Journal of Computer Applications
计算机应用, 2020, 40(4): 1016 - 1022
ISSN 1001-9081
CODEN JYIIDU
2020-04-10
http://www.joca.cn

文章编号:1001-9081(2020)04-1016-07
DOI:10.11772/j.issn.1001-9081.2019081475

基于视觉误差与语义属性的零样本图像分类

徐　戈[1],肖永强[2,3,4],汪　涛[1,2],陈开志[2,3,4],廖祥文[2,3,4*],吴运兵[2,3,4]

(1. 闽江学院 计算机与控制工程学院,福州 350108;　2. 福州大学 数学与计算机科学学院,福州 350116;
3. 福建省网络计算与智能信息处理重点实验室(福州大学),福州 350116;
4. 数字福建金融大数据研究所,福州 350116)

(*通信作者电子邮箱liaoxw@fzu.edu.cn)

摘　要:在图像分类的实际应用过程中,部分类别可能完全没有带标签的训练数据。零样本学习(ZSL)的目的是将带标签类别的图像特征等知识迁移到无标签的类别上,实现无标签类别的正确分类。现有方法在测试时无法显式地区分输入图像属于已知类还是未知类,很大程度上导致未知类在传统设定下的ZSL和广义设定下的ZSL(GZSL)上的预测效果相差甚远。为此,提出一种融合视觉误差与属性语义信息的方法来缓解零样本图像分类中的预测偏置问题。首先,设计一种半监督学习方式的生成对抗网络架构来获取视觉误差信息,由此预测图像是否属于已知类;然后,提出融合属性语义信息的零样本图像分类网络来实现零样本图像分类;最后,测试融合视觉误差与属性语义的零样本图像分类方法在数据集AwA2和CUB上的效果。实验结果表明,与对比模型相比,所提方法有效缓解了预测偏置问题,其调和指标H在AwA2(Animal with Attributes)上提升了31.7个百分点,在CUB(Caltech-UCSD-Birds-200-2011)上提升了8.7个百分点。

关键词:零样本学习;图像分类;生成对抗网络;视觉误差;属性语义

中图分类号:TP391　　**文献标志码:**A

Zero-shot image classification based on visual error and semantic attributes

XU Ge[1], XIAO Yongqiang[2,3,4], WANG Tao[1,2], CHEN Kaizhi[2,3,4], LIAO Xiangwen[2,3,4*], WU Yunbing[2,3,4]

(1. *College of Computer and Control Engineering, Minjiang University, Fuzhou Fujian* 350108, *China*;
2. *College of Mathematics and Computer Science, Fuzhou University, Fuzhou Fujian* 350116, *China*;
3. *Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing (Fuzhou University), Fuzhou Fujian* 350116, *China*;
4. *Digital Fujian Financial Big Data Institute, Fuzhou Fujian* 350116, *China*)

Abstract: In the practical applications of image classification, some categories may have no labeled training data at all. The purpose of Zero-Shot Learning (ZSL) is to transfer knowledge such as image features of labeled categories to unlabeled categories and to correctly classify the unlabeled categories. However, the existing state-of-the-art methods cannot explicitly distinguish the input image belonging to the known categories or unknown categories, which leads to a large performance gap for unlabeled categories between the traditional ZSL prediction and the Generalized ZSL (GZSL) prediction. Therefore, a method of fusing of visual error and semantic attributes was proposed to alleviate the prediction bias problem in zero-shot image classification. Firstly, a semi-supervised learning based generative adversarial network framework was designed to obtain visual error information, so as to predict whether the image belongs to the known categories. Then, a zero-shot image classification network combining semantic attributes was proposed to achieve zero-shot image classification. Finally, the performance of zero-shot image classification algorithm combining visual error and semantic attributes was tested on AwA2 (Animal with Attributes) and CUB (Caltech-UCSD-Birds-200-2011) datasets. The experimental results show that, compared to the baseline models, the proposed method can effectively alleviate the prediction bias problem, and has the harmonic index H increased by 31.7 percentage points on AwA2 dataset and 8.7 percentage points on CUB dataset.

Key words: Zero-Shot Learning (ZSL); image classification; generative adversarial network; visual error; semantic attribute

收稿日期:2019-09-03;**修回日期:**2019-10-23;**录用日期:**2019-11-01。　**基金项目:**国家自然科学基金资助项目(61772135,U1605251,61703195);中国科学院网络数据科学与技术重点实验室开放课题基金资助项目(CASNDST201708,CASNDST201606);模式识别国家重点实验室开放课题基金资助项目(201900041);福建省自然科学基金面上项目(2017J01755);赛尔网络下一代互联网技术创新项目(NGII20160501)。

作者简介:徐戈(1978—),男,福建福州人,副教授,博士,主要研究方向:人工智能、自然语言处理;　肖永强(1994—),男,福建龙岩人,硕士研究生,主要研究方向:多模态、人工智能;　汪涛(1987—),男,福建福州人,讲师,博士,主要研究方向:场景理解、目标检测与分割;　陈开志(1983—),男,福建福州人,讲师,博士,主要研究方向:自然语言处理、图像识别、深度学习算法;　廖祥文(1980—),男,福建福州人,教授,博士,主要研究方向:观点挖掘、情感分析、自然语言处理;　吴运兵(1976—),男,福建福州人,副教授,硕士,主要研究方向:知识表示与知识发现。

0　引言

在进行图像分类的过程中,图像类别数往往较多。如果每次都采用人工去标注每个类别标签,工作量会非常巨大。例如,在文字识别过程中,单字类别数量成千上万,对每一个单字类别都进行标注几乎是一件难以完成的工作。针对这样的情况,如何利用机器学习训练分类器,在不标注新类别的情况下仍然可以对其分类值得研究。显然传统的分类器并不能实现这样的目的,所以研究者提出零样本学习(Zero-Shot Learning, ZSL)概念[1]。

零样本图像分类思想属于迁移学习的范畴,通过训练已有的视觉类别信息然后迁移到新图像类别,并实现对其分类。人类能够通过学习将已见过的视觉信息与语义信息建立联系,并通过这样的联系去判别新图像类别,使得我们具备识别未见过的图像类别的能力。在零样本学习图像分类中,未知的图像类别并没有已知的源图像训练样本,但可以通过对已知图像类别相关知识的学习,然后迁移到未知的图像类别当中,从而对未知的图像类别样本进行结果预测。属性作为中层语义特征,可以建立跨类别、多层级关联。属性语义描述通常由人工标注得到,因而能够有效反映出类别之间的关系以及表示每一个类别。对零样本图像分类研究有助于解决目标类别数据标签缺失问题、促进对人类智能推理的深入研究,具有重要的理论研究意义和实际应用价值。

根据测试时的图像类别不同,零样本学习通常分为两种:传统的零样本学习(ZSL)和广义零样本学习(Generalized ZSL, GZSL)。其中:传统零样本学习只测试模型在未知类上的分类正确率;而广义零样本学习的测试类别既包括已知类,也包括未知类。对比方法在提升未知类上的分类性能的同时影响到了已知类上的分类性能,在已知类和未知类之间存在较严重的预测偏置。受此启发,本文提出一种基于视觉误差预测的方法来判别输入图像是否属于已知类,并根据其结果将图像嵌入不同的类别语义空间进行预测。

1　相关工作

在早期的方法中,Lampert 等[2-3]提出一种直接属性预测模型(Direct Attribute Prediction, DAP),该学习方法先训练已知样本的视觉特征,以及相对应类别的属性特征,这样可以学习得到一个分类器预测属性,从而得到测试样本每一个未见过类别的后验概率,最后使用最大后验估计法,对未知类测试图像进行结果预测。Romera-Paredes 等[4]提出了一个较为通用的神经网络框架,该网络将特性、属性和类之间的关系建模为两个线性层网络,其中顶层的权重不是由环境来学习,而是由环境来给出。通过将其作为域自适应方法,进一步给出了这种方法的泛化误差的学习界限。第一线性网络层建立图像特征和属性特征之间的联系,第二线性网络层建立属性特征和类别标签之间的联系。

针对各个属性特征的某些语义关系,一些文献对其进行了不同程度的研究。比如,Liu 等[5]提出了一个统一的框架进行属性的学习,属性关系和属性分类共同提高属性预测的能力。具体来说,将属性关系学习和属性分类器设计成一个共同的目标函数,通过它不仅可以预测属性,还可以自动发现数据中属性之间的关系。巩萍等[6]提出了一种基于属性关系图正则化特征选择的零样本图像分类方法,通过训练样本中类别和属性矩阵来计算属性之间的正负相关性,该属性相关性直接预测属性。

也有一些文献提出将样本的图像特征和类别的语义特征同时映射到一个公共空间的方法。比如,文献[7]提出了一种结构化联合嵌入(Structured Joint Embedding, SJE)方法,利用映射空间矩阵将图像特征和语义特征嵌入到一个公共的特征空间,并基于公共特征空间计算各模态特征内积和,内积和最大的即为预测可能性最高的类别。Xian 等[8]对 SJE 进行改进,提出了一种隐藏嵌入(Latent Embedding, Lat Em)模型,在零样本分类中学习图像和类别嵌入之间的兼容性,通过结合深度变量来增强双线性兼容性模型,使用基于得分排序的目标函数训练模型,该目标函数惩罚给定图像的真实类别的错误得分排序。Akata 等[9]提出的联合嵌入框架将多个文本语义部分映射到一个公共空间,具体操作是先将图像划分成多个块,然后为各块获得相应的文本特征,最后用联合嵌入模型将它们映射到公共的特征空间当中。

近年来利用深度学习充分融合属性语义特征与图像特征进行分类,不需要考察每一类之间的距离度量,可以直接端到端训练分类。文献[10]设计了深度嵌入模型,选择合适的嵌入方式,对多个语义模态(属性和文本)融合后端到端学习,体现了一种自然机制,速度和正确率都有所改观。文献[11]设计出来的网络使用的是一种比较的思想,通过比较来对新类别进行分类,在性能上比文献[10]方法略有提升。文献[12]设计的神经网络是生成对抗的方式,利用均匀分布或高斯分布噪声对文本与属性特征产生新的图像特征(包含已知的和未知的),从而对图像特征进行分类。文献[13]设计了 3 个网络,分别是 FNet(提取图像特征)、ZNet(定位感兴趣区域并放大)和 ENet(特征映射),并联合训练 3 个网络,其中 ZNet 是参考文献[14]方法利用重复注意力机制来定位。

本文采用深度学习方法重新设计神经网络,充分融合属性语义特征与图像特征进行端到端训练。训练包括两个步骤:1)训练半监督学习方式的生成对抗网络,提取图像视觉误差信息;2)训练零样本图像分类,使用属性语义信息指导图像分类,能够识别未知类。

本文的主要工作如下:

1)提出一种半监督学习方式的生成对抗网络,采用编码器-解码器-编码器(encoder-decoder-encoder)架构,得到已见过类别和未见过类别在网络深层次抽象特征的误差值。该误差值的作用在于区分已知类和未知类,防止在 GZSL 下预测发生偏置问题。

2)提出一种零样本图像分类网络,采用图像特征与属性语义特征完全拼接(feature concatenation),得到的模型既能在传统设定下又能在广义设定下获取良好的分类表现。

3)通过视觉误差参数设定阈值的选择,可根据实际需要调整,缩小零样本图像分类在广义设定下预测类别的搜索空间,具体可分为两段和三段视觉误差。

2　模型方法

2.1　问题描述

本问题描述分成两部分:一部分是零样本学习,主要描述如何融合属性语义实现识别目标数据,包括具体识别已知类和未知类;另一部分是视觉异常识别,主要描述如何通过训练已知类来预测辨别未来数据中已知类和未知类,以下使用图像数据为例进行描述。

如图 1 所示,假设有一个带标签的源数据集 $D^s=\{(x_i^s,y_i^s)\}_{i=1}^{N_s}$,每张图片 x_i^s 与相应的标签 y_i^s 对应,其中 $y_i^s\in y^s=$

$\{(y_i)\}_{i=1}^{S}$，S 表示源类中类的个数；目标数据集 $D^t=\{(x_i^t,y_i^t)\}_{i=1}^{N_t}$，每张图片 x_i^t 与相应的标签 y_i^t 对应，其中 $y_i^t\in y^t=\{(y_i)\}_{i=1}^{T}$，$T$ 表示目标类中类的个数。在训练过程中不使用标签 y_i^t，该标签在测试时用于评估。

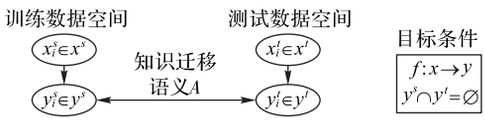

图1 零样本学习

Fig. 1 Zero-shot learning

用于类别知识迁移的描述包括训练类别数据属性语义描述 A_{train}，以及测试类别数据属性语义描述 A_{test}。这里的每个类别 $y_i\in Y$ 都可以表示为一个语义特征向量 $\boldsymbol{a}_i\in A$ 的形式表示，比如“黑白色”“有条纹”“有四条腿”“食肉的”“有羽毛”“有尾巴”等，属性语义特征向量每个维度都代表着一种高级的属性描述，当这个类别具备某些属性时，则在其对应的维度上会被设置成非零值（具体可参看3.1节中的数据集）。对于全集类别，每个类别属性语义向量维度都是一定的，一般经过多个人工标注后就可以得到该类别准确的属性语义描述信息。

在ZSL中，希望运用已知类数据 x_i^s 和 y_i^s 来训练模型，然后对训练好的模型除了可以识别已知类还能识别未知类，即具备预测 x_i^t 得到 y_i^t 的能力。外部知识迁移信息需要被模型知道才能得到希望的目标，这里的类别描述 A_{train} 和 A_{test} 建立起 y_i^t 中已知类和未知类的联系。

如图2所示，训练数据为已知类，而测试数据既有已知类也有未知类，异常识别问题是希望通过训练已知类去识别已知类和未知类。可以将已知类当作正常样本，而未知类当作异常样本，异常识别问题被理解为二分类问题。由于训练数据没有未知类，所以该训练方式属于半监督学习。

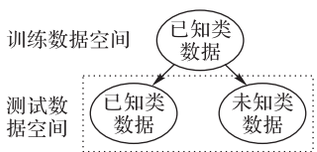

图2 视觉异常识别

Fig. 2 Visual anomaly recognition

2.2 利用生成对抗网络提取视觉误差信息

网络设计思路借鉴GANomaly[15]模型，该生成对抗网络为半监督异常识别：编码器-解码器-编码器（Encoder-Decoder-Encoder）流水线内的新型对抗自动编码器，捕获图像在深度特征空间内的训练数据分布；是一种有效且新颖的异常识别方法，可以在统计和计算上提供更好的性能，如图3模型框架所示。

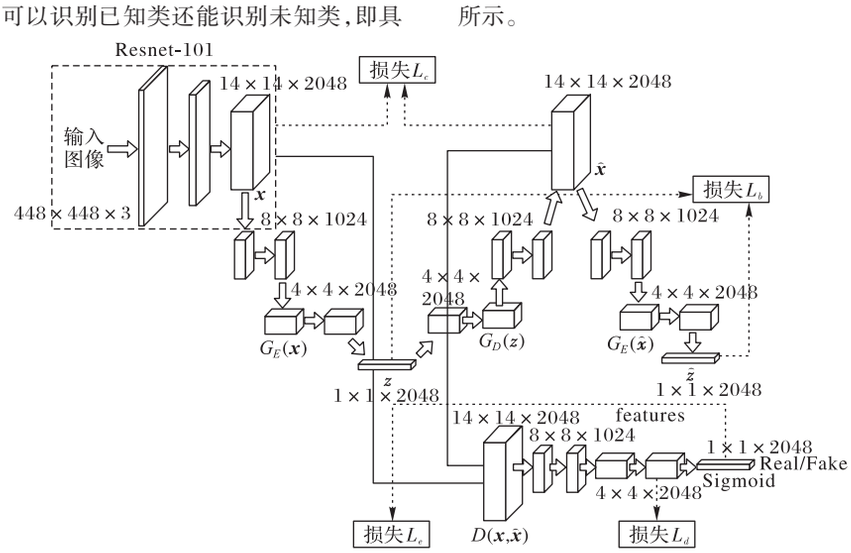

图3 半监督学习方式的生成对抗网络框架

Fig. 3 Generative adversarial network framework of semi-supervised learning mode

整个框架由4部分组成：

1）视觉特征网络。采用深度残差网络，去掉最后的分类层和池化层，最后提取到深度特征图像（feature map）。

2）生成网络由编码器 $G_E(\boldsymbol{x})$ 和解码器 $G_D(z)$ 构成。对于送入深度特征图数据 $\boldsymbol{x}$ 经过编码器 $G_E(\boldsymbol{x})$ 得到深度视觉向量 z，z 经过解码器 $G_D(z)$ 得到 $\boldsymbol{x}$ 的重构数据 $\hat{\boldsymbol{x}}$。

3）判别网络 $D(\boldsymbol{x},\hat{\boldsymbol{x}})$。网络与编码器 $G_E(\boldsymbol{x})$ 相同，对原始图像 $\boldsymbol{x}$ 判为真，重构图像 $\hat{\boldsymbol{x}}$ 判为假，不断优化重构数据 $\hat{\boldsymbol{x}}$ 与源数据 $\boldsymbol{x}$ 的差距，理想情况下重构数据与源数据是完全一样的。

4）重构编码网络 $G_E(\hat{\boldsymbol{x}})$。对重构图像 $\hat{\boldsymbol{x}}$ 再编码，由编码器 $G_E(\hat{\boldsymbol{x}})$ 得到重构图像编码的深度视觉向量 $\hat{z}$，使用 L_b 可以不断缩小深度视觉向量 z 与 $\hat{z}$ 的差距，理想情况下它们是完全一样的。

在训练阶段，整个模型均是通过已见过类别的正常样本做训练。也就是编码器 $G_E(\boldsymbol{x})$、解码器 $G_D(z)$ 和重构编码器 $G_E(\hat{\boldsymbol{x}})$ 都适用于正常样本。测试阶段，当模型在测试阶段接收到一个异常样本，此时模型的编码器、解码器将不适用于异常样本，此时得到的编码后深度视觉向量 z 和重构编码器得到的深度视觉向量 $\hat{z}$ 的差距是大的。这时候规定这个差距是一个分值，通过设定阈值 φ，一旦深度视觉向量之间的均方误差大于设定阈值 φ，模型就认定送入的样本 $\boldsymbol{x}$ 为未见过类别的异常样本。对于目标测试数据，经过源数据训练，根据阈值 φ，目标数据可以正确被区分已知类和未知类。

2.3 利用属性语义信息实现零样本图像分类

融合属性语义特征实现零样本图像分类网络作为基准实验（Baseline），它由视觉特征网络、属性语义转换网络和视觉-属性语义衔接网络构成。视觉特征网络采用在ImageNet[16]预

训练 1 000 类别的 Resnet-101[17]模型,去掉最后分类层获取 2 048 个特征向量。视觉-属性语义衔接网络将高维属性语义特征嵌入到视觉特征。由于不同的数据集的属性语义特征维度不同且特征数量维度比较低,所以需要引入属性语义转换网络,实现特征从低维空间映射到高维空间,平衡属性语义特征影响程度与视觉特征的影响程度,该子网络采用了双层的线性激活层。图 4 展示了属性语义转换子网络与视觉-属性语义衔接子网络。

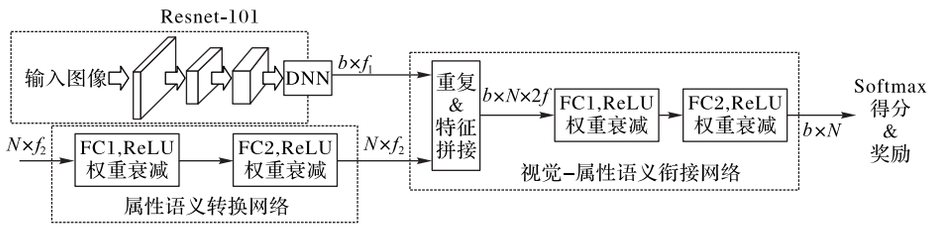

图 4　视觉-属性语义嵌入网络

Fig. 4　Visual-attribute semantic embedding network

属性语义转换网络用于产生所有类别的属性特征表征。在训练阶段,将源数据已知类别数 N^s 的属性语义特征映射到高维;在预测阶段,将目标数据已知和未知总类别数 N^{s+t} 的属性语义特征映射到高维,规定与 DNN 视觉特征的维度数相同。当确定要输入训练的类别属性特征 $N\times f_2$(f_1 和 f_2 表示特征维度)后将其送入属性语义转换网络,特征映射到 $N\times f_2'$,再使用 Repeat 产生和图像表征一样的批大小 b,即张量维度变为 $b\times N\times f_2'$,通过特征拼接(Feature concatenation)操作完全拼接视觉-属性语义特征输出张量维度 $b\times N\times 2f$,最后送入视觉-属性语义衔接网络,视觉-属性语义衔接网络也同样使用了两层的线性激活层,得到 $b\times N$。对于送入的源数据,最后计算的是分类得分。

2.4　融合视觉误差信息提升零样本图像分类

对于生成对抗网络,首先,生成器网络中采用编码器-解码器-编码器子网络使得模型能够将输入图像映射到较低维度的矢量,该较低维度矢量用于重建所生成的输出图像;然后,在训练期间最小化这些图像与深度视觉向量之间的距离有助于学习正常样本的数据分布,而在预测阶段图像编码的深度视觉空间下进行对比,源数据图像在深度视觉空间下差异下,目标数据在深度视觉空间下差异大;最后将训练好的深度视觉空间特征差异融入属性语义训练指导零样本图像分类训练中。

融合视觉误差信息的方式在这里直接用目标数据中每一个实例样本的均方误差值。首先训练好生成对抗网络,并设定一个好的阈值 φ(一般取经验值 0.2,然后微调),使尽可能得到较强区分目标数据中已见过类别和未见过类别的能力;然后训练零样本图像分类网络,输入的数据不使用无标签的目标数据,反之是训练全量数据下的零样本图像分类;接着分别训练好生成对抗网络与零样本图像分类网络后进行测试,测试阶段在视觉-语义嵌入网络融入视觉误差信息,利用视觉误差信息确定目标数据中已见过类别和未见过类别,根据阈值 φ 来决定视觉-语义嵌入网络在已见过类别空间和未见过类别空间搜索。

视觉-语义衔接网络在重复与特征拼接操作时候被视觉误差信息选定了已见过类别空间和未见过类别空间,这样再进行两层的线性激活层分类可以有效防止类别预测偏置,即:未知类被预测到已知类,而已知类也可能被预测到未知类。

在理想情况下,生成对抗网络能够完美区分目标数据中已知类和未知类,但由于异常数据等问题,实际情况仍然会有交叉预测样本。所以在融入视觉误差信息时,差分阈值 φ 可以根据实际情况来调整。另外差分阈值的选定可以不止一个,比如也可以是两个,将误差区域分成三段,如图 5 所示,第一段是可以很大程度划分出已见过类别的数据,第二段是可以很大程度划分出未见过类别的数据,第三段介于第一段与第二段之间。第一段在已知语义空间预测,第二段在未知语义空间预测,第三段在全集类别语义空间预测。这样可调节分段阈值的好处就在于进一步提高目标数据的预测,同时也有助于实际应用的需要。

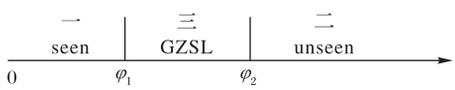

图 5　三段类别样本分布空间

Fig. 5　Sample distribution space of three-segment category

2.5　模型优化

模型优化主要是损失函数,包含半监督生产对抗网络和视觉-属性语义嵌入网络。

视觉-属性语义嵌入网络在训练阶段执行的是强监督学习,所以定义如下损失函数 L_a:

$$L_a=\frac{1}{N_s}\sum_{i=1}^{N_s}L_a\left(\boldsymbol{x}_i^s\right)+\lambda\|\boldsymbol{W}\|^2 \tag{1}$$

一般地,传统的强监督分类器的损失函数包括分类损失 L_a 和正则化项损失 $\|\boldsymbol{W}\|^2$,L_a 采用交叉熵,正则化用 L_2 范数来约束训练,该正则是为了缓解神经网络在训练学习过程中某些神经元不被激活停滞问题。λ 用于损失函数中正则化权重,通过实验多次修改训练验证来确定。在训练阶段,所有标注的源数据被打乱再一起作为训练数据。模型使用 Adam 算法进行优化。

半监督学习方式的生成对抗网络框架的损失函数设置如图 3 所示,由四部分损失构成,分别是:深度视觉向量误差 L_b、原图与重构图误差 L_c、特征匹配误差 L_d 以及真假判别损失 L_e。

重构图像编码得到的深度视觉向量 $\hat{\boldsymbol{z}}$,这部分对于正常数据而言,希望得到的 $\hat{\boldsymbol{z}}$ 与原始数据直接编码得到的 $\boldsymbol{z}$ 无差别最好,也就是对于 L_b 最好的得分判断。当然这是最理想的状态,但是对于正常数据还是希望误差越小越好,所以引入了:

1)深度视觉向量间的误差优化函数 L_b:

$$L_b=\frac{1}{2}\|\boldsymbol{z}-\hat{\boldsymbol{z}}\|_2 \tag{2}$$

2)用于在深度图像特征图面上减小源数据 $\boldsymbol{x}$ 和重构数据 $\hat{\boldsymbol{x}}$ 差距的损失函数 L_c:

$$L_c=\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_1 \tag{3}$$

3)用于在判别网络中深度图像特征层面做优化,设置一个特征匹配误差损失函数 L_d:

$$L_d = \frac{1}{2}\left\| f(\boldsymbol{x}) - f(\hat{\boldsymbol{x}}) \right\|_2 \tag{4}$$

4)为了在判别网络中对判别是源数据$\boldsymbol{x}$还是重构数据$\hat{\boldsymbol{x}}$，设置一个二进制交叉熵损失函数L_e：

$$L_e = y \log D(\boldsymbol{x}) + (1 - y)\log(1 - D(\hat{\boldsymbol{x}})) \tag{5}$$

其中：$y \in \{0, 1\}$，“1”表示优化判断是源数据$\boldsymbol{x}$，而“0”表示判断为重构数据$\hat{\boldsymbol{x}}$。

训练过程只有已知类数据参与，模型只对已知类数据可以做到较好的编码解码，所以送入未知类数据在编码解码下会出现编码得到的深度视觉向量差异大从而使得差距分值φ大，判断为未见过类别即是异常。对于模型，整个损失函数的优化分为生成器L_G和判别器L_D表示为：

$$\begin{cases} L_G = \omega_b L_b + \omega_c L_c \\ L_D = \omega_d L_d + \omega_e L_e \end{cases} \tag{6}$$

其中：ω_b、ω_c、ω_d和ω_e是调节各损失的参数，可以根据具体实验设置。式(6)展示了各损失函数之间关系，模型训练分别学习生成器和判别器，协同交替迭代更新。

3 实验与结果分析

3.1 实验数据集和评价指标

数据集来源与评价指标主要参考文献[12]，所属领域分别是鸟类与动物类，每一类的图像数据相对均衡，数据划分比例也基本均衡。

1)CUB(Caltech-UCSD-Birds-200-2011)数据集。全部都为鸟类的图像，总共200个类，其中：150类为训练集，50类为测试集。类别的属性语义特征维度为312，共有11 788张图片，划分情况见表1所示。

表1 CUB和AwA2数据集划分情况

Tab. 1 Division of CUB and AwA2 datasets

数据集划分	CUB数据集		AwA2数据集	
	已知类(150类)	未知类(50类)	已知类(40类)	未知类(10类)
训练集	7 057	—	23 527	—
测试集	1 764	2 967	5 882	7 913

图6 CUB和AwA2数据集样例

Fig. 6 Samples of CUB and AwA2 datasets

2)AwA(Animal with Attributes)数据集。

提出ZSL定义的作者，给出的数据集都是动物的图片，包括50个类别的图片，其中：40个类别数据作为训练集，另外10个类别数据作为测试集。每一个类别的属性语义特征维度为85，总共37 322张图片，本文使用AwA2数据集，划分情况见表1所示。

评估指标有四个，分别是T_1、ACC_{senn}、ACC_{unseen}和H。

1)正确率。

ZSL：只评估在未知类下的正确率T_1(传统设定下)。

GZSL(ACC_{seen}，ACC_{unseen})：分别评估在全集类别下预测已知与未知类的正确率(广义设定下)。

2)调和指标H值：

$$H = \frac{2 \times ACC_{\text{seen}} \times ACC_{\text{unseen}}}{ACC_{\text{seen}} + ACC_{\text{unseen}}} \times 100\% \tag{7}$$

3.2 对比模型

为了验证模型的有效性，本文选取了以下模型作为基准实验。

1)深度嵌入模型(Deep Embedding Model，DEM)[11]：该方法使用深度ZSL模型成功的关键是选择合适的嵌入空间，即使用视觉空间作为嵌入空间，而不是嵌入语义空间或中间空间。在这个空间里，随后的最近邻搜索将受到中心问题的影响，因此变得更有效。该模型设计还提供了用于端到端方式联合融合和优化的多个语义模态(例如，属性和句子描述)的自然机制。

2)GAZSL[12]：该方法将关于未知类(例如维基百科文章)的噪声文本描述作为输入，并为该类生成视觉特征。通过添加伪数据，将零样本学习自然地转换为传统的分类问题。另外，为了保持所生成特征的类间区分，提出了视觉枢轴正则化作为显式监督。

3)关系网络(Relation Network，RN)[13]：是从头开始端到端训练的。在元学习期间，它学习深度距离度量以比较一批中的少量图像，每个图像旨在模拟少数样本设置。一旦经过训练，RN就能够通过计算查询图像与每个新类的少数样本之间的关系得分来对新类别的图像进行分类，而无需进一步更新网络。

3.3 实验设置

将本文提出的融合视觉误差信息提升零样本图像分类方法分别在AwA2和CUB数据集中进行实验，对比不同方法所得到的分类结果的调和指标H值。所有基准实验的实验环境均为单机环境，处理器为Intel Xeon CPU E5-2620 v4@2.10 GHz，操作系统为Ubuntu 16.04.1 LTS，内存为64 GB RAM，GPU为GeForce GTX 1080Ti，显存为11 GB，开发平台为Python 2.7.13，所有算法实现语言为Python，使用深度学习框架PyTorch。实验基本参数设置包括零样本图像分类网络和生成对抗网络，分别如表2和表3所示。

表2 零样本图像分类网络实验参数设置

Tab. 2 Setting of experimental parameters for zero-shot image classification network

参数	设置	参数	设置
学习率	0.000 01	batch_size大小	64
迭代次数	200 000	正则化系数λ	0.000 01

表3 生成对抗网络实验基本参数设置

Tab. 3 Setting of basic experimental parameters for generative adversarial network

参数	设置	参数	设置
学习率	0.000 2	ω_b、ω_c、ω_d、ω_e	1，50，1，1
迭代次数	1 000	深度视觉向量长度	2 048
batch_size大小	64		

训练半监督学习方式的生成对抗网络，随着训练不断迭代进行，已知类和未知类区分程度越来越明显，其损失下降也将变得缓和。分别将AwA2数据集和CUB数据集的测试数据使用均方误差计算在生成网络下$\boldsymbol{z}$和$\hat{\boldsymbol{z}}$的误差结果，如图7~8所示。

分别从实验结果中取阈值φ的合理区间，AwA2选择

[0.32,0.38]闭区间,CUB选择[0.10,0.12]闭区间,再通过微调可以确定具体值。在零样本图像分类实验中,仅使用有标签数据,而假设情况下的测试数据用于Baseline效果测试。

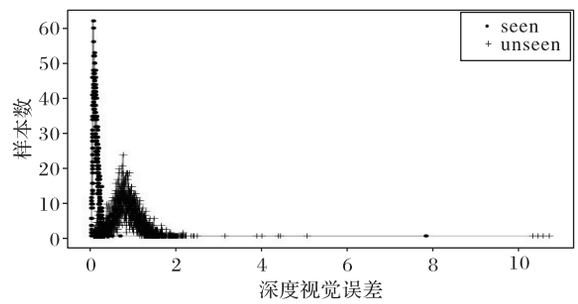

图7　AwA2数据集视觉误差检测结果

Fig. 7　Visual error detection results of AwA2 dataset

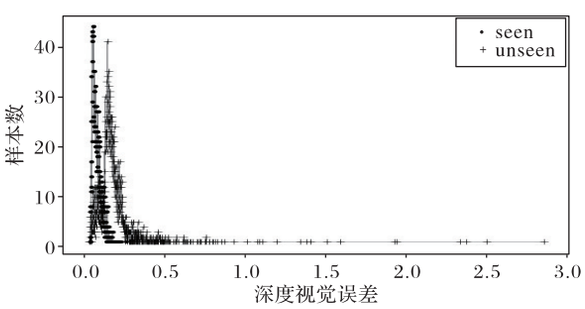

图8　CUB数据集视觉误差检测结果

Fig. 8　Visual error detection results of CUB dataset

比较生成对抗网络中使用判别网络视觉误差,分别将AwA2数据集和CUB数据集的测试数据使用均方误差计算$f(\boldsymbol{x})$和$f(\hat{\boldsymbol{x}})$的误差结果。采用判别网络计算的视觉误差在已知类和未知类上几乎同分布,使判别网络难于区分已知类和未知类。由此确定,半监督生成网络与判别网络相比,半监督生成网络生成的深度视觉误差更能够区别已知类和未知类。

在使用生成对抗来获取视觉误差时,有两个问题需要确定:1)不增加判别网络$D(\boldsymbol{x},\hat{\boldsymbol{x}})$对区分已知类和未知类是否有影响;2)不增加重构编码网络$G_E(\hat{\boldsymbol{x}})$而直接用判别网络$D(\boldsymbol{x},\hat{\boldsymbol{x}})$来提取视觉误差对区分已知类和未知类的效果如何。这两个问题通过控制变量法来确定。以AwA2数据集和CUB数据集为例,问题1)将生成对抗剔除判别网络,只训练生成网络和重构编码网络;问题2)剔除重构编码网络,只训练生成网络和判别网络。

对问题1)的实验情况,在同等条件下,增加判别网络会比不增加判别网络学习区分已知类和未知类更快,且更加稳定(可将阶段性学习过程可视化观察)。值得注意的原因是判别网络的对抗学习的确可以促进生成网络的学习能力。对问题2)的实验情况,在同等条件下,不增加重构编码网络会比增加重构编码网络对区分已知类和未知类的能力差。值得注意的原因是减少了深度视觉向量间的误差损失会降低网络学习能力。

3.4　结果分析

利用生成对抗提升零样本图像分类方法将生成对抗网络实现异常检测结果的视觉误差信息融入到零样本图像分类当中,实验分别在AwA2数据集和CUB数据集上对比基准实验效果。Baseline是只利用属性语义信息实现零样本图像分类结果,GAN_enhance_ZSL是融合了视觉误差信息和属性语义信息后的提升结果。提升结果有两组数据:第一组,实验微调阈值φ,将视觉误差分成两段,确定AwA2数据集的具体值是0.33,确定CUB数据集的具体值是0.11。第二组,视觉误差分成三段,确定AwA2数据集视觉误差三段区间为[0,0.27]、(0.27,0.40]、(0.40,∞),确定CUB数据集视觉误差三段区间为[0,0.10]、(0.10,0.13]、(0.13,∞)。

表4　对比基准实验效果　　单位:%

Tab. 4　Comparison with the benchmark experiment　　unit:%

模型		AwA2(类数为50)				CUB(类数为200)			
		ZSL	GZSL			ZSL	GZSL		
		T_1	ACC_{unseen}	ACC_{seen}	H	T_1	ACC_{unseen}	ACC_{seen}	H
DEM		67.1	30.5	86.4	45.1	51.7	19.6	57.9	29.2
GAZSL		**69.0**	27.0	82.4	40.6	**55.8**	23.9	60.6	34.3
RN		64.2	30.0	93.4	45.3	55.6	38.1	61.4	47.0
Baseline		68.2	41.3	91.4	56.9	55.1	42.5	55.5	48.1
GAN_enhance_ZSL	两段	68.2	61.6	94.3	74.5	55.1	48.2	60.8	53.9
	三段	68.2	**64.9**	**94.5**	**77.0**	55.1	**49.6**	**63.4**	**55.7**

对于本文Baseline只使用属性语义信息的实验,在AwA2数据集上,对比DEM网络,ACC_{unseen}提升10.8个百分点,而ACC_{seen}提升5.0个百分点;对比RN网络,ACC_{unseen}提升11.3个百分点,但ACC_{unseen}下降2.0个百分点。在CUB数据集上,对比RN网络,ACC_{unseen}提升4.4个百分点,但ACC_{unseen}下降5.9个百分点。在T_1指标上表现良好,而调和指标H在AWA2数据集优于CUB数据集,值得注意的原因是AWA2数据集训练数据集样本数相对较多且类别相对少。

对比Baseline,融入视觉误差信息后ACC_{seen}和ACC_{unseen}指标均有所提升。在AWA2数据集上,保证ACC_{seen}提升2.9个百分点情况下,ACC_{unseen}还提升20.3个百分点,调和指标H提升17.6个百分点;在CUB数据集上,保证ACC_{seen}提升5.3个百分点情况下,ACC_{unseen}还提升6.7个百分点,调和指标H提升了5.8个百分点。

对于GAN_enhance_ZSL实验,第二组实验分成三段的视觉误差明显优于第一组实验分成两段的视觉误差,主要原因与使用预测类别空间相一致。在AwA2数据集上,同样对比RN网络,在保证ACC_{seen}提升1.1个百分点情况下ACC_{unseen}提升34.9个百分点,调和指标H提升31.7个百分点。在CUB数据集上,对比未使用测试数据的RN网络,在保证ACC_{seen}提升2.0个百分点情况下ACC_{unseen}提升1.7个百分点,调和指标H提升8.7个百分点。在传统定义下的指标T_1,相比GAZSL略低,但仍表现良好。

只使用属性语义信息的Baseline实验中,T_1指标总是远大于ACC_{unseen},显然预测发生极大偏置现象,具体发生在未知类预测为已知类,而当融入视觉误差信息后这种现象随即缓解。

另外只使用属性语义信息的Baseline虽然在AWA2数据集和CUB数据集的ACC_{unseen}指标相比基准实验都提高，但是需要降低ACC_{seen}的代价。融入视觉误差信息后，实验效果明显有大幅度提升，说明使用半监督学习方式的生成对抗网络不仅有效促进神经网络辨别已知类和未知类，还能缩小零样本图像分类在广义下预测类别空间，有效防止预测发生偏置问题。对比AwA2与CUB结果，AwA2的提升比CUB提升多很多，值得注意的原因是数据集本身的区别，动物类的形态迥异，而鸟类形态大都相似，所以使用半监督学习方式的生成对抗网络区别已知类和未知类对类别差异大的效果更好。

4 结语

本文针对现有零样本图像分类方法仅利用属性语义信息来指导分类而未使用视觉误差信息，导致模型无法显式地辨别已知类和未知类，存在较大的预测偏置的问题，提出融合视觉误差信息与属性语义信息来提升零样本图像分类。当模型学习完已知类的源图像数据后，对于目标图像数据中已知类视觉误差小，而未知类视觉误差大。利用视觉误差信息在一定程度上有助于零样本图像分类模型将目标图像数据中已知图像类别和未知图像类别区分开来，还能缩小零样本图像分类在广义下预测类别的搜索空间，可以有效防止预测类别发生偏置问题。视觉误差信息的获取采用半监督学习方式的生成对抗网络，零样本图像分类采用融合属性语义信息训练。实验结果表明，与基准模型相比，本文所提出来的融合视觉误差信息后的零样本图像分类在AwA2数据集和CUB数据集的评价指标上均有大幅度提升趋势，验证了本文方法在提升零样本图像分类问题上的有效性。

参考文献（References）

[1] LAROCHELLE H, ERHAN D, BENGIO Y. Zero-data learning of new tasks [C]// Proceedings of the 23rd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2008: 646-651.

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

[3] 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.

[4] ROMERA-PAREDES B, TORR P H S. An embarrassingly simple approach to zero-shot learning [C]// Proceedings of the 32nd International Conference on Machine Learning. New York: JMLR. org, 2015: 2152-2161.

[5] LIU M, ZHANG D, CHEN S. Attribute relation learning for zero-shot classification[J]. Neurocomputing, 2014, 139: 34-46.

[6] 巩萍,程玉虎,王雪松. 基于属性关系图正则化特征选择的零样本分类[J]. 中国矿业大学学报, 2015, 44(6): 1097-1104. (GONG P, CHEN Y H, WANG X S. Zero-shot classification based on attribute correlation graph regularized feature selection[J]. Journal of China University of Mining and Technology, 2015, 44(6): 1097-1104.)

[7] AKATA Z, REED S, WALTER D, et al. Evaluation of output embeddings for fine-grained image classification [C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 2927-2936.

[8] XIAN Y, AKATA Z, SHARMA G, et al. Latent embeddings for zero-shot classification [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 69-77.

[9] AKATA Z, MALINOWSKI M, FRITZ M, et al. Multi-cue zero-shot learning with strong supervision [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 59-68.

[10] ZHANG L, XIANG T, GONG S. Learning a deep embedding model for zero-shot learning [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 3010-3019.

[11] ZHU Y, ELHOSEINY M, LIU B, et al. A generative adversarial approach for zero-shot learning from noisy texts [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 1004-1013.

[12] SUNG F, YANG Y, ZHANG L, et al. Learning to compare: relation network for few-shot learning [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 1199-1208.

[13] LI Y, ZHANG J, ZHANG J, et al. Discriminative learning of latent features for zero-shot recognition [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7463-7471.

[14] FU J, ZHENG H, MEI T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 4476-4484.

[15] AKCAY S, ATAPOUR-ABARGHOUEI A, BRECKON T P. GANomaly: semi-supervised anomaly detection via adversarial training [C]// Proceedings of the 2018 Asian Conference on Computer Vision, LNCS 11363. Cham: Springer, 2018: 622-637.

[16] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[17] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.

This work is partially supported by the National Natural Science Foundation of China (61772135, U1605251, 61703195), the Open Fund of the Key Laboratory of Network Data Science and Technology of the Chinese Academy of Sciences (CASNDST201708, CASNDST 201606), the Open Fund of the State Key Laboratory of Pattern Recognition (201900041), the Surface Program of Fujian Natural Science Foundation (2017J01755), the CERNET Innovation Project (NGII20160501).

XU Ge, born in 1978, Ph. D., associate professor. His research interests include artificial intelligence, natural language processing.

XIAO Yongqiang, born in 1994, M. S. candidate. His research interests include multi-modal, artificial intelligence.

WANG Tao, born in 1987, Ph. D., lecturer. His research interests include scene understanding, object detection and segmentation.

CHEN Kaizhi, born in 1983, Ph. D., lecturer. His research interests include natural language processing, image recognition, depth learning algorithm.

LIAO Xiangwen, born in 1980, Ph. D., professor. His research interests include opinion mining, sentiment analysis, natural language processing.

WU Yunbing, born in 1976, M. S., associate professor. His research interests include knowledge representation and knowledge discovery.