软件学报 ISSN 1000-9825, CODEN RUXUEW
2024,35(12):5671−5686 [doi: 10.13328/j.cnki.jos.007068] [CSTR: 32375.14.jos.007068]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于中心差分卷积和注意力的空域彩色图像隐写分析*

魏康康[1,2], 骆伟祺[1,2], 刘明林[3]

[1](中山大学 计算机学院, 广东 广州 510006)

[2](广东省信息安全技术重点实验室, 广东 广州 510006)

[3](郑州大学 网络空间安全学院, 河南 郑州 450002)

通信作者: 骆伟祺, E-mail: luoweiqi@mail.sysu.edu.cn

摘　要: 目前, 大多数已发表的图像隐写分析方法都是针对灰度图像设计的, 因此这些方法无法有效检测广泛应用于社交媒体的彩色图像. 为解决这一问题, 提出一种基于中心差分卷积和注意力增强的彩色图像隐写分析方法. 首先设计一个包含预处理, 特征提取和特征分类这 3 个阶段的主干流. 在预处理阶段, 对输入的彩色图像进行颜色通道分离, 并串联各通道经过 SRM 滤波后的残差图. 在特征提取阶段, 构建 3 个基于中心差分卷积的卷积块来提取更深层的隐写分析特征图. 在分类阶段, 使用全局协方差池化和带有丢弃操作的两个全连接层来对载体和载密图像进行分类. 此外, 为了进一步增强主干流在不同时期的特征表达能力, 在主干流的前期和后期分别引入一个残差空间注意力增强模块和一个通道注意力增强模块. 其中, 残差空间注意力增强模块首先使用 Gabor 滤波核对输入图像进行通道分离卷积再串联相应的残差, 然后通过空间注意力机制获取残差特征图的有效信息. 而通道注意力增强模块则通过获取通道间的依赖关系来增强模型最后的特征分类能力. 进行大量的对比实验, 结果表明所提出方法可以显著提高对彩色图像隐写的检测性能, 并取得当前最好的结果. 此外, 还进行相应的消融实验来验证所提出的网络架构的合理性.

关键词: 隐写分析; 隐写; 彩色图像; 卷积神经网络; 注意力机制

中图法分类号: TP391

中文引用格式: 魏康康, 骆伟祺, 刘明林. 基于中心差分卷积和注意力的空域彩色图像隐写分析. 软件学报, 2024, 35(12): 5671−5686. http://www.jos.org.cn/1000-9825/7068.htm

英文引用格式: Wei KK, Luo WQ, Liu ML. Spatial Color Image Steganalysis Based on Central Difference Convolution and Attention. Ruan Jian Xue Bao/Journal of Software, 2024, 35(12): 5671−5686 (in Chinese). http://www.jos.org.cn/1000-9825/7068.htm

Spatial Color Image Steganalysis Based on Central Difference Convolution and Attention

WEI Kang-Kang[1,2], LUO Wei-Qi[1,2], LIU Ming-Lin[3]

[1](School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China)

[2](Guangdong Key Laboratory of Information Security Technology, Guangzhou 510006, China)

[3](School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450002, China)

Abstract: Currently, most of the published image steganalysis methods are designed for grayscale images, which cannot effectively detect color images widely used in social media. To solve this problem, this study proposes a color image steganalysis method based on central difference convolution and attention enhancement. The proposed method first designs a backbone flow consisting of three stages: preprocessing, feature extraction, and feature classification. In the preprocessing stage, the input color image is color channel-separated, and the residual images after SRM filtering are concatenated through each channel. In the feature extraction stage, the study constructs three convolutional blocks based on central difference convolution to extract deeper steganalysis feature maps. In the classification stage, the

* 基金项目: 国家自然科学基金 (61972430)
收稿时间: 2023-04-08; 修改时间: 2023-07-06, 2023-09-06; 采用时间: 2023-10-10; jos 在线出版时间: 2024-01-24
CNKI 网络首发时间: 2024-01-26

© 中国科学院软件研究所 http://www.jos.org.cn

study uses global covariance pooling and two fully connected layers with dropout operation to classify the cover and stego images. Additionally, to further enhance the feature expression ability of the backbone flow at different stages, it introduces a residual spatial attention enhancement module and a channel attention enhancement module at the early and late stages of the backbone flow, respectively. Specifically, the residual spatial attention enhancement module first uses Gabor filter kernels to perform channel-separated convolution on the input image and then obtains the effective information of the residual feature map through the spatial attention mechanism. The channel attention enhancement module enhances the final feature classification ability of the model by obtaining the dependence relationship between channels. A large number of comparative experiments have been conducted, and the results show that the proposed method can significantly improve the detection performance of color image steganography and achieve the best results currently. In addition, the study also conducts corresponding ablation experiments to verify the rationality of the proposed network architecture.

Key words: steganalysis; steganography; color image; convolutional neural network (CNN); attention mechanism

与传统的加密技术不同, 数字隐写在不明显改变人们视听觉感知的情况下, 将秘密信息隐藏于数字媒体中, 并通过公开信道将含有秘密信息的媒体进行传输, 以实现隐秘通讯的目的[1]. 随着互联网及数字图像处理技术的飞速发展, 以数字图像为载体的隐写研究已引起国内外学者的高度关注. 然而, 如果隐写术被不法分子利用, 将可能引发严重后果. 作为隐写的对立面, 隐写分析拟通过检测由于隐写引入的嵌入痕迹, 以发现带有秘密信息的图像. 隐写和隐写分析两者相互竞争的同时相互促进. 由于现代图像隐写方法[2−7]可根据图像内容和各颜色通道间的相互关系自适应地进行信息嵌入, 从而极大提升了隐写分析的检测难度.

早期隐写分析主要基于人工经验来设计隐写分析特征, 最为典型的方法是利用富模型[8]提取图像各种残差的统计特性, 然后结合集成分类器进行隐写分析. 随着深度学习技术的发展, 近年来研究人员将其应用到图像隐写分析. 基于深度学习的隐写分析方法无需过度依赖人工经验, 而是以数据驱动的方式构建一个端到端的网络来实现分类. 其中, 最为常见的方法[9−16]是利用多层卷积层来提取图像残差特征和全连接层来进行特征分类的结构, 取得了比传统隐写分析方法更好的检测性能.

然而, 目前绝大多数基于深度学习的隐写分析网络都是针对灰色图像设计, 因而无法有效检测彩色图像的隐写. 我们通过大量对比实验工作表明: 直接将已有的灰度图像隐写分析器用于检测彩色载密图像, 其检测性能往往不够理想. 其主要原因在于彩色图像隐写嵌入过程会不可避免地改变了各个颜色通道间的关系, 而基于灰度图像的隐写分析器并不能充分发掘隐写前后彩色通道间的统计差异. 从目前研究现状看, 针对彩色图像的隐写分析的工作相对较少. 另外, 已有的隐写分析方法对于检测彩色图像隐写的检测准确率仍然没有达到较好的结果. 因此, 彩色图像隐写分析存在着较大的性能提升空间.

为了有效提升彩色隐写分析方法的检测准确率, 本文提出了一种基于中心差分卷积和注意力增强的彩色图像隐写分析模型. 与现有相关彩色图像隐写分析模型所使用的普通卷积层, 全局平均池化层和全连接层等结构进行构建模型不同, 所提出的方法主要采用中心差分卷积层和全局协方差池化层等结构来进行隐写分析特征提取和分类, 此外, 所提方法同时引入了残差空间注意力和通道注意力增强模块来有效地提升模型对彩色隐写图像的检测表现. 本文的主要贡献如下.

1) 引入中心差分卷积应用于彩色图像隐写分析, 基于中心差分卷积设计了一个由预处理阶段, 特征提取阶段和特征分类阶段构成的彩色图像隐写分析模型主干流.

2) 引入注意力机制到彩色图像隐写分析以进一步增强主干流的特征表达能力. 在主干流的前期和后期, 我们分别构建了一个残差空间注意力增强模块和一个通道注意力增强模块进行合并.

3) 为了验证本文方法的有效性, 在彩色图像数据集 ALASKA II 的实验结果表明所提出的彩色图像隐写分析方法对于多种彩色图像隐写算法能达到目前最优的检测准确率. 另外, 我们也设置了相应的消融实验来验证所设计的隐写分析模型的合理性.

本文第 1 节介绍隐写分析的相关方法和研究现状. 第 2 节详细介绍本文构建的基于中心差分卷积和注意力的彩色图像隐写分析模型. 第 3 节介绍所用的数据集和应用细节以及通过对比实验验证所提模型的有效性. 最后, 第 4 节总结全文并给出未来的工作展望.

© 中国科学院软件研究所 http://www.jos.org.cn

1 隐写分析相关工作

现阶段的图像隐写分析任务主要是用来判别数字图像中是否含有隐写嵌入的秘密信息, 因此本质上是一个图像二分类任务[17,18]. 图像隐写分析最初依赖人工经验对图像进行统计分析建模来获取手工隐写分析特征, 而后结合集成分类器进行分类. 近年来随着研究人员将深度学习方法应用到隐写分析, 不需要过度依赖人工经验, 而是以数据驱动的方式来构建一个端到端的深度网络进行分类. 根据图像隐写分析的发展历程, 已有的图像隐写分析方法可以分为基于手工特征的传统方法和基于深度学习的方法两类. 接下来我们将分别从这两类方法出发, 详细介绍这两类方法中的典型算法.

1.1 基于手工特征的隐写分析方法

传统的隐写分析方法一般将特征提取和分类作为两个独立的部分分别进行设计. 首先通过对图像中因为隐写引入的嵌入信息伪迹产生变化的统计量来构建隐写分析特征, 而后将构建的特征通过集成分类器进行最后的分类. Fridrich 等人[8]提出一种富模型方法, 首次将多种线性和非线性高通滤波核对图像进行建模得到相应的残差图像并对其进行截断和量化操作, 然后对处理后的残差图像的共生矩阵进行组合构建出了 34 671 维的 SRM 特征, 取得了良好的分类表现. 而后, Holub 等人[19]将 SRM 中提取出来的噪声残差图像进行随机投影, 使用投影后的直方图构建出了 12 870 维的 PSRM 隐写分析特征, 相比于 SRM 降低了维数的同时增加了检测准确率.

上述方法都是基于灰度图像而设计的, 并且它们不能直接用来检测彩色图像隐写方法生成的载密图像. 因此, 研究人员引入了对相邻彩色通道之间像素提取特征的彩色图像隐写分析方法. 首先, Goljan 等人[20]提出了第 1 个彩色图像富模型隐写分析特征集 CRMQ1, 该方法是对 SRM 的扩展并加入了由颜色通道计算得到的共生矩阵统计特征来捕捉彩色通道间像素相关性. 另外, 该方法的特征维度只有 5 404 维, 远小于 SRM, 在 BOSSBase[21]上的实验表明了该方法相比于对独立和合并颜色通道进行检测的 SRM 表现更好. 随后, Abdulrahman 等人[22]提出了一种在 CRMQ1 的基础上添加新的特征的方法 SGRM, 所添加的特征通过应用可转向的高斯滤波器然后计算像素的共生矩阵得到, 其中高斯滤波器在 18 种不同方向上倾斜来获取不同的特征. 该方法与 CRMQ1 相比, 虽然特征维度更大达到了 22 563 维, 但对于多种自适应隐写算法都取得了更好的检测性能.

1.2 基于深度学习的隐写分析方法

随着深度学习技术的发展, 研究人员开始尝试将深度学习方法应用到图像隐写分析任务并提出了多种隐写分析模型[9−16]. 在 2014 年, Tan 等人[9]分析了 SRM 与卷积神经网络之间的联系, 首次提出一个包含 9 层卷积结构的隐写分析网络 TanNet. 该网络首先使用了卷积层来模拟 SRM 滤波器组, 尽管该网络的性能略差于 SRM, 但给后续的基于深度学习的隐写分析研究奠定了一定的基础. Boroumand 等人[12]提出了一种通用的残差网络 SRNet 应用于隐写分析, 该网络首次使用了 64 个随机初始化的滤波核代替预定义的高通滤波器来对输入图像提取残差信息, 并取得了比先前的隐写分析器都要更好的检测表现. Deng 等人[13]将全局协方差池化引入隐写分析任务并提出了一个轻量级的隐写分析模型 CovNet, 其检测性能在多数情况下优于 SRNet 并且训练耗时远小于 SRNet.

除了上述针对于灰度图像的隐写分析模型, 基于深度学习面向于彩色图像的隐写分析方法也相继被提出. Zeng 等人[23]首次提出一种具有宽卷积结构的彩色图像隐写分析网络 WISERNet, 该方法在底层不采用常规卷积而是引入逐通道卷积来获取高通残差, 然后经过 3 个卷积层和 3 个全连接层组成的结构进行分类. 该方法在 BOSSBase 数据集上进行了充足的实验, 结果表明该模型有着比传统彩色隐写分析方法 CRMQ1 和多种灰度图像隐写分析网络更好的检测效果. 最近, Wei 等人[24]提出了一种适用于空域彩色图像和 JPEG 彩色图像的通用彩色图像隐写分析网络 UCNet, 该方法在预处理阶段将各个颜色通道的高通残差进行串联, 有效保存了隐写嵌入修改. 另外, 该模型设计了 3 种不同层级结构的卷积层来进行特征提取, 在多种彩色图像隐写体系下的实验结果表明该方法相比于已有的隐写分析方法都有着目前最优的检测准确率.

2 基于中心差分卷积和注意力的彩色图像隐写分析

如图 1 所示, 其中, ©表示串联操作, ⊙表示逐元素相乘再相加, ⊗表示逐元素相乘, ⊕表示逐元素相加, C, H, W

© 中国科学院软件研究所 http://www.jos.org.cn

分别表示通道数, 特征图的高度和宽度, dropout 为丢弃操作. 本文提出的彩色图像隐写分析模型主要由 3 个部分组成, 分别是模型的主干流, 残差空间注意力增强模块和通道注意力增强模块. 下面我们将对所提模型的各部分进行详细介绍.

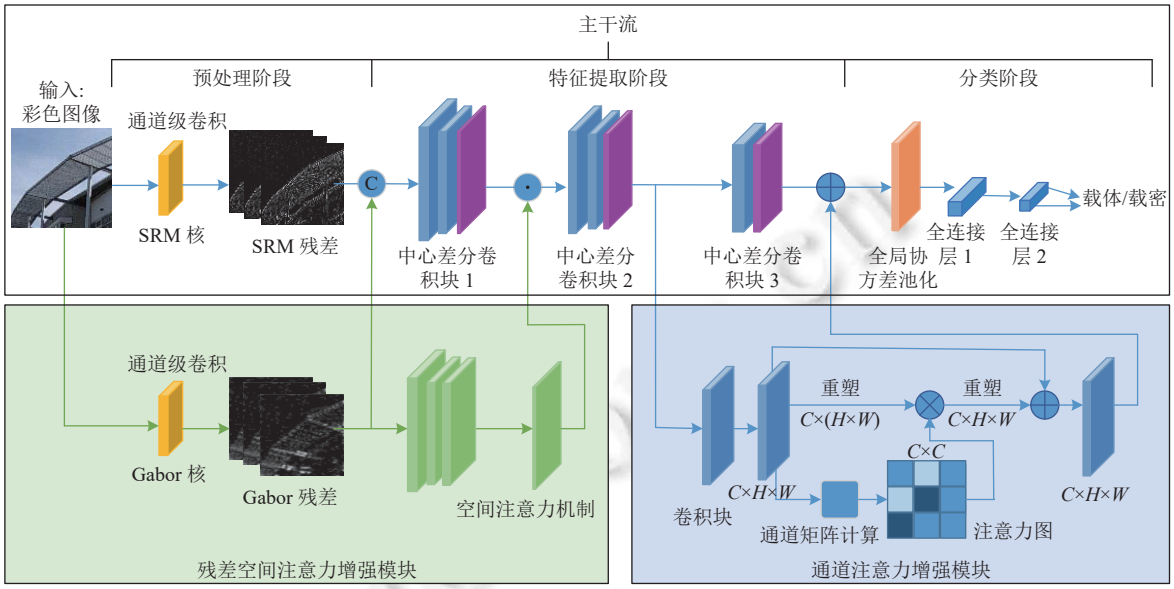


图 1 基于中心差分卷积和注意力增强的彩色图像隐写分析模型

2.1 所提模型的主干流

在本节, 首先我们介绍所提出的模型主干流, 模型主干流包含 3 个阶段, 分别是预处理阶段, 特征提取阶段和最后的分类阶段. 预处理阶段目的是用来获取各个通道的高通残差并保留更多的隐写嵌入修改, 特征提取阶段则采用了基于中心差分卷积结构的卷积层来提取更深层次的隐写分析特征, 而分类阶段则是使用池化层和全连接层来对隐写分析特征进行分类结果输出. 对这 3 个阶段, 具体的结构和实现细节分别如下.

● 主干流的预处理阶段

预处理阶段主要用来提取图像的高频信息, 提高输入图像的信噪比, 需要注意的是, 这里的信息表示由隐写嵌入信息后引入的伪迹, 噪声则表示图像自身内容. 一般来说, 由于隐写在进行嵌入时对图像像素值的修改会破坏相邻像素间的相关性, 若能获取这种被破坏的相关性将有利于后续的特征提取. 因此, 引入高通滤波和通道级卷积等先验知识有利于网络前期的收敛. 预处理阶段的流程主要为: 首先将输入彩色图像进行通道分离为 R, G 和 B 这 3 个通道, 然后对这 3 个通道分别使用 30 个固定的 SRM 高通滤波核进行卷积得到相应的高频残差并采取截断阈值为 5 的截断操作来减小残差的动态范围, 最后将各个通道的 30 个 SRM 高通残差串联在一起得到 90×256×256 尺寸的特征图, 其中 90 为特征图的维度, 256 为特征图的宽度和高度.

该阶段的操作流程主要参考于已有的彩色隐写分析网络 UCNet[24], 采用这种处理方式的优势在于使用 SRM 高通滤波器能有效提取高通残差信息, 有利于后期的隐写分析特征提取. 同时, 由于各个通道间有着内在的相关性, 隐写噪声的引入会对这些相关性带来一定的修改, 因此不同于正常卷积对 3 个通道采用直接求和的形式, 而是在得到各个通道的高通残差后进行串联在一起, 能够有效地保留各个通道间的隐写嵌入修改信息, 从而有利于后续的特征提取阶段.

● 主干流的特征提取阶段

主干流的特征提取阶段主要用来对预处理阶段输出的残差特征图进一步提取更深层次的隐写分析特征. 在该阶段, 我们构建了 3 个基于中心差分卷积的卷积块来提取更充分的特征, 并且表 1 列出了具体的参数设置. 第 1 个

© 中国科学院软件研究所 http://www.jos.org.cn

卷积块类型由 2 个基于中心差分卷积的卷积层和 1 个步长为 2 的平均池化层构成, 两层卷积的输入和输出通道分别为 (186, 64) 和 (64, 64). 第 2 个卷积块类型的结构和第 1 个卷积块相似, 差异在于两层中心差分卷积的输入和输出通道分别为 (64, 128) 和 (128, 128). 第 3 个卷积块类型包含 1 个中心差分卷积层和 1 个步长为 2 的平均池化层, 其中输入和输出通道为 (128, 256).

表 1　所提出的彩色图像隐写分析模型的具体参数配置

模型结构	具体阶段	输入核尺寸	输出特征图尺寸
		(宽×高)×深度	(宽×高)×通道
主干流	预处理阶段	(5×5)×30	(256×256)×(90+96)
	特征提取阶段-卷积块1	(3×3)×90	(128×128)×64
	特征提取阶段-卷积块2	(3×3)×64	(64×64)×128
	特征提取阶段-卷积块3	(3×3)×128	(32×32)×256
	分类阶段-全局协方差池化层	(32×32)×256	256×(256+1)/2
	分类阶段-全连接层1	256×(256+1)/2	2048
	分类阶段-丢弃层	2048	2048
	分类阶段-全连接层2	2048	2
残差空间注意力增强模块	Gabor滤波	(5×5)×32	(256×256)×96
	空间注意力机制	(3×3)×96 (3×3)×64 (256×256)×1 (7×7)×2	(128×128)×64 (128×128)×64 (128×128)×2 (128×128)×1
通道注意力增强模块	卷积块	(64×64)×128	(32×32)×256
	通道矩阵运算	(32×32)×256	256×256
	后续运算	256×256/256×(32×32)	(32×32)×256

现有的基于卷积神经网络的隐写分析方法大多数都是使用普通卷积进行设计的. 在卷积神经网络中, 普通卷积的操作主要分为两步: 如图 2(a) 所示, 在输入特征图 x 中采样一个局部感受野 R, 然后将权重与对应的特征值相乘再求和, 输出的特征图 y 可表示如下.

$$y(p_0)=\sum_{p_n\in R} w(p_n)\cdot x(p_0+p_n) \tag{1}$$

其中, p_0 为输入特征图 x 和输出特征图 y 当前的位置, p_n 为局部感受野 R 上的位置, w 为对应权重参数. 区别于普通卷积, 中心差分卷积[25,26]最初被提出应用在人脸活体检测任务并取得了比普通卷积更好的结果. 具体地说, 中心差分卷积能够通过聚合强度和梯度信息来更好地捕获特征图中的细节信息, 能增强网络对细节纹理的捕捉能力和模型的表达能力. 如图 2(b) 所示, 中心差分卷积是在普通卷积的基础上多加了一步中心差分操作, 也就是将输入特征图当前的位置扩展为与局部感受野一样大小, 然后再将采样后的位置与扩展后的进行相减, 最后再经过普通卷积计算. 输出的特征图 y 的具体计算过程可表示如下:

$$y(p_0)=\sum_{p_n\in R} w(p_n)\cdot (x(p_0+p_n)-x(p_0)) \tag{2}$$

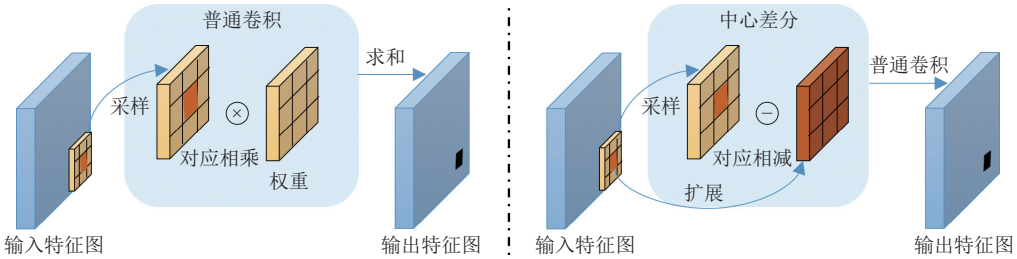
与此同时, 为了考虑梯度信息并且保留强度信息, 一般将中心差分和普通卷积进行融合, 过程如下所示:

$$\mathrm{y}(p_0)=\theta\cdot\sum_{p_n\in R} w(p_n)\cdot (x(p_0+p_n)-x(p_0))+(1-\theta)\cdot\sum_{p_n\in R} w(p_n)\cdot x(p_0+p_n) \tag{3}$$

其中, 参数 θ 的取值范围为 [0, 1], 在所提的隐写分析模型中我们将其设置为 0.7.

此外, 这是首次将中心差分卷积代替普通卷积结构应用在隐写分析任务. 由于目前的彩色自适应隐写算法都是基于内容自适应进行嵌入秘密信息的, 大多数嵌入修改集中在纹理丰富的区域. 因此, 相比于普通卷积, 能够良好地捕捉细节纹理信息的中心差分卷积更加有利于提取深层次的隐写分析特征.

© 中国科学院软件研究所 http://www.jos.org.cn



(a) 普通卷积　　(b) 中心差分卷积

图 2　普通卷积和中心差分卷积的操作示意图

● 主干流的分类阶段

在分类阶段, 最后的输出是预测载体图像或载密图像. 现有的基于深度学习的隐写分析模型一般在分类输出前使用全局平均池化层来计算每个特征图的平均值, 以此来聚合高层次卷积特征并且进一步融合相邻区域的信息和降低特征维度. 然而, 与一阶统计特征 (如均值) 相比, 高阶统计特征 (如协方差[27]) 一般包含更多有用的信息. 因此, 我们采用一个全局协方差池化层用来将特征提取阶段的输出特征图转换为特征向量并将其作为主干流分类阶段的池化层.

然后, 两个全连接层被用来对特征向量进行进一步的分类. 第 1 个全连接层的输入和输出单元分别为 256×(256+1)/2 和 2 048, 该层主要用于将高维特征映射到较低维度的特征向量, 以降低计算复杂度. 第 2 个全连接层的输入和输出单元分别为 2 048 和 2, 用于将特征向量映射到最终的输出类别. 同时, 我们在两个全连接层之间设置了一层丢弃 (dropout) 操作, 也就是前向传播的过程中, 让神经元以一定概率暂时将其从网络中丢弃, 这种策略有利于检测性能的提升和有效地缓解模型过拟合现象的发生. 在我们所提出的隐写分析模型中, 所设置的丢弃概率为 0.7. 另外, 为了验证采用的全局协方差池化层和全连接层的丢弃操作的必要性, 我们在第 3.4 节进行了相应的消融实验, 以此证明所设计的分类阶段结构的合理性.

2.2　残差空间注意力增强模块

为了进一步增强模型主干流前期的特征表达能力, 我们在主干流的前期设计了一个残差空间注意力增强模块然后与主干流合并到一起, 用来对前期的 SRM 高通残差和隐写分析特征进行特征增强. 该模块主要由通道级的 Gabor 滤波[28]卷积和空间注意力机制组成, 采用这两者组成前期特征增强模块的原因分别如下. 首先, 区别于在主干流使用的 SRM 滤波器, 在该模块我们使用 Gabor 滤波器来提取相应的高通残差, 而后在主干流的前期通过串联的连接方式合并到模型主干流, 合并后得到一个 186 维的高通残差特征图, 能够增强模型主干流的残差特征多样性, 有利于主干流能从更丰富的高通残差中提取隐写分析特征. 其次, 受注意力机制的启发, 所采用的空间注意力机制是为了关注残差特征图中特征更明显的区域[29], 能够更有利于提取多样的隐写分析特征.

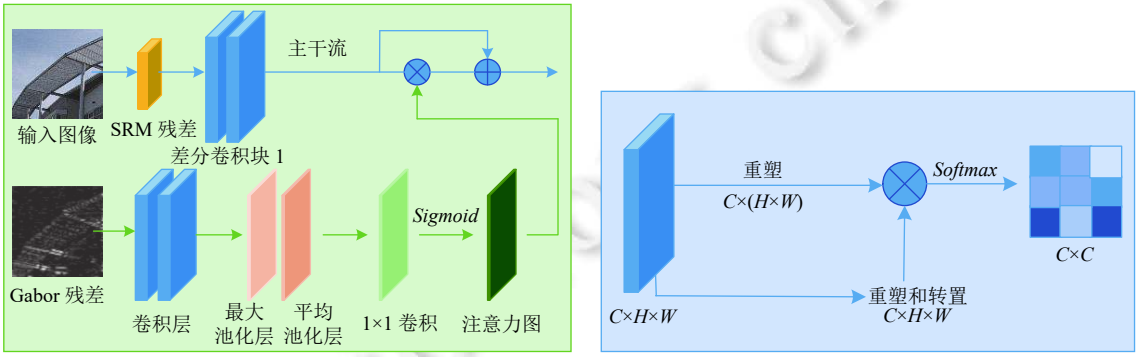
总的来说, 残差空间注意力增强模块的具体结构可表示为如下. 首先, 为了获取区别于主干流的 30 个 SRM 高通残差, 我们采用了 32 个 Gabor 核来提取输入图像的 Gabor 高通残差特征, 依然是对输入的彩色图像使用通道级卷积得到每个通道有着 32 个截断后的 Gabor 高通残差, 最后将 3 个通道的残差串联成 96 维的特征图形式. 然后, 为了进一步获取 Gabor 残差特征图的隐写修改伪迹, 我们采用空间注意力机制来进一步提取更深层的特征. 为了得到空间注意力图, 采用的是将特征图沿通道轴进行平均池化和最大池化操作, 然后将它们串联起来生成一个有效的特征表示. 如图 3(a) 所示, 我们所使用的空间注意力结构可表示为: 首先是两个普通卷积层, 第 1 层卷积层的卷积核为 3×3, 步长为 2, 并附带着 BN 和 ReLU 层. 第 2 层卷积层的卷积核为 3×3, 步长为 1. 接着对经过卷积后得到的特征图进行平均池化和最大池化, 再将平均池化特征图和最大池化特征图进行串联到一起. 最后, 再经过一个卷积核为 7×7 的卷积层并附带 *Sigmoid* 函数计算来得到最后的注意力权重图. 该残差空间注意力图的计算过程可表示如下:

$$R = Sigmoid(Conv([avg_f, max_f])) \tag{4}$$

© 中国科学院软件研究所 http://www.jos.org.cn

其中, f 是经过两层卷积后输出的特征图, avg_f 和 max_f 分别是对 f 进行平均池化和最大池化后获得的特征图, $Conv$ 表示卷积操作, R 是输出的注意力图, $[\cdot]$ 表示串联操作.

值得注意的是, 我们首先将获得的 96 维 Gabor 残差输入到主干流与 90 维 SRM 残差进行串联为 186 维的高通残差特征, 以此加强主干流的残差特征多样性, 从而有利于主干流后续的隐写分析特征提取. 其次, 将残差空间注意力增强模块合并到模型主干流时, 我们采用的是先将该模块输出的注意力图与主干流的特征图逐元素相乘后再与之前的主干流特征进行相加的方式. 具体而言, 先将模块输出的注意力图与主干流的特征图逐元素相乘, 得到加权的主干流特征图, 然后将加权的主干流特征图与之前的主干流特征进行相加. 这种方式在一定程度上保留了初始特征和增加了注意力特征的丰富性. 在这个过程中, 合并的位置是在主干流的第 1 个中心差分卷积块之后, 随后将计算好后的特征图送入主干流后续的中心差分卷积块和特征分类阶段中去.



(a) 空间注意力机制 (b) 通道矩阵计算流程

图 3 空间注意力机制和通道矩阵计算流程示意图

2.3 通道注意力增强模块

另外, 区别于在模型主干流前期的残差空间注意力增强模块, 为了进一步增强模型主干流后期的通道间的依赖性, 我们构建了一个通道注意力增强模块并与模型主干流进行合并, 从而有利于最后分类阶段对载体和载密图像的判别. 具体来说, 主干流后期的更深层特征的每个通道都可以被看作是一个特定的类别且是相互关联的, 而通道注意力关注的是哪些特征是有意义的[30,31]. 因此, 受已有通道注意力方法[30]的启发, 我们使用了一个通道注意力增强模块来模拟各通道之间的相互依赖关系. 通过利用通道之间的内在关系, 可以在模型主干流后期强化相应的通道特征并改善特征表示, 更有利于后续的特征分类.

图 1 和图 3(b) 表示了通道注意力增强模块的具体结构, 另外表 1 列出了该模块的具体参数设置. 总的来说, 我们直接从原始的输入特征 $X \in R^{C\times H\times W}$ 来计算得到最终的输出注意力图 $Y \in R^{C\times H\times W}$. 该通道注意力增强模块的结构可概括为: 首先, 将主干流的第 2 个中心差分卷积块的输出特征图经过一个和中心差分卷积块 3 相同结构的卷积块, 然后将该卷积块输出的特征 $A \in R^{C\times H\times W}$ 经过通道矩阵计算和另外的重塑合并操作. 其中通道矩阵计算流程如图 3(b) 所示, 具体流程是先将 A 重塑为 $A_{\text{reshape}} \in R^{C\times(H\times W)}$ 形式, 然后再单独对 A 进行先重塑后转置的方式来得到 $A_{\text{reshape}}^{\text{T}} \in R^{(H\times W)\times C}$ 结构, 最后对两者进行矩阵相乘计算, 接着应用 $Softmax$ 函数层来获得相应的注意力图 $Z \in R^{C\times C}$. 该过程可以表示为:

$$Z = Softmax(A_{\text{reshape}} \times A_{\text{reshape}}^{\text{T}}) \tag{5}$$

在进行完通道矩阵计算操作得到注意力图后, 我们再进行另外的重塑合并操作. 该重塑合并操作过程首先将 A 重塑为 $A_{\text{reshape}} \in R^{C\times(H\times W)}$ 特征形式, 然后将 A_{reshape} 与注意力图 Z 进行矩阵相乘, 然后将得到的特征图 N 再进行重塑为 $N \in R^{C\times H\times W}$ 形式, 并将该特征图乘以设置的尺度参数 β, 最后将特征 A 与此输出进行逐元素求和运算来获得最终的输出通道注意力图 C, 该过程可以表示为:

$$C = \beta \cdot (Z \times A_{\text{reshape}}) + A \tag{6}$$

© 中国科学院软件研究所 http://www.jos.org.cn

其中, β 是从 0 开始逐渐学习的权重. 公式 (6) 表明, 每个通道的最终特征是所有通道的特征和原始特征的加权和, 它是特征图之间长距离语义依赖的表示, 并且有助于提高通道特征的可辨别性.

为了有效地计算通道注意力, 需要对输入特征图的空间维度进行重塑转换, 有效地利用了通道间的原始和转换后的信息. 同时, 注意到在计算通道间的关系之前, 我们使用了一层卷积层和一层平均池化层结构的卷积块来对初始输入的特征图进行卷积和下采样, 这样能减少计算资源的需求并且可以维持不同通道映射间的关系. 此外, 与一般的通过全局平均池化或单一的通道注意力机制的工作不同, 我们在该模块利用的是所有对应位置的空间信息来构建通道间的相关性, 这样更有利于后续的分类阶段.

2.4 与现有隐写分析方法的差异性

为了体现所提出的彩色图像隐写分析模型与现有的基于深度学习的隐写分析模型结构的差异性, 如表 2 所示, 我们列出了相关隐写分析模型结构的组成细节和适用场景. 主要对比的现有隐写分析模型有适用于灰度图像的 CovNet[13]和 SRNet[12], 以及适用于彩色图像的 WISERNet[23]和 UCNet[24].

表 2 所提方法与已有的基于深度学习的隐写分析模型结构的差异性

隐写分析器	预处理阶段	特征提取阶段	分类阶段	适用场景
CovNet[13]	30个固定的SRM核和截断处理	10层普通卷积层	1个全局协方差池化层和1个全连接层	灰度图像
SRNet[12]	64个可学习的随机滤波核	22层普通卷积层	1个全局平均池化层和1个全连接层	灰度图像
WISERNet[23]	30个可学习的SRM核	3层普通卷积层	1个全局平均池化层和3个全连接层	彩色图像
UCNet[24]	62个固定的SRM核和Gabor核, 以及截断处理	11层普通卷积层	1个全局平均池化层和1个全连接层	彩色图像
所提方法	30个固定的SRM核和截断处理	5层中心差分卷积层和2个基于注意力机制的特征增强模块	1个全局协方差池化层和2个全连接层附加丢弃操作	彩色图像

从表 2 中, 我们可以有以下 3 点发现: (1) 在预处理阶段, 只有 SRNet 采用了 64 个可学习的随机核, 其他方法都是采用了已有的高通核, 所提出的方法与 CovNet 同样采用的是 30 个固定的 SRM 核并进行了截断处理; (2) 在特征提取阶段, 现有的方法都是采用了多层结构的普通卷积层, 而所提方法则首次采用了中心差分卷积层结构并且构建了两个基于注意力机制的特征增强模块来强化隐写分析特征的提取; (3) 在分类阶段, 已有的方法都是采用一个全局平均/协方差池化层和全连接层进行分类, 而所提方法则在设计的两个全连接层之间增加了一个丢弃层的操作, 以此来有效地增加模型的分类能力.

相比于目前已有的表现最好的彩色隐写分析方法 UCNet, 所提方法与其差异主要在于: 首先, UCNet 采用的是 62 个 SRM 和 Gabor 组成的高通滤波器, 而所提方法使用的是 30 个固定的 SRM 核. 其次, UCNet 采用的是 3 种不同的卷积层级类型结构, 而所提方法使用了 5 层中心差分卷积层以及 2 个基于注意力机制的特征增强模块. 最后, UCNet 使用的是 1 个全局平均池化层和 1 个全连接层进行分类, 而所提方法采用了 1 个全局协方差池化层和有着丢弃操作的两个全连接层组成的结构.

3 实验结果与分析

3.1 数据集与评价指标

在实验中, 我们采用了图像隐写及隐写分析领域常用的彩色数据集, 也就是来自 ALASKA II 隐写分析挑战赛 (The ALASKA Steganalysis Challenge) [32]的图像. 我们从该数据集随机选取 20 000 张彩色载体图像进行后续的实验, 图像尺寸为 256×256. 我们采用 3 种典型的彩色图像隐写体系进行嵌入, 分别是 GINA[5], ACMP[6]和 CMD-C[7], 嵌入率分别为 0.4, 0.3 和 0.2 bpc (bit per channel). 因此, 对于一组特定的隐写算法和嵌入率, 我们可以得到 20 000 个载体图像-载密图像对. 和已有的方法[20]一样, 我们随机从 20 000 对图像中随机选取 14 000 对图像作为训练集, 1 000 对作为验证集, 剩余的 5 000 对作为测试集. 另外, 为了缓解基于深度学习的图像隐写分析方法的过拟合现象, 我们

© 中国科学院软件研究所 http://www.jos.org.cn

采用和现有方法[12,13,24]相同的策略, 在训练过程中使用随机镜像和镜像 90°旋转两种数据增强方法来增加训练数据的样本量.

在本文中, 我们采用常用的隐写分析评价指标检测准确率 (ACC) 来评估所提出的隐写分析模型性能, 该评价指标也广泛用于已有的隐写分析工作[9−16]. 检测准确率的计算可由虚警率和漏检率计算而来, 虚警率 P_{FA} 主要指的是载体图像被错误分类为载密图像的比例, 漏检率 P_{MD} 指的是载密图像被错误分类为载体图像的比例, 其中, 虚警率 P_{FA} 和漏检率 P_{MD} 的计算公式分别如下所示:

$$P_{FA}=\frac{FP}{TN+FP} \tag{7}$$

$$P_{MD}=\frac{FN}{TP+FN} \tag{8}$$

其中, FP 指的是载体图像被错误分类为载密图像的数量, TN 指的是载体图像被正确分类为载体图像的数量, FN 表示载密图像被错误分类为载体图像的数量, TP 指的是载密图像被正确分类为载密图像的数量. 最后, 检测准确率 ACC 可由虚警率 P_{FA} 和漏检率 P_{MD} 计算得到, 具体的计算方式可表示为:

$$ACC=1-\frac{1}{2}(P_{FA}+P_{MD}) \tag{9}$$

除了检测准确率之外, 我们也采用了另一评价指标 (AUC) 来衡量模型的检测表现, 该指标是由计算 ROC 曲线下的所占面积而得到. ROC 曲线是将假正例率 (false positive rate, FPR) 作为横轴和真正例率 (true positive rate, TPR) 作为纵轴的曲线表示, 横纵轴表示可分别定义为:

$$FPR=\frac{FP}{TN+FP} \tag{10}$$

$$TPR=\frac{TP}{TP+FN} \tag{11}$$

AUC 值代表的则是相应方法所绘制的 ROC 曲线下所占的面积, 所占面积越大表示性能越好, 能良好地反映隐写分析器对载体和载密样本的分类能力.

3.2 参数设置

在本文中, 所提出的模型是基于 PyTorch 深度学习框架实现的, 实验的硬件平台是采用两张 12 GB 显存的 NVIDIA 2080Ti 显卡进行的. 在训练阶段, 我们使用了动量为 0.9 且权重衰减为 5×10^{-4} 的随机梯度下降 (stochastic gradient descent, SGD) 优化器来优化网络参数. 对于卷积块中的卷积层, 我们采用了 He 初始化方法[33]进行初始化, 并禁用了偏置. 训练的批大小设置为 30 (即 30 对载体-载密图像), 训练集在每个迭代周期都会被随机打乱. 网络完整的训练迭代周期为 230, 初始学习率为 0.02, 在训练过程中, 学习率会在第 80、130 和 170 个迭代周期时下降为原来的 1/10. 在最后 60 个迭代周期中, 在验证集上性能最好的参数将被作为训练结果用于测试. 另外, 对于更低的嵌入率, 我们没有采用已有隐写分析方法常用的迁移学习[34]策略, 而是对每个低嵌入率从头开始训练, 相关的讨论可参考第 3.5 节. 关于所提出的隐写分析模型的具体参数配置可参考表 1 以及详细的实现细节可参考本文的代码 (代码已发布在 https://github.com/revere7/CANet).

3.3 实验结果

为了进一步验证本方法的有效性, 在本节中, 我们将所提出的彩色图像隐写分析模型与现有的一些典型隐写分析方法进行了性能对比. 主要包括了一种传统的彩色图像隐写分析器 CRMQ1[20], 两种典型的现有经过修改后的灰度图像隐写分析器 CovNet[13]和 SRNet[12], 以及两种彩色图像隐写分析网络 WISERNet[23]和 UCNet[24]. 为了公平的比较, 和已有方法[24,35,36]一样, 我们对相应的灰度图像隐写分析器 (也就是 CovNet 和 SRNet) 做了细微的修改, 在第 1 层将 $H\times W$ 核修改为 $3\times H\times W$ 核, 其中 H 和 W 分别是卷积核的高和宽, 3 为彩色图像的通道数. 实验过程中, 我们使用了 3 种彩色图像隐写嵌入策略 GINA[5]、ACMP[6]和 CMD-C[7]并各自结合两种隐写算法 HILL[3]和 S-UNIWARD[2], 共在 6 种不同的隐写方法上进行实验. 另外, 所采用的嵌入率为 0.4 bpc, 0.3 bpc 和 0.2 bpc. 相应的实验结果分别如表 3–表 5 所示, 其中所有的实验结果都是基于两次随机数据划分进行实验的平均结果.

© 中国科学院软件研究所 http://www.jos.org.cn

表 3 在 GINA 策略下所提方法和 5 种相关的彩色图像隐写分析器的检测准确率比较 (%)

隐写分析器	GINA-HILL[5]			GINA-S-UNIWARD[5]		
	0.4 bpc	0.3 bpc	0.2 bpc	0.4 bpc	0.3 bpc	0.2 bpc
CRMQ1[20]	69.63	65.11	60.63	70.74	67.13	61.80
CovNet[13]	67.62	65.48	61.85	63.90	58.80	52.88
SRNet[12]	76.80	72.90	66.30	70.05	65.20	59.20
WISERNet[23]	71.30	66.55	59.50	73.63	68.25	56.90
UCNet[24]	81.50	77.75	73.46	76.10	73.35	67.30
所提方法	84.55*	82.15*	78.70*	80.75*	77.50*	71.50*

注: “*”表示在相应情况下最好的结果

表 4 在 ACMP 策略下所提方法和 5 种相关的彩色图像隐写分析器的检测准确率比较 (%)

隐写分析器	ACMP-HILL[6]			ACMP-S-UNIWARD[6]		
	0.4 bpc	0.3 bpc	0.2 bpc	0.4 bpc	0.3 bpc	0.2 bpc
CRMQ1[20]	73.61	69.37	63.78	75.42	71.96	66.31
CovNet[13]	78.49	71.33	64.80	70.55	66.75	62.55
SRNet[12]	86.83	81.45	74.40	83.20	78.85	68.45
WISERNet[23]	79.48	70.23	61.75	77.50	71.05	63.30
UCNet[24]	89.52	84.81	79.55	85.60	81.05	76.50
所提方法	90.50*	86.25*	82.10*	87.64*	82.60*	77.25*

注: “*”表示在相应情况下最好的结果

表 5 在 CMD-C 策略下所提方法和 5 种相关的彩色图像隐写分析器的检测准确率比较 (%)

隐写分析器	CMD-C-HILL[7]			CMD-C-S-UNIWARD[7]		
	0.4 bpc	0.3 bpc	0.2 bpc	0.4 bpc	0.3 bpc	0.2 bpc
CRMQ1[20]	71.95	67.29	62.42	73.68	68.83	63.80
CovNet[13]	70.56	66.62	62.85	66.45	62.95	58.30
SRNet[12]	76.20	73.45	68.40	75.96	68.48	66.60
WISERNet[23]	74.96	69.12	60.45	76.15	70.40	63.68
UCNet[24]	82.85	80.75	78.05	80.05	76.85	72.37
所提方法	85.85*	83.55*	79.75*	82.90*	79.50*	76.15*

注: “*”表示在相应情况下最好的结果

● 在 GINA 彩色图像隐写策略下的性能表现. 表 3 展示了各种隐写分析方法分别对 GINA-HILL 和 GINA-S-UNIWARD 隐写算法下的检测准确率. 通过表 3 可以看出, 在各种嵌入率下, 我们所提出的模型的检测准确率都要优于其他隐写分析方法, 其中 CovNet 平均表现最差, 而 UCNet 有着已有方法的最优性能. 相比目前最好的彩色图像隐写分析网络 UCNet, 我们所提出的方法在 GINA-HILL 和 GINA-S-UNIWARD 下分别平均提升了 4.23% 和 4.33% 的检测准确率, 提升幅度是显著的.

● 在 ACMP 彩色图像隐写策略下的性能表现. 表 4 展示了已有的隐写分析方法对 ACMP-HILL 和 ACMP-S-UNIWARD 隐写算法的检测准确率. 从表 4 可以看出, 我们所提出的隐写分析模型有着最优的检测性能, 其中 CovNet 有着最差的平均表现, 而 UCNet 则有着已有方法的最好表现. 相比于 UCNet, 所提出方法在 ACMP-HILL 和 ACMP-S-UNIWARD 下分别平均提升了 1.66% 和 1.45% 的检测准确率.

● 在 CMD-C 彩色图像隐写策略下的性能表现. 表 5 列出了各种隐写分析网络分别对 CMD-C-HILL 和 CMD-C-S-UNIWARD 隐写算法下的检测准确率. 从中可以看出, 我们所提的模型的检测表现依然是最优的, 其中

© 中国科学院软件研究所 http://www.jos.org.cn

CovNet 平均表现依然最差, UCNet 平均表现最好. 相比于 UCNet, 所提彩色图像隐写分析器在 CMD-C-HILL 和 CMD-C-S-UNIWARD 下检测准确率分别平均提升了 2.50% 和 3.09%.

总的来说, 从上述对 6 种不同的彩色图像隐写算法的检测结果表明, 我们所提出的彩色图像隐写分析模型相比于已有的方法都有着更好的检测准确率. 即使和目前表现最优的彩色图像隐写分析网络 UCNet 相比, 所提方法在六种彩色图像隐写算法下的检测准确率平均提升幅度也达到了 2.92% 左右. 因此, 在本节通过充足的实验证明了我们所提出的彩色图像隐写分析模型的有效性.

3.4 消融实验

为了验证所提出的彩色图像隐写分析器所设计的结构合理性, 我们进一步对比了当所提模型采用不同配置下的实验检测结果. 该消融实验主要探究的是所提出的隐写分析器中的各类结构存在的必要性和主干流分类阶段所使用的不同配置带来的影响. 在该实验下, 我们采用的是 GINA-HILL 和 ACMP-S-UNIWARD 两种隐写算法, 嵌入率为 0.4 bpc. 相关实验的具体描述及实验结果如下.

● 所提出的隐写分析器中各类结构的影响. 我们基于所提出的模型探究了 4 种不同结构对检测性能带来的影响, 分别是: 在主干流的特征提取阶段使用如图 2(a) 的普通卷积, 在主干流的特征提取阶段使用如图 2(b) 的中心差分卷积, 使用中心差分卷积的主干流加上残差空间注意力增强模块, 以及使用中心差分卷积的主干流加上通道注意力增强模块 4 种不同结构. 相应的实验结果如表 6 所示.

表 6　所提方法对不同结构的消融实验检测准确率比较 (%)

不同结构	GINA-HILL	ACMP-S-UNIWARD
主干流 (特征提取阶段使用普通卷积)	82.60	86.50
主干流 (特征提取阶段使用中心差分卷积)	83.15	86.88
主干流 (特征提取阶段使用中心差分卷积) + 残差空间注意力增强模块	83.75	87.40
主干流 (特征提取阶段使用中心差分卷积) + 通道注意力增强模块	84.15	87.23
所提方法	84.55*	87.64*

注: “*”表示在相应情况下最好的结果

通过表 6 的结果, 我们可以观察到以下 3 点现象: (1) 使用中心差分卷积的主干流的检测表现要好于使用普通卷积的主干流, 在 GINA-HILL 和 ACMP-S-UNIWARD 两种隐写方法下检测准确率提升幅度分别达到了 0.55% 和 0.38%; (2) 采用中心差分卷积的主干流加上残差空间注意力增强模块相比于只使用中心差分卷积的主干流提升幅度分别有 0.60% 和 0.52%, 采用中心差分卷积的主干流加上通道注意力增强模块相比只使用中心差分卷积的主干流也能分别有 1.00% 和 0.35% 的提升幅度; (3) 我们所提出的隐写模型结构包含中心差分卷积的主干流, 残差注意力增强模块和通道注意力增强模块, 将其共同使用也就是我们所提出的彩色图像隐写分析器, 达到了检测性能最好的表现. 因此, 也表明了我们所构建的模型主干流和两个注意力增强模块的合理性和必要性.

● 所提出的隐写分析器中分类阶段结构的影响. 在本实验中, 我们探究的是模型主干流的分类阶段的不同配置对检测性能带来的影响. 首先, 我们测试了不同的池化层类型, 一种是被广泛用于基于深度学习隐写分析网络中的全局平均池化层, 另一种是在所提出的隐写分析器中所使用的全局协方差池化层. 然后, 我们对在分类阶段的两个全连接层之间有无丢弃层操作进行了实验, 测试了丢弃层的存在所带来的结果影响. 相应的实验结果如表 7 所示.

通过表 7 的结果, 我们可以有以下两点发现: (1) 在所提出的模型中使用全局协方差池化相比于全局平均池化可以提高隐写分析器的检测性能, 在 GINA-HILL 和 ACMP-S-UNIWARD 两种隐写方法下提升幅度分别达到了 2.25% 和 0.74%; (2) 在所提出的模型的分类阶段的两个全连接层之间采用丢弃操作可以有效地提升检测的表现, 在两种隐写算法下平均提升幅度分别达到了 0.86% 和 0.44%. 因此, 在模型的分类阶段采用全局协方差池化和丢弃操作是有必要的, 能够对模型的检测性能带来一定的提升.

● 所提出的隐写分析器主干流的中心差分卷积参数 θ 的影响. 在本实验中, 我们探究的是模型主干流所使用

© 中国科学院软件研究所 http://www.jos.org.cn

的中心差分卷积融合时的参数 θ 的选择对检测性能带来的影响. 参数 θ 是用来将中心差分卷积和普通卷积进行融合时的参数, 能在考虑到梯度信息的同时并保留强度信息. 我们选取了参数 θ 的 4 种取值, 分别是 0.3, 0.5, 0.7 和 0.9 进行实验. 相应的实验结果如表 8 所示.

表 7　分类阶段消融实验检测准确率比较 (%)

不同结构	GINA-HILL	ACMP-S-UNIWARD
全局平均池化	82.30	86.90
无丢弃层	83.69	87.20
所提方法	84.55*	87.64*

注: “*”表示在相应情况下最好的结果

表 8　参数 θ 的消融实验检测准确率比较 (%)

参数 θ 的值	GINA-HILL	ACMP-S-UNIWARD
0.3	84.20	87.60
0.5	84.45	87.05
0.7	84.55*	87.64*
0.9	78.95	85.25

注: “*”表示在相应情况下最好的结果

通过表 8 的结果, 我们可以有以下两点发现: (1) 在所提模型的中心差分卷积的融合参数 θ 值为 0.7 时有着最好的检测表现, 相比于其他 3 种取值, 在 GINA-HILL 和 ACMP-S-UNIWARD 两种隐写方法下平均提升幅度分别达到了 2.02% 和 1.01%; (2) 当融合参数 θ 值为 0.9 时, 所提模型有着最差的表现, 相比于最好的取值 0.7, 在 GINA-HILL 和 ACMP-S-UNIWARD 下的下降幅度分别达到了 5. 50% 和 2.39%. 因此, 在所提模型的中心差分卷积的融合参数 θ 值我们设置为 0.7, 能够有着最好的模型检测性能.

● 所提出的隐写分析器中主干流和残差空间注意力增强模块所使用的滤波器的影响. 在本实验中, 我们探究的是模型主干流和残差空间注意力增强模块使用不同的滤波器对检测性能带来的影响. SRM 滤波器和 Gabor 滤波器是隐写分析中最为常用的来提取高通残差的滤波器. 因此, 我们在所提方法也使用到这两种滤波器. 我们在主干流和残差空间注意力增强模块分别使用了 SRM/Gabor 和 Gabor/SRM 滤波器共 4 种组合进行了实验. 相应的实验结果如表 9 所示.

表 9　主干流和残差空间注意力增强模块所使用滤波器的消融实验检测准确率比较 (%)

主干流	残差空间注意力增强模块	GINA-HILL	ACMP-S-UNIWARD
SRM	SRM	83.30	87.00
SRM	Gabor	84.55*	87.64*
Gabor	Gabor	79.35	83.25
Gabor	SRM	83.05	86.70

注: “*”表示在相应情况下最好的结果

通过表 9 的结果, 我们可以有以下两点发现: (1) 当模型主干流使用 SRM 滤波器和残差空间注意力增强模块使用 Gabor 滤波器时, 所提方法有着最好的检测表现, 相比于主干流和残差空间注意力增强模块都使用 SRM 滤波器, 在 GINA-HILL 和 ACMP-S-UNIWARD 两种隐写方法下提升幅度分别达到了 1.25% 和 0.64%; (2) 当主干流和残差空间注意力增强模块都使用 Gabor 滤波器时检测性能最差, 说明两者都使用 Gabor 滤波器提取到的高通残差不利于模型后续的隐写分析特征提取. 因此, 我们在所提的模型中的主干流和残差空间注意力增强模块分别使用的是 SRM 滤波器和 Gabor 滤波器, 有利于提升模型的检测表现.

3.5　低嵌入率训练策略选择

为了检测更低嵌入率下 (也就是 0.3 bpc 和 0.2 bpc) 的隐写算法, 已有的基于深度学习的隐写分析方法通常使用从头训练或者从更高的嵌入率下进行迁移学习的策略. 因此, 在所提出的网络训练过程中, 我们分别测试了这两种训练策略对检测低嵌入率下的隐写算法带来的影响. 在该实验下, 我们采用的是 GINA-HILL 和 ACMP-S-UNIWARD 两种隐写方法, 对更低嵌入率进行从头训练采用的是和 0.4 bpc 下相同的训练方法, 而迁移学习采用的训练迭代周期为 120, 其中学习率衰减迭代周期设置为当第 50、80、100 次迭代时衰减为原来的 1/10. 我们在这两种训练策略下分别进行了对比实验, 以此来确定所提出的隐写分析模型下的低嵌入率下所采用的训练策略. 相关的对比实

© 中国科学院软件研究所 http://www.jos.org.cn

验结果如表 10 所示.

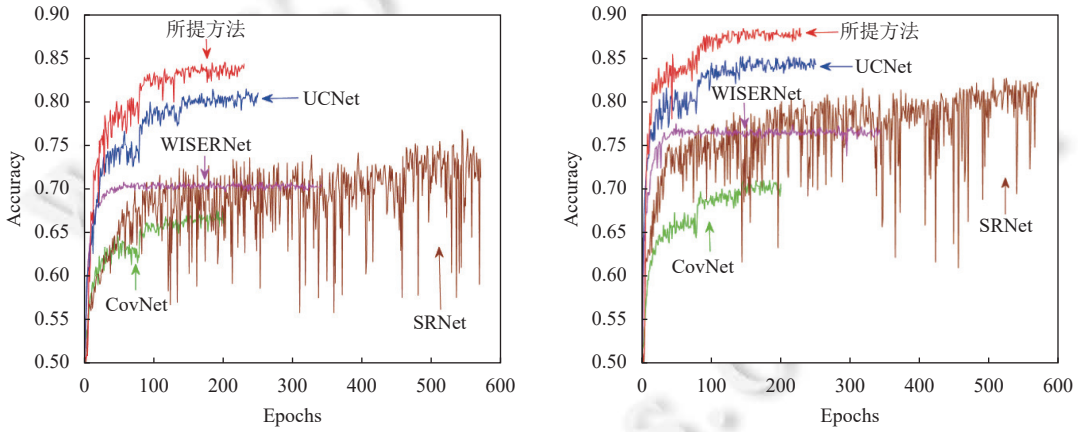
从表 10 的实验结果来看, 我们有以下两点发现: (1) 在 GINA-HILL 和 ACMP-S-UNIWARD 两种隐写方法下的 0.3 bpc 和 0.2 bpc, 采用从头训练策略的检测表现多数情况下是优于迁移学习策略的, 整体平均提升幅度为 0.04%; (2) 在 ACMP-S-UNIWARD 的 0.2 bpc 下, 迁移学习的性能比从头训练更好, 这种情况表明了迁移学习训练时的稳定性还有待提升. 因此, 为了在更低嵌入率下有着普遍更好的性能, 我们对于更低嵌入率下采用的是从头训练策略.

表 10　低嵌入率下采用迁移学习和从头训练的检测准确率比较 (%)

隐写方法	迁移学习			从头训练	
	0.4 bpc	0.3 bpc	0.2 bpc	0.3 bpc	0.2 bpc
GINA-HILL	84.55	81.75	78.20	82.15	78.70
ACMP-S-UNIWARD	87.64	82.40	78.20	82.60	77.25

3.6　相关隐写分析模型收敛表现和 ROC 曲线

在本节, 我们首先比较了所提出的隐写分析模型与所对比的基于深度学习的隐写分析方法的收敛性能. 在该实验下, 我们采用的是 0.4 bpc 下的 GINA-HILL 和 ACMP-S-UNIWARD 两种隐写方法进行测试. 图 4 显示了在训练时期相关的基于深度学习隐写分析方法的验证集下的检测准确率变化曲线. 从图 4 可以看出, 所提出的隐写分析方法的收敛性能最好且所需的训练迭代周期较少, 而 SRNet 的训练迭代周期最多且验证集检测准确率波动较大. 另外, UCNet 的收敛性能有着较好的表现, WISERNet 表现一般的同时收敛起伏较为稳定, CovNet 虽然有着最少的训练迭代次数但收敛表现最差.



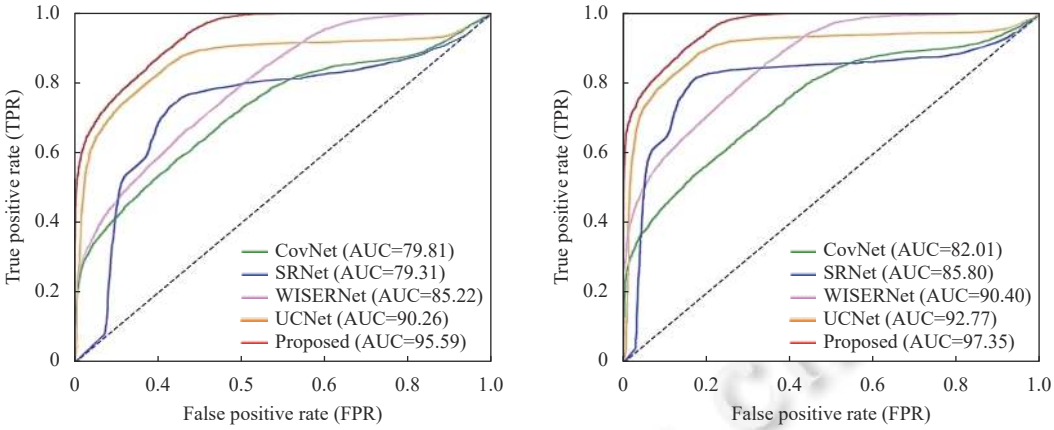
(a) GINA-HILL 下的收敛表现　　(b) ACMP-S-UNIWARD 下的收敛表现

图 4　对 GINA-HILL 和 ACMP-S-UNIWARD 隐写算法下的相关隐写分析方法的验证集收敛示意图

其次, 为了更全面地展示所提方法的有效性, 我们比较了所提出的彩色图像隐写分析模型与相关对比的基于深度学习的隐写分析方法的 ROC 曲线及其 AUC 值, 该实验是在 0.4 bpc 下的 GINA-HILL 和 ACMP-S-UNIWARD 两种隐写方法的测试集下进行的. 图 5 显示了所提隐写分析方法和相关隐写分析方法的 ROC 变化曲线图以及相应的 AUC 值.

从图 5, 我们可以观察到以下情况: (1) 所提出的隐写分析方法 (在图中表示为“Proposed”) 的 ROC 曲线最靠近 ROC 图的左上角表现最好, UCNet 也有着较好的表现, 而 CovNet 的 ROC 曲线表现最差; (2) 对于相应的 AUC 值, 所提出的模型在 GINA-HILL 和 ACMP-S-UNIWARD 下的值分别达到了 95.59% 和 97.35%, 相比于表现最差的 CovNet 分别有着 15.78% 和 15.34% 的提升幅度.

© 中国科学院软件研究所 http://www.jos.org.cn



(a) GINA-HILL 下的 ROC 曲线及其 AUC(%) 值　　(b) ACMP-S-UNIWARD 下的 ROC 曲线及其 AUC(%) 值

图 5　对 GINA-HILL 和 ACMP-S-UNIWARD 隐写算法下的相关隐写分析方法的 ROC 曲线图

4 总　结

现有的图像隐写分析模型通常基于普通卷积的层级结构进行构建, 并且大多数针对灰度图像, 这使得它们不太适用于社交媒体中广泛存在的彩色图像. 本文提出了一种针对彩色图像的隐写分析模型, 与现有的相关方法不同, 该模型采用中心差分卷积来更好地捕捉特征细节信息, 并在模型的不同阶段采用了残差空间注意力增强模块和通道注意力增强模块, 以增强整体的特征学习能力. 实验结果表明, 所提出的模型能够显著提升彩色图像隐写分析的检测表现, 并且取得了目前最好的结果. 此外, 我们还进行了相应的消融实验来验证所提出的模型结构的合理性.

本文虽然在基于深度学习的彩色图像隐写分析方法上取得了一定的提升, 但随着彩色图像隐写算法的不断发展和应用场景的变化, 彩色图像隐写分析任务仍有以下问题有待进一步解决: (1) 在社交媒体中, JPEG 彩色图像格式存在更为广泛, 目前关于彩色 JPEG 图像隐写分析算法的研究较少. 因此, 构建基于彩色 JPEG 图像的深度学习隐写分析模型仍值得进一步研究; (2) 目前基于深度学习的彩色图像隐写分析模型都是基于卷积结构进行设计. 卷积通常获取的是局部特征, 如何有效引入能够获取全局特征的视觉 Transformer 结构[37]并与卷积进行结合, 应用到彩色图像隐写分析同样值得被考虑; (3) 本文针对的是固定为 256×256 尺寸大小的彩色图像, 对检测实际场景中存在的非固定尺寸和大尺寸隐写图像仍有着一定的局限性. 因此, 针对现实中的非固定尺寸和大尺寸图像隐写分析[38−40], 设计更精细化的特征提取和数据增强等策略来提升彩色图像隐写分析网络架构的实用性具有重要的应用价值.

References:

[1] Huang DZ, Zhang JF, Zhang R, Li PC, Guo YB. New system of multi-modal information hiding based on big data environment. Acta Electronica Sinica, 2017, 45(2): 477–484 (in Chinese with English abstract). [doi: 10.3969/j.issn.0372-2112.2017.02.029]

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

[3] Li B, Wang M, Huang JW, Li XL. A new cost function for spatial image steganography. In: Proc. of the 2014 IEEE Conf. on Image Processing. Paris: IEEE, 2014. 4206–4210. [doi: 10.1109/ICIP.2014.7025854]

[4] Liu ML, Fan HY, Wei KK, Luo WQ, Lu W. Adversarial robust image steganography against lossy JPEG compression. Signal Processing, 2022, 200: 108668. [doi: 10.1016/j.sigpro.2022.108668]

[5] Wang YF, Zhang WM, Li WX, Yu XZ, Yu NH. Non-additive cost functions for color image steganography based on inter-channel correlations and differences. IEEE Trans. on Information Forensics and Security, 2020, 15: 2081–2095. [doi: 10.1109/TIFS.2019.2956590]

[6] Liao X, Yu YB, Li B, Li ZP, Qin Z. A new payload partition strategy in color image steganography. IEEE Trans. on Circuits and Systems for Video Technology, 2020, 30(3): 685–696. [doi: 10.1109/TCSVT.2019.2896270]

[7] Tang WX, Li B, Luo WQ, Huang JW. Clustering steganographic modification directions for color components. IEEE Signal Processing

© 中国科学院软件研究所 http://www.jos.org.cn

Letters, 2016, 23(2): 197–201. [doi: 10.1109/LSP.2015.2504583]

[8] Fridrich J, Kodovsky J. Rich models for steganalysis of digital images. IEEE Trans. on Information Forensics and Security, 2012, 7(3): 868–882. [doi: 10.1109/TIFS.2012.2190402]

[9] Tan SQ, Li B. Stacked convolutional auto-encoders for steganalysis of digital images. In: Proc. of the 2014 Signal and Information Processing Association Annual Summit and Conf. Siem Reap: IEEE, 2014. 1−4. [doi: 10.1109/APSIPA.2014.7041565]

[10] Xu GS, Wu HZ, Shi YQ. Structural design of convolutional neural networks for steganalysis. IEEE Signal Processing Letters, 2016, 23(5): 708–712. [doi: 10.1109/LSP.2016.2548421]

[11] Ye J, Ni JQ, Yi Y. Deep learning hierarchical representations for image steganalysis. IEEE Trans. on Information Forensics and Security, 2017, 12(11): 2545–2557. [doi: 10.1109/TIFS.2017.2710946]

[12] Boroumand M, Chen M, Fridrich J. Deep residual network for steganalysis of digital images. IEEE Trans. on Information Forensics and Security, 2019, 14(5): 1181–1193. [doi: 10.1109/TIFS.2018.2871749]

[13] Deng XQ, Chen BL, Luo WQ, Luo D. Fast and effective global covariance pooling network for image steganalysis. In: Proc. of the 2019 ACM Workshop on Information Hiding and Multimedia Security. Paris: ACM, 2019. 230–234. [doi: 10.1145/3335203.3335739]

[14] Zhang R, Zhu F, Liu JY, Liu GS. Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis. IEEE Trans. on Information Forensics and Security, 2020, 15: 1138–1150. [doi: 10.1109/TIFS.2019.2936913]

[15] Shen J, Liao X, Qin Z, Liu XC. Spatial steganalysis of low embedding rate based on convolutional neural network. Ruan Jian Xue Bao/Journal of Software, 2021, 32(9): 2901–2915 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5980.htm [doi: 10.13328/j.cnki.jos.005980]

[16] 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 with English abstract). http://www.jos.org.cn/1000-9825/6358.htm [doi: 10.13328/j.cnki.jos.006358]

[17] Zhai LM, Jia J, Ren WX, Xu YB, Wang LN. Recent advances in deep learning for image steganography and steganalysis. Journal of Cyber Security, 2018, 3(6): 2−12 (in Chinese with English abstract). [doi: 10.19363/J.cnki.cn10-1380/tn.2018.11.01]

[18] 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 with English abstract). http://www.jos.org.cn/1000-9825/6135.htm [doi: 10.13328/j.cnki.jos.006135]

[19] Holub V, Fridrich J. Random projections of residuals for digital image steganalysis. IEEE Trans. on Information Forensics and Security, 2013, 8(12): 1996–2006. [doi: 10.1109/TIFS.2013.2286682]

[20] 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. Atlanta: IEEE, 2014. 185–190. [doi: 10.1109/WIFS.2014.7084325]

[21] Bas P, Filler T, Pevný T. "Break our steganographic system": The ins and outs of organizing BOSS. In: Proc. of the 13th Int'l Workshop on Information Hiding. Prague: Springer, 2011. 59−70. [doi: 10.1007/978-3-642-24178-9_5]

[22] Abdulrahman H, Chaumont M, Montesinos P, Magnier B. Color image steganalysis based on steerable gaussian filters bank. In: Proc. of the 4th ACM Workshop on Information Hiding and Multimedia Security. Vigo Galicia: ACM, 2016. 109–114. [doi: 10.1145/2909827.2930799]

[23] Zeng JS, Tan SQ, Liu GQ, Li B, Huang JW. WISERNet: Wider separate-then-reunion network for steganalysis of color images. IEEE Trans. on Information Forensics and Security, 2019, 14(10): 2735–2748. [doi: 10.1109/TIFS.2019.2904413]

[24] Wei KK, Luo WQ, Tan SQ, Huang JW. Universal deep network for steganalysis of color image based on channel representation. IEEE Trans. on Information Forensics and Security, 2022, 17: 3022–3036. [doi: 10.1109/TIFS.2022.3196265]

[25] Yu ZT, Zhao CX, Wang ZZ, Qin YX, Su Z, Li XB, Zhou F, Zhao GY. Searching central difference convolutional networks for face anti-spoofing. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5295–5305. [doi: 10.1109/CVPR42600.2020.00534]

[26] Yu ZT, Wan J, Qin YX, Li XB, Li SZ, Zhao GY. NAS-FAS: Static-dynamic central difference network search for face anti-spoofing. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2021, 43(9): 3005–3023. [doi: 10.1109/TPAMI.2020.3036338]

[27] Li PH, Xie JT, Wang QL, Gao ZL. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 947–955. [doi: 10.1109/CVPR.2018.00105]

[28] Song XF, Liu FL, Yang CF, Luo XY, Zhang Y. Steganalysis of adaptive JPEG steganography using 2D Gabor filters. In: Proc. of the 3rd ACM Workshop on Information Hiding and Multimedia Security. Portland: ACM, 2015. 15–23. [doi: 10.1145/2756601.2756608]

[29] Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional block attention module. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 3–19. [doi: 10.1007/978-3-030-01234-2_1]

© 中国科学院软件研究所 http://www.jos.org.cn

[30] Fu J, Liu J, Tian HJ, Li Y, Bao YJ, Fang ZW, Lu HQ. Dual attention network for scene segmentation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3141–3149. [doi: 10.1109/CVPR.2019.00326]

[31] Wang QL, Wu BG, Zhu PF, Li PH, Zuo WM, Hu QH. ECA-Net: Efficient channel attention for deep convolutional neural networks. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 11531–11539. [doi: 10.1109/CVPR42600.2020.01155]

[32] Cogranne R, Giboulot Q, Bas P. The ALASKA steganalysis challenge: A first step towards steganalysis. In: Proc. of the 2019 ACM Workshop Information Hiding and Multimedia Security. Paris: ACM, 2019. 125–137. [doi: 10.1145/3335203.3335726]

[33] He KM, Zhang XY, Ren SQ, Sun J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 1026–1034. [doi: 10.1109/ICCV.2015.123]

[34] Bengio Y, Louradour J, Collobert R, Weston J. Curriculum learning. In: Proc. of the 26th Annual Int'l Conf. on Machine Learning. Montreal: ACM, 2009. 41–48. [doi: 10.1145/1553374.1553380]

[35] Yousfi Y, Butora J, Fridrich J, Giboulot Q. Breaking ALASKA: Color separation for steganalysis in JPEG domain. In: Proc. of the 2019 ACM Workshop Information Hiding Multimedia Security. Paris: ACM, 2019. 138–149. [doi: 10.1145/3335203.3335727]

[36] Yousfi Y, Butora J, Khvedchenya E, Fridrich J. ImageNet pre-trained CNNs for JPEG steganalysis. In: Proc. of the 2020 IEEE Int'l Workshop Information Forensics Security. New York: IEEE, 2020. 1–6. [doi: 10.1109/WIFS49906.2020.9360897]

[37] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021.

[38] Tsang CF, Fridrich J. Steganalyzing images of arbitrary size with CNNs. Electronic Imaging, 2018, 30(7): 121-1−121-8. [doi: 10.2352/issn.2470-1173.2018.07.mwsf-121]

[39] You WK, Zhang H, Zhao XF. A siamese CNN for image steganalysis. IEEE Trans. on Information Forensics and Security, 2021, 16: 291–306. [doi: 10.1109/tifs.2020.3013204]

[40] Li H, Wang JW, Xiong N, Zhang Y, Vasilakos AV, Luo XY. A siamese inverted residuals network image steganalysis scheme based on deep learning. ACM Trans. on Multimedia Computing, Communications, and Applications, 2023, 19(6): 214. [doi: 10.1145/3579166]

附中文参考文献:

[1] 黄殿中, 张静飞, 张茹, 李鹏超, 郭云彪. 基于大数据环境的多模态信息隐藏新体系. 电子学报, 2017, 45(2): 477–484. [doi: 10.3969/j.issn.0372-2112.2017.02.029]

[15] 沈军, 廖鑫, 秦拯, 刘绪崇. 基于卷积神经网络的低嵌入率空域隐写分析. 软件学报, 2021, 32(9): 2901–2915. http://www.jos.org.cn/1000-9825/5980.htm [doi: 10.13328/j.cnki.jos.005980]

[16] 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 [doi: 10.13328/j.cnki.jos.006358]

[17] 翟黎明, 嘉炬, 任魏翔, 徐一波, 王丽娜. 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. [doi: 10.19363/J.cnki.cn10-1380/tn.2018.11.01]

[18] 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 [doi: 10.13328/j.cnki.jos.006135]

魏康康(1995－), 男, 博士生, 主要研究领域为隐写分析, 隐写, 深度学习.

刘明林(1991－), 男, 博士, 讲师, 主要研究领域为隐写, 隐写分析, 对抗样本.

骆伟祺(1980－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为数字媒体伪造与检测, 信息隐写与分析.

© 中国科学院软件研究所 http://www.jos.org.cn