计算机科学
COMPUTER SCIENCE

基于Transformer的图文跨模态检索算法

杨晓宇, 李超, 陈舜尧, 李浩亮, 殷光强

引用本文

杨晓宇, 李超, 陈舜尧, 李浩亮, 殷光强基于Transformer的图文跨模态检索算法[J]. 计算机科学, 2023, 50(4): 141-148.

YANG Xiaoyu, LI Chao, CHEN Shunyao, LI Haoliang, YIN Guangqiang. Text-Image Cross-modal Retrieval Based on Transformer [J]. Computer Science, 2023, 50(4): 141-148.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220100083

基于 Transformer 的图文跨模态检索算法

杨晓宇　李　超　陈舜尧　李浩亮　殷光强
电子科技大学公共安全技术研究中心　成都 611731
(yangxy@std.uestc.edu.cn)

摘　要　随着互联网多媒体数据的不断增长,文本图像检索已成为研究热点。在图文检索中,通常使用相互注意力机制,通过将图像和文本特征进行交互,来实现较好的图文匹配结果。但是,这种方法不能获取单独的图像特征和文本特征,在大规模检索后期需要对图像文本特征进行交互,消耗了大量的时间,无法做到快速检索匹配。然而基于 Transformer 的跨模态图像文本特征学习取得了良好的效果,受到了越来越多的关注。文中设计了一种新颖的基于 Transformer 的文本图像检索网络结构(HAS-Net),该结构主要有以下几点改进:1)设计了一种分层 Transformer 编码结构,以更好地利用底层的语法信息和高层的语义信息;2)改进了传统的全局特征聚合方式,利用自注意力机制设计了一种新的特征聚合方式;3)通过共享 Transformer 编码层,使图片特征和文本特征映射到公共的特征编码空间。在 MS-COCO 数据集和 Flickr30k 数据集上进行实验,结果表明跨模态检索性能均得到提升,在同类算法中处于领先地位,证明了所设计的网络结构的有效性。

关键词:Transformer;跨模态检索;特征分层提取;特征聚合;特征共享
中图法分类号　TP399

Text-Image Cross-modal Retrieval Based on Transformer

YANG Xiaoyu,LI Chao,CHEN Shunyao,LI Haoliang and YIN Guangqiang
Center for Public Security Technology,University of Electronic Science and Technology of China,Chengdu 611731,China

Abstract　With the growth of Internet multimedia data,text image retrieval has become a research hotspot. In image and text retrieval,the mutual attention mechanism is used to achieve better image-text matching results by interacting image and text features. However,this method cannot obtain image features and text features separately,and requires interaction of image and text features in the later stage of large-scale retrieval,which consumes a lot of time and is not able to achieve fast retrieval and matching. However,the cross-modal image text feature learning based on Transformer has achieved good results and has received more and more attention from researchers. This paper designs a novel Transformer-based text image retrieval network structure (HAS-Net),which mainly has the following improvements:a hierarchical Transformer coding structure is designed to better utilize the underlying grammatical information and high-level semantic information;the traditional global feature aggregation method is improved,and the self-attention mechanism is used to design a new feature aggregation method;by sharing the Transformer coding layer,image features and text features are mapped to a common feature coding space. Finally,experiments are conducted on the MS-COCO and Flickr30k datasets,the cross-modal retrieval performance has been improved,and it is in a leading position among similar algorithms. It is proved that the designed network structure is effective.

Keywords　Transformer,Cross-modal retrieval,Hierarchical feature extraction,Feature aggregation,Feature share

1　引言

目前,跨模态检索受到越来越多的关注,其目的是从不同的模态中搜索语义相似的样本,特别是互联网上图像内容的爆炸式增长给图像文本的准确检索带来了巨大的挑战。本文重点研究了文本到图像的检索,也希望能对其他跨模态检索任务有所启发。

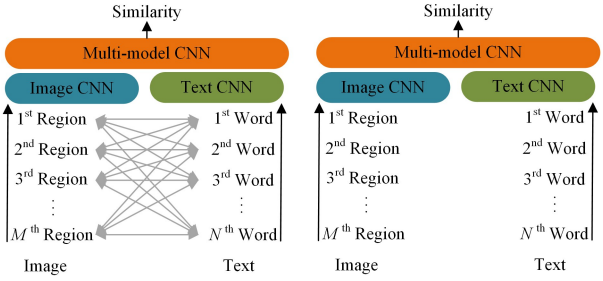
文本图像检索本质是要对图像和文本这两个模态的样本分别进行编码以得到其语义表示,同时还需利用相应的相似性计算方法来计算这些语义表示之间的相似度。现有的文本图像检索方法主要包括跨模态相似性度量方法和公共空间特征学习方法,其大致结构如图 1 所示。

如图 1(a)所示,跨模态相似度度量方法的主要思路是将图文特征进行融合,再经过隐层,目的是让隐层学习到可以度量跨模态相似度的函数。其优点是检索精度较高,因为图文信息融合之后提供了很多或是互补或是对齐的特征信息;

到稿日期:2022-01-10　返修日期:2022-07-05
基金项目:深圳市科技计划项目(JSGG20220301090405009)
This work was supported by the Shenzhen Science and Technology Program(JSGG20220301090405009).
通信作者:殷光强(yingq@uestc.edu.cn)

缺点是检索速度较慢。当用户输入一个文本查询，系统需要在线地将系统中的所有图像与文本成对地输入模型中，才能得到文本与每个图像的相似度分数。

如图 1(b)所示，公共空间特征学习方法是将图像和文本映射到一个公共空间中，得到多模态表示即最后一层的表示，从而直接使用余弦距离计算其相似度。图像和文本相互独立，没有交互，希冀于学习到一个优秀的表示就可以进行相似度度量。这种方法的优点是检索效率高，系统可以提前得到图像和文本的语义表示，进行离线保存，用户输入一个文本查询，直接与保存好的图像表示进行相似度计算即可；缺点是由于缺少交互，检索精度相对较低。



(a)跨模态相似度度量方法 (b)公共空间特征学习方法

图 1 图文检索方法

Fig. 1 Image-Text retrieval methods

在公共空间特征学习架构中，本文提出了一种新的基于 Transformer 的图文检索网络结构(Hierarchical Aggregation Sharing-Network，HAS-Net)，其贡献有以下几点：

(1)特征分层提取(H)：根据 Transformer 结构中不同编码层的注意力分配特性，不同层级提取的特征关注点不相同[1-2]。例如，较低层中的特征倾向于使用基本句法表示对更多局部内容进行编码；更高层的特征可以捕获更复杂的语义信息，通常会产生更高级别的语义表示，如文献[3-4]所述。基于这些特点，本文提出了分层 Transformer 编码结构来实现文本图像跨模态检索。

(2)特征聚合模块(A)：全局特征与局部特征之间有着复杂的关联关系，而传统的特征聚合方式，如求和、平均，以及文献[5]中使用 GRU 或 LSTM 等，都不足以描述特征之间的复杂关系。本文利用自注意力机制，充分挖掘全局特征和局部特征之间的关系，将局部特征通过自注意力机制进行特征聚合，得到更具表现力的全局特征。

(3)特征共享模块(S)：公共空间特征学习方法是在图像文本模型独立不交互的情况下获取高质量、高语义的跨模态表示。本文通过共享多个 Transformer 编码层，使图片特征和文本特征映射到公共的特征编码空间，从而得到高质量、高语义的跨模态表示。

2 相关工作

2.1 文本图像检索

图像和文本之间的跨模态检索问题是如何推断图像和句子之间的相似度，通常，计算这种跨模态的相似度的常用方法是将图像和文本投影到一个共同特征表示空间中，在该空间中定义一种相似度的度量。

在处理图像特征时，常用的方法是基于 CNN 或 CNN 的变种，并且通常会利用图像分类任务来完成预训练。比如，一些工作利用了不同类型的卷积神经网络来进行图像特征的处理。Eisenschtat 等[6]将文本和图像特征分别通过 encoder 映射到共同空间，然后用 L2 计算文本和图像之间的相似性。Faghri 等[7]的工作利用 GRU 将文本与图像映射到同一子空间，并针对 hard negative 样本改进了损失函数。Gu 等[8]将生成过程结合到跨模态特征嵌入中，通过该方法可以学习全局抽象特征和局部层次特征。许多工作通过 GRU 或者 LSTM 循环神经网络来提取文本特征。Huang 等[9]通过 LSTM 提取文本特征，并提出了语义增强图片及语句匹配模型。Li 等[5]提出首先建立图像区域之间的联系，再使用图卷积网络进行推理，生成具有语义关系的特征。

上述方法通常是描述文本或图像的全局特征，可能缺乏对一些有用信息的细节描述，同时也可能产生许多冗余的信息。为了解决这一问题，一些研究尝试使用区域信息来进行图像区域和词之间的细粒度对齐，一些工作利用 Faster-RCNN[10]来提取区域级的图像特征。其中，Chen 等[11]提出基于重复注意记忆的迭代匹配多步对比图像和文本之间的对应信息，从而逐步探索这种细粒度的对应关系。Wang 等[12]通过整合目标位置线索来增强图像文本联合嵌入的学习，以此增强目标位置信息。也有一些工作通过 ResNet 直接提取句子的细粒度信息。Ji 等[13]采用 ResNet-101 的 Faster R-CNN 网络对每一个图像产生 k 个目标区域，提取每个目标对象的特征，使用图像区域和句子中的单词作为上下文来发现完整的潜在对齐，并推断出图像文本的相似性。Xu 等[14]针对高级语义不能严格对应于单个图像区域提出了语义一致性跨模态注意机制(Cross-modal Attention with Sematic Consistency，CASC)，用于图像文本匹配。

2.2 基于 Transformer 的文本图像检索

最近，在很多自然语言处理任务中，如文本分类、上下文预测，Transformer[15]结构都有不俗的表现。BERT 模型[16]使用了掩蔽语言模型和下一句预测的联合训练方法，加上 12 层的 Transformer，证明了注意力机制在产生准确的上下文感知描述方面是十分有效的。因此，一些文本和图像的匹配工作使用 BERT 来提取词向量特征，其中 Qu 等[17]提出了一种新的网络 CAMERA 来解决图像-文本匹配的多视角描述问题。Wei 等[18]将文本和图片的特征进行融合，文本和图片分别与融合后的特征进行 Attention 操作，利用 Attention 结果进行匹配操作。还有一些工作受到自注意力机制的启发，在图像和文本的模式上使用了类似于 BERT 的方法，比如 ViLBERT[19]和 VL-BERT[20]。

目前 Transformer 已经成为了 NLP 领域的主流，但在图像识别领域的应用还在探索之中，其中 Parmar 等[21]在局部领域采用了自注意力机制，Cordonnier 等[22]从图像中提取 2×2 大小的区域，并在此区域上使用完全注意力机制；Dosovitskiy 等[23]将图像切片作为 Transformer 的序列输入。在跨模态图文检索中，一些工作使用 Transformer 对图文特征进行编码，如 TERN[24]，TERAN[25]，MMCA[26]，Unicoder-VL[27]等，其中 MMCA[26]通过连接作为 Transformer 输入的

两个模态特征来探索通道间的关系，TERN[24] 和 TERAN[25] 只探索通道内关系，Cao 等[28] 提出了全局关系感知注意网络，该网络可以联合考虑局部特征、全局特征以及它们之间的关系。

3 本文方法

本文提出了一个新颖的基于 Transformer 的图文检索网络结构(HAS-Net)。如图 2 所示。相比以往的结构，HASNet 主要通过分层 Transformer 编码结构来提取基本语法信息和高级语义信息(见 3.1 节)；其改变了传统的特征聚合方式，利用自注意力机制进行特征聚合(见 3.2 节)；通过共享 Transformer 编码层，使跨模态特征表示映射到同一个特征空间(见 3.3 节)；最后使用 hinge-based triplet loss 损失函数来训练所提模型(见 3.4 节)。

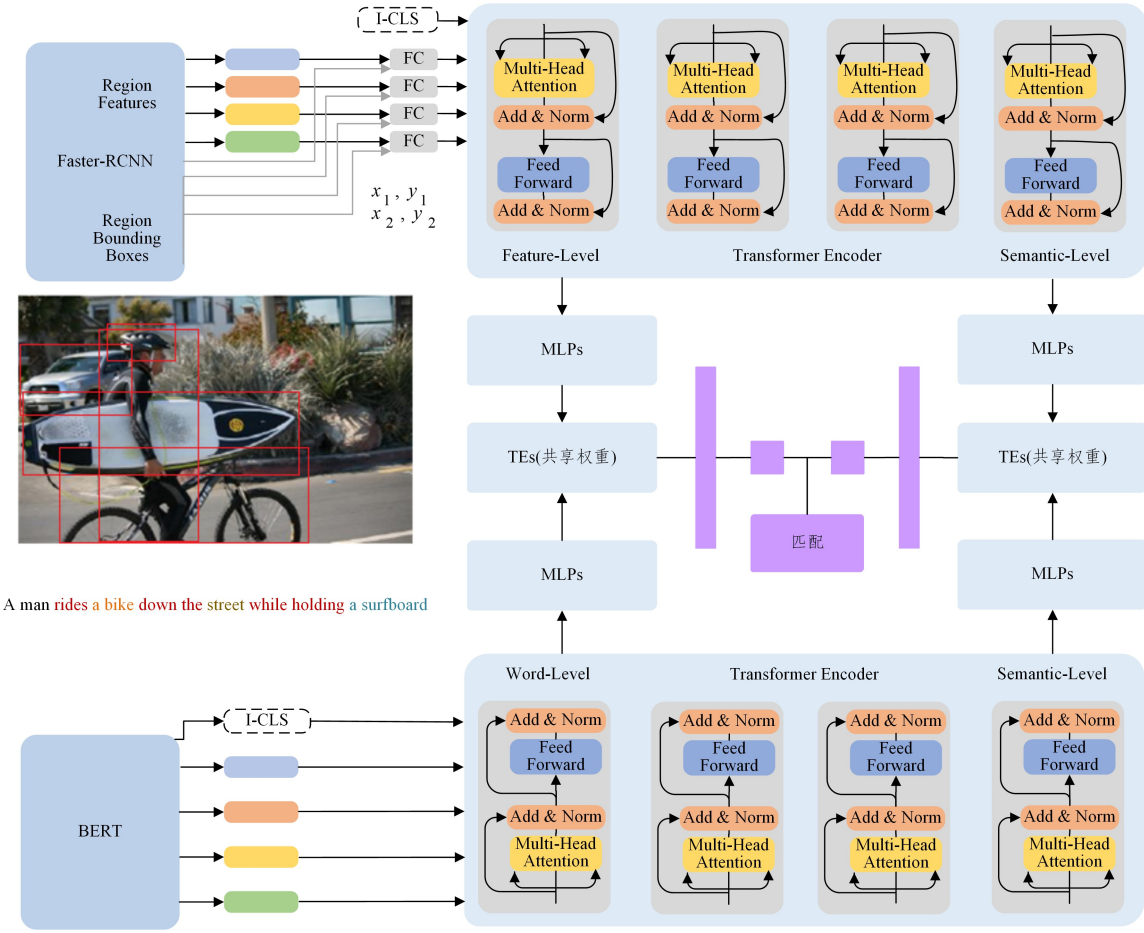


图 2 HAS-Net 网络结构

Fig. 2 HAS-Net structure

3.1 特征分层提取

Transformer 不同网络层的输出具有不同层次的特性，底层 Transformer 侧重于编码相对简单的基本语法信息，高层 Transformer 则侧重于编码相对复杂的高级语义信息。而现有基于 Transformer 的跨模态检索方法并没有充分利用这一特性。

3.1.1 图像编码

(1) Image Feature-level Feature：正如文献[29-30]所述，在基于 Transformer 的体系结构中，底层的特征捕获了描述基本语法信息的低级模式。我们在图像特征编码器的第一层获得了这些图像特征。文献[31]证明了非线性变换可以大幅度提高模型学习到的特征表示的质量，因此本文对图像特征进行非线性投影，采用文献[31]中的 MLPs 进行非线性变换，最终得到了图像的基本语法表示。

(2) Image Semantic-level Feature：基于 Transformer 的体系结构中的高层特性捕获了具有更复杂语义的高级表示。我们在图像特征编码器的最后一层获得了图像的高级语义特征表示，然后利用 MLPs 进行非线性变换，得到图像的高级语义表示。

3.1.2 文本编码

(1) Text Word-level Feature：与 Image Feature-level Feature 类似，我们从文本编码器的第一层获取文本的词级特征，然后利用 MLPs 进行非线性变换，得到了文本的词级特征表示。

(2) Text Semantic-level Feature：与 Image Semantic-level Feature 类似，我们从文本编码器的最后一层获取文本的语义特征，然后利用 MLPs 进行非线性变换，得到了文本的语义特征表示。

3.2 特征聚合模块

我们在图像区域和文本单词序列的开头引入了一个特殊的 token，其可以在 Transformer 编码层中传输全局信息，因此，我们将图像区域的数量扩展到了 $n+1$，将文本单词的

数量扩展到了 $m+1$。设置 I-CLS 为零向量，设置 T-CLS 为第一个单词提取到的特征。在 Transformer 编码层中，全局信息和局部信息根据自注意力机制更新得到。为了建立最后得出的全局信息和局部信息之间的关联，引入了自注意力机制的特征聚合方式，如图 3 所示。首先将全局特征和局部特征分别经过全连接层得到嵌入向量，随后将嵌入向量对位元素相乘，然后将得到的向量送入 FC 层，计算归一化后的权重分数，再对权重分数施以 Softmax 激活函数，最后点乘局部特征向量，得到加权后的向量，将加权后的向量相加得到最终聚合向量结果。Bs 代表 Batchsize 的大小，n 代表图像区域的数量或文本句子的长度。

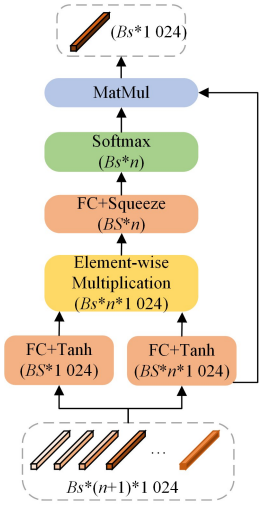


图 3 自注意力机制特征聚合图

Fig. 3 Self-attention mechanism feature aggregation

3.3 特征共享模块

跨模态检索面临的主要挑战是“异构鸿沟”。异构鸿沟指由于图像和文本的表示形式不一致，两者数据处于不同的分布空间，无法直接度量相似性，文本样本和图片样本之间有一条明显的分界线。要在这样的分布中只用内积距离来度量两个模态的语义相关性是十分困难且不准确的，因此我们通过共享多个 Transformer 编码层(TEs)，让图片特征和文本特征映射到公共的特征编码空间，得到高质量、高语义的跨模态表示。TEs 结构如图 4 所示。

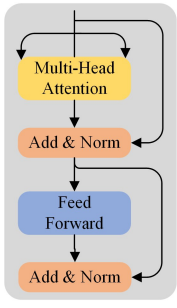


图 4 TEs 结构图

Fig. 4 TEs structure

3.4 Loss 损失函数

为了在相同的公共空间中匹配图像和文本，本文使用了 hinge-based triplet loss 损失函数，将注意力集中在难以匹配的样本上。使用的损失函数为：

$$L_\gamma(v,t)=[\gamma-S_r(v,t)+S_r(v,t^-)]_+ +[\gamma-S_r(v,t)+S_r(v^-,t)]_+ \quad (1)$$

其中，γ 是 margin 参数，在本实验中设置为 0.2，$[x]_+=\max(0,x)$；$S(\cdot)$ 代表联合嵌入空间的相似函数，在实验中采用内积作为相似函数；t^- 和 v^- 代表每个 Batchsize 中最难训练的样本。

由于我们提取了两个层级的特征，因此最终的 loss 函数为：

$$L=\alpha L_F+L_s \quad (2)$$

其中，L_F 代表底层特征匹配损失；L_s 代表高层特征匹配损失；α 为超参数，用于衡量 L_F 和 L_s 的重要占比，其取值范围为 0～1。

4 实验结果

我们使用 PyTorch 作为实验框架，实现本文设计的图文检索网络模型。选用 Tesla T4 显卡进行训练和测试，其具有 16 GB 显存，可高速进行数据计算和处理。表 1 列出了本实验的主要软硬件环境。

表 1 实验环境

Table 1 Experimental environment

项目名字	配置说明
CPU	Intel © Xeon E5-2678V3 CPU
GPU	NVIDIA Tesla T4 16 GB
内存	RECC4 2666MHz 32 GB
操作系统	Ubuntu 18.04 X86_64
实验框架	PyTorch 1.6.0 cuda 10.2

4.1 数据集及评价指标

4.1.1 数据集介绍

为了验证本文方法的有效性，我们在两个使用最广泛的基准数据集 MS-COCO[32] 和 Flickr30k[33] 上进行实验。这两个数据集都由多个图像文本对组成，每张图片对应 5 个描述性的句子。Flickr30k 数据集更多地关注人们日常活动中的事件以及场景中的人物、动物等，MS-COCO 图像集大致可划分为 3 个类别，分别为标志性的场景图像、标志性的对象图像以及非标志性的场景图像。MS-COCO 数据集包含 123 287 张图片，Flickr30k 数据集包含31 783 张图片。因为 MSCOCO 数据集的图像数量远远多于 Flickr30k，因此其被视为适用于检测模型泛化能力的数据集。

4.1.2 评价指标

文本到图像的匹配通常采用召回率 Recall 作为评价指标，其定义为对系统所有相关的文件中能被模型检索找回的文件数。本文为了更一步细分比较模型的检索能力，采用 Recall@1，Recall@5 和 Recall@10 作为评价指标。Recall@1 计算方法是首先选定任意一个文本作为输入，计算该文本与所有图像的相似度，对模型检索的图像结果进行排序，选择相似度最高的 Top1 检索图像作为模型的最终结果。使用召回率 Recall@1 评价指标具有重要意义，因为只选择 Top1 作为最终的结果进行评判，判断标准较为严苛，往往能更好地反应出模型的检索性能。类似地，对于 Recall@5 和 Recall@10 评价指标，对所有检索的图像进行相似度排序后，选择 Top5 和 Top10 的图像作为检索结果，然后计算这

些结果中正确图像所占的比率。具体地，Recall@k 的计算方法可表示为：

$$Recall@k=\frac{1}{N}\sum_{i=1}^{N}R(gt_i,k) \tag{3}$$

$$R(gt_i,k)=\begin{cases}1, & gt_i\in\{pred_1^i,pred_2^i,\cdots,pred_k^i\}\\ 0, & gt_i\notin\{pred_1^i,pred_2^i,\cdots,pred_k^i\}\end{cases} \tag{4}$$

4.2 实验参数

对于文本数据，我们使用 HuggingFace 预先训练的 BERT 模型，提取到的是 768 维的文本特征。对于图像数据，我们使用在 GitHub 上免费提供的 MS-COCO 数据集和 Flickr30k 数据集上已经保存好的自底向上的特征，所有的图像特征都为 2048 维度。在实验中，我们选用每幅图像置信度得分排名前 36 的特征。

在特征提取阶段，我们首先使用 4 个非共享权重的 Transformer 编码层进行图像与文本低级语法特征和高级语义特征的提取，随后通过 MLPs 非线性变换，最后经过两个 Transformer 共享权重层。所有的 Transformer 前馈层都是 2048 维的，并且衰减设置为 0.1。

受限于硬件资源，每个批次 Batchsize 的大小设置为 80，采用 Adaptive Moment Estimation(Adam)进行网络训练。训练过程中，权重衰减项(Weight Decay)设置为 0.1；初始学习率(Initial Learning Rate)设置为 0.000 002；模型的总迭代次数为 30，在迭代次数达到 15 时，将学习率设置为原来的 1/10。

4.3 参数实验

4.3.1 α 参数实验

参数 α 越大，代表低级语法信息在图文检索的过程中影响越大。对于 $\alpha\in\{0.01,0.03,0.1,0.3,1,3,10\}$，在 MS-COCO 数据集和 Flickr30k 数据集上进行实验，Rank1 随着参数 α 的变化结果如表 2 所列。

表 2 Rank1-α 变化

Table 2 Variation of Rank1-α

(单位：%)

α	MS-COCO		Flickr30k	
	Text retrieval	Image retrieval	Text retrieval	Image retrieval
0.01	67.6	55.6	63.1	47.9
0.03	68.2	55.9	63.5	48.6
0.10	68.9	56.1	64.2	48.8
0.30	69.0	56.1	**64.8**	**49.1**
1.00	**69.6**	**56.3**	64.5	48.9
3.00	69.1	56.1	63.4	48.3
10.00	68.6	56.0	62.8	47.6

从表 2 可以看出，随着 α 取值变大，*Rank*@1 有一定的提升，这说明引入低级语义信息有助于提升文本到图像匹配的性能。但是当 α 达到一定值后，*Rank*@1 开始下降，表明低级语义信息比例过大会影响整体特征的表达。因此，本文中，对于 MS-COCO 数据集，我们选定 *Rank*@1 达到最高时 α 的取值为 1，对于 Flickr30k 数据集，我们选定 *Rank*@1 达到最高时 α 的取值为 0.3。

4.3.2 Transformer 分层实验

为了验证低层特征在图文检索过程中的有效性，我们设计了一组对比实验，一组只选用高层特征做匹配，另一组选用高层特征和底层特征，高层特征和底层特征在检索过程中的占比相同。最终得出的结果如表 3、表 4 所列。

表 3 MS-COCO 数据集 Transformer 分层实验结果

Table 3 Transformer hierarchical experiment results on MS-COCO

(单位：%)

Transformer 特征提取	Text retrieval			Image retrieval		
	R@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10
不分层	63.7	90.5	96.2	51.9	85.6	93.6
分层	69.6	93.0	97.5	56.3	87.4	94.1

从表 3 可以看出，对于 MS-COCO 数据集，在图文跨模态检索过程中只选用高层特征，文本检索 *Rank*@1 为 63.7%，图像检索 *Rank*@1 为 51.9%；而同时选用底层特征和高层特征，文本检索 *Rank*@1 提升了 5.9%，图像检索 *Rank*@1 提升了 4.4%。

表 4 Flickr30k 数据集 Transformer 分层实验结果

Table 4 Transformer hierarchical experiment results on Flickr30k

(单位：%)

Transformer 特征提取	Text retrieval			Imagere trieval		
	R@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10
不分层	53.2	79.4	86.0	41.1	71.9	81.2
分层	64.8	88.3	92.5	49.1	77.6	86.2

从表 4 可以看出，对于 Flickr30k 数据集，在图文跨模态检索过程中只选用高层特征，文本检索 *Rank*@1 为 53.2%，图像检索 *Rank*@1 为 41.1；而同时选用底层特征和高层特征，文本检索 *Rank*@1 提升了 11.6%，图像检索 *Rank*@1 提升了 8%。由此证明了底层特征在文本到图像的匹配过程中起到了一定的作用，从而证明了本文方法的有效性。

4.3.3 聚合方式对比实验

当我们在 transformer 编码层共享权重层之后，得到了全局特征和局部特征。为了建立全局特征和局部特征之间的关联，最简单的方法是将所有特征求和或平均，或者舍弃局部特征，或者使用循环神经网络进行特征聚合。而本文的方法是引入自注意力机制，建立局部特征和全局特征之间的关系，实现特征聚合，并通过实验证明了自注意机制聚合方式的有效性，实验结果如表 5、表 6 所列。

表 5 MS-COCO 数据集特征聚合方法对比实验

Table 5 Comparison of feature aggregation methods on MS-COCO

(单位：%)

聚合方式	Text retrieval			Image retrieval		
	R@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10
Sum	68.6	92.4	97.2	56.0	87.3	94.1
First	67.5	92.4	97.1	54.8	86.7	93.7
Gated	67.3	92.6	97.2	55.5	87.3	93.9
GRU	65.7	92.1	97.1	53.5	86.4	93.7
self-Attention	**69.6**	**93.0**	**97.5**	**56.3**	**87.4**	**94.1**

由表 5 可以看出，在 MS-COCO 数据集上，使用求和的特征聚合方法，文本检索 *Rank*@1 为 68.6%，图像检索 *Rank*@1 为 56.0%；只选用全局特征的方式，文本检索 *Rank*@1 为 67.5%，图像检索 *Rank*@1 为 54.8%；使用门函数的特征聚合方式，文本检索 *Rank*@1 为 67.3%，图像检索 *Rank*@1 为 55.5%；使用循环神经网络的特征聚合方式，文本检索

Rank@1 为 65.7%，图像检索 *Rank*@1 为 53.5%。而使用自注意力机制，文本检索 *Rank*@1 达到了 69.6%，图像检索 *Rank*@1 达到了 56.3%。

表 6 Flickr30k 数据集特征聚合方法对比实验

Table 6 Comparison of feature aggregation methods on Flickr30k

（单位：%）

聚合方式	Text retrieval			Image retrieval		
	R@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10
Sum	61.8	86.6	91.8	48.5	77.2	85.2
First	62.1	86.5	91.7	46.9	77.2	85.5
Gated	59.7	84.0	90.8	47.5	75.9	84.1
GRU	57.7	82.7	89.9	44.8	74.1	83.0
self-Attention	**64.8**	**88.3**	**92.5**	**49.1**	**77.6**	**86.2**

由表 6 可以看出，在 Flickr30k 数据集上，使用求和的特征聚合方法，文本检索 *Rank*@1 为 61.8%，图像检索 *Rank*@1 为 48.5%；只选用全局特征的方式，文本检索 *Rank*@1 为 62.1%，图像检索 *Rank*@1 为 46.9%；使用门函数的特征聚合方式，文本检索 *Rank*@1 为 59.7%，图像检索 *Rank*@1 为 47.5%；使用循环神经网络的特征聚合方式，文本检索 *Rank*@1 为 57.7%，图像检索 *Rank*@1 为 44.8%。而使用自注意力机制，文本检索 *Rank*@1 达到了 64.8%，图像检索 *Rank*@1 达到了 49.1%。相比于以往的特征聚合方式，本文方法的效果得到了提升，证明了自注意力机制特征聚合方式在跨模态检索过程中是有效的。

4.3.4 共享权重实验

为了验证共享权重在图文跨模态检索过程中的有效性，我们设计了一组对比实验，一组选用共享权重，另外一组则不选用，以此来验证共享权重将跨模态特征表示映射到同一编码空间是否有效。最终得出的结果如表 7、表 8 所列。从表 7 可以看出，在 MS-COCO 数据集上，在图文跨模态检索过程中不选用共享权重的方法，文本检索 *Rank*@1 为 62.7%，图像检索 *Rank*@1 为 49.7%；而选用共享权重的方法，文本检索 *Rank*@1 提升了 6.9%，图像检索 *Rank*@1 提升了 6.6%。

表 7 MS-COCO 数据集共享权重实验结果

Table 7 Shared weight experiment results on MS-COCO

（单位：%）

方法	Text retrieval			Image retrieval		
	R@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10
不共享	62.7	90.3	95.8	49.7	83.9	92.6
共享	69.6	93.0	97.5	56.3	87.4	94.1

表 8 Flickr30k 数据集共享权重实验结果

Table 8 Shared weight experiment results on Flickr30k

（单位：%）

方法	Text retrieval			Image retrieval		
	R@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10
不共享	57.3	82.4	89.7	43.0	73.7	82.8
共享	64.8	88.3	92.5	49.1	77.6	86.2

从表 8 可以看出，在 Flickr30k 数据集上，在图文跨模态检索过程中不选用共享权重的方法，文本检索 *Rank*@1 为 57.3%，图像检索 *Rank*@1 为 43.0%；而选用共享权重的方法，文本检索 *Rank*@1 提升了 7.5%，图像检索 *Rank*@1 提升了 6.1%。由此证明了共享权重将跨模态特征表示映射到共同编码空间的方法是有效的。

4.4 与其他实验对比

在 MS-COCO 数据集和 Flickr30k 数据集上，将所提模型与图像检索任务的代表性技术进行比较，结果如表 9、表 10 所列。

表 9 在 MS-COCO 数据集上与其他方法的实验结果对比

Table 9 Experimental results comparison of the proposed method and other methods on MS-COCO dataset

（单位：%）

类型	模型	1K test						5K test					
		Text retrieval			Image retrieval			Text retrieval			Image retrieval		
		R@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10
①	VSE++[7]	64.6	90.0	95.7	52.0	84.3	92.0	41.3	71.1	81.2	30.3	59.4	72.4
	CAMP[34]	72.3	94.8	98.3	58.5	87.9	95.0	50.1	82.1	89.7	39.0	68.9	80.4
	SCAN[35]	72.7	94.8	98.4	58.8	88.4	94.8	50.4	82.2	90.0	38.6	69.3	80.4
	IMRAM[11]	76.7	95.6	98.5	61.7	89.1	95.0	53.7	—	91.0	39.7	—	79.8
	VSRN[5]	76.2	94.8	98.2	62.8	89.7	95.1	53.0	81.1	89.4	40.5	70.6	81.1
②	MMCA[26]	74.8	95.6	97.7	61.6	89.8	95.2	54.0	82.5	90.7	38.7	69.7	80.8
	TERAN[25]	80.2	96.6	99.0	67.0	92.2	96.9	59.3	85.8	92.4	45.1	74.6	84.4
③	TERN[24]	63.7	90.5	96.2	51.9	85.6	93.6	38.4	69.5	81.3	28.7	59.7	72.7
	Ours(HAS)	**69.6**	**93.0**	**97.5**	**56.3**	**87.4**	**94.1**	**43.5**	**75.4**	**85.5**	**33.0**	**64.2**	**76.5**

方法①中的 VSE++，SCAN，CAMP，IMRAM 和 VSRN 都是用卷积神经网络提取图像特征，使用 GRU 提取文本特征之后对图像特征和文本特征进行注意力交互；方法②中的 MMCA 和 TERAN 利用 Transformer 提取图像文本特征后进行注意力交互。这两种算法类型虽然精度高，但是需要大量的时间进行注意力交互，在实际应用中不可行；方法③中的 TERN 是采用 Transformer 进行图像和文本的特征提取，没有进行特征交互，因此效果较差，但是匹配速度更快；本文提出的 HAS 算法模型采用 Transformer 进行图像和文本的特征提取，且没有进行跨模态特征的交互，和 TERN 算法属于同一种类型。

TERN 模型在 MS-COCO 1K 测试集上文本检索 *Rank*@1 为 63.7%，图像检索 *Rank*@1 为 51.9%；在 MS-COCO 5K 测试集上文本检索 *Rank*@1 为 38.4%，图像检索 *Rank*@1 为 28.7%；在 Flickr30k 测试集上文本检索 *Rank*@1 为 53.2%，图像检索 *Rank*@1 为 41.1%。而本文提出的模型 HAS 在 MS-COCO 1K 测试集上文本检索 *Rank*@1 为 69.6%，图像检索 *Rank*@1 为 56.3%；在 MS-COCO 5K 测试集上文本

检索 *Rank*@1 为 43.5%，图像检索 *Rank*@1 为 33.0%；在 Flickr30k 测试集上文本检索 *Rank*@1 为 64.8%，图像检索 *Rank*@1 为 49.1%。本文模型效果有所提升的原因在于我们不仅使用了高层的语义信息，也考虑了底层的语法信息，并且利用自注意力机制挖掘局部特征和全局特征的关联，通过共享权重将特征映射到相同的编码空间。

表 10 在 Flickr30k 数据集上与其他方法的实验结果对比

Table 10 Experimental results comparison of the proposed method and other methods on Flickr30k

（单位：%）

类型	模型	Text retrieval			Image retrieval		
		R@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10
①	VSE++[7]	52.9	80.5	87.2	39.6	70.1	79.5
	SCAN[35]	67.4	90.3	95.8	48.6	77.7	85.2
	CAMP[34]	68.1	89.7	95.2	51.5	77.1	85.3
	VSRN[5]	71.3	90.6	96.0	54.7	81.8	88.2
	IMRAM[11]	74.1	93.0	96.6	53.9	79.4	87.2
②	MMCA[26]	74.2	92.8	96.4	54.8	81.4	87.8
	TERAN[25]	79.2	94.4	96.8	63.1	87.3	92.6
③	TERN[24]	53.2	79.4	86.0	41.1	71.9	81.2
	Ours(HAS)	**64.8**	**88.3**	**92.5**	**49.1**	**77.6**	**86.2**

结束语 本文设计了一种新颖的基于 Transformer 的文本图像匹配网络结构（HAS-Net），设计了一种分层 transformer 编码结构，以更好地利用底层的语法信息和高层的语义信息；通过共享 Transformer 编码层，让图片特征和文本特征映射到公共的特征编码空间；改进了传统的全局特征聚合方式，利用自注意力机制设计一种新的特征聚合方式。最后在 MS-COCO 数据集和 Flickr30k 数据集上进行实验，结果表明，所提方法在同类型算法中性能均得到提升。下一步可以考虑进行跨模态特征交互，牺牲检索速度来提升检索精度。

参 考 文 献

[1] HAO Y，DONG L，WEI F，et al. Visualizing and Understanding the Effectiveness of BERT[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong：Association for Computational Linguistics，2019：4141-4150.

[2] TENNEY I，DAS D，PAVLICK E. BERT Rediscovers the Classical NLP Pipeline[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence：Association for Computational Linguistics，2019：4593-4601.

[3] GABEUR V，SUN C，ALAHARI K，et al. Multi-modal transformer for video retrieval[C]// Proceedings of the 16th European Conference Computer Vision（ECCV）. Glasgow：Springer，2020：214-229.

[4] PATRICK M，HUANG P，ASANO Y，et al. Support-set bottlenecks for video-text representation learning[C]// Proceedings of the 9th International Conference on Learning Representations（ICLR）. Austria：OpenReview，2021：1-18.

[5] LI K，ZHANG Y，LI K，et al. Visual Semantic Reasoning for Image-Text Matching[C]// 2019 IEEE International Conference on Computer Vision（ICCV）. 2019：4653-4661.

[6] EISENSCHTAT A，WOLF L. Linking Image and Text with 2-Way Nets[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition（CVPR）. 2017：1855-1865.

[7] FAGHRI F，FLEET D J，KIROS J R，et al. VSE++：Improving Visual-Semantic Embeddings with Hard Negatives[C]// British Machine Vision Conference（BMVC）. 2018：12-21.

[8] 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（CVPR）. 2018：7181-7189.

[9] HUANG Y，WANG W，WANG L. Instance-aware Image and Sentence Matching with Selective Multimodal LSTM[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition（CVPR）. 2017：7251-7262.

[10] REN S，HE K，GIRSHICK R，et al. Faster R-CNN：Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2017，39(6)：1137-1149.

[11] CHEN H，DING G，LIU X，et al. IMRAM：Iterative Matching With Recurrent Attention Memory for Cross-Modal Image-Text Retrieval[C]// Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Seattle：IEEE，2020：12652-12660.

[12] WANG Y X，YANG H，QIAN X M，et al. Position Focused Attention Network for Image-Text Matching[C]// Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence（IJCAI）. Macao：AAAI，2019：3792-3798.

[13] JI Z，WANG H，HAN J，et al. SMAN：Stacked Multimodal Attention Network for Cross-Modal Image-Text Retrieval[J]. IEEE Transactions on Cybernetics，2020(99)：1-12.

[14] XU X，WANG T，YANG Y，et al. Cross-Modal Attention with Semantic Consistence for Image-Text Matching[J]. IEEE Transactions on Neural Networks and Learning Systems，2020(99)：1-14.

[15] ASHISH V，NOAM S，NIKI P，et al. Attention is all you need[J]. Advances in Neural Information Processing Systems，2017(1)：5998-6008.

[16] DEVLIN J，CHANG M W，LEE K，et al. BERT：Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis：Association for Computational Linguistics，2019：4171-4186.

[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. Seattle：ACM，2020：1047-1055.

[18] WEI X，ZHANG T，LI Y，et al. Multi-Modality Cross Attention Network for Image and Sentence Matching[C]// Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Seattle：IEEE，2020：10938-10947.

[19] LU J，BATRA D，PARIKH D，et al. ViLBERT：Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-

Language Tasks[C]// Proceedings of International Conference on Neural Information Processing Systems. Vabcouver: IEEE, 2019: 13-23.

[20] SU W, ZHU X, CAO Y, et al. VL-BERT: Pre-training of Generic Visual-Linguistic Representations[C]// International Conference on Learning Representations(ICLR). 2020.

[21] PARMAR N, VASWANI A, USZKOREIT J, et al. Image Transformer[J]. International Conference on Machine Learning, 2018(80): 4052-4061.

[22] CORDONNIER J, LOUKAS A, JAGGI M. On the Relationship between Self-Attention and Convolutional Layers[C]// International Conference on Learning Representations(ICLR). 2020.

[23] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[C]// International Conference on Learning Representations(ICLR). 2021.

[24] MESSINA N, FALCHI F, ESULI A, et al. Transformer Reasoning Network for Image-Text Matching and Retrieval[C]// International Conference on Learning Representations (ICLR). 2020: 5222-5229.

[25] MESSINA N, AMATO G, ESULI A, et al. Fine-grained Visual Textual Alignment for Cross-Modal Retrieval using Transformer Encoders[C]// CoRR. 2020.

[26] WEI X, ZHANG T, LI Y, et al. Multi-Modality Cross Attention Network for Image and Sentence Matching[C]// Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Seattle: IEEE, 2020: 10938-10947.

[27] LI G, DUAN N, FANG Y, et al. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training [J]. AAAI Conference on Artificial Intelligence. 2020: 11336-11344.

[28] CAO L, QIAN S, ZHANG H, et al. Global Relation-Aware Attention Network for Image-Text Retrieval[C]// Proceedings of International Conference on Multimedia Retrieval. Taiwan: ACM, 2021: 19-28.

[29] PETERS M, NEUMANN M, ZETTLEMOYER L, et al. Dissecting Contextual Word Embeddings: Architecture and Representation[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018: 1499-1509.

[30] VIG J. A Multiscale Visualization of Attention in the Transformer Model[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Florence: Association for Computational Linguistics, 2019: 37-42.

[31] CHEN T, KORNBLITH S, NOROUZI M, et al. A Simple Framework for Contrastive Learning of Visual Representations [C]// International Conference on Machine Learning(ICML). 2020: 1597-1607.

[32] LIN T, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[J]. European Conference Computer Vision(ECCV), 2014, 8693: 740-755.

[33] YOUNG P, LAI A, HODOSH M, et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions[J]. Transactions of the Association for Computational Linguistics, 2014, 2: 67-78.

[34] LEE K H, XI C, GANG H, et al. Stacked Cross Attention for Image-Text Matching[C]// 15th European Conference Computer Vision(ECCV). 2018: 212-228.

[35] WANG Z, LIU X, LI H, et al. CAMP: Cross-Modal Adaptive Message Passing for Text-Image Retrieval[C]// 2019 IEEE International Conference on Computer Vision(ICCV). 2019: 5763-5772.

YANG Xiaoyu, born in 1996, postgraduate. His main research interests include deep learning, computer vision and cross-modal retrieval.

YIN Guangqiang, born in 1982, master, professor. His main research interests include network security, computer vision, signal processing and intelligent manufacturing.

(责任编辑:何杨)