软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(8):3739−3751 [doi: 10.13328/j.cnki.jos.006960]

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

基于锚点的无监督跨模态哈希算法*

胡 鹏[1], 彭 玺[1], 彭德中[1,2]

[1](四川大学 计算机学院, 四川 成都 610065)

[2](成都瑞贝英特信息技术有限公司, 四川 成都 610094)

通信作者: 彭玺, E-mail: pengx.gm@gmail.com

摘 要: 基于图的无监督跨模态哈希学习具有存储空间小、检索效率高等优点, 受到学术界和工业界的广泛关注, 已成为跨模态检索不可或缺的工具之一. 然而, 图构造的高计算复杂度阻碍其应用于大规模多模态应用. 主要尝试解决基于图的无监督跨模态哈希学习面临的两个重要挑战: 1) 在无监督跨模态哈希学习中如何高效地构建图? 2) 如何解决跨模态哈希学习中的离散值优化问题? 针对这两个问题, 分别提出基于锚点图的跨模态学习和可微分哈希层. 具体地, 首先从训练集中随机地选择若干图文对作为锚点集, 利用该锚点集作为中介计算每批数据的图矩阵, 以该图矩阵指导跨模态哈希学习, 从而能极大地降低空间与时间开销; 其次, 提出的可微分哈希层可在网络前向传播时直接由二值编码计算, 在反向传播时亦可产生梯度进行网络更新, 而无需连续值松弛, 从而具有更好的哈希编码效果; 最后, 引入跨模态排序损失, 使得在训练过程中考虑排序结果, 从而提升跨模态检索正确率. 通过在 3 个通用数据集上与 10 种跨模态哈希算法进行对比, 验证了提出算法的有效性.

关键词: 无监督哈希学习; 跨模态检索; 锚点图; 可微分哈希; 公共汉明空间

中图法分类号: TP301

中文引用格式: 胡鹏, 彭玺, 彭德中. 基于锚点的无监督跨模态哈希算法. 软件学报, 2024, 35(8): 3739–3751. http://www.jos.org.cn/1000-9825/6960.htm

英文引用格式: Hu P, Peng X, Peng DZ. Anchor-based Unsupervised Cross-modal Hashing. Ruan Jian Xue Bao/Journal of Software, 2024, 35(8): 3739–3751 (in Chinese). http://www.jos.org.cn/1000-9825/6960.htm

Anchor-based Unsupervised Cross-modal Hashing

HU Peng[1], PENG Xi[1], PENG De-Zhong[1,2]

[1](College of Computer Science, Sichuan University, Chengdu 610065, China)

[2](Chengdu Ruibei Yingte Information Technology Co. Ltd., Chengdu 610094, China)

Abstract: Thanks to the low storage cost and high retrieval speed, graph-based unsupervised cross-modal hash learning has attracted much attention from academic and industrial researchers and has been an indispensable tool for cross-modal retrieval. However, the high computational complexity of graph structures prevents its application in large-scale multi-modal applications. This study mainly attempts to solve two important challenges facing graph-based unsupervised cross-modal hash learning: 1) How to efficiently construct graphs in unsupervised cross-modal hash learning? 2) How to handle the discrete optimization in cross-modal hash learning? To address such two problems, this study presents anchor-based cross-modal learning and a differentiable hash layer. To be specific, the study first randomly samples some image-text pairs from the training set as anchor sets and uses the anchor sets as the agent to compute the graph matrix of each batch of data. The graph matrix is used to guide cross-modal hash learning, thus remarkably reducing the space and time cost; second, the proposed differentiable hash layer directly adopts binary coding for computation during network forward propagation and produces gradient to update the network without continuous-value relaxation during backpropagation, thus embracing better hash encoding

* 基金项目: 国家自然科学基金 (62102274, 62176171, U21B2040 U19A2078); 四川省科技计划 (2021YFS0389, 2022YFQ0014, 2022YFSY0047, 2022YFH0021); 中央高校基本科研业务费专项资金 (YJ202140); 中国博士后科学基金 (2021M692270)
收稿时间: 2021-08-30; 修改时间: 2022-10-13; 采用时间: 2023-04-28; jos 在线出版时间: 2023-09-06
CNKI 网络首发时间: 2023-09-07

performance. Finally, the study introduces cross-modal ranking loss to consider the ranking results in the training process and improve the cross-modal retrieval accuracy. To verify the effectiveness of the proposed algorithm, the study compares the algorithm with 10 cross-modal hash algorithms on three general data sets.

Key words: unsupervised hashing learning; cross-modal retrieval; anchor graph; differentiable hashing; common Hamming space

1 引 言

随着互联网和多媒体技术的飞速发展, 网络上迅速产生了大量的多媒体数据. 如何从这些海量的多媒体数据中检索到人们感兴趣的知识, 具有广泛的应用前景, 但同时也是一个巨大挑战. 跨模态检索是用一个模态中的数据查询另一模态检索库中与之语义相关的数据, 其灵活的检索方式吸引了众多学术界和工业界研究者的关注. 但由于不同模态的数据类型和结构上存在巨大的差异 (被称为“异构鸿沟”), 导致无法直接度量不同模态间的相似性, 为跨模态检索带来巨大挑战. 为了削减甚至消除异构鸿沟, 许多跨模态检索方法在过去多年被提出[1−4]. 然而, 这些方法大多是连续值方法, 其在大规模跨模态检索中面临着计算和存储成本高的问题. 因此, 如何在弥合跨模态差异的同时, 降低跨模态的存储与检索开销具有重要的理论意义和应用价值.

近年来, 跨模态哈希学习被成功用于压缩特征存储大小和降低检索复杂度. 一方面, 为降低特征的存储空间, 跨模态哈希学习将不同的模态映射到一个公共汉明空间中, 在该空间中不同模态的数据可由二值编码进行表示[5−7]. 另一方面, 为提高检索效率, 样本间的相似度可以直接由汉明距离高效计算得到, 而汉明距离可以用位运算 (即: 异或运算) 代替浮点运算进行计算[8−10]. 按照是否需要语义监督信息, 现有的跨模态哈希学习主要可以分为有监督的跨模态哈希学习方法[7,11−13]和无监督的跨模态哈希学习方法[6,10,14]. 有监督的跨模态哈希学习方法在良好标注的语义信息的指导下将不同的模态映射到一个公共的汉明空间中, 由于具有语义信息的指导, 这些方法往往可取得良好的效果. 然而, 对大规模的数据进行标注是费时且昂贵的[15,16], 并且同时标注多个模态将成倍地增加标注成本. 因此, 无监督的跨模态哈希学习方法近年来受到国内外研究者的密切关注, 此类方法可从大量易于获取的未标记数据中学习到跨模态判别信息, 其灵活的低成本学习方式具有很高的应用价值. 本文主要研究无监督的跨模态哈希学习方法.

无监督跨模态哈希学习主要利用图文对中成对的相关信息弥合跨模态差异, 进而将不同的模态映射到一个公共的汉明空间中[7,11]. 尽管已有的无监督方法取得了显著进展, 然而大多方法主要利用图文对的相关性, 往往忽略了多模态数据中潜在的流形结构信息[11,17]. 为了挖掘多模态数据中潜藏的结构信息, 近年来一些基于图的跨模态哈希算法被提出并取得了较好的性能[12,17]. 但是, 这些基于图的跨模态哈希方法需要在整个训练集上构建图矩阵, 具有很高的时间复杂度 $O(kN^2)$ 和空间复杂度 $O(N^2)$, 其中 N 为训练集的图文对个数、 k 为任意点的最近邻个数. 因此, 现有的基于图的跨模态哈希算法难以高效应对大规模多模态数据. 此外, 由于直接优化二值编码是一个 NP 难题 (NP-hard problem), 为解决该问题, 现有方法主要采用: (1) 连续值松弛[12,17,18], 将二值编码由连续值代替进行优化, 该松弛会导致训练的目标与优化方式不一致, 使得算法性能下降. (2) 逼近二值编码[14,19], 虽然该类方法的优化方式与哈希目标一致, 但是在优化过程中依然存在松弛问题, 即在训练过程中算法的输出依然为连续值参与优化, 也同样会使算法性能退化.

针对上述问题, 本文提出了基于锚点的跨模态哈希算法 (anchor-based unsupervised cross-modal hashing, AUCMH), 该方法无需在整个训练集上构建图, 因此具有更低的时间和空间复杂度. 如图 1 所示, 其中 CNN 与 BoW 分别为对应的骨干网络和特征提取器, $\mathcal{L}_g$ 为基于锚点图的跨模态损失, $\mathcal{L}_r$ 为跨模态三元排序损失. 首先, 为解决构建图的成本过高问题, 本文提出一种基于锚点的跨模态图方法. 该方法仅利用少量的跨模态锚点样本构建锚图, 然后利用锚图构建每批数据的子图, 该子图为整个训练集上的一个子图, 其构建的时间复杂度 $O(kmn)$ 和空间复杂度 $O(mn)$ 均远低于在整个训练集上构建的复杂度, 其中 n 为一批数据中图文对的个数, m 为锚点集中图文对的个数, 且 $n < m \ll N$. 其次, 为解决跨模态哈希学习中的离散优化问题, 本文提出一种可微分哈希层. 该神经层通过强制网络的输出为二值编码, 可使在网络优化过程中直接对二进制编码进行优化, 而不需连续值松弛, 从而使得推理与训练保持一致, 进而改善跨模态哈希的性能. 最后, 为让跨模态哈希学习与下游任务 (即: 跨模态检索) 一致, 本文引入三元排序损失使得在整个跨模态哈希学习过程中考虑跨模态排序, 从而引导跨模态哈希学习与下游跨模态检索任务保持一致.

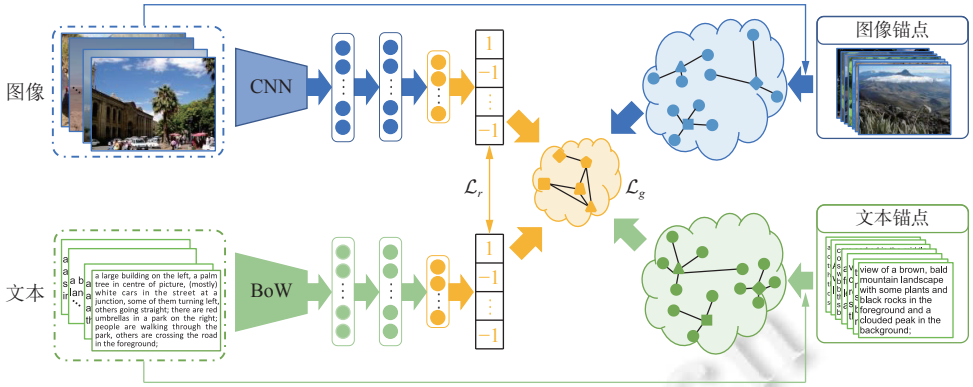


图 1 算法整体框架图

本文主要贡献可概括如下.

● 本文提出了一种基于锚点的无监督跨模态哈希学习方法. 该方法可高效地从跨模态数据中学到一个公共汉明空间, 从而在该空间中实现跨模态检索. 文章通过大量实验证明了提出算法的有效性.

● 为解决哈希学习中二值编码不可微问题, 本文提出了一种可微分哈希层. 该神经层能使得神经网络前向传播时采用离散方式进行计算, 且可使用反向传播进行神经网络权值更新. 由于该方法没有采用连续值松弛的方式, 故可取得更好的性能.

● 本文提出了一种新颖的基于锚点的子图构造方法, 该方法可利用少量的锚点构造子图, 从而指导跨模态哈希学习. 相较传统的基于图的方法, 该方法可极大地降低时间与空间开销.

2 相关工作

近几十年来, 无数的研究者致力于充分挖掘出多模态数据中的跨模态相关性, 从而消除不同模态之间的差异, 以便将不同的模态投影到一个公共的汉明空间中. 本节将从有监督和无监督两个方面对跨模态哈希学习进行回顾, 以简要介绍目前跨模态哈希学习的研究现状.

2.1 有监督的跨模态哈希学习

有监督的跨模态哈希学习旨在利用良好标注的多模态标签数据指导模型学习特定模态的映射, 同时将不同模态投影到一个公共的汉明空间中[5,10,14,19]. 由于排序信息有利于构建检索系统, 因此很多工作将排序损失引入到跨模态哈希学习中, 强制具有相同语义的跨模态对具有更高的排序[6,20,21]. 具体地, 文献 [20] 提出了一种基于排序的哈希框架, 通过共同学习两组线性子空间 (每个模态一组), 对特征在不同线性子空间中的排序顺序最大程度地保留, 从而将来自不同模态的数据映射到一个公共汉明空间, 以便在该空间中可使用汉明距离测量跨模态相似性. 在文献 [21] 中, 作者提出了一种基于排名的深度跨模态散列方法 (ranking-based deep cross-modal hashing, RDCMH), 该方法首先利用数据的特征和标签信息推导出半监督语义排序表, 然后将该语义排序信息集成到深度跨模态哈希中, 并联合优化深度表征和哈希函数的参数. 文献 [22] 提出了一种保持排序的哈希 (rank-order preserving hashing, RoPH) 方法, 该方法具有一种基于回归的排序保留损失, 且该损失函数具有大边距特性并易于优化. 文献 [6] 基于三元组提出了一种用于跨模态检索的深度哈希网络 (triplet-based deep hashing, TDH), 该方法利用三元组标签描述 3 个实例之间的相对关系作为监督, 以捕获跨模态实例之间更一般的语义相关性. 为了解决无法在训练时同时获得所有模态的问题, 文献 [14] 提出了一种可分离变分哈希网络方法 (separated variational hashing networks, SVHN), 该方法分别从各自的模态数据中学习到二值语义编码, 因此各模态可以在空间或时间上独立的情况下完成模型训练. 文献 [23] 提出了一种模态不变非对称网络 (modality-invariant asymmetric networks, MIAN) 架构, 该架构探索了概率模态对齐框架下的不对称模态内和模态间相似性保持. 文献 [24] 提出了一种具有双向关系推理

的深度归一化跨模态哈希方案, 该方案通过考虑双向关系, 即一致和不一致关系来构建多级语义相似度矩阵. 虽然这些有监督方法利用良好标注的跨模态数据能取得不错的性能, 但是其主要依赖于大量良好标注的数据. 由于数据标注是一项昂贵且费时的任务, 故限制了有监督方法在真实场景下的应用范围.

2.2 无监督跨模态哈希学习

无监督跨模态哈希学习通过利用跨模态相关性从无类别标注的多模态数据中学习到统一的二值编码. 由于其无需对多模态数据进行类别标注, 故具有更广泛的应用场景, 深受研究者的喜爱[12,17,25,26]. 文献 [18] 提出了一种跨视图哈希 (cross-view hashing, CVH) 方法, 通过使用多模态数据的模态内和模态间相似性为每个模态学习公共的哈希码. 文献 [11] 提出了一种融合相似哈希方法 (fusion similarity hashing, FSH). 该方法将跨模态的基于图的融合相似性嵌入到公共汉明空间中, 进而在此基础上引入具有交替优化的图哈希方案, 以学习嵌入这种融合相似性的二进制编码. 但是上述方法均为浅层方法, 它们难以捕获多模态数据中的高阶非线性语义, 难以获得理想的效果. 为解决该问题, 近年来一些基于深度神经网络 (deep neural network, DNN) 的方法被提出. 这些方法利用 DNN 的高度非线性来捕获跨模态数据中的高层语义信息, 从而获得当前最优的效果. 例如, 文献 [26] 提出无监督耦合循环生成对抗式哈希网络 (unsupervised coupled cycle generative adversarial hashing networks, UCH) 通过使用外循环和内循环网络来学习统一的二值表征. 为利用输入数据中内在的结构信息, 文献 [12] 提出了一种无监督的生成对抗式跨模态散列 (unsupervised generative adversarial cross-modal hashing, UGACH) 方法, 该方法可充分利用生成对抗网络 (generative adversarial network, GAN) 的无监督表示学习能力来挖掘跨模态数据的潜在流形结构信息, 从而得到更好的公共哈希编码. 类似地, 文献 [17] 提出了一种用于无监督跨模态检索的多路径生成对抗式哈希方法 (multi-pathway generative adversarial hashing, UGACH), 该方法充分利用了生成对抗网络的无监督表示学习能力, 以挖掘跨模态数据中潜在流形结构信息. 文献 [27] 提出了一种新的自适应跨模态相关挖掘和结构语义维护策略以学习二值表示, 并设计了一种非对称的保持相似性的二值优化算法, 以减少二值化后的信息损失. 但是, 这些基于图的方法需要预先在整个训练集上构建图, 具有非常高的时间和空间复杂度, 难以用于大规模数据.

3 基于锚的跨模态哈希算法

本节将详细介绍提出的基于图的无监督跨模态哈希方法. 具体内容包括对跨模态哈希的问题描述, 其次是介绍提出的可微分哈希层, 之后介绍如何基于锚点图进行跨模态学习, 最后是结合排序的跨模态哈希学习.

3.1 问题描述

为更清楚地描述跨模态哈希问题, 本文有如下定义: 令 $\mathcal{D}=\{i_j,t_j\}_{j=1}^{N}$ 表示一个具有 N 个图文对的跨模态数据集, 其中 i_j 表示第 j 个图像样本, t_j 表示第 j 个文本样本. t_j 通常为 i_j 的文本描述或文本标签, 共同构成一个图文对.

跨模态哈希学习的目的是将不同的模态样本以二进制进行编码, 因此需要从不同的模态数据中学习到特定于模态的映射函数分别将图像模态 $\mathcal{I}=\{i_j\}_{j=1}^{N}$ 和文本模态 $\mathcal{T}=\{t_j\}_{j=1}^{N}$ 投影到一个公共的判别汉明空间中, 即 $\mathcal{I}\xrightarrow{f_i}H_i$ 和 $\mathcal{T}\xrightarrow{f_t}\mathcal{H}_t$, 其中 $f_i(\cdot;\Theta_i)$ 和 $f_t(\cdot;\Theta_t)$ 分别为图像和文本的哈希函数; Θ_i 和 Θ_t 分别为这些函数对应的网络参数; $\mathcal{H}_{\rangle}=\left\{h_j^i\right\}_{j=1}^{N}$ 和 $\mathcal{H}_{\sqcup}=\left\{h_j^t\right\}_{j=1}^{N}$ 分别为图像和文本模态的二值表征; $h_j^i=f_i(i_j)\in\{-1,+1\}^L$; $h_j^t=f_t(t_j)\in\{-1,+1\}^L$; L 为哈希编码的长度. 在学到的该汉明空间里, 本文希望具有相同语义的样本尽可能地靠近, 否则尽可能地相互远离, 即相关样本比不相关样本具有更大的相似度. 为方便计算, 本文使用内积 $<h^i,h^t>$ 来计算汉明距离: $d(h^i,h^t)=\frac{1}{2}(L-\langle h^i,h^t\rangle)$. 相应地, 可用内积 $<h^i,h^t>$ 来计算 h^i 和 h^t 之间的相似度. 此外, 对于 L 长度的二进制编码向量 h, 其二范数恒为一个常数, 即 $\|h\|\equiv\sqrt{L}$, 因此也可用余弦相似度来计算不同哈希码的相似度: 即 $\cos(x,y)=\frac{\langle h^i,h^t\rangle}{\|h^i\|\|h^t\|}=\frac{1}{L}\langle h^i,h^t\rangle$, 其中 $\|x\|$ 指 x 的二范数.

3.2 可微分哈希层

在本文中, 不同模态的神经网络由多层全连接构成, 除了最后一层网络, 其他每层全连接层之后接一层 ReLU.

最后一层全连接层接本文提出的可微分哈希层. 该可微分哈希层由 tanh 和符号函数 $\mathrm{sign}(\cdot)$ 构成. 具体地, 以一个样本为例, 将其输入神经网络得到最后一层全连接的输出, 再将该输出输入 tanh 将其限定到 $[-1,1]$ 之间, 然后再将 tanh 输出向量进行归一化后输入 $\mathrm{sign}(\cdot)$ 得到二值化的编码. $\mathrm{sign}(\cdot)$ 函数的定义如下:

$$\mathrm{sign}(x)=\begin{cases}+1, & x>0\\ -1, & x\leqslant 0\end{cases} \tag{1}$$

由于符号函数 $\mathrm{sign}(\cdot)$ 的不可导使得网络不能直接采用梯度下降法进行优化. 为解决该问题, 本文采用直通式估算器 (straight-through estimator, STE)[28,29]进行优化. 具体地, 在网络的前向传播时, 我们直接将 $\mathrm{sign}(\cdot)$ 函数加到网络的输出层后, 以二值化网络的输出 x, 然后利用二值化后的表征 $\mathrm{sign}(x)$ 计算损失以优化网络; 在反向传播时, 由于 $\mathrm{sign}(x)$ 不能够直接优化, 因此我们令 $\dfrac{\partial \mathrm{sign}(x)}{\partial x}=1$ 以计算梯度, 从而可利用梯度下降法更新网络. 换言之, 在反向传播时, 我们视 $\mathrm{sign}(x)\approx x$ 进行梯度计算.

该层网络接到各模态网络的最后一层全连接层之后, 旨在将全连接层的输出转化为哈希编码. 与现有的跨模态哈希算法不同, 本文的可微分哈希层没有采用连续值松弛或哈希逼近的方式, 而是在整个网络的优化过程中均是哈希值直接参与运算. 从而, 本文的方法能更加准确地计算出在汉明空间中的损失值, 获得更好的性能 (如第 4.4 节消融实验所示).

3.3 基于锚点图的跨模态学习

为利用各模态数据中的固有结构信息, 现有的无监督跨模态哈希方法一般在整个训练集上构建一个结构图, 但是该计算方式具有很高的时间 $\mathrm{O}(kN^2)$ 和空间复杂度 $\mathrm{O}(N^2)$, 其中 k 为任意点的最近邻个数. 因此, 限制了该类方法处理大规模的跨模态数据. 受文献 [30,31] 启发, 本文提出了一种基于锚点的图学习方法, 从训练集中采样少量数据作为锚点来构建图, 从而可大大地降低构建图的开销.

首先, 从训练集中随机地选择 m 个图文对作为锚集 $\mathcal{A}=\left\{i_{A_j},t_{A_j}\right\}_{j=1}^{m}=\left\{a_j^i,a_j^t\right\}_{j=1}^{m}$, 其中 $\{A_j\}_{j=1}^{m}$ 是不重复的随机索引, 用于从多模态数据集中随机地选取 m 个图文对作为锚集. 此外, 为处理大规模多模态数据, 本文采用批处理的训练方式对神经网络进行优化. 具体地, 从跨模态数据集 $\mathcal{D}$ 中随机地选取 n 个图文对, 构成一个批次的图文对 $\mathcal{B}=\{i_{B_j},t_{B_j}\}_{j=1}^{n}$, 其中 $\{B_j\}_{j=1}^{n}$ 为数据集 $\mathcal{D}$ 中 n 个不重复的随机索引, 被用于从训练集中随机地采样一小批数据. 为简化表述, 本文将该批次写为 $\mathcal{B}=\{i_j,t_j\}_{j=1}^{n}$.

通过上述锚点可构造该批数据的子图. 具体地, 首先采用余弦相似度计算批数据中的任意数据点 x_l 与锚点 $a_j^x(x\in\{i,t\})$ 的相似度:

$$S_{jl}^{x}=\begin{cases}\cos\left(a_j^x,x_l\right), & a_j^x\in\mathcal{N}_k(x_l)\\ 0, & \text{其他}\end{cases} \tag{2}$$

其中, $\mathcal{N}_k(x)$ 表示 x 在 $\mathcal{A}$ 中的 k 个最近邻样本集, 相似度矩阵 $S^x=[S_1^x,S_2^x,\ldots,S_n^x]\in R^{m\times n}$. 然后, 利用 S^x 计算批数据内的相似性子图 W^x:

$$W_{jl}^{x}=\cos\left(S_j^x,S_l^x\right) \tag{3}$$

对该相似性子图 W^x 进行归一化, 则有 $G^x=(D^x)^{-1}W^x$, 其中 D^x 为对角阵, 且其对角元素满足 $D_{jj}^{x}=\sum_{l=1}^{n}W_{jl}^{x}$.

最后, 利用该批数据中的图像和文本的相似图可以构造一个公共图:

$$P=\frac{1}{2}(G^i+G^t) \tag{4}$$

显然, 该公共图满足 $\sum_{l=1}^{n}P_{jl}=1$. 因此, P_{jl} 可以视为第 j 个图文对 $\{i_j,t_j\}$ 与第 l 个图文对 $\{i_l,t_l\}$ 的相关概率. 相应地, 可计算出在该批数据中的跨图像和文本的相关概率:

$$Q_{jl}^{it}=\frac{\exp\left(\cos\left(h_j^i,h_l^t\right)\right)}{\sum_{p=1}^{n}\exp\left(\cos\left(h_j^i,h_p^t\right)\right)} \tag{5}$$

类似地, 不难得到 Q^{ti}. 由于期望公共汉明空间中的相似图或概率应与数据固有的相似图逼近, 即 Q^{ii} 和 Q^{tt} 逼近 P. 故可利用 KL 散度 (Kullback-Leibler divergence) 以度量汉明空间的相似图与输入空间的相似图之间的距离, 即: $\mathcal{L}' = D_{KL}(P\|Q^{it}) + D_{KL}(P\|Q^{ti}) = H(P,Q^{it}) + H(P,Q^{ti}) - 2H(P)$, 其中 $H(x)$ 为 x 的熵, 且 $H(P)$ 为常数. 因此, 可以得到如下损失函数:

$$\mathcal{L}_g = -\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\left(P_{jl}\log Q_{jl}^{it} + P_{jl}\log Q_{jl}^{ti}\right) \tag{6}$$

3.4 跨模态排序学习

为使跨模态哈希学习与其下游跨模态检索任务保持一致, 本文在哈希学习过程中引入跨模态排序. 具体地, 本文采用三元排序损失 (triplet ranking loss) 使得相关样本的相似度始终比非相关样本之间的相似度更大, 从而捕获跨模态数据中的排序信息. 首先, 定义一个三元排序损失集: $\mathcal{R}^{it} = \left\{g\left(h_j^i, h_l^t\right)g\left(h_j^i, h_l^t\right) > 0; j \neq l; j,l = 1,2,\ldots,n\right\}$, 其中 $g\left(h_j^i, h_l^t\right) = \gamma + \cos\left(h_j^i, h_j^t\right) - \cos\left(h_j^i, h_l^t\right)$, γ $(0 < \gamma < 1)$ 为一实数, 用以限制相关样本相似度与非相关样本相似度之差不低于 γ, 从而保证在跨模态排序中相关样本始终排在非相关样本之前. 类似地, 不难得到 $\mathcal{R}^{ti}$. 基于上述目标, 可得到以文搜图和以图搜文的平均三元排序损失如下:

$$\mathcal{L}_r = \frac{1}{|\mathcal{R}^{it}|}\sum_{j=1}^{|\mathcal{R}^{it}|}\mathcal{R}_j^{it} + \frac{1}{|\mathcal{R}^{ti}|}\sum_{j=1}^{|\mathcal{R}^{ti}|}\mathcal{R}_j^{ti} \tag{7}$$

其中, $|\mathcal{R}|$ 指 $\mathcal{R}$ 中元素的个数.

3.5 算法优化

综上所述, 算法的一次前向传播过程是: 随机采样一批跨模态数据输入神经网络经可微哈希层输出二值编码, 之后通过公式 (6) 和公式 (7) 计算各自损失, 同时以一个平衡因子 β $(0 < \beta < 1)$ 进行组合得到最终损失值:

$$L = \beta\mathcal{L}_g + (1-\beta)\mathcal{L}_r \tag{8}$$

最终, 利用梯度下降法 (如: ADAM[32]) 进行参数更新, 优化过程如算法 1 所示.

算法 1. AUCMH 的网络优化过程.

输入: 图文对训练集 $D = \{x_i, y_i\}_{i=1}^{N}$、哈希编码长度 L、批大小 n、平衡参数 β、约束参数 γ、锚点个数 m、最近邻个数 k 以及学习率 α;

1. 随机初始化不同模态的网络参数 Θ_i, Θ_t.
2. 随机地从 D 中选择 m 个图文对,构成一个锚点集 $\mathcal{A} = \left\{i_{A_j}, t_{A_j}\right\}_{j=1}^{m} = \left\{a_j^i, a_j^t\right\}_{j=1}^{m}$.
3. **while** 未收敛 **do**
4. 随机地从 D 中选择 n 个图文对,构成一个批次的图文对 $\mathcal{B} = \left\{i_{B_j}, t_{B_j}\right\}_{j=1}^{n} = \{i_j, t_j\}_{j=1}^{n}$.
5. 对于每个模态, 用 k 近邻的方式计算 $\mathcal{B}$ 与 $\mathcal{A}$ 的相似度图, 得到 G^i 和 G^t, 如公式 (2) 和公式 (3) 所示.
6. 利用计算得到的 G^i 和 G^t, 计算公共图 $P = \frac{1}{2}(G^i + G^t)$.
7. 利用 P 和公式 (5), 公式 (6), 计算基于锚点图的跨模态损失 $\mathcal{L}_g$.
8. 利用公式 (7) 计算跨模态三元排序损失 $\mathcal{L}_r$.
9. 采用随机梯度下降法, 最小化公式 (8) 中的损失 $\mathcal{L}$ 以对网络参数进行更新:
   $\Theta_x = \Theta_x - \alpha(\beta\nabla_{\Theta_x}(\mathcal{L}_g) + (1-\beta)\nabla_{\Theta_x}(\mathcal{L}_r))(x \in \{i,t\})$.
10. **end while**

输出: 优化得到的 AUCMH 模型.

4 实验探究

本文分别在 3 个广泛使用的多模态数据集 (即, MIRFLICKR-25K[33]、IAPR TC-12[34] 和 NUS-WIDE[35]) 上展

开实验从而评估提出的 AUCMH 的有效性.

4.1 数据集

本文采用文献 [26] 一样的数据集划分方法. 具体地, 将每个数据集分为检索库和查询集两部分, 其中检索库与查询集没有交集. 值得注意的是, 对于无监督方法, 检索库的所有数据均作为其训练集; 而对于有监督的对比算法, 将从检索库中随机选取 5 000 个图文对作为其训练集; 此外, 随机地从训练集中选取 2 000 个图文对作为验证集. 各数据集的统计信息见表 1. 接下来, 对不同数据集进行简要介绍.

表 1 本文所使用的数据集的统计信息

数据集	类别数	检索库	训练集		查询集	模态	特征维度和类型
			有监督	无监督			
MIRFLICKR-25K	24	18 015	5 000	18 015	2 000	图像文本	4 096维 VGG, 1 386维 BoW
IAPR TC-12	255	18 000	5 000	18 000	2 000	图像文本	4 096维 CNN-F, 2 912维 BoW
NUS-WIDE	21	184 457	5 000	184 457	2 100	图像文本	4 096维 VGG, 1 000维 BoW

4.1.1 MIRFLICKR-25K[33]

该数据集由 25 000 个图文对构成, 其中每个图文对包括一张图片和该图所对应的多个文本标签, 同时每个图文对按照其语义信息被标注为一个 24 维的多标签向量. 该原始数据集中具有一些没有标签的文本对, 将这些没有标签的文本对剔除之后, 剩余 20 015 个文本对用于本文的验证实验. 为公平比较, 在该数据集中, 每张图片由预训练好的 19 层 VGGNet[36]提取出 4 096 维特征向量进行表示, 对应的文本由 2 912 维 BoW 向量表示.

4.1.2 IAPR TC-12[34]

该数据集由 20 000 个图文对组成, 每个图文对由 255 个独立语义类别的多标签进行标注. 与其他的数据集不同, IAPR TC-12 所有的样本均进行了标注, 无需进行剔除, 因此将整个数据集用于实验. 每张图片由预训练好的 CNN-F[36]提取出 4 096 维特征向量进行表示, 对应的文本由 2 912 维 BoW 向量表示.

4.1.3 INUS-WIDE[35]

该数据集具有 269 498 张网页图片, 每张图片具有对应的文本标签, 共同组成一个图文对. 根据其语义信息, 一个图文对可标注为 81 个类别中的一个或多个标签. 此外, 该数据集不同类别分布非常不均匀, 故本文仅选择属于 10 个最常见类别的样本进行实验, 共有 186 557 个图文对. 与 MIRFLICKR-25K 类似, 每张图片由预训练好的 19 层 VGGNet[36]提取出 4 096 维特征向量进行表示, 对应的文本由 1 000 维 BoW 向量表示.

4.2 实验设置

本文采用跨模态检索任务以评估算法的性能. 类似于文献 [7,37], 首先从数据集中随机选择若干图文对作为查询集, 然后将剩下的图文对作为检索数据库以供检索. 为评估跨模态哈希算法的有效性, 本文采用两种不同的跨模态检索任务:

• 以图搜文 (图像→文本): 给定任意一个图像样本作为查询输入, 首先计算其与文本检索库中所有样本的汉明距离, 然后按照距离从小到大对相应的文本样本进行排序, 该排序结果为该图像样本的检索结果.

• 以文搜图 (文本→图像): 与以图搜文类似, 给定任意一个文本样本作为查询输入, 首先计算其与图像检索库中所有样本的汉明距离, 然后按照距离从小到大对相应的图像样本进行排序, 该排序结果为该文本样本的检索结果.

其中, 若两个跨模态样本具有至少一个相同语义类别时, 则为相关样本, 反之则不相关. 为定量地评估检索结果的准确性, 本文采用常用的平均精度均值 (mean average precision, MAP) 作为检索性能的评估指标. MAP 为所有查询结果的平均精度 (average precision, AP) 的均值, 其被广泛地用来衡量检索结果的准确率. 除了 MAP, 本文还采用 Precision-Recall 曲线作为另一个评估标准来直观地对算法性能进行评估. 值得注意的是, 与现有大多数方法不同, 本文的 MAP 值将在所有返回的检索结果上进行计算, 以更全面地评估检索性能.

在对比实验部分, 本文选用 10 种经典的跨模态哈希方法作为对比算法. 其中, FOMH[38]和 DCH[37]为有监督跨模态哈希方法, CVH[18]、LSSH[39]、CMFH[40]、FSH[11]、UCH[26]、UGACH[12]、DJSRH[7]、MGAH[17]和 UKD-SS[13]

为无监督的跨模态哈希方法. 需要说明的是, FOMH[38]、DCH[37]、CVH[18]、LSSH[39]、CMFH[40]、FSH[11]和UCH[26]为浅层的跨模态哈希算法, UGACH[12]、DJSRH[7]、MGAH[17]和 UKD-SS[13]为基于深度学习的跨模态哈希方法. 所有对比方法均采用作者提供的默认参数进行评估. 对于本文的 AUCMH, 除了最后一层全连接层接可微分哈希层之外, 每层全连接层均接一层 ReLU. 本文从检索库中随机选取了 2 000 个图文对作为验证集, 以选择参数 β. 其他参数根据经验固定为: $n=256$ 、 $m=4096$ 、 $\alpha=0.0001$.

4.3 对比实验

为了评估提出算法的有效性, 本文在 3 个常用的跨模态数据集 (即: MIRFLICKR-25K[33]、IAPR TC-12[34]和NUS-WIDE[35]) 上与 10 个跨模态哈希算法进行对比. 为了全面地评估跨模态哈希算法的性能, 在实验中本文分别对比了哈希码长度 L 为 16、32 和 64 时的性能, 其实验结果如表 2–表 4 和图 2 所示. 对表 2–表 4 中的实验结果进行分析, 可以得到如下的结论.

● 随着哈希码长度的增加, 二值特征中所能保留的信息量也相应增加, 从而哈希算法能够取得更好的性能. 此外, 本文的算法能在低比特的情况下取得较好的性能 (例如: 在 MIRFLICKR-25K 和 IAPR TC-12 上, AUCMH 在 16 比特下的 MAP 已经超过大多数对比算法在 64 比特下的结果), 这证明了本方法具有更大的存储和检索优势.

表 2　在 MIRFLICKR-25K 数据集上针对 MAP 分数的性能比较

方法	图像→文本			文本→图像		
	16比特	32比特	64比特	16比特	32比特	64比特
CVH[18]	0.620	0.608	0.594	0.629	0.615	0.599
LSSH[39]	0.597	0.609	0.606	0.602	0.598	0.598
CMFH[40]	0.557	0.557	0.556	0.553	0.553	0.553
FSH[11]	0.581	0.612	0.635	0.576	0.607	0.635
FOMH[38]	0.575	0.640	0.691	0.585	0.648	0.719
DCH[37]	0.596	0.602	0.626	0.612	0.623	0.653
UGACH[12]	0.685	0.693	0.704	0.673	0.676	0.686
DJSRH[7]	0.652	0.697	0.700	0.662	0.691	0.683
MGAH[17]	0.685	0.693	0.704	0.673	0.676	0.686
UKD-SS[13]	0.714	0.718	0.725	0.715	0.716	0.721
AUCMH	**0.736**	**0.742**	**0.744**	**0.719**	**0.723**	**0.730**

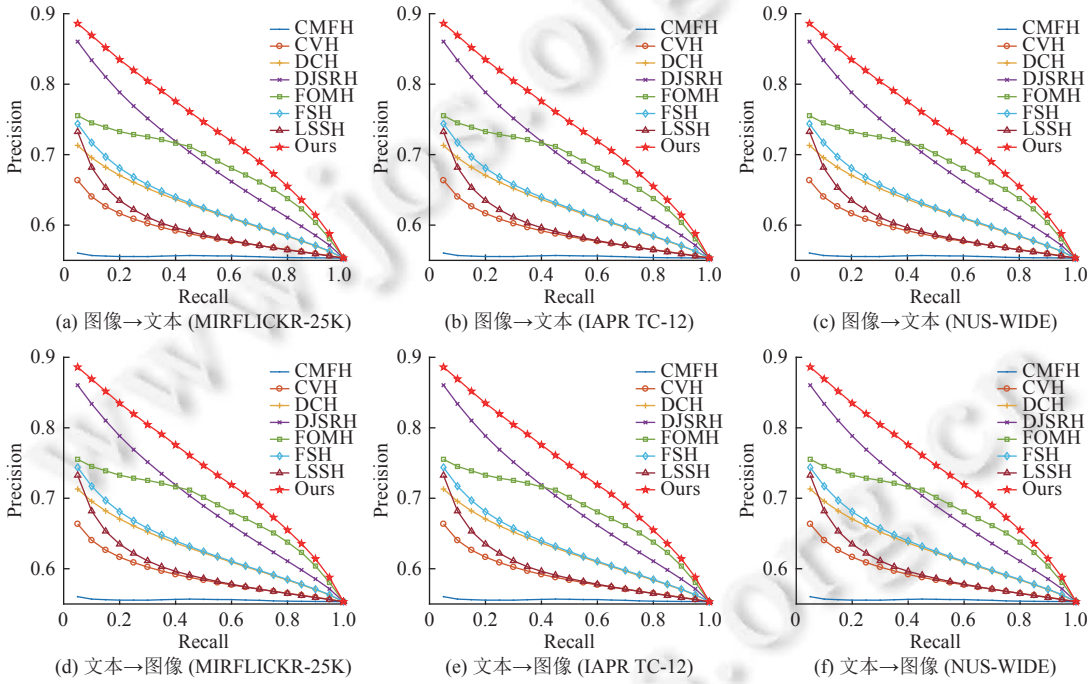
表 3　在 IAPR TC-12 数据集上针对 MAP 分数的性能比较

方法	图像→文本			文本→图像		
	16比特	32比特	64比特	16比特	32比特	64比特
CVH[18]	0.392	0.378	0.366	0.398	0.384	0.372
LSSH[39]	0.372	0.386	0.396	0.367	0.380	0.392
CMFH[40]	0.312	0.314	0.314	0.306	0.306	0.306
FSH[11]	0.377	0.392	0.417	0.383	0.399	0.425
FOMH[38]	0.312	0.316	0.317	0.311	0.315	0.322
DCH[37]	0.336	0.336	0.344	0.350	0.358	0.374
UGACH[12]	0.462	0.467	0.469	0.447	0.463	0.468
DJSRH[7]	0.409	0.412	0.470	0.418	0.436	0.467
AUCMH	**0.478**	**0.486**	**0.496**	**0.476**	**0.487**	**0.496**

● 整体上, 基于深度神经网络的方法比传统的浅层方法具有更好的性能. 这表明深度神经网络能够捕获高非线性的特征, 有利于跨模态哈希算法学习到更好的二值表征, 从而提高算法检索性能.

表 4 在 NUS-WIDE 数据集上针对 MAP 分数的性能比较

方法	图像→文本			文本→图像		
	16比特	32比特	64比特	16比特	32比特	64比特
CVH[18]	0.487	0.495	0.456	0.470	0.475	0.444
LSSH[39]	0.442	0.457	0.450	0.473	0.482	0.471
CMFH[40]	0.339	0.338	0.343	0.306	0.306	0.306
FSH[11]	0.557	0.565	0.598	0.569	0.604	0.651
FOMH[38]	0.305	0.305	0.306	0.302	0.304	0.300
DCH[37]	0.392	0.422	0.430	0.379	0.432	0.444
UGACH[12]	0.613	0.623	0.628	0.603	0.614	0.640
DJSRH[7]	0.502	0.538	0.527	0.465	0.532	0.538
MGAH[17]	0.613	0.623	0.628	0.603	0.614	0.640
UKD-SS[13]	0.614	0.637	0.638	0.630	0.656	0.657
AUCMH	**0.616**	**0.634**	**0.646**	**0.652**	**0.669**	**0.696**



(a) 图像→文本 (MIRFLICKR-25K) (b) 图像→文本 (IAPR TC-12) (c) 图像→文本 (NUS-WIDE)

(d) 文本→图像 (MIRFLICKR-25K) (e) 文本→图像 (IAPR TC-12) (f) 文本→图像 (NUS-WIDE)

图 2 在 MIRFLICKR-25K、IAPR TC-12 和 NUS-WIDE 上的 Precision-Recall 曲线 (哈希编码长度为 64)

● 由于有监督方法可以直接从有标签的数据中学习到判别信息, 即使在标记数据较少时, 其也能取得较为稳定的结果. 但是, 其性能依然比大多数的无监督方法低, 这表明通过利用大量的无标记数据, 无监督方法可弥补判别信息缺失带来的性能差异. 由于在实际应用中, 获取良好标注数据昂贵且费时, 相反获取大量无标记数据更容易, 因此无监督方法更具实用性.

● AUCMH 比现有的基于图的方法具有更好的性能. 这表明基于锚点的图方法不仅可以极大地降低图构建的时间和空间复杂度, 而且能够保证算法的性能.

4.4 消融实验

为全面地评估本文所提出算法的性能, 本文分别在 MIRFLICKR-25K 和 IAPR TC-12 数据集上进行了消融实验, 以验证不同模块的有效性. 本文分别设计了以下不同的 AUCMH 的变体以观察不同模块的重要性.

● AUCMH (无 $\mathcal{L}_r$): 不具备 $\mathcal{L}_r$ 的 AUCMH 变体, 该变体主要为观察 $\mathcal{L}_r$ 对性能的影响.

● AUCMH (无 $\mathcal{L}_g$): 不具备 $\mathcal{L}_g$ 的 AUCMH 变体, 该变体主要为观察 $\mathcal{L}_g$ 对性能的影响.

● AUCMH (无哈希层): 不具备哈希层的 AUCMH 变体, 该变体主要为观察哈希层对性能的影响.

为公平比较, 所有变体与本文的 AUCMH 采用相同的神经网络结构和参数进行训练, 其实验结果如表 5 所示. 通过观察和分析该实验结果, 不难得出如下结论.

表 5　在不同数据集上的消融实验 (最高分数加粗显示)

数据集	方法	图像→文本			文本→图像		
		16	32	64	16	32	64
MIRFLICKR-25K	AUCMH (无 $\mathcal{L}_r$)	0.725	0.734	0.740	0.714	0.721	0.726
	AUCMH (无 $\mathcal{L}_g$)	0.673	0.686	0.699	0.666	0.692	0.700
	AUCMH (无哈希层)	0.724	0.730	0.737	0.711	0.711	0.721
	AUCMH	**0.736**	**0.742**	**0.744**	**0.719**	**0.723**	**0.730**
IAPR TC-12	AUCMH (无 $\mathcal{L}_r$)	0.464	0.473	0.474	0.460	0.469	0.474
	AUCMH (无 $\mathcal{L}_g$)	0.438	0.450	0.456	0.444	0.451	0.454
	AUCMH (无哈希层)	0.463	0.475	0.486	0.462	0.477	0.484
	AUCMH	**0.478**	**0.486**	**0.496**	**0.476**	**0.487**	**0.496**

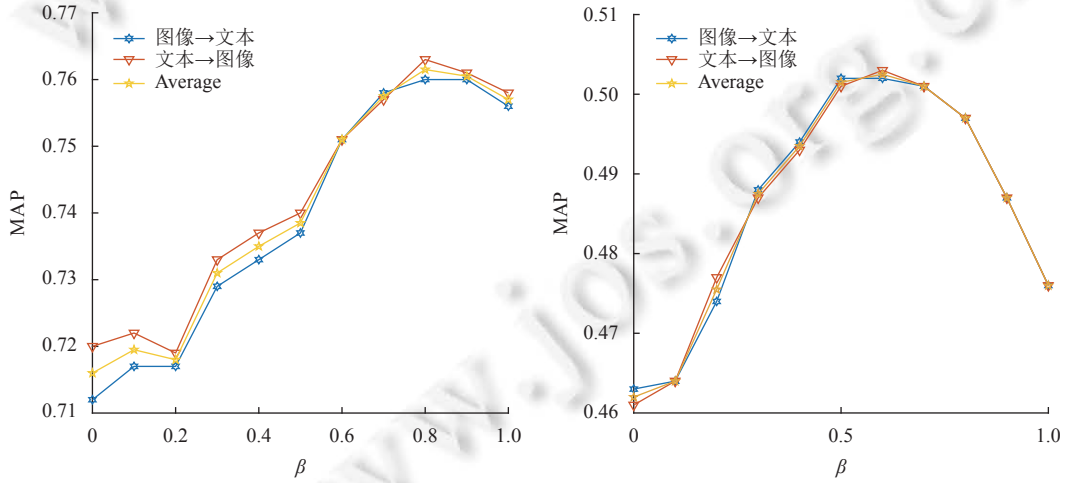
● 所有的模块对算法均有一定的贡献, 缺失任意一个模块均会造成算法性能的下降.

● 基于锚点的图损失函数 $\mathcal{L}_g$ 对性能具有较大影响. 当仅采用 $\mathcal{L}_g$ 时算法依然可以取得较好的性能, 证明了本文提出的基于锚点计算相似性图的有效性.

● 采用连续值松弛 (无哈希层) 会导致跨模态哈希检索性能降低. 这证实了连续值松弛会导致算法性能下降以及提出的无松弛哈希算法的有效性.

4.5　参数分析

为探讨参数 β 对算法性能的影响, 本文在 MIRFLICKR-25K 和 IAPR TC-12 两个数据集的验证集上分别进行了参数分析实验, 其实验结果如图 3 所示. 从图 3 可以看出基于锚点的图损失 $\mathcal{L}_g$ 对性能具有更大的影响, 与第 4.4 节的消融实验的结论一致. 此外, 算法的性能首先随 β 的增大而提高, 当达到峰值之后, β 再增加时其性能将降低. β 的最优取值对于不同的数据集是不同的, 因此在实验中本文首先在验证集上进行参数分析以获取 β 的最优取值.



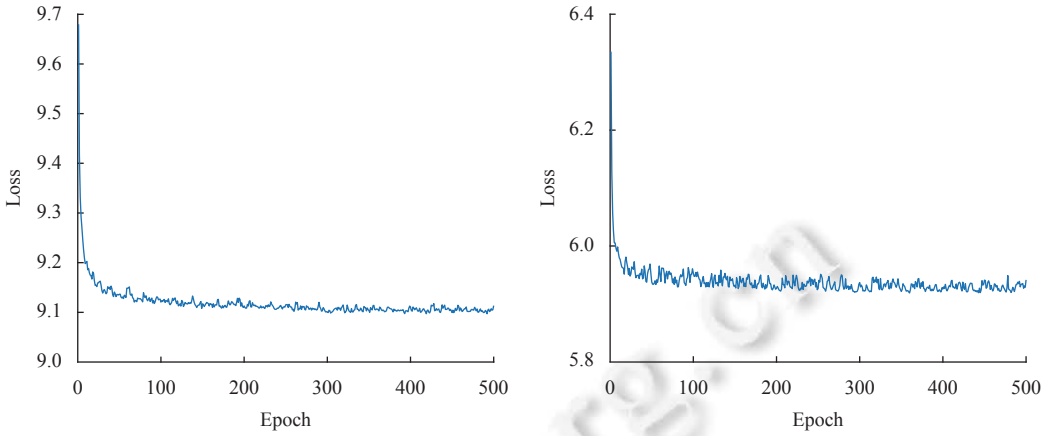
(a) MIRFLICKR-25K　　(b) IAPR TC-12

图 3　在 MIRFLICKR-25K 和 IAPR TC-12 验证集上的参数分析实验结果 (编码长度为 64)

4.6　收敛性分析

为探究 AUCMH 的收敛性, 本文分别在 MIRFLICKR-25K 和 IAPR TC-12 两个数据集上进行了收敛性分析实

验. 图 4 展示了算法的损失值随着迭代周期的变化. 从图中的结果可以看出本文的算法在迭代前期其损失值快速下降, 然后下降趋势逐渐变缓, 最终在第 100 个训练周期附近收敛.

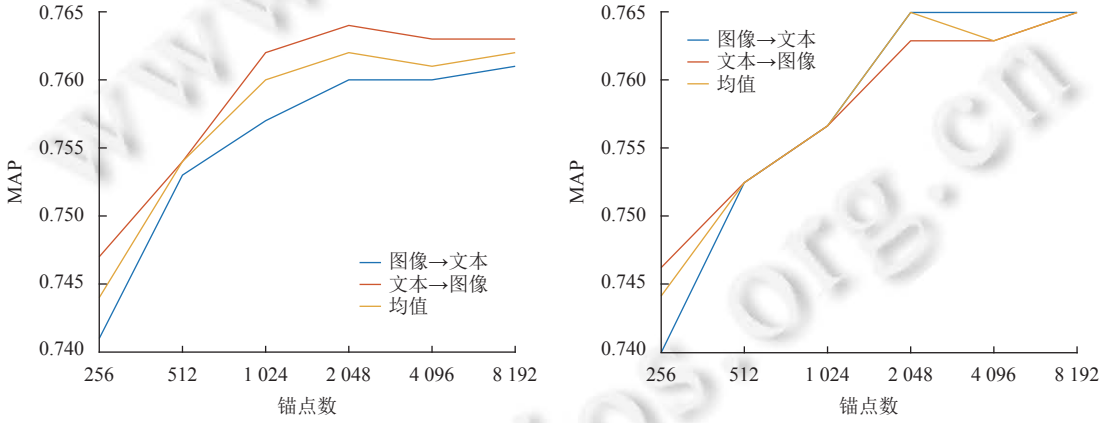


(a) MIRFLICKR-25K (b) IAPR TC-12

图 4 在 MIRFLICKR-25K 和 IAPR TC-12 数据集上的收敛性分析 (编码长度为 64)

4.7 锚点数影响分析

为探究锚点的个数对 AUCMH 性能的影响, 本文在分别在 MIRFLICKR-25K 和 IAPR TC-12 两个数据集上进行了锚点数对性能影响的分析实验, 其实验结果如图 5 所示. 从图中的结果可以看出在不同的数据集上随着锚点数的变化其性能逐渐提高, 然后上升趋势逐渐变缓至稳定. 即, 在锚点数达到一定数量之后, 其继续增加所带来的性能增益将减缓, 甚至不再提高. 因此, 证明了取少量的锚点数以提高构图效率的可行性.



(a) MIRFLICKR-25K (b) IAPR TC-12

图 5 在 MIRFLICKR-25K 和 IAPR TC-12 验证集上对锚点数影响的分析实验结果 (编码长度为 64)

5 总 结

本文提出了一种基于锚点图的跨模态哈希学习算法, 该方法随机采样一些图文对作为锚点集以构建结构图, 从而降低了构建图的时间和空间复杂度. 此外, 本文提出了可微分的哈希层, 在不采用任何连续值松弛的情况下进行离散值优化, 从而获得更好的性能. 在 3 个通用基准数据集上与 10 个跨模态哈希算法进行对比, 证明了本文提出算法的效果和效率. 在未来, 将探索如何在不具有图文对信息与类别信息的同时实现更具挑战性的无监督跨模态哈希学习.

References:

[1] Hu P, Zhen LL, Peng DZ, Liu P. Scalable deep multimodal learning for cross-modal retrieval. In: Proc. of the 42nd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Paris: ACM, 2019. 635–644. [doi: 10.1145/3331184.3331213]

[2] Xu X, Lu HM, Song JK, Yang Y, Shen HT, Li XL. Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval. IEEE Trans. on Cybernetics, 2020, 50(6): 2400–2413. [doi: 10.1109/TCYB.2019.2928180]

[3] Deng C, Xu XX, Wang H, Yang ML, Tao DC. Progressive cross-modal semantic network for zero-shot sketch-based image retrieval. IEEE Trans. on Image Processing, 2020, 29: 8892–8902. [doi: 10.1109/TIP.2020.3020383]

[4] Jin L, Li ZC, Tang JH. Deep semantic multimodal hashing network for scalable image-text and video-text retrievals. IEEE Trans. on Neural Networks and Learning Systems, 2023, 34(4): 1838–1851. [doi: 10.1109/TNNLS.2020.2997020]

[5] Lin ZJ, Ding GG, Han JG, Wang JM. Cross-view retrieval via probability-based semantics-preserving hashing. IEEE Trans. on Cybernetics, 2017, 47(12): 4342–4355. [doi: 10.1109/TCYB.2016.2608906]

[6] Deng C, Chen ZJ, Liu XL, Gao XB, Tao DC. Triplet-based deep hashing network for cross-modal retrieval. IEEE Trans. on Image Processing, 2018, 27(8): 3893–3903. [doi: 10.1109/TIP.2018.2821921]

[7] 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]

[8] Cao ZJ, Long MS, Wang JM, Yu PS. HashNet: Deep learning to hash by continuation. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 5609–5618. [doi: 10.1109/ICCV.2017.598]

[9] Chen ZX, Yuan X, Lu JW, Tian Q, Zhou J. Deep hashing via discrepancy minimization. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6838–6847. [doi: 10.1109/CVPR.2018.00715]

[10] Hu P, Peng X, Zhu HY, Lin J, Zhen LL, Peng DZ. Joint versus independent multiview hashing for cross-view retrieval. IEEE Trans. on Cybernetics, 2021, 51(10): 4982–4993. [doi: 10.1109/TCYB.2020.3027614]

[11] Liu H, Ji RR, Wu YJ, Huang FY, Zhang BC. Cross-modality binary code learning via fusion similarity hashing. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6345–6353. [doi: 10.1109/CVPR.2017.672]

[12] Zhang J, Peng YX, Yuan MK. Unsupervised generative adversarial cross-modal hashing. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018. 539–546. [doi: 10.1609/aaai.v32i1.11263]

[13] Hu HT, Xie LX, Hong RC, Tian Q. Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 3120–3129. [doi: 10.1109/CVPR42600.2020.00319]

[14] Hu P, Wang X, Zhen LL, Peng DZ. Separated variational hashing networks for cross-modal retrieval. In: Proc. of the 27th ACM Int'l Conf. on Multimedia. Ottawa: ACM, 2019. 1721–1729. [doi: 10.1145/3343031.3351078]

[15] Hu P, Zhu HY, Peng X, Lin J. Semi-supervised multi-modal learning with balanced spectral decomposition. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 99–106. [doi: 10.1609/aaai.v34i01.5339]

[16] 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, 2021. 5399–5409. [doi: 10.1109/CVPR46437.2021.00536]

[17] Zhang J, Peng YX. Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval. IEEE Trans. on Multimedia, 2020, 22(1): 174–187. [doi: 10.1109/TMM.2019.2922128]

[18] Kumar S, Udupa R. Learning hash functions for cross-view similarity search. In: Proc. of the 22nd Int'l Joint Conf. on Artificial Intelligence. Barcelona: AAAI, 2011. 1360–1365. [doi: 10.5591/978-1-57735-516-8/IJCAI11-230]

[19] Jiang QY, Li WJ. Deep cross-modal hashing. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 3270–3278. [doi: 10.1109/CVPR.2017.348]

[20] Li K, Qi GJ, Ye J, Hua KA. Linear subspace ranking hashing for cross-modal retrieval. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(9): 1825–1838. [doi: 10.1109/TPAMI.2016.2610969]

[21] Liu XW, Yu GX, Domeniconi C, Wang J, Ren YZ, Guo MZ. Ranking-based deep cross-modal hashing. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI, 2019. 4400–4407. [doi: 10.1609/aaai.v33i01.33014400]

[22] Ding K, Fan B, Huo CL, Xiang SM, Pan CH. Cross-modal hashing via rank-order preserving. IEEE Trans. on Multimedia, 2017, 19(3): 571–585. [doi: 10.1109/TMM.2016.2625747]

[23] Zhang Z, Luo HY, Zhu L, Lu GM, Shen HT. Modality-invariant asymmetric networks for cross-modal hashing. IEEE Trans. on Knowledge and Data Engineering, 2023, 35(5): 5091–5104. [doi: 10.1109/TKDE.2022.3144352]

[24] Sun CC, Latapie H, Liu GW, Yan Y. Deep normalized cross-modal hashing with bi-direction relation reasoning. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 4937–4945. [doi: 10.1109/CVPRW56347.

2022.00541]
[25] Li ZC, Tang JH. Weakly supervised deep metric learning for community-contributed image retrieval. IEEE Trans. on Multimedia, 2015, 17(11): 1989–1999. [doi: 10.1109/TMM.2015.2477035]
[26] Li C, Deng C, Wang L, Xie D, Liu XL. Coupled CycleGAN: Unsupervised hashing network for cross-modal retrieval. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI, 2019. 176–183. [doi: 10.1609/aaai.v33i01.3301176]
[27] Li L, Zheng BH, Sun WW. Adaptive structural similarity preserving for unsupervised cross modal hashing. In: Proc. of the 30th ACM Int'l Conf. on Multimedia. Lisboa: ACM, 2022. 3712–3721. [doi: 10.1145/3503161.3548431]
[28] Bengio Y, Léonard N, Courville A. Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv:1308.3432, 2013.
[29] Su SP, Zhang C, Han K, Tian YH. Greedy hash: Towards fast optimization for accurate hash coding in CNN. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montreal: Curran Associates Inc., 2018. 806–815.
[30] Liu W, He JF, Chang SF. Large graph construction for scalable semi-supervised learning. In: Proc. of the 27th Int'l Conf. on Machine Learning. Haifa: Omnipress, 2010. 679–686.
[31] Liu JJ, Zhang ST, Liu W, Deng C, Zheng YJ, Metaxas DN. Scalable mammogram retrieval using composite anchor graph hashing with iterative quantization. IEEE Trans. on Circuits and Systems for Video Technology, 2017, 27(11): 2450–2460. [doi: 10.1109/tcsvt.2016.2592329]
[32] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego, 2015. 1–13.
[33] 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]
[34] Escalante HJ, Hernández CA, Gonzalez JA, López-López A, Montes M, Morales EF, Enrique Sucar L, Villaseñor L, Grubinger M. The segmented and annotated IAPR TC-12 benchmark. Computer Vision and Image Understanding, 2010, 114(4): 419–428. [doi: 10.1016/j.cviu.2009.03.008]
[35] Rasiwasia N, Costa Pereira J, Coviello E, Doyle G, Lanckriet GRG, Levy R, Vasconcelos N. A new approach to cross-modal multimedia retrieval. In: Proc. of the 18th ACM Int'l Conf. on Multimedia. Firenze: ACM, 2010. 251–260. [doi: 10.1145/1873951.1873987]
[36] Chatfield K, Simonyan K, Vedaldi A, Zisserman A. Return of the devil in the details: Delving deep into convolutional nets. In: Proc. of the 2014 British Machine Vision Conf. Nottingham: BMVA Press, 2014.
[37] Xu X, Shen FM, Yang Y, Shen HT, Li XL. Learning discriminative binary codes for large-scale cross-modal retrieval. IEEE Trans. on Image Processing, 2017, 26(5): 2494–2507. [doi: 10.1109/TIP.2017.2676345]
[38] Lu X, Zhu L, Cheng ZY, Li JJ, Nie XS, Zhang HX. Flexible online multi-modal hashing for large-scale multimedia retrieval. In: Proc. of the 27th ACM Int'l Conf. on Multimedia. Nice: ACM, 2019. 1129–1137. [doi: 10.1145/3343031.3350999]
[39] Zhou JL, Ding GG, Guo YC. Latent semantic sparse hashing for cross-modal similarity search. In: Proc. of the 37th Int'l ACM SIGIR Conf. on Research & Development in Information Retrieval. Gold Coast: ACM, 2014. 415–424. [doi: 10.1145/2600428.2609610]
[40] Ding GG, Guo YC, Zhou JL, Gao Y. Large-scale cross-modality search via collective matrix factorization hashing. IEEE Trans. on Image Processing, 2016, 25(11): 5427–5440. [doi: 10.1109/TIP.2016.2607421]

胡鹏(1990－), 男, 博士, 副研究员, 博士生导师, CCF 专业会员, 主要研究领域为机器学习, 多媒体分析.

彭德中(1975－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为盲信号处理, 神经网络.

彭玺(1983－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为机器学习, 多媒体分析.