doi:10.11887/j.cn.202203011 http://journal.nudt.edu.cn

上下文感知的深度弱监督图像哈希表示学习方法*

刘 萌,周 迪,田传发,齐孟津,聂秀山
(山东建筑大学 计算机科学与技术学院,山东 济南 250101)

摘 要:针对现有深度监督图像哈希表示学习方法依赖于图像的类别信息,难以在现实中被广泛应用问题,利用与图像相关的标签信息作为监督信息,提出上下文感知的深度弱监督图像哈希表示学习方法。该方法一方面通过自适应捕获图像区域特征的相关上下文来增强它们的表示能力,另一方面通过引入判别损失来提高学习到的哈希码表示的判别性。在现有两个公开数据集上的大量实验结果证明了该方法的有效性。

关键词:图像哈希;弱监督学习;图像检索;区域上下文建模;判别损失
中图分类号:TP311 **文献标志码**:A **开放科学(资源服务)标识码(OSID)**:
文章编号:1001-2486(2022)03-085-08

听语音 聊科研 与作者互动

Context-aware deep weakly supervised image hashing learning method

LIU Meng, ZHOU Di, TIAN Chuanfa, QI Mengjin, NIE Xiushan
(School of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, China)

Abstract: Existing deep supervised image hashing approaches rely on substantial labeled image data, which is very difficult to be widely applied in reality. By utilizing tags associated with images as the supervision information, a context-aware deep weakly supervised image hashing method was proposed. The method enhanced the image region representations by adaptively capturing the relevant context information of image region features, and raised the discrimination of the learnt hash codes by introducing a discrimination loss. Extensive experiments on two public datasets show the effectiveness of the method.

Keywords: image hashing; weakly supervised learning; image retrieval; region context modeling; discrimination loss

随着社交网络和移动智能手机的快速发展,大量的图片被网民记录和分享。为了规避海量图像所带来的巨大存储成本,同时满足高效的图像检索需求,图像哈希表示学习方法引起了越来越多的研究兴趣[1]。早期的图像哈希表示学习方法大多采用手工设计的局部特征,因此这些方法的性能很大程度上取决于它们使用的特征或它们设计的特征提取方法。近年来,随着深度神经网络在图像表示中的发展[2],深度图像哈希表示学习方法得到了广泛的研究,其有效地将深度卷积神经网络的优势与哈希技术的低计算成本和存储能力相结合,例如非对称深度监督哈希方法[3]、深度锚图哈希方法[4]、深度增量哈希网络[5]、半监督自步对抗哈希方法[6]、基于局部归一化指数函数损失的深度哈希方法[7]以及自适应局部多视图哈希方法[8]。关于深度图像哈希表示学习方法的详细介绍,可参见文献[9]。

尽管现有深度图像哈希表示学习方法取得了令人瞩目的进步,但它们中的大多数都是有监督的学习方法,十分依赖大量类别标注信息,而这对于现实世界的应用而言是一件成本很高的事情。社交网络中的图像通常具有用户提供的标签信息,而这些标签信息在一定程度上可以描述图像的语义信息。例如,图1中的第一个图像,它的标签信息“trees”“sky”“clouds”和“blue”都描绘了图像的内容。与此同时,它的标签信息“sky”和“clouds”也是该图像的类别信息。更重要的是,与图像类别相比,图像的标签信息更容易获得。鉴于此,研究弱监督图像哈希表示学习方法——利用图像的标签作为监督信息,而不是图像类别信息来学习哈希函数,是十分必要的。

然而,通过图像的标签信息学习图像哈希表

* **收稿日期**:2021-06-07
基金项目:国家自然科学基金资助项目(62006142)
作者简介:刘萌(1991—),女,黑龙江哈尔滨人,教授,博士,硕士生导师,E-mail:mengliu.sdu@gmail.com;
聂秀山(通信作者),男,教授,博士,硕士生导师,E-mail:niexiushan19@sdjzu.edu.cn

图像			
类别	sky, clouds, grass	plants, flowers	sky
标签	city, blue, trees, light, sea, sky, orange, white, green, colors, clouds, digital, wonderful, landscape, star, amazing, nice, long, heaven, day, cityscape, gulf, wide, vivid, super, fisheye, arab, shining, ultra	flowers, white, beautiful, stars, lights, lily	church, architecture, steeple, spire

图 1 NUS-WIDE 数据集中部分图像展示

Fig. 1 Illustrating some image samples from the NUS-WIDE dataset

示并非易事,具体原因如下:①在社交媒体平台上,用户提供的标签信息可能与图像类别不直接相关。例如,图 1 中第三个图像的标签是“church”“architecture”“steeple”和“spire”,但是它的类别是“sky”。因此,如何从标签中挖掘有效的监督信息成为一个关键问题。②为增强图像与相应的语义标签之间的匹配,图像细粒度语义信息起着至关重要的作用。如何全面理解图像内容并捕获有益语义信息为一个亟待解决的问题。

虽然在监督图像哈希表示学习方法方面取得了一些成功,但弱监督图像哈希表示学习任务仍是一个尚未解决的问题。据知,文献[10]首次研究了基于深度学习的弱监督哈希表示学习方法,它提出了利用标签嵌入或二进制标签向量的深度弱监督哈希表示学习模型。尽管取得了良好的性能,但是该模型有几个关键的缺点:①其假设若两幅图像具有至少一个相同的文本标签,则它们是相似的,并基于此假设设计了二元标签向量模型。但是,这个假设是不恰当的。如图 1 所示,第一张图像和第二张图像共享红色的标签信息“white”,但它们的类别完全不同。②其将图像进行整体编码,而这样一个紧凑的全局表示很难捕获图像中的细粒度语义细节。例如,若想充分理解图 1 中第一张图像的内容,需对其涉及的属性信息(如“blue”)和实体信息(如“trees”)进行理解。为了解决上述问题,本文提出了一种上下文感知的深度弱监督图像哈希表示(context-aware deep weakly supervised image hashing,IDEA)学习方法。具体来说,本文设计了一种新的图像编码器,它可以自适应地捕捉有意义的图像区域上下文信息来增强图像表示。此外,本文引入判别损失来加强图像与标签之间的对齐,继而提升哈希码的表示能力。

1 相关工作

1.1 有监督的图像哈希表示学习

目前,在有监督的图像哈希表示学习方面,有许多研究成果。例如:文献[11]将哈希表示学习表述为一个多分类任务,并通过最大化原始空间和汉明空间分类顺序一致性来学习哈希函数;文献[12]提出了一种两阶段的监督哈希表示学习方法用于图像检索;文献[13]提出了一种深度监督哈希表示学习方法;由于现有的松弛方法对松弛的误差界没有理论保证,文献[14]证明了当损失函数为 Lipschitz 连续时,二进制优化问题可以松弛为有界约束的连续优化问题,并提出了一种二进制优化哈希学习方法;此外,为了提升哈希码的判别性,文献[15]提出了一种判别式深度哈希学习框架,该框架集成了特征提取、哈希学习和类别预测;针对人脸图像检索任务,文献[16]提出了一种基于分类和量化误差的深度哈希算法。

由于现有的深度监督哈希方法大多采用对称策略来学习深度哈希函数,其训练通常耗时较长,难以适应于大规模数据场景。鉴于此,文献[3]提出了一种非对称深度监督哈希方法用于大规模最近邻搜索,即它以非对称的方式处理查询点和数据库点。为充分利用可用的有标记信息,文献[4]提出了一种深度锚图哈希框架。文献[5]提出了一种深度增量哈希网络,其以增量方式学习哈希码。为了学习能有效地保持图像标签信息的鉴别哈希码,文献[7]提出了基于局部归一化指数函数损失的深度哈希方法。

1.2 无监督图像哈希表示学习

现有的无监督图像哈希表示学习方法大致可以分为两类:基于浅层学习的方法和基于深度学习的方法。作为浅层学习的代表,文献[17]中提出了一种谱哈希方法。但是,此方法是基于主投影来构造哈希函数,因此生成的哈希码不是非常准确且效率不高。为了解决这个问题,文献[18]提出一种半监督哈希表示学习框架;文献[19]提出了一种简单而有效的交替最小化算法,通过寻找零中心数据的旋转来学习哈希码。随着深度学习技术的发展,一些基于深度学习的无监督图像哈希表示学习算法被提出。其中,文献[20]提出一种深度哈希表示学习方法,该方法利用 GIST 特征作为神经网络的输入;文献[21]旨在学习旋转不变的哈希码。文献[8]提出了一种无监督自适应局部多

视图哈希方法,用来处理部分视图哈希问题,以实现高效的社交图像检索。

1.3　弱监督图像哈希表示学习

近年来,有监督和无监督的图像哈希表示学习已经取得了很大的进展,而在弱监督图像哈希表示学习方面却鲜有尝试。弱监督图像哈希表示学习的目的是在训练时仅仅利用图像的标签信息,而不使用图像的类别信息。文献[22]中提出一种弱监督多模态哈希学习方法,但它依赖于手工设计的特征,如 GIST 特征、颜色直方图和尺度不变特征变换(scale-invariant feature transform, SIFT),限制了其性能。与此不同的是,文献[23]设计了一个两阶段弱监督深度哈希框架,包括弱监督预训练和监督微调。虽然在跨模态哈希领域中一些工作试图利用标签信息和图像来学习哈希空间[24-25],但它们的目标是为不同模态的输入(图像和标签)学习一个通用的哈希空间。这与弱监督图像哈希表示学习的目标(即为图像学习一个哈希空间)完全不同。关于跨模态哈希和单模态哈希学习区别的详细讨论可以参考文献[26]。

2　符号与问题定义

2.1　符号定义

本文分别使用大写的黑斜体字母(如 $\boldsymbol{X}$)和小写的黑斜体字母(如 $\boldsymbol{b}$)表示矩阵和向量,分别使用白斜体字母(如 N)、花体(如 $\mathscr{T}$)和希腊字母(如 λ)表示标量、集合、参数或函数,$\boldsymbol{W}_{i,j}$表示矩阵 $\boldsymbol{W}$ 第 i 行第 j 列的元素。sgn(·)为基于元素的符号函数,其对于正数输出“+1”,对于负数输出“-1”。如果没有明确说明,所有向量都是列向量。

2.2　问题定义

假设有 N 个给定样本 $\{\boldsymbol{I}_i, \mathscr{T}_i, \boldsymbol{L}_i\}_{i=1}^{N}$,其中 $\boldsymbol{I}_i$ 表示第 i 张图像,$\mathscr{T}_i$ 表示第 i 张图像的标签集合,$\boldsymbol{L}_i \in \{0,1\}^K$ 表示二值标签向量,K 表示类别数目。本文的目标是以 $\{\boldsymbol{I}_i, \mathscr{T}_i\}_{i=1}^{N}$ 为输入,学习一个哈希函数为每张图像输出一个哈希表示向量 $\boldsymbol{b}_i$。换句话说,图像的类别信息在训练阶段是不可以用的,它们只在测试阶段被使用。更重要的是,本文学习到的哈希函数需要确保语义相似的图像具有相似的哈希码表示。

3　学习方法

如图 2 所示,本文提出的学习方法主要包含三个部分:①图像编码器生成上下文强化的视觉表示;②标签编码器输出标签嵌入;③损失函数。

3.1　图像编码器

图像编码器包括区域特征表示和上下文强化的视觉表示两部分。

1)区域特征表示:采用预训练的 ResNet-50 网络作为主干网络(如图 2 所示),该网络以 224×224×3 大小的图像作为输入,通过 Conv1、Conv2_x、Conv3_x、Conv4_x、Conv5_x 五个模块后,输出大小为 7×7×2 048 的特征映射,上述过程可总结为如下公式:

$$\boldsymbol{X}_i = \theta(\boldsymbol{I}_i) \tag{1}$$

其中,θ 表示去除最后全连接层的 ResNet-50 网络,$\boldsymbol{X}_i \in \mathbb{R}^{7\times7\times2\,048}$ 表示第 i 张图像的区域特征表示。

为得到图像表示,一种直接的方式是使用平均池化方法,但这可能会引入噪声信息,即无意义的区域表示信息。为了解决这个问题,可以采用注意机制为特征映射中不同区域学习注意力分数,然后通过自适应地聚合区域表示,得到图像表示。然而,上述注意机制完全忽略了视觉区域上下文信息在判断区域起着至关重要的作用。

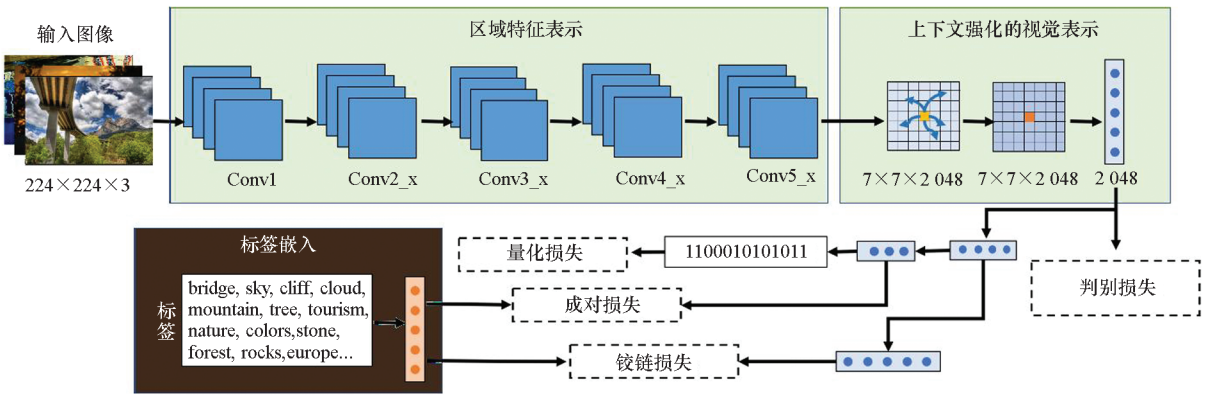


图 2　本文提出的 IDEA 方法流程

Fig. 2　Framework of the proposed model IDEA

2)上下文强化的视觉表示:为了更好地利用上下文区域,首先采用自注意力机制来捕获每个区域的上下文信息,具体过程如下:

$$\begin{cases} \boldsymbol{A}^{i} = \hat{\boldsymbol{X}}_i^{\mathrm{T}} \hat{\boldsymbol{X}}_i \\ \boldsymbol{A}_{i,j}^{(i)} = \dfrac{\exp(\boldsymbol{A}_{i,j}^{(i)})}{\sum_{k=1}^{49} \exp(\boldsymbol{A}_{i,j}^{(i)})} \\ \tilde{\boldsymbol{X}}_i = \boldsymbol{A}^{(i)} \hat{\boldsymbol{X}}_i \end{cases} \tag{2}$$

其中,$\hat{\boldsymbol{X}}_i \in \mathbb{R}^{49\times 2\,048}$ 为 $\boldsymbol{X}_i$ 变形后的结果,$\boldsymbol{A}^{(i)} \in \mathbb{R}^{49\times 49}$ 为区域注意力分数矩阵,$\tilde{\boldsymbol{X}}_i \in \mathbb{R}^{49\times 2\,048}$ 为上下文信息强化后的区域特征表示矩阵。

由于上下文信息强化后的区域特征表示与原始区域特征表示之间是互补关系,为充分表示图像的视觉区域,本文将上下文感知的区域特征与原始区域特征矩阵相加。由于每一个图像区域均可能蕴含有效的语义信息,故本文采用平均池化操作聚合局部的区域特征表示,使得到的全局图像表示尽可能保留全部有效语义信息。上述过程表示为:

$$\begin{cases} \bar{\boldsymbol{X}}_i = \hat{\boldsymbol{X}}_i + \tilde{\boldsymbol{X}}_i \\ \boldsymbol{x}_i = avg_pooling(\bar{\boldsymbol{X}}_i) \end{cases} \tag{3}$$

其中,$\boldsymbol{x}_i \in \mathbb{R}^{2\,048}$ 表示强化后的视觉表示向量。为了得到图像的哈希表示,本文将上下文增强的图像表示输入一个多层感知器中,具体如下:

$$\begin{cases} \tilde{\boldsymbol{h}}_i = \theta_1(\boldsymbol{W}_1 \boldsymbol{x}_i + \boldsymbol{b}_1) \\ \boldsymbol{h}_i^1 = \theta_2(\boldsymbol{W}_2 \tilde{\boldsymbol{h}}_i + \boldsymbol{b}_2) \end{cases} \tag{4}$$

其中:θ_1 和 θ_2 分别表示 ReLU 和 Sigmoid 激活函数;$\boldsymbol{W}_1$、$\boldsymbol{W}_2$、$\boldsymbol{b}_1$ 和 $\boldsymbol{b}_2$ 为参数矩阵和向量,$\boldsymbol{h}_i^1$ 为第 i 张图像长度为 l 的哈希码表示向量。

3.2 标签编码器

假设第 i 张图像的标签集合 $\mathcal{T}_i$ 中包含 M 个语义标签,其中第 j 个标签表示为 T_i^j。本文使用 word2vec 模型来提取每一个标签的词嵌入表示向量,表示为 $\boldsymbol{t}_i^j \in \mathbb{R}^d, j \in \{1,2,\cdots,M\}$。由于每一张图像对应不止一个文本标签,故本文采用平均池化得到整体的标签表示向量,具体操作如下:

$$\begin{cases} \boldsymbol{V}_i = [\boldsymbol{t}_i^1, \boldsymbol{t}_i^2, \cdots, \boldsymbol{t}_i^M] \\ \boldsymbol{t}_i = avg_pooling(\boldsymbol{V}_i) \end{cases} \tag{5}$$

其中,$\boldsymbol{t}_i \in \mathbb{R}^d$ 为最终文本标签表示向量。

3.3 损失函数

本文 IDEA 方法的优化目标函数为:

$$L = \lambda_1 L_1 + L_2 + \lambda_3 L_3 + \lambda_4 L_4 \tag{6}$$

其中,λ_1、λ_3 和 λ_4 为平衡参数。

具体地,L_1 为量化损失函数,旨在使输出的哈希表示向量 $\boldsymbol{h}_i^1$ 的元素值接近 0 或 1,它的表示形式如下:

$$L_1 = -\sum_{i=1}^{N} \frac{1}{L} (\boldsymbol{h}_i^1 - 0.5\boldsymbol{I})^{\mathrm{T}} (\boldsymbol{h}_i^1 - 0.5\boldsymbol{I}) \tag{7}$$

其中,$\boldsymbol{I}$ 表示元素值全为 1 且长度为 l 的向量。

L_2 为成对损失函数,用于约束具有相似标签表示的图像拥有相似的哈希码表示,其具体表示如下:

$$L_2 = \sum_{i=1}^{N}\sum_{j=1}^{N} \left[\frac{1}{L} (\boldsymbol{h}_i^1 - \boldsymbol{h}_j^1)^{\mathrm{T}} (\boldsymbol{h}_i^1 - \boldsymbol{h}_j^1) - \frac{1}{2}\left(1 - \frac{\boldsymbol{t}_i^{\mathrm{T}} \boldsymbol{t}_j}{\|\boldsymbol{t}_i\| \|\boldsymbol{t}_j\|}\right) \right]^2 \tag{8}$$

其中,$\boldsymbol{h}_i^1$ 和 $\boldsymbol{h}_j^1$ 分别为图像 I_i 和 I_j 的哈希码表示。

L_3 为铰链损失,旨在消除模态语义鸿沟,其具体形式如下:

$$L_3 = \sum_{i=1}^{N} \sum_{j\neq i} \max[0, m + \boldsymbol{t}_j^{\mathrm{T}} \boldsymbol{h}_i^2 - \boldsymbol{t}_i^{\mathrm{T}} \boldsymbol{h}_i^2] \tag{9}$$

$$\boldsymbol{h}_i^2 = \theta_3(\boldsymbol{W}_3 \tilde{\boldsymbol{h}}_i + \boldsymbol{b}_3) \tag{10}$$

其中,θ_3 为 Tanh 激活函数,$\boldsymbol{W}_3$ 和 $\boldsymbol{b}_3$ 为参数矩阵和向量,$\boldsymbol{h}_i^2$ 为与标签表示同维度的图像嵌入表示,m 为预定义参数。

L_4 为判别损失,通过促进图像表示准确地生成相应的标签信息,来增强图像表示的判别性,公式形式如下:

$$L_4 = \frac{1}{N} \sum_{i=1}^{N} -\sum_{c=1}^{L_c} y_{ic} \log(p_{ic}) \tag{11}$$

$$p_{ic} = \theta_4(\boldsymbol{W}_4 \boldsymbol{x}_i + \boldsymbol{b}_4) \tag{12}$$

其中:L_c 为文本标签类别数目;p_{ic} 为预测得到的第 i 张图像包含第 c 个文本标签的概率;$\boldsymbol{W}_4$ 和 $\boldsymbol{b}_4$ 为参数矩阵和向量;θ_4 为 Softmax 函数,用来归一化预测结果;如果图像 I_i 包含第 c 个文本标签,则 y_{ic} 值为 1,否则为 0。

在测试阶段,本文首先利用式(4)对测试图像提取哈希表示向量 $\boldsymbol{h}^1$;然后,对其进行如下量化得到哈希码表示:

$$\boldsymbol{b} = \frac{1}{2}[\mathrm{sgn}(\boldsymbol{h}^1 - 0.5\boldsymbol{I}) + 1] \tag{13}$$

4 实验与结果

4.1 数据集

在两个广泛使用的公开图像数据集,即 MIR-FLICKR25K[27] 和 NUS-WIDE[28] 上进行大量的实验。其中,NUS-WIDE 数据集是从 Flickr 上收集的大规模社交图像数据集,它包含了 269 648 张图像和 5 018 个文本标签信息,这些图像被手工

标记为 81 个类别[29]。与文献[1]类似,本文只考虑了最高频的 21 个图像类别,得到 194 541 张图像。本文从中随机选择 120 000 张图像,其中 100 000张图像作为训练集,其余的为测试集。MIR-FLICKR25K 是一个相对较小的数据集,共有 25 000张图像和 1 386 个用户提供的标签。类似地,这些图像被手工标记为 38 个类别。本文仅保留至少包含一个文本标签的图像,过滤后共获得了 20 015 张图像,从中随机选取 16 000 张图片用于训练,2 000 张图片用于测试。

训练阶段:将训练数据集中的图像信息以及相应的文本标签信息输入 IDEA 网络中,进行参数学习。

测试阶段:仅需将测试图像输入训练好的 IDEA 网络中,得到的哈希向量表示 $\boldsymbol{h}^1$ 经过式(13)的量化操作,即可得到二值的图像哈希表示,用于下游的检索任务。

4.2　实验设置

1)评价指标:为了衡量本文方法和基线方法的性能,本文采用全类平均精度(mean average precision, mAP)作为评价指标。

2)实验细节设置:在开源深度学习库 Keras 上使用 Tensorflow 作为后端实现本文的方法,并采用动量设置为 0.9 的随机梯度下降算法作为优化器。Conv1 ~ Conv5_x 的学习率设置为 0.001,其他层设置为 0.01。批处理大小被设置为 50,两个全连接层的大小分别为 256 和 300。目标函数中有 3 个平衡参数,即 λ_1、λ_3 和 λ_4,本文采用网格搜索策略仔细调节并选择最优参数。具体地,首先使用自适应步长在[0,1 000]范围内执行粗粒度的网格搜索。之后,在每个参数的近似范围内,使用较小的步长在较小的范围内进行微调。最终,本文的 3 个平衡参数分别设置为 1.0、1.0 和 0.01。word2vec 模型是在 Wikipedia documents 上预先训练,输出维度为 300 的向量。

3)基线方法:与几种最先进的无监督和弱监督图像哈希方法进行比较,包括 SH[17]、PCAH[18]、ITQ[19]、DH[20]、DeepBit[21]、LSH[30]、SpH[31]、DSH[32]、AGH[33]、UH-BDNN[34]、WDHT-BTV[10]和 WDHT[10]。

4.3　结果对比分析

表 1 和表 2 显示了本文方法 IDEA 与基线方法在两个数据集上的性能比较。通过分析两个表格中的结果,可以得到以下发现:①在非深度学习方法中,随着哈希码的长度增加,虽然 LSH 的表现越来越好,但它的性能是最差的,这主要是因为它忽略了数据的分布信息;除了 AGH 之外,ITQ 比其他基线方法表现得都好,这充分表明了数据分布的重要性;而 AGH 的表现优于 ITQ 且超过了其他非深度学习方法,这是因为 ITQ 忽略了局部邻域关系的重要性。②基于深度学习方法,如 DH、UH-BDNN 和 DeepBit 的性能不如非深度学习方法,这是因为 DH 和 UH-BDNN 过度依赖手工

表 1　在 5 000 返回结果上的 mAP 性能比较

Tab. 1　Performance comparison in terms of mAP scores over 5 000 retrieved images

模型	NUS-WIDE				MIR-FLICKR25K			
	12 bit	24 bit	32 bit	48 bit	12 bit	24 bit	32 bit	48 bit
ITQ(非深度)	0.632 9	0.629 9	0.594 0	0.647 8	0.690 8	0.706 4	0.668 4	0.699 6
PCAH(非深度)	0.576 6	0.504 6	0.490 0	0.490 4	0.643 0	0.630 6	0.637 2	0.651 6
LSH(非深度)	0.350 1	0.409 3	0.416 9	0.454 6	0.573 6	0.604 9	0.595 4	0.623 9
DSH(非深度)	0.591 9	0.598 2	0.571 3	0.579 1	0.695 5	0.707 1	0.683 4	0.660 3
SpH(非深度)	0.464 5	0.464 5	0.446 5	0.447 2	0.596 6	0.581 1	0.582 8	0.579 0
SH(非深度)	0.562 3	0.503 3	0.489 6	0.453 3	0.660 5	0.640 5	0.629 1	0.621 3
AGH(非深度)	0.655 1	0.645 9	0.627 4	0.622 5	0.686 2	0.700 5	0.699 8	0.685 3
DH(深度)	0.473 3	0.460 1	0.462 0	0.476 3	0.603 3	0.619 5	0.613 5	0.618 0
UH-BDNN(深度)	0.592 3	0.591 5	0.590 2	0.609 7	0.665 4	0.668 4	0.667 2	0.669 9
DeepBit(深度)	0.546 3	0.554 8	0.562 4	0.561 0	0.589 0	0.602 7	0.609 0	0.608 6
WDHT-BTV(深度)	0.620 2	0.627 0	0.624 7	0.624 9	0.636 5	0.632 6	0.637 3	0.635 2
WDHT(深度)	0.670 9	0.680 5	0.695 5	0.676 0	0.734 6	0.743 0	0.703 4	0.705 4
IDEA(深度)	**0.705 2**	**0.741 1**	**0.750 7**	**0.759 0**	**0.777 4**	**0.785 6**	**0.793 7**	**0.798 5**

表2 在50 000返回结果上的mAP性能比较

Tab. 2 Performance comparison in terms of mAP scores over 50 000 retrieved images

模型	NUS-WIDE				MIR-FLICKR25K			
	12 bit	24 bit	32 bit	48 bit	12 bit	24 bit	32 bit	48 bit
ITQ(非深度)	0.529 5	0.522 7	0.493 2	0.517 5	0.641 8	0.655 0	0.625 3	0.650 4
PCAH(非深度)	0.456 6	0.420 9	0.401 6	0.397 1	0.609 8	0.603 3	0.608 5	0.616 9
LSH(非深度)	0.330 8	0.368 2	0.372 6	0.391 8	0.570 8	0.588 5	0.584 3	0.601 5
DSH(非深度)	0.506 5	0.511 8	0.490 2	0.480 7	0.656 1	0.659 3	0.644 0	0.642 2
SpH(非深度)	0.382 9	0.395 9	0.390 7	0.384 7	0.586 0	0.578 5	0.578 9	0.578 9
SH(非深度)	0.450 3	0.402 9	0.400 6	0.373 1	0.625 1	0.615 7	0.604 4	0.596 0
AGH(非深度)	0.535 0	0.522 6	0.497 0	0.479 1	0.637 8	0.648 4	0.647 3	0.634 6
DH(深度)	0.403 6	0.387 4	0.393 2	0.401 4	0.583 3	0.594 5	0.593 2	0.594 2
UH-BDNN(深度)	0.498 2	0.499 6	0.483 2	0.485 3	0.632 4	0.627 9	0.627 4	0.625 8
DeepBit(深度)	0.422 5	0.424 7	0.435 9	0.431 0	0.597 4	0.603 2	0.607 7	0.611 5
WDHT-BTV(深度)	0.480 9	0.475 0	0.479 3	0.470 2	0.606 4	0.608 7	0.607 7	0.609 8
WDHT(深度)	0.625 8	0.639 7	0.660 6	0.647 0	0.687 0	0.695 0	0.666 7	0.662 1
IDEA(深度)	**0.652 1**	**0.672 4**	**0.672 0**	**0.687 8**	**0.728 5**	**0.736 8**	**0.741 6**	**0.749 3**

设计的特征,而DeepBit不能充分利用视觉语义信息。③WDHT-BTV和WDHT的性能优于其他基于深度学习的哈希表示学习方法,其中WDHT的mAP超过了WDHT-BTV,这是因为WDHT-BTV的假设不完全正确,导致引入的监督信息不准确,继而影响性能。④本文的IDEA方法达到了最佳的性能,特别地,与WDHT相比,IDEA在两个数据集上均取得了性能提升,这充分反映了捕获细粒度视觉语义信息和判别损失的重要性。

除此之外,本文设计了几种变体方法,以进一步验证IDEA方法的有效性,具体如下:

1)IDEA-L:去除了方法中的判别损失,即设置$\lambda_4 = 0$。

2)IDEA-A:擦除上下文感知的视觉表示学习部分,通过平均池化区域特征得到图像表示,即$\boldsymbol{x}_i = avg_pooling(\overline{\boldsymbol{X}}_i)$。

在NUS-WIDE和MIR-FLICKR25K数据集上对这些变体方法进行实验,实验结果总结在表3和表4中。综合分析这些实验结果,可以发现:①IDEA-L在两个数据集上的检索结果均降低,这表明去除判别损失会对结果造成影响,揭示了判别损失的优势。②IDEA在两个数据集上的性能远高于IDEA-A,这表明平均池化操作不足以充分捕获图像中的细粒度语义信息,验证了图像编码器模块的有效性。③在两个数据集上,IDEA方法检索方面无论是在5 000返回结果还是50 000返回结果上均取得最优结果,这充分验证了增强上下文视觉信息和考虑判别损失的重要性。

4.4 总结与分析

本文提出方法IDEA与现有主流基线方法的对比实验充分验证了IDEA方法的有效性。与此同时,IDEA与其相应变体方法间的对比实验也充分反映了判别损失以及利用上下文信息增强视觉表示的必要性。但是,本文所提出的IDEA为一个弱监督的图像哈希表示学习方法,即训练数据为图像以及图像的文本标签信息,并不依赖图像类别信息。所以,图像文本标签的质量对其学习性能起着至关重要的作用。具体地,如果文本标签与图像内容的语义信息较为匹配,则IDEA可学习到非常鲁棒的图像哈希表示;反之,如果训练图像的文本标签过于嘈杂甚至全部与图像语义信息无关,那么IDEA可能无法学习到具有判别性的图像哈希表示(即语义不同的两张图像学习到相似的哈希表示)。未来,将通过引入外部知识或设计文本标签过滤机制等方式,来解决标签噪声问题,以进一步提升IDEA性能和使用范围。

表 3　变体方法在两个数据集上 5 000 返回结果的性能比较

Tab. 3　Performance comparison among model variants in terms of the mAP scores over 5 000 retrieved images

模型	NUS-WIDE				MIR-FLICKR25K			
	12 bit	24 bit	32 bit	48 bit	12 bit	24 bit	32 bit	48 bit
IDEA-L	0.701 8	0.726 6	0.733 0	0.746 2	0.758 8	0.764 7	0.767 6	0.790 7
IDEA-A	0.702 7	0.737 1	0.737 2	0.744 3	0.759 7	0.774 7	0.781 3	0.782 0
IDEA	0.723 4	0.741 1	0.750 7	0.759 0	0.777 4	0.785 6	0.793 7	0.798 5

表 4　变体方法在两个数据集上 50 000 返回结果的性能比较

Tab. 4　Performance comparison among model variants in terms of the mAP scores over 50 000 retrieved images

模型	NUS-WIDE				MIR-FLICKR25K			
	12 bit	24 bit	32 bit	48 bit	12 bit	24 bit	32 bit	48 bit
IDEA-L	0.631 6	0.649 3	0.660 9	0.673 6	0.704 2	0.722 3	0.715 5	0.733 5
IDEA-A	0.642 5	0.663 6	0.665 5	0.674 8	0.717 8	0.724 5	0.728 5	0.732 5
IDEA	0.657 2	0.672 4	0.672 0	0.687 8	0.728 5	0.736 8	0.741 6	0.749 3

5　结论

本文提出了一个上下文感知的深度弱监督哈希表示学习方法,用于大规模图像检索。特别地,为了更好地利用每个视觉区域的上下文信息并增强它们的表示,本文设计了一个上下文感知的视觉表示提取模块,来动态计算每个视觉区域的视觉注意及其上下文信息。同时,本文引入了一个判别损失来强制图像表示重新生成相应的标签,从而提高图像表示和哈希表示的判别性。为了评估本文提出的方法,本文在两个公共数据集上进行了大量的实验。结果表明,与最先进的基线相比,本文方法可以获得更好的性能。

参考文献(References)

[1] 徐文娟, 易波. 基于离散曲波变换的图像 Hash 算法[J]. 中国图象图形学报, 2011, 16(8): 1374 – 1378.
XU W J, YI B. Image Hash based on discrete curvelet transform[J]. Journal of Image and Graphics, 2011, 16(8): 1374 – 1378. (in Chinese)

[2] 何磊, 钱炜祺, 易贤, 等. 基于转置卷积神经网络的翼型结冰冰形图像化预测方法[J]. 国防科技大学学报, 2021, 43(3): 98 – 106.
HE L, QIAN W Q, YI X, et al. Graphical prediction method of airfoil ice shape based on transposed convolution neural networks [J]. Journal of National University of Defense Technology, 2021, 43(3): 98 – 106. (in Chinese)

[3] JIANG Q Y, LI W J. Asymmetric deep supervised hashing[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018: 3342 – 3349.

[4] CHEN Y D, LAI Z H, DING Y J, et al. Deep supervised hashing with anchor graph [C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 9796 – 9804.

[5] WU D Y, DAI Q, LIU J, et al. Deep incremental hashing network for efficient image retrieval [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 9069 – 9077.

[6] JIN S, ZHOU S C, LIU Y, et al. SSAH: semi-supervised adversarial deep hashing with self-paced hard sample generation[J]. Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020, 34(7): 11157 – 11164.

[7] TU R C, MAO X L, GUO J N, et al. Partial-softmax loss based deep hashing[C]//Proceedings of the Web Conference 2021, 2021: 2869 – 2878.

[8] ZHENG C Q, ZHU L, CHENG Z Y, et al. Adaptive partial multi-view hashing for efficient social image retrieval [J]. IEEE Transactions on Multimedia, 2021, 23: 4079 – 4092.

[9] 刘颖, 程美, 王富平, 等. 深度哈希图像检索方法综述[J]. 中国图象图形学报, 2020, 25(7): 1296 – 1317.
LIU Y, CHENG M, WANG F P, et al. Deep hashing image retrieval methods[J]. Journal of Image and Graphics, 2020, 25(7): 1296 – 1317. (in Chinese)

[10] GATTUPALLI V, ZHUO Y X, LI B X. Weakly supervised deep image hashing through tag embeddings [C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 10367 – 10376.

[11] WANG J F, WANG J D, YU N H, et al. Order preserving hashing for approximate nearest neighbor search [C]// Proceedings of the 21st ACM International Conference on Multimedia, 2013: 133 – 142.

[12] XIA R, PAN Y, LAI H, et al. Supervised hashing for image retrieval via image representation learning[C]//Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014, 28: 2156 – 2162.

[13] LIU H M, WANG R P, SHAN S G, et al. Deep supervised hashing for fast image retrieval [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2064 – 2072.

[14] DAI Q, LI J G, WANG J D, et al. Binary optimized hashing[C]// Proceedings of the 24th ACM International Conference on Multimedia, 2016: 1247 – 1256.

[15] LIN J, LI Z C, TANG J H. Discriminative deep hashing for

scalable face image retrieval[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017: 2266 – 2272.

[16] TANG J H, LI Z C, ZHU X. Supervised deep hashing for scalable face image retrieval[J]. Pattern Recognition, 2018, 75: 25 – 32.

[17] WEISS Y, TORRALBA A, FERGUS R. Spectral hashing[C]// Proceedings of the 22st Conference on Neural Information Processing Systems, 2008.

[18] WANG J, KUMAR S, CHANG S F. Semi-supervised hashing for large-scale search[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(12): 2393 – 2406.

[19] GONG Y C, LAZEBNIK S, GORDO A, et al. Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2916 – 2929.

[20] LIONG V E, LU J W, WANG G, et al. Deep hashing for compact binary codes learning[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015: 2475 – 2483.

[21] LIN K, LU J W, CHEN C S, et al. Learning compact binary descriptors with unsupervised deep neural networks[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1183 – 1192.

[22] TANG J H, LI Z C. Weakly supervised multimodal hashing for scalable social image retrieval[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(10): 2730 – 2741.

[23] GUAN Z Y, XIE F, ZHAO W Q, et al. Tag-based weakly-supervised hashing for image retrieval[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018: 3776 – 3782.

[24] JIANG Q Y, LI W J. Deep cross-modal hashing[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3270 – 3278.

[25] XU X, SHEN F M, YANG Y, et al. Learning discriminative binary codes for large-scale cross-modal retrieval[J]. IEEE Transactions on Image Processing, 2017, 26(5): 2494 – 2507.

[26] WANG J D, ZHANG T, SONG J K, et al. A survey on learning to hash[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 769 – 790.

[27] HUISKES M J, LEW M S. The MIR Flickr retrieval evaluation[C]// Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, 2008: 39 – 43.

[28] JIN L, LI Z C, PAN Y H, et al. Weakly-supervised image hashing through masked visual-semantic graph-based reasoning[C]//Proceedings of the 28th ACM International Conference on Multimedia, 2020: 916 – 924.

[29] CHUA T S, TANG J H, HONG R C, et al. NUS-WIDE: a real-world web image database from National University of Singapore[C]//Proceedings of the ACM International Conference on Image and Video Retrieval, 2009.

[30] CHARIKAR M S. Similarity estimation techniques from rounding algorithms[C]//Proceedings of the 34th Annual ACM Symposium on Theory of Computing, 2002: 380 – 388.

[31] HEO J P, LEE Y, HE J F, et al. Spherical hashing[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012: 2957 – 2964.

[32] JIN Z M, LI C, LIN Y, et al. Density sensitive hashing[J]. IEEE Transactions on Cybernetics, 2014, 44(8): 1362 – 1371.

[33] LIU W, MU C, KUMAR S, et al. Discrete graph hashing[C]// Proceedings of the 28th Conference on Neural Information Processing Systems, 2014.

[34] DO T T, DOAN A D, CHEUNG N M. Learning to hash with binary deep neural network[C]//Proceedings of the 14th European Conference on Computer Vision, 2016.