智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于改进TransGAN的零样本图像识别方法

翟永杰, 张智柏, 王亚茹

引用本文:
翟永杰,张智柏,王亚茹. 基于改进TransGAN的零样本图像识别方法[J]. 智能系统学报, 2023, 18(2): 352–359.
ZHAI Yongjie,ZHANG Zhibai,WANG Yaru. An image recognition method of zero –shot learning based on an improved TransGAN[J]. *CAAI Transactions on Intelligent Systems*, 2023, 18(2): 352–359.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202111002

DOI: 10.11992/tis.202111002
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20221025.1351.002.html

基于改进 TransGAN 的零样本图像识别方法

翟永杰,张智柏,王亚茹
(华北电力大学 自动化系,河北 保定 071003)

摘 要:零样本学习算法旨在解决样本极少甚至缺失情况下的图像识别问题。生成式模型通过生成缺失类别的图像,将此问题转化为传统的基于监督学习的图像识别,但生成图像的质量不稳定、容易出现模式崩塌,影响图像识别准确性。为此,通过对 TransGAN 模型进行改进,提出基于改进 TransGAN 的零样本图像识别方法。将 TransGAN 的生成器连接卷积层进行降维,并进一步提取图像特征,使生成图像特征和真实图像特征更加接近,提高特征的稳定性;同时,对判别器加入非线性激活函数,并进行结构简化,使判别器更好地指导生成器,并减小计算量。在公共数据集上的实验结果表明,所提方法的图像识别准确率较基线模型提高了 29.02%,且具有较好的泛化性能。

关键词:零样本学习;生成对抗网络;TransGAN;深度学习;图像识别;图像特征;卷积层;非线性激活函数
中图分类号:TP18 **文献标志码**:A **文章编号**:1673−4785(2023)02−0352−08

中文引用格式:翟永杰, 张智柏, 王亚茹. 基于改进 TransGAN 的零样本图像识别方法 [J]. 智能系统学报, 2023, 18(2): 352–359.
英文引用格式:ZHAI Yongjie, ZHANG Zhibai, WANG Yaru. An image recognition method of zero -shot learning based on an improved TransGAN[J]. CAAI transactions on intelligent systems, 2023, 18(2): 352–359.

An image recognition method of zero -shot learning based on an improved TransGAN

ZHAI Yongjie, ZHANG Zhibai, WANG Yaru
(Department of Automation, North China Electric Power University, Baoding 071003, China)

Abstract: Zero-shot learning algorithms aim to address the challenge of image recognition with limited or even missing samples. By transforming the problem into a supervised learning task through the use of generative models, the method generates images of missing classes. However, the quality of generated images can be inconsistent and is susceptible to pattern collapse, affecting image recognition accuracy. To address this issue, we propose an improved zero-shot learning image recognition method based on an improved TransGAN. The generator of TransGAN is linked to a convolutional layer for dimensionality reduction, leading to a more effective extraction of image features and improved stability. Moreover, the addition of a nonlinear activation function to the discriminator and simplifying its structure enhances its ability to guide the generator and reduces computational requirements. Experiment results on public datasets show that our proposed method increases image recognition accuracy by 29.02% compared to the baseline model and demonstrates improved generalization performance.

Keywords: zero -shot learning; generative adversarial network; TransGAN; deep learning; image recognition; image feature; convolutional layer; nonlinear activation function

目前,深度卷积神经网络在图像识别领域的研究取得了优异的成果,深度学习具有提取特征能力强、实时性快、识别精度高等特点,使得有监督的图像识别算法的识别精度已经达到甚至超过了人眼的识别精度。然而,深度神经网络实现高精度识别结果需要大量已标注的数据集进行训练。随着互联网技术的发展和人类对自然的探索不断深入,各种各样的数据增长迅速,人工对这些数据进行筛选、标注往往费时费力。同时自然界中的物种呈现长尾效应,一部分样本数据的获取量十分稀少,甚至无法获取。针对以上问题,Larochellec 等[1] 于 2008 年首次提出了零样本学

收稿日期:2021−11−01. **网络出版日期**:2022−10−25.
基金项目:国家自然科学基金面上项目 (U21A20486, 61871182);河北省自然科学基金青年科学基金项目 (F2021502008);中央高校基本科研业务费专项资金面上项目 (2021MS081).
通信作者:王亚茹. E-mail: wangyaru@ncepu.edu.cn.

习 (zero shot learning, ZSL) 的概念, 即利用可见类对不可见类进行识别。2009 年 Palatucci 等[2]在神经网络中应用零样本学习, 之后零样本学习逐渐深入, 旨在于解决不可见类的识别问题。

在零样本学习方法中, 属性信息的应用大大促进了零样本学习的发展。属性信息作为一种具有人工先验知识的类别描述, 对类别进行类别语义表征。零样本学习在训练阶段只包含可见类的图像特征和所有类别的语义特征, 通过搭建视觉模态和语义模态的映射关系, 实现对不可见类的识别[3-5]。Lampert 等[6]提出了直接属性预测模型 (direct attribute prediction model, DAP) 与间接属性预测模型 (indirect attribute prediction model, IAP)。DAP 模型使用训练数据直接学习图像特征与属性之间的映射关系, 再根据属性去识别不可见类所属的类别; IAP 模型则使用可见类数据学习图像特征到已知类的映射, 学习多个分类器, 再构建类别与属性之间的映射。Romera-paredes 等[7]提出了 ESZSL 模型, 在图像和标签之间建立属性空间, 并进一步建立了特征与属性的映射关系。Qiao 等[8]通过使用在线文本源并抑制噪声, 匹配文本和视觉特征完成对不可见类的识别。尽管零样本学习得到了广泛研究, 但由于视觉模态与语义模态存在着语义鸿沟, 建立跨模态的映射关系容易造成语义特征信息的丢失, 零样本学习方法仍存在着可见类与不可见类的领域偏移问题, 因此基于跨模态迁移的零样本学习模型对不可见类的识别精度不够理想[9-10]。

近年来, 生成对抗网络飞速发展, 在图像生成、文本生成、风格迁移等方面表现优异。GAN 由生成器和判别器组成, 生成器生成虚假图像欺骗判别器, 判别器则试图去区分虚假图像和真实图像, 两部分互相对抗学习, 使得虚假图像逼近于真实图像。针对零样本识别精度较低的问题, 一些学者提出了基于生成对抗网络 (generative adversarial network, GAN)[11]的零样本学习。Zhu 等[12]使用维基百科文本生成虚假图像特征, 并引入 GAN 作为辅助分类器。Chen 等[13]将 CycleGAN 应用于零样本中, 提高了属性与图像之间的映射关联。然而上述基于 GAN 的零样本方法的生成器未能学习到具有足够多样性的图像数据, 且生成的虚假特征不稳定, 导致零样本图像识别精度有限。

TransGAN[14]使用 transformer 结构作为生成器与判别器, 应用自注意力机制保证图像生成的质量, 可有效创建高分辨率图像。本文提出一种基于改进 TransGAN 的零样本图像识别方法, 对 TransGAN 模型进行改进, 提高零样本图像识别的准确率。对生成器加入两层卷积层, 提高生成图像特征的稳定性, 提高其与真实图像特征的相似度; 对判别器进行结构简化, 减小了网络计算内存, 并加入非线性激活函数, 使得判别器更好地指导生成器。实验结果验证了所提方法的有效性。

1　TransGAN 模型

GAN 的主要思想是通过生成器和判别器不断进行二元博弈, 最终学习真实样本分布。GAN 一经提出便吸收了卷积神经网络 (convolutional neural networks, CNN) 中批量归一化、池化等优点。为了减轻 GAN 训练阶段的不稳定性, 产生了 SGAN、AdaGAN、WGAN 等改进模型[15-16]。Transformer 最初在自然语言处理上表现优秀, Dosovitskiy 等[17]将其应用在图像分类方向并得到了较好的效果。TransGAN 的主要思想与 GAN 相同, 主要的不同是只使用两个 Transformer 构造作为生成器和判别器, 而抛弃了 CNN 结构。其中, 生成器逐步减小嵌入维度并同时提高特征分辨率, 判别器则在 Patch 级进行判别。TransGAN 结构示意图如图 1 所示。

TransGAN 中生成器以一维随机噪声作为输入, 经过多层感知机 (multi-layer perceptron, MLP) 得到一个维度为 $H\times W\times C$ 的向量, 该向量与可学习的位置编码相结合, 之后通过分段式设计, 逐步增加特征图分辨率并减小维度, 直到满足目标分辨率, 其中采用金字塔结构减少计算量和计算机内存需求。具体操作如图 2 所示, 首先将一维向量 $HW\times C$ 变形为二维特征图 $(H\times W)\times C$, 之后采用上采样处理, 得到 $(2H\times 2W)\times C/4$ 的特征图, 再将其变形为一维向量 $4HW\times C/4$。重复多次, 最终将通道数投影到 3 并获得 RGB 图像, 即为虚假图像。

判别器输入真实图像或虚假图像并判断其真假。TransGAN 设计了多尺度判别器, 如图 1 所示, 输入图像会在不同阶段进行不同尺度的切分并输入进判别器, 当切分的 patch 较小时, 判别器可以更有效地处理图像的细节特征, 较大的 patch 可以处理图像的结构信息。输入图像在第 1 阶段切分为 $4n\times 4n$ 个 patch, 经过线性变换和位置编码后作为第 1 阶段输入。第 2 阶段切分为 $2n\times 2n$ 个 patch 并与第 1 阶段输出进行拼接, 再通过平均池化层进行下采样, 作为第 2 阶段输入, 如此重复。最后通过全连接层输出真假预测。但当输入图像尺寸较大时, 模型对硬件的计算能力要求较大。

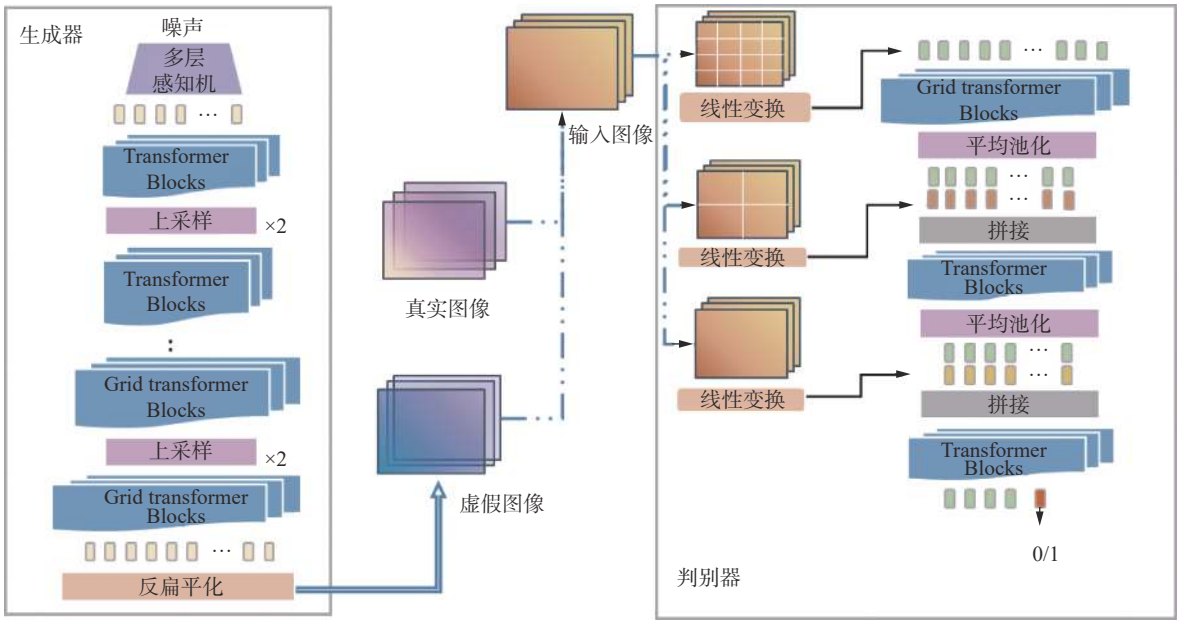


图 1　TransGAN 结构

Fig. 1　Structure of TransGAN

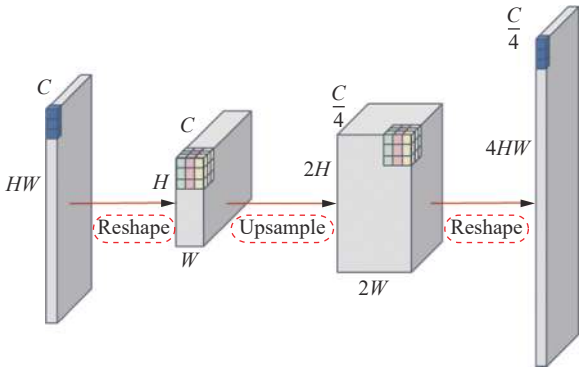


图 2　特征维度变换流程图

Fig. 2　Feature size processing flowchart

TransGAN 中 Transformer Blocks 由多个 Transformer 编码器组成，单个 Transformer 编码器结构如图 3 所示。

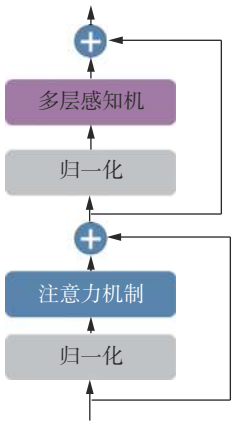


图 3　Transformer 编码器结构示意

Fig. 3　Structure of transformer encoder

Grid Transformer Blocks 与其区别在于：图像特征在输入自注意力模块前进行了分割，相较于 Transformer Blocks 节省了计算量。Grid Transformer Blocks 被应用在处理分辨率较高的特征图上，能够在保证生成高分辨率图像的同时，减小自注意力机制的计算量。

2　基于改进 TransGAN 的零样本图像识别模型

2.1　零样本图像识别模型

零样本图像识别将数据集 D 分为可见类 S 和不可见类 U，将可见类 S 作为训练集，不可见类 U 作为测试集。训练集 $S=\{(x,y,c(y))\,|\,x\in X_s, y\in Y_s, c(y)\in C\}$，$x\in X_s$ 表示神经网络提取的可见类图像特征，$y\in Y_s$ 为可见类的标签，测试集 $U=\{(x,y,c(y))\,|\,x\in X_u, y\in Y_u, c(y)\in C\}$，其中 $X_s\cap X_u=\varnothing$，$X_s\cup X_u=X$，$Y_s\cap Y_u=\varnothing$，$c(y)$ 表示对应的 y 类的属性信息，由人工标注得到，用来描述对应类别物体的颜色、形状等属性。ZSL 的任务是训练分类器 $f: X_u\rightarrow Y_u$。

基于 TransGAN 的零样本图像识别模型结构如图 4 所示。具体过程为，给定真实的可见类图像数据集，利用残差网络 ResNet101(residual network)[18] 提取出真实图像特征 $x\in X_s$。将随机高斯噪声 z 与可见类属性信息 $c(y)$ 的拼接向量输入改进 TransGAN 模型的生成器，输出相应的虚假可见类图像特征 x'，即 $G(z,c(y))=x'$。分别将真实可见类图像特征 x 和虚假可见类图像特征 x' 与可见类属性信息 $c(y)$ 进行拼接，然后将两组拼接向量输入判别器。判别器用于区分真实图像特征和虚假图像特征。

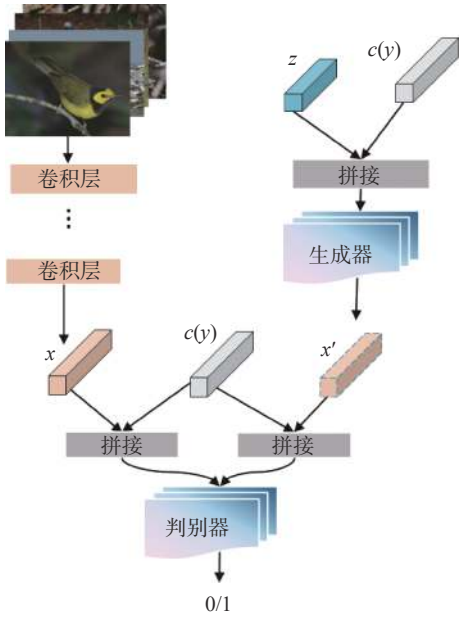


图 4　基于改进 TransGAN 的 ZSL 模型结构

Fig. 4　Model structure of ZSL based on improved TransGAN

模型中生成器不生成完整的虚假图像, 而是生成图像特征。因此对 TransGAN 进行改进, 具体见 2.2 节。改进的 TransGAN 模型训练完成后, 对于不可见类 U, 将高斯噪声 z 和不可见类属性信息 $c(y)$ 的拼接向量输入到生成器, 其中 $y \in Y_u$, 生成不可见类的虚假图像特征 $x_u = G(z, c(y))$, 然后对传统的基于监督学习的 Softmax 分类器进行训练, 训练完成后, 对不可见类测试集图像进行分类识别。

2.2　改进 TransGAN 模型

由于原 TransGAN 生成的是完整图像, 为使 TransGAN 生成图像特征并减小计算量, 本文对 TransGAN 中的生成器和判别器进行了改进。改进后的生成器和判别器网络结构如图 5 所示。为保持与真实图像特征维度 x 一致, 本文在生成器的最后加入两层卷积层 (convolutional layer, Conv) 来进一步提取特征, 并减小生成器输出的特征维度。相较于全连接层, 卷积层采用稀疏连接方式, 训练参数更少, 并且卷积层不会破坏像素与像素之间的空间结构。同时由于判别器输入为一维向量, 本文简化了图 1 中的判别器, 对输入向量只进行第一阶段处理, 减小计算量。生成器与判别器中各神经层输入及输出维度如表 1 和表 2 所示。

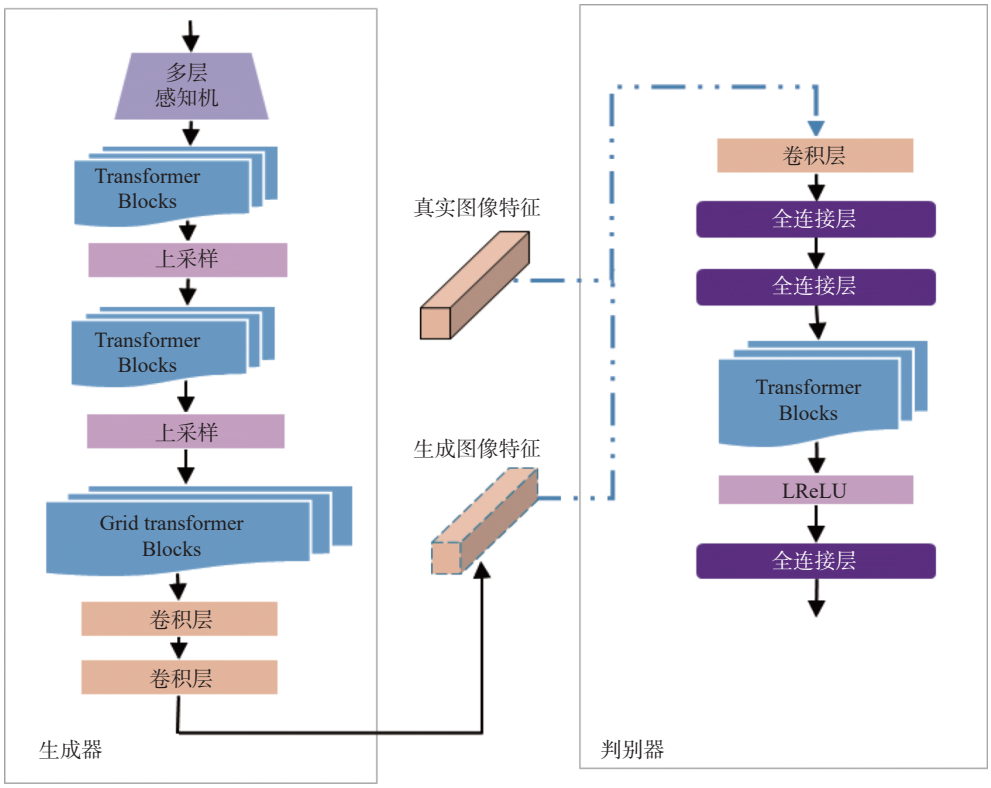


图 5　生成器 (左) 和判别器 (右) 网络结构

Fig. 5　Structure of generator(left) and discriminator(right)

生成器和判别器实际上是极小极大的博弈[19], 损失函数为

$$\min_G \max_D L_{\mathrm{GAN}} = \mathrm{E}[\log_2(D(x, c(y)))] + \mathrm{E}[\log_2(1 - D(x', c(y)))] \tag{1}$$

判别器在生成器更新参数前被训练到最优, 判别器试图将损失最大化, 生成器试图将损失最小化。随着网络迭代, 判别器逐渐趋于饱和, 理论上最终可以实现损失为零。

表 1　生成器网络结构

Table 1　Generator network structure

阶段	层	输入尺寸	输出尺寸
1	MLP	1 336	(8×8)×128
2	Block	(8×8)×128	(8×8)×128
	Block	(8×8)×128	(8×8)×128
	Block	(8×8)×128	(8×8)×128
3	Upsample	(8×8)×128	(16×16)×32
	Block	(16×16)×32	(16×16)×32
	Block	(16×16)×32	(16×16)×32
	Block	(16×16)×32	(16×16)×32
	Block	(16×16)×32	(16×16)×32
4	Upsample	(16×16)×32	(32×32)×8
	Block	(32×32)×8	(32×32)×8
	Block	(32×32)×8	(32×32)×8
5	Conv	(32×32)×8	512×8
	Conv	512×8	512×4

表 2　判别器网络结构

Table 2　Discriminator network structure

阶段	层	输入尺寸	输出尺寸
1	Conv	2048+1 024	3 072×64
	FC	3 072×64	1 560×64
	FC	1 560×64	256×64
2	Add CLS Token	256×64	(256+1)×64
	Block	(256+1)×64	(256+1)×64
	Block	(256+1)×64	(256+1)×64
	Block	(256+1)×64	(256+1)×64
	Block	(256+1)×64	(256+1)×64
	Block	(256+1)×64	(256+1)×64
3	FC	1×64	1

由于真实图像特征 $x \in X_s$ 由卷积神经网络提取，无法保证有效地训练判别器。为了使判别器可以对输入数据进行准确判断，同时使生成器可以生成更近似于真实图像特征 x 的虚假特征 x'，对生成的虚假图像特征 x' 采用负对数似然损失函数[20]。最终的损失函数为

$$L = -\mathrm{E}[\log(D(x', c(y)))] + \min_G \max_D L_{\mathrm{GAN}} \tag{2}$$

通过迭代训练，最终可以通过类别的属性特征得到与真实图像特征相似的虚假图像特征用于分类器训练，进而对测试集图像进行识别。

3　实验与分析

3.1　实验环境与数据集

本文所有的模型在 Ubuntu 16.04 操作系统下运行，使用 NVIDIA GeForce 1080Ti GPU 进行训练。使用 Pytorch 深度学习框架实现，设置学习率为 0.1×10^{-3}，选择 Adam 作为生成器和判别器的优化策略，网络参数随机初始化。

本文选择 FLO[21] 和 CUB[22] 作为实验数据集。FLO 数据集为 Nilsback 等于 2008 年提出的花卉数据集，共计包含图像 8 189 幅，总共 102 类，每类由 40~258 张图像组成，每类图片包含人工标注的 1 024 维属性特征。CUB 数据集为加州理工学院于 2010 年提出的细粒度数据集，共计图像 11 788 幅，包含 200 类鸟类子类，每个类别包含 312 维不同的属性。其中，训练集和测试集无交集。本文使用平均 top-1 识别准确率 acc(accuracy)[23] 作为各模型性能的评价指标。

3.2　实验结果与分析

对于 FLO 数据集，本文基于改进 TransGAN 的零样本图像识别模型的图像识别结果如表 3 所示。其中，生成器 (阶段 5) 一列表示表 1 中生成器网络结构中所对应的阶段 5 部分的不同网络结构，判别器 (阶段 3) 一列表示表 2 中判别器网络结构中所对应的阶段 3 部分的不同网络结构。第 1 组实验为基线模型。实验中，生成器和判别器的注意力机制设置 head 数为 1。

表 3　不同网络结构下的识别结果

Table 3　Recognition results under different network structures

序号	生成器(阶段5)	判别器(阶段3)	准确率/%
1	FC+FC	FC	26.61
2	Conv+Conv	FC	51.06
3	Conv+Conv+FC	FC	47.21
4	Conv+Conv	Sigmoid+FC	52.67
5	**Conv+Conv**	**LReLU+FC**	**55.63**
6	FC+FC	LReLU+FC	36.85

零样本学习的主要目的是对不可见样本进行识别，所以使用测试集正确率为性能验证指标不断对模型进行改进。对于生成器模型，在第 1 组和第 6 组实验中，生成器后加入两层全连接层用于降低特征维度，使生成特征与真实特征维度一致；在第 2 组、第 4 组和第 5 组中使用两层卷积层进行降维，第 3 组中使用了两层卷积层和一层全连接层。对于判别器模型，第 1 组、第 2 组和第 3 组只使用全连接层输出判别结果；第 4 组在全连接层前加入 Sigmoid 函数；第 5 组和第 6 组中在全连接层前使用 LReLU 函数

对比第 1 组与第 2 组实验可以发现，当判别

器结构相同时, 生成器后加入两层全连接层的网络模型, 准确率远低于加入两层卷积层的网络模型; 对比第 1 组和第 3 组实验, 可以发现生成器中添加卷积层后, 实验准确率提升了 20.6%; 对比第 2 组和第 3 组实验, 在卷积层后添加全连接层并没有进一步提高准确率。对此, 原因有以下两个方面:

1) 本文没有在所有像素点上使用全连接层, 而是在通道数上使用全连接层进行降维, 造成了一部分特征信息的损失, 破坏了生成图像特征的空间结构; 卷积层由于参数共享和稀疏连接, 减少了计算消耗, 可以在像素级进行处理, 保留了图像特征中像素点之间的空间信息。

2) 真实图像特征通过深度卷积神经网络提取, 生成器中使用卷积可以在一定程度上贴合真实图像特征分布。

为此, 本文计算了生成图像特征和真实图像特征之间的余弦相似度, 计算公式为

$$d_{AB}=0.5+\frac{\sum_{i=1}^{n}(\boldsymbol{A}_i\times\boldsymbol{B}_i)}{2\times\sqrt{\sum_{i=1}^{n}(\boldsymbol{A}_i)^2}\times\sqrt{\sum_{i=1}^{n}(\boldsymbol{B}_i)^2}} \tag{3}$$

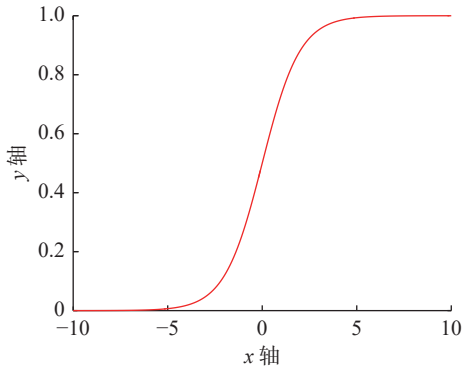
式中: $\boldsymbol{A}$和$\boldsymbol{B}$是两个 n 维向量; d_{AB}为两个向量的余弦相似度。d_{AB}越接近 1, 表示两个向量越相似, d_{AB}接近 0, 表示两个向量相似度不高。本文针对表 3 中的第 5 组和第 6 组实验, 计算了生成特征与真实特征之间的余弦相似度, 实验结果如表 4 所示。第 5 组和第 6 组实验中的判别器结构相同, 均在判别器后加入 LReLU 函数和全连接层, 第 5 组实验中网络模型在生成器后加入两层卷积层, 第 6 组实验中在生成器后加入两层全连接层。可以发现使用卷积层的生成特征与真实特征的余弦相似度更接近 1, 特征更加相似, 实验的准确率也较高。

表 4　生成特征与真实特征的余弦相似度

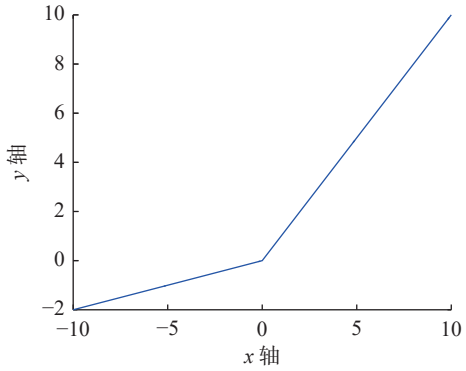
Table 4　Cosine degree of the generated feature and the real feature

序号	生成器(阶段5)	准确率/%	余弦相似度
4	Conv+Conv	54.97	0.9566
6	FC+FC	35.44	0.5178

对比表 3 中第 2 组、第 4 组、第 5 组实验数据, 可以发现判别器中非线性激活函数的应用提高了模型的识别准确率。判别器加入 Sigmoid 函数后, 模型的准确率提高了 1.61%, 再更改为 LReLU 后, 准确率又提高了 2.96%。Sigmoid 和 LReLU 函数图像如图 6 所示。如图所示, Sigmoid 接近饱和区域时, 在进行反向传播时容易出现梯度消失的现象, 而 LReLU 函数会使一部分神经元的输入为负, 可以增加网络的稀疏性, 减小过拟合发生的可能。



(a) Sigmoid



(b) LReLU

图 6　激活函数

Fig. 6　Activation function

3.3　与其他方法的对比

为验证本文方法的有效性, 在保持网络结构和设置不变的情况下, 在 CUB 数据集上进行对比实验, 不同网络模型的图像识别结果如表 5 所示, 其中对比方法 ESZSL[7]、ZSLNS[8]、WAC[24]、GAZSL[12]、CANZSL[13] 的实验数据来自于相应文献, 本文训练集和测试集类别的选取与相应对比算法相同。

由表 5 中实验数据可知, 本文算法较传统零样本识别方法 ESZSL、ZSLNS、WAC 图像识别精度明显提升, 分别为 21.3%、20.7% 和 16.3%。与生成式模型 GAZSL 和 CANZSL 相比, 在训练集类别数为 180 类时, 本文算法的识别准确率比 GAZSL 提高 6.1%, 比 CANZSL 提高 4.0%; 当训练集类别数为 120 类时, 本文算法的识别准确率比 GAZSL 提高 11.8%, 比 CANZSL 提高 7.8%, 这表示在可见类类别数较少时, 本文算法仍能有效地生成稳定的虚假图像特征, 提高图像识别准确性, 证明本文算法具有较高的准确性和较好的泛化性。

表 5　不同网络模型的图像识别结果

Table 5　Image recognition results of different network models

模型	训练集类别数	测试集类别数	准确率/%
ESZSL	180	20	28.5
ZSLNS	180	20	29.1
WAC	180	20	33.5
GAZSL	180	20	43.7
GANZSL	180	20	45.8
本文算法	**180**	**20**	**49.8**
GAZSL	120	80	10.3
CANZSL	120	80	14.3
本文算法	**120**	**80**	**22.1**

本文进一步使用 Vision Transformer (ViT)[17] 网络提取数据集的真实图像特征，对两种不同生成器结构进行对比实验，实验结果如图 7 所示。

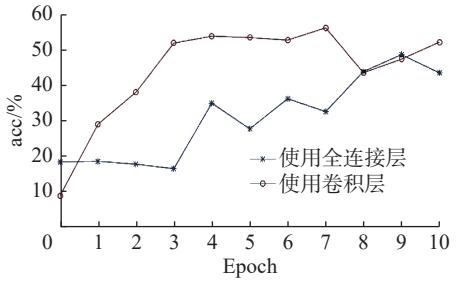


图 7　ViT 提取特征实验结果

Fig. 7　Chart of results with ViT feature extraction

图 7 中，生成器中添加了两层全连接层的生成器网络与表 3 第 6 组实验中生成器结构相同。生成器中添加卷积层所取得的准确率，为了匹配 ViT 提取的真实图像特征维数，在表 3 第 5 组生成器结构的基础上添加了一层全连接层，使得真实图像特征与生成图像特征维数相同。

可以看出在前 10 个 epoch 中，只添加了全连接层的生成器结果最高达到 48.51%，而使用了卷积层的生成器所取得结果较高，最高精确率达 56.06%，高出 7.55%，较 ResNet101 提取的真实图像特征的实验提高了 6.26%。

本文使用 TransGAN 提取一维图像特征，使用卷积层降低特征维度，相较于全连接层减少了信息损失，可以更好地提取特征。同时，ViT 网络中使用了 Transformer 结构，提取的真实图像特征可以更好地贴合本文中生成器的生成特征，因此实验具有较高的准确率。

4　结束语

本文针对生成式零样本图像识别模型的生成图像质量不稳定问题进行研究，将无卷积结构的生成对抗网络 TransGAN 应用于零样本学习，并对其进行改进，使 TransGAN 将先验属性特征生成稳定的图像特征而非完整的图像。对生成器的输出使用卷积层降维并进一步强化提取特征，更好地匹配了生成图像特征和真实图像特征；对判别器进行简化并引入非线性激活函数，减小了网络的计算量并使生成器得到更理想的指导。实验结果表明，本文基于改进 TransGAN 的零样本图像识别方法相比于基线模型以及其他多个对比算法，显著提升了图像识别准确率，验证了本文方法的有效性。

本文的主要工作是针对 TransGAN 在零样本学习方法上进行的，今后将进一步优化 TransGAN 的网络结构，缩小生成特征与真实特征的差异，并在运行速度方面进行优化。

参考文献：

[1] LAROCHELLE H, ERHAN D, BENGIO Y. Zero-data learning of new tasks[C]//AAAI'08: Proceedings of the 23rd national conference on Artificial intelligence - Volume 2. New York: ACM, 2008: 646−651.

[2] PALATUCCI M, POMERLEAU D, HINTON G, et al. Zero-shot learning with semantic output codes[C]//NIPS'09: Proceedings of the 22nd International Conference on Neural Information Processing Systems. New York: ACM, 2009: 1410−1418.

[3] XU Xing, SHEN Fumin, YANG Yang, et al. Matrix tri-factorization with manifold regularizations for zero-shot learning[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2007−2016.

[4] LIU Mingxia, ZHANG Daoqiang, CHEN Songcan. Attribute relation learning for zero-shot classification[J]. Neurocomputing, 2014, 139: 34–46.

[5] 翟永杰, 吴童桐. 基于语义空间信息映射加强的零样本学习方法 [J]. 计算机应用与软件, 2020, 37(12): 113–118,196.
ZHAI Yongjie, WU Tongtong. Zero shot learning based on semantic spatial information mapping enhancement[J]. Computer applications and software, 2020, 37(12): 113–118,196.

[6] LAMPERT C H, NICKISCH H, HARMELING S. Learning to detect unseen object classes by between-class attribute transfer[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 951−958.

[7] ROMERA-PAREDES B, TORR P H S. An Embarrassingly Simple Approach to Zero-Shot Learning[M]// Visual Attributes. Cham: Springer, 2017: 11−30.

[8] QIAO Ruizhi, LIU Lingqiao, SHEN Chunhua, et al. Less is more: zero-shot learning from online textual documents with noise suppression[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 2249−2257.

[9] FU Yanwei, HOSPEDALES T M, XIANG Tao, et al. Transductive multi-view zero-shot learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(11): 2332–2345.

[10] SHIGETO Y, SUZUKI I, HARA K, et al. Ridge regression, hubness, and zero-shot learning[M]//Machine Learning and Knowledge Discovery in Databases. Cham: Springer International Publishing, 2015: 135−151.

[11] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014: 27.

[12] ZHU Y, ELHOSEINY M, LIU B, et al. A generative adversarial approach for zero-shot learning from noisy texts[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City: IEEE, 2018: 1004−1013.

[13] CHEN Z, LI J, LUO Y, et al. Canzsl: Cycle-consistent adversarial networks for zero-shot learning from natural language[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Salt Lake City: IEEE, 2020: 874−883.

[14] JIANG YIFAN, CHANG SHIYU, WANG ZHANGYANG. TransGAN: two pure transformers can make one strong GAN, and that can scale up[EB/OL].(2021−02−04)[2021−10−31].https://arxiv.org/abs/2102.07074.

[15] 王正龙, 张保稳. 生成对抗网络研究综述 [J]. 网络与信息安全学报, 2021, 7(4): 68–85.
WANG Zhenglong, ZHANG Baowen. Survey of generative adversarial network[J]. Chinese journal of network and information security, 2021, 7(4): 68–85.

[16] ADLER J, LUNZ S. Banach Wasserstein Gan[EB/OL].(2018−06−18)[2021−10−31].https://arxiv.org/abs/1806.06621.

[17] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL].(2020−10−22)[2021−10−31].https://arxiv.org/abs/2010.11929.

[18] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas: IEEE, 2016: 770−778.

[19] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of Wasserstein GANs[EB/OL].(2017−12−25)[2020−02−02].https://arxiv.org/abs/1704.00028.

[20] XIAN Yongqin, LORENZ T, SCHIELE B, et al. Feature generating networks for zero-shot learning[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 5542−5551.

[21] NILSBACK M E, ZISSERMAN A. Automated flower classification over a large number of classes[C]//2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. Bhubaneswar: IEEE, 2008: 722−729.

[22] WAH C, BRANSON S, WELINDER P, et al. The caltech-ucsd birds-200-2011 dataset[R]. California Institute of Technology, 2011.

[23] 刘靖祎, 史彩娟, 涂冬景, 等. 零样本图像分类综述 [J]. 计算机科学与探索, 2021, 15(5): 812–824.
LIU Jingyi, SHI Caijuan, TU Dongjing, et al. Survey of zero-shot image classification[J]. Journal of frontiers of computer science and technology, 2021, 15(5): 812–824.

[24] ELHOSEINY M, SALEH B, ELGAMMAL A. Write a classifier: Zero-shot learning using purely textual descriptions[C]// Proceedings of the IEEE International Conference on Computer Vision. Sydney: IEEE, 2013: 2584−2591.

作者简介:

翟永杰, 教授, 博士, 主要研究方向为电力视觉。主持国家自然科学基金面上项目 1 项, 河北省自然科学基金项目 1 项, 主持横向科研项目 12 项, 参与国家重点研发计划项目 1 项, 授权发明专利 10 项, 获得山东省科技进步一等奖 1 项。编著 1 部, 参编教材 1 部、著作 3 部, 发表学术论文 30 余篇。

张智柏, 硕士研究生, 主要研究方向为零样本学习与人工智能。

王亚茹, 讲师, 博士, 主要研究方向为模式识别与计算机视觉、数据挖掘、电力视觉。发表学术论文 10 余篇。