基于卷积痕迹挖掘的 GAN 生成假脸图片检测

罗正军,张丽丽
(南京航空航天大学 经济与管理学院,江苏 南京 211106)

摘 要:虚假人脸的生成与篡改带来的安全威胁已引起广泛关注。针对目前大部分研究需分别对不同 GAN(生成对抗网络)生成的虚假图片训练检测模型,很难提取通用特征检测虚假图片,导致模型泛化能力不足的问题,提出了一种基于虚假图片生成过程中转置卷积层造假痕迹挖掘的 GAN 生成人脸虚假图片检测模型。首先基于虚假图片像素局部相关性和对比损失函数改进图片特征向量提取框架,再利用粒子群算法改进最大期望算法构成 EM-PSO(最大期望-粒子群)算法优化特征向量求解过程,进而获取模型在图片 RGB 三通道计算得到的特征向量。通过支持向量机和图片特征向量实现虚假图片检测。实验结果表明:在由 FFHQ 真实人脸数据、StyleGAN 和 StyleGAN2 生成的假脸数据构成的模型训练数据集上,模型检测准确率最高可达 94.25%,AUC 值可达 0.99,模型检测准确率明显优于 VGG16 模型在此数据集上的检测准确率,由此验证了该模型的有效性。

关键词:生成对抗网络;对比损失;卷积痕迹;假脸图片;特征提取;虚假检测
中图分类号:TP309 **文献标识码**:A **文章编号**:1673-629X(2022)07-0052-06
doi:10.3969/j.issn.1673-629X.2022.07.009

Research on GAN-Generated False Face Image Detection Based on Convolution Trace Mining

LUO Zheng-jun, ZHANG Li-li
(School of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: The threat caused by the generation and tampering of fake faces has attracted widespread attention. But most of the current research needs to train detection models respectively for the fake images generated by different GANs (Generative Adversarial Networks) since they are hard to find the common discriminative features for judging the fake images, resulting in insufficient model generalization ability. A GAN-generated fake face images detection model based on forensics trace mining in transposed convolution layer is proposed for the insufficient generalization ability of existing models. Firstly, the image feature vector extraction framework is optimized based on the principle of GAN-generated image pixel local correlation and contrastive loss function. And then the particle optimization swarm algorithm and the maximum expectation algorithm are used to compose the EM-PSO (maximum expectation-particle swarm) algorithm, which will optimize model solution process and obtain the feature vector from RGB three-channel in convolution process. Finally, support vector machine uses feature vectors to detect fake images. Experiments were performed on data set composed of FFHQ real face data and fake face data generated by StyleGAN and StyleGAN2. The results demonstrate that the detection accuracy of the proposed model reaches 94.25%, and the AUC value reaches 0.99. The detection accuracy of proposed model is superior to the VGG16 model, verifying the effectiveness of the proposed model.

Key words: generative adversarial networks (GAN); contrastive loss; convolutional trace; fake face images; feature extraction; fake detection

0 引 言

随着以生成对抗网络为代表的深度学习生成技术不断提高,虚假图片的仿真度愈来愈高,人脸的生成与篡改所带来的安全威胁已引起广泛关注。Choi 等[1]利用 Inception V3、VGG16、VGG19 和 ResNet50 四种网络结构提取图片特征。Li 等[2]在 CelebA 数据集上测试了 GAN 自有判别器和 VGG16 网络分类能力。Andreas 等[3]创建并公开了大规模假脸图像数据集

收稿日期:2021-06-28 **修回日期**:2021-10-29
基金项目:国家自然科学基金(71373123);中央高校基本科研业务费专项资金资助(ND2021002)
作者简介:罗正军(1972-),男,硕士,副教授,硕导,研究方向为管理信息系统、数据分析、系统仿真;张丽丽(1996-),女,硕士研究生,研究方向为大数据分析。

FaceForensics++,并在此数据集上测试了多种 CNN 模型。Dang 等[4]构建 CNN 模型提取图片特征,分类准确率优于 VGG 模型。Fu 等[5]设计了基于 CNN 的双通道结构,从原始 RGB 图像和预处理图像的高通分量中提取图片特征。Tariq 等[6]基于浅层神经网络模型用来检测 GAN 生成的人脸图片和人工创建的人脸图片。Bonettini 等[7]利用本福德定律、Luca[8]利用转置卷积痕迹、Scott 等[9]基于颜色饱和度和曝光情况提取图片特征,再对 GAN 生成虚假图片进行分类。

GAN 能够按帧生成虚假视频,因此可以将图片检测技术推广到 DeepFake 虚假视频的检测中[10-11]。李旭嵘等[12]提出了基于 EfficientNet 的双流网络检测模型,赵磊等[13]提出了一种基于时空特征一致性的检测模型,陈鹏等[14]利用全局时序特征和局部空间特征检测虚假视频。胡永健等[15]利用神经分割网络预测篡改区域进行假脸视频跨库检测。张怡暄等[16]利用视频相邻帧人脸图像差异的特征来进行预测。

考虑到已有模型较难提取通用特征用于区分多种不同 GAN 生成的虚假图片,导致现有模型泛化能力不足。该文改进了 Luca 等人[8]的图片特征提取模型,基于对比损失函数和图片像素局部相关性优化图片特征提取过程,同时结合梯度下降和粒子群算法优化模型求解过程。最后在 FFHQ 数据集上进行模型检测性能测试,该模型在检测准确率和 AUC 取值方面有着很好的表现,检测准确率明显优于 VGG16 模型。

1　GAN 生成机理

生成对抗网络[17]是一种生成模型,其训练过程是对抗博弈的过程,模型结构如图 1 所示。

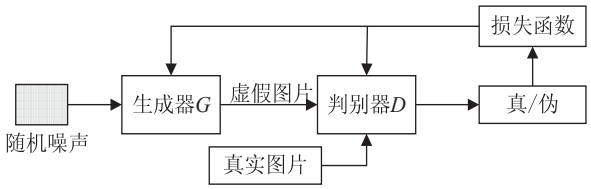


图 1　GAN 模型结构

GAN 的主要结构包括一个生成器 G(Generator)和一个判别器 D(Discriminator),生成器的目标函数为式(1),作用是根据输入的一组向量生成图片,判别器的目标函数为式(2),作用是区分从训练数据抽取的真实图片和从生成器抽取的图片。在迭代过程中判别器学习正确地区分真假图片,生成器试图欺骗判别器让其相信生成的图片是真的,当判别器的识别能力达到一定程度却无法正确识别数据来源时,就得到了一个能够生成清晰度和分辨率较高图片的生成器。

$$G^* = \arg\min_G \max_D V(G,D) \tag{1}$$

$$D^* = \arg\max_D V(D,G) \tag{2}$$

生成器主要由转置卷积层、池化层和全连接层构成,转置卷积层是一个上采样过程,利用一个卷积核对输入的多维数据进行卷积计算进而得到一个维度更高的输出数据。从 GAN 图片生成过程来看,生成器中各层所执行的操作都会对像素产生影响,特别是进行上采样的转置卷积层。在一张图片像素的计算过程中,卷积核是始终保持不变的,会造成图片像素的内在关系。

2　检测模型设计与实现

检测模型的总体结构如图 2 所示。

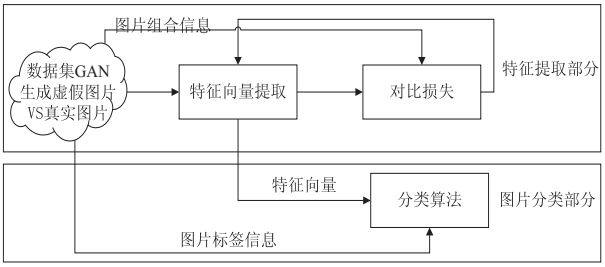


图 2　图片检测模型结构

检测模型主要由特征提取模块、对比损失模块以及机器学习分类模块三个主要部分构成。在特征提取模块完成特征向量求解过程,满足模型约束条件,满足约束条件的两张图片的卷积核及其相应的图片标签信息在对比损失模块完成欧氏距离的计算获得优化后的特征向量。最后利用机器学习算法对图片进行分类。

2.1　基于像素局部相关性思想构建特征向量提取模型

GAN 生成虚假图像的像素点之间存在的局部相关性可以用式(3)表示。即存在一个 $N \times N$ 的卷积核 K 使像素点 $I_0(x,y)$ 周边的像素点经过卷积计算和该像素点存在近似关系。

$$I_0(x,y) \sim I_1(x,y)$$

$$I_1(x,y) = \sum_{s,t=-\alpha}^{\alpha} k_{s,t} * I_0[x+s, y+t] \tag{3}$$

本模型的目的是在图片的 RGB 每个通道中各寻找一个 $N \times N$ 的卷积核能够表征图片像素之间的局部相关关系,而该卷积核 K 作为图片的特征向量会使 GAN 生成的图片区别于真实图片。用方差表示新估计的图片与原始图片之间的近似程度,如公式(4)所示。

$$E = \frac{1}{N} \sum_{x,y} \left(I(x,y) - \overline{I(x,y)} \right)^2 \tag{4}$$

其中,$\overline{I(x,y)}$ 表示图片像素的均值。对于一张图片所有的像素点而言,卷积计算得到的图片像素点的方差需落在原图片像素点方差 E_0 的某一邻域内。由此实现对特征向量取值范围的约束。

基于对比损失函数优化特征向量提取过程。对比损失(contrastive loss)通常用在传统的孪生神经网络中,可以通过计算两个样本特征的欧氏距离来有效表达成对样本的匹配程度,还可以用于训练特征提取的模型。对数据集中的所有图片采用随机匹配的方式,每两张图片之间进行对比,通过对比损失计算,评价两张图片的相似度,按照同类图片相似度高,异类图片相似度低的原则,继续迭代计算卷积核,进而实现特征向量提取过程的优化。对比损失函数的表达式如公式(5):

$$L(W,(Y,K_1,K_2)) = \frac{1}{2}(YD_k^2 + (1 - Y)\max(m - D_k,0)^2) \quad (5)$$

其中,D_k 表示特征向量 K_1 和 K_2 的欧氏距离,Y 是两张图片是否为一类的标签,$Y=1$ 代表两个样本都是虚假图片或都是真实图片,$Y=0$ 代表两个样本不属于一类。当 $Y=1$ 时,损失函数为 $\frac{1}{2}D_k^2$,即相似的样本,如果欧氏距离较大,则说明当前的模型较差,需降低损失。当 $Y=0$ 时,损失函数为 $\frac{1}{2}\max(m - D_k,0)^2$,即不相似的样本,如果欧氏距离较小,则说明模型较差,应该增大其欧氏距离,降低损失。综上,在应用中应求解对比损失函数的最小值。

经过优化后,模型需要求解的损失函数为公式(6),需要满足约束条件(7)~(10),其中 Δ 为方差的可变动范围,且 $k_{0,0}$ 一直为 0。因此,模型分别在 RGB 三通道求解得到 $N \times N$ 的卷积核 K,最终形成 $N \times N - 3$ 维的特征向量。

$$\min L(W,(Y,K_i,K_j)) \quad (6)$$

$$E_i - \Delta \leqslant \frac{1}{N}\sum_{x,y}(I_i(x,y) - \overline{I_i(x,y)})^2 \leqslant E_i + \Delta \quad (7)$$

$$E_j - \Delta \leqslant \frac{1}{N}\sum_{x,y}(I_j(x,y) - \overline{I_j(x,y)})^2 \leqslant E_j + \Delta \quad (8)$$

$$k_{i_{0,0}} = 0 \quad (9)$$

$$k_{j_{0,0}} = 0 \quad (10)$$

2.2 基于 EM-PSO 算法的优化求解设计

模型求解主要可以分为两部分,一部分是对比损失部分特征向量的求解,另一部分是约束条件方差 E 的求解。

在对比损失部分的迭代求解中综合考虑方差 E 的约束和损失函数 L 最小的目标,图片 i 和图片 j 分别按照公式(11)和(12)迭代更新卷积核 K_i,K_j。在对比损失部分按照排列组合的方式会有 C_n^2 种组合方案(n 表示训练数据集中图片的张数),为了简化实验,本模型针对每一张图片 i,随机读取一张图片 j,循环完所有的图片 i 构成一次完整的迭代,获得数据集中所有图片的特征向量 $K_1,K_2,\cdots,K_m$。

$$\text{Image}_i: K_{t+1} = K_t - \lambda_1(\frac{\partial E}{\partial k_{s,t}}) - \lambda_2(\frac{\partial L}{\partial k_{s,t}}) \quad (11)$$

$$\text{Image}_j: K_{t+1} = K_t - \lambda_1(\frac{\partial E}{\partial k_{s,t}}) - \lambda_2(\frac{\partial L}{\partial k_{s,t}}) \quad (12)$$

在方差求解部分利用 EM 算法更新均值 $\bar{I}$ 和方差 E。对于卷积核 K 的求解,利用基于 EM-PSO 算法实现优化求解,即在迭代过程中综合利用梯度下降和粒子群算法。

EM 算法主要依赖梯度下降完成迭代过程,梯度下降法的计算过程就是沿梯度的负方向搜索最小值,但在多维数据的求解中只考虑局部梯度易陷入局部最优造成方差早熟收敛的问题,使方差无法收敛到约束范围内。

PSO 算法常用于求解优化问题,初始化一群随机粒子(随机解),通过迭代找到最优解。该算法在迭代求解过程中既考虑了每个粒子的历史最优,又考虑了种群的全局最优,但忽略了目标函数本身的变化规律。在求解过程中可以通过求解梯度的方式获知函数本身的变化规律,为了能充分利用这两种方法的优势,并克服早熟收敛的缺陷,在迭代过程中综合梯度下降和粒子群算法,利用粒子群算法的全局思想改善梯度下降易陷入局部最优的缺点,形成 EM-PSO 算法。即在各个种群的迭代过程中,将种群受自身的最佳历史位置影响改为每个种群受梯度下降方向影响,如公式(13)所示。EM-PSO 算法的整体求解过程如算法 1 的伪代码所示。

$$K_t = K_{t-1} - \lambda_1(\frac{\partial E_{t-1}}{\partial k_{s,t}}) + \lambda_2(K_{\text{gbest}} - K_{t-1}) \quad (13)$$

$$\frac{\partial E}{\partial k_{i,j}} = \frac{2}{N}\sum_{x,y}(I_i - \bar{I})(\frac{\partial I_i}{\partial k_{i,j}} - \frac{\partial I_1}{\partial k_{i,j}}) = \frac{2}{N}\sum_{x,y}(\sum_{s,t=\alpha}^{\alpha} k_{s,t} * I_0[x+s,y+t] - \frac{1}{N}\sum_{x,y}(\sum_{s,t=\alpha}^{\alpha} k_{s,t} * I_0[x+s,y+t]))(I_0[x+i,y+j] - \frac{1}{N}\sum_{i=1}^{N} I_i[x+i,y+j]) \quad (14)$$

算法 1:EM-PSO 求解算法。

输入:训练集图片 I,模型迭代次数 iter,种群数 M。

输出:卷积核 K(特征向量)。

(1)初始化种群

(2)初始化全局最优值 E_{gbest},全局最优向量 K_{gbest}

(3)计算图片 I 的方差 E_0

(4)For $t=0$ to iter do:

(5)　For $i=0$ to M do:

(6)　按照公式(14)计算梯度 Gra_i^t

(7) 按照公式(13)更新 K_i^t

(8)　$I^t(x,y)=\sum_{s,t=-\alpha}^{\alpha}k_{s,t}^t * I_0[x+s,y+t]$

(9)　$\overline{I^t}=\frac{1}{N}\sum_{x,y}I^t(x,y)$

(10)　end For

(11)更新全局最优值 E_{best},全局最优向量 K_{gbest}

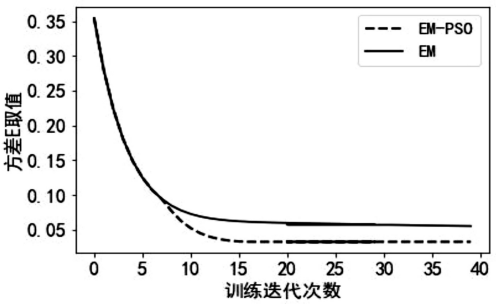
(12)　if $E_0-\Delta<E_{\mathrm{gbest}}$ and $E_{\mathrm{gbest}}<E_0+\Delta$:

(13)　break

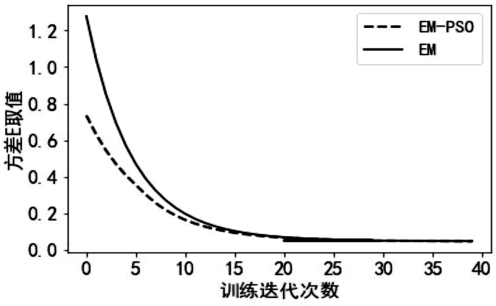
(14)　end For

2.3　EM-PSO 算法实验结果分析

随机选取一张数据集图片进行实验,比较直接利用梯度下降(EM)求解和混合梯度下降和粒子群算法(EM-PSO)求解时,方差 E 随迭代次数的变化情况,如图 3 所示。



(a)



(b)

图 3　方差 E 随训练次数变化趋势

从图 3(a)中曲线可以看出,直接利用梯度下降求解时,算法初始化值对 EM 算法迭代过程会造成较大影响,且易出现提前收敛现象,使方差无法收敛到约束范围内。EM-PSO 算法会一定程度上优于 EM 算法,且收敛速度较快,但 EM-PSO 算法仍存在一些缺点,比如每一次迭代花费的时间久,且存在随机生成的初始化值使得 EM-PSO 和 EM 算法的收敛效果相似,如图 3(b)所示。

3　实验结果与分析

3.1　实验数据

为了验证提出检测模型的有效性,从 FFHQ 高清数据集中选取 2 000 张图片进行实验,并选取 StyleGAN、StyleGAN2 生成的虚假人脸图片各 2 000 张构成模型检测数据集。其中 80% 作为训练集,20% 作为测试集,图片大小均为 1 024×1 024×3。

3.2　评判标准

针对分类模型,准确率 Accuracy 通常用来评估一个模型的全局分类正确情况,准确率越高则模型的分类正确率越高。准确率的计算公式如下:

$$\mathrm{Accurary}=\frac{N_{\mathrm{TP}}+N_{\mathrm{TN}}}{N_{\mathrm{TP}}+N_{\mathrm{TN}}+N_{\mathrm{FP}}+N_{\mathrm{FN}}} \tag{15}$$

式中,N_{TP} 为真正例(true positive,TP),表示被正确分类的 GAN 图片的数量;N_{TN} 为真负例(true negative,TN),表示被正确分类的真实图片的数量;N_{FP} 为假正例(false positive,FP),表示被错误分类的 GAN 图片的数量;N_{FN} 为假负例(false negative,FN),表示被错误分类的真实图片的数量。

为了更加全面地评估模型,还采用 ROC 曲线和 AUC 值作为评价指标,ROC 曲线描述在分类混淆矩阵中 FPR 与 TPR 两个量之间的相对变化情况。横轴为假正例率(false positive rate,FPR),纵轴是真正例率(true positive rate,TPR),即可以通过 ROC 曲线的变化情况考察 GAN 生成的虚假图片是否可以被正确分类。由于存在样本不均衡实验,ROC 曲线这种模型评价方法较准确率更能评价模型的分类性能。AUC 被定义为 ROC 曲线与坐标轴围成的面积,可以直观地评价分类器的好坏,值越大越好。

3.3　实验过程及结果分析

在方差求解部分,损失函数为方差,实验将模型的迭代次数 iter 设置为 40 次,约束条件 Δ 的取值范围为原始图片像素点方差的 10%。在对比损失部分,在 RGB 每个通道上利用 3×3 的卷积核求解特征向量,设置变化阈值,当 $L_t<0.001$ 时,停止迭代。在获得图片的特征向量后,以此作为输入,利用机器学习分类算法完成检测实验,通过对比实验最终选取支持向量机 SVM 作为模型分类器,其中核函数选取 RBF 函数。

实验 1:在不均衡样本上进行检测。

当迭代次数 $C=0$ 时,即当方差满足约束条件就停止迭代获得特征向量。从表 1 可以看出,检测准确率为 67.08%,AUC 值为 0.47,模型检测效果较差。从表 1 和图 4 可以看出,随着迭代效果次数的增加,模型的拟合效果越来越好,在 $C=5$ 时,模型检测准确率为 94.25%,AUC 值为 0.99。

为了研究迭代次数对检测准确率的影响,增加迭代次数至 10 次。实验结果表明,模型检测准确率随迭代次数的增加而提高,当 $C\geqslant 8$ 时,模型检测准确率变化趋于平稳,最高可达 99%。

表 1 不均衡样本模型检测准确率

迭代次数 C / 指标	0	1	2	3	4	5
准确率/%	67.08	69.42	78.67	94.25	93.08	94.25
AUC 值	0.47	0.74	0.89	0.94	0.98	0.99

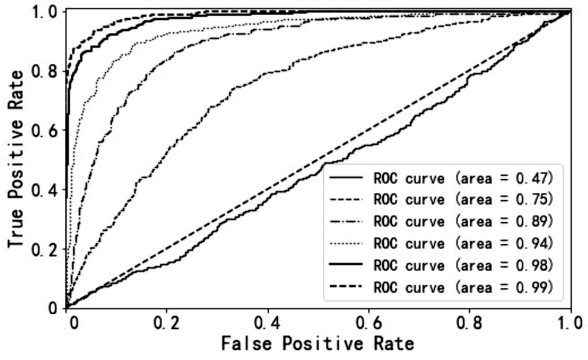


图 4 不平衡样本模型检测 ROC 曲线和 AUC 值

实验 2:在均衡样本上进行检测。

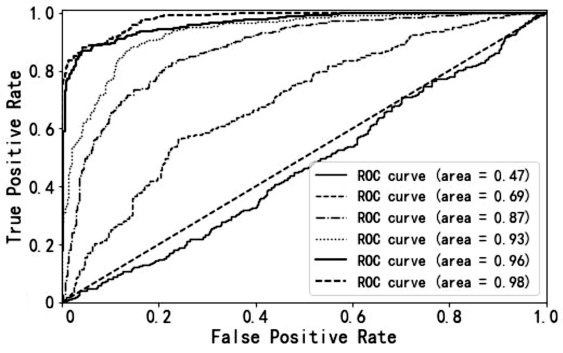
分别将 StyleGAN 和 StyleGAN 2 生成的图片与真实图片 FFHQ 数据集进行分类实验。

StyleGAN VS FFHQ:从表 2 可以看出,当迭代次数为 5 时,利用 SVM 进行分类,准确率为 90.62%,对应的 AUC 值为 0.98。

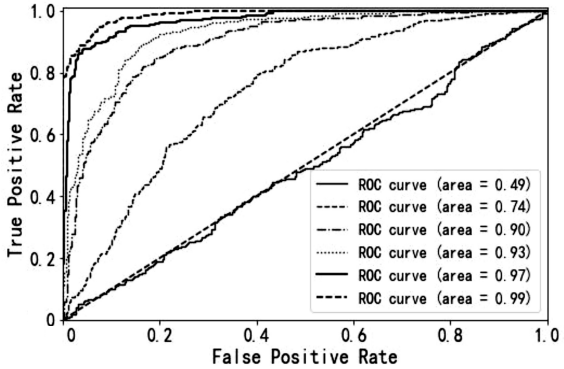
表 2 均衡样本模型检测准确率

迭代次数 C	StyleGAN 准确率/%	StyleGAN AUC 值	StyleGAN2 准确率/%	StyleGAN2 AUC 值
0	48.00	0.47	48.00	0.49
1	64.00	0.69	68.75	0.74
2	80.25	0.87	82.75	0.90
3	86.12	0.93	85.50	0.93
4	90.87	0.96	90.88	0.97
5	90.62	0.98	92.25	0.99

StyleGAN2 VS FFHQ:从表 2 可以看出,当迭代次数为 5 时,利用 SVM 进行分类,准确率为 92.25%,对应的 AUC 值为 0.99。迭代过程中除迭代次数为 3 时,检测效果均优于 StyleGAN 与 FFHQ,且从图 5 的 ROC 曲线对比情况可以看出,在区分 GAN 虚假图片的过程中,模型能更准确检测 StyleGAN2 生成的虚假图片。



(a) StyleGAN VS FFHQ



(b) StyleGAN2 VS FFHQ

图 5 均衡样本模型检测 ROC 曲线和 AUC 值

为了研究卷积核大小对检测结果的影响,对 5×5 的卷积核的分类效果进行实验,实验结果如表 3 所示。当迭代次数为 3 时,模型检测效果最佳,准确率最高为 90.50%,但检测准确率均低于卷积核为 3×3 时的检测结果。

表 3 卷积核为 5×5 的模型检测准确率

迭代次数 C	不均衡实验	StyleGAN VS FFHQ	StyleGAN2 VS FFHQ
0	67.08	47.50	50.38
1	69.08	61.50	67.13
2	83.50	79.75	81.00
3	90.50	84.13	88.75
4	86.17	79.13	83.13
5	89.08	86.00	86.75

为了验证模型的有效性,以 VGG16 模型展开对比实验,在不均衡实验中,VGG16 模型检测准确率为 67%,在均衡实验中,最佳检测结果为 50.28%,文中模型明显优于 VGG16 模型。

4 结束语

针对目前的检测模型大多侧重于针对某一种生成对抗模型生成的虚假图片,从 GAN 生成过程出发,基于 GAN 生成图片像素的局部相关性和对比损失原理构建了适用于多种 GAN 生成人脸虚假图片混合数据集的检测模型,研究了卷积核在分类过程中的作用,并在求解过程中结合梯度下降和粒子群位置更新方法改进 EM 算法迭代求解过程,利用 PSO 算法的全局思想克服梯度下降造成的早熟收敛问题,改善了收敛效果。

对 StyleGAN 和 StyleGAN2 生成的人脸图片进行了分类检测。实验结果表明,3×3 的卷积核分类准确率优于 5×5 的卷积核分类准确率,其中 3×3 的卷积核在迭代次数达到 5 次时,在不均衡样本实验中模型检测准确率可达 94.25%,AUC 取值可达 0.99,且不均衡样本实验效果略优于均衡样本实验。

生成对抗技术不止有 StyleGAN 和 StyleGAN2,还有诸多其他典型的生成对抗技术,因此下一步的研究重点是在其他生成对抗技术生成的虚假图片数据集上验证本模型是否具有通用性,并改进检测模型,使模型具有更广泛的应用。

参考文献:

[1] CHOI H, CHOI E. Discrimination of facial image generated via GAN[C]//2018 international conference on software security and assurance. Seoul, Korea: IEEE, 2018: 77-80.

[2] LI H, CHEN H, LI B, et al. Can forensic detectors identify GAN generated images[C]//2018 Asia-Pacific signal and information processing association annual summit and conference. Honolulu, HI, USA: IEEE, 2018: 722-727.

[3] RÖSSLER A, COZZOLINO D, VERDOLIVA L, et al. FaceForensics++: learning to detect manipulated facial images[C]// 2019 IEEE/CVF international conference on computer vision. Seoul, Korea: IEEE, 2019: 1-11.

[4] DANG L, HASSAN S, IM S, et al. Deep learning based computer generated face identification using convolutional neural network[J]. Applied Sciences, 2018, 8(12): 2610.

[5] FU Y, SUN T, JIANG X, et al. Robust GAN-face detection based on dual-channel CNN network[C]//2019 12th international congress on image and signal processing, biomedical engineering and informatics. Suzhou, China: IEEE, 2019: 1-5.

[6] TARIQ S, LEE S, KIM H, et al. GAN is a friend or foe? a framework to detect various fake face images[C]//Proceedings of the 34th ACM/SIGAPP symposium on applied computing. New York, USA: ACM, 2019: 1296-1303.

[7] BONETTINI N, BESTAGINI P, MILANI S, et al. On the use of Benford's law to detect GAN-generated images[EB/OL]. (2020)[2020-12-20]. https://arxiv.org/abs/2004.07682.

[8] GUARNERA L, GIUDICE O, BATTIATO S. DeepFake detection by analyzing convolutional traces[C]//2020 IEEE/CVF conference on computer vision and pattern recognition workshops. Seattle, WA, USA: IEEE, 2020: 2841-2850.

[9] MCCLOSKEY S, ALBRIGHT M. Detecting GAN-generated imagery using saturation cues[C]//2019 IEEE international conference on image processing. Taipei: IEEE, 2019: 4584-4588.

[10] 暴雨轩,芦天亮,杜彦辉. 深度伪造视频检测技术综述[J]. 计算机科学,2020,47(9):283-292.

[11] 李旭嵘,纪守领,吴春明,等. 深度伪造与检测技术综述[J]. 软件学报,2021,32(2):496-518.

[12] 李旭嵘,于　鲲. 一种基于双流网络的 Deepfakes 检测技术[J]. 信息安全学报,2020,5(2):84-91.

[13] 赵　磊,葛万峰,毛钰竹,等. 基于时空特征一致性的 Deepfake 视频检测模型[J]. 工程科学与技术,2020,52(4):243-250.

[14] 陈　鹏,梁　涛,刘　锦,等. 融合全局时序和局部空间特征的伪造人脸视频检测方法[J]. 信息安全学报,2020,5(2):73-83.

[15] 胡永健,高逸飞,刘琲贝,等. 基于图像分割网络的深度假脸视频篡改检测[J]. 电子与信息学报,2021,43(1):162-170.

[16] 张怡暄,李　根,曹　纭,等. 基于帧间差异的人脸篡改视频检测方法[J]. 信息安全学报,2020,5(2):49-72.

[17] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014, 27: 2672-2680.