软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(10):3874–3890 [doi: 10.13328/j.cnki.jos.006358]

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

基于少样本学习的通用隐写分析方法*

李大秋[1], 付章杰[1,2], 程 旭[1], 宋 晨[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)

通信作者: 付章杰, E-mail: fzj@nuist.edu.cn

摘 要: 近年来, 深度学习在图像隐写分析任务中表现出了优越的性能. 目前, 大多数基于深度学习的图像隐写分析模型为专用型隐写分析模型, 只适用于特定的某种隐写术. 使用专用隐写分析模型对其他隐写算法的隐写图像进行检测, 则需要该隐写算法的大量载密图像作为数据集对模型进行重新训练. 但在实际的通用隐写分析任务中, 隐写算法的大量载密图像数据集是难以得到的. 如何在极少隐写图像样本的情况下训练通用隐写分析模型是一个极大的挑战. 对此, 受少样本学习领域研究成果的启发, 提出了基于转导传播网络的通用隐写分析方法. 首先, 在已有的少样本学习分类框架上改进了特征提取部分, 设计了多尺度特征融合网络, 使少样本分类模型能够提取到更多的隐写分析特征, 使其可用于基于秘密噪声残差等弱信息的分类任务; 其次, 针对少样本隐写分析模型难收敛的问题, 提出了预训练初始化的方式得到具有先验知识的初始模型; 然后, 分别训练了频域和空域的少样本通用隐写分析模型, 通过自测和交叉测试, 结果表明, 检测平均准确率在 80%以上; 接着, 在此基础上, 采用数据集增强的方式重新训练了频域、空域少样本通用隐写分析模型, 使少样本通用隐写分析模型检测准确率与之前相比提高到 87%以上; 最后, 将得到的少样本通用隐写分析模型分别与现有的频域和空域隐写分析模型的检测性能进行比较, 结果显示, 空域上少样本通用隐写分析模型在常用的少样本环境下的检测准确率稍低于 SRNet 和 ZhuNet, 频域上少样本通用隐写分析模型在常见的少样本环境下的检测准确率已超越现有的频域隐写分析模型. 实验结果表明, 基于少样本学习的通用隐写分析方法对未知隐写算法的检测具有高效性和鲁棒性.

关键词: 隐写术; 隐写分析; 少样本学习; 深度学习

中图法分类号: TP306

中文引用格式: 李大秋, 付章杰, 程旭, 宋晨, 孙星明. 基于少样本学习的通用隐写分析方法. 软件学报, 2022, 33(10): 3874–3890. http://www.jos.org.cn/1000-9825/6358.htm

英文引用格式: Li DQ, Fu ZJ, Cheng X, Song C, Sun XM. Universal Steganalysis Based on Few-shot Learning. Ruan Jian Xue Bao/Journal of Software, 2022, 33(10): 3874–3890 (in Chinese). http://www.jos.org.cn/1000-9825/6358.htm

Universal Steganalysis Based on Few-shot Learning

LI Da-Qiu[1], FU Zhang-Jie[1,2], CHENG Xu[1], SONG Chen[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)

Abstract: In recent years, deep learning has shown excellent performance in image steganalysis. At present, most of the image steganalysis models based on deep learning are special steganalysis models, which are only applied to a specific steganography. To detect the stego images of other steganographic algorithms using the special steganalysis model, a large number of stego images encoded by the steganographic algorithms are regarded as datasets to retrain the model. However, in practical steganalysis tasks, it is difficult to obtain a large number of encoded stego images, and it is a great challenge to train the universal steganalysis model with very few stego image samples. Inspired by the research results in the field of few-shot learning, a universal steganalysis method is proposed based on

* 基金项目: 江苏省基础研究计划(自然科学基金)(BK20200039); 国家自然科学基金(U1836110, 61802058)
收稿时间: 2020-12-28; 修改时间: 2021-02-22; 采用时间: 2021-03-15; jos 在线出版时间: 2021-10-20

transductive propagation network. First, the feature extraction network is improved based on the existing few-shot learning classification framework, and the multi-scale feature fusion network is designed, so that the few-shot classification model can extract more steganalysis features for the classification task based on weak information such as secret noise residue. Second, to solve the problem that steganalysis model based on few-shot learning is difficult to converge, the initial model with prior knowledge is obtained by pre-training. Then, the steganalysis models based on few-shot learning in frequency domain and spatial domain are trained respectively. The results of self-test and cross-test show that the average detection accuracy is above 80%. Furthermore, the steganalysis models based on few-shot learning in frequency domain and spatial domain are retrained by means of dataset enhancement, so that the detection accuracy of the steganalysis models based on few-shot learning is improved to more than 87% compared with the previous steganalysis model based on few-shot learning. Finally, the proposed steganalysis model based on few-shot learning is compared with the existing steganalysis models in frequency domain and spatial domain, the result shows that the detection accuracy of the universal steganalysis model based on few-shot learning is slightly below those of SRNet and ZhuNet in spatial domain and is beyond that of existing best steganalysis model in frequency domain under the experimental setup of few-shot learning. The experimental results show that the proposed method based on few-shot learning is efficient and robust for the detection of unknown steganographic algorithms.

Key words: steganography; steganalysis; few-shot learning; deep learning

移动互联网和大数据技术的发展, 一方面给人们的生活、工作带来了便利, 另一方面也产生了诸多的安全问题[1]. 如何保障国家以及个人的信息安全, 成为越来越重要的问题. 为了解决这些潜在的安全隐患问题, 信息隐藏技术应运而生. 信息隐藏通过利用人类的感觉器官对数字信号的感觉冗余, 将秘密信息以较为隐蔽的方式嵌入到另一个公开载体中, 使秘密信息得到保护[2], 其在隐藏秘密信息内容的同时, 也隐藏了秘密信息的存在. 信息隐藏技术作为一种信息安全技术, 在当今时代已被许多领域所采用. 隐写术作为信息隐藏技术的重要方法, 保证了网络通信过程中数据的安全性. 隐写术是将秘密信息隐藏在图像等数字载体中的一门艺术和科学. 相应地, 为了防止隐写术的非法使用, 阻止秘密信息的非法传递, 隐写分析技术同时受到了人们的广泛关注[3]. 由于隐写术对载体图像进行秘密信息嵌入后会导致载体图像的统计特性发生改变, 隐写分析算法通过提取并分析图像的统计特性, 判断图像中是否嵌入秘密信息, 从而在一些可疑的数字图像中准确地找出载密图像[4]. 传统的隐写分析算法都是通过手工提取隐写分析特征, 然后将其送入分类器进行分类. 手工提取隐写分析特征十分依赖人工设计的滤波核, 这个过程耗时耗力且没有统一的滤波核设计标准, 不利于非专业人士的操作. 因此, 如何改进隐写分析特征提取, 成为研究者们研究的一个难题.

近年来, 基于深度学习的方法在计算机视觉、自然语言处理等各个研究领域取得了突出的成果[5]. 随着深度学习方法的广泛应用, 研究者将深度学习应用到图像隐写分析领域. 深度学习技术的优点在于特征提取的自动化、标准化过程, 这大大减弱了隐写分析方法对于手工设计滤波核的依赖性. 因此, 基于深度学习的隐写分析方法受到研究者的重点关注. 到目前为止, 已有大量研究者们提出了各种基于深度学习的隐写分析方法[6−17], 然而目前, 基于深度学习的隐写分析方法主要是针对特定的隐写算法所提出的, 利用特定隐写算法中存在的安全漏洞对其进行有针对性的检测, 其大多数偏向于提高模型的检测精度. 这些方法对于实际情况下的通用隐写分析任务而言还存在较大的问题, 其在一定程度上忽略了实际场景下的隐写分析检测条件. 在实际应用的场景下, 隐写分析器需要对未接触过的隐写术进行检测, 这就要求深度学习隐写分析模型具有通用检测能力. 在深度学习隐写分析模型训练的过程中, 一般需要隐写术大量的载密图像对模型进行训练. 在实际场景下, 由于要检测的隐写术是未知的, 这时要得到大量的载密图像作为数据集训练隐写分析模型是难以实现的. 考虑到深度学习隐写分析方法的实用性, 特别是在只有未知隐写术的少量含密图像样本的情况下, 基于少样本学习的通用隐写分析方法值得进一步加以研究.

已有的少样本分类的研究成果表明[18−25]: 可以在目标类样本数量极少的情况下, 通过少样本分类模型对目标类样本进行有效的区分. 少样本分类模型在训练过程中不需要接触目标类的大量标记样本, 这与实际场景的限制条件相符. 受这些研究工作和最近的一些深度学习隐写分析的启发, 本文提出了一种自适应隐写的少样本通用隐写分析模型. 据我们所知, 到目前为止, 少样本学习在图像隐写分析领域还没有相关的研究. 本文的方法借鉴了已有的少样本学习和隐写分析方法, 在少样本转导传播网络[25]的基础上, 主要通过高通滤波核、可分离卷积以及多尺度感受野卷积的结构改进了预处理和特征提取网络, 使少样本分类网络能够更好地

提取到载密图像的隐写特征, 进而使其能够应对隐写分析任务. 本文中展现了与已有频域、空域隐写分析算法[8,10,11,13,14,16,17,26]的性能对比, 比较结果表明了所提方法的有效性与鲁棒性. 与最近的深度学习隐写分析方法相比, 本文提出的方法并非仅针对特定的图像隐写术, 对于模型未知的图像隐写术不需要重新训练模型, 也能表现出较好的检测性能. 除此之外, 本文提出的少样本通用隐写分析方法适用于少样本环境下的隐写分析检测任务, 这对于隐写分析技术的实际应用具有重大意义. 本文的主要贡献在于:

(1) 提出了基于少样本学习的通用隐写分析算法, 在预处理和特征选择阶段, 改进了传统的少样本算法框架, 使其能够对自适应隐写算法进行通用隐写分析. 在实际应用中, 可在只有数张未知隐写算法载密图像的情况下, 也能对该隐写算法进行有效的检测;

(2) 针对少样本通用隐写分析模型难以拟合的问题, 提出了预训练方式以初始化少样本隐写分析模型, 使其初步具有少样本分类能力. 在此基础上, 使用隐写数据集进行微调, 从而使模型能够更好地拟合到秘密噪声的特征空间;

(3) 在已有的转导传播网络的基础上, 改进了特征提取模块, 设计了多尺度特征融合网络, 使少样本分类模型能够提取更多的隐写分析特征, 用于基于噪声残差等弱信息的分类任务, 并进一步使用增强数据集的方式训练隐写分析模型, 提高了初始模型的检测精度.

本文第 1 节简要介绍传统的通用隐写分析、深度学习隐写分析以及少样本学习领域的相关工作. 第 2 节介绍所提出的少样本通用隐写分析模型框架. 第 3 节介绍实验中使用的数据集及实验结果. 第 4 节主要分析少样本通用隐写分析模型进行预训练初始化的必要性以及通过对比实验验证本文提出的特征提取模块的有效性. 最后, 第 5 节给出总结和未来的工作展望.

1 相关工作

通用隐写分析技术发展至今, 可分为基于手工设计特征的通用隐写分析方法以及基于深度学习的隐写分析方法. 基于手工设计特征的通用隐写分析方法过于依赖手工设计的隐写分析特征提取滤波器. 基于深度学习的隐写分析方法大多只针对特定的一种或几种隐写术, 有较好的检测效果, 但不能很好地应对通用的隐写分析任务. 近几年提出的少样本学习分类方法可以用于少样本环境下的通用分类任务, 然而不能直接将少样本分类模型用于基于残差噪声等弱信息的隐写分析分类任务. 接下来将简要介绍传统的通用隐写分析、深度学习隐写分析、少样本学习领域的相关成果以及与本文提出的方法的关联性.

1.1 传统的通用隐写分析方法

到目前为止, 在传统领域已有一些通用的隐写分析方法被提出[27−32]. 在 2015 年, Farid 等人首次采用正交镜像滤波分解图像的方法, 得到多个高频子带系数的偏度、峰值、方差和平均值等特征向量, 最后分析得到的特征向量对图像进行分类[27,28], 该方法对常见隐写术有较好的检测效果. 到 2003 年, Avcibas 等人通过选择图像质量进行多变量分析的方法来检测图像中的水印[29], 首次提出了基于图像质量度量的通用隐写分析方法, 该方法的主要缺陷在于算法复杂度过高且缺乏合适的度量指标. 同年, Harmsen 等人发现隐写后的图像直方图特征函数的质心会发生改变, 基于此现象, 提出了检测图像直方图特征函数质心的通用隐写分析算法[30], 该方法的主要贡献在于发现了新的隐写分析特征. 于 2005 年, Fridrich 等人提出了基于 JPEG 图像的通用隐写分析算法, 该方法提取图像在预处理前和预处理后的变换域来统计特征向量, 接着将计算统计向量间的差值作为隐写分析特征, 最后通过得到的隐写分析特征来训练分类器[31]. 该方法的提出, 给接下来的研究起到了极大的借鉴作用. 同年, Shi 等人将提取图像的小波直方图特征函数统计矩作为隐写分析特征, 据此对图像进行隐写分析检测载密图像[32]. 经测试, 该方法对大多数隐写术都有较好的检测效果.

随着自适应隐写算法的不断发展, 传统的基于手工设计特征的隐写分析方法越来越具有局限性, 想要使得隐写分析器在性能上继续获得突破, 需要借助其他领域的新技术. 深度学习隐写分析器作为一种端到端的网络结构, 摒弃了传统的手工设计高通滤波器进行特征提取的方式, 而是通过构建由多层线性网络层和非线性激活单元组成的可学习模型, 使其在大量的载密图像中自动挖掘图像统计特性中的隐写分析特征, 从而从

数据中得到有效的特征表达, 极大地简化了隐写分析任务的复杂性, 进一步促进了自适应隐写分析领域的发展. 由于新兴的深度学习等机器学习技术在诸多领域取得了极大的成功, 越来越多的研究人员将深度学习技术应用到隐写分析领域, 并取得了较好的成果.

1.2 基于深度学习的隐写分析方法

早在 2014 年, Tan 等人提出使用自动编码器进行隐写分析模型预训练[6], 该模型可对 HUGO 隐写算法进行检测, 但其检测性能还不是很高. 到 2015 年, Qian 等人又提出了基于卷积神经网络的隐写分析检测模型[7], 在该模型中, 隐写分析特征通过卷积网络被自动地提取出来, 并通过最后的分类网络进行载密图像的检测. 之后, 在 2016 年, Xu 等人对 Qian 等人提出的隐写分析模型进行了改进, 提出了 XuNet 隐写分析检测模型[8]. XuNet 模型在预处理阶段加上了绝对值层(ABS)和批处理层(BN), 并利用了残差图像的对称性提取有效特征进行分类. 随后, Qian 等人在之前研究的基础上, 利用迁移学习的方式训练低嵌情况下的检测模型[9]. 他们的迁移学习方法是先用高嵌入率的载密图像对模型进行预训练, 然后在预训练好的模型上继续用低嵌入率的载密图像进行微调训练. 到 2017 年, Ye 等人提出了改进非线性激活函数的卷积网络来训练隐写分析器[13], 在 YeNet 模型中, 使用一组高通滤波核作为预处理模块, 并提出使用截断线性单元作为激活函数, 目的是更好地提取隐写特征. 在 2018 年, Fridrich 等人提出了基于卷积残差网络的隐写分析检测模型 SRNet[17]. SRNet 模型没有使用高通滤波核对输入图像进行预处理操作, 而是使用 He 等人提出的残差模型来构建卷积神经网络检测模型[33]. 该模型取得了比之前模型都要好的检测性能. 2019 年, Zhu 等人借鉴卷积神经网络中新提出来的可分离卷积、金字塔池化等新技术, 建立了新的深度学习隐写分析模型 ZhuNet[16]. 该模型仍然采用预处理模块, 并在模型中使用残差模块里的恒等连接结构, 最后取得了超越 SRNet 的检测性能. 图 1 展示了目前较流行的基于深度学习的自适应隐写分析框架, 从图 1 可以看出, 基于深度学习的隐写分析模型主要分为预处理和卷积神经网络提取特征并加以分类两个部分. 预处理部分是利用高通滤波核对输入图像进行滤波处理, 提取出图像内容特征中隐藏的隐写分析特征, 前人提出的大多数隐写分析模型都包含预处理部分. 大部分研究者关注的重点也多在于接下来的卷积神经网络的设计上, 因为不同规格的卷积及池化操作对隐写分析特征的提取效果也是不同的, 所以设计出一个较好的卷积神经网络对基于深度学习的隐写分析方法至关重要.

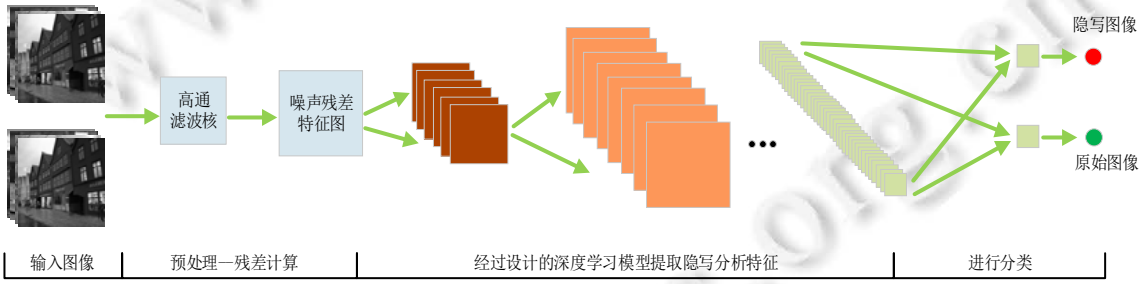


图 1 基于深度学习的自适应隐写分析框架

上述深度学习隐写分析方法的鲁棒性较差, 大多只针对特定的隐写术有较好的检测性能, 且不能应用于少样本的通用隐写分析任务场景. 与以上这些基于深度学习的隐写分析方法相比较, 本文提出的方法不仅可对模型未知的隐写算法进行有效的检测, 而且可在少样本环境下, 高效地完成通用隐写分析检测任务.

1.3 基于少样本学习的分类方法

为了使深度学习分类模型在少量样本上也能很好地泛化, Li 等人于 2006 年首次提出了少样本学习方法[18]. 目前, 大多数少样本学习都是基于表征学习、度量学习和元学习的方法, 主要包括匹配网络[19]、原型网络[20]、关系网络[21]、孪生网络[22]等. 最近, 少样本学习领域又出现了一些新的思路, 即用图网络来建模处理不同类样本之间的关系. 少样本学习在得到查询集样本的标签之前需要对查询集样本和支持集样本进行嵌入向量的距离度量, 目前, 所有的少样本学习方法本质上都是在探究如何去学习查询集样本与支持集样本之

间的关系. 而图网络是典型的建模学习节点间关系的网络结构, 所以用图网络建模少样本学习的分类任务是很合适的.

2018 年, Victor 等人首次将图网络用于少样本学习[23], 提出了少样本图神经网络分类模型. 该模型使用卷积神经网络卷积来提取查询集和支持集的样本特征, 并通过每个样本特征之间的欧式距离构建图邻接矩阵. 之后, 使用图卷积操作得到查询集样本标签的预测值, 最后与其真实标签计算损失, 进行反向传播, 更新模型参数. Victor 等人提出的少样本图神经网络分类模型性能一举超越之前所有的少样本学习方法, 达到了最好的分类效果. 2019 年, Kim 等人在 Victor 等人提出的模型基础上进行改进, 提出了基于边特征的少样本图网络分类模型. Kim 等人认为: Victor 等人的图网络模型构建图邻接矩阵时仅考虑到节点特征, 没有显式考虑类内的相似性和类间的差异性[24]. 实验结果表明, Kim 等人基于边特征的少样本图网络模型的性能超越了 Victor 等人的少样本图网络模型. 同年, Liu 等人提出了转导传播网络模型用于少样本分类任务. 转导传播网络首次把标签传播和图网络结合起来进行查询集的标签学习[25]. 转导传播网络中图网络的构建同样依赖于卷积神经网络对支持集和查询集进行特征提取, 并构建图邻接矩阵. 与之前不同的是: 计算特征间欧式距离的公式是改进的高斯相似函数, 在模型的训练过程中会通过子卷积网络去学习相似函数的长度比例参数. Liu 等人认为, 衡量样本特征间距离的相似函数对整个模型的性能也起到十分重要的作用, 所以需要对长度比例参数进行建模学习, 从而得到更好的相似函数来计算样本特征间的距离. 在初始化图结构之后, 使用标签传播公式直接得出查询集的类标签. 标签传播公式本质上是以图邻接矩阵中各样本之间的距离为依据, 把支持集样本类标签赋值给查询集样本. 同样, Liu 等人的转导传播网络模型的性能也超越了 Victor 等人的少样本图网络模型, 且从实验结果看, Liu 等人的方法比 Kim 等人的方法在 5-way 5-shot 少样本设置下的分类准确率提高了 3%.

已有的这些少样本学习方法只能应用于传统的少样本分类场景, 无法直接应用于隐写分析任务. 因为计算机视觉的图像分类任务和隐写分析任务有较大的区别. 这二者的关注点是不同的, 图像分类、物体检测等任务关注的是图像中的内容信息, 而隐写分析则正好相反. 图像内容中存在着人类视觉感知系统不敏感的冗余信息, 图像隐写算法就是利用这些冗余信息隐藏秘密信息, 图像隐写分析主要是获得图像中的秘密噪声残差等各种弱信息, 使分类模型能够拟合到秘密噪声的特征空间, 所以还需要在少样本分类框架上进一步加以改进, 使其能够适应隐写分析检测任务.

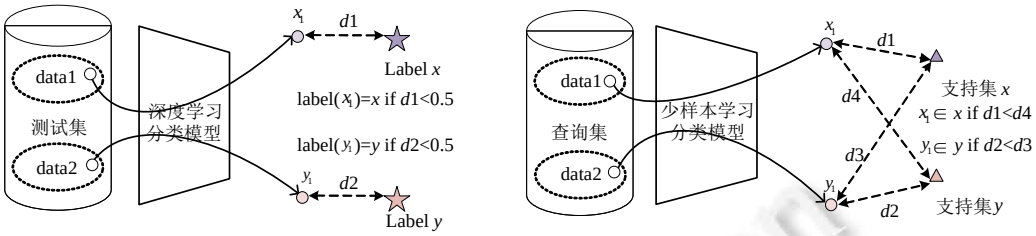
2 基于少样本学习的通用隐写分析方法

2.1 总体框架

少样本学习分类任务的主要挑战是, 如何从少量训练样本中识别新的类别. 与一般深度学习方法的最大区别在于: 在少样本学习分类任务中, 目标类别的样本数量极其有限. 在这种情况下, 当用少量的训练样本去训练一般的深度卷积网络分类模型时, 不足以使其拟合到类特征表征空间. 这是因为, 训练一个性能较好的深度卷积神经网络需要大量的带标记样本数据. 为了解决这个问题, 少样本学习框架在分类时把需要分类的样本作为查询集, 把目标类的少量标记样本作为支持集, 支持集作为类分布空间中的类中心, 用于辅助查询集以得到正确的类标签. 在训练时, 使用大量非目标类的训练数据模拟目标类的支持集和查询集进行模型的训练. 少样本分类模型从这些非目标类的数据中学习足够多的先验知识, 最后使模型得到强大的度量能力, 从而能够根据支持集和查询集样本嵌入向量的欧式距离来确定查询集向量是否与支持集向量是同类样本.

图 2 比较了深度学习二分类模型(如图 2(a)所示)和少样本学习二分类模型(如图 2(b)所示)的测试过程. 对于一般的深度学习二分类模型, 测试样本 x_1 输入模型后会得到属于某一类的概率值 P, 把测试样本 x_1 的嵌入向量与类标签的相对距离记为 $d1$, 令 $d1=1-P$. 若 P 值大于 0.5, 即 $d1<0.5$, 则样本 x_1 属于类 x. 而在少样本学习二分类模型中, 查询集样本 x_1 输入模型中并不直接得到其属于某一类的概率, 而是会使样本 x_1 的嵌入向量分别与支持集中的标记样本 x 和 y 的嵌入向量作距离度量. 若查询集样本 x_1 与支持集样本 x 的距离 $d1$ 小于其与支持集样本 y 的距离 $d4$, 则样本 x_1 归属于支持集 x 所属的类, 从而使样本 x_1 的类标签与样本 x 保持一致. 从图 2 所示的测试过程对比中可以看出, 少样本学习分类模型在分类时会在支持集样本的辅助下显式度量类间

距离作为分类依据. 这也是少样本学习分类模型能够适用于少样本环境下图像分类的关键因素. 同样, 这对于少样本环境下的隐写分析二分类任务也具有同样的作用. 因此, 本文在已有的少样本转导传播网络[25]的基础上进行改进, 提出了少样本通用隐写分析网络.



(a) 深度学习分类模型 (b) 少样本学习分类模型

图 2 深度学习分类模型和少样本学习分类模型的测试过程

如图 3 所示: 本文提出的少样本通用隐写分析网络框架主要由隐写分析特征提取、图构造、标签传播、损失计算这 4 部分组成, 其中, 图构造、标签传播及损失计算等模块的设计受少样本转导传播网络[25]启发. 在少样本转导传播网络框架的基础上, 主要改进了其特征嵌入部分. 改造后的特征嵌入部分由预处理、多尺度特征融合和特征嵌入子网络组成, 这 3 个子结构一起组成输入图像的隐写分析特征提取模块.

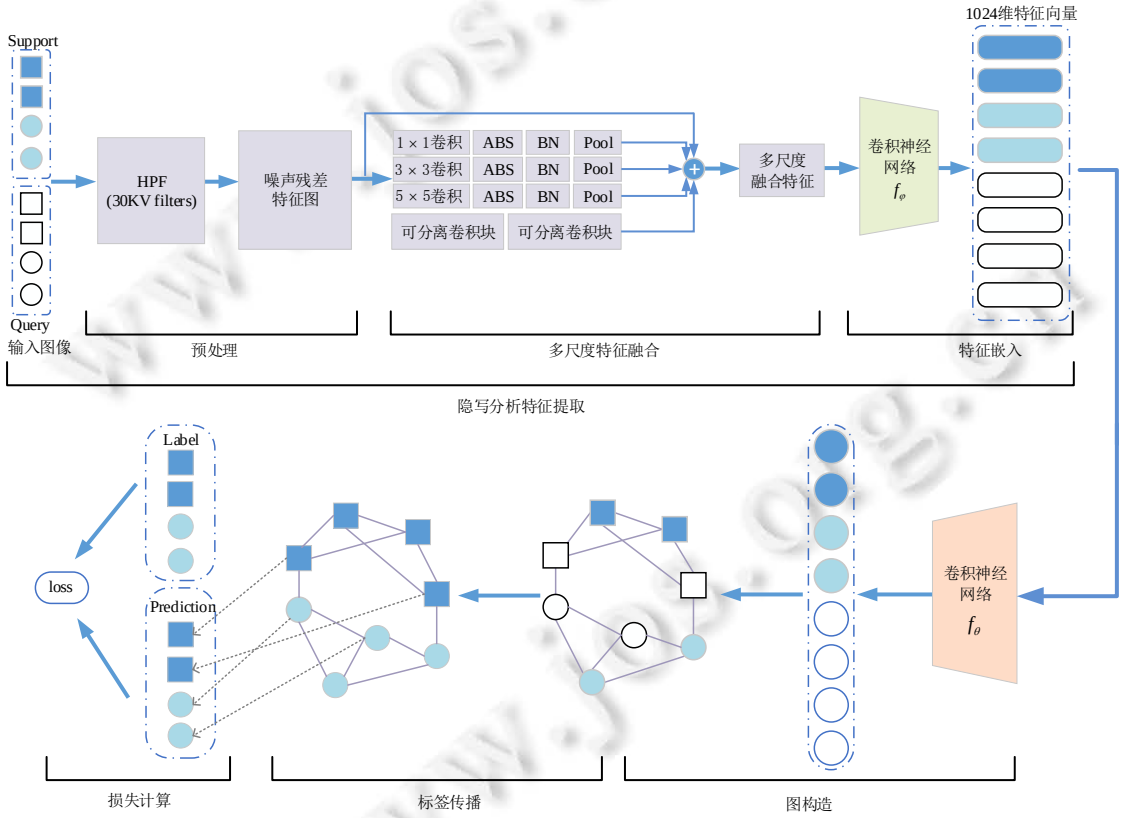


图 3 少样本通用隐写分析模型整体框架

2.2 少样本隐写分析特征提取

在深度学习隐写分析领域中, 将图像隐写分析预处理层加入到深度学习模型中最早由 Qian 等人[7]提出, 该层的设计源于传统的隐写分析方法. 当图像经过隐写后, 相当于在载体图像中加入一些十分微弱的隐写噪

声信号. 为了减少图像的内容特征对检测隐写噪声带来的干扰, 预处理层对深度学习隐写分析模型而言显得非常重要. 如果不经过预处理操作直接把数字图像输入到网络模型中训练隐写分析器, 通常网络会难以收敛. 所以在少样本隐写分析模型的特征提取部分, 同样设计了预处理层提取隐写分析特征. 隐写分析特征提取部分的主要任务就是要把图像的隐写分析统计特征最大程度地提取出来, 从而为接下来用少样本学习分类网络进行特征分类提供依据. 在预处理部分, 使用目前常用的高通滤波核[26]作为残差特征图提取器, 可在很大程度上保留高频隐写噪声, 并抑制图像内容的干扰. 与 YeNet[13]和 ZhuNet[16]类似, 为了尽可能多地提取到秘密噪声特征信息, 预处理层使用了 30 层高通滤波核. 模型对输入图像进行预处理操作后, 可以得到对应的噪声残差特征图, 如图 4 中的预处理部分所示.

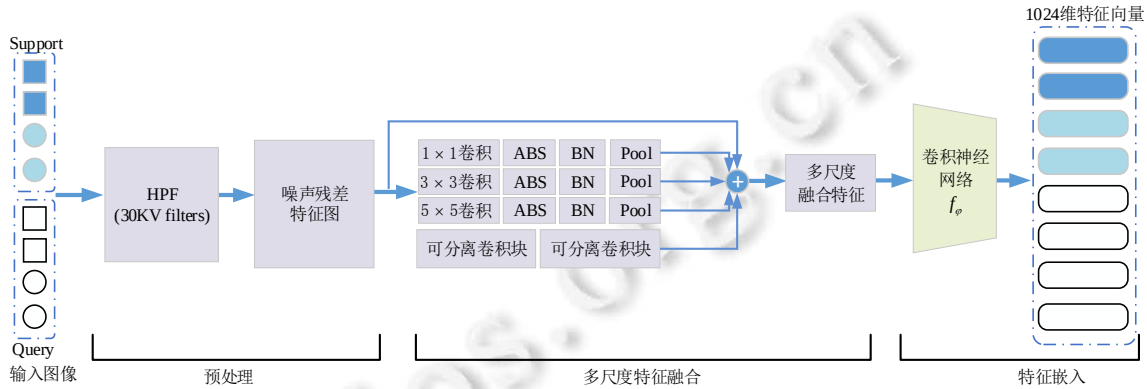


图 4 少样本通用隐写分析特征提取网络

得到噪声残差图后, 考虑的就应该是如何在这些残差特征图中得到最有效的隐写分析特征, 从而更好地使载密图像和载体图像的残差特征在各自的特征空间进行特征嵌入. 已有研究者验证了跨层连接可以有效缓解深层网络模型训练困难的问题[33]. 同时, 跨层连接结构也在一定程度上减少了层级之间信息传递过程的损失. 图像隐写分析利用输入图像的统计特性变化来判定图像是否载密, 因此, 层间信息传递时的损失会加大检测的困难性. 对数字图像隐写分析任务而言, 减少训练过程中模型的信息损失具有重要意义. 另外, 使用多种尺度的卷积核对噪声残差图进行特征提取有利于不同尺度的隐写噪声特征的提取[34]. 所以这里使用了 1×1、3×3、5×5 的卷积核同时对噪声残差图进行卷积. 可分离卷积由 Chollet 等人提出[35], 可分离卷积可在不同通道上对图像信息进行提取, 有利于对多通道特征的最大程度地提取聚合. 最后, 如图 4 中多尺度特征融合子网络所示, 将多尺度卷积特征、可分离卷积特征以及跨层的噪声残差特征进行多尺度特征融合.

隐写分析特征提取部分的最后一步是对前面提到的多尺度融合特征进行特征嵌入操作. 特征嵌入网络是由 4 个卷积块组成的子卷积神经网络, 如图 5 所示.

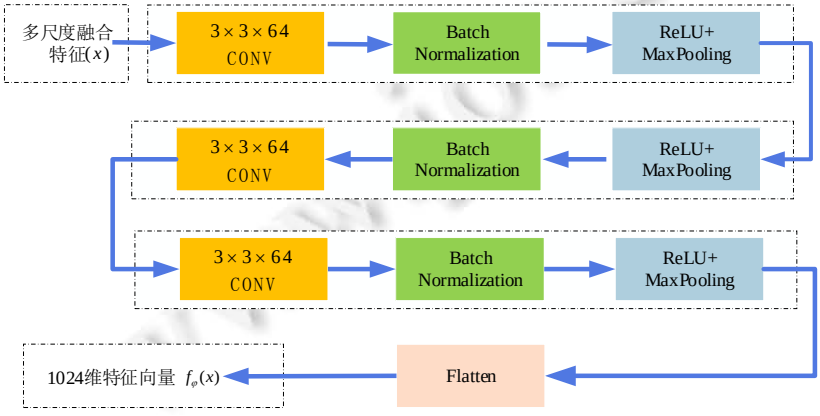


图 5 少样本通用隐写分析特征嵌入子网络

在特征嵌入子卷积神经网络中, 每个卷积块由 3×3 的卷积核、批处理归一化[36]及最大池化操作组成, 通过图 5 所示的特征嵌入子网络, 可将载密图像、载体图像的多尺度融合特征进行各自类空间中的嵌入操作, 最终得到每个图像样本各自类空间中的 1 024 维特征向量.

2.3 少样本图网络构造

将隐写分析特征提取部分得到的 1 024 维特征向量作为图构造网络的输入, 从而得到对图构造至关重要的长度比例参数σ. 图网络构造的核心在于得到表述所有节点之间距离的邻接矩阵. 如图 4 所示: 假设每次输入模型的样本包括支持集 4 张图像、查询集 4 张图像, 一共 8 张图像样本. 这 8 张图像样本即为 8 个节点, 这 8 个节点通过隐写分析特征提取模块得到 8 个不同的特征向量. 图网络构造部分首先要计算这 8 个特征向量两两之间的高斯距离, 如公式(1), 再通过聚合每两个节点之间的高斯距离建立起这 8 个节点的邻接矩阵:

$$W_{ij}=\exp\left(-\frac{1}{2}d\left(\frac{f_\varphi(x_i)}{\sigma_i},\frac{f_\varphi(x_j)}{\sigma_j}\right)\right) \tag{1}$$

公式(1)中的参数σ对高斯距离公式起到一个调控的作用, 如何调节该参数, 使高斯距离公式更适合于这 8 个节点之间的距离衡量是很重要的. 如图 6 中的图构造子网络就是为了对σ参数建模训练出一个较好的拟合值, 从而通过图邻接矩阵 W 构造一个较好的初始图, 这将有利于下一步的标签传播过程. 在得到初始的图邻接矩阵 W 后, 对该邻接矩阵做拉普拉斯正则化, 即 $S=D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$, 其中, D 表示邻接矩阵 W 的度矩阵. 对计算节点间距离的公式进行建模, 将有利于在标签传播时减小支持集和查询集节点之间的类内距离以及增大支持集和查询集节点之间的类间距离.

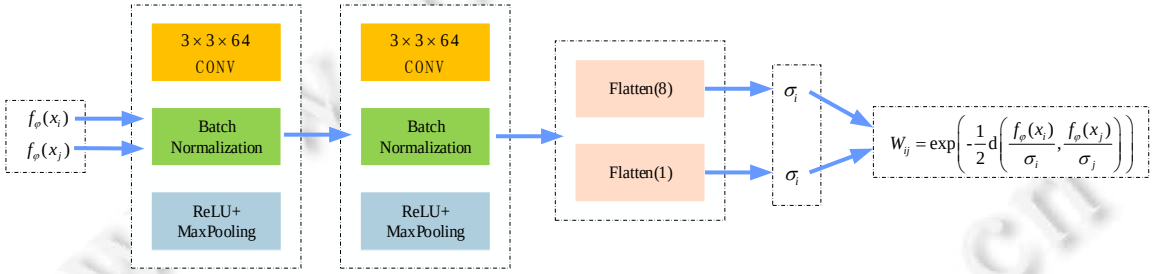


图 6 少样本通用隐写分析图构造子网络

2.4 标签传播与损失计算

在得到初始图邻接矩阵后, 就以该邻接矩阵中各节点之间距离度量为依据, 把支持集各节点的标签通过已有的标签传播公式传递给查询集各节点. 该标签传播公式可见公式(2):

$$F=(I-\alpha S)^{-1}Y \tag{2}$$

其中, F 为传播后的标签矩阵; S 为正则化处理过的邻接矩阵; I 为单位矩阵; α为标签传播参数, 控制着传播的信息总量且$\alpha\in(0,1)$; Y 为初始标签矩阵. 初始标签矩阵 Y 和 F 是一个$(N\times K+T)\times N$ 的矩阵, N 是类别的数量, K 是每个类支持集的样本数量, T 是每个类查询集的样本数量. 在标签矩阵 Y 和 F 中, 行数表示支持集和查询集中所有的样本数量. 支持集每个样本的行值即为该样本标签的 one-hot 编码, 而查询集的行值是网络的输出值, 初始化为 0. 少样本隐写分析网络模型经前向传播, 通过标签传播公式后, 查询集节点的标签就会得到更新. 为了方便计算更新后查询集样本的预测标签与其真实标签之间的损失值, 使用 softmax 将更新后的标签矩阵 F 进行概率值的转换, 如公式(3):

$$F(y_i=j\mid i)=\frac{\exp(F_{ij}^*)}{\sum_{j=1}^{N}\exp(F_{ij}^*)} \tag{3}$$

公式(3)中 y_i 是查询集样本的预测标签, F_{ij}^* 是由公式(2)计算的传播后的标签矩阵.

将预测标签与真实标签进行交叉熵的计算, 从而得到查询集样本节点的分类损失, 如公式(4):

$$J = -\sum_{j=1}^{T} y_j^* \log(P(y_i = j \mid i)) \tag{4}$$

公式(4)中的 y_j^* 是查询集样本的真实标签且 $y_j^* \subset \{0,1\}$. 最后, 可使用梯度下降法对模型进行端到端的参数更新, 直至模型收敛.

3 实验与结果

3.1 参数设置

在实验中, 少样本通用隐写分析模型使用由 Kingma 等人提出的 Adam 优化器[37], 模型的初始学习率为 10^{-3}, 标签传播参数α设置为 0.99. 载体图像和载密图像的 Batchsize 各为 20 个样本, 即每类支持集 5 个样本, 查询集 15 个样本. 在少样本学习中, 称每个 Batchsize 为一个情景. 二分类且每类支持集为 5 个样本的实验设置为 2-way 5-shot, 二分类且每类支持集为 1 个样本的实验设置为 2-way 1-shot. 在本文的实验过程中, 分别做了 2-way 5-shot 和 2-way 1-shot 的实验. 每次训练时会把载体图像和载密图像都随机打乱, 然后, 在每个情景训练时, 每类选取的 20 个样本按顺序遍历, 每次按顺序取 20 个样本, 直到整个训练集遍历结束. 实验平台的配置为 Windows 10 操作系统, GPU 为 NVIDIA 2080Ti 显卡, 内存为 16 G RAM, CPU 为 Intel Core (TM) i5-7500 处理器.

3.2 数据集与评价指标

表 1 描述了实验所使用的数据集的具体信息. 实验中, 针对空域隐写算法, 选取了 Bossbase 1.01 自然图像库作为载体图像类. Bossbase 图像数据集包含了 10 000 张自然图像, 它是隐写术和隐写分析领域最常用的基准数据集, 专门用于隐写及隐写分析领域的标准数据集之一. 然后用 HUGO、SUNIWARD 和 MIPOD 这 3 种空域自适应隐写算法对 Bossbase 数据集进行隐写操作, 隐写嵌入率为 0.4 bpp. 得到对应的 3 种自适应隐写术的载密图像. 训练集的载体和载密图像各为 5 000 张, 测试集的每类也是各 5 000 张图像. 针对频域隐写算法, 选取了 ALASKA2 彩色图像库作为载体图像. ALASKA2 彩色图像数据集来自最近流行的 ALASKA2 图像隐写分析挑战赛, ALASKA2 彩色图像数据集中带有 75 000 张载体彩色图像, 选取其中 10 000 张载体彩色图像, 使用 JMiPOD、JUNIWARD、UERD 这 3 种频域隐写算法进行相应的图像频域隐写操作, 嵌入率为 0.4 bpnzAC, JEPG 图像的质量因子为 95. 设置训练集的载密图像和载体图像各为 5 000 张彩色图像. 同样, 测试集的载体图像和载密图像也各为 5 000 张. 另外, 为了对少样本隐写分析模型进行初始化, 选取 MiniImageNet 图像数据集来进行模型的预训练. MiniImageNet 数据集是 ImageNet 数据集的一个子集, 是专门用于少样本图像识别任务的标准数据集之一[38]. 它由从 ImageNet 中随机选取的 100 个类组成, 每个类包含 600 个样本. 另外, 当在实验中将图像输入到少样本隐写分析模型训练时, 会将图像大小统一调整为 256×256(像素).

表 1 实验所用数据集描述

数据集	数据量(张)	位深度	图像类型	图片大小(像素)	格式
Bossbase 1.01	10 000	8	灰度图	256×256	PNG
ALASKA2	75 000	24	彩图	512×512	JPEG
MiniImageNet	60 000	24	彩图	256×256	PKL

一般用于评价隐写分析模型的性能指标有误检率(ERR)或者准确率(ACC): 误检率体现了隐写分析模型在隐写分析检测任务中的出错概率; 相反, 准确率体现的是隐写分析模型正确区分载体图像和载密图像的数学概率. 本文统一使用检测准确率作为少样本隐写分析模型的性能评价指标. 由于隐写分析实际上是二分类任务, 故可将载体图像类作为阳性类, 将载密图像类作为阴性类. 正确预测的阳性样本和阴性样本数量之和占所有阳性样本和阴性样本数量总和之比即为检测准确率, 如公式(5)所示:

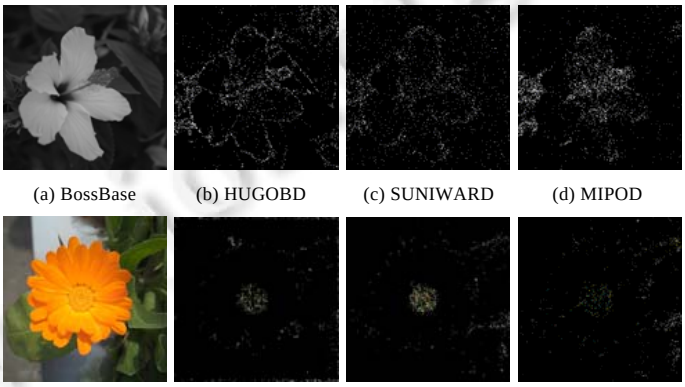
$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

其中, TP 代表预测的真阳样本数量, TN 代表预测的真阴样本数量, FP 代表预测的假阳样本数量, FN 代表预测

的假阴样本数量.

3.3 实验结果

为了分析通用隐写分析任务的可行性, 本文选取了 6 种具有代表性的空域和频域自适应隐写算法 HUGOBD[39]、SUNIWARD[40]、MIPOD[41]、UERD[42]、JUNIWARD[40]及 JMIPOD[43]进行实验分析. 用 HUGOBD、SUNIWARD 和 MIPOD 这 3 种空域隐写术分别对原始载体 BossBase 图像库中任意选取的图像进行信息嵌入, 嵌入率为 0.4 bpp. 最后得到不同的隐写术隐写的载密图像. 对这 3 种空域隐写算法隐写的载密图像分别与对应的载体图像进行 SUB 操作, 得到 3 幅残差图像. 在 BossBase 图像库中任意选取的图像如图 7(a)所示, 这 3 种空域隐写术所对应的载密图像和原始载体图像之间的残差图像如图 7(b) –图 7(d)所示. 类似地, 用 UERD、JUNIWARD 及 JMIPOD 这 3 种频域隐写术分别对原始载体 ALASKA2 图像库中任意选取的彩色图像在 DCT 域进行信息嵌入, 嵌入率为 0.4 bpnzAC, JEPG 图像的质量因子为 95. 最后, 同样得到不同的频域隐写术对应的载密图像. 对这 3 种频域隐写算法隐写的载密图像分别与对应的载体图像进行 SUB 操作, 得到 3 幅残差图像. 在 ALASKA2 彩色图像库中任意选取的彩色图像如图 7(e)所示, 这 3 个频域隐写术所对应的载密图像和原始载体图像之间的残差图像如图 7(f)–图 7(h)所示.



(a) BossBase (b) HUGOBD (c) SUNIWARD (d) MIPOD

(e) ALASKA2 (f) UERD (g) JUNIWARD (h) JMIPOD

图 7 BossBase 及 Alaska2 图像库中任意选取的两幅载体图像

由图 7 所知: 虽然隐写算法不同, 但是这些自适应隐写术在嵌入秘密信息时对图像的改变总是发生在近似相同的区域, 即图像的边缘及纹理较复杂的区域. 事实上, 为了减小对图像统计特性的改变, 现有的自适应隐写术都将秘密信息尽可能地嵌入到图像边缘或纹理比较复杂的区域. 类似空间域自适应隐写术, 频域隐写算法同样如此. 虽然频域隐写算法大多是通过修改量化后的 DCT 系数来嵌入秘密信息, 但是某个 DCT 系数被修改后也会将这种影响扩散至整个空域的对应区域中. 频域隐写算法也会给原始载体图像带来空域上的统计特性的改变. 频域隐写术的目的也是使修改后的载密图像总失真最小, 隐写时, 不同区域载密能力不同, 通过不同区域所提取的特征有效性也不一致, 这与在空域把隐秘信息隐藏在纹理复杂区域的特点是一致的. 故无论是频域还是空域隐写术, 其隐秘信息修改位置一般都集中在某一区域. 所以在理论上, 通过这个特性是可以训练出通用型隐写分析器去检测的未知隐写算法. 本文所提出的少样本隐写分析算法实际上也是使少样本隐写分析模型去学习这些自适应隐写术的隐写共性特征, 从而可以去完成通用型的少样本隐写分析任务.

在分别用频域和空域隐写数据集对少样本通用隐写分析模型进行训练之前, 首先对模型参数进行了初始化. 初始化方式是使用 MiniImageNet 数据集对少样本通用隐写分析模型进行预训练. 除了将每个情景中的类别参数由 5 改为 2, 预训练参数设置与转导传播网络[25]中的基本一致. 预训练过程的准确率和损失值的变化曲线如图 8 所示. 如图 8 所示: 当使用少样本通用隐写分析模型用于传统的少样本分类任务(2-way 5-shot)时, 模型在 10 个 epochs 的训练中就能得到收敛. 至此已经得到通过预训练初始化后的少样本隐写分析模型, 该模型已经具备初步的少样本分类能力. 在此基础上, 继续进行频域和空域少样本通用隐写分析模型的训练.

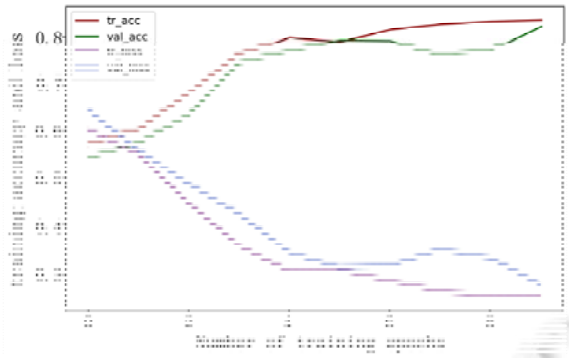
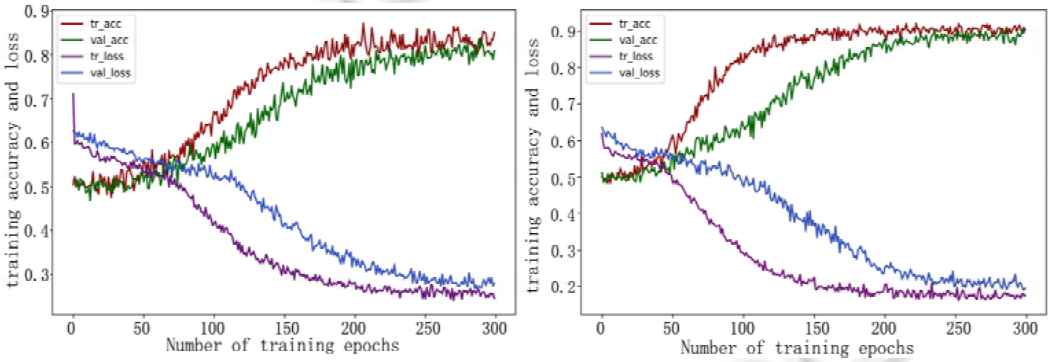


图 8　少样本隐写分析模型预训练过程

对于频域少样本通用隐写分析模型，分别使用 JMIPOD、JUNIWORD、UERD 这 3 种隐写数据集进行训练，训练后得到 3 个不同的隐写分析模型. 类似地，对于空域少样本通用隐写分析模型，分别使用 MIPOD、SUNIWARD、HUGO 这 3 种隐写数据集进行训练，训练后也得到 3 个不同的空域少样本隐写分析模型. 如图 9(a)、图 9(b)所示，分别展现了频域和空域隐写分析模型的训练结果. 图 9(a)、图 9(b)训练的结果是基于 JMiPOD 和 MiPOD 隐写数据集训练的. 通过对比可以发现，本文提出的少样本隐写分析模型对于空域隐写算法更具优势. 因为在训练的第 50 个轮次之前，空域少样本隐写分析模型就已经在加速收敛，且在第 300 个训练轮次左右时，损失值和检测准确率已接近平稳；而频域少样本隐写分析模型是在训练的第 50 个轮次后才加速收敛，且在第 300 个训练轮次左右，损失曲线和准确率曲线还存在较大程度的波动现象.



(a) 频域　　(b) 空域

图 9　少样本隐写分析模型训练过程

接下来在测试过程中，分别使用 3 个少样本隐写分析模型进行自测和交叉测试. 自测和交叉测试的测试结果可见表 2 和表 3.

表 2　频域隐写分析自测和交叉测试结果

隐写分析模型	自适应隐写术	2-way 5-shot (ACC)	2-way 1-shot (ACC)
Model-JMIPOD	JMIPOD	0.86	0.53
	JUNIWARD	0.81	0.51
	UERD	0.80	0.52
Model-JUNIWARD	JMIPOD	0.82	0.51
	JUNIWARD	0.84	0.53
	UERD	0.83	0.52
Model-UERD	JMIPOD	0.79	0.50
	JUNIWARD	0.83	0.51
	UERD	0.84	0.52

表 3 空域隐写分析自测和交叉测试结果

隐写分析模型	自适应隐写术	2-way 5-shot (ACC)	2-way 1-shot (ACC)
Model-JMIPOD	MIPOD	0.89	0.51
	SUNIWARD	0.87	0.50
	HUGO	0.88	0.50
Model-JUNIWARD	MIPOD	0.89	0.50
	SUNIWARD	0.89	0.50
	HUGO	0.87	0.50
Model-UERD	MIPOD	0.90	0.50
	SUNIWARD	0.89	0.51
	HUGO	0.91	0.51

由表 2 和表 3 可知, 在 2-way 5-shot 实验设置的少样本隐写模型的自测和交叉测试的平均准确率分别为 0.847 和 0.813. 总体的准确率都能达到 0.80 以上, 这在少样本的分类任务中已经是一个较好的结果. 本文补充了常见的 2-way 1-shot 少样本实验设置下的实验, 2-way 1-shot 是在支持集样本只有一个已知标签的样本的情况下的实验设置. 测试结果在表 2、表 3 的最后一列中. 这里, 对于原始载体和含密图像的分类准确率总体能达到 0.50 以上.

为了进一步提高少样本隐写分析模型的通用检测精度, 使用数据集增强的方式重新训练了频域和空域的少样本通用隐写分析模型. 通过之前的实验结果, 在频域和空域数据集中挑选了少样本隐写分析模型在交叉测试中表现较好的两个隐写数据集. 然后, 通过数据集拼接的方式增强数据集潜在的通用隐写分析特征空间. 拼接的方式是, 将挑选出来的数据集载密图像各取 2 500 张图像组成新的载密图像集. 最终在频域数据集中挑选了 UERD 和 JUNIEARD 这两种载密图像, 在空域数据集中挑选了 MIPOD 和 HUGO 这两种隐写书隐写的载密图像. 创建好新的增强数据集后, 重新训练了频域和空域上的少样本的通用隐写分析模型, 新的少样本通用隐写分析模型的测试结果见表 4. 由表 4 可见, 增强后的少样本隐写分析模型无论是对频域隐写术的检测性能, 还是对空域隐写术的检测性能都有了一定的提高. 特别是对于 2-way 1-shot 的实验设置, 模型的检测性能得到了明显的提高.

表 4 数据集增强后的频域、空域少样本隐写分析模型测试结果

隐写分析模型	自适应隐写术	2-way 5-shot (ACC)	2-way 1-shot (ACC)
Model (UERD+JUNIWARD)	JMIPOD	0.87	0.61
	JUNIWARD	0.88	0.59
	UERD	0.91	0.60
Model (MIPOD+HUGO)	MIPOD	0.89	0.58
	SUNIWARD	0.87	0.57
	HUGO	0.90	0.56

最后, 本文将得到的少样本通用隐写分析模型分别与现有的频域和空域的通用隐写分析模型进行了比较, 见表 5 和表 6.

表 5 与现有的空域隐写分析模型检测性能对比的结果

隐写分析模型	空域隐写术	检测准确率(ACC)
SRM[26]	WOW	0.791
	SUNIWARD	0.795
XuNet[8]	WOW	0.793
	SUNIWARD	0.728
YeNet[13]	WOW	0.768
	SUNIWARD	0.688
YedroudjNet[14]	WOW	0.842
	SUNIWARD	0.829
ZhuNet[16]	WOW	0.935
	SUNIWARD	0.919
SRNet[17]	WOW	0.901
	SUNIWARD	0.877
OurModel	WOW	**0.885** (2-way 5-shot)
	SUNIWARD	**0.872** (2-way 5-shot)

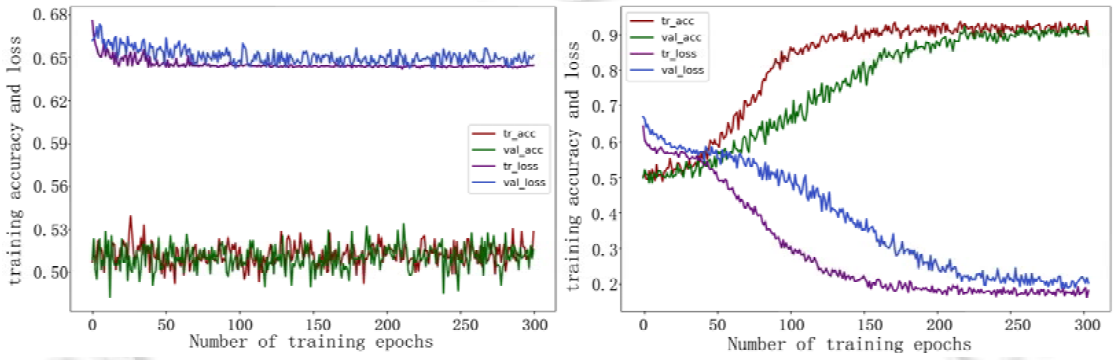
表 6 与现有的频域隐写分析模型检测性能对比的结果

隐写分析模型	频域隐写术	检测准确率(ACC)
XuNet-JPEG[10]	JUNIWARD	0.802
PNet[11]	JUNIWARD	0.757
SRNet[17]	JUNIWARD	0.824
SCA-SRNet[17]	JUNIWARD	**0.866**
OurModel	JUNIWARD	**0.880** (2-way 5-shot)

从表 5 的对比结果可见: 与目前通用的空域隐写模型相比较, 本文提出的少样本方法在 2-way 5-shot 实验设置下, 除了 ZhuNet 和 SRNet, 其他模型的检测精度都逊色于少样本隐写分析模型. 由表 6 可知: 与目前通用的频域隐写分析算法相比, 本文提出的频域隐写分析方法在 2-way 5-shot 实验设置下比目前的方法都更加优越.

4 讨 论

前文提到: 在训练少样本隐写分析模型时, 首先要使用 MiniImageNet 数据集对模型进行预训练, 使模型参数得到初始化. 对于预训练过程的必要性, 本文设置了如图 10 所示的对比实验. 图 10(a)所示为不进行预训练, 直接用隐写数据集训练少样本隐写分析模型的训练结果; 图 10(b)所示为通过预训练后再用隐写数据集进行微调的模型训练结果.



(a) 不进行预训练技巧 (b) 通过预训练

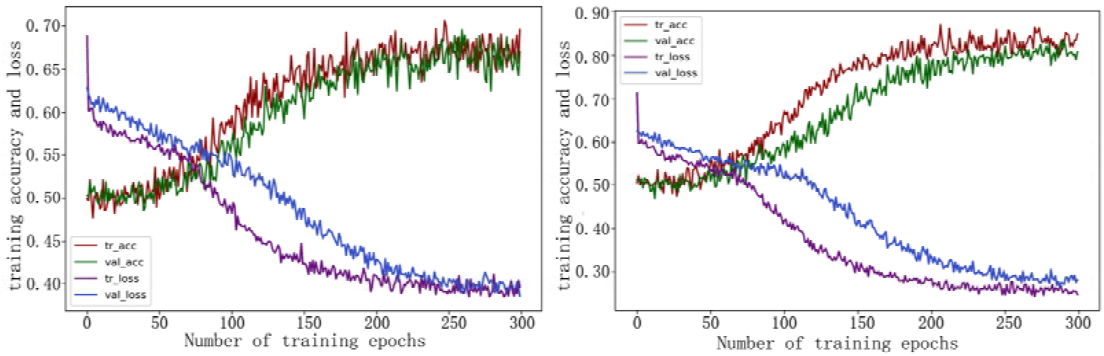
图 10 不进行预训练技巧和通过预训练的模型训练过程的对比

从图 10(a)可以看出: 在不进行预训练操作时, 训练少样本隐写分析模型会遇到模型不收敛的问题, 即训练时模型的准确率总在 0.5 上下波动, 模型的损失值下降到 0.60 附近会趋近于平稳浮动, 不能如图 10(b)那样, 通过平缓期后快速下降到 0.2 附近. 从图 10 的对比结果中可以明显地看出: 若少样本隐写分析模型不经过预训练过程, 则少样本模型将不能应对隐写分析任务. 这是因为少样本隐写分析模型初始化后初步获得进行图像分类的能力, 这类似于人类在学习深入细致的专业知识前, 首先要获得基础的先验知识. 少样本隐写分析模型在初步拟合少样本图像分类任务之后, 对于接下来的隐写分析任务才能更好地进行优化. 如果一开始就让少样本隐写分析模型直接进行隐写分析任务的训练, 就会导致模型的优化器容易面临不动点问题, 即模型的训练损失曲线保持在较高的水平, 使优化过程变得异常困难. 对于一般的深度学习图像分类任务或者原始的少样本分类任务而言, 模型的优化器是不太容易遇到不动点问题的, 这是由隐写分析任务与一般的图像分类任务的差别所决定的.

因为隐写分析任务实质上也是分类任务, 不过它是对载体图像和载密图像进行分类的二分类任务. 载密图像通过在载体图像中加入的微弱秘密噪声产生, 这导致模型分类的依据从图像内容的特征空间转移到秘密噪声的特征空间. 一方面, 秘密噪声的特征是微弱的, 隐写分析模型就是要尽可能地提取更多的秘密噪声特征, 并使分类器拟合到该秘密噪声的特征空间; 另一方面, 载体图像和载密图像一般是肉眼不可察觉的, 载

密图像中隐藏的秘密特征远远少于基于图像本身的内容特征, 这就导致分类器在优化过程中不可避免地受到图像内容特征的干扰. 所以, 本文提出的少样本隐写分析模型在少样本的情境中训练隐写分析分类器的难度比单独训练隐写分析器或训练少样本分类器都要大得多. 经实验验证, 使用预训练初始化的方式确实对少样本隐写分析模型的训练起到关键作用.

为了验证模型中提出的特征提取模块的有效性, 本文设计了如图 11 所示的对比实验, 图 11(a)所示为去除特征提取模块后的少样本隐写分析模型的训练结果, 图 11(b)所示为带有特征提取模块的模型训练结果.



(a) 去除特征提取模块 (b) 带有特征提取模块

图 11 去除特征提取模块和带有特征提取模块的模型训练过程对比

从图 10(a)可以看出: 在去除对尺度特征提取模块后, 训练时模型的准确率最后收敛于 0.7 以下, 模型的损失值下降到 0.40 附近会趋近于平稳地浮动, 而不能像图 11(b)所示那样, 通过训练结束后模型准确率收敛于 0.8 以上, 模型损失快速下降到 0.30 以下. 另外, 从图 11(a)、图 11(b)的对比结果可以明显地看出: 去除特征提取模块后, 模型的准确率曲线和损失曲线的交叉点相较于完整的模型有一定的后移, 这说明去除特征提取模块后模型的收敛速度减缓, 其训练难度加大. 经过如图 11 所示的实验验证, 本文提出的特征提取模块对于少样本环境下的隐写分析任务是有效的.

5 总结与展望

本文首次提出了结合少样本学习的通用隐写分析算法. 在只有数张未知隐写算法隐写的载密图像的情况下, 少样本通用隐写分析方法也能对该隐写算法及其他隐写算法进行有效的检测. 本文提出的方法在已有的少样本分类模型的基础上, 主要改进了特征提取模块, 结合残差连接、深度可分离卷积和多重感受野卷积组成多尺度特征融合模块, 使少样本分类模型提取更多的隐写分析特征用于基于噪声残差等弱信息的分类任务. 另外, 针对少样本隐写分析模型在训练时难以拟合的问题, 采用了预训练的方法对模型进行初始化. 这避免了少样本通用隐写分析模型在进行少样本隐写分析任务训练时难以优化的问题. 本文提出的少样本通用隐写分析方法可以有效地应用于实际的隐写分析通用检测任务, 这将有利于维护社会稳定, 保障国家和个人的信息安全.

本文提出的方法是将少样本学习应用于隐写分析任务的首次尝试, 实验结果表明, 本文方法在单样本环境下表现的检测性能还不够高, 在今后的研究中, 将进一步提高少样本通用隐写分析方法在单样本情景下的检测性能.

另外, 接下来, 我们将继续研究少样本隐写分析模型的跨域检测能力, 即只需通过频域(空域)隐写数据集训练出少样本隐写分析模型, 使其可针对空域(频域)的隐写术进行有效的检测. 目前, 少样本隐写分析模型在跨域检测任务上的效果还比较差, 后续将考虑在模型的预处理阶段提取多域特征进行特征融合, 使其最终能够拟合到多域隐写特征的类空间, 使少样本隐写分析模型具备跨域检测能力. 这将大大提高隐写分析技术的

实用性, 且对于隐写分析技术的实际应用将具有重要的指导意义.

References:

[1] Feng DG, Zhang M, Li H. Big data security and privacy protection. Chinese Joural of Computers, 2014, 37(1): 246−258 (in Chinese with English abstract).

[2] 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]

[3] 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).

[4] Liu J, Ke Y, Lei Y. Recent advances of image steganography with generative adversarial networks. arXiv preprint arXiv: 1907. 01886, 2019.

[5] Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks, 2015, 61: 85−117.

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

[7] Qian Y, Dong J, Wang W, Tan T. Deep learning for steganalysis via convolutional neural networks. In: Proc. of the Media Watermarking, Security, and Forensics 2015. Int'l Society for Optics and Photonics, 2015. 9409: 94090J.

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

[9] Qian Y, Dong J, Wang W. Learning and transferring representations for image steganalysis using convolutional neural network. In: Proc. of the 2016 IEEE Int'l Conf. on Image Processing (ICIP). Phoenix: IEEE, 2016. 2752−2756.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] Li FF, Rob F, Pietro P. One-shot learning of object categories. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2006, 28(4): 594−611.

[19] Vinyals O, Blundell C, Lillicrap T. Matching networks for one shot learning. In: Advances in Neural Information Processing Systems. 2016. 3630−3638.

[20] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. 2017. 4077−4087.

[21] Sung F, Yang Y, Zhang L. Learning to compare: Relation network for few-shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 1199−1208.

[22] Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition. In: Proc. of the ICML Deep Learning Workshop. 2015. 2.

[23] Satorras VG, Estrach JB. Few-shot learning with graph neural networks. In: Proc. of the Int'l Conf. on Learning Representations. 2018.

[24] Kim J, Kim T, Kim S. Edge-labeling graph neural network for few-shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 11−20.

[25] Liu Y, Lee J, Park M. Learning to propagate labels: Transductive propagation network for few-shot learning. In: Proc. of the Int'l Conf. on Learning Representations. 2019.

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

[27] Farid H. Detecting steganographic messages in digital images. Technical Report, TR2001-412, Dartmouth College, 2001.

[28] Farid H. Detecting hidden messages using higher-order statistical models. In: Proc. of the IEEE Int'l Conf. on Image Processing, Vol.2. 2002. 905−908.

[29] Ismail A, Nasir M, Bulent S. Steganalysis using image quality metrics. IEEE Trans. on Image Processing, 2003, 12(2): 221−229.

[30] Harmsen JJ, Pearlman WA. Steganalysis of additive noise modelable information hiding. In: Proc. of the SPIE, Security, Steganography, and Watermarking of Multimedia Contents V, Vol.5020. 2003. 131−142.

[31] Fridrich J. Feature-based steganalysis for JPEG images and its implications for future design of steganographic schemes. In: Proc. of the 6th Information Hiding Workshop. Toronto: Springer-Verlag, 2005. 67−81.

[32] Shi YQ, Xuan GR, Zou DK. Steganalysis based on moments of characteristic functions using wavelet decomposition. In: Proc. of the IEEE Prediction-error Image, and Neural Network (ICME 2005). Amsterdam, 2005. 268−270.

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

[34] Szegedy C, Liu W, Jia Y. Going deeper with convolutions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 1−9.

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

[36] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proc. of the Int'l Conf. on Machine Learning. 2015. 448−456.

[37] Diederik PK, Jimmy B. Adam: A method for stochastic optimization. In: Proc. of the Int'l Conf. on Learning Representations (ICLR), Vol.5. 2015.

[38] Alex K, Ilya S, Geoffrey EH. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. 2012. 1097−1105.

[39] Filler T, Fridrich J. Gibbs construction in steganography. IEEE Trans. on Information Forensics and Security, 2010, 5(4): 705−720. [doi: 10.1109/TIFS.2010.2077629]

[40] Holub V, Fridrich J, Denemark T. Universal distortion function for steganography in an arbitrary domain. EURASIP Journal on Information Security, 2014. [doi: 10.1186/1687-417X-2014-1]

[41] Sedighi V, Cogranne R, Fridrich J. Content-adaptive steganography by minimizing statistical detectability. IEEE Trans. on Information Forensics and Security, 2016, 11 (2): 221−234. [doi: 10.1109/TIFS.2015.2486744]

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

[43] Cogranne R, Giboulot Q, Bas P. Steganography by minimizing statistical detectability: The cases of JPEG and color images. In: Proc. of the ACM Information Hiding and MultiMedia Security (IH & MMSec). 2020. 161−167.

附中文参考文献:

[1] 冯登国, 张敏, 李昊. 大数据安全与隐私保护. 计算机学报, 2014, 37(1): 246−258.

[2] 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]

[3] 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.

李大秋(1996－), 男, 硕士, 主要研究领域为信息安全, 隐写术与隐写分析, 深度学习.

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

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.

宋晨(1997－), 男, 硕士, 主要研究领域为计算机视觉, 目标检测, 卷积神经网络.

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