基于深度学习的图像隐写方法研究

原创

岁月漫长_ 已于 2022-03-30 15:27:03 修改 2403 收藏 2

文章标签： 深度学习 人工智能 计算机视觉

于 2022-02-21 21:33:21 首次发布

本文链接：https://blog.csdn.net/qq_40859587/article/details/123051299

版权

论文：付章杰, 王帆, 孙星明, 等. 基于深度学习的图像隐写方法研究[J]. 计算机学报, 2020, 43(9): 1656-1672.

图像隐写工具箱：http://dde.binghamton.edu/download/

摘要

2014年GAN出现，2016年第一个基于深度学习的隐写模型——SGAN

四类基于深度学习的隐写模型：

1）基于生成载体式

2）基于嵌入载体式

3）基于合成载体式

4）基于映射关系式

引言

原始空域隐写算法：LSB，±1嵌入

自适应隐写：S-UNIWARD、HILL、**WOW**、**HUGO**

隐写分析

隐写分析的具体过程分为两步：**特征提取**和**分类器训练**，通常先利用**高通滤波器**获取**残差图像**，再利用各种统计模型提取隐写分析特征。

传统隐写分析方法： Spatial Rich Model，SRM、maxSRM

基于深度学习的隐写模型：

表 1　基于深度学习的隐写方法分类

方法	具体实现	优缺点	隐写模型
生成载体式深度学习隐写模型	利用深度学习生成更加适合隐写的载体图像，再利用传统隐写算法实现信息的隐藏和提取	深度学习与隐写术相结合；载体图像不真实，安全性不高	SGAN[15]、SSGAN[16]等
嵌入载体式深度学习隐写模型	利用深度学习在自然载体图像上完成或辅助完成秘密信息的嵌入和提取	扩大隐写容量、提升隐写安全性；载体图像中的改动嵌入痕迹较大；	Deep Steganography[17]、HiDDeN[18]、ASDL-GAN[19]、SteganoGAN[20]、ADV-EMB[21]等
合成载体式深度学习以下内模型	利用深度学习对已有载体图像上进行二次改动生成新的图像并完成秘密信息的嵌入和提取	秘密信息的完整提取；提取时需要依靠额外的 mask	Liu's model[22]、SGSRGAN[23]
映射关系式深度学习隐写模型	秘密信息与随机噪声或目标对象之间建立映射关系，再利用深度学习完成秘密信息的隐藏和提取	载体图像没有嵌入痕迹；隐写容量小、秘密信息难以恢复	Hu's model[24]、Zhang's model[25]、Meng's model[26]

2.1 基于生成载体式深度学习隐写模型

2017 年 Volkhonskiy 等提出的 SGAN 模型中被首次提出，模型首先将**随机噪声**作为输入，通过 DCGAN生成尽可能真实的载体图像，再使用传统的 **±1嵌入**算法实现秘密信息的隐藏，最后将含密图像作为隐写分析网络的输入，在**隐写分析对抗训练过程**中提升其抗检测能力。但训练过程不稳定且生成的载体图像质量较差。
2018年SSGAN，将载体图像的生成网络替换成 Wasserstein GAN(简称为WGAN), WGAN 的损失函数使用对距离分布更加敏感的 EM 距离(又称为 Wasserstein distance)来提供更加有意义的梯度，从而生成更加符合真实分布的载体图像。

表 2　基于生成载体式深度学习隐写方法分类

方法	实现细节	优点	缺点
SGAN[15]	利用 DCGAN 生成适合隐写的载体图像再与隐写分析对抗，提升隐写安全性	生成的载体图像更加适合隐写	生成载体图像质量较差
SSGAN[16]	在 SGAN 基础上，将 DCGAN 替换成 WGAN	相较于 SGAN，生成载体图像质量更好，抵抗 GNCNN 的检测能力增强	未对隐写算法本身做出改进 生成图像质量一般

2.2 基于嵌入载体式深度学习隐写模型

2.2.1 自动学习嵌入改变概率以及嵌入代价

现有的自适应隐写算法都是基于最小嵌入失真框架设计的。
S-UNIWARD 使用的是一种与嵌入域无关的通用失真函数；
HILL 是在像素内嵌入一些信息的效果，为像素分配成本，定义失真函数域，使用加权范数将像素压缩到特征空间；
WOW 则是根据复杂区域将秘密信息嵌入到载体图像中，如果图像的区域在结构上比另一区域更加复杂，则该区域像素值被更改.
Li 等人在 2017 年提出利用生成对抗网络自动**学习嵌入失真代价**，从而寻找最小失真嵌入位置的隐写模型 ASDL-GAN，由三部分组成：**生成器，嵌入模拟器，判别器**。

1. 生成器将真实图像（又称为载体图像）作为输入，生成“非 0”的嵌入改动概率图；
2. 把概率图放入嵌入模拟器（TES）“模拟”秘密数据的嵌入，生成与载体图像同尺寸大小的改动位置映射图，该图每个像素点的取值范围都在{-1，0，1}之间；
3. 将载体图像与改动位置映射图进行点对点相加，生成“模拟”的含密载体，判别器检测载体图像是否含有秘密消息.

问题：收敛慢

2.2.2 编码-解码网络

先将多个数据（比如:文字、数据、图像等）先融合，然后再进行抽离或提取.
Hayes 等人提出的 SteGAN，利用Alice、Bob 和 Eve三方对抗游戏，分别用来进行信息隐藏、信息提取与隐写分析。但未充分考虑到含密图像的图像质量以及与真实载体图像之间存在的差距。
中科院Wang 等人[31]提出了 SsteGAN 模型，该模型在SteGAN 的基础上添加了一个 Dev 方，Alice 方通过与 Dev 方的对抗训练缩减含密图像与原载体图像之间的距离，然后生成更加真实的含密图像。
Zhu 等人HiDDeN，编码-解码网络由卷积网络构成，考虑到含密载体在通信传输过程中的安全问题在，编码和解码网络之间加入了噪声层进行噪声建模. 编码网络经由噪声层“模拟”生成带有噪声的含密图像，解码网络将其作为输入进行秘密信息的提取 。但隐藏容量最高只能达到 0.2bpp。
Zhang等人[20]在 2019 年提出了 SteganoGAN（以图藏比特数据），该隐写模型可以在载体图像中隐藏任意二进制比特数据，SteganoGAN 模型由三部分组成：编码器，解码器以及评价方.
Baluja 等人[17]在2017 年首次提出**以图藏图**的深度学习隐藏网络Deep Steganography.
Baluja 等人则以突破隐写容量为目标不断研究，在 2019 年又提出了“一图藏多图”.
Atique 等人[37]从另一角度出发，将秘密图像由3 通道的**彩色图像更换为单通道的灰度图像**，Zhang 等人[38] 提出的 ISGAN 模型在Y 通道嵌入灰度图像。

问题：

- 图像质量下降，颜色失真；
- 解码网络无法百分百的实现秘密消息的重新提取；
- 编码-解码网络这种端对端隐写模型并不适用于现实情况，现有编码网络大多直接将输出的**浮点型张量**作为解码网络的输入，从而实现秘密图像的重建. 但是，在现实世界中发送方必须要将编码网络输出的**张量转换成图像**后才能发送给接收方进行提取. 因此，这个过程必然会造成张量信息的丢失从而造成解码网络提取准确度的下降.

2.2.3 对抗样本

利用对抗攻击或者对抗噪声去欺骗基于深度学习的隐写分析二分类器

基于嵌入载体式深度学习隐写模型优缺点对比

表 3　基于嵌入载体式深度学习隐写模型优缺点对比

方法	实现细节	优点	缺点
ASDL-GAN[19] UT-SCA-GAN[29]	依据与隐写分析对抗网络自动学习嵌入失真代价和嵌入改变概率，实现信息隐藏.	自动学习最小嵌入失真代价	相较于传统自适应隐写算法,抗隐写分析检测能力没有很大提升
Deep Steganography[17] Stegnet[32] Atique's model[37] ISGAN[38]	利用编码-解码网络实现以图藏图的隐写模型	扩大了隐写容量	含密图像颜色失真; 重构秘密图像质量有损
SteGAN[30] SsteGAN[31] HiDDeN[18] SteganoGAN[20]	利用编码-解码网络实秘密二进制数据及文字信息的隐藏和提取,并添加第三方网络进行对抗训练，增强隐写的安全性	提升隐写安全性和鲁棒性	难以抵抗基于深度学习隐写分析模型的检测
Ma's model[40] Zhang's model[42] ADV-EMB[21]	利用与隐写分析网络在对抗训练过程中,生成对抗扰动或对抗梯度图 ，干扰隐写分析的判别结果	抵抗基于深度学习的隐写分析模型的检测	模型结构庞大，训练耗时久; 含密图像泛化性差

2.3 基于合成载体式深度学习隐写模型

利用 GAN 网络进行图像合成（比如图像修复、图像拼接、前景生成等）的同时完成信息隐藏。

表 4　基于合成载体式深度学习隐写模型优缺点对比

方法	实现细节	优点	缺点
Liu's model[22]	利用卡丹格将秘密信息隐藏到图像破损区域，利用 DCGAN 进行图像修复	提供一种新型的隐写方法，在图像修复过程中完成隐写扩大了可用作隐写的载体图像类型	信息提取准确度不稳定
SGSRGAN[23]	利用 MC-GAN 生成适合隐写的复杂前景区域，LSBM 隐写后与前景与载体图像进行拼接合成	LSBM 隐写算法安全性提升且含密图像质量高	信息提取需要依赖额外提供 mask，容易引起起攻击方的怀疑

基于合成载体式深度学习隐写方法目前有两种实现途径：

1. 是先利用生成对抗网络对原始载体图像进行二次改动，生成全新的图像，并在新生成的图像区域利用传统隐写算法完成秘密信息的嵌入；
2. 另一种是在原始载体图像受损区域中利用传统隐写算法嵌入秘密信息，然后将图像放入生成对抗网络完成对图像的修复.

但是，该类隐写方法在秘密信息提取时都需要依靠额外的 mask（比如前景图像轮廓、卡丹格等）才能完成，这对隐写模型的安全性造成了一个潜在的威胁. 同时，该类隐写模型中生成对抗网络性能的好坏对于含密图像质量有着很大影响。

2.4 基于映射关系式深度学习隐写模型

表 5　基于映射关系式深度学习隐写模型优缺点对比

方法	实现细节	优点	缺点
Hu's model[24]	将随机噪声与二进制秘密信息之间构建映射关系，再通过 DCGAN 将噪声作为输入生成载体图像（含密图像），并构建提取网络恢复噪声序列，最后通过映射关系恢复处秘密信息	生成的载体图像中没有嵌入和修改痕迹	隐写容量小 生成的图像不够真实和自然 秘密消息不能完全正确提取
Zhang's model[25]	在 Hu 的基础上，利用预训练好的 CycleGAN 对生成图像进行风格迁移，并在恢复原始生成图像过程时，直接将噪声向量作为输出	风格迁移后的图像具有更高的安全性	秘密信息的提取准确度低
Meng's model[26]	利用 faster RCNN 检测在载体图像上检测目标物体的颜色，类型等属性与秘密消息构建映射关系	相较于传统无载体隐藏方法，能够更加自动化的获取映射关系对应的秘密信息	需要建立大型数据集构建寻找与映射关系相对应的载体图像

Hu 等人[24]依据 SWE（无嵌入隐写）又称作无载体隐写思想提出了一个基于 DCGAN 的隐写网络模型，，将秘密信息 m 转换成二进制序列，与[-1,1]范围的 100 维随机噪声向量 z 构建映射关系，然后将该随机噪声作为DCGAN 网络的输入，经过对抗训练输出尽量符合自然图像统计分布的生成图像，该图像也可看作含密载体图像. 接收方利用由反卷积层组成的提取网络提取出秘密信息.

Zhang 等人[25]在 Hu 等人的基础上，利用 CycleGAN 对噪声生成的图像 C 进行了风格迁移，并对风格迁移后的图像 C'进行恢复. 在恢复原始生成图像 C 的过程中，该隐写模型直接将恢复的噪声向量 z 作为输出，最后接收方通过映射关系完成噪声向量到秘密信息的转换.

Meng 等人[26]利用 VGG-19 目标检测网络以及映射关系，将秘密信息与图像中多个目标的类别、颜色等特征形成映射关系. 该隐写方法首先选择符合秘密消息对应映射关系的自然载体图像，再使用 VGG-19 检测出目标所在位置以及适合隐写的安全区域，最后使用隐写算法完成对秘密消息的隐藏和提取。

问题：隐写容量很低

深度学习隐写模型问题：

生成的载体图像不够真实自然。可以结合更加先进的 BEGAN 或者 PG-GAN 等网络，生成高清、纹理细节丰富的载体图像。

利用 FGSM 攻击算法生成对抗含密图像耗时久，泛化性差。目前隐写模型中生成对抗样本采用的方法都是—FGSM。

基于编码-解码网络大容量隐写模型安全性差，难以抵抗隐写分析模型的检测。利用**对抗样本**能够有效提升基于编码-解码网络的大容量隐写模型的安全性。

提取网络无法实现秘密信息的无损恢复。针对基于映射关系式和基于编码-解码网络。

基于对抗样本的大容量隐写方法

论文思路：首先利用隐藏网络将秘密图像 S 编码到载体图像 C 中，然后将生成的普通含密图像 C'放入对抗噪声生成网络，让它与目标隐写分析对抗来生成对抗噪声. 最后将叠加了对抗噪声的含密图像C''放入解码网络中提取秘密图像.

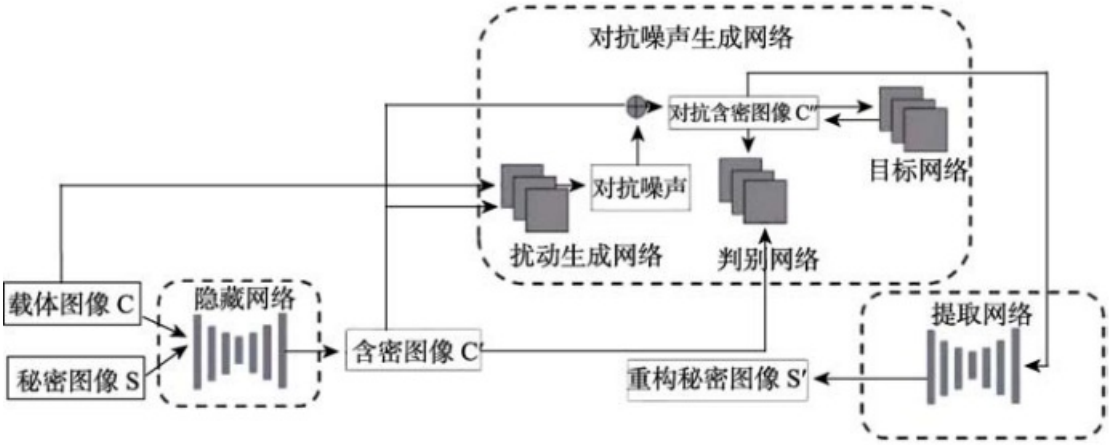

图 12　对抗隐写图像的隐藏和提取过程 CSDN @岁月漫长_

对抗噪声生成网络也包含三个子网络：噪声生成网络、判别网络、目标网络。
噪声生成网络 G 将含密图像 Stego 作为输入，生成能够成功干扰隐写分析分类的对抗噪声.
判别网络 D 用于缩小对抗含密图像与含密图像之间的差别，减少对抗噪声对于含密图像质量造成的影响.
目标网络 f 即预训练好的隐写分析网络通过输出的概率值，引导对抗噪声的生成，最终使得叠加了对抗噪声的含密图像 C''尽可能被分类为载体图像.