基于多粒度匹配的文本引导服装图像检索

肖华兴,马丽丽,陈金广
(西安工程大学 计算机科学学院,陕西 西安 710048)

摘 要:文本引导的图像检索是将查询图像与文本条件集成为多模态查询。现有的方法通过构建更先进的细粒度度量学习来提升性能,但这可能会使模型在文本条件不够精确的情况下对目标图像过拟合,并使得检索结果特征单调。针对该问题,提出了基于特征增强和多粒度匹配的文本引导的服装图像检索方法。首先,根据目标特征的分布,产生服从正态分布的噪声,使其产生小幅度的类内抖动;然后,根据目标特征的波动对增强特征施加约束,波动越大,则对增强特征的惩罚越大,由此得到粗粒度匹配损失;最后,优化学习策略,使用随着训练迭代不断衰减的动态权重将粗粒度与细粒度损失进行统一。通过该方法降低模型对潜在目标图像的排斥,提高特征识别的多样化。在两个公开服装数据集 FashionIQ 和 Shoes 上的大量实验表明,该方法能够提高召回率,并且检索结果更丰富。

关键词:文本引导;图像检索;特征增强;多粒度匹配;多模态融合
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2024)07-0024-07
doi:10.20165/j.cnki.ISSN1673-629X.2024.0119

Text Guided Clothing Image Retrieval Based on Multi-granularity Matching

XIAO Hua-xing,MA Li-li,CHEN Jin-guang
(School of Computer Science,Xi'an Polytechnic University,Xi'an 710048,China)

Abstract:Text guided image retrieval integrates query images and text conditions into a multimodal query. Existing methods improve performance by constructing more advanced fine-grained metric learning,but this may cause the model to overfit the target image under imprecise text conditions and make the retrieval results feature monotonous. To address this issue,we propose a text guided clothing image retrieval method based on feature enhancement and multi granularity matching. Firstly,based on the distribution of target features,noise following a normal distribution is generated,causing small intra-class jittering. Then,constraints are imposed on the enhanced features based on the fluctuations of the target features. The larger the fluctuations,the greater the penalty for the enhanced features,resulting in coarse-grained matching losses. Finally,we optimize the learning strategy by using dynamic weights that continuously decay with training iterations to unify coarse-grained and fine-grained losses. The proposed method reduces the model's rejection of potential target images and improves the diversity of feature recognition. Extensive experiments on two publicly available clothing datasets,FashionIQ and Shoes,have shown that the proposed method can improve recall rates and provide richer retrieval results.

Key words:text guided;image retrieval;feature enhancement;multi-granularity matching;multi-modal fusion

0 引 言

传统的图像检索一般可分为图像-图像检索[1]和图像-文本检索[2],前者以图搜图,后者进行图文匹配,然而这种以单一模态作为输入的查询对用户意图的表达力有限。图文组合的多模态查询则能提供更多样化的信息,更能体现用户需求。图文组合的多模态查询也被称为文本引导的图像检索。文本引导的图像检索系统尤其适合基于对话的交互式检索场景,而这样的场景在电商领域极为常见。例如,用户希望改变查询图像某一服装属性而保持其它服装属性不变[3],那么用户可以基于这张图像输入想得到的属性的修改描述以获得更准确的检索结果。文本引导的服装图像检索在电商领域有巨大的发展潜力,因而具有重要的应用价值和研究意义。

文本引导的图像检索任务已经取得一定的进展。文献[4]最早提出解决该任务的方法,是将查询图像

收稿日期:2023-10-22 **修回日期**:2024-02-23
基金项目:陕西省自然科学基础研究计划项目(2023-JC-YB-568);陕西省教育厅科研计划项目(22JP028);陕西省计算机学会 & 翔腾公司基金项目(XT-QC-202309-119287)
作者简介:肖华兴(1997-),男,硕士研究生,研究方向为服装图像检索;通信作者:陈金广(1977-),男,博士,教授,研究方向为多目标跟踪。

的全局表示和文本表示通过残差连接和门控模块进行融合。与之类似的方法还有文献[5-6],都使用了门控机制。而文献[7-8]根据文本特征使用“两步修改”的方式对参考图像的表示进行调整。文献[9-12]使用了多级表示的方法进行多模态特征的融合。文献[13-14]都使用了双子网的结构,通过更高的计算量换取更高的性能。文献[15]提出了额外的约束来训练多模态网络。文献[16]通过文本、图像语义增强模块在特征融合时对组合特征进行语义增强。文献[17]提出动态的多专家协作网络(Adaptive Multi-expert Collaborative,AMC)来处理该任务,根据不同的图像文本查询组合对处理节点赋予不同的激活值。

然而目前基于文本引导的图像检索方法仍存在问题,这些问题和对应的解决思路如下:第一个问题,同一查询图像通常可能会与多个文本条件相关联,分别对应不同的目标图像,即,对同一查询图像,可能会受到多用户意图干扰的问题。借用 AMC 网络的动态交互层来解决这一问题,将多个不同结构的网络节点作为专家,并使用路由节点针对不同的组合查询生成不同的权重。然而,AMC 网络专注于一对一的细粒度匹配,当文本条件不够精确时,可能会造成对潜在目标图像的排斥。这就是第二个问题。在文本条件的表达不够精确或者是模棱两可的情况下,现有的方法只能有效识别数据集中唯一标注的目标图像或与其高度相似的其他图像。也就是说这些方法可能会导致模型对目标图像特征过拟合,而这种过拟合不仅会影响精度,而且会使模型的检索结果特征相对单调。

针对上述两个问题,该文在相关工作的基础上,提出了多粒度匹配的文本引导的服装图像检索方法。总的来说,贡献如下:

(1)将噪声增强目标特征的方法应用到 AMC 网络中,使模型能有效抵抗多用户意图干扰,针对不同查询组合进行自适应调整,同时使模型在面对不够精确的粗粒度检索时,减少对潜在目标图像的排斥。

(2)优化了学习策略中的目标函数,对于细粒度匹配的 infoNCE(info Noise Contrastive Estimation)损失[18],引入噪声增强特征后,再根据目标特征的波动施加约束,波动越大,对增强特征的惩罚越大,通过这种方法使其转变为粗粒度损失。然后在训练时,使用不断衰减的动态权重来减弱粗粒度匹配,同时增强细粒度匹配;并且移除了与噪声起相反效果的 BSC(Batch-based Similarity Consistency)损失。

1　方　法

1.1　问题描述

文本引导的图像检索任务如图 1 所示。该任务的输入是由一张查询图像和用户提供的文本条件组成,文本条件描述了用户想得到的属性特征,目标是要在图像库中检索到与查询图像相似但尽可能符合文本描述的图像。将该任务形式化如下:给定一批三元组 $D=\{(I_s,T_s,I_t)_i\}_{i=1}^{N}$,其中 I_s 表示查询图像,T_s 表示查询文本,I_t 表示目标图像,N 表示三元组的数量;要执行的任务是将查询图像特征与查询文本特征融合,使其尽可能接近目标图像。公式表示为:

$$H(I_s,T_s)\rightarrow F(I_t) \tag{1}$$

其中,H 表示多模态查询的变换空间,F 表示目标图像的变换空间。

图 1　文本引导的图像检索示意图

1.2　模型总体框架

模型总体框架如图 2 所示。使用了 AMC 网络的动态交互层,增添了噪声增强模块,并采用了新的学习策略。该模型可以分为 4 个部分:(1)图像编码器;(2)文本编码器;(3)图像文本合成模块;(4)噪声增强模块。源图像(即查询图像)与目标图像的编码器相同且共享权重,图像文本合成模块由两个动态交互层组成,噪声增强模块对目标特征进行处理。将在后续详细介绍每个模块及损失函数。

1.3　图像和文本编码器

查询图像和目标图像使用共享权重的 CNN,将其变换表示为 $f_{img}(\cdot)$,为方便后续处理,目标表示需要再经过平均池化以降维,表示为:

$$\begin{cases} f_s^I = f_{img}(I_s) \\ f_t = \mathrm{pool}_{avg}(f_{img}(I_t)) \end{cases} \tag{2}$$

其中,$f_s^I \in \mathbb{R}^{H\times W\times D}$,$f_t \in \mathbb{R}^{H\times D}$ 分别为查询图像与目标图像的中间表示。而对于查询文本,对其进行一定的预处理,将查询文本拆分为单词,并统计每个单词出现的次数,再使用嵌入层获取对应的词向量,最后使用文本编码器得到文本特征,将这些变换表示为 $f_{text}(\cdot)$,则:

$$f_s^t = f_{text}(T_s) \tag{3}$$

其中,$f_s^t \in \mathbb{R}^D$ 表示变换后的文本特征。

1.4　图像文本合成模块

图像文本合成模块包括两个动态交互层,动态交互层的设计采用了 AMC 网络的方法,由 1 个路由节点和 3 个专家节点组成,这些专家节点用于处理图像、文本特征。第一个节点是 NIN(Normalized Identity

Node),即归一化身份节点。使用该节点保留和归一化上一层的输出。公式表示为:

$$F_1(X) = N(X) \tag{4}$$

其中,$F_1(\cdot)$ 表示第一个节点 NIN 执行的函数,X 表示输入当前节点的中间特征,$N(\cdot)$ 表示层归一化。

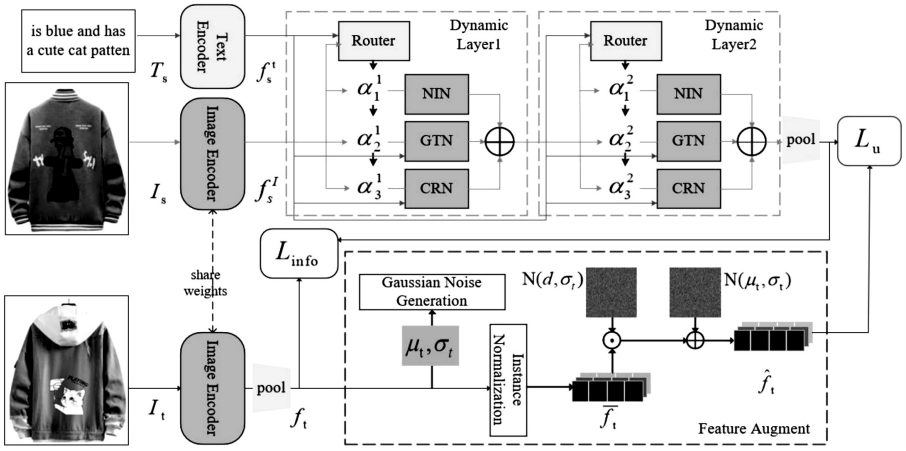

图 2 整体模型结构

第二个节点是 GTN(Global Transformation Node),即全局转换节点,作用是根据文本特征,有选择性地抑制或突出当前图像特征,达到对视觉表示整体上的修改,也可以说是对视觉局部特征的粗粒度的调整。具体的做法是根据文本特征 f_s^t 生成移动向量 β 和缩放向量 γ,公式表示为:

$$\begin{cases}\gamma = \boldsymbol{W}_\gamma f_s^t + b_\gamma \\ \beta = \boldsymbol{W}_\beta f_s^t + b_\beta\end{cases} \tag{5}$$

其中,$\boldsymbol{W}_\gamma, \boldsymbol{W}_\beta \in \mathbb{R}^{D\times D}$ 为可学习的参数矩阵,b_γ,b_β 为偏置参数,通过线性层来实现,然后使用移动向量 β 和缩放向量 γ 对输入当前节点的中间特征 X 应用仿射变换,公式表示为:

$$F_2(X) = N(\gamma \odot (X) + \beta) \tag{6}$$

其中,$F_2(\cdot)$ 表示第二个节点 GTN 执行的函数,$\odot$ 表示矩阵按元素相乘。

第三个节点是 CRN(Cross - modal Reasoning Node),即跨模态推理节点,作用是根据查询文本表示对图像特征进行视觉语义上的细粒度调整。具体操作是先将文本特征 f_s^t 与输入当前节点的中间特征 X 拼接,将拼接后的特征再输入到一个线性层中,公式表示为:

$$X_c = W[X; f_s^t] \tag{7}$$

其中,$[;]$ 表示矩阵在特征维度进行拼接,$W \in \mathbb{R}^{2D\times D}$ 表示该线性层执行的操作,然后将拼接后的特征 X_c 映射到 $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ 这 3 个不同的空间中,公式表示为:

$$\begin{cases}\boldsymbol{Q} = X_c \boldsymbol{W}_Q \\ \boldsymbol{K} = X_c \boldsymbol{W}_K \\ \boldsymbol{V} = X_c \boldsymbol{W}_V\end{cases} \tag{8}$$

其中,$\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V \in \mathbb{R}^{D\times D}$,是可学习的参数矩阵,通过线性层来实现,然后构建多头自注意力,公式表示为:

$$\text{head}_i = \text{softmax}\left(\frac{Q_i \boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{d_k}}\right) V_i \tag{9}$$

其中,head_i 表示第 i 个头,d_k 代表矩阵 $\boldsymbol{K}$ 的维度,将所有分头拼接得到多头自注意力函数,公式表示为:

$$\text{MutiHead}(X) = [\text{head}_1; \cdots; \text{head}_h] \tag{10}$$

其中,X 为输入当前节点的中间特征,h 为总头数。则第三个节点 CRN 执行的函数可以表示为:

$$F_3(X) = N(\text{FFN}(\text{MutiHead}(X)) + \text{MutiHead}(X)) \tag{11}$$

其中,$F_3(\cdot)$ 表示第三个节点 CRN 执行的函数,FFN(·) 为一个两层的带 ReLu 激活函数的多层感知机,MutiHead(·) 为多头自注意力函数。

路由节点的作用是根据当前查询表示文本特征,对当前层的 3 个专家节点,生成相应的路径激活值,将 3 个专家节点的输出作加权聚合。每个动态交互层的路径激活值可以表示为:

$$\alpha^l = R^l(X^{l-1}, f_s^t) \tag{12}$$

其中,$R^l(\cdot,\cdot)$ 表示第 l 个动态交互层的路由函数,α^l 为第 l 个动态交互层输出的激活值,X^{l-1} 为第 $l-1$ 个动态交互层输出的中间特征表示,f_s^t 为文本特征。第 l 个动态交互层的输出 X^l 可以表示为:

$$X^l = \begin{cases} f_s^t, & l = 0 \\ \sum_{i=1}^{3} \alpha_i^l P_i^l, & l \geqslant 1 \end{cases} \tag{13}$$

其中,$X^l \in \mathbb{R}^{H\times W\times D}$,$\alpha_i^l, P_i^l$ 分别表示第 l 个动态交互层的第 i 个节点的激活值和输出特征。每个动态层的路由函数的实现具体为:首先将第 l 个动态交互层的输出 X^l 与文本特征 f_s^t 拼接,表示为:

$$r = [X^l; f_s^t] \in \mathbb{R}^{K\times 2D} \tag{14}$$

其中, r 为拼接后的特征, $K = H \times W$,将 r 在维度 K 上的所有分量 r_i 累加,再经过线性层和层归一化,公式表示为:

$$\psi = N(\boldsymbol{W}_1(\frac{1}{K}\sum_{i=1}^{K} r_i)) \tag{15}$$

其中, $\boldsymbol{W}_1 \in \mathbb{R}^{2D\times\frac{D}{2}}$ 为参数矩阵,表示线性层执行的操作, $N(\cdot)$ 为层归一化, ψ 为得到的特征,则第 l 个动态交互层的路由函数 $R^l(\cdot,\cdot)$ 可以表示为:

$$R^l(X^{l-1}, f_s^t) = \sigma(\boldsymbol{W}_2(\xi(\psi))) \tag{16}$$

其中, X^{l-1} 为第 $l-1$ 个动态交互层输出的中间特征表示, f_s^t 为文本特征, σ 为 sigmoid 函数, σ 为 Relu 函数 , $\boldsymbol{W}_2 \in \mathbb{R}^{2D\times 3}$ 为参数矩阵,由线性层实现。

1.5　噪声增强模块

噪声增强模块的作用是对目标特征添加噪声,目的是让模型在输入查询文本不够精确的条件下,能够识别更多潜在的目标图像,或者说使模型减少对目标特征的过拟合。该文只添加少量的噪声,使目标特征产生小范围的抖动以保证噪声增强后的特征与原目标特征仍属于同一类,但又具有细节差异。噪声增强后的特征 $\hat{f}_t$ 可以表示为:

$$\hat{f}_t = \alpha \cdot \bar{f}_t + \beta \tag{17}$$

其中, $\bar{f}_t$ 是对目标特征 f_t 应用一维实例归一化后获得, $\alpha \sim N(d,\sigma_t)$, $\beta \sim N(\mu,\sigma_t)$ 分别为噪声参数,它们服从正态分布,与目标特征 f_t 形状相同,其中 d 为超参数, μ,σ_t 分别为目标特征的均值、标准差。通过这种方式,使噪声增强后的特征与原目标特征的分布相接近。

1.6　损失函数

现有的方法大多采用基于批的分类(batch-based classification)损失,也称为 InfoNCE 损失,是常用的一种对比学习损失,表示为:

$$L_{\text{info}}(f_s, f_t) = \frac{1}{B}\sum_{i=1}^{B} -\log\frac{\exp(\kappa(f_s^i, f_t^i))}{\sum_{j=1}^{B}\exp(\kappa(f_s^i, f_t^j))} \tag{18}$$

其中, f_s 为合成的图像文本特征, f_t 为目标特征, B 为批大小, f_s^i, f_t^i 分别为一批三元组中的第 i 个图文组合特征和目标特征, κ 为相似核,采用余弦相似度。该损失函数会驱使 $\kappa(f_s^i, f_t^i)$ 尽可能大,同时 $\kappa(f_s^i, f_t^j)$, $(i \neq j)$ 尽可能小。

InfoNCE 损失仅关注于一对一的细粒度匹配,而添加噪声后,目标特征产生类内抖动,相当于变为一对多的粗粒度匹配,需要对其做调整以进行统一。调整后的损失为:

$$L_u(f_s, \hat{f}_t, \sigma) = \frac{L_{\text{info}}(f_s, \hat{f}_t)}{2\sigma^2} + \frac{1}{2}\log\sigma^2 \tag{19}$$

其中, σ 为目标特征的标准差。若目标特征波动较大,则 InfoNCE 损失的权重会变小,反之,若目标特征变换较小,则 L_u 接近原始的 L_{info} 。然后将 L_u , L_{info} 进行统一:

$$L_{\text{total}} = \gamma L_u(f_s, \hat{f}_t, \sigma) + (1-\gamma)L_{\text{info}}(f_s, f_t) \tag{20}$$

γ 表示为:

$$\gamma = \exp(-\gamma_0 \cdot \frac{p_1 + c}{p_2}) \tag{21}$$

其中, p_1 表示当前训练的轮数, p_2 表示训练的总轮数, γ_0, c 为固定参数, $\gamma \in [0,1]$ 。随着训练轮数的增加, γ 会不断变小,即粗粒度匹配的权重不断减小,而细粒度匹配权重不断增加。这样做的目的是使模型在不失精度的同时能够产生更多样性的结果。

2　实　验

2.1　数据集

在时尚服装数据集 FashionIQ[19] 和 Shoes[20] 上评估该方法。FashionIQ 包含 77 684 张服装图像,分为 3 个类别:Dress, Shirt, Toptee。文本条件的平均长度为 10.69 个单词。由于该数据集的测试集的真值未发布,使用验证集代替。与现有的工作[17]保持相同的训练集、测试集划分,结果共有 18 000 个训练三元组和 6 016个测试三元组。Shoes 的训练集包含 10K 张鞋子图像,测试集包含 4 658 张鞋子图像。文本条件描述了查询图像与目标图像间的细粒度视觉差异,平均长度 5.32 个单词。

2.2　实验设置

图像编码器采用在 ImageNet 上进行预训练的 ResNet50,但删除了最后的分类层以保留原始特征。文本编码器由一个嵌入层和 LSTM 网络组成。该嵌入层由 GloVe 词嵌入进行初始化。对于训练,使用 Adam 优化器,权重衰减因子为 1e-6,训练周期为 60 轮。学习率为 1e-4,每 40 轮衰减 10 倍。批大小 B 为 32,编码维度为 1 024。超参数 $\gamma_0 = 2, c = 1, d = 1$ 。实验环境为 Python 3.8, PyTorch1.12.1, 实验设备为 NVIDIA Geforce RTX 3080 Laptop GPU。

实验使用的评价指标为 Recall@K,即按与目标图像的相似度排序的前 K 个检索结果的召回率,也称为查全率,是指前 K 个检索结果中的相关结果数与库中所有相关结果数的比率。计算公式为:

$$\text{Recall@K} = \frac{1}{n}\sum_{i=1}^{n}\text{score} \tag{22}$$

其中, n 为查询的总数,对于每次查询,若前 K 个检索结果中存在目标图像,则此次查询的 score 记为 1,否则记为 0。Recall@K 为测试集中所有查询的 score 的平均。使用 Recall@K 指标可以有效地评估检索模型

的查找能力,即能否在前 K 个检索结果中找到用户感兴趣的目标图像。

2.3 实验结果与分析

表 1 显示了在数据集 FashionIQ 上的结果,该模型在绝大部分指标上都优于现有的方法,只在 Dress 子集的 R@50 指标上略低于 AMC 网络。总体上,在平均召回率 R@10,R@50 上,该方法都实现了最先进的性能,并且对于平均召回率 R@10 提升比较明显,而 R@50 提升相对较小。能够得到性能提升与该模型的结构设计有关,因为每次添加的噪声幅度都较小,增强后的目标特征变化较小,以此来模拟各种潜在的目标图像(未被标记但又符合查询条件)。随着训练迭代,粗粒度损失比重不断减小逐渐趋近于 0,同时细粒度损失比重不断增大逐渐趋近于 1,这样得到的模型与未添加噪声的模型相比首先性能不会下降,因为未添加噪声的情况就相当于只有细粒度损失。在这种情况下,该模型提高了泛化能力,增强了识别潜在目标图像的能力,同时随着训练迭代,使得模型相较于潜在目标图像不断提高对目标图像的优先级,即,将目标图像排序在更靠前的位置。所以,该模型得到了性能的提升,且 R@K 的 K 值较小时,提升相对更明显。

表 1 数据集 FashionIQ 上的结果

Methods	Dress		Toptee		Shirt		Average	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
VAL[9]	21.12	42.19	25.64	49.49	21.03	43.44	22.60	45.04
DATIR[10]	21.90	43.80	27.20	51.60	21.90	43.70	23.70	46.40
JPM[15]	21.38	45.15	27.78	51.70	22.81	45.18	23.99	47.34
MAAF[11]	23.80	48.60	27.90	53.60	21.30	44.20	24.30	48.80
MCR[12]	26.20	51.20	29.70	56.40	22.40	46.00	26.10	51.20
CoSMo[7]	25.64	50.30	29.21	57.46	24.90	49.18	26.58	52.31
DCNet[13]	28.95	56.07	30.44	58.29	23.95	47.30	27.78	53.89
SAC w/BERT[8]	26.52	51.01	32.70	61.23	28.02	51.86	29.08	54.70
CLVC-Net[14]	29.85	56.47	33.50	64.00	28.75	54.76	30.70	58.41
AMC[17]	31.18	**59.20**	35.60	65.88	29.78	57.56	32.19	60.88
Ours	**31.88**	58.85	**36.51**	**66.45**	**30.42**	**57.80**	**32.93**	**61.03**

在数据集 Shoes 上的结果如表 2 所示。该模型在所有指标上都实现了最高的性能。与基准模型 AMC 相比,在 R@1,R@10 指标上提升相对较多,在 R@50 上提升相对较少,这与在数据集 FashionIQ 上得到的结果类似,说明在检索时,该模型能够将相似度较高的目标图像排在更靠前的位置。

表 2 数据集 Shoes 上的结果

Method	Shoes			
	R@1	R@10	R@50	Average
VAL[9]	16.49	49.10	73.53	46.37
CoSMo[7]	16.72	48.36	75.64	46.91
MAAF[11]	16.45	49.95	76.36	47.58
DATIR[10]	17.20	51.10	75.60	47.97
MCR[12]	17.85	50.95	77.24	48.68
SAC w/ BERT[8]	18.50	51.73	77.28	49.17
CLVC-Net[14]	17.64	54.39	17.64	50.50
AMC[17]	25.85	64.31	86.19	58.78
Ours	**26.77**	**65.00**	**86.31**	**59.36**

2.4 消融实验

如表 3 所示,在数据集 FashionIQ 上进行了消融研究,以探索不同配置对模型的影响。对噪声增强模块进行消融,移除之后,平均召回率 R@10 和 R@50 都有下降,而 R@10 下降得更多,说明使用噪声增强目标特征的方法对提升性能是有效的。移除噪声增强模块之后,再添加 BSC 损失,这种架构即为 AMC 网络,可以看出添加该损失后,性能有所提升,但仍低于文中

模型,而在文中模型的基础上添加 BSC 损失,结果性能下降。这是因为 BSC 损失是用来加强一对一的关系,而对目标特征添加噪声,增加了一对多的关系,两者的效果相互矛盾。

2.5　定性比较

图 3、图 4 展示了文中方法与 AMC 网络分别在数据集 FashionIQ 和数据集 Shoes 上的定性结果比较,图中列出了前五的检索结果。图 3 的查询图像为身着条纹长裙的女性,查询文本要求“有细带和不同的图案,有更多秋天元素的色彩且更长”。可以看出文中方法将目标图像排在了更靠前的位置,并且检索结果满足了绝大部分的要求,而 AMC 网络第五个检索结果是蓝紫色的裙子,不符合秋天色彩这一要求。图 4 的文本描述是“高过脚裸且有黑红对比”,虽然文中模型与 AMC 网络都将目标图像排在了首位,但文中模型的检索结果的第三、四张图像都符合“高过脚裸”这一描述,且有相近的棕红对比。而 AMC 网络的后三张图像相同且不符合“高过脚裸”这一描述。总体上,文中方法能够将符合用户需求的图像排在更靠前的位置,检索结果相对更符合文本描述且更具多样性。

表 3　FashionIQ 数据集上的消融实验

Methods	Dress		Toptee		Shirt		Average	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
w/o Augment	30.19	58.60	34.93	65.47	28.80	57.36	31.30	60.47
w/o Augment+bsc_loss	31.18	59.20	35.60	65.88	29.78	57.56	32.19	60.88
Ours+bsc_loss	29.64	58.05	**36.86**	65.98	28.94	57.45	31.81	60.49
Ours	**31.88**	58.85	**36.51**	**66.45**	30.42	**57.80**	**32.93**	**61.03**

图 3　数据集 FashionIQ 上的检索结果比较

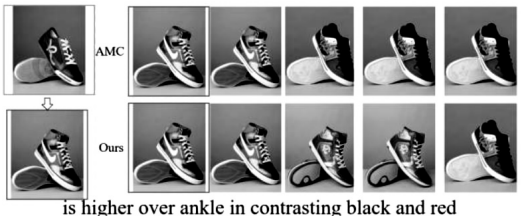

图 4　数据集 Shoes 上的检索结果比较

然而文中方法仍存在一定的局限性,例如对标签过度的依赖,需要对数据集做繁复且精确的注释,因为添加的噪声完全由目标图像特征的分布决定,所以不能用于无监督的多模态融合算法[21]。并且对目标特征进行噪声增强后对性能的提升幅度有限,特别是在 Recall@K 的 K 值较大时。

3　结束语

在这项工作中,专注于文本引导的图像检索任务,提出了一种噪声增强方法,并将其用于自适应多专家协作网络中。具体地,对目标特征进行采样获得服从正态分布的噪声,然后在学习策略上,根据目标特征的波动,对用于细粒度检索的 InfoNCE 损失做一定的惩罚,使其转变为用于一对多匹配的粗粒度损失,最后使用动态权重将其统一。在数据集 FashionIQ 和 Shoes 上验证了该方法的有效性,特别是该方法能将符合用户需求的图像排在更靠前的位置,且检索结果更具多样性。

参考文献:

[1] 张舜尧,李华旺,张永合,等. 基于独立注意力机制的图像检索算法[J]. 计算机科学,2023,50(S1):328-333.

[2] 缪岚芯,雷　雨,曾鹏鹏,等. 基于粒度感知和语义聚合的图像-文本检索网络[J]. 计算机科学,2022,49(11):134-140.

[3] 丁安安. 融合多模态特征的细粒度服装图像检索[D]. 上海:东华大学,2023.

[4] NAM V,JIANG L,SUN C,et al. Composing text and image for image retrieval-an empirical odyssey[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach:IEEE,2019:6439-6448.

[5] 宣益亮. 基于深度学习的服装图像检索方法研究与实现[D]. 上海:东华大学,2022.

[6] 王依凡. 基于多模态特征融合的图像文本检索方法研究[D]. 成都:电子科技大学,2023.

[7] LEE S,KIM D,HAN B. Cosmo:content-style modulation for image retrieval with text feedback[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville:IEEE,2021:802-812.

[8] SURGAN J,PINKESH B,PRANIT C,et al. SAC:semantic attention composition for text-conditioned image retrieval[C]//Proceedings of the IEEE/CVF winter conference on computer vision and pattern recognition. Hawaii:IEEE,

2022:4021-4030.
[9] CHEN Y,GONG S,LORIS B. Image search with text feedback by visiolinguistic attention learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle:IEEE,2020:3001-3011.
[10] GU C,BU J,ZHANG Z,et al. Image search with text feedback by deep hierarchical attention mutual information maximization[C]//Proceedings of the 29th ACM international conference on multimedia. Virtual:ACM,2021:4600-4609.
[11] ERIC D,JACK C,SIMAO H,et al. Modality-agnostic attention fusion for visual search with text feedback[J]. arXiv:2007.00145,2020.
[12] ZHANG G,WEI S,PANG H,et al. Heterogeneous feature fusion and cross-modal alignment for composed image retrieval[C]//Proceedings of the 29th ACM international conference on multimedia. Virtual:ACM,2021:5353-5362.
[13] JONGSEOK K,YU Y,HOESEONG K,et al. Dual compositional learning in interactive image retrieval[C]//Proceedings of the AAAI conference on artificial intelligence. Virtual:AAAI,2021:1771-1779.
[14] WEN H,SONG X,YANG X,et al. Comprehensive linguistic-visual composition network for image retrieval[C]//Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. Virtual:ACM,2021:1369-1378.
[15] YANG Y,WANG M,ZHOU W,et al. Cross-modal joint prediction and alignment for composed query image retrieval[C]//Proceedings of the 29th ACM international conference on multimedia. Virtual:ACM,2021:3303-3311.
[16] 杨 帆,宁 博,李怀清,等. 基于语义增强特征融合的多模态图像检索模型[J]. 浙江大学学报:工学版,2023,57(2):252-258.
[17] ZHU H,WEI Y,ZHAO Y,et al. Amc:adaptive multi-expert collaborative network for text-guided image retrieval[J]. ACM Transactions on Multimedia Computing,Communications and Applications,2023,19(6):1-22.
[18] HAN X,YU L,ZHU X,et al. Fashionvil:fashion-focused vision-and-language representation learning[C]//Proceedings of the 17th European conference on computer vision. Tel Aviv:Springer,2022:634-651.
[19] WU H,GAO Y,GUO X,et al. The fashioniq dataset:retrieving images by combining side information and relative natural language feedback[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville:IEEE,2021:11302-11312.
[20] GUO X,WU H,CHENG Y,et al. Dialog-based interactive image retrieval[C]//Proceedings of the neural information processing systems. Montreal:Curran,2018:676-686.
[21] 毛鑫鑫. 基于多模态融合的图像相似性度量算法的研究与应用[D]. 株洲:湖南工业大学,2023.