文章编号:1674-2974(2022)04-0035-12 DOI:10.16339/j.cnki.hdxbzkb.2022280

HCGAN:一种基于GAN的高容量信息隐藏算法

张克君[1,2],李旭[1],于新颖[2†],冯丽雯[1],秦昊聪[1],张健毅[1]
(1. 北京电子科技学院 网络空间安全系,北京市 100071;
2. 北京邮电大学 网络空间安全学院,北京市 100876)

摘 要:针对现有信息隐藏算法存在隐写容量低、信息提取困难以及安全性差等问题,本文提出了一种基于生成对抗网络的高容量信息隐藏算法(High Capacity Information Hiding Algorithm Based on GAN,HCGAN).在秘密信息嵌入方面,使用基于Im-Residual结构的编码器将秘密信息嵌入载体图像中,避免了秘密信息嵌入时由卷积层提取特征导致的信息损失.在秘密信息提取方面,使用基于稠密结构的解码器从含秘图像中提取出秘密信息,利用特征复用来增加秘密信息的提取率.在抗隐写分析方面,利用基于隐写分析的鉴别器与基于Im-Residual结构的编码器进行对抗训练,以提高含秘图像的抗隐写分析能力.实验表明,经过对抗训练后,HCGAN在2 bpp嵌入率下比WOW和S-UNIWARD在0.4 bpp嵌入率下具有更低的隐写分析检测率.

关键词:信息隐藏;深度学习;生成对抗网络;自编码器;卷积神经网络
中图分类号:TN915.08 **文献标志码**:A

HCGAN: A High Capacity Information Hiding Algorithm Based on GAN

ZHANG Kejun[1,2],LI Xu[1],YU Xinying[2†],FENG Liwen[1],QIN Haocong[1],ZHANG Jianyi[1]
(1. Department of Cyberspace Security,Beijing Electronic Science and Technology Institute,Beijing 100071,China;
2. School of Cyberspace Security,Beijing University of Posts and Telecommunications,Beijing 100876,China)

Abstract:Aiming at the problems of low steganographic capacity, difficult information extraction, and poor security in existing information hiding algorithms, this paper proposes a high capacity information hiding algorithm based on GAN(HCGAN). For secret information embedding, an Im-Residual structure-based encoder is applied to embed the secret information into the carrier image, avoiding the information loss caused by the feature extraction of the convolution layer. For secret information extraction, a dense structure-based decoder is utilized to extract secret information from the secret image, and feature reuse is used to increase the extraction rate of secret information. In terms of anti-steganalysis, the discriminator based on steganalysis and the encoder based on Im-Residual structure are used for adversarial training to improve the anti-steganalysis ability of the secret image. Experiments show that after adversarial training, HCGAN has a lower steganalysis detection rate at an embedding rate of 2bpp than the WOW and S-

* **收稿日期**:2021-11-29
基金项目:北京高校高精尖学科建设项目(20210086Z0401), Advanced Discipline Construction Project of Beijing Universities (20210086Z0401); 国家重点研发计划网络空间安全重大专项课题资助(2018YFB0803601), National Key Research and Development Program on Cyberspace Security (2018YFB0803601)
作者简介:张克君(1972—),男,山东临沂人,北京电子科技学院教授,博士
† 通信联系人,E-mail: Xinying_334@bupt.edu.cn

UNIWARD algorithms at an embedding rate of 0.4bpp.

Key words: information hiding; deep learning; generative adversarial networks; autoencoder; convolutional neural network

随着网络信息技术的快速发展和移动终端设备的不断普及,数字多媒体成为互联网通信的主要载体.数字多媒体的广泛应用使得网络中充斥着大量对国家、企业或个人而言十分敏感的信息.为保证这些信息能被安全地传递,发送方通常需要对信息进行加密,同时加密后的密文只能被持有特定密钥的接收者解密,但密文的伪随机性会暴露秘密信息的存在,从而引来攻击者的怀疑和攻击.信息隐藏技术通过将敏感的秘密信息伪装成普通信息来隐藏秘密信息,为上述问题提供了解决方案,对信息安全有着重要意义[1].

在信息隐藏过程中,发送方首先将秘密信息嵌入载体中形成含秘载体,随后把含秘载体上传到公共信道[2].特定的接收方从公共信道中下载含秘载体,使用密钥或提取算法从中提取出秘密信息.在秘密信息的传递过程中,除了通信双方之外,其他人虽然可以下载含秘载体但无法获取其中的秘密信息.信息隐藏技术的这种隐蔽性能够有效阻止攻击者对秘密信息的感知和破坏,进而保证了秘密信息在传递过程中的安全[3].

信息隐藏技术和加密技术的结合既防止了秘密信息的存在被知晓,又保护了秘密信息的内容,但仍不能保证秘密信息的绝对安全.发送方在嵌入秘密信息时会不可避免地对载体造成干扰,隐写分析就可以对这些干扰进行分析并判断出载体是否含有秘密信息.信息隐藏和隐写分析互为对手,在相互对抗中不断发展进步.为了对抗隐写分析,信息隐藏已经从早期的最低有效位(Least Significant Bit,LSB)替换算法[4]发展到现在根据载体内容来嵌入秘密信息的自适应信息隐藏算法.在不断进步的信息隐藏技术的推动下,隐写分析也很快实现了从专用到通用,从低维特征到高维特征的跨越.

不断进步的隐写分析算法虽然促进了信息隐藏算法的发展,却对信息隐藏技术的安全性提出了挑战.近年来,深度学习技术在运算速度更快的硬件支持下迅速发展,不断涌现出各种网络模型.深度置信网络[5]及反向传播算法的提出更是推动了深度学习技术在计算机视觉、自然语言处理和自动驾驶等领域的发展.深度学习强大的学习能力和特征表达能力,使得隐写分析算法的设计不再需要丰富的专业知识.在基于深度学习的隐写分析算法被提出之后,隐写分析对含秘载体的检测能力越来越强,这迫使信息隐藏领域的研究者开始研究更加安全的信息隐藏算法.

在对信息隐藏技术的探索中,Goodfellow 等人于2014年提出的生成对抗网络[6](Generative Adversarial Networks,GAN)受到了信息隐藏技术研究者的关注.GAN包含两个神经网络,其中一个网络能够生成符合真实样本分布的生成样本,另一个网络用来区分真实样本和生成样本,这两个网络互相对抗并交替训练.GAN这种对抗博弈的方法和信息隐藏技术与隐写分析算法互相对抗并不断发展的情况十分相似,因此信息隐藏技术的研究者期望将GAN用于信息隐藏中以得到更强的安全性和更高的隐写容量.迄今为止,基于GAN网络的信息隐藏算法并没有完全实现上述目标,它们在隐写容量、安全性和秘密信息提取准确率等方面还存在不足.因此,研究基于GAN的信息隐藏技术并使其在隐写容量和安全性等方面获得提升有着重要的实际意义.

本文提出了一种新的基于GAN的高容量信息隐藏算法(HCGAN).该算法能够在高容量的情况下保持较高的秘密信息提取率,并且有较强的抗隐写分析能力.HCGAN算法由三部分组成:①基于Im-Residual结构的编码器,发送方能够通过编码器将秘密信息嵌入载体图像中形成隐秘图像;②基于稠密结构的解码器,接收方可以通过解码器将秘密从隐秘图像中提取出来;③基于隐写分析的鉴别器,用来进行对抗性训练以提高隐秘图像的抗隐写分析能力.本文的主要贡献如下:

1)提出了基于Im-Redisual结构的编码器,其中Im-Residual结构是基于Residual结构进行优化所形成的,Im-Residual结构的编码器采用多次嵌入的思想能够降低秘密信息因特征提取而产生的信息损失,从而提高秘密信息提取率.

2)设计了基于稠密结构的解码器,利用Dense结构特征复用的能力提取出更多有用特征,以进一步提高解码器的秘密信息提取率.

3)提出了一种基于隐写分析的鉴别器,通过GAN的对抗训练思想进一步提升HCGAN信息隐藏算法的抗隐写分析能力.

1 相关工作

1.1 经典的信息隐藏算法

根据所使用的方法不同,信息隐藏技术可以分为经典的信息隐藏算法和基于深度学习的信息隐藏算法.

最低有效位替换算法作为早期的信息隐藏算法,只对像素中最不影响图像视觉效果的最低有效位进行修改[4]. LSB替换算法能保证载体的视觉质量,但在嵌入过程中没有考虑载体的统计特征.因此,隐写分析算法可以轻易地根据直方图异常等特征识别出通过LSB替换算法获得的含秘载体.像素值差分[7](Pixel Value Differencing,PVD)使用两个连续像素之间的差来表示秘密信息,比LSB替换算法更安全,但是也存在统计特征的异常.此后,研究者们设计信息隐藏算法时就注意保持某些统计特征不变. LSB匹配算法对LSB替换技术进行改进,如果秘密比特和载体图像的LSB不匹配,则在相应的像素值上随机加减1[8]. 随机加减1的方法可以保持载体图像的统计直方图不变,但最低位和次低位的统计特征异常仍会使LSB匹配算法不够安全[8].除了可以在空间域嵌入秘密信息以外,也可以在各种变换域嵌入秘密信息.离散余弦变化(Discrete Cosine Transform,DCT)就是一种变换域.与傅立叶变换类似,DCT可以将图像从空间域转换到频率域.JPEG格式的图像就使用了DCT,其对每个颜色成分使用余弦变换来将大小为8×8的连续像素块转换为64个余弦系数. JSteg[9]、OutGuess[10]等方法都是在DCT的基础上嵌入秘密信息的信息隐藏算法.

信息隐藏算法能够根据载体内容自适应地嵌入秘密信息来提高算法的安全性.2010年,Pevný等人[11]提出的HUGO就是一种自适应的信息隐藏算法.HUGO首先通过嵌入信息对载体图像的影响来定义失真函数,随后使用加权范数函数将像素空间压缩为特征空间,并在失真函数最小的位置嵌入秘密信息[11]. 2012年,Holub等人[12]提出的WOW(Wavelet Obtained Weights)算法可以根据图像内容在纹理复杂的区域嵌入更多的信息.2014年Holub等人[13]提出的S-UNIWARD算法是一种与嵌入域无关的通用失真函数.虽然HUGO、WOW和S-UNIWARD的失真函数各不相同,但三个算法的最终目的都是通过最小化失真函数来自适应地嵌入秘密信息.此后提出的经典的信息隐藏算法或是改进失真函数[14],或是改进隐写编码[15].这些改进虽然提高了信息隐藏的安全性,但对信息隐藏算法的隐写容量提高非常有限.

1.2 基于深度学习的信息隐藏算法

基于深度学习的信息隐藏算法被提出后,经典信息隐藏算法的安全就面临着严重的威胁.将深度学习应用于信息隐藏领域以对抗基于深度学习的隐写分析算法成为信息隐藏领域中新的研究方向.特别是生成对抗网络,其优秀的图像生成能力和博弈对抗的思想引起了信息隐藏领域研究者们的关注.

2017年,Hayes等人[16]提出将GAN与信息隐藏相结合的Ste-GAN-ography算法.该算法使用三个普通的神经网络分别作为嵌入秘密信息的生成器、提取秘密信息的提取器和进行隐写分析的鉴别器,训练过程和一般的生成对抗网络相同. 2017年,Tang等人[17]提出了ASDL-GAN算法.该算法首先学习一个概率映射矩阵来为秘密信息寻找合适的嵌入位置,然后使用传统的STC方法嵌入秘密信息.Yang等人[18]在ASDL-GAN的基础上进行改进,使用Tanh-simulator激活函数、基于U-NET[19]的生成器和SCA[20]判别器来减少训练时间并增强抗隐写分析能力.Liu等人[21]于2018年提出了一种直接使用ACGAN生成器进行无载体信息隐藏的Stego-ACGAN算法.该算法通过建立图像类别与文本之间的映射字典,将秘密信息表示为图像类别信息,然后将图像类别信息输入生成器生成含秘图像,从而实现无载体的信息隐藏[21-22]. 2019年,Zhang等人[23]提出一种基于约束采样的含密载体生成方法.该方法首先训练一个生成器,然后通过数字化卡登格子对生成器进行约束采样,以获得满足条件的含密图像.

自2017年提出在载体图像中隐藏秘密图像的应用场景后,图像信息隐藏这一方向迅速成为信息隐藏领域的研究热点[24]. 2017年,Baluja等人[24]提出使用神经网络在图像中寻找合适的位置并嵌入秘密图像信息.该方法对编码过程进行训练,将整个秘密图像嵌入载体图像中并使秘密图像信息分散到图像的每个比特中. 2018年,Zhang等人[25]提出将封面图像分为Y、U、V三个通道,并通过编码器将灰度化的

秘密图像隐藏到封面图像的Y通道中.2020年,Duan等人[26]提出了一种结合图像椭圆曲线密码学和深度学习的大容量信息隐藏算法.该算法利用离散余弦变换对秘密图像进行变换,再用椭圆曲线密码术对变换后的图像进行加密,最后使用SegNet[27]网络来隐藏和提取秘密信息.

图像信息隐藏算法带来的高隐写容量是在损失部分秘密图像数据的前提下达成的.这些算法之所以能隐藏秘密图像信息,是因为图像能够在损失部分数据后不影响信息的呈现.如果要隐藏较多的文本或其他类型的秘密信息,使用现有的图像信息隐藏算法会导致较低的秘密信息提取准确率和较弱的抗隐写分析能力.

本文对图像信息隐藏算法进行研究后发现引起秘密信息损失的主要原因是卷积层只能提取有限特征.因此,本文设计了基于Im-Redisual结构的编码器,通过减少编码过程中的信息损失来提高秘密信息提取率.本文利用基于稠密结构的解码器和基于隐写分析的鉴别器分别来提高秘密信息提取率和抗隐写分析能力.

2 基于GAN的高容量信息隐藏算法HCGAN

为解决信息隐藏算法存在的隐写容量低和抗隐写能力弱的问题,本文提出了一种基于GAN的高容量信息隐藏算法(HCGAN).如图1所示,HCGAN算法主要由三部分组成:基于Im-Residual结构的编码器、基于稠密结构的解码器和基于隐写分析的鉴别器.基于Im-Residual结构的编码器将输入的秘密信息和载体进行编码以得到含秘载体.发送方得到含秘载体后将其上传到公共信道中.接收方从公共信道中获得含秘载体后,使用基于稠密结构的解码器从含秘载体中解码出秘密信息.基于隐写分析的鉴别器的任务是区分含秘载体和普通载体,并在训练时提供梯度以优化基于Im-Residual结构的编码器.需要注意的是,基于隐写分析的鉴别器只在模型训练时使用.模型只需使用基于Im-Residual结构的编码器和基于稠密结构的解码器就可以完成秘密信息的传递.本节将详细介绍HCGAN模型的三个组成部分及其总体架构,并描述其训练过程.

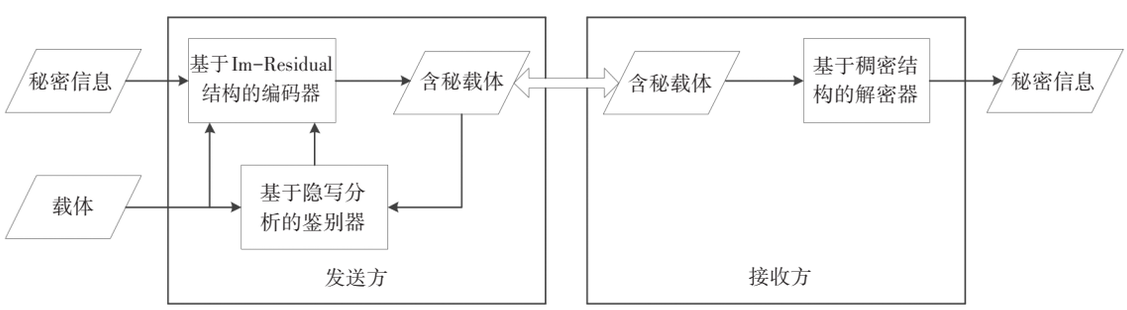


图1 基于GAN的高容量信息隐藏模型的整体框架

Fig.1 Overall framework of high capacity information hiding model based on GAN

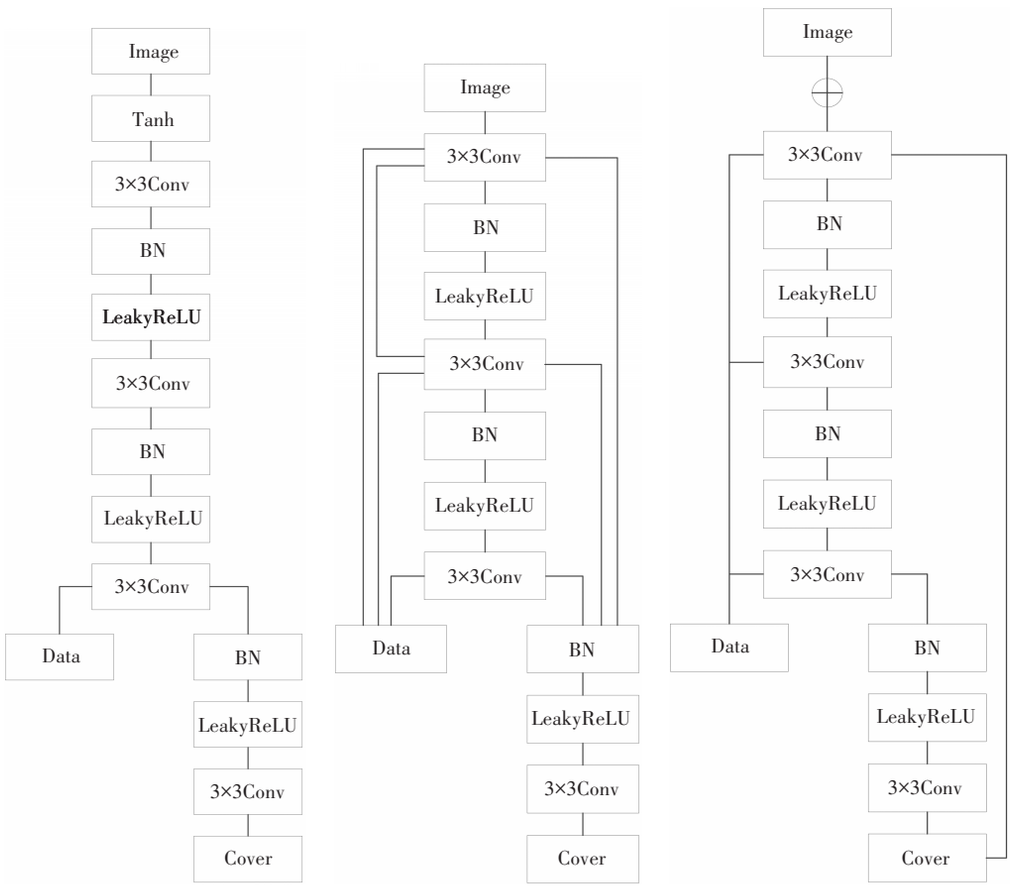
2.1 基于Im-Residual结构的编码器

为提高模型的隐写容量,本文在残差网络结构的基础上进行改进,得到了一种被称为Im-Residual的网络基础结构.Im-Residual结构通过增加秘密信息的直接映射来减少秘密信息在特征提取过程中的损失.基于Im-Residual结构的编码器更适用于信息隐藏,并且能够增加载体中可以嵌入的秘密信息.

编码器的目标是将秘密信息安全地嵌入载体图像中,从而方便解码器从中提取出秘密信息.如图2(a)所示,基于深度学习的信息隐藏算法的编码器一般使用多个基础卷积层+LeakyReLU激活函数+BN层结构作为卷积神经网络.因为卷积神经网络是通过卷积核提取特征,再将提取出来的特征归一化处理后作为下一个卷积核的输入,所以在此过程中不可避免地会因为提取特征而损失信息.对图像类型的秘密信息而言,损失一部分不重要的信息是可以接受的,并且不会影响秘密信息内容的传递.但是对于其他类型如文本类型的秘密信息而言,损失一部分信息是不可接受的.因此,为了使编码器结构能适用于其他类型的秘密信息,Kevin A. Zhang等提出使用Dense结构的卷积神经网络来降低编码器对于信息的损失,如图2(b)所示.但Dense结构也会导致输入增加,进而增加需训练的参数[28].

在研究和分析了Dense结构之后,我们发现其能

减少信息损失的核心原因不是Dense结构带来的特征的多次重复使用,而是对秘密信息的多次嵌入.因此,本文提出了基于Im-Residual结构的编码器,如图2(c)所示.在残差结构的基础上通过短路多次嵌入秘密信息,优化了编码器模型,减少了约1/3的训练参数,最终得到了更优的提取率和视觉性能.Dense结构和Im-Residual结构的参数对比如表1所示,表中D为嵌入率.



(a)基础的编码器　(b)Dense结构的编码器　(c)Im-Redisual结构的编码器

图2　三种编码器对比图

Fig.2　Comparison of three encoders

表1　Dense结构和Im-Residual结构的参数对比

Tab.1　Comparison of parameters between Dense structure and Im-Residual structure

网络层数	Dense	Im-Residual
Conv1	3×3×3×32	3×3×32
Conv2	3×3×(32+D)×32	3×3×(32+D)×32
Conv3	3×3×(2×32+D)×32	3×3×(32+D)×32
Conv4	3×3×(3×32+D)×3	3×3×(32+D)×3
总参数量	31 164+1 755D	20 160+1 755D

在本文中,基于Im-Residual结构的编码器通过Im-Residual结构的卷积神经网络将秘密信息嵌入载体图像中以得到含秘图像,如公式(1)所示:

$$S = C + E_{\text{Im-Redisual}}(C, M) \tag{1}$$

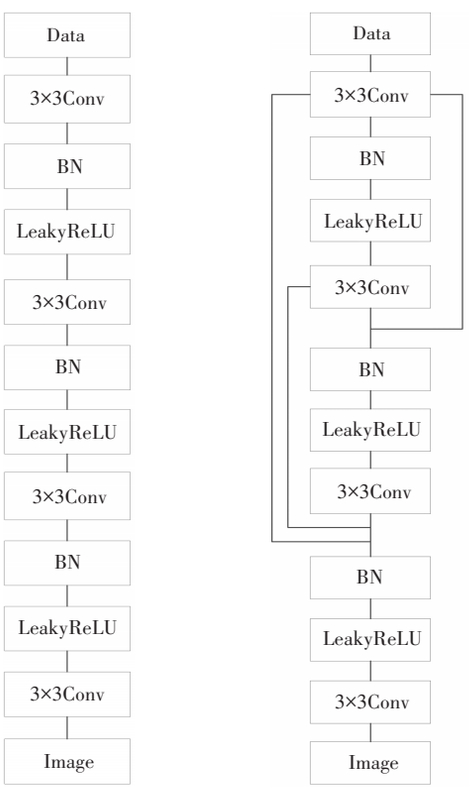
式中:C为$W \times H \times 3$的彩色载体图像,$M \in \{0,1\}^{D \times w \times H}$为将要嵌入的秘密信息,$S$为$W \times H \times 3$的彩色含秘图像.

2.2　基于稠密结构的解码器

在基于Im-Residual结构的编码器将更多的秘密信息嵌入载体中之后,如何从含秘载体中更好地提取出秘密信息就成了后续研究的重点.Im-Residual结构可以使用更少量参数完成将少量信息嵌入图片等含有大量信息的载体中,以减少过拟合风险并完成编码器工作.但在需要从大量信息中提取出少量信息的解码器场景中,Im-Residual结构就无法提供足够的信息供后续网络层使用.因此为提高解码器的提取能力,本文设计出了一种基于稠密结构的解码器.稠密连接网络中每一个网络层输出的特征映射会直接叠加到后面网络层的输入中,以

实现特征的复用.特征的复用保证了后面网络层的信息总是多于前面网络层的信息.基于稠密结构的解码器就是利用稠密结构特征复用的能力对含秘载体的特征进行多次处理以提高秘密信息的提取准确率.

基于深度学习的信息隐藏算法的解码器使用的也是基本的卷积层+LeakyReLu+BN层结构,如图3(a)所示.然而,网络层数的增加会带来信息损失的问题.稠密结构通过将卷积层提取出的特征叠加在之后所有的卷积层的输入中,来实现各层特征的复用并降低信息的损失.因此,本文设计出基于稠密结构的解码器来对含秘图像中的秘密信息进行提取,如图3(b)所示.实验结果显示基于稠密结构的解码器在高容量的信息隐藏算法中表现出更优的秘密信息提取率,并能指引编码器实现更优的视觉效果.



(a)基本的解码器 (b)基于稠密结构的解码器

图3 两种解码器对比图

Fig.3 Comparison of two decoders

本文模型将基于Im-Residual结构的编码器和基于稠密结构的解码器组合起来形成了一种自编码器.这种自编码器能够实现高隐写容量的信息隐藏,其过程如公式(2)所示:

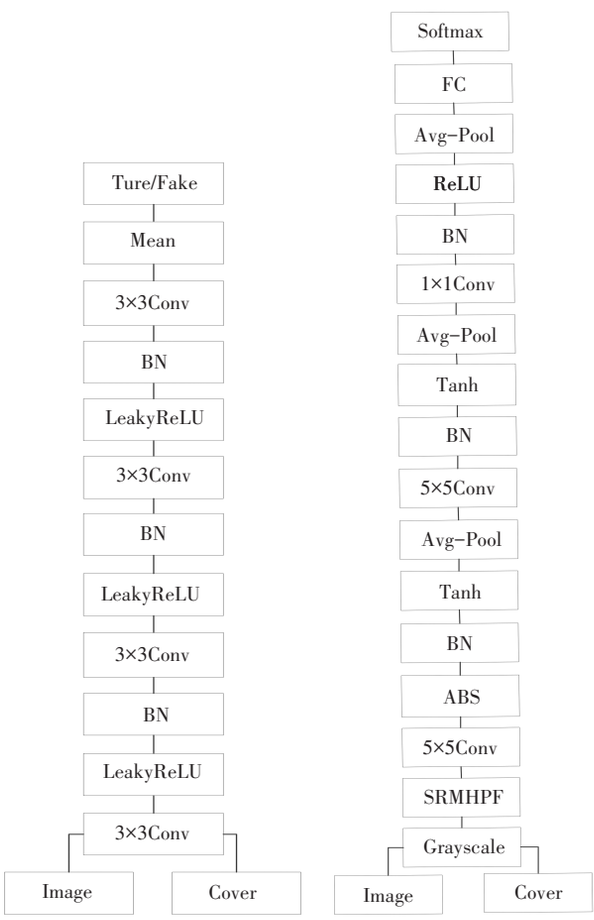
$$\tilde{M} = D_{\text{Dense}}(E(C,M)) \tag{2}$$

式中:$E(C,M)$代表$W\times H\times 3$的彩色含秘图像,$\tilde{M}\in\{0,1\}^{D\times w\times H}$是基于稠密结构的解码器从含秘图像中提取出的秘密信息,D为嵌入率,表示我们想要在载体图像的每个像素中嵌入的秘密信息比特数.

2.3 基于隐写分析的鉴别器

基于Im-Residual结构的编码器和基于稠密结构的解码器所构成的自编码器能够实现高隐写容量的信息隐藏,但是在含秘图像中嵌入大量的秘密信息会导致其抗隐写分析能力降低.因此,为了提高含秘图像的抗隐写分析能力,本文提出了基于隐写分析的鉴别器.该鉴别器利用其鉴别结果和GAN的对抗性来训练编码器,从而生成更安全的含秘图像.

本文在XuNet的基础上进行改进,使用全部的SRM滤波核和KV核作为预处理层,并在模型训练时将SRM滤波核的参数固定.相比于KV核,使用全部的SRM滤波核能够提供更多的噪声信息以便后面的神经网络进行特征提取和分类.为了适用于彩色含秘图像,本文在对含秘图像进行分析前加入了灰度层来将彩色含秘图像变为灰度含秘图像,以便后续的处理.基础的鉴别器和本文提出的基于隐写分析的鉴别器对比图如图4所示.



(a)基础的鉴别器 (b)基于隐写分析的鉴别器

图4 两种鉴别器对比图

Fig.4 Comparison of two discriminators

利用GAN的对抗博弈能力,本文提出的基于隐写分析的鉴别器和基于Im-Residual结构的编码器实现了相互对抗并且相互提供梯度,最终提高了基于Im-Residual结构的编码器输出含秘图像的抗隐写分析能力.在训练过程中,我们将含秘图像或载体图像输入基于隐写分析的鉴别器得到对应的鉴别结果,如公式(3)所示:

$$p = S((C + E_{\text{Im-Redisual}}(C, M)) \text{ or } C) \tag{3}$$

式中:C表示尺寸为$3 \times W \times H$的彩色载体图像,M表示需嵌入的秘密信息,p表示分类结果,一般用输入图像是载体图像的概率来表示.

2.4 HCGAN的整体架构和损失函数设计

将基于Im-Redisual结构的编码器、基于稠密结构的解码器和基于隐写分析的鉴别器组合起来就是本文提出的HCGAN算法的整体架构,如图5所示.其中,⊕代表像素相加.首先,发送者将秘密信息M和彩色载体图像C输入编码器网络来得到隐秘图像S.之后,接收者将隐秘图像S输入解码器网络得到解密后的信息$\tilde{M}$.基于隐写分析的鉴别器仅在训练时使用,用来判断S或C是否为隐秘图像,并通过对抗性训练优化编码器网络以提高隐秘图像S的抗隐写分析能力.

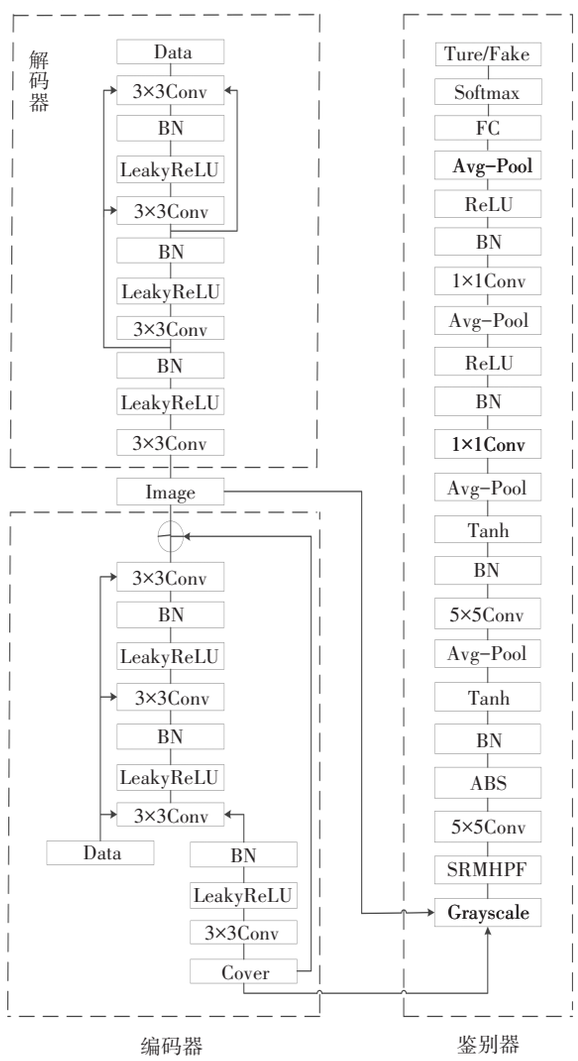


图5　HCGAN信息隐藏算法架构

Fig.5　HCGAN information hiding algorithm architecture

本文提出的基于GAN的高容量信息隐藏模型包含编码器、解码器和鉴别器三个部分,因此需要对这三个部分分别设计损失函数

2.4.1　相似性损失

基于Im-Residual结构的编码器作为嵌入过程需要将秘密信息嵌入载体图像中形成含秘图像,因此,首先要保证在载体图像和含秘图像之间不存在肉眼可以感知的失真.图像之间的相似性可以使用SSIM和PSNR来表述.为了便于计算,本文使用均方误差(也称L2损失)作为载体图像和含秘图像相似性的评价指标,如公式(4)所示:

$$l_{\text{mse}} = E_{C\sim P_C}\frac{1}{3 \times W \times H}\|C - E(C, M)\|_2^2 \tag{4}$$

式中:C表示载体图像,W和H表示载体图像的宽和高,P_C表示载体图像的分布,M表示秘密信息,$E(C, M)$则表示编码器网络的输出,即含秘图像.

2.4.2　准确率损失

准确率损失指的是基于稠密结构的解码器从含秘图像中提取出的秘密信息和原本的秘密信息之间的差距.交叉熵是一种机器学习的损失函数,可以用来衡量两个随机变量之间的差别,因此本文选择交叉熵函数作为算法的准确率损失,如公式(5)所示:

$$l_{\text{accuracy}} = E_{C\sim P_C}\text{CrossEntropy}\left(D\left(E\left(C, M\right)\right), M\right) \tag{5}$$

式中:P_C表示载体图像的分布,$D\left(E\left(C, M\right)\right)$表示解码器提取出来的秘密信息.交叉熵函数在这里用CrossEntropy表示,如公式(6)所示:

$$\text{CrossEntropy} = -\left[y\log\hat{y} + (1 - y)\log(1 - \hat{y})\right] \tag{6}$$

式中:y表示真实分类标签,在二分类中y可以为0或1,$\hat{y}$表示预测样本标签为1的概率.因此,交叉熵函数也可以表示为公式(7):

$$\text{CrossEntropy} = \begin{cases} -\log\hat{y}, & y = 1 \\ -\log(1 - \hat{y}), & y = 0 \end{cases} \tag{7}$$

2.4.3　安全性损失

基于隐写分析的鉴别器的目标是区分含秘图像和载体图像,同时具有一定的隐写分析的能力.因此和一般的GAN一样,本文使用鉴别器的输出作为安全性损失,如公式(8)所示:

$$l_{\text{security}} = E_{C\sim P_c}S(E(C, M)) \tag{8}$$

式中:$S(E(C, M))$表示鉴别器的输出.

2.4.4 编码器的损失函数

基于Im-Residual结构的编码器是模型的核心,其性能的优劣对上述三个损失都有很大的影响,因此本文将基于Im-Residual结构的编码器的损失函数定义为公式(9):

$$L_E = l_{\text{mse}} + l_{\text{accuracy}} + l_{\text{security}} \tag{9}$$

2.4.5 解码器的损失函数

与基于Im-Residual结构的编码器相比,基于稠密结构的解码器的损失函数就比较简单,只与准确率损失有关,如公式(10)所示:

$$L_D = l_{\text{accuracy}} \tag{10}$$

2.4.6 鉴别器的损失函数

基于隐写分析的鉴别器被用来区分载体图像和含秘图像,并在训练过程中给基于Im-Residual结构的编码器提供梯度.因此在训练整个网络时,基于隐写分析的鉴别器训练至关重要,它所提供的梯度直接影响着基于Im-Residual结构的编码器性能和含秘图像的质量.因此,本文使用Wasserstein距离作为基于隐写分析的鉴别器的损失函数,如公式(11)所示:

$$l_S = E_{C\sim P_c}S(C) - E_{C\sim P_c}S(E(C, M)) \tag{11}$$

每次迭代使用损失函数来判断神经网络的好坏,并使用梯度下降的方法来更新神经网络的参数.

3 实验分析

信息隐藏算法的性能可以从三个角度进行评估:图像中可以隐藏的数据量,即容量;含秘图像与载体图像之间的相似度,即失真;隐写分析工具避免检测的能力,即安全性.因此,本节分别设计实验验证容量、失真和安全性.若无特殊说明,所有实验均使用Pytorch框架实现,并在单个NVIDIA GTX1080 Ti GPU上进行训练.ADAM优化器用于最小化损失,学习率为0.000 4.

3.1 评价标准

3.1.1 峰值信噪比(Peak Signal-to-Noise Ratio, PSNR)

实验使用PSNR和SSIM[26]来评价含秘图像和载体图像的相似度.PSNR代表峰值信噪比,给定载体图像C、含秘图像S和可能的最大像素值MAX_1,PSNR的计算方式如公式(13)所示:

$$\text{MSE} = \frac{1}{W \times H}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}(C_{i,j} - S_{i,j})^2 \tag{12}$$

$$\text{PSNR} = 10 \times \log_{10}\left(\frac{\text{MAX}_1^2}{\text{MSE}}\right) \tag{13}$$

3.1.2 结构相似性(Structural SIMilarity, SSIM)

SSIM代表结构相似性,由载体图像C的采样x和含秘图像S的采样y之间的三个比较衡量——亮度、对比度和结构来计算,如公式(14)所示:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + r_1)(2\sigma_{xy} + r_2)}{(\mu_x^2 + \mu_y^2 + r_1)(\sigma_x^2 + \sigma_y^2 + r_2)} \tag{14}$$

式中:μ_x和μ_y分别是x和y的均值,σ_x和σ_y分别是x和y的方差,σ_{xy}是x和y的协方差.$r_1 = (k_1\text{MAX}_1)^2$、$r_{12} = (k_2\text{MAX}_1)^2$为两个常数,一般设$k_1$为0.01、$k_2$为0.03.

3.1.3 RS-BPP评价方法

为了测量基于深度学习的信息隐藏算法的有效容量,本文采用Zhang等人提出的RS-BPP[28]评价方法.该方法表示图像中可以可靠传输的平均比特数除以图像的大小.因此,RS-BPP可以直接与传统的隐写技术进行比较,其值等于$1 - 2p$,p为提取错误率.

3.2 失真和容量分析

本文在LUSN-bed数据集中对不同编码器-解码器结构在嵌入率$D \in \{1, 2, \cdots, 6\}$的情况下进行训练,并分别使用PSNR、SSIM和RS-BPP对各个算法的失真和容量在测试集中进行测试,结果如图6和表2所示.

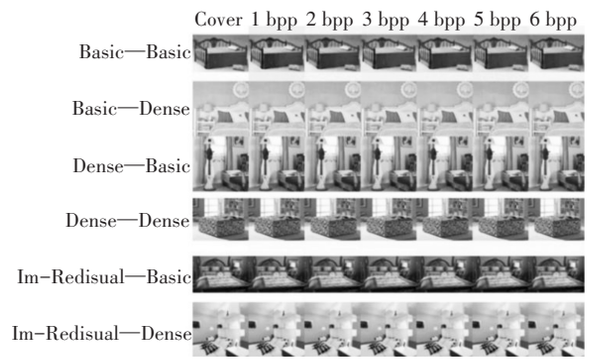


图6 不同编码器-解码器在不同嵌入率下的隐秘图像

Fig.6 Secret images of different encoders-decoders at different embedding rates

从图6中可以看出,就算是在6 bpp的嵌入率的情况下,隐秘图片也未显示出肉眼可见的失真.通过对比表2中高嵌入量下的基础提取器和基于Dense结构的提取器的结果可以发现,无论使用哪种编码

表2　不同编码器-解码器结构下各个算法的失真和容量

Tab.2　Distortion and capacity of each algorithm under different encoder-decoder structures

Decoder	*D*	Accuracy			PSNR/SSIM			RS-BPP		
		Basic	Dense	Im-Residual	Basic	Dense	Im-Residual	Basic	Dense	Im-Residual
Basic	1	0.96	1.00	1.00	29/0.84	44/0.98	44/0.98	0.93	0.99	0.99
	2	0.94	0.99	0.99	29/0.81	42/0.97	41/0.96	1.68	1.97	1.96
	3	0.87	0.97	0.97	30/0.83	39/0.94	38/0.93	2.26	2.84	2.83
	4	0.78	0.82	0.82	30/0.83	38/0.91	38/0.93	2.24	2.53	2.56
	5	0.71	0.76	0.76	29/0.83	40/0.95	40/0.94	2.11	2.62	2.64
	6	0.68	0.71	0.75	30/0.85	41/0.95	40/0,94	2.16	2.50	3.04
Dense	1	0.97	1.00	**1.00**	30/0.88	44/0.98	**45/0.98**	0.93	0.99	**0.99**
	2	0.95	0.99	**0.99**	31/0.86	41/0.96	**40/0.97**	1.79	1.96	**1.96**
	3	0.88	0.97	**0.98**	30/0.83	38/0.94	**39/0.93**	2.31	2.81	**2.86**
	4	0.78	0.91	**0.92**	30/0.84	38/0.92	**38/0.92**	2.24	3.32	**3.36**
	5	0.72	0.83	**0.85**	30/0.83	38/0.93	**39/0.92**	2.20	3.29	**3.53**
	6	0.68	0.77	**0.77**	30/0.86	40/0.94	**40/0.93**	2.16	3.25	**3.29**

器,基于Dense结构的解码器总是有更高的容量,可证明使用Dense结构确实可以提高解码器的提取准确率.同样,相同条件下基于Im-Redisual结构的编码器也显示了比其他编码器更高的容量和较小的失真.由表2可知,无论在哪种嵌入率下,HCGAN模型总是显示出较优的性能.

为了研究神经网络是如何嵌入秘密信息的,本文在6 bpp嵌入率下将HCGAN算法得到的隐秘图像减去载体图像得到残差图像并将其灰度化,如图7所示.从图7可以看出,HCGAN算法能够自适应地寻找嵌入位置,并在纹理复杂的区域嵌入更多的信息,在平滑的区域嵌入较少的信息.

为了进一步验证HCGAN算法的泛化能力,本文采用CelebA和COCO数据集在1 bpp、3 bpp、6 bpp的嵌入率下对本文模型进行试验,结果如表3和表4所示.由表可知,本文提出的算法适用于各种不同的图像集并表现良好.

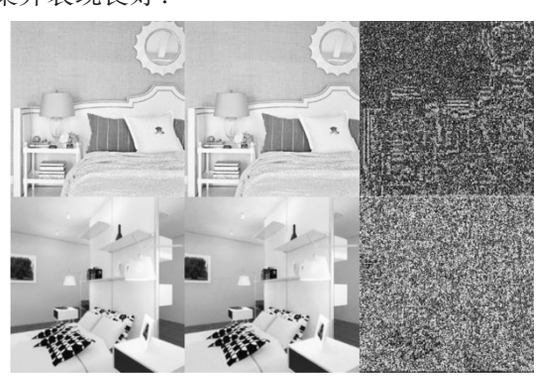


图7　载体图像(左)、隐秘图像(中)、残差图像(右)

Fig.7　Carrier image (left), hidden image(middle), residual image (right)

表3　解码器为Dense结构下不同数据集的表现

Tab.3　The performance of different data sets under the Dense structure

Dataset	Encoder	Accuracy			RSNR			SSIM			RS-BPP		
		1 bpp	3 bpp	6 bpp	1 bpp	3 bpp	6 bpp	1 bpp	3 bpp	6 bpp	1 bpp	3 bpp	6 bpp
CelebA	Im-Residual	**1.00**	**0.97**	**0.82**	**40**	**38**	**39**	**0.96**	**0.94**	**0.92**	**0.99**	**2.84**	**3.85**
	Dense	0.99	0.98	0.71	40	37	37	0.96	0.94	0.94	0.99	2.86	2.53
	Basic	0.97	0.91	0.67	30	30	30	0.88	0.87	0.87	0.94	2.47	2.07
COCO	Im-Residual	**1.00**	**0.97**	**0.80**	**37**	**37**	**37**	**0.94**	**0.92**	**0.90**	**0.99**	**2.84**	**3.61**
	Dense	1.00	0.97	0.76	38	36	37	0.95	0.90	0.72	0.99	2.82	3.07
	Basic	0.97	0.85	0.68	27	26	27	0.80	0.75	0.78	0.94	2.13	2.24

表4 解码器为Basic结构下不同数据集的表现

Tab.4 The performance of different data sets under the Basic structure

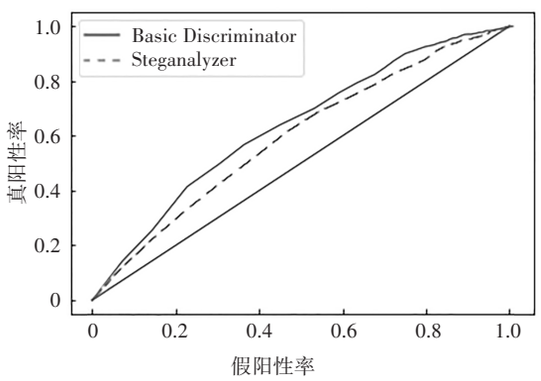
Dataset	Encoder	Accuracy			RSNR			SSIM			RS-BPP		
		1 bpp	3 bpp	6 bpp	1 bpp	3 bpp	6 bpp	1 bpp	3 bpp	6 bpp	1 bpp	3 bpp	6 bpp
CelebA	Im-Residual	**1.00**	**0.97**	**0.80**	**40**	**39**	**38**	**0.97**	**0.95**	**0.93**	**0.99**	**2.82**	**3.67**
	Dense	1.00	0.94	0.70	40	38	40	0.97	0.92	0.95	0.99	2.65	2.41
	Basic	0.98	0.90	0.68	30	30	31	0.88	0.85	0.88	0.95	2.43	2.13
COCO	Im-Residual	**1.00**	**0.97**	**0.80**	**39**	**37**	**37**	**0.95**	**0.91**	**0.92**	**0.99**	**2.82**	**3.60**
	Dense	0.99	0.97	0.74	38	36	37	0.95	0.91	0.92	0.99	2.80	2.89
	Basic	0.97	0.85	0.68	27	27	26	0.80	0.75	0.77	0.94	2.12	2.19

3.3 安全性分析

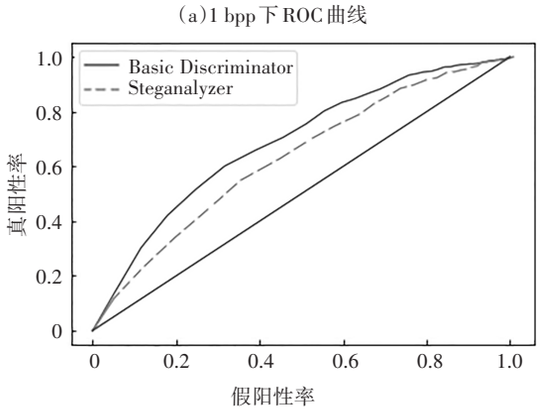
安全性是信息隐藏算法的另一种重要指标,其主要是指抗隐写分析能力,通常使用隐写分析工具的检测率进行评估.本节使用一个流行的开源隐写分析工具StegExpose(结合了几种现有的隐写分析技术,包括样本对、RS分析、卡方攻击和初级集)进行实验,来验证HCGAN模型生成不可检测的含秘图像的能力.为了进行对照实验,本文采用LUSN-bed数据集分别在1 bpp、3 bpp、6 bpp嵌入率下训练HCGAN模型和其他结构不变只改变鉴别器为普通卷积结构鉴别器的信息隐藏算法模型.模型训练完成后,分别给各个模型输入1 000个测试集中的图像得到1 000个隐秘图像,然后使用StegExpose工具对2 000个图像进行分析并画出ROC曲线,如图8所示.需注意的是,图中的直线表示随机判断的ROC曲线.

从图8中可以看出,无论是1 bpp、3 bpp还是6 bpp嵌入率的信息隐藏模型,使用隐写分析器作为鉴别器的HCGAN算法的抗隐写分析能力总是优于其他结构不变只改变鉴别器为普通卷积结构鉴别器的模型.即使在嵌入率为6 bpp的情况下,StegExpose工具的检测率也低于0.7,可以看出本文算法能够更有效地抵抗基于深度学习的隐写分析算法.

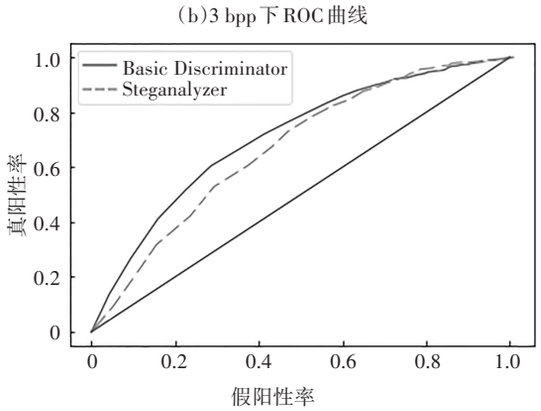
为了验证本文算法具有抵抗基于深度学习的隐写分析能力,以及避免由于对XuNet的针对性训练而引起的检测不可信问题,本文使用Yedroudj等人[29]于2018年提出的Yedroudj-Net隐写分析模型来检测本文模型的抗隐写分析能力,结果如表5所示.需要注意的是,Yedroudj-Net在训练时采取半盲检测方式,其训练集中包含各种嵌入率下分别由普通卷积结构的鉴别器和基于隐写分析的鉴别器引导的基于Im-Residual结构的编码器所生成的含秘图像.



(a)1 bpp下ROC曲线



(b)3 bpp下ROC曲线



(c)6 bpp下ROC曲线

图8 不同嵌入率下的ROC曲线

Fig.8 ROC curve under different embedding rates

从表5中可以看出,在1 bpp、2 bpp等较低嵌入率下基于隐写分析的鉴别器引导基于Im-Residual结构的编码器生成的含秘图像有着较低的检测准确率.在5 bpp和6 bpp的高嵌入率下,由于载体图像中嵌入了大量秘密信息所带来的大量扰动,从而使Yedroudj-Net隐写分析器能够实现接近100%的检测准确率.

表5　Yedroudj-Net隐写分析网络的检测准确率

Tab.5　The detection accuracy of Yedroudj-Net steganalysis network　%

鉴别器类型	嵌入率					
	1 bpp	2 bpp	3 bpp	4 bpp	5 bpp	6 bpp
基础卷积结构的鉴别器	66	80	84	98	99	99
基于隐写分析的鉴别器	62	72	83	97	99	99

最后,本文在Yedroudj-Net隐写分析模型下比较了HCGAN算法和经典信息隐藏算法WOW、S-UNIWARD在相似安全性下的嵌入率,结果如表6所示.注意,这里的训练集包含各嵌入率下分别由普通卷积结构的鉴别器和基于隐写分析的鉴别器所生成的含秘图像(作为负向样本)和载体(作为正向样本).从表6中可以看出,本文提出的信息隐藏算法能够在嵌入率为2.0 bpp下显示出比WOW和S-UNIWARD算法在0.4 bpp的嵌入率下更低的隐写分析检测率,即更高的安全性.这进一步验证了本文提出的信息隐藏模型能够实现高隐写容量并且在高隐写容量下保持较强的安全性.

表6　相似安全性下各个算法的嵌入率对比

Tab.6　Comparison of the embedding rate of each algorithm under similar security

算法	Yedroudj-Net检测率/%	嵌入率
WOW	85	0.4 bpp
S-UBIWARD	78	0.4 bpp
HCGAN	72	2.0 bpp

4　总结与展望

本文提出了一种基于GAN的高容量信息隐藏算法HCGAN,该算法能够以较高的提取率隐藏高容量秘密信息,并具有一定的抗隐写分析能力.实验表明,HCGAN算法能有效嵌入3 bpp以上高容量的秘密信息,并且在高容量下依旧具有较强的安全性.

信息隐藏技术常因新的创意或者其他领域的进展而有长足的发展.结合本文工作,本节提出了未来研究信息隐藏技术的一些思路.本文提出的模型利用了GAN的对抗性,但没有关注GAN能通过噪声生成符合真实图像分布的图像生成能力.因此将GAN的图像生成能力用于信息隐藏领域是后续研究的一个方向.本文提出的模型虽然拥有较高的安全性,但还是有可能被隐写分析检测出来.在近几年对于神经网络的研究中,有部分研究者发现对一张已经正确分类的图片进行细微的像素修改后,DNN会将其错误地分类为其他类型[30].在今后的工作中,将这种对抗性样本引入信息隐藏技术,使基于深度学习的隐写分析将含秘图像错误分类,以此来进一步提高含秘图像的抗隐写分析能力.

参考文献

[1] 张新鹏,钱振兴,李晟.信息隐藏研究展望[J].应用科学学报,2016,34(5):475-489.
ZHANG X P,QIAN Z X,LI S. Prospect of digital steganography research[J].Journal of Applied Sciences,2016,34(5):475-489.(In Chinese)

[2] 高华玲.信息隐藏关键技术研究综述[J].电子世界,2016(9):146.
GAO H L. Review on key technologies of information hiding [J]. Electronics World, 2016(9):146-146. (In Chinese)

[3] FRIDRICH J. Steganography in digital media [M]. Cambridge: Cambridge University Press,2009.

[4] CHAN C K,CHENG L M. Hiding data in images by simple LSB substitution[J].Pattern Recognition,2004,37(3):469-474.

[5] HINTON G E,SALAKHUTDINOV R R.Reducing the dimensionality of data with neural networks[J].Science,2006,313(5786):504-507.

[6] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M,*et al*. Generative Adversarial Networks[C]//Advances in Neural Information Processing Systems. Montreal, 2014:20672-2680.

[7] WU D C, TSAI W H. A steganographic method for images by pixel-value differencing[J]. Pattern Recognition Letters,2003,24(9/10):1613-1626.

[8] MIELIKAINEN J. LSB matching revisited[J]. IEEE Signal Processing Letters,2006,13(5):285-287.

[9] HSU C T,WU J L. Hidden digital watermarks in images[J]. IEEE Transactions on Image Processing,1999,8(1):58-68.

[10] BHATTACHARYYA S, SANYAL G. A robust image steganography using DWT difference modulation (DWTDM)[J]. International Journal of Computer Network and Information Security, 2012,4(7):27-40.

[11] PEVNÝ T, FILLER T, BAS P. Using high-dimensional image models to perform highly undetectable steganography[M]//Information Hiding. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010:161-177.

[12] HOLUB V, FRIDRICH J. Designing steganographic distortion using directional filters[C]//2012 IEEE International Workshop on Information Forensics and Security. Costa Adeje, Spain: IEEE, 2012:234-239.

[13] HOLUB V, FRIDRICH J, DENEMARK T. Universal distortion function for steganography in an arbitrary domain[J]. EURASIP Journal on Information Security, 2014, 2014:1.

[14] GUO L J, NI J Q, SU W K, *et al*. Using statistical image model for JPEG steganography: uniform embedding revisited[J].IEEE Transactions on Information Forensics and Security, 2015, 10(12): 2669-2680.

[15] DIOUF B, DIOP I, KEITA K W, *et al*. Adaptive linear programming of polar codes to minimize additive distortion in steganography[C]//2016 SAI Computing Conference (SAI). London, UK: IEEE, 2016:1086-1092.

[16] HAYES J, DANEZIS G. Generating Steganographic Images via Adversarial Training[C]//Advances in Neural Information Processing System. Long Beach, 2017: 1964-1963.

[17] TANG W X, TAN S Q, LI B, *et al*. Automatic steganographic distortion learning using a generative adversarial network[J]. IEEE Signal Processing Letters, 2017, 24(10): 1547-1551.

[18] YANG J H, LIU K, KANG X G, *et al*. Spatial image steganography based on generative adversarial network[EB/OL]. 2018: arXiv: 1804.07939[cs.MM].https://arxiv.org/abs/1804.07939.

[19] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. 2015:234-241.

[20] DENEMARK T D, BOROUMAND M, FRIDRICH J. Steganalysis features for content-adaptive JPEG steganography[J].IEEE Transactions on Information Forensics and Security, 2016, 11(8): 1736-1746.

[21] LIU M M, ZHANG M Q, LIU J, *et al*. Coverless information hiding based on generative adversarial networks[J]. Journal of Applied Sciences, 2018, 36(2): 371-382.

[22] 吴建斌,康子阳,刘逸雯,等. 基于图像分类的无载体信息隐藏方法[J].湖南大学学报(自然科学版),2019,46(12):25-32.
WU J B, KANG Z Y, LIU Y W, *et al*. Coverless information hiding algorithm based on image classification[J].Journal of Hunan University (Natural Sciences), 2019, 46(12): 25-32.(In Chinese)

[23] ZHANG Z, LIU J, KE Y, *et al*. Generative steganography by sampling[J]. IEEE Access, 2019, 7: 118586-118597.

[24] BALUJA S. Hiding images in plain sight: Deep steganography[C]// Advances in Neural Information Processing Systems. Long Beach: Neural Information Processing Systems, 2017: 2069-2079.

[25] ZHANG R, DONG S Q, LIU J Y. Invisible steganography via generative adversarial networks[J]. Multimedia Tools and Applications, 2019, 78(7): 8559-8575.

[26] DUAN X T, GUO D D, LIU N, *et al*. A new high capacity image steganography method combined with image elliptic curve cryptography and deep neural network[J].IEEE Access, 2020, 8: 25777-25788.

[27] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[28] ZHANG K A, CUESTA-INFANTE A, XU L, *et al*. SteganoGAN: high capacity image steganography with GANs[EB/OL]. 2019: arXiv: 1901.03892[cs.CV].https://arxiv.org/abs/1901.03892

[29] YEDROUDJ M, COMBY F, CHAUMONT M. Yedroudj-net: an efficient CNN for spatial steganalysis[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, AB, Canada: IEEE, 2018: 2092-2096.

[30] NGUYEN A, YOSINSKI J, CLUNE J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015: 427-436.