介绍

我们的模型由四个子模型组成，一个编码器E负责将输入图片编码成代表内容的特征c和代表风格的特征d；一个解码器G负责根据输入的特征d和c将图像还原；一个判别器D与G就进行对抗训练，用于指导G的图像生成，使得G生成的图像尽可能真实；另外还有一个分类器C与生成器G进行协同训练，使得生成的图像满足正确的特定分类。

假设有两组不同的数据x1和x2，其标签为y1和y2，在这里我们将x1和x2分别使用编码器编码，分别提取出x1和x2的代表内容的特征c1,c2和代表风格的特征d1,d2，即：

c1,d1= E(x1)

c2,d2= E(x2)

然后我们使用解码器G将两组图像重构，即使得解码器G可以根据特征c和d还原出原有的图像内容：

x1’=G(c1,d1)

x2’=G(c2,d2)

另外我们将特征d1和d2进行交换，即c1和d2进行新的组合，以及c2和d1进行新的组合，并根据这两组组合将图像再次使用G解码重构，假设生成的图像为g1，g2即：

g1=G(c1,d2)

g2=G(c2,d1)

生成的图像g1具有标签y2，生成的图像g2具有标签y1。

为了保证生成的图像g1和g2，确实充分反映了c1，c2和d1，d2中的信息，我们对g1和g2做了更多约束，即g1和g2进行再次编码，并将分解出的特征d1和d2再次交换，新的组合应当可以还原回原有的图像x1和x2，即：

c1‘，d2’=E(g1)

c2’ , d1’=E(g2)

x1’’ = G(c1’,d1’)

x2’’ = G(c2’,d2’)

根据训练要求我们定义了若干组loss函数：

**重构loss**：重构loss主要有两组，第一组为，使用编码器-解码器生成图像，则生成的图像应当与原有的图像相同：

Lrecon=E x~p(x)=[||G(Ec(x),Es(x))-x||1]

另外一组为将d1和d2特征进行交换后，使用解码器生成的g1和g2，同时将g1和g2进行编码，并将新的d1和d2重新交换回原有的位置并解码生成的新图像应当与原有的图像相同：

Lcycle1=

E x1~p(x1) x2~p(x2)

[||G(Ec(G(Ec(x1),Es(x2))),

Es(G(Ec(x2),Es(x1))))||1]

Lcycle2=

E x1~p(x1) x2~p(x2)

[||G(Ec(G(Ec(x2),Es(x1))),

Es(G(Ec(x1),Es(x2))))||1]

**对抗生成网络loss：**在生成器和判别器的对抗训练中，我们使用了WGAN-GP 的loss函数定义来对网络做训练，在这里我们将原始图像x1 x2作为真实图像，将原始图像通过编码器解码器生成的图像x1’ x2’ g1 g2 以及二次重构的图像 x1’‘ x2’‘作为生成图像用于对抗训练：



**分类网络loss：**另外为了使得生成的图像g1和g2向指定的类别转换，我们加入了分类网络的loss对生成图像g1和g2的类别进行约束,对于真实图像有：

L cls=E x,y [-log C(y|x)]

其中在真实图像部分，我们使用真实图像数据x对C做分类训练

对于生成图像有：

L cls’=E g,y [-log C(y|g)]

其中在生成图像的部分，我们固定C，对生成g的网络做训练。

整体上来看即分类器使用真实数据进行训练来提升分类能力，并对生成的图像以梯度传递的方式提供指导意见。

实验

1 数据集

在实验中，我们选取了CelebA人脸数据集作为训练数据，CelebA数据集共有约20W张人脸图像，每张图像包含了40种标签，包含了男女，老少，表情，胡须，装饰物等信息，在本实验中，我们选取了CelebA的20w数据作为训练集，并重点使用了其中4个标签：男/女，是否化妆，年轻衰老，是否笑。我们用这4组标签进行了网络训练，训练完成的网络中，特征d则应当编码了控制这四组标签的信息，另外我么对CelebA数据集还做了中心采样，将数据集规范成一个正方形并将图片重新调整为128\*128大小，并在训练过程中，将像素规约到了-1到1范围，以使得网络方便训练。

2 网络结构

在网络结构上，编码器E和解码器G共同构成了一个U-net结构，并将U-net结构有压缩成的最小尺寸部分作为图像的特征d的编码，将编码器每层卷积产生的输出作为特征c的编码，在解码阶段，解码器G接受编码器E的每层卷积的输出以及最小尺寸的特征，并按照U-net网络解码的方式将图像还原，具体的网络结构如下表：（此处插入一个表格）

判别器D和分类器G则使用了相同的卷积网络结构，使用了步长为2的卷积将特征图逐渐成倍减小，最终判别器D产生1个输出用于判断图像是否真实，分类器C产生4个输出用于对图像进行分类，具体的网络结构如下表：（此处插入一个表格）

3 训练参数

在这个实验中，每一个batch的数据包含两张图片和对应的标签，每一步训练的batch size为4，每25000个iteration为一个epoch，共训练20个epoch。对于参数，我们使用了xavier初始化初始化了所有网络；在训练过程中，我们使用了Adam优化器优化所有的网络模块，其参数分别为0.5和0.99，所有的模块的学习率均为0.001，学习率的衰减系数设定为0.001。

4 实验结果

我们训练网络，并对每个epoch的网络训练得到的网络进行储存，然后在测试过程中我们加载已经训练完成的模型，然后一次输入两张图片，首先提取第一张图片的特征d1和特征s1，然后提取第二张图片的特征d2和特征s2，然后我们使用d1和s1来作为编码，使用解码器G来恢复原有的图像查看图像自编码效果，另外我们使用d1和s2来作为编码，使用解码器G来生成新的图像，我们可以发现，生成的新的图像仍然是第一张图的人物，同时迁移了第二张图的男女，化妆，表情，老少等信息。实验结果如图：（插图并说明）

另外为了展示我们生成图像的多样性，我们提取了多张图像的s信息，并将这些信息于一张图片的d信息，进行组合，我们发现，由于不同图片提取出的s信息不同，生成的若干组图片显示了不同的人物特点。实验结果如图：（插图并说明）

最后我们还使用第一张图像的的特征d1和第二张图像的特征d2进行了线性插值处理，共做了8组插值，并将新的插值特征与第一张图的特征s1来一起用作解码，生成了新的若干组图像，这些图像展示了人脸的渐变过程。实验结果如图：（插图并说明），可以看出我们的网络结构具有优良的线性插值特性。