计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(8):171−178 [doi: 10.15888/j.cnki.csa.008041]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于柯西分布的深度哈希跨媒体检索①

田　枫[1], 李　闯[1], 刘　芳[1], 李婷玉[2], 张　蕾[2], 刘志刚[1]

[1](东北石油大学 计算机与信息技术学院, 大庆 163318)
[2](中国石油天然气股份有限公司 冀东油田分公司, 唐山 063004)
通讯作者: 刘　芳, E-mail: lfliufang1983@126.com

摘　要: 针对深度哈希跨媒体检索方法中, 语义相似的媒体对象的哈希码在汉明空间内的分布不合理问题, 提出了一种新的深度哈希跨媒体检索模型. 该模型是在汉明空间内利用柯西分布对现有的深度哈希跨媒体关联损失进行改进, 使得语义相似的媒体对象哈希码距离较小, 语义不相似的媒体对象哈希码较大, 进而提高模型的检索效果. 同时, 本文给出了一种高效的模型求解方法, 采用交替迭代方式获得模型的近似最优解. 在 Flickr-25k 数据集, IAPR TC-12 数据集和 MS COCO 数据集上的实验结果表明, 该方法可以有效的提高跨媒体检索性能.

关键词: 跨媒体检索; 哈希学习; 柯西函数; 汉明距离; 汉明空间

引用格式: 田枫,李闯,刘芳,李婷玉,张蕾,刘志刚.基于柯西分布的深度哈希跨媒体检索.计算机系统应用,2021,30(8):171–178. http://www.c-s-a.org.cn/1003-3254/8041.html

Cross-Media Retrieval of Deep Hash Based on Cauchy Distribution

TIAN Feng[1], LI Chuang[1], LIU Fang[1], LI Ting-Yu[2], ZHANG Lei[2], LIU Zhi-Gang[1]

[1](School of Computer &Information Technology, Northeast Petroleum University, Daqing 163318, China)
[2](Jidong Oilfield Branch, Petro China Co. Ltd., Tangshan 063004, China)

Abstract: This study proposes a new cross-media retrieval model of deep hash to solve the unreasonable distribution of the hash codes of semantically similar media objects in Hamming space in the existing retrieval methods. In this model, the cross-media association loss of deep hash is improved by the Cauchy distribution in Hamming space, making the hash codes of semantically similar media objects in a short distance and those of semantically dissimilar ones far apart. Thus, the retrieval effect of the model is improved. Furthermore, an efficient model-solving method is presented in this study, and the approximate optimal solution of the model is obtained by alternating iteration. The experimental results on Flickr-25k, IAPR TC-12, and MS COCO datasets show that this method can effectively improve the performance of cross-media retrieval.

Key words: cross-media retrieval; hash learning; Cauchy distribution; Hamming distance; Hamming space

随着互联网技术的快速发展, 图像, 文本, 视频, 音频, 三维模型等多媒体数据量越来越多, 多媒体信息检索[1] 发展迅速, 其中跨媒体检索是研究热点. 跨媒体检索[2] 是指任意使用一种媒体数据对其他媒体数据在语义层面进行相关性检索, 实现多媒体数据在语义上的互通. 其难点在于, 不同媒体类型的数据表示形式不一

① 基金项目: 国家自然科学基金 (61502094, 61702093); 中央支持地方高校改革发展资金人才培养支持计划 (140119001); 黑龙江省省属本科高校基本科研业务费项目 (KYCXTD201903);黑龙江省高等教育教学改革研究项目 (SJGY20180079, SJGY20190098);东北石油大学引导性创新基金 (2020YDL-11)
Foundation item: National Natural Science Foundation of China (61502094, 61702093); The Talent Training Support Program Funded By the Central Government to Support the Reform and Development of Local Colleges and Universities (140119001); Basic Research Foundation of Heilongjiang Provincial Higher Educations (KYCXTD201903); Higher Education Teaching Reform Research Program of Heilongjiang Province (SJGY20180079, SJGY20190098); Guiding Innovation Fund of Northeast Petroleum University (2020YDL-11)
收稿时间: 2020-11-20; 修改时间: 2020-12-21; 采用时间: 2021-01-07; csa 在线出版时间: 2021-07-31

致, 导致它们之间存在异构性. 而且, 不同媒体类型的数据特征维度高, 导致检索效率低是具有挑战性的问题. 针对此问题, 哈希学习将不同媒体数据从高维表示空间映射到低维汉明空间, 同时将原始数据的相关性尽可能保留到汉明空间, 使在同一语义下的不同媒体数据具有相似的哈希码. 因此, 哈希学习成为研究跨媒体检索的一类代表性方法.

目前主流的跨媒体哈希检索方法主要分为两类:一类是无监督跨媒体哈希方法和有监督跨媒体哈希方法. 其无监督跨媒体哈希是指不使用语义标签信息进行学习, 而是通过捕捉底层数据的结构, 分布以及拓扑信息来学习哈希函数. 例如媒体间哈希 (Inter-Media Hashing, IMH)[3], 协同矩阵分解哈希 (Collective MatrixFactorization Hashing, CMFH)[4], 跨媒体相似检索的潜在语义稀疏哈希 (Latent Semantic Sparse Hashing for cross modal similarity search , LSSH)[5] 等方法. 有监督跨媒体哈希方法主要利用语义标签信息的指导学习哈希函数. 如跨视角哈希 (Cross View Hashing, CVH)[6], 最大语义关联跨媒体检索 (Semantic Correlation Maximization, SCM)[7], 语义保留哈希跨媒体检索 (Semantics Preserving Hashing, SePH)[8] 等方法, 而以上这些方法尽管利用语义标签信息减轻了不同媒体类型数据之间的异构差距, 但是在哈希函数学习的过程中没有使用深层次的特征表示. 深度学习利用神经网络强化媒体之间相关性学习, 可以大幅度提升检索效果. 深度视觉语义哈希 (Deep Visual-Semantics Hashing, DVSH)[9], 通过利用 CNN 和 LSTM 分别提取图像表示和文本表示, 为图像和文本数据分别学习哈希函数, 同时保留了模态内和模态间的相关性. 深度跨模态哈希 (Deep Hashing Cross Modal Retrieval, DCMH)[10] 是这类方法的一个代表, 它是一个端到端的框架, 将图像和文本的特征学习与哈希学习统一起来, 将不同模态间的相关性保留到哈希码, 实现比较好的效果. 再如, 基于三元组的跨模态深度哈希方法[11], 利用 Triplet 损失函数学习图像和文本之间的相似性, 增强对模态间相关性的学习.

综上所述, 为了使得语义相似的媒体对象哈希码的距离较小, 语义不相似的媒体对象哈希码的距离较大, 使得汉明空间和语义空间具备结构性保持, 进而提高模型的检索效果, 本文提出基于柯西分布的深度哈希跨媒体检索方法, 该方法使用基于柯西函数的损失函数, 减小同类别下哈希码之间距离的同时, 增加不同类别间哈希码的距离, 从而提高模型的检索效果.

1 本文方法

本文方法的整体框架示意如图 1 所示, 通过神经网络为不同媒体类型的数据学习哈希函数, 再利用哈希函数将不同媒体类型的数据映射到一个公共的汉明空间, 得到统一的哈希码. 在公共的汉明空间内, 不同于现有的基于交叉熵的关联损失函数, 本文引入基于柯西分布的跨媒体损失函数, 它不但能够缩小语义相似媒体对象的哈希码之间的距离, 而且可以增大语义不相似的媒体对象的哈希码距离, 从而提高跨媒体检索效果.

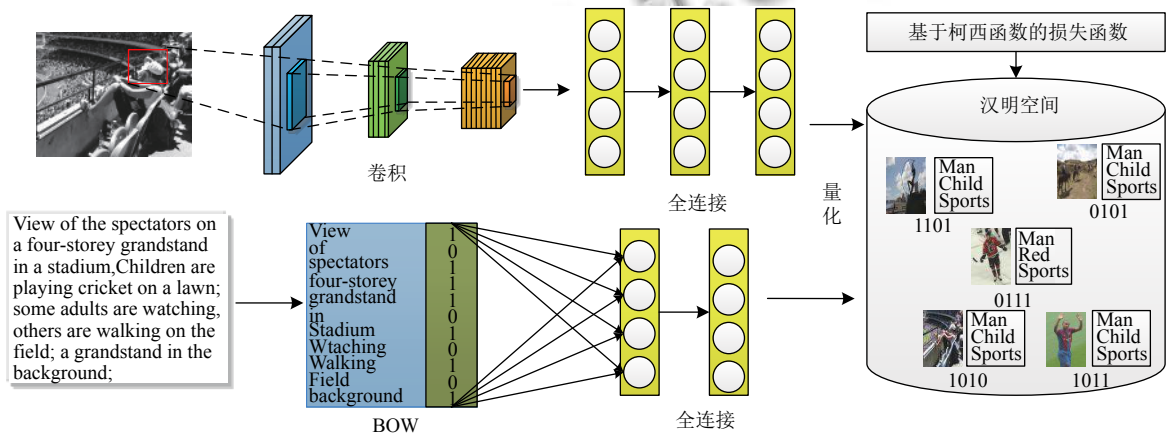


图 1 整个算法的流程示意图

1.1 形式化描述

本文以图像和文本为例进行介绍, 令 $X=\{x_i\}_{i=1}^{n}$ 表示图像集合, x_i 表示第 i 张图像, $Y=\{y_j\}_{j=1}^{n}$ 表示文本集合, y_j 表示第 j 张图像所对应的文本, S 表示图像文本

对的相似矩阵, 如果$S_{ij}=1$, 表示图像和文本相似, 他们至少有一个共同的类, 否则, $S_{ij}=0$, 表示图像和文本不相似, 他们分别属于不同的类.

本文的主要任务是为不同媒体类型的数据学习哈希函数. 设$g^x(x)\in\{-1,1\}^{k\times n}$表示图像的哈希函数, $g^y(y)\in\{-1,1\}^{k\times n}$表示文本的哈希函数, k 表示哈希码的长度. 而哈希码是通过哈希函数将数据映射成二进制码, 则图像的哈希码$b_i^x=g^x(x)$, 文本的哈希码$b_i^y=g^y(x)$. 同时, 本文使用汉明距离表示汉明空间内哈希码之间的相似性, 距离越小哈希码相似程度越高. 若$S_{ij}=1$, 表示哈希码与之间的距离较小, 若$S_{ij}=0$, 表示哈希码与之间的距离较大.

1.2 网络结构

本文的网络框架主要分为两部分, 一部分用于提取图像特征, 另一部分用于提取文本特征.

对于图像数据, 我们对 ResNet-34[12] 做了一些改变, 网络配置如表 1 所示, 总共有 10 层, 其中前 8 层为卷积层, 第 9 为全连接层, 第 10 层是将图像特征映射到汉明空间, 而在第 10 层的特征维度应该与哈希码的长度一致, 每个卷积层内参数的含义如表 1 所示.

表 1 图像神经网络配置

层数	配置
Conv1	kernel:64×7×7,stride:2×2,BN,max_pool:2×2,
Layer1	$\begin{bmatrix}\text{kernel}:64\times3\times3,\text{stride}:1\times1,\\ \text{kernel}:64\times3\times3,\text{stride}:1\times1,\end{bmatrix}\times3$
Layer2	kernel:128×1×1,stride:1×1,
Layer1	$\begin{bmatrix}\text{kernel}:128\times3\times3,\text{stride}:1\times1,\\ \text{kernel}:128\times3\times3,\text{stride}:1\times1,\end{bmatrix}\times4$
Layer2	kernel:256×1×1,stride:1×1,
Layer1	$\begin{bmatrix}\text{kernel}:256\times3\times3,\text{stride}:1\times1,\\ \text{kernel}:256\times3\times3,\text{stride}:1\times1,\end{bmatrix}\times6$
Layer2	kernel:512×1×1,stride:1×1,
Layer1	$\begin{bmatrix}\text{kernel}:512\times3\times3,\text{stride}:1\times1,\\ \text{kernel}:512\times3\times3,\text{stride}:1\times1,\end{bmatrix}\times3$
Layer2	Avg_pool:1×1
Fc1	2048
Fc2	k

“kernel num*size*size”描述了关于卷积核的信息, “num”表示输出通道数, size*size 表示卷积核的大小.

“stride size*size”描述了关于卷积操作的步长, “stride”表示步长大小

“BN[13]”表示对网络层进行归一化

“max_pool:size*size”描述了下采样的大小,

“avg_pool:size*size”描述了下采样的大小.

每一个全连接层的数字. 例如“4096”表示这个全连接层的输出维度, k 表示哈希码长度.

对于一个图像样本 x_i, 本文方法获得哈希码h_i^x是通过阈值函数获得, 即$h_i^x=\text{sgn}(f^x(x_i,\theta_x))$, θ_x为图像网络参数, 由于 sgn 函数它是一个离散的函数, 不能进行反向传播, 由于 tanh 函数的取值范围为 [−1, 1], 同时也能够减少图像网络输出层的值与h_i^x的误差, 因此本文在图像神经网络的输出值使用 tanh 函数.

对于文本数据, 我们使用词袋模型对文本数据进行预处理, 再输入两层玻尔兹曼机获得句子的深度特征表示, 文本神经网络配置如表 2 所示, 前两层的激活函数使用 ReLU, 最后一层使用 tanh 函数, 同时特征长度与哈希码的长度保持一致.

表 2 文本神经网路配置

层数	配置
Layer1	8192
Layer2	4096
Fc1	k

对于每一个文本 y_j, 本文方法获得的哈希码h_j^y是通过阈值函数获得, 即$h_j^y=\text{sgn}(f^y(y_j,\theta_y))$, θ_y为文本网络参数, 与图像神经网络输出层的设置一样, 由于阈值函数不能反向传播, 对文本神经网络输出层的值使用 tanh 函数.

1.3 基于柯西函数的相似度学习

令$\{x_i,y_j\}$表示一组图像和文本数据对, s_{ij}表示x_i与y_j的相似关系, h_i^x和h_j^y分别表示x_i与y_j的哈希码, 由条件概率可知:

$$p\left(s_{ij}|\left(h_i^x,h_j^y\right)\right)=\begin{cases}\sigma(\varphi_{ij}), & s_{ij}=1\\ 1-\sigma(\varphi_{ij}), & s_{ij}=0\end{cases}$$
$$=\sigma(\varphi_{ij})^{s_{ij}}(1-\sigma(\varphi_{ij}))^{s_{ij}} \quad (1)$$

其中, $\varphi_{ij}=dist\left(h_i^x,h_j^y\right)$, $dist\left(h_i^x,h_j^y\right)$表示图像$x_i$的哈希码$h_i^x$与文本$y_j$的哈希码$h_j^y$之间的汉明距离, $\sigma(\varphi_{ij})$是一个概率激活函数. 其次, 结合式 (1), 对图像数据的哈希码h_i^x和文本数据的哈希码h_j^y的负极大似然估计为:

$$L=-\log p(S|(h^x,h^y))$$
$$=\sum_{i=1}^{n}\sum_{j=1}^{n}\log p\left(s_{ij}|h_i^x,h_j^y\right)$$
$$=-\sum_{i=1}^{n}\sum_{j=1}^{n}s_{ij}\log\sigma(\varphi_{ij})+(1-s_{ij})\log(1-\sigma(\varphi_{ij})) \quad (2)$$

根据式 (1), 式 (2), 若$dist\left(h_i^x,h_j^y\right)$越小, h_i^x和h_j^y之间的语义相似度越高, 代表图像x_i与文本y_j具有相似的

语义.

目前方法大多数使用 Sigmoid 函数作为式 (2) 的实现, Sigmoid 函数的定义如下:

$$\sigma(\varphi_{ij}) = \frac{1}{1+\mathrm{e}^{-\varphi_{ij}}} \tag{3}$$

将哈希码之间的汉明距离φ_{ij}映射为 0 和 1 之间的相似度. 图 2 中显示了 Sigmoid 函数的输出随着φ_{ij}的变化情况. 如图 2 所示, 当φ_{ij}小于 $k/2$(k 为哈希码长度) 时, Sigmoid 映射后的相似度值区分能力较弱, 只有当φ_{ij}接近于 $k/2$ 时, 区分能力才较强. 该分析结果说明, Sigmoid 函数对跨媒体检索性能影响较大.

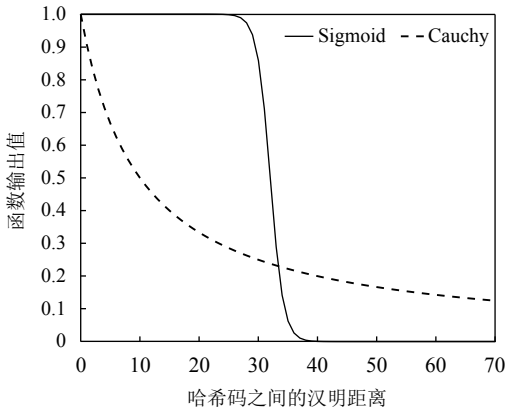


图 2 Sigmoid 函数与 Cauchy 分布的输出随汉明距离φ_{ij}的变化情况

如图 2 所示, 采用 Sigmoid 函数, 当两个媒体对象的汉明距离较小时, 其相似度区分能力很弱. 但是如果采用柯西 (Caughy) 分布作为式 (2) 中$\sigma(\varphi_{ij})$的实现, 当汉明距离小于 $k/2$ 时, 两个函数的输出存在明显的差异, Caughy 分布的输出使得映射后得相似度值区分能力较强, 进而可提高语义相近得媒体对象得检索性能.

$$\sigma(\varphi_{ij}) = \frac{\gamma}{\gamma+\varphi_{ij}} \tag{4}$$

其中, $\varphi_{ij} = dist\left(h_i^x, h_j^y\right) = k*\left(1-\cos\left(h_i^x, h_j^y\right)\right)/2$.

综上所述, 将式 (4) 带入式 (2) 并化简后, 可得改进后得损失函数:

$$L_{cau} = -\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}\log\left(dist\left(h_i^x, h_j^y\right)/\gamma\right) + \log\left(1+\gamma/dist\left(h_i^x, h_j^y\right)\right) \tag{5}$$

1.4 哈希码学习

根据 1.2 节可知, 由于图像和文本神经网络输出层使用 tanh 函数, 图像和文本特征向量的取值范围是 [−1, 1], 所以哈希码存在量化误差, 使用跨媒体哈希码, 需要学习哈希函数, 则哈希码量化损失表示为式 (6):

$$L_q = \alpha\left(\left\|h^x-\mathbf{B}\right\|_F^2 + \left\|h^y-\mathbf{B}\right\|_F^2\right) \quad \text{s.t.} \quad \mathbf{B}\in\{-1,1\}^{k\times n} \tag{6}$$

其中, α为平衡损失函数的参数.

结合式 (4), 式 (6), 得到本文方法的目标函数为:

$$\min_{\mathbf{B},\theta_x,\theta_y} L = \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}\log\left(dist\left(h_i^x, h_j^y\right)/\gamma\right) + \log\left(1+\gamma/dist\left(h_i^x, h_j^y\right)\right) + \alpha\left(\left\|h_x-\mathbf{B}\right\|_F^2 + \left\|h_y-\mathbf{B}\right\|_F^2\right) \quad \text{s.t.} \quad \mathbf{B}\in\{-1,1\}^{k\times n} \tag{7}$$

1.5 目标函数求解

由于目标函数是一个非凸问题, 若在求解一个变量的同时固定其他变量, 此时目标函数就变成凸优化问题, 可以使用梯度求导解决此问题, 因此本文采用一种交替迭代求解的策略获得目标函数的近似最优解, 具体的求解过程如下所示.

(1) 更新θ_x, 固定θ_y, $\mathbf{B}$ 时, 利用反向传播算法学习提取图像特征的 CNN 网络参数θ_x, 对于每一个图像样本x_i, 梯度计算的公式为:

$$\frac{\partial L}{\partial h_i^x} = \sum_{i,j=1}^{n} s_{ij}\left(\frac{\gamma}{dist\left(h_i^x, h_j^y\right)} + \frac{dist\left(h_i^x, h_j^y\right)}{\gamma+dist\left(h_i^x, h_j^y\right)}\right) + 2\alpha\left(b-h_i^x\right) \tag{8}$$

同时, 利用反向传播算法计算$\partial L/\partial\theta_x$.

(2) 更新θ_y, 固定θ_x, $\mathbf{B}$ 时, 还是利用反向传播算法学习提取文本特征的神经网络参数θ_y, 对于每一个文本样本 y_j, 梯度的计算公式为:

$$\frac{\partial L}{\partial h_i^y} = \sum_{i,j=1}^{n} s_{ij}\left(\frac{\gamma}{dist\left(h_i^x, h_j^y\right)} + \frac{dist\left(h_i^x, h_j^y\right)}{\gamma+dist\left(h_i^x, h_j^y\right)}\right) + 2\alpha\left(b-h_i^y\right) \tag{9}$$

同时, 利用反向传播算法计算$\partial L/\partial\theta_y$.

(3) 更新 $\mathbf{B}$, 固定θ_x, θ_y时; 目标函数式 (7) 可以重写为式 (10) 为:

$$\min_{\mathbf{B}} L = \alpha\left(\left\|\mathbf{h}^x-\mathbf{B}\right\|_F^2 + \left\|\mathbf{h}^y-\mathbf{B}\right\|_F^2\right) \quad \text{s.t.} \quad \mathbf{B}\in\{-1,1\}^{k\times n} \tag{10}$$

对式 (10) 进行进一步整理可得如下公式:

$$\min_{\mathbf{B}} L = \alpha\left(2\mathbf{B}^{\mathrm{T}}\mathbf{B} - 2(\mathbf{h}^x+\mathbf{h}^y)^{\mathrm{T}}\mathbf{B} + (\mathbf{h}^x)^{\mathrm{T}}\mathbf{h}^x + (h^y)^{\mathrm{T}}h^y\right) \tag{11}$$

显然, 在上述公式中$tr((\mathbf{h}^y)^{\mathrm{T}}\mathbf{h}^y)$和$tr(\mathbf{B}^{\mathrm{T}}\mathbf{B})$都为常数值, 因此 **B** 的解为:

$$\mathbf{B}=\mathrm{sgn}\pi(\alpha(\mathbf{h}^x+\mathbf{h}^y)) \tag{12}$$

迭代该算法, 直到满足收敛准则.

1.6 外样本扩展

对于那些不在训练集的样本点, 首先将它们转化为哈希码. 特别地, 给一个图像的样本查询点x_q, 与之对应的图像哈希码b_q^x通式 (13) 得到:

$$b_q^x=g^x(x_q)=\mathrm{sgn}(f(x_q,\theta_x)) \tag{13}$$

同理, 对于一个文本的样本查询点y_q, 与之对应的文本哈希码b_q^y可由式 (14) 得到.

$$b_q^y=g^y(y_q)=\mathrm{sgn}(f(y_q,\theta_y)) \tag{14}$$

如 1.1 节所述, 本文方法只是以图文互相检索为例, 事实上, 本文可以扩展为任意两个媒体进行检索, 主要区别在于获取特征的方法.

2 实验结果与分析

本文的基于柯西分布的深度哈希 跨媒体检索方法在 Flickr-25k[14], IAPR TC-12[15], MS_COCO[16] 三个标准数据集上进行试验, 并与最大语义关联哈希 (SCM)[7], 深度跨模态哈希方法 (DCMH)[10], 在图像检索文本, 文本检索图像两个任务进行了性能分析.

2.1 数据集

Flickr-25k 数据集共包含 25 015 张图像组成, 每张图像都有几个文本标记相关联, 每张图片大概有标记 8 个或者 9 个, 数据集总共 24 个类别标签, 都是由人工标注的. 本文选组标记单词出现次数高于 20 的样本作为实验数据, 最终实验数据为 20 015 个图像文本对.

IAPR TC-12 数据集共包含 20 000 张图像以及相对应的文本句子, 总共 275 个类别标签, 通过对数据集预处理之后, 去除没有类别标签的数据, 实验数据总共挑选 19 998 个图像文本对.

MS COCO 数据集共包含 82 785 张训练集图像和 40 504 张验证集图像, 同时每张图像都有 5 条描述的句子, 80 个类别标签, 在本次实验中, 去掉没有类别标签和没有文本的描述的图像, 同时选取最能描述图像的句子作为文本数据, 最终实验数据有 122 218 个图像文本对.

2.2 实验环境的设置及评价指标

本文的实验在深度学习框架 PyTorch 上进行, 对于图像, 使用 ImageNet[17] 的预训练模型初始化图像特征提取网络 ResNet-34, 并对输出层网络参数进行随机初始化, 对于文本, 使用词袋模型对文本数据进行预处理, 然后输入到多层玻尔兹曼机中, 获得其深度特征表示.

本文使用 Rmscrop 对训练网络模型, 学习参数配置如下: 图像网络的初始化学习率为 0. 0 001, 文本网络的初始化学习率 0. 0003, 学习率每训练 15 次迭代后学习率变为当前值的 1/2, 式 (7) 中参数$\alpha=1$, $\gamma=10$.

使用平均精度均值 (Mean Average Precision, MAP) 评价模型, 具体地, 存在一个查询样本 q 及其返回结果的列表, 平均准确率 (Average Precision, *AP*) 的定义为:

$$AP(q)=\frac{1}{N_q}\sum_{m=1}^{n_q}P(m)I(m) \tag{15}$$

其中, N_q表示查询样本 q 在数据库中真正与之相关的样本数目, n_q是查询样本 q 检索数据库返回的结果总数, $P(m)$表示前 m 个检索结果的平均精度, $I(m)$=1表示第 m 个检索样本与查询样本相似, 否则, $I(m)$=0表示第 m 个检索样本与查询样本不相似. 所有查询样本 *AP* 的平均值即为 MAP.

2.3 实验结果及分析

本文方法与其他基准模型在 Flickr-25k, IAPR TC-12, MSCOCO 数据集上 MAP 的结果如表 3 所示. 本次实验主要有两个任务: (1) Text-Image: 表示为图像检索文本, (2) Image-Text: 表示为文本检索图像与当前最好的模型 DCMH[10] 相比, 在 Flickr-25k 数据集上的图像检索文本的任务, 本文方法在哈希码为 16 位时提高了 2.02%, 32 位时提高了 2.11%, 64 位时提高了 1.57%; 同时在文本检索图像时, 本文方法在哈希码 16 位时提高了 3.01%, 32 位时提高了 2.98%, 64 位时提高了 3.41%; 在 IAPR TC-12 数据集上的文本检索图像时, 本文方法在哈希码为 16 位时提高了 3.45%, 在 32 位时 3.88%, 在 64 位时提高了 5.32%, 同时在图像检索文本的任务, 本文方法在哈希码为 16 位时提高了 12.61%, 32 位时提高了 10.29%, 64 位时提高了 13.45%; 在 MSCOCO 数据集上的文本检索图像时, 本文方法在哈希吗 16 位时提高了 8.68%, 32 位时提高了 7.71%, 64 位时提高

了 8.53%, 同时在图像检索文本任务, 本文方法在哈希码为 16 位时提高了 6.80%, 32 位时提高了 4.31%, 64 位时提高了 5.47%. 以上的数据表明了本文方法可以学习到更有判别能力的哈希码.

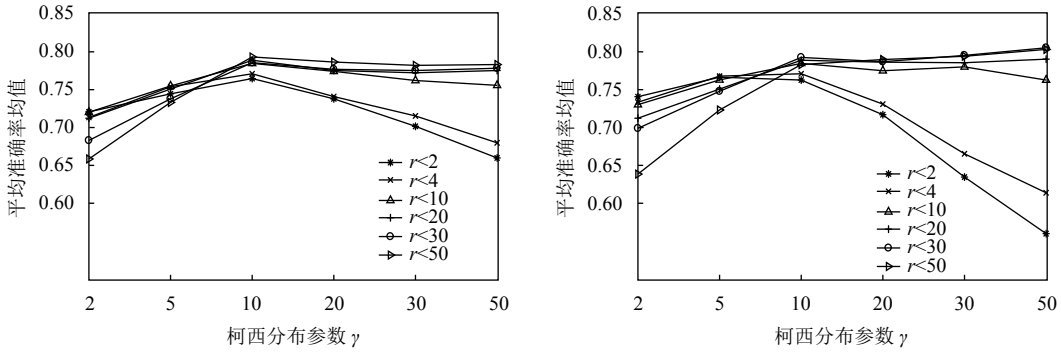
表 3 在 Flickr-25k, IAPR TC-12, MSCOCO 数据集上的 MAP 值

任务	方法	MIRFlickr-25k (bits)			IAPR TC-12 (bits)			MSCOCO (bits)		
		16	32	64	16	32	64	16	32	64
Text	SCM	0.5531	0.5712	0.5778	0.3567	0.3611	0.3870	0.3667	0.3756	0.3870
To	DCMH	0.7464	0.7561	0.7672	0.5185	0.5378	0.5468	0.4753	0.4852	0.4975
Image	Ours	**0.7688**	**0.7787**	**0.7934**	**0.5364**	**0.5587**	**0.5759**	**0.5166**	**0.5266**	**0.5359**
Image	SCM	0.5531	0.5712	0.5778	0.3567	0.3611	0.3815	0.4107	0.4163	0.4335
To	DCMH	0.7345	0.7533	0.7767	0.4852	0.4975	0.4844	0.5028	0.5238	0.5319
Text	Ours	**0.7508**	**0.7692**	**0.7889**	**0.5464**	**0.5487**	**0.5515**	**0.5370**	**0.5464**	**0.5610**

本文方法与 DCMH[10] 都是以监督式的深度学习为基础的. DCMH 方法是基于交叉熵的关联损失函数, 使用 Sigmoid 函数表示不同媒体对象哈希码的语义相似度, 只有汉明距离在 $k/2$ 周围时, 不同媒体对象哈希码的语义相似度才具有判别力, 而本文方法通过引入柯西分布提出基于柯西分布的关联损失函数, 使不同媒体对象哈希码的距离更小, 获取更具有判别力的语义相似度, 进而提升跨媒体哈希检索效果.

2.4 Caughy 参数对性能的影响

为了够验证 Caughy 参数γ与汉明空间内聚集区域大小的关系, 设置r= 2,5,10,20,30,50, 设置哈希码长度为 64 位, 设置哈希码聚集的区域半径为r= 2,4,10,20,30,50. 在 Flickr-25k 数据集实验结果如图 3 所示, 当γ= {2,5,10} 时, 模型检索准确率呈上升趋势, 当γ= {10,20,50}时, r= {2,5}时, 模型的检索准确率在下降, 模型在r= 10时模型比较稳定.



(a) 图像检索文本 (b) 文本检索图像

图 3 不同的汉明距离在不同γ下的准确率

另外, 本文在表 4 和表 5 分别展示了本文方法的文本检索图像和图像检索文本两个任务在 Flickr-25k 数据集上的一些例子. 在表 4 和表 5 中, 最左边的一列代表查询样本的标签, 中间列代表查询样本, 最右边的一列代表检索结果, 哈希码的长度为 64 bit. 表 4 展示文本检索图像的例子, 中间列为图像, 最右边列为图像检索文本的结果, 该结果通过计算查询图像的哈希码与被检索文本哈希码之间的汉明距离, 再按照汉明距离从小到大按顺序排列, 获得与查询图像最相似的文本. 同理. 表 5 展示图像检索像的例子, 中间列为文本, 最右边列为文本检索图像的结果.

3 结论

本文提出了一种基于柯西分布的深度哈希跨媒体检索模型, 它能够产生质量较高哈希码. 通过在 Flickr-25k, IAPR TC-12 和 MSCOCO 三个数据集上与现有方法的对比, 证明本文方法在跨媒体图文检索任务上的有效性. 但本文方法只是图文之间的检索, 下一步工作将他们应用到其他媒体类型数据, 例如图像与视频相互检索, 文本与视频相互检索.

表 4 文本检索图像的例子

标签	查询文本	检索结果
bird,plant_life,animals,flower	explore,naturesfinest,impressedbeauty,bird	
structures,clouds,river,sunset,water,	sky,blue,clouds,white,pink,lake,museum,ice,iceland	

表 5 图像检索文本的例子

标签	查询图像	检索结果
structuressunset		1. sunset, city, sony, sp, mobile, sãopaulo2. sunset, city, nyc, newyork, manhattan, brooklyn3. red, sunset, yellow, black, orange, india, evening, grey, pier, sunday4.hdr, downtown, buildings, day, 'skyscraper, sunny, shanghai
plant_life,clouds,tree,sky		1. clouds, landscape, italy, italia, e500, nuvole2. explore, sky, bw, clouds, explored, grass, wood, road, canoneos400d3. light, landscape, geotagged, england4. sky, landscape, tree, grey, horizon

参考文献

1 庄毅, 庄越挺, 吴飞. 一种支持海量跨媒体检索的集成索引结构. 软件学报, 2008, 19(10): 2667–2680.

2 卓昀侃, 綦金玮, 彭宇新. 跨媒体深层细粒度关联学习方法. 软件学报, 2019, 30(4): 884–895. [doi: 10.13328/j.cnki.jos.005664]

3 Song JK, Yang Y, Yang Y, *et al*. Inter-media hashing for large-scale retrieval from heterogeneous data sources. Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. New York, NY, USA. 2013. 785–796.

4 Ding GG, Guo YC, Zhou JL. Collective matrix factorization hashing for multimodal data. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 2083–2090.

5 Zhou JL, Ding GG, Guo YC. Latent semantic sparse hashing for cross-modal similarity search. Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. Montreal, QC, Canada. 2014. 415–424.

6 Kumar S, Udupa R. Learning hash functions for cross-view similarity search. Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. Barcelona, Spain. 2011. 1360–1365.

7 Zhang DQ, Li WJ. Large-scale supervised multimodal hashing with semantic correlation maximization. Proceedings of the 28th AAAI Conference on Artificial Intelligence, Quebec City, QC, Canada. 2014. 2177–2183.

8 Lin ZJ, Ding GG, Hu MQ, *et al*. Semantics-preserving hashing for cross-view retrieval. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 3864–3872.

9 Cao Y, Long MS, Wang JM, *et al*. Deep visual-semantic hashing for cross-modal retrieval. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. East Lansing, MI, USA. 2016. 1445–1454.

10 Jiang QY, Wu JL. Deep cross-modal hashing. Proceedings of

the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 3270–3278.

11 Deng C, Chen ZJ, Liu XL, *et al*. Triplet-based deep hashing network for cross-modal retrieval. IEEE Transactions on Image Processing, 2018, 27(8): 3893–3903. [doi: 10.1109/TIP.2018.2821921]

12 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 770–778.

13 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lile, France. 2015. 448–456.

14 Huiskes MJ, Lew MS. The Mir flickr retrieval evaluation. Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval. British Columbia, Vancouver, BC, Canada. 2008. 39–43.

15 Escalante HJ, Hernández CA, González JA, *et al*. 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]

16 Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland. 2014. 740–755.

17 Russakovsky O, Deng J, Su H, *et al*. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211–252. [doi: 10.1007/s11263-015-0816-y]