计算机科学
COMPUTER SCIENCE

基于GAN的语义对齐网络半监督跨模态哈希方法

刘华咏, 朱婷

引用本文

刘华咏, 朱婷. 基于GAN的语义对齐网络半监督跨模态哈希方法[J]. 计算机科学, 2025, 52(6): 159-166.

LIU Huayong, ZHU Ting. Semi-supervised Cross-modal Hashing Method for Semantic Alignment Networks Basedon GAN [J]. Computer Science, 2025, 52(6): 159-166.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240400022

基于 GAN 的语义对齐网络半监督跨模态哈希方法

刘华咏　朱　婷
华中师范大学计算机学院　武汉 430079
(lhywuhee@ccnu.edu.cn)

摘　要　监督方法在跨模态检索中已有不少成果,是比较热门的方法。然而,这类方法过于依赖标记的数据,没有充分利用无标签数据所包含的丰富信息。为了解决这一问题,人们开始研究无监督方法,但是仅依靠未标记数据的效果并不理想。对此,提出了基于 GAN 的语义对齐网络半监督跨模态哈希方法(GAN-SASCH)。该模型基于生成对抗网络,结合了语义对齐的概念。生成对抗网络分为两个模块,分别是生成器和判别器,生成器学习拟合未标记数据的相关性分布并生成虚假的数据样本,判别器则用于判断数据对样本是来自数据集还是生成器。通过这两个模块之间展开极大极小的对抗博弈游戏,不断提升生成对抗网络的性能。语义对齐能充分利用不同模态之间的相互作用和对称性,统一不同模态的相似性信息,有效地指导哈希代码的学习过程。除此之外,还引入了自适应学习优化参数以提升模型性能。在 NUS-WIDE 和 MIRFLICKR25K 数据集上,对比了所提方法与 9 种相关前沿方法,使用 MAP 与 PR 图两种评价指标验证了所提方法的有效性。

关键词:跨模态哈希;生成对抗网络;语义对齐;半监督;自适应学习
中图分类号　TP391

Semi-supervised Cross-modal Hashing Method for Semantic Alignment Networks Based on GAN

LIU Huayong and ZHU Ting
School of Computer Science,Central China Normal University,Wuhan 430079,China

Abstract　Supervised methods have achieved a lot of results in cross-modal retrieval and have become popular methods. However,these methods rely too much on labeled data and do not make full use of the rich information contained in unlabeled data. To solve this problem,unsupervised methods have been studied,but when relying solely on unlabeled data,the results are not ideal. Therefore,this paper proposes a semi-supervised cross-modal hashing method for semantic alignment networks based on GAN (GAN-SASCH). This model is based on generative adversarial networks that incorporate the concept of semantic alignment. The generative adversarial network is divided into two modules. The generator learns to fit the correlation distribution of the unlabeled data and generates a spurious data sample,and the discriminator is used to determine whether the data pair sample comes from the dataset or the generator. By developing a very small adversarial game between these two modules,the performance of the generative adversarial network is continuously improved. Semantic alignment can make full use of the interaction and symmetry between different modalities,unify the similarity information of different modalities,and effectively guide the learning process of hash code. In this paper,adaptive learning optimization parameters are also introduced to improve the performance of the model. On NUS-WIDE and MIRFLICKR25K datasets,we compare the proposed method with 9 related frontier methods,and verify the effectiveness of the proposed method by using two evaluation indicators,MAP and PR map.

Keywords　Cross-modal hash,Generative adversarial network,Semantic alignment,Semi-supervised,Adaptive learning

1　引言

随着网络的迅猛发展,不同形式的媒体数据呈爆炸式增长,导致需要检索的数据规模越来越大。在大规模数据中进行快速准确的检索是跨模态检索面临的重大挑战,对此,人们提出了使用哈希方法的解决方案。哈希一词意为切割和混合,哈希函数通过切割和混合信息生成哈希结果[1]。哈希方法的主要目的是将高维特征转换为二进制码,从而能够将不同模态但包含相似信息的数据转化为相似的二进制码[2]。跨模态哈希方法具有两个显著优势:一是哈希码支持位操作,可快速有效地计算汉明距离;二是哈希码能大大减少存储空间的需求。

跨模态检索面临的主要挑战在于不同模态之间的异质性,该特性导致不同模态数据的语义不同,因此无法直接进行比较[3]。为解决这一问题,跨模态哈希方法应运而生,并得到了大量的研究。传统的哈希方法可以分为以下 3 种:1)监督

到稿日期:2024-04-02　返修日期:2024-09-26
基金项目:教育部人文社会科学研究项目(21YJA870005)
This work was supported by the Humanities and Social Sciences Research Project of the MoE(21YJA870005).
通信作者:朱婷(zhuting_ccnu@163.com)

方法依靠哈希函数来维持由标签提供的语义相关性，如量化相关哈希方法(QCH)[4]、共正则化哈希方法(CRH)[5]、语义保留的散列哈希方法(SePH)[6]和优化知识蒸馏监督哈希方法(OKD)[7]；2)无监督方法通过把多模态数据映射到公共空间中，使不同模态的数据之间的关系最大化，如矩阵分解哈希方法(CMFH)[8]、跨视图哈希方法(CVH)[9]、可预测的双视图哈希方法(PDH)[10]和无监督三元哈希方法(CUTHash)[11]；3)半监督方法则结合了有标签数据和无标签数据的语义信息，如生成对抗半监督哈希方法(SCHGAN)[2]。然而，传统的哈希算法侧重于对模态间的语义联系进行挖掘，这将导致哈希码的性能下降。由于深度学习的特征表示能力很强，同时也蕴含着丰富的语义信息[12]，因此基于深度学习的哈希方法在检索的有效性和精确度上有很大的突破。例如，基于深度学习的跨模态哈希方法有深度哈希方法(DCMH)[13]、跨模态汉明哈希方法(CMHH)[14]、多标签语义保持的深度哈希方法(MLSPH)[15]和深度在线哈希方法(DOCHCM)[16]。

然而，上述深度哈希方法都是基于监督跨模态方法的，监督方法训练使用的数据需要对应的标签，但手工对大量的样本进行标记既费时又费力。尽管无监督方法能够极大地降低开销，但其执行效率和准确性有待提高。有效地挖掘无标签数据所蕴含的语义信息，是提升跨模态检索准确率的关键。传统的半监督方法能够有效地使用未标记数据指导哈希代码，但由于缺乏对语义信息的充分挖掘，其在检索精度上与监督方法相比仍有较大差距。

近年来，生成对抗网络在图像合成、目标检测和图像分类等领域得到了广泛的应用。基于 GAN 能够有效地描述数据的分布，为跨模态检索提供了广阔的发展空间。受到语义对齐模型在提取无标签图像语义方面的启发，本文提出了一种新的基于 GAN 的语义对齐网络半监督跨模态哈希方法(GAN-SASCH)。该方法利用生成对抗网络和语义对齐算法更好地指导哈希代码的学习，从而在检索的有效性和精确度上有更进一步的突破。本文的主要贡献总结为以下几点。

1)提出了基于 GAN 的语义对齐网络的半监督跨模态哈希方法，在该方法中，生成对抗网络的生成器学习拟合未标记数据的相关性分布并生成假数据对，而判别器负责判断这些数据对是否为真实。这两个模块之间进行极大极小的对抗博弈，能够提升生成对抗网络的性能。这种方式能够不断地优化生成对抗网络，以更好地实现语义对齐任务。

2)在 GAN-SASCH 中，语义对齐网络充分利用了不同模态的相互作用和对称性，统一了不同模态中的相似性信息，将重建的特征与原始特征对齐，以更好地指导哈希代码学习。

3)自适应学习是一种能够动态调整学习率的方法，它使得梯度较大的参数学习率下降速度更快，而梯度较小的参数学习率下降速度更慢，从而有效优化参数，提升模型性能。

在 NUS-WIDE 和 MIRFLICKR25K 数据集上，将本文方法与 9 种相关前沿方法进行了对比实验，使用 MAP 与 PR 曲线两种评价指标验证了本文方法的有效性。

2 相关工作

2.1 跨模态哈希方法

2.1.1 无监督跨模态哈希方法

无监督跨模态哈希方法的核心思想与 CCA[17]类似，都是将多个模态的媒体数据投射到公共的汉明空间中，以此获得它们之间的最大相关性。例如，协同矩阵分解哈希方法(CMFH)[8]使用潜在因子模型并结合矩阵分解，可以从多个模态样本中提取出统一的哈希码。另外，Hu 等提出的协同重构嵌入方法(CRE)[18]采用了不同的模态特定模型来处理异构类型的数据。此外，无监督对比学习哈希方法(UCCH)[19]基于对比学习方法设计，针对跨模态哈希进行优化。用于无监督跨模态哈希的相似图相关重构网络方法(SGRN)[20]利用关系图重构模块得到相似关系图，从而实现相似度的对齐。

尽管无监督方法不依赖标注数据，但由于缺乏标签信息，检索精度通常较低。

2.1.2 监督跨模态哈希方法

监督跨模态哈希方法由于利用了标签，因此可以极大限度地使用语义信息，从而达到比无监督方法更高的检索准确率。例如，语义保留哈希方法(SePH)[6]通过最小化 KL-散度来逼近学习到的哈希码与汉明空间之间的关系，以保留数据的语义信息。另外，离散潜在因子哈希方法(DLFH)[21]是基于离散潜在因子模型的跨模态哈希方法，能够有效地学习到模态间的潜在因子信息。此外，增强型离散多模态哈希方法(EDMH)[22]在离散约束下，同时学习二进制代码和哈希函数，以实现更好的跨模态检索效果。有监督对比离散哈希方法(SCDH)[23]将无监督的对比方法应用到监督方法中，并且直接生成离散哈希码，增强了哈希码的特征表示能力。

监督方法因为使用的训练数据带有标签，所以在保持哈希码的相似性的同时，提高了检索精度，在实际应用中通常具有更好的性能表现。

2.1.3 半监督跨模态哈希方法

半监督跨模态哈希方法的提出旨在平衡性能和标签需求之间的冲突，能够同时利用有标签数据和无标签数据的语义信息。例如，Zhang 等提出的半监督多图哈希方法(MGH)[24]通过多视图模型复合离散哈希学习，结合了多个视图的信息以提高哈希码学习的效果。另外，生成对抗半监督哈希方法(SCHGAN)[2]利用生成对抗网络来指导哈希代码的学习，并结合基于强化学习的算法来优化模型的训练。半监督跨模态图卷积网络(MCGCN)[25]包含两个模态异质性通道和一个跨模态通道，分别用于学习每个模态的异质性和共享表示。

相比于监督方法和无监督方法，半监督方法在实践中往往能够节省成本，具有更大的实际应用价值。然而，目前半监督跨模态哈希方法在利用语义信息方面仍存在提升空间，研究成果相对较少。

2.1.4 基于深度学习的跨模态哈希方法

深度学习的特征表示能力强，同时也蕴含着丰富的语义信息[12]，因此在跨模态哈希方法中承担着重要的角色。例如，深度跨模态哈希方法(DCMH)[13]利用两个深度神经网络分别从图像和文本中学习特征，实现了有效的跨模态哈希编码。另外，深度三联哈希方法(TDH)[26]引入图正则化技术来保持汉明空间中哈希码的语义相似性，进一步提升了检索效果。深度监督跨模态哈希方法(DSCMR)[27]通过将多个模态数据映射到一个共同的表征空间，实现了对多个模态样本相似度的直接比较。基于深度表示学习的混合 DAER 方法(DEAR)[28]构建的双重注意力网络和增强关系网络能够精确

提取细粒度的权重信息，提高相似度的计算精度。

众多研究表明，基于深度学习的监督方法在跨模态检索方面已经取得了非常好的成果。

2.2 生成对抗网络

生成对抗网络(GAN)[29]最初被提出，是为了通过对抗训练的方式估计生成模型的一个全新框架。GAN 由两部分组成：生成器模型 G 用于学习拟合未标记数据的相关性分布并生成假数据样本，判别器模型 D 用于区分真实数据样本和生成的假数据样本。这两个模型进行对抗训练，相互博弈，以更好地学习数据的表示方式。模型框架如图 1 所示。

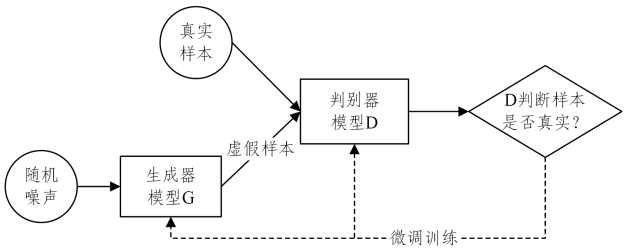


图 1 生成对抗网络模型框架图

Fig. 1 Framework of generative adversarial network model

在跨模态检索领域，对抗跨模态检索方法(ACMR)[30]通过施加三重约束最小化语义相同的图像和文本之间的差距，同时最大化不同语义的图像和文本之间的距离。另外，Peng 等提出的跨模态生成对抗网络(CM-GAN)[31]致力于模拟不同模态数据的联合分布，探索模态间和模态内的相关性。此外，Bai 等提出的深度对抗离散哈希方法(DADH)[32]采用对抗性训练来学习跨模态的特征。双重注意生成对抗网络(DA-GAN)[33]是具有双重注意力机制的对抗性语义表征模型，能够更精确地表征高层语义关联。

传统的深度哈希方法通常使用预定义的损失函数来约束相应的哈希码，以强制缩小模态之间的间隙，但这意味着大部分有用的信息会被中和，导致哈希码难以捕捉模态之间的一致性。将深度学习与 GAN 相结合，有助于保持固有的模态一致性。

2.3 语义对齐

语义对齐(Semantic Alignment)[32]是一种将不同模态的数据映射到共享的语义空间的方法，旨在通过学习模态之间的相互关系和语义表达，使得不同模态的数据在语义层面上能够有效匹配和比较。在跨模态检索领域，对语义对齐尚未有很多深入的探索。Wen 等提出的双语义关系注意网络(DSRAN)[34]由语义关系模块和联合语义关系模块组成，能够同时学习不同层次的语义关系。另外，Zhang 等提出的基于图的语义对齐网络(GSAN)[35]通过将不同模态特征与文本数据的语义嵌入进行对齐，实现了学习公共表示的目的。

语义对齐能够充分利用跨模态的相互作用和对称性，统一不同模态间的相似性信息，从而解决了由于模态异构性差异而导致二进制哈希码的语义描述与特征的语义描述不一致的问题。但是在实际情况中，源数据集(训练集)与目标数据集(测试集)很少来自同一分布。上述语义对齐仅关注语义和模态之间的鸿沟，而忽略了数据集之间的域差异，因此，仅仅依靠语义对齐学习的特征无法有效地适用于目标数据集。

GAN 拥有强大的数据表达能力，其中生成器生成拟合数据分布，判别器判断数据来源，展现了优秀的数据泛化能力。因此，将语义对齐与生成对抗网络相结合，能够有效地减少不同数据集之间的域差异。

3 GAN-SASCH

3.1 网络结构

GAN-SASCH 方法的整体框架由生成器模型和判别器模型组成，如图 2 所示。生成器 G 接受标记和未标记的图像和文本作为输入，在特征层进行语义对齐，然后在哈希层进行比较，最终输出生成的虚假数据对。判别器 D 接受真实数据对和生成的虚假数据对的输入，用于判断这些数据对是真实的还是生成的。生成器 G 和判别器 D 之间展开极小极大游戏，进行对抗博弈训练，使得模型逐步学习并优化。此外，使用自适应学习来调整参数，确保模型能够达到最佳性能。

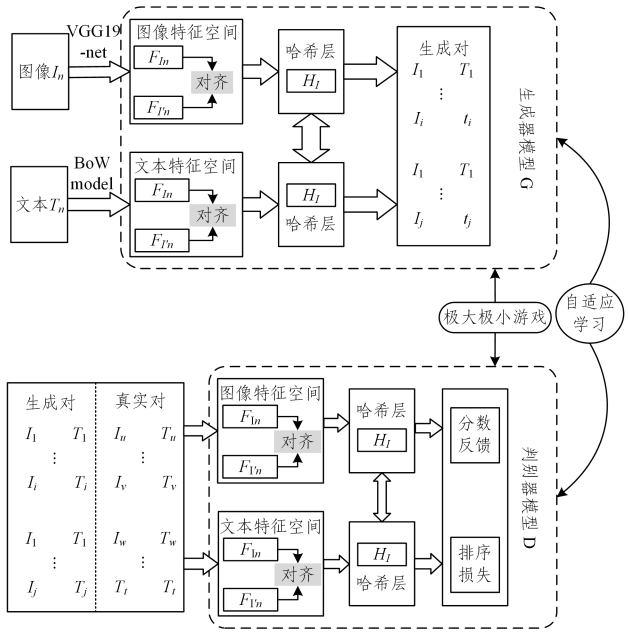


图 2 GAN-SASCH 模型的整体框架图

Fig. 2 Overall framework of GAN-SASCH model

生成器 G 采用双路径结构，同时接受两种类型的数据输入。在特征层使用语义对齐来对齐原始特征和重构特征，充分利用不同模态中的共现信息来指导哈希码的学习。生成器 G 的两条路径都包含由全连接层组成的哈希层，将全连接层作为哈希函数将特征映射到哈希码中。

$$H(x)=\mathrm{sigmoid}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{F}+v) \tag{1}$$

其中，$\boldsymbol{F}$ 为提取的特征值，$\boldsymbol{W}$ 为哈希码学习的权值，v 为偏差参数。由于哈希码 $H(x)\in[0,1]$是连续的实值，因此需要使用一个阈值函数来获得二进制码。

$$B_k(x)=g(H(x))=sgn(H_k(x)-0.5),k=1,2,\cdots,q \tag{2}$$

最后，将哈希码重构为成对特征向量：

$$\boldsymbol{F}'_*=Dec(B_*;\delta_*),*\in\{I,T\} \tag{3}$$

其中，$Dec(.;.)$表示哈希码解码器，B 代表的是二进制码表达式，δ 是图文解码器要学习的参数。因此可以得到重构损失：

$$\mathcal{L}_{\mathrm{F}}=\sum\|\boldsymbol{F}'_*-\boldsymbol{F}_*\|^2,*\in\{I,T\} \tag{4}$$

判别器 D 的结构与生成器 G 的结构基本一致，其主要作用是区分真实存在的数据对和 G 生成的数据对，最后对判断结果进行打分。这个分数会反馈给 G，作为 G 工作的评判标准。通过对抗性训练，D 的反馈可以帮助 G 不断优化自身生成的数据对，使其更接近真实数据对。

3.2 损失函数

损失函数由两部分组成，分别是生成对抗损失和语义对齐损失。

3.2.1 生成对抗损失

用 I 表示图像，用 T 表示文本，那么跨模态数据集表示为 $D=\{I,T\}$，$I\in\mathcal{R}^I$，$T\in\mathcal{R}^T$，并进一步划分训练集 D_{tra} 和测试集 D_{tes}。D_{tra} 又分为有标签的训练集 $D_{\text{tra}}^{\text{L}}=\{I_{\text{tra}}^{\text{L}},T_{\text{tra}}^{\text{L}}\}$ 和没有标签的 $D_{\text{tra}}^{\text{U}}=\{I_{\text{tra}}^{\text{U}},T_{\text{tra}}^{\text{U}}\}$。测试集 $D_{\text{tes}}=\{I_{\text{tes}},T_{\text{tes}}\}$。

生成器 G 的目标函数定义为：$p_\theta(i^{\text{U}}\mid q_t,r)$ 和 $p_\theta(t^{\text{U}}\mid q_i,r)$，当给定一个文本查询 q_t，G 试图从 $I_{\text{tra}}^{\text{U}}$ 中选择相关的图像 I^{U}，学习拟合未标记数据的相关性分布 $p_{\text{true}}(i^{\text{U}}\mid q_t,r)$。判别器 D 的目标函数定义为：$f_\varphi(i,q_t)$ 和 $f_\varphi(t,q_i)$，预测查询和候选数据对的相关性得分，学习准确区分真实对和生成对。文本搜索图像的过程可表示为：

$$V(G,D)=\min_\theta\max_\varphi\sum_{k=1}^{n}E_{i\sim p_{\text{true}}(i^{\text{U}}\mid q_t^k,r)}[\log(D(i^{\text{L}}\mid q_t^k))]+E_{i\sim p_\theta(i^{\text{U}}\mid q_t^k,r)}[\log(1-D(i^{\text{U}}\mid q_t^k))] \tag{5}$$

类似地，图像查询文本任务的对抗过程表达式如下：

$$V(G,D)=\min_\theta\max_\varphi\sum_{k=1}^{n}E_{t\sim p_{\text{true}}(i^{\text{U}}\mid q_t^k,r)}[\log(D(t^{\text{L}}\mid q_i^k))]+E_{t\sim p_\theta(i^{\text{U}}\mid q_t^k,r)}[\log(1-D(t^{\text{U}}\mid q_i^k))] \tag{6}$$

$$\mathcal{L}_{\text{adv}}=\sum_{k=1}^{n}(E_{i\sim p_{\text{true}}(i^{\text{U}}\mid q_t^k,r)}[\log(\text{sigmoid}(f_\varphi(i^{\text{L}},q_t^k)))]+E_{i\sim p_\theta(i^{\text{U}}\mid q_t^k,r)}[\log(1-\text{sigmoid}(f_\varphi(i^{\text{U}},q_t^k)))]) \tag{12}$$

通过观察式(5)和式(6)，发现两个任务的公式是具有对称性的，因此后面的公式都以文本搜索图像为例。其中，$p_\theta(i^{\text{U}}\mid q_t,r)$ 定义为一个 softmax 函数：

$$p_\theta(i^{\text{U}}\mid q_t,r)=\frac{\exp(-\parallel H_T(q_t)-H_I(i^{\text{U}})\parallel^2)}{\sum_{i^{\text{U}}}\exp(-\parallel H_T(q_t)-H_I(i^{\text{U}})\parallel^2)} \tag{7}$$

$D(i^{\text{U}}\mid q_t)$ 和 $D(t^{\text{U}}\mid q_i)$ 为生成器模型 D 预测查询和候选数据对的相关性得分，我们定义了一个 sigmoid 函数：

$$D(i^{\text{U}}\mid q_t)=\text{sigmoid}(f_\varphi(i^{\text{U}},q_t))=\frac{\exp(f_\varphi(i^{\text{U}},q_t))}{1+\exp(f_\varphi(i^{\text{U}},q_t))} \tag{8}$$

$$D(i^{\text{L}}\mid q_t)=\text{sigmoid}(f_\varphi(i^{\text{L}},q_t))=\frac{\exp(f_\varphi(i^{\text{L}},q_t))}{1+\exp(f_\varphi(i^{\text{L}},q_t))} \tag{9}$$

其中，$f_\varphi(i^{\text{U}},q_t)$ 和 $f_\varphi(i^{\text{L}},q_t)$ 被定义为三元组排名损失：

$$f_\varphi(i^{\text{U}},q_t)=\max(0,m_i+\parallel H_T(q_t)-H_I(i^+)\parallel^2-\parallel H_T(q_t)-H_I(i^{\text{U}})\parallel^2) \tag{10}$$

$$f_\varphi(i^{\text{L}},q_t)=\max(0,m_i+\parallel H_T(q_t)-H_I(i^{\text{L}})\parallel^2-\parallel H_T(q_t)-H_I(i^-)\parallel^2) \tag{11}$$

其中，i^+ 表示语义相似的图像；i^- 则表示语义不同的图像；m_i 是一个边缘参数，通常设置为 1。最后，可以得到对抗损失 $\mathcal{L}_{\text{adv}}$ 的表达式：

$$\mathcal{L}_{\text{adv}}=\sum_{k=1}^{n}(E_{i\sim p_{\text{true}}(i^{\text{U}}\mid q_t^k,r)}[\log(\text{sigmoid}(f_\varphi(i^{\text{L}},q_t^k)))]+E_{i\sim p_\theta(i^{\text{U}}\mid q_t^k,r)}[\log(1-\text{sigmoid}(f_\varphi(i^{\text{U}},q_t^k)))]) \tag{12}$$

3.2.2 语义对齐损失

语义对齐过程包括排序对齐和相似度对齐。

1)排序对齐。汉明距离是用于衡量两个等长二进制码之间的差异性的指标，在二进制码用作特征向量的情况下，汉明距离可以由二进制码之间的角距离决定。首先计算图文对的余弦相似矩阵：

$$\boldsymbol{S}_{x,y}^B(k,w)=\cos(B_{x,k},B_{y,w}),x,y\in\{I,T\} \tag{13}$$

最后，计算排序对齐损失 $\mathcal{L}_{\text{R}}$，其由对角元素量化误差和对角元素对称损失组成：

$$\mathcal{L}_{\text{R}}=\sum_{k=1}^{n}\parallel 1-\boldsymbol{S}_{I,T}^B(k,k)\parallel^2+\frac{1}{2}\sum_{k=1}^{n}\sum_{w=1}^{n}\parallel\boldsymbol{S}_{I,T}^B(k,w)-\boldsymbol{S}_{I,T}^B(w,k)\parallel^2 \tag{14}$$

2)相似度对齐。二进制码保留了来自不同模态的原始邻域关系，这些邻域关系的语义描述是排序的关键。这里也使用余弦相似函数来计算图文特征的相似矩阵 $\boldsymbol{S}_{x,y}^F$：$x,y\in\{I,T\}$。首先将二进制码的相似信息与特征向量的相似性信息进行对齐：

$$\mathcal{L}_{\text{intra}}=\sum\parallel k\boldsymbol{S}_{x,x}^F-\boldsymbol{S}_{x,x}^B\parallel^2,x\in\{I,T\} \tag{15}$$

其中，k 是权重因子，目的是提高相似性对齐的存在性。然后，从模态间的角度进行对齐，进一步统一相似性信息：

$$\mathcal{L}_{\text{inter}}=\sum\parallel k(\boldsymbol{S}_{I,I}^F+(1-\alpha)\boldsymbol{S}_{T,T}^F)-\boldsymbol{S}_{x,y}^B\parallel^2,(x,y)\in\{I,T\} \tag{16}$$

其中，α 是权重因子，用于调节不同模态间邻域关系的重要性。最后，将模态间和模态内的相似性对齐损失合并，结合式(4)，可以得到语义对齐损失：

$$\mathcal{L}_{\text{SA}}=\mathcal{L}_{\text{F}}+\mathcal{L}_{\text{R}}+\mathcal{L}_{\text{intra}}+\beta\mathcal{L}_{\text{inter}} \tag{17}$$

其中，β 是平衡模态间和模态内的权重因子。

3.2.3 整体损失函数

结合式(12)和式(17)，可以得到总的损失函数为：

$$\mathcal{L}=\mathcal{L}_{\text{adv}}+\omega\mathcal{L}_{\text{SA}} \tag{18}$$

其中，ω 是语义对齐损失的权重因子，用于调整语义对齐损失在总体损失中的权重。

3.3 优化

在训练生成器 G 时，保持判别器 D 的参数不变；在训练判别器 D 时，保持生成器 G 的参数不变。优化后的表达式如下：

$$\theta^*=\arg\min_\theta v(G,D) \tag{19}$$

$$\varphi^*=\arg\max_\varphi V(G,D) \tag{20}$$

4 实验与结果分析

4.1 实验设置

在实验开始前，需要选定验证实验结果的数据集、对比方法，以及评价标准。

4.1.1 数据集

1)NUS-WIDE 数据集[36]共包含 81 个类，共有 269498 幅图像，每幅图像都有对应的标签。在实验中，可以将数据集分成 15000 对训练集和 5000 对测试集，并且将该数据集中的 1%作为查询集，其余数据作为检索数据集。为了表示每张图像，可以使用从 VGG19-Net 中提取的 4096 个深度特征，而每个文本可以使用 1000-D BoW(词袋模型)表示。

2)MIRFLICKR25K 数据集[37]是从 Flickr 中收集的包含 25000 张图像的数据集,每张图像都有对应的标签。该数据集一共有 24 个语义标签,并使用其中的至少 1 个作为图像的标签。与 NUS-WIDE 数据集划分类似,可以将 MIRFLICKR25K 数据集分成 15000 对训练集和 5000 对测试集,选择该数据集中 5%的数据作为查询集,其余部分作为检索数据集。与 NUS-WIDE 数据集相同,图像可以使用从 VGG19-Net 中提取的深度特征表示,而文本可以使用 BoW 表示方法。

4.1.2 对比方法

1)语义保留散列哈希(SePH)[6]将训练样本的语义矩阵转化成一个概率分布,并通过最小化 KL-散度将其与汉明空间中学习到的哈希码进行逼近。

2)深度跨模态哈希(DCMH)[13]利用两个深度神经网络分别从图像和文本中学习特征。为了维持模态间的语义相似度,采用了负对数似然损失。

3)跨模态汉明哈希(CMHH)[14]设计了一种基于指数分布的成对焦点损失,以惩罚汉明距离超过阈值的实例。

4)自约束注意力哈希(SCAHN)[38]为每个哈希代码分配不同的权重,并将特征学习的中间层的标签及其特征加入到哈希函数学习中。

5)多标签语义保持深度哈希(MLSPH)[15]采用样本中的多个标签对原样本进行语义相似度计算,并通过存储机制维持多标签的语义相似度。

6)CLIP 增强型网络哈希(CLIP4CMR)[39]研究了 CLIP 增强网络对跨模态检索性能的影响。CLIP 是当前具有代表性的视觉语言预训练模型。

7)生成对抗半监督哈希(SCHGAN)[2]利用生成对抗网络来指导哈希代码的学习,并通过基于强化学习的算法来驱动模型的训练。

8)深度对抗多标签哈希(DAMCH)[40]通过多标签和深度特征共同建立邻接矩阵。

9)对称一致哈希(SCH)[41]提出了一种新的两步哈希技术。

4.1.3 评价标准

1)平均准确率(MAP)值是所有训练过程中的平均准确度的平均值,其表达式如下:

$$AP=\frac{1}{R}\sum_{k=1}^{n}\frac{k}{R_k}\times rel_k \tag{21}$$

其中,n 是数据集的规模,R 为相关图像数量,R_k 为前 k 个返回的图像数量。如果排名第 k 位的图像为相关,则 $rel_k=1$,否则 $rel_k=0$。

2)精度-召回(P-R)曲线是以 precision(精准率)和 recall(召回率)这两个为变量而形成的曲线,其中 recall 为横坐标,precision 为纵坐标。P-R 曲线表示检索到的排名表在一定召回水平上的精度,常用于衡量信息检索性能。

4.2 实验结果

为了验证本文方法的有效性,开展了对比实验,使用的数据集是 NUS-WIDE 和 MIRFLICKR25K。与 4.3 节介绍的其他 9 种方法进行比较,评估了不同哈希码长度(16 bit,32 bit,64 bit 和 128 bit)的 MAP 值,并分别给出了图像检索文本(I2T)和文本检索图像(T2I)这 2 个检索任务的结果。最终的实验结果如表 1 和表 2 所列。此外,还进一步比较了哈希码长度为 64 bit 的情况下不同方法在 2 个检索任务上的 PR 曲线,具体结果如图 3 和图 4 所示。

表 1 在 NUS-WIDE 数据集上的两个检索任务的 MAP 分数

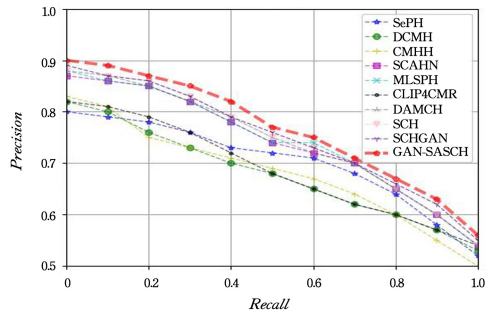
Table 1 MAP scores for two retrieval tasks on NUS-WIDE dataset

Methods	I2T_MAP				T2I_MAP			
	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit
SePH[6]	0.5902	0.6072	0.6394	0.6327	0.5964	0.6260	0.6534	0.6471
DCMH[13]	0.6311	0.6580	0.6582	0.6637	0.6504	0.6906	0.6931	0.6957
CMHH[14]	0.6614	0.6735	0.6738	0.6741	0.6769	0.6770	0.6870	0.6879
SCAHN[38]	0.6544	0.6614	0.6740	0.6738	0.6720	0.6802	0.6913	0.6910
MLSPH[15]	0.6487	0.6706	0.6815	0.6802	0.6629	0.6842	0.6998	0.6973
CLIP4CMR[39]	0.6079	0.6168	0.6294	0.6385	0.6215	0.6303	0.6421	0.6547
DAMCH[40]	0.6560	0.6709	0.6751	0.6774	0.6552	0.6710	0.6739	0.6763
SCH[41]	0.6579	0.6722	0.6773	0.6796	0.7265	0.7301	0.7569	0.7632
SCHGAN[2]	0.7031	0.7140	0.7222	0.7389	0.7182	0.7232	0.7469	0.7624
GAN-SASCH	**0.7293**	**0.7354**	**0.7461**	**0.7578**	**0.7438**	**0.7542**	**0.7639**	**0.7752**

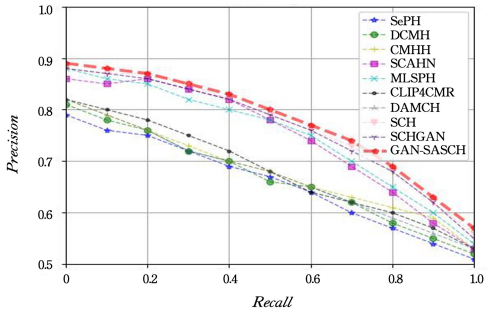
表 2 在 MIRFLICKR25K 数据集上的两个检索任务的 MAP 分数

Table 2 MAP scores for two retrieval tasks on MIRFLICKR25K dataset

Methods	I2T_MAP				T2I_MAP			
	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit
SePH[6]	0.7091	0.7213	0.7014	0.7149	0.6848	0.7085	0.7192	0.7108
DCMH[13]	0.7262	0.7278	0.7435	0.7389	0.7620	0.7703	0.7784	0.7562
CMHH[14]	0.7322	0.7311	0.7459	0.7462	0.7335	0.7229	0.7357	0.7359
SCAHN[38]	0.8146	0.8282	0.8298	0.8261	0.8025	0.8105	0.8182	0.8157
MLSPH[15]	0.8055	0.8222	0.8333	0.8320	0.7843	0.8044	0.8144	0.8132
CLIP4CMR[39]	0.7610	0.7723	0.7846	0.7892	0.7809	0.7932	0.8058	0.8101
DAMCH[40]	0.8025	0.8142	0.8269	0.8316	0.7951	0.8025	0.8092	0.8136
SCH[41]	0.7213	0.7326	0.7389	0.7452	0.8248	0.8429	0.8571	0.8620
SCHGAN[2]	0.7380	0.7453	0.7572	0.7683	0.7711	0.7902	0.7934	0.8041
GAN-SASCH	**0.8284**	**0.8390**	**0.8495**	**0.8507**	**0.8471**	**0.8566**	**0.8683**	**0.8694**



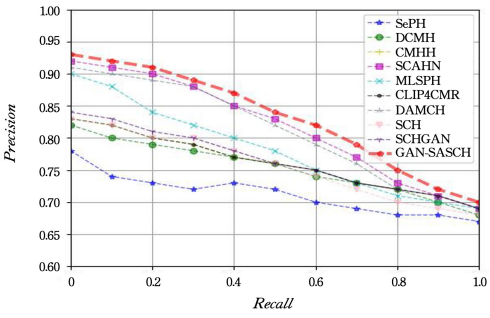
(a) NUS-WIDE_I2T



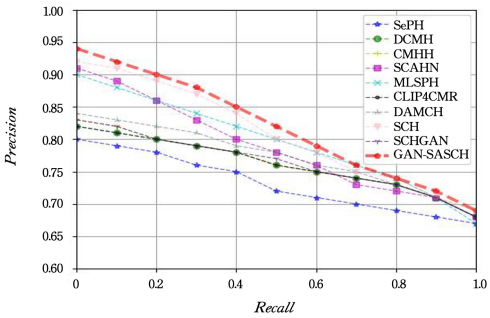
(b) NUS-WIDE_T2I

图 3 在 NUS-WIDE 数据集上的哈希码长度为 64 bit 的两个检索任务的 PR 曲线

Fig. 3 PR curves for two retrieval tasks with a hash code length of 64 bit on the NUS-WIDE dataset



(a) MIRFLICKR25K_I2T



(b) MIRFLICKR25K_T2I

图 4 在 MIRFLICKR25K 数据集上的哈希码长度为 64 bit 的两个检索任务的 PR 曲线

Fig. 4 PR curves for two retrieval tasks with hash code length of 64 bit on MIRFLICKR25K dataset

4.3 结果分析

根据 4.2 节中的结果，本文所提出的方法在跨模态检索任务中表现优秀，获得了最佳的检索精度。通过分析实验结果，可以得出 GAN-SASCH 具有以下性能优势。

1）如表 1 所列，在 NUS-WIDE 数据集上，哈希码长度为 128 bit 时，与 SCHGAN 相比，I2T 任务的 MAP 值增长了约 0.02，与 SCH 相比，T2I 任务的 MAP 值增长了约 0.01；哈希码长度 64 bit 时，和 SCHGAN 相比，I2T 任务的 MAP 值从 0.722 2 提升到了 0.746 1，与 SCH 相比，T2I 任务的 MAP 值从 0.765 9 提升到了 0.763 9；哈希码长度 32 bit 时，与 SCHGAN 相比，I2T 任务的精确度增长了 0.025 4，与 SCH 相比，T2I 任务的精确度增长了 0.024 1；哈希码长度 16 bit 时，与 SCHGAN 相比，I2T 任务的 MAP 值提升了约 0.026，与 SCH 相比，T2I 任务的 MAP 值提升了约 0.017 3。

2）如表 2 所列，在 MIRFLICKR25K 数据集上，哈希码长度为 128 bit 时，I2T 任务中相比 MLSPH 方法精确度提升了约 0.02，T2I 任务中相比 SCH 方法精确度提升了约 0.007；哈希码长度 64 bit 时，I2T 任务中的 MAP 值比 MLSPH 增长了 1.62%，T2I 任务的 MAP 值比 SCH 增长了 0.74%；哈希码长度为 32 bit 时，与 SCAHN 相比，I2T 任务中的精确度提升了 1.08%，与 SCH 相比，T2I 任务的精确度提升了 1.317%；哈希码长度 16 bit 时，与 SCAHN 相比，I2T 任务的 MAP 值提升了 1.38%，与 SCH 相比，T2I 任务的 MAP 值提升了 2.23%。

3）根据图 3 和图 4，在不同的检索任务（I2T 和 T2I）中，GAN-SASCH 在 64 bit 哈希码长度下展现出了较好的性能，具有较高的精度和召回率，其有效性得到进一步验证。

4.4 消融实验

为了验证自适应学习和语义对齐网络在跨模态哈希检索上的性能，进行了消融实验。GAN-SASCH-1 为仅移除自适应学习的方法，GAN-SASCH-2 为仅移除语义对齐部分的方法，它们的参数设置都与 GAN-SASCH 参数设置一致，哈希码长度都为 64 bit。实验结果如表 3 所列。

表 3 在 NUS-WIDE 和 MIRFLICKR25K 数据集上的消融实验的 MAP 分数

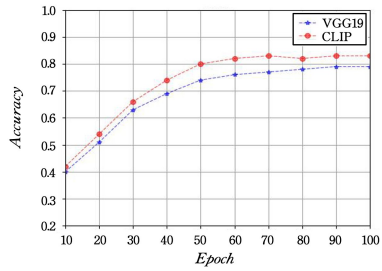
Table 3 MAP scores of ablation experiments on NUS-WIDE and MIRFLICKR25K datasets

Methods	NUS-WIDE			MIRFLICKR25K		
	GAN-SASCH-1	GAN-SASCH-2	GAN-SASCH	GAN-SASCH-1	GAN-SASCH-2	GAN-SASCH
I2T_MAP	0.740 5	0.712 2	**0.746 1**	0.831 9	0.806 8	**0.849 5**
T2I_MAP	0.759 8	0.738 4	**0.763 9**	0.851 1	0.824 1	**0.868 3**

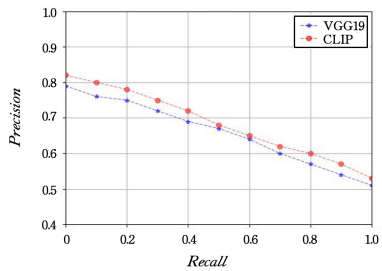
根据表 3 的结果：仅移除自适应学习会导致两个检索任务（I2T 和 T2I）的 MAP 值降低约 1%左右，这是因为自适应学习能够根据参数的不同情况灵活调整，提高了模型在训练过程中的效率和稳定性；仅移除语义对齐后，两个任务的 MAP 值甚至降低了 3%～4%，这是因为语义对齐技术能够充分利用不同模态的相互作用和对称性，统一不同模态的相似性信息，将重建的特征与原始特征对齐，其与 GAN 结合之后进一步减少了不同数据集之间的域差异，能够更好地指导哈希函数的学习，从而提升了模型的性能表现。因此，自适应学习和语义对齐技术在跨模态哈希检索任务中的重要性得到了验证，它们对模型性能的提升起到了关键作用。

5 展望与讨论

虽然本文方法表现不错，实验环境成本较低，但训练过程相对缓慢。哈希方法因为其简化的二进制特征表示导致效率下降，尽管采用了轻量级架构，但特征提取的二次执行仍然需要较长的时间来完成。因此，影响效率的关键在于特征提取阶段，尤其是图像特征的质量。CLIP 是一种多模态预训练神经网络，其利用大量的图文数据进行预训练，以学习图像与文本之间的对齐关系。CLIP 的图像和文本编码器对图文数据进行编码，提取出高质量的特征。为了探讨后续是否可以结合 CLIP 进行有效展望，分别将 CLIP 的图像编码器与 VGG19 提取的特征注入到同一个图像分类模型中，并对结果进行了定量分析。通过对图 5 所示准确率与召回率的分析，可以发现使用 CLIP 提取的图像特征相较于 VGG19 提取的图像特征效果更佳。因此，未来的研究工作可以考虑在特征提取阶段引入 CLIP 的图像与文本编码器，提取高质量的特征，从而提高下游任务的性能。



(a) Score



(b) P-R

图 5 对 CLIP 与 VGG19 提取的特征在图像分类任务上的定量分析

Fig. 5 Quantitative analysis of features extracted by CLIP and VGG19 on image classification tasks

结束语 本文提出了一种新的半监督跨模态哈希方法——基于 GAN 的语义对齐网络半监督跨模态哈希(GAN-SASCH)。生成对抗网络的生成器学习拟合未标记数据的相关性分布并生成虚假数据对样本，判别器用于判断真实数据对样本和虚假数据对样本。通过这两个模块的对抗训练，不断提升生成对抗网络的性能。另外，引入了语义对齐技术，充分利用不同模态数据的相互作用和对称性，统一相似性信息，将重建的特征与原始特征对齐，结合 GAN 进一步减少不同数据集之间的域差异，能够更好地指导哈希函数学习，提升了模型的性能。此外，在训练过程中使用自适应学习参数，在保持训练的同时动态调整学习率，提高模型性能。最后，在两个公共数据集上与其他 9 种相关前沿方法进行了对比实验，结果均验证了本文方法的有效性。

参 考 文 献

[1] CHI L H, ZHU X Q. Hashing techniques: a survey and taxonomy[J]. Association for Computing Machinery, 2017, 50(1): 1-36.

[2] ZHANG J, PENG Y X, YUAN M K. SCH-GAN: semi-supervised cross-modal hashing by generative adversarial network [J]. IEEE Transactions on Cybernetics, 2020, 50(2): 489-502.

[3] CHEN N, DUAN Y X, SUN Q F. Cross-modal search research literature review[J]. Computer Science and Exploration, 2021, 15(8): 1390-1404.

[4] WU B T, YANG Q, ZHENG W S, et al. Quantized correlation hashing for fast cross-modal search[C]// Proceedings of the 24th International Conference on Artificial Intelligence. 2015: 3946-3952.

[5] ZHEN Y, YEUNG D Y. Co-regularized hashing for multimodal data[C]// Proceedings of the 25th International Conference on Neural Information Processing Systems. 2012: 1376-1384.

[6] LIN Z J, DING G G, HU M Q, et al. Semantics-preserving hashing for cross-view retrieval[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3864-3872.

[7] ABID H, HENG C L, MEHBOOB H, et al. A gradual approach to knowledge distillation in deep supervised hashing for large-scale image retrieval[J]. Computers and Electrical Engineering. 2024, 120(PC): 109799-109799.

[8] DING G G, GUO Y C, ZHOU J L, et al. Large-scale cross-modality search via collective matrix factorization hashing[J]. IEEE Transactions on Image Processing. 2016, 25(11): 5427-5440.

[9] KUMAR S, UDUPA R. Learning hash functions for cross-view similarity search[C]// Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. 2011: 1360-1365.

[10] RASTEGARIM, CHOI J, FAKHRAEI S, et al. Predictable dual-view hashing[C]// Proceedings of the 30th International Conference on International Conference on Machine Learning. 2013: 1328-1336.

[11] LI Y Q, LU Z W, LIU C. Unsupervised Triplet Hashing Method Based on Contrastive Learning [J]. Application Research of Computers, 2023, 40(5): 1434-1440

[12] PENG L K, LU X M, XU Q B. Research progress on cross-modal hash retrieval based on deep learning[J]. Journal of Data Communications, 2022, 208(3): 32-38.

[13] JIANGQ Y, LI W J. Deep Cross-modal hashing[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3270-3278.

[14] CAO Y, LIU B, LONG M S, et al. Cross-modal hamming hashing [C]// Proceedings of the European Conference on Computer Vision. 2018: 202-218.

[15] 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.

[16] XIE Y C, ZENG X H, WANG T H, et al. Deep online cross-modal hashing by a co-training mechanism[J]. Knowledge-Based Systems, 2022, 257: 109888.

[17] HARDOOND R, SZEDMAK S, SHAWE-TAYLOR J. Canonical correlation analysis: an overview with application to learning methods[J]. Neural Computation, 2004, 16(12): 2639-2664.
[18] HUM Q, YANG Y, SHEN F M, et al. Collective reconstructive embeddings for cross-modal hashing[J]. IEEE Transactions on Image Processing, 2019, 28(6): 2770-2784.
[19] HU P, ZHU H Y, LIN J, et al. Unsupervised contrastive cross-modal hashing[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 3877-3889.
[20] YAOD, LI Z X, LI B, et al. Similarity graph-correlation reconstruction network for unsupervised cross-modal hashing[J]. Expert Syst. Appl., 2024, 273: 1-13.
[21] JIANGQ Y, LI W J. Discrete latent factor model for cross-modal hashing[J]. IEEE Transactions on Image Processing, 2019, 28(7): 3490-3501.
[22] CHENY, ZHANG H, TIAN Z B, et al. Enhanced discrete multi-modal hashing: more constraints yet less time to learn[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(3): 1177-1190.
[23] LI Z, YAO T, WANG L L, et al. Supervisedcontrastive discrete hashing for cross-modal retrieval[J]. Knowledge-Based Systems, 2024, 295: 1-13.
[24] ZHANG C, ZHENG W S. Semi-supervised multi-view discrete hashing for fast image search[J]. IEEE Transactions on Image Processing, 2017, 26(6): 2604-2617.
[25] WU F, LI S S, GAO G W, et al. Semi-supervised cross-modal hashing via modality-specific and cross-modal graph convolutional networks[J]. Pattern Recognation, 2023, 136(C): 1-10.
[26] DENGC, CHEN Z J, LIU X L, et al. Triplet-based deep hashing network for cross-modal retrieval[J]. IEEE Transactions on Image Processing, 2018, 27(8): 3893-3903.
[27] ZHENL L, HU P, WANG X, et al. Deep supervised cross-modal retrieval[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 10386-10395.
[28] HUANG Z, HU H W, SU M. Hybrid DAER based cross-modal retrieval exploiting deep representation learning. Entropy[J]. Entrpoy, 2023, 25(8): 1216-1234.
[29] GOODFELLOWI J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. 2014: 2672-2680.
[30] WANGB K, YANG Y, XU X, et al. Adversarial cross-modal retrieval[C]// Proceedings of the 25th ACM International Conference on Multimedia. 2017: 154-162.
[31] PENG Y X, QI J W. CM-GANs: cross-modal generative adversarial networks for common representation learning[J]. Association for Computing Machinery, 2019, 15(22): 1-24.
[32] ANDREJ K, LI F F. Deep visual-semantic alignments for generating image descriptions[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3128-3137.
[33] CAI L W, ZHU L, ZHANG H Y, et al. DA-GAN: Dualattention generative adversarial network for cross-modal retrieval[J]. Future Internet, 2022, 14(2): 43-43.
[34] WEN K, GU X, CHENG Q. Learning dual semantic relations with graph attention for image-text matching[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(7): 2866-2879.
[35] ZHANGL, CHEN L T, OU W H, et al. Semi-supervised cross-modal retrieval with graph-based semantic alignment network [J]. Computers and Electrical Engineering, 2022, 102(C): 1-19.
[36] CHUAT 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] HUISKESM J, LEW M S. The mir flickr retrieval evaluation [C]// Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval. 2008: 39-43.
[38] 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.
[39] ZENG Z X, MAO W J. A comprehensive empirical study of vision-language pre-trained model for supervised cross-modal retrieval[J]. arXiv: 2201.02772, 2022.
[40] YANG X H, WANG Z, LIU W H, et al. Deep adversarial multi-label cross-modal hashing algorithm [J]. International Journal of Multimedia Information Retrieval, 2023, 12: 1-12.
[41] NI H M, FANG X Z, KANG P P, et al. SCH: Symmetric consistent hashing for cross-modal retrieval[J]. Signal Processing, 2024, 215(C): 1-12.

LIU Huayong, born in 1978. Ph.D, associate professor, is a member of CCF (No. 35656M). His main research interests include cross modal retrieval, computer vision and deep learning.

ZHU Ting, born in 2001, postgraduate. Her main research interests include cross modal retrieval and deep learning.

(责任编辑:何杨)