文章编号：1003-0530(2022)02-0285-15

# 基于堆叠交叉注意力的图像文本跨模态匹配方法

王红斌[1,2] 张志亮[1,2] 李华锋[1,2]
（1. 昆明理工大学信息工程与自动化学院，云南昆明 650500；
2. 昆明理工大学云南省人工智能重点实验室，云南昆明 650500）

**摘 要：**图像文本跨模态匹配是计算机视觉与自然语言处理交叉领域的一项重要任务，然而传统的图像文本跨模态匹配方法要么只考虑到全局图像与全局文本匹配，要么只考虑到局部图像与局部文本匹配，无法全面有效的考虑局部和全局信息，导致提取出来的特征信息不完善。或者只是简单的对全局图像与全局文本特征进行提取，局部细节信息无法凸显，导致全局特征无法充分表达其全局语义信息。针对该问题，本文提出一种基于堆叠交叉注意力的图像文本跨模态匹配方法。该方法在考虑局部图像与局部文本匹配的同时，将堆叠交叉注意力引进全局图像与全局文本匹配，通过注意力来进一步挖掘全局特征信息，让全局图像与全局文本特征得到优化，从而提升图像文本跨模态检索的效果。在 Flickr30K 和 MS-COCO 两个公共数据集上进行了实验验证，模型的总体性能 R@sum（Recall@sum）较 baseline（SCAN）分别提高了 3.9% 与 3.7%。该模型与 SCAN 模型相比，R@sum 表现较好。由此表明本文提出方法在图像文本跨模态检索任务上的有效性，并且与现有方法相比具有一定的优越性。

**关键词：**跨模态匹配；局部细节信息；全局语义信息；堆叠交叉注意力；图像文本特征
**中图分类号：**TP391.3 **文献标识码：**A **DOI：**10.16798/j.issn.1003-0530.2022.02.008

**引用格式：**王红斌,张志亮,李华锋. 基于堆叠交叉注意力的图像文本跨模态匹配方法[J]. 信号处理,2022,38(2):285-299. DOI:10.16798/j.issn.1003-0530.2022.02.008.
**Reference format:** WANG Hongbin, ZHANG Zhiliang, LI Huafeng. Image-text cross-modal matching method based on stacked cross attention[J]. Journal of Signal Processing,2022,38(2):285-299. DOI:10.16798/j.issn.1003-0530.2022.02.008.

# Image-text Cross-modal Matching Method Based on Stacked Cross Attention

WANG Hongbin[1,2] ZHANG Zhiliang[1,2] LI Huafeng[1,2]
（1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China; 2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, Yunnan 650500, China）

**Abstract:** Cross-modal matching of image-text is an important task in the intersection of computer vision and natural language processing. However, traditional image-text cross-modal matching methods either only consider global image and global text matching, or only consider local image and local text matching, it cannot fully and effectively consider local and global information, resulting in imperfect feature information extracted. Or simply extracting global image and global text features, local details cannot be highlighted, resulting in global features unable to fully express their global semantic information. To solve this problem, this paper proposes a cross-modal matching method for image-text based on stacked cross attention. While considering the matching of local images and local text, this method introduces stacked cross attention into global image and global text matching, and further mines global feature information through attention, so that

收稿日期：2021-05-06；修回日期：2021-06-23
基金项目：国家自然科学基金(61966021,61966020)

global image and global text features are optimized, thereby improving image-text the performance of cross-modal retrieval. Experimental verification was carried out on the two public datasets of Flickr30K and MS-COCO. Experiments were carried out on the two public datasets of Flickr30K and MS-COCO. The overall performance of the model R@sum (Recall@sum) was improved by 3.9% and 3.7% respectively compared with the baseline (SCAN). Compared with the SCAN model, the R@sum performance of this model is better. This shows that the method proposed in this paper is effective in the task of cross-modal retrieval of image-text, and it has certain advantages compared with the existing methods.

**Key words:** cross-modal matching; local detail information; global semantic information; stacked cross attention; image-text features

## 1 引言

大数据时代的到来,生活中出现了许多不同模态的数据,例如:文本、图像、视频以及语音正在以前所未有的速度增长。图像文本跨模态检索是计算机视觉与自然语言处理交叉领域的一项重要任务。所谓图像文本跨模态检索,是指依据给定的图像去检索与图像内容描述相吻合的文本,或者根据给定的文本去检索与文本内容描述相吻合的图像[1-6]。以前的任务都是输入文本去检索相关的文本信息[7-8],或者输入图像去检索相关的图像信息[9-10],这两个任务都是在同一种模态下进行的,我们称其为“单模态”[11]。而使用图像去检索相关的文本信息,利用文本去检索相关的图像信息,由于输入信息与获得信息的类型不同,所以我们把这个任务称作为“跨模态”[12]。它能应用于搜索领域,输入文本描述能检索出对应的图像,或者输入图像能检索出对应的文本描述。因此研究图像文本跨模态检索是一项重要而有意义的工作。

传统的图像文本跨模态检索要么只提取全局特征,进行全局图像与全局文本匹配[13],要么只提取局部特征并进行局部图像与局部文本匹配[14],导致提取出的特征信息并不完善。图像文本跨模态检索的效果主要取决于对特征信息的挖掘是否充分、全面。因此,传统的方法集中通过提出的注意力去充分挖掘局部特征的信息,从而提升图像文本跨模态检索的效果。但是,仍然有两大问题急需解决:1)由于图像文本跨模态检索无法全面有效的考虑局部和全局信息,导致提取出来的特征信息不完善。2)只是简单的对全局图像与全局文本特征进行提取,局部细节信息无法凸显,导致全局特征无法充分表达其全局语义信息。

针对以上问题,本文提出了一种基于堆叠交叉注意力的图像文本跨模态匹配方法。该方法是在利用注意力挖掘局部特征信息的同时;利用注意力去挖掘全局特征信息来提升图像文本跨模态检索的效果。本文的主要贡献有:

1)综合了局部图像与局部文本匹配以及全局图像与全局文本匹配的优势,有效的考虑局部和全局信息,使得提取出来的特征信息更加完善,同时捕捉到细粒度信息的对齐和全局信息的对齐。

2)将堆叠交叉注意力引进全局图像与全局文本匹配,通过注意力来进一步挖掘全局特征信息,让全局图像与全局文本特征得到优化,使其更能表达全局语义信息,从而提升图像文本跨模态检索的效果。

3)本文方法在Flickr30K和MS-COCO两个公共数据集上进行了实验验证,实验结果表明本文模型的R@1召回率与SCAN模型相比表现较好。本文方法在Flickr30K数据集上,模型的总体性能R@sum(Recall@sum)较baseline(SCAN)提高了3.9%,R@1图像检索文本较baseline提高了0.9%,R@1文本检索图像较baseline提高了1.6%。在MS-COCO数据集上,模型的总体性能R@sum较baseline提高了3.7%,R@1图像检索文本较baseline提高了1.7%,R@1文本检索图像较baseline提高了0.6%。

## 2 相关工作

图像文本跨模态检索其目的是为了探索图像与文本之间的潜在对齐。目前,根据已有图像文本跨模态匹配方法的特点,可粗略地分为五类:1)基于全局的图像文本跨模态匹配,2)基于局部的图像文本跨模态匹配,3)基于全局和局部的图像文本跨模态匹配,4)基于注意力的图像文本跨模态匹配,5)基于哈希变换的图像文本跨模态匹配。

### 2.1 基于全局的图像文本跨模态匹配

基于全局的图像文本跨模态匹配方法一般是提取图像与文本的全局特征并进行全局图像与全局文本匹配,达到提升模型性能的目的。具体地,为解决全局图像和全局文本特征无法充分表达其全局语义信息的问题,Faghri等人[15]在三元组损失函数中引进了难分样本,能够学习出比较好的映射矩阵并更好的

度量图像与文本的相关性。Zheng等人[16]提出了可微调的视觉和文本表示，利用微调之后的全局图像特征与全局文本特征进行匹配学习，从而提升图像文本跨模态检索的效果。Huang等人[17]提出了一种语义增强的图像与文本匹配模型，它能通过学习语义概念来提高图像的表示，然后按照正确的语义顺序组织它们。虽然这类方法考虑了全局特征信息，能较好的实现全局信息的对齐，但它们都缺乏局部图像特征与局部文本特征之间的匹配，没有挖掘到局部特征信息，最终影响图像文本跨模态检索的效果。

### 2.2　基于局部的图像文本跨模态匹配

基于局部的图像文本跨模态匹配方法是近几年较为流行的方法，他们通常是提取图像与文本的局部特征并进行局部图像与局部文本匹配，达到提升模型性能的目的。具体地，为解决图像与文本的局部特征无法得到充分优化的问题，Lee等人[18]提出了一种堆叠交叉注意力用于捕捉图像区域和文本单词之间的潜在对齐。Zheng等人[19]提出了自注意力视觉语义嵌入方法，利用多尺度特征融合技术，从不同的表示尺度获得不同层次的语义概念来进行图像与文本匹配。Wang等人[20]提出了跨模态自适应信息传递模型，能自适应地控制信息传递的信息流，考虑了细粒度的跨模态交互，实现图像文本细粒度信息的对齐。Ji等人[21]提出利用多模态记忆增强注意力网络来定位有意义的语义部分，并且利用记忆网络来捕捉长短时上下文知识，实现图像文本细粒度信息的对齐。虽然这类方法考虑了局部特征信息，能较好的实现细粒度信息的对齐，但它们都缺乏全局图像特征与全局文本特征之间的匹配，没有挖掘到全局特征信息，最终影响图像文本跨模态检索的效果。

### 2.3　基于全局和局部的图像文本跨模态匹配

基于全局和局部的图像文本跨模态匹配方法，由于模型结构复杂且对模型的兼容性要求较高，近几年较为稀少，他们通常是同时提取图像与文本的全局特征和局部特征，进行全局图像与全局文本匹配以及局部图像与局部文本匹配，达到提升模型性能的目的。具体地，为解决全局图像与全局文本匹配以及局部图像与局部文本匹配无法很好的兼容进行共同学习问题，Gu等人[22]提出将生成过程结合到跨模态特征嵌入中，通过该方法不仅可以学习到全局抽象特征，还能学习到局部底层特征。Li等人[23]和Ma等人[24]提出了一种融合两级相似性的跨媒体图像文本检索方法，构建了跨媒体两级网络，进一步探索图像与文本更好的匹配。但是这些方法没有利用注意力对局部特征以及全局特征进行共同优化，导致局部图像与局部文本匹配以及全局图像与全局文本匹配无法很好的兼容进行共同学习，从而影响图像文本跨模态检索的召回率。

### 2.4　基于注意力的图像文本跨模态匹配

基于注意力的图像文本跨模态匹配方法是利用注意力对局部或者全局特征进行优化并进行图像与文本匹配，针对局部特征或者全局特征未能得到充分优化问题，Zheng等人[19]提出了自注意力视觉语义嵌入方法，利用多尺度特征融合技术，从不同的表示尺度获得不同层次的语义概念来进行图像与文本匹配。Ji等人[21]提出利用多模态记忆增强注意力网络来定位有意义的语义部分，并且利用记忆网络来捕捉长短时上下文知识，实现图像文本细粒度信息的对齐。但是这些方法对全局图像与全局文本特征的优化程度仍然不足，导致全局特征无法充分表达其全局语义信息，从而影响图像文本跨模态检索的召回率。

### 2.5　基于哈希变换的图像文本跨模态匹配

基于哈希变换的图像文本跨模态匹配方法是利用不同模态的样本对信息，学习不同模态的哈希变换，将不同模态特征映射到一个汉明二值空间，然后在汉明空间实现快速的跨模态检索。具体地，为解决模态特征二值化过程中出现精度损失问题，Tang等人[25]提出了一种邻近鉴别哈希方法实施邻近相似性搜索，通过利用局部鉴别信息来学习鉴别哈希函数，能够保证相似的图像可以被编码成相同的哈希位。Li等人[26]提出一种弱监督深度指标学习方法，它使用一种渐进的学习方式，通过联合利用来自视觉内容和用户提供的社会图像标签中的异构数据结构来发现知识。Jin等人[27]提出一种新的深度语义多模态哈希网络，它利用二维卷积神经网络作为骨干网络来捕捉空间信息用于图像文本检索，使用三维卷积神经网络作为骨干网络来捕捉时空信息用于视频文本检索。Li等人[28]提出了一种新的语义引导哈希方法以及加上二进制矩阵分解，通过同时探索弱监督丰富的社会贡献信息和底层数据结构来执行更有效的最近邻图像搜索。这类方法都是为了解决模态特征二值化的过程中出现精度损失。而本文模型需解决的问题是图像与文本特征的实值化不完善，与上述方法解决的问题有较大的区别。

从上述相关工作可以看出，为解决图像与文本特征的实值化不完善问题，现有研究工作要么只考虑到全局图像与全局文本匹配，要么只考虑到局部图像与局部文本匹配，无法全面有效的考虑局部和

全局信息，导致提取出来的特征信息不完善。或者只是简单的对全局图像与全局文本特征进行提取，局部细节信息无法凸显，导致全局特征无法充分表达其全局语义信息。针对该问题，本文提出了一种基于堆叠交叉注意力的图像文本跨模态匹配方法，该方法在考虑局部图像与局部文本堆叠交叉注意力匹配的同时，将堆叠交叉注意力引进全局图像与全局文本匹配，通过注意力来进一步挖掘全局特征信息，让全局图像与全局文本特征得到优化，使其更能表达全局语义信息，从而提升图像文本跨模态检索的效果。

## 3 本文模型

本文受Lee等人[18]提出的堆叠交叉注意力(baseline模型)的启发，针对baseline模型无法全面有效的考虑局部和全局信息，导致提取出来的特征信息不完善问题。我们提出了一种新的基于堆叠交叉注意力的图像文本跨模态匹配方法，该方法在局部图像与局部文本堆叠交叉注意力匹配的基础上，首先通过堆叠交叉注意力进一步细化局部特征的权重分布并优化全局特征。其次，通过改进后的相似度融合模块全面有效的考虑局部和全局信息。以下我们将首先介绍baseline模型SCAN，然后详细介绍本文模型。

### 3.1 SCAN模型

Lee等人[18]在ECCV2018提出SCAN模型，该模型分别对图像和文本应用交叉注意力机制，学习比较好的局部图像和局部文本表示，使用局部图像和局部文本作为上下文来获取完整的潜在对齐，然后再在共享的子空间中利用Max of Hinges loss度量图像和文本之间的相似性，如下图1所示。

SCAN模型由跨模态图像文本表示、局部图像到局部文本交叉注意力、局部文本到局部图像交叉注意力、相似度融合四个模块组成，由于该模型仅仅考虑了局部图像特征与局部文本特征之间的匹配，缺乏全局图像特征与全局文本特征之间的匹配，没有挖掘到全局特征信息，导致提取的特征信息不完善。针对该问题，本文提出了一种新的基于堆叠交叉注意力的图像文本跨模态匹配方法，具体模型设计与算法见下节。

### 3.2 本文模型的设计

本文模型在分别对图像和文本应用交叉注意力机制，学习比较好的局部图像和局部文本表示。同时，利用改进后的局部图像到局部文本交叉注意力和改进后的局部文本到局部图像交叉注意力，进一步细化局部特征的权重分布并优化全局特征，如下图2所示。

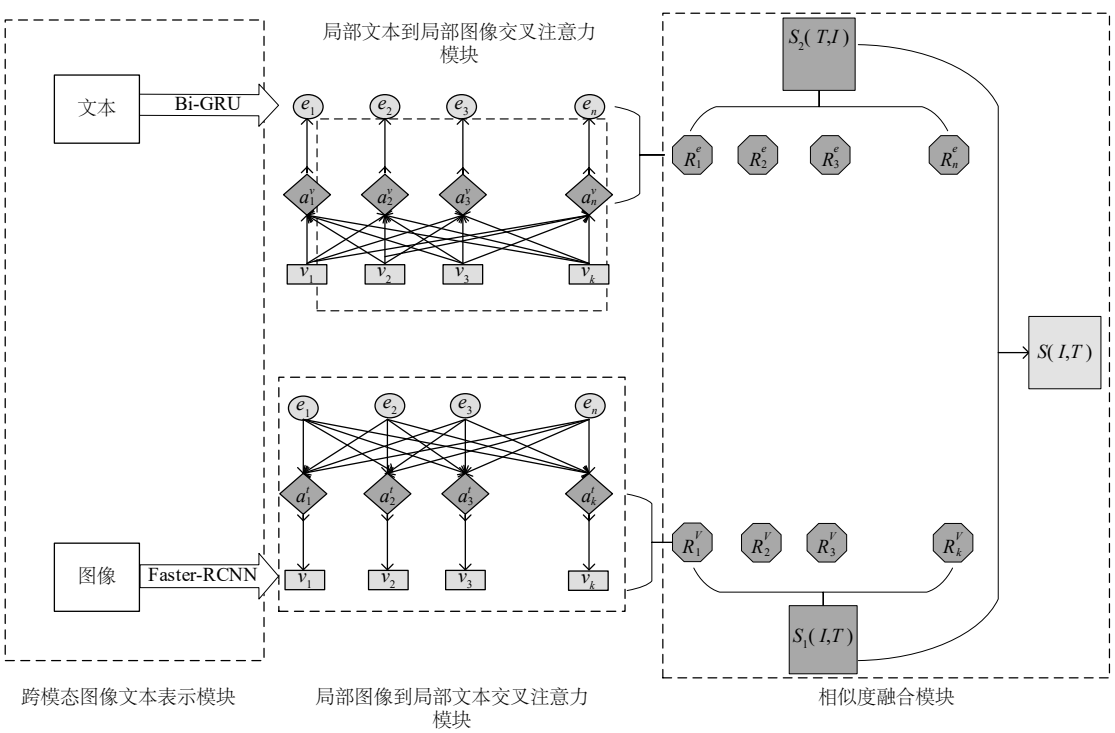


图1 SCAN模型图

Fig. 1 SCAN model map

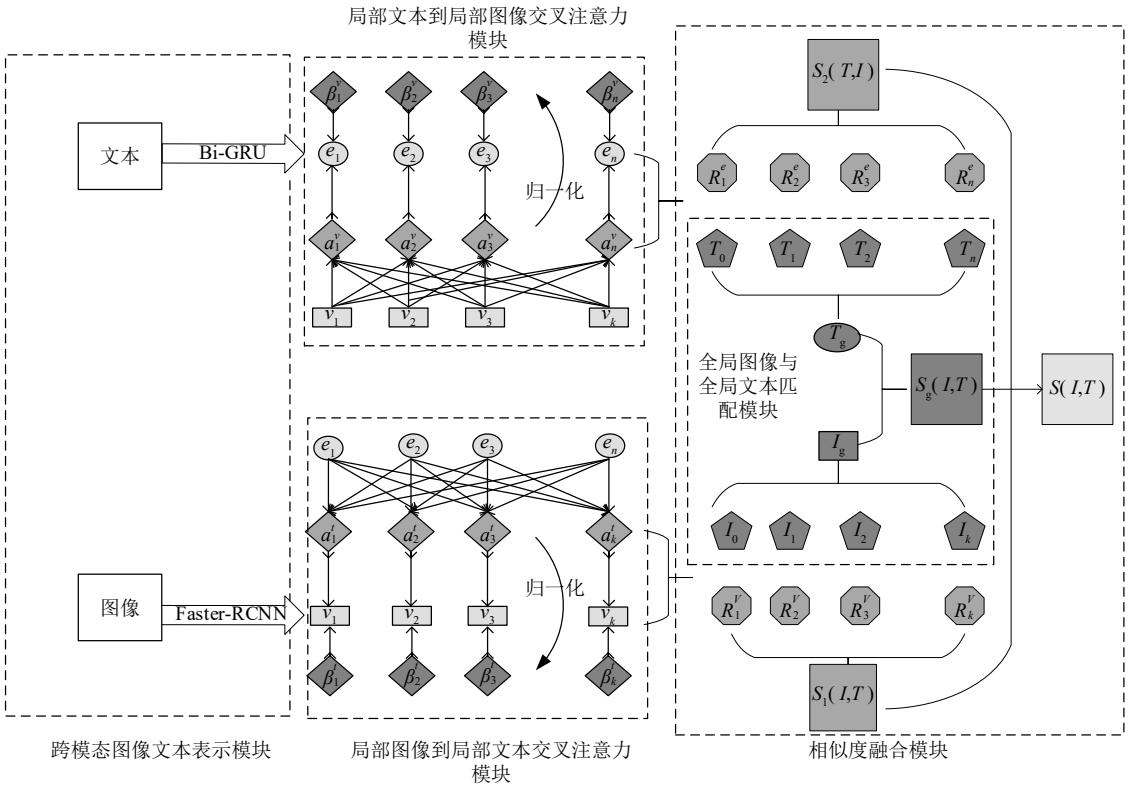


图2　基于堆叠交叉注意力的图像文本跨模态匹配模型图

Fig. 2　Image-text cross-modal matching model map based on stacked cross attention

本文模型主要设计为跨模态图像文本表示、改进后的局部图像到局部文本交叉注意力、改进后的局部文本到局部图像交叉注意力、全局图像与全局文本匹配、改进后的相似度融合五个模块。我们分别叙述改进后的局部图像到局部文本交叉注意力以及改进后的局部文本到局部图像交叉注意力。进一步阐述本文模型设计出全局图像与全局文本匹配模块的有效性以及改进后的相似度融合模块是如何有效的综合局部和全局信息。

### 3.2.1　跨模态图像文本表示

通过对图像与文本分别进行编码，学习出具有上下文信息的局部图像与局部文本表示。为了实现这一目标，我们将文本中的第$j$个单词编码成对应的one-hot向量，再将每个one-hot向量通过嵌入矩阵$\boldsymbol{W}_e$嵌入到300维空间，$\boldsymbol{x}_j=\boldsymbol{W}_e\boldsymbol{w}_j, j\in[1,n]$，其中$\boldsymbol{W}_e$是可学习的嵌入矩阵。然后，为了使得局部文本表示具有上下文信息，我们利用双向GRU对来自文本的前向与后向两个方向的信息进行编码，分别得到第$j$个单词的前向隐藏状态与后向隐藏状态，如公式(1)、(2)所示。

$$\overrightarrow{\boldsymbol{h}_j}=\overrightarrow{\mathrm{GRU}}(\boldsymbol{x}_j), j\in[1,n] \tag{1}$$

$$\overleftarrow{\boldsymbol{h}_j}=\overleftarrow{\mathrm{GRU}}(\boldsymbol{x}_j), j\in[1,n] \tag{2}$$

为了得到第$j$个局部文本特征，我们通过对$\overrightarrow{\boldsymbol{h}_j}$与$\overleftarrow{\boldsymbol{h}_j}$求平均，如公式(3)所示。

$$\boldsymbol{e}_j=\frac{(\overrightarrow{\boldsymbol{h}_j}+\overleftarrow{\boldsymbol{h}_j})}{2}, j\in[1,n] \tag{3}$$

为了获得图像的显著性区域，我们利用Anderson等人[29]提出的Bottom-up attention对图像进行检测。然后，我们使用在Visual Genomes数据集[30]上预训练的Faster-RCNN[31]提取出前$k$个最为显著的局部图像特征，再通过线性映射将局部图像特征映射到D维空间，与局部文本特征的维度保持一致。

$$\boldsymbol{v}_i=\boldsymbol{W}_v\boldsymbol{f}_i+\boldsymbol{b}_v \tag{4}$$

式(4)中，$\boldsymbol{f}_i$表示第$i$个局部图像特征，$\boldsymbol{W}_v$和$\boldsymbol{b}_v$表示可学习的参数，$\boldsymbol{v}_i$表示D维空间的第$i$个局部图像特征。

### 3.2.2　局部图像到局部文本交叉注意力

由于局部图像到局部文本交叉注意力只是简单的对局部图像应用交叉注意力机制，学习出比较好的局部图像表示，但是没有细化每个局部图像特征对应的权重，以及未利用权重对局部图像特征做进一步的优化。针对该问题，我们对局部图像到局

部文本交叉注意力进行了改进。

我们在SCAN模型工作的基础上，为了得到每个局部图像特征关于整个文本的关注文本向量，我们将每个局部文本特征与其对应的权重进行加权求和，如公式(5)所示。

$$\boldsymbol{a}_i^t = \sum_{j=1}^{n} \boldsymbol{\alpha}_{ij} \boldsymbol{e}_j \tag{5}$$

我们通过计算每个局部图像特征与其关注文本向量的相似度来确定每个局部图像特征的重要性。

$$\boldsymbol{R}_i^v = \frac{\boldsymbol{v}_i^T \boldsymbol{a}_i^t}{\|\boldsymbol{v}_i\| \|\boldsymbol{a}_i^t\|} \tag{6}$$

式(6)中，$\boldsymbol{R}_i^v$表示第$i$个局部图像特征与整个文本的相似度。为了得到局部图像到局部文本交叉注意力模块的相似度。我们对这$k$个相似度进行Log-SumExp pooling(LSE)。

$$S_1(\boldsymbol{I}, \boldsymbol{T}) = \log\left(\sum_{i=1}^{k} \exp(\lambda_2 \boldsymbol{R}_i^v)\right)^{(1/\lambda_2)} \tag{7}$$

式(7)中，$\lambda_2$是决定如何放大最相关局部图像特征$\boldsymbol{v}_i$与关注文本向量$\boldsymbol{a}_i^t$重要性的超参数。为了细化每个局部图像特征的重要性，我们对$k$个关注文本向量进行归一化，如公式(8)所示。

$$\boldsymbol{\beta}_i^t = [\boldsymbol{a}_i^t]_+ / \sqrt{\sum_{i=1}^{k} [\boldsymbol{a}_i^t]_+^2} \tag{8}$$

### 3.2.3 局部文本到局部图像交叉注意力

由于局部文本到局部图像交叉注意力只是简单的对局部文本应用交叉注意力机制，学习出比较好的局部文本表示，但是没有细化每个局部文本特征对应的权重，以及未利用权重对局部文本特征做进一步的优化。针对该问题，我们对局部文本到局部图像交叉注意力进行了改进。

同样地，我们在SCAN模型工作的基础上，为了得到每个局部文本特征关于整个图像的关注图像向量，我们将每个局部图像特征与其对应的权重进行加权求和，如公式(9)所示。

$$\boldsymbol{a}_j^v = \sum_{i=1}^{k} \boldsymbol{\alpha}'_{ji} \boldsymbol{v}_i \tag{9}$$

我们通过计算每个局部文本特征与其关注图像向量的相似度来确定每个局部文本特征的重要性。

$$\boldsymbol{R}_j^e = \frac{\boldsymbol{e}_j^T \boldsymbol{a}_j^v}{\|\boldsymbol{e}_j\| \|\boldsymbol{a}_j^v\|} \tag{10}$$

式(10)中，$\boldsymbol{R}_j^e$表示第$j$个局部文本特征与整个图像的相似度。为了得到局部文本到局部图像交叉注意力模块的相似度，我们对这$n$个相似度进行平均池化，如公式(11)所示。

$$S_2(\boldsymbol{T}, \boldsymbol{I}) = \frac{\sum_{j=1}^{n} \boldsymbol{R}_j^e}{n} \tag{11}$$

为了细化每个局部文本特征的重要性，我们对$n$个关注图像向量进行归一化，如公式(12)所示。

$$\boldsymbol{\beta}_j^v = [\boldsymbol{a}_j^v]_+ / \sqrt{\sum_{j=1}^{n} [\boldsymbol{a}_j^v]_+^2} \tag{12}$$

### 3.2.4 全局图像与全局文本匹配

由于SCAN模型的方法只是利用堆叠交叉注意力来捕捉图像区域与文本单词之间的潜在对齐，缺乏全局图像特征与全局文本特征之间的匹配，没有挖掘到全局特征信息，如图3所示。针对该问题，我们在局部图像与局部文本堆叠交叉注意力匹配的基础上，考虑了全局图像与全局文本匹配，为了使得全局图像特征与全局文本特征能充分表达其全局语义信息，我们将堆叠交叉注意力引进全局图像与全局文本匹配，如图4所示。

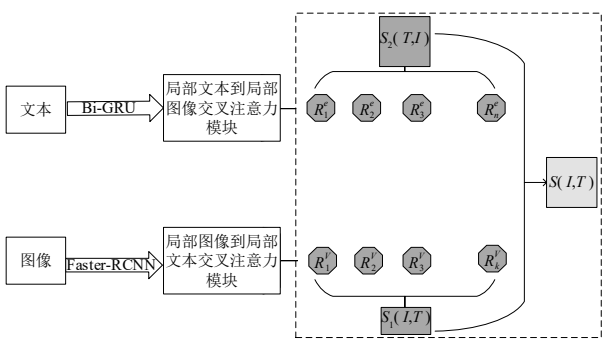


图3 相似度融合模型图

Fig. 3 Similarity fusion model map

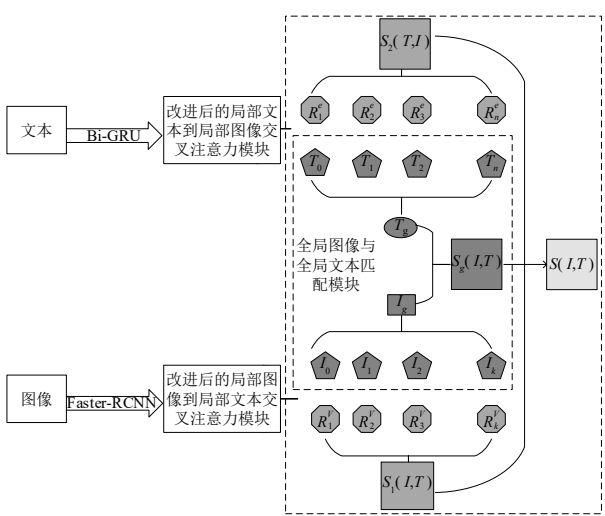


图4 改进后的相似度融合模型图

Fig. 4 Improved similarity fusion model map

这一小节，在全局图像与全局文本匹配中考虑堆叠交叉注意力，通过注意力来进一步挖掘全局特征信息，让全局图像与全局文本特征得到优化。在局部图像特征之间采用自注意力机制，利用 $\bar{\boldsymbol{I}} = 1/k \sum_{i=1}^{k} \boldsymbol{v}_i$ 作为查询和聚合所有的局部图像特征来获得全局图像特征 $\boldsymbol{I}_0$。同理，利用 $\bar{\boldsymbol{T}} = 1/n \sum_{j=1}^{n} \boldsymbol{e}_j$ 作为查询和聚合所有的局部文本特征来获得全局文本特征 $\boldsymbol{T}_0$。

我们首先根据改进后的局部图像到局部文本交叉注意力，得到每个局部图像特征对应的权重。然后为了使局部图像特征得到优化，我们将每个局部图像特征与其对应的权重相乘。

$$\boldsymbol{I}_i = \boldsymbol{v}_i \cdot \boldsymbol{\beta}_i^t \tag{13}$$

式(13)中，$\boldsymbol{\beta}_i^t$ 表示第 $i$ 个局部图像特征对应的权重。为了使全局图像特征能充分表达其全局语义信息并合理的调整与局部图像特征分布一致的全局图像特征，我们将 $\boldsymbol{I}_0$ 和优化之后的 $k$ 个局部图像特征进行拼接，然后 Log-SumExp pooling。

$$\boldsymbol{I}_g = \log\left(\sum_{i=0}^{k} \exp(\lambda_4 \boldsymbol{I}_i)\right)^{(1/\lambda_4)} \tag{14}$$

式(14)中，$\lambda_4$ 是决定如何放大最相关图像特征 $\boldsymbol{v}_i$ 与其对应权重 $\boldsymbol{\beta}_i^t$ 重要性的超参数。我们首先根据改进后的局部文本到局部图像交叉注意力，得到每个局部文本特征对应的权重。为了使局部文本特征得到优化，我们将每个局部文本特征与其对应的权重相乘。

$$\boldsymbol{T}_j = \boldsymbol{e}_j \cdot \boldsymbol{\beta}_j^v \tag{15}$$

式(15)中，$\boldsymbol{\beta}_j^v$ 表示第 $j$ 个局部文本特征对应的权重。为了使全局文本特征能充分表达其全局语义信息并合理的调整与局部文本特征分布一致的全局文本特征，我们将 $\boldsymbol{T}_0$ 和优化之后的 $n$ 个局部文本特征进行拼接，然后平均池化，如公式(16)所示。

$$\boldsymbol{T}_g = \frac{\sum_{j=0}^{n} \boldsymbol{T}_j}{n+1} \tag{16}$$

我们计算全局图像特征与全局文本特征的余弦相似度来确定全局图像与全局文本的相似性，如公式(17)所示。

$$S_g(\boldsymbol{I}, \boldsymbol{T}) = \cos ine(\boldsymbol{I}_g, \boldsymbol{T}_g) \tag{17}$$

#### 3.2.5　相似度融合

相似度融合目的是能够有效的综合局部和全局信息。为了使提取的特征信息更加完善，我们首先对局部图像到局部文本交叉注意力模块的相似度与局部文本到局部图像交叉注意力模块的相似度求平均来确定局部图像与局部文本匹配的相似性，使得局部特征信息提取的更加完善，如公式(18)所示。

$$S_l(\boldsymbol{I}, \boldsymbol{T}) = (S_1(\boldsymbol{I}, \boldsymbol{T}) + S_2(\boldsymbol{T}, \boldsymbol{I}))/2 \tag{18}$$

然后，为了全面有效的考虑局部和全局信息，我们将设计出的全局图像与全局文本匹配引进相似度融合模块，如图4所示。为了合理的融合局部图像与局部文本匹配以及全局图像与全局文本匹配的相似度。我们引入两个超参数来调整它们的比例，如公式(19)所示。

$$S(\boldsymbol{I}, \boldsymbol{T}) = uS_l(\boldsymbol{I}, \boldsymbol{T}) + vS_g(\boldsymbol{I}, \boldsymbol{T}) \tag{19}$$

#### 3.2.6　目标函数

为了让匹配的图像文本对相似度远远大于不匹配的图像文本对相似度，我们使用 Max of Hinges loss 作为目标函数来优化模型，如公式(20)所示。

$$l(\boldsymbol{I}, \boldsymbol{T}) = [\alpha - S(\boldsymbol{I}, \boldsymbol{T}) + S(\boldsymbol{I}, \boldsymbol{T}_h')]_+ + [\alpha - S(\boldsymbol{I}, \boldsymbol{T}) + S(\boldsymbol{I}_h', \boldsymbol{T})]_+ \tag{20}$$

式(20)中，$\alpha$ 是余量，$[\boldsymbol{x}]_+$ 表示 $\max(x, 0)$，$S(\boldsymbol{I}, \boldsymbol{T})$ 是匹配的图像文本对，$S(\boldsymbol{I}, \boldsymbol{T}_h')$、$S(\boldsymbol{I}_h', \boldsymbol{T})$ 是不匹配的图像文本对。其中最难图像负样本、最难文本负样本由公式(21)、(22)获得。

$$\boldsymbol{I}_h' = \arg\max_{M \neq I} S(\boldsymbol{M}, \boldsymbol{T}) \tag{21}$$

$$\boldsymbol{T}_h' = \arg\max_{D \neq T} S(\boldsymbol{I}, \boldsymbol{D}) \tag{22}$$

### 3.3　本文模型的算法表

本文模型的总体训练步骤算法表如下算法 1 所示。

## 4　实验结果与分析

为了验证本文所提出方法在图像文本跨模态检索上的有效性，我们在 Flickr30K 和 MS-COCO 数据集上进行了实验。并且结合本文模型做了3类对比实验，一类是与其他基线模型的性能对比实验，验证该模型的有效性和优越性。另一类是本文模型的消融测试实验，验证该模型的合理性。最后一类是本文模型的参数测试实验，验证不同参数对该模型的影响。

### 4.1　实验数据集

我们在 Flickr30K 和 MS-COCO 数据集上对提出的模型进行了实验验证，这两个数据集的图像是普通图像，文本是英文句子，同时之前的研究工作都是基于这两个数据集来进行实验的，因此本文实验选用这两个数据集来进行对比测试实验、消融测试实验以及参数测试实验。Flickr30K 包含从 Flickr 网站收集的31000张图像，每张图像对应5个文本。我们按照 Faghri 等人[15]以及 Lee 等人[18]的分割方式，使用29000张图像作为训练样本，1000张图像

**算法 1** 图像文本跨模态匹配的训练步骤。

**输入：** 图像 F，文本 T，训练轮数 $N$，训练的图像文本总对数 $M$，批量大小 $K$，图像的局部共享数 $H$，文本的局部共享数 $G$。

for $n$=1:$N$ do
for $m$=1:$M$ do
Faster-RCNN 和 Bi-GRU
图像文本对:(F,T)
{图像编码} F →V,式(3)
{文本编码} T →E,式(4)
for $g$=1:$K$ do
更新局部图像与局部文本特征的参数
更新全局图像与全局文本特征的参数
end for
for $i$=1: $H$ do
for $j$=1:$G$ do
计算局部图像与局部文本特征的相似度
计算全局图像与全局文本特征的相似度
end for
end for
end for
end for

**函数：** 局部匹配、全局匹配

/*Bi-GRU and Faster-RCNN*/
{文本编码} T →E, 式(3)
{图像编码} F →V, 式(4)
局部的堆叠交叉注意力匹配

$$s_{ji}=\frac{\boldsymbol{e}_j^T\boldsymbol{v}_i}{\|\boldsymbol{e}_j\|\|\boldsymbol{v}_i\|},\ j\in[1,n],\ i\in[1,k]$$

$$s_{ij}=\frac{\boldsymbol{v}_i^T\boldsymbol{e}_j}{\|\boldsymbol{v}_i\|\|\boldsymbol{e}_j\|},\ i\in[1,k],\ j\in[1,n]$$

计算全局匹配相似度 $S_g(\boldsymbol{I},\boldsymbol{T})$,式(17)
计算局部匹配相似度 $S_l(\boldsymbol{I},\boldsymbol{T})$,式(18)
计算最终图像文本匹配相似度 $S(\boldsymbol{I},\boldsymbol{T})$,式(19)
/*匹配*/
更新模型参数,式(20)。

**输出：** 图像检索文本:R@1、R@5、R@10、R@sum
文本检索图像:R@1、R@5、R@10、R@sum

作为验证样本，1000 张图像作为测试样本。MS-COCO 总共包含 123287 张图像，每张图像对应 5 个文本。我们按照 Faghri 等人[15]以及 Lee 等人[18]的分割方式，使用 113287 张图像作为训练样本，1000 张图像作为验证样本，1000 张图像作为测试样本(以及训练样本不变，5000 张图像作为验证样本，5000 张图像作为测试样本)。

### 4.2 参数设置

我们的对比测试实验、消融测试实验以及参数测试实验都是在 Inter i9-10900X CPU、2080ti GPU、64GB 内存的硬件平台以及 python3.6.4、pytorch 1.3.0 的软件环境下进行的。对于图像，我们采用 Faster-RCNN 网络来提取局部图像特征。我们的交并比阈值设置为 0.7。我们提取出每个图像经过最后池化层的前 $k$ 个局部图像特征，然后使用一个全连接层将局部图像特征转换成 1024 维向量。对于文本，我们使用双向 GRU 将文本编码成一系列局部文本特征。通过对双向 GRU 输出的 1024 维前向隐藏状态与后向隐藏状态求平均来得到局部文本特征。其他参数根据经验以及参数测试实验设置如下：$k$ 设为 36，维度 $D$ 设为 1024。在 Flickr30K 数据集上，批量大小(batch size)设为 64，训练轮数(training epochs)设为 30，优化器使用 Adam，前 15 个 epoch 的学习率设为 0.0002，后 15 个 epoch 的学习率设为 0.00002，损失函数中的余量 $\boldsymbol{\alpha}$ 设为 0.2，参与训练和测试的超参数 $\lambda_1$ 设为 4，$\lambda_2$ 设为 5，$\lambda_3$ 设为 9，$\lambda_4$ 设为 5，$u$ 设为 1，$v$ 设为 1。在 MS-COCO 数据集上的实现细节和超参数的设置与在 SCAN 模型上的一样。

### 4.3 评价指标

本文采用检索召回率的评价指标对模型的性能进行分析，主要评价指标包括 R@1(Recall@1)、R@5(Recall@5)、R@10(Recall@10)、R@sum(Recall@ sum)等，各项指标的具体描述如表 1 所示。

### 4.4 实验与分析

为了验证本文方法的有效性，本文结合本文模型分别在两个数据集上做了 3 类实验，一类是本文模型与不同基线方法的对比实验，实验结果如表 2、表 3、表 4 所示。另一类是验证本文模型全局图像与全局文本匹配模块的有效性以及全局特征未使用注意力、使用了自注意力和局部特征对应的权重未进行归一化的对比实验，实验结果如表 5、表 6、表 7 所示。最后一类是验证不同参数对本文模型的影响。实验结果如表 8、表 9 所示。

#### 4.4.1 对比测试实验分析

对比测试实验目的是通过本文模型与不同基线方法进行对比，验证本文模型的有效性和优越性。我们分别在 Flickr30K 和 MS-COCO 两个公共数据集上进行了对比测试实验，实验结果如表 2、表 3、表 4 所示。

从表 2、表 3、表 4 中可以看出，我们的模型在 Flickr30K 和 MS-COCO 两个公共数据集上，图像检索文本和文本检索图像的召回率相对于 Lee 等人[18]

提出的baseline(SCAN)模型有明显的提升。例如，在Flickr30K数据集上，文本检索图像的R@1是50.2，相对于SCAN模型的召回率提高了1.6%，在MS-COCO数据集(1K测试图像)上，图像检索文本的R@1是74.4，相对于SCAN模型的召回率提高了1.7%，在MS-COCO数据集(5K测试图像)上，图像检索文本的R@1是51.0，相对于SCAN模型的召回率提高了0.6%。在与其他模型对比，我们的模型在Flickr30K和MS-COCO两个公共数据集上，图像检索文本和文本检索图像的总体性能在R@sum上达到了最好。该方法之所以能取得如此效果，是根据以下原因：基于堆叠交叉注意力的全局图像与全局文本匹配，发挥了一定的作用，因为全局图像特征与全局文本特征是经过堆叠交叉注意力优化的，优化之后的全局图像特征与全局文本特征更能表达其全局图像与全局文本的全局语义信息，有利于实现图像与文本更准确的对齐。

本文模型在Flickr30K和MS-COCO两个公共数据集上，图像检索文本和文本检索图像的总体性能R@sum都优于Zheng等人[19]、Wang等人[20]、Ji等人[21]以及Li等人[23]提出的模型。虽然Zheng等人[19]提出的自注意力视觉语义嵌入方法以及Wang等人[20]提出的跨模态自适应信息传递模型考虑了细粒度信息的对齐，但是它们都缺乏全局图像特征与全局文本特征之间的匹配，没有挖掘到全局特征信息，就捕捉不到全局语义信息的对齐，最终影响图像文本跨模态检索的效果。Ji等人[21]提出的模型利用结构简单的卷积神经网络直接对图像的特征进行提取过于草率，导致提取的图像特征包含过多的干扰信息，无法准确的实现图像信息与文本信息的对齐。Li等人[23]提出的融合两级相似性的跨媒体图像文本检索方法，由于只是简单的利用自注意力机制对局部特征以及全局特征进行优化，无法准确的凸显局部特征和全局特征的重要程度，所

表1　评价指标

Tab. 1　Evaluation index

| 指标 | 描述 |
|---|---|
| R@1 | 图像检索文本或者文本检索图像中正确结果出现在前1排名的查询的比例 |
| R@5 | 图像检索文本或者文本检索图像中正确结果出现在前5排名的查询的比例 |
| R@10 | 图像检索文本或者文本检索图像中正确结果出现在前10排名的查询的比例 |
| R@sum | 图像检索文本和文本检索图像的总体性能(该行所有召回率之和) |

表2　模型在Flickr30K实验结果下的召回率和R@sum(%)

Tab. 2　Model recall and R@sum under Flickr30K experimental results(%)

| 方法 | 图像检索文本 | | | 文本检索图像 | | | R@sum |
|---|---|---|---|---|---|---|---|
| 评价指标 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| Karpathy等人[14](2017) | 22.2 | 48.2 | 61.4 | 15.2 | 37.7 | 50.5 | 235.2 |
| Niu等人[3](2017) | 38.1 | _ | 76.5 | 27.7 | _ | 68.8 | 211.1 |
| Huang等人[4](2017) | 42.5 | 71.9 | 81.5 | 30.2 | 60.4 | 72.3 | 358.8 |
| Eisenschtat等人[5](2017) | 49.8 | 67.5 | _ | 36.0 | 55.6 | _ | 208.9 |
| Faghri等人[15](2017) | 52.9 | 80.5 | 87.2 | 39.6 | 70.1 | 79.5 | 409.8 |
| Nam等人[6](2017) | 55.0 | 81.8 | 89.0 | 39.4 | 69.2 | 79.1 | 413.5 |
| Zheng等人[16](2020) | 55.6 | 81.9 | 89.5 | 39.1 | 69.2 | 80.9 | 416.2 |
| Huang等人[17](2018) | 55.5 | 82.0 | 89.3 | 41.1 | 70.5 | 80.1 | 418.5 |
| Lee等人[18](2018)(baseline) | 67.4 | 90.3 | 95.8 | 48.6 | 77.7 | 85.2 | 465.0 |
| Zheng等人[19](2019) | 67.2 | 88.3 | 94.2 | 49.8 | **78.7** | **86.2** | 464.4 |
| Wang等人[20](2019) | 68.1 | 89.7 | 95.2 | **51.5** | 77.1 | 85.3 | 466.9 |
| Ji等人[21](2020) | 58.1 | 82.8 | 90.1 | 44.7 | 72.4 | 81.1 | 429.2 |
| Li等人[23](2020) | 68.2 | 89.1 | 94.5 | 43.4 | 73.5 | 82.5 | 451.2 |
| 本文 | **68.3** | **91.2** | **96.0** | 50.2 | 77.9 | 85.3 | **468.9** |

表3 模型在MS-COCO(1K测试图像)实验结果下的召回率和R@sum(%)

Tab. 3 Model recall and R@sum under MS-COCO (1K Test Images) experimental results(%)

| 方法 | 图像检索文本 | | | 文本检索图像 | | | R@sum |
|---|---|---|---|---|---|---|---|
| 评价指标 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| Karpathy等人[14](2017) | 38.4 | 69.9 | 80.5 | 27.4 | 60.2 | 74.8 | 351.2 |
| Niu等人[3](2017) | 43.9 | _ | 87.8 | 36.1 | _ | 86.7 | 254.5 |
| Vendrov等人[13](2016) | 46.7 | _ | 88.9 | 37.9 | _ | 85.9 | 259.4 |
| Huang等人[4](2017) | 53.2 | 83.1 | 91.5 | 40.7 | 75.8 | 87.4 | 431.7 |
| Eisenschtat等人[5](2017) | 55.8 | 75.2 | _ | 39.7 | 63.3 | _ | 234.0 |
| Faghri等人[15](2017) | 64.6 | 90.0 | 95.7 | 52.0 | 84.3 | 92.0 | 478.6 |
| Zheng等人[16](2020) | 65.6 | 89.8 | 95.5 | 47.1 | 79.9 | 90.0 | 467.9 |
| Gu等人[22](2018) | 68.5 | _ | 97.9 | 56.6 | _ | 94.5 | 317.5 |
| Huang等人[17](2018) | 69.9 | 92.9 | 97.5 | 56.7 | 87.5 | 94.8 | 499.3 |
| Lee等人[18](2018)(baseline) | 72.7 | 94.8 | 98.4 | 58.8 | 88.4 | 94.8 | 507.9 |
| Zheng等人[19](2019) | 70.8 | 93.2 | 97.6 | 56.9 | 87.6 | 94.4 | 500.5 |
| Wang等人[20](2019) | 72.3 | 94.8 | 98.3 | 58.5 | 87.9 | **95.0** | 506.8 |
| Ji等人[21](2020) | 70.4 | 91.7 | 96.8 | 58.4 | 87.1 | 94.0 | 498.4 |
| Li等人[23](2020) | 68.9 | 94.1 | 98.0 | 58.6 | 88.2 | 94.9 | 502.7 |
| 本文 | **74.4** | **95.8** | **98.5** | **59.4** | **88.5** | **95.0** | **511.6** |

表4 模型在MS-COCO(5K测试图像)实验结果下的召回率和R@sum(%)

Tab. 4 Model recall and R@sum under MS-COCO (5K Test Images) experimental results(%)

| 方法 | 图像检索文本 | | | 文本检索图像 | | | R@sum |
|---|---|---|---|---|---|---|---|
| 评价指标 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| Vendrov等人[13](2016) | 23.3 | _ | 84.7 | 31.7 | _ | 74.6 | 214.3 |
| Faghri等人[15](2017) | 41.3 | _ | 81.2 | 30.3 | _ | 72.4 | 225.2 |
| Zheng等人[16](2020) | 41.2 | 70.5 | 81.1 | 25.3 | 53.4 | 66.4 | 337.9 |
| Gu等人[22](2018) | 42.0 | _ | 84.7 | 31.7 | _ | 74.6 | 233.0 |
| Huang等人[17](2018) | 42.8 | 72.3 | 83.0 | 33.1 | 62.9 | 75.5 | 369.6 |
| Lee等人[18](2018)(baseline) | 50.4 | 82.2 | 90.0 | 38.6 | 69.3 | 80.4 | 410.9 |
| Zheng等人[19](2019) | 46.7 | 76.3 | 86.1 | 34.0 | 64.8 | 77.0 | 384.9 |
| Wang等人[20](2019) | 50.1 | 82.1 | 89.7 | 39.0 | 68.9 | 80.2 | 410.0 |
| Ji等人[21](2020) | 48.9 | 75.2 | 84.4 | 38.3 | 65.7 | 76.9 | 389.4 |
| 本文 | **51.0** | **82.5** | **90.2** | **39.2** | **69.4** | **80.4** | **412.7** |

以局部图像与局部文本匹配以及全局图像与全局文本匹配无法很好的兼容进行共同学习,从而导致图像文本跨模态检索的召回率不高。而本文模型是基于baseline模型利用Faster-RCNN网络对图像的特征进行提取,能够提取出前$k$个最为显著的图像信息作为局部图像特征;同时将堆叠交叉注意力引进全局图像与全局文本匹配,使得局部图像与局部文本匹配以及全局图像与全局文本匹配能够很好的兼容进行共同学习,从而提高图像文本跨模态检索的召回率。

另外,本文模型在Flickr30K和MS-COCO两个公共数据集上,图像检索文本以及文本检索图像的部分召回率指标与目前最前沿的模型相比略微有点低。经过分析得出,本文模型的局部图像与局部

文本特征和全局图像与全局文本特征的优化程度仍然不足，所以局部图像与局部文本匹配以及全局图像与全局文本匹配无法更好的兼容进行共同学习，导致图像检索文本以及文本检索图像的部分召回率指标略微有点下降。

最后，在模型的时间复杂度方面对SCAN模型的复杂度与本文模型的复杂度进行分析，$N$表示算法的第一层循环语句的执行次数，$M$表示嵌套在第一层循环语句内的第二层循环语句执行次数，$K$表示嵌套在第二层循环语句内的第三层循环语句执行次数，$H$表示嵌套在第二层循环语句内的另一个第三层循环语句执行次数，$G$表示嵌套在第三层循环语句内的第四层循环语句执行次数，SCAN模型的算法代码渐进时间复杂度为$O(1)+O(N)+O(NM+NM)+O(NMK+NMK+NMK)+O(NMHG+NMHG+NMHG+NMHG)+O(NHG+NHG+NHG+NHG)=O(1)+O(N)+O(2NM)+O(3NMK)+O(4NMHG)+O(4NHG)=O(4NMHG)\approx \boldsymbol{O}(\boldsymbol{N}^4)$，本文模型的算法代码渐进时间复杂度为$O(1)+O(N)+O(NM+NM)+O(NMK+NMK+NMK)+O(NMHG+NMHG+NMHG+NMHG)+O(NHG+NHG+NHG+NHG)=O(1)+O(N)+O(2NM)+O(3NMK)+O(4NMHG)+O(4NHG)=O(4NMHG)\approx \boldsymbol{O}(\boldsymbol{N}^4)$，所以本文模型的时间复杂度与SCAN模型的时间复杂度相当。

### 4.4.2　消融测试实验分析

消融测试实验目的是通过删除或者替换本文模型的某些模块来进行对比，验证本文模型的合理性。(image-text)表示局部图像到局部文本交叉注意力模块。(image-text +global-global)表示局部图像到局部文本交叉注意力模块加上全局图像与全局文本匹配模块。(text-image)表示局部文本到局部图像交叉注意力模块。(text-image+global-global)表示局部文本到局部图像交叉注意力模块加上全局图像与全局文本匹配模块。no-attention表示本文模型的全局特征未使用注意力进行优化，直接对$k$个局部图像特征和$n$个局部文本特征分别进行池化作用来得到对应的全局特征。self-attention表示本文模型的全局特征使用了自注意力进行优化，$k$个局部图像特征和$n$个局部文本特征分别通过自注意力，达到凸显局部特征对应的重要程度，然后再分别进行池化作用得到对应的全局特征。no-normalize表示本文模型的局部特征对应的权重未利用归一化进行优化，其$k$个局部图像特征和$n$个局部文本特征对应的关注文本向量和关注图像向量没有进行归一化，关注文本向量以及关注图像向量作为权重和其对应的局部图像特征以及局部文本特征直接相乘，来优化局部图像特征和局部文本特征，然后再分别进行池化作用得到对应的全局特征。本文模型的全局特征是使用了堆叠交叉注意力来进行优化的。实验结果如表5、表6、表7所示。

表5　模型在Flickr30K上消融测试实验结果(%)

Tab. 5　Model ablation test results on Flickr30K(%)

| 方法 | 图像检索文本 | | | 文本检索图像 | | | R@sum |
|---|---|---|---|---|---|---|---|
| 评价指标 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| (image-text)(baseline) | 67.9 | 89.0 | 94.4 | 43.9 | 74.2 | 82.8 | 452.2 |
| (image-text+global-global) | 67.2 | 90.7 | 95.3 | 48.5 | 76.6 | 84.7 | 463.0 |
| (text-image)(baseline) | 61.8 | 87.5 | 93.7 | 45.8 | 74.4 | 83.0 | 446.2 |
| (text-image+global-global) | 66.9 | 89.9 | 94.7 | 49.8 | 77.0 | 84.4 | 462.7 |
| no-attention | 60.4 | 86.6 | 92.5 | 43.2 | 72.9 | 81.6 | 437.2 |
| self-attention | 63.4 | 87.3 | 93.3 | 44.5 | 73.8 | 82.6 | 444.9 |
| no-normalize | 66.5 | 89.6 | 94.8 | 48.6 | 75.5 | 84.0 | 459.0 |
| 本文 | **68.3** | **91.2** | **96.0** | **50.2** | **77.9** | **85.3** | **468.9** |

从表5、表6、表7中可以看出，在Flickr30K数据集上，(image-text+global-global)的总体性能R@sum相对于(image-text)baseline模型的总体性能R@sum提高了10.8%。(text-image+global-global)的总体性能R@sum相对于(text-image)baseline模型的总体性能R@sum提高了16.5%。在MS-COCO数据集(1K测试图像)上，(image-text+global-global)的总体性能R@sum相对于(image-text)baseline模型的总体性能R@sum提高了11.7%。(text-image+global-global)的总体性能R@sum相对于(text-image)baseline模型的总体性能R@sum提高了7.9%。在MS-COCO数据集(5K测试图像)上，(image-text+global-global)的总

表6 模型在MS-COCO(1K测试图像)上消融测试实验结果(%)

Tab. 6 Model ablation test results on MS-COCO (1K Test Images)(%)

| 方法 | 图像检索文本 | | | 文本检索图像 | | | R@sum |
|---|---|---|---|---|---|---|---|
| 评价指标 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| (image-text)(baseline) | 69.2 | 93.2 | 97.5 | 54.4 | 86.0 | 93.6 | 493.9 |
| (image-text+global-global) | 73.9 | 95.2 | 97.9 | 56.9 | 87.5 | 94.2 | 505.6 |
| (text-image)(baseline) | 70.9 | 94.5 | 97.8 | 56.4 | 87.0 | 93.9 | 500.5 |
| (text-image+global-global) | 73.2 | **95.8** | 98.1 | 59.0 | 87.7 | 94.6 | 508.4 |
| no-attention | 66.3 | 93.1 | 97.4 | 52.0 | 84.5 | 92.3 | 485.6 |
| self-attention | 68.1 | 93.5 | 97.6 | 53.1 | 85.5 | 93.0 | 490.8 |
| no-normalize | 72.0 | 94.6 | 97.9 | 57.1 | 86.9 | 93.7 | 502.2 |
| 本文 | **74.4** | **95.8** | **98.5** | **59.4** | **88.5** | **95.0** | **511.6** |

表7 模型在MS-COCO(5K测试图像)上消融测试实验结果(%)

Tab. 7 Model ablation test results on MS-COCO(5K Test Images)(%)

| 方法 | 图像检索文本 | | | 文本检索图像 | | | R@sum |
|---|---|---|---|---|---|---|---|
| 评价指标 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| (image-text)(baseline) | 46.4 | 77.4 | 87.2 | 34.4 | 63.7 | 75.7 | 384.8 |
| (image-text+global-global) | 48.5 | 79.3 | 88.6 | 36.3 | 65.2 | 77.1 | 395.0 |
| no-attention | 40.8 | 71.9 | 83.5 | 30.5 | 60.1 | 72.4 | 359.2 |
| self-attention | 42.2 | 73.7 | 84.2 | 31.3 | 61.2 | 73.3 | 365.9 |
| no-normalize | 47.0 | 77.8 | 87.4 | 35.9 | 64.6 | 76.8 | 389.5 |
| 本文 | **51.0** | **82.5** | **90.2** | **39.2** | **69.4** | **80.4** | **412.7** |

体性能R@sum相对于(image-text)baseline模型的总体性能R@sum提高了10.2%。说明在局部图像与局部文本堆叠交叉注意力匹配的基础上,考虑全局图像与全局文本匹配使得模型提取的特征信息更加完善。证明了全局图像与全局文本匹配模块的有效性。

由表5、表6、表7可知,本文模型图像检索文本和文本检索图像的召回率以及它们的R@sum均优于表5、表6、表7中其他消融模型。当本文模型的全局特征未使用注意力优化时,全局图像与全局文本特征无法充分表达其全局语义信息,所以召回率和R@sum比较低。由此可以看出模型的全局特征使用注意力进行优化的重要性。当本文模型的全局特征使用自注意力进行优化时,召回率和R@sum相对于no-attention均有所提升,但由于自注意力无法准确的确定局部特征重要性的权重分布,未能充分优化全局特征,所以召回率和R@sum相对于本文模型略微有点低。当本文模型的局部特征对应的权重未进行归一化时,可能导致部分局部特征对应的权重偏大或者偏小,不利于特征的优化,所以召回率和R@sum相对于本文模型略微有点低。由此可以看出模型局部特征对应的权重进行归一化的重要性。而本文模型的全局特征是使用堆叠交叉注意力以及局部特征对应的权重进行归一化来共同优化的,所以召回率和R@sum提升的比较明显。因此,证明将堆叠交叉注意力引进全局图像与全局文本匹配进而优化全局图像与全局文本特征是合理的。

#### 4.4.3 参数分析

公式(19)中,$u$和$v$是调整局部图像文本匹配与全局图像文本匹配的比例对最终图像文本匹配相似度分数影响的两个超参数。为了验证这两个超参数对本文模型的影响,我们在Flickr30K和MS-COCO两个数据集上进行了参数测试实验。实验结果显示在表8、表9中。

从表8、表9中可以看出,当$u$=1.0,$v$=1.0时,本文模型的召回率和R@sum达到了最好。相对于$u$=1.0,$v$=1.0,当$u$逐渐下降且$v$逐渐上升时,虽然模型文本检索图像的部分召回率有所上升,但是图像检索文本的召回率和R@sum出现小幅度的下降。说明局部图像与局部文本匹配的比例低于全局图像与全局文本匹配的比例,会对本文模型产生负面影响。当$u$逐渐上升且$v$逐渐下降时,模型图像检索文

表8　模型在Flickr30K上参数测试实验结果(%)

Tab. 8　Model parameter test results on Flickr30K(%)

| 不同参数 | 图像检索文本 | | | 文本检索图像 | | | R@sum |
|---|---|---|---|---|---|---|---|
| 评价指标 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| $u$=0.4 $v$=1.6 | 67.5 | 91.0 | 95.6 | 49.8 | 77.5 | **85.3** | 466.7 |
| $u$=0.6 $v$=1.4 | 67.5 | 90.9 | 95.7 | 50.5 | 77.2 | 85.1 | 466.9 |
| $u$=0.8 $v$=1.2 | 66.0 | 90.7 | 95.9 | **50.8** | 77.8 | **85.3** | 466.5 |
| $u$=1.0 $v$=1.0 | **68.3** | **91.2** | 96.0 | 50.2 | **77.9** | **85.3** | **468.9** |
| $u$=1.2 $v$=0.8 | 67.2 | 90.8 | 95.8 | 49.2 | 77.5 | 84.9 | 465.4 |
| $u$=1.4 $v$=0.6 | 67.2 | 91.1 | **96.3** | 49.0 | 75.6 | 84.3 | 463.5 |
| $u$=1.6 $v$=0.4 | 67.6 | 90.2 | 95.2 | 48.7 | 76.1 | 84.5 | 462.3 |

表9　模型在MS-COCO(1K测试图像)上参数测试实验结果(%)

Tab. 9　Model parameter test results on MS-COCO (1K Test Images)(%)

| 不同参数 | 图像检索文本 | | | 文本检索图像 | | | R@sum |
|---|---|---|---|---|---|---|---|
| 评价指标 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| $u$=0.4 $v$=1.6 | 73.8 | 95.1 | 98.1 | **59.7** | 88.3 | 94.5 | 509.5 |
| $u$=0.6 $v$=1.4 | 73.4 | 95.7 | 98.1 | 59.2 | **88.6** | 94.6 | 509.6 |
| $u$=0.8 $v$=1.2 | 73.4 | 95.6 | 98.0 | 59.0 | 88.2 | 94.4 | 508.6 |
| $u$=1.0 $v$=1.0 | **74.4** | 95.8 | **98.5** | 59.4 | 88.5 | **95.0** | **511.6** |
| $u$=1.2 $v$=0.8 | 72.7 | 95.4 | **98.5** | 58.2 | 87.7 | 94.1 | 506.6 |
| $u$=1.4 $v$=0.6 | 72.3 | 94.6 | 98.1 | 57.8 | 87.8 | 94.0 | 504.6 |
| $u$=1.6 $v$=0.4 | 71.1 | **96.1** | 98.2 | 57.2 | 86.9 | 93.9 | 503.4 |

本与文本检索图像的召回率以及R@sum都出现小幅度的下降。说明局部图像与局部文本匹配的比例高于全局图像与全局文本匹配的比例，也会对本文模型产生负面影响。所以本文模型选取$u$=1.0，$v$=1.0作为调整局部图像文本匹配与全局图像文本匹配比例的超参数，有效的综合了局部图像与局部文本匹配以及全局图像与全局文本匹配的优势。

### 4.5　可视化

为了从更直观的角度说明本文模型相对于SCAN模型的优越性，图5与图6分别展示了SCAN模型与本文模型在Flickr30K数据集上的跨模态特征可视化。在图中，红色圆圈代表图像特征，黑色圆圈代表文本特征。算法的检索性能越好，图像文本对的特征距离就越近，分布的就越稠密。算法的检索性能越差，图像文本对的特征距离就越远，分布的就越稀疏。

可以看出图5中映射到同一空间的图像特征与文本特征距离较远，分布的较稀疏，而相比于图5，图6中映射到同一空间的图像特征与文本特征距离较近，分布的较稠密，说明本文模型能够学习出更好的映射空间，拉近图像文本对之间

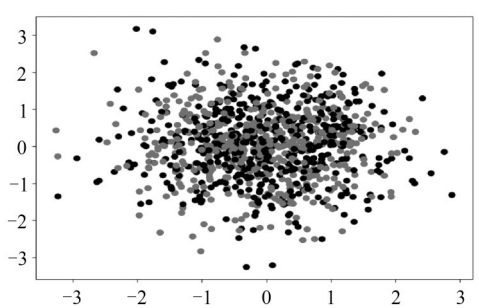


图5　SCAN模型的可视化

Fig. 5　Visualization of SCAN model

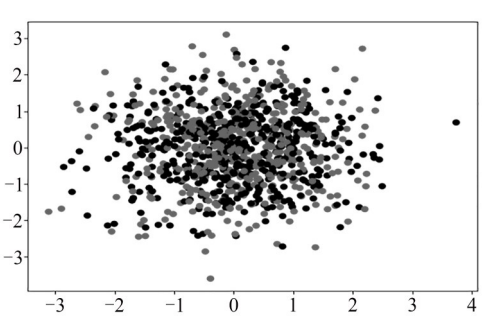


图6　本文模型的可视化

Fig. 6　Visualization of our model

的距离,从而更好的学习图像特征与文本特征的相关性,进一步说明了本文模型相对于SCAN模型的优越性。

## 5 结论

本文针对图像文本跨模态的匹配问题展开研究,提出一种基于堆叠交叉注意力的图像文本跨模态匹配方法。综合了局部图像与局部文本匹配以及全局图像与全局文本匹配的优势,有效解决了提取的特征信息不完善问题;同时将堆叠交叉注意力引进全局图像与全局文本匹配,通过堆叠交叉注意力来进一步挖掘全局特征信息,让全局图像与全局文本特征得到优化,有效解决了全局特征无法充分表达其全局语义信息的问题。实验结果表明,本文与SCAN模型相比,在Flickr30K数据集上,模型的总体性能R@sum提高了3.9%。在MS-COCO数据集上,模型的总体性能R@sum提高了3.7%。

**参考文献**


[1] 董永亮, 柴旭清. 基于潜在语义的双层图像-文本多模态检索语义网络[J]. 计算机工程, 2016,42(7):299-303,309.
DONG Yongliang, CHAI Xuqing. Two-layer image-text semantic network for multi-modal retrieval based on latent semantic[J].Computer Engineering, 2016, 42(7): 299-303,309. (in Chinese)

[2] SONG Y, SOLEYMANI M, et al. Polysemous visual semantic embedding for cross-modal retrieval [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA. IEEE, 2019: 1979-1988.

[3] NIU Zhenxing, ZHOU Mo, WANG Le, et al. Hierarchical multimodal lstm for dense visual-semantic embedding[C]// 2017 IEEE International Conference on Computer Vision. Venice, ITALY. IEEE, 2017: 1899-1907.

[4] HUANG Yan, WANG Wei, WANG Liang. Instance-aware image and sentence matching with selective multimodal lstm[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu HI. IEEE, 2017: 7254-7262.

[5] EISENSCHTAT A, WOLF L. Linking image and text with 2-way nets[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI. IEEE, 2017: 1855-1865.

[6] NAM H, HA J W, KIM J H. Dual attention networks for multimodal reasoning and matching[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI. IEEE, 2017: 2156-2164.

[7] BERRY M W, BROWNE M. Understanding search engines: mathematical modeling and text retrieval[M]. Society for Industrial and Applied mathematics, 2005.

[8] LUAN Y, EISENSTEIN J, TOUTANOVA K, et al. Sparse, dense, and attentional representations for text retrieval[J]. ArXiv Preprint ArXiv: 2005.00181, 2020.

[9] 赵师亮,吴晓富,张索非. 基于PCB特征加权的行人重识别算法[J].信号处理,2020,36(8):1300-1307.
ZHAO Shiliang, WU Xiaofu, ZHANG Suofei. Pedestrian re-recognition algorithm based on PCB feature weighting [J]. Journal of Signal Processing, 2020, 36(8): 1300-1307. (in Chinese)

[10] SUH Y, WANG J, TANG S, et al. Part-aligned bilinear representations for person re-identification[C] //15th European Conference on Computer Vision. Munich, GERMANY, 2018, 11218: 418-437.

[11] LU D, LIN Q, CHEN B , et al. A single-modal linear ultrasonic motor based on multi vibration modes of PZT ceramics[J]. Ultrasonics, 2020, 107(9):106158.

[12] 张天,靳聪,帖云,等.面向跨模态检索的音频数据库内容匹配方法研究[J].信号处理,2020,36(6):966-976.
ZHANG Tian, JIN Cong, TIE Yun, et al. Research on content matching method of audio database for cross-modal retrieval[J]. Journal of Signal Processing, 2020, 36(6): 966 -976. (in Chinese)

[13] VENDROV I, KIROS R, FIDLER S, et al. Order-embeddings of images and language [C]//4th International Conference on Learning Representations, 2016.

[14] KARPATHY A, LI F F. Deep visual-semantic alignments for generating image descriptions [J]//2017 IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 664-676.

[15] FAGHRI F, FLEET D J, KIROS J R, et al. VSE++: Improving visual-semantic embeddings with hard negatives [EB/OL]. https://arxiv.org/pdf/1707.05612.pdf.[2022-2-15].

[16] ZHENG Zhedong, ZHENG Liang, GARRETT M, et al. Dual-path convolutional image-text embedding with instance loss[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2020.

[17] HUANG Yan, WU Qi, WANG Liang. Learning semantic concepts and order for image and sentence matching[C]// 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT. IEEE, 2018: 6163-

6171.

[18] LEE K H, CHEN X, HUA G, et al. Stacked cross attention for image-text matching[J]//15th European Conference on Computer Vision. Munich, GERMANY. 2018, 11208: 212-228.

[19] ZHENG Zhuobin, BEN Youcheng, YUAN Chun. Multi-scale visual semantics aggregation with self-attention for end-to-end image-text matching [C]// Asian Conference on Machine Learning, 2019: 940-955.

[20] WANG Zihao, LIU Xihui, LI Hongsheng, et al. CAMP: Cross-modal adaptive message passing for text-image retrieval [C]// 2019 IEEE/CVF International Conference on Computer Vision. Seoul, SOUTH KOREA. IEEE, 2019: 5763-5772.

[21] JI Zhong, LIN Zhigang, WANG Haoran, et al. Multi-modal memory enhancement attention network for image-text matching[J]. IEEE Access, 2020, 8: 38438-38447.

[22] GU J, CAI J, JOTY S, et al. Look, imagine and match: improving textual-visual cross-modal retrieval with generative models[C]//2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT. IEEE, 2018: 7181-7189.

[23] LI Zhixin, LING Feng, ZHANG Canlong, et al. Combining global and local similarity for cross-media retrieval[C]//2020 IEEE International Conference on Data Science and Advanced Analytics. IEEE Access, 2020, 8: 21847-21856.

[24] 马子轩. 基于注意力机制及对抗学习的图像—文本检索研究[D]. 杭州：浙江大学，2020.
MA Zixuan. Research on image-text retrieval based on attention mechanism and adversarial learning [D]. Hangzhou: Zhejiang University, 2020. (in Chinese)

[25] TANG Jinhui, LI Zechao, WANG Meng, et al. Neighborhood discriminant hashing for large-scale image retrieval [J]. IEEE Transactions on Image Processing, 2015, 24(9): 2827-2840.

[26] LI Zechao, TANG Jinhui. Weakly supervised deep metric learning for community-contributed image retrieval[J]. IEEE Transactions on Multimedia, 2015, 17(11): 1989-1999.

[27] JIN Lu, LI Zechao, TANG Jinhui. Deep semantic multimodal hashing network for scalable image-text and video-text retrievals[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020.

[28] LI Zechao, TANG Jinhui, ZHANG Liyan, et al. Weakly-supervised semantic guided hashing for social image retrieval [J]. International Journal of Computer Vision, 2020, 128(8-9): 2265-2278.

[29] ANDERSON P, HE X, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//2018 IEEE/CVE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT. IEEE, 2018: 6077-6086.

[30] KRISHNA R, ZHU Y, GROTH O, et al. Visual genome: connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision, 2017, 123(1): 32-73.

[31] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: towards real-time object detection with region proposal networks[J]//2017 IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017, 39(6): 1137-1149.

## 作者简介

**王红斌**　男，1983年生，云南曲靖人。昆明理工大学信息工程与自动化学院硕士生导师，副教授，主要研究方向为智能信息系统、自然语言处理、数据分析。
E-mail：whbin2007@126.com

**张志亮**　男，1999年生，湖南衡阳人。昆明理工大学信息工程与自动化学院硕士研究生，主要研究方向为计算机视觉、自然语言处理、跨模态图像文本检索。
E-mail：1915873420@qq.com

**李华锋(通讯作者)**　男，1984年生，安徽阜阳人。昆明理工大学信息工程与自动化学院硕士生导师，教授，主要研究方向为计算机视觉、机器学习、模式识别。
E-mail：lhfchina99@kust.edu.cn