软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(11):4973−4988 [doi: 10.13328/j.cnki.jos.006742]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

融合多重实例关系的无监督跨模态哈希检索*

李志欣, 侯传文, 谢秀敏

(广西多源信息挖掘与安全重点实验室 (广西师范大学), 广西 桂林 541004)

通信作者: 李志欣, E-mail: lizx@gxnu.edu.cn

摘　要: 大多数跨模态哈希检索方法仅使用余弦相似度进行特征匹配, 计算方式过于单一, 没有考虑到实例的关系对于性能的影响. 为此, 提出一种基于多重实例关系图推理的方法, 通过构造相似度矩阵, 建立全局和局部的实例关系图, 充分挖掘实例之间的细粒度关系. 在多重实例关系图的基础上进行相似度推理, 首先分别进行图像模态和文本模态关系图内部的推理, 然后将模态内的关系映射到实例图中进行推理, 最后执行实例图内部的推理. 此外, 为了适应图像和文本两种模态的特点, 使用分步训练策略训练神经网络. 在 MIRFlickr 和 NUS-WIDE 数据集上实验表明, 提出的方法在 *mAP* 指标上具有很明显的优势, 在 Top-k-Precision 曲线上也获得良好的效果. 这也说明所提方法对实例关系进行深入挖掘, 从而显著地提升检索性能.

关键词: 关系图推理; 跨模态哈希检索; 相似度矩阵; K 近邻; 分步训练策略

中图法分类号: TP301

中文引用格式: 李志欣, 侯传文, 谢秀敏. 融合多重实例关系的无监督跨模态哈希检索. 软件学报, 2023, 34(11): 4973–4988. http://www.jos.org.cn/1000-9825/6742.htm

英文引用格式: Li ZX, Hou CW, Xie XM. Unsupervised Cross-modal Hash Retrieval Fusing Multiple Instance Relations. Ruan Jian Xue Bao/Journal of Software, 2023, 34(11): 4973–4988 (in Chinese). http://www.jos.org.cn/1000-9825/6742.htm

Unsupervised Cross-modal Hash Retrieval Fusing Multiple Instance Relations

LI Zhi-Xin, HOU Chuan-Wen, XIE Xiu-Min

(Guangxi Key Lab of Multi-source Information Mining and Security (Guangxi Normal University), Guilin 541004, China)

Abstract: Most cross-modal hash retrieval methods only use cosine similarity for feature matching, employ one single calculation method, and do not take into account the impact of instance relations on performance. For this reason, the study proposes a novel method based on reasoning in multiple instance relation graphs. Global and local instance relation graphs are generated by constructing similarity matrices to fully explore the fine-grained relations among the instances. Similarity reasoning is then conducted on the basis of the multiple instance relation graphs. For this purpose, reasoning is performed within the relation graphs in the image and text modalities, respectively. Then, the relations within each modality are mapped to the instance graphs for reasoning. Finally, reasoning within the instance graphs is performed. Furthermore, the neural network is trained by a step-by-step training strategy to adapt to the features of the image and text modalities. Experiments on the MIRFlickr and NUS-WIDE datasets demonstrate that the proposed method has distinct advantages in the metric mean average precision (*mAP*) and obtains a favorable Top-k-Precision curve. This also indicates that the proposed method deeply explores instance relations and thereby significantly improves the retrieval performance.

Key words: relation graph reasoning; cross-modal hash retrieval; similarity matrix; K-nearest neighbor (KNN); step-by-step training strategy

随着信息技术的发展, 不同模态的数据实现了海量的增长, 包括图像、文本、视频、音频等. 如何对数据进行有效处理以获得有价值的信息成为关键课题, 特别是在不同模态数据之间实现跨模态语义映射和检索显得尤其重

* 基金项目: 国家自然科学基金 (61966004, 61866004); 广西自然科学基金 (2019GXNSFDA245018)
收稿时间: 2022-02-12; 修改时间: 2022-04-30; 采用时间: 2022-07-18; jos 在线出版时间: 2023-06-16
CNKI 网络首发时间: 2023-06-19

要[1]. 例如: 对文本、图像和音频中的“太阳”语义, 搜索相应的源数据并建立联系, 对于数据存储和管理是很有意义的. 跨模态检索的目的就在于发现不同模态数据中具有相同语义数据的联系, 在工业界和学术界都引起了巨大的兴趣.

传统的跨模态检索方法[2−8]使用的是实值特征. 如果将实值特征转化为二值哈希码进行跨模态检索, 就是跨模态哈希检索方法. 与深度学习中那些体积巨大的模型相比, 这种参数量少, 训练时间短的模型具有很大的优势, 也具备了将来实际应用的可能性. 在跨模态哈希检索的方法[9−18]中, 深度联合语义重构哈希 (deep joint-semantics reconstructing hashing, DJSRH) 方法[14]创造性地构造了相似度矩阵, 通过特征相乘得到包含若干个实例相似关系的矩阵. 该方法的基本思路是分别为实值特征和哈希特征构造相似度矩阵, 并使两者对应的相似度数值对齐. 在此方法的基础上, 出现了很多改进工作, 例如基于联合模态分布的相似度哈希 (joint-modal distribution-based similarity hashing, JDSH) 方法[15]、深度语义对齐哈希 (deep semantic-alignment hashing, DSAH) 方法[16]和深度语义保持重构哈希 (deep semantic-preserving reconstruction hashing, DSPRH) 方法[17]等. 但是, 这些方法也存在着一些不足之处: (1) 相似度矩阵考虑了一对一的实例关系, 但没有考虑与其他实例的关系. (2) 特征间相似度的计算方法过于单一, 只使用了 cos(·,·) 函数, 没有从多个角度描述相似度. (3) 相似度矩阵只显示所有实例之间的相似度, 而没有进一步分析相似度信息, 寻找规律和进行更细粒度的操作. (4) 以往的方法都是基于整个数据集构建静态关系图, 这种完全基于全局的关系图并不一定能有效挖掘细粒度的相似信息. (5) 跨模态检索的基本方法是将图像和文本特征映射到公共空间中. 即使编码器功能强大, 也不可避免地会产生语义损失, 从而导致性能下降.

为了缓解以上问题, 本文提出了基于多重实例关系图推理的跨模态哈希检索 (cross-modal hash retrieval based on multiple instance relation graph reasoning, IRGR) 方法, 能够根据原始的图像文本对来构建实例关系并进行推理, 以获得更加准确的匹配, 其过程如图 1 所示. 首先, 使用某种评价指标可对实例间的关系进行一定程度的配对, 但是这样不可避免地会产生一定的误差; 然后, 将模态内的实例关系进行推理, 如实例 A 与实例 C 的关系和实例 C 与实例 E 的关系, 可以推理出实例 A 与实例 E 的关系; 同样, 模态间的实例关系也可以进行推理, 如图像实例 A 与图像实例 C 的关系和图像实例 C 与文本实例 C 的关系, 可以推理出图像实例 A 与文本实例 C 的跨模态关系; 最后, 通过这种推理方法可以挖掘出更加合适的实例匹配关系.

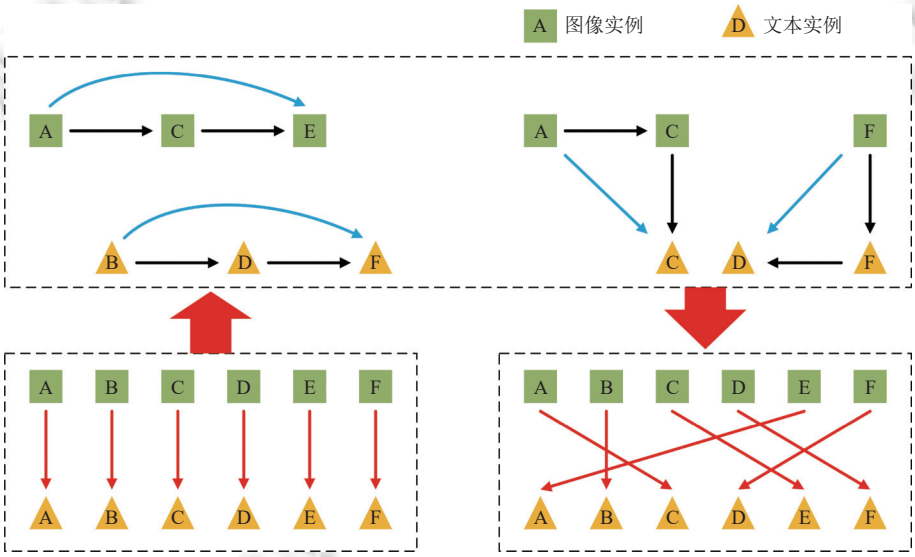


图 1 在两种模态中基于相邻关系进行推理的过程

IRGR 方法的贡献主要包括: (1) 在相似度矩阵的基础上, 使用 K 近邻 (K-nearest neighbor, KNN) 方法构建局部相似度关系图, 将全局关系图和局部关系图进行组合, 得到说明性更强的实例关系. 由于建立关系图时基于训练批次内的所有实例, 所以对实例关系的考虑更加细致. (2) 构造了图像关系实例图、文本关系实例图和总体实例

图, 依次使用模态内图推理、模态间图推理以及实例图推理 3 种方法, 充分考虑了 3 种关系图和相邻实例的关系, 使得相似度矩阵包含的信息更加准确充分. (3) 根据图像和文本两种模态各自的特点, 提出了一种新的分步训练策略. 该策略先将两种模态数据分别训练, 得到图像网络和文本网络, 然后再共同进行训练.

通过在 MIRFlickr 数据集和 NUS-WIDE 数据集上的实验可以证明, IRGR 方法与目前先进的跨模态哈希检索方法相比在性能上具有很大的优势. 这说明 IRGR 方法能够缓解当前跨模态哈希检索方法的不足之处, 同时也验证了 IRGR 方法的逻辑性和创新性.

1 相关工作

1.1 跨模态哈希检索方法

跨模态哈希检索方法是在源数据及其标签信息的基础上, 利用实值特征和哈希特征, 通过构造新的损失函数和训练方法, 尽可能挖掘图像和文本模态之间的相似度信息.

可扩展判别离散哈希 (scalable discriminative discrete hashing, SDDH) 方法[19]在目标函数中使用了正交约束和平衡约束, 从而保持了数据的异构相似度. 同时使用标签信息进行语义嵌入辅助生成哈希码, 进一步挖掘相似度信息以减少误差. 基于四元组的深度跨模态哈希 (quadruplet-based deep cross-modal hashing, QDCMH) 方法[20]使用图像模态的 1 个实例和文本模态的 3 个实例构造出一个四元组损失函数, 同时利用文本模态的一个实例和图像模态的 3 个实例构造出同样的四元组损失函数, 基于该损失函数构造了语义相关性保持模块. 然后考虑哈希码的生成及其表示学习, 构造了跨模态哈希检索模型. 在线标签一致性哈希 (online label consistent hashing, OLCH) 方法[21]利用多类分类方法获取语义标签, 首先学习目前的数据块, 然后随着数据块的更新而更新哈希函数, 并使用前向后向分裂方法保证稀疏性. 非线性监督离散哈希 (nonlinear supervised discrete hashing, NSDH) 方法[22]使用三层网络提取图像和文本特征, 实现了各种模态的语义增强和哈希码生成, 同时将标签矩阵和相似度信息参与到哈希码的学习中, 使用新的训练方式生成哈希码. 最大化共享潜在因子 (maximized shared latent factor, MSLF) 方法[23]利用标签信息获得不同模态的共享因子、单一模态的个体因子以及实例之间的相互关系, 并综合考虑这 3 种信息, 最大化不同模态的共享因子, 从而获得有效的哈希码.

1.2 基于相似度矩阵的方法

一些跨模态哈希检索方法利用特征相乘的方法分别构造实值特征和哈希特征的相似度矩阵, 并在损失函数中对这两种矩阵进行比较, 而对相似度矩阵作进一步处理可获得更细粒度的相似度信息.

DJSRH 方法[14]使用一般的深度模型提取图像和文本特征, 将训练批次中所有实例的特征相乘, 得到实例之间的相似度矩阵. 在将实值特征转化为哈希特征后, 构建了哈希特征的相似度矩阵. 在损失函数中比较两种矩阵, 使不同类型特征的相似度矩阵在语义上对齐, 并在此基础上训练跨模态哈希检索模型. 深度图相邻一致性保持网络 (deep graph-neighbor coherence preserving network, DGCPN)[24]基于整个数据集建立静态的 KNN 图, 在训练时提取训练批次内对应的相似度信息, 并结合 $\cos(\cdot,\cdot)$ 函数计算出的相似度信息构建实值特征相似度矩阵. 然后再计算哈希特征的相似度矩阵, 并在训练中依次使用实值特征和哈希特征. DSAH 方法[16]使用编码器-解码器结构将图像实值特征转换为图像哈希特征和文本实值特征, 对于文本特征也是类似的处理过程. 然后, 根据实值特征和哈希特征构建各自的相似度矩阵, 并在损失函数中比较实值特征和哈希特征之间以及哈希特征内部的相似度关系.

1.3 基于相似度推理的方法

基于实例构建的关系图或相似度矩阵可以进一步对实例关系进行处理和分析, 从而挖掘出更有价值的相似度信息.

相似度推理度量 (similarity inference metric, SIM) 方法[25]初始化查询节点和画廊节点之间的边以及画廊节点和画廊节点之间的边, 使用其他节点寻找查询节点和画廊节点之间的最短路径, 并根据三角形原理寻找画廊节点之间的最短路径. 然后更新查询节点和画廊节点之间的路径, 取前几条最短路径的平均值作为节点之间的距离. 高阶非局部哈希 (high-order nonlocal hashing, HNH) 方法[26]构建了图像特征和文本特征各自的相似度矩阵, 将两个

矩阵组合起来得到整体的相似度矩阵. 由于矩阵中的相似度关系是一一对应的, 通过考虑与其他实例的距离可使对应的相似度关系包含更全面的相邻关系. 语义重建跨模态哈希 (semantic-rebased cross-modal hashing, SRCH) 方法[27]建立图像和文本实例各自的 KNN 关系图, 只有当两个关系图中存在相同的实例关系时, 才会在新的实例关系图中采用这种实例关系. 无监督生成对抗跨模态哈希 (unsupervised generative adversarial cross-modal hashing, UGACH) 方法[28]构造图像和文本实例关系的无向图, 如果图像模态实例和文本模态实例共存, 则图像实例 A 和图像实例 B 之间的关系也存在于文本实例 A 和文本实例 B 中.

2 IRGR 方法

IRGR 方法的框架结构如图 2 所示. 首先, IRGR 方法基于每个训练批次内的实例, 为图像特征、文本特征以及实例分别构建相似度矩阵. 通过 KNN 方法得到每种相似度矩阵所对应的局部关系图. 其次, 使用图推理方法对 3 种局部关系图进行处理, 即图像和文本模态内部的图推理、关系图间的推理和实例图内部的推理, 然后将相似度矩阵与局部关系图相结合. 此外, IRGR 使用哈希特征构造哈希相似度矩阵, 与实值特征的关系图进行比较, 并采用分步训练策略进行训练, 即首先分别训练各种模态的网络, 然后再统一训练跨模态网络.

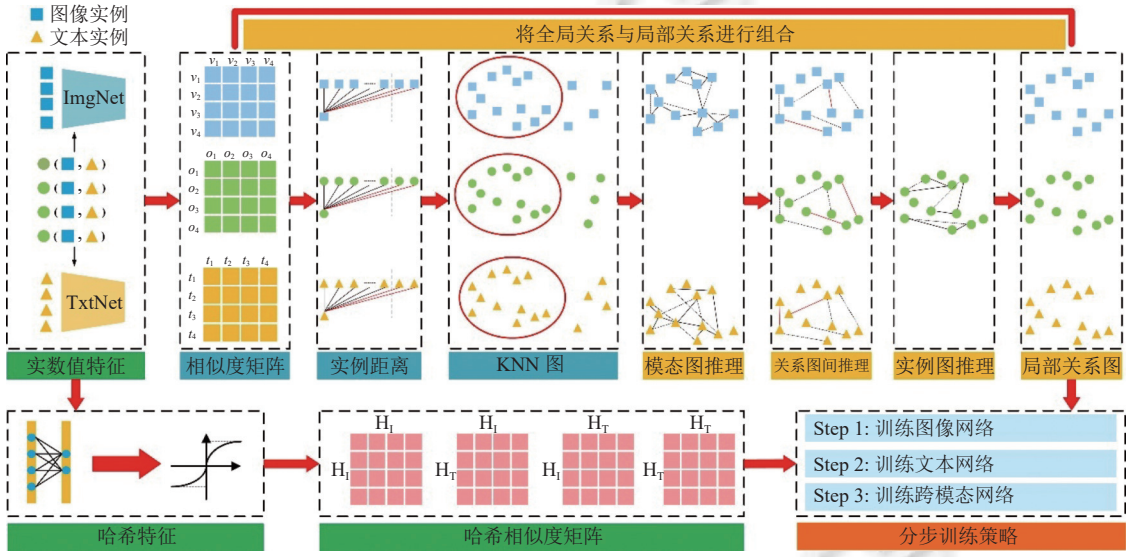


图 2 跨模态哈希检索方法 IRGR 的框架结构

2.1 问题定义

假设有 m 个实例, 其集合可以表示为 $O=\{o_i|1\leqslant i\leqslant m\}$, 每个实例可以用一个共现的图像-文本对 $o_i=(v_i, t_i)$ 来描述. 对于实值特征, 定义为 $\boldsymbol{F}_*\in\mathbb{R}^{m\times d}$, $*\in\{\mathrm{I}, \mathrm{T}\}$, 其中 m 表示实例的个数, d 表示维度, I 和 T 表示图像和文本数据. 对于哈希特征, 定义为 $\boldsymbol{B}_*\in\{-1, +1\}^{m\times c}$, $*\in\{\mathrm{I}, \mathrm{T}\}$, 其中 m 表示实例的个数, c 表示维度. 本文使用 sgn(·) 函数将实值特征转化为哈希特征, 使用 cos(·,·) 函数计算特征之间的相似度, 即:

$$\mathrm{sgn}(z)=\begin{cases}-1, & z\leqslant 0\\ 1, & z>0\end{cases} \tag{1}$$

$$\cos(x,y)=\frac{xy}{\|x\|\|y\|} \tag{2}$$

其中, $\|\cdot\|$表示向量的 L_2 范数.

2.2 特征提取

IRGR 方法使用预训练模型 AlexNet[29]提取图像特征 $\boldsymbol{F}_\mathrm{I}$, 使用词袋模型提取文本特征 $\boldsymbol{F}_\mathrm{T}$, 公式如下:

$$\boldsymbol{F}_*=E(*,\theta_*),\ *\in\{\mathrm{I},\mathrm{T}\} \tag{3}$$

其中, $E(\cdot,\cdot)$ 表示特征提取器, θ_*表示参数. 在跨模态哈希检索[30−36]中, 哈希特征的维度是哈希编码长度, 通常设为预定义的固定值. 所以, 可通过全连接层将实值特征的维度降为哈希编码长度, 再使用 sgn(·) 函数将实值特征转化为哈希特征, 即图像哈希特征 $\boldsymbol{B}_{\mathrm{I}}$ 和文本哈希特征 $\boldsymbol{B}_{\mathrm{T}}$, 公式如下:

$$\boldsymbol{B}_{*}=\mathrm{sgn}(\boldsymbol{F}_{*}),\ *\in\{\mathrm{I},\mathrm{T}\} \tag{4}$$

但是, 由于哈希特征的元素为–1 或+1 的形式, 会导致反向传播的梯度消失. 为此, 可使用 tanh(·) 函数代替 sgn(·) 函数来解决此问题. 因为当 tanh(·) 函数趋于极限时, 会无限逼近 sgn(·) 函数, 即:

$$\lim_{z\to\infty}\tanh(\delta z)=\mathrm{sgn}(z) \tag{5}$$

其中, 参数 δ 会随着训练次数的增加而变大.

2.3 构造相似度矩阵

实值特征和哈希特征经过 L_2 正则化处理后, 特征转化为 $\boldsymbol{F}'_{\mathrm{I}}$, $\boldsymbol{F}'_{\mathrm{T}}$, $\boldsymbol{B}'_{\mathrm{I}}$, $\boldsymbol{B}'_{\mathrm{I}}$. 在此基础上, 通过以下公式构造原始特征和哈希特征的相似度矩阵:

$$\boldsymbol{S}_{wx}=2\cos(\boldsymbol{F}'_{w},\boldsymbol{F}'_{x})-1,\quad w,x\in\{\mathrm{I},\mathrm{T}\} \tag{6}$$

$$\tilde{\boldsymbol{S}}=\beta\boldsymbol{S}_{\mathrm{II}}+(1-\beta)\boldsymbol{S}_{\mathrm{TT}},\quad \beta\in[0,1] \tag{7}$$

$$\begin{aligned}\boldsymbol{S}&=C(\boldsymbol{S}_{\mathrm{II}},\boldsymbol{S}_{\mathrm{TT}})=(1-\eta)\tilde{\boldsymbol{S}}+\eta\frac{\tilde{\boldsymbol{S}}\tilde{\boldsymbol{S}}^{\top}}{m}\\&=(1-\eta)[\beta\boldsymbol{S}_{\mathrm{II}}+(1-\beta)\boldsymbol{S}_{\mathrm{TT}}]+\frac{\eta}{m}[\beta^2\boldsymbol{S}_{\mathrm{II}}\boldsymbol{S}_{\mathrm{II}}^{\top}+\beta(1-\beta)\boldsymbol{S}_{\mathrm{II}}\boldsymbol{S}_{\mathrm{TT}}^{\top}+(1-\beta)\beta\boldsymbol{S}_{\mathrm{TT}}\boldsymbol{S}_{\mathrm{II}}^{\top}+(1-\beta)^2\boldsymbol{S}_{\mathrm{TT}}\boldsymbol{S}_{\mathrm{TT}}^{\top}]\end{aligned} \tag{8}$$

$$\boldsymbol{B}_{yz}=\cos(\boldsymbol{B}'_{y},\boldsymbol{B}'_{z}),\ y,z\in\{\mathrm{I},\mathrm{T}\} \tag{9}$$

其中, m 代表分批次训练中的实例个数, η 是调节参数. 通过这种方法构造了几种不同类型的特征相似度矩阵, 如表 1 所示.

表 1 不同类型的相似度矩阵

相似度矩阵	关系类型	矩阵类型
$\boldsymbol{S}$	实例-实例	实值相似度矩阵
$\boldsymbol{S}_{\mathrm{II}}$	图像特征-图像特征	实值相似度矩阵
$\boldsymbol{S}_{\mathrm{TT}}$	文本特征-文本特征	实值相似度矩阵
$\boldsymbol{S}_{\mathrm{IT}}$, $\boldsymbol{S}_{\mathrm{TI}}$	图像特征-文本特征	实值相似度矩阵
$\boldsymbol{B}_{\mathrm{II}}$	图像特征-图像特征	哈希相似度矩阵
$\boldsymbol{B}_{\mathrm{TT}}$	文本特征-文本特征	哈希相似度矩阵
$\boldsymbol{B}_{\mathrm{IT}}$, $\boldsymbol{B}_{\mathrm{TI}}$	图像特征-文本特征	哈希相似度矩阵

2.4 局部关系图

相似度矩阵 $\boldsymbol{S}$ 以实例个数为维度, 矩阵中的数值 $\boldsymbol{S}(i,j)=\boldsymbol{S}(j,i)$ 代表了实例 i 与实例 j 的相似度信息. 同时, 也可以将矩阵 $\boldsymbol{S}$ 看作邻接矩阵或者图, 图中的每个节点就是实例, 各边的权重即实例之间的相似度, 与矩阵 $\boldsymbol{S}$ 的数值直接对应. 在相似度矩阵中, 矩阵 $\boldsymbol{S}$、$\boldsymbol{S}_{\mathrm{II}}$ 和 $\boldsymbol{S}_{\mathrm{TT}}$ 分别代表了实例的相似度矩阵、图像的相似度矩阵和文本的相似度矩阵. 考虑到根据实例之间的相邻关系可以进一步对相似度信息进行细化, 本文通过将部分实例的相似度增大, 使得实例的局部关系更加突出.

此前, DGCPN 方法[24]使用数据集中所有的实例来建立关系图, 而 IRGR 方法则基于一个训练批次内的实例来建立关系图, 不同组的实例所构建的关系图各不相同. 通过这种学习到的关系图, 将节点之间的关系建模为条件概率问题, 可以在更加细粒度的层面表示实例之间的关系. IRGR 方法与 DGCPN 方法的不同之处在于: DGCPN 方法利用整个数据集构造出一个静态的关系图, 其中节点数目很多; 而 IRGR 方法对每一个训练批次内的实例构造关系图, 从而得到多个内容各不相同的关系图, 而且图的节点数目较少 (节点数目与批次大小相等). 相比 DGCPN 方法, IRGR 方法构造的关系图更侧重于局部关系, 能够获得实例之间更细粒度的相似度. 此外, 从训练的

角度来说, 节点数目较少的关系图有助于降低训练难度. 实际上, IRGR 方法为图像关系、文本关系和实例关系分别构造了关系图, 这样就使得不同类型的相似度关系更加突出, 比只使用实例关系的 DGCPN 方法更加的全面. 总体来说, IRGR 方法从不同的角度对关系图进行了处理, 使其包含的语义信息更加丰富, 从而有助于检索性能的提升.

假设将一个训练批次内的所有实例构造图 $G=\{O, E\}$, 其中 $O=\{o_i|1\leqslant i\leqslant m\}$代表实例集合, 每个实例包括图像和文本, 即 $o_i=(v_i, t_i)$, 则实例之间的相似度计算问题就转换为节点之间的相似度计算问题. 然后, 进一步将节点之间的关系图建模为条件概率问题. 假设 l_i 表示 v_i、t_i 和 o_i 的标签, 则有:

$$P(l_i,l_j)=P(l_i,l_j|O) \tag{10}$$

在计算两个节点之间的概率关系时, 仅将另一个节点作条件, 即:

$$P(l_i,l_j)=\sum_{q=1}^{M}P(l_i^F=l_q^F|o_i,o_q)P(l_j^F=l_q^F|o_j,o_q) \tag{11}$$

其中, l_*^F 是 o_* 的虚拟标签, 用来讨论节点之间的相似度. $P(l_i^F=l_q^F|o_i,o_q)$ 表示 o_i 和 o_q 具有相同虚拟标签的可能性. 假设每个节点都与它最近邻的 k 个节点相关, 则可将 $P(l_i^F=l_q^F|o_i,o_q)$ 定义为:

$$P(l_i^F=l_q^F|o_i,o_q)=\begin{cases}\dfrac{d(o_i,o_q)}{\sum_{o_p\in Ne(o_i,k)}d(o_i,o_p)}, & o_q\in Ne(o_i,k)\\ 0, & \text{其他}\end{cases} \tag{12}$$

其中, $Ne(o_i, k)$ 表示 o_i 的 k 个最近邻的集合, $d(o_i, o_q)$ 是 o_i 和 o_q 之间的余弦相似度.

以相似度矩阵为标准, 若定义 $d(o_i, o_q)=\boldsymbol{S}(i, q)=\boldsymbol{S}(q, i)$, 通过求得的 $P(l_i=l_j)$ 就能构造 o_i 和 o_j 之间的局部实例关系图 G_{O}, 对应的矩阵为 $\boldsymbol{S}_{G_{\mathrm{O}}}$. 若定义 $d(o_i, o_q)=\boldsymbol{S}_{\mathrm{II}}(i, q)=\boldsymbol{S}_{\mathrm{II}}(q, i)$, 则可构造局部图像关系图 G_{I}, 对应的矩阵为 $\boldsymbol{S}_{G_{\mathrm{I}}}$. 定义 $d(o_i, o_q)=\boldsymbol{S}_{\mathrm{TT}}(i, q)=\boldsymbol{S}_{\mathrm{TT}}(q, i)$, 则可构造局部文本关系图 G_{T}, 对应的矩阵为 $\boldsymbol{S}_{G_{\mathrm{T}}}$.

2.5 关系图推理

根据得到的 3 种关系图, 分别对图像关系、文本关系和实例关系进行推理, 可以弥补彼此之间的不足. 关系图 G_{I} 和 G_{T} 分别表示图像和文本模态内的关系, 而实例关系图 G_{O} 表示模态间的关系, 可以使用模态内的关系辅助推理模态间的关系. 3 种关系图的分步推理过程如图 3 所示, 以实例 A 和实例 B 之间的相似度为例, I 表示第 1 步模态内推理得到的结果, II 表示第 2 步模态间推理得到的结果, III 表示第 3 步实例关系图内部推理得到的结果.

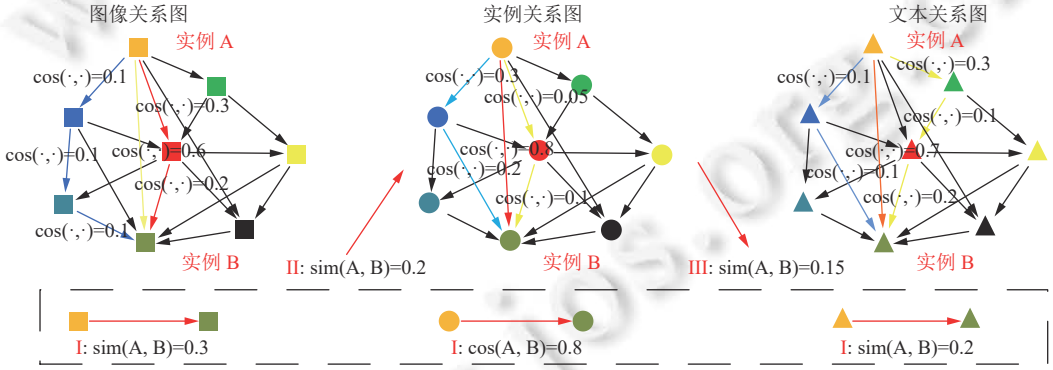


图 3 3 种关系图分步进行模态内图推理、模态间图推理以及实例图推理的过程

第 1 步. 基于关系图 G_{I} 和 G_{T} 进行模态内关系推理. 图 3 中展示了图像关系图和文本关系图内部的推理, 在两个节点之间关系可以通过其他节点进行推理得到. 基于图像关系图 G_{I}, 给定节点 o_i 和 o_j, $R(o_i, o_j)$ 被定义为从 o_i 到 o_j 的最小相似度路径. 具体来说, 假设 $\Omega(o_i, o_j)$ 表示包含从 o_i 到 o_j 的所有可能路径的集合, 则对于任意路径 $L\in\Omega(o_i, o_j)$, $L=(l_1, l_2, \ldots, l_n)$, 其中, $n=length(L)$, $l_1=o_i$, $l_n=o_j$, 则 $R(o_i, o_j)$ 的计算公式为:

$$R(o_i,o_j)=\min_{L\in\Omega(o_i,o_j)}\sum_{t=1}^{n-1}\boldsymbol{S}_{G_{\mathrm{I}}}(l_t,l_{t+1}) \tag{13}$$

如果通过若干个节点, 可以使得 o_i 和 o_j 的相似度变小, 则表明 o_i 和 o_j 之间的相似度应当符合与其他节点之间的关系. 而使用 cos(·,·) 函数计算特征相似度, 则仅考虑了 o_i 和 o_j 的相似度, 而没有考虑与其他节点的关系. 换句话说, 若存在 o_k 使下式成立:

$$\boldsymbol{S}_{G_{\mathrm{I}}}(o_i,o_k)+\boldsymbol{S}_{G_{\mathrm{I}}}(o_k,o_j)<\boldsymbol{S}_{G_{\mathrm{I}}}(o_i,o_j),\quad \forall 1\leqslant i,k,j\leqslant m \tag{14}$$

则通过图推理可以得到新的关系图:

$$\boldsymbol{S}_{G_{\mathrm{I}}}(o_i,o_j)=R(o_i,o_j)=\min_{1\leqslant k\leqslant m}\{\boldsymbol{S}_{G_{\mathrm{I}}}(o_i,o_k)+\boldsymbol{S}_{G_{\mathrm{I}}}(o_k,o_j)\} \tag{15}$$

文本关系图 G_{T} 也通过类似的方法进行推理:

$$\boldsymbol{S}_{G_{\mathrm{T}}}(o_i,o_k)+\boldsymbol{S}_{G_{\mathrm{T}}}(o_k,o_j)<\boldsymbol{S}_{G_{\mathrm{T}}}(o_i,o_j),\quad \forall 1\leqslant i,k,j\leqslant m \tag{16}$$

$$\boldsymbol{S}_{G_{\mathrm{T}}}(o_i,o_j)=R(o_i,o_j)=\min_{1\leqslant k\leqslant m}\{\boldsymbol{S}_{G_{\mathrm{T}}}(o_i,o_k)+\boldsymbol{S}_{G_{\mathrm{T}}}(o_k,o_j)\} \tag{17}$$

第 2 步. 基于 3 种关系图进行模态间图推理, 也就是实例关系图 G_{O} 中两个实例之间关系可以利用图像或者文本模态的关系进行推理. 在 3 种关系图中, $\boldsymbol{S}_{G_{\mathrm{O}}}(o_i,o_j)$ 表示实例 o_i 和 o_j 的关系, $\boldsymbol{S}_{G_{\mathrm{I}}}(o_i,o_j)$ 表示以图像特征计算的实例关系, $\boldsymbol{S}_{G_{\mathrm{T}}}(o_i,o_j)$ 表示以文本特征计算的实例关系, 各自的侧重点不一样. 如果在公式 $\boldsymbol{S}_{G_{\mathrm{O}}}(o_i,o_j)=\boldsymbol{S}_{G_{\mathrm{O}}}(o_i,o_k)+\boldsymbol{S}_{G_{\mathrm{O}}}(o_k,o_j)$ 中, 将 $\boldsymbol{S}_{G_{\mathrm{O}}}(o_k,o_j)$ 替换成 $\boldsymbol{S}_{G_{\mathrm{I}}}(o_k,o_j)$ 或者 $\boldsymbol{S}_{G_{\mathrm{T}}}(o_k,o_j)$ 可使得 $\boldsymbol{S}_{G_{\mathrm{O}}}(o_i,o_j)$ 的相似度变小, 则说明模态内的相似度信息证明实例 o_i 和 o_j 之间的相似度并不像表面上那么大. 这种推理方式将实例关系、图像关系和文本关系统一进行了考量, 相当于利用模态内的实例关系辅助计算模态间的实例关系, 即:

$$\boldsymbol{S}_{G_{\mathrm{O}}}(o_i,o_k)+\boldsymbol{S}_{G_{\mathrm{I}}}(o_k,o_j)<\boldsymbol{S}_{G_{\mathrm{O}}}(o_i,o_j),\quad \forall 1\leqslant i,k,j\leqslant m \tag{18}$$

$$\boldsymbol{S}_{G_{\mathrm{O}}}(o_i,o_j)=R(o_i,o_j)=\min_{1\leqslant k\leqslant m}\{\boldsymbol{S}_{G_{\mathrm{O}}}(o_i,o_k)+\boldsymbol{S}_{G_{\mathrm{I}}}(o_k,o_j)\} \tag{19}$$

$$\boldsymbol{S}_{G_{\mathrm{O}}}(o_i,o_k)+\boldsymbol{S}_{G_{\mathrm{T}}}(o_k,o_j)<\boldsymbol{S}_{G_{\mathrm{O}}}(o_i,o_j),\quad \forall 1\leqslant i,k,j\leqslant m \tag{20}$$

$$\boldsymbol{S}_{G_{\mathrm{O}}}(o_i,o_j)=R(o_i,o_j)=\min_{1\leqslant k\leqslant m}\{\boldsymbol{S}_{G_{\mathrm{O}}}(o_i,o_k)+\boldsymbol{S}_{G_{\mathrm{T}}}(o_k,o_j)\} \tag{21}$$

第 3 步. 在实例关系图 G_{O} 中进行实例内部的图推理, 即:

$$\boldsymbol{S}_{G_{\mathrm{O}}}(o_i,o_k)+\boldsymbol{S}_{G_{\mathrm{O}}}(o_k,o_j)<\boldsymbol{S}_{G_{\mathrm{O}}}(o_i,o_j),\quad \forall 1\leqslant i,k,j\leqslant m \tag{22}$$

$$\boldsymbol{S}_{G_{\mathrm{O}}}(o_i,o_j)=R(o_i,o_j)=\min_{1\leqslant k\leqslant m}\{\boldsymbol{S}_{G_{\mathrm{O}}}(o_i,o_k)+\boldsymbol{S}_{G_{\mathrm{O}}}(o_k,o_j)\} \tag{23}$$

经过模态内图推理和模态间图推理, 将 3 个关系图中的所有实例关系统一起来, 使相似度信息更加准确, 比之前单纯使用 cos(·,·) 函数来计算两个实例之间的相似度更具合理性. 相似度矩阵 $\boldsymbol{S}$、$\boldsymbol{S}_{\mathrm{II}}$、$\boldsymbol{S}_{\mathrm{TT}}$ 是一对一的实例关系, 局部关系图 G_{O}、G_{I}、G_{T} 是基于全局的关系, 将这两者相结合可使得实例关系更加全面, 也更有利于挖掘相似度信息.

IRGR 方法定义了参数 α 用于调节相似度矩阵的权重, 参数 δ 用于调节局部关系图的权重, 最终的相似度矩阵计算公式为:

$$\boldsymbol{S}=\alpha\boldsymbol{S}+\delta\boldsymbol{S}_{G_{\mathrm{O}}} \tag{24}$$

$$\boldsymbol{S}_{\mathrm{II}}=\alpha\boldsymbol{S}_{\mathrm{II}}+\delta\boldsymbol{S}_{G_{\mathrm{I}}} \tag{25}$$

$$\boldsymbol{S}_{\mathrm{TT}}=\alpha\boldsymbol{S}_{\mathrm{TT}}+\delta\boldsymbol{S}_{G_{\mathrm{T}}} \tag{26}$$

2.6 分步训练策略

图像和文本属于两种不同的模态, 但是之前的工作将这两种模态的数据统一进行训练, 忽略了他们的各自的规律与特征. 因此, IRGR 方法采用分步训练的方法, 首先分别训练图像网络和文本网络, 然后再统一训练整体网络, 并定义了参数 λ 用于调节损失函数范围.

第 1 步. 训练图像网络, 损失函数如下:

$$\min\lambda(\|\boldsymbol{S}-\boldsymbol{B}_{\mathrm{II}})\|_{\boldsymbol{F}}^2+\|\boldsymbol{S}_{\mathrm{II}}-\boldsymbol{B}_{\mathrm{II}}\|_{\boldsymbol{F}}^2) \tag{27}$$

其中, 参数 λ 用于调节损失函数, $\boldsymbol{B}_{\mathrm{II}}$ 可看作基于哈希特征所构造的图像关系图. 因此, 在损失函数中既考虑了实例关系图 $\boldsymbol{S}$ 和 $\boldsymbol{B}_{\mathrm{II}}$ 的关系, 也考虑了图像关系图 $\boldsymbol{S}_{\mathrm{II}}$ 与和 $\boldsymbol{B}_{\mathrm{II}}$ 的关系.

第 2 步. 训练文本网络, 损失函数如下:

$$\min \lambda(\|\boldsymbol{S}-\boldsymbol{B}_{\mathrm{TT}})\|_{\boldsymbol{F}}^2+\|\boldsymbol{S}_{\mathrm{TT}}-\boldsymbol{B}_{\mathrm{TT}}\|_{\boldsymbol{F}}^2) \tag{28}$$

该损失函数的构造方法与训练图像网络类似.

第 3 步. 将图像和文本网络共同训练, 损失函数如下:

$$\min \|\boldsymbol{B}_{\mathrm{IT}}-\boldsymbol{B}_{\mathrm{TI}}\|_{\boldsymbol{F}}^2+\left\|K_{\mathrm{diag}}-\boldsymbol{B}'_{\mathrm{IT}}\right\|_{\boldsymbol{F}}^2+\|\boldsymbol{S}_{\mathrm{IT}}-\boldsymbol{B}_{\mathrm{IT}}\|_{\boldsymbol{F}}^2+\|\boldsymbol{S}_{\mathrm{TI}}-\boldsymbol{B}_{\mathrm{TI}}\|_{\boldsymbol{F}}^2 \tag{29}$$

其中, $\boldsymbol{B}'_{\mathrm{IT}}$ 表示 $\boldsymbol{B}_{\mathrm{IT}}$ 的对角线向量. 参数 $\boldsymbol{K}_{\mathrm{diag}}$ 用于与 $\boldsymbol{B}'_{\mathrm{IT}}$ 进行比较, 以获得更好的实验结果. 如果图像特征和文本特征相同, 则 $\boldsymbol{B}_{\mathrm{IT}}$ 与 $\boldsymbol{B}_{\mathrm{TI}}$ 相同, 对角线元素值为 1. 此外, 将 $\boldsymbol{B}_{\mathrm{IT}}$、$\boldsymbol{B}_{\mathrm{TI}}$、$\boldsymbol{S}_{\mathrm{IT}}$ 和 $\boldsymbol{S}_{\mathrm{TI}}$ 作为训练目标, 这些矩阵同时考虑了图像和文本的特征. $\boldsymbol{S}_{\mathrm{IT}}$ 与 $\boldsymbol{B}_{\mathrm{IT}}$ 或者 $\boldsymbol{S}_{\mathrm{TI}}$ 与 $\boldsymbol{B}_{\mathrm{TI}}$ 之间进行实值特征与哈希特征的语义对齐计算, 能够互相弥补各自相似度信息的缺失.

综上所述, IRGR 方法的具体训练过程可描述如算法 1.

算法 1. IRGR 方法的训练过程.

输入: 图像数据集 I, 文本数据集 T, 批次大小 m, 哈希码长度 c, 最大训练批次 N, 参数值 $\eta, k, \lambda, \alpha, \beta, K_{\mathrm{diag}}$;
输出: 特征提取函数 $F_*=E(*,\theta_*), *\in\{\mathrm{I},\mathrm{T}\}$.

过程:

1. **for** n=1 to N **do**
2. 从图像集 I 和文本集 T 中提取实值特征和哈希特征.
3. 构造实值特征和哈希特征各自的相似度矩阵.
4. 使用 $\boldsymbol{S}$、$\boldsymbol{S}_{\mathrm{II}}$、$\boldsymbol{S}_{\mathrm{TT}}$ 分别构造关系图 G_{O}、G_{I}、G_{T}.
5. 基于关系图 G_{I} 和 G_{T} 分别进行模态内推理.
6. 基于关系图 G_{I}、G_{T} 和 G_{O} 进行关系图间推理.
7. 基于关系图 G_{O} 进行实例图内部推理.
8. 利用公式 (24)–公式 (26) 将相似度矩阵 $\boldsymbol{S}$、$\boldsymbol{S}_{\mathrm{II}}$、$\boldsymbol{S}_{\mathrm{TT}}$ 与关系图 G_{O}、G_{I}、G_{T} 进行融合.
9. 使用损失函数公式 (27) 单独训练图像网络并更新参数 θ_{I}.
10. 使用损失函数公式 (28) 单独训练文本网络并更新参数 θ_{T}.
11. 使用损失函数公式 (29) 共同训练图像和文本网络并更新参数 θ_{I} 和 θ_{T}.
12. **end for**
13. **return** $F_*=E(*,\theta_*),\ *\in\{\mathrm{I},\mathrm{T}\}$.

3 实验结果分析

3.1 数据集与评估指标

MIRFlickr 数据集[37]由 25 000 个实例构成, 每个实例都包含图像和标签. 文本标签分为 24 个语义类别, 每幅图像可以有多个语义标签进行标注. NUS-WIDE 数据集[38]包含 269 648 幅图像, 文本标签共 81 个类别, 图像和对应的标签构成了实例. 在实验中, 这两个数据集的查询集包含 2 000 个实例, 训练集包含 5 000 个实例, 其他实例作为数据库.

实验采用最常用的精确率 precision 和召回率 recall 作为评估指标, 并可以利用求出的精确率和召回率得到对应的 Top-k-Precision 曲线 (横坐标: Top-k; 纵坐标: precision).

平均精度 AP 值为通过求积分得到的 precision-recall 曲线下方的面积, mAP 值就是将所有类别的 AP 值取平均值, 计算公式为:

$$AP(i)=\int_0^1 p(r)\mathrm{d}r \tag{30}$$

$$mAP = \frac{\sum_{i=1}^{M} AP(i)}{M} \tag{31}$$

其中, i 指的是类别序号, M 指的是类别的总数.

3.2 实验设置

IRGR 方法基于 PyTorch 1.3.1 和 Python 3.7 在 RTX2080Ti 的 GPU 上进行实验, 使用随机梯度下降法进行梯度优化, 动量为 0.9, 权重衰减为 0.000 5. 在测试阶段, 通过计算 *mAP* 值和 Top-k-Precision 曲线进行评估. 此外, 总体评价标准定义为图像检索文本的 *mAP* 值和文本检索图像的 *mAP* 值两者之和.

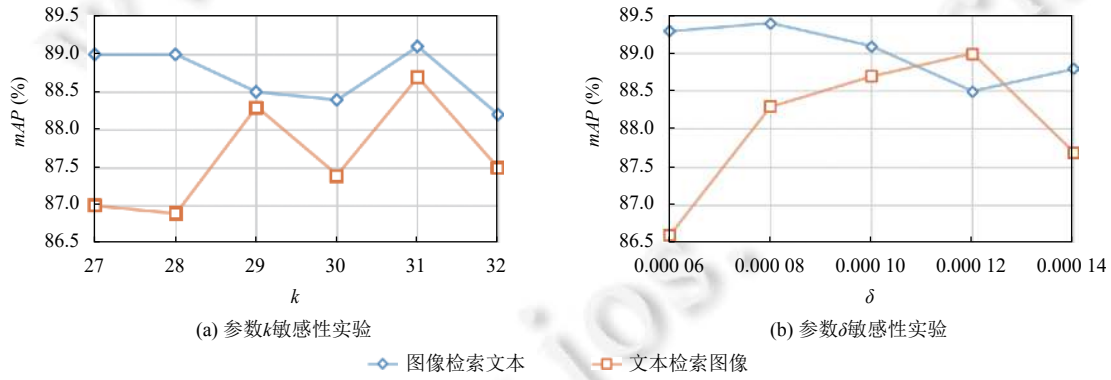
参数设置需要综合考虑跨模态哈希检索系统的最终实验性能和训练效率. 当数据集总共训练 10 次时, 模型的训练效率较高且检索性能最好. 由于 JDSH 方法[15]、DSAH 方法[16]、DGCPN 方法[24]和 IRGR 方法均采用了相同的基线模型 DJSRH 方法[14], 故 IRGR 方法借鉴了这些方法的训练参数, 并在实验过程中对参数进行了微调. 经过验证, ImgNet 和 TxtNet 的学习率分别设为 0.001 和 0.01. 此外, 当批次小于 32 时, 系统性能会下降; 当批次大于 32 时, 则会使得训练效率下降, 而且对于性能提升没有明显作用. 所以, 批次大小设为 32 有利于平衡系统的性能和训练效率.

在训练 MIRFlickr 数据集时, 公式 (7) 中的 β 设为 0.9; 在训练 NUS-WIDE 数据集时, β 设为 0.6. 此外, 公式 (8) 中的 η 设为 0.4, 公式 (24)–公式 (26) 中的 α 设为 1.5. 公式 (27) 和公式 (28) 的 λ 设为 0.1, 公式 (29) 中 K_{diag} 的元素值设为 1.5. 最后, 通过参数敏感性实验的验证, 将公式 (12) 中的 k 设为 31, 公式 (24)–公式 (26) 中的 δ 设为 0.000 1.

3.3 参数敏感性实验

为了验证参数以获得更好的效果, 在 MIRFlickr 数据集上进行了 IRGR 的参数敏感性实验, 哈希码长度设为 128 位. 参数 k 表示局部关系图的阈值, 在训练时设置批次大小为 32 时, 固定参数 δ 为 0.000 1, 分别设置 k 为 27、28、29、30、31、32. 参数 δ 表示经过推理关系图的范围, 固定参数 k 为 31, 分别设置 δ 为 0.000 06、0.000 08、0.000 1、0.000 12、0.000 14.

参数敏感性实验结果如图 4 所示, 当局部关系图的阈值 k 为 31, 推理关系图的范围 δ 为 0.000 1 时, IRGR 取得了最好的效果, 即图像检索文本的 *mAP* 值和文本检索图像的 *mAP* 值两者之和最大.



(a) 参数k敏感性实验　　(b) 参数δ敏感性实验

图 4　参数敏感性实验结果

3.4 实验结果对比

为了验证 IRGR 方法的效果, 与目前最先进的跨模态哈希检索方法进行了对比, 包括: 平均近似哈希 (average approximate hashing, AAH) 方法[39]、基于分层语义交互的深度哈希网络 (hierarchical semantic interaction-based deep hashing network, HSIDHN) 方法[40]、HNH 方法[26]、判别结构保持哈希 (discriminative structure preserving hashing, DSPH) 方法[41]、快速判别离散哈希 (fast discriminative discrete hashing, FDDH) 方法[42]、SDDH 方法[19]、NSDH 方法[22]、基于多标签语义保持的深度跨模态哈希 (multi-label semantics preserving based deep cross-modal

hashing, MLSPH) 方法[43]、基于多注意力的语义深度哈希 (multi-attention based semantic deep hashing, MSDH) 方法[44]、MSLF 方法[23]和 QDCMH 方法[20].

表 2 和表 3 展示了在 MIRFlickr 和 NUS-WIDE 数据集上各种方法的性能对比 ("—"表示在原文献中未提供该项数据, 粗体表示最佳性能). 从表中可以看到, 无论是在 MIRFlickr 还是 NUS-WIDE 数据集, IRGR 方法都取得了良好的性能. 以 MIRFlickr 数据集为例, 当基于图像检索文本时, HNH 方法效果最好, IRGR 方法其次, IRGR 方法比其他方法至少高出了 2.34%, 3.15%, 4.33%, 12.28%. 基于文本检索图像时, IRGR 方法的效果最好, HNH 方法其次, IRGR 方法比其他方法至少高出了 1.13%, 2.30%, 1.40%, 1.40%. 虽然在单独进行某一种检索时, IRGR 的效果不一定是最好的, 但是将两种检索方式的 mAP 值相加, 则 IRGR 方法和 HNH 方法的效果最好. 可以看出, 与目前最先进的方法相比, IRGR 方法在大数据集上的效果很有竞争力, 这说明了 IRGR 方法具有良好的有效性和鲁棒性.

表 2 MIRFlickr 数据集上各种方法的性能对比

方法	图像检索文本				文本检索图像			
	16位	32位	64位	128位	16位	32位	64位	128位
AAH[39]	0.7145	0.7230	0.7271	0.7283	0.8137	0.8198	0.8251	0.8281
DSPH[41]	0.6473	0.6610	0.6703	—	0.6581	0.6781	0.6818	—
FDDH[42]	—	0.7392	0.7578	0.7631	—	0.8022	0.8250	0.8357
HNH[26]	—	**0.8830**	**0.8950**	**0.9020**	—	0.8540	0.8680	0.8780
HSIDHN[40]	0.7978	0.8097	0.8179	—	0.7802	0.7946	0.8115	—
MLSPH[43]	0.8076	0.8235	0.8337	—	0.7852	0.8041	0.8146	—
MSDH[44]	0.7836	0.7905	0.8017	—	0.7573	0.7635	0.7813	—
MSLF[23]	0.6988	0.7175	0.7222	0.7294	0.7572	0.7763	0.7892	0.7959
NSDH[22]	0.7363	0.7561	0.7656	0.7712	0.7836	0.8014	0.8183	0.8229
QDCMH[20]	0.7635	0.7688	0.7713	—	0.7762	0.7725	0.7859	—
SDDH[19]	0.7210	0.7394	0.7454	0.7494	0.7917	0.8132	0.8241	0.8328
IRGR	**0.8310**	0.8550	0.8770	0.8940	**0.8250**	**0.8770**	**0.8820**	**0.8920**

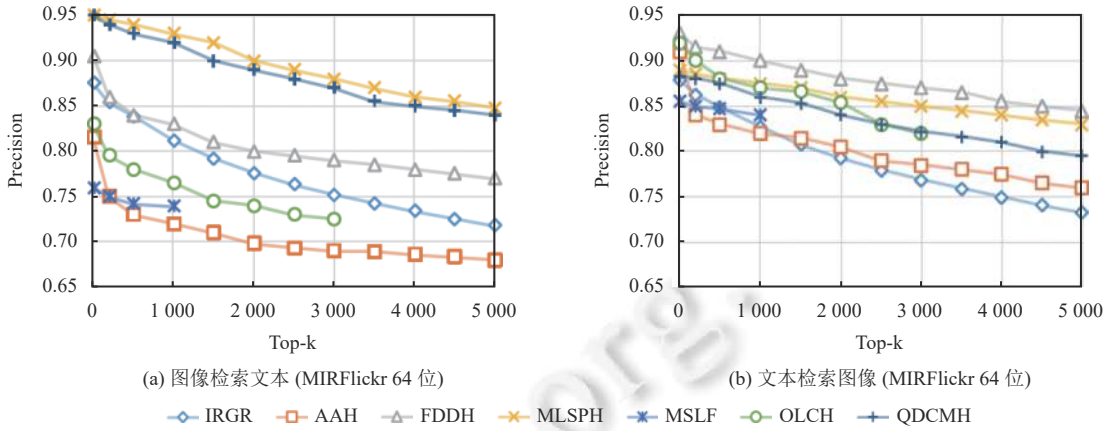
表 3 NUS-WIDE 数据集上各种方法的性能对比

方法	图像检索文本				文本检索图像			
	16位	32位	64位	128位	16位	32位	64位	128位
AAH[39]	0.6409	0.6439	0.6515	0.6549	0.7379	0.7533	0.7595	0.7629
FDDH[42]	—	0.6970	0.6910	0.7118	—	**0.8133**	**0.8111**	**0.8244**
HNH[26]	—	**0.8020**	**0.8160**	**0.8470**	—	0.7760	0.7960	0.8020
HSIDHN[40]	0.6498	0.6787	0.6834	—	0.6396	0.6529	0.6792	—
MLSPH[43]	0.6405	0.6604	0.6734	—	0.6433	0.6633	0.6724	—
MSDH[44]	0.6633	0.6859	0.7155	—	0.6359	0.6632	0.6934	—
MSLF[23]	0.6213	0.6339	0.6374	0.6482	0.7212	0.7427	0.7578	0.7765
NSDH[22]	0.6418	0.6604	0.6732	0.6791	**0.7658**	0.7892	0.7939	0.8011
SDDH[19]	0.6510	0.6564	0.6670	0.6733	0.7638	0.7790	0.7945	0.7990
IRGR	**0.7560**	0.7930	**0.8160**	0.8390	0.7500	0.7830	0.8040	0.8170

需要指出的是, HNH 方法和 IRGR 方法均采用了分别为两种模态构造相似度矩阵的方法, 并在此基础上对相似度矩阵作进一步的处理, 但两种方法对相似度矩阵的处理方式不一样. 实验结果也证明, HNH 方法和 IRGR 方法在所有对比方法中性能最佳. 这说明构造相似度矩阵有利于性能的提升, 同时也说明了 IRGR 方法对相似度矩阵进行细粒度处理在逻辑上具有合理性且显著提升了实验性能.

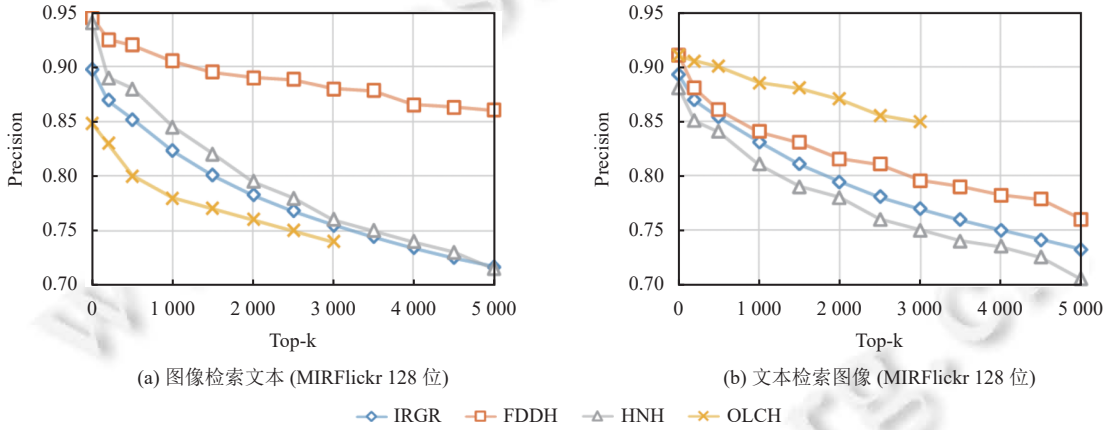
为了进一步对比 IRGR 方法与其他跨模态哈希检索方法, 使用了 Top-k-Precision 曲线测试效果. 此处对比了

一些主流方法在 MIRFlickr 和 NUS-WIDE 数据集上哈希码长度分别为 64 位和 128 位的结果, 如图 5–图 8 所示. 从图中曲线可以看到, 在 MIRFlickr 数据集和 NUS-WIDE 数据集上, 尽管 IRGR 方法在相同断点上的精确率不一定是最高的, 但是具有相对较高的性能, 在同类方法中具有很大的优势. 而且 IRGR 方法的综合性能与最佳方法相比差距不大, 总体上具备良好的竞争力.



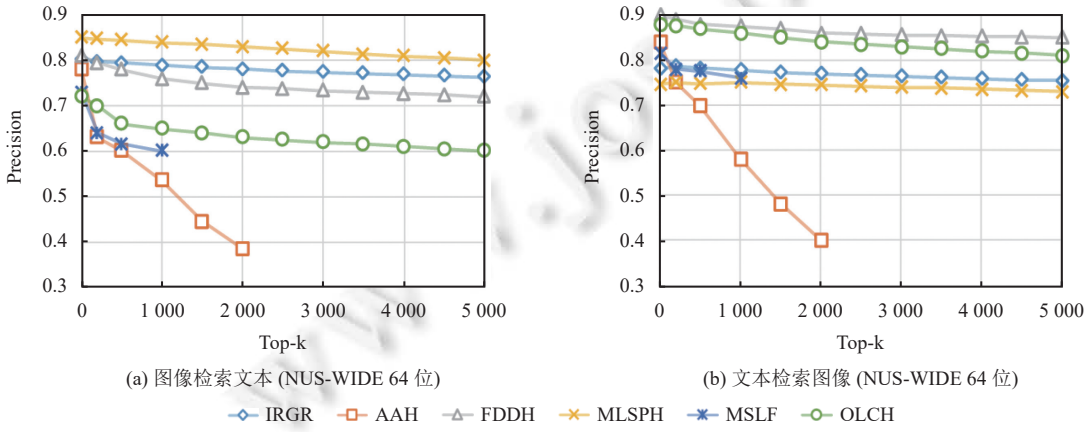
(a) 图像检索文本 (MIRFlickr 64 位)　(b) 文本检索图像 (MIRFlickr 64 位)

图 5　在 MIRFlickr 数据集 (64 位) 上的 Top-k-Precision 曲线



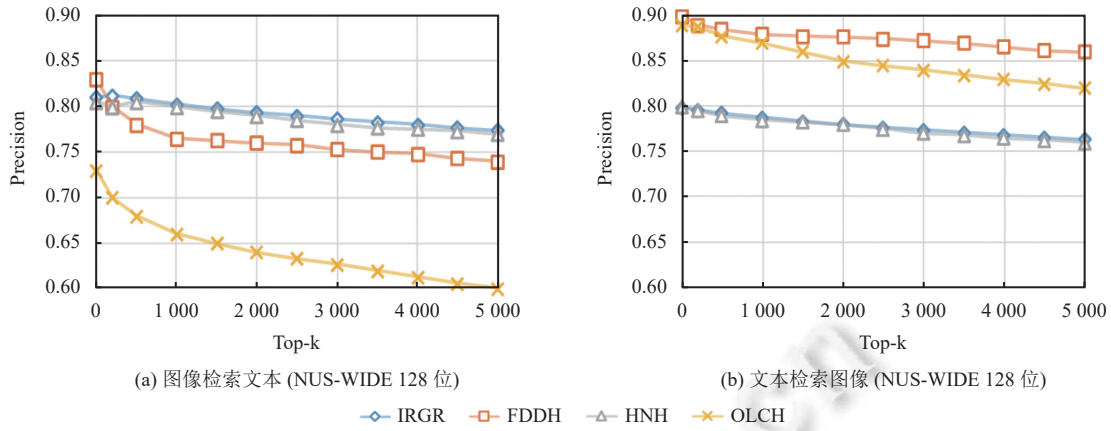
(a) 图像检索文本 (MIRFlickr 128 位)　(b) 文本检索图像 (MIRFlickr 128 位)

图 6　在 MIRFlickr 数据集 (128 位) 上的 Top-k-Precision 曲线



(a) 图像检索文本 (NUS-WIDE 64 位)　(b) 文本检索图像 (NUS-WIDE 64 位)

图 7　在 NUS-WIDE 数据集 (64 位) 上的 Top-k-Precision 曲线



(a) 图像检索文本 (NUS-WIDE 128 位)　　(b) 文本检索图像 (NUS-WIDE 128 位)

图 8　在 NUS-WIDE 数据集 (128 位) 上的 Top-k-Precision 曲线

3.5　消融实验

为了证明 IRGR 方法各个创新点的效果, 在 MIRFlickr 数据集上进行了消融实验, 哈希码长度设为 128 位. 评价标准为图像检索文本的 *mAP* 值和文本检索图像的 *mAP* 值两者之和.

在 IRGR 的基础上分别进行各个模块的消融实验, 其配置如表 4 所示. IRGR-1 方法消融了分步训练策略, IRGR-2 方法消融了关系图推理. 由于关系图推理是在局部关系图的基础上进行的, 所以 IRGR-3 方法既消融了局部关系图, 也消融了关系图推理; IRGR-4 方法消融了局部关系图, 仅使用相似度矩阵进行关系图推理. 表 5 给出了关系图推理内部的消融实验, IRGR-5 方法消融了模态内推理, IRGR-6 方法消融了模态间推理, IRGR-7 方法消融了实例图推理.

表 4　IRGR 消融实验配置

方法	局部关系图	关系图推理	分步训练策略
IRGR	√	√	√
IRGR-1	√	√	×
IRGR-2	√	×	√
IRGR-3	×	×	√
IRGR-4	×	√	√

表 5　图推理过程消融实验配置

方法	模态内推理	模态间推理	实例图推理
IRGR-5	×	√	√
IRGR-6	√	×	√
IRGR 7	√	√	×

表 6 展示了消融实验的实验结果. 从表中数据可见, 尽管在图像检索文本时, IRGR-1 方法的效果要比 IRGR 方法略好一点, 但在文本检索图像时的效果则差一些. 以图像检索文本的 *mAP* 值和文本检索图像的 *mAP* 值两者之和作为总体评价标准来看, IRGR-1 方法的性能稍低, 表明分步训练策略有助于模型性能的提升. IRGR-2 方法的实验结果表明关系图推理对于提升模型性能有积极作用. IRGR-3 方法在消融了局部关系图之后, 无法使用关系图推理, 使得模型效果显著下降. 即便 IRGR-4 方法采用相似度矩阵进行关系图推理, 也无法获得最佳效果. 这两种方法体现了局部关系图的作用. IRGR-5 方法、IRGR-6 方法和 IRGR-7 方法表明 IRGR 方法的 3 种推理均有各自的效果.

表 6　消融实验结果

方法	图像检索文本	文本检索图像	方法	图像检索文本	文本检索图像
IRGR	0.894	0.892	IRGR-4	0.892	0.871
IRGR-1	0.900	0.872	IRGR-5	0.894	0.873
IRGR-2	0.881	0.883	IRGR-6	0.881	0.867
IRGR-3	0.576	0.576	IRGR-7	0.877	0.876

3.6　与 DJSRH 方法的对比

IRGR 方法以 DJSRH 方法[14]作为基线模型, 直接采用了 DJSRH 方法中提取图像和文本特征的方式, 部分实验参数也与 DJSRH 方法一致. 在实验中, 两者采用的实验环境相同, 参数值也相同. 所以, 将两者进行单独对比, 更

能说明 IRGR 创新的效果.

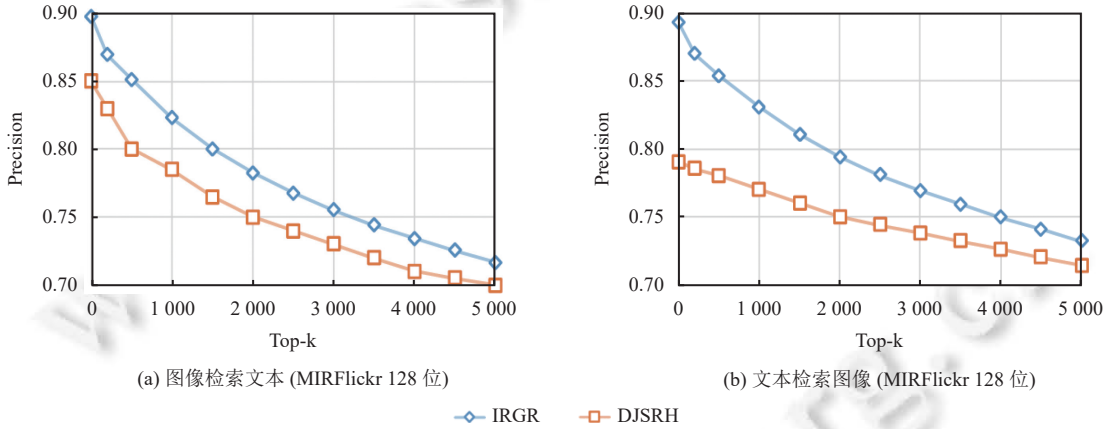
表 7 和表 8 分别展示了在 MIRFlickr 和 NUS-WIDE 数据集上两种方法的实验对比结果. 可以看到, IRGR 方法比 DJSRH 方法在 *mAP* 指标上有了明显的提升. 在 MIRFlickr 数据集上, 图像检索文本时分别提升了 2.1%, 1.2%, 1.5%, 1.8%, 文本检索图像时分别提升了 3.9%, 5.5%, 4.7%, 4.5%. 在 NUS-WIDE 数据集上, 图像检索文本时分别提升了 3.2%, 2.0%, 1.8%, 2.2%, 文本检索图像时分别提升了 3.8%, 3.9%, 3.3%, 2.8%. 图 9 和图 10 分别展示了在 MIRFlickr 和 NUS-WIDE 数据集上哈希码长度设为 128 位时两种方法的 Top-k-Precision 曲线. 可以看出, IRGR 方法在相同断点上的 precision 值比 DJSRH 方法更高, 具有明显的性能优势.

表 7　在 MIRFlickr 数据集上 IRGR 与 DJSRH 的对比结果

方法	图像检索文本				文本检索图像			
	16位	32位	64位	128位	16位	32位	64位	128位
DJSRH	0.810	0.843	0.862	0.876	0.786	0.822	0.835	0.847
IRGR	0.831	0.855	0.877	0.894	0.825	0.877	0.882	0.892

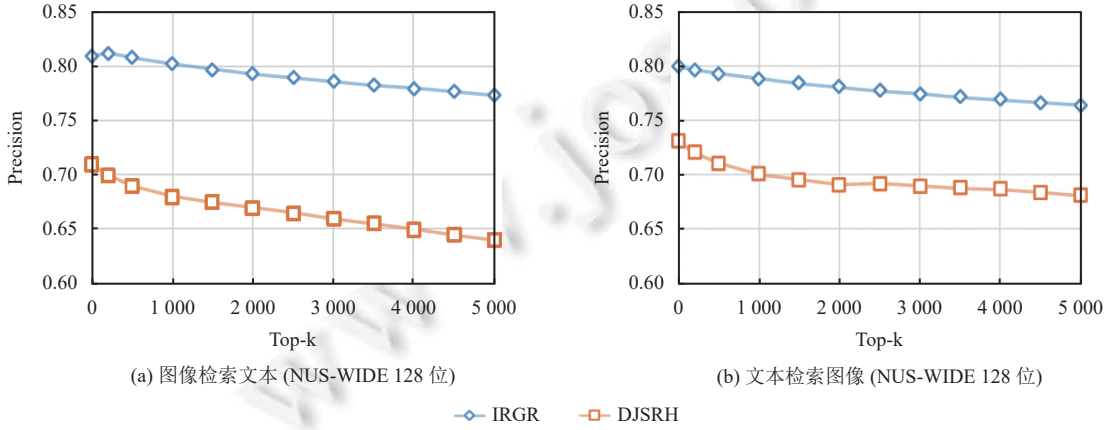
表 8　在 NUS-WIDE 数据集上 IRGR 与 DJSRH 的对比结果

方法	图像检索文本				文本检索图像			
	16位	32位	64位	128位	16位	32位	64位	128位
DJSRH	0.724	0.773	0.798	0.817	0.712	0.744	0.771	0.789
IRGR	0.756	0.793	0.816	0.839	0.750	0.783	0.804	0.817



(a) 图像检索文本 (MIRFlickr 128 位)　(b) 文本检索图像 (MIRFlickr 128 位)

图 9　在 MIRFlickr 数据集上 IRGR 对比 DJSRH 的 Top-k-Precision 曲线



(a) 图像检索文本 (NUS-WIDE 128 位)　(b) 文本检索图像 (NUS-WIDE 128 位)

图 10　在 NUS-WIDE 数据集上 IRGR 对比 DJSRH 的 Top-k-Precision 曲线

4 结 论

本文提出新的跨模态哈希检索方法 IRGR, 对此前的静态关系图进行改进, 采用了基于训练批次的动态关系图进行推理. 在全局关系图的基础上, 采用 KNN 的方法构造了局部关系图, 并将两种关系图进行组合. 为了进一步挖掘实例之间的关系, 采用图推理的方式, 依次进行模态内推理, 模态间推理以及实例图内部推理, 充分考虑了每个实例与其他实例之间的关系. 此外, 基于不同模态数据独有的特点, 设计了分步训练策略. IRGR 方法与国际上一些先进的方法[15−17,24]都是以相似度矩阵[14]作为共同的出发点, 而 IRGR 融合了关系图的相似度评价标准, 并进一步提出了关键的图推理模块, 从而取得了显著的效果. 这充分说明了 IRGR 方法的创新具有坚实的逻辑性和合理性, 并有效地提升了系统的检索性能.

References:

[1] Li ZX, Wei HY, Zhang CL, Ma HF, Shi ZZ. Research progress on image captioning. Journal of Computer Research and Development, 2021, 58(9): 1951–1974 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2021.20200281]

[2] Chun S, Oh SJ, de Rezende RS, Kalantidis Y, Larlus D. Probabilistic embeddings for cross-modal retrieval. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE Computer Society, 2021. 8411–8420. [doi: 10.1109/CVPR46437.2021.00831]

[3] Hu P, Peng X, Zhu HY, Zhen LL, Lin J. Learning cross-modal retrieval with noisy labels. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE Computer Society, 2021. 5399–5409. [doi: 10.1109/CVPR46437.2021.00536]

[4] Jing LL, Vahdani E, Tan JX, Tian YL. Cross-modal center loss for 3D cross-modal retrieval. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE Computer Society, 2021. 3141–3150. [doi: 10.1109/CVPR46437.2021.00316]

[5] Yu T, Yang Y, Li Y, Liu L, Fei HL, Li P. Heterogeneous attention network for effective and efficient cross-modal retrieval. In: Proc. of the 44th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2021. 1146–1156. [doi: 10.1145/3404835.3462924]

[6] Wang K, Herranz L, van de Weijer J. Continual learning in cross-modal retrieval. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops. Nashville: IEEE Computer Society, 2021. 3623–3633. [doi: 10.1109/CVPRW53098.2021.00402]

[7] Wang X, Hu P, Zhen LL, Peng DZ. DRSL: Deep relational similarity learning for cross-modal retrieval. Information Sciences, 2021, 546: 298–311. [doi: 10.1016/j.ins.2020.08.009]

[8] Li ZX, Ling F, Zhang CL, Ma HF. Cross-media image-text retrieval with two level similarity. Acta Electronica Sinica, 2021, 49(2): 268–274 (in Chinese with English abstract). [doi: 10.12263/DZXB.20191037]

[9] Zhan YW, Wang YX, Sun Y, Wu XM, Luo X, Xu XS. Discrete online cross-modal hashing. Pattern Recognition, 2022, 122: 108262. [doi: 10.1016/j.patcog.2021.108262]

[10] Li JZ. Deep semantic cross modal hashing based on graph similarity of modal-specific. IEEE Access, 2021, 9: 96064–96075. [doi: 10.1109/ACCESS.2021.3093357]

[11] Yuan X, Wang GZ, Chen ZK, Zhong FM. CHOP: An orthogonal hashing method for zero-shot cross-modal retrieval. Pattern Recognition Letters, 2021, 145: 247–253. [doi: 10.1016/j.patrec.2021.02.016]

[12] Zhang XW, Lin JZ, Zhou Y. DHLBT: Efficient cross-modal hashing retrieval method based on deep learning using large batch training. Int'l Journal of Software Engineering and Knowledge Engineering, 2021, 31(7): 949–971. [doi: 10.1142/S0218194021500297]

[13] Li ZX, Ling F, Tang ZJ, Ma HF, Shi ZP. Unsupervised cross-media hashing retrieval based on multi-head attention network. Scientia Sinica Informationis, 2021, 51(12): 2053–2368 (in Chinese with English abstract). [doi: 10.1360/SSI-2020-0264]

[14] Su SP, Zhong ZS, Zhang C. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 3027–3035. [doi: 10.1109/ICCV.2019.00312]

[15] Liu S, Qian SS, Guan Y, Zhan JW, Ying L. Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval. In: Proc. of the 43rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2020. 1379–1388. [doi: 10.1145/3397271.3401086]

[16] Yang DJ, Wu DY, Zhang WQ, Zhang HS, Li B, Wang WP. Deep semantic-alignment hashing for unsupervised cross-modal retrieval. In: Proc. of the 2020 Int'l Conf. on Multimedia Retrieval. Dublin: ACM, 2020. 44–52. [doi: 10.1145/3372278.3390673]

[17] Cheng SL, Wang LJ, Du AY. Deep semantic-preserving reconstruction hashing for unsupervised cross-modal retrieval. Entropy, 2020, 22(11): 1266. [doi: 10.3390/e22111266]

[18] Li ZX, Hou CW, Xie XM. Enhancing cross-modal hash retrieval with multiple similarity matrices. Journal of Computer-aided Design & Computer Graphics, 2022, 34(6): 933–945 (in Chinese with English abstract). [doi: 10.3724/SP.J.1089.2022.19044]

[19] Qin JY, Fei LK, Zhu J, Wen J, Tian CW, Wu S. Scalable discriminative discrete hashing for large-scale cross-modal retrieval. In: Proc. of the 2021 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Toronto: IEEE, 2021. 4330–4334. [doi: 10.1109/ICASSP39728.2021.9413871]

[20] Liu H, Xiong J, Zhang N, Liu FM, Zou XT. Quadruplet-based deep cross-modal hashing. Computational Intelligence and Neuroscience, 2021, 2021: 9968716. [doi: 10.1155/2021/9968716]

[21] Yi JH, Liu X, Cheung YM, Xu X, Fan WT, He Y. Efficient online label consistent hashing for large-scale cross-modal retrieval. In: Proc. of the 2021 IEEE Int'l Conf. on Multimedia and Expo. Shenzhen: IEEE, 2021. 1–6. [doi: 10.1109/ICME51207.2021.9428323]

[22] Yang Z, Yang L, Raymond OI, Zhu L, Huang WT, Liao ZF, Long J. NSDH: A nonlinear supervised discrete hashing framework for large-scale cross-modal retrieval. Knowledge-based Systems, 2021, 217: 106818. [doi: 10.1016/j.knosys.2021.106818]

[23] Wang S, Zhao H, Nai K. Learning a maximized shared latent factor for cross-modal hashing. Knowledge-based Systems, 2021, 228: 107252. [doi: 10.1016/j.knosys.2021.107252]

[24] Yu J, Zhou H, Zhan YB, Tao DC. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI, 2021. 4626–4634. [doi: 10.1609/aaai.v35i5.16592]

[25] Jia MX, Zhai YP, Lu SJ, Ma SW, Zhang J. A similarity inference metric for RGB-infrared cross-modality person re-identification. In: Proc. of the 29th Int'l Joint Conf. on Artificial Intelligence. Yokohama: IJCAI.org, 2021. 1026–1032.

[26] Zhang PF, Luo YD, Huang Z, Xu XS, Song JK. High-order nonlocal hashing for unsupervised cross-modal retrieval. World Wide Web, 2021, 24(2): 563–583. [doi: 10.1007/s11280-020-00859-y]

[27] Wang WW, Shen YM, Zhang HF, Yao YZ, Liu L. Set and Rebase: Determining the semantic graph connectivity for unsupervised cross-modal hashing. In: Proc. of the 29th Int'l Joint Conf. on Artificial Intelligence. Yokohama: IJCAI.org , 2020. 119.

[28] Zhang J, Peng YX, Yuan MK. Unsupervised generative adversarial cross-modal hashing. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence and the 30th Innovative Applications of Artificial Intelligence Conf. and the 8th AAAI Symp. on Educational Advances in Artificial Intelligence. New Orleans: AAAI, 2018. 67.

[29] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84–90. [doi: 10.1145/3065386]

[30] Zhang DL, Wu XJ, Yu J. Label consistent flexible matrix factorization hashing for efficient cross-modal retrieval. ACM Trans. on Multimedia Computing, Communications, and Applications, 2021, 17(3): 90. [doi: 10.1145/3446774]

[31] Song G, Tan XY, Zhao J, Yang M. Deep robust multilevel semantic hashing for multi-label cross-modal retrieval. Pattern Recognition, 2021, 120: 108084. [doi: 10.1016/j.patcog.2021.108084]

[32] Shen X, Zhang HF, Li LB, Zhang Z, Chen DB, Liu L. Clustering-driven deep adversarial hashing for scalable unsupervised cross-modal retrieval. Neurocomputing, 2021, 459: 152–164. [doi: 10.1016/j.neucom.2021.06.087]

[33] Chen SB, Wu S, Wang L, Yu ZY. Self-attention and adversary learning deep hashing network for cross-modal retrieval. Computers & Electrical Engineering, 2021, 93: 107262. [doi: 10.1016/j.compeleceng.2021.107262]

[34] Fang XZ, Liu ZH, Han N, Jiang L, Teng SH. Discrete matrix factorization hashing for cross-modal retrieval. Int'l Journal of Machine Learning and Cybernetics, 2021, 12(10): 3023–3036. [doi: 10.1007/s13042-021-01395-5]

[35] Li HX, Zhang C, Jia XY, Gao Y, Chen CL. Adaptive label correlation based asymmetric discrete hashing for cross-modal retrieval. IEEE Trans. on Knowledge and Data Engineering, 2021. [doi: 10.1109/TKDE.2021.3102119]

[36] Fang YZ. Robust multimodal discrete hashing for cross-modal similarity search. Journal of Visual Communication and Image Representation, 2021, 79: 103256. [doi: 10.1016/j.jvcir.2021.103256]

[37] Huiskes MJ, Lew MS. The MIR Flickr retrieval evaluation. In: Proc. of the 1st ACM Int'l Conf. on Multimedia Information Retrieval. Vancouver: ACM, 2008. 39–43. [doi: 10.1145/1460096.1460104]

[38] Chua TS, Tang JH, Hong R, Li HJ, Luo ZP, Zheng YT. NUS-WIDE: A real-world web image database from National University of Singapore. In: Proc. of the 2009 ACM Int'l Conf. on Image and Video Retrieval. Santorini: ACM, 2009. 48. [doi: 10.1145/1646396.1646452]

[39] Fang XZ, Jiang KH, Han N, Teng SH, Zhou GX, Xie SL. Average approximate hashing-based double projections learning for cross-modal retrieval. IEEE Trans. on Cybernetics, 2022, 52(11): 11780–11793. [doi: 10.1109/TCYB.2021.3081615]

[40] Chen SB, Wu S, Wang L. Hierarchical semantic interaction-based deep hashing network for cross-modal retrieval. PeerJ Computer

Science, 2021, 7: e552. [doi: 10.7717/peerj-cs.552]
[41] Zhang DL, Wu XJ, Yu J. Learning latent hash codes with discriminative structure preserving for cross-modal retrieval. Pattern Analysis and Applications, 2021, 24(1): 283–297. [doi: 10.1007/s10044-020-00893-6]
[42] Liu X, Wang XZ, Cheung YM. FDDH: Fast discriminative discrete hashing for large-scale cross-modal retrieval. IEEE Trans. on Neural Networks and Learning Systems, 2022, 33(11): 6306–6320. [doi: 10.1109/TNNLS.2021.3076684]
[43] Zou XT, Wang XZ, Bakker EM, Wu S. Multi-label semantics preserving based deep cross-modal hashing. Signal Processing:Image Communication, 2021, 93: 116131. [doi: 10.1016/j.image.2020.116131]
[44] Zhu LP, Tian GY, Wang BY, Wang WJ, Zhang D, Li CY. Multi-attention based semantic deep hashing for cross-modal retrieval. Applied Intelligence, 2021, 51(8): 5927–5939. [doi: 10.1007/s10489-020-02137-w]

附中文参考文献:

[1] 李志欣, 魏海洋, 张灿龙, 马慧芳, 史忠植. 图像描述生成研究进展. 计算机研究与发展, 2021, 58(9): 1951–1974. [doi: 10.7544/issn1000-1239.2021.20200281]
[8] 李志欣, 凌锋, 张灿龙, 马慧芳. 融合两级相似度的跨媒体图像文本检索. 电子学报, 2021, 49(2): 268–274. [doi: 10.12263/DZXB.20191037]
[13] 李志欣, 凌锋, 唐振军, 马慧芳, 施智平. 基于多头注意力网络的无监督跨媒体哈希检索. 中国科学: 信息科学, 2021, 51(12): 2053–2368. [doi: 10.1360/SSI-2020-0264]
[18] 李志欣, 侯传文, 谢秀敏. 利用多重相似度矩阵增强跨模态哈希检索. 计算机辅助设计与图形学学报, 2022, 34(6): 933–945. [doi: 10.3724/SP.J.1089.2022.19044]

李志欣(1971－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为计算机视觉, 机器学习, 跨媒体计算.

谢秀敏(1997－), 女, 硕士, 主要研究领域为跨媒体检索, 机器学习.

侯传文(1996－), 男, 硕士, 主要研究领域为跨媒体检索, 机器学习.