基于 Star-Gan 的人脸互换算法

易　旭,白　天

(中国科学技术大学 软件学院,安徽 合肥 230026)

摘　要:基于深度学习的人脸互换算法会因背景人脸环境的明亮程度、人脸表情、姿势等因素影响换脸效果,现阶段的人脸互换模型存在固有的弊端。采用 Patch-Gan(Generative Adversarial Networks)的判别器结构能通过全卷积网络增强人脸局部一致性的效果。生成器将 U-net 结构的编码器的特征输出作为输入,能考虑多层信息细节。整体模型架构采用 Star-Gan 的模型,引入实例归一化层能保证图像的独立性。最后在 Face-Forensics ++ 人脸互换数据集上进行验证,结果表明,融合的模型有较好的生成效果和细节。

关键词:深度学习;人脸互换;对抗生成网络

中图分类号:TP183　　**文献标识码**:A　　DOI: 10.19358/j.issn.2096-5133.2020.05.003

引用格式:易旭,白天. 基于 Star-Gan 的人脸互换算法[J]. 信息技术与网络安全,2020,39(5):12-16.

Face swap algorithm based on Star-Gan

Yi Xu, Bai Tian

(School of Software Engineering, University of Science and Technology of China, Hefei 230026, China)

Abstract: The effect of face swap algorithm based on deep learning will be affected by the brightness of the background face environment, facial expression, posture and other factors. There are inherent disadvantages in the current face swap models. The discriminator structure based on Patch-Gan can enhance the local consistency of human face through full convolution network. The generator takes the feature output of the U-net encoder as the input, and considers the multi-layer information details. Among them, Star-Gan model is adopted as the overall model architecture, and case normalization layer is introduced to ensure the image independence. Finally, it is validated on Face-Forensics ++ face exchange data set. The results show that the optimized model has better generation effect and details.

Key words: deep learning; face swap; generative adversarial networks

0　引言

随着深度学习技术的兴起,图像处理相关的研究有了一项强有力的技术支持。人脸互换在图像处理方面作为一个里程碑式的技术,意味着计算机能够理解人脸图像。如何通过对抗生成网络实现人脸互换,提升生成效果是现如今计算机视觉的一大热点。

对于传统的方法 Face-swap[1],人脸互换只是把目标人脸截取,粘贴到原始人脸上面,使用图像融合的相关算法(如泊松融合)消除边界,后续的改进一般是在图像融合方面进行突破。

近年来,随着深度神经网络技术的成熟,KORSHUNOVA I[2]提出基于深度学习的人脸互换,将两个人脸的身份信息看成是两个不同图片风格,为一个目标人物训练一个深度神经网络提取人脸特征,换脸其实就是替换人脸的高维隐空间向量,而后再用训练好的人脸生成器进行生成,这种方式要求同一身份大量的人脸数据,其训练得到的模型只适用于这两个身份。YUVAL N[3]提出先使用 3DMM 模型拟合人脸,再互换人脸,解决了需要大量同一身份人脸图片的问题,但 3DMM 仍然有人脸匹配失败的问题,最终导致模型出错。NATSUME R[4-5]提出了 FSnet 和 RSGAN,使用编码器学习整体人脸的编码,对所有的人脸只学习一个单一的人脸身份编码器网络,但由于输出的编码是一个高维的人脸身份向量,特征信息依然高度纠缠。

本文借鉴前人的思想,使用 Star-Gan 模型作为生成器,利用 Arcface[6]身份编码器提取人脸高维身份特征,针对人脸细节的生成,使用基于 U-net[7]的人脸特征编码器模型为多层级的输入,解决人脸特征纠缠的问题,使用 Patch-Gan 的思想改造判别器

网络结构,引入实例归一化层提升生成效果。

1 Star-Gan 模型[8]

Star-Gan 模型的目的是解决人脸在多个域之间的转换问题,通过使用循环损失保证生成图像和背景图像的一致性,判别器保证生成图像的真实性,域分类器保证转换的有效性。调节多个损失,在生成效果上当时达到了较优的水平。

Star-Gan 可以作为一个不错的条件生成模型框架用于其他的生成任务。即将它的条件输入替换成身份图片的身份编码。该模型包括 3 个下采样模块、6 个残差模块和 3 个上采样模块。网络的输入是背景人脸图片,在残差层添加身份人脸的高维身份信息。最后端输出换脸图片。图 1 是整体网络架构图,详细的损失细节见 2.4 节。

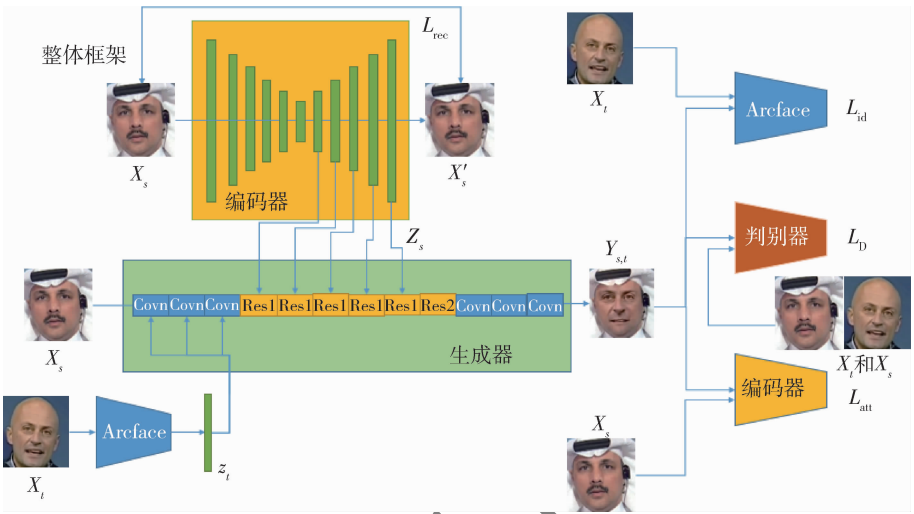

图 1 模型整体框架图

1.1 数据预处理

为了增强本模型的鲁棒性及泛化能力,在保证数据标签不变的情况下增大数据集。本文使用图像翻转、添加高斯噪声的方法扩充数据,提高模型鲁棒性。

(1)图像翻转

使用图像水平翻转扩充数据,可直接将数据集扩充一倍。如图 2 所示,左边的九张图为数据集中的原始图片,右边为水平翻转图片。

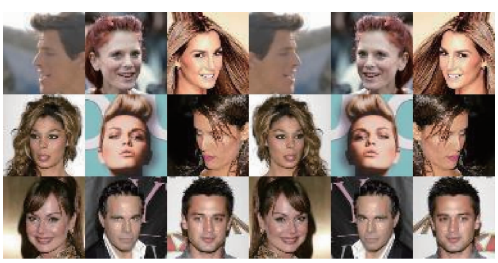

图 2 图像水平翻转效果

(2)添加高斯噪声

对图像训练集进行图像预处理,首先将图像像素归一化为 0 ~ 1 数值,如式(1)所示,I_{xy} 代表归一化的数值,I^R_{xy}、I^G_{xy}、I^B_{xy} 分别代表图片的三个通道。然后叠加范围为 0 ~ 0.05 的高斯噪声,并在 0,1 之间截断,如式(2)所示。最后得到的图像的像素点所对应的值的范围为[0,1]。

$$I_{xy} = [I^R_{xy}, I^G_{xy}, I^B_{xy}], I_{xy} \in [0,1] \quad (1)$$

$$I_{xy} = \min(0, \max(1, I_{xy} + 0.05N(0,1))) \quad (2)$$

1.2 实例归一化(Instance Normalization)

传统的 GAN 使用批归一化(Batch Normalization)的处理方式,虽然能提高收敛速度,但是却会降低最终的生成效果,因为它注重的是一个 batch 的数据,而更少地考虑单一图像本身的一致性。Star-Gan 使用实例归一化,将输入的图像四个维度记为[N, C, H, W],N 代表 Batch_size,C 代表通道数,H 和 W 分别是图像的长和宽。实例归一化是在 H 和 W 维度上进行归一化,这样的好处是它更加关注图像本身的一致性,具体处理如式(3) ~ 式(5)所示,三个公式中,t 表示 batch 的数目,i 表示通道数,j 和 k 表示图像的像素位置,l 和 m 表示图像的长和宽,式(3)中 ε 是一个极小的常数,防止出现除以 0 的计算。通过式(3)处理原始图像 x,计算得出归一化后的图像 y;式(4)计算得到原始图像 x 本身像素

的均值;式(5)计算得到原始图像 x 本身像素的方差。

$$y_{tijk} = \frac{x_{tijk} - \mu_{ti}}{\sqrt{\sigma_{ti}^2 + \varepsilon}} \tag{3}$$

$$\mu_{ti} = \frac{1}{HW}\sum_{l=1}^{W}\sum_{m=1}^{H} x_{tilm} \tag{4}$$

$$\sigma_{ti}^2 = \frac{1}{HW}\sum_{l=1}^{W}\sum_{m=1}^{H} (x_{tilm} - \mu_{ti})^2 \tag{5}$$

1.3 残差层

Resnet18[9]通过在 block 的输出增加残差层,使得网络有了逼近恒等映射的能力,解决了网络传播过程中信息丢失的问题,由于残差层的存在,避免了深度乘法对梯度的影响。Star-Gan 通过引入残差层训练提升了深度学习训练效果,同时解决了由于网络层数加深,出现的梯度消失现象。具体残差块如图 3 所示。

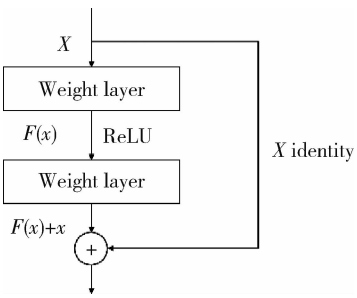

图 3 Resnet18 残差层

图 3 中,X 是上一层的输入,通过两层的卷积层学习到新的特征输出 $F(x)$,最后的输出是 $F(x)$ 和输入 X 的叠加,梯度可以从两条支路进行传播,避免过深的网络导致的梯度消失。

2 基于 Star-Gan 的人脸互换模型

2.1 U-net 结构编码器

通过一个类 U-net 的编码解码结构提取人脸身份特征,如图 4 所示,使用特征重构损失约束编码器学习图片各个维度的特征。$\boldsymbol{X}_{特征}$ 是输入图片,$\boldsymbol{X}_{重构}$ 是输出图片,中间层的特征输出作为生成器的图片高维特征输入。

2.2 Patch-Gan 结构

Patch-Gan 与传统的 GAN 十分相似,不同的是它的判别器是一个全卷积网络,其输出不是一个数字,而是一个二维矩阵,这样的好处是它能够保证图片的局部一致性。判别器具体结构如图 5 所示。

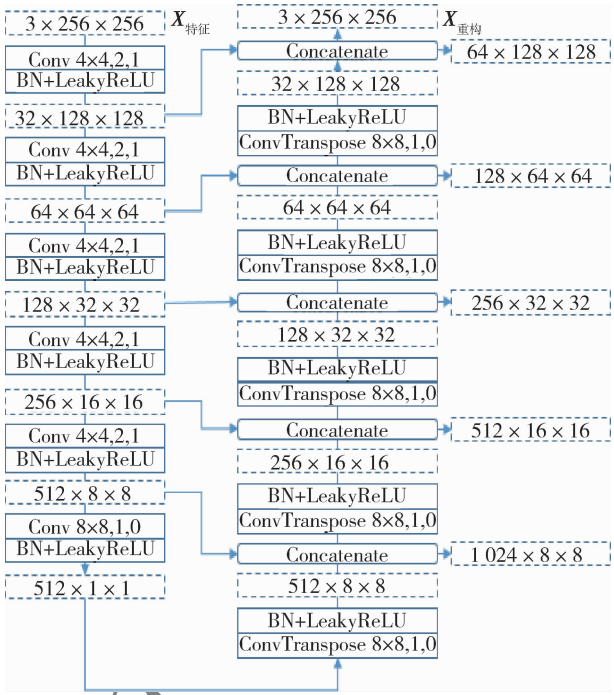

图 4 编码器详细结构

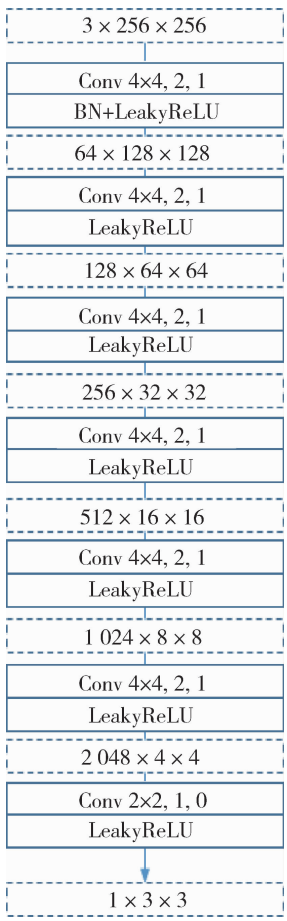

图 5 判别器详细结构

2.3 Arcface 人脸身份判别器

本文使用 Arcface 最后一层的特征作为人脸身份信息的特征表示,并将它作为生成器前三层的身份特征输入。

2.4 整体损失函数和训练细节

固定 Arcface 网络,网络训练时不更新参数,特征编码器损失使用 MSE 损失训练确保每一层网络能够提取出不同维度的特征。解码后能够还原初始图片。

$$L_{\text{rec}} = \| \boldsymbol{X}_s - \boldsymbol{X}'_s \|^2 \tag{6}$$

其中,$\boldsymbol{X}_s$ 表示换脸图片,$\boldsymbol{X}'_s$ 表示换脸重构图片。

为了保证生成器具有生成能力,使用了三个损失函数:对抗损失 L_{D}、身份损失 L_{id} 和特征损失 L_{att},保证生成图片 $\boldsymbol{Y}_{s,t}$ 的真实度。具体公式如式(7)~式(9)所示。

式(7)中 $\boldsymbol{X}$ 是从真实图片中的采样,来自于 $\boldsymbol{X}_s$ 和 $\boldsymbol{X}_t$,$\boldsymbol{Z}_s$ 和 $\boldsymbol{Z}_t$ 是图片分别经过 Arcface 和特征编码器提取的特征向量。

$$L_{\text{D}} = -E_{x\sim p_r}[\log D(\boldsymbol{X})] - E_{z\sim p_z}[\log(1 - D(G(\boldsymbol{Z}_s, \boldsymbol{Z}_t, \boldsymbol{X}_t)))] \tag{7}$$

式(8)中 $\boldsymbol{Y}_{s,t}$ 表示换脸后的图片,z_{id} 表示将人脸图片输入 Arcface 得到的身份特征,用两张图片的余弦距离作为身份相似度。

$$L_{\text{id}} = 1 - \cos(z_{\text{id}}(\boldsymbol{X}_s), z_{\text{id}}(\boldsymbol{Y}_{s,t})) \tag{8}$$

提取多层级的人的身份特征后,使用 L_{att} 损失函数评估身份特征的保持。因为使用了 5 层的身份特征,此时式(9)中 $n=5$。

$$L_{\text{att}} = \frac{1}{2}\sum_{k=1}^{n} \| z_{\text{att}}(\boldsymbol{Y}_{s,t}) - z_{\text{att}}(\boldsymbol{X}_t) \|_2^2 \tag{9}$$

为了达到人脸互换的目的,需要最小化上述所有三个损失,最终损失 L 如式(10)所示,其中 λ_{id} 和 λ_{att} 是训练时的两个超参数,调节三个损失的比例,这里设定 $\lambda_{\text{id}} = 10$ 和 $\lambda_{\text{att}} = 1$。

$$L = L_{\text{D}} + \lambda_{\text{id}} L_{\text{id}} + \lambda_{\text{att}} L_{\text{att}} \tag{10}$$

3 实验结果及分析

实验环境为:Linux16.04 的 64 位操作系统;显卡型号为 RTX1080ti,并行化处理,每张显卡拥有 12 GB显存;使用 Pytorch1.0.1 环境;Python3.6 包括 python-open-cv 库以及 Scipy。

因人脸互换的输出判别较为主观,本文将输出图像与 FF++ 的人脸互换模型中的 Faceswap 和 Deepfake 模型进行比较。

3.1 人脸互换泛化能力测试

初始训练时使用最新的 FFHQ 人脸数据集进行训练,Arcface 模型使用原 Arcface 论文提供的预训练模型。为了加速模型训练,将 FFHQ 的图片缩放到 3×256×256 进行训练,图 6 是最终输出结果,最上面一行为身份人脸 $\boldsymbol{X}_s$,最左边是特征人脸,可以看出本文模型在很好地保持背景人脸的背景风格的同时可以很好地保持人脸的身份特征,且较少出现图像失真的情况。即使在双方差异很大的情况下,模型仍然能够保持良好的生成情况。第二列身份图片和背景脸部朝向有很大的差距;第三列、第五列和第六列改变了性别,但生成器仍然能够得到他们的合成图片;第四列和第五列分别测试了年龄上的差异性,结果显示模型在两张图片差异较大的情况下仍然能够处理相应的人脸图片,具有较好的泛化能力。

图 6 模型基于 FFHQ 数据集训练人脸互换结果

3.2 人脸互换效果比较

为了衡量本文模型的生成效果,将模型生成的图片和 Deepfake 以及 Faceswap 的生成图片进行比较,此时不使用人脸数据集 FFHQ 进行训练,仅使用 FaceForensics ++[10] 数据集中视频人脸数据进行训练和替换。训练时对视频帧进行采样,每 1 秒采样 10 帧,为了保持背景尽量一致,保持 Arcface 固定不变,训练时提升 $L1$ 损失的权重 $\lambda_{\text{att}} = 20$,训练效果如图 7 所示。

图 7 模型与 Faceswap、Deepfake 视觉效果对比

由图 7 可以发现 Faceswap 方法的人脸会出现强烈的失真和脸部变形,其主要是由于 Faceswap 的方法着重于将身份人脸直接替换到背景人脸,导致前后帧不一致从而形成失真。而 Deepfake 着重于人脸框的替换,这样的好处是其不需要关注背景的信息,但显而易见地,它趋向于模糊化人脸,目标人

脸无法很好地融合进背景人脸中。

3.3 使用 Face-net[11] 比较人脸身份的保持

Face-net 是一个人脸识别的框架,它可以通过计算人脸图片的高维特征的余弦距离衡量两张人脸之间的身份相似性。表 1 是将图 7 的五种图片分别与身份人脸比较,输入 Face-net 网络后计算得到的欧氏距离。距离越小表示身份越相似。经过大量数据的测试,可以认为两张人脸是同一个人的阈值约为 1.1,即余弦距离低于 1.1,则可以认为两张图片是同一个人,当图片完全一致时,距离为 0。

表 1 模型与 Faceswap,Deepfake 身份保持对比

	特征人脸	身份人脸	Faceswap	Deepfake	本文模型
身份人脸	1.360 0	0.000 0	1.133 0	0.851 2	0.830 8
特征人脸	0.000 0	1.360 0	0.950 5	1.363 2	1.374 4

由于人脸互换模型的重点是保持身份特征,所以只需要关注表 1 第一行身份人脸和其他所有人脸的相似度。本文模型最大限度地保持了人脸身份,Deepfake 也比较出色地完成了人脸互换,但是由于它只考虑到脸部中心区域,因此最后的效果比本文模型略差。而 Faceswap 只用了简单的扣取加替换方式,导致生成的人脸的一致性较低,与身份人脸的相似性较低。第二行显示出本文模型转换的图片与特征人脸的身份差距较大,高于其他两个人脸互换模型,证明本文模型更少地依赖特征人脸的身份信息。

4 结论

本文基于 Star-Gan 构造了一个人脸互换模型,针对人脸身份独立设计了一个编码器用于提取身份,使用多层级的思想改良了生成器 Star-Gan,实现了一个人脸互换的模型。在 FaceForensics ++ 数据集上的实验效果表明,该模型在生成效果和人脸身份保持上优于现有的人脸互换模型。但该模型仍有缺陷,其虽然有良好的泛化能力,但针对初始身份人脸数量较少时,模型效果仍然有待提升。

参考文献

[1] BITOUK D, KUMAR N, DHILLON S, et al. Face swapping:automatically replacing faces in photographs[J]. Proceedings of ACM Siggraph,2008,27(3):1-8.

[2] KORSHUNOVA I, SHI W, DAMBRE J, et al. Fast face-swap using convolutional neural networks[C]. The IEEE International Conference on Computer Vision (ICCV), 2017:3677-3685.

[3] YUVAL N,IACOPO M,ANH T T,et al. On face segmentation,face swapping, and face perception[C]. 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition,2018:98-105.

[4] NATSUME R, YATAGAWA T, MORISHIMA S. RSGAN: face swapping and editing using face and hair representation in latent spaces[J]. arXiv:1804.03447. 2018.

[5] NATSUME R, YATAGAWA T, MORISHIMA S. FSNet: an identity-aware generative model for image-based face swapping[C]. Asian Conference on Computer Vision,2018:117-132.

[6] DENG J, GUO J, XUE N, et al. Arcface: additive angular margin loss for deep face recognition[C]. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2019:4690-4699.

[7] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]. Medical Image Computing and Computer-Assisted Intervention, MICCAI, 2015: 234-241.

[8] CHOI Y, CHOI M, KIM M, et al. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation[C]. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2018:8789-8797.

[9] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 4690-4699.

[10] RSSLER A, COZZOLINO D, VERDOLIVA L, et al. FaceForensics ++ :learning to detect manipulated facial images [C]. The IEEE International Conference on Computer Vision (ICCV),2019:1-11.

[11] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016:2818-2826.

(收稿日期:2020-03-30)

作者简介:

易旭(1995 -),男,硕士研究生,主要研究方向:对抗生成网络、深度学习。

版权声明