文章编号：1672-3961(2025)04-0029-11 DOI：10.6040/j.issn.1672-3961.0.2024.024

基于多粒度对齐网络的图像-文本匹配方法

王旭峰[1]，周迪[1]，张风雷[1]，宋雪萌[2]，刘萌[1*]
（1.山东建筑大学计算机科学与技术学院，山东 济南 250101；2.山东大学计算机科学与技术学院，山东 青岛 266237）

摘要：为精准匹配图像与文本数据，提出一种多粒度对齐网络（multi-granularity alignment network，MGAN）。通过对比语言-图像预训练模型和基于Transformer的双向编码器模型，分别提取图像块级、区域级和全局级3个不同粒度的信息，弥补匹配信息单一的缺陷。根据各级信息的特性，采用多级对齐机制。在区域级对齐上，结合多视角总结策略，让MGAN有效应对图像和文本之间的一对多描述问题；在图像块级对齐上，引入跨模态相似性交互建模模块，进一步增强图像与文本之间的细节交互。在Flickr30K和MS-COCO两个公开数据集上的大量试验结果表明，MGAN具有更高的匹配性能，验证了多粒度对齐网络方法的有效性。

关键词：图像-文本匹配；跨模态检索；多粒度；多视角；跨模态相似性交互
中图分类号：TP391 **文献标志码**：A
引用格式：王旭峰，周迪，张风雷，等. 基于多粒度对齐网络的图像-文本匹配方法[J]. 山东大学学报（工学版），2025，55(4)：29-39.
WANG Xufeng，ZHOU Di，ZHANG Fenglei，et al. Multi-granularity alignment network for image-text matching[J]. Journal of Shandong University（Engineering Science），2025，55(4)：29-39.

Multi-granularity alignment network for image-text matching

WANG Xufeng[1]，ZHOU Di[1]，ZHANG Fenglei[1]，SONG Xuemeng[2]，LIU Meng[1*]
（1. College of Computer Science and Technology，Shandong Jianzhu University，Jinan 250101，Shandong，China；2. College of Computer Science and Technology，Shandong University，Qingdao 266237，Shandong，China）

Abstract：To precisely match image and text data，a multi-granularity alignment network（MGAN）was proposed. By adopting a contrastive language-image pre-training model and a Transformer-based bidirectional encoder model，MGAN extracted information at three different granularities：patch level，regional level，and global level，addressing the shortcomings of single-granularity information matching. A multi-level alignment mechanism was employed based on the characteristics of information at each level. At the regional level，a multi-view summarization module was integrated，allowing MGAN to effectively handle the one-to-many description problems between images and texts. At the patch level，a cross-modal similarity interaction modeling module was introduced to further enhance the detailed interactions between images and texts. Extensive experimental results on the publicly available datasets Flickr30K and MS-COCO demonstrated that MGAN achieved promising performance，confirming the effectiveness of the multi-granularity alignment network approach.

Keywords：image-text matching；cross-modal retrieval；multi-granularity；multi-view；cross-modal similarity interaction

0 引言

随着大数据时代的来临，视觉和文本数据的爆炸性增长激发了对多模态表示、理解与推理能力的深入研究。在此背景下，图像-文本匹配成为多模态研究领域的一个关键任务，核心是揭示图像与文本之间的复杂语义关联，量化其相似性。当前，在多媒体识别[1-3]及跨模态检索[4-5]等多模态交互领域，研究人员通过计算机视觉技术取得一系列突破，展示技术的创新性，指明研究的发展方向。但图像-文本匹配仍面临如何准确揭示和处理图文之

收稿日期：2024-01-31
基金项目：国家自然科学基金资助项目（62376140，U23A20315，62236003）；山东省优秀青年科学基金资助项目（ZR2022YQ59）；山东省高等学校青创科技支持计划资助项目（2023KJ128）
第一作者简介：王旭峰（1997— ），男，江苏徐州人，硕士研究生，主要研究方向为多媒体计算和信息检索。E-mail：xufeng_wang@163.com
*通信作者简介：刘萌（1991— ），女，黑龙江尚志人，教授，硕士生导师，博士，主要研究方向为多媒体计算和信息检索。
E-mail：mengliu.sdu@gmail.com

间复杂语义关联的挑战。

为了应对这一挑战,研究者提出多种方法,根据处理信息的粒度,大致分为全局级别匹配方法[6-10]和局部级别匹配方法[11-16]。全局级别匹配方法着重于图像与文本数据的宏观理解与表示,通过将图像与文本映射至共同的向量空间,实现二者间的对齐。全局级别匹配方法虽能捕捉到与全局图像线索相关的文本主题,但可能因过度依赖全局线索而在精确匹配上出错。局部级别匹配方法聚焦于局部特征的学习与理解,构建更精细的表示。在这一方向上,不少研究致力于提升对图像与文本间细致语义信息的理解,以减小模态间的差异。但局部级别匹配方法可能会过度关注图像细节,忽略图像背景信息,造成细节属性准确但上下文不一致的错误匹配。此外,单一图像多角度描述处理也是一大挑战。为了整合不同粒度的优势并提升匹配性能,很多研究者尝试解决该问题,却依然面临识别区域准确性的挑战。局部级信息的有效性在很大程度上取决于局部识别的准确性,识别不精确可能影响后续图像分析的准确度,导致匹配错误。因此,深入探索多视角问题仍是领域内亟待解决的课题。

为了应对现有图像-文本匹配方法中的局限性,本研究提出一种多粒度对齐网络(multi-granularity alignment network, MGAN),综合考虑图像与文本匹配过程中的各个维度,将不同粒度的信息层次(区域级、图像块级及全局级)有机结合,弥补单一粒度匹配的缺陷。在区域级,MGAN 采用多视角总结策略,有效解决由于单一视角导致的信息偏颇问题;通过引入图像块级的图像信息,捕捉图像与文本间更为细节的语义交互,对模型在多模态学习和推理过程中的性能至关重要;结合全局级的图像信息,使模型在学习图像细节属性的同时理解全局背景信息,提高对上下文背景的识别准确性。此外,为了进一步提升细粒度信息处理能力,本研究设计一个跨模态相似性交互模块,通过强化图像块与文本单词之间的相似性,深化模型对细节交互的理解,为实现准确的跨模态对齐和交互打下坚实的基础。MGAN 通过聚合来自 3 个不同层次的对齐分数,形成全面的匹配机制。通过广泛试验,MGAN 的有效性已在多个公开数据集上得到验证,证明了其在图像-文本匹配任务中的优越性能。

1 相关工作

1.1 图像-文本匹配

基于全局级别匹配方法重点在于对数据内蕴含的全局语义及上下文要素进行对齐,典型做法是使用卷积神经网络捕获整体图像语义,通过循环神经网络解析文本描述中的语义信息。将全局特征映射到共享空间,以距离度量方式衡量图像-文本对的相似性。例如:文献[6]将卷积神经网络与 Skip-gram 模型结合,建立深度视觉语义嵌入模型;文献[7]将长短期记忆网络与 OxfordNet 结合,提出跨模态语义对齐的统一视觉-语义嵌入模型。然而,此类全局级别匹配方法可能在探究模态内丰富的细粒度语义时存在局限,在图像-文本匹配精确度上受限。

基于局部级别匹配方法旨在挖掘文本中的词汇与图像局部视觉区域之间的关系。由文献[11]提出的深度视觉语义对齐模型是基于局部级别匹配方法的先驱工作。注意力机制在检索领域中展现出巨大潜力,文献[13]将跨模态注意力引入图像-文本匹配,提出双向语义注意力网络,提取最相关的图像特征和文本特征。

研究者们不断挖掘和优化局部语义信息,进一步细化出基于区域特征的匹配方法[17-26],在局部特征注重细节信息的基础上,加强对上下文环境信息的理解。文献[17]将深度视觉语义对齐模型进一步优化,通过图卷积网络推理图像区域间的关系,利用门控记忆机制增强区域特征的全局语义推理,提出视觉语义推理网络;文献[26]提出跨模态语义一致性注意力图像-文本匹配,旨在增强区域与全局对齐的效果。尽管基于区域特征的方法在探索细粒度语义方面表现出色,但可能会忽视整体上下文所提供的信息。因此,将全局与区域信息相结合已成为提升匹配任务性能的关键。

1.2 视觉-语言预训练模型

在自然语言处理领域,先预训练后微调的方法已取得显著进步,其中代表性的工作为基于 Transformer[27] 的双向编码器模型(bidirectional encoder representations from Transformers, BERT)[28]。由于语言是跨模态任务中的核心模态,故语言预训练模型的引入促进了视觉-语言预训练(vision-language pretraining, VLP)模型的发展。VLP 模型通过大规模数据集学习视觉与语言间的集成表示,在多种视觉语言任务上提升了性能。

目前,预训练模型框架主要有单流模型和双流模型两种。单流模型(如文献[29]提出的 VideoBERT)将 Transformer 结构拓展至 VLP 领域,处理视觉和文本特征时采用统一框架,将图像与文本视作联合表示,通过共享参数促进跨模态细粒度对齐,但在推理时需要所有模态信息同时输入,微

调复杂度较高。双流模型(如对比语言-图像预训练模型(contrastive language-image pre-training, CLIP)[30])有效融合了视觉与语言理解,在图像分类、检索、生成和视觉问答等任务上具有多样性和卓越表现。由于 CLIP 拥有卓越的泛化性能,本研究决定将 CLIP 整合进框架中,用于提取图像块级别的视觉信息,以优化图像与文本间的对齐和匹配。

2　多粒度对齐网络

多粒度对齐网络如图 1 所示,由 3 个核心部分构成:多级图像编码器、文本编码器和多级对齐机制。在图像-文本匹配领域,传统方法通常专注于区域级或全局图像表示与文本描述的对齐,忽视了不同粒度级别信息间的互动。单一的区域级匹配可能会由于依赖预先识别区域的质量而使性能受限,若这些区域未能有效捕捉与文本描述相关的图像细节,则匹配准确性会大幅下降。

为解决这一挑战,本研究引入图像块级信息作为区域级信息的补充,进一步融入全局级别信息,提出一种多级图像编码器,协同利用区域、图像块和全局图像线索,以实现更精准和全面的图像-文本匹配。

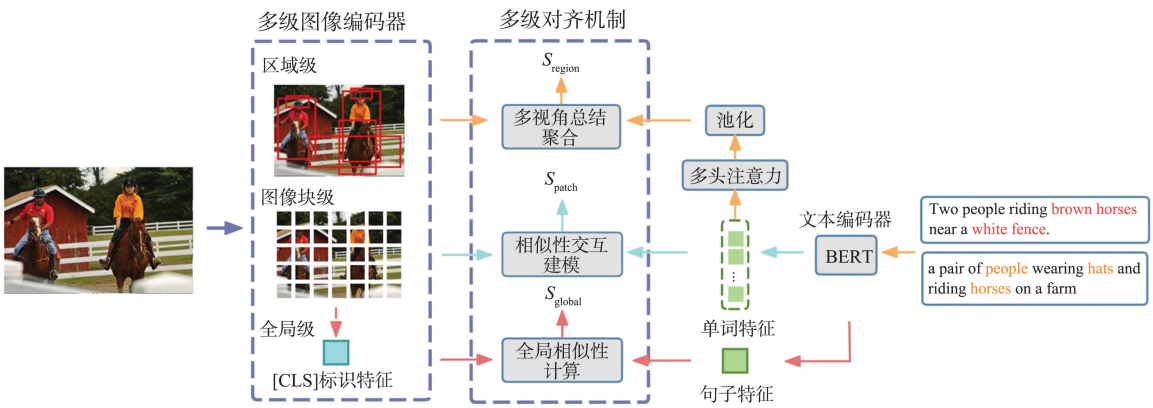


图 1　多粒度对齐网络

Fig.1　Multi-granularity alignment network model

2.1　多级图像编码器

2.1.1　区域级图像编码

在区域级图像编码中,采用自底向上的注意力机制确定图像中最显著的相关区域。给定图像数据 $\boldsymbol{I}$,提取前 r 个具有最高置信度的区域。通过一个全连接层将这些特征映射到 D 维空间中,将每个区域的特征表示为 $\boldsymbol{f}_i \in \mathbf{R}^{2\,048}$,每个区域的位置向量 $\boldsymbol{l}_i=(x_i \quad y_i \quad w_i \quad h_i)$,其中 x_i 和 y_i 为坐标信息,w_i 和 h_i 为相关区域的宽和高。为了能更全面地利用空间属性,在原有的位置向量 $\boldsymbol{l}_i$ 基础上加入宽高比和区域面积,整体进行归一化处理,得到新的位置向量 $\tilde{\boldsymbol{l}}_i$,通过全连接层和一个 sigmoid 激活函数进行绝对位置编码,即

$$\tilde{\boldsymbol{l}}_i=\sigma(\boldsymbol{W}_{\mathrm{p}}\hat{\boldsymbol{l}}_i+\boldsymbol{b}_{\mathrm{p}}), \tag{1}$$

式中:σ 为激活函数;$\boldsymbol{W}_{\mathrm{p}}$ 为权重矩阵,$\boldsymbol{W}_{\mathrm{p}} \in \mathbf{R}^{D\times 6}$;$\hat{\boldsymbol{l}}_i$ 为加入宽高比与区域面积的位置向量,$\hat{\boldsymbol{l}}_i=\left(\frac{x_i}{w} \quad \frac{y_i}{h} \quad \frac{w_i}{w} \quad \frac{h_i}{h} \quad \frac{w_i}{h_i} \quad \frac{w_i h_i}{wh}\right)$;$\boldsymbol{b}_{\mathrm{p}}$ 为偏置向量。

区域的特征表示和位置表示分别整合成矩阵 $\boldsymbol{V}=[\boldsymbol{f}_1 \quad \boldsymbol{f}_2 \quad \cdots \quad \boldsymbol{f}_r] \in \mathbf{R}^{r\times D}$ 和 $\boldsymbol{L}=[\tilde{\boldsymbol{l}}_1 \quad \tilde{\boldsymbol{l}}_2 \quad \cdots \quad \tilde{\boldsymbol{l}}_r] \in \mathbf{R}^{r\times D}$,将二者融合为最终区域特征 $\bar{\boldsymbol{V}}=\boldsymbol{V}\odot\boldsymbol{L} \in \mathbf{R}^{r\times D}$,其中⊙为特征融合操作。

2.1.2　图像块级和全局级图像编码

本研究采用基于视觉 Transformer 的对比语言-图像预训练模型(contrastive language-image pre-training vision Transformer, CLIP-ViT)从图像中提取出图像块级和全局级特征。给定图像 $\boldsymbol{I}$,将其划分为多个尺寸一致的非重叠图像块;将一个[CLS]标识作为附加学习目标附在图像块序列的前端,以此捕获全局图像特征;将此序列输入 CLIP-ViT 模型中,得到图片的图像块特征序列

$$\boldsymbol{P}=[\boldsymbol{p}_1 \quad \boldsymbol{p}_2 \quad \cdots \quad \boldsymbol{p}_M] \in \mathbf{R}^{M\times d_{\mathrm{p}}}, \tag{2}$$

式中,$\boldsymbol{p}_i$ 为第 i 个图像块特征表示,M 为图像被分割成的图像块数,d_{p} 为每个图像块特征维度。[CLS]标识特征包含图像的全局信息,该全局表示记为 $\boldsymbol{p}_{\mathrm{cls}} \in \mathbf{R}^{d_{\mathrm{p}}}$。

为了将图像块特征和[CLS]标识特征转换至所需的维度 D,使用一个多层感知器(multilayer perceptron, MLP)对获取的特征进行维度转换,将图像块级别特征和全局特征转换至 D 维空间,将转换后得到的特征分别表示为 $\bar{\boldsymbol{P}}$ 和 $\boldsymbol{P}^*_{\mathrm{cls}}$,$\bar{\boldsymbol{P}} \in \mathbf{R}^{M\times D}$,$\boldsymbol{P}^*_{\mathrm{cls}} \in \mathbf{R}^{D}$。

2.2 文本编码器

使用 WordPiece 标记化策略处理包含 L 个单词的文本序列,为后续的文本编码做好预处理准备。采用 BERT 模型对文本句子的语义进行深度挖掘,捕获文本上下文特征,得到一个包含 L 个标记的特征序列 $\boldsymbol{E}=[\boldsymbol{e}_1\ \ \boldsymbol{e}_2\ \ \cdots\ \ \boldsymbol{e}_L]$,其中每个单词特征 $\boldsymbol{e}_i\in\mathbf{R}^{768}$。

为了与图像特征的维度保持一致,每个单词特征嵌入同样需要映射到 D 维的特征空间中,该过程由一个全连接层实现,将文本序列特征转换为新的维度,产生最终的文本特征表示 $\boldsymbol{T}=[\boldsymbol{t}_1\ \ \boldsymbol{t}_2\ \ \cdots\ \ \boldsymbol{t}_L]\in\mathbf{R}^{L\times D}$,其中 $\boldsymbol{t}_i$ 为句中每个单词的特征表示。值得注意的是,BERT 模型内置一个专用的[CLS]标识,用于编码整个文本序列的全局语义信息。这个[CLS]标识的特征同样通过上述全连接层进行转换,得到 D 维的全局文本表示 $\boldsymbol{t}^*_{\text{cls}}$。与图像块级特征 $\bar{\boldsymbol{P}}$ 类似,文本特征 $\boldsymbol{T}$ 中包含每个单词的特征,而全局文本表示 $\boldsymbol{t}^*_{\text{cls}}$ 则与全局图像特征 $\boldsymbol{P}^*_{\text{cls}}$ 保持一致,提供文本描述的全局信息。图像特征与文本特征之间的这种相同性质,为接下来实现多级对齐机制提供有利条件。

2.3 多级对齐

通过对每种粒度的信息特征进行分析,MGAN 提供了一个多级对齐模块,旨在精确计算图像特征与文本特征之间的跨模态相似度。

2.3.1 区域级对齐

在区域级,保持图像与文本间的语义一致性是实现精准匹配的关键。面对图像和文本之间的一对多描述问题,引入多视角聚合策略,丰富不同视角下对图像的描述能力,可以很好地处理多模态数据的复杂性与多样性。文献[25]在解决多视角描述问题上取得卓越成果,因此本研究采用与其类似的多视角聚合策略。给定区域级图像特征 $\bar{\boldsymbol{V}}\in\mathbf{R}^{r\times D}$,通过多视角总结模块 F_{mvs} 生成 N 个不同视角的表征,转换后的特征向量

$$\tilde{\boldsymbol{V}}=F_{\text{mvs}}(\bar{\boldsymbol{V}})\in\mathbf{R}^{N\times D}。\tag{3}$$

为了保证整个过程中特征一致,沿用文献[29]中的方法获取文本特征,即将多头注意力应用于文本序列 $\boldsymbol{T}=[\boldsymbol{t}_1\ \ \boldsymbol{t}_2\ \ \cdots\ \ \boldsymbol{t}_L]$,通过一个平均池化操作得到区域级文本特征 $\tilde{\boldsymbol{t}}$。

在对齐过程中,计算每个视角与对应文本特征间的相似度分数。选取每一对图像-文本对相似度最高的分数作为该对的最终对齐分数 S_{region},提升语义对齐精确度。S_{region} 的计算式为

$$S_{\text{region}}=\max_{i=1,2,\cdots,N}\cos(\tilde{\boldsymbol{v}}_i,\tilde{\boldsymbol{t}}),\tag{4}$$

式中,$\cos(\cdot,\cdot)$ 为向量间的余弦相似度,$\tilde{\boldsymbol{v}}_i$ 为第 i 个视角的图像特征。

2.3.2 图像块级对齐

在图像块级,将图像特征与文本特征之间的细节信息精确对齐是一个难题,涉及选择与文本内容相关的图像块及捕捉其与单词之间的内在关联。

图像块的选择关键在于识别与文本内容最相关的部分,防止不相关内容干扰对齐过程。受文献[31]中方法的启发,设计一个基于 MLP 的图像块选择器 F_{selector},可以精确识别图像特征集中相关性最高的 K 个图像块。被选中的图像块级特征

$$\tilde{\boldsymbol{P}}=F_{\text{selector}}(\bar{\boldsymbol{P}})\in\mathbf{R}^{K\times D}。\tag{5}$$

通过该选择模块可以有效减少不必要的冗余及噪声干扰,降低所需计算资源。

细化图像块与单词的交互层面,计算图像块级特征 $\tilde{\boldsymbol{P}}$ 中每一行向量与 $\boldsymbol{T}=[\boldsymbol{t}_1\ \ \boldsymbol{t}_2\ \ \cdots\ \ \boldsymbol{t}_L]$ 中每个单词可能组合的余弦相似度。相似度矩阵

$$\boldsymbol{C}_{\text{PW}}=[[\cos(\tilde{\boldsymbol{p}}_k,\boldsymbol{t}_i)]_{k=1}^{K}]_{i=1}^{L}\in\mathbf{R}^{K\times L},\tag{6}$$

式中 $\tilde{\boldsymbol{p}}_k$ 为单个图像块特征。为了更精确地建模图像块与单词间的交互关系,仅使用余弦相似度可能不足以捕捉模态内的细节交互信息。$\boldsymbol{C}_{\text{PW}}$ 的每行描绘特定图像块与所有单词的关系,每列则表达每个单词与所有图像块的关联。为了深入挖掘交互信息并提高对模态间数据的理解,设计一个专用于加强图像块-单词交互的相似性交互模块。

相似性交互模块建模图像块与单词之间的交互过程如图 2 所示。将 $\boldsymbol{C}_{\text{PW}}$ 的每一行和每一列分解为独立的相似性向量,应用相似性交互模块聚合相似度分数,形成新的向量。

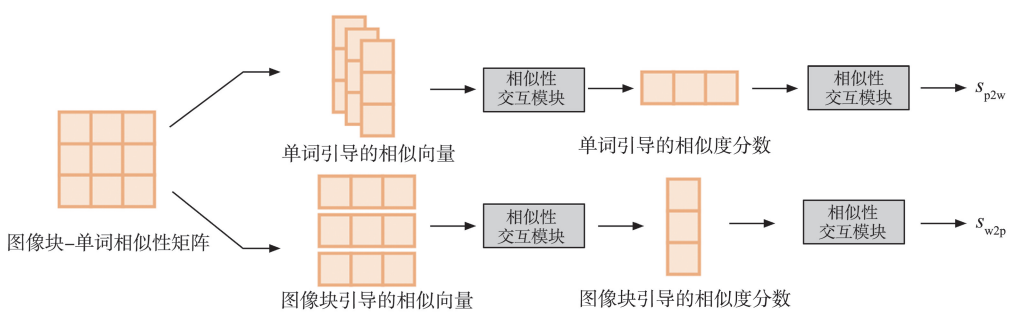


图 2 相似性交互模块

Fig.2 The similarity interaction modeling module

在单词到图像块的方向上,模型将一个 Softmax 函数和一个线性层 Linear_1 应用于 $\boldsymbol{C}_{\mathrm{PW}}$ 的行向量 $\boldsymbol{c}_{\mathrm{p}}^k \in \mathbf{R}^L$,通过另一个 Softmax 函数计算相似度分数权重 $\boldsymbol{\alpha}_{\mathrm{p}}^k$,即 $\boldsymbol{\alpha}_{\mathrm{p}}^k = \mathrm{Softmax}(\mathrm{Linear}_1(\mathrm{Softmax}(\boldsymbol{c}_{\mathrm{p}}^k)))$,得到每个图像块分数

$$s_{\mathrm{p}}^k = (\boldsymbol{\alpha}_{\mathrm{p}}^k)^{\mathrm{T}} \boldsymbol{c}_{\mathrm{p}}^k \circ \tag{7}$$

将计算后的图像块分数 $\boldsymbol{S}_{\mathrm{p}} = [\, s_{\mathrm{p}}^1 \quad s_{\mathrm{p}}^2 \quad \cdots \quad s_{\mathrm{p}}^k \,]$ 送到另一个相似性交互模块,计算单词到图像块方向的匹配分数

$$s_{\mathrm{w2p}} = \boldsymbol{\beta}_{\mathrm{p}}^{\mathrm{T}} \boldsymbol{S}_{\mathrm{p}}, \tag{8}$$

式中,$\boldsymbol{\beta}_{\mathrm{p}}$ 为单词到图像块方向的相似度分数权重,$\boldsymbol{\beta}_{\mathrm{p}} = \mathrm{Softmax}(\mathrm{Linear}_2(\mathrm{Softmax}(\boldsymbol{S}_{\mathrm{p}})))$。

通过上述过程,模型能够识别和量化从单词到图像块的对应关系,以更丰富、更细致的视角理解二者之间的相互关系。

在图像块到单词的方向上,模型处 $\boldsymbol{C}_{\mathrm{PW}}$ 的列向量 $\boldsymbol{c}_{\mathrm{w}}^l \in \mathbf{R}^K$,采用与单词到图像块方向相同的 Softmax 和线性层处理策略计算相似度权重 $\boldsymbol{\alpha}_{\mathrm{w}}^l$,$\boldsymbol{\alpha}_{\mathrm{w}}^l = \mathrm{Softmax}(\mathrm{Linear}_3(\mathrm{Softmax}(\boldsymbol{c}_{\mathrm{w}}^l)))$,得到每个单词分数

$$s_{\mathrm{w}}^l = (\boldsymbol{\alpha}_{\mathrm{w}}^l)^{\mathrm{T}} \boldsymbol{c}_{\mathrm{w}}^l \circ \tag{9}$$

将得到的单词分数 $\boldsymbol{S}_{\mathrm{w}} = [\, s_{\mathrm{w}}^1 \quad s_{\mathrm{w}}^2 \quad \cdots \quad s_{\mathrm{w}}^L \,]$ 送到另一个相似性交互模块,提取图像块到单词方向的匹配分数

$$s_{\mathrm{p2w}} = \boldsymbol{\beta}_{\mathrm{w}}^{\mathrm{T}} \boldsymbol{S}_{\mathrm{w}}, \tag{10}$$

式中,$\boldsymbol{\beta}_{\mathrm{w}}$ 为图像块到单词方向的相似度分数权重,$\boldsymbol{\beta}_{\mathrm{w}} = \mathrm{Softmax}(\mathrm{Linear}_4(\mathrm{Softmax}(\boldsymbol{S}_{\mathrm{w}})))$。

通过上述过程,模型获得从图像块到单词关系的识别和量化能力,丰富了彼此之间信息关系的理解与学习。

为了综合利用模型从图像块到单词和从单词到图像块两个方向上学习的关系理解,采用均值融合方法计算图像块级别的最终匹配分数 S_{patch},$S_{\mathrm{patch}} = (s_{\mathrm{p2w}} + s_{\mathrm{w2p}})/2$。

2.3.3　全局级对齐

全局级对齐重点在于利用[CLS]标识特征进行图像和文本之间的整体语义对齐。在对齐之前,为了简化后续相似性计算过程,确保特征在同一标准下进行对比,对 $\boldsymbol{P}_{\mathrm{cls}}^*$ 和 $\boldsymbol{t}_{\mathrm{cls}}^*$ 用标准化函数 L_2 进行处理,得到图像与相应文本之间的全局相似性分数

$$S_{\mathrm{global}} = \cos(\boldsymbol{P}_{\mathrm{cls}}^*, \boldsymbol{t}_{\mathrm{cls}}^*) \circ \tag{11}$$

通过计算全局级特征之间的余弦相似度,捕捉图像与文本之间的整体语义一致性,提升模型对全局语义关系的理解。

将来自全局级、区域级和图像块级的相似性分数聚合,综合考虑各级别对齐对整体匹配具有不同的影响,将通过平衡参数 α、β 和 λ 调整各级别对齐的贡献,计算出最终匹配分数

$$S = \alpha S_{\mathrm{region}} + \beta S_{\mathrm{global}} + \lambda S_{\mathrm{patch}} \circ \tag{12}$$

2.4　损失函数

本研究将多样性正则化损失 $L_{\mathrm{Diversity}}$ 和排名损失 L_{Rank} 组合,得到最终损失

$$L = L_{\mathrm{Diversity}} + \lambda_2 L_{\mathrm{Rank}}, \tag{13}$$

式中 λ_2 为平衡参数,用于调和多样性正则化和排名损失之间的比重。排名损失 L_{Rank} 的目标是加强图像-文本对 $(\boldsymbol{I}, \boldsymbol{T})$ 间的匹配,特别是强化难以区分的负样本。该过程的计算式为

$$L_{\mathrm{Rank}} = [\, \alpha_2 - S(\boldsymbol{I}, \boldsymbol{T}) + S(\boldsymbol{I}, \hat{\boldsymbol{T}}) \,]_+ + [\, \alpha_2 - S(\boldsymbol{I}, \boldsymbol{T}) + S(\hat{\boldsymbol{I}}, \boldsymbol{T}) \,]_+, \tag{14}$$

式中:α_2 为边缘参数;$S(\cdot, \cdot)$ 为图像-文本对之间的相似度;$\hat{\boldsymbol{I}}$ 为批样本中最难分辨的图像负样本,$\hat{\boldsymbol{I}} = \underset{\boldsymbol{i} \neq \boldsymbol{I}}{\mathrm{argmax}}\, S(\boldsymbol{i}, \boldsymbol{T})$,其中 $\boldsymbol{i}$ 为图像;$\hat{\boldsymbol{T}}$ 为批样本中最难分辨的文本负样本,$\hat{\boldsymbol{T}} = \underset{\boldsymbol{j} \neq \boldsymbol{T}}{\mathrm{argmax}}\, \boldsymbol{S}(\boldsymbol{I}, \boldsymbol{j})$,其中 $\boldsymbol{j}$ 为文本。

为了确保多视角描述下特征的多样性,减少不必要的重复,引入文献[25]中的 $L_{\mathrm{Diversity}}$,在维持多角度图像对齐的同时,最小化不必要的冗余。这种正则化有助于在各种视角中实现表征平衡,保持模型的有效性和效率。

3　试验分析

为了验证 MGAN 模型的有效性,本研究进行大量试验,对于图像-文本匹配任务,有两种方向上的匹配任务:图像到文本的匹配,即从给定图像中匹配符合描述的句子;文本到图像的匹配,即从给定文本描述中找出与之最匹配的图像。

3.1　数据集

(1)Flickr30K 数据集[32]。包含 31 783 张来自 Flickr 平台的图像,每张图像配有 5 个描述性句子,遵循各类方法都普遍使用的分割方式,将 1 000 张图像作为验证和测试,其余用于训练。

(2)MS-COCO 数据集[33]。包含 123 287 张图像,与 Flickr30K 类似,每张图像配有 5 个手动标注的句子,按照常用的划分方式,选取 113 287 张图像用于训练,两组各 5 000 张图像用于验证和测试。评估在两种设置下进行:MS-COCO 1K,平均划分出 5 个不同子集,每个子集包含 1 000 张测试图像;MS-COCO 5K,对完整的 5 000 张测试图像进行评估。

3.2 评估指标

性能评估指标采用信息检索领域广泛引用的召回率 R_K(其中 $K=1, 5, 10$),衡量在前 K 个位置上查询结果的准确性和相关性。

3.3 试验细节

3.3.1 优化器与训练配置

模型采用 Adam 优化器,样本批量大小设置为128,初始学习率为 0.000 1。对 Flickr30K 数据集的模型训练进行 30 个轮次,每 10 个轮次更新一次学习率;对 MS-COCO 数据集的训练进行 40 个轮次,每 20 个轮次调整一次学习率。

3.3.2 特征向量与模型框架选择

图像和文本特征向量投影至 2 048 维的公共空间,以平衡计算和保留语义信息。基于 CLIP-ViT 模型采用 ViT-B/32 的版本。区域级特征提取过程使用的视角数量设置为 12,以确保图像数据的全面性。

3.3.3 文本编码器

采用 BERT 模型,对最后一层进行微调,增强对文本语义的理解。为了能够直观比较,将没有解冻 BERT 的方法称为 MGAN*,解冻了最后一层的完整方法称为 MGAN。

3.3.4 参数配置

式(12)中 α、β 和 λ 分别设定为 1.0、0.6 和 0.2,式(13)中 λ_2 设为 0.01,式(14)中 α_2 设为 0.2,以平衡模型目标函数中各组成部分的影响。

3.4 性能对比

为展示 MGAN 模型的有效性,本研究将其与一系列先进的基线模型进行比较,包括堆叠交叉注意力(stacked cross attention, SCAN)模型[20]、跨模态自适应消息传递(cross-modal adaptive message passing, CAMP)模型[14]、视觉语义推理(visual semantic reasoning, VSRN)模型[17]、多模态交叉注意力网络(multi-modality cross attention network, MMCA)[15]、图结构化匹配网络(graph structured matching network, GSMN)[12]、双语义关系注意网络(dual semantic relations attention network, DSRAN)[34]、上下文感知多视图摘要网络(context-aware multi-view summarization network, CAMERA)[25]、相似图推理与过滤(similarity graph reasoning and filtration, SGRAF)模型[35]、跨层次一致性的概念语法跨模态对齐(conceptual and syntactical cross-modal alignment with cross-level consistency, CSCC)模型[36]、视觉语义推理++(visual semantic reasoning++, VSRN++)模型[18]、基于图结构的双模态表示(graph-based dual-modal representation, GraDual)模型[37]、生成标签融合网络(generative label fused network, GLFN)[38]、基于注意力与外部知识嵌入的双向生成网络(bidirectional knowledge-assisted embedding and attention-based generation, BiKA)[39]、基于递归对应和聚合调节器(recurrent correspondence and aggregation regulator, RCAR)模型[40]。这些基线模型的试验结果直接从原论文中引用。考虑部分基线模型采用集成方法(即独立训练两个模型并对结果进行平均化),为了试验的公平与严谨性,本研究采用类似的集成策略更准确地评估 MGAN 与其他先进方法的性能对比。

在 Flickr30K 数据集上不同方法的试验结果如表 1 所示,其中最优结果加粗表示。由表 1 可知:相较于 MGAN*,MGAN 展现出更卓越的性能,表明当 BERT 模型的最后一层被解冻并微调时,BERT 强大的自适应学习能力对提升匹配性能起到关键作用;与 GLFN、BiKA 和 RCAR 相比,MGAN 在性能上仍然具有明显优势;在图像到文本和文本到图像的匹配任务中,MGAN 在所有评估指标上均超越之前的最佳结果。上述试验结果验证了引入图像块级信息和相似性交互模块能够提升整个匹配任务的有效性。

表 1 Flickr30K 数据集上的性能对比

Table 1 Performance comparison on the Flickr30K

单位:%

模型	图像-文本方向			文本-图像方向		
	R_1	R_5	R_{10}	R_1	R_5	R_{10}
SCAN[20]	67.4	90.3	95.8	48.6	77.7	85.2
CAMP[14]	68.1	89.7	95.2	51.5	77.1	85.3
VSRN[17]	71.3	90.6	96.0	54.7	81.8	88.2
MMCA[15]	74.2	92.8	96.4	54.8	81.4	87.8
GSMN[12]	76.4	94.3	97.3	57.4	82.3	89.0
DSRAN[34]	77.8	95.1	97.6	59.2	86.0	91.9
CAMERA[25]	78.0	95.1	97.9	60.3	85.9	91.7
SGRAF[35]	77.8	93.4	96.5	61.2	86.7	91.5

表 1(续)

模型	图像-文本方向			文本-图像方向		
	R_1	R_5	R_{10}	R_1	R_5	R_{10}
CSCC[36]	78.8	96.1	97.6	60.5	86.1	91.3
VSRN++[18]	79.2	94.6	97.5	60.6	85.6	91.4
GraDual[37]	78.3	96.0	98.0	60.4	86.7	92.0
GLFN[38]	75.1	93.8	97.2	54.5	82.8	89.9
BiKA[39]	75.2	91.6	97.4	54.8	82.5	88.6
RCAR[40]	78.7	94.6	97.6	59.5	84.0	89.5
MGAN*	81.2	96.6	98.7	63.5	88.0	94.0
MGAN	**84.7**	**97.4**	**99.3**	**68.9**	**91.1**	**95.4**

MGAN 在 MS-COCO 数据集的两种设置下与其他基线方法的性能对比如表 2、3 所示,其中最优结果加粗表示。由表 2、3 可知:在 MS-COCO 数据集上,多数方法的性能通常低于在 Flickr30K 数据集上的表现,可能归因于 Flickr30K 和 MS-COCO 数据集在构建上存在一定的差异性,Flickr30K 数据集特别强调详细和精确的标注,图像通常聚焦于特定的场景或活动,为每一幅图像提供丰富形象的描述,MS-COCO 数据集的语义信息较为复杂,模型理解语义更加困难;尽管 MS-COCO 数据集的图像更加复杂和抽象,无论在图像到文本还是文本到图像的匹配任务下,MGAN 都能在 1K 和 5K 设置中展现出较高的性能,表明 MGAN 在处理不同规模和复杂度数据集时具有强大的稳定性及学习能力;与 GLFN、BiKA、RCAR 对比,解冻并微调 BERT 之后的 MGAN 实现了最高的匹配准确性,证明在图像-文本匹配任务中引入图像块级信息和深度微调语言模型具有独特的优势,能有效提升模型语义理解和匹配性能。

表 2　MS-COCO 1K 数据集上的性能对比

Table 2　Performance comparison on the MS-COCO 1K　　单位:%

模型	图像-文本方向			文本-图像方向		
	R_1	R_5	R_{10}	R_1	R_5	R_{10}
SCAN[20]	72.7	94.8	98.4	58.8	88.4	94.8
CAMP[14]	72.3	94.8	98.3	58.5	87.9	95.0
VSRN[17]	76.2	94.8	98.2	62.8	89.7	95.1
MMCA[15]	74.8	95.6	97.7	61.6	89.8	95.2
GSMN[12]	78.4	96.4	98.6	63.3	90.1	95.7
DSRAN[34]	78.3	95.7	98.4	64.5	90.8	95.8
CAMERA[25]	77.5	96.3	98.8	63.4	90.9	95.8
SGRAF[35]	79.6	96.2	98.5	63.2	90.7	96.1
CSCC[36]	78.8	96.1	99.0	66.6	92.5	96.4
VSRN++[18]	77.9	96.0	98.5	64.1	91.0	96.1
GraDual[37]	77.0	96.4	98.6	65.3	91.9	96.4
GLFN[38]	78.4	96.0	98.5	62.6	89.6	95.4
BiKA[39]	77.6	96.5	98.6	62.8	90.3	95.8
RCAR[40]	80.6	96.6	98.6	64.1	90.5	95.8
MGAN*	80.8	97.6	99.2	64.5	91.8	96.1
MGAN	**83.4**	**97.6**	**99.3**	**67.2**	**92.3**	**96.9**

表 3　MS-COCO 5K 数据集上的性能对比

Table 3　Performance comparison on the MS-COCO 5K　　单位:%

模型	图像-文本方向			文本-图像方向		
	R_1	R_5	R_{10}	R_1	R_5	R_{10}
SCAN[20]	50.4	82.2	90.0	38.6	69.3	80.4
CAMP[14]	50.1	82.1	89.7	39.0	68.9	80.2
VSRN[17]	53.0	81.1	89.4	40.5	70.6	81.1
MMCA[15]	54.0	82.5	90.7	38.7	69.7	80.8
GSMN[12]	—	—	—	—	—	—
DSRAN[34]	55.3	83.5	90.9	41.7	72.7	82.8

表 3(续)

模型	图像-文本方向			文本-图像方向		
	R_1	R_5	R_{10}	R_1	R_5	R_{10}
CAMERA[25]	55.1	82.9	91.2	40.5	71.7	82.5
SGRAF[35]	57.8	—	91.6	41.9	—	81.3
CSCC[36]	55.6	83.6	91.2	40.8	73.2	84.3
VSRN++[18]	54.7	82.9	90.9	42.0	72.2	82.7
GraDual[37]	—	—	—	—	—	—
GLFN[38]	—	—	—	—	—	—
BiKA[39]	54.5	83.4	91.4	40.2	70.9	80.7
RCAR[40]	59.6	85.8	92.4	42.5	71.7	81.8
MGAN*	57.7	86.9	92.6	41.7	71.9	82.4
MGAN	**61.8**	**87.5**	**93.1**	**44.7**	**75.3**	**85.2**

注:“—”表示该模型没有进行 MS-COCO 5K 设置的试验。

3.5 消融试验

为了分析和理解每个对齐级别在整体匹配中的作用和重要性,本研究通过有选择地组合不同的对齐,构建出不同的模型变体,设计一系列消融试验。在具体试验设置中,构建 3 种主要的模型变体,每种变体分别缺少一种对齐级别:区域级与全局级对齐组合,旨在探索在没有细粒度图像块级别对齐时,模型能否依然有效地捕获图像与文本间的关联;区域级与图像块级对齐组合,测试在缺乏全局视角时模型性能的变化,以此评估全局信息对理解图像-文本整体意义的重要性;全局级与图像块级对齐组合,了解在不考虑区域级别信息时,全局和局部信息对上下文信息的依赖程度。为了验证本研究提出的相似性交互模块的有效性,用标准的聚合方法代替该模块,构建出另外一个模型变体。消融试验的结果如表 4 所示,其中最优结果加粗表示。

表 4 消融试验性能比较

Table 4 Performance comparison in ablation study 单位:%

模型	图像-文本方向			文本-图像方向		
	R_1	R_5	R_{10}	R_1	R_5	R_{10}
MGAN*	**81.2**	**96.6**	**98.7**	**63.5**	**88.0**	**94.0**
区域级与全局级	75.1	94.8	97.3	58.7	83.8	90.0
区域级与图像块级	76.4	94.2	96.7	54.4	81.5	89.2
全局级与图像块级	75.6	93.7	96.6	56.9	84.3	91.1
消去相似性交互模块	78.8	95.7	98.0	60.0	86.7	92.5

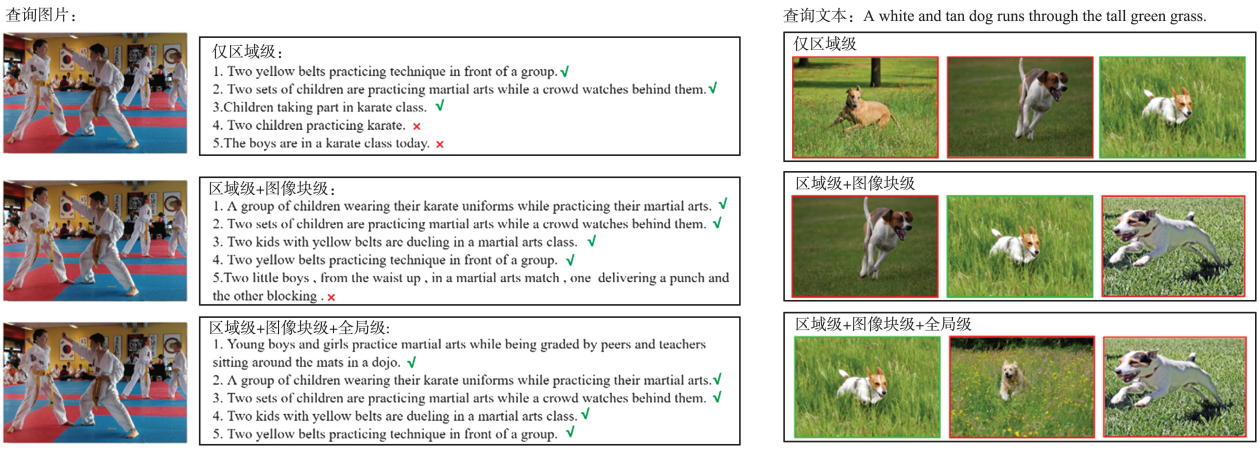
由表 4 可以看出:包含所有 3 个粒度级别(区域级、图像块级和全局级)的 MGAN* 模型在各项评估指标上均表现最优,证明在提升图像-文本匹配性能中区域级、图像块级和全局级特征信息之间互补的必要性;通过比较仅包含 2 个级别对齐的模型变体,可以明显看出每种级别粒度对整体性能的贡献,区域级与图像块级的组合及全局级与图像块级的组合在性能上表现相对更为接近,表明图像块级在匹配过程中的重要性;任何 2 个级别的组合都无法超过 MGAN* 的性能,进一步证明整合 3 个级别特征信息的重要性;当缺少相似性交互模块时,模型的匹配性能表现不佳,验证了该模块在捕捉模态内交互方面的卓越性,在提升匹配性能中起到关键作用。

3.6 结果可视化

为了更直观地展现 MGAN 模型的性能及不同粒度对齐对匹配过程的具体影响,基于不同对齐级别的匹配结果可视化如图 3 所示。

基于特定图像查询描述语句的结果如图 3(a)所示。当模型采用图像块级细粒度对齐和区域级基础对齐时,能够识别出复杂的描述细节,如空手道服和黄带;在加入全局级对齐后,模型的理解能力进一步增强,可以识别出如道场的整体上下文线索,表明多粒度对齐在理解复杂场景中的有效性。

基于特定文本描述匹配图像时的前 3 个最高排名结果如图 3(b)所示。模型匹配到的图像与查询语句在语义上高度匹配。使用仅区域级对齐的模型在识别图像的细节属性上存在不足,比如忽视了小狗的白色外观特征;当引入图像块级对齐后,这一问题得到明显改善,模型能够匹配到与文本描述精确匹配的图像;在进一步与全局级对齐结合后,模型的性能再一次得到增强,能够捕捉如高草丛的整体场景背景。这些结果验证了 MGAN 在利用多粒度信息上的高效性和优越性。



(a) 图像到文本的匹配　　(b) 文本到图像的匹配

图 3　定性试验结果

Fig.3　Qualitative results

综上所述,通过详细分析图 3 中的可视化结果可以得出:MGAN 模型在图像-文本匹配任务中可以有效利用多粒度信息对齐策略提高它在匹配任务中的性能。

4　结论

本研究提出一种新的图像-文本匹配方法 MGAN,整合不同粒度级别的信息,与现有方法仅侧重区域级或全局级信息形成显著对比。采用 CLIP 模型提取全局和局部图像特征,设计一个相似性交互模块,以增强特征之间的交互和协同作用。在 Flickr30K 和 MS-COCO 这两个重要的数据集上进行大量试验,MGAN 展现出卓越的性能,超越多个现有模型。但是弥补模态间的语义差距依旧是一个挑战。利用 CLIP 及 BERT 虽然可以极大地提升性能,但 MGAN 的训练速度有些不尽人意。未来将更深入地研究多粒度信息的细节差别,挖掘出多粒度信息的全部潜力,继续完善技术,更好地弥合不同模态之间的语义差距,确保不同模态的信息之间更加细致准确地对齐。优化算法逻辑以实现更高速的训练也是未来的主要研究课题之一。

参考文献:

[1] WEI Y, WANG X, GUAN W, et al. Neural multimodal cooperative learning toward micro-video understanding [J]. IEEE Transactions on Image Processing, 2019, 29: 1-14.

[2] HU Y, ZHAN P, XU Y, et al. Temporal representation learning for time series classification[J]. Neural Computing and Applications, 2021, 33: 3169-3182.

[3] WEI Y, WANG X, NIE L, et al. MMGCN: multi-modal graph convolution network for personalized recommendation of micro-video[C]//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM, 2019: 1437-1445.

[4] CHEN H, DING G, LIN Z, et al. Cross-modal image-text retrieval with semantic consistency[C]//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM, 2019: 1749-1757.

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

[6] FROME A, CORRADO G S, SHLENS J, et al. DeViSE: a deep visual-semantic embedding model[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. Red Hook, USA: ACM, 2013: 2121-2129.

[7] KIROS R, SALAKHUTDINOV R, ZEMEL R S. Unifying visual-semantic embeddings with multimodal neural languagemodels[EB/OL]. (2014-11-10)[2024-01-31]. https://arxiv.org/abs/1411.2539

[8] LIU Y, GUO Y, BAKKER E M, et al. Learning a recurrent residual fusion network for multimodal matching[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 4107-4116.

[9] WANG L, LI Y, LAZEBNIK S. Learning deep structure-preserving image-text embeddings[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 5005-5013.

[10] SARAFIANOS N, XU X, KAKADIARIS I A. Adversarial representation learning for text-to-image matching[C]// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 5814-5824.

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

[12] LIU C, MAO Z, ZHANG T, et al. Graph structured network for image-text matching[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 10921-10930.

[13] HUANG F, ZHANG X, ZHAO Z, et al. Bi-directional spatial-semantic attention networks for image-text matching[J]. IEEE Transactions on Image Processing, 2018, 28(4): 2008-2020.

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

[15] WEI X, ZHANG T, LI Y, et al. Multi-modality cross attention network for image and sentence matching[C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 10941-10950.

[16] WANG B, YANG Y, XU X, et al. Adversarial cross-modal retrieval[C]// Proceedings of the 25th ACM International Conference on Multimedia. Mountain View, USA: ACM, 2017: 154-162.

[17] LI K, ZHANG Y, LI K, et al. Visual semantic reasoning for image-text matching[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 4654-4662.

[18] LI K, ZHANG Y, LI K, et al. Image-text embedding learning via visual and textual semantic reasoning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(1): 641-656.

[19] WANG Y, YANG H, QIAN X, et al. Position focused attention network for image-textmatching[EB/OL]. (2019-07-23)[2024-01-31]. https://arxiv. org/abs/1907.09748

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

[21] DENG Y, ZHANG F, CHEN X. Collaborative attention network model for cross-modal retrieval[J]. Computer Science, 2020, 47(4): 54-59.

[22] CHEN T, LUO J. Expressing objects just like words: recurrent visual embedding for image-text matching[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020: 10583-10590.

[23] ZHANG J, HE X, QING L, et al. Cross-modal multi-relationship aware reasoning for image-text matching[J]. Multimedia Tools and Applications, 2022, 81: 12005-12027.

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

[25] 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, USA: ACM, 2020: 1047-1055.

[26] 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, 31(12): 5412-5425.

[27] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: ACM, 2017: 6000-6010.

[28] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional Transformers for language understanding[EB/OL]. (2019-05-24)[2024-01-31]. https://arxiv.org/abs/1810.04805

[29] SUN C, MYERS A, VONDRICK C, et al. VideoBERT: a joint model for video and language representation learning[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 7464-7473.

[30] 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. Vienna, Austria: ICML, 2021: 8748-8763.

[31] LIU Y, XIONG P, XU L, et al. TS2-Net: token shift and selection Transformer for text-video retrieval[C]// Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022: 319-335.

[32] 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.

[33] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014: 740-755.

[34] WEN K, GU X, CHENG Q. Learning dual semantic relations with graph attention for image-text matching [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(7): 2866-2879.

[35] DIAO H, ZHANG Y, MA L, et al. Similarity reasoning and filtration for image-text matching[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2021: 1218-1226.

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

[37] LONG S, HAN S C, WAN X, et al. GraDual: graph-based dual-modal representation for image-text matching [C]//Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE, 2022: 3459-3468.

[38] ZHAO G, ZHANG C, SHANG H, et al. Generative label fused network for image-text matching[J]. Knowledge-Based Systems, 2023, 263: 110280.

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

[40] DIAO H, ZHANG Y, LIU W, et al. Plug-and-playregulators for image-text matching[J]. IEEE Transactions on Image Processing, 2023, 32: 2322-2334.

(编辑:孙亚彤)

(上接第 28 页)

[30] IIZUKA S, SIMO-SERRA E, ISHIKAWA H. Globally and locally consistent image completion[J]. ACM Transactions on Graphics, 2017, 36(4): 1-14.

[31] LI C, WAND M. Precomputed real-time texture synthesis with Markovian generative adversarial networks [C]//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, Netherlands: Springer, 2016: 702-716.

[32] LIU G, REDA F A, SHIH K J, et al. Image inpainting for irregular holes using partial convolutions[C]// Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018: 89-105.

[33] DENG Y, HUI S, MENG R, et al. Hourglass attention network for image inpainting[C]//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022: 483-501.

(编辑:孙亚彤)