计算机科学
COMPUTER SCIENCE

基于注意力和视觉语义推理的枸杞虫害检索

韩会珍, 刘立波

引用本文

韩会珍, 刘立波. 基于注意力和视觉语义推理的枸杞虫害检索[J]. 计算机科学, 2022, 49(11A): 211200087-6.

HAN Hui-zhen, LIU Li-bo. Lycium Barbarum Pest Retrieval Based on Attention and Visual Semantic Reasoning [J]. Computer Science, 2022, 49(11A): 211200087-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211200087

基于注意力和视觉语义推理的枸杞虫害检索

韩会珍　刘立波
宁夏大学信息工程学院　银川 750021
(hhz52122@163.com)

摘　要　针对传统作物虫害检索模态单一的问题,将注意力与视觉语义推理相结合,对常见的17种枸杞虫害进行图文跨模态检索研究。首先利用Faster R-CNN+ResNet101实现注意力机制来提取枸杞虫害图像局部细粒度信息;接着,引入视觉语义推理,建立图像区域连接并采用图卷积网络(GCN)进行区域关系推理来增强区域表示;然后,进一步进行全局语义推理,选择具有判别性的特征,过滤掉不重要的内容,以捕获更多的关键语义信息;最后通过模态交互深入挖掘枸杞虫害图像和文本不同模态间的语义关联。在自建的枸杞虫害数据集上,采用平均准确率均值(MAP)作为评价指标对所提方法进行对比实验和消融实验。实验结果表明,图检文和文检图的平均*MAP*值达到了0.522,与8种主流方法相比提升了0.048~0.244,具有更好的检索效果。

关键词:跨模态检索;注意力机制;细粒度;视觉语义推理;枸杞虫害
中图法分类号　TP391

Lycium Barbarum Pest Retrieval Based on Attention and Visual Semantic Reasoning

HAN Hui-zhen and LIU Li-bo
School of Information Engineering,Ningxia University,Yinchuan 750021,China

Abstract　Aiming at the problem that traditional retrieval model on pest has a single mode,this paper uses a cross-modal retrieval method for 17 kinds of common lycium pests in image and text modal,which integrates attention mechanism and visual semantic reasoning. First,use Faster R-CNN+ResNet101 to realize the attention mechanism to extract local fine-grained information of wolfberry pest images. Then further introduce vision semantic reasoning to build the image region connections and use convolutional network GCN for region relation reasoning to enhance area representation. In addition,global semantic reasoning is performed by enhancing semantic correlation between regions,selecting discriminant features and filtering out unimportant information to capture more key semantic information. Finally,the semantic association between different modalities of lycium barbarum pest image and text is deeply explored through modal interaction. On the self-built lycium barbarum pest dataset,the average accuracy(MAP) is used as the evaluation index to carry out comparative experiment and ablation experiment. Experimental results demonstrate that the averaged *MAP* of the proposed method in the self-built lycium pest dataset achieves 0.522,compared with the eight mainstream methods,the average *MAP* of the method improves by 0.048 to 0.244,and it has better retrieval effect.

Keywords　Cross-modal retrieval,Attention mechanism,Fine-grained,Visual semantic reasoning,Lycium barbarum pest

1　引言

宁夏枸杞产业是全区九大重点产业之一,种植面积广,产品远销多个国家地区。但在栽植和生产过程中,虫害问题严重影响了宁夏枸杞产业的发展。随着互联网的迅速发展,病虫害图像、文本等多模态数据呈爆炸式增长[1-3],图像和文本两种模态数据经常同时产生、相互关联并相互补充。Fan[4]以马铃薯和柑橘病虫害图像为研究对象,提出了基于关键特征点的病虫害ROI快速自动检测算法。Chen等[5]针对林业信息检索智能性不高,提出了基于林业病虫害领域本体的语义检索模型。Li等[6]提出了关键字、主题词等字词匹配的枸杞病虫害信息资源检索方法,提升了检索质量。以上方法在农作物虫害检索中取得了很好的成效,但存在检索模态单一的问题,不能很好地将虫害不同模态信息进行展示。

如何通过计算机视觉等先进信息技术实现图像和文本信息间跨模态检索[7-12],对满足日益增长的病虫害多样化检索具有重要研究意义。Hotelling等[13]利用典型相关分析(Canonical Correlation Analysis,CCA),最大化不同模态间的相关性构建公共语义空间,进而实现相似性度量。Andrew等[14]提出深度典型关联分析方法,通过充分学习数据的非线性特征来构建更有效的公共语义空间。Wang等[15]采用基于对抗学习的跨模态检索方法,保留数据间语义的可区分性。Zhen等[16]采用深度监督的跨模态学习体系结构,以弥合不同模式之间的异质性差距。Huang等[17]将每个图像区域

基金项目:国家自然科学基金(61862050);宁夏自然科学基金(2020AAC03031)
This work was supported by the National Natural Science Foundation of China(61862050) and Ningxia Natural Science Fundation of China (2020AAC03031).
通信作者:刘立波(liulib163.com)

分类为对象和语义关系的多标签，可以捕获局部区域内的语义概念。Lee 等[18]分别在图像和文本两端利用注意力机制以更好地推断图像区域和单词之间潜在对应关系。研究图像和文本的跨模态检索，主要探索图文不同模态数据间的语义相关性[19-20]。上述方法取得了很好的检索结果，但未考虑图像局部区域和全局语义关联。由于枸杞虫害图像中虫害面积占比小，图像比文本具有更多的细粒度信息，而文本又包含了很多比图像更强的语义描述，因此如何更好地提取局部细粒度信息进而关联全局语义信息十分重要。

通过上述分析，针对农业领域的枸杞虫害信息检索模态单一的问题，本文以常见的尺蠖、负泥虫、红长蝽等 17 种虫害为主要研究对象，结合图文跨模态检索技术，采用注意力机制和视觉语义推理对枸杞虫害图像进行局部区域和全局语义关联，使最终的图像表示能捕获更多关键语义信息（如 cicadella viridis 和 green leaves 等对象概念，以及 eating the leaves, lying on the stalk, broken 等高级语义），有助于实现更好的图片文本匹配性能。本文主要对枸杞虫害图像处理展开研究。

2 模型设计

本文模型框架如图 1 所示，主要由 3 个部分组成，分别是图像编码模块、文本编码模块以及模态交互模块。

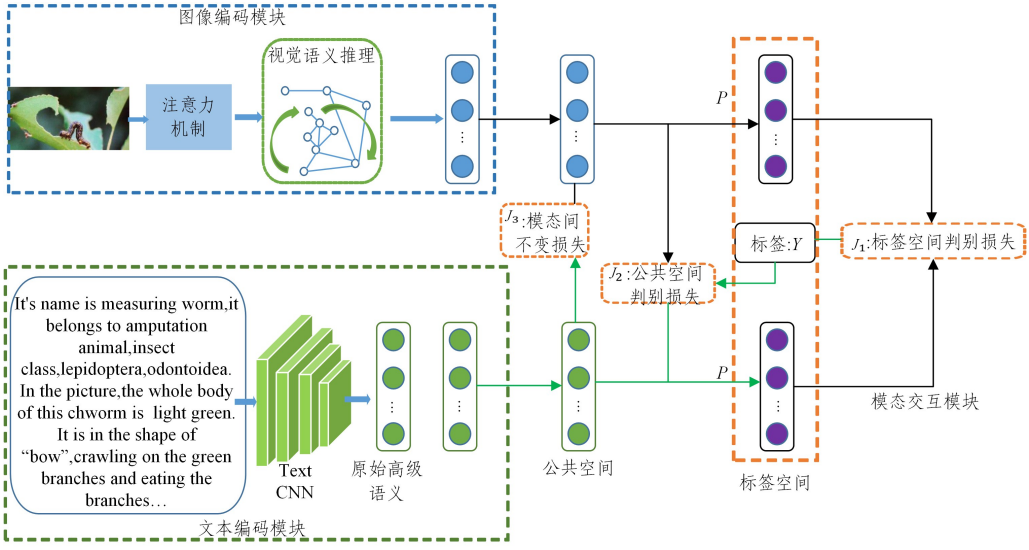


图 1 模型框架图

Fig. 1 Modal framework diagram

对于图像模态，结合枸杞虫害图像特点，首先采用注意力机制对图像进行局部区域特征提取，以更好地获取目标虫害局部细粒度信息。接着，引入视觉语义推理探索枸杞虫害图像局部和全局区域的语义相关性，进行区域关系推理来增强区域表示；然后进一步进行全局语义推理来增加整个枸杞虫害图像场景的描述，得到图像的最终表示。针对文本模态，使用在 Google News 上预训练的 word2vec 模型得到给定文本的词向量序列，每个词向量为 300 维，生成原始的文本高级语义表示，获得较为全面的上下文语义信息。然后，最小化模态间不变损失，采用权值共享策略消除多模态差异，同时在模态交互模块最小化标签空间和公共表示空间的判别损失，使学习的共同表征具有显著区分性，从而提升跨模态检索的性能。

本文对枸杞虫害数据进行符号化定义，首先令 $\boldsymbol{V}=[v_1, v_2, \cdots, v_n]$，$U=[u_1, u_2, \cdots, u_n]$，$Y=[y_1, y_2, \cdots, y_n]$ 分别表示为图片表示矩阵、文本表示矩阵和标签表示矩阵。$\boldsymbol{v}_i$ 是在公共空间学习到第 i 个实例的图片表示，u_j 是在公共空间学习到第 j 个实例的文本表示。

2.1 枸杞虫害图像局部细粒度信息提取

由于枸杞虫害图像背景大多是绿色，且虫害面积相对较小，获取局部目标枸杞虫害图像信息是一个挑战。为了获取更有效的虫害图像局部细粒度特征，引入注意力机制能很好地专注于枸杞虫害图像局部细粒度部分。本文主要采用在视觉基因组上预训练的 Faster R-CNN 目标检测方法结合 ResNet101 来实现自下而上的注意力机制，提取图像局部区域特征。针对图像获取兴趣区域，对每个兴趣区域应用目标检测器，得到边界框特征和对应的视觉特征，进而更好地检测到目标虫害等区域特征，捕获更多局部细粒度信息。每个图像可以用一组特征 $\boldsymbol{V}=\{v_1, \cdots, v_k\}$，$v_i \in \boldsymbol{R}^D$ 表示，采用自下而上的注意力使得每个特征 v_i 对该图像中的目标对象或局部区域进行编码，学习具有丰富语义的视觉特征表示。最后，在平均池化层之后提取特征，输出图片表示的局部细粒度特征向量。其结构如图 2 所示。

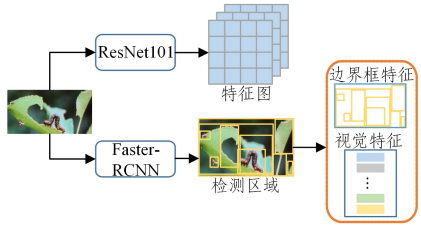


图 2 枸杞虫害图片注意力结构图

Fig. 2 Attention structure diagram of lycium barbarum pest picture

2.2 视觉语义推理的引入

为了更进一步获取关联局部细粒度信息和全局语义信息的视觉特征，针对枸杞虫害图片虫害姿态、图像背景、不同生长时期身体特征等特点，基于枸杞虫害图像局部细粒度信息，加入视觉语义推理进行局部和全局语义关联，学习图像中的对象概念和高级语义关系，使得最终图像表示能捕获更多的

关键语义信息。视觉语义推理结构如图 3 所示。

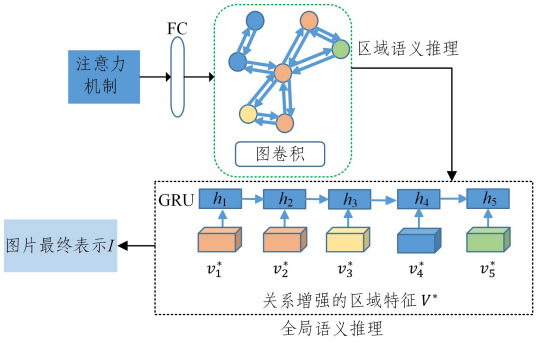


图 3　基于注意力的视觉语义推理结构图

Fig. 3　Diagram of attention-based visual semantic inference structure

2.2.1　区域关系推理

为了深入探索枸杞虫害图像局部区域间的语义相关性,针对枸杞虫害图像特点,建立一个区域关系推理模型,通过考虑图像区域之间的语义相关性来增强基于区域的表示。

首先,通过测量嵌入空间中图像区域之间的成对亲和力来构建区域间关系,计算每对区域的亲和边来得到亲和矩阵 $\boldsymbol{R}$。如果存在连接这两个图像区域的高亲和分数的边,则这两个图像区域具有强语义关系并且高度相关,表达式如下:

$$\boldsymbol{R}_{(v_i,v_j)}=\varphi(v_i)^{\mathrm{T}}\phi(v_j) \tag{1}$$

其中,$\varphi(v_i)=W_\varphi v_i$,$\phi(v_j)=W_\phi v_j$ 是两个嵌入函数,权重参数 W_φ 和 W_ϕ 通过反向传播学习。

接着,构建一个全连接关系图 $G_r=(V,E)$,V 是被检测区域的集合,边缘集 E 由亲和矩阵 $\boldsymbol{R}$ 描述。

为了更好地获取区域关系增强表示,采用图卷积网络(Graph Convolutional Network,GCN)对全连接图进行推理,利用图关系定义的邻节点计算得到各个节点的响应,然后在原始 GCN 上添加残差连接,表达式如下:

$$V^*=\boldsymbol{W}_r(\boldsymbol{R}V\boldsymbol{W}_g)+V \tag{2}$$

其中,$\boldsymbol{W}_g$ 是维度为 $D\times D$ 的图卷积层权重矩阵,$\boldsymbol{W}_r$ 是残差结构的权重矩阵,对亲和矩阵 $\boldsymbol{R}$ 逐行归一化,输出 $V^*=\{v_1^*,\cdots,v_k^*\}$,$v_i^*\in\boldsymbol{R}^D$ 是图像区域节点关系增强后的表示。

2.2.2　全局语义推理

基于对枸杞虫害图像区域关系推理获取的增强表示进一步进行全局语义推理,选择具有判别性的特征来过滤掉不重要的内容,从而捕获更多的关键语义信息。将区域特征序列 $V^*=\{v_1^*,\cdots,v_k^*\}$,$v_i^*\in\boldsymbol{R}^D$ 依次输入到 GRUs 模型中,在推理过程中,整个场景的描述将会在记忆单元 m_i 中逐渐增长和更新。i 是推理步骤,更新门 z_i 主要分析当前输入区域特征 v_i^* 和整个场景的描述 m_{i-1} 来决定更新其记忆单元的程度。更新门 z_i 的计算式如下:

$$z_i=\sigma_z(W_z v_i^*+U_z m_{i-1}+b_z) \tag{3}$$

其中,σ_z 是 sigmoid 激活函数,W_z,U_z 和 b_z 是权重和偏置。

新增内容会增加整个枸杞虫害图片场景的描述,表达式如下:

$$\widetilde{m}_i=\sigma_m(W_m v_i^*+U_z(r_i\circ m_{i-1})+b_m) \tag{4}$$

其中,σ_m 是 tanh 激活函数,W_m,U_z 和 b_m 是权重和偏差,$\circ$ 是逐元素乘法。根据 v_i^* 和 m_{i-1} 之间的推理来过滤掉不重要的内容。r_i 的计算类似更新门,表达式如下:

$$r_i=\sigma_r(W_r v_i^*+U_r m_{i-1}+b_r) \tag{5}$$

其中,σ_r 是 sigmoid 激活函数,W_r,U_r 和 b_r 是权重和偏差。

当前步骤的整个场景 m_i 的描述是在先前描述 m_{i-1} 和新内容 $\widetilde{m}_i$ 之间使用更新门 z_i 的线性插值,表达式如下:

$$m_i=(1-z_i)\circ m_{i-1}+z_i\circ\widetilde{m}_i \tag{6}$$

由于每个 v_i^* 都包含全局语义推理信息,对 m_i 的更新实际上是同时考虑当前局部区域和全局语义相关性,将序列 V^* 末端的记忆单元作为整个图像的最终表示 I。

在区域关系推理的基础上,本文进一步进行全局语义推理,选择判别信息,逐渐增加对枸杞虫害图像场景的关键语义描述,进而生成全局表示。

2.3　模态语义对齐

在获得了包含关键语义信息的枸杞虫害图像特征和原始文本的高级语义特征之后,在图像和文本子网络中分别建立两个全连接层,并对第二层的全连接层建立权值共享约束来学习图像和文本的交叉模态关联。然后,通过线性分类器 P 预测在公共表示空间中投影的样本的语义标签。在图像模态网络和文本模态网络的顶部连接一个线性层,该分类器获取公共空间中训练数据的表示,并为每个样本生成 c 维向量的预测标签。标签空间判别损失计算式如下:

$$J_1=\frac{1}{n}\|\boldsymbol{P}^T V-Y\|_{\mathrm{F}}+\frac{1}{n}\|\boldsymbol{P}^T-Y\|_{\mathrm{F}} \tag{7}$$

其中,$\|\cdot\|_{\mathrm{F}}$ 表示 Frobenius 范数,$\boldsymbol{P}$ 是线性分类器的映射矩阵。

此外,在公共表示空间直接测量来自枸杞虫害图文的所有样本的判别损失,公共空间判别损失的计算式如下:

$$\begin{aligned}J_2=&\underbrace{\frac{1}{n^2}\sum_{i,j=1}^{n}(\log(1+\mathrm{e}^{\Gamma_{ij}})-S_{ij}^{\alpha\beta}\Gamma_{ij})+}_{\text{inter-modalities}}\\&\underbrace{\frac{1}{n^2}\sum_{i,j=1}^{n}(\log(1+\mathrm{e}^{\Phi_{ij}})-S_{ij}^{\alpha\alpha}\Phi_{ij}+}_{\text{image modality}}\\&\underbrace{\frac{1}{n^2}\sum_{i,j=1}^{n}(\log(1+\mathrm{e}^{\Theta_{ij}})-S_{ij}^{\beta\beta}\Theta_{ij})}_{\text{text modality}}\end{aligned} \tag{8}$$

其中,$\Gamma_{ij}=\frac{1}{2}\cos(v_i,u_j)$,$\Phi_{ij}=\frac{1}{2}\cos(v_i,v_j)$,$\Theta_{ij}=\frac{1}{2}\cos(u_i,u_j)$,$S_{ij}^{\alpha\beta}=1\{v_i,u_j\}$,$S_{ij}^{\alpha\alpha}=1\{v_i,v_j\}$,$S_{ij}^{\beta\beta}=1\{u_i,u_j\}$,$\cos(\cdot)$ 是用于计算两个输入向量之间相似度的余弦函数,$1\{\cdot\}$ 是一个函数,如果两个元素是类内样本的表示,则该函数值为 1,否则为 0。式(8)中第一项是模态间样本相似性的负对数似然,似然函数定义公式如下:

$$p(S_{ij}^{\beta\beta}\mid u_i,v_j)=\begin{cases}\delta(\Gamma_{ij}), & \text{if } S_{ij}^{\alpha\beta}=1\\1-\delta(\Gamma_{ij}), & \text{otherwise}\end{cases} \tag{9}$$

其中,$\delta(\Gamma_{ij})=\frac{1}{1+\mathrm{e}^{-\Gamma_{ij}}}$ 是 sigmoid 函数,最小化这个负对数似然函数等于最大化似然。余弦相似度 $\cos(v_i,u_j)$ 越大,$p(1\mid v_i,u_j)$ 就会越大,这就意味着样本归类为相似。

为了消除跨模态差异,最小化所有图像-文本对表示之间的距离。采用的模态不变性损失可被表示为:

$$J_3=\frac{1}{n}\|V-U\|_{\mathrm{F}} \tag{10}$$

结合式(7)、式(8)和式(10)得到最终损失函数计算式如下：

$$J=J_1+\lambda J_2+\eta J_3 \tag{11}$$

其中，超参数 λ 和 η 控制最后两个分量的贡献，n 是输入实例的数量。

综上，本文针对枸杞虫害图像特点，首先采用注意力机制提取虫害图像局部细粒度特征，接着引入视觉语义推理关联图像局部和全局语义信息，获取枸杞虫害图像关键语义信息，然后挖掘不同模态间语义相关关系来提高枸杞虫害图文跨模态检索性能。

3 结果与分析

3.1 实验环境与实现细节

实验中使用的平台操作系统为 Ubuntu 16.04LTS，GPU 为 NVIDIA Quadro p5000。采用 Python3.6.12 语言编程，结合深度学习框架 Pytorch1.10，gensim3.8.0 及 Caffe 实现算法。

模型中有两个子网络，一个用于图像模态，一个用于文本模态。对于图像模态，将枸杞虫害图像输入到预训练的 Faster R-CNN+ResNet101 来实现自下而上的注意力，模型被训练得到图像中的对象概念和局部细粒度信息。接着，采用 GCN 构建全连接图，得到区域关系增强表示 V^*，进而将区域特征序列输入到 GRUs 模型中以保留重要的信息，并过滤掉不重要的信息，得到图片最终表示 I。对于文本模态，使用在 Google News 上预训练的 word2vec 模型得到文本的词向量序列，每个词向量为 300 维，之后输入到 Text CNN 来得到文本原始高级语义表示。最后，在每个子网络设置两个具有 ReLU 激活函数的全连接层，两层的隐藏单元数分别为 2048 和 1024，共享两个子网络第二层的权值来学习图片和文本模态的相关性。实验训练采用 Adam 优化器，训练时将 α 设置为 0.001，β 设置为 0.1，学习率为 0.0001，训练次数是 500 epochs。

3.2 实验数据预处理

实验数据集以常见的 17 种枸杞虫害的图片为研究对象，首先通过团队实地调研拍照、书本收集以及网络爬虫技术共获取 1900 张枸杞虫害图像样本，均为.jpg 格式。然后参照跨模态检索常用的 Wikipedia 数据集文本结构为基准，并结合枸杞虫害特色以枸杞虫害学术命名、分布范围、形态特征、生活习性、危害特点、防治方法和图片整体视觉内容等为主要描述，利用网络渠道并借助专家力量为每类枸杞虫害中所有图片撰写对应英文文本描述，得到 1900 个图像-文本对。由于数据集不足以支撑模型训练，故采用数据增强的方式分别对图片和文本进行数据扩充。对图片采用亮度、对比度、旋转、翻转进行数据增强，对文本采用 EDA(Easy Data Augmentation)[21]数据增强方式。数据增强后共包含 9500 个枸杞虫害图片-文本对。所有样本被划分为 17 个类别，每种虫害对应一个类别。以跨模态检索常用的 Wikipedia 数据集为参考，构建枸杞虫害图像-文本对，按照 8∶2 的比例划分数据集，随机选取 7600 个图像-文本对作为训练集，1900 个图像-文本对作为测试集。枸杞虫害数据集示例如图 4 所示。

图 4 枸杞虫害数据集示例

Fig. 4 Example of lycium barbarum pest data set

3.3 评价指标

为了验证本文方法的有效性，在枸杞虫害数据集上实现两种跨模态检索任务对模型进行衡量：1)使用枸杞虫害图像查询检索枸杞虫害文本(I2T)；2)使用枸杞虫害文本查询检索枸杞虫害图片(T2I)。

本文采用跨模态检索研究中广泛使用的性能评价标准均值平均精度(Mean Average Precision，MAP)作为枸杞虫害图文跨模态检索的评价指标。MAP 是对每个类别的平均精度 AP 再求平均值，取值范围为 0～1，取值越大，表示准确率越高，AP 的计算式如下：

$$AP=\frac{1}{R}\sum_{k=1}^{n}\frac{R_k}{k}\times rel_k \tag{12}$$

其中，R 表示测试集中与给定查询样本相似的样本数量，n 表示测试集的样本总数，R_k 表示前 k 个检索结果中与查询样本相似的样本数量，当第 k 个检索结果与查询样本相似时，rel_k 为 1，反之则为 0。

3.4 对比方法

本文方法与 8 种跨模态检索主流方法在枸杞虫害图文数据集上进行了对比，其中包括 3 种传统方法，分别是典型相关分析 CCA[13]、核典型相关分析(Kernel Canonical Correlation Analysis，简称 KCCA)[22]、跨模态联合表示学习方法(Learning Cross-media Joint Representation With Sparse And Semisupervised Regularization，JRL)[23]，5 种深度学习方法，分别是深度典型相关分析(Deep Canonical Correlation Analysis，DCCA)[14]、深度语义匹配(Deep Semantic Matching，Deep-SM)[24]、对抗式跨模态检索(Adversarial Cross-Modal Retrieval，ACMR)[15]、2019 年提出的深度监督跨模态检索(Deep Supervised Cross-modal Retrieval，DSCMR)[16]以及 2021 年提出的深度关联相似性学习(Deep Relational Similarity Learning For Cross-modal Retrieval，DRSL)[25]。

传统的跨模态检索主要对文本和图片之间的检索进行研究，将不同模态的特征投影到一个共同的潜在子空间，然后在该子空间中最小化不同模态样本对之间的距离来学习投影子空间，从而实现跨模态检索。CCA 利用典型相关分析将文本和图片从各自原来的空间映射到公共子空间。KCCA 是 CCA 的一种扩展，它使用核函数将特征投影到一个高维空间，能更好地处理特征集合非线性的情景。JRL 主要是架构数据间的语义关联及数据间对应的语义标签整合到一起，统一优化构建公共子空间。深度学习的方法是利用深度学习的特征抽取能力，在底层提取不同模态的有效表示，在高层建立不同模态的语义关联，进而提出 DCCA 框架，主要用来学习两种模态之间的非线性转换，使结果是高度线性相关的。Deep-SM 提出一种深层语义匹配的方法来解决带一个或多个

标签的样本的跨模态检索问题。ACMR 将生成对抗的方法引入实值跨模态检索来探索更有效的公共子空间。DSCMR 采用一个深度监督的跨模态学习体系结构,以弥合不同模式之间的异质性差距。DRSL 是通过关系网络进行文本和图片的关系交互,直接学习不同模态数据对之间的相似性来消除异质性差距。本文方法与对比方法的结果如表 1 所列。

表 1　在枸杞虫害数据集上不同方法的结果对比

Table 1　Comparsion of different methods on lycium barbarum pest data set

Method	I2T	T2I	Average
CCA[13]	0.281	0.275	0.278
KCCA[22]	0.359	0.313	0.336
DCCA[14]	0.386	0.377	0.382
JRL[23]	0.403	0.380	0.392
Deep-SM[24]	0.427	0.391	0.409
ACMR[15]	0.455	0.430	0.443
DSCMR[16]	0.463	0.459	0.461
DRSL[25]	0.482	0.465	0.474
Our Method	0.535	0.509	0.522

表 1 中列出了在两个检索任务上 I2T 和 T2I 的 *MAP* 值及平均 *MAP* 值。从表中数据可以看出,本文方法超过了所有对比方法,其在 I2T 和 T2I 任务上 *mAP* 分别为 0.535 和 0.509,平均 *MAP* 为 0.522,在两个检索任务上与对比方法的最高性能相比分别提高了 5.3%和 4.4%,体现了本文方法有较好的检索性能。

采用准确率-召回率(Precision-Recall)曲线进一步验证方法的性能,Recall 为横坐标,Precision 为纵坐标。其中,Precision 值越大表明该方法的性能越好。

枸杞虫害图像检索枸杞虫害文本(I2T)和枸杞虫害文本检索枸杞虫害图片(T2I)的 Precision-Recall 曲线如图 5 和图 6 所示。

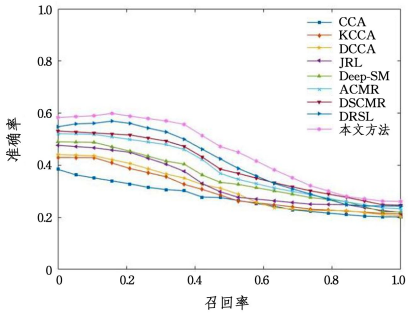


图 5　I2T 的 Precision-Recall 曲线

Fig. 5　I2T's Precision-Recall curve

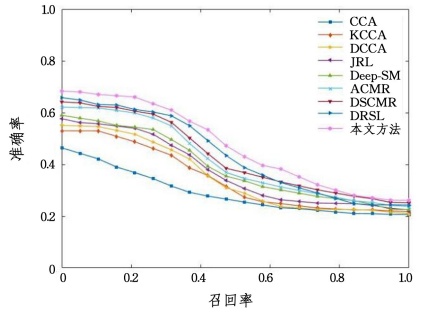


图 6　T2I 的 Precision-Recall 曲线

Fig. 6　T2I's Precision-Recall curve

由图 5 可以看出,本文方法的明显优于其他对比方法。图 6 中,召回率值在 0～0.6 时,本文方法的准确率明显高于其他方法,召回率在 0.6～1 时,本文方法的准确率与其他方法的最高性能基本持平。综合来看,在枸杞虫害图文数据集上,本文方法图检文和文检图性能均优于其他对比方法。

3.5　消融实验

为了验证注意力机制和视觉语义推理对本文方法各个部分的影响,采用消融实验进行实验分析。消融实验结果如表 2 所列。其中 A 表示注意力机制模块,V 表示视觉语义推理模块,NA 表示不包含注意力机制模块,NV 表示不包含视觉语义推理模块,Ours(A+V)表示本文方法。

表 2　在枸杞虫害数据集上消融实验的 *MAP* 结果

Table 2　*MAP* results of ablation experiments on lycium barbarum pest data set

Method	I2T	T2I	Average
NA+NV	0.463	0.459	0.461
A+NV	0.516	0.460	0.488
Ours(A+V)	0.535	0.509	0.522

从表 2 的结果可以看出,引入注意力机制模块可以更好地提取到枸杞虫害局部细粒度信息,视觉语义推理模块能更好地推理局部关联全局的关键语义信息,从而更好地进行模态间语义交互。

表 3 列出了基于不同特征维度的 *MAP* 值,探索不同特征维度对模型性能的影响。分别将隐空间设置为 512 维、1024 维、2048 维进行实验。从表 3 可以看出,维度设置为 2048 维时模型性能最好,且各特征维度表现出的性能差异较小,表明模型稳定性较好。

表 3　特征维度对模型的影响

Table 3　Influence of feature dimension on model

Feature dimension	I2T	T2I	Average
512 维	0.529	0.491	0.510
1024 维	0.523	0.505	0.516
2048 维	0.535	0.509	0.522

结束语　针对传统虫害检索模态单一的问题,本文以 17 种常见枸杞虫害为研究对象,提出了基于注意力和视觉语义推理的枸杞虫害图文跨模态检索方法。通过注意力机制对枸杞虫害图像进行局部细粒度信息提取,进而引入视觉语义推理关联局部和全局语义信息,更好地提取枸杞虫害图像关键语义信息。实验在枸杞虫害图文数据集上对本文方法和 8 种主流方法进行了对比实验,本文方法在图检文和文检图的平均 *MAP* 值达到了 0.522,与 8 种主流方法相比提升了 0.048～0.244,体现了本文方法的优越性。下一步将针对文本检索图像的准确率提升和枸杞虫害的中文文本描述的图文检索做进一步研究。

参 考 文 献

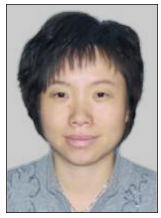
[1]　HE H X. Research on the Pain Points and Paths of My Country's Smart Agriculture Development in the Internet Era[J]. Agricultural Economics,2021(6):15-17.

[2]　LIU Q P. Occurrence and control techniques of major diseases and insect pests of Chinese wolfberry[J]. Modern Horticulture,2018(12):42.

[3]　CHANG X. Agricultural and forestry pests and diseases and meteorological information remote monitoring system [D]. Beijing:

Beijing University of Technology, 2020.
[4] FAN Z J. Research and implementation of image retrieval methods for crop diseases and insect pests[D]. Mianyang: Southwest University of Science and Technology, 2018.
[5] CHEN Z F. Research on Semantic Retrieval of Forestry Diseases and Pests Domain Ontology [D]. Harbin: Northeast Forestry University, 2017.
[6] LI G F, LI W J. A semantic retrieval model based on the domain ontology of wolfberry diseases and insect pests[J]. Computer Technology and Development, 2017, 27(9): 48-52.
[7] OU W H, LIU B, ZHOU Y H, et al. A review of cross-modal retrieval research[J]. Journal of Guizhou Normal University(Natural Science Edition), 2018, 36(2): 118-124.
[8] GONG Y, KE Q, ISARDM, et al. A multi-view embedding space for modeling internet images, tags, and their semantics[J]. International Journal of Computer Vision, 2014, 106(2): 210-233.
[9] ZHAI X, PENG Y, XIAO J. Learning cross-media joint representation with sparse and semisupervised regularization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2013, 24(6): 965-978.
[10] RASIWASIA N, COSTA PEREIRA J, COVIELLO E, et al. A new approach to cross-modal multimedia retrieval[C]// Proceedings of the 18th ACM International Conference on Multimedia. New York: ACM Press, 2010: 251-260.
[11] RASHTCHIAN C, YOUNG P, HODOSH M, et al. Collecting image annotations using amazon's mechanical turk[C]// Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Stroudsburg: Association for Computational Linguistics Press, 2010: 139-147.
[12] 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.
[13] HOTELLING H. Relations between two sets of variates[M]// Kotz S, Johnson N L Breakthroughs in statistics. New York: Springer Press, 1992: 162-190.
[14] ANDREW G, ARORA R, BILMES J, et al. Deep canonical correlation analysis[C]// Proceedings of the 30th International Conference on Machine Learning. Atlanta: Machine Learning Research Press, 2013: 1247-1255.
[15] WANG B, YANG Y, XU X, et al. Adversarial cross-modal retrieval[C]// Proceedings of the 25th ACM International Conference on Multimedia. New York: ACM Press, 2017: 154-162.
[16] ZHEN L, HU P, WANG X, et al. Deep supervised cross-modal retrieval[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 10394-10403.
[17] HUANG Y, WU Q, SONG C F, et al. Learning semantic concepts and order for image and sentence matching[C]// CVPR. 2018.
[18] LEE K H, CHEN X, HUA G, et al. . Stacked cross attention for image-text matching[C]// ECCV. 2018.
[19] LI Z X, LING F, ZHANG C L, et al. Cross-media image text retrieval based on two-level similarity [J]. Chinese Journal of Electronics, 2021, 49(2): 268-274.
[20] LI K, ZHANG Y, LI K, et al. Visual Semantic Reasoning for Image-Text Matching[C]// 2019 IEEE/CVF International Conference on Computer Vision(ICCV). 2019: 4653-4661.
[21] WEI J, ZOU K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks[J]. arXiv: 1901. 11196, 2019.
[22] HARDOON D R, SZEDMAK S, SHAWE-TAYLOR J. Canonical correlation analysis: An overview with application to learning methods[J]. Neural Computation, 2004, 16(12): 2639-2664.
[23] ZHAI X, PENG Y, XIAO J. Learning cross-media joint representation with sparse and semisupervised regularization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2013, 24(6): 965-978.
[24] WEI Y, ZHAO Y, LU C, et al. Cross-modal retrieval with CNN visual features: A new baseline[J]. IEEE transactions on cybernetics, 2016, 47(2): 449-460.
[25] WANG X, HU P, ZHEN L, et al. DRSL: Deep Relational Similarity Learning for Cross-modal Retrieval [J]. Information Sciences, 2021, 546: 298-311.

HAN Hui-zhen, born in 1995, postgraduate. Her main research interests include information retrieval and so on.



LIU Li-bo, born in 1974, Ph.D, professor, is a member of China Computer Federation. Her main research interests include intelligent information processing and so on.