

一种融合注意力机制的跨模态图文检索算法

杨迪, 吴春明

(西南大学计算机与信息科学学院, 重庆 400700)

摘要:随着不同模态的数据爆发式增长,跨模态检索成为信息检索领域的重要研究课题。由于语义相同事物在不同模态下底层特征异构,如何科学度量它们之间的相似性成为跨模态检索研究首先要解决的重要问题。当前主流的图文检索方法通过模型将异构特征映射到公共空间再进行相似性度量,这些工作主要可分为两种思路,一是从全局特征角度来实现全局信息对齐,二是从局部特征入手来实现细粒度信息对齐,但前者容易丢失局部细节信息,而后者容易导致语义信息不完善。为此,该文提出一种融合注意力机制的跨模态图文检索算法。首先,利用 Vision Transformer 和 Bert 模型获得包含上下文信息的图像和文本特征,再利用注意力机制获得模态内局部的图像和文本特征;其次,通过注意力机制得到模态间全局的图像和文本特征;最后,将这些优化的特征与基础特征融合来进行跨模态检索。该算法既充分利用了不同模态的细粒度特征,又更好地兼顾了全局信息,因而能取得更好的检索精度,通过在 Wikipedia 数据集上的大量对比实验,证明了该算法的有效性。

关键词:图文检索;跨模态;注意力机制;全局特征;局部特征

中图分类号: TP391

文献标识码: A

文章编号: 1673-629X(2023)11-0143-06

doi: 10.3969/j.issn.1673-629X.2023.11.021

A Cross-modal Image and Text Retrieval Algorithm for Integrating Attention Mechanism

YANG Di, WU Chun-ming

(School of Computer and Information Science, Southwest University, Chongqing 400700, China)

Abstract: With the explosive growth of data in different modes, cross-modal retrieval has become an important research topic in the field of information retrieval. Because the underlying features of semantically identical objects are different in different modes, how to measure the similarity between them scientifically has become the first important issue to be solved in the research of cross-modal retrieval. The current mainstream image and text retrieval methods map the heterogeneous features to the public space through the model and then measure the similarity. These works can be divided into two main ideas. One is to achieve global information alignment from the perspective of global features, and the other is to achieve fine-grained information alignment from the perspective of local features. However, the former is easy to lose local details, while the latter is easy to lead to incomplete semantic information. Therefore, we propose a cross-modal image and text retrieval algorithm that integrates attention mechanism. Firstly, we use Vision Transformer and Bert model to obtain image and text features containing context information, and then use attention mechanism to obtain local image and text features within the mode. Secondly, we use attention mechanism to obtain global image and text features between modes, and finally fuse these optimized features with basic features to conduct cross-mode retrieval. The proposed algorithm not only makes full use of the fine-grained features of different modes, but also gives better consideration to the global information, so it can achieve better retrieval accuracy. Through a large number of comparative experiments on Wikipedia data sets, the effectiveness of the proposed algorithm is proved.

Key words: image-text retrieval; cross-modal; attention mechanism; global feature; local feature

0 引言

随着移动设备智能化,社交软件的普及,人们可以更加便捷地生成各种不同模态的多媒体数据(图像、

文本、视频、音频等)。面对这些海量数据,人们的检索需求从传统的单模态检索转变为跨模态检索。跨模态检索是指给定一种模态的查询样本,得到与查询样

收稿日期: 2022-12-27

修回日期: 2023-04-28

基金项目: 重庆自然科学基金(cstc2019jcyj-msxmX0130)

作者简介: 杨迪(1995-),男,硕士研究生,CCF会员(98062G),研究方向为深度学习技术;通信作者: 吴春明(1972-),男,博士,副教授,研究方向为智能信息获取。

本语义相似的其他模态的样本^[1],如文本/视频检索图像,图像/视频检索文本,该技术的关键在于如何有效提取不同模态数据的特征,并将这些特征以适宜的方法进行相似性度量。以图文检索为例,图像由像素构成,文本由单词序列组成,它们之间的相似度不能直接比较,这种底层特征异构所带来的“语义鸿沟”是跨模态检索首先要解决的重要问题。

传统的跨模态检索主要采用典型相关性分析(Canonical Correlation Analysis, CCA)方法,如 Yan 等人^[2]利用该方法来寻找图像和句子的最大相关性。随着深度学习技术的发展,跨模态检索普遍解决方案变为从不同模态提取特征,再将这些特征映射到深度空间中,在该空间进行距离计算,经过学习之后,该空间鼓励相似样本对互相靠近,不相似样本对互相远离。Wang 等人^[3]利用 CNN 和 WCNN 分别提取图像和文本特征,证明这种基于深度神经网络提取的特征能有效提高检索精度。Dong 等人^[4]提出图卷积网络(Graph Convolutional Network, GCN),利用样本的邻接关系重构样本表示并基于局部图重构节点特征,从而获取隐藏的高级语义信息,但节点更新较为复杂,计算代价巨大。Peng 等人^[5]提出了一种跨模态生成对抗网络(Cross-Modal Generative Adversarial Networks, CM-GAN),利用生成模型和判别模型互相博弈来生成更具细粒度的多模态特征表示。Bahdanau^[6]首次将注意力机制应用到机器翻译领域,该机制能聚焦重要部分而忽略不重要部分的特性,使得其在计算机视觉和自然语言处理领域取得了一系列成绩,学者们也开始将注意力机制应用到跨模态检索领域。Nam 等人^[7]提出双重注意力网络(Dual Attention Networks, DANs),利用视觉和文本注意力机制来捕获图像区域和单词之间的相互关系;Lee 等人^[8]提出堆叠交叉注意力方法来捕捉图像区域和单词的潜在对齐;Li 等人^[9]提出 DMASA 方法,利用多种自注意力机制从不同角度提取图像文本细粒度特征。

上述工作都在一定程度上提升了检索效果,但也存在两个主要问题:一是仅考虑了局部特征或者全局特征的一种,导致特征关键语义不够全面,信息表征不够完善;二是忽略了模态间有效交互,由于不同模态所含信息量不等,这会导致特征语义表达不够充分。针对这些问题,该文提出了一种融合注意力机制的图文检索算法。首先,利用 ViT 和 Bert 模型得到包含上下文信息的图像和文本特征;其次,利用注意力机制融合不同模态信息即用文本信息来表示图像,用图像信息表示文本;再次,将注意力机制引进到特征提取过程,利用融合不同模态信息的特征向量来获得新的全局特征表示和局部特征表示;最后,融合新的全局特征向

量、局部特征向量和原始特征向量来表征数据。由于该方法更好地融合了全局和局部特征,因而取得了更好的检索精度,通过在 Wikipedia 数据集上与 6 种经典方法的对比实验,证明了该方法的有效性。

1 网络结构

整个模型结构如图 1 所示,包含图像编码模块、文本编码模块、交互模块 3 个部分。其中,图像编码模块负责图像特征提取,首先将图像分成块并加入位置信息编码,通过输入 ViT 模型得到全局特征和局部特征,作为图像的基础特征表示;文本编码模块负责文本特征提取,首先将文本数据通过词嵌入方式转为词向量,输入 Bert 模型得到文本的全局表示和单体表示,作为文本的基础特征表示;交互模块又分为模态内注意模块和模态间注意模块,为了挖掘语义相似不同模态数据间的内在联系,该文利用两个模块分别获取图像和文本新的局部特征和新的全局特征。最后,将这些特征与基础特征拼接,作为图像和文本的最终特征表示。

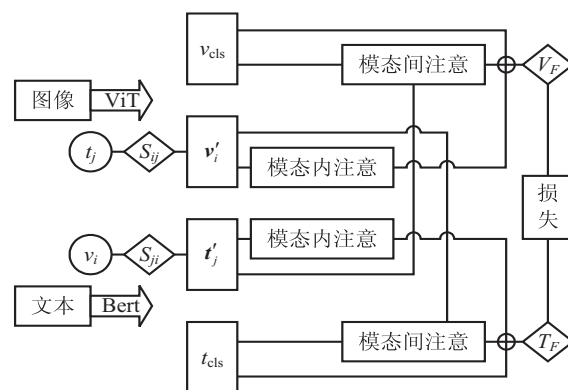


图 1 模型结构

1.1 图像编码模块

Transformer 模型的自注意力机制能对长距离依赖问题建模,能充分利用上下文信息从而获得有效的全局信息,因此,文中图像特征提取过程使用基于 Transformer 编码器的 ViT 模型。ViT^[10]是 Google 团队提出的基于 Transformer 的一种图像分类模型,该模型将二维的图像数据转换成一维块序列使得 Transformer 能处理图像。具体来说,将输入图像的像素调整为 224×224,把图像分割成大小为 16×16、数量为 196 的 patch 块,加入位置信息编码并将其按顺序展平转化为向量,输入预训练好的 ViT 模型,得到输入图像的特征表示 $V = \{v_{cls}, v_1, \dots, v_i, \dots, v_n\}$,其中 v_{cls} 表示图像的整体信息, n 为图像块的数量, v_i 为第 i 个图像块的特征向量。特征提取整体过程如图 2 所示。

1.2 文本编码模块

在图文检索中,文本通常以句子或长段落形式存在,而 Bert 模型的双向编码结构使得其在提取长文本

数据特征方面有着突出优势。Bert^[11]模型也是基于Transformer的自然语言处理模型,该模型使用Transformer Encoder作为特征提取器,具有强大的语义信息提取能力。因此,该文利用Bert模型进行文本特征的提取。如图3所示,首先,将文本数据通过word2vec模型转化为词向量,然后,输入到预训练好的Bert模型得到文本特征表示 $T = \{t_{cls}, t_1, \dots, t_j, \dots, t_l\}$,其中, t_{cls} 为文本的全局表示, l 为文本长度, t_j 为第 j 个词的特征向量。

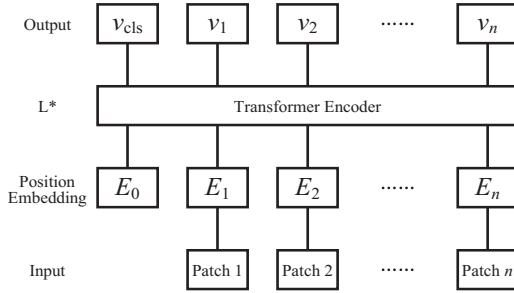


图2 图像特征提取

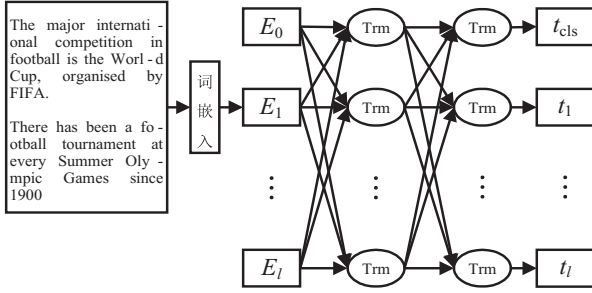


图3 文本特征提取

1.3 交互模块

注意力机制能选择性地关注重要信息,能为其赋予更高的权重,因而能有效提取关键特征。在图像和文本编码模块,利用注意力机制对图像和文本的基础特征进行了提取,但这种注意计算仅局限在同一模态内,即图像块到图像块的注意和单词到单词的注意,然而语义相似的图像和文本数据所包含的信息量不等,不同模态所关注的内容也不尽相同,因此在进行注意计算的时候应充分考虑不同模态间的相互影响,即图像块到单词的注意和单词到图像块的注意。因此,为了融合不同模态的特征并挖掘不同模态的内在联系,在本模块中,结合注意力机制分别设计了模态内注意模块和模态间注意模块,用来寻找新的局部特征映射和全局特征映射。

图像数据相比于文本数据具有更多的细节信息,文本数据比图像数据有更多的语义描述,为了凸显它们的内在关系,该文用文本信息来表征图像,用图像信息来表征文本。首先计算每个图像块和每个单词的相似性:

$$S_{ij} = \mathbf{v}_i^T \mathbf{t}_j \quad (1)$$

每个图像块的文本表示为:

$$\mathbf{v}_i' = \sum_{j=1}^l \frac{\exp(S_{ij})}{\sum_{j=1}^l \exp(S_{ij})} \mathbf{t}_j \quad (2)$$

同理,每个单词的图像表示为:

$$\mathbf{t}_j' = \sum_{i=1}^n \frac{\exp(S_{ij})}{\sum_{i=1}^n \exp(S_{ij})} \mathbf{v}_i \quad (3)$$

其中, \exp 是以自然数 e 为底的指数函数。在Transformer中,通过点乘的方式来计算两个向量的相似性,而这里的图像和单词相似性矩阵乘与该方式本质上一致。

1.3.1 模态内注意模块

模态内注意模块的目的是生成融合另一模态信息的局部特征表示。以图像数据为例,首先算出由文本表示的图像块向量的平均值 $\bar{\mathbf{v}}$,并以该值为基准与图像向量 \mathbf{v}_i' 做注意计算,得到新的代表局部信息的特征向量 \mathbf{V}_p ,计算公式如下:

$$\mathbf{V}_p = \sum_{i=1}^n f(\bar{\mathbf{v}}, \mathbf{v}_i') = \sum_{i=1}^n \alpha_i \mathbf{v}_i' \quad (4)$$

其中, $f(\cdot)$ 是注意力机制函数, $\bar{\mathbf{v}}$ 是 \mathbf{v}_i' 的平均值,局部特征向量 \mathbf{V}_p 是通过用文本表示的图像向量 \mathbf{v}_i' 加权求和得来,权重 α_i 是通过单层感知机和点乘得来。每个图像块向量的注意力权重代表各自的重要程度,新的局部特征向量因此能够表达图像中的关键部分,同时由于 \mathbf{v}_i' 和 $\bar{\mathbf{v}}$ 是通过文本表示而来,最后的局部特征也融合了文本信息。

$$\bar{\mathbf{v}} = \frac{1}{n} \left(\sum_{i=1}^n \mathbf{v}_i' \right) \quad (5)$$

$$\alpha_i = \text{softmax}(s(\mathbf{v}_i', \bar{\mathbf{v}})) = \frac{\exp(s(\mathbf{v}_i', \bar{\mathbf{v}}))}{\sum_{i=1}^n \exp(s(\mathbf{v}_i', \bar{\mathbf{v}}))} \quad (6)$$

$$s(\mathbf{v}_i', \bar{\mathbf{v}}) = [\text{Tanh}(\mathbf{W}_1 \mathbf{v}_i' + \mathbf{b}_1)]^T \bar{\mathbf{v}} \quad (7)$$

以相同的方法计算融合图像信息的文本局部特征向量 \mathbf{T}_p 。

1.3.2 模态间注意模块

模态间注意模块的目的是生成融合另一模态信息的全局特征表示。还是以图像数据为例,该文利用基础图像全局特征 \mathbf{v}_{cls} 与用图像信息表示的文本特征向量 \mathbf{t}_j' 做注意计算,得到新的代表全局信息的特征向量 \mathbf{V}_w ,计算公式如下:

$$\mathbf{V}_w = \sum_{j=1}^l f(\mathbf{v}_{cls}, \mathbf{t}_j') = \sum_{j=1}^l \beta_j \mathbf{t}_j' \quad (8)$$

β_j 的计算过程与公式(6)(7)相同。同理可以得到融合图像信息的文本全局特征向量 \mathbf{T}_w 。

最后,融合新的局部特征向量 V_p 、 T_p ,新的全局特征向量 V_w 、 T_w 及基础全局特征向量 v_{cls} 、 t_{cls} 作为图像文本的最终特征表示 V_F 、 T_F ,分别见公式(9)和(10),其中 $[\cdot]$ 表示向量的拼接。

$$V_F = F_{\text{fusion}}[V_p; V_w; v_{cls}] \quad (9)$$

$$T_F = F_{\text{fusion}}[T_p; T_w; t_{cls}] \quad (10)$$

1.4 损失函数

为了保证共享空间中语义相似的图像-文本对距离足够近,不相似图像-文本对的距离足够远,该文采用三元组排序损失函数^[12]。对于图像数据集,构建三元组 $\{V_F, T_{F^+}, T_{F^-}\}$,其中 (V_F, T_{F^+}) 表示锚点 V_F 的正样本对, (V_F, T_{F^-}) 表示负样本对,即与图像样本语义不相似的文本对,以相同的方式构建文本三元组 $\{T_F, V_{F^+}, V_{F^-}\}$ 。通过最小化相似样本对之间的距离,同时最大化不相似样本对的距离,保证图像文本模态的一致性。由于跨模态检索任务是双向检索,因此损失函数定义为:

$$L = [S(V_F, T_{F^+}) - S(V_F, T_{F^-}) + \lambda]_+ + [S(T_F, V_{F^+}) - S(T_F, V_{F^-}) + \lambda]_+ \quad (11)$$

其中, λ 是一个常量,用来保证相似样本对得分比不相似样本对得分大于一个固定值, $[x]_+ = \max(0, x)$ 。 S 函数表示图像文本对的相似性得分,以 (V_F, T_{F^+}) 为例,具体的计算公式为:

$$S(V_F, T_{F^+}) = \frac{V_F \cdot T_{F^+}}{\|T_{F^+}\| \|V_F\|} \quad (12)$$

算法流程如表 1 所示。

表 1 跨模态检索算法

跨模态检索训练流程
输入:图像集 V ,文本集 T
输出:图像文本的最优特征表示
1. 选出 n 对匹配的图像文本对 (v_i, t_i)
2. For $i = 1, 2, \dots, n$ do
3. 通过 ViT 和 Bert 得到图像文本编码
4. 通过公式(2)进行图像表示文本,用公式(3)进行文本表示图像
5. 利用公式(4)和(8)分别计算融合另一模态信息的局部特征向量和全局特征向量
6. 融合不同特征向量
7. 构建三元组并计算损失
8. end for

2 实验结果与分析

2.1 数据集

Wikipedia^[13]是跨模态检索研究普遍使用的数据集,来源于维基百科中的代表文章,并基于对应文章补充相关图像,整个数据集共有 2 866 个图像文本对,这些文本以短段落(至少 70 个字)描述图像,包含 10 个

语义类。

2.2 评价指标

该文采用跨模态检索研究中通常采用的精确率-召回率(Precision-recall)曲线和平均精度均值 mAP(Mean Average Precision)作为评价指标。

PR 曲线横坐标为召回率,纵坐标为精确率,纵坐标值越大表示该方法性能越好。精确率 P 、召回率 R 计算公式如下:

$$P = \frac{a}{a+b} \quad (13)$$

$$R = \frac{a}{a+c} \quad (14)$$

其中, a 表示检索返回中的正样本数量, b 表示检索返回中的负样本数量, c 表示数据集中没有返回的正样本数量。

mAP 是 AP 的平均值,该指标综合考虑了排序信息和精确率^[14]。取值越接近 1 代表方法性能越好。给出查询数据和 n 个检索结果,AP 计算公式如下:

$$AP = \frac{1}{R} \sum_{i=1}^n P(i) \delta(i) \quad (15)$$

其中, R 是测试集中的正样本数量, $P(i)$ 表示前 i 个检索结果的精确率,若检索结果为正样本,则 $\delta(i) = 1$,否则为 0。 Q 代表查询次数,最终 mAP 值公式为:

$$mAP = \frac{1}{Q} \sum_{q=1}^Q AP \quad (16)$$

2.3 模型对比分析

基于验证文中算法有效性的目的,选取了 KCCA^[15]、DCCA^[16]、SCM^[17]、ACMR^[18]、DSCMR^[19]、DMTL^[20]共 6 种方法进行对比实验。其中,KCCA 利用核函数改变特征维度再进行关联分析,解决了 CCA 不能处理非线性关系的不足;DCCA 将深度神经网络与 CCA 相结合,从两个视图学习非线性投影,比 KCCA 模型更为简洁;SCM 是在 CCA 基础上将无监督相关和有监督语义结合的匹配算法;ACMR 将对抗机制引入到语义融合层面,丰富了特征空间内容,并利用三元组约束保证语义相同的不同模态表示差异最小;DSCMR 充分利用标签信息有效学习了不同模态公共表示,并通过最小化标签空间和公共表示空间的判别损失,以监督模型学习判别特征;DMTL 由两个多模态特定的神经网络和一个联合学习模块组成,是一种迁移已标记类别的知识,以提高在未标记的新类别上检索性能的学习方法。

实验结果如表 2 所示。

由表 2 可知,文中方法的平均 mAP 达到了 0.699,不管是图像模态检索文本还是文本模态检索图像,均高于其他方法。对比 DMTL 方法,文中方法图像检索文本的 mAP 值从 0.633 提高到 0.687,文本检索图像

的 mAP 值从 0.652 提高到 0.711, 平均 mAP 值从 0.642 提高到 0.699。整体来看, 基于深层结构方法在检索效果上大于浅层结构方法, 这得益于深度学习强大的特征学习能力, 可以有效捕捉样本间非线性关系, 从而获取更能代表数据的关键特征。现用的跨模态检索方法大多将不同模态的数据映射到公共空间, 这些方法只是简单将图像文本全局特征或局部特征对齐, 而文中方法利用注意力机制充分挖掘同一模态内细粒度局部信息和不同模态间交互全局信息, 全面考虑两种信息从而提高了模型的检索准确率。

表 2 跨模态检索方法 mAP(Wikipedia 数据集)

方法	图像检索文本	文本检索图像	平均值
KCCA	0.289	0.273	0.281
DCCA	0.362	0.386	0.374
SCM	0.326	0.298	0.312
ACMR	0.462	0.489	0.475
DSCMR	0.545	0.527	0.536
DMTL	0.633	0.652	0.642
文中	0.687	0.711	0.699

为了进一步验证文中方法的有效性, 在数据集上绘制所有对比方法的 PR 曲线, 如图 4 所示。

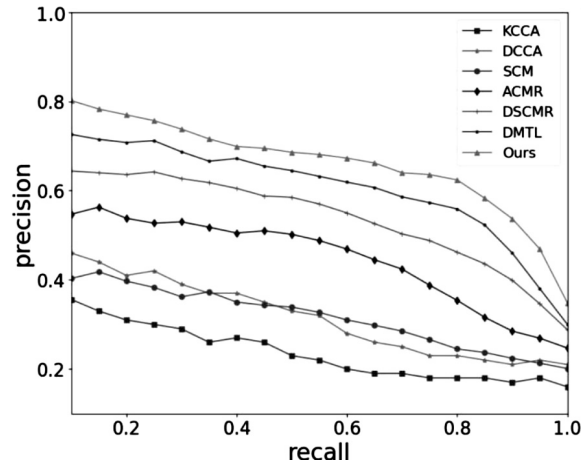


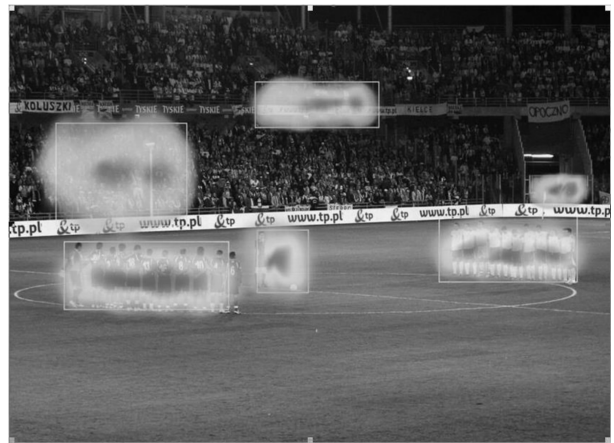
图 4 图像检索文本 PR 曲线

由图 4 可知, 文中方法明显优于其他对比方法, 当召回率值为 0.4 时, 仅 DMTL 方法的精确率与文中方法基本持平, 当召回率为其他值时, 文中方法的精确率

均高于其他方法。

2.4 注意力可视化

为了更加直观地表现在交互模块中图像对文本和文本对图像的注意, 该文进行了注意力可视化分析, 结果如图 5 所示。



The major international competition in football is the World Cup, organized by FIFA

图 5 注意力可视化

由图 5 可知, 图像对文本的注意力主要集中在单词“football”“World”“Cup”和“FIFA”上, 即图中文字描述划线部分。文本对图像的注意力权重主要集中在球员、球迷和场地等部分, 即图中标注区域。

2.5 对照实验

考虑到特征提取器及注意力机制对整个检索模型性能的影响, 该文通过改变特征提取器类型和是否添加注意力机制等方式进行了一系列对照实验。为公平起见, 对于图像特征提取器为 CNN 类的实验, 该文均采用预训练好的 VGG16 的最后一个池化层作为图像特征向量。

实验结果如表 3 所示, 由方法二四六和方法一三五对比得知, 添加了注意力机制的方法在检索效果上显著优于没有添加注意力机制的方法。这是因为融合注意力的方法能选择性地关注不同模态数据间的重要信息部分, 进而提取到更完善的语义特征。通过方法一和三、二和四比较得知, 图像特征提取器为 ViT 模型类的方法与为 CNN 类的方法效果存在差异, 但差距并

表 3 对照实验 mAP 结果对比(Wikipedia 数据集)

	特征提取		注意力	图像检索文本	文本检索图像	平均值
方法一	CNN	LSTM		0.463	0.437	0.450
方法二	CNN	LSTM	✓	0.526	0.544	0.535
方法三	VIT	LSTM		0.482	0.465	0.473
方法四	VIT	LSTM	✓	0.594	0.578	0.586
方法五	VIT	Bert		0.579	0.553	0.566
方法六	VIT	Bert	✓	0.687	0.711	0.699

不明显。通过方法六和四、五和三对比得知,文本特征提取器为 Bert 类模型的方法比为 LSTM 类方法的效果更好,一是因为 Bert 模型是双向编码模型,能同时考虑上下文信息,具有更强大的语义提取能力,二是因为 Wikipedia 数据集中多以长文本为主,在处理长距离依赖问题上,Bert 模型有着更为优秀的表现。

3 结束语

针对图文检索研究,该文提出了一种融合注意力机制的跨模态检索算法。为了综合考虑全局特征和局部特征对检索效果的影响,基于注意力机制提取语义表达更充分的全局特征和局部特征,并将这些特征有机融合,使得模态数据特征信息表达更完善;同时,为了挖掘语义相似但模态不同的数据内在关系,通过注意力机制融合不同模态信息,从而提取更好的特征表示。实验证明,提出的算法优于目前已知方法,未来将针对文本描述为中文的图文检索做进一步研究。

参考文献:

- [1] 刘颖,郭莹莹,房杰,等.深度学习跨模态图文检索研究综述[J].计算机科学与探索,2022,16(3):489-511.
- [2] YAN F,MIKOLAJCZYK K. Deep correlation for matching images and text[C]//Computer vision & pattern recognition. Boston:IEEE,2015:3441-3450.
- [3] WANG J,HE Y,KANG C,et al. Image-text cross-modal retrieval via modality-specific feature learning[C]//Proceedings of the 5th ACM on international conference on multimedia retrieval. Shanghai:ACM,2015:347-354.
- [4] DONG X F,LIU L,ZHU L,et al. Adversarial graph convolutional network for cross-modal retrieval[J]. IEEE Transactions on Circuits and Systems for Video Technology,2022,32(3):1634-1645.
- [5] PENG Y,QI J,YUAN Y. CM-GANs:cross-modal generative adversarial networks for common representation learning[J]. ACM Transactions on Multimedia Computing Communications and Applications,2019,15(1):22.1-22.24.
- [6] BAHDANAU D,CHO K,BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv:1409.0473,2014.
- [7] NAM H,HA J W,KIM J. Dual attention networks for multimodal reasoning and matching[C]//IEEE conference on computer vision & pattern recognition. Hawaii:IEEE,2016:2156-2164.
- [8] 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:Springer,2018:201-216.
- [9] LI W,ZHENG Y,ZHANG Y,et al. Cross-modal retrieval with dual multi-angle self-attention[J]. Journal of the Association for Information Science and Technology,2021,72(1):46-65.
- [10] DOSOVITSKIY A,BEYER L,KOLESNIKOV A,et al. An image is worth 16x16 words:transformers for image recognition at scale[J]. arXiv:2010.11929,2020.
- [11] DEVLIN J,CHANG M W,LEE K,et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv:1810.04805,2018.
- [12] HOFFER E,AILON N. Deep metric learning using Triplet network[J]. arXiv:1412.6622v4,2014.
- [13] RASIWASIA N,PEREIRA J C,COVIELLO E,et al. A new approach to cross-modal multimedia retrieval[C]//Proceedings of the 18th international conference on multimedia 2010. Firenze:ACM,2010:251-260.
- [14] WANG K,YIN Q,WANG W,et al. A comprehensive survey on cross-modal retrieval[J]. arXiv:1607.06215,2016.
- [15] HARDOON R,SZEDMAK S,SHAW-TAYLOR J N. Canonical correlation analysis:an overview with application to learning methods[J]. Neural Computation,2004,16(12):2639-2664.
- [16] ANDREW G,ARORA R,BILMES J,et al. Deep canonical correlation analysis[C]//International conference on international conference on machine learning. Atlanta:JMLR.org,2013:2284-2292.
- [17] PEREIRA J C,COVIELLO E,DOYLE G,et al. On the role of correlation and abstraction in cross-modal multimedia retrieval[J]. IEEE Trans Pattern Anal Mach Intell,2014,36(3):521-535.
- [18] WANG B,YANG Y,XU X,et al. Adversarial cross-modal retrieval[C]//2017 ACM international conference on multimedia. New York:ACM,2017:154-162.
- [19] ZHEN L,HU P,WANG X,et al. Deep supervised cross-modal retrieval[C]//2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR). Long Beach:IEEE,2019:10394-10403.
- [20] ZHEN L,HU P,PENG X,et al. Deep multimodal transfer learning for cross-modal retrieval[J]. IEEE Transactions on Neural Networks and Learning Systems,2020,33(2):798-810.