第53卷第5期
2021年10月

南京航空航天大学学报
Journal of Nanjing University of Aeronautics & Astronautics

Vol. 53 No. 5
Oct. 2021

DOI:10.16356/j.1005-2615.2021.05.004

基于K-means的深度跨模态哈希量化优化方法

吴家皋[1,2], 杨　璐[1,2], 翁玮薇[1,2], 刘林峰[1,2]

(1.南京邮电大学计算机学院,南京 210023; 2.江苏省大数据安全与智能处理重点实验室,南京 210023)

摘要: 互联网应用的普及使得多模态数据快速增长,跨模态检索技术已成为相关领域的关键技术之一。针对现有跨模态哈希算法存在的网络结构和量化方法等方面的问题,本文在新的深度跨模态哈希检索模型之上,提出了一种基于K-means的深度跨模态哈希量化优化方法(K-means-based quantitative-optimization for deep cross-modal hashing,KQDH)。该方法通过K-means聚类算法对多模态数据特征向量分类,并通过集体量化方式来控制量化误差,使得哈希码更好地表示出多模态特征。实验结果表明,该方法能在多模态数据之间保持相似性并最大程度地捕获语义信息,从而提高跨模态检索的准确性和效率。

关键词: 多模态;哈希算法;聚类算法;特征向量

中图分类号: TP391.4　　**文献标志码:** A　　**文章编号:** 1005-2615(2021)05-0684-08

K-means Based Quantitative-Optimization Method for Deep Cross-Modal Hashing

WU Jiagao[1,2], *YANG Lu*[1,2], *WENG Weiwei*[1,2], *LIU Linfeng*[1,2]

(1. School of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; 2. Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, Nanjing 210023, China)

Abstract: Multi-modal data are growing rapidly with the popularity of Internet applications, and cross-modal retrieval technology has become one of the key technologies in related research areas, where the cross-modal hash algorithm has been paid more and more attention because of its simplicity and efficiency recently. Due to the problems of existing algorithms in the network structure and quantization method, based on a new deep cross-modal hash retrieval model, a K-means based quantitative-optimization method for deep cross-modal hashing (KQDH) is proposed, which classifies the feature vectors of multi-modal data by K-means clustering algorithm, controls the quantization error by the collective quantization method, and makes the hash code better represent the multi-modal features. Experiments show that the proposed method can preserve the similarity between multi-modal data and capture semantic information to the greatest extent, and improve the accuracy and efficiency of cross-modal retrieval.

Key words: multi-modal; hash algorithm; clustering algorithm; feature vector

互联网应用的普及使得文本、图像和视频等多模态数据在网络上快速增加。如何有效地利用各种模态数据进行信息检索已成为相关领域的研究热点,而跨模态检索则是其中的关键技术之一[1-3]。跨模态检索是指在不同模态的数据中进行检索,即通过一种模态的数据检索出语义相似的其他模态数据。由于传统的跨模式检索方法在处理海量、高维多模态数据时存在计算量大、效率低等问题,因此,研究人员关注到信息检索领域中常用的哈希算法,提出了跨模态哈希方法。跨模态哈希方法使用

基金项目: 国家自然科学基金(41571389,61872191)资助项目。

收稿日期: 2020-10-29;**修订日期:** 2020-12-05

通信作者: 吴家皋,男,副教授,E-mail:jgwu@njupt.edu.cn。

引用格式: 吴家皋,杨璐,翁玮薇,等. 基于K-means的深度跨模态哈希量化优化方法[J]. 南京航空航天大学学报,2021,53(5):684-691. WU Jiagao, YANG Lu, WENG Weiwei, et al. K-means based quantitative-optimization method for deep cross-modal hashing[J]. Journal of Nanjing University of Aeronautics & Astronautics,2021, 53(5):684-691.

哈希码将图像和文本映射到同一向量空间，从而有效地降低了算法复杂度、提高了检索效率，目前已广泛应用于大规模跨模态检索中。

跨模态哈希方法通常分为两类：监督哈希方法和无监督哈希方法。有监督哈希方法包括语义相关最大化方法（Semantic correlation maximization，SCM）[4]、语义保留哈希方法（Semantics-preserving hashing，SePH）[5]等。无监督哈希方法有潜在语义稀疏哈希方法（Latent semantic sparse hash，LSSH）[6]、媒体间哈希方法（Inter-media hashing，IMH）[7]、语义主题多模态哈希方法（Semantic topic multimodal hashing，STMH）[8]以及交叉视图哈希方法 CVH（Cross-view hashing，CVH）[9]等。但是，大多数早期的哈希方法都是基于浅层结构和人工特征提取的，无法描述不同模态之间复杂的非线性关系。

近年来，深度跨模态哈希方法利用深度神经网络的优势来捕获不同模态之间的相关性，因此比基于浅层结构的哈希方法更有效。深度哈希经典算法——深度跨模态哈希（Deep cross-modal hash，DCMH）[10] 使用深度神经网络模型实现端到端的特征学习和哈希码学习，通过保留标记信息语义关联构造的不同模态之间的关系以学习哈希码。但是，DCMH仅使用单独的量化来生成次优的哈希二进制代码，并且难以保持特征值和哈希代码之间的最佳兼容性，这可能导致检索结果不准确。但是，当前的深度跨模态哈希算法仍然存在一些不足。首先，由于实际数据集中存在大量语义相似的数据对，因此这些方法仅使用单独的网络来提取每个模态的特征，而无法在不同模态之间建立准确的关联。其次，哈希码生成与公共表示学习是分离的，这大大降低了哈希码的学习准确性。

为了解决上述问题，本文提出了一种基于K-means的深度跨模态哈希量化优化方法（K-means-based quantitative-optimization for deep cross-modal hashing，KQDH）。该方法通过一种全新的量化方式来控制量化误差，减少计算量的同时，使得哈希码更好地表示出多模态特征。K-means聚类算法是一种经典的基于迭代思想的聚类算法。相比与其他聚类算法，K-means算法更加简单、高效，使用也最为广泛。

1　相关工作

跨模态检索的目的是检索拥有相似语义信息的不同模态的数据结果。典型的跨模态检索任务包括：以文检图、以图检文等。因为哈希方法具有提高检索效率、降低存储空间占用的优点，所以将哈希算法用于跨模态检索已经成为这几年的研究热点。

机器学习一般可分为无监督方法[11-13]和有监督方法[14-16]。无监督哈希方法是指学习没有语义标签的哈希函数。LSSH[6]是一种典型的无监督方法，该方法利用字典表示和矩阵分解来学习各种模态数据的隐空间，并利用隐空间中相应的低维表示系数进行量化得到哈希码。但LSSH算法复杂度高且训练时间长。IMH[7]研究视图和视图之间的一致性，是一种基于图的无监督方法，它采用最小化二进制编码距离来保持多模态数据之间的相似性，通过线性回归模型将不同视图的特征映射到常见的汉明空间哈希函数。STMH[8]方法在充分考虑数据隐式语义信息的情况下训练编码过程。该方法通过隐语义空间投影找到语义主题与数据的关联，能直接解析出二进制哈希码。CVH[9]是单模态哈希方法在多模态中的扩展，其目标是尽量减小相似数据间的距离，并增大非相似数据间的距离。该方法可以通过公式得到目标优化问题，并通过放松二值约束，求解获得哈希函数。有监督哈希方法可以利用现有的监视信息（例如语义标签或语义关联）来增强数据相关性并缩小不同模型中的语义差距。语义相关最大化（Semantic correlation maximization，SCM）[4]使用标签信息描述语义关联，获得相似性矩阵，然后通过二进制代码对其进行重构。语义保留哈希（Semantics-preserving hashing，SePH）[5]使用给定的语义相似度矩阵作为监督信息，并将其转换为概率分布。通过最小化Kullback-Leibler距离，并且将逻辑回归用于学习，对哈希函数的每种模态进行非线性预测，此方法可以提高检索准确性，但会降低检索速度。

但是，大多数早期的哈希方法都是基于浅层结构和人工特征提取的，无法描述不同模态之间复杂的非线性关系，不能有效发现其内在相关性。近年来，面向跨模态的深度模型研究表明，它们可以有效利用模态之间的异构关系。使用深层结构实现跨模态检索的代表性工作包括如下：深度跨模态哈希算法（Deep cross-modal hashing，DCMH）[10]通过保留标记信息语义关联构造的不同模态之间的关系以学习哈希码，从而能在深度神经网络模型上同时实现端到端的特征学习和哈希码学习；自我监督的对抗式哈希网络（Self-supervised adversarial hashing networks，SSAH）[17]将对抗生成网络应用于跨模态哈希检索中。该算法通过生成器和判别器的对抗训练来获得具有一致性的语义哈希码；成对关系引导的深度哈希（Pairwise relationship guid-

ed deep hashing, PRDH)[14]方法通过端到端深度学习架构生成紧凑的哈希码,可以有效地捕获各种模态之间的内在联系。该算法的体系结构集成了不同类型的成对约束,以分别从模内视图和模间视图鼓励哈希码的相似性。而且,附加的去相关约束被引入该架构,从而增强了每个散列比特的判别能力。但是,这些方法并未探索量化技术来最大程度地减少量化误差并提高深度表示的可量化性。

2 系统模型

在跨模态检索中,输入的数据和被检索到的数据分别来自不同的模态,本文主要研究图像和文本这两种模态数据。设图像数据集的大小为 N_x,其中的每个图像数据点表示为 $\{\boldsymbol{x}_i\}_{i=1}^{N_x}$,$\boldsymbol{x}_i \in \mathbf{R}^{G_x}$,表示图像模态的 G_x 维特征向量;文本数据集的大小为 N_y,其中的每个文本数据点表示为 $\{\boldsymbol{y}_j\}_{j=1}^{N_y}$,$\boldsymbol{y}_j \in \mathbf{R}^{G_y}$,表示文本模态的 G_y 维特征向量,且 $N_x = N_y = N$。进一步,将文本和图片通过相似矩阵 $\boldsymbol{S}$ 连接,当 $S_{ij}=1$ 表示 $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_j$ 是相似的,反之,当 $S_{ij}=0$ 时表示 $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_j$ 是不相似的。在监督哈希中,可以根据数据点的语义标签构造 $\boldsymbol{S}=\{S_{ij}\}$。KQDH 的目标是共同学习特定于模态的哈希码 $\boldsymbol{b}_i^x \in \{0,1\}^D$ 和 $\boldsymbol{b}_i^y \in \{0,1\}^D$,将每一个数据点 $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_j$ 编码成长度为 D 的二进制编码,并且使得在给定的每个数据点的相似矩阵 $\{\boldsymbol{S}_{ij}\}$ 中传递的相似性信息能被最大程度地保留。

用于提取多模态特征和量化的混合深度网络结构如图 1 所示,该网络由图像网络、文本网络和量化过程组成。图像网络采用扩展的 AlexNet,这是一个深度卷积神经网络(Convolutional neural networks, CNN),它由 5 个卷积层 conv1~conv5 和 3 个全连接层 fc6~fc8 组成。将 fc8 层替换为 D 个节点的全连接层,这会将 fc7 的特征表示转换为 D 维特征表示 $\boldsymbol{h}_i^x$。文本网络采用 Word2Vec 模型将文本信息转换为词向量,同样使用 3 个全连接层 fc1~fc3,其中最后 1 层 fc3 替换为 D 个节点的全连接层,将词向量转换为 D 维特征表示 $\boldsymbol{h}_i^y$。为了使得到的 D 维特征表示 $\boldsymbol{h}_i^x$ 和 $\boldsymbol{h}_i^y$ 能更好地被量化为二进制编码,使用 tanh 激活函数 $a(\boldsymbol{h})=\tanh(\boldsymbol{h})$ 来生成非线性降维特征表示。损失函数用于控制跨模态学习和量化误差,以将其最小化,用于进行高效的跨模态检索。

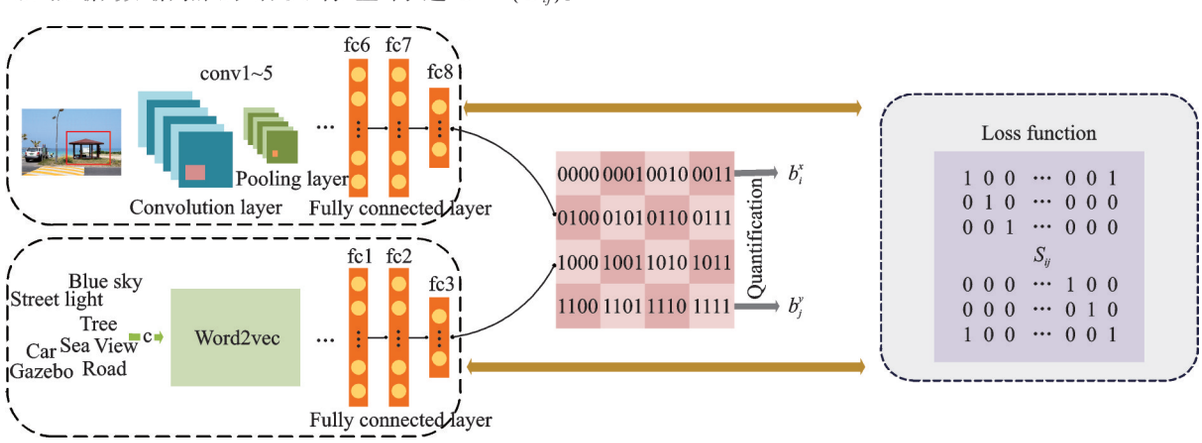


图 1 KQDH 网络结构

Fig.1 Network architecture of KQDH

使用集体量化的方法来保持特征表示 $\boldsymbol{h}_i^x$ 和 $\boldsymbol{h}_i^y$ 与二进制哈希码 $\boldsymbol{b}_i^x$ 和 $\boldsymbol{b}_i^y$ 之间的相似性,其中的关键思想是将提取到的特征向量按其分量顺序均匀地分解为 M 个相同长度的子向量,再通过 K-means 算法对所有的子向量进行分类,并对同一类别中的子向量进行集体量化,映射在同一码本中。在本文中,M 的取值与 D 成正比,从而获得适当的量化码长度和聚类数。

跨模态哈希的目标是学习 2 个哈希函数:$h^{(x)}(\boldsymbol{x}) \in \{0,1\}^D$ 用于图像模态;$h^{(y)}(\boldsymbol{y}) \in \{0,1\}^D$ 用于文本模态。这 2 个哈希函数应保留相似矩阵 $\boldsymbol{S}$ 中的跨模态相似性,即:若 $S_{ij}=1$,则二进制编码 $\boldsymbol{b}_i^x = h^{(x)}(\boldsymbol{x}_i)$ 和 $\boldsymbol{b}_j^y = h^{(y)}(\boldsymbol{y}_j)$ 之间应保持较小距离;否则,相应的距离应非常大。

3 算法描述

3.1 K-means 量化优化方法

K-means 聚类算法是一种迭代聚类算法,其主要过程是:随机选择 K 个对象作为初始聚类中心,并计算每个对象与每个聚类中心之间的距离。再把每个对象都归类到最近的聚类中心。聚类中心及其所属对象就组成一个聚类代。根据现有聚类对象重新计算新的聚类中心,重复聚类过程,直到满足某终止条件。

本文通过图1神经网络训练获得的特征 $\boldsymbol{h}_i^x$ 和 $\boldsymbol{h}_i^y$ 为 D 维的特征向量，将其按分量顺序均匀地划分为 M 个子向量：$\boldsymbol{h}_i^x=\{\boldsymbol{h}_{i1}^x;\boldsymbol{h}_{i2}^x;\cdots;\boldsymbol{h}_{iM}^x\}$，$\boldsymbol{h}_{im}^x\in\mathbf{R}^{D_x/M}$，$\boldsymbol{h}_j^y=\{\boldsymbol{h}_{j1}^y;\boldsymbol{h}_{j2}^y;\cdots;\boldsymbol{h}_{jM}^y\}$，$\boldsymbol{h}_{im}^y\in\mathbf{R}^{D_y/M}$，$D_x=D_y=D$，映射所有的子向量 $\hat{\boldsymbol{H}}=\{\boldsymbol{h}_{im}^x,\boldsymbol{h}_{jm}^y\}$，$i,j\in[1,N]$，$m\in[1,M]$ 到相同的向量空间。取聚类数 $K=2^{\frac{D}{M}}$，通过K-means算法分类后将所有的子向量量化为 $\boldsymbol{b}_i^x,\boldsymbol{b}_i^y$。$\hat{\boldsymbol{H}}$ 中的每一个向量用 $\boldsymbol{h}_p$ 来表示。

算法1是本文提出的KQDH算法的伪代码描述。

算法1　KQDH算法

输入： $\hat{\boldsymbol{H}},K$

输出： $\boldsymbol{b}_i^x,\boldsymbol{b}_j^y$

从 $\hat{\boldsymbol{H}}$ 中随机选择 K 个样本作为原始的聚类中心 $\boldsymbol{\omega}_\tau,\tau\in\{1,\cdots,K\}$

使 $\boldsymbol{C}_\tau\leftarrow\varnothing,\tau\in\{1,\cdots,K\}$

repeat

for each $\boldsymbol{h}_p\in\hat{\boldsymbol{H}}$ do

//计算 $\boldsymbol{h}_p$ 与 $\boldsymbol{\omega}_\tau$ 之间的欧氏距离并进行聚类//

$\tau^*\leftarrow\arg\min_{\tau\in\{1,\cdots,K\}}\left|\boldsymbol{h}_p-\boldsymbol{\omega}_\tau\right|$

$\boldsymbol{C}_{\tau^*}\leftarrow\boldsymbol{C}_{\tau^*}\cup\{\boldsymbol{h}_p\}$

end for

for $\tau=1$ to K do

//计算新的聚类中心//

$$\boldsymbol{\omega}_\tau'\leftarrow\frac{\sum_{\boldsymbol{h}_p\in\boldsymbol{C}_\tau}\boldsymbol{h}_p}{\left|\boldsymbol{C}_\tau\right|}$$

$\boldsymbol{\omega}_\tau\leftarrow\boldsymbol{\omega}_\tau'$

end for

until $E<$ threshold

$\forall\boldsymbol{h}_{im}^x\in\boldsymbol{C}_\tau$，使 $\boldsymbol{b}_{im}^x\leftarrow\mathrm{cod}\,\boldsymbol{e}_\tau$

$\forall\boldsymbol{h}_{jm}^y\in\boldsymbol{C}_\tau$，使 $\boldsymbol{b}_{jm}^y\leftarrow\mathrm{cod}\,\boldsymbol{e}_\tau$

$\boldsymbol{b}_i^x\leftarrow\{\boldsymbol{b}_{i1}^x;\boldsymbol{b}_{i2}^x;\cdots;\boldsymbol{b}_{iM}^x\}$

$\boldsymbol{b}_j^y\leftarrow\{\boldsymbol{b}_{j1}^y;\boldsymbol{b}_{j2}^y;\cdots;\boldsymbol{b}_{jM}^y\}$

return $\boldsymbol{b}_i^x,\boldsymbol{b}_j^y$

聚类算法的误差函数采用欧氏距离，有

$$E=\sum_{\tau=1}^{K}\sum_{\boldsymbol{h}_p\in\boldsymbol{C}_\tau}\left|\boldsymbol{h}_p-\boldsymbol{\omega}_\tau\right|^2\tag{1}$$

式中：$\boldsymbol{C}_\tau$ 为第 τ 个簇，是输入向量集合的不相交的子集；$\boldsymbol{\omega}_\tau$ 为聚类的中心。

在本文模型中，映射到向量空间的点通过K-means聚类算法划分为 K 个聚类，每个聚类都有一个对应的码本，每个子向量都对应于该聚类的码本。每个聚类的 code_τ 是长度为 $L=\dfrac{D}{M}$ 且值为 $\tau-1$ 的二进制代码，即有

$$\boldsymbol{b}_{im}^x=\mathrm{code}_\tau,\boldsymbol{b}_{jm}^y=\mathrm{code}_\tau\tag{2}$$

$$\boldsymbol{b}_i^x=\{\boldsymbol{b}_{i1}^x;\boldsymbol{b}_{i2}^x;\cdots;\boldsymbol{b}_{iM}^x\},\boldsymbol{b}_j^y=\{\boldsymbol{b}_{j1}^y;\boldsymbol{b}_{j2}^y;\cdots;\boldsymbol{b}_{jM}^y\}\tag{3}$$

将该量化过程表示为

$$\boldsymbol{b}_i^x=\eta(\boldsymbol{h}_i^x),\boldsymbol{b}_j^y=\eta(\boldsymbol{h}_j^y)\tag{4}$$

算法1的复杂度和K-mean聚类算法的复杂度一致，其时间复杂度与样本数据量、样本维度、聚类数和聚类迭代次数成正比。这里，样本数据量为 $\hat{\boldsymbol{H}}$ 中所有子向量的数量 $2NM$，样本维度为子向量的长度 D/M，聚类数为 $K=2^{D/M}$，设聚类迭代次数为 T，则算法1的时间复杂度可表示为 $O(2NM\cdot D/M\cdot 2^{D/M}\cdot T)$，若将 D、M 和 T 都视为常量，则为 $O(N)$。同理，算法1的空间复杂度也为 $O(N)$。因此，算法1具有较低的（线性）复杂度。

3.2　哈希码学习与优化

3.2.1　哈希码学习

采用 $\boldsymbol{h}_i^x=f(\boldsymbol{x}_i;\boldsymbol{\varphi}_x)\in\mathbf{R}^D$ 表示学习到的图像数据点 $\boldsymbol{x}_i$ 的特征，为相应的CNN网络的输出；$\boldsymbol{h}_j^y=t(\boldsymbol{y}_j;\boldsymbol{\varphi}_y)\in\mathbf{R}^D$ 表示学习到的文本数据点 $\boldsymbol{y}_i$ 的特征，为相应的Word2Vec的输出。$\boldsymbol{\varphi}_x,\boldsymbol{\varphi}_y$ 分别表示CNN和Word2Vec网络的参数。$\boldsymbol{B}^x=\{\boldsymbol{b}_1^x;\boldsymbol{b}_2^x;\cdots;\boldsymbol{b}_N^x\}$ 是数据集中所有图像哈希码组成的矩阵，$\boldsymbol{B}^y=\{\boldsymbol{b}_1^y;\boldsymbol{b}_2^y;\cdots;\boldsymbol{b}_N^y\}$ 是数据集中所有文本哈希码组成的矩阵，$\boldsymbol{H}^x=\{\boldsymbol{h}_1^x;\boldsymbol{h}_2^x;\cdots;\boldsymbol{h}_N^x\}$ 是数据集中所有图像特征向量组成的矩阵，$\boldsymbol{H}^y=\{\boldsymbol{h}_1^y;\boldsymbol{h}_2^y;\cdots;\boldsymbol{h}_N^y\}$ 是数据集中所有文本特征向量组成的矩阵。在实验中发现若将两种模态对应数据点的哈希码设置为相同，则可以获得更好的性能。因此，本文设 $\boldsymbol{B}=\boldsymbol{B}^x=\boldsymbol{B}^y$。则KQDH的总体目标函数定义为

$$\begin{aligned}\min_{\boldsymbol{B},\boldsymbol{\varphi}_x,\boldsymbol{\varphi}_y}J&=J_1+\gamma J_2+\mu J_3=\\&-\sum_{i,j=1}^{N}(S_{ij}\theta_{ij}-\log^{p(S_{ij}|\boldsymbol{h}_i^x,\boldsymbol{h}_j^y)})+\\&\gamma(\left\|\boldsymbol{B}-\boldsymbol{H}^x\right\|_{\mathrm{F}}^2+\left\|\boldsymbol{B}-\boldsymbol{H}^y\right\|_{\mathrm{F}}^2)+\mu(\left\|\boldsymbol{H}^x\boldsymbol{1}\right\|_{\mathrm{F}}^2+\\&\left\|\boldsymbol{H}^y\boldsymbol{1}\right\|_{\mathrm{F}}^2)\quad\text{s.t.}\quad\boldsymbol{B}\in\{0,1\}^{D\times N}\end{aligned}\tag{5}$$

式中 γ 和 μ 为超参数。

第1项 $J_1=-\sum_{i,j=1}^{N}(S_{ij}\theta_{ij}-\lg^{p(S_{ij}|\boldsymbol{h}_i^x,\boldsymbol{h}_j^y)})$ 是用于保持多模态相似性的负对数似然函数，该函数定义为

$$p(S_{ij}\left|\boldsymbol{h}_i^x,\boldsymbol{h}_j^x\right.)=\begin{cases}\sigma(\theta_{ij}) & S_{ij}=1\\1-\sigma(\theta_{ij}) & S_{ij}=0\end{cases}\tag{6}$$

式中：$\theta_{ij}=\dfrac{1}{2}\boldsymbol{h}_i^{x\mathrm{T}}\boldsymbol{h}_j^y$；$\sigma(\theta_{ij})=\dfrac{1}{1+\mathrm{e}^{-\theta_{ij}}}$。

容易发现，J_1 项可以使 $\boldsymbol{h}_i^x$ 和 $\boldsymbol{h}_j^y$ 之间的相似性在 $S_{ij}=1$ 时较大，而在 $S_{ij}=0$ 时较小。也就是说能保留图像数据点 $\boldsymbol{x}_i$ 和文本数据点 $\boldsymbol{y}_j$ 之间的相似性。

第2项 $J_2=\left\|B-\boldsymbol{H}^x\right\|_{\mathrm{F}}^2+\left\|B-\boldsymbol{H}^y\right\|_{\mathrm{F}}^2$，其中，$\|\cdot\|_{\mathrm{F}}$为矩阵的Frobenius范数。由于$\boldsymbol{H}^x$和$\boldsymbol{H}^y$可以保留多模态数据点的相似性，因此也可以期望哈希码B保留这种相似性。这一项的主要作用是减少量化得到的哈希码和提取到的特征之间的量化损失。

第3项 $J_3=\left\|\boldsymbol{H}^x\boldsymbol{1}\right\|_{\mathrm{F}}^2+\left\|\boldsymbol{H}^y\boldsymbol{1}\right\|_{\mathrm{F}}^2$主要是保证所有训练样本上哈希码的每一位0和1格式是平衡的。

3.2.2 优化方法

使用和DCMH相同的交替学习策略[10]来学习$\boldsymbol{\varphi}_x$，$\boldsymbol{\varphi}_y$和B。每当一个参数被优化时，其他参数则被固定，主要步骤如下。

步骤1 在固定$\boldsymbol{\varphi}_y$和B的情况下优化$\boldsymbol{\varphi}_x$。

同时使用随机梯度下降(Stochastic gradient descent，SGD)和反向转播(Back propagation，BP)算法，以优化图像模态的CNN参数$\boldsymbol{\varphi}_x$。对于每个采样点$\boldsymbol{x}_i$，计算梯度有

$$\frac{\partial J}{\partial \boldsymbol{h}_i^x}=\frac{1}{2}\sum_{j=1}^{N}(\sigma(\theta_{ij})\boldsymbol{h}_j^y-S_{ij}\boldsymbol{h}_j^y)+2\gamma(\boldsymbol{h}_i^x-\boldsymbol{b}_i^x)+2\mu\boldsymbol{H}^x\boldsymbol{1} \tag{7}$$

然后可以对$\frac{\partial J}{\partial \boldsymbol{h}_i^x}$用链式规则来计算$\frac{\partial J}{\partial \boldsymbol{\varphi}_x}$，基于BP算法可以更新参数$\boldsymbol{\varphi}_x$。

步骤2 在固定$\boldsymbol{\varphi}_x$和B的情况下优化$\boldsymbol{\varphi}_y$。

这里仍然采用与步骤1同样算法来优化文本模态的Word2Vec参数$\boldsymbol{\varphi}_y$。对于每个采样点$\boldsymbol{y}_j$，计算梯度有

$$\frac{\partial J}{\partial \boldsymbol{h}_j^y}=\frac{1}{2}\sum_{i=1}^{N}(\sigma(\theta_{ij})\boldsymbol{h}_i^x-S_{ij}\boldsymbol{h}_i^x)+2\gamma(\boldsymbol{h}_j^y-\boldsymbol{b}_j^y)+2\mu\boldsymbol{H}^y\boldsymbol{1} \tag{8}$$

然后可以对$\frac{\partial J}{\partial \boldsymbol{h}_j^y}$使用链式规则来计算$\frac{\partial J}{\partial \boldsymbol{\varphi}_y}$，基于BP算法可以更新参数$\boldsymbol{\varphi}_y$。

步骤3 在固定$\boldsymbol{\varphi}_x$和$\boldsymbol{\varphi}_y$的情况下优化B。

当$\boldsymbol{\varphi}_x$和$\boldsymbol{\varphi}_y$固定时，可以将式(5)中的问题重新表述为

$$\max_B \operatorname{tr}(B^{\mathrm{T}}(\gamma(\boldsymbol{H}^x+\boldsymbol{H}^y)))=\operatorname{tr}(B^{\mathrm{T}}\boldsymbol{V}) \quad \text{s.t. } B\in\{0,1\}^{D\times N} \tag{9}$$

式中：$\boldsymbol{V}=\gamma(\boldsymbol{H}^x+\boldsymbol{H}^y)$；$\operatorname{tr}(\cdot)$表示矩阵的迹线。

因此，在模型训练过程中，通过对步骤1～3的反复迭代，系统将不断更新图像、文本网络参数并优化哈希码，直到模型收敛或完成训练轮数。

3.2.3 样本外扩展

对于不在训练集中的任何数据点q，若其图像模态为$\boldsymbol{x}_q$，文本模态为$\boldsymbol{y}_q$，则可以利用式(10)和式(11)，通过正向传播生成哈希码$\boldsymbol{b}_q^{(x)}$和$\boldsymbol{b}_q^{(y)}$，即有

$$\boldsymbol{b}_q^{(x)}=h^{(x)}(\boldsymbol{x}_q)=\eta(f(\boldsymbol{x}_q;\boldsymbol{\varphi}_x)) \tag{10}$$

$$\boldsymbol{b}_q^{(y)}=h^{(y)}(\boldsymbol{y}_q)=\eta(t(\boldsymbol{y}_q;\boldsymbol{\varphi}_y)) \tag{11}$$

从而使得本模型可用于跨模态检索。

4 测试与性能分析

为了验证KQDH的有效性，在2个常用的数据集上进行了充分的实验，设计了2种类型的跨模态检索任务来评估其性能：(1)图像-文本：使用图像查询相关文本；(2)文本-图像：使用文本查询相关图像。

4.1 实验数据集

MIRFLICKR-25K[18]数据集包含从社交摄影网站Flickr收集的25 000个实例。每个图像都标记有38个语义概念和一些关联的文本标签，并使用24个类别标签中的一个或多个手动进行注释。在实验中共选择了20 015个数据点，其中不少于20个文本标签。在这个数据集中图片的特征向量维度是150维，而文本特征向量维度则是50维。

NUS-WIDE[19]是一个超大的数据集，由实际网页图片组成，包括269 648个实例以及带有相关文本标记的图像，其中有81种基本事实概念可供人工注释以进行检索评估。在将该数据集作为算法输入之前，首先对该数据集做预处理，从中选取了10种较为常见的标签作为图片的标注，将其余的数据从该数据集里去除，最后得到用于实验的186 577个图像/文本对。

4.2 实验结果及分析

在实验中，将所提出的算法与5种最具代表性的跨模态哈希方法进行比较，包括SCM[4]、STMH[8]、LSSH[6]、CVH[9]和DCMH[10]。这些方法所涵盖的技术层面较广，其中CM、STMH、LSSH和CVH是基于浅层结构，而DCMH和本文的方法则基于深层结构；同时，无监督算法有STMH、LSSH和CVH，而SCM、DCMH和本文的方法都是有监督算法。在实验中，这些方法中的所有参数都是根据原始文献进行设置。

对于MIRFLICKR-25K数据集，随机抽取10 000个实例作为训练集。为了进行测试，使用该数据集的2 000个实例作为测试集，其余则为检索集。对于NUS-WIDE数据集，随机采样了10 500个实例作为训练集。同样地，该数据集的测试集大小为2 100个实例。AlexNet网络已在ImageNet数据集上进行预训练，并在训练本文的模型时进行微调。在实验中，令超参数$\gamma=0.3$和$\mu=0.1$，这样

能获得较好的性能，具体在后面给出实验评价。此外，训练随机采样批次 mini-batch 设为 128，每次实验的训练次数设为 500。所有实验都重复 5 次，然后取平均作为实验结果。

实验中所有比较方法的性能都使用平均精度（Mean average precision，MAP）和准确度（Precision）-召回率（Recall）曲线直接进行了评估[20]。MIRFLICKR-25K 和 NUS-WIDE 中哈希码长度 D 为 16、32 和 64 位，M 分别取值 4、8、16，L=4，K=16，所有方法在 2 个数据集的 MAP 值如表 1，2 所示。此外，图 2 和图 3 显示了两个数据集上所有方法的确切召回曲线。图中哈希码码长度均为 64。通过比较分析，本文方法比其他方法更有效。

表 1　各方法 MAP 在 MIRFLICKR-25K 数据集上的比较

Table 1　Comparison of MAP with different methods on MIRFLICKR-25K dataset

任务	方法	MAP		
		D=16 位	D=32 位	D=64 位
图像-文本	KQDH	**0.732 1**	**0.738 4**	**0.743 1**
	DCMH	0.727 4	0.729 8	0.738 1
	SCM	0.624 3	0.630 5	0.635 2
	STMH	0.594 3	0.598 1	0.599 3
	LSSH	0.585 0	0.588 7	0.591 0
	CVH	0.605 2	0.606 1	0.603 5
文本-图像	KQDH	**0.761 5**	**0.769 8**	**0.780 2**
	DCMH	0.760 1	0.764 2	0.770 2
	SCM	0.615 0	0 .6 216	0.628 3
	STMH	0.586 5	0.590 2	0.592 8
	LSSH	0.594 2	0.596 1	0.599 2
	CVH	0.604 9	0.605 1	0.598 3

表 2　各方法 MAP 在 NUS-WIDE 数据集上的比较

Table 2　Comparison of MAP with different methods on NUS-WIDE dataset

任务	方法	MAP		
		D=16 位	D=32 位	D=64 位
图像-文本	KQDH	**0.632 6**	**0.638 9**	**0.647 0**
	DCMH	0.623 1	0.627 7	0.640 1
	SCM	0.478 6	0.479 3	0.485 4
	STMH	0.429 3	0.435 0	0.440 1
	LSSH	0.383 1	0.387 9	0.391 6
	CVH	0.379 7	0.387 1	0.389 3
文本-图像	KQDH	**0.662 5**	**0.667 6**	**0.681 6**
	DCMH	0.657 1	0.659 7	0.674 0
	SCM	0.445 3	0.450 2	0.459 9
	STMH	0.376 9	0.389 3	0.400 5
	LSSH	0.409 9	0.412 4	0.411 8
	CVH	0.368 2	0.386 9	0.377 6

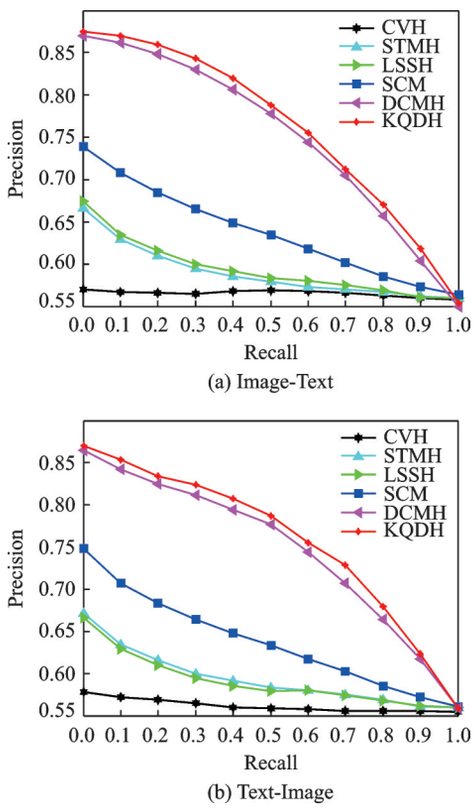


图 2　MIRFLICKR-25K 数据集的精确度-召回率曲线

Fig.2　Precision-recall curves on MIRFLICKR-25K datasets

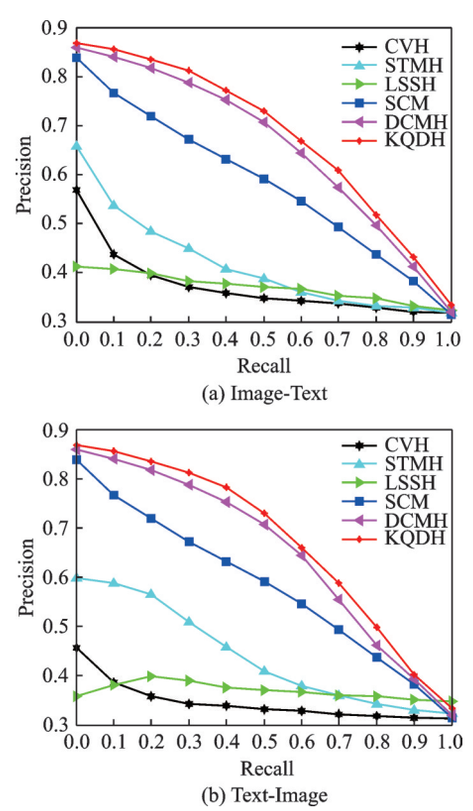
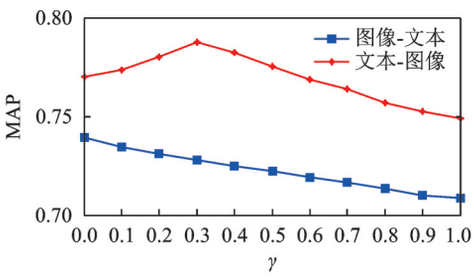


图 3　NUS-WIDE 数据集的精确度-召回率曲线

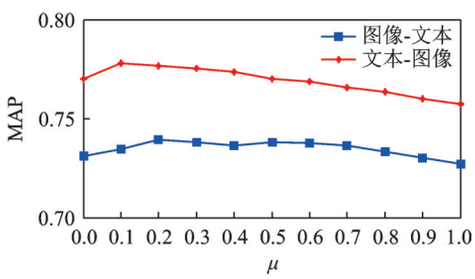
Fig.3　Precision-recall curves on NUS-WIDE datasets

进一步研究超参数 γ 和 μ 对模型性能的影响。图 4 为在 MIRFLICKR-25K 数据集上设置不同 γ 和 μ 值时的 MAP 曲线。从图 4(a)可以看到，当

$0<\gamma<1$时，文本检索图像的MAP值在$\gamma=0.3$时取到最大值，图像检索文本的MAP值随着γ值的增大而减少，故本文取超参数$\gamma=0.3$；同样，由图4(b)可知，当$0<\mu<1$时，文本检索图像的MAP值在$\mu=0.1$时取到最大值，图像检索文本的MAP值随着μ值的变化而小幅度振荡，故本文取超参数$\mu=0.1$。



(a) Influence of γ on MAP



(b) Influence of μ on MAP

图4 超参数的影响

Fig.4 Influence of the hyper-parameters

5 结 论

本文提出了一种用于跨模态检索的深度哈希量化优化方法KQDH，将图像和文本的特征向量映射到相同的向量空间中，并设计了基于K-means的量化结构来控制哈希码的质量，以更准确地描述特征与哈希码之间的相关性。实验结果表明本文方法在跨模态检索中具有较好的性能。在将来的工作中，将继续改进所提出的模型及算法，并将其推广到具有音频、视频等更多模态数据的跨模态检索中。

参考文献：

[1] DENG C, CHEN Z, LIU X, et al. Triplet-based deep hashing network for cross-modal retrieval[J]. IEEE Transactions on Image Processing, 2018, 27(8): 3893-3903.

[2] WANG K Y, HE R, WANG L, et al. Joint feature selection and subspace learning for cross-modal retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(10): 2010-2023.

[3] WU Y L, WANG S H, HUANG Q M. Online asymmetric similarity learning for cross-modal retrieval[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE Computer Society Press, 2017: 3984-3993.

[4] ZHANG D Q, LI W J. Large-scale supervised multimodal hashing with semantic correlation maximization[C]//Proceedings of 28th AAAI Conference on Artificial Intelligence. Quebec, Canada: AAAI, 2014: 2177-2183.

[5] LIN Z J, DING G G, HU M Q, et al. Semantics-preserving hashing for cross-view retrieval[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE Computer Society Press, 2015: 3864-3872.

[6] ZHOU J L, DING G G, GUO Y C. Latent semantic sparse hashing for cross-modal similarity search[C]//Proceedings of 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. Queensland, Australia: ACM, 2014: 415-424.

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

[8] WANG D, GAO X B, WANG X M, et al. Semantic topic multimodal hashing for cross-media retrieval[C]//Proceedings of 24th International Joint Conference on Artificial Intelligence. Argentina: IJCAI, 2015: 3890-3896.

[9] KUMAR S, UDUPA R. Learning hash functions for cross-view similarity search[C]//Proceedings of 22nd International Joint Conference on Artificial Intelligence. Catalonia, Spain: IJCAI, 2011: 1360-1365.

[10] JIANG Q Y, LI W J. Deep cross-modal hashing[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE Computer Society Press, 2017: 3232-3240.

[11] BARLOW H B. Unsupervised learning[J]. Neural Computation, 1989, 1(3): 295-311.

[12] DANON D, AVERBUCH-ELOR H, FRIED O, et al. Unsupervised natural image patch learning[J]. Computational Visual Media, 2019, 5(3): 229-237.

[13] ZHANG J, PENG Y. Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval[J]. IEEE Transactions on Multimedia, 2020, 22(1): 174-187.

[14] YANG E, DENG C, LIU W, et al. Pairwise relation-

ship guided deep hashing for cross-modal retrieval[C]//Proceedings of 31st AAAI Conference on Artificial Intelligence. California, USA: AAAI, 2017: 1618-1625.

[15] ZHEN L, HU P, WANG X. Deep supervised cross-modal retrieval[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE Computer Society Press, 2019: 10386-10395.

[16] SHEN F M, SHEN C H, LIU W, et al. Supervised discrete hashing[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE Computer Society Press, 2015: 37-45.

[17] LI C, DENG C, LI N, et al. Self-supervised adversarial hashing networks for cross-modal retrieval[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Utah, USA: IEEE Computer Society Press, 2018: 4242-4251.

[18] HUISKES M J, LEW M S. The MIR flickr retrieval evaluation[C]//Proceedings of 1st ACM International Conference on Multimedia Information Retrieval. British Columbia, Canada: ACM, 2008: 39-43.

[19] CHUA T S, TANG J H, HONG R, et al. Nus-wide: A real-world web image database from national university of singapore[C]//Proceedings of ACM International Conference on Image and Video Retrieval. Fira, Greece: ACM, 2009: 48.

[20] LIU W, MU C, KUMAR S, CHANG S F. Discrete graph hashing[C]//Proceedings of 27th International Conference on Neural Information Processing Systems. Montreal, Canada: ACM, 2014: 3419-3427.

（编辑：刘彦东）