深度学习在图像隐写术与隐写分析领域中的研究进展

翟黎明, 嘉 炬, 任魏翔, 徐一波, 王丽娜

武汉大学 国家网络安全学院 空天信息安全与可信计算教育部重点实验室 武汉 中国 430072

摘要 隐写术与隐写分析是信息安全领域的热门研究方向, 近年来得到了广泛的研究与快速的发展。随着深度学习新技术的兴起, 深度学习也被引入到隐写术与隐写分析领域, 并在方法和性能上取得了一系列突破性的研究成果。为推进基于深度学习的隐写术与隐写分析的研究, 本文对目前的主要方法和代表性工作进行了归纳与探讨。对于图像隐写术与隐写分析这两个领域, 本文分别各自比较了传统方法和与相关深度学习方法的异同, 详细介绍了目前主要的基于深度学习的图像隐写术与隐写分析的基本原理和方法, 最后讨论了基于深度学习的图像隐写术与隐写分析仍需要解决的问题及未来的研究趋势。

关键词 隐写术; 隐写分析; 深度学习; 生成对抗网络; 卷积神经网络
中图法分类号 TP391 DOI 号 10.19363/J.cnki.cn10-1380/tn.2018.11.01

Recent advances in deep learning for image steganography and steganalysis

ZHAI Liming, JIA Ju, REN Weixiang, XU Yibo, and WANG Lina

Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China

Abstract steganography and steganalysis are hot research directions in the area of information security, and they have been widely researched and rapidly developed in recent decades. With the rise of new technologies for deep learning, steganography and steganalysis based on deep learning have achieved a series of breakthrough in methods and performance. In order to promote the research of steganography and steganalysis based on deep learning, typical methods and representative work are summarized and discussed in this paper. For image steganography and steganalysis, the similarities and differences between conventional methods and deep learning based methods are compared respectively, and the basic principles and methods of image steganalysis and steganography based on deep learning are introduced in detail. Finally, the problems to be solved and the future research directions are discussed.

Key words steganography; steganalysis; deep learning; generative adversarial network; convolutional neural network

1 引言

隐写术与隐写分析是信息安全领域中的研究热点, 近年来受到学术界越来越多的关注[1]。

隐写术是一种用于隐秘通讯的方法与技术, 它通过把秘密信息嵌入在数字图像等多媒体载体中而尽可能地不改变载体的视觉和统计特性, 以达到掩盖“正在进行秘密通讯”的目的[2]。互联网中广泛传播的多媒体数据为信息隐藏提供了丰富的隐藏载体; 网络上已公开的信息隐藏软件多达上千种, 这些软件简单易用, 降低了信息隐藏的应用门槛。隐写术在提高隐蔽通讯安全性的同时, 也容易为不法分子所利用。伴随着世界各国反恐和维稳形势日趋严峻, 隐写分析技术受到高度重视。

作为隐写术的反向检测手段, 隐写分析的目标是根据载体的统计特性来判断其中是否隐藏有秘密信息, 进而估计嵌入的信息长度、识别隐写工具、估计隐写秘钥, 最终提取秘密信息[3]。其中, 判断载体中隐藏信息的有无是阻断隐蔽通讯的关键, 这被称作被动隐写分析, 也是当前学术界研究的重点。

通讯作者: 王丽娜, 博士, 武汉大学教授, Email: lnwang@whu.edu.cn。
本课题得到国家自然科学基金项目(No. U1536204; No. 61876134; No. U1536114)资助。
收稿日期: 2018-08-30; 修改日期: 2018-09-27; 定稿日期: 2018-09-28

隐写术与隐写分析的研究一直在对抗中相互促进、相互借鉴。隐写术从早期的 LSB 隐写方法开始, 通过各种修改方式和隐藏策略来提高安全性。与之相对应的则是特定隐写分析的出现, 它根据隐写方法的嵌入特点开展针对性的检测。随后隐写术与纠错编码理论相结合, 以减少载体元素的修改数量来降低被检测的风险。隐写方法的多样性与安全性又促进了通用隐写分析的发展。通用隐写分析采用机器学习的方法, 提取通用性的隐写分析特征, 结合支持向量机、神经网络等分类器来训练通用的检测模型, 以达到对多种隐写方法的检测目标。基于机器学习的通用隐写分析使得隐写术采用新的隐藏策略——内容自适应策略——来提升安全性。自适应隐写方法设计失真代价来衡量每个载体元素的适宜修改程度, 在嵌入信息的同时降低总体失真。为了更好的检测自适应隐写方法, 隐写分析通过构造更高维度的特征(富模型)以全面捕获隐写嵌入带来的统计异常, 并使用集成分类器来降低高维特征的训练难题。隐写失真代价设计与隐写分析特征的构造分别是当前隐写术与隐写分析研究的主要问题, 这二者往往也会相互转化, 例如利用隐写分析特征来设计失真代价, 利用失真代价构造选择通道来增强隐写分析特征。

深度学习是机器学习领域中的新兴方向, 旨在通过模拟人脑自动地学习数据各个层次的抽象特征, 从而更好地反映数据的本质特征[4]。自 2006 年, Hinton[5]提出一种基于概率图模型的多层受限波尔兹曼机(Restricted Boltzmann Machine, RBM)后, 深度学习已成为图像处理和计算机视觉领域的主导工具。近年来, 深度学习在图像处理、自然语言处理和语音识别等领域取得了一系列突破性进展, 已经成为学术界的研究热点, 特别是卷积神经网络(Convolutional Neural Network, CNN)[6]、深度置信网络(Deep Belief Network, DBN) [7]、层叠自动编码机(Stacked Auto-Encoder, SAE)[8]、长短时记忆网络(Long-Short Term Memory)[9]、生成对抗网络(Generative Adversarial Network, GAN)[10]等深度模型在各领域内的大量突破性成果的涌现[11-12]。

近年来传统隐写术与隐写分析的发展已经进入瓶颈期, 而深度学习的兴起又为该领域注入了新的活力。众多学者根据隐写与隐写分析的特点, 对深度学习网络结构进行相应的改进, 将深度学习技术与隐写术和隐写分析相结合, 取得了不少创造性的成果。本文将对近几年来深度学习在隐写术与隐写分析领域中的研究发展状况进行梳理与归纳总结, 讨论其技术特点、存在问题以及未来研究趋势。由于数字图像是目前最为流行的隐藏载体, 因此本文也主要讨论面向数字图像的基于深度学习的隐写术与隐写分析方法。

本文组织结构如下: 第2节比较生成对抗网络与隐写术的特点, 并讨论代表性的基于深度学习的隐写方法; 第3节先对比卷积神经网络与隐写分析过程的异同, 然后介绍基于卷积神经网络的隐写分析方法; 第 4 节总结全文并展望未来研究方向。

2 基于深度学习的隐写术

2.1 隐写术与生成对抗网络和对抗样本的比较

根据图像嵌入域的不同, 图像隐写术主要包括空域隐写术与 JPEG 隐写术两类。而根据嵌入策略的不同, 图像隐写术又可大致分为非自适应隐写术和自适应隐写术两类。非自适应隐写术的思想是对载体图像修改的越少隐写安全性就越高。非自适应隐写术通常与纠错编码(隐写码)相结合来实现具体的嵌入过程, 常见的隐写码有矩阵编码[13]、湿纸码(Wet Paper Codes, WPC)[14]、BCH 码(Bose Chaudhuri Hocquenghem)等[15]。非自适应隐写术对所有载体元素无差别的选择修改, 而自适应隐写术则考虑载体图像的自身属性, 根据图像纹理复杂区域难于建模的特点, 有选择地将秘密信息嵌入到载体纹理复杂或者边缘丰富的区域, 提高了载密图像的抗隐写分析检测能力。自适应隐写术首先设计修改失真代价(每个载体元素因修改所引起的损失), 然后结合 STC 码(Syndrome Trellis Codes)[16]来完成具体的嵌入过程。常见的自适应隐写术主要有、HUGO[17], WOW (Wavelet Obtained Weights)[18]、UNIWARD (UNIversal WAvelet Relative Distortion)[19]、HILL (High-pass, Low-pass, and Low-pass) [20] 、CPP (controversial pixels prior)[21]、EBS (Entropy Block Steganography)[22]、UED (Uniform Embedding Distortion)[23] 、UERD (Uniform Embedding revisited Distortion)[24]等。

自适应隐写术是当前学界研究的主流, 而深度学习与隐写术的结合也促进了隐写术新的发展。基于深度学习的隐写术可以大致分为两类, 一类采用生成对抗网络, 另一类则借鉴对抗样本的思想。本节将从这两个方面对比传统隐写术与基于深度学习的隐写术的异同。

隐写术的安全性通常由载密图像的抗隐写分析检测能力来衡量。载体图像与载密图像在视觉和统

计特性方面要尽可能的相近, 使隐写分析检测模型难以区分两类图像。生成对抗网络是 Goodfellow 等[10]在 2014 年提出的一种生成模型, 其主要思想是训练生成器和判别器网络, 以数字图像为例, 生成器尝试生成图像的近似分布并且尽可能的将其合成真实图像, 而判别器则努力将真实图像与虚假图像区分开。隐写术和隐写分析之间的对抗特点与生成对抗网络有一定的共通之处, 这为二者的结合提供了研究思路。

隐写术/隐写分析与生成对抗网络结构的对比如图 1 的前两个框图所示。隐写术与生成器都要“生成”新的图像, 而隐写分析与判别器都要判断“生成”的图像与原图像是否可区分。所不同的是隐写术有信息提取的过程, 且其与隐写分析之间的对抗是相对独立、无交互的; 而生成对抗网络中的生成器与判别器之间的对抗则通过多轮迭代反馈来达到相互优化。

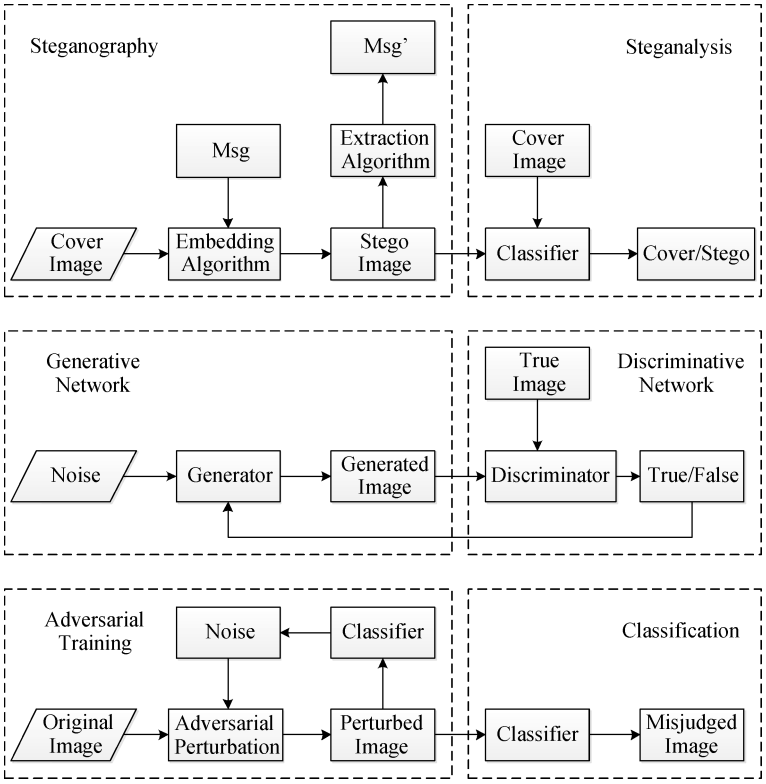


图 1　隐写术/隐写分析与生成对抗网络、对抗样本

Figure 1　steganography/steganalysis vs. generative adversarial network and adversarial examples

对抗样本是针对机器学习模型的缺陷, 通过对样本添加感官上不易察觉的修改, 使分类器接受并做出错误的分类判断, 这与隐写术同样有很大的相似性。

隐写术/隐写分析与对抗样本的对比如图 1 的第一个框图和第三个框图所示。隐写术与对抗训练都要对图像进行修改(嵌入信息或添加噪声)。隐写术修改图像的目的除了传递信息还要保证修改操作不被分类器(隐写分析)检测到, 而对抗样本则要保证修改后的图像被错误分类。

2.2　基于深度学习的隐写方法

当前基于生成对抗网络的隐写术主要集中于图像空域。Volkhonskiy 等[25]最早提出基于生成对抗网络的隐写模型 SGAN(Steganographic GAN)。SGAN 采用深度卷积生成对抗网络(Deep Convolutional Generative Adversarial Network, DCGAN)[26] 构造而成, 由以下三部分组成:

1) 生成器(G): 目标是生成接近自然的图像;

2) 判别器(D): 目标是区分真实图像和生成图像;

3) 判别器(S): 目标是区分载体图像和载密图像。

SGAN 的模型结构如图 2 所示。生成器 G 与判别器 D 之间的对抗学习是得到用于隐写的载体图像, 并且使载体图像更接近于真实的自然图像。然后利用传统隐写方法对生成的载体图像进行嵌入, 得到载密图像。生成器 G 与判别器 S(隐写分析)之间的对抗学习是为了使载体图像与载密图像更相近, 用以提高隐写安全性。

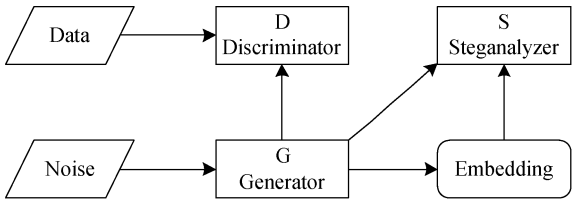


图 2　SGAN 模型结构图

Figure 2　SGAN model structural diagram

在 SGAN 的基础上, Shi 等[27]提出另一种基于生成对抗网络的隐写模型 SSGAN(Secure Steganography Based on GAN)。SSGAN 的模型结构与 SGAN 相类似, 采用 WGAN(Wasserstein GAN)[28]代替 DCGAN, 具有更快的训练速度和更高的生成图像质量。

SGAN 与 SSGAN 利用生成对抗网络生成载体图像, 而 Hayes 等[29]提出的 HayesGAN 模型则利用对抗学习直接生成载密图像。HayesGAN 由以下三部分组成:

1) 生成器(G): 目标是生成载密图像;

2) 判别器(D): 目标是能够提取嵌入的秘密信息;

3) 判别器(S): 目标是区分载体图像和载密图像。

HayesGAN 的模型结构如图 3 所示。载体图像和秘密信息作为生成器 G 的输入, 生成载密图像。判别器 D 提取嵌入的秘密信息, 评价提取的正确性。判别器 S(隐写分析)则评价生成的载密图像的隐写安全性。HayesGAN 的信息提取过程通过训练网络(判别器 D)来实现, 由于误差的存在, 不能保证完全正确提取嵌入的秘密信息。

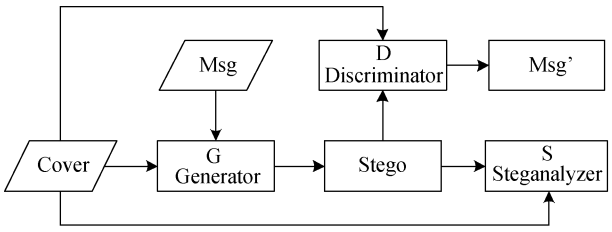


图 3　HayesGAN 模型结构图

Figure 3　HayesGAN model structural diagram

Zhu 等[30]借鉴 HayesGAN 的思想, 提出另一种生成对抗网络 HiDDeN(Hiding Data With Deep Networks)。该模型与 HayesGAN 具有相似的网络结构, 其特点是利用对抗样本(Adversarial examples)对图像变化鲁棒的特性, 使得生成的载密样本在多种载体攻击下(高斯模糊、像素丢失、裁剪和 JPEG 压缩)仍能以较高的准确率提取出嵌入的信息。其他与 HayesGAN 相近的方法还有[31-32]等。

Tang 等[33]把生成对抗网络与自适应隐写方法相结合, 提出 ASDL-GAN(automatic steganographic distortion learning framework with GAN)来学习隐写失真代价。ASDL-GAN 由以下两部分组成:

1) 生成器(G): 目标是生成图像的修改概率图;

2) 判别器(D): 目标是区分载体图像和载密图像。

生成器 G 以载体图像作为输入, 模拟嵌入过程, 输出图像的修改概率图。载体图像根据修改概率图得到载密图像, 判别器 D(隐写分析)把这两类图像作为输入, 评价载密图像的安全性。经过多轮迭代求得合适的修改概率图, 把修改概率转换为载体图像的隐写失真代价, 利用 STC 码进行嵌入。ASDL-GAN 利用生成对抗网络生成的是失真代价, 不是载体/载密图像, 实际隐写过程仍然应用传统的自适应隐写框架。

在嵌入率为 0.1bpp 和 0.4bpp 的条件下, 采用 SRM 和 Xu-Net 对 ASDL-GAN 模型生成的载密图像进行检测, 其错误率如下表 1 所示。从表 1 实验结果可知, 经过多轮对抗学习, ASDL-GAN 安全性不断增强, 但安全性尚未超越 S-UNIWARD。

表 1　采用 SRM 和 Xu-Net 对生成图像的检测错误率

Table 1　detection error rates of SRM and Xu-Net for generated images

隐写算法	0.1bpp		0.4bpp	
	SRM	Xu-Net	SRM	Xu-Net
20000 次迭代	25.86%	26.92%	13.50%	9.01%
60000 次迭代	27.70%	32.56%	15.44%	14.23%
100000 次迭代	29.35%	36.49%	16.39%	14.52%
140000 次迭代	32.64%	39.72%	16.81%	15.72%
180000 次迭代	33.02%	40.04%	17.40%	16.20%
S-UNIWARD	40.02%	42.53%	20.22%	20.01%

Yang 等[34]ASDL-GAN 对进行了改进, 提出使用 Tanh 模拟器(Tanh-simulator)作为激活函数来代替 ASDL-GAN 中的 TES(ternary embedding simulator), 用以解决 TES 难以反向传播的问题; 同时在判别器 D 的设计中考虑选择通道, 使得学习到的失真代价能够抵御基于选择通道的隐写分析检测。Yang 等[34]的改进模型明显提高了隐写安全性, 在 BOSSbase 图像库上取得了比 ASDL-GAN 和 S-UNIWARD 更低的检测错误率。

以上基于对抗生成网络的隐写术主要是从载体或隐写失真代价方面“被动”的增强自身对隐写分析的抵御能力; 一些学者则借鉴对抗样本[35-36]的思想, 通过“主动”干扰隐写分析采用的机器学习模型来提高隐写安全性。Zhang 等[37]针对基于深度学习的

隐写分析方法(详见第 3 章), 提出利用深度学习模型训练过程中的梯度, 给载体图像加入特定的噪声来得到增强的载体图像, 使之能够“误导”深度学习的分类(载密图像识别为载体图像); 然后在增强的载体图像上采用传统的自适应隐写框架实现信息嵌入。该方法能够有效对抗基于深度学习的隐写分析方法; 但是对于传统的基于高维特征的隐写分析方法, 其安全性能会有所下降。

与 Zhang 等[37]利用对抗样本构造载体图像不同, Tang 等[38]则利用对抗样本来调整隐写失真代价。对于三进制嵌入(ternary embedding), 该方法的目标是选择合适的修改方向, 使得载密图像尽可能被识别为载体图像。其具体做法是先计算反向传播过程中图像每个位置的梯度(预设类别标签为载体图像), 如果修改操作的方向与梯度反方向相同则减小其失真代价, 反之则增加其失真代价。隐写失真代价调整完后, 同样采用自适应隐写框架完成信息嵌入。该方法根据梯度方向调整隐写失真代价, 在嵌入过程中没有引入额外的修改噪声, 它不仅能抵御基于深度学习的隐写分析, 还能在传统的高维特征隐写分析的检测下保持较高的安全性。

基于与 Tang 等[38]同样的思想, Ma 等[39]也根据梯度方向来选择图像元素的修改方向。不同的是, Ma 等[39]先保持原有失真代价不变, 然后利用二进制嵌入(binary embedding)得到修改的位置, 再根据梯度方向来确定最终的修改方向(+1 或–1), 这在一定程度上与使用边信息(side information)[19,22,24]的隐写嵌入过程相类似。该方法同样可以对抗基于深度学习的隐写分析和传统的高维特征隐写分析。

2.3 基于深度学习的隐写术总结

图像隐写术的安全性表现在载体图像与载密图像的不可区分性。为此, 基于生成对抗网络的隐写术借鉴的对抗学习思想, 使之应用于隐写过程中的不同环节, 以达到隐写更加“自然”的目的。根据对抗学习在隐写过程中的不同作用, 基于生成对抗网络的隐写方法主要可以分为如下三种:

1) 对抗学习生成载体图像(cover): 模型包含一个生成器和两个判别器, 通过构造载体图像来提高隐写安全性, 实际隐写过程仍然采用传统的隐写术, 秘密信息能够正确提取。

2) 对抗学习生成载密图像(stego): 模型包含一个生成器和两个判别器, 模型直接生成载密图像, 在载密图像生成过程中实现信息嵌入, 但不能保证正确提取嵌入的信息。

3) 对抗学习生成隐写失真代价: 模型包含一个生成器和一个判别器, 模型生成隐写失真代价, 并使用传统的自适应隐写框架进行信息的嵌入和提取。

对于基于对抗样本的隐写术而言, 提高载体图像与载密图像的不可区分性在于迫使隐写分析检测模型产生误判。因此, 寻找机器学习模型的弱点(例如梯度下降方法的脆弱性), 使之与隐写修改操作相结合, 是此类隐写术的研究目标。根据对抗样本在隐写过程中的不同应用, 基于对抗样本的隐写方法主要可以分为如下两种:

1) 对抗样本生成载体图像(cover): 根据反向传播过程中的梯度对载体图像加入噪声实现增强, 使增强后的载体图像在隐写嵌入前后都被隐写分析检测模型识别为载体图像。

2) 对抗样本调整隐写失真代价: 根据反向传播过程中的梯度来调整隐写失真代价, 进而选择合适的修改方向, 使得修改难以被深度学习模型所识别。

这两种基于对抗样本的隐写方法也都采用传统的自适应隐写框架来实现信息的嵌入和提取。

3 基于深度学习的隐写分析方法

3.1 隐写分析与卷积神经网络的比较

通用隐写分析过程一般分为两个阶段: 特征构造与分类器训练。由于隐写嵌入操作修改的是图像的高频信号, 在特征构造阶段, 往往先使用高通滤波器计算残差图像, 再使用各种统计模型来提取隐写分析特征。常见的空域图像隐写分析方法有 SPAM[40]、SRM[41]、PSRM[42]、TLBP[43]等。针对自适应隐写方法的特点, 通过估计待测图像的修改概率图, 为残差特征分配不同的权重, 这被称为自适应隐写分析, 代表性的方法有 tSRM[44]、maxSRM[45]、σ SRM[46]等。早期的 JPEG 图像隐写分析遵循在嵌入域构造特征的原则, 通常基于 DCT 系数来计算残差和提取特征, 例如 PEV[47]、JRM[48]等。后来 JPEG 图像的隐写分析则根据解压缩的信号放大和分块相位的特点, 在 JPEG 的解压缩空域结合相位信息来构造残差特征, 典型方法有 DCTR[49]、PHARM[50]、GFR[51]及其自适应特征构造方法 SCA[52]。

良好的特征表示对隐写分析的检测准确性起到至关重要的作用, 因此目前通用隐写分析方法的研究主要集中在特征的设计和提取上。然而传统的隐写分析特征一般由研究者依据人工经验、启发式设计完成, 需要一定的领域知识。而且隐写分析的特征构造与分类器训练是独立进行, 这种两段式过程难以使特征与分类器同步优化。

针对隐写分析面临的上述问题, 部分学者开始将深度学习与隐写分析相结合, 利用深度学习模型具有的模拟复杂表示的能力, 来达到自动学习有效特征表达的目的。同时利用深度学习的端到端学习过程, 把特征构造和分类器训练在一个结构中同步完成。

在众多的深度学习模型中, 卷积神经网络是最具代表性的一个, 也是隐写分析使用最多的网络模型。卷积神经网络的模型结构与传统隐写分析的过程有很大的相似之处, 其对比如图 4 所示。隐写分析的第一个步骤是残差获取, 由多个高通滤波器组成, 用以获得残差图像(反映图像的噪声信号)。与之相对应, 卷积神经网络的第一层是卷积层, 由多个卷积核组成, 其作用类似于低通滤波器, 用以获得特征图(反映图像的关键内容表示)。隐写分析的第二和第三个步骤分别是量化和截断, 目的是缩小残差范围和保留主要残差信息。而卷积神经网络的卷积层之后分别是激活函数和池化层, 其作用分别是非线性变换和特征图的压缩降维。其中, 池化与量化的作用类似, 激活函数与截断操作一样更多关注零值附近的信息。对于隐写分析, 经过残差的处理之后, 需要通过一定的统计方法对残差进行汇总规约得到特征; 而卷积神经网络则在特征表示的最后通过若干个全连接层获得合并的特征图。最后, 隐写分析与卷积神经网络的末尾都连接一个分类器, 输出分类结果。

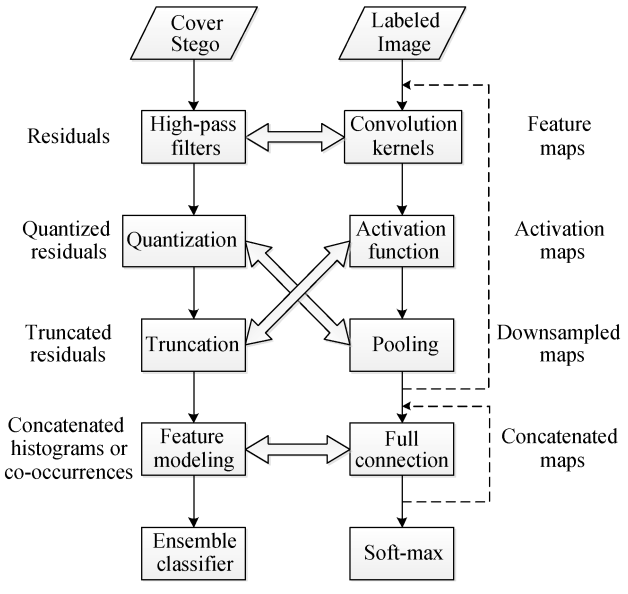


图 4 隐写分析与卷积神经网络

Figure 4 steganalysis vs. convolutional neural network

3.2 基于深度学习的隐写分析模型

Tan 等[53]在 2014 年首次提出了基于深度学习的隐写分析方法, 简称为 TanNet。TanNet 具有 4 层卷积神经网络结构, 包括三个卷积层和一个全连接层。TanNet 采取如下三种训练方式:

1) 模型一: 随机初始化第一层卷积核。

2) 模型二: 使用 KV 核初始化第一层卷积核(KV 核乘以随机初始化的卷积核)。

3) 模型三: 使用 KV 核初始化第一层卷积核以及使用栈式卷积自动编码器预训练每个卷积层。

这三种模型与 SPAM 和 SRM 对嵌入率为 0.4 bpp(bit per pixel)的 HUGO[20]隐写方法进行检测, 检测错误率如表 2 所示。

表 2 五种隐写分析模型对 HUGO 的检测错误率

Table 2 detection error rates of five steganalytic models for HUGO

模型一	模型二	模型三	SPAM	SRM
48%	43%	31%	42%	14%

由实验结果可知, 随机初始化卷积核, 模型基本无检测能力, 但KV核初始化卷积核有助于提升检测效果, 而且卷积核参数的预训练可以进一步提升检测效果。其中模型三的检测能力已经超过 SPAM, 但尚不及 SRM。TanNet 的初步尝试显示了深度学习在隐写分析领域中的潜力。

2015 年, Qian 等[54]提出了 QianNet。该网络包括一个预处理层, 五个卷积层和三个全连接层。其中预处理层采用固定的 KV 核, 用以获取残差图像进行后续学习, 减少图像本身内容的干扰。根据隐写噪声的特点, 提出高斯激活函数并应用于卷积层, 而在全连接层使用 ReLU 激活函数。由于最大池化容易丢失高频残差信号, QianNet 使用均匀池化来减少信息损失。

与SRM的检测能力相比, 在BOSSbase图像库上, QianNet 比 SRM 的检测正确率低 3% ~ 5%; 而对于 ImageNet 图像库, QianNet 与 SRM 的检测水平相当。

在 TanNet 与 QianNet 提出之后, 研究者们更多关注如何把深度学习与隐写分析的特点相结合。Xu 等[55]在 2016 年提出了基于卷积神经网络的 XuNet, 该网络同样在前端添加一个固定的高通滤波层(KV 核)。由于隐写的高频噪声信号关于 0 对称且符号无关, XuNet 在第一个卷积层采用零值偏置参数, 同时添加 ABS(绝对值)层来缩小特征图的范围。增加 BN 层来提高收敛速度, 避免陷入局部最小值。此外, XuNet 采用混合激活函数, 在网络前端使用 tanH 提高特征的学习能力, 在网络深层使用ReLU减少反向

传播的难度。最后, XuNet 还在网络后几层使用 1×1 卷积核和全局平均池化, 用以防止网络模型过拟合和信息损失。XuNet 对卷积神经网络的改造使之获得了较大的性能提升, 在 BOSSbase 图像库上对 S-UNIWARD[19]和 HILL[20]的检测, XuNet 的检测准确率基本超过 SRM。

2017 年, YNYNet[56]的提出标志着深度学习在隐写分析领域取得重大突破。YNYNet 是一个 10 层的卷积神经网络, 主要包括以下三个特点: 1) 网络第一层采用 30 个 SRM 卷积核来初始化参数。这些卷积核能够迭代更新以学习更多有效的特征表示, 且更多的卷积核能够提升特征学习的多样性。2) 提出新的激活函数 TLU。TLU 能够更好的适应隐写噪声的分布, 收敛速度快, 学习到的卷积核的区分性更好。3) 首次在深度学习中引入选择通道。借鉴传统自适应隐写分析的思想, 把选择通道与卷积神经网络相结合, 能够提升对自适应隐写嵌入的检测效果。此外, YNYNet 还验证了大规模实验样本对于深层网络训练的重要性。在对 WOW[18]、S-UNIWARD[19]和 HILL[20]的测试上, YNYNet 及其自适应版本的检测能力都显著超越 SRM[41]和 maxSRMd2[45]。

其他基于深度学习的隐写分析研究则包括深度学习模型的迁移学习、信号增强, 以及实际应用等方面。Qian 等[57]认为卷积神经网络难以学习图像的全局统计信息, 提出通过迁移学习方法, 利用辅助领域(传统隐写分析特征)中的知识来增强对模型全局统计信息的学习能力。之后, Qian 等[58]为解决低嵌入率样本难以训练的问题, 提出另一种迁移学习方法, 从高嵌入率样本的卷积神经网络模型迁移到低嵌入率样本的模型, 减少学习成本和节约训练时间, 提高了模型的检测能力。

隐写噪声是弱信号, 为了提高隐写分析模型的识别能力, 这些残差信息一方面需要增强, 另一方面则需要避免在训练过程中丢失。Wu 等[59]提出利用深度残差网络来构造隐写分析模型, 通过残差学习来保持和增强隐写的弱信号, 并用残差结构来避免深层网络易出现的梯度弥散现象。Xu [60]等则提出一种重叠池化(Overlapped pooling)来解决传统池化过程中丢失信息过多的问题, 同时利用卷积神经网络的集成学习方法来提高其检测能力。

当前基于深度学习的隐写分析方法主要应用于固定尺寸的小图像(例如 256×256), 隐写信号的特点又决定了图像不适合采取缩放操作。针对这个问题, Pibre 等[61]提出在卷积神经网络中采用大尺寸卷积核(与输入的残差图像尺寸相同)。Tsang 等[62]则提出在卷积神经网络的最后一个卷积层与全连接层之间增加一个统计矩提取层。卷积层输出的特征图维度虽然随图像大小而变, 而统计矩提取层则可接受任意维度的输入并输出固定维度的特征图送入全连接层。

以上分析的基于深度学习的隐写分析模型都是针对空域的隐写方法, Chen 等[63]与 Xu 等[64]在 2017 年把深度学习引入到 JPEG 图像的隐写分析。Chen 等[63]借鉴传统 JPEG 隐写分析的思想, 基于 JPEG 图像的不同相位来训练不同的卷积神经网络模型, 并使用两种卷积核——KV 核和催化核(catalyst kernel)——分别分析空域的高频残差信号和 JPEG 隐写嵌入带来的噪声信号。Xu 等[64]提出一个 20 层的深层卷积神经网络来检测 J-UNIWARD 隐写方法[19], 并利用残差结构来解决梯度弥散问题。

2018 年, Zeng 等[65]提出了一种混合的深度学习隐写分析模型 ZengNet 来检测 JPEG 图像隐写方法。ZengNet 基于卷积神经网络, 主要有以下三个特点: 1) 人工设计残差特征。JPEG 图像解压到空域, 空域统计特性的改变来自块效应和隐写噪声。人工设计残差特征可以把这二者区分开来, 使模型只对隐写噪声进行学习。2)在深度学习结构中引入量化和截断操作。借鉴传统隐写分析的思想, 进一步细化手工设计残差特征, 对不同的残差特征用不同的子模型进行学习。3)混合模型: 采用多个子模型来丰富特征学习的多样性。ZengNet 能够应用于大规模图像样本集, 在对 J-UNIWARD[19]、UED[23]和 UERD[24]的检测方面, 其检测准确性明显高于传统隐写分析方法。

3.3 深度学习隐写分析总结

隐写分析属于模式识别的范畴, 深度学习在模式识别领域中的成功应用也带动了隐写分析技术的发展。与图像分类等模式识别任务相比, 图像隐写分析有两点显著的不同: 1)在检测对象方面, 隐写分析更关注图像高频信号而非图像本身内容; 2)在检测方法方面, 隐写分析更多使用全局统计方法而非局部统计方法。正是因为上述两点差异, 深度学习在与隐写分析的结合过程中也呈现出其独特之处, 主要表现在如下四个方面:

1) 计算有效的残差信号

采用固定的高通滤波器(KV 核或 SRM 卷积核), 有的作为独立的预处理层, 核参数不更新[54-55,57-60,64-65]; 有的用来初始化第一层卷积层, 核参数可学习更新[53,56,62-63]。

2) 减少残差信号的损失

对于隐写分析而言, 所有残差信号同等重要,

均匀池化可以综合所有残差信息, 而最大池化会造成一定的信息损失, 因此模型通常采用均匀池化代替最大池化[54-58,60,62], 或是采用重叠池化[54,57-58,60]。由于均匀池化会混淆池化范围内的统计信息, 有些模型甚至禁用池化层[61]或在网络前端禁用池化层[56,62-64]。

3) 提高残差信号的多样性

采用多个高通滤波核[56,62]来学习更多的特征表达, 或者使用混合网络模型[63,65]。

4) 获取更多的图像全局统计信息

采用大尺寸卷积核[61]、增大池化范围[65]、全局池化[55,60,63-64], 迁移传统隐写分析特征知识[57]。

总之, 基于深度学习的隐写分析的发展过程, 就是深度学习模型与传统隐写分析方法相互借鉴与融合的过程。

4 结语与展望

深度学习给隐写术与隐写分析领域带来颠覆性的变化。本文简要回顾了隐写术与隐写分析的研究背景与发展历程, 分别从隐写术与隐写分析两个方面介绍深度学习方法与传统方法的区别、代表性深度学习模型的特点及发展现状。

基于深度学习的隐写术与隐写分析研究方兴未艾, 但也面临着许多待解决的问题, 未来研究可以从以下方面开展。

在隐写术方面:

1) 嵌入与提取的一致性。与自适应隐写方法相比, 利用生成对抗网络直接生成载密图像能够把所有环节在一个过程完成。然而由于训练中不可避免地存在误差, 导致信息不能完全提取正确。因此, 保证嵌入与提取的一致性对于实际应用十分必要。

2) 嵌入效率。对于基于生成对抗网络的隐写术而言, 复杂的网络结构往往需要更多的时长来完成嵌入, 对硬件资源也有较高的需求。从应用上考虑, 为了在低配置终端上实现高效的信息隐写, 需要在网络结构方面研究嵌入效率问题。

在隐写分析方面:

1) 端到端学习。深度学习在其他领域已经实现完全的端到端学习, 而由于隐写分析任务的特殊性, 目前面向隐写分析的深度学习模型仍然需要一定的人工干预。因此, 实现端到端学习的深度学习模型对于隐写分析有重要意义。

2) 特征学习。传统隐写分析主要依靠于特征构造, 深度学习虽然可以自动学习有效的特征, 但仍是对已知隐写方法的修改模式的学习。由于隐写分析相对于隐写术始终具有滞后性, 如何学习反映隐写嵌入带来异常的更本质特征, 是值得研究的课题。

3) 小样本量训练。深度学习需要大量的训练集才能得到良好的检测效果。然而大样本量训练耗时耗力, 且有时大量样本不易获得。利用小样本量训练有效的基于深度学习的隐写分析模型是十分迫切的现实需求。

4) 载体失配。一种观点是扩大训练样本的来源和数量可以缓解载体失配问题, 但这又与小样本量的训练需求相矛盾。另外在大数据时代, 数据多源异构, 静态的训练集始终难以覆盖动态的检测对象, 深度学习应该从本质上解决载体失配问题。

5) 主动隐写分析。当前深度学习主要应用于被动隐写分析, 即判断隐写操作的有无, 而在提取攻击等方面尚未有应用。隐写内容的提取本质上属于密码破译, 而深度学习在密码破译方面已有不少研究成果, 这对于隐写分析提取攻击具有一定的借鉴意义。

总体而言, 深度学习为解决隐写和隐写分析问题提供了新的理念及技术。本文通过对深度学习在图像隐写术与隐写分析领域中的研究进展的回顾及分析讨论, 希望能够给该领域的研究人员带来新的思路及研究启发, 共同促进深度学习技术在隐写术与隐写分析领域的进一步发展与繁荣。

参考文献

[1] Li B, He J, Huang J, et al. A survey on image steganography and steganalysis[J]. *Journal of Information Hiding and Multimedia Signal Processing*, 2011, 2(2): 142-172.

[2] Cheddad A, Condell J, Curran K, et al. Digital image steganography: Survey and analysis of current methods[J]. *Signal processing*, 2010, 90(3): 727-752.

[3] Nissar A, Mir A H. Classification of steganalysis techniques: A study[J]. *Digital Signal Processing*, 2010, 20(6): 1758-1770.

[4] Schmidhuber J. Deep learning in neural networks: An overview[J]. *Neural networks*, 2015, 61: 85-117.

[5] Salakhutdinov R, Mnih A, Hinton G. Restricted Boltzmann machines for collaborative filtering[C]//*Proceedings of the 24th international conference on Machine learning. ACM*, 2007: 791-798.

[6] Sahiner B, Chan H P, Petrick N, et al. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images[J]. *IEEE transactions on Medical Imaging*, 1996, 15(5): 598-610.

[7] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. *science*, 2006, 313(5786): 504-507.

[8] Poultney C, Chopra S, Cun Y L. Efficient learning of sparse representations with an energy-based model[C]//*Advances in neural information processing systems*. 2007: 1137-1144.

[9] Hochreiter S, Schmidhuber J. Long short-term memory[J]. *Neural*

computation, 1997, 9(8): 1735-1780.

[10] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. *arXiv preprint arXiv*:1511.06434, 2015.

[11] Le Q V, Ngiam J, Coates A, et al. On optimization methods for deep learning[C]//*Proceedings of the 28th International Conference on International Conference on Machine Learning. Omnipress*, 2011: 265-272.

[12] Oquab M, Bottou L, Laptev I, et al. Learning and transferring mid-level image representations using convolutional neural networks [C]//*Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014: 1717-1724.

[13] Westfeld A. F5—a steganographic algorithm[C]//*International workshop on information hiding. Springer, Berlin, Heidelberg*, 2001: 289-302.

[14] Fridrich J, Goljan M, Lisonek P, et al. Writing on wet paper[J]. *IEEE Transactions on signal processing*, 2005, 53(10): 3923-3935.

[15] Sachnev V, Kim H J, Zhang R. Less detectable JPEG steganography method based on heuristic optimization and BCH syndrome coding[C]//*Proceedings of the 11th ACM workshop on Multimedia and security. ACM*, 2009: 131-140.

[16] 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.

[17] Pevny T, Filler T, Bas P. Using high-dimensional image models to perform highly undetectable steganography[C]//*International Workshop on Information Hiding. Springer, Berlin, Heidelberg*, 2010: 161-177.

[18] Holub V, Fridrich J. Designing steganographic distortion using directional filters[C]//*Information Forensics and Security* (*WIFS*), *2012 IEEE International Workshop on. IEEE*, 2012: 234-239.

[19] Holub V, Fridrich J. Digital image steganography using universal distortion[C]//*Proceedings of the first ACM workshop on Information hiding and multimedia security. ACM*, 2013: 59-68.

[20] Li B, Tan S, Wang M, et al. Investigation on cost assignment in spatial image steganography[J]. *IEEE Transactions on Information Forensics and Security*, 2014, 9(8): 1264-1277.

[21] Zhou W, Zhang W, Yu N. A new rule for cost reassignment in adaptive steganography[J]. *IEEE Transactions on Information Forensics and Security*, 2017, 12(11): 2654-2667.

[22] Wang C, Ni J. An efficient JPEG steganographic scheme based on the block entropy of DCT coefficients[C]//*Acoustics, Speech and Signal Processing* (*ICASSP*), *2012 IEEE International Conference on. IEEE*, 2012: 1785-1788.

[23] Guo L, Ni J, Shi Y Q. Uniform embedding for efficient JPEG steganography[J]. *IEEE transactions on Information Forensics and Security*, 2014, 9(5): 814-825.

[24] Guo L, Ni J, Su W, 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.

[25] Volkhonskiy D, Nazarov I, Borisenko B, et al. Steganographic generative adversarial networks[J]. *arXiv preprint arXiv*:1703.05502, 2017.

[26] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv*:1511.06434, 2015.

[27] Shi H, Dong J, Wang W, et al. Ssgan: Secure steganography based on generative adversarial networks[C]//*Pacific Rim Conference on Multimedia. Springer*, Cham, 2017: 534-544.

[28] Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *ArXiv e-prints*, January 2017

[29] Hayes J, Danezis G. Generating steganographic images via adversarial training[C]//*Advances in Neural Information Processing Systems*. 2017: 1954-1963.

[30] Zhu J, Kaplan R, Johnson J, et al. HiDDeN: Hiding Data with Deep Networks[C]//*Proceedings of the European Conference on Computer Vision (ECCV)*. 2018: 657-672.

[31] Hu D, Wang L, Jiang W, et al. A Novel Image Steganography Method via Deep Convolutional Generative Adversarial Networks[J]. *IEEE Access*, 2018, 6: 38303-38314.

[32] Shi H C, Zhang X Y. Synchronization Detection and Recovery of Steganographic Messages with Adversarial Learning[J]. *arXiv preprint arXiv*:1801.10365, 2018

[33] Tang W, Tan S, Li B, et al. Automatic steganographic distortion learning using a generative adversarial network[J]. *IEEE Signal Processing Letters*, 2017, 24(10): 1547-1551.

[34] Yang J, Liu K, Kang X, et al. Spatial Image Steganography Based on Generative Adversarial Network[J]. *arXiv preprint arXiv:*1804.07939, 2018.

[35] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. *arXiv preprint arXiv*:1312.6199, 2013.

[36] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. *arXiv preprint arXiv*:1412.6572, 2014.

[37] Zhang Y, Zhang W, Chen K, et al. Adversarial Examples Against Deep Neural Network based Steganalysis[C]//*Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security. ACM*, 2018: 67-72.

[38] Tang W, Li B, Tan S, et al. CNN Based Adversarial Embedding with Minimum Alteration for Image Steganography[J]. *arXiv preprint arXiv*:1803.09043, 2018.

[39] Ma S, Guan Q, Zhao X, et al. Weakening the Detecting Capability of CNN-based Steganalysis[J]. *arXiv preprint arXiv*:1803.10889, 2018.

[40] 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.

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

[42] 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.

[43] Li B, Li Z, Zhou S, et al. New steganalytic features for spatial image steganography based on derivative filters and threshold LBP operator[J]. *IEEE Transactions on Information Forensics and Security*, 2018, 13(5): 1242-1257.

[44] Tang W, Li H, Luo W, et al. Adaptive steganalysis against WOW embedding algorithm[C]//*Proceedings of the 2nd ACM workshop on Information hiding and multimedia security. ACM*, 2014: 91-96.

[45] Denemark T, Sedighi V, Holub V, et al. Selection-channel-aware rich

model for steganalysis of digital images[C]//*Information Forensics and Security* (*WIFS*), *2014 IEEE International Workshop on. IEEE*, 2014: 48-53.

[46] Denemark T, Fridrich J, Comesaña-Alfaro P. Improving selection-channel-aware steganalysis features[J]. *Electronic Imaging*, 2016, 2016(8): 1-8.

[47] Pevny T, Fridrich J. Merging Markov and DCT features for multi-class JPEG steganalysis[C]//*Security, Steganography, and Watermarking of Multimedia Contents IX. International Society for Optics and Photonics*, 2007, 6505: 650503.

[48] Kodovsky J, Fridrich J. Steganalysis of JPEG images using rich models[C]//*Media Watermarking, Security, and Forensics 2012. International Society for Optics and Photonics*, 2012, 8303: 83030A.

[49] Holub V, Fridrich J. Low-complexity features for JPEG steganalysis using undecimated DCT[J]. *IEEE Transactions on Information Forensics and Security*, 2015, 10(2): 219-228.

[50] Holub V, Fridrich J. Phase-aware projection model for steganalysis of JPEG images[C]//*Media Watermarking, Security, and Forensics 2015. International Society for Optics and Photonics*, 2015, 9409: 94090.

[51] Song X, Liu F, Yang C, et al. Steganalysis of adaptive JPEG steganography using 2D Gabor filters[C]//*Proceedings of the 3rd ACM workshop on information hiding and multimedia security. ACM*, 2015: 15-23.

[52] 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.

[53] Tan S, Li B. Stacked convolutional auto-encoders for steganalysis of digital images[C]//*Signal and Information Processing Association Annual Summit and Conference* (*APSIPA*), 2014 Asia-Pacific. IEEE, 2014: 1-4.

[54] Qian Y, Dong J, Wang W, et al. Deep learning for steganalysis via convolutional neural networks[C]//*Media Watermarking, Security, and Forensics 2015. International Society for Optics and Photonics*, 2015, 9409: 94090J.

[55] Xu G, Wu H Z, Shi Y Q. Structural design of convolutional neural networks for steganalysis[J]. *IEEE Signal Processing Letters*, 2016, 23(5): 708-712.

[56] 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.

[57] Qian Y, Dong J, Wang W, et al. Learning representations for steganalysis from regularized cnn model with auxiliary tasks[C]//*Proceedings of the 2015 International Conference on Communications, Signal Processing, and Systems. Springer*, Berlin, Heidelberg, 2016: 629-637.

[58] Qian Y, Dong J, Wang W, et al. Learning and transferring representations for image steganalysis using convolutional neural network[C]//Image Processing (ICIP), *2016 IEEE International Conference on. IEEE*, 2016: 2752-2756.

[59] Wu S, Zhong S, Liu Y. Deep residual learning for image steganalysis[J]. Multimedia tools and applications, 2018, 77(9): 10437-10453.

[60] Xu G, Wu H Z, Shi Y Q. Ensemble of CNNs for steganalysis: An empirical study[C]//*proceedings of the 4th aCM workshop on information Hiding and Multimedia security. ACM*, 2016: 103-107.

[61] Couchot J F, Couturier R, Guyeux C, et al. Steganalysis via a convolutional neural network using large convolution filters for embedding process with same stego key[J]. *arXiv preprint arXiv*:1605.07946, 2016.

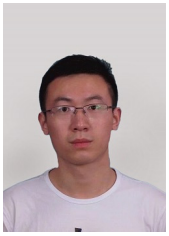
[62] Tsang C F, Fridrich J. Steganalyzing Images of Arbitrary Size with CNNs[J]. *Electronic Imaging,* 2018, 2018(7): 1-8.

[63] Chen M, Sedighi V, Boroumand M, et al. JPEG-phase-aware convolutional neural network for steganalysis of JPEG images[C]//*Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. ACM*, 2017: 75-84.

[64] Xu G. Deep convolutional neural network to detect J-UNIWARD [C]//*Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. ACM*, 2017: 67-73.

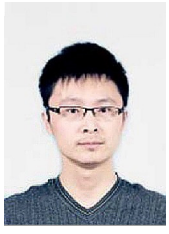
[65] Zeng J, Tan S, Li B, et al. Large-scale jpeg image steganalysis using hybrid deep-learning framework[J]. *IEEE Transactions on Information Forensics and Security*, 2018, 13(5): 1200-1214.

翟黎明 于2011年在山西大学计算机科学与技术专业获得学士学位。现在武汉大学网络空间安全专业攻读博士学位。研究领域为多媒体内容安全。研究兴趣包括: 隐写术、信隐写分析。Email: limingzhai@whu.edu.cn



嘉炬 于2017年在华中师范大学通信与信息系统专业获得硕士学位。现在武汉大学网络空间安全专业攻读博士学位。研究领域为多媒体内容安全。研究兴趣包括: 隐写术、信隐写分析。Email: jiaju123@whu.edu.cn

任魏翔 于2016年在贵州大学计算机软件与理论专业获得硕士学位。现在武汉大学网络空间安全专业攻读博士学位。研究领域为多媒体内容安全。研究兴趣包括: 隐写术、信隐写分析。Email: renweixiang@whu. edu.cn



徐一波 于2012年在合肥工业大学计算机软件与理论专业获得硕士学位。现在武汉大学网络空间安全专业攻读博士学位。研究领域为多媒体内容安全。研究兴趣包括: 隐写术、信隐写分析。Email: xu_yi_bo@163.com

王丽娜 于 2001 年在东北大学获得博士学位。现任武汉大学国家网络安全学院教授。研究领域为多媒体安全、云计算安全、网络安全。研究兴趣包括: 隐写术、信隐写分析、虚拟化、数字信号处理与识别等。Email: lnwang@whu.edu.cn