工程科学学报
Chinese Journal of Engineering

嵌入共识知识的因果图文检索方法

梁彦鹏 刘雪儿 马忠贵 李卓

Causal image-text retrieval embedded with consensus knowledge

LIANG Yanpeng, LIU Xueer, MA Zhonggui, LI Zhuo

引用本文:
梁彦鹏, 刘雪儿, 马忠贵, 李卓. 嵌入共识知识的因果图文检索方法[J]. 工程科学学报, 2024, 46(2): 317–328. doi: 10.13374/j.issn2095–9389.2023.05.28.001
LIANG Yanpeng, LIU Xueer, MA Zhonggui, LI Zhuo. Causal image–text retrieval embedded with consensus knowledge[J]. *Chinese Journal of Engineering*, 2024, 46(2): 317–328. doi: 10.13374/j.issn2095–9389.2023.05.28.001

在线阅读 View online: https://doi.org/10.13374/j.issn2095–9389.2023.05.28.001

工程科学学报, 第 46 卷, 第 2 期: 317−328, 2024 年 2 月
Chinese Journal of Engineering, Vol. 46, No. 2: 317−328, February 2024
https://doi.org/10.13374/j.issn2095-9389.2023.05.28.001; http://cje.ustb.edu.cn

嵌入共识知识的因果图文检索方法

梁彦鹏, 刘雪儿, 马忠贵✉, 李　卓

北京科技大学计算机与通信工程学院, 北京 100083
✉通信作者, E-mail: zhongguima@ustb.edu.cn

摘　要　跨模态图像−文本检索是一项在给定一种模态(如文本)的查询条件下检索另一种模态(如图像)的任务. 该任务的关键问题在于如何准确地测量图文两种模态之间的相似性, 在减少视觉和语言这两种异构模态之间的视觉语义差异中起着至关重要的作用. 传统的检索范式依靠深度学习提取图像和文本的特征表示, 并将其映射到一个公共表示空间中进行匹配. 然而, 这种方法更多地依赖数据表面的相关关系, 无法挖掘数据背后真实的因果关系, 在高层语义信息的表示和可解释性方面面临着挑战. 为此, 在深度学习的基础上引入因果推断和嵌入共识知识, 提出嵌入共识知识的因果图文检索方法. 具体而言, 将因果干预引入视觉特征提取模块, 通过因果关系替换相关关系学习常识因果视觉特征, 并与原始视觉特征进行连接得到最终的视觉特征表示. 为解决本方法文本特征表示不足的问题, 采用更强大的文本特征提取模型 BERT(Bidirectional encoder representations from transformers, 双向编码器表示), 并且嵌入两种模态数据之间共享的共识知识对图文特征进行共识级的表示学习. 在 MS-COCO 数据集以及 MS-COCO 到 Flickr30k 上的跨数据集实验, 证明了本文方法可以在双向图文检索任务上实现召回率和平均召回率的一致性改进.

关键词　因果推断; 图像−文本检索; 跨模态; 计算机视觉; 自然语言处理
分类号　TP391.1; TP391.4

Causal image-text retrieval embedded with consensus knowledge

LIANG Yanpeng, *LIU Xueer*, *MA Zhonggui*✉, *LI Zhuo*

School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
✉Corresponding author, E-mail: zhongguima@ustb.edu.cn

ABSTRACT　Crossmodal image-text retrieval involves retrieving relevant images or texts based on a query condition from the opposite modality. Its primary challenge lies in precisely quantifying the similarity metric used for feature matching between the two distinct modalities, playing an important role in mitigating the visual-semantic disparities between the heterogeneous realms of visual and linguistic domains. It has extensive applications in domains such as e-commerce product search and medical image retrieval. Traditional retrieval paradigms depend on harnessing deep learning techniques for extracting feature representations from images and texts. Crossmodal image-text retrieval learns semantic feature representations of disparate modal data by harnessing the formidable feature–extraction ability, subsequently mapping them into a shared semantic space for semantic alignment. However, this approach primarily depends on superficial data correlations, lacking the capacity to reveal the latent causal relationships underpinning the data. Moreover, owing to the inherent "black-box" nature of deep learning, the interpretability of model predictions often eludes human comprehension. In addition, an undue reliance on training data distributions impairs the generalization performance of the model. Consequently, the existing methods suffer the challenge of representing high-level semantic insights while maintaining interpretability. Causal inference, which endeavors to ascertain the causal effect of specific phenomena by isolating confounding factors by means of intervention, presents a novel avenue for enhancing the generalization capability and interpretability of deep models. Recently,

收稿日期: 2023−05−28
基金项目: 中央高校基本科研业务费专项资金资助项目(FRF-DF-20-12, FRF-GF-18-017B)

researchers have sought to combine visual and linguistic tasks with the principles of causal inference. Accordingly, we introduce causal inference and embeds consensus knowledge into the bedrock of deep learning, and a novel causal image-text retrieval methodology with embedded consensus knowledge is proposed. Specifically, causal intervention is introduced into the visual feature extraction module, replacing correlated relationships with causal counterparts to cultivate common causal visual features. These features are then fused with the primal visual features acquired through bottom-up attention, resulting in a definitive visual feature representation. This study adopts the potent textual feature extraction ability of bidirectional encoder representations from transformers to address the shortfall in textual feature representation. Shared consensus knowledge between the two modal data is entwined, allowing for consensus-level feature representation learning image-text features. Empirical validation on the dataset MS-COCO and crossdataset experiments on the dataset Flickr30k substantiate the capacity of the proposed method to consistently enhance recall and mean recall in bidirectional image-text retrieval tasks. In summary, this pioneering approach endeavors to bridge the gap between visual and textual representations by combining causal inference principles and shared consensus knowledge within the framework of deep learning, thereby promising enhanced generalization and interpretability.

KEY WORDS causal inference; image-text retrieval; crossmodality; computer vision; natural language processing

在最近几年, 深度学习在计算机视觉[1] 和自然语言处理[2−3] 领域均取得了巨大的成功. 对视觉和语言的理解是人类感知现实世界的基础, 人类可以很容易地描述给定图像中的内容或者根据一段描述选择对应的图像. 为了使机器能更全面地理解视觉和语言模态, 将视觉特征与文本特征进行匹配以实现跨模态的图文检索[4] 引起了学术界和工业界的广泛关注.

跨模态图文检索在电子商务产品搜索[5]、医学图像检索[6] 等领域有着广泛的应用. 现有的大部分研究都集中在利用深度学习方法进行图文检索, 通过其强大的特征提取能力学习不同模态数据的特征表示, 然后映射到一个公共表示空间中进行匹配. 然而, 这种基于深度学习的图文检索方法更多地依赖数据表面的相关关系, 而无法挖掘数据背后真实的因果关系[7]; 同时由于深度学习的“黑盒特性”, 模型最终的预测结果对人类来说往往不具备可解释性[8−9], 并且过度依赖训练过程中的数据分布也导致模型的泛化性能较差.

虽然利用相关关系建模的深度学习取得了一系列的成就, 但是相关并非因果, 真正的知识是根据因果关系得到的知识[10]. 因果推断可以通过前后门调整、反事实等操作去除混杂因素的影响[11], 进一步学习隐藏在相关关系背后的因果关系, 从而给出更具可靠性和可解释性的结果. 为此, 本文在深度学习的基础上引入因果推断和共识知识, 构建嵌入共识知识的因果图文检索方法. 具体而言, 在视觉特征提取部分利用因果干预提取常识因果视觉特征, 并与原始的使用 BUTD(Bottom-up and top-down attention)[12] 提取的特征连接. 为了在文本检索图像任务上提升召回率和平均召回率, 我们在引入因果推断的基础上, 针对文本特征的提取做出改进, 引入更强大的文本特征提取模型 BERT(Bidirectional encoder representations from transformers)[13] 学习文本特征表示, 并在模型外部嵌入共识知识, 通过学习图像和文本之间的共识减少图文特征间的语义差异, 从而提高双向图文检索的性能.

本文的主要贡献如下: (1)首次尝试将因果推断中的干预机制引入视觉特征提取模块进行跨模态图文检索. 与现有的基于深度学习的方法不同, 本文提出的模型利用因果干预去除混杂因素以得到常识因果视觉特征. 学习隐藏在相关关系背后的因果关系, 以进一步增强模型的鲁棒性和可解释性. (2)引入外部共识知识, 通过图卷积神经网络学习语料库中的共识表示并进一步生成共识级图文特征表示, 使用注意力机制进行共识级特征与原始实例级特征的融合, 进一步增强模型的检索与泛化性能. (3)在 MS-COCO 数据集上的实验结果证明了本文所提出的嵌入共识知识的因果图文检索方法相比于基准方法可以在双向图文检索任务上实现召回率($R@k$)和平均召回率(mR)的一致性改进. 在 MS-COCO[14] 到 Flickr30k[15] 数据集上的跨域实验证明了本文方法的泛化性能.

1 相关工作

1.1 跨模态图文检索

基于深度学习的图文检索方法利用深度学习分别提取图文模态数据特征的有效表示, 通过计算其相似性进行匹配. 根据图文特征提取方式和匹配方式的不同可以将现有的图文检索方法分为 3 种: (1)基于全局特征匹配的方法[16]: 从整张图像

和文本语句中提取整体的语义信息, 然后将图文特征映射到公共表示空间中, 利用损失函数进行优化使图文模态进行语义对齐和匹配. (2)基于局部特征匹配的方法[17−19]: 此类方法更关注细粒度的图像区域和文本单词的对齐, 通过局部对齐最终整合得到整体的图文相似性. (3)基于外部知识的方法[20−22]: 方法(1)和(2)都是依赖实例级图像文本对进行特征表示学习, 而没有考虑任何外部知识, 导致模型在推理图文数据间的高级语义关系时存在一定的欠缺, 此类方法通过引入场景图或模态间共享的共识知识等外部先验信息来增强语义表示. 例如, 文献 [21] 通过构建多模态知识图和多模态图卷积网络, 利用两个模态实体之间的隐含语义关系来增强图像和文本的嵌入. Wang 等[22] 提出了共识感知视觉语义嵌入(Consensus-aware visual-semantic embedding, CVSE), 将图文两种模态数据之间共享的共识知识整合到图文匹配中, 提高了模型的检索精度和泛化能力. 但由于外部知识与图文匹配任务数据集之间的域差异, 可能会对匹配结果产生影响. 总而言之, 基于深度学习的跨模态图文检索方法得到了广泛地应用, 但是这些方法仅仅是建模数据间虚假的相关关系而无法推断出具有说服力的因果关系, 而且模型在可解释性和泛化能力方面的问题并没有得到有效地解决.

1.2 因果推断在视觉和语言任务中的应用

因果推断[10] 旨在通过干预措施去除混杂因素来追求特定现象的因果效应, 成为提高深度学习模型泛化能力和可解释性的新方法. 其已经成功地应用于心理学、经济学[23] 和流行病学[24], Angrist 和 Imbens 也因“对因果关系分析的方法学贡献”而获得 2021 年的诺贝尔经济学奖. 近年来, 研究者们尝试将因果推断引入视觉和语言任务中, 包括目标检测[25]、视觉问答[26] 和图像字幕[27] 等. 具体来说, 文献 [28] 从因果关系的角度制定了 OOD(Out-of-distribution)推荐问题, 用户特征转移被表示为干预, OOD 推荐旨在估计干预后的交互概率, 并利用反事实推理来减轻过时交互的影响. 文献 [29] 通过因果干预代替常规似然以无监督的方式学习视觉常识特征. 文献 [30] 引入了一种新的域自适应模型来探索特定目标在不同天气条件下的不变特征, 用以在多种不利的天气条件下进行自动驾驶场景下[31] 的目标检测. 文献 [32] 提出通过反事实推理和因果干预, 减轻多标签分类任务中的上下文偏见. 文献 [33] 提出了因果注意模块(Causal attention module, CaaM),该模块以无监督的方式自我注释混杂因素, 并且多个 CaaM 模块可以堆叠并集成在常规 CNN(Convolutional neural network, 卷积神经网络)和 Transformer[34] 模型中以学习更鲁棒的视觉因果特征. 文献 [35] 提出一种新的因果干预训练方法, 通过去除“坏”的上下文信息而保留“好”的以训练更好的图像分类器. 文献 [36] 针对后门准则需要明确识别混杂因素的缺点, 提出了无混杂识别的因果视觉特征学习 CICF(Confounder identification-free causal visual feature learning, CICF), 基于前门准则对不同样本之间的干预进行建模, 然后从优化的角度基于实例级别的干预来近似全局范围的干预效果. 文献 [37] 提出了一种利用图像−文本匹配偏差进行多模态假新闻检测的因果推理框架, 可以应用于任何以视觉和文本特征作为输入的假新闻检测模型. 与这些工作类似, 本文将因果推断中的干预机制引入图文检索中, 充分利用图像区域的上下文信息提取常识因果视觉特征, 并与原始特征连接, 将图像局部信息与全局信息结合, 提高检索结果的准确度.

2 嵌入共识知识的因果图文检索方法

嵌入共识知识的因果图文检索方法由实例级图文特征提取模块、嵌入共识知识模块、共识级图文特征表示与融合模块、图文特征匹配和损失计算四大模块构成, 整体框架如图 1 所示. 在实例级图文特征提取模块中, 图像特征提取部分采用通过因果干预提取的常识因果特征与 BUTD 模型提取的原始图像特征连接的方法, 文本特征提取模块使用 BERT 模型; 在嵌入共识知识模块, 从图文检索数据集的注释文本中获取语料库, 在语料库中筛选得到共识词后, 通过图卷积网络学习共识词之间的语义关联, 生成共识知识概念表示; 在共识级图文特征表示与融合模块, 共识级图文特征表示由共识知识概念表示和实例级图文特征表示通过 Transformer 机制生成, 然后将实例级图文特征表示与共识级图文特征表示进行融合得到融合图文特征表示; 在图文特征匹配和损失计算模块, 图文匹配值利用余弦函数对图文特征间的相似性进行计算, 并采用引进难负样本的三元组损失函数对模型进行优化.

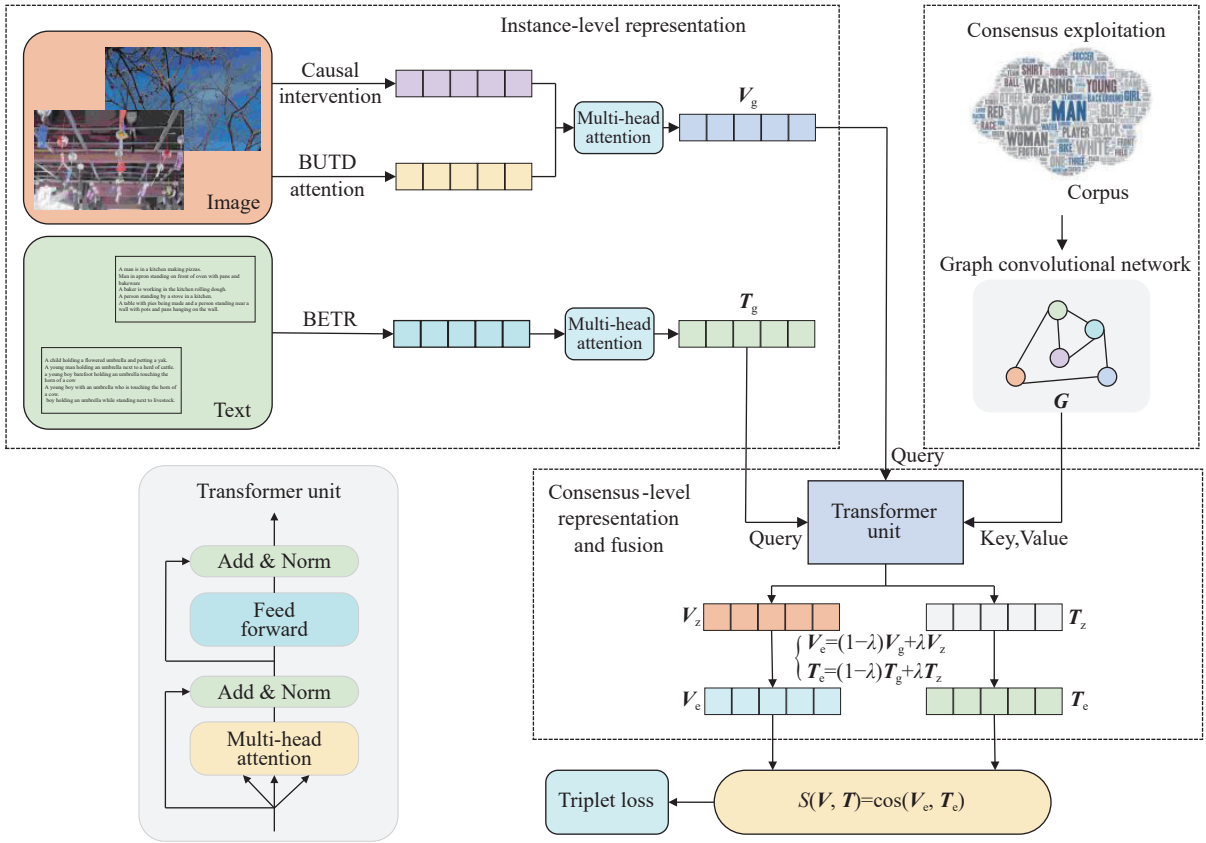
2.1 引入因果推断的图像特征提取

为了避免由于观察偏差导致模型基于一些共现信息做预测, 而忽略了常识性的因果关系, 本文通过因果干预提取蕴含常识的因果视觉特征, 并与原始图像特征进行融合. 基于共现信息学习到

的原始相关性特征可以告诉我们“是什么”、“在哪里”, 而常识因果特征则可以告诉我们“为什么”, 能够更好地利用图像中的上下文信息对图像特征进行表示学习.

常识因果视觉特征利用引入因果干预的 Faster R-CNN[38] 学习, 通过使用因果干预 $P(Y|\mathrm{do}(X))$ 替代传统的 $P(Y|X)$ 预测 RoI(Region of interest)区域的上下文对象作为代理任务, 使常识因果视觉特征提取模块能够学习到“椅子可以被坐”这样的常识性知识而不是仅仅学习到传统的物体共现现象“椅子与桌子一同出现”, 常识因果视觉特征提取模块如图 2 所示.

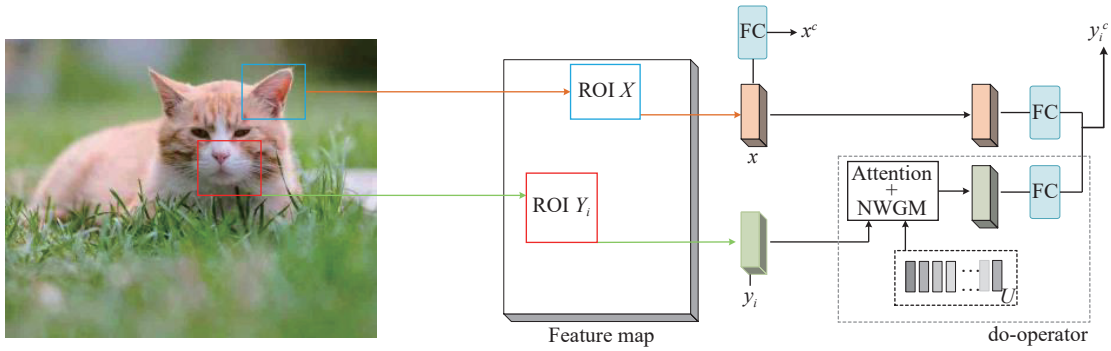
具体来说, 给定一张图像 $\boldsymbol{V}$, 常识因果视觉特征提取模块是以给定 RoI $\boldsymbol{X}$ 的类别 x^c 去预测 RoI $\boldsymbol{Y}_i$ 的类别 y_i^c 为代理任务进行训练. 首先图像被送入到以 ResNet-101[39] 为骨干的 Faster R-CNN 生成特



Notes: $\boldsymbol{V}_g$, $\boldsymbol{T}_g$ is instance-level visual and textual representations; $\boldsymbol{V}_z$, $\boldsymbol{T}_z$ is consensus-level visual and textual representations; $\boldsymbol{V}_e$, $\boldsymbol{T}_e$ is fused visual and textual representations; $\boldsymbol{G}$ is consensus-aware concept representation.

图 1　嵌入共识知识的因果图文检索整体框架

Fig.1　Framework of causal image-text retrieval embedded with consensus knowledge



Notes: FC is Fully Connected Layer; $\boldsymbol{x}$ and $\boldsymbol{y}_i$ is the features of ROI $\boldsymbol{X}$ and ROI $\boldsymbol{Y}_i$, x_c and y_i^c is the class label of ROI $\boldsymbol{X}$ and ROI $\boldsymbol{Y}_i$, and c represents its category; NWGM is normalized weighted geometric mean; $\boldsymbol{U}$ is confounder dictionary.

图 2　利用因果干预提取常识因果视觉特征

Fig.2　Using causal intervention to extract causal visual features with common sense

征图. 与 Faster R-CNN 不同, 去掉了区域提案网络, 直接利用 ground-truth 边界框通过 RoI Align[40] 层提取目标级表示. 然后每两个 RoI 特征 $\boldsymbol{X}$ 和 $\boldsymbol{Y}_i$ 被送入到两个平行的子分支: 自预测器和上下文预测器. 假设 $\boldsymbol{X}$ 作为主要关注的目标, $\boldsymbol{Y}=\{\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_i,\cdots,\boldsymbol{Y}_K\}$ 为 K 个上下文目标, 自预测器后接一个全连接层预测 $\boldsymbol{X}$ 的类标签 x^c, 上下文预测器使用 do 运算预测每个上下文目标的类标签 y_i^c. 模型最终的训练损失分为自预测器的损失和上下文预测器的损失, 自预测器的损失定义为:

$$L_{\text{self}}(\boldsymbol{p},x^c)=-\log(p[x^c]) \tag{1}$$

其中, x^c 是 RoI $\boldsymbol{X}$ 的 ground-truth 类标签, $\boldsymbol{p}=(p[1],p[2],\cdots,p[N])$ 为 $\boldsymbol{X}$ 在 N 个类别上的离散概率分布. 上下文预测的损失定义为:

$$L_{\text{cxt}}(\boldsymbol{p}_i,y_i^c)=-\log(p_i[y_i^c]) \tag{2}$$

其中 $\boldsymbol{p}_i=P(\boldsymbol{Y}_i|\text{do}(\boldsymbol{X}))$, y_i^c 为第 i 个上下文目标的 ground-truth 类标签.

因此 RoI $\boldsymbol{X}$ 的总的多任务损失为:

$$L(\boldsymbol{X})=L_{\text{self}}(\boldsymbol{p},x^c)+\frac{1}{K}\sum_{i=1}^{K}L_{\text{cxt}}(\boldsymbol{p}_i,y_i^c) \tag{3}$$

在利用 do 运算计算 $\boldsymbol{p}_i$ 时, 由于难以对现实世界的混杂因素进行收集统计, 因此在实际中将其近似为一个固定的混杂因素字典 $\boldsymbol{U}=[\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_N]\in\mathbb{R}^{N\times d}$, 其中 N 是数据集中的图像类别数, d 是 RoI 区域的特征维度, $\boldsymbol{u}_j$ 是数据集中第 j 个类别样本的平均 RoI 特征. 给定 RoI $\boldsymbol{X}$ 的特征 $\boldsymbol{x}$ 和其上下文 RoI $\boldsymbol{Y}_i$ 的类标签 y_i^c, 因果干预公式可以表示为:

$$P(\boldsymbol{Y}_i|\text{do}(\boldsymbol{X}))=\sum_{j=1}^{N}P(\boldsymbol{Y}_i|\boldsymbol{X},\boldsymbol{u}_j)P(\boldsymbol{u}_j)=\sum_{j=1}^{N}P(y_i^c|\boldsymbol{x},\boldsymbol{u}_j)P(\boldsymbol{u}_j) \tag{4}$$

由于标签预测网络的最后一层是 SoftMax 层, 所以:

$$P(y_i^c|\boldsymbol{x},\boldsymbol{u})=\text{Soft}\max(f_y(\boldsymbol{x},\boldsymbol{u})) \tag{5}$$

其中 $f_y(\cdot)$ 为分类器, 计算 N 个类别的 logits. 因此最终的干预公式可以表示为:

$$P(\boldsymbol{Y}_i|\text{do}(\boldsymbol{X}))\coloneqq E_{\boldsymbol{u}}[\text{Soft}\max(f_y(\boldsymbol{x},\boldsymbol{u}))] \tag{6}$$

由于 $E_{\boldsymbol{u}}$ 需要昂贵的抽样, 因此应用 NWGM[41] 近似上述期望:

$$E_{\boldsymbol{u}}[\text{Soft}\max(f_y(\boldsymbol{x},\boldsymbol{u}))]\overset{\text{NWGM}}{\approx}\text{Soft}\max(E_{\boldsymbol{u}}[f_y(\boldsymbol{x},\boldsymbol{u})]) \tag{7}$$

对 $\boldsymbol{Y}_i$ 的影响同时来自于 $\boldsymbol{x}$ 和混杂因素 $\boldsymbol{U}$, 最后的全连接层在使用线性模型时 $f_y(\boldsymbol{x},\boldsymbol{u})=\boldsymbol{W}_1\boldsymbol{x}+\boldsymbol{W}_2\cdot g_y(\boldsymbol{u})$, 其中 $\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 代表全连接层权重矩阵, $\boldsymbol{W}_1,\boldsymbol{W}_2\in\mathbb{R}^{N\times d}$, $E_{\boldsymbol{u}}[f_y(\boldsymbol{x},\boldsymbol{u})]$ 可以表示为:

$$E_{\boldsymbol{u}}[f_y(\boldsymbol{x},\boldsymbol{u})]=\boldsymbol{W}_1\boldsymbol{x}+\boldsymbol{W}_2\cdot E_{\boldsymbol{u}}[g_y(\boldsymbol{u})] \tag{8}$$

$E_{\boldsymbol{u}}[g_y(\boldsymbol{u})]$ 计算如下:

$$E_{\boldsymbol{u}}[g_y(\boldsymbol{u})]=\sum_{j=1}^{N}\left[\text{Soft}\max\left(\frac{\boldsymbol{q}^{\text{T}}\boldsymbol{K}}{\sqrt{\sigma}}\right)\odot\boldsymbol{U}\right]P(\boldsymbol{u}_j) \tag{9}$$

其中, $\boldsymbol{q}=\boldsymbol{W}_3\boldsymbol{y}_i$, $\boldsymbol{K}=\boldsymbol{W}_4\boldsymbol{U}^{\text{T}}$, $P(\boldsymbol{u}_j)$ 是先验概率, 通常假设 $P(\boldsymbol{u}_j)=1/N$, $\boldsymbol{y}_i$ 是 RoI $\boldsymbol{Y}_i$ 的特征, 并且 $\odot$ 是元素级乘, σ 是 $\boldsymbol{W}_3$, $\boldsymbol{W}_4$ 的第一个维度作为一个比例因子.

2.2 共识知识嵌入

为解决外部知识与图文检索任务数据集之间的域差异, 嵌入共识知识的语料库来自图文检索数据集的文本描述, 通常一张图像对应着 5 句文本描述, 这是非常庞大的语料库, 因此本文选择了在语料库中出现频率为 Top-q 的单词作为共识词, 并将其分为实体、属性、动作三类, 按照名词归类为实体、形容词归类为属性、动词归类为动作的原则进行划分, 之后采用 Glove 模型[42] 进行共识词嵌入并将其表示为 $\boldsymbol{Z}$.

由于共识词间的关系可以通过它们的共现频率来衡量, 例如电脑和鼠标经常一起出现, 这就可以看作是一种简单的共识. 即如果共识词 $\boldsymbol{z}_i$ 和 $\boldsymbol{z}_j$ 同时出现则将他们视作一个共现对. 根据共现频率构建出共识词之间的共现矩阵 $\boldsymbol{R}$, 其中 $\boldsymbol{R}_{ij}$ 表示 $\boldsymbol{z}_i$ 和 $\boldsymbol{z}_j$ 的共现次数. 为了更好的地利用共识知识, 可以根据共现矩阵 $\boldsymbol{R}$ 构建共识词之间的相关矩阵 $\boldsymbol{E}$, 用来捕获各共识词之间的内在相关性(不仅仅是共现关系):

$$\boldsymbol{E}_{ij}=\frac{\boldsymbol{R}_{ij}}{N_i} \tag{10}$$

其中, N_i 是 $\boldsymbol{z}_i$ 在语料库中的出现次数. 由于该方法是通过统计图文检索数据集中共识词的共现关系得到, 可能会偏离真实场景的数据分布, 产生数据偏差而影响之后的泛化能力; 而且, 由于共识词间共现频率产生的统计模式很容易受到长尾分布的影响. 也就是说, 仅仅偶尔出现的共现现象不足以作为共识使用, 只有出现次数较多, 具有一定普遍性的共现关系才能作为共识, 因此设计一个尺度函数用来调整相关矩阵 $\boldsymbol{E}$:

$$\boldsymbol{P}_{ij}=f(\boldsymbol{E}_{ij})=s^{E_{ij}-a}-s^{-a} \tag{11}$$

其中, s 和 a 是两个超参数, 参数 s 可以放大或缩小 $\boldsymbol{E}$ 中的值, 有助于调整 $\boldsymbol{E}$ 更好地匹配实际数据分布, 增强其泛化能力; 参数 a 可以有助于抵消长尾

分布带来的潜在偏差, 进而构建一个更平衡和无偏的$\boldsymbol{E}$. 而且, 为防止相关矩阵$\boldsymbol{E}$过度拟合训练数据影响其泛化能力, 应用二进制操作调整矩阵$\boldsymbol{P}$:

$$\boldsymbol{O}_{ij}=\begin{cases}0, & \text{if } \boldsymbol{P}_{ij}\leqslant\eta\\ 1, & \text{if } \boldsymbol{P}_{ij}>\eta\end{cases} \tag{12}$$

其中, $\boldsymbol{O}$是二值化的$\boldsymbol{P}$矩阵, 可以看作共识词之间的邻接矩阵, 0 代表无边, 表示共识词之间不存在关系, 1 代表有边, 表示共识词之间存在关系. η表示一个阈值参数, 可以过滤一些较少出现的共现关系.

最后, 在共识词的特征表示方面, 由于图卷积网络(Graph convolutional networks, GCN)[43] 可以从图结构的数据上学习映射函数, 通过基于节点的邻域传播信息来更新节点的嵌入表示. 因此使用多个堆叠的 GCN 层来学习共识词的表示, 它在共识词之间引入更高阶的邻域信息来对共识词之间的相关关系进行建模. 具体来说, 给定实例化的共识词表示$\boldsymbol{Z}$以及$\boldsymbol{O}$, 第 l 层的嵌入特征计算为:

$$\boldsymbol{T}^{(l+1)}=\mathrm{ReLU}(\hat{\boldsymbol{A}}\boldsymbol{T}^{l}\boldsymbol{W}^{l}) \tag{13}$$

其中, $\boldsymbol{T}^{0}$表示图卷积网络第 0 层的输入, 即$\boldsymbol{T}^{0}=\boldsymbol{Z}$; $\hat{\boldsymbol{A}}$是归一化的对称矩阵, 即$\hat{\boldsymbol{A}}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{O}\boldsymbol{D}^{-\frac{1}{2}}$; $\boldsymbol{W}^{l}$是训练过程中需要学习的权重矩阵; ReLU为非线性激活函数. 取图卷积网络最后一层的输出, 得到最终的共识知识概念表示$\boldsymbol{G}=\{\boldsymbol{g}_1,\boldsymbol{g}_2,\cdots,\boldsymbol{g}_i,\cdots,\boldsymbol{g}_q\}\in\mathbb{R}^{q\times d}$, 其中$d$表示联合嵌入空间的维度, $\boldsymbol{g}_i$表示共识词$\boldsymbol{z}_i$的嵌入表示.

2.3　图文特征提取与融合

首先提取实例级图文特征表示, 再根据实例级图文特征表示与共识知识概念$\boldsymbol{G}$通过 Transformer 机制得到共识级图文特征表示, 最后将两者融合得到最终的融合图文特征表示.

(1)实例级图文特征表示.

通过 BUTD[12] 模型提取原始图像特征, 引入因果干预的图像特征提取模块提取图像常识因果特征, 之后再将两种特征连接, 可得到实例级的局部图像特征表示, 为获取到更具有鲁棒性的全局特征, 本文采用多头注意力机制[32], 将所提取的局部图像特征作为注意力机制的 Key 和 Value 项, 计算局部图像特征的平均值$\bar{\boldsymbol{v}}=\dfrac{1}{i}\sum_{k=1}^{i}\boldsymbol{v}_k$作为注意力机制的 Query 项, 最终计算得到实例级全局图像特征$\boldsymbol{V}_{\mathrm{g}}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_i\}$.

实例级文本特征的提取利用预训练的 BERT[13] 模型, 类似地, 将所提取的局部文本特征作为注意力机制的 Key 和 Value, 并且计算局部文本特征的平均值$\bar{\boldsymbol{t}}=\dfrac{1}{j}\sum_{k=1}^{j}\boldsymbol{t}_k$作为注意力机制的 Query 项, 最终计算得到实例级全局文本特征$\boldsymbol{T}_{\mathrm{g}}=\{\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_j\}$.

(2)共识级图文特征表示.

使用一个 Transformer 单元对实例级全局图像特征$\boldsymbol{V}_{\mathrm{g}}$和实例级全局文本特征$\boldsymbol{T}_{\mathrm{g}}$使用共识知识概念表示$\boldsymbol{G}$进行增强, 得到最终的共识级图文特征表示:

$$\begin{cases}\boldsymbol{V}_{\mathrm{z}}=\mathrm{FFN}(\mathrm{MultiHead}(\boldsymbol{V}_{\mathrm{g}},\boldsymbol{G}))\\ \boldsymbol{T}_{\mathrm{z}}=\mathrm{FFN}(\mathrm{MultiHead}(\boldsymbol{T}_{\mathrm{g}},\boldsymbol{G}))\end{cases} \tag{14}$$

其中, FFN(·)表示由两层感知器实现的前馈网络. 且

$$\mathrm{MultiHead}(\boldsymbol{X},\boldsymbol{Y})=\mathrm{Concat}(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_H)+\boldsymbol{X} \tag{15}$$

其中, $\boldsymbol{X}=\boldsymbol{V}_{\mathrm{g}}$或者$\boldsymbol{X}=\boldsymbol{T}_{\mathrm{g}}$, $\boldsymbol{Y}=\boldsymbol{G}$, Concat(·)表示特征维度的连接操作, H表示注意力机制头部的数量, $\boldsymbol{h}_i$使用点积注意力来计算:

$$\boldsymbol{h}_i=\mathrm{Attention}(\boldsymbol{X}\boldsymbol{W}_i^{\mathrm{Q}},\boldsymbol{Y}\boldsymbol{W}_i^{\mathrm{K}},\boldsymbol{Y}\boldsymbol{W}_i^{\mathrm{V}}) \tag{16}$$

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Soft\,max}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_j}}\right)\boldsymbol{V} \tag{17}$$

其中, $\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$分别表示注意力机制中的 Query、Key 和 Value, d_j是$\boldsymbol{Q}$和$\boldsymbol{K}$的通道数; $\boldsymbol{W}_i^{\mathrm{Q}}$, $\boldsymbol{W}_i^{\mathrm{K}}$和$\boldsymbol{W}_i^{\mathrm{V}}$均为需要学习的权重矩阵.

(3)图文特征融合模块.

根据实例级图文特征表示与共识级图文特征表示, 最终融合后的图文特征可表示为:

$$\begin{cases}\boldsymbol{V}_{\mathrm{e}}=(1-\lambda)\boldsymbol{V}_{\mathrm{g}}+\lambda\boldsymbol{V}_{\mathrm{z}}\\ \boldsymbol{T}_{\mathrm{e}}=(1-\lambda)\boldsymbol{T}_{\mathrm{g}}+\lambda\boldsymbol{T}_{\mathrm{z}}\end{cases} \tag{18}$$

其中, λ是实例级图文特征表示与共识级图文特征表示的融合超参数.

2.4　图文特征匹配与损失计算

得到融合后的图文特征表示后, 采用余弦函数进行匹配值的计算, 表达式如下:

$$S(\boldsymbol{V},\boldsymbol{T})=\cos(\boldsymbol{V}_{\mathrm{e}},\boldsymbol{T}_{\mathrm{e}}) \tag{19}$$

采用引入难负样本的三元组损失函数:

$$\begin{aligned}l_{\mathrm{h}}(\boldsymbol{V},\boldsymbol{T})=&[\beta-S(\boldsymbol{V},\boldsymbol{T})+S(\boldsymbol{V},\hat{\boldsymbol{T}}_{\mathrm{h}})]_{+}+\\ &[\beta-S(\boldsymbol{V},\boldsymbol{T})+S(\hat{\boldsymbol{V}}_{\mathrm{h}},\boldsymbol{T})]_{+}\end{aligned} \tag{20}$$

其中, β表示正样本对与负样本对需要保持的最小间隔; $\boldsymbol{V}$和$\boldsymbol{T}$分别表示匹配正确的图文对; $\hat{\boldsymbol{T}}_{\mathrm{h}}=\arg\max_{\boldsymbol{t}\neq\boldsymbol{T}}S(\boldsymbol{V},\boldsymbol{t})$表示在批数据样本中与图像$\boldsymbol{V}$匹配分值最高且并不匹配的文本; $\hat{\boldsymbol{V}}_{\mathrm{h}}=\arg\max_{\boldsymbol{v}\neq\boldsymbol{V}}S(\boldsymbol{T},\boldsymbol{v})$则表示在批数据样本中与文本$\boldsymbol{T}$匹配分值最高且并不匹配的图像; $[x]_{+}=\max(x,0)$.

3 实验结果分析

3.1 数据集与评价指标

数据集采用 MS-COCO[14], MS-COCO 是一个常用于图文检索的公开数据集, 它包含 123287 幅图像, 每张图像对应 5 句不同的描述文本. 为保证实验的公平性与可比较性, 对数据集的划分与其他图文检索算法一致, 即 113287 张训练图像, 5000 张验证图像及 5000 张测试图像. 同时, 对于 5000 张测试图像, 考虑两种评估机制: MS-COCO 1K, 在 5 个 1000 张测试图像上计算检索结果并平均. MS-COCO 5K, 在完整的 5000 张测试图像上计算最终检索结果;

评价指标使用标准的 $R@k(k=1, 5, 10)$ 在测试集上计算检索得分, 为提供更全面的评估, 我们还使用了 R_{sum} 和 mR.

3.2 实验环境和参数设置

所有的实验都是在一台搭载 RTX 3090 GPU 的计算机上使用 PyTorch 实现. 使用在 ImageNet 上预训练的 ResNet101 作为 Faster R-CNN 的骨干网络提取常识因果视觉特征, 维度为 1024 维. BUTD 提取的原始图像特征维度为 2048 维, 将两种图像特征连接得到 3072 维的全局图像特征表示, 然后经过一个全连接层最终得到 1024 维的图像特征表示.

文本特征提取使用拥有 12 层, 12 个头和 768 个隐藏单元的预训练 BETR 基础版本, 最终得到 768 维的词嵌入向量. 图文特征公共表示空间维度为 1024. 在计算图文全局特征时采用的多头注意力机制的头数为 8. 在嵌入共识模块, 采用在维基百科数据集训练的 300 维的 Glove 来表示初始语义概念, 选取的共识词大小为 300. 图卷积网络使用了两个图卷积层, 嵌入维度分别设为 512 和 1024, 在公式 (11) 和 (12) 中, $s=5$, $a=0.02$, $\eta=0.3$, 共识级图文特征表示模块中注意力头数 H 设为 1, λ 为 0.05.

在模型训练时, 使用 Adam 优化器, 在 MS-COCO 数据集进行训练时, batch size 为 64, 训练 epoch 为 30, 学习率为 0.0005, 每经过 15 个 epoch 学习率衰减为原来的 10%, 三元组损失函数中的 β 为 0.2, 并且为了避免出现过拟合现象, 全连接层都以 0.1 的概率随机丢弃一些特征.

3.3 主要实验结果

我们在 MS-COCO 数据集上进行了实验并与 CVSE[22]、SCO[44]、PVSE[45]、SGM[46]、VSE++[47]、SCAN[48]、RAAN[49]、SHAN[50]、DREN[51]、ALGRL[52]、NCR[53]、VSRN++[54]、ReSG[55]、SAGRL[56]、GLFN[57]、VSE∞[58]、DRCE[59] 和 CMRN[60] 在 MS-COCO 1K 和 5K 测试集上进行比较, 结果如表 1 和表 2 所示.

我们的方法同对比的图文检索方法相比取得了不错的结果, 比如在 1K 测试集上的 R_{sum} 为 523.4%, 相比于 CVSE 提升了 9.8%, 与最近的方法 RANN 和 GLFN 相比分别提升了 4.8% 和 2.9%. 在 5K 测试集上, 所提出的方法同样在评价指标上得到了提升, 与最近的方法 ALGRL 和 VSRN++相比, R_{sum} 分别提升了 1.1% 和 1.4%, 充分体现了本文所提方法的有效性.

3.4 消融研究

全局特征提取方式: 为获得更鲁棒的全局特征, 针对图像和文本全局特征的提取方式, 使用平均池化和注意力机制两种方式进行消融研究, 并使用 mR 作为评价指标, 结果如图 3(a) 所示. 其中 A 代表平均池化方式, S 代表注意力机制, A+S 即表示图像全局特征提取采用平均池化而文本全局特征提取采用注意力机制. 在嵌入外部共识的前提下, 图像和文本同时使用注意力机制提取全局特征能够得到最好的结果, 相比于同时使用平均池化来说, mR 指标能够提升 0.8%.

共识词数量 q: 我们对从语料库中选定的共识词数量 q 进行了消融实验, 分别选其为 0、100、200、300 和 400, 结果如图 3(b) 所示. 可见并不是一味地增加共识词数量就能够提高模型性能, 当 q 选定为 300 时 mR 指标达到了最大, 增加到 400 时反而会导致性能的下降, 这可能是由于 q 为 300 时已经覆盖了语料库中的大部分词汇, 当选定为 400 时反而会引入不必要且没有意义的词汇.

融合参数 λ: 本次消融实验中分别选取 0、0.0025、0.05、0.075 和 0.1 进行对比分析, mR 指标随 λ 的变化结果趋势如图 3(c) 所示. 可以明显看到 λ 的选取会极大影响模型的性能, 当 λ 选取为 0.1 时, mR 指标会急剧下降, 这可能是由于共识级图文表示占比多了反而忽略检索目标本身的信息导致无法得到正确的检索结果.

3.5 泛化性分析

MS-COCO[[14] 和 Flickr30k[15]] 都是相对大规模的图文数据集, 在图像内容分布、文本语言风格和标注质量等方面存在一定的差异, 跨数据集的泛化问题需要考虑如何处理这些差异, 模型需要学会泛化到新数据集的能力并且不会产生灾难性遗忘, 以确保模型在不同的数据集上都能够有效工作. 因此跨数据集泛化问题在图文检索任务中非常具有挑战性, 需要更多的创新方法和技术来解

表 1 MS-COCO 1K 测试集评估结果

Table 1 Evaluation results on the test set MS-COCO 1K

Methods	Image Backbone	Text Backbone	I2T Retrieval			T2I Retrieval				
			R@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10	R_{sum}	mR
VSE++	ResNet-152	GRU	64.6	90.0	95.7	52.0	84.3	92.0	478.6	79.77
SCAN	Faster R-CNN	Bi-GRU	72.7	94.8	98.4	58.8	88.4	94.8	507.9	84.65
PVSE	ResNet-152	Bi-GRU	69.2	91.6	96.6	55.2	86.5	93.7	492.8	82.13
SCO	ResNet-152	LSTM	71.3	93.8	98.0	58.2	88.8	95.3	505.4	84.23
SGM	Faster R-CNN	Bi-GRU	73.4	93.8	97.8	57.5	87.3	94.3	504.1	84.02
CVSE	Faster R-CNN	Bi-GRU	74.8	95.1	98.3	59.9	89.4	95.2	513.6	85.60
NCR	Faster R-CNN	Bi-GRU	78.7	95.8	98.5	63.3	90.4	95.8	522.5	87.08
SHAN	Faster R-CNN	Bi-GRU	76.8	96.4	98.7	62.6	89.6	95.8	519.8	86.63
VSE∞	Faster R-CNN	Bi-GRU	78.5	96.0	98.7	61.7	90.3	95.6	520.8	86.80
SAGRL	Faster R-CNN	Bi-GRU	75.5	95.9	99.0	59.8	89.1	95.0	514.3	85.72
CMRN	Faster R-CNN	GRU	73.9	93.9	97.9	60.4	88.5	94.0	508.6	84.77
DERN	Faster R-CNN	Bi-GRU	78.5	96.2	99.0	62.0	89.7	96.2	521.6	86.93
ALGRL	Faster R-CNN	BETR	77.8	96.1	99.0	63.9	91.1	96.0	523.9	87.32
RAAN	Faster R-CNN	Bi-GRU	76.8	96.4	98.3	61.8	89.5	95.8	518.6	86.43
GLFN	Faster R-CNN	BETR+ Bi-GRU	78.4	96.0	98.5	62.6	89.6	95.4	520.5	86.75
DRCE	Faster R-CNN	Bi-GRU	79.1	96.4	99.0	63.6	90.3	95.9	524.3	87.38
VSRN++	Faster R-CNN	BETR	77.9	96.0	98.5	64.1	91.0	96.1	523.6	87.27
Our*	Faster R-CNN	Bi-GRU	76.3	95.7	98.6	59.6	88.8	94.9	513.9	85.65
Our	Faster R-CNN	BETR	78.6	96.4	98.9	62.8	90.4	96.3	523.4	87.23

Note: "*" indicates only use causal intervention.

表 2 MS-COCO 5K 测试集评估结果

Table 2 Evaluation results on the test set MS-COCO 5K

Methods	Image Backbone	Text Backbone	I2T Retrieval			T2I Retrieval				
			R@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10	R_{sum}	mR
VSE++	ResNet-152	GRU	41.3	71.1	81.2	30.3	59.4	72.4	355.7	59.28
SCAN	Faster R-CNN	Bi-GRU	50.4	82.2	90.0	38.6	69.3	80.4	410.9	68.48
PVSE	ResNet-152	Bi-GRU	45.2	74.3	84.5	32.4	63.0	75.0	374.4	62.40
SCO	ResNet-152	LSTM	45.7	76.0	86.4	36.8	67.0	78.8	390.7	65.12
SGM	Faster R-CNN	Bi-GRU	50.0	79.3	87.9	35.3	64.9	76.5	393.9	65.65
CVSE	Faster R-CNN	Bi-GRU	51.1	80.1	89.3	37.0	68.0	79.7	405.2	67.53
SAGRL	Faster R-CNN	Bi-GRU	51.7	82.9	90.4	39.6	69.9	81.1	415.6	69.27
ReSG	Faster R-CNN	Bi-GRU	55.8	83.0	91.0	42.0	72.4	82.1	426.3	71.05
ALGRL	Faster R-CNN	BETR	55.2	83.9	91.4	40.7	71.9	82.6	425.7	70.98
DRCE-A	Faster R-CNN	Bi-GRU	56.2	83.0	90.9	40.3	69.5	80.6	420.5	70.08
VSRN++	Faster R-CNN	BETR	54.7	82.9	90.9	42.0	72.2	82.7	425.4	70.90
Our*	Faster R-CNN	Bi-GRU	53.2	83.0	90.3	39.2	69.5	80.5	415.7	69.28
Our	Faster R-CNN	BETR	55.3	84.3	91.7	42.4	71.7	81.4	426.8	71.13

Note: "*" indicates only use causal intervention.

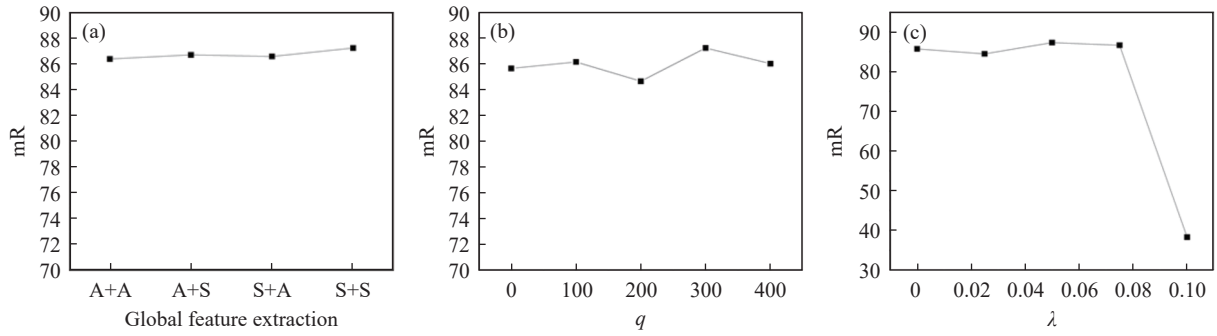


图 3 消融实验. (a) mR 受全局特征提取方式的影响; (b) mR 受共识词数量 q 的影响; (c) mR 受融合参数 λ 的影响

Fig.3 Ablation experiments: (a) mR is affected by the manner of global feature extraction; (b) mR is affected by the number of consensus words q; (c) mR is affected by the fusion parameter λ

决这些问题, 比如需要更强大的特征提取和表示学习方法. 跨数据集泛化问题本身就是一个具有挑战性的问题, 仅有较少的工作研究了图文检索任务的跨数据集泛化. 最近的一些工作将因果推断方法引入领域自适应和分布外泛化[28], 通过因果干预学习目标在不同上下文的不变特征, 这与本文所采用的方法类似, 本文为了进一步验证引入因果干预与嵌入共识模块是否能够有效推理图文模态数据间的语义关联, 设计了跨数据集实验来测试本文方法的泛化能力. 具体来说, 通过将在 MS-COCO 数据集训练好的模型直接迁移到 Flickr30k 数据集进行实验得到其评价指标, 最终实验结果如表 3 所示.

由表 3 可以看出, 在跨数据集实验中, 本文方法的性能超过了对比的方法, 在双向检索任务中均获得了突出的表现, 图像检索文本任务的 R@1 达到了 62.3%, 文本检索图像任务的 R@1 达到了 42.2%, 并且总体指标 R_{sum} 相比于 SCAN 和 CVSE 分别提升了 40.7% 和 18%. 这些结果充分表明了结合因果推断与外部共识后能够显著提高模型的学习能力, 并且学习到的知识可以在跨域异构数据之间共享, 从而显著提高模型的泛化能力.

3.6 信息可视化

图 4 展示了图像检索文本的部分结果示例, "*" 表示仅使用因果干预, 红色文本代表错误的召回文本. 可见在嵌入共识后, 相比于仅使用因果推断的方法来说, 引入外部共识知识后召回率得到了一定提升, 并且检索得分最高的文本能够更全面地描述图像信息, 获得更准确的检索结果.

图 5 展示了部分文本检索图像的结果示例, 检索结果中红色框和绿色框分别标记错误和正确的召回对象, 第一行代表仅使用因果推断. 对于检索文本 "A group of people on skis stand in a line.", 嵌入共识知识后的方法不仅能够检索到正确图像, 并且该图像位于检索得到图像中的首位, 说明引入图像和文本的共识能更好地推理图文模态数据间的语义关联, 从检索得到的图像内容也可以看出, 嵌入共识后的模型更容易理解 "group" 这个词汇, 而仅使用因果推断的方法得到的结果关注重点可能更多的在于 "people" 和 "skis", 无法理解 "group", 这进一步证实了本文方法的有效性.

4 结论与展望

相关并非因果, 针对目前传统的基于深度学习的跨模态图文检索方法无法建模数据背后的因果关系的问题, 本文将因果干预引入传统的图文检索方法, 通过学习图像背后的因果关系加强图文数据间的逻辑关联. 同时嵌入外部共识知识以

表 3 MS-COCO 与 Flickr30k 跨数据集泛化评估结果

Table 3 Evaluation results on crossdataset generalization from MS-COCO to Flickr30k

Methods	I2T Retrieval			T2I Retrieval				
	R@1	R@5	R@10	R@1	R@5	R@10	R_{sum}	mR
VSE++	40.5	67.3	77.7	28.4	55.4	66.6	335.9	55.98
SCAN	49.8	77.8	86.0	38.4	65.0	74.4	391.4	65.23
CVSE	56.4	83.0	89.0	39.9	68.6	77.2	414.1	69.02
Our	62.3	85.1	91.2	42.2	71.3	80.0	432.1	72.02

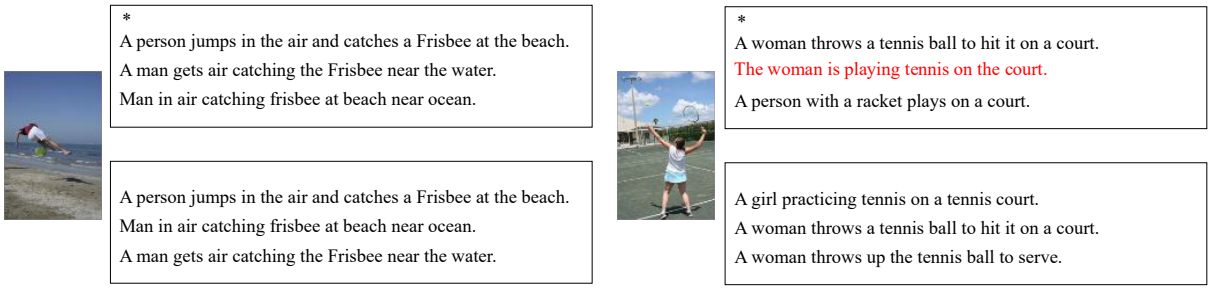


图 4　图像检索文本可视化

Fig.4　Visualization of I2T retrieval

图 5　文本检索图像可视化

Fig.5　Visualization of T2I retrieval

加强图像和文本间的共识, 同时考虑模态内关联和模态间关联, 最终提升模型的鲁棒性与泛化性. 在 MS-COCO 数据集以及 Flickr30k 数据集上的跨域进行实验, 证明了本文方法在双向图文检索方面的有效性. 本文仅在视觉特征提取模块引入了因果干预, 而没有考虑文本单词之间的因果关系, 因此在未来的工作中将会同时考虑两种模态背后的因果关系, 并考虑如何更加精确地描述混杂因素. 此外还可以在模型轻量化和融合知识图谱等方向进行探索.

参　考　文　献

[1] Han K, Wang Y H, Chen H T, et al. A survey on vision transformer. *IEEE Trans Pattern Anal Mach Intell*, 2023, 45(1): 87

[2] Otter D W, Medina J R, Kalita J K. A survey of the usages of deep learning for natural language processing. *IEEE Trans Neural Netw Learn Syst*, 2021, 32(2): 604

[3] Ma Z G, Ni R Y, Yu K H. Recent advances, key techniques and future challenges of knowledge graph. *Chin J Eng*, 2020, 42(10): 1254
(马忠贵, 倪润宇, 余开航. 知识图谱的最新进展、关键技术和挑战. 工程科学学报, 2020, 42(10): 1254)

[4] Kaur P, Pannu H S, Malhi A K. Comparative analysis on cross-modal information retrieval: A review. *Comput Sci Rev*, 2021, 39: 100336

[5] Li S, Lv F Y, Jin T W, et al. Embedding-based product retrieval in Taobao search // *The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Singapore, 2021: 3181

[6] Hu B, Vasu B, Hoogs A. X-MIR: Explainable medical image retrieval // 2022 *IEEE/CVF Winter Conference on Applications of Computer Vision*. Waikoloa, 2022: 1544

[7] Chen H, Du K Q, Yang X Y, et al. A review and roadmap of deep learning causal discovery in different variable paradigms [J/OL]. *arXiv preprint* (2022-09-14) [2023-05-28]. https://arxiv.org/abs/2209.06367

[8] Rudin C, Chen C F, Chen Z, et al. Interpretable machine learning: Fundamental principles and 10 grand challenges [J/OL]. *arXiv preprint* (2021-07-10) [2023-05-28]. https://arxiv.org/abs/2103.11251

[9] Räuker T, Ho A, Casper S, et al. Toward transparent AI: A survey on interpreting the inner structures of deep neural networks [J/OL]. *arXiv preprint* (2023-01-27) [2023-05-28]. https://arxiv.org/abs/2207.13243

[10] Ma Z G, Xu X H, Liu X E. Three analytical frameworks of causal inference and their applications. *Chin J Eng*, 2022, 44(7): 1231
(马忠贵, 徐晓晗, 刘雪儿. 因果推断三种分析框架及其应用综述. 工程科学学报, 2022, 44(7): 1231)

[11] Judea P. *Causality: Models, Reasoning, and Inference*. 2nd Ed. Beijing: China Machine Press, 2022
(朱迪亚·珀尔. 因果论 模型、推理和推断. 2 版. 北京: 机械工业出版社, 2022)

[12] Anderson P, He X D, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering //

2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, 2018: 6077

[13] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding // *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, 2019: 4171

[14] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context // *European Conference on Computer Vision*. Cham, 2014: 740

[15] Plummer B A, Wang L W, Cervantes C M, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models // 2015 *IEEE International Conference on Computer Vision*. Santiago, 2015: 2641

[16] Matsubara T. Target-oriented deformation of visual-semantic embedding space. *IEICE Trans Inf Syst*, 2021, E104(1): 24

[17] Wu J, Wu C, Lu J, et al. Region reinforcement network with topic constraint for image-text matching. *IEEE Trans Circuits Syst Video Technol*, 2021, 32(1): 388

[18] Zhang K, Mao Z D, Liu A N, et al. Unified adaptive relevance distinguishable attention network for image-text matching. *IEEE Trans Multimed*, 2023, 25: 1320

[19] Ji Z, Wang H R, Han J G, et al. SMAN: Stacked multimodal attention network for cross-modal image-text retrieval. *IEEE Trans Cybern*, 2022, 52(2): 1086

[20] Dong X F, Zhang H X, Zhu L, et al. Hierarchical feature aggregation based on transformer for image-text matching. *IEEE Trans Circuits Syst Video Technol*, 2022, 32(9): 6437

[21] Feng D D, He X T, Peng Y X. MKVSE: Multimodal knowledge enhanced visual-semantic embedding for image-text retrieval. *ACM Trans Multimedia Comput Commun Appl*, 2023, 19(5): 1

[22] Wang H R, Zhang Y, Ji Z, et al. Consensus-aware visual-semantic embedding for image-text matching // *Proceedings of the European Conference on Computer Vision*. Cham, 2020: 18

[23] Imbens G W, Wooldridge J M. Recent developments in the econometrics of program evaluation. *J Econ Lit*, 2009, 47(1): 5

[24] Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: Methods, interpretation and bias. *Int J Epidemiol*, 2013, 42(5): 1511

[25] Huang W Q, Jiang M, Li M, et al. Causal intervention for object detection // 2021 *IEEE* 33*rd International Conference on Tools with Artificial Intelligence* (*ICTAI*). Washington, 2021: 770

[26] Niu Y L, Tang K H, Zhang H W, et al. Counterfactual VQA: A cause-effect look at language bias // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville, 2021: 12695

[27] Yang X, Zhang H W, Cai J F. Deconfounded image captioning: A causal retrospect. *IEEE Trans Pattern Anal Mach Intell*, 2022: 1

[28] Wang W J, Lin X Y, Feng F L, et al. Causal representation learning for out-of-Distribution recommendation // *Proceedings of the ACM Web Conference* 2022. Lyon, 2022: 3562

[29] Wang T, Huang J Q, Zhang H W, et al. Visual commonsense R-CNN // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, 2020: 10757

[30] Zhang H, Xiao L Q, Cao X C, et al. Multiple adverse weather conditions adaptation for object detection *via* causal intervention. *IEEE Trans Pattern Anal Mach Intell*, 2022, PP(99): 1

[31] Ma Z G, Li Z, Liang Y P. Overview and prospect of communication-sensing-computing integration for autonomous driving in the Internet of vehicles. *Chin J Eng*, 2023, 45(1): 137 (马忠贵, 李卓, 梁彦鹏. 自动驾驶车联网中通感算融合研究综述与展望. 工程科学学报, 2023, 45(1): 137)

[32] Liu R Y, Liu H, Li G, et al. Contextual debiasing for visual recognition with causal mechanisms // 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, 2022: 12745

[33] Wang T, Zhou C, Sun Q R, et al. Causal attention for unbiased visual recognition // 2021 *IEEE/CVF International Conference on Computer Vision*. Montreal, 2021: 3071

[34] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need // 31*st Conference on Neural Information Processing Systems*. Long Beach, 2017

[35] Qin W, Zhang H W, Hong R C, et al. Causal interventional training for image recognition. *IEEE Trans Multimed*, 2023, 25: 1033

[36] Li X, Zhang Z Z, Wei G Q, et al. Confounder identification-free causal visual feature learning [J/OL]. *arXiv preprint* (2022-10-09) [2023-05-28].https://arxiv.org/abs/2111.13420

[37] Hu L M, Chen Z W, Zhao Z W, et al. Causal inference for leveraging image-text matching bias in multi-modal fake news detection. *IEEE Trans Knowl Data Eng*, 2023, 35(11): 11141

[38] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell*, 2017, 39(6): 1137

[39] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition // 2016 *IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, 2016: 770

[40] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN // 2017 *IEEE International Conference on Computer Vision*. Venice, 2017: 2980

[41] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention // *Proceedings of the* 32*nd International Conference on International Conference on Machine Learning*. Lille, 2015: 2048

[42] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation // *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing*. Doha, 2014: 1532

[43] Sanchez B, Lengeling, Reif E, et al. A gentle introduction to graph neural networks [J/OL]. *Distill* (2021-09-02) [2023-05-28]. https://distill.pub/2021/gnn-intro

[44] Huang Y, Wu Q, Wang W, et al. Image and sentence matching *via* semantic concepts and order learning. *IEEE Trans Pattern Anal*

Mach Intell, 2020, 42(3): 636

[45] Song Y L, Soleymani M. Polysemous visual-semantic embedding for cross-modal retrieval // 2019 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, 2019: 1979

[46] Wang S J, Wang R P, Yao Z W, et al. Cross-modal scene graph matching for relationship-aware image-text retrieval // *Proceedings of the* 2020 *IEEE/CVF Winter Conference on Applications of Computer Vision*. Snowmass, 2020: 1497

[47] Faghri F, Fleet D J, Kiros J, et al. VSE++: Improving visual-semantic embeddings with hard negatives // *Proceedings of the British Machine Vision Conference*. Newcastle, 2018: 1

[48] Lee K H, Chen X, Hua G, et al. Stacked cross attention for image-text matching // *Proceedings of the European Conference on Computer Vision*. Cham, 2018: 212

[49] Wang Y, Su Y T, Li W H, et al. Rare-aware attention network for image–text matching. *Inf Process Manag*, 2023, 60(3): 103280

[50] Ji Z, Chen K X, Wang H R. Step-wise hierarchical alignment network for image-text matching // 30*th International Joint Conference on Artificial Intelligence* (*IJCAI*-2021). Montreal, 2021: 765

[51] Yang S, Li Q, Li W H, et al. Dual-level representation enhancement on characteristic and context for image-text retrieval. *IEEE Trans Circuits Syst Video Technol*, 2022, 32(11): 8037

[52] Tian M X, Wu X X, Jia Y D. Adaptive latent graph representation learning for image-text matching. *IEEE Trans Image Process*, 2023, 32: 471

[53] Huang Z Y, Niu G H, Liu X, et al. Learning with noisy correspondence for cross-modal matching // 35*th Conference on Neural Information Processing Systems* (*NeurIPS* 2021). Piscataway, 2021: 29406

[54] Li K P, Zhang Y L, Li K, et al. Image-text embedding learning via visual and textual semantic reasoning. *IEEE Trans Pattern Anal Mach Intell*, 2023, 45(1): 641

[55] Liu X, He Y, Cheung Y M, et al. Learning relationship-enhanced semantic graph for fine-grained image-text matching. *IEEE Trans Cybern*, 2022, PP(99): 1

[56] Qi X F, Zhang Y, Qi J Q, et al. Self-attention guided representation learning for image-text matching. *Neurocomputing*, 2021, 450: 143

[57] Zhao G S, Zhang C F, Shang H, et al. Generative label fused network for image-text matching. *Knowl Based Syst*, 2023, 263: 110280

[58] Chen J C, Hu H X, Wu H, et al. Learning the best pooling strategy for visual semantic embedding // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville, 2021: 15748

[59] Wang Y, Su Y T, Li W H, et al. Dual-path rare content enhancement network for image and text matching. *IEEE Trans Circuits Syst Video Technol*, 2023, 33(10): 6144

[60] Zhang J, He X H, Qing L B, et al. Cross-modal multi-relationship aware reasoning for image-text matching. *Multimed Tools Appl*, 2022, 81(9): 12005