计算机科学
COMPUTER SCIENCE

基于粒度感知和语义聚合的图像-文本检索网络

缪岚芯，雷雨，曾鹏鹏，李晓瑜，宋井宽

引用本文

缪岚芯，雷雨，曾鹏鹏，李晓瑜，宋井宽. 基于粒度感知和语义聚合的图像-文本检索网络[J]. 计算机科学, 2022, 49(11): 134-140.

MIAO Lan-xin, LEI Yu, ZENG Peng-peng, LI Xiao-yu, SONG Jing-kuan. Granularity-aware and Semantic Aggregation Based Image-Text Retrieval Network[J]. Computer Science, 2022, 49(11): 134-140.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于 Bi-LSTM 的期货市场关联交易行为检测方法

Related Transaction Behavior Detection in Futures Market Based on Bi-LSTM

计算机科学, 2022, 49(7): 31-39. https://doi.org/10.11896/jsjkx.210400304

基于注意力机制的细粒度语义关联视频-文本跨模态实体分辨

Fine-grained Semantic Association Video-Text Cross-modal Entity Resolution Based on Attention Mechanism

计算机科学, 2022, 49(7): 106-112. https://doi.org/10.11896/jsjkx.210500224

基于多路径特征提取的实时语义分割方法

Real-time Semantic Segmentation Method Based on Multi-path Feature Extraction

计算机科学, 2022, 49(7): 120-126. https://doi.org/10.11896/jsjkx.210500157

指静脉识别技术研究综述

Survey on Finger Vein Recognition Research

计算机科学, 2022, 49(6A): 1-11. https://doi.org/10.11896/jsjkx.210400056

基于特征分离的红外与可见光图像融合算法

Infrared and Visible Image Fusion Based on Feature Separation

计算机科学, 2022, 49(5): 58-63. https://doi.org/10.11896/jsjkx.210200148

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220600010

基于粒度感知和语义聚合的图像-文本检索网络

缪岚芯[1] 雷 雨[1] 曾鹏鹏[1] 李晓瑜[2] 宋井宽[1]
1 电子科技大学计算机科学与工程学院(网络空间安全学院) 成都 611731
2 电子科技大学信息与软件工程学院 成都 610054
(miaolanxin.lisa@gmail.com)

摘 要 图像-文本检索是视觉-语言领域中的基本任务,其目的在于挖掘不同模态样本之间的关系,即通过一种模态样本来检索具有近似语义的另一种模态样本。然而,现有方法大多高度依赖于将图像特定区域和句中单词进行相似语义关联,低估了视觉多粒度信息的重要性,导致了错误匹配以及语义模糊嵌入等问题。通常,图片包含了目标级、动作级、关系级以及场景级的粗、细粒度信息,而这些信息无显式多粒度标签,难以与模糊的文本表达直接一一对应。为了解决此问题,提出了一个粒度感知和语义聚合(Granularity-Aware and Semantic Aggregation,GASA)网络,用于获得多粒度视觉特征并缩小文本和视觉之间的语义鸿沟。具体来说,粒度感知的特征选择模块挖掘视觉多粒度信息,并在自适应门控融合机制和金字塔空洞卷积结构的引导下进行了多尺度融合。语义聚合模块在一个共享空间中对来自视觉和文本的多粒度信息进行聚类,以获得局部表征。模型在两个基准数据集上进行了实验,在 MSCOCO 1k 上 $R@1$ 优于最先进的技术 2%以上,在 Flickr30K 上 $R@Sum$ 优于之前最先进的技术 4.1%。

关键词:图文匹配;跨模态检索;特征提取;语义聚类;多粒度信息提取
中图法分类号 TP391

Granularity-aware and Semantic Aggregation Based Image-Text Retrieval Network

MIAO Lan-xin[1],LEI Yu[1],ZENG Peng-peng[1],LI Xiao-yu[2] and SONG Jing-kuan[1]
1 School of Computer Science and Engineering,University of Electronic Science and Technology of China,Chengdu 611731,China
2 School of Information and Software Engineering,University of Electronic Science and Technology of China,Chengdu 610054,China

Abstract Image-text retrieval is a basic task in visual-language domain,which aims at mining the relationships between different modalities. However,most existing approaches rely heavily on associating specific regions of an image with each word in a sentence with similar semantics and underappreciate the significance of multi-granular information in images,resulting in irrelevant matches between the two modalities and semantically ambiguous embedding. Generally,an image contains object-level,action-level,relationship-level or even scene-level information that is not explicitly labeled. Therefore,it is challenging to align complex visual information with ambiguous descriptions. To tackle this issue,this paper proposes a granularity aware and semantic aggregating(GASA) network to obtain multi-visual representations and narrow the cross-modal gap. Specifically,the granularity-aware feature selection module selects copious multi-granularity information of images and conducts a multi-scale fusion,guided by an adaptive gated fusion mechanism and a pyramid structure. The semantic aggregation module clusters the multi-granularity information from visual and textual clues in a shared space to obtain the residual representations. Experiments are conducted on two benchmark datasets,and the results show our model outperforms the state-of-the-arts by over 2% on $R@1$ of MSCOCO 1k. Besides,our model outperforms the state-of-the-art by 4.1% in terms of Flickr30k on $R@Sum$.

Keywords Image-text matching,Cross-model retrieval,Feature extraction,Semantic aggregation,Multi-granularity information extraction

1 引言

随着互联网上多媒体数据的海量增长,单一模态的检索已经无法满足用户需求,跨模态检索应运而生[1-2]。图像-文本检索是其中重要的任务之一,旨在实现图片和文本两个模态之间的信息交互,通过一种模态样本来检索具有近似语义的另一种模态样本[3]。它促进了许多下游任务的发展,包括图像标注[4-6]以及视觉问答[7-10]。其核心挑战在于减小

到稿日期:2022-05-31 返修日期:2022-08-02
基金项目:国家自然科学基金(62122018,61872064)
This work was supported by the National Natural Science Foundation of China(62122018,61872064).
通信作者:宋井宽(jingkuan.song@gmail.com)

两种异构模态之间的语义差距，有效约束视觉和文本语义表征的一致性。

早期的研究工作主要采用全局匹配的方法[11-12]，来构造可学习的视觉-文本的联合嵌入空间，在该公共空间中直接度量两种异构模态全局向量的相关性。此后，研究者们发现，全局特征匹配仅利用了粗粒度信息，易造成大量视觉语义信息丢失，因此转而采用局部特征匹配的方式[13-16]。例如，Lee 等[14]使用自下而上的注意力方案，将图像编码为区域级特征，将文本编码为单词级特征，度量每个区域和单词之间的相似度。这些方法取得了可喜的成绩，它们提取图像中的细粒度表征，通过损失函数约束语义联合空间的嵌入，显式度量文本和图片两个模态不同粒度间的相似性。

但是，图 1 中通常包含多个对象，且它们所属类别相似，如都是“boys”，仅在属性、动作或者场景等信息上有所区别，而常用的区域特征难以凸显这些信息的判别性，如“rushing around”作为描述动作的短语，无法简单地对应图片的区域特征，导致图片匹配到不相关的动作“kicking”。同时，数据集中正负样本（P 和 N）存在一定的语义模糊问题，如“Boys in orange”和“Boys in black”两者语义相似，但后者才正确对应图片中对象的属性。换而言之，模糊的文本表达和缺少显示粒度标签的图像会大大影响现有匹配方法的准确性。为了解决这个问题，一些研究者通过多尺度表征去学习不同粒度间的对齐[17]，采用递归注意力机制存储单元的迭代匹配方式[18]或者可学习的树型结构[19]来衡量粒度间的相似性。受此启发，本文通过自适应门控融合机制和金字塔结构感知图片中的不同视觉粒度。同时，受 NetVLAD 相关工作的启发[20]，本文将包含多粒度信息的视觉线索和文本线索映射到一个共享空间中，通过语义聚类的方式隐式地计算粒度间的相似性。

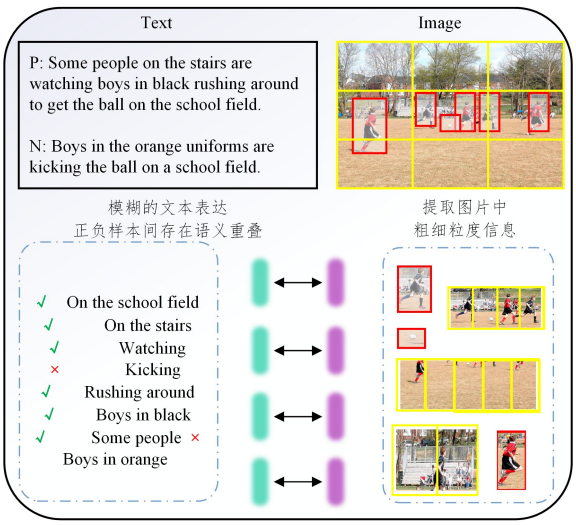


图 1　本文解决的问题

Fig. 1　Problem addressed in this paper

具体来说，本文提出了一个新颖的粒度感知和语义聚合（GASA）模型。该模型的整体框架主要由两部分组成：1）粒度感知的特征选择模块，旨在灵活地将网格特征和物体特征融合起来，以获得自适应视觉表征，在此基础上使用多层空洞金字塔卷积操作，从自适应视觉表征中挖掘多粒度信息；2）语义聚合模块，以联合嵌入的方式对来自视觉线索的多粒度信息和文本线索信息进行语义聚合，以减小两种模态之间的差距。

综上所述，本文的贡献主要为 3 个方面：

（1）提出了一个粒度感知的特征选择器，结合自适应门控融合机制和金字塔空洞卷积结构来提取图像的多粒度信息。

（2）设计了一个语义聚合模块，聚合多粒度线索的视觉信息和文本信息，进而减少异构模态的语义差距。整个过程具备可解释性，且易于实现。

（3）在两个基准数据集 Flickr30K 和 MS-COCO 上进行了性能实验，与最先进的方法相比，模型取得了有竞争力的结果。

2　相关工作

2.1　图像文本匹配

近年来，图像-文本检索引起了越来越多的关注。该任务需要对两个模态特征进行编码，根据一定的对齐机制来计算最终的跨模态相似度。现有方法主要分为两种：全局匹配方法和局部匹配方法。

全局匹配方法[11-12]将整个图像和整个句子以端到端的方式映射到一个联合语义空间，并直接计算相似度。例如，VSE＋＋[12]分别通过 VGG 和 GRU 网络获得图像和文本的全局特征。然而，它只考虑了图像和文本的全局上下文对齐关系，忽略了图像区域和单词的对齐。

局部匹配方法则通过细粒度跨模态交互，解决了上述全局匹配中存在的问题。典型的局部匹配方法倾向于计算所有的区域和单词对的相似度。Lee 等[14]提出了一种堆叠交叉注意来挖掘区域-单词的局部对齐，将图像区域和单词作为上下文，并通过平均或 LogSumExp 来聚集实现局部对齐，以衡量整体的跨模态相关性。PFAN[21]将区域的位置信息纳入考虑范围，并设计了一个位置信息整合方案，以实现更好的匹配。考虑到句子中总是有指代关系的词，Li 等[16]提出了基于 GCN 的 VSRN 模型，通过区域关系推理和全局语义推理增强图片表征。Wang 等[22]引入了共同出现的概念作为共识知识，以提高关于图像和文本之间高层对应关系的合理性。

本文工作结合了全局匹配和局部匹配的优势，重点是改善图像表征，获取多粒度信息。与以往的工作不同，本文采用了自适应门控机制和金字塔空洞卷积来捕捉图像中存在的多粒度信息。

2.2　NetVLAD

VLAD（Vector of Locally Aggregated Descriptors）是一个流行的特征计算方式，NetVLAD[23]则采用深度学习方法对 VLAD 进行了改进。在单模态中，NetVLAD[24]可用于点云检索任务，编码提取到的点云特征，使其更具全局表达性；其也可用于场景识别[25]任务，将一系列相似的候选图像区分开来，结合局部特征与全局特征的优势有效应对环境以及

视角变化。在多模态中，NetVLAD 可用于视频和文本检索[20]，将多个模态中的局部特征与共享中心进行聚类，这些中心提供共享的语义主题，可以弥合不同模态之间的差距。

受以上相关工作的启发，本文结合 NetVLAD 的优势，并加入深度可分离卷积，通过更少量的乘法运算来学习聚合更多潜在的语义信息，将多粒度视觉线索和文本线索归纳为共享空间中的数个残差表征，隐式地学习两种模态之间的潜在对应关系。

3 粒度感知和语义聚合模型

GASA 的整体框架如图 2 所示，其主要由粒度感知的特征选择器和语义聚合模块两部分组成。首先介绍了视觉表征和文本表征的生成方式，其次细致地阐述了模型的框架结构，最后对模型的损失函数进行了定义。

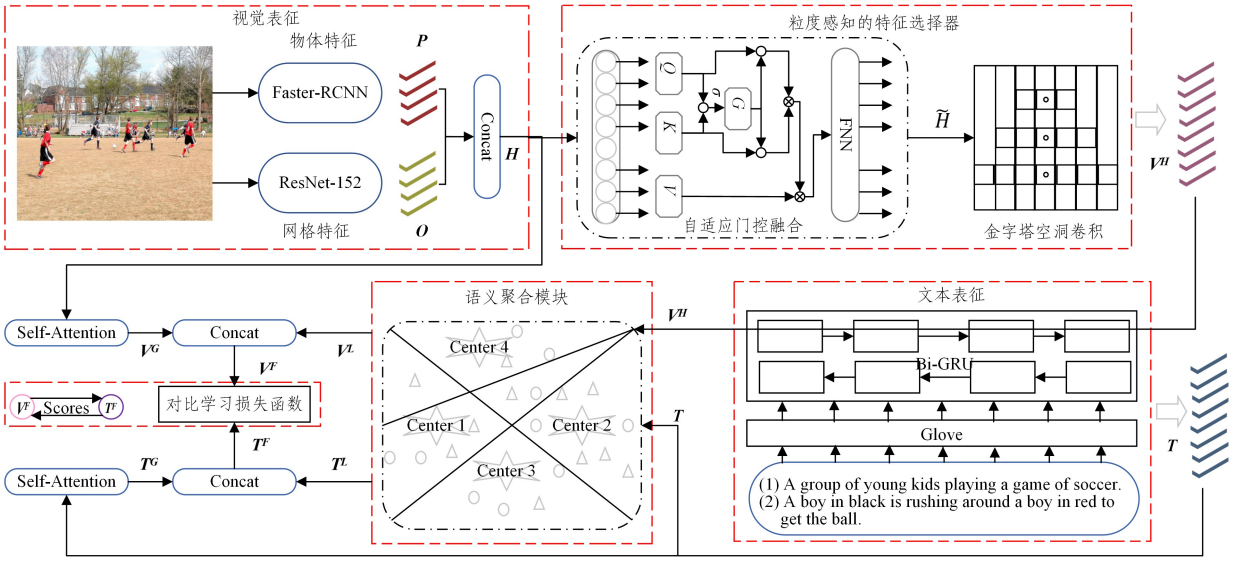


图 2 本文方法的整体框架图

Fig. 2 Overall framework of GASA

3.1 视觉表征和文本表征

视觉特征主要包含物体特征和网格特征。本文使用自下而上的注意力(Bottom-up Attention)模型[26]提取图像中置信度最高的感兴趣区域(Region of Interest, ROI)作为物体特征。具体来说，给定一个图像 I，通过预训练模型 Faster R-CNN 提取一组特征序列$\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_k\}$，其中$\boldsymbol{f}_k\in\mathbb{R}^D$，$k$ 代表图片中物体的数量。由于物体特征边界框并没有捕捉到动作级、关系级以及场景级等多粒度信息，因此网格特征被引入作为图片多样化信息的补充。本文使用预训练模型 ResNet-152[27]获得图片的网格特征$\boldsymbol{g}_j\in\mathbb{R}^D$，网格大小为 $N\times N(N=7)$。之后通过线性变化将上述两种特征映射到联合嵌入空间，如式(1)所示：

$$\boldsymbol{o}_k=W_o\boldsymbol{f}_k+b_o,\boldsymbol{p}_j=W_p\boldsymbol{g}_j+b_p \tag{1}$$

其中，W_o,b_o,W_p,b_p为可学习参数。最终，物体特征和网格特征归一化之后的向量分别表示为：$\boldsymbol{O}=[\boldsymbol{o}_1,\cdots,\boldsymbol{o}_k]\in\mathbb{R}^{k\times d}$，$\boldsymbol{P}=[\boldsymbol{p}_1,\cdots,\boldsymbol{p}_j]\in\mathbb{R}^{j\times d}$。将物体特征 $\boldsymbol{O}$ 和网格特征 $\boldsymbol{P}$ 拼接起来，得到中间视觉特征 $\boldsymbol{H}=[\boldsymbol{h}_1,\cdots,\boldsymbol{h}_g]\in\mathbb{R}^{g\times d}$，其中 $g=k+j$，表示一个图像中物体个数和网格数的总和。

文本特征通过一个双向 GRU(Bi-GRU)文本编码器获得。给定一个含有N_w个单词的句子，通过 Glove[28]对句子进行初始化处理，得到相应的词嵌入序列$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{N_w}\}$。将这些词嵌入序列输入到文本编码器 Bi-GRU 中，得到如下结果：

$$\vec{w}_i=\overrightarrow{GRU}(\boldsymbol{x}_i,\overrightarrow{w_{i-1}},),\overleftarrow{w}_i=\overleftarrow{GRU}(\boldsymbol{x}_i,\overleftarrow{w_{i+1}}) \tag{2}$$

其中，$\vec{w}_i,\overleftarrow{w}_i$分别表示之前隐藏层的状态。最终，文本表征 $\boldsymbol{T}$ 是前向和后向 GRU 隐藏层状态的平均值，其计算过程如下：

$$\boldsymbol{w}_i=(\vec{w}_i+\overleftarrow{w}_i)/2,\boldsymbol{T}=\frac{1}{N_w}\sum_{i=1}^{N_w}\boldsymbol{w}_i \tag{3}$$

此外，本文采用自注意机制来获得增强之后的全局视觉特征和全局文本特征。这种自注意力机制集中关注片段级特征信息，增强两种模态的潜在语义信息。全局视觉特征$\boldsymbol{V}^G\in\mathbb{R}^d$的计算式如下：

$$a_g^v=\frac{\exp(W^v\boldsymbol{h}_g)}{\sum_{g=1}^{j+k}\exp(W^v\boldsymbol{h}_g)},\boldsymbol{V}_G=\sum_{g=1}^{j+k}a_g^v\boldsymbol{h}_g \tag{4}$$

其中，$\boldsymbol{h}_g$表示中间视觉特征，采用相同的方式可以得到全局文本特征 $\boldsymbol{T}^G\in\mathbb{R}^d$。

3.2 粒度感知的特征选择器

粒度感知的特征选择通过自适应门控融合和金字塔空洞卷积，从中间视觉特征中系统地学习丰富的多粒度信息。自适应门控融合部分结合了多头注意力机制和门控操作，允许来自网格特征和物体特征中的不同视觉信息多次组合，同时控制内部信息流，自适应地捕获上下文信息，精炼出更具有适应性的、效果增强的视觉表征。同时，为了整合效果增强的视觉表征中存在的多粒度信息，采用金字塔空洞卷积来筛选出多个尺度的重要视觉表征。

多头注意力机制一般通过点积从不同方面获取注意力特征，其原始公式如下：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{d_k})\boldsymbol{V} \tag{5}$$

其中，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 分别表示查询(query)、键值(key)和对应值(value)的信息，d_k 表示 query 的维度。与一般多头注意力机制

不同，本文增加了一个门控操作，用于有效地平衡中间视觉特征 $\boldsymbol{H}$ 中不同内容的相对重要性。通过线性映射，将中间视觉特征 $\boldsymbol{H}$ 转化为查询向量和键值向量，两者点对点相乘得到门控 $\boldsymbol{G}$(Gate)。其计算式如下：

$$\boldsymbol{G}=\mathrm{sigmoid}((\boldsymbol{Q}W_1+b_1)\cdot(\boldsymbol{K}W_2+b_2)) \tag{6}$$

其中，W_1，b_1，W_2，b_2 都是可学习参数，$\odot$ 表示点乘。最后效果增强的视觉表征 $\widetilde{\boldsymbol{H}}\in\mathbb{R}^{(k+j)\times d}$ 可以通过式(7)计算得到。

$$\begin{aligned}&\widetilde{\boldsymbol{Q}}=\boldsymbol{Q}\odot\boldsymbol{G},\widetilde{\boldsymbol{K}}=\boldsymbol{K}\odot\boldsymbol{G}\\&\widetilde{\boldsymbol{H}}=Attention(\widetilde{\boldsymbol{Q}},\widetilde{\boldsymbol{K}},\boldsymbol{V})\end{aligned} \tag{7}$$

其中，$\boldsymbol{V}$ 通过对中间视觉特征 $\boldsymbol{H}$ 进行线性映射得到。

为了从效果增强的视觉表征 $\widetilde{\boldsymbol{H}}$ 中选择多粒度信息，本文采用了金字塔空洞卷积结构[29]。此方法原用于解决图像语义分割中像素点级别的密集预测分类，通过空洞卷积进行多尺度上下文信息聚合而不降低特征图的大小，且扩张卷积支持感受野的指数增长。本文利用此技术，将物体和网格视觉特征进行多尺度上下文信息聚合，深入挖掘多粒度信息。具体来说，金字塔空洞卷积结构包含了从上到下不同大小的平行空洞卷积核，卷积核的大小以降序排列。通过调整卷积核空洞率来控制感受野的大小，以便检测到不同粒度的信息。每层的卷积输出结果被定义为 m_s，将每层输出结果拼接起来，本文得到一个中间分数向量 $\boldsymbol{M}=[m_1,\cdots,m_s]\in\mathbb{R}^{s\times d}$。此外，本文采用了一个全连接层和 softmax 操作来处理中间分数向量，并得到聚合之后的向量。

$$\bar{\boldsymbol{m}}=\frac{\exp(W_m m_s+b_m)}{\sum_{s=1}^{N_m}\exp(W_m m_s+b_m)} \tag{8}$$

其中，W_m，b_m 是可学习参数，N_m 是金字塔空洞卷积的层数。金字塔空洞卷积分数向量最终表示为 $\bar{\boldsymbol{M}}\in\mathbb{R}^{(k+j)\times N_m}$。粒度感知的视觉线索可以通过式(9)计算得到。

$$\boldsymbol{V}^H=\bar{\boldsymbol{M}}^{\mathrm{T}}\otimes\widetilde{\boldsymbol{H}} \tag{9}$$

其中，$\boldsymbol{V}^H\in\mathbb{R}^{N_m\times d}$，$\otimes$ 表示矩阵乘法。

3.3 语义聚合模块

语义聚合模块旨在将多粒度视觉线索和文本线索嵌入到一个共同语义空间，隐式地进行信息聚合，并减小模态间的语义差距。本文采用了一个即插即用的基于 NetVLAD[23] 改造的网络，用于聚类生成两个模态的局部向量。对于一张给定的图片，其视觉特征可表示为 $\boldsymbol{X}\in\mathbb{R}^{N\times D}$，使用 NetVLAD 将视觉特征 $\boldsymbol{X}$ 聚类到 $\boldsymbol{Y}$，可以表示为 $\boldsymbol{Y}=f_{\mathrm{NetVLAD}}(\boldsymbol{X})$，其中 $\boldsymbol{Y}\in\mathbb{R}^{L\times D}$，它通过软分配加权的方式计算每个向量 $\boldsymbol{x}_i$ 和 L 个可学习的聚类中心之间的残差，具体的计算式如下：

$$\begin{aligned}&f_{\mathrm{NetVLAD}}(\boldsymbol{X})=\left\|\sum_{i=1}^{N}\bar{a}_l(\boldsymbol{x}_i)(\boldsymbol{x}_i-c_l)\right\|_2^2\\&\bar{a}_l=\frac{\exp(w_l^{\mathrm{T}}\boldsymbol{x}_i+b_l)}{\sum_{l=1}^{L}\exp(w_l{}^{\mathrm{T}}\boldsymbol{x}_i+b_l)}\in(0,1)\end{aligned} \tag{10}$$

其中，W_l，b_l 是权重和偏差，$\bar{a}_l$ 表示软分配操作所得的权重，c_l 表示第 l 个可学习的聚类中心。

与 NetVLAD 不同的是，本文将原来的卷积层替换为深度可分离卷积，通过更少的乘法运算来学习聚合更多的潜在语义信息，修改之后的网络被命名为 S-NetVLAD。这里将 $\boldsymbol{V}^H$ 称为视觉粒度感知线索，将 $\boldsymbol{T}$ 称为文本粒度感知线索，这些线索被聚合到本地共享语义空间中，并表示为残差向量。具体地，将 S-NetVLAD 作为文本和视觉的共享参数模块，来聚合 $\boldsymbol{V}^H$ 和 $\boldsymbol{T}$ 到 L 个可学习的语义共享中心。共享的 S-NetVLAD 提高了局部特征聚合过程的可解释性，并且易于插入。最后，在每个语义共享中心上，经过归一化处理之后的局部残差向量被表示为：

$$\boldsymbol{T}^L=f_{\text{S-NetVLAD}}(\boldsymbol{T}),\boldsymbol{V}^L=f_{\text{S-NetVLAD}}(\boldsymbol{V}^H) \tag{11}$$

其中，$\boldsymbol{T}^L\in\mathbb{R}^{L\times d}$ 和 $\boldsymbol{V}^L\in\mathbb{R}^{L\times d}$，分别表示文本局部特征和视觉局部特征。

3.4 损失函数

本节介绍了模型的损失函数。首先将视觉和文本的全局特征和局部特征分别拼接到一起，作为最终特征表示 $\boldsymbol{V}^F$ 和 $\boldsymbol{T}^F$，结合了局部特征和全局特征的优势，最大限度地减小了它们的信息损失。其次，利用一个对比学习损失函数来拉近正样本对，推开负样本对。具体来说，在小批量训练集中，正样本对 $(\boldsymbol{v}^+,\boldsymbol{t}^+)$ 最难的负样本对 $(\boldsymbol{v}^+,\boldsymbol{t}^-)$ 以及 $(\boldsymbol{v}^-,\boldsymbol{t}^+)$ 将被推离，且距离远远大于预设好的边界限制 Δ(margin)。最终，损失函数的计算式如下：

$$\begin{aligned}L_{\mathrm{rank}}=&[\Delta+s(\boldsymbol{v}^+,\boldsymbol{t}^-)-s(\boldsymbol{v}^+,\boldsymbol{t}^+)]_+\\&+[\Delta+s(\boldsymbol{v}^-,\boldsymbol{t}^+)-s(\boldsymbol{v}^+,\boldsymbol{t}^+)]_+\end{aligned} \tag{12}$$

其中，$s(\cdot)$ 表示用余弦相似度计算两个样本之间的距离，在一个小批量训练集中，最困难负样本可以表示为：$t^-=\arg\max_{j\neq t}s(v,j)$，$v^-=\arg\max_{i\neq v}s(i,t)$。

4 实验及结果分析

本文在两个公开数据集 MSCOCO 和 Flickr30K 上评估了所提模型 GASA。本节首先介绍了数据集、评价指标以及实验环境和实施细节，然后进行了量化分析，将本文方法与最先进的方法进行比较，展示了模型的整体性能，最后用消融实验验证了每个模块的有效性。

4.1 实验设置

4.1.1 数据集

本文主要采用两个数据集进行实验。数据集 MS-COCO[30] 是最流行的大规模图文匹配任务的数据集之一，它包括 123 287 个图像-文本对，每张图像有 5 个人工注释，其中 113 287 张用于训练、5 000 张用于验证、5 000 张用于测试。数据集 Flickr30K[31] 由网站收集的 31 783 张图片组成，其内容涵盖事件、活动和场景，每张图片都有 5 个文本描述，平均句子长度为 10.5 个单词，其中 1 000 张图片用于测试，1 000 张图片用于验证，其余 28 783 张图片用于训练。

4.1.2 评价指标

本文采用 K 召回率($R@K$)和召回率之和($R@\mathrm{sum}$)来衡量模型的性能。$R@K$ 表示在前 top K 检索结果中，正确样本出现在前 K 个返回结果的样例占总样例的比例，本文中使用 $K=1,5,10$。$R@sum$ 表示所有图像检索和文本检索的召回率总和。K 召回率和召回率之和越大，模型的性能就越好。

4.1.3 实验环境和实施细节

本文提出的 GASA 模型在 MSCOCO 和 Flickr30K 上分别训练了 30 个 epoch 和 50 个 epoch，仅需 1 张 NVIDIA GeForce 3090 24GB 显卡。在训练过程中，一个小批量数据集大小为 128，损失函数中的 Δ 被固定为 0.2，初始化学习率为 0.0005，每 5 轮迭代之后学习率衰减为原来的 1/2。对于文本编码，使用 Glove 模型来嵌入给定句子中的每个单词，其维度为 300。对于图像编码，一张图片中检测到的物体数量 k 固定为 36，网格总数固定为 49，即 $N=7$。物体特征和网格特征的向量维度 $D=2048$，文本和视觉联合空间维度 $d=1024$。对于 MS-COCO，金字塔空洞卷积层数 $N_m=36$，语义共享中心 $L=12$。对于 Flickr30k，$N_m=12$，$L=5$。

4.2 定量指标实验结果

本文在 MS-COCO 和 Flickr30K 两个数据集上进行了测试，与现有多种方法进行了定量指标比较，其中包含经典模型 SCAN[14]（2018），VSRN[16]（2019）等，以及新模型 CVSE[22]（2020），SMFEA[19]（2021）等。该方法在图像检索和文本检索上都取得了最优性能，结果如表 1 所列。表 1 中，$R@K$ 指标在大规模数据集 MS-COCO 上取得了较大的优势，说明了 GASA 在复杂场景中更具备泛化能力和稳定性。在 MS-COCO 1K 数据集上，GASA 在大多数评价指标上都优于目前最优的模型，在文本检索和图像检索方面的 $R@1$ 结果分别为 77.9% 和 63.4%。与 CAMERA * 相比，在文本检索方面，$R@1$，$R@5$，$R@10$ 取得了明显的提升（2.0%，1.0%，0.5%），在图像检索上也有明显的改善（1.1%，0.6%，0.8%）。本文在视觉多粒度信息提取方面与 CAMERA * 有相似之处，但 CAMERA * 并没有对多粒度语义信息进行聚类，这导致其性能不如模型 GASA。此外，GASA 的召回率总和超过了所有对比模型，说明了 GASA 具有明显的性能优势，能够充分缩小模态间的语义鸿沟。在 MS-COCO 5K 数据集上，GASA 模型也表现良好，如表 2 所列，在文本检索和图像检索的 $R@1$ 上分别为 56.7% 和 42.3%，远超过了其他对比模型。

表 1 GASA 和对比方法在 MS-COCO 1K 和 Flickr30K 上的结果

Table 1 Results of GASA and comparative methods on MS-COCO 1K and Flickr30K

（单位：%）

Method	MS-COCO 1K							Flickr30K						
	Text Retrieval			Image Retrieval			$R@sum$	Text Retrieval			Image Retrieval			$R@sum$
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$		$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
VSE++[12]	64.6	90.0	95.7	52.0	84.3	92.0	478.6	52.9	80.5	87.2	39.6	70.1	79.5	409.8
SCAN[14]	72.7	94.8	98.4	58.8	88.4	94.8	493.9	67.4	90.3	95.8	48.6	77.7	85.2	465.0
VSRN[13]	76.2	94.8	98.2	62.8	89.7	95.1	516.8	70.4	89.2	93.7	53.0	77.9	85.7	469.9
PFAN[21]	76.5	96.3	99.0	61.6	89.6	95.2	510.7	70.0	91.8	95.0	50.4	78.7	86.1	472.0
CAMP[32]	72.3	94.8	98.3	58.5	87.9	95.0	506.8	68.1	89.7	95.2	51.5	77.1	85.3	466.9
CAMERA *[17]	75.9	95.5	98.6	62.3	90.1	95.2	517.6	72.0	92.0	96.6	54.5	81.9	89.0	485.9
CVSE[22]	74.8	95.1	98.3	59.9	89.4	95.2	512.3	73.5	92.1	95.8	52.9	80.4	87.8	482.5
CAAN[33]	75.5	95.4	98.5	61.3	89.7	95.2	515.6	70.1	91.6	97.2	52.8	79.0	87.9	478.6
ABGR[34]	73.0	94.7	98.3	59.5	89.4	95.2	510.1	72.3	91.8	95.1	53.7	80.1	87.2	480.2
SMFEA[19]	75.1	95.4	98.3	62.5	90.1	96.2	517.6	73.7	92.5	96.1	54.7	82.1	88.4	487.5
IMRAM[18]	76.7	95.6	98.5	61.7	89.1	95.0	516.6	74.1	93.0	96.6	53.9	79.4	87.2	484.2
GASA	**77.9**	**96.5**	**98.8**	**63.4**	**90.7**	**96.0**	**523.3**	**74.9**	**92.7**	**96.8**	**55.3**	**82.5**	**89.3**	**491.6**

注：CAMERA * 表示为了公平起见，本文选择没有文字增强的结果

表 2 GASA 和对比方法在 MS-COCO 5K 上的结果

Table 2 Results of GASA and comparative methods on MS-COCO 5K

（单位：%）

Method	MS-COCO 5K					
	Text Retrieval			Image Retrieval		
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$
VSE++[12]	41.3	69.1	77.9	31.4	60.0	71.2
SCAN[14]	50.4	82.2	90.0	38.6	69.3	80.4
VSRN[13]	53.0	81.1	89.4	40.5	70.6	81.1
PFAN[21]	50.8	83.9	89.1	39.5	69.5	80.8
CAMP[32]	50.1	82.1	89.7	39.0	68.9	80.2
CAMERA	53.1	81.3	89.8	39.0	70.5	81.5
CVSE[22]	52.5	83.3	90.9	41.2	70.3	82.9
CAAN[33]	52.5	83.3	90.9	41.2	70.3	82.9
SMFEA[19]	54.2	—	89.9	41.9	—	83.7
IMRAM[18]	53.7	83.2	91.0	39.7	69.1	79.8
GASA	**56.7**	**84.8**	**91.8**	**42.3**	**71.2**	**83.1**

GASA 和对比方法在 Flickr30K 数据集上的研究结果如表 1 所列。本文的 GASA 在文本检索的 $R@1$ 标准上达到了 74.9%，超过次优模型 0.8%。在图像检索和文本检索方面，本文模型超过了之前最好的方法 IMRAM，特别是在图像检索上，其在 $R@1$，$R@5$，$R@10$ 上分别高出 1.4%，3.1%，2.1%。上述结果证明了利用多粒度信息和语义聚合的有效性和必要性。针对文本检索，GASA 模型在指标 $R@5$，$R@10$ 上略低于 IMRAM，原因是该模型在训练阶段采用复杂的损失函数进行多次迭代匹配，筛选出了最困难的负样本，导致训练参数过拟合。该方式针对小数据集有一定效果，但在大数据集上无效，为了保证训练过程尽可能简单，本文没有采取此类方式。尽管如此，本文的 GASA 模型仍然在大多数评价指标上取得了最好的结果。

4.3 消融实验结果

为了研究 GASA 中不同模块对模型性能的贡献，本文在 MS-COCO 1K 数据集上进行了消融实验，结果如表 3 所列。

表 3 消融实验的结果

Table 3 Results of ablation study

（单位：%）

Method	Text-to-Image Retrieval		
	$R@1$	$R@5$	$R@10$
a) baseline	72.8	95.1	98.4
b) w/o GA	73.8	95.3	98.6
c) w/o SA	75.3	96.1	98.8
d) GASA	**77.9**	**96.5**	**98.8**

表 3 中，a)baseline 表示以 VSE＋＋网络为基线，将物体特征和网格特征作为视觉输入，文本特征由 Bi-GRU 编码得到，只使用自注意机制来获得全局特征；b)w/o GA 表示从 GASA 中删除了粒度感知的特征选择器；c)w/o SA 表示移除 GASA 中的语义聚合模块；d)GASA 表示论文中提出的完整模型。比较 b)和 d)，将 GA 从完整 GASA 模型中去除，文本到图像的检索结果在 $R@1$ 上减少了 4.1%，验证了 GA 模块在获取详细视觉区域的多粒度信息方面具有重要作用。与 d)相比，c)在 $R@1$ 上减少了 2.6%，这说明语义聚合模块有助于将两个模态的多粒度描述符在同一语义空间中进行有效聚类。与基线模型 a)相比，GASA 在 $R@1$ 上比基线模型增加了 5.1%，清晰地展示了本文模型在跨模态检索上具有良好性能，并表明多粒度信息挖掘和语义聚合是相辅相成的方法。

结束语 本文提出了一个粒度感知和语义聚合(GASA)模型。该模型挖掘图像中存在的多粒度信息，聚合两个模态中的相同语义信息，隐式地计算视觉线索与文本线索的相似度，并缩小两种模态之间的语义差距。具体来说，一个粒度感知的特征选择器被引入，在自适应门控融合机制和金字塔空洞卷积结构的指导下，充分提取筛选视觉中的多粒度信息。此外，语义聚合模块将视觉线索和文本线索联合嵌入到一个共同语义空间中，缩小模态间的语义鸿沟。实验结果表明，与最先进的方法相比，本文模型取得了非常有竞争力的结果。

在现有研究方法的基础上进行分析与展望发现，还存在以下具有挑战性的内容。

(1)挖掘文本多粒度信息。图文匹配任务中，图像和句子间的语义不匹配通常发生在短语级别，除了给图像自动构建多粒度语义标签，引入额外短语级监督信息也有重要意义。显式分层次匹配名词短语、动词短语等文本特征和视觉多粒度信息，有利于提升匹配结果。

(2)构建模态内多粒度信息关系图。构建两个模态的场景图，提取视觉与文本中多种形式的实体和三元组作为标签。基于这些多粒度的语义标签，匹配模型能更好地识别出细粒度不匹配的语义单元。

下一步的工作将探索如何提取文本的多层次多粒度信息，构建模态内不同粒度信息的关系图，实现更精细的分层次图文匹配。

参 考 文 献

[1] ZENG P,GAO L,LYU X,et al. Conceptual and syntactical cross-modal alignment with cross-level consistency for image-text matching [C]// Proceedings of the 29th ACM International Conference on Multimedia. 2021:2205-2213.

[2] FENG X,HU Z Y,LIU C H. Survey of Research Progress on Cross modal Retrieval [J]. Computer Science, 2021,48(8):13-23.

[3] FENG Y G,CAI G Y. Cross-modal Retrieval Fusing Multilayer Semantics[J]. Computer Science,2019,46(3):227-233.

[4] WANG W,CHEN Z,HU H. Hierarchical attention network for image captioning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019:8957-8964.

[5] YAN Y,ZHUANG N,NI B,et al. Fine-grained Video Captioning via Graph-based Multi-granularity Interaction Learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2022,44(2):666-683.

[6] LEI Y,HE Z,ZENG P,et al. Hierarchical Representation Network With Auxiliary Tasks For Video Captioning[C]// 2021 IEEE International Conference on Multimedia and Expo(ICME). IEEE,2021.

[7] SEO A,KANG G C,PARK J,et al. Attend What You Need: Motion-Appearance Synergistic Networks for Video Question Answering[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing(Volume 1:Long Papers). 2021:6167-6177.

[8] WANG H,GUO D,HUA X S,et al. Pairwise VLAD Interaction Network for Video Question Answering[C]// Proceedings of the 29th ACM International Conference on Multimedia. 2021:5119-5127.

[9] GAO L,ZENG P,SONG J,et al. Structured two-stream attention network for video question answering[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 6391-6398.

[10] GAO L,LEI Y,ZENG P,et al. Hierarchical Representation Network With Auxiliary Tasks for Video Captioning and Video Question Answering[J]. IEEE Transactions on Image Processing,2022,31:202-215.

[11] KARPATHY A,JOULIN A,LI F. Deep fragment embeddings for bidirectional image sentence mapping[J]. Advances in Neural Information Processing Systems,2014,2:1889-1897.

[12] FAGHRI F,FLEET D J,KIROS J R,et al. Vse＋＋:Improving visual-semantic embeddings with hard negatives[C]// Proceedings of the British Machine Vision Conference(BMVC). 2018.

[13] DENG Y J,ZHANG F L,CHEN X Q,et al. Collaborative Attention Network Model for Cross-modal Retrieval[J]. Computer Science,2020,47(4):54-59.

[14] LEE K H,CHEN X,HUA G,et al. Stacked cross attention for image-text matching[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018:201-216.

[15] WU Y,WANG S,SONG G,et al. Learning fragment self-attention embeddings for image-text matching[C]// Proceedings of the 27th ACM International Conference on Multimedia. 2019: 2088-2096.

[16] LI K,ZHANG Y,LI K,et al. Visual semantic reasoning for image-text matching[C]// Proceedings of the IEEE/CVF International conference on computer vision. 2019:4654-4662.

[17] QU L,LIU M,CAO D,et al. Context-aware multi-view summarization network for image-text matching[C]// Proceedings of the 28th ACM International Conference on Multimedia. 2020: 1047-1055.

[18] CHEN H,DING G,LIU X,et al. Imram:Iterative matching with recurrent attention memory for cross-modal image-text retrieval[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:12655-12663.

[19] GE X,CHEN F,JOSE J M,et al. Structured Multi-modal Fea-

ture Embedding and Alignment for Image-Sentence Retrieval [C]// Proceedings of the 29th ACM International Conference on Multimedia. 2021:5185-5193.

[20] WANG X, ZHU L, YANG Y. T2vlad: global-local sequence alignmentfor text-video retrieval[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:5079-5088.

[21] WANG Y, YANG H, QIAN X, et al. Position focused attention network for image-text matching [C] // Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. 2019:3792-3798.

[22] WANG H, ZHANG Y, JI Z, et al. Consensus-aware visual-semantic embedding for image-text matching[C]// European Conference on Computer Vision. Cham: Springer, 2020:18-34.

[23] ARANDJELOVIC R, GRONAT P, TORII A, et al. NetVLAD: CNN architecture for weakly supervised place recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:5297-5307.

[24] UY M A, LEE G H. Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:4470-4479.

[25] HAUSLER S, GARG S, XU M, et al. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:14141-14152.

[26] ANDERSON P, HE X, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:6077-6086.

[27] JIANG H, MISRA I, ROHRBACH M, et al. In defense of grid features for visual question answering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:10267-10276.

[28] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP). 2014:1532-1543.

[29] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[J]. arXiv:1511.07122, 2015.

[30] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]// European Conference on Computer Vision. Cham: Springer, 2014:740-755.

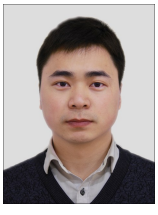
[31] PLUMMER B A, WANG L, CERVANTES C M, et al. Flickr 30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015:2641-2649.

[32] WANG Z, LIU X, LI H, et al. Camp: Cross-modal adaptive message passing for text-image retrieval[C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:5764-5773.

[33] ZHANG Q, LEI Z, ZHANG Z, et al. Context-aware attention network for image-text retrieval[C]// Proceedings of theIEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:3536-3545.

[34] ZHONG X, YANG Z, YE M, et al. Auxiliary bi-level graph representation for cross-modal image-text retrieval[C]// 2021 IEEE International Conference on Multimedia and Expo(ICME). IEEE, 2021:1-6.

MIAO Lan-xin, born in 1998, postgraduate. Her main research interests include cross-modal retrieval, computer vision and machine learning.



SONG Jing-kuan, born in 1986, Ph. D, professor, associate editor. His main research interests include large-scale multi-media retrieval, image/video segmentation and image/video understanding using hashing, graph learning, and deep learning techniques.

(责任编辑:喻藜)