智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

特征融合的装修案例跨模态检索方法

亢洁, 刘威

引用本文:
亢洁,刘威. 特征融合的装修案例跨模态检索方法[J]. 智能系统学报, 2024, 19(2): 429–437.
KANG Jie, LIU Wei. A cross–modal retrieval algorithm of decoration cases on feature fusion[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(2): 429–437.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202207030

DOI: 10.11992/tis.202207030
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20231116.1417.008

特征融合的装修案例跨模态检索方法

亢洁, 刘威
(陕西科技大学 电气与控制工程学院, 陕西 西安 710021)

摘　要: 目前家装客服系统中主要依靠人工方式进行装修案例检索, 导致该系统不能满足用户对咨询服务快捷、及时的需求而且人力成本高, 故提出一种基于特征融合的装修案例跨模态检索算法。针对多模态数据的语义信息挖掘不充分, 模型检索精度低等问题, 对现有的风格聚合模块进行改进, 在原始模块中引入通道注意力机制, 以此来为每组装修案例中不同图片的特征向量添加合适的权重, 从而增强包含更多有用信息的重要特征并削弱其他不重要的特征。同时, 为充分利用多模态信息, 设计一种适用于检索场景下的多模态特征融合模块, 该模块能够自适应地控制 2 种不同模态的特征向量进行一系列的融合操作, 以实现跨模态数据间的知识流动与共享, 从而生成语义更丰富、表达能力更强的特征向量, 进一步提升模型的检索性能。在自建的装修案例多模态数据集上将该方法与其他方法进行比较, 试验结果表明本文方法在装修案例检索上具有更优越的性能。

关键词: 家装客服系统; 装修案例检索; 跨模态检索; 风格聚合; 多模态; 特征融合; 通道注意力机制; 语义信息
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2024)02−0429−09

中文引用格式: 亢洁, 刘威. 特征融合的装修案例跨模态检索方法 [J]. 智能系统学报, 2024, 19(2): 429–437.
英文引用格式: KANG Jie, LIU Wei. A cross-modal retrieval algorithm of decoration cases on feature fusion[J]. CAAI transactions on intelligent systems, 2024, 19(2): 429–437.

A cross-modal retrieval algorithm of decoration cases on feature fusion

KANG Jie, LIU Wei
(School of Electrical and Control Engineering, Shaanxi University of Science & Technology, Xi'an 710021, China)

Abstract: Currently, home decoration customer service systems chiefly depend on manual decoration case retrieval, which leads to the system not meeting user demand for fast and timely consulting service and high labor costs. Thus, we propose a feature fusion-based cross-modal retrieval algorithm for decoration cases. Aiming at the problems of insufficient semantic information mining of multimodal data and low accuracy of model retrieval, the existing style aggregation module is improved. The channel attention mechanism is introduced into the original module to add suitable weights to the feature vectors of different pictures in each group of decoration cases, thereby improving the important features that include more helpful information and weakening other unimportant features. Conversely, a multimodal feature fusion module is developed for retrieval scenarios to make full use of multimodal information. The module can adaptively control a series of fusion operations of feature vectors from two different modalities to achieve knowledge flow and sharing between cross-modal data, thereby producing feature vectors with richer semantics and stronger expressive ability to improve the retrieval performance of the model further. Our developed algorithm is compared with other methods on a self-built multimodal dataset of decoration cases, and results show that the algorithm performs better in decoration case retrieval.

Keywords: home decoration customer service system; decoration case retrieval; cross-modal retrieval; style aggregation; multimodal; feature fusion; channel attention mechanism; semantic information

随着互联网技术的高速发展, 互联网家装平台应运而生, 作为其重要组成部分的客服系统是家装企业与客户在线沟通的桥梁。装修案例检索

收稿日期: 2022−07−20.　**网络出版日期:** 2023−11−16.
基金项目: 陕西省重点研发计划项目(2021GY-022).
通信作者: 亢洁. E-mail: kangjie@sust.edu.cn.

是家装客服系统中一项必不可少的功能，其通过文本信息检索与之相关的图像信息[1-2]，而目前人工方式的装修案例检索人力成本高、实时性差，大大降低了企业的服务质量。因此急需一种装修案例的智能检索系统，该系统可以采用跨模态图文检索的方法来完成，从而降低企业的人力成本，并满足用户对咨询服务快捷、及时的需求。

目前，学习公共子空间是跨模态检索领域中最主流的特征关联方法，其依据是语义相同的不同模态数据之间具有潜在的相关性，这使得构建公共子空间并将不同模态数据的特征映射到这个空间直接进行相似性度量成为可能[3-11]。其中，代表性的工作有典型相关分析 (canonical correlation analysis, CCA)[12]、基于核的典型相关分析算法 (Kernel canonical correlation analysis, KCCA)[13]、多视角判别分析 (multi-view discriminant analysis, MvDA)[14] 及带视角一致性的多视角判别分析 (MvDA with view consistency, MvDA-VC)[15] 等。这些方法是基于传统统计分析的方法，其通过优化统计值来为多模态数据学习公共子空间的线性投影矩阵。近年来，相关学者开始将深度学习应用到跨模态检索领域，提出了许多基于深度学习的跨模态检索方法[16-20]。Wang 等[21] 提出了一种正则化的深度神经网络 (regularized deep neural network, RE-DNN)，分别利用卷积神经网络和词袋模型来提取输入图像和文本的特征，并通过一个五层的神经网络将视觉特征与文本特征投影到公共子空间中，进而实现不同模态间相似性的度量。Xu 等[22] 提出了一种用于跨模态检索的深度对抗度量学习 (deep adversarial metric learning, DAML)方法，利用对抗损失来减小不同模态数据在所学子空间中的异构鸿沟。Zhang 等[23] 提出了一种跨模态关系引导网络 (cross-modal relation guided network, CRGN)，利用残差网络来学习图像特征，并通过对图像不同区域之间的关系进行建模，最终生成一个基于全局信息的图像特征。然而，这些方法并不能直接应用于装修案例智能匹配的场景中。因为这些方法建立的是单个文本与单张图片之间的关联关系，当利用文本检索图片时，其检索结果是多张与文本描述类似的图片；而在本场景中，需要建立单个文本与一组图片 (装修案例)之间的关联关系，当利用文本检索装修案例时，希望返回的结果是多组与文本风格语义相似的装修案例。亢洁等[24] 提出了一种面向装修案例智能匹配的跨模态检索方法，通过一种风格聚合模块来生成代表一组装修案例整体风格的特征表示，从而利用后续网络建立文本信息与装修案例之间潜在的语义关联，以实现两者间的跨模态匹配。虽然该方法实现了装修案例的智能匹配，但其仍然存在一些问题需要改进。首先，在装修案例多模态数据集的图像样本中，每组案例都包含一定数量的装修图片。其中，有些图片的风格特点非常突出，而有些图片的风格则不够明显。例如，一组中式风格的装修案例中，描述书房、客厅等区域的图片可能具有浓重的中式风格，而描述浴室、厨房等区域的图片则可能没有特别明确的风格指向。所以，当模型在使用装修案例中的图片生成代表装修案例整体的风格特征时，应考虑到每张图片的重要程度是不一样的，模型需要关注那些风格鲜明的图片，以生成风格语义更加显著的特征表示。其次，文本描述的语义层次高，但包含的信息有限；而图像的语义层次低，但包含了丰富的信息。当文本与图像进行匹配时，文本信息的准确性可以指导图像特征的生成，图像信息的复杂性也可以对文本特征进行增强。因此，当利用文本和图像特征学习公共向量空间时，应考虑 2 种模态之间的信息交互，以获得表达能力更强的特征表示，从而提高跨模态数据间的检索精度。

为解决上述问题，本研究在文献 [24] 所提方法的基础上提出了一种基于特征融合的装修案例跨模态检索方法，该算法利用深度神经网络提取文本和图像的特征，并将两者进行多模态特征融合，然后投影到一个公共的表示空间，以此来建立文本与图像之间的对应关系，从而完成通过指定文本检索相应风格的装修案例这一任务。

1　基于特征融合的装修案例跨模态检索方法

本研究算法包括文本特征学习模块、融合改进风格聚合模块的图像特征学习模块、多模态特征融合模块和公共向量空间学习模块 4 个模块。在融合改进风格聚合模块的图像特征学习模块中设计了一种改进的风格聚合模块，便于后续网络建立文本与装修案例之间的潜在关系。同时，为了充分利用多模态信息，设计了一种适用于检索场景下的多模态特征融合模块，该模块能够自适应地控制 2 种不同模态的特征向量进行一系列的融合操作，以实现跨模态数据间的知识流动与共享，从而生成语义更丰富、表达能力更强的特征向量，进一步提升模型的检索性能。

1.1　模型整体框架

本研究所提模型的整体框架如图 1 所示，从左到右从上到下主要包括 4 个模块。

1）文本特征学习模块，输入的文本信息利用在维基百科中文数据集上预先训练的 BERT[25] 网络对文本信息进行特征提取。

2）融合改进风格聚合模块的图像特征学习模块，利用在 ImageNet 数据集上预先训练的 VGG19[26] 网络对输入的图像信息进行特征提取；图像特征学习时将 VGG19 网络 block1 层输出的特征图作为输入图片的纹理特征，然后，再利用改进的风格聚合模块对纹理特征进行处理。

3）多模态特征融合模块，将特征学习模块得到的特征向量$\boldsymbol{h}_i^{\alpha}$和$\boldsymbol{h}_i^{\beta}$输入多模态特征融合模块进行跨模态间的信息交互，输出融合后的文本特征$\boldsymbol{v}_i'$和图像特征$\boldsymbol{w}_i'$。

4）公共向量空间学习模块，通过全连接层f_{c1}和f_{c2}为特征向量$\boldsymbol{v}_i'$和$\boldsymbol{w}_i'$学习一个潜在的公共向量空间，并得到两种模态在该空间中的特征表示$\boldsymbol{v}_i$和$\boldsymbol{w}_i$，两者可以直接进行相似性的比较。同时，分别在 2 个子网络的末端连接 2 个线性分类器f_{c3}和f_{c4}，利用标签信息来学习区分特征。

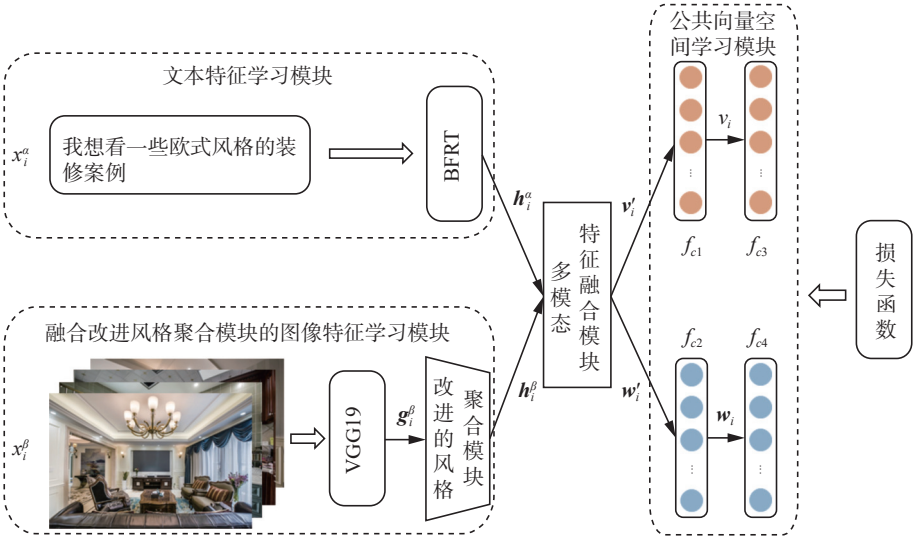

图 1　基于特征融合的装修案例跨模态检索方法的整体框架

Fig. 1　An overall framework for a cross-modal retrieval algorithm of decoration cases based on feature fusion

假设数据集中包含 n 个文本−图像对，用$M=\{(x_i^{\alpha},x_i^{\beta})\}_{i=1}^{n}$表示，其中$x_i^{\alpha}$表示第 i 个样本中的文本信息，与客服系统中用户输入的文本对应；x_i^{β}表示第 i 个样本中的图像信息，与客服系统中被检索的装修案例对应。每个样本对$\left(x_i^{\alpha},x_i^{\beta}\right)$都对应有各自的标签向量，用$\boldsymbol{y}_i=[y_{i1}\ \ y_{i2}\ \ \cdots\ \ y_{ic}]\in\mathbf{R}^c$表示，其中 c 表示输入样本的总类别数。当第 i 个样本属于第 j 类时$y_{ij}=1$，否则$y_{ij}=0$。接下来，以第 i 个输入样本为例来介绍整个模型的工作流程。

1.2　文本特征学习模块

将第 i 个输入样本中的查询文本x_i^{α}，利用在维基百科中文数据集上预先训练的 BERT[25] 网络对文本信息进行特征提取。其中，将 BERT 网络输出中 [CLS] 标志位对应的一个 768 维的向量作为查询文本x_i^{α}的语义特征表示，记作$\boldsymbol{h}_i^{\alpha}$。

1.3　融合改进风格聚合模块的图像特征学习模块

将第 i 个输入样本中的图像信息x_i^{β}，利用在 ImageNet 数据集上预先训练的 VGG19[26] 网络对图像信息进行特征提取。其中，将 VGG19 网络 block1 层输出的特征图作为输入图片的纹理特征，则含有 k 张图片的装修案例x_i^{β}的纹理特征表示为$\boldsymbol{g}_i^{\beta}=\left\{\boldsymbol{g}_{i1}^{\beta},\boldsymbol{g}_{i2}^{\beta},\cdots,\boldsymbol{g}_{ik}^{\beta}\right\}$，其中$\boldsymbol{g}_{ik}^{\beta}$表示第 i 个样本的装修案例中第 k 张图片的纹理特征，大小为64×112×112。然后，再利用改进的风格聚合模块对纹理特征$\boldsymbol{g}_i^{\beta}$进行处理，可以获得一个大小为64×64的特征图，最终将其展开成一个 4096 维的向量作为装修案例x_i^{β}的风格特征表示，记作$\boldsymbol{h}_i^{\beta}$。

下面重点介绍改进的风格聚合模块。

装修案例跨模态检索模型的核心是借助风格语义来建立查询文本与装修案例之间的对应关系。文献 [24] 设计了一种风格聚合模块，该模块通过对一组装修案例中所有图片的纹理特征进行整合，生成了代表该组装修案例整体风格的特征表示，使得后续网络可以学习查询文本与装修案例之间的潜在语义联系。然而，该模块没有考虑不同图片的风格特征对于最终结果的重要程度应

该是不一样的。其中，风格鲜明的图片包含更多的风格语义信息，可以使最终结果的风格语义更加显著；而风格模糊的图片可能削弱这种影响。所以，为了获得更好的特征表示，本研究设计了改进的风格聚合模块，在文献 [24] 风格聚合模块中引入通道注意力机制，即 SE（squeeze-and-excitation）[27] 模块，来学习不同图片对于最终结果的重要程度。其能够通过全局信息自动学习输入特征图各通道的权重，为重要通道的特征赋予更大的权值，减小其他通道特征的权值。

改进的风格聚合模块如图 2 所示。模块的输入是一组装修案例中所有图片的纹理特征 $\boldsymbol{g}_i^\beta=\left\{\boldsymbol{g}_{i1}^\beta,\boldsymbol{g}_{i2}^\beta,\cdots,\boldsymbol{g}_{ik}^\beta\right\}$，首先通过计算每张图片纹理特征的格拉姆（Gram）矩阵，可以得到对应图片的风格特征[28]，再对所有图片的风格特征进行拼接，生成一组大小为 $k\times64\times64$ 的特征图，该特征图含有 k 个通道，每个通道代表一张图片的风格特征。接着，将其输入 SE 模块，学习不同图片的风格特征在整个装修案例中的重要程度，并通过分配不同的权重来对不同通道的特征进行调整，最终获得一组风格语义更加显著的特征图，大小为 $k\times64\times64$。最后，利用卷积层进一步整合特征图中的风格语义信息，输出一个大小为 64×64 的特征图作为该组装修案例的风格特征表示。

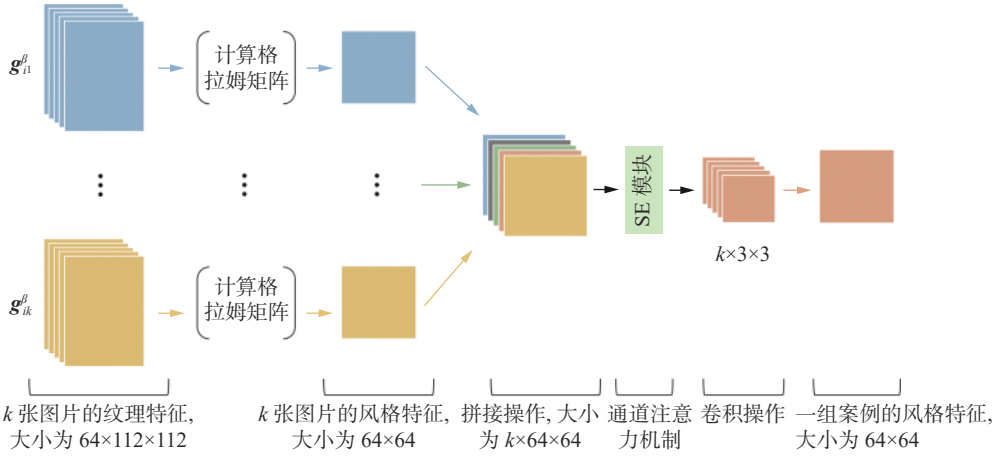

图 2　改进的风格聚合模块结构

Fig. 2　Improved style aggregation module structure diagram

1.4　多模态特征融合模块

将 1.2 和 1.3 节中得到的特征向量 $\boldsymbol{h}_i^\alpha$ 和 $\boldsymbol{h}_i^\beta$ 输入给图 1 中的多模态特征融合模块进行跨模态间的信息交互，输出融合后的文本特征 $\boldsymbol{v}_i'$ 和图像特征 $\boldsymbol{w}_i'$。

为了解决跨模态信息间的异构性问题，多数方法将文本特征和图像特征独立地投影到一个公共向量空间中，以比较它们之间的相似性。然而，这种方法在计算跨模态特征间的相似性之前，缺乏不同模态间的信息交互，忽略了多模态数据间的冗余性和互补性，信息没有被充分利用，使得文本与装修案例之间的匹配可能没有达到最优结果[29]。因此，本研究提出了一种多模态特征融合模块（multimodal feature fusion module，MFFM）来实现跨模态检索中不同模态间的信息流动与共享，以生成表达能力更强的特征，从而进一步提升模型的性能。MFFM 的结构如图 3 所示，具体的融合过程描述如下。

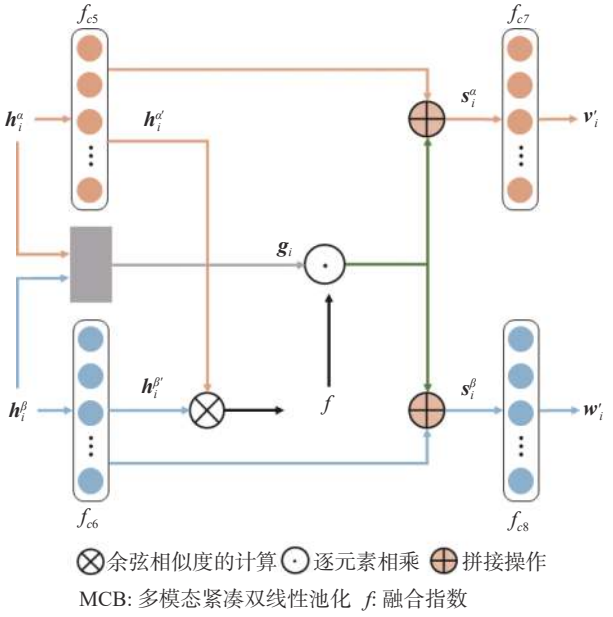

图 3　多模态特征融合模块（MFFM）结构图

Fig. 3　Multimodal feature fusion module（MFFM）structure diagram

MFFM 的目的是对输入的多模态特征 $\boldsymbol{h}_i^\alpha$ 和 $\boldsymbol{h}_i^\beta$ 进行融合操作。其中，$\boldsymbol{h}_i^\alpha$ 表示第 i 个样本中查询

文本的语义特征，是一个 768 维的向量；$\boldsymbol{h}_i^{\beta}$表示第 i个样本中装修案例的风格特征，是一个 4 096 维的向量。首先，利用多模态紧凑双线性池化（multimodal compact bilinear pooling，MCB）[30] 方法对输入特征$\boldsymbol{h}_i^{\alpha}$和$\boldsymbol{h}_i^{\beta}$进行初步融合，输出融合后的特征表示$\boldsymbol{g}_i$，是一个 1 024 维的向量。

然后，采用门控机制来实现对输入特征融合程度的控制。传统的多模态特征融合方式一般是基于文本和图像所表达的信息是相同的，即两者是匹配的。但是在跨模态检索过程中，文本信息与所检索的图像信息可能是不匹配的。试验表明，不匹配的文本−图像对之间的直接融合可能会生成无意义的特征表示，从而妨碍模型的推理能力，降低模型的检索性能[29]。因此，MFFM 通过引入融合指数（Fusion Value）来自适应地控制 2 种模态间的信息融合程度，从而尽可能地鼓励匹配的文本−图像对间的融合，并抑制不匹配的文本−图像对间的融合。融合指数f的计算公式为

$$\begin{cases} f = \cos\_similarity\left(\boldsymbol{h}_i^{'\alpha}, \boldsymbol{h}_i^{'\beta}\right) \\ \boldsymbol{h}_i^{'\alpha} = \boldsymbol{h}_i^{\alpha}\boldsymbol{W}_1 \\ \boldsymbol{h}_i^{'\beta} = \boldsymbol{h}_i^{\beta}\boldsymbol{W}_2 \end{cases} \tag{1}$$

式中，cos_similarity(·,·)是用于计算 2 个向量余弦相似度的函数。余弦相似度的计算需要 2 个特征向量的维数一致，所以通过具有激活函数 ReLU 的全连接层f_{c5}和f_{c6}分别将输入特征$\boldsymbol{h}_i^{\alpha}$和$\boldsymbol{h}_i^{\beta}$转化为 1 024 维的特征向量$\boldsymbol{h}_i^{'\alpha}$和$\boldsymbol{h}_i^{'\beta}$，$\boldsymbol{W}_1$和$\boldsymbol{W}_2$分别表示全连接层$f_{c5}$和$f_{c6}$中的可学习参数。$\boldsymbol{h}_i^{'\alpha}$与$\boldsymbol{h}_i^{'\beta}$间的相似性越高，则融合指数$f$的值越大；反之，则$f$的值越小。又因为在融合多模态特征的同时，也应该保留自身模态的原始信息，所以将融合后的特征向量通过残差连接的方式与原始的特征向量进一步融合，输出 2 048 维的特征向量$\boldsymbol{s}_i^{\alpha}$和$\boldsymbol{s}_i^{\beta}$，计算公式分别为

$$\boldsymbol{s}_i^{\alpha} = (f \odot \boldsymbol{g}_i) \oplus \boldsymbol{h}_i^{'\alpha} \tag{2}$$

$$\boldsymbol{s}_i^{\alpha} = (f \odot \boldsymbol{g}_i) \oplus \boldsymbol{h}_i^{'\alpha} \tag{3}$$

式中：⊙表示逐元素相乘；⊕表示拼接操作。当 2 种输入模态的信息属于同一类别时，f的值较大，增强此时的融合程度；反之，则f的值较小，削弱此时的融合程度。

最后，通过具有激活函数 ReLU 的全连接层f_{c7}和f_{c8}分别对拼接后的特征向量$\boldsymbol{s}_i^{\alpha}$和$\boldsymbol{s}_i^{\beta}$进行学习并降维，得到 1 024 维的文本特征$\boldsymbol{v}_i'$和图像特征$\boldsymbol{w}_i'$，计算公式分别为

$$\boldsymbol{v}_i' = \boldsymbol{s}_i^{\alpha}\boldsymbol{U}_1 \tag{4}$$

$$\boldsymbol{w}_i' = \boldsymbol{s}_i^{\beta}\boldsymbol{U}_2 \tag{5}$$

其中，$\boldsymbol{U}_1$和$\boldsymbol{U}_2$分别表示全连接层f_{c7}和f_{c8}中的可学习参数。

1.5　公共向量空间学习模块

通过图 1 中的全连接层f_{c1}和f_{c2}为从 1.4 节中得到的特征向量$\boldsymbol{v}_i'$和$\boldsymbol{w}_i'$学习一个潜在的公共向量空间，并得到 2 种模态在该空间中的特征表示$\boldsymbol{v}_i$和$\boldsymbol{w}_i$，两者可以直接进行相似性比较。同时，分别在 2 个子网络的末端连接 2 个线性分类器f_{c3}和f_{c4}，利用标签信息来学习区分特征。其中，f_{c1}和f_{c2}的隐藏单元数量均为 512，且权值共享；线性分类器f_{c3}和f_{c4}的隐藏单元数量均为 8，且权值共享。本研究采用文献 [24] 中提出的损失函数对模型进行训练。

2　试验验证

2.1　数据集

为了验证所提模型的性能，本研究从某互联网家装企业获取到部分用户的查询语料和相应的装修案例，通过对这些数据进行整理，最终构建了一个关于装修案例的多模态数据集，部分数据如图 4 所示。

图 4　装修案例多模态数据集中的部分样本

Fig. 4　Sample of decoration case multimodal dataset

装修案例多模态数据集共包含 7 200 个样本对，并根据风格语义设置了 8 个类别标签，分别为中式、欧式、美式、日式、地中海、现代简约、古典和田园，每个样本对共用一个类别标签。数据集中的每个样本对均包含文本信息和图像信息两部分，其中文本信息是一个句子，其内容主要是查找某种风格的装修案例，句子的平均长度为 10.43 个字；图像信息是一组装修案例，该组装修案例的装修风格与样本对的标签一致，每组装修案例中包含 9~13 张数量不等的图片。实验时，按照 90% 和 10% 的比例将数据集随机划分为训练集和测试集。

2.2　试验环境与评价指标

2.2.1　试验环境

本研究所提模型是在 PyTorch 深度学习框架中搭建的，硬件试验平台的 CPU 为 Intel Core i7-8700，内存为 16 GB，GPU 为 11 GB 的 NVIDIA GeForce GTX 2080Ti，编程语言为 Python3.6.12。本次模型训练一共设置了 500 轮，一轮训练中每批量数据大小设置为 100，学习率设置为10^{-4}。模型使用 Adam 优化器进行参数更新，其中一阶矩估计的指数衰减率$\beta_1=0.5$，二阶矩估计的指数衰减率$\beta_2=0.999$，同时设置一个用于数值稳定的小常数$\varepsilon=10^{-8}$，防止在参数更新中出现除以 0 的情况。SE 模块中的降维系数r为 1。

2.2.2　评价指标

模型经过训练后，可以得到不同模态信息在公共向量空间中的特征表示。基于此，本研究通过计算文本特征与图像特征之间的余弦距离来衡量两者间的相似性，并采用在图文检索领域中广泛使用的两种评价指标：召回率（Recall@N）和平均精度均值（mean average precision，mAP）对本研究所提检索模型的性能进行评估。同时，鉴于应用场景的实际需求，本研究只考虑以文本信息检索图像信息这一任务中模型的性能。评价指标 Recall@N 和 mAP 的值越大，则表明模型的检索性能越好。

2.3　试验结果与分析

2.3.1　不同多模态特征融合方法对模型性能的影响

本研究通过引入多模态特征融合模块（MFFM）来对输入的文本特征和图像特征进行融合，以实现跨模态间的信息交互，生成表达能力更强的特征向量。为了研究不同多模态特征融合方法对模型性能的影响，本次试验对以下 5 种特征融合方法进行了对比，分别表示为：MFFM1（$\boldsymbol{h}_i^{\prime\alpha}$和$\boldsymbol{h}_i^{\prime\beta}$逐元素相加）、MFFM2（$\boldsymbol{h}_i^{\prime\alpha}$和$\boldsymbol{h}_i^{\prime\beta}$逐元素相乘）、MFFM3（$\boldsymbol{h}_i^{\alpha}$和$\boldsymbol{h}_i^{\beta}$拼接）、MFFM4（$\boldsymbol{h}_i^{\alpha}$和$\boldsymbol{h}_i^{\beta}$采用 MCB 方法进行融合，输出融合后的特征向量$\boldsymbol{g}_i$）以及 MFFM（第 1.3 节提出的融合方法），其中 MFFM1、MFFM2 和 MFFM3 都通过全连接层输出融合后的特征向量$\boldsymbol{g}_i$，且 MFFM1、MFFM2、MFFM3 和 MFFM4 中均不包括门控机制。另外，本次实验的基准模型（Baseline）为本研究所提模型（不包含 MFFM）。

采用不同多模态特征融合方法的模型在装修案例多模态数据集上的试验结果如表 1 所示。可以看出，当模型使用 MFFM 方法时，Recall@5、Recall@10、Recall@15 和 mAP 的值最高，其相较于 Baseline 分别提升了 0.052、0.069、0.009 和 0.097，此时模型的检索性能最好。其次，MFFM4 方法与 MFFM 方法相比，缺少了门控机制，当模型使用 MFFM4 方法时，其在 Recall@5、Recall@10、Recall@15 和 mAP 上的值相较于 MFFM 分别降低了 0.032、0.041、0.049 和 0.055。由此可知，MFFM 方法通过门控机制可以更加合理地控制多模态间的特征融合，从而显著提升模型的性能。同时，当模型使用 MFFM1、MFFM2 或 MFFM3 中任意一种方法时，其在 Recall@5、Recall@10、Recall@15 和 mAP 上的值相较于 Baseline 仅有略微的提升，此时模型的性能与最优模型相差较大。因此，本研究所提模型采用 MFFM 方法来完成多模态特征融合的任务。

表 1　不同的多模态特征融合方法的对比试验

Table 1　Comparative experiments of different multimodal feature fusion methods

方法	Recall@5	Recall@10	Recall@15	mAP
Baseline	0.616	0.673	0.717	0.755
MFFM1	0.625	0.691	0.733	0.770
MFFM2	0.627	0.694	0.739	0.774
MFFM3	0.618	0.685	0.730	0.765
MFFM4	0.636	0.701	0.758	0.797
MFFM	**0.668**	**0.742**	**0.807**	**0.852**

注：加黑数字为评价指标最好，下同。

2.3.2　与现有方法的对比试验

为了验证本研究提出的基于特征融合的装修案例跨模态检索方法的有效性，在装修案例多模态数据集上，将其与多种常见的图文检索方法进行比较，包括 KCCA[13]、MvDA-VC[15]、CMPM[16]（cross-modal projection matching）、CAMP[29]（cross-modal adaptive message passing）。其中，KCCA 和 MvDA-VC 是基于传统统计分析的方法，CMPM 和 CAMP 是基于深度学习的方法，且 CAMP 中对多模态特征进行了融合。同时，由于本次实验的 Baseline 为文献 [24] 所提的装修案例跨模态检索模型，所以将本研究所提模型与

Baseline、Baseline + SE、Baseline + MFFM 也进行了对比。为了公平起见，所有参与比较的方法均采用 BERT 和 VGG19 预训练模型来获取样本中文本和图像的特征表示。

表 2 为不同方法在装修案例多模态数据集上的召回率和平均精度均值，可以发现，本研究所提模型相较于 Baseline，在 Recall@5、Recall@10、Recall@15 和 mAP 上的值分别有了 0.071、0.104、0.142 和 0.138 的提升，其相较于次优方法 Baseline + MFFM 也分别提升了 0.023、0.033、0.046 和 0.046，此时模型的检索性能最好，且全面优于其他方法。其次，Baseline + SE 和 Baseline + MFFM 在 Recall@5、Recall@10、Recall@15 和 mAP 上的值均优于 Baseline，这说明本研究通过引入 SE 模块和 MFFM 有效提升了 Baseline 的性能。同时，CAMP 相较于 Baseline，在 Recall@5、Recall@10、Recall@15 和 mAP 上的值分别提升了 0.037、0.059、0.079 和 0.082，且与 Baseline + MFFM 相比，差距不大，这表明在跨模态检索模型中引入合适的特征融合模块可以使模型具有更好的性能。另外，KCCA 和 MvDA-VC 这 2 种方法在 Recall@5、Recall@10、Recall@15 和 mAP 上的值均低于其他方法，这说明相较于传统方法学习到的线性关系，深度学习方法可以为多模态数据建立非线性的高阶映射关系，从而实现更好的检索性能。

表 2　不同方法的对比试验

Table 2　Comparative experiments of different methods

方法	Recall@5	Recall@10	Recall@15	mAP
KCCA	0.323	0.429	0.493	0.558
MvDA-VC	0.411	0.525	0.572	0.595
CMPM	0.548	0.610	0.639	0.663
CAMP	0.634	0.697	0.744	0.796
Baseline	0.597	0.638	0.665	0.714
Baseline + SE	0.616	0.673	0.717	0.755
Baseline + MFFM	0.645	0.709	0.761	0.806
本研究方法	**0.668**	**0.742**	**0.807**	**0.852**

2.3.3　可视化公共子空间中的特征表示

本次实验的 Baseline 为文献 [24] 提出的装修案例跨模态检索模型。本研究所提模型是在 Baseline 的基础上进行改进的，为了更加直观地研究该模型的有效性，采用 t-SNE 方法分别将测试样本在 2 种模型学习到的公共子空间中的特征表示投影到二维空间中进行可视化。测试样本在两种模型所学公共子空间中的原始文本特征 $\boldsymbol{v}_i$ 和原始图像特征 $\boldsymbol{w}_i$ 均是 512 维的特征向量，其可视化后的结果如图 5 所示，同一类别的样本用相同的颜色标记。

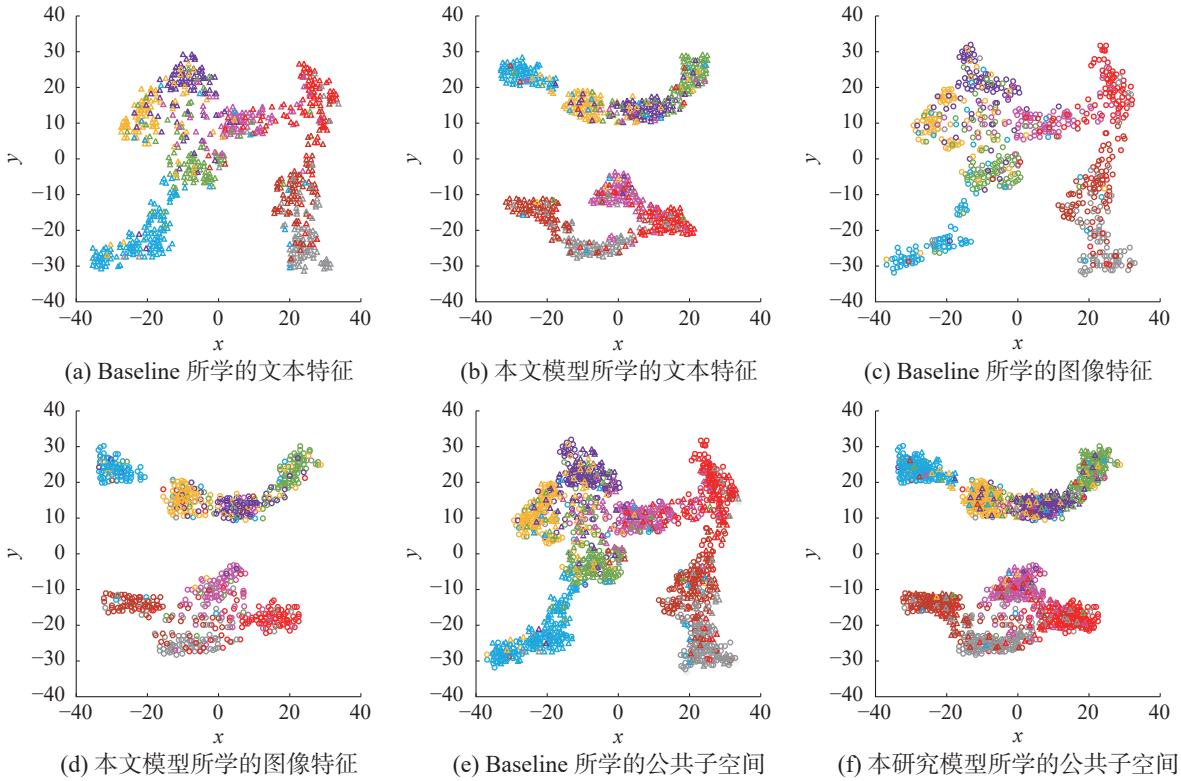

图 5　测试样本在不同公共子空间中投影后的可视化结果

Fig. 5　Visualization results of the test data projected in different common subspaces

图 5(a)和图 5(c)显示了 Baseline 所学公共子空间中文本特征和图像特征的二维分布。图 5(b)和图 5(d)则显示了本研究模型所学公共子空间中文本特征和图像特征的二维分布。通过以上结果可以看出，首先 2 种模型所学的公共子空间都能够对来自不同语义类别的样本进行区分，并形成了多个语义簇。其次，通过对比可以发现，相较于图 5(a)和图 5(c)，图 5(b)和图 5(d)的语义簇更加集中，且不同语义类别的混合程度更低，这表明本研究模型所学公共子空间的辨别能力要优于 Baseline 所学的公共子空间，与表 2 所示的结果一致。此外，图 5 中(e)和图 5(f)分别显示了 2 种模型所学公共子空间中测试样本特征表示的二维分布。可以看出，来自文本模态和图像模态的样本很好地混合在一起，这表明 2 种模型都可以有效地消除跨模态间的差异。

3　结束语

面向装修案例智能检索的现实场景，本研究提出了一种基于特征融合的装修案例跨模态检索方法。首先，该方法对风格聚合模块进行了改进，通过引入 SE 模块来自动学习每组装修案例中不同图片的风格特征对装修案例整体风格特征的重要程度，以增强重要特征并削弱其他不重要的特征，从而生成风格语义更加显著的特征表示。其次，为了实现多模态间的信息交互，设计了一种多模态特征融合模块，其能够自适应地控制跨模态间的特征融合，以生成含有更多语义信息、表达能力更强的特征表示。在自建的装修案例多模态数据集上对本研究方法进行了测试与评价，与 Baseline 相比，其在 Recall@5、Recall@10、Recall@15 和 mAP 上的值分别有了 0.071、0.104、0.014 和 0.138 的提升，证明本研究方法使文本与装修案例的匹配达到了更优的结果。

现实场景中，装修案例检索仍然存在许多更加具体、更加复杂的应用需求。在未来的工作中，将构建文本信息更加丰富的多模态数据集，并研究相应的跨模态检索模型，进一步优化家装客服系统中装修案例检索这一功能。

参考文献：

[1] 刘颖, 郭莹莹, 房杰, 等. 深度学习跨模态图文检索研究综述 [J]. 计算机科学与探索, 2022, 16(3): 489–511.
LIU Ying, GUO Yingying, FANG Jie, et al. Survey of research on deep learning cross-modal image-text retrieval[J]. Journal of frontiers of computer science and technology, 2022, 16(3): 489–511.

[2] 徐文婉, 周小平, 王佳. 跨模态检索技术研究综述 [J]. 计算机工程与应用, 2022, 58(23): 12–23.
XU Wenwan, ZHOU Xiaoping, WANG Jia. Overview of cross-modal retrieval technology[J]. Computer engineering and applications, 2022, 58(23): 12–23.

[3] 宫大汉, 陈辉, 陈仕江, 等. 一致性协议匹配的跨模态图像文本检索方法 [J]. 智能系统学报, 2021, 16(6): 1143–1150.
GONG Dahan, CHEN Hui, CHEN Shijiang, et al. Matching with agreement for cross-modal image-text retrieval[J]. CAAI transactions on intelligent systems, 2021, 16(6): 1143–1150.

[4] 刘卓锟, 刘华平, 黄文美, 等. 视听觉跨模态表面材质检索 [J]. 智能系统学报, 2019, 14(3): 423–429.
LIU Zhuokun, LIU Huaping, HUANG Wenmei, et al. Audiovisual cross-modal retrieval for surface material[J]. CAAI transactions on intelligent systems, 2019, 14(3): 423–429.

[5] ZHANG Qi, LEI Zhen, ZHANG Zhaoxiang, et al. Context-aware attention network for image-text retrieval[C]// Computer Vision and Pattern Recognition.Seattle: IEEE, 2020: 3536−3545.

[6] DONG Jianfeng, LI Xirong, XU Chaoxi, et al. Dual encoding for video retrieval by text[EB/OL]. (2020−10−10)[2022−01−01]. http://arxiv.org/abs/2009.05381.

[7] RAMACHANDRAM D, TAYLOR G W. Deep multimodal learning: a survey on recent advances and trends[J]. IEEE signal processing magazine, 2017, 34(6): 96–108.

[8] 彭良康, 卢向明, 徐清波. 基于深度学习的跨模态哈希检索研究进展 [J]. 数据通信, 2022(3): 32–38.
PENG Liangkang, LU Xiangming, XU Qingbo. Research progress of cross-modal hash retrieval based on deep learning[J]. Data communications, 2022(3): 32–38.

[9] 许炫淦, 房小兆, 孙为军, 等. 语义嵌入重构的跨模态哈希检索 [J]. 计算机应用研究, 2022, 39(6): 1645–1650,1672.
XU Xugeng, FANG Xiaozhao, SUN Weijun, et al. Semantics embedding and reconstructing for cross-modal hashing retrieval[J]. Application research of computers, 2022, 39(6): 1645–1650,1672.

[10] 冯霞, 胡志毅, 刘才华. 跨模态检索研究进展综述 [J]. 计算机科学, 2021, 48(8): 13–23.
FENG Xia, HU Zhiyi, LIU Caihua. Survey of research progress on cross-modal retrieval[J]. Computer science, 2021, 48(8): 13–23.

[11] 尹奇跃, 黄岩, 张俊格, 等. 基于深度学习的跨模态检索综述 [J]. 中国图象图形学报, 2021, 26(6): 1368–1388.
YIN Qiyue, HUANG Yan, ZHANG Junge, et al. Survey

on deep learning based cross-modal retrieval[J]. Journal of image and graphics, 2021, 26(6): 1368–1388.

[12] 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.

[13] ZHANG Hong, LIU Yun, MA Zhigang. Fusing inherent and external knowledge with nonlinear learning for cross-media retrieval[J]. Neurocomputing, 2013, 119: 10–16.

[14] KAN Meina, SHAN Shiguang, ZHANG Haihong, et al. Multi-view discriminant analysis[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 38(1): 188-194.

[15] YANG Xiangfei , LI Chunna, SHAO Yuanhai. Robust multi-view discriminant analysis with view-consistency[J]. Information sciences, 2022, 596: 153–168.

[16] PENG Yuxin, QI Jinwei, HUANG Xin, et al. CCL: cross-modal correlation learning with multigrained fusion by hierarchical network[J]. IEEE transactions on multimedia, 2018, 20(2): 405–420.

[17] ZHANG Ying, LU Huchuan. Deep cross-modal projection learning for image-text matching[C]//European Conference on Computer Vision. Cham: Springer, 2018: 707−723.

[18] 陈曦, 彭姣, 张鹏飞, 等. 基于预训练模型和编码器的图文跨模态检索算法 [J]. 北京邮电大学学报, 2023, 46(5): 112–117.
CHEN Xi, PENG Jiao, ZHANG Pengfei, et al. Cross-modal retrieval algorithm for image and text based on pretrained models and encoders[J]. Journal of Beijing University of Posts and Telecommunications, 2023, 46(5): 112–117.

[19] ZHEN Liangli, HU Peng, WANG Xu, et al. Deep supervised cross-modal retrieval[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition . Long Beach: IEEE, 2020: 10386−10395.

[20] ZHANG Yifan, ZHOU Wengang, WANG Min, et al. Deep relation embedding for cross-modal retrieval[J]. IEEE transactions on image processing, 2021, 30: 617–627.

[21] WANG Cheng, YANG Haojin, MEINEL C. Deep semantic mapping for cross-modal retrieval[C]//2015 IEEE 27th International Conference on Tools with Artificial Intelligence. Vietri sul Mare: IEEE, 2016: 234−241.

[22] XU Xing, HE Li, LU Huimin, et al. Deep adversarial metric learning for cross-modal retrieval[J]. World wide web, 2019, 22(2): 657–672.

[23] ZHANG Qi, LEI Zhen, ZHANG Zhaoxiang, et al. Context-aware attention network for image-text retrieval[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 3533−3542.

[24] 亢洁, 刘威. 面向装修案例智能匹配的跨模态检索方法 [J]. 智能系统学报, 2022, 17(4): 714–720.
KANG Jie, LIU Wei. A crossmodal retrieval method for intelligent matching of decoration cases[J]. CAAI transactions on intelligent systems, 2022, 17(4): 714–720.

[25] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018−11−11)[2022−01−01]. https://arxiv.org/abs/1810.04805.pdf.

[26] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]// International Conference on Learning Representations. Cambridge: MIT Press, 2015: 1768−1776.

[27] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7132−7141.

[28] GATYS L A, ECKER A S, BETHGE M. Image style transfer using convolutional neural networks[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 2414−2423.

[29] WANG Zihao, LIU Xihui, LI Hongsheng, et al. CAMP: cross-modal adaptive message passing for text-image retrieval[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2020: 5763−5772.

[30] FUKUI A, PARK D H, YANG D, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding[EB/OL]. (2016−06−06)[2022−01−01]. https://arxiv.org/abs/1606.01847.pdf.

作者简介:

亢洁, 副教授, 主要研究方向为机器学习、模式识别。近几年主持和参与教学科研项目 20 余项, 授权发明专利 2 项, 发表学术论文 20 余篇。E-mail: kangjie@sust.edu.cn。

刘威, 硕士研究生, 主要研究方向为数字图像处理、多模态表示学习。E-mail: 535473833@qq.com。