计算机科学
COMPUTER SCIENCE

基于文本引导图像语义融合的跨模态哈希检索

顾宝程, 刘立

引用本文

顾宝程, 刘立. 基于文本引导图像语义融合的跨模态哈希检索[J]. 计算机科学, 2023, 50(11A): 221100191-6.

GU Baocheng, LIU Li. Cross-modal Hash Retrieval Based on Text-guided Image Semantic Fusion[J]. Computer Science, 2023, 50(11A): 221100191-6.

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)

一种融合CNN和Swin Transformer的医学显微图像分割模型

Medical Microscopic Image Segmentation Model Based on CNN Structure and Swin Transformer

计算机科学, 2023, 50(11A): 230200119-8. https://doi.org/10.11896/jsjkx.230200119

基于局部特征与全局表征耦合的2D人体姿态估计

Coupling Local Features and Global Representations for 2D Human Pose Estimation

计算机科学, 2023, 50(11A): 221100007-5. https://doi.org/10.11896/jsjkx.221100007

基于深度哈希学习的知识库问答检索框架

Deep Hashing-based Retrieval Framework for KBQA

计算机科学, 2023, 50(11): 227-233. https://doi.org/10.11896/jsjkx.220900206

基于多粒度的Transformer目标检测算法

Transformer Object Detection Algorithm Based on Multi-granularity

计算机科学, 2023, 50(11): 143-150. https://doi.org/10.11896/jsjkx.230600028

基于时间感知Transformer的交通流预测方法

Time-aware Transformer for Traffic Flow Forecasting

计算机科学, 2023, 50(11): 88-96. https://doi.org/10.11896/jsjkx.221000201

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221100191

基于文本引导图像语义融合的跨模态哈希检索

顾宝程　刘　立
南华大学计算机学院　湖南 衡阳 421001
(1254427045@qq.com)

摘　要　基于哈希的跨模态检索算法具有存储消耗低和搜索效率高的特点,跨模态哈希检索在多媒体数据中的应用成为当前的研究热点。目前对于跨模态哈希检索的主流方法是研究模态间哈希码的学习能力,忽视了不同模态之间的特征学习能力以及语义融合能力。将Clip中的图像-文本匹配问题转换为像素-文本匹配问题,文本特征经过Transformer解码器查询图片特征,鼓励文本特征学习到最相关的图片像素级信息,并将像素-文本匹配得分引导图片模态的特征学习,挖掘出不同模态之间的更深层次的相关联的语义信息,并引入二元交叉熵损失函数来提升模态之间的语义融合能力,在高维特征映射到低维的汉明空间时能够得到高质量的二值哈希码。在MIRFLICKR-25K和NUS-WIDE数据集上进行对比实验,实验结果表明所提算法模型在不同长度的哈希码条件下的检索效果均优于目前主流的算法。

关键词:哈希;Clip;Transformer;二元交叉熵;跨模态检索
中图法分类号　TP391

Cross-modal Hash Retrieval Based on Text-guided Image Semantic Fusion

GU Baocheng and LIU Li
School of Computing,University of South China,Hengyang,Hunan 421001,China

Abstract　Hash-based cross-modal retrieval algorithm is characterized by low storage consumption and high search efficiency,and the application of cross-modal hash retrieval in multimedia data has become a current research hot-spot. At present,the mainstream method for cross-modal hash retrieval is to study the learning ability of intermodal hash codes,ignoring the feature learning ability and semantic fusion ability between different modes. This paper transforms the image-text matching problem in Clip into pixel-text matching problem,the text features query image features through Transformer decoder,encourage text features to learn the most relevant image pixel level information,and the pixel-text matching score guide image modal feature learning,dig out the deeper related semantic information between different modalities,and introduce binary cross-entropy loss function to improve the semantic fusion ability between modalities. High-quality binary hash codes can be obtained when high-dimensional features are mapped to a low-dimensional Hamming space. Comparative experiments are carried out on MIRFLICKR-25K and NUS-WIDE datasets,and the experimental results show that the present algorithm model performs better than the current mainstream algorithms under hash codes of different lengths.

Keywords　Hash,Clip,Transformer,Binary cross-entropy,Cross-modal retrieval

1　引言

随着互联网技术的发展,文本、图片、视频等多元化的信息的数量正迎来爆炸式增长。以往的单一模态数据的检索方式已经不能满足当下网民对多元数据的检索需求。因此,多模态的检索方式应运而生。但不同模态的特征提取和表示的方式不同,不同模态之间的相似性很难衡量,如何快速和高效地得到多模态的检索结果是一个亟待解决的难题。

跨模态哈希方法(CMH)将高维实例压缩为潜在空间中的紧凑二进制码,对相似的数据产生类似的二进制码,然后用有效的XOR操作度量跨模态数据的相似性。早期的跨模态哈希检索使用手工制作的特征生成哈希码,提取特征的过程和哈希学习的过程是独立开的,所以实际应用的结果不是很理想。由于深度神经网络能够很好地学习到数据间的非线性关系,因此深度神经网络被应用到跨模态哈希方法学习中,以提高二进制代码的表示能力。深度跨模态哈希学习由哈希码生成部分和特征学习部分组成,前者着重解决如何使不同模态生成的哈希码能够准确地反映它们之间的相似性,后者则是解决如何使提取的特征具有区别性和一致性,以减少二值化过程中的信息丢失。

为了减少不同模态在二值化过程中的信息丢失,本文的贡献如下:

1)将Clip[1]中的图像-文本匹配问题转换为像素-文本匹配问题,即通过Transformer解码器的注意力机制实现文本特征与相关图像像素的匹配,鼓励文本特征学习到最相关的图片像素级信息,使用像素-文本匹配得分引导图片模态的特征学习。

2)通过二元交叉熵损失函数来优化Transformer解码的

通信作者:刘立(liuleelap@163.com)

过程，提高不同模态之间的语义融合能力。

在两个广泛使用的数据集 MIRFLICKR-25K 和 NUS-WIDE 上的实验结果表明，本文模型优于目前主流方法。

2 相关工作

跨模态哈希检索方法可分为无监督、半监督和有监督。无监督方法通过探索模态数据之间的相似性来学习公共汉明空间。如深度联合语义重构哈希(DJSRH)[2]将语义相似性融合成一个统一的亲和矩阵，以捕获输入多模态实例的潜在相关性。无监督交叉模态哈希(UDCMH)[3]在网络训练过程中结合了矩阵分解和拉普拉斯限制，并对哈希码进行了明确的约束，以保留原始数据的域特征，从而获得更好的性能。半监督利用少量的标签数据和大量的无标签数据，使得模型能够对无标签数据进行更好的分类。例如半监督图卷积哈希网络(SGCH)[4]通过图卷积网络探索高阶模态内相似性，同时将语义信息从标记样本传播到未标记数据，并通过孪生神经网络将学习到的特征投射到一个公共的汉明空间，有效保留模态内和模态间的相似性。多视图图跨模态哈希(MGCH)[5]框架利用图形推理模块处理多视图图的输出，以半监督的方式生成哈希码。监督学习方法在训练过程中从类别标签中获取信息，通常具有更好的表现。

本文着重研究基于深度学习的有监督跨模态图像文本检索。跨模态哈希方法的目标是将不同模态的特征映射到同一汉明空间，如何提取具有代表性的语义特征和关系成为了关键的步骤之一。最近针对如何获取不同模态的语义特征和关系，人们进行了大量的探索。如将语义学习与对抗学习相结合，其中自监督对抗性哈希(SSAH)[6]首次尝试使用对抗学习来解决跨模态哈希问题，将自监督语义学习与对抗学习相结合，以保持跨模态数据之间的语义相关性。深度对抗离散哈希(DADH)[7]不仅在特征学习中引入了对抗训练，而且在哈希学习中也引入了对抗训练，并使用快速离散哈希算法来减少定量损失。也有研究将语义学习和图卷积(GCN)相结合，如图卷积网络哈希(GCH)[8]引入了一种 GCN 来弥合模态差距，改进了跨模态检索。图卷积网络离散哈希(GCDH)[9]使用 GCN 来弥合不同类型数据之间的信息差距，将每个标签表示为单词嵌入，嵌入被视为一组相互依赖的对象分类器，从这些分类器中获得预测的标签，以增强跨模式的特征表示。此外，有研究通过相关矩阵操作来获取相关语义信息，如语义约束矩阵分解哈希方法(SCMFH)[10]以及联合和单独矩阵分解哈希(JIMFH)[11]都通过矩阵因子分解哈希的方法来学习模态语义表示的相关性。

上述方法虽然在跨模态图像文本检索任务上取得了不错的效果，但是忽略了图像和文本数据在细粒度层面上的语义关联。注意力机制 Transformer[12]在自然语言[13-14]和图像处理[15-16]领域获得了巨大的成功。将注意力机制 Transformer 应用到跨模态中可以挖掘不同模态间的上下文信息，以获得判别性更强的语义表示[17-18]。位感知语义转换器哈希(BSTH)[19]使用双向 Transformer 结构，同时考虑上下文信息和语义信息，获得更准确和丰富的特征表示。可微跨模态哈希(DCHMT)[20]采用了一种特殊的多头 Transformer，可以同时考虑不同模态之间的关系和相互作用，从而得到更加准确和丰富的特征表示。“预训练＋微调”的方法促进了下游任务的 SOTA。跨模态哈希检索模型大多使用传统的 ImageNet(VGG 或 ResNet)预训练模型提取全局特征，考虑的仅仅是图片实例级的信息，而图片大部分是多标签的，导致在二值化的过程中保留的信息不够全面，进而导致跨模态检索效果不理想。Clip 是一个新型的跨模态预训练模型，其基于对比学习方法进行图像-文本匹配。Clip 模型为了能够更好地学习到图像和文本之间的语义关系，其一共训练了 4 亿个图片-文本对，因此 Clip 模型已经被转移到下游任务中[21-23]并取得了很好的效果。本文对图片进行像素级的细粒度建模，采用预训练模型减少计算成本；通过迁移 Clip 预训练模型中的知识来提取图像和文本特征，再利用 Transformer Decoder 将带有上下文信息的图像特征和文本特征进行交叉融合，并利用融合后的得分信息来引导不同模态的语义学习。此外，本文通过引入二元交叉熵损失函数来保证模型能够更好地学习到像素和文本之间的信息。

3 本文方法

3.1 问题描述

本文研究图像和文本之间的检索。使用 $O=\{X_i,Y_i,L_i\}_{i=1}^{n}$ 来表示数据集，其中 n 表示样本集的个数，X_i 代表第 i 个图像，Y_i 代表第 i 个文本，L_i 代表第 i 个图像-文本对的标签。假如数据集中有 j 个标签，其中单独的标签是由 0 或 1 组成，则 L_i 可以表示为 $L_i=[L_{i1},L_{i2},L_{i3},\cdots,L_{ij}]$。本文将所有实例的图像特征值定义为 V，文本特征值定义为 T，图片-文本对应的标签定义为 L。使用成对多标签相似性矩阵 $\boldsymbol{S}$ 来描述两个实例之间的语义相似性，其中 $\boldsymbol{S}_{ij}=1$ 表示在语义上 O_i 与 O_j 相似，若 $\boldsymbol{S}_{ij}=0$ 则相反。在多标签设置中，两个实例(O_i 和 O_j)由多个标签进行注释。因此，如果 O_i 和 O_j 共享至少一个标签，则定义为 $S_{ij}=1$，否则定义为 $S_{ij}=0$。

3.2 模型概述

跨模态哈希检索的模型架构如图 1 所示，主要由 3 部分组成：第一部分为图片和文本的编码模块，通过使用 Clip 模型分别对图片和文本编码；第二部分为图片和文本特征交叉增强模块，该模块将文本的标签进行编码然后和图片编码进行交叉融合，将融合后的信息通过残差的方式来增强文本特征的学习，再将学习到的相似度 score map 与图片编码进行合并，将具有语言先验知识的知识融入图片原有特征信息中；第三部分为哈希生成部分以及模型损失函数约束部分，采用损失函数来提高不同模态特征转化哈希码的能力。

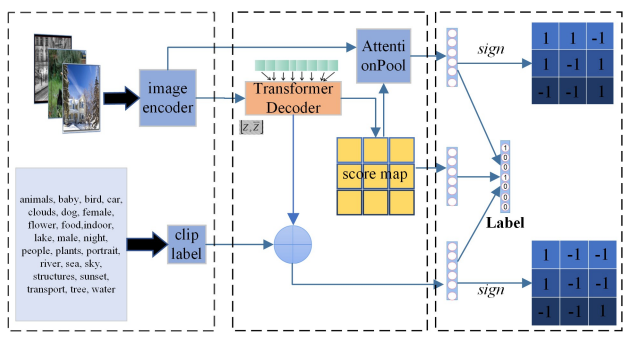

图 1 本文模型结构

Fig. 1 Structure of the proposed model

3.3 图像和文本特征学习

Clip 模型使用(ResNet)或(Vit)对图片进行特性提取，本文采用 ResNet 作为 Clip 模型的编码器。从 ResNet 模型获取 4 个 bock 输出的特征图，表示为 $\{x_i\}_{i=1}^{4}$。首先将 $x_4 \in \mathbb{R}^{H_4 W_4 * C}$ 进行全局平均池化，如式(1)所示：

$$\overline{x_4} = GAP(x_4) \tag{1}$$

其中，H_4 代表特征图的高度，W_4 代表特征的宽度，C 代表特征图的通道数。

接着将全局平均池化后的 $\overline{x_4}$ 和 x_4 进行拼接操作(concat)，然后将拼接后的特征输入到多头注意力层(MHSA)中，通过多头注意力层可以很好地获取到图片的内部特征，其过程如式(2)所示：

$$\begin{cases} \boldsymbol{Q},\boldsymbol{K},\boldsymbol{V} = f([\overline{x_4}, x_4]) \\ head = \mathrm{softmax}\left(\dfrac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \\ [\bar{z}, z] = Concat(head_1, head_2, \cdots, head_h) \end{cases} \tag{2}$$

其中，$\bar{z} \in \mathbb{R}^{1\times C}$，$z \in \mathbb{R}^{H_4 W_4 \times C}$，$h$ 为多头注意力层的个数，本文 h 的初始值设定为 32。z 为具有空间信息的特征，$\bar{z}$ 为全局特征。通过对数据集的 K 个 class 进行编码，获取 class 的特征。Clip 的文本编码使用 transformer 获得，可以表示为 $t = [t_1, t_2, \cdots, t_i]$，$t \in \mathbb{R}^{K\times C}$，其中 i 为数据集类别的个数，t_i 生成过程可以使用式(3)进行表示。

$$t_i = clip_{\mathrm{transformer}}(t_i) \tag{3}$$

通过将图片特征值 $[\bar{z}, z]$ 与类别文本生成的 t 进行 Transformer Decoder 操作，从而获取文本标签和图片特征在像素级别上的最优的关联信息，如式(4)所示：

$$m_t = TransDecoder(t, [\bar{z}, z]) \tag{4}$$

文本的特征学习的过程如下：首先通过 Clip 对图片对应的标签进行编码，然后通过残差的方式更新文本的特征。其过程如式(5)所示：

$$\begin{cases} T_i = clip(Y_i) \\ T_i = \lambda m_t + T_i \end{cases} \tag{5}$$

其中，λ 为超参数，λ 初始值为 0.001。这样既可以极大地保留原始标签的特征，又能融合图片和标签融合后的特征，提升文本特征的学习能力。

图片的特征学习的过程如下：将标签特征 t 与具有空间信息的 z 经过 L_2 归一化后，将标签特征进行转置相乘，获取标签与图片特征的匹配相关度，如式(6)所示。最后将具有文本先验知识的得分 S 与 x_4 相互拼接得到新的特征 x_4'，这样图片特征便拥有了文本标签的知识，从而提高了图片的特征提取能力，如式(7)所示。最后再次进行多头注意力(MHSA)操作生成最终图片特征。

$$S = L_2\_norm(z) * L_2\_norm(t)^{\mathrm{T}} \tag{6}$$

$$x_4' = [x_4, S], x_4' \in \mathbb{R}^{H_4 W_4 \times (C+K)} \tag{7}$$

3.4 哈希学习

在提取到图片和文本特征后，为了利用多标签注释中丰富的语义相关性，我们在特征提取模块后使用多层感知机(MLP)来获取模态内部的丰富语义信息。将经过感知机后的图片和文本特征通过激活函数(tanh)将特征值固定在 $-1\sim1$ 的范围内。最后使用 sign 函数将激活后的值转变为哈希码。其过程如式(8)所示：

$$\begin{cases} H^I = sign(\tanh(V)) \\ H^T = sign(\tanh(T)) \end{cases} \tag{8}$$

其中，H^I 和 H^T 为图片和文本的哈希符号表示。

对于两个哈希码 H_i^I 和 H_i^T，可以通过汉明距离来衡量两者的哈希码的相似性，如式(9)所示：

$$dist_H = \frac{1}{2}(K - \langle H_i^I, H_i^T \rangle) \tag{9}$$

其中，K 为哈希码的长度，$\langle H_i^I, H_i^T \rangle$ 表示两个哈希码的内积。考虑到图片或者文本被离散化为 -1 或 1 的哈希码，生成的哈希长度是不固定的。内积的结果会受到维度的影响，同时余弦相似度关注的是两个向量之间的夹角，因此在高维和低维时都能够维持相同的相似度。所以在汉明空间的相似度评估中，本文采用余弦距离 $\cos(\cdot,\cdot)$ 作为替代。如式(10)所示：

$$\begin{cases} dist_H = \dfrac{K}{2}(1 - \mathrm{COS}(H_i^I, H_i^T)) \\ \mathrm{COS}(H_i^I, H_i^T) = \dfrac{\langle H_i^I, H_i^T \rangle}{\| H_i^I \| \, \| H_i^T \|} \end{cases} \tag{10}$$

三元组损失函数的目标是最小化不同模态语义相似项之间的距离，最大化不相似项之间的距离。本文的三元损失如式(11)所示：

$$J_{tri} = \sum_{i,j,k}^{n} \max(\cos(H_i^I, H_k^{T^-}) - \cos(H_i^I, H_j^{T^+}) + m, 0) + \max(\cos(H_i^T, H_k^{I^-}) - \cos(H_i^T, H_j^{I^+}) + m, 0) \tag{11}$$

其中，m 为超参数，初始值为 0.3；H_i^I 表示图片哈希码；$H_j^{T^+}$ 表示与图片语义相似的文本实例，$H_k^{T^-}$ 则表示相反的实例。文本处理方式与图片相同。

虽然余弦三元组损失函数可以很好地保留模态的语义信息，但仍有一些问题需要进一步考虑，如图 2 所示，它只优化了两个矢量的夹角，而没有考虑矢量的实际位置。因此，本文加入了余弦量化损失，如式(11)所示：

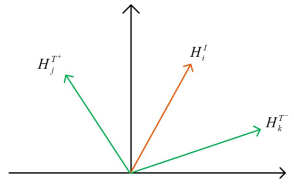

图 2 余弦量化损失案例

Fig. 2 Cosine quantization loss case

$$J_{\cos} = \sum_{i}^{n}\left(\frac{1}{1+\exp(\cos(|H_i^I|,1))} + \frac{1}{1+\exp(\cos(|H_i^T|,1)}\right) \tag{12}$$

为了提升不同模态特征的学习能力，本文模型中增加了一个标签层；为了约束模型学习，本文增加了标签预测损失。如式(13)所示：

$$J_{\mathrm{lab}} = \frac{1}{n}\sum_{i=1}^{n}(\| \boldsymbol{Q}^{\mathrm{T}} H_i^T - L_i \|_2 + \| \boldsymbol{Q}^{\mathrm{T}} H_i^I - L_i \|_2) \tag{13}$$

为了能够让文本模态和图片模态在进行交叉融合时学习到更加深层次的关系，本文将 score map 进行自适应池化(AdaptivePooling)后经过全连接层，使用 Sigmoid 激活后通过二元交叉熵损失进行约束。其过程如式(14)所示：

$$J_{\mathrm{bce}} = -\frac{1}{n}\sum_{i}^{n} L_i * \log(S_i) + (1 - L_i) * \log(1 - S_i) \tag{14}$$

其中，S_i 为经过激活函数后的标签概率值。

联合以上的所有损失得到本文所提模型最后的损失，如式(15)所示：

$$J = J_{tri} + J_{COS} + J_{lab} + J_{bce} \tag{15}$$

4 本文方法

4.1 数据集及评价指标

MIRFLICKR-25K 数据集：原始数据集中包含 25000 个图片文本标签对，每张图片被标记了一个或多个标签，文本标签的总类别数为 24。通过去除一些无效数据，最终保留了 20015 个图片文本标签对。本文通过随机抽样的方式抽取了 10000 对数据作为模型的训练数据集，同样采用随机的方式抽取 2000 对数据作为查询集。

NUS-WIDE 数据集：原始数据集中包含 269648 个图片文本标签对。同样，每张图片被标记了一个或多个标签，文本标签的总类别数为 81。本文采用前 10 个标签数量最多的图片作为实验的数据集，最终使用了 186577 个图片文本标签对。同样通过随机抽样的方式抽取了 10000 对数据作为模型的训练数据集，同样采用随机的方式抽取 2000 对数据作为查询集。

跨模态哈希算法的检索性能使用平均准确率（mAP）进行评估，其公式如式(16)所示：

$$\begin{aligned} mAP &= \frac{1}{M}\sum_{i=1}^{M} AP(q_i) \\ AP(q_i) &= \frac{1}{N}\sum_{r=1}^{R} p(k)d(k) \end{aligned} \tag{16}$$

其中，R 为用于检索数据的个数；N 为查询后与查询数据相关的个数；$p(k)$ 为检索前 k 个数据的精度；$d(k)$ 的值为 1 或者 0，表示检索与数据集是否相关；M 表示查询数据集。

4.2 实验对比

本文算法一共选取了 6 个近几年在深度跨模态哈希方法中取得了不错效果的模型来进行对比实验。这 6 种模型为自监督对抗性哈希（SSAH）[6]、深度对抗性离散哈希（DADH）[7]、可微跨模态哈希（DCHMT）[20]、自约束和基于注意力哈希（SCAHN）[24]、多标签语义保留的深度跨模态散列哈希（MLSPH）[25]、对手引导的非对称哈希（AGAH）[26]。

本次实验在 GPU NVIDIA RTX2070 8GB，CPU 为 Intel i7-10700 4.60GHz，RAM 为 32GB 的硬件环境上进行。

表 1 和表 2 列出了不同模型在两个数据集上输出哈希长度为 8，16，32 任务下的 mAP，表中 I2T 代表图像检索文本，T2I 代表文本检索图像。

表 1 MIRFLICKR-25K 数据集上对比结果

Table 1 Comparison results on MIRFLICKR-25K dataset

任务	算法	mAP		
		8 bit	16 bit	32 bit
I2T	DCHMT	0.8914	0.9036	**0.9198**
	AGAH	0.8018	0.8584	0.8707
	SSAH	0.8761	0.8794	0.9030
	SCAHN	0.8180	0.8324	0.8559
	MLSPH	0.8116	0.8359	0.8491
	DADH	0.8149	0.8259	0.8275
	OURS	**0.8963**	**0.9078**	0.9134
T2I	DCHMT	0.8761	0.8904	0.9032
	AGAH	0.8270	0.8454	0.8615
	SSAH	0.8221	0.8314	0.8499
	SCAHN	0.8061	0.8240	0.8390
	MLSPH	0.8058	0.8304	0.8443
	DADH	0.8185	0.8303	0.8314
	OURS	**0.9111**	**0.9211**	**0.9250**

表 2 NUS-WIDE 数据集上对比结果

Table 2 Comparison results on NUS-WIDE dataset

任务	算法	mAP		
		8 bit	16 bit	32 bit
I2T	DCHMT	**0.8295**	**0.8320**	**0.8417**
	AGAH	0.5378	0.5980	0.5998
	SSAH	0.6134	0.6251	0.6280
	SCAHN	0.7107	0.7124	0.7334
	MLSPH	0.5173	0.5217	0.5335
	DADH	0.5408	0.5544	0.5572
	OURS	0.8275	0.8285	0.8403
T2I	DCHMT	0.8264	**0.8439**	**0.8502**
	AGAH	0.5481	0.5914	0.6369
	SSAH	0.6160	0.6267	0.6272
	SCAHN	0.7183	0.7226	0.7360
	MLSPH	0.5637	0.6016	0.6177
	DADH	0.5290	0.5334	0.5466
	OURS	**0.8309**	0.8336	0.8419

从表 1 中可以得出，在数据集 MIRFLICKR-25K 上进行图像检索文本和文本检索图像任务中，本文模型在不同哈希码长度中都取得了最好的结果。在图像检索文本任务中，当哈希码长度为 8 和 16 时，本文模型比对比模型中的最优模型（DCHMT）检索结果分别提升 0.5% 和 0.46%。当哈希码长度为 32 时，DCHMT 模型稍优于本文模型。在文本检索图像任务中，当哈希码长度为 8，16，32 时，本文模型比 DCHMT 检索结果分别提升 3.95%，3.45%，2.41%。

从表 2 中可以得出，在数据集 NUS-WIDE 上进行图像检索文本和文本检索图像任务中，本文模型在不同哈希码长度中同样取得了较优的检索性能。在图像检索文本任务中，在哈希码为 8，16，32 长度时，本文模型稍逊色于 DCHMT 模型。在文本检索图像任务中，在哈希码为 8 长度时，本文模型比 DCHMT 模型检索结果提升 0.54%。当哈希码长度为 16 和 32 时，DCHMT 模型稍优于本文模型。

通过实验结果可以发现，可微跨模态哈希（DCHMT）与本文模型都使用了 Transformer 注意力机制，在跨模态哈希检索任务中都取得了不错的结果，证明 Transformer 可以有效地提升不同模态特征提取能力。

4.3 消融分析

为了证明本文提出的“预训练＋微调”的方式可以提升跨模态检索能力，以及利用 Transformer 解码器可以提高不同模态之间的语义融合能力，我们设计了消融模块。消融实验在 MIRFLICKR-25K 数据集上进行，哈希长度为 16。

首先使用常规的特征提取（ResNet 和 Word2vec）方式代替本文使用的预训练模型 clip，实验结果表明跨模态预训练模型可以有效提升跨模态哈希检索的性能。结果如表 3 所列。

表 3 不同特征提取方式结果

Table 3 Results of feature extraction methods

	I2T	T2I
no_clip	0.7854	0.7983
use_clip	0.9078	0.9211

通过在模型中是否添加 Transformer 解码器的实验，证明了 Transformer 解码器可以提高不同模态之间的语义融合能力。其结果如表 4 所列。

表 4　是否使用 Transformer Decoder 的实验结果

Table 4　Experimental results of whether or not Transformer Decoder is used

	I2T	T2I
no_Decoder	0.846 1	0.858 7
use_Decoder	0.907 8	0.921 1

4.4　结果可视化和结果分析

图 3 和图 4 给出了在标签文本和图片进入 Transformer Decoder 之后，对每个标签对应的图片相关度的得分进行可视化的结果。通过将模型部分内容进行可视化，进一步解释说明本模型在检索中获得较好的提升的原因。通过图 3 和图 4 可以观察到，文本模态和图像模态得到了很好的融合，并且学习到了图片对于标签的语义关系。以图 3 为例，图片对应的标签为“person”和“sky”，通过热力图可看出模型很好地学习到了标签和图片之间的关系。除此之外，从图片中发现即使有水和云的存在，本文模型也能够获取到相应的关系。其次，本模型是基于“预训练＋微调”的方式建立模型的，它充分利用了原模型（clip）的跨模态的先验知识，能够在将文本特征和图片特征转化为哈希码之前，尽可能地保留各自的特征。

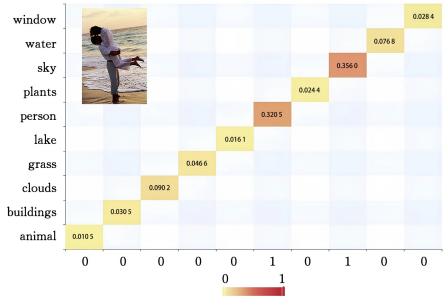

图 3　在 NUS-WIDE 数据集上标签和图片关系热力图

Fig. 3　Label and picture relationship heat map on NUS-WIDE dataset

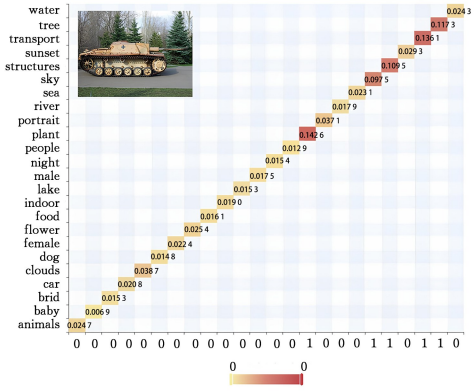

图 4　在 MIRFLICKR-25K 数据集上标签和图片关系热力图

Fig. 4　Label and image relationship heat map on MIRFLICKR-25K dataset

为了进一步观察本文所提出模型的优越性，对 MIRFLICKR-25K 数据集上的部分查询结果进行了可视化，具体如表 5 所列。我们将图片和文本转化为 8 bits 后，通过计算不同模态之间的汉明距离，选出了前 5 个检索结果进行展示。本文算法有效地学习到了不同模态之间的细粒度匹配关系。

表 5　在 MIRFLICKR-25K 数据集上查询可视化

Table 5　Query visualization on MIRFLICKR-25K dataset

Query	Results
	animals，baby，bird，car，clouds，dog，female，flower，food，**indoor**，lake，**male**，night，**people**，plant，portrait，river，sea，sky，structures，sunset，transport，tree，water
indoor，male，night，people	
	animals，baby，bird，car，clouds，dog，female，flower，food，indoor，lake，male，night，people，**plant**，portrait，river，sea，**sky**，**structures**，sunset，transport，tree，water
plant，sky，tree	

结束语　本文提出了一种基于文本引导图像语义融合的跨模态哈希检索算法，通过“预训练＋微调”的方式，将大模型知识迁移到本文算法模型中，提升了初始模态特征值的提取能力，同时节省了模型迭代的次数，让模型快速收敛。本文引入注意力机制，通过文本标签引导图像来充分挖掘不同模态间的语义关系，用融合后的得分信息来提示不同模态的语义的学习。在跨模态哈希检索领域通用的数据集上进行了对比分析，结果说明了本文算法的有效性。在两大公开数据集上进行了实验对比分析，表明本文提出的算法优于其他算法。本文的主要工作是学习高质量的模态特征，下一步工作会联合哈希码的学习，从而进一步提升跨模态哈希检索的能力。

参 考 文 献

[1]　RAO Y，ZHAO W，CHEN G，et al. Denseclip：Language-guided dense prediction with context-aware prompting[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022：18082-18091.

[2]　SU S P，ZHONG Z S，ZHANG C. Deep Joint-Semantics Reconstructing Hashing For Large-Scale Unsupervised Cross-Modal Retrieval[C]// IEEE International Conference on Computer Vision. 2019：3027-3035.

[3]　HOANG T，DO T T，NGUYEN T V，et al. Unsupervised Deep Cross-modality Spectral Hashing[J]. IEEE Transactions on Image Processing，2020，29：8391-8406.

[4]　SHEN Z，ZHAI D，LIU X，et al. Semi-Supervised Graph Convolutional Hashing Network For Large-Scale Cross-Modal Retrieval[C]// 2020 IEEE International Conference on Image Processing(ICIP). IEEE，2020：2366-2370.

[5]　SHEN X，ZHANG H，LI L，et al. Semi-supervised cross-modal hashing with multi-view graph representation[J]. Information Sciences，2022，604：45-60.

[6]　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.

[7]　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.

[8]　XU R，LI C，YAN J，et al. Graph Convolutional Network Hashing for Cross-Modal Retrieval[C]// IJCAI. 2019：982-988.

[9] CONG B, CHAO Z, QING M, et al. Graph Convolutional Network Discrete Hashing for Cross-Modal Retrieval[J/OL]. https://doi.org/10.1109/TNNLS.2022.3174970.

[10] LI W, XIONG H, OU W, et al. Semantic Constraints Matrix Factorization Hashing for cross-modal retrieval[J]. Computers and Electrical Engineering, 2022, 100: 107842.

[11] WANG D, WANG Q, HE L, et al. Joint and individual matrix factorization hashing for large-scale cross-modal retrieval[J]. Pattern Recognition, 2020, 107: 107479.

[12] ASHISH V, NOAM S, NIKI P, et al. Attention Is All You Need [C]// Conference on Neural Information Processing Systems. 2017: 5998-6008.

[13] BROWN T, MANN B, RYDERN, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems. 2020: 1877-1901.

[14] ZHANG Z, WU Y, ZHAO H, et al. Semantics-aware BERT for language understanding[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 9628-9635.

[15] LIU Z, LIN Y, CAOY, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[16] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[C]// International Conference on Learning Representations. 2021.

[17] LU J, BATRA D, PARIKH D, et al. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks[C]// Conference on Neural Information Processing Systems. 2019: 13-23.

[18] LI G, DUAN N, FANG Y, et al. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 11336-11344.

[19] TAN W, ZHU L, GUAN W, et al. Bit-aware Semantic Transformer Hashing for Multi-modal Retrieval[C]// Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022: 982-991.

[20] TU J F, LIU X L, LIN Z X, et al. Differentiable Cross-modal Hashing via Multimodal Transformers[C]// ACM International Conference on Multimedia. 2022: 453-461.

[21] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]// International Conference on Machine Learning. PMLR, 2021: 8748-8763.

[22] LUO H, JI L, ZHONG M, et al. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning[J]. Neurocomputing, 2022, 508: 293-304.

[23] TANG M, WANG Z, LIU Z, et al. Clip4caption: Clip for video caption[C]// Proceedings of the 29th ACM International Conference on Multimedia. 2021: 4858-4862.

[24] WANG X, ZHOU X, BAKKER E M, et al. Self-constraining and attention-based hashing network for bit-scalable cross-modal retrieval[J]. Neurocomputing, 2020, 400: 255-271.

[25] ZOU X, WANG X, BAKKER E M, et al. Multi-Label Semantics Preserving Based Deep Cross-Modal Hashing[J]. Signal Processing: Image Communication, 2021, 93: 116131.

[26] GU W, GU X, GU J, et al. Adversary Guided Asymmetric Hashing for Cross-Modal Retrieval[C]// International Conference on Multimedia Retrieval. 2019: 159-167.

GU Baocheng, born in 1994, postgraduate. His main research interests include computer vision and cross-modal retrieval.

LIU Li, born in 1971, Ph.D, professor. His main research interests include digital image processing and embedded, etc.