基于多示例学习图卷积网络的隐写者检测

钟圣华 [1] 张 智 [1, 2]

摘 要 隐写者检测通过设计模型检测在批量图像中嵌入秘密信息进行隐蔽通信的隐写者, 对解决非法使用隐写术的问题具有重要意义. 本文提出一种基于多示例学习图卷积网络 (Multiple-instance learning graph convolutional network, MILGCN) 的隐写者检测算法, 将隐写者检测形式化为多示例学习 (Multiple-instance learning, MIL) 任务. 本文中设计的共性增强图卷积网络 (Graph convolutional network, GCN) 和注意力图读出模块能够自适应地突出示例包中正示例的模式特征, 构建有区分度的示例包表征并进行隐写者检测. 实验表明, 本文设计的模型能够对抗多种批量隐写术和与之对应的策略.

关键词 图像隐写者检测, 图卷积网络, 多示例学习, 示例包表征

引用格式 钟圣华, 张智. 基于多示例学习图卷积网络的隐写者检测. 自动化学报, 2024, **50**(4): 771−789

DOI 10.16383/j.aas.c220775

Steganographer Detection via Multiple-instance Learning Graph Convolutional Networks

ZHONG Sheng-Hua[1] ZHANG Zhi[1, 2]

Abstract Steganographer detection aims to solve the problem of illegal use of batch steganography by designing models to detect steganographers who embed secret information in images for covert communication. This paper proposes a novel steganographer detection algorithm called as multiple-instance learning graph convolutional network (MILGCN) to formalize steganography detection as a multiple-instance learning (MIL) task. The commonness enhancement graph convolutional network (GCN) and attention graph readout module designed in this paper can adaptively highlight the positive instance pattern and construct distinguishable bag representations for steganographer detection. Experiments show that the designed model can successfully attack a variety of batch steganography and the corresponding strategies.

Key words Image steganograher detection, graph convolutional network (GCN), multiple-instance learning (MIL), bag of instances representation

Citation Zhong Sheng-Hua, Zhang Zhi. Steganographer detection via multiple-instance learning graph convolutional networks. *Acta Automatica Sinica*, 2024, **50**(4): 771−789

图像隐写者检测技术是一项通过对用户传播的图像进行综合分析、侦测, 来发现那些试图将隐秘信息隐藏在图片中进行隐秘通信的隐写者的信息安全技术. 在真实的社交网络中, 隐写者检测十分困难. 一方面, 基于图像的隐写算法可以帮助隐写者在不改变图像外观的前提下, 将隐秘信息嵌入图像中. 另一方面, 隐写者使用隐写术和有效载荷[1] 等相关参数往往是无法预知的, 这进一步增加了隐写者检测的难度. 与试图捕获载密图像和载体图像之间的差异的隐写分析方法不同, 隐写者检测更关注隐写者与非隐写者之间的差异.

现有的隐写者检测方法中, 为了模拟真实场景中隐写者数量远远少于正常用户的情况, 常常假设在测试的过程中, 用户中只有一个隐写者存在, 采用异常检测或排序的方法将预测的隐写者概率最高的用户作为隐写者进行输出.

因此, 通用的隐写者检测方法通常由两部分组成: 特征提取和基于特征的聚类或离群值检测. Ker 等[1] 首次将隐写者检测转换为聚类问题进行研究, 从每张图像中提取 PEV-274 特征[2], 并使用最大平均差异计算每对用户之间的距离, 再通过层次聚类算法来区分隐写者与非隐写者. 此后, Ker 等[3−4] 进一步改进之前的工作, 用局部离群值因子方法代替层次聚类算法, 计算用户的异常程度并进行排序, 异常值最高的用户被检测为隐写者. 2016 年,

收稿日期 2022-09-30 录用日期 2023-04-12
Manuscript received September 30, 2022; accepted April 12, 2023
广东省自然科学基金 (2023A1515012685, 2023A1515011296), 国家自然科学基金 (62002230, 62032015) 资助
Supported by Natural Science Foundation of Guangdong Province (2023A1515012685, 2023A1515011296) and National Natural Science Foundation of China (62002230, 62032015)
本文责任编委 赫然
Recommended by Associate Editor HE Ran
1. 深圳大学计算机与软件学院 深圳 518060 2. 香港理工大学电子计算学系 香港 999077
1. College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060 2. Department of Computing, The Hong Kong Polytechnic University, Hong Kong 999077

Li 等[5] 提出使用高阶联合特征作为图像的隐写分析特征, 并集成多个层次聚类器来检测隐写者. 2017 年, Zheng 等[6] 首次提出一种基于深度学习方法的隐写者检测框架, 使用深度残差网络来提取图像特征, 最后使用聚合性层次聚类算法识别隐写者. 2018 年, Zheng 等[7] 进一步改进特征提取模型, 并提出一种用于隐写者检测任务的多分类深度神经网络, 与传统方法和其他深度学习方法相比, 该模型在标准数据集上实现了最好的性能. 尽管这些方法的特征提取部分有所不同, 但是在用户表征、用户之间相似性的度量和可疑用户的检测等步骤没有本质差异. 在这些方法中, 每名用户的表征由其所分享的所有图像的特征分布拼接而成, 在此基础上, 计算用户的特征分布之间的相似度, 找出与其他用户差异较大的用户, 进而确定隐写者. 2020 年以来, Zhang 等[8−9] 将用户分享的图像及其相关关系建模成图, 提出相似性累积图卷积单元, 能够增强相似特征分布, 从而发现载密图像构成的子图, 对其进行加权, 以获得更有效的用户表征, 这也是迄今为止唯一使用图神经网络模型进行隐写者检测的方法.

本文将隐写者检测形式化成多示例学习 (Multiple-instance learning, MIL) 任务, 并提出基于多示例学习图卷积网络的隐写者检测算法 (Steganographer detection algorithm based on multiple-instance learning graph convolutional network, MILGCN). 该算法通过共性增强图卷积网络 (Graph convolutional network, GCN) 有效增加正示例的共性特征, 通过注意力示例包表征模块自适应地构建更具有区分力的示例包表征, 并设计多示例学习损失约束. 与现有算法相比, 提升了空域和频域、已知和未知隐写术等多种隐写策略情况下的隐写者检测准确率. 相比于 Zhang 等[8−9] 的工作, 本文提出一种新的基于图的用户表征模型, 能够针对不同嵌入策略做到对分享的图像数量鲁棒. 相比于文献 [8−9] 中基于规则构建的图重建和边池化方法, 本文提出自适应的图构建和归一化方法, 并通过损失进行约束, 自适应地攻击不同隐写策略; 相比于文献 [8−9] 中将节点视为同等重要的展平读出和平均读出, 本文进一步设计新的图读出方式, 能够载密图像构建具有区分力的图表征.

本文内容安排如下: 第 1 节回顾图神经网络的相关工作; 第 2 节给出基于多示例学习的通用隐写者检测方法的详细介绍; 第 3 节给出一系列实验, 以验证提出方法的有效性; 第 4 节对全文工作进行总结, 并给出进一步的研究思路.

1 图神经网络的相关工作

最近, 基于深度学习的方法已经在图结构数据的分类和聚类任务上获得了成功的应用, 实际应用领域包括恶意软件分析、图像分类、动作识别、物体分类等[10]. 其中, 图卷积网络通过将适用于分析欧氏数据结构的传统卷积神经网络泛化到图等非欧结构数据上, 实现了图结构上的卷积运算, 取得了显著的研究进展并获得了广泛的关注. 在这些方法中, 图卷积方法通常被分为两类, 即基于谱域的图卷积方法和基于空域的图卷积方法[9].

谱域图卷积方法源于谱图理论, 可以看作傅立叶变换在图上的推广[11]. 该类方法通过对拉普拉斯矩阵的特征分解, 定义了图上的拉普拉斯变换和拉普拉斯逆变换. 该工作通过傅立叶变换将图变换到谱域进行卷积, 再通过傅立叶逆变换将卷积结果变换回图. 在此基础上, Defferrard 等[11] 进一步提出一种卷积核的多项式近似方法, 将节点聚合信息的范围限制在 k 阶邻居节点内. 在此基础上, Kipf 和 Welling[12] 将节点聚合信息的范围限制在一阶邻域内, 再次对卷积计算进行近似, 提出最为常用的图卷积网络.

与基于谱域的方法不同, 基于空域的方法并未将图映射到傅立叶域, 而是将卷积操作形式化为一种“块级操作”, 基于这种块级操作, 卷积操作通过聚合块级区域 (图上的每个节点及其邻域) 的信息构建新的节点特征表示. 2017 年, Hamilton 等[13] 提出 GraphSAGE 并将图结构数据上的卷积操作形式化为三个主要步骤, 利用对节点排列具有不变性的函数 (如均值、和、极大值等) 聚合节点的邻域信息. 2018 年, Ying 等[14] 提出一个可微分的图池化网络 DiffPool, 来生成层次化的图的表征. 通过可微分的方式, 模型能够自适应地对图中的节点进行池化, 从而得到新的表征. 最近, 注意力机制在图深度学习领域取得了突出的成就, 这类方法使用注意力机制为重要的节点、邻居节点和特征赋予更高的权重. 2018 年, Veličković 等[15] 提出图注意力网络 (Graph attention network, GAT), 在该模型中, 注意力机制被用于计算在聚合来自邻居节点的信息时不同邻居节点的权重值. 近年来的研究表明, 基于谱域的图卷积方法和基于空域的图卷积方法并不是完全对立的, 一些谱域的方法能够形式化为在空域应用某种卷积核进行卷积.

2 基于多示例学习图卷积网络的隐写者检测框架

多示例学习是有监督学习中的一种特殊形式.

相比于对一系列独立标注的样本进行分类, 在多示例学习任务中, 学习器以一系列被标注的包 (Bag) 作为输入, 每个包包含若干个未标注的样本, 即示例 (Instance)[16−19]. 如果包中的所有示例都为负样本, 那么这个包则被标注为负包. 另一方面, 如果包中含有至少一个正示例, 则包被标记为正包. 在隐写者检测任务中, 只要用户分享的图像中包含至少一张载密图像, 则这个用户就应当被检测为隐写者. 只有当用户分享的所有图像都为载体图像时, 用户应当被判别为正常用户. 毋庸置疑, 隐写者检测任务的目标和多示例学习在本质上是一致的. 因此, 本文将隐写者检测任务形式化为多示例学习任务.

本文提出的基于多示例学习图卷积网络的隐写者检测算法, 将社交网络中的用户及其分享的图像作为输入, 区分隐写者和正常用户. 如图 1 所示, 在多示例学习的框架下包含 4 个主要组成部分, 即特征提取网络、共性增强图卷积网络、注意力示例包表征网络和分类网络. 需要注意的是, 本文提出的模型在训练和测试过程中有所不同. 在训练过程中, 以示例包的分类作为目标, 所以构建的是分类网络. 在测试过程中, 采用与以往隐写者检测一致的实验设置, 即将预测的隐写者概率最高的用户作为隐写者输出.

首先使用基于多类别深度神经网络的隐写者检测 (Multiclass deep neural networks based steganographer detection, MDNNSD)[7], 针对用户分享的每一张图像提取特征, 得到用户对应的特征集合 $\{x_1, \cdots, x_n\}$. 这里, 将提取到的特征表示为橙色圆角矩形, 其中第 i 个圆角矩形表示用户分享的第 i 张图像所对应的特征. 接着, 将特征作为示例包中的示例, 将特征集合作为示例包, 构建图结构. 其中, 紫色圆柱表示示例所对应的特征, 圆柱的高度表示特征的维度, 紫色圆柱之间的连线代表图结构中的连接示例特征的边. 在图结构上进行卷积, 实现示例特征的分析和降维, 降维的过程使用箭头指示. 在多次卷积后, 可以得到分析和降维后的特征集合 $\{t_1, \cdots, t_n\}$. 这里, 将得到的示例特征按顺序排列, 绘制为新的紫色圆柱. 在此基础上, 使用注意力图读出计算每个示例特征的重要程度, 并根据重要程度对示例特征进行加权, 加权后的结果使用不同深浅的紫色圆柱表示. 最终, 对加权后的示例特征进行汇总, 得到图的表征, 并使用箭头指示得到的示例包表征 u_x. 之后, 将示例包表征输入由蓝色示意图表示的多层感知器中, 得到隐写者和正常用户. 在表 1 中, 给出了模型中使用的主要变量的介绍.

2.1　基于多分类扩张残差网络进行示例表征

本文致力于设计新的隐写者检测方法, 而对单张图像的表征, 本文直接使用目前最前沿隐写者检测中的特征提取方法, 即 MDNNSD[7], 来从每张分享图像中提取特征向量. 具体来说, 在 Zheng 等[7] 的工作中, 使用包含特征提取器 $\mu : M_{\text{cha}\times h\times w}(\mathbf{R}) \to M_{1\times d}(\mathbf{R})$ 和分类器 $\pi : M_{1\times d}(\mathbf{R}) \to M_{1\times \text{cls}}(\mathbf{R})$ 的多分类扩张残差网络. 其中, cha, h, w 分别表示输入多分类扩张残差网络的图像的通道数、高度和宽度, d 表示 MDNNSD 提取的特征维度数, cls 表示 MDNNSD 预测得到的样本属于某类的概率中的类别数. 这里, 本文使用 MDNNSD 中的特征提取器 μ 对用户 x 分享的第 i 张图像 I_i 进行特征提取, 得到用户 x 分享的第 i 张图像 I_i 的特征 $x_i = \mu(I_i)$. 对于分享了 n 张图像的用户 x, 本文使用用户分享的所有图像对应的特征向量 $\{x_1, \cdots, x_n\}$ 共同构成用户 x 所对应的示例包的表征.

2.2　使用共性增强图卷积网络进行示例分析

隐写者嵌入秘密信息的过程中有多种批量隐写

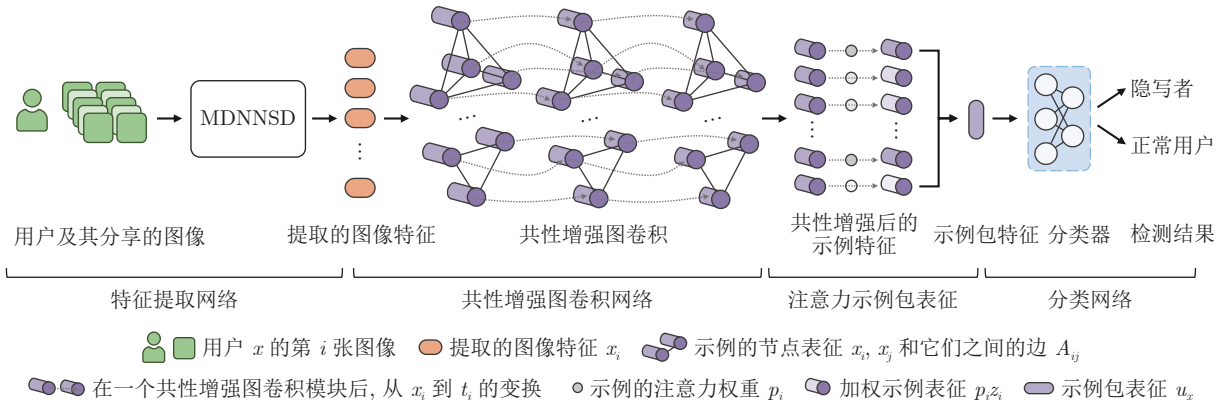


图 1　基于多示例学习图卷积网络的隐写者检测框架

Fig. 1　Steganographer detection framework based on multiple-instance learning graph convolutional network

表 1 使用的变量符号及对应说明
Table 1 The variable symbols and their corresponding descriptions

变量	含义
B_x	用户 x 对应的示例包
x_i	示例包 B_x 内第 i 个示例
m	当前数据集中示例包的总数量
n	当前示例包中示例的总数量
v_i	共性增强图卷积模块的第 i 个输入示例特征
h_i	对 v_i 使用 f 函数进行特征提取后得到的示例特征向量
H	由示例特征向量构成的示例包的矩阵表示 $H=[h_1, \cdots, h_n]^{\mathrm{T}}$
A_{ij}	图中第 i 个与第 j 个示例节点之间边的权重
N_i	图中第 i 个示例节点的所有邻居节点
A	示例包 B_x 所构成图的邻接矩阵
r_i	进行图卷积后 h_i 所对应的示例特征向量
R	由示例特征向量构成的示例包的矩阵表示 $R=[r_1, \cdots, r_n]^{\mathrm{T}}$
s_i	进行图归一化后 r_i 所对应的示例特征向量
S	由示例特征向量构成的示例包的矩阵表示 $S=[s_1, \cdots, s_n]^{\mathrm{T}}$
t_i	共性增强图卷积模块的第 i 个输出示例特征
T	由示例特征向量构成的示例包的矩阵表示 $T=[t_1, \cdots, t_n]^{\mathrm{T}}$
f	特征提取函数
g	注意力计算函数
z_i	共性增强图卷积模块的输出, 也是注意力读出模块的第 i 个输入示例特征
Z_x	由共性增强图卷积模块得到的示例特征向量构成的用户 x 对应的示例包的矩阵表示
u_x	用户 x 对应的示例包的特征向量表征
p_i	当前示例包中第 i 个示例对示例包表征的贡献
ρ_i	第 i 个示例包的预测结果
Y_i	第 i 个示例包的真实标签
L	本文设计的损失函数
L_{bag}	本文设计的多示例学习分类损失
L_{entropy}	本文设计的熵正则损失
$L_{\mathrm{contrastive}}$	本文设计的对比学习损失
λ_1, λ_2, λ_3	超参数, 用于调整 L_{bag}, L_{entropy}, $L_{\mathrm{contrastive}}$ 的权重

策略. 对于固定的有效载荷, 隐写者可能将有效载荷分散在多数分享图像中, 以降低每张载密图像的嵌入信息量, 从而降低被发现的可能; 也可能挑选少数分享图像, 集中嵌入秘密信息, 通过大量的载体图像遮掩载密图像的存在. 然而, 现有的隐写者检测工作通常将用户分享的每张图像视为从相同分布中独立采样的个体, 用户分享的所有图像的特征分布作为用户的特征表征, 从而计算用户之间的差异并将异常用户作为隐写者. 而当载密图像占比较小或是载密图像包含的嵌入信息较少时, 隐写者的表征将与正常用户的表征极为相近, 无法有效检测出隐写者.

为了解决该问题, 依托于本文对隐写者检测的多示例学习的形式化, 提出共性增强图卷积网络进行示例分析, 利用批量图像间的相关关系增强用户表征中的正示例的模式特征. 相关工作已经表明, 在多示例学习任务中, 示例包中的示例并不是独立存在的, 每个示例都和包中的其他示例有千丝万缕的联系. 对示例间的依赖关系进行建模, 有利于提取并增强相互关联的正示例的共性特征, 凸显出其与负示例共性特征的差异. 在隐写者检测任务中, 示例 (用户所分享图像) 相关关系无疑也是客观存在的, 例如, 隐写者传播的载密图像也可能共享相似的隐写方法或者有效载荷. 利用这一特点, 本文提出一种示例分析方法, 增强示例包 (用户) 中正示例 (载密图像) 的共有模式特征, 使之与负示例 (正常图像) 的模式特征区分开, 进而能够简化发现包含正示例的正示例包 (隐写者) 的任务.

然而, 这些关联关系属于非欧的数据结构, 无法使用基于欧氏空间的深度学习方法进行建模. 因此, 目前仍鲜有基于深度学习的相关工作利用这种关联关系完成多示例学习任务.

本文提出用图对这种数据结构进行表示, 将包 B_x 中的示例 $\{x_1, \cdots, x_n\}$ 作为节点, 将节点之间的关联关系建模为边, 进而使用图卷积来处理节点间的关联信息. 在此基础上, 提出共性增强图卷积模块, 并使用 λ_b 个共性增强图卷积模块对示例包中的示例进行分析. 具体而言, 共性增强图卷积模块的结构如图 2(a) 所示, 下面将具体介绍共性增强图卷积模块的构成.

1) 构建示例包的图结构

共性增强图卷积模块的第一步是构建示例包的图结构. 示例包的图结构可以分为两部分: 一部分是选取示例表征中与特定模式相关的特征作为图结构中的图节点, 另一部分是根据图节点之间的关联关系创建图结构中的边. 只有构建的图结构准确表达示例间关联关系时, 图卷积模块才能利用示例间的关联关系完成学习任务.

目前, 图卷积领域的研究已经在图构建问题上取得了一定进展, 大多数工作将给定的特征向量作为节点, 致力于针对节点特征提出相似度度量方法, 例如欧氏距离、余弦相似度、热核函数等方法, 从而构建节点间的边. 然而, 相似度度量方法的设计往往依赖于先验知识和专家系统, 面对复杂多样的数据特征, 很难通过事先设计来表达不同语义、含义抽象的多种关联关系. 因此, 本文提出一种可学习的节点表征和相似度度量方法, 通过反向传播训练

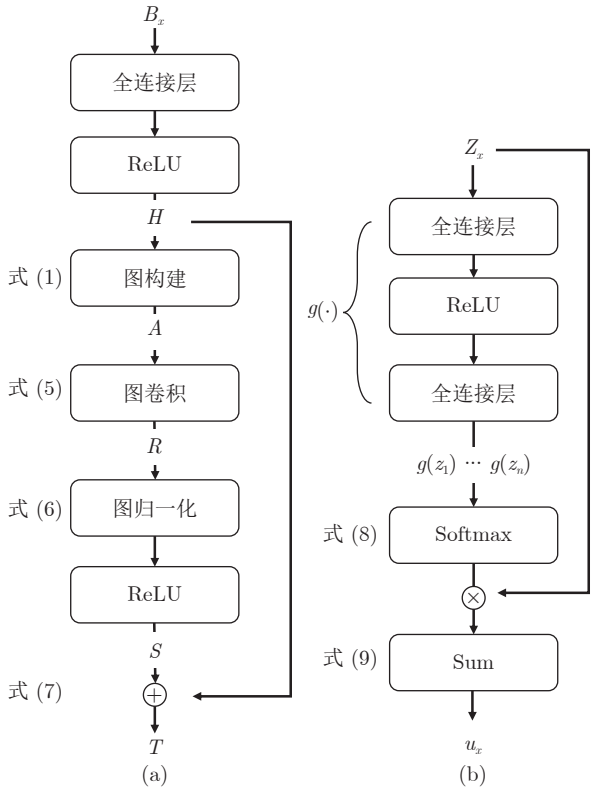


图 2　隐写者检测框架中两个模块 ((a) 共性增强图卷积模块; (b) 注意力读出模块)

Fig. 2　Two modules in steganographer detection framework ((a) The commonness enhancement graph convolutional network module; (b) The attention readout module)

模型自适应地构建图结构.

具体而言, 对于共性增强图卷积模块的第 i 个输入示例特征 v_i, 首先设计特征提取函数 $f: M_{1\times d}(\mathbf{R}) \to M_{1\times d'}(\mathbf{R})$ 对 d 维特征向量 v_i 进行变换, 得到 d' 维特征向量 $h_i = f(v_i)$ 作为节点 i 的特征表示. 其中, 特征提取函数 f 包括一个全连接层和一个线性整流单元 (Rectified linear unit, ReLU). 在训练过程中, 通过优化函数 f 中的参数, 模型将自适应地提取 v_i 中与检测目标相关的模式特征.

接着, 使用向量内积计算不同节点中与任务相关的模式特征的交互, 作为节点间关联关系的表征. 具体来说, 对于节点 i 和节点 j, 使用 $f(v_i)f^{\mathrm{T}}(v_j)$ 表示节点之间的关联关系, 并且使用 Sigmoid 函数将其归一化到 (0, 1) 范围内. 数学上, 对于一个有 n 个示例的示例包, 本文使用 $n\times n$ 维的邻接矩阵 A 表示图结构中的边. 对于不同示例之间可能存在不同程度的关联关系, 本文使用 A_{ij} 表示边的权重

$$A_{ij} = \frac{1}{1+\mathrm{e}^{-h_i h_j^{\mathrm{T}}}} \tag{1}$$

由上式可知, A_{ij} 是一个 0 到 1 之间的实数, 其中 0 表示示例 i 与示例 j 完全不相关, 而 1 表示两节点完全相关.

2) 利用图卷积聚合示例间的共性特征

共性增强图卷积模块的第二步是根据步骤 1) 中构建的图结构, 对示例包中的示例进行卷积运算, 聚合当前示例和具有相近模式特征的示例, 并提取公共的模式特征.

参照现有图卷积领域的研究成果[12], 本文类比欧氏空间中的卷积运算定义了示例包对应的图结构上的卷积运算. 具体而言, 将与当前示例节点处于同一包内的其他示例节点定义为卷积运算的邻域. 接着, 使用上文中计算得到的示例节点之间边的权重作为权, 对邻域内的示例节点的特征向量进行加权平均, 将得到的新特征向量作为当前示例节点的表征. 由式 (1) 可知, 与当前节点模式特征相关联的邻居节点将具有更大的权重. 为了在卷积运算中保留当前节点自身的特征, 本文在卷积运算的基础上, 以常数 1 作为权重, 为当前节点添加了自环, 共同参与卷积运算. 最终, 节点 i 的聚合结果 r_i 可以由下式计算得到

$$r_i = \sum_{j\in N_i\cup\{i\}} A_{ij}h_j + 1\times h_i \tag{2}$$

其中, N_i 表示节点 i 的一阶邻居节点. 不难发现, 式 (2) 等价于

$$R = (A+\mathbf{1})H \tag{3}$$

其中, H 表示用户分享图像的特征构成的矩阵, R 为聚合后的特征矩阵. 在式 (3) 中, 如果示例节点 i 与示例包中的较多示例相关, 那么 r_i 将会较大. 否则, 如果示例节点 i 与其他示例节点都不相关, 那么 r_i 将会较小. 相关工作表明, 在这种情况下, 特征矩阵 H 与邻接矩阵 A 相乘会使得特征向量 h_i 的尺度发生变化. 经过多次卷积之后, 这种数值变化可能会造成数值不稳定、梯度爆炸或弥散等问题. 因此, 本文限制了 h_i 的数值区间, 并对聚合结果进行归一化. 在式 (3) 的基础上, 卷积层被形式化为

$$R = (D+\mathbf{1})^{-\frac{1}{2}}(A+\mathbf{1})(D+\mathbf{1})^{-\frac{1}{2}}H \tag{4}$$

其中, D 是图上的度矩阵, $D(i,i) = \sum_j A(i,j)$, 通过使用度矩阵进行对称归一化, 能够消除示例包中与当前示例相关的数量对卷积运算结果的数值范围产生的影响, 从而有利于模型收敛.

最后, 参照现有图卷积领域的研究成果[12], 本文类比欧氏空间中卷积运算的权重共享机制, 使用一个参数矩阵 W 对聚合结果进行仿射变换, 提取聚合结果中与隐写者检测相关的示例间共享的模式特

征, 数学上可以表示为

$$R = (D+\mathbf{1})^{-\frac{1}{2}}(A+\mathbf{1})(D+\mathbf{1})^{-\frac{1}{2}}HW \tag{5}$$

其中, W 表示对聚合结果进行仿射变换的可训练的参数矩阵, 特别地, 本文使用 Xavier 方法对其进行了初始化.

3) 使用图归一化解决深层图卷积过平滑问题

相比于单层图卷积网络, 多层图卷积网络能够显著地大幅度提高模型性能. 然而, 多层图卷积会导致输出的特征产生过平滑的问题, 使得不同类别的节点特征变得不可区分. 在深度学习领域中, 现有工作普遍认为归一化技术能够有效解决深层网络的训练问题. 其中, 批量归一化被广泛应用在深度学习模型中, 在批量维度进行归一化操作. 然而, 这些方法无法被直接应用在图卷积网络中. 近年来, 一些现有工作提出了图上的归一化方法. 目前, 图归一化被认为能够有效地抑制深层图卷积过程中的过平滑问题. 因此, 本文使用一种用于多示例学习的新型图归一化方法.

一方面, 本文不希望在图卷积运算后, 示例包内的节点表征趋于一致, 即示例包内的节点的特征分布塌缩到向量空间中的一点, 因此本文提出在图卷积运算后对包内示例节点的特征分布进行重整, 重新将示例包中示例的特征向量缩放到模长为 1 的向量空间内. 另一方面, 为防止均值偏移过大导致出现神经元坏死现象, 本文同时矫正示例特征向量所处向量空间的均值. 数学上, 归一化后的示例节点的特征向量 s_i 可以表示为

$$s_i = \frac{r_i}{\|r_i\|_2} - \frac{1}{n}\sum_{j\in N_i\cup\{j\}} r_j \tag{6}$$

其中, N_i 为图中第 i 个示例节点的所有邻居节点.

实验过程中发现, 在多示例学习中, 示例包内示例特征的均值信息和尺度信息分别独立地对模型性能产生影响. 在一个示例包内, 当所有或大多数示例的特征向量都带有正示例的模式或者负示例的模式时, 示例特征的均值能够有效帮助区分该示例包. 在一个示例包内, 当其同时包含数量相当的正示例和负示例时, 示例包内示例特征的方差较大, 则说明示例包内同时包含两种示例, 即必然包含正示例, 因此示例包的方差也能够有效区分该示例包. 综合上述两点, 相比于传统的先进行均值归一化, 再将均值归一化的结果进行尺度归一化的图归一化方法, 本文将均值归一化和尺度归一化解耦, 同时基于输入节点特征向量 r_j 进行归一化.

接着, 本文使用 ReLU 对示例包内归一化后的示例节点特征 s_i 进行激活, 得到示例包和示例表征, $S = [\mathrm{ReLU}(s_1), \cdots, \mathrm{ReLU}(s_n)]^{\mathrm{T}}$.

4) 使用残差连接计算新的示例表征

由于图像内容的差异, 图中的一些节点与其他同类节点特征的相似度较小, 因此, 其对应的特征可能会在多层卷积中丢失. 为缓解该问题, 本文参照相关工作构建一个残差连接层, 将共性增强图卷积模块的输入特征 H 直接与图归一化的输出特征 S 相连. 数学上, 可以将残差连接层定义为

$$T = H + S \tag{7}$$

其中, T 表示残差连接操作的输出特征矩阵, $T = [t_1, \cdots, t_n]^{\mathrm{T}}$, 其中 t_i 为共性增强图卷积模块输出的第 i 个示例特征.

2.3 基于注意力的示例包表征

在使用共性增强图卷积网络进行示例分析后, 为对示例包进行正确检测, 还需要根据示例特征构建示例包的特征表示. 本文将用户 x 经过共性增强图卷积模块的输出表示为 $Z_x = [z_1, \cdots, z_n]^{\mathrm{T}}$. 其中, z_i 对应于第 i 张图片进行示例分析后的特征表示. 现有的隐写者检测工作通常将用户分享的每张图片视为同等重要的个体, 根据用户分享的所有图像的特征分布作为用户的特征分布, 计算用户之间的差异并检测异常用户作为隐写者. 然而, 在载密图像占比较小或是载密图像包含的嵌入信息较少的情况下, 隐写者的表征将与正常用户的表征极为相近, 无法有效检测出隐写者.

为解决该问题, 本文依托对隐写者检测任务的多示例学习形式化, 提出注意力示例包表征模块, 自适应地构建更具有区分力的示例包表征. 与多示例学习的定义一致, 只要示例包中包含至少一个正示例, 该示例包就应当被判定为正示例包. 因此, 带有正示例模式特征的示例包, 在表征和示例包检测的过程中应当被予以更多的关注. 基于上述考虑, 本文设计注意力读出模块 (图 2(b)), 根据示例中包含的正负示例模式对示例包中的示例赋予不同权值, 形成示例包的表征.

为自适应地识别示例包中符合正示例模式特征的示例, 并对其施加更显著的注意力, 本文使用注意力计算函数 $g(\cdot)$ 来根据示例的特征分布计算其对应示例在示例包表征过程中的重要程度. 其中注意力计算函数 $g(\cdot)$ 包括两个全连接层, 在第一个全连接层后, 使用 ReLU 作为激活函数. 在训练过程中, 模型通过调整 $g(\cdot)$ 中的训练参数, 将能够针对符合正示例模式特征的示例输出更高的数值.

在此基础上, 使用 Softmax 函数对示例包中所有示例的重要程度进行归一化, 得到每个示例在表

征示例包过程中的权重得分 p_i

$$p_i = \frac{\mathrm{e}^{g(z_i)}}{\sum_{j}^{n} \mathrm{e}^{g(z_j)}} \tag{8}$$

接着, 使用式 (8) 中计算得到的权重值作为权, 以示例包中每个示例特征的加权平均作为示例包的表征 u_x

$$u_x = \sum_{i}^{n} p_i z_i \tag{9}$$

相比于现有工作中使用平均、合并等方法建立的示例包表征, 本文提出的示例包表征方法能够自适应地学习并对正示例赋予更高的权重, 从而提高示例包表征的区分能力.

2.4 基于多种损失约束的示例包分类

在得到基于注意力的示例包表征 u_x 之后, 将其送入由一个全连接层和一个 Softmax 激活函数构成的分类器, 最终完成示例包的分类任务. 需要注意的是, 本文提出的模型其训练过程和测试过程有所不同, 在训练过程中, 以示例包的分类作为目标, 而在测试的过程中, 采用与以往隐写者检测一致的实验设置, 即将预测的隐写者概率最高的用户作为隐写者输出.

本文设计的模型 MILGCN 在优化过程中是端到端的, 即, 使用多个损失函数加权求和, 并将得到的损失进行反向传播, 更新网络中的所有模块. 在训练分类模型的过程中, 所使用的损失函数由三部分构成

$$L = \lambda_1 L_{\text{bag}} + \lambda_2 L_{\text{entropy}} + \lambda_3 L_{\text{contrastive}} \tag{10}$$

其中, L_{bag} 为多示例分类损失, L_{entropy} 为熵正则损失, $L_{\text{contrastive}}$ 为对比学习损失. λ_1, λ_2, λ_3 为三个超参数, 分别表示多示例分类损失、熵正则损失和对比学习损失作为训练目标的权重. 下面分别介绍三个损失函数.

多示例分类损失: 多示例分类损失用于指导模型对隐写者和正常用户进行正确分类. 复杂的样本特征可能会导致这些示例 (图像) 的预测结果有较大偏差, 对于最终分类结果的意义有限, 并不能得到令人满意的包的预测结果. 与此相反, 相比于注重示例的准确性, 多示例学习更注重包的预测标签是否符合真实标签. 因此, 本文提出一个多示例分类损失来优化图卷积神经网络. 如式 (11) 所示, 使用包 (用户) 的二分类逻辑损失来优化模型, 使最终输出结果能够对隐写者和正常用户进行分类

$$L_{\text{bag}} = \sum_{i}^{m} \frac{\ln(1 + \mathrm{e}^{-Y_i \rho_i})}{m} \tag{11}$$

其中, Y_i 表示第 i 个用户 (包) 的标签, 模型预测其为隐写者的概率为 ρ_i, m 表示用户 (包) 的总数量. 多示例分类损失忽略单个示例类别的不确定性, 而更重视全局信息, 这增强了整个框架的鲁棒性.

熵正则损失: 熵正则损失用于指导模型在根据示例表征构建示例包表征的过程中, 利用先验信息抓取带有正示例模式特征的示例, 构建示例包的表征. 现有的隐写者检测相关工作已经表明, 隐写者在分享图像和嵌入秘密信息的过程中有多种策略选择. 对于固定的有效载荷, 隐写者可能将有效载荷分散在多数分享图像中, 每张载密图像的嵌入信息较少, 这时, 关注尽可能多的分享图像, 即使用分享图像特征的均值作为用户表征能够更有效地检测隐写者. 隐写者也可能将有效载荷分散在少数分享图像中, 每张载密图像的嵌入信息较多, 这时, 关注尽可能少的分享图像, 使用最有嫌疑的图像特征作为用户表征则更为高效.

尽管本文设计的注意力模块能够自适应地学习攻击不同的隐写策略, 能够通过学习判别应当关注的示例, 但是, 当隐写者检测人员明确知晓训练集合中隐写者使用的嵌入策略时, 本文也提供了损失函数融入隐写者检测人员对于嵌入策略的先验知识, 帮助模型更快收敛, 提升模型性能

$$L_{\text{entropy}} = \lambda_e \sum_{i}^{n} \frac{-p_i \ln(p_i)}{n} \tag{12}$$

其中, λ_e 为超参数, 当隐写者检测人员认为隐写者更可能将有效载荷分散在较少的分享图像中时, λ_e 将被设置成 $+1$. 通过最小化 L_{entropy}, 模型将学习给出低熵的示例包构建方式, 即, 以更大的概率对尽可能少的正示例赋予更大权重, 作为示例包的表征. 这种损失设计有利于模型在正示例占比较低时, 仍能分辨出其中正示例的模式特征. 反之, 当隐写者检测人员认为隐写者更可能将秘密信息以较低有效载荷分散在较多的分享图像中时, λ_e 将被设置成 -1. 通过最小化 L_{entropy}, 模型将学习给出高熵的示例包构建方式, 即, 使用尽可能多的正示例特征相互佐证, 检测隐写者.

对比学习损失: 对比学习损失用于引导模型在提取示例特征、构建示例包的过程中, 尽量保留示例间的关联关系和差异性. 对比学习旨在将包含同类模式特征 (例如正示例特征) 的示例表征映射到向量空间中相近的区域, 对包含不同类别模式特征的示例表征进行区分, 映射到相距较远的区域, 从

而帮助构建有效的图结构进行示例分析和示例包表征.

正如上文中提到的, 由于图像特征的复杂性, 直接对示例的模式特征进行约束可能会导致因图像内容的差异而得到不令人满意的映射函数. 因此, 相比于直接使用示例特征进行引导, 本文受到多示例学习的启发, 通过引导模型保留同类示例包间的相似性、不同类示例包间的差异性, 间接地迫使模型保留作为示例包组成部分的主要示例特征间的相似性和差异性. 最终, 提出对比学习损失

$$L_{\text{contrastive}} = \sum_{i}^{m} \phi(i, j) \left\| u_i - u_j \right\|_2 - (1 - \phi(i, j)) \| u_i - u_j \|_2 \tag{13}$$

其中, $\phi(i, j)$ 为指示函数, 用于指示示例包 i 和示例包 j 是否为同一类别的示例包. 当示例包 i 和示例包 j 同时为正示例包或负示例包时, $\phi(i, j) = 1$; 当示例包 i 和示例包 j 中的一个为正示例包、一个为负示例包时, $\phi(i, j) = 0$. 当示例包属于同一类别时, 对比学习损失最小化两示例包表征间的欧氏距离 $\| u_i - u_j \|_2$, 提高同类示例包特征及其包含示例特征的相似性. 当示例包不属于同一类别时, 对比学习损失最小化两示例包表征间的欧氏距离的相反数 $-\| u_i - u_j \|_2$, 提高不同类别示例包特征及其包含示例特征的差异性. 因此该损失可以最大化不同类别示例包间的差异, 最小化相同类别示例包间的差异, 以区分不同类别的示例和示例包.

3 实验评估

为验证所提出的模型, 本文在两个隐写者检测基准数据集上进行一系列验证实验. 首先, 使用隐写者检测基准数据集对所提出的 MILGCN 模型在隐写者使用不同隐写术时的检测性能进行验证. 在空域中, 本文在标准数据集 BOSSbase ver 1.01[20] 上进行实验, 对不同方法检测使用空域隐写术的隐写者的性能进行对比分析. 在频域中, 使用 JPEG 算法在 80 质量因子的参数下对 BOSSbase ver 1.01 上的图像进行压缩, 将其中的图像转换为 JPEG 图像, 并在得到的 BOSSbase ver 1.01 JPEG 基准数据集上进行实验[21], 对不同方法检测使用频域隐写术的隐写者的性能进行对比分析. 与之前隐写者检测的已有工作一致, 本文使用的方法和对比方法都采用真阳性率进行评估.

本文在实验过程中对比分析了大量不同隐写者检测算法. 其中包括: 结合传统隐写分析模型 SRMQ1[22] 与层次聚类的隐写者检测框架 SRMQ1_SD, 目前最前沿的隐写者检测方法 MDNNSD, 结合基于生成对抗网络 (Generative adversarial network, GAN) 的隐写方法 SSGAN[23] 与层次聚类的隐写者检测框架得出的 SSGAN_SD, 多示例学习任务中的前沿方法 MILNN[24], 著名的基于图的模型 GAT[15]、GraphSAGE[13]、AGNN[25]、GCN[12] 和 DiffPool[14], 以及迄今为止唯一用于隐写者检测任务的基于图的深度学习模型 SAGCN. 其中, MDNNSD 是基于深度学习提取特征, 再与传统的聚类方法结合的隐写者检测方法[7]. 相比于其他现有隐写者检测方法, 例如 PEV_SD 和基于最前沿隐写分析方法 XuNet[26] 构建的隐写者检测框架 XuNet_SD, MDNNSD 取得了最好的隐写者检测性能. 因此, 本文选取 MSD 作为代表, 进行着重的对比分析. 近年来, 基于 GAN 的隐写术和隐写分析技术取得了巨大的成功, 因此, 本文将已有的基于 GAN 的隐写方法 SSGAN 与层次聚类的隐写者检测方法结合, 得到隐写者检测框架 SSGAN_SD, 来对本文提出的模型与基于 GAN 的隐写者检测模型进行对比. 特别地, 在对比过程中, 本文使用 GAN 方法学习得到的判别器被用作特征提取器, 用来从用户分享的图像中提取隐写分析特征, 并在此基础上使用 MMD 距离度量用户表征之间的差异, 并使用层次分析算法来检测隐写者. 为说明本文设计的 MILGCN 在图结构和图卷积方向上的性能提升, 本文对比了 Zhang 等[8−9] 设计的 SAGCN 模型, 并使用著名的基于图的方法 (例如 GAT、GraphSAGE、AGNN、GCN 和 DiffPool) 中的图卷积层替换 SAGCN 框架中的核心部分——相似度累积图卷积单元, 来与其他基于图的深度学习算法进行比较.

就实验设置而言, 在对本文提出的网络结构进行验证时, 将 MILGCN 中共性增强图卷积模块的数量 λ_b 设置为 2 个, 并使用 $\lambda_1 = 1.0$ 作为多示例分类损失的权重, $\lambda_2 = 0.01$ 作为熵正则损失的权重, $\lambda_3 = 0.01$ 作为对比学习损失的权重. 这三种损失函数的权重设置遵循重要性原则, 多示例分类损失是为之后的检测服务的, 是本文的任务目标, 因此其权重比其他两种大. 另外, 将权重值在 $\{-10, -1, -0.10, -0.01, 0.01, 0.10, 1, 10\}$ 范围中进行搜索, 选取较好的结果. 在训练阶段, 本文使用随机梯度下降法来最小化损失函数, 初始学习率设置为 0.001, 动量设置为 0.9. 本文还使用 L2 正则化以避免过拟合并提高泛化能力, 并将正则化项的权重设置为 0.01. 本文的实验设置与基于图卷积的隐写者检测方法 SAGCN[8−9] 基本相同, 唯一的区别是本文中设

计的 MILGCN 模型无须使用序列训练方法, 即可有效收敛并检测使用不同有效载荷的隐写者. 在实验过程中, 给定基准数据集 BOSSbase ver 1.01 和 BOSSbase ver 1.01 JPEG, 首先使用实验中所探究的隐写术, 从 0.5 到 0.05 的嵌入率, 生成载体图像对应的载密图像. 对于每种嵌入率, 分别有 20 000 张载体图像和载密图像. 每个批次的训练中生成 50 个正常用户, 每个用户分享 200 张从 20 000 张载体图像中有放回抽样得到的载体图像; 生成 50 个隐写者, 每个隐写者分享 200 张从 20 000 张载密图像中有放回抽样得到的载密图像. 本文在训练集上训练 80 轮, 每轮训练包含 600 个批次, 每个批次包含 100 个用户样本, 因此共 60 000 个用户混合在一起随机打乱, 作为训练集训练 MILGCN. 在测试阶段, 每次实验中的用户数量被设置为 100, 包括 1 名隐写者和 99 名正常用户, 每个用户分享 200 张图像. 该实验设置模拟了真实环境中经常出现的现象, 即在众多用户中只有 1 名隐写者或者不存在隐写者. 在不同实验中, 使用不同的隐写术和有效载荷生成隐写者分享的载密图像, 使用的隐写术包括: S-UNIWARD[27]、HUGO-BD[28]、WOW[29]、MiPod[30]、J-UNIWARD[27]、nsF5[31]、UERD[32] 等. 与之前已有的隐写者检测工作一致, 所有的统计实验都重复 100 次, 并在文中汇报平均的检测准确率, 即真阳率. 衡量嵌入率的单位分为空域的 bpp (Bit per pixel) 和频率域的 bpnzAC (Bit per non zero DCT coefficient), 前者表示每像素嵌入的比特数, 后者表示每个非零的 DCT 系数中嵌入的比特数. 此外, 本文中的全部模型及实验使用 PyTorch 在 4 张 Tesla V100 上完成.

3.1 基于空域的隐写者检测性能评估与对比

3.1.1 已知隐写术情况下的隐写者检测

在本节中, 本文在最常使用的隐写者检测基准数据集 BOSSbase ver 1.01 上进行初步实验, 在已知测试阶段图像隐写者所使用的隐写术的条件下, 对所提出的模型的效率和效果进行检验. 具体而言, 隐写者检测基准数据集 BOSSbase ver 1.01 中包含 10 000 张自然图像的灰度图, 其尺寸为 512 × 512 像素. 根据之前隐写分析和隐写者检测任务的通用实验设置[7−9, 33], 本文将数据集中的每张图像切分为 4 个互相不重叠子图, 其中子图的尺寸为 256 × 256 像素. 与前文描述的将所有载荷的样本进行训练不同, 在此实验的训练阶段, 只使用 S-UNIWARD 隐写术在给定有效载荷下对每张载体图像嵌入秘密信息得到对应的载密图像, 进而将得到的载体图像和载密图像组合为用户作为训练样本. 在训练阶段的每个批次中, 本文采样 50 名正常用户, 每名正常用户分享 200 张载体图像, 并采样 50 名隐写者, 每名隐写者分享 200 张载密图像, 每轮训练采样 100 个批次的用户对模型进行优化.

同样地, 测试样本由剩余的 20 000 张图像使用同样的方法构成. 在测试阶段的每次实验中, 被测方法被要求从 100 名用户中检测出 1 名隐写者. 特别地, 本文在 20 000 张载体图像中随机选择 200 张载体图像对应的载密图像, 构成 1 名图像隐写者, 剩余的 19 800 张载体图像被随机选择并分配给 99 个正常用户. 为验证在隐写者检测人员已知隐写者使用的图像隐写术时本文提出方法的有效性, 在本节实验中保持测试阶段隐写者使用的图像隐写术与训练阶段相同. 即, 在训练阶段, 本文隐写者使用 S-UNIWARD 作为图像隐写术嵌入秘密信息生成载密图像, 在生成的 60 000 个用户中, 隐写者使用的有效载荷包括 0.05 bpp、0.1 bpp、0.2 bpp、0.3 bpp、0.4 bpp 和 0.5 bpp. 在测试阶段, 隐写者分享的每张图像中同样只包含使用 S-UNIWARD 图像隐写术嵌入的秘密信息. 本文分别在 5 种不同的有效载荷上进行测试, 其中包括 0.05 bpp、0.1 bpp、0.2 bpp、0.3 bpp 和 0.4 bpp.

如表 2 所示, 本文首先对比隐写者检测中的相关工作. 根据相关工作[7], MDNNSD 和 XuNet_SD 这些隐写分析方法直接用于隐写者检测任务的检测结果对比, 会取得更好的检测效果. 然而, 与现有的在实验环境中进行研究的隐写者检测任务不同, 在实际应用中, 图像隐写者通常会使用批量隐写术, 将秘密信息分散在多张图像中, 使用较小的嵌入率在每张图像中嵌入秘密信息, 从而隐蔽自己. 所以本文更关注在困难情况下, 即当有效载荷较低时, 不同方法是否能够检测出图像隐写者. 因此, 本文进一步对比图卷积神经网络在隐写者检测中的应用效果. 正如上文所提到的, 当有效载荷小于 0.1 bpp 时, SAGCN、MILGCN 和其他基于图深度学习的模型都能够取得比 MDNNSD 更好的性能. 与之相对的是, 本文提出的 MILGCN 进一步超越了 SAGCN, 在有效载荷为 0.05 bpp 的困难情况下取得了更好的隐写者检测效果. 尽管生成对抗网络在很多领域取得了突破进展[34−41], 但 SSGAN_SD 效果较差, 这主要是因为 SSGAN 原本设计的目的是通过对抗训练得到一种自适应的隐写术而不是得到隐写分析模型, 因此在设计过程中, 相对于其他工作, SSGAN_SD 的判别器结构简单、判别能力差.

正如上文所提到的, 本文创新性地提出将隐写

表 2 已知隐写者使用相同图像隐写术 (S-UNIWARD) 时的隐写者检测准确率 (%), 嵌入率从 0.05 bpp 到 0.4 bpp

Table 2 Steganography detection accuracy rate (%) when steganographers use the same image steganography (S-UNIWARD), while the embedding payload is from 0.05 bpp to 0.4 bpp

模型		嵌入率 (bpp)				
		0.05	0.1	0.2	0.3	0.4
前沿	MDNNSD	4	54	100	100	100
	XuNet_SD	2	2	71	100	100
基于 GAN	SSGAN_SD	0	1	1	2	4
基于 GNN	GAT	2	3	3	3	4
	GraphSAGE	28	88	100	100	100
	AGNN	24	99	100	100	100
	GCN	19	96	100	100	100
	SAGCN	72	100	100	100	100
基于 MIL	MILNN_self	15	87	100	100	100
	MILNN_git	18	96	100	100	100
本文	MILGCN-MF	47	**100**	**100**	**100**	**100**
	MILGCN	**74**	**100**	**100**	**100**	**100**

者检测任务形式化为多示例学习问题, 从而开辟了将现有多示例学习方法应用于隐写者检测任务的路径. 因此, 本文进一步将此前 Pevný 等[24] 提出的最优的多示例学习神经网络应用在隐写者检测问题中进行对比. 具体而言, 一方面, 本文沿用 Pevný 等[24] 在论文中汇报的设置, 自行在隐写者检测问题上复现其提出的方法, 将该方法称为 MILNN-self, 并提供了 MILNN-self 的隐写者检测性能. 另一方面, 本文也参考并应用 Github 社区中实现的 Pevný 等[24] 提出的方法, 并提供了其在隐写者检测任务上的实验结果, 后者称其为 MILNN-git. 值得注意的是, MILNN-git 中的模型实现与 Pevný 等[24] 的工作略有不同, 其模型包含了一个隐藏层、一个平均最大池化层和两个输出单元, 除了最后一个输出单元使用 Tanh 作为激活函数外, 所有隐藏层使用 Sigmoid 作为激活函数, 模型在训练过程中最小化交叉熵损失. 实验结果表明, MILNN-git 能够在隐写者检测问题上取得比 MDNNSD 更好的效果. 同时, 无论隐写者使用的有效载荷如何变化, 本文提出的多示例学习图卷积网络都强于任何版本的 MILNN 多示例学习模型.

除此以外, 本文还比较了传统图归一化方法和本文提出的新型图归一化方法. 具体而言, 将本文中使用的归一化方法修改为传统图归一化方法 (即先进行均值归一化再将均值归一化的结果进行尺度归一化), 得到 MILGCN-MF 模型作为对比方法. 从表 2 中可以看出, 在隐写者使用的有效载荷较高时 (大于等于 0.1 bpp), MILGCN-MF 和 MILGCN 的性能没有差异, 都能够有效检测出 100 名用户中包含的隐写者, 这表明在简单情况下, 两种归一化方法都能够有效地帮助模型构建具有区分能力的示例表征. 但是, 当有效载荷较低时 (为 0.05 bpp), MILGCN-MF 的性能产生了约 30% 的下滑. 当隐写者分享的有效载荷较低时, 正示例包和负示例包分布相近, 特征差异不明显. 在这种情况下, 使用原有的图归一化方法导致性能下降, 意味着先进行均值归一化再进行方差归一化使得部分特征的区分能力丢失. 与此相对, 本文提出的图归一化方法能够记录原始特征中的均值和方差, 在归一化的过程中保留和识别特征分布间的差异, 得到更好的检测效果.

最近还有一些新的隐写分析方法被提出, 如果不采用层次聚类的方法将其改造为隐写者检测方法, 而是使用其隐写分析模型的输出概率作为每张图像是载密图像的概率, 并使用用户的所有图像预测为载密图像的平均概率作为该用户为隐写者的概率, 也可以得到与最新的隐写分析方法相比较的结果. 具体而言, 在训练阶段, 隐写者使用 S-UNIWARD 作为图像隐写术, 隐写者检测模型学习在正常用户中检测使用 0.2 bpp 嵌入率对分享图像嵌入秘密信息的隐写者. 这里, 隐写者分享的图像均为载密图像, 正常用户分享的图像均为载体图像. 在测试阶段, 模型需要从 100 名用户中找出 1 名隐写者. 其中, 每名用户分享 200 张图像, 隐写者仍然使用 S-UNIWARD 作为图像隐写术, 使用 0.2 bpp 嵌入率对分享图像嵌入秘密信息. 根据实验的不同, 隐写者分享的载密图像可能占比其分享的所有图像的 10%、30%、50%、90% 或 100%. 本文进行 100 次实验, 并记录模型在隐写者检测过程中的平均准确率, 如表 3 所示. 在 SRNet-AVG 方法中, 本文将 SRNet[42] 直接用于隐写者检测, 预测用户分享的每一张图像可能是载密图像的概率, 并使用该用户的每一张图像预测为载密图像的平均概率作为该用户为隐写者的概率. 在 SRNet-MILGCN 方法中, 使用本文提出的 MILGCN 作为隐写者检测算法, 将 SRNet 提取的特征作为 MILGCN 的输入预测用户为隐写者的概率. 由表 3 可见, 无论使用哪种方法, 模型都能够在载密图像占比超过 10% 的情况下有效检测出隐写者. 这是由于隐写者使用的隐写术和有效载荷已知, 训练好的 SRNet 能够较为准确地检测出使用该隐写术和载荷嵌入秘密信息的载密图像. 当隐写者分享的载密图像较多时, 数量众多的载密图像容易被 SRNet 检测到, 从而使得隐写者区

表 3　当测试阶段隐写者使用相同隐写术 (S-UNIWARD) 和分享的载密图像数量占总图像数量为 10% 到 100% 时, SRNet-AVG 和 SRNet-MILGCN 的检测成功率 (%)

Table 3　The accurate rate (%) of SRNet-AVG and SRNet-MILGCN when the number of shared secret images is from 10% to 100% of the total number of images and the steganographer uses the same steganography (S-UNIWARD) in test

方法	占比 (%)					
	10	30	50	70	90	100
SRNet-AVG	26	100	100	100	100	100
SRNet-MILGCN	35	100	100	100	100	100

别于正常用户. 而当载密图像占其分享的所有图像比例较低时, SRNet-AVG 和 SRNet-MILGCN 都产生了较大的性能下滑, 相比于 SRNet-AVG, 本文提出的 SRNet-MILGCN 取得了接近 10% 的性能提升, 表明本文设计的模型能够更好地对抗将隐秘信息嵌入较少图像的批量隐写策略. 这是由于 MILGCN 能够利用共性增强图卷积网络和注意力图读出模块自适应地突出示例包中正示例的模式特征, 构建示例包表征进行分类. 在这种设计下, 正示例所对应的载密图像的模式特征更容易被模型发现.

3.1.2　对用户分享图像数量变化的探究

在真实应用场景的社交网络中, 用户分享的图像数量是海量的. 因此, 本文进一步松弛现有工作在实验中对用户分享的图像数量的约束, 在隐写者检测任务上进行实验来研究用户分享的图像数量对隐写者检测性能的影响. 具体而言, 本节保持其他设置与第 3.1.1 节中的实验设置相同, 验证当用户分享图像的数量为 100、200、400 和 600 张时, 模型检测使用不同有效载荷在分享图像中嵌入秘密信息的隐写者的性能.

如表 4 所示, 当隐写者使用大于 0.1 bpp 的有效载荷在分享图像中嵌入秘密信息时, 无论隐写者分享的图像数量如何变化, 本文提出的 MILGCN 和 Zhang 等[8−9] 设计的 SAGCN 都能够准确地检测出隐写者, 这说明本文提出的方法对用户分享图像数量的变化具有鲁棒性. 当有效载荷等于或小于 0.1 bpp 时, 发现所有方法对隐写者的检测性能随着用户分享图像数量的减少而下降. 这种现象是由于当隐写者分享的图像数量或有效载荷较少时, 隐写者在分享图像中嵌入的秘密信息的总量较少, 这使得隐写者的特征分布与正常用户的特征分布差异较小, 更加难以区分. 同时, 由于隐写者和用户都分享了较少的图像, 图像数量的不足使得图像间的关联信息十分有限, 模型难以利用这些关联关系增强

表 4　当用户分享不同数量的图像时, 使用 MILGCN 和 SAGCN 进行隐写者检测的准确率 (%), 嵌入率从 0.05 bpp 到 0.4 bpp

Table 4　Steganography detection accuracy rate (%) of MILGCN and SAGCN when users share different numbers of images, while the embedding payload is from 0.05 bpp to 0.4 bpp

	数量 (张)	嵌入率 (bpp)				
		0.05	0.1	0.2	0.3	0.4
MILGCN	100	**35**	**96**	**100**	**100**	**100**
	200	**74**	**100**	**100**	**100**	**100**
	400	**96**	**100**	**100**	**100**	**100**
	600	**100**	**100**	**100**	**100**	**100**
SAGCN	100	31	96	100	100	100
	200	72	100	100	100	100
	400	91	100	100	100	100
	600	91	100	100	100	100

用户所对应的示例包中正示例的公共特征, 进而导致模型无法区分隐写者与众多正常的用户.

尽管如此, 无论用户分享的图像数量如何变化, 本文设计的 MILGCN 都能够超过此前最佳隐写者检测算法 SAGCN, 取得令人满意的效果. 值得注意的是, 当用户分享的图像数量为 600 张时, 即使隐写者使用非常低的有效载荷, 即 0.05 bpp 的有效载荷嵌入秘密信息, 本文提出的 MILGCN 依然能够聚合分散在示例包中正示例的模式特征, 对用户形成具有区分能力的示例包表征, 并对隐写者进行正确检测.

3.1.3　隐写术未知情况下对使用比例嵌入策略的隐写者检测

在本节中, 进一步放宽两个基础假设. 一方面, 检测者在对社交网络中用户分享的图像进行分析时, 无法预知隐写者使用的图像隐写术. 因此, 本文放宽了图像隐写者在训练阶段和测试阶段使用的图像隐写术为相同隐写术的假设, 探讨在隐写术错配情况下本文提出模型的隐写者检测性能. 另一方面, 在真实应用场景中, 隐写者可能使用批量图像隐写策略, 在分享的图像中同时分享载体图像和载密图像以防止自己被发现. 因此, 本文放宽了隐写者分享的图像都是载密图像的假设, 探究隐写者分享的图像中载密图像占某一比例时模型的隐写者检测性能.

具体而言, 在训练阶段, 隐写者仍使用 S-UNIWARD 作为图像隐写术. S-UNIWARD 在隐写、隐写分析、隐写者检测任务中被认为是更为通用的设置, 常常用来作为训练模型中载密数据所使用的

隐写术, 在本文对比方法 SAGCN 中, 也采用其作为隐写术未知情况下的实验设置. 隐写者检测模型学习在正常用户中检测使用 0.5 bpp、0.4 bpp、0.3 bpp、0.2 bpp、0.1 bpp 或 0.05 bpp 嵌入率对分享图像嵌入秘密信息的隐写者. 这里, 隐写者分享的图像均为载密图像, 正常用户分享的图像均为载体图像.

在测试阶段, 隐写者将其使用的图像隐写术替换为与训练阶段不同的隐写术, 特别地, 本文中研究探讨的错配隐写术包括现阶段图像隐写术和隐写者检测领域普遍使用的 HUGO-BD、WOW、HILL 和 MiPOD 等. 在每次实验中, 模型需要从 100 名用户中找出 1 名隐写者. 其中, 每名用户分享 200 张图像, 隐写者使用的有效载荷为 0.2 bpp, 根据实验的不同, 隐写者分享的载密图像可能占比其分享的所有图像的 10% 、30% 、50% 、90% 或100%. 本文进行 100 次实验, 并记录模型在检测隐写者过程中的平均准确率.

图 3 展示了当隐写者在测试阶段使用不同图像隐写术时不同方法检测隐写者的性能. 容易发现, 在图像隐写术错配的情况下, 无论隐写者选择使用何种图像隐写术, 当隐写者分享的载密图像占其分享图像的数量 100% 时, 所有基于图的模型都能够正确检测出隐写者. 此外, 还可以发现所有隐写者检测方法的性能都随着隐写者分享的载密图像占比的升高而升高, 这是由于当隐写者分享的载密图像占比较低时, 隐写者分享的图像大部分为自然图像, 即载体图像, 在这种情况下, 隐写者的表征将与正常用户非常接近, 模型难以正确区分隐写者和正常用户. 然而, 即使在隐写者分享的载密图像占比低于 50% 的情况下, 本文提出的 MILGCN 依然能够取得较好的效果. 特别的是, 在大多数情况下, MILGCN 的检测准确率曲线覆盖了其他方法的检测准确率曲线, 这说明 MILGCN 能够取得比 SAGCN 及其他基于图的隐写者检测方法更好的性能. 这是由于本文提出的 MILGCN 是针对隐写者检测问题

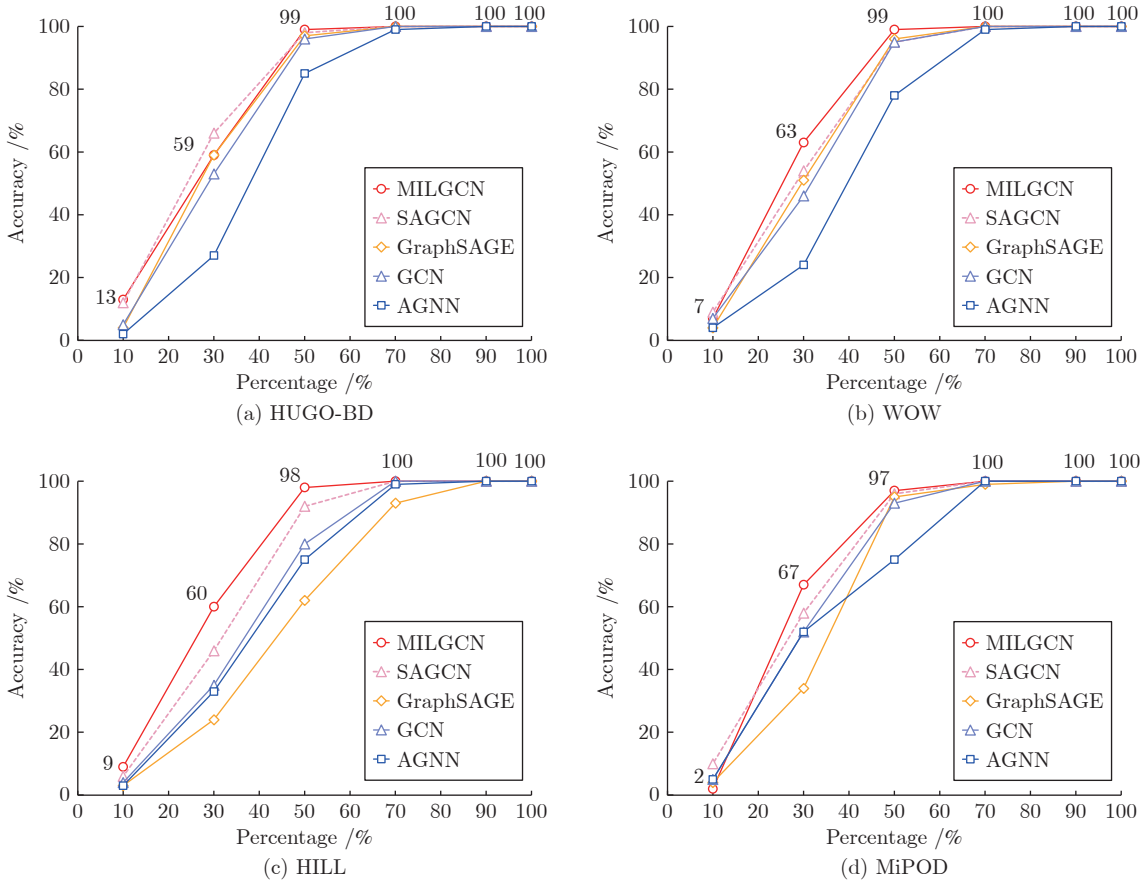


图 3 当测试阶段隐写者使用不同隐写术、分享的载密图像数量占总图像数量的 10% 到 100% 时, 不同的基于图的隐写者检测方法检测准确率

Fig. 3 The accurate rate of different graph-based steganographer detection methods when the number of shared secret images is from 10% to 100% of the total number of images and the steganographer uses different steganography in test

的多示例学习形式化提出的, 在设计过程中考虑了正示例占比较低的情况, 能够使用共性增强图卷积网络来增强示例包中正示例的公共模式特征, 并使用基于注意力的示例包表示选择识别到的正示例模式特征作为示例包表征, 从而区分正负示例包. 实验结果表明, 本文提出的方法对不同隐写术具有良好的泛化性, 在隐写者分享的载密图像占比较低时也能取得较好的检测性能. 从图 3 中还可以发现, GCN 的性能稍差于 SAGCN, 而当隐写者分享的载密图像占比较低时, AGNN 检测隐写者的准确率都是最低的, 这种差异源自于图结构的构建方式不同. MILGCN 依托于示例节点特征间的内积计算节点的关联关系, 在损失函数约束下, 正示例的特征分布相近, 并与负示例的特征分布区分开, 因而具有共性的示例节点关联关系更强, 在此基础上, 在图卷积过程中, 正示例间的共性特征被增强, 从而使得模型能够构建具备区分力的用户表征. 而 GCN 和 SAGCN 基于热核函数计算节点之间的相似度, 当隐写者分享的图像中同时包含载体图像和载密图像时, 载密图像和载体图像所对应图结构中的节点特征相似度较小, 而载密图像之间或载体图像之间的节点特征相似度较大, 在图卷积过程中, 相似的节点可以识别和累积载密图像节点间共有的模式特征, 从而区分隐写者和正常用户的分布差异. 与之相反的是, AGNN 在构建图的过程中, 利用的是注意力机制, 在训练过程中, 隐写者分享的图像均为载密图像, 这导致 AGNN 缺少对构建载体图像与载密图像关联关系的泛化能力, 因而性能下降. 此外, GraphSAGE 使用采样方法和平均聚合器来构建节点的邻域并进行卷积运算, 如果载体图像与载密图像之间差异较大, 其能够学习到边的权重, 累积同类节点的特征. 但对于譬如 HILL 这样难以攻击的隐写术, 载密图像与载体图像之间特征分布相近, 所以 GraphSAGE 的效果也会有所下降, 而本文提出的 MILGCN 泛化性能最好.

除此以外, 探讨一下本文的方法在什么比例下将完全失效. 首先对“完全失效”进行定义, 当从 100 个用户中检测 1 名隐写者的成功率接近 1% 时, 模型的检测性能等同于随机猜测的性能, 可以认为模型完全失效. 在此基础上, 本文添加了实验, 探讨在用户分享的载密图像占比较低时, 模型的检测性能是如何接近完全失效的. 这里, 保持本节中实验的设置不变, 测试隐写者分享的载密图像占其分享的所有图像 5% 时模型的检测性能. 本文进行 100 次实验, 并记录模型在检测隐写者过程中的平均准确率, 实验结果如表 5 所示.

表 5　在隐写术错配情况下, 当分享的载密图像数量占比 5% 时, MILGCN 取得的隐写者检测准确率 (%)

Table 5　Steganography detection accuracy rate (%) in the case of steganography mismatch when the number of shared secret images is 5% of the total number of images

测试隐写术	HUGO-BD	WOW	HILL	MiPOD
检测准确率	6	4	5	5

可以发现, MILGCN 在隐写术错配的情况下, 对于较为容易攻击的隐写术 HUGO-BD, 即使在载密图像占比仅为 5% 的情况下, 仍有 6% 的成功率能够检测出隐写者. 当隐写者使用 MiPOD 隐写术时, MILGCN 仅能够以 3% 的成功率检测出载密图像占比为 5% 的隐写者, 检测性能接近完全失效的性能. 可见 5% 的载密图像占比是非常困难的检测场景, 因此建议将本文提出的方法用于隐写者分享的图像占比超过 5% 的场景中.

上文中都使用 S-UNIWARD 作为训练数据的图像隐写术, 这主要是因为本文希望在训练模型时, 即使只使用一种隐写术得到载密图像, 训练得到的模型也可以在其他隐写术得到的载密图像组成的测试集上取得不错的结果. 而 S-UNIWARD 在隐写、隐写分析、隐写者检测任务中, 被认为是更为通用的设置, 因而常常拿它来作为训练模型中载密数据所使用的隐写术, 在本文的重要对比方法 SAGCN 中, 也采用其作为隐写术未知情况下的实验设置, 综合上述原因, 本文也采用了这样的设置. 在表 6 中, 添加了使用 HILL 作为训练数据的隐写术实验. 选择 HILL 的原因如下: 1) 在隐写术中, HILL 被认为是安全性能最高, 也是最难攻破的隐写术, 各种隐写分析算法在其上的准确率较低; 2) 与 S-UNIWARD 相比, HILL 的嵌入概率图完全不同, 因此, 可以进一步检测本文所提方法的泛化能力. 具体而言, 在训练阶段, 隐写者使用 HILL 作为图像隐写术, 隐写者检测模型学习在正常用户中检测使用 0.5 bpp、0.4 bpp、0.3 bpp、0.2 bpp、0.1 bpp 或 0.05 bpp 嵌入率对分享图像嵌入秘密信息的隐写者. 这里, 隐写者分享的图像均为载密图像, 正常用户分享的图像均为载体图像. 在测试阶段, 隐写者使用 HUGO-BD、WOW、HILL、MiPOD 等多种隐写术, 嵌入率为 0.2 bpp. 在每次实验中, 模型需要从 100 名用户中找出 1 名隐写者. 其中, 每名用户分享 200 张图像, 根据实验的不同, 隐写者分享的载密图像可能占比其分享的所有图像的 10% 或 30%. 本文进行 100 次实验, 并记录模型在检测隐写者过程中的平均准确率.

为更好地展示本文示例包表征过程与结果, 对

表 6 训练模型使用 HILL 作为隐写术, 分享的载密图像数量占比 10% 或 30%, MILGCN 取得的隐写者检测准确率 (%)

Table 6 Steganography detection accuracy rate (%) when the steganography used for training is HILL and the number of shared secret images is 10% or 30% of the total number of images

载密图像比例	测试隐写术			
	HUGO-BD	WOW	HILL	MiPOD
10%	9	6	7	4
30%	37	48	49	47

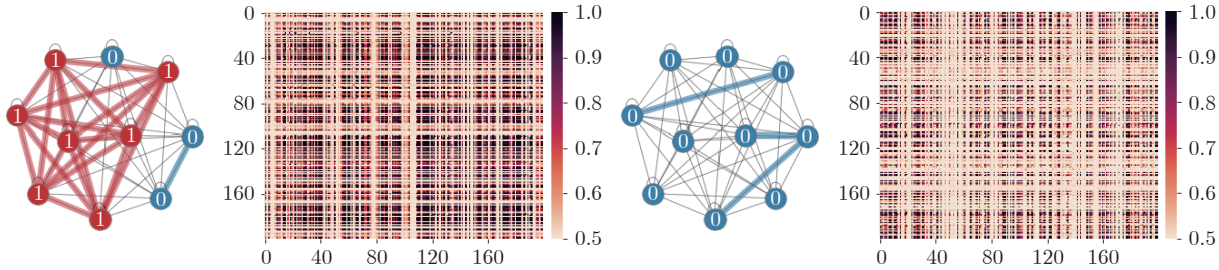
用户所对应示例包的图表征进行可视化. 首先, 绘制不同用户对应示例包的图结构示意图. 如图 4(a) 所示, 在隐写术未知情况下对使用比例嵌入策略的隐写者检测的实验中, 随机选择一名使用 MiPOD 作为隐写术的隐写者, 其分享的载密图像占比 70%. 绘制这名用户分享图像所对应示例包的图结构示意图. 其中, 节点表示分享的图像, 节点上的数字 1 表示该图像为载密图像, 数字 0 表示该图像为载体图像. 节点间使用边相连, 带有颜色的加粗边表示具有紧密关联关系的图像. 通常, 载密图像之间关联紧密 (对应红色加粗边), 都具有隐写所带有的模式特征, 部分正常图像可能由于具有相同的纹理 (对应蓝色加粗边), 也相互关联. 这时, 该图结构在第一层图卷积中的邻接矩阵如图 4(b) 所示. 其中, 深红色代表较大的邻接矩阵元素值, 白色代表较小的邻接矩阵元素值. 可以发现, 当用户为隐写者且分享的载密图像较多时, 模型能够学习构建邻接矩阵, 将较多的载密图像相互关联, 提取和增强其共有的正示例模式特征. 为与隐写者的图结构进行对比, 随机选择一名正常用户进行可视化. 如图 4(c) 所示, 正常用户分享的图像都为载体图像 (用数字 0 标注), 部分载体图像间具有相同的模式特征. 该图结构在第一层图卷积中的邻接矩阵如图 4(d) 所示, 可以发现, 相比于隐写者对应的图结构, 正常用户的图结构中因为不存在载密图像, 因此节点间的关联关系较弱, 部分载体图像间存在关联关系, 共同增强了负示例特征. 可视化结果表明本文所提模型能够按照预期工作, 从而区分隐写者和正常用户.

除此以外, 本文还进一步测试将隐写者检测问题形式化为分类问题情况下的正检率和误检率

$$\text{正检率} = \frac{TP}{TP+FP} \tag{14}$$

$$\text{误检率} = \frac{FP}{FP+TN} \tag{15}$$

其中, TP 表示模型预测为隐写者中的真实隐写者数量, FP 表示模型预测为隐写者中的正常用户数量, TN 表示模型预测为正常用户中的正常用户数量. 具体而言, 在训练阶段隐写者使用 S-UNIWARD 作为图像隐写术, 嵌入率为 0.4 bpp, 在测试阶段隐写者分享的载密图像占比为 0%、50% 和100%, 阈值为 0.5、0.7、0.9, 本文使用被测方法预测每名用户属于隐写者的概率, 并将概率高于阈值的用户输出为隐写者. 可以发现, 当隐写者分享的载密图像占比为 0% 时, 正检率都不存在 (因为此时隐写者也表现为正常用户, 用户中不存在应当被检测到的隐写者), 阈值为 0.5 时, 误检率为 7.98%; 阈值为 0.7 时, 误检率为 0.87%; 阈值为 0.9 时, 误检率为 0%. 当隐写者分享的载密图像占比为 50% 时, 阈值为 0.5 时, 正检率为 10.29%, 误检率为 8.10%; 阈值为 0.7 时, 正检率为 37.33%, 误检率为 0.95%; 阈值为 0.9 时, 正检率为 NaN (表示非数), 误检率为 0%. 当隐写者分享的载密图像占比为 100% 时, 阈值为 0.5 时, 正检率为 10.79%, 误检率为 8.34%; 阈值为 0.7 时, 正检率为 58.82%, 误检率为 0.71%; 阈值为 0.9 时, 正



(a) 隐写者的示例包的图结构示意图
(a) Graph structure of the bag of instances for a steganographer

(b) 隐写者在第一层图卷积中的邻接矩阵可视化
(b) Visualization of the adjacency matrix of a steganographer in the first layer of graph convolution

(c) 正常用户的示例包的图结构示意图
(c) Graph structure of the bag of instances for a normal user

(d) 正常用户在第一层图卷积中的邻接矩阵可视化
(d) Visualization of the adjacency matrix of a normal user in the first layer of graph convolution

图 4 隐写者和正常用户所对应图结构的可视化

Fig. 4 Visualization of graph structures corresponding to steganographer and normal user

检率为 NaN, 误检率为 0%. 从该结果可以发现, 当隐写者分享的载密图像占比为 50% 和 100% 时, 模型的正检率随着阈值的提高而提高, 当阈值为 0.7 时, 模型的正检率以较大差值超过了阈值为 0.5 时的正检率. 这主要受益于本文提出的网络结构和损失函数设计, 正常用户被预测为隐写者的概率相对较低, 随着阈值的提高这些用户被判别为正常用户, 而隐写者被预测为隐写者的概率相对较高, 尽管阈值提高, 这些用户仍然能被判别为隐写者. 而当阈值为 0.9 时, 模型将所有用户预测为正常用户, 此时 $TP + FP$ 为零, 故正检率不为实数. 这意味着虽然相对于正常用户, 隐写者被正确预测为隐写者的概率较高, 但当阈值提高到 0.9 时, 模型对隐写者的预测概率小于阈值. 这是由于相对于丰富的图像内容而言, 图像中嵌入的秘密信息难以被识别, 模型很难以较高的置信度来区分隐写者与正常用户. 值得注意的是, 将其形式化为分类问题时首先需要判定未知用户是否存在隐写者, 并在某个阈值水平下正确检测出隐写者. 这需要对隐写者预测的置信度具有较高的绝对水平, 而原有的隐写者检测定义只需要模型在隐写者预测的置信度上具有较高的相对水平 (相对于正常用户) 即可. 因此, 使用分类问题来度量更加困难, 计划在未来工作中进一步研究和探讨如何提升隐写者预测置信度的绝对水平.

3.2 基于频域的隐写者检测性能评估与对比

在本节中, 本文进一步考虑隐写者检测中的跨域问题. 现有工作通常使用 BOSSbase ver 1.0.1 的 JPEG 版本来检测当图像隐写者使用频域隐写术在 JPEG 图像的频率系数中嵌入信息时不同隐写者检测方法的检测性能. 因此, 在本节中, 将上文中的实验设置扩展到频率域. 具体而言, 首先将 BOSSbase ver 1.0.1 中的图像切分成 4 个互不重叠的子图, 每张子图的尺寸为 256×256 像素. 进一步, 利用 MATLAB 中的 imwrite 函数, 使用 JPEG 算法在 80 质量因子参数下对每张被切分的子图进行压缩, 并使用频域图像隐写术生成对应的载密图像. 在隐写者检测过程中, 与 Holub 等[21] 在其工作中对 JPEG 域中图像的隐写分析过程一致, 输入图像被解压缩至空域, 接着 MDNNSD 被用于从解压缩的图像中提取特征. 与在空域中的实验设置相似, 在训练阶段, 首先从所有子图样本中随机选择 20000 张载体图像, 并使用 J-UNIWARD 隐写术在不同嵌入率的有效载荷下生成载密图像, 特别地, 本文中使用的有效载荷包括 0.4 bpnzAC、0.3 bpnzAC、0.2 bpnzAC、0.1 bpnzAC 和 0.05 bpnzAC. 进而, 本文将得到的载体图像和载密图像组合为用户作为训练样本. 在训练阶段的每个批次中, 本文采样 50 名正常用户, 每名正常用户分享 200 张载体图像, 并采样 50 名图像隐写者, 每名隐写者分享 200 张载密图像, 每轮训练采样 100 个批次的用户对模型进行优化. 在测试阶段, 本文在每个实验中随机选择 100 个用户, 每个用户分享 200 张图像, 隐写者检测模型被要求从 100 个用户中检测出唯一的 1 名隐写者. 因此, 本文在 20000 张载体图像中随机选择 200 张载体图像对应的载密图像, 构成 1 名图像隐写者, 剩余的 19800 张载体图像被随机选择并分配给 99 个正常用户. 在此基础上, 本文统计不同隐写者检测模型的检测准确率, 所有统计实验进行 100 次, 最终得到所有统计结果的平均值.

3.2.1 已知隐写术情况下的隐写者检测

为验证在隐写者检测人员已知隐写者使用的图像隐写术时本文方法的有效性, 在本节实验中保持测试阶段隐写者使用的图像隐写术与训练阶段相同, 使用不同模型对使用不同嵌入率在图像中嵌入秘密信息的隐写者进行检测并记录检测模型的性能. 具体而言, 在测试阶段, 隐写者分享的每张图像中只包含使用 J-UNIWARD 图像隐写术嵌入的秘密信息. 同时, 与空域隐写者检测的实验设置类似, 本文在 5 种不同的有效载荷上进行测试, 其中包括 0.4 bpnzAC、0.3 bpnzAC、0.2 bpnzAC、0.1 bpnzAC 和 0.05 bpnzAC.

如表 7 所示, 本文首先对比隐写者检测与隐写分析领域的相关工作. 其中, JRM 富特征提取模型和 PEV-274 特征提取模型是经典的用于 JPEG 图像的隐写分析模型, 在隐写分析任务中取得了良好的性能. 本文将 JRM 与层次聚类算法结合在一起, 构成基于 JRM 的隐写者检测框架 JRM_SD, 以评估其在隐写者检测任务上的性能. 同时, 本文也将 PEV 与层次聚类算法结合在一起, 构成基于 PEV 的隐写者检测框架 PEV_SD, 进行对比分析. 从表 7 中可以看出, PEV_SD 无法有效地检测出使用 J-UNIWARD 作为图像隐写术的隐写者, 这是由于 PEV_SD 方法基于 PEV-274 特征进行分析, 无法检测出使用内容自适应图像隐写术 J-UNIWARD 嵌入秘密信息的载密图像并构建具有区分力的特征. 与 PEV_SD 相比, JRM_SD 能够更有效地攻击使用 J-UNIWARD 作为图像隐写术的隐写者, 但检测性能仍然有限. 当隐写者使用 0.4 bpnzAC 嵌入率嵌入秘密信息时, 尽管 JRM_SD 联合使用了最前沿的富特征提取模型和一个优化检测器, 也只能获得接近 50% 的检测准确率.

接着, 本文进一步将图神经网络领域的相关工

表 7 已知隐写者使用相同图像隐写术 (J-UNIWARD) 时的隐写者检测准确率 (%), 嵌入率从 0.05 bpnzAC 到 0.4 bpnzAC

Table 7 Steganography detection accuracy rate (%) when steganographer use the same image steganography (J-UNIWARD) and the embedding payload is from 0.05 bpnzAC to 0.4 bpnzAC

模型	嵌入率 (bpnzAC)				
	0.05	0.1	0.2	0.3	0.4
JRM_SD	11	17	25	31	48
PEV_SD	0	0	1	1	5
GraphSAGE	13	68	100	100	100
AGNN	13	84	100	100	100
GCN	16	88	100	100	100
SAGCN	17	92	100	100	100
MILGCN	**25**	**92**	**100**	**100**	**100**

作应用在 JPEG 图像的隐写者检测任务中, 对比图神经网络领域的相关工作. 从表 7 中可以发现, 基于图神经网络的隐写者检测方法的性能远超于现有隐写者检测与隐写分析领域相关工作的性能, 表明图神经网络在利用用户分享的图片间的关联关系进行隐写者检测的有效性. 其中, SAGCN 仅次于本文提出的 MILGCN, 取得了超过其他基于图神经网络的隐写者检测方法的性能. 与之相对的, 本文提出的 MILGCN 在隐写者使用较低的嵌入率嵌入信息的困难情况下, 相比 SAGCN 的性能有了进一步提升. 这是由于当嵌入率较低时, 隐写者与正常用户的特征分布极为相近, 而本文提出的 MILGCN 通过基于多示例学习的设计增加了其区分性. 具体而言, 本文提出的共性增强图卷积网络能够有效增加正示例的共性特征, 在此基础上, 基于注意力机制的图表征能够根据识别的正示例模式特征构建具有区分力的图表征, 进而正确检测出隐写者.

此外还测试了使用比例策略进行嵌入时的情况, 实验的训练部分与本节的实验设置相同, 测试部分: 隐写者使用 J-UNIWARD 隐写术, 嵌入率为 0.2 bpnzAC, 分享的载密图像占其分享的所有图像的 10% 、30% 、50% 、70% 、90% 或 100%. 本文进行 100 次实验, 并记录模型在检测隐写者过程中的平均准确率. 可以发现, 在载密图像数量占比大于等于 70% 的情况下, 本文的方法都能达到 100% 的准确率; 在占比 50% 的情况下, 准确率为 94%; 占比 30% 的情况下, 准确率为 47%; 占比 10% 的情况下, 准确率为 30%. MILGCN 的性能随着隐写者分享的载密图像占比的升高而升高, 这种现象与空域上的实验结果一致, 是由于当隐写者分享的载密图像占比较低时, 隐写者分享的图像大部分为自然图像, 即载体图像. 在这种情况下, 隐写者的表征将与正常用户非常接近, 模型难以正确区分隐写者与正常用户. 反之, 随着用户分享的载密图像占比升高, 隐写者的表征与正常用户差异变大, 模型将能够更好地区分二者. 此外, 相比于检测使用空域隐写术的隐写者, 可以发现在检测使用频域隐写术的隐写者时, 模型只能够在载密图像占比较高时保持较高的检测性能, 当载密图像占比较低时, 模型的检测准确率会产生更大的下降. 这是由于频域隐写术的检测比空域隐写术的检测更为困难, 因此在载密图像占比较低时, 隐写者与正常用户的区分性会比空域隐写术设置下更低.

3.2.2 隐写术未知情况下的隐写者检测

正如在上文中提到的, 隐写者检测人员在对社交网络中用户分享的图像进行分析时, 无法预先知道隐写者使用的图像隐写术. 因此, 本文放宽了图像隐写者在训练阶段和测试阶段使用的图像隐写术为相同隐写术的假设, 探讨在隐写术错配情况下本文所提模型的检测性能. 具体而言, 在训练阶段隐写者使用 J-UNIWARD 作为隐写术嵌入秘密信息, 以此设置进行训练得到模型. 在测试阶段, 隐写者分享的每张图像中包含与训练阶段使用不同的图像隐写术, 如 nsF5、UERD. 同时, 与空域隐写者检测的实验设置类似, 本文在 5 种不同的有效载荷上进行测试, 其中包括 0.4 bpnzAC、0.3 bpnzAC、0.2 bpnzAC、0.1 bpnzAC 和 0.05 bpnzAC.

在表 8 中, 可以发现, 当有效载荷在 0.1 ~ 0.4 bpnzAC 区间时, 尽管在测试阶段隐写者使用的图像隐写术与训练阶段不同, 但是无论隐写者使用哪种隐写术, 本文方法仍能够获得超过 90% 的检测准确率. 通过将表 7 与表 8 进行对比, 可以发现在隐写术错配时, 隐写者检测模型的性能甚至超过了在训练数据集和测试数据集中使用相同的 J-UNIWARD 作为图像隐写术时的性能. 这是由于 J-UNIWARD 作为近年来提出的前沿的内容自适应图像隐写术, 相比于 nsF5 和 UERD 更难攻破, 因此, 当 J-UNIWARD作为图像隐写术时, 载密图像与载体图像的特征分布极为相近、难以区分, 不利于基于特征分布构建有效的图结构并提取载密图像的公共模式特征. 相比而言, 使用 nsF5 和 UERD 作为图像隐写术嵌入秘密信息的载密图像的特征分布与载体图像的特征分布差异较大. 特别地, 还能够发现在隐写术错配的情况下, 本文方法也能超过其他隐写者检测方法取得最好的性能, 这说明 MILGCN 确实能够有效提取正示例共享的模式特征, 即识别的载密图像模式特征, 并与载体图像的

表 8 当测试阶段隐写者使用 nsF5 或 UERD 等图像隐写术嵌入秘密信息时, 不同方法的隐写者检测准确率 (%), 嵌入率从 0.05 bpnzAC 到 0.4 bpnzAC

Table 8 Steganography detection accurate rate (%) of different methods when steganographer uses nsF5 or UERD as image steganography in the testing phase and the embedding payload is from 0.05 bpnzAC to 0.4 bpnzAC

隐写术	模型	嵌入率 (bpnzAC)				
		0.05	0.1	0.2	0.3	0.4
nsF5	PEV_SD	0	1	9	52	93
	GraphSAGE	21	91	100	100	100
	AGNN	20	90	100	100	100
	GCN	24	90	100	100	100
	SAGCN	**29**	**92**	100	100	100
	MILGCN	22	90	**100**	**100**	**100**
UERD	GraphSAGE	25	91	100	100	100
	AGNN	29	94	100	100	100
	GCN	33	96	100	100	100
	SAGCN	33	98	100	100	100
	MILGCN	**42**	**99**	**100**	**100**	**100**

模式特征区分开.

当隐写者使用的有效载荷为 0.05 bpnzAC 时, 尽管本文提出的方法在性能上稍强于其他基于图的方法, 但所有方法都无法有效检测出隐写者. 这是由于秘密信息在 JPEG 的压缩过程中丢失了, 当有效载荷较低时, 载密图像与载体图像非常相似. 因此, 对于模型而言, 很难区分隐写者和正常图像的特征. 与空域中隐写者检测方法相比, 频域中隐写者检测框架的性能相对较低, 这也说明在压缩到 JPEG 格式过程中, 存在信息的丢失.

3.2.3 计算复杂度分析

在这一部分, 提供本文方法 MILGCN 和其他几种对比方法的计算复杂度分析. 首先使用 pt-flops 软件包计算不同方法的计算复杂度, 并使用浮点运算数量来度量. 具体来说, 给定一个模型, 构建分享 200 张图像的 100 个用户样本作为输入, 计算其中哪一个用户为隐写者. 在计算过程中, 对于指数、开根号、乘法、加法、比较等操作, 均将其计数为一次浮点运算, 而对于数据复制、移动等操作, 则不进行计数. 最终, 得到计算单个样本的浮点运算次数并进行统计报告. 除此以外, 还度量不同方法在推断过程中所需的 CPU 时间, 以说明设计方法在推理和部署方面的可用性. 这里, 本文运行 200 个批次样本的推理, 并统计样本推理所需的平均时间. 其中, 每个批次包含 100 个用户样本. 表 9 中, 还提供了不同模型所包含的参数量.

表 9 计算复杂度分析

Table 9 The analysis of computational complexity

方法名称	批次平均运行时间 (s)	单个样本浮点运算数 (千兆次)	参数量 (千个)
MILNN	0.001	0.003	12.92
GCN	0.830	2.480	67.97
SAGCN	2.210	7.410	67.94
MILGCN	0.020	0.070	74.18

从表 9 中可以发现, 尽管本文提出的 MILGCN 具有相对较多的参数, 但仍与基于图的方法 GCN 和 SAGCN 保持在同一量级上. 与此同时, 还可以发现, 虽然参数量相对较大, 但 MILGCN 所需的浮点运算数和批次平均运行时间要比基于图的方法 GCN 和 SAGCN 小很多. 这是由于这些方法在构建图的过程中, 使用欧氏距离度量特征集合中两两图像特征对的关联关系, 具有较高的时间复杂度. 而本文通过自适应的学习方法取代了基于欧氏距离的图构建方法, 大幅降低了运算量. 这使得本文提出的 MILGCN 不但在检测性能上占有优势, 也在运行效率上远超其他基于图的方法, 接近于简单模型 MILNN 的水平.

4 结论与未来工作

近年来, 隐写者在分散秘密信息嵌入图像的策略选择上越来越多样. 因此, 本文扩展基于图的隐写者检测方法, 进一步探究应对不同嵌入策略的通用隐写者检测方案. 本文提出一种基于多示例学习图卷积网络的隐写者检测算法, 将隐写者检测形式化为多示例学习任务. 在多示例学习任务的形式化下, 用户分享的图像对应于示例, 用户对应于示例包. 在此基础上, 本文设计多示例学习图卷积网络, 在正示例占比较低或特征分布与负示例相近的情况下, 能够识别和区分正示例的模式特征. 其中, 本文设计的共性增强图卷积能够自适应地突出示例包中正示例的模式特征, 而注意力图读出模块能够自适应地构建示例包表征, 并根据具有区分力的表征对用户进行正确检测. 实验表明, 本文设计的模型能够对抗多种批量隐写术和隐写策略. 在未来工作中, 将研究当载密图像数量占比很小情况下有效的隐写者检测方法.

References

1 Ker A D, Pevný T. A new paradigm for steganalysis via clustering. In: Proceedings of the SPIE 7880, Media Watermarking, Security, and Forensics III. Bellingham, USA: SPIE, 2011. 312–324

2 Pevný T, Feidrich J. Merging Markov and DCT features for multi-class JPEG steganalysis. In: Proceedings of the SPIE 6505, Security, Steganography, and Watermarking of Multime-

dia Contents IX. Bellingham, USA: SPIE, 2007. 1−13

3 Ker A D, Pevný T. Identifying a steganographer in realistic and heterogeneous data sets. In: Proceedings of the SPIE 8303, Media Watermarking, Security, and Forensics. Bellingham, USA: SPIE, 2012. 1−13

4 Ker A D, Pevný T. The steganographer is the outlier: Realistic large-scale steganalysis. *IEEE Transactions on Information Forensics and Security*, 2014, **9**(9): 1424−1435

5 Li F Y, Wu K, Lei J S, Wen M, Bi Z Q, Gu C H. Steganalysis over large-scale social networks with high-order joint features and clustering ensembles. *IEEE Transactions on Information Forensics and Security*, 2016, **11**(2): 344−357

6 Zheng M J, Zhong S H, Wu S T, Jiang J M. Steganographer detection via deep residual network. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). Piscataway, USA: IEEE, 2017. 235−240

7 Zheng M J, Zhong S H, Wu S T, Jiang J M. Steganographer detection based on multiclass dilated residual networks. In: Proceedings of the ACM on International Conference on Multimedia Retrieval. New York, USA: ACM, 2018. 300−308

8 Zhang Z, Zheng M J, Zhong S H, Liu Y. Steganographer detection via enhancement-aware graph convolutional network. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). Piscataway, USA: IEEE, 2020. 1−6

9 Zhang Z, Zheng M J, Zhong S H, Liu Y. Steganographer detection via a similarity accumulation graph convolutional network. *Neural Networks: The Official Journal of the International Neural Network Society*, 2021, **136**: 97−111

10 Ning X, Tian W J, Yu Z Y, Li W J, Bai X, Wang Y B. HCFNN: High-order coverage function neural network for image classifican. *Pattern Recognition*, 2022, **131**: Article No. 108873

11 Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. In: Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS). New York, USA: Curran Associates Inc., 2016. 3844−3852

12 Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv: 1609.02907, 2016.

13 Hamilton W L, Ying R, Leskovec J. Inductive representation learning on large graphs. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS). New York, USA: Curran Associates Inc., 2017. 1025−1035

14 Ying R, You J X, Morris C, Ren X, Hamilton W L, Leskovec J. Hierarchical graph representation learning with differentiable pooling. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS). New York, USA: Curran Associates Inc., 2018. 4805−4815

15 Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. In: Proceedings of the 6th International Conference on Learning Representation. Vancouver, Canada: ICLR Press, 2018. 1−12

16 Gao W, Wan F, Yue J, Xu S C, Ye Q X. Discrepant multiple instance learning for weakly supervised object detection. *Pattern Recognition: The Journal of the Pattern Recognition Society*, 2022, **122**: Article No. 108233

17 Tang X C, Liu M Z, Zhong H, Ju Y Z, Li W L, Xu Q. MILL: Channel attention-based deep multiple instance learning for landslide recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2021, **17**(2s): 1−11

18 Yuan M, Xu Y T, Feng R X, Liu Z M. Instance elimination strategy for non-convex multiple-instance learning using sparse positive bags. *Neural Networks*, 2021, **142**: 509−521

19 Su Z Y, Tavolara T E, Carreno-Galeano G, Lee S J, Gurcan M N, Niazi M K K. Attention2majority: Weak multiple instance learning for regenerative kidney grading on whole slide images. *Medical Image Analysis*, 2022, **79**: Article No. 102462

20 Bas P, Filler T, Pevný T. "Break our steganographic system": The ins and outs of organizing BOSS. In: Proceedings of the International Workshop on Information Hiding. Prague, Czech Republic: Springer, 2011. 59−70

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

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

23 Shi H C, Dong J, Wang W, Qian Y L, Zhang X Y. SSGAN: Secure steganography based on generative adversarial networks. In: Proceedings of the Pacific Rim Conference on Multimedia. Harbin, China: Springer, 2017. 534−544

24 Pevný T, Somol P. Using neural network formalism to solve multiple-instance problems. arXiv preprint arXiv: 1609.07257, 2016.

25 Thekumparampil K K, Wang C, Oh S, Li L J. Attention-based graph neural network for semi-supervised learning. arXiv preprint arXiv: 1803.03735, 2018.

26 Xu G S, Wu H Z, Shi Y Q. Structural design of convolutional neural networks for steganalysis. *IEEE Signal Processing Letters*, 2016, **23**(5): 708−712

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

28 Filler T, Fridrich J. Gibbs construction in steganography. *IEEE Transactions on Information Forensics and Security*, 2010, **5**(4): 705−720

29 Holub V, Fridrich J. Designing steganographic distortion using directional filters. In: Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS). Piscataway, USA: IEEE, 2012. 234−239

30 Sedighi V, Cogranne R, Fridrich J. Content-adaptive steganography by minimizing statistical detectability. *IEEE Transactions on Information Forensics and Security*, 2016, **11**(2): 221−234

31 Fridrich J, Pevný T, Kodovský J. Statistically undetectable JPEG steganography: Dead ends challenges, and opportunities. In: Proceedings of the 9th Workshop on Multimedia and Security. Dallas, USA: ACM, 2007. 3−14

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

33 Qian Y L, Dong J, Wang W, Tan T N. Deep learning for steganalysis via convolutional neural networks. In: Proceedings of the SPIE 9409, Media Watermarking, Security, and Forensics. Bellingham, USA: SPIE, 2015. 1−10

34 Wei Xing, Li Jia, Sun Xiao, Liu Shao-Fan, Lu Yang. Cross-view image generation via mixture generative adversarial network. *Acta Automatica Sinica*, 2021, **47**(11): 2623−2636
(卫星, 李佳, 孙晓, 刘邵凡, 陆阳. 基于混合生成对抗网络的多视角图像生成算法. 自动化学报, 2021, **47**(11): 2623−2636)

35 Hu Ming-Fei, Zuo Xin, Liu Jian-Wei. Survey on deep generative model. *Acta Automatica Sinica*, 2022, **48**(1): 40−74
(胡铭菲, 左信, 刘建伟. 深度生成模型综述. 自动化学报, 2022, **48**(1): 40−74)

36 Dong Yin-Peng, Su Hang, Zhu Jun. Interpretability analysis of deep neural networks with adversarial examples. *Acta Automatica Sinica*, 2022, **48**(1): 75−86

(董胤蓬, 苏航, 朱军. 面向对抗样本的深度神经网络可解释性分析. 自动化学报, 2022, **48**(1): 75−86)

37 Yu Zheng-Fei, Yan Qiao, Zhou Yun. A survey on adversarial machine learning for cyberspace defense. *Acta Automatica Sinica*, 2022, **48**(7): 1625−1649
(余正飞, 闫巧, 周鋆. 面向网络空间防御的对抗机器学习研究综述. 自动化学报, 2022, **48**(7): 1625−1649)

38 Zhao Bo-Yu, Zhang Chang-Qing, Chen Lei, Liu Xin-Wang, Li Ze-Chao, Hu Qing-Hua. Generative model for partial multi-view clustering. *Acta Automatica Sinica*, 2021, **47**(8): 1867−1875
(赵博宇, 张长青, 陈蕾, 刘新旺, 李泽超, 胡清华. 生成式不完整多视图数据聚类. 自动化学报, 2021, **47**(8): 1867−1875)

39 Zhang Bo-Wei, Zheng Jian-Fei, Hu Chang-Hua, Pei Hong, Dong Qing. Missing data generation method based on flow model and its application in remaining life prediction. *Acta Automatica Sinica*, 2023, **49**(1): 185−196
(张博玮, 郑建飞, 胡昌华, 裴洪, 董青. 基于流模型的缺失数据生成方法在剩余寿命预测中的应用. 自动化学报, 2023, **49**(1): 185−196)

40 Wei G, Guo J, Ke Y Z, Wang K, Yang S, Sheng N. A three-stage GAN model based on edge and color prediction for image outpainting. *Expert Systems With Applications*, 2023, **214**: Article No. 119136

41 Wang Y F, Dong X S, Wang L X, Chen W D, Zhang X J. Optimizing small-sample disk fault detection based on LSTM-GAN model. *ACM Transactions on Architecture and Code Optimization*, 2022, **19**(1): 1−24

42 Boroumand M, Chen M, Fridrich J. Deep residual network for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 2019, **14**(5): 1181−1193

钟圣华 深圳大学计算机与软件学院副教授. 主要研究方向为多媒体内容分析, 情感脑机接口. 本文通信作者.
E-mail: csshzhong@szu.edu.cn
(**ZHONG Sheng-Hua** Associate professor at the College of Computer Science and Software Engineering, Shenzhen University. Her research interest covers multimedia content analysis and affective brain-machine interface. Corresponding author of this paper.)

张 智 深圳大学计算机与软件学院研究助理, 香港理工大学电子计算学系博士研究生. 主要研究方向为隐写者检测, 脑电信号分析.
E-mail: zhi271.zhang@connect.polyu.hk
(**ZHANG Zhi** Research assistant at the College of Computer Science and Software Engineering, Shenzhen University; Ph.D. candidate in the Department of Computing, The Hong Kong Polytechnic University. Her research interest covers steganographer detection and electroencephalography signal analysis.)