

注意力机制和自编码器构造的零水印算法^①



李西明¹, 蔡河鑫¹, 陈志浩¹, 马 莎¹, 杜治国¹, 吕红英²

¹(华南农业大学 数学与信息学院, 广州 510642)

²(华南农业大学 电子工程学院, 广州 510642)

通信作者: 马 莎, E-mail: shamahb@163.com

摘 要: 零水印技术为保护图像版权的有效手段之一. 然而, 现有的许多零水印算法大多采用传统的数学理论进行人工提取特征, 在结合神经网络进行图片特征提取的零水印方向上并没有广泛研究. 目前神经网络在图像特征提取上已经取得了很好的成绩, 充分利用卷积自编码器和注意力机制, 提出了一种用于构造零水印的深度注意自编码器模型 (attention mechanism and autoencoder, AMAE). 首先是利用带有注意力的卷积神经网络构建自编码器, 然后对自编码器进行训练; 其次, 利用训练好的编码器输出的特征构造图像的整体特征; 最后, 将获得的特征图进行二值模式处理得到特征二值矩阵, 再与水印图像异或运算得到零水印, 并在知识产权信息数据库进行注册, 零水印一旦注册, 原图像便处于水印技术的保护下. 在训练过程中, 借鉴对抗训练的思想, 对模型进行加噪训练, 这提高了模型的鲁棒性. 实验结果表明, 本文的零水印算法在旋转、噪声和滤波等攻击下, 提取水印图像与原水印图像的归一化系数 (normalized correlation, NC) 值均超过 0.9, 证明了提出算法的有效性和优越性.

关键词: 自编码器; 零水印; 鲁棒性; 注意力机制; 对抗训练; 深度学习

引用格式: 李西明, 蔡河鑫, 陈志浩, 马莎, 杜治国, 吕红英. 注意力机制和自编码器构造的零水印算法. 计算机系统应用, 2022, 31(9): 257-264.
<http://www.c-s-a.org.cn/1003-3254/8668.html>

Zero-watermarking Algorithm Constructed by Attention Mechanism and Autoencoder

LI Xi-Ming¹, CAI He-Xin¹, CHEN Zhi-Hao¹, MA Sha¹, DU Zhi-Guo¹, LYU Hong-Ying²

¹(College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China)

²(College of Electronic Engineering, South China Agricultural University, Guangzhou 510642, China)

Abstract: Zero-watermarking technology is an effective means of protecting image copyright. However, most of the existing zero-watermarking algorithms use traditional mathematical theories to extract features manually, and extensive research on zero-watermarking extracting image features with neural networks is still to be conducted. At present, neural networks have achieved favorable results in image feature extraction. A deep attention mechanism and autoencoder (AMAE) model is proposed for constructing zero-watermarks by making full use of a convolutional autoencoder and the attention mechanism. Specifically, an attention-based convolutional neural network is used to construct an autoencoder, which is then trained. Subsequently, the global features of the image are constructed with the features output from the trained encoder. Finally, binary pattern processing of the obtained feature image is conducted to acquire the binary feature matrix. An XOR operation with the image to be watermarked is then performed to obtain a zero-watermark, which is then registered into the intellectual property database. Once the zero-watermark is registered, the original image is under the protection of watermarking technology. During training, the idea of adversarial training is drawn on to train the model with noise, which improves the robustness of the model. The experimental results show that the normalized correlation

① 基金项目: 国家自然科学基金 (61872152, 61872409); 2018 年广东省农业厅省级乡村振兴战略专项 (粤农计 (2018)54 号); 广东省基础与应用基础重大项目 (2019B030302008, 2020A151010751); 广州市科技计划 (201902010081)

收稿时间: 2021-12-02; 修改时间: 2021-12-31; 采用时间: 2022-01-13; csa 在线出版时间: 2022-06-17

(NC) values of the extracted watermarked image and the original one to be watermarked both exceed 0.9 under rotation, noise, filtering, and other attacks, which proves the effectiveness and superiority of the proposed algorithm.

Key words: autoencoder; zero-watermarking; robustness; attention mechanism; combat training; deep learning

随着科学技术的发展,数字媒体已经得到广泛应用,越来越多的数字产品在网络中传播,但是,科技给人们带来便利的同时,也给数字产品的信息安全和版权保护带来了日益严峻的问题.为了保护数字信息的安全问题以及版权问题,研究人员提出了数字水印^[1-3].数字水印技术是一种安全可靠的方案,同时也是信息安全领域研究的一个热点.然而在当时,传统数字水印算法存在算法透明性和鲁棒性之间矛盾的问题,原因有二:透明性意味着需要嵌入较弱的水印信号;而更强的水印信号则可以提高算法的鲁棒性.为了解决这个问题,温泉等人^[4]提出了零水印的概念,零水印的思想是利用原始载体图像的内部特征进行构造水印,而不需要修改载体图像的信息,这保证了原始载体图像的完整性.零水印被提出后,也成为了研究的热点之一.郝世博^[5]结合了离散小波变换和奇异值分解,通过比较特征矩阵的每一个系数与特征矩阵均值的大小关系来构造零水印信息,然而该算法对于旋转攻击的鲁棒性较差.张海涛等人^[6]提出的基于超混沌的图像零水印算法解决了零水印鲁棒性不高的问题,但是对于噪声攻击的鲁棒性较差.为了解决该问题,文献[7]和文献[8]都提出了相应的解决方案,文献[7]结合离散小波变换和奇异值分解来构造特征矩阵;文献[8]则结合张量展开、奇异值分解和离散余弦变换来构造图像的特征矩阵.以上的工作都是使用传统的人工方法进行提取特征,而且提取的是图像的几何特征,然而神经网络却可以模拟人类视觉机制,从而提取到图像的视觉特征.此外,若算法需要改进,传统方法可能需要付出更多的努力去探索更优化的方法,基于神经网络的方法则只需要对神经网络的结构参数进行修改并进行重新训练.然而,目前对于利用神经网络进行图像特征提取的零水印方案并没有得到广泛的研究.Fierro-Radilla等人^[9]提出利用卷积神经网络(convolutional neural network, CNN)提取的特征来构造特征矩阵,虽然该算法可以抵抗多种攻击和常见的图像处理,然而卷积神经网络本身也存在缺陷:一是对于传统CNN来说,图像中的权值都是一样的;二是图像中可能存在的干扰会影响卷

积神经网络的分类结果,这是通过影响卷积神经网络对特征的提取所致,这也是卷积神经网络容易受到对抗样本攻击的可能原因.目前,提高模型稳健性的方法有3类:对抗训练、修改模型和添加模型.基于对抗训练^[10]的防御方法在训练过程中加入新的对抗样本,使得神经网络能够更好地了解对抗样本的特征,提高了模型的鲁棒性.

人类视觉注意力机制可以帮助人们快速聚焦目标物体的关键特征,而忽略次要特征^[11-14],引入注意力机制理论上也能使神经网络在图像关键区域投入更多的注意力.受以上思想的启发,本文提出了一种基于注意力机制和卷积自编码器的零水印算法,利用卷积自编码器重构数据的能力来提取图像特征,并结合注意力机制实现对关键特征的稳健提取,训练过程中采用对抗训练,增强了模型的鲁棒性,实验表明,本文算法在受到多种已知攻击的情况下仍能提取图像的稳健特征.

1 理论知识

1.1 自编码器

传统自编码器的概念最开始是由Rumelhart等人^[15]提出的,随后,Bourlard等人^[16]对自编码进行了详细的解释.早期,关于新型自编码器提出的进展还比较缓慢,并且该编码器还只是单层,到了2010年,Vincent等人^[17]又提出了深度去噪自编码器.紧接着,卷积自编码器^[18]、变分自编码器^[19]、循环自编码器^[20]相继被提出.

自编码器是一种无监督学习的人工神经网络,主要应用于数据降维和特征学习.它可以给出比原始数据更好的特征描述,此外,它具有较强的特征学习能力.自编码器包括两部分:一个是编码器,一个是解码器.编码器从原始输入数据提取特征,而解码器则从特征信息中重建原始输入数据,并且使得构建的数据尽可能的等于原始数据.

典型的3层自编码器如图1所示,它由输入层、输出层和一个隐藏层组成.输入层用于原始数据的输入,输出层用于输出特征数据经重构后的数据,而隐藏

层用于特征提取. 输入层和输出层的神经元个数相等, 隐藏层的神经元个数少于输入层和输出层神经元个数. 自编码器通过简单的学习能够使得输出尽可能复制输入, 但重构后的数据与原始数据存在一定的误差. 要使得输出尽可能地等于输入, 则要求隐藏层提取的原始数据特征要更具代表性.

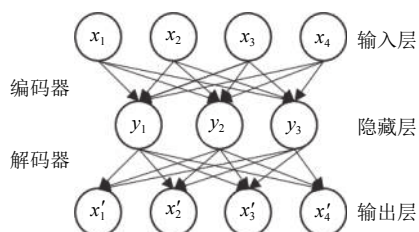


图1 典型3层自编码器

设编码器函数用 encoder 表示, 解码器函数用 decoder 表示, 编码器函数提取到的数据特征为 feature , 即数据特征, 输入数据用 x 表示, 输出数据用 x' 表示. 编码器的作用是将输入数据 x 变换成数据特征 feature , 而解码器是将 feature 转换成输出数据 x' , 整个编码器训练过程就是不断调整参数, 使得 x' 尽可能接近 x . 自编码器可以用图2表示.

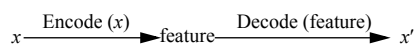


图2 自编码器工作图

1.2 卷积注意力机制

注意力机制是模拟人类视觉而提出的, 它不仅能告诉网络模型需要注意什么特征, 而且也能增强特定区域的表征, 过滤不重要的信息. 本算法使用的注意力模块基于卷积注意力模块 (convolutional block attention module, CBAM), 是由 Woo 等人^[21] 首次提出的, 该方案不同于通道注意力机制^[22], 能显著提升模型的特征表达能力.

如图3所示, 给定一个中间特征图 F 作为输入, CBAM 依次经过通道注意力模块和空间注意力模块. 首先是通道注意力模块. 将中间特征图 F 对每个通道进行最大池化和平均池化得到最大池化特征 $F_{\max}$ 和平均池化特征 F_{avg} . 然后将二者分别输入到含有一个隐藏层的多层感知机 (multilayer perceptron, MLP) 中, 使用元素求和法来合并输入的特征向量, 最后经过激活函数得到通道注意力 $Mc(F)$. 通道注意力的计算过程如

式(1):

$$Mc(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \quad (1)$$

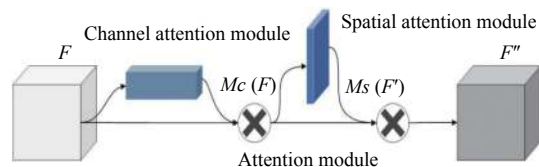


图3 卷积注意力模块

最后, 进入空间注意力模块. 先沿着通道轴应用平均池化和最大池化操作, 生成平均池化特征图和最大池化特征图: $F_{\text{avg}}^s \in R^{1 \times H \times w}$ 和 $F_{\max}^s \in R^{1 \times H \times w}$. 然后这两个特征图通过一个标准的卷积层和 Sigmoid 函数后生成二维空间注意力图 $Ms(F')$. 空间注意力计算过程可以用式(2)进行表示:

$$\begin{cases} Ms(F) = \sigma(f^{7 \times 7}([AvgPool(F); MaxPool(F)])) \\ Ms(F') = \sigma(f^{7 \times 7}([F_{\text{avg}}^s; F_{\max}^s])) \end{cases} \quad (2)$$

其中, $AvgPool(F)$ 表示平均池化, $MaxPool(F)$ 表示最大池化, σ 表示 Sigmoid 函数, $f^{7 \times 7}$ 表示卷积核尺寸为 7×7 的卷积运算. 整个卷积注意力模块的过程可以用式(3)来进行概括:

$$\begin{cases} F' = Mc(F) \otimes F \\ F'' = Ms(F') \otimes F' \end{cases} \quad (3)$$

其中, $\otimes$ 表示按元素计算的乘法.

2 AMAE 架构设计

本节介绍了基于注意力机制和自编码器的网络架构的详细设计, 同时说明了训练过程.

2.1 网络架构描述

本实验的目标是设计一个能够提取稳健图像特征的网络架构, 这也意味着我们设计的网络需要具备抵抗噪声干扰等的能力, 该网络架构是基于自编码器提出的. 自编码器具有重构数据的能力, 能够很好地提取数据的特征, 此外, 相对于传统的图像特征提取的方法, 卷积神经网络可以更好地提取图像的特征, 所以本实验用卷积层和池化层代替了传统自编码器的全连接层, 而由于加了干扰的图片会影响卷积神经网络的特征提取, 再加上卷积神经网络也无法聚焦在图片的关键特征上, 所以我们需要让卷积自编码器能够将注意力更

多地关注于关键特征,从而保证算法的稳健性. 综上,我们在卷积自编码的基础上加入注意力机制,这可以让网络过多关注图片的关键特征,而忽视图片上存在

的类似于噪声的无关特征.

网络的架构图如图4所示,分为注意力机制编码器和注意力机制解码器两部分.

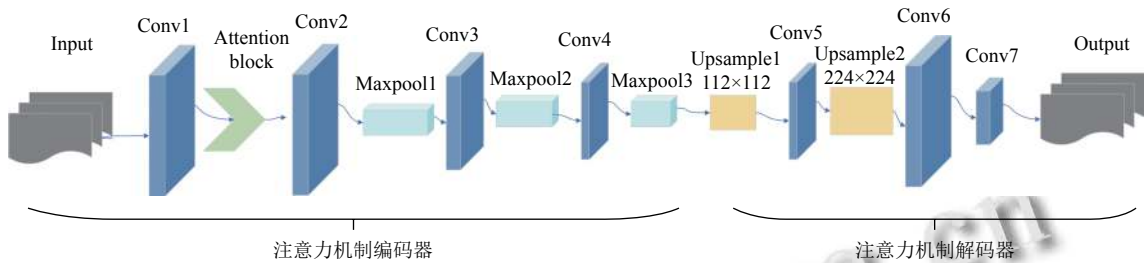


图4 网络架构图

注意力机制编码器: 将大小为 (N, C, H, W) 的数据input输入到编码器中, 首先是经过一个包含卷积核大小为 $(3, 3)$ 的卷积层 Conv1, 然后将其输入到注意力模块中, 输出具有显著特征的掩膜图像 $T1$, 大小为 $(N, 32, 112, 112)$, 紧接着再经过3层卷积层和最大池化层, 分别是 Conv2、Maxpool1、Conv2、Maxpool1、Conv2、Maxpool1, 变为 $(N, 16, 56, 56)$ 的张量 $T2$. 最后将 $T2$ 输入到解码器.

注意力机制解码器: $T2$ 先经过采用了最近邻算法的上采样 Upsample1 和卷积核大小为 $(3, 3)$ 的卷积层 Conv5 得到大小为 $(N, 16, 112, 112)$ 的张量 $T3$, 再经过采用了最近邻算法的上采样 Upsample1 和卷积核大小为 $(3, 3)$ 的卷积层 Conv6 得到大小为 $(N, 16, 224, 224)$ 的张量 $T4$, 最后经过一层卷积核大小为 $(3, 3)$ 的卷积层 Conv7 实现数据的复原, 得到 output. 网络的结构参数如表1.

表1 网络参数表

种类	In_channels	Out_channels	Kernel_size	Stride	Padding
Conv1	3	32	(3, 3)	1	1
Conv2	3	32	(3, 3)	1	1
Maxpool1	—	—	(2, 2)	2	0
Conv3	32	32	(3, 3)	1	1
Maxpool2	—	—	(2, 2)	2	0
Conv4	32	16	(3, 3)	1	1
Maxpool3	—	—	(2, 2)	2	0
Conv5	16	16	(3, 3)	1	1
Conv6	16	32	(3, 3)	1	1
Conv7	32	3	(3, 3)	1	1

2.2 训练过程

为了提高网络模型的鲁棒性, 我们在训练网络的时候借鉴了对抗训练的思想, 目标是使用随机初始化

的权重来训练一个具有鲁棒性的网络模型, 实验中添加的扰动权重值在 0-1 之间随机选取, 添加扰动后的图像样本如图5所示, 训练集由原始数据集和添加扰动的数据集组成, 并且随着迭代次数的增加, 扰动数据集数量也会随之增加.

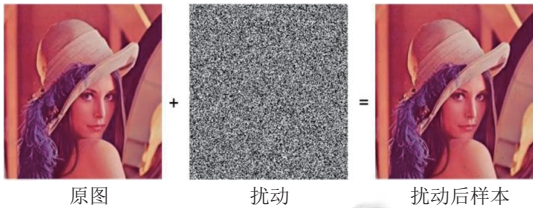


图5 扰动样本生成图

网络训练的最终目的是使输入无限接近于输出, 详细的训练过程如下: 训练集合 D 上的数据 x, y 为模型的输出, 网络的损失函数为 $L(x, y, \theta)$, 本实验使用的是交叉熵函数, θ 为网络模型的参数, Δx 为扰动. 训练的前将扰动 Δx 加入到部分数据 x 中, 并随着迭代次数的增加, 添加扰动的数据比例会上升, 其目标是使得 $L(x + \Delta x, y, \theta)$ 越来越大, 也就是说该扰动尽可能让神经网络重构后的数据与训练集数据差别越来越大. 在利用原始样本都构造出 $x + \Delta x$ 扰动样本后, 训练的目标就是利用梯度下降法来找到能够最小化网络输出和输入差 $E_{(x,y)} \sim D[L(x, y, \theta)]$ 的合适参数. 在不断的迭代过程中, 持续地优化参数, 整个优化过程中是最大化和最小化交替执行, 详细公式如式 (4). 此过程类似于生成式对抗网络 (generative adversarial networks, GAN), 然而也不同于 GAN, 因为该训练过程的输入扰动过程为最大化过程, 调整参数为最小化过程.

该模型是在 ImageNet 数据集上进行训练, 迭代次数为 100 次, 训练步长为 0.001, 该网络训练的损失值变化如图 6 所示, 可见随着训练迭代次数的增加, 损失值逐渐下降, 并在迭代次数大于 20 后, 趋于稳定, 可以见得本方案的模型训练速度很快, 很快就达到拟合状态。

$$\min_{\theta} E_{(x,y)} \sim D \left[\max_{\Delta X \in \Omega} L(x + \Delta x, y, \theta) \right] \quad (4)$$

其中, Ω 为扰动空间。

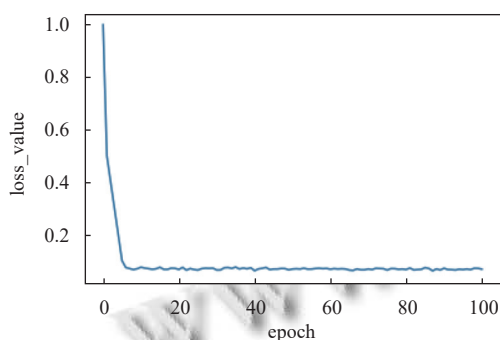


图 6 损失值变化图

3 AMAE 零水印算法

本节给出基于注意力机制和自编码器的零水印算法的详细说明。

3.1 零水印算法描述

自编码器的零水印算法分为零水印的构造和水印的提取两部分。水印构造包括 3 个步骤, 分别是提取图像特征、获取二值矩阵、构造零水印。水印提取包括 3 个步骤, 分别是获取待测图像特征、获取二值矩阵、恢复水印。

3.1.1 水印构造

该部分从宿主照片提取特征构造零水印, 如图 7(a) 所示, 包括以下步骤, 如算法 1 所示。

算法 1. 零水印构造算法

1) 提取图像特征

当基于注意力机制的卷积自编码器训练好后, 该网络便能够提取图像的稳定特征, 我们利用编码器的输出来构造图像的特征, 将一张需要提取特征的大小为 (3, 224, 224) 的图像输入到编码器, 从网络架构图中可以看出, 编码器的输出大小为 (16, 56, 56), 也就是 16 张大小为 (56, 56) 的特征图, 再将这 16 张特征图融合成大小为 (224, 224) 的特征图 A, 如图 8(a), 融合策略如下:

首先将特征图标号为 1, 2, ..., 6, 然后每 4 张大小为 (56, 56) 特征图拼接成大小为 (224, 224) 的特征子图 F_1, F_2, F_3, F_4 , 最后利用式 (5) 进行加权融合得到最终的特征图 A。

$$A = \sum_{i=1}^4 \frac{1}{4} F_i \quad (5)$$

2) 获取二值矩阵

利用矩阵 A 的每个元素的值 $A_{x,y}$ 与矩阵 A 的均值 T 的大小关系构造二值矩阵 C, 如式 (6) 所示。

$$C_{x,y} = \begin{cases} 1, & \text{if } A'_{x,y} > T \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

3) 构造零水印

将矩阵 C 与水印图像矩阵 W 进行异或运算得到零水印 M, 紧接着, 零水印 M 在知识产权信息数据库进行注册, 使版权信息得到保存。一般认为, 一旦零水印得到注册, 也意味着该载体图像处于水印技术的保护中, 当发现有侵权现象时, 可取出零水印进行版权认证, 从而实现对图像所有权的保护。M 的计算方法如式 (7) 所示:

$$M = XOR(C, W) \quad (7)$$

3.1.2 水印提取

水印的提取即为水印构造的逆过程, 如图 7(b) 所示, 包括以下步骤, 如算法 2 所示。

算法 2. 零水印提取算法

1) 获得待测图像特征

将待检测载体图片作为自编码器输入, 然后取编码器输出构造载体图片特征矩阵 A'。

2) 获取二值矩阵

利用矩阵 A' 的每个元素的值 $A'_{x,y}$ 与矩阵 A' 的均值 T' 的大小关系构造二值矩阵 C', 如式 (8) 所示。

$$C'_{x,y} = \begin{cases} 1, & \text{if } A'_{x,y} > T' \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

3) 恢复水印

将矩阵 C' 与零水印 M 进行异或运算得到所恢复的水印图像 W', 如式 (9) 所示。

$$W' = XOR(C', M) \quad (9)$$

3.2 实验结果分析

本文实验是在 PyCharm 实验平台上进行仿真, 并利用 PyTorch 框架进行编码实现。在本文的零水印算法中, 引入归一化相关系数 NC, 它是衡量提取到的水印图像与原始水印图像之间相近程度的一个度量工具, NC 值的范围在 0 到 1 之间, 该值越接近于 1, 表明提取出来的水印越接近原始水印, NC 值的计算方法如式 (10) 所示。

$$NC(W, W') = \frac{\sum_{i=1}^M \sum_{j=1}^N W(i, j) W'(i, j)}{\sqrt{\sum_{i=1}^M \sum_{j=1}^N W(i, j)^2} \sqrt{\sum_{i=1}^M \sum_{j=1}^N W'(i, j)^2}} \quad (10)$$

其中, W 是指原始水印图, W' 指的是从待测载体图像上提取出来的水印图像。

本实验使用的水印图片和载体图片如图 8(b) 与图 8(c) 所示。

3.2.1 鲁棒性分析

本文分别对载体图像进行不同种类和不同强度的

攻击, 攻击样例如图 9, 并计算不同的 NC 值, 实验结果如表 2 所示.

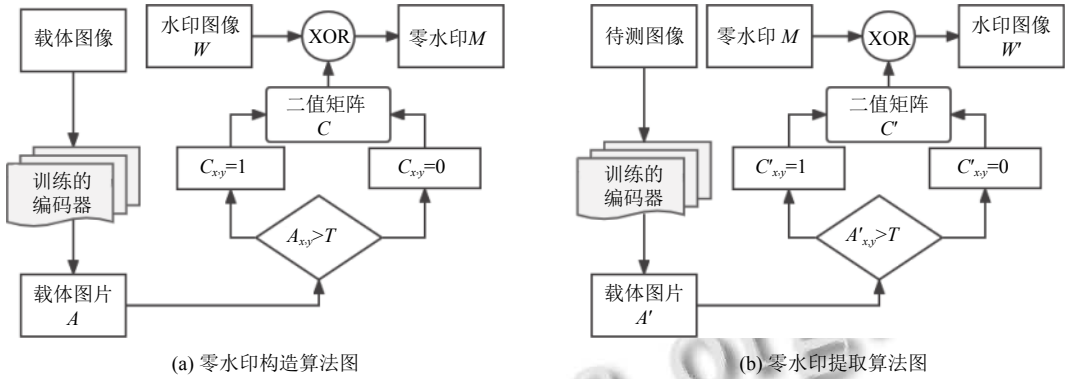


图 7 零水印算法图



图 8 特征图、水印图和载体图

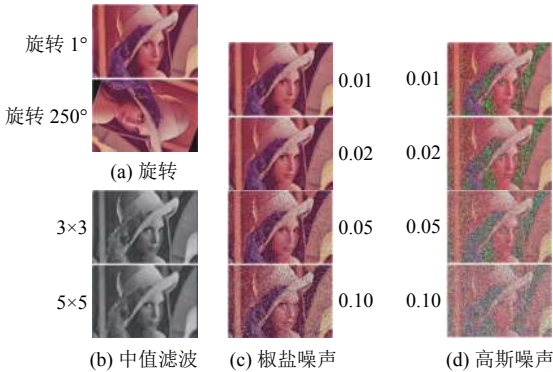


图 9 攻击样例图

从表 2 可知, 载体图片不管是受到几何攻击还是非几何攻击, 提取出的水印的 NC 值均超过 0.9, 由此可见, 该算法能很好地抵抗多种攻击, 具有良好的鲁棒性.

3.2.2 实验对比分析

(1) 神经网络实验对比分析

为了验证本模型的有效性, 我们做了 3 组对比实验. 分别是: 对抗训练与正常训练情况下的对比, 训练集减半情况下的对比, 添加注意力机制与不添加注意力机制的对比, 实验所得的 NC 值分别对应于表 3 中的 $NC1$ 、 $NC2$ 和 $NC3$.

从实验结果来看, 不加注意力机制和不进行对抗训练的模型在抵抗攻击的时候并没本论文提出的模型

表现得好, 这也体现出本实验提出的模型的有效性, 此外, 减半数据集后的实验结果相较于完整数据集的结果, NC 值相差在 0.01 以内, 这也说明了本方案小数据集上, 也可以得到良好的零水印.

(2) 零水印算法实验结果对比

本文结合注意力机制和编码器的输出来构造特征矩阵, 文献 [23] 在小波变换域提取的低频区域进行分块的奇异值分解, 再利用分块的最大奇异值来构造特征矩阵, 文献 [24] 则是选择在时域上使用非均匀 NURP 来进行特征矩阵的构造, 不同方法的实验结果如表 4 所示.

表 2 待测图像攻击后提取出的水印的 NC 值

攻击	强度	NC
旋转攻击	向左 1°	0.999 871 238 720 675
	向左 250°	0.999 906 354 162 537
椒盐噪声	0.01	0.999 847 827 466 87
	0.02	0.999 625 461 862 405
	0.05	0.998 138 559 240 155
	0.1	0.993 036 408 583 17
中值滤波	3×3	0.999 982 943 456 889
	5×5	0.999 894 649 974 479
高斯噪声	0.01	0.998 665 089 891 95
	0.02	0.996 836 913 182 36
	0.05	0.986 264 330 489 146
	0.1	0.968 769 108 575 163

从表 4 可知, 在抵抗噪声攻击方面, 文献 [23,24] 的性能明显比本文实验差, 尤其是在抵抗高斯噪声方面, 本实验比文献 [24] 的 NC 值高了近 0.1. 此外, 从整体来看, 本文实验的 NC 值普遍都要比文献 [23,24] 的高, 而且均在 0.9 以上, 通过对比实验可以得出本文算法的鲁棒性更强.

表 3 NC 值对比

攻击	强度	NC	NC1	NC2	NC3
旋转攻击	向左1°	0.999 971 238 720 675	0.976 383 348 457 811	0.992 611 451 767 347	0.952 666 901 814 73
	向左250°	0.999 806 354 162 537	0.964 463 953 067 063	0.990 153 713 252 474	0.923 468 835 288 987
椒盐噪声	0.01	0.999 847 827 466 87	0.993 506 117 824 039	0.998 148 862 541 674	0.983 790 268 492 007
	0.02	0.999 625 461 862 405	0.990 249 594 617 071	0.996 270 787 849 92	0.975 743 025 449 604
	0.05	0.998 138 559 240 155	0.978 117 269 860 505	0.990 650 652 723 307	0.961 218 828 997 796
	0.1	0.993 036 408 583 17	0.951 747 756 614 324	0.977 962 568 272 038	0.936 765 637 825 744
中值滤波	3×3	0.999 982 943 456 889	0.985 331 454 110 338	0.997 691 407 955 363	0.964 086 093 220 478
	5×5	0.999 894 649 974 479	0.984 713 472 474 983	0.997 374 585 525 778	0.957 603 472 473 661
高斯噪声	0.01	0.998 665 089 891 95	0.983 907 830 734 522	0.993 882 875 435 024	0.957 985 267 432 987
	0.02	0.996 836 913 182 36	0.978 294 252 425 983	0.991 040 544 250 835	0.949 748 801 418 943
	0.05	0.986 264 330 489 146	0.962 518 446 778 142	0.979 218 367 353 522	0.936 634 961 984 841
	0.1	0.968 769 108 575 163	0.940 625 823 827 975	0.957 841 244 816 34	0.924 099 153 158 844

表 4 不同文献 NC 值对比

攻击	强度	文献[23]	文献[24]	本文
旋转攻击	向左1°	—	—	0.999 971 238 720 675
	向右1°	—	—	0.999 906 354 162 537
椒盐噪声	0.01	0.996 8	0.994 2	0.999 847 827 466 87
	0.02	0.992 3	0.937 7	0.999 625 461 862 405
	0.05	—	0.972 4	0.998 138 559 240 155
	0.1	—	0.962 1	0.993 036 408 583 17
中值滤波	3×3	0.996 8	0.936 5	0.999 982 943 456 889
	5×5	0.992 9	0.996 8	0.999 894 649 974 479
高斯噪声	0.01	0.991 1	—	0.998 665 089 891 95
	0.02	0.985 2	—	0.996 836 913 182 36
	0.05	—	0.901 0	0.986 264 330 489 146
	0.1	—	0.852 1	0.968 769 108 575 163

4 结论与展望

结合卷积注意力模块和卷积编码器,提出了一种用于构造零水印的深度注意自编码器模型,该算法利用卷积自编码器重构数据的能力对图像特征进行提取,并结合注意力机制,对关键位置给予更多的关注,忽视了类似于噪声等攻击的无关特征.在训练的过程中,结合对抗训练,进一步提高了模型的鲁棒性,从而保证了图像特征的稳定提取.实验结果表明,在载体图片受到几何攻击和非几何攻击下所提取出的水印 NC 值均在 0.9 以上,该算法具有很好的鲁棒性.是否添加注意力机制、是否减半数据集和是否进行对抗训练等 3 个不同的对比实验可以证明提出的自编码器模型的有效性,从与传统方法的实验结果对比,可以发现,相较于传统方法,该算法鲁棒性更好.

参考文献

- van Schyndel RG, Tirkel AZ, Osborne CF. A digital watermark. Proceedings of the 1st International Conference on Image Processing. Austin: IEEE, 1994. 86–90.
- Cox IJ, Kilian J, Leighton FT, *et al.* Secure spread spectrum watermarking for multimedia. IEEE Transactions on Image Processing, 1997, 6(12): 1673–1687.
- Boland FM, O’Ruanaidh JJK, Dautzenberg C. Watermarking digital images for copyright protection. Proceedings of the 5th International Conference on Image Processing and Its Applications. Edinburgh: IET, 1995. 326–330.
- 温泉, 孙钊锋, 王树勋. 零水印的概念与应用. 电子学报, 2003, 31(2): 214–216. [doi: 10.3321/j.issn:0372-2112.2003.02.015]
- 郝世博. 针对彩色图像的新型零水印算法. 计算机应用与软件, 2013, 30(6): 298–301, 311. [doi: 10.3969/j.issn.1000-386x.2013.06.079]
- 张海涛, 张思博. 基于超混沌的图像零水印算法. 计算机应用研究, 2019, 36(11): 3387–3390, 3409.
- 齐向明, 张晶, 谭昕奇. 基于低频奇异值均值的强鲁棒零水印算法. 计算机工程, 2019, 45(12): 214–221.
- Jiang FF, Gao TG, Li D. A robust zero-watermarking algorithm for color image based on tensor mode expansion.

- Multimedia Tools and Applications, 2020, 79(11): 7599–7614.
- 9 Fierro-Radilla A, Nakano-Miyatake M, Cedillo-Hernandez M, *et al.* A robust image zero-watermarking using convolutional neural networks. Proceedings of the 2019 7th International Workshop on Biometrics and Forensics. Cancun: IEEE, 2019. 1–5.
- 10 Kurakin A, Goodfellow IJ, Bengio S. Adversarial machine learning at scale. arXiv: 1611.01236, 2016.
- 11 Wang XD, Cai ZW, Gao DS, *et al.* Towards universal object detection by domain attention. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7281–7290.
- 12 Fan H, Ling HB. Siamese cascaded region proposal networks for real-time visual tracking. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7944–7953.
- 13 Burt PJ. Attention mechanisms for vision in a dynamic world. Proceedings of the 9th International Conference on Pattern Recognition. Rome: IEEE, 1988. 977–987.
- 14 Itti L, Koch C. A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research, 2000, 40(10–12): 1489–1506.
- 15 Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature, 1986, 323(6088): 533–536.
- 16 Bourlard H, Kamp Y. Auto-association by multilayer perceptrons and singular value decomposition. Biological Cybernetics, 1988, 59(4–5): 291–294.
- 17 Vincent P, Larochelle H, Lajoie I, *et al.* Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. The Journal of Machine Learning Research, 2010, 11(12): 3371–3408.
- 18 Masci J, Meier U, Cireşan D, *et al.* Stacked convolutional auto-encoders for hierarchical feature extraction. Proceedings of the 21st International Conference on Artificial Neural Networks. Espoo: Springer, 2011. 52–59.
- 19 Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv: 1312.6114, 2014.
- 20 Cho K, van Merriënboer B, Gulcehre C, *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014. 1724–1734.
- 21 Woo S, Park J, Lee JY, *et al.* CBAM: Convolutional block attention module. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 3–19.
- 22 Hu J, Shen L, Albanie S, *et al.* Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011–2023.
- 23 刘万军, 孙思宇, 曲海成, 等. 一种抗几何旋转攻击零水印算法. 计算机应用研究, 2019, 36(9): 2803–2808.
- 24 Shen ZL, Kintak U. A novel image zero-watermarking scheme based on non-uniform rectangular. Proceedings of 2017 International Conference on Wavelet Analysis and Pattern Recognition. Ningbo: IEEE, 2017. 78–82.

(校对责编: 孙君艳)