基于深度学习的人脸表情迁移方法

刘 剑 金泽群
(沈阳建筑大学信息与控制工程学院 沈阳 110168)

摘 要 针对人脸表情迁移生成图像质量不高、训练过程较长且生成速度较慢的问题,文中提出了一种基于生成式对抗网络的人脸表情迁移方法,使表情迁移更加快速和自然。首先,利用卷积神经网络进行人脸特征提取,并将图像从高维空间映射到浅层空间,在浅层空间中利用生成式对抗网络模型对人脸表情特征进行判别;然后,通过最近邻上采样层和卷积层组合结构将图像从浅层空间映射到高维空间,并在此过程中通过加入表情标签特征图对人脸表情进行改变。与 Fader Networks 相比,所提方法的网络模型参数量减少 43.7%,训练时间缩短了 36%。实验结果表明,所提方法有效地提高了人脸表情迁移生成图像的速度和质量。

关键词 人脸表情迁移,生成式对抗网络,计算机视觉,深度学习
中图法分类号 TP183 **文献标识码** A

Facial Expression Transfer Method Based on Deep Learning

LIU Jian JIN Ze-qun
(Faculty of Information and Control Engineering, Shenyang Jianzhu University, Shenyang 110168, China)

Abstract In order to solve the problems of low image quality, long training process and slow generation speed of face expression transfer, this paper proposed a facial expression transfermethod based on generative adversarial network to make expression transfer faster and more natural. Firstly, the facial features are extracted by using convolutional neural network, and the images are mapped from high-dimensional space to shallow space. In the shallow space, the facial expression features are discriminated by using the Generative Adversarial Networks. Then the nearest neighbors up-sampling and convolutional neural networks are used to mapthe image from the shallow space to the high-dimensional space, and in this process, the face expression is changed by adding the facial expression feature maps into neural networks. Compared with Fader Networks, the network model parameter amount of the proposed method is reduced by 43.7% and training time is reduced by 36%. The experimental results show that the proposed method can effectively improve the quality and the speed of generated images.

Keywords Face expression transfer, Generative adversarial networks, Computer vision, Deep learning

1 引言

近年来,人脸表情迁移技术在影视娱乐、虚拟现实、人机交互等众多领域得到了广泛的应用。如何对人脸表情进行简单、快速的迁移,已经成为了计算机视觉与图形学领域的热点研究课题,如何自然地改变一个人的表情具有很大的挑战性。

人脸表情迁移问题是近年来计算机视觉与图像处理领域的热点问题。人脸表情变化不仅体现在面部特征的运动形变上,还包括面部细节的改变。传统的方法对面部细节的处理能力较弱,导致生成图像的质量较差。近期提出的生成式对抗网络 GAN(Generate Adversarial Net)[1] 在生成图像方面具有显著的优势。生成式对抗网络建立从图像到图像(Image-to-Image)的非线性映射,并通过判别器对生成图像和真实图像进行判别,从而提高生成图像的质量[2]。目前实现人脸表情迁移的方法有很多种,主要分为传统算法和深度学习算法两种。传统算法主要是基于图像渐变的方法和基于特征偏移的方法[3-5],通过对多人面部表情的变形和组合来对表情进行改变,这些方法生成图像的质量较低,面部表情不协调。近年来,深度学习在计算机视觉与图形学领域有了极大的发展[6-8],goodfellow 等人提出的生成对抗网络极大地提高了生成图像的质量。对于人脸表情迁移课题,研究者们也做出了很多尝试,如 Guillaume Lample 等提出的 Fader Networks[9] 首次将生成对抗网络应用于人脸表情迁移问题,取得了不错的效果,但是由于存在训练不稳定、计算复杂度高并且生成图像质量较差等问题,并未得到广泛的应用。

本文提出一种由深度自编码器和生成式对抗网络融合成的网络模型进行人脸表情迁移的方法。采用深度自编码器(DAE)建立一个端到端的深度神经网络模型,并通过上采样层和卷积层组合的方式代替转置卷积层,从而有效消除棋盘格效应,提高生成图像的质量。在上采样过程中加入表情标签信息,可以对生成人脸的表情进行改变,从而达到生成高质量可迁移表情人脸图像的目的。

本文受国家自然科学基金(61272253),辽宁省自然科学基金(201602616),辽宁省教育厅科学研究项目(L2015443),住建部项目(2015-K2-015)资助。

刘 剑(1963—),女,博士,教授,主要研究方向为视觉图像、智能控制等,E-mail:jeanliu@163.com;**金泽群**(1994—),硕士生,主要研究方向为视觉图像。

2　本文算法

本文提出一种由深度自编码器和生成式对抗网络融合成的网络模型。深度自编码器能够建立表示向量从高维空间到低维空间再到高维空间的映射关系，实现了端到端的神经网络模型映射。生成式对抗网络通过生成和判别低维空间表示向量，判别低维空间表示向量所对应的标签信息，生成无标签信息的低维空间向量，最后通过深度自编码器中的标签特征图控制生成图像的表情，实现人脸表情的迁移。

2.1　深度自编码器网络

自编码器是表示学习的一种经典算法。自编码器由一个编码器(Encoder)函数和一个解码器(Decoder)函数组成，编码器函数将高维向量映射到低维空间，而解码器函数将低维向量映射到高维空间[10]。深度自编码器(DAE)是由深度神经网络组成的自编码器，编码器由一个包含多个隐含层的卷积神经网络组成，通过稀疏编码将输入高维空间向量映射到低维空间向量，对高维信息进行特征提取，并通过低维向量进行表示。解码器是由上采样层和卷积层组成的神经网络，作用是将低维空间中的特征向量映射到高维空间，在上采样的过程中加入限制条件可以对映射过程进行约束。深度自编码器在给定足够多的隐藏单元的情况下，能够以任意精度近似任何从输入到编码的映射。

本文提出了一种改进的深度自编码器。如图 1 所示，给定输入图像和标签，编码器将映射到浅层表征，解码器经过训练，将浅层表征重建为高维空间向量。编码器中去掉了广泛应用在神经网络中的批归一化层，并在解码器中加入标签特征图来影响生成图片的特征，通过大量数据的训练来获取人脸表情迁移编码器和解码器的权重系数，从而建立表情迁移前后的映射关系。

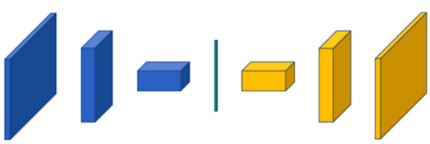


图 1　深度自编码器模型示意图

批归一化自 Sergey Ioffe 等人提出后，便被广泛应用在神经网络模型中，并取得了良好的效果。批归一化在神经网络训练的过程中，通过计算每个批次数据的均值和标准差，将输入数据进行归一化处理，使其均值为 0，标准差为 1。批归一化公式如式(1)所示：

$$\hat{x}^{(k)}=\frac{x^{(k)}-E[x^{(k)}]}{\sqrt{Var[x^{(k)}]+\varepsilon}} \tag{1}$$

其中，$x^{(k)}$ 表示输入数据的第 k 维，$E[x^{(k)}]$ 表示该维的平均值，$\sqrt{Var[x^{(k)}]}$ 表示标准差。经过以上公式对原始数据进行归一化处理后，原始数据分布会被强行破坏，批归一化同时设置两个可以进行学习的参数 γ 和 β，以减轻数据被破坏的程度，尽可能的还原原始数据的分布情况，如式(2)所示：

$$y^{(k)}=\gamma^{k}\hat{x}^{(k)}+\beta^{(k)} \tag{2}$$

虽然批归一化通过改变输入的分布来解决每个隐含层输入数据分布差距较大导致的训练不稳定等问题，能够使神经网络的损失函数值迅速收敛，但是在对输入数据进行归一化的同时，数据的分布被改变，原始数据的信息会有一定的损失，从而降低生成图像的质量，尤其在生成模型中，使用批归一化处理会影响生成图像的质量。在本文所提出的模型中，编码器去掉了批归一化层，通过实验确定超参数的设置，尽可能减少训练过程中对原始数据的改变，提高生成图像的质量；同时使得模型参数减少了 43.7%，并提高了生成图像的速度。

本文提出模型的解码器由上采样层、卷积层和标签特征图层组成，代替了传统生成模型中经常采用的转置卷积层(Transposed Convolution Layer)。转置卷积又称为反卷积，经常在生成模型等神经网络中实现从低维空间到高维空间的映射，但是不合理的卷积核大小和步长会导致生成像素相互重叠的现象，被称为棋盘效应(Checkerboard Artifacts)。有两种方法可以消除棋盘效应：1)严格的计算卷积核与步长之间的相互关系，用到的转置卷积核的大小可以被步长整除，从而避免重叠效应；2)将转置卷积操作分解为上采样和卷积两步，首先利用最近邻插值或者双线性插值的方法将特征图进行缩放，然后再进行卷积操作，对缩放后的图像进行调整。

本文采用最近邻插值和卷积相结合的方法，首先利用最近邻插值对图像进行上采样，然后通过卷积操作建立从低维空间到高维空间的非线性映射，最终通过加入标签特征图的方式进行人脸表情迁移。

如图 2 所示，在模型上采样的过程中加入标签特征图，标签特征图的大小与当前层特征图的大小相同，叠加在多通道的特征图上作为下一层隐含层的输入；在训练的过程中，通过给标签特征层赋予不同的值来更新模型参数，从而使模型具有人脸表情迁移的能力，最终可以通过标签特征图的值来对解码器进行约束，使生成的图像具有明显的表情特征。本文深度自编码器模型的参数如表 1 所列。

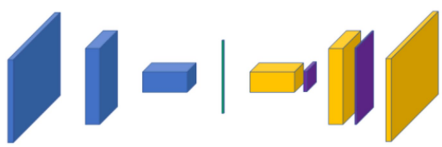


图 2　标签特征示意图

表 1　深度自编码器模型参数表

Operation	Kernel	Feature maps	Neuron number
256×256 inputs			256×256×3
Convolution	4×4	8	128×128×8
Convolution	4×4	16	64×64×16
Convolution	4×4	32	32×32×32
Convolution	4×4	64	16×16×64
Convolution	4×4	128	8×8×128
Convolution	4×4	256	4×4×256
Convolution	4×4	512	2×2×512
Up Scale		512+1	2×2×513
Convolution	4×4	512+1	8×8×513
Up Scale		256+1	8×8×257
Convolution	4×4	256+1	16×16×257
Up Scale		128+1	16×16×129
Convolution	4×4	128+1	32×32×129
Up Scale		64+1	32×32×65
Convolution	4×4	64+1	64×64×65
Up Scale		32+1	64×64×33
Convolution	4×4	32+1	128×128×33
Up Scale		16+1	128×128×17
Convolution	4×4	3	256×256×3

2.2　生成式对抗网络

生成式对抗网络(Generative Adversarial Networks，

GAN)是 Goodfellow 等在 2014 年提出的一种生成式模型。生成式对抗网络基于博弈论中二人零和博弈(即二人利益之和为零,一方所得是另一方的损失),由一个生成器和判别器组成:生成器捕捉真实数据样本的潜在分布,并且生成新的数据样本;判别器是一个分类网络,判别输入的数据是来自真实样本还是生成的样本。生成器和判别器可以采用各种深度神经网络模型进行尝试[11]。生成式对抗网络的优化过程是一个极小极大博弈(Minmax Game)问题,优化过程最终希望逼近纳什均衡,使生成器估计样本数据的分布,并生成新的数据样本[12]。

$$\min_G \max_D V(D,G)=E_{x\sim P_{data}(x)}[\log D(x)]+E_{z\sim P_z(z)}[\log(1-D(G(z)))] \tag{3}$$

如式(3)所示,用可微分函数 G 和 D 来分别表示生成器和判别器,它们的输入分别为真实数据 x 和由编码器产生的低维空间特征向量 z;$G(x)$ 为由 G 生成的低维空间特征向量。D 的目标是判别低维空间特征向量对应的标签信息,而 G 的目标是使自己生成的伪数据 $G(x)$ 在 D 上的表现 $D(G(x))$ 和真实数据在 D 上的表现一致,生成器和判别器相互对抗,最终判别器无法判断生成器输出的正确标签,则生成器训练完成[13]。

本文提出的模型如图 3 所示,将深度自编码器中的编码器作为生成对抗网络的生成器,将编码器生成的低维空间向量 z 作为判别器的输入。判别器是一个由 3 个全连接层组成的分类网络,用于判别低维空间向量对应的标签信息 y,生成器和判别器相互对抗,最终判别器无法判断由生成器生成的低维空间向量 z 的标签信息[14],不含标签信息的低维空间向量再经有新的标签信息约束的解码器最终生成可控制表情的高质量图像。

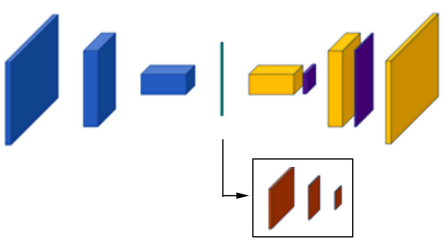


图 3 生成式对抗网络模型示意图

2.3 损失函数

本文所提模型的损失函数来自深度自编码器和生成式对抗网络两个部分,x 为输入原始图像,经过深度自编码器 E_{AE} 输出为与之相对的低维空间表示向量 z,低维向量 z 由解码器 D_{AE} 经过上采样和卷积等操作最终映射到高维空间,输出与原始图像大小相同经过表情迁移的图像。

深度自编码器采用均方误差函数(Mean Square Error)作为损失函数,以此来衡量生成图像与原图像的误差[15]。均方误差的计算如式(4)所示:

$$Loss_{ae}=\frac{1}{m}\sum \| D_{AE}(E_{AE}(x),y)-x \|_2^2 \tag{4}$$

均方误差函数广泛应用于深度学习网络模型中,作为衡量生成网络模型的损失函数,计算成本低,并且能够使深度神经网络稳定地收敛。本文采用均方误差函数作为总体损失函数的一部分,参与整个神经网络的训练。

生成式对抗网络对浅层空间向量 z 进行生成,自编码器 D_{AE} 同时作为生成式对抗网络的生成器,判别器 D_{GAN} 为分类网络,输出为浅层空间向量 z 的标签结果,所以本文采用 softmax 函数作为生成式对抗网络判别器的损失函数。softmax 函数如式(5)所示:

$$P(i)=\frac{\exp(\theta_i^T z)}{\sum_{k=1}^{K}\exp(\theta_k^T z)} \tag{5}$$

判别器 D_{GAN} 输出浅层空间特征向量 z 所属标签 y 的概率 $P_{dis}(y|E_{AE}(z))$。判别器的损失函数可以表示为:

$$Loss_{dis}=-\frac{1}{m}\sum \log P_{dis}(y|E_{AE}(x)) \tag{6}$$

模型的总体损失函数由深度自编码器和对抗式生成网络两部分组成,两部分损失函数在训练过程中对最终结果的影响不同,需要设置超参数对两部分损失函数的比例进行调节,如式(7)所示:

$$Loss_{Total}=Loss_{ae}+\mu Loss_{dis} \tag{7}$$

其中,μ 用来调节损失函数两部分的比例。损失函数的总体表达式如式(8)所示:

$$Loss_{Total}=\frac{1}{m}\sum \| D_{AE}(E_{AE}(x),y)-x \|_2^2-\mu\sum \log P_{dis}(y|E_{AE}(x)) \tag{8}$$

超参数 μ 用来调节生成式对抗网络在整体网络中的影响程度,μ 的值越大,表示生成式对抗网络对生成图像的影响也越大,但过大的系数会造成生成图像的质量不稳定,所以需要通过实验来确定超参数的取值。

3 实验与分析

深度神经网络的训练数据集来源于 CelebA 数据集[16]。CeleA 数据集由超过 200 000 张 178×218 的带标签图像组成,图像的标签包含 40 个类别。首先将大小为 178×218 的原始图像裁剪至 178×178 的图像,再经过缩放扩大为 256×256 大小的图像,用于深度神经网络的训练及测试。把整个数据集分为训练数据集和测试数据集,其中训练数据集包含 180000 张图片,测试数据集包含 20000 张图片。将图像的像素值归一化到[-1,1],并通过改变图像明暗度等进行数据增强[17]。

本文通过 Tensorflow 开源深度学习框架搭建神经网络模型进行训练,训练过程包括 3 个阶段。

1)首先只对深度自编码器进行训练,生成式对抗网络不参与训练,建立深度自编码器中从图像到图像(Image-to-Image)的非线性映射。

2)将超参数 μ 设置为一个较小的值,本文采用 0.0001,使生成式对抗网络与深度自编码器共同进行训练,超参数 μ 取值较小可以保证初始阶段整体训练的稳定性,生成式对抗网络对整体模型的影响过大会造成生成图像质量不稳定。

3)将超参数 μ 设置为一个较大的值,本文采用 0.01,以加大生成式对抗网络对整个模型的影响,最终得到人脸表情迁移的图像。

本文采用由 Kingma 等[18]提出的 Adam 优化器进行训练,batchsize 设置为 32,学习率设置为 0.01,并在每个阶段结束后进行指数衰减,衰减系数为 0.1,训练过程中每个阶段各进行 100000 次迭代,共进行 300000 次迭代,最终完成模型的训练。

与之前生成式深度神经网络相比，本文提出的人脸表情迁移模型用上采样层和卷积层组合取代反卷积层，并且取消了批归一化层，显著减少了模型参数量。如图 4 所示，相较于 Fader Networks，本文所提模型的参数量减少了 43.7%。

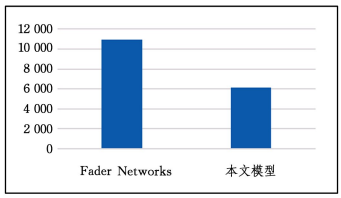


图 4　模型参数量的对比

较少的参数量可以加快模型的训练，并且使其更易于部署，最终模型的训练速度提升了 36%，如图 5 所示。

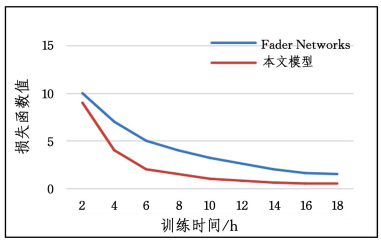
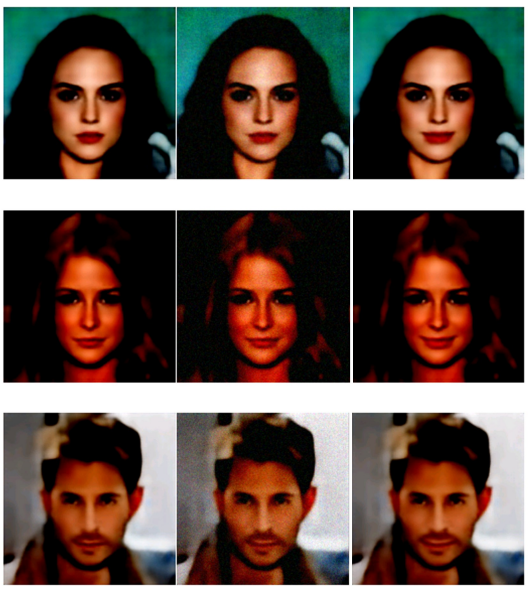


图 5　模型训练时间的对比

本文提出的模型在减少模型参数量、提高训练速度的同时，提升了生成图像的质量，有效消除了由转置卷积层带来的棋盘格效应，如图 6 所示。



(a)原始图像　　(b)Fader Networks　　(c)本文算法

图 6　模型生成结果的对比

图 6 中(a)为原始输入图像，(b)为原始图像送入 Fader Networks 经过人脸表情迁移所得到的图像，(c)为本文算法得到的实验结果。Fader Networks 训练得到的人脸表情迁移效果并不明显，只有微小的变化。与之相比，本文提出的模型人脸表情迁移效果更加明显。

结束语　针对目前人脸表情迁移生成图像质量较低、生成速度较慢等问题，本文提出了一种改进的基于生成式对抗网络的人脸表情迁移模型，通过去除批归一化层，用上采样层和卷积层结合代替转置卷积层。与 Fader Networks 网络相比，本文方法将网络模型参数量减少 43.7%，训练时间减少 36%；与此同时，生成图像的质量有所提高。本文所提方法在人脸表情迁移方向进行了新的尝试，并且在影视娱乐、虚拟现实、人机交互等领域具有较强的实际意义。

参 考 文 献

[1] IOFFE S, SZEGEDY C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[J]. arXiv: 1502.03167v3, 2015.

[2] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[J]. arXiv: 1611.07004, 2016.

[3] 王娅，侯进，王献. 基于顶点权重的网格简化在虚拟人脸中的应用[J]. 计算机仿真，2014，31(2)：329-334.

[4] 雷腾. 虚拟人眼的运动与表情合成的研究[D]. 成都：西南交通大学，2014.

[5] 李俊龙，章登义，黄珺. Kinect 驱动的人脸动画合成技术研究[J]. 计算机工程，2015，41(3)：237-241.

[6] ZHANG H, XU T, LI H, et al. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks[J]. arXiv: 1612.03242, 2016.

[7] ZHANG Z, SONG Y, QI H. Age progression/regression by conditional adversarial autoencoder[C]// The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

[8] ZHAO J, MATHIEU M, LECUN Y. Energy-based generative adversarial network[C]// 5th International Conference on Learning Representations (ICLR). 2017.

[9] LAMPLE G, ZEGHIDOUR N. Fader Networks: Manipulating Images by Sliding Attributes[J]. arXiv: 1706.00409v2, 2017.

[10] HINTON G, KRIZHEVSKY A, WANG S. Transforming auto-encoders[C]// Artificial Neural Networks and Machine Learning (ICANN 2011). 2011: 44-51.

[11] PERARNAU G, VAN DE WEIJER J, RADUCANU B, et al. Invertible conditional gans for image editing[J]. arXiv: 1611.06355, 2016.

[12] RATLIFF L J, BURDEN S A, SASTRY S S. Characterization and computation of local Nash equilibria in continuous games[C]// Proceedings of the 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton). Monticello, IL, USA, IEEE, 2013: 917-924.

[13] ARJOVSKY M, OTTOU L. Towards principled methods for training generative adversarial networks[C]// ICLR. 2017.

[14] SHEN W, LIU R. Learning residual images for face attribute manipulation[C]// The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[15] ANTIPOV G, BACCOUCHE M, DUGELAY J L. Face aging with conditional generative adversarial networks[J]. arXiv: 1702.01983, 2017.

[16] LIU Z W, LUO P, WANG X G, et al. Deep learning face attributes in the wild[C]// Proceedings of International Conference on Computer Vision (ICCV). 2015.

[17] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[C]// AISTATS. 2010.

[18] KINGMA D, BA J. Adam: A method for stochastic optimization[J]. arXiv: 1412.6980, 2014.