基于知识蒸馏的跨模态哈希

王锦荟, 金露, 李泽超*, 唐金辉

南京理工大学计算机科学与工程学院, 南京 210094

* E-mail: zechao.li@njust.edu.cn

收稿日期: 2021-05-22; 接受日期: 2021-11-04; 网络版发表日期: 2022-03-30

科技创新2030“新一代人工智能”重大项目(编号: 2018AAA0102002)、国家自然科学基金项目(批准号: U20B2064, 61925204, 62102181)和中国博士后科学基金项目(编号: 2020TQ0153, 2020M681607)资助

摘要 由于计算和存储的高效性, 哈希被广泛地用于大规模跨模态检索. 现有跨模态哈希方法分别对单模态数据生成哈希码, 忽略了模态内和模态间的上下文信息, 无法充分挖掘多媒体数据的潜在关联信息. 为此, 本文提出一种基于知识蒸馏的跨模态哈希方法. 该方法首先利用基于Transformer的教师网络从图像和文本数据中捕获模态内和模态间的上下文信息, 进而得到包含丰富视觉-语义关联信息的联合表示, 并将联合表示投影到低维的汉明空间以得到判别性较高的二值哈希码. 此外, 该方法利用知识蒸馏技术将教师网络学到的多模态数据潜在关联信息迁移到学生网络, 从而让学生网络生成的哈希码最大程度保留多模态关联信息. 该方法在MIRFLICKR-25K, NUS-WIDE和MS-COCO数据集上进行验证, 实验结果表明该方法的跨模态检索性能优于目前的主流方法.

关键词 哈希, 跨模态检索, 知识蒸馏, Transformer

1 引言

图像、文本、视频、音频等社交网络多媒体作为信息的载体广泛地存在于互联网, 为人们的生活提供了极大的便利. 随着互联网技术的快速发展, 社交网络多媒体数据呈现爆炸式增长^[1-5]. 面对海量社交网络多媒体数据, 用户很难从繁杂多样的数据中高效搜索到感兴趣的内容. 因此, 如何有效地从海量社交网络多媒体数据中查询到所需信息变得愈发具有挑战性. 哈希通过哈希函数将数据映射为低维的二值哈希码. 在汉明空间通过异或操作计算数据之间的汉明距离能够显著降低时间和空间复杂度, 实现快速检索.

现有的跨模态哈希方法根据使用的特征可以分为

非深度方法和深度方法. 非深度方法^[6-11]使用手工制作的特征生成哈希码, 例如, 潜在语义稀疏哈希(latent semantic sparse hashing, LSSH)^[7]、语义主题多模态哈希(semantic topic multimodal hashing, STMH)^[9]、语义保持哈希(semantics-preserving hashing, SePH)^[10]、线性子空间排序哈希(linear subspace ranking hashing, LSRH)^[11]等. 在非深度方法中, 特征提取和哈希学习相互独立, 使人工制作的特征不能生成最优的哈希码, 从而限制检索性能. 近年来, 深度学习以优越的性能受到了广泛的关注. 不同于人工特征, 深度学习使用卷积神经网络捕获数据的高层语义信息, 能够学习判别性强的特征表示. 因此, 研究人员提出了一系列的基于深度学习的跨模态哈希^[12-20]方法, 诸如多模态相似性

引用格式: 王锦荟, 金露, 李泽超, 等. 基于知识蒸馏的跨模态哈希. 中国科学: 技术科学, 2022, 52: 713-726

Wang J H, Jin L, Li Z C, et al. Crossmodal knowledge distillation hashing (in Chinese). Sci Sin Tech, 2022, 52: 713-726, doi: 10.1360/SST-2021-0214

保持哈希(multimodal similarity-preserving hashing, MMNN)^[12]、深度跨模态哈希(deep cross-modal hashing, DCMH)^[14]、自监督对抗哈希(self-supervised adversarial hashing, SSAH)^[16]等。

现有的深度跨模态哈希方法首先利用神经网络分别对不同模态数据进行特征表示学习,然后将不同模态的特征表示投影到共享空间以得到可比较的哈希码。这些方法在特征学习的过程中仅仅对单模态的信息编码,使学到的特征不能捕获模态之间的互补信息,进而限制哈希码的检索性能。为此,文献[6,7,18]提出融合多模态信息学习联合的哈希码。考虑到实际应用场景只能利用单模态信息进行检索,所以这些方法并不能适用于解决实际应用需求。此外,考虑到图像和文本包含了丰富的潜在关联语义信息,合理地挖掘这些关联信息有助于深度理解多媒体内容,实现精准检索。

近年来,知识蒸馏(knowledge distillation, KD)被广

泛用于模型压缩和知识迁移^[21-30]。知识蒸馏基于“教师-学生网络”的思想将训练好的教师网络的知识迁移到学生网络,通过最小化教师网络和学生网络的预测分布的熵促使学生网络最大化地保留教师网络的知识。通常情况下,知识蒸馏利用一个更小的学生网络逼近一个更大的教师网络。受此启发,本文设计教师网络用于学习多模态数据的多源信息(例如,潜在关联信息、模态间互补信息、模态特定的语义信息等),同时利用知识蒸馏将教师网络学习到的知识迁移到学生网络,从而解决上述问题。

基于此,本文提出一种基于知识蒸馏的跨模态哈希(cross-modal knowledge distillation hashing, CKDH)方法。该方法主要包含教师哈希网络和学生哈希网络,如图1所示。具体而言,教师哈希网络将基于Transformer的双流网络作为主干网络,通过引入协同注意力机制编码模态间的上下文信息,从而挖掘出不同模态之间的互补信息和潜在关联知识。类似地,教师哈希网

图1 基于知识蒸馏的跨模态哈希方法框架图。教师哈希网络捕捉多模态数据的上下文信息用于生成联合哈希码。学生哈希网络通过知识蒸馏最大程度保留联合哈希码的多模态关联信息,用于学习模态特定的哈希码

Figure 1 The framework of the proposed cross-modal knowledge distillation hashing. The teacher hash network captures the context information in multi-modal data to generate joint hash codes. Then, the student hash network learns the model-specific hash codes by retaining the associated multi-modal information from joint hash codes through knowledge distillation.

络同时引入自注意力机制编码模态内的上下文信息, 从而挖掘模态特定的语义信息. 最后, 教师哈希网络融合多模态信息进行哈希学习, 生成联合哈希码. 此外, 学生哈希网络利用较小的神经网络(例如, AlexNet^[31]和多层感知机)生成哈希码. 该方法利用知识蒸馏思想约束学生哈希码, 使其与教师哈希网络的联合哈希码一致, 从而让学生哈希网络能够很好地保留教师网络包含的多模态数据知识, 进而生成判别性高的哈希码. 在三个通用的跨模态检索数据集上对该方法进行了验证, 实验结果表明该方法的跨模态检索性能优于现有的主流方法.

本文的主要贡献如下:

(1) 本文在跨模态哈希中利用Transformer捕获多模态数据的模态内和模态间上下文信息学习联合哈希码;

(2) 本文利用知识蒸馏思想, 将基于Transformer的教师哈希网络学到的多源知识迁移到较小的学生哈希网络, 从而使学生哈希网络能够很好地保留多模态信息, 提高哈希码的判别性;

(3) 在三个广泛使用的跨模态检索数据集上进行了充分的实验验证, 实验结果表明该方法能够有效提高跨模态检索性能.

2 相关工作

跨模态哈希方法^[6-20]将不同模态的特征投影到一个共同的汉明空间生成低维的二值哈希码, 从而实现不同模态数据的相似性度量. LSSH^[7]利用稀疏编码和矩阵分解来获得图像和文本的潜在语义特征, 将学习到的潜在语义特征映射到同一联合空间来计算样本相似性. STMH^[9]将图像和文本信息转换成语义主题特征, 通过计算文本或图像中是否包含主题来生成哈希码. SePH^[10]通过最小化样本标签的相似度分布和样本哈希码相似度分布的KL散度(Kullback-Leibler divergence)学习统一的哈希码, 再使用核逻辑回归分别对图像和文本学习哈希函数. LSRH^[11]利用特征空间的排序结构学习基于排序的哈希函数. 近年来, 深度学习在诸多领域展现出了不俗的学习潜力, 其学习到的高层语义表示打破了传统手工制作特征的限制, 因此诞生了一系列深度跨模态哈希工作^[12-20]. MMNN^[12]利用孪生网络联合保持模态间和模态内的相似性来学习

跨模态哈希函数. DCMH^[14]将深度特征提取和离散哈希码学习融合到一个统一框架, 进行端到端的跨模态哈希学习. SSAH^[16]将对抗学习以自监督的方式引入到跨模态哈希当中, 利用两个独立的网络来最大化不同模态特征表示的语义相关性和一致性. 深度联合语义重构哈希(deep joint-semantics reconstructing hashing, DJSRH)^[19]通过构造联合语义关联矩阵来保持原始数据的邻域关系, 使哈希码最大限度地重构该矩阵来学习哈希函数. 基于分类与哈希检索任务之间的联系, 同等引导判别性哈希(equally-guided discriminative hashing, EGDH)^[20]构建了一个共同语义保持分类器来指导不同模态哈希函数的训练, 进而生成能够保持语义结构的可区分哈希码. 这些跨模态哈希方法虽然保留了数据的相似性关系, 但忽略了模态内和模态间的上下文信息, 致使学习到的哈希码不能很好地捕获多模态数据的潜在关联信息.

随着注意力机制的兴起, Transformer^[32,33]在自然语言处理领域受到越来越多的关注. 不同于传统的卷积神经网络, Transformer利用注意力机制捕获数据之间的上下文关系从而使网络的特征表达能力更强. 近年来, Transformer被广泛用于图像分类、目标检测、表示学习等任务^[34-39]. ViT^[34]首先将图像切割为数个子图构成子图序列, 通过自注意力编码图像的全局信息来学习特征表示. DETR^[35]提出将卷积神经网络与Transformer结合, 通过推理检测对象和全局图像的上下文关系得到预测的检测目标集合. 为了学习具有上下文信息的多模态表示, ViLBERT^[36]通过协同注意力机制对不同模态信息进行匹配和融合. 不同于ViLBERT, UNITER^[37]和Unicoder-vl^[38]将图像和文本信息编码到同一个特征空间, 利用自注意力机制学习联合表示. 这些基于Transformer的工作通过注意力机制挖掘数据的高层语义知识, 学习数据的上下文信息以获得判别性更强的语义表示. 本文利用Transformer从多源数据中捕获模态内和模态间的上下文信息, 探索多模态数据的互补性与相关性以获得判别性更高的哈希码.

知识蒸馏通过使网络结构较简单的学生网络尽可能地保留网络结构较复杂的教师网络的知识, 从而使教师网络的知识迁移到学生网络^[21-30]. Hinton等人^[21]提出通过最小化相对熵指导学生模型保留教师网络的知识. 概率知识迁移(probabilistic knowledge transfer,

PKT)^[23]通过匹配数据在特征空间的分布实现知识蒸馏. 保持相似性知识蒸馏(similarity-preserving knowledge distillation, SPKD)^[24]提出将教师网络学习到的样本间相似性结构知识迁移到学生网络. 关系型知识蒸馏(relational knowledge distillation, RKD)^[25]将教师网络捕获的样本间距离和角度关系进行蒸馏. 相关一致性知识蒸馏(correlation congruence for knowledge distillation, CCKD)^[26]旨在将教师网络中的实例相关性传递给学生网络. 变分知识蒸馏(variational information distillation, VID)^[27]将知识蒸馏问题转换为教师和学生网络之间的互信息最大化. 文献[30]将知识蒸馏与跨模态哈希结合起来, 将教师网络中的样本相似度知识传递给学生网络, 来指导学生网络的哈希学习过程. 本文利用知识蒸馏, 将多模态数据的上下文关联知识蒸馏到轻量级的学生网络当中, 在简化模型加速计算的同时生成具有更多语义信息的哈希码.

3 跨模态知识蒸馏哈希

本文所提出的跨模态知识蒸馏哈希框架主要由教师哈希网络和学生哈希网络组成. 教师哈希网络引入基于Transformer的注意力机制来捕获模态内和模态间上下文信息以学习哈希码. 学生哈希网络通过知识蒸馏来学习教师哈希网络嵌入的多模态关联知识, 由此来生成更具有判别力的哈希码.

3.1 问题描述

假设由 N 个图像-文本样本对构成跨模态数据集 $O=\{\mathbf{x}_i, \mathbf{y}_i, \mathbf{a}_i\}_{i=1}^N$, 其中 $\mathbf{x}_i$ 和 $\mathbf{y}_i$ 分别表示第 i 个样本对的图像和文本. 假设总共有 C 个类别, $\mathbf{a}_i=[a_{i1}, a_{i2}, \dots, a_{ic}]$ 表示第 i 个样本对的标签. 如果样本 i 属于第 c 个类别, $a_{ic}=1$, 否则 $a_{ic}=0$. 基于样本类别标签可以得到 N 个样本的相似性矩阵 $\mathbf{S} \in \{0, 1\}^{N \times N}$, 若样本 i 和样本 j 至少有一个类别标签相同, 则样本 i 和样本 j 对应的样本相似度 $S_{ij}=1$, 反之, $S_{ij}=0$. 本文所提出方法的目标是利用深度模型为图像和文本模态数据分别生成长度为 K 的二进制哈希码 $\mathbf{b}_i^X = f_{\theta_X}(\mathbf{x}_i) \in \{-1, +1\}^K$ 和 $\mathbf{b}_i^Y = f_{\theta_Y}(\mathbf{y}_i) \in \{-1, +1\}^K$, 其中, θ_X 和 θ_Y 分别表示图像和文本网络的参数.

3.2 教师哈希网络

教师哈希网络使用基于Transformer的双流网络作为模型主干网络, 如图1所示. 每个分支都由数个基于自注意力的Transformer编码器(transformer encoder, TE)^[32]和基于协同注意力的Transformer编码器(co-attention transformer encoder, CO-TE)^[36]组成. 教师哈希网络通过注意力机制可以计算不同特征之间的语义相关性. 具体来说, 注意力机制通过查询、键和值三个向量实现计算, 查询和键之间的点积决定了值向量上的注意力分布. TE由单模态特征生成的查询、键和值来实现自注意力. CO-TE打破TE单模态的限制, 将某一模态的查询向量与其他模态的键和值向量进行计算, 从而实现多模态的协同注意力机制.

教师哈希网络采用6个TE来提取文本特征, 再由图像和文本分支的6组CO-TE与TE组合的联合学习模块对模态内和模态间的上下文信息进行学习, 挖掘多模态数据的多源信息. 利用协同注意力机制, CO-TE能够挖掘图像和文本模态之间的关联知识. 类似地, TE通过自注意力机制编码单模态内的上下文信息, 能够捕获模态特定的语义信息. 因此, 教师哈希网络利用自注意力机制和协同注意力机制能够从图像和文本数据中捕获互补和关联知识以学习视觉-语义联合表示. 利用这种联合表示生成的哈希码包含更加丰富的模态特定的语义信息和潜在关联知识. 值得注意的是, 由于需要同时使用图像和文本模态的信息, 直接将教师哈希网络学习到的哈希码用于跨模态检索任务将不符合实际应用场景. 因此, 本文提出利用知识蒸馏将教师哈希网络学到的多模态的互补和关联知识迁移到学生哈希网络.

教师哈希网络需要对原始的样本信息进行预处理. 对于教师哈希网络的文本分支, 处理方法与BERT^[33]类似, 将文本输入拆解为独立单词并和特殊标记[CLS], [SEP]组成单词序列(特殊标记分别使用单词CLS和SEP进行初始化), 再把它们与自身的位置信息(即在单词序列中的位置)进行联合编码得到文本特征序列 $V^Y = \{\mathbf{v}_{\text{CLS}}^Y, \mathbf{v}_1^Y, \dots, \mathbf{v}_{s-2}^Y, \mathbf{v}_{\text{SEP}}^Y\}$, s 表示文本特征序列长度. 教师哈希网络的图像分支首先对图像进行目标检测获得数个图像区域特征和对应的位置特征. 然后, 将区域特征取均值得到特殊标记[IMG]特征, 且将其作为区域特征序列的头部特征. 最后, 把区域特征和

与之对应的位置特征联合编码得到图像特征序列 $V^X = \{\mathbf{v}_{\text{IMG}}^X, \mathbf{v}_1^X, \dots, \mathbf{v}_{r-1}^X\}$, r 表示图像特征序列的长度. 将特征序列进行嵌入表示后, 通过 TE 和 CO-TE 模块的多模态关联学习, 得到关联特征序列 $Z^X = \{\mathbf{z}_{\text{IMG}}^X, \mathbf{z}_1^X, \dots, \mathbf{z}_{r-1}^X\}$ 和 $Z^Y = \{\mathbf{z}_{\text{CLS}}^Y, \mathbf{z}_1^Y, \dots, \mathbf{z}_{s-2}^Y, \mathbf{z}_{\text{SEP}}^Y\}$. 经过池化得到图像特征 $\mathbf{z}_{\text{IMG}}^X$ 和文本特征 $\mathbf{z}_{\text{CLS}}^Y$. 在教师哈希网络的哈希层分别通过映射 $\mathbf{z}_{\text{IMG}}^X$ 和 $\mathbf{z}_{\text{CLS}}^Y$ 来生成图像和文本模态的哈希码 $\mathbf{b}^{X,T}$ 和 $\mathbf{b}^{Y,T}$, 其中 T 表示教师哈希网络. 即教师哈希网络通过对图像 $\mathbf{x}_i$ 和文本信息 $\mathbf{y}_i$ 分别进行预处理得到图像特征序列 V_i^X 和文本特征序列 V_i^Y , 其生成哈希码表示为

$$\begin{aligned} \mathbf{b}_i^{X,T}, \mathbf{b}_j^{Y,T} &= f_{\sigma}(V_i^X, V_j^Y), \\ \text{s.t. } \mathbf{b}_i^{X,T}, \mathbf{b}_j^{Y,T} &\in \{-1, +1\}^K, \end{aligned} \quad (1)$$

其中, σ 表示教师哈希网络的参数.

为了使教师哈希网络的哈希码能够保留样本的相似性, 本文最小化相似性度量损失 L_p^T 使哈希码的相似度和样本相似度一致, 如式(2):

$$L_p^T = L_{NL}^{U,T} + L_{NL}^{X,T} + L_{NL}^{Y,T}, \quad (2)$$

其中, L_{NL} 表示负对数似然函数, X , Y 和 U 分别代表图像模态、文本模态和跨模态. $L_{NL}^{U,T}$, $L_{NL}^{X,T}$ 和 $L_{NL}^{Y,T}$ 定义为

$$L_{NL}^{U,T} = -\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \left(S_{ij} \Delta_{ij}^{U,T} - \log(1 + e^{\Delta_{ij}^{U,T}}) \right), \quad (3)$$

$$L_{NL}^{X,T} = -\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \left(S_{ij} \Delta_{ij}^{X,T} - \log(1 + e^{\Delta_{ij}^{X,T}}) \right), \quad (4)$$

$$L_{NL}^{Y,T} = -\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \left(S_{ij} \Delta_{ij}^{Y,T} - \log(1 + e^{\Delta_{ij}^{Y,T}}) \right), \quad (5)$$

其中, $\Delta_{ij}^{U,T} = \frac{1}{2}(\mathbf{b}_i^{X,T})^T \mathbf{b}_j^{Y,T}$, $\Delta_{ij}^{X,T} = \frac{1}{2}(\mathbf{b}_i^{X,T})^T \mathbf{b}_i^{X,T}$, $\Delta_{ij}^{Y,T} = \frac{1}{2}(\mathbf{b}_i^{Y,T})^T \mathbf{b}_i^{Y,T}$ 分别表示跨模态哈希码相似度、图像哈希码相似度和文本模态的哈希码相似度, T 表示转置.

为了增加哈希码的判别性, 本文将得到哈希码作为特征表示直接应用于分类任务. 最小化分类损失使得哈希码能够很好地预测类别标签. 这里, 分类损失 L_c^T 定义为

$$L_c^T = L_{CE}^{X,T} + L_{CE}^{Y,T}, \quad (6)$$

其中, L_{CE} 表示交叉熵函数, $*$ 表示图像或文本模态, 定义为

$$L_{CE}^{*,T} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C \left(a_{ij} \log(p_{ij}^{*,T}) - (1 - a_{ij}) \log(1 - p_{ij}^{*,T}) \right), \quad (7)$$

其中, $p_{ij}^{*,T}$ 表示教师哈希网络第 i 个样本的第 j 类预测分数. 哈希码 $\mathbf{b}_i^{X,T}$ 和 $\mathbf{b}_j^{Y,T}$ 经过分类层得到分类预测 $\mathbf{p}_i^{X,T}$ 和 $\mathbf{p}_j^{Y,T}$, 将它们代入上式即可求得分类损失.

此外, 教师哈希网络期望图像和文本哈希码一致, 引入一致性约束, 最小化它们之间的差异:

$$L_g^T = \frac{1}{N} \sum_{i=1}^N \|\mathbf{b}_i^{X,T} - \mathbf{b}_i^{Y,T}\|_2^2, \quad (8)$$

其中, $\|\cdot\|_2^2$ 表示 L2 范数.

综上, 教师哈希网络的优化问题被建模为

$$\min L_{\text{tea}} = \beta_1 L_p^T + \beta_2 L_c^T + \beta_3 L_g^T, \quad (9)$$

其中, β_1 , β_2 和 β_3 是用来平衡各项损失的超参数.

3.3 学生哈希网络

学生哈希网络由文本学生哈希网络和图像学生哈希网络组成. 文本学生哈希网络将三层全连接层堆叠组成网络主干, 使用二进制文本标签向量作为文本输入信息. 图像学生哈希网络舍弃了较为耗时的目标检测网络与基于注意力机制的 Transformer, 仅仅使用一个简单的卷积神经网络(AlexNet^[31])作为网络主干. 学生哈希网络的哈希层和分类层的配置和教师哈希网络相似.

类似于教师哈希网络, 学生哈希网络的优化函数同样考虑了保留模态间和模态内样本的相似性以及通过分类提高哈希码的判别性. 由此, 学生哈希网络将生成哈希码 $\mathbf{b}^{X,S}$ 和 $\mathbf{b}^{Y,S}$ 及分类预测 $\mathbf{p}^{X,S}$ 和 $\mathbf{p}^{Y,S}$, 其中, S 表示学生哈希网络. 学生哈希网络损失函数定义为

$$\min L_{\text{stu}} = \gamma_1 L_p^S + \gamma_2 L_c^S, \quad (10)$$

其中, L_p^S 和 L_c^S 分别表示学生哈希网络类似于式(2)和式(6)的相似性损失与分类损失, γ_1 和 γ_2 是用于平衡损失的超参数.

通过优化式(10)中的问题, 学生哈希网络能够保持样本的模态间和模态内相似性, 但是由于无法利用Transformer来关注多模态数据中的上下文信息, 因此无法生成包含更多语义信息的哈希码。

3.4 知识蒸馏

教师哈希网络虽然能够捕获多模态信息之间的潜在关联, 但是由于需要同时输入多源信息生成联合哈希码, 其不满足实际对单模态信息编码进行检索的应用场景(例如, 仅依靠输入的文本信息检索图像)。相比于教师哈希网络, 学生哈希网络的模型结构复杂性较低。但是它无法利用Transformer挖掘多模态数据之间的互补信息和关联知识。

为了解决上述问题, 本文引入知识蒸馏思想将教师哈希网络学习到的多模态互补信息和关联知识转移到学生哈希网络中。通过知识蒸馏, 原本知识相对匮乏的学生哈希网络在教师哈希网络的辅导下学习生成具有多模态关联信息的哈希码。特别地, 本文定义如下知识蒸馏损失:

$$L_{kd}^S = \frac{1}{N} \sum_{i=1}^N \left(\left\| \mathbf{b}_i^{X,T} - \mathbf{b}_i^{X,S} \right\|_2^2 + \left\| \mathbf{b}_i^{Y,T} - \mathbf{b}_i^{Y,S} \right\|_2^2 \right). \quad (11)$$

通过式(11), 学生哈希网络生成的哈希码最大程度地与基于Transformer的教师哈希网络学习到的哈希码相逼近, 从而使学生哈希网络能够快速准确地汲取教师哈希网络的知识, 进而生成富有多源语义信息的哈希码。综上, 学生哈希网络的优化问题被建模为

$$\min L_{stu} = \gamma_1 L_p^S + \gamma_2 L_c^S + \gamma_3 L_{kd}^S, \quad (12)$$

其中, γ_3 是用于平衡蒸馏损失的超参数。

3.5 整体目标损失

模拟教师教授学生知识的过程, 分别对式(9)和(12)中的问题进行优化以实现哈希学习的知识蒸馏。然而哈希码的二值特性使上述优化问题变成NP-hard问题。为解决此问题, 将网络输出的哈希码松弛为连续值, 将损失函数式(9)和(12)中的二进制哈希码 $\mathbf{b}$ 替换为连续的 $\tilde{\mathbf{b}}$ 。为了使松弛后连续的哈希码接近+1或-1, 定义如下量化损失:

$$L_r^\# = \frac{1}{N} \sum_{i=1}^N \left(\left\| \tilde{\mathbf{b}}_i^{X,\#} - \mathbf{1}_{\text{col}} \right\|_2^2 + \left\| \tilde{\mathbf{b}}_i^{Y,\#} - \mathbf{1}_{\text{col}} \right\|_2^2 \right), \quad (13)$$

其中, $\#$ 表示 T 或 S , $\mathbf{1}_{\text{col}} \in \mathbb{R}^K$ 表示全1的列向量。此外, 引入式(14)使生成的哈希码的每位均值为零, 每位哈希码有相同的概率为+1或者-1:

$$L_b^\# = \left\| \frac{1}{N} \mathbf{1}_{\text{row}} (\tilde{\mathbf{B}}^{X,\#})^T \right\|_2^2 + \left\| \frac{1}{N} \mathbf{1}_{\text{row}} (\tilde{\mathbf{B}}^{Y,\#})^T \right\|_2^2, \quad (14)$$

其中, $\tilde{\mathbf{B}} \in \mathbb{R}^{K \times N}$ 表示 $\tilde{\mathbf{b}}$ 组成的矩阵, $\mathbf{1}_{\text{row}} \in \mathbb{R}^{1 \times K}$ 表示全1的行向量。

综上, 将式(13)和(14)分别添加到教师哈希网络和学生哈希网络的优化问题当中, 可将式(9)和(12)中需要优化的问题分别重新定义为式(15)和(16):

$$\min L_{tea} = \beta_1 L_p^T + \beta_2 L_c^T + \beta_3 L_g^T + \delta_1 L_b^T + \delta_2 L_r^T, \quad (15)$$

$$\min L_{stu} = \gamma_1 L_p^S + \gamma_2 L_c^S + \gamma_3 L_{kd}^S + \delta_1 L_b^S + \delta_2 L_r^S, \quad (16)$$

其中, δ_1 和 δ_2 是用于平衡损失的超参数。通过优化式(15)和(16)中的问题, 教师哈希网络能够将模态内和模态间的互补信息和潜在关联知识蒸馏到轻量级的学生哈希网络当中, 使学生哈希网络生成富有多模态上下文信息的哈希码。

4 实验

为了验证所提出的跨模态知识蒸馏哈希方法的有效性, 我们在三个广泛使用的多模态图像数据集MIRFLICKR-25K, NUS-WIDE和MS-COCO上对所提方法进行性能比较和评估。

4.1 数据集

MIRFLICKR-25K^[40]: 该数据集有25000幅图像, 且每幅图像对应一个或多个文本标签。每个图像-文本对使用38个类别标签进行标注。本文删除了没有对应文本标签和类别标签的样本。在保留的17515个图像-文本对中, 随机选择2000对样本作为测试集, 剩余15515个图像-文本对作为训练集和数据库。

NUS-WIDE^[41]: 该数据集包含了269648幅图像及其对应的文本信息。该数据集具有81种类别标签, 本文选取了包含样本数最多的21种类别标签, 并舍弃了没有对应文本信息或语义标签的图像文本对, 最终得到190421个样本。在选取的21种标签中, 从每种标签随机选取100个样本作为测试集, 其余样本形成数据

库, 并从数据库中每种标签随机选取500个样本作为训练集. 即测试集和训练集的样本总数分别是2100和10500.

MS-COCO^[42]: 该数据集被划分为80个类别, 总共包含82783张训练集图片和40504张验证集图片, 每张图片拥有五句不同的文本描述信息. 本文删除了没有类别标签和文本信息的样本, 最终总共得到122218个样本. 在保留的样本中随机选择5000个和80000个样本分别作为测试集和数据库, 从数据库中随机选取10000个样本形成训练集.

4.2 数据集对比哈希方法和评价指标

为了评估所提出CKDH方法的性能, 本文将其与一些现有的先进方法进行比较. 这些方法包括4种非深度哈希方法(LSSH^[7], STMH^[9], SePH^[10], LSRH^[11])和3种深度哈希方法(MMNN^[12], DCMH^[14], SSAH^[16]). 本文采用平均精度均值(mean average precision, MAP)、精度召回(precision-recall, PR)和前 N 个返回样本的精度(precision@ N , P@ N)作为评价准则衡量不同哈希方法的性能.

4.3 实验设置

本文所提方法主要由教师哈希网络和学生哈希网络构成. 教师哈希网络使用ViLBERT^[36]模型学习图像和文本联合特征表示, 网络中包含了数个TE和CO-TE模块. 我们在教师哈希网络主干尾部添加两个全连接层用于哈希函数学习与分类预测. 哈希层将学习到的特征表示用于生成哈希码, 并使用tanh函数将哈希层的输出限制在-1~+1之间. 紧跟哈希层之后的分类层将生成的哈希码用于分类预测, 以提升哈希码的判别性, 本文使用sigmoid函数将分类预测分数限制在0~1之间. 哈希层和分类层的输出神经元个数分别设置为哈希码长度和类别标签数目. 为了更好地证明我们方法的有效性, 学生哈希网络采用了与当前大多数跨模态哈希方法较为相似的网络结构, 便于与其他同类模型比较. 文本的学生哈希网络主干由三层全连接层构成, 前两层输出神经元个数为4096, 第三层输出神经元个数为1024, 每个全连接层后使用relu函数作为激活函数. 图像学生哈希网络使用AlexNet^[31]网络提取视觉特征, 并将其网络的最后一层替换成与文本学生哈希网络主干第三层相同的全连接层. 在学生哈希网

络主干之后分别添加与教师哈希网络类似的哈希层和分类层.

对于CKDH的教师哈希网络, 将经过预训练的目标检测网络Faster R-CNN^[43]生成的图像区域特征序列作为图像输入, 序列长度 $s=37$. 文本输入为若干个单词组成的单词序列, 序列长度 $r=20$ (由于MS-COCO数据集的文本信息较多, 该数据集下 $r=35$). 对于未达到序列长度的特征序列, 使用空信息进行填充. 对于CKDH的学生哈希网络和其他对比方法, 在MIRFLICKR-25K, NUS-WIDE和MS-COCO三个数据集上, 所有的图像在输入网络时分辨率调整为 224×224 . 本文分别用1386维和1000维的二进制文本标签向量表示MIRFLICKR-25K数据集和NUS-WIDE数据集中的文本, 使用2000维的高频单词的词频向量表示MS-COCO数据集的文本.

教师哈希网络使用预训练的ViLBERT模型进行网络参数初始化, 图像学生哈希网络使用预训练的AlexNet进行初始化, 其余所有网络层参数进行随机初始化. 对于CKDH的其余超参数, $\beta_1=10$, $\beta_2=\beta_3=1$, $\gamma_1=\gamma_2=\gamma_3=1$, $\delta_1=1$, $\delta_2=0.01$. 网络输入的批量大小设置为128, 使用随机梯度下降优化器进行网络更新. 本文采用两阶段训练法, 对教师哈希网络进行训练后固定其参数, 再进行学生哈希网络的知识蒸馏训练. 教师哈希网络主干的学习率设置为 2×10^{-5} , 主干尾部的哈希层以及分类层的学习率设置为 10^{-3} . 学生哈希网络主干的学习率设置为 10^{-4} , 主干尾部的哈希层以及分类层的学习率设置为 10^{-2} .

4.4 实验结果与分析

表1~3展示了本文提出的CKDH与现有方法在8, 16, 24, 32位哈希码上的对比结果. 通过表中数据可看出, CKDH在不同长度的哈希码以及不同数据集上相较于其他方法均取得了提升. 以MIRFLICKR-25K数据集为例, CKDH相对于LSSH, STMH, SePH, LSRH方法, 在文本检索图像(text query image)任务上分别取得了39.5%, 35.6%, 13.6%, 14.7%的平均性能提升, 在图像检索文本(image query text)任务上分别取得了39.2%, 37.8%, 8.6%, 16.9%的平均性能提升. 虽然非深度的方法同样使用了深度特征进行测试, 但由于无法端到端地学习特征提取与哈希码生成, 其性能相比于深度方法依然存在一定差距. 相较于MMNN,

表 1 各方法的8, 16, 24, 32位哈希码在MIRFLICKR-25K数据集上的平均精度均值(%)^{a)}

Table 1 The MAP (%) results for all the methods on MIRFLICKR-25K dataset with 8, 16, 24, and 32 hash bits

方法	文本检索图像				图像检索文本			
	8-bit	16-bit	24-bit	32-bit	8-bit	16-bit	24-bit	32-bit
LSSH	60.18	60.00	60.65	60.20	60.87	61.02	61.53	61.78
STMH	59.82	62.95	62.57	62.60	59.62	62.90	62.52	62.69
SePH	73.36	73.75	74.16	74.79	77.22	78.51	79.19	79.52
LSRH	70.22	72.44	75.19	75.43	70.40	72.07	74.17	75.59
MMNN	73.91	74.39	72.30	74.85	71.66	72.64	74.85	73.82
DCMH	72.17	73.95	74.15	74.38	73.49	75.27	75.98	76.79
SSAH	78.37	79.91	80.62	80.34	73.75	75.89	76.39	76.02
CKDH	82.13	84.38	84.82	84.91	83.38	85.30	85.95	86.75

a) 粗体表示最优值

表 2 各方法的8, 16, 24, 32位哈希码在NUS-WIDE数据集上的平均精度均值(%)^{a)}

Table 2 The MAP (%) results for all the methods on NUS-WIDE dataset with 8, 16, 24, and 32 hash bits

方法	文本检索图像				图像检索文本			
	8-bit	16-bit	24-bit	32-bit	8-bit	16-bit	24-bit	32-bit
LSSH	46.44	46.42	47.84	47.53	45.68	46.60	47.88	48.43
STMH	45.62	45.95	49.40	50.72	49.58	49.94	52.73	54.02
SePH	56.37	56.70	60.01	60.50	58.59	57.92	61.78	63.45
LSRH	55.83	53.94	60.16	59.91	48.70	54.58	56.91	58.20
MMNN	58.19	59.05	60.64	63.47	54.49	55.09	57.15	58.51
DCMH	59.93	61.01	62.97	64.54	57.23	59.98	58.62	58.91
SSAH	63.52	65.67	65.52	65.79	61.14	62.62	63.62	63.51
CKDH	65.09	66.93	68.09	67.23	62.23	64.19	64.63	64.73

a) 粗体表示最优值

表 3 各方法的8, 16, 24, 32位哈希码在MS-COCO数据集上的平均精度均值(%)^{a)}

Table 3 The MAP (%) results for all the methods on MS-COCO dataset with 8, 16, 24, and 32 hash bits

方法	文本检索图像				图像检索文本			
	8-bit	16-bit	24-bit	32-bit	8-bit	16-bit	24-bit	32-bit
LSSH	43.05	44.68	47.39	48.31	40.68	40.64	43.55	42.83
STMH	44.13	44.63	46.30	46.09	43.75	46.09	47.38	47.02
SePH	49.62	49.58	49.53	50.42	46.31	46.13	47.06	46.16
LSRH	45.94	46.90	47.67	47.27	46.86	47.98	49.31	48.88
MMNN	50.33	50.57	51.69	51.58	50.91	51.57	51.55	51.72
DCMH	51.84	51.85	51.56	52.79	50.44	51.26	51.01	51.83
SSAH	51.03	54.04	54.89	54.88	53.10	54.86	55.17	55.84
CKDH	55.92	58.95	59.92	60.40	58.43	61.15	62.05	62.17

a) 粗体表示最优值

DCMH, SSAH三种深度方法, CKDH在文本检索图像和图像检索文本任务上分别取得了13.8%, 14.1%, 5.3%和16.5%, 13.2%, 13%的平均性能提升, 这显示出CKDH相比于现有的深度方法也具有一定优势. 类似地, 在NUS-WIDE数据集上, 相比于LSSH, STMH, SePH, LSRH, MMNN, DCMH和SSAH, CKDH在文本检索图像和图像检索文本任务上分别领先了42%, 39.7%, 14.5%, 16.5%, 10.9%, 7.7%, 2.6%和35.7%, 24.1%, 5.9%, 17.5%, 13.6%, 9%, 2%. 同样地, 在MS-COCO数据集上, CKDH相较于其他方法在文本检索图像和图像检索文本任务上取得了平均19.9%和29.3%的性能提升.

由于CKDH的教师哈希网络使用基于Transformer的网络架构, 它通过协同注意力机制发现不同模态之间的共性与关联, 同时通过自注意力机制探索特定模态下的上下文信息, 进而学习到融合了多模态信息的联合哈希码. 学生哈希网络经过知识蒸馏将这些知识汲取到自身的网络当中, 在仅使用简单网络架构和单模态信息输入的情况下就能学习到相比其他方法更加优质的哈希码. CKDH在三个数据集上的性能提升有力地证明了基于知识蒸馏学习到的具有模态内和模态

间上下文信息的哈希码更加具有判别性. 同时我们注意到, CKDH在MIRFLICKR-25K和MS-COCO数据集上图像检索文本的性能提升优于文本检索图像的提升. 由于相对于文本模态, 图像内容比较复杂, 模型更易于从文本模态捕获语义知识生成判别性高的哈希码, 使得图像检索文本的性能较优于文本检索图像的性能.

除MAP之外, 我们还使用了其他评价指标来评估哈希检索性能. 图2展示了不同方法的不同长度哈希码在MIRFLICKR-25K, NUS-WIDE和MS-COCO数据集上的前5000个返回样本的精度(P@5000)曲线. 图3展示了各方法在MIRFLICKR-25K, NUS-WIDE和MS-COCO数据集上16位哈希码长度下不同数量下的返回样本的精度曲线. 图4展示了各方法在所有数据集下16位哈希码的PR曲线. 可以看到CKDH在不同数据集上都取得了有竞争力的结果.

为了进一步验证CKDH的有效性, 我们设计消融实验, 将去除知识蒸馏损失的方法(CKDH-S*)、保留蒸馏损失的完整方法(CKDH)和教师哈希网络(CKDH-T)的MAP进行对比. 表4中展示了CKDH的16位哈希码在三个测试数据集上舍弃和保留知识蒸馏时的MAP,

图2 各方法在所有数据集上的P@5000曲线

Figure 2 P@5000 curves for all the methods on all the datasets.

图3 各方法在所有数据集上16位哈希码长度下不同数量返回样本的精度曲线

Figure 3 Precision curves in terms of top returned samples for all the methods on all the datasets with 16 hash bits.

图4 各方法在所有数据集上的16位哈希码的PR曲线

Figure 4 PR curves for all the methods on all the datasets with 16 hash bits.

即是否让学生哈希网络从教师哈希网络进行学习。可以看到在保持其余设置一样的情况下, 加入知识蒸馏

使CKDH在MIRFLICKR-25K数据集上文本检索图像和图像检索文本任务性能分别提升了1.8%和5.6%, 在

表 4 16 位哈希码的平均精度均值(%)在所有数据集上的消融实验**Table 4** Ablation study of the MAP (%) results on all the datasets with 16 hash bits

任务	方法	MIRFLICKR-25K	NUS-WIDE	MS-COCO
文本检索图像	CKDH-S*	82.85	62.02	54.12
	CKDH	84.38	66.93	58.95
	CKDH-T	92.68	74.08	67.42
图像检索文本	CKDH-S*	80.81	60.38	55.33
	CKDH	85.30	64.19	61.15
	CKDH-T	92.64	74.21	67.48

NUS-WIDE和MS-COCO数据集的对应任务上也分别取得了7.9%, 6.3%和8.9%, 10.5%的提升. 同时我们也对比了学生哈希网络与教师哈希网络的MAP, 可以看出虽然学生哈希网络能够学习教师哈希网络中的多模态关联知识. 但是由于不能直接融合多模态信息, 其性能依然与教师哈希网络存在一定的差距. 上述实验结果印证了本文工作的出发点, 即学生哈希网络的简单网络架构较难挖掘出真正具有判别性的信息. 而基于注意力机制的教师哈希网络凭借其探索模态内和模态间上下文信息的能力, 能学习到更有判别力的哈希码, 从而在引入知识蒸馏后学生哈希网络能够获得更大的检索精度提升. 另外, 对比教师哈希网络与学生哈希网络在NUS-WIDE数据集上对一个图像-文本对生成16位哈希码的所需时间, 可以得出由于教师哈希网络需要对原始图像提取区域目标特征, 其所需时间(1442.05 ms)大约是学生哈希网络(2.98 ms)的484倍. 通过知识蒸馏, 学生哈希网络在保证一定性能的情况下极大地缩短了生成哈希码的时间, 明显提升了检索

速度. 综上, 实验结果证明了CKDH可以有力提高哈希检索性能.

5 结论

本文提出了一种基于知识蒸馏的跨模态哈希方法(cross-modal knowledge distillation hashing, CKDH). 该方法将Transformer引入教师哈希网络, 使其通过注意力机制学习到多模态数据模态内与模态间的上下文信息, 进而获得视觉-语义的联合哈希码. 通过知识蒸馏, 教师哈希网络中的多模态互补关联知识被提炼到学生哈希网络当中, 使学生哈希网络在仅使用简单网络和单模态信息的情况下, 最大程度地保留丰富的多模态潜在关联知识, 从而生成具有判别性的哈希码. 为了验证所提出跨模态知识蒸馏哈希的有效性, 该方法与多种现有方法在三个广泛使用的多媒体数据集上进行了对比实验, 实验结果证明了所提出方法的优越性和有效性.

参考文献

- Li W J, Zhou Z H. Learning to hash for big data: Current status and future trends (in Chinese). *Chin Sci Bull*, 2015, 60: 485–490 [李武军, 周志华. 大数据哈希学习: 现状与趋势. 科学通报, 2015, 60: 485–490]
- Liu C, Du J P, Zhou N. A cross-media search method for social networks based on adversarial learning and semantic similarity (in Chinese). *Sci Sin Inf*, 2021, 51: 779–794 [刘翀, 杜军平, 周南. 一种基于对抗学习和语义相似度的社交网络跨媒体搜索方法. 中国科学: 信息科学, 2021, 51: 779–794]
- Hu D, Nie F P, Li X L. Deep linear discriminant analysis hashing (in Chinese). *Sci Sin Inf*, 2021, 51: 279–293 [胡迪, 聂飞平, 李学龙. 基于深度线性判别分析的哈希技术. 中国科学: 信息科学, 2021, 51: 279–293]
- Li Z, Tang J, Zhang L, et al. Weakly-supervised semantic guided hashing for social image retrieval. *Int J Comput Vis*, 2020, 128: 2265–2278
- Li Z, Tang J, Mei T. Deep collaborative embedding for social image understanding. *IEEE Trans Pattern Anal Mach Intell*, 2019, 41: 2070–2083
- Ding G G, Guo Y C, Zhou J L. Collective matrix factorization hashing for multimodal data. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, 2014. 2075–2082

- 7 Zhou J L, Ding G G, Guo Y C. Latent semantic sparse hashing for cross-modal similarity search. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. Gold Coast, 2014. 415–424
- 8 Zhang D Q, Li W J. Large-scale supervised multimodal hashing with semantic correlation maximization. In: Proceedings of the AAAI Conference on Artificial Intelligence. Quebec, 2014. 2177–2183
- 9 Wang D, Gao X B, Wang X M, et al. Semantic topic multimodal hashing for cross-media retrieval. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence. Buenos Aires, 2015. 3890–3896
- 10 Lin Z J, Ding G G, Hu M Q, et al. Semantics-preserving hashing for cross-view retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, 2015. 3864–3872
- 11 Li K, Qi G J, Ye J, et al. Linear subspace ranking hashing for cross-modal retrieval. *IEEE Trans Pattern Anal Mach Intell*, 2016, 39: 1825–1838
- 12 Masci J, Bronstein M M, Bronstein A M, et al. Multimodal similarity-preserving hashing. *IEEE Trans Pattern Anal Mach Intell*, 2013, 36: 824–830
- 13 Cao Y, Long M S, Wang J M, et al. Correlation hashing network for efficient cross-modal retrieval. In: Proceedings of British Machine Vision Conference. London, 2017
- 14 Jiang Q Y, Li W J. Deep cross-modal hashing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017. 3232–3240
- 15 Cao Y, Liu B, Long M S, et al. Cross-modal hamming hashing. In: Proceedings of the European Conference on Computer Vision. Salt Lake, 2018. 202–218
- 16 Li C, Deng C, Li N, et al. Self-supervised adversarial hashing networks for cross-modal retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake, 2018. 4242–4251
- 17 Jin L, Li Z, Tang J. Deep semantic multimodal hashing network for scalable image-text and video-text retrievals. *IEEE Trans Neural Netw Learn Syst*, 2020, 1–14
- 18 Tu R C, Mao X L, Ma B, et al. Deep cross-modal hashing with hashing functions and unified hash codes jointly learning. *IEEE Trans Knowl Data Eng*, 2020, 34: 560–572
- 19 Su S P, Zhong Z S, Zhang C. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In: Proceedings of the IEEE International Conference on Computer Vision. Seoul, 2019. 3027–3035
- 20 Shi Y F, You X G, Zheng F, et al. Equally-guided discriminative hashing for cross-modal retrieval. In: Proceedings of the International Joint Conference on Artificial Intelligence. Macao, 2019. 4767–4773
- 21 Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. *Comput Sci*, 2015, 14: 38–39
- 22 Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: Proceedings of the International Conference on Learning Representations. Toulon, 2017. 1–13
- 23 Passalis N, Tefas A. Learning deep representations with probabilistic knowledge transfer. In: Proceedings of the European Conference on Computer Vision. Munich, 2018. 268–284
- 24 Tung F, Mori G. Similarity-preserving knowledge distillation. In: Proceedings of the IEEE International Conference on Computer Vision. Seoul, 2019. 1365–1374
- 25 Park W, Kim D, Lu Y, et al. Relational knowledge distillation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 3967–3976
- 26 Peng B Y, Jin X, Liu J H, et al. Correlation congruence for knowledge distillation. In: Proceedings of the IEEE International Conference on Computer Vision. Seoul, 2019. 5007–5016
- 27 Ahn S, Hu S X, Damianou A, et al. Variational information distillation for knowledge transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 9163–9171
- 28 Li H, Liu Q, Wang J F, et al. Method of oil and gas reservoir detection based on geological knowledge distillation learning (in Chinese). *Sci Sin Inf*, 2021, 51: 40–55 [李徽, 刘淇, 王喆锋, 等. 基于地质知识蒸馏学习的油气储集层识别方法. *中国科学: 信息科学*, 2021, 51: 40–55]
- 29 Gao H, Tian Y L, Xu F Y, et al. Survey of deep learning model compression and acceleration (in Chinese). *J Softw*, 2021, 32: 68–92 [高晗, 田育龙, 许封元, 等. 深度学习模型压缩与加速综述. *软件学报*, 2021, 32: 68–92]
- 30 Hu H T, Xie L X, Hong R C, et al. Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 3123–3132

- 31 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems, 2012, 25: 1097–1105
- 32 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems. Long Beach, 2017. 6000–6010
- 33 Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, 2019. 4171–4186
- 34 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations, 2021. 1–21
- 35 Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision. 2020, 213–229
- 36 Lu J S, Batra D, Parikh D, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Proceedings of Advances in Neural Information Processing Systems. Vancouver, 2019. 13–23
- 37 Chen Y C, Li L, Yu L, et al. Uniter: Universal image-text representation learning. In: Proceedings of the European Conference on Computer Vision, 2020. 104–120
- 38 Li G, Duan N, Fang Y, et al. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *AAAI*, 2020, 34: 11336–11344
- 39 Du P F, Li X Y, Gao Y L. Survey on multimodal visual language representation learning (in Chinese). *J Softw*, 2021, 32: 327–348 [杜鹏飞, 李小明, 高雅丽. 多模态视觉语言表征学习研究综述. *软件学报*, 2021, 32: 327–348]
- 40 Huiskes M J, Lew M S. The MIR Flickr retrieval evaluation. In: Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval. Vancouver, 2008. 39–43
- 41 Chua T S, Tang J H, Hong R C, et al. Nus-wide: A real-world web image database from national university of Singapore. In: Proceedings of the ACM International Conference on Image and Video Retrieval. Santorini, 2009. 1–9
- 42 Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context. In: Proceedings of the European Conference on Computer Vision. Zurich, 2014. 740–755
- 43 Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell*, 2016, 39: 1137–1149

Crossmodal knowledge distillation hashing

WANG JinHui, JIN Lu, LI ZeChao & TANG JinHui

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

Hashing is widely used for large-scale crossmodal retrieval due to its high computation and storage efficiency. Most of the existing crossmodal hashing methods separately generate hash codes for single-modal data, ignoring the context information within and between modalities. Therefore, the learned hash codes cannot preserve the potential dependence between multimodal data. To address this issue, this paper proposes a crossmodal hashing method based on knowledge distillation. The Transformer-based teacher network is first employed to capture the inter- and intra-modal context information from images and texts to further learn the joint representations that can contain rich visual semantic-associated information. The teacher network then generates discriminate hash codes by projecting the joint representations into a compact hamming space. In addition, knowledge distillation is leveraged to transfer the potentially associated knowledge embedded in the teacher network to the student network, such that the hash codes generated by the student network can preserve the associated information of the multimodal data as much as possible. The proposed method is evaluated on the MIRFLICKR-25K, NUS-WIDE, and MS-COCO datasets. The experimental results show that the performance of the proposed method is better than that of the other compared methods, demonstrating its superiority.

hashing, crossmodal retrieval, knowledge distillation, Transformer

doi: [10.1360/SST-2021-0214](https://doi.org/10.1360/SST-2021-0214)