基于迁移知识的跨模态双重哈希

钟建奇，林秋斌，曹文明*
（深圳大学电子与信息工程学院，广东深圳 518060）

摘 要： 随着社交网络的普及和多媒体数据的急剧增长，有效的跨模态检索引起了人们越来越多的关注．由于哈希有效的检索效率和低存储成本，其被广泛用于跨模态检索任务中．然而，这些基于深度学习的跨模态哈希检索方法大多数是利用图像网络和文本网络各自生成对应模态的哈希码，难以获得更加有效的哈希码，无法进一步减小不同模态数据之间的模态鸿沟．为了更好地提高跨模态哈希检索的性能，本文提出了一种基于迁移知识的跨模态双重哈希（Cross-modal Dual Hashing based on Transfer Knowledge，CDHTK）．CDHTK 通过结合图像网络、知识迁移网络以及文本网络进行跨模态哈希检索任务．对于图像模态，CDHTK 融合图像网络和知识迁移网络各自生成的哈希码，进而生成具有判别性的图像哈希码；对于文本模态，CDHTK 融合文本网络和知识迁移网络各自生成的哈希码，从而生成有效的文本哈希码．CDHTK 通过采用预测标签的交叉熵损失、生成哈希码的联合三元组量化损失以及迁移知识的差分损失来共同优化哈希码的生成过程，从而提高模型的检索效果，在 2 个常用的数据集（IAPR TC-12，MIR-Flickr 25K）上进行的实验验证了 CDHTK 的有效性，比当前最先进的跨模态哈希方法（Adaptive Label correlation based asymm Etric Cross-modal Hashing，ALECH）分别高出 6.82% 和 5.13%．

关键词： 跨模态；图像-文本检索；双重哈希；迁移知识
基金项目： 国家自然科学基金（No.617714322）；深圳市基础研究基金（No.JCYJ20220531100814033）
中图分类号： TP391 **文献标识码：** A **文章编号：** 0372-2112(2025)01-0209-12
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20240032

Cross-Modal Dual Hashing Based on Transfer Knowledge

ZHONG Jian-qi, LIN Qiu-bin, CAO Wen-ming*
(*School of Electronic and Information Engineering*, *Shenzhen University*, *Shenzhen*, *Guangdong* 518060, *China*)

Abstract: With the popularity of social networks and the rapid growth of multimedia data, efficient cross-modal retrieval has attracted more and more attention. Hashing is widely used in cross-modal retrieval tasks due to its high retrieval efficiency and low storage cost. However, most of these deep learning-based cross-modal hashing retrieval methods utilize image networks and text networks to respectively generate corresponding modal hash codes, making it difficult to obtain more efficient hash codes and unable to further reduce the modal gap between different modal data. To better improve the performance of cross-modal hashing retrieval, this paper proposes a cross-modal dual hashing based on transfer knowledge (CDHTK). CDHTK performs cross-modal hashing retrieval tasks by combining an image network, a transfer knowledge network, and a text network. For the image modality, CDHTK combines the hash codes generated separately by the image network and the knowledge transfer network to generate discriminative hash codes. For the text modality, CDHTK fuses the hash codes generated separately by the text network and the knowledge transfer network to generate efficient hash codes. CDHTK employs a combination of cross-entropy loss for label prediction, joint triplet quantization loss for hash code generation, and differential loss for transfer knowledge to jointly optimize the hash code generation process, thereby improving the retrieval performance of the model. Experiments on two commonly used data sets (IAPR TC-12, MIR-Flickr 25K) verified the effectiveness of CDHTK, which outperforms the current state-of-the-art cross-modal hashing method ALECH (Adaptive Label correlation based asymmEtric Cross-modal Hashing) by 6.82% and 5.13%, respectively.

Key words: cross-modal; image-text retrieval; dual hashing; transfer knowledge

收稿日期：2024-01-08；修回日期：2024-04-19；责任编辑：王天慧
*通讯作者：曹文明

Foundation Item(s): National Natural Science Foundation of China (No.617714322);Fundamental Research Foundation of Shenzhen (No.JCYJ20220531100814033)

1 引言

随着各种多媒体数据(例如文本、图像、音频以及视频)的爆炸式增长,高效的跨模态检索引起了人们越来越多的关注. 跨模态检索[1]利用一个模态数据(例如文本)来检索另一模态中语义相关的实例(例如图像). 然而,各种模态的不同分布和表征方式导致多模态数据之间存在异构性差异,这使得有效的跨模态检索非常具有挑战性.

为了统一不同模态的表征方式,并缩小各种模态之间的的语义鸿沟,现有的检索方法可以分为两类. 第一类是基于实值表示的方法. 第二类是基于哈希表示的方法. 研究人员提出了许多基于实值的表征方法,包括子空间学习[2,3]、主题模型[4,5]和深度模型[6,7]. 然而,计算的复杂性和检索的低效性是基于实值表示方法的问题. 为了降低计算的复杂性并提高检索的效率,跨模态检索任务经常使用基于哈希表示的方法[8]. 哈希方法将高维的特征向量转换为汉明空间中低维的紧凑哈希码. 在汉明空间中,一个实例与其语义上相似实例之间的汉明距离比与其语义上不相似实例之间的汉明距离更小.

根据是否使用标签信息,传统的跨模态哈希方法可以分为无监督方法和监督方法. 无监督的跨模态哈希方法利用原始的特征来探索多模态数据的分布,如IMH(Inter-Media Hashing)[9]、CMFH(Collective Matrix Factorization Hashing)[10]、FedUCH(Federated Unsupervised Cross-modal Hashing)[11]. 由于无监督方法仅利用共同出现的信息来学习跨模态数据的哈希函数,因此学习到的哈希码的质量较差.

与无监督方法相比,有监督的跨模态哈希方法利用语义标签或语义相关信息来提取跨模态的相关性以产生可区分的哈希码,例如SEPH(SEmantic Preserving Hash)[12]、LDCLA(a Linear Discriminative Cross-modal hashing Learning Algorithm)[13]、SCM(Semantic Correlation Maximization)[14]、Ranking-based Supervised Discrete Cross-modal Hashing(RSDCH)[15]. 然而,大多数传统的跨模态哈希方法都是基于人为设计的特征,它们的缺点是哈希函数学习过程与特征学习过程无关,即这两个过程可能无法最佳地互相协调.

随着近期深度学习的重大发展,深度卷积神经网络[16~18]在许多计算机视觉任务中取得了巨大的成功. 基于深度学习的跨模态检索方法构建了一个端到端的架构,可以同时学习特征和哈希码. 此外,基于深度学习的跨模态哈希方法的检索效果优于传统哈希方法的检索效果,例如DCMH(Deep Cross-Modal Hashing)[19]、PRDH(Pairwise Relationship guided Deep Hashing)[20]、SSAH(Self-Supervised Adversarial Hashing)[21]、AGAH(Adversary Guided Asymmetric Hashing)[22]、SDCH(Semantic Deep Cross-modal Hashing)[23]、MDCH(Mask Deep Cross-modal Hashing)[24]、TEACH(aTtEntion-Aware deep Cross-modal Hashing)[25]、$DDCH_{ms}$(Deep Discrete Cross-modal Hashing with multiple supervision)[26]、LtCMH(Long-tail Cross-Modal Hashing)[27]、ALECH(Adaptive Label correlation based asymmEtric Cross-modal Hashing)[28]. 然而,这些基于深度学习的跨模态哈希检索方法大多数是利用图像网络和文本网络各自生成对应模态的哈希码[19~25],比如,DCMH[19]通过将特征学习和哈希码学习集成到同一框架中,利用图像网络和文本网络各自生成对应模态的哈希码,提出了深度跨模态哈希. PRDH[20]利用了来自模态之间和模态内部的两种类型的成对检索图像网络和文本网络各自损失. SSAH[21]构建了一个自监督的语义网络,通过对抗学习来最大化不同模态的语义相关性. AGAH[22]利用注意力机制和对抗学习产生有效的哈希码. 由于对抗性学习的不稳定性和语义信息的不足,SSAH[21]和AGAH[22]降低了它们的检索性能. $DDCH_{ms}$[26]提出了一个端到端学习框架来学习离散哈希码,通过引入类相似度学习的类级哈希码与标签矩阵来探索实例(图像、文本)-标签和类信息的内在相关性,从而加强图像与文本不同模态之间的学习.

因此,仅仅利用图像网络和文本网络各自生成图像哈希码和文本哈希码的这些基于深度学习的跨模态哈希检索方法难以获得更加有效的哈希码,无法进一步减小不同模态数据之间的模态鸿沟. 而融合图像和文本共同优化哈希码的难点在于:(1)语义差异的处理,图像和文本之间存在着语义上的差异,如何将它们融合到一个共同的特征空间中是一个挑战;(2)特征空间的不一致性,不同模态的特征空间可能具有不同的分布特性和度量方式,如何统一这些特征空间以获得更一致的哈希码表示是一个挑战. 一些工作[27,28]尝试解决以上问题,比如,针对第一个问题,LtCMH[27]利用两种模态数据的特征与个体性/共性语义特征融合后共同生成一个统一的哈希码. 针对上述第二个问题,ALECH[28]充分利用了标签信息和成对语义相似性,学习一个哈希函数映射矩阵 $\boldsymbol{W}$,将该模态的原始特征映射到之前学习的公共哈希码 $\boldsymbol{B}$,确保了不同模态的哈希码在海明空间中是相似的. 与这些方法不同,本文提出以知识迁移学习为基础,利用知识迁移模型泛化能力强、目标领域数据依赖程度底等特点,深度挖掘融合图

像和文本共同优化哈希码的作用. 针对上述第一个问题,本文提出在利用图像网络和文本网络各自生成哈希码能力的基础上,引入知识迁移网络,将跨模态信息融合到哈希码的生成过程中,学习和利用跨模态信息的共性和相关性,以缓解不同模态之间的语义差异,从而提高哈希码的一致性和可区分性,从而获得更有效的哈希码. 针对第二个问题,本文提出多损失函数的综合优化策略,采用了预测标签的交叉熵损失、联合三元组量化损失和迁移知识的差分损失等多个损失函数进行哈希码的优化. 这些损失函数共同作用下,可以在不同层面上对哈希码生成过程进行调节和优化,从而提高哈希码的表征能力和检索效果.

综合以上设计,本文提出了一种基于迁移知识的跨模态双重哈希(Cross-modal Dual Hashing based on Transfer Knowledge, CDHTK),其创新之处主要体现在三个方面:

(1)构造了一个跨模态图像-文本哈希检索模型,结合图像网络、知识迁移网络以及文本网络完成跨模态哈希检索任务;

(2)针对图像模态,融合图像网络和知识迁移网络各自生成的哈希码,进而生成具有判别性的图像哈希码;针对文本模态,融合文本网络和知识迁移网络各自生成的哈希码,从而生成有效的文本哈希码;

(3)通过采用预测标签的交叉熵损失、生成哈希码的联合三元组量化损失以及迁移知识的差分损失来共同优化哈希码的生成过程,从而提高模型的检索效果.

2　相关工作

2.1　传统跨模态哈希方法

跨模态哈希因其在大规模信息检索中具有有效性和低存储成本的特点而备受关注. 传统跨模态哈希方法主要依赖于手工设计的特征,例如,集体矩阵分解哈希(Collective Matrix Factorization Hashing, CMFH)[10]利用集体矩阵分解技术从手工制作的特征中导出跨视图哈希函数,从而通过融合多个视图信息源来提高搜索准确性. 潜在语义稀疏散列(Latent Semantic Sparse Hashing, LSSH)[29]集成了稀疏编码和矩阵分解来融合各种潜在语义表示,产生有区别的二进制代码. 此外,LSSH 采用迭代方法来探索图像和文本表示之间的互相关性. 判别式跨模态哈希(Deep Cross-modal Hashing, DCH)[30]开发了一种有效的离散优化算法来共同学习模态特定的哈希函数和统一的二进制代码. 通过直接学习有区别的二进制代码,同时保留离散约束,从而提高检索准确性. 交替共量化(Alternating Co-Quantization, ACQ)[31]为一种模态的数据与其他模态的有用连接生成二进制量化器,从而帮助提高哈希检索性能. 语义保留哈希(probability-based Semantics-Preserving Hashing, SePH)[32],根据训练数据的给定语义亲和力生成统一哈希码,然后利用核逻辑回归学习从特征到哈希码的非线性投影. 语义相关性最大化(Semantic Correlation Maximization, SCM)[14]将语义标签无缝整合到哈希学习过程中,具有高度的可扩展性和有效性. 尽管这些方法在弥合异构差距和增强跨模态检索方面表现出色,但在不考虑特征学习和哈希编码之间的反馈循环时,它们的性能往往无法令人满意.

2.2　基于深度学习的跨模态哈希方法

深度学习被广泛用来缩小不同模态之间的异构鸿沟. 因具有强大的表示能力和反馈机制,基于深度学习的跨模态哈希方法比传统跨模态哈希方法表现出更好的检索性能. DCMH[19]在深度的端到端框架中学习特征和哈希码. 然而,在 DCMH[19]中仅使用了跨模态信息,因此其检索性能稍差. PRDH[20]利用了来自模态之间和模态内部的两种类型的成对检索损失. 虽然 PRDH[20]使用了模态内部的检索信息,但它无法探索图像模态和文本模态之间的底层语义信息. SSAH[21]构建了一个自监督的语义网络,通过对抗学习来最大化不同模态的语义相关性. AGAH[22]利用注意力机制和对抗学习产生有效的哈希码. 由于对抗性学习的不稳定性和语义信息的不足,SSAH[21]和 AGAH[22]降低了它们的检索性能. SDCH[23]通过构建哈希码分支和语义分类分支学习富含语义信息的哈希码. MDCH[24]和 TEACH[25]利用掩码信息增加图像特征的语义信息,从而增强哈希码的区分能力. $DDCH_{ms}$[26]提出了一个端到端学习框架来学习离散哈希码,通过引入类相似度学习的类级哈希码与标签矩阵来探索实例(图像、文本)-标签和类信息的内在相关性,从而加强图像与文本不同模态之间的学习. 上述基于深度学习的跨模态哈希检索方法大多数是利用图像网络和文本网络各自生成对应模态的哈希码. 一些工作[27,28]开始尝试融合图像和文本信息共同优化哈希码,比如,LtCMH[27]提出 LtCMH(长尾跨模态哈希)来处理不平衡的多模态数据,利用两种模态数据的特征与个体性/共性特征融合后共同生成一个统一的哈希码. ALECH[28]提出一种基于自适应标签相关性的非对称跨模态哈希方法,尝试从标签具有多层语义这一特性着手,充分挖掘多标签语义的标签信息和潜在相关性.

然而,如何进一步提高哈希码的区分性,提高检索性能,仍然是一个具有挑战性的工作. 本文通过构建一个知识迁移网络额外生成对应模态的辅助哈希码,并与原来的图像网络或文本网络产生的哈希码进行融合,从而生成更具有判别性的哈希码.

3 基于迁移知识的双重哈希

3.1 符号定义

小写粗体字母表示向量，如 $\boldsymbol{a}$. 大写粗体字母表示矩阵，如 $\boldsymbol{A}$. 矩阵 $\boldsymbol{A}$ 的第 i 行和第 j 列的元素记为 $\boldsymbol{A}_{ij}$. $\boldsymbol{A}_{i*}$ 表示矩阵 $\boldsymbol{A}$ 第 i 行的元素. $\boldsymbol{A}_{*j}$ 表示矩阵 $\boldsymbol{A}$ 第 j 列的元素. $\|\boldsymbol{A}\|_{\mathrm{F}}$ 表示矩阵 $\boldsymbol{A}$ 的 Frobenius 范数. sign() 是符号函数，其表述如下：

$$\operatorname{sign}(a)=\begin{cases}1, & a\geqslant 0\\ -1, & a<0\end{cases} \tag{1}$$

3.2 问题定义

假设跨模态数据集具有 n 对训练实例对，每对训练实例对都具有一个图像模态实例 $\boldsymbol{P}_{i*}$ 和一个文本模态实例 $\boldsymbol{Q}_{i*}$. $\boldsymbol{L}\subset\{0,1\}_{n\times c}$ 表示训练集的标签，其中 c 是类别的数量. 此外，利用跨模态相似矩阵 $\boldsymbol{S}$ 来测量两个模态实例之间的相似性，其中 $S_{ij}=1$ 表示图像 $\boldsymbol{P}_{i*}$ 与文本 $\boldsymbol{Q}_{j*}$ 至少有一个共同的标签，否则 $S_{ij}=0$.

跨模态哈希的目的是将图像和文本转换为统一的哈希码：$\boldsymbol{B}^{(p,q)}\subset\{-1,+1\}_{n\times h}$，其中 h 是哈希码的长度. 为了测量两个不同模态实例的哈希码之间的相似性，汉明距离 $D(\boldsymbol{B}_{i*}^{(p)},\boldsymbol{B}_{j*}^{(q)})$ 表示如下：

$$D\left(\boldsymbol{B}_{i*}^{(p)},\boldsymbol{B}_{j*}^{(q)}\right)=\frac{1}{2}\left(h-<\boldsymbol{B}_{i*}^{(p)},\boldsymbol{B}_{j*}^{(q)}>\right) \tag{2}$$

其中，$<\boldsymbol{B}_{i*}^{(p)},\boldsymbol{B}_{j*}^{(q)}>$ 表示两个哈希码 $\boldsymbol{B}_{i*}^{(p)}$ 和 $\boldsymbol{B}_{j*}^{(q)}$ 之间的内积. 在实验中，使用两个哈希码之间的余弦相似性替换它们的内积，表示如下：

$$D\left(B_{i*}^{(p)},B_{j*}^{(q)}\right)=\frac{h}{2}\left(1-\cos\left(B_{i*}^{(p)},B_{j*}^{(q)}\right)\right) \tag{3}$$

其中，$\cos\left(\boldsymbol{B}_{i*}^{(p)},\boldsymbol{B}_{j*}^{(q)}\right)=\dfrac{<\boldsymbol{B}_{i*}^{(p)},\boldsymbol{B}_{j*}^{(q)}>}{\left\|\boldsymbol{B}_{i*}^{(p)}\right\|\cdot\left\|\boldsymbol{B}_{j*}^{(q)}\right\|}$.

3.3 哈希码学习

如图 1 所示，本文提出的 CDHTK 模型包括三个网络，分别是图像网络、知识迁移网络以及文本网络. 接下来，将逐一介绍各个网络.

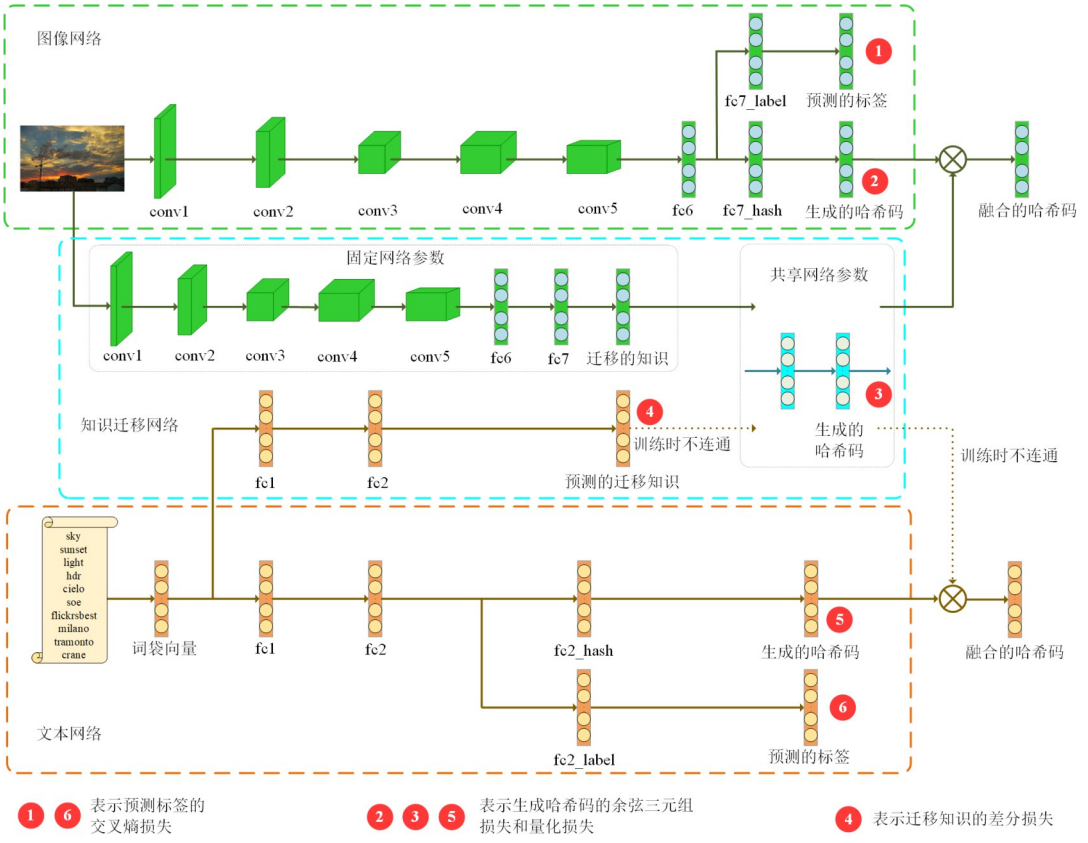


图 1 基于迁移知识的跨模态双重哈希结构图

3.3.1 图像网络

针对图像网络，CDHTK 利用 CNN-F[12] 网络架构作为基础网络生成图像特征. 具体而言，去掉原始 CNN-F 网络的最后一层，并在 fc6 层之后，构建两个分支，分别生成预测的标签和哈希码. 具体结构如图 1 所示. 具体的网络参数如表 1 所示，其中卷积核的参数表示为“通道大小×卷积核的长×卷积核的宽”. 对于图像网络中的激活函数，CDHTK 分别使用 tanh(·) 和 sigmoid(·) 生成连续的哈希码和预测的标签. 其他网络层的激活函数使用 ReLU(·)[33].

表1　图像网络的网络层配置

网络层	配置
conv1	卷积核64×11×11
conv2	卷积核256×5×5
conv3	卷积核256×3×3
conv4	卷积核256×3×3
conv5	卷积核256×3×3
fc6	4 096
fc7_label	4 096
fc8_label	c,类别数目
fc7_hash	4 096
fc8_hash	h,哈希码长度

假设图像网络预测的标签是 $\hat{\boldsymbol{L}}_{i*}^{(p)}=g^{(p)}\left(\boldsymbol{P}_{i*};\boldsymbol{\theta}^{(p)},\boldsymbol{\theta}^{(p\_\text{label})}\right)$,其中,$\boldsymbol{\theta}^{(p_\text{label})}$是预测标签分支的参数.为了充分利用训练数据的标签信息,引导模型学习语义相关的特征表示,从而生成具有区分性的哈希码,CDHTK对预测的图像标签施加了交叉熵损失,基于标签的交叉熵损失可以约束图像网络生成能较好预测实例标签的特征表示,这种特征表示自然包含了语义相关的信息,从而为生成判别性哈希码奠定基础,即

$$L_1=-\frac{1}{nc}\sum_{i=1}^{n}\sum_{j=1}^{c}\left(\boldsymbol{L}_{ij}\log\hat{\boldsymbol{L}}_{ij}^{(p)}+\left(1-\boldsymbol{L}_{ij}\right)\log\left(1-\hat{\boldsymbol{L}}_{ij}^{(p)}\right)\right) \tag{4}$$

其中,n表示训练集的样本数量,c表示类别数目,$\hat{\boldsymbol{L}}_{ij}^{(p)}$是预测的图像标签.

假设图像网络生成的哈希码是 $\boldsymbol{H}_{i*}^{(p)}=f^{(p)}\left(\boldsymbol{P}_{i*};\boldsymbol{\theta}^{(p)},\boldsymbol{\theta}^{(p\_\text{hash})}\right)$,其中,$\boldsymbol{\theta}^{(p)}$是CNN-F网络的前6层参数,$\boldsymbol{\theta}^{(p_\text{hash})}$是生成哈希码分支的参数.另外,为了方便网络进行反向传播,对于网络生成的哈希码,除了特别声明之外,均指连续值,其范围在(–1, 1). 为了使最终生成的哈希码满足哈希检索所需的相似度量和二值约束,从而生成具有判别性的图像哈希码,CDHTK对生成的图像哈希码 $\boldsymbol{H}_{i*}^{(p)}$ 利用联合三元组量化损失进行优化.联合三元组量化损失包括余弦三元组损失以及量化损失,其中,余弦三元组损失要求语义相似实例对之间的哈希码相似度大于语义不相似实例对的哈希码相似度,直接体现了哈希检索的目标需求,量化损失则是为了使生成的哈希码在符号函数的约束下,尽可能接近于期望的二值编码,满足哈希存储和检索的需求.具体如下:

$$L_2=\sum_{i,j,k}\max\left(\cos\left(\boldsymbol{H}_{i*}^{(p)},\boldsymbol{H}_{j*}^{(q+)}\right)-\cos\left(\boldsymbol{H}_{i*}^{(p)},\boldsymbol{H}_{k*}^{(q-)}\right)+m,0\right)+\frac{1}{nh}\left\|\boldsymbol{B}-\boldsymbol{H}^{(p)}\right\|_{\text{F}}^{2} \tag{5}$$

其中,$\boldsymbol{H}_{j*}^{(q+)}$、$\boldsymbol{H}_{k*}^{(q-)}$表示文本网络生成的哈希码.而且 $\boldsymbol{H}_{j*}^{(q+)}$ 是 $\boldsymbol{H}_{i*}^{(p)}$ 的正样本,即 $\boldsymbol{H}_{i*}^{(p)}$ 与 $\boldsymbol{H}_{j*}^{(q+)}$ 至少共享一个标签.$\boldsymbol{H}_{k*}^{(q-)}$ 是 $\boldsymbol{H}_{i*}^{(p)}$ 的负样本,即 $\boldsymbol{H}_{i*}^{(p)}$ 与 $\boldsymbol{H}_{k*}^{(q-)}$ 没有共享标签.m 表示余弦三元组损失的边距.另外,$\boldsymbol{B}=\text{sign}(\alpha\boldsymbol{H}^{(p)}+(1-\alpha)\boldsymbol{H}^{(t_p)}+\boldsymbol{H}^{(q)})$,其中 $\boldsymbol{H}^{(p)}$ 指的是图像网络生成的图像哈希码,$\boldsymbol{H}^{(t_p)}$ 指的是图像的迁移网络生成的辅助图像哈希码,$\boldsymbol{H}^{(q)}$ 指的是文本网络生成的哈希码.值得注意的是,如未特殊说明,下文中的 $\boldsymbol{B}$ 表示的含义均与此处的意思一致.

综合图像网络的交叉熵损失、联合三元组量化损失,图像网络总的损失函数如下:

$$L_{\text{txt}}=a_1L_1+a_2L_2 \tag{6}$$

其中,a_1 和 a_2 为分别为图像网络交叉熵损失、联合三元组量化损失的权重参数.

3.3.2　文本网络

CDHTK利用文本的词袋向量(Bag of Word, BoW)输入文本网络.首先,CDHTK设计了两层的网络(BoW→8 192→4 096).之后,CDHTK构建两个分支,分别生成预测的标签和哈希码.具体结构如图1所示.具体的网络参数如表2所示.

表2　文本网络的网络层配置

网络层	配置
fc1	8 192
fc2	4 096
fc3_label	4 096
fc4_label	c,类别数目
fc3_hash	4 096
fc4_hash	h,哈希码长度

对于文本网络中的激活函数,CDHTK分别使用tanh(·)和sigmoid(·)生成连续的哈希码和预测的标签.其他网络层的激活函数使用ReLU(·).

假设文本网络预测的标签是 $\hat{\boldsymbol{L}}_{i*}^{(q)}=g^{(q)}\left(\boldsymbol{Q}_{i*};\boldsymbol{\theta}^{(q)},\boldsymbol{\theta}^{(q\_\text{label})}\right)$,其中 $\boldsymbol{\theta}^{(q\_\text{label})}$ 是预测标签分支的参数.为了充分有效地利用标签信息,促进后续生成区分性强的文本哈希码,与图像网络一致,CDHTK对预测的文本标签施加了交叉熵损失,即

$$L_3=-\frac{1}{nc}\sum_{i=1}^{n}\sum_{j=1}^{c}\left(\boldsymbol{L}_{ij}\log\hat{\boldsymbol{L}}_{ij}^{(q)}+\left(1-\boldsymbol{L}_{ij}\right)\log\left(1-\hat{\boldsymbol{L}}_{ij}^{(q)}\right)\right) \tag{7}$$

其中,n表示训练集的样本数量,c表示类别数目,$\hat{\boldsymbol{L}}_{ij}^{(q)}$是预测的文本标签.

假设文本网络生成的哈希码是 $\boldsymbol{H}_{i*}^{(q)}=f^{(q)}\left(\boldsymbol{Q}_{i*};\boldsymbol{\theta}^{(q)},\boldsymbol{\theta}^{(q\_\text{hash})}\right)$,其中 $\boldsymbol{\theta}^{(q)}$ 是文本网络的前两层参数,$\boldsymbol{\theta}^{(q\_\text{hash})}$ 是生成哈希码分支的参数.为了满足哈希编码的二值约束,并使得语义相似的文本实例对之间的哈希码相似度大于语义不相似的实例对,从而生成具有判别性的文本哈希码,使生成的哈希码可高效存储和检索,与图像网络一致,CDHTK对生成的文本哈希码

利用联合三元组量化损失，即

$$L_4=\sum_{i,j,k}\max\left(\cos\left(\boldsymbol{H}_{i*}^{(q)},\boldsymbol{H}_{j*}^{(p+)}\right)-\cos\left(\boldsymbol{H}_{i*}^{(q)},\boldsymbol{H}_{k*}^{(p-)}\right)+m,0\right)+\frac{1}{nh}\left\|\boldsymbol{B}-\boldsymbol{H}^{(q)}\right\|_{\mathrm{F}}^{2} \tag{8}$$

其中，$\boldsymbol{H}_{i*}^{(q)}$表示文本网络生成的哈希码，$\boldsymbol{H}_{j*}^{(p+)}$、$\boldsymbol{H}_{k*}^{(p-)}$表示图像网络生成的哈希码. 而且$\boldsymbol{H}_{j*}^{(p+)}$是$\boldsymbol{H}_{i*}^{(q)}$的正样本，即$\boldsymbol{H}_{i*}^{(q)}$与$\boldsymbol{H}_{j*}^{(p+)}$至少共享一个标签. $\boldsymbol{H}_{k*}^{(p-)}$是$\boldsymbol{H}_{i*}^{(q)}$的负样本，即$\boldsymbol{H}_{i*}^{(q)}$与$\boldsymbol{H}_{k*}^{(p-)}$没有共享标签. m表示余弦三元组损失的边距.

综合文本网络的交叉熵损失、联合三元组量化损失，文本网络总的损失函数如下：

$$L_{\mathrm{txt}}=a_3L_3+a_4L_4 \tag{9}$$

其中，a_3和a_4为分别为文本网络交叉熵损失、联合三元组量化损失的权重参数.

3.3.3 知识迁移网络

在图像的知识迁移网络中，CDHTK固定原始的CNN-F网络参数. 输入原始图像$\boldsymbol{P}_{i*}$，可以得到1 000维的迁移知识$\boldsymbol{T}_{i*}$. 之后，CDHTK将得到的迁移知识输入图像-文本共享网络（1 000→4 096→h），从而得到辅助的图像哈希码，具体结构如图1所示.

假设图像的知识迁移网络生成的辅助图像哈希码是$\boldsymbol{H}_{i*}^{(t_p)}=f^{(t_p)}\left(\boldsymbol{P}_{i*};\boldsymbol{\theta}^{(t_p)},\boldsymbol{\theta}^{(t_{p_\mathrm{hash}})}\right)$，文本的知识迁移网络生成的辅助文本哈希码是$\boldsymbol{H}_{i*}^{(t_q)}=f^{(t_q)}\left(\boldsymbol{Q}_{i*};\boldsymbol{\theta}^{(t_q)},\boldsymbol{\theta}^{(t_{p_\mathrm{hash}})}\right)$，其中，$\boldsymbol{\theta}^{(t_p)}$是原始的CNN-F网络参数，$\boldsymbol{\theta}^{(t_q)}$是文本预测迁移知识所需要的网络参数，$\boldsymbol{\theta}^{(t_{p_\mathrm{hash}})}$是图像-文本共享网络的参数.

由于图像特征的学习对最终的跨模态检索性能尤为关键，CDHTK更关注利用知识迁移网络增强图像特征的学习，而文本特征的优化可以依赖于文本网络本身的训练. 因此，为了更有效地促进图像特征和文本特征的对齐，从而学习到更优的跨模态表示，与图像网络一致，CDHTK利用联合三元组量化损失对生成的辅助图像哈希码进行优化，即

$$L_5=\max\left(\cos\left(\boldsymbol{H}_{i*}^{(t_p)},\boldsymbol{H}_{j*}^{(t_p+)}\right)-\cos\left(\boldsymbol{H}_{i*}^{(t_p)},\boldsymbol{H}_{k*}^{(t_p-)}\right)+m,0\right)+\frac{1}{nh}\left\|\boldsymbol{B}-\boldsymbol{H}^{(t_p)}\right\|_{\mathrm{F}}^{2} \tag{10}$$

其中，$\boldsymbol{H}_{j*}^{(t_p+)}$、$\boldsymbol{H}_{k*}^{(t_p-)}$表示知识迁移网络生成的辅助图像哈希码，$\boldsymbol{H}_{j*}^{(t_p+)}$是$\boldsymbol{H}_{i*}^{(t_p)}$的正样本，即$\boldsymbol{H}_{i*}^{(t_p)}$与$\boldsymbol{H}_{j*}^{(t_p+)}$至少共享一个标签，$\boldsymbol{H}_{k*}^{(t_p-)}$是$\boldsymbol{H}_{i*}^{(t_p)}$的负样本，即$\boldsymbol{H}_{i*}^{(t_p)}$与$\boldsymbol{H}_{k*}^{(t_p-)}$没有共享标签；$m$表示余弦三元组损失的边距. 从式（10）中可以看出，文本的知识迁移网络生成的辅助文本哈希码并没有参与$\boldsymbol{\theta}^{(t_{p_\mathrm{hash}})}$的优化（图1中表示为“训练时不连通”状态）.

此外，在知识迁移网络中，输入文本$\boldsymbol{Q}_{i*}$，经过三层网络（BoW→8 192→4 096→1 000）生成预测的迁移知识$\hat{\boldsymbol{T}}_{i*}$. CDHTK利用迁移知识的差分损失拟合图像固定的迁移知识，即

$$L_6=\frac{1}{1000n}\sum_{i=1}^{n}\left\|\boldsymbol{T}_{i*}-\hat{\boldsymbol{T}}_{i*}\right\|^{2} \tag{11}$$

综合辅助哈希码的联合三元组量化损失，以及迁移知识的差分损失，知识迁移网络总的损失函数如下：

$$L_t=a_5L_5+a_6L_6 \tag{12}$$

其中，a_5和a_6为分别为知识迁移网络联合三元组量化损失、差分损失的权重参数.

结合图像网络、文本网络以及知识迁移网络的损失函数，CDHTK总的损失函数如下：

$$L=L_{\mathrm{img}}+L_{\mathrm{txt}}+L_t \tag{13}$$

其中，本文采用的交叉熵损失约束生成语义相关特征，为生成判别性哈希码奠定基础. 联合三元组量化损失则直接约束生成的哈希码满足哈希检索的需求. 三者共同构成整个端到端的深度模型的优化目标，在相互作用下，可以使所学习到的哈希码兼具语义判别性和检索实用性. 在训练过程中，CDHTK利用Adam优化器对该损失函数进行优化. 整体优化过程总结在算法1中.

3.4 训练集外样本的哈希码生成

在训练结束之后，对于测试集和被检索集中的样本点，当给定一个图像模态实例$\boldsymbol{p}$，CDHTK分别利用图像网络$f^{(p)}$和图像的知识迁移网络$f^{(t_p)}$生成它对应的图像哈希码$\boldsymbol{h}^{(p)}$和辅助图像哈希码$\boldsymbol{h}^{(t_p)}$，再按照一定的比例进行融合，之后经过符号函数sign(·)生成最后离散的图像哈希码$\boldsymbol{b}^{(p)}$. 上述过程可按照如下式进行表示：

$$\boldsymbol{h}^{(p)}=f^{(p)}\left(\boldsymbol{p};\boldsymbol{\theta}^{(p)},\boldsymbol{\theta}^{(p\_\mathrm{hash})}\right) \tag{14}$$

$$\boldsymbol{h}^{(t_p)}=f^{(t_p)}\left(\boldsymbol{p};\boldsymbol{\theta}^{(t_p)},\boldsymbol{\theta}^{(t_{p\_\mathrm{hash}})}\right) \tag{15}$$

$$\boldsymbol{b}^{(p)}=\mathrm{sign}\left(\alpha\boldsymbol{h}^{(p)}+(1-\alpha)\boldsymbol{h}^{(t_p)}\right) \tag{16}$$

当给定一个文本模态实例$\boldsymbol{q}$，CDHTK分别利用文本网络$f^{(q)}$和文本的知识迁移网络$f^{(t_q)}$生成它对应的文本哈希码$\boldsymbol{h}^{(q)}$和辅助文本哈希码$\boldsymbol{h}^{(t_q)}$，再按照一定的比例进行融合，之后经过符号函数sign(·)生成最后离散的文本哈希码$\boldsymbol{b}^{(q)}$. 上述过程可按照如下公式进行表示：

$$\boldsymbol{h}^{(q)}=f^{(q)}\left(\boldsymbol{q};\boldsymbol{\theta}^{(q)},\boldsymbol{\theta}^{(q\_\mathrm{hash})}\right) \tag{17}$$

$$\boldsymbol{h}^{(t_q)}=f^{(t_q)}\left(\boldsymbol{q};\boldsymbol{\theta}^{(t_q)},\boldsymbol{\theta}^{(t_{q\_\mathrm{hash}})}\right) \tag{18}$$

$$\boldsymbol{b}^{(q)}=\mathrm{sign}\left(\alpha\boldsymbol{h}^{(q)}+(1-\alpha)\boldsymbol{h}^{(t_q)}\right) \tag{19}$$

4 实验

4.1 数据集

IAPR TC-12数据集[34]由20 000对图像-文本实例

算法1　基于迁移知识的跨模态双重哈希

输入:参与训练的图像数据$\boldsymbol{P}$,文本数据$\boldsymbol{Q}$,标签数据$\boldsymbol{L}$,预定义哈希长度h,学习率l_r.

输出:文本和图像网络参数$\boldsymbol{\theta}^{(p_\text{label})}$,$\boldsymbol{\theta}^{p}$,$\boldsymbol{\theta}^{(p_\text{hash})}$,$\boldsymbol{\theta}^{q_\text{label}}$,$\boldsymbol{\theta}^{q}$,$\boldsymbol{\theta}^{q_\text{hash}}$,$\boldsymbol{\theta}^{(t_p)}$,$\boldsymbol{\theta}^{(t_q)}$和$\boldsymbol{\theta}^{(t_p_\text{hash})}$生成的哈希码$\boldsymbol{B}_p\boldsymbol{B}_q$.

初始化:网络参数$\boldsymbol{\theta}^{(p_\text{label})}$,$\boldsymbol{\theta}^{p}$,$\boldsymbol{\theta}^{(p_\text{hash})}$,$\boldsymbol{\theta}^{q_\text{label}}$,$\boldsymbol{\theta}^{q}$,$\boldsymbol{\theta}^{q_\text{hash}}$,$\boldsymbol{\theta}^{(t_p)}$,$\boldsymbol{\theta}^{(t_q)}$,设置训练批次$n$=64,迭代次数epoch=50

REPEAT

REPEAT

从图像数据中$\boldsymbol{P}$随机选取n个样本参与训练;

根据式(6)计算图像网络的梯度;

采用反向传播对网络参数$\boldsymbol{\theta}^{(p_\text{label})}$,$\boldsymbol{\theta}^{p}$,$\boldsymbol{\theta}^{(p_\text{hash})}$进行更新

UNTIL 达到设定迭代次数

REPEAT

REPEAT

从文本数据中$\boldsymbol{Q}$随机选取n个样本参与训练;

根据式(9)计算文本网络的梯度;

采用反向传播对网络参数$\boldsymbol{\theta}^{(t_q)}$,$\boldsymbol{\theta}^{(t_q)}$,$\boldsymbol{\theta}^{(t_p_\text{hash})}$进行更新

UNTIL 达到设定迭代次数

REPEAT

REPEAT

从图像数据中$\boldsymbol{P}$和文本数据$\boldsymbol{Q}$中随机选取n个样本参与训练;

根据式(12)计算知识迁移网络的梯度;

采用反向传播对网络参数$\boldsymbol{\theta}^{(t_q)}$,$\boldsymbol{\theta}^{(t_q)}$,$\boldsymbol{\theta}^{(t_p_\text{hash})}$进行更新

UNTIL 达到设定迭代次数

UNTIL 完成迭代

对组成,并用255个标签进行标注.每个文本实例都用2 912维的词袋向量进行表示.

MIR-Flickr 25K数据集[35]是从Flickr网站收集的25 000张图像组成.每张图像都标有其关联的文本解释.根据DCMH的实验装置,在本文的实验中选择了20 015对图像-文本实例.每个文本实例用1 386维词袋向量进行表示.此外,每对图像-文本实例都用24个标签中的至少一个进行标注.

NUS-WIDE数据集[36]由269 648张网络图片组成,每张图片用81个标签进行标注.本实验使用属于21个最常见类别的图像-文本对,共有195 834对.此外,每个文本实例用1 000维的词袋向量进行表示.

对于IAPR TC-12和MIR-Flickr 25K数据集,随机选择2 000对图像-文本实例,并将其剩余的实例对作为被检索集.此外,从被检索集中选择10 000对图像-文本作为训练集.

对于NUS-WIDE数据集,随机选择2 100对图像-文本作为查询集,剩余的实例对作为被检索集.此外,从被检索集中随机选择10 500对作为训练集.

4.2　评价指标

本实验使用两种常用的检索评价指标以评估跨模态哈希检索的性能,即汉明排序和哈希查找.汉明排序根据数据库中的实例与给定查询实例之间的距离按升序重新对数据库中的实例进行排序.平均精度的均值(Mean Average Precision,MAP)通常用于评估汉明排序.哈希查找指返回与查询实例在某个汉明距离内的所有实例,这可以通过精确率-召回率(Precision-Recall,PR)曲线来度量.

4.3　实验细节

对于图像网络,其Adam优化器的初始学习率是10^{-5};对于文本网络和知识迁移网络,其Adam优化器的初始学习率均是10^{-4}.一共训练50次,每5次,优化器的学习率变为原来的0.9倍.每次训练的样本批次是32.式(5)、式(8)、式(10)中余弦三元组损失的边距m取值为0.3.式(16)、式(19)中的比例系数α取值为0.6.CDHTK是基于PyTorch框架实现功能.实验平台是带有8块2080Ti的GPU服务器.对于实验结果,本文呈现的是5次平均.

4.4　本文算法与其他优秀算法的性能比较与分析

表3展示了本文所提出的CDHTK方法与其他对比方法在三个常用数据集上的MAP结果.各方法均采用CNN-F网络作为图像网络的基础网络.对于表3中其他对比方法的MAP结果,本文尽可能地引用它们原文中的数值.本文重新复现了SDCH和MDCH方法,并呈现了5次MAP的平均值.

对于IAPR TC-12数据集,CDHTK方法的MAP值比SDCH、MDCH、ALECH方法的MAP值整体平均分别高3.95%、4.74%、6.82%;对于MIR-Flickr 25K数据集,CDHTK方法的MAP值比SDCH、MDCH、ALECH方法的MAP值整体平均分别高1.72%、2.48%、5.13%;对于NUS-WIDE数据集,CDHTK方法的MAP值比SDCH、MDCH方法的MAP值整体平均分别高4.93%、5.08%.这些指标展示了本文提出的CDHTK方法的优越性能.同时,我们也注意到,在NUS-WIDE数据集上,CDHTK方法在图像检索文本任务上的MAP值平均比最新算法ALECH方法高0.27%,但在文本检索图像任务上,ALECH方法效果优于CDHTK方法.这可能是因为ALECH方法提出的不对称哈希码生成策略造成的,这种策略可能更适合处理跨模态的异构检索任务,导致在图像到文本和文本到图像的检索任务上效果存在一定的差异性.

图2展示了CDHTK方法与其他方法在三个常用数据集上的精确率-召回率曲线.哈希码长度是16,各方法均使用CNN-F网络作为它们图像网络的基础网络.对于MIR-Flickr 25K数据集,CDHTK的PR曲线位于最

表3 汉明排序的MAP结果

任务	方法	出版处	IAPR TC-12			MIR-Flickr 25K			NUS-WIDE		
			16 bit	32 bit	64 bit	16 bit	32 bit	64 bit	16 bit	32 bit	64 bit
图像检索文本	DCMH[19]	CVPR 2017	0.452 6	0.473 2	0.484 4	0.741 0	0.746 5	0.748 5	0.590 3	0.603 1	0.609 3
	SSAH[21]	CVPR 2018	—	—	—	0.780 0	0.785 0	0.800 0	0.628 0	0.633 0	0.639 0
	AGAH[22]	ICMR 2019	—	—	—	0.792 0	0.794 0	0.806 0	0.645 0	0.660 0	0.651 0
	SDCH[23]	Neurocomputing 2020	0.534 0	0.565 8	0.587 4	0.797 3	0.811 5	0.822 4	0.644 6	0.660 4	0.674 4
	MDCH[24]	TMM 2021	0.530 3	0.560 2	0.581 4	0.785 1	0.799 6	0.809 7	0.635 7	0.658 5	0.664 4
	TEACH[25]	ICMR 2021	—	—	—	0.784 0	0.794 0	0.801 0	0.651 0	0.664 0	0.670 0
	DDCHms[26]	Neurocomputing 2022	0.463 2	0.489 1	0.512 8	0.739 4	0.745	0.757 5	0.632 1	0.649 5	0.666 2
	LtCMH[27]	AAAI 2023				0.745 0	0.753 0	0.781 0	0.635 0	0.654 0	0.678 0
	ALECH[28]	TKDE 2023	0.479 8	0.507 6	0.521 2	0.741 4	0.749 2	0.749 3	0.660 4	0.673 4	0.677 4
	CDHTK	本文	**0.556 4**	**0.579 3**	**0.592 0**	**0.816 9**	**0.827 8**	**0.833 0**	**0.659 7**	**0.678 1**	**0.678 9**
文本搜索图像	DCMH[19]	CVPR 2017	0.518 5	0.537 8	0.546 8	0.782 7	0.790 0	0.793 2	0.638 9	0.651 1	0.657 1
	SSAH[21]	CVPR 2018	—	—	—	0.783 0	0.791 0	0.803 0	0.654 0	0.647 0	0.666 0
	AGAH[22]	ICMR 2019	—	—	—	0.789 0	0.790 0	0.805 0	0.631 0	0.642 0	0.634 0
	SDCH[23]	Neurocomputing 2020	0.546 4	0.587 4	0.612 4	0.793 6	0.804 0	0.814 2	0.675 2	0.686 5	0.685 4
	MDCH[24]	TMM 2021	0.545 4	0.581 0	0.609 2	0.792 7	0.805 6	0.814 3	0.661 2	0.673 5	0.680 4
	TEACH[25]	ICMR 2021	—	—	—	0.787 0	0.798 0	0.804 0	0.673 0	0.687 0	0.689 0
	DDCHms[26]	Neurocomputing 2022	0.516 9	0.540 0	0.564 0	0.759 6	0.774 2	0.784 7	0.673 2	0.687 2	0.703 9
	LtCMH[27]	AAAI 2023				0.770 0	0.795 0	0.802 0	0.608 0	0.644 0	0.688 0
	ALECH[28]	TKDE 2023	0.570 2	0.616 9	0.645 6	0.807 6	0.816 9	0.821 5	**0.785 7**	**0.794 3**	0.804 6
	CDHTK	本文	**0.572 3**	**0.617 9**	**0.651 3**	**0.806 8**	**0.818 1**	**0.823 9**	0.685 5	0.723 5	**0.805 3**

上方,表明CDHTK在该数据集上的效果最好,该结论与使用MAP值进行比较的结果一致. 另外,针对IAPR TC-12和NUS-WIDE数据集,CDHTK的P1R曲线与坐标轴的正方向所围成的面积并不是最大的. 这可能是由于CDHTK对于某些查询实例,其与找到的样本之间的汉明距离小于5的比较少. 这种情况导致了CDHTK的MAP值虽然大于其他方法的MAP值,但是CDHTK的PR曲线与坐标轴所围成的面积并不是最大.

4.5 消融性实验

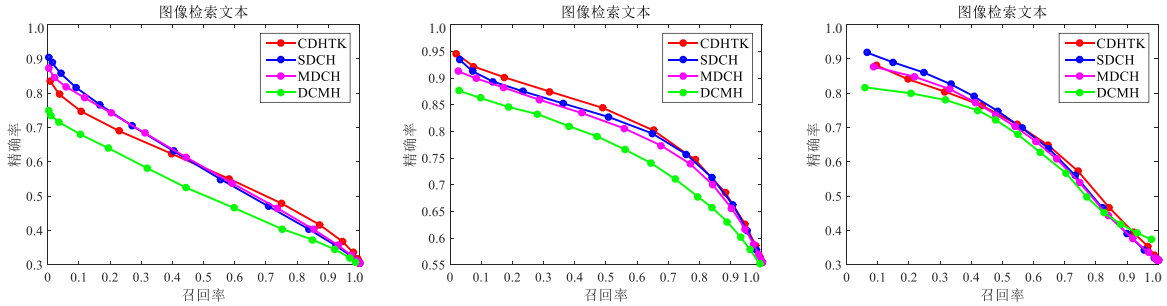
4.5.1 CDHTK方法消融实验

为了进一步分析本文提出的CDHTK方法的优越性,本节设计了CDHTK的几种变体并进行消融性实验. CDHTK-1没有使用知识迁移网络生成辅助的图像哈希码和文本哈希码. CDHTK-2没有利用知识迁移网络生成文本的辅助哈希码. CDHTK-3没有采用知识迁移网络生成图像的辅助哈希码.

表4~表6分别展示了CDHTK及其三种变体在三个常用数据集上的消融性实验. 从这三个表可以看出,如果没有使用知识迁移网络生成辅助的图像哈希码和文本哈希码,那么会导致最差的MAP结果. 如果去掉知识迁移网络文本的辅助哈希码或者图像的辅助哈希码,那么会导致较差的MAP结果. 综上所述,本文所构造的基于迁移知识的双重哈希策略对于提高跨模态哈希检索的性能是必不可少的.

表4 在IAPR TC-12数据集上的消融性实验

方法	实验策略	图像搜索文本			文本搜索图像			全体平均	同比提高
		16 bit	32 bit	64 bit	16 bit	32 bit	64 bit		
CDHTK-1	没有使用迁移知识生成的辅助哈希码	0.549 9	0.572 7	0.584 8	0.561 3	0.591 7	0.603 6	0.577 3	2.95%
CDHTK-2	没有使用文本的辅助哈希码	0.556 4	0.581 1	0.593 3	0.566 8	0.595 4	0.608 3	0.583 6	1.90%
CDHTK-3	没有使用图像的辅助哈希码	0.550 6	0.573 5	0.588 1	0.570 4	0.597 4	0.609 2	0.581 5	2.24%
CDHTK	—	**0.556 4**	**0.579 3**	**0.592 0**	**0.572 3**	**0.617 9**	**0.651 3**	**0.584 9**	—



(a) IAPR TC-12　(b) MIR-Flickr 25K　(c) NUS-WIDE

(d) IAPR TC-12　(e) MIR-Flickr 25K　(f) NUS-WIDE

图2　精确率-召回率曲线

表5　在MIR-Flickr 25K数据集上的消融性实验

方法	实验策略	图像搜索文本			文本搜索图像			全体平均	同比提高
		16 bit	32 bit	64 bit	16 bit	32 bit	64 bit		
CDHTK-1	没有使用迁移知识生成的辅助哈希码	0.812 6	0.823 3	0.828 3	0.800 4	0.812 5	0.818 0	0.815 8	0.64%
CDHTK-2	没有使用文本的辅助哈希码	0.817 4	0.827 0	0.832 7	0.801 2	0.812 7	0.818 2	0.818 2	0.35%
CDHTK-3	没有使用图像的辅助哈希码	0.810 4	0.821 2	0.827 3	0.805 6	0.816 3	0.821 7	0.817 1	0.49%
CDHTK	—	**0.816 9**	**0.827 8**	**0.833 0**	**0.806 8**	**0.818 1**	**0.823 9**	**0.821 1**	—

表6　在NUS-WIDE数据集上的消融性实验

方法	实验策略	图像搜索文本			文本搜索图像			全体平均	同比提高
		16 bit	32 bit	64 bit	16 bit	32 bit	64 bit		
CDHTK-1	没有使用迁移知识生成的辅助哈希码	0.650 2	0.666 1	0.673 4	0.673 8	0.711 7	0.791 8	0.694 5	1.51%
CDHTK-2	没有使用文本的辅助哈希码	0.657 5	0.670 7	0.676 8	0.675 7	0.712 1	0.791 7	0.697 4	1.10%
CDHTK-3	没有使用图像的辅助哈希码	0.654 2	0.669 7	0.676 4	0.677 8	0.714 6	0.794 0	0.697 8	1.05%
CDHTK	—	**0.659 7**	**0.678 1**	**0.678 9**	**0.685 5**	**0.723 5**	**0.805 3**	**0.705 2**	—

4.5.2　计算代价及模型大小分析

为了进一步分析本文提出的CDHTK方法,本文展示了所提CDHTK的训练时间成本以及模型大小. 为此,本文在三个常用数据集上记录了CDHTK的几种变体的训练时间. 表7只展示了哈希码长度为16的情况,而其他哈希码长度的训练成本是相似的. 实验结果如表7所示. 从图中可以看到,虽然引入迁移网络导致CDHTK的模型大小相比于CDHTK-1、CDHTK-2、CDHTK-3有所增长,但训练时间成本增长在可接受范围. 最重要的是,CDHTK模型性能有所提升.

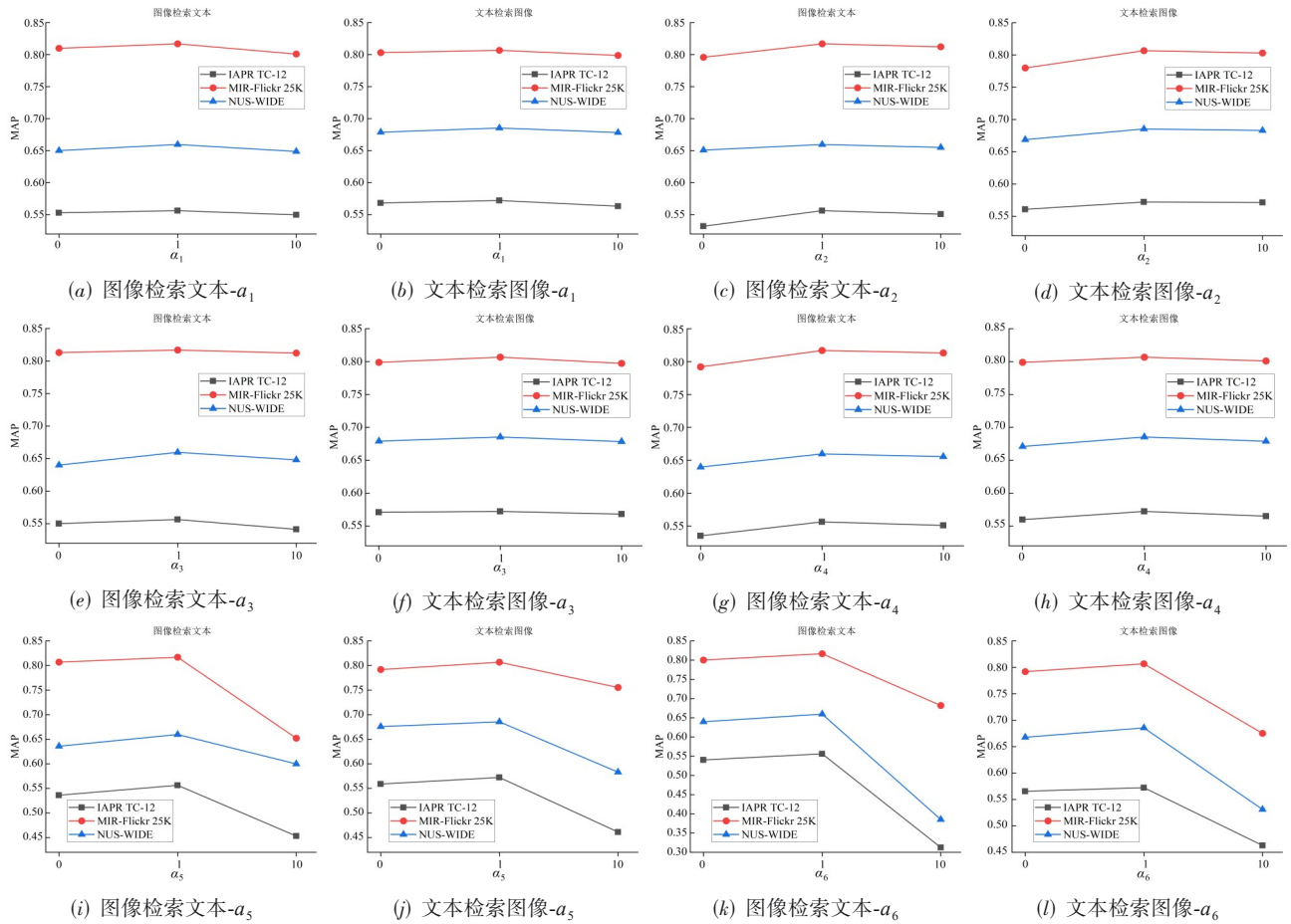
表7　CDHTK在三个数据的训练时间及模型大小

方法	实验策略	IAPR TC-12/s	MIR-Flickr 25K/s	NUS-WIDE/s	模型大小/M
CDHTK-1	没有使用迁移知识生成的辅助哈希码	0.27	0.39	2.35	152.33
CDHTK-2	没有使用文本的辅助哈希码	0.65	1.25	3.52	217.33
CDHTK-3	没有使用图像的辅助哈希码	0.64	1.45	3.65	218.13
CDHTK	—	0.85	1.83	4.51	283.13

4.5.3 参数敏感性分析

模型中参数的不同取值会直接影响模型性能，因此有必要对参数敏感性进行分析. 在本文中，CDHTK利用预测标签的交叉熵损失、生成哈希码的联合三元组量化损失以及迁移知识的差分损失，共同优化哈希码的生成过程. 本节针对以上损失函数的权重$\{a_i\}(i=1,2,\cdots,6)$进行消融性实验，分别对应式(4)、式(5)、式(7)、式(8)、式(10)、式(11)的权重取值. 为了探索每个参数的最佳取值(取值集合范围：$\{0.1,1,10\}$)，对三个数据集进行实验并测试一个参数，同时固定其他参数. 哈希码长度为16. 图3展示了不同参数下模型在三个常用数据集上的MAP.

从图3中可以看到，MAP结果在不同的$\{a_1,a_3\}$值下保持稳定，说明CDHTK对权重参数$\{a_1,a_3\}$并不敏感. 同时，可以看到MAP结果会随着$\{a_2,a_4\}$变化出现一定的波动. 当$\{a_2,a_4\}$取值为0.1时，MAP结果有所下降，随着权重比重增大时(=1时)，MAP结果达到最好. $\{a_2,a_4\}$取值表示式(5)、式(8)中生成哈希码的联合三元组量化损失函数的权重，这表明本文引入的联合三元组量化损失函数能够提高模型的检索能力. 值得注意的是，MAP结果在权重参数$\{a_5,a_6\}$取1时最好. 当取值大于1时，MAP结果呈现快速下降. 一个可能的原因是，当知识迁移网络损失函数权重$\{a_5,a_6\}$设置得太高时，模型忽略了观察到的标签的可靠信息，导致标签信息挖掘不完整.



(a) 图像检索文本-a_1 (b) 文本检索图像-a_1 (c) 图像检索文本-a_2 (d) 文本检索图像-a_2

(e) 图像检索文本-a_3 (f) 文本检索图像-a_3 (g) 图像检索文本-a_4 (h) 文本检索图像-a_4

(i) 图像检索文本-a_5 (j) 文本检索图像-a_5 (k) 图像检索文本-a_6 (l) 文本检索图像-a_6

图3 损失函数权重参数敏感性分析

5 总结

本文提出了一种基于迁移知识的跨模态双重哈希方法. 该方法构造了一个跨模态图像-文本哈希检索模型，通过结合图像网络、知识迁移网络以及文本网络完成跨模态哈希检索任务. CDHTK利用知识迁移网络生成辅助的图像哈希码和文本哈希码，之后通过与图像网络和文本网络各自生成的哈希码进行融合，从而生成了具有判别性的哈希码. 虽然CDHTK通过采用预测标签的交叉熵损失、生成哈希码的联合三元组量化损失以及迁移知识的差分损失来共同优化哈希码的生成过程，从而提高模型的检索效果；但是，图像网络和文本网络采用了经典但较为简单的CNN-F网络和词袋模型，没有使用更先进的网络结构和文本表示方式，这可

能会在一定程度上限制方法的性能上限. 因此,在下一步的工作中,探索将本文方法与更先进的网络结构(如Transformer等)和文本表示方式(如BERT等)相结合,以进一步提升跨模态哈希的性能,同时深入研究损失函数的构建机理,探索更合理有效的损失函数组合,以获得更优的哈希码质量.

参考文献

[1] 李志欣, 凌锋, 张灿龙, 等. 融合两级相似度的跨媒体图像文本检索[J]. 电子学报, 2021, 49(2): 268-274.
LI Z X, LING F, ZHANG C L, et al. Cross-media image-text retrieval with two level similarity[J]. Acta Electronica Sinica, 2021, 49(2): 268-274. (in Chinese)

[2] SHARMA A, KUMAR A, DAUME H, et al. Generalized Multiview Analysis: A discriminative latent space[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2012: 2160-2167.

[3] JING X Y, HU R M, ZHU Y P, et al. Intra-view and inter-view supervised correlation analysis for multi-view feature learning [C]//Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. New York: ACM, 2014: 1882-1889.

[4] JIA Y Q, SALZMANN M, DARRELL T. Learning cross-modality similarity for multinomial data[C]//2011 International Conference on Computer Vision. Piscataway: IEEE, 2011: 2407-2414.

[5] ZHENG Y, ZHANG Y J, LAROCHELLE H. Topic modeling of multimodal data: An autoregressive approach[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014: 1370-1377.

[6] WANG J, HE Y H, KANG C C, et al. Image-text cross-modal retrieval via modality-specific feature learning[C]// Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. New York: ACM, 2015: 347-354.

[7] FROME A, CORRADO G, SHLENS J, et al. Devise: A deep visual-semantic embedding model[C]//2013 the Advances in Neural Information Processing System. Massachusetts: MIT Press, 2013: 2121-2129.

[8] 姚涛, 孔祥维, 付海燕, 等. 基于映射字典学习的跨模态哈希检索[J]. 自动化学报, 2018, 44(8): 1475-1485.
YAO T, KONG X W, FU H Y, et al. Projective dictionary learning hashing for cross-modal retrieval[J]. Acta Automatica Sinica, 2018, 44(8): 1475-1485. (in Chinese)

[9] SONG J K, YANG Y, YANG Y, et al. Inter-media hashing for large-scale retrieval from heterogeneous data sources[C]// Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2013: 785-796.

[10] DING G G, GUO Y C, ZHOU J L. Collective matrix factorization hashing for multimodal data[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014: 2083-2090.

[11] 朱磊, 李京智, 王天时, 等. 联邦无监督跨模态哈希[J]. 中国科学: 信息科学, 2023, 53(11): 2180-2201.
ZHU L, LI J Z, WANG T S, et al. Federated unsupervised cross-modal Hashing[J]. Scientia Sinica (Informationis), 2023, 53(11): 2180-2201. (in Chinese)

[12] LIN Z J, DING G G, HU M Q, et al. Semantics-preserving hashing for cross-view retrieval[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2015: 3864-3872.

[13] 严双咏, 刘长红, 江爱文, 等. 语义耦合相关的判别式跨模态哈希学习算法[J]. 计算机学报, 2019, 42(1): 164-175.
YAN S Y, LIU C H, JIANG A W, et al. Discriminative cross-modal hashing with coupled semantic correlation[J]. Chinese Journal of Computers, 2019, 42(1): 164-175. (in Chinese)

[14] ZHANG D, LI W J. Large-scale supervised multimodal hashing with semantic correlation maximization[C]//Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. New York: ACM, 2014: 2177-2183.

[15] 李慧琼, 王永欣, 陈振铎, 等. 基于排序的监督离散跨模态哈希[J]. 计算机学报, 2021, 44(8): 1620-1635.
LI H Q, WANG Y X, CHEN Z D, et al. Ranking-based supervised discrete cross-modal hashing[J]. Chinese Journal of Computers, 2021, 44(8): 1620-1635. (in Chinese)

[16] CHATFIELD K, SIMONYAN K, VEDALDI A, et al. Return of the devil in the details: Delving deep into convolutional nets[C]//Proceedings of the British Machine Vision Conference 2014. London: British Machine Vision Association, 2014: 1-12.

[17] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2015-04-10]. https://arxiv.org/abs/1409.1556.

[18] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[19] JIANG Q Y, LI W J. Deep cross-modal hashing[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 3270-3278.

[20] YANG E K, DENG C, LIU W, et al. Pairwise relationship guided deep hashing for cross-modal retrieval[C]// Proceedings of the Thirty-First AAAI Conference on Arti-

ficial Intelligence. New York: ACM, 2017: 1618-1625.
[21] LI C, DENG C, LI N, et al. Self-supervised adversarial hashing networks for cross-modal retrieval[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 4242-4251.
[22] GU W, GU X Y, GU J Z, et al. Adversary guided asymmetric hashing for cross-modal retrieval[C]//Proceedings of the 2019 on International Conference on Multimedia Retrieval. New York: ACM, 2019: 159-167.
[23] LIN Q B, CAO W M, HE Z H, et al. Semantic deep cross-modal hashing[J]. Neurocomputing, 2020, 396: 113-122.
[24] LIN Q B, CAO W M, HE Z Q, et al. Mask cross-modal hashing networks[J]. IEEE Transactions on Multimedia, 2020, 23: 550-558.
[25] YAO H L, ZHAN Y W, CHEN Z D, et al. TEACH: Attention-aware deep cross-modal hashing[C]//Proceedings of the 2021 International Conference on Multimedia Retrieval. New York: ACM, 2021: 376-384.
[26] YU E, MA J H, SUN J D, et al. Deep discrete cross-modal hashing with multiple supervision[J]. Neurocomputing, 2022, 486: 215-224.
[27] GAO Z J, WANG J, YU G X, et al. Long-tail cross modal hashing[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(6): 7642-7650.
[28] LI H X, ZHANG C, JIA X Y, et al. Adaptive label correlation based asymmetric discrete hashing for cross-modal retrieval[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(2): 1185-1199.
[29] ZHOU J L, DING G G, GUO Y C, et al. Latent semantic sparse hashing for cross-modal similarity search[C]//Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. New York: ACM, 2014: 415-424.
[30] XU X, SHEN F M, YANG Y, et al. Learning discriminative binary codes for large-scale cross-modal retrieval[J]. IEEE Transactions on Image Processing, 2017, 26(5): 2494-2507.
[31] IRIE G, ARAI H, TANIGUCHI Y. Alternating co-quantization for cross-modal hashing[C]//2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2015: 1886-1894.
[32] LIN Z J, DING G G, HAN J G, et al. Cross-view retrieval via probability-based semantics-preserving hashing[J]. IEEE Transactions on Cybernetics, 2017, 47(12): 4342-4355.
[33] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems. Massachusetts: MIT Press, 2012: 1097-1105.
[34] ESCALANTE H J, HERNÁNDEZ C A, GONZALEZ J A, et al. The segmented and annotated IAPR TC-12 benchmark[J]. Computer Vision and Image Understanding, 2010, 114(4): 419-428.
[35] HUISKES M J, LEW M S, HUISKES M J, et al. The MIR flickr retrieval evaluation[C]//Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval. New York: ACM, 2008: 39-43.
[36] CHUA T S, TANG J H, HONG R C, et al. NUS-WIDE: A real-world web image database from National University of Singapore[C]//Proceedings of the ACM International Conference on Image and Video Retrieval. New York: ACM, 2009: 1-9.

作者简介

钟建奇 男，1997年出生，湖南衡阳人．现为深圳大学电子与信息工程学院信息与通信工程在读博士研究生．主要研究方向为深度学习、跨模态检索、人体动作理解．
E-mail: zhongjianqi2017@email.szu.edu.cn

林秋斌 男，1994年出生，广东潮汕人．深圳大学电子与信息工程学院信息与通信工程专业博士．主要研究方向为深度学习、跨模态检索．
E-mail: 2170269126@email.szu.edu.cn

曹文明 男，1965年出生，江苏淮阴人．现为深圳大学电子与信息工程学院教授，博士生导师．主要研究方向为多媒体信息处理、模式识别和人工智能算法．
E-mail: wmcao@szu.edu.cn