软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.007543]

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

基于 CLIP 引导标签优化的弱监督图像哈希 *

李泽超[1], 金露[1], 王浩骅[1], 唐金辉[2]

[1](南京理工大学 计算机科学与工程学院,江苏 南京 210094)

[2](南京林业大学,江苏 南京 210037)

通讯作者: 金露, E-mail: lu.jin@njust.edu.cn

摘 要: 在大规模图像检索任务中,图像哈希技术通常依赖大量人工标注数据来训练深度哈希模型,但高昂的人工标注成本限制了其实际应用.为缓解对人工标注的依赖,现有研究尝试利用网络用户提供的文本作为弱监督信息,引导模型从图像中挖掘和文本关联的语义信息.然而,用户标签中普遍存在噪声,限制了这些的方法的性能.多模态预训练基础模型(如 CLIP)具备较强的图像-文本对齐能力.受此启发,本文利用 CLIP 来优化用户标签,并提出一种 CLIP 引导标签优化的弱监督哈希方法(CLIP-guided Tag Refinement Hashing, CTRH).该方法包含三个主要内容:标签置换模块,标签赋权模块和标签平衡损失函数.标签置换模块通过微调 CLIP 挖掘图像关联的潜在标签.标签赋权模块利用优化后的文本和图像进行跨模态全局语义交互,学习判别性的联合表示.针对用户标签的分布不平衡问题,本文设计了一种标签平衡损失,通过动态加权增强模型对难样本的表征学习.在 MirFlickr 和 NUS-WIDE 两个通用数据集上与最先进的方法对比验证了所提方法的有效性.

关键词: 图像检索;弱监督哈希;预训练多模态基础模型;标签优化

中图法分类号: TP311

中文引用格式: 李泽超,金露,王浩骅,唐金辉. 基于 CLIP 引导标签优化的弱监督图像哈希.软件学报. http://www.jos.org.cn/1000-9825/7543.htm

英文引用格式: Li ZC, Jin L, Wang HH, Tang JH. Weakly Supervised Hashing for Image Retrieval via CLIP-Guided Tag Refinement. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7543.htm

Weakly Supervised Hashing for Image Retrieval via CLIP-Guided Tag Refinement

LI Ze-Chao[1], JIN Lu[1], WANG Hao-Hua[1], TANG Jin-Hui[2]

[1](School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)

[2](Nanjing Forestry University, Nanjing 210037, China)

Corresponding author: Jin Lu, E-mail: lu.jin@njust.edu.cn

Abstract: Image hashing typically rely on large-scale manually annotated data to train deep hashing models. However, the high cost of manual annotation limits their practical application. To alleviate this dependency, recent studies have explored using user-provided textual tags as weak supervision to guide hash model capturing semantic information. Nevertheless, the inherent noise in user-generated tags often hinders model performance. Multimodal pre-trained foundation models, such as CLIP, exhibit strong image-text alignment capabilities. Inspired by this, we propose a CLIP-guided Tag Refinement Hashing (CTRH) framework that leverages CLIP to optimize noisy user tags for weakly supervised hashing. The proposed method consists of three key components: a tag replacement module, a tag weighting module, and a tag-balanced loss function. The tag replacement module fine-tunes CLIP to discover potential image-relevant tags. The tag weighting module performs cross-modal global semantic interaction among the refined text and images to learn discriminative joint representations. To address the tag imbalance problem, we design a tag-balanced loss that dynamically reweights training samples to enhance representation learning for hard instances. Extensive experiments conducted on two benchmark datasets, MirFlickr and NUS-WIDE, demonstrate that our method consistently outperforms state-of-the-art approaches, validating its effectiveness.

* 基金项目: 国家自然科学基金(62425603, 62372233); 江苏省基础研究计划攀登项目（BK20240011)
收稿时间:2025-05-26; 修改时间: 2025-07-11; 采用时间: 2025-09-05; jos 在线出版时间: 2025-09-23

Key words: image retrieval; weakly supervised hashing; multimodal pre-trained foundation models; tag refigment

在当今信息爆炸的时代,图像作为信息传递的重要媒介,广泛分布于社交媒体,商业网站以及科研数据库等各类平台,催生了海量的图像数据.如何从这些海量图像中高效并且精准地检索到与用户需求最相关的内容,已成为计算机视觉,信息检索和人工智能领域的热点研究方向之一[1].

图像检索旨在根据用户输入的查询图像,从数据库中检索出与之相关的图像.其中,哈希技术因其能将维度数据转换为低维度的二值哈希码,在降低存储和计算成本方面表现突出,被广泛应用于图像检索任务中[2].近年来,深度学习技术的发展推动了深度哈希方法的广泛研究[3-6],通过端到端的特征学习与哈希编码联合优化,有效提升了检索性能.然而,深度哈希方法通常依赖大规模的人工标注数据,获取成本高昂,限制了其实用性.为此,研究者提出了多种无监督深度哈希方法[7-10],虽然无需依赖标签,但由于缺乏语义引导,模型学习到的哈希码判别能力有限.

为缓解上述问题,已有研究[11-14]尝试利用网络用户提供的文本标签作为弱监督信息进行模型训练.这类标签在一定程度上蕴含图像的语义内容,有助于降低对精确人工标注的依赖,并提升检索性能.然而,由于这些标签往往包含与图像内容相关性较低的噪声信息,限制了模型的检索性能.为了降低噪声标签的影响,Wang 等人[13]提出对用户标签进行优化,通过最小化图像与其标签嵌入之间相似性矩阵的负迹值,迭代式的将与图像具有最大相似性的标签加入当前标签集合,同时从集合中移除相似性最小的标签.由于没有考虑图像和文本之间的交互,限制了文本标签优化的质量.此外,Du 等人[14]通过注意力机制来学习图像和文本之间的关联,对噪声标签赋值较低的权重,并通过加权得到联合表示来引导图像哈希码的学习.这些方法虽然获得了一定的性能提升,但是并不能有效的解决噪声标签问题,使得学习到的联合表示是次优的.

近年来,预训练多模态基础模型[15-18]被广泛用于跨模态检索,视觉问答,分类等任务.其中,Contrastive language-image pretraining (CLIP)模型[18]因其出色的图像-文本对齐能力而备受关注.CLIP 基于从互联网收集的 4 亿对图文数据进行训练,所学得的多模态表示融合了丰富的视觉概念和语言语义知识.受此启发,本文充分挖掘 CLIP 蕴含的多模态知识与跨模态对齐能力,对用户提供的文本进行重标注与语义补全,并结合优化后的文本集,通过多模态交互机制增强图文联合表示的语义表达能力.

为此,本文提出了一种基于 CLIP 引导标签优化的哈希模型(CLIP-guided Tag Refinement Hashing,CTRH).该方法充分利用 CLIP 所蕴含的视觉语义知识以及其跨模态理解能力,来引导弱监督标签的优化和哈希学习过程.CTRH 方法由三个核心部分组成:弱监督标签置换模块,弱监督标签赋权模块和标签平衡损失.弱监督标签置换模块通过微调 CLIP 模型并融合其预训练模型参数,实现对图像标签的推理,通过图像与标签库之间的相似性匹配,生成更为准确的标签集合.弱监督标签赋权模块将主干网络提取到的图像特征与 CLIP 文本编码器得到的文本特征映射到统一特征空间,以增强图像与文本之间的交互,从而获得更加丰富的多模态联合表示.由于用户文本标签分布不均匀,本文引入了标签平衡损失通过对样本对动态加权,使得模型更加关注尾部类标签的学习.在哈希学习部分,本文利用相似性保留损失,铰链损失和量化损失,引导哈希码有效捕获联合特征蕴含的语义知识.为了验证所提方法的有效性,本文在 MirFlickr 和 NUS-WIDE 两个通用数据集上与多种先进方法进行了对比实验.实验结果表明,CTRH 在检索性能方面均取得了提升,充分验证了其在弱监督哈希检索任务中的优越性.

1 相关工作

本文从弱监督哈希学习和预训练多模态基础大模型进行相关研究工作的介绍.

1.1 弱监督深度哈希学习

弱监督深度哈希方通过使用网络用户提供的弱监督标签,既有效降低了人工标注成本,又在图像检索任务中取得了良好的性能.目前,较新的弱监督深度哈希图像检索方法通常采用诸如 AlexNet[19]等深度模型来提取图像特征,并使用 word2vec 模型[20]提取文本特征作为弱监督信号.其中,最早提出的方法是基于标签嵌入的弱监督深度哈希方法(Weakly Supervised Deep Image Hashing Through Tag Embedding,WDHT)[11].该方法平等的对

待每个文本标签,使用平均聚合标签向量作为弱监督信号,然而该策略容易受到噪声标签的干扰,影响模型性能.基于掩蔽视觉语义图推理的方法(Masked Visual-Semantic Graph-based Reasoning,MGRN)[12]通过构建视觉-语义图,并训练图神经网络预测随机遮蔽的标签,从而以自监督方式学习图像与标签的联合表示.尽管在性能上有所提升,但该方法需处理复杂的图结构信息,导致计算开销显著增加.基于迭代标签优化的方法(Deep Enhanced Weakly-Supervised Hashing with Iterative Tag Refinement,EWSH)[13]根据图像内容对用户提供的标签进行优化,以获得更为准确的弱监督信息.然而,该方法在优化过程中未充分考虑图像与标签之间的模态交互,限制了模型对跨模态语义关联的挖掘能力.基于重构跨模态注意力的方法(Weakly Supervised Hashing with Reconstructive Cross-modal Attention,WSHRCA)[14]则通过哈希码重构图像特征,以显式更新哈希表示,同时引入跨模态注意力机制优化特征学习.该方法在一定程度上提升了性能,并降低了噪声标签的影响.但其对噪声标签的处理仍较为有限,仅通过削弱权重因子进行调整,哈希学习过程仍可能受噪声干扰.

1.2 预训练多模态基础大模型

最近几年,预训练多模态基础大模型取得了显著进展.通过联合训练大规模图像与文本数据,这类视觉-语言模型在图像理解,文本生成,跨模态检索等多种任务中展现出卓越性能[21-22].其中,ViLT(Vision-and-Language Transformer)[15]是一种基于 Transformer[23],用于自监督学习的视觉和语言交互式模型...该模型无需依赖卷积神经网络或区域级监督,直接通过最大似然估计优化图像与文本的跨模态匹配关系,从而学习到通用且高质量的多模态表示.UNITER(Universal Image-Text Representation Learning)[16]采用条件掩蔽策略,增强视觉与语言之间的对齐效果,并设计了四种预训练任务,从不同角度挖掘多模态关联性,提升表示能力.Oscar(Object-Semantics Aligned Pre-training)[17]则引入目标检测器提取图像中的对象标签作为额外语义信息,结合对比学习与负采样策略,进一步加强图像与文本之间的语义关联.CLIP(Contrastive Language-Image Pretraining)[18]则开创性地采用大规模对比学习框架,将图像与文本嵌入映射到统一语义空间中,从而实现高效的跨模态语义对齐.由于其训练不依赖下游任务的标签,CLIP 具备强大的零样本迁移能力,能够广泛适用于多种下游任务.受此启发,本文引入 CLIP 所蕴含的多模态先验知识与跨模态对齐能力,对用户提供的弱标签进行优化与补全,增强联合表示的语义表达能力,从而提升哈希检索的性能.

2 本文所提方法

本文所提的 CTRH 方法主要包含弱监督标签置换模块、弱监督标签赋权模块和哈希学习,下面具体介绍相关内容.

2.1 问题描述

在弱监督哈希图像检索任务中,通常将数据集划分为训练集,数据库集和查询集.假设训练集记为 $S_{tr}=\{(x_i,T_i)\,|\,i=1,\cdots,N_{tr}\}$,其中 x_i 表示第 i 张图片,N_{tr} 为训练集中图像的总数,T_i 表示与 x_i 相关联的弱监督文本标签,,,由网络用户提供.具体地,$T_i=[T_{i1},\cdots,T_{iC}\ ]$,这里 C 表示图像 x_i 所关联的文本标签数量.数据集和查询集分别表示为 $S_{db}=\{(x_i,y_i)\,|\,i=1,\cdots,N_{db}\}$ 和 $S_q=\{(x_i,y_i)\,|\,i=1,\cdots,N_q\}$,其中 y_i 表示图像 x_i 的真值标签,N_{db} 和 N_q 分别为数据库集和查询集中的图像数量.

给定图像 x_i,本文利用编码器抽取图像特征,并通过哈希映射得到对应的哈希码,表示为:

$$h_i=\sigma(f_{hash}(Encoder(x_i))) \tag{1}$$

这里 *Encoder* 为编码器骨干网络,$f_{hash}(\cdot)$ 表示哈希层将高维特征投影到低维哈希空间,$\sigma(\cdot)$ 表示 sigmoid 激活函数,$h_i\in R^L$ 表示长度为 L 的哈希码.

2.2 CTRH的总体框架图

CTRH 的整体框架如图 1所示,主要包括三个模块: 弱监督标签置换,弱监督标签赋权和哈希学习. 在标签置换模块中,为增强 CLIP 对下游任务的适应能力,本文使用下游数据对其进行微调,并将微调后的参数与原始 CLIP 参数加权融合,基于融合模型对图像进行标签推理,以获得更准确的文本标签.在标签赋权模块中,使用

CLIP 文本编码器提取标签语义特征,并结合多头自注意力机制,实现图像与文本特征的深度交互,生成更具判别力的多模态表示.在哈希学习模块中,设计了量化损失、相似性保持损失和铰链损失,联合引导哈希码学习多模态语义信息.同时,引入标签平衡损失,通过动态调整样本对权重,提升对尾部类别的关注度.

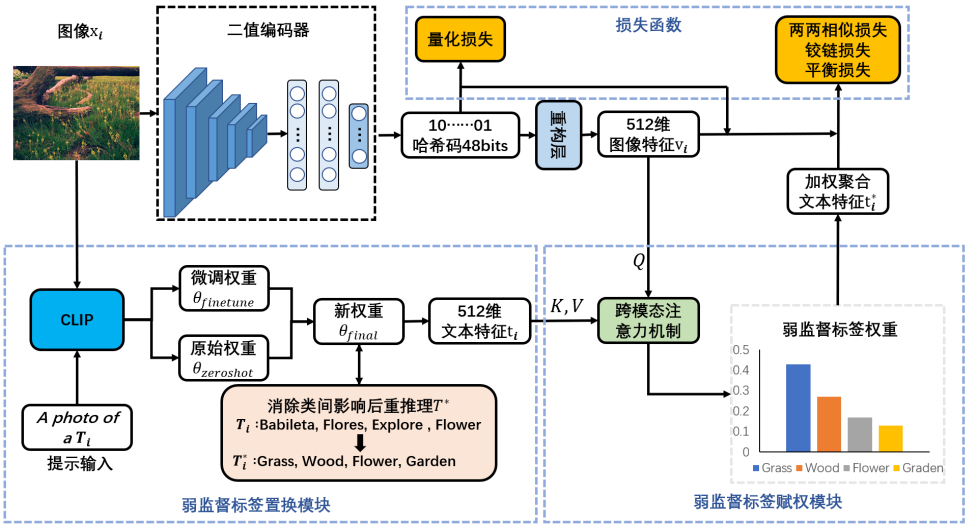

图 1 CTRH 的整体框架示意图

2.3 弱监督标签置换模块

在理解图像内容的基础上对噪音标签进行置换是弱监督哈希的关键.基于此目的,本文使用具有较好跨模态理解能力的大规模视觉语言模型 CLIP 来引导这一过程.本模块的流程大致可以分为以下几个步骤:

(1) 微调 CLIP 模型.虽然 CLIP 模型拥有强大的零样本预测能力,但为了使其更加适用下游任务,本文首先对 CLIP 模型进行微调.对于图像 x_i 使用提示模版“*a photo of* T_{i1}, T_{i2},⋯ and T_{iC}”作为其对应的文本描述.微调过程中,本文设置较小的学习率并且 10 轮训练后将学习率降低为原来的 0.1 倍,以避免对预训练模型的权重做出过大的改动.同时,对大模型的训练过程采用混合精度来减少显存的使用并提高计算效率.此外,本文使用交叉熵损失对图像编码器和文本编码器进行优化.这里将微调后的权重记为 $\theta_{finetune}$,原始权重记为 $\theta_{zeroshot}$.

(2) 权重融合.在上述微调 CLIP 模型的过程中,由于使用的是带有噪声的图片文本对,若直接依赖该数据进行训练,可能会破坏 CLIP 原有的视觉-语义知识结构.如图 2 所示,虽然微调后的模型可以去除原始文本标签中较明显的噪声标签,但由于使用的数据集中的动植物图像带有“花园”和“户外”等如图中红色标注所示的文本,微调后的模型在推理阶段会更倾向于预测这些标签.然而,这些用户标签其实并不能很好的表示图像的内容.相比之下,若将微调后的权重与原始预训练权重进行融合后再进行推理,则更容易获得如图中绿色部分所示的,更具语义针对性的标签.为此,本文采用线性插值策略进行权重融合,即以加权方式结合微调后的模型参数与原始 CLIP 预训练参数,具体表达式为:

$$\theta_{final} = (1-\alpha)\theta_{zeroshot} + \alpha\theta_{finetune} \tag{2}$$

其中,α 是插值系数可以控制模型权重的贡献比例.适当的插值比例可以使得模型既能学习到新任务的特征,又能保留预训练模型原始的知识.尤其是在弱监督任务中,权重融合可以避免受到噪声标签的较大影响.

	弱监督标签	zoo, mariposa, butterfly ……
	微调后预测标签	butterfly, garden, nature, outdoor
	融合后预测标签	butterfly, wild life, nature, plant
	真值标签	animals, flower, plant life

图 2 利用融合权重的 CLIP 模型推理的文本标签示例

(3) 消除类间影响.在推理阶段,本文使用上述的融合权重 θ_{final} 对图像标签进行预测,并选择预测概率排名前 K 的标签替换原始的文本标签.为了提升标签替换的灵活性与适应性,本文根据原始文本标签在数据集中的常见标签集中的出现的次数动态调整 K 值;若原始标签中无任何标签出现在常见标签集中,则将 K 值设置为原始标签的数量.具体定义如下:

$$K=\begin{cases} Count(T_i \bigcap T_{common}),\text{if } Count(T_i \bigcap T_{common})>0 \\ Count(T_i),\text{if } Count(T_i \bigcap T_{common})=0 \end{cases} \tag{3}$$

其中,T_{common} 表示常见文本标签的集合,$Count(\cdot)$ 表示计数函数.在利用上述方法推理得到前K个优化标签的过程中,本文观察到显著标签对其他标签的预测存在干扰.如图 3 所示,当输入图像中存在与“人”相关的显著语义内容时,与其相关的文本标签在预测结果中会获得较高的置信度分数,从而主导整个预测排序.这些显著标签会压制其他语义相关但不显著的标签的得分,导致原本与图像内容密切相关的标签(如“天空”等)未能进入前个标签的候选集,如图中红色区域所示.

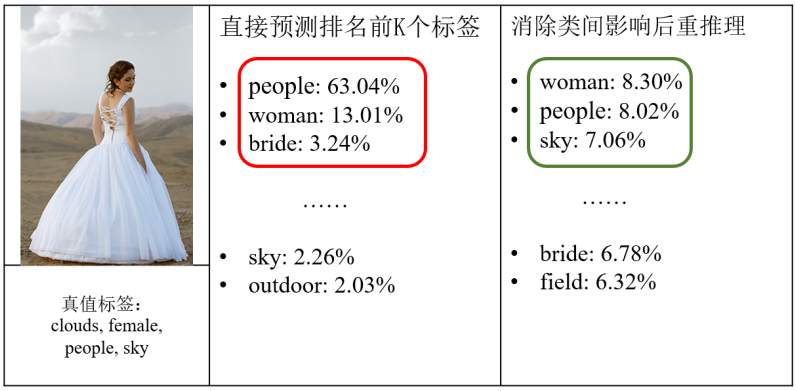

图 3 消除类间影响得到的文本标签示例

为消除推理阶段不同类别之间的相互干扰,本文提出对各类别之间的冗余特征进行建模与抑制.具体而言,首先利用融合权重 θ_{final} 得到的图像特征 $F_{im}\in\mathbb{R}^{1\times 50\times 512}$ 与文本标签的特征 $F_{txt}\in\mathbb{R}^{C\times 512}$,分别进行 L2 归一化处理.随后,通过对图像特征与文本特征维度扩充,进行逐元素相乘,构建原始的跨模态交叉表征 $F_o\in\mathbb{R}^{1\times 50\times C\times 512}$.该过程可表示为:

$$F_o=\text{expand}(\frac{F_{im}}{\|F_{im}\|_2})\odot\text{expand}(\frac{F_{txt}}{\|F_{txt}\|_2}) \tag{4}$$

这里,$\odot$ 表示向量逐元素相乘,$\text{expand}(\cdot)$ 表示维度扩充操纵.之后,使用 CLIP 图像特征 F_{im} 中的类别标记特征 $F_c\in\mathbb{R}^{1\times 512}$ 和文本特征 $F_{txt}\in\mathbb{R}^{C\times 512}$ 计算相似度分数 $s\in\mathbb{R}^{1\times C}$:

$$s = soft\max(\frac{F_c}{\|F_c\|_2}(\frac{F_{txt}}{\|F_{txt}\|_2})^T) \tag{5}$$

通过各类别的相似度分数和均值分数的比值计算权重 $w \in \mathbb{R}^{1\times C}$:

$$w = \frac{s}{mean(s)} \tag{6}$$

这里, $mean(\cdot)$ 表示求均值运算.然后对每一个特征加权,并在类别维度求均值作为该类别受其他类别影响的冗余特征 $F_r \in \mathbb{R}^{1\times 50\times 512}$:

$$F_r = mean(F_o \odot \text{expand}(w)) \tag{7}$$

通过对原始的融合 F_o 去除受各类别影响的冗余特征 F_r 可以得到新的融合表征 F_n:

$$F_n = F_o - \text{expand}(F_r) \tag{8}$$

利用上述得到的新的融合表征 $F_n \in \mathbb{R}^{1\times 50\times C\times 512}$ 进行推理,可以减少显著标签对其他标签的影响.本文对融合表征的图像维度和特征维度求均值,利用 $F_n \in \mathbb{R}^{1\times C}$ 预测,取分数最高的前 K 个标签作为最后的置换标签.如图 3 中绿色区域所示,重新推理后得到的标签中不会出现某些显著类别预测概率分数很高但其他类别都很低的情况,这也使得与图片内容相关的非显著类同样可以被取到.

2.4 弱监督标签赋权模块

弱监督标签赋权模块采用 Transformer 模型中的多头自注意力机制来捕获图像和优化标签之间的全局依赖关系.本文首先将哈希码通过重构层将其映射为 512 维的特征 v:

$$v = f_{re}(h) \tag{9}$$

这里 $f_{re}(\cdot)$ 表示重构层.同时,利用融合权重后的 CLIP 文本编码器将优化后的标签转化为 512 维的文本特征 t:

$$t = CLIP^{text}(T) \tag{10}$$

这里 $CLIP^{text}$ 表示为 CLIP 的文本编码器.对于 MSA 中的每个头 $\varphi \in \{1,2,\cdots,h\}$,都有三个权重矩阵 $W_\varphi^Q, W_\varphi^K, W_\varphi^V$ 用于计算 Q_φ, K_φ 和 V_φ:

$$Q_\varphi = W_\varphi^Q v, K_\varphi = W_\varphi^K t, V_\varphi = W_\varphi^V t \tag{11}$$

接着利用缩放点积注意力得到相似度分数并和 V_φ 相乘得到如下表示 Z_φ:

$$Z_\varphi = \text{softmax}(\frac{Q_\varphi K_\varphi^T}{\sqrt{d_k}})V_\varphi \tag{12}$$

其中, $\sqrt{d_k}$ 是缩放因子用来避免梯度消失或者梯度爆炸的问题, d_k 是 K_φ 的维度.将所有头的输出拼接在一起并通过一个线性层得到输出特征 Z,计算表达式为:

$$Z = Concat(Z_1, Z_2, \cdots, Z_h)W_0 \tag{13}$$

其中, W_0 表示线性层的权重矩阵.为了增强局部特征和非线性表达能力,这里使用一个前馈网络进一步处理和转换输出的特征,并进行残差连接得到最终的文本表示 t^*:

$$t^* = Z + FFN(Z) \tag{14}$$

这里 FFN 表示前馈网络,主要包括两个全连接层,一个激活层和一个 Dropout.第一个全连接层将 512 维的文本特征扩展到更高的维度,为激活函数提供更多非线性变换的可能.激活层采用 ReLU 激活函数引入非线性,增加模型的表达能力.Dropout 层可以提高模型的泛化能力.

2.5 哈希学习

为了学习具有判别性的哈希码,本文利用量化损失,铰链损失和相似性保留损失保留文本特征中蕴含的丰富的 CLIP 的预训练多模态知识.此外,引入标签平衡损失用来缓解标签不平衡问题.

(1) 量化损失.该损失通过约束哈希码的每位接近于 0.5,使得模型输出的值近似二值化,计算表达式如下所示:

$$L_q = \sum_{i=1}^{N_b} \frac{1}{L} \left\| h_i - 0.5I \right\|_2^2 \tag{15}$$

其中, h_i 表示图像 x_i 的哈希码, N_b 表示最小批次的大小,L 代表哈希码中的比特数量,I 是与哈希码同维度的全 1 向量.量化损失的意义在于确保连续特征在二值化时尽量保留语义信息.同时,引导模型稳定优化,避免二值化带来梯度消失的问题.

(2) 铰链损失.该损失的作用是通过在高维特征空间对齐公式(14)中文本表示 t^* 和公式(9)中哈希码重构的图像表示 v 来增强哈希码的判别能力.该损失的计算表达式如下所示:

$$L_{hinge} = \sum_{i=1}^{N_b} \sum_{j=1, j\neq i}^{N_b} \max(0, \varepsilon + \frac{t_j^* v_i^T}{\left\| t_j^* \right\|_2 \left\| v_i \right\|_2} - \frac{t_i^* v_i^T}{\left\| t_i^* \right\|_2 \left\| v_i \right\|_2}) \tag{16}$$

其中, ε 是边界参数用于控制匹配对和非匹配对之间的最小相似度差.

(3) 相似性保留损失.该损失项旨在促使样本在哈希空间中保留其在文本语义空间中的相似性关系,从而使生成的哈希码能够有效表达原始文本的语言信息.其具体表达式为:

$$L_{pair} = \sum_{i=1}^{N_b} \sum_{j=1}^{N_b} [\frac{1}{L} \left\| h_i - h_j \right\|_2^2 - \frac{1}{2}(1 - \frac{t_i^* t_j^{*T}}{\left\| t_i^* \right\|_2 \left\| t_j^* \right\|_2})]^2 \tag{17}$$

该损失促使样本在哈希空间的汉明距离与余弦距离具有相似的分布.

(4) 标签平衡损失.由于文本标签分布不均匀,导致模型更加倾向于拟合头部类标签.为此,本文引入标签平衡损失通过对样本对动态加权,使得模型更加关注尾部类标签的学习.在正样本比例较低时赋予较高的权重,同时使模型更加聚焦于困难样本.标签平衡损失计算表达式如下式所示:

$$L_b = -\frac{1}{m} \sum_{i=1}^{m} \sum_{c=1}^{C} \beta T_{ic}^* [1 - \sigma(\frac{t_c^* v_i^T}{\left\| t_c^* \right\|_2 \left\| v_i \right\|_2})]^\gamma \log[\sigma(\frac{t_c^* v_i^T}{\left\| t_c^* \right\|_2 \left\| v_i \right\|_2})] \tag{18}$$

其中, β 是平衡因子,C 是经过优化后的新标签集 T^* 中的类别数, T_{ic}^* 表示图片 x_i 在 T^* 中的独热标签向量,参数 γ 作为调节因子用于调整对较难样本的关注程度, $\sigma(\cdot)$ 表示 sigmoid 激活函数.

最终,本文的总体目标函数如下式所示:

$$L = \lambda_1 L_q + \lambda_2 L_{hinge} + \lambda_3 L_{pair} + \lambda_4 L_b \tag{19}$$

其中, λ 是各损失对应的权重因子.

3 实验分析

为了验证本文所提方法的有效性,在弱监督图像检索领域常用的两个数据集 MirFlickr[24]和 NUS-WIDE[25]上进行对比实验验证.

3.1 实验数据集介绍

MirFlickr 数据集包含了 25000 张图片,分为 38 个通用类别.每张图片归属于一个或者多个类别,并且包含了不定数量的由用户提供的弱监督标签.这些弱监督标签中较为常见的有 1386 种,模型在训练过程中只能使用用户提供的弱监督标签,对精细标注的 38 种真值标签是不可见的.本文从中随机选取 2000 张图片作为查询集,其

余图片作为数据库集和训练集.

NUS-WIDE 数据集包含 269498 张图片,81 个通用类别.与 MirFlickr 数据集相似,NUS-WIDE 数据集的每张图片是多标签的,并且拥有不定数量的弱监督标签.本文选用了 81 个类别中最常见的 10 个类别,总共包括 181365 张图片.从中随机选择 5000 张图片作为查询集,其余图片作为数据库集,并从数据库集中选择 10500 张图片作为训练集.

3.2 评价指标和对比方法

为了评估所提方法的有限性,本文与一些先进的方法进行对比.这些方法三种无监督的方法(LSH[26],SH[27]和 ITQ[28]),和四种近期最具有代表性的弱监督方法(WDHT[11],MGRN[12],EWSH[13]和 WSHRCA[14]).本文采用了图像检索领域常用的四种评估方法:前 5000 个检索结果的平均精度(mAP@5000),所有检索结果的平均精度(mAP@all),前 N 个返回值准确率曲线(P@N)和准确率回归曲线(PR).

3.3 实验设置

本文利用预训的 AlexNet 网络作为主干网络提取特性特征,接着连接一个用 Sigmoid 函数激活的哈希全连接层输出哈希码.重构层 f_{re} 包含一个由 Leaky ReLU 函数激活的全连接层.大规模视觉语言模型 CLIP 使用的权重是 ViT-B-32,微调过程中采用 Adam 作为优化器,学习率设置为 10^{-6} 并且每 10 轮降低为原来的 0.1 倍,一阶和二阶矩估计的指数衰减率分别是 0.9 和 0.98,L2 正则化参数为 0.001,训练轮次设置为 30 次迭代.参数 α 分别设置为对于 MirFlickr 数据集设置为 0.5,NUS-WIDE 数据集设置为 0.25. CTRH 使用 SGD 优化器,对特征提取的主干网络 AlexNet 学习率设置为 10^{-4} ,哈希层和重构层的学习率设置为 10^{-3} ,弱监督标签赋权模块的学习率为 2×10^{-4} .在 MirFlickr 数据集, $\alpha=0.5$, $\varepsilon=0.7$, 而在 NUS-WIDE 数据集上 $\alpha=0.25$, $\varepsilon=0.9$.其他参数两个数据集上钧设置为: $\lambda_1=\lambda_3=1$, $\lambda_2=10$, $\lambda_4=15$.

3.4 实验结果分析

由于目前基于弱监督学习的深度哈希图像检索方法较少,因此除了弱监督的方法之外,还和三种无监督的方法进行比较.为了公平起见,所有方法的主干网络均使用 AlexNet 来提取特征,本文所提方法的视觉分支不会使用到任何 CLIP 模型的图像特征.哈希码的长度分别设置为 16bits,32bit3,48bits 和 64bits.各方法在两个数据集中不同哈希码长度的 mAP@all 如表 1 所示.

表 1 CTRH 和其他方法在两个数据集上的 mAP@all(%)结果,最优结果已加粗表示

方法	主干网络	MirFlickr				NUS-WIDE			
		16bits	32bits	48bits	64bits	16bits	32bits	48bits	64bits
传统无监督哈希方法									
LSH	-	56.23	57.75	57.72	59.13	36.34	38.25	40.21	42.67
SH	-	59.54	58.73	58.78	58.96	41.16	43.84	42.47	41.75
ITQ	-	64.82	65.57	66.03	66.17	54.32	56.16	56.94	57.35
深度弱监督哈希方法									
WDHT	AlexNet	66.43	68.62	69.43	69.02	55.92	63.34	60.72	62.34
MGRN	AlexNet	69.78	70.45	70.81	70.60	62.08	63.52	64.13	64.01
EWSH	AlexNet	70.12	71.63	71.83	72.08	64.32	64.14	65.08	64.87
WSHRCA	AlexNet	72.03	72.76	73.79	72.91	65.95	67.41	68.09	68.23
CTRH	AlexNet	**72.21**	**74.03**	**75.38**	**75.37**	**67.79**	**67.63**	**68.49**	**68.96**

从实验结果中可以看出,本文的方法 CTRH 在不同哈希码长度设置中对于两大通用数据集 MirFlickr 和 NUS-WIDE 的检索结果均要优于现有的深度弱监督哈希方法.在 MirFlickr 数据集中,本文的方法对比目前该领域效果最好的方法在不同哈希码长度的设置下均有一定提升,平均提高了 1.4%.由此可见,引入 CLIP 模型可以有效引导弱监督标签的优化

.并且随着哈希码长度的增加,本文发现 CTRH 模型对比其他方法的提升也逐渐升高,这说明了本文的方法为哈希码的表示提供了更多有意义的信息.其中,当哈希码长度为 48 比特时取得了最好的平均检索精度.这也说明哈希码的长度并不是越长越好,64 比特长度的哈希码中可能存在一些冗余的信息,使得其精度并不如 48 比

特的哈希码.对于更加复杂的 NUS-WIDE 数据集中,本文的方法在不同哈希码长度的设置下也平均提高了0.8%.其中,在哈希码长度为 16 比特时提升了 1.9%,但是随着哈希码长度的增加,提升幅度均低于平均水平.这可能是因为 NUS-WIDE 数据集的图片内容更为复杂,这使得模型在视觉分支中提取到更多的冗余特征,从而影响整体的哈希学习.

对于用户而言,在所有检索返回项中大家更加关心的是前一部分的检索结果,因此在检索任务中前 5000 个返回结果的平均精度 mAP@5000 是更加重要的评价指标如表 2 所示.

表 2 CTRH 和其他方法在两个数据集上的 mAP@5000(%)结果,最优结果已加粗表示

方法	主干网络	MirFlickr				NUS-WIDE			
		16bits	32bits	48bits	64bits	16bits	32bits	48bits	64bits
传统无监督哈希方法									
LSH	-	57.24	59.19	60.03	62.53	41.09	47.17	52.69	57.81
SH	-	60.17	62.76	63.66	63.59	56.02	63.62	62.81	62.38
ITQ	-	70.15	71.47	71.99	72.21	65.98	66.90	67.56	68.72
深度弱监督哈希方法									
WDHT	AlexNet	70.83	74.32	74.91	74.70	60.64	69.51	70.39	68.86
MGRN	AlexNet	75.35	76.51	77.04	76.89	73.59	76.85	78.04	78.55
EWSH	AlexNet	75.29	76.99	77.57	78.16	72.26	73.72	75.86	77.37
WSHRCA	AlexNet	78.08	79.22	80.32	79.46	75.66	77.80	78.68	79.81
CTRH	AlexNet	**78.80**	**82.66**	**83.33**	**82.98**	**78.19**	**78.50**	**78.98**	**80.61**

在两个通用数据集中,使用 16bits,32bits,48bits 和 64bits 长度的哈希码对现有方法进行测试后发现,本文方法准确率的提升更为显著.在 MirFlickr 和 NUS-WIDE 数据集中,不同哈希码长度设置下对比目前最优的方法分别提升了 2.7%和 1.1%.这说明了本文的方法对前一部分检索结果的优化更为明显,这是因为在提供更加准确的弱监督信号后,模型可以更好的对图像内容进行哈希学习.但是对于 NUS-WIDE 数据集中较长哈希码的情况,提升幅度同样低于平均水平.

本文绘制了 MirFlickr 和 NUS-WIDE 数据集在 4 种不同哈希码长度的下各方法的 P@N 曲线,如图 4 和图 5 所示.可见,对于在排名前 1000 的返回值内本文的方法在两个数据集上都优于现有的方法.

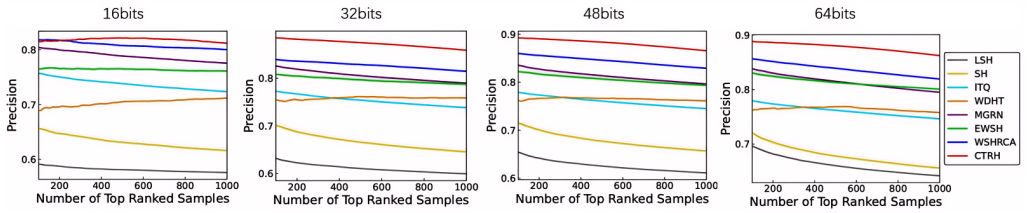

图 4 各方法在 MirFlickr 数据集上不同哈希码长度的 P@N 曲线

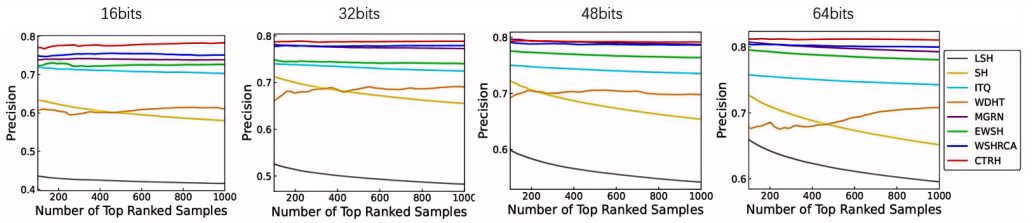

图 5 各方法 NUS-WIDE 数据集上不同哈希码长度的 P@N 曲线

本文还评估了各方法在 MirFlickr 和 NUS-WIDE 数据集上不同哈希码长度的 PR 曲线,如图 6 和图 7 所示.可见,本文的方法对比目前该领域最好的方法有一定的提升.

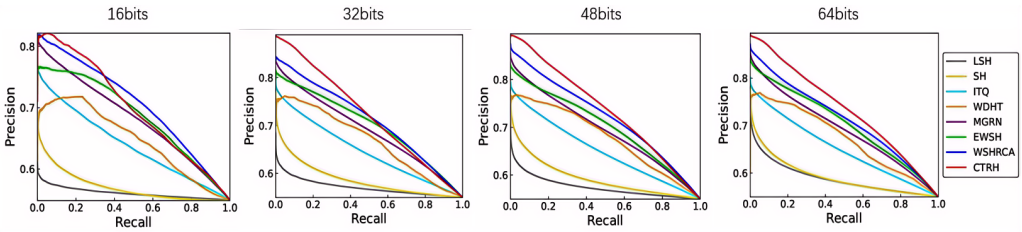

图 6 各方法在 MirFlickr 数据集上不同哈希码长度的 PR 曲线

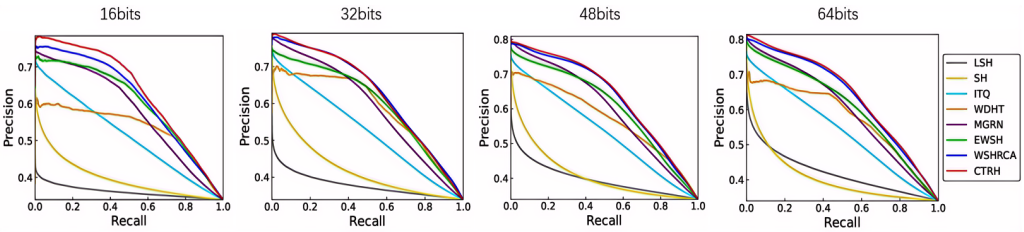

图 7 各方法在 NUS-WIDE 数据集上不同哈希码长度的 PR 曲线

为了更加直观的展示哈希图像检索的可视化结果,本文将输入查询图像后得到的前十个返回结果汇总如图 8 所示.其中,绿色框标注的图片代表检索正确的结果,红色框标注的代表检索有误的图片.从中可见,本文的方法在弱监督哈希图像检索领域可以根据图像内容取得不错的检索效果.

图 8 前十个检索结果的效果展示

除了对检索性能的定量分析外,在哈希检索任务中,时间与空间开销同样是衡量方法实用性的重要指标,尤其是在本文引入了弱监督标签置换和标签赋权等结构模块后,需进一步验证其在实际检索过程中的效率表现.值得说明的是,这些模块仅在训练阶段参与优化,而在测试阶段,模型仅保留轻量的哈希编码网络用于高效检索.为评估检索过程中的时间与空间效率,本文在 MirFlickr 和 NUS-WIDE 两个数据集上, 上分别测试了不同模型生成 64 比特哈希码的平均时间和存储开销.具体地,我们对每个数据集中的所有图像进行了前向传播推理,记录单张图像的平均推理时间和最大显存使用,结果如表 3 所示.从结果中可以看出, CTRH 在推理阶段与其他深度弱监督哈希方法相比,在时间和空间开销方面均未带来额外负担,反而在两个指标上略具优势.这进一步说明,本文提出的方法在保持检索效率的同时,兼顾了计算资源的友好性.

表 3 CTRH 和其他方法在两个数据集上的平均时间和空间开销比较,最优结果已加粗表示

方法	主干网络	MirFlickr (64bits)		NUS-WIDE (64bits)	
		时间开销（ms）	空间开销（M）	时间开销（ms）	空间开销（M）
WDHT	AlexNet	1.44	232.18	1.16	232.18
MGRN	AlexNet	1.37	229.05	1.11	229.05
EWSH	AlexNet	1.38	232.58	1.14	232.58
WSHRCA	AlexNet	1.40	228.89	1.13	228.89
CTRH	AlexNet	1.35	228.95	1.13	228.95

3.5 消融实验分析

为了验证 CTRH 模型中所提出的每个模块的有效性,本文在 MirFlickr 和 NUS-WIDE 数据集中 4 种不同哈希码长度的情况下进行了消融实验.表 4 和表 5 分别展示了 mAP@5000 和 mAP@all 的性能指标.其中,第一行代表去除所有优化方法只保留哈希学习框架的结果.第二行在其基础上增加了弱监督标签置换模块,可见效果得到了明显的提升,尤其在噪声较多的 NUS-WIDE 以及哈希码长度很小这样的极端情况下效果更为明显.第三行在第二行的基础上增加了弱监督标签赋权模块,通过进一步优化对更新后标签的使用,其准确率同样得到了提升.第四行是本文所提出的方法,在第三行的基础上引入了平衡损失,有利于模型更好的学习困难样本.

表 4 CTRH 模型添加不同模块后 mAP@5000(%)对比,最优结果已加粗表示

方法	主干网络	MirFlickr				NUS-WIDE			
		16bits	32bits	48bits	64bits	16bits	32bits	48bits	64bits
基准模型	AlexNet	64.99	73.53	74.25	75.58	34.52	65.85	69.15	70.01
增加 TR	AlexNet	72.78	76.17	77.74	78.08	68.77	74.09	76.09	77.30
增加 TW	AlexNet	77.89	81.74	82.63	82.09	77.58	77.62	78.04	79.83
CTRH	AlexNet	**78.80**	**82.66**	**83.33**	**82.98**	**78.19**	**78.50**	**78.98**	**80.61**

表 5 CTRH 模型添加不同模块后 mAP@all(%)对比,最优结果已加粗表示

方法	主干网络	MirFlickr				NUS-WIDE			
		16bits	32bits	48bits	64bits	16bits	32bits	48bits	64bits
基准模型	AlexNet	63.60	68.01	68.32	69.52	34.00	58.23	58.97	59.22
增加 TR	AlexNet	68.90	70.12	71.25	71.64	61.96	62.86	63.67	64.75
增加 TW	AlexNet	71.62	73.17	74.85	74.62	67.02	66.89	67.64	68.23
CTRH	AlexNet	**72.21**	**74.03**	**75.38**	**75.37**	**67.79**	**67.63**	**68.49**	**68.96**

为了分析 CTRH 的参数敏感性,本文使用 48bits 长度的哈希码分别在 MirFlickr 和 NUS-WIDE 数据集上对插值系数 α 、边界参数 ε 和平衡损失比例参数 λ_4 的取值进行分析,如图 9 所示. 其中,对于插值系数 $\alpha=0.5$ 时,在 MirFlickr 数据集中效果最好,这说明了微调过程中学到的新知识和原始的知识同样重要.而对于噪声更大的 NUS-WIDE 数据集而言,则 $\alpha=0.25$ 效果更好.这说明在 NUS-WIDE 数据集中进行微调时,过多的噪声标签不仅没有使得 CLIP 更加适应下游数据,反而破坏了 CLIP 模型中原有的视觉语言知识.当铰链损失中边界值 $\varepsilon=0.7$ 时,在 MirFlickr 数据集中效果最好.而对于噪声更大的 NUS-WIDE 数据集则需要当 $\varepsilon=0.9$ 时效果最好.在对损失函数超参数的探讨中,本文这里只对新引入平衡损失的超参数 λ_4 的取值进行实验.当 $\lambda_4=15$ 时在两个数据集中的总体效果最好.

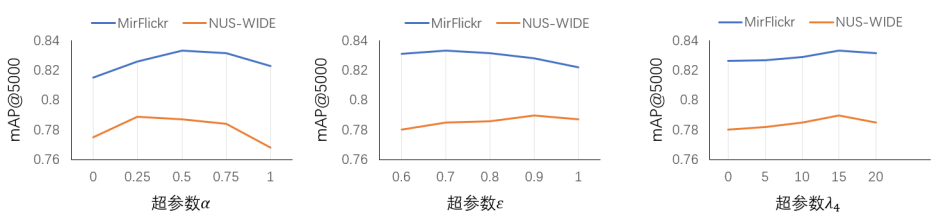

图 9 消融实验结果

为了分析 CTRH 在不同主干网络上性能,除了 AlexNet,本文还进行了更先进的骨干网络的实验,包括 ResNet-50、ViT-B-16 和 CLIP (Vit-B-32)的视觉编码器.对于 ResNet-50 和 ViT-B-16,我们在训练过程中进行了端到端微调,以便更好地适配当前任务；而对于 CLIP,我们采用了部分参数冻结的方式,仅微调其投影层,以保留其语义知识.表 6 和表 7 分别展示了 mAP@5000 和 mAP@all 的性能指标. 结果表明,本文方法在更强主干网络的支持下取得了进一步性能提升,同时也验证了所提方法在不同视觉编码器下的通用性与鲁棒性.

表 6 CTRH 在两个数据集上使用不同主干网络的 mAP@all(%)结果,最优结果已加粗表示

方法	主干网络	MirFlickr				NUS-WIDE			
		16bits	32bits	48bits	64bits	16bits	32bits	48bits	64bits
CTRH	AlexNet	72.21	74.03	75.38	75.37	**67.79**	67.63	68.49	**68.96**
CTRH	ResNet-50	74.11	76.73	78.09	77.51	66.02	65.44	67.49	68.34
CTRH	ViT-B-16	76.60	77.13	77.69	77.81	61.81	65.56	66.93	68.14
CTRH	CLIP(ViT-B-32)	**79.79**	**78.06**	**78.96**	**80.82**	67.57	**69.91**	**69.41**	68.35

表 7 CTRH 在两个数据集上使用不同主干网络的 mAP@5000(%)结果,最优结果已加粗表示

方法	主干网络	MirFlickr				NUS-WIDE			
		16bits	32bits	48bits	64bits	16bits	32bits	48bits	64bits
CTRH	AlexNet	78.80	82.66	83.33	82.98	78.19	78.50	78.98	80.61
CTRH	ResNet-50	84.50	86.30	87.89	87.85	78.46	79.91	80.81	81.75
CTRH	ViT-B-16	87.97	**88.28**	**88.38**	88.02	77.98	80.63	81.70	82.13
CTRH	CLIP(ViT-B-32)	**88.17**	86.76	87.34	**89.26**	**80.50**	**81.54**	**81.86**	**82.50**

为了分析 CTRH 在不同预训练多模态基础模型上的性能,除了 CLIP,本文还尝试使用 OpenCLIP,在 MirFlickr 和 NUS-WIDE 数据集上进行实验.实验结果如表 8 和表 9 所实验结果表明,CLIP 在不同哈希码长度下依然取得更优的 Map 性能,整体表现优于 OpenCLIP,展现出更强的稳定性和鲁棒性.我们认为,这一现象的主要原因在于：CTRH 所面向的是弱监督多标签图文哈希检索场景,其中标签信息稀疏、语义噪声大,这可能使得 OpenCLIP 在大规模语义分布预训练中学到的泛化特征未能充分发挥优势.相比之下,CLIP 在图文对齐上的特性更稳定、对语义噪声更具鲁棒性,因此在此类任务中仍具有优势.综上所述,采用 CLIP 模型是一个合理且有效的选择,能够很好支撑本文所提出的哈希学习框架.

表 8 CTRH 在两个数据集上使用不同预训练多模态基础模型的 mAP@all(%)结果,最优结果已加粗表示

方法	VL 模型	MirFlickr				NUS-WIDE			
		16bits	32bits	48bits	64bits	16bits	32bits	48bits	64bits
CTRH	CLIP	**72.21**	**74.03**	**75.38**	**75.37**	**67.79**	**67.63**	**68.49**	**68.96**
CTRH	OpenCLIP	70.60	72.32	72.77	73.17	64.25	66.17	66.71	66.64

表 9 CTRH 在两个数据集上使用不同预训练多模态基础模型的 mAP@5000(%)结果,最优结果已加粗表示

方法	VL 模型	MirFlickr				NUS-WIDE			
		16bits	32bits	48bits	64bits	16bits	32bits	48bits	64bits
CTRH	CLIP	**78.80**	**82.66**	**83.33**	**82.98**	**78.19**	**78.50**	**78.98**	**80.61**
CTRH	OpenCLIP	77.25	79.22	80.03	80.34	73.11	75.42	76.71	77.27

4 总 结

本文提出了一种 CLIP 引导标签优化的弱监督哈希方法,通过挖掘 CLIP 蕴含的丰富的多模态知识对用户标签进行优化,从而增强图像与文本的联合表示,并进一步地引导哈希模型生成具有高判别性的哈希码.为缓解噪声标签对模型性能的影响,本文设计了一个弱监督标签置换模块,通过微调 CLIP 并融合预训练参数来进行推理,生成和图像内容一致的文本标签.进一步地,本文引入了弱监督标签赋权模块,利用多头自注意力机制实现优化标签和图像的全局交互,有效提升联合表示的表征能力. 针对文本标签分布不均的问题,本文设计了标签

平衡损失,通过动态加权策略加强对尾类样本的学习.最后,本文联合引入相似性保留损失,铰链损失和量化损失,进一步引导哈希码学习联合特征蕴含的语义知识.在 MirFlickr 和 NUS-WIDE 上的实验结果验证了本文所提方法的有效性.

References:

[1] Huang XY, Sun B, Yang ZY, Zhu YY, Tian Q. 2021. Locality-sensitive hashing approach based on semantic space for visual retrieval. Journal of Image and Graphics,26(07):1568-1582

[2] Li ZX, Ling F, Tang ZJ, Ma HF, Shi ZP. Unsupervised cross-media Hashing retrieval based on multi-head attention network (in Chinese). Sci Sin Inform, 2021, 51: 2053–2068.

[3] Xia RK, Pan Y, Lai HJ, Liu C, Yan SC. Supervised Hashing for Image Retrieval via Image Representation Learning. In: AAAI conference on artificial intelligence. 2014.28(1): 2156-2162.

[4] Hussain A, Li HC, Ali D, Ali M, Abbas F, Hussain M. An optimized deep supervised hashing model for fast image retrieval. Image and vision computing, 2023. 104668.

[5] Yang F, Ding X, Liu YF, Ma FM, Cao J. Scalable semantic-enhanced supervised hashing for cross-modal retrieval. Knowl. Based Syst.2022, 251:109176.

[6] Shi NF, Fu C, Tie M, Zhang WC, Wang XW, Sham CW. Attention-based deep supervised hashing for near duplicate video retrieval. Neural Computing and Applications, 2024, 361: 5217-5230.

[7] Karaman S, Lin XD, Hu XF, Chang, SF. Unsupervised Rank-Preserving Hashing for Large-Scale Image Retrieval. In: International Conference on Multimedia Retrieval. ACM, 2019: 192-196.

[8] Xiong S, Pan, L, Ma X, Hu Q, Beckman E. Unsupervised deep hashing with multiple similarity preservation for cross-modal image-text retrieval. International Journal of Machine Learning and Cybernetics .2024, 15(10): 4423-4434.

[9] Yang Z, Deng X, Long J. Fast unsupervised consistent and modality-specific hashing for multimedia retrieval. Neural Computing & Applications, 2023, 35(8): 6207-6223.

[10] Venkateswara H, Eusebio J, Chakraborty S, Panchanathan, S. Deep Hashing Network for Unsupervised Domain Adaptation. In: IEEE conference on computer vision and pattern recognition. 2017: 5018-5027.

[11] Gattupalli V, Zhuo Y, Li B. Weakly supervised deep image hashing through tag embeddings. In: IEEE Conference on Computer Vision and Pattern Recognition. 2019: 10375-10384.

[12] Jin L, Li ZC, P YH, Tang JH. Weakly-Supervised Image Hashing through Masked Visual-Semantic Graph-based Reasoning. In: 28th ACM International Conference on Multimedia. 2020: 916-924.

[13] Wang M, Zhou WG, Tian Q, Li HQ. Deep Enhanced Weakly-Supervised Hashing with Iterative Tag Refinement. IEEE Transactions on Multimedia, 2021, PP(99):1-1.

[14] Du YC, Wang M, Lu ZB, Zhou WG, Li HQ. Weakly Supervised Hashing with Reconstructive Cross-modal Attention. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19(6): 1-19.

[15] Kim WJ, Son BY, Kim I. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In: International conference on machine learning. PMLR, 2021: 5583-5594.

[16] Chen YC, Li LJ, Yu LC, Kholy AE, Ahmed F, Gan Z, Cheng Y, Liu JJ. UNITER: UNiversal Image-Text Representation Learning. In: European Conference on Computer Vision, Springer, Cham, 2020: 104-120.

[17] Li XJ, Yin X, Li CY, Zhang PC, Hu XW, Zhang L, Wang LJ, Hu HD, Dong L, Wei FR, Choi YJ, Gao JF. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In: European conference on computer vision. Cham: Springer International Publishing, 2020: 121-137.

[18] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning Transferable Visual Models From Natural Language Supervision. In: International conference on machine learning. PmLR, 2021: 8748-8763.

[19] Chatfield K, Simonyan K, Vedaldi A, Zisserman A. Return of the Devil in the Details: Delving Deep into Convolutional Nets. In: Proceedings of the British Machine Vision Conference, 2014: 1-12

[20] Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781, 2013.

[21] Yin J, Zhang ZD, Gao YH, Yang ZW, Li L, Xiao M, Sun YQ, Yan CG. Survey on Vision-language Pre-training. Journal of Software, 2023, 34(5): 2000-2023 (in Chinese).

[22] Zhang HY, Wang TB, Li MZ, Zhao Z, Pu SL, Wu F. Comprehensive review of visual-language-oriented multimodal pre-training methods. Journal of image and graphics, 2022, 27(9): 2652-2682.

[23] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention Is All You Need. In: Advances in Neural Information Processing Systems,2017,30(1):6000-6010.

[24] Huiskes MJ, Lew MS. The MIR flickr retrieval evaluation. In: 1st ACM international conference on Multimedia information retrieval. 2008: 39-43

[25] Chua TS, Tang JH, Hong RC, Li HJ, Luo ZP, Zheng YT. NUS-WIDE: A real-world web image database from National University of Singapore. In: ACM international conference on image and video retrieval. 2009: 1-9.

[26] Indyk P, Motwani R. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In: Thirtieth Annual ACM Symposium on Theory of Computing. 1998: 604-613.

[27] Weiss Y, Torralba A, Fergus R. Spectral Hashing. In: Advances in Neural Information Processing Systems, 2008, 21.

[28] Gong YC, Lazebnik S, Gordo A, Perronnin F. Iterative Quantization: A Procrustean Approach to Learning Binary Codes for Large-scale Image Retrieval. IEEE Transactions on Pattern Analysis and Machine Lntelligence, 2012, 35(12): 2916-2929.

附中文参考文献:

[1] 黄小燕, 孙彬, 杨展源, 朱映映, 田奇. 面向视觉搜索的空间局部敏感哈希方法. 中国图象图形学报, 2021, 26(07):1568-1582.

[2] 李志欣, 凌锋, 唐振军, 马慧芳, 施智平, 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, 52(12): 2053-2068.

[21] 殷炯, 张哲东, 高宇涵, 杨智文, 李亮, 肖芒, 孙垚棋, 颜成钢, 视觉语言预训练综述. 软件学报, 2022, 34(5): 1-24.

[22] 张浩宇, 王天保, 李孟择, 赵洲, 浦世亮, 吴飞, 视觉语言多模态预训练综述. 中国图象图形学报, 2022, 27(9): 2652-2682.