doi:10.3772/j.issn.1002-0470.2019.07.002

基于生成对抗网络的隐写术设计①

陈 璐②* 毛玮韵* 苏 磊* 赵 磊*** 孙志庆**

(*国网上海市电力公司电力科学研究院 上海 200437)

(**上海赛璞乐电力科技有限公司 上海 200437)

摘 要 随着数字多媒体的发展,网络数字媒体已经逐渐成为人们传递和获取信息的主要方式,同时,以数字媒体为载体的隐写术也得到了空前的发展。新型隐写术层出不穷,然而,据统计目前的隐写术在大部分情况下都被非法使用,因此,设计安全的隐写方法迫在眉睫。本文提出一种基于生成对抗网络(GAN)的隐写术,主要包括生成网络和判别网络,生成网络主要生成用于隐写的图像载体,判别网络主要区分原始图像和生成图像以及生成图像和生成图像经过嵌入得到的隐写图像。同时,在 CelebA 人脸数据集上进行了实验,验证本文提出方法的有效性和鲁棒性。

关键词 隐写术, 隐写分析, 生成对抗网络(GAN)

0 引 言

隐写术(Steganography)是信息隐藏的主要分支之一。信息隐藏是利用人类的感觉器官对数字信号的感觉冗余,将一组或者多组秘密信号隐藏到载体信息中,让攻击者在其不影响宿主信号的感觉效果和使用价值的情况下难以判断秘密信息是否存在,而使信息更加难以截获,从而保证信息传递的安全性。随着科学技术的发展[1],尤其是数字媒体技术的广泛应用[2,3],信息隐藏技术有了进一步的发展。隐写术经常应用在秘密通信中,尤其是在当今社交媒体纷繁芜杂的图像和视频环境中。因此,设计一种安全的隐写机制至关重要。

隐写分析(Steganalysis),是与隐写相对抗的技术,主要研究如何区分载体图像和隐写图像。在进行隐写分析的过程中,不需要了解嵌入密钥的知识,甚至不需要嵌入算法的相关知识。隐写分析可以通过其本身的性质,发现隐蔽通信的信道,阻止隐写分析方进行隐蔽通信。同时,还可以指导隐写算法的设计,提高隐写方法的安全性。

隐写术由于其本身的特性,目前经常被犯罪分子用来进行隐藏犯罪证据、非法通信甚至是泄漏机密等。目前,隐写术的研究主要关注安全性能和信息隐藏容量的问题,其目的就是用来隐蔽通信,因此,隐写方法的安全性至关重要。在现代隐写术中,一般对隐写方法的安全性进行两种评估:一种是使用一些主流的隐写分析检测方法进行评价;另一种是对图像质量进行评价检测,比如峰值信噪比(PSNR)、均方误差(mean square error,MSE)等。本文主要采用主流的隐写分析方法进行检测评估。

当人们设计隐写算法时,通常会启发式地从隐写分析方面进行考虑。例如,消息应该嵌入到更安全的区域,如图像的噪声和纹理区域中。本文提出了一种基于生成对抗网络(generative adversarial networks, GANs)的隐写方法,来实现安全的隐写。在生成对抗网络的框架下,使用生成网络来生成载体图像,然后使用目前最先进的嵌入算法对生成的图

① 国家自然科学基金(61501457)资助项目。

② 男,1981 年生,博士生,高级工程师;研究方向:数据挖掘和输变电设备在线监测;联系人,E-mail: shihaichao@iie.ac.cn

(收稿日期:2018-07-22)

像进行嵌入,并利用判别网络对图像的真假性以及是否为经过隐写的图像进行区分。

本文第 1 节概述了隐写方法、隐写分析方法以及生成对抗网络的发展。第 2 节对本文提出的方法进行简要分析。第 3 节给出了实验结果并进行分析。第 4 节对隐写和隐写分析领域进行了总结和展望。

1　相关工作

图 1 所示为隐写和隐写分析模型图。要嵌入的对象称为载体对象(图像、音频、视频等),准备嵌入的信息称为嵌入载体对象。通过隐写方法将秘密信息嵌入到载体对象中,就会得到载体加密对象,而接收方在掌握了载体加密密钥之后才能正确恢复原始图像,得到加密的信息。而密钥算法和载体对象对隐写分析方法都是未知的,信息嵌入位置更是无从下手,因此隐写分析方法只能对载体加密图像进行分析检测。

隐写术的主要关注点在隐藏容量和安全性两个方面。为了实现隐蔽通信,首要任务就是保证安全性。隐写方法的常用载体包括图像、视频、音频、文本等,在这些载体中,基于图像数字媒体的隐写方法发展最快,主要分为两类:一类是空域图像隐写术,直接对图像的像素值进行修改来实现嵌入。主要包括基于修改最低有效位(LSB)的方法,LSB matching[4]以及改进的 LSB matching 方法等。如果采用 LSB 作为隐写方法,则图像的统计特征将被破坏,隐写分析者就会很容易检测到秘密信息。为了方便和易于实现的目的,LSB 算法将秘密信息隐藏到每个像素的图像通道中的最低有效位。在大多数情况下,对 LSB 算法的修改也称为 ±1 嵌入[5]。它在最低有效位的像素中随机地加 1 或减 1,所以最后一位的比特位也会匹配其需要的位数。目前最流行、也最安全的空域图像隐写算法是图像内容自适应算法,包括 HUGO(highly undetectable SteGO)[6],WOW(wavelet obtained weights)[7], S-UNIWARD (spatial universal wavelet relative distortion)[8]等。HUGO 是一种通过在像素中嵌入一些信息来为像素分配代价,进而定义失真函数的隐写方法。它使用加权范数函数来表示特征空间。HUGO 被认为是最安全的隐写技术之一,本文将使用 HUGO 算法进行嵌入。WOW 是另一种内容自适应隐写方法,根据图像区域的纹理复杂度,将信息嵌入到原始图像中。在 WOW 算法中,图像区域越复杂,在这个区域像素值被修改得就越多。S-UNIWARD 引入一种独立于嵌入域的通用失真函数。这些内容自适应的图像隐写方法的实现细节多种多样,但最终目标都是致力于最小化损失函数,尽量将信息嵌入到噪声区域或者是复杂的纹理区域中,避免嵌入到平滑的图像区域中,以便更快更安全地实现秘密信息的嵌入。另一类是在图像的变换域进行隐写,常见的有基于 JPEG 图像的隐写方法,通过修改量化 DCT 系数进行数据嵌入的方法。这类方法主要有 OutGuess[9]、nsF5[10]以及内容自适应算法 J-UNIWARD 等。Outguess 算法将图像的统计特性保持,尽可能地在对图像统计特性影响小的系数上嵌入秘密信息,这种方法通过不同的种子值来生成不同的随机序列进而确定待修改的区域,选取影响最小的序列,并将秘密信息编码后进行嵌入。nsF5的嵌入效率比较高,也是在JPEG

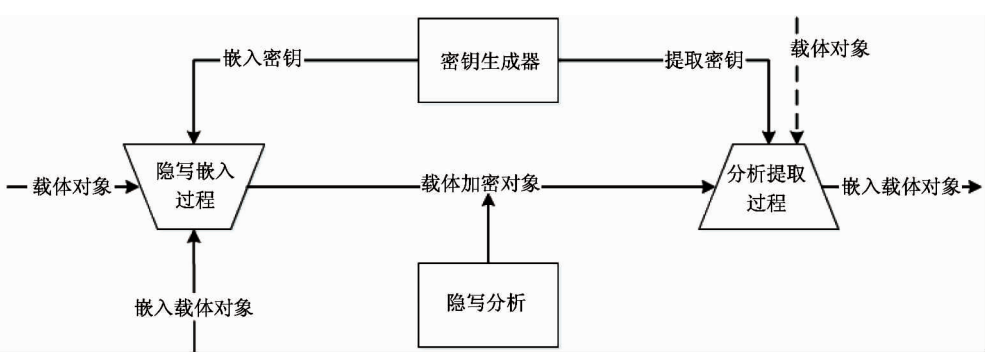

图 1　隐写和隐写分析模型

上嵌入,使用 Wet Paper 编码的方式,使得嵌入图像带来的修改不会导致系数变为零,安全性得到了提高。J-UNIWARD 同 S-UNIWARD 的原理一致,是在 JPEG 域上进行嵌入,先将图像解压缩到空域,再进行小波变换,进而从小波滤波得到的残差图像中计算失真函数,进行嵌入。

隐写分析目前最常用的方法就是使用模式识别的方法,可以将隐写分析看成是载体对象和载密对象两类的模式分类问题,本质上是一个二分类的问题。按照应用范围,一般将隐写分析分为两种类型:针对性隐写分析以及通用隐写分析。针对性隐写分析就是在已知隐写算法的情况下,检测图像中是否含有隐写信息;通用隐写分析是在隐写算法未知的情况下,对图像进行检测。通用隐写分析方法由于其对隐写方法的泛化性能较好,不需要知道隐写算法就可以进行检测,已成为目前主流的研究方向。通用的隐写分析方法是对图像中某些统计特性的破坏,例如像素之间的相关性等。近几年,隐写分析方法逐渐得到发展,主要有 SPAM[11], NIP[12], Rich Model[13]等。而在隐写分析的分类器设计方面,也是主要依赖于模式识别技术的发展,常用的分类器有支持向量机(SVM),集成分类器(Ensemble Classifier)等。

近年来,生成对抗网络已被成功应用于生成式和判别式卷积神经网络的图像生成任务中。Goodfellow 等人[14]提出了生成对抗网络的理论框架,并在不凭借任何监督信息的前提下生成图像。后来,Radford 等人[15]提出了一种深度卷积生成对抗网络(DCGANs)。然而,由于早期的生成对抗网络结果有些嘈杂和模糊,在训练网络时梯度往往会消失。为了解决梯度消失的情况,研究人员提出了 WGAN[16], WGAN 使用 Wasserstein 距离来代替 Jensen-Shannon 散度,提出将数据集的分布与生成网络学习到的分布进行比较,得出生成样本的质量是和网络的收敛性有关的,并且使得网络训练的速度显著提高。条件生成对抗网络(conditional GANs)[17]是对原始 GAN 的条件进行了扩展,通过额外地将类别标签输入到 GAN 的生成器和判别器中,而不是只输入噪声。Info-GAN[18]对 GAN 引入了一个新的概念,将输入的噪声 z 分为两部分:一部分是无法被解释的连续噪声信号;另一部分称为 C,表示潜在的属性,可以解释为面部表达,诸如眼睛的颜色,是否戴着眼镜等属性。

传统上,按照启发式的方式进行设计是不会完全考虑隐写分析方面。为了提高隐写的安全性,我们从隐写分析的方面考虑。受到生成对抗网络的启发,并基于隐写术和生成对抗网络的发展[19],本文提出一种基于生成对抗网络的隐写术模型。在 WGAN 的框架下,使用生成网络生成载体图像,在对载体图像使用隐写算法嵌入后,利用判别网络对图像的真假性以及是否为经过隐写的图像进行区分。

本文的主要创新点概括如下:

感知性 使用 WGAN 而不是其他方式的生成对抗网络来生成载体图像,以获得具有更高视觉质量的生成图像并确保更快的训练过程。

安全性 使用目前比较先进的隐写分析网络 GNCNN 来评估生成的图像是否适合做隐写。

多样性 还使用 GNCNN[20]来与生成网络进行对抗,从而使生成的图像更适合于嵌入。

2 基于生成对抗网络的隐写术设计

2.1 对抗学习

对抗训练采用博弈论的方法进行学习,并与无监督方式相结合,共同训练模型[21,22]。对抗学习的模型被训练为彼此进行竞争,不断地改善每个模型的输出结果。在生成对抗网络中,生成模型试图通过训练由噪声生成图像,判别模型相当于一个二分类器,区分生成网络生成的样本和真实样本。生成网络在欺骗判别网络的基础上,更新自身权重,同时通过区分真假样本来更新判别器的权重。对抗训练的模型可被描述为以下博弈竞争:

$$\min_G \max_D J(D,G) = E_{x\sim p_{data}(x)} \log(D(x)) + E_{z\sim p_{noise}(z)} \log(1 - D(G(z))) \tag{1}$$

其中,$D(x)$ 表示 x 是真实图像而不是生成图像的概率,$G(z)$ 是经过噪声输入后得到的生成图像。

在这个过程中,G 和 D 两个网络同时训练:

- 生成网络的输入是噪声分布为 $p_{noise}(z)$ 的噪声 Z,将其转变成数据分布 $p_{data}(x)$,生成与真实数据分布一致的数据样本。
- 判别网络的输入是真实数据和由生成网络生成的数据,并且区分真实数据和生成数据。

为了解决最小最大化的博弈问题,在每次随机梯度的优化迭代中,首先对判别网络使用梯度上升进行优化,对生成网络使用梯度下降进行参数更新。用 ω_N 表示神经网络 N,优化过程可以表示如下:

- 判别网络 D 固定,通过以下式子更新生成网络 G。

$$\omega_G \leftarrow \omega_G - \gamma_G \nabla_G J \tag{2}$$

$$\nabla_G J = \frac{\partial}{\partial \omega_G} E_{z\sim p_{noise}(z)} \log(1 - D(G(z,\omega_G), \omega_D)) \tag{3}$$

- 生成网络 G 固定,通过以下式子更新判别网络 D。

$$\omega_D \leftarrow \omega_D + \gamma_D \nabla_D J \tag{4}$$

$$\nabla_D J = \frac{\partial}{\partial \omega_D} \{E_{x\sim p_{data}(x)} \log(D(x,\omega_D)) + E_{z\sim p_{noise}(z)} \log(1 - D(G(z,\omega_G), \omega_D))\} \tag{5}$$

2.2　模型设计

本文提出的一种基于生成对抗网络的隐写方法,包含一个生成网络,用来生成载体图像;一个判别网络,具有两个功能,分别用作区分真图和假图以及区分生成图像和经过嵌入得到的隐写图像。模型结构如图 2 所示。

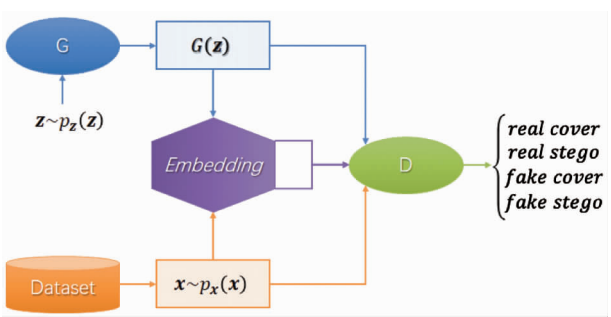

图 2　基于生成对抗网络的隐写方法结构图

对于输入噪声 z 的生成器 G,想要 G 生成接近于真实图片的并且可以被用作隐写载体的图像,同时让生成网络 G 和判别网络 D 相互博弈,最终达到纳什均衡。对于判别网络 D 具有两个功能,不仅要同生成网络相互对抗,难以区分生成的图像和真实图像,还要充当隐写分析网络,区分出隐写图和原始图像。因此,使用 $Stego(x)$ 来表示对生成的图像进行嵌入得到的隐写图像,该过程的目标函数表示如下:

$$\min_G \max_D J = \alpha\left(\begin{array}{c} E_{x\sim p_{data}(x)} \log(D(x)) \\ + E_{z\sim p_{noise}(z)} \log(1 - D(G(z))) \end{array}\right) + (1-\alpha) E_{z\sim p_{noise}(z)}\left[\begin{array}{c} \log D(Stego(G(z))) \\ + \log(1 - D(G(z))) \end{array}\right] \tag{6}$$

为了让实验结果更均衡,控制生成的图像的真实性和隐写分析的评估之间的权衡,我们让参数 $\alpha \in [0,1]$。证明当 $\alpha \leqslant 0.7$ 时,结果更接近噪声分布。

2.2.1　生成网络 G

生成网络 G,是用来生成隐写载体图像的。我们首先使用一个全连接层对噪声进行映射,然后是 4 个 fractionally-strided 卷积层和一个双曲正切激活函数层。网络结构如图 3 所示。

2.2.2　判别网络 D

判别网络,不仅可以用来评估生成图像质量的好坏,还可以用来对隐写图和非隐写图像分类。本文使用 1 个高通滤波层,4 个卷积网络层,然后是 1 个全连接层。网络结构如图 4 所示。

2.2.3　更新规则

使用随机梯度下降(stochastic gradient descent)进行优化更新,更新过程如下。

- 对于生成网络 G,计算如下:

$$\omega_G \leftarrow \omega_G - \gamma_G \nabla_G J$$

$$\nabla_G J = \frac{\partial}{\partial \omega_G} E_{z\sim p_{noise}(z)}[\log(1 - D(G(z,\omega_G), \omega_D))] + \frac{\partial}{\partial \omega_G}(1-\alpha) E_{z\sim p_{noise}(z)}[\log(D(Stego(G(z,\omega_G), \omega_D)))] + \frac{\partial}{\partial \omega_G}(1-\alpha) E_{z\sim p_{noise}(z)}[\log(1 - D(G(z,\omega_G), \omega_D))] \tag{7}$$

其中,ω_G 代表生成网络 G 的参数,γ_G 代表 G 的参数的折扣系数,$G(x)$ 代表生成的图像,$Stego(x)$ 代表生成的图像经过嵌入得到的隐写图像。

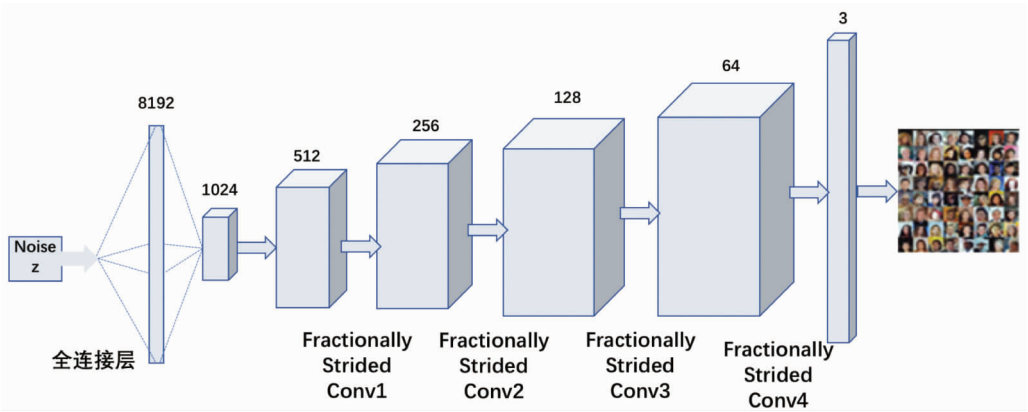

图 3　生成网络结构图

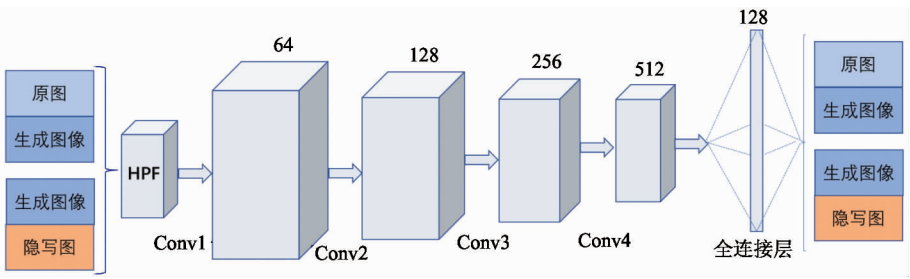

图 4　判别网络结构图

- 对于判别网络 D，计算如下：

$$\omega_D \leftarrow \omega_D + \gamma_D \nabla_D J$$

判别网络用于区分真图和假图时的计算过程如下：

$$\nabla_D J = \frac{\partial}{\partial \omega_D}\left\{\begin{array}{c} E_{z\sim p_{data}(x)}[\log D(x,\omega_D)] \\ + E_{z\sim p_{noise}(z)}[\log(1 - D(G(z,\omega_G),\omega_D))]\end{array}\right\} \tag{8}$$

判别网络用于区分生成图像和将信息嵌入到生成图像得到的隐写图像时的计算过程如下：

$$\nabla_D J = \frac{\partial}{\partial \omega_D} E_{z\sim p_{noise}(z)}\left[\begin{array}{c}\log D(Stego(G(z,\omega_G)),\omega_D) \\ + \log(1 - D(G(z,\omega_G),\omega_D))\end{array}\right] \tag{9}$$

其中，ω_D 代表判别网络 D 的参数，γ_D 代表 D 的参数的折扣系数，$D(x)$ 代表对输入到判别网络的图像进行判别的结果。

3　实　验

3.1　训练数据及软硬件环境

本文中所有模型均在 CelebA[23] 人脸数据集上进行，包含 202 599 张人脸图片，来自 10 177 个不同身份的人。所有的实验均在 TensorFlow[24] 平台上进行，为加快运行效率，使用一块 Titan X 显卡。首先对图像进行预处理，所有图像都被裁剪成大小为 64 ×64 像素。为了便于进行隐写分析，选取 90% 的数据作为训练集，将其余的数据作为测试集。用 TRAIN 表示训练集，用 TEST 表示测试集。使用 *Stego*(*x*)表示嵌入秘密信息的隐写算法，这样就构成了两个数据集，一个是 TRAIN + *Stego*(TRAIN)，*Stego*(TRAIN)表示训练集经过嵌入得到的图像。另一个数据集是 TEST + *Stego*(TEST)。最终，得到 380 000 张图像来做隐写训练，20 000 张图像作为测试集使用。

3.2　实验设置

为了体现所使用 WGAN 模型的优势，同 DC-

GAN(深度卷积生成对抗网络)进行了比较。同时使用 HUGO 算法对生成的图像进行嵌入,嵌入率为 0.4 bpp。两种生成对抗网络生成的载体图像如图 5 所示。

图 5　在 CelebA 数据集上,训练 7 个 epoch 生成的图像样本,左边为 WGAN 生成的图像样本,右边为 DCGAN 生成的图像样本

实验结果显示,WGAN 在训练时的收敛速度快,效果更加明显,观察数据如表 1 所示。

表 1　WGAN 和 DCGAN 运行时间对比

方法	时间(min)
WGAN	227.5
DCGAN	240.3

本实验使用 RMSProp 进行优化,学习率为 2×10^{-4},更新参数设置为 $\beta_1=0.5$ 和 $\beta_2=0.99$。在每一个 mini-batch 中,更新判别网络 D 的参数一次,更新生成网络 G 的参数两次。除了判别网络,还使用一个独立的隐写分析网络,来对比本文提出方法生成的隐写图像进行隐写分析的结果,并且使用如下的滤波器:

$$F^{(0)}=\begin{pmatrix}-1&2&-2&2&-1\\2&-6&8&-6&2\\-2&8&-12&8&-2\\2&-6&8&-6&2\\-1&2&-2&2&-1\end{pmatrix}$$

此外,使用 RMSProp 优化算法对该隐写分析网络进行训练,学习率设置为 5×10^{-6},更新参数设置为 $\beta_1=0.9$ 和 $\beta_2=0.99$,并且使用二值交叉熵来计算损失。本文分别在真实图像和生成图像上进行了实验,在真实图像上进行嵌入,并且使用隐写分析网络进行判别。然后在生成图像上进行嵌入,再使用隐写分析网络进行判别。

实验结果如表 2 所示。通过实验结果可以得出,即使普通的 WGAN 生成的图像经过嵌入,得到的隐写图像也会比较容易被隐写分析网络区分。本文提出的方法提高了隐写分析网络的分类错误率,这就意味着本文提出的方法生成的图像可以作为更加安全的图像载体进行隐写操作。

表 2　隐写分析网络在真实图像上训练得到的分类准确率

图像类型	使用 WGAN	使用 DCGAN
真实图像	0.87	0.92
生成图像	0.72	0.90

接下来的一组实验,验证在不同种子值(控制实验的可重复性)设置下的生成图像的安全性,使用不同的实验设置来进行实验。在这组实验中,使用 Qian[20] 的网络作为隐写分析网络,图像由本文提出的方法进行生成。实验输入是一些固定种子值的噪声分布 $p_{noise}(z)$,实验设置如下:

(1) 使用同样的种子值;

(2) 使用随机选择的种子值;

(3) 使用随机选择的种子值,并且在 WGAN 训练的过程中进行参数微调。

表 3 不同实验设置下的隐写分析网络的分类准确率

实验设置	分类准确率
(1)	0.87
(2)	0.72
(3)	0.71

如表 3 所示,通过使用不同的种子值来生成图像,也会比较容易欺骗隐写分析网络。

3.3 实验结果

本文提出的方法,在理论和实验上进行了证明。本文提出的方法可以使用随机密钥进行嵌入,同时,实验结果表明:一方面生成的图像更难以检测,表明其安全性能更高;另一方面生成的图像的视觉质量更好、更真实。

4 结 论

本文提出了一种基于生成对抗网络的隐写术模型,可以生成更加安全的载体图像,并用来做隐写。基于 WGAN 模型,对网络结构进行了阐述。本文提出的模型可以较为高效地生成更高视觉质量的图像,并且模型适合使用随机密钥嵌入生成图像,使得到的隐写图像比较难以检测。同时,本文使用 CelebA 数据集评估了本文模型的性能。结果表明本文提出的模型方法在应对隐写分析检测的分类准确性上,具有明显的优势。因此,认为这种方法将会在社交网络自适应隐写算法中得到应用,通过对隐写术进一步的探索,将会出现更多有效的方法。

-参考文献

[1] Liu Y, Zhang X, Zhu X, et al. ListNet-based object proposals ranking[J]. *Neurocomputing* (*NEUCOM*), 2017, 267:182-194

[2] Zhang X Y. Simultaneous optimization for robust correlation estimation in partially observed social network[J]. *Neurocomputing* (*NEUCOM*), 2017, 205: 455-462

[3] Zhang X Y, Wang S P, Yun X C. Bidirectional active learning: a two-way exploration into unlabeled and labeled dataset[J]. *IEEE Transactions on Neural Networks and Learning Systems* (*TNNLS*), 2015, 26(12): 3034-3044

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

[5] Andrew D K. Resampling and the detection of ISB matching in color bitmaps[C]. In: Proceedings of Security, Steganography, and Watermarking of Multimedia Contents VII, San Jose, USA, 2005. 1-15

[6] Pevnỳ T, Filler T, Bas P. Using high-dimensional image models to perform highly undetectable steganography[C]. In: International Workshop on Information Hiding, Canada, 2010. 161-177

[7] Holub V, Fridrich J. Designing steganographic distortion using directional filters [C]. In: IEEE International Workshop on Information Forensics & Security, Tenerife, Spain, 2012. 234-239

[8] Holub V, Fridrich J, Denemark T. Universal distortion function for steganography in an arbitrary domain[J]. *EURASIP Journal on Information Security*, 2014(1):1

[9] Provos N. Defending against statistical steganalysis[C]. In: Conference on Usenix Security Symposium, USA, 2001. 24-24

[10] Fridrich J, Pevnỳ T. Statistically undetectable JPEG steganography: deadends challenges, and opportunities[C]. In: Proceedings of the 9th Workshop on Multimedia & Security, Dallas, USA, 2007. 3-14

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

[12] Guan Q X, Dong J, Tan T N. An effective image steganalysis method based on neighborhood information of pixels[C]. In: Proceedings of International Conference on Image Processing, Brussels, Belgium, 2011. 2721-2724

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

[14] Goodfellow I, Adadie J, Mirza M, et al. Generative adversarial nets[C]. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada , 2014. 2672-2680

[15] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks [C]. In: Proceedings of International Conference on Learning Representations, Puerto Rico, 2016. 1-16

[16] Arjovsky M, Chintala S, Bottou L. Wasserstein genera-

tive adversarial networks[C]. In: Proceedings of the 34th International Conference on Machine Learning, Australia, 2017. 214-223

[17] Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier gans[C]. In :Proceedings of the 34th International Conference on Machine Learning, Australia, 2016. 2642-2651

[18] Chen X, Duan Y, Houthooft R, et al. InfoGAN: interpretable representation learning by information maximizing generative adversarial nets[C]. In: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems, Spain, 2016. 2172-2180

[19] Shi H C, Dong J, Wang W, et al. SSGAN: Secure Steganography Based on Generative Adversarial Networks [C]. In: Advances in Multimedia Information Processing, Pacific-Rim Conference on Multimedia, China, 2017. 534-554

[20] Qian Y L, Dong J, Wang W, et al. Deep learning for steganalysis via convolutional neural networks[C]. In: Media Watermarking, Security, and Forensics, USA, 2015. 94090-94090J

[21] Zhang X Y. Interactive Patent classification based on multi-classifier fusion and active learning[J]. *Neurocomputing* (*NEUCOM*), 2014, 127: 200-205

[22] Zhang X Y, Wang S P, Zhu X B, et al. Update vs. upgrade: modeling with indeterminate multi-class active learning[J]. *Neurocomputing* (*NEUCOM*), 2015, 162: 163-170

[23] Liu Z W, Luo P, Wang X G, et al. Deep learning face attributes in the wild[C]. In: IEEE International Conference on Computer Vision, Chile, 2015. 3730-3738

[24] Abadi M, Agarwal A, Barham P, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems[J]. *arXiv*:1603.04467, 2016

Design of steganography based on generative adversarial networks

Chen Lu[*], Mao Weiyun[*], Su Lei[*], Zhao Lei[*,**], Sun Zhiqing[**]

([*]State Grid Shanghai Electric Power Company Electric Power Research Institute, Shanghai 200437)

([**]Shanghai Saipule Power Technology Co., Ltd., Shanghai 200437)

Abstract

With the development of digital multimedia, the network digital media has gradually become the main way for people to transmit and acquire information. Therefore, the steganography based on digital media has also experienced an unprecedented development. However, it is estimated that the current steganography is illegally used in most cases. Therefore, the design of secure steganography is imminent. In this paper, a novel strategy of steganography based on generative adversarial networks is proposed, which includes a generative network and a discriminative network. The former mainly generates image carrier for steganography, while the latter is utilized to distinguish the original images from generated images, as well as the generated images from the stegos obtained by the generated images. At the same time, experiments are conducted on the CelebA dataset to verify the effectiveness and robustness of the proposed method.

Key words: steganography, steganalysis, generative adversarial network (GAN)