doi:10.6041/j.issn.1000-1298.2022.02.032

融合注意力机制的枸杞虫害图文跨模态检索方法

刘立波　赵斐斐
（宁夏大学信息工程学院，银川 750021）

摘要：针对现有农作物病虫害检索模态较为单一问题，以 17 种常见的枸杞虫害图像和文本描述为研究对象，将跨模态检索引入枸杞虫害检索领域，提出一种融合注意力机制的枸杞虫害图文跨模态检索方法。首先，借助 Transformer 模型和循环神经网络分别获取带有上下文信息的细粒度图像和文本特征序列；然后，利用注意力机制对特征序列进行聚合以挖掘图像和文本的显著性语义信息；最后，为了深入挖掘不同模态间语义关联，采用跨媒体联合损失函数对模型进行约束。试验结果表明，本文方法在自建的枸杞虫害图文跨模态数据集上平均精度均值平均值达到了 0.458。与现有的 8 种方法相比，平均精度均值平均值提高了 0.011 ~ 0.195，优于所有对比方法，可为农作物病虫害多样化检索提供技术支撑和算法参考。

关键词：枸杞虫害；注意力机制；图文检索；跨模态
中图分类号：S567.19；TP391.41　**文献标识码**：A　**文章编号**：1000-1298(2022)02-0299-10　**OSID**：

Cross-modal Image and Text Retrieval Method for *Lycium barbarum* Pests by Integrating Attention Mechanism

LIU Libo　ZHAO Feifei
(*School of Information Engineering*, *Ningxia University*, *Yinchuan* 750021, *China*)

Abstract: In recent years, with the change of climatic conditions and the introduction of cultivation techniques, the planting area of Lycium has gradually expanded. It has become one of the important economic crops in Ningxia and even the entire northwestern region. Lycium is a multi-insect host and has poor resistance to insect pests. It is very susceptible to insect infestation, which has a huge impact on yield and quality, causing serious economic losses. Therefore, it is very important to quickly and accurately retrieve and obtain various information about Lycium pests and provide timely and accurate control for the development of the industry. To address the problem that the present retrieval system on crop pests owns only the single mode, the cross-modal retrieval for images and texts in Lycium pest dataset was introduced, which had 17 kinds of common pests, and a cross-modal image and text retrieval method with the attention mechanism was proposed. Firstly, the transformer and the LSTM were used to obtain text and image fine-grained feature sequences with the context information, respectively. Then, the attention mechanism was leveraged to aggregate feature sequences to capture the salient semantic information in texts and images. Finally, in order to explore the semantic correlation between different modalities, the cross-media joint loss was used to constrain the proposed model. The experiment showed that the averaged MAP of the proposed method in the self-built Lycium pest dataset achieved 0.458. Compared with the existing eight methods, the averaged MAP of the method was improved by 0.011 ~ 0.195, outperforming all these methods. The proposed method can provide technical support and algorithm reference for diversified retrieval requirements of crop pests.

Key words: *Lycium barbarum* pests; attention mechanism; image and text retrieval; cross-modal

0　引言

枸杞具有免疫调节、滋肾、润肺、补肝等功效，在国内外市场备受青睐。同时，作为防风固沙和改良盐碱地的先锋树，枸杞兼具生态与经济价值，随着气候条件的变化和栽培技术的引进，近年来种植面积

收稿日期：2020-12-28　修回日期：2021-07-09
基金项目：国家自然科学基金项目(61862050)和宁夏自然科学基金项目(2020AAC03031)
作者简介：刘立波(1974—)，女，教授，博士后，主要从事人工智能及农业信息化研究，E-mail：liulib@163.com

逐渐扩大[1],已成为宁夏乃至整个西北地区重要的经济作物之一[2-3]。枸杞属于多虫寄主且抗虫害能力较差,极易遭受虫害侵扰,并呈现逐年加重趋势,对于产量及品质影响巨大,造成了严重的经济损失。因此快速准确检索得到枸杞虫害多方面信息并给予及时精准防治,对于避免虫害进一步扩散进而提高枸杞产量与品质,推进枸杞产业带动区域经济发展至关重要。

传统的农作物病虫害检索主要通过人眼查看病虫害目标区域的颜色、纹理、虫子体态等特征,与农作物病虫害图像信息手册进行人工对比来实现[4]。该方法依赖个人经验以及肉眼观察,导致主观性强、误判率高并且耗费时间和精力[5]。随着精准农业和智慧农业的发展,农作物病虫害信息量爆炸式增长,其数据也因自身特点呈现多模态形式,图像和文本两种模态数据经常同时产生、相互关联并互相补充。如何通过计算机视觉、图像处理等先进信息技术,从这些不同模态且语义关联的数据中获取有价值的信息,进而实现图像文本信息间的跨模态检索,对满足人们日益增长的农作物病虫害信息多样化检索需求具有重要意义。

现有研究[4,6-8]在农作物病虫害检索任务中都取得了很好的成效,但均存在检索模态单一的问题,即仅能够以图像检索图像,或者以文本检索文本,很难将农作物病虫害不同模态信息进行展示。随着农业数据化信息及形式的多样化[9],研究人员更加注重不同模态信息的互检及模态的综合分析,而跨模态检索(Cross-modal retrieval)正是兼具多模态数据之间的相似互检这一特性,并融合图像、文本等多个模态对数据进行高效互查与量化,使其不断成为多媒体信息检索中的一个研究热点[10],被广泛应用在医疗、交通、艺术等领域[11-12]。在农业领域,由于农作物病虫害信息模态更加多样化[13],图像或文本的单模态检索显然已经不能满足人们的需求,对于经验不足的人员,仅凭农作物病虫害图像、文本等单模态信息并不能全面且直观、形象地了解想要检索的内容[14]。跨模态检索能够实现不同模态之间信息互检,获得更加多元化的农作物病虫害信息,从而对农作物病虫害的及时防治提供帮助。但目前跨模态检索尚未在农业领域应用,因此将跨模态检索引入农业领域实现农作物病虫害的跨模态检索更能满足农业发展的现实需求。

由于不同模态在进行某些特定特征与语义交互学习时,往往存在细节信息不互补或者高级语义不平衡的现象,导致模态间的映射关系不对等,造成不同模态间特征描述缺失或者语义关联匮乏。比如,枸杞虫害图像和枸杞虫害文本之间的模态学习,图像具有比文本更多的细节信息,而文本又包含了很多比图像更强的语义描述。因而,为了解决上述问题,通过引入具有模拟人类视觉系统功效的注意力机制,能够更加突出图像与文本中更具区分性的重点部位,来缓解这种模态间的不对等以及不平衡性。

本文以 17 类枸杞虫害图像和与其相对应的枸杞虫害文本为研究对象,针对现有方法检索模态单一的问题,将跨模态检索技术引入枸杞虫害检索中,利用注意力机制对图像和文本数据进行特征提取,使模型能够集中于图像和文本中必要细粒度部分,学习图像与文本的显著性语义信息,从而挖掘两者之间的语义关联,针对枸杞虫害的图文跨模态检索,期望获得更高的实时性和更丰富的内容,为农作物病虫害检索提供新思路。

1 融合注意力机制的图文跨模态检索模型

1.1 数据符号化表示及模型框架概览

本文对数据进行符号化定义,首先令 $V=\{v_i\}_{i=1}^{n}$ 表示图像数据集,令 $T=\{t_i\}_{i=1}^{n}$ 表示文本数据集,且 t_i 为 v_i 的文本描述,共同组成相似图像文本对 v_i-t_i。除此之外,令 $\{c_i\}_{i=1}^{n}$ 表示图像文本对所属类别,其中 $c_i=[b_{i1},b_{i2},\cdots,b_{im}]$,$m$ 为枸杞虫害类别总数,b_{ij} 取 1 则表示第 i 个图像文本对属于第 j 个类别,取 0 则表示不属于该类别。

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

图 1 中,y_j 为单词特征,α_j 为单词注意力权重,d_i^t 为文本 t_i 对应注意力特征,s_i^t 为文本 t_i 对应最终特征,h_j 为图像子区域特征,β_j 为图像子区域注意力权重,d_i^v 为图像 v_i 对应注意力特征,s_i^v 为图像 v_i 对应最终特征。

对于图文跨模态检索任务来说,给定任一相似的图像文本,其中的内容往往只存在一部分相似性,不可能完全相似,该任务的这一特点便促使模型需要首先将数据拆分为多个部分,探索数据不同部分之间的语义关联,进而挖掘数据中所包含的细粒度信息。再者,在本文自建的枸杞虫害数据集中,这样的局部相似性往往集中在图像中包含害虫的区域,以及文本中包含对虫害特点进行描述的部分,这又进一步要求模型能够提取图像和文本中的显著性语义信息。本文首先通过文本编码模块和图像编码模块获取各数据的细粒度特征序列,并基于注意力机制对序列进行聚合以获取文本和图像的显著性语义特征,接着通过模态交互模块提高文本和图像特征的判别力,并对文本和图像特征进行语义对齐,保证

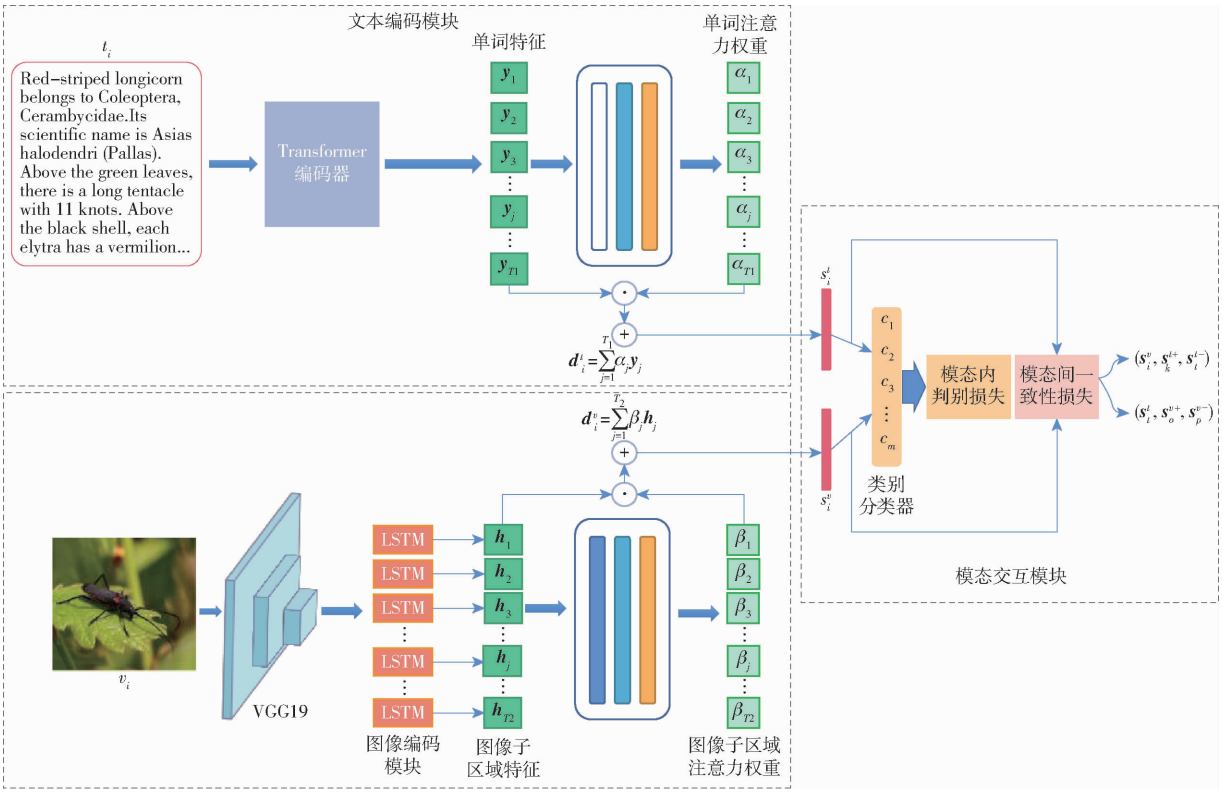


图 1　模型框架图

Fig. 1　Frame of model

文本数据和图像数据之间的模态间一致性。

对于文本模态，首先通过 word2vec 方式来获取文本中每个单词的词向量作为文本的细粒度特征序列。接着通过 Transformer 模型获取包含文本上下文信息的细粒度特征序列，使序列中每一个元素既包含其本身独有的信息，又包含与整个数据的关系，增强可判别性。然后通过注意力机制获取序列中每个元素的注意力权重，即每个元素对数据的重要性，并基于所得权重对序列元素进行加权求和以得到包含了显著性语义信息的文本特征。同样地，对于图像模态，首先通过 VGG19 网络提取该网络最后一个池化层的特征图谱，并将其拆分为 49 个子区域特征，形成图像的细粒度特征序列。接着通过 LSTM 网络获取包含图像上下文信息的细粒度图像特征序列。然后通过注意力机制以同样的方式得到图像的特征表示。最后，通过模态交互模块的模态内判别损失以及模态间一致性损失来共同引导模型的训练。

1.2　文本编码模块

在处理枸杞虫害文本内容时，采用 Transformer 编码器对文本进行编码，Transformer[15] 是 Google 团队在 2017 年提出的一种自然语言处理（NLP）经典模型，是一种新的、基于注意力机制来实现的特征提取器，可以用于代替卷积神经网络（CNN）和循环神经网络（RNN）来提取序列的特征，其结构如图 2 所示。图中，N 为 Transformer 的层数。

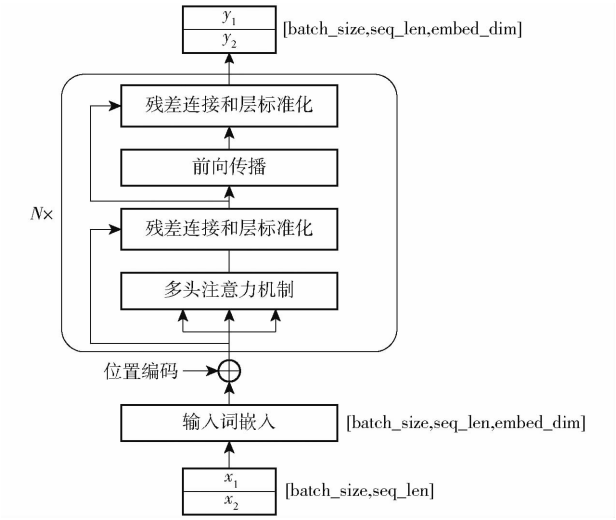


图 2　Transformer 模型的编码器结构

Fig. 2　Encoder structure of Transformer model

以 1 层 Transformer 模型为例，该模型可以被简单表示为

$$\boldsymbol{y}_i = \text{transformer}(\boldsymbol{x}_i) \tag{1}$$

式中　$\boldsymbol{x}_i$——输入的词向量

$\boldsymbol{y}_i$——输入的词向量经过模型编码后的输出向量

在对文本信息进行编码时，对于给定的一个枸杞虫害文本 $\boldsymbol{t}_i$，设每个文本 $\boldsymbol{t}_i$ 中包含 T_1 个单词，可表示为 $\boldsymbol{t}_i = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_{T_1}]$。在本文所提数据集文

本语料库中添加 Wikipedia 语料库，基于 skip-gram 与 Negative Sampling 策略构建 word2vec 模型，将 $\boldsymbol{t}_i$ 中的单词 $\boldsymbol{w}_j$ 转换为一个词向量，记为 $\boldsymbol{x}_j$，得到该文本的细粒度特征序列，再将得到的序列送入 Transformer 编码器中以获取序列中每个单词包含了文本上下文信息的特征向量 $\boldsymbol{y}_j$，得到包含了文本上下文信息的细粒度特征序列 $Y_i=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_j,\cdots,\boldsymbol{y}_{T_1}]$。具体公式为

$$\boldsymbol{x}_j=\text{word2vec}(\boldsymbol{w}_j)\quad(j\in[1,T_1]) \tag{2}$$

$$\boldsymbol{y}_j=\text{transformer}(\boldsymbol{x}_j)\quad(j\in[1,T_1]) \tag{3}$$

接着采用类似的方法[16]实现注意力机制，将 Transformer 编码器输出的文本特征序列 Y_i 送入前馈神经网络中，然后利用 softmax 函数计算得到序列中每个 $\boldsymbol{y}_j$ 的注意力权重 α_j，对应的注意力权重序列可被表示为 $\alpha=[\alpha_1,\alpha_2,\cdots,\alpha_{T_1}]$，具体计算公式为

$$\alpha=\text{softmax}(W_{ta}Q^t) \tag{4}$$

$$Q^t=\tanh(W_a^tY_i) \tag{5}$$

其中 W_a^t 和 W_{ta} 为各自层的权重。在得到文本 $\boldsymbol{t}_i$ 每个单词的注意力权重后，对所有单词的特征向量与其注意力权重进行加权求和得到最后的文本特征 $\boldsymbol{d}_i^t$（维度为 256）。特征序列中每一元素的注意力权重体现了其在整个文本中的重要性，最后的文本特征也因此能够突出文本内容中的显著部分。$\boldsymbol{d}_i^t$ 的具体计算公式为

$$\boldsymbol{d}_i^t=\sum_{j=1}^{T_1}\alpha_j\boldsymbol{y}_j \tag{6}$$

1.3 图像编码模块

在处理枸杞虫害图像时，首先将图像 v_i 的尺寸调整为 256 像素×256 像素，并将其输入到 VGG19 网络中获取该网络最后一个池化层的特征图谱，该特征图谱的尺寸为 7×7×512，3 个参数分别表示特征图谱的高、宽以及通道数，由此将该特征图谱看作图像 49（7×7）个子区域对应的特征，每个子区域可被表示为 512 维的特征向量，将这 49 个子区域连接起来便可构成图像的特征序列，可以表示为 $\boldsymbol{r}=[\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_j,\cdots,\boldsymbol{r}_{T_2}]$，$T_2$ 为图像区域总数，$\boldsymbol{r}_j$ 为第 j 个区域所对应的特征向量。然后利用长短期记忆网络[17]（Long short term memory，LSTM）获取包含了图像上下文信息的细粒度特征序列 $H_i=[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_j,\cdots,\boldsymbol{h}_{T_2}]$。

LSTM 是一种特殊的循环神经网络，通过记忆单元学习长期依赖关系和更新门的能力较强，同时保留了之前的时间步长信息，能够有效解决一般的 RNN 存在的长期依赖问题，图 3 为 LSTM 计算单元内部结构。

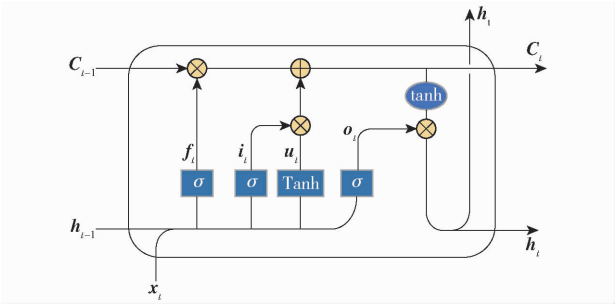


图 3 LSTM 单元架构

Fig. 3 Architecture of LSTM unit

LSTM 单元计算公式为

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_i[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_i) \tag{7}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f) \tag{8}$$

$$\boldsymbol{u}_t=\tanh(\boldsymbol{W}_C[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_C) \tag{9}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_o[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_o) \tag{10}$$

$$\boldsymbol{C}_t=\boldsymbol{f}_t\boldsymbol{C}_{t-1}+\boldsymbol{i}_t\boldsymbol{u}_t \tag{11}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\tanh(\boldsymbol{C}_t) \tag{12}$$

其中 $\boldsymbol{f}_t$ 为遗忘门，表示 $\boldsymbol{C}_{t-1}$ 的哪些特征被用于计算 $\boldsymbol{C}_t$，$\boldsymbol{f}_t$ 是一个向量，向量的每个元素均位于[0,1]范围内；$\boldsymbol{u}_t$ 表示单元状态更新值，由输入数据 $\boldsymbol{x}_t$ 和隐节点 $\boldsymbol{h}_{t-1}$ 经由一个神经网络层得到，单元状态更新值的激活函数通常使用 tanh。$\boldsymbol{i}_t$ 为输入门，同 $\boldsymbol{f}_t$ 一样也是元素介于[0,1]区间内的向量，由 $\boldsymbol{x}_t$ 和 $\boldsymbol{h}_{t-1}$ 经由 Sigmoid 计算得到。$\boldsymbol{i}_t$ 用于控制 $\boldsymbol{u}_t$ 的哪些特征用于更新 $\boldsymbol{C}_t$，使用方式与 $\boldsymbol{f}_t$ 相同。最后为了计算预测值和生成下个时间片完整的输入，需要计算隐节点的输出 $\boldsymbol{h}_t$，$\boldsymbol{h}_t$ 由输出门 $\boldsymbol{o}_t$ 和单元状态 $\boldsymbol{C}_t$ 得到，其中 $\boldsymbol{o}_t$ 计算方式与 $\boldsymbol{f}_t$ 和 $\boldsymbol{i}_t$ 相同。σ 为 Sigmoid 激活函数，其定义为

$$\sigma(x)=\frac{1}{1+\mathrm{e}^{-x}} \tag{13}$$

接着同样利用注意机制将从 LSTM 得到的特征序列 H_i 送入前馈神经网络中，利用 softmax 函数计算对应的注意力权重序列 $\beta=[\beta_1,\beta_2,\cdots,\beta_{T_2}]$，计算公式为

$$\beta=\text{softmax}(W_{va}Q^v) \tag{14}$$

$$Q^v=\tanh(W_a^vH_i) \tag{15}$$

其中 W_a^v 和 W_{va} 为各自层的权重。则最后包含图像显著性语义信息的特征 $\boldsymbol{d}_i^v$（维度为 512）的计算公式为

$$\boldsymbol{d}_i^v=\sum_{j=1}^{T_2}\beta_j\boldsymbol{h}_j \tag{16}$$

1.4 模态交互模块

在获得包含显著性语义信息的枸杞虫害文本和图像特征之后，通过全连接网络将得到的文本和图像数据表征 $\boldsymbol{d}_i^t$ 及 $\boldsymbol{d}_i^v$ 分别映射到同一个隐空间中，

记为 $\boldsymbol{s}_i^t$ 和 $\boldsymbol{s}_i^v$（维度均为 1 024），并利用余弦相似度计算文本与图像之间相似度距离。

在整个模态交互模块中，采用类似文献[18]中提出的跨媒体联合损失函数，对枸杞虫害图像和文本两种媒体类型数据间的语义关联进行约束。

使映射入隐空间后的文本与图像特征向量通过一个分类器，进而通过模态内判别损失约束模型训练过程，使得到的图像和文本特征在各自模态内保持语义类别方面的可判别性。该分类器以图像和文本特征作为输入，预测各特征所属语义类别的概率分布，本文通过计算特征真实标签与所得概率分布间的交叉熵来构建模态内判别损失，具体公式为

$$L_{\text{sem}} = -\frac{1}{n}\sum_{i=1}^{n}\left(c_i\left(\log\hat{p}_i(\boldsymbol{s}_i^v)+\log\hat{p}_i(\boldsymbol{s}_i^t)\right)\right) \tag{17}$$

式中　L_{sem}——所有图像文本对语义类别分类的交叉熵损失

n——图像文本对总数

c_i——真实类别标签

$\hat{p}_i$——分类器预测得到图像文本对中每一项（图像或文本）属于某一类别所对应的概率分布

引入模态间一致性损失来保证相似图像文本对在特征空间中的距离足够近，不相似图像文本对在特征空间中的距离足够远，最终确保不同模态间数据特征表示的一致性。以图像数据集中每幅图像 $\boldsymbol{s}_i^v$ 为基点，从文本数据集中分别随机抽取与其相似（具有相同语义标签）的文本 $\boldsymbol{s}_k^{t^+}$ 及与其不相似（具有不同语义标签）的文本 $\boldsymbol{s}_l^{t^-}$，构建图像三元组样本集 $\{(\boldsymbol{s}_i^v,\boldsymbol{s}_k^{t^+},\boldsymbol{s}_l^{t^-})\}_i$ 的样本对，同时以文本数据集中每个文本 $\boldsymbol{s}_i^t$ 为基点，以相同方式构建文本三元组样本集 $\{(\boldsymbol{s}_i^t,\boldsymbol{s}_o^{v^+},\boldsymbol{s}_p^{v^-})\}_i$。分别以图像以及文本三元组样本集作为输入，通过最小化每个三元组样本中相似样本对之间的距离，同时最大化不相似样本对之间的距离来保证图像和文本模态间的一致性。具体的模态间一致性损失为

$$L_{\text{related}}^v = \sum_{i,k,l}\left[D(\boldsymbol{s}_i^v,\boldsymbol{s}_k^{t^+})+\lambda\max(0,\mu-D(\boldsymbol{s}_i^v,\boldsymbol{s}_l^{t^-}))\right] \tag{18}$$

$$L_{\text{related}}^t = \sum_{i,o,p}\left[D(\boldsymbol{s}_i^t,\boldsymbol{s}_o^{v^+})+\lambda\max(0,\mu-D(\boldsymbol{s}_i^t,\boldsymbol{s}_p^{v^-}))\right] \tag{19}$$

式中　L_{related}^v——图像模态间一致性损失

L_{related}^t——文本模态一致性损失

λ——平衡参数　　μ——边缘约束

$D(\cdot,\cdot)$——两向量间的余弦距离

以 $D(\boldsymbol{s}_i^v,\boldsymbol{s}_i^t)$ 为例，其计算公式为

$$D(\boldsymbol{s}_i^v,\boldsymbol{s}_i^t) = \frac{\boldsymbol{s}_i^v\boldsymbol{s}_i^t}{\|\boldsymbol{s}_i^v\|_2\|\boldsymbol{s}_i^t\|_2} \tag{20}$$

总的模态间一致性损失 L_{related} 由 L_{related}^v 和 L_{related}^t 组成，记为

$$L_{\text{related}} = L_{\text{related}}^v + L_{\text{related}}^t \tag{21}$$

因此，模态交互模块中跨媒体联合损失函数 L 可表示为

$$L = \varepsilon_1 L_{\text{sem}} + \varepsilon_2 L_{\text{related}} \tag{22}$$

其中超参数 ε_1 和 ε_2 用于平衡各损失项在训练时对模型的影响。

综上，本文首先引入注意力机制提取枸杞虫害图像与文本数据自身所蕴含的显著性语义信息，接着通过最小化跨媒体联合损失函数来探索枸杞虫害图像与文本特征间的语义关联，挖掘不同模态间语义相关关系，最终达到提升枸杞虫害图文跨模态检索准确率的目的。

2　数据预处理

2.1　试验数据准备与数据集构建

以尺蠖、大青叶蝉、负泥虫、木虱、蚜虫、蓟马等 17 种常见枸杞虫害为研究对象，通过实地调研拍照、书本收集以及网络爬虫技术共获取 9 380 幅包含 17 类枸杞虫害图像样本，图像样本均为.jpg 格式。根据图文跨模态检索数据集构建的需要，充分利用网络渠道并借助专家力量为每类枸杞虫害中所有虫害图像撰写对应的文本描述，图 4 为自建枸杞虫害数据集部分类别图像及文本示例。给 17 类枸杞虫害分配所属类别标签，标签 0 为尺蠖，标签 1 为大青叶蝉，标签 2 为负泥虫，以此类推到标签 16 蛀果蛾。以跨模态检索常用的 Wikipedia 数据集结构为基准，构建枸杞虫害图像-文本对列表，按照 8∶2 的比例将自建的枸杞虫害数据集划分训练集与测试集。Wikipedia 数据集图像-文本对列表格式为“文本名称 图像名称 所属类别标签”，自建的枸杞虫害数据集图像-文本对列表格式为“图像相对路径 文本相对路径 所属类别标签”。

2.2　数据增强与扩充

针对自建枸杞虫害数据集学习样本少，在复杂网络中容易发生过拟合问题，本研究采用数据增强技术对原始数据集进行扩充。数据增强可使原始数据集更具多样性，从而减少过拟合现象，进一步提升训练模型的泛化能力。

在处理图像样本时，通过对原始图像进行垂直翻转、调整亮度、随机裁剪以及旋转操作扩增枸杞虫害图像样本，部分图像扩增前后结果如图 5

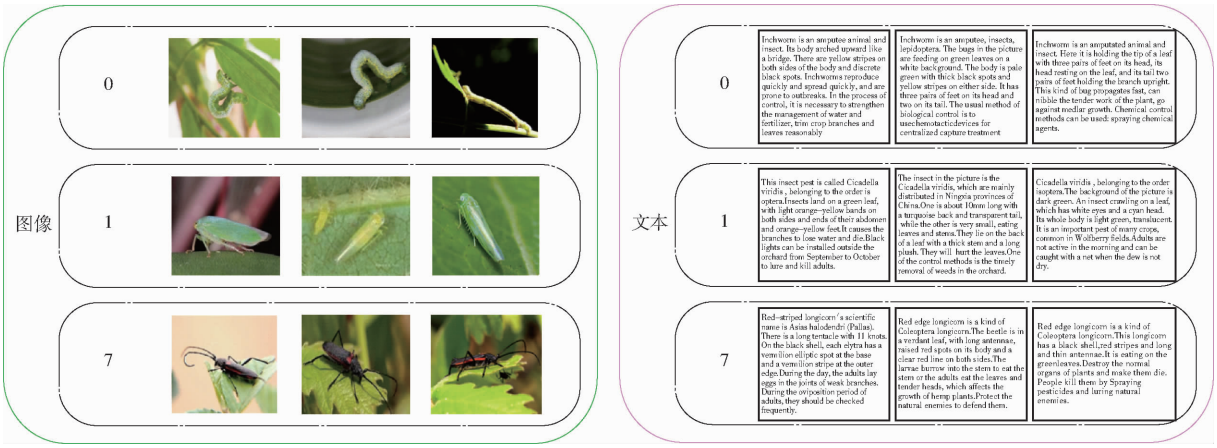


图 4 自建枸杞虫害数据集部分类别图像及对应文本示例

Fig. 4 Some category images and corresponding text examples of self-built Lycium pest dataset

图 5 枸杞虫害图像数据增强操作

Fig. 5 Lycium pests image data enhancement operation

所示。

在处理文本样本时,现有自然语言处理的数据样本增强扩充主要有 2 种方法:加噪和回译。加噪即在原始文本数据基础之上通过替换词、删除词等方式生成新的与原始数据相似的文本样本;回译即将原始文本数据翻译为其他语言,再将翻译得到的结果再次翻译回原始语言。本文采用文本分类任务的简单数据增强(Easy data augmentation for text classification tasks,EDA)方法通过加噪的思路对文本进行处理,该方法中主要的加噪形式有同义词替换、随机插入、随机交换以及随机删除,其处理效果如图 6 所示。

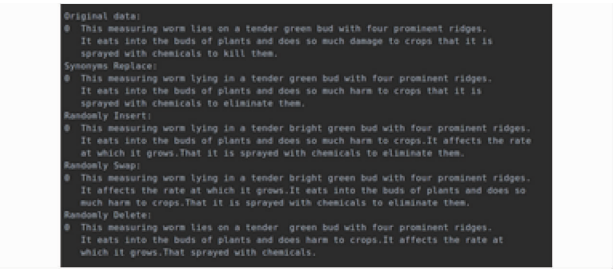


图 6 枸杞虫害文本数据增强操作

Fig. 6 Lycium pests text data enhancement operation

将扩增后的文本内容与扩增后的图像相对应,更新枸杞虫害样本内容,得到扩增后的枸杞虫害样本集,按照 8∶2的比例划分训练集与测试集,并用新得到的枸杞虫害数据集代替原始数据集进行后续试验。

3 试验与结果分析

3.1 试验环境设置

试验在宁夏大学高性能计算平台上进行,平台操作系统为 Ubuntu 16.04 LTS。加载软件环境有 gcc、cuda 9.0 和 Python 3.6.10。GPU 为 NVIDIA quadro p 5000。采用深度学习框 Tensorflow 1.13.1。

3.2 评价指标和对比方法

为了充分验证本文方法的可行性及准确率,在自建的枸杞虫害数据集上,使用枸杞虫害图像检索枸杞虫害文本以及枸杞虫害文本检索枸杞虫害图像 2 个任务对模型准确率进行衡量。

采用平均精度均值(Mean average precision,MAP)和准确率(Precision)-召回率(Recall)曲线作为枸杞虫害跨模态检索准确率的评价指标。

相同试验环境下通过试验对比了 2 种传统跨模态检索方法:典型相关分析[19](Canonical correlation analysis, CCA)和核典型相关分析[20](Kernel canonical correlation analysis,KCCA),以及 6 种基于深度神经网络的方法:深度典型相关分析[21](Deep canonical correlation analysis,DCCA)、端到端的深度典型相关分析[22](End-to-end DCCA)、深度语义匹配[23](Deep semantic matching,Deep-SM)、通信自编码器[24](Correspondence autoencoder,Corr-AE)、对抗式跨模态检索[25](Adversarial cross-modal retrieval,ACMR)、特定模态的跨模态相似度测

量[26]（Modality-specific cross-modal similarity measurement，MCSM）。传统的跨模态检索方法 CCA 通过学习映射矩阵，最大化公共空间中不同模态投影特征之间的相关性。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，能更好的处理特征集合非线性的情景。DCCA 是 CCA 的一个非线性延伸，能够同时学习 2 个数据视图间的非线性投影，使得到的映射特征高度非线性相关。End-to-end DCCA 采用 GPU 以及减少过拟合的方法可以应对原始 DCCA 框架的不足。Deep－SM 采用 2 种不同的卷积神经网络进行深度语义匹配，实现跨模态检索。Corr－AE 由 2 个耦合在编码层的自编码器网络组成，可以同时对重构误差和相关损耗进行建模。ACMR 在不同模态之间互相作用获得一个有效的共享子空间，能够有效解决一个模态的一项数据可能存在多个语义不同项的问题。MCSM 为不同模态数据构建独立的语义空间，通过端到端框架直接从每个语义空间生成特定于模态的跨模态相似度。为公平对比，所有对比方法的图像端输入均为从预训练 VGG19 网络中提取的 4096 维深度特征，文本端则首先通过 word2vec 提取词向量然后将文本中所有词向量的平均值作为输入。除了 CCA、KCCA、DCCA、End-to-end DCCA 为 20 维，其他所有方法最后的特征维数均为 1 024。

3.3　结果对比

本文方法与所对比方法的结果如表 1 所示。实验结果表明，本文方法的平均精度均值平均值达到了 0.458。不论通过图像检索文本，还是通过文本检索图像，本方法的平均精度均值均高于对比方法。在所有对比方法中，MCSM 方法表现最好，而本文方法相较于该方法平均精度均值平均值提高了 0.011。ACMR 方法和 Corr－AE 方法效果相当。本文方法比 Deep－SM、End-to-end DCCA 以及 DCCA 方法的平均精度均值平均值分别提高了 0.045、0.069、0.074，而 CCA 方法的平均精度均值平均值只有 0.263。

表 1　不同方法平均精度均值结果对比

Tab.1　Comparison of results of different methods

方法	图像检索文本	文本检索图像	平均值
本文方法	0.482	0.433	0.458
MCSM	0.466	0.427	0.447
ACMR	0.442	0.414	0.428
Corr－AE	0.431	0.420	0.426
Deep－SM	0.420	0.405	0.413
End-to-end DCCA	0.401	0.377	0.389
DCCA	0.403	0.364	0.384
KCCA	0.378	0.370	0.374
CCA	0.274	0.252	0.263

现有的跨模态检索方法大多将来自其自身特征空间的不同模态的数据平均投影到一个单一的公共空间中，以找到它们之间的潜在对齐方式，学习它们之间的内在联系。但这些方法只能粗略捕捉枸杞虫害图像与文本之间的对应关系，无法探索图像与文本数据中的细粒度信息。而本文方法有针对性地融合注意力机制对枸杞虫害图像和文本分别进行处理，能够充分挖掘枸杞虫害图像与文本之间复杂的跨媒体关联，从而提高检索准确率。

为进一步证实本文方法的有效性，在枸杞虫害数据集上的 Precision－Recall 曲线如图 7 和图 8 所示。由图 7 可知，本文方法在图像检索文本任务中性能明显优于其他对比方法。对于图 8 中文本检索图像任务，召回率取 0.3～1.0 时，本文方法的准确率略低于某些对比方法，但基本与性能最优方法持平，而在召回率取 0～0.3 时，本文方法的准确率明显高于其他对比方法。综合来看，本文方法在文本检索图像任务中的检索性能也优于其他对比方法。

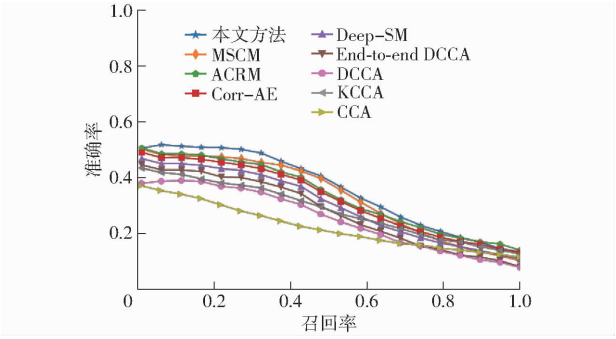


图 7　图像检索文本任务的 Precision－Recall 曲线

Fig.7　Precision－Recall curves of image to text task

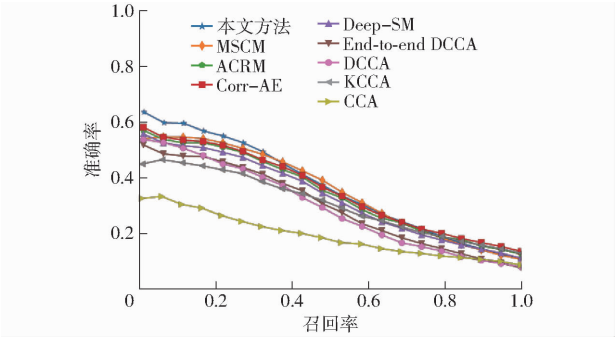
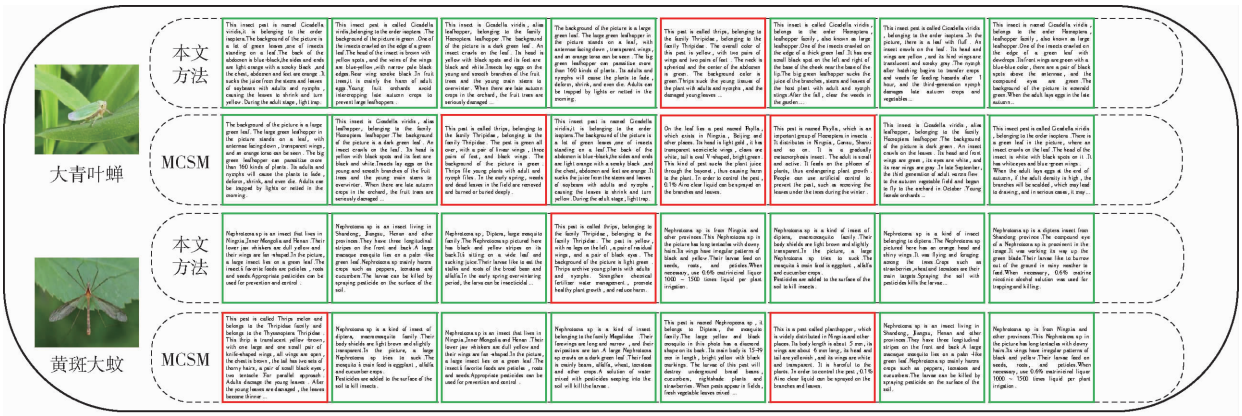


图 8　文本检索图像任务的 Precision－Recall 曲线

Fig.8　Precision－Recall curves of text to image task

图 9 给出了本文方法与 MCSM 方法在自建的枸杞虫害数据集上的跨模态检索结果对比示例，带有绿色边框的表示正确的检索结果，带有红色边框的表示错误结果。以大青叶蝉的图像检索文本任务为例，在本文方法检索得到的前 8 个文本中，检索正



(a) 图像检索文本

(b) 文本检索图像

图 9 本文方法与比较方法 MCSM 在自建枸杞虫害数据集上的跨模态检索结果示例

Fig. 9 Examples of cross-modal retrieval results of proposed method and comparison method MCSM on self-built Lycium pest dataset

确的有 7 个,错误的有 1 个,而通过 MSCM 方法检索得到正确结果有 5 个,错误结果有 3 个。从图 9 中可看出,不论在图像检索文本任务上,还是在文本检索图像任务上,本文方法的跨模态检索表现均略优于 MCSM。

3.4 消融试验

为了进一步验证注意力机制对所提方法各个部分的影响,本文进行了消融试验,结果如表 2 所示。其中 I 表示图像编码模块,T 表示文本编码模块,A 表示图像或文本编码模块中包含注意力机制,NA 则表示不包含注意力机制。从表 2 可以看出,融合注意力机制的模型可以突出枸杞虫害图像和文本内容中较为重要的细粒度局部信息,更好地为 2 种模态间的关联关系建模,提高检索准确率。

另外,为进一步证明模型的鲁棒性,本文探索了隐空间中特征维度对本文模型的影响。在构建网络时分别将隐空间设置为 256、512、1 024 维并进行试验,结果如表 3 所示。从表 3 可以看出,在特征维度取 1 024 时,模型性能最佳,但在特征维度取 256 或 512 时,模型性能并没有出现显著下降。

表 2 自建枸杞虫害数据集上基线试验的平均精度均值结果

Tab. 2 MAP results of baseline experiment on self-constructed *Lycium barbarum* pest dataset

方法	图像检索文本	文本检索图像	平均值
本文方法(I(A)+T(A))	0.482	0.433	0.458
I(A)+T(NA)	0.460	0.416	0.438
I(NA)+T(A)	0.438	0.429	0.434
I(NA)+T(NA)	0.377	0.361	0.369

表 3 不同特征维度的平均精度均值结果

Tab. 3 MAP results of different feature dimensions

特征维度	图像检索文本	文本检索图像	平均值
256	0.465	0.407	0.436
512	0.469	0.428	0.449
1 024	0.482	0.433	0.458

4 结论

(1)针对现有农作物病虫害识别与检索方法识别或检索模态较为单一的问题,本文以枸杞尺蠖、大青叶蝉、负泥虫、木虱、蚜虫、蓟马等共 17 类枸杞虫害为研究对象,提出了一种融合注意力机制的枸杞虫害图文跨模态检索方法。根据跨模态检索任务需

要，构建枸杞虫害数据集，然后通过图像编码模块以及文本编码模块分别对图像和文本信息进行细粒度处理，经过模态交互模块中损失函数的约束，深入挖掘不同模态间的语义相关关系，实现跨模态检索任务，并在自建的枸杞虫害数据集上对本文方法以及一些经典方法的性能进行了对比分析。

（2）提出的融合注意力机制的枸杞虫害图文跨模态检索模型，将跨模态检索引入枸杞虫害检索中，为枸杞虫害多模式数据检索提供了有效而强大的方法，相比于传统的基于单模式的技术更加方便且检索结果更加直观、丰富。

（3）通过在模型中融入注意力机制，能够挖掘数据中的细粒度信息，捕捉数据的显著性语义信息，从而提升检索性能，与 8 种现有方法相比，本文方法平均精度均值平均值提高了 0.011 ~ 0.195，优于所有对比方法。

参 考 文 献

[1] 李坤梁. 差异化竞争视角下中宁枸杞产业的出口策略研究[J]. 商场现代化，2018(14)：64 - 66.
LI Kunliang. Research on the export strategy of Zhongning wolfberry industry from the perspective of differentiated competition [J]. Market Modernization, 2018(14):64 - 66. (in Chinese)

[2] 许盼盼. 枸杞抗盐种质资源筛选与抗盐基因的克隆鉴定[D]. 杨凌：西北农林科技大学，2018.
XU Panpan. Screening of salt-tolerant germplasm resources of *Lycium barbarum* and cloning and identification of salt-tolerant genes[D]. Yangling: Northwest A&F University, 2018. (in Chinese)

[3] 徐峰. 宁夏枸杞产业竞争力研究[D]. 银川：宁夏大学，2017.
XU Feng. Research on the competitiveness of Ningxia wolfberry industry [D]. Yinchuan: Ningxia University, 2017. (in Chinese)

[4] 范振军. 农作物病虫害图像检索方法研究与实现[D]. 绵阳：西南科技大学，2018.
FAN Zhenjun. Research and realization of image retrieval method for crop diseases and pests [D]. Mianyang: Southwest University of Science and Technology, 2018. (in Chinese)

[5] 汪京京，张武，刘连忠，等. 农作物病虫害图像识别技术的研究综述[J]. 计算机工程与科学，2014，36(7)：1363 - 1370.
WANG Jingjing, ZHANG Wu, LIU Lianzhong, et al. A review of research on image recognition technology of crop diseases and insect pests[J]. Computer Engineering and Science, 2014, 36(7):1363 - 1370. (in Chinese)

[6] 陈志飞. 林业病虫害领域本体的语义检索研究[D]. 哈尔滨：东北林业大学，2017.
CHEN Zhifei. Research on semantic retrieval of ontology in the field of forestry diseases and pests[D]. Harbin: Northeast Forestry University, 2017. (in Chinese)

[7] 李贯峰，李卫军. 一个基于枸杞病虫害领域本体的语义检索模型[J]. 计算机技术与发展，2017，27(9)：48 - 52.
LI Guanfeng, LI Weijun. A semantic retrieval model based on the domain ontology of *Lycium barbarum* diseases and pests[J]. Computer Technology and Development, 2017, 27(9):48 - 52. (in Chinese)

[8] 刘月娥. 基于内容的作物害虫图像检索技术研究与系统实现[D]. 杨凌：西北农林科技大学，2006.
LIU Yue'e. Content-based crop pest image retrieval technology research and system implementation[D]. Yangling: Northwest A&F University, 2006. (in Chinese)

[9] WANG K, YIN Q, WANG W, et al. A comprehensive survey on cross-modal retrieval[J]. arXiv preprint arXiv:1607.06215, 2016.

[10] 邵杰. 基于深度学习的跨模态检索[D]. 北京：北京邮电大学，2017.
SHAO Jie. Cross-modal retrieval based on deep learning[D]. Beijing: Beijing University of Posts and Telecommunications, 2017. (in Chinese)

[11] 徐峰，马小萍，刘立波. 基于生成对抗网络的甲状腺超声图像文本跨模态检索方法[J]. 生物医学工程学杂志，2020，37(4)：641 - 651.
XU Feng, MA Xiaoping, LIU Libo. Cross-modal retrieval method of thyroid ultrasound image text based on generative adversarial network[J]. Journal of Biomedical Engineering, 2020, 37(4):641 - 651. (in Chinese)

[12] 邵阳雪，孟伟，孔德珍，等. 基于深度学习的特种车辆跨模态检索方法[J]. 计算机科学，2020，47(12)：205 - 209.
SHAO Yangxue, MENG Wei, KONG Dezhen, et al. Cross-modal retrieval method for special vehicles based on deep learning [J]. Computer Science, 2020, 47(12):205 - 209. (in Chinese)

[13] 许景辉，邵明烨，王一琛，等. 基于迁移学习的卷积神经网络玉米病害图像识别[J]. 农业机械学报，2020，51(2)：230 - 236，253.
XU Jinghui, SHAO Mingye, WANG Yichen, et al. Image recognition of corn diseases based on transfer learning convolutional neural network[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(2): 230 - 236, 253. (in Chinese)

[14] 翟肇裕，曹益飞，徐焕良，等. 农作物病虫害识别关键技术研究综述[J]. 农业机械学报，2021，52(7)：1 - 18.
ZHAI Zhaoyu, CAO Yifei, XU Huanliang, et al. Review of key techniques for crop disease and pest detection[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(7):1 - 18. (in Chinese)

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems, 2017: 5998 - 6008.
[16] 林阳,初旭,王亚沙,等.融合自注意力机制的跨模态食谱检索方法[J].计算机科学与探索,2020,14(9):1471 - 1481.
LIN Yang, CHU Xu, WANG Yasha, et al. Cross-modal recipe retrieval method fused with self-attention mechanism[J]. Computer Science and Exploration, 2020,14(9):1471 - 1481. (in Chinese)
[17] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735 - 1780.
[18] 綦金玮,彭宇新,袁玉鑫.面向跨媒体检索的层级循环注意力网络模型[J].中国图象图形学报,2018,23(11):1751 - 1758.
QI Jinwei, PENG Yuxin, YUAN Yuxin. Hierarchical cyclic attention network model for cross-media retrieval[J]. Chinese Journal of Image and Graphics,2018,23(11):1751 - 1758. (in Chinese)
[19] RASIWASIA N, COSTA P, COVIELLO E, et al. A new approach to cross-modal multimedia retrieval[C]// Proceedings of the 18th ACM International Conference on Multimedia, 2010: 251 - 260.
[20] 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.
[21] ANDREW G, ARORA R, BILMES J, et al. Deep canonical correlation analysis[C]// International Conference on Machine Learning. PMLR, 2013: 1247 - 1255.
[22] YAN F, MIKOLAJCZYK K. Deep correlation for matching images and text[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2015: 3441 - 3450.
[23] 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.
[24] FENG F, WANG X, LI R. Cross-modal retrieval with correspondence autoencoder[C]// Proceedings of the 22nd ACM International Conference on Multimedia,2014: 7 - 16.
[25] WANG B, YANG Y, XU X, et al. Adversarial cross-modal retrieval[C]// Proceedings of the 25th ACM International Conference on Multimedia,2017: 154 - 162.
[26] PENG Yuxin, QI Jinwei, YUAN Yuxin. Modality-specific cross-modal similarity measurement with recurrent attention network [J]. IEEE Transactions on Image Processing, 2018, 27(11): 5585 - 5599.

(上接第 298 页)

[27] COVINGTON A K, BATES R G, DURST R A. Definition of pH scales, standard reference values, measurement of pH and related terminology[J]. Pure and Applied Chemistry, 1985, 57(3): 531 - 542.
[28] GRAHAM D J, JASELSKIS B, MOORE C E. Development of the glass electrode and the pH response[J]. Journal of Chemical Education, 2013, 90(3): 345 - 351.
[29] 何东健,刘畅,熊虹婷. 奶牛体温植入式传感器与实时监测系统设计与试验[J]. 农业机械学报, 2018, 49(12): 195 - 202.
HE Dongjian, LIU Chang, XIONG Hongting. Design and experiment of implantable sensor and real-time detection system for temperature monitoring of cow[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(12): 195 - 202. (in Chinese)
[30] 中华人民共和国国家质量监督检验检疫总局,中国国家标准化管理委员会. GB/T 30121—2013 工业铂热电阻及铂感温元件[S]. 北京: 中国标准出版社, 2013.
[31] NOGAMI H, ARAI S, OKADA H, et al. Minimized bolus-type wireless sensor node with a built-in three-axis acceleration meter for monitoring a cow's rumen conditions[J]. Sensors, 2017, 17(4): 687.
[32] BEWLEY J M, GROTT M W, EINSTEIN M E, et al. Impact of intake water temperatures on reticular temperatures of lactating dairy cows[J]. Journal of Dairy Science, 2008, 91(10): 3880 - 3887.
[33] ALZAHAL O, RUSTOMO B, ODONGO N E, et al. Technical note: a system for continuous recording of ruminal pH in cattle [J]. Journal of Animal Science, 2007, 85(1): 213 - 217.
[34] 赵小伟,王加启,卜登攀,等. 奶牛瘤胃 pH 不同测定方法的比较研究[J]. 中国畜牧兽医, 2013, 40(11): 213 - 216.
ZHAO Xiaowei, WANG Jiaqi, BU Dengpan, et al. Comparative study on two different test methods in rumen pH of dairy cows [J]. China Animal Husbandry & Veterinary Medicine, 2013, 40(11): 213 - 216. (in Chinese)
[35] ALZAHAL O, ALZAHAL H, STEELE M, et al. The use of a radiotelemetric ruminal bolus to detect body temperature changes in lactating dairy cattle[J]. Journal of Dairy Science, 2011, 94(7): 3568 - 3574.