Speech Steganalysis for LSB Steganography*

YANG Yong[1,2], DENG Wei[2], LI Jing[3*], ZHU Zuofu[1], JIA Weiwei[1], WANG Chen[1], ZHANG Yan[1]
(1.School of Information Engineering, Xuzhou College of Industrial Technology, Xuzhou Jiangsu 221140, China;
2.Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education,
Southeast University, Nanjing Jiangsu 210096, China;
3.School of Information Engineering, Nanjing Audit Unversity, Nanjing Jiangsu 211815, China)

Abstract: The least significant bit (LSB) steganography is analyzed. In the current network environment, the detection of low embedding rate speech steganographic signal is still a very challenging topic. In recent years, neural network model has achieved remarkable performance in many subjects. The mainstream architectures of neural networks include convolution neural networks (CNNs) and recurrent neural networks (RNNs), which use different ways to understand various signals. An appropriate method is proposed to combine the advantages of the two architectures, and then a new model i.e., CNN-LSTM network is constructed to detect LSB based steganography. In the proposed model, BiLSTM is used to capture the long-term context information of the signal, and CNN is used to capture the local and global features. The experimental results show that the model achieves better effect than the contrast method for LSB based steganalysis.

Key words: speech steganalysis; deep neural network; least significant bit
EEACC: 6130 **doi**: 10.3969/j.issn.1005-9490.2023.03.035

针对 LSB 隐写的语音隐写分析方法*

杨 勇[1,2], 邓 伟[2], 李 晶[3*], 朱作付[1], 贾伟伟[1], 王 晨[1], 张 岩[1]
(1.徐州工业职业技术学院信息工程学院, 江苏 徐州 221140;
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.南京审计大学信息工程学院, 江苏 南京 211815)

摘 要: 对常用的最低有效位(LSB)隐写技术进行了隐写分析。在目前的网络环境中, 检测低嵌入率语音隐写信号仍然是一个非常有挑战性的课题。近年来, 神经网络模型在许多课题中都取得了显著的性能。神经网络的主流架构包括卷积神经网络(CNNs)和循环神经网络(RNNs), 这两种网络采用了不同的方式来理解各种信号。提出了一种合适的方法来结合这两种架构的优点, 然后构造了一个新的模型, 即 CNN-LSTM 网络来检测基于 LSB 的隐写方法。在所提出的模型中, 使用双向长短时记忆循环神经网络(BiLSTM)来捕获信号的长时间上下文信息, 然后使用 CNN 捕获局部特征和全局特征。实验结果表明, 该模型相较对比方法对于基于 LSB 的隐写分析达到了更好的效果。

关键词: 语音隐写分析; 深度神经网络; 最低有效位隐写
中图分类号: TN912.3; TP183 **文献标识码**: A **文章编号**: 1005-9490(2023)03-0790-05

由于静态载体不能满足客户对实时双工交互的需求, 流媒体的动态载体如 VoIP (Voice over IP) 是最适合秘密通信[1]的载体之一。国际电信联盟(International Telecommunication Union, ITU)已经为 VoIP 应用程序[2]定义了高速语音编解码器 G.711。最低有效位(LSB)隐写术可以通过直接修改每个语音采样点的最低有效位来有效地应用于波形编码器中隐藏信息。像 Skype 和 Gtalk 这样的流媒体软件很容易添加隐写模块, 这意味着可以随意建立秘密的通信通道。因此, 对 VoIP 隐写分析的研究具有重要的意义。

然而, 低嵌入率的 LSB 隐写术对载波结构的改变非常小, 而且 VoIP 语音流是实时性的, 在 VoIP 中很难检测到 LSB 隐写术。目前的隐写分析方法虽然取得了很大的进展, 但在 VoIP 场景下还远远不能令人满意。注意到 CNN 能够从时间或空间数据中学习局部响应, 但缺乏学习序列相关性的能力, 而 RNN 能够处理任何长度的序列, 并捕捉长期的上下文相关性[3]。

本文提出了一种新的 CNN-LSTM 模型来检测 VoIP 流的 LSB 隐写。在该模型中, 使用双向长短时

项目来源: 国家自然科学基金项目(52005267)
收稿日期: 2021-10-25 **修改日期**: 2022-02-18

记忆循环神经网络从语音中获取长期上下文信息,使用不同核大小的 CNN 层提取语音帧的局部特征。最后,全连接层和 softmax 层作为分类器输出类概率。此外,我们采用声码器后的量化索引序列作为模型的输入。下文的组织如下:首先详述了提出的 CNN-LSTM 网络结构的具体细节,然后给出了实验结果,并对模型进行了分析。最后,对本文的工作做出总结。

1　深度神经网络方案

所提出的 CNN-LSTM 模型的体系结构如图 1 所示。接下来,本文将从下到上详细解释模型中的每一层。

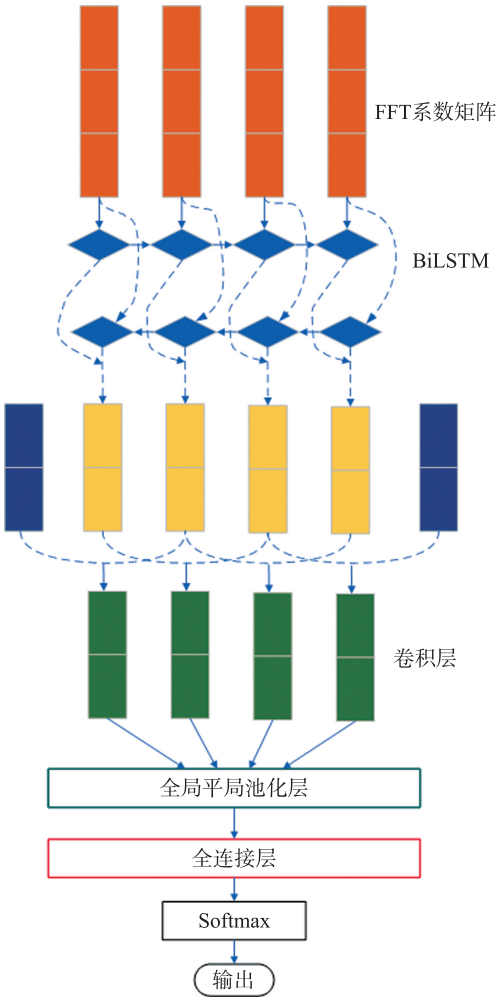

图 1　网络结构原理图

1.1　双向长短时记忆网络

语音帧的短时傅里叶变换系数作为网络输入。本文提出的网络结构的第一层是 BiLSTM(双向长短时记忆网络)层,该层的输入为短时傅里叶变换系数。给定一个输入序列 $x=(x_1,\cdots,x_T)$,标准的 RNN 层通过下面的公式来计算隐藏向量序列 $h=(h_1,\cdots,h_T)$ 和本层的输出向量序列 $y=(y_1,\cdots,y_T)$:

$$\begin{aligned} h_t &= H(W_h\cdot[h_{t-1},x_t]+\boldsymbol{b}_h) \\ y_t &= \boldsymbol{W}_{hy}\cdot h_t+\boldsymbol{b}_y \end{aligned} \tag{1}$$

式中:$t=1,\cdots,T$。$\boldsymbol{W}$ 表示权重矩阵,$\boldsymbol{b}$ 表示偏置向量。H 是隐藏层函数,其通常是 sigmoid 函数的一个元素级应用。

RNN 网络的特点在于能够将之前的信息与当前的任务联系起来,比如使用之前的帧可能会提高对当前帧的理解。而在语音信号处理中,所有语音都是一次性转录的,这就不得不利用未来语境。因此,文献[4]中提出一种双向 RNN(Bidirectional Recurrent Neural Network,BiRNN),通过两个独立的隐含层在前后两个方向上处理数据,然后将数据前馈到同一个输出层。BiRNN 通过后向层从 $t=T$ 迭代到 1,前向层从 $t=1$ 迭代到 T,然后更新输出层,计算前向隐藏序列 $\vec{h}$,后向隐藏序列 $\overleftarrow{h}$ 和最终的输出序列 y:

RNN 在多个时序处理的课题中都取得了很好的效果。然而,RNN 很难处理长期依赖[5]。文献[6]中的长短期记忆(Long-Short Term Memory,LSTM)架构使用特制的存储单元来存储信息,很好地解决了这一问题。对于文献[7]中使用的 LSTM 版本,H 是由以下复合函数实现的:

$$i_t=\sigma(\boldsymbol{W}_i\cdot[h_{t-1},x_t]+\boldsymbol{b}_i) \tag{2}$$

$$f_t=\sigma(\boldsymbol{W}_f\cdot[h_{t-1},x_t]+\boldsymbol{b}_f) \tag{3}$$

$$q_t=\tanh(\boldsymbol{W}_q\cdot[h_{t-1},x_t]+\boldsymbol{b}_q) \tag{4}$$

$$o_t=\sigma(\boldsymbol{W}_o\cdot[h_{t-1},x_t]+\boldsymbol{b}_o) \tag{5}$$

$$c_t=f_t\odot c_{t-1}+i_t\odot q_t \tag{6}$$

$$h_t=o_t\odot\tanh(c_t) \tag{7}$$

这里 σ 是 logistic sigmoid 函数。LSTM 有三种门,分别是输入门 i_t、遗忘门 f_t 和输出门 o_t。另一种标记 c_t 作为单元激活载体。所有门和激活向量的维度与隐藏向量 h 相同。单元和门向量的权重矩阵都是对角矩阵,所以每个门向量中的元素 m 只接收来自单元向量中的元素 m 的输入。

将 BiRNN 与 LSTM 相结合,得到双向的 LSTM[4](BiLSTM),这种网络既能发现和利用远距离序列的信息,又能充分利用语音载体的前后双向的信息。在该模型中,双向 LSTM 层也可以作为一个转换器,逐步建立语音数据的更高层次表示。在这一层之后,会生成一个新的有代表性的帧序列向量,作为后续网络层的输入。

1.2　卷积层

卷积层是本文提出模型的第二部分。CNN 由于能够捕捉空间或时间结构的局部关联,在计算机视觉[8]、语音识别[9]和自然语言处理[10]中取得了优异的性能。在语音建模方面,CNN 也被应用于许多

语音任务中,它可以建模时间和频谱局部相关并保证增益平移不变的特性,并在文献[11]中有成功的尝试。在本文提出的模型中选择使用不同窗口大小的一维卷积来捕获不同尺度的特征。

一维卷积涉及到一个滤波器向量在输入序列上滑动并检测不同位置的特征。设 $h \in R^d$ 为双向 LSTM 层生成的第 i 帧的 d 维帧向量。设 $h \in R^{L*d}$ 表示输入段,L 为帧数。设 k 为滤波器的长度,向量 $m \in R^{k*d}$ 为进行卷积运算的滤波器。对于语音帧的第 j 点,用一个窗向量 $\boldsymbol{w}_j$ 表示如下:

$$\boldsymbol{w}_j = [h_j, h_{j+1}, \cdots, h_{j+k-1}] \tag{8}$$

式中:逗号表示帧向量连接。一个滤波器 m 与每个位置的窗向量进行有效卷积,生成特征映射 $h \in R^{(L-k+1)*d}$;特征图种的每个向量 $\boldsymbol{g}_j$ 的运算公式如下:

$$\boldsymbol{g}_j = f(w_j \circ m + b) \tag{9}$$

式中:$\circ$ 表示一个元素的乘法,$b \in R$ 是一个偏置项,f 是一个非线性变换函数,可以是 sigmoid、双曲正切、ReLU 等。本文研究选择 ReLU 作为非线性函数。

在本文提出的模型中,使用多个过滤器来生成多个特征映射。对于具有相同长度的 n 个过滤器,生成的 n 个特征映射可以重新排列为每个窗口 w_j 的特征表示:

$$W = [g_1; g_2; \cdots; g_n] \tag{10}$$

式中:分号表示列向量连接,g_i 是第 i 个过滤器生成的特征图。$W \in R^{(L-k+1)\times n}$ 中的每一行 W_j 是对位置 j 的窗向量的 n 个滤波器生成的新特征表示。

通常在卷积后的特征映射上使用池化层来降维或选择最重要的特征,一般使用平均池化或最大池化。在本文提出的模型中,采用全局平均池化来降低特征维数。

模型的最后一部分是全连接层,利用 softmax 激活函数作为分类器来判断样本是否含隐藏信息。此外,为了加快收敛速度和克服过拟合问题,本文还在模型中加入了批归一化层[12]和 Dropout[13]。

2 实验分析

2.1 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 LSB 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。此外,数据集中的语音片段长度是影响检测精度的另一个因素。本文的实验中,将载体语音数据集和隐写语音数据集中的样本分别切割为 0.1 s、0.5 s、1 s 和 3 s 来测试不同持续时间下的模型性能。

模型中的超参数是通过试验集上的交叉验证来选择的。具体来说,BiLSTM 隐藏状态的维数为 64,CNN 滤波器的窗口大小分别为 3、4、5。每个 CNN 过滤器的数量是 128。全连通层的尺寸为 64,全连通层的 dropout rate 为 0.6。批处理大小为 256,高位训练单元设置为 100。我们使用 Adam[14] 作为网络训练的优化器,深度神经网络模型由 Keras 实现。通过分类准确率来评估模型的性能,分类准确率定义为正确分类的样本数与样本总数的比值。

2.2 实验评估

本文选择三种方法:文献[15]中的 IDC 和文献[16]中的 QCCN 作为基线方法来证明本文提出模型的有效性。本文所提模型在测试集中的性能如表 1 所示。根据表 1,可以看出本文的模型优于所有对比方法。对比不同的模型,我们可以得出 IDC 和 QCCN 使用的传统机器学习算法表现较差。同时,相较对比的深度学习方法 RNN-SM,本文模型在检测精度上有了显著的提高。

表 1 本文模型在不同样本时长和嵌入率下的实验准确率

样本长度	分析方法	嵌入率/%		
		1	10	30
0.1 s	IDC	60.36	67.25	72.34
	QCCN	63.59	68.29	74.61
	本文方法	65.46	70.52	76.16
0.5 s	IDC	63.58	70.61	75.76
	QCCN	65.74	72.39	77.92
	本文方法	70.68	75.81	79.63
1 s	IDC	70.13	74.94	80.66
	QCCN	70.37	76.35	82.64
	本文方法	75.56	79.16	85.75
3 s	IDC	74.87	78.58	84.51
	QCCN	75.25	80.36	85.37
	本文方法	80.31	83.67	87.57

语音片段的时长是影响 LSB 隐写方法的一个重要因素。因此,将嵌入率固定在 10%,并研究语音长度对隐写分析的影响。图 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

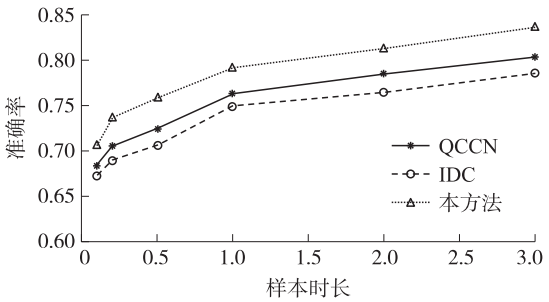

图 2 检测准确率与样本时长关系图

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。

嵌入率是影响检测精度的另一个重要因素。如图 3 展示了在样本时长为 1 s 时,准确率与嵌入率的关系。当嵌入率较低时,随着嵌入率的增加,准确率显著提高。当嵌入率在 40%以上时,检测准确率在 95%以上。同时,在嵌入率较低的情况下,该模型显著提高了检测精度。隐写算法通常采用低嵌入率策略,以避免容易被发现,这给隐写分析带来了挑战。本文提出的模型在低嵌入率下的优异性能使其在现实场景中更加实用。

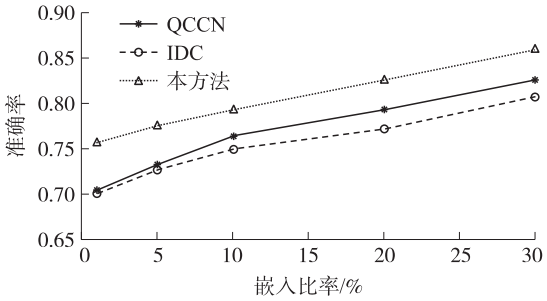

图 3　检测准确率与嵌入率关系图

2.3　模型调整对比

在这一节试图调整模型的结构来验证该模型中不同部分的功能,同时佐证本文使用的的模型结构的优越性。不同调整下模型的性能如表 2 所示。

表 2　不同模型调整策略下的实验结果

模型调整	准确率/%
本文提出的模型	79.16
只用 BiLSTM	76.78
只用 CNN	66.12
将 CNN 放在 BiLSTM 之前	75.32
使用典型的 LSTM 代替 BiLSTM	76.35

实验中的片段长度为 1 s,嵌入率为 10%。首先,从表中可以看出将 CNN 和 BiLSTM 结合在一个模型中的优越性,其可以充分利用不同结构的优势,BiLSTM 层在整个体系结构中起着至关重要的作用。同时,由于 CNN 模型不使用任何上下文信息,而使用 LSTM 的模型只捕获语音载体的前向信息,但使用 BiLSTM 的模型能同时捕获语音的前向和后向信息,因此捕获语音中方向信息的有效性非常重要。而且在 BiLSTM 之后放置 CNN 是非常有效的,这主要是因为输入矩阵中的帧向量维数较低,CNN 可以捕捉到局部特征,而 BiLSTM 作为变换将帧向量映射到高维,为后续的模块生成更具区分度的特征向量。

3　结束语

本文提出了一种结合 CNN 和 LSTM 的 LSB 隐写分析的方法。本文提出的模型架构充分利用了 LSTM 和 CNN 两种主流架构,其中双向 LSTM 用于捕获语音的较长持续时间的上下文信息,并生成更具区分度的特征帧向量。而之后的 CNN 被用来捕获局部特征以及全局和时序的语音特征。实验证明,相较其他针对 LSB 的隐写对比分析方法,本文提出的的模型结构能够达到更好的检测准确率,该模型是一种实用的有效模型,可以进一步研究。

参考文献:

[1] Huang Y F,Tang S Y. Covert Voice over Internet Protocol Communications Based on Spatial Model[J]. SCIENCE CHINA(Technological Sciences),2016,59(1):117-127.

[2] 黄永峰,李松斌. 网络隐蔽通信及其检测技术[M]. 北京:清华大学出版社,2016:1-10.

[3] Geetha S,Ishwarya N,Kamaraj N. Audio Steganalysis with Hausdorff Distance Higher Order Statistics Using a Rule Based Decision Tree Paradigm[J]. Expert Systems with Applications,2010,37(12):7469-7482.

[4] Schuster M,Paliwal K K. Bidirectional Recurrent Neural Networks[J]. IEEE Transactions on Signal Processing,1997,45(11):2673-2681.

[5] Bengio Y,Simard P,Frasconi P. Learning Long-Term Dependencies with Gradient Descent is Difficult[J]. IEEE Transactions on Neural Networks,1994,5(2):157-166.

[6] Graves A,Schmidhuber J. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural network Architectures[J]. Neural Networks,2005,18(5):602-610.

[7] Felix A G,Nicol N S,Jãijrgen S. Learning Precise Timing with LSTM Recurrent Networks.[J] Journal of Machine Learning Research,2002,3(1):115-143.

[8] He K M,Gkioxari G,Dollar P,et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2020,42(2):386-397.

[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] Yang Z L,Huang Y F,Jiang Y R,et al. Clinical Assistant Diagnosis for Electronic Medical Record Based on Convolutional Neural Network. Scientific Reports,2018 8(1):6329.

[11] Palaz D, Magimai-Doss M, Collobert R. Convolutional

Neural Networks-based Continuous Speech Recognition Using Raw Speech Signal[C]// IEEE International Conference on Acoustics, Speech, and Signal Processing South Brisbane, QLD, Australia, 2015: 4295-4299.

[12] Ioffe E, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[C]//32nd International Conference on Machine Learning, Lille, France, 2015: 448-456.

[13] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting [J]. Journal of Machine Learning Research, 2014, 15 (1): 1929-1958.

[14] Kingma D P, Ba J. Adam: A Method for Stochastic Optimization[C]//International Conference on Learning Representations, San Diego, CA, USA, 2015: ArXiv: 1412.6980.

[15] Li S B, Tao H Z, Huang Y F. 2012. Detection of Quantization Index Modulation Steganography in G. 723. 1 bit Stream Based on Quantization Index Sequence Analysis [J]. Journal of Zhejiang University-Science C(Computers & Electronics), 2012, 13(8): 624-634.

[16] Li S B, Jia Y Z, Kuo C C J. Steganalysis of QIM Steganography in Low-Bit-Rate Speech Signals[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2017, 25(5): 1011-1022.

杨 勇(1981—),男,河北涉县人,工学博士,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;

李 晶(1982—),女,汉族,天津蓟县人,工学博士,中级工程师,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;

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;

王 晨(1992—),男,江苏邳州人,硕士,助教,研究方向为虚拟现实、数据挖掘,wangc@mail.xzcit.cn;

张 岩(1993—),女,江苏沛县人,工学硕士,研究实习员,研究方向为人工智能。18362981195@163.com。