◆人工智能算法与应用◆

基于特征增强的对抗哈希跨模态检索*

何　沛[1],王　萌[2**],王　卓[1],卢光云[3]

(1.广西科技大学理学院,广西柳州　545000;2.广西科技大学数字启迪学院,广西柳州　545000;3.柳州工学院,广西柳州　545616)

摘要:在跨模态检索任务中,哈希方法由于其检索效率高效、储存成本低廉而被广泛应用。但是,这些方法很少关注如何去弥补主体网络将高维特征转换为哈希码的过程中所丢失的特征信息。为解决这些问题,本文提出了一种特征增强对抗跨模态哈希(Feature Boosting Adversarial Hashing for Cross-Modal,FBAH)方法。FBAH方法将子空间学习与对抗学习相结合,来减少不同模态数据的差异性。另外,构造一种类残差模块,它可以将筛选出具有区别性的特征绕过主体网络直接输入到哈希空间进行特征增强。这样,生成的哈希码能够具有更多的原始特征信息。最后,通过带有分支网络的线性分类器在标签空间进行两种方式的预测,并最小化与真实标签的差距来保证语义的不变性。本文选择两个跨模态检索任务中常用的大型数据集进行大量实验,结果表明FBAH方法的性能优于目前7种较为先进的跨模态哈希方法。

关键词:特征增强　跨模态检索　稀疏矩阵　哈希子空间学习　对抗学习

中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1005-9164(2022)04-0691-09

DOI:10.13656/j.cnki.gxkx.20220919.009

随着多媒体技术的发展,网络上每天都会产生大量的多媒体数据。由于不同类型的数据表现形式不同,如何使用一种类型的数据去灵活检索其他不同类型但包含了相同对象的数据成为了一个巨大的挑战。

不同模态的特征投影具有异质性差异,跨模态哈希方法通过将其映射到同一汉明空间去度量其相似性,是当前最普遍的跨模态方法之一。其中比较关键的一步是如何提取具有代表性的特征。早期哈希方法一般是根据数据集来手工设计特征[1,2],此类方法泛化性较差,导致其在实际应用中无法达到较好的检索性能。深度学习的发展带来了更优秀的特征提取方式,近年来出现了很多深度跨模态哈希(Deep Cross-Modal Hashing,DCMH)方法[3-5],其中包括很多具有代表性的方法[6,7]。比如,深度跨模态哈希方法[6]将特征学习和哈希码学习集成到同一个框架中,用端到端的深度神经网络将不同模态的数据变换到一个公共空间来实现跨模态检索,但是这些深度跨模

收稿日期:2022-04-04

*广西中青年教师基础能力提升项目"基于语义的跨媒体检索方法研究(2019KY1095)"资助。

【作者简介】

何　沛(1995-),男,在读硕士研究生,主要从事跨模态检索研究。

【**通信作者】

王　萌(1979-),男,副教授,主要从事自然语言理解、机器学习等方面研究,E-mail:mwang007@163.com。

【引用本文】

何沛,王萌,王卓,等.基于特征增强的对抗哈希跨模态检索[J].广西科学,2022,29(4):691-699.

HE P,WANG M,WANG Z,et al.Feature Boosting Adversarial Hashing for Cross-Modal Retrieval [J].Guangxi Sciences,2022,29(4):691-699.

态哈希方法依旧存在许多需要解决的问题。首先,许多哈希方法在从特征提取到生成哈希码的过程中使用了深度网络,在这个过程中可能会丢失部分语义信息,如何增强哈希码中包含的特征信息是一个需要解决的问题。其次,不同模态的数据具有异质性差异,如何更好地弥合异质性差异,也是值得注意的一点。另外,许多哈希方法都在更加有效地利用标签信息,使得数据中的语义信息能够被充分利用。

针对以上问题,本文提出了一种特征增强对抗跨模态哈希(Feature Boosting Adversarial Hashing for Cross-Modal,FBAH)方法,采用子空间学习和对抗学习相结合的方式来学习特征表示和哈希码。首先,受到 JFSSL 方法[8]的启发,选择网络末端的哈希空间作为子空间,利用线性网络去学习投影矩阵,将不同模态的特征映射到哈希空间,通过最小化差异来保持特征表示在不同模态间的分布一致性。其次,受到残差网络[9]和一些语义增强策略[10]的启发,FBAH 方法构造了一种类残差结构来实现特征增强,使得后续生成的哈希码包含更多语义信息。此外,为了保证能够生成高质量的哈希码,在哈希空间通过对抗学习来保证不同模态哈希码的一致性,引入余弦三元组约束[11]来保证特征间的相似性以及可区分性,采用一个带有分支的线性分类器进行两个维度的标签判别,以保证模态内语义不变性。

1 相关工作

子空间学习是跨模态检索方法的主流方法之一,其核心是将不同模态的数据映射到一个公共子空间来进行特征学习。近些年来,出现了很多基于子空间学习的跨模态方法,比如典型相关分析(CCA)[12]、偏最小二乘(PLS)[13]等方法。这些方法力求寻找一个合适的子空间[8,13],比如具有标签信息的标签空间[8]或者将图像模态映射到文本模态之中[14],通过映射到同一空间来减少不同模态数据的异质性差异。不同于上述子空间方法,本文尝试选择哈希空间作为子空间。哈希空间可以包含文本和图像的特征信息,通过子空间学习能够使不同模态的特征表示尽可能近似,以便进行跨模态检索。

解决不同模态数据的异质性是跨模态检索中的核心问题之一。对抗性跨模态检索(Adversarial Cross-Modal Retrieval,ACMR)[15]首次将对抗学习思想引入到跨模态检索之中,通过构建特征投影器和模态分类器来进行对抗学习,去尝试解决模态间数据的异质性,当模态分类器无法分辨特征投影器输入特征的归属模态时,则训练结束。因为对抗学习的有效性,Li 等[16]也尝试利用对抗学习来解决跨模态哈希问题。自监督对抗哈希(Self-Supervised Adversarial Hashing,SSAH)将自监督语义学习与对抗学习相结合,利用语义信息和不同模态子网络输出特征的对抗过程,来实现自监督过程,最终达到提升跨模态检索精度的目的。Bai 等[17]也在深度对抗离散哈希(Deep Adversarial Discrete Hashing,DADH)中,将对抗学习同时应用于特征学习模块和哈希学习模块之中,进行两次对抗学习,来保证跨模态特征表示的分布一致性。研究表明,通过充分利用对抗学习可以有效地弥合异质性差异[18]。本文将对抗学习引入哈希空间,来保证模态间数据的一致性。

相对于无监督的跨模态检索方法[19,20],有监督的跨模态方法[11,17,20]能够利用标签信息来区分具有不同语义的特征,并保证具有相同语义特征的结构不变,来进一步提高检索性能。比如,Gu 等[11]提出的对抗引导非对称哈希(Adversary Guided Asymmetric Hashing for Cross-Modal Retrieval,AGAH)方法,采用对抗学习引导的多标签注意模块去学习特征表示,并且利用多标签二进制码映射,使哈希码具有多标签语义信息,获得了很好的检索效果。Zhen 等[21]提出的深度监督跨模态检索(Deep Supervised Cross-modal Retrieval,DSCMR),通过线性分类器预测并判别投影特征的语义类别,来保证特征投影的语义结构不变,并根据语义对特征投影的余弦距离进行放缩,来提高特征间的辨识度。不同于一般有监督的哈希方法,本文使用带有分支的线性分类器进行标签预测,从两个维度实现模态内的语义保持。

2 FBAH 模型

2.1 问题定义

不失一般性,假定有 n 个图像和文本的特征对。图像模态的输入记为 $X=\{x_i\}_{i=1}^{n}$,其中,x_i 是第 i 张图像的特征向量。同样地,将文本模态的输入记为 $Y=\{y_i\}_{i=1}^{n}$,y_i 是第 i 张图像文本描述的特征向量。另外,这 n 个图像和文本对的标签记为 $L=\{l_i\}_{i=1}^{n}$,其中,$l_i=\{l_{i1},l_{i2},\cdots,l_{ic}\}\in R^c$,$c$ 是类别的数目。如果某对实例属于第 j 类,那么 $l_{ij}=1$,否则 $l_{ij}=0$。FBAH 模型的目标是学习图像模态和文本模态的两个哈希函数:H_I 和 H_T,以及两个哈希函数生成的哈希码 $B^I\in\{-1,+1\}^K$ 和 $B^T\in\{-1,+1\}^K$,K 是

哈希码的长度。

2.2 FBAH 方法的框架

FBAH 方法的整体框架如图 1 所示，可以分为特征学习和哈希学习两个部分。

在特征学习部分，首先对输入的原始高维图像特征 X 和文本特征 Y 进行子空间学习。在保持哈希码与高维特征一致性的同时，通过稀疏投影矩阵选择其中具有较大区分度的特征和 m_y。主网络则将两种高维特征通过 3 层全连接层和 1 个哈希层将高维特征映射到哈希空间。

在哈希学习部分，首先将通过主网络输出的低维特征 f^X 和 f^Y 与通过特征选择投影过来的特征信息 m^X 和 m^Y 结合，形成被增强的特征 h^X 和 h^Y。为了生成二进制码的一致性，引入对抗学习来消除模态间的异质性差异。然后，采用带有分支的线性分类器，通过两种维度预测并判别标签信息，来保证模态内语义的一致性。最后，引入余弦三元组约束来保证哈希空间中相同语义特征的相似性以及不同语义特征的区分性。

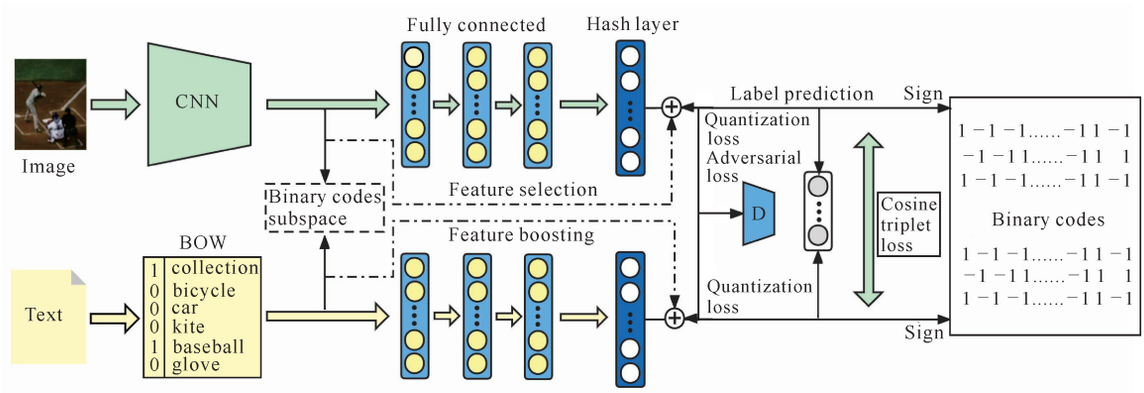

图 1 FBAH 方法的框架

Fig. 1 Frame of the FBAH method

2.3 特征增强与子空间学习

首先使用在 ImageNet 上预训练好的 CNN-F 模型[22]提取原始图片的 4 096 维高维特征，然后将 3 个全连接层和 1 个哈希层作为图像特征学习的主网络，将高维特征映射到哈希空间。对于文本特征，使用 Bow 模型提取原始高维文本特征，同样使用 3 层全连接与 1 个哈希层作为文本特征学习的主网络。函数 $f^X=F_X(X;\theta_X)$ 和 $f^Y=F_Y(Y;\theta_Y)$ 分别表示图像和文本的特征投影。其中 f^X 和 f^Y 分别表示图像和文本主网络的输出，θ_X 和 θ_Y 是两个函数的参数。如图 2 所示，本节可以分为子空间学习和特征增强两个模块。

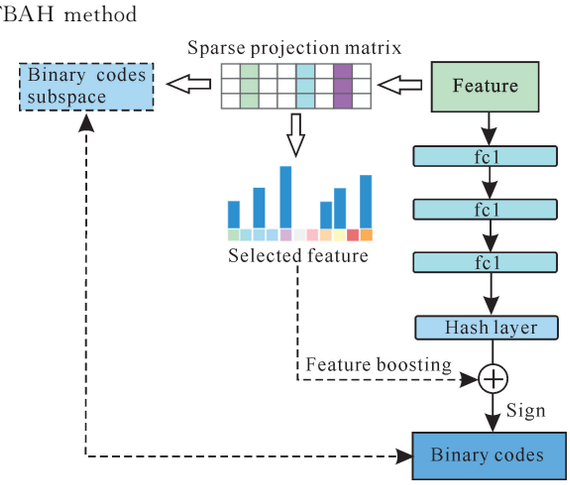

图 2 特征增强与子空间学习

Fig. 2 Feature boosting and subspace learning

2.3.1 子空间学习

为了缩小不同模态数据的异质性以及保证最终所产生的哈希码与原始特征具有一致性，FBAH 方法选择将最终生成的统一哈希码所在的空间作为一个哈希子空间。接下来，学习每个模态的投影矩阵，将不同模态的数据映射到哈希子空间。在哈希子空间中，可以测量不同模态数据与最终生成哈希码的相似性。也就是说，这是一个最小化问题，将不同模态数据映射到哈希空间的投影矩阵的目标函数定义如下：

$$\min(\|XU_x^{\mathrm{T}}-B\|_F+\|YU_y^{\mathrm{T}}-B\|_F), \quad (1)$$

其中，B 是哈希子空间中的哈希码，U_x 和 U_y 分别是图像模态和文本模态的投影矩阵。随着哈希码的不断更新，哈希子空间也会随之更新。哈希子空间学习使得模态间异质性缩小，生成质量更高的哈希码，而高质量的哈希码也会生成更好的哈希子空间，二者形成一个良性循环。

2.3.2 特征增强

经过主体网络的非线性变换,语义信息不可避免地会出现一部分丢失,为了解决这个问题,FBAH 方法通过两步去解决。首先,采用 L_{21} 范数获得稀疏的投影矩阵,同时对不同的特征空间起到特征选择的作用[8]。

方程(1)中的目标函数可以进一步写成

$$L_F = \| XU_x^{\mathrm{T}} - B \|_F + \| YU_y^{\mathrm{T}} - B \|_F + \mu_1 \| U_x \|_{21} + \mu_2 \| U_y \|_{21}, \quad (2)$$

其中,μ_1 和 μ_2 用来控制矩阵稀疏性参数。

接下来,需要考虑的是如何将选择出来的有区分度的特征融合进主网络输出的特征之中。受到 Resnet 方法[9]的启发,本文设计了一个类残差结构,FBAH 方法将全连接层和哈希层视为一个主网络,将通过稀疏投影矩阵筛选出来的特征与主网络生成的低维特征融合,来达到特征增强的目的。

2.4 哈希学习

经过特征增强后的特征记为 $h^* = f^* + \sigma m^*$,$* \in (X,Y)$,哈希码通过 sign 函数生成:$B^X = \mathrm{sign}(h^X)$,$B^Y = \mathrm{sign}(h^Y)$,其中信号函数被定义为

$$\mathrm{sign}(x) = \begin{cases} +1, x > 0 \\ -1, x \leqslant 0 \end{cases}^{\circ} \quad (3)$$

同时,为了在模态间生成一致的哈希码,获得一个包含文本和图像两个模态信息的哈希子空间,引入 $B = B^X = B^Y$,其中,$B = \mathrm{sign}(h^X + h^Y)$。最后通过最小化量化损失来保证哈希码与 h^* 的一致性,量化损失 L_q 被定义为

$$\min \| B - h^* \|_F^2, \mathrm{s.t.}\, B \in \{-1,1\}^{N \times K}。 \quad (4)$$

高质量的哈希码应该保证模态内部的判别性和模态间的一致性。接下来采用 3 个损失函数对哈希码进行约束。

2.4.1 对抗学习

为了消除不同模态增强后特征投影之间的异质性,FBAH 方法选择引入对抗学习。在对抗学习过程中,构建一个由 3 层前馈神经网络构成的判别器 D。将一个未知特征投影输入判别器 D,D 会尝试去区分输入特征投影的模态种类,即尽可能去区分输入的特征投影来自图像模态还是文本模态。同时两种来自不同模态的特征投影则尽可能去混淆判别器 D,即使其无法判别输入特征所归属的模态种类。将对抗性损失定义为 L_{adv},即

$$L_{\mathrm{adv}} = -\frac{1}{n}\sum_{i=1}^{n}(\log(D(h_i^x;\theta_D)) + \log(1 - D(h_i^y;\theta_D))), \quad (5)$$

其中,L_{adv} 可以看作是所有样本模态分类的交叉熵损失,$D(h_i^x)$ 和 $D(h_i^y)$ 则分别代表增强后的图像和文本特征的分类得分,θ_D 为判别器的参数。当判别器的分数越高,则输入特征越可能来自图像模态;分数越小,则越可能来自文本模态。

2.4.2 双重标签预测

通过双重标签预测来保证模态内不同样本特征具有语义区分性。在图像和文本模态主网络的末端连接一个带有 Softmax 函数分支的线性分类器,该分类器会在公共空间中为每一个样本生成两种不同维度的 c 维预测标签向量。

其中,分类器生成的标签投影矩阵与真实标签的误差可以被定义为

$$L_{l1}(P) = \frac{1}{n}(\| P^{\mathrm{T}}X - L \|_F + \| P^{\mathrm{T}}Y - L \|_F), \quad (6)$$

其中,$\|\ \|$ 是 Frobenius 范数,P 是线性分类器的投影矩阵。

Softmax 激活函数同时输出每个语义类别的概率分布。使用 $\hat{p}$ 来表示预测每个样本类别的概率分布,用 l_i 代表每个样本的真实标签,损失函数使用所有样本语义分类的交叉熵损失表示为

$$L_{l2}(P) = -\frac{1}{n}(l_i \cdot (\log\hat{p}_i(h_i^x) + \log\hat{p}_i(h_i^y)))。 \quad (7)$$

总的标签预测损失为 $L_l = \eta_1 L_{l1} + \eta_2 L_{l2}$。

2.4.3 余弦三元组损失

使用余弦三元组损失去保证实例间的相似性结构,使得具有相同语义的样本距离最近,不同语义的样本距离最远。以图像模态实例为例,构造一种三元组损失形式为 $(h_i^x, h_j^{y+}, h_k^{y-})$,其中文本实例 h_j^{y+} 与图像 h_i^x 有相近的语义,而 h_k^{y-} 则相反。测量相似度的方法是使用余弦距离,图像模态的余弦三元组损失被定义为

$$L_{\mathrm{ctl}}^x = \sum_{i,j,k}\max(\cos(h_i^x, h_j^{y+}) - \cos(h_i^x, h_k^{y-}) + m, 0), \quad (8)$$

同样地,文本模态的余弦三元组损失被定义为

$$L_{\mathrm{ctl}}^y = \sum_{i,j,k}\max(\cos(h_i^y, h_j^{x+}) - \cos(h_i^y, h_k^{x+}) + m, 0), \quad (9)$$

其中,m 为边缘参数。

总的余弦三元组损失为 $L_{\mathrm{ctl}} = L_{\mathrm{ctl}}^x + L_{\mathrm{ctl}}^y$。

2.5 优化

优化的总体目标函数为

$$\min_{\theta,\theta_D,B,P,U} L_{\text{Total}} = L_F + L_{\text{ctl}} + L_l + \beta L_q - \gamma L_{\text{adv}}, \text{s.t.} B^* \in \{-1,1\}^{N\times K}, \quad (10)$$

其中,参数 θ 是主体网络的参数,θ_X 和 θ_Y 是参数 θ 的一部分,分别表示图像模态和文本模态网络的参数;θ_D 是对抗网络的参数;P 代表标签预测分类器的参数;U 表示稀疏投影矩阵的参数,U_X 和 U_Y 是参数 U 的一部分,分别表示图像模态和文本模态网络的参数。本文采用 Adam 优化算法[23]来优化总体目标函数,优化过程的细节在以下算法中总结。

算法 FBAH 模型的算法

输入:图片集合 X,文本集合 Y,标签集合 L;

输出:网络参数 θ,判别器参数 θ_D,线性分类器参数 P,稀疏投影矩阵 U 和二进制码 B 。

①初始化参数和二进制码 B;

② 重复

③ for t iteration do

④ 通过后向传播算法更新参数 θ_D:

$$\theta_D \leftarrow \theta_D - \tau \nabla_{\theta_D} \frac{1}{n} L_{\text{adv}};$$

⑤ 通过后向传播算法更新参数 P:

$$P \leftarrow P - \tau \nabla_P \frac{1}{n} L_l;$$

⑥ 通过后向传播算法更新 U:

$$U \leftarrow U - \tau \nabla_U \frac{1}{n} L_F;$$

⑦ 通过后向传播算法更新 θ:

$$\theta \leftarrow \theta - \tau \nabla_\theta \frac{1}{n} L_{\text{Total}}$$

⑧ end for

⑨ end

⑩ 通过(4)更新 B

⑪ until convergence

3 验证实验

采用 MIRFlickr25K[24] 和 NUSWIDE[25] 两个被广泛应用于跨模态检索任务中的数据集进行实验,来验证 FBAH 方法的有效性。实验在 PyTorch 环境使用 GeForce RTX 2080 Ti GPU 进行,batch_size 设定为 128,学习率为 $5e-5$,$\eta_1 = 1.5e+3$,$\eta_1 = 1.5e+4$,$\beta = \gamma = 1$。

3.1 数据集及评价标准

MIRFlickr25K 数据集[24]包含 24 个类别,共计 25 000 个图像-文本对。其中每个实例最少具有 1 个标签。在实验中,笔者选择的 20 015 个实例至少有 20 个文本标签。文本模态中的每个实例被表示为 1 386 维的词袋向量(BOW)。本文的测试集为 2 000 幅图像-文本对,均为从 MIRFlickr25K 数据集中随机选择。另外,本文将其余的样本作为检索数据库,然后将从中随机抽取的 10 000 个实例作为训练集。

NUSWIDE 数据集[25]包含 81 个类别,共有 269 648 个图像,每张图片均带有相关的文本描述。每个实例至少有 1 个标签。与 AGAH 方法[11]一致,本文选择了包含最多样本的 21 个类别,共 195 834 个图像-文本对。本文的测试集包含 2 100 个图像-文本对,均为从数据集中随机抽取。另外,本文将其余的样本作为检索数据库,并从中随机抽取 10 000 个实例作为训练集。其中,每个实例的文本用 1 000 维词袋向量表示。

评价标准:本文采用跨模态检索中广泛应用的两个指标作为评价标准。

平均查准率均值(mean Average Precision,mAP):对查询样本和所有返回的检索样本之间进行余弦相似度的计算,综合考虑了排序信息和精度。

基线:与 7 种当前先进的哈希方法作比较,包括 SePH[26]、DCMH[6]、PRDH[7]、CHN[27]、SSAH[16]、AGAH[11]、DADH[17]。

3.2 对比实验

详细实验结果见表 1 和表 2,FBAH 方法在这两个数据集上的 mAP 值明显优于所比较的哈希方法,这证实了它的跨模态检索的有效性。

在 MIRFlickr25K 数据集上,对于 Image2Text 任务,FBAH 比其他方法中表现最好的方法在 16 bit、32 bit、64 bit 3 种不同的哈希码位数上,分别提升了 1.14%、2.31%和 2.26%;对于 Text2Image 任务,FBAH 方法则分别提升了 1.08%、2.40%和 2.63%。总体而言,随着哈希码位数的增加,FBAH 方法的检索效果也随之增加,这可能是因为较长的哈希码能够容纳更多的特征信息。

FBAH 方法在 NUSWIDE 数据集上的 mAP 值有明显的提升,对于 Image2Text 任务,在 16 bit、32 bit、64 bit 3 种不同的哈希码位数上,FBAH 方法分别提升了 3.93%、2.59%和 3.11%,对于 Text2Image 任务在则分别提升了 5.67%、4.83%和 4.83%。总体而言,FBAH 方法在 NUS-WIDE 数据集上有着更显著的提升。另外,在哈希码为 16 bit 和

64 bit 上的提升大于 32 bit。可能是其他方法的哈希码在 16 bit 时,由于不能容纳足够的有效信息导致哈希码质量不高。同样地,在更长的 64 bit 位时,其他方法因为哈希码包含的特征信息稀疏而导致哈希码质量不高,而 FBAH 方法可以在不同位数时都生成较高质量的哈希码。

表 1　不同编码长度下在 MIRFlickr25K 数据集上的 mAP 比较

Table 1　Comparison of mAP on MIRFlickr25K dataset under different encoding lengths

方法 Method	图像检索文本 Image retrieval text			文本检索图片 Text retrieval picture			平均 Average		
	16 bit	32 bit	64 bit	16 bit	32 bit	64 bit	16 bit	32 bit	64 bit
SePH	0.670 7	0.673 6	0.674 7	0.690 7	0.696 1	0.699 0	0.680 7	0.684 9	0.679 8
DCMH	0.720 1	0.728 7	0.741 8	0.753 3	0.759 3	0.773 0	0.724 4	0.744 0	0.757 4
PRDH	0.722 1	0.739 5	0.755 0	0.754 8	0.763 8	0.777 2	0.738 5	0.751 7	0.766 1
CHN	0.763 2	0.777 6	0.789 2	0.763 6	0.780 7	0.790 6	0.763 4	0.779 2	0.789 9
SSAH	0.770 9	0.784 0	0.792 4	0.768 1	0.773 7	0.783 8	0.769 5	0.778 9	0.788 1
AGAH	0.792 3	0.794 5	0.806 9	0.788 7	0.790 4	0.804 9	0.790 5	0.792 5	0.805 9
DADH	0.802 0	0.807 2	0.817 9	0.792 0	0.795 9	0.806 4	0.797 0	0.801 6	0.812 2
FBAH	0.813 4	0.830 3	0.840 5	0.802 8	0.819 9	0.832 7	0.808 1	0.825 1	0.836 6

表 2　不同编码长度下在 NUSWIDE 数据集上的 mAP 比较

Table 2　Comparison of mAP on NUSWIDE dataset under different encoding lengths

方法 Method	图像检索文本 Image retrieval text			文本检索图片 Text retrieval picture			平均 Average		
	16 bit	32 bit	64 bit	16 bit	32 bit	64 bit	16 bit	32 bit	64 bit
SePH	0.506 5	0.514 0	0.518 9	0.533 4	0.543 7	0.549 9	0.520 0	0.528 9	0.534 4
DCMH	0.565 7	0.600 7	0.600 1	0.533 6	0.587 1	0.591 6	0.549 7	0.593 9	0.595 9
PRDH	0.592 9	0.633 3	0.624 3	0.593 9	0.609 8	0.600 6	0.593 4	0.621 6	0.612 5
CHN	0.602 8	0.608 0	0.642 2	0.608 5	0.626 3	0.611 2	0.605 7	0.617 2	0.626 7
SSAH	0.602 0	0.621 9	0.646 3	0.612 3	0.636 9	0.639 8	0.607 2	0.629 4	0.643 1
AGAH	0.645 5	0.660 0	0.651 2	0.631 3	0.642 2	0.633 6	0.638 4	0.651 1	0.651 2
DADH	0.649 2	0.666 2	0.666 4	0.650 1	0.667 9	0.680 8	0.657 7	0.667 1	0.673 6
FBAH	0.688 5	0.692 1	0.697 5	0.706 8	0.716 2	0.729 1	0.697 7	0.704 2	0.713 3

3.3　消融实验

为了验证 FBAH 模型各个模块的有效性,设计了 4 个消融实验来进行验证:(a)FBAH-1 表示不进行特征筛选,但保留类残差结构,将未经筛选的特征与主网络输出的特征进行融合。(b)FBAH-2 表示不进行特征增强,即去掉类残差结构。(c)FBAH-3 表示去掉标签分类器。(d)FBAH-4 表示去掉特征学习部分子空间学习的损失。

表 3 展示了 4 个消融实验的结果。对于 FBAH-1 来说,去掉特征筛选,导致许多无关信息被融合进新的特征之中,所以精度会有所下滑。FBAH-2 的结果说明了设计的特征增强模块的有效性。FBAH-3 的结果表明,去掉标签预测模块,可能会导致部分特征与标签无法对齐,使精度有所下滑。而 FBAH-4 的结果则说明了以哈希码构筑的子空间学习的有效性。

表 3　64 位哈希码下消融实验中 mAP 结果对比

Table 3　mAP results of the ablation experiments with 64 bit hash codes

方法 Method	MIRFlickr25K		NUSWIDE	
	图像检索文本 Image retrieval text	文本检索图像 Text retrieval picture	图像检索文本 Image retrieval text	文本检索图像 Text retrieval picture
FBAH-1	0.838 8	0.827 5	0.691 4	0.716 7
FBAH-2	0.836 6	0.829 4	0.694 3	0.715 3
FBAH-3	0.838 9	0.827 9	0.696 7	0.698 7
FBAH-4	0.836 8	0.828 0	0.688 4	0.713 3
FBAH	0.840 5	0.832 7	0.697 5	0.729 1

4　结论

本文提出了一种新的跨模态哈希方法,被称为特征增强对抗跨模态哈希(Feature Boosting Adversarial Hashing for Cross-Modal,FBAH)方法。FBAH 方法不仅可以消除模态间的异质性差异,还能够通过特征增强来弥补主网络造成的特征损失。笔者将包含了文本和图像两个模态信息的哈希子空间作为一个学习对象进行子空间学习,在减少模态间差异的同时,还能够使得生成的哈希码与高维特征的相似度更高。此外,笔者设计了一个类残差结构去弥补特征在深度网络中损失的信息,使得提升后的特征包含更多高维信息。为了生成高质量的哈希码,使用 3 种约束来保证生成哈希码模态间的一致性、模态内的判别性以及项目间的相似性。在两个基准数据集上进行验证,结果表明,在跨模态检索任务中,FBAH 方法优于当前 7 种先进的方法(表 2)。当然,本文还存在一些尚待改善的部分,比如特征融合和特征提取的方法尚待完善,后续将采取效果更好的特征提取和特征融合方式来提升精度、完善方法。

参考文献

[1] BRONSTEIN M M,BRONSTEIN A M,MICHEL F,et al. Data fusion through cross-modality metric learning using similarity-sensitive hashing [C]//Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco,CA,USA:IEEE,2010:3594-3601.

[2] DING G G,GUO Y C,ZHOU J L. Collective matrix factorization hashing for multimodal data [C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus,OH,USA:IEEE,2014:2075-2082.

[3] CAO Y,LONG M,WANG J,et al. Deep visual-semantic hashing for cross-modal retrieval [C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco,California,USA:ACM,2016:1445-1454.

[4] ERIN LIONG V,LU J W,TAN Y P,et al. Cross-modal deep variational hashing [C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice,Italy:IEEE,2017:4077-4085.

[5] YAN C,BAI X,WANG S,et al. Cross-modal hashing with semantic deep embedding [J]. Neurocomputing,2019,337:58-66.

[6] JIANG Q Y,LI W J. Deep cross-modal hashing [C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu,HI,USA:IEEE,2017:3232-3240.

[7] YANG E,DENG C,LIU W,et al. Pairwise relationship guided deep hashing for cross-modal retrieval [C]//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco,California,USA:AAAI Press,2017:1618-1625.

[8] WANG K Y,HE R,WANG L,et al. Joint feature selection and subspace learning for cross-modal retrieval [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2016,38(10):2010-2023.

[9] HE K M,ZHANG X Y,REN S Q,et al. Deep residual learning for image recognition [C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas,NV,USA:IEEE,2016:770-778.

[10] ZHONG F M,CHEN Z K,MIN G Y,et al. A novel strategy to balance the results of cross-modal hashing [J]. Pattern Recognition,2020,107:107523. DOI:10.1016/j.patcog.2020.107523.

[11] GU W,GU X,GU J,et al. Adversary guided asymmetric hashing for cross-modal retrieval [C]//Proceedings of the 2019 on International Conference on Multimedia Retrieval. Ottawa ON,Canada:ACM,2019:159-167.

[12] HARDOON D R,SHAWE-TAYLOR J. Canonical correlation analysis:An overview with application to learning methods [J]. Neural Computation,2004,16(12):2639-2664.

[13] ROSIPAL R,KRAMER N. Overview and recent advances in partial least squares [C]//International Statistical and Optimization Perspectives Workshop "Sub-

space,Latent Structure and Feature Selection". Bohinj, Slovenia:Springer,2005:34-51.

[14] ZHAI X H,PENG Y X,XIAO J G. Learning cross-media joint representation with sparse and semisupervised regularization [J]. IEEE Transactions on Circuits and Systems for Video Technology,2014,24(6):965-978.

[15] WANG B K,YANG Y,XU X,et al. Adversarial cross-modal retrieval [C]//Proceedings of the 25th ACM International Conference on Multimedia. Mountain View, California,USA:ACM,2017:154-162.

[16] LI C,DENG C,LI N,et al. Self-supervised adversarial hashing networks for cross-modal retrieval [C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT,USA:IEEE,2018:4242-4251.

[17] 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. Dublin,Ireland:ACM,2020:525-531.

[18] PENG Y X,QI J W. CM-GANs:Cross-modal generative adversarial networks for common representation learning [J]. ACM Transactions on Multimedia Computing,Communications,and Applications (TOMM), 2019,15(1):22.

[19] HU D,NIE F P,LI X L. Deep binary reconstruction for cross-modal hashing [J]. IEEE Transactions on Multimedia,2019,21(4):973-985.

[20] SONG J K,YANG Y,YANG Y,et al. Inter-media hashing for large-scale retrieval from heterogeneous data sources [C]//Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. New York,USA:ACM,2013:785-796.

[21] ZHEN L L,HU P,WANG X,et al. Deep supervised cross-modal retrieval [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, California, USA: IEEE, 2019: 10394-10403.

[22] CHATFIELD K,SIMONYAN K,VEDALDI A,et al. Return of the devil in the details: Delving deep into convolutional nets [C]//Proceedings of the British Machine Vision Conference. Nottingham, United Kingdom:BMVA Press,2014:1-11.

[23] KINGMA D P,BA L J. ADAM:A method for stochastic optimization [C]//Proceedings of the 2015 International Conference on Learning Representations(ICLR). San Diego,California,USA:arXiv.org,2015:1-15.

[24] HUISKES M J,LEW M S. The MIR flickr retrieval evaluation [C]//Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval. Vancouver, British Columbia, Canada: ACM, 2008: 39-43.

[25] 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. Santorini,Fira,Greece:ACM,2009:1-9.

[26] LIN Z J,DING G G,HU M Q,et al. Semantics-preserving hashing for cross-view retrieval [C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE,2015:3864-3872.

[27] CAO Y,LONG M S,WANG J M,et al. Correlation hashing network for efficient cross-modal retrieval [C]//Proceedings of the British Machine Vision Conference (BMVC). London, United Kingdom: BMVA Press,2017:1-8.

Feature Boosting Adversarial Hashing for Cross-Modal Retrieval

HE Pei[1],WANG Meng[2],WANG Zhuo[1],LU Guangyun[3]
(1. College of Science,Guangxi University of Science and Technology,Liuzhou,Guangxi,545000,China;2. Tus College of Digit,Guangxi University of Science and Technology,Liuzhou,Guangxi,545000,China;3. Liuzhou Institute of Technology,Liuzhou,Guangxi,545616,China)

Abstract:In cross-modal retrieval tasks,hashing method is widely used because of its high retrieval efficiency and low storage cost. However,these methods pay little attention on how to compensate for the loss of feature information in the process of transforming high-dimensional features into hash codes. To solve these problems,this article proposes a Feature Boosting Adversarial Hashing for Cross-Modal (FBAH). FBAH combines subspace learning with adversarial learning to reduce the difference of different modal data. In addition,a module similar to residual structure is constructed,which can bypass the main network and directly input the selected distinguishing features into the hash space for feature boosting. In this way,the generated hash code can have more original feature information. Finally,the linear classifier with branch network is used to make two kinds of prediction in label space,and the gap with the real label is minimized to ensure the invariance of semantics. In this article,two large datasets commonly used in cross-modal retrieval tasks are selected for a large number of experiments. The results show that the performance of FBAH is superior to seven existing advanced cross-modal hashing methods.

Key words:feature boosting;cross-modal retrieval;sparse matrix;hashing subspace learning;adversarial learning

责任编辑:陆媛峰