

基于深度学习的多模态服装风格检索

陈佳芸

天津工业大学软件学院, 天津

收稿日期: 2023年2月18日; 录用日期: 2023年3月20日; 发布日期: 2023年3月29日

摘要

近年来虽然深度学习在服装检索领域有了很多不错的成果, 但是研究者们对服装风格研究却很少。消费者往往通过自己喜欢的风格来检索搭配的服装, 或者是消费者更愿意检索到与自己穿衣风格相似的服装。现有的服装风格研究者只是将服装风格进行分类, 通过用户输入图像为消费者识别喜欢的风格, 然而这样的检索结果只能返回与该图像风格相似的服装, 而不能与输入的图像完成搭配。因此本文从服装风格的整体兼容性出发, 将每件服装单品视作单词, 按照Word2vec中的单词相似性概念, 分别提出了基于文本的风格检索模型以及基于图像的风格检索模型。最后将获取到的两种模态信息进行特征融合, 提出了一个多模态风格检索模型。实验结果表明, 在Polyvore多模态数据集上, 按照前人研究者的服装风格相似性评判标准, 多模态融合的服装风格检索方法比单模态风格检索以及其他多模态混合风格检索方法所获取的结果列表的平均相似度更佳。

关键词

服装风格, 多模态, 检索

Multimodal Fashion Style Retrieval Based on Deep Learning

Jiayun Chen

School of Software, Tiangong University, Tianjin

Received: Feb. 18th, 2023; accepted: Mar. 20th, 2023; published: Mar. 29th, 2023

Abstract

Although deep learning has achieved many good results in the field of Fashion retrieval in recent years, researchers have little research on Fashion style. Consumers often search the matching clothing through their favorite style, or consumers are more willing to search the clothing similar

to their own style. The existing Fashion style researchers only classify the Fashion style and identify the favorite style for consumers through the user's input image. However, such retrieval results can only return the clothing similar to the image style, but can not match with the input image. Therefore, starting from the overall compatibility of Fashion style, this paper regards each piece of clothing as a word, and proposes text-based style retrieval model and image-based style retrieval model respectively according to the concept of word similarity in Word2vec. Finally, a multimodal style retrieval model is proposed based on feature fusion of the two modal information obtained. The experimental results show that on the Polyvore multimodal data set, according to the previous researchers' Fashion style similarity evaluation criteria, the multimodal fusion Fashion style retrieval method has better average similarity than the single modal style retrieval and other multimodal hybrid style retrieval methods.

Keywords

Fashion Style, Multimodality, Retrieval

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

消费者在购买平台挑选衣服时不仅仅讲究是否合体，而且对衣服的款式、展示的风格有着更高的需求。用户也更偏向于选购自己喜欢的风格的服装。近年来对服装的检索多是基于用户购买历史数据来为消费者推荐服装[1]，如果消费者没有历史购买数据，或者数据较为杂乱，那么如何在没有这些背景信息的情况下了解服装风格不但能够帮助消费者选择衣服，而且能够协助设计者和制造商更好地设计消费者所要求的衣服。识别一件服装的风格是十分具有挑战性的，近年来，一个较为普遍的办法是将服装预定义为如图 1 所示，类似 a：学院风，b：休闲风，c：朋克风等类型的风格集。最近使用该方法的是[2] [3] [4]，风格集是一些不同类别但风格相似的集合。但是使用人工来预定义风格的方法，几乎不可能精确的定义一组足以描述风格这种微妙的属性，在大量投入工作量的同时，还可能出现错误的风格分类。另外还有研究者通过主题模型来分析一件服装潜在的风格[5]。然而这些研究只是为消费者识别出服装的风格，而我们要做的是为消费者找出风格相似或者风格兼容可搭配的其他服装单品。



Figure 1. Style set

图 1. 风格集

在本文中,借鉴[6]在家具风格上的思想,我们提出了一个文本风格检索模型来检索相似的服装风格,将一整套服装集作为一个风格集,使用词向量模型 Word2vec 的思想,其中里面的每一件服装单品文本描述作为中心词,与之同一个风格集中的其他服装单品作为其上下文信息,将这些信息输入模型进行训练,获取到最终的模型。由于文字可表达的信息有限,本文另外提出一个图像风格检索模型,使用 Vgg16 提取图像特征,将选取的服装单品图像表示为一个向量,将同一风格兼容集的其他上下文项目映射为其权重向量。利用 softmax 函数将目标项向量与各上下文权重向量的内积化为概率,训练整个网络使概率之和最大化。最后,提出一个多模态融合模型,将文本与图像风格检索所获取到的特征进行融合,利用融合特征进行相似性检索。通过该模型在一个多模态数据集上的实验结果来看,本文提出的结合多模态特征融合的服装风格检索模型所得到的检索结果相较于单模态以及其他的多模态混合方法所获取的检索结果更佳。

2. 相关知识

2.1. 卷积神经网络

卷积神经网络是为了处理图像而生,其设计原理是把图像转换为像素值数字矩阵,由此,将图像转换为计算机可识别的形式,从而得到图像相应的具体特征空间。卷积神经网络由于引入权值共享和局部感受野等特性,可以在保证网络鲁棒性的前提下,使用卷积方法大大减少计算步骤。池化层可以有效的减小参数矩阵的尺寸,可以加快计算速度和防止过拟合的作用,全连接操作就是将不同维度特征连接在一起,整合特征提取层获取的较为重要的权重信息。如图 2 所示,这就是一个典型的神经网络结构。

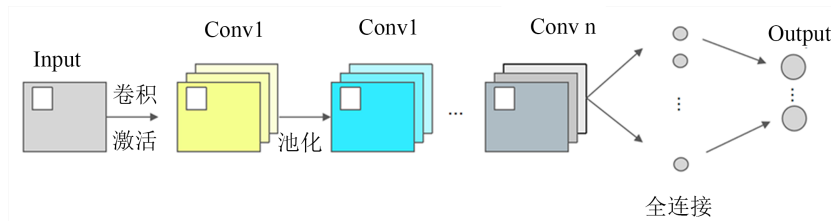


Figure 2. Convolution neural network diagram
图 2. 卷积神经网络图

2.2. Word2vec 模型

Word2vec [7]将词汇表中的每个词表示为一个嵌入空间中的低维密集向量,并尝试从训练语料库中学习词向量,通过词之间的空间距离衡量词之间的相似性,例如,两个词在所表示的嵌入空间中越近,它们在语义和语法上就越相似。这些单词表示是基于分布假设学习的,该假设上下文相似的单词往往具有相似的含义。在这种假设下,word2vec 提出了两种不同的模型架构:上下文 Bag-Of-Words (CBOW)和跳字模型 Skip-Gram with Negative Sampling (SGNS),用于从周围上下文中预测目标单词,接下来对 CBOW 做介绍:

CBOW 模型共有三层结构,如图 3 所示,分别是输入层 Input layer、隐藏层 Hidden layer、输出层 Output layer。该模型的具体计算过程如下描述,将目标词的上下文中的词的独热编码首先输入输入层;然后在隐藏层中,将输入的这些词同时乘以同一个矩阵后,将不同词得到的向量通过加权平均得到一个向量,最后将这个向量乘以矩阵后得到最终的一个向量。在输出层是一个 softmax 回归分类器,将最终得到的向量经过 softmax 归一化后,在该层的每个结点将会输出一个概率向量,将预测值最高的向量作为预测的结果向量,将其与该词的真实标签向量做误差计算。一般使用交叉熵损失函数来做计算,

在经过每次前向传播与反向传播误差，从而不断变换上述中和的值。使得结果向量与真实标签向量误差最小化，在进行预测时，通过向前传播可以得到预测的中心字的结果。

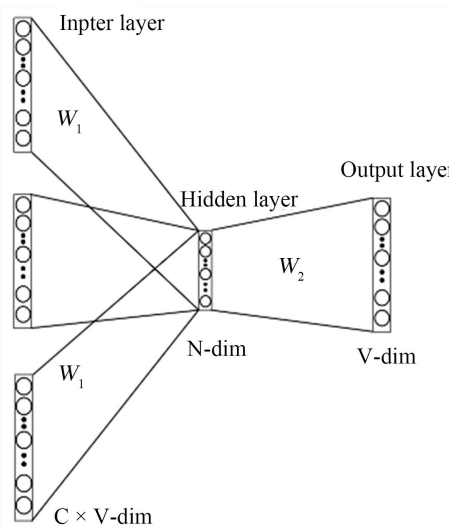


Figure 3. CBOW model diagram
图 3. CBOW 模型图

由于 softmax 做计算比较复杂，使用负采样方法提高计算效率。负采样的思路是将由上下文词生成中心词或由中心词生成上下文词的过程拆分为两种独立事件的组合：正样本事件和负样本事件。

在 CBOW 模型中，假设每个词都用维度为 d 的向量表示，从数据集所得词典为 \mathbb{R}^d 。对从词典中于所选中的第 i 个目标词，使用 $\mathbf{v}_i \in \mathbb{R}^d$ 和 $\mathbf{u}_i \in \mathbb{R}^d$ 表示第 i 个词的上下文词和该词的向量表示。假设由上下文词 $W_o = \{w_{o_1}, w_{o_2}, \dots, w_{o_{2m}}\}$ 生成中心词 W_c 是由两种独立的事件来近似，而：

正样本事件：上下文词与中心词出现在同一个窗口内，概率 $P(D=1|W_o, w_c) = \sigma(\mathbf{u}_c^T \bar{\mathbf{v}}_o)$ ；

负样本事件：上下文词 W_o 与噪声词 w_k 同时出现在窗口内，噪声词 w_k 按指的是非上下文词对应的中心词，其分布 $P(w)$ 由随机生成。设噪声词 w_k 在词典中的索引为 h_k ，因为噪声词也可以是中心词，则噪声词 w_k 的向量可表示为 \mathbf{u}_{h_k} ，则概率可以写为 $P(D=0|W_o, w_k) = 1 - \sigma(\mathbf{u}_{h_k}^T \bar{\mathbf{v}}_o) = \sigma(-\mathbf{u}_{h_k}^T \bar{\mathbf{v}}_o)$ 。

所以 $P(w_c | W_o) = \frac{\exp(\mathbf{u}_c^T \bar{\mathbf{v}}_o)}{\sum_{i=1}^{|V|} \exp(\mathbf{u}_i^T \bar{\mathbf{v}}_o)}$ 可以改写为正样本事件与负样本事件的联合概率，如式(2-1)所示：

$$\begin{aligned}
 P(w_c | W_o) &= P(D=1|W_o, w_c) \prod_{k=1}^K P(D=0|W_o, w_k) \\
 &= \sigma(\mathbf{u}_c^T \bar{\mathbf{v}}_o) \prod_{k=1}^K \sigma(-\mathbf{u}_{h_k}^T \bar{\mathbf{v}}_o)
 \end{aligned} \tag{2-1}$$

其损失函数为式(2-2)所示：

$$\begin{aligned}
 \text{loss} &= -\log P(w_c | W_o) \\
 &= -\log \left[\sigma(\mathbf{u}_c^T \bar{\mathbf{v}}_o) \prod_{k=1}^K \sigma(-\mathbf{u}_{h_k}^T \bar{\mathbf{v}}_o) \right] \\
 &= -\log \left[\sigma(\mathbf{u}_c^T \bar{\mathbf{v}}_o) \right] - \sum_{k=1}^K \log \left[\sigma(-\mathbf{u}_{h_k}^T \bar{\mathbf{v}}_o) \right]
 \end{aligned} \tag{2-2}$$

3. 模型介绍

3.1. 英文缩写

为了实现利用词向量模型进行风格检索的功能, 借鉴[6]的思想, 选择使用 Word2Vec [7]中的连续词袋模型(Continuous Bag-of-Words, CBOW), 由 2.2 中介绍的其原理, 类比于 cbow 中的中心词与上下文信息, 将每一件服装单品的文本描述作为一个中心词, 一套风格兼容集中的其他服装单品文本描述作为这个词的上下文信息, 如图 4 所示, 用该风格兼容集中服装图像的文字描述来形成上下文信息, 按照 CBOW 模型的思想先把上下文信息输入输入层中, 再将中心词作为输出向量, 来训练 CBOW 模型。得到这样一个训练好的模型之后, 利用这个模型来计算风格相似度。总结来说, 如果两件服装单品出现在一个风格兼容集中, 那么它们的风格是相似的, 如果两件服装单品不在同一兼容集中, 但是它们所在风格兼容集里其他服装单品是相同或者相似的, 即对应 CBOW 里有相同或者相似的上下文信息, 那么两个服装单品也是相似的。

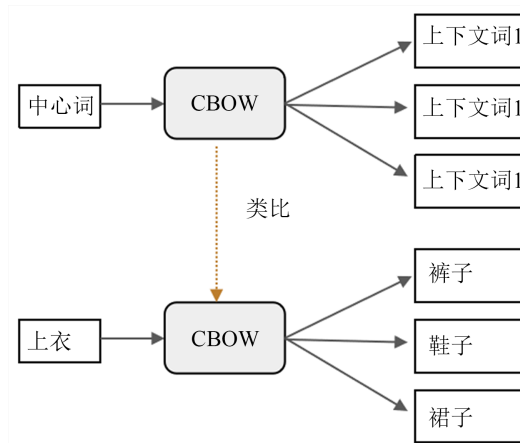


Figure 4. Analogy
图 4. 类比图

3.2. 基于图像的风格检索模型

基于文本所表达的特征有限, 本文另外训练一个基于图像上文风格检索模型, 参考 3.1 中词向量检索风格的概念, 在使用图像检索相似风格任务时, 使用两个经典的 Vgg16 网络模型, 一个用于将选取的服装单品图像投影为一个表示向量, 另一个用于将同一风格兼容集的其他上下文项目映射为其权重向量。利用 softmax 函数将目标项向量与各上下文权重向量的内积化为概率, 训练整个网络使概率之和最大化。具体来说, 让图像中第 i 个项目为 I_i , 将两种类型的网络参数化, 对于输入的图像 I_i 定义为 $f_{in} : I_i \mapsto u_i$, 对于上下文项目 I_c 定义为 $f_{ctx} : I_c \mapsto v_c$ 。两个网络使用相同的 VGGNet 结构, 两个网络不共享参数, 因为这两个网络具有不同的目的。两者的目标是学习 f_{in} 和 f_{ctx} , 使项目在样式集中同时出现的对数概率最大化。设 F 为所有时尚单品的集合。假设的训练数据集 $D \subset \mathcal{P}(F)$ 中的每个风格集 S 都具有一致的风格。迭代 S 中的 i 项作为输入, S 中除了 i 之外的项作为上下文项。模型的总体目标是最大化 D 中每个 S 的平均 log 概率。公式如式 3-1 所示:

$$\frac{1}{|S|} \sum_{i \in S} \sum_{c \in S \setminus \{i\}} \log \left(\frac{\exp(u_i^\top v_c)}{\sum_{j \in F} \exp(u_i^\top v_j)} \right) \quad (3-1)$$

其中 $u_i = f_{in}(I_i), v_c = f_{ctx}(I_c)$ and $v_j = f_{ctx}(I_j)$ 。

与文本计算一样，由于计算 softmax 计算量非常大，因此同样也引入负采样的思想，随机选取 k 个不同于训练样本 S 中的项目，对每个 (i, c) 形成负样本集 $N_{S,i,c} \subset F \setminus S$ ，将目标函数变成如式 3-2 所示：

$$\frac{1}{|S|} \sum_{i \in S} \sum_{c \in S \setminus \{i\}} \left(\log \sigma(u_i^\top v_c) + \sum_{j \in N_{S,i,c}} \log \sigma(-u_i^\top v_j) \right) \quad (3-2)$$

使得同一风格集中使用图像进行检索时也能获取到其同一风格兼容集中的图像，同时，也能使得两个有相似上下文信息的服装具有风格相似性。服装图像风格检索模型如图 5 所示。

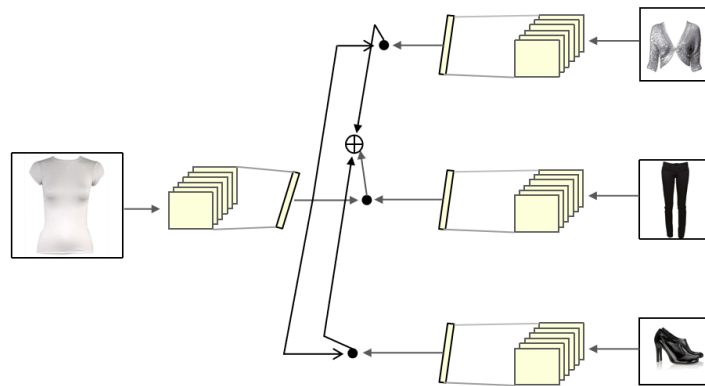


Figure 5. Image style retrieval of clothing
图 5. 服装图像风格检索图

3.3. 基于多模态融合的服装风格相似性

多模态服装风格检索模型如图 6 所示：分为三个模块，第一个为特征提取模型，分别使用 Vgg16 模型提取图像特征，使用 Word2vec 中的 CBOW 模型获取文本特征，在特征融合模块，最后将两者特征使用一种特征融合的方式，即级联，将获取的两者特征融合。最后相似性检索模块中，softmax 层通过计算，对最终结果进行预测，按照预测的相似度值的高低，返回预测检索结果列表。

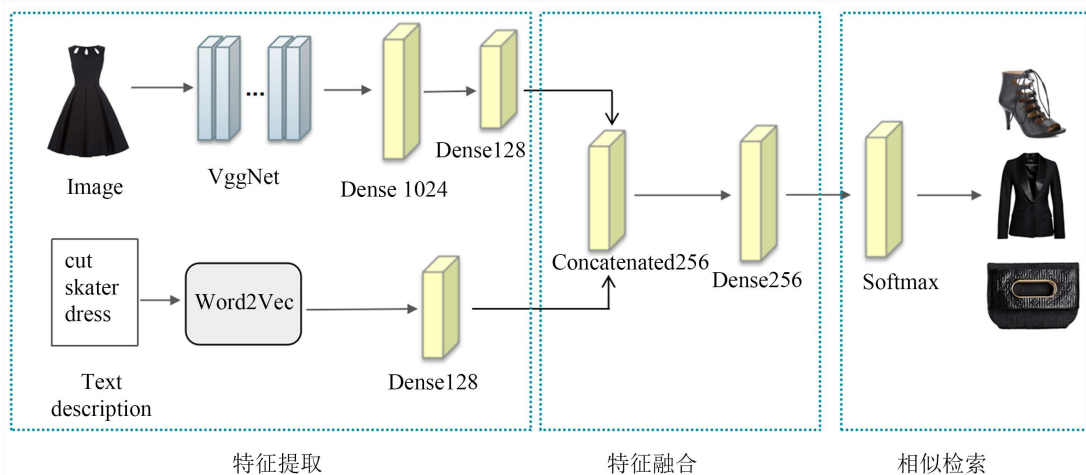


Figure 6. Multimodal clothing style search diagram
图 6. 多模态服装风格检索图

4. 实验

4.1. 实验概况

对于利用文本进行风格相似性检索时, 采用 word2vec 中的 CBOW 模型, 将数据集中服装单品文本描述看作是一个单词, 将一套服装的文本描述看作是一句话, 如此来训练该风格检索模型。同时用负采样方法对模型中的复杂计算进行加速, 词向量维度设置为 400, 窗口大小设置为 5, 和其他文本分类任务不同, 本实验对于所有出现过的单词都考虑进去。

对于本文提出的图像风格检模块, 使用经典的 Vggnet 作为特征提取器, Vggnet 的结构 5 层卷积层、3 层全连接层和 softmax 输出层构成, 每个层与层之间使用最大池化层分开, 所有隐藏层的激活单元都采用 ReLU 函数。通过 CBOW 的思想对网络进行训练, 最终将图像映射为一个 1024 维向量。网络模型中使用自适应矩估计(Adaptive Moment Estimation)优化器以及批归一化(Batch Normalization)进行训练, 收敛速度较快且相对稳定。

4.2 数据源

训练和测试数据来自 Polyvore 数据集[8]。该数据集中共收集了 21,889 套服装, 每一套服装的里面有单独的服装单品对应图像, 一套服装里平均有 5 个服装单品。此次实验训练集样本和测试集样本划分为训练集 17,316 套服装, 测试集 4573 件。每一件服装单品在其套装都有相应的图像和文字说明。

4.3 评估指标

相似性评分: 在这项服装风格检索任务中, 沿用了以往研究者的相似性度量, 如式 4-1 所示, 首先, P 是目录中所有可能的产品项目的集合。然后将 C 表示为包含风格兼容项目的所有集合的集合。检索两个项目 p_1 和 p_2 之间的相似性函数, 确定它们是否合适。风格相似函数 s_c 如式(4-1)所示:

$$s_c(p_1, p_2) = \frac{|\{C_i \in C : p_1 \in C_i \wedge p_2 \in C_i\}|}{\max_{p \in \{p_1, p_2\}} |\{C_j \in C : p \in C_j\}|} \quad (4-1)$$

由于数据集中每件服装与其他服装的关联度不是很高, 研究者添加了一个额外的相似性度量, 如式(4-2)所示, 该度量直接从描述的名称重叠中导出:

$$s_n(p_1, p_2) = \mathbb{1} \{W_{p_1} \cap W_{p_2} \neq \emptyset\} \quad (4-2)$$

综上所述, 如果满足以下两个条件之一, 则被评估的配对被认为是相似的, 即项目在同一套服装中共同出现或者是两个项目的名称在形式上重叠, 记其为, 其计算过程如式(4-3)所示:

$$s_{\max}(p_1, p_2) = \max(s_c(p_1, p_2), s_n(p_1, p_2)) \quad (4-3)$$

列表内相似性: 将最终得到的长度为 k 的生成结果列表 R 的平均列表内相似性定义为如式(4-4)所示:

$$s(p_1, p_2) = \max(s_c(p_1, p_2), s_n(p_1, p_2)) \text{AILS}(R) = \binom{k}{2}^{-1} \sum_{p_i \in R} \sum_{p_j \in R, p_i \neq p_j} s(p_i, p_j) \quad (4-4)$$

即在生成的项目列表中的所有可能的对上计算的平均相似性得分。通过这样做, 可以评估生成项目列表的整体风格相似性, 风格检索结果的顺序无关紧要。

4.4. 实验结果

在将图像与文本风格相似性融合之前, 分别对风格评估标准中的相似性结果进行对比, 分别是 s_c 、 s_n

以及 s_{\max} 。分别从图像与文本两个方式从数据集中检索相似风格，结果返回 k 值为 5~20 之间，比较了图像风格检索以及文本风格检索 s_c 、 s_n 以及两者风格检索最大值 s_{\max} ，两种风格检索得到的相似性相差不大，如图 7 所示。因此，将 4.3 中的列表相似性计算作为最终的风格相似性评判是合理的。

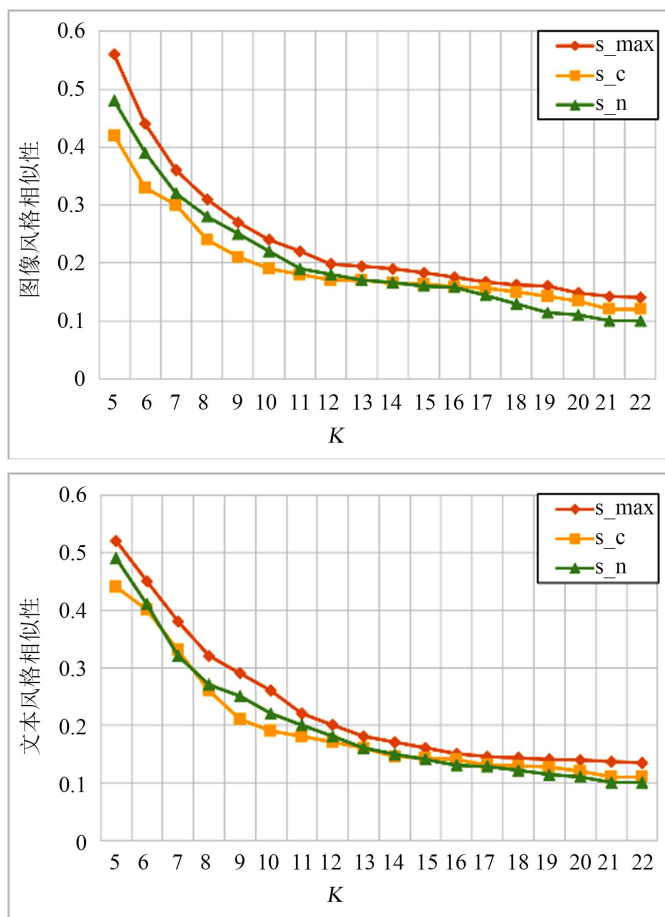


Figure 7. Structure of style retrieval
图 7. 风格检索结构图

对于多模态融合的风格相似性，对比单模态风格检索，以及对比其他多模态融合方法，包括后期融合，即指的是将视觉与文本分别按照相似性顺序返回二分之一一个结果混合的最终结果，以及尝试还原了文献[9]和文献[10]的融合方法：

- A. 文本风格检索；
- B. 图像风格检索；
- C. 后期融合；
- D. 对目标项目使用图像相似性检索(此时的检索与风格无关)之后，再使用文本上下文风格检索，最终将查询文本作为检索条件(此时的检索与风格无关)，最后将查询文本作为检索条件，检索出相似的结果；
- E. 先对目标项目使用图像检索以及文本检索，对检索的最佳结果进行文本检索(此时的检索与风格无关)，将这三部分的结果各取一部分，与目标项目的图像特征距离进行排列，得到最终结果。
- F. 本文的方法，对一个目标的项目分别提取图像与文本特征之后，进行特征融合，通过与融合特征

距离进行排列，得到最终检索结果。

为了比较，本文将相似度按照六个类别的方式进行呈现，分别是 dress、coats、shoes、pants、hats、bags。对于 D、E 则将类别作为查询文本，此时将返回的列表长度设置为 10，进行风格相似度的对比分析，表 1 为结果。分析单模态的上下文风格检索，图像风格检索比文本风格检索精度略高，表明了利用卷积神经网络进行服装图像特征提取的强大能力，同时，也证明使用 word2vec 网络架构的思想，所搭建的神经网络模型来完成服装风格检索任务是可以实现的。另外，通过与后期融合以及其他研究所采用的混合方法对比，本文提出的多模态服装风格检索获得了最高的风格相似性。对表中类别进行详细分析时，发现 bags、hats、shoes 这三类的服装单品相似度比较高，可能是因为这三种是作为配饰，类别并不是很多，没有复杂的颜色以及图案。

Table 1. Comparison of fashion style search results

表 1. 服装风格检索结果比较表

类别	A	B	C	D	E	F
dress	0.27	0.27	0.30	0.24	0.23	0.33
coats	0.21	0.18	0.22	0.27	0.28	0.26
shoes	0.28	0.24	0.27	0.27	0.29	0.30
pants	0.22	0.21	0.23	0.16	0.18	0.24
hats	0.31	0.27	0.32	0.30	0.32	0.32
bags	0.32	0.30	0.31	0.32	0.29	0.36
平均	0.27	0.25	0.27	0.26	0.27	0.30

为了对多个模态的结果有更直观的呈现,将本文的检索结果进行整理进行结果展示。检索文本在图像下标注，检索图像如图 8 所示。



Figure 8. Feature fusion result diagram (text retrieval content: women hooded faux leather jacket)

图 8. 特征融合结果图(文本检索内容: women hooded faux leather jacket)

通过上图 8 进行的可视化展示，证明本文提出的多模态服装风格检索在视觉上有一定的相似性，对于该检索，是比较休闲的风格，所匹配的服装在颜色或者搭配上风格都较为相似。

5. 结束语

本章设计了一个多模态融合的服装风格检索方法。能检索出服装风格上的相似结果。介绍了所使用的 Word2vec 模型的原理以及风格检索模型的整体框架，额外添加了一个服装图像风格检索模型。介绍了融合服装图像和文本两种模态的混合风格检索策略，最后将提出的模型应用于 Polyvore 服装多模态数据集，从获取的服装检索结果来看，本章提出的模型有一定的风格检索效果。

参考文献

- [1] Gharaei, N.Y., Dadkhah, C. and Daryoush, L. (2021) Content-Based Clothing Recommender System Using Deep Neural Network. 2021 26th International Computer Conference, Computer Society of Iran (CSICC), Tehran, 3-4 March 2021, 1-6.
- [2] 李扬, 黄荣, 董爱华. 基于改进 Bilinear-CNN 的服装图像风格识别[J]. 东华大学学报(自然科学版), 2021, 47(3): 90-95.
- [3] Yang, S.-Y., Zhong, Y.-Q. and Wang, X. (2021) Clothing Recommendation with Style Recognition. 2021 14th Textile Bioengineering and Informatics Symposium Proceedings (TBIS 2021), Rube, 6-9 July 2021, 589-596.
- [4] Hu, M.Y. and Zhong, Y.Q. (2020) Convolutional Neural Networks for Fashion Style Classification. *Proceedings of 2020 SAISTA*, Beijing, 12-14 August 2020, 208-212.
- [5] Zhu, M., Jiang, N., Feng, X., et al. (2022) Style Analysis of Clothing from Fashion Shows Based on Deep Learning. 2022 5th International Conference on Data Science and Information Technology (DSIT), Shanghai, 22-24 July 2022, 1-6. <https://doi.org/10.1109/DSIT55514.2022.9943872>
- [6] Tautkute, I., Moejko, A., Stokowiec, W., et al. (2017) What Looks Good with My Sofa: Multimodal Search Engine for Interior Design. 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), Prague, 3-6 September 2017, 1275-1282
- [7] Kenter, T., Borisov, A. and De Rijke, M. (2016) Siamese Cbow: Optimizing Word Embeddings for Sentence Representations. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1, 941-951. <https://doi.org/10.18653/v1/P16-1089>
- [8] Han, X., Wu, Z., Jiang, Y.G., et al. (2017) Learning Fashion Compatibility with Bidirectional LSTMs. *Proceedings of the 25th ACM international conference on Multimedia*, Silicon Valley, 23-27 October 2017, 1078-1086.
- [9] Tautkute, I., Trzcinski, T., Skorupa, A.P., et al. (2019) Deep Style: Multimodal Search Engine for Fashion and Interior Design. *IEEE Access*, 7, 84613-84628. <https://doi.org/10.1109/ACCESS.2019.2923552>
- [10] 苏卓, 柯司博, 王若梅, 等. 深度多模态融合服装风格检索[J]. 中国图象图形学报, 2021, 26(4): 857-871.