中国科学: 信息科学 2021 年 第 51 卷 第 12 期: 2053–2068
SCIENTIA SINICA Informationis
论文

基于多头注意力网络的无监督跨媒体哈希检索

李志欣[1*], 凌锋[1], 唐振军[1], 马慧芳[2], 施智平[3]

1. 广西师范大学广西多源信息挖掘与安全重点实验室, 桂林 541004
2. 西北师范大学计算机科学与工程学院, 兰州 730070
3. 首都师范大学信息工程学院与交叉科学研究院, 北京 100048
* 通信作者. E-mail: lizx@gxnu.edu.cn

收稿日期: 2020–08–22; 修回日期: 2020–10–18; 接受日期: 2020–11–18; 网络出版日期: 2021–12–09

国家自然科学基金 (批准号: 61663004, 61966004, 61866004, 61962008, 61762078, 61876111) 和广西自然科学基金 (批准号: 2019GXNSFDA245018) 资助项目

摘要 跨媒体哈希检索将不同媒体数据编码到公共二值哈希空间中, 从而可以有效地测量不同模态样本之间的相关性. 为了进一步提高检索性能, 提出基于多头注意力网络的无监督跨媒体哈希检索方法. 首先, 利用多头注意力网络生成哈希码矩阵, 使图像和文本能获得更好的匹配. 其次, 构造一个辅助相似度矩阵, 用以整合来自不同模态的原始邻域信息. 通过辅助相似度矩阵与哈希码矩阵的协同学习, 能够捕获不同模态之间和相同模态内部的潜在联系. 此外, 设计了两种损失函数训练网络模型, 并使用批量归一化和更换哈希码生成函数的策略对模型进行优化, 使模型的训练速度得到大幅提升. 在 3 个数据集上的实验表明, 本方法的平均性能比目前国际上先进的无监督方法有显著提升, 充分证明了本方法的有效性和优越性.

关键词 卷积神经网络, 多头注意力网络, 跨媒体哈希检索, 无监督学习, 协同学习, 辅助相似度矩阵, 批量归一化

1 引言

跨媒体检索是人工智能与信息检索研究中至关重要的一个环节, 旨在根据任意类型的媒体数据从不同的多媒体数据中搜索语义相关的实例 [1]. 例如, 使用图像来检索相关的文本描述或根据文本检索相关的视频和音频. 随着多媒体数据的日益增长, 使用传统方法 (如关键词法) 对多媒体数据进行生硬匹配难以取得好的效果, 利用机器学习技术处理多媒体数据成为一个好的选择. 二值向量存储效率高且能使用汉明距离 (Hamming distance) 进行度量, 对具有类似二值特征的样本进行编码在跨媒体检索中变得越来越重要. 由于图像和文本是最常用最具代表性的媒体数据, 图像和文本的跨媒体检索

引用格式: 李志欣, 凌锋, 唐振军, 等. 基于多头注意力网络的无监督跨媒体哈希检索. 中国科学: 信息科学, 2021, 51: 2053–2068, doi: 10.1360/SSI-2020-0264
Li Z X, Ling F, Tang Z J, et al. Unsupervised cross-media Hashing retrieval based on multi-head attention network (in Chinese). Sci Sin Inform, 2021, 51: 2053–2068, doi: 10.1360/SSI-2020-0264

也可以相对容易地引申到视频音频等其他媒体, 所以跨媒体图像 – 文本检索成为引人注目的研究课题 [2~5].

不同模态数据之间存在的 “异构鸿沟” 限制了不同模态样本之间相关性的直接度量, 因此跨媒体哈希 (Hash) 检索方法提出将原始数据嵌入到通用二值哈希空间中, 不同模态样本之间的相关性可以有效地利用汉明距离进行度量. 跨媒体哈希检索方法可以大致分为无监督方法和有监督方法两类. 无监督跨媒体哈希检索方法通常将不同模态的数据映射到一个公共哈希空间, 仅利用输入图像 – 文本对的共现 (co-occurence) 信息来最大化它们在公共哈希空间中的相关性. 有监督跨媒体哈希检索方法则进一步利用语义标签为语义相关的跨媒体数据学习更多一致的哈希码, 从而能显著缩减异构鸿沟并获得良好的检索性能. 但是, 当前的有监督跨媒体哈希检索方法虽然性能较高但是实用性有限, 因为现实中存在海量的没有语义标签的多媒体数据, 而人工标注费时费力, 故很多情况下需要采用无监督方法进行学习.

由于深度神经网络 (deep neural network, DNN) [6, 7] 可以生成更多语义相关的特征和哈希码, 因此使用深度跨媒体哈希方法能够进一步提高检索性能, 并具有占用内存小、查询速度快等优点. 在基于 DNN 的跨媒体检索中, 大多数方法主要关注对不同模态的单个实例进行成对关系建模, 没有充分利用整体语义信息, 并忽略了相同模态内部的相关性. 如何学习有效的公共空间以缩小语义相关实例之间的距离并扩大语义不相关实例之间的距离是跨媒体检索迫切需要解决的问题. 为此, 本文基于多头注意力网络 [8] 处理不同模态的原始数据, 能够更好地生成二值哈希码, 同时构造辅助相似度矩阵, 并通过与哈希码矩阵的协同学习进行整体关系建模, 可以获得各个实例之间的潜在语义相似度, 从而提高检索精度. 此外, 训练过程中利用批量归一化 [9] 进行加速, 增强了本文方法的效率和实用性.

本文提出的方法利用原始的图像和文本生成辅助相似度矩阵, 用于指导训练样本生成哈希码矩阵. 然后, 利用辅助相似度矩阵和哈希码矩阵进行协同学习, 以保证检索的准确性. 实验表明, 辅助相似度矩阵在提升性能方面起着至关重要的作用, 它能够拟合原始数据特征, 与基于 DNN 生成的哈希码相互补充, 实现特征协同学习 [10, 11]. 在生成哈希码矩阵和辅助相似度矩阵的过程中, 多头注意力网络能够为图像和文本两种模态数据生成一致的表示, 起到了至关重要的作用. 因此, 本文方法称为 “基于多头注意力网络的无监督跨媒体哈希检索” (unsupervised cross-media Hashing retrieval based on multi-head attention network, UMHA), 主要贡献包括:

(1) 提出利用多头注意力网络为图像和文本两种不同模态的数据生成哈希码, 构建公共哈希空间, 得到哈希码矩阵, 从而为跨媒体哈希检索奠定基础.

(2) 提出辅助相似度矩阵以结合不同模态的信息, 整合来自不同模态的原始邻域关系, 因而能够捕获实例之间的潜在语义相似度. 辅助相似度矩阵与哈希码矩阵协同学习, 能获得更优的检索结果.

(3) 对于生成哈希码的网络, 设计了两种损失函数, 并使用 tanh 函数优化训练过程, 能够大幅提升训练速度, 比传统的拉普拉斯 (Laplace) 约束方案更适合于分批训练.

在 3 个数据集上进行的一系列实验表明, UMHA 的性能显著高于当前先进的跨媒体图像 – 文本检索方法. 此外, 消融实验证明了 UMHA 中各个组件的有效性. 值得注意的是, UMHA 是无监督的方法, 一旦训练完成, 就能够应用到大规模数据中, 因而具有实际应用的潜力.

2 相关工作

跨媒体检索的关键在于跨媒体语义映射, 也就是根据不同模态数据的特性设计能有效将特征空间映射到语义空间的模型或算法, 使得不同模态数据能够在公共的语义空间上进行相似度计算或匹配,

从而具备跨媒体特性. 图像/视频描述生成 [12] 等相关任务也包含跨媒体语义映射的环节, 所以其编解码学习架构、特征映射和注意机制等关键技术普遍应用于跨媒体检索系统中. 此外, 跨媒体检索还涉及异构数据的相关性分析、排位函数学习和相关反馈等技术.

跨媒体哈希检索的基本思想是利用哈希变换将不同模态样本的特征映射到一个公共的汉明二值空间, 然后基于汉明空间实现快速的相似度计算和检索. 根据是否使用深度学习技术, 跨媒体哈希检索方法可大致分为浅层方法 [13~22] 和深层方法 [23~36]. 浅层方法通常采用传统方法学习得到哈希码并构建公共空间. 跨模态相似度敏感哈希方法 (cross-modal similarity-sensitive Hashing, CMSSH) [13] 根据两种不同模态的数据学习两组哈希函数, 以确保不同模态的两个相关数据点对应的哈希码是相似的. 但 CMSSH 只保留了模态间的相关性, 忽略了模态内的相似性. 跨视图哈希方法 (cross-view Hashing, CVH) [14] 将频谱哈希从传统的单模态设置扩展到多模态场景, 通过最小化同类数据的距离和最大化异类数据的距离来生成哈希码, 同时保持了模态间的相关性和模态内的相似性. 媒体间哈希方法 (inter-media Hashing, IMH) [15] 研究来自不同数据源的多个媒体类型之间的相关性, 并处理可扩展性问题. IMH 整合一个线性回归模型来学习哈希函数, 能够为新数据点有效地生成哈希码. 集体矩阵分解哈希方法 (collective matrix factorization Hashing, CMFH) [16] 通过利用潜在因子模型的集体矩阵分解为不同模态数据学习统一的哈希码. CMFH 不仅支持跨视图搜索, 还通过合并多个视图信息源提高搜索精度. 潜在语义稀疏哈希方法 (latent semantic sparse Hashing, LSSH) [17] 分别利用稀疏编码和矩阵分解来提取图像和文本的潜在特征, 然后再将这些特征映射到公共空间并量化为统一的二值码. 可预测双视图哈希方法 (predictable dual-view Hashing, PDH) [18] 将数据样本的邻近性嵌入到原始空间中, 以保持预生成二值码的可预测性, 并利用基于块坐标下降算法的迭代方法优化目标函数. 组合相关性量化方法 (composite correlation quantization, CCQ) [19] 学习将不同模态转换为同构潜在空间的相关性最大化映射, 同时学习将同构潜在特征转换为紧凑二值码的复合量化器. 语义相关性最大化方法 (semantic correlation maximation, SCM) [20] 基于标签构造语义相似度矩阵, 并学习哈希函数来保持该矩阵, 将语义标签无缝地集成到哈希学习过程中. 深层方法基于深度学习方法生成哈希码, 可以探索不同模态间复杂的非线性相关性, 在缩减模态鸿沟与提取深度语义特征方面具有更卓越的能力. 大多数方法 [23~27] 采用基于 DNN 的有监督学习框架生成哈希码; 也有一些方法 [28, 29] 采用自动编码器等方法来探索多模态学习, 为异构数据生成统一的潜在二值码. 此外, 另一些方法基于无监督学习策略生成哈希码, 有些方法 [30~33] 关注 DNN 训练过程的优化, 期望能够生成更有效的哈希码; 有些方法 [34~36] 使用对抗学习策略来训练 DNN, 试图捕获不同模态相关实例的特征分布并利用生成对抗网络 (generative advesarial net, GAN) [37, 38] 缩小其距离. 深度视觉语义哈希方法 (deep visual-semantic Hashing, DVSH) [23] 在一个端到端的深度学习体系结构中生成图像和语句的紧凑哈希码, 捕捉视觉数据与自然语言之间内在的跨模态对应关系. DVSH 是一种混合的深层架构, 它包含一个视觉语义融合网络和一个哈希网络, 前者用于学习图像和文本语句的联合嵌入空间, 后者用于学习哈希函数以生成紧凑的哈希码. 深度跨模态哈希方法 (deep cross-modal Hashing, DCMH) [24] 将特征学习和哈希码学习集成到同一框架中. DCMH 是一个端到端的学习框架, 对每种模态数据都使用一个 DNN 进行从零开始的特征学习. 深度二值重构方法 (deep binary reconstruction, DBRC) [30] 首次利用原始邻域矩阵和跨媒体哈希码矩阵进行有效结合, 实现了端到端的训练, 并取得了良好的效果. 无监督深度跨模态哈希方法 (unsupervised deep cross-modal Hashing, UDCMH) [31] 将矩阵分解和拉普拉斯约束组合到网络训练中, 利用显式约束哈希码以保留原始数据的邻域结构, 获得了更优越的检索结果. 自监督对抗性哈希方法 (self-supervised adversarial Hashing, SSAH) [34] 尝试以一种自监督的方式将对抗性学习引入跨模态哈希生成. SSAH 利用两个对抗性网络来最大化不同模态之间的语义相关性和一致性, 同时利用一

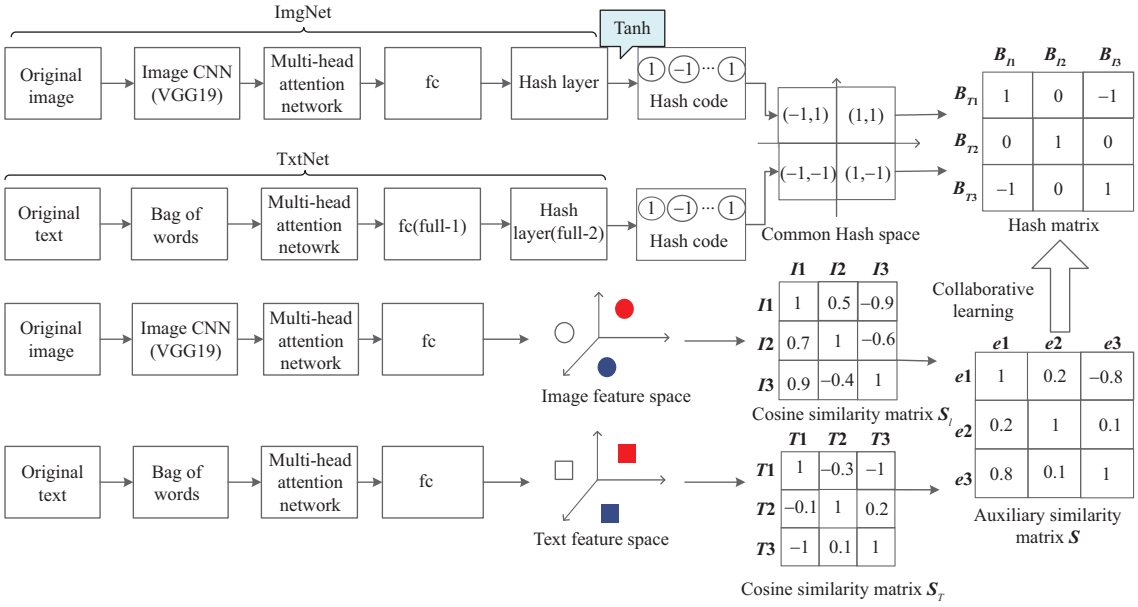

图 1　(网络版彩图) UMHA 模型结构

Figure 1　(Color online) Model structure of UMHA

个自监督语义网络以多标签标注的形式发现高层语义信息. 多路径生成对抗哈希方法 (multi-pathway generative adversarial Hashing, MGAH) [36] 利用 GAN 的强大能力进行无监督表示学习, 充分挖掘了跨模态数据的底层流形结构. MGAH 提出了多路径 GAN, 以无监督的方式生成跨模态哈希码, 并提出了一种基于相关图的方法来捕捉不同模态下的流形结构, 使得同一流形内不同模态的数据具有更小的汉明距离, 以提高检索精度.

综上所述, 目前几乎没有工作探究多头注意力网络 [8] 在跨媒体哈希检索上的应用. 本文首次将多头注意力网络应用到无监督跨媒体哈希检索中, 并取得了良好的效果. 此外, 探讨了对图像网络的每个卷积层进行批量归一化 [9], 能在跨媒体哈希检索上加速训练并控制数值区间, 使得 DNN 的训练稳定性更好, 不容易产生梯度爆炸, 从而有利于最终的实验结果.

3　跨媒体检索模型

UMHA 构建的检索模型从不同模态的数据中提取跨媒体的哈希码矩阵和辅助相似度矩阵, 可以为跨媒体学习提供丰富而又全面的语义信息, 其模型结构如图 1 所示. 由图可见, 上方的网络又分为图像网络 ImgNet 和文本网络 TxtNet, 分别用卷积神经网络 (convolutional neural network, CNN) [6, 7] 和词袋 (bag of word, BOW) 模型 [24] 处理图像和文本, 通过多头注意力网络生成图像和文本的哈希码, 进而构建公共哈希空间得到哈希码矩阵, 体现了不同模态数据之间的相关性; 下方的网络也对图像和文本分别处理, 获得图像和文本模态内部的余弦相似度矩阵, 进而生成辅助相似度矩阵. 最后辅助相似度矩阵与哈希码矩阵协同学习以计算跨媒体实例的相关性, 从而提升检索性能.

由于 UMHA 侧重于分批训练, 因此变量也用分批方式表示. 本文使用 $E = \{e_1, e_2, \ldots, e_n\}$ 代表每个批次 E 中包含 n 个实例 (n 表示批次大小), 每个实例由图像 – 文本对 $e_k = [i_k, t_k]$ 来表示.

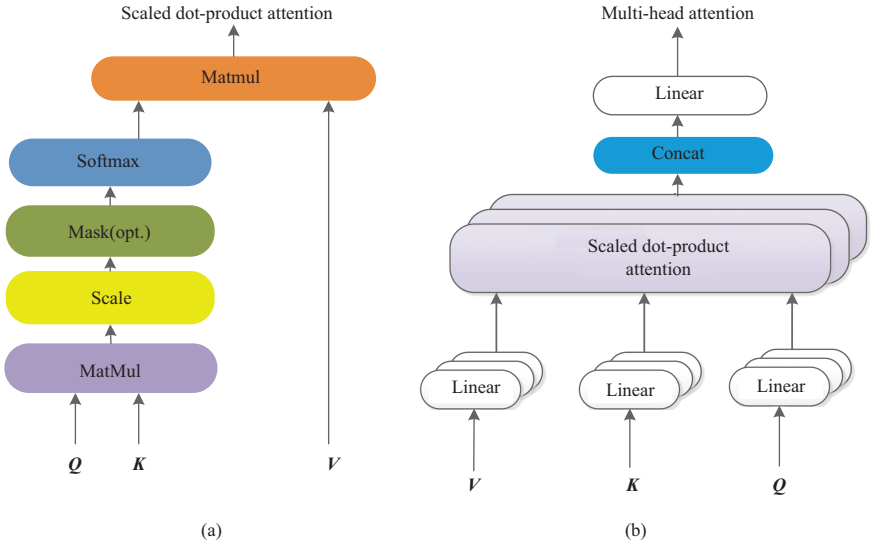

图 2 (网络版彩图) 多头注意力网络结构

Figure 2 (Color online) Structure of multi-head attention network. (a) Scaled dot-product attention; (b) multi-head attention

3.1 多头注意力网络

多头注意力网络 [8] 采用两层感知机实现, 在编码时不需要额外的条件, 对每个实例仅编码一次, 从而大大减少了推理时的计算开销. 多头注意力网络结构如图 2 所示.

图 2(a) 是缩放点积注意力的结构, 输入包括 d_q 维的查询 $\boldsymbol{Q}$、d_k 维的键 $\boldsymbol{K}$ 和 d_v 维的值 $\boldsymbol{V}$. 首先计算查询 $\boldsymbol{Q}$ 和所有键 $\boldsymbol{K}$ 的点积, 再除以 $\sqrt{d_k}$, 然后应用 softmax 函数获得值 $\boldsymbol{V}$ 的权重. 输出可通过下式计算:

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}. \tag{1}$$

图 2(b) 是多头注意力网络结构, 允许模型共同关注来自不同位置的不同表征子空间信息, 平均值会抑制注意力集中于一个头. 其计算公式如下:

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Concat}(\mathbf{head}_1,\ldots,\mathbf{head}_h)\boldsymbol{W}^O, \tag{2}$$

这里 $\mathbf{head}_i = \text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V)$, 参数矩阵 $\boldsymbol{W}_i^Q \in \mathbb{R}^{d_{\text{model}}\times d_q}$, $\boldsymbol{W}_i^K \in \mathbb{R}^{d_{\text{model}}\times d_k}$, $\boldsymbol{W}_i^V \in \mathbb{R}^{d_{\text{model}}\times d_v}$ 及 $\boldsymbol{W}^O \in \mathbb{R}^{hd_v\times d_{\text{model}}}$.

3.2 生成公共哈希空间

为生成图像的哈希码表示, 将每个输入图像的大小调整为 224×224, 送到 CNN 以获取高维图像特征. 本文使用的网络与 VGG19 [7] 具有相同的配置, 并在每个卷积层后进行批量归一化 [9]. 进行归一化处理的原因有两个: 一方面, DNN 的本质是学习数据分布的规律, 一旦训练数据与测试数据的分布不同, 那么网络的泛化能力就会大大降低; 另一方面, 如果每批次的训练数据分布各不相同, 那么 DNN 在每次迭代都需要学习以适应不同的分布, 会大大降低网络的训练速度.

VGG19 的 fc7 层生成 4096 维的特征向量, 可作为图像的原始高级语义表示, 随后将这个特征向量输入多头注意力网络进行处理. 给定局部特征 $\boldsymbol{\varphi}(\boldsymbol{x}) \in \mathbb{R}^{n\times D}$ 对应键 $\boldsymbol{K}$, 权重矩阵 $\boldsymbol{\omega}_1 \in \mathbb{R}^{A\times D}$ 对应

表 1　处理文本数据的网络配置

Table 1　Deep neural network configuration for processing text data

Layer	Configuration
Attention network	BOW vector length 1000
Full connected layer-1	4096
Full connected layer-2	Hash code length d

查询 $\boldsymbol{Q}$, $\boldsymbol{\omega}_2 \in \mathbb{R}^{n\times A}$ 对应值 $\boldsymbol{V}$, 根据经验令 $A = D/2$, 则可以计算 n 个注意力图 $\boldsymbol{\rho} \in \mathbb{R}^{n\times n}$:

$$\boldsymbol{\rho} = \mathrm{softmax}(\boldsymbol{\omega}_2 \tanh(\boldsymbol{\omega}_1 \boldsymbol{\varphi}(\boldsymbol{x})^{\mathrm{T}})). \tag{3}$$

然后逐行应用 softmax 函数, 使 n 行注意系数的每一行加起来都等于 1. 最后, 将注意力图 $\boldsymbol{\rho} \in \mathbb{R}^{n\times n}$ 与局部特征 $\boldsymbol{\varphi}(\boldsymbol{x}) \in \mathbb{R}^{n\times D}$ 相乘, 并进一步应用非线性变换以获得 n 个局部引导表示 $\boldsymbol{\gamma}(\boldsymbol{x}) \in \mathbb{R}^{n\times H}$, 即

$$\boldsymbol{\gamma}(\boldsymbol{x}) = \sigma((\boldsymbol{\rho}\boldsymbol{\varphi}(\boldsymbol{x}))\boldsymbol{\omega}_3 + \boldsymbol{b}_3), \tag{4}$$

其中 $\boldsymbol{\omega}_3 \in \mathbb{R}^{D\times H}$, $\boldsymbol{b}_3 \in \mathbb{R}^H$, 激活函数使用 sigmoid 函数 $\sigma(\cdot)$.

为了生成文本的哈希码表示, 采用向量空间的 TF-IDF (term frequency-inverse document frequency) 特征来表示文本, 并利用 BOW 模型进行处理统一表示为 1000 维的向量. 然后, 通过多头注意力网络得到文本的语义表示, 并将这个表示向量作为两个全连接层的输入. 表 1 给出了文本网络各层的配置, 其中 3 个层的激活函数分别是 sigmoid 函数、ReLU 函数与恒等函数, 4096 维向量是为了对齐 VGG19 神经网络的输出维数, 也就是图 1 中 fc 层的输出维数, 最终得到与图像表示对齐的文本哈希码输出.

对于生成哈希码的网络, 如果将最后一个隐藏层的输出表示为 $\boldsymbol{H} \in \mathbb{R}^{n\times d}$ (在 ImgNet 中表示为 $\boldsymbol{H}_I$, 在 TxtNet 中表示为 $\boldsymbol{H}_T$), 可以通过以下方式生成严格的二值哈希码:

$$\boldsymbol{B} = \mathrm{sgn}(\boldsymbol{H}) \in \{-1, +1\}^{n\times d}, \tag{5}$$

其中 $\mathrm{sgn}(\cdot)$ 是符号函数, d 表示编码长度, 取值为 16, 32, 64 和 128. 于是, 图像和文本两种不同模态的原始数据经过各自的处理, 生成了公共的哈希空间, 为计算不同模态实例的相关性提供了基础.

令 $\boldsymbol{B}_I \in \{-1, +1\}^{n\times d}$ 和 $\boldsymbol{B}_T \in \{-1, +1\}^{n\times d}$ 表示图像 i_k 和文本 t_k 经过 ImgNet 和 TxtNet 最终生成的二值码, 则 $\boldsymbol{B}_I$ 和 $\boldsymbol{B}_T$ 可视为超立方体顶点的特征向量. 从这个角度来看, 相邻的顶点对应于相似的哈希码. 也就是说, 两个二值码之间的汉明距离可以由它们的角距离表示. 因此, 为了描述汉明空间中的邻域结构, 本文计算成对的余弦相似度矩阵 $\cos(\boldsymbol{B}_I, \boldsymbol{B}_T) \in [-1, +1]^{n\times n}$ 来表示每一个图像 i 和文本 j 的哈希码之间的互余弦关系, 即 $(\cos(\boldsymbol{B}_I, \boldsymbol{B}_T))_{ij} = \frac{\boldsymbol{B}_{Ii}\boldsymbol{B}_{Tj}^{\mathrm{T}}}{\|\boldsymbol{B}_{Ii}\|_2 \|\boldsymbol{B}_{Tj}\|_2} \in [-1, +1]$, 其中 $\boldsymbol{B}_{Ii}$ 表示 $\boldsymbol{B}_I$ 中的第 i 行, $\boldsymbol{B}_{Tj}$ 表示 $\boldsymbol{B}_T$ 中的第 j 行. 这个矩阵反映了所生成二值码之间的角度关系, 等同于汉明距离关系, 也就是表达了跨媒体哈希码之间的相似度.

3.3　构造辅助相似度矩阵

对于每个随机抽样训练批次 $\{e_k = [i_k, t_k]\}_{k=1}^n$, 原始图像 i_k 和原始文本 t_k 分别经过图像卷积神经网络、文本词袋模型网络和多头注意力网络的一系列变换后得到输出, 再经过全连接网络后得到相应的特征表示. 图像和文本的最终特征分别用 $\boldsymbol{F}_I \in \mathbb{R}^{n\times p_i}$ 和 $\boldsymbol{F}_T \in \mathbb{R}^{n\times p_t}$ 表示. 对 $\boldsymbol{F}_I$ 和 $\boldsymbol{F}_T$ 中的每一行进行归一化, 得到 $\widehat{\boldsymbol{F}}_I$ 和 $\widehat{\boldsymbol{F}}_T$, 然后计算余弦相似度矩阵 $\boldsymbol{S}_I = \widehat{\boldsymbol{F}}_I \widehat{\boldsymbol{F}}_I^{\mathrm{T}} \in [-1, +1]^{n\times n}$ 和 $\boldsymbol{S}_T = \widehat{\boldsymbol{F}}_T \widehat{\boldsymbol{F}}_T^{\mathrm{T}} \in [-1, +1]^{n\times n}$, 分别用于描述输入图像和文本的原始邻域结构.

通过学习原始数据邻域结构得到的二值哈希码, 能够对深度哈希网络的无监督训练进行有效改进 [31]. 具体地说, 对于跨媒体检索任务, 给定分批输入实例 e_k, 可以计算余弦相似度矩阵 $\boldsymbol{S}_I$ 和 $\boldsymbol{S}_T$, 以不同的形式描述原始的图像和文本语义结构, 然后利用这两个余弦相似度矩阵来指导哈希码学习 i_k 和 t_k, 这个过程中, 如何进行 $\boldsymbol{S}_I$ 和 $\boldsymbol{S}_T$ 的训练对提升算法整体性能占有重要地位.

来自不同模态的相似度矩阵通常彼此互补, 对其进行有效整合可以获取更准确的邻域描述. 为此, 本文提出使用辅助相似度矩阵 $\boldsymbol{S}=\varepsilon(\boldsymbol{S}_I,\boldsymbol{S}_T)\in[-1,+1]^{n\times n}$ 来整合邻域信息, 其中每个元素 $\boldsymbol{S}_{ij}\in[-1,+1]$ 表示输入实例 e_i 和 e_j 之间捕获的潜在语义相关性.

为了引入组合函数 ε 计算辅助相似度矩阵 $\boldsymbol{S}$, 首先通过加权求和方式合并 $\boldsymbol{S}_I$ 和 $\boldsymbol{S}_T$, 即

$$\tilde{\boldsymbol{S}}=\alpha\boldsymbol{S}_I+(1-\alpha)\boldsymbol{S}_T, \tag{6}$$

其中 $\alpha\in[0,1]$ 为权值参数, $\tilde{\boldsymbol{S}}\in[-1,+1]^{n\times n}$. $\tilde{\boldsymbol{S}}$ 中的每一行表示每个实例的新关系, 记录每个实例与其他实例之间的关联信息. 然后, 基于两个语义相关实例应与其他实例共享相同关联的原理计算 $\tilde{\boldsymbol{S}}\tilde{\boldsymbol{S}}^{\mathrm{T}}$, 实现图像和文本实例之间的邻域描述. 也就是说, 它们在 $\tilde{\boldsymbol{S}}$ 中各自行之间的点积结果具有很大的辅助作用. 最后, 设计的辅助相似度矩阵通过下式计算:

$$\begin{aligned}\boldsymbol{S}&=\varepsilon(\boldsymbol{S}_I,\boldsymbol{S}_T)=(1-\theta)\tilde{\boldsymbol{S}}+\theta\frac{\tilde{\boldsymbol{S}}\tilde{\boldsymbol{S}}^{\mathrm{T}}}{\boldsymbol{n}}\\&=(1-\theta)[\alpha\boldsymbol{S}_I+(1-\alpha)\boldsymbol{S}_T]+\frac{\theta}{n}\left[\alpha^2\boldsymbol{S}_I\boldsymbol{S}_I^{\mathrm{T}}+\alpha(1-\alpha)\boldsymbol{S}_I\boldsymbol{S}_T^{\mathrm{T}}+\alpha(1-\alpha)\boldsymbol{S}_T\boldsymbol{S}_I^{\mathrm{T}}+(1-\alpha)^2\boldsymbol{S}_T\boldsymbol{S}_T^{\mathrm{T}}\right].\end{aligned} \tag{7}$$

式 (7) 把邻域信息矩阵 $\tilde{\boldsymbol{S}}$ 进行整合, 然后除以批量大小 n 进行归一化, 得到 $\frac{\tilde{\boldsymbol{S}}\tilde{\boldsymbol{S}}^{\mathrm{T}}}{n}\in[-1,+1]^{n\times n}$. 式中 θ 是权衡参数, 用于调整图像和文本的重要性.

与单独的协同训练方式相比, 式 (6) 与 (7) 以更高级的方式组合了不同媒体的关联信息. 辅助相似度矩阵 $\boldsymbol{S}\in[-1,+1]^{n\times n}$ 从不同角度描述图像和文本的关系 ($\boldsymbol{S}_I$, $\boldsymbol{S}_T$ 和高阶邻域描述 $\tilde{\boldsymbol{S}}\tilde{\boldsymbol{S}}^{\mathrm{T}}$), 这对于捕获输入实例之间潜在的语义关系非常有帮助. 因此, 使用辅助相似度矩阵学习与不同模态数据语义相关的二值码作为自我监督的信号, 可以有效地学习得到一致的表示形式, 从而提高检索性能. 此外, 辅助相似度矩阵包含了大量的跨媒体关联信息, 要判断图像 i_j 和文本 t_k 是否相关, 可通过一些语义信息 (例如类标签) 的关联进行判断. 例如, 如果 $\boldsymbol{S}_j\boldsymbol{S}_k^{\mathrm{T}}=1$ 说明图像 i_j 和文本 t_k 共享相同的类标签, 则可认为它们相关. 否则, 如果 $\boldsymbol{S}_j\boldsymbol{S}_k^{\mathrm{T}}=0$ 或负数则说明图像 i_j 和文本 t_k 来自不同的类, 可认为它们不相关.

3.4 模型训练与优化

由于构造了辅助相似度矩阵 $\boldsymbol{S}$ 为批量输入实例挖掘潜在的语义关系, 所以可通过最小化矩阵 $\boldsymbol{S}$ 和待学习哈希码相似度 $\cos(\boldsymbol{B}_I,\boldsymbol{B}_T)$ 之间的误差来学习语义相关的二值码, 即

$$\begin{aligned}&\mathrm{loss1}=\min_{\boldsymbol{B}_I,\boldsymbol{B}_T}\|\eta\boldsymbol{S}-\cos(\boldsymbol{B}_I,\boldsymbol{B}_T)\|_{\mathrm{F}}^2,\\&\text{s.t. }\boldsymbol{S}=\varepsilon(\boldsymbol{S}_I,\boldsymbol{S}_T)\in[-1,+1]^{n\times n}.\end{aligned} \tag{8}$$

关于这个损失函数要注意两点: (1) 添加了超参数 η, 使图像和文本的语义结合方式更加灵活; (2) 提出了新的方法进行批次训练, 比拉普拉斯约束方案 [31] 更兼容特定的相似度矩阵.

3.4.1 超参数 η 的影响

本文希望模型能生成合适的哈希码来与辅助相似度矩阵 $\boldsymbol{S}$ 进行协同学习. 显然, 在跨媒体哈希空间中, 3 位的哈希码有 $2^3=8$ 种情况, 分别对应三维立方体的 8 个顶点. 那么, n 维的哈希码对应 2^n 种

情况. 简单起见, 这里以 2 位哈希码为例. 在 2 位哈希码的情况下, 哈希码只能在 $(+1,+1)$, $(+1,-1)$, $(-1,+1)$ 和 $(-1,-1)$ 的位置上, 对应的余弦相似度只能是 “−1”, “0” 和 “+1” 关系. 当希望让辅助相似度矩阵结构 $\boldsymbol{S}\in[-1,+1]^{n\times n}$ 与这些 2 位哈希码进行协同学习, 可在汉明空间中将 $\boldsymbol{S}$ 中的原始相似度范围 $[0.5,1]$ 分配为 “+1” 关系, 即促使相应的图文对采用相同的二值码. 同样, $(-0.5,0.5)$ 将被分配为 “0”, $[-1,-0.5]$ 将被分配为 “−1”. 但是, 上述量化过程太僵硬, 无法学习合理的哈希码. 调整辅助相似度矩阵 $\boldsymbol{S}$ 可能更好地与二值码相似度矩阵 $\cos(\boldsymbol{B}_I,\boldsymbol{B}_T)$ 进行协同学习, 而不是让辅助相似度矩阵 $\boldsymbol{S}$ 进行僵硬的匹配.

为解决这个问题, 添加了一个超参数 η 来实现辅助相似度矩阵的缩放, 能够与二值码相似度矩阵进行更好的匹配, 并最小化它们之间的距离. 以关系 “+1” 为例, $\eta>1$ 表示扩展原始范围 $[0.5,1]$, 从而能够以 “+1” 关系量化更多图像 – 文本对, 并因此具有相同的哈希码, 而 $\eta<1$ 表示相反地缩小 “+1” 范围. 因此, 所提出的式 (8) 中的参数 η 有助于调节 $\boldsymbol{S}$ 的量化区域, 从而提高了本文框架的灵活性.

3.4.2 分批训练方法

现有的拉普拉斯约束方案 $\mathrm{Tr}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{B})=\sum_{i,j}\boldsymbol{S}_{ij}\|\boldsymbol{B}_i-\boldsymbol{B}_j\|^2$ 仅以加权方式约束二值码以保留原始相似度顺序. 当 $\boldsymbol{S}_{12}>\boldsymbol{S}_{13}$, 推测出 $\boldsymbol{B}_1$ 应该更类似于 $\boldsymbol{B}_2$ 而不是 $\boldsymbol{B}_3$, 而这种相对顺序对每个随机抽样训练批次中的样本组成极为敏感 [31]. 例如, 假设 $\boldsymbol{S}_{12}=0.5$, $\boldsymbol{S}_{13}=0.2$, $\boldsymbol{S}_{23}=0.2$, 则在具有 3 个样本的当前批次中, $\boldsymbol{B}_1$ 应该比 $\boldsymbol{B}_3$ 更类似于 $\boldsymbol{B}_2$. 而由于具体相似度还未完全学习, 并且 $\boldsymbol{S}_{12}$ 是关联于 $\boldsymbol{B}_1$ 和 $\boldsymbol{B}_2$, 因此 $\boldsymbol{B}_1$ 和 $\boldsymbol{B}_2$ 可能会被限制采用过多的相似或者相同的哈希码. $\boldsymbol{S}_{12}$ 虽然是当前批次中相对最大的一个, 但是显然不适合整体相邻结构. 因为 $\boldsymbol{S}_{12}\neq 1$, 过多的相似或者相同的哈希码对于随后哈希码的学习十分不利. 综上所述, 拉普拉斯约束方案不可避免地在训练阶段带来了较高的时间和空间复杂性, 而且学习到的语义信息并不完整.

UMHA 方法使用完整的 n 行 n 列的辅助相似度矩阵来学习整个训练样本的二值码, 以确保算法精度. 所提出的式 (8) 中的损失函数最大程度地匹配了 $\boldsymbol{S}$ 中的特定相似度值而不是它们的相似度顺序, 因此它对每个随机采样训练批次的组成都不敏感, 更适合于按批输入方式, 从而使哈希网络能够实现端到端分批训练. 与拉普拉斯约束方案相比, 式 (8) 中的损失函数不仅显著降低了算法复杂度, 而且由于每批次中辅助相似度矩阵与哈希网络之间的协同学习, 优化了生成的哈希码, 有助于实现更好的编码性能. 优化式 (8) 的主要困难在于对二值码 $\boldsymbol{B}_I$ 和 $\boldsymbol{B}_T$ 施加的离散约束. 哈希网络可以利用式 (5) 生成哈希码. 但是, 在反向传播的过程中, 对于所有非零输入, 符号函数的梯度为零, 这将破坏梯度, 使梯度无法返回到顶层. 也就是说, 由于符号函数 $\mathrm{sgn}(\cdot)$ 不可微分, 存在梯度消失的问题, 因而难以通过反向传播算法更新梯度, 不利于模型的优化. 为了解决这个梯度消失的问题, 采用可缩放的 tanh 函数 [30] 进行处理, 即

$$\boldsymbol{B}=\tanh(\mu\boldsymbol{H})\in\{-1,+1\}^{n\times d},\tag{9}$$

其中 $\mu\in\mathbb{R}^+$. tanh 函数在训练阶段, 将随着 μ 值的增加而替换产生哈希码的 $\mathrm{sgn}(\boldsymbol{H})\in\{-1,+1\}^{n\times d}$ 函数, 其主要观察结果是 $\lim_{\mu\to\infty}\tanh(\mu\boldsymbol{x})=\mathrm{sgn}\,\boldsymbol{x}$. 因此, 收紧 tanh 函数会生成一系列平滑的优化问题. 而随着 μ 值的增加, 将收敛到原始的难处理的二值码问题. 本文在实验部分专门探究了 μ 值的大小, 首先将 μ 值初始化为 1, 然后经过反向传播算法使得 μ 值在 ImgNet 和 TxtNet 中自动学习, 最后在 3 个数据集上 μ 的数值都大约收敛于 28∼29.

3.4.3 最终目标函数

为构建最终的目标函数并获得良好的训练结果, 不仅要考虑不同模态 ($\boldsymbol{B}_I$ 和 $\boldsymbol{B}_T$) 之间的信息, 还

要考虑相同模态 ($\boldsymbol{B}_I$ 和 $\boldsymbol{B}_I$, $\boldsymbol{B}_T$ 和 $\boldsymbol{B}_T$) 内部的信息, 并采用和辅助相似度矩阵协同学习的方式结合到一起成为不同实例的综合信息. 因此, 最终的目标函数设计为

$$\begin{aligned} &\text{loss} = \min_{\boldsymbol{B}_I, \boldsymbol{B}_T} \|\eta \boldsymbol{S} - \cos(\boldsymbol{B}_I, \boldsymbol{B}_T)\|_{\mathrm{F}}^2 + \lambda_1 \|\eta \boldsymbol{S} - \cos(\boldsymbol{B}_I, \boldsymbol{B}_I)\|_{\mathrm{F}}^2 + \lambda_2 \|\eta \boldsymbol{S} - \cos(\boldsymbol{B}_T, \boldsymbol{B}_T)\|_{\mathrm{F}}^2, \\ &\text{s.t } \boldsymbol{S} = \varepsilon(\boldsymbol{S}_I, \boldsymbol{S}_T) \in [-1, +1]^{n \times n}, \quad \boldsymbol{B}_I, \boldsymbol{B}_T \in \{-1, +1\}^{n \times d}, \end{aligned} \tag{10}$$

其中 λ_1 和 λ_2 是权衡参数, 以平衡图像和文本两种不同模态之间和相同模态内部信息的结合. ε 是式 (7) 中的组合函数, 用于整合 $\boldsymbol{S}_I$ 和 $\boldsymbol{S}_T$. 实验部分会给出 η, λ_1 和 λ_2 的具体数值.

需要注意的是, 在测试阶段图像和文本顺序已经打乱, 不能使用辅助相似度矩阵, 而且在训练阶段哈希网络也已训练完毕, 不再需要辅助相似度矩阵进行辅助. 与传统的有监督学习方法不同, 这种方法只要利用好训练阶段输入的图像文本对的顺序, 是一种隐性无监督学习方法, 而不是显性的利用类别标签来表示图像 – 文本对是正样本还是负样本. 最后通过反向传播算法可以更新模型参数.

4 实验结果分析

与其他方法一样, 本文的评估指标使用前 50 个检索结果来计算 mAP (mean average precision), 记为 mAP@50. 本文提出的 UMHA 方法在 Wikipedia[1)], NUS-WIDE[2)]和 MIRFLICKR-25K[3)]这 3 个数据集上获得了优越的性能, 普遍高于一些国际先进的方法. 以使用 128 位哈希码为例, 比较 UMHA 与 UDCMH 方法 [31] 得到的 mAP@50. 在 Wikipedia 数据集上, UMHA 方法基于图像查询文本的检索性能要高出 11.1%, 基于文本查询图像的检索性能则低了 0.6%. 在 NUS-WIDE 数据集上, UMHA 方法基于图像查询文本的检索性能提高了 28.3%, 基于文本查询图像的检索性能提高了 11.2%. 在 MIRFLICKR-25K 数据集上, UMHA 方法基于图像查询文本的检索性能提高了 22.6%, 基于文本查询图像的检索性能提高了 16.6%.

4.1 数据集与参数设置

实验使用 Wikipedia, NUS-WIDE, MIRFLICKR-25K 3 个基准数据集, 全部采用 Wu 等 [31] 的分割方法. Wikipedia 数据集包含 2866 个图像 – 文本对 (每幅图像只有 1 个文本描述), 一共是 10 个类别, 其中训练集有 2173 个图像 – 文本对, 测试集有 693 个图像 – 文本对. NUS-WIDE 数据集有 186557 个图像 – 文本对, 从 81 个类别中选出 10 个最大类别, 本文将 NUS-WIDE 数据集的 1% 作为测试集, 其余作为检索集 (也称数据库). MIRFLICKR-25K 数据集有 24 个类别, 25000 个图像 – 文本对, 本文将 MIRFLICKR-25K 数据集中 2000 个图像 – 文本对作为测试集, 并将其余的作为检索集. NUS-WIDE 和 MIRFLICKR-25K 的训练集分别从其数据库中随机抽取 5000 个实例构建. 各个数据集的划分情况如表 2 所示.

本文实验在 PyTorch1.0.0 上进行, 使用 Python2.7, 占用一块 NVIDIA Tesla P100 显卡 16280 MB 内存中的 13485 MB, 批次设置为 16. 使用随机梯度下降算法进行梯度优化, 其动量项为 0.9, 权重衰减为 0.0005. 经过实验验证, 最后对所有 3 个数据集的超参数取 $\theta = 0.4$, $\eta = 1.5$. 对于 Wikipedia 数据集, 取 $\alpha = 0.3$, $\lambda_1 = \lambda_2 = 0.3$; 对于 NUS-WIDE 数据集, 取 $\alpha = 0.6$, $\lambda_1 = \lambda_2 = 0.1$; 对于 MIRFlickr-25K 数据集, 取 $\alpha = 0.9$, $\lambda_1 = \lambda_2 = 0.1$. 此外, 在 NUS-WIDE 和 MIRFLICKR 数据集上运行时, ImgNet 的学

1) http://www.svcl.ucsd.edu/projects/crossmodal/
2) https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html
3) http://press.liacs.nl/mirflickr/mirdownload.html

表 2 3 个基准数据集的划分 (%)

Table 2 Patition of the three benchmark datasets (%)

Dataset	Training set	Testing set
Wikipedia	2173	693
NUS-WIDE	5000	1866
MIRFLICKR-25K	5000	2000

表 3 在 Wikipedia 数据集上的实验结果 (%)

Table 3 Experimental results on the Wikipedia dataset (%)

Method	Text to image				Image to text			
	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
CVH [14]	25.2	23.5	17.1	15.4	17.9	16.2	15.3	14.9
IMH [15]	46.7	47.8	45.3	45.6	20.1	20.3	20.4	19.5
CMFH [16]	59.5	60.1	61.6	62.2	25.2	25.3	25.9	26.3
LSSH [17]	56.9	59.3	59.3	59.5	19.7	20.8	19.9	19.5
DBRC [30]	57.4	58.8	59.8	59.9	25.3	26.5	26.9	28.8
UDCMH [31]	**62.2**	63.3	64.5	**65.8**	30.9	31.8	32.9	34.6
UMHA (Ours)	61.4	**63.8**	**64.7**	65.2	**43.2**	**45.2**	**45.4**	**45.7**

习率设置为 0.001, TxtNet 的学习率设置为 0.01. 对于包含实例更少的 Wikipedia 数据集, 为 ImgNet 和 TxtNet 都设置 0.01 的学习率. 使用预训练的参数固定 ImgNet 的卷积层, 仅更新完全连接层的参数.

对 ImgNet 使用批量归一化, 可使模型的训练速度提高 5 倍. 本文把训练数据打乱以防止分批训练的时候某一个样本经常被挑选到. 当然每个批次中的样本顺序是固定的, 每幅图像对应相应的文本. 对数据集的预处理采用数据增强的方法, 包括对图片大小进行缩放, 使输入像素统一为 224×224. 图像归一化参数包括均值和标准差, 实验中均值设为 $[0.485, 0.456, 0.406]$, 标准差设为 $[0.229, 0.224, 0.225]$.

4.2 在 Wikipedia 数据集上的实验结果

表 3 展示了在 Wikipedia 数据集上的实验结果, 可以看到在 Wikipedia 数据集上所有的方法效果都不是很理想, 这是由于 Wikipedia 数据集较小而导致 DNN 训练不够充分. 不过, UMHA 方法在大多数情况下的性能优于其他方法, 这也在一定程度上证明 UMHA 方法在小数据集上有良好的表现.

4.3 在 NUS-WIDE 数据集上的实验结果

在 NUS-WIDE 数据集上的实验结果如表 4 所示. 对于大多数方法来说, 并不是哈希码位数越大结果就越好. 由于这些方法只利用了较浅层次的特征完成训练, 所以选择合适的哈希位数对性能至关重要. 而对于 UMHA 方法, 越大的哈希码位数承载的信息越多, 理论上哈希码位数越多实验效果越好. 所以, UMHA 方法采用 128 位哈希码在 3 个数据集上都可以得到最好的结果. 从表 4 中的数据还可以看到, UMHA 方法的性能明显高于其他方法, 这说明 UMHA 方法在较大规模数据集上表现得更为出色.

表 4 在 NUS-WIDE 数据集上的实验结果 (%)

Table 4 Experimental results on the NUS-WIDE dataset (%)

Method	Text to image				Image to text			
	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
CMSSH [13]	51.9	49.8	45.6	48.8	51.2	47.0	47.9	46.6
CVH [14]	47.4	44.5	41.9	39.8	45.8	43.2	41.0	39.2
IMH [15]	47.8	48.3	47.2	46.2	47.0	47.3	47.6	45.9
CMFH [16]	43.9	41.6	37.7	34.9	51.7	55.0	54.7	52.0
LSSH [17]	51.7	61.7	64.2	66.3	48.1	48.9	50.7	50.7
PDH [18]	48.9	51.2	50.7	51.7	47.5	48.4	48.0	49.0
CCQ [19]	49.9	49.6	49.2	48.8	50.4	50.5	50.6	50.5
SCM [20]	51.8	51.0	51.7	51.8	51.7	51.4	51.8	51.8
DBRC [30]	45.5	45.9	46.8	47.3	42.4	45.9	44.7	44.7
UDCMH [31]	63.7	65.3	69.5	71.6	51.1	51.9	52.4	55.8
MGAH [36]	60.3	61.4	64.0	64.1	61.3	62.3	62.8	63.1
UMHA (Ours)	**76.4**	**81.4**	**81.2**	**82.8**	**77.9**	**81.6**	**83.3**	**84.1**

表 5 在 MIRFLICKR-25K 数据集上的实验结果 (%)

Table 5 Experimental results on the MIRFLICKR-25K dataset (%)

Method	Text to image				Image to text			
	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
CMSSH [13]	61.2	60.4	59.2	58.5	61.1	60.2	59.9	59.1
CVH [14]	59.1	58.3	57.6	57.6	60.6	59.9	59.6	59.8
IMH [15]	60.3	59.5	58.9	58.0	61.2	60.1	59.2	57.9
CMFH [16]	64.2	66.2	67.6	68.5	62.1	62.4	62.4	62.7
LSSH [17]	63.7	65.9	65.9	67.2	58.4	59.9	60.2	61.4
PDH [18]	62.7	62.8	62.8	62.9	62.3	62.4	62.1	62.6
CCQ [19]	62.8	62.8	62.2	61.8	63.7	63.9	63.9	63.8
SCM [20]	66.1	66.4	66.8	67.0	63.6	64.0	64.1	64.3
DBRC [30]	61.8	62.6	62.6	62.8	61.7	61.9	62.0	62.1
UDCMH [31]	69.2	70.4	71.8	73.3	68.9	69.8	71.4	71.7
MGAH [36]	67.3	67.6	68.6	69.0	68.5	69.3	70.4	70.2
UMHA (Ours)	**86.4**	**86.4**	**88.2**	**89.9**	**87.7**	**90.6**	**91.3**	**94.3**

4.4 在 MIRFLICKR-25K 数据集上的实验结果

在 MIRFLICKR-25K 数据集上的实验结果如表 5 所示. 从表中数据可见, UMHA 方法在所有性能指标上都显著高于其他先进方法, 充分证明了 UMHA 方法的有效性. 综合 3 个数据集上的实验结果可以看出, 相对于规模较小的数据集, UMHA 方法在大型数据集上的实验性能更优越.

4.5 精确率 – 召回率曲线

图 3 给出了在 MIRFLICKR-25K 数据集上绘制得到的精确率 – 召回率 (precision-recall, P-R) 曲

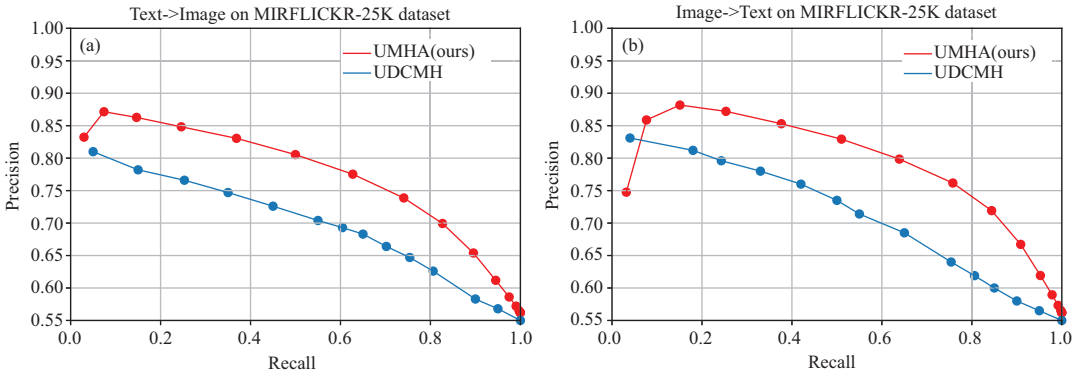

图 3　在 MIRFLICKR-25K 数据集上绘制的 P-R 曲线

Figure 3　P-R curves drawn on MIRFLICKR-25K dataset. (a) Retrieval from text to image; (b) retrieval from image to text

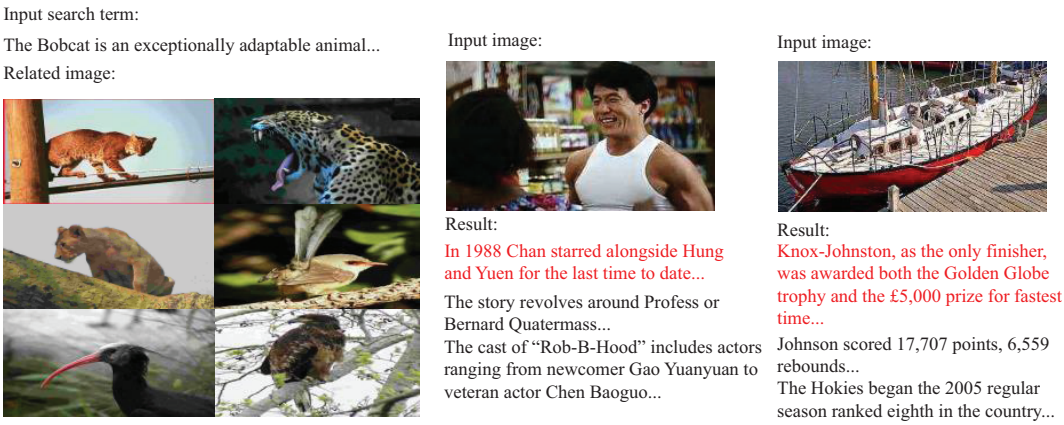

图 4　部分检索结果展示

Figure 4　Presentation of partial retrieval results

线, 记录了 UMHA 和 UDCMH 方法执行文本到图像检索和图像到文本检索两个任务时的精确率和召回率, 其中 UMHA 采用 16 位哈希码. 由图可见, 刚开始模型未训练充分所以精确率并不高, 然后精确率逐渐升高, 最后随着召回率的增长而慢慢下降, 在完全召回时精确率能达到约 55%, 高于 UDCMH 方法.

4.6　检索结果展示

图 4 展示了 Wikipedia 数据集上的部分检索结果, 红色边框的图像和红色文本代表正确的检索结果. 由于 Wikipedia 数据集的文本很长, 所以用 "..." 来表示. 可以看到, UMHA 方法通常经过一两次检索就能找到正确的结果, 充分证明了其有效性. UMHA 方法利用多头注意力网络可以准确地关注到图像中的局部实体, 这与人类的视觉注意系统非常相似. 这种机制不仅可以达到增强语义特征的效果, 也能使 UMHA 方法检索到的事物更加准确和丰富.

从前面的实验结果可以看出, UMHA 在 "基于图像检索文本" 时的性能要高于 "基于文本检索图像". 这个现象可以说明两个问题: 一方面, UMHA 对图像的细粒度语义识别性能还不够好. 正如图 4 中所示, 系统能够检索出与 "animal" 相关的图像, 但不能有效地检索出多幅与 "Bobcat" 相关的图像,

表 6 在 MIRFlickr-25K 数据集上的消融实验结果 (%)

Table 6 Ablation experimental results on the MIRFLICKR-25K dataset (%)

Method	Text to image	Image to text
	128 bits	128 bits
UMHA-1	83.1	86.4
UMHA-2	86.0	88.6
UMHA-3	86.2	91.3
UMHA-4	87.6	91.5
UMHA-5	87.9	92.0
UMHA	**89.9**	**94.3**

而文本中经常包含这种对应细粒度语义的单词. 另一方面, UMHA 对文本的语义分析不如对图像的语义分析那么准确, 利用 BOW 模型和多头注意力网络将文本转化为语义向量的过程中相对更容易丢失信息. 这会导致一部分复杂文本的语义表示不够完整, 从而降低基于文本检索图像的性能.

4.7 消融实验

消融实验采用 128 位哈希码在 MIRFLICKR-25K 数据集上进行, 结果如表 6 所示.

表中 UMHA-1 方法没有使用辅助相似度矩阵, 结果表明辅助相似度矩阵至关重要, 能与哈希码矩阵进行协同学习. UMHA-2 方法仅使用式 (8) 中的 loss1 作为损失函数, 结果证明式 (8) 仅仅使用了部分语义信息, 没有式 (10) 整合的语义信息全面. UMHA-3 方法没有使用多头注意力网络, 结果表明多头注意力网络能增强模型对语义信息的理解, 对性能提升有很大帮助. UMHA-4 方法使用 $\tanh(\boldsymbol{x})$ 而没有使用 $\tanh(\mu\boldsymbol{x})$ 产生哈希码, 结果表明使用自适应学习的 $\tanh(\mu\boldsymbol{x})$ 激活函数能有效提升性能. UMHA-5 方法没有在卷积层后面统一进行批量归一化, 结果表明批量归一化能很好地适应本文方法, 具有重要作用. 最后 UMHA 方法包含了所有组件, 实验结果证明这些组件结合起来可以获得最好的效果.

5 结论

本文提出了基于多头注意力网络的无监督跨媒体哈希检索方法 UMHA, 采用无监督方法构建检索模型, 具备实际应用于大规模跨媒体检索的潜力. UMHA 利用深浅层并联的思想, 基于更深层的哈希网络构建哈希矩阵, 利用相对较浅层的图像和文本特征网络构建辅助相似度矩阵, 并利用两个矩阵进行协同学习, 获取大量图像和文本的语义信息, 使 UMHA 能够捕获不同模态之间和相同模态内部的潜在联系, 从而缩减异构鸿沟. 在 3 个基准数据集上的大量实验证明了 UMHA 方法的优越性和鲁棒性, 并通过消融实验研究了 UMHA 方法各个组件的有效性. 在未来的工作中, 会考虑将半监督学习方法与生成对抗网络融入跨媒体检索的框架, 争取利用少量的类别标签获得更好的检索性能.

参考文献

1 Peng Y X, Huang X, Zhao Y Z. An overview of cross-media retrieval: concepts, methodologies, benchmarks, and challenges. IEEE Trans Circ Syst Video Technol, 2018, 28: 2372–2385

2 Feng F X, Wang X J, Li R F. Cross-modal retrieval with correspondence autoencoder. In: Proceedings of the 22nd ACM International Conference on Multimedia, 2014. 7–16

3 Huang Y, Wu Q, Song C, et al. Learning semantic concepts and order for image and sentence matching. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 6163–6171
4 Zheng Z D, Zheng L, Garrett M, et al. Dual-path convolutional image-text embedding with instance loss. ACM Trans Multimedia Comput Commun Appl, 2020, 16: 51
5 Chen Z, Du H, Chen Y F, et al. Cross-modal video moment retrieval based on visual-textual relationship alignment. Sci Sin Inform, 2020, 50: 862–876 [陈卓, 杜昊, 吴雨菲, 等. 基于视觉 – 文本关系对齐的跨模态视频片段检索. 中国科学: 信息科学, 2020, 50: 862–876]
6 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2012. 1097–1105
7 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556
8 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2017. 5998–6008
9 Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. 2015. ArXiv:1502.03167
10 Lai H J, Pan Y, Liu Y, et al. Simultaneous feature learning and Hash coding with deep neural networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3270–3278
11 Li W-J, Wang S, Kang W-C. Feature learning based deep supervised Hashing with pairwise labels. 2015. ArXiv:1511.03855
12 Li Z X, Wei H Y, Huang F C, et al. Combine visual features and scene semantics for image captioning. Chinese J Comput, 2020, 43: 1624–1640 [李志欣, 魏海洋, 黄飞成, 等. 结合视觉特征和场景语义的图像描述生成. 计算机学报, 2020, 43: 1624–1640]
13 Bronstein M M, Bronstein A M, Michel F, et al. Data fusion through cross-modality metric learning using similarity-sensitive Hashing. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010. 3594–3601
14 Kumar S, Udupa R. Learning Hash functions for cross-view similarity search. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, 2011. 1360–1365
15 Song J K, Yang Y, Yang Y, et al. Inter-media Hashing for large-scale retrieval from heterogeneous data sources. In: Proceedings of ACM SIGMOD International Conference on Management of Data, 2013. 785–796
16 Ding G G, Guo Y C, Zhou J. Collective matrix factorization Hashing for multimodal data. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2014. 2083–2090
17 Zhou J, Ding G G, Guo Y C. Latent semantic sparse Hashing for cross-modal similarity search. In: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2014. 415–424
18 Rastegari M, Choi J, Fakhraei S, et al. Predictable dual-view Hashing. In: Proceedings of International Conference on Machine Learning, 2013. 1328–1336
19 Long M S, Cao Y, Wang J M, et al. Composite correlation quantization for efficient multimodal retrieval. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2016. 579–588
20 Zhang D, Li W-J. Large-scale supervised multimodal Hashing with semantic correlation maximization. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014. 2177–2183
21 Wu B, Yang Q, Zheng W-S, et al. Quantized correlation Hashing for fast cross-modal search. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence, 2015. 3946–3952
22 Lin Z J, Ding G G, Hu M Q, et al. Semantics-preserving Hashing for cross-view retrieval. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3864–3872
23 Cao Y, Long M S, Wang J M, et al. Deep visual-semantic Hashing for cross-modal retrieval. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. 1445–1454
24 Jiang Q-Y, Li W-J. Deep cross-modal Hashing. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 3232–3240
25 Liong V E, Lu J, Tan Y-P, et al. Cross-modal deep variational Hashing. In: Proceedings of IEEE International Conference on Computer Vision, 2017. 4077–4085
26 Shen Y M, Liu L, Shao L, et al. Deep binaries: encoding semantic-rich cues for efficient textual-visual cross retrieval. In: Proceedings of IEEE International Conference on Computer Vision, 2017. 4097–4106
27 Cao Y, Liu B, Long M S, et al. Cross-modal hamming Hashing. In: Proceedings of European Conference on Computer Vision, 2018. 202–218
28 Wang W, Ooi B C, Yang X Y, et al. Effective multi-modal retrieval based on stacked auto-encoders. Proc VLDB Endow, 2014, 7: 649–660
29 Cao Y, Long M S, Wang J M, et al. Correlation autoencoder Hashing for supervised cross-modal search. In: Proceedings of ACM International Conference on Multimedia Retrieval, 2016. 197–204

30 Hu D, Nie F P, Li X L. Deep binary reconstruction for cross-modal Hashing. IEEE Trans Multimedia, 2019, 21: 973–985
31 Wu G S, Lin Z J, Han J G, et al. Unsupervised deep Hashing via binary latent factor models for large-scale cross-modal retrieval. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018. 2854–2860
32 Wang D X, Cui P, Ou M D, et al. Deep multimodal Hashing with orthogonal regularization. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence, 2015. 2291–2297
33 Su S P, Zhong Z S, Zhang C. Deep joint-semantics reconstructing Hashing for large-scale unsupervised cross-modal retrieval. In: Proceedings of IEEE International Conference on Computer Vision, 2019. 3027–3035
34 Li C, Deng C, Li N, et al. Self-supervised adversarial Hashing networks for cross-modal retrieval. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 4242–4251
35 Zhang X, Lai H J, Feng J S. Attention-aware deep adversarial Hashing for cross-modal retrieval. In: Proceedings of European Conference on Computer Vision, 2018. 614–629
36 Zhang J, Peng Y X. Multi-pathway generative adversarial Hashing for unsupervised cross-modal retrieval. IEEE Trans Multimedia, 2020, 22: 174–187
37 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2014. 2672–2680
38 Wang J, Yu L T, Zhang W N, et al. IRGAN: a minimax game for unifying generative and discriminative information retrieval models. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017. 515–524

Unsupervised cross-media Hashing retrieval based on multi-head attention network

Zhixin LI[1*], Feng LING[1], Zhenjun TANG[1], Huifang MA[2] & Zhiping SHI[3]

1. *Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin* 541004*, China*;
2. *College of Computer Science and Engineering, Northwest Normal University, Lanzhou* 730070*, China*;
3. *College of Information Engineering and Academy for Multidisciplinary Studies, Capital Normal University, Beijing* 100048*, China*
* Corresponding author. E-mail: lizx@gxnu.edu.cn

Abstract The cross-media Hash retrieval encodes different media data into a common binary Hash space, which can effectively measure the correlation between different modal samples. In order to further improve the retrieval performance, this paper proposes an unsupervised cross-media Hash retrieval method based on multi-head attention network. First, we use a multi-head attention network to generate a Hash code matrix, which makes the images and texts match better. Second, an auxiliary similarity matrix is constructed to integrate the original neighborhood information from different modalities. Through the collaborative learning of auxiliary similarity matrix and Hash code matrix, our method can capture the potential correlations between different modalities and within the same modality. In addition, we design two loss functions to train the model, and adopt strategies of batch normalization and replacing Hash code generation functions to optimize the model, which greatly improves the training speed of the model. Experiments on three datasets show that the average performance of our method is significantly higher than many state-of-the-art unsupervised methods, which fully proves the effectiveness and superiority of our method.

Keywords convolutional neural network, multi-head attention network, cross-media Hashing retrieval, unsupervised learning, collaborative learning, auxiliary similarity matrix, batch normalization

Zhixin LI was born in 1971. He is a professor and Ph.D. supervisor of the College of Computer Science and Information Engineering, Guangxi Normal University. His research interests include image understanding, machine learning, cross-media computing, and natural language processing.

Feng LING was born in 1993. He received his M.S. degree from computer science and information engineering, Guangxi Normal University in 2019. His research interests include machine learning and cross-media retrieval.

Zhenjun TANG was born in 1979. He is a professor and Ph.D. supervisor of the College of Computer Science and Information Engineering, Guangxi Normal University. His research interests include multimedia Hashing and image processing.

Huifang MA was born in 1981. She is a professor and M.S. supervisor of the College of Computer Science and Engineering, Northwest Normal University. Her research interests include data mining and natural language processing.