智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

面向装修案例智能匹配的跨模态检索方法

亢洁, 刘威

引用本文:
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.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202106012

DOI: 10.11992/tis.202106012
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20220419.0836.002.html

面向装修案例智能匹配的跨模态检索方法

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: 根据用户输入的文本信息为其实时推送相应风格的装修案例是家装客服系统中的重要功能。然而, 目前该功能的实现主要依赖于人工方式, 不仅不能满足用户对咨询服务快捷、及时的需求, 还增加了企业的人力成本。为此, 提出了一种面向装修案例智能匹配的跨模态检索方法。针对现有算法难以直接建立文本与装修案例之间的对应关系这一问题, 设计了一种风格聚合模块, 可以获取一组装修案例统一的风格特征, 从而便于后续网络建立文本与装修案例之间的潜在语义关联, 实现两者间的跨模态匹配。同时, 在关注图像模态中难易样本分类问题的基础上, 构建了一种双重损失函数对模型进行训练。实验结果表明, 本文提出的方法在装修案例多模态数据集上取得了较好的检索效果。

关键词: 文本信息; 风格; 装修案例; 家装客服系统; 智能匹配; 跨模态检索; 风格聚合; 双重损失函数
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2022)04−0714−07

中文引用格式: 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.
英文引用格式: 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.

A crossmodal retrieval method for intelligent matching of decoration cases

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

Abstract: An important function in the customer service system for home decoration is providing users with decoration cases of corresponding styles in real-time based on the text information input by users. However, the current realization of this function mainly relies on the manual method, which not only fails to meet users' demand for quick and timely consulting services but also increases the labor cost of enterprises. This paper proposes a crossmodal retrieval method for intelligent matching of decoration cases to that end. Aiming at the problem that the existing algorithms cannot directly establish the correspondence between texts and decoration cases, a style aggregation module is designed to obtain the uniform style feature of a set of decoration cases, to facilitate the subsequent network to establish a potential semantic relationship between texts and decoration cases and realize crossmodal matching between them. Simultaneously, a dual loss function is constructed to train the model based on the problem of classifying difficult and easy samples in the imaging modality. The experimental results show that the method proposed in this paper achieves better retrieval results on the multimodal dataset of decoration cases.

Keywords: text information; style; decoration cases; the customer service system for home decoration; intelligent matching; crossmodal retrieval; style aggregation; dual loss function

“互联网+”背景下, 家装企业利用互联网平台搭建客服系统, 以此树立企业形象、营销企业产品, 并为用户提供与装修相关的咨询和服务。用户在使用客服系统的过程中, 经常需要检索不同风格的装修案例, 因此装修案例检索是客服系统中一项重要的功能。目前, 在家装企业的客服系统中, 关于装修案例的检索主要通过人工方式来实现, 即客服人员根据用户需求为其推送具有相

收稿日期: 2021−06−08.　**网络出版日期:** 2022−04−19.
基金项目: 陕西省重点研发计划项目(2021GY-022); 西安市科技计划项目(2019216514GXRC001CG002-GXYD1.7); 国家留学基金项目(201708615011).
通信作者: 亢洁. E-mail: kangjie@sust.edu.cn.

应风格标签的装修案例。这种方法不仅增加了人力资源的消耗，而且实时性较差，在一定程度上降低了企业的服务质量。装修案例检索归根结底是一个通过文本信息检索图像信息的过程，随着信息检索技术的高速发展[1-4]，这一任务可以通过跨模态图文检索的方法来完成。这类方法仅利用文本和图像自身包含的内容信息，就能实现图像与文本间的相互匹配[5]，这使得客服系统能够根据用户输入的话语自动检索相应风格的装修案例并推送给用户，从而降低企业的人力成本，实现信息的实时回复。

目前，构建公共子空间已经成为跨模态图文检索的主流方法，其核心思想是对不同模态数据间的关系进行建模，学习一个公共的表示空间，在该空间中可以直接对来自不同模态的样本进行比较[6]。其中，代表性工作有典型相关分析 (canonical correlation analysis, CCA)[7]，多视角判别分析（multi-view discriminant analysis, MvDA）[8]，联合表示学习算法（joint representation learning, JRL）[9]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学习公共空间的投影矩阵。近年来，由于深度学习在单模态信息处理中的优异表现，相关学者开始将其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[10-12]。其中代表性的工作包括 Wei 等[13] 提出了一种深度语义匹配（deep semantic matching, Deep-SM）方法来解决带有一个或多个标签的跨模态图文检索问题。笔者利用在 ImageNet 上预先训练的卷积神经网络来提取图像特征，验证了使用卷积神经网络提取的图像特征在跨模态检索中容易获得更好的结果。Wang 等[14] 提出了一种对抗跨模态检索（adversarial cross-model retrieval, ACMR）方法，其以对抗学习的思想来拟合不同模态数据的分布，同时对投影空间施加三元组约束，以最小化不同模态中语义相同的样本间的差距，并最大化语义不同的样本间的距离。Zhen 等[15] 提出了一种深度监督跨模态检索（deep supervised cross-modal retrieval, DSCMR）方法，从 3 个角度考虑并设计了损失函数，使得网络学习到的公共空间具有更强的判别能力，显著提升了跨模态检索的性能。然而，上述方法并不能直接应用于家装客服系统中。因为在这些方法中，一张图片就是一类信息，当使用文本检索图片时，输出结果是多张与输入文本相似的图片；而本场景中，包含多张图片的一组装修案例表示一类信息，当用户输入文本进行检索时，希望得到的结果应该是多组与输入文本相似的装修案例。同时，由于一些风格的装修案例在视觉上非常接近，如美式和欧式，而另一些装修案例在视觉上的差别较大，如现代简约和古典，所以装修案例多模态数据集中的图像数据存在难易样本不均衡的问题。

为解决上述问题，本文提出了一种面向装修案例智能匹配的跨模态检索方法，该方法主要有以下 3 个创新点：

1）为了缓解人力资源的消耗，实现客服系统中装修案例自动检索的功能，本文提出了一种面向家装领域客服系统的跨模态图文检索模型。

2）结合应用场景，本文提出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风格特征进行处理，得到可以代表这组装修案例整体风格的一个特征表示，使得装修图片可以按组与文本信息建立联系。

3）针对装修案例多模态数据集中图像样本难易不均衡的问题，设计了一种双重损失函数来对模型进行监督学习。

1　方法模型

本文所提出的模型利用深度神经网络提取文本和图像的特征，并将两者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。本节首先介绍了模型的整体框架，之后分别对模型中用到的风格聚合模块和损失函数进行了介绍。

1.1　跨模态图文检索模型的整体框架

本文所提模型的整体框架如图 1 所示，其包含两个子网络，分别用于处理文本信息和图像信息。假设数据集中包含 n 个文本−图像对，用 $M=\left\{\left(x_i^{\alpha},x_i^{\beta}\right)\right\}_{i=1}^{n}$ 表示，其中 x_i^{α} 表示第 i 个样本中的文本信息，与客服系统中用户输入的话语对应；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 $\boldsymbol{y}_i=[y_{1i}\ y_{2i}\ \cdots\ y_{ci}]\in\mathbf{R}^c$ 表示，其中 c 表示输入样本的类别数。当第 i 个样本属于第 j 类时 $y_{ji}=1$，否则 $y_{ji}=0$。下面，本文以第 i 个输入样本为例来介绍整个模型的工作流程。

首先，模型需要提取输入样本中不同模态信息的特征。针对文本信息 x_i^{α}，本文采用在维基百科中文数据集上预先训练的 BERT（bidirectional encoder representation from transformers）模型[16] 来提取输入文本的语义特征，将模型输出中 [CLS] 标志位对应的一个 768 维的向量作为整个文本的

特征表示，记作$\boldsymbol{h}_i^{\alpha}$。图像信息$x_i^{\beta}$是包含 k 张图片的一组装修案例，本文希望得到这组图片的整体风格特征。由文献 [17] 可知，纹理可以描述一个图像的风格，而卷积神经网络浅层的特征图含有大量的纹理信息，所以图像经过卷积神经网络提取出的浅层特征可以作为该图像风格的特征表示。因此，本文采用在 ImageNet 上预先训练的 VGG19[18] 来提取输入图像的纹理信息，将网络 block1 层输出的特征图作为对应图像的纹理特征，纹理特征的大小为64×112×112，则一组装修案例的纹理特征表示为$\boldsymbol{g}_i^{\beta}=\left\{\boldsymbol{g}_{i1}^{\beta},\boldsymbol{g}_{i2}^{\beta},\cdots,\boldsymbol{g}_{ik}^{\beta}\right\}$，其中$\boldsymbol{g}_{ik}^{\beta}$是指一组装修案例中第 k 张图片的纹理特征。之后，使用本文提出的风格聚合模块对$\boldsymbol{g}_i^{\beta}$进行处理，得到一个大小为64×64的特征图，最终将其展开成一个 4 096 维的向量作为整组装修案例的风格特征表示，记作$\boldsymbol{h}_i^{\beta}$。

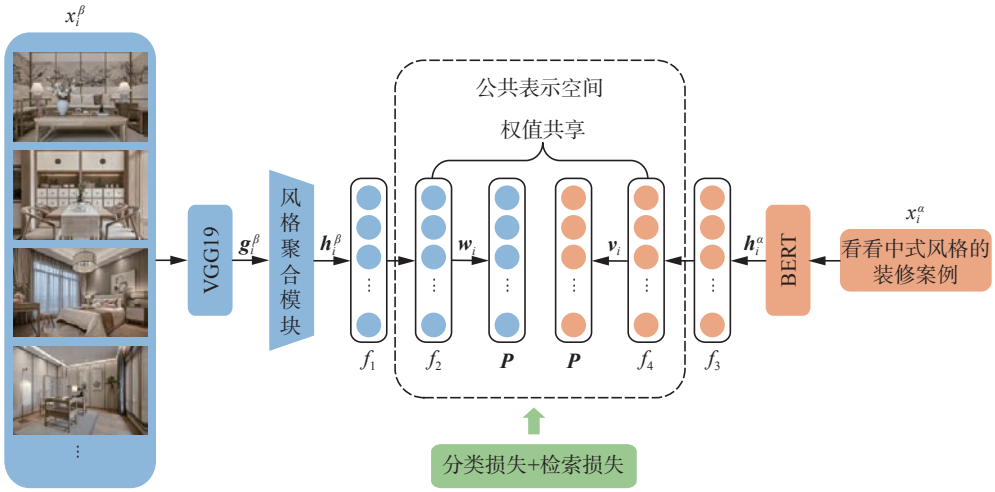

图 1　所提模型的整体框架

Fig. 1　Framework of the proposed method

接着，在两个子网络的后面分别添加两个具有激活函数 ReLU 的全连接层，用f_1、f_2、f_3和f_4表示。利用损失函数对模型进行监督学习，通过这些全连接层将提取到的文本特征$\boldsymbol{h}_i^{\alpha}$和图像特征$\boldsymbol{h}_i^{\beta}$投影到一个公共子空间中，为文本和图像生成统一的特征表征形式，从而使两者可以直接进行相似性比较。其中，全连接层f_1和f_3的隐藏单元数量均为 1 024，全连接层f_2和f_4的隐藏单元数量均为 512，且权值共享。数学上等同于为这两种模态信息分别学习一种映射关系，表示为：$\boldsymbol{v}_i=f\left(x_i^{\alpha};\Phi_{\alpha}\right)\in\mathbf{R}^d$和$\boldsymbol{w}_i=g\left(x_i^{\beta};\Phi_{\beta}\right)\in\mathbf{R}^d$，其中$\boldsymbol{v}_i$和$\boldsymbol{w}_i$分别表示第 i 个样本中的文本和图像信息在公共空间中的特征向量，d 是它们特征向量的维数，Φ_{α}和Φ_{β}分别代表对应函数中的可训练参数。最后，将参数矩阵为$\boldsymbol{P}$的线性分类器分别连接至 2 个子网络的末端，利用标签信息来区分不同类别的特征。

1.2　风格聚合模块

为了建立文本与对应风格的装修案例之间的联系，实现通过文本信息检索装修案例的任务，本文提出了一种风格聚合模块，如图 2 所示。该模块通过对一组装修案例中所有图片的纹理特征$\boldsymbol{g}_i^{\beta}$进行处理，最终获得该组图片统一的风格特征表示，从而便于后续网络学习装修案例与对应文本之间的语义关系。

风格聚合模块具体的工作内容如下：

1）生成格拉姆（Gram）矩阵：由文献 [17] 可知，图像经过卷积神经网络提取到的浅层特征含有更多的纹理特征，这些特征可以用来描述一个图像的风格。之后，通过计算这些特征的 Gram 矩阵，可以度量各个特征之间的相关性，从而得知哪些特征是同时出现的，哪些特征是此消彼长的等。同时，Gram 矩阵的对角线元素还反映了每个特征在图像中的重要程度，所以 Gram 矩阵可以被用于表征图像的风格。本模块的输入是 k 张图片经过 VGG19 第一个 block 层输出的所有特征图$\boldsymbol{g}_i^{\beta}=\left\{\boldsymbol{g}_{i1}^{\beta},\boldsymbol{g}_{i2}^{\beta},\cdots,\boldsymbol{g}_{ik}^{\beta}\right\}$，其中每张图片对应特征图的大小为64×112×112。之后，计算每张图片对应特征图的 Gram 矩阵，得到 k 张图片的风格特征，大小为64×64。Gram 矩阵计算为

$$G_{ij}=\sum_{m=1}^{H\times W}F_{im}F_{jm}\quad(i,j=1,2,\cdots,C)$$

式中：C、H、W 分别表示输入特征图的通道数、高和宽；F_{im}表示输入特征图第 i 个通道的第 m 个元素；F_{jm}表示输入特征图第 j 个通道的第 m 个元素；G_{ij}表示 Gram 矩阵中第 i 行第 j 列的元素。

2）风格聚合：经过第一步运算后，最终得到一

组装修案例中 k 张图片的风格特征，大小为64×64。接着，将全部的风格特征拼接成一组k×64×64的特征图，并使用大小为k×3×3的卷积核对其进行处理，学习这一组特征图整体的风格特征，获得了一个大小为64×64的特征图，其聚合了一组装修案例中所有图片的风格信息。最后，为了方便后续网络的使用，将得到的特征图展开成一个4096维的向量作为一组装修案例的风格特征。

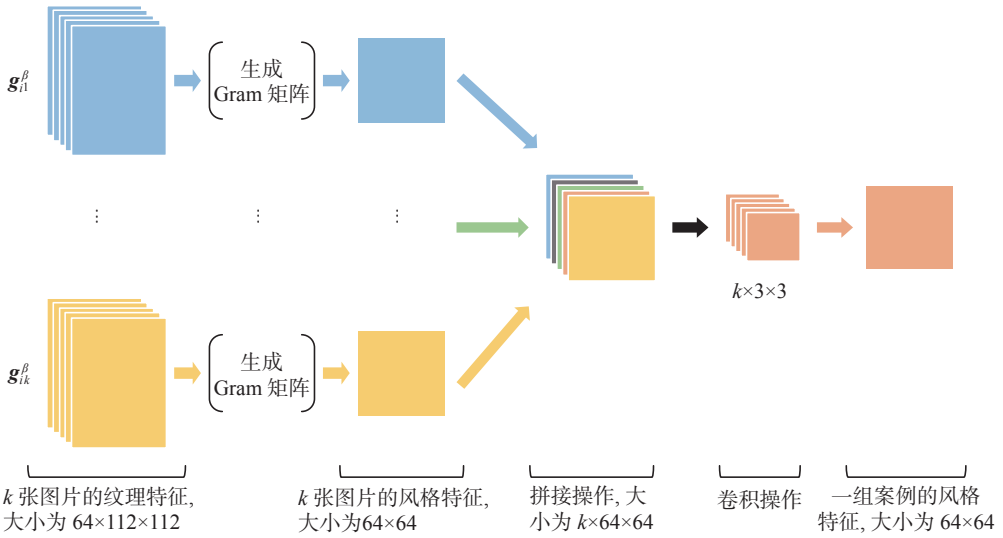

图 2　风格聚合模块

Fig. 2　The module of style aggregation

1.3　损失函数

本文提出的图文检索模型为来自不同模态的特征学习一个公共的向量空间，使得语义类别相同的样本在这个空间是相似的，语义类别不同的样本是不相似的。为此，本文设计了一种双重损失函数 L，它由分类损失L_1和检索损失L_2组成。本文通过最小化 L 来训练网络参数，下面对上述损失函数进行详细介绍。

首先，考虑单模态内的分类损失L_1。对于数据集中的图像样本，因为一些风格的装修案例在视觉感官上非常相似，如美式和欧式，属于不好区分的困难样本；而另一些装修案例在视觉感官上的差别较大，如现代简约和古典，属于容易区分的简单样本，所以数据集中存在图像样本难易不均衡的问题。为此，受文献 [19] 的启发，利用损失函数l_f来优化图像样本的分类，即

$$\begin{cases} l_f = -\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(1-q_i')^{\gamma}\ln(q_i') \\ q_i' = \boldsymbol{y}_i^{\mathrm{T}}\boldsymbol{q}_i \end{cases}$$

式中：$\boldsymbol{y}_i=[y_{i1}\ y_{i2}\ \cdots\ y_{ic}]$表示第 i 个输入样本对应的标签向量；$\boldsymbol{q}_i=[q_{i1}\ q_{i2}\ \cdots\ q_{ic}]$表示模型对第 i 个输入样本的预测概率分布；q_i'表示第 i 个输入样本属于真实类别的概率；n 为输入样本的个数；γ是超参数，用于调节难易样本的分类损失。对于数据集中的文本样本，使用标签平滑的交叉熵损失l_e来学习文本分类，则单模态内的分类损失L_1为

$$L_1 = l_f + l_e \tag{1}$$

之后，考虑跨模态间的检索损失L_2。通过约束公共空间中文本特征和图像特征的相似程度，来消除跨模态差异，优化网络模型。损失函数为

$$L_2 = \frac{1}{n^2}\sum_{i,j=1}^{n}\left(\ln\left(1+\mathrm{e}^{\Gamma_{ij}}\right)-S_{ij}^{\alpha\beta}\Gamma_{ij}\right)+\frac{1}{n}\|\boldsymbol{V}-\boldsymbol{W}\|_F \tag{2}$$

式中：n 为输入样本的个数；$\Gamma_{ij}=\dfrac{1}{2}\cos(\boldsymbol{v}_i,\boldsymbol{w}_j)$；$S_{ij}^{\alpha\beta}=1\{\boldsymbol{v}_i,\boldsymbol{w}_j\}$，$1\{\cdot\}$是一种指示函数，当函数中的两个元素属于同一类别时，其值为 1，反之为 0；$\|\cdot\|_F$表示矩阵的 F 范数；$\boldsymbol{V}=[\boldsymbol{v}_1\ \boldsymbol{v}_2\ \cdots\ \boldsymbol{v}_n]$表示所有文本样本在公共空间中的特征矩阵；$\boldsymbol{W}=[\boldsymbol{w}_1\ \boldsymbol{w}_2\ \cdots\ \boldsymbol{w}_n]$表示所有图像样本在公共空间中的特征矩阵。公式（2）中的第 1 项是根据似然函数重新定义的跨模态负对数似然函数，用于度量不同模态数据间的相似性，似然函数定义为

$$p\left(S_{ij}^{\alpha\beta}|\boldsymbol{v}_i,\boldsymbol{w}_j\right)=\begin{cases}\delta(\Gamma_{ij}), & S_{ij}^{\alpha\beta}=1\\ 1-\delta(\Gamma_{ij}), & S_{ij}^{\alpha\beta}=0\end{cases}$$

式中：$\delta(\Gamma_{ij})=\dfrac{1}{1+\mathrm{e}^{-\Gamma_{ij}}}$是 sigmoid 函数。通过最小化这个似然函数，可以在$S_{ij}^{\alpha\beta}=1$时，提高$\boldsymbol{v}_i$和$\boldsymbol{w}_j$间的相似性；在$S_{ij}^{\alpha\beta}=0$时，降低$\boldsymbol{v}_i$和$\boldsymbol{w}_j$间的相似性。公式（2）中的第 2 项通过计算公共空间中文本–图像对之间的距离损失来关联同一类别不同模态的信息。

综合式 (1) 和式 (2), 可以得到本模型最终的损失函数:

$$L = L_1 + \lambda L_2$$

式中λ是超参数, 负责控制两类损失的贡献程度。

2　实验验证

2.1　数据集和评估标准

为了贴合真实的应用场景, 本文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。本文创建的数据集共包含 7 200 个文本–图像对, 并根据装修风格设置了 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。在一个样本对中, 文本信息是一个与样本标签语义相同的句子, 句子的平均长度为 10.43 个字; 图像信息是与样本标签类别相同的一组装修案例, 不同装修案例中包含 9~13 张数量不等的图片。在实验时, 本文按照 90% 和 10% 的比例将数据集随机划分为训练集和测试集。

本文通过计算文本特征和图像特征之间的余弦值来度量两者的相似性, 并采用了在图文检索中广泛使用的 2 种评估标准: 召回率(Recall@N)和平均精度均值(mean average precision, mAP)对检索算法的性能进行评价。Recall@N 表示输入文本信息后得到的跨模态检索结果中, 前 N 个图像中出现与文本信息类别相同的概率。mAP 与召回率不同, 其综合考虑了所有的检索结果, 对每个测试样本的平均精度进行了再平均, 反映了检索模型的整体性能。上述 2 种评估指标的值越大, 则表示模型的检索能力越强。

2.2　训练与参数设置

根据本文应用场景的实际需求, 本文只考虑以文本检索图像任务中模型的性能。本文实验均在 Python 3.6.12 上进行, 硬件平台为 Intel Coreli7-8700CPU, 内存为 16 GB, GPU 为 11 GB 的 NVIDIA GeForce GTX 2080Ti。模型使用 Adam 优化器进行训练, 学习率为10^{-4}, 平滑常数β_1和β_2分别为 0.5 和 0.999, batch size 为 100, epoch 为 500。本文提出的损失函数中的超参数γ为 2, λ为 0.2。

2.3　实验结果及分析

2.3.1　与现有模型的对比实验

为了验证本文所提模型的有效性, 在本文自建的装修案例多模态数据集上, 将本文提出的模型与多种常见的图文检索模型进行比较, 包括 CCA[7], MvDA[8], JRL[9], CCL[10], ACMR[14] 和 DSCMR[15]。其中, CCA, MvDA 和 JRL 是基于传统统计分析的方法, 其余 3 种是基于深度学习的方法。为了公平, 文本和图像样本均采用 BERT 和 VGG19 预训练模型来提取特征。表 1 所示为不同模型在装修案例多模态数据集上的召回率和平均精度均值, 由实验结果可知, 本文方法相较于次优方法, 在 Recall@5, Recall@10, Recall@15 和 mAP 上分别有了 4.1%, 2%, 3.4% 和 4.4% 的提升, 证明了本文方法在装修案例多模态数据集上的检索性能全面优于其它方法。同时可以看出, 采用深度学习方法学习到的公共表示空间较于传统统计分析的方法具有更好的辨识能力, 可以为跨模态数据建立更强的语义联系, 实现更好的检索性能。

表 1　不同模型的对比实验

Table 1　Comparison of results using different methods

方法	Recall@5	Recall@10	Recall@15	mAP
CCA	0.292	0.357	0.379	0.429
MvDA	0.330	0.415	0.496	0.566
JRL	0.404	0.512	0.558	0.587
CCL	0.468	0.575	0.597	0.609
ACMR	0.534	0.593	0.612	0.635
DSCMR	0.556	0.618	0.631	0.670
本文算法	**0.597**	**0.638**	**0.665**	**0.714**

2.3.2　不同风格特征对模型性能的影响

本文利用在 ImageNet 上预先训练的 VGG19 来提取输入图像的纹理信息。为了研究由 VGG19 不同卷积层的输出特征生成的风格特征对本文所提模型性能的影响, 本文通过提取以下 5 个不同层的图像特征来获得装修案例的风格特征表示, 这 5 个层分别是: block1、block2、block3、block4 和 block5。为了保证这些特征可以在统一的网络中参与计算, 使用卷积操作对 block2、block3、block4 和 block5 输出的特征进行降维, 使得输出特征的通道数都变为 64, 从而获得大小相同的风格特征。表 2 所示为本文模型在使用不同风格特征时, 其在数据集上的召回率和平均精度均值。可以看出, 当模型使用 block1 层的特征生成风格特征时, Recall@5, Recall@10, Recall@15 和 mAP 的值是最高的, 此时模型的检索性能也是最好的。而随着模型使用的卷积层越深, 模型的检索性能也随之下降。这是因为卷积网络的浅层特征包含更多的纹理信息, 深层特征则包含更多的内容信息。而在不同风格的装修案例中, 图片的内

容信息包含很多重叠的内容，如客厅、卧室、厨房和卫生间等，因此包含更多内容信息的深层特征不利于装修案例的跨模态检索。

表 2　不同风格特征的对比实验

Table 2　Comparison of results using different style features

输出层	Recall@5	Recall@10	Recall@15	mAP
block1	**0.597**	**0.638**	**0.665**	**0.714**
block2	0.592	0.632	0.657	0.706
block3	0.583	0.622	0.648	0.695
block4	0.578	0.617	0.644	0.689
block5	0.576	0.617	0.641	0.685

2.3.3　不同损失函数对模型性能的影响

1）损失函数中的超参数分析

本文提出的损失函数 L 包含两部分，分别是单模态内的分类损失 L_1 和跨模态间的检索损失 L_2，并通过超参数 λ 来调节两类损失的贡献程度。图 3 表示本文所提模型使用含有不同 λ 值的损失函数 L 训练后，其在数据集上的 mAP 值。可以看出，当 λ 为 0 时，损失函数只包含分类损失 L_1，没有考虑跨模态样本间的检索损失；当 λ 为 0.2 时，使用损失函数 L 训练的模型在数据集上的 mAP 值最高，此时模型的检索性能最好。之后，随着 λ 值的增大，模型在数据集上的 mAP 值呈下降趋势，模型的检索性能变差。

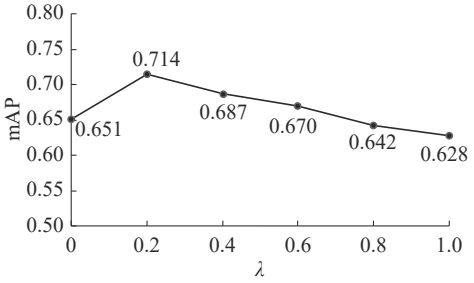

图 3　参数 λ 的对比实验

Fig. 3　Comparison of results using different λ

2）不同损失函数的对比实验

单模态内的分类损失 L_1 由文本分类损失和图像分类损失组成。为了解决数据集中图像样本难易不均衡的问题，本文利用损失函数 l_f 来优化图像样本的分类，同时使用标签平滑的交叉熵损失 l_e 来学习文本分类。为了研究不同损失函数对本文所提模型性能的影响，本文对以下 4 种形式的损失函数进行了评估，分别是：$L_1(l_e+l_e)$（仅使用 l_e 学习文本和图像分类）、$L_1(l_e+l_f)$（使用 l_e 学习文本分类，并使用 l_f 学习图像分类）、L_2（仅使用跨模态检索损失）和 $L_1(l_e+l_f)+\lambda L_2$（使用本文提出的损失函数，其中 λ 设为 0.2）。表 3 所示为使用不同损失函数训练本文模型后，模型在困难样本（如欧式和美式，中式和古典）上的 mAP 和所有样本上的 mAP。可以发现，使用分类损失 $L_1(l_e+l_f)$ 训练的模型，其在困难样本上的 mAP 高于使用 $L_1(l_e+l_e)$ 训练的模型，这说明分类损失 $L_1(l_e+l_f)$ 缓解了数据集中图像样本难易不均衡的问题，提高了模型整体的检索性能。同时，通过比较 $L_1(l_e+l_f)$、L_2 和 $L_1(l_e+l_f)+\lambda L_2$ 可知，单独一种损失函数训练的模型在数据集上的 mAP 都低于共同训练的模型，证明只有同时考虑单模态内的分类损失 $L_1(l_e+l_f)$ 和跨模态间的检索损失 L_2，使用 $L_1(l_e+l_f)+\lambda L_2$ 损失函数训练的模型才具有更好的跨模态检索性能。

表 3　不同损失函数的对比实验

Table 3　Comparison of results using different loss function

损失函数	mAP				mAP
	欧式	美式	中式	古典	
$L_1(l_e+l_e)$	0.505	0.511	0.566	0.572	0.602
$L_1(l_e+l_f)$	0.583	0.592	0.624	0.633	0.651
L_2	0.413	0.405	0.465	0.471	0.504
$L_1(l_e+l_f)+\lambda L_2$	**0.652**	**0.655**	**0.682**	**0.690**	**0.714**

3　结束语

本文针对家装客服系统中装修案例的检索问题，提出了一种基于深度学习的装修案例跨模态检索方法。该方法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处理，得到该组装修案例统一的风格特征表示，方便网络建立查询语句与装修案例之间的联系。同时，本文提出了一种改进的损失函数，用于学习多模态数据在公共空间中的特征表示，并提升了数据集中图像难样本的分类效果。实验结果表明，本文所提方法在自建的数据集上有较好的检索效果，可以将其应用在家装客服系统中，以实现装修案例自动检索的功能。

在未来的工作中，将会构建类型更加多样、内容更加具体的数据集，研究针对某一特定装修案例的跨模态检索模型，进一步完善家装客服系统中装修案例检索这一功能。

参考文献：

[1] CAO Da, YU Zhiwang, ZHANG Hanling, et al. Video-based cross-modal recipe retrieval[C]//MM '19: Proceed-

ings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019: 1685−1693.

[2] 李佳敏, 刘兴波, 聂秀山, 等. 三元组深度哈希学习的司法案例相似匹配方法 [J]. 智能系统学报, 2020, 15(6): 1147–1153.
LI Jiamin, LIU Xingbo, NIE Xiushan, et al. Triplet deep Hashing learning for judicial case similarity matching method[J]. CAAI transactions on intelligent systems, 2020, 15(6): 1147–1153.

[3] MORIK M, SINGH A, HONG J, et al. Controlling fairness and bias in dynamic learning-to-rank[C]//SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2020: 429−438.

[4] WU Fei, JING Xiaoyuan, WU Zhiyong, et al. Modality-specific and shared generative adversarial network for cross-modal retrieval[J]. Pattern recognition, 2020, 104: 107335.

[5] 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, 2019: 5764−5773.

[6] PENG Yuxin, HUANG Xin, ZHAO Yunzhen. 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.

[7] RASIWASIA N, PEREIRA J C, COVIELLO E, et al. A new approach to cross-modal multimedia retrieval[C]// MM '10: Proceedings of the 18th ACM international conference on Multimedia. New York: ACM, 2010: 251−260.

[8] KAN Meina, SHAN Shiguang, ZHANG Haihong, et al. Multi-view discriminant analysis[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(1): 188–194.

[9] ZHAI Xiaohua, PENG Yuxin, XIAO Jianguo. Learning cross-media joint representation with sparse and semisupervised regularization[J]. IEEE transactions on circuits and systems for video technology, 2013, 24(6): 965–978.

[10] PENG Yuxin, QI Jinwei, HUANG Xin, et al. CCL: cross-modal correlation learning with multigrained fusion by hierarchical network[J]. IEEE transactions on multimedia, 2017, 20(2): 405–420.

[11] ZHANG Yifan, ZHOU Wengang, WANG Min, et al. Deep relation embedding for cross-modal retrieval[J]. IEEE transactions on image processing, 2020, 30: 617–627.

[12] CHEN Shizhe, ZHAO Yida, JIN Qin, et al. Fine-grained video-text retrieval with hierarchical graph reasoning[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 10635−10644.

[13] WEI Yunchao, ZHAO Yao, LU Canyi, et al. Cross-modal retrieval with CNN visual features: a new baseline[J]. IEEE transactions on cybernetics, 2016, 47(2): 449–460.

[14] WANG Bokun, YANG Yang, XU Xing, et al. Adversarial cross-modal retrieval[C]//MM '17: Proceedings of the 25th ACM international conference on Multimedia. New York: ACM, 2017: 154−162.

[15] 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, 2019: 10386-10395.

[16] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. New York: arXiv, 2018. (2018−10−11)[2021−06−06].https://arxiv.org/abs/1810.04805.

[17] 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.

[18] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. San Diego: Mendeley, 2015: 1768−1776.

[19] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2999−3007.

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。

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。