文章编号：2096-1472(2022)-10-01-06
DOI:10.19644/j.cnki.issn2096-1472.2022.010.001

跨模态数据融合综述

祁铧颖，贺 萍

(河北经贸大学信息技术学院，河北 石家庄 050061)
✉2227568724@qq.com; pinghe@heuet.edu.cn

摘 要：大数据时代下，人们热衷于在社交媒体上以图文等结合的方式发布状态数据。为了更好地整合海量社交媒体数据，首先，介绍了单模态、多模态、跨模态数据融合的最新研究进展，并在此基础上比较其优势与存在的问题；其次，在跨模态应用方面，对跨模态图文检索、跨模态推荐系统、跨模态情感分析、跨模态人机对话系统及其目前存在的问题进行了分析与总结；最后，得出跨模态数据融合能够实现数据合理有效利用的结论，并提出对跨模态数据融合未来发展的展望。

关键词：单模态融合；多模态融合；跨模态融合；深度学习
中图分类号：TP301 **文献标识码**：A

www.rjgcz.com

Overview of Cross-Modal Data Fusion

QI Huaying, HE Ping

(*School of Information Technology*, *Hebei University of Economics and Business*, *Shijiazhuang* 050061,*China*)
✉2227568724@qq.com; pinghe@heuet.edu.cn

Abstract: In the big data era, people are keen to release status data by means of pictures and texts on social media. In order to better integrate massive social media data, this paper first introduces the latest research progress of single-mode, multi-mode and cross-mode data fusion, and compares their advantages and existing problems on this basis. Secondly, in terms of cross-modal application, the cross-modal image and text retrieval, recommendation system, sentiment analysis, human-machine dialogue system, and the existing problems are analyzed and summarized. Finally, it is concluded that cross-modal data fusion can realize reasonable and effective data utilization, and the future development prospect of cross-modal data fusion is proposed.

Keywords: single-mode fusion; multi-modal fusion; cross-modal fusion; deep learning

1 引言(Introduction)

随着智能设备的不断更新迭代，越来越多的人选择在微博、抖音等社交平台上以文本、图像、短视频等形式发表个人观点。面对海量的社交媒体数据，若能够实现其合理利用，则将在网络监管等方面产生极大的价值[1]。不同于特征简单的数据拼接[2]，数据融合不但包括数据显示融合，而且包括数据深层次的语义交互及不同数据之间关系的建模。目前，数据融合的对象包括单模态、多模态和跨模态三种类型数据。

通过对三种不同类型数据的梳理，发现单模态数据融合主要集中于文本、图像数据[3]；多模态数据融合主要集中在文本、图像、音频和视频等数据，包括模态的两两结合或者三种模态相结合[4]；跨模态数据融合同样是对不同模态的数据进行融合，但它通过跨模态关系抽取来获取不同模态间的交互信息。目前，跨模态数据融合是高效利用海量媒体数据的有效手段之一[5]。

基金项目：河北省教育厅科学研究计划项目(重点项目，ZD2019017).

2 单模态数据融合方法(Single-mode data fusion method)

模态是指数据的存在形式，单模态是模态中的一种形式，是对文本或图像单独进行处理的一种方法。以基于文本的单模态融合为例，对文本单模态数据进行处理步骤如下：(1)源数据预处理；(2)向量表示；(3)特征提取；(4)模型分类/回归；(5)输出结果(图1)。

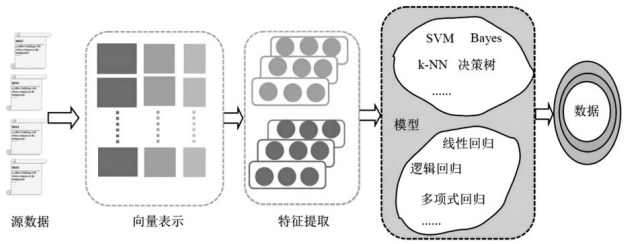

图1 文本单模态数据融合

Fig.1 Single-mode data fusion of the text

首先对源数据进行预处理，对文本进行分词、删除标签、归一化等清洗工作，并将数据序列化；然后在文本分析阶段，将获取的向量表示的文本嵌入同一维度的矩阵中，之后提取文本特征；最后经过模型的分类回归处理得出结果。

文本特征的提取结果对文本挖掘起到至关重要的作用。PANG等人[6]首次使用基于机器学习的方法解决文本情感分类问题，该方法使用电影评论作为数据，效果优于传统人工标注。张庆庆等人[7]针对特征向量维度高的问题，提出BPSO与随机子空间方法结合的选择性集成算法，使文本情感分类的准确率得到提升，模型具有较优的泛化能力。但是，基于机器学习的方法不能充分获取文本信息，需要经过大量的训练语料和复杂的样本特征抽取的过程才能得到较准确的分类结果。随着深度学习的发展，基于深度学习方法体系可以自动学习语义特征，避免出现庞大的特征提取过程，常用的方法如使用循环神经网络(Recurrent Neural Networks，RNN)提取文本特征。卢强等人[8]将非情感词、情感词、否定词和程度词等中文语法规则以正则化的形式同双向长短期记忆网络(Bidirectional Long Short Term Memory，Bi-LSTM)模型相结合，应用于中文文本分类。在此基础上，卢强等人[8]在模型中引入了注意力机制和双向门控循环单元(Bidirectional Gated Recurrent Unit，Bi-GRU)，能够更好地提取文本关键信息，捕捉文本上下文语义，应用于情感分析中。YUE等人[9]提出了一个Word2vec-BiLSTM-CNN混合模型，该模型得益于卷积神经网络(Convolutional Neural Networks，CNN)提取特征的能力和Bi-LSTM学习文本短期双向依赖性的特性，证明了混合网络模型在短文本中的性能优于单一结构神经网络，但不足之处是它需要比现有基线模型更多的训练数据和训练时间。

对图像单模态数据进行处理步骤如下：(1)源数据预处理；(2)特征提取；(3)模型分类/回归；(4)输出结果(图2)。

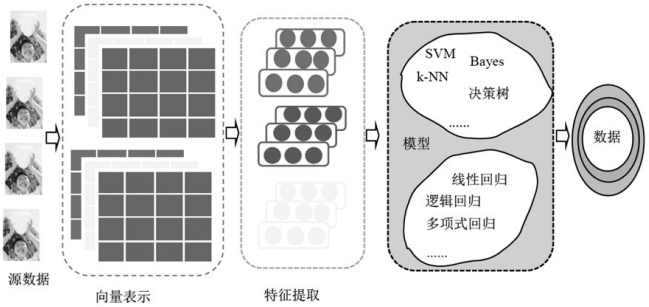

图2 图像单模态数据融合

Fig.2 Single-mode data fusion of the image

首先对源数据进行预处理，对图像进行去均值、标注、归一化等清洗工作，剔除不相关的区域；然后在图像处理阶段，将获取的向量表示的图像嵌入同一维度的矩阵中，之后提取图像特征作为下一层的输入；最后经过模型的分类回归处理得出结果。

人眼在图像处理方面具有很大优势，传统的基于机器学习的视觉特征提取包括底层特征提取和中高层特征提取。COLOMBO等人[10]在图像和视频数据中根据颜色、温暖度、位置和大小等特征分割不同的区域，获取比物体本身更高的语义层次，应用于艺术类图片的情感分析中。仅从图像底层特征中提取的信息是有偏差的，黄崑等人[11]提取了图像物理感知层、情感接受层与喜好层三层不同特征进行情感分类。随着深度学习的发展，可以自动提取图像特征，常用的方法如使用CNN提取图像特征。在此基础上，李志义等人[12]使用改进的卷积网络模型，将图像的底层特征融入图像的高层语义中，提高了情感分析的准确率和模型的泛化能力，但存在信息冗余的不足。蔡国永等人[13]使用注意力机制以关注图像中的显著区域，融合高层特征和底层特征用于情感分类。

单模态数据融合的优势与问题：单一模态的数据融合具有计算复杂度低、表达更简练的优点，例如在线评论中短文本的情感分类有利于完成用户的推送服务。但是现代社会的信息具有丰富的特征，仅靠单一模态往往容易造成收集信息不全面从而导致判断的不准确性，所以多模态数据融合方法应运而生，以更好地满足人们以图文、音频、视频等多模态形式表达自己的需求。

3 多模态数据融合方法(Multi-modal data fusion method)

多模态是指两个或两个以上模态的各种组合形式。多模态数据融合步骤如下：(1)单模态特征提取；(2)特征融合；(3)模型分类/回归；(4)结果输出(图3)。

www.rjgcz.com

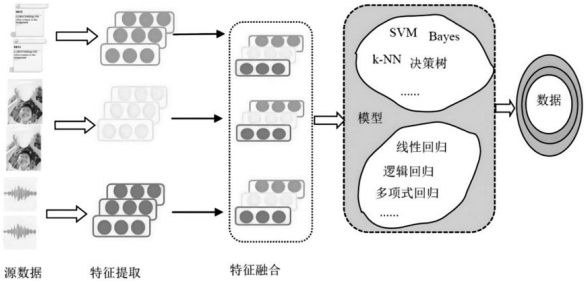

图3 多模态数据融合

Fig.3 Multi-modal data fusion

文本、图像、音频是多模态研究的三种主要异构数据类型，首先提取不同数据源的特征；然后将提取到的不同特征进行融合，在融合阶段可以采用特征级融合、决策级融合、混合融合三种不同的融合方式；最后进行模型分类、回归等操作得出结果。

3.1 基于显式融合方式的多模态融合

3.1.1 特征级融合

特征级融合是对从原始数据中提取的特征信息进行融合的一种方式，适用于模态之间高度相关的情况。YANG等人[14]采用特征级融合的方法对音频和视频特征进行处理，用于语音识别，但存在特征提取难度大的不足，并不是最理想的融合方法。随着深度学习的发展，林淑瑞等人[15]分别采用特征级融合和决策级融合对音频和视频进行融合，应用于情感分析研究，证明了多模态融合优于单模态融合，且决策级融合结果略高于特征级融合结果，但当特征数量增大时，很难获得不同特征之间的交叉相关性。模糊推理法、产生式规则法等都是特征级融合的常见方法。

3.1.2 决策级融合

决策级融合是对数据经过进一步评估或推理得到的局部决策信息进行融合的一种方式。决策级融合可以融合异质数据，但存在容易造成信息损失的缺点。谢豪等人[16]充分挖掘图文之间的相似性信息，运用多层语义决策级融合模型，对社交媒体中的图文进行情感分类。周红标[17]在单一模态情感识别的基础上，将基于隐马尔可夫的语音情感识别结果和基于最小二乘支持向量机的脉搏情感识别结果进行决策级融合，使情感分类准确率大大提高。贝叶斯概率推理法、D-S证据推理法等都是决策级融合的常见方法。

3.1.3 混合融合

混合融合是上述特征级融合和决策级融合的结合，它保留了上述两种融合策略的优势，但同时也使模型更加复杂，学习难度加大。由于深度学习模型具有灵活性和多样性的结构特点，比较适合使用混合融合方法，但也需要根据具体问题选择最合适的融合方法。例如，周新民等人[18]提出了一种深度多模态融合模型，该模型兼顾特征级融合、决策级融合及混合融合多层级数据融合方式，实现了多模态数据的有效融合。混合融合一般应用于网络安全、遥感影像等领域。

3.2 基于隐式融合方式的多模态融合

3.2.1 双线性融合

与简单的向量整合操作(如元素相加或拼接)只考虑多模态特征之间的简单相关性不同，双线性池化方法由于两模态之间的二阶相互作用而成为信息融合研究的关键。但是，直接使用双线性池化方法会导致学习参数和计算资源的急剧增加，针对这一问题，人们提出了多种解决方法，如多模态紧合型双线性池化(MCB)、多模态低秩双线性池化(MLB)、多模态因子化双线性池化(MFB)和多模态Tucker Fusion (MUTAN)。

3.2.2 条件随机场

条件随机场在自然语言处理方面具有较好的效果，广泛应用于标注场景。为提高多模态数据分类效果，JIANG等人[19]提出了一种基于监督学习的多模态数据分类方法——多模态隐条件随机场(Multi-modal Hidden Conditional Random Field，M-HCRF)，捕捉输入数据的潜在信息，从而促进多模态数据的分类。

3.2.3 基于张量的融合

多模态情感分析是一个越来越受欢迎的研究领域，它将传统的基于语言的情感分析定义扩展到一个多模态设置，其中其他相关的模式伴随着语言。ZADEH等人[20]针对在线视频中口语的不稳定特性及伴随的手势和声音问题，引入一种新的张量融合网络模型，融合语言、视觉和听觉，应用于情感分析中。实验证明，张量融合能较好地保留多模态数据中各模态之间的交互信息，但存在信息冗余的不足。

3.2.4 基于注意力的融合

近年来，在自然语言处理过程中，注意力机制得到了广泛应用[21-22]。得益于深度学习的发展，郭可心等人[21]以文本内容为驱动，充分利用图像的中低层特征，并在多层图像特征中引入注意力机制，从多个角度挖掘图文之间的情感共现，但其只在单一模态中使用注意力机制，忽略了不同模态间的交互作用。范涛等人[22]提出一种基于多模态联合注意力机制模型，即通过词引导的注意力机制引导图像注意，使用图像引导的注意力机制引导词注意，将文本和图像进行融合，应用于网民情感分析，相对减少了冗余与噪声。

多模态数据融合的优势与问题：与传统的单一模态相比，多模态数据融合负责将多个模态的信息进行有效整合，提供了更加丰富的信息，且信息之间具有互补的优势。但因为每种模态数据看待事物的角度不同，所以存在一些交叉，要想保存各模态信息与各模态间的交互信息，就必然造成数据的冗余，由此跨模态应运而生。

www.rjcz.com

4 跨模态数据融合方法(Cross-modal data fusion method)

随着人工智能的发展，跨模态数据融合逐渐走进人们的视野。跨模态数据融合步骤如下：(1)跨模态关系抽取；(2)各模态特征提取；(3)跨模态特征融合；(4)模型分类/回归；(5)输出结果(图4)。

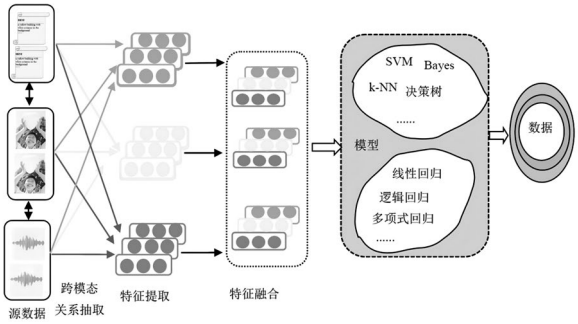

图4 跨模态数据融合

Fig.4 Cross-modal data fusion

跨模态数据融合主要是文本、图像和音频的融合，首先从源数据中进行跨模态关系抽取，获取各个模态的交互信息；其次结合模态本身得到相互独立的文本特征、音频特征、图像特征，得到的特征更加全面、准确，这是对多模态的精简和补充，结果也更准确；最后将得到的各模态特征进行融合，应用于情感分析等众多领域。

4.1 基于深度学习的跨模态融合

4.1.1 基于卷积神经网络的跨模态融合

CNN因其无须手动提取特征，权值共享且可以直接输入网络的优势，被广泛应用于语音识别和图像处理中[23]，而跨模态融合的核心任务在于跨模态关系抽取和联合特征学习。2016年，AlphaGO战胜李世石，深度学习得到迅猛发展，蔡国永等人[24]最早使用CNN对图文进行情感预测，并将结果通过决策级融合后进行分类，但图像和文本的表达具有差异。在此基础上，申自强[25]针对图文之间情感互斥问题提出一种基于图文融合的跨模态社交媒体方法，使用CBOW(Continuous Bag of Words)和CNN分别提取文本与图像特征，但该方法只是提取了图像中的高级语义特征，而中、低级语义特征同样影响着情感分类的准确性。针对这个问题，陈巧红等人[26]在图像特征提取中利用VGG13网络外接卷积层，分别获取高、低、中层次的图像特征，对文本图像进行跨模态融合，应用于情感分析。

4.1.2 基于循环神经网络的跨模态融合

21世纪初，RNN发展为深度学习算法之一[27]，LSTM作为RNN的变形，克服了RNN长期依赖的缺点。安迅[28]利用LSTM提取文本特征，使其拥有更强的记忆功能，对文本特征和图像特征进行跨模态的特征对齐，使用Softmax进行情感分类。虽然LSTM能够保留较长时间的信息，但是文本的语义同时与上下文有着密切联系。针对这类问题，陈巧红等人[26]利用BERT词嵌入加Bi-GRU提取文本特征，GRU通过添加更新门与重置门，能够对输入的时序信息选择性地保留与遗忘，从而捕捉到重要的长期信息。同时Bi-GRU更好地结合了文本的上下文语义，将文本特征和图像特征进行注意力融合，应用于情感分析中，但存在计算复杂度大的问题。

4.1.3 基于残差网络的跨模态融合

深度学习虽然可以提取复杂的数据特征，但是会出现梯度消失的问题，使用残差网络可以使神经网络的层数超越之前的约束，达到几十层、上百层甚至上千层，为高级语义特征提取和分类提供了可行性。王茂等人[29]使用LSTM提取问题词特征，使用ResNet-152提取图像特征，并利用改进的残差通道自注意力增强图像特征，将图文特征进行跨模态动态融合，应用于视觉问答领域，提高了视觉问答的准确性。在此基础上，吴琼[30]在图像特征提取时使用预训练的152层残差网络，在文本特征提取时使用LSTM，将图文两种特征进行拼接，再与图像的局部和全局特征加权求和，得到最终的视觉表示；接着通过紧凑双线性融合，对融合后特征进行跨模态的情感分类，从文本、图像及二者融合三方面研究情感变化，融合结果更加准确，但增加了计算的复杂度。

4.2 基于注意力的跨模态融合

在深度学习中引入注意力机制是使机器更贴合人类行为的重要方法，最早是在图像情感分析中应用注意力机制[31]。针对传统的单模态情感分析存在的不足且在特征提取时容易忽略重要特征的情感信息的问题，陈小敏等人[32]提出一种基于注意力机制的TAI-CNN(Text And Image-Convolutional Neural Networks)图文跨模态情感分类方法，解决了上述问题，但该模型仍存在改进空间。陈巧红等人[26]提出一种基于多层跨模态注意力融合的图文情感分析模型，如图5所示。

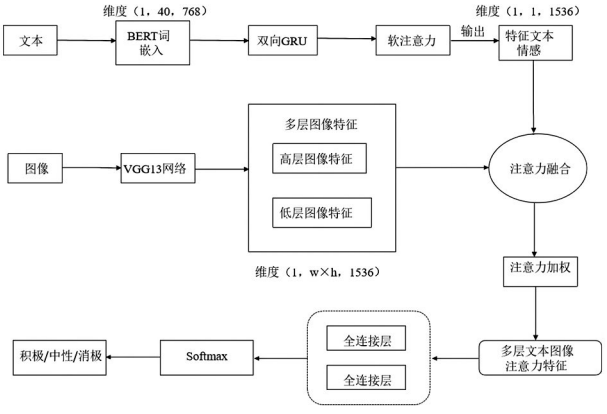

图5 多层跨模态注意力融合模型

Fig.5 Multi-layer cross-modal attention fusion model

www.rigzz.com

该模型使用Bi-GRU完成对文本情感的表示，并使用软注意力机制给予文本中表示情感的关键词更多的权重；使用VGG13网络外接卷积层，获取不同层次的图像特征，将文本特征与多层图像特征进行注意力融合，并将图文融合特征输入多层感知机及Softmax分类器进行情感分类。但基于情感分析的跨模态融合不应只局限于文本和图像之间，王雨竹等人[33]针对文本、视觉、音频三种模态，结合跨模态上下文信息并使用注意力机制过滤冗余信息，将融合信息应用在情感分析领域。

跨模态数据融合的优势与问题：跨模态数据融合与多模态数据融合相比，两者的相同之处是数据都来自所有模态，两者的差别是跨模态融合的数据只用于某一个模态，而多模态融合的数据用于所有模态。融合后的跨模态数据相比于单一模态数据具有去冗余、全面、多元及互补的优点，从而实现数据的合理有效利用。今后跨模态数据融合会不断得到改进，获得全面发展。

5 跨模态的应用(Cross-modal applications)

5.1 在跨模态图文检索上的应用

跨模态图文检索能更好地丰富用户对同一事物的认知，如以文搜图或以图搜文，其核心任务是特征提取和语义相似性度量。为提高跨模态图文检索的精度和速率，朱路等人[34]提出了一种基于语义自编码哈希学习的跨模态检索方法，不仅减小了不同模态数据之间的异构鸿沟，而且降低了存储空间。当在检索中出现新数据时，会使计算量大大增加，江朝杰等人[35]针对检索数据的不断迭代，提出增量跨模态检索方法，实验证明其在新增类样本集上具有良好的检索精度。

随着网络数据表现形式越来越多样化，跨模态图文检索逐渐成为研究热点。跨模态图文检索具有检索精度高和应用范围广等优势，目前主要应用于公安领域舆情分析、个性化推荐、医学数据存储等方面。减小跨模态之间的语义鸿沟仍是未来的一个研究方向。

5.2 在推荐系统上的应用

推荐系统最早的一个独立研究领域是新闻推荐[36]，目前还有很大的发展空间。滑瑾[5]从跨模态语义分析的角度展开研究，通过引入外部信息源，解决了现实网络中的数据稀疏问题，完成高质量推荐。推荐系统性能优越的关键就是了解用户的潜在兴趣，贾冬柏[37]从跨模态角度对用户兴趣进行分析，结合艾宾浩斯遗忘曲线，提出构建基于兴趣衰减和多模态特征融合的兴趣挖掘模型来获取用户兴趣所在，实现微博用户个性化推荐。

跨模态数据在推荐系统中的应用使推荐精准度得到提升，目前，基于跨模态数据融合的推荐系统得到了信息检索、计算机视觉等多个学科研究者的关注。相信随着跨模态数据融合技术的不断发展，推荐系统的精度会越来越高，能够更好地服务于大众。

5.3 在情感分析上的应用

情感分析作为近几年的热点话题，得到了广大研究者的青睐。目前，基于注意力机制的图文跨模态情感分析成为研究主流。陈巧红等人[26]将文本与图像进行注意力融合，应用于情感分析。ZHOU等人[38]通过使图文间的语义和情感交互被分层提取，显著呈现不同模态之间的语义关联，用于图文情感分析，解决了噪声问题和联合理解上的差异问题，提高了情感分析结果的准确性。

与单一模态情感分析相比，跨模态能够获得更加全面的数据，提高情感分析的准确度。目前，基于跨模态数据融合的情感分析主要应用于政府部门监控舆情、微博情感分析及微博用户情绪异常预警中[39]。但跨模态数据融合不应只包括图像、文本，基于图文、音频、视频等模态的融合进行情感分析是未来的一个研究方向。

5.4 在人机对话系统上的应用

跨模态学习作为一种能让机器更加贴近人类行为的技术，已经成为人工智能相关领域研究者关注的热点问题。可视问答作为跨模态的一个主要示例，也受到越来越多的关注。LAO等人[40]提出了共同注意学习方法和跨模态多步融合网络，该模型通过句子引导词注意力和问题引导图注意力提取不同模态的特征，通过跨模态多步骤融合网络来生成多步骤特征并实现两种模态的多个交互，它区别于大多数现有的特征融合方法仅专注于两种模态之间的复杂交互建模，通过使用池化方法将最终的融合特征应用于答案预测，实现较为精准的视觉问答。

人机对话系统不断朝着智能化的方向发展[41]，基于文本的细粒度分析和对图像的关键区域提取信息，不仅解决了对话内容的单调性，还能充分结合用户的个性化特征，确保能够准确识别用户意图。针对个性化人机对话系统，如何有效融合文本、视觉和听觉等模态构造人机对话系统将成为未来的研究方向。

6 结论(Conclusion)

跨模态数据融合近年来才得到关注，随着更多研究者的关注，其关键技术也将得到提升，可以从以下三方面着手进行研究。

(1)针对社交媒体中网络新语层出不穷，可考虑在跨模态融合中建立词典，降低文本模态的噪音，提升跨模态融合

www.rjgcz.com

性能。

(2)针对视觉数据读取与用户对相同信息的理解存在差异的问题，利用注意力模型对齐不同模态特征之间的细粒度将是未来跨模态融合研究的热点问题。

(3)目前，跨模态融合主要集中在文本和图像数据上，而视频数据可以更好地表达人们的不同状态。因此，整合文本、图像、视频等不同模式是未来研究需要考虑的问题。

参考文献(References)

[1] 陈龙,管子玉,何金红,等.情感分类研究进展[J].计算机研究与发展,2017,54(06):1150-1170.

[2] 裴红星,刘金达,葛佳隆,等.图像拼接技术综述[J].郑州大学学报(理学版), 2019,51(04):1-10,29.

[3] CHEN J, YAN S, WONG K C. Verbal aggression detection on Twitter comments: Convolutional neural network for short-text sentiment analysis[J]. Neural Computing and Applications, 2020, 32(15):10809-10818.

[4] 任泽裕,王振超,柯尊旺,等.多模态数据融合综述[J].计算机工程与应用,2021,57(18):49-64.

[5] 滑瑾.基于跨模态语义挖掘与生成对抗网络的推荐系统研究[D].南昌:华东交通大学,2020.

[6] PANG B, LEE L, VAITHYANATHANS. Thumbs up?Sentiment classification using machine learning techniques[C]// PHILLIPS W, RILOFF E. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. Philadelphia, USA: ACM, 2002:79-86.

[7] 张庆庆,刘西林.基于BPSO随机子空间的文本情感分类研究[J].数据分析与知识发现,2017(5):71-81.

[8] 卢强,朱振方,徐富永,等.融合语法规则的Bi-LSTM中文情感分类方法研究[J].现代图书情报技术,2019,3(11):99-107.

[9] YUE W, LI L. Sentiment analysis using Word2vec-CNN-BiLSTM classification[C]// GUETL C, SALEH I, CARAVOLO P, et al. 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS). Pairs, France: IEEE, 2020:1-5.

[10] COLOMBO C, BIMBO A D. Semantics in visual information retrieval[J]. IEEE MultiMedia, 1999, 6(3):38-53.

[11] 黄崑,赖茂生.图像情感特征的分类与提取[J].计算机应用,2008,28(3):659-661,668.

[12] 李志义,许洪凯,段斌.基于深度学习CNN模型的图像情感特征抽取研究[J].图书情报工作,2019,63(11):96-107.

[13] 蔡国永,储阳阳.基于双注意力多层特征融合的视觉情感分析[J].计算机工程,2021,47(09):227-234.

[14] YANG X C, FENG S, WANG D L,et al. Image-text multimodal emotion classification via multi-view attentional network[J]. IEEE Transactions on Multimedia, 2021, 23:4014-4026.

[15] 林淑瑞,张晓辉,郭敏,等.基于音视频的情感识别方法研究[J].信号处理,2021,37(10):1889-1898.

[16] 谢豪,毛进,李纲.基于多层语义融合的图文信息情感分类研究[J].数据分析与知识发现,2021,5(06):103-114.

[17] 周红标.融合语音和脉搏的多模态情感识别研究[J].微电子学与计算机,2015,32(06):5-9.

[18] 周新民,胡宜桂,刘文洁,等.基于多模态多层级数据融合方法的城市功能识别研究[J].计算机科学,2021,48(09):50-58.

[19] JIANG X, WU F, ZHANG Y, et al. The classification of multi-modal data with hidden conditional random field[J]. Pattern Recognition Letters, 2015, 51:63-69.

[20] ZADEH A, CHEN M, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis[C]// PALMER M, HWA R, RIEDEL S. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2017:1103-1114.

[21] 郭可心,张宇翔.基于多层次空间注意力的图文评论情感分析方法[J].计算机应用,2021,41(10):2835-2841.

[22] 范涛,吴鹏,王昊,等.基于多模态联合注意力机制的网民情感分析研究[J].情报学报,2021,40(06):656-665.

[23] GUO M, ZHOU C, LIU J. Jointly learning of visual and auditory: A new approach for RS image and audio cross-modal retrieval[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019, 12(11): 4644-4654.

[24] 蔡国永,夏彬彬.基于卷积神经网络的图文融合媒体情感预测[J].计算机应用,2016,36(2):428-431,477.

[25] 申自强.一种基于图文融合的跨模态社交媒体情感分析方法[J].软件导刊,2019,18(01):9-13,16.

[26] 陈巧红,孙佳锦,孙麒,等.基于多层跨模态注意力融合的图文情感分析[J].浙江理工大学学报(自然科学版),2021,47(01):85-94.

[27] SCHMIDHUBER J. Deep learning in neural networks: An overview[J]. Neural Networks, 2015, 61:85-117.

www.rjgczz.com

[28] 安迅.基于跨模态融合的图文情感分析方法研究[D].北京:北京工业大学,2020.

[29] 王茂,彭亚雄,陆安江.面向视觉问答的跨模态交叉融合注意网络[J/OL].计算机应用.[2021-12-02].http://kns.cnki.net/kcms/detail/51.1307.TP.20210708.1522.002.html.

[30] 吴琼.基于卷积神经网络的图文情感分析技术研究[D].泉州:华侨大学,2017.

[31] SONG K, YAO T, LING Q, et al. Boosting image sentiment analysis with visual attention[J]. Neurocomputing, 2018, 312:218-228.

[32] 陈小敏,许华虎,方迪恺,等.基于注意力机制的TAI-CNN图文跨模态情感分类方法[J].计算机应用与软件,2021,38(05):196-200,293.

[33] 王雨竹,谢珺,陈波,等.基于跨模态上下文感知注意力的多模态情感分析[J].数据分析与知识发现,2021,5(4):49-59.

[34] 朱路,邓芳,刘坤,等.基于语义自编码哈希学习的跨模态检索方法[J/OL].数据分析与知识发现,2021,5(12):110-122.

[35] 江朝杰,杨良怀,高楠,等.增量跨模态检索方法[J].小型微型计算机系统,2021,42(10):2234-2240.

[36] ADOMAVICIUS G, TUZHILIN A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6):734-749.

[37] 贾冬柏.基于跨模态特征融合的用户兴趣挖掘算法的研究与实现[D].北京:北京邮电大学,2020.

[38] ZHOU T, CAO J, ZHU X, et al. Visual-textual sentiment analysis enhanced by hierarchical cross-modality interaction[J]. IEEE Systems Journal, 2020(99):1-12.

[39] 凌海彬.基于多特征融合的微博情感分析研究[D].桂林:桂林电子科技大学,2019.

[40] LAO M, GUO Y, WANG H, et al. Cross-modal multistep fusion network with co-attention for visual question answering[J]. IEEE Access, 2018, 6:31516-31524.

[41] 杨帆,饶元,丁毅,等.面向任务型的对话系统研究进展[J].中文信息学报,2021,35(10):1-20.

作者简介：

祁铧颖(1998-)，女，硕士生.研究领域：跨模态数据融合，情感分析.

贺　萍(1982-)，女，博士，副教授.研究领域：算法优化，深度学习，广度学习.本文通讯作者.

(上接第19页)

表2 算法运行精度对比

Tab.2 Comparison of algorithm running accuracy

算法	不同样本个数的运行精度/%								
	50	100	200	500	800	1,000	2,000	2,500	3,000
SOM	87.5	85.0	82.2	81.5	80.9	79.1	80.0	77.2	75.0
改良算法	92.4	91.5	90.5	87.2	88.5	86.3	85.9	83.2	81.0

由表2可知，在限制了迭代次数及batch_size的情况下，改良算法的运行结果在推荐精度上有了5%—8%的提升，符合视频推荐应用对精度要求更高的需求。

5 结论(Conclusion)

SOM和其他算法的组合在很多领域应用效果较好。本文使用SOM和K-means的组合算法来进行视频推荐，并设计了一个视频推荐系统，系统由数据采集与处理、算法优化和可视化三部分构成。实验结果表明，在限制batch_size和迭代次数的情况下，使用K-means优化的SOM算法虽然在运行效率上有所降低，但在运行精度上有了5%—8%的提升，在视频推荐应用方面，推荐的准确性远比运行时间更符合实际用户需求。因此，改良的SOM算法适合应用在视频推荐系统上。

参考文献(References)

[1] 刘方爱,王倩倩,郝建华.基于深度神经网络的推荐系统研究综述[J].山东师范大学学报(自然科学版),2021,36(4):325-336.

[2] 杨占华,杨燕.SOM神经网络算法的研究与进展[J].计算机工程,2006,32(16):201-202.

[3] 李寒,陶涵境,崔立昊.基于SOM-K-means算法的番茄果实识别与定位方法[J].农业机械学报,2021,52(1):23-29.

[4] ZHANG L, JIA J D, GUI G. Deep learning based improved classification system for designing tomato harvesting robot[J]. IEEE Access, 2018, 6:67940-67950.

[5] 高文欣,刘升,肖子雅.闪电分叉过程算法优化的K-means聚类[J].运筹与管理,2021,33(12):93-100.

[6] 孟宪颖,毛应爽.基于Python爬虫技术的商品信息采集与分析[J].软件,2021,42(11):128-130.

[7] 王素格,杨安娜,李德玉.基于汉语情感词表的句子情感倾向分类研究[J].计算机工程与应用,2009,45(24):153-155.

[8] 王嘉菲,朱志锋.基于协同过滤算法的视频智能推荐系统[J].湖北大学学报(自然科学版),2019,41(2):202-207.

作者简介：

付丽梅(1977-)，女，硕士，副教授.研究领域：计算机图像处理，移动应用开发.

www.rigczz.com