Speech Steganalysis Based on Multichannel Convolutional Neural Network*

LI Jing[3*], *DENG Wei*[2], *YANG Yong*[1,2], *ZHU Zuofu*[1], *JIA Weiwei*[1]

(1.*School of Information Engineering, Najing Audit Unversity, Nanjing Jiangsu* 211815, *China*;
2.*Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Nanjing Jiangsu* 210096, *China*;
3.*School of Information Engineering, Xuzhou College of Industrial Technology, Xuzhou Jiangsu* 221140, *China*)

Abstract: In this study, three commonly used speech steganography methods are analyzed. In the current network environment, the steganography technology in VoIP is a huge threat to communication monitoring. In recent years, neural network model has achieved remarkable performance in many subjects. An improved neural network is used to detect steganographic signals. The difference of STFT along time axis and frequency axis is constructed, and three channels are constructed as input data. Then an improved CNN network structure is used to capture deep-level features. This network structure introduces the concept structure to extract features of different scales on the same convolution layer, and uses global average pooling to replace the full connection layer. At the same time, it improves the generalization ability of the network. The experimental results show that the model achieves better detection results for the three steganography methods.

Key words: speech steganalysis, deep neural network, multichannel convolution
EEACC: 7210 **doi**: 10.3969/j.issn.1005-9490.2022.05.015

基于多通道卷积神经网络的语音隐写分析方法*

李 晶[3*], 邓 伟[2], 杨 勇[1,2], 朱作付[1], 贾伟伟[1]

(1.南京审计大学信息工程学院, 江苏 南京 211815; 2.东南大学水声信号处理教育部重点实验室, 江苏 南京 210096;
3.徐州工业职业技术学院信息工程学院, 江苏 徐州 221140)

摘 要: 本研究对三种常用的语音隐写方法进行了隐写分析。在目前的网络环境中, VoIP 中的隐写技术对通信监控是一个巨大的威胁。近年来, 神经网络模型在许多课题中都取得了显著的性能。利用一种改进的神经网络对多种隐写方法实现隐写信号的检测。构建短时傅立叶变换的沿时间轴、频率轴的差分以构建三通道作为输入数据, 然后利用一种改进的 CNN 网络结构进行深层次特征的捕获, 这种网络结构引入了 Inception 结构在同一卷积层上提取各种不同尺度的特征, 使用全局平均池化来代替全连接层, 在降低参数的同时提升了网络的泛化能力。实验结果表明, 该模型相较对比方法对于三种隐写方法均达到了较好的检测效果。

关键词: 语音隐写分析; 深度神经网络; 多通道卷积
中图分类号: TN912.34; TP183 **文献标识码**: A **文章编号**: 1005-9490(2022)05-1105-05

隐写术是一种古老的技术, 旨在以不可检测的方式将秘密信息嵌入图像[1]、文本[2]和声音[3]等载体中。相反, 隐写分析的目的是检测可疑载体中的隐藏信息。VoIP 通常是指在互联网上传输语音通信, 随着网络融合趋势的不断发展, VoIP 已成为互联网上流行的流媒体和通信服务。VoIP 的网络流量巨大且不断增长。同时, 由于 VoIP 的实时性和大规模特性, 通过篡改 VoIP 软件产生的语音流来实现非法隐蔽通信的情况非常普遍, 这些语音流可能被黑客、恐怖分子等恶意利用。因此, VoIP 中的隐写技术对通信监控是一个巨大的威胁。因此, 有必要研究 VoIP 隐写技术的隐写分析方法。

项目来源: 国家自然科学基金项目(52005267)
收稿日期: 2021-08-26 **修改日期**: 2021-10-22

最低有效位隐写算法(Least Significant Bit, LSB)[4],直接扩频隐写算法[3]和相位索引调制隐写算法[5]是传统音频隐写的主要方法。LSB 通过修改载体信号的最低有效位来嵌入秘密数据,直接扩频隐写方法通常是将秘密信息隐藏在复杂变换的变换域中,相位索引调制隐写方法是由 Chen[6]首先提出的,它通过修改语音编码过程中的量化向量来隐藏秘密数据,具有较少的附加失真和更高的隐藏能力[7-8]。这三种隐写方法都常用于在 VoIP 流中嵌入秘密信息。

近年来,神经网络成功应用于许多领域,在计算机视觉[9]、音频识别[10]、自然语言处理[11]等方面取得了突破,为了进一步提高各种系统的性能,已经开发了许多深度学习模型。Chen 等人[12]在利用神经网络进行各种音频隐写分析方面进行了一些显著的研究[13]。采用卷积神经网络在时域检测 LSB 隐写。Wang 等人[13]提出了一种卷积神经网络(Convolutional Neural Networks,CNN)模型在熵码域检测 MP3 隐写术。上述两种方案在不同领域的目标隐写术的隐写分析中均优于传统方法。同时,Lin[14]使用 2 层长短期记忆网络(Long Short-Term Memory,LSTM)在量化索引序列中捕获更多的相关性,提升了针对量化索引调制(Quantization Index Modulation,QIM)隐写术的检测效果。

本文利用一种改进的神经网络对多种隐写方法实现隐写信号的检测。首先对隐写信号和载体进行时频分析,构建短时傅里叶变换(Short Time Fourier Transform,STFT)时频特征图,以反映不同类型信号的频率随时间的变化。然后,构建 STFT 的沿时间轴、频率轴的差分以构建三通道输入数据。其次,提出了一种改进的 CNN 网络结构,提取输入图像各种尺度的有效特征对隐写信号进行检测。最后通过实验验证了本文提出的改进 CNN 网络的有效性。

本文的其余部分组织如下。首先详解了提出的多通道卷积神经网络结构的细节,然后给出和分析了实验结果,并对模型进行了讨论。最后,对本文做出总结。

1 传统卷积神经网络

卷积神经网络是对传统普通神经网络的改进。对于输入为图像的神经网络来说,图像的每一个像素都可以看做一个特征值。若采用全连接的网络连接方式,会造成参数过多的情况。卷积神经网络通过引入卷积层和池化层实现了局部特征提取和层次特征提取的功能,并使用网络权值参数共享机制降低了整个网络的参数数量。下面对卷积神经网络的结构进行概述。

卷积神经网络主要由输入层、卷积层、池化层、全连接层和输出层组成,如图 1 所示。

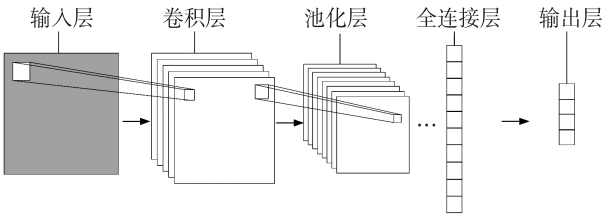


图 1 卷积神经网络结构原理图

卷积层是整个卷积神经网络的核心,其主要功能是通过卷积运算来提取图像特征。CNN 的卷积层通过利用图像的区域性、局部相关性,将传统网络的全连接改动为局部连接,从而降低网络的参数数量。局部连接的过程是通过卷积运算来实现的,卷积运算是是一种特殊的线性运算,其可由如下的数学公式来表示:

$$x_i^l = \sum_{j \in N_i} W_{ij}^l * x_j^{l-1} + b_i^l \tag{1}$$

式中:x_i^l 表示第 l 层的第 i 个特征图,N_i 表示前一层的输入特征图集合,$*$ 表示卷积运算,W_{ij}^l 表示第 l 层的第 i 个特征图与第 $l-1$ 层的第 j 个特征图连接的卷积核参数,b_i^l 表示偏置项。

池化操作也称为欠采样或者降采样操作,主要用来缩小特征图的尺寸进而减少计算量。池化层的出现可以加快网络的计算速度,并且在一定程度上解决平移鲁棒性、防止模型过拟合。然而,池化操作也会同时损失输入特征图的空间位置精度,所以池化操作是在计算量和空间位置精度之间的权衡。池化层之后接全连接层以构建网络最后的输出。

这里可以将隐写信号看成是短时平稳的,即对任一时刻 t,都可在该时刻附近小范围时段内对信号进行频谱分析。对一系列 t 值进行连续的频谱分析即可以得到隐写信号的二维频谱图。二维频谱图的每一列代表了一帧信号的 512 点 FFT,每一帧持续时间为 1.024 ms,帧重叠率为 0.5。考虑到载体信号 STFT 谱图的频率分布相对稳定和单一性,以及 CNN 强大的图像学习分类能力,这里为卷积神经网络设计了三通道的输入以提升模型准确率。三通道输入分别为:

(1)STFT 幅度谱;

(2)STFT 幅度谱沿时间轴的一阶差分谱;

(3)STFT 幅度谱沿频率轴的一阶差分谱。

在这里,对 STFT 幅度谱做差分可以提取隐写信号和载体信号频谱图的变化特性,比如一些图像

边缘特征等。差分谱作为辅助联合原始的 STFT 谱图一起构成了本章网络的多通道输入数据。

2　改进的多通道卷积神经网络

本文通过借鉴 GoogleNet、VGGNet 等诸多经典 CNN 网络结构的设计思想,提出了一种对传统卷积神经网络的改进方法。

如图 2 所示,展示了本章所采用的改进 CNN 的网络结构图。网络的输入数据是帧重叠率为 50% 的载体信号和隐写信号的 STFT 谱图、STFT 沿时间轴的一阶差分、STFT 沿频率轴的一阶差分构成的三通道图像,图像的尺寸为 257×129 pixel。

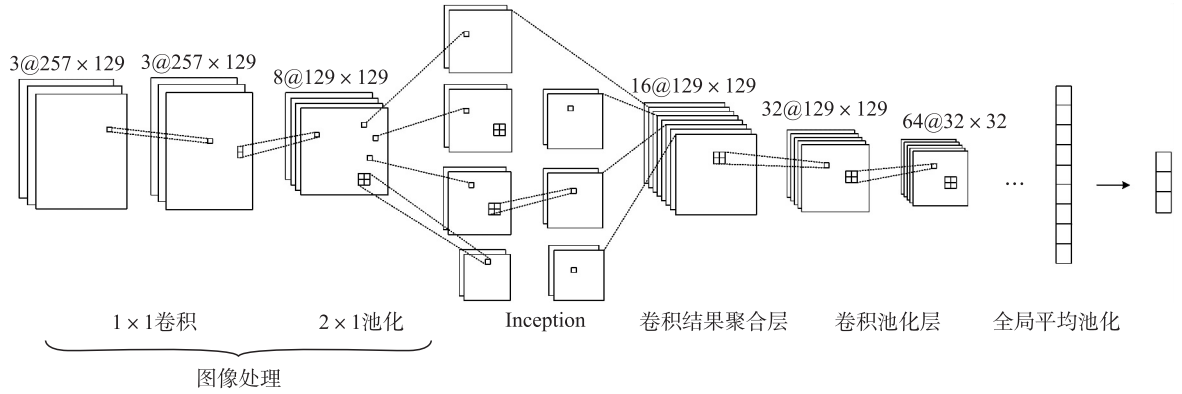


图 2　改进的多通道卷积神经网络结构原理图

在输入图像预处理上,由于输入的三通道图像均是长、宽比为 2:1 的图片,首先利用卷积核尺寸为 1×1 的卷积层来对输入数据进行非线性变换,然后使用尺寸为(2,1)的池化操作将前一层输入的矩形特征图整理成 1:1 的特征图以利于后续的操作。

在卷积核的构成上,引入 Inception 结构在同一层上使用多种尺度的卷积核提取网络各种层级的特征。此外由于不同卷积之间的特征并不进行交叉计算,可以在一定程度上较少网络的计算量。

在特征图通道数上,借鉴 VGG-Net 的思想,每经过一个池化层之后,使通道数目翻倍。最后,使用全局平均池化(Global Average Pooling,GAP)代替传统卷积神经网络中的全连接层,以减轻网络发生过拟合。全连接层即使采用的 dropout 操作来防止过拟合,也并不会减少网络的实际计算量。而 GAP 是对整个网络从结构上做正则化以防止过拟合。

3　实验分析

3.1　实验设置

本文的实验采用 TIMIT 语音数据集,其中的语音样本就作为实验的原始载体语音。对原始载体语音数据集中的每个样本,采用语音隐写算法嵌入秘密数据,这样就得到了隐写语音。将有嵌入秘密数据的样本定义为负类,将没有嵌入数据的样本定义为正类,这样就得到了完整的实验数据集。本实验采用 tensorflow 深度学习框架来搭建并训练网络。网络的模型参数如表 1 所示。

此网络采用 ReLU 激活函数[16],代表的是一种“修正线性单元”,可以在一定程度上避免梯度消失问题。此外,本实验使用 Adam 优化算法来进行训练。对于池化操作,采取补零的方法对其进行 padding。

表 1　网络模型参数

No	Layer	Kernel Size	Channels
1	Convoluaiton	(1,1)	3
2	Pooling	(2,1)	8
3-1	Convolution	(1,1)	8
3-2	Convolution	(1,1),(3,3)	8
3-3	Convolution	(1,1),(5,5)	8
3-4	Max Pooling--Convolution	(3,3),(1,1)	8
4	Concatenate	-	32
5	Convolution	(2,2)	48
6	Pooling	(2,2)	96
7	Convolution	(2,2)	128
8	Pooling	(3,3)	256
9	Global Average Pooling	-	-
10	Softmax	(1,*X*)	-

3.2　实验评估

本文对三种常用的隐写技术进行分析实验,评估本文提出的语音隐写分析方法在检测这三种常用的隐写方法生成的隐写信号时的性能。分别使用直接扩频隐写(Direct Sequence Spread Spectrum,DSSS)、最低有效位隐写(Least Significant Bit,LSB)以及相位索引调制隐写(Quantization Index Modulation,QIM)这三种隐写方法在原始载体语音中嵌入隐藏信息,得到和载体语音对应的隐写语音,从而载体语音和对应的隐写语音一同构成完成的数据集用于模型训练和测试。具体地,用 QIM 隐写方法对载体音频进行信息

嵌入,嵌入域为时域,量化步长为 0.001。LSB 算法只在数据点的最低一位嵌入信息;DSSS 嵌入强度为 0.001,嵌入段长可变。

语音片段的时长是影响隐写分析方法的一个重要因素。因此固定隐写方法的嵌入数据的比率,分析样本时长对隐写分析方法的影响。如表 2 所示的结果显示,随着样本时间长度的增加,精度也会增加。这一现象是因为较长的时间序列提供了更多对码字相关性的观察,因此,隐写语音和载体语音的之间的差异更明显,从而更容易分类。此外,当样本长度较小时,增加样本长度可以显著提高准确率。随着样本长度的增加,增加样本长度对模型准确率的提升逐渐减小。

表 2 针对三种隐写方法在不同样本时长下的检测准确率

样本时长/s	隐写方法	检测准确率/%
0.5	QIM	72.34
	DSSS	78.36
	LSB	70.12
1.0	QIM	76.71
	DSSS	82.26
	LSB	78.45
2.0	QIM	79.38
	DSSS	84.12
	LSB	81.23
3.0	QIM	81.89
	DSSS	85.49
	LSB	82.98

3.3 模型对比

此外,为了进一步探究本文提出模型的有效性,现将其常用的语音隐写分析方法进行对比实验,表 3 展示了不同模型在检测 QIM 隐写信号和载体信号时的性能,其中语音样本时长为 2 s,嵌入率与上述保持一致。

表 3 不同模型对比实验结果 单位:%

模型	不同信号类别检测准确率		平均检测准确率
	隐写信号	载体信号	
KNN	69.58	68.16	68.87
DNN	71.49	75.29	73.39
稀疏表示	72.53	76.69	74.61
SVM	79.49	77.93	78.71
CNN	82.36	77.70	80.03
改进 CNN	82.64	81.14	81.89

从表 3 可以看出,不管是对隐写信号还是原始载体信号,改进 CNN 的检测准确率均是最高,对隐写信号和载体信号分别达到了 82.64%和 81.14%,且整体平均识别率比支持向量机(Support Vector Machine,SVM)高出了 3%以上。传统的信号分类方法中 SVM 的性能最优,达到了 78.71%,比 K 紧邻(K Nearest Neighbors,KNN)、稀疏表示分别高了 10%、4%。分类器 KNN 在隐写信号中的检测效果较差,检测准确率仅为 68.87%。稀疏表示的检测准确率率虽有所提升为 74.61%,但分类效果仍然不如人意且训练时间较长。

与同样作为神经网络的 DNN 相比,由于 CNN、改进 CNN 网络输入为 STFT 时频图,可以有效地提升隐写信号整体时频变换的特性,所以总体识别效果更优,平均识别率均达到了 80%以上。此外,本章提出的改进 CNN 网络由于引入了 Inception 结构,可以在同一层上同时捕捉多种尺度的图像特征,识别率较传统 CNN 网络提升了 1.86%。

4 结束语

本章提出了一种改进的多通道卷积神经网络,并将其应用于语音隐写分析。首先构建了语音隐写信号的 STFT 谱图以及其沿时间轴、频率轴的一阶差分作为网络输入。其次,在网络结构上,引入了 Inception 结构在同一卷积层上提取各种不同尺度的特征,使用全局平均池化来代替全连接层,在降低参数的同时提升了网络的泛化能力。最后通过对比实验验证了改进 CNN 网络的性能,实验结果表明,改进 CNN 网络对语音隐写信号的检测准确率远高于传统识别方法。

参考文献:

[1] Fridrich J. Steganography in Digital Media:Principles,Algorithms,and Applications[M]. Cambridge:University Press,2009:108-220.

[2] Rekik S,Guerchi D,Hamam H,et al. Audio Steganography Coding Using the Discrete Wavelet Transforms[J]. International Journal of Computer Science & Security,2012,6(1):283-290.

[3] Yang Z L,Guo X Q,Chen Z M,et al. RNN-Stega:Linguistic Steganography Based on Recurrent Neural Networks[J]. IEEE Transactions on Information Forensics and Security,2018,14(5):1280-1295.

[4] Xiao H,Xiao B,Huang Y. Implementation of Covert Communication Based on Steganography[C]//2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing. IEEE,Harbin,China,2008:1512-1515.

[5] Chen B,Wornell G W. Quantization Index Modulation:A Class of Provably Good Methods for Digital Watermarking and Information Embedding[J]. IEEE Transactions on Information Theory,2001,47(4):1423-1443.

[6] Xiao B,Huang Y F,Tang S Y. An Approach to Information Hiding in Low Bit-Rate Speech Stream[C]//IEEE GLOBECOM Global Telecommunications Conference. IEEE,New Orleans,LA,USA,

2008:1-5.

[7] Wu Z J, Wei G, Wei Y. LPC Parameters Substitution for Speech Information Hiding[J]. The Journal of China Universities of Posts and Telecommunications, 2009, 16(6): 103-112.

[8] He K M, Gkioxari G, Dollar P, et al. Mask R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2961-2969.

[9] Hinton G, Deng L, Yu D, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[10] Gehring J, Auli M, Grangier D, et al. Convolutional Sequence to Sequence Learning[C]//International Conference on Machine Learning. PMLR, Sydney Australia, 2017: 1243-1252.

[11] Chen B, Luo W, and Li H. 2017. Audio Steganalysis with Convolutional Neural Network[C]//Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. New York, USA, 2017: 85-90.

[12] Wang Y T, Yang K, Yi X W, et al. CNN-Based Steganalysis of MP3 Steganography in the Entropy Code Domain[C]//Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security, New York, USA, 2018. 55-65.

[13] Lin Z N, Huang Y F, Wang J L. RNN-SM: Fast Steganalysis of VoIP Streams Using Recurrent Neural Network[J]. IEEE Transactions on Information Forensics & Security, 2018, 99(2018): 1-1.

[14] Deng A D, Cao H, Tong H, et al. Recognition of Acoustic Emission Signal Based on the Algorithms of TDNN and GMM[J]. Applied Mathematics & Information Sciences, 2014, 8(2): 907-916.

[15] LeCun Y, Bottou L, Bengio Y, et al. Gradient-Based Learning Applied to Document Recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

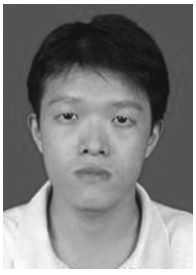
[16] Krizhevsky A, Sutskever I, Hinton G E. Imagenet Classification with Deep Convolutional Neural Networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

李　晶(1982—),女,汉族,天津蓟县人,工学博士,现为东南大学信息学院博士后,工程师,主要研究方向为信号与信息处理、模式识别、机器学习,230149053@seu.edu.cn;

邓　伟(1996—),男,湖南新宁人,工学硕士,研究方向为语音信号处理220184451@seu.edu.cn;

杨　勇(1981—),男,河北涉县人,工学博士,现为东南大学信息学院博士后,副教授,研究方向为矿井信息处理,信号与信息处理,YongYang@cumt.edu.cn;

朱作付(1968—),男,江苏新沂人,教授,研究方向为算法设计,软件工程,65112208@qq.com;



贾伟伟(1981—),男,山西临汾人,硕士,研究方向为嵌入式应用,jiaww@mail.xzcit.cn。