计算机应用
*Journal of Computer Applications*
ISSN 1001-9081,CN 51-1307/TP

# 《计算机应用》网络首发论文

题目：融合残差网络和极限梯度提升的音频隐写检测模型
作者：陈朗，王让定，严迪群，林昱臻
收稿日期：2020-06-19
网络首发日期：2020-10-22
引用格式：陈朗，王让定，严迪群，林昱臻．融合残差网络和极限梯度提升的音频隐写检测模型[J/OL]．计算机应用．
https://kns.cnki.net/kcms/detail/51.1307.TP.20201021.0856.010.html

Journal of Computer Applications 计算机应用 | ISSN 1001-9081 CODEN JYIIDU | 2020-10-20 http://www.joca.cn

DOI:10.11772/j.issn.1001-9081.2020060775

# 融合残差网络和极限梯度提升的音频隐写检测模型

陈 朗，王让定*，严迪群，林昱臻

(宁波大学 信息科学与工程学院，浙江 宁波 315211)

(*通信作者电子邮箱 wangrangding@nbu.edu.cn)

**摘 要:** 针对目前音频隐写检测方法对基于校验网格编码（STC）的音频隐写检测准确较低的问题，考虑到卷积神经网络在抽象特征提取上的优势，提出一种融合深度残差网络和极限梯度提升（XGBoost）的音频隐写检测模型。首先，利用固定参数的高通滤波器预处理输入的音频，并通过三个卷积层提取特征，其中第一个卷积层使用了截断线性单元（TLU）激活函数，使得模型适应低信噪比下的隐写信号分布；其次，通过五个阶段的残差块和池化操作进一步提取抽象特征；最后，经过全连接层和 Dropout 层将提取的高维特征作为 XGBoost 模型的输入进行分类。分别对 STC 隐写和最低有效位匹配（LSBM）隐写进行检测，实验结果表明，所提出的模型在 0.5bps、0.2bps、0.1bps 三种嵌入率下,即音频每个采样值平均修改的比特数分别为 0.5、0.2、0.1 时，对子校验矩阵高度为 7 的 STC 隐写的平均检测准确率分别为 73.27%、70.16%、65.18%，对 LSBM 隐写的平均检测准确率分别为 86.58%、76.08%、72.82%，相较于传统提取手工特征的隐写检测方法和深度学习隐写检测方法，对两种隐写算法的平均检测准确率均提高了 10 个百分点以上。

**关键词:** 深度残差网络；极限梯度提升；校验网格编码隐写；最低有效位匹配隐写；音频隐写检测

**中图分类号:TP391.4** **文献标志码：A**

# Audio steganography detection model combing residual network and extreme gradient boosting

CHEN Lang, WANG Rangding*, YAN Diqun, LIN Yuzhen

(*Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo Zhejiang* 315211, *China*)

**Abstract:** Aiming at the problem that the current audio steganography detection methods have got low accuracies in detecting audio steganography based on Syndrome-Trellis Codes (STC), considering the advantages of convolutional neural network in extracting abstract feature, a model for audio steganography detection which combined deep residual network and eXtreme Gradient Boosting (XGBoost) was proposed. Firstly, a fixed-parameter high-pass filter was used to preprocess the input audio, and features were extracted through three convolutional layers. Truncated Linear Unit (TLU) activation function was applied in the first convolutional layer to make the model adapt to the distribution of steganographic signals with low signal-to-noise ratio. Then, abstract features were further extracted by five-stage residual blocks and pooling operations. Finally, the extracted high-dimensional features were classified as inputs to the XGBoost model through fully connected layers and dropout layers. The STC steganography and the Least Significant Bit Matching (LSBM) steganography were detected by the proposed methods respectively. When the embedding rates are 0.5 bit per sample, 0.2 bit per sample and 0.1 bit per sample, that is to say, the average number of bits modified for per audio sample equals to 0.5, 0.2 and 0.1 respectively. The experimental results show that the proposed model achieves average detection accuracies of 73.27%, 70.16% and 65.18% respectively for the STC steganography with a submatrix height of 7, and the average detection accuracies of the LSBM steganography are 86.58%, 76.08% and 72.82% respectively. Compared with the traditional steganography detection methods

**收稿日期**: 2020-06-19; **修回日期**: 2020-08-17; **录用日期**: 2020-09-08。

**基金项目**: 国家自然科学基金资助项目(U1736215, 61672302, 61901237)；浙江省自然科学基金资助项目(LY20F020010, LY17F020010)；浙江省移动网应用技术重点实验室开放基金资助项目(F2018001)；宁波大学王宽诚幸福基金资助项目；宁波大学研究生科研创新基金资助项目(IF2020131)。

**作者简介**: 陈朗(1997—)，男，四川巴中人，硕士研究生，主要研究方向：多媒体信息安全、信息隐藏、隐写分析；王让定(1962—)，男，甘肃天水人，教授，博士，CCF 会员，主要研究方向：多媒体信息安全、信息隐藏、隐写分析；严迪群(1979—)，男，浙江余姚人，副教授，博士，CCF 会员，主要研究方向：多媒体信息安全、数字取证；林昱臻(1994—)，男，浙江宁波人，硕士研究生，主要研究方向：多媒体信息安全、隐写分析。

based on extracting handcrafted features and deep learning steganography detection methods, average detection accuracies of the two steganography algorithms both have increased by more than 10 percent points.

**Keywords:** deep residual network; extreme gradient boosting; syndrome-trellis codes steganography; least significant bit matching steganography; audio steganography detection

## 0 引言

伴随着互联网和多媒体处理技术的飞速发展，互联网平台上涌现出越来越多的数字媒体应用，数字媒体安全问题变得日益突出，已成为数字经济等领域迫切需要破解的难题。在信息安全领域，保障信息安全的技术有密码术和信息隐藏。然而，应用密码术获得的密文在公开信道上传输，很容易引起第三方的监听进而采取手段破坏隐蔽通信；信息隐藏则很好地解决了密码术存在的问题，信息隐藏将密信以“暗度陈仓”的方式隐藏在另一非机密载体中进行传输，从而可实现隐蔽通信。广义的信息隐藏包括隐写术和隐写分析，隐写术利用人类感官对数字信号的感知冗余性将秘密信息隐藏到数字媒体中，目的是避免引起检查者的注意。相反的，隐写分析则是通过分析载体对像和嵌密对象在感知和统计特征上的差异来判别载体中是否隐藏密信。

近二十多年来，研究人员在采用传统方法进行音频隐写分析方面做了大量工作，主要集中于手工特征的设计上。Johnson 等[1]通过短时傅里叶变换(Short-Time Fourier Transform, STFT)提取音频统计特征，以支持向量机作为分类器。实验结果显示，对较低嵌入率下最低有效位（Least Significant Bit, LSB）隐写算法并不能有效检测。Kraetzer 等[2]第一次提出了基于梅尔倒谱(Mel-cepstrum)的网络协议(Internet Protocol, IP)语音隐写检测方法，以梅尔倒谱特征作为支持向量机的输入。Liu 等[3]提出基于时域导数频谱和梅尔倒谱的音频隐写检测方法，以傅里叶频谱统计和二阶差分滤波器获取的梅尔倒谱系数作为提取的手工特征，采用支持向量机作为分类器，实验结果显示，相较于仅提取梅尔倒谱特征的方法检测准确率有明显提高。Liu 等[4]将梅尔倒谱系数和根据二阶差分滤波器获取的马尔科夫转移矩阵作为提取的手工特征，并研究了音频信号复杂度和检测准确率之间的关系，实验结果表明，对 LSB 等低隐蔽性隐写算法检测准确率达到 90%以上。Geetha 等[5]设计了一种基于音频质量测度的特征，计算载体音频和嵌密音频的 Hausdorff 距离作为失真度测量特征。王昱洁等[6]提出了一种基于改进离散余弦变换（Modified Discrete Cosine Transform, MDCT）量化系数统计特征的高级音频编码（Advanced Audio Coding, AAC）隐写检测方法，从 MDCT 量化系数中提取了广义高斯分布模型的参数、量化系数分布直方图的频域统计矩、帧内和帧间 MDCT 系数的隐马尔科夫转移矩阵的部分数据作为手工设计的特征，采用支持向量机作为分类器，实验结果表明，对基于 MDCT 量化系数的直接扩频隐写算法检测效果较好。王昱洁等[7]提出了一种基于模糊 C 均值(Fuzzy C-means Clustering, FCC)和单类支持向量机(One Class Support Vector Machine, OC-SVM)的音频隐写检测方法，首先提取短时傅里叶频谱统计特征和基于音频质量测度的特征，然后对提取的特征进行 FCC 聚类从而得到 C 个聚类，最后以单类支持向量机作为分类器，实验结果表明，对几种经典隐写算法在满嵌时总体检测率达到 85.1%。Han 等[8]提出了一种基于线性预测的音频隐写检测方法，提取了线性预测系数、线性预测残差、线性预测频谱和线性预测频谱系数，将它们的最大值作为支持向量机的输入特征，实验结果表明，对四种低隐蔽性传统隐写算法的检测准确率均在 96%以上。Ren 等[9]提出了一种基于帧内和帧间的改进离散余弦变换系数差异的高级音频编码音频隐写检测方法。

随着人工智能时代的到来，不同的深度学习网络相继提出，在许多领域已经达到了目前最好的效果，尤其是在语音识别和计算机视觉领域。Chen 等[10]首次提出了针对 LSB 匹配隐写的卷积神经网络(Convolutional Neural Network, CNN)，该方法的检测准确率高于手工提取特征的传统方法。随后，Lin 等[11]提出了一种改进的卷积神经网络，首先采用四阶差分滤波器计算残差进行预处理，其中滤波器参数在网络中设置为可训练的，再将预处理后的音频数据输入到网络中，采用迁移学习的策略进行低嵌入率下音频隐写检测。Wang 等[12]提出了一种在熵编码域（Entropy Code Domain）对 MP3 进行隐写分析的卷积神经网络，使用量化改进的离散余弦变换(Quantified Modified Discrete Cosine Transform, QMDCT)特征作为网络的输入的浅层特征，准确率相比于以前的隐写分析方法提升了 20%以上。Yang 等[13]融合了卷积神经网络和循环神经网络(Recurrent Neural Network, RNN)的优势，提出了一种联合卷积神经网络和长短期记忆(Long Short-Term Memory, LSTM)网络的模型（CNN-LSTM）用于检测网络协议语音流（Voice over Internet Protocol, VoIP）上基于量化索引调制（Quantization Index Modulation, QIM）隐写算法，其中的 LSTM 用于学习音频载体的上下文信息，CNN 用于学习分类特征。

然而，上述方法不能有效检测高隐蔽性音频信息隐藏算法，如 STC 框架[14]下的音频隐写算法。基于此，本文提出了一种融合深度残差网络和极限梯度提升(eXtreme Gradient Boosting, XGBoost)[15]的音频隐写检测模型(ResNet-XGBoost)。该方法首先采用残差网络模型提取输入音频的高维抽象特征，再用 XGBoost 模型对残差网络提取的高维抽象特征进行分类。最后，实验中对 STC 音频隐写算法在三种不同嵌入率下的检测准确率相较于传统提取手工特征的检测方法和基于深度学习的检测方法均有显著提升。

## 1 STC 框架

STC 框架下的隐写算法属于一种基于校验网格编码的最小化嵌入失真隐写算法，具备嵌入容量大、安全性高等优势。隐写设计者先定义失真函数，采用维特比算法获取合适的原始载体修改模式，使得嵌入秘密信息后的总体失真接近理论最小值。接收方不必知道失真函数，仅根据校验矩阵便可提取出密信。

假设原始载体为二进制向量 $\boldsymbol{X}\in\{0,1\}^n$，嵌密载体 $\boldsymbol{Y}\in\{0,1\}^n$，秘密信息 $\boldsymbol{M}\in\{0,1\}^m$，且 $n\geq m$。定义嵌入函数：$Emb:\{0,1\}^n\times\{0,1\}^m\to\{0,1\}^m$，提取函数：$Ext:\{0,1\}^n\to\{0,1\}^m$ 且满足如下条件：

$$Ext(Emb(\boldsymbol{X},\boldsymbol{M}))=\boldsymbol{M}\quad\forall\boldsymbol{X}\in\{0,1\}^n,\forall\boldsymbol{M}\in\{0,1\}^m \tag{1}$$

构造一个矩阵(亦即校验矩阵，其结构如图 1 所示)，使得嵌密载体满足：

$$\boldsymbol{H}\boldsymbol{Y}^{\mathrm{T}}=\boldsymbol{M} \tag{2}$$

若嵌入操作相互独立即互相不影响，则嵌入密信后引起的载体失真为加性失真。定义嵌入密信后的总失真函数为：

$$D(\boldsymbol{X},\boldsymbol{Y})=\sum_{i=1}^{n}\rho_i|x_i-y_i| \tag{3}$$

其中：$D(\boldsymbol{X},\boldsymbol{Y})$ 为嵌入密信后的总失真；$i$ 为当前载体位置索引；$n$ 为载体长度；$\rho_i$ 为当前元素变化引起的失真，在复杂区域取值较小，在平滑区域取值较大。

STC 隐写的目标是使得总失真函数尽可能小，即求解最优的 $\boldsymbol{Y}$,既满足 $\boldsymbol{H}\boldsymbol{Y}^{\mathrm{T}}=\boldsymbol{M}$ 又使 $D(\boldsymbol{X},\boldsymbol{Y})$ 最小，其嵌密和提取过程的数学表达为：

$$Emb(\boldsymbol{X},\boldsymbol{M})=arg\min_{\boldsymbol{Y}\in\boldsymbol{C}(\boldsymbol{M})}D(\boldsymbol{X},\boldsymbol{Y}) \tag{4}$$

$$Ext(\boldsymbol{Y})=\boldsymbol{H}\boldsymbol{Y}^{\mathrm{T}} \tag{5}$$

其中，$\boldsymbol{H}\in\{0,1\}^{m\times n}$ 为奇偶校验矩阵，是收发双方共享的参数；$\boldsymbol{C}(\boldsymbol{M})$ 是伴随式 $\boldsymbol{M}$ 的陪集，通俗地理解就是满足式(4)、(5)的可行解，即 $\boldsymbol{C}(\boldsymbol{M})=\left\{\boldsymbol{z}\in\{0,1\}^n\mid\boldsymbol{H}\boldsymbol{z}=\boldsymbol{M}\right\}$。校验矩阵 $H$ 是由子校验矩阵 $\hat{H}_{h\times w}$ 按主对角线顺序错行排列而成。对于子校验矩阵的尺寸参数选取，$h\in[2,32]$，一般来说，$h$ 越大，STC 编码时间复杂度越高，嵌入容量越大，安全性也越高，通常取 $6\leq h\leq 15$。$w$ 由嵌入率 $\alpha$ 决定，当 $\frac{1}{\alpha}$ 为整数时，$w=\frac{1}{\alpha}$；否则，$w$ 可取两个值 $w_1=\left\lfloor\frac{1}{\alpha}\right\rfloor, w_2=\left\lfloor\frac{1}{\alpha}\right\rfloor+1$。

STC 编码过程在网格图上完成，求解 $\boldsymbol{Y}$ 时可将所有满足条件 $\boldsymbol{H}\boldsymbol{Y}^{\mathrm{T}}=\boldsymbol{M}$ 的可行解 $\boldsymbol{Y}$ 在网格图中用一条路径表示。STC 编码在所有路径中选择具有最小权重的路径，该路径对应一个最佳矢量 $\boldsymbol{Y}$，使 $\boldsymbol{H}\boldsymbol{Y}^{\mathrm{T}}=\boldsymbol{M}$，并且使得嵌入总失真函数 $D(\boldsymbol{X},\boldsymbol{Y})$ 达到最小。维特比算法包括前向计算和反向计算，利用该算法可求解出最佳 $\boldsymbol{Y}$。

STC 音频隐写一般将 STC 编码和失真代价结合。首先利用失真函数计算载体中每个元素被修改后的失真代价，再应用 STC 编码确定需要修改的位置并进行嵌密操作，从而得到嵌密音频。STC 音频隐写流程如图 2 所示。

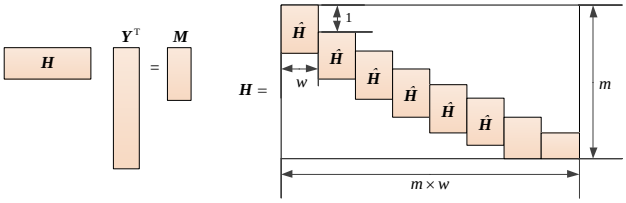



图 1　校验矩阵示意图

Fig.1　Diagram of check matrix

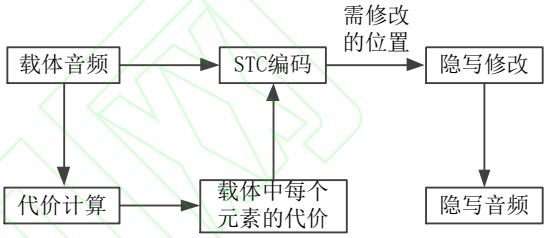



图 2　STC 音频隐写流程

Fig. 2　Flowchart of audio steganography based on STC

## 2　极限梯度提升算法

XGBoost 是[15]一种改进的梯度提升算法，改进点在于采用二阶导数优化目标函数。它融合多个弱分类器进而演化成强分类器，基分类器为分类回归树(Classification and Regression Trees, CART)。

XGBoost 的目标函数由训练损失和正则化项两部分组成，定义如下：

$$Obj=\sum_{i=1}^{n}l(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \tag{6}$$

其中: $f_k$ 为第 $k$ 棵树模型的函数表达式；$y_i$ 为第 $i$ 个样本 $x_i$ 的真实标签; $\hat{y}_i$ 为第 $i$ 个样本 $x_i$ 的预测值。而 XGBoost 是一个加法模型，故预测值为每棵树预测值的累加之和，即 $\hat{y}_i=\sum_{k=1}^{K}f_k(x_i), f_k\in F$。

将 $K$ 棵树的复杂度之和作为正则化项，用于防止模型过拟合。假设第 $t$ 次迭代训练的树模型为 $f_t$，则有：

$$\hat{y}_i^{(t)}=\sum_{k=1}^{t}f_k(x_i)=\hat{y}_i^{(t-1)}+f_t(x_i) \tag{7}$$

将式(7)代入式(6)，得：

$$Obj^{(t)}=\sum_{i=1}^{n}l\left(y_i,\hat{y}_i^{(t)}\right)+\sum_{i=1}^{n}\Omega(f_i)=\sum_{i=1}^{n}l\left(y_i,\hat{y}_i^{(t-1)}+f_t(x_i)\right)+\Omega(f_t) \tag{8}$$

将损失函数 $l$ 进行二阶泰勒展开，得：

$$Obj^{(t)} \cong \sum_{i=1}^{n}\left[l\left(y_i,\hat{y}_i^{(t-1)}+f_t(x_i)\right)\right]= \sum_{i=1}^{n}\left[l\left(y_i,\hat{y}_i^{(t-1)}\right)+g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t) \tag{9}$$

其中：$g_i$ 为损失函数 $l$ 关于 $\hat{y}^{(t-1)}$ 的一阶偏导数；$h_i$ 为损失函数 $l$ 关于 $\hat{y}^{(t-1)}$ 的二阶偏导数；$\Omega(f_t)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2$ 。

记叶子节点 $I_j=\{i\mid q(x_i)=j\}$，目标函数最终简化为，

$$Obj^{(t)}=\sum_{j=1}^{T}\left[\left(\sum_{i\in I_j}g_i\right)w_j+\frac{1}{2}\left(\sum_{i\in I_j}h_i\right)w_j^2\right]+\gamma T \tag{10}$$

对于式(10)，其最小值点和最小值分别为：

$$w_j^{*}=-\frac{\sum_{i\in I_j}g_i}{\left(\sum_{i\in I_j}h_i\right)+\lambda} \tag{11}$$

$$Obj=-\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j}g_i\right)^2}{\left(\sum_{i\in I_j}h_i\right)+\lambda}+\gamma T \tag{12}$$

XGBoost 模型在训练过程中，当建立第 $t$ 棵树时，采用贪心策略进行树结点的分裂。每次分裂成左右两个叶子节点后，会给损失函数带来增益，其定义如下：

$$Gain=Obj_{L+R}-(Obj_L+Obj_R)= \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L}g_i\right)^2}{\left(\sum_{i\in I_L}h_i\right)+\lambda}+\frac{\left(\sum_{i\in I_R}g_i\right)^2}{\left(\sum_{i\in I_R}h_i\right)+\lambda}-\frac{\left(\sum_{i\in I}g_i\right)^2}{\left(\sum_{i\in I}h_i\right)+\lambda}\right]-\gamma \tag{13}$$

若 $Gain>0$，则将此次分裂结果加入模型构建中。

## 3 基于模型融合的音频隐写检测

### 3.1 ResNet-XGBoost 模型总体结构

考虑到卷积神经网络在特征提取上的优势，本文提出了一种基于 ResNet-XGBoost 模型的音频隐写检测方法。该方法思路是利用残差网络提取高维抽象特征，并将其作为 XGBoost 分类器的输入，结合实验结果和网络结构调整得到最终的模型结构如图 3 所示。首先，输入到模型中的音频通过高通滤波器的预处理后，利用第一阶段的三个卷积层进行初步特征提取；接下来，经过五个阶段的残差块和池化操作进一步提取深层次特征；最后，经过全连接层和 Dropout[16] 层将最终提取的特征作为 XGBoost 分类器的输入进行分类，进而输出模型分类结果。

### 3.2 截断线性单元

在神经网络中应用非线性激活函数可增强特征学习能力。到目前为止，Sigmoid、Tanh、修正线性单元(Rectified Linear Unit, ReLU)等非线性激活函数已相继提出，其中尤以 ReLU 在计算机视觉等领域应用较好的泛化能力。ReLU 表达式如下：

$$f(x)=\begin{cases}0, & x<0\\ x, & x\geq 0\end{cases} \tag{14}$$

ReLU 能适应计算机视觉领域被检测目标的数据分布（被检测信号信噪比较高），但在隐写分析任务中由于隐写噪声微弱即被检测信号信噪比低，若采用 ReLU 激活函数，会使得模型不能提取有效的特征。因此，引入了如下的线性截断单元（Truncated Linear Unit，TLU）[17]激活函数。

$$f(x)=\begin{cases}-T, & x<-T\\ x, & -T\leq x\leq T\\ T, & x>T\end{cases} \tag{15}$$

其中，超参数 T 为线性截断阈值，根据实验确定。

本文方法在阶段 1 的第一个卷积层的卷积核大小为 1×1，卷积核个数为 8，且使用了 TLU 激活函数，实验中发现T取 3时，检测效果较好；第二个卷积层的卷积核大小为 ，卷积核个数为 8；第三个卷积层的卷积核大小为 ，卷积核个数为 16。

### 3.3 池化层

为消除特征中的冗余信息和减少来自卷积层的参数个数，通常在卷积层后进行池化操作。池化也可被认为是一种卷积操作，分为最大池化和平均池化。最大池化输出滑动窗口的最大值，用于提取纹理特征；平均池化输出滑动窗口的平均值，用于保留背景信息。根据隐写分析任务的特点和实验效果，本文使用了平均池化，可在一定程度上防止过拟合，为使模型更好地学习到残差，需在第 6 阶段后汇聚高维特征，因此，除了第 6 阶段残差块后使用了全局平均池化，其余阶段均使用平均池化，池化窗口大小为 $1\times 3$，步长为 2。

### 3.4 残差层

残差网络最早由 He 等[18]提出并应用于图像识别，设计了一个多达 152 层的深度卷积神经网络，在 ImageNet 大规模视觉识别挑战赛图像分类任务上识别错误率仅为 3.57%，夺得了图像分类组第一名。残差网络在卷积神经网络的基础上做出了些许改进，有效避免了梯度消失和梯度爆炸的问题，解决了卷积神经网络层数过深时准确率下降的缺陷，残差块基本结构如图 4 所示。假设某段神经网络的输入为 $x$ ，期望输出为 $H(x)$，则在残差块中学习目标为 $F(x)=H(x)-x$，其中 $F(x)$ 为示残差。

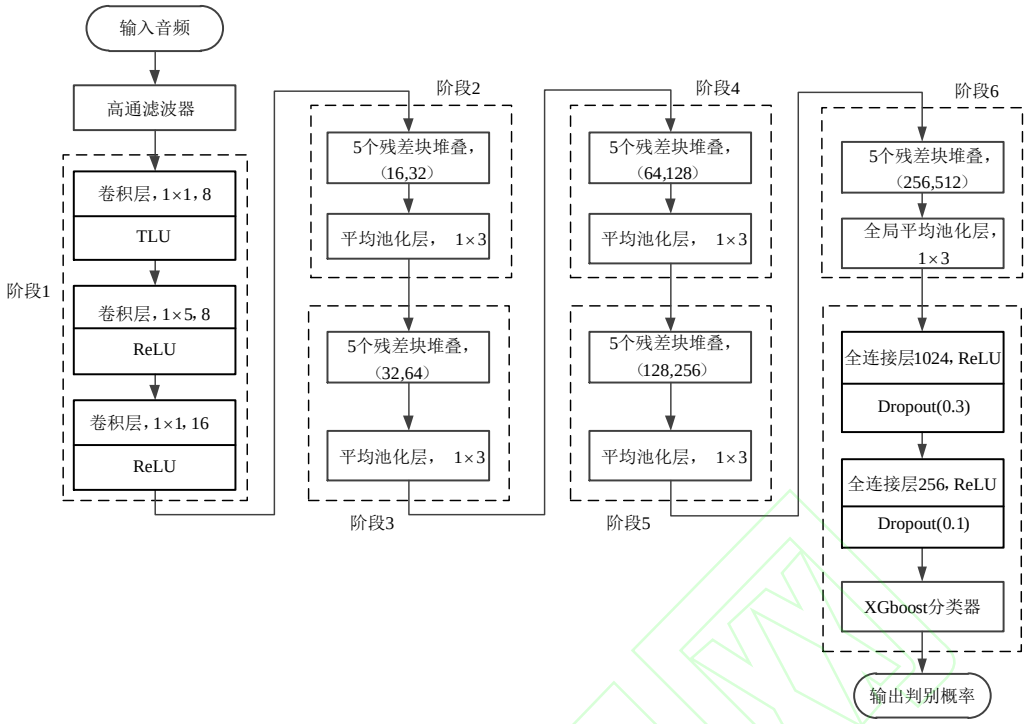



图 3　本文模型总体结构

Fig. 3　Overall structure of proposed model

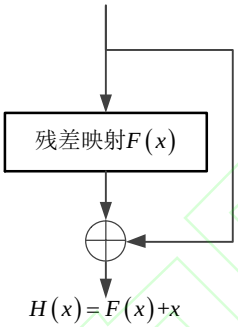



图 4　残差块基本结构

Fig. 4　Basic structure of residual block

将输入音频 $x$ 记为：

$$x=\begin{cases}c+0 & ，原始载体\\ c+m & ，含密载体\end{cases} \tag{16}$$

其中：$c$ 为载体音频；0 代表零信号，即无隐写；$m$ 为嵌入秘密信息后产生的微弱信号。

将 $x$ 输入到残差学习块中，恒等映射块将 $x$ 传输到输出结点，同时，残差学习块 $F(x)$ 将学习到残差 $m$，并将其与输入 $x$ 汇合输出。因此，可将残差 $m$ 类比为隐写噪声，利用残差学习即可学习到隐写噪声的数据分布，使得残差网络非常适合于隐写分析任务。

本文残差网络模型采用的残差块如图 4 所示。残差映射部分由 3 组基本单元构成，其中基本单元包括 1×1 卷积核、1×5 卷积核、批规范化和 tanh 激活函数。在该残差块中，输入维度为 D，经过残差学习后，输出维度为 2D。在本文残差网络模型中，阶段 2~阶段 6 中“5 个残差块堆叠”后面括号内的参数分别表示输入和输出维度，以阶段 2 为例，(16，32) 表示残差块的输入维度为 16，输出维度为 32；每个阶段的残差层由 5 个结构如图 5 所示的残差块堆叠而成，其输入维度和输出维度按图 3 中残差层参数设置。

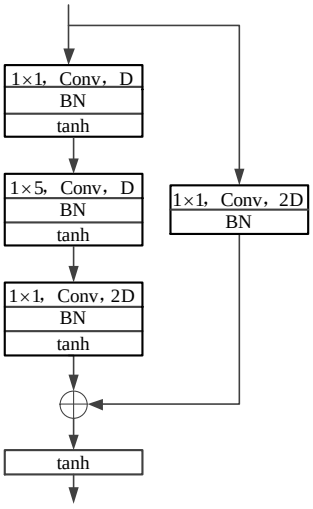



图 5　文中采用的残差块

Fig. 5　Residual block used in the paper

### 3.5　全连接层和 Dropout 层

本文所提出模型的残差网络中有两个全连接层，并且分别在每个全连接层后加入了一个 Dropout[16]层，其中第一个全连接层的神经元个数为 1024，第二个全连接层的神经元个数为 256。

在训练深度神经网络时难免出现过拟合现象，为解决这一问题，通常的思路是采用模型集成，即训练多个模型进行组合，但是训练多个模型会耗费巨大的计算量。因此，Dropout

作为一种典型的正则化手段应运而生，一定程度上能有效防止出现过拟合。通俗地讲，Dropout 就是在神经网络前向传播过程中，以一定概率 $p \in \{0,1\}$ 抑制神经元，在测试时，激活所有神经元，但需将每个神经元的输出乘以 $p$ 。

假设一个神经网络有 $L$ 个隐藏层，$l \in \{1,2,...,L\}$ 为隐藏层数序号，$\boldsymbol{z}^{(l)}$ 为第 $l$ 层的输入向量，$\boldsymbol{y}^{(l)}$ 为第 $l$ 层的输出向量，$\boldsymbol{w}^{(l)}$ 和 $\boldsymbol{b}^{(l)}$ 分别为第 $l$ 权重和偏置。则若在前向传播过程中不使用 Dropout，对隐藏节点 $i$ 有：

$$\boldsymbol{z}_i^{(l+1)} = \boldsymbol{w}_i^{(l+1)} \boldsymbol{y}^{l} + \boldsymbol{b}_i^{(l+1)} \tag{17}$$

$$\boldsymbol{y}_i^{(l+1)} = f\left(\boldsymbol{z}_i^{(l+1)}\right) \tag{18}$$

其中，$l \in \{0,1,..., L-1\}$，$f$ 为激活函数。

若在前向传播过程中使用 Dropout，对隐藏节点 $i$ 有：

$$r_j^{(l)} \sim Bernoulli\left(p\right) \tag{19}$$

$$\tilde{\boldsymbol{y}}^{(l)} = r^{(l)} * \boldsymbol{y}^{(l)} \tag{20}$$

$$\boldsymbol{z}_i^{(l+1)} = \boldsymbol{w}_i^{(l+1)} \tilde{\boldsymbol{y}}^{(l)} + \boldsymbol{b}_i^{(l+1)} \tag{21}$$

$$\boldsymbol{y}_i^{(l+1)} = f\left(\boldsymbol{z}_i^{(l+1)}\right) \tag{22}$$

其中：*Bernoulli* 函数的作用是以概率 $p$ 随机生成一个由 0、1 组成的向量；$\boldsymbol{y}^{(l)}$ 为第 $l$ 层隐藏层的输出。

实验中通过多次调参发现，第一个 Dropout 层抑制概率取 0.3，第二个 Dropout 层抑制概率取 0.1 时，提取的特征输入到 XGBoost 分类器中的检测效果最好。

### 3.6 XGBoost 分类器

本文方法的思路是先利用残差网络提取高维抽象特征，再将提取到的输入到 XGBoost 分类器中进行分类。XGBoost 分类器融合多个分类回归树进行分类，具有较好的泛化能力，其参数设置显得尤为重要。实验中多次调参得出 XGBoost 分类器最优参数组合：学习率取 0.1，树的深度取 150，训练轮数取 1000，其余均为默认的参数，此时检测性能可达到最优。在 XGBoost 模型训练过程中，使用了 logistic 和 hinge 两种损失函数，其表达式分别为：

$$l\left(z\right) = \frac{1}{1+\mathrm{e}^{-z_1}} \tag{23}$$

$$l\left(z\right) = \max\left(0, 1-z_2\right) \tag{24}$$

其中，$z_1, z_2$ 均为分类器的预测输出。

对于 XGBoost 分类器来讲，使用 logistic 损失函数时，分类器输出给定输入样本属于正类的概率；使用 hinge 损失函数时，分类器输出给定输入样本的分类结果。在实验中发现，在保持除了损失函数外的所有参数不变的条件下，损失函数为 hinge 时的分类准确率明显高于 logistic。

## 4 实验与结果分析

### 4.1 实验设置

实验数据集来源于 TIMIT 语料库[19]，其中包含 6300 个采样频率为 16kHz、量化位数为 16 的无压缩单声道音频，将每段语音切割成时长为 1s 的音频片段，选取其中 15000 个作为用于隐写的原始音频，并分别利用 LSBM 和 STC 隐写算法嵌入密信，从而生成 15000 个嵌密音频。

实验时将 15000 个原始音频和 15000 个嵌密音频组合在一起，合计 30000 个音频片段，其中，24000 个为训练集，6000 个用于验证与测试。残差网络模型的训练批次大小为 64，输入到 XGBoost 分类器中的高维抽象特征维度是 256。实验硬件环境是内存大小为 11GB 的 NVDIA GTX1080Ti GPU，软件环境是 Tensorflow 和 Keras 深度学习框架。

### 4.2 模型微调对实验结果的影响

#### 4.2.1 高通滤波器阶数对检测准确率的影响

本文采用的高通滤波器为差分滤波器[4]，在训练过程中发现,将其参数设为固定时检测效果较好。隐写相当于向原始载体中添加了微弱的噪声，需比较嵌入密信前后的载体差异即计算残差，以此提高信噪比，使得特征学习更充分。为研究滤波器阶数对检测准确率的影响，分别在嵌入率为 0.5bps 时对 LSBM 和 STC 进行检测，实验结果如图 6 所示。

实验结果表明，高通滤波器阶数取 6 时，隐写检测准确率最高，当滤波器阶数大于 6 时，隐写检测准确率反而下降。6 阶高通滤波器参数设置如下：

$F_1$=[1,-1, 0,  0,  0, 0,0];
$F_2$=[1,-2, 1,  0,  0, 0,0];
$F_3$=[1,-3, 3, -1,  0, 0,0];
$F_4$=[1,-4, 6, -4,  1, 0,0];
$F_5$=[1,-5,10,-10, 5,-1,0];
$F_6$=[1,-6,15,-20,15,-6,1].

#### 4.2.2 模型结构对检测准确率的影响

通过修改用于提取特征的残差网络的不同结构参数，比较每种修改情况下对 LSBM 隐写算法的检测准确率，其中 LSBM 嵌入率为 0.5bps。修改的网络结构参数如表 1 所示。为比较修改不同的网络结构参数后模型检测准确率的波动范围，对于每个不同的模型进行 10 次实验，实验结果如图 7 的箱形图所示。本文模型平均检测准确率为 86.58%，与其他 9 种模型相比，检测准确率最高且较稳定；模型#2 的平均检测准确率为 83.48%，原因可能是移除高通滤波器后，模型不能有效学习到残差；模型#3 的平均检测准确率为 82.59%，原因可能是移除 TLU 激活函数后，模型提取的特征数据缺失规范化；模型#4 和#5 的检测准确率介于 81.12%和 86.85%之间且波动范围较大；模型#6 的检测准确率在 75.12%和 87.25%之间，有时检测准确率高于模型#1，但有时检测准确率陷入局部最小值；模型#7 的检测准确率波动范围小，但平均检测

准确率仅为 79.14%，原因是梯度消失；模型#8 的检测平均检测准确率为 85.6%，原因是可能发生过拟合；模型#9 的平均检测准确率为 74.72%，在 9 种模型中是最低的，恰恰印证了 XGBoost 分类器的有效性。XGBoost 分类器能提高检测准确率的一个重要原因是，XGBoost 模型在训练过程中，不断优化当前决策树，并融合多颗当前最优决策树进行分类，是集成学习的一种体现。综上所述，本文模型收敛快，检测准确率最高，因此，相较于其他 8 种模型，本文模型是最有效的。

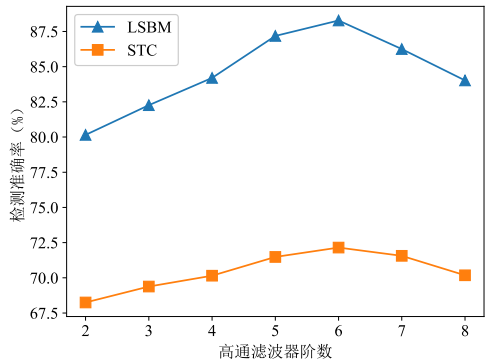



图 6　高通滤波器不同阶数对应的检测准确率

Fig. 6　Detection accuracy corresponding to different orders of high-pass filter

表 1　修改的网络结构参数

Tab. 1 Modified network structure parameters

| 序号 | 网络结构参数 |
|---|---|
| #1 | 本文模型 |
| #2 | 移除高通滤波层 |
| #3 | 移除 TLU 激活函数 |
| #4 | 第一阶段最末尾卷积层后加上平均池化 |
| #5 | 将残差块中的激活函数 Tanh 换成 ReLU |
| #6 | 网络各阶段均使用最大池化 |
| #7 | 移除残差块的短连接 |
| #8 | 移除 Dropout 层 |
| #9 | 移除 XGBoost 分类器 |

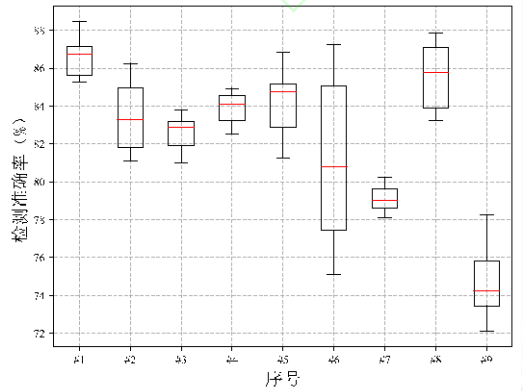



图 7　不同模型对嵌入率为 0.5bps 的 LSBM 隐写检测准确率的箱形图

Fig. 7　Box plots of the detection accuracy acquired from different models for detection LSBM steganography with 0.5bps

### 4.3　结果对比分析

为进一步证实本文方法的优势所在，对比了文献[3]、文献[4]、文献[20]分别提出的传统基于手工特征的音频隐写检测方法和文献[11]提出的深度学习检测方法对 LSBM 和 STC 在不同嵌入率下的平均检测准确率，其中平均检测准确率为 10 次重复实验结果取平均值。对于 STC 嵌密，为研究嵌入率相同即子校验矩阵宽度相同而子校验矩阵高度不同对隐写检测准确率的影响，由于在 STC 隐写中，子校验矩阵高度取值范围为 $6 \le h \le 15$，为简化实验，分别在子校验矩阵高度为 7、8和9 时进行嵌密。实验结果如表 2~表 5 所示。

从表 2 可知，本文方法对三种不同嵌入率下的 LSBM 隐写算法检测准确率相较于文献[3]、文献[4]、文献[20]和文献[11]的方法有明显提升。本文方法对 0.5bps 的 LSBM 隐写检测准确率比文献[3]、文献[4]、文献[20]和文献[11]的方法分别提高了 28.46 个百分点、16.23 个百分点、15.22 个百分点、10.16 个百分点。

表 3~表 5 显示了不同方法对子校验矩阵高度分别为 7、8 和 9 时的 STC 隐写算法检测准确率，可知本文方法对不同嵌入率下的 STC 隐写算法检测准确率明显文献[3]、文献[4]、文献[20]、文献[11]的方法，且已达到了目前最好的效果。以 STC 隐写子校验矩阵高度 $h=7$ 时为例，本文方法对 0.5bps 的 STC 隐写算法检测准确率比文献[3]、文献[4]、文献[20]和文献[11]的方法分别提高了 22 个百分点以上、12.75 个百分点、12.25 个百分点、11.07 个百分点。文献[3]方法不能检测 STC 隐写算法的一个重要原因是，STC 隐写会使得嵌入密信后引起的失真最小即隐写噪声极其微弱，以此保证 STC 隐写具有很高的隐蔽性，使得基于传统手工特征的音频隐写检测方法很难提取到有效的特征。通过表 3~表 5 的实验结果可知，就 STC 隐写而言，子校验矩阵高度越高，安全性越高，即越抗隐写检测。

同时，对于 0.5bps 的 LSBM 隐写和 STC 隐写（子校验矩阵高度 $h=7$ ）检测，文献[11]方法和本文方法的 ROC 曲线如图 8 所示。对于图中的 ROC 曲线，横轴表示假阳率 FPR，亦即虚警率；横轴表示真阳率 TPR，亦即检测率；虚线表示随机猜测，判断载体隐写和未隐写的概率均为 50%。ROC 曲线与坐标轴围成图形的面积 AUC 表征了检测器对此类隐写对象的适用性程度。由图 8 可知，本文方法在两种隐写算法下的 AUC 均大于文献[11]方法，即本文检测器性能更优相较于文献[11]方法。

表 2　不同方法对 LSBM 隐写检测准确率　单位：%

Tab. 2　Detection accuracy on LSBM steganography with different methods　unit：%

| 检测 | 嵌入率（bps） |
|---|---|

| 方法 | 0.5 | 0.2 | 0.1 |
|---|---|---|---|
| 文献[3]方法 | 58.12 | 低于 51 | 低于 51 |
| 文献[4]方法 | 70.35 | 65.26 | 58.64 |
| 文献[20]方法 | 71.36 | 64.18 | 60.72 |
| 文献[11]方法 | 76.42 | 67.24 | 62.31 |
| 本文方法 | **86.58** | **76.08** | **72.82** |

表 3 不同方法对 h=7 时的 STC 隐写检测准确率 单位：%

Tab. 3 Detection accuracy on STC steganography with different methods when h equals to 7 unit：%

| 检测方法 | 嵌入率（bps） | | |
|---|---|---|---|
| | 0.5 | 0.2 | 0.1 |
| 文献[3]方法 | 低于 51 | 低于 51 | 低于 51 |
| 文献[4]方法 | 60.52 | 56.08 | 低于 51 |
| 文献[20]方法 | 61.02 | 57.32 | 51.86 |
| 文献[11]方法 | 62.20 | 58.24 | 52.42 |
| 本文方法 | **73.27** | **70.16** | **65.18** |

表 4 不同方法对 h=8 时的 STC 隐写检测准确率 单位：%

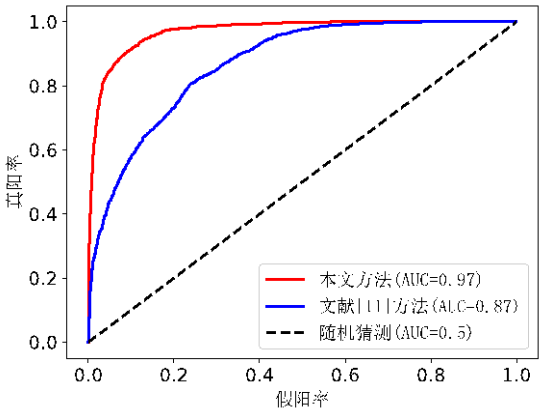
Tab.4 Detection accuracy on STC steganography with different methods when h equals to 8 unit：%

| 检测方法 | 嵌入率（bps） | | |
|---|---|---|---|
| | 0.5 | 0.2 | 0.1 |
| 文献[3]方法 | 低于 51 | 低于 51 | 低于 51 |
| 文献[4]方法 | 57.36 | 52.15 | 低于 51 |
| 文献[20]方法 | 58.19 | 53.09 | 低于 51 |
| 文献[11]方法 | 60.13 | 55.24 | 低于 51 |
| 本文方法 | **70.35** | **68.02** | **65.47** |

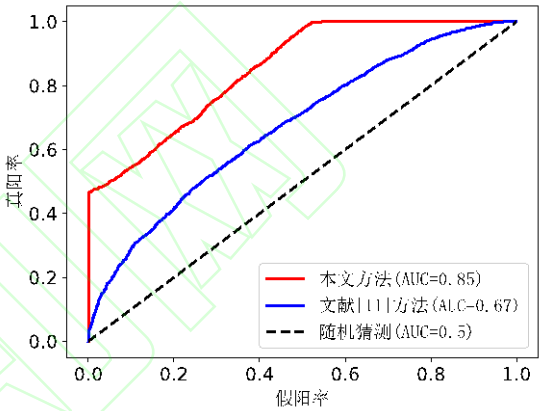
表 5 不同方法对 h=9 时的 STC 隐写检测准确率 单位：%

Tab. 5 Detection accuracy on STC steganogaphy with different methods when h equals to 9 unit：%

| 检测方法 | 嵌入率（bps） | | |
|---|---|---|---|
| | 0.5 | 0.2 | 0.1 |
| 文献[3]方法 | 低于 51 | 低于 51 | 低于 51 |
| 文献[4]方法 | 53.05 | 低于 51 | 低于 51 |
| 文献[20]方法 | 54.32 | 低于 51 | 低于 51 |
| 文献[11]方法 | 55.26 | 51.03 | 低于 51 |
| 本文方法 | **69.15** | **66.71** | **62.08** |




(a) 两种深度学习检测方法对 LSBM 隐写的 ROC 曲线




(b) 两种深度学习检测方法对 STC 隐写的 ROC 曲线

图 8 两种深度学习检测方法对 LSBM 隐写和 STC 隐写的 ROC 曲线

Fig. 8 ROC curves for detecting LSBM steganography and STC steganography by two deep learning detection methods

## 5 结语

针对现有隐写检测方法对 STC 音频隐写算法难以检测的问题，本文提出了一种融合深度残差网络和极限梯度提升的音频隐写检测模型。该方法首先构建了一种用于提取高维抽象特征的深度残差网络模型，再将利用其提取的特征作为 XGBoost 模型的输入进行分类。实验中比较了本文方法和其他四种经典音频隐写检测方法对 LSBM 隐写算法和 STC 隐写算法的检测准确率，实验结果表明，相较于传统提取手工特征的检测方法和基于深度学习的检测方法，本文方法检测准确率最高且有明显提升，STC 隐写中子校验矩阵高度较高时的检测准确率低于子校验矩阵高度较低时，证明了子校验矩阵高度对隐写检测准确率有直接影响，子校验矩阵高度越高，隐写算法越难以检测。同时，本文方法采用深度残差网络提取高维抽象特征并用 XGBoost 模型分类，相较传统的隐写检测方法通常集中于依靠经验设计手工特征，本文方法则从一个新的角度出发提取特征，避免了耗费大量精力着眼于复杂的手工特征设计上。

后续工作将研究基于 STC 框架的自适应音频隐写的检测算法，考虑到自适应隐写嵌密集中于复杂区域上，研究重点则聚焦于隐写信号的提取和放大，然后采用本文方法进行隐写检测。

## 参考文献


[1] JOHNSON M K, LYU S, and FARID H. Steganalysis of recorded speech [C]// Proceedings of the 2005 International Society for Optics and Photonics. San Jose: SPIE, 2005: 664-672.

[2] KRAETZER C, DITTMANN J. Mel-cepstrum-based steganalysis for VoIP steganography [C]// Proceedings of the 2007 International Society for Optics and Photonics. Bellingham: SPIE, 2007: Article No. 650505.

[3] LIU Q, SUNG A H, and QIAO M. Temporal derivative-based spectrum and mel-cepstrum audio steganalysis [J]. IEEE Transactions on Information Forensics and Security, 2009, 4(3): 359-368.

[4] LIU Q, SUNG H A, and QIAO M. Derivative-based audio steganalysis [J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2011, 7(3): 18:1-18:19.

[5] GEETHA S, ISHWARYA N, KAMARAJ N. Audio steganalysis with Hausdorff distance higher order statistics using a rule based decision tree paradigm [J]. Export Systems with Applications, 2010, 37(12): 7469-7482.

[6] 王昱洁，杨萍，蒋薇薇. 一种基于 MDCT 量化系数统计特征的 AAC 音频隐写分析方法[J]. 合肥工业大学学报, 2015, 38, (10)：1348-1352. (WANG Y J, YANG P, JIANG W W. A steganalysis method of AAC audio based on statistical features of MDCT quantized coefficients [J]. Journal of Hefei University of Technology, 2015, 38(10)：1348-1352.)

[7] 王昱洁，蒋薇薇. 基于模糊 C 均值聚类与单类支持向量机的音频隐写分析方法[J]. 计算机应用，2016, 36(3): 647-652.(WANG Y J, JIANG W W. Audio steganalysis method based on fuzzy-C means clustering and one class support vector machine[J]. Journal of Computer Applications, 2016, 36(3): 647-652.)

[8] HAN C, XUE R, ZHANG R, et al. A new audio steganalysis method based on linear prediction [J]. Multimedia Tools and Applications, 2017, 77: 15431-15455.

[9] REN Y, XIONG Q, WANG L. A steganalysis scheme for AAC audio based on MDCT difference between intra and inter frame [C]// Proceedings of the 2017 International Workshop on Digital Watermarking. Berlin: Springer, 2017:217-231.

[10] CHEN B, LUO W, LI H. Audio steganalysis with convolutional neural network [C]// Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2017: 85-90.

[11] LIN Y, WANG R, YAN D, et al. Audio steganalysis with improved convolutional neural network [C]// Proceedings of the 7th ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2019: 210-215.

[12] WANG Y, YANG K, YI X, et al. CNN-based steganalysis of MP3 steganography in entropy code domain [C]// Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2018: 55-65.

[13] YANG H, YANG Z, HUANG Y. Steganalysis of VoIP streams with CNN-LSTM network [C]// Proceedings of the 7th ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2019: 204-209.

[14] FILLER T, JUDAS J, FRIDRICH J. Minimizing additive distortion in steganography using syndrome-trellis codes [J]. IEEE Transactions on Information Forensics and Security, 2011, 6(3): 920-935.

[15] CHEN T, GUESTRIN C. XGBoost: a scalable tree boosting system [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 785-794.

[16] 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: 1929-1958.

[17] YE J, NI J, YI Y. Deep Learning Hierarchical Representations for image steganalysis [J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2545-2557.

[18] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.

[19] GAROFOLO J S, LAMEL L F, FISHER W M, et al. DARPA TIMIT: acoustic-phonetic continuous speech corpus [R]. Gaithersburg, MD: National Institute of Standards and Technology, 1993.

[20] LUO W, LI H , YAN Q, et al. Improved audio steganalytic feature and its applications in audio forensics [C]// Proceedings of the 2018 ACM Transactions on Multimedia Computing, Communications, and Applications. New York: ACM, 2018: Article No. 43.


This work is partially supported by the National Natural Science Foundation of China (U1736215, 61672302, 61901237), the Natural Science Foundation of Zhejiang Province (LY20F020010, LY17F020010), the Open Foundation of the Mobile Network Application Technology Key Laboratory of Zhejiang Province (F2018001), the K.C. Wong Magna Fund in Ningbo University, the Scientific Research Foundation of Graduate School of Ningbo University (IF2020131).

**CHEN Lang**, born in 1997, M.S. candidate. His research interests include multimedia information security, information hiding, steganalysis.

**WANG Rangding**, born in 1962, Ph. D., professor. His research interests include multimedia information security, information hiding, steganalysis.

**YAN Diqun**, born in 1979, Ph. D., associate professor. His research interests include multimedia information security, digital forensics.

**LIN Yuzhen**, born in 1994, M.S. candidate. His research interests include multimedia information security, steganalysis.