软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(2):551−578 [doi: 10.13328/j.cnki.jos.006135]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于深度学习的图像隐写分析综述*

陈君夫[1], 付章杰[1,2], 张卫明[3], 程 旭[1], 孙星明[1]

[1](南京信息工程大学 计算机与软件学院,江苏 南京 210044)

[2](鹏城实验室,广东 深圳 518055)

[3](中国科学技术大学 信息科学技术学院,安徽 合肥 230026)

通讯作者: 付章杰, E-mail: fzj@nuist.edu.cn

摘 要: 隐写术及隐写分析是信息安全领域研究热点之一.隐写术的滥用造成许多安全隐患,如非法分子利用隐写进行隐蔽通信完成恐怖袭击.传统隐写分析方法的设计需要大量先验知识,而基于深度学习的隐写分析方法利用网络强大的表征学习能力自主提取图像异常特征,大大减少了人为参与,取得了较好的研究效果.为了促进基于深度学习的隐写分析方法研究,对目前隐写分析领域的主要方法和突破性工作进行了分析与总结.首先,比较了传统隐写分析方法与基于深度学习的隐写分析方法的差异;然后根据训练方式的不同,将基于深度学习的隐写分析模型分为两类——半学习隐写分析模型与全学习隐写分析模型,详细介绍了基于深度学习的各类隐写分析网络结构与检测效果;其次,分析和总结了对抗样本对深度学习安全带来的挑战,并阐述了基于隐写分析的对抗样本检测方法;最后,总结了现有基于深度学习的隐写分析模型存在的优缺点,并探讨了基于深度学习的隐写分析模型的发展趋势.

关键词: 隐写术;隐写分析;卷积神经网络;深度学习;对抗样本

中图法分类号: TP391

中文引用格式: 陈君夫,付章杰,张卫明,程旭,孙星明.基于深度学习的图像隐写分析综述.软件学报,2021,32(2):551−578. http://www.jos.org.cn/1000-9825/6135.htm

英文引用格式: Chen JF, Fu ZJ, Zhang WM, Cheng X, Sun XM. Review of image steganalysis based on deep learning. Ruan Jian Xue Bao/Journal of Software, 2021,32(2):551−578 (in Chinese). http://www.jos.org.cn/1000-9825/6135.htm

Review of Image Steganalysis Based on Deep Learning

CHEN Jun-Fu[1], FU Zhang-Jie[1,2], ZHANG Wei-Ming[3], CHENG Xu[1], SUN Xing-Ming[1]

[1](School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China)

[2](Peng Cheng Laboratory, Shenzhen 518055, China)

[3](School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China)

Abstract: Steganography and steganalysis are one of the research hotspots in the field of information security. The abuse of steganography has caused many potential safety hazards. For example, illegal elements use steganography for covert communications to carry out terrorist attacks. The design of traditional steganalysis methods requires a large amount of prior knowledge, and the steganalysis methods based on deep learning use the powerful representation learning ability of the network to autonomously extract abnormal image features, which greatly reduces human participation and achieves good results. To promote the research of steganalysis technology based on deep learning, this study analyzes and summarizes the main methods and work in the field of steganalysis. Firstly, this study analyzes and compares the differences between traditional steganalysis and deep learning-based steganalysis. Furthermore, according to the different training methods, the steganalysis models based on deep learning are divided into two categories: semi-learning steganalysis model and full-learningsteganalysis model. The network structure and detection effect of various types of steganalysis based on deep

* 基金项目: 国家重点研发计划(2018YFB1003205); 国家自然科学基金(U1836110, U1836208, 61802058, 61911530397)

Foundation item: National Key Research and Development Program of China (2018YFB1003205); National Natural Science Foundation of China (U1836110, U1836208, 61802058, 61911530397)

收稿时间: 2020-05-30; 修改时间: 2020-07-10; 采用时间: 2020-08-27; jos 在线出版时间: 2020-09-10

learning are introduced in detail. In addition, the challenges that the adversarial samples pose to deep learning security are analyzed and summarized, the detection method of adversarial samples is expounded based on steganalysis. Finally, this study summarizes the pros and cons of existing steganalysis models based on deep learning and discusses its development trends.

Key words: steganography; steganalysis; convolution neural networks; deep learning; adversarial examples

多媒体技术的普及与应用,一方面给社会带来了不少便利,另一方面也带来了许多风险,如信息泄露、恶意篡改、隐私窃取等.人们越来越注重多媒体传播过程中的信息安全和隐私保护问题.现有的通信安全保障主要分为加密和信息隐藏:加密主要对秘密信息本身进行操作,但经过特殊处理后的明文更加容易受到第三方的怀疑;而信息隐藏则隐藏秘密数据的存在性,使秘密数据在不引起第三方的怀疑下进行隐蔽通信[1].因此,信息隐藏这种具有伪装特性的通信安全保障受到了越来越多的关注[2].在囚徒模型中,可以很好地阐述隐写术中各方的角色:Alice 和 Bob 是监狱中不同牢房的犯人,他们之间的通信需要在狱警 Eve 的监视下完成;同时,Eve 能够看见他们的通信内容.为了降低狱警 Eve 防范心的同时完成通信,隐写术孕育而生.Alice 将想要传达的秘密信息进行隐写操作隐藏在载体当中,Bob 则需要将含密载体中的秘密信息进行提取,狱警 Eve 时刻监视 Alice 和 Bob 的通信,一旦发现任何可疑信息就断绝双方通信[3].隐写术是一门关于信息隐藏的科学,所谓信息隐藏指的是不让除预期的接受者之外的任何人知晓信息的传递事件.隐写术的英文叫做 Steganography[2],来源于特里特米乌斯的讲述密码学与隐写术的著作《Stegano-graphia》,该书名起源于希腊语,意为“隐秘书写”[4],如图 1 所示为图像隐写的一般过程.

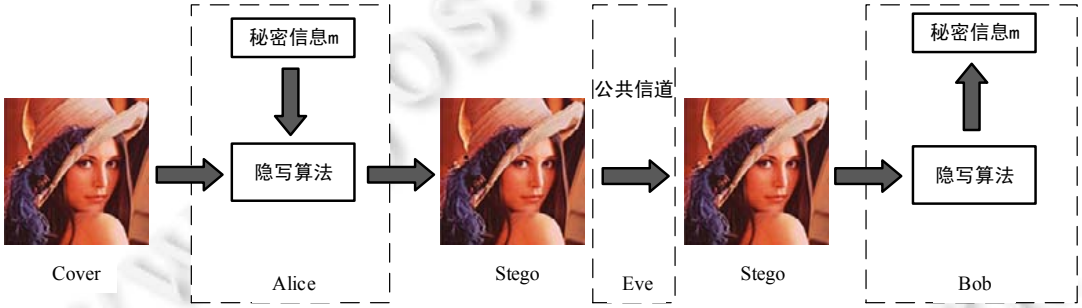

Fig.1 General process of steganography

图 1 隐写的一般过程

随着信息隐藏技术的不断推广,隐写术逐渐成为一把双刃剑,在其为人们的通信安全提供保障的同时,不法分子利用其获取个人利益或应用于恐怖袭击.2001 年,美国的主流媒体 CNN 就刊登过一则利用隐写术进行隐秘通信从而犯罪的新闻.在 2007 年哥伦比亚毒枭以及 2011 年全能神邪教等案件中都出现了隐写术的影子.由此可见,非法和恶意使用隐写术已经造成了非常巨大的社会危害,所以隐写分析研究油然而生.这对于打击恐怖分子的恐怖行动、维护社会安定和保障国家信息安全具有十分重要的意义.但是隐写分析本身非常依赖人工设计的滤波核,并且对于图像本身的纹理属性与细节属性没有一个较好的统筹概念,根据不同的图库可能会有不同的滤波核的设计.如何减少甚至避免人为设计成为了一个难题.

深度学习自 2006 年 Hinton 提出的受限玻尔兹曼机(restricted Boltzmann machine,简称 RBM)[5]之后,就成为了机器学习中不可或缺的新兴技术,通过模拟人脑神经元,可以自动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性.现如今,深度学习已经成为图像处理和计算机视觉(computer vision,简称 CV)领域中的主要工具.其中比较热门的研究网络热点——卷积神经网络(convolution neural network,简称 CNN)[6]、深度置信网络(deep belief network,简称 DBN)[7]、层叠自动编码机(stacked auto-encoder,简称 SAE)[8]、循环神经网络(recurrent neural network,简称 RNN)[9]等各种网络在深度学习的各个领域中不断涌现.虽然 U-Net[10],ResNet[11],DenseNet[12]同属于卷积神经网络,但不同的网络结构会产生截然不同的效果和应用[13].在 2014 年,Goodfellow 提出的生成对抗网络(generative adversarial networks,简称 GAN)[14]通过构建判别器与生成器的对抗博弈环境,

最终达到两者的纳什平衡[15],不仅为深度学习开启了新的篇章,也给隐写术与深度学习网络的结合提供了机遇.GAN网络由于其纷繁复杂的变形网络[16−18]和独有的创造力,被众多国内外学者应用于隐写术.传统隐写术和隐写分析的发展也因为深度学习的出现与发展到达了新的高度.

本文在深入隐写分析模型的基础上,首先将现有的隐写分析模型按照其针对隐写操作类型分为专用型隐写分析模型与通用型隐写分析模型,如图 2 中隐写分析模型分类所示.由于专用型隐写分析模型仅针对特定的隐写算法且对于不匹配的或者未知的隐写算法检测效果较差,随着各式各类的自适应隐写算法的不断涌现,专用型隐写分析模型显得力不从心,也逐渐退出历史舞台,通用型隐写分析模型也逐渐成为主流隐写分析模型.接着,本文将通用型隐写分析模型按照其采用的技术基础分为传统隐写分析模型与基于深度学习的隐写分析模型:传统隐写分析模型需要一定的先验知识和根据数据而设定的滤波核;基于深度学习的隐写分析方法利用网络强大的表征学习能力自主提取图像异常特征,大大减少了人为参与,且取得了显著的检测效果.将现有的基于深度学习的隐写分析模型按照不同的预处理层训练方式分为以下两个大类:(1) 半学习隐写分析模型;(2) 全学习隐写分析模型,并在此基础上根据不同的网络架构模式将上述两个模型再细分为基于深度网架构与基于宽度网络架构两个分支.然后讨论对抗样本的出现为信息隐藏提供的新思路与方法,将对抗样本与隐写方法结合的对抗隐写方法分为在隐写前对载体进行操作与在隐写过程中内容进行操作,分析了基于隐写分析的对抗样本的检测方法.最后,本文总结目前基于深度学习隐写分析模型存在的问题并展望未来的发展方向.

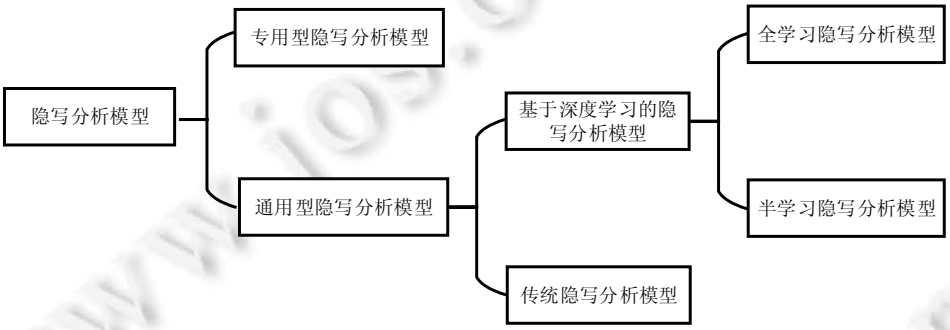

Fig.2 Steganalsis model classification

图 2 隐写分析模型分类

本文第 1 节介绍信息隐藏领域的相关知识,并将基于深度学习的隐写分析按照预处理层的差异分为半学习隐写分析与全学习隐写分析.第 2 节、第 3 节对基于深度的半学习隐写分析与全学习隐写分析进行详细介绍并对比.第 4 节对于基于深度学习的隐写分析进行对比与总结.第 5 节介绍对抗样本和基于隐写分析的对抗样本检测.第 6 节对于基于深度学习的隐写分析网络进行总结与未来展望.

1 相关知识

1.1 隐写术

隐写术是在尽可能不破坏图像本身各种性质的情况下,在多媒体载体中嵌入秘密信息的技术.隐写术最重要的特点是不可检测性,其目的是使通信双方能够进行隐蔽通信,而不被其他用户察觉通信痕迹.图像隐写是隐写术中的一个重要分支,由于数字图像具有信息冗余度大的特性,因此在其中隐藏秘密信息时难以被肉眼察觉,是一个理想的秘密信息载体.LSB(least significant bit)[19]作为早期的隐写方法,是一种基于图片最低有效位修改并储存信息的隐写方法.利用人眼对于色彩差异的不敏感性,将秘密信息通过一定的嵌入方法放入图片的最低有效位,从而将我们所需要隐藏的信息通过一定方法放入图片的最低有效位上.除此以外,LSB 还有一种变化形式 LSB 匹配(LSB matching,简称 LSBM)[20].二者之间的差距在于:LSB 对于最低有效位进行的是替换操作;LSBM 采用的则是随机±1 原则,采用三元伴随式矩阵编码(syndrome-trellis codes,简称 STC)[21]嵌入秘密信息.应

用 LSB 算法的图像格式需为位图形式,即图像不能经过压缩,所以 LSB 算法多应用于 png,bmp 等空域图像中.图 3 是 LSB 类隐写流程图,可以看到,载体图像 Lena(戴帽子的女人)在隐写前后并不存在明显的差距.

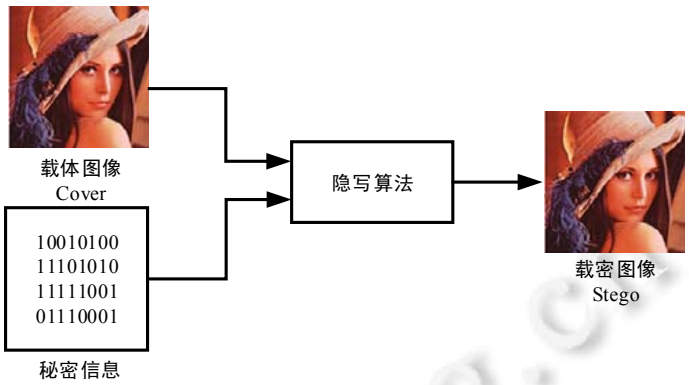

Fig.3 LSB steganography process

图 3 LSB 隐写过程

无论是 LSB 还是 LSBM,都是比较简单的隐写方法,都是一种非自适应的隐写算法.非自适应隐写术的思想是:对载体图像中像素内容修改地越少,隐写算法抗隐写分析能力就越强.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[22]、湿纸码(wet paper code,简称 WPC)[23]、BCH 码(Bose Chaudhuri Hocquenghem)[24]等.非自适应隐写术对载体图像整体进行修改而不考虑单独像素间的关联性;自适应隐写术则考虑载体图像的自身属性,例如图片内容的纹理信息、边缘信息,根据图像纹理复杂区域难于检测的特点,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[19].常见的自适应隐写算法有 HUGO[25]、WOW[26]、UNIWARD[27]、HILL[28]等,各类自适应隐写算法都与 STC[21]编码方法结合,差异在于失真函数的不同.图 4 是自适应隐写术的操作流程.

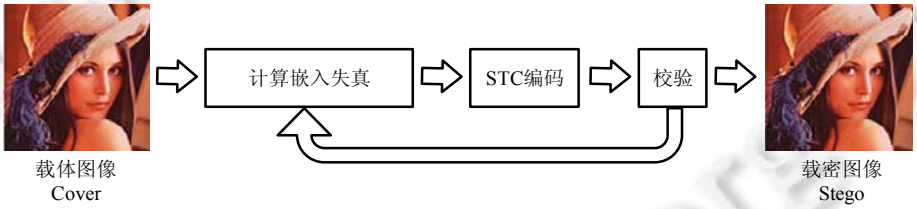

Fig.4 Adaptive steganography process

图 4 自适应隐写过程

不仅在空域上存在隐写算法,在频域即 JPEG 域上的隐写方法也很多,其中较早且具有代表性的是 Jsteg[29]、F5[22]、J-UNIWARD[30]、UED[31]、UERD[32].根据图像经过离散余弦变换之后得到 63 个 AC 系数和 1 个 DC 系数,由于 DCT 分块后得到的结构信息存储在 DC 系数中,所以对于中频的 AC 系数的修改并不会引起结构上可见的变化,这样修改 AC 系数既可以达到隐写的目的,也不会对图像结构造成太大的破坏,保证了隐写的隐蔽性.

随着生成对抗网络的出现,国内外研究学者们将信息隐藏和生成对抗网络相结合,利用生成对抗网络对于复杂图像的学习能力,学习不同像素之间的结构关系,从而生成一些更适合隐写或者更符合隐写特点的载体图像.2016 年,Volkhonskiy[33]提出了 SGAN 的隐写模型,结合已有的 DCGAN 网络生成更加符合隐写规则的载体图像.2018 年,ASDL-GAN[34]和 UT-6HPF-GAN[35]相继被提出,将对抗网络应用在修改概率图的生成上.不仅如此,由于生成对抗网络需要一个‘对手’共同进步,通常将基于深度学习的隐写分析模型作为生成对抗网络中的对立方,这样两个模型可以在对抗学习中共同进步.这种新型的隐写方法不仅减少了人为参与,还可以有效提升隐写安全性.但是仍然存在一些问题,例如网络架构不稳定、GAN 网络的不可逆性导致隐写内容无法准确提取等.

1.2 隐写分析

隐写方法的多样性与安全性,推动了隐写分析的发展.隐写分析是检测隐写术的一种手段,根据隐写分析的发展趋势,我们可以将其分为 3 个不同的阶段.

- 第 1 个阶段:判断载密图像(stego)中是否隐藏秘密信息,即判断数字图像是载体图像(cover)还是载密图像.这是现在大多数隐写分析模型最重要的步骤,也被称为盲隐写分析.
- 第 2 个阶段:判断载密图像中秘密信息的容量和秘密信息隐藏的位置等(多为纹理复杂处或者图像边缘处).
- 第 3 个阶段:从载密图像中提取秘密信息,这个阶段需要具体了解隐写方法、隐写位置、隐写容量等各种信息[36].

3 个阶段呈现出一种递进的关系,只有前一层做了充足的准备,才可以在最后提出载密图像中的秘密信息.

针对于早期的非自适应与纠错码结合的隐写算法,隐写分析器可以通过简单的统计分析和直方图分布来有效检测图片.针对 LSB 和 LSBM 这两种空域隐写算法,修改最低有效位会在一定程度上破坏相邻像素之间的关联性.根据这一特性,存在相应的专用型隐写分析模型[37,38].专用型隐写分析是指隐写分析一方在已知隐写具体算法的情况下所设计的特用的隐写分析模型,数字图像在嵌入秘密信息后,载体图像的某种统计特性特征会发生相应的改变.通用隐写分析在未知载体图像和隐写算法的基础上,检测图像是否含有秘密信息.相对于通用型隐写分析,专用型隐写分析的准确率更高但具局限性.

2000 年,Westfeld 等人[39]最早提出了针对 LSB 隐写的统计检验法,之后,研究者们相继提出了 RS 分析法、DIH 分析法、WS 分析法,提高了嵌入率的估计精度.2005 年,Andrew 设定了特征直方公式(HCF),这是第一个专用灰度图 LSB 的隐写分析[37].2008 年,Liu 等人[40]采用图像最低两位平面的相关性作为特征检测 LSBM 隐写,该方法考虑到了 LSBM 隐写对图像低位平面造成的影响.Tan[41]提出了一种基于 B 样条函数的专用分析方法.除此之外,研究人员还提出了针对 BPCS、PVD 等隐写方式的专用隐写分析方法.Bohme[42]将对于 LSB 专用隐写分析的方法迁移到频域图像上,提出了一种针对于 Jsteg 的专用隐写分析算法.2014 年,Xia[43]等人通过分析相邻像素之间的关联性,设计出针对于 LSBM 隐写算法的专用隐写分析器.随着自适应隐写算法的出现,各类隐写算法的抗隐写分析能力逐渐增强,这对隐写分析的要求也越来越高.2011 年,Gul 等人和 Luo 等人分别提出了针对于 HUGO 的专用隐写分析模型[44,45].2014 年,Tang 等人[46]提出了针对于 WOW 这种自适应隐写的隐写分析策略,并且这种策略可以根据不同的隐写算法应用于空域和频域.

随着隐写算法的逐渐增强以及各式各样隐写算法的不断涌现,通用型隐写分析模型逐渐壮大,特征的维数从低维开始慢慢发展到上万维.在空域上,从 686 维的 SPAM[47]发展到 34 671 维的 SRM[48];而在频域上,也从 8 000 维的 DCTR[49]发展到 12 600 维的 PHARM[50].富模型 Rich Model[51]中的分类器采用机器学习领域中比较常用的分类器[52],例如支持向量机[53]、集成分类器[54]、FLD[55]等.传统的隐写分析步骤包括特征提取、特征增强、二分类决策模型训练这 3 个部分.传统的隐写分析模型有 SPAM[47]、SRM[48]、DCTR[49]、PHARM[50]和 GFR[56],这些都基于人工计算的特征提取方式.空域隐写分析通过分析数字图像的统计特性,来检测图像是否嵌入秘密信息;而频域隐写分析由于不同的 DCT 与量化矩阵,则需要分析 DCT 系数关系而进行判别.SRM 通过建立不同的子模型,首先对训练样本中的图片空域特征信息进行提取并计算残差;然后对得到的残差信息进行截断与量化,计算相应的共生矩阵;最后再利用机器学习的方式训练分类器.但隐写分析研究并不仅分析图像中是否隐写内容,并且分析出可能的隐写方法、隐写修改的区域,最后,通过推测隐写方法和隐写位置截取秘密信息.

自适应隐写算法根据图片最小失真函数,结合 STC[21]使用进行隐写.这使得隐藏的秘密信息越来越难以发现,所以图像中秘密信息的有效特征越来越难以获取,原有的隐写分析特征一般是由专业的研究人员依赖自己的先验经验和不断启发式尝试计算得出.隐写分析的特征提取和机器学习二分类训练是分开的,前者通过手工设计,后者通过机器学习方法完成,两步操作无法同时进行优化,很难达到一个异构平衡状态.

在传统隐写分析的发展过程中,正是因为上述问题的存在,再加上深度学习蓬勃发展,所以国内外的学者将隐写分析和深度学习结合起来.这样既可以不用专业研究人员手工设计特征提取方式,又可以利用深度学习端

到端的学习过程,使得特征提取和判别器可以同时训练.依赖深度学习可以模拟人脑学习复杂的结构信息,从而提取出数字图像中的特征信息.

1.3 数据集与评价指标

隐写术和隐写分析所采用的数据集多为 BOWS2(https://photogallery.sc.egov.usda.gov/)和 BOSSbase-v1.01 (ttp://agents.fel.cvut.cz/stegodata/),两款数据集都是 512×512 的一万张灰度图,数据集中包含生活、景点、建筑等多种类型图片.BOSSbase1.01 是 Fridrich 团队 2011 年所创建的用于隐写分析竞赛的专用数据集,采用 7 种不同类型的数码相机拍摄得到的图像用于隐写和隐写分析,可以防止单个数码相机拍摄出现相机指纹,使判别器学习出现偏差.Pevny 和 Fridrich 为了举办 HUGO 隐写分析竞赛,专门构建了 BOSSbase0.92(http://agents.fel.cvut.cz/boss/index.php?mode=VIEW\&tmpl=materials/)图像库,包含 10 000 张未经任何压缩处理的 512×512 像素的图片.表 1 是 BOSSbase 内不同数码相机的拍摄图片序号.

Table 1 Source of BOSSbase image datasets

表 1 BOSSbase 图像数据拍摄来源

图片序号	相机型号
1–1354	CanonEOS 400D
1355–1415	CanonEOS 40D
1416–2769	CanonEOS 7D
2770–4811	CanonEOSDIGITALREBELXsi
4812–6209	PENTAXK20D
6210–7242	NIKON D70
7243–10212	M9 digital camera

BOWS2 数据集始创于 2008 年用于水印竞赛,由于其特征分布于内容与 BOSS 数据集相似,自 2017 年后,被信息隐藏领域广泛地使用,当作 BOSS 数据集的补充.UCID(uncompressed colour image database,http://vision.doc.ntu.ac.uk/)是一种彩色图片数据集,数据内的图片已经标好了预设的正确选框.由于图片的处理过程中没有采取任何压缩方式,图像中的各种信息都得以有效的保存.UCID 是一种通用型基准数据集,并且还可以应用在测试图像压缩能力和色彩质量等方面.除此之外,还有一种 NRCS(NRCS photo gallery,https://photogallery.sc.egov.usda.gov/)的图像数据集.表 2 是这几类数据集各项信息的对比展示,其中的 SIPI(USC-SIPI image database, http://sipi.usc.edu/database/)中有一张著名的图像:Lena,即图 3 中的示例图.

Table 2 Comparison of different datasets

表 2 各类数据库对比

数据集	数据量(张)	位深度	图像类型	图片大小	格式
BOSSbase	10 000	8	灰度图	512×512	PGM
BOWS2	10 000	8	灰度图	512×512	PGM
UCID	1 338	24	彩图	512×384/384×512	TIF
NRCS	N/A	24	彩图	1500×2100	TIF/JPEG
SIPI	N/A	8/24	灰/彩图	256×256/512×512/1024×1024	TIFF

不同的数据集之间存在一定的相似性,较为常用的数据集是 BOSSbase[57]和 BOWS2[58],这两类数据集不仅属性相似,图片的内容也存在一定的相似性,所以在隐写分析模型需要对数据集进行增强操作时,通常混用两个数据集进行网络训练.图 5 是 BOSSbase 几张示例图.

通常,比较隐写分析网络的检测效果,采用误检率 Err 或准确率 Acc 作为模型效果的衡量标准.隐写分析的目标是从数字图像中检测载密图像,因此将载密图像作为阳性类,载体图像作为阴性类.假设载体图像和载密图像的数量分别为 C 和 S,其中被正确分类的载体图像与载密图像的样本数为 N 和 P,在评价隐写分析模型时,通常会用到如下几种指标:

$$Acc=\frac{P+N}{C+S} \tag{1}$$

$$Err = 1 - \frac{P+N}{C+S} \tag{2}$$

其中,$P+N$ 为被隐写分析判别正确的样本总数,$C+S$ 为所有参与测试的样本总数,并且满足 Err 与 Acc 之和为 1.

P_{FA} 代表虚警率(false alarm ratio),即代表载体图像被误判成载密图像的比率.

$$P_{FA} = \frac{C-N}{C} \tag{3}$$

P_{MD} 代表漏检率(missed detection ratio),即代表载密图像被误判成载体图像的比率.

$$P_{MD} = \frac{S-P}{S} \tag{4}$$

P_E 代表最小平均错误率(minimum average decision error ratio),即在虚警率发生变化时,两类错误平均值的最小值.

$$P_E = \frac{1}{2}\min_{P_{FA}}(P_{FA} + P_{MD}(P_{FA})) \tag{5}$$

$MD5$ 代表当 P_{FA} 为 5%情况下的误检率.

$$MD5 = P_{MD}(P_{FA}=0.05) \tag{6}$$

$FA50$ 代表当 P_{MD} 为 50%情况下的虚警率.

$$FA50 = P_{FA}(P_{MD}=0.5) \tag{7}$$

公式(6)、公式(7)为在 ALASKA 隐写分析挑战赛[59]中的评判标准.

Fig.5 Part of pictures in BOSSbase datasets

图 5 BOSSbase 数据集中部分图片

2 半学习隐写分析

在众多纷繁的深度学习模型中,卷积神经网络由于其特有的网络属性,可以精确地对数字图像进行操作,是最具有代表性的一种深度学习网络.通过卷积计算的方式,可以获取图像中细致的图像信息,与传统特征提取异曲同工.不仅如此,由于不同层的网络参数是可以训练的,深度学习还可以通过大量的数据学习一种捕捉细致特征的手段.不同的网络层具有不同的效果,例如传统的量化和截断,在作用上可以用正则化层和激活函数来替代,激活函数 Sigmoid[60]、TanH[61]、ReLU[62]等都是在深度学习隐写分析中所常用的.在训练二分类模型上,传统隐写分析与基于深度学习的隐写分析的差距并不大,依赖的都是机器学习的方式,都是训练一个分类器,最后输出二分类结果.

在本节中,半学习是指在隐写分析网络利用固定滤波核作为独立的一个预处理层,并且内部的权重参数不参与反向传播,其他的网络层则是依赖深度学习方法去优化.在本节中,按照网络的架构分为深度网络模型与宽度网络模型.

2.1 基于深度网络的半学习隐写分析模型

2015 年,Qian 等人[63]提出了一种新的网络,称为 GNCNN(Gaussian-Neuron CNN),图 6 是 GNCNN 网络与传统隐写分析之间的对比图.

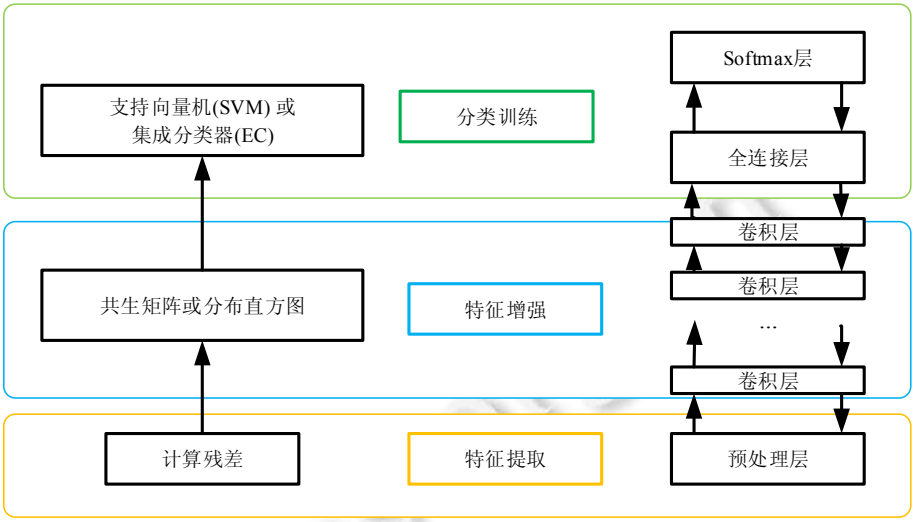

Fig.6 Traditional steganalysis and GNCNN structure

图 6 GNCNN 与传统隐写分析结构

该网络结构包括一个预处理层、5 个卷积层和 3 个全连接层,预处理层将卷积层中的卷积核替换成固定的高通滤波核,获取数字图像的高维残差信息,辅助网络进行学习.这样不仅仅加快了隐写分析网络的训练,而且将不必要的图像内容信息移除,减少了图像低维信息干扰.在实验过程中,加入了固定的高通滤波核的 GNCNN 网络在训练速度和训练结果上都会优于使用在预处理层中随机初始化卷积核的网络.由于经过高通滤波器后得到的信息多为高频残差信息,最大池化容易丢失高频残差图像信息,导致网络难以拟合,所以在 GNCNN 中,使用平均池化操作来减少残差信息的丢失.Qian 根据隐写噪声的特点提出了高斯激活函数,替代卷积层中的 ReLU 激活函数.如下是 GNCNN 中所采用高斯激活函数.

$$F(x)=1-\mathrm{e}^{-\frac{x^2}{\sigma^2}} \tag{8}$$

其中,σ是用来衡量函数曲线宽度的参数.该公式可以将数值较小的输入转换成一个正数,并且这种激活函数也是第一次在深度网络中应用,故该网络也因此命名为 GNCNN 或是 QianNet 隐写分析模型.

对不同嵌入率下的空域自适应隐写算法,GNCNN 的表现见表 3.

Table 3 Comparison of experimental results under different steganography algorithms of traditional steganalysis and GNCNN

表 3 GNCNN 与传统隐写分析在不同隐写算法下实验结果对比

BPP	HUGO			WOWO			S-UNIWARD		
	GNCNN	SRM	SPAM	GNCNN	SRM	SPAM	GNCNN	SRM	SPAM
BOSSbase									
0.3	33.8%	29.6%	42.9%	34.3%	31.2%	42.2%	35.9%	34.3%	40.0%
0.4	28.9%	25.2%	39.1%	29.3%	25.7%	38.2%	30.9%	29.3%	35.1%
0.5	25.7%	21.4%	35.7%	24.8%	22.1%	34.9%	26.3%	24.8%	30.6%
ImageNet database									
0.4	33.6%	32.5%	–	34.1%	34.7%	–	34.7%	34.4%	–

从表 3 的实验结果中可以看出,GNCNN 的检测效果较优于 SPAM 较弱于 SRM.在各类的隐写算法上都满

足这样一个条件:随着嵌入率(bit per pixel,简称 BPP)的提升,即隐写容量的增加、载密图像中嵌入的秘密信息增加,隐写分析的准确率就会越高.BOSSbase 是由 10 000 张经过裁减的灰度图所组成的专用数据集;表 3 中最下一行的 ImageNet[64]数据集则是由互联网中大量彩图组成,在彩图隐写分析上,GNCNN 已经与 SRM 的检测效果非常接近.在 BOSSbase 数据集上,通过大量数据测试发现:GNCNN 仅仅比 SRM 的检测正确率低 3%~5%;而对于彩图这种通道数较多的数据集而言,GNCNN 与 SRM 的隐写检测水平相近.这是因为相对于灰度图的隐写,彩图不同通道间具有关联性且包含的信息更多,因此彩图隐写也更容易被检测,对于网络自学习的参数权重要求较低.相对于其他基于深度学习的隐写分析而言,GNCNN 由于网络模型较为简单,在隐写分析的准确率上存在局限性.

2016 年,Xu 等人提出了 Xu-Net[65]网络.Xu-Net 在网络框架上仍然沿用了 GNCNN 的网络架构特点,例如依旧采用全局池化操作,减少残差图像信息的丢失.同样在网络前端添加了一个固定的高通滤波层,即 KV 核作为预处理层,如下所示.

$$K_{kv}=\frac{1}{12}\begin{pmatrix}-1 & 2 & -2 & 2 & -1\\ 2 & -6 & 8 & -6 & 2\\ -2 & 8 & -12 & 8 & -2\\ 2 & -6 & 8 & -6 & 2\\ -1 & 2 & -2 & 2 & -1\end{pmatrix} \tag{9}$$

公式(9)表示的滤波核是从 SRM[48]的 30 个高通滤波核中挑选出来的,在区分高维特征即纹理复杂度时具有较好的效果.高通滤波器是一种中心对称的结构,这样可以有效地提取出像素点与周围像素之间的信息差距,使得隐写分析模型可以有效地获取像素之间的共生矩阵,重新排列得到信噪特征,从而帮助隐写分析模型更好地检测,各类不同的滤波核在处理相同的数字图像时会有不同的效果.

从表 4 的测试结果中可以看出:Xu-Net 与 SRM 在相同的隐写方法下具有相似的检测效果,甚至超过传统的 SRM 方法,并且远远超过了同样是基于深度学习的 GNCNN[63]的检测效果.Xu-Net 的提出与实验结果,正式宣告基于深度学习的隐写分析模型已经可以与传统隐写分析模型进行较量.

Table 4 Comparison of detection accuracy of Xu-Net and SRM on S-UNIWARD and HILL

表 4 Xu-Net 与 SRM 在 S-UNIWARD 与 HILL 下准确率对比

Algorithm	BPP (%)			
	0.1bit/pixel		0.4bit/pixel	
	CNN	SRM	CNN	SRM
S-UNIWARD	57.33	59.25	80.24	79.53
HILL	58.44	56.44	79.24	75.24

Xu-Net 网络根据经过预处理层的残差高频噪声信号具有关于 0 对称且与符号无关的特性,在第 1 个卷积层采用添加 ABS(absolute layer)层来收敛特征图的范围,从原来无意义的正负区间缩小到正向区间.添加 BN 层(batch normalization layer)进行批处理,使得训练数据符合正态分布.这样可以提升训练时的收敛速度,也可以避免训练时出现梯度弥散或梯度爆炸现象,导致训练结果陷入局部最小值.最后采用 1×1 的卷积核将特征信息集聚,并且防止模型存在过拟合的情况.

2017 年,Xu[66]在原有 Xu-Net 基础上提出了一种基于 JPEG 域的隐写分析网络,并命名为 Xu-Net-JPEG,采用 20 层的全卷积网络证明了深度学习网络可以在复杂领域击败基于特征的隐写分析方法,同时也证明了深度网络比宽度网络更容易提取隐写噪声.这种网络结构依赖固定的 DCT 内核和特征图组的阈值设定,为了防止过深的卷积层会使网络在训练时出现梯度弥散或者梯度爆炸的情况,在网络中采用与 ResNet 相同的跳接结构,这在后续的 SRNet[67]中也有相应的考虑.同年,Chen 等人[68]也在 Xu-Net 的基础上提出一种带有 JPEG 相位感知的频域隐写分析网络 VNet 与 PNet(VNet 结构较小且精度相对于 PNet 相差较小).VNet 不仅沿用了 Xu-Net 中的预处理层,还在其基础上额外添加了 3 个滤波核作为固定的预处理层,分别为点高通滤波核(point high-pass filter)、二维水平 Gabor 滤波核和二维垂直 Gabor 滤波核,用以学习一些具有方向特性的隐写噪声,其中,点高通滤波核

在预处理层中起到“催化剂”的作用.PNet与VNet借鉴了DCTR等频域隐写分析的先验知识,在网络框架中还添加了JPEG相位感知模块,用以学习频域的信噪比信息,从而提升隐写检测精度.

这是因为JPEG编码会将8×8的像素分块作为基础操作单元,各个不同的JPEG块内与块间的系数都具有较强的关联性,这种特性被称为相位特性.相位感知对JPEG图像对应位置的点进行统计与合并,这样可以较好地描述块间相关性与相位特性的变化,从而提升频域隐写分析模型的准确性.在文献[50–51,56]中,都可以见到相同的操作.

从图7的左半部分可以看出:每一个特征图在经过该模块都会被下采样为64张代表不同DCT系数的统计特征图(一张图代表一个相位),这也是VNet或是PNet适用于频域隐写分析的重要原因.VNet经过相位分离模块后会将得到的特征图继续放入一个线性网络中进行训练(PNet则会放入64个并行子网络中进行训练),最后输入全连接层并输出判别结果.

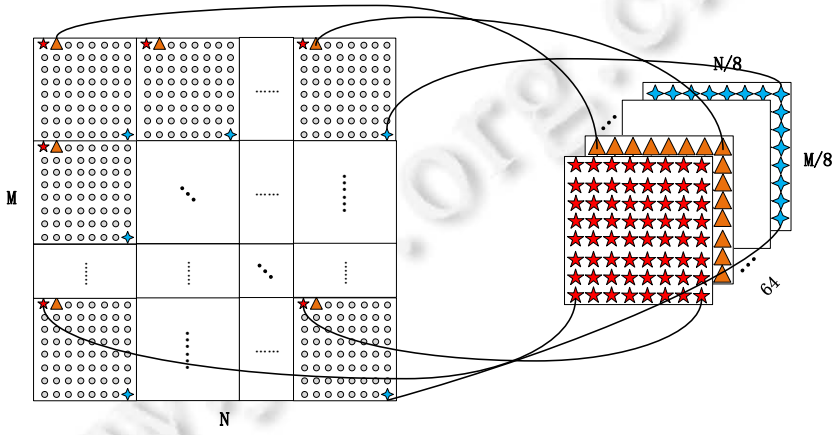

Fig.7 Phase split module

图7 相位分离模块

2018年,Yedroudj等人[69]提出了一个采用Alex-Net[70]理念的网络,并命名为Yedroudj-Net.该网络沿用了传统的SRM中的所有高通滤波核,并且所有滤波核的权重不参与网络训练过程中的反向传播.Yedroudj-Net在网络结构上与Xu-Net和Ye-Net[71]存在不少相似处,预处理层采用与Ye-Net相同的30个滤波核,但不同的是Ye-Net的预处理层参与训练过程.Yedroudj-Net在除去预处理层的其他网络架构上与Xu-Net极为相似,都采取了绝对值层与批正则化层.Yedroudj-Net还使用了Ye-Net模型提出的截断激活函数(truncation activation function,简称TLU),并且在最后的判别网络部分采用了3层全连接层.Yedroudj-Net与SRM,Xu-Net和Ye-Net的误检率对比见表5.

Table 5 Comparison of detection error of Yedroudj-Net and other steganalysis models

表5 Yedroudj-Net与其他隐写模型的误检率对比

隐写分析器	WOW (%)		S-UNIWARD (%)	
	0.2bit/pixel	0.4bit/pixel	0.2bit/pixel	0.4bit/pixel
SRM+EC	36.5	25.5	36.6	24.7
Yedroudj-Net	27.8	14.1	36.7	22.8
Xu-Net	32.4	20.7	39.1	27.2
Ye-Net	33.1	23.2	40.0	31.2

Yedroudj-Net在Xu-Net与Ye-Net的基础上,降低了7个百分点的误检率.这是因为该网络在Xu-Net模型基础上延用了SRM中的30个滤波核作为图像的预处理层和Ye-Net中的截断激活函数,但网络最后的3层全连接层会使得网络的收敛速度变慢,且易受到对抗样本的攻击.在Yedroudj-Net网络的基础上,Deng等人在2019年[72]首次将全局协方差池化[73]引入基于深度学习的隐写分析领域,并且为了加速该网络的拟合速度,在训练过

程中采用了迭代计算平方根的方法帮助网络加速拟合,取得了优异的效果.

2.2 基于宽网络的半学习隐写分析模型

2017 年,Zeng 等人[74]首次提出一种基于深度学习的频域图像隐写分析模型(后称 Zeng's model).Zeng's model 首先采用 25 个固定的 DCT 基础块作为预处理层,对频域图像进行处理;经过预处理层后,得到一个 25 层的特征图,再放入与 DCTR 取值相同的量化截断值层;然后,将这些经过截断与量化操作的 25 通道的信噪比信息放入与 Xu-Net 结构相似的子网络中分别运算;最后,将 25 个不同子网络提取的长度为 125 的一维信息进行级联,放入全连接层中进行判别.

从表 6 的 Zeng's model 与其他隐写分析模型对比的实验结果可以看出:Zeng's model 在检测的精确率上略优于 DCTR 且略劣于 PHARM;虽然 Xu-Net 模型在较低嵌入率(bit per non-zero AC DCT coefficient,简称 bpnzac)下的效果并不是很好,但也证明了基于深度学习的隐写分析模型不再是针对单独某个域具有检测能力.Zeng's model 的出现,也标志了在深度学习所拥有的强大算力在频域隐写分析这一领域崭露头角,也为后来的频域隐写分析模型打下了基础.

Table 6 Comparison of detection error of Zeng's model and other steganalysis models

表 6 Zeng's model 与其他隐写分析模型的误检率对比

Model	bpnzac (%)			
	0.1	0.2	0.3	0.4
Xu-Net	49.7	47.5	46.2	43.6
DCTR	48.2	44.9	42.3	38.2
Zeng's model	48.2	43.7	40.2	35.5
PHARM	47.5	43.8	39.1	33.2

2018 年,Zeng[75]又在 Zeng's model 的基础上提出一种将 JPEG 域转化成为空域图像后,再进行隐写检测的模型,考虑到太宽的网络不仅难以训练,而且会使得网络获取太多的冗余信息,减少原有 Zeng's model 上子模块的数量.实验结果表明[75]:在网络收敛速度与精确率,都相对于 Zeng's model 有了较为显著的提升.

2018 年,Li 等人[76]提出一种名为 ReST-Net 的结构,该网络在 Xu-Net 模型的基础之上融合宽度网络思想,采用 Inception[77]结构.ReST-Net 希望通过 3 个子模型的并行,可以获取更多的经过预处理的图像信息.在不同的子模型中采用 Sigmoid、ReLU、TanH 这 3 类函数不同组合方式的应用,以获取具有不同结构的载密图像信息,从而从多方面获取隐写痕迹.ReST-Net 的 3 种子网络采用不同的滤波器:Subnet#1 选用 16 个不同参数组合大小为 6×6 的 Gabor 滤波核作为预处理层;Subnet#2 选用 16 种不同的 SRM 滤波核作为一种线性的预处理方式;Subnet#3 则先采用 SRM 滤波核进行线性处理,再将得到的预处理信息投入到经过不同角度旋转过的 SRM 滤波核内进行非线性处理,最后输出 14 个非线性特征图.

如表 7 的实验结果所示,不仅 ReST-Net 本身,ReST-Net 的 3 个子网络在检测准确率上相较于 Xu-Net 都存在明显的提升.

Table 7 Comparison of accuracy of Xu-Net and ReST-Net with subnets

表 7 ReST-Net 及其子网络与 Xu-Net 准确率对比

CNN scheme	S-UNIWARD (BPP) (%)				
	0.1	0.2	0.3	0.4	0.5
Xu-Net	57.33	66.67	73.68	80.24	83.54
Subnet#1	66.42	69.58	76.13	84.62	85.87
Subnet#2	64.15	68.73	76.44	84.28	86.17
Subnet#3	60.24	66.37	74.75	84.28	86.17
ReST-Net	65.67	71.35	78.78	85.44	87.93

在 ReST-Net 中,Li 认为,不同 Subnet 之间的组合也会对隐写分析的准确率产生不一样的影响.ReST-Net 采取如下 6 种不同的模型组合方式.

1) 仅采用一个子网络共同使用 Gabor、SRM 线性和 SRM 非线性滤波核(将 3 个子网络融合成 1 个).

2) 与方式 1)类似,采用单子网络结构,不同的是不采用 Gabor 滤波器.

3) 采用 ReST-Net 中 Subnet#1 与 Subnet#2 的组合.

4) 采用 ReST-Net 中 Subnet#1 与 Subnet#3 的组合.

5) 采用 ReST-Net 中 Subnet#2 与 Subnet#3 的组合.

6) 采用 4 个并行子网络,将原本的 Subnet#1 拆分成两个子网络,与 Subnet#2 和 Subnet#3 共同使用.

在表 8 的消融实验中,X 代表未经过修改的 ReST-Net 模型,训练集所采用的隐写算法都是 S-UNIWARD.针对于第 1 种方案,虽然网络总体的层数并没有减少,但仅使用一个子网络时,这种串连结构在检测准确率上不如并联结构.除此以外,上述结果表明,并联网络数量越多检测效果越好.当子网络的数量从 3 个增到 4 个时,检测准确率的增长并不明显,但会消耗大量的服务器算力资源并且使得网络的收敛速度更慢.ReST-Net 考虑到这个原因,并且权衡其中的利弊,最后仅使用 3 个子网络.

Table 8 Detection accuracies of six cases of subnets are used (%)

表 8 6 类子网络组合的检测准确率(%)

Case	I	II	III	IV	V	VI	X
0.1bit/pixel	61.95	60.91	64.48	63.12	64.72	67.13	65.67
0.4bit/pixel	82.76	81.85	84.93	84.49	84.07	85.73	85.44

2.3 半学习模型小结

上述的方法都是半学习隐写分析模型,将 SRM 的滤波核或者固定的处理方式放入网络的预处理层中对图像进行处理,固定其中滤波核内的权重参数,充分融合 SRM 这类非深度学习隐写分析的特点,再依赖深度学习强有力的拟合能力进行训练[63,65,66,68,69,72,76,78],在发展中逐渐超越了传统隐写分析模型的检测能力.

在空域上,其他研究者根据深度学习的网络不断改进,对于隐写分析的网络做出相应的改变.Qian[79]提出了对于模型采用深度学习增强方式进行迁移学习,对于原始图像的信号做信号增强等操作,并把网络架构和现实应用相结合.Qian 等人[80]提出了由于神经网络的训练存在单一性和随机性很难让网络学习到图像的全局信息,通过迁移学习方法,利用传统隐写分析方法与特征分析来增强隐写分析模型对于全局统计信息的学习能力.但是迁移学习也会导致许多的效果受到限制,不仅如此,由于载体图像与载密图像之间的差异较小,如果是 0.1bit/pixel 甚至更低的迁入率,就容易导致网络结构难以收敛的问题.这是因为通过特征提取步骤得到的像素间差异和共生矩阵结构相似太大.为了解决这样的问题,Qian 提出了另外一种迁移学习方法,让隐写分析网络从高嵌入率的样本集中学习到如何区分载体图像与载密图像之间的差异,然后将已经训练好的网络迁移到低嵌入率的样本空间中.这样可以有效地减少训练成本与时间,提高隐写分析模型的检测效果.在频域上,Chen 也在 PNet 中提出了迁移想法,但在 UED 隐写算法上训练的模型迁移到 J-UNIWARD 隐写算法的图像上会出现过拟合的现象,且在测试效果远低于传统频域隐写分析的检测效果.

3 全学习隐写分析

在本节中将介绍全学习隐写分析模型,全学习网络是指在训练过程中,预处理层中的参数会随着网络反向传播一起更新.在本节中,按照网络的架构分为深度网络模型与宽度网络模型.

3.1 基于深度网络的全学习隐写分析模型

2014 年,Tan 等人[81]首次将隐写分析与深度学习相结合,激发了基于深度学习的隐写分析新浪潮,并且给这种网络结构简称为 TanNet.该网络结构一共只有 4 层网络,分别由 3 层卷积层和一层全连接层组合而成.Tan 提出了 3 种不同的方案,用以证明将深度学习与隐写分析相结合方法的可行性与有效性.

1) 随机初始化第 1 层卷积核.因为 SRM 有各种各样不同滤波核的存在,Tan 认为,通过这种随机初始化的卷积核会存在比人工设计的卷积核效果更好的可能.

2) 使用滤波核初始化第 1 层卷积核(滤波核乘以随意初始化的卷积核).

3) 使用滤波核作为初始化第 1 层卷积和以及使用栈式卷积自动编码器与训练每个卷积层.

在 BOSSBase 数据集下,使用 HUGO 自适应隐写算法对上述 3 种模型进行负载为 0.4bit/pixel 的比较实验.评价一个隐写分析器是否有效,需要在大量的数据集上测试得出最后的评价指标.

表 9 是 TanNet[61]和 SPAM、SRM 在传统数据集上的测试结果.

Table 9 Comparison detection error of different proposals in TanNet

表 9 TanNet 不同方案误检率对比

第 1 方案	第 2 方案	第 3 方案	SPAM	SRM
0.48	0.43	0.31	0.42	0.14

从表 9 的实验结果可以得出:第 3 方案的方法是最好的,相较于 SPAM 这种特征维数较少的传统隐写分析方案,在误检率上提升了 9%.3 种不同的方案得到 3 组不同的数据,说明网络架构本身和隐写分析的效果并没有太大的关系.滤波核的初始化可以提升检测的成功率,但是相较于比较强力的 SRM 而言还是有所不及.这仅仅是深度学习与隐写分析的初步结合,是简单的结合与尝试,不仅证明了深度学习这项技术是可以应用在隐写分析上的,而且给予了未来深度学习有望超过 SRM 的一个观念,但是在网络架构上,比较简单依赖卷积层与全连接层的结构.

2017 年,Ye 等人[71]提出了 Ye-Net 网络,直接将传统的 SRM 中的特征提取中滤波核与深度学习网络结合,利用 SRM 的 30 个高通滤波核共同工作,然后得到了一张通道数为 30 的残差叠加图像.将其放入隐写分析网络中进行训练,让网络可以有效地学习到更多特征信息的残差信息,让网络自己学习矩阵的构建模式与构建大小,利用卷积神经网络来代替 SRM 中的计算残差图像和提取共生矩阵的方法.

不仅在计算残差时添加了各式各样的滤波核,而且相较于之前 Xu-Net 中的网络使用了混合激活函数,还提出了一种新型的截断(truncated linear unit,简称 TLU)激活函数,用以模仿 SRM 中的截断操作.TLU 函数具有更好地适应隐写噪声的分布、收敛速度快等特点,这样使得经过卷积后的特征图具有更好的区分性.因为在 Ye-Net 网络中没有添加绝对值层,所以需要采用截断 TLU 激活函数,更好地方便函数收敛.这是因为隐写算法采用的三元 STC 编码嵌入,得到的噪声残差图像会存在±1 和 0 的三元取值图,利用 TLU 函数可以让网络无论在+1 还是–1 的时候都可以有效地学习到数据进行梯度下降加速损失函数的收敛,更好地找到一个全局最小值点:

$$F(x)=\begin{cases}-T, & x<-T\\ x, & -T\leqslant x\leqslant T\\ T, & x>T\end{cases} \tag{10}$$

公式(10)是 TLU 函数的具体公式,T 作为截断数值,将绝对值大于 T 的数值赋值为 T,用来限制整个激活函数的数值情况,防止数值间差异太大.表 10 就 T 的具体数值进行讨论.

Table 10 Comparison of experimental detection error of TLU and ReLU activation functions on Ye-Net

表 10 在 Ye-Net 上 TLU 与 ReLU 激活函数的实验误检率结果对比

隐写算法	ReLU	TLU		
		T=3	T=7	T=∞
WOW	0.213 6	0.198 2	0.196 6	0.217 0
S-UNIWARD	0.293 7	0.254 0	0.262 4	0.299 0
HILL	0.297 1	0.276 1	0.281 2	0.295 5

通过在 3 类传统隐写方法上的大量隐写检测实验对比可以看出:TLU 激活函数与 ReLU 激活函数在 T=∞时,误检率是相近的.值得注意的是:当 T=∞时,TLU 激活函数的表现形式就是一个线性函数了.TLU 激活函数中 T 的数值经过实验被证明:在 T=3 或者 T=7 时网络的检测效果最好.

Ye-Net 首次在隐写分析网络的训练过程中添加了通道选择感知,并且通过大量实验也证明这种方法存在一定的优势,可以帮助隐写分析网络更好地收敛和更好地实验检测效果.将选择通道与卷积神经网络相结合,能

够提升对自适应隐写算法的准确率,在纹理复杂处和细节处检测效果更好.

此外,Ye-Net 还验证了数据集对于深度神经网络的训练会产生巨大的影响,大规模的实验样本可以提升网络训练的稳定性.在 WOW[26]、S-UNIWARD[27]、HILL[28]这种基于空域的隐写算法的测试上,Ye-Net 及其对应的网络架构检测能力都以明显的优势超越了 SRM 和 maxSRMd2[78].此时是基于深度学习的隐写分析技术第一次超越传统的隐写分析技术,这也是基于深度学习的隐写技术发展史上的里程碑.在训练网络的过程中,Ye-Net 采用的梯度下降方法时并不是批次梯度下降(batch gradient descent,简称 BGD)[82],而是采用 AdaDelta[83]作为梯度下降的优化器.

在文献[71]中,Ye 等人认为:在深度学习的隐写分析模型训练中,训练集的大小对于训练的结果会有一定的影响.训练集的大小和训练结果的关系见表 11.

Table11 Detection error comparison of Ye-Net, SRM and MaxSRMd2 under data enhancement

表 11 Ye-Net、SRM 和 MaxSRMd2 在数据增强下的误检率对比

隐写检测	BOSSbase	BOSSbase+BOWS2	BOSSbase+BOWS2+AUG
SRM	0.326 6	0.322 8	N/A
MaxSRMd2	0.242 4	0.232 5	N/A
TLU-CNN	0.336 4	0.269 3	0.198 2

注:BOSSbase 与 BOWS2 数据量为 10 000 张,AUG 为数据增强手段

随着训练集数据量的增大,Ye-Net 误检率也会逐渐变小.针对于基于深度学习的隐写检测,可以利用例如旋转、翻转等数据增强的办法来增加训练集的数据量.但是根据横向对比可以看出,数据增强对与传统的隐写分析的准确率没有明显影响.

从表 12 的实验结果中可以看出:加了 TLU 激活函数的误检率结果更低,在不同的空域隐写算法内,都至少降低了 3 个百分点.这对于传统的隐写分析是具有改革效果的,但是相较于其他隐写分析模型而言,Ye-Net 模型的架构较为简单相较于 Xu-Net 模型所做到的检测效果的提升并不明显.在实际训练过程中,Ye-Net 由于其预处理层的学习操作,会使得网络本身出现更加难以收敛、复现效果差等问题.

Table 12 Comparison of Ye-Net and other models'detection error on different steganographic algorithms

表 12 Ye-Net 等模型在不同隐写算法上的误检率对比

算法	嵌入率	SRM	TLU-CNN	MaxSRMd2
WOW	0.1	0.406 6	0.300 0	0.316 3
	0.5	0.180 0	0.093 8	0.133 1
S-UNIWARD	0.1	0.423 2	0.335 0	0.380 6
	0.5	0.184 8	0.100 3	0.173 2
HILL	0.1	0.453 0	0.356 0	0.389 4
	0.5	0.236 3	0.156 1	0.211 5

2018 年,Boroumand 等人[67]提出了一个 48 层基于深度学习的隐写分析器——SRNet,该网络利用了残差网络模拟传统 SRM 在筛选特征的过程.SRNet 不仅可以应用于空域,在 JPEG 域上也有不错的效果.SRNet 的成功,也证明了深度学习网络并不需要的过多的先验知识.Jessica 作为传统隐写分析领航者的一员,认为深度学习具备强大的学习能力,如果单纯依赖传统的隐写分析的滤波核势必会对于网络最后收敛结果存在一定的限制,影响了深度学习的拟合能力.所以 Jessica 提高了网络结构的层数,采用残差结构方式解决网络层数较高时出现的在反向传播过程中的梯度爆炸与梯度弥散情况,帮助网络在训练过程中更容易收敛到一个全局最优解或者全局较优解.但正是因为这种全靠网络依赖方向传播的拟合方式,导致 SRNet 模型在训练过程中所需要耗费的时间也更长,更容易在训练过程中出现损失不动点情况.

从图 8 的 SRNet 网络结构图中可以观察到:SRNet 的前 7 层不使用下采样层(pooling),能有效避免降低隐写信号的能量;同时使用残差结构,有利于网络学习到相应的"隐写噪声残差"特征图和原本依赖计算而产生的共生矩阵.其中,整体网络结构(图 8)的前两层采用的是 Type1 单元层,即线性网络结构,依赖的是深度学习本身的

拟合能力,从而做到提取图像中的细节部分增强;Type2 单元层采用了残差块的理念,将图像数据跳跃连接防止训练过程中的梯度问题;Type3 和 Type4 分别与 Type2 和 Type1 在结构上相似.其中,

- BN 代表 Batch Normalization[84],代表批归一化操作将数据归一化帮助训练.
- 平均池化(average pooling,简称 AP)和全局平均池化(global average pooling,简称 GAP)操作不仅可以缩小计算图并减少运算量,还可以防止采用最大值池化产生的残差信息丢失问题.
- GAP 操作将得到的数据放入全连接层(fully conneted,简称 FC)帮助网络的训练.

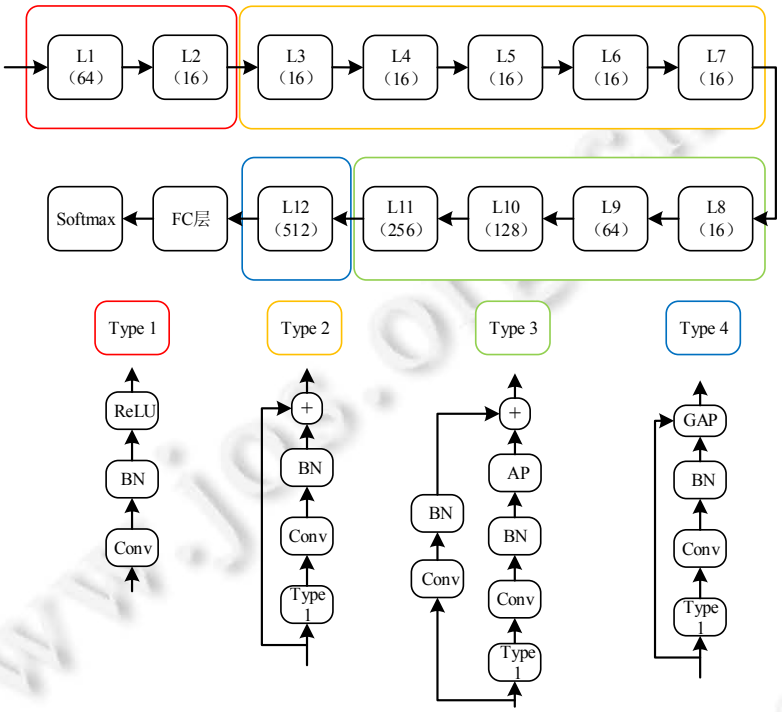

Fig.8 Network structure diagram of SRNet

图 8 SRNet 的网络结构图

SRNet 通过设计多种不同的实验方案,将 SRNet 针对空域隐写术和传统隐写分析比较,SRNet 已经远远超过了 maxSRM 的结果.

在表 13 和表 14 中 SRNet 在空域和频域的测试结果可以看出:SRNet 与 maxSRM 在各类空域隐写算法上的检测效果相比,SRNet 已经远远优于传统的隐写检测方法.SRNet 不仅采纳了不少其他先驱者的想法和工作,也利用迁移学习来证明训练的效果具有泛型能力,得出 SRNet 不仅仅在空域上有效,在 JPEG 域上也有不俗的效果.其中,QF 代表品质系数,系数越高代表图片的质量越高细节更丰富,损失的信息越少.通过实验的结果对比,SRNet 的隐写检测错误率远远低于当时的网络模型,具有绝对的隐写分析的优势.SRNet 是不同于 GNCNN,Xu-Net 这类空域的隐写分析模型:首先,SRNet 依赖残差网络本身对于信息的跳跃利用比较高的原因,实现了既可以在空域上分析隐写,也可以在频域上做到有效地检测;其次,SRNet 在预处理层上是不同于其他的基于深度学习的隐写分析网络,其他的隐写分析模型是将传统的隐写分析研究的滤波核放入网络的第 1 层作为预处理层,而 SRNet 则纯粹依赖深度学习的拟合能力.

再做不同隐写分析网络训练时间对照实验,可以发现:在 SRNet 训练网络中添加一层 HPF 作为高通滤波层,也是具有很不错的效果,可以加速其网络的收敛性.适当地添加几层高通滤波,可以有效地加快网络的训练速度.SRNet 主要依赖网络本身的学习能力,将特征提取、特征增强、二分类训练这些难题都交给网络本身去训练.但是随之而来的问题也让人不得不重视,那就是网络本身训练时间的问题.Jessica 所提出的 SRNet 的训练轮数

(eopchs)也将超过600轮,这个轮数可以通过添加高通滤波核来加快收敛.但是随着高通滤波核个数的添加,就又会导致一系列的问题,就是网络学习到参数不够,让网络变得不那么优秀.SRNet 通过避免启发式的元素,限制了网络本身的灵活性.

Table 13 Comparison of detection error results between SRNet and traditional steganalysis model maxSRM

表 13 SRNet 与传统隐写分析模型 maxSRM 的误检率结果对比

算法	BPP	SRNet		maxSRM	
		$P_{FA}(0.5)$	$P_{FA}(0.3)$	$P_{FA}(0.5)$	$P_{FA}(0.3)$
S-UNIWARD	0.2	0.022 2	0.003 2	0.124 4	0.033 6
	0.4	0.000 6	0.000 2	0.020 0	0.003 4
HILL	0.2	0.048 8	0.010 0	0.160 8	0.043 6
	0.4	0.004 2	0.001 0	0.048 2	0.011 8

Table 14 Detection error result of SRNet on frequency domain

表 14 SRNet 在频域上的误检率结果

嵌入算法	隐写分析网络	*QF*=75					*QF*=95				
		0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
J-UNI	SRNet	0.32	0.189	0.115	0.067	0.039	0.428	0.344	0.252	0.176	0.115
	SCA-SRNet	0.269	0.163	0.092	0.058	0.032	0.371	0.324	0.235	0.134	0.110
UED-JC	SRNet	0.131	0.057	0.029	0.019	0.009	0.304	0.203	0.126	0.088	0.050
	SCA-SRNet	0.125	0.052	0.027	0.015	0.008	0.277	0.167	0.107	0.033	0.039

3.2 基于宽度网络的全学习隐写分析模型

2019 年,Zhu 等人[85]提出了 Zhu-Net 网络.Zhu-Net 相对于之前的隐写分析网络做出了较大的改进,首次在预处理层提出改进的 3×3 的滤波核,在预处理使用 25 个 3×3 滤波核与 5 个 5×5 滤波核组合代替原有 30 个 5×5 的滤波核,这样预处理层的参数减少,从而更容易拟合模型.

在预处理层的初始化上,Zhu-Net 采用了与 Ye-Net 相似的方法,利用 SRM 中手工设计的滤波核对预处理层进行初始化操作,但仅保留其中最为有效的 5 个滤波核,其他的滤波核都用 3×3 的卷积核代替,并且这些权重也随着网络传播过程中而不断更新的.针对于权重问题,Zhu 进行了对照实验,得出结果:在训练过程中,随着 Epoch 轮数的增加,可优化的预处理层在整体的二元交叉熵损失上数值更小,并且这种数值上的差异会不断增加.表 15 是 Zhu-Net 在不同权重优化方案上的误检率.

Table 15 Detection error results of Zhu-Net different preprocessing layer processing schemes

表 15 Zhu-Net 不同预处理层处理方案的误检率结果

隐写算法	固定预处理层	优化预处理层
S-UNIWARD(0.2)	0.324	0.285
S-UNIWARD(0.4)	0.169	0.153
WOW(0.2)	0.243	0.233
WOW(0.4)	0.130	0.118

从表 15 中的误检率结果可以看出:Zhu-Net 对预处理层中的滤波核权重采取优化策略,有助于整体网络的学习,帮助网络收敛.这种收敛效果会随着算法完善性与嵌入的降低变得更加明显.Zhu 为了使得网络对于信噪比信息更加敏感,学习到更加有效的信息,所以网络中都仅采用 ReLU 作为每一层的激活函数,并将空间金字塔池化[86]引进隐写分析,代替全连接层前的全局平均池化操作.2017 年,Baluja[87]将空间金字塔池化引入隐写领域,凭借自编码网络结构完成将彩图藏入彩图的任务,开启了以图藏图的新型隐写模式.

不仅如此,Zhu-Net 不同于其他基于深度学习的隐写分析模型的架构模式,采取了与 Inception[77]和 Xception[88]相似的架构模式,使用两种不同的深度分离卷积模块,获取空域残差特征与通道残差特征信息.

从表 16 中 Zhu-Net 与 Yedroudj-Net 和 SRNet 的误检率对比可以看出:Zhu-Net 凭借其优异的检测准确率,无论在半学习模型还是全学习模型,都取得了最先进的水准.Zhu-Net 相对于其他网络也有巨大的不同.

1) 预处理层精细化,利用 3×3 滤波核代替原有的 5×5 滤波核,从而减少参数数量,加速收敛速度.

2) 采用深度可分离网络,对于预处理层中得到的信息再次精细化.

3) 利用空间金字塔池化代替全局平均池化,使得进入全连接层网络的信息更具有代表性.

Table 16 Comparison of detection error between Yedroudj-Net, SRNet and Zhu-Net

表 16 Yedroudj-Net、SRNet 和 Zhu-Net 误检率对比

算法	嵌入率	Yedroudj-Net	SRNet	Zhu-Net
WOW	0.2	0.206	0.286	0.233
	0.4	0.158	0.099	0.065
S-UNIWARD	0.2	0.305	0.228	0.176
	0.4	0.171	0.123	0.081
HILL	0.2	0.338	0.329	0.262
	0.4	0.208	0.183	0.152

以上全学习模型是依赖深度学习本身的学习来完成的[67,71,81,85].关于预处理层的初始化方式也有不同,其中,TanNet 采用随机初始化,SRNet 采用 Heinitializer 初始化与训练层权重,Ye-Net 与 Zhu-Net 采用人工设计的方式初始化权重.SRNet 由于其庞大的参数量,收敛速度缓慢,收敛时间相对于 Zhu-Net 要多消耗一倍时间.

3.3 全学习模型小结

上述的方法都是全学习隐写分析模型,不使用传统隐写分析中的滤波核作为预处理层,并在网络训练过程中对预处理层中的权重进行更新,利用深度学习的强有力的拟合能力进行训练[67,71,81,85,89].

全学习模型相对于传统隐写分析与半学习模型具有更高的检测精度,但是所需要的训练时间更长,也更容易出现过拟合的情况.在全学习模型的检测过程中,我们发现训练好的网络具有数据集特异性,如果测试集与训练集之间不是同一类型图片,那么测试效果就会降低许多.所以在文献[67,71,85]中,都采用了混合数据集与数据增强的手段来防止网络出现过拟合.

相对于依赖手工设计的传统隐写分析而言,基于深度学习的隐写分析网络利用深度学习本身强大的学习能力,从纷繁复杂的像素信息中选择最为重要的残差信息.这个网络优化的过程是通过损失函数和梯度反向传播来实现的,相较于传统隐写分析不同的是:无论是特征提取还是特征增强的过程,都是建立在网络层结构来辅助实现的;由于网络结构和初始化参数等问题,基于深度学习的隐写分析器具有不确定性和可复现性较差的特点,相同的环境下,在不同时间段的训练可能产生不同的结果.

4 隐写分析总结

从图 9 中隐写分析发展历程可以看出:从 2014 年 TanNet 的提出后,隐写分析也逐渐变成信息安全的热点研究方向.本文根据预处理层是否参与训练,将基于深度学习的隐写分析分为半学习模型与全学习模型.全学习模型[67,71,81,85]相对于半学习模型在训练中更难以收敛,这是因为半学习模型可以通过第 1 层的预处理获得有效的残差信息.首先,这种预处理在一定程度上抑制了图像内容,缩小了动态范围;然后增加了弱 stego 信号(如果存在)与图像信号之间的信噪比,从而帮助网络更有效地获取残差信息[90].但全学习模型在检测精度上略优于半学习模型,这依赖于深度学习本身强大的学习能力.

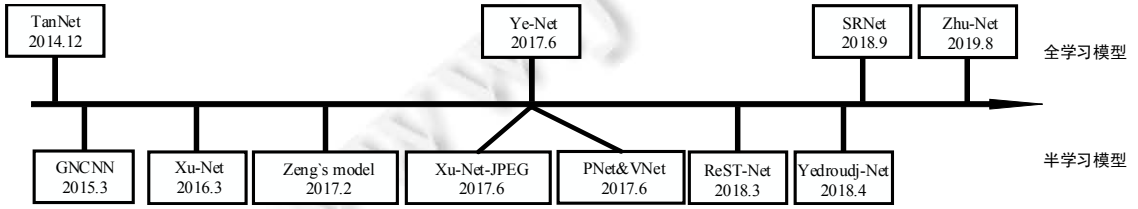

Fig.9 Map of the development of mainstream steganalysis models

图 9 主流隐写分析模型发展图

半学习隐写分析模型都将传统隐写分析中的滤波核作为深度学习网络的预处理层,固定滤波核的参数.与深度学习的网络层参数不同,卷积核权重不受反向传播所影响[63,65,66,68,69,76,91].这种结构将传统隐写分析中SRM或DCTR的滤波核与深度学习网络相结合,故称为半学习隐写分析模型.半学习隐写分析模型相较于全学习模型,所需要的拟合时间更短,并且也拥有不俗的精度.故在应用层面,半学习隐写分析模型更具有参考意义与研究价值.

在表17中,TanNet在这几类空域隐写分析模型中,模型层数最少但是参数量最大.这是因为在全连接层中采用了过多的隐藏神经元,使得网络收敛速度慢且不容易收敛.在我们的设备上无法完成训练,故测试时间尚且为空.众多模型中,Xu-Net参数最少且收敛时间较短,这得益于半学习模式与较少的网络层,但检测效果却不够优秀.虽然GNCNN网络拥有最短的收敛时间,但是模型效果却不如SRM.Yedroudj-Net,Zhu-Net,SRNet这三者的参数量依次递增,预期拟合时间也逐渐增长.其中,Zhu-Net对于参数的拟合速度最快,达到了331.7万/小时.虽然SRNet拥有更多的参数量,但误检率却高于Zhu-Net模型.这意味着在隐写分析领域,盲目地添加网络层数与参数量,并不可以提升网络的拟合效果.由于Yedroudj-Net固定了预处理层中的权重,拟合速度更快,但会造成检测精度上的损失.但是,在复现上述3个模型的过程中,我们发现SRNet拥有其他两个模型所不具备的特点:具有一定的跨数据集迁移能力,并且对于对抗样本具有一定的抗攻击能力.

Table 17 Comparison of parameters, fitting time and test time of different steganographic networks

表17 不同隐写网络的参数、拟合时间、测试时间对比

模型	参数量(万)	拟合时间(h)	测试时间(s)
TanNet	>1000	>72	–
GNCNN	6.3	3.02	20
Xu-Net	1.4	3.2	17
Zeng`s model	811.1	>72	58
YeNet	10.6	3.86	27
Xu-Net-JPEG	252.9	11.2	35
VNet	30.0	4.2	23
ReST-Net	227	5.7	37
Yedroudj-Net	44.5	4.8	25
SRNet	477.6	17.38	41
Zhu-Net	287.1	8.95	33

基于深度学习的隐写分析网络与其他的图像分类模型存在差异:(1) 隐写分析模型所观察的图像更加细致,注重的并不是图像轮廓信息而是高频信息特征,从而提取有效的毗连信息构建关系模型;(2) 隐写分析需要统计全局像素间的差异信息,而不仅仅是考虑局部像素间的差异,判别条件具有统筹性和全局性的特点.

深度学习的训练对于数据集的也具有一定要求,其中,文献[67,71,81]都采用了增大数据量与数据增强的手段帮助网络收敛.不能盲目扩大数据集,要考虑网络可能出现难以拟合的情况.如何实现数据增强,需要根据网络结构来选择.我们在进行对比隐写分析实验的过程中发现:不同训练集的模型,在进行跨数据集检验时会出现较大的误差.

在表18中,本文对各类模型的特点进行了总结,表中误检率计算方式为*Err*公式(1).对于空域隐写分析模型,采用*BPP*=0.4的S-UNIWARD隐写算法;频域隐写分析模型采用质量因子95且*bpnzac*=0.4的J-UNIWARD隐写算法.全学习模型的预处理层权重都是可以更新的,这使得网络拟合所需要的时间更长.而半学习模型在滤波核数量上不及其他网络,且预处理层的参数固定,所以在拟合速度上更具优势.SRNet依赖深度学习本身的学习能力和迁移学习、数据增强等训练技巧,取得了较好的检测效果,但是网络的训练时间和训练轮数都远超其他网络.ReST-Net与Zhu-Net在网络结构上相较于其他隐写分析差异较大,首次将多个子模型采用并联的方式结合.传统隐写分析中的滤波核,不仅在设计上依赖研究人员的先验知识,还会使其存在一定的“知识”偏差,需要研究人员的不断尝试探索.而半学习模型和全学习模型都建立在深度学习的强大算力上,所以基于深度学习的隐写分析模型在检测率上逐渐取得优势.在上述各类隐写分析网络的训练过程中,SRNet在训练过程所需要的收敛时间更多,其他采用传统滤波核初始化的网络收敛时间更短.隐写分析网络中,超参数对Cover与Stego间信噪比信息的学习效果影响网络收敛,例如BN层的参数初始化、网络权重的初始化方式等[92].BN层对于隐写分

析乃至计算机视觉都有非常巨大的影响,Zeng在文献[74]中通过消融实验发现,缺失BN层的网络难以训练甚至无法正确判别数字图像.Chen在文献[68]中对BN层的作用进行了论述,提出BN层对于深度学习本身的重要性,并且提出了一种对于BN层处理的训练方式.

Table 18 Comparison of different steganalysis network based on deep learning

表 18 不同基于深度学习隐写网络的对比

	预处理层	激活函数	模型特点	池化方式	初始方式	误检率(%)
TanNet	40 个滤波核	Sigmoid	线性	Max pooling	Evenly distribution	31
GNCNN	1 个滤波核	Gaussian/Tanh	线性	Average pooling	Xavier	25.2
Xu-Net	1 个滤波核	TanH/ReLU	线性	Average pooling	Xavier	24
Xu-Net-JPEG	16 个滤波核	ReLU	跳接	Average pooling	Xavier	23.6
VNet	4 个滤波核	ReLU/Tanh	线性	Average pooling	Normal distribution	16.9
Ye-Net	30 个滤波核	TLU/ReLU	线性	Mean pooling	Xavier	22
Zeng`s model	25 个滤波核	ReLU	并行	Average pooling	Xavier	35.4
Yedroudj-Net	30 个滤波核	TLU/ReLU	线性	Average pooling	Xavier	20
ReST-Net	44 个滤波核	Sigmoid/TANH/RELU	并行	Average pooling	Normal distribution	16
SRNet	–	ReLU	跳接	Average pooling	Kaiming	18
Zhu-Net	30 个滤波核	ReLU	并行+跳接	SPP	Xavier	15.3

在激活函数方面,GNCNN采用独有的高斯激活函数,Zhu-Net与Ye-Net采用截断激活函数,其他隐写分析模型中激活函数大致相同.在文献[93]中,Pibre认为:池化层是一种低通滤波操作,如果池化操作与预处理层距离相隔层数太近,会对预处理层得到的高通滤波信息造成不可恢复的损坏.所以在文献[67,71,81]中,预处理层和池化操作都距离较远,保障得到的高通残差信息不被损坏[94].

以上方法大多都是空域的隐写分析方法,也有不少针对于JPEG域的隐写分析模型.由于经过JPEG压缩过后会发生图像损失丢失一部分信息,但相应的图像大小却会缩小很多,因此在现如今的社交通信中,大多会采用JPEG压缩来提升通信速度.因为频域上的隐写会将图片转化成8×8的小块(JPEG phase),然后在系数中进行修改,所以早期的JPEG图像隐写分析通常从JPEG的处理方式来着手,利用DCT系数来计算残差和提取特征.2016年,张等人[95]提出一种最低有效位特征拓展方法,通过构造高阶共生矩阵的方式辅助判别.后来的JPEG图像隐写分析则根据解压缩过程中放大信号和分块相位的特点.Chen[68]于2017年将深度学习应用在JPEG[69]域图像隐写分析,通过分析JPEG压缩的各种操作,例如采样、分块、DCT变换、ZigZag扫描、量化等操作,并且采用了不同的卷积核,更有效地帮助网络获取空域和频域的像素信息和隐写噪声.相似的基于深度学习应用在JPEG域上的隐写分析算法还有文献[96–98],但相较于之前的其他模型,都会显得不太“智能”,因为需要研究人员借鉴先验知识设计专用的残差特征矩阵,抑或是DCT变化系数与量化矩阵等.就未来针对JPEG域的深度隐写分析提出如下方案:(1) 用更加有效的过滤器去替代DCT,或是采用一些带有先验知识的公式;(2) 减少池化层的使用或是用卷积层代替池化,以防止信噪信息的丢失,能有平均池化就不用最大池化;(3) 不断尝试各种网络结构的应用与优化.该网络也将残差结构[11]引入,这样可以有效地缓解梯度弥散问题[66].

总而言之,无论是空域还是频域的隐写分析,都是为了抑制图像内容,同时获取隐写噪声信息.那如何将基于深度学习的隐写分析本身的特点与传统隐写分析的特点结合起来,两种技术不断交融、共同发展,就成为一个问题.传统隐写分析模型可以通过设计新的滤波核提升其检测效果,但是基于深度学习的隐写分析模型会因为深度网络训练而受到对抗样本的攻击,导致检测准确率直线下降.在实验数据迁移测试中,我们发现:全学习隐写分析具备更强大的数据迁移能力与泛化能力,不同域的隐写分析模型也具有一定的迁移能力.

5 隐写分析检测对抗样本

随着深度学习的快速发展与进步,深度学习也被应用在许多条件严格的环境下.然而,深度学习对于通过一系列“精心”设计的输入样本,它的结果就可能是脆弱的、错误的,这种样本也被称为对抗样本.对抗样本对人类是很容易分辨的,但却能在测试或部署阶段,很容易地糊弄深度神经网络.当应用深度神经网络到对安全有严格要求的环境中时,处理对抗样本造成的脆弱性变成已成了一个重要的任务.

5.1 对抗样本

对抗样本是深度学习中非常有趣的一个现象.攻击者希望添加一个不被人类察觉的扰动,让训练好的深度学习网络将攻击过后的图片错误分类.这是因为判别网络的工作依赖卷积层获取的大量图像参数,而深度学习模型的输入和输出大多是线性的,微小的扰动经过网络层的强化,就会使网络的判别产生偏差.

在图 10 中,首先将左侧大熊猫的图片设为 x,训练好的判别网络给予的置信度为 57.7%;通过添加一层置信度为 8.2%的线虫噪声图扰动,最后将得到的两个图片叠加,通过判别网络得到一个置信度为 99.3%的长臂猿结果.中间的噪声图通过判别网络进行梯度计算得出:

$$sign(\nabla_x J(\theta,x,y)) \tag{11}$$

在公式(11)中,$sign(\cdot)$表示计算出梯度的方向,$sign(\cdot)$函数内的则是损失函数的梯度.ε在图 10 中代表超参数且数值为 0.07,用来保证图片质量,防止图片添加对抗噪声之后质量受到影响.从噪声扰动与添加扰动后的熊猫图可知,深度学习网络中学到的知识与预期效果存在一些偏差.

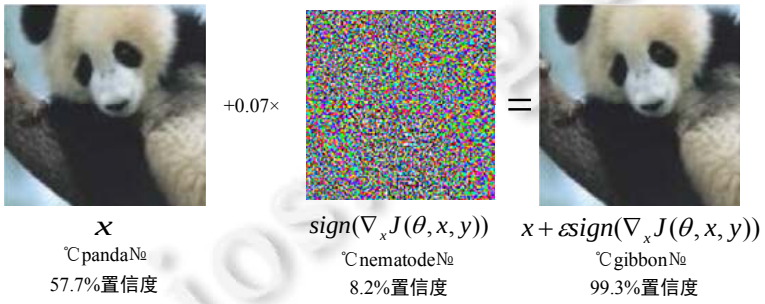

Fig.10 Processing of adding adversarial examples

图 10 对抗样本的添加过程

给原有图像添加相应网络梯度噪声的生成对抗样本的方法被称为快速梯度法(fast gradient sign method,简称 FGSM)[99],它是一种基于梯度的攻击方法.后来,在此基础上不断改进,出现了迭代梯度法(iterative gradient sign method,简称 IGSM)[89].通过迭代的方式不断修改扰动,直到可以改变网络的判别结果.这种方法添加的扰动更小,但会消耗更多的计算时间.基于优化的攻击方法(Carlini & Wagner method,简称 C&W)[100]可以分为有目标攻击和无目标攻击,损失函数中的距离控制可以分为一范式距离、二范式距离和无穷范式距离.其中,二范式距离加上无目标攻击的效果最好,并且可以作为一种黑盒攻击的方式.单像素攻击(one pixel attack)[101]与上述的攻击表现形式不一致,只会修改原图中的一个像素点,但改动的数值较大,容易被人眼观察.

5.2 对抗隐写

对抗样本及其变种的出现,使深度学习的安全受到了极大的挑战.隐写术与隐写分析不可避免地受到影响,对抗隐写应运而生.国内外学者认为:对抗样本可以干扰隐写分析网络判别,可以在秘密信息嵌入前[102]和秘密信息嵌入过程中[103−105]添加对抗样本.2018 年,Ma 等人[103]提出了 AEN 模型,将对抗样本与传统隐写结合起来,在隐写的过程中添加对抗样本.该模型有效地提升了传统空域隐写的安全性,提高了载密图像抵抗基于深度学习的隐写分析的能力.Zhang 等人[102]利用生成式对抗网络生成一种“增强”载体,用以抵抗基于深度学习的隐写分析.“增强”载体经过隐写术之后仍然携带对抗样本的效果,但是这些扰动的添加会使图像更容易受到其他隐写分析模型的检测.2019 年,Li 等人[104]提出了对抗嵌入的方法 ADV-EMB:首先,将图片像素随机分为普通像素和可修改像素;然后,在失真损失函数的基础上,对每个像素进行有效的权重修改,在隐写过程中,将对抗样本融入;最后,使目标隐写分析模型误判.Pevny 等人[105]在 ASO[106]与 ADV-EMB 的基础上,将所有的基于深度学习的隐写分析模型组成一个集合,利用对抗训练的思想与迭代方式建立一个自适应隐写的损失函数,使 Xu-Net 的检测准确率下降了 13 个百分点.

目前,对抗样本在隐写术中的应用较少,未来将会出现不同的对抗样本与不同的隐写术相结合的方法,可以

利用 SRM 与 SPAM 这类传统隐写分析模型检测对抗样本.

5.3 基于隐写分析的对抗样本检测

2018 年,Pascal[107]认为对抗样本与隐写内容存在相似之处并将它们进行对比,提出将传统隐写分析模型应用于对抗样本检测.

表 19 中,主要不同在于目标图来源:用于深度学习判别网络的图片都是真实的自然图像,而用于隐写分析的载密图片添加了攻击扰动.两种算法的攻击目标相同,都针对图像本身独立的像素进行攻击;攻击的方式不同,对抗样本采用修改方式使得图像越过决策边界,而隐写分析则将秘密信息嵌入到载体图像.对抗样本根据反向梯度传播计算得出,可以认为对抗样本是一种带方向的隐写内容,但修改内容对网络的判别存在影响.在基于卷积神经网络的判别模型中,一方面对抗样本的存在使网络判别不够准确;另一方面,将对抗样本加入训练过程进行“投毒”训练,可以使判别网络更加稳定、更具有鲁棒性和迁移能力.根据表中的对比可知:对抗样本的添加,会破坏像素间的相关性.故可以借鉴传统隐写分析的检测理念,通过共生矩阵分析像素之间的相关性.传统隐写分析模型会将带有对抗样本的载体图像判定为载密图像,这使得经过加密的载密图像更容易被判别.因此,可以使用传统隐写分析方法判别一张图片中是否添加对抗样本.

Table 19 Similarities and differences between deep learning and steganalysis

表 19 深度学习与隐写分析的异同

	安全深度学习	隐写分析
攻击方式	对抗样本	载密图像
攻击者	网络结构	隐写者
攻击算法	修改	嵌入
攻击标准	内部参数	自适应准则
原始分布	其他分类	载体图像
目标分布	目标分类	载密图像
攻击目的	混淆原图与目标图	混淆原图与目标图
目标图来源	外部获取	由攻击生成

Liu 等人[108]提出将传统的隐写分析方法应用在对抗样的检测中,利用传统的 SPAM 和 SRM 来检测图像是否添加对抗样本.不仅如此,文献[108]分析了对抗样本的特征属性,提出了增强型的 SPAM 和 SRM 方法.参考传统隐写术通过最小失真函数得到嵌入概率图的理念,Liu 通过模拟 N 分类问题中 L 种对抗样本的添加方法:假设判别器是 N 分类判别器,首先随机选择其中 L 个分类,然后生成 L 张不同的修改概率图,将得到的修改概率图转换成一个二值化矩阵,最后计算 L 张图像像素点的平均值,得到一张平均化的修改概率图.在之后的研究中,可以利用注意力机制帮助网络训练,提高检测精度.

5.4 隐写分析检测对抗样本小结

对抗样本的存在,为深度学习的发展敲响了警钟.一味地追求网络层数的叠加,会使得模型的拟合能力增强的同时,也会操作决策边界的精细化,对抗样本也就越有效.如何在保证网络精度不受到损失的情况,如何提升网络本身的鲁棒性.

对抗样本由于网络的变化而具有多样性,所以对抗样本的检测存在以下困难:(1) 对抗样本的修改更加细微,需要更强大的残差计算方法放大像素间的差异;(2) 对抗样本的位置根据目标网络的模型参数确定,因此修改位置变化很大.为解决上述的两个问题,需要研究对抗样本的产生机理,从而优化判别器的网络结构,提高检测对抗样本的能力.

对抗样本可以视为一种带方向的隐写内容,这使得载体图像在 SRM 与 SPAM 这类不依赖深度学习网络的隐写分析模型更容易被检测出来.因此,可以利用基于深度学习的隐写分析模型与传统隐写分析建立多角度投票机制来检测对抗样本.

6 总结与展望

6.1 总 结

本文从基于深度学习的隐写分析模型这个方面对近期的图像隐写分析模型进行了总结与归纳.深度学习网络与隐写分析方法结合,可以在训练过程中不仅自动学习图像的信噪比信息,还可以在一定程度上完成结构信息的统计.因此,将深度学习与信息隐藏领域相结合,不仅增强了隐写分析方法的检测能力,还提升了隐写分析方法鲁棒性.隐写分析模型判断依据是图片的信噪比信息,基于深度学习的隐写分析模型极易出现过拟合现象,对抗样本的出现,使得基于深度学习隐写分析模型不得不防止网络出现的过拟合问题.随着基于深度学习的隐写分析模型快速发展,神经网络强大的特征提取能力使隐写分析模型的判别能力不断增强,但是基于深度学习的隐写分析模型存在着这样几个问题.

1) 拟合速度慢.例如,Zhu-Net 需要至少 8 个小时的训练时间,而 SRNet 需要至少 22 个小时的训练时间,如果训练一个经过数据增强的训练集,通常在 GPU 上训练 1 周以上才最后收敛.其中,全学习隐写分析模型所需要的时间更长且不容易收敛,极易受到局部最小值的干扰.
2) 迁移能力弱.由于隐写算法的不断进化、自适应隐写算法与基于深度学习的隐写算法的出现,隐写分析网络不得不学习更加精细化的信噪比信息,这也导致了模型跨数据集检测能力弱.不同的数据集之间采用相同的算法,也会因为不同的相机指纹或是不同光照因素甚至拍摄角度等问题,导致模型迁移能力较差检测效果弱的问题.
3) 预处理层依赖强.经过实验测试,基于深度学习的半学习隐写分析模型十分依赖预处理层操作,如果预处理操作不符合网络拟合条件或是不能提取有效的信噪比信息,那网络模型会在不动点上停留很久.但现有的全学习模型除了 SRNet 外都采用传统的隐写分析高通滤波核作为预处理层的初始化方式,所以现有基于深度学习的隐写分析模型并不是一个端到端的学习模式.
4) 参数要求苛刻.在实验过程中我们发现:不同的损失函数与不同的学习率,甚至 BN 层中的超参数对于网络的训练都有着巨大的影响,有些参数不仅在最后的检测精度上产生改变,甚至直接影响网络的收敛时间甚至是否收敛.

在文献[59,90,109,110]中发现:基于深度学习的隐写方法为了保证能够有效地抵抗隐写分析模型的检测,通过在隐写模型训练中加入隐写分析网络进行对抗训练,提升隐写模型的抗检测能力.随着深度学习的应用越来越普及,深度学习的安全性也愈发重要.所以,建立一个安全、有效的深度学习网络,使其增强抵抗对抗样本的能力也变得重要.对抗样本的出现,对于深度学习的应用来说是一项巨大的挑战,基于深度学习的隐写分析研究可以有效检测图像中的对抗样本.在未来的研究过程中,不断提高隐写分析模型的检测精度,从而保证深度学习技术在应用中的安全问题.近年来,研究真彩色图像的隐写分析也逐渐走进人们的视野里,如何检测不同大小的 JPEG 图像,也有了不少的研究[59,106,110,111].

6.2 展 望

基于深度学习的隐写分析研究方兴未艾,但是仍然存在一些问题有待改进.在未来的研究过程中,可以针对这几个方面对基于深度学习的隐写分析进行研究.

1) 实现完全端到端的学习模式.基于深度学习的隐写分析技术并不是端对端的学习模式,这是因为隐写分析模型具有一定的特殊性,所以各类的模型训练还需要一定的人工干预措施.可以在全学习隐写分析模型的基础上,依赖深度学习本身强大的计算能力,支持实现端到端的学习模式.
2) 提升网络拟合速度.基于深度学习的隐写分析模型都是依赖深度学习网络本身大量的参数来帮助获取特征,所以研究有效的处理方式和更具有方向性的图像处理方法,可以帮助网络拟合.但是基于深度学习的隐写分析网络由于有大量的网络参数,在训练网络的过程中存在不确定性,并且非常依赖网络参数的训练,网络本身训练的轮数也相对比较久.所以,如何对网络架构进行蒸馏,提升网络拟合速度,也成了亟待解决的问题.

3) 小规模数据学习.由于深度学习的训练效果和训练数据集规模息息相关,一个优秀的训练网络都建立在大量的数据集上.但增加训练集数量又会导致网络拟合时间呈几何倍数增张,所以如何在小规模数据集训练的基础上避免过拟合现象的出现,就成为了一个问题.该问题可以通过以下几种方法解决:(1) 实现零样本或少样本学习;(2) 使用不同数据集之间跨领域自适应的迁移学习模型;(3) 收集更多的图片数据集,完成数据集的有效扩充.

4) 多模型融合.目前,GAN网络仅在隐写方面有较多的应用.可以利用GAN网络独有的创造性,将其应用于隐写分析检测模型中.在检测载密图像之前,生成辅助检测的隐写位置图,利用注意力机制帮助隐写分析检测载密图像,加速隐写分析网络的收敛.

References:

[1] Xiang SJ, Luo XR. Reversible data hiding in encrypted image based on homomorphic public key cryptosystem. Ruan Jian Xue Bao/Journal of Software, 2016,27(6):1592−1601 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5007.htm [doi: 10.13328/j.cnki.jos.005007]

[2] Simmons GJ. The prisoners' problem and the subliminal channel. In: Proc. of the Advances in Cryptology. Boston: Springer-Verlag, 1984. 51−67.

[3] Reeds J. Solved: The ciphers in book III of Trithemius's Steganographia. Cryptologia, 1998,22(4):291−317.

[4] Li YF, Ding LP, Wu JZ, Cui Q, Liu XH, Guan B, Wang YJ. Survey on key issues in networks covert channel. Ruan Jian Xue Bao/Journal of Software, 2019,30(8):2470−2490 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5859.htm [doi: 10.13328/j.cnki.jos.005859]

[5] Sutskever I, Hinton GE, Taylor GW. The recurrent temporal restricted Boltzmann machine. In: Proc. of the Advances in Neural Information Processing Systems. 2009. 1601−1608.

[6] Sahiner B, Chan HP, Petrick N, Wei D, Mark A, Dorit D, Mitchell M. Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images. IEEE Trans. on Medical Imaging, 1996,15(5): 598−610.

[7] Chen Y, Zhao X, Jia X. Spectral-spatial classification of hyperspectral data based on deep belief network. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2015,8(6):2381−2392.

[8] Masci J, Meier U, Cireşan D, Schmidhuber J. Stacked convolutional auto-encoders for hierarchical feature extraction. In: Proc. of the Int'l Conf. on Artificial Neural Networks. Berlin, Heidelberg: Springer-Verlag, 2011. 52−59.

[9] Mikolov T, Karafiát M, Burget L, Kombrink S. Recurrent neural network based language model. In: Proc. of the 11th Annual Conf. of the Int'l Speech Communication Association. 2011. 2877−2080.

[10] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proc. of the Int'l Conf. on Medical Image Computing and Computer-assisted Intervention. Cham: Springer-Verlag, 2015. 234−241.

[11] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 770−778.

[12] Iandola F, Moskewicz M, Karayev S, Girshick R, Darrell T, Keutzer K. DenseNet: Implementing efficient ConvNet descriptor pyramids. arXiv preprint arXiv:1404.1869, 2014.

[13] Zhai LM, Jia J, Ren WX, *et al*. Progress in deep learning in the field of image steganography and steganalysis. Journal of Cyber Security, 2018,3(6):2−12 (in Chinese with English abstract).

[14] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the Advances in Neural Information Processing Systems. 2014. 2672−2680.

[15] Aumann R, Brandenburger A. Epistemic conditions for Nash equilibrium. Econometrica: Journal of the Econometric Society, 1995,63(5):1161−1180.

[16] Mirza M, Osindero S. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014.

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

[18] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. arXiv preprint arXiv:1701.07875, 2017.

[19] Chan CK, Cheng LM. Hiding data in images by simple LSB substitution. Pattern Recognition, 2004,37(3):469−474.

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

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

[22] Westfeld A. F5—A steganographic algorithm. In: Proc. of the Int'l Workshop on Information Hiding. Berlin, Heidelberg: Springer-Verlag, 2001. 289−302.

[23] Fridrich J, Goljan M, Lisonek P, Soukal D. Writing on wet paper. IEEE Trans. on Signal Processing, 2005,53(10):3923−3935.

[24] Sachnev V, Kim HJ, Zhang R. Less detectable JPEG steganography method based on heuristic optimization and BCH syndrome coding. In: Proc. of the 11th ACM Workshop on Multimedia and Security. 2009. 131−140.

[25] Pevný T, Filler T, Bas P. Using high-dimensional image models to perform highly undetectable steganography. In: Proc. of the Int'l Workshop on Information Hiding. Berlin, Heidelberg: Springer-Verlag, 2010. 161−177.

[26] Holub V, Fridrich J. Designing steganographic distortion using directional filters. In: Proc. of the 2012 IEEE Int'l Workshop on Information Forensics and Security (WIFS). IEEE, 2012. 234−239.

[27] Holub V, Fridrich J. Digital image steganography using universal distortion. In: Proc. of the 1st ACM Workshop on Information Hiding and Multimedia Security. 2013. 59−68.

[28] Li B, Tan S, Wang M, Huang J. Investigation on cost assignment in spatial image steganography. IEEE Trans. on Information Forensics and Security, 2014,9(8):1264−1277.

[29] Zhang T, Ping X. A fast and effective steganalytic technique against JSteg-like algorithms. In: Proc. of the 2003 ACM Symp. on Applied Computing. 2003. 307−311.

[30] Holub V, Fridrich J, Denemark T. Universal distortion function for steganography in an arbitrary domain. EURASIP Journal on Information Security, 2014,2014:Article No.1.

[31] Guo L, Ni J, Shi YQ. Uniform embedding for efficient JPEG steganography. IEEE Trans. on Information Forensics and Security, 2014,9(5):814−825.

[32] Guo L, Ni J, Su W, Tang C, Shi Y. Using statistical image model for JPEG steganography: Uniform embedding revisited. IEEE Trans. on Information Forensics and Security, 2015,10(12):2669−2680.

[33] Volkhonskiy D, Borisenko B, Burnaev E. Generative adversarial networks for image steganography. 2017. https://openreview.net/forum?id=H1hoFU9xe&noteId=H1hoFU9xe

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

[35] Yang J, Ruan D, Huang J, *et al*. An embedding cost learning framework using GAN. IEEE Trans. on Information Forensics and Security, 2019,15:839−851.

[36] Wang SZ, Zhang XP, Zhang WM. Recent advances in image-based steganalysis research. Chinese Journal of Computers, 2009, 32(7):1247−1263 (in Chinese with English abstract).

[37] Ker AD. Steganalysis of LSB matching in grayscale images. IEEE Signal Processing Letters, 2005,12(6):441−444.

[38] Chandramouli R, Memon N. Analysis of LSB based image steganography techniques. In: Proc. of the 2001 Int'l Conf. on Image Processing. IEEE, 2001. 1019−1022.

[39] Westfeld A, Pfitzmann A. Attacks on steganographic systems. In: Proc. of the Int'l Workshop on Information Hiding. Berlin, Heidelberg: Springer-Verlag, 1999. 61−76.

[40] Liu Q, Sung AH, Chen Z, Xu J. Feature mining and pattern classification for steganalysis of LSB matching steganography in grayscale images. Pattern Recognition, 2008,41(1):56−66.

[41] Tan S. Steganalysis of LSB matching revisited for consecutive pixels using B-spline functions. In: Proc. of the Int'l Workshop on Digital Watermarking. Berlin, Heidelberg: Springer-Verlag, 2011. 16−29.

[42] Böhme R. Weighted stego-image steganalysis for JPEG covers. In: Proc. of the Int'l Workshop on Information Hiding. Berlin, Heidelberg: Springer-Verlag, 2008. 178−194.

[43] Xia Z, Wang X, Sun X, Liu Q, Xiong N. Steganalysis of LSB matching using differences between nonadjacent pixels. Multimedia Tools and Applications, 2016,75(4):1947−1962.

[44] Gul G, Kurugollu F. A new methodology in steganalysis: Breaking highly undetectable steganograpy (HUGO). In: Proc. of the Int'l Workshop on Information Hiding. Berlin, Heidelberg: Springer-Verlag, 2011. 71−84.

[45] Luo X, Song X, Li X, Zhang W, Lu J, Yang C, Liu F. Steganalysis of HUGO steganography based on parameter recognition of syndrome-trellis-codes. Multimedia Tools and Applications, 2016,75(21):13557−13583.

[46] Tang W, Li H, Luo W, Huang J. Adaptive steganalysis against WOW embedding algorithm. In: Proc. of the 2nd ACM Workshop on Information Hiding and Multimedia Security. 2014. 91−96.

[47] Jindal N, Liu B. Review spam detection. In: Proc. of the 16th Int'l Conf. on World Wide Web. 2007. 1189−1190.

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

[49] Holub V, Fridrich J. Low-complexity features for JPEG steganalysis using undecimated DCT. IEEE Trans. on Information Forensics and Security, 2014,10(2):219−228.

[50] Holub V, Fridrich J. Phase-aware projection model for steganalysis of JPEG images. In: Proc. of the Int'l Society for Optics and Photonics, Media Watermarking, Security, and Forensics, Vol.9409. 2015.

[51] Goljan M, Fridrich J, Cogranne R. Rich model for steganalysis of color images. In: Proc. of the 2014 IEEE Int'l Workshop on Information Forensics and Security (WIFS). IEEE, 2014. 185−190.

[52] Zhang XP, Qian ZX, Li S. Prospect of digital steganography research. Journal of Applied Sciences, 2016,34(5):476−484. (in Chinese with English abstract).

[53] Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. Neural Processing Letters, 1999,9(3):293−300.

[54] Shen HB, Chou KC. Ensemble classifier for protein fold pattern recognition. Bioinformatics, 2006,22(14):1717−1722.

[55] Liu C, Wechsler H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Trans. on Image Processing, 2002,11(4):467−476.

[56] Xia C, Guan Q, Zhao X, Xu Z, Ma Y. Improving GFR steganalysis features by using Gabor symmetry and weighted histograms. In: Proc. of the 5th ACM Workshop on Information Hiding and Multimedia Security. 2017. 55−66.

[57] Bas P, Filler T, PevnýT. "Break our steganographic system": The ins and outs of organizing BOSS. In: Proc. of the Int'l Workshop on Information Hiding. Berlin, Heidelberg: Springer-Verlag, 2011. 59−70.

[58] Bas P, Furon T. BOWS-2 contest (break our watermarking system). 2008. http://bows2.eclille.fr/

[59] Yousfi Y, Butora J, Fridrich J, Giboulot Q. Breaking ALASKA: Color separation for steganalysis in JPEG domain. In: Proc. of the ACM Workshop on Information Hiding and Multimedia Security. 2019. 138−149.

[60] Yin X, Goudriaan JAN, Lantinga EA, Vos J, Spiertz HJ. A flexible sigmoid function of determinate growth. Annals of Botany, 2003,91(3):361−371.

[61] Fan E. Extended tanh-function method and its applications to nonlinear equations. Physics Letters A, 2000,277(4-5):212−218.

[62] Xu B, Wang N, Chen T, Li M. Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505. 00853, 2015.

[63] Qian Y, Dong J, Wang W, Tan T. Deep learning for steganalysis via convolutional neural networks. In: Proc. of the Int'l Society for Optics and Photonics, Media Watermarking, Security, and Forensics, Vol.9409. 2015.

[64] Deng J, Dong W, Socher R, Li L, Li K, Li F. Imagenet: A large-scale hierarchical image database. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition. IEEE, 2009. 248−255.

[65] Xu G, Wu HZ, Shi YQ. Structural design of convolutional neural networks for steganalysis. IEEE Signal Processing Letters, 2016, 23(5):708−712.

[66] Wallace GK. The JPEG still picture compression standard. IEEE Trans. on Consumer Electronics, 1992,38(1):30−44.

[67] Boroumand M, Chen M, Fridrich J. Deep residual network for steganalysis of digital images. IEEE Trans. on Information Forensics and Security, 2018,14(5):1181−1193.

[68] Chen M, Sedighi V, Boroumand M, Fridrich J. JPEG-phase-aware convolutional neural network for steganalysis of JPEG images. In: Proc. of the 5th ACM Workshop on Information Hiding and Multimedia Security. 2017. 75−84.

[69] Yedroudj M, Comby F, Chaumont M. Yedroudj-net: An efficient CNN for spatial steganalysis. In: Proc. of the 2018 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018. 2092−2096.

[70] Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keuter K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv preprint arXiv:1602.07360, 2016.

[71] Ye J, Ni J, Yi Y. Deep learning hierarchical representations for image steganalysis. IEEE Trans. on Information Forensics and Security, 2017,12(11):2545−2557.

[72] Deng X, Chen B, Luo W, Luo D. Fast and effective global covariance pooling network for image steganalysis. In: Proc. of the ACM Workshop on Information Hiding and Multimedia Security. 2019. 230−234.

[73] Li P, Xie J, Wang Q, Gao Z. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 947−955.

[74] Zeng J, Tan S, Li B, Huang J. Pre-training via fitting deep neural network to rich-model features extraction procedure and its effect on deep learning for steganalysis. Electronic Imaging, 2017,2017(7):44−49.

[75] Zeng J, Tan S, Li B, Huang J. Large-scale JPEG image steganalysis using hybrid deep-learning framework. IEEE Trans. on Information Forensics and Security, 2017,13(5):1200−1214.

[76] Li B, Wei W, Ferreira A, Tan S. ReST-net: Diverse activation modules and parallel subnets-based CNN for spatial image steganalysis. IEEE Signal Processing Letters, 2018,25(5):650−654.

[77] Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. 2017. 4278−4284.

[78] Denemark T, Sedighi V, Holub V, Cogranne R, Fridrich J. Selection-channel-aware rich model for steganalysis of digital images. In: Proc. of the 2014 IEEE Int'l Workshop on Information Forensics and Security (WIFS). IEEE, 2014. 48−53.

[79] Qian Y, Dong J, Wang W, Tan T. Learning and transferring representations for image steganalysis using convolutional neural network. In: Proc. of the 2016 IEEE Int'l Conf. on Image Processing (ICIP). IEEE, 2016. 2752−2756.

[80] Qian Y, Dong J, Wang W, Tan T. Learning representations for steganalysis from regularized cnn model with auxiliary tasks. In: Proc. of the 2015 Int'l Conf. on Communications, Signal Processing, and Systems. Berlin, Heidelberg: Springer-Verlag, 2016. 629−637.

[81] Tan S, Li B. Stacked convolutional auto-encoders for steganalysis of digital images. In: Proc. of the 2014 Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. (APSIPA). IEEE, 2014. 1−4.

[82] Pearlmutter B. Gradient descent: Second order momentum and saturating error. In: Proc. of the Advances in Neural Information Processing Systems. 1992. 887−894.

[83] Zeiler MD. Adadelta: An adaptive learning rate method. arXiv preprint arXiv:1212.5701, 2012.

[84] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015.

[85] Zhang R, Zhu F, Liu J, Liu G. Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis. IEEE Trans. on Information Forensics and Security, 2019,15:1138−1150.

[86] He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015,37(9):1904−1916.

[87] Baluja S. Hiding images in plain sight: Deep steganography. In: Advances in Neural Information Processing Systems. 2017. 2069−2079.

[88] Chollet F. Xception: Deep learning with depthwise separable convolutions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 1251−1258.

[89] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016.

[90] Wei LX, Gao PX, Liu J, Liu MM. Image steganalysis based on convolution neural network. Application Research of Computers, 2019,36(1):235−238 (in Chinese with English abstract).

[91] Xu G. Deep convolutional neural network to detect J-UNIWARD. In: Proc. of the 5th ACM Workshop on Information Hiding and Multimedia Security. 2017. 67−73.

[92] Sedighi V, Fridrich J. Histogram layer, moving convolutional neural networks towards feature-based steganalysis. Electronic Imaging, 2017,2017(7):50−55.

[93] Pibre L, Pasquet J, Ienco D, Chaumont M. Deep learning is a good steganalysis tool when embedding key is reused for different images, even if there is a cover sourcemismatch. Electronic Imaging, 2016,2016(8):1−11.

[94] Chaumont M. Deep learning in steganography and steganalysis from 2015 to 2018. arXiv preprint arXiv:1904.01444, 2019.

[95] Zheng GH, Feng GR, Yu J, Cheng H, Zhang XP. JPEG steganalysis based on LSBD etection and enhanced features. Journal of Applied Sciences, 2016,34(6):670−676 (in Chinese with English abstract).

[96] Xu G, Wu HZ, Shi YQ. Ensemble of CNNs for steganalysis: An empirical study. In: Proc. of the 4th ACM Workshop on Information Hiding and Multimedia Security. 2016. 103−107.

[97] Zeng J, Tan S, Liu G, Li B, Huang J. WISERNet: Wider separate-then-reunion network for steganalysis of color images. IEEE Trans. on Information Forensics and Security, 2019,14(10):2735−2748.

[98] Yang J, Shi YQ, Wong EK, Kang XG. JPEG steganalysis based on densenet. arXiv preprint arXiv:1711.09335, 2017.

[99] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[100] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proc. of the 2017 IEEE Symp. on Security and Privacy (SP). IEEE, 2017. 39−57.

[101] Su J, Vargas DV, Sakurai K. One pixel attack for fooling deep neural networks. IEEE Trans. on Evolutionary Computation, 2019, 23(5):828−841.

[102] Zhang Y, Zhang W, Chen K, Liu J, Liu Y, Yu N. Adversarial examples against deep neural network based steganalysis. In: Proc. of the 6th ACM Workshop on Information Hiding and Multimedia Security. 2018. 67−72.

[103] Ma S, Zhao X, Liu Y. Adaptive spatial steganography based on adversarial examples. Multimedia Tools and Applications, 2019, 78(22):32503−32522.

[104] Tang W, Li B, Tan S, Barni M, Huang J. CNN-based adversarial embedding for image steganography. IEEE Trans. on Information Forensics and Security, 2019,14(8):2074−2087.

[105] Bernard S, Pevný T, Bas P, Klein J. Exploiting adversarial embeddings for better steganography. In: Proc. of the ACM Workshop on Information Hiding and Multimedia Security. 2019. 216−221.

[106] Kouider S, Chaumont M, Puech W. Adaptive steganography by oracle (ASO). In: Proc. of the 2013 IEEE Int'l Conf. on Multimedia and Expo (ICME). IEEE, 2013. 1−6.

[107] Schöttle P, Schlögl A, Pasquini C, Bohme R. Detecting adversarial examples—A lesson from multimedia security. In: Proc. of the 2018 26th European Signal Processing Conf. (EUSIPCO). IEEE, 2018. 947−951.

[108] Liu J, Zhang W, Zhang Y, Hou D. Detection based defense against adversarial examples from the steganalysis point of view. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 4825−4834.

[109] Hussain I, Zeng J, Tan S. A survey on deep convolutional neural networks for image steganography and steganalysis. KSII Trans. on Internet and Information Systems, 2020,14(3):1228−1248.

[110] Wang YJ, Niu K, Yang XY. Information hiding scheme based on generative adversarial network. Journal of Computer Application, 2018,38(10):2923−2928 (in Chinese with English abstract).

[111] Cogranne R, Giboulot Q, Bas P. The alaskasteganalysis challenge: A first step towards steganalysis. In: Proc. of the ACM Workshop on Information Hiding and Multimedia Security. 2019. 125−137.

附中文参考文献:

[1] 项世军,罗欣荣.同态公钥加密系统的图像可逆信息隐藏算法.软件学报,2016,27(6):1592−1601. http://www.jos.org.cn/1000-9825/5007.htm [doi: 10.13328/j.cnki.jos.005007]

[4] 李彦峰,丁丽萍,吴敬征,崔强,刘雪花,关贝,王永吉.网络隐蔽信道关键技术研究综述.软件学报,2019,30(8):2470−2490. http://www.jos.org.cn/1000-9825/5859.htm [doi: 10.13328/j.cnki.jos.005859]

[13] 翟黎明,嘉炬,任魏翔,等.深度学习在图像隐写术与隐写分析领域中的研究进展.信息安全学报,2018,3(6):2−12.

[36] 王朔中,张新鹏,张卫明.以数字图像为载体的隐写分析研究进展.计算机学报,2009,32(7):1247−1263.

[52] 张新鹏,钱振兴,李晟.信息隐藏研究进展.应用科学学报,2016,34(5):476−484.

[90] 魏立线,高培贤,刘佳,刘明明.基于卷积神经网络的图像隐写分析方法.计算机应用研究,2019,36(1):235−238.

[95] 郑国华,冯国瑞,余江,程航,张新鹏.基于 LSB 检测的 JPEG 隐写分析特征增强方法.应用科学学报,2016,34(6):670−676.

[110] 王耀杰,钮可,杨晓元.基于生成对抗网络的信息隐藏方案.计算机应用,2018,38(10):2923−2928.

陈君夫(1997−),男,硕士生,主要研究领域为信息安全,隐写术与隐写分析.

付章杰(1983−),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为信息隐藏,数据安全.

张卫明(1976−),男,博士,教授,博士生导师,主要研究领域为信息隐藏,媒体内容安全.

程旭(1983−),男,博士,副教授,CCF 专业会员,主要研究领域为计算机视觉,模式识别.

孙星明(1963−),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数字取证,人工智能安全.