计算机科学
COMPUTER SCIENCE

基于双门控-残差特征融合的跨模态图文检索

张昌凡, 马远远, 刘建华, 何静

引用本文

张昌凡, 马远远, 刘建华, 何静.基于双门控-残差特征融合的跨模态图文检索[J]. 计算机科学, 2023, 50(6A): 220700030-7.

ZHANG Changfan, MA Yuanyuan, LIU Jianhua, HE Jing. Dual Gating-Residual Feature Fusion for Image-Text Cross-modal Retrieval [J]. Computer Science, 2023, 50(6A): 220700030-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220700030

基于双门控-残差特征融合的跨模态图文检索

张昌凡[1]　马远远[1]　刘建华[2]　何　静[1]
1 湖南工业大学电气与信息工程学院　湖南 株洲 412007
2 湖南工业大学轨道交通学院　湖南 株洲 412007
(zhangchangfan@263.net)

摘　要　由于互联网和社交媒体的快速发展,跨模态检索引起了广泛关注,跨模态检索学习的目的是实现不同模态的灵活检索。不同模态数据之间存在异质性差距,不能直接计算不同模态特征的相似度,使得跨模态检索任务的准确率很难提高。为缩小图像和文本数据间的异质性差距,文中提出了一种双门控-残差特征融合的跨模态图文检索方法(DGRFF),该方法通过设计门控特征和残差特征来融合图像模态和文本的特征,能够从相反的模态中获得更有效的特征信息,使得语义特征信息更全面。同时,采用对抗损失来对齐两个模态特征的分布,以保持融合特征模态不变性以及在公共潜在空间中得到更有辨识力的特征表示。最后,联合标签预测损失、跨模态相似性损失和对抗损失对模型进行训练学习。在 Wikipedia 和 Pascal Sentence 数据集上进行实验,结果证明,DGRFF 在跨模态检索任务上获得了良好的效果。

关键词:跨模态检索;异质性差距;门控特征;残差特征;特征融合
中图法分类号　TP391

Dual Gating-Residual Feature Fusion for Image-Text Cross-modal Retrieval

ZHANG Changfan[1],MA Yuanyuan[1],LIU Jianhua[2] and HE Jing[1]
1 School of Electrical and Information Engineering,Hunan University of Technology,Zhuzhou,Hunan 412007,China
2 School of Rail Transit,Hunan University of Technology,Zhuzhou,Hunan 412007,China

Abstract　Due to the rapid development of the Internet and social media,cross-modal retrieval has attracted extensive attention. The purpose of cross-modal retrieval is to achieve flexible retrieval of different modalities. The heterogeneity gap between different modal suggests that the similarity of different modal features cannot be calculated directly,making it difficult to improve the accuracy of cross-modal retrieval. This paper proposes an image-text cross-modal retrieval method for dual gating-residual feature fusion(DGRFF),to narrow the heterogeneity gap between the image and text. By designing gating features and residual features to fusion the features of image modality and text modality,this method can gain more effective feature information from the opposite modality,making semantic feature information more comprehensive. At the same time,the adversarial loss is adopted to align the feature distribution of the two modalities,to maintain the modality invariance of the fusion feature and obtain a more recognizable feature representation in the public potential space. Finally,the model is trained by combining label prediction loss,cross-modal similarity loss and adversarial loss. Experiments on Wikipedia and Pascal Sentence datasets show that DGRFF performs well on cross-modal retrieval tasks.

Keywords　Cross-modal retrieval,Heterogeneity gap,Gating features,Residual features,Feature fusion

1　引言

随着互联网移动设备的快速发展,数据形式变得多种多样,人们在日常的互联网应用中面临的大数据已由单模态数据形式转移到多模态数据形式,例如图像、文本、音频等多模态数据,多模态数据可以增强人们对同一事物的全方位认识。跨模态检索[1]是多模态数据的应用之一,通过一个模态(图像)检索到与之相关联的另一个模态(文本),由于不同模态的数据之间的异构差异性导致难以学习不同模态数据之间的相关性。跨模态检索的目标是弥合不同模态之间的异构差异,通过对不同模态之间的关系进行建模,学习公共表示,再使用度量函数来衡量跨模态数据之间的相似性,从而实现有效的跨模态检索学习。

目前已经提出了一些方法,这些方法大多数集中在表示学习上,通常将不同模态数据嵌入到公共空间中,直接度量不同模态数据之间的相似性。这些方法可以分为两类:浅层方法[2-4]和深度学习方法[5-7]。浅层方法大多通过优化统计值来学习不同模态数据的映射矩阵,然后将映射矩阵投影到一个

基金项目:国家自然科学基金(52172403,62173137,52272347);湖南省自然科学基金(2021JJ50001,2021JJ30217)
This work was supported by the National Natural Science Foundation of China(52172403,62170137,52272347) and Natural Science Foundation of Hunan Province,China(2021JJ50001,2021JJ30217).
通信作者:何静(hejing@263.net)

公共空间，如典型相关性分析(Canonical Correlation Analysis，CCA)[8]是浅层跨模态检索方法之一，它的目标是通过最大化两个模态投影之间的相关性来寻找公共子空间；继而又在CCA上进行改进，提出核典型相关性分析方法(Kernel Canonical Correlation Analysis，KCCA)[9]，它是一种无监督的方法，通过引入核函数将CCA进行扩展，利用核函数将跨模态数据的相关性最大化后投影到语义空间，解决数据的非线性问题。虽然它们在跨模态检索上表现出了良好的性能，但是这类方法在子空间学习到的是线性映射，所学习的空间有一定的复杂度，且学习到的特征是浅层的特征表示，无法有效地学习到不同模态的高维语义信息。

随着深度学习的流行，深度神经网络(Deep Neural Networks，DNN)模型在图像识别、自然语言处理等单模态任务中展示了处理不同模态数据的能力，进而这些模型逐渐被用到跨模态检索任务之中。相比浅层方法，深度学习方法拥有强大的特征提取能力，能够提取不同模态的有效表示，建立高层语义相关性。深度学习的跨模态检索方法可以分为实值表示学习方法和二进制表示学习方法，实值表示学习能直接学习不同模态的特征，旨在学习一个实值公共表示空间。Lee等[10]提出了一种堆叠交叉注意力方法(Stacked Cross Attention，SCAN)，用于发现图像和文本完整的潜在对齐，使得图像-文本匹配更具可解释性，并使用句子中的图像区域和单词作为上下文推断图像-文本两者的相似性。Cornia等[11]采用一种不同于学习联合多模态嵌入空间的方法，将图文跨模态检索转变为图文数据相互转换的问题，利用文本特征和图片特征的相互翻译建立重构约束。Wang等[12]提出一种对抗性跨模态检索(Adversarial Cross-Modal Retrieval，ACMR)，通过将生成对抗网络用到跨模态检索学习中，特征投影器在子空间中生成模态不变性的特征表示，而模态分类器则区分来自特征投影器生成的不同模态的特征表示。Ren等[13]提出模态内自注意力距离度量方法(Intra-modal Self-attention Distance，ISD)，通过测量文本和图像之间的语义距离来量化两者关系的一致性。Peng等[14]提出一种特定于模态特定特征的跨模态相似性距离度量方法(Modality-specific Cross-modal Similarity Neasurement，MCSM)，为每个模态构建独立的语义空间，通过循环注意力网络充分利用一个模态的模态特定特征，另外一个模态通过联合注意力投影到公共空间中。不同于实值表示学习的方法，二进制表示学习旨在将不同模态的数据映射到一个共同的汉明空间中，如Cao等[15]提出一种端到端的深度语义哈希模型(Deep Visual-Semantic Hashing，DVSH)，该模型构成一个学习图像和文本句子联合嵌入的视觉语义融合网络，以及利用哈希函数生成两个特定模态的紧凑哈希编码网络。Lin等[16]提出了语义深度跨模态哈希方法(Semantic Deep Cross-modal Hashing，SDCH)，利用语义标签分支对特征学习部分进行改进，生成更具辨识度的哈希编码。上述基于深度学习的方法在跨模态检索任务上获得了良好的效果。但是，在跨模态数据中，如何更进一步缩小跨模态数据的异质性差距还需要继续进行研究。

在实值表示学习的基础上，本文提出了一种双门控-残差特征融合的跨模态图文检索方法，为了更好地缩小跨模态数据之间的异质性差距，采用双门控-残差特征融合的方法将特征进行融合，充分利用图像模态和文本模态的特征信息，使得图文特征信息的更全面。另外，为减小不同模态数据之间的分布差异，使用对抗损失对多模态融合特征进行模态对齐，以保证融合特征的模态不变性，本文工作的主要贡献如下：

(1)采用双门控-残差特征融合方法，构建图像融合特征的门控特征和残差特以及文本融合特征的门控特征和残差特征，以进行跨模态检索；

(2)使用对抗损失来调整两种模态特征的分布，使得图像融合特征和文本融合特征的分布难以区分，从而保持两者之间的模态不变性；

(3)在两个基准数据集上得到的实验结果表明，本文方法明显优于浅层方法，并且相比深度学习的几个主流方法获得了良好的效果。

本文第2节回顾了跨模态学习的相关工作；第3节介绍了所提方法，包括问题定义、模型框架、门控-残差特征融合网络、目标函数定义；第4节提供了实验结果与分析；最后总结全文。

2 相关工作

跨模态检索的目标是从不同的模态中检索到相关实例。检索到相关实例的方法主要是学习一个公共表示空间，然后直接测量来自不同模态数据间的相似性。Hao等[17]提出了一种对抗性跨模态嵌入框架(Adversarial Cross-Modal Embedding，ACME)，用于解决跨模态检索任务，其通过权重共享的方式，将图像特征和文本特征表示到一个公共空间得到最终嵌入特征，并将其用于跨模态检索任务中。Wu等[18]提出了一种模态特定特征和模态共享特征生成对抗网络(Modality-Specific and Shared Generative Adversarial Network，MS2GAN)，用于跨模态检索学习，由一个子网络学习每个模态的特定特征，另一个公共子网络用来学习每个模态的共享特征，也就是学习一个潜在公共空间，最后将特定特征和共享特征表示联合用于检索任务中。Zhen等[19]提出了深度监督跨模态检索(Deep Supervised Cross-modal Retrieval，DSCMR)，旨在找到一个公共空间，在公共的特征空间中直接对来自不同模态的样本进行相似性度量，同时还建立一个标签空间来指导学习。这些方法通过学习一个潜在的公共空间，在跨模态检索任务中得到了不错的效果。

虽然通过学习潜在公共空间的方式能够缩小不同模态数据之间的异质性差距，但想要进一步缩小不同模态数据之间的异质性差距，提高检索性能，还需要进行探索。一些方法通过多模态融合的方式来解决这一问题，Guo等[20]在图像-音频检索任务中，设计了一个深度视觉音频网络(Deep Visual-Audio Network，DVAN)将图像和音频进行融合，通过学习一个多模态特征来判断图像音频对的相关性，以增强图像与音频的匹配度。Beltran等[21]提出了一种深度多模态嵌入模型(Deep Multimodal Embeddings，DME)，它将视觉问答(Visual-Question-Answering，VQA)用于深度多模态学习，通过结合视觉特征和文本特征得到多模态融合特征，将其进行信息检索，大大提高了检索性能。Dong等[22]提出一种跨模态图像注意力方法，用于生成每个样本的图注意力表示，图像注意力机制可从相反模态中获得更有效的特征，同时还采用特征融合方法缩小异构鸿沟。Wang[23]将关系网络整合到跨模态学习中，提出了深度关系相似性学习(Deep Relational Simi-

larity Learning，DRSL）跨模态检索方法，将图像和文本融合后输入到关系网络中计算图像-文本的相似度，提高跨模态检索性能。Vo 等[24]在图像检索任务中，针对学习目标图像与源图像源文本之间的相似性度量问题，提出了一种组合图像和文本的方法（Text Image Residual Gating，TIGR），利用文本特征来修改图像特征，得到源图像融合特征，最终与目标图像进行相似性度量。Dong 等[25]将 TIRG 方法用到中国传统绘画的图像检索任务中，取得了不错的效果，表明融合不同模态的特征能够有效地减小不同模态间的语义鸿沟，进一步提高检索性能。

另外，来自不同模态的编码特征分布可能不一致，在解决不同模态数据分布不一致的问题上，Peng 等[26]提出了一种跨模态生成对抗网络（Cross-modal Generative Adversarial Networks，CM-GANS），利用其来模拟不同模态数据的联合分布，以促进跨模态相关学习。Hao 等[17]提出一种对抗性跨模态嵌入方法，用于解决食品领域跨模态检索任务，使用对抗性损失来对齐两种模态的嵌入，从而无法区分两个模态的特征分布。

根据上述方法，本文考虑将学习潜在公共空间中的特征表示和特征融合方法相结合，同时利用对抗损失来减小两个模态的分布差异并保持模态不变性，进一步减小跨模态数据间的异质性差距，提高检索精度。

3 所提方法

3.1 问题定义

本节将详细介绍本文的问题定义、模型框架、门控-残差特征融合网络公式定义以及目标函数定义。

本文在两个广泛使用的数据上进行实验，数据集包含图像数据和文本数据，它们分别表示为 X 和 Y。令集合 $Q=\{o_i,l_i\}_{i=1}^{n}$ 为 n 个跨模态数据集的图像-文本对，其中 $o_i=(x_i,y_i)$ 上的每个实例包括一个图像特征向量 $\boldsymbol{x}_i\in R^{d_x}$ 和文本特征向量 $\boldsymbol{y}_i\in R^{d_y}$，$d_x$ 和 d_y 表示图像和文本两个模态各自的维度。$l_i=\{l_{i1},l_{i2},\cdots,l_{ic}\}$ 是对应于 o_i 的语义标签，其中 c 是类别数量。如果 o_i 属于第 c 个类别，那么 $l_{ic}=1$，否则 $l_{ic}=0$。图像特征矩阵、文本特征矩阵以及标签矩阵分别为 $X=[x_1,x_2,\cdots,x_n]\in R^{n\times d_x}$，$Y=[y_1,y_2,\cdots,y_n]\in R^{n\times d_y}$ 和 $L=[l_1,l_2,\cdots,l_n]\in R^{n\times c}$。

3.2 模型框架

模型框架如图 1 所示，模型框架主要包含 4 个部分：1）输入数据；2）特征表示网络；3）多模态融合网络；4）标签预测。

本文方法中图像和文本两个模态分别采用两种不同的特征提取模型。首先，图像的原始特征表示是由 19 层 VGG-Net[27]的 fc7 层中得到图像特征向量，维度为 4 096；原始文本特征表示由词袋预训练模型（Bag of Words，BoW）提取，得到文本特征向量，维度为 300。其次，将原始图像特征和原始文本特征分别输入到由全连接层组成的 ImageNet 和 TextNet 网络中，得到最终的图像特征表示和文本特征表示，维度为 1 024，随后将 1 024 维的图像特征和文本特征输入到融合网络中进行特征融合，得到图像融合特征和文本融合特征，再将两个融合特征在最后一层进行权重共享，确保同一类别的图像和文本在公共空间中具有相似的表示。最后，将图像融合特征和文本融合特征进行标签预测，以确保类别具有区分性。

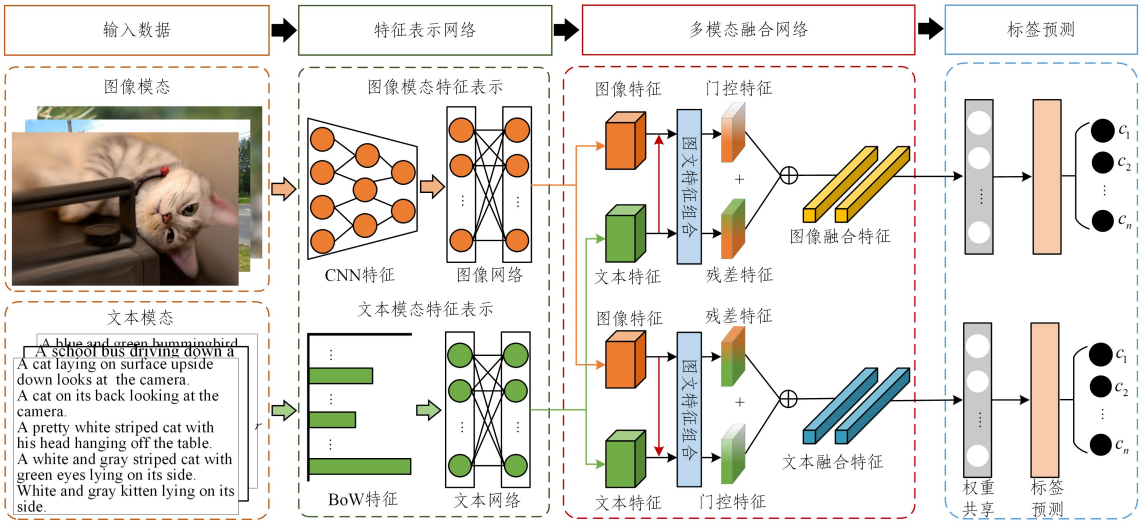


图 1 所提方法的框架图

Fig. 1 Framework diagram of the proposed method

3.3 门控-残差特征融合网络

多模态融合可有效利用不同模态数据的特征信息，使得多模态特征表示的特征信息更全面，本文采用 Vo 等[24]提出的特征融合方法获得融合特征。该方法主要通过文本特征来“修改”图像特征，进而得到一个图像融合特征。Dong 等[25]提及在设计门控特征和残差特征的过程中，门控特征则需要尽可能保留原始图像数据中的特征元素，而残差特征则需要结合文本数据中更多的特征元素。可以说使用此方法能够从相反的模态中得到更有用的特征，最后将两者进行融合。本文不仅用文本特征来“修改”图像特征，还需要用图像特征来“修改”文本特征，得到重构的融合特征，即图像融合特征和文本融合特征，将两者用于跨模态检索学习。

（1）图像融合特征

门控特征的计算方式为：

$$f_{gx}(\phi_x,\phi_y)=\sigma(W_{g2}*RELU(W_{g1}*([\phi_x,\phi_y])))\odot\phi_x \tag{1}$$

其中，σ 为 sigmoid 函数，W_{g1} 和 W_{g2} 为 3 * 3 的卷积滤波器，$RELU$ 是线性矫正单元，$\odot$ 是元素的乘积符号，$\boldsymbol{\phi}_x$ 是图像

特征向量，ϕ_y 是文本特征向量。

残差特征的计算方式为：

$$f_{rx}(\phi_x,\phi_y)=W_{r2}*RELU(W_{r1}*([\phi_x,\phi_y])) \tag{2}$$

其中，W_{r1} 和 W_{r2} 为 3*3 的卷积滤波器。

将式(1)和式(2)组合得到：

$$Q_x=l_{gx}f_{gx}(\phi_x,\phi_y)+l_{rx}f_{rx}(\phi_x,\phi_y) \tag{3}$$

其中，Q_x 为图像融合特征，l_{gx} 和 l_{rx} 分别为 f_{gx} 和 f_{rx} 的权重。

(2)文本融合特征

门控特征与残差特征的计算方式与式(1)和式(2)同理：

$$f_{gy}(\phi_x,\phi_y)=\sigma(W_{g4}*RELU(W_{g3}*([\phi_x,\phi_y])))\odot\phi_y \tag{4}$$

$$f_{ry}(\phi_x,\phi_y)=W_{r4}*RELU(W_{r3}*([\phi_x,\phi_y])) \tag{5}$$

其中，W_{g3}，W_{g4}，W_{r3} 和 W_{r4} 都是 3*3 的卷积滤波器。

将式(4)和式(5)组合得到：

$$Q_y=l_{gy}f_{gy}(\phi_x,\phi_y)+l_{ry}f_{ry}(\phi_x,\phi_y) \tag{6}$$

其中，Q_y 为文本融合特征，l_{gy} 和 l_{ry} 分别为 f_{gy} 和 f_{ry} 的权重。

在门控-残差特征融合网络中采用全连接层作为融合网络，因此 W_{g1}，W_{g2}，W_{r1}，W_{r2}，W_{g3}，W_{g4}，W_{r3}，W_{r4} 的值设置为 1。

3.4 目标函数

为优化网络模型，本文主要从模态间不变以及模态内可判别性来定义目标函数。如何使融合的多模态特征在模态内仍然具有区分性，本文使用两个线性分类器来对多模态融合特征进行标签预测，目标函数的定义如下：

$$\varsigma_1=\frac{1}{n}\|B^{\mathrm{T}}Q_X-L\|_{\mathrm{F}}+\frac{1}{n}\|B^{\mathrm{T}}Q_Y-L\|_{\mathrm{F}} \tag{7}$$

其中，$Q_X=[Q_{x_1},Q_{x_2},\cdots,Q_{x_n}]$ 为图像融合特征，$Q_Y=[Q_{y_1},Q_{y_2},\cdots,Q_{y_n}]$ 为文本融合特征，B^{T} 为模态内可判别网络中的参数，$\|\cdot\|_F$ 为 Frobenius 范数。

同时，使用跨模态相似性损失函数计算公共潜在空间中的图像模态和文本模态所有样本的相似性，即图像-图像、文本-文本、图像-文本的相似性。目标函数的定义如下：

$$\zeta(u,v)=\frac{1}{n^2}\sum_{i,j=1}^{n}(\log(1+\mathrm{e}^{d(u_i,v_j)})-S(u_i,v_j)d(u_i,v_j)) \tag{8}$$

其中，$d(u_i,v_j)$ 表示 u_i 和 v_j 之间的相似性距离度量，$d(\cdot)$ 为余弦函数；$S(u_i,v_j)$ 表示预测标签的相似度，如果预测标签相似则为 1，预测标签不相似则为 0。

每个样本对的损失总和为：

$$\varsigma_2=\zeta(Q_X,Q_X)+\zeta(Q_Y,Q_Y)+\zeta(Q_X,Q_Y) \tag{9}$$

其中，$\zeta(Q_X,Q_X)$ 为图像-图像的相似性损失，$\zeta(Q_Y,Q_Y)$ 为文本-文本的相似性损失，$\zeta(Q_X,Q_Y)$ 为图像-文本的相似性损失。

不同模态的编码特征会使得编码特征的分布不一致，需要对齐图像和文本的编码特征分布，减小配对样本在公共空间中的特征分布差异，在本文实验中采用 WGAN 对抗损失来实行跨模态对齐，从而保持模态间的不变性，目标函数的计算式如下：

$$\varsigma_3=E_{x\sim p_{\mathrm{image}}}[\log D_M(Q_X(Q_x))]+E_{y\sim p_{\mathrm{text}}}[\log(1-D_M(Q_Y(Q_y)))] \tag{10}$$

最小-最大优化方法为：

$$\min_{Q_X,Q_Y}\max_{D_M}\varsigma_3$$

联合式(7)、式(9)－式(10)，总目标函数为：

$$\varsigma=\varsigma_1+\beta*\varsigma_2+\gamma*\varsigma_3 \tag{11}$$

DGRFF 检索模型的训练过程如算法 1 所示。

算法 1 DGRFF 检索模型算法

输入：图像集 X，文本集 Y，标签集 L，批量值 b，学习率 α，最大 epoch 数 N。

输出：图像模态和文本模态的最终融合特征表示 Q_X 和 Q_Y

1. 通过网格搜索算法设置参数 $\beta,\gamma,l_{gx},l_{rx},l_{gy},l_{ry}$
2. for i=1 开始 do
   2.1. 从图文集合中随机抽取 b 个图文样本对 x_i,y_i，获得最终图文对特征表示
   2.2. 计算式(11)的结果并通过 Adam 优化器更新参数
   2.3. 更新图像融合特征 Q_x 和文本融合特征 Q_y 的权重 $l_{gx},l_{rx},l_{gy},l_{ry}$
3. 获得 Q_X,Q_Y

4 实验结果与分析

为了验证所提方法的有效性，本文在两个广泛使用的 Wikipedia[28] 数据集和 Pascal Sentence[29] 数据集上进行实验。在实验中，首先将本文方法与几种主流方法进行比较，以评估所提方法的性能，本文使用网格搜索算法设置模型中所涉及的权重参数。其次，对所提方法进行消融分析，评估各个模块在整个模型中的作用，同时对潜在空间特征表示进行可视化以及收敛性分析。

4.1 实施细节

在实验中，本文遵循文献[14]等所提的数据集划分策略，采用 19 层 VGGNet 网络提取图像特征，输出 4096 维的原始图像特征表示。采用 BoW 模型提取文本特征，输出 300 维的原始文本特征表示。同时本文的实验是在 Pytorch3 上实现的。在训练中，学习率为 10^{-4}，在 Wikipedia 数据集和 Pascal Sentence 数据集上各训练 300 个 epoch。

4.2 评价指标

本文通过计算每个图像(文本)和文本(图像)的特征之间的余弦相似性所返回的均值平均精度(Mean Average Precision，MAP)分数来展现最终的评估指标。其综合考虑了排序信息和准确率，是跨模态检索研究中广泛使用的性能评价标准[12]。在本文的实验中，进行两种类型的跨模态检索任务：1)使用图片查询检索文本(Img2Txt)；2)使用文本查询检索图像(Txt2Img)两个任务。

MAP 可以通过计算所有查询项的平均精度(Average Precision，AP)的平均值来得到，AP 的定义如下：

$$AP=\frac{1}{P}\sum_{k=1}^{N}\frac{P_k}{k}\times rel_k \tag{12}$$

其中，P 表示测试集中相关样本的数量，N 表示测试集中的样本数，P_k 表示前 k 个返回结果相关样本的数目。如果第 k 个样本与查询的样本相关，则 $rel_k=1$，否则 $rel_k=0$。

由式(12)可得 MAP 的定义如下：

$$MAP=\frac{1}{R}\sum_{k=1}^{R}AP(i) \tag{13}$$

其中，R 是查询样本的数量，$AP(i)$ 是第 i 个实例的 AP 值。

4.3 结果对比与分析

为验证本文方法的有效性，将本文方法与几种主流方法

进行比较,即 CCA[8],LCFS[30],GMA[31],ACMR[12],DRSL[23],DSCMR[19],S2MC[32] 和 MARS[33],其中 CCA,LCFS,GMA 为浅层的方法,其余 5 种方法为深度学习的方法。本文通过直接引用文献[30-32]中的结果或根据相应的模型提供的源代码进行复现得到对比结果。在所有实验中,数据集都采用 19 层 VGGNet 模型和 BoW 模型提取的图文特征。

表 1 和表 2 列出了本文方法与 8 种主流方法在两个基准数据集上的 MAP 分数,结果表明,在 Img2Txt、Txt2Img 两个任务以及两个任务的平均值 Average 中,本文方法的 MAP 分数都明显优于其他方法。具体来说,本文方法在 Wikipedia 数据集上的 Img2Txt 和 Txt2Img 上的 MAP 分数与次优方法比较分别提高了 0.272 和 0.256,对于在 Pascal sentence 数据集上的 Img2Txt 和 Txt2Img 的 MAP 分数与次优方法比较分别提高了 0.112 和 0.099。相比浅层方法,基于深度学习的方法能够学习到更有效的语义信息,提高了检索性能。

表 1　Wikipedia 数据集上的 MAP 分数比较

Table 1　Comparison of MAP scores on Wikipedia dataset

Method	Img2Txt	Txt2Img	Average
CCA[8]	0.236	0.208	0.222
LCFS[30]	0.280	0.214	0.247
GMA[31]	0.272	0.232	0.253
ACMR[12]	0.509	0.535	0.522
DRSL[23]	0.458	0.435	0.447
DSCMR[19]	0.493	0.438	0.466
S2MC[32]	0.521	0.470	0.496
MARS[33]	0.486	0.402	0.444
DGRFF w/o fusion	0.500	0.440	0.470
DGRFF	**0.793**	**0.791**	**0.792**

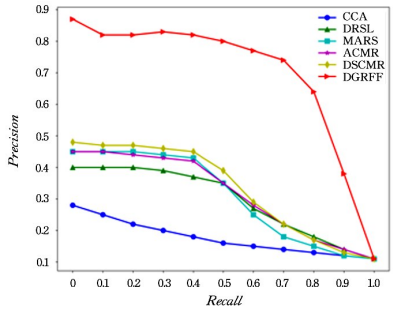
表 2　Pascal sentence 数据集上的 MAP 分数比较

Table 2　Comparison of MAP scores on Pascal sentence dataset

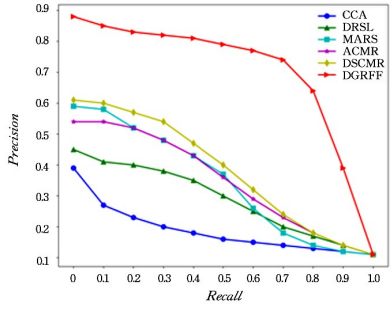
Method	Img2Txt	Txt2Img	Average
CCA[8]	0.457	0.451	0.454
LCFS[30]	0.344	0.267	0.306
GMA[31]	0.427	0.339	0.383
ACMR[12]	0.663	0.675	0.669
DRSL[23]	0.617	0.630	0.624
DSCMR[19]	0.725	0.723	0.724
S2MC[32]	0.739	0.722	0.731
MARS[33]	0.730	0.736	0.733
DGRFF w/o fusion	0.704	0.714	0.709
DGRFF	**0.851**	**0.835**	**0.843**

表 1 和表 2 中的 DGRFF w/o fusion 表示没有融合模块的 DGRFF 变体,由结果可以看到没有融合模块之后,在 Wikipedia 数据集上的 Img2Txt 任务的 MAP 分数由 0.793 降到了 0.500 以及 Txt2Img 任务的 MAP 分数由 0.791 降到了 0.440;在 Pascal sentence 数据集上的 Img2Txt 任务的 MAP 分数由 0.851 降到了 0.704 以及 Txt2Img 任务的 MAP 分数由 0.835 降到了 0.714。

图 2(a)和图 2(b)分别给出了在 Wikipedia 数据集上绘制 Img2Txt 和 Txt2Img 两个任务的 Precision-Recall 曲线。Precision 为查准率和 Recall 为查全率,两者相互矛盾。查准率高时查全率低,查全率高时查准率反而低,在查全率相同的情况下,本文方法与其他方法相比获得了更高的查准率。



(a)Img2Txt



(b)Txt2Img

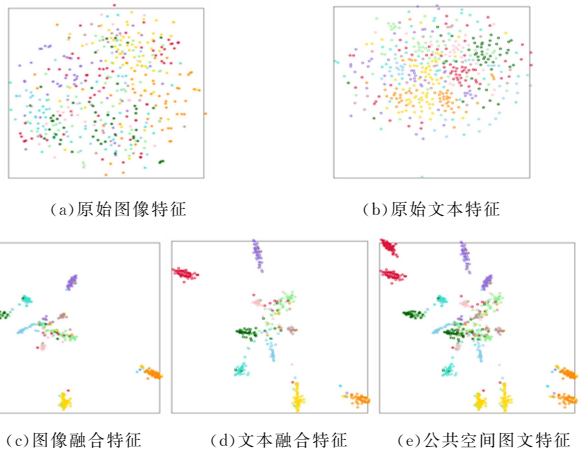
图 2　Wikipedia 数据集上的 Precision-Recall 曲线

Fig. 2　Precision-Recall curve on Wikipedia dataset

4.4　所提方法的进一步分析

4.4.1　表示学习可视化

为了更直观地研究本文方法的有效性,本文采用 t-SNE[34] 方法将 Wikipedia 数据集上的测试集数据中的图像模态和文本模态原始特征以及最终的高级特征表示嵌入到二维空间中进行可视化。星号'*'表示图像模态的样本,圆圈表示文本模态的样本。图 3(a)和图 3(b)分别为原始图像特征和原始文本特征,原始图像特征由 19 层 VGGNet 得到 4 096 维特征,原始文本特征由 Bow 模型得到 300 维特征。图 3(c)—3(e)分别为融合后的图像特征表示、融合后的文本特征表示以及公共空间中的融合图像特征和文本特征表示。从图中可以观察到,原始图像和原始文本的分布有很大的区别,具有不同类别的标签样本没有被很好地分离,而学习特征表示的分布表明图像和文本模态的分布更好地混合到了一起,并且被分为几个语义集群。但是也有少部分来自不同语义类别的特征表示混合在一起,这可能会带来不相关的检索结果。



(a)原始图像特征　(b)原始文本特征

(c)图像融合特征　(d)文本融合特征　(e)公共空间图文特征

图 3　Wikipedia 数据集上的 t-SNE 可视化

Fig. 3　T-SNE visualization on Wikipedia dataset

4.4.2 消融分析

为了评估本文方法中相关组件的有效性，将本文所提方法分为3个变体，第一个是没有标签预测损失的变体 DGRFF w/o ζ_1，第二个是没有所有模态样本的相似性损失的变体 DGRFF w/o ζ_2，第三个是没有模态不变性损失的变体 DGRFF w/o ζ_3。表3和表4列出了本文方法以及3个变体在 Wikipedia 和 Pascal sentence 上的对比结果。

表3 Wikipedia 数据集上的消融实验结果

Table 3 Ablation experiment results on Wikipedia dataset

Method	Img2Txt	Txt2Img	Average
DGRFF w/o ζ_1	0.746	0.750	0.748
DGRFF w/o ζ_2	0.788	0.772	0.780
DGRFF w/o ζ_3	0.776	0.771	0.773
DGRFF	**0.793**	**0.791**	**0.792**

表4 Pascal Sentence 数据集上的消融实验结果

Table 4 Ablation experiment results on Pascal Sentence dataset

Method	Img2Txt	Txt2Img	Average
DGRFF w/o ζ_1	0.800	0.797	0.798
DGRFF w/o ζ_2	0.838	0.833	0.835
DGRFF w/o ζ_3	0.830	0.827	0.829
DGRFF	**0.851**	**0.835**	**0.843**

从表中可以看到，本文方法在两个数据集中表现最好，3个变体的 MAP 分数明显低于完整方法的 MAP。这表明了标签预测损失、模态相似性损失、模态不变性损失3个组件对跨模态检索任务的有效性。此外，去除每个组件后，每个检索任务的性能都有所下降，去除融合模块的 MAP 分数的变化最为显著，说明多模态融合对检索性能提升的影响很大。标签预测损失对性能的影响稍逊于融合模块，但其在提升检索性能方面也存在很大的影响，另外去除模态相似性组件和去除模态不变性组件的 MAP 分数下降不如前面两个组件明显，但也说明了这两个组件对检索任务性能提升的重要性。表3和表4中的消融实验结果不仅体现了各个组件的有效性，同时也为本文方法所得结果提供了有效支撑。

4.4.3 收敛性分析

图4给出了在 Pascal Sentence 数据集上的目标函数值在不同 epoch 的变化曲线。从图中可以观察到，随着 epoch 数增加，目标函数值在训练过程中的曲线相对平滑，并且能迅速下降，很快地收敛到一个点。

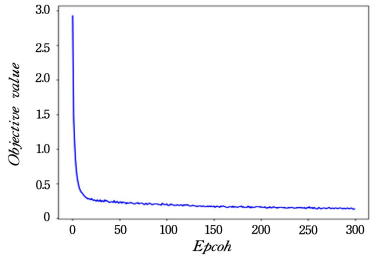


图4 目标函数值的迭代变化曲线

Fig. 4 Iterative curve of objective function value

结束语 本文提出了一种双门控-残差特征融合的跨模态图文检索方法。所提方法利用门控-残差特征融合网络来提高图像模态特征和文本模态特征的交互性，以获得更完备的特征表示，从而进一步弥合图像文本模态数据间的异质性差距；又通过最小化联合损失函数来优化模型，提高了跨模态检索的性能。本文主要考虑图像文本数据间的异质性差距问题，未来的研究还会更进一步研究跨模态图文检索，如图像关键信息和文本关键信息中的对齐匹配。

参 考 文 献

[1] PENG Y X, HUANG X, ZHAO Y Z. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 28(9): 2372-2385.

[2] ZHANG L, MA B P, LI G R, et al. Generalized semi-supervised and structured subspace learning for cross-modal retrieval[J]. IEEE Transactions on Multimedia, 2017, 20(1): 128-141.

[3] ZHANG L, MA B P, LI G R, et al. PL-ranking: A novel ranking method for cross-modal retrieval[C]// Proceedings of the 24th ACM International Conference on Multimedia. New York: ACM, 2016: 1355-1364.

[4] PENG X, HUANG Z Y, LV J C, et al. COMIC: Multi-view clustering without parameter selection[C]// International Conference on Machine Learning. New York: PMLR, 2019: 5092-5101.

[5] WEI Y C, ZHAO Y, LU C Y, et al. Cross-modal retrieval with CNN visual features: A new baseline[J]. Piscataway: IEEE Transactions on Cybernetics, 2016, 47(2): 449-460.

[6] ZENG D H, OYAMA K. Learning joint embedding for cross-modal retrieval[C]// 2019 International Conference on Data Mining Workshops(ICDMW). IEEE, 2019: 1070-1071.

[7] QIANG B H, ZHAO T, WANG Y F, et al. Cross-modal Retrieval Based on Stacked Bimodal Auto-Encoder[C]// 2019 Eleventh International Conference on Advanced Computational Intelligence(ICACI). IEEE, 2019: 256-261.

[8] RASIWASIA N, PEREIRA J C, COVIELLO E, et al. A new approach to cross-modal multimedia retrieval[C]// Proceedings of the 18th ACM International Conference on Multimedia. New York: ACM, 2010: 251-260.

[9] HWANG S J, GRAUMAN K. Accounting for the relative importance of objects in image retrieval[C]// BMVC. 2010: 5.

[10] LEE K H, CHEN X, HUA G, et al. Stacked cross attention for image-text matching[C]// Proceedings of the European Conference on Computer Vision(ECCV). Berlin: Springer, 2018: 201-216.

[11] CORNIA M, BARALDI L, TAVAKOLI H R, et al. A unified cycle-consistent neural model for text and image retrieval[J]. Multimedia Tools and Applications, 2020, 79(35): 25697-25721.

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

[13] REN S H, LIN J Y, ZHAO G X, et al. Learning relation alignment for calibrated cross-modal retrieval[J]. arXiv: 2105.13868, 2021.

[14] PENG Y X, QI J W, YUAN Y X. Modality-specific cross-modal similarity measurement with recurrent attention network[J]. IEEE Transactions on Image Processing, 2018, 27(11): 5585-5599.

[15] CAO Y, LONG M S, WANG J M, 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. New York: ACM, 2016: 1445-1454.

[16] LIN Q B,CAO W M,HE Z H,et al. Semantic deep cross-modal hashing[J]. Neurocomputing,2020,396:113-122.

[17] WANG H,SAHOO D,LIU C H,et al. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE,2019:11572-11581.

[18] WU F,JING X Y,WU Z Y,et al. Modality-specific and shared generative adversarial network for cross-modal retrieval[J]. Pattern Recognition,2020,104:107335.

[19] 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. IEEE,2019:10394-10403.

[20] GUO M,YUAN Y,LU X Q. Deep cross-modal retrieval for remote sensing image and audio[C]// 2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing(PRRS). IEEE,2018:1-7.

[21] BELTAN L V B,CAICEDO J C,JOURNET N,et al. Deep multimodal learning for cross-modal retrieval:One model for all tasks[J]. Pattern Recognition Letters,2021,146:38-45.

[22] DONG X F,ZHANG H X,DONG X,et al. Iterative graph attention memory network for cross-modal retrieval[J]. Knowledge-Based Systems,2021,226:107138.

[23] WANG X,HU P,ZHENG L L,et al. DRSL:Deep relational similarity learning for cross-modal retrieval[J]. Information Sciences,2021,546:298-311.

[24] VO N,JIANG L,SUN C,et al. Composing text and image for image retrieval-an empirical odyssey[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE,2019:6439-6448.

[25] DONG Z H,WAN J,LI C Y,et al. Feature Fusion based Cross-modal Retrieval for Traditional Chinese Painting[C]// 2020 International Conference on Culture-oriented Science & Technology(ICCST). IEEE,2020:383-387.

[26] 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):1-24.

[27] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science,2014.

[28] PEREIRA J C,COVIELLO E,DOYLE G,et al. On the role of correlation and abstraction in cross-modal multimedia retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,36(3):521-535.

[29] RASHTCHIAN C,YOUNG P,HODOSH M,et al. Collecting image annotations using amazon's mechanical turk[C]// Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Stroudsburg,PA:ACL,2010:139-147.

[30] WANG K,HE R,WANG W,et al. Learning coupled feature spaces for cross-modal matching[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway:IEEE,2013:2088-2095.

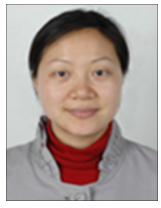
[31] SHARMA A,KUMAR A,DAUME H,et al. Generalized Multiview analysis:A discriminative latent space[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE,2012:2160-2167.

[32] LI M Y,LI Y,HUANG S L,et al. Semantically supervised maximal correlation for cross-modal retrieval[C]// 2020 IEEE International Conference on Image Processing(ICIP). IEEE,2020:2291-2295.

[33] WANG Y B,PENG Y X. MARS:Learning Modality-Agnostic Representation for Scalable Cross-media Retrieval[J]. IEEE Transactions on Circuits and Systems for Video Technology,2022,32(7):4765-4777.

[34] LAURENS V D M,HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research,2008,9(11):2579-2605.

ZHANG Changfang, born in 1960, Ph.D, professor, Ph. D supervisor. His main research interests include fault diagnosis on electrical machines and industrial process control.



HE Jing, born in 1971, Ph.D, professor, master supervisor. Her main research interests include fault diagnosis on mechatronics machinesand industrial process control.