基于半监督编码生成对抗网络的图像分类模型

付 晓 [1] 沈远彤 [1] 李宏伟 [1] 程晓梅 [1]

摘 要 在实际应用中, 为分类模型提供大量的人工标签越来越困难, 因此, 近几年基于半监督的图像分类问题获得了越来越多的关注. 而大量实验表明, 在生成对抗网络 (Generative adversarial network, GANs) 的训练过程中, 引入少量的标签数据能获得更好的分类效果, 但在该类模型的框架中并没有考虑用于提取图像特征的结构, 为了进一步利用其模型的学习能力, 本文提出一种新的半监督分类模型. 该模型在原生成对抗网络模型中添加了一个编码器结构, 用于直接提取图像特征, 并构造了一种新的半监督训练方式, 获得了突出的分类效果. 本模型分别在标准的手写体识别数据库 MNIST、街牌号数据库 SVHN 和自然图像数据库 CIFAR-10 上完成了数值实验, 并与其他半监督模型进行了对比, 结果表明本文所提模型在使用少量带标数据情况下得到了更高的分类精度.

关键词 深度学习, 生成对抗网络, 图像分类, 半监督学习

引用格式 付晓, 沈远彤, 李宏伟, 程晓梅. 基于半监督编码生成对抗网络的图像分类模型. 自动化学报, 2020, **46**(3): 531−539

DOI 10.16383/j.aas.c180212

A Semi-supervised Encoder Generative Adversarial Networks Model for Image Classification

FU Xiao[1] SHEN Yuan-Tong[1] LI Hong-Wei[1] CHENG Xiao-Mei[1]

Abstract The semi-supervised image classification task has attracted more and more attention recently owing to the problem that adequate labeled data is hard to acquire from industrial applications. Meanwhile, considerable works demonstrate that the improved generative adversarial networks (GANs) can achieve great classification performance with only few labeled images. Intuitively, GAN is a generative model, there is no semantic feature extractor in the main framework. In order to further utilize the ability of GANs, we propose to add an encoder in the framework to extract features of images directly, and simultaneously to use a new semi-supervised training method to train this new image classification model. The classification results of experiments have shown the state-of-the-art accuracy performance in semi-supervised MNIST, SVHN and CIFAR-10.

Key words Deep learning, generative adversarial network (GAN), image classification, semi-supervised learning

Citation Fu Xiao, Shen Yuan-Tong, Li Hong-Wei, Cheng Xiao-Mei. A semi-supervised encoder generative adversarial networks model for image classification. *Acta Automatica Sinica*, 2020, **46**(3): 531−539

随着互联网的普及和智能信息处理技术的迅速发展, 大规模图像资源不断涌现, 面对海量的图像信息, 如何准确地归类整理图像内容变得尤为重要, 所以图像分类问题成为近年来的研究重点. 图像分类就是根据图像的不同特征将不同类别的图像区分开来, 因此一个好的特征提取方法是影响图像分类效果的重要因素. 最近, 机器学习方法在图像处理的各个领域都取得了很大的成功, 特别是图像分类领域. 大量实验证明, 机器学习方法提取的特征较传统手工方法提取的特征在图像分类上能获得更好的分类效果[1].

机器学习方法一般分为三大类: 有监督学习、无监督学习以及半监督学习. 由于有监督学习方法需要大量的人工标注, 而在一般的实际应用中, 提供大量的标签数据无疑会消耗庞大的人力物力, 所以在无监督学习的基础之上, 结合有监督训练的半监督学习成为学者们的研究热点. Suddarth 等[2]于 1990 年第一次提出在无监督学习过程中引入预测值和训练集真实标签之间的误差, 将无监督训练得到的神经网络作为其他图像处理问题的初始参数, 进而完成了不同的图像任务. 而在深度学习算法兴起之后, 可供选择的无监督深度学习算法有很多, 例如深度自编码网络[3]、生成对抗网络 (Generative adversarial networks, GANs)[4], 以及把每一个样本当成单独的一个类别进行训练的卷积神经网络 (Co-

收稿日期 2018-04-12 录用日期 2018-08-30
Manuscript received April 12, 2018; accepted August 30, 2018
国家自然科学基金 (61601417) 资助
Supported by National Natural Science Foundation of China (61601417)
本文责任编委 金连文
Recommended by Associate Editor JIN Lian-Wen
1. 中国地质大学数学与物理学院 武汉 430074
1. College of Mathematics and Physics, China University of Geosciences, Wuhan 430074

nvolutional neural network, CNN)[5] 等. 将有监督算法与上述的无监督方法进行结合, 均能得到效果不错的半监督学习模型. 例如利用自编码网络性质构造的阶梯网络[6], 该模型由一个无监督网络连接一个有监督网络组成, 它能有效地从数据信息中筛选出与分类任务相关的信息. 还有学者利用对抗的思想, 对样本施加对抗性噪声, 并训练模型使加噪样本和未加噪样本的输出结果类似, 从而使模型具备学习无标注样本的能力, 完成半监督学习[7].

在大部分无监督学习方法中, 生成模型是一个不错的选择, 一般的生成模型都有隐性学习原始图像信息的能力, 很多通过优化生成模型搭建的半监督框架都取得了良好的分类效果[8]. 近年来, 由于 GANs 具有从简单的隐变量分布中模拟产生任意复杂数据的能力, 很多学者选择对原本的 GANs 进行优化, 以期获得在半监督图像分类领域更好的效果. 例如, 改变网络的训练误差, 通过数据的不确定熵信息对分类器进行训练的策略 GANs[9]; 改变模型中的鉴别器结构, 将输出层直接连接分类器, 使数据分为原始类别和一个假图像类, 训练得到半监督分类鉴别器[10]. 还有学者提出了一个实用的贝叶斯公式, 使 GANs 进行半监督式学习[11]. 但在这些半监督 GANs 框架中有隐性学习图像信息的结构, 而没有考虑直接从隐变量中提取图像特征.

为了更好地应用 GANs 的特征学习能力, 优化图像分类的效果, 本文提出一种半监督编码生成对抗网络 (Semi-supervised encoder GAN, SSE-GAN). 此网络在原 GANs 模型中添加一个编码器结构作为生成结构的逆运算, 从而获得原始数据的本质特征, 并将此特征用于图像分类. 由于生成图像的过程就是通过对图像本质特征的逐步提取, 进而学习图像表达以及产生图像数据, 可以认为, 生成器的这种从内而外的学习方式学习到的特征准确和全面, 而作为其逆运算的编码器同时保留下这些图像特征的信息, 所以使用这种保留图像特征的编码器结构进行图像分类比使用鉴别器更加准确. 本文还将有监督与无监督学习相结合, 构造了一种新的半监督训练方法, 进一步提高了图像分类的准确度.

1 生成对抗网络

生成对抗网络 (GANs) 近两年引起了机器学习界的广泛关注, 其主要思想是构造两个模型来模拟人类博弈游戏, 其中一个模型是生成器, 主要负责将随机隐变量映射成图像; 另一个模型是鉴别器, 主要负责辨别输入的图像是来自图像库还是来自生成器. 在训练 GANs 的过程中, 通过最大化真实图像与生成图像分布之间的差异来优化鉴别器, 而最小化这个差异来优化生成器. 整个模型通过反复不断地对抗训练, 最终达到生成图像成功误导鉴别器的目的, 即生成器完美地模拟了真实数据的分布.

优化 GANs 模型时, 鉴别器相当于一个函数 D, 它的输入是图像, 输出是该图像来自真实图像库中的概率, 而生成器相当于一个从随机隐变量空间到真实图像空间的映射 G, 所以 GANs 的损失函数为[4]

$$V(D,G)=E_{x\sim p_{\text{data}}}[\ln D(x)]+E_{z\sim p_z}[\ln(1-D(G(z)))] \tag{1}$$

其中, x 为真实图像库中的图像, p_{data} 为其分布, z 为随机隐变量, p_z 为其分布, 一般为高斯白噪声分布, $D(x)$ 代表真实图像输入鉴别器后的输出概率值, $D(G(z))$ 对应的则是生成图像通过鉴别器后的输出概率值, 其中 $G(z)$ 为隐变量通过生成器得到的生成图像. 该损失函数通过式 (1) 的形式, 将真实图像鉴别概率的对数期望与负的生成图像鉴别概率的对数期望相加, 实现了对生成图像的分布与真实图像分布之间的差异度量. 而 GANs 模型利用该损失函数优化模型参数时, 使用的是对抗训练的方式, 即最小化该损失来训练生成器中的参数, 最大化该损失来训练鉴别器中的参数. 对损失函数进行化简得:

$$V(D,G)=\int p_{\text{data}}(x)[\ln D(x)]+p_g(x)[1-\ln(D(x)]\mathrm{d}x \tag{2}$$

其中, p_g 代表生成图像的分布, 当 G 的参数固定时, 对上述公式求导, 可得:

$$\frac{\partial V}{\partial D_G}=\frac{p_{\text{data}}}{D}-\frac{p_g}{1-D} \tag{3}$$

上述导数等于 0 时, 求得最佳的 D 为

$$\begin{aligned}&\frac{p_{\text{data}}}{D}=\frac{p_g}{1-D}\Leftrightarrow\\&p_{\text{data}}-Dp_{\text{data}}=Dp_g\Leftrightarrow\\&D(p_{\text{data}}+p_g)=p_{\text{data}}\Leftrightarrow\\&D_{G^*}=\frac{p_{\text{data}}}{p_{\text{data}}+p_g}\end{aligned} \tag{4}$$

所以当生成器的生成图像与真实数据图像的分布一致时, 鉴别器将以 50 % 的概率判断某个输入图像是否来自真实图像分布. 但是这个损失函数在训练时非常不稳定, 大量的实验表明, 生成器与鉴别器的学习能力若能始终保持对应平衡, 损失函数将更易于收敛[12]. 在优化生成器时需要鉴别器有一定的辨别能力, 因此应先优化鉴别器再优化生成器; 但是当鉴别器的辨别能力过强时, 它又不能给生成器的参数提供有效的梯度, 所以鉴别器和生成器需要循环交替地进行优化.

虽然 GANs 作为一个生成模型, 具有非常卓越的生成能力, 可以模拟非常复杂的图像分布, 但是把它用于图像分类则还应该进一步将图像特征给予明确的表述. 所以若要将 GANs 用于解决图像分类问题, 则需要为其添加一个提取特征的结构.

2　半监督编码生成对抗网络

为了利用 GANs 的学习能力, 提高图像分类的准确率, 本文提出了半监督编码生成对抗网络 (SSE-GAN) 模型. 该模型是一个半监督图像分类模型, 其主要思想是在生成器的对应位置添加一个编码器, 通过半监督训练的方式训练该编码器, 使之能直接提取图像特征. 若将这个编码器看作一个映射, 则这个函数的主要作用是将数据从图像空间映射到特征空间. 由于模型中的编码器主要是模拟生成器的逆运算, 所以在模型优化时, 生成器获取真实图像数据分布的同时, 编码器也能够模拟真实数据对应的随机隐变量分布即图像特征. 因此在 SSE-GAN 中, 鉴别器的输入将不再仅是图像数据, 而是图像数据以及对应的特征信息.

2.1　编码生成对抗网络

在 GANs 模型中添加一个编码器结构, 然后利用特征与图像共同输入鉴别器的方法训练模型, 本文称之为编码生成对抗网络. 因为在编码生成对抗网络中, 鉴别器需要接受图像和特征两种不同空间维度的输入, 一般的做法是将特征与图像直接结合输入神经网络, 即特征数据通过复制扩充直接与图像数据相结合[13]. 这些结合形式虽然仍能对网络进行训练, 但从客观角度上来说会导致大量不必要的计算损失.

从流形学习[14] 的角度来说, 直接将二维数据通过复制与三维数据结合, 即强行将二维数据平面贴在三维曲面上, 自然会一定程度上造成数据的不贴和. 由于函数 D 进行映射时, 相当于将数据由原本的流形状态展开成平面, 而将上述强行结合的数据展开为平面时, 必不可少的会出现褶皱情况, 为了抚平这种数据褶皱, 在调整函数 D 的参数时需要大量预先计算. 而在 SSE-GAN 中, 隐变量并不直接与图像结合, 而是如图 1 所示先将图像做流形结构展开, 然后将其特征与之结合. 这种特殊的结合方式称为流形一致结合, 这种结合方式可以去除数据的不平整现象, 减少网络的预计算过程.

在具体进行流形一致结合时, 还应考虑到隐变量与图像特征之间存在的差异导致批量初始化 (Batch normalization, BN)[15] 操作出现异常, 所以需对隐变量及图像特征进行 L2 范数归一化[16], 再相互结合并进行 BN 操作, 即第 l 层神经元输出为

$$\boldsymbol{X}_{l+1}=f\left\{\boldsymbol{W}_l\mathrm{BN}\left[\mathrm{Con}\left(\frac{\boldsymbol{X}_l}{\boldsymbol{m}_1},\frac{\boldsymbol{Z}}{\boldsymbol{m}_2}\right)\right]+\boldsymbol{b}_l\right\} \tag{5}$$

其中, $\boldsymbol{m}_1$ 表示图像特征 $\boldsymbol{X}_l$ 的模, $\boldsymbol{m}_2$ 表示随机隐变量 $\boldsymbol{Z}$ 的模, 函数 $\mathrm{Con}(\boldsymbol{X}_l,\boldsymbol{X}_2)$ 表示将矩阵 $\boldsymbol{X}_l$ 与矩阵 $\boldsymbol{X}_l$ 按照列进行合并, $\boldsymbol{W}_l$ 和 $\boldsymbol{b}_l$ 分别表示第 l 层网络的权值与阈值, f 代表网络使用的激活函数. 由于在合并之前进行了 L2 范数归一化, 此时反向求导调整第 $l-1$ 层参数时, 要在原本导数基础之上除以图像特征的模. SSE-GAN 模型使用上述的流行一致结合方式将特征与图像结合, 可以大幅减少模型的预计算过程. 本文在 MNIST 数据库[17] 上进行了对比实验, 经过多次重复试验发现, 使用流形一致结合方式的编码 GANs 模型达到收敛时, 网络的总迭代次数约为 5 次, 而使用传统图像与特征直接结合方式的模型达到完全收敛, 所需的迭代次数约为 9 次, 从上述结果可以看出这种流形一致结合方式能大幅度地有效提高整个模型的收敛速度.

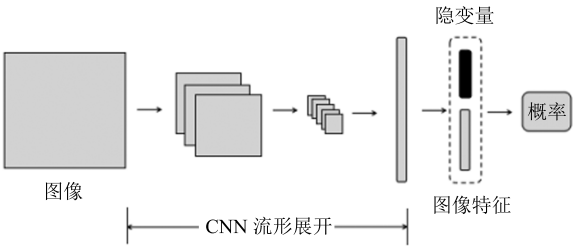


图 1　SSE-GAN 模型中流形一致结合方式

Fig. 1　The manifold agreement combination method in SSE-GAN

2.2　半监督网络训练

SSE-GAN 采用半监督的学习方法训练编码生成对抗网络, 其损失函数定义为

$$\begin{aligned}V(D,G,Enc)=&E_{x\sim p_{\mathrm{data}}}[\ln D(x,Enc(x))]+\\&E_{z\sim p_z}[\ln(1-D(G(z),z))]+\\&E_{x,y\sim p_{\mathrm{data}}}[\ln(p_{Enc}(y|x))]\end{aligned} \tag{6}$$

其中, y 是少量带标样本数据的标签, x 为图像库的采样样本, p_{data} 为图像库样本分布, $Enc(x)$ 是样本通过编码器得到的特征, z 为服从高斯分布的随机隐变量, $G(z)$ 是生成器从随机隐变量中产生的样本, 而函数 $D(x,z)$ 表示在本模型中鉴别器将接受图像和特征两个输入, 并且输入的图像数据和其对应的特征采取第 2.1 节提出的流形一致结合方式进行结合. 网络采取循环交替优化的方式训练, 固定 G 的参数时, 同样可以得到最佳的函数 D 表示为

$$D_{G^*,Enc^*}=\frac{p_{EX}}{p_{EX}+p_{GZ}} \tag{7}$$

其中, $p_{EX}=p_{\text{data}}(x)p_{Enc}(z|x)$, $p_{GZ}=p_Z(z)p_G(x|z)$. 其实损失函数中的前两部分与原本的 GANs 损失函数类似, 所以由上可知, 当 $p_{EX}=p_{GZ}$ 时, 网络收敛, 而此时则有

$$\begin{aligned} p_G=&\int p_{x\sim G,z\sim N(0,1)}(x,z)\mathrm{d}z=\\ &\int p_z(z)p_G(x|z)\mathrm{d}z=\\ &\int p_{\text{data}}(x)p_{Enc}(z|x)\mathrm{d}z=\\ &\int p_{x\sim\text{data},z\sim Enc}(x,z)\mathrm{d}z=p_{\text{data}} \end{aligned}\tag{8}$$

也即当网络收敛时, 生成器仍能很好地模拟真实数据的分布, 所以可以通过观察网络迭代训练时生成器产生的图像质量来判断网络是否收敛. 当网络收敛时有 $p_{GZ}=p_Z(z)p_G(x|z)=p_{\text{data}}(x)p_{Enc}(z|x)$, 因此由 $p_G=p_{\text{data}}$ 可以得到 $p_Z=p_{Enc}$, 即编码器也学习到了生成器输入特征的分布. 所以 SSE-GAN 的损失函数前两部分的效果则是利用对抗训练的原理, 获得数据无监督学习特征, 而 SSE-GAN 的损失函数最后一部分则是有标签数据的交叉熵损失.

本模型这种将有监督损失与无监督损失相结合共同调整网络参数的训练方式, 能更进一步提高网络的学习能力, 同时还能使本模型的学习过程更加接近人类学习过程. 模型在具体实现半监督训练时, 只需要在检测到数据含有标签时加上交叉熵损失, 而无标签数据只使用损失函数中前两部分的损失即可. SSE-GAN 框架如图 2 所示, 模型首先通过生成器对随机隐变量映射得到生成图像, 再通过编码器对图像库中图像映射得到图像特征, 最后将真实和生成的图像与其对应特征共同输入鉴别器, 通过鉴别器判别输入是否来自真实数据. 反向调整编码器和生成器的参数, 使生成图像不断逼近数据库的图像, 同时编码器获取图像的本质特征.

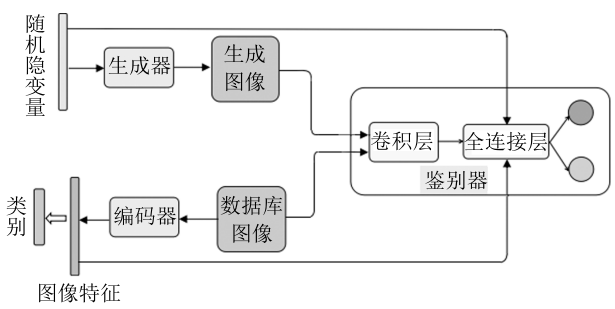


图 2 SSE-GAN 框架图

Fig. 2 The framework of SSE-GAN

2.3 具体算法步骤

本文提出的 SSE-GAN 模型算法具体步骤如下:

步骤 1. 在高斯白噪声中随机采样生成隐变量 z, 并将随机隐变量输入生成器 G, 得到生成图像 $G(z)$;

步骤 2. 将生成图像 $G(z)$ 与隐变量 z 按照流形一致结合方式结合, 共同输入鉴别器得到输出概率值 $D(G(z),z)$;

步骤 3. 随机在图像库中采样得到真实图像 x, 并将真实图像输入编码器 Enc, 得到图像特征 $Enc(x)$;

步骤 4. 将图像库图像 x 与其对应的特征 $Enc(x)$ 按照流形一致结合方法结合, 共同输入鉴别器得到输出概率值 $D(x,Enc(x))$;

步骤 5. 若采样的真实图像 x 无标签, 则通过式 (6) 的前两部分得到损失函数, 若真实图像 x 有标签, 则添加标签误差作为损失函数, 利用 Adam 梯度下降方法调整鉴别器的参数;

步骤 6. 固定鉴别器的参数, 利用损失函数调整生成器与编码器的参数;

步骤 7. 重复步骤 1 ∼ 6, 直至网络收敛;

步骤 8. 将测试图像输入到编码器中, 编码器的输出即图像的类别.

3 实验与分析

实验部分检验了 SSE-GAN 对图像的特征分类能力, 在训练结束之后, 直接利用模型中的编码器作为数据的分类器对测试数据进行分类测试. 为了展示 SSE-GAN 对各类复杂数据都有很好的学习能力, 本文实验在 MNIST 数据库, SVHN 数据库[18] 和 CIFAR-10 数据库[19] 三个数据库上进行测试. 由第 2.2 节的论证可知, 当模型中生成器的生成图像与原图像库图像基本一致时, 网络趋于收敛, 所以实验部分还会展示生成器的生成结果与原图像的对比情况. 实验部分中 SSE-GAN 模型主要由 CNN 堆叠组成, 其中编码器结构与生成器网络结构完全相反, 即编码器网络的高层神经元结构与生成器网络底层神经元结构相同, 并且编码器和生成器中每个全连接层后都添加了 Dropout 处理[20] 以避免网络过拟合.

3.1 MNIST 图像库分类实验

本节实验部分采用 MNIST 手写体识别图像数据库, 该数据库中的图像均是大小为 28 像素 × 28 像素的手写体数字图像. 数据库中共有 60 000 个训练数据, 10 000 个测试数据. 该数据库的图像均为黑白单通道图像, 初始化数据时直接将图像数据归一化到区间 $[0,1]$, 且在实验中生成器的最后一层激活函数采用 sigmoid 函数, 使生成图像与原图像在数

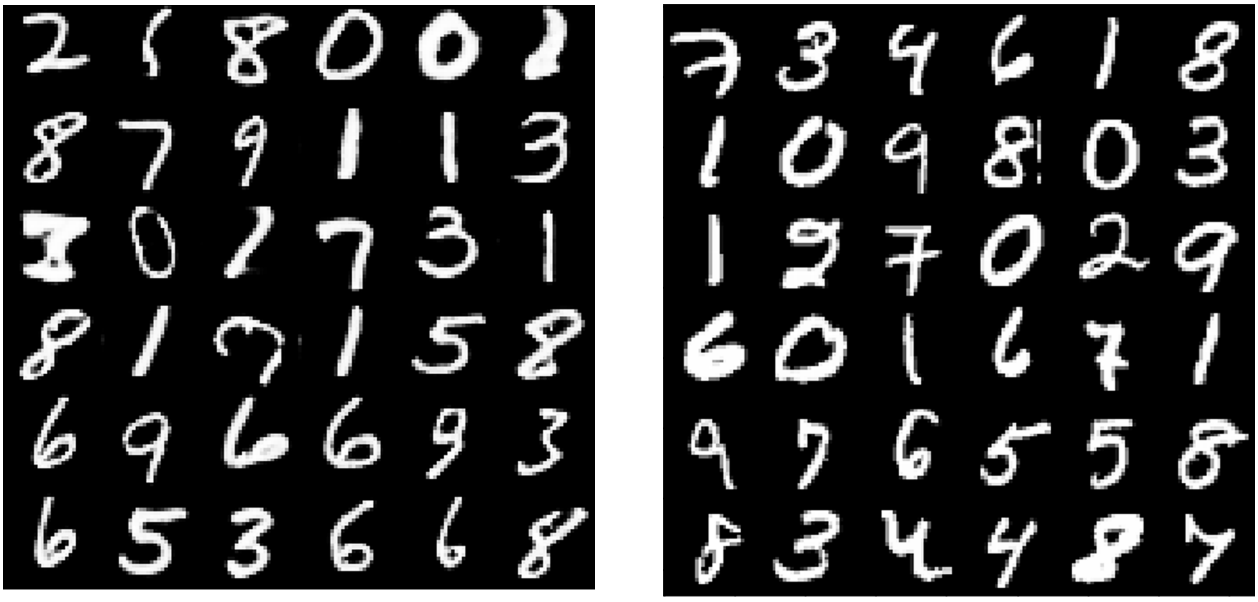
值空间中更一致. 本实验使用的 SSE-GAN 模型结构类似经典 DCGAN[12] 的结构, 其中鉴别器从低到高依次由 3 层卷积神经结构和 3 层全连接神经结构组成, 网络的输入为 $28\times28\times1$ 的图像数据, 卷积层的通道个数从底层到高层依次为 32, 64 和 128, 且第一个卷积操作只改变输入图像通道数而不改变其尺寸, 全连接层神经元的个数从底层到高层分别为 6 272 (最后一个卷积层输出向量化变形得到), 1 024 和 100; 生成器从底层到高层依次由 2 层全连接神经结构和 3 层卷积神经结构组成, 网络的输入为含有 100 个元素的一维向量, 全连接层神经元的个数从底层到高层分别为 1 024 和 6 272, 卷积层的通道个数从底层到高层依次为 128, 64 和 1; 模型中编码器的网络结构与生成器的结构相反; 网络中所有卷积层的卷积核大小均为 5×5; 实验模型中添加的 Dropout 操作系数为 0.5.

实验选取 1 000 个带标数据, 训练迭代 15 次时模型收敛. 图 3 是模型收敛后, SSE-GAN 生成器的生成图像与原图像库中的部分图像对比. 由模型的迭代次数可以看出模型的收敛速度很快, 而且由于使用了少量标签数据, SSE-GAN 模型变得更稳定, 生成器的生成图像多样性更加丰富, 不会出现网络的训练崩塌到一个数据点的现象.

表 1 显示了在 MNIST 数据库上, SSE-GAN 模型与其他经典半监督模型在选取不同数目带标数据时的分类准确率. 对比可知本文提出的模型能获得更高的分类准确度, 也说明 SSE-GAN 模型框架中, 编码器对图像数据类别特征的学习能力很强. 由于某些方法在原文部分实验中未给出具体参数, 因此表 1 中部分结果空缺. 但是在 MNIST 数据库中, 图像均为简单的灰度数字图像, 一般简单的分类模型均能达到不错的分类精度, 因此, 接下来的实验将模型应用到其他更复杂的图像中, 来说明 SSE-GAN 模型的分类能力.

3.2 SVHN 图像库分类实验

本节实验采用 SVHN 图像库, 该数据库中的数据均为大小为 32 像素 $\times$ 32 像素的彩色街牌号图像, 每张图像上有一个或者多个数字, 但是图像的类别以正中间的数据为基准. 数据库中共有 73 257 个训练数据, 26 032 个测试数据. 由于该数据库的图像均为彩色多通道图像, 直接归一到区间则会导致模型的不稳定[12], 所以初始化数据时需要将图像数据归一化到区间, 因此, 为了使生成图像与原图像在数值空间中更匹配, 本实验中生成器的最后一层激活函数采用 tahn 函数. 本实验使用的 SSE-GAN 模型结构与 MNIST 实验中的结构类似, 由于使用的输入数据库尺寸不同, 所以本实验使用的模型比第 3.1 节所述模型在鉴别器、生成器和编码器中都多一个 256 通道的卷积神经结构, 这个结构将被添加在上述网络的全连接层和卷积层之间. 实验模型中添加的 Dropout 操作系数同样设置为 0.5.



(a) SSE-GAN 生成的 MNIST 图像
(a) The generated MNIST image of SSE-GAN

(b) 原 MNIST 图像库图像
(b) The image from MNIST database

图 3　模型收敛后生成图像与原 MNIST 数据库图像对比

Fig. 3　The generated image and the image from MNIST database after model converges

表 1 MNIST 数据库上不同数量带标数据的半监督训练分类准确率

Table 1 Using different number of labeled data when semi-supervised training on MNIST

模型	带标数据个数及对应分类准确率 (%)		
	100	1 000	全部数据
Ladder-network[6]	98.14	99.06	–
Cat-GAN[9]	98.09	99.11	99.40 ± 0.03
Improved-GAN[10]	98.58	99.15	99.40 ± 0.02
ALI[21]	98.77	99.16	99.45 ± 0.01
GAR[22]	98.92	99.21	99.55 ± 0.03
SSE-GAN	**99.10**	**99.23**	$\mathbf{99.61 \pm 0.03}$

实验随机在训练数据中添加少量标签数据进行训练时, 网络可以较快收敛. 图 4 展示了利用 500 个带标数据时, SSE-GAN 模型的生成器生成图像和原数据库图像对比, 可以看出生成图像已经与原图像库中的部分图像基本匹配.

表 2 分别在 500 个和 1 000 个带标数据情况下, 与目前常见的半监督分类算法的准确率进行对比, 结果可以看出 SSE-GAN 的分类精度较高, 说明该模型具备强大的特征学习能力, 网络的编码器可以充当很好的分类器的角色. 相比 GANs 模型, 本模型虽然增加了一个网络结构, 但是训练过程中引入了少量标签数据, 这使得网络收敛的速度有所提高, 所以仍然具有很高的训练效率. 对比表 2 和表 1 的数据可以看到, 在处理较复杂的数据时, SSE-GAN 模型对数据的分布拟合的更加准确, 因此在少量数据的情况下, SSE-GAN 模型应用于复杂数据时得到的分类精度提升更加明显.

表 2 SVHN 数据库上不同数量带标数据的半监督训练分类准确率

Table 2 Using different number of labeled data when semi-supervised training on SVHN

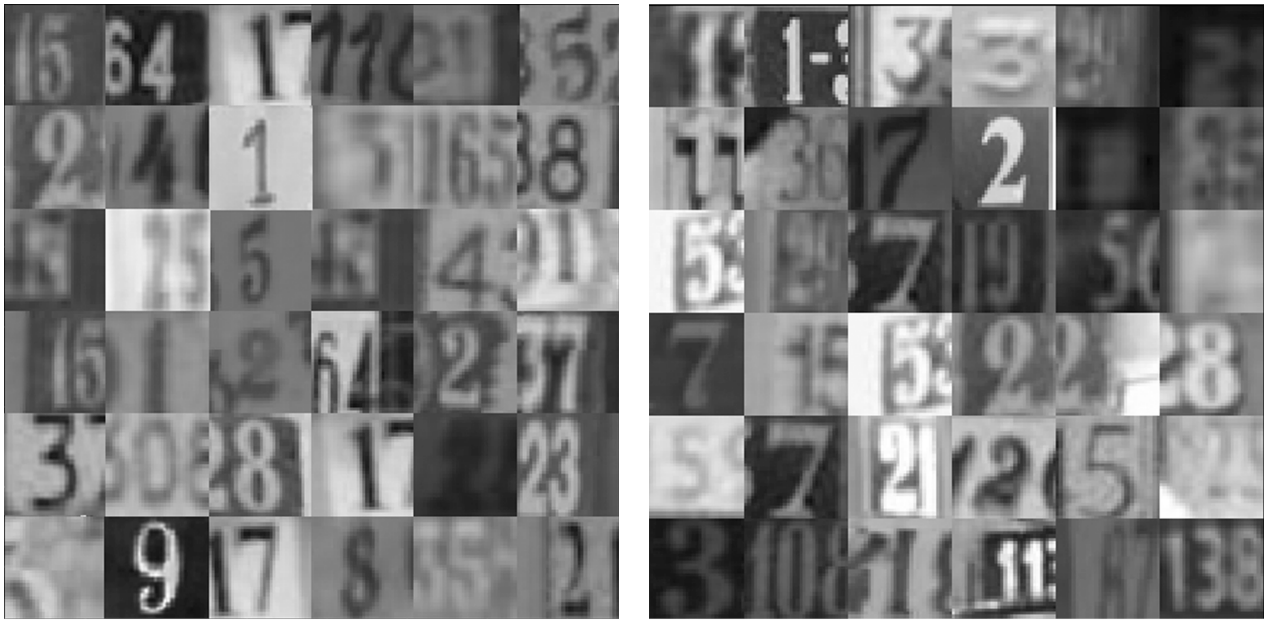
模型	带标数据个数及对应分类准确率 (%)	
	100	1 000
Ladder-network[6]	75.50	87.06
Cat-GAN[9]	77.68	88.90
Improved-GAN[10]	–	90.78
Virtual Adversarial[23]	79.71	90.99
Adversarial Training[7]	79.99	91.11
Bayesian GAN[11]	80.53	92.01
GAR[[22]	80.87	92.08
SSE-GAN	**81.08**	**92.92**

3.3 CIFAR-10 图像库分类实验

第 3.1 节和第 3.2 节的实验所使用的数据库均是数字识别, 为了验证 SSE-GAN 能够应用于更复杂的数据, 本节实验采用 CIFAR-10 数据库, 这个数据库中的数据均为自然图像, 图像中存在很多精致细节, 而且同一类型的图像之间差异比较大, 所以此数据库对分类模型的鲁棒性要求很高. 数据库中的图像大小为 32 像素 × 32 像素, 且都是彩色多通道图像. 数据库中共有 50 000 个训练数据, 10 000 个测试数据, 由于本数据库仍然是彩色图像, 所以其预处理方式与第 3.2 节实验相同. 本实验使用的输入数据库尺寸与第 3.2 节实验相同, 所以实验使用的模型与第 3.2 节实验中 SSE-GAN 模型结构相同. 但是本实验使用的数据为自然图像数据, 图像数据中细节信息非常丰富, 因此为了使模型中生成器的训练更加稳定, 特别将生成器网络中 Dropout 操作的系数设置为 0.3.

随机选取 2 000 个有标签数据对网络进行训练之后, SSE-GAN 生成器的生成图像与原数据库图像对比见图 5, 可以明显观察到 SSE-GAN 模型对于自然图像数据也具备较强的学习能力, 但是对图像细节信息的学习有些欠缺. 有些生成图像比较模糊, 其主要原因是整个网络利用半监督的学习机制训练, 在生成器的训练过程中, 会有少量的空间信息损失, 网络参数在调整的时候趋向于保留与分类任务相关的信息, SSE-GAN 模型用于图像的分类, 仍然具有高的分类准确率.

表 3 是不同带标数据与目前常见算法分类准确率的对比, 从表 3 的分类准确率对比可以看出, 与其他的 GANs 相关网络相比, 本模型的分类效果更为突出. 与此同时, SSE-GAN 不仅对数据的学习能力强, 而且只是采用最基础的交叉熵结合原损失的形式, 在面对不同的图像处理任务时, 模型还有很好的普适性. 同时, SSE-GAN 模型在 CIFAR-10 数据库上优异的分类表现, 也证实了该模型是鲁棒性很强



(a) SSE-GAN 生成 SVHN 图像
(a) The generated SVHN image of SSE-GAN

(b) 原 SVHN 图像库图像
(b) The image from SVHN database

图 4　模型收敛后生成图像与原 SVHN 数据库图像对比

Fig. 4　The generated image and the image from SVHN database after model converges

(a) SSE-GAN 生成 CIFAR-10 图像
(a) The generated CIFAR-10 image of SSE-GAN

(b) 原 CIFAR-10 图像库图像
(b) The image from CIFAR-10 database

图 5　模型收敛后生成图像与原 CIFAR-10 数据库图像对比

Fig. 5　The generated image and the image from CIFAR-10 database after model converges

的模型.

4　结束语

本文提出了一种用于图像分类的半监督模型 SSE-GAN, 该模型添加一个编码器结构作为 GANs 模型中生成器的逆运算, 直接提取图像数据特征用于分类. 同时, 该模型还利用无监督与半监督损失共同训练网络的形式, 构造出分类精度高的半监督分

表 3 CIFAR-10 数据库上不同数量带标数据的半监督训练分类准确率

Table 3 Using different number of labeled data when semi-supervised training on CIFAR-10

模型	带标数据个数及对应分类准确率 (%)		
	1 000	2 000	4 000
Ladder-network[6]	–	76.52	79.31
Cat-GAN[9]	–	78.83	80.42
improved-GAN[10]	77.17	79.39	81.37
ALI[21]	80.02	80.91	81.48
Adversarial training[7]	81.25	82.88	83.61
Bayesian GAN[11]	81.89	83.13	84.20
GAR[[22]	82.10	83.35	84.94
SSE-GAN	**82.34**	**83.66**	**85.14**

类器. 实验表明面对各类复杂的图像数据, SSE-GAN 均可利用少量的标签数据训练得到效果显著的分类器. 大部分优化的 GANs 模型均使用鉴别器对图像进行分类, 而 SSE-GAN 模型则是通过添加编码器的形式, 逆向利用了 GANs 中的生成器来进行图像分类, 使网络能够直接学习本质特征, 降低了信息在处理过程中损失程度, 所以网络的分类效果更好. 且本模型中提出的半监督损失函数的结构也具有一定的普适性, 可以通过改变有监督部分的误差使其能应用于其他多个图像处理任务. 值得指出的是, SSE-GAN 模型与大部分 GANs 模型相似, 均存在训练参数过多的问题, 虽然本文提出的流形一致结合方式能一定程度上加快模型的收敛速度, 但是模型的训练耗时仍远大于传统图像分类方法, 因此后期的主要工作是通过对网络损失进行改进, 进一步提高网络的收敛速度.

References

1 Zhang Hao-Kui, Li Ying, Jiang Ye-Nan. Deep learning for hyperspectral imagery classification: the state of the art and prospects. *Acta Automatica Sinica*, 2018, **44**(6): 961−977
(张号逵, 李映, 姜晔楠. 深度学习在高光谱图像分类领域的研究现状与展望. 自动化学报, 2018, **44**(6): 961−977)

2 Suddarth S C, Kergosien Y L. Rule-injection hints as a means of improving network performance and learning time. *Neural Networks. EURASIP 1990*, 1990. 120−129

3 Li Min, Yu Long, Tian Sheng-Wei, Ibrahim T, Zhao Jian-Guo. Coreference resolution of uyghur noun phrases based on deep learning. *Acta Automatica Sinica*, 2017, **43**(11): 1984−1992
(李敏, 禹龙, 田生伟, 吐尔根·依布拉音, 赵建国. 基于深度学习的维吾尔语名词短语指代消解. 自动化学报, 2017, **43**(11): 1984−1992)

4 Wang Kun-Feng, Zuo Wang-Meng, Tan Ying, Qin Tao, Li Li, Wang Fei-Yue. Generative adversarial networks: from generating data to creating intelligence. *Acta Automatica Sinica*, 2018, **44**(5): 769−774
(王坤峰, 左旺孟, 谭营, 秦涛, 李力, 王飞跃. 生成式对抗网络: 从生成数据到创造智能. 自动化学报, 2018, **44**(5): 769−774)

5 Dosovitskiy A, Fischer P, Springenberg J T, Riedmiller M, Brox T. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, **38**(9): 1734−1747

6 Rasmus A, Valpola H, Honkala M, Berglund M, Raiko T. Semi-supervised learning with ladder networks. arXiv: 1507.02672, 2015.

7 Miyato T, Maeda S, Ishii S, Koyama M. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, DOI: 10.1109/TPAMI.2018.2858821

8 Kingma D P, Rezende D J, Mohamed S, Welling M. Semi-supervised learning with deep generative models. In: Proceedings of the 2014 Neural Information Processing Systems. Massachusetts, USA: MIT Press, 2014. 3581−3589

9 Springenberg J T. Unsupervised and semi-supervised learning with categorical generative adversarial networks. arXiv: 1511.06390, 2015.

10 Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. In: Proceedings of the 2016 Neural Information Processing Systems. Massachusetts, USA: MIT Press, 2016. 1−10

11 Saatchi Y, Wilson A G. Bayesian GAN. In: Proceedings of the 2017 Neural Information Processing Systems. Massachusetts, USA: MIT Press, 2017. 1−16

12 Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. In: Proceedings of the 2016 International Conference on Learning Representations. Piscataway, USA: IEEE, 2016. 1−16

13 Donahue J, Krähenbühl P, Darrell T. Adversarial feature learning. In: Proceedings of the 2017 International Conference on Learning Representations. Piscataway, USA: IEEE, 2017. 111−128

14 Tenenbaum J B, de Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000, **290**(5500): 2319−2323

15 Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 2015 International Conference on Machine Learning. Piscataway, USA: IEEE, 2015. 11−21

16 Zheng L, Wang S J, Tian L, He F, Liu Z Q, Tian Q. Query-adaptive late fusion for image search and person re-identification. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2015. 1741−1750

17 LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, **86**(11): 2278−2324

18 Netzer Y, Wang T, Coates A, Bissacco A, Wu B, Na A Y. Reading digits in natural images with unsupervised feature learning. In: Proceedings of the 2011 Neural Information Processing Systems. Massachusetts, USA: MIT Press, 2011. 5−16

19 Krizhevsky A. Learning Multiple Layers of Features from Tiny Images [Ph. D. dissertation], University of Toronto, Toronto, Canada, 2009.

20 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In: Proceedings of the 2012 Neural Information Processing Systems. Massachusetts, USA: MIT Press, 2012. 1106−1114

21 Dumoulin V, Belghazi I, Poole B, Mastropietro O, Lamb A, Arjovsky M, et al. Adversarially Learned Inference. In: Proceedings of the 2017 International Conference on Learning Representations. Piscataway, USA: IEEE, 2017. 111−128

22 Kilinc O, Uysal I. GAR: an efficient and scalable graph-based activity regularization for semi-supervised learning. *Neurocomputing*, 2018, **296**: 46−54

23 Miyato T, Maeda S, Koyama M, Nakae K, Ishii S. Distributional smoothing with virtual adversarial training. In: Proceedings of the 2016 International Conference on Learning Representations. Piscataway, USA: IEEE, 2016. 1−12

付 晓 中国地质大学 (武汉) 数学与物理学院硕士研究生. 2015 年获得中国地质大学 (武汉) 数学与物理学院学士学位. 主要研究方向为深度学习与图像处理. E-mail: cugfuxiao@163.com
(**FU Xiao** Master student at the College of Mathematics and Physics, China University of Geosciences. She received her bachelor degree from China University of Geosciences in 2015. Her research interest covers deep learning and image processing.)

沈远彤 中国地质大学 (武汉) 数学与物理学院教授. 主要研究方向为小波分析理论与应用, 数字图像处理. 本文通信作者. E-mail: whsyt@163.com
(**SHEN Yuan-Tong** Professor at the College of Mathematics and Physics, China University of Geosciences. His research interest covers theory and application of wavelet analysis and digital image processing. Corresponding author of this paper.)

李宏伟 中国地质大学 (武汉) 数学与物理学院教授. 主要研究方向为信息处理与智能计算. E-mail: hwli@cug.edu.cn
(**LI Hong-Wei** Professor at the College of Mathematics and Physics, China University of Geosciences. His research interest covers information processing and intelligent computing.)

程晓梅 中国地质大学 (武汉) 数学与物理学院硕士研究生. 2016 年获得山东大学 (威海) 数学与统计学院统计系学士学位. 主要研究方向为深度学习与图像处理. E-mail: 13016471716@163.com
(**CHENG Xiao-Mei** Master student at the College of Mathematics and Physics, China University of Geosciences. She received her bachelor degree from Shandong University in 2016. Her research interest covers deep learning and image processing.)