激光与光电子学进展 先进成像

基于多尺度融合和投影匹配约束的跨模态哈希方法

邓万宇，赵怡娜*，杨婉祯，张博，李昊，叶书齐
西安邮电大学计算机学院，陕西 西安 710121

摘要 大多数基于深度学习的跨模态哈希方法直接通过神经网络学习不同模态数据的统一哈希码。这些方法忽略了单模态数据不同尺度包含不同语义信息这一影响数据特征表示的因素以及低维特征在弥合模态鸿沟上的重要性。基于上述问题，提出一种基于多尺度融合和投影匹配约束的跨模态哈希方法(MFPMC)。通过设计图像多尺度融合网络和文本多尺度融合网络来获取不同模态数据的低维特征，引入低维特征投影匹配约束和对抗训练来保证低维特征在模态间分布的一致性，同时用包含丰富语义信息的低维特征作为哈希函数的输入，进一步构建模态内哈希码损失、模态间哈希码损失、量化损失、标签嵌入损失来约束哈希函数及哈希码的学习，以此保证生成具有判别性的离散二进制哈希码。在MIRFlickr-25K和NUS-WIDE两个基准的跨模态检索数据集上的实验表明：所提方法比现有的几种哈希方法具有更好的检索性能。

关键词 跨模态哈希检索；多尺度融合；低维特征；投影匹配约束
中图分类号 TP391.3 **文献标志码** A **DOI**：10.3788/LOP202259.2410006

Cross-Modal Hash Method Based on Multi-Scale Fusion and Projection Matching Constraint

Deng Wanyu, Zhao Yina*, Yang Wanzhen, Zhang Bo, Li Hao, Ye Shuqi
School of Computer Science & Technology, Xi'an University of Posts & Telecommunications, Xi'an 710121, *Shaanxi, China*

Abstract Most cross-modal Hash methods based on deep learning learn unified Hash codes of different-modality data directly through neural networks. However, these methods ignore the factor that different scales of single-modality data contain different semantic information, which affects the data feature representation, and the importance of low-dimensional features in bridging the "heterogeneity" gap. Based on the above problems, we propose a new cross-modal Hash retrieval method (MFPMC) based on multi-scale fusion and projection matching constraint, which obtains low-dimensional features of different-modality data by designing the image multi-scale fusion network and text multi-scale fusion network. Moreover, it introduces the low-dimensional feature projection matching constraint and adversarial training to ensure the distribution consistency of low-dimensional features among different modalities. Simultaneously, low-dimensional features containing rich semantic information are used as inputs for the Hash function. Furthermore, inter-modal Hash code, intra-modal Hash code, quantization, and label-embedding losses are constructed to constrain the learning of Hash function and Hash codes to ensure the generation of discriminative discrete binary Hash codes. Experiments on two benchmark cross-modal retrieval datasets (MIRFlickr-25K and NUS-WIDE) reveal that the proposed method outperforms other methods in terms of retrieval performance.

Key words cross-modal hash retrieval; multi-scale fusion; low-dimensional feature; projection matching constraint

1 引 言

近年来，伴随着信息技术的快速发展，数据存在的形式也从单一的文本转变为图像、视频、音频、3D模型等多种媒体类型，获取相似语义但不同类型的数据已经成为当下信息检索的新趋势，研究人员将这一新的检索方式称为跨模态检索[1-3]。具体来说，每一种类型的数据称为一种模态，跨模态检索就是通过一种模态的数据(例如：文本)找到语义上与之相似的另一模态数据(例如：图像)的过程。为了满足大规模数据搜索

收稿日期：2021-09-17；**修回日期**：2021-10-19；**录用日期**：2021-10-27
通信作者：*3065783275@qq.com

中存储成本低、搜索速度快的需求，跨模态哈希检索受到了广泛的关注。在早期的研究中，许多都是针对单模态哈希检索的方法[4-8]，这些方法面对的是同一类型的数据，不存在空间异构的问题，因而并不适用于跨模态哈希检索，因为跨模态哈希检索面临的是不同类型的数据，并且这些数据处于不同的空间中，难以进行统一的表示和衡量。因此，如何对不同模态数据进行有效的表示以及解决不同模态数据之间的语义鸿沟是跨模态哈希检索面临的巨大挑战。

目前，按照特征提取方式的不同，跨模态哈希方法可以划分为两大类：基于手工提取特征的跨模态哈希方法[9-14]、基于深度学习的跨模态哈希方法[15-20]。基于手工提取特征的方法主要通过手工提取不同模态的特征（如尺度不变特征变换 SIFT、空间包络特征 GIST），然后再进行跨模态的相关性学习和检索。Ding 等[9]提出了基于多模态数据的集体矩阵分解哈希（CMFH），通过矩阵分解的方式获取不同模态数据的特征进而来学习不同模态数据统一的哈希码。Wang 等[10]提出了跨媒体语义主题多模态哈希（STMH），通过对文本模态的数据和图像模态的数据分别进行聚类和鲁棒性矩阵分解得到文本的多个语义主题和图像的多个概念，然后充分利用数据的语义标签信息生成离散二进制哈希码。这一类方法有两个局限性：首先是提取不同模态数据特征的方法不能充分挖掘不同模态数据的潜在特征；其次是这类方法将特征提取和哈希码学习分为两个步骤来完成，容易造成两者不兼容的问题。近几年来，深度学习在特征提取方面取得巨大进步，为跨模态哈希检索性能的提升提供了一个很好的契机，吸引了许多科研人员在深度学习这个领域对跨模态哈希检索展开新的研究。Jiang 等[15]提出了深度跨模态哈希（DCMH），将不同模态数据的特征提取与哈希码的学习通过神经网络集成到一个端到端的框架中，从而直接生成哈希码，解决了提取的特征与生成的哈希码不兼容的问题。Cao 等[17]提出了跨模态汉明哈希（CMHH），与 DCMH 不同之处在于，CMHH 是通过设计基于指数分布的双焦损失函数来惩罚汉明距离大于汉明半径阈值的相似图像-文本对的，从而直接生成紧凑、高度集中的哈希码。这一类方法采用深度卷积神经网络提取不同模态数据的特征，与基于手工提取特征的跨模态哈希方法相比，解决了手工提取特征表达能力有限的问题。此外，该方法还可以将特征提取融入到哈希码的学习过程中，保证哈希码的准确性，从而获得更好的检索效率。但值得注意的是，基于深度学习的跨模态哈希方法仍然存在一些普遍的缺点。这类方法直接从神经网络最后的全连接层中提取特征作为生成的哈希码，一方面忽略了不同尺度数据包含着的不同语义信息在不同模态特征提取中的重要性，另一方面也忽略了低维特征包含的丰富语义信息对哈希码生成及模态差异弥合的重要性。

针对上述基于深度学习的跨模态哈希方法忽略的两个问题，本文提出了一种基于多尺度融合和投影匹配约束的跨模态哈希算法（MFPMC）。首先，在图像神经网络和文本神经网络中分别设计了一个图像多尺度融合模型和文本多尺度融合模型，以挖掘单个模态数据不同尺度所包含的不同语义信息，提高神经网络对提取的不同模态数据特征的表达能力。其次，考虑到直接生成哈希码所造成的大量信息丢失以及低维特征包含的丰富语义信息有助于弥合模态差异的优势，引入了低维特征投影匹配约束。最后，为了保证哈希码的有效性，还引入了标签嵌入技术。

2 所提方法内容

在这里，本文只讨论图像模态和文本模态两种类型的数据，当然，所提方法也可以应用到其他模态的数据上。所提 MFPMC 的框架如图 1 所示，该框架将图像多尺度融合模型（IMFM）、文本多尺度融合模型（TMFM）、低维特征学习和哈希码学习集成到一个统一的端到端架构中。不同模态的数据经过多尺度融合网络后通过低维特征投影匹配约束（LFPMC）和对抗学习得到不同模态数据的低维特征，然后将不同模态的低维特征作为哈希函数的输入，进行约束学习，最终得到统一的二进制哈希码。

2.1 问题描述

先介绍一些跨模态哈希的形式化定义以及使用的一些符号。假设训练数据集中包含 n 个样本数据，可以表示成 $\boldsymbol{O}=\{\boldsymbol{O}_i\}_{i=1}^{n}$，其中 $\boldsymbol{O}_i=(\boldsymbol{x}_i,\boldsymbol{y}_i,\boldsymbol{l}_i)$ 表示第 i 个样本数据，$\boldsymbol{x}_i$ 和 $\boldsymbol{y}_i$ 分别表示第 i 个实例中的图像样本数据和文本样本数据，$\boldsymbol{l}_i=[l_{i1},l_{i2},\cdots,l_{ic}]$ 表示第 i 个样本数据 $\boldsymbol{O}_i$ 的标签，c 为总类别数。如果实例 $\boldsymbol{O}_i$ 属于第 j 个类别，则 $l_{ij}=1$，否则 $l_{ij}=0$。$\boldsymbol{S}$ 表示成对多标签相似度矩阵，用来描述两个实例之间真实的语义相似度。$\boldsymbol{f}^{\mathrm{x}}=F^{\mathrm{x}}(\boldsymbol{X};\theta_{F^{\mathrm{x}}})$ 表示从图像多尺度融合神经网络中学习到的低维特征，$\boldsymbol{f}^{\mathrm{y}}=F^{\mathrm{y}}(\boldsymbol{Y};\theta_{F^{\mathrm{y}}})$ 表示从文本多尺度融合神经网络中学习到的低维特征，$\theta_{F^{\mathrm{x}}}$ 和 $\theta_{F^{\mathrm{y}}}$ 分别是图像多尺度融合神经网络的参数和文本多尺度融合神经网络的参数。此外，定义 $\boldsymbol{H}^{\mathrm{x}}\in\mathbf{R}^{r\times n}$ 和 $\boldsymbol{H}^{\mathrm{y}}\in\mathbf{R}^{r\times n}$ 分别为经过哈希函数得到的图像模态和文本模态的伪哈希码，它们是连续的实值。由于跨模态哈希的目标是为两种模态生成统一的二进制离散哈希码 $\boldsymbol{B}^{\mathrm{x,y}}\in\{-1,+1\}^{r\times n}$，其中 r 为二进制哈希码的长度，因此需要对生成的连续实值的伪哈希码进行离散化，最终才能得到离散二进制哈希码。跨模态哈希检索则是利用汉明距离通过比特运算计算二进制哈希码之间相似度的过程，从而找到与待查询样本语义相似的另一模态的数据。

2.2 基于多尺度融合的网络架构

基于多尺度融合的网络架构由两个神经网络组

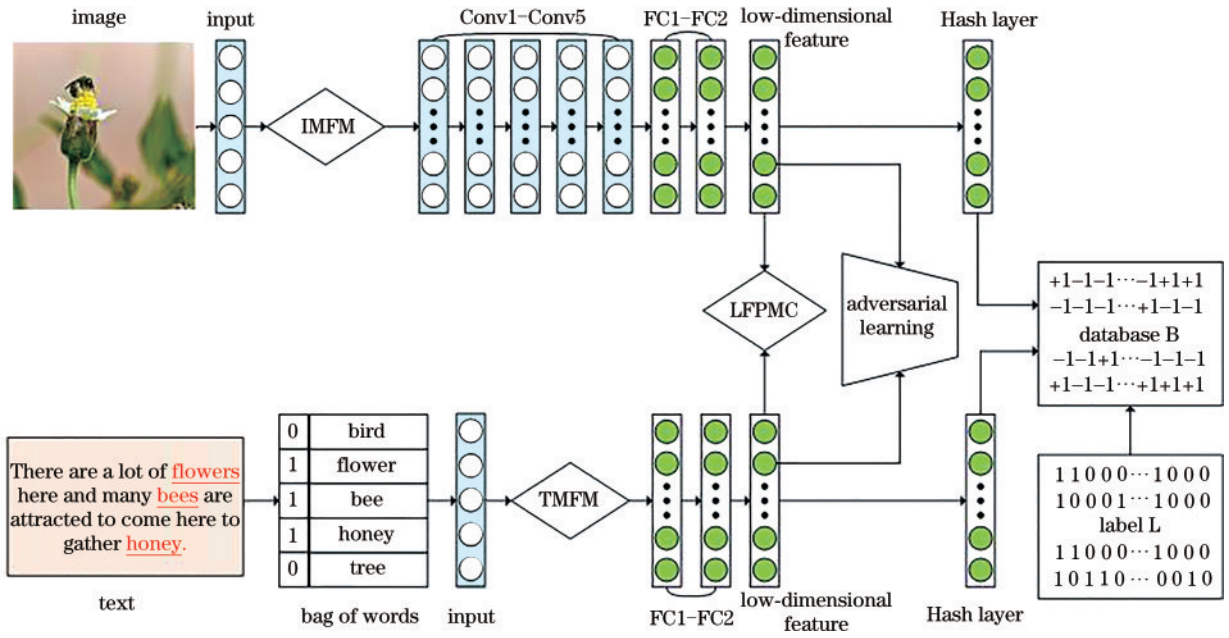

图 1　所提 MFPMC 的框架图

Fig. 1　Framework of proposed MFPMC

成，一个用于图像模态，另一个用于文本模态。其设计目的：一方面是为了解决神经网络在提取特征过程中忽略单个模态数据不同尺度包含丰富语义信息的问题；另一方面也是为了解决图像样本数据和文本样本数据直接输入神经网络中所造成的部分信息丢失的问题。不同模态多尺度融合网络具体的设计缘由及设计思路如下。

1）基于图像的多尺度融合神经网络

图像多尺度融合神经网络是在图像多尺度融合模型（IMFM）和 CNN-F 神经网络[21]的基础上构建的。受 PSPnet[22]以及行人重识别在特征融合方面所做工作[23-25]的启发，本文提出 IMFM。IMFM 由 3 个平均池化层和 3 个 1×1 卷积层组成。将原始图像作为 IMFM 中每个池化层的输入，通过这些池化层将得到的多尺度特征输入到一个 1×1 卷积层，然后对结果进行融合作为 CNN-F 的输入，从而得到图像模态的低维特征。IMFM 成功地解决了传统 CNN-F 对输入图像尺寸大小的限制以及一些信息丢失导致特征学习不可靠的问题。IMFM 的详细设置如表 1 所示。

2）基于文本的多尺度融合神经网络

对于文本模态来说，词袋向量（BOW）被广泛用于表示文本，这很容易导致稀疏性。为了解决这一问题，所提方法在文本多尺度融合网络中采用了文本多尺度融合模型（TMFM），该模型的设计是基于 SSAH[26]中的方法。TMFM 由 5 个平均池化层和 5 个 1×1 卷积层组成。用 BOW 表示的文本依次通过池化层和卷积层，获得文本的多尺度特征，然后将这些多尺度特征融合并输入到三层前馈神经网络中，最终获得具有丰富语义信息的文本低维特征。表 2 显示了 TMFM 的详

表 1　IMFM 的详细参数设置

Table 1　Detailed parameter settings for IMFM

Input	Layer	Kernel size	Stride	Output
Original image	Average pooling 1	5×5	5×5	Ipool 1
Ipool 1	1×1Conv	1×1	1×1	IMs-feture 1
Original image	Average pooling 2	10×10	10×10	Ipool 2
Ipool 2	1×1Conv	1×1	1×1	IMs-feature 2
Original image	Average pooling 3	15×15	15×15	Ipool 3
Ipool 3	1×1Conv	1×1	1×1	IMs-feature 3

表 2　TMFM 的详细参数设置

Table 2　Detailed parameter settings for TMFM

Input	Layer	Kernel size	Stride	Output
Bow vector	Average pooling 1	1×50	1×50	Tpool 1
Tpool 1	1×1Conv	1×1	1×1	TMs-feture 1
Bow vector	Average pooling 2	1×30	1×30	Tpool 2
Tpool 2	1×1Conv	1×1	1×1	TMs-feature 2
Bow vector	Average pooling 3	1×15	1×15	Tpool 3
Tpool 3	1×1Conv	1×1	1×1	TMs-feature 3
Bow vector	Average pooling 4	1×10	1×10	Tpool 4
Tpool 4	1×1Conv	1×1	1×1	TMs-feature 4
Bow vector	Average pooling 5	1×5	1×5	Tpool 5
Tpool 5	1×1Conv	1×1	1×1	TMs-feature 5

细参数设置。

2.3 低维特征学习

哈希码码长通常小于128位，这意味着大多数有用的信息都被中和了，使得生成的哈希码无法捕获固有的模态一致性。相比之下，低维特定的模态特征包含更丰富的信息，有助于弥合不同模态间的差距。基于此，本文通过引入对抗训练和低维特征投影匹配约束（LFPMC）来学习从多尺度融合网络中得到的不同模态的低维特征。

1）对抗训练

对抗训练的目的是使图像模态和文本模态学习到的低维特征表示分布尽可能相同。为了达到这一目标，定义了一个判别器（D），用于区分低维特征是来自图像模态还是文本模态。所提方法假定来自图像模态的低维特征为真，来自文本模态的低维特征为假，则对抗性损失可表示为

$$L_{\mathrm{adv}}=-\frac{1}{n}\sum_{i=1}^{n}\left\{\log_{10}\left[D\left(\boldsymbol{f}_i^{\mathrm{x}};\theta_D\right)\right]+\log_{10}\left[1-D\left(\boldsymbol{f}_j^{\mathrm{y}};\theta_D\right)\right]\right\},\tag{1}$$

式中：θ_D为判别器的参数；$\boldsymbol{f}_i^{\mathrm{x}}$表示第$i$张图像的低维特征；$\boldsymbol{f}_j^{\mathrm{y}}$表示第$j$个文本的低维特征。

2）低维特征投影匹配约束

采用低维特征投影匹配约束，主要是为了保证不同模态的数据嵌入到低维特征空间中时语义信息不会丢失，同时也是为了弥合不同模态数据之间的异质鸿沟，进一步使不同模态生成的哈希码在同一个空间中能够更好地度量相似性。低维特征投影匹配约束的本质是用Kullback-Leibler（KL）散度来最小化不同模态数据低维特征投影分布与真实标签投影分布，从而尽最大可能关联低维特征空间中不同模态的语义一致性。

真实的标签投影分布反映了不同模态数据真实的语义相似度，因此可以通过构建成对相似度矩阵$\boldsymbol{S}$将其转换为概率分布。标签构造的成对相似度矩阵是基于这样的事实的：如果图像样本数据和文本样本数据至少有一个相同的标签，则两个样本数据是相似的；否则，它们就不相似。对于多标签数据来说，这不是一个很好的衡量两个样本数据之间相似性信息的方法。为了更好地度量标签之间的相似度，利用余弦相似度来构建成对相似度矩阵。如果两个样本数据相似，则对应的标签相似度为

$$\tilde{S}_{ij}=\frac{\sum_{k=1}^{c}l_{ik}l_{jk}}{\sqrt{\sum_{k=1}^{c}l_{ik}^2}\sqrt{\sum_{k=1}^{c}l_{jk}^2}},\tag{2}$$

式中：$\tilde{S}_{ij}$表示两个样本数据之间真实的语义相似度；l_{ik}表示类标签向量$\boldsymbol{l}_i$中的第k个元素；l_{jk}表示类标签向量$\boldsymbol{l}_j$中的第k个元素。为了简化式（2），定义

$$\tilde{l}_{ik}=\frac{l_{ik}}{\sqrt{\sum_{k=1}^{c}l_{ik}^2}},\tag{3}$$

$$\tilde{l}_{jk}=\frac{l_{jk}}{\sqrt{\sum_{k=1}^{c}l_{jk}^2}}。\tag{4}$$

由式（3）和式（4）可以得到简化的$\tilde{S}_{ij}=\tilde{\boldsymbol{l}}_i^{\mathrm{T}}\tilde{\boldsymbol{l}}_j$，因此成对相似度矩阵最终可以定义为$\boldsymbol{S}=2\tilde{\boldsymbol{S}}-\boldsymbol{I}=2\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{L}}-\boldsymbol{I}$，这里$\boldsymbol{I}\in\{1\}$，$\tilde{\boldsymbol{L}}$表示所有标签经过式（3）和式（4）计算得到的标签矩阵。然后，用成对相似度矩阵$\boldsymbol{S}$的归一化输出P_{ij}表示真实的标签投影分布，定义为

$$P_{ij}=\frac{S_{ij}}{\sum_{k=1}^{n}S_{ik}}。\tag{5}$$

虽然多尺度融合网络使不同模态的数据都能够用数字化的形式来表示，但是在本质上不同模态的数据还是处于异构空间中的，直接度量不同模态之间的相似性不能很好地挖掘不同模态之间的相关性。因而借助数学中投影的概念，将一种模态的数据映射到另一模态所在空间中，依据投影向量越大，两种模态的低维特征向量越相似，投影向量越小，两种模态的特征向量相似度越小来衡量异构数据之间的相似性。而低维特征投影分布则是将投影向量用Softmax函数归一化输出得到的。因此，图像模态的低维特征投影到文本模态低维特征空间中的分布为

$$q_{ij}=\frac{\exp\left[\left(\boldsymbol{f}_i^{\mathrm{x}}\right)^{\mathrm{T}}\overline{\boldsymbol{f}_j^{\mathrm{y}}}\right]}{\sum_{k=1}^{n}\exp\left[\left(\boldsymbol{f}_i^{\mathrm{x}}\right)^{\mathrm{T}}\overline{\boldsymbol{f}_k^{\mathrm{y}}}\right]},\tag{6}$$

式中：$\overline{\boldsymbol{f}_j^{\mathrm{y}}}=\frac{\boldsymbol{f}_j^{\mathrm{y}}}{\left\|\boldsymbol{f}_j^{\mathrm{y}}\right\|_{\mathrm{F}}^2}$表示文本的归一化单位特征向量；$\left(\boldsymbol{f}_i^{\mathrm{x}}\right)^{\mathrm{T}}\overline{\boldsymbol{f}_j^{\mathrm{y}}}$表示第$i$幅图像的低维特征在第$j$个归一化文本低维特征上的投影向量。

KL散度主要用来比较两个概率分布之间的接近程度，在本文中则用来衡量标签投影分布P_{ij}和图像低维特征到文本低维特征空间的投影分布q_{ij}之间的接近程度，这样能够使不同模态之间的相似性通过向真实的语义信息靠拢得到加强。因此，图像模态到文本模态的低维特征投影匹配约束定义为

$$L_{p(\mathrm{i}->\mathrm{t})}=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}q_{ij}\log_{10}\left(\frac{q_{ij}}{P_{ij}+\delta}\right),\tag{7}$$

式中：δ值很小，用来防止分母为零；$p(\mathrm{i}->\mathrm{t})$是一个整体，表示从图像（image）到文本（text）的投影（projection）；$L_{p(\mathrm{i}->\mathrm{t})}$表示从图像模态到文本模态的低维特征投影匹配约束。同样，从文本模态到图像模态的低维特征投影匹配约束定义为$L_{p(\mathrm{t}->\mathrm{i})}$，其定义类似于$L_{p(\mathrm{i}->\mathrm{t})}$。因此，图像模态和文本模态的低维特征投影匹配约束损失定义为

$$L_p = L_{p(t->i)} + L_{p(i->t)}。\tag{8}$$

2.4 哈希码学习

跨模态哈希的最终目的是使不同模态的数据生成的哈希码能够在公共汉明空间中得到有效的表达。所提方法中图像模态的哈希函数和文本模态的哈希函数均由一个全连接层构成。用不同模态的低维特征作为哈希函数的输入，最终得到图像模态的伪哈希码 $\boldsymbol{H}^{x}$ 和文本模态的伪哈希码 $\boldsymbol{H}^{y}$。为了保证学习到具有判别力的哈希码，需要定义不同的损失函数对不同模态的哈希函数进行学习。

1）模态间哈希码损失

大多数方法使用标签构造的成对相似度矩阵 $\boldsymbol{S}$ 来约束不同模态数据的哈希码，以保证模态之间的相似性，以此来构建模态间哈希码的损失。

为了实现鲁棒的跨模态哈希检索，利用跨模态相似度的负似然对数有效地捕捉不同模态之间的异构相似度。似然函数定义为

$$g\left(S_{ij}|\boldsymbol{H}^{x}_{*i}, \boldsymbol{H}^{y}_{*j}\right)=\begin{cases}\mathrm{sig}\left(\psi_{ij}\right), & S_{ij}\neq 0\\ 1-\mathrm{sig}\left(\psi_{ij}\right), & S_{ij}=0\end{cases},\tag{9}$$

式中：$\psi_{ij}=\frac{1}{2}\left(\boldsymbol{H}^{x}_{*i}\right)^{\mathrm{T}}\boldsymbol{H}^{y}_{*j}$；$\mathrm{sig}\left(\psi_{ij}\right)=\frac{1}{1+\mathrm{e}^{-\psi_{ij}}}$。因此，模态间的损失定义为

$$J_1=-\sum_{i,j=1}^{n}\left[S_{ij}\psi_{ij}-\log_{10}\left(1+\mathrm{e}^{\psi_{ij}}\right)\right]+\alpha\left(\left\|\boldsymbol{H}^{x}\boldsymbol{I}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{H}^{y}\boldsymbol{I}\right\|_{\mathrm{F}}^{2}\right),\tag{10}$$

式中：α 为超参数；$\boldsymbol{H}^{x}_{*i}=h^{x}\left(\boldsymbol{f}_i^{x};\theta_x\right)$ 为第 i 张图像的伪哈希码；$\boldsymbol{H}^{y}_{*j}=h^{y}\left(\boldsymbol{f}_j^{y};\theta_y\right)$ 为第 j 个文本的伪哈希码；θ_x 和 θ_y 分别表示图像哈希函数和文本哈希函数的参数；$\alpha\left(\left\|\boldsymbol{H}^{x}\boldsymbol{I}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{H}^{y}\boldsymbol{I}\right\|_{\mathrm{F}}^{2}\right)$ 这一项用于使生成的哈希码中 -1 和 1 的数量尽可能保持平衡。

2）模态内哈希码的损失

为了使同一模态的数据生成的哈希码与其对应标签的语义信息保持一致，通过线性映射将哈希码映射到公共语义表示空间，以减少模态内的语义损失。模态内哈希码的损失定义为

$$J_2=\left\|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{H}^{x}-\boldsymbol{L}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{H}^{y}-\boldsymbol{L}\right\|_{\mathrm{F}}^{2}+\beta\left\|\boldsymbol{W}\right\|_{\mathrm{F}}^{2},\tag{11}$$

式中：$\boldsymbol{W}$ 为映射矩阵；$\beta\left\|\boldsymbol{W}\right\|_{\mathrm{F}}^{2}$ 是对 $\boldsymbol{W}$ 的惩罚项；β 是正则化的权重因子；$\left\|\bullet\right\|_{\mathrm{F}}^{2}$ 表示 Frobenius 范数；$\boldsymbol{L}$ 表示标签向量构成的矩阵。

3）量化损失

考虑到两个网络输出的伪哈希码是连续变量而最终所要的哈希码是离散变量，本文将连续变量松弛为离散变量，通过最小化实值嵌入伪哈希码和离散二进制哈希码之间的量化损失，使生成的离散哈希码保持原有的语义信息。量化损失定义为

$$J_3=\left\|\boldsymbol{B}^{x}-\boldsymbol{H}^{x}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{B}^{y}-\boldsymbol{H}^{y}\right\|_{\mathrm{F}}^{2}。\tag{12}$$

在实验中发现，如果将两种模态中相似样本数据对应的哈希码设置为相同，则在训练优化过程中可以获得更好的性能。因此加上另一个约束条件 $\boldsymbol{B}=\boldsymbol{B}^{x}=\boldsymbol{B}^{y}$，则式(12)可以等价地转换为

$$J_4=\left\|\boldsymbol{B}-\boldsymbol{H}^{x}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{B}-\boldsymbol{H}^{y}\right\|_{\mathrm{F}}^{2}。\tag{13}$$

4）标签嵌入损失

大多数有监督跨模态哈希检索方法仅仅对生成的伪哈希码通过标签语义信息进行约束，忽略了最终生成的离散二进制哈希码在语义上更应该与其对应的标签语义信息保持一致的内容。基于此，引入了标签嵌入技术，具体来说就是通过线性映射将标签信息映射到离散二进制哈希码所在的汉明空间，使离散二进制哈希码所表达的语义信息与映射之后的标签语义信息越接近越好。标签嵌入损失定义为

$$J_5=\left\|\boldsymbol{B}-\boldsymbol{L}\boldsymbol{V}\right\|_{\mathrm{F}}^{2}+\lambda\left\|\boldsymbol{V}\right\|_{\mathrm{F}}^{2},\tag{14}$$

式中：$\boldsymbol{V}$ 表示映射矩阵；$\lambda\left\|\boldsymbol{V}\right\|_{\mathrm{F}}^{2}$ 是对 $\boldsymbol{V}$ 的惩罚项；λ 是正则化的权重因子。综合式(1)～(14)，可以得到最终的目标函数为

$$\min_{\theta_{F^x},\theta_{F^y},\theta_D,\theta_x,\theta_y,\boldsymbol{W},\boldsymbol{V},\boldsymbol{B}} J_{\mathrm{total}}=L_{\mathrm{adv}}+\xi L_{\mathrm{p}}+\gamma J_1+\tau J_2+\eta J_4+\mu J_5,\tag{15}$$

式中：ξ、γ、τ、η、μ 是控制每个损失项目权重的平衡参数。

2.5 优化

由于最终的目标函数 J_{total} 是非凸函数，难以直接进行优化，故而采用交替学习策略来学习参数 θ_{F^x}、θ_{F^y}、θ_D、θ_x、θ_y、$\boldsymbol{W}$、$\boldsymbol{V}$、$\boldsymbol{B}$。首先，固定矩阵变量 $\boldsymbol{W}$、$\boldsymbol{V}$ 和 $\boldsymbol{B}$，然后采用随机梯度下降优化算法对多尺度融合神经网络的参数进行优化。利用随机梯度下降算法计算总的损失函数 J_{total} 的梯度后，利用反向传播算法对网络的参数 θ_{F^x}、θ_{F^y}、θ_D、θ_x 和 θ_y 进行更新。由于随机梯度下降算法在深度学习中应用广泛，所以在这里就不进行描述了。训练完所有数据后，通过固定多尺度融合神经网络的参数和任意两个矩阵变量，更新另一个矩阵变量，直到三个矩阵变量都被更新，然后开始下一次训练，直到总损失函数收敛为止。

3 实验结果分析

3.1 数据集

选择两个标准的跨模态数据集，即 NUS-WIDE 数据集[27]和 MIRFlickr-25K 数据集[28]，来评估所提方法。

NUS-WIDE 数据集包含 269648 个与文本标记 web 图像相关联的实例。选取 21 个常见概念，包括 195834 个图像-文本对，并从该数据集中随机抽取 2100 个图像-文本对作为测试集，其余作为查询集，然后又从查询集中随机抽取 10500 个图像-文本对作为

训练集。每个文本样本数据都由一个1000维的词向量包表示。对于基于手工提取特征的跨模态哈希方法，每个图像由一个500维视觉词袋向量表示。对于基于深度学习的跨模态哈希方法，用原始的图像样本数据直接作为图像神经网络的输入。

MIRFlickr-25K数据集包含25000个图像-文本对。在实验中，选择那些被至少20个文本标记的样本，最终得到20015个图像-文本对。从该数据集中随机选取2000个图像-文本对作为测试集，其余作为查询集，然后又从查询集中随机选取10000个图像-文本对作为训练集。每个文本样本数据表示为一个1386维的词向量包。对于基于手工提取特征的跨模态哈希方法，每个图像样本由512维SIFT特征向量表示。对于基于深度学习的跨模态哈希方法，用原始的图像样本数据直接作为图像神经网络的输入。

3.2 基线方法和评价指标

为了证明所提方法的有效性，采用两种在信息检索领域常用的评价指标来进行评估，这两种评价指标分别是平均精度均值(mAP)和准确率-查全率曲线(P-R曲线)。在实验中，进行两种类型的跨模态检索任务：Image2Text表示通过图像样本数据检索到在文本查询集中与之语义相似的文本样本数据；Text2Image表示通过文本样本数据检索到在图像查询集中与之语义相似的图像样本数据。

为了比较，选取了几种具有代表性的方法作为基线对比实验对象，这些方法包括SCM[11]、SePH[12]、STMH[10]、CMFH[9]、DCMH[15]、PRDH[16]、CMHH[17]。对于基于深度学习的跨模态哈希方法，统一采用CNN-F提取图像特征。对于选取的基线对比实验对象，对所有超参数进行了初始化。对于每一种跨模态哈希方法，均进行5次实验，然后取5次实验结果的平均值作为最终的实验结果。所提方法在实验中的超参数设置为 $\alpha=0.8$，$\beta=0.01$，$\xi=10$，$\tau=1$，$\gamma=1.2$，$\eta=0.8$，$\mu=1.2$。

3.3 性能评估

对所提MFPMC与所选取的其他不同哈希方法的mAP值进行了比较，mAP值越大表示检索性能越好。在实验中，哈希码的长度被设置为16，32，64 bit，最佳性能得分的方法对应的mAP值以粗体显示。表3显示了所提方法与其他基线实验方法在MIRFlickr-25K数据集上mAP值的比较。表4给出了所提方法与其他基线实验方法在NUS-WIDE数据集上mAP值的比较。从表3和表4可以看出：对于Image2Text任务，当哈希码长为16 bit时，与次优方法相比，所提方法在MIRFlickr-25K和NUS-WIDE数据集上的mAP值分别提高了1.99个百分点和1.23个百分点；当哈希码长为32 bit时，与次优方法相比，所提方法在MIRFlickr-25K和NUS-WIDE数据集上的mAP值分别提高了2.21个百分点和1.38个百分点；当哈希码长度为64 bit时，与次优方法相比，所提方法在MIRFlickr-25K和NUS-WIDE数据集上的mAP值分别提高了2.37个百分点和1.40个百分点。对于Text2Image任务，当哈希码长为16 bit时，与次优方法相比，所提方法在MIRFlickr-25K和NUS-WIDE数据集上的mAP值分别提高了1.57个百分点和0.91个百分点；当哈希码长为32 bit时，与次优方法相比，所提方法在MIRFlickr-25K和NUS-WIDE数据集上的mAP值分别提高了1.56个百分点和0.88个百分点；当哈希码长为64 bit时，与次优方法相比，所提方法在MIRFlickr-25K和NUS-WIDE数据集上的mAP值分别提高了1.14个百分点和0.88个百分点。由此可以证明：所提方法在两个基准数据集上均实现了最优的检索性能。此外还可以看出：相同条件下，哈希码长度为64 bit时，所提方法的检索准确性高于哈希码长度为16 bit、32 bit的情况，由此可以证明，在一定条件下，哈希码长度越长，包含的语义信息越丰富，相应的检索性能也越好。

表3 不同方法在MIRFlickr-25K数据集上的mAP值比较

Table 3 Comparison of mAP values of different methods on MIRFlickr-25K dataset

Method	Image2Text			Text2Image		
	16 bit	32 bit	64 bit	16 bit	32 bit	64 bit
SCM	0.6157	0.6213	0.6268	0.6102	0.6284	0.6292
SePH	0.6481	0.6453	0.6596	0.6457	0.6476	0.6508
STMH	0.5877	0.5901	0.6001	0.5863	0.5877	0.5879
CMFH	0.5780	0.5827	0.5861	0.5784	0.5878	0.5889
DCMH	0.7219	0.7332	0.7450	0.7526	0.7576	0.7704
PRDH	0.7052	0.7125	0.7208	0.7607	0.7739	0.7784
CMHH	0.7302	0.7387	0.7444	0.7320	0.7283	0.7301
MFPMC	**0.7501**	**0.7608**	**0.7687**	**0.7764**	**0.7895**	**0.7898**

表4 不同方法在NUS-WIDE数据集上的mAP值比较

Table 4 Comparison of mAP values of different methods on NUS-WIDE dataset

Method	Image2Text			Text2Image		
	16 bit	32 bit	64 bit	16 bit	32 bit	64 bit
SCM	0.4905	0.4946	0.4995	0.4598	0.4660	0.4701
SePH	0.5324	0.5350	0.5529	0.5078	0.5095	0.5177
STMH	0.4354	0.4471	0.4544	0.3895	0.4098	0.4187
CMFH	0.3925	0.3958	0.3990	0.3956	0.3955	0.3978
DCMH	0.5257	0.5375	0.5458	0.5792	0.5875	0.5944
PRDH	0.5919	0.6058	0.6116	0.6155	0.6287	0.6349
CMHH	0.5530	0.5697	0.5559	0.5739	0.5786	0.5639
MFPMC	**0.6042**	**0.6196**	**0.6256**	**0.6246**	**0.6375**	**0.6437**

进一步为了证明所提方法的有效性，以哈希码长度为16 bit作为基准条件，在MIRFlickr-25K和NUS-

WIDE数据集上分别针对Image2Text和Text2Image任务绘制了准确率-查全率曲线，如图2～5所示，横坐标表示查全率(recall)，纵坐标表示准确率(precision)。

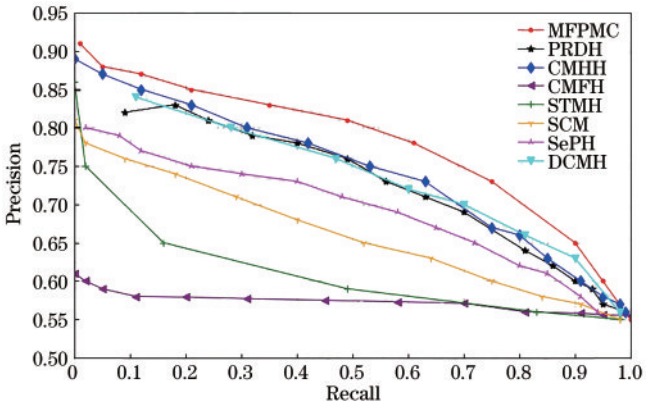

图2　哈希码长度为16 bit时，在MIRFlickr-25K数据集上Image2Text的P-R曲线

Fig. 2　P-R curves of Image2Text on MIRFlickr-25K dataset when Hash code length is 16 bit

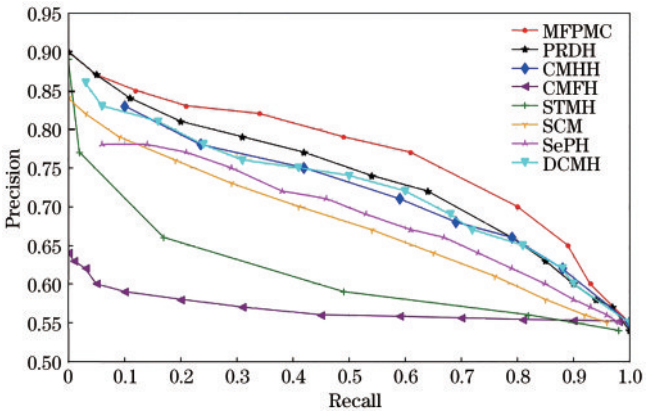

图3　哈希码长度为16 bit时，在MIRFlickr-25K数据集上Text2Image的P-R曲线

Fig. 3　P-R curves of Text2Image on MIRFlickr-25K dataset when Hash code length is 16 bit

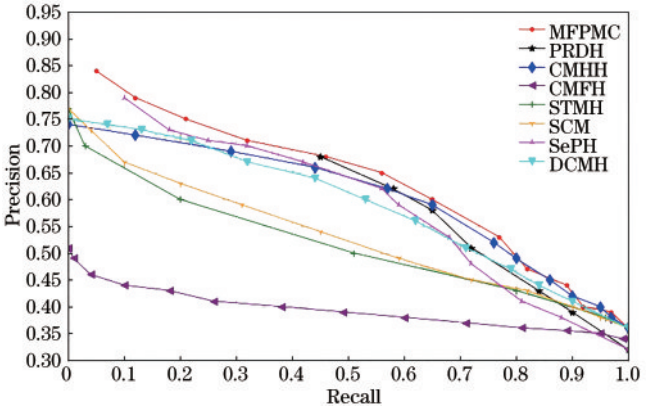

图4　哈希码长度为16 bit时，在NUS-WIDE数据集上Image2Text的P-R曲线

Fig. 4　P-R curves of Image2Text on NUS-WIDE dataset when Hash code length is 16 bit

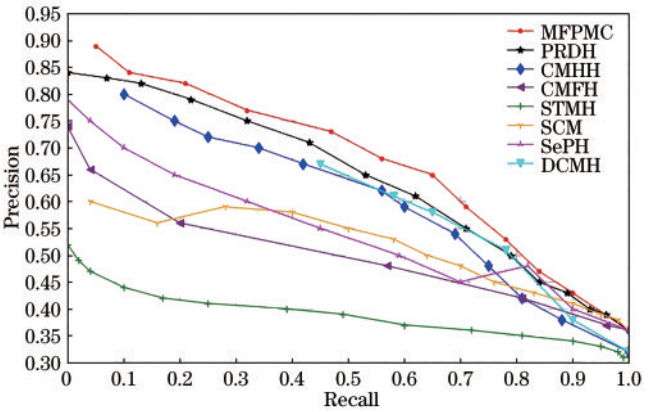

图5　哈希码长度为16 bit时，在NUS-WIDE数据集上Text2Image的P-R曲线

Fig. 5　P-R curves of Text2Image on NUS-WIDE dataset when Hash code length is 16 bit

在同一幅P-R曲线图中，曲线越靠外围且与其他曲线没有交叉，则该曲线所对应的方法性能越好；如果有交叉，则曲线与坐标轴围成的截面面积越大，对应方法的性能越好。从图2～5可以看出：无论是Image2Text任务还是Text2Image任务，无论是在MIRFlickr-25K数据集上还是在NUS-WIDE数据集上，所提方法对应的P-R曲线均在最外围，因此所提方法的检索性能优于其他方法，这与通过mAP指标得出的结论一致。

3.4　消融实验

为了进一步证明所提多尺度融合模型以及引入的低维特征投影匹配约束的有效性，当生成哈希码长度为64 bit时，在两个基准数据集上进行了消融实验。Base表示没有加入多尺度融合模型同时也没有引入低维特征投影匹配约束；IMFM表示图像多尺度融合模型；TMFM表示文本多尺度融合模型；LFPMC表示引入的低维特征投影匹配约束。消融实验的结果如表5所示。从表5可以看出：IMFM、TMFM和LFPMC对不同数据集的检索性能有不同的提升幅度。

1）通过表5中Base、Base ＋ IMFM、Base ＋ TMFM、Base ＋ IMFM ＋ TMFM的对比结果可以看出：在Image2Text任务上，相较于Base，加入图像多尺度融合模块、文本多尺度融合模块、图像＋文本多尺度融合模块在MIRFlickr-25K数据集上的mAP值分别提升了0.62个百分点、0.74个百分点、1.51个百分点，在NUS-WIDE数据集上的mAP值分别提升了1.04个百分点、0.97个百分点、2.03个百分点；在Text2Image任务上，加入这些模块在MIRFlickr-25K数据集上的mAP值分别提升了0.57个百分点、0.92个百分点、1.81个百分点，在NUS-WIDE数据集上的mAP值分别提升了1.64个百分点、2.29个百分点、4.03个百分点。由此可以证明：所提多尺度融合模型确实有助于跨模态哈希检索性能的提升。

表 5　消融实验的 mAP 值对比

Table 5　Comparison of mAP values of ablation experiments

Task	Method	MIRFlickr-25K	NUS-WIDE
Image2Text	Base	0.7250	0.5630
	Base＋IMFM	0.7312	0.5734
	Base＋TMFM	0.7324	0.5727
	Base＋IMFM＋TMFM	0.7401	0.5833
	Base＋LFPMC	0.7568	0.5974
	Base＋IMFM＋TMFM＋LFPMC	0.7687	0.6256
Text2Image	Base	0.7341	0.5605
	Base＋IMFM	0.7398	0.5769
	Base＋TMFM	0.7433	0.5834
	Base＋IMFM＋TMFM	0.7522	0.6008
	Base＋LFPMC	0.7698	0.6294
	Base＋IMFM＋TMFM＋LFPMC	0.7898	0.6437

2）通过表 5 中 Base ＋ IMFM ＋ TMFM、Base ＋ IMFM ＋ TMFM ＋ LFPMC 的对比结果可以看出：在 Image2Text 任务上，引入低维特征投影匹配约束后，在 MIRFlickr-25K、NUS-WIDE 数据集上的 mAP 值分别提升了 2.86 个百分点、4.23 个百分点；在 Text2Image 任务上，加入该模块在 MIRFlickr-25K、NUS-WIDE 数据集上的 mAP 值分别提升了 3.76 个百分点、4.29 个百分点。这一方面说明引入的低维特征投影匹配约束确实有助于跨模态哈希检索性能的提升，另一方面也可以间接证明低维特征中包含的丰富语义信息是不容忽视的。

3）通过表 5 中 Base、Base ＋ IMFM ＋ TMFM、Base ＋ LFPMC 的对比结果可以看出：在 Image2Text 任务上，相较于 Base，加入图像＋文本多尺度融合模块、引入低维特征投影匹配约束在 MIRFlickr-25K 数据集上的 mAP 值分别提升了 1.51 个百分点、3.18 个百分点，在 NUS-WIDE 数据集上的 mAP 值分别提升了 2.03 个百分点、3.44 个百分点；在 Text2Image 任务上，加入这些模块在 MIRFlickr-25K 数据集上的 mAP 值分别提升了 1.81 个百分点、3.57 个百分点，在 NUS-WIDE 数据集上的 mAP 值分别提升了 4.03 个百分点、6.89 个百分点。由此可以证明：相较于引入的多尺度融合模型，引入的低维特征投影匹配约束对跨模态哈希检索性能的提升更为显著。

3.5　参数分析

对于大多数基于深度学习的跨模态哈希方法来说，超参数的设置是影响模型性能的重要因素，因此需要选择合适的超参数以获得最佳性能。对此，对所提方法中涉及的重要超参数 ξ、τ、γ、η、μ 进行了分析，其中 ξ、τ、γ、η、μ 分别代表了低维特征投影匹配约束损失、模态内哈希码损失、模态间哈希码损失、量化损失、标签嵌入损失占总损失的权重。在 MIRFlickr-25K 数据集上，以哈希码长度为 64 bit 为基准条件，对 ξ、τ、γ、η、μ 5 个重要超参数进行了分析，并绘制了如图 6～10

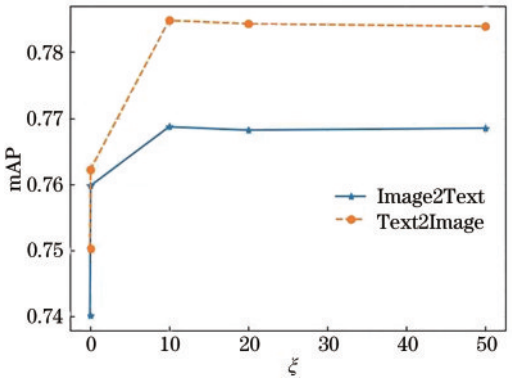

图 6　在 MIRFlickr-25K 数据集上超参数 ξ 对 mAP 的影响

Fig. 6　Influence of hyper-parameter ξ on mAP on MIRFlickr-25K dataset

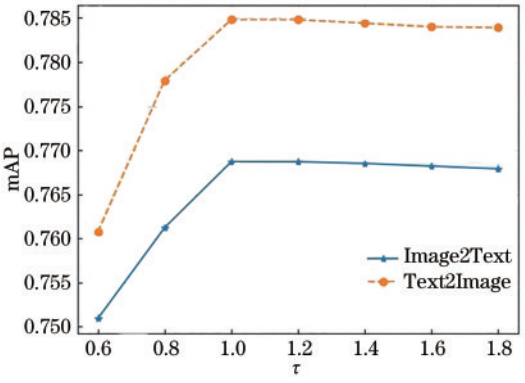

图 7　在 MIRFlickr-25K 数据集上超参数 τ 对 mAP 的影响

Fig. 7　Influence of hyper-parameter τ on mAP on MIRFlickr-25K dataset

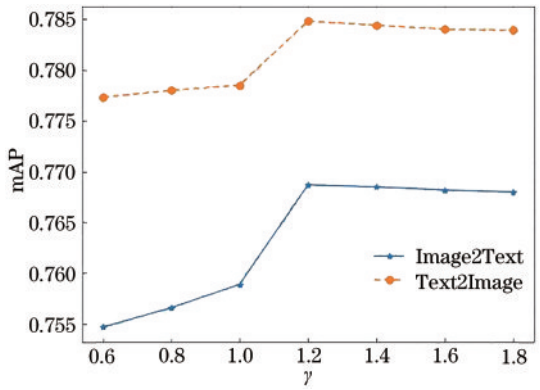

图 8 在 MIRFlickr-25K 数据集上超参数 γ 对 mAP 的影响

Fig. 8 Influence of hyper-parameter γ on mAP on MIRFlickr-25K dataset

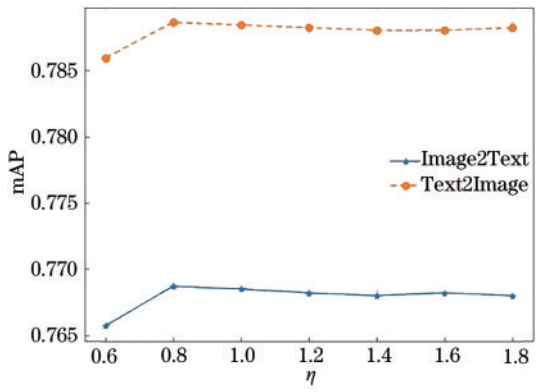

图 9 在 MIRFlickr-25K 数据集上超参数 η 对 mAP 的影响

Fig. 9 Influence of hyper-parameter η on mAP on MIRFlickr-25K dataset

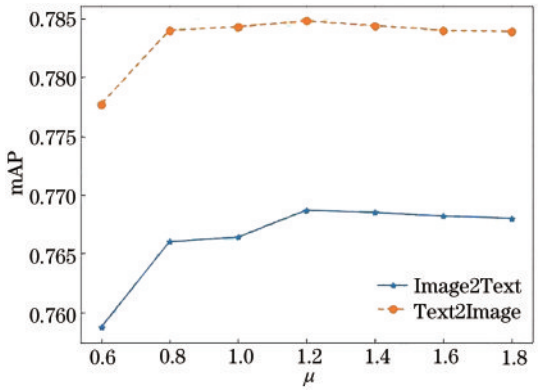

图 10 在 MIRFlickr-25K 数据集上超参数 μ 对 mAP 的影响

Fig. 10 Influence of hyper-parameter μ on mAP on MIRFlickr-25K dataset

所示的变化曲线。从图 6～10 可以看出：当超参数 ξ、τ、γ、η、μ 分别为 10、1.0、1.2、0.8、1.2 时，所提方法检索性能达到最佳。此外，这些超参数的值反映了各部分损失函数对于整个算法的重要程度，进一步可以得出：低维特征投影匹配约束损失对算法的影响最大，说明了低维特征投影匹配约束对跨模态哈希检索准确率的提升至关重要。

4 结 论

提出了一种基于多尺度融合和投影匹配约束的跨模态哈希算法，即 MFPMC。该方法主要针对大多数基于深度学习的跨模态哈希方法忽略单模态数据不同尺度包含不同语义信息和低维特征包含重要语义信息对缩小模态差异的重要性这两个问题。设计了基于图像特征训练网络和文本特征训练网络的多尺度融合模型，该模型可以提取图像模态的多尺度低维特征和文本模态的多尺度低维特征。在此基础上，为了使不同模态数据的低维特征之间的语义结构与真实语义空间的结构保持一致，引入了低维特征投影匹配约束和对抗训练。同时，构建模态内哈希码损失、模态间哈希码损失、量化损失来约束哈希码的学习，此外，为了进一步获得更具判别性的哈希码，还引入了标签嵌入技术。在两个基准数据集上进行实验，MFPMC 的检索性能均有所提升。但本文只探讨了图像与文本这两种模态之间的检索方法。在后续的工作中，将进一步改进所提检索算法，并将其应用到更多形式的多媒体数据中，包括图像、文本、音频和视频等。

参 考 文 献

[1] Zhu L, Tian G, Wang B, et al. Multi-attention based semantic deep hashing for cross-modal retrieval[J]. Applied Intelligence, 2021, 51(8): 5927-5939.

[2] Wu J, Xie X, Nie L, et al. Reconstruction regularized low-rank subspace learning for cross-modal retrieval[J]. Pattern Recognition, 2021, 113: 107813.

[3] Cheng Q R, Gu X D. Bridging multimedia heterogeneity gap via Graph Representation Learning for cross-modal retrieval[J]. Neural Networks, 2021, 134:143-162.

[4] Zhang J, Peng Y. Query-adaptive image retrieval by deep-weighted hashing[J]. IEEE Transactions on Multimedia, 2018, 20(9): 2400 - 2414.

[5] Ahmad J, Muhammad K, Baik S W. Medical image retrieval with compact binary codes generated in frequency domain using highly reactive convolutional features[J]. Journal of medical systems, 2018, 42(2): 1-19.

[6] Lu X, Song L, Xie R, et al. Deep binary representation for efficient image retrieval[J]. Advances in Multimedia, 2017, 2017:1-10

[7] Duan L, Zhao C, Miao J, et al. Deep hashing based fusing index method for large-scale image retrieval[J]. Applied Computational Intelligence and Soft Computing, 2017, 2017:250-257.

[8] Ye D, Li Y, Tao C, et al. Multiple feature hashing learning for large-scale remote sensing image retrieval[J]. ISPRS International Journal of Geo-Information, 2017, 6(11): 364.

[9] Ding G G, Guo Y C, Zhou J L, et al. Large-scale cross-modality search via collective matrix factorization hashing [J]. IEEE Transactions on Image Processing, 2016, 25(11): 5427-5440.

[10] Wang D, Gao X B, Wang X M, et al. Semantic topic multimodal hashing for cross-media retrieval[C]// Proceedings of the 24th International Conference on Artifificial Intelligence, July 25-31, 2015, Buenos Aires, Argentina. New York: AAAI Press, 2015: 3890-3896.

[11] Zhang D Q, Li W J. Large-scale supervised multimodal hashing with semantic correlation maximization[C]// Proceedings of the 28th AAAI Conference on Artificial Intelligence, July 27-31, 2014, Quebec City, Quebec, Canada. New York: AAAI Press, 2014: 2177-2183.

[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, June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 3864-3872.

[13] Gao J, Zhang W, Zhong F, et al. UCMH: unpaired cross-modal hashing with matrix factorization[J]. Neurocomputing, 2020, 418: 178-190.

[14] Xiong H, Ou W, Yan Z, et al. Modality-specific matrix factorization hashing for cross-modal retrieval[J]. Journal of Ambient Intelligence and Humanized Computing, 2020: 1-15.

[15] Jiang Q Y, Li W J. Deep cross-modal hashing[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 3270-3278.

[16] Yang E, Deng C, Liu W, et al. Pairwise relationship guided deep hashing for cross-modal retrieval[C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA. USA: IEEE Press, 2017: 1618-1625.

[17] Cao Y, Liu B, Long M S, et al. Cross-modal hamming hashing[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11205: 207-223.

[18] Liu X, Cheung Y, Hu Z, et al. Adversarial tri-fusion hashing network for imbalanced cross-modal retrieval[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2020, 5(4): 607-619.

[19] Wang X, Zou X, Bakker E M, et al. Self-constraining and attention-based hashing network for bit-scalable cross-modal retrieval[J]. Neurocomputing, 2020, 400: 255-271.

[20] Yan C, Bai X, Wang S, et al. Cross-modal hashing with semantic deep embedding[J]. Neurocomputing, 2019, 337: 58-66.

[21] 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, September, 2014, Nottingham, UK. London: British Machine Vision Association, 2014: 1-5.

[22] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6230-6239.

[23] 刘可文, 房攀攀, 熊红霞, 等. 基于多级特征的行人重识别[J]. 激光与光电子学进展, 2020, 57(8): 081503.
Liu K W, Fang P P, Xiong H X, et al. Person re-identification based on multi-layer feature[J]. Laser & Optoelectronics Progress, 2020, 57(8): 081503.

[24] 李聪, 蒋敏, 孔军. 基于多尺度注意力机制的多分支行人重识别算法[J]. 激光与光电子学进展, 2020, 57(20): 201001.
Li C, Jiang M, Kong J. Multi-branch person re-identification based on multi-scale attention[J]. Laser & Optoelectronics Progress, 2020, 57(20): 201001.

[25] 李思瑶, 刘宇红, 张荣芬. 多尺度特征融合的细粒度图像分类[J]. 激光与光电子学进展, 2020, 57(12), 121002.
Li S Y, Liu Y H, Zhang R F. Fine-grained image classification based on multi-scale feature fusion[J]. Laser & Optoelectronics Progress, 2020, 57(12): 121002.

[26] Li C, Deng C, Li N, et al. Self-supervised adversarial hashing networks for cross-modal retrieval[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City. New York: IEEE Press, 2018: 4242-4251.

[27] Bronstein M M, Bronstein A M, Michel F, et al. Data fusion through cross-modality metric learning using similarity-sensitive hashing[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 13-18, 2010, San Francisco, CA, USA. New York: IEEE Press, 2010: 3594-3601.

[28] Kumar S, Udupa R. Learning hash functions for cross-view similarity search[C]//Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, July 16-22, 2011, Barcelona, Catalonia, Spain. New York: AAAI Press, 2011: 1360-1365.