计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20190386
56(10): 2250-2261, 2019

基于卷积神经网络的JPEG图像隐写分析参照图像生成方法

任魏翔　翟黎明　王丽娜　嘉　炬
(空天信息安全与可信计算教育部重点实验室(武汉大学)　武汉　430072)
(武汉大学国家网络安全学院　武汉　430072)
(renweixiang@whu.edu.cn)

Reference Image Generation Algorithm for JPEG Image Steganalysis Based on Convolutional Neural Network

Ren Weixiang, Zhai Liming, Wang Lina, and Jia Ju
(*Key Laboratory of Aerospace Information Security and Trusted Computing* (*Wuhan University*), *Ministry of Education*, *Wuhan* 430072)
(*School of Cyber Science and Engineering*, *Wuhan University*, *Wuhan* 430072)

Abstract　As the opponent of image steganography, the image steganalysis is to detect the secret message in images concealed by steganography algorithms. Recently, state-of-the-art JPEG image steganalysis schemes are changing from complex handcrafted feature-based ones to deep learning-based ones. Although the deep learning steganalysis for detecting JPEG steganography achieves great advancement, there still exists room for improvement. As it is verified that side information could promote the steganography detection accuracy, we seek the method to further improve the accuracy of content-adaptive steganography detection in JPEG domain from the perspective of side information offering for the deep learning steganalysis scheme. The proposed method utilizes convolutional neural networks to generate reference images from the input data. And the reference image is treated as the side information for the deep learning-based JPEG image steganalysis model. The proposed method can be pre-trained or trained together with the steganalysis model. Experimental results on classic content-adaptive steganography algorithms in JPEG domain named J-UNIWARD and JC-UED verifies the proposed method could enhance the detection ability compared with the deep learning steganalysis model without the aid of the proposed method to a certain extent. The proposed method could boost the detection accuracy for deep learning-based JPEG steganalysis model by 6 percentage points at most.

Key words　JPEG steganalysis; side information; reference image; convolution neural network (CNN); JPEG adaptive steganography

收稿日期:2019-06-11;**修回日期**:2019-07-31

基金项目:国家自然科学基金重点项目(U1536204);NSFC-通用技术基础研究联合基金项目(U1836112);国家自然科学基金项目(61876134,61872275)

This work was supported by the Key Program of the National Natural Science Foundation of China (U1536204), the United Basic Research Foundation of NSFC-General Technology (U1836112), and the National Natural Science Foundation of China (61876134, 61872275).

通信作者:王丽娜(lnwang@whu.edu.cn)

摘 要 基于深度学习的 JPEG 数字图像隐写分析模型检测能力已超越基于人工设计特征隐写分析模型,但检测能力仍存在提升空间.以进一步提升 JPEG 隐写分析模型的检测能力为目标,借助深度学习方法,为基于深度学习的 JPEG 隐写分析模型提供辅助信息,从数据输入角度,探索进一步提升隐写分析模型检测能力的途径.基于卷积神经网络,构建隐写分析参照图像生成模型,对待检测图像进行变换,从而获得对应参照图像.之后,将待检测图像与对应参照图像作为隐写分析模型的输入数据,进一步挖掘待检测图像中存在的隐写分析相关信息.为验证所提出算法的有效性,进行针对 JPEG 自适应隐写算法的对比实验.实验结果表明:所设计的参照图像生成模型能够提升现有基于深度学习的隐写分析模型检测能力,提升效果最多可达 6 个百分点.

关键词 JPEG 隐写分析;辅助信息;参照图像;卷积神经网络;JPEG 自适应隐写算法

中图法分类号 TP309.7

随着信息技术的飞速发展,多媒体文件在网络中广泛传输,导致多媒体信息安全问题日益严重.信息隐藏的研究成果对解决多媒体信息安全问题具有重要作用.

信息隐藏领域的研究可分为隐写(steganography)、隐写分析(steganalysis)两方面.如图 1 所示,隐写是将有意义的信息隐藏在另一个称为载体 C(cover)的信息中得到隐蔽载体 S(stego)的技术[1].隐写分析是隐写术的对抗技术,对可疑的载体信息进行攻击,达到检测、破坏、甚至提取秘密信息的技术[1].

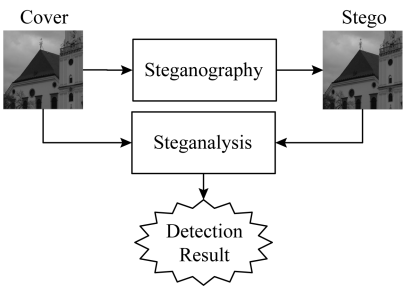


Fig. 1 Steganography and steganalysis

图 1 隐写与隐写分析

针对数字图像的隐写术通过对图像中的像素或者 DCT 系数等图像元素进行微量的修改,从而实现在图像中隐藏秘密信息的目的.然而隐写过程中对图像元素的修改操作会对图像元素值的统计分布造成一定的扰动.因此,当前数字图像隐写算法主要通过合理地选择对图像进行修改的位置[2],以及尽量减少所需修改的图像元素[3-4]来达到最小化隐写操作对原始图像所造成影响的目的.当前具有较高安全性的数字图像隐写算法主要为内容自适应隐写算法,该类算法能够有选择地对图像元素进行修改,以减小秘密信息嵌入过程对图像造成影响.该类算法设计过程可分为失真代价函数设计和隐写码的设计 2 部分.其中失真代价函数用于衡量秘密信息的嵌入过程中对原始图像载体元素的统计规律造成的影响.而隐写码是指在基本嵌入的基础上,为了减小隐写嵌入失真(distortion)(亦称代价)而进行的编码[5].在 STC 隐写码[6]提出后,面向数字图像的隐写算法的设计工作主要集中在失真代价函数的设计[7-13].然而,隐写的安全性不仅取决于隐写算法、嵌入率等因素,而且在很大程度上收到隐写过程中载体图像自身属性的影响[14].

当前通用图像隐写分析方法的相关研究可分为基于隐写分析特征的检测方法,以及基于深度学习的检测方法.前者通过构建能够描述图像载体统计规律或特性,并且对隐写嵌入操作敏感的隐写分析特征,结合机器学习方法对载体进行检测[15-19].其中,隐写分析特征依据对图像载体.后者基于深度学习技术,构建用于执行隐写分析任务的深度学习模型,以实现隐写信号的端到端检测.目前,基于深度学习的隐写分析相关研究已成信息隐藏领域的热点研究问题,并取得了较好的成果[20-29].

选择信道感知的隐写分析方法以提升针对自适应图像隐写算法检测效果为目标,利用自适应隐写算法对图像各元素修改的概率,将检测重点集中在更易被隐写算法修改的图像区域,从而提升针对自适应隐写算法的检测能力.其中,待检测图像各元素被特定自适应隐写算法修改的概率可视为隐写分析模型的辅助信息.当前选择信道感知的思想已运用于基于特征的隐写分析模型,以及基于深度学习的隐写分析模型.

由于 JPEG 图像在互联网中的广泛应用,因此针对 JPEG 图像的研究具有较高的实际意义.当前基于深度学习的 JPEG 图像隐写分析的研究工作主

要集中于负责隐写检测的深度学习模型的相关研究[26-29].基于深度学习隐写分析模型辅助信息的产生方式主要基于研究者专业经验进行设计[23-24].本文借助深度学习方法,从隐写分析辅助信息产生方式的角度,探索提升针对 JPEG 图像的深度学习隐写分析模型检测效果的新途径.

本文构建具有卷积层和反卷积层的 16 层卷积神经网络,对待检测图像进行变换,得到待检测图像所对应的参照图像,将参照图像和待检测图像一同作为 JPEG 图像隐写分析模型的输入数据,从而基于现有针对 JPEG 图像深度学习隐写分析模型,进一步提升 JPEG 图像深度学习隐写分析模型检测能力.

本文的主要贡献有 3 个方面:

1) 提出针对 JPEG 图像隐写分析的参照图像生成模型,为针对 JPEG 图像的深度学习隐写分析模型提供辅助信息,从辅助信息的角度探索提升深度学习隐写分析模型检测能力的途径;

2) 隐写分析参照图像生成模型基于深度卷积神经网络,能够通过训练,学习有利于提升隐写检测能力的辅助信息生成方式;

3) 针对多种嵌入率、隐写算法的对比实验结果表明,所提出的针对 JPEG 图像隐写分析的参照图像生成模型生成的参照图像能够提升针对 JPEG 图像的深度学习隐写分析模型的检测能力.

1 相关工作

由于 JPEG 图像在互联网中的广泛使用,针对 JPEG 图像的隐写分析具有较高的研究价值.因此,基于深度学习的 JPEG 隐写分析方法也是隐写分析研究领域的热点问题.

在深度学习隐写分析开始引起学者关注期间,Xu[26],Chen 等人[27],Zeng 等人[28]分别提出了针对 JPEG 图像的深度学习隐写分析模型.其中,Xu[26]构建了具有 20 层的深度卷积网络模型,该模型为减少信息丢失,不使用池化操作,并加入跳转链接(shortcut connection),以防止梯度消失现象发生.Chen 等人[27]构建的深度学习模型包含 64 个卷积神经网络,其中每个神经网络与 JPEG 图像的 1 个相位相对应,从而实现了网络模型对 JPEG 图像不同相位数据的分离处理.Zeng 等人[28]构建了 1 种针对 JPEG 隐写分析的混合卷积经网络模型.该模型首先利用 25 个 DCT 变换基对待检测图像进行预处理,之后对预处理结果进行多种不同的量化和截断处理,并将处理的结果分别作为不同深度卷积神经网络的输入.Yang 等人[29]基于 DenseNet[30]提出了针对 JPEG 图像的深度学习隐写分析模型.

1.1 JPEG 图像深度学习隐写分析处理过程

卷积神经网络(convolutional neural network, CNN)能够提取图像中相邻元素之间存在的相关性.因此,当前针对 JPEG 图像的深度学习隐写分析模型主要基于卷积神经网络,并将 JPEG 图像的解压(不取整)结果作为输入数据.除此之外,不同针对 JPEG 图像的深度学习隐写分析模型在卷积神经网络部分都采用了卷积层(convolutional layer)、批标准化(batch normalization)、ReLU(rectified linear unit)激活函数、平均池化(avarage pooling)等操作.

卷积神经网络中的卷积层由多个卷积核构成,各卷积核的有卷积权重和偏置组成,卷积权重和偏置参与神经网络的训练.卷积层中的各卷积核分别利用卷积权重对输入数据进行卷积操作,并将结果与对应的偏置相加,之后将卷积核的处理结果进行合并,合并结果作为该卷积层的输出.具体处理过程:

$$output^l = output^{l-1} * \boldsymbol{W}_l + \boldsymbol{B}_l, \tag{1}$$

其中,$output^l$ 和 $output^{l-1}$ 分别为第 l 个卷积层的输出数据和输入数据;$\boldsymbol{W}_l$ 为第 l 个卷积层的卷积核参数矩阵;$*$ 为卷积操作算子;$\boldsymbol{B}_l$ 为第 l 个卷积层的偏置.

为了提升深度学习隐写分析模型在训练过程中的收敛速度,采用批标准化(batch normalization, BN)层[31],对卷积层输出的特征图进行标准化处理.BN 层的处理过程:

$$u = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{2}$$

$$v = \frac{1}{n}\sum_{i=1}^{n} (X_i - u)^2, \tag{3}$$

$$\hat{X}_i = \frac{X_i - u}{\sqrt{v+\varepsilon}}, \tag{4}$$

$$y_i = \gamma * \hat{X}_i + \beta, \tag{5}$$

其中,γ 和 β 为可训练参数,参与神经网络的训练;ε 参数为大于 0 的常量;$\hat{X}_i$ 为当前批处理层的最小批量输入数据中的 1 个特征图.

深度神经网络中的激活函数保证了深度学习隐写分析模型进行非线性特性的学习,而 ReLU(rectified linear unit, ReLU)激活函数在深度学习模型中得到了广泛的应用.ReLU 激活函数的具体处理过程为

$$f(x) = \begin{cases} x, & x > 0, \\ 0, & x \leqslant 0, \end{cases} \tag{6}$$

其中,x 为 ReLU 激活函数的输入;$f(x)$为其输出结果.

平均池化属于 1 种下采样操作,对矩阵 $\boldsymbol{X}$ 进行平均池化操作具体过程:

$$avg\_pooling(\boldsymbol{X})=\frac{1}{n^2}(\boldsymbol{K}*\boldsymbol{X}), \tag{7}$$

其中,$\boldsymbol{X}$ 为池化操作待处理的矩阵;池化操作的窗口大小为 $n\times n$;$\boldsymbol{K}$ 为大小为 $n\times n$,并且元素全为 1 的矩阵,"$*$"为卷积操作算子.

相对于最大池化等其他池化操作,能够保留隐写算法的嵌入操作在载体中引入的微弱扰动.因此,基于深度学习的隐写分析模型主要采用平均池化作为其池化操作.

1.2 深度学习隐写分析辅助信息产生方式

为进一步提升针对 JPEG 图像的深度学习隐写分析模型检测能力,Yang 等人[23]和 Ye 等人[24]提出的深度学习隐写分析模型在输入数据的基础上,附加以相应的辅助信息.当前基于深度学习的隐写分析模型的辅助信息主要来源于待检测图像各元素被自适应隐写算法修改的概率.

如图 2 所示,图像中位于纹理复杂区域的元素对应的大于位于平滑区域元素的修改概率.以上现象表明自适应隐写算法更倾向于在图像中的纹理复杂区域进行嵌入操作.

产生以上现象的原因为图像纹理复杂区域的元素值的统计分布相比于平滑区域更为复杂,隐写算法在这些统计规律复杂区域造成的扰动更加不易被觉察.因此,对隐写算法修改概率更高的区域进行更有侧重地检测,有利于提升隐写检测效果待检测图像的元素被自适应隐写算法修改概率可利用损失函数计算得到.具体计算方式为

$$\beta_{i,j}=\frac{1}{2+\mathrm{e}^{\lambda\rho_{i,j}}}, \tag{8}$$

其中,$\beta_{i,j}$ 为位于图像位置(i,j)的元素被修改的概率值;λ 为大于零的数值,具体数值由隐写相对嵌入率决定.

(a) Original image (b) Embedding probability map

Fig. 2 An image and the corresponding embedding probability map

图 2 一张图像及其对应的修改概率图

2 JPEG 图像隐写分析参照图像生成模型

虽然利用图像各元素被修改的概率作为辅助信息,能够提升针对 JPEG 图像的深度学习隐写分析模型的检测能力,但当前基于深度学习隐写分析模型所使用辅助信息的产生方法有待进一步改进,隐写分析模型的输入数据中隐含的与隐写分析有关的信息有待进一步挖掘.

本文从隐写分析辅助信息的角度,探索进一步提升深度学习隐写分析模型检测能力的途径,提出 JPEG 图像隐写分析参照图像生成模型,尝试更加深入的挖掘待检测图像中与隐写分析有关的信息.首先,基于 U-NET[32]构建具有跳转连接(skip connection)的深度卷积神经网络,对待检测图像进行处理,生成用于 JPEG 图像隐写分析的参照图像.之后,将生成的参照图像作为隐写分析的辅助信息,与隐写分析模型相结合.从而更加充分地挖掘待检测图像中与隐写有关的信息,在现有隐写分析模型的基础上,进一步提升深度学习隐写分析模型的检测能力的目的.

2.1 JPEG 图像隐写分析模型的关系

JPEG 图像隐写分析参照图像生成模型将待检测图像进行处理和变换,以提升隐写分析模型的检测能力为目的,为针对 JPEG 图像的深度学习隐写分析模型提供辅助信息.

如图 3 所示,JPEG 图像隐写分析参照图像生

成模型与隐写分析模型相互独立,以待检测图像为输入数据.而隐写分析模型的输入数据为待检测图像和 JPEG 图像隐写分析参照图像生成模型的输出数据.最终检测结果由隐写分析模型输出.

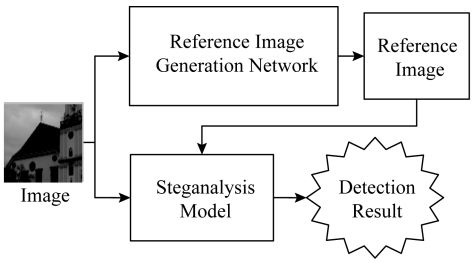


Fig. 3 Relations between the proposed method and steganalysis model

图 3 本文方法与隐写分析模型的关系

2.2 所提出参照图像生成模型结构

JPEG 图像隐写分析参照图像生成模型,由 8 个卷积层与 8 个反卷积层(deconvolution)组成,共 16 层.每个卷积层以及反卷积层的输出依次经过批标准化(BN)和 LReLu(leaky ReLu, LReLU)激活函数的处理,分别记为 CONV-BN-LReLU 和 deCONV-BN-LReLU.

本文提出模型的反卷积层又称为转置卷积,记为 deCONV.本文模型通过反卷积层实现上采样目的,将经由多个卷积层输出的抽象特征图进行进一步变换,最终得到与输入图像相同大小的参照图像.此外,为抑制梯度消失现象,在模型的卷积层与对称位置的反卷积层之间添加跳转连接(skip connection).

具体网络结构如图 4 所示,包括 2 种类型的操作组合:$GT1$,$GT2$.网络各层结构的具体参数如表 1 所示,其中,列 2 和列 3 代表对应网络层的输入数据和输出数据尺寸.输入数据和输出数据尺寸参数格式为 $a\times(b\times c)$,a 为数据的通道数,b 为数据的高度,c 为数据的宽度.卷积核尺寸列为对应网络层中卷积核的参数.卷积核的具体参数格式为 $n\times(h\times w)\times s$,$n$ 为卷积核的个数,h 为卷积核的高度,w 为卷积核的宽度,s 为卷积操作的步长.Process 列表示对应层包含的操作,ADD(·)为按位相加操作,"·"是与当前层相加的网络层标号.

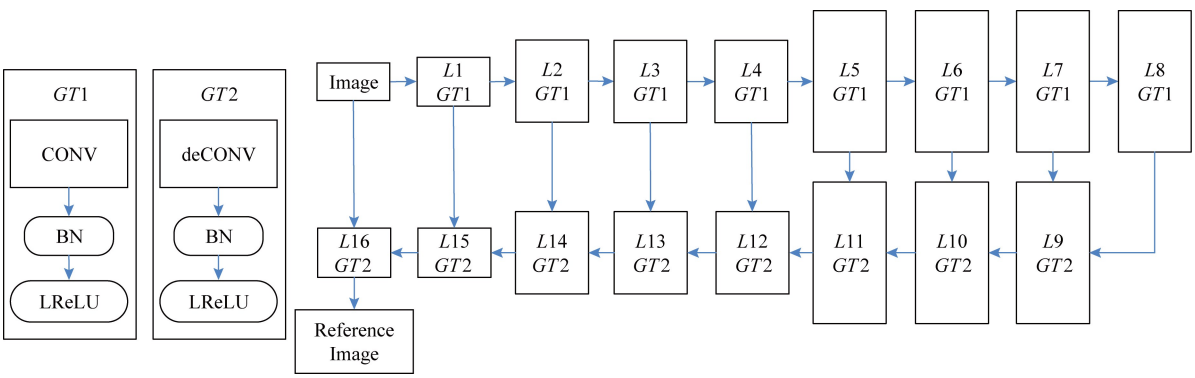


Fig. 4 Structure of reference image generation network for JPEG image deep learning steganalysis model

图 4 JPEG 图像隐写分析参照图像生成模型结构图

Table 1 Configuration Details of Reference Image Generation Network

表 1 参照图像生成模型的参数细节

Layer	Input Size $a\times(b\times c)$	Output Size $a\times(b\times c)$	Kernel Size $n\times(h\times w)\times s$	Process
$L1$	1×(256×256)	16×(256×256)	16×(3×3)×1	CONV-BN-LReLu
$L2$	16×(256×256)	32×(256×256)	32×(3×3)×1	CONV-BN-LReLu
$L3$	32×(256×256)	64×(128×128)	64×(3×3)×2	CONV-BN-LReLu
$L4$	64×(128×128)	64×(64×64)	64×(3×3)×2	CONV-BN-LReLu
$L5$	64×(64×64)	128×(32×32)	128×(3×3)×2	CONV-BN-LReLu
$L6$	128×(32×32)	128×(16×16)	128×(3×3)×2	CONV-BN-LReLu
$L7$	128×(16×16)	128×(8×8)	128×(3×3)×2	CONV-BN-LReLu
$L8$	128×(8×8)	128×(4×4)	128×(3×3)×2	CONV-BN-LReLu
$L9$	128×(4×4)	128×(8×8)	128×(5×5)×2	DECONV-BN-LReLu-ADD($L8$)
$L10$	128×(8×8)	128×(16×16)	128×(5×5)×2	DECONV-BN-LReLu-ADD($L7$)

Continued (Table 1)

Layer	Input Size $a\times(b\times c)$	Output Size $a\times(b\times c)$	Kernel Size $n\times(h\times w)\times s$	Process
$L11$	128×(16×16)	128×(32×32)	128×(5×5)×2	DECONV-BN-LReLu-ADD($L6$)
$L12$	128×(32×32)	64×(64×64)	64×(5×5)×2	DECONV-BN-LReLu-ADD($L5$)
$L13$	64×(64×64)	64×(128×128)	64×(5×5)×2	DECONV-BN-LReLu-ADD($L4$)
$L14$	64×(128×128)	32×(256×256)	32×(5×5)×2	DECONV-BN-LReLu-ADD($L3$)
$L15$	32×(256×256)	16×(256×256)	16×(5×5)×1	DECONV-BN-LReLu-ADD($L2$)
$L16$	16×(256×256)	1×(256×256)	1×(5×5)×1	DECONV-BN-LReLu-ADD($L1$)

Note: a is the channel of the data, b is the height of the data and c is the width of the data. Besides, n is the number of the convolutional kernel, h is the height of the convolutional kernel, w is width of the convolutional kernel and s is the stride for the convolution.

2.3 所提出参照图像生成模型训练方式

本文提出的隐写分析参照图像生成模型可采取2种训练策略:第1种训练策略为预训练策略,记为Pre-training;第2种训练策略为共同训练方式,记为Together.

第1种训练方式,即预训练方式.首先,对本文所提出参照图像生成模型进行预训练,预训练过程如图5所示.隐写分析参照图像生成模型预训练的损失函数为生成的参照图像与待检测图像对应的cover图像之间的相似程度,该损失函数的计算方法:

$$\varphi(\boldsymbol{I})=\boldsymbol{I}*\boldsymbol{K}, \tag{9}$$

$$\varphi(\boldsymbol{R})=\boldsymbol{R}*\boldsymbol{K}, \tag{10}$$

$$LOSS=\frac{1}{MN}\sum_{i,j}(\varphi(\boldsymbol{I})_{i,j}-\varphi(\boldsymbol{R})_{i,j})^2, \tag{11}$$

其中,$1\leqslant i\leqslant M$,$1\leqslant j\leqslant N$;M,N 为图像的宽度和长度;$\boldsymbol{I}$ 为待检测图像对应的原始图像;$\boldsymbol{R}$ 为本文提出的方法产生的隐写分析参照图像;$\boldsymbol{K}$ 是预训练过程中使用的滤波器;$\varphi(\boldsymbol{I})$和$\varphi(\boldsymbol{R})$分别是对 $\boldsymbol{I}$ 和 $\boldsymbol{R}$ 进行滤波的结果;$\varphi(\boldsymbol{I})_{i,j}$ 和 $\varphi(\boldsymbol{R})_{i,j}$ 分别为 $\varphi(\boldsymbol{I})$和$\varphi(\boldsymbol{R})$位于(i,j)位置的元素值;$LOSS$ 为预训练的损失函数.

首先采用 YeNet[24] 中所使用的 Rich Model[16] 中的30个高通滤波器 $\boldsymbol{K}$ 分别对生成图像 $\boldsymbol{R}$ 和输入图像所对应的cover图像 $\boldsymbol{I}$ 进行滤波操作;然后,计算以上2种图像的滤波结果 $\varphi(\boldsymbol{I})$和 $\varphi(\boldsymbol{R})$之间的均方误差,并将其作为参照图像 $\boldsymbol{R}$ 与待检测图像对应的cover图像 $\boldsymbol{I}$ 之间相似程度的度量和预训练的损失函数 $LOSS$.

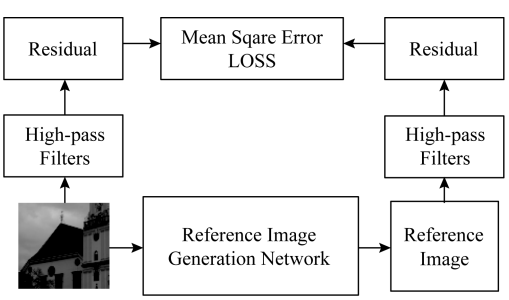


Fig. 5 Pre-training strategy

图5 Pre-training 训练策略

预训练完成之后,将参照图像生成模型与隐写分析模型相结合,以提升隐写分析任务分类准确率为目标,进行训练.通过对参照图像生成模型进行预训练,保证参照图像与待检测图像所对应的cover图像在残差噪声上尽可能接近.

第2种训练策略,属于1种共同训练方式,如图6所示,将参照图像生成模型与隐写分析模型作为1个整体,以提升隐写分析检测准确率为目标,进行共同训练.

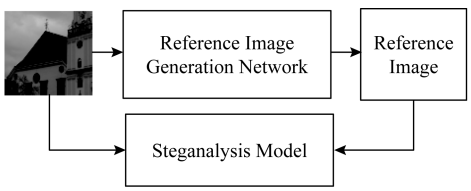


Fig. 6 Together training strategy

图6 Together 训练策略

2.4 参照图像与隐写分析模型的结合方式

JPEG图像隐写分析参照图像生成模型输出的参照图像将作为深度学习隐写分析模型的辅助信息,与待检测图像一同作为深度学习隐写分析模型的输入,以提升深度学习隐写分析模型的检测能力.

本文所提出方法生成的参照图像可以按照2种方式与隐写分析模型进行结合,如图7所示.

第1种结合方式,记为Combine方式,该种方式是将参照图像与对应的待检测图像沿着图像的 x 轴进行连接(concation)操作,合成结果为1个宽度为待检测待检测图像2倍的矩阵,并将该合成结果作为隐写分析模型的输入.

第2种合结合方式,记为Channel方式,将参照

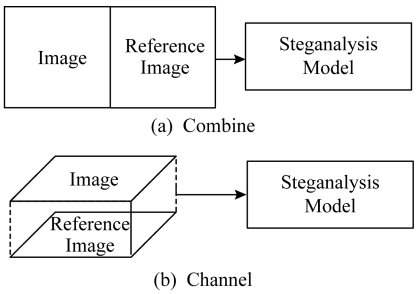


Fig. 7 Reference images input methods

图 7 参照图像输入方式

图像和待检测图像沿 z 轴进行合并.分别作为隐写分析模型的输入数据,在隐写分析模型对 2 种数据进行相应处理,并将处理的结果进行合并.

3 实验与分析

为验证本文方法有效性,基于 JPEG 图像隐写分析参照图像生成模型和 J-Xune[26],构建隐写分析模型,并与 J-Xunet 的检测效果进行对比.本文使用 Tensorflow 深度学习框架实现提出的模型以及对比模型.本论文的数值计算得到了武汉大学超级计算中心的计算支持和帮助.

3.1 数据集合

本文实验采用的数据集合来源于 BossBase v1.01[33]中的 10 000 图像.如图 8 所示,本文在实验中将 BossBase v1.01 中每张 512×512 大小的图像从 4 个对角以及图像的中间部分截取的 5 张大小为 256×256 的图像,从而产生 50 000 张空域图像.之后对 50 000 张空域图像分别以质量因子为 75 和 95,进行 JPEG 压缩,得到对应不同质量因子的 JPEG 图像,作为本文实验中涉及的 cover 图像.

Fig. 8 Generation of image samples in the experiments

图 8 实验数据产生方式

本文所涉及的对比实验的训练、验证、测试数据的 stego 图像由 JC-UED[9]和 J-UNIWARD[10]和隐写算法分别以相对嵌入率 0.1bpnzAC(bits per none zero AC), 0.2bpnzAC, 0.3bpnzAC, 0.4bpnzAC, 0.5bpnzAC 对 50 000 张 256×256 大小的 cover 图像进行嵌入得到.

对比实验的数据构成如图 9 所示,将 BossBase v1.01 中 8 000 张原始图像对应的 40 000 张 256×256 大小的 cover 图像以及对应的 40 000 张 stego 图像,共 80 000 张图像作为模型的训练数据集合;将 BossBase v1.01 中与训练测试集合不同的 1 000 张图像所对应的 5 000 张 256×256 大小 cover 图像以及对应的 stego 图像,共 10 000 张图像作为验证数据集合;将 BossBase v1.01 中其余 1000 张图像对应的 5 000 张 256×256 大小图像及其对应的 stego 图像,共 10 000 张图像作为实验的测试数据集合.

所构建的数据集合保证了本文提出的模型以及对比方法采用相同的训练、验证、测试数据集合,并且训练、验证、测试数据集合的原图像互不重复.

3.2 模型参数设置

本文所提出模型在训练过程中都采用 Adam[34]优化算法进行目标函数的优化.其中,学习率初始值为 0.001,学习率每 5 000 次训练衰减为原来的 0.9.训练过程中,每个 mini-batch 包括 16 对 cover 和 stego 图像,即 32 张图像;每种模型训练的最大训练迭代次数为 20 万,即 80 个 epoch.

此外,本文实验所涉及的对比算法采用相同参数设置.

3.3 模型训练策略与结合方式的选择

为了确定 JPEG 图像隐写分析参照图像生成模型的训练策略,以及与深度学习隐写分析模型的结合方式,分别采用 Together, Pre-training 训练方式以及 Combine 和 Channel 结合方式进行包含 1 种情况的对比实验.

该对比实验中,将本文所提出的隐写分析参照凸显生成模型与 J-Xunet[26]隐写分析模型相结合.各种情况下模型的最大训练迭代次数为 12.5 万次,并选取最后 5 个 epoch 的验证结果的平均值作为最终的检测准确率.检测算法为 J-UNIWARD[10],相对嵌入率为 0.4bpnzAC,JPEG 图像载体为质量因子为 75.分别采用 2 种模型结合方式,以及 2 种不同的训练策略,得到的 4 种模型在验证数据集合中的检测准确率如表 2 所示.其中,加粗数据为最高的准确率.模型在验证集合上的准确率在训练过程中的变化情况如图 10 所示.其中,横坐标为训练的 epoch,纵坐标为各个 epoch 训练完成之后的测试准确率.

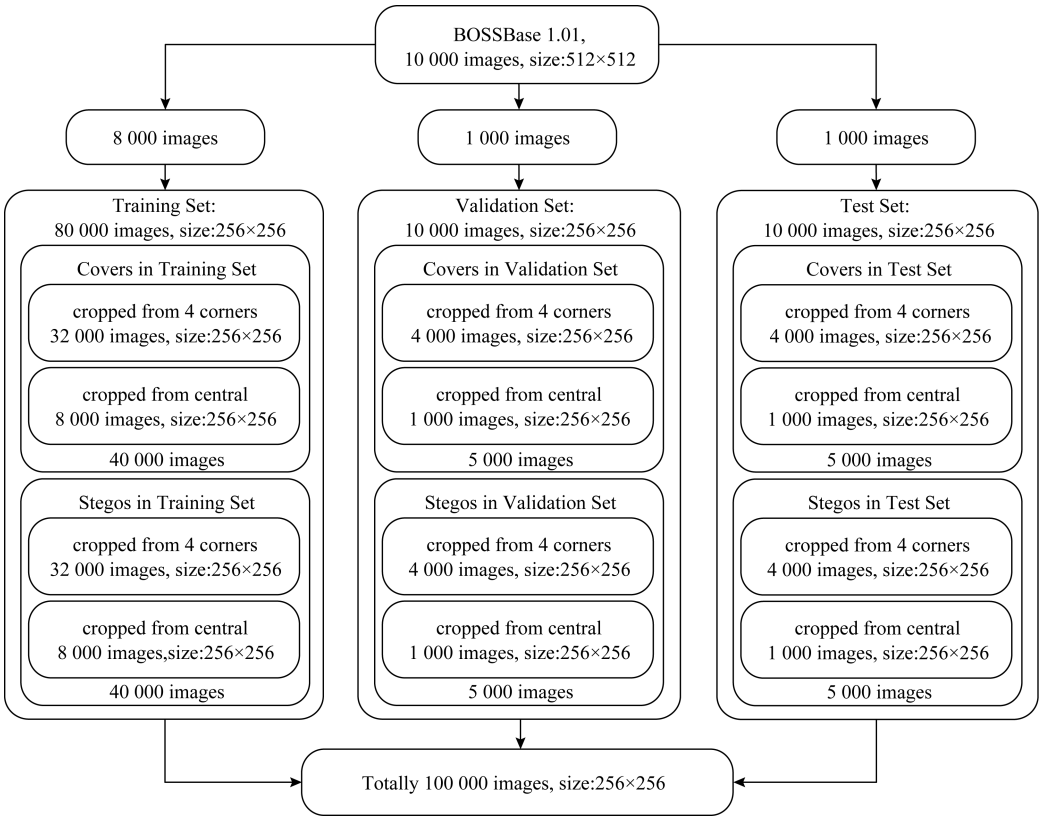


Fig. 9 Data sets for the experiments

图 9 对比实验数据集合构成

Table 2 Detection Accuracy when Different Training Strategies and Combination Ways are Applied

表 2 不同训练策略、结合方式在验证数据集的检测效果

%

Training Strategy	Combine	Channel
Together	**83.45**	49.30
Pre-training	78.75	80.16

Note: The bold numbers are the best performance in experiments.

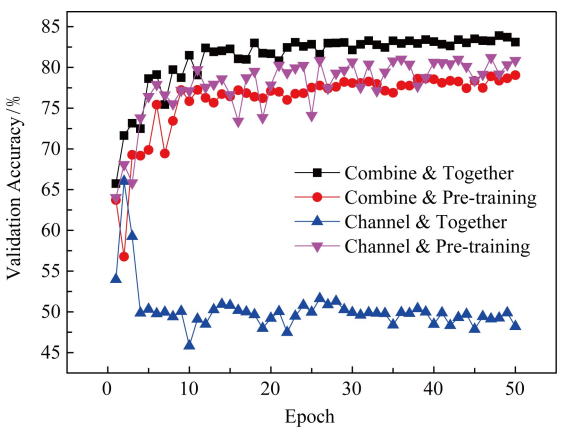


Fig. 10 Effect of training strategies and combination ways on verification results

图 10 训练策略、结合方式对验证结果的影响

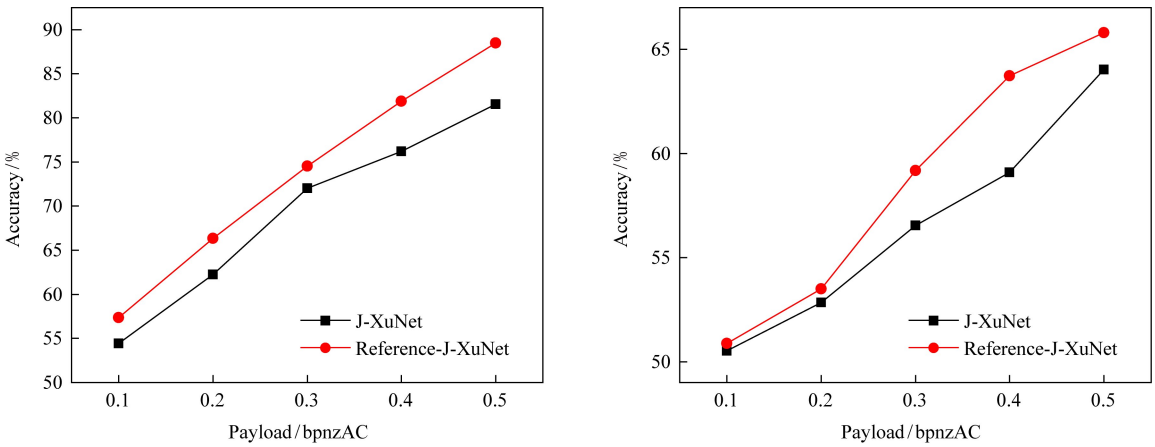
实验结果表明:采用 Combine 方式将参照图像生成模型与隐写分析模型结合,能够获得具有更高检测准确率的模型.当本文方法与 J-Xunet 隐写分析模型采用 Combine 方式进行结合,并采用 Together 训练方式得到的隐写分析模型具有最高的检测准确率.因此,本文对比实验采用 Combine 方式将参照图像生成模型产生的参照图像与待检测图像结合,并采用 Together 训练方式进行训练.

3.4 对比实验与结果分析

本文提出的 JPEG 隐写分析参照图像生成模型与 J-Xunet[26] 深度学习隐写分析模型结合,从而为针对 JPEG 图像的深度学习隐写分析模型提供辅助信息,所构成的隐写分析模型记为 Reference-J-Xunet.为验证本文提出方法的有效性,利用第 3.1 节中所构建的隐写分析数据集合,进行 Reference-J-Xunet 与 J-XuNet[26] 之间的隐写分析检测能力对比实验.

每种深度学习隐写分析模型采用与 3.2 节中相同的参数设置.每种隐写分析模型训练完成之后,取训练过程中最后 5 个 epoch 所保存模型的测试结果的平均值作为最终检测准确率.除此之外,每种隐写

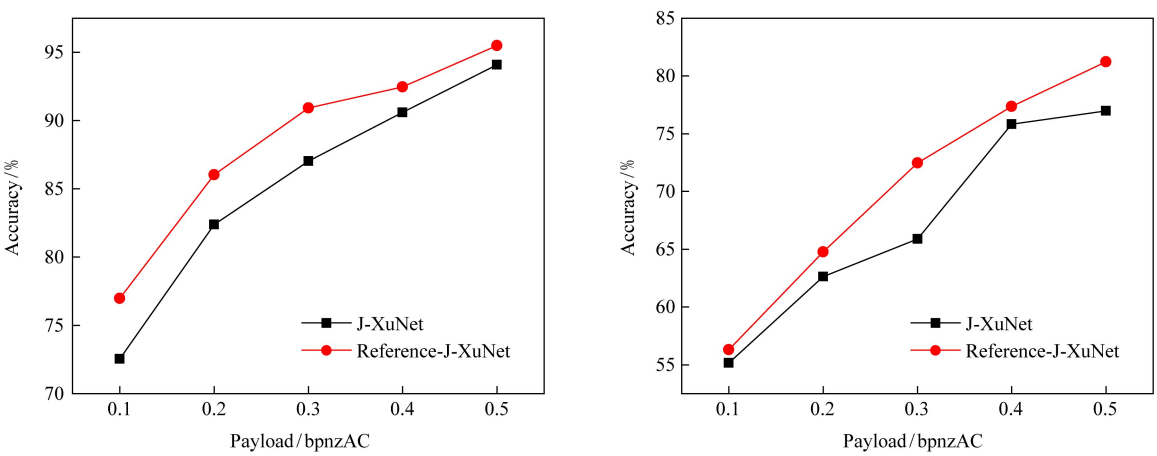
分析模型各进行以上 3 次测试,并取 3 次测试准确率的平均结果作为最终的测试结果.具体测试结果如图 11、图 12 以及表 3 所示.其中,图 11 和图 12 中的圆点表示本文提出方法针对不同载体质量因子、不同嵌入率,以及不同隐写算法嵌入的样本的隐写分析准确率,而正方形表示 J-XuNet 相应的隐写分析准确率.表 3 中,行 2 数据为待检测样本的相对嵌入率,单位为 bpnzAC(bits per non-zero AC DCT coefficient).此外,表 3 中的加粗数据为针对相同算法相同嵌入率情况下最高检测准确率.



(a) J-UNIWARD QF75 (b) J-UNIWARD QF95

Fig. 11 Detection accuracy comparison between J-XuNet and Reference-J-XuNet for J-UNIWARD QF75 and QF95

图 11 J-XuNet 与 Reference-J-XuNet 对于 J-UNIWARD 隐写算法的检测准确率对比



(a) J-UNIWARD QF75 (b) J-UNIWARD QF95

Fig. 12 Detection accuracy comparison between J-XuNet and Reference-J-XuNet for JC-UED

图 12 J-XuNet 与 Reference-J-XuNet 对于 JC-UED 隐写算法的检测准确率对比

Table 3 Detection Results of Proposed Method and Prior Art for J-UNIWARD and UED-JC

表 3 本文方法与对比算法对于不同隐写算法的检测结果 %

Embedding Method	Detecor	QF75					QF95				
		0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
J-UNIWARD	Reference-J-Xunet	**57.352**	**66.357**	**74.512**	**81.867**	**88.508**	**50.88**	**53.49**	**59.176**	**63.718**	**65.802**
	J-Xunet	54.438	62.246	72.026	76.205	81.572	50.52	52.838	56.544	59.089	64.024
JC-UED	Reference-J-Xunet	**76.966**	**86.024**	**90.916**	**92.472**	**95.486**	**56.296**	**64.762**	**72.478**	**77.378**	**81.236**
	J-Xunet	72.544	82.388	87.044	90.604	94.096	55.146	62.644	65.904	75.838	76.984

Note: The bold numbers are the best performance in experiments.

在载体质量因子为 95 时,相对嵌入率为 0.1bpnzAC 和 0.2bpnzAC,隐写算法为 J-UNIWARD 的情况下,以及载体质量因子为 95 时,相对嵌入率为 0.1bpnzAC,隐写算法为 JC-UED 的情况下,检测难度较大.为了保证模型在以上情况下的检测效果,本文对比实验采用调优策略进行以上情况下模型的训练.

具体训练过程为:

1) 采用载体质量因子为 95,相对嵌入率为 0.5bpnzAC的样本,对模型进行训练,共训练 80 个 epoch.

2) 利用当前情况下的样本,继续训练上一步训练过程中得到的模型,即对模型进行调优.调优过程的最大 epoch 数为 80.

在隐写样本的嵌入方法为 J-UNIWARD[10] 算法、载体图像质量因子为 75、隐写样本的相对嵌入率分别为 0.1bpnzAC,0.2bpnzAC,0.3bpnzAC,0.4bpnzAC,0.5bpnzAC 的情况下,本文提出方法相对于对比方法的隐写分析准确率分别提升了2.914,4.111,2.486,5.662,6.936 个百分点.而当载体图像具有较高质量因子,如质量因子为 95 时,隐写分析难度相对于低质量因子图像更大.因此,实验中载体质量因子为 95 时,隐写分析准确率相对于载体质量因子为 75 有所下降.但本文提出方法依然能够提升隐写分析模型的检测能力.在相对嵌入率为 0.2bpnzAC,0.3bpnzAC,0.4bpnzAC,0.5bpnzAC 的情况下,本文方法的隐写分析准确率相对于对比方法分别提升 0.652,2.632,4.629,1.778 个百分点.

当隐写样本采用 JC-UED[9] 算法进行嵌入的情况下,本文提出方法同样能够提升深度学习隐写分析模型的检测能力.当载体图像的质量因子为 75 且相对嵌入率分别为 0.1bpnzAC,0.2bpnzAC,0.3bpnzAC,0.4bpnzAC,0.5bpnzAC 的情况下,本文方法隐写分析准确率相对于对比方法分别提升 4.422,3.636,3.872,1.868,1.39 个百分点.在载体质量因子为 95 且相对嵌入率分别为 0.1bpnzAC,0.2bpnzAC,0.3bpnzAC,0.4bpnzAC,0.5bpnzAC 的情况下,本文方法的隐写分析准确率相对于对比方法分别提升 1.15,2.118,6.574,1.54,4.252 个百分点.

当载体图像质量因子为 95、隐写样本相对嵌入率为 0.1bpnzAC,隐写算法为 J-UNIWARD 时,对比试验在模型训练过程中采用了调优策略.虽然本文算法相对于对比方法略有提升,但是本文方法和对比方法的隐写检测准确率均接近 50%.

实验结果表明:本文方法能够为基于深度学习的 JPEG 图像隐写分析模型提供有利于提升检测能力的辅助信息,提升深度学习隐写分析模型的检测能力.但在载体图像具有较高质量因子、较低嵌入率情况下提升效果有待加强.

4 总 结

本文从隐写分析模型辅助信息的角度,探索进一步提升 JPEG 图像隐写分析模型检测能力的途径.构建了基于卷积神经网络的隐写分析参照图像生成模型为,以更加充分地挖掘待检测图像中与隐写分析有关的信息.本文提出方法对待检测图像进行变换,并与隐写分析模型一同参与训练,保证参照图像生成模型以提升隐写分析检测能力为目标,为隐写分析模型提供辅助信息.对比实验结果表明:本文提出模型,能够为针对 JPEG 图像的深度学习隐写分析模型提供有助于提升检测能力的辅助信息.但在载体为高质量因子 JPEG 图像,相对嵌入率较低的情况下,提升效果有待加强.

在未来,将针对隐写分析检测对象的载体属性与深度学习隐写分析模型的特性,对 JPEG 图像隐写分析参照图像生成模型进行改进与优化,改善本文构建的模型在载体具有高质量因子,相对嵌入率较低情况下的检测能力,探索进一步提升针对 JPEG 图像的深度学习隐写分析模型检测能力的途径.

参 考 文 献

[1] Wang Lina, Zhang Huanguo, Ye Dengpan. Information Hiding Technology and Application [M]. Wuhan: Wuhan University Press, 2003 (in Chinese)
(王丽娜, 张焕国, 叶登攀. 信息隐藏技术与应用[M]. 武汉: 武汉大学出版社, 2003)

[2] Zhang Zhan, Liu Guangjie, Dai Yuewei, et al. A self-adaptive image steganography algorithm based on cover-coding and Markov model [J]. Journal of Computer Research and Development, 2012, 49(8): 1668-1675 (in Chinese)
(张湛, 刘光杰, 戴跃伟, 等. 基于隐写编码和 Markov 模型的自适应图像隐写算法[J].计算机研究与发展, 2012, 49(8): 1668-1675)

[3] Han Tao, Zhu Yuefei, Lin Sisi, et al. Modified matrix encoding based on the spatial distortion model and its improvement [J]. Journal of Computer Research and Development, 2014, 51(7): 1467-1475 (in Chinese)

(韩涛, 祝跃飞, 林斯思, 等. 基于空域失真模型的修正矩阵编码及其改进[J].计算机研究与发展, 2014, 51(7): 1467-1475)

[4] Bao Zhenkun, Zhang Weiming, Cheng Sen, et al. ±1 Steganographic codes by applying syndrome-trellis codes to dynamic distortion model in pixel chain [J]. Journal of Computer Research and Development, 2014, 51(8): 1739-1747 (in Chinese)
(包震坤, 张卫明, 程森, 等. 基于像素链动态失真和校验格码的±1 隐写编码[J]. 计算机研究与发展, 2014, 51(8): 1739-1747)

[5] Zhao Xianfeng, Zhang Hong. Principles and Techniques of Seganography [M]. Beijing: Science Press, 2018 (in Chinese)
(赵险峰, 张弘. 隐写学原理与技术[M]. 北京: 科学出版社, 2018)

[6] Filler T, Judas J, Fridrich J, et al. Minimizing additive distortion in steganography using syndrome-trellis codes [J]. IEEE Transactions on Information Forensics and Security, 2011, 6(3): 920-935

[7] Pevný T, Filler T, Bas P. Using high-dimensional image models to perform highly undetectable steganography [G] // LNCS 6387: Proc of the 12th Int Conf on Information Hiding. Berlin: Springer, 2010: 161-171

[8] Holub V, Fridrich J. Designing steganographic distortion using directional filters [C] //Proc of the 2012 IEEE Int Workshop on Information Forensics and Security. Piscataway, NJ: IEEE, 2012: 234-239

[9] Guo Linjie, Ni Jiangqun, Shi Yunqing. Uniform embedding for efficient JPEG steganography [J]. IEEE Transactions on Information Forensics and Security, 2014, 9(5): 814-825

[10] Holub V, Fridrich J, Denemark T. Universal distortion function for steganography in an arbitrary domain [OL]. [2019-06-01]. https://doi.org/10.1186/1687-417X-2014-1

[11] Fridrich J, Kodovsky J. Multivariate Gaussian model for designing additive distortion for steganography [C] //Proc of the 2013 IEEE ICASSP. Piscataway, NJ: IEEE, 2016: 26-31

[12] Li Bin, Wang Ming, Huang Jiwu, et al. A new cost function for spatial image steganography [C] //Proc of the 2014 IEEE Int Conf on Image Processing (ICIP). Piscataway, NJ: IEEE, 2014: 4206-4210

[13] Sedighi V, Cogranne R, Fridrich J. Content-adaptive steganography by minimizing statistical detectability [J]. IEEE Transactions on Information Forensics and Security, 2016, 11(2): 221-234

[14] Wang Lina, Wang Kaige, Xu Yibo, et al. An evaluation of carrier security for image steganography based on residual co-occurrence probability [J]. Journal of Computer Research and Development, 2018, 55(12): 2664-2673 (in Chinese)
(王丽娜, 王凯歌, 徐一波, 等. 基于残差共生概率的隐写图像载体安全性评价[J].计算机研究与发展, 2018, 55(12): 2664-2673)

[15] Pevny T, Bas P, Fridrich J. Steganalysis by subtractive pixel adjacency matrix [J]. IEEE Transactions on Information Forensics and Security, 2010, 5(2): 215-224

[16] Fridrich J, Kodovsky J. Rich models for steganalysis of digital images [J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882

[17] Holub V, Fridrich J. Random projections of residuals for digital image steganalysis [J]. IEEE Transactions on Information Forensics and Security, 2013, 8(12): 1996-2006

[18] Denemark T, Sedighi V, Holub V, et al. Selection-channel-aware rich model for steganalysis of digital images [C] //Proc of the 2014 IEEE Int Workshop on Information Forensics and Security (WIFS). Piscataway, NJ: IEEE, 2014: 48-53

[19] Tang Weixuan, Li Haodong, Luo Weiqi, et al. Adaptive steganalysis based on embedding probabilities of pixels [J]. IEEE Transactions on Information Forensics and Security, 2016, 11(4): 734-745

[20] Tan Shunquan, Li Bin. Stacked convolutional auto-encoders for steganalysis of digital images [C/L] //Proc of the 2014 Signal and Information Processing Association Annual Summit and Conf. Piscataway, NJ: IEEE, 2014. [2019-06-01]. https://ieeexplore.ieee.org/document/7041565

[21] Qian Yinlong, Dong Jing, Wang Wei, et al. Deep learning for steganalysis via convolutional neural networks [G] // LNCS 9409: Proc of the 2015 SPIE Electronic Imaging. Berlin: Springer, 2015

[22] Xu Guanshuo, Wu Hanzhou, Shi Yunqing. Structural design of convolutional neural networks for steganalysis [J]. IEEE Signal Processing Letters, 2016, 23(5): 708-712

[23] Yang Jianhua, Liu Kai, Kang Xiangui, et al. Steganalysis based on awareness of selection-channel and deep learning [C] // Proc of the 16th Int Workshop on Digital Forensics and Watermarking. Berlin: Springer, 2017: 263-272

[24] Ye Jian, Ni Jiangqun, Yi Yang. Deep learning hierarchical representations for image steganalysis [J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2545-2557

[25] Boroumand M, Chen M, Fridrich J, et al. Deep residual network for steganalysis of digital images [J]. IEEE Transactions on Information Forensics and Security, 2019, 14(5): 1181-1193

[26] Xu Guanshuo. Deep convolutional neural network to detect J-UNIWARD [C] //Proc of the 5th ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2017: 67-73

[27] Chen Mo, Sedighi V, Boroumand M, et al. JPEG-phaseaware convolutional neural network for steganalysis of JPEG images [C] // Proc of the 5th ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2017, 75-84

[28] Zeng Jishen, Tan Shunquan, Li Bin, et al. Large-scale JPEG image steganalysis using hybrid deep-learning framework [J]. IEEE Transactions on Information Forensics and Security, 2017, 13(5): 1200-1214

[29] Yang Jianhua, Shi Yunqing, Wong E K, et al. JPEG steganalysis based on DenseNet [J]. arXiv preprint arXiv: 1711.09335, 2017

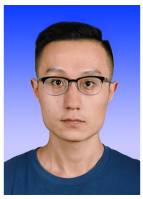
[30] Huang Gao, Liu Zhuang, Maaten L V D, et al. Densely connected convolutional networks [C] //Proc of the 2017 IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2017: 2261-2269

[31] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [C] // Proc of the 32nd Int Conf on Machine Leaning-Volume 37. New York: ACM, 2015: 448-456

[32] Ronneberger O, Fischer P, Brox T, et al. U-Net: Convolutional networks for biomedical image segmentation [G] //LNCS 9351: Proc of MICCAI 2015. Berlin: Springer, 2015: 234-241

[33] Bas P, Filler T, Pevný T. "Break our steganographic system": The ins and outs of organizing BOSS [G] //LNCS 9351: Proc of MICCAI 2015. Berlin: Springer, 2015: 234-241

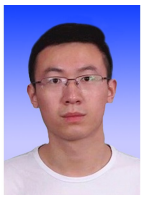
[34] Kingma D P, Ba J. Adam: A method for stochastic optimization [J]. arXiv preprint arXiv: 1412.6980, 2014



Ren Weixiang, born in 1987. PhD candidate. His main research interests include information hiding, machine learning and deep learning.

Zhai Liming, born in 1988. PhD candidate. His main research interests include steganography and steganalysis. (limingzhai@whu.edu.cn)

Wang Lina, born in 1964. PhD, professor. Member of CCF. Her main research interests include system security and steganalysis.



Jia Ju, born in 1990. PhD candidate. His main research interests include steganography and steganalysis.