融合图像与文本特征的组合检索方法

秦钰淑[1],杨良怀[1*],朱艳超[2],龚卫华[1]
(1. 浙江工业大学计算机学院,浙江杭州 310023;2. 中国电子口岸数据中心杭州分中心,浙江杭州 310008)

摘 要: 随着电商领域图像数据的爆炸式增长,针对目标图像的检索成为信息检索研究中的挑战性工作. 现有的传统图像检索模型仅依靠单一文本描述或相似图像,难以准确捕捉用户的检索意图,导致检索结果不理想. 为了解决该难题,本文提出了一种融合图像与文本特征的组合检索方法,采用Swin Transformer(SwinT)提取参考图像的多层特征,将图像与文本特征在多个层级上进行融合,使文本特征能够多层次、细粒度地修改参考图像特征,以更接近目标图像特征. 然后,将修改后的图像特征与目标图像特征嵌入到一个空间中进行相似性度量,并采用基于批次的分类损失来优化检索性能. 在Fashion200k、MIT-States和CSS这3个数据集上的实验结果表明,相较于现有主流方法,本文方法在性能上平均提升了5个百分点.

关键词: 图像文本组合检索;图像特征;文本特征;特征融合
基金项目: 浙江省重点研发计划"领雁"项目(No.2022C01088)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2025)02-0558-10
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240679

A Combined Retrieval Method by Fusing Image and Text Features

QIN Yu-shu[1], YANG Liang-huai[1*], ZHU Yan-chao[2], GONG Wei-hua[1]
(1. *Computer Science and Technology*, *Zhejiang University of Technology*, *Hangzhou*, *Zhejiang* 310023, *China*;
2. *China Electronic Port Data Center*, *Hangzhou Branch*, *Hangzhou*, *Zhejiang* 310008, *China*)

Abstract: With the explosive growth of image data in the field of e-commerce, target image retrieval has become a challenging work in information retrieval research. The existing traditional image retrieval models only rely on a single text description or similar image, which is difficult to accurately capture the user's retrieval intention, resulting in unsatisfactory retrieval results. In order to solve this problem, this paper proposes a combined retrieval method that fuses image and text features. Swin Transformer (SwinT) is used to extract the multi-layer features of the reference image, and the image and text features are fused at multiple levels, so that the text features can modify the reference image features at multi-level and fine-grained, and get closer to the target image features. Then, the modified image features and the target image features are embedded in a space for similarity measurement, and the batch-based classification loss is used to optimize the retrieval performance. Experimental results on Fashion200k, MIT-States and CSS datasets show that the proposed method improves the performance by 5 percentage points on average compared with the existing mainstream methods.

Key words: combined image and text retrieval; image features; text features; features fusion
Foundation Item(s): Zhejiang Provincial Key Research and Development Program ("Ling Yan" Project) (No.2022C01088)

1 引言

随着互联网技术的发展和数字设备的普及,图像检索[1,2]在搜索引擎和电子商务等领域得到了迅速发展和应用. 然而,传统图像检索面临的挑战是难以仅凭文本描述或相似图像来准确捕捉用户的检索意图. 这主要源于人们在文本描述时的表达能力与习惯差异,这些差异往往导致信息传递中的信息遗漏和语义混淆. 此外,随着图像数据的爆炸式增长,相似图片的数量激增,这进一步加大了检索结果与用户真实需求之间的偏差,使得检索结果中充斥着众多相似却并非用户所需的图像. 为了解决上述问题,现有的主流解决方案是用户先利用已找到的图像作为初始参考,然后结合文本来进一步细化对参考图像所期望的修改. 通过

收稿日期:2024-07-22;修回日期:2024-11-01;责任编辑:孙瑶
*通讯作者:杨良怀

这种方式,可以更准确地定位到用户真正满意的检索结果,从而提高图像检索的准确性和效率. 如图1所示,假设用户以已有的时装图像为参照,并通过文本来表达差异,旨在检索到与之相关的图像. 因此,如何有效地将图像与文本模态数据进行组合,以实现精准的图像检索,是目前国内外研究者广泛关注的研究热点.

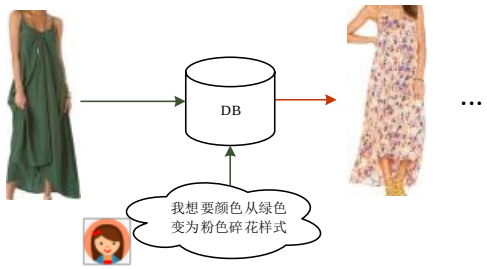


图1　图像文本组合检索示例图

现有的多模态检索研究大致分为2类:跨模态检索[3~5]和多模态组合检索[6,7]. 跨模态检索旨在以一种模态的数据去检索另一种模态的相关数据,其核心任务是数据特征提取和不同模态数据之间的相关性度量. 多模态组合检索指以多种模态组合的数据去检索一种或多种模态的相关数据,除了包含跨模态检索的核心任务,还增加了对不同模态数据的融合. 基于此,本文将采用基于图像和文本的多模态组合检索方式.

现有的图像文本组合检索方法按照数据融合方式又分为:全局融合方法[8~10]、局部融合方法[11,12]以及全局-局部融合方法[13~17]等. 全局融合方法侧重使用参考图像的全局特征,局部融合方法侧重使用参考图像的局部特征,全局-局部方法组合侧重将参考图像的全局和局部特征相结合. 这些研究结果表明,全局-局部融合的方式在整体上要优于其他组合方式. 受此启发,本文将采用全局-局部融合的方法,提取多层图像特征与文本特征进行融合. 基本思想是:参考图像经过SwinT[18](Swin Transformer)提取得到4个阶段的多层图像特征,修改文本通过LSTM(Long Short Term Memory)[19]提取得到文本特征,将多层图像特征与文本特征进行融合,与目标图像特征共同嵌入同一向量空间进行相似性度量训练.

本文的主要贡献归纳如下:

(1)提出了一种融合图像与文本特征的组合检索方法,该方法充分利用了图像和文本2种模态信息,以克服单一模态检索的局限性,从而提供了更加全面的综合结果.

(2)不同于以往将图像整体处理的方法,本文使用SwinT提取了4个阶段的图像特征,将图像与文本特征在多个层级上进行融合,以获得更为丰富的视觉语义信息.

(3)在Fashion200k[20]、MIT-States[21]和CSS[8]这3个数据集上进行了性能比较,实验结果表明:所提组合检索方法不仅相较于传统的单模态检索方法具有明显优势,还超越了目前主流的图像文本组合检索方法.

2　相关工作

2.1　多模态学习架构

多模态学习架构根据视觉编码(Visual Embed, VE)、文本编码(Textual Embed, TE)以及模态融合(Modality Interaction, MI)的计算复杂度,大致分为4类:(a)VE>TE>MI;(b)VE=TE>MI;(c)VE>MI>TE;(d)MI>VE=TE. 根据ViLT[22]、CLIP[23]、ALBEF[24]、VLMO[25]等研究可知:

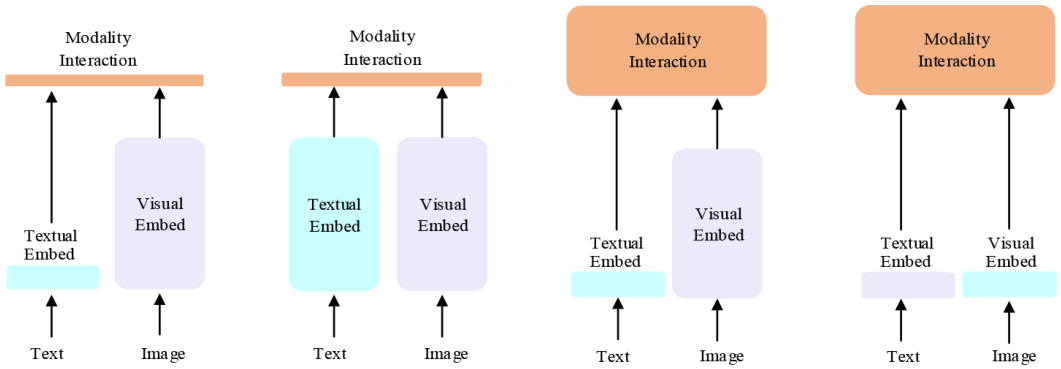
(1)图2(c)中VE>MI>TE的多模态学习架构比图2(*a*)、图2(*b*)、图2(*d*)的模型架构实验效果更好. 鉴于图像相较于文本而言具有更高的复杂度,对于图像处理的计算需求通常会大于文本处理. 因此,文本提取模型的设计应更为轻量化,以便于高效处理文本数据. 而模态融合模块的计算则需在图像与文本之间找到一个平衡点,既不过于繁重,也不过于简化,以确保整体检索系统的效率和性能.

(2)目前视觉转换器在同时处理文本和图像方面有所尝试,但其效果尚未达到理想状态. 因此,为了确保特征提取的准确性和效率,图像和文本的特征提取过程仍需分别进行,以便更精确地捕捉各自的信息特点.

基于以上研究,本文采用类似图2(*c*)的多模态学习架构,图像特征提取器采用较为重量级的SwinT[18],文本特征提取器采用较为轻量级的LSTM[19],并进行了中等量级的模态融合. ViLT[22]、CLIP[23]、ALBEF[24]、VLMO[25]等多模态大模型需要在大规模的数据上进行训练,依赖于庞大的计算机资源,并且人工标注成本高昂,而本文提出的方法成本则相对较低.

2.2　图像-文本特征提取器

图像特征提取器:随着计算视觉的发展,越来越多的视觉转换器被提出,例如ViT[26]视觉转换器和SwinT[18]转换器. 它们使用自注意力机制来获取图像中的全局和局部依赖关系,从而提高图像检索、分类、检测等任务的性能. ViT的核心思想是利用转换器的自注意力机制,来获取图像中不同区域之间的长距离依赖关系,从而提高图像的表示能力. ViT不需要考虑图像的局部结构和平移不变性,也不需要使用卷积神经网络,而是直接将图像视为一个一维的序列,以简化模型的结构和计算. ViT的局限性是需要将图像划分为固定大小的块(patch),会导致图像的细节信息丢失,并且自注意力机制涉及所有块之间的计算,会导致计算复杂



(*a*) VE>TE>MI (*b*) VE=TE>MI (*c*) VE>MI>TE (*d*) MI>VE=TE

图2 多模态学习架构[22]

度和显存占用较高.SwinT通过设计层级式的转换器来学习图像中不同尺度的特征,其表示方法是用移动窗口计算的.移动窗口方案将自我注意的计算限制在非重叠的局部窗口,同时也允许跨窗口连接,从而带来更大的效率.相对于ResNet[27]和ViT,SwinT通过引入滑动窗口的自注意力机制、分层的金字塔结构和一些优化技巧,在图像特征提取上取得了更好的性能.

文本特征提取器:LSTM[19]是一种常见的RNN(Recurrent Neural Network)类型,它可以学习文本序列中的长期依赖关系,从而提取有用的文本特征.LSTM的主要特点是具有3个门(输入门、遗忘门和输出门),它们可以控制信息的流动.BERT(Bidirectional Encoder Representation from Transformers)[28]抛弃了传统的单向语言模型或者2个单向语言模型的浅层结合方法,而是采取了一种掩码语言模型用于双向语言特征的生成.GPT(Generative Pre-trained Transformer)系列[29]是一种基于转换器的自回归语言模型,它的主要思想是使用自回归方式生成文本序列.在文本处理中,GPT系列通常被用来生成文本,例如自动文本摘要和机器翻译等.相对于LSTM,BERT和GPT系列更适用于处理大规模的自然语言文本数据,可以通过预训练来学习通用的文本表示.而LSTM的参数量相对较少,训练和推理的速度较快,占用的内存和计算资源较少,因此本文采用LSTM作为文本编码器.

2.3 图像-文本特征融合方法

现有的图像文本融合方法大致分为全局融合方法、局部融合方法、全局-局部融合方法以及其他融合方法.

全局融合方法.TIRG[8]门控残差方法通过文本修改参考图像的全局特征,而不是传统的特征融合,门控表示输入的参考图像特征是输出合成特征的参考,添加的残差连接表示在该特征空间中的修改.Relationship[9]关系推理方法使用卷积神经网络的最后一层全局图像特征,同时结合循环神经网络提取文本特征,采用级联和多层感知机操作来学习多模态关系.MRN[10]提出一种对联合残差映射使用逐元素乘法的视觉问答方法,通过线性变换最后一层全局图像特征来获得图像文本融合输出.

局部融合方法.LBF[11]采用局部有界特性进行图像文本组合检索,首先对图像中的一组局部区域进行特征提取,然后使用带有自我注意层的单独分支处理局部图像集和修改文本;在跨模态模块利用注意力机制将每个单词与图像中的每个实体关联起来,学习查询图像和修改文本的联合表示.VAL[12]采用基于视觉语言注意力学习的文本联合图像检索方法,使用注意力机制将文本与参考图像的局部特征图融合,特点是在卷积神经网络内部多级插入多个复合转换器(composite transformer),以组合图像特征和文本语义.

全局-局部融合方法.DCNET[13]基于双网络模型进行图像文本组合检索,由2个网络组成,组合网络与TIRG[8]一致,组合参考图像和修改文本的复合特征,使其与目标图像匹配,校正网络模拟参考图像和目标图像之间的差异,使这个差异与修改文本匹配.DCNET对参考图像的局部特征和全局特征做了简单融合.CLVC-Net[14]是一种兼顾局部和全局特征,并且两者相互学习增强的网络模型.它设计了2个合成模块:细粒度局部合成模块和细粒度全局合成模块,针对多模态数据的融合.CoSMo[15]是一种图像-文本合成器,能够对参考图像的内容和风格进行调制,有选择地保留和修改参考图像的特征.FashionVLP[16]是一种基于视觉转换器的模型,将大型图像文本语料库中包含的先验知识引入时尚图像检索领域,并结合来自多个上下文级别的视觉信息(整体图像、裁剪服装局部区域、时尚地标和兴趣区域)来有效捕获时尚相关信息.Comquery-Former[17]是一种全局-局部对齐的多模态检索方法,引入了一个高效的全局-局部对齐模块来缩小组合查询与目标图像之间的距离.它不仅考虑了全局联合嵌入空间的差异,而且迫使模型关注局部细节差异.

其他融合方法.MPC[7]是一种多模态概率组合器

方法,允许在任意模态下对灵活数量的查询进行合成嵌入. 它的概率性质允许编码语义和给定输入的模糊性,从而很好地捕捉到文本查询中的多语义信息. 通用条件图像相似度基准(GeneCIS)[30]通过仅针对零样本测试的设计评估一组开放的相似性条件,利用Combiner架构[31]来融合文本条件和图像特征. FiLM[32]使用仿射变换将文本特征注入到参考图像特征中,并基于该特征检索目标图像.

3　方法

融合图像与文本特征的组合检索方法网络架构如图3所示. 输入为参考图像及其修改文本,参考图像经过SwinT提取4阶段特征,修改文本经过LSTM提取文本特征,目标图像经过SwinT提取目标图像特征. 将文本特征与多层参考图像特征进行融合,与目标图像特征共同嵌入到同一特征向量空间,采用基于批次的分类损失函数进行训练. 接下来,将从图像-文本特征提取、图像-文本特征融合和目标函数3个方面介绍.

3.1　图像-文本特征提取

本文使用预训练在ImageNet-1K上的SwinT作为图像提取器,如图3所示,提取了参考图像4阶段特征. 这样做的优点是:(1)可以实现多尺度的图像文本匹配,从低层到高层,逐渐捕捉更多的语义信息;(2)可以提高检索的召回率,因为不同阶段的图像特征可以与文本特征以不同的方式进行组合;(3)可以提高模型的鲁棒性,适应不同的检索场景,因为不同阶段的图像特征可以根据参考图像和修改文本的内容和长度进行自适应选择.

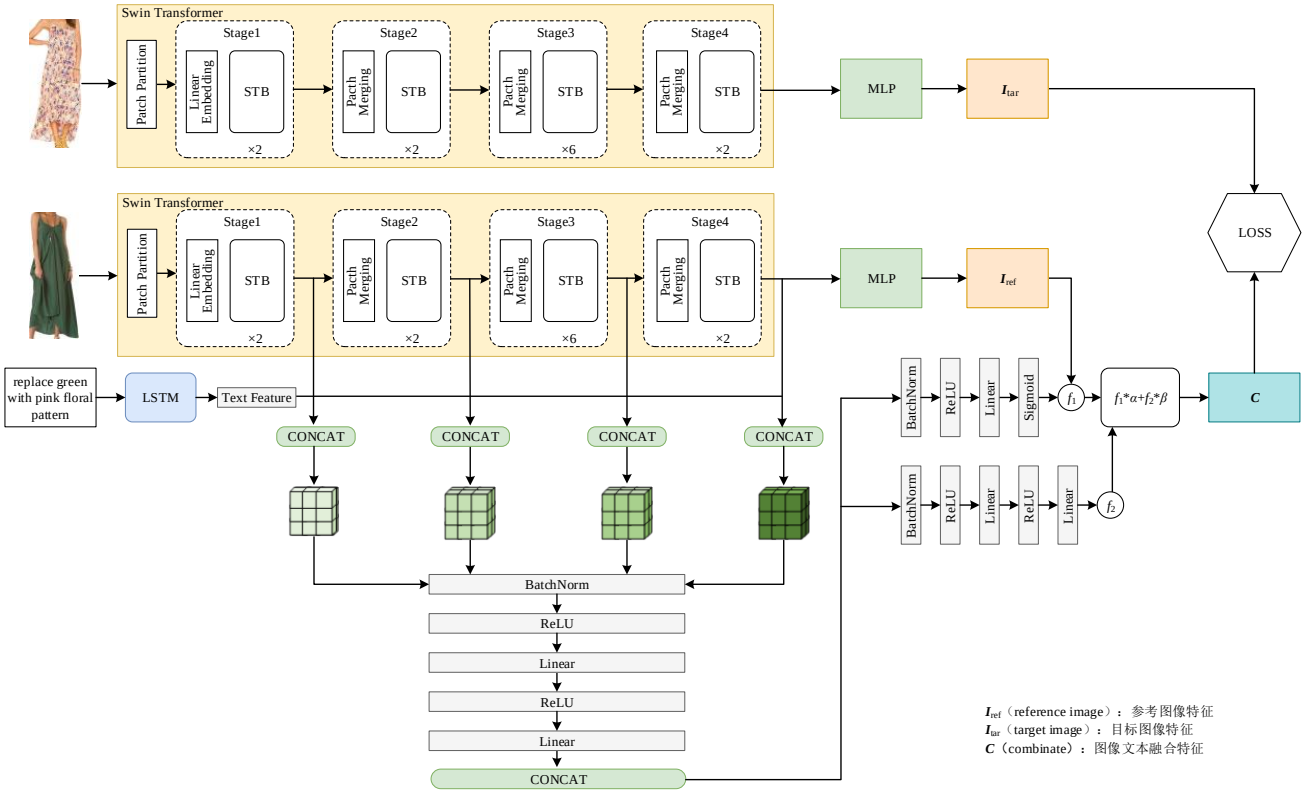


图3　融合图像与文本特征的组合检索方法网络架构图

图像的特征提取过程分为4个阶段:

Stage1. 将图像分割成若干个4×4的小块,每个小块作为1个标记,经过线性嵌入和2个STB(Swin Transformer Block)的处理. 这个阶段的输出是1个$H/4\times W/4\times C$的特征图,其中H、W为图像的长度和高度,C是线性嵌入的维度.

Stage2. 将Stage1的输出通过块合并(Patch Merging)层,将相邻的4个标记合并成1个更大的标记,然后经过2个STB的处理. 这个阶段的输出是1个$H/8\times W/8\times 2C$的特征图.

Stage3. 将Stage2的输出通过块合并层,将相邻的4个标记合并成1个更大的标记,然后经过6个STB的处理. 这个阶段的输出是1个$H/16\times W/16\times 4C$的特征图.

Stage4. 将Stage3的输出通过块合并层,将相邻的4个标记合并成1个更大的标记,然后经过2个STB的处理. 这个阶段的输出是1个$H/32\times W/32\times 8C$的特征图.

STB结构如图4所示,1个STB由1个基于移动窗口的多头自注意力模块(MSA)组成,后面跟着1个2层的

多层感知机(MultiLayer Perceptron,MLP). 在每个MSA模块和每个MLP模块之前,都进行了层归一化(LayerNorm,LN);在每个模块之后,都应用了1个残差连接. MSA模块分别是窗口化自注意力(W-MSA)和移动窗口自注意力(SW-MSA),它们可以有效地提取图像的局部和全局特征,同时降低计算复杂度.

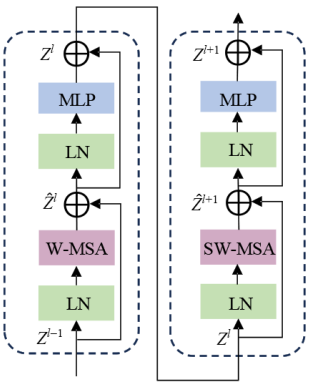


图4 2个连续的STB结构[18]

将参考图像ref和目标图像tar输入SwinT模型中,分别得到4阶段参考图像特征$\boldsymbol{I}_1$、$\boldsymbol{I}_2$、$\boldsymbol{I}_3$、$\boldsymbol{I}_4$,经过MLP的参考图像特征$\boldsymbol{I}_{\text{ref}}$和目标图像特征$\boldsymbol{I}_{\text{tar}}$:

$$\boldsymbol{I}_1,\boldsymbol{I}_2,\boldsymbol{I}_3,\boldsymbol{I}_4=\text{Swin}(\text{ref}) \tag{1}$$

$$\boldsymbol{I}_{\text{ref}}=\text{MLP}(\text{Swin}(\text{ref})) \tag{2}$$

$$\boldsymbol{I}_{\text{tar}}=\text{MLP}(\text{Swin}(\text{tar})) \tag{3}$$

使用word2id将文本转换为词典列表,将得到的文本词向量text输入到LSTM神经网络中进行词编码,从而得到整个文本的特征向量$\boldsymbol{w}$:

$$\boldsymbol{w}=\text{LSTM}(\text{text}) \tag{4}$$

3.2 图像-文本特征融合

经过特征提取器后得到图像4个阶段特征$\boldsymbol{I}_1$、$\boldsymbol{I}_2$、$\boldsymbol{I}_3$、$\boldsymbol{I}_4$,将文本特征$\boldsymbol{w}$通过张量的广播机制,使其维度分别与4个阶段的图像特征相同,得到$\boldsymbol{w}_1$、$\boldsymbol{w}_2$、$\boldsymbol{w}_3$、$\boldsymbol{w}_4$. 接着进行了Concat连接,得到图像文本连接特征:

$$\boldsymbol{C}_1,\boldsymbol{C}_2,\boldsymbol{C}_3,\boldsymbol{C}_4=\text{Concat}\left([\boldsymbol{I}_1,\boldsymbol{w}_1],[\boldsymbol{I}_2,\boldsymbol{w}_2],[\boldsymbol{I}_3,\boldsymbol{w}_3],[\boldsymbol{I}_4,\boldsymbol{w}_4]\right) \tag{5}$$

然后,进行BN-ReLU-Linear-ReLU-Linear操作,它由1个BN层、2个ReLU层和2个Linear层组成. BN层可以对输入数据进行归一化;ReLU层是激活函数操作,可以增加网络的非线性,提高模型的表达能力;Linear层是一种全连接层,可以对输入数据进行线性变换. 经过BN-ReLU-Linear-ReLU-Linear网络单元得到图像文本融合特征:

$$\begin{aligned}&\boldsymbol{C}_1',\boldsymbol{C}_2',\boldsymbol{C}_3',\boldsymbol{C}_4'\\&=\text{Linear}\left(\text{ReLU}\left(\text{Linear}\left(\text{ReLU}\left(\text{BN}\left(\boldsymbol{C}_1,\boldsymbol{C}_2,\boldsymbol{C}_3,\boldsymbol{C}_4\right)\right)\right)\right)\right)\end{aligned} \tag{6}$$

随后对$\boldsymbol{C}_1'$、$\boldsymbol{C}_2'$、$\boldsymbol{C}_3'$、$\boldsymbol{C}_4'$进行了Concat连接操作,得到4阶段图像文本融合特征$\boldsymbol{C}_m$:

$$\boldsymbol{C}_m=\text{Concat}\left(\boldsymbol{C}_1',\boldsymbol{C}_2',\boldsymbol{C}_3',\boldsymbol{C}_4'\right) \tag{7}$$

受TIRG[8]方法的启发,所提方法旨在从现有的特征中创建出一个新的特征,通过文本对参考图像特征进行修改,而非传统的特征融合方式. 在此过程中,$\boldsymbol{f}_1$负责将输入的图像特征作为最终融合特征的参考,用以增强参考图像特征. $\boldsymbol{f}_2$表示在这个特征空间中文本对图像的修改. 最终得到图像文本融合特征$\boldsymbol{C}$:

$$\boldsymbol{f}_1=\boldsymbol{I}_{\text{ref}}\odot\text{Sigmoid}\left(\text{Linear}\left(\text{ReLU}\left(\text{BN}\left(\boldsymbol{C}_m\right)\right)\right)\right) \tag{8}$$

$$\boldsymbol{f}_2=\text{Linear}\left(\text{ReLU}\left(\text{Linear}\left(\text{ReLU}\left(\text{BN}\left(\boldsymbol{C}_m\right)\right)\right)\right)\right) \tag{9}$$

$$\boldsymbol{C}=\boldsymbol{f}_1*\alpha+\boldsymbol{f}_2*\beta \tag{10}$$

其中,$\odot$表示点积,Sigmoid是常用的激活函数,具有将输入值映射至接近0与1之间区间内的功能;α和β表示可以学习的参数,它们之间没有特定的约束. 整个学习过程分为5个步骤:(1)在训练开始之前,α和β会被初始化为随机值;(2)模型根据当前α和β计算预测结果$\boldsymbol{C}$,采用基于批次的分类损失函数,计算预测结果$\boldsymbol{C}$与目标特征$\boldsymbol{I}_{\text{tar}}$之间的损失;(3)通过反向传播算法计算$\alpha$和$\beta$的梯度;(4)使用随机梯度下降算法,根据计算出的梯度更新α和β,目的是最小化损失函数;(5)重复步骤(2)~(4),直到模型收敛,或者达到预定的训练轮次.

3.3 目标函数

本文采用一种基于批次的分类损失函数来训练模型. 具体而言,每个批次内的样本会与同一批次内的其他样本进行比较. 训练的主要目标是最大化修改图像与目标图像之间的相似度,并有效地区分不相似图像的特征,以达到损失函数的最小化. 对于一批包含B个图像-文本的融合特征数据,基于批次的分类损失函数定义为

$$L=-\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(\text{sim}\left(\boldsymbol{C}^i,\boldsymbol{I}_{\text{tar}}^i\right)\right)}{\sum_{j=1}^{B}\exp\left(\text{sim}\left(\boldsymbol{C}^i,\boldsymbol{I}_{\text{tar}}^j\right)\right)} \tag{11}$$

其中,sim为相似性度量,采用点积进行相似度计算.

4 实验

为了验证所提方法的有效性,在Fashion200k[20]、MIT-States[21]和CSS[8]这3个数据集上开展了实验比较,使用参考图像和修改文本作为查询,输出目标图像. 检索指标采用K级召回率(R@K). R@K反映了在前K个检索结果中,实际与查询相关的结果所占的比例. K是指按相似度排序的推荐列表中返回的前K个结果.

在3个数据集上共同采用的对比方法,包括Image only、Text only、Concat、Relationship[9]、MRN[10]、FiLM[32]、

TIRG[8]以及LBF[11]方法. 针对Fashion200k数据集,拓展了对比方法的范围,新增了VAL[12]、DCNET[13]、CLVC-Net[14]、CoSMo[15]、FashionVLP[16]以及ComqueryFormer[17]方法. 在MIT-States数据集上,新增了GeneCIS[30]方法进行比较研究.

在实验中采用了PyTorch深度学习框架作为基础,使用在ImageNet-1K数据集上预先训练的SwinT作为图像编码器,该编码器输出的特征维度达到1 024. 至于文本编码部分,选用了LSTM网络,其中隐藏层的维度设置为1 024,初始权重通过随机数进行初始化. 在训练过程中,默认执行了共计48万次的迭代,每次迭代的batchsize为32,初始学习率设定在0.01,随着迭代次数的累加,学习率会逐步递减. 在实验中使用的操作系统是Ubuntu,实验平台软硬件参数信息如表1所示. 经过统计,模型的浮点运算次数(FLOPs)为32.38 GFLOPs,参数量Params为176.06×10^6.

表1　环境配置信息

组件	型号
操作系统	Ubuntu 22.04 64位
深度学习框架	PyTorch 1.7.0 on Python 3.7.16
CPU	Intel® Xeon(R) CPU E5-2699A v4 @ 2.40 GHz × 88
GPU	NVIDIA Corporation GA102GL [RTX A5000] × 2
CUDA	CUDA 11.0
内存	512 GB

4.1　数据集介绍

Fashion200k[20]、MIT-States[21]和CSS[8]数据集的划分见表2,分别包含了训练集和测试集.

表2　数据集划分

数据集	训练集	测试集
Fashion200k[20]	172 049	29 789
MIT-States[21]	43 207	10 546
CSS[8]	19 012	19 057

(1)Fashion200k是一个具有挑战性的时尚商品数据集,大约由20万张时尚商品图像组成,商品类别丰富,包含连衣裙、外套、裤子、半袖、衬衫等. 其中172 049张用于训练,29 789张用于测试. 每张图片都带有类似属性的产品英文描述,例如"black knit medium length dress"(黑色针织中长裙). 查询的创建如下:选择描述中存在一个单词差异的一对产品作为查询图像和目标图像,修改文本是另一个单词.

(2)MIT-States是一个用于描述物品状态的数据集,由6万多张图像组成,每张图像都带有一个英文名词标签和一个英文形容词标签,例如"clear sky"(晴朗的天空). 数据集共有245个名词和115个形容词,平均每个名词被它所提供的9个形容词修饰,其中43 207张用于训练,10 546张作为测试. 实验中,对具有相同对象标签和不同状态标签的图像对采样,它们分别作为查询图像和目标图像使用,修改文本是目标图像物体的状态描述. 因此,应该检索与查询图像相同对象的图像,但具有修改文本描述的新状态.

(3)CSS数据集是一个合成的图像数据集,其中包含几种不同布局的不同几何对象(球体、立方体等). CSS是在CLEVR平台之上产生的. 它分别包含大约19 012张训练图像和19 057张测试图像. 该数据集的修改文本分为3类:向图像中添加新对象、从图像中删除对象,以及更改图像中当前对象的属性. 添加对象需要指定要放置在场景中的新对象(其颜色、大小、形状和位置),删除对象需要指定从场景中移除的对象,更改对象需要指定要更改的对象及其新属性值.

4.2　实验结果

4.2.1　Fashion200k数据集实验结果

Fashion200k数据集实验结果如表3所示,最佳结果以粗体显示,"—"表示该方法在此类别上没有发布结果.

表3　Fashion200k数据集实验结果　　单位:%

方法	R@1	R@10	R@50
Image only	5.6	24.9	47.8
Text only	1.7	12.6	22.1
Concat	11.8	41.2	57.8
Relationship[9]	13.0	40.5	62.4
MRN[10]	13.4	40.0	61.9
FiLM[32]	12.9	39.5	61.9
TIRG[8]	14.1	42.5	63.8
DCNET[13]	—	46.9	67.6
LBF[11]	17.8	48.4	68.5
VAL[12]	21.2	49.0	68.8
CLVC-Net[14]	22.6	53.0	72.2
CoSMo[15]	**23.3**	50.4	69.3
FashionVLP[16]	—	49.9	70.5
ComqueryFormer[17]	—	52.2	72.2
ours	22.4	**54.4**	**72.9**

实验结果表明:

(1)本文方法在R@10和R@50上均优于其他模型,验证了本文方法在Fashion200k数据集上的有效性. 这是因为使用SwinT提取图像特征具有显著的优势,通过将包含更多语义信息的4阶段图像特征与文本特征进行融合,实现了模型性能的进一步提升.

(2)单纯的图像检索(Image only)和文本检索(Text only)相较于其他组合检索方法表现较差,组合检索方法能够整合不同模态信息,其综合性能超越了单模态检索方法.

一些可视化检索结果如图5所示，例如输入的是一张白色衬衫图像，修改文本为“replace with gray”（替换为灰色），输出为相关的图像，绿色框代表符合查询的检索结果.

图5 Fashion200k检索结果示例

4.2.2 MIT-States数据集实验结果

MIT-States数据集实验结果如表4所示，最佳结果以粗体显示. 实验结果表明：

（1）本文方法在MIT-States数据集上的R@1、R@5和R@10指标均优于其他主流模型，相较于GeneCIS方法，分别实现了2.7、3.6和2.8个百分点的性能提升，验证了本文方法在MIT-States数据集上的有效性. 融合多层图像特征与文本特征相比其他方法能够捕获到更为丰富的语义信息，从而实现了更高的性能表现.

表4 MIT-States数据集实验结果 单位:%

方法	R@1	R@10	R@50
Image only	7.1	18.3	24.6
Text only	9.7	20.6	32.1
Concat	13.9	31.7	42.8
Relationship[9]	12.3	31.9	42.9
MRN[10]	11.9	30.5	41.0
FiLM[32]	10.1	27.7	38.3
TIRG[8]	12.2	32.2	43.1
LBF[11]	14.7	35.3	46.5
GeneCIS[30]	15.8	37.5	49.4
ours	**18.5**	**41.1**	**52.2**

（2）在MIT-States数据集上，单模态检索仍然表现较差，而将图像和文本2种模态组合，文本能够对图像进行修改，可以避免落入单一模态的局限困境.

一些可视化检索结果如图6所示，例如输入的是一张晴朗的天空图像，修改文本为“change state to cloudy”（改变状态为多云），输出为相关的图像，绿色框代表符合查询的检索结果.

4.2.3 CSS数据集实验结果

CSS数据集实验结果如表5所示，最佳结果以粗体显示. 本文方法在实验中使用了数据集的3D和2D版本.

图6 MIT-States检索结果示例

实验结果表明：在3D图像检索3D图像（3D→3D）任务中，本文方法与对比的最好方法相比，R@1提升了6.1个百分点. 在2D图像检索3D图像（2D→3D）任务中，本文方法与对比方法相比，获得了较大的提升，比LBF高了22.8个百分点. 使用SwinT提取多层图像特征，并将其与文本特征进行多模态融合，使模型更好地捕捉到2D图像中的细节和语义信息，从而提高了检索精度. 一些可视化检索结果如图7所示，例如输入的是一张包含多个几何对象的图片，修改文本为“add small yellow circle to bottom-right”（在右下角添加一个小的黄色圆形物体），输出为相关的图像，绿色框代表符合查询的检索结果.

表5 CSS数据集实验结果 单位:%

方法	3D→3D R@1	2D→3D R@1
Image only	10.1	9.3
Text only	0.7	0.7
Concat	61.9	29.7
Relationship[9]	62.1	30.6
MRN[10]	60.1	26.8
FiLM[32]	65.6	43.7
TIRG[8]	73.7	46.6
LBF[11]	79.2	55.7
ours	**85.3**	**78.5**

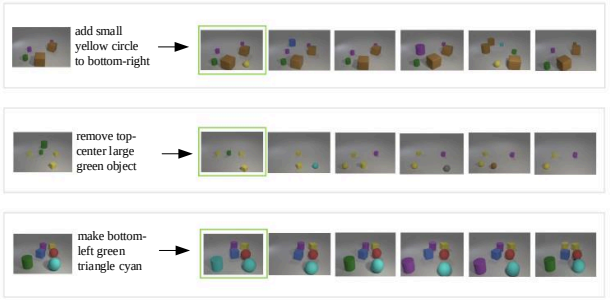


图7 CSS检索结果示例

本文方法在3个数据集上超越了目前较为先进的方法，验证了本文方法的有效性. 在3个数据集上的实验结果表明：将多层图像特征与文本特征进行融合能够有效提升检索性能；SwinT在图像特征提取上的强大性能，为模型的高效表现提供了重要支持；相比单一模

态检索,图像文本组合检索通过整合多模态信息,显著提升了检索的准确性和鲁棒性.

4.3 消融研究

在Fashion200k数据集上,分别对图像特征提取器、文本提取器以及融合方法做消融研究. 相比其他数据集,Fashion200k在规模和类别多样性方面具有明显优势. 其图像和文本描述来源于真实的时尚商品展示,具有较高的挑战性和现实应用价值,因此在Fashion200k数据集上进行消融研究更具代表性和普适性.

4.3.1 图像特征提取器消融研究

本文采用了ResNet、ViT以及SwinT的4个版本(SwinT-T(iny)、SwinT-S(mall)、SwinT-B(ase)、SwinT-L(arge))作为图像特征提取器,并使用SwinT进一步提取了参考图像的4阶段特征. 同时,进一步探讨了batchsize对实验结果的影响. 需要注意的是,由于实验设备的条件限制,在使用SwinT-L版本时,其batchsize最高为26. 实验结果如表6所示.

表6 图像特征提取器消融研究结果

图像特征提取器	batchsize	R@1/%	R@10/%	R@50/%
ResNet-18	32	17.1	42.9	64.6
ResNet-50	32	19.3	48.6	67.5
ViT	32	20.6	51.2	70.3
SwinT-T	32	17.3	46.8	66.2
SwinT-S	32	19.7	50.4	70.1
SwinT-B	18	17.3	48.2	65.3
	24	18.6	50.3	69.8
	32	21.3	52.9	71.6
	36	19.7	51.8	70.3
SwinT-L	18	18.1	50.7	68.7
	24	19.7	51.4	70.3
	26	20.1	52.2	71.8
SwinT-B+4阶段	32	**22.4**	**54.4**	**72.9**

表6结果表明:

(1)SwinT-B版本要优于ResNet和ViT,体现了SwinT-B在图像特征提取任务中具有更好的表现能力和更高的准确性,这种优势主要得益于SwinT独特的自注意力机制和层次化结构设计.

(2)在条件相同的情况下,SwinT-L会优于SwinT-T、SwinT-S和SwinT-B. 那么可以推断,如果实验条件允许,采用SwinT-L版本(batchsize=32)提取图像4阶段特征可能会取得更好的效果.

(3)通过SwinT-B版本的实验效果可以发现,batchsize的大小会对实验结果有影响,在32以内,模型效果会随着batchsize的增加而增加.

(4)采用SwinT提取4阶段特征后进行融合要优于只使用最后阶段特征进行融合,在batchsize=32时,R@1、R@10和R@50分别提高了1.1、1.5和1.3个百分点,这是因为4阶段特征有着更多的语义信息,包含了图像的局部和全局特征信息,从而更有利于模型进行检索.

4.3.2 文本特征提取器消融研究

本文采用了Bert和LSTM进行消融研究,图像提取器均采用提取4阶段的SwinT-B. 如表7所示,实验结果表明:采用Bert的实验效果要比LSTM的效果差,原因在于Fashion200k数据集的文本较短,信息含量不足,而Bert模型则是针对长文本数据集进行预训练的. 对于简短的文本,Bert可能会导致过拟合和领域不适配问题. 此外,Bert模型训练的资源消耗也相当巨大. 因此,采用更加轻量的LSTM模型,可以实现对计算资源的有效平衡.

表7 文本特征提取器消融研究结果 单位:%

文本特征提取器	R@1	R@10	R@50
Bert	19.7	49.8	68.6
LSTM	**22.4**	**54.4**	**72.9**

4.3.3 融合方法消融研究

本文采用了简单融合(Concat)、注意力机制融合以及多层融合方法进行消融研究,图像提取器均采用提取4阶段特征的SwinT-B,文本提取器采用LSTM,实验结果如表8所示. 实验结果表明:采用简单的融合方法,实验效果要比其他方法差,多层融合方法则优于其他方法,相较于注意力机制融合方法,在R@1、R@10和R@50上分别提高了3.1、4.7和4.4个百分点,验证了本文多层融合方法的有效性. 目前比较流行的基于注意力机制的融合方法,与多层融合方法相比存在一定的差距,这进一步体现了多层融合方法能够更有效地整合不同层次的特征和信息.

表8 融合方法消融研究结果 单位:%

融合方法	R@1	R@10	R@50
Concat	11.8	41.2	57.8
注意力机制	19.3	49.7	68.5
多层融合	**22.4**	**54.4**	**72.9**

5 结束语

本文提出了一种融合图像与文本特征的组合检索方法,为图像文本组合检索提供了新的解决思路. 该方法利用SwinT提取了多层图像特征,并将这些特征与文本特征嵌入到一个共同空间中进行多模态融合,使得文本能够更细粒度地修改图像. 在Fashion200k、MIT-States和CSS这3个数据集上的实验结果表明:该方法超越了现有的主流方法,使用SwinT提取图像特征具有显著优势,多层图像特征与文本特征的融合能够捕获更丰富的语义信息,从而实现更高的性能. 未来,将探

索生成模型在图像文本组合检索中的应用,通过图像和文本的相互作用生成全新的图像,进而检索目标图像.

参考文献

[1] DUBEY S R. A decade survey of content based image retrieval using deep learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(5): 2687-2704.

[2] LI X Q, YANG J S, MA J W. Recent developments of content-based image retrieval (CBIR) [J]. Neurocomputing, 2021, 452: 675-689.

[3] ZHANG Q, LEI Z, ZHANG Z X, et al. Context-aware attention network for image-text retrieval[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 3533-3542.

[4] 李志欣, 凌锋, 张灿龙, 等. 融合两级相似度的跨媒体图像文本检索[J]. 电子学报, 2021, 49(2): 268-274.
LI Z X, LING F, ZHANG C L, et al. Cross-media image-text retrieval with two level similarity[J]. Acta Electronica Sinica, 2021, 49(2): 268-274. (in Chinese)

[5] 冯霞, 胡志毅, 刘才华. 跨模态检索研究进展综述[J]. 计算机科学, 2021, 48(8): 13-23.
FENG X, HU Z Y, LIU C H. Survey of research progress on cross-modal retrieval[J]. Computer Science, 2021, 48(8): 13-23. (in Chinese)

[6] 冯奕, 周晓松, 李传艺, 等. 基于多模态特征融合嵌入的相似广告检索方法[J]. 计算机学报, 2022, 45(7): 1500-1516.
FENG Y, ZHOU X S, LI C Y, et al. A multi-modal feature fusion embedding method for similar ad retrieving[J]. Chinese Journal of Computers, 2022, 45(7): 1500-1516. (in Chinese)

[7] NECULAI A, CHEN Y B, AKATA Z. Probabilistic compositional embeddings for multimodal image retrieval[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2022: 4546-4556.

[8] VO N, JIANG L, SUN C, et al. Composing text and image for image retrieval - an empirical odyssey[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 6432-6441.

[9] SANTORO A, RAPOSO D, BARRETT D G T, et al. A simple neural network module for relational reasoning[EB/OL]. (2017-06-05)[2024-07-22]. https://arxiv.org/abs/1706.01427v1.

[10] KIM J H, LEE S W, KWAK D, et al. Multimodal residual learning for visual QA[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: ACM, 2016: 361-369.

[11] HOSSEINZADEH M, WANG Y. Composed query image retrieval using locally bounded features[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 3593-3602.

[12] CHEN Y B, GONG S G, BAZZANI L. Image search with text feedback by visiolinguistic attention learning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 2998-3008.

[13] KIM J, YU Y, KIM H, et al. Dual compositional learning in interactive image retrieval[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(2): 1771-1779.

[14] WEN H K, SONG X M, YANG X, et al. Comprehensive linguistic-visual composition network for image retrieval[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2021: 1369-1378.

[15] LEE S, KIM D, HAN B. CoSMo: Content-style modulation for image retrieval with text feedback[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 802-812.

[16] GOENKA S, ZHENG Z H, JAISWAL A, et al. FashionVLP: Vision language transformer for fashion retrieval with feedback[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 14085-14095.

[17] XU Y H, BIN Y, WEI J W, et al. Multi-modal transformer with global-local alignment for composed query image retrieval[J]. IEEE Transactions on Multimedia, 2023, 25: 8346-8357.

[18] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 9992-10002.

[19] GREFF K, SRIVASTAVA R K, KOUTNÍK J, et al. LSTM: A search space odyssey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(10): 2222-2232.

[20] HAN X T, WU Z X, HUANG P X, et al. Automatic spatially-aware fashion concept discovery[C]//2017 IEEE In-

ternational Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 1472-1480.

[21] ISOLA P, LIM J J, ADELSON E H. Discovering states and transformations in image collections[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2015: 1383-1391.

[22] KIM W, SON B, KIM I. ViLT: Vision-and-language transformer without convolution or region supervision[C]//Proceedings of the International Conference on Machine Learning. Stockholm: PMLR, 2021: 5583-5594.

[23] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the International Conference on Machine Learning. Stockholm: PMLR, 2021: 8748-8763.

[24] LI J N, SELVARAJU R R, GOTMARE A D, et al. Align before fuse: Vision and language representation learning with momentum distillation[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. New York: ACM, 2024: 9694-9705.

[25] BAO H B, WANG W H, DONG L, et al. VLMo: Unified vision-language pre-training with mixture-of-modality-experts[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New York: ACM, 2024: 32897-32912.

[26] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. (2020-10-20) [2024-07-22]. https://arxiv.org/abs/2010.11929.

[27] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[28] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis: ACL, 2019: 4171-4186.

[29] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[EB/OL]. (2020-05-28) [2024-07-22]. https://arxiv.org/abs/2005.14165.

[30] VAZE S, CARION N, MISRA I. GeneCIS: A benchmark for general conditional image similarity[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 6862-6872.

[31] BALDRATI A, BERTINI M, URICCHIO T, et al. Conditioned and composed image retrieval combining and partially fine-tuning CLIP-based features[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2022: 4955-4964.

[32] PEREZ E, STRUB F, DE VRIES H, et al. FiLM: Visual reasoning with a general conditioning layer[C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New York: ACM, 2018: 3942-3951.

作者简介

秦钰淑　女,1999年4月出生于山西省长治市. 现为浙江工业大学硕士研究生. 主要研究方向为多模态检索.
E-mail: qinyushu1999@163.com

朱艳超　女,1981年11月出生于浙江省丽水市. 毕业于大连海事大学. 主要研究方向为信息系统.
E-mail: dongdong7981@126.com

杨良怀　男,1967年出生于浙江省新昌县. 现为浙江工业大学计算机学院教授. 主要研究方向为数据科学与工程.
E-mail: yanglh@zjut.edu.cn

龚卫华　男,1977年生于湖北省武汉市汉阳区. 现为浙江工业大学计算机学院副教授. 主要研究方向为机器学习、社会网络.
E-mail: whgong@sohu.com