智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

融合VAE和StackGAN的零样本图像分类方法

张冀, 曹艺, 王亚茹, 赵文清, 翟永杰

引用本文:
张冀,曹艺,王亚茹,赵文清,翟永杰. 融合VAE和StackGAN的零样本图像分类方法[J]. 智能系统学报, 2022, 17(3): 593–601.
ZHANG Ji,CAO Yi,WANG Yaru,ZHAO Wenqing,ZHAI Yongjie. Zero-shot image classification method combining VAE and StackGAN[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(3): 593–601.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202107012

DOI: 10.11992/tis.202107012
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20211210.2326.006.html

融合 VAE 和 StackGAN 的零样本图像分类方法

张冀[1], 曹艺[1], 王亚茹[2], 赵文清[1], 翟永杰[2]
(1. 华北电力大学 计算机系, 河北 保定 071003; 2. 华北电力大学 自动化系, 河北 保定 071003)

摘 要: 零样本分类算法旨在解决样本极少甚至缺失类别情况下的分类问题。随着深度学习的发展, 生成模型在零样本分类中的应用取得了一定的突破, 通过生成缺失类别的图像, 将零样本图像分类转化为传统的基于监督学习的图像分类问题, 但生成图像的质量不稳定, 如细节缺失、颜色失真等, 影响图像分类准确性。为此, 提出一种融合变分自编码 (variational auto-encoder, VAE) 和分阶段生成对抗网络 (stack generative adversarial networks, StackGAN) 的零样本图像分类方法, 基于 VAE/GAN 模型引入 StackGAN, 用于生成缺失类别的数据, 同时使用深度学习方法训练并获取各类别的句向量作为辅助信息, 构建新的生成模型 stc-CLS-VAEStackGAN, 提高生成图像的质量, 进而提高零样本图像分类准确性。在公用数据集上进行对比实验, 实验结果验证了本文方法的有效性与优越性。

关键词: 深度学习; 零样本学习; 图像分类; 变分自编码器; 生成对抗网络; 分阶段网络; 句向量; 辅助信息
中图分类号: TP18 **文献标志码**: A **文章编号**: 1673−4785(2022)03−0593−09

中文引用格式: 张冀, 曹艺, 王亚茹, 等. 融合 VAE 和 StackGAN 的零样本图像分类方法 [J]. 智能系统学报, 2022, 17(3): 593–601.
英文引用格式: ZHANG Ji, CAO Yi, WANG Yaru, et al. Zero-shot image classification method combining VAE and StackGAN[J]. CAAI transactions on intelligent systems, 2022, 17(3): 593–601.

Zero-shot image classification method combining VAE and StackGAN

ZHANG Ji[1], CAO Yi[1], WANG Yaru[2], ZHAO Wenqing[1], ZHAI Yongjie[2]
(1. Department of Computer, North China Electric Power University, Baoding 071003, China; 2. Department of Automation, North China Electric Power University, Baoding 071003, China)

Abstract: The zero-shot classification algorithm is designed to solve the classification problem in case of a few samples or even missing categories. With the development of deep learning, the application of the generation model in zero-shot classification has made a breakthrough. By generating images of missing categories, the zero-shot image classification is transformed into a traditional image classification problem based on supervised learning. However, the generated samples are unstable in quality, including missing details and color distortion, thus affecting the accuracy of image classification. To this end, the zero-shot image classification method combining variational auto-encoding (VAE) and stack generative adversarial networks (StackGAN) is proposed. Based on the VAE/GAN model, StackGAN is introduced to generate the data of missing categories. Meanwhile, the deep learning method is used to train and obtain the sentence vectors of each category as auxiliary information and build a new generation model stc–CLS–VAEStackGAN to improve the quality of generated images and subsequently improve the classification accuracy of the zero-shot images. A comparative experiment was conducted on the public dataset, and the experimental results verified the effectiveness and superiority of the method proposed herein.

Keywords: Deep learning; Zero-shot learning; Image classification; Variational autoencoder; Generative adversarial network; Staged network; Sentence vector; Auxiliary information

近年来, 深度学习算法在机器学习领域取得了高速发展, 尤其在图像识别领域, 计算机的识别精度已经达到甚至超过人类识别精度, 但需要消耗大量的人力物力以获得足够数量的人工标注数据[1-2]。在很多实际应用中, 大量有标签的数据难以获取, 物体种类也处于不断增长的趋势, 这就要求计算机训练过程不断增加新样本及新物体种类[3]。如何在样本标签数据不足甚至完全缺失的情况下利用计算机和已有知识对其进行分类识别, 成为深度学习应用研究中亟须解决的问题。为此, 零样本学习应运而生。

零样本学习 (zero-shot learning, ZSL) 也称作

收稿日期: 2021−07−07. **网络出版日期**: 2021−12−14.
基金项目: 国家自然科学基金面上项目 (61773160); 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.

零样本分类，是指根据一些已有类别标签 (seen classes) 的样本数据，辅以相关常识信息或先验知识 (辅助信息)，用于训练某种学习模型，对训练数据或标注完全缺失的类别 (unseen classes) 进行预测和识别的一类技术[4-7]。零样本分类中训练集和测试集的类别是不相交的，这明显区别于传统的基于监督学习的分类任务。该类方法可以看作视觉数据与文本等其他模态数据间的一种跨模态学习。零样本分类方法的发展主要包括 3 个阶段[8-9]：

1）早期的零样本分类方法大多为基于直接语义预测的方法，其中直接属性预测模型 (direct attribute prediction, DAP) 是零样本分类的先驱工作，通过建立视觉数据与属性特征之间的关系，对无标签数据进行分类[10-12]。此类模型虽然在零样本分类领域取得了一定的成果，但依赖于类别的属性特征，极易受到人工属性标注的影响。

2）为更好地解决基于直接语义预测方法存在的问题，出现了基于嵌入模型的方法，其核心思想是将不同模态的数据映射到某一个公共空间中，再根据相似性度量进行零样本分类。Du 等[13]提出基于类内类间约束的语义映射模型，实现类别间知识的有效迁移。陈祥凤等[14]提出基于度量学习改善 SAE 的零样本分类算法，以缓解跨领域漂移。吴晨等[15]提出将融合的语义词向量线性映射到图像特征空间完成分类。谢于中等[16]利用典型相关分析将跨模态特征映射至公共特征空间实现图像的零样本分类。但在训练类与测试类之间建立联系较为困难，且存在领域漂移问题[17]。

3）为解决上述问题，基于深度网络进行视觉样本生成的零样本分类方法涌现出来。Xian 等[18]通过构建生成模型，以类别属性作为辅助信息生成未见类别对应的视觉特征。Sariyildiz 等[19]提出新的损失函数来提高生成样本质量。Mandal 等[20]引入非条件判别器来判别图像属于可见类还是未见类，提高分类准确率。Xian 等[21]提出一个结合 VAE 和生成对抗网络 (generative adversarial networks, GAN) 优势的条件生成模型来进行数据生成。Kim 等[22]提出零样本生成模型 ZSGAN，通过学习可见类和未见类图像与属性之间的关系生成未见类图像。Verma 等[23]提出一种基于类属性条件设置的元学习方法 ZSML（zero-shot meta-learning），将生成器模块和带有分类器的判别器模块分别同元学习代理相关联，利用少量可见类样本训练模型。Ma 等[24]提出一种相似度保持损失，使 GAN 的生成器减小生成样本与真实样本之间的距离，利用相似度消除异常的生成样本。Liu 等[25]提出一种双流生成式对抗网络合成具有语义一致性和明显类间差异的视觉样本，同时保留用于零样本学习的类内多样性。Liu 等[26]提出一种包含两个端到端模型的跨类生成对抗网络用于提高生成的未见类样本的质量。Tang 等[27]提出一种结构对齐的生成对抗网络，以缓解语义差距和领域漂移等问题。Li 等[28]通过构建基于增强语义特征的生成网络来合成未见类别的可分离视觉表示。Gao 等[29]提出一种 VAE 与 GAN 相结合的生成模型 (Zero-VAE-GAN)，以缓解领域漂移问题。但已有模型生成图像的质量仍然不稳定，图像细节的效果较差，影响零样本图像分类的效果。

为此，本文使用 VAE/GAN 变体模型作为生成模型主体，同时将 StackGAN 模型引入其中，得到生成模型 stc-CLS-VAEStackGAN，利用 StackGAN 分阶段生成图像的特点提高图像质量，同时使用深度学习方法对各类别文本信息进行句向量的提取，将其作为辅助信息约束变分自编码器和生成对抗网络的生成工作。

1　相关工作

1.1　VAE/GAN 模型

变分自编码器 (variational auto-encoder, VAE)[30]的作用是求解给定输入空间和特征空间之间的映射，使得输入特征的重建误差达到最小。但 VAE 只能得到一个平均的结果，这也是导致其生成图像质量较低，图像较模糊的原因。研究人员提出 VAE/GAN 模型[31]，将 VAE 与 GAN 进行融合，判别器的加入使得 VAE 产生的图像变得清晰。然而，常规 VAE/GAN 中编码器得到的隐变量并不完全符合期望样式。于是，出现了 VAE/GAN 的变体模型，其结构如图 1 所示，它改变了判别器结构，使其能更精细地鉴别输入图像的种类。

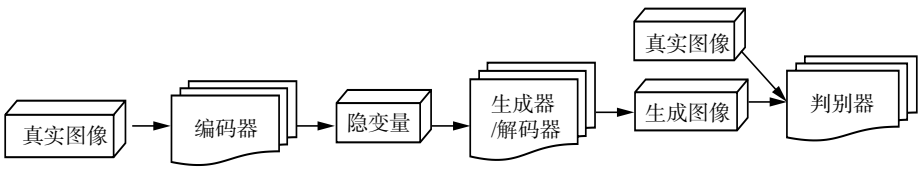


图 1　VAE/GAN 模型结构

Fig. 1　Structure of VAE/GAN model

真实图像作为编码器输入, 通过编码得到隐变量作为解码器的输入以生成图像, 此时 VAE 希望真实图像和生成图像之间的差异越小越好。判别器则需判别输入的图像属于真实数据分布p_{data}还是生成数据分布p_{G}。在 VAE 和 GAN 的共同作用下, 生成更加相似且清晰的图像。

1.2　StackGAN 模型

根据文字描述生成高质量图像的任务是计算机视觉领域的一个挑战。条件生成对抗网络 (conditional GAN, CGAN) 可以生成和文本比较相关的图像, 但是分辨率不够, 细节部分缺失严重, 不够生动具体。如果简单地增加更多的采样层来提高分辨率, 会导致模型不稳定或者生成一些奇形怪状的图像, 这种现象在分辨率提高时会更加严重。分阶段生成对抗网络 StackGAN 本质上是 2 个 CGAN 的堆叠[32], 其结构如图 2 所示, 采用分阶段的方式 (两个阶段) 生成高分辨率且置信度高的图像。

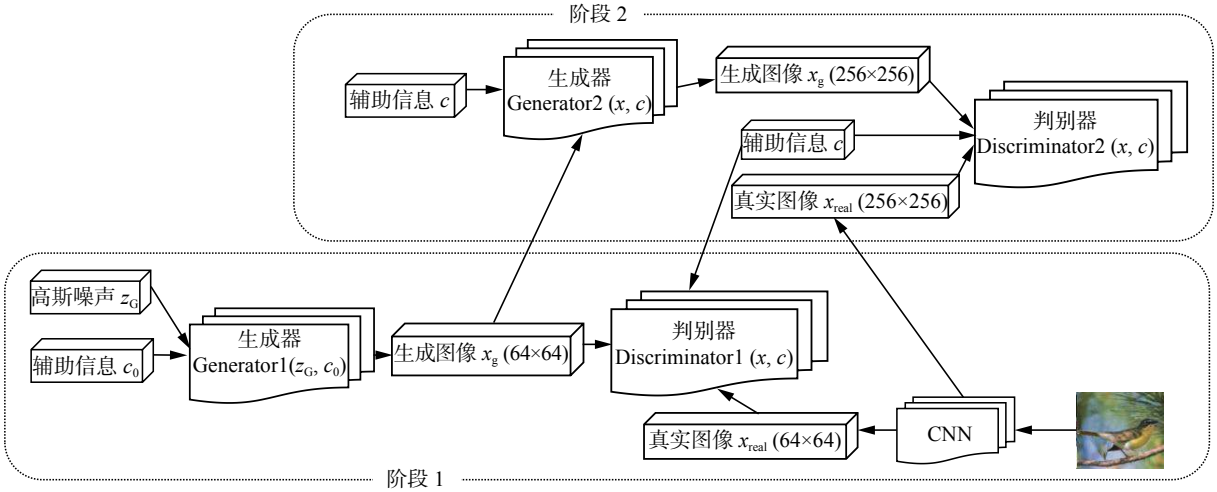


图 2　StackGAN 模型结构

Fig. 2　Structure of StackGAN model

阶段 1: 主要用于生成粗略的形状和颜色, 通过文本嵌入从中采样出服从$N(\mu_0(\varphi_t),\sum_0(\varphi_t))$分布的辅助信息$c_0$, 并随机采样高斯噪声$z_{\text{G}}$, 二者作为阶段 1 的输入, 用来训练生成器 Generator1 和判别器 Discriminator1, 分别对应如下目标函数:

$$\max L_{\text{D}_1}=E_{(x_{\text{real}},t)\sim p_{\text{data}}}[\log(D_1(x_{\text{real}},\varphi_t))]+E_{z_{\text{G}}\sim p_{z_{\text{G}}},t\sim p_{\text{data}}}[\log(1-D_1(G_1(z_{\text{G}},c_0),\varphi_t))]$$

$$\min L_{\text{G}_1}=E_{z_{\text{G}}\sim p_{z_{\text{G}}},t\sim p_{\text{data}}}[\log(1-D_1(G_1(z_{\text{G}},c_0),\varphi_t))]+\lambda D_{\text{KL}}(N(\mu_0(\varphi_t),\sum\nolimits_0(\varphi_t))\|N(0,I))$$

式中: G_1和D_1分别表示 Generator1 和 Discriminator1 的函数; 真实图像x_{real}和文本描述 t 源自于p_{data}; z_{G}表示服从高斯分布的噪声向量; λ为正则化参数。

阶段 2: 在 阶段 1 的基础上, 修正低分辨率图像的缺陷, 完善被忽略的文本信息细节, 生成高分辨率图像。以辅助语义向量以及阶段 1 的输出$s_0=G_1(z_{\text{G}},c_0)$作为输入来训练生成器 Generator2 和判别器 Discriminator2, 其目标函数分别为

$$\max L_{\text{D}_2}=E_{(x_{\text{real}},t)\sim p_{\text{data}}}[\log(D_2(x_{\text{real}},\varphi_t))]+E_{s_0\sim p_{G_1},t\sim p_{\text{data}}}[\log(1-D_2(G_2(s_0,c),\varphi_t))]$$

$$\min L_{\text{G}_2}=E_{s_0\sim p_{G_1},t\sim p_{\text{data}}}[\log(1-D_2(G_2(s_0,c),\varphi_t))]+\lambda D_{\text{KL}}(N(\mu(\varphi_t),\sum(\varphi_t))\|N(0,I))$$

式中: G_2和D_2分别表示 Generator2 和 Discriminator2 的函数; s_0源自于阶段 1 的生成数据分布p_{G_1}。

同时, 本模型引入条件增强来从独立高斯分布$N(\mu(\varphi_t),\sum(\varphi_t))$中随机采样产生额外的条件变量$\lambda$, 使得在给定较少文本–图像数据对时, 能够产生更多的训练样本, 增加生成样本的随机性。

1.3　零样本生成模型 f-CLSWGAN

基于生成对抗网络的零样本生成模型 f-CLSWGAN 由生成网络、判别网络和分类网络 3 部分构成。生成网络部分采用 WGAN 模型, 以各类别的属性信息$c(y)$和噪声z_{G}共同作为输入, 生成未知类图像x_{g}。属性信息$c(y)$、未知类图像x_{g}以及真实图像x_{real}作为判别网络的输入, 判别真伪后产生一个损失值用以优化生成网络。生成网络和判别网络二者相互对抗学习。同时, 生成的图像作为分类网络的输入进行分类, 同样产生一个损失值, 两部分共同指导生成网络进行优化, 最终生成足够接近真实图像的未知类图像。

2　模型改进

基于生成模型的零样本图像分类任务流程如图 3 所示。本文主要针对生成模型部分进行改进, 以进一步提高生成的未见类图像质量。

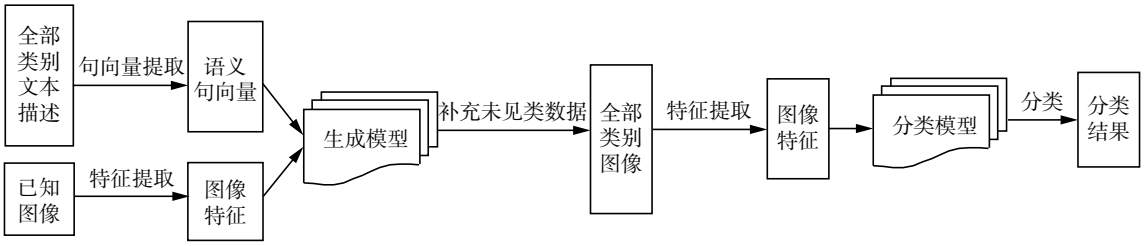


图 3　基于生成模型的零样本图像分类流程图

Fig. 3　Flow chart of zero-shot image classification based on generative model

2.1　VAE-StackGAN 模型

针对现有 VAE/GAN 模型训练不稳定，生成图像质量低的问题，本文基于 VAE/GAN 变体模型引入 StackGAN，得到 VAE-StackGAN 模型，如图 4 所示。VAE-StackGAN 模型中真实图像作为 VAE 编码器 Encoder 的输入，通过编码得到隐变量z_{V}。z_{V}作为 VAE 解码器 Decoder 的输入，得到重构图像x_{rec}。随机噪声z_{G}和辅助信息c_0作为 StackGAN 模型阶段 1 生成器 Generator1 的输入，得到生成图像x_{g}。VAE 和 StackGAN 两生成模型通过解码器和生成器共享参数进行连接。对于阶段 1 的判别器 Discriminator1，将真实图像x_{real}、x_{rec}、x_{g}和c_0作为输入，通过更精细地学习x_{rec}和x_{g}之间的细微差别，使得编码器得到更接近z_{G}的隐变量z_{V}。则 StackGAN 中阶段 1 的目标函数为

$$\begin{aligned}\min L_{\mathrm{G}_1} = {} & E_{z_{\mathrm{G}}\sim p_{z_{\mathrm{G}}},t\sim p_{\mathrm{data}}}[\log(1-D_1(G_1(z_{\mathrm{G}},c_0),\varphi_t))]+\\ & E_{z_{\mathrm{V}}\sim p_{(z_{\mathrm{V}}|x_{\mathrm{real}})},t\sim p_{\mathrm{data}}}[\log(1-D_1(G_1(\mathrm{Enc}(x_{\mathrm{real}}),c_0),\varphi_t))]+\\ & \lambda D_{\mathrm{KL}}(N(\mu_0(\varphi_t),\sum\nolimits_0(\varphi_t))\|N(0,I))\end{aligned}$$

$$\begin{aligned}\max L_{\mathrm{D}_2} = {} & E_{(x_{\mathrm{real}},t)\sim p_{\mathrm{data}}}[\log(D_2(x_{\mathrm{real}},\varphi_{\mathrm{t}}))]+\\ & E_{s_{\mathrm{G}}\sim p_{G_1},t\sim p_{\mathrm{data}}}[\log(1-D_2(G_2(s_{\mathrm{G}},c),\varphi_{\mathrm{t}}))]+\\ & E_{s_{\mathrm{V}}\sim p_{G_1},t\sim p_{\mathrm{data}}}[\log(1-D_2(G_2(s_{\mathrm{V}},c),\varphi_{\mathrm{t}}))]\end{aligned}$$

式中Enc(·)表示 VAE 编码器 Encoder。

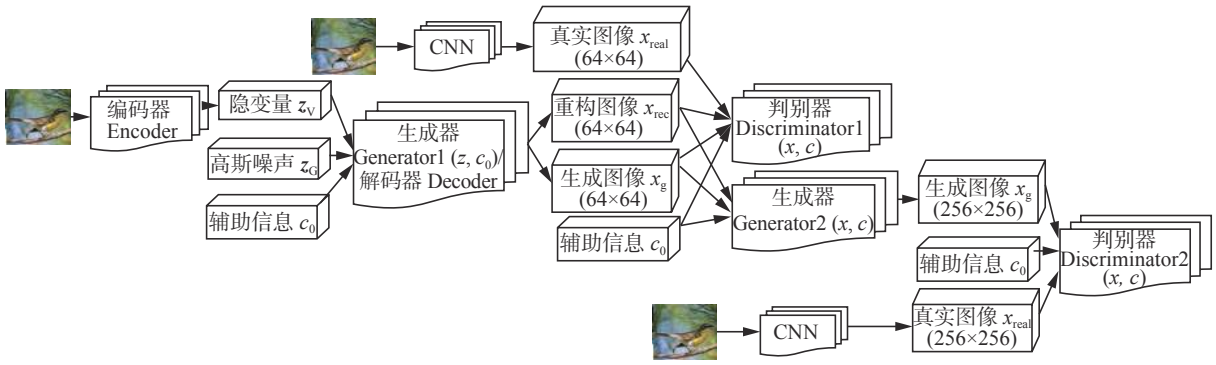


图 4　VAE-StackGAN 网络结构

Fig. 4　Network structure of VAE-StackGAN

VAE 和 StackGAN 中的阶段 1 为 VAE-StackGAN 模型的主要部分，生成包含基本轮廓和颜色的图像，而 StackGAN 中阶段 2 的生成器 Generator2 以阶段 1 生成的x_{rec}和x_{g}以及阶段 2 辅助语义信息c作为输入，在此基础上根据c对阶段 1 产生图像的细节部分进行细化，以生成更真实的图像x_{g}。将x_{g}和x_{real}以及c一同作为阶段 2 判别器 Discriminator2 的输入，经过判别得到一个分类结果，产生的损失值用来指导生成器优化。阶段 2 的目标函数为

$$\begin{aligned}\min L_{\mathrm{G}_2} = {} & E_{s_{\mathrm{G}}\sim p_{G_1},t\sim p_{\mathrm{data}}}[\log(1-D_2(G_2(s_{\mathrm{G}},c),\varphi_{\mathrm{t}}))]+\\ & E_{s_{\mathrm{V}}\sim p_{G_1},t\sim p_{\mathrm{data}}}[\log(1-D_2(G_2(s_{\mathrm{V}},c),\varphi_{\mathrm{t}}))]+\\ & \lambda D_{\mathrm{KL}}(N(\mu(\varphi_{\mathrm{t}}),\sum(\varphi_{\mathrm{t}}))\|N(0,I))\end{aligned}$$

$$\begin{aligned}\max L_{\mathrm{D}_2} = {} & E_{(x_{\mathrm{real}},t)\sim p_{\mathrm{data}}}[\log(D_2(x_{\mathrm{real}},\varphi_{\mathrm{t}}))]+\\ & E_{s_{\mathrm{G}}\sim p_{G_1},t\sim p_{\mathrm{data}}}[\log(1-D_2(G_2(s_{\mathrm{G}},c),\varphi_{\mathrm{t}}))]+\\ & E_{s_{\mathrm{V}}\sim p_{G_1},t\sim p_{\mathrm{data}}}[\log(1-D_2(G_2(s_{\mathrm{V}},c),\varphi_{\mathrm{t}}))]\end{aligned}$$

式中：$s_{\mathrm{G}}=G_1(z_{\mathrm{G}},c_0)$表示随机噪声通过生成器生成的图像；$s_{\mathrm{V}}=G_1(\mathrm{Enc}(x_{\mathrm{real}}),c_0)$表示通过解码器重构得到的图像。

VAE-StackGAN 模型的目标函数为

$$\min_{G_1,G_2}\max_{D_1,D_2} L_{\text{VAE - StackGAN}} = \min_{G_1,G_2}\max_{D_1,D_2} L_{\text{StackGAN}} + L_{\text{VAE}}$$

VAE-StackGAN 模型改变了判别器结构，使其可以更精细地判别出生成图像的类别，学习重构图像x_{rec}与生成图像x_{g}之间的差异，以优化生成器的参数，但编码器产生的隐变量z_{V}和生成器输入的随机噪声z_{G}仍存在差异。为此，本文在目标函数中增加正则项，使用 KL 散度帮助优化z_{V}和z_{G}之间的距离。此时，VAE-StackGAN 模型的目标函数为

$$\begin{aligned}\min_{G_1,G_2}\max_{D_1,D_2} L_{\text{VAE-StackGAN}} = {} & \min_{G_1,G_2}\max_{D_1,D_2} L_{\text{StackGAN}} + L_{\text{VAE}} +\\ & D_{\mathrm{KL}}(p(z_{\mathrm{V}}|x_{\mathrm{real}})\|p(z_{\mathrm{G}}))\end{aligned}$$

式中: L_{VAE}为 VAE 模型的目标函数; L_{StackGAN}为 StackGAN 模型的目标函数; $p(z_G)$为 StackGAN 的噪声分布。假设$p(z_G)$为服从高斯分布的随机噪声, 采用 KL 散度计算两分布之间的距离, 目的是使z_{V}不断向期望的形式更新。

2.2　零样本图像分类方法

在 VAE-StackGAN 模型的基础上构建本文零样本图像分类方法, 包括句向量的提取、图像的生成以及分类器的训练 3 部分。

首先是句向量的提取。本文采用循环卷积神经网络对各类别的文本信息进行无监督学习, 提取各类别的语义句向量, 并作为辅助信息输入到生成模型中。每轮训练选取文本中不同的句向量以保证后续生成图像的多样性。

其次是图像的生成。零样本生成模型 f-CLS-WGAN 采用条件生成对抗网络, 使用基于属性的语义辅助信息结合 Wasserstein 距离计算方法作为生成模型, 生成未见类数据特征。但仅采用生成对抗网络得到的效果并不理想, 因此本文采用 VAE-StackGAN 作为零样本生成模型的主体网络。并与语义句向量进行结合, 得到零样本生成模型 stc-CLS-VAEStackGAN。

将语义句向量代替属性信息。编码器对可见类图像进行编码, 生成器根据语义信息生成未见类图像, 判别器根据相应类别的句向量对输入图像进行判别。同时保留 f-CLSWGAN 中的分类器部分, 生成的图像也作为分类器的输入, 进行类别的判断, 此处分类器设置的目的是约束生成器生成具有分类特征的图像, 辅助生成模型进行优化。此时基于语义句向量的零样本生成模型的目标函数为

$$\min_{G_1,G_2}\max_{D_1,D_2} L_{\mathrm{VAE\text{-}StackGAN}}+\beta L_{\mathrm{CLS}}$$

最后是分类器的训练。随着对分类精度要求的不断提高, 模型深度越来越深, 模型复杂度也越来越高。但模型过于庞大会导致响应速度慢、内存不足等问题, 因此本文使用深度可分离卷积 MobileNet 网络, 结合 SoftMax 回归模型进行分类。

在生成模型部分生成了未见类图像, 填充了数据集, 因此现有数据集包括可见类与未见类图像。生成模型的加入将零样本图像分类任务转变成传统图像分类任务, 使用多类别分类器进行训练, 训练好的分类器即可对测试集图像进行分类。

3　实验结果与讨论

3.1　数据集与实验配置

本文采用零样本学习领域常用的 CUB (caltech-uCSD-Birds-200-2011)[33] 和 AwA (animals with attributes)[34] 数据集进行模型训练及测试, 并采用两个数据集默认的可见类与未见类划分方式。CUB 数据集共包含 200 个鸟类类别, 共 11 788 张图片, 类别间差异较小, 属于精细分类, 可见类包含 150 个类别, 其中的 100 个类别为训练集, 50 个类别为验证集; 未见类包含 50 个类别, 为测试集。AwA 数据集中共包含 50 个动物类别, 共 30 475 张图片, 可见类包含 40 个类别, 其中的 27 个类别为训练集, 13 个类别为验证集; 未见类包含 10 个类别, 为测试集。可见类与未见类样本不重叠。

3.2　评价指标

GAN 网络生成图像的任务中, 评价模型表现的一项重要指标是初始评分 (inception score, IS)[35], 可以用来评价模型生成图像的清晰度和多样性。对于一个清晰的图像, 它属于某一类的概率应该非常大, 而属于其他类的概率应该很小, 同时如果一个模型能生成足够多样的图像, 那么它生成的图像在各个类别中的分布应该是平均的。IS 计算公式为

$$\mathrm{IS}(G)=\exp(E_{x\sim p_G}D_{\mathrm{KL}}(p(y|x)\|p(y)))$$

$$D_{\mathrm{KL}}(P\|Q)=\sum_i P(i)\log\frac{P(i)}{Q(i)}$$

式中: $x\sim p_G$表示生成器生成图像; $p(y|x)$表示该图像属于各个类别的概率分布; $p(y)$表示生成器生成的全部图像在所有类别上的边缘分布。

对于本文方法及对比方法的零样本图像分类结果, 采用类别的平均分类准确率进行评价。该指标是评价模型有效性的较好指标, 能更好地反映数据集类内数据量不均衡时分类模型的识别效果, 削弱实验的随机性。平均分类准确率公式为

$$f(acc)=\frac{1}{|Y|}\sum_{i=0}^{|y|}\left(\frac{N_{\mathrm{C}}^{Y-i}}{N_{\mathrm{T}}^{Y-i}}\right)$$

式中: Y为总体类别个数; N_{T}为当前类别样本数量; N_{C}为当前类别分类正确样本数量。

3.3　本文生成模型定性与定量分析

本文通过在 VAE/GAN 变体模型中引入 StackGAN 网络, 得到 VAE-StackGAN 生成模型, 用于生成未见类图像。

基于 CUB 数据集, 对比 VAE-StackGAN、VAE/GAN 和 StackGAN 模型的生成图像效果, 并计算 IS 指标, 计算结果如表 1 所示。从表 1 中可见, VAE-StackGAN 模型的 IS 值较 StackGAN 和 VAE/GAN 分别提升 0.26 和 0.33。该指标值在图像质量方面与人类的感知高度相关, 更高的分数意味着更好的图像质量。因此, VAE-StackGAN 模型明显优于 StackGAN 和 VAE/GAN 模型。

表 1　不同方法在 CUB 数据集上的 IS 指标

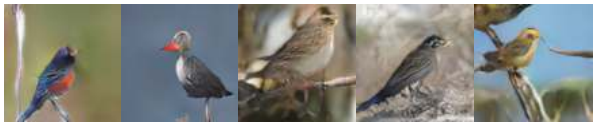
Table 1　Inception score of different methods on CUB dataset

方法	IS指标
VAE/GAN	3.63 ± 0.06
StackGAN	3.70 ± 0.04
VAE-StackGAN	3.96 ± 0.03

图 5 为不同方法的生成图像对比，可以更直观地看出，本文生成模型 VAE-StackGAN 具有更好的性能，相比于原始模型更加注重目标颜色的细节，从而使得生成的目标图像更加逼真，更加符合其相应类别特征。因此，整体来看，本文模型生成图像的效果优于 StackGAN 模型。

(a) 真实图像

(b) StackGAN 生成图



(c) VAE-stackGAN 生成图

图 5　生成图像效果对比

Fig. 5　Comparison of generate figures

3.4　实验结果评估

基于 CUB 和 AwA1 数据集，将本文方法 stc-CLS-VAEStackGAN 用于零样本图像分类，平均分类准确率随迭代次数变化的曲线如图 6 所示。随着迭代次数的增加，平均分类准确率逐渐提升并且趋于稳定，尽管个别值在小范围内有所波动，但处于正常波动范围内，证明了本模型具有较好的收敛性。

将 stc-CLS-VAEStackGAN 与现有其他生成模型进行零样本图像分类对比，使用类别平均分类准确率作为评价指标。stc-CLS-VAEStackGAN 为基于生成模型的方法，因此同样选取基于生成模型且具有代表性、较新提出的模型作为对比方法，包括 f-CLSWGAN[18]、FD-fGAN-Attention[36]、Zero-VAE-GAN[29]、ZSML[23]、SPGAN[24] 和 SAGAN[27]。需说明的是，以上对比方法的相应文献中所采用的数据集以及训练集、测试集的划分均与本文相同，因此这些方法的平均分类准确率指标值采用相应文献所提供的数值。

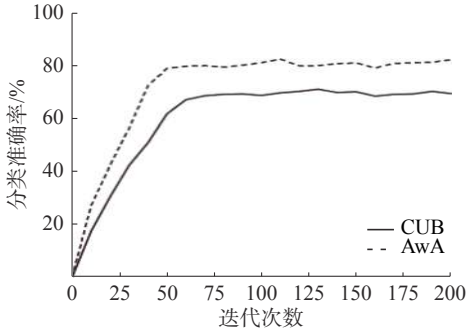


图 6　平均分类准确率随迭代次数的变化

Fig. 6　Variation of the average classification accuracy with the number of iterations

表 2 为不同方法在 CUB 和 AwA 两个数据集上的平均分类准确率指标值。从表中可以看出，本文生成模型 stc-CLS-VAEStackGAN 在 AwA 数据集上的平均分类准确率明显高于所有对比算法，在 CUB 数据集上的平均分类准确率虽略低于 ZSML，但均明显高于其他对比算法。stc-CLS-VAEStackGAN 基于 VAE/GAN 的变体模型，通过两种互补的生成模型 VAE 和 GAN 分别捕获不同的数据分布，弥补了各自的缺点，其中变体判别器可以更加精细地学习重构图像与生成图像之间差异，提高生成图像效果。将 StackGAN 引入其中，采用分阶段方法细化图像生成过程，进一步提高了图像生成质量。同时采用句向量代替属性信息，使得语义描述更加准确，生成图像更加多样化。对比实验结果验证了本文方法的有效性。

表 2　不同方法的平均分类准确率对比

Table 2　Comparison of the average classification accuracy of different methods　%

方法	CUB	AwA
f-CLSWGAN	57.3	68.2
FD-fGAN-Attention	58.5	71.6
Zero-VAE-GAN	54.8	71.4
SPGAN	58.6	71.5
ZSML	69.6	73.5
SAGAN	58.1	71.6
stc-CLS-VAEStackGAN	68.8	79.2

3.5　算法鲁棒性测试

为测试本文图像生成模型 VAE-StackGAN 的鲁棒性，将数据集中各类别的文本信息进行加噪处理，包括标签关键字删除、替换、交换顺序等，并根据处理后的文本信息提取类别句向量作为辅

助信息, 分别采用 VAE-StackGAN 模型和原始模型 StackGAN 进行图像生成。表 3 显示了两种模型生成相应类别图像的 IS 评价指标值。对文本信息进行上述加噪处理后, 可能造成一些关键信息的缺失或异常, 因此从表 3 中可以看出, 文本信息加噪后, 两种模型生成图像的 IS 评价指标均有所下降, 但 VAE-StackGAN 模型的 IS 指标下降幅度小于原始模型 StackGAN, 验证了 VAE-StackGAN 模型的鲁棒性有所提高。

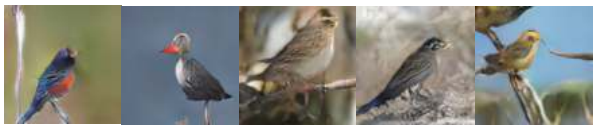
表 3　生成模型鲁棒性测试

Table 3　Robustness test of generative models

方法	文本信息	IS指标
VAE-StackGAN	不加噪	3.96 ± 0.03
	加噪	3.87 ± 0.05
StackGAN	不加噪	3.70 ± 0.04
	加噪	3.59 ± 0.03

图 7 为文本信息加噪前后, VAE-StackGAN 模型的生成图像。整体来看, 加入噪声后生成的图像在细节方面可能会有缺失或者存在多余的噪声背景, 但仍具有相应类别的特征, 整体效果并没有发生巨大的差距。

(a) 真实图像



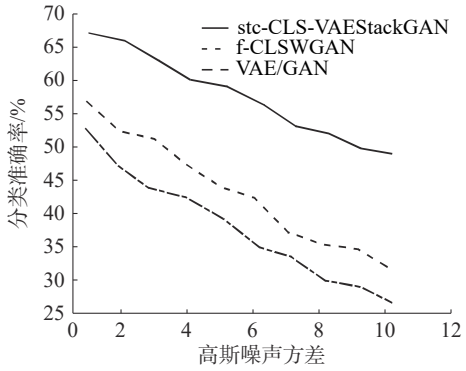
(b) VAE-stackGAN 生成图

(c) VAE-stackGAN 加噪后生成图

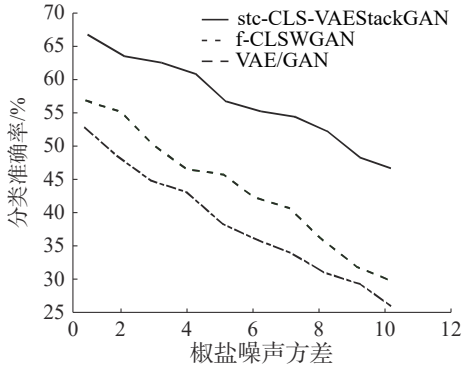
图 7　生成模型鲁棒性测试

Fig. 7　Robustness test of generative models

为测试本文零样本图像分类方法 stc-CLS-VAEStackGAN 的鲁棒性, 对实验中扩充后的图像数据集样本进行加噪声处理, 加入高斯噪声用于模拟实际应用中较为常见的由不良照明引起的图像模糊问题; 加入椒盐噪声, 随机改变一些像素值以产生黑白相间的亮暗点, 用于模拟实际应用中较常见的由图像切割引起的噪点问题, 然后进行图像分类, 评价分类方法的准确率。stc-CLS-VAEStackGAN 是在 f-CLSWGAN 和 VAE/GAN 的基础上改进得到的, 因此图 8 对比了这 3 种方法在图像加噪后的分类准确率。由图 8 可见, 在任意噪声强度处, stc-CLS-VAEStackGAN 的分类准确率均明显高于其他两种方法; 随着加入高斯噪声的方差逐渐增大, 以及椒盐噪声的增强, 3 种方法的图像分类准确率均有所下降, 但 CLS-VAESStackGAN 的准确率下降程度较为平缓。上述结果验证了, 本文 stc-CLS-VAEStackGAN 方法不仅提高了零样本图像分类准确率, 而且具有较好的鲁棒性。



(a) 高斯噪声对分类准确率的影响



(b) 椒盐噪声对分类准确率的影响

图 8　零样本图像分类方法鲁棒性的测试

Fig. 8　Robustness test of zero-shot image classification methods

4　结束语

本文提出一种融合 VAE 与 StackGAN 的零样本图像分类方法, 通过生成未见类的图像填充图像数据集, 将零样本图像分类任务转变为传统的基于监督学习的图像分类任务。基于 VAE/GAN 的变体模型引入 StackGAN, 采用分阶段方法细化图像生成过程, 提高图像生成质量; 并通过提取类别的句向量信息代替其属性信息, 增加语义描述的准确性, 使生成图像更加多样化, 进而提高零样本图像分类的准确性。在现有公用数据集上进行对比实验, 实验结果验证了本文方法的有效性。

参考文献:

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60: 84–90.

[2] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436–444.

[3] BIEDERMAN I. Recognition-by-components: a theory of human image understanding[J]. Psychological review, 1987, 94(2): 115–147.

[4] PALATUCCI M, POMERLEAU D, HINTON G E, et al. Zero-shot Learning with semantic output codes[C]// Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Vancouver: NIPS, 2009: 1410−1418.

[5] 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.

[6] ROHRBACH M, STARK M, SCHIELE B. Evaluating knowledge transfer and zero-shot learning in a large-scale setting[C]// Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011: 1641−1648.

[7] HABIBIAN A, MENSINK T, SNOEK C G M. Video2vec embeddings recognize events when examples are scarce[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(10): 2089–2103.

[8] 冀中, 汪浩然, 于云龙, 等. 零样本图像分类综述: 十年进展 [J]. 中国科学:信息科学, 2019, 49(10): 1299–1320.
JI Zhong, WANG Haoran, YU Yunlong, et al. A decadal survey of zero-shot image classification[J]. Scientia sinica (informationis), 2019, 49(10): 1299–1320.

[9] 张鲁宁, 左信, 刘建伟. 零样本学习研究进展 [J]. 自动化学报, 2020, 46(1): 1–23.
ZHANG Luning, ZUO Xin, LIU Jianwei. Research and development on zero-shot learning[J]. Acta automatica sinica, 2020, 46(1): 1–23.

[10] 冀中, 孙涛, 于云龙. 一种基于直推判别字典学习的零样本分类方法 [J]. 软件学报, 2017, 28(11): 2961–2970.
JI Zhong, SUN Tao, YU Yunlong. Transductive discriminative dictionary learning approach for zero-shot classification[J]. Journal of software, 2017, 28(11): 2961–2970.

[11] WANG Xuesong, CHEN Chen, CHENG Yuhu. Zero-shot learning by exploiting class-related and attribute-related prior knowledge[J]. IET computer vision, 2016, 10(6): 483–492.

[12] 赵鹏, 汪纯燕, 张思颖, 等. 一种基于融合重构的子空间学习的零样本图像分类方法 [J]. 计算机学报, 2021, 44(2): 409–421.
ZHAO Peng, WANG Chunyan, ZHANG Siying, et al. A zero-shot image classification method based on subspace learning with the fusion of reconstruction[J]. Chinese journal of computers, 2021, 44(2): 409–421.

[13] DU Yujiao, XIAO Bo, XU Wenchao, et al. Destination prediction for sharing-bikes' trips[C]//2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC). Guiyang: IEEE, 2018: 198−202.

[14] 陈祥凤, 陈雯柏. 度量学习改进语义自编码零样本分类算法 [J]. 北京邮电大学学报, 2018, 41(4): 69–75.
CHEN Xiangfeng, CHEN Wenbai. Improving semantic autoencoder zero-shot classification algorithm by metric learning[J]. Journal of Beijing university of posts and telecommunications, 2018, 41(4): 69–75.

[15] 吴晨, 袁昱纬, 王宏伟, 等. 基于词向量融合的遥感场景零样本分类算法 [J]. 计算机科学, 2019, 46(12): 286–291.
WU Chen, YUAN Yuwei, WANG Hongwei, et al. Word vectors fusion based remote sensing scenes zero-shot classification algorithm[J]. Computer science, 2019, 46(12): 286–291.

[16] 冀中, 谢于中, 庞彦伟. 基于典型相关分析和距离度量学习的零样本学习 [J]. 天津大学学报(自然科学与工程技术版), 2017, 50(8): 813–820.
JI Zhong, XIE Yuzhong, PANG Yanwei. Zero-shot learning based on canonical correlation analysis and distance metric learning[J]. Journal of Tianjin University (science and technology edition), 2017, 50(8): 813–820.

[17] XIAN Yongqin, LAMPERT C H, SCHIELE B, et al. Zero-shot learning-A comprehensive evaluation of the good, the bad and the ugly[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 41(9): 2251–2265.

[18] 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.

[19] SARIYILDIZ M B, CINBIS R G. Gradient matching generative networks for zero-shot learning[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 2163−2173.

[20] MANDAL D, NARAYAN S, DWIVEDI S K, et al. Out-of-distribution detection for generalized zero-shot action recognition[C]//2019 IEEE/CVF Conference on Com-

puter Vision and Pattern Recognition (CVPR). Long Beach : IEEE, 2019: 9977−9985.

[21] XIAN Yongqin, SHARMA S, SCHIELE B, et al. F-VAEGAN-D2: a feature generating framework for any-shot learning[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach : IEEE, 2019: 10267−10276.

[22] KIM H, LEE J, BYUN H. Unseen image generating domain-free networks for generalized zero-shot learning[J]. Neurocomputing, 2020, 411: 67–77.

[23] VERMA V K, BRAHMA D, RAI P. Meta-learning for generalized zero-shot learning[J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(4): 6062–6069.

[24] MA Yuanbo, XU Xing, SHEN Fumin, et al. Similarity preserving feature generating networks for zero-shot learning[J]. Neurocomputing, 2020, 406: 333–342.

[25] LIU Huan, YAO Lina, ZHENG Qinghua, et al. Dual-stream generative adversarial networks for distributionally robust zero-shot learning[J]. Information sciences, 2020, 519: 407–422.

[26] LIU Jinlu, ZHANG Zhaocheng, YANG Gang. Cross-class generative network for zero-shot learning[J]. Information sciences, 2021, 555: 147–163.

[27] TANG C, HE Z, LI Y, ET Al. Zero-shot learning via structure-aligned generative adversarial network[J]. IEEE transactions on neural networks and learning systems, 2021(99): 1–14.

[28] LI Zhiqun, CHEN Qiong, LIU Qingfa. Augmented semantic feature based generative network for generalized zero-shot learning[J]. Neural networks:the official journal of the international neural network society, 2021, 143: 1–11.

[29] GAO RUI, HOU XINGSONG, QIN JIE, et al. Zero-VAE-GAN: generating unseen features for generalized and transductive zero-shot learning[J]. IEEE transactions on image processing, 2020, 29: 3665–3680.

[30] KINGMA D P, WELLING M. Auto-Encoding Variational Bayes[EB/OL]. (2014−05−01)[2022−03−10]https://arxiv.org/abs/1312.6114v2.

[31] LARSEN A B L, SØNDERBY S K, LAROCHELLE H, et al. Autoencoding beyond pixels using a learned similarity metric[EB/OL]. (2016−02−10)[2022−03−10] https://arxiv.org/abs/1512.09300.

[32] ZHANG Han, XU Tao, LI Hongsheng, et al. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks[EB/OL]. (2017−08−05)[2022−03−10]https://arxiv.org/abs/1612.03242v2.

[33] Wah C, Branson S, Welinder P, et al. The Caltech-UCSD Birds-200-2011 Dataset[J]. California institute of technology, 2011.

[34] LAMPERT C H, NICKISCH H, HARMELING S. Attribute-based classification for zero-shot visual object categorization[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 36(3): 453–465.

[35] CHE TONG, LI YANRAN, JACOB A P, et al. Mode regularized generative adversarial networks[EB/OL]. (2017−03−02)[2022−03−10]https://arxiv.org/abs/1612.02136 .

[36] 魏宏喜, 张越. 基于生成对抗网络的零样本图像分类[J]. 北京航空航天大学学报, 2019, 45(12): 2345–2350. WEI Hongxi, ZHANG Yue. Zero-shot image classification based on generative adversarial network[J]. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(12): 2345–2350.

作者简介：

张冀，副教授，博士，主要研究方向为计算机测控、故障诊断、信息融合、图像处理、深度学习。出版规划教材 2 部。发表学术论文 20 余篇。

曹艺，硕士研究生，主要研究方向为计算机视觉。

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。主持河北省自然科学基金青年基金项目 1 项，参与国家自然科学基金面上项目 2 项、横向科研项目多项。发表学术论文 10 余篇。