN)[6]，深度强化学习(Deep Reinforcement Learning，DRL)[7]和深度置信网络(Deep Belief Network，DBN)[8]等。这些深度学习算法通过对输入数据的处理和分析，不需要数据标签即可学习特征。

深度学习强大的特征提取能力使得其在图像处理、文本识别、人脸检测、自动驾驶等多个领域取得了重大的突破。

2.2 网络安全检测

网络安全检测是一种主动的网络安全实时保护。它通过检测网络过程中的内部和外部攻击及失误操作，能够有效地弥补网络防火墙的缺陷，对防火墙进行高效地补充和提高。网络安全检测一般有异常检测和误用检测两种。异常检测的检测率偏低，但是不需要检测的先验知识，因此能够检测到突变的，以及未曾有过的入侵行为，是目前的主流研究方向；误用检测利用已有的入侵检测行为，无法对突变的或者全新的攻击行为进行识别。

网络入侵检测是一种常见的网络安全检测方法。网络入侵检测的目的在于能够主动防御从内部或者网络来的攻击，通过相应的手段来及时准确的预测出会产生的入侵行为，并做出相应的防御措施，从而减少安全工作人员的维护压力并提高系统的安全防护能力。近年来，很多研究人员对网络安全检测方法进行了深入的学习和探索。Shone等人[9]利用非对称深度自动编码器对入侵检测进行建立深度学习分类模型；钱亚冠等人[10]利用毒性攻击，改变支持向量机(Support Vector Machine，SVM)机器学习过程，降低了对攻击流量的识别率；潘建国等人[11]用支持向量机进行预训练获取入侵检测判定规则并应用于物联网中每个节点；石乐义等人[12]根据相关信息熵的算法进行特征选择，再利用深度学习方法学习数据特征；吴亚丽等人[13]利用稀疏降噪自编码网络自动提取入侵数据的最优特征，提高了模型的鲁棒性。

3 方法设计(Method design)

本文提出一种非对称分解卷积的网络安全检测方法。在最近的入侵检测领域中，传统统计学习方法与机器学习方法相辅相成，包括马尔科夫链、决策树、决策森林、贝叶斯分类及隐马尔可夫等。但总体而言，需要计算的数据量庞大、实时性低，因此难以满足网络安全检测的要求。本文提出的基于非对称分解卷积的网络安全检测方法，将不同形式的对称卷积因式分解为非对称结构并应用于网络安全检测。

3.1 KDDCUP99数据集

本文使用KDDCUP99数据集[5]作为训练和测试数据集。该数据集由大量的异常连接数据与正常连接数据构成，是用于评估网络安全检测模型的一套标准数据集。很多入侵检测的研究工作在此数据集进行测试。

KDDCUP99的训练集由不同的攻击类型与正常类型构成，测试集中包含17中未知攻击。样本共分为五大类，分别为：探测攻击、拒绝服务攻击、远程对本地攻击、用户对管理员攻击和正常。KDDCUP99的每个样本包括了41个固定特征属性与一个类标识符。这些固定特征属性里，32个特征属性为连续的，9个特征属性为离散类型。例如其中的一条连接记录如下：

0，udp，http，sf，289，2478，0，0，0，0，0，1，0，0，0，0，0，0，0，0，0，0，6，6，0，0，0，0，1，0，0，29，255，1，0，0.02，0.03，0，0，0，0，normal

该记录加上最后的标签，共有42个特征属性。其中第1—10位为基本特征，第11—22位为内容特征，第23—41位为流量统计特征。

3.2 数据预处理

原始的KDDCUP99数据集为异构数据集，因而需要对数据进行预处理，以便于对攻击类型分类。其中，对离散型和连续性数据进行不同的处理方法与归一化。离散的特征变量采用独热编码(one-hot)处理，而英文单词(如数据集的协议与服务类型)的编码则使用数值代替。协议类型和编码如表1所示。

表1 协议类型的独热编码

Tab.1 One-hot coding of protocol

协议类型	编码
ICMP	1000
TCP	0100
UDP	0010
其他协议	0001

经过处理后的数据集虽然已经可以进行训练，但为了加快模型的收敛速度与学习能力，需要减少样本的数值之差。将样本数据进行归一化，使样本数据全部在[0,1]区间范围内。

首先将特征数值进行标准化。标准化公式为：

$$n' = \frac{n - AVG}{STAD} \tag{1}$$

其中，n为原始特征值，n'为标准化后的特征值，AVG为特征值的平均值，STAD为平均绝对误差。

然后，对标准化后的数进行归一化处理，采用公式如下：

$$n'' = \frac{n' - n_{\min}}{n_{\max} - n_{\min}} \tag{2}$$

$$n_{\min} = \min\{n'\} \tag{3}$$

$$n_{\max} = \max\{n'\} \tag{4}$$

其中，n'为标准化之后的特征值。

3.3 模型结构与分析

基于非对称分解卷积的网络安全检测采用了非对称卷积

模型架构。本文将较大的对称卷积分解为多个小的非对称卷积，以提高训练精度与速度。使用非对称分解卷积的网络安全检测有三个优点：(1)通过非对称卷积，增加了网络的非线性；(2)由于卷积被分解，由此可以降低过拟合的概率；(3)加速网络计算。

基于非对称分解卷积的网络安全检测方法如算法1所示。

算法1基于非对称分解卷积的网络安全检测 (Network Security Detection Based on Asymmetric Decomposed Convolution，ADC-NSD)

输入：KDDCUP99网络连接数据D

输出：训练完成的网络参数θ'

1. 初始化：神经网络结构参数θ初始化为随机较小数值
2. 初始化：抽样的样本数量minibatch为32
3. Repeat(对每次神经网络值传递)
4. 对非字符值数据进行独热操作
5. 将数据标准化
6. 将数据进行min-max归一化
7. 在数据集中抽取minibatch样本数据S
8. S输入全对称卷积层并激活输出S'
9. S'输入非对称分解卷积层并激活池化输出S''
10. S''输入全连接层并输出结果
11. 以Adam梯度更新反向传播
12. 更新θ
13. Until到达预期训练次数
14. 返回训练完成后的网络参数θ'

卷积神经网络模型是由卷积、激活和池化等一系列的线性，以及非线性的变化模块所构成。数据集经过数次的变换可以得到更为抽象的特征，从而提高泛化能力。从某些方面来说，模型深度越深，层次越多，特征的可辨别能力也随之提高，从而性能也相应变好。然而，一般比较大的空间滤波器的卷积在计算方面需要的成本极其昂贵，因此本文将全对称卷积进行因式分解，将原本对称高维卷积分解为非对称低维度卷积，以节省内存和计算量，从而快速提高网络性能。本文3×3的卷积分解示意图如图1所示。

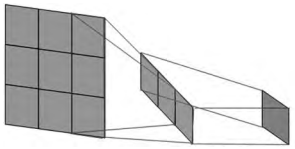

图1 将3×3的卷积分解为1×3的卷积与3×1的卷积

Fig.1 Decomposing the convolution of 3×3 into the convolution of 1×3 and the convolution of 3×1

通过将对称卷积进行结构拆分为非对称卷积，一方面可以节约大量参数，另一方面也加快了运算速度，从而能够分析更多样的空间特征并提高特征的多元化。具体的卷积神经网络示意图如图2所示。

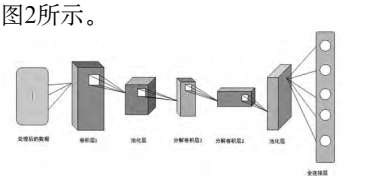

图2 非对称分解卷积

Fig.2 Asymmetric decomposed convolution

在图2中，第一层卷积层为全对称卷积层，采用5×5的卷积核，步长为2；第二层与第三层为分解卷积层，分别为1×3的卷积核与3×1的卷积核，最后全连接层输出数据。该神经网络结构节约了训练成本，由于激活层的输出是高度相关的，通过分解使其在聚合前能够有效降维，从而加快了训练速度，提升了性能。

4 实验结果和分析(Experimental results and analysis)

实验模型选取0.0005为学习率，优化算法为Adam。本文选取了传统的统计学习方法与机器学习方法与本文模型(ADC-NSD)进行对比，其中包括支持向量机(SVM)、决策树(Decision Tree，DT)、K近邻算法(K-Nearest Neighbor，KNN)、卷积神经网络(CNN)。分别测试了不同方法的精确率、准确率、召回率与F1值，如表2所示。

表2 不同机器学习的方法对比

Tab.2 Comparison of different machine learning methods

方法	精确率(%)	准确率(%)	召回率(%)	F1值(%)
SVM	82.42	74.03	80.31	77.04
DT	91.94	98.92	86.06	92.04
KNN	85.22	89.61	75.65	82.04
CNN	94.78	90.32	94.23	92.23
ADC-NSD	98.72	99.92	94.61	97.19

从表2可以看得到不同的机器学习方法与本文的模型对于KDDCUP99数据集的训练效果。传统的统计学习方法比如支持向量机、决策树和K近邻算法模型的测试数据比较低，然而神经网络算法能获得比较好的效果。通过对比传统深度卷积网络和非对称分解卷积，可以看出非对称分解卷积可以取得比较好的实验效果。本文的非对称分解卷积在所有模型里表现最好，展现了本模型的可靠性和高效性。

5 结论(Conclusion)

相比传统的统计学习方法，深度卷积模型在处理网络安全检测过程中具有更独特的优势。对于大规模数据，深度卷积网络能够更精细地提取到数据特征。

本文在深度卷积网络的基础上提出了非对称分解卷积，通过对完全对称卷积进行分解为非对称卷积，快速地提高了网络性能，对于空间特征处理的更加细腻，从而对大规模的网络安全检测数据的特征提取更加智能。最后，本文以KDDCUP99数据集为测试数据，比较了多种机器学习方法。实验结果表明，本文提出的非对称分解卷积具有较好的安全检测效果。

参考文献(References)

[1] Ravi D, Wong C, Deligianni F, et al. Deep Learning for Health Informatics[J]. IEEE Journal of Biomedical and Health Informatics, 2017, 21(1):4-21.

[2] Shvets A, Rakhlin A, Kalinin A, et al. Automatic Instrument Segmentation in Robot-Assisted Surgery using Deep Learning[C]. International Conference on Machine Learning and Applications, 2018: 624-628.

[3] 圣文顺,孙艳文.卷积神经网络在图像识别中的应用[J].软件工程,2019,22(2):13-16.

[4] Klein G, Kim Y, Deng Y, et al. OpenNMT: Open-Source Toolkit for Neural Machine Translation[C]. Meeting of the

Association for Computational Linguistics, 2017: 67−72.
[5] 郭成华.基于KDDCUP99数据集的入侵检测系统的设计与实现[J].网络安全技术与应用,2017(12):60−63.
[6] Karras T, Laine S, Aila T, et al. A Style−based Generator Architecture for Generative Adversarial Networks[C]. Computer Vision and Pattern Recognition, 2019: 4401−4410.
[7] 李广创,程良伦.基于深度强化学习的机械臂避障路径规划研究[J].软件工程,2019,22(3):12−15.
[8] 满忠昂,刘纪敏,孙宗锟.基于局部二值模式与深度置信网络的人脸识别[J].软件工程,2020,23(5):13−16.
[9] Tang A, Mhamdi L, Mclernon D, et al. Deep learning approach for Network Intrusion Detection in Software Defined Networking[C]. International Conference on Wireless Networks, 2016: 258−263.
[10] 钱亚冠,卢红波,纪守领,等.一种针对基于SVM入侵检测系统的毒性攻击方法[J].电子学报,2019,47(1):59−65.
[11] 潘建国,李豪.基于实用拜占庭容错的物联网入侵检测方法[J].计算机应用,2019,39(6):1742−1746.
[12] 石乐义,朱红强,刘祎豪,等.基于相关信息熵和CNN−BiLSTM的工业控制系统入侵检测[J].计算机研究与发展,2019,56(11):2330−2338.
[13] 吴亚丽,李国婷,付玉龙,等.基于自适应鲁棒性的入侵检测模型[J].控制与决策,2019(11):2330−2336.

作者简介：
冯仁君(1989−)，男，硕士，工程师.研究领域：软件智能化，软件项目管理，信息安全.
吴　吉(1981−)，女，学士，工程师.研究领域：信息安全，计算机应用.
王震宇(1981−)，男，硕士，工程师.研究领域：信息安全，计算机应用.
景栋盛(1981−)，男，硕士，高级工程师.研究领域：信息安全，软件智能化.

(上接第14页)

实验结果表明，本文方法显著优于原HDFS和Har方法，在重复读取时由于命中缓存，读取时间进一步缩短，且读取时间不随文件数量的增大而提高，在文件数量越多时优势越明显。原因在于随着HDFS存储的小文件增多，HDFS检索元数据的性能下降，而本文方法由于使用HashMap保存索引信息，在通过文件名查找索引信息时直接获取，时间复杂度为O(1)。在读取重复文件时，由于首次读取已将索引信息添加到内存中的HashMap对象中，再次读取时直接从内存中的HashMap中查询索引信息，读取效率进一步提高。

6 结论(Conclusion)

本文通过分析Hadoop存储模块HDFS的架构并引出其存储海量小文件的不足，进而提出一种基于分类合并思想的改进方案。即将小文件按照文件类型(扩展名)分类合并为大文件，并为每一类型的合并文件添加索引，将索引保存在HashMap中，最终写入到索引文件。同时为了进一步提高读取文件的速度，本方案设置了基于HashMap的缓存机制，即在内存中缓存已读文件的索引信息。实验表明，本方案在写入速度、NameNode内存占用，以及读取速度上均显著优于原HDFS。

参考文献(References)

[1] 冯贵兰,李正楠,周文刚.大数据分析技术在网络领域中的研究综述[J].计算机科学,2019,46(06):1−20.
[2] 王佳隽,吕智慧,吴杰,等.云计算技术发展分析及其应用探讨[J].计算机工程与设计,2010,31(20):4404−4409.
[3] 夏靖波,韦泽鲲,付凯,等.云计算中Hadoop技术研究与应用综述[J].计算机科学,2016,43(11):6−11;48.
[4] 郭建华,杨洪斌,陈圣波.基于HDFS的海量视频数据重分布算法[J].计算机科学,2016,43(S1):480−484.
[5] 李建江,崔健,王聃,等.MapReduce并行编程模型研究综述[J].电子学报,2011,39(11):2635−2642.
[6] 郑通,郭卫斌,范贵生.HDFS中海量小文件合并与预取优化方法的研究[J].计算机科学,2017,44(S2):516−519.
[7] 李三淼,李龙澍.Hadoop中处理小文件的四种方法的性能分析[J].计算机工程与应用,2016,52(09):44−49.
[8] 刘斌.Hadoop小文件编程处理的性能优化[J].工业控制计算机,2018,31(12):47−48.
[9] 谭台哲,向云鹏.Hadoop平台下海量图像处理实现[J].计算机工程与设计,2017,38(04):976−980.
[10] 段隆振,洪新利,邱桃荣.基于MapFile的HDFS小文件存取优化[J].南昌大学学报(工科版),2017,39(02):175−178.
[11] 游小容,曹晟.海量教育资源中小文件的存储研究[J].计算机科学,2015,42(10):76−80.
[12] 赵晓永,杨扬,孙莉莉,等.基于Hadoop的海量MP3文件存储架构[J].计算机应用,2012,32(06):1724−1726.
[13] 赵洋.淘宝TFS深度剖析[J].数字化用户,2013,19(03):58−59.
[14] Bo Dong, Qinghua Zheng, Feng Tian, et al. An optimized approach for storing and accessing small files on cloud storage[J]. Journal of Network and Computer Applications, 2012, 35(6): 1847−1862.

作者简介：
秦加伟(1995−)，男，硕士生.研究领域：大数据，大数据分析.
刘　辉(1979−)，男，硕士，副教授.研究领域：软件工程，信息系统.
方木云(1968−)，男，博士，教授.研究领域：软件工程，信息系统.