·人工智能与模式识别·　　文章编号：1000-3428(2023)10-0064-08　　文献标志码：A　　中图分类号：TP18

基于深度互学习的多标记零样本分类

袁志祥，王雅卿，黄　俊

（安徽工业大学 计算机科学与技术学院，安徽 马鞍山 243032）

摘　要：目前已有大量方案解决零样本图像分类问题，但对多标记零样本图像分类问题的研究很少，在现有的解决方案中，模型在训练时除了利用已标注的数据集和给定的先验知识外，只利用图像区域信息或只利用标签语义信息。基于深度互学习技术，提出一种能同时利用图像区域和标签语义两种信息的解决方法。设计两个子网络，将子网络1用于增强图像视觉特征，通过多头自注意机制关联图像中不同区域的特征信息，得到基于区域的视觉特征表示，再将该特征表示映射到语义空间中，并输出预测概率分布；使子网络2用于融合标签语义信息与图像视觉特征，通过计算标签和图像区域特征的相关性，得到基于语义的视觉特征表示，将特征表示映射到语义空间中输出概率分布。最后引入深度互学习技术，利用两个子网络得到的概率分布为对方提供训练经验以进行互相学习，该过程中子网络在训练自身分类性能的同时也学习对方的训练经验，有效提升多标记零样本图像分类的性能。实验结果表明，所提方法在MS COCO数据集上的F1值相比Deep0Tag方法提升了5.2个百分点。

关键词：深度学习；图像分类；多标记学习；零样本学习；互学习

开放科学（资源服务）标志码（OSID）：

中文引用格式：袁志祥，王雅卿，黄俊. 基于深度互学习的多标记零样本分类[J]. 计算机工程，2023，49(10)：64-71.

英文引用格式：YUAN Z X，WANG Y Q，HUANG J. Multi-label zero-shot classification based on deep mutual learning[J]. Computer Engineering，2023，49(10)：64-71

Multi-Label Zero-Shot Classification Based on Deep Mutual Learning

YUAN Zhixiang，WANG Yaqing，HUANG Jun

（School of Computer Science and Technology，Anhui University of Technology，Maanshan 243032，Anhui，China）

【Abstract】Numerous methods have been proposed to solve the zero-shot image classification problem；however，there are limited studies on the multi-label zero-shot image classification problem. In the existing solutions，in addition to the use of the basic settings of the labeled dataset and the given prior knowledge，the model either only uses the image region information or only the label semantic information. Based on deep mutual learning technology，this study proposes a solution that utilizes both the image region and label semantic information. Two sub-networks are designed. Sub-network 1 is used to enhance the visual features of the image，whereas the multi-head self-attention mechanism is used to associate the feature information of different regions in the image to obtain a region-based visual feature representation and then map the feature representation to the semantic space to output the predicted probability. Sub-network 2 is used to fuse the label semantic information and image visual features by calculating the correlation between the labels and image region features to obtain a semantic-based visual feature representation，and then map the feature representation to the semantic space to output a probability distribution. Finally，the deep mutual learning technology is introduced，and the probability distribution obtained by the two sub-networks is used to provide training experience for mutual learning. In this process，the sub-network refers to the training experience of the other sub-network while training its own classification performance，which effectively improves the performance of multi-label zero-shot image classification. The experimental results show that the F1 value of the proposed method on the MS COCO dataset increased by 5.2 percentage points compared to the Deep0Tag method.

【Key words】deep learning；image classification；multi-label learning；zero-shot learning；mutual learning

DOI：10. 19678/j. issn. 1000-3428. 0065806

基金项目：国家自然科学基金（61806005）；安徽省高校科学研究重点项目（KJ2021A0372，KJ2021A0373）；安徽省高校优秀青年人才支持计划项目（gxyqZD2022032）。

作者简介：袁志祥（1973—），男，副教授，CCF会员，主研方向为机器学习、Petri网理论；王雅卿，硕士研究生；黄　俊，副教授、博士。

收稿日期：2022-09-20　　**修回日期：**2022-12-05　　**E-mail：**zxyuan@ahut.edu.cn

0　概述

传统的图像分类问题主要属于单标记学习领域的问题，即一个对象只有一个类别标签。而在很多应用中目标对象并没有那么简单，一个对象可以属于好几个类别。例如，在图像分类中，一张图片里可能包含多个物体；在文本分类中，一篇新闻可能涵盖多个主题；在视频分类中，一个电影可能属于多个类型。

由于以往的单标记学习方法只能预测类别单一的样本，使用效果有待改善，因此人们将注意力转移到了多标记学习上。多标记学习的主要任务是通过训练数据，学习高效的分类模型，为输入样本预测可能的类别标记集合。随着数据集的扩大和深度学习方法越来越成熟，多标记学习问题得到解决，但在实际应用场景中，大部分数据集依然没有类别标记，这要求模型能够识别训练过程中从未见过的类别，于是多标记零样本学习应运而生。多标记零样本学习模拟了人类学习未知事物的过程，利用以往学习到的先验知识为目标样本推理预测多个未见过的新类别。然而目前的零样本问题几乎也都属于单标记学习领域，在多标记方向上的研究很少。

本文针对多标记零样本分类问题，提出一种基于深度互学习技术的解决方案。该方案包含3个模块，其中一个子网络利用图像中每个区域与其他区域的关联信息来增强图像本身的特征，挖掘图像中存在的类别标签，包括已知和未知；另一个子网络将标签的语义信息与图像的每个区域特征相融合，在训练过程中引入标签语义使知识可以很好地从已知标签转移到未知标签；另一个是深度互学习模块，该模块能使两个子网络在训练过程中做深度互学习，即他们在训练自身分类性能的同时还能互相学习对方的训练经验，从而达到互相促进、共同进步的目的。

1　相关工作

1.1　多标记学习

随着数据集的扩大和深度学习方法的逐渐成熟，图像分类领域取得了显著的发展。多标记分类的任务是为输入图像预测多个标签，通过为每个标签学习一个二元分类器[1]完成，但它有两个缺点，一是在处理大量标签时增加了计算的复杂性，二是不包含标签之间的相关性。近年来大多数多标记学习方法均聚焦在挖掘标签之间的相关性上，比如文献[2]通过对标记空间进行属性聚类来挖掘标签的局部相关性；文献[3]利用余弦相似性来计算标签的全局和局部相关性；文献[4-5]采用图神经网络建立标签之间的依赖关系；文献[6]基于先验知识的词嵌入将标签转化为嵌入的标签向量后，再利用标签之间的相关性；文献[7-9]使用基于注意力机制的方法来解决多标记问题，通过编码图像的每个区域，使训练过程中的模型能注意到图像中的每个部分。

1.2　多标记零样本学习

虽然上述大多数方法在传统的多标记学习中都能取得很好的成绩，但不能直接应用到多标记零样本学习。由传统多标记学习方法训练得到的模型只能识别和预测它学习过程中见过的类别，见过的类别越多，即训练数据越多，该模型的分类性能就越好。尽管研究人员为了科研工作标记了大量数据集，但在现实生活中的数据依旧是未标记占绝大多数，导致以往的训练方法很难有效地解决实际问题，于是人们开始关注多标记零样本学习。

随着对零样本图像分类的广泛研究，模型在很大程度上克服了对未知类别数据进行分类的局限。零样本学习依赖于已知类别与未知类别之间相关联的语义信息，这通常是利用相关先验知识得到的，比如属性、词向量、文本描述等。零样本学习的解决方式主要分为两种，一种是将图像视觉特征和标签语义向量结合起来学习，如文献[10]提出的ALE(Attribute Label Embedding)模型，首先提取图像视觉特征及类别标签的语义向量，引入一个双线性评分函数，通过衡量视觉特征嵌入语义空间的兼容度来预测输入图像的类别。文献[11]提出LDF(Latent Discriminative Features)模型，能够发现图像中的判别性区域，并将图像的判别性区域特征与图像的全局特征进行联合学习，提升分类的准确率。另一种通过生成模型来生成未知标签的特征，再将其当做传统的监督学习进行训练，比如基于生成对抗网络(Generative Adversarial Network，GAN)[12]的方法和进一步对生成对抗网络进行优化的GMMN[13]方法等。

上述方法在零样本领域取得了巨大的成功，但这些解决方案并不能直接用到多标记零样本分类问题中。多标记零样本分类任务是为输入图像预测多个已知标签和未知标签。目前，对多标记零样本学习问题的研究较少，比较典型的有文献[14]中结合知识图谱的框架来描述多标签之间的关系，以此来建模已知类和未知类之间的相互依赖，但它需要访问已知和未知标签之间的先验知识图；相似地，还有文献[15]介绍的融合图卷积神经网络(Graph Convolutional Network，GCN)的多标记零样本学习框架，也是利用图来学习标签相互依赖的分类器；文献[16]提出一种基于生成模型的多标记零样本学习方法，它提出的CLF(Cross-Level feature Fusion)方法结合了ALF(Attribute-Level Fusion)标签依赖性和FLF(Feature-Level Fusion)特定类判别性的优点，并将其集成到常用生成模型框架中进行预测分类。还有一些基于注意机制的解决方案，例如文献[17]介绍了多模态注意，它可用于为每个标签产生特定的注意，并通过标签语义推广到未知标签，但是对数千个标签需要计算数千个注意，这会导致巨大的时间和内存消耗。文献[18]提出一种共享多注意框架，该框架为一幅图像学习所有类别共享的多个注意力模块，利用得到的多个注意力权重对图像的区域特征进行加权；而后文献[19]在其上进行优化，提

出双层注意模块，通过融合图像的区域和全局信息来增强图像视觉特征。这两个模型的缺点在于在训练过程中只单独关注到图像特征，包括利用区域特征与区域特征之间的关联以及区域特征与全局特征之间的关联，并没有引入标签语义信息参与训练。

以上目前存在的多标记零样本学习方法在训练过程中除了利用一般图像分类任务所给定的基础信息（已标记的样本和类别先验知识）外，要么就只利用图像区域信息，要么就只利用标签语义信息。而本文提出的基于深度互学习技术的解决方案，在两个子网络互相学习的过程中，不仅可以起到互相促进、互相增强的效果，而且可以同时将图像区域信息和标签语义信息一起引入到训练过程中，这样得到的模型既能识别未知类，又能更全面地挖掘图像中存在的已知和未知标记。

1.3 深度互学习

文献[20]介绍了深度互学习，其灵感来源于模型蒸馏算法。模型蒸馏算法需要有教师网络和学生网络，教师网络向学生网络单方向传递它自身所学到的知识，即教师网络单方面教学生网络，并不能从学生网络上学到东西。而且在做蒸馏的时候，要有一个训练好的网络当教师，但深度互学习是将多个子网络同时进行训练，这些子网络不仅被真实标签值监督来训练自身的预测性能，而且能通过学习其他子网络的训练经验来进一步提高预测能力。在模型训练时，多个子网络之间都在不断分享训练经验，互相学习、互相增强，从而实现共同进步。

2 本文方案

2.1 问题定义

本文用C^S表示已知类别集合，其中S表示已知类别个数；用C^U表示未知类别集合，其中U表示未知类别个数。已知类别表示在训练过程中出现过的类别，而未知类别表示训练过程中没有出现过，只包含在测试数据集中的类别。$C^{S+U}\triangleq C^S\cup C^U$表示包括已知和未知类别的集合。$(I_1,Y_1),(I_2,Y_2),\cdots,(I_N,Y_N)$表示$N$个训练样本，其中$I_i$表示第$i$个训练图像，$Y_i\subseteq C^S$表示第$i$个训练图像对应的标签集合。由于未知类没有对应的训练图像，本文假设给定标签描述的语义向量$\{\boldsymbol{V}_c\}_{c\in C^{S+U}}$，给定的标签语义向量可以是属性或者词嵌入。传统多标记零样本分类的任务是为给定图像I_i预测其存在的多个未知标签$Y_i\subseteq C^U$；广义多标记零样本分类的任务是为给定图像I_i预测其存在的多个已知和未知标签$Y_i\subseteq C^{S+U}$。

2.2 多标记零样本学习模型

本文提出一种基于深度互学习技术的方案来解决多标记零样本图像分类问题，框架如图1所示，该模型由两个子网络和一个深度互学习模块组成。具体过程为：给定图像I_i，经过卷积神经网络获得图像特征$\boldsymbol{x}_i$。在区域特征与区域特征相关联的子网络中将$\boldsymbol{x}_i$输入到多头自注意机制，得到图像中各区域特征之间的相关性权值$\boldsymbol{r}_m$，m为多头自注意机制的投影头，最终利用图像中各区域相关信息得到基于区域的特征$\boldsymbol{F}_i$，将$\boldsymbol{F}_i$映射到语义空间中，计算每个标签的置信度分数；在区域特征与标签语义相关联的子网络中，通过计算标签语义$\boldsymbol{V}=\{\boldsymbol{V}_1,\boldsymbol{V}_2,\cdots,\boldsymbol{V}_S\}$与图像特征$\boldsymbol{x}_i$的相关性权重，对标签语义与图像特征进行融合，最终得到基于语义的特征$\boldsymbol{F}_g$，将$\boldsymbol{F}_g$映射到语义空间中，计算每个标签的置信度分数；最后加上深度互学习模块，引入一种损失函数对整个模型进行约束，使得两个子网络能够一边训练自身的分类性能，一边学习对方的训练经验。

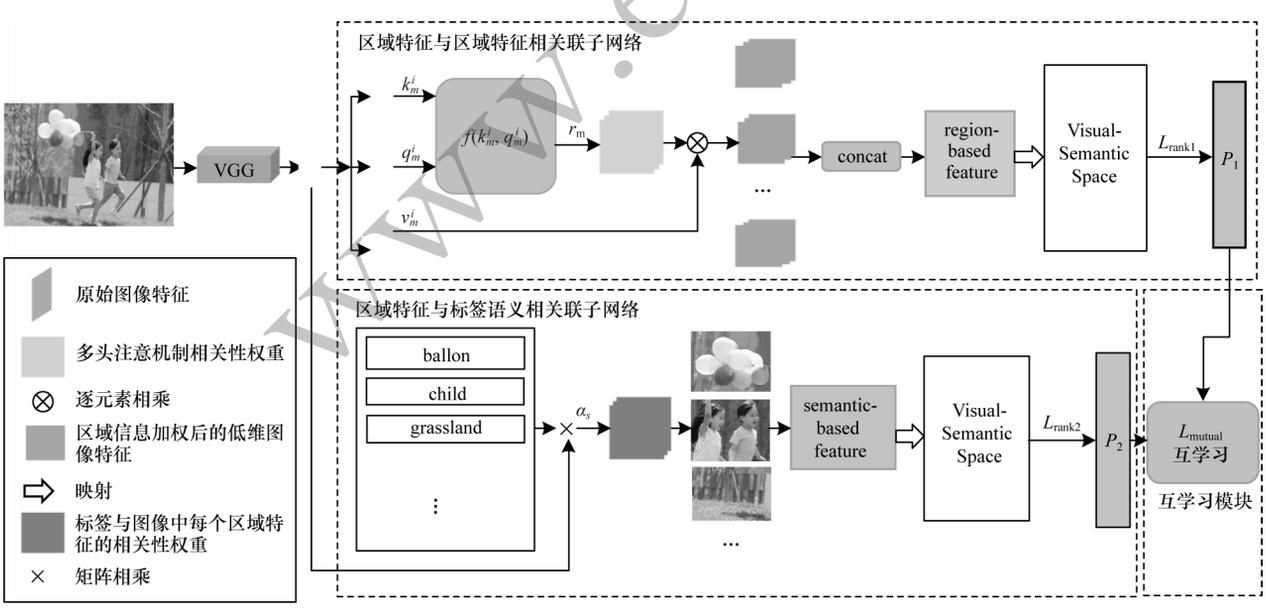

图1 基于深度互学习的多标记零样本学习模型

Fig.1 Multi-label zero-shot learning model based on deep mutual learning

2.2.1　关联区域特征与区域特征的子网络

在该子网络中引入多头自注意力机制来关联图像各个区域，相当于利用各区域的相关信息增强图像自身的特征，得到基于区域的视觉特征表示。

首先从卷积神经网络中提取得到原始图像特征 $\boldsymbol{x}_i \in \mathbb{R}^{h\times w\times d}$，本文可以将其分成 $h\times w$ 个 d 维区域特征，即 $\boldsymbol{x}_i=\left\{\boldsymbol{x}_i^1,\boldsymbol{x}_i^2,\cdots,\boldsymbol{x}_i^{hw}\right\}$，其中 $\boldsymbol{x}_i^r \in \mathbb{R}^{1\times d}$ 表示图像 i 的第 r 个区域。然后将原始图像特征 $\boldsymbol{x}_i \in \mathbb{R}^{h\times w\times d}$ 映射到低维空间（$d'=d/M$），使用 M 个投影头为图像的每个区域创建查询向量（query）、键向量（key）和值向量（value）。则原始特征经过3种映射可得到：

$$\boldsymbol{q}_m^i=\boldsymbol{x}_i\boldsymbol{W}_m^Q \tag{1}$$

$$\boldsymbol{k}_m^i=\boldsymbol{x}_i\boldsymbol{W}_m^K \tag{2}$$

$$\boldsymbol{v}_m^i=\boldsymbol{x}_i\boldsymbol{W}_m^V \tag{3}$$

其中：$\boldsymbol{q}_m^i \in \mathbb{R}^{h\times w\times d'}$；$\boldsymbol{k}_m^i \in \mathbb{R}^{h\times w\times d'}$；$\boldsymbol{v}_m^i \in \mathbb{R}^{h\times w\times d'}$；$m$ 表示自注意机制的第 m 个投影头，$m\in\{1,2,\cdots,M\}$；$\boldsymbol{W}_m^Q$，$\boldsymbol{W}_m^K$，$\boldsymbol{W}_m^V$ 表示可学习的映射权重，查询向量 $\boldsymbol{q}_m^i$、键向量 $\boldsymbol{k}_m^i$ 和值向量 $\boldsymbol{v}_m^i$ 由图像的区域特征分别做3种映射转换得到的。

计算每个区域的查询向量与图像中所有 $h\times w$ 个区域的键向量之间的相关性，可得到图像各个区域的相关权值：

$$\boldsymbol{r}_m=\sigma\left(\frac{\boldsymbol{q}_m^i\boldsymbol{k}_m^{i\,\mathrm{T}}}{\sqrt{d'}}\right) \tag{4}$$

其中：$\boldsymbol{r}_m \in \mathbb{R}^{hw\times hw}$；$\sigma$ 函数用来对权重值做归一化处理。利用得到的权值对值向量进行加权：

$$\boldsymbol{\alpha}_m=\boldsymbol{r}_m\boldsymbol{v}_m^i \tag{5}$$

其中：$\boldsymbol{\alpha}_m \in \mathbb{R}^{h\times w\times d'}$，表示从第 m 个投影头得到的 $h\times w$ 个 d' 维加权区域特征。在多头自注意机制中，图像原始特征的通道数将会从 d 维被切片成 M 个 d' 维，经过计算加权后再合并这些低维特征，得到最终基于区域的特征表示 $\boldsymbol{F}_i$：

$$\boldsymbol{F}_i=\left[\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots\boldsymbol{\alpha}_M\right]\boldsymbol{W}_f \tag{6}$$

其中：$\boldsymbol{W}_f \in \mathbb{R}^{d\times d}$ 表示可学习的权重参数。

本文将加权后的特征表示 $\boldsymbol{F}_i \in \mathbb{R}^{h\times w\times d}$ 也分成 $h\times w$ 个区域，即 $\boldsymbol{F}_i=\left\{\boldsymbol{F}_i^1,\boldsymbol{F}_i^2,\cdots,\boldsymbol{F}_i^{hw}\right\}$，其中 $\boldsymbol{F}_i^r \in \mathbb{R}^{1\times d}$ 表示图像 i 中第 r 个区域的加权特征。

最后将得到的 $\boldsymbol{F}_i$ 映射到语义空间中，可以计算所有标签在图像 i 中的置信度分数，即图像 i 中存在这些标签的概率。计算如下：

$$S_i^c=\max_{r=1,2,\cdots,hw}\boldsymbol{F}_i^r\boldsymbol{\theta}^{c\,\mathrm{T}} \tag{7}$$

其中：c 表示第 c 个标签；$\boldsymbol{\theta}^c \in \mathbb{R}^{1\times d}$ 为 c 的分类器参数；将图像的每个区域特征都与标签 c 的分类器参数做计算，取结果中最大值作为标签 c 的置信度分数 S_i^c。

每个标签的分类器参数都取决于其对应的语义向量，可表示为：

$$\boldsymbol{\theta}^c=\boldsymbol{V}_c\boldsymbol{W}_v \tag{8}$$

其中：$\left\{\boldsymbol{V}_c\right\}_{c\in C^S} \in \mathbb{R}^{1\times d_a}$ 表示已知标签 c 的语义向量，d_a 表示语义向量维度；$\boldsymbol{W}_v \in \mathbb{R}^{d_a\times d}$ 是可学习的权重参数。

如果图像中存在一个标签，那么该标签在图像上的置信度分数一定大于其他不存在的标签，据此引入损失函数作为一种约束，对该子网络进行优化：

$$L_{\mathrm{rank1}}=\sum_i\sum_{c\in y_i,c'\notin y_i}\max\left(1+S_i^{c'}-S_i^c,0\right) \tag{9}$$

其中：y_i 表示图像 i 中所存在标签的集合；S_i^c 表示标签 c 的置信度分数；$S_i^{c'}$ 表示标签 c' 的置信度分数。

2.2.2　关联区域特征与标签语义的子网络

首先将从卷积神经网络中提取到的原始图像特征和所有的标签语义向量输入该子网络，计算每个标签与给定图像中每个区域特征的相关性权重，利用相关性权重融合标签语义信息与图像视觉特征，获得基于语义的视觉特征表示。

$$e_c^r=\boldsymbol{V}_c\boldsymbol{W}_g\boldsymbol{x}_i^{r\,\mathrm{T}} \tag{10}$$

其中：$\boldsymbol{W}_g \in \mathbb{R}^{d_a\times d}$ 是可学习的权重参数；e_c^r 表示标签 c 对图像 i 中第 r 个区域的相关性权重。

$$\alpha_c^r=\frac{\exp\left(e_c^r\right)}{\sum_{c=1}^{S}\exp\left(e_c^r\right)} \tag{11}$$

其中：α_c^r 将式（10）得到的相关性权重做归一化处理。

$$\boldsymbol{F}_c=\sum_{r=1}^{hw}\alpha_c^r\boldsymbol{x}_i^r \tag{12}$$

其中：$\boldsymbol{F}_c \in \mathbb{R}^{1\times d}$ 表示图像 i 中所有区域经标签 c 加权后的特征，则 $\boldsymbol{F}_g=\left\{\boldsymbol{F}_1,\boldsymbol{F}_2,\cdots,\boldsymbol{F}_S\right\}$ 表示图像 i 经所有标签加权后的特征，即基于语义的视觉特征表示。

然后将 $\boldsymbol{F}_g$ 同样映射到语义空间中，可以计算所有标签在图像 i 中的置信度分数，表达式如式（13）所示：

$$\phi_i^c=\boldsymbol{F}_c\boldsymbol{\theta}^{c\,\mathrm{T}} \tag{13}$$

其中：c 表示第 c 个标签；$\boldsymbol{\theta}^c \in \mathbb{R}^{1\times d}$ 由式（8）得到。

同样规定，如果图像中存在一个标签，那么该标签在图像上的置信度分数一定大于其他不存在的标签，据此引入损失函数作为约束，对该子网络进行优化：

$$L_{\mathrm{rank2}}=\sum_i\sum_{c\in y_i,c'\notin y_i}\max\left(1+\phi_i^{c'}-\phi_i^c,0\right) \tag{14}$$

其中：y_i 表示图像 i 中所存在标签的集合；ϕ_i^c 表示标签 c 的置信度分数；$\phi_i^{c'}$ 表示标签 c' 的置信度分数。

2.2.3　两种子网络互相学习

为约束提出的两个子网络，使它们在整个训练过程中相互学习、相互促进，本文提出一种互学习损失函数。由于子网络学习到的训练经验可以通过最后输出的概率分布表现出来，所以本文将每个子网络得到的概率分布引入互学习损失，在互学习过程中让两个概率分布应尽可能接近，保持一致性。

在一般情况下，用KL散度（Kullback-Leibler divergence）来计算概率分布之间的差别，概率分布越相似，散度值就越小，表达式如下：

$$D_{\mathrm{KL}}\left(P_1(x_i)\|P_2(x_i)\right)=\sum_{i=1}^{N}P_1(x_i)\log_a\left(\frac{P_1(x_i)}{P_2(x_i)}\right) \tag{15}$$

其中：$D_{\mathrm{KL}}\left(P_1(x_i)\|P_2(x_i)\right)$表示$P_1$和$P_2$之间的散度值；$N$表示训练样本的个数；$P_1(x_i)=\{S_i^1,S_i^2,\cdots,S_i^S\}$为关联区域特征与区域特征的子网络得到的预测概率；$P_2(x_i)=\{\phi_i^1,\phi_i^2,\cdots,\phi_i^S\}$为关联区域特征与标签语义的子网络得到的预测概率。

KL散度的缺点是P_1与P_2之间的散度值和P_2与P_1之间的散度值不相等。所以本文模型采用JS散度(Jensen-Shannon divergence)作为互学习损失，JS散度为KL散度的变体，表达式如下：

$$L_{\mathrm{mutual}}=\sum_{i=1}^{N}\left[\frac{D_{\mathrm{KL}}(P_1(x_i)\|P_2(x_i))+D_{\mathrm{KL}}(P_2(x_i)\|P_1(x_i))}{2}\right] \tag{16}$$

最后，本文定义模型总的损失函数如式(17)所示：

$$L_{\mathrm{total}}=L_{\mathrm{rank\,1}}+L_{\mathrm{rank\,2}}+\lambda L_{\mathrm{mutual}} \tag{17}$$

其中：λ是一个控制互学习损失的系数。

2.2.4 多标记零样本预测

利用得到的模型对多标记零样本图像分类任务进行预测：首先从CNN网络中得到测试样本I_i的原始特征，再将原始特征分别输入到关联区域特征与区域特征的子网络和关联区域特征与标签语义的子网络，输出基于区域和基于语义的两种特征表示，将两种表示分别做映射，在语义空间中计算标签的置信度分数，得到S_i^c和ϕ_i^c。最后，本文引入一组权重$(\alpha,1-\alpha)$融合这两个子网络输出的预测值，可得到测试样本I_i的最终标签预测，表达式如下：

$$C=\underset{c\in C^U/C^{U+S}}{\arg\operatorname{top}k}\left(\alpha S_i^c+(1-\alpha)\phi_i^c\right) \tag{18}$$

其中：top$k$表示按照预测值大小排序的操作；arg top$k$表示取前k个预测值作为测试样本I_i的预测标签的操作；当$c\in C^U$时，表示标签c属于只包含未知类别的集合，即是未知标签，此时该任务属于传统多标记零样本分类；当$c\in C^{U+S}$时，表示标签c属于同时包含未知类别和已知类别的集合，即可能是未知标签也可能是已知标签，此时该任务属于广义多标签零样本分类。

3 实验结果与分析

3.1 数据集及实验细节

实验中采用多标记零样本分类常用的两个数据集NUS-WIDE[21]和MS COCO[22]。NUS-WIDE数据集中有81个人工标注的标签被用作未知类，925个用户自动标记的标签被用作已知类；本文参考文献[23]对MS COCO数据集中的标签进行划分，分成了48个已知类和17个未知类。数据集的具体信息见表1。

表1 数据集的具体信息

Table 1 Specific information of the data set 单位：个

数据集	样本总数	已知类数量	未知类数量
NUS-WIDE	129 431	925	81
MS COCO	82 783	48	17

为评估本文方法的有效性，使用mAP和每个图像的前K个预测的F1得分作为评价标准。

本文的实验配置与文献[24]保持一致，所有实验均使用预训练的VGG-19网络对图像进行特征提取。输入图片尺寸为224×224像素，本文提取最后一个卷积层输出的特征，尺寸大小为14×14×512像素，将其看作14×14个区域的特征。使用基于维基文章训练得到的GloVe[25]模型来提取标签的语义向量，其中每个标签的向量维度等于300。

本文将多头注意机制的投影头个数M设置为8。当模型训练时，在NUS-WIDE数据集上使用ADAM优化器，(β_1,β_2)设为(0.5,0.999)，学习率设为0.006，批量大小设为256，训练20轮；在MS COCO数据集上使用SGD优化器，动量值设为0.9，学习率设为0.001，批量大小设为32，训练20轮。

3.2 多标记零样本图像分类

3.2.1 NUS-WIDE数据集上的实验结果

为评估本文方法的性能，本文在NUS-WIDE数据集上做了传统多标记零样本(ZS)图像分类实验和广义多标记零样本(GZS)图像分类实验。将本文方法与Fast0Tag[24]、CONSE[26]、LabelEM[27]、One Attention per Label[17]、One Attention per Cluster[18]和LESA[18]进行对比，这些对比方法在NUS-WIDE数据集上的实验结果由本文直接引入文献[18]中的结果获得。文献[26]介绍的CONSE是最基本的零样本学习模型，利用CNN计算给定图像的预测标签，再将其输入Word2Vec模型得到对应的类别向量，最后与真实的类别向量计算相似度。文献[27]介绍的LabelEM是基于嵌入的方式解决零样本学习模型，将类别标签嵌入到给定的属性向量空间中，再引入兼容函数，计算图像特征和嵌入标签的兼容度。文献[24]介绍的Fast0Tag是最开始用于解决多标记零样本图像分类问题的模型，利用图像-标签的关联，提出对于给定图像，相关标签的词向量在词向量空间中沿着一个主方向排在不相关的标签前面，该方法通过估计图像的主方向来解决图像标记问题。表2展示了在数据集NUS-WIDE上两种分类实验的结果比较，表中加粗数字表示该组数据最大值。对于传统多标记零样本分类，LESA方法提出一种共享多注意框架，为一幅图像学习所有$K\in\{3,5\}$类别共享的多个注意力模块，得到加权注意特征后，再将注意特征映射到语义空间进行预测分类。LESA方法的分类性能在各方面都优于之前的方法。对比LESA方法，本文

方法在ZS任务上的F1（$K=3$）分数、F1（$K=5$）分数、mAP分别提高了1.4、1.1、1.9个百分点。对于广义多标记零样本分类，与LESA方法相比，本文方法的mAP、F1（$K=3$）分数、F1（$K=5$）分数分别提高了1.4、0.2、0.8个百分点。实验结果表明，在NUS-WIDE数据集上，本文方法在传统多标记零样本（ZS）图像分类实验和广义多标记零样本（GZS）图像分类实验中，性能都可以达到最佳。

表2　在NUS-WIDE数据集上的传统多标记零样本和广义多标记零样本分类性能比较

Table 2　Comparison of classification performance between traditional multi-label zero-shot and generalized multi-label zero-shot on NUS-WIDE data set　%

方法	任务	K=3			K=5			mAP
		P	R	F1	P	R	F1	
CONSE	ZS	17.5	28.0	21.6	13.9	37.0	20.2	9.4
	GZS	11.5	5.1	7.0	9.6	7.1	8.1	2.1
LabelEM	ZS	15.6	25.0	19.2	13.4	35.7	19.5	7.1
	GZS	15.5	6.8	9.5	13.4	9.8	11.3	2.2
Fast0Tag	ZS	22.6	36.2	27.8	18.2	48.4	26.4	15.1
	GZS	18.8	8.3	11.5	15.9	11.7	13.5	3.7
One Attention per Label	ZS	20.9	33.5	25.8	16.2	43.2	23.6	10.4
	GZS	17.9	7.9	10.9	15.6	11.5	13.2	3.7
One Attention per Cluster(M=10)	ZS	20.0	31.9	24.6	15.7	41.9	22.9	12.9
	GZS	10.4	4.6	6.4	9.1	6.7	7.7	2.6
LESA(M=10)	ZS	25.7	41.1	31.6	19.7	52.5	28.7	19.4
	GZS	23.6	10.4	14.4	19.8	14.6	16.8	5.6
本文方法	ZS	**26.8**	**42.9**	**33.0**	**20.5**	**54.6**	**29.8**	**21.3**
	GZS	**23.8**	**10.5**	**14.6**	**20.8**	**15.3**	**17.6**	7.0

3.2.2　MS COCO数据集上的实验结果

MS COCO数据集曾被用于多标记零样本目标检测，近年来开始用在多标记零样本图像分类任务中。将本文方法与Fast0Tag[24]、CONSE[26]、Deep0Tag[28]进行对比，这些对比方法在MS COCO数据集上的实验结果将参考文献[23]中的结果。文献[28]介绍的Deep0Tag是一种基于多示例框架来解决多标记零样本学习问题的模型，能够自动定位相关图像区域和建模图像标记（端到端），从多个尺度发现图像中场景信息，并兼顾全局和局部图像细节。表3展示了在MS COCO数据集上广义多标记零样本（GZS）图像分类的结果，主要将$K=3$处的F1分数及每个F1分数的P值和R值进行比较。

在MS COCO数据集中，参照文献[23]工作结果对已知类别和未知类别进行划分，本文模型在传统多标记零样本（ZS）图像分类任务中的性能不占优势，这是因为本文模型对一些复杂和抽象的类别如baseball bat、baseball glove、microwave、dining table、sink、fire hydrant等难以预测。但在广义多标记零样本分类中，本文模型性能依然可以达到最好。

以往提出的多标记零样本分类方法中大多基于目标检测等模块，可以在MS COCO数据集上达到较好的效果。通过对比本文方法和传统方法，发现本文方法即便不使用任何额外的检测模块，性能也可以达到最优，实验结果如表3所示。对比Deep0Tag方法，本文方法的P值、R值、F1分数分别提高了5.4、4.9、5.2个百分点。

表3　MS COCO数据集上的广义多标记零样本分类性能比较

Table 3　Comparison of classification performance of generalized multi-label zero-shot on MS COCO data set　%

方法	P	R	F1
CONSE方法	23.8	28.8	26.1
Fast0Tag方法	38.5	46.5	42.1
Deep0Tag方法	43.2	52.2	47.3
本文方法	48.6	57.1	52.5

3.3　消融实验

本文还在NUS-WIDE数据集上进行了消融实验：仅使用关联区域特征与区域特征的子网络1训练、仅使用关联区域特征与标签语义的子网络2训练，仅使用本文方法训练，将其得到的实验结果进行对比。表4展示了在传统多标记零样本（ZS）和广义多标记零样本（GZS）分类实验上三者的F1分数和mAP的对比，表中加粗数字表示该组数据最大值。对于传统多标记零样本分类，当仅使用子网络2时，相对于仅使用子网络1的F1（$K=3$）分数、F1（$K=5$）分数、mAP值分别提高了4.8、3.5、8.4个百分点。而本文方法相对于仅使用子网络2的F1（$K=3$）分数、F1（$K=5$）分数、mAP值分别提高了2.7、1.8、1.5个百分点。对于广义多标记零样本分类，当仅使用子网

络1时，相对于仅使用子网络2的F1($K=3$)分数、F1($K=5$)分数分别提高了1.1、1.0个百分点。而本文方法相对于仅使用子网络1的F1($K=3$)分数、F1($K=5$)分数分别提高了0.9、1.3个百分点。

表4 在NUS-WIDE数据集上3种方法的分类性能对比

Table 4 Comparison of classification performance of the three methods on NUS-WIDE data set %

方法	任务	F1		mAP
		(K=3)	(K=5)	
子网络1	ZS	25.5	24.5	11.4
	GZS	13.7	16.3	3.9
子网络2	ZS	30.3	28.0	19.8
	GZS	12.6	15.3	**7.5**
本文方法	ZS	**33.0**	**29.8**	**21.3**
	GZS	**14.6**	**17.6**	**7.0**

上述结果说明了关联区域特征与标签语义的子网络2在传统多标记零样本分类任务中表现更好，这是因为在传统多标记零样本分类任务中，测试数据集只包含未知标签，训练过程中只将每个标签的语义信息融入到图像区域，知识能很好地从已知标签转移到未知标签，所以在只识别未知标签的任务中表现较好；而关联区域特征与区域特征的子网络1在广义多标记零样本分类任务中表现更好，这是因为在广义多标记零样本分类任务中，测试数据集既包含已知标签又包含未知标签，将图像中各区域的特征信息相互关联之后，更容易挖掘图像中存在的标签，包括已知标签和未知标签。而本文方法在两种类型任务中的表现都能达到最好，证明了两种子网络在训练过程中进行深度互学习的有效性。

3.4 超参数分析

本文在NUS-WIDE数据集上进行实验，分析互学习损失系数λ的影响，实验结果如图2所示，其中F1_ZS_3表示在ZS分类实验中排名前三的预测结果的F1分数。对比实验结果发现，当$\lambda=0.01$时，本文模型性能达到最佳。

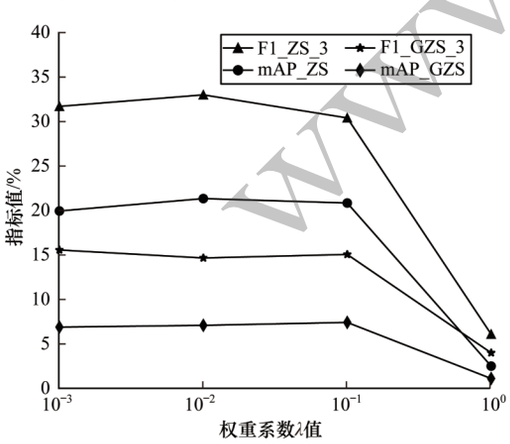

图2 不同互学习损失系数λ对模型性能的影响

Fig.2 Effect of different mutual learning loss coefficients λ on model performance

本文还通过实验分析了2个子网络权重系数组合$(\alpha,1-\alpha)$对模型预测性能的影响。在数据集NUS-WIDE上，实验结果如图3所示，对比结果发现$\alpha=0.3$即权重组合系数为(0.3,0.7)时，本文模型性能达到最佳；在数据集MS COCO上，实验结果如图4所示，对比结果发现$\alpha=0.2$即权重组合系数为(0.2,0.8)时，本文模型性能达到最佳。

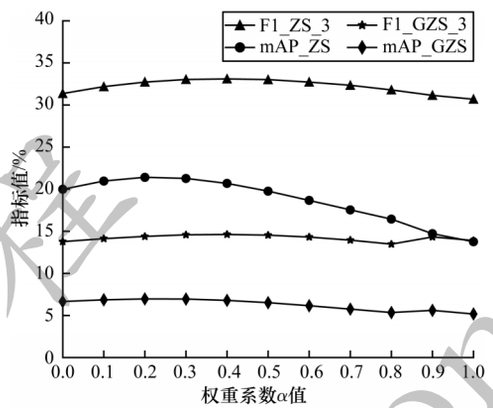

图3 NUS-WIDE数据集上不同权重系数α对模型性能的影响

Fig.3 Effect of different weight coefficients α on model performance on NUS-WIDE data set

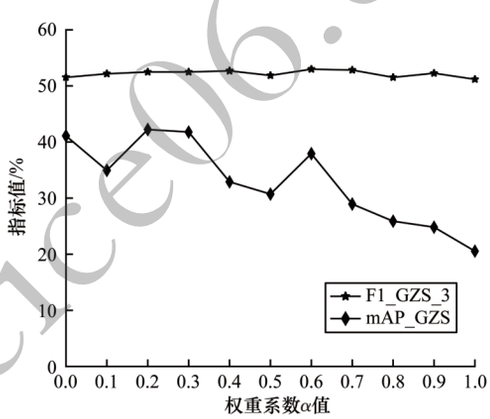

图4 MS COCO数据集上不同权重系数α对模型性能的影响

Fig.4 Effect of different weight coefficients α on model performance on MS COCO data set

4 结束语

为了解决多标记零样本图像分类问题，本文提出基于深度互学习的方法，使图像区域信息和标签语义信息同时参与到模型训练中，增强图像本身的视觉特征。建立标签与图像特征之间的关系，且在训练过程中让2个子网络互相学习对方的训练经验，互相促进。最后在对输入样本做预测时，使用一个组合权重系数融合两个子网络分别得到的预测值。本文还在两个数据集上进行传统多标记零样本分类和广义多标记零样本分类两种类型的实验，与以往研究方法的结果进行对比，证明所提方法的有效性。由于深度互学习并不局限于两个子网络进行互相学习，因此下一步也可以设计多个子网络，从不

同的研究方向和技术切入，让各个子网络做不同的工作，互相弥补、促进，提高分类性能。

参考文献

[1] BOUTELL M R，LUO J B，SHEN X P，et al. Learning multi-label scene classification[J]. Pattern Recognition，2004，37(9)：1757-1771.

[2] 蔡亚萍，杨明. 一种利用局部标记相关性的多标记特征选择算法[J]. 南京大学学报(自然科学)，2016，52(4)：693-704.
CAI Y P，YANG M. A multi-label feature selection algorithm by exploiting label correlations locally[J]. Journal of Nanjing University (Natural Science)，2016，52(4)：693-704. (in Chinese)

[3] 朱赛赛，贾修一，李泽超. 一种基于全局和局部标记相关性的多标记分类算法[J]. 电子学报，2020，48(12)：2345-2351.
ZHU S S，JIA X Y，LI Z C. Exploiting global and loca label correlations for multi-label classification[J]. Acta Electronica Sinica，2020，48(12)：2345-2351. (in Chinese)

[4] KIPF T，WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. [2022-08-10]. https://arxiv. org/pdf/1609. 02907. pdf.

[5] CHEN Z M，WEI X S，WANG P，et al. Multi-label image recognition with graph convolutional networks[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2020：5172-5181.

[6] CHEN T S，XU M X，HUI X L，et al. Learning semantic-specific graph representation for multi-label image recognition[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Washington D. C.，USA：IEEE Press，2020：522-531.

[7] WANG Z X，CHEN T S，LI G B，et al. Multi-label image recognition by recurrently discovering attentional regions[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C.，USA：IEEE Press，2017：464-472.

[8] YOU R C，GUO Z Y，CUI L，et al. Cross-modality attention with semantic graph embedding for multi-label classification[J]. Proceedings of the AAAI Conference on Artificial Intelligence，2020，34(7)：12709-12716.

[9] LIXIA X，DI J，RONGGUI W，et al. Multi-label classification based on attention mechanism and semantic dependencies[EB/OL]. [2022-08-10]. http://en. cnki. com. cn/Article_en/CJFDTotal-GDGC201909003. html.

[10] AKATA Z，PERRONNIN F，HARCHAOUI Z，et al. Label-embedding for attribute-based classification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2013：819-826.

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

[12] 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 H X，ZHANG Y. Zero-shot image classification based on generative adversarial network[J]. Journal of Beijing University of Aeronautics and Astronautics，2019，45(12)：2345-2350. (in Chinese)

[13] LI Y，SWERSKY K，ZEMEL R. Generative moment matching networks[EB/OL]. [2022-08-10]. https://arxiv. org/abs/1502. 02761.

[14] LEE C W，FANG W，YEH C K，et al. Multi-label zero-shot learning with structured knowledge graphs[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2018：1576-1585.

[15] GOA B，GYAB C，CD D，et al. Multi-label zero-shot learning with graph convolutional networks[J]. Neural Networks，2020，132：333-341.

[16] GUPTA A，NARAYAN S，KHAN S，et al. Generative multi-label zero-shot learning[EB/OL]. [2022-08-10]. https://arxiv. org/abs/2101. 11606.

[17] KIM J H，JUN J，ZHANG B T. Bilinear attention networks[EB/OL]. [2022-08-10]. https://arxiv. org/abs/1805. 07932.

[18] HUYNH D，ELHAMIFAR E. A shared multi-attention framework for multi-label zero-shot learning[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2020：8773-8783.

[19] NARAYAN S，GUPTA A，KHAN S，et al. Discriminative region-based multi-label zero-shot learning[C]// Proceedings of IEEE/CVF International Conference on Computer Vision. Washington D. C.，USA：IEEE Press，2022：8711-8720.

[20] ZHANG Y，XIANG T，HOSPEDALES T M，et al. Deep mutual learning[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2018：4320-4328.

[21] CHUA T S，TANG J H，HONG R C，et al. NUS-WIDE：a real-world web image database from National University of Singapore[C]//Proceedings of the ACM International Conference on Image and Video Retrieval. New York，USA：ACM Press，2009：1-9.

[22] LIN T Y，MAIRE M，BELONGIE S，et al. Microsoft COCO：common objects in context[C]///Proceedings of European Conference on Computer Vision. Berlin，Germany：Springer，2014：740-755.

[23] BEN-COHEN A，ZAMIR N，BEN BARUCH E，et al. Semantic diversity learning for zero-shot multi-label classification[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Washington D. C.，USA：IEEE Press，2022：620-630.

[24] ZHANG Y，GONG B Q，SHAH M. Fast zero-shot image tagging[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2016：5985-5994.

[25] PENNINGTON J，SOCHER R，MANNING C. Glove：global vectors for word representation[C]//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg，USA：Association for Computational Linguistics，2014：1532-1543.

[26] NOROUZI M，MIKOLOV T，BENGIO S，et al. Zero-shot learning by convex combination of semantic embeddings[EB/OL]. [2022-08-10]. https://arxiv. org/abs/1312. 5650.

[27] AKATA Z，PERRONNIN F，HARCHAOUI Z，et al. Label-embedding for image classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2016，38(7)：1425-1438.

[28] RAHMAN S，KHAN S，BARNES N. Deep0Tag：deep multiple instance learning for zero-shot image tagging[J]. IEEE Transactions on Multimedia，2020，22(1)：242-255.

编辑　赖玉玲