第 39 卷 第 4 期
2024 年 8 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 39 No. 4
Aug. 2024

DOI:10.13364/j.issn.1672-6510.20230177
数字出版日期：2024–06–21；数字出版网址：http://link.cnki.net/urlid/12.1355.n.20240620.1627.011

融合语义增强和位置编码的图文匹配方法

赵婷婷，常玉广，郭 宇，陈亚瑞，王 媛
（天津科技大学人工智能学院，天津 300457）

摘 要：图文匹配是跨模态基础任务之一，其核心是如何准确评估图像语义与文本语义之间的相似度。现有方法是通过引入相关阈值，最大限度地区分相关和无关分布，以获得更好的语义对齐。然而，对于特征本身，其语义之间缺乏相互关联，且对于缺乏空间位置信息的图像区域与文本单词很难准确对齐，从而不可避免地限制了相关阈值的学习导致语义无法准确对齐。针对此问题，本文提出一种融合语义增强和位置编码的自适应相关性可学习注意力的图文匹配方法。首先，在初步提取特征的基础上构造图像（文本）无向全连通图，使用图注意力去聚合邻居的信息，获得语义增强的特征。然后，对图像区域的绝对位置信息编码，在具备了空间语义的图像区域与文本单词相似性的基础上获得最大程度区分的相关和无关分布，更好地学习两个分布之间的最优相关边界。最后，通过公开数据集 Flickr 30 k 和 MS-COCO，利用 Recall@K 指标对比实验，验证本文方法的有效性。

关键词：跨模态图文匹配；图注意力；位置编码；相关性阈值
中图分类号：TP391.4 **文献标志码**：A **文章编号**：1672-6510（2024）04-0063-10

Image-Text Matching Method Combining Semantic Enhancement and Position Encoding

ZHAO Tingting，CHANG Yuguang，GUO Yu，CHEN Yarui，WANG Yuan
（College of Artificial Intelligence，Tianjin University of Science and Technology，Tianjin 300457，China）

Abstract：Image-text matching is one of the basic cross-modal tasks. Its core is how to accurately evaluate the similarity between image semantics and text semantics. Existing methods maximize the distinction between relevant and irrelevant distributions by introducing a correlation threshold to obtain better semantic alignment. However，for the features themselves，there is a lack of correlation between their semantics，and it is difficult to accurately align image areas and text words that lack spatial location information，which inevitably limits the learning of relevant thresholds and results in the inability to accurately align semantics. To address this problem，in this article we propose an image-text matching method that combines semantic enhancement and positional coding with adaptive correlation learnable attention. Specifically，an undirected fully connected graph of images（texts）is first constructed based on preliminary feature extraction，and graph attention is used to aggregate neighbor information to obtain semantically enhanced features. Then，the absolute position information of the image area is encoded，and the most differentiated relevant and irrelevant distributions are obtained based on the similarity between the image area and the text words with spatial semantics，so as to better learn the optimal correlation between the two distributions. boundary. Finally，through the public datasets Flickr 30 k and MS-COCO，the effectiveness of the method proposed in this article was verified with the use of the Recall@K indicator comparison experiment.

Key words：cross-modal image-text matching；graph attention；position encoding；relevance threshold

引文格式：
赵婷婷，常玉广，郭宇，等. 融合语义增强和位置编码的图文匹配方法[J]. 天津科技大学学报，2024，39（4）：63–72.
ZHAO T T，CHANG Y G，GUO Y，et al. Image-text matching method combining semantic enhancement and position encod-

收稿日期：2023–09–25；**修回日期**：2024–02–18
基金项目：国家自然科学基金项目（61976156）；天津市企业科技特派员项目（20YDTPJC00560）
作者简介：赵婷婷（1986—），女，内蒙古赤峰人，副教授，tingting@tust.edu.cn

ing[J]. Journal of Tianjin university of science and technology, 2024, 39(4): 63-72.

视觉和文本是描述现实场景最常用的信息表达方式，将这两种最常见的模态联系起来，对于人工智能理解现实世界至关重要。图像–文本匹配是在这两种异构模态之间建立联系的桥梁，是跨模态领域的基础研究之一，被广泛应用于视觉问答[1-2]和图像字幕[3]。这种匹配任务旨在通过给定的文本描述搜索图像，或者通过图像查询找到相应的文本。尽管此项工作已取得了一定的进展，但图像–文本匹配仍然面临挑战，即如何准确学习语义对齐，以找到模态之间的相关共享语义以衡量相似性。

现有图文匹配方法可以分为局部匹配和全局匹配两类。全局匹配侧重于学习模态之间的全局对齐，即将整个图像和文本映射到一个共享的嵌入空间中。该领域侧重于网络模型的设计[4-6]以及不同的优化策略[7-10]。局部匹配则着重于学习局部片段的对齐，进而推断图像和文本的相似性，即在视觉区域和文本单词之间进行匹配[11]。早期工作[12-13]探索了局部对齐，但没有考虑到每个片段的不同重要性，然而每个片段在图像和文本中都可能传达不同的重要信息。如果只关注局部对齐而忽略不同片段的重要性，可能会丢失关键的信息，导致匹配模型无法准确地捕捉到图像和文本之间的语义关系。最近，基于注意力的匹配通过差异化关注特定片段，发现了所有细粒度的单词–区域对齐，取得了更优性能，成为图像–文本匹配的主流方法。其中，堆叠交叉注意力[14]（stacked cross attention，SCAN）以及其各种改进的变体是此类方法的代表之一。基于注意力的匹配方法包含两个关键步骤：首先，在联合嵌入空间中学习片段（区域和单词）特征；随后，利用注意力发现单词–区域片段之间所有语义对齐，并基于另一种模态的查询片段找到这一模态与之相关的片段特征，即找到模态之间的共享语义度量图像–文本相似度。

然而，上述方法学习到的片段特征之间缺乏语义关联。针对同一模态，其他片段特征往往对当前片段特征具有积极或消极的影响，而缺乏语义关联就导致了单词–区域对齐时存在偏差，影响最终的匹配效果。在上述注意力对齐过程中，为了减少无关片段对匹配性能的影响，现有方法通常将直观的零设为隐含的相关性阈值。具体而言，现有的注意力方法通过 ReLU 操作（将负值设为零并保持正值）缩小得分低于零的片段的注意力，并将几乎全部注意力分配给得分大于零的片段。相关性阈值始终被固定为零，因此注意力无法准确区分不相关片段，从而干扰了共享语义。Zhang 等[15]提出了一种全新的自适应相关性可学习注意力机制，第一次将相关性阈值和特征学习整合到一个统一的框架中。图像文本匹配实例如图 1 所示，对于图像描述高度相似的情况，现有方法很难正确匹配，图像区域语义高度重合，仅仅通过语义信息将图像区域与文本单词对齐存在性能瓶颈。

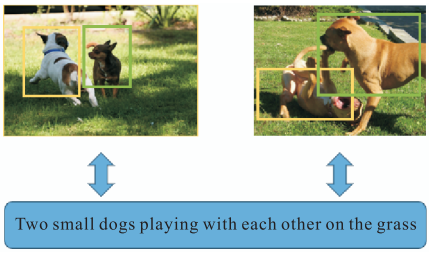

图 1　图像文本匹配实例

Fig. 1　Example of image-text matching

针对上述问题，本文提出一种融合语义增强和位置编码的自适应相关性可学习注意力机制。首先，将片段特征视为节点并构建图像区域和文本单词的全连通图，通过引入图注意力机制学习节点之间的语义信息，增强片段特征的语义表达能力。然后，对图像区域的位置信息进行编码，得到其空间特征。在得到了空间位置关系的基础上，通过最小化相关性区分的错误概率自适应学习最优阈值，进一步优化相关和不相关单词–区域片段特征的相似度分布。阈值将以更高的可区分性改善特征学习，促进两个分布更好地分离，从而学习更好的语义对齐测量图像–文本相似性。最后，在公开数据集 Flickr 30 k[16]和 MS-COCO[17]上进行实验验证。

1　相关工作

1.1　图像–文本匹配

图像–文本匹配已经得到了显著的发展，根据匹配方式可将其分为两类方法：全局匹配，其倾向于学习全局对准，即将所述图像或文本表示为整体特征以测量相似性；局部匹配，其集中于局部片段之间的细粒度对准，即通过所有单词–区域对的相关性推断整体图像–文本相似性，本文属于后者。

全局匹配学习方法使用预训练的深度神经网络

提取图像和文本的全局特征，并学习在视觉和文本特征之间建立共同的嵌入空间。Frome 等[18]提出了一个视觉–语义嵌入模型，通过卷积神经网络（CNN）提取图像的视觉特征，并通过 Skip-Gram 方法提取文本的语义特征。近年来，大多数方法采用各种循环神经网络（RNN）架构捕捉语言的长距离的上下文信息。Kiro 等[19]使用 LSTM 提取文本的全局语义表示。Faghri 等[9]使用 CNN 和 GRU 分别提取图像和文本的全局特征，并设计了一种新的目标函数，通过硬负样本挖掘进一步提高匹配准确性。然而，主要对象在图像–文本对的全局表示中起主导作用，而次要对象大多被忽略，因此全局匹配学习方法不能准确地学习图像和文本的对应关系。

基于局部匹配的方法通过提取局部模式对齐图像区域和文本单词。Karpathy 等[12]使用预训练的 R-CNN 在图像中检测视觉区域，然后学习句子中的单词与图像中的区域之间的相似性。受到自下而上的注意力机制[20]的启发，Lee 等[14]提出了一种堆叠的交叉注意力模型，用于聚合图像区域和文本单词之间的局部相似性匹配结果。受此工作启发，研究者提出了诸多基于注意力的方法：Chen 等[21]采用多步对齐的迭代交叉注意力匹配；也有研究人员[22]通过循环神经网络对图像和文本进行对称处理，以捕捉图像对象的上下文信息；Yang 等[23]提出使用图神经网络探索准确的语义对齐；Liu 等[24]提出通过设计的标识规则仅关注一部分区域或单词。阈值有利于提高匹配性能，通常采用零阈值减少无关片段的干扰[14]。此外，Liu 等[24]进一步证明了消除无关片段对于语义对齐是必要的。Zhang 等[15]通过设计一种自适应相关区分注意力机制学习阈值，更好地区分无关片段。然而，仅仅利用图像区域的语义特征去学习阈值依然很难将非常相似的图片与文本准确对齐。因此，本文提出一种基于位置编码的自适应相关性可学习注意力机制，空间信息的引入进一步促进单词–区域片段特征相似度分布中相关分布与无关分布的分离，以进一步实现更好的语义对齐学习。

1.2　图形表示学习

图形表示在视觉–语言任务中被广泛用于建模实体之间的关系，包括图像字幕生成[25]、视觉问答[26]和视觉常识推理[27]。在图像–文本匹配方法中，相关工作通过使用图结构增强视觉和文本内容的单体表示，并取得了良好的性能。Li 等[28]提出了一种视觉语义推理方法，学习图像中对象区域之间的关系以生成增强的视觉表示。Liu 等[29]提出了一个图匹配网络，用于在图像和文本之间进行节点级和结构级的匹配。Wang 等[30]构建了场景图表示图像和文本，并在场景图上执行图像和文本之间的对象级和关系级匹配。然而，以上方法都是通过图卷积聚合邻居节点特征，但是不同邻居节点的语义信息对中心节点的重要性不同，图卷积共享权重导致无法区分邻居节点的重要性。因此，本文提出在模型训练的过程中，自适应地学习节点之间的关系，通过引入图注意力机制[31]学习邻居的权重以实现更好的邻居聚合，使增强后的节点特征学习到更丰富的语义，从而可以进一步促进细粒度级的单词–区域语义对齐。

2　融合语义增强和位置编码的图文匹配方法

融合语义增强和位置编码的图文匹配方法由两个模块组成：语义增强模块和基于位置编码的自适应相关性可学习模块，整体网络结构框架如图 2 所示。

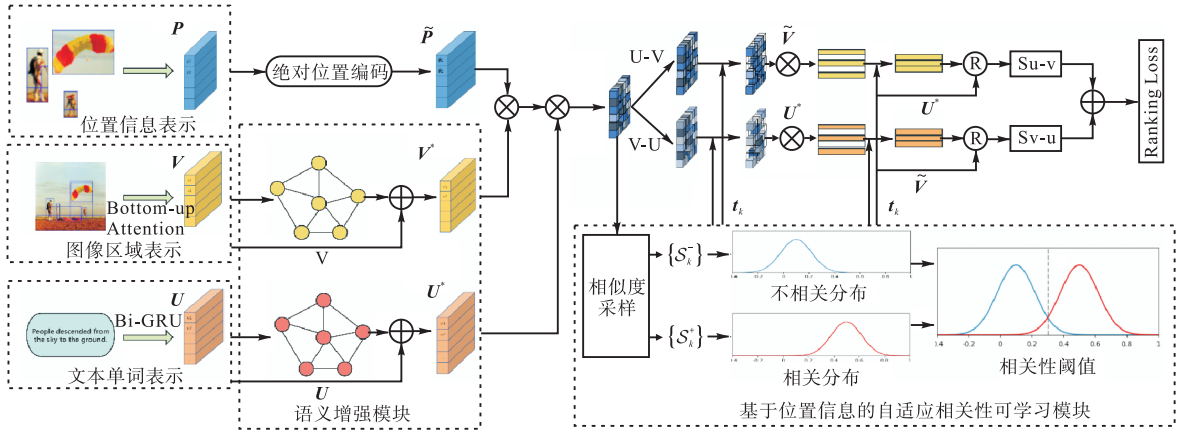

图 2　总体框架图

Fig. 2　Overall framework

语义增强模块分别构造图像区域的全连通图和文本单词的全连通图，通过图注意力机制获得聚合邻居语义的增强特征。基于位置编码的自适应相关性可学习模块通过将图像区域的位置信息编码，这里位置编码采用的是绝对位置编码，相较于相对编码可以从全局角度捕获空间特征。融合了空间特征的图像区域特征具有更强的表达能力，对于单词–区域片段特征相似度分布更精准，有利于相关性阈值的学习。

2.1　特征提取

图像区域特征：给定一幅图像，通过利用自下而上的注意力优势，将其表示为一组显著区域特征 $[v_1,v_2,\cdots,v_n]$，n 表示显著区域个数。显著对象和其他区域是通过在 Visual Genome[32]上预训练的 Faster-RCNN 进行检测的，选择了前 n 个（$n=36$）提议框。通过预训练的 ResNet-101[33]对检测到的区域进行编码，再经过一个平均池化层得到特征向量 $\boldsymbol{x}_j$。接下来，使用一个全连接层将每个区域映射到一个 1 024 维的特征向量 $\boldsymbol{v}_j$，$j\in[1,n]$。表示为

$$\boldsymbol{v}_j=W_v\boldsymbol{x}_j+\boldsymbol{b}_v \tag{1}$$

其中：W_v 是全连接层的参数，$\boldsymbol{b}_v$ 是偏置向量。

文本单词特征：给定一段包含 m 个单词的文本，将其特征表示为 $[u_1,u_2,\cdots,u_m]$，其中每个单词与一个特征向量相关联。首先将文本中的单词表示为 one-hot 编码，然后通过嵌入矩阵将单词编码为 300 维向量 $\boldsymbol{y}_i$，接着使用双向门控循环单元（BiGRU）以整合前向和后向上下文信息，得到 1 024 维的单词特征 u_i。

2.2　语义增强模块

图像全连通图：为了增强图像区域的语义表达能力，构建一个无向全连通图 $G_v=(V_v,E_v)$ 学习图像区域之间的语义关联，其中 V_v 表示图像区域集合，E_v 表示图像区域之间的关系集合。在实验过程中，对图像区域特征执行自注意力，通过图像区域之间的语义相似性初始化边。$\boldsymbol{A}^v\in\mathbb{R}^{n\times n}$ 表示邻接矩阵，n 是图像区域的个数。$\boldsymbol{A}^v$ 所有元素初始化为 1，并对每一行进行归一化。$\boldsymbol{S}^v\in\mathbb{R}^{n\times n}$ 是任意区域之间的相似性矩阵，将 $\boldsymbol{A}^v$ 与 $\boldsymbol{S}^v$ 逐元素相乘就得到了边权矩阵 $\boldsymbol{W}_e^v$，即

$$\boldsymbol{W}_e^v=\boldsymbol{S}^v\odot\boldsymbol{A}^v \tag{2}$$

$$s_{ij}^v=\frac{\exp\left(\mathrm{LeakyReLU}\left(\left(\boldsymbol{W}_q^v\boldsymbol{v}_i\right)^{\mathrm{T}}\left(\boldsymbol{W}_k^v\boldsymbol{v}_j\right)\right)\right)}{\sum_{j=0}^{n}\exp\left(\mathrm{LeakyReLU}\left(\left(\boldsymbol{W}_q^v\boldsymbol{v}_i\right)^{\mathrm{T}}\left(\boldsymbol{W}_k^v\boldsymbol{v}_j\right)\right)\right)} \tag{3}$$

其中：$\odot$ 表示逐元素乘法，$\boldsymbol{W}_q^v$ 和 $\boldsymbol{W}_k^v$ 是可学习的参数。

将图像区域特征与边权相乘，并经过一个激活函数得到语义增强后的特征，之后引入残差结构，目的是保留图像区域特征自身的信息加速模型训练，最后得到最终的增强特征 $\boldsymbol{V}^*$，即

$$\boldsymbol{V}^*=\sigma\left(\boldsymbol{W}_e^vW^v\boldsymbol{V}\right)+\boldsymbol{V} \tag{4}$$

其中：$\sigma(\cdot)$ 表示激活函数，W^v 是可学习的参数。

文本全连通图：为了增强文本单词的语义表达能力，构建一个无向全连通图 $G_u=(V_u,E_u)$ 学习文本单词之间的语义关联。V_u 表示文本单词集合，E_u 表示单词之间关系集合。表示邻接 $\boldsymbol{A}^u\in\mathbb{R}^{m\times m}$ 矩阵，m 是一句文本中的单词个数，并同 $\boldsymbol{A}^v$ 一样进行初始化和归一化。$\boldsymbol{S}^u\in\mathbb{R}^{m\times m}$ 是任意单词之间的相似性矩阵，将 $\boldsymbol{A}^u$ 和 $\boldsymbol{S}^u$ 逐元素相乘得到边权矩阵 $\boldsymbol{W}_e^u$，即

$$\boldsymbol{W}_e^u=\boldsymbol{S}^u\odot\boldsymbol{A}^u \tag{5}$$

$$s_{ij}^u=\frac{\exp\left(\mathrm{LeakyReLU}\left(\left(\boldsymbol{W}_q^u\boldsymbol{u}_i\right)^{\mathrm{T}}\left(\boldsymbol{W}_k^u\boldsymbol{u}_j\right)\right)\right)}{\sum_{j=0}^{m}\exp\left(\mathrm{LeakyReLU}\left(\left(\boldsymbol{W}_q^u\boldsymbol{u}_i\right)^{\mathrm{T}}\left(\boldsymbol{W}_k^u\boldsymbol{u}_j\right)\right)\right)} \tag{6}$$

其中：$\boldsymbol{W}_q^u$ 和 $\boldsymbol{W}_k^u$ 是可学习的参数。

同理可得到文本单词的增强特征 $\boldsymbol{U}^*$，即

$$\boldsymbol{U}^*=\sigma\left(\boldsymbol{W}_e^uW^u\boldsymbol{U}\right)+\boldsymbol{U} \tag{7}$$

其中：W^u 是可学习的参数。

2.3　基于位置编码的自适应相关性可学习模块

对图像区域的位置信息编码可以使其获得空间语义，加强了图像区域特征的表征能力。位置信息可以由 $p_i=(x_i,y_i,w_i,h_i)$ 表示，x_i 和 y_i 表示图像区域左上角坐标，w_i 和 h_i 表示图像区域的宽和高。

为了从全局角度捕捉空间特性，选择绝对位置编码而不是相对位置编码[34]。对于原始位置向量 $\boldsymbol{p}_i$ 添加两个额外维度，即宽高比和面积，并归一化它们以获得新的位置向量 $\hat{\boldsymbol{p}}_i\in\mathbb{R}^6$，表示为

$$\hat{\boldsymbol{p}}_i=\left(\frac{x_i}{w},\frac{y_i}{h},\frac{w_i}{w},\frac{h_i}{h},\frac{w_i}{h_i},\frac{w_ih_i}{wh}\right) \tag{8}$$

然后经过全连接层得到绝对位置编码 $\tilde{\boldsymbol{p}}_i$，即

$$\tilde{\boldsymbol{p}}_i=\sigma\left(\boldsymbol{W}_p\hat{\boldsymbol{p}}_i+\boldsymbol{b}_p\right) \tag{9}$$

其中：$\sigma(\cdot)$ 表示激活函数，$\boldsymbol{W}_p$ 是可学习的参数，$\boldsymbol{b}_p$ 为可学习的偏置矩阵。将绝对位置编码与增强语义后的图像区域特征融合即可得到具有空间语义的特征，即 $\tilde{\boldsymbol{V}}=\boldsymbol{V}^*\odot\tilde{\boldsymbol{P}}$，$\odot$ 表示矩阵点乘运算。

自适应相关可学习注意力机制就是使用相关性阈值准确地排除不相关的内容，使注意力更集中在相

关内容上，促进了语义对齐学习。为了学习相关和不相关单词–区域片段相似度之间的区别，需要采样构建两种分布的相似性集合，即

$$\begin{aligned}\mathcal{S}_k^+ &= \left[s_1^+, s_2^+, s_3^+, \cdots, s_i^+, \cdots\right] \\ \mathcal{S}_k^- &= \left[s_1^-, s_2^-, s_3^-, \cdots, s_i^-, \cdots\right]\end{aligned} \tag{10}$$

其中：$\mathcal{S}_k^+$ 和 $\mathcal{S}_k^-$ 被认为是相关和不相关的单词–区域片段对的标签，并且它们在训练过程中被动态更新，$k \in [1,b]$ 表示 mini-batch 中的更新索引（b 是批大小）。

当采样相关和不相关的单词–区域片段对的相似性时，面临一个问题：有图像–文本实例级注释，没有片段级单词–区域匹配注释。本文通过对单词–区域片段对分配伪相似性标签解决这个问题。采样策略基于这样一个事实：图像的真实描述应该与图像内容完全相关。因此，对于匹配的图像–文本对，采样每个单词与所有图像区域之间的最大相似度。不过这里的图像区域融合了绝对位置编码，具备了空间语义的图像区域与单词的最大相似度更能准确反映两者的对应程度。最大相似度采样表示为

$$s_i^+ = \max\left(\left\{\cos\left(u_i^*, \tilde{v}_j^+\right)\right\}_{j=1}^n\right) \tag{11}$$

对于不匹配的词，来自错误图像的所有图像区域与它无关。这些不匹配的单词–区域对的最大相似性提供了最有效的区分约束，因为它反映了不匹配片段的相似性的上限。因此依然采用最大相似度采样，表示为

$$s_i^- = \max\left(\left\{\cos\left(u_i^*, \tilde{v}_j^-\right)\right\}_{j=1}^n\right) \tag{12}$$

为了保证伪标签的准确性，避免低质量的伪标签影响模型整体的性能，基于所计算的相似性排名的正确性，即最终的图文匹配的相似性分数 $S(U,V)$ 中排名第一的图片是正确的，就认为收集到的伪标签是可靠的并保留在采样集中。另一方面，如果 $S(U,V)$ 中排名第一的图片是错误的，则认为收集到的伪标签是不可靠的，需要直接丢弃。

基于构造的两个集合 $\mathcal{S}_k^+$ 和 $\mathcal{S}_k^-$ 在每 k 个时间步更新，可以得到关于单词–区域片段对相似度 S 的相关和不相关概率分布，即

$$F_k^+(s) = \frac{1}{\sigma_k^+\sqrt{2\pi}} \mathrm{e}^{\left[-\frac{\left(s-\mu_k^+\right)^2}{2\left(\sigma_k^+\right)^2}\right]} \tag{13}$$

$$F_k^-(s) = \frac{1}{\sigma_k^-\sqrt{2\pi}} \mathrm{e}^{\left[-\frac{\left(s-\mu_k^-\right)^2}{2\left(\sigma_k^-\right)^2}\right]} \tag{14}$$

其中：$\left(\mu_k^+, \sigma_k^+\right)$ 和 $\left(\mu_k^-, \sigma_k^-\right)$ 分别是两个分布的均值和标准差。

目标是找到一个最佳的阈值，可以实现最小的错误概率，最大限度地区分相关和不相关的分布。因此，最优阈值学习可以被公式化为最小错误概率问题。

$$\min_t \int_{-\infty}^t F_k^+(s)\mathrm{d}s + \int_t^{+\infty} F_k^-(s)\mathrm{d}s \quad t \geqslant 0 \tag{15}$$

其中：t 是决策变量，$t \geqslant 0$ 是相关单词–区域片段对的充分条件。

由于公式（15）是可微的，可以从 t 的一阶最优解条件出发，即 $F_k^+(t) - F_k^-(t) = 0$，满足此条件的最优阈值计算为

$$t_k = \left[\left(\mu_k^-\sigma_k^{+2} - \mu_k^+\sigma_k^{-2} + \sigma_k^+\sigma_k^- \cdot \sqrt{\left(\mu_k^+ - \mu_k^-\right)^2 + 4\left(\sigma_k^{+2} - \sigma_k^{-2}\right)\ln\frac{\sigma_k^-}{\sigma_k^+}}\right) \Big/ \left(\sigma_k^{+2} - \sigma_k^{-2}\right)\right]_+ \tag{16}$$

其中 $[\cdot]_+ \equiv \max(\cdot, 0)$。

最优阈值 t_k 的显式相关性区分可以很容易地集成以形成一个统一的学习框架，这使阈值能够调整特征学习，有利于学习更多的区分片段特征以更好地分离它们。这样可以准确地找到模态之间的共享语义测量图像–文本相似性。通过设计一个掩码函数允许模型在学习过程中根据相关和不相关的分布边界自适应地聚合共享语义。

计算所有单词和图像区域之间的相似度，即

$$s_{ij} = \frac{u_i^* \tilde{v}_j^{\mathrm{T}}}{\| u_i^* \| \| \tilde{v}_j \|}, i \in [1,m], j \in [1,n] \tag{17}$$

其中：u^* 是语义增强后的单词表征，$\tilde{v}_j$ 是语义增强并且融入了空间语义的图像区域表征。

对于每一个作为查询的单词去查询相关的图像区域内容，使用相关性阈值去区分相关的图像区域，即单词–区域相似性被重新调节为

$$\hat{s}_{ij} = \begin{cases} s_{ij} - t_k, & s_{ij} > t_k \\ -\infty, & s_{ij} \leqslant t_k \end{cases}, j = 1, \cdots, n \tag{18}$$

经过归一化后可以计算出所有图像区域对应的注意权 $\left\{w_{ij}\right\}_{j=1}^n$。

$$\mathrm{Mask}^{t_k}\left(w_{ij}\right) = \begin{cases} \dfrac{\mathrm{e}^{\lambda s_{ij}}}{\sum_{j=1}^n \mathrm{e}^{\lambda s_{ij}}} & s_{ij} > t_k \\ 0 \leftarrow \mathrm{e}^{-\infty} & s_{ij} \leqslant t_k \end{cases} \tag{19}$$

只有大于相关阈值的图像区域被分配注意力，而其他

不相关区域的注意力权重通过将相似性掩蔽为–∞而被设置为零。λ 是缩放参数。同理可以得到图像到文本方向的掩码函数 $\mathrm{Mask}^{t_k}\left(w_{ji}\right)$，即以图像区域作为查询。

该掩码函数简单有效，因为自适应学习阈值 t_k 使得注意力能够准确地对齐相关片段内容，并且排除错误的语义对齐，从而获得更纯的共享语义以测量图像–文本相似性。图像–文本匹配目的是准确地找到模态之间的共享语义测量图像–文本相似性。这反映在两个检索方向上，即文本到图像和图像到文本。基于掩码函数可以得到两个检索方向的共享语义，即

$$S_j^U=\sum_{i=1}^{m}\mathrm{Mask}^{t_k}\left(w_{ji}\right)u_i^* \tag{20}$$

$$S_i^V=\sum_{j=1}^{n}\mathrm{Mask}^{t_k}\left(w_{ij}\right)\tilde{v}_j \tag{21}$$

为了进一步减少推理时间，避免不相关查询片段的相似度计算 $R(\cdot)$，设计了一个掩码函数，这里依然从文本到图像的检索方向说明。首先对于每个查询词 u_i^*，计算学习阈值 t_k 和 u_i^* 与所有图像区域 $\left\{\tilde{v}_j\right\}_{j=1}^{n}$ 的最大相似度之间的差为

$$s(u_i^*)=\max\left(\{s_{ij}\}_j^n\right)-t_k \tag{22}$$

当 $s(u_i^*)\leqslant 0$，u_i^* 是不相关查询，因此可以设计一个指示函数：

$$f(u_i^*)=\begin{cases}1, & s(u_i^*)\leqslant 0\\ 0, & s(u_i^*)>0\end{cases} \tag{23}$$

通过指示函数可以得到不相关查询片段的索引集合 $\mathcal{X}$，其表示为 $\mathcal{X}=\left\{f(u_i^*)\cdot i\right\}_{i=1}^{m}$。最终共享语义的计算细化为

$$\mathrm{Mask}\left(R\left(S_i^V,u_i^*\right)\right)=\begin{cases}R\left(S_i^V,u_i^*\right), & i\notin\mathcal{X}\\ s\left(u_i^*\right), & i\in\mathcal{X}\end{cases} \tag{24}$$

最终图像–文本相似度由两个方向相似度分数的和组成，即

$$S(U,V)=\frac{1}{m}\sum_{i=1}^{m}\mathrm{Mask}\left(R\left(S_i^V,u_i^*\right)\right)+\frac{1}{n}\sum_{j=1}^{n}\mathrm{Mask}\left(R\left(S_j^U,\tilde{v}_j\right)\right) \tag{25}$$

其中 $R(\cdot)$ 表示相似性计算，实验中采用的是余弦相似度。

2.4　损失函数

根据现有的方法[20-21]，本文采用的端到端训练的损失函数是双向三元组排序损失，它通过固定的边界约束对齐的图像–文本对的相似性高于未对齐的对的相似性，专注于优化产生最高损失的最难的未对齐样本。给定真实的图像–文本对 (U,V) 及其所有不匹配的对 (U,V') 和 (U',V)，最难的未对齐样本通过 $V'=\arg\max_{p\neq V}S(U,p)$ 和 $U'=\arg\max_{q\neq U}S(q,V)$ 进行选择。因此，目标函数可以表示为

$$L=\sum_{(U,V)}\left[\gamma-S(U,V)+S\left(U,V'\right)\right]_+ +\left[\gamma-S(U,V)+S\left(U',V\right)\right]_+ \tag{26}$$

其中 γ 是边界超参数，$[x]_+\equiv\max(x,0)$。

3　实　验

3.1　实验设置

为了验证效果，在两个基准数据集上进行了大量实验。Flickr 30 k[16]共有 31 000 张图片和 155 000 个句子。按照文献[14]的相同协议，将 Flickr 30 k 分为 1 000 张测试图片、1 000 张验证图片和 29 000 张训练图片。MS-COCO[17]包含 123 287 张图片和 616 435 个句子，将其分为 5 000 张测试图片、5 000 张验证图片和 113 287 张训练图片。在 MS-COCO 的结果上，对 1 K 测试图片的 5 折平均和完整的 5 K 测试图片进行了测试。

采用 Recall@K 评价指标，即排名前 K 个查询结果的正确个数与所有正确结果数的比率（召回率），其中 K 取{1, 5, 10}。

所有实验均在 NVIDIA GeForce RTX 4090 GPU 上进行。Adam 优化器用于模型优化，初始学习率为 0.000 5，每 8 个 epoch 衰减 10%。mini-batch 大小设置为 128，epoch 设置为 20，缩放参数 λ 设置为 20，边界超参数 γ 设置为 0.2。

3.2　定量结果

下面是提出的模型与当前最先进的方法进行量化对比的结果。这里直接引用这些方法原论文中的结果。

表 1 展示了模型在 Flickr 30 k 测试数据集上与最先进方法的对比结果，可以看出本文模型在两个检索方向的评价指标均优于基线模型，在 Rsum 上相对提高了 7.4%。CAMP[35]、BFAN、PFAN[34]、VSRN[28]、SGM[29]、GSMN[30]、IMRAM[21]和 UARDA[15]均是在典型方法 SCAN[14]上改进的。本文模型相较于基线模型 UARDA 在 R@1 上分别提升了 2.4% 和 2.0%，同时与最新的先进方法 HREM[36]和 CHAN[37]对比，在 Rsum 上达到最先进水平，表 2 是在更大的数据集

MS-COCO 进行的 1 K 测试集对比实验，可以看到本文模型依然优于基线模型，在 Rsum 上相对基线模型提高了 1.1%。表 3 是在完整的测试集上进行对比实验的结果，由于部分论文没有提供相关数据，因此只与提供了数据的论文进行了对比。无论文本检索还是图像检索在 R@1 上均优于最先进方法。这验证了本文方法在图像–文本匹配上的有效性。

表 1　在 Flickr 30 k 数据集上的对比结果

Tab. 1　Comparison results on Flickr 30 k datasets

方法	文本检索			图像检索			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
SCAN	67.4	90.3	95.8	48.6	77.7	85.2	465.0
CAMP	68.1	89.7	95.2	51.5	77.1	85.3	466.9
BFAN	68.1	91.4	—	50.8	78.4	—	288.7
PFAN	70.0	91.8	95.0	50.4	78.7	86.1	472.0
VSRN	70.4	89.2	93.7	53.0	77.9	85.7	469.9
SGM	71.8	91.7	95.5	53.5	79.6	86.5	478.6
GSMN	76.4	94.3	97.3	57.4	82.3	89.0	496.8
IMRAM	74.1	93.0	96.6	53.9	79.4	87.2	484.2
UARDA	77.8	95.0	97.6	57.8	82.9	89.2	500.3
HREM	79.5	94.3	97.4	59.3	85.1	91.2	506.8
CHAN	79.7	94.5	97.3	60.2	85.3	90.7	507.8
本文方法	80.2	96.0	98.1	59.8	83.6	90.0	507.7

表 2　在 MS-COCO 上的 1 K 测试集对比结果

Tab. 2　Comparison results of 1 K test sets on MS-COCO

方法	文本检索			图像检索			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
SCAN	72.7	94.8	98.4	58.8	88.4	94.8	507.9
CAMP	72.3	94.8	98.4	58.5	87.9	95.0	506.8
BFAN	74.9	95.2	—	59.4	88.4	—	317.9
PFAN	76.5	96.3	99.0	61.6	89.6	95.2	518.2
VSRN	76.2	94.8	98.2	62.8	89.7	95.1	516.8
SGM	73.4	93.8	97.8	57.5	87.3	94.3	504.1
GSMN	78.4	96.4	98.6	63.3	90.1	95.7	522.5
IMRAM	76.7	95.6	98.5	61.7	89.1	95.0	516.6
UARDA	78.6	96.5	98.9	63.9	90.7	96.2	524.8
HREM	80.0	96.0	98.7	62.7	90.1	95.4	522.8
CHAN	79.7	96.7	98.7	63.8	90.4	95.8	525.0
本文方法	80.3	96.6	98.7	64.0	90.1	96.2	525.9

表 3　在 MS-COCO 上的 5 K 测试集对比结果

Tab. 3　Comparison results of 5 K test sets on MS-COCO

方法	文本检索			图像检索			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
SCAN	50.4	82.2	90.0	38.6	69.3	80.4	410.9
CAMP	50.1	82.1	89.7	39.0	68.9	80.2	410.0
VSRN	53.0	81.1	89.4	40.5	70.6	81.1	415.7
SGM	50.0	79.3	87.9	35.3	64.9	76.5	393.9
IMRAM	53.7	83.2	91.0	39.7	69.1	79.8	416.5
UARDA	56.2	83.8	91.3	40.6	69.5	80.9	422.3
HREM	58.9	85.3	92.1	40.0	70.6	81.2	428.1
CHAN	60.2	85.9	92.4	41.7	71.5	81.7	433.4
本文方法	60.6	85.0	91.8	41.8	71.3	81.1	431.6

3.3　定性结果

图 3 和图 4 可视化地展示了该模型在 Flickr 30 k 数据集中用图像检索文本和用文本检索图像的结果。

1.A man in jeans climbed up a light pole to change the light bulb while the light is on.
2.Man with beard and glasses wearing a jacket, holding onto a light pole and light.
3.A boy is hanging on a light pole while opening his mouth and touching the light.
4.A man has climbed to the top of a red light pole and is touching the light.
5.A man is climbing a red pole.

1.A lady rides her pony in a victory lap after competing in an English riding competition.
2.This woman is riding a beautiful and well groomed horse.
3.A young lady equestrian riding a brown and white horse.
4.A woman is riding a brown and white horse.
5.A young girl rides horseback.

序号 1—5 为结果排名，其中不匹配的文本被标记为红色。

图 3　文本检索结果可视化

Fig. 3　Visualization of text retrieval results

从左到右是检索结果排名，其中不匹配的图像具有红色框，匹配的图像具有绿色框。

图 4　图像检索可视化

Fig. 4　Visualization of image retrieval

为了检验模型的图文匹配性能，分别对文本检索和图像检索两个方向在 Flickr 30 k 数据集进行了可视化实验。图 3 是文本检索结果可视化，用指定的图像查询与之相关的文本，按照相关程度依次排序，其中不匹配的文本被标注为红色。图 4 是图像检索结果可视化，通过给出查询文本检索出与之相关的图像，对于正确的结果用绿色框出来。从实验结果可以看出，本文模型可以准确找出与查询最匹配的图像或者文本，这是因为通过语义的增强以及融合了空间语义的自适应相关性可学习的注意力，使模型对于相似语义的区分能力得到了进一步的提升。在模态间进行信息交互时有效地过滤冗余信息，使单词–区域片段特征之间的语义对齐更加准确，大大提高了图文匹配的精确度。

3.4　消融实验

为了验证网络内部模块的有效性，在 Flickr 30 k 数据集上进行了消融实验，实验通过在单一基线的基

础上分别增加了语义增强和绝对位置编码，结果见表 4。

表 4　在 Flickr 30 k 数据集上的消融实验

Tab. 4　Ablation experiments on the Flickr 30 k datasets

方法	文本检索			图像检索		
	R@1	R@5	R@10	R@1	R@5	R@10
基线	77.8	95.0	97.6	57.8	82.9	89.2
语义增强	79.0	95.9	97.9	58.9	83.3	89.3
绝对位置编码	80.1	95.1	97.7	57.4	82.5	89.0
本文模型	80.2	96.0	98.1	59.8	83.6	90.0

本文模型是在基线模型 UARDA 上的进一步改进，通过消融实验证明模块的可行性。仅仅对片段特征进行语义增强，从实验结果可以看出各个指标均有所提升，尤其是在 R@1 上相对提升了 1.2%，证明了方法的有效性。然后是仅仅加入绝对位置编码的自适应相关性可学习注意力，在文本检索方向上，R@1 相对提高了 2.3%，空间语义的加入进一步丰富了图像区域的语义表达，促进了局部语义的对齐。最后是完整的模型，两者结合进一步提高了图像–文本匹配的准确性。

为了验证单词–区域片段特征单一方向的相似度匹配性能，在 Flickr 30 k 数据集上进行了消融实验，结果见表 5。结果表明单一方向的相似度匹配会降低准确率，而结合单词–区域片段特征两个方向的相似度匹配可以获得更好的性能。

表 5　在 Flickr 30 k 数据集上的消融实验

Tab. 5　Ablation experiments on the Flickr 30 k datasets

方法	文本检索			图像检索		
	R@1	R@5	R@10	R@1	R@5	R@10
V-U	74.5	94.8	97.5	55.0	80.8	87.9
U-V	79.4	95.0	97.3	58.4	83.3	89.6
本文模型	80.2	96.0	98.1	59.8	83.6	90.0

4　结　语

本文提出一种适用于图像–文本匹配任务的融合语义增强和位置编码的自适应相关性可学习注意力框架。通过引入图注意力机制，对图像（文本）特征进行语义增强，提高局部语义对齐的准确性。在原有方法的自适应相关性可学习注意力的基础上融入了绝对位置编码，具备空间语义后可以得到更具区分性的相关和无关分布，从而学习到更好的相关性阈值。综合实验表明，本文方法与现有方法相比具有显著的优越性。在未来的工作中可以考虑将这种方法应用到其他的多模态场景中。

参考文献：

[1] YU J，ZHANG W，LU Y，et al. Reasoning on the relation：enhancing visual representation for visual question answering and cross-modal retrieval[J]. IEEE Transactions on multimedia，2020，22(12)：3196–3209.

[2] JIANG M，CHEN S，YANG J，et al. Fantastic answers and where to find them：immersive question-directed visual attention[C]//IEEE. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York：IEEE，2020.

[3] TAN J H，CHAN C S，CHUAH J H. Comic：toward a compact image captioning model with attention[J]. IEEE Transactions on multimedia，2019，21(10)：2686–2696.

[4] GU J，CAI J，JOTY S，et al. Look，imagine and match：improving textual-visual cross-modal retrieval with generative models[C]//IEEE. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE，2018.

[5] YU L，GUO Y，BAKKER E M，et al. Learning a recurrent residual fusion network for multimodal matching[C]//IEEE. Proceedings of the IEEE International Conference on Computer Vision. New York：IEEE，2017.

[6] MA L，LU Z，SHANG L，et al. Multimodal convolutional neural networks for matching image and sentence[C]//IEEE. Proceedings of the IEEE International Conference on Computer Vision. New York：IEEE，2015.

[7] WU Y L，WANG S H，HUANG Q M. Learning semantic structure-preserved embeddings for cross-modal retrieval [C]//IEEE. Proceedings of the 26th ACM International Conference on Multimedia. New York：IEEE，2018.

[8] WANG L，LI Y，LAZEBNIK S. Learning deep structure-preserving image-text embeddings[C]//IEEE. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE，2016.

[9] FAGHRI F，FLEET D J，KIROS J R，et al. VSE++：improving visual-semantic embeddings with hard negatives[EB/OL]. [2023–09–01].https://arxiv.org/pdf/1707.05612.pdf.

[10] NIKOLAOS S，XU X，KAKADIARIS I A. Adversarial representation learning for text-to-image matching [C]//IEEE. Proceedings of the IEEE/CVF International Conference on Computer Vision. New York：IEEE，2019.

[11] KARPATHY A，JOULIN A，LI F F. Deep fragment

embeddings for bidirectional image sentence mapping. [EB/OL]. [2023-09-01]. https://arxiv.org/pdf/1406.5679.pdf.

[12] KARPATHY A, LI F F. Deep visual-semantic alignments for generating image descriptions[C]//IEEE. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2015.

[13] NIU Z, ZHOU M, WANG L, et al. Hierarchical multimodal LSTM for dense visual-semantic embedding[C]// IEEE. Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE, 2017.

[14] LEE K H, XI C, GANG H, et al. Stacked cross attention for image-text matching[EB/OL].[2023-09-01]. https://doi.org/10.48550/arXiv.1803.08024.

[15] ZHANG K, MAO Z D, LIU A A, et al. Unified adaptive relevance distinguishable attention network for image-text matching[J]. IEEE Transactions on multimedia, 2022, 25: 1320-1332.

[16] 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]// IEEE. Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE, 2015.

[17] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: common objects in context[C]//ECCV. Computer Vision-ECCV 2014: 13th European Conference. Berlin: Springer International Publishing, 2014: 740-755.

[18] FROME A, CORRADO G S, SHLENS J, et al. DeVISE: a deep visual-semantic embedding model[C]// NIPS. Advances in Neural Information Processing Systems. Denver: NIPS, 2013.

[19] KIRO R, SALAKHUTDINOV R, ZEMEL R S. Unifying visual-semantic embeddings with multimodal neural language models[EB/OL]. [2023-09-01]. https://doi.org/10.48550/arXiv.1411.2539.

[20] ANDERSON P, HE X, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//IEEE. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2015.

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

[22] CHEN T L, LUO J B. Expressing objects just like words: recurrent visual embedding for image-text matching[C]//AAAI. Proceedings of the AAAI Conference on Artificial Intelligence. California: AAAI, 2020.

[23] YANG X, DENG C, DANG Z, et al. SelfSAGCN: self-supervised semantic alignment for graph convolution network[C]//IEEE. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021.

[24] LIU C X, MAO Z D, LIU A A, et al. Focus your attention: a bidirectional focal attention network for image-text matching[C]//ACM. Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019.

[25] ZHONG Y W, WANG L W, CEN J S, et al. Comprehensive image captioning via scene graph decomposition [C]//ECCV. Computer Vision-ECCV 2020: 16th European Conference. Berlin: Springer International Publishing, 2020.

[26] HUANG D, CHEN P, ZENG R, et al. Location-aware graph convolutional networks for video question answering[C]//AAAI. Proceedings of the AAAI Conference on Artificial Intelligence. California: AAAI, 2020.

[27] YU W J, ZHOU J W, YU W H, et al. Heterogeneous graph learning for visual commonsense reasoning [EB/OL]. [2023-09-01]. https://doi.org/10.48550/arXiv.1910.11475.

[28] LI K P, ZHANG Y L, LI K, et al. Visual semantic reasoning for image-text matching[C]//IEEE. Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2019.

[29] LIU C X, MAO Z D, ZHANG T Z, et al. Graph structured network for image-text matching[C]//IEEE. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020.

[30] WANG S, WANG R, YAO Z, et al. Cross-modal scene graph matching for relationship-aware image-text retrieval[C]//IEEE. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. New York: IEEE, 2020.

[31] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[EB/OL]. [2023-09-01]. https://doi.org/10.48550/arXiv.1710.10903.

[32] KRISHNA R, ZHU Y K, GROTH O, et al. Visual genome: connecting language and vision using crowd-

sourced dense image annotations[J]. International journal of computer vision, 2017, 123: 32–39.

[33] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//IEEE. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016.

[34] WANG Y X, YANG H, QIAN X M, et al. Position focused attention network for image-text matching [EB/OL]. [2023–09–01]. https://doi.org/10.48550/arXiv.1907.09748.

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

[36] FU Z, MAO Z, SONG Y, et al. Learning semantic relationship among instances for image-text matching[C]// IEEE. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2023: 15159–15168.

[37] PAN Z, WU F, ZHANG B. Fine-grained image-text matching by cross-modal hard aligning network[C]// IEEE. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2023: 19275–19284.

责任编辑：郎婧

（上接第 45 页）

京：中国轻工业出版社, 1994.

[21] MCDONALD T P, WALMSLEY A R, HENDERSON P. Asparagine 394 in putative helix 11 of the galactose-H^+ symport protein (GalP) from *Escherichia coli* is associated with the internal binding site for cytochalasin B and sugar[J]. Journal of biological chemistry, 1997, 272(24): 15189–15199.

[22] JECKELMANN J M, ERNI B. Transporters of glucose and other carbohydrates in bacteria[J]. Pflugers archiv: European journal of physiology, 2020, 472(9): 1129–1153.

[23] BANARES A B, NISOLA G M, VALDEHUESA K N G, et al. Engineering of xylose metabolism in *Escherichia coli* for the production of valuable compounds[J]. Critical reviews in biotechnology, 2021, 41(5): 649–668.

[24] DESAI T A, RAO C V. Regulation of arabinose and xylose metabolism in *Escherichia coli*[J]. Applied and environmental microbiology, 2010, 76(5): 1524–1532.

[25] 孙金凤, 田康明, 沈微, 等. 大肠杆菌不同菌株木糖代谢差异性的遗传本质[J]. 食品与发酵工业, 2017, 43(10): 68–73.

责任编辑：郎婧