·人工智能与模式识别·　　文章编号：1000-3428(2022)11-0062-07　　文献标志码：A　　中图分类号：TP391.3

基于模态特异及模态共享特征信息的多模态细粒度检索

李　佩[1,2]，陈乔松[1,2]，陈鹏昌[1,2]，邓　欣[1,2]，王　进[1,2]，朴昌浩[1,2]

（1.重庆邮电大学 计算机科学与技术学院，重庆 400065；2.数据工程与认知计算重庆市重点实验室，重庆 400065）

摘　要：跨模态检索的目标是用户给定任意一个样本作为查询样例，系统检索得到与查询样例相关的各个模态样本，多模态细粒度检索在跨模态检索基础上强调模态的数量至少大于两个，且待检索样本的分类标准为细粒度子类，存在多模态数据间的异构鸿沟及细粒度样本特征差异小等难题。引入模态特异特征及模态共享特征的概念，提出一种多模态细粒度检索框架MS2Net。使用分支网络及主干网络分别提取不同模态数据的模态特异特征及模态共享特征，将两种特征通过多模态特征融合模块进行充分融合，同时利用各个模态自身的特有信息及不同模态数据间的共性及联系，增加高维空间向量中包含的语义信息。针对多模态细粒度检索场景，在center loss函数的基础上提出multi-center loss函数，并引入类内中心来聚集同类别且同模态的样本，根据聚集类内中心来间接聚集同类别但模态不同的样本，同时消减样本间的异构鸿沟及语义鸿沟，增强模型对高维空间向量的聚类能力。在公开数据集FG-Xmedia上进行一对一与一对多的模态检索实验，结果表明，与FGCrossNet方法相比，MS2Net方法mAP指标分别提升65%和48%。

关键词：信息检索；多模态检索；细粒度检索；多模态表征学习；深度学习

开放科学（资源服务）标志码（OSID）：

中文引用格式：李佩，陈乔松，陈鹏昌，等.基于模态特异及模态共享特征信息的多模态细粒度检索[J].计算机工程，2022，48(11)：62-68，76.

英文引用格式：LI P，CHEN Q S，CHEN P C，et al. Multi-modal fine-grained retrieval based on modal specific and modal shared feature information[J].Computer Engineering，2022，48(11)：62-68，76.

Multi-Modal Fine-Grained Retrieval Based on Modal Specific and Modal Shared Feature Information

LI Pei[1,2]，CHEN Qiaosong[1,2]，CHEN Pengchang[1,2]，DENG Xin[1,2]，WANG Jin[1,2]，PIAO Changhao[1,2]

（1.College of Computer Science and Technology，Chongqing University of Posts and Telecommunications，Chongqing 400065，China；
2.Chongqing Key Laboratory of Data Engineering and Visual Computing，Chongqing 400065，China）

【Abstract】The goal of cross-modal retrieval is that the user gives any sample as a query sample；then，the system retrieves and feeds back various modal samples related to the query sample. Multi-modal fine-grained retrieval emphasizes that the number of modalities is greater than two and the granularity of classification is the fine-grained sub-category. This paper introduces the concepts of modal specific features and modal shared features and proposes the MS2Net framework. The branch network and backbone network are used to extract the modal specific features and modal shared features of different modal data. Then，the two features are fully fused through the Multi-Modal Feature fusion Module（MMFM）. Meanwhile，the semantic information contained in the high-dimensional space vector is greatly increased by using the unique information of each mode and the commonness and relationship between different modal data. In addition，for the multi-modal fine-grained retrieval scenario，based on center loss，this paper proposes multi-center loss，introduces the inner-class center to gather the samples of the same category and the same mode，and then indirectly gathers the samples of the same category but different modes by aggregating the inner-class center. This reduces the heterogeneous gap and semantic gap between the samples. It clearly enhances the clustering ability of the model to high-dimensional spatial vectors. Finally，the experimental results of one-to-one and one-to-multimodal retrieval on the FG-Xmedia public dataset show that，compared with the FGCrossNet method，the MS2Net method improves the mAP index by 65% and 48%，respectively.

【Key words】information retrieval；multi-modal retrieval；fine-grained retrieval；multi-modal representation learning；deep learning

DOI：10. 19678/j. issn. 1000-3428. 0063185

基金项目：国家自然科学基金（61806033）；国家社会科学基金西部项目（18XGL013）。

作者简介：李　佩（1996—），男，硕士研究生，主研方向为多模态检索、机器视觉；陈乔松，副教授、博士；陈鹏昌，硕士；邓　欣，副教授、博士；王　进，教授、博士；朴昌浩，教授、博士后。

收稿日期：2021-11-09　　**修回日期：**2021-12-29　　**E-mail：**2307848408@qq.com

0　概述

在移动互联网时代,人们能够随时随自由地通过网络发布信息、传递信息和接收信息,这些信息中通常包含文字、音频、图片、视频等多模态数据。飞速增长的多模态数据带来了大量的跨模态检索应用需求,但这些跨模态检索需求不能由以文检文等单模态检索技术来解决,因此亟需发展适用于跨模态检索的理论、方法和技术。

近年来,深度神经网络在计算机视觉[1-2]、自然语言处理[3-4]、语音识别[5-6]等各个领域都取得了显著的成果,展现出了深度学习模型在处理不同模态信息时具有的优异特征提取能力。当前,基于深度学习的多模态检索逐渐成为多模态检索方法的主流。

在传统的多模态检索模型[7-9]中,一般针对不同的模态使用不同的神经网络提取特征向量,或者使用一个主干网络同时提取不同模态的特征向量。前者着重利用模态特异信息,但难以提取模态间联系与不同模态样本的共性,后者着重提取模态间联系与共性,但共性与联系只是所有数据的一小部分,造成了大量有效的模态特异信息的损失[10]。

针对以上问题,本文提出一种多模态细粒度检索框架MS2Net。通过提取并融合不同模态细粒度样本的模态特异信息及模态公共信息,得到包含丰富语义信息及模态间联系与共性的特征向量,并通过改进discriminate loss[11]、center loss[12]、triplet loss[13]等损失函数,将其组成适合多模态细粒度检索任务的目标函数。

1　相关工作

1.1　跨模态检索

跨模态检索的目标是用户给定任意一个样本作为查询样例,系统检索得到并反馈与查询样例相关的各个模态样本。目前主流的基于深度学习的跨模态检索方法过程是使用不同的分支网络提取不同模态数据的特征向量,再将这些不同模态数据提取出的特征向量映射到一个高维公共空间中,在该高维公共空间中,对不同模态的公共空间向量进行直接比较得到最佳匹配项。这种方法利用神经网络优异的特征提取能力消减了不同模态数据间的异构鸿沟,并利用在高维空间中的聚类函数消减不同模态数据间的语义鸿沟,达到较优的检索效果。在此基础上,文献[14]利用上述基本框架提取公共空间向量,并通过多目标函数的方式增强了公共空间向量的多模态检索性能,具体方法是通过监督学习保证分支网络提取的特征向量的质量,并设计目标函数减小同类别样本的类内差异,增加不同类样本的类间差异。文献[15]通过设置模态内的注意力机制及模态间注意力机制,建立图像中部分位置与文本中单词的强联系,来增强不同模态的相同类别样本间的语义相似性。文献[16]在公共空间中引入了对抗神经网络GAN中的对抗思想,使得图像向量与文字向量尽可能地相融合。

1.2　细粒度检索

跨模态细粒度检索相较于跨模态检索的最大困难是样本类间差异小、类内差异大。为了解决这个问题,文献[17]通过对输入信息进行预处理的方式,建立样本不同部位图像与文字之间的强监督学习,进行多模态表征学习,并用于跨模态细粒度检索。文献[18]验证了使用单一主干网络不仅可以提取用于各模态数据分类的模态特异信息,还可以提取出不同模态数据间的联系,实现跨模态检索。同时,通过多任务目标函数的协作减小类内差异,增大类间差异,实现最优的多模态细粒度检索性能。

1.3　多模态表征学习

优秀的多模态表征学习能够有效地提取不同模态样本的有效信息,使得特征向量含有丰富的原始样本中的语义信息,能极大地提升后序检索工作的准确性。文献[19]在ReID任务中同时利用了模态特异特征及模态共享特征,通过两个分支分别提取各个模态特异信息,将分支信息进行多损失函数约束的转换得到模态公共特征,再进行充分的特征融合,在ReID任务中取得了较好的性能。文献[20]使用卷积神经网络提取图像特征,同时利用一个双阶段特征提取网络提取文本特征,具体是在第1个阶段使用两个LSTM分支分别提取食物实体以及长句子的特征,最后通过正则化联合两个特征得到菜单文本特征。

1.4　高维公共空间方法

将样本映射到高维公共空间是各种多模态任务中的重要方法,可以有效地化解不同模态数据之间的异构鸿沟,而对于检索任务,公共空间向量聚类效果的好坏,直接决定了多模态检索的效果。文献[21]通过监督学习保证特征向量的质量,再通过增大不同类别但相同模态样本对及不同类别不同模态样本对之间的类间距离,减少相同类别不同模态样本对之间的距离,对公共空间向量进行聚类。文献[22]则通过类别单词的预训练嵌入向量作为锚点,将提取出的对应类别的样本对的特征向量以这个锚点作为中心进行聚类。为了更有效地提高triplet loss的训练效果,文献[23]通过L2正则化将高纬空间向量限制在一个球形空间中,并通过预训练的音频锚点,保证了类别中心的稳定,增强了聚类效果。

2　本文方法

2.1　公式化

本文提出方法实现了跨图像、文本、音频及视频的检索,假设有N对 image-video-audio-text的四元样本对,设为$\Psi=(x_i^{\mathrm{I}},x_i^{\mathrm{V}},x_i^{\mathrm{A}},x_i^{\mathrm{T}})_{i=1}^{N}$,其中:$x_i^{\mathrm{I}}$为图像样本;$x_i^{\mathrm{V}}$为视频样本;$x_i^{\mathrm{A}}$为音频样本;$x_i^{\mathrm{T}}$为文本样本。它们

都属于第i个四元样本对，每个样本对都被赋予一个语义标签向量label，$\boldsymbol{y}_i=(y_i^{\mathrm{I}},y_i^{\mathrm{V}},y_i^{\mathrm{A}},y_i^{\mathrm{T}})\in\mathbb{R}^c$，其中：$y_i^{\mathrm{I}}$、$y_i^{\mathrm{V}}$、$y_i^{\mathrm{A}}$、$y_i^{\mathrm{T}}$分别为图像样本、视频样本、音频样本及文字样本的类别标签。

MS2Net首先使用不同的特征分支网络对应不同的模态样本学习一个特征提取方法$u_i^m=f(x_i^m;P^m)\in\mathbb{R}^{d_1},m\in\{\mathrm{I},\mathrm{V},\mathrm{A},\mathrm{T}\}$，得到模态特异特征$\boldsymbol{u}_i=\{u_i^{\mathrm{I}},u_i^{\mathrm{V}},u_i^{\mathrm{A}},u_i^{\mathrm{T}}\}$，其中：$x_i^m$代表$m$模态下的第$i$个样本对；$P^m$代表$m$模态特异分支网络的可训练参数；$d_1$代表模态特异特征的向量维度。

然后使用一个特征共享主干网络学习一个提取四元样本对的模态共享特征的特征提取方法$\boldsymbol{v}_i=f(x_i^{\mathrm{I}},x_i^{\mathrm{V}},x_i^{\mathrm{A}},x_i^{\mathrm{T}},P)\in\mathbb{R}^{d_2}$，得到模态共享特征$\boldsymbol{v}_i=\{v_i^{\mathrm{I}},v_i^{\mathrm{V}},v_i^{\mathrm{A}},v_i^{\mathrm{T}}\}$，其中：$P$为特征共享主干网络的可训练参数；$d_2$代表模态公共特征的向量维度。

通过模型融合模块分别将模态特异$\boldsymbol{u}_i$与模态共享特征$\boldsymbol{v}_i$融合后，得到模态融合特征$\boldsymbol{z}_i=\{z_i^{\mathrm{I}},z_i^{\mathrm{V}},z_i^{\mathrm{A}},z_i^{\mathrm{T}}\}$，最后将模态融合特征映射到公共空间中得到$\Theta_i=\{\theta_i^{\mathrm{I}},\theta_i^{\mathrm{V}},\theta_i^{\mathrm{A}},\theta_i^{\mathrm{T}}\}\in\mathbb{R}^{d_3}$，即为样本集合$\Psi$在公共空间中的表现形式，其中$d_3$为公共空间特征向量的维度。

2.2 MS2Net结构

2.2.1 多模态特征提取

若要实现多模态检索，则首要任务是必须有效地提取多模态数据的特征。相较于传统方法通常只提取模态特异特征，多模态特征应同时具备模态特异信息及模态共享信息，本文使用模态共享主干网络及模态特异分支网络分别提取这两种特征信息。

由于在数据预处理阶段，图像、视频及音频模态的数据都被转换成了四维矩阵的图像形式，文本模态数据也通过卷积神经网络升维成了四维矩阵，因此本文采用计算机视觉领域常见的基准网络作为主干网络进行监督分类实验，目的是为了选择对样本特征提取能力更强的网络，更利于模型学习到共享信息。

MS2Net的特征提取部分由1个主干网络及4个分支网络构成，主干网络层数过深或者结构过于复杂会导致模型无法收敛或者训练时间超出预期，ResNet-50网络结构简单，参数量不大，更适合用于进行端到端学习，考虑到分类精度和训练难度的平衡，本文最终采用在ImageNet上预训练的ResNet-50作为主干网络的初始状态。

MS2Net框架包含4个分支网络，为了减少整个网络的参数量，需要选择轻量级网络，同时MixNet-S[24]中的深度混合卷积包含不同大小的卷积核，不仅保证了提取图像特征时精度与参数量的平衡，还在提取矩阵形式的声音及文本特征时，能够提取到不同范围内的时序特征，在多种轻量级网络中进行监督分类的效果最佳。

MS2Net网络结构如图1所示。

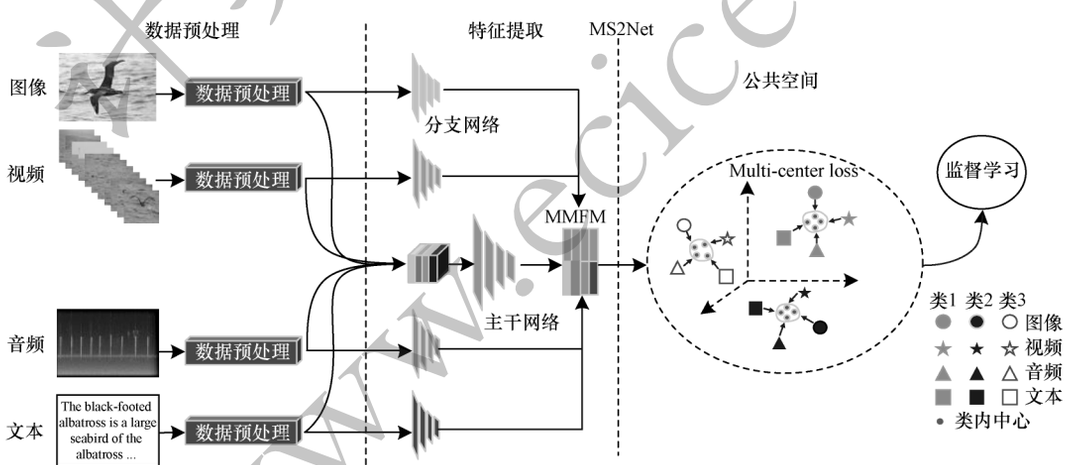


图1 MS2Net网络结构

Fig.1 MS2Net network structure

监督分类实验结果如表1所示（加粗数字为最优值）。可以看出，MixNet-S在单独一个模态的F1-score指标明显高于ResNet-50，但是当MixNet-S作为主干网络进行多模态分类时，平均F1-score指标不及ResNet-50的50%。

从实验结果可以看出，主干网络ResNet-50提取的是各个模态共有的粗粒度共有特征，能同时对4个模态的数据进行分类，而分支网络专注于提取单个模态的细粒度特异特征信息，因此在特定模态分类效果上明显优于主干网络，结合主干网络的粗粒度特征及分支网络的细粒度特征，经过多模态特征融合模块之后，MS2Net的效果相较于主干网络ResNet-50提升了36.2%，进一步佐证了模态特异特征及模态共享特征思想的有效性。同时，模态共享主干网络及模态特异分支网络可以根据实际情况进行更换，模态特异分支还可以根据实际数据中的模态数量进行增减，使得整个网络具有良好的可扩展性及鲁棒性。

表1　多跨模态监督分类的F1-score值

Table 1　F1-score values of multi-span modal supervised classification

方法	Image	Video	Audio	Text
ResNet-50	0.795	0.433	0.581	0.166
MixNet-S分支	0.832	0.465	0.627	0.334
MS2Net	**0.872**	**0.548**	**0.686**	**0.584**

2.2.2　多模态特征融合

在得到多模态特征向量后,需要有效地利用多模态特征向量。相较于传统方法将不同模态样本特征进行拼接后直接传入映射层中,本文的方法在多模态特征融合更注重于不同模态信息的充分混合及高效利用,在多模态融合模块中,先将各个模态的模态特异特征与模态公共特征进行拼接,通过注意力机制[25],使得模型能更有效地选择信息。最终对于每个模态产生一个模态融合特征,即 $\boldsymbol{z}_i=\{z_i^{\mathrm{I}},z_i^{\mathrm{V}},z_i^{\mathrm{A}},z_i^{\mathrm{T}}\}$。映射层使用全连接层,将模态融合特征映射到 d_3 维度,即 $\boldsymbol{\Theta}_i=\{\theta_i^{\mathrm{I}},\theta_i^{\mathrm{V}},\theta_i^{\mathrm{A}},\theta_i^{\mathrm{T}}\}\in\mathbb{R}^{d_3}$。

通过以上的操作,将4种不同模态的数据分别提取了模态共享特征及模态特异特征,并对这两个特征进行了特征融合,最后将其融合之后的特征映射到公共空间中,并通过multi-center loss函数进行聚类。

多模态特征融合模块结构如图2所示。

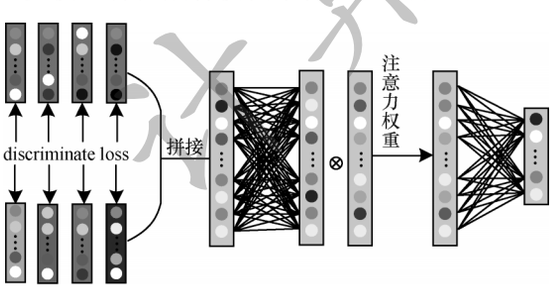


图2　多模态特征融合模块结构

Fig.2　Structure of multi-modal feature fusion module

2.3　损失函数

损失函数的目标是引导上述网络结构来学习一种语义映射,可以将同类别的样本映射到公共空间中距离相近的区域,即使这些样本属于不同的模态。

为了使得模型能区分细粒度样本的所属类别,本文方法首先通过ohem loss函数来进行监督学习,并且通过将易分类样本cross entropy loss函数值置0,使得模型更加专注于学习相似样本间的细节差异。其次为了使提取出来的模态共享特征与模态特异特征包含不同的信息,使用discriminate loss函数来加大这两种信息的差异,最后通过multi-center loss函数对映射到公共空间中的向量进行聚类。相较于传统的center loss函数,multi- center loss函数给每个模态数据都分配一个类别中心,通过类别中心之间的距离减小,间接地聚合不同模态数据。以上3个loss函数的组合,使得本文方法在高位空间中克服了不同模态数据间的语义鸿沟及异构鸿沟,使不同模态下的同类别样本相较于同模态的不同类别样本能够呈现更相似的特征向量。

2.3.1　supervised loss函数

监督学习是保证网络提取特征向量的质量的重要手段,同时为了驱使网络学习细粒度样本间的细节信息,本文方法选用ohem loss函数,如式(1)所示,即在原始的cross-entropy函数的基础上,对每个batch的样本只学习损失较大的一部分样本,使得模型更加专注于提取不同样本间的细节差异。

$$\mathcal{L}_{\mathrm{s}}=-\frac{1}{n}\sum_{i=1}^{N}\mathrm{l}(y\log_a(p_i)-(1-y)\log_a(1-p_i)) \tag{1}$$

其中:y 是样本的label值;p_i 是网络的输出值;$\mathrm{l}(\cdot)$ 为指示函数,如果loss值大于阈值,则值为1,否则值为0。

2.3.2　discriminate loss函数

为了避免主干网络及分支网络学习到过于相近的特征,本文方法使用discriminate loss函数,如式(2)所示,保证主干网络提取的信息与分支网络提取的信息既包含足够多的样本特征信息,又具有足够的差异性。

$$\mathcal{L}_{\mathrm{d}}=\frac{1}{n^2}\sum_{m=1}^{\mathrm{I,V,A,T}}\sum_{i=1}^{N}\max(0,\zeta-\log_a(1-\delta(\boldsymbol{\Phi}_i))) \tag{2}$$

其中:$\boldsymbol{\Phi}_i=\frac{1}{2}\cos(\boldsymbol{u}_i,\boldsymbol{v}_i)$ 为特征相似性函数,用于计算模态公共特征向量和模态特异特征向量的相似性;$\delta(\cdot)$ 为sigmoid函数,即 $y=\frac{1}{1+\mathrm{e}^{-x}}$;$\zeta$ 为一个自定义阈值,当模态共享特征向量和模态特异特征向量的相似性小于该阈值时,模型不再计算其的loss。随着 $\boldsymbol{v}_i$ 与 $\boldsymbol{u}_i$ 的相似度越来越高,即 $\delta(\boldsymbol{\Phi}_i)$ 越大,$\mathcal{L}_{\mathrm{dis}}$ 值也越大。因此,discriminate loss函数在确保模态公共特征向量和模态特异特征向量质量的基础上,同时保证这两个向量的差异性,使得MS2Net提取的特征向量包含原始样本中更全面的信息。

2.3.3　multi-center loss函数

为了更好地提高多模态细粒度检索效果,不同模态的相同类别样本向量,相较于同模态不同类别样本向量在公共空间中应该更加相邻,但是在多模态场景下,不同模态样本在公共空间中概率分布相差较大,传统center loss函数的聚类能力有限,所以本文针对多模态聚类场景设计了multi-center loss函数,通过在每个类别的所有模态中都分别引入一个类内中心,提升整体的聚类效果,计算公式如式(3)所示:

$$\mathcal{L}_{\mathrm{m}}=\frac{1}{4}\sum_{m=1}^{\mathrm{I,V,A,T}}\sum_{i=1}^{N}(\|\boldsymbol{x}_i^m-\boldsymbol{c}_y^m\|_2^2+\mu D(\boldsymbol{C}_y)) \tag{3}$$

其中:$\boldsymbol{x}_i^m$ 表示 m 模态的第 i 个样本向量;$\boldsymbol{c}_y^m$ 表示该样本向量在 m 模态下的类内中心;$\boldsymbol{C}_y$ 为 y 类别下的所有模态的类内中心;μ 是权重参数;$D(\cdot)$ 为类内中心的

距离函数。

$D(\cdot)$计算公式如式(4)所示：

$$D(\boldsymbol{C}_y)=\frac{1}{4}\sum_{m_2=1}^{\mathrm{I,V,A,T}}\|\boldsymbol{c}_y^{m}-\boldsymbol{c}_y^{m_2}\|_2^2 \tag{4}$$

其中：$\boldsymbol{c}_y^{m}$、$\boldsymbol{c}_y^{m_2}$表示同一类别下不同模态的类内中心。

multi-center loss函数的思想可以概括为：先将相同模态、相同类别的样本分别聚集到该模态下的类内中心周围，再通过减小不同模态、相同类别的类内中心的距离，间接地将不同模态、相同类别的样本聚集到一起。

综上所述，最终目标函数定义如式(5)所示：

$$\mathcal{L}=\alpha\mathcal{L}_{\mathrm{s}}+\beta\mathcal{L}_{\mathrm{d}}+\gamma\mathcal{L}_{\mathrm{m}} \tag{5}$$

目标函数由3个经典的损失函数改进组成，更加适合于多模态细粒度检索的场景，在下文的消融实验中将证明该目标函数的有效性。

3 实验

为了验证本文提出方法的有效性，在公开数据集FG-Xmedia上进行实验，首先和之前的最优结果进行对比，然后对MS2Net进行更深入的分析，包括各个组件的消融实验以及通过降维的公共空间向量可视化。

3.1 数据集

本文使用的数据集是当前唯一的跨四模态细粒度检索公开数据集FG-Xmedia，该数据集的图像部分采用CUB-200-2011[26]，视频部分采用YouTube Bird[27]，并从公开数据集上收集了相应的文本和音频部分。总地来讲，该数据集包含了“鸟类”这一粗粒度大类中的200种细粒度子类的4个模态的信息，分别为11 788张图像、8 000段文本、18 350个视频及12 000段音频。

数据集的划分方法如下：对于图片数据集，训练集包含5 994张图片，测试集包含5 794张图片；对于视频数据集，训练集包含12 666个视频，测试集包含5 684个视频；对于音频数据集，训练集和测试集同时包含6 000份音频；对于文本数据集，训练集和测试集同时包含4 000份文本。

3.2 数据处理

为了降低模型训练难度及尽量消除数据的异构性，使得主干网络可以统一提取不同模态的公共特征，音频数据通过傅里叶变换转化为频谱图，文字通过词嵌入转换为一维词向量，再通过卷积操作拼接成为类图像的矩阵形式。视频通过抽帧的方式转换为图像形式，这样4种数据都可以被组织成为相同的矩阵形式传入到主干网络中。

3.3 评价指标

对于多模态细粒度检索任务，本文使用平均精度均值(mean Average Precision，mAP)分数来衡量其性能，首先对于每一个查询样例计算其平均准确率，然后再计算这些平均准确率的平均值，作为mAP查询分数。

为了更加全面地了解模型的检索效果，需要分别计算每个模态对于其他单个模态的检索分数，并计算每个模态对其他所有模态的检索分数。

3.4 实现细节

MS2Net使用ResNet-50作为主干网络，将Mixnet-L作为分支网络，数据经过预处理后由这两个网络提取特征向量，之后通过一个512维的全连接层映射到公共空间，公共空间向量再通过一个200维的全连接层，以类别为标签进行监督学习。

网络使用PyTorch编写，并通过RTX-2070显卡训练。在训练过程中使用SGD优化器，初始学习率设置为3e−3，学习率衰减策略使用PyTorch框架提供的ReduceLROnPlateau策略，并训练200轮。

3.5 实验结果

表2所示为某模态数据作为查询样本对其他某单一模态的样本进行检索的mAP值结果(加粗数字为最优值)，箭头左右两端分别代表查询样本的模态及待检索的目标模态，可以看到MS2Net在大多数模态的检索任务中都明显优于FGCrossNet算法。在I→T、T→I、T→V、V→T、A→T等场景例中，MS2Net相较于FGCrossNet算法提升超过100%。整体上一对一模态检索的平均检索mAP值相较FGCrossNet算法提升了65%。

表2 一对一跨模态检索的mAP值

Table 2 mAP values of one-to-one cross-modal retrieval

算法	I→V	I→A	I→T	V→I	V→A	V→T	A→I	A→V	A→T	T→I	T→V	T→A	均值
MS2Net	**0.836**	**0.629**	**0.547**	**0.817**	**0.631**	**0.570**	**0.627**	**0.657**	**0.426**	**0.537**	**0.405**	**0.570**	**0.604**
FGCrossNet[18]	0.606	0.526	0.210	0.629	0.437	0.195	0.553	0.443	0.159	0.255	0.181	0.208	0.366
MHTN[28]	0.281	0.195	0.116	0.306	0.204	0.186	0.196	0.290	0.127	0.124	0.138	0.185	0.204
ACMR[16]	0.477	0.119	0.162	0.536	0.162	0.138	0.128	0.068	0.028	0.075	0.015	0.081	0.162
JRL[29]	0.435	0.085	0.160	0.517	0.160	0.126	0.115	0.065	0.035	0.190	0.028	0.095	0.160
GSPH[30]	0.417	0.089	0.140	0.512	0.159	0.126	0.129	0.073	0.024	0.179	0.024	0.109	0.159
CMDN[31]	0.377	0.009	0.099	0.446	0.105	0.081	0.017	0.010	0.005	0.123	0.007	0.078	0.105

表3所示为某一个模态数据作为查询对其他所有模态检索的mAP值(加粗数字为最优值)，该项测试中T→ALL相较于FGCrossNet算法提升超过100%，其他场景也有明显提升。整体上多模态的平

均检索mAP值相较FGCrossNet算法提升48%，充分证明了MS2Net在多模态细粒度检索任务上的有效性。

表3　一对多跨模态检索的mAP值

Table 3　mAP values of one-to-all cross-modal retrieval

算法	I→ALL	V→ALL	A→ALL	T→ALL	均值
MS2Net	**0.710**	**0.456**	**0.538**	**0.732**	**0.609**
FGCrossNet[18]	0.549	0.196	0.416	0.485	0.412
MHTN[28]	0.208	0.142	0.237	0.341	0.232
GSPH[30]	0.387	0.103	0.075	0.312	0.219
JRL[29]	0.344	0.080	0.069	0.275	0.192
CMDN[31]	0.321	0.071	0.014	0.229	0.159
ACMR[16]	0.245	0.039	0.041	0.279	0.151

3.6　消融实验

为了验证MS2Net中各个组件的有效性，本文进行了消融实验，结果如表4和表5所示（加粗数字为最优值），MS2Net表示利用MS2Net网络提取并直接拼接使用模态公共特征及模态特异特征得到的结果，MMFM表示引入注意力机制优化了特征融合过程，并使用单独的映射层输出的512维的向量作为公共空间向量。由消融实验结果可知，经过特征融合和注意力机制之后单模态检索性能提升13.5%，多模态检索性能提升19%，multi-center loss函数表示使用本文提出的multi-center loss函数替代传统的center loss函数对公共空间向量进行聚类。一对一跨模态检索实验与一对多跨模态检索实验性能分别提升10%与8%。

表4　一对一跨模态检索的消融实验结果

Table 4　Ablation experiment results of one-to-one cross-modal retrieval

算法	I→V	I→A	I→T	V→I	V→A	V→T	A→I	A→V	A→T	T→I	T→V	T→A	均值
MMFE	0.793	0.450	0.453	0.485	**0.631**	0.793	0.464	0.464	0.278	0.485	0.492	0.275	0.483
MMFE+MMFM	0.817	0.619	0.544	0.478	0.437	0.817	0.620	0.607	0.319	0.478	0.473	0.327	0.548
MMFE+MMFM+Multi-Center	**0.836**	**0.629**	**0.547**	**0.537**	0.304	**0.836**	**0.627**	**0.657**	**0.426**	**0.537**	**0.570**	**0.405**	**0.604**

表5　一对多跨模态检索的消融实验结果

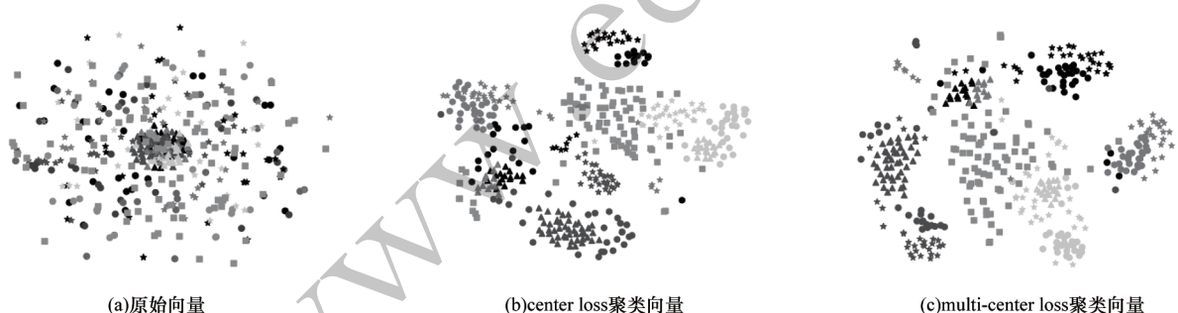
Table 5　Ablation experiment results of one-to-all cross-modal retrieval

算法	I→ALL	V→ALL	A→ALL	T→ALL	均值
MMFE	0.601	0.362	0.337	0.594	0.473
MMFE+MMFM	0.683	0.383	0.509	0.675	0.562
MMFE+MMFM+Multi-Center	**0.710**	**0.456**	**0.538**	**0.732**	**0.609**

消融实验的结果验证了联合模态间共性与模态特性思路及本文提出的multi-center loss函数的有效性。同时也可看出，高维公共空间向量相较于同类别数相等维度的特征向量，在检索任务中更合适作为原始样本的表达形式。

3.7　聚类效果可视化

为展示改进后multi-center loss的聚类效果，本文对其进行了可视化分析，聚类结果可视化结果如图3所示。



(a)原始向量　(b)center loss聚类向量　(c)multi-center loss聚类向量

图3　降维可视化图聚类结果示意图

Fig.3　Schematic diagram of clustering result of dimensionality reduction visual view

为了直观地感受不同方法产生的公共空间向量的聚类情况，本文统一使用VTSne算法对数据集中前5个类别产生的高维向量进行降维。

在上述图像中，相同的灰度代表样本属于相同的类别，相同的形状表示样本属于相同的模态。多模态细粒度检索的困难就是类间差异小和类内差异大，这在图3(a)中可以清晰地看到，各个模态的样本相互混杂，公共空间向量没有以类别为中心聚集，图3(b)为采用了MS2Net及center loss函数之后的结果，因为特征向量质量的提升与类别中心的引入，聚类效果相较于之前有明显提升，图3(c)为采用multi-center loss函数替换center loss函数，可以看到，同类别之间的公共空间向量，相较于图3(b)聚集的更加紧密，不同类别的公共空间向量在高维空间中的分布也更加稀疏，验证了multi-center loss函数在多模态数据聚类中的有效性。

4 结束语

本文提出一种多模态细粒度检索方法，该方法包括MS2Net多模态特征提取框架及相应的目标函数。MS2Net通过利用模态公共特征及模态特异特征提升公共空间向量的性能，同时给出目标函数组合，通过监督学习保证特征的质量和模态公共特征及模态特异特征的可区分性，从而进行有效的多模态特征向量聚类。消融实验结果表明，MS2Net性能明显提高，验证了各组件的有效性。由于当前多模态检索的数据集稀少，且数据集达标成本明显高于单模态数据集，下一步利用当前的公开数据集对多模态特征提取网络进行预训练，并加入判别器网络判断样本对类别，以提高在无标签检索数据集上的检索效果。

参考文献

[1] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[EB/OL]. [2021-10-08]. https://arxiv.org/abs/2010.11929.

[2] TOLSTIKHIN I, HOULSBY N, KOLESNIKOV A, et al. MLP-mixer: an all-MLP architecture for vision[EB/OL]. [2021-10-08]. https://arxiv.org/abs/2105.01601.

[3] ZHANG Y, WALLACE B. A sensitivity analysis of(and practitioners' guide to) convolutional neural networks for sentence classification[EB/OL]. [2021-10-08]. https://arxiv.org/abs/1510.03820.

[4] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2021-10-08]. https://arxiv.org/abs/1810.04805.

[5] CHIU C C, SAINATH T N, WU Y H, et al. State-of-the-art speech recognition with sequence-to-sequence models[C]// Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2018: 4774-4778.

[6] GULATI A, QIN J, CHIU C C, et al. convolution-augmented transformer for speech recognition[EB/OL]. [2021-10-08]. https://arxiv.org/abs/2005.08100.

[7] XU R C, NIU L, ZHANG J F, et al. A proposal-based approach for activity image-to-video retrieval[J]. Artificial Intelligence, 2020, 34(7): 12524-12531.

[8] XU X, SONG J K, LU H M, et al. Modal-adversarial semantic learning network for extendable cross-modal retrieval[C]//Proceedings of 2018 ACM on International Conference on Multimedia Retrieval. New York, USA: ACM Press, 2018: 46-54.

[9] JIANG Q Y, LI W J. Deep cross-modal hashing[C]// Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 3270-3278.

[10] CHEN Y C, LI L J, YU L C, et al. UNITER: UNiversal image-TExt representation learning[C]//Proceedings of ECCV'20. Berlin, Germany: Springer, 2020: 104-120.

[11] ZHEN L L, HU P, WANG X, et al. Deep supervised cross-modal retrieval[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2019: 10386-10395.

[12] WEN Y D, ZHANG K P, LI Z F, et al. A discriminative feature learning approach for deep face recognition[C]// Proceedings of European Conference on Computer Vision. Berlin, German: Springer, 2016: 499-515.

[13] SCHROFF F, KALENICHENKO D, PHILBIN J. FaceNet: a unified embedding for face recognition and clustering[C]// Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015: 815-823.

[14] GU J X, CAI J F, JOTY S, et al. Look, imagine and match: improving textual-visual cross-modal retrieval with generative models[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 7181-7189.

[15] ZHANG Q, LEI Z, ZHANG Z X, et al. Context-aware attention network for image-text retrieval[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 3533-3542.

[16] WANG B K, YANG Y, XU X, et al. Adversarial cross-modal retrieval[C]//Proceedings of the 25th ACM International Conference on Multimedia. New York, USA: ACM Press, 2017: 154-162.

[17] HE X T, PENG Y X. Fine-grained visual-textual representation learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(2): 520-531.

[18] HE X T, PENG Y X, XIE L. A new benchmark and approach for fine-grained cross-media retrieval[C]// Proceedings of the 27th ACM International Conference on Multimedia. New York, USA: ACM Press, 2019: 1740-1748.

[19] LU Y, WU Y, LIU B, et al. Cross-modality person re-identification with shared-specific feature transfer[C]// Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 13376-13386.

[20] WANG H, SAHOO D, LIU C H, et al. Cross-modal food retrieval: learning a joint embedding of food images and recipes with semantic consistency and attention mechanism [J]. IEEE Transactions on Multimedia, 2022, 24(3): 2515-2525.

[21] UDANDARAO V, MAITI A, SRIVATSAV D, et al. COBRA: contrastive bi-modal representation algorithm[EB/OL]. [2021-10-08]. https://arxiv.org/abs/2005.03687.

[22] NARAYANA P, PEDNEKAR A, KRISHNAMOORTHY A, et al. HUSE: hierarchical universal semantic embeddings [EB/OL]. [2021-10-08]. https://arxiv.org/abs/1911.05978.

[23] XIONG C Y, ZHANG D Y, LIU T, et al. Voice-face cross-modal matching and retrieval: a benchmark[EB/OL]. [2021-10-08]. https://arxiv.org/abs/1911.09338.

[24] TAN M X, LE Q V. MixNet: mixed depthwise convolutional kernels[EB/OL]. [2021-10-08]. https://arxiv.org/abs/1907.09595.

[25] XU K, BA J, KIROS R, et al. Show, attend and tell: neural image caption generation with visual attention[C]// Proceedings of International Conference on Machine Learning. New York, USA: ACM Press, 2015: 2048-2057.

（下转第76页）

（上接第68页）

[26] GÖERING C, RODNER E, FREYTAG A, et al. Nonparametric part transfer for fine-grained recognition [C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2014: 2489-2496.

[27] ZHU C, TAN X, ZHOU F, et al. Fine-grained video categorization with redundancy reduction attention [C]// Proceedings of ECCV'18. Berlin, Germany: Springer, 2018: 139-155.

[28] HUANG X, PENG Y X, YUAN M K. MHTN: modal-adversarial hybrid transfer network for cross-modal retrieval [J]. IEEE Transactions on Cybernetics, 2020, 50(3): 1047-1059.

[29] ZHAI X H, PENG Y X, XIAO J G. Learning cross-media joint representation with sparse and semi-supervised regularization [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(6): 965-978.

[30] MANDAL D, CHAUDHURY K N, BISWAS S. Generalized semantic preserving hashing for n-label cross-modal retrieval [C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 2633-2641.

[31] PENG Y, HUANG X, QI J. Cross-media shared representation by hierarchical learning with multiple deep networks [C]//Proceedings of IEEE IJCAI'16. Washington D. C., USA: IEEE Press, 2016: 3846-3853.

编辑 索书志