文章编号：2095-2163(2022)07-0202-04 中图分类号：TP391 文献标志码：A

# 面向电商的多模态商品检索引擎设计

潘 巍，林榆森，施自凯，林世翔
（哈尔滨华德学院 数据科学与人工智能学院，哈尔滨 150025）

**摘 要：**本文根据电商用户无法有效的检索出符合自身偏好商品的问题，设计出一种面向电商平台的多模态商品检索引擎模型（MCFR-Net）。该模型分别采用 Transformer Transducer 模型提取音频特征、Transformer 模型提取文本特征、Twins-PCPVT 模型提取图像特征；利用 MCB 特征融合技术，将多模态特征进行融合；对融合后的特征向量进行商品的相似度计算，根据相似度阀值来检索商品数据。实验在 KDD Cup 2020 多模态商品数据集上，将本文提出的模型与 LSTM-DSSM 和 DELF 进行对比实验。结果表明，本文提出的算法是一种有效的商品检索模型。

**关键词：**多模态；商品检索；特征融合

## Design of multimodal commodity retrieval engine for e-commerce

PAN Wei, LIN Yusen, SHI Zikai, LIN Shixiang
(School of Data Science and Artificial Intelligence, Harbin Huade University, Harbin 150025, China)

**【Abstract】** According to the problem that users cannot effectively retrieve the commodity that meet their preferences on the e-commerce platform, this paper designs a multimodal commodity retrieval model (MCFR net) for e-commerce platform. It uses Transformer Transducer model to extract audio features, Transformer model to extract text features and Twins-PCPVT model to extract image features respectively for common data for e-commerce. Then MCB feature fusion technology is used to fuse the above multimodal features. Finally, the fused feature vectors are used to calculate the commodity similarity, and the commodity data are retrieved according to the similarity threshold. A multimodal commodity retrieval engine scheme is designed for e-commerce platform. The proposed model is tested through a series of experiments with KDD Cup 2020 multimodal data sets. Compared with LSTM-DSSM and DELF, we are concluded that the proposed model is effective and efficient.

**【Key words】** multimodal; commodity retrieval; feature fusion

## 0 引 言

近年来，随着互联网技术的发展以及人们对生活便利的需求，网购电商平台得到了飞速的发展，网上消费和选购心仪商品成为大多数人的首选。当前，随着网购模式的快速普及，虽然电商平台已经储备了海量的用户商品购买行为数据，但人们在网上购物时，若想在电商平台中有效检索出符合自身偏好的商品却越来越难，该问题的存在使得电商平台的现有检索系统面临着巨大挑战。此外，经常使用电商购物平台（如淘宝，京东和亚马逊）的用户还会发现，这些平台仅支持语音搜索、文本搜索、图片搜索等单模态检索方式，不能满足用户精准定位的商品需求[1-2]。

## 1 多模态商品检索引擎需求分析

商品搜索引擎以多模态商品检索条件数据作为输入，并将这些数据和数据库里的用户行为信息一起提取特征，进行多模态融合得到融合后的特征向量，并把特征向量，构建出一种全新的多模态个性化商品检索引擎，主要应用于电商的多模态商品检索、商品个性化推荐和商品问答机器人。目的是为了提高检索的准确度，提高用户购物体验的满意度。可以说检索引擎为用户带来了线上购物的极大便利，对相关的电商平台带来了巨大的盈利效益。

多模态的商品检索引擎需要处理多种数据类型的数据，如图片，音频和文本信息。如何将多模态数据进行特征表达和融合是其要解决的关键性问题。此外，在现有的电商商品交易系统中存在海量的多

**基金项目：**2020 年度黑龙江省自然科学基金联合引导项目（LH2020C001）；2020 年度黑龙江省高等教育教学改革研究一般研究项目（SJGY20200270）；黑龙江省教育科学“十三五”规划 2020 年度重点课题（GJB1320092）。

**作者简介：**潘 巍（1976-），男，博士，讲师，主要研究方向：人工智能应用、机器学习与数据挖掘；林榆森（2000-），男，本科生，主要研究方向：人工智能应用；施自凯（1999-），男，本科生，主要研究方向：人工智能应用；林世翔（2000-），男，本科生，主要研究方向：人工智能应用。

**收稿日期：**2022-03-01

模态数据，若能从中自动提取出商品特征，有效的检索出用户偏好的商品集合也是其有待解决的重要问题。相比传统的机器学习方法，深度学习可通过多个隐含层的仿射变换来自动提取多种类型数据的特征，并且对于海量数据处理任务表现出极好的学习泛化能力[3]。从而利用深度学习来构建多模态的商品检索引擎是最好的选择。

## 2　基于深度学习的特征表示和提取技术

在多模态的商品检索引擎中，主要提取文字、图像和声音数据的特征，并有效的将其融合。其特征提取可依赖于深度学习技术来完成。

### 2.1　基于深度学习模型的图像处理技术

近年来，基于深度学习的卷积神经网络，在图像识别方面获得了巨大的成功，其可以通过多层卷积操作来获得图像特征的深度表达，如 ResNet、LeNet5、AlexNet、Inception Net 等[4-6]。Vision Transformer 的提出，相较于卷积神经网络来说，使用了一种自注意力机制，该模型的学习能力超越了前面所提到的基于深度学习的神经网络模型[7]。2021 年 3 月，微软公布了 Swin Transformer 模型，该模型使用移动窗口来计算多尺度的图像特征，并减少了模型的计算复杂度[8]。同年，美团和阿德莱德大学提出了 Twins Transformer( Twins-PCPVT )，其设计出空间自注意力机制，使其在图像分类、目标检测和语义分割任务上超越了 Swin Transformer 模型[9]。Twins-PCPVT 通过将 PVT 中的位置编码替换为 CPVT 中提出的条件位置编码 CPE，使其在分类和下游任务上直接获得大幅度的性能提升。尤其是在稠密任务上，由于条件编码 CPE 支持输入可变长度，使得对于图像的处理上，可以灵活处理来自不同空间尺度的特征。

### 2.2　基于深度学习模型的音频处理技术

众所周知，早期的语音识别系统通常由两部分组成：一是利用输入的 waveform，人为提取 MFCC 特征；二是通过分类模型来对声音进行识别。随着深度神经网络技术的发展，可以通过 CNN、DNN、LSTM 等深度神经网络结构来自动化提取特征，相对于非端对端模型，减少了工程的复杂度，并广泛的应用到语音识别中获得了良好的效果。

2006 年以来，虽然基于深度学习的 CTC 模型（如 LSTM-CTC、RNN-CTC 等）在语音识别声学建模上获得了巨大的成功，但该模型也存在如下问题：一是缺乏语言模型建模能力，不能整合语言模型进行联合优化；二是不能构造模型输出之间的依赖关系[10]。针对 CTC 的不足，Alex Graves[11] 提出了 RNN-T 模型。RNN-T 模型巧妙的将语言模型与声学模型整合在一起，同时进行联合优化。2020 年 2 月，谷歌团队提出了 Transformer Transducer[12]。其是一款在 RNN-T 模型基础上，把 LSTM encoding 替换为 transformer encoders 的模型，利用有限宽度的上下文时序信息，在基本不损失精度的条件下，可以满足流式语音识别的要求，获得了巨大成功。

### 2.3　基于深度学习模型的文本处理技术

近年来，NLP 自然语言处理在文本识别方面获得了巨大的成功，可以通过文本嵌入技术来获得文本特征的深度表达。例如 Skip-Gram、Word2vec 和 GloVe 等等[13]。基于深度学习的文本处理任务存在很多模型，如 ABCNN、IndRNN 和 TextCNN 模型等[14]。在此基础上，2017 年谷歌公司提出了基于多头注意力机制的 Transformer 的模型，该模型并没有沿用典型的循环神经网络结构，而是通过多头注意力来学习文本的语义，并在性能方面超越了其它模型[15]。

### 2.4　多模态特征融合技术

众所周知，对于多模态任务，如 VQA、视觉定位等，都需要融合两个模态的特征[16]。近年来，多模态融合最常用的方法是拼接（concatenation）、按位乘（element-wise product）、按位加（element-wise sum）[3]。而多模态紧凑双线性池（MCB）的作者认为，这些简单的操作融合效果不如外积，不足以建模两个模态间的复杂关系[17]。MCB 将外积的结果映射到低维空间中，使其计算更为方便。双线性池化首先对特征提取，得到特征映射每个位置的特征向量进行向量外积计算，然后对所有位置外积计算的结果进行平均池化得到特征向量；最后经过 L2 范数标准化得到最后的特征。

## 3　基于深度学习的多模态的商品检索引擎

根据深度学习的特点，本文设计了一种全新的基于深度学习的多模态商品检索引擎。其整体结构框架如图 1 所示。该引擎的工作流程如下：首先采用深度学习模型对用户偏好信息中的文本和图片信息进行特征提取，即对商品数据库中的文本和图片进行特征提取；然后对用户输入的检索条件（如文本、音频和图片）信息进行特征提取；计算两种商品特征的相似度，选取相似度超过一定阈值的商品，组成用户偏好商品集合；之后求得商品数据库内的商

品信息和用户检索查询之间的商品特征向量相似度,选取相似度超过一定阈值的商品组成用户检索查询的商品集合。如果上述两个集合有交集,在交集中根据商品特征相似度,选取前 $k$ 个商品作为多模态商品检索的结果;否则,就将用户检索查询的商品集合中根据商品特征相似度选取前 $k$ 个商品作为多模态商品检索的结果。

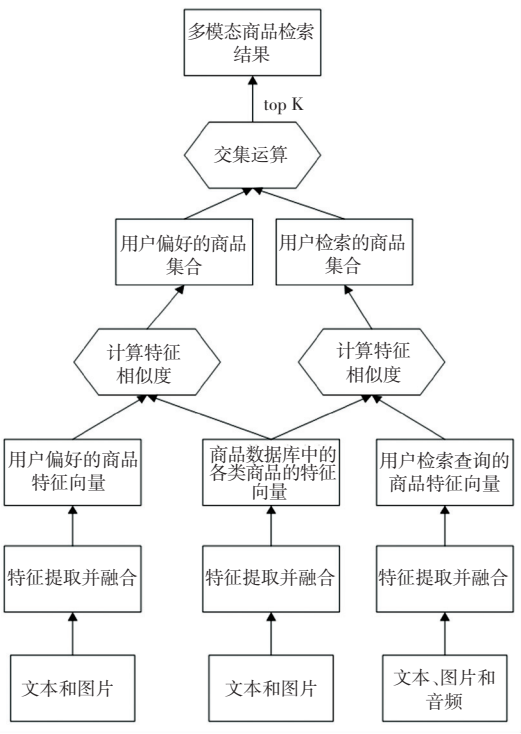


**图 1 多模态的商品检索引擎结构图**

**Fig. 1 Structure diagram of multimodal commodity retrieval engine**

多模态特征融合信息主要分为两类:一类是对于用户检索的条件包含了音频、文本和图像的特征融合;还有一类是对于商品数据库信息和用户偏好信息的融合(如文本和图像特征融合)。

关于音频、文本和图像的特征提取以及融合如图 2 所示。首先采用 LCMV 算法对音频进行增强,然后将音频分成 Refiner 段,再把 Refiner 段带入 Transformer Transducer 模型得到文本转换信息;然后把这些文本信息和用户检索查询的文本进行串联拼接,再对拼接后的文本进行Skip-Gram嵌入分词得到 Tokenization,将其带入 Transformer 模型得到文本的特征向量;之后图像的处理也是如此,先将图像进行分割成 Patch Projection,再将其带入 Twins-PCPVT 模型进行特征提取,得到图像的特征向量;最后再将文本的特征向量和图像的特征向量采用多头注意力机制和全连接层的处理,得到了处理后的商品信息的文本和图像的特征向量,再将这两个特征向量带入 MCB 模型进行融合。

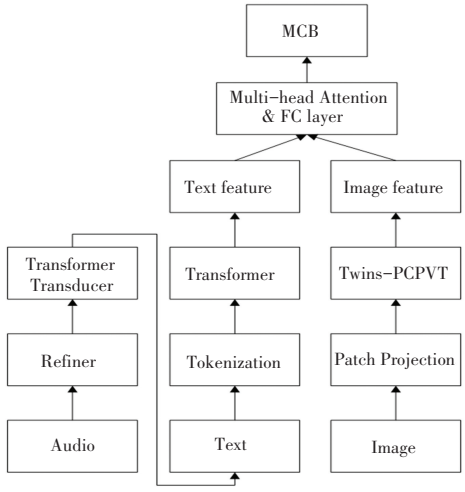


**图 2 多模态特征融合的结构图**

**Fig. 2 Structure diagram of multi-modal feature fusion**

## 4 实验及分析

本文选用 KDD Cup 2020 挑战赛中的多模态商品数据集[18],该数据集中包含用户文本检索数据和图像检索数据,从中选取 10 万条数据作为实验数据集,在此基础上添加了用户的偏好信息,并为数据集中 50%的样本添加了商品语音检索信息,构造出实验所需的多模态商品数据集(MCDB)。实验环境为 Ubuntu13.04 操作系统,Intel i9 109000x 处理器,内存 32G,GPU RTX3090 32G,硬板 1T。实验中使用 Python3.6 和 Pytorch1.10 深度学习框架编写程序来实现模型并对上述模型进行训练。

为了验证本文模型的有效性,选取具有代表性的文本检索模型 LSTM-DSSM[19] 和图像检索模型 DELF[20] 与本文提出的多模态商品检索模型 MCFR-Net 模型进行运行时间效率对比实验,实验结果见表 1。

**表 1 单模态与多模态商品检索模型的时间效率**

**Tab. 1 Time efficiency of unimodal and multimodal commodity retrieval models** s

| 时间 | LSTM-DSSM | DELF | MCFR-Net-1 | MCFR-Net-2 | MCFR-Net-3 |
|---|---|---|---|---|---|
| 训练时间 | $6.49\times10^5$ | $7.28\times10^5$ | $1.21\times10^6$ | $1.32\times10^6$ | $1.42\times10^6$ |
| 平均测试时间 | 0.83 | 0.86 | 0.90 | 0.92 | 0.93 |

表 1 中，MCFR-Net-1 表示利用图像和文本进行商品检索的模型，MCFR-Net-2 表示利用声音和文本进行商品检索的模型，MCFR-Net-3 表示利用图像、文本和声音进行商品检索的模型。在模型训练阶段，随机选取 MCDB 数据集上的 80%样本进行训练，其余的作为测试样本。通过表 1 可以看出，本文提出的 MCFR-Net 模型相比 LSTM-DSSM 和 DELF 模型需要更多的训练时间才能使模型收敛，但对于测试样本的平均测试时间不存在明显差异。

根据检索召回率对比 LSTM-DSSM、DELF 和 MCFR-Net 的性能，实验结果见表 2。

**表 2　单模态与多模态商品检索模型的召回率**

**Tab. 2　Recall rate of unimodal and multimodal commodity retrieval models**　%

| 检索排序结果数量 | LSTM-DSSM | DELF | MCFR-Net-1 | MCFR-Net-2 | MCFR-Net-3 |
|---|---|---|---|---|---|
| 1 | 31.3 | 45.2 | 52.6 | 53.9 | 55.3 |
| 5 | 51.4 | 68.6 | 76.5 | 77.2 | 78.1 |
| 10 | 63.7 | 80.4 | 85.2 | 85.6 | 86.4 |

通过表 2 可以看出，随着检索排序结果数量的增加，各模型的检索召回率都有明显提升。DELF 模型的性能要高于 LSTM-DSSM 模型，而本文提出的 MCFR-Net 系列模型的召回率明显高于 DELF 和 LSTM-DSSM，并且 MCFR-Net-3 模型的性能最好。

## 5　结束语

综上所述，本文设计了一种全新的多模态商品检索引擎，采用深度学习和特征融合技术实现了多模态数据同时应用在一次搜索行为中。实验证明，面对多种多样的信息来源（如语音，图像和文本）时，可以使用多模态检索引擎模型来提升搜索的准确性，解决了单模态检索模型特征表示能力有限和准确性较低的问题。

## 参考文献


[1] 丁婷. 阿里集团双品牌战略拓展海外电商市场的合力与挑战[J]. 对外经贸实务，2020(6)：4.

[2] 葛静茹，焦世奇，翁朝霞. “区块链+跨境电商”的发展机遇及挑战[J]. 现代经济信息，2019(24)：2.

[3] NGIAM J，KHOSLA A，KIM M，et al. Multimodal Deep Learning[C]// International Conference on Machine Learning. DBLP，2009：102-108

[4] B D K J A，B ZZ A，B K H A. Multi angle optimal pattern-based deep learning for automatic facial expression recognition [J]. Pattern Recognition Letters，2020，139：157-165

[5] Ivo M，Baltruschat，et al. Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification [J]. Scientific Reports，2019：58-66

[6] LI Y . Research on Chinese Entity Relation Extraction Method Based on Deep Learning[C]// 2021 International Conference on Communications，Information System and Computer Engineering (CISCE). 2021：210

[7] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，et al. An Image is Worth 16x16 Words：Transformers for ImageRecognition at Scale[J]. 2020：121-127

[8] LIU Z，LIN Y，CAO Y，et al. Swin Transformer：Hierarchical Vision Transformer using Shifted Windows[J]. 2021：106

[9] CHU X，TIAN Z，WANG Y，et al. Twins：Revisiting the Design of Spatial Attention in Vision Transformers[J]. 2021：86

[10] LV Z，KANG J，ZHANG W Q，et al. An LSTM-CTC based verification system for proxy-word based OOV keyword search [C]// IEEE International Conference on Acoustics. IEEE，2017：5655-5659

[11] CHIU C C，NARAYANAN A，HAN W，et al. RNN-T Models Fail to Generalize to Out-of-Domain Audio：Causes and Solutions [C]// 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE，2021：135-139

[12] DALMIA S，LIU Y，RONANKI S，et al. Transformer-Transducers for Code-Switched Speech Recognition[C]// 2021 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). IEEE，2021：311-316

[13] Yoav Goldberg，车万翔，郭江，等. 基于深度学习的自然语言处理[J]. 中文信息学报，2021，35(8)：1.

[14] 郝星跃. 基于多头注意力机制的 BiGRU-CNN 文本情感分析[J]. 计算机科学与应用，2022，12(1)：12.

[15] LI Z，LI Z，ZHANG J，et al. Bridging Text and Video：A Universal Multimodal Transformer for Video-Audio Scene-Aware Dialog[J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2021 (99)：1-3.

[16] 王保加，潘海为，谢晓芹，等. 基于多模态特征的医学图像聚类方法[J]. 计算机科学与探索，2018，12(3)：12.

[17] 陈锻生，吴琼，吴扬扬，等. 一种基于紧凑双线性融合的图文跨模态情感分类方法：，CN107066583A[P]. 2017.

[18] 李光宇. 基于深度神经网络的多模态信息检索[J]. 计算机应用与软件，2022，39(1)：7.

[19] PALANGI H. Deep learning for sequence modelling：applications in natural languages and distributed compressive sensing. 2017：73-75

[20] NOH H，ARAUJO A，SIM J，et al. Large-Scale Image Retrieval with Attentive Deep Local Features[J]. 2016：40-42