基于文本生成与迭代匹配的图像-文本检索

潘莹莹[1)], 马青[2)*], 白琮[1)]

[1)] (浙江工业大学计算机科学与技术学院 杭州 310023)
[2)] (浙江工业大学理学院 杭州 310023)
(maqing@zjut.edu.cn)

摘 要: 跨模态图文检索由于图像和文本的特征表示方式不同面临着模态异构问题, 传统的公共空间方法难以度量图像和文本的相似性. 为此, 提出了基于文本生成与迭代匹配的跨模态图像文本检索框架, 它包含了特征融合模块和文本生成模块. 特征融合模块通过迭代融合的方式, 多次对齐图像和文本, 在不同的迭代步骤中聚合细粒度信息, 捕获图像和文本之间的局部关联信息, 优化了局部公共嵌入空间; 文本生成模块采用特征转换的思路, 由图像模态中的特征映射到文本模态中的句子特征, 通过图文信息交互增强了图像和文本的整体语义相关性, 优化了全局公共嵌入空间, 挖掘出图像与文本更深层的语义信息, 以提高跨模态图像文本检索模型的性能. 在 Flickr30K 和 COCO 数据集上进行了实验, 并与现有的模型进行比较, 结果表明, 该框架在 Flickr30K 和 COCO 上的整体性能分别提升了 0.7%和 1.2%. 在文本检索任务的召回指标中, 最高可以提升 3.4%; 在图像检索任务的召回指标中, 最高可以提升 4.6%. 消融实验也证明了其中特征融合模块以及文本生成模块的有效性.

关键词: 跨模态图文检索; 文本生成; 公共空间; 特征融合; 迭代匹配
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023-00363

Iterative Matching with Text Generation for Cross-Modal Image-Text Retrieval

Pan Yingying[1)], Ma Qing[2)*], and Bai Cong[1)]

[1)] (*College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou* 310023)
[2)] (*College of Science, Zhejiang University of Technology, Hangzhou* 310023)

Abstract: Cross-modal image-text retrieval faces the problem of modal heterogeneity due to the different feature representations of images and texts, and it is difficult for traditional public space methods to measure the similarity of images and texts. Therefore, a cross-modal image-text retrieval framework based on text generation and iterative matching is proposed, which includes a feature fusion module and a text generation module. The feature fusion module which can optimizes the local public embedding space aligns images and texts multiple times through iterative fusion, aggregates fine-grained information in different iterative steps, and captures local association information between images and texts. The text generation module adopts the idea of feature transformation, which maps the features in the image modal to the sentence features in the text mode, enhances the overall semantic correlation of image and text through image and text information interaction. To improve the performance of cross-modal image-text retrieval model, the text generation module not only optimizes the global public embedding space, but also excavates deeper semantic information of images and texts. Comparing with existing models, the experiments on Flickr30K and COCO datasets show that the overall performance of the framework on Flickr30K and COCO is improved by 0.7% and

收稿日期: 2023-06-20; 修回日期: 2023-08-12. **基金项目**: 国家自然科学基金(62102365, 61976192); 浙江省杰出青年科学基金(LR21F020002). **潘莹莹**(1998—), 女, 硕士研究生, 主要研究方向为跨模态检索; **马青**(1982—), 女, 博士, 讲师, 论文通信作者, 主要研究方向为跨模态检索; **白琮**(1981—), 男, 博士, 教授, 博士生导师, CCF 会员, 主要研究方向为多媒体信息处理与智慧气象.

1.2%, respectively. In the recall index of text retrieval tasks, it can be increased by up to 3.4%. In the recall index of image retrieval tasks, it can be improved by up to 4.6%. The ablation experiments also prove the effectiveness of the feature fusion module and the text generation module.

Key words: cross-modal image-text retrieval; text generation; public spaces; feature fusion; iterative matching

随着大数据时代的到来, 文本、图像和视频等不同模式的数据正在以前所未有的速度增长, 图像和文本呈现出底层特征异构、而高层语义相关的特点, 异构数据的匹配需求使得图文检索成为一个新的需求. 在这种情况下, 如何解决图像-文本间的模态异构性, 进而关联大量的跨模态信息是当前图文检索任务面临的巨大挑战. 跨模态特征融合与特征转换是实现图像-文本信息关联的重要途径. 这 2 种方法都针对图像和文本的模态异构性, 增强图像和文本特征之间的交互, 同时构建图像和文本之间的信息关联, 进而挖掘它们之间的语义一致性. 因此, 图像-文本特征融合与特征转换模型已经得到了充分的关注和广泛研究.

一些研究[1-5]致力于图文特征融合, 通过全局或是局部的特征融合得到更多对齐的图文特征信息, 从而达到优化公共空间的目的; 还有一些研究[6-11]致力于图文特征转换, 不仅研究图像和文本的对齐, 还探索图文数据的特征转换, 将图像和文本序列映射到同一公共空间, 使 2 种模态语义相近的对象尽量靠近, 从而应用于图像字幕生成和跨模态检索等任务. 虽然这些特征融合与特征转换的方法已经能学习图像-文本的关联语义信息, 但尚无法挖掘出具有细粒度相似性的图像或具有单词级相似性的句子, 因此在复杂语义的理解上具有局限性.

事实上, 人在观察一幅图像时, 第一眼得到的信息通常都会是一些低级语义, 如物体和对象; 再通过不断地观察, 可以挖掘出图像中的更高层语义, 如关系和情感等; 最后, 人会在脑海中生成一段对于这幅图像相关的文字描述. 如果借鉴这种人类的视觉习惯, 将一种模态的信息表示到另一种模态上, 有利于学习一个更好的公共空间表示.

受此概念的启发, 本文提出了面向图像-文本检索的基于文本生成与迭代匹配(iteration matching with text-generation, IMTG)的框架, 通过模拟人类视觉的 3 个习惯, 实现了“观察−联想−匹配”的过程. IMTG 主要包括特征融合模块和文本生成模块 2 个核心模块. 特征融合模块就类比于“观察”行为, 采取迭代的方式进行图像文本特征融合, 多次对齐图像和文本, 在不同的迭代步骤中聚合细粒度信息, 很好地捕获图像和文本之间的局部关联信息. 而文本生成模块则类比于“联想”行为, 由图像模态中的局部特征映射到文本模态中的句子特征, 通过图文信息交互增强了图像和文本的整体语义相关性. 最后, “匹配”行为通过将生成模块引入文本-图像特征嵌入中, 实现图像文本的匹配与检索.

1　相关方法介绍

1.1　图像-文本特征融合方法

现有的图文特征融合方法大多数都致力于探索跨模态图像文本中潜在的细粒度视觉-语言对应关系, 从而优化特征公共空间. Kim 等[12]提出了最简单的多模态 Transformer 方法: ViLT, 它使用预训练的 ViT[13]来初始化交互的 Transformer, 直接对图像-文本特征进行交互融合. 而 Radford 等[14]使用大规模的文本-图像配对预训练、提出 CLIP 方法, 突破了文本-图像之间的限制, 并且可以将其直接迁移到 ImageNet 上, 完全不需要图像标签微调即可实现 zero-shot 分类. Andrej 等[9]为每幅图像和每段文本提取了局部的特征(即图像块和文本单词), 并在公共空间中对齐, 提出了每个片段对之间的对齐方法. Lee 等[2]提出了一种堆叠交叉注意模型, 将每个小片段与来自另一种模式的所有小片段对齐, 从而增强了局部对齐和全局对齐. Huang 等[3]将注意机制引入到图文嵌入中, 设计了一种注意力方案, 可以选择性地关注图像和文本对, 然后动态聚合图像-文本对之间的相似度, 得到图像和文本之间的全局相似度. 然而, 由于上述方法忽视了图像和文本的语义差异, 因此不能抓住图像块-单词对之间的最优成对关系. 针对上述问题, 本文提出的 IMTG 框架中采用了一种迭代匹配的特征融合方法, 通过注意力机制和蒸馏在不同的迭代步骤中捕获图像和文本之间的关联语义信息, 并使用了三元组损失优化公共语义空间, 进而挖掘出更细微的图文对关系.

1.2　特征转换方法

特征转换的研究有着非常广泛的应用. 一些

研究[15-16]让不同模态的数据相互转换来解决跨模态行人重识别问题. Kiros 等[17]提出的框架使用了门控循环单元(gated recurrent unit, GRU)作为句子编码器, 并将图像和句子映射到一个公共空间, 采用排序损失训练模型用于图像字幕的生成. Donahue 等[6]提出了一种结合传统卷积神经网络(convolutional neural network, CNN)和长短期记忆(long short-term memory, LSTM)网络的新网络结构LRCN(long-term recurrent convolutional network)来处理视频-文本的映射问题. 黄远等[18]用CNN来生成图像描述, 并通过对抗训练来生成图像的句子描述. Kottur 等[10]以图像词嵌入的想法提出了一种用于图像字幕的联合图像字幕嵌入模型, 但其基于卡通图像, 很难应用于真实的图像. 本文提出了一种结合视觉网络和序列生成网络的文本生成模型, 以处理图像特征序列到文本单词序列的映射问题, 以特征转换的思想来增强图像文本之间的信息交互, 使得更多图像和文本的整体语义信息被挖掘到.

2 本文方法实现

本节详细介绍 IMTG 整体框架. 如图 1 所示, 该框架主要由特征融合模块和文本生成模块组成. 首先由特征融合模块获取较充分的细粒度的图像文本特征; 然后通过文本生成模块由图像特征生成文本描述, 增强图像和文本的整体语义相关性, 最后共同优化特征融合模块和文本生成模块. 其中, 真实文本特征为通过 Bi-GRU 提取到的标签文本特征, 生成文本特征是由迭代融合后的图像特征经过文本生成模块后得到的文本特征.

2.1 特征提取

2.1.1 图像特征提取

本文使用经过预训练的 Faster R-CNN 提取图像特征[2], 对一幅图像 $\boldsymbol{I}$, Faster R-CNN 识别出 a 个包含语义信息的区域, 并提取出每一个区域对应的一个 d 维的特征 $\boldsymbol{f}_i \in \mathbb{R}^d$, 最后将特征通过一个线性投影层变换成 d 维的向量 $\boldsymbol{v}_i = \boldsymbol{W}_{\mathrm{v}} \boldsymbol{f}_i + \boldsymbol{b}_{\mathrm{v}}$. 其中, $\boldsymbol{W}_{\mathrm{v}}$ 和 $\boldsymbol{b}_{\mathrm{v}}$ 是可学习的参数. 最终, 图像 $\boldsymbol{I}$ 的特征可以表示为 $\boldsymbol{V}_0 = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_a]$.

2.1.2 文本特征提取

为了实现图像和文本的细粒度对齐, 本文通过 Bi-GRU 提取文本的单词级特征[2]. 对一段包含 b 个单词的文本 $\boldsymbol{T}$, 每个单词均用连续的嵌入向量 $\boldsymbol{e}_j = \boldsymbol{W}_{\mathrm{e}} \boldsymbol{w}_j, j \in [1, b]$ 表示. 其中, $\boldsymbol{W}_{\mathrm{e}}$ 是可学习的嵌入矩阵; $\boldsymbol{w}_j$ 是单词的热编码(one-hot). 通过 Bi-GRU 获取文本前向和后向 2 个方向的信息分别为

$$\overrightarrow{\boldsymbol{h}_j} = \overrightarrow{\mathrm{GRU}}\left(\boldsymbol{e}_j, \overrightarrow{\boldsymbol{h}_{j-1}}\right),$$

$$\overleftarrow{\boldsymbol{h}_j} = \overleftarrow{\mathrm{GRU}}\left(\boldsymbol{e}_j, \overleftarrow{\boldsymbol{h}_{j+1}}\right).$$

其中, $\overrightarrow{\boldsymbol{h}_j}$ 和 $\overleftarrow{\boldsymbol{h}_j}$ 分别是前向 GRU 和后向 GRU 的隐藏状态. 每个单词的特征向量可表示为 $\boldsymbol{t}_j = \left(\overrightarrow{\boldsymbol{h}_j} + \overleftarrow{\boldsymbol{h}_j}\right)/2$, $\boldsymbol{t}_j \in \mathbb{R}^d$. 因此, 文本 $\boldsymbol{T}$ 的特征表示为 $\boldsymbol{T}_0 = [\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_b]$.

2.2 特征融合模块

令某一模态的一个样本(一幅图像或一段文本)为 $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m]$, 另一个模态的对应样本则为 $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n]$. 其中, $\boldsymbol{x}_j, \boldsymbol{y}_j \in \mathbb{R}^d$ 表示图像块或单词的特征. 本模块分为注意力和蒸馏 2

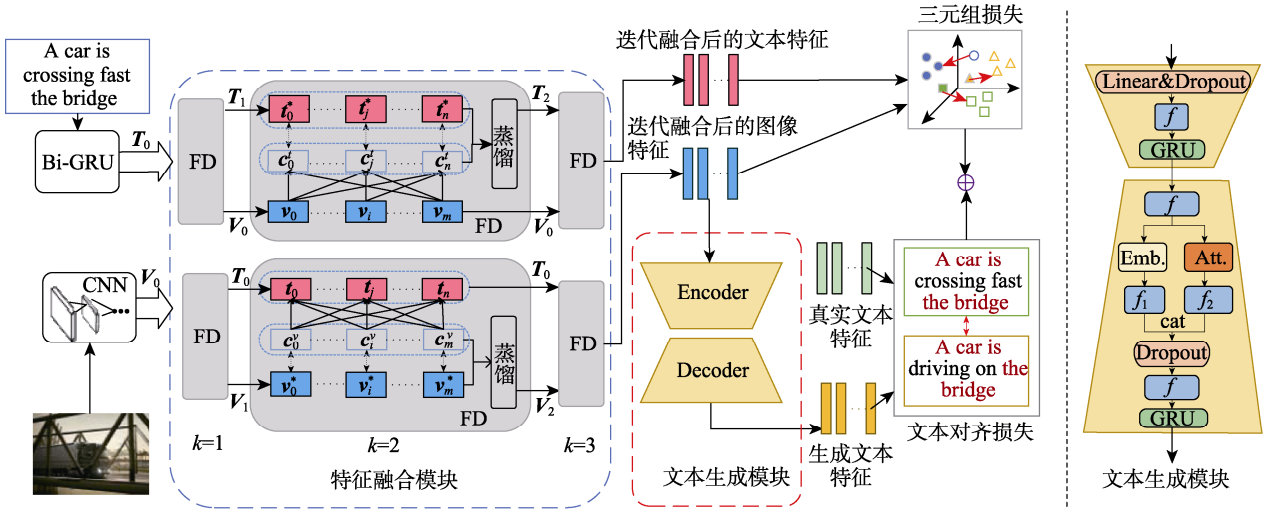


图 1 IMTG 整体框架图

个步骤进行特征融合.

Step1. 使用余弦函数计算每对特征 $\left(\boldsymbol{x}_i,\boldsymbol{y}_j\right)$ 的相似性 $s_{ij}=\mathrm{sim}\left(\boldsymbol{x}_i,\boldsymbol{y}_j\right)=\dfrac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{y}_j}{\left\|\boldsymbol{x}_i\right\|\cdot\left\|\boldsymbol{y}_j\right\|}$；然后对 s_{ij} 进行相似度归一化，即 $\overline{s_{ij}}=\mathrm{ReLU}\left(s_{ij}\right)\Big/\sqrt{\sum_{i=1}^{m'}\mathrm{ReLU}\left(s_{ij}\right)^2}$. 其中，$\mathrm{ReLU}\left(x\right)=\max\left(0,x\right)$.

对特征 $\boldsymbol{x}_i$，采用自适应的方法，通过 Softmax 函数得到每一个 $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_j$ 的关系权重，即注意力值

$$\alpha_{ij}=\exp\left(\varphi\overline{s_{ij}}\right)\Big/\sum_{j=1}^{n'}\exp\left(\varphi\overline{s_{ij}}\right).$$

其中，φ 为用于调整注意力分布的温度系数[2]；α_{ij} 为 $\boldsymbol{X}$ 中第 i 个特征 $\boldsymbol{x}_i$ 与 $\boldsymbol{Y}$ 中第 j 个特征 $\boldsymbol{y}_j$ 的关系权重.

然后再通过注意力机制让每个图像块和每个单词都进行交互，以加权求和的方式总结出 $\boldsymbol{x}_i$ 在 $\boldsymbol{Y}$ 中的上下文关系 $\boldsymbol{c}_i^x=\sum_{j=1}^{n}\boldsymbol{\alpha}_{ij}\boldsymbol{y}_j$. 其中，$\boldsymbol{c}_i^x$ 是 $\boldsymbol{x}_i$ 与 $\boldsymbol{Y}$ 对齐后的特征. $\boldsymbol{X}$ 通过注意力机制与 $\boldsymbol{Y}$ 对齐后，得到对齐特征集 $\boldsymbol{C}^x=\left[\boldsymbol{c}_1^x,\boldsymbol{c}_2^x,\cdots,\boldsymbol{c}_m^x\right]$. 通过上述过程，每一个 $\boldsymbol{c}_i^x$ 都捕捉到了 $\boldsymbol{x}_i$ 的相关语义信息以及整个 $\boldsymbol{Y}$ 的相关语义信息.

令上述步骤为 $f\left(\boldsymbol{X},\boldsymbol{Y}\right)=\boldsymbol{C}^x$，即表示 $\boldsymbol{X}$ 通过上述特征融合过程与 $\boldsymbol{Y}$ 实现对齐，得到 $\boldsymbol{C}^x$.

Step2. 聚合 $\boldsymbol{X}$ 和 $\boldsymbol{C}^x$ 得到更好的特征 $\boldsymbol{x}_i^*=d\left(\boldsymbol{x}_i,\boldsymbol{c}_i^x\right)$. 其中，函数 d 是一个门控机制，它可以保留历史信息，同时删除冗余信息，使蒸馏后的 $\boldsymbol{x}_i^*$ 更有助于下一次迭代提取图像和文本的关键信息.

将上述的 2 个步骤表示为函数

$$\mathrm{FD}\left(\boldsymbol{X},\boldsymbol{Y}\right)=d\left[f\left(\boldsymbol{X},\boldsymbol{Y}\right),\boldsymbol{X}\right]=\boldsymbol{X}^* \tag{1}$$

为了优化公共空间，挖掘出更多细粒度信息，同时提高对关键信息的利用率，如图 1 蓝框所示，该模块采取式(1)的特征融合策略[1]对图像和文本分别进行 k 次迭代. 即由文本 $\boldsymbol{T}_0$ 指导的图像特征迭代融合表示为

$$\boldsymbol{V}_k=\mathrm{FD}\left(\boldsymbol{V}_{k-1},\boldsymbol{T}_0\right)=d\left[f\left(\boldsymbol{V}_{k-1},\boldsymbol{T}_0\right),\boldsymbol{V}_{k-1}\right].$$

同理，由图像 $\boldsymbol{V}_0$ 指导的文本特征迭代融合表示为

$$\boldsymbol{T}_k=\mathrm{FD}\left(\boldsymbol{T}_{k-1},\boldsymbol{V}_0\right)=d\left[f\left(\boldsymbol{T}_{k-1},\boldsymbol{V}_0\right),\boldsymbol{T}_{k-1}\right].$$

其中，k 表示迭代次数；$\boldsymbol{V}_k$ 和 $\boldsymbol{T}_k$ 分别表示图像 $\boldsymbol{V}_0$ 和文本 $\boldsymbol{T}_0$ 的第 k 次迭代融合特征. 特征融合模块不断地用更新 k 次的特征与另一个模态的原始特征进行对齐与融合，在多次迭代步骤中聚合细粒度信息，捕获图像和文本之间的局部关联信息.

2.3　文本生成模块

如图 1 中的红框所示，文本生成模块是由 2 层 GRU 组成，第 1 层将图像特征“编码”为固定维度的隐向量，第 2 层将第 1 层的隐向量“解码”为一个句子(即一组单词序列).

编码部分是通过图像特征转换 $\varnothing\boldsymbol{E}$ 将每个图像特征输入 $\boldsymbol{v}_i$ 转变为一个固定维度的向量表示 $\varnothing\boldsymbol{E}\left(\boldsymbol{v}_i\right)$，即 $\varnothing_i=\varnothing\boldsymbol{E}\left(\boldsymbol{v}_i\right)$，简写为 $\varnothing_i\in\mathbb{R}^d$. 在处理完所有的视觉输入后，得到的序列特征空间表示为 $\left[\varnothing_1,\varnothing_2,\cdots,\varnothing_T\right]$. 其中，$T$ 为总步长数，视觉转换权重 $\boldsymbol{E}$ 是为了学习如何挑选出与序列分类问题相关的图像输入片段. 图像特征转换 $\varnothing\boldsymbol{E}$ 类比于深度 CNN 的某一激活层，它关注原来图像的信息，是固定步长的并且是独立的，因而可以利用 $\varnothing\boldsymbol{E}$ 来实现图像到序列模型的过渡.

解码部分利用一个 GRU 序列模型预测文本描述 $\boldsymbol{S}$ 中的下一个单词. 具体地，每一个时间步长对应一个 GRU 单元，并使用函数 f_θ 来表示，即

$$\boldsymbol{h}_t=f_\theta\left(\boldsymbol{x}_t,\boldsymbol{h}_{t-1}\right).$$

而在每一个时间步长中，输入当前时间步对应的单词 $\boldsymbol{x}_t$ 和前一个 GRU 单元生成的隐藏状态 $\boldsymbol{h}_{t-1}$，生成包含了上文信息的隐藏状态向量 $\boldsymbol{h}_t$，用于预测当前时间步 t 的单词，并输入下一个 GRU 单元.

此外，为了获得当前时间步生成的单词，基于上文所述的当前时间步的隐藏状态 $\boldsymbol{h}_t$，使用 Softmax 函数获得对应单词的概率分布. 具体地，首先通过向量点积计算当前隐藏状态 $\boldsymbol{h}_t$ 与每个单词嵌入向量的相似度，并将其作为每个单词类别的生成概率

$$\boldsymbol{p}\left(\boldsymbol{w}\middle|\boldsymbol{h}_t\right)=\exp\left(\boldsymbol{W}_w\boldsymbol{h}_t\right)\Big/\sum_{w'\in D}\exp\left(\boldsymbol{W}_{w'}\boldsymbol{h}_t\right) \tag{2}$$

其中，$\boldsymbol{W}_w$ 是单词 $\boldsymbol{w}$ 的可学习嵌入向量；D 是包含 N 个单词的词汇表(即图像数据集中单词的集合). 将词汇表中的 N 个单词看成 N 个类别，在第 t 个时间步长上，根据由 N 个单词组成并且以 one-hot 编码的样本真实单词分布向量 $\boldsymbol{L}_t=\left[\boldsymbol{l}_0,\boldsymbol{l}_1,\cdots,\boldsymbol{l}_{N-1}\right]$，通过式(2)就可以获得当前时间步长 t 上生成单词的概率分布向量 $\boldsymbol{P}_t=\left[\boldsymbol{p}_0,\boldsymbol{p}_1,\cdots,\boldsymbol{p}_{N-1}\right]$，其中的最大值 $\boldsymbol{p}_{\max t}$ 即为每个生成时间步 t 上对应的正确单词 $\boldsymbol{s}_t$ 的生成概率.

最终，完成 T 个时间步的计算. 由每个时间步隐藏状态序列 $\left\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_T\right\}$ 得到每个时间步生成单词的概率分布向量序列 $\left\{\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_T\right\}$；然后取每个时长步的 $\boldsymbol{p}_{\max t}$ 对应的正确单词 $\boldsymbol{s}_t$，最终生成

了对应的生成文本序列 $\boldsymbol{S}=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_T\}$.

2.4 损失函数

2.4.1 三元组损失

为了对匹配的图像-文本对进行聚类, 将不匹配的图像-文本对在嵌入空间中分离, 本文使用三元组损失对模型进行训练.

在特征迭代融合模块中, 每一次迭代都计算一个模态的局部特征与另一个模态扩展的局部特征相似度, 将它们求和得到整体的图像文本相似度

$$F_k(\boldsymbol{V},\boldsymbol{T})=\frac{1}{m}\sum_{i=1}^{m}\mathrm{sim}\left(\boldsymbol{v}_i,\boldsymbol{c}_{ki}^{v}\right)+\frac{1}{n}\sum_{j=1}^{n}\mathrm{sim}\left(\boldsymbol{c}_{kj}^{t},\boldsymbol{t}_j\right).$$

其中, sim 是余弦函数; m 和 n 分别是图像区域数量和文本单词数量.

经过 K 次迭代后将所有的局部相似度求和, 得到图像和文本之间总的相似度

$$F(\boldsymbol{V},\boldsymbol{T})=\sum_{k=1}^{K}F_k(\boldsymbol{V},\boldsymbol{T}).$$

最后, 为了使匹配的图像-文本对聚类、不匹配的文本对分离, 本文采用只考虑小区域强负样本的三元组损失[19]

$$L_{\mathrm{triplet}}(\boldsymbol{V},\boldsymbol{T})=\sum_{\hat{\boldsymbol{T}}}\left[\alpha-F(\boldsymbol{V},\boldsymbol{T})+F(\boldsymbol{V},\hat{\boldsymbol{T}})\right]_+ +\sum_{\hat{\boldsymbol{V}}}\left[\alpha-F(\boldsymbol{V},\boldsymbol{T})+F(\hat{\boldsymbol{V}},\boldsymbol{T})\right]_+.$$

其中, $[x]_+\equiv\max(x,0)$; α 是边缘因子; 第 1 个加和取自于给定一个图像 $\boldsymbol{I}$ 的所有负句子 $\hat{\boldsymbol{T}}$; 第 2 个加和取自于给定一个句子 $\boldsymbol{T}$ 的所有负图像 $\hat{\boldsymbol{V}}$.

2.4.2 文本对齐损失

为了通过图像和文本的特征转换增强图像文本间的交互, 进而挖掘出更多图像文本整体语义信息, 本文提出了单词级别的文本对齐损失函数训练文本生成模块, 以捕捉到每个隐层空间中准确的文本描述.

通过文本生成模块的图像转换权重 $\boldsymbol{E}$ 和序列组件权重 $\boldsymbol{W}_{\mathrm{w}}$, 将文本生成模块设计为通过最小化生成文本分布和真实分布之间的交叉熵来进行训练, 如 $L_{\mathrm{align}}(\boldsymbol{E},\boldsymbol{W}_{\mathrm{w}})=-\sum_{t=1}^{T}\sum_{i=0}^{N-1}\boldsymbol{L}_t(i)\lg\left(\boldsymbol{P}_t(i);\boldsymbol{E},\boldsymbol{W}_{\mathrm{w}}\right)$. 其中, T 为总时间步长; N 为词汇表中的单词总数; $\boldsymbol{L}_t(i)$ 是第 t 个时间步长上样本的真实单词分布 one-hot 向量; $\boldsymbol{P}_t(i)$ 是第 t 个时间步长上的单词概率分布向量. 生成文本的损失计算是单词级别的, 通过每个时间步长的累加得到整个句子 $\boldsymbol{S}$ 的损失值; 然后再使用梯度下降方法反向传播优化 $\boldsymbol{E}$ 和 $\boldsymbol{W}_{\mathrm{w}}$, 实现了句子级别的优化.

2.4.3 联合损失

为了兼顾特征融合模块和文本生成模块对模型的影响, 本文使用三元组损失和文本对齐损失的聚合损失 $L_{\mathrm{total}}=L_{\mathrm{triplet}}+\mu\times L_{\mathrm{align}}$ 优化 IMTG 框架. 其中, μ 用于平衡三元组损失和文本对齐损失.

3 实 验

3.1 数据集和评估指标

本文使用了 2 个基准数据集: (1) Flickr30K 由 31 000 幅图像和 158 915 个英文文本组成. 每幅图像都标注有 5 个英文文本. 与对比方法[1-2]一样, 本文使用 29 000 幅图像进行训练, 1 000 幅图像验证, 1 000 幅图像测试. (2) MS COCO 中包含约 123 287 幅图像, 每幅图像至少有 5 个英文文本. 与所对比方法[1-2]一样, 本文使用 122 287 幅图像来训练所有模型, 1 000 幅图像用于验证, 另外 1 000 幅图像用于测试.

本文采用了 R@1, R@5 和 R@10 评估指标[1-2]来衡量数据集中文本检索和图像检索性能. 为了更好地评估本文方法的有效性, 还统计了所有评价度量的总和 R@sum[3].

3.2 实验结果

3.2.1 对比实验

本文分别在数据集 Flickr30K 和 COCO 上将 IMTG 与其他模型进行了比较, 实验结果分别如表 1 和表 2 所示. 从所有评估指标可以看出, 本文提出的 IMTG 的检索性能有较大提升.

本文方法在 Flickr30K 和 COCO 上的 R@sum 分别以 4.7%和 5.0%的提升显著优于 JEL[23], 以 6.1%和 7.3%的提升显著优于 SCAN*[2], 以 1.2%和 5.6%优于的提升 VSRN*[17], 以 2.1%和 2.2%的提升优于 IMRAM[1], 以 2.4%和 2.3%的提升显著优于 SHAN[25], 以 2.6%和 1.8%的提升显著优于 DIME[26], 以 1.1%和 1.3%的提升显著优于 SGRAF[20], 以 0.7%和 1.2%的提升显著优于 UARDA[27]. 由此可见, IMTG 不仅能够在小数据集 Flickr30K 上达到最优的性能, 而且在大规模数据集 COCO 上也有很好的检索表现.

此外, 为了更直观地展示 IMTG 框架的优越性, 图 2 比较了 2 种图文检索方法 IMRAM[1]和 IMTG 的实例, 其中绿色标记表示检索正确, 红色

表 1　Flickr30K 数据集上 12 种方法比较　%

方法	文本检索(i2t)			图像检索(t2i)			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
DPC[20]	53.4	79.8	88.5	37.7	67.2	78.5	405.1
SCO[21]	53.3	80.1	87.2	39.6	87.9	83.2	431.3
SCAN*[2]	65.4	87.6	93.8	46.3	75.6	83.1	451.8
VSRN*[22]	69.8	88.6	96.1	53.1	79.4	86.3	473.3
IMRAM[1]	68.6	91.5	95.5	52.2	77.6	84.5	469.1
DCCA[23]	35.6	52.3	69.3	38.7	55.8	69.6	321.3
JEL[24]	57.2	85.2	91.2	**57.4**	79.7	86.7	457.8
SHAN[25]	69.1	90.1	95.2	51.0	79.3	83.5	468.2
DIME[26]	67.9	89.2	94.6	51.2	80.1	83.9	466.9
SGRAF[27]	70.3	91.2	95.3	52.3	78.9	85.8	473.8
UARDA[28]	**70.8**	91.5	95.6	52.5	79.6	85.7	475.7
IMTG	70.4	**91.7**	**96.1**	54.3	**80.0**	86.7	**479.2**

注．最优值以粗体显示．

表 2　COCO 数据集上 12 种方法比较　%

方法	文本检索(i2t)			图像检索(t2i)			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
DPC[20]	63.1	87.6	94.1	44.9	77.2	87.8	454.7
SCO[21]	67.8	90.7	95.5	54.5	85.1	92.8	486.4
SCAN*[2]	70.1	92.7	96.5	56.7	86.3	92.7	495.0
PVSE[29]	66.1	91.0	94.5	53.1	84.2	93.0	481.9
VSRN*[22]	73.9	92.4	96.3	60.1	87.6	93.0	503.3
IMRAM[1]	75.3	96.2	98.5	63.0	90.6	96.2	520.1
DCCA[23]	46.7	59.3	79.4	48.3	68.5	78.2	380.4
JEL[24]	65.3	91.5	97.3	**66.2**	90.1	95.7	506.1
SHAN[25]	76.8	96.3	98.7	62.6	89.6	95.8	519.5
DIME[26]	77.9	95.9	98.3	63	90.5	96.2	521.8
SGRAF[27]	**79.5**	96.2	98.5	63.2	90.7	96.1	524.3
UARDA[28]	78.6	96.5	98.9	63.9	90.7	96.2	524.8
IMTG	77.8	**96.5**	**99.2**	65.9	**93.5**	**98.3**	**531.2**

注．最优值以粗体显示．

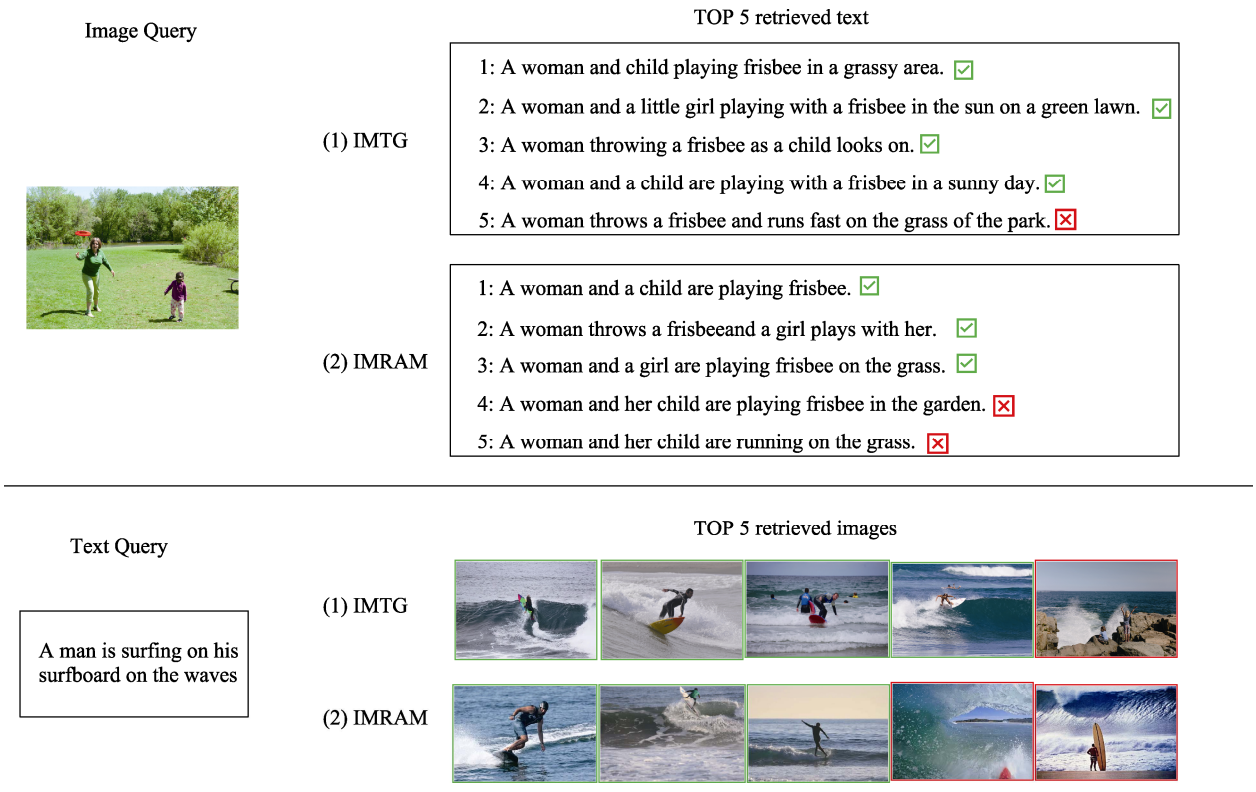


图 2　IMTG 和 IMRAM[1]检索方法的实例

标记表示检索错误．可以看出，IMTG 比 IMRAM[1]对图像和文本的语义理解更加丰富，有更好的图文检索能力．

3.2.2　消融实验

为了检验文本生成模块和特征融合模块的有效性，本文在 Flickr30K, COCO 上对 IMTG 的 4 种变体进行消融实验．如表 3 所示，Full 表示完整的 IMTG 模型；w/o|TG 表示缺少文本生成模块的模型；w/o|VF 表示只进行图像特征迭代融合的模型；w/o|TF 表示只进行文本特征迭代融合的模型．

(1) 文本生成模块的有效性．由表 3 的前 2 行可以看出，相比缺少文本生成模块的模型，完整的 IMTG 模型在 Flickr30K 和 COCO 上的整体性能 R@sum 分别提高了 2.1%和 2.2%；在文本检索任务

的召回指标中, 完整的IMTG模型最大的相对性能提升分别为 2.7%(Flickr30K 中的 R@1)和 3.4%(COCO 中的R@1); 在图像检索任务中, 完整的 IMTG 模型最大的相对性能提升分别为 4.0%(Flickr30K 中的 R@1)和 4.6%(COCO 中的 R@1). 这些评估检索性能指标的显著提升, 表明了 IMTG 框架中文本生成模块的有效性, 也说明了公共表示空间能够得以优化.

表 3 IMTG 在数据集 Flickr30K 和 COCO 上的消融实验对比 %

模型	方法			Flickr30K							COCO						
	文本生成模块	图像特征融合	文本特征融合	文本检索(i2t)			图像检索 (t2i)			R@sum	文本检索 (i2t)			图像检索 (t2i)			R@sum
				R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
Full	✓	✓	✓	**70.4**	**91.7**	96.1	**54.3**	**80.0**	**86.7**	**479.0**	**77.8**	**96.5**	99.2	**65.9**	**93.5**	**98.3**	**531.2**
w/o|TG	×	✓	✓	68.6	91.5	95.5	52.2	77.6	84.5	469.1	75.3	96.2	98.9	63.0	90.6	96.2	520.1
w/o|VF	✓	✓	×	70.0	90.6	95.5	50.4	78.2	85.8	470.5	77.5	95.4	98.5	63.2	91.3	97.3	523.2
w/o|TF	✓	×	✓	67.8	91.1	**96.5**	51.6	78.2	85.0	470.2	75.2	95.9	**99.5**	63.4	91.8	97.2	523.0

注. 最优值以粗体显示.

(2) 特征融合模块的有效性. 由表 3 的后 2 行可以看出, 相比完整的 IMTG 模型, 只进行图像特征迭代融合的模型在 Flickr30K 和 COCO 上的整体性能 R@sum 分别下降了 1.8%和 1.5%; 而只进行文本特征迭代融合的模型在 Flickr30K 和 COCO 上的整体性能 R@sum 分别下降了 1.9%和 1.6%. 这说明了 ITMG 框架的特征融合模块对检索是有效的.

3.2.3 超参实验

(1) 迭代次数对检索性能的影响. 表 4 所示为当 k 不同时, IMTG 的图文检索性能差异. 值得关注的是, 完整的 IMTG 模型在 $k=1$ 时的检索性能与缺少文本模块的模型在 $k=3$ 时的检索性能相近, 在 $k=2$ 时超过 w/o|TG 在 $k=3$ 时的检索性能, 说明文本生成模块对检索性能提升有着很好的促进作用.

表 4 Flickr30K 数据集上迭代融合次数对性能的影响 %

方法	k	文本检索(i2t)			图像检索(t2i)			R@sum
		R@1	R@5	R@10	R@1	R@5	R@10	
w/o|TG	1	55.9	83.4	91.4	34.9	64.9	74.9	405.4
	2	65.9	90.1	95.3	51.4	**77.8**	**84.9**	465.4
	3	**68.6**	**91.7**	**95.5**	**52.2**	77.6	84.5	**470.1**
Full	1	67.4	89.4	94.8	49.3	77.4	85	463.3
	2	70.0	90.2	95.2	53.1	79.5	86.3	474.3
	3	**70.4**	**91.5**	**96.1**	**54.3**	**80.0**	**86.7**	**479.2**

注. 最优值以粗体显示.

此外, 图 3 展示了当 k 不同时, IMTG 框架对于提升图像检索性能的效果. 值得关注的是, $k=1$ 时, 文本生成模块的加入使得整体检索效果提升得最明显, 与 $k=1$ 时相比, $k=2$ 时 R@sum 相对提升了 14.3%; 而当 $k=3$ 时, R@sum 相对提升了 1.9%, 虽然也有较好的提升, 但是提升的效果没有 $k=1$ 时明显. 这是因为在迭代特征融合的初期, 特征转换的策略能够快速地挖掘出新的图像和文本关联语义信息, 而在迭代特征融合的后期, 由于语义空间逐渐丰富, 文本生成模块受到了一定程度的能力限制.

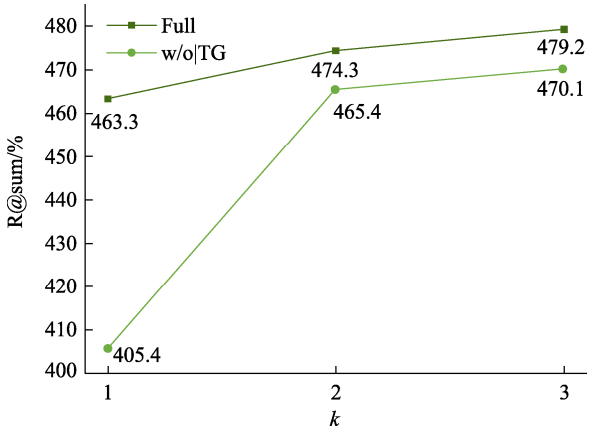
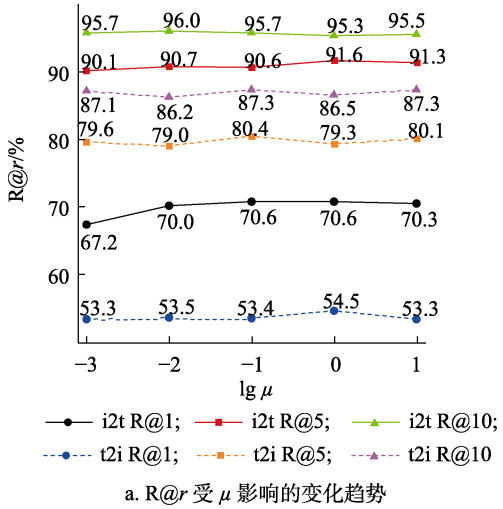
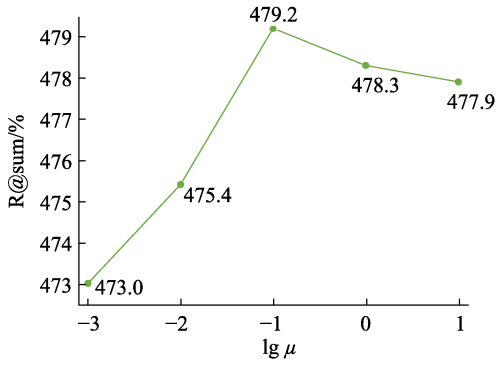


图 3 迭代次数不同时整体检索的提升效果差异

(2) 损失平衡参数 μ 对检索性能的影响. 图 4 展示了在 Flickr30K 上平衡参数 μ 对 IMTG 的影响. 如图 4a 所示, 当 $\mu=0.1$ 时, IMTG 的各项指标 R@r 最优. 如图 4b 所示, μ 从 0.001 变化到 10, 可以看出, 当 $\mu=0.1$ 时, IMTG 的 R@sum 最优; 当 μ 的值减小为 0.01 和 0.001 时, R@sum 的下降幅度明显大于 μ 增大的幅度, 表明平衡参数 μ 对应的文本生成模块对检索效果的提升有非常重要的作用.



a. R@r 受 μ 影响的变化趋势



b. R@sum受 μ 影响的变化趋势

图 4　文本/图像检索性能受损失平衡参数 μ 的影响

4　结　论

本文提出一种基于文本生成与迭代匹配的跨模态检索框架 IMTG, 以解决跨模态图文检索的公共特征空间表达不足的问题. 其首先获取较充分的细粒度的图像文本特征, 然后通过文本生成模块由图像特征生成文本描述, 增强图像和文本的整体语义相关性, 最后共同优化特征融合模块和文本生成模块. 在 Flickr30K 和 COCO 上的多组实验结果表明, IMTG 的性能显著地超越了对比模型, 并且证明了在跨模态检索任务中, 优化公共嵌入空间的重要性以及文本生成模块的有效性. 相比其他方法, 本文方法具有明显的优势: (1) 优化了局部公共嵌入空间. 采用迭代匹配的方法, 融合图像和文本特征, 充分地获取图像文本的关联语义信息. (2) 优化了全局公共嵌入空间. 生成模块实现了从图像到文本的映射, 增强了图像和文本的整体语义相关性, 使更多的全局语义信息被挖掘到. 未来将改进图文特征转化的思路在图文检索中的运用, 由文本特征生成图像特征的方法可以和由图像特征生成文本特征的方法一样运用于跨模态图文检索研究中.

参考文献(References):

[1] Chen H, Ding G G, Liu X D, *et al*. IMRAM: iterative matching with recurrent attention memory for cross-modal image-text retrieval[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 12652- 12660

[2] Lee K H, Chen X, Hua G, *et al*. Stacked cross attention for image-text matching[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 212-228

[3] Huang Y, Wang W, Wang L. Instance-aware image and sentence matching with selective multimodal LSTM[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 7254-7262

[4] Jiang X Y, Wu F, Li X, *et al*. Deep compositional cross-modal learning to rank via local-global alignment[C] //Proceedings of the 23rd ACM International Conference on Multimedia. New York: ACM Press, 2015: 69-78

[5] Karpathy A, Joulin A, L F F. Deep fragment embeddings for bidirectional image sentence mapping[C] //Proceedings of the 28th International Conference on Neural Information Processing Systems. New York: ACM Press, 2014: 1889-1897

[6] Donahue J, Anne Hendricks L, Guadarrama S, *et al*. Long-term recurrent convolutional networks for visual recognition and description[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 2625-2634

[7] Venugopalan S, Xu H J, Donahue J, *et al*. Translating videos to natural language using deep recurrent neural networks[C] //Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver: Association for Computational Linguistics, 2015: 1494-1504

[8] Gu J X, Cai J F, Joty S, *et al*. Look, imagine and match: improving textual-visual cross-modal retrieval with generative models[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 7181-7189

[9] Andrej k, Li F F. Deep visual-semantic alignments for generating image descriptions[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 3128-3137

[10] Kottur S, Vedantam R, Moura J M F, *et al*. VisualWord2vec (vis-w2v): learning visually grounded word embeddings using abstract scenes[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 4985-4994

[11] Cui Huailei, Liu Li, Zhang Huaxiang, *et al*. Detail preserving image generation method based on semantic consistency[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(10): 1497-1505(in Chinese)

(崔怀磊, 刘丽, 张化祥, 等. 基于语义一致性的细节保持图像生成方法[J]. 计算机辅助设计与图形学学报, 2022, 34(10): 1497-1505)

[12] Kim W, Son B, Kim I. ViLT: vision-and-language transformer without convolution or region supervision[C] //Proceedings of the 38th International Conference on Machine Learning. New York: PMLR, 2021: 5583-5594

[13] Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale[C] //Proceedings of the 9th International Conference on Learning Representations. Ithaca: OpenReview.net, 2021: Article No.21

[14] Radford A, Kim J W, Hallacy C, *et al*. Learning transferable visual models from natural language supervision[C] //Proceedings of the 38th International Conference on Machine Learning. New York: PMLR, 2021: 8748-8763

[15] Wang G A, Zhang T Z, Cheng J, *et al*. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 3622-3631

[16] Wang Z X, Wang Z, Zheng Y Q, *et al*. Learning to reduce dual-level discrepancy for infrared-visible person re-identification[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 618-626

[17] Kiros R, Salakhutdinov R, Zemel R S. Unifying visual-semantic embeddings with multimodal neural language models[OL]. [2023-06-20]. https://arxiv.org/abs/1411.2539

[18] Huang Yuan, Bai Cong, Li Hongkai, *et al*. Image captioning based on conditional generative adversarial nets[J]. Journal of Computer-Aided Design & Computer Graphics, 2020, 32(6): 911-918(in Chinese)
(黄远, 白琮, 李宏凯, 等. 基于条件生成对抗网络的图像描述生成方法[J]. 计算机辅助设计与图形学学报, 2020, 32(6): 911-918)

[19] Faghri F, Fleet D J, Kiros J R, *et al*. VSE++: improving visual-semantic embeddings with hard negatives[C] //Proceedings of the British Machine Vision Conference. London: BMVA Press, 2018: Article No.12

[20] Zheng Z D, Zheng L, Garrett M, *et al*. Dual-path convolutional image-text embeddings with instance loss[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2020, 16(2): Article No.51

[21] Huang Y, Wu Q, Song C F, *et al*. Learning semantic concepts and order for image and sentence matching[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 6163-6171

[22] Li K P, Zhang Y L, Li K, *et al*. Visual semantic reasoning for image-text matching[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 4653-4661

[23] Andrew G, Arora R, Bilmes J, *et al*. Deep canonical correlation analysis[C] //Proceedings of the 30th International Conference on International Conference on Machine Learning. New York: ACM Press, 2013: III-1247-III-1255

[24] Xie Z W, Liu L, Wu Y Z, *et al*. Learning text-image joint embedding for efficient cross-modal retrieval with deep feature engineering[J]. ACM Transactions on Information Systems, 2022, 40(4): Article No.74

[25] Ji Z, Chen K X, Wang H R. Step-wise hierarchical alignment network for image-text matching[C] //Proceedings of the 30th International Joint Conference on Artificial Intelligence. Freiburg: IJCAI, 2021: 765-771

[26] Qu L G, Liu M, Wu J L, *et al*. Dynamic modality interaction modeling for image-text retrieval[C] //Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2021: 1104-1113

[27] Diao H W, Zhang Y, Ma L, *et al*. Similarity reasoning and filtration for image-text matching[C] //Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021: 1218-1226

[28] 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, 2023, 25: 1320-1332

[29] Song Y L, Soleymani M. Polysemous visual-semantic embedding for cross-modal retrieval[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 1979-1988