

基于细粒度语义推理的跨媒体双路对抗哈希学习模型

曹晓雯, 梁美玉, 鲁康康

引用本文

曹晓雯, 梁美玉, 鲁康康. [基于细粒度语义推理的跨媒体双路对抗哈希学习模型](#)[J]. 计算机科学, 2022, 49(9): 123-131.

CAO Xiao-wen, LIANG Mei-yu, LU Kang-kang. [Fine-grained Semantic Reasoning Based Cross-media Dual-way Adversarial Hashing Learning Model](#)[J]. Computer Science, 2022, 49(9): 123-131.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[蒙汉神经机器翻译研究综述](#)

Survey of Mongolian-Chinese Neural Machine Translation

计算机科学, 2022, 49(1): 31-40. <https://doi.org/10.11896/jsjcx.210900006>

[融合深度典型相关分析和对抗学习的跨模态检索](#)

Cross-modal Retrieval Combining Deep Canonical Correlation Analysis and Adversarial Learning

计算机科学, 2021, 48(9): 200-207. <https://doi.org/10.11896/jsjcx.200600119>

[基于细粒度差异特征的文本匹配方法](#)

Text Matching Method Based on Fine-grained Difference Features

计算机科学, 2021, 48(8): 60-65. <https://doi.org/10.11896/jsjcx.200700008>

[基于生成对抗网络的图像修复](#)

Image Inpainting Based on Generative Adversarial Networks

计算机科学, 2018, 45(12): 229-234. <https://doi.org/10.11896/j.issn.1002-137X.2018.12.038>

[一种基于 Spark 的大规模语义数据分布式推理框架](#)

Spark Based Large-scale Semantic Data Distributed Reasoning Framework

计算机科学, 2016, 43(Z11): 93-96. <https://doi.org/10.11896/j.issn.1002-137X.2016.11A.020>

基于细粒度语义推理的跨媒体双路对抗哈希学习模型

曹晓雯 梁美玉 鲁康康

北京邮电大学计算机学院(国家示范性软件学院)智能通信软件与多媒体北京市重点实验室 北京 100876

(xwcao@bupt.edu.cn)

摘要 跨媒体哈希因其优越的搜索效率和较低的存储成本而在跨媒体搜索任务中受到广泛关注。然而,现有方法无法充分保持多模态数据的高阶语义相关性和多标签语义信息,从而导致学习到的哈希编码的质量下降。为了解决上述问题,提出了基于细粒度语义推理的跨媒体双路对抗哈希(Semantic Reasoning Based Cross-media Dual-way Adversarial Hashing Learning Model,SDAH)学习模型,通过最大程度地挖掘不同模态间的细粒度语义关联,产生紧凑且一致的跨媒体统一高效哈希语义表示。首先,提出了基于跨媒体协同注意力机制的细粒度跨媒体语义关联学习和推理方法,基于跨媒体注意力机制协同学习图像和文本的细粒度隐含语义关联,获取图像和文本的显著性语义推理特征;然后,建立了跨媒体双路对抗哈希网络,通过联合学习模态内和模态间的语义相似性约束,并通过双路对抗学习机制更好地对齐不同模态哈希码的语义分布,产生更高质量和更具判别性的跨媒体统一哈希表示,促进了跨媒体语义融合,提升了跨媒体搜索性能。在两个公开数据集上与现有方法的对比实验结果验证了所提方法在各种跨媒体搜索场景下的优越性能。

关键词: 语义推理;哈希学习;跨媒体搜索;对抗学习;跨媒体语义融合

中图法分类号 TP391

Fine-grained Semantic Reasoning Based Cross-media Dual-way Adversarial Hashing Learning Model

CAO Xiao-wen, LIANG Mei-yu and LU Kang-kang

Beijing Key Laboratory of Intelligent Communication Software and Multimedia, School of Computer(National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract Cross-media hashing has received extensive attention in cross-media searching tasks due to its superior searching efficiency and low storage cost. However, existing methods cannot adequately preserve the high-level semantic relevance and multi-label of multi-media data. In order to solve the above problems, this paper proposes a fine-grained semantic reasoning based cross-media dual-way adversarial hashing learning model(SDAH), which generates compact and consistent cross-media unified efficient hash semantic representations by maximizing fine-grained semantic associations between different medias. First, a fine-grained cross-media semantic association learning and inference method based on the cross-media collaborative attention mechanism is proposed. The cross-media attention mechanism collaboratively learns the fine-grained implicit semantic associations of images and texts, and obtains the salient semantic inference features of images and texts. Then, a cross-media dual-way adversarial hashing network is established to jointly learn the intra-modality and inter-modality semantic similarity constraints, and better to align the semantic distributions of different media hash codes through a two-way adversarial learning mechanism, which generates higher-quality and more discriminative cross-media unified hash representation, facilitates the process of cross-media semantic fusion and improves the cross-media searching performance. Experimental results compared with existing methods on two public datasets verify the performance superiority of the proposed method in various cross-media search scenarios.

Keywords Semantic reasoning, Hash learning, Cross-media search, Adversarial learning, Cross-media semantic fusion

随着社交网络和搜索引擎中多媒体数据的快速增长,不同媒体类型(如图像或文本)信息之间的搜索成为了一大研究热点。例如,当用户使用一种模态,如文本,去搜索感其兴趣的话题时,通常会希望得到与此话题相关的其他模态,如图像

或视频等信息的反馈。然而,不同模态的结构分布和语义表示是不一致的,它们之间存在异质语义鸿沟,这也是跨媒体搜索面临的重大挑战。跨媒体哈希技术为实现高效跨媒体大数据搜索提供了有效的解决方案,通过对不同模态的高维数据

到稿日期:2022-06-02 返修日期:2022-07-05

基金项目:国家自然科学基金(61877006,62192784);中国人工智能学会-华为 MindSpore 学术奖励基金(CAAIXSJLJJ-2021-007B)

This work was supported by the National Natural Science Foundation of China(61877006,62192784) and CAAI-Huawei MindSpore Open Fund(CAAIXSJLJJ-2021-007B).

通信作者:梁美玉(meiyu1210@bupt.edu.cn)

信息学习紧凑的二进制哈希表示,可进行高效语义匹配和搜索,拥有存储消耗低、搜索速度快的优点。早期的哈希学习方法大多是在手工制作的特征基础上,通过学习线性函数生成哈希码,这些方法虽然易于实现,但是随着多媒体数据越来越复杂多样,利用这些方法对跨媒体数据的复杂语义相关性进行建模也变得愈加困难,从而影响跨媒体搜索的性能。随着深度学习的快速发展和成功应用,深度神经网络凭借其强大的非线性特征学习能力被成功应用在跨媒体哈希学习方法中,并取得了优越的性能。但是这些跨媒体哈希学习方法仍然存在一些局限性。首先,这些方法通常都由特征学习部分和哈希码生成部分组成,二者对于最终生成的哈希码质量都至关重要,然而现有的大多数方法主要集中在哈希码生成部分,忽略了特征学习部分的关键性,对不同模态所包含的关键语义信息没有进行全面充分的挖掘;其次,一个模态中的数据点在该模态中存在着多个语义相关或者语义不相关的信息点,其中语义相关的信息点和给定的数据点应该具有更高的语义特征相似度;同时,现有的大型数据集往往都伴随着多标签语义结构。因此,不仅要尽可能地保持不同模态数据的语义结构信息,还要借助多标签语义信息生成

更具判别性、更高质量的哈希码。

为了解决上述问题,本文构建了基于细粒度语义推理的跨媒体双路对抗哈希学习模型(SDAH),模型框架如图 1 所示。该模型分为两个部分,即跨媒体细粒度语义关联学习和融合网络、融合网络和跨媒体对抗哈希学习网络。在跨媒体细粒度语义关联学习和融合网络中,对于图像模态,首先对图像基于 Faster-RCNN 技术进行区域特征学习,经过连接层堆栈的处理后输入到语义推理模块,对图像的区域特征向量序列进行高级语义推理,最大程度地挖掘图像区域之间的语义关联;其次,对于文本模态,为了获取文本模态紧凑、丰富的语义特征向量,基于 BERT 深度语言模型对文本集合进行语义推导,随后输入到语义推理模块;最后通过全连接层的映射,将图像和文本模态的特征向量映射到统一的语义空间,使用 TE_s (Transformer Encoders)加速跨媒体语义融合过程,最终得到经过细粒度语义推理的多模态特征向量。在跨媒体对抗哈希学习网络中,将特征学习模块得到的语义特征向量作为哈希学习层的输入,在跨媒体对抗学习、模态内以及模态间的语义相似性约束下,生成更具判别性和更高质量的哈希码,从而充分保持哈希码的多标签语义相似性和跨媒体语义不变性。

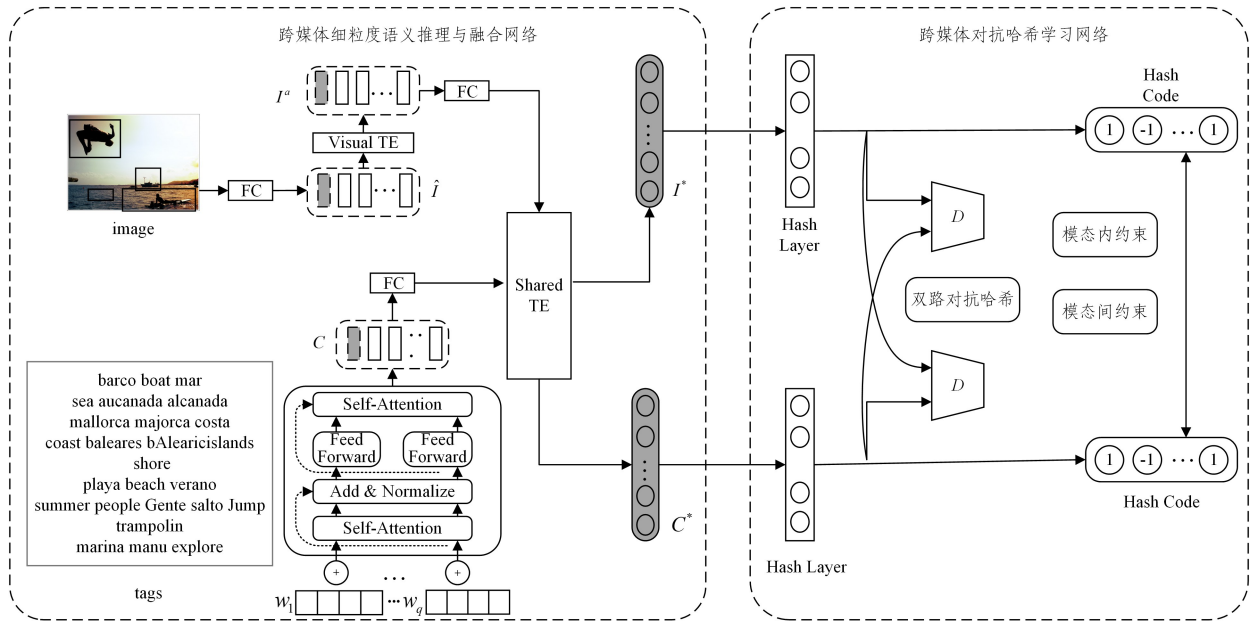


图 1 提出的 SDAH 模型的框架图

Fig. 1 Framework of the proposed SDAH model

本文的主要贡献包括以下 4 个方面:

- (1) 构建了基于细粒度语义推理的跨媒体对抗哈希学习模型,通过将跨媒体细粒度语义推理学习和对抗哈希学习集成在一个统一的端到端框架下来联合学习跨媒体语义相似性,获取跨媒体统一高效哈希语义表示,促进跨媒体语义融合,提升跨媒体搜索的性能。
- (2) 提出了基于跨媒体深度注意力机制的细粒度语义推理和融合方法,结合语义共享与融合的跨媒体注意力机制学习图像、文本的细粒度语义信息,获取图像和文本的显著性语义推理特征。
- (3) 建立了跨媒体双路对抗哈希网络,通过联合学习模态内和模态间的语义相似性约束,并通过双路对抗学习机制

更好地对齐不同模态哈希码的语义分布,生成更高质量和更具判别性的跨媒体统一哈希表示。

- (4) 在两个广泛使用的跨媒体公开数据集上对所提模型进行了评估,实验结果表明所提模型优于几种最先进的方法;此外,通过消融实验验证了所提细粒度语义推理和双路对抗哈希学习机制可以显著提高跨媒体语义表示学习能力和搜索性能。

1 相关工作

跨媒体哈希方法大致可以分为无监督方法^[1-4]和监督方法^[5-10]。无监督哈希通过挖掘训练样本的模态内和模态间语义关联关系来学习跨媒体统一哈希表示。例如,文献[2]将

图像和文本嵌入到潜在特征空间中,通过计算余弦相似度来捕捉图像和文本之间的内在相关性。文献[3]提出了一个跨图注意力模型来显式地学习共享的语义概念,可以很好地利用它来指导每个模态的特征学习过程,并促进共同的嵌入学习。监督方法利用标签信息或相似度矩阵来获得更具辨别力的哈希函数。文献[5]的方法(SCM)将语义标签集成到哈希学习过程中,从而得到带有标签向量的相似度矩阵。文献[6]提出了一种语义保持的哈希方法(SePH),通过最小化 KL 散度来学习二进制代码。由于大多数跨媒体哈希方法都是基于浅层架构构建的,因此它们无法描述跨媒体的复杂相关性。

最近用于跨媒体哈希的深度学习方^[11-15]在保持不同模态之间的相似性方面表现出了更好的性能。深度跨媒体哈希方法(DCMH)^[11]在端到端深度学习框架中同时执行特征学习和哈希码学习。文献[12]采用对抗学习引导的多标签注意力模块来增强特征学习部分,并且提出了利用多标签二进制码映射的非对称哈希方法(AGAH)。文献[13]提出了一种自约束和基于注意力的哈希网络(SCAHN),将早期和后期的标签约束及其融合特征集成到哈希表示学习中。尽管这些方法已经实现了较好的搜索性能,但它们大多不能保持所有图像-文本对之间的更高级别的语义相关性。在深度跨媒体哈希方法中,生成对抗网络(Generative Adversarial Network, GAN)^[16]在跨媒体搜索方面也取得了显著进展。对抗性跨媒体搜索(ACMR)^[17]是第一个将 GAN 应用于跨媒体哈希的实值方法。自监督对抗哈希(SSAH)^[18]方法将多标签信息作为网络输入并学习标签表示以最大化不同模态之间的语义相关性。其他使用对抗学习的方法,如文献[19-21],也取得了可观的结果。但是这些方法都忽视了特征学习的重要性,仍然使用传统的手工特征提取方法,丢失了大量的多模态关键语义信息。

在多模态特征学习方面,近年来部分研究人员专注于研究用于高级理解的关系推理模型。文献[22]提出了一种将感知与推理分开的架构。他们引入被称为 Relation Network (RN)的特定层来解决视觉问答问题,该层专门用于比较实体对,使用 4 层卷积神经网络学习对象表示,并通过长短时记忆网络(LSTM)生成问题嵌入。最近,文献[23-24]扩展了 RN,以生成用于关系感知图像搜索的紧凑特征。然而,上述研究工作并没有涉及跨媒体学习领域。其他解决方案试图更多地使用象征性的推理方式,如文献[25-26]引入了组合方法,能够通过动态构建推理图来显式建模推理过程。最近的工作使用图卷积网络(Graph Convolutional Network, GCN)来推理概念之间的语义关联,如文献[27]使用 GCN 对图像区域进行推理以进行图像描述,而文献[28-29]使用基于注意力机制的 GCN 从普通图像生成场景图。同时,文献[27-29]使用了基于细粒度的特征学习方法,尽管取得了优秀的跨媒体搜索性能,但是缺乏对哈希检索的研究,并且以牺牲时间为代价。因此,本文构建了基于细粒度语义推理的跨媒体双路对抗哈希学习(SDAH)模型,在充分利用多标签语义信息的同时,实现了跨媒体语义融合,并在此基础上利用双路对抗学习机制生成更具判别性和更高质量的跨媒体统一哈希表示,提升了跨媒体搜索的综合性能。

2 基于细粒度语义推理的跨媒体双路对抗哈希学习模型

本节首先介绍本文模型的问题定义和符号说明,然后具体介绍本文模型中各个模块的工作原理,最后给出模型涉及的算法优化过程。

2.1 问题定义

假设给定训练集 S 由 n 个样本对组成,即 $S = \{v_i, t_i\}_{i=1}^n$, 其中 v_i 和 t_i 分别是第 i 个样本对中的图像和文本, d 和 f 分别为图像特征向量和文本特征向量的维数。语义标签矩阵定义为 $L = \{l_i\}_{i=1}^n$, $l_i = [l_{i1}, \dots, l_{ic}]$, c 是训练集中的语义类别数目,如果第 i 个样本对属于第 j 个类别,那么 $l_{ij} = 1$,反之 $l_{ij} = 0$ 。本文模型通过学习两个哈希函数 $h^v(v)$ 和 $h^t(t)$, 分别为图像模态和文本模态生成哈希码 $b^v \in \{-1, +1\}^K$ 和 $b^t \in \{-1, +1\}^K$, K 代表哈希码的长度,以此保持模态内和模态间的相似性。在接下来的部分, $\text{sign}(\cdot)$ 是元素符号函数,定义如下:

$$\text{sign}(x) = \begin{cases} +1, & x > 0 \\ -1, & x \leq 0 \end{cases}$$

2.2 基于跨媒体协同注意力机制的细粒度语义推理和融合网络

2.2.1 多模态特征学习

特征学习模块分为图像特征学习和文本特征学习。对于图像特征学习,本文采用 Faster-RCNN 网络^[30]对图像 v_i 进行区域特征学习,采用从每个图像中提取的自下而上的特征作为图像特征表示,记为 $I_i = \{r_i^0, r_i^1, \dots, r_i^p\}$, p 表示图像区域的个数, r_i^0 为零向量,用于后期的语义推理。为了在后续的图像视觉推理过程中包含空间感知要素,本文使用边界框坐标矩阵协同调节视觉推理的早期处理过程。首先计算图像中每个区域的归一化坐标和归一化面积,如式(1)所示:

$$c = \left\{ \frac{x_1}{W}, \frac{y_1}{W}, \frac{x_2}{H}, \frac{y_2}{H}, \frac{(x_2 - x_1)(y_2 - y_1)}{WH} \right\} \quad (1)$$

其中, W 和 H 分别表示各个边界框的宽和高, x_1, x_2, y_1, y_2 是边界框的坐标值。然后将 c 与特征 I 连接,最后通过 Linear-ReLU-Linear 全连接层映射,从而获得包含空间感知的自下而上的视觉特征 $\hat{I}_i = \{\widehat{r_i^0}, \widehat{r_i^1}, \dots, \widehat{r_i^p}\}$ 。

对于文本特征学习,本文使用 BERT 模型^[31]提取文本 t_i 的词嵌入表示,基于多层 Transformer 编码器处理句子中的单词,并通过强大的自注意力机制挖掘单词之间的语义关系,最终得到文本词向量,表示为: $C_i = \{w_i^0, w_i^1, \dots, w_i^q\}$, 其中 q 是每张图片对应标题中单词的个数, w_i^0 包含了 q 个单词之间的语义信息。

2.2.2 跨媒体细粒度语义推理

本文使用 Transformer Encoder(TE)^[32]对包含空间感知的图像特征 $\hat{I}$ 和基于自注意力的文本特征 C 进行细粒度语义推理。以图像模态的非共享 TE 推理模块,即图中的 visual TE 为例,基于 TE 的工作原理,说明细粒度语义推理的过程。受文献[33]的启发,以单个图像的空间感知特征 $\hat{I} = \{\hat{r}_0, \dots, \hat{r}_p\}$ 为例,

$\hat{r}_1, \dots, \hat{r}_p$ 作为 TE 的输入向量,通过式(2)生成注意力矩阵:

$$\begin{aligned} Att(Q, K, V) &= \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \\ Q &= W^Q I, W^Q \in \mathbb{R}^{d_i \times d_i} \\ K &= W^K I, W^K \in \mathbb{R}^{d_k \times d_i} \\ V &= W^V I, W^V \in \mathbb{R}^{d_v \times d_i} \end{aligned} \quad (2)$$

其中, $d_i = p + 1$, $\sqrt{d_k}$ 用于缓解 softmax 函数的梯度消失问题,以防内积产生更大的值。在本模型中, $d_k = d_v$ = 每个区域特征的向量维数。 W^Q, W^K 和 W^V 均为在训练模型过程中生成的参数矩阵。首先对生成的注意力矩阵进行归一化,然后通过前馈神经网络 FFN,并采用 ReLU 函数激活,最后再次进行归一化处理,从而生成自注意力感知的图像特征 I^a , I^a 中的每一个向量都保存了 p 个区域的语义信息。

为了后续哈希码生成的一致性和连续性,将自注意力感知的图像特征 I^a 和基于自注意力的文本特征 C 使用全连接层映射到公共语义空间,随后使用共享的 TE 堆栈层加强不同模态信息的语义约束,为后续的双路对抗哈希学习提供高质量的语义推理特征。在每一次推理步骤中, p 个图像区域的特征信息都会通过 TE 的自注意力机制更新。因此,我们用最后一个 TE 层输出中的 r_0 和 w_0 分别代表图像模态和文本模态的语义推理特征,分别记为 I^* 和 C^* 。

2.3 跨媒体双路对抗哈希学习网络

在跨媒体特征学习模块的基础上,本模块通过双路对抗哈希学习、模态间语义相似性以及模态内语义一致性约束来指导生成更具判别性和更高质量的哈希码。哈希学习网络的图像输入和文本输入分别是图像推理特征 I^* 和文本推理特征 C^* ,然后经过 tanh 函数激活的图像和文本哈希层,得到图像和文本模态的输出,分别记为: $X = h^v(I^*; \theta_{I_H}^v) = \{x_i\}_{i=1}^n$ 和 $Y = h^l(C^*; \theta_{I_H}^l) = \{y_i\}_{i=1}^n, x_i, y_i \in \mathbb{R}^k$ 。通过 sign 函数获得最终的二进制代码, $b^v = \text{sign}(x)$ 和 $b^l = \text{sign}(y)$ 。

2.3.1 双路对抗哈希学习

为了生成跨媒体的模态不变表示,本文首先为每个模态设计两个判别器 D_v 和 D_t 。对于图像模态判别器 D_v ,将文本网络作为图像哈希码的生成器,从图像特征进行哈希映射得到的哈希码被认为是真实的图像哈希码,而从文本特征进行哈希映射得到的哈希码被标记为假图像哈希码。图像模态判别器的目标是区分输入的图像哈希码是否为真,文本模态判别器的作用与 D_v 类似。本文使用两个分别具有参数 $\theta_{D_v}^v$ 和 $\theta_{D_t}^v$ 的 3 层多层感知器 (Multiayer Perceptron, MLP) 来实现两个判别器。在双路对抗的博弈学习中,图像哈希网络尝试生成文本哈希码混淆文本模态判别器,文本哈希网络试图生成图像哈希码混淆图像模态判别器。将双路对抗损失标记为 L_{adv} , L_{adv} 是所有图像和文本实例的模态分类的交叉熵损失。双路对抗损失 L_{adv} 的表达式如式(3)所示:

$$\begin{aligned} L_{adv} &= L_{adv}^v + L_{adv}^t \\ &= -\frac{1}{n} \left[\sum_{i=1}^n \log(D_v(x_i; \theta_{D_v}^v)) + \log(1 - D_v(y_i; \theta_{D_v}^v)) + \right. \\ &\quad \left. \sum_{i=1}^n \log(D_t(y_i; \theta_{D_t}^v)) + \log(1 - D_t(x_i; \theta_{D_t}^v)) \right] \end{aligned} \quad (3)$$

其中, L_{adv}^* 表示图像或文本模态的对抗损失, $*$, $*$ $\in \{x, y\}$

表示第 i 个图像或文本实例的哈希码表示。

2.3.2 模态间语义相似性学习

以图像模态为例,三元组的形式为: (x_i, y_j^+, y_k^-) , 其中 y_j^+ 和 x_i 具有语义相关性, y_k^- 则相反。考虑到当每个数据点与多个类标签都具有语义相关性时,数据点之间的相似性关系和只有单个类标签的情况不同,为了确保多级语义结构相关性可以被最大程度地保持,受到基于语义排名的方法^[34]的启发,本文使用归一化折扣累积增益 (NDCG) 评估标准,其定义如式(4)所示:

$$NDCG@p = \frac{1}{Z} \sum_{i=1}^p \frac{2^{rel_i} - 1}{\log(1 + i)} \quad (4)$$

其中, p 是排名列表中的截断位置; Z 是确保正确排名的 NDCG 得分为 1 的归一化常数; rel_i 是排名列表中第 i 个数据点的相似度,可以根据从一种模态的查询 q 出发,在另一种模态查询到的 i 所共享的公共标签数目计算得到;排名列表可以在 Hamming 空间中计算。受文献[19]的启发,本文定义模态间语义相似性损失函数如式(5)所示:

$$\begin{aligned} L_{inter} &= \tilde{L}_{I \rightarrow T} + \tilde{L}_{T \rightarrow I} \\ &= \sum_{i,j,k} \omega(r_j, r_k) (\max(\cos(x_i, y_k^-)) - \cos((x_i, y_j^+) + \\ &\quad \text{margin}, 0)) + \sum_{i,j,k} \omega(r_j, r_k) (\max(\cos(x_i, y_k^-) - \\ &\quad \cos((x_i, y_j^+) + \text{margin}, 0)) \end{aligned} \quad (5)$$

其中, $\omega(r_j, r_k)$ 是权重因子,其值越高,查询点与来自另一个模态的数据点的语义相关性越高。

2.3.3 模态内语义一致性学习

由于二进制哈希量化过程中不可避免地会产生误差,为了尽量减少错误并防止信息丢失,使二进制哈希码尽可能地保持原始特征空间中的信息,设计模态内语义一致性损失函数如式(6)所示:

$$\min_{\theta_{I_H}^v, \theta_{I_H}^l} L_{intra} = \|b^v - x\|_F^2 + \|b^l - y\|_F^2 \quad (6)$$

其中, $b^v = \text{sign}(x)$, $b^l = \text{sign}(y)$, $\theta_{I_H}^v$ 和 $\theta_{I_H}^l$ 分别是图像哈希层和文本哈希层的网络参数。

2.4 目标函数及优化

本文提出的基于细粒度语义推理的跨媒体双路对抗哈希学习模型的总体目标函数如式(7)所示:

$$L = \alpha L_{inter} + \beta L_{intra} + \gamma L_{adv} \quad (7)$$

模型最终通过式(8)进行优化:

$$\min_{\theta, \theta_{D_v}^v, \theta_{D_t}^v, \theta_{I_H}^v, \theta_{I_H}^l} L = \alpha L_{inter} + \beta L_{intra} + \gamma L_{adv} \quad (8)$$

其中, θ 表示特征学习网络的所有参数。

所有网络参数都是基于反向传播 (BP) 算法学习的。首先,初始化所有超参数。然后训练对抗哈希网络:固定哈希层的参数,训练判别器;固定判别器的参数,训练哈希层,迭代直到模型收敛或达到最大 epoch 数。所提模型的优化学习过程如算法 1 所示。

算法 1 SDAH 模型训练过程的优化算法

输入:训练集 S , 标签集 L

输出:图像哈希码 B^v , 文本哈希码 B^l , 网络参数 $\theta, \theta_{D_v}^v, \theta_{D_t}^v, \theta_{I_H}^v, \theta_{I_H}^l$

1. Initialize parameters: Hyperparameters α, β, γ, m , learning rate: lr , batch size: bs , maximum epoch number: $Epoch$
2. repeat

3. for t iteration do;

4. Update θ_D^* by descending their gradients;

$$\theta_D^* \leftarrow \theta_D^* - \rho \nabla_{\theta_D^*} \frac{1}{n} L_{adv}, *, * \in \{v, t\}$$

5. Update θ_H^* by ascending their gradients;

$$\theta_H^* \leftarrow \theta_H^* + \rho \nabla_{\theta_H^*} \frac{1}{n} L_{adv}, *, * \in \{v, t\}$$

6. Update θ by BP algorithm;

$$\theta \leftarrow \theta - \rho \nabla_{\theta} (L - L_{adv})$$

7. end for

8. Update B^v, B^t

9. $e = e + 1$

10. until convergence or $e = E_{po}$

3 实验与结果

3.1 数据集

MIRFlickr25K 数据集^[35]: 包含从 Flickr 网站收集的 25 000 个图像-文本对,每个实例都由 24 个类别中的一个或多个标签进行注释。进行数据清洗后得到 22 371 个图像-文本对,从中随机选择 2 000 个图像-文本对作为查询集,其余的作为搜索数据库,在搜索数据库中随机选择 10 000 个样本实例进行训练。

NUSWIDE 数据集^[36]: 包含 269 648 个图像-文本对,每对都由 81 个概念的一个或多个标签进行注释。经过数据清洗之后,从中选择包含 21 个最常见概念的 193 622 个图像-文本对。随机选择 2 100 个图像-文本对作为查询集,其余的作为搜索数据库,然后从搜索数据库中随机选择 10 000 个样本实例进行训练。

3.2 实验设置

本文使用 HuggingFace^[37] 在英语句子的掩码语言任务上预训练的 BERT 模型,这些预训练的 BERT 嵌入矩阵都是 768 维。在提取图像特征时,本文使用文献[38]中的方法对 MIRFlickr25K 数据集和 NUSWIDE 数据集进行了特征提取,为每张图像选择前 36 个最显著的区域特征,得到的区域特征为 2 048 维。在图像特征的视觉先验推理过程中,本文使用 4 个非共享的 TE 堆栈层。在产生最终的细粒度语义推理特征之前,本文使用 2 个共享的 TE 堆栈层实现跨媒体语义融合,该堆栈的权重在图像和文本特征层之间共享。所有 TE 的前馈网络均为 2 048 维,dropout 设置为 0.1,在进入

最后一个共享权重的 TE 堆栈层之前,图像特征向量和文本特征向量均被线性投影至 1 024 维,这也是最终公共语义空间的特征向量维数。本文提出的 SDAH 模型在 MIR-Flickr25K 上的参数设置为: $\alpha=10, \beta=\gamma=1, lr=0.000\,03, margin=0.3$ 。在 NUSWIDE 上的参数设置为: $\alpha=10, \beta=\gamma=1, lr=0.000\,01, margin=0.3$ 。在两个数据集上的迭代次数均为 40 次。

3.3 评估标准

基于以上两个数据集,本文基于两种跨媒体搜索任务来验证提出的模型的性能:1) I2T(用图像搜索文本数据库); 2) T2I(用文本搜索图像数据库)。使用平均精度均值(Mean Average Precision, mAP)、准确度 precision@K,即仅考虑搜索出的前 K 个样本,以及精确率-召回率曲线(PR 曲线)对模型进行性能评估。

3.4 实验结果与分析

3.4.1 不同长度哈希码的跨媒体搜索性能比较

本文在上述两个数据集上,通过实验比较了本文提出的模型与语义保留哈希(Semantics Preserving Hashing, SePH)^[6]、具有语义相关性最大化的大规模监督多模态哈希(Large-Scale Supervised Multimodal Hashing with Semantic Correlation Maximization, SCM)^[5]、广义语义保留哈希(Generalized Semantic Preserving Hashing, GSPH)^[9]、深度跨媒体哈希(Deep Cross-Modal Hashing, DCMH)^[11]、自监督对抗哈希网络(Self-Supervised Adversarial Hashing Networks, SSAH)^[18]、对抗引导的非对称哈希(Adversary Guided Asymmetric Hashing, AGAH)^[12]、基于自约束和注意力的哈希网络(Self-Constraining and Attention-based Hashing Network, SCAHN)^[13]、基于多标签语义保持的深度跨媒体哈希(Multi-Label Semantics Preserving based Deep Cross-Modal Hashing, MLSPH)^[14] 8 个跨媒体哈希学习方法。

表 1 列出了本文提出的模型与 8 个基线方法在 MIR-Flickr25K 和 NUSWIDE 上的 mAP 分数,最高的 mAP 分数用加粗字体表示。可以看出,随着哈希码长度的增加,mAP 分数也越来越高;本文提出的模型在两个数据集上的两种跨媒体搜索任务都取得了最好的 mAP 分数。图 2 给出了哈希码长为 64 位时两个标准数据集在以上两种搜索任务的精确率-召回率(PR)曲线。可以观察到,SDAH 的表现更好,在所有召回级别上都获得了更高的准确率分数。

表 1 各方法在两个基准数据集上的 mAP 分数

Table 1 mAP scores for two cross-media retrieval tasks based on two benchmark datasets

(单位:bits)

Task	Method	MIRFlickr25K			NUSWIDE		
		16	32	64	16	32	64
I2T	SePH	0.711 0	0.709 0	0.716 9	0.653 8	0.667 8	0.672 0
	SCM	0.628 1	0.628 6	0.636 7	0.428 0	0.437 4	0.409 0
	GSPH	0.606 8	0.619 1	0.623 0	0.401 5	0.415 1	0.421 4
	DCMH	0.720 1	0.728 7	0.741 8	0.565 7	0.600 7	0.600 1
	SSAH	0.770 9	0.784 0	0.792 4	0.602 0	0.621 9	0.646 3
	AGAH	0.792 3	0.794 5	0.806 9	0.645 5	0.660 0	0.651 2
	SCAHN	0.816 8	0.830 1	0.832 8	0.642 9	0.651 0	0.663 5
	MLSPH	0.807 6	0.823 5	0.833 7	0.640 5	0.660 4	0.673 4
	SDAH	0.822 4	0.830 5	0.842 8	0.664 3	0.686 1	0.695 8

(续表)

Task	Method	MIRFlickr25K			NUSWIDE		
		16	32	64	16	32	64
T2I	SePH	0.7127	0.7261	0.7309	0.6600	0.6676	0.6854
	SCM	0.6068	0.6089	0.6108	0.4280	0.4374	0.4090
	GSPH	0.6282	0.6458	0.6503	0.4995	0.5233	0.5351
	DCMH	0.7553	0.7593	0.7730	0.5336	0.5871	0.5916
	SSAH	0.7681	0.7737	0.7838	0.6123	0.6369	0.6398
	AGAH	0.7887	0.7904	0.8049	0.6313	0.6422	0.6336
	SCAHN	0.8034	0.8105	0.8193	0.6501	0.6575	0.6685
	MLSPH	0.7852	0.8041	0.8146	0.6433	0.6633	0.6724
	SDAH	0.8128	0.8223	0.8322	0.6830	0.7053	0.7202

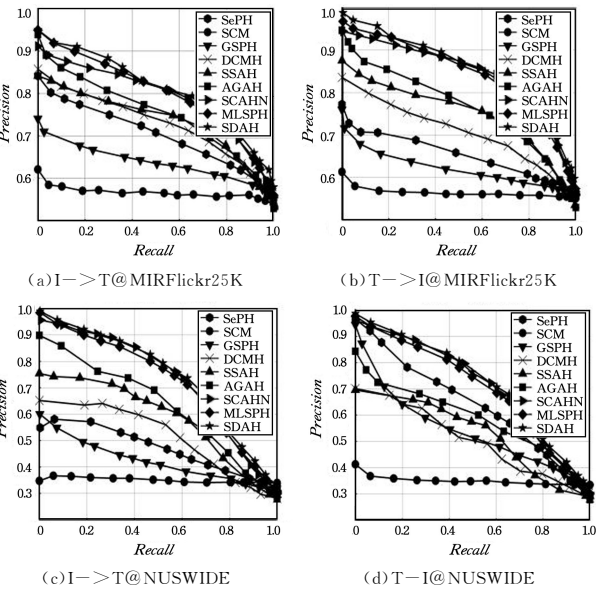


图2 两个数据集上的精确率-召回率曲线
Fig. 2 Precision-recall curves on two datasets

3.4.2 不同 Top-k 排序下的跨媒体搜索结果

本实验将 SDAH 与最先进的算法在不同 Top-k 的搜索结果下进行了比较,包括 1,100,200,300,500,600,700,800,900,1000。在不同的 Top-k 下,各基线方法在 MIRFlickr25K 上的两个跨媒体搜索任务(I2T 和 T2I)的 mAP 性能表现如图 3 所示。

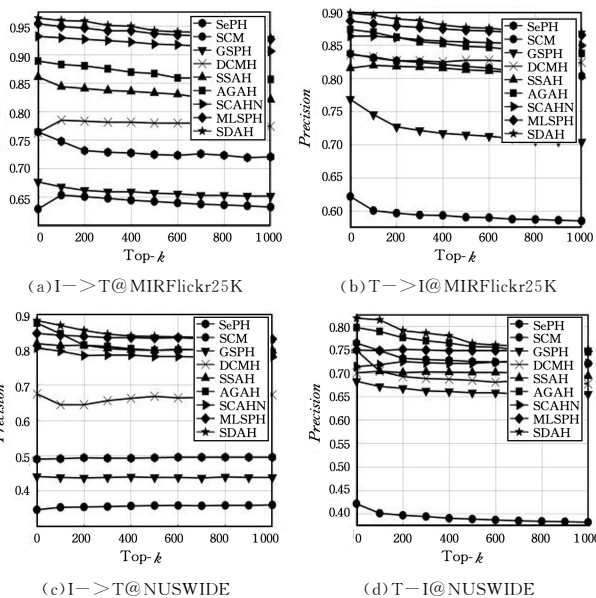


图3 两个数据集上的 Top-k 曲线
Fig. 3 Top-k precision curves on two datasets

从图中可以看出,在任意 k 值下 SDAH 方法在两个搜索任务上的表现都优于其他方法。当 k 值在 1~500 时 SDAH 方法具有非常明显的优势,这也说明了所提方法中引入 ND-CG 指标对最终的搜索性能是有帮助的,语义相关性最强的结果会优先展示。

3.5 模型消融性实验

为了证明本文提出组件的有效性,我们在哈希码长为 64 位的情况下设计了 5 个消融实验。

- (1)SDAH-v1:该对比实验去掉了用于公共语义空间推理的最后两层 TE 堆栈,即图 1 中的 visual TE,其余模块和组件与 SDAH 保持一致。
- (2)SDAH-v2:该对比实验去掉了用于图像区域间推理的非共享 TE 堆栈,即图 1 中的 shared TE,意味着经过图像特征提取后的特征向量直接进入公共投影层,其余模块和组件与 SDAH 相同。
- (3)SDAH-v3:该对比实验将 visual TE 和 shared TE 都用 GCN 推理机制进行了替换,其余部分和 SDAH 相同。
- (4)SDAH-v4:该对比实验去掉了文本单路对抗哈希,即去除了目标函数中的文本对抗哈希损失函数,其余模块和组件与 SDAH 相同。
- (5)SDAH-v5:该对比实验去掉了图像单路对抗哈希,即去除了目标函数中的图像对抗哈希损失函数,其余模块和组件与 SDAH 相同。

表 2 列出了消融实验的结果。

表 2 使用 64 位哈希码进行消融实验的 mAP 结果				
Method	MIRFlickr25k		NUSWIDE	
	I2T	T2I	I2T	T2I
SDAH-v1	0.8061	0.7934	0.6554	0.6705
SDAH-v2	0.8072	0.7910	0.6578	0.6697
SDAH-v3	0.8013	0.7897	0.6328	0.6177
SDAH-v4	0.8189	0.8005	0.6564	0.6702
SDAH-v5	0.8139	0.8009	0.6579	0.6703
SDAH	0.8224	0.8128	0.6643	0.6830

首先,通过 SDAH-v1 以及 SDAH-v2 两项实验可以看到,本文提出的细粒度语义推理方法对提高跨媒体搜索性能以及促进跨媒体语义融合有显著效果,这是因为 TE 堆栈在高级语义推理方面具有强大的效果,可以很好地挖掘到不同模态的细粒度特征,最大程度促进跨媒体语义融合过程,为后续生成哈希码提供了高质量的模态特征向量。此外,通过 SDAH-v3 这项实验可以看出,GCN 推理削弱了跨媒体检索性能。值得注意的是,GCN 推理还需要更多的时间成本,需要制作额外的邻接点文件,而 SDAH 则无须花费此成本。

其次,通过 SDAH-v4 以及 SDAH-v5 两项实验可以看出本文提出的双路对抗哈希对最终的跨媒体搜索结果均有益,去掉任何一路对抗哈希都会削弱最后的搜索性能。这是因为通过双路对抗哈希可以更好地对齐跨媒体哈希码,在前述推理特征向量的基础上可以产生更具判别性的哈希码,从而提升搜索性能。

此外,我们还通过迁移 SDAH 模型中的细粒度语义推理方法进一步证明了该模块的可移植性和可用性。将 SDAH

模型中的细粒度语义推理模块应用到 AGAH^[12] 和 DADH^[19] 的特征学习模块,哈希学习组件和原文保持一致,通过表 3 可以看出,该模块具有优秀的迁移能力,其中迁移后模型的性能结果用加粗的字体表示。可以看到,SDAH 方法中的细粒度语义推理机制在两个跨媒体搜索任务上都不同程度地提升了原模型的搜索性能,其中,在 NUSWIDE 数据集上的作用更显著,这也说明该模块对大型搜索集的跨媒体搜索有着更大的应用潜力。

表 3 SDAH 中细粒度语义推理模块的迁移能力评估
Table 3 Transferability evaluation of fine-grained semantic reasoning module in SDAH
(单位:bits)

Task	Method	MIRFlickr25K			NUSWIDE		
		16	32	64	16	32	64
I2T	AGAH	0.7923	0.7945	0.8069	0.6455	0.6600	0.6512
	AGAH *	0.8168	0.8305	0.8331	0.6566	0.6749	0.6859
	DADH	0.8020	0.8072	0.8179	0.6492	0.6662	0.6664
	DADH *	0.8064	0.8219	0.8339	0.6520	0.6801	0.6872
T2I	AGAH	0.7887	0.7904	0.8049	0.6313	0.6422	0.6336
	AGAH *	0.8087	0.8104	0.8189	0.6830	0.7053	0.7202
	DADH	0.7920	0.7959	0.8064	0.6501	0.6679	0.6808
	DADH *	0.7999	0.8162	0.8269	0.6688	0.6907	0.7114

3.6 参数分析和语义可视化实验

(1) 哈希码的语义可视化

为了说明 SDAH 模型学习到的哈希码具有有效性和语义一致性,我们将测试数据集按照 1:1 的比例随机划分(每个模态使用 1000 个样本点),然后使用 t-SNE 工具在哈希码长为 64 位的情况下,在 MirFlickr25k 数据集上可视化了 SDAH 模型生成的哈希码分布图,如图 4 所示,灰色点表示图像模态生成的哈希码,白色点表示文本模态生成的哈希码。可以看出,SDAH 模型不仅使两种模态的哈希码在模态内保持了很好的语义对齐,并且能有效地将样本点分为不同的语义判别簇,在模态间也实现了很好的语义对齐。

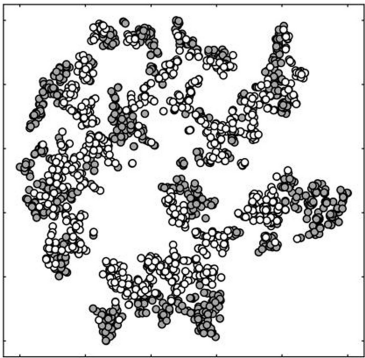


图 4 哈希码的语义可视化
Fig. 4 Semantic visualization of hash codes

(2) 三元组损失函数边际参数 $margin$ 的性能评估

本实验分析了超参数 $margin$ 对 SDAH 模型目标函数的影响。图 5 显示了在 64 位哈希码的情况下,不同的 $margin$ 值对模型在 MIRFlickr25K 数据集上 mAP 性能的影响。可以看出,模型的 mAP 随着 $margin$ 的增加而增加,当 $margin$ 值超过 0.3 时开始下降。当 $margin$ 值较小时,模型更难区分相似对和不相似对,因此 mAP 值相对较小。当 $margin$ 值为

0.3 时, mAP 在两个搜索任务中取得了最好的性能。因此,本文将所有实验 $margin$ 设置为 0.3。

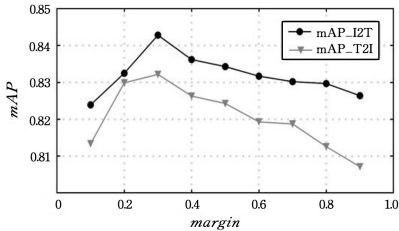


图 5 超参数 $margin$ 对跨媒体搜索性能的影响
Fig. 5 Influence of hyperparameter $margin$ on cross-media search performance

(3) 学习率对 SDAH 模型性能的影响

本实验分析了学习率 lr 对模型性能的影响。图 6 给出了当码长为 64 位时,不同 lr 对 SDAH 模型在 MIRFlickr25K 数据集上跨媒体搜索性能的影响。可以看出,当 $lr=0.00003$ 时,在 I2T 和 T2I 两个搜索任务中取得的 mAP 性能最高。因此,本实验设置 $lr=0.00003$ 。

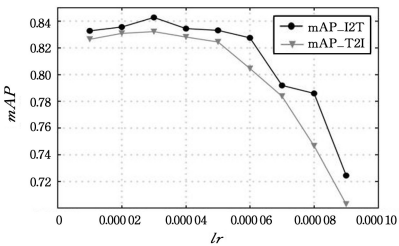


图 6 学习率 lr 对 SDAH 模型性能的影响
Fig. 6 Influence of learning rate lr on SDAH model performance

(4) 不同迭代次数下的跨媒体搜索性能分析

本实验分析了不同 $epoch$ 大小下的跨媒体搜索性能。在 MIRFlickr25K 数据集上进行了实验,实验结果如图 7 所示。可以看出,随着 $epoch$ 的增加,模型逐渐达到收敛。当 $epoch$

约为 20 时, mAP 变化减小; 当 $epoch$ 约为 30 时, 跨媒体搜索模型的收敛趋于稳定。

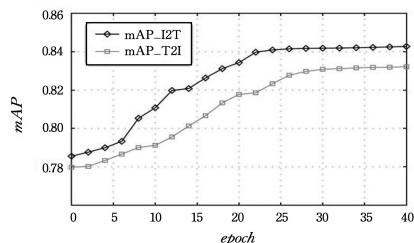


图 7 参数 $epoch$ 对模型性能的影响

Fig. 7 Influence of parameter $epoch$ on model performance

结束语 本文提出了一种基于细粒度语义推理的跨媒体双路对抗哈希学习(SDAH)模型, 该模型最大程度地挖掘了不同模态的细粒度语义关联信息, 促进了跨媒体语义融合, 产生了语义一致性和判别性更强的多模态语义特征信息; 提出了基于跨媒体协同注意力机制的细粒度跨媒体语义关联学习和推理方法, 学习图像和文本的细粒度隐含语义关联, 获取图像和文本的统一语义表示; 建立了跨媒体双路对抗哈希网络, 通过联合学习模态内和模态间的语义相似性约束, 并通过双路对抗学习机制更好地对齐不同模态哈希码的语义分布, 生成更高质量和更具判别性的跨媒体统一哈希表示, 提升了跨媒体搜索性能。在两个公开数据集上与现有方法的对比实验结果验证了所提方法在各种跨媒体搜索场景下的优越性能。

参考文献

- [1] LIU S, QIAN S S, GUAN Y, et al. Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval[C] // Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020;1379-1388.
- [2] ZHANG Y F, ZHOU W G, WANG M, et al. Deep relation embedding for cross-modal retrieval[J]. IEEE Transactions on Image Processing, 2020, 30: 617-627.
- [3] HE Y, LIU X, CHEUNG Y M, et al. Cross-Graph Attention Enhanced Multi-Modal Correlation Learning for Fine-Grained Image-Text Retrieval[C] // Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021;1865-1869.
- [4] ZHANG P F, DUAN J S, HUANG Z, et al. Joint-teaching: Learning to Refine Knowledge for Resource-constrained Unsupervised Cross-modal Retrieval[C] // Proceedings of the 29th ACM International Conference on Multimedia. 2021;1517-1525.
- [5] ZHANG D Q, LI W J. Large-scale supervised multimodal hashing with semantic correlation maximization[C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2014; 2177-2183.
- [6] LIN Z J, DING G G, HU M Q, et al. Semantics-preserving hashing for cross-view retrieval[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3864-3872.
- [7] LIU X B, NIE X S, SUN H L, et al. Modality-specific structure preserving hashing for cross-modal retrieval[C] // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2018;1678-1682.
- [8] LIANG M Y, DU J P, YANG C X, et al. Cross-Media Semantic Correlation Learning Based on Deep Hash Network and Semantic Expansion for Social Network Cross-Media Search[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(9): 3634-3648.
- [9] DEVRAJ M, KUNAL N C, SOMA B. Generalized semantic preserving hashing for cross-modal retrieval[J]. IEEE Transactions on Image Processing, 2018, 28(1): 102-112.
- [10] CHEN Z D, WANG Y X, LI H Q, et al. A two-step cross-modal hashing by exploiting label correlations and preserving similarity in both steps[C] // Proceedings of the 27th ACM International Conference on Multimedia. 2019;1694-1702.
- [11] JIANG Q Y, LI W J. Deep cross-modal hashing [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017;3232-3240.
- [12] GU W, GU X Y, GU J Z, et al. Adversary guided asymmetric hashing for cross-modal retrieval[C] // Proceedings of the 2019 International Conference on Multimedia Retrieval. 2019: 159-167.
- [13] WANG X Z, ZOU X T, BAKKER E M, et al. Self-constraining and attention-based hashing network for bit-scalable cross-modal retrieval[J]. Neurocomputing, 2020, 400: 255-271.
- [14] ZOU X T, WANG X Z, BAKKER E M, et al. Multi-label semantics preserving based deep cross-modal hashing[J]. Signal Processing: Image Communication, 2021, 93: 116131.
- [15] YAO H L, ZHAN Y W, CHEN Z D, et al. TEACH: Attention-Aware Deep Cross-Modal Hashing [C] // Proceedings of the 2021 International Conference on Multimedia Retrieval. 2021: 376-384.
- [16] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. arXiv:1406. 2661, 2014.
- [17] WANG B L, YANG Y, XU X, et al. Adversarial cross-modal retrieval[C] // Proceedings of the 25th ACM International Conference on Multimedia. 2017;154-162.
- [18] LI C, DENG C, LI N, et al. Self-supervised adversarial hashing networks for cross-modal retrieval[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018;4242-4251.
- [19] BAI C, ZENG C, MA Q, et al. Deep adversarial discrete hashing for cross-modal retrieval[C] // Proceedings of the 2020 International Conference on Multimedia Retrieval. 2020;525-531.
- [20] HAN L G, MIN M R, STATHOPOULOS A, et al. Dual projection generative adversarial networks for conditional image generation[C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021;14438-14447.
- [21] KARRAS T, AITTLA M, HELLSTEN J, et al. Training generative adversarial networks with limited data[J]. Advances in Neural Information Processing Systems, 2020, 33: 12104-12114.
- [22] SANTORO A, RAPOSO D, BARRETT D G, et al. A simple neural network module for relational reasoning[J]. arXiv:1706. 01427, 2017.
- [23] MESSINA N, AMATO G, CARRARA F, et al. Learning visual

- features for relational CBIR[J]. International Journal of Multimedia Information Retrieval, 2020, 9(2): 113-124.
- [24] MESSINA N, AMATO G, CARRARA F, et al. Learning relationship-aware visual features[C] // Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018: 486-501.
- [25] HU R H, ANDREAS J, ROHRBACH M, et al. Learning to reason; End-to-end module networks for visual question answering [C] // Proceedings of the IEEE International Conference on Computer Vision. 2017: 804-813.
- [26] ZHENG W F, LIU X J, NI X B, et al. Improving visual reasoning through semantic representation[J]. IEEE Access, 2021, 9: 91476-91486.
- [27] WANG J B, WANG W, WANG L, et al. Learning visual relationship and context-aware attention for image captioning[J]. Pattern Recognition, 2020, 98: 107075.
- [28] YANG L, HU H, LU X L, et al. Constrained lstm and residual attention for image captioning[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2020, 16(3): 1-18.
- [29] LI Y K, OUYANG W L, ZHOU B, et al. Factorizable net: an efficient subgraph-based framework for scene graph generation [C] // Proceedings of the European Conference on Computer Vision (ECCV). 2018: 335-351.
- [30] REN S Q, HE K M, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems, 2017, 39(6): 1137-1149.
- [31] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 4171-4186.
- [32] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. arXiv:1706.03762 2017.
- [33] MESSINA N, FALCHI F, ESULI A, et al. Transformer reasoning network for image-text matching and retrieval[C] // 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 5222-5229.
- [34] ZHAO F, HUANG Y, WANG L, et al. Deep semantic ranking based hashing for multi-label image retrieval[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1556-1564.
- [35] 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.
- [36] 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: 1-9.
- [37] WOLF T, DEBUT L, SANH V, et al. Huggingface's transformers: State-of-the-art natural language processing[J]. arXiv: 1910.03771, 2019.
- [38] ANDERSON P, HE X D, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6077-6086.



CAO Xiao-wen, born in 1998, master. Her main research interests include deep learning and cross-modal retrieval.



LIANG Mei-yu, born in 1985, associate professor, master supervisor. Her main research interests include artificial intelligence, data mining, multimedia information processing and computer vision.

(责任编辑:何杨)