基于深度多模态表示学习的短视频流行度预测研究

Research on Micro-Video Popularity Prediction Based on Deep Multimodal Representation Learning

专业类别：电子信息

研究方向（领域）：多媒体信息处理

作者姓名：叶徐清

指导教师：吕卫

企业导师：江海

答辩日期	2021 年 12 月 10 日		
答辩委员会	姓名	职称	工作单位
主席	井佩光	副教授	天津大学
委员	宋丹	副教授	天津大学
	李东	高级工程师	日电信息系统（中国）有限公司

天津大学自动化学院

二〇二一年十二月

摘　要

近年来，随着移动互联网的迅速发展和智能手机的普及，短视频数据正在激增，但其中只有少部分短视频因为受到大量的观看、点赞、评论和转发而流行起来，大多数很快就会被遗忘。短视频数据具有多模态特性和丰富的语义信息，而这些复杂的数据之间必然存在相关性和独立性，因此有必要对短视频数据进行有效的特征表征和语义理解。本文基于对短视频多模态特征表示和预测模型的创新，针对短视频流行度预测完成了以下研究工作：

本文设计了一种基于深度多模态融合的短视频流行度预测算法，该算法利用自注意力机制网络探寻各个模态之间的相关性，解决了短视频原始模态特征之间的维度差异、数据缺失等问题。同时利用算法网络层中的批量归一化层进行深度信道交换，具体地，信道交换的过程是通过批量归一化的尺度因子引导不同模态的子网络之间动态交换模态信息。融合各个模态以获取更好的多模态统一特征表示，从而更好的表示短视频。通过在公开数据集上的实验结果证明了模型的有效性。

本文提出了一种基于双向深度编码网络的短视频流行度预测模型，该模型同时考虑多模态融合和单模态监督的建模并将其整合为一个双向深度编码网络。多模态融合模块利用模态相关性解决原始特征之间的数据缺失和维度差异等问题以获取更全面特征表示，单模态监督模块利用模态独立性监督多模态特征融合。联合训练多模态融合和单模态监督任务，充分学习模态间相关性和独立性信息以增强算法的泛化能力，提高短视频流行度预测精度。通过对公开数据集的实验表明，证明了本文提出的模型的有效性和优越性。

关键词：模态关联性，多模态融合，特征表示，短视频，流行度预测

ABSTRACT

Recently, with the rapid development of the mobile Internet and the popularization of smart phones, micro-video data is increasing rapidly, but only a small part of micro-videos can become popular due to a large number of watching, likes, comments and reposts, and most of them will be forgotten quickly. Micro-video data has multimodal characteristics and rich semantic information, meaning that the correlation and independence properties are embedded in micro-videos. Therefore, it is necessary to learn effective feature representations to benefit semantic understanding of micro-videos. Based on the innovation of short video multi-modal feature representation and prediction, this article has completed the following research works for micro-video popularity prediction.

This thesis proposed a deep multimodal fusion based on micro-video popularity prediction method. The method is to uses the self-attention mechanism network to explore the correlations between various modalities to solve the problems of unbalanced dimensions and data missing. The batch normalization layer in the proposed method is used for deep channel exchange. Specifically, the process of channel exchange is to guide the dynamic exchange of modal information between different modal sub-networks through the batch normalization scale factor. The fusion of various modalities is to obtain a unified feature representation to better represent micro-video. The effectiveness of the model is proved by the experimental results on the public dataset.

This thesis proposed a micro-video popularity prediction model based on bidirectional deep encoding network, which considers both multimodal fusion and unimodal supervision modeling, and integrates them into a bidirectional deep encoding network. The multimodal fusion module uses modal relevance to solve the problems of dimensionality difference and missing data to obtain a better feature representation; The unimodal supervision module uses modal differences to supervise the fusion of multimodal features. Joint training of multi-modal fusion and unimodal supervision tasks, fully learning the consistency and difference of multimodal information to improve the generalization ability of the algorithm. Extensive experiments have demonstrated the effectiveness of the proposed model.

KEY WORDS: Modal relevance, Multi-modal fusion, Feature representation, Micro-video, Popularity prediction

目 录

第 1 章 绪论

1.1 研究背景

近几年来，随着智能手机和移动互联网的普及和发展，短视频已经成为用户生成内容(User-generated contents, UGCs)[1]的一种新趋势，并在各种社交平台上广泛传播。短视频的时长通常在 5 分钟以内，有些短视频平台的作品甚至只有 10 秒左右，例如，Vine[1]和 Snapchat[2]平台将视频的长度分别缩短到 6 秒和 10 秒以内。尽管短视频时长比较短，但是短视频通常描述了一个相对简单但完整的故事。与传统的长视频相比，短视频更易于制作和传播，这使得短视频具有非常强的社交属性。同时短视频因其时间短、数据流量少的特性持续吸引大量的用户。在互联网信息化时代，这些短视频平台让普通民众拥有更多展示自己的机会，人人都可以进行短视频内容创作，呈现出人人自媒体的局面。因此，短视频正在逐步替代传统的图文为代表的内容消费，成为新闻播报、社交网络平台中等图文媒体中的主导内容形式之一。

短视频相比于传统的长视频，短视频在很短的时间内就可以展示了一个完整的内容。短视频对碎片化的时间利用更合理，使用场景也更多样化，满足快节奏的现代社会需求，这使得短视频具备了更多的社交属性；同时，相比于文字和图片，短视频能够让用户同时拥有视觉和听觉感受，这使得用户体验感更好。这种快速而舒适的消遣娱乐方式越来越受到广大用户的青睐，因此短视频的用户正在爆炸式增长。根据《2020 年中国网络视听发展研究报告》[2]指出，在抖音、快手等短视频平台的加速渗透下，我国短视频用户规模已达 8.18 亿，因此，每天都有大量的短视频发布到平台上，但只有很小一部分因受到大量的观看、喜欢、评论和转载被广泛传播而流行起来，而大多数很快就会被遗忘。

随着人工智能技术的兴起，众多的短视频平台都在将大数据分析、人工智能等先进的技术引入到短视频的内容生产、广告营销、运营推广和监管审核等环节中，在给用户带来更好体验的同时还能有效的控制运营成本，为企业在未来的发展中带来更大的竞争力。当前，由于短视频用户的快速增长，短视频产业已经逐渐转向为用户提供更精准和个性化、多样化的服务，因此，针对短视频的语义理解技术正在不断细化。

[1] https://vine.co.
[2] https://www.snapchat.com.

回归任务是机器学习中的三大基本模型中的重要一环，旨在确定某些变量之间的定量关系，即建立合适的数学模型并估计参数。短视频的流行度预测问题就是一个机器学习的回归任务，通过短视频的视觉模态、音频模态、文本模态和社交属性模态特征来预测相对应的短视频流行度得分。短视频流行度预测在提高网络舆情预测能力，加深用户群体行为理解等方面有重要的应用价值。受学术界和工业界短视频发展趋势的推动，本文致力于解决短视频在社交网络上的流行度预测问题。但是短视频流行度的预测面临着巨大的挑战，主要体现在一下三个方面：

- 多模态信息间的相关性：短视频的视觉、音频、文本和社交属性模态都是对同一个短视频的描述，所以这些模态信息在语义层面上有一定的相关性。所以需要一种有效的多模态特征融合框架，来探寻不同模态之间的共享关联的信息。
- 多模态信息间的独立性：尽管短视频可以通过视觉、音频、文本和社会属性等模态特征来呈现，但是，不同的模态信息分别从不同的角度描述短视频，每个模态都蕴含其他模态不具备的特有信息。因此，非常需要在统一框架中综合考虑所有异构特征的内在语义信息。
- 模态信息的噪声：噪声源于很多的外部因素，各种类型的噪声都会影响最终的预测结果。例如文本模态信息，根据统计数字[3]超过 11%的短视频不提供文本描述，同时还有很多短视频创作者为了博人眼球，取一些吸引人但是和短视频内容不相关的标题；还有些短视频创作者拍摄的作品质量差，有很严重的抖动和分辨率低等问题，这些都是外部不可避免的噪声。

相比之下，短视频的内容本身确保了稳定的信息源，所以需要利用不同角度下的模态特征之间的相关性和独立性来学习一个多模态统一特征表示。

1.2 研究现状

随着短视频平台的蓬勃发展，在短视频的研究上已经有了一些开创性研究。由于短视频数据天然的语义特性和社交属性，目前关于短视频的研究主要可以分为两个大类：短视频语义理解和短视频推荐。

在短视频语义理解领域上，Redi 等人[4]通过分析使视频具有创造性的视听特征，研究了短视频的创造性。创造性的概念，相对于美感或趣味性等相关概念，尚未从多媒体内容的自动分析的角度进行研究。Zhang 等人[5]提出了一种新的多任务多模态算法模型，该模型能够从多模态特征中共同学习公共空间，并利用预定义的 Foursquare 层次结构来规范场地类别之间的相关性，来解决短视频的场地类别估计问题。Wang 等人[6]调查并总结了通过语义特征促进短视频相关性的方

法。Nguyen 等人[7]建了一个新的短视频数据集，并引入了视点特定和时间演化的短视频理解模型。

在短视频推荐领域上，基于内容过滤方法[8]在短视频中得到应用，根据过去观看的视频记录向用户推荐视频。Wei 等人[9]通过设计了一个多模态图卷积网络(Multi-modal Graph Convolution Network, MMGCN)来探索用户和短视频的特定模态表示，在每个模态中构造一个 user-item 二部图，以更好的为用户推荐更合适的短视频。Ma 等人[10]通过提取 user-item 交互的特征，以及描述短视频上下文和视觉内容的辅助特征来更好的推荐合适的短视频。Liu 等人[11]提出了一种新颖的用户-视频共同注意网络(User-Video Co-Attention Network, UVCAN)框架，它可以使用注意机制从用户和短视频方面学习多模态信息。此外，UVCAN 以堆叠注意力网络方式对用户和短视频进行推理。

由于短视频流行度预测在推荐、广告、和许多其他应用中的重要作用，短视频流行度预测在工业界和学术界都受到了相当大的关注。由于短视频兴起的比较晚，前期关于流行度的预测主要都是集中在文本[12]、图像[13]、和传统的长视频[14]上。对于文本的流行度主要聚焦在探索流行度与内容以及当时的社会环境之间的关系。例如，Ma 等人[15]提出了一个新模型，通过从用户生成的社交图中提取 7 个内容特征，来探索 Twitter 上新出现的标签的流行程度。对于图像而言，基于内容的特征、上下文特征和社会上下文特征是必不可少的，并且在其流行度预测过程中起着关键作用。例如，Yamaguchi 等人[16]研究了影响图像流行的因素，如视觉、文本和社会因素。同样，对于视频，与文本和图像不同，视频是整合来自不同来源的信息。大多数预测视频受欢迎程度的方法都构建了一个潜在的公共特征空间，在这个空间中不同的特征可以很好地集成。Trzcinski 等人[17]提出了一个支持向量回归模型，通过时间和视觉线索预测在线视频的流行程度。Chen 等人[2]提出一种基于直推式多模态学习的算法用于解决短视频的语义流行度预测。Jing 等人[18]提出一种基于低秩多视角嵌入的直推式的学习方法用于解决短视频流行度预测问题。方法中采用了四种不同类型的特征来综合描述短视频，包括视觉、听觉、文本和社交属性模态。通过低阶表示和多视角学习来处理不完整和有噪声的信息，然后将流行度预测作为一个回归问题来表述。Xie 等人[19]将短视频多模态特征中的不确定性因素考虑其中，并提出一种多模态变分编解码框架解决短视频的流行度预测问题。该框架将所有不确定因素视为从特征到流行程度映射的随机性。

尽管针对短视频流行度预测研究已经取得了不错的研究成果，但是这些研究中对于短视频多模态统一表示学习时只考虑多模态信息之间的相互性，将四个模态通过多模态的融合框架融合成一个统一的特征表示，而对不同模态信息间的独

立性考虑的不够充分。一方面，多模态的融合已经被证明在学习多模态统一特征表征中是有效的，另一方面，多模态的融合应该充分的考虑不同模态的独立性，这就需要设计一个能够充分考虑多模态信息之间相关性和独立性的框架，利用该框架能够更好的利用多模态信息去学习一个更完备的多模态统一特征表示。增强算法模型的鲁棒性，提升短视频流行度的预测性能。

1.3 本文的主要工作

本文针对短视频的流行度预测，基于对短视频多模态特征表示和预测模型的创新，从短视频多模态关联性出发，协同利用多模态信息之间的相关性和独立性。提出了以下方案：

- 本文设计了一种基于深度多模态融合的短视频流行度预测算法，该算法利用自注意力机制网络探寻各个模态信息间的相关性，解决短视频原始模态特征之间的维度差异、数据缺失等问题。同时利用算法网络层中的批量归一化层进行深度信道交换，具体地，信道交换的过程是通过批量归一化的尺度因子引导不同模态的子网络之间动态交换模态信息。融合各个模态以获取更好的多模态统一特征表示，从而更好的表示短视频。通过在公开数据集上的实验结果证明了模型的有效性。
- 本文提出了一种基于双向深度编码网络的短视频流行度预测模型，该模型同时考虑多模态融合和单模态监督的建模，并将其整合为一个双向深度编码网络。多模态融合模块利用模态相关性解决原始特征之间的数据缺失和维度差异等问题以获取更全面特征表示，单模态监督模块利用模态独立性监督多模态特征融合。联合训练多模态融合和单模态监督任务，充分学习模态间相关性和独立性信息以增强算法的泛化能力，通过对公开数据集的实验表明，证明了本文提出的模型的有效性和优越性。

本文通过深度多模态表示学习，充分利用了所有模态信息，使得学习到的多模态统一特征表示的表征能力大大增强，提高了短视频流行度的预测精度。

1.4 本文的组织结构

本文按照研究背景和意义、国内外的研究现状、提出问题、构建算法模型、模型优化和实验验证分析的顺序来组织全文结构。全文一共分为 5 个章节，组织结构如下：

第一章是绪论部分，主要分析和总结了短视频和短视频流行度相关研究的发

展背景和实际意义，结合目前国内外的研究现状分析短视流行度预测方向研究的困难和挑战，同时给出本文的主要贡献和文章的整体结构。

第二章主要介绍了在短视频流行度预测方向上相关技术背景，分析其算法的有效性和不足之处。

第三章中，针对短视频流行度预测任务，本文设计了一种基于深度多模态融合的短视频流行度预测算法，该算法融合了各个模态信息，获取到更好的多模态统一特征表示。

第四章中，针对短视频流行度预测的任务，本文提出了一种基于双向深度编码网络的流行度预测模型。该模型同时考虑多模态融合和单模态监督的建模并将其整合为一个双向深度编码网络。通过对公开数据集的实验表明，证明了该模型的有效性和优越性。

第五章，总结与展望。本章全面概括了本文提出的两种应用于短视频流行度预测的算法，总结本篇论文的贡献和不足之处，在此基础上，对未来工作的研究方向进行了讨论和展望。

第 2 章 相关工作

本章节主要是介绍与本文内容相关的技术方法：包括特征提取、多模态表示学习等。

2.1 特征提取

特征提取和表示是多媒体信息处理的关键，目的是为了提取一些关键信息用于后续的算法模型。由于短视频可以由视觉特征、音频特征、文本特征和社交属性特征来进行描述，因此本文将对短视频分别提取四个模态的特征。

2.1.1 视觉特征

“一幅画抵得上千言万语”，作为人类，人类视觉系统根据所看到的和已有的背景知识，从一张图片中获取作者想要表达的故事。但是计算机可以从图像中发现语义概念吗？答案是肯定的，然而，计算机想要理解一张图片需要做的第一步就是提取有效的视觉特征表示，并根据这些特征建立模型。因此，如何提取图像的底层视觉特征和提取什么样的特征在图像处理的各个任务中有着至关重要的作用。

最常见的视觉特征包括颜色、纹理和形状等[20-21]，大多数的凸显标注和检索系统都是基于这些特征构建的。

- 色彩特征：颜色是图像最重要的特征之一。颜色特征是根据特定的颜色空间或模型定义的。颜色空间包括 RGB，YUV，HSV。一旦指定了颜色空间，就可以从图像或区域中提取颜色特征，颜色特征包括颜色直方图[22]、颜色矩[23]、颜色相干向量[24]和颜色相关图[25]等。
- 纹理特征：纹理特征是一种非常有用的图像特征，它可以广泛的表征图像，因为人类的视觉系统使用纹理进行识别和解释。根据提取纹理特征的领域，可以大致分为空间纹理特征提取方法[26]和光谱纹理特征提取方法[27]，它们各有优劣。空间纹理提取方法通过计算像素统计信息，或者在原始图像中找到局部像素结构来提取纹理特征,这样提取的纹理特征是有意义的且易于理解，可以从任何形状中提取而不丢失信息，但是在噪声和失真方面比较敏感；光谱纹理提取方法将图像变换到频域，然后从频域图像特征

中计算纹理特征，这样提取的纹理特征是健壮的且计算比较少，但是没有语义层面的意义。

- 形状特征：形状是人类识别现实世界物体的重要线索，其目的是编码不同方向上的直线或者几何图形。形状特征提取技术大致可以分为两类[28]，即基于轮廓的方法和基于区域的方法。轮廓的方法从形状的边界计算形状特征，基于区域的方法从整个区域提取特征。

2.1.2 音频特征

音频特征提取是当前音频信号处理研究与开发的基石之一，其要点是提取音频中有用的特征或描述符，以便解释其中包含的信息，从音频信息中提取有用信息的问题在统计理论、声音工程领域等方面得到了很多的应用[29-30]。音频特征是可以从音频信号中提取上下文的信息，但是由于其非结构化信息使得音频特征提取非常复杂。

低级音频特征可以直接从音频信号中计算出来，通常是逐帧计算的，如过零率、光谱质心或信号能量等。这些特征相比于更高级的特征，基本不包含感知相关性，如音乐的弦和键，它们具有更强的语义意义。在文献[31]中提出的块级特征(Block-level feature, BLF)技术中，分析了四种描述符类型：频谱方面（频谱模式，频谱模式增量，delta 变化等）、谐波方面（相关模式）、节奏方面（对数波动模式）和音调方面（频谱对比模式）。另一方面，在声音工程领域，通常建议使用的是描述符[32]：音频持续时间（以秒为单位）、频率（以赫兹为单位）、响度或声音强度（以分贝为单位）、混响（以秒为单位）和方向性。Peeters 等人[33]进一步定义了 54 个音频特征。此项目提供了一系列要素的定义，将这些要素分组，并将其作为基于框架的要素进行关联，然后可以处理这些音频特征，以识别音频信号的某些特定方面。

2.1.3 文本特征

提取文本信息的文本特征提取是表示文本信息的提取，它是大量文本处理的基础[34-35]。从一些有效的特征中选择一组特征来降低文本特征空间的维度过程称为特征提取[36]。在文本特征提取的过程中，删除不相关或者多余的特征以获取更好的特征表示。文本特征提取方法包括过滤、融合、映射和聚类等。

深度学习具有识别非结构化数据的优势，大多数人都熟悉图像、声音、视频和文本等媒体，所有这些都属于此类数据。深度学习在自然语言理解[37]的各种任务中产生了非常好的结果，特别是主题分类、情感分析、问答和语言翻译[38-39]。

它的深层架构特性使深度学习能够解决更复杂的人工智能任务。

2.1.4 社交属性特征

由于短视频的兴起是以传播及社交为导向的，社交属性特征则通常作为不可或缺的因素。社交属性特征一般是指在线社交网络中群体在属性上所表现出来的特征，通常表现为内部个体之间具有相似的属性。例如，短视频的创作者发布视频后吸引的观看者，这些观看者有一部分将会追随该短视频的创作者。创作者同时也会观看其他创作者的短视频，这表示这部分群体有相似的兴趣爱好。

2.2 多模态表示学习

为了让人工智能能够更好的理解周围的世界，它需要能够解释和推理多模态信息。多模态表示学习旨在解决不同模态信息之间的异质性，充分利用多模态之间相关性和独立性信息，以学到更加全面的多态统一特征表示。

由于多模态数据是从不同角度上描述同一个对象，所以在语义内容上是互补的或是补充的，因此多模态数据比单模态数据信息更丰富。但是，由于在多模态数据中，不同的模态特征向量位于不相等的子空间中，这将导致所提取的特征向量表示会不同。这就是不同模态间的异质性差距，异质性会阻碍这些提取的多模态数据的后续使用[40]。所以，需要对多模态数据进行整合，获取一个多模态统一的特征表示。

2.2.1 多模态特征融合

现实世界包括多种模态——看到的物体、听到的声音、感觉到的质地和闻到的气味等等，这些都是某件事正在发生的模态。多模态信息主要拥有以下四个特性：

- 不同的模态存在格式的不同，捕捉的频率也不同。例如，一个短视频捕获的帧速率可能与获取的音频样本的速率不同，甚至两个视频源的帧速率也不一样。因此，融合过程需要解决这种由于异步问题导致的不一致信息以获取更好的特征表示。
- 不同类型的模态信息处理的时间的不同也需要采用相对应的融合策略。
- 这些模态之间是关联的，既有相关性，又有独立性。相关性可以在不同级别上感知。例如，从不同模态提取的低级特征表示之间的相关性，以及基于不同模态获得的语义级之间的相关性。但是，模态之间的独立性也很重要，因为它可以为决策网络提供额外的线索。所以，当融合多种模态特征

时，这种相关性和独立性可能同样提供基于特定场景或上下文的有价值的信息。

- 不同的模态通常在完成不同任务时具有不同的重要程度。例如，在短视频流行度预测时，社交属性特征被证明是最重要的线索。

多模态融合旨在建立能够处理和关联多种模态的信息的模型。但是由于不同模态特征信息之间异构性，多模态特征融合的领域存在一些独特且复杂的挑战，主要表现在以下五个方面：

- 表示：第一个挑战是在表示多模态数据中，如何充分利用不同模态之间的相关性和独立性信息。由于多模态信息之间的异构性，使得构建多模态统一的特征表示变得异常困难。例如，文字表示具有很强的象征性的，而视频一般直接被表示成信号。
- 转换：第二个挑战是如何将某一个模态下的特征表示转换到另一个模态下的特征表示。不同模态特征之间不仅仅有异构性，而且不同模态信息之间的关系往往是主观的或是开放的，并没有统一量化的表示。例如，对于一张图片，可以从很多角度来进行文字描述。
- 对齐：第三个挑战是如何确定不同模态中的要素之间的直接关系，并且在提取不同模态特征的时候导致这些特征的维度并不对齐。例如，短视频中画面-语音-字幕的对齐，为了应对这一挑战，模型需要衡量不同模态之间的相似性，并处理可能的长期依赖性和模糊性。
- 融合：第四个挑战是如何将来自不同模态信息结合起来进行预测。来自不同模态的信息在预测时重要程度并不是等同的且不同模态信息之间的噪声表现形式也不同，同时还存在某个模态信息的缺失情况。
- 共同学习：第五个挑战是如何转移在多模态信息、多模态统一特征表示和预测模型之间的有用信息。共同学习探索从一种模态学习知识如何帮助在不同模态训练上计算模型。

由于传感器的日益发展，利用从不同来源或结构获得的异构数据进行分类和回归的多模态融合已经成为机器学习的一个重要问题。多模态特征融合的目的是关联和处理来自多种模态的信息，可以为模型决策提供更多的信息，从而提高了决策总体结果的准确率。例如，在检测体育视频中的事件时，视觉特征和文本特征信息的融合会变得更加有效[41]，仅仅是单一模态的信息是做不到的，所以需要设计合适多模态融合的算法模型。接下来将从多模态的融合层次和融合方法展开详细介绍。

2.2.1.1 融合层次

不同模态的融合通常在两个层面上进行：特征层面的早期融合、决策层面的后期融合和混合融合。

- 早期融合：也是运用最广泛的融合策略，在特征层面融合各个模态的信息。早期融合优势主要是可以利用来自不同模态特征之间的相关性，以获得一个更完备的多模态统一特征表示。另外早期融合只需要对组合特征向量进行一个一个学习阶段[42]。但是，当多模态的数量增多是，早期融合将变得困难。
- 后期融合：决策级融合[43]是在语义空间中将各个模态特征融合到一起。其优势主要在于各个模态的特征映射到语义空间后，原本不同的表示会变得相同，这是由于子空间中各模态的信息往往是相同的。因此，在融合多模态特征时会更加方便且精确，这在早期融合中是难以实现的[44]。后期融合策略的另一个优点是它允许使用最合适的方法来分析每种单一模态，这提供了比早期融合更大的灵活性。但是，后期融合方法的没能使用模态之间的特征级的相关性。此外，由于使用不同的分类器来获得局部决策，它们的学习过程变得非常耗时。
- 混合融合：将这两种方法组合使用被称为混合融合方法[45]。混合融合方法可以利用早期和晚期融合策略的优势。因此，许多研究人员已经使用混合融合策略来解决各种多媒体数据的分析问题[46]。

2.2.1.2 融合的方法

融合的方法主要分为三大类：基于规则的方法、基于分类的方法和基于估计的方法。这种分类是基于这些方法的本质，它本质上意味着问题空间的分类。例如，通过基于估计的方法解决估计参数的问题。同样，基于特定观察获得决策的问题可以通过基于分类或者基于规则的方法来解决。

- 基于规则的融合方法[47]：这类方法包括多种用来组合多模态信息的基本规则，包括线性加权融合、最大、最小和多数投票等。Jain 等人[48]的工作从理论上介绍了这些规则。除了这些规则之外，还有为特定应用程序透视图构建的自定义规则。如果不同模态之间的时间对齐质量良好，则基于规则的方案通常表现良好。
- 基于分类的融合方法：这类方法包括SVM[49]、贝叶斯推理[50](Bayesian inference)、动态贝叶斯网络[51]、神经网络[52](Neural networks)和最大熵模型[53](Maximum entropy model)。这些技术已经用于将多模态分类为预

定义的类别之一。

- 基于估计的融合方法：估计的类别包括卡尔曼滤波器、扩展卡尔曼滤波器和粒子滤波器[54]等融合方法。这些方法主要用于基于多模态数据，从而更好地估计移动对象的状态。例如，对于目标跟踪任务，融合音频和视频等多种模态来估计物体的位置。

2.2.2 多模态特征表示

根据不同的多模态融合的基础框架，多模态特征表示主要分为联合表示、协调表示和编解码表示。

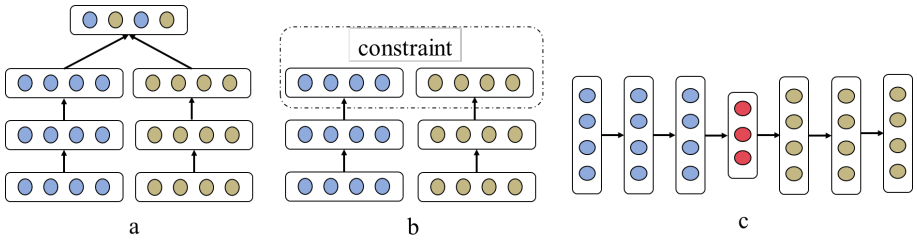

图 2-1 多模态特征表示的三种框架。(a)联合表示框架的目的学习一个共享的语义子空间；(b)协调表示框架在某种约束条件下学习每个模态的独立但协调的表示；(c)编解码表示框架在保持语义一致的情况下将一种模态特征转换成另一种模态特征。

2.2.2.1 联合表示

联合表示将单模态的特征投影到一个共享的语义子空间中，在子空间中融合多模态特征以获取多模态统一的特征表示[55]。如图 2-1(a)所示，每个模态特征单独经过各自的子网络编码之后被映射到共享的语义子空间中，在该子空间中，模态共享的信息将融合为一个统一的向量表示。尽管该模型框架是为了训练和推理多模态提供的并行数据，但是也需要能够处理某些模态中的部分数据丢失的问题。以便有更多的训练数据可以在数据从一种或多种模态中丢失的情况下，被利用或下游的任务的性能不至于受到很大的影响。为此，一种广泛的解决方法是通过模态关联性学习解决模态间数据缺失、模态干扰等问题。Hazarika 等人[56]针对多模态情感分析种的模态异构问题，提出一种将不同模态特征分解为不变模态特征表示与特定模态特征表示的学习框架。井等人[57]提出了一种基于低秩多视角嵌入学习的短视频流行度预测算法。

为了模型有更好的性能，期望融合后的特征向量能够包含不同模态各自独立部分的语义信息。独立属性无法自动保证，这是由于联合表示倾向于在忽略模态特定信息的同时保留模态之间的共享语义信息。所以为了充分利用不同模态之间

的特定信息，Jiang 等人[58]通过对网络参数施加一个迹范数正则化以揭示多模态特征的隐藏独立性和多样性。在多模态自动编码器[59]中使用重构损耗可以被视为一个正则化项，起到了保持模态独立性的作用。

与其他框架相比，联合表示的优点主要有两个，一个是不需要对模态映射施加约束，因此可以方便的融合多种模态特征；另一个是共享的公共子空间趋于模态不变，这有助于将信息从一个模态转移到另一个模态中[60]。但是，该框架的缺点之一就是不能用于推断每个模态的独立表示。

2.2.2.2 协调表示

协调表示的主要目的是为了协调子空间中每个模态学习独立但受约束的表示。如图 2-1(b)所示，由于不同模态中包含的信息是不平等的，学习独立表示有助于保持独有且有用的模态特定信息[61]。协调表示方法主要分为基于跨模态相似性和基于跨模态相关性。基于跨模态相似性的方法旨在学习一个公共子空间，其中可以直接测量来自不同模态的向量的距离[62]，跨模态相似方法学习相似性度量约束下的协调表示，模型的学习目标是保持模态间和模态内的相似结构，期望与相同语义或对象相关的跨模态相似距离尽可能小，而与语义不相关的距离尽可能大；而基于跨模态相关性的方法旨在学习一个共享子空间，使得来自不同模态的表示集的相关性被最大化[63]。

协调表征在每种模态中保持独立且有用的模态特定信息[61]。由于在独立的网络中对不同的模态进行重新编码，其优点之一是每个模态可以单独推断。这一特性也有利于跨模态的迁移学习，即跨不同的模态迁移知识。该框架的一个缺点是，在大多数情况下，很难学习两种以上模态的表示。

2.2.2.3 编解码表示

编解码表示是通过一个编解码器来学习一种模态映射到另外一种模态的中间表示。编码器-解码器框架已广泛用于多模态翻译任务，如图像标题[64]、视频描述[65]、和图像合成[66]。如图 2-1(c)所示，编码器-解码器框架主要由编码器和解码器两个组件组成。编码器将原始模态映射为潜在向量 $\mathbf{z}$，然后，基于向量 $\mathbf{z}$，解码器将生成新的目标模态样本。

注意力机制已经被广泛用于编解码表示框架中。这是由于注意力机制不仅仅使用编码器最后一步产生的单个向量，还使用了中间表示[67-68]。由于注意力机制能够选择突出的特征，它已经成功的应用在各种神经网络中。例如，Fan 等人[69]基于显著性转移感知注意机制提出了显著性转移感知模块(Saliency-Shift Aware VSOD, SSAV)，并被用来在时间域上动态地捕捉显著性的转移。Chu 等人[70]提出

了空间注意力图的概念并且通过给损失加上权重的方式控制网络参数更新的方式，解决了在多目标跟踪时多个目标相互遮挡而导致的跟踪错误的问题。Sharma 等人[71]将软注意力机制应用到卷积特征图，对输入帧的不同区域的卷积特征进行加权平均，组成为视频的全局特征，然后将其应用在视频中动作识别的任务。Girdhar 等人[72]将注意力机制应用到卷积神经网络当中的池化层，使得网络能够自动将注意点放在与动作类别息息相关的感兴趣区域。Yao 等人[73]将三维卷积神经网络提取后的视频特征利用软注意力机制将其加权并输入到 RNN/LSTM 中，并且通过解码器得到其对于视频的描述。

与其他框架相比，编码器-解码器框架的优点之一是能够在原始模态表示的基础上生成目标模态条件的新样本。但是，该框架的缺点是每个编码器-解码器只能对其中一种模态进行编码。

第 3 章 基于深度多模态融合的短视频流行度预测

由于传感器的日益发展，利用从不同来源或结构获得的异构数据进行回归或是分类的多模态特征融合已经成为机器学习中的一个比较重要的任务。现实生活中，人们处在这些领域相互交融的环境，人们所看到的视觉信息、听到的音频信息以及相关的文本内容信息都是这些领域上的模态形式。对于短视频而言，本文提取了短视频的视觉模态、音频模态、文本模态和社交属性模态这四个模态特征。如何充分挖掘和利用短视频多模态之间的相关性和独立性信息，是获取短视频多模态统一特征表示的关键。近几年来，由于深度学习(Deep Learning, DL)在图像识别、语音分析、机器翻译和自然语言处理等领域上取得了非凡的成果。所以本文使用是深度学习相关的算法来解决多模态融合任务。

首先，假设有一个含有 N 个带有流行度得分的短视频样本的数据库，并对该数据集提取了 K 个不同模态特征，因此，得到特征矩阵 $\mathbf{x}=[\mathbf{x}_1;\mathbf{x}_2;\cdots\mathbf{x}_K]$，其中 $\mathbf{x}_u\in\mathbb{R}^{D\times N}$ 表示第 u 个模态的特征，D 表示第 u 个模态特征对应的维度。对于短视频流行度得分，将其表示为 $y=[y_1,y_2,\cdots y_N]\in\mathbb{R}^N$，$y_i$ 表示第 i 个样本对应的流行度得分。因此，输入模型的第 i 个数据可以表示为 $\mathbf{x}^{(i)}=\{\mathbf{x}_u^{(i)}\in\mathbb{R}^D\}_{u=1}^K$，将批量(batch-size)的大小设为 n，深度多模态融合的目标是学习一个多层的深度学习网络 $f(\mathbf{x}^{(i)})$（本文主要使用的是 MLP 网络结构），其输出是 $\hat{y}_i$，在回归问题中期望输出的预测值 $\hat{y}_i$ 和真实值 y_i 之间的差值越小越好，所以可以通过最小化预测损失函数来实现这一期望：

$$\min_f \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\left(\hat{y}_i=f(\mathbf{x}^{(i)}),y_i\right). \tag{3-1}$$

3.1 问题描述

由于短视频数据具有多模态特性，且不同模态信息之间存在关联性，因此，如何充分挖掘和利用短视频多模态之间的相关性和独立性信息，来获取短视频多模态统一的特征表示是多模态融合的主要问题。首先，与经典的多模态任务类似，短视的多模态特征从不同的角度下共同描述了一个短视频，所以不同模态特征在语义层面上是相互关联的，即多模态信息之间的相关性。因此，在多模态统一表示学习中需要将不同模态信息之间的相关性信息结合起来以获取一个更全面的多模态统一特征表示。其次，不同模态的信息分别从不同的视图上对短视频进行

描述，不同模态信息之间都蕴含了特有信息，所以每个模态特征之间存在独立性。同时，短视频的兴起是以传播和社交为导向的，模态信息的不确定性更为明显，传统的多模态统一特征表示方法并不适合直接用于短视频的多模态统一特征表示问题。例如，短视频的“时长短”以及为了传播和社交的“强编辑”等特性容易导致某个单一模态表征能力不足、模态信息缺失以及模态信息与主旨不一致导致的噪声问题。因此，本章提出了一种基于深度多模态融合的短视频流行度预测算法。

3.2 算法模型

针对短视频多模态的特点，本节通过一个深度信道交互模型来融合这四个模态特征得到一个多模态统一的特征表示。模型主要通过深度信道交换模块来实现多模态融合，融合模态之间的相关性和独立性信息，获取一个更全面的多模态统一特征表示。然后利用多模态统融合后的统一特征表示来预测短视频的流行度得分。算法模型框架如图 3-1 所示：

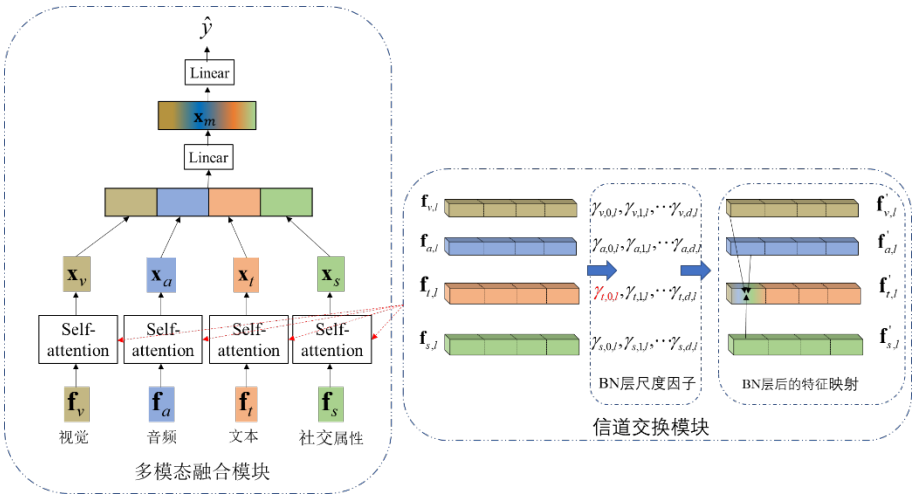

图 3-1 基于深度信道交换的多模态融合模型。左边为多模态融合模块，$\mathbf{f}_v$、$\mathbf{f}_a$、$\mathbf{f}_t$和$\mathbf{f}_s$为输入，分别代表视觉、音频、文本和社交属性模态特征，其中 Self-attention 表示自注意力机制网络；右边为信道交换模块，应用在多模态融合模块中的 BN 层中。BN 层中尺度因子标红的表示当前通道的$\gamma_{u,d,l} < \theta$，需要进行信道交换。

3.2.1 多模态融合模块

多模态融合模块首先利用一个自注意力机制(Self-attention)网络来探寻不同模态之间的相关性来解决原始模态特征中的维度不统一、数据缺失等问题，然后将补全的特征向量直接级联起来映射到一个低维空间作为融合后的多模态统一特征表示。

多模态关联性学习主要是采用自注意力机制对原始特征进行编码。多模态关联性学习包括两个子模块，分别是特征映射模块和关联度计算模块。利用多模态关联性学习探寻不同模态信息之间的相关性，解决原始特征之间存在的数据缺失、特征维度不统一等问题。

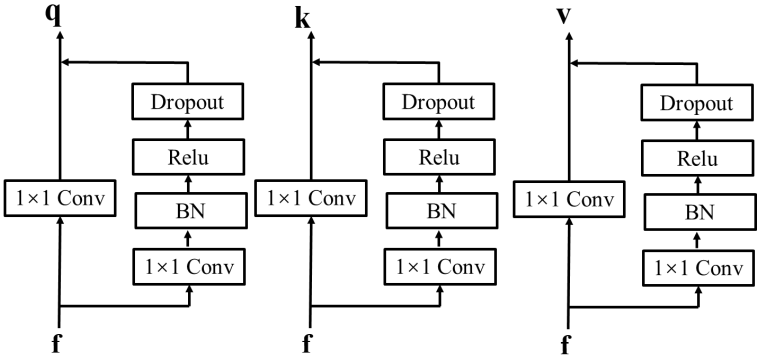

图 3-2 特征映射模块示意图。1×1 的卷积网络对原始特征进行升维（降维）操作；BN 层主要用于信道交换。

尽管短视频的多模态信息之间存在比较强的相关性，但是由于短视频的“时长短”以及为了传播和社交的“强编辑”等特性容易导致某个单一模态表征能力不足、模态信息缺失以及模态信息与主旨不一致导致的噪声问题。因此，直接获取的原始特征并不合适直接用于短视频的流行度预测。针对这一问题，采用自注意力机制进行模态关联性学习，利用不同模态间的交互进行特征编码。具体地，首先将四个不同模态的原始特征向量通过特征映射模块，如图 3-2 所示，每一个原始特征向量在映射时会综合考虑其线性映射和非线性映射。针对线性映射，利用 1×1 的卷积网络对原始特征进行升维（降维）操作；在非线性映射中，在保留 1×1 的卷积网络的同时，添加了非线性激励函数 Relu、批量归一化层(Batch Normalization, BN)和 Dropout 层。Relu 层为了增加非线性映射，并且使网络更加稀疏的同时提高特征的泛化能力。BN 层中为信道交换提供了条件，关于信道交换的条件和过程将在 3.2.2 节深度信道交换中详细描述。将线性映射的特征向量和非线性映射的特征向量相加得到最终的输出向量。经过线性和非线性的映射增加了所提取特征的泛化能力。升维（降维）是为了特征对齐，解决视觉、音频、文本和社交属性特征之间维度差异比较大的问题。最终，经过特征映射模块后得到三个向量，它们分别是查询(Query)向量、键(Key)向量和值(Value)向量，并将其简略表示为 $\mathbf{q}$、$\mathbf{k}$、$\mathbf{v}$。

由于四个模态都是对于同一个短视频的描述，所以经过提取的特征在语义内容上具有一定的相似性。针对这种情况，本文使用残差映射的方式进行特征映射，通过构建三个相互独立但是结构相同的网络模块将特征 $\mathbf{f}$ 映射为对应的查询向

量$\mathbf{q}$、键向量$\mathbf{k}$和值向量$\mathbf{v}$，然后使用点积注意力的方式去计算模态之间的相关度，通过利用当前模态的查询向量去探寻与所有模态特征的键向量的和的相关性，然后将得到的相关性得分作为各个模态特征的值向量的权重。这种方法的目的是使得短视频的四个模态在语义内容相似的部分在经过自注意力网络之后特征差异性尽可能的小。这样通过键向量$\mathbf{k}$的交互能够让各个模态在语义内容上相似的部分相互补充的同时，还保持了不同模态信息的独立性。具体计算如图3-3所示，增加权重后的特征向量Attention($\mathbf{f}_u$)记为：

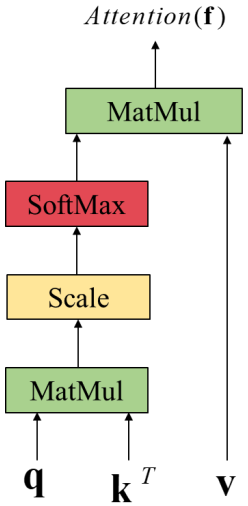

图 3-3 关联度计算模块示意图

$$\text{Attention}(\mathbf{f}_u)=\frac{\mathbf{q}_u\mathbf{k}^T}{\sqrt{d_k}}\mathbf{v}_u, \tag{3-2}$$

其中$u\in\{v,a,t,s\}$，$\mathbf{k}=(\mathbf{k}_v+\mathbf{k}_a+\mathbf{k}_t+\mathbf{k}_s)$，$\mathbf{q}_u$、$\mathbf{v}_u$分别是$\mathbf{f}_u$经过自注意力机制网络后的查询向量和值向量，$\mathbf{k}_v$、$\mathbf{k}_a$、$\mathbf{k}_t$和$\mathbf{k}_s$分别为视觉模态特征、音频模态特征、文本模态特征和社交属性模态特征经过自注意力机制网络后的键向量，d_k是查询向量的维度。借鉴残差网络的结构，构建由多层自注意力机制模块组成的仿残差网络模块，按照公式(3-3)规则进行更新：

$$\mathbf{x}_{n+1}=\mathbf{x}_n+\text{Attention}\left(\mathbf{x}_n\right). \tag{3-3}$$

用自注意力机制网络对原始特征进行编码，最终得到具有统一维度表示的各个模态特征向量$\{\mathbf{x}_v,\mathbf{x}_a,\mathbf{x}_t,\mathbf{x}_s\}\in\mathbb{R}^D$，$D$是特征向量统一后的维度。

然后，将上述得到的四个模态特征向量级联起来投影到低维特征向量中，得到$\mathbf{x}_m$：

$$\mathbf{x}_m=\text{Linear}([\mathbf{x}_v;\mathbf{x}_a;\mathbf{x}_t;\mathbf{x}_s];\Theta_1), \tag{3-4}$$

其中 Linear 表示线性回归函数，Θ_1为待学习参数。最后，将多模态融合后的特

征向量 $\mathbf{x}_m$ 用来预测短视频的流行度得分 $\hat{y}$：

$$\hat{y} = \text{Linear}(\mathbf{x}_m; \Theta_2), \tag{3-5}$$

其中，Θ_2 为待学习参数。

3.2.2 深度信道交换模块

深度信道交换网络是一种无参数的多模态融合框架，可以在不同模态之间的网络层中动态交换信道信息。具体地，信道交换的条件和过程是由单模态信道中的重要性来动态的交换信道，这个重要性是由批量归一化(Batch Normalization, BN)层的尺度因子来衡量的。每个单模态的子网络中都包含独立的 BN 层，这使得多模态网络框架和单模态网络框架一样紧凑且易于训练。接下来将详细描述 BN 层进行信道交换的条件和过程。

3.2.2.1 批量归一化

批量归一化和普通的数据归一化的原理类似，都是将分散的数据变得有规律的一种算法，也是深度学习中优化神经网络的一个常用的模块。BN 层的作用主要是让数据变得更同分布，而同分布的数据会使得机器学习的算法在训练时更易收敛且效果更好[74]。在神经网络中，数据分布对模型训练有很大的影响。例如，某个神经元 x_1 的值为 1，另外一个神经元 x_2 的值为 20，某个参数权重 w 的值为 1，计算当前层的输出可以得到 $wx_1 = 0.1$ 和 $wx_2 = 2$。当前的输出经过一层激励函数时，若当前的激励函数是 tanh 时，可得到 $\tanh(wx_1) \approx 0.1$ 和 $\tanh(wx_2) \approx 1$，但是接近于 1 的部分已经到达激励函数的饱和区域，接下来无论 x_2 的值在怎么扩大，经过 tanh 激励函数后的输出都是接近于 1。在这种情况下，说明神经网络在初始阶段对于特征中比较大的值不再敏感了，这就是激活函数的梯度饱和效应。

激活函数都存在梯度饱和区域，当激活函数的输入过大或者过小时，进入激活函数的梯度饱和区域，这样的数值会导致激活函数的梯度值趋近于 0，使得网络收敛的比较慢，甚至导致梯度消失。在加入 BN 层后，可以有效的解决这类问题。它的策略是在调用激活函数之前将 $wx+b$ 的值归一化到梯度值比较大的区域中，如公式(3-6)所示。

$$z = h(BN(wx+b)), \tag{3-6}$$

其中 h 表示激活函数，BN 表示批量归一化，BN 层应在激活函数层之前使用。

3.2.2.2 BN 的训练过程

批量处理的大小为 n 时，在前向传播过程中，每个网络节点都有 n 个输出，BN 层的作用就是对该层每个节点的这 n 个输出先进行归一化然后再输出：

$$\mathrm{BN}_{\gamma,\beta}:\mathbf{x}\rightarrow\hat{\mathbf{x}}', \tag{3-7}$$

其中 γ 和 β 是待学习的参数。如图 3-4 所示：

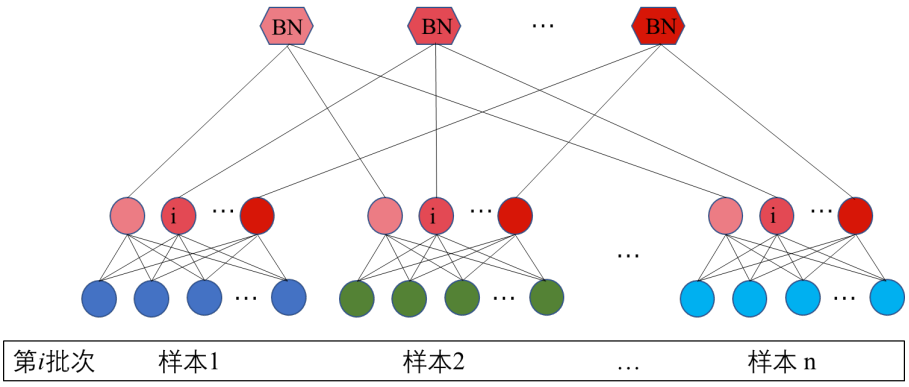

图 3-4 神经网络 BN 层示意图

BN 层的操作主要分为两步：

首先归一化，首对 n 个输入向量 $\mathbf{x}$ 进行归一化处理，得到 $\hat{\mathbf{x}}$：

$$\hat{\mathbf{x}}=\frac{\mathbf{x}-\mu}{\sqrt{\sigma^2+\varepsilon}}, \tag{3-8}$$

其中 μ 表示在第 i 信道上这个批量中所有样本数据的均值，σ 表示在第 i 信道上这个批量中所有样本数据的标准差，ε 是一个小常数以避免除以 0。简单的对每层的输入进行归一化可能会改变其可以表示的内容。例如，在网络中间学到的数据特征分布原本就分布在激活函数的两侧，但是在经过归一化处理之后，将数据分布强行变为了标准正态分布，使得数据变换成分布在激活函数的中间部分，这就相当于当前层所学习到的特征分布被破坏了。因此，在 Bn 层中引入了尺度和平移变换。

然后尺度和平移变化，缩放并平移到新的分布 $\hat{\mathbf{x}}'$：

$$\hat{\mathbf{x}}'=\gamma\hat{\mathbf{x}}+\beta, \tag{3-9}$$

其中 γ 是尺度变换因子，β 是平移变换因子。BN 层通过 γ 和 β 将 $\mathbf{x}$ 映射到 $\hat{\mathbf{x}}'$，γ 和 β 是待学习的参数。当 $\gamma=\delta$，$\beta=\mu$ 时，就可以恢复原始的数据分布。

关于 BN 层的算法伪代码如算法 1 所示：

算法 1：BN 层算法
输入：一个批量的数据 $\{\mathbf{x}_{1\cdots n}\}$
计算批量平均值和方差：$\mu \leftarrow \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ 计算批量方差:$\sigma^2 \leftarrow \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i\text{-}\mu\right)^2$ 归一化：$\hat{\mathbf{x}}_i \leftarrow \frac{\mathbf{x}_i-\mu}{\sqrt{\sigma^2+\varepsilon}}$ 尺度和平移变化：$\hat{\mathbf{x}}_i^{'} \leftarrow \gamma\hat{\mathbf{x}}_i+\beta$
输出：$\hat{\mathbf{x}}_i^{'}$，γ，β

3.2.2.3 深度信道交互

BN 层广泛用于深度学习中，用来消除协变量移位并提高模型泛化能力。图 3-1 是深度信道的框架，在深度多模态学融合的网络中加入了 BN 层，以实现信道交换。在这一小节中，将介绍具体的信道交换的过程和条件。

在经过模态关联性学习后，四个模态维度被统一到 D 维，一个批量中有 n 个样本，所以，针对某个单模态子网络中的 BN 层，可以得到：

$$\mathbf{f}_{u,d,l}^{'}=\gamma_{u,d,l}\frac{\mathbf{f}_{u,d,l}-\mu_{u,d,l}}{\sqrt{\sigma_{u,d,l}^2+\varepsilon}}+\beta_{u,d,l}, \tag{3-10}$$

其中 $u\in\{v,a,t,s\}$，$\mathbf{f}_{u,d,l}$ 表示 u 模态的子网络的第 l 层中 d 信道中的特征向量，$\mu_{u,d,l}$ 和 $\sigma_{u,d,l}$ 分别表示当前批量数据中 d 信道所有向量的均值和方差，$\gamma_{u,d,l}$ 和 $\beta_{u,d,l}$ 分别表示可训练的尺度因子和偏移量，ε 是一个小常数以避免除以 0。$\mathbf{f}_{u,d,l}^{'}$ 在经过一个非线性的激活函数后作为 l+1 层的输入。

在公式(3-10)中，尺度因子 $\gamma_{u,d,l}$ 评估了训练过程中的输入 $\mathbf{f}_{u,d,l}$ 和输出 $\mathbf{f}_{u,d,l}^{'}$ 之间的相关性，当 $\gamma_{u,d,l}\to 0$ 时，则最后的损失函数相对于 $\mathbf{f}_{u,d,l}^{'}$ 的梯度也将趋于 0，这就意味着 $\mathbf{f}_{u,d,l}^{'}$ 对最终的预测的影响比较小。因此该信道中的信息变得多余，在最后的预测结果上并没有什么作用，在这种情况下，将其他信道中信息代替当前信道，使得尺度因子比较小的信道不至于在后续的网络中失去作用。替换条件由公式(3-11)给出：

$$\mathbf{f}_{u,d,l}^{'}=\begin{cases}\gamma_{u,d,l}\dfrac{\mathbf{f}_{u,d,l}-\mu_{u,d,l}}{\sqrt{\sigma_{u,d,l}^2+\varepsilon}}+\beta_{u,d,l} & \gamma_{u,d,l}\geq\theta\\ \dfrac{1}{K-1}\displaystyle\sum_{u^{'}\neq u}^{K}\gamma_{u^{'},d,l}\dfrac{\mathbf{f}_{u^{'},d,l}-\mu_{u^{'},d,l}}{\sqrt{\sigma_{u^{'},d,l}^2+\varepsilon}}+\beta_{u^{'},d,l} & \text{其他}\end{cases}, \tag{3-11}$$

其中 θ 是一个小常数，表示如果当前信道的尺度因子小于阈值 θ，则用其他信道

的平均值替换当前信道。也就是说当某一种模态的一个信道对最终的预测影响很小时，那么就将其替换成其他模态下的当前信道特征的平均值特征表示。并且，在信道交换时，模型只选择尺度因子最小的一个模态信道进行替换，因此，如果当多模态特征经过 BN 层后，有多个模态信道的尺度因子都小于θ，模型只对尺度因子最小的那个模态信道特征进行替换，而其他模态信道的特征不替换。将所有模态应用公式(3-11)，然后将它们输入非线性的激活函数，最终将非线性函数的输出作为下一层网络的输入。

综上所述，为了对网络进行批量归一化，在网络中的激活函数层之前插入 BN 变换，所以原先以 **x** 作为输入的激活函数层，现在替换成 BN(**x**) 作为输入。为使 BN 层能够尽可能的发现无效的信道，所以对尺度因子使用 $\mathcal{L}_1$ 范数约束，使其稀疏化，所以该模型的最终损失函数为：

$$\mathcal{L}=\frac{1}{M}\sum_{i=1}^{M}\left((\hat{y}_i-y_i)^2\right)+\lambda\sum_{u=1}^{K}\sum_{d=1}^{D}\sum_{l=1}^{L}\left|\gamma_{u,d,l}\right|, \tag{3-12}$$

其中 M 为测试集的样本个数，$\hat{y}_i$ 表示模型预测的短视频流行度得分，y_i 表示真实的短视频流行度得分，K 表示模态个数，D 表示信道个数，L 表示网络层的层数，λ 是平衡因子。

3.3 实验设置

3.3.1 实验数据及设置

本文中使用的短视频流行度预测数据集是由新加坡国立大学媒体实验室构建 [3]。这个数据集总共包含 303242 个从在线短视频分享网站 Vine 收集的用户生成的短视频，这些视频由 98166 个用户上传。其中约 75%的视频长度为 6 至 7 秒。由于受欢迎程度与在线社交互动高度相关，因此在计算短视频的最终受欢迎程度分数时，需要考虑评论数、转发数、喜欢数和浏览循环数四类统计数据的平均值，并将其归一化，最终分数在 0 到 1 之间。

本次实验对数据集做了 10 轮随机实验，在每一轮中，本文实验用 90%的样本进行训练，剩下的用于测试，最终取 10 次测试的平均结果。模型采用随机梯度下降(Stochastic gradient descent, SGD)算法优化，根据经验将学习率设为 0.01。然后根据网格搜索的方法对模型中的θ、λ进行最优化搜索，首先确定一个粗略的范围，然后再粗略的范围中更加精细的网格搜索，最终得到θ=0.05，λ=0.4。训练和测试都是在 GPU 上完成的，显卡配置是 GeForce RTX 3090。算法模型运

[3] http://acmmm2016.wixsite.com/micro-videos.

行环境具体如下：python3.6；pytorch1.7.1；numpy1.19.5。

3.3.2 特征提取

在本节中，将提取短视频的视觉模态、音频模态、文本模态和社交属性模态这四个模态特征，分别用$\mathbf{f}_v$、$\mathbf{f}_a$、$\mathbf{f}_t$和$\mathbf{f}_s$来表示。

3.3.2.1 视觉模态特征

由于短视频“时长短”的特性，所以一个短视频的主旨会贯彻视频始终，因此基于关键帧的表示策略在表述其固有主题上是稳健的。受此特性的启发，本文采用平均池化操作融合从关键帧中提取的特征，并用以表示短视频的视觉特征。

- 颜色直方图特征(Color Features)：如文献[75]所述，简单的图像特征和与流行度预测的相关性比较小。由于颜色直方图可以通过显示醒目的颜色来吸引更多的注意力，因此颜色被分组为 50 种不同的颜色，本文直接使用其模型提取了 50 维的颜色直方图特征向量。
- 目标特征(Object Features)：由于深度卷积神经网络在视觉理解任务中的强大性能[76]，训练有素的 AlexNet ImageNet 模型被用于直接表示关键帧。最后一个完全连接层 fc7 的输出作为 1000 路 softmax 的输入，并生成 1000 个类别标签上的分布，该分布被视为对象特征的最终表示。所以本文直接用训练好的 AlexNet ImageNet 模型来提取 1000 维的目标特征。
- 情感特征(SentiBank Features)：根据文献[77-78]研究表明，各种情绪概念对于流行度的预测是有影响的。Chen 等人[78]训练了一个 DeepSentiBank 的深度 CNN 模型，最后一个完全连接层的输出作为 2089 路 softmax 的输入，并生成 2089 个情绪概念上的分布作为高级情感特征，用于视觉情感概念的分类。本文直接用其模型提取的 2089 维的情感特征。
- 美学特征(Aesthetic Features)：美学规定了人类感知的高度主观性。美学研究[79]表明，高美学质量使某些图像比其他图像更具吸引力。在 Bhattacharya 等人[80]的视频美学评估之后，提取了包括暗通道、清晰度、锐度、眼睛敏感度、低景深、白平衡、色彩统计的 149 维视觉统计特征作为美学特征。

本文直接将上述四种类型的视觉特征直接级联起来，形成一个 3288 维的特征向量。

3.3.2.2 音频模态特征

音频特征对于各种任务至关重要，例如跨媒体关联计算中，音频信息可以为视觉信息提供补充信息，特别是对于视频中的视觉信息不足或者噪声干扰比较大时。目前已经有研究者在短视频中嵌入音频信息来提高学习性能[3-5]。例如，Chen等人[3]以声学特征为输入，研究了声学特征对短视频流行度预测的影响。Zhang等人[5]基于堆栈去噪自动编码器提取声学特征，以更全面地表示短视频。按照Chen 等人[3]的模型设置，本文使用从音频通道中提取的 552 维特征来表示短视频的声学模态，包括梅尔频率倒谱系数、能量熵、信号能量、过零率、频谱滚降、频谱质心，和光谱通量。声学特征的范围是连续的实数值。

3.3.2.3 文本模态特征

额外的文本信息为从不同方面理解短视频内容提供了新的线索。最近在社交媒体流行度预测方面的工作已经将文本信息视为改进预测性能的一个不可或缺的组成部分。在正常情况下，短视频的文本信息包含了短视频的主题、类型和情感等方面的内容。这对短视频流行度的预测至关重要。例如，Mishne 等人[81]利用情境的情绪值作为指标，预测电影在销售方面的受欢迎程度。Sentence2Vector[4]是一个经典的文本特征提取工具，用于生成短视频主题表示的 100 维特征。斯坦福大学 CoreNLP tools[5]提供了一种文本情感分析工具。通过利用情绪分析工具，每个短视频被分配到一个情绪分数，该分数是一个 0 到 4 之间的整数，分别对应于“非常消极”、“消极”、“中性”、“积极”和“非常积极”。在分析数据之前，本文首先分别对每个文本特征类型进行规范化，并将它们连接在一起,形成一个 101 维向量并将其作为最终的文本模态特征。

3.3.2.4 社交属性模态特征

虽然一些相关工作已经证明低级视觉特征和高级语义特征能够在一定程度上预测流行度，但社交线索是决定短视频传播范围的重要因素。例如，短视频的关注者，当一个短视频发布者有了大量关注者时，该发布者发布一个短视频往往会吸引更多的潜在观众，从而有一个更高的流行度得分。因此，本文编码了 4 种类型的社交线索：

- 跟随者计数：给定发布者的关注者和被关注者的数量；
- 播放计数：短视频上传后的播放次数和发布者的所有短视频总播放次数；

[4] https://github.com/klb3713/sentence2vec.
[5] http://stanfordnlp.github.io/CoreNLP/.

- 发帖数：每个发布者的总发贴数；
- 推特验证：反映发布者是否为经过验证的用户的二进制值。

将上述所有数据级联，形成一个由总播放量、当前样本的播放量、关注者数量，被关注者数量、总发帖数量和是否验证构成的 6 维的向量，并作为最终社交属性模态的特征表示。

3.3.3 评价指标

本文用归一化均方误差(normalized Mean Squared Error, nMSE)[82]来衡量预测值和真实值之间的一致性：

$$\mathrm{nMSE}=\frac{1}{M\sigma^2}\sum_{i=1}^{M}(y_i-\hat{y}_i)^2 \tag{3-13}$$

其中 M 是测试样本的个数，σ 是真实的短视频流行度得分的标准差，y_i 是短视频流行度得分的真实值，$\hat{y}_i$ 是短视频流行度得分的预测值。nMSE 的值越小说明性能越好。

3.4 实验结果和分析

为了全面验证所提出的算法，在实验中，从以下五个方面对所提出的算法模型进行验证：

- 收敛性分析：基于本文的算法模型，测试算法的收敛性。
- 模块分析：为了验证算法模型中不同模块的有效性，通过删除相关模块后比较其预测性能。
- 特征分析：为评估模态特征对短视频流行度预测得分的贡献，本文考虑不同模态之间的性能比较。
- 参数分析：在 nMSE 指标下分析 λ 对于本文提出的模型的性能影响。
- 与现有的方法比较：通过与几种主流算法的性能比较，验证了该方法的有效性。

3.4.1 收敛性分析

快速收敛是评价一个算法是否是可训练的重要指标。为了验证本章所提出算法的收敛性，随机选取一次实验，将 nMSE 作为评价指标，通过观察 nMSE 数值随迭代次数变化的趋势判断该算法是否收敛。如图 3-5 展示了基于深度信道交换的多模态统一特征表示算法的收敛性。

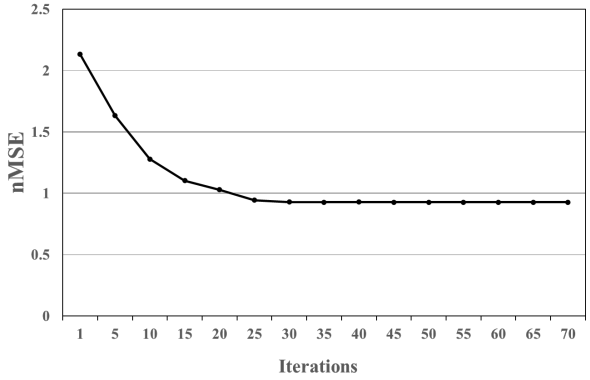

图 3-5 收敛曲线：nMSE 随模型迭代次数的变化

在图 3-5 中，x 轴表示算法的迭代次数，y 轴表示对应迭代次数下的 nMSE 的数值。从图中可以看出，nMSE 随着迭代次数的增加而逐渐减小，且在迭代到 25 次左右达到稳定。这证明了本文提出的算法模型能够经过训练而收敛，验证了算法的可行性。

3.4.2 模块分析

本章节提出深度信道交换的算法模型是将通过一个自注意力机制网络去探寻不同模态之间的关联性，同时利用 BN 层在不同模态信道之间进行信道交换来实现多模态的融合。因此，这一小节将通过删除相关的模块然后比较预测性能，验证该模型所提出的两个模块的有效性。实验设计分为以下两个策略：

- 删除自注意力(Self-attention)机制网络：用一个简单的线性回归函数代替自注意力机制网络，不考虑原始模态特征的数据缺失、干扰和维度差异性等问题。
- 删除深度信道交换(Channel Exchanging)：在算法模型的网络层中，不考虑不同模态的子信道中信息失效问题。

表 3-1 中“√”表示使用该模块，“×”表示删除该模块。

表 3-1 模型的不同成分对短视频流行度预测性能的贡献

	第一组	第二组	第三组
Self-attention	×	√	√
Channel Exchanging	×	×	√
nMSE	0.984	0.956	0.927

表 3-1 展示了深度多模态统一特征表示算法模型在缺少不同模块情况下的预

测性能。从表中可以看出在删除自注意力机制网络、删除深度信道交换模块后，nMSE 指标均有所上升，即算法模型性能变差。因此，本节提出的两个模块都是有效的。在第一组删除自注意力机制网络和不使用深度信道交互的情况下，模型性能最差，说明原始的模态特征缺失不适合直接用于预测。模态特征的数据缺失、干扰和维度差异性等问题对模型预测性能的影响比较大。在第二组加入自注意力机制网络而不用 BN 层进行深度信道交换的情况下，算法模型性能有所提升，这是由于自注意力机制网络能够很好的解决原始模态特征中的数据缺失、模态干扰和维度差异性等问题。第三组在第二组的情况下加入深度信道交换后，算法模型的性能最好，这表明深度信道交换在算法模型的网络层中能够很好的处理信息失效问题，并将其他模态中的有效信息融合到信息失效的模态信道中，便于训练的同时提升算法的泛化能力，所以提高了算法模型的预测性能。

3.4.3 特征分析

为了评估多模态中的不同模态特征对短视频流行度预测得分的贡献，本小节将短视频的四个模态——视觉模态、音频模态、文本模态和社交属性模态分别表示成 V、A、T 和 S。具体地，实验中在固定其他情况下，在四个模态特征中选取其中三个然后实验。实验结果记录在表 3-2 中。

表 3-2 不同模态对短视频流行度预测的影响

	V+A+T	V+A+S	V+T+S	A+T+S	V+A+T+S
Top50	0.362	0.348	0.356	0.355	0.358
Top100	0.264	0.284	0.287	0.273	0.297
Top200	0.226	0.253	0.262	0.239	0.266
Bottom200	0.207	0.188	0.197	0.186	0.217
Bottom100	0.195	0.176	0.182	0.183	0.21
Bottom50	0.194	0.174	0.173	0.172	0.196
nMSE	0.962	0.933	0.941	0.946	**0.927**

在本次实验中，将测试集中的短视频样本根据其流行度得分的真实值从大到小进行排名，本次实验中重点关注前 50 名、前 100 名和前 200 名以及后 50 名、后 100 名和后 200 名的短视频流行度得分的预测结果。在本章中，Top50 表示短视频流行度得分真实值为前 50 名的短视频流行度得分的预测结果的平均值，Bottom50 表示短视频流行度得分真实值为后 50 名的短视频流行度得分的预测结果均值，其他表达方式与此类似。

表3-2展示了深度多模态统一特征表示模型在缺少不同模态特征的情况下的预测性能。通过对表3-2中短视频流行度得分预测结果与短视频流行度得分真实值之间进行比较与分析，可以看出不同范围的流行度预测平均得分为Top50>Top100>Top200>Bottom200>Bottom100>Bottom50，这说明预测结果是合理的。也就是说短视频流行度预测得分的趋势与短视频流行度得分的真实值的趋势保持一致，这说明预测结果是符合逻辑的。

通过对表中数据的分析，可以得到如下结论：四个模态特征在缺失社交属性这一模态特征后，模型的预测性能最差，这是说明社交属性特征在短视频流行度的预测上至关重要。这是由于社交属性模态中包含有非常重要的信息，例如关注者的数量、视频播放量等直接或间接影响短视频流行度的得分；在缺失视觉模态特征时，模型的预测性能也比较差，但是性能比缺失社交属性模态特征的性能提高了0.016，这说明视觉模态特征在预测短视频流行度预测上非常重要，但是重要程度会略低于社交属性模态特征；在缺失文本模态特征时，模型性能下降比较少，这说明文本模态信息对短视频流行度预测的影响比较小，这是由于有很多短视频样中并没有包含文本信息，同时文本信息中还存在很多不相关的信息；将所有视图的模态特征结合在一起时，获得最佳的性能，这说明利用不同模态特征表示提供的互补信息是有效的。此外，按照所有模态信息的贡献程度从大到小排列，可以得到：社会属性模态>视觉模态>音频模态>文本模态。

3.4.4 参数敏感性分析

在本章节提出的深度多模态统一特征表示算法模型中，参数θ在深度信道交换中发挥重要作用，正则化参数λ是权衡稀疏约束项的权重，在本小节评估了θ和λ的不同取值对短视频流行度得分预测性能的影响。

在进行参数选择时，首先在一个比较大的范围是寻优，然后再细分最优区间。例如在确定参数θ时，首先在[0,1]内寻优，之后经过网格搜索，确定了更为精细的范围{0.01，0.02，0.05，0.1，0.2，0.5}。

图3-6展示了在nMSE指标下分析λ对于本章节提出的模型的性能影响，从图中可以看出，nMSE随着λ的增大先减小在增大，在正则化参数λ取值为0.3时，模型性能达到最优状态，同时在λ在[0,0.4]区间里面波动比较大，这是由于正则化稀疏项对于模型的深度信道交换部分有着重要影响，这也说明深度信道交互模块对多模态统一特征表示算法模型是非常重要的。

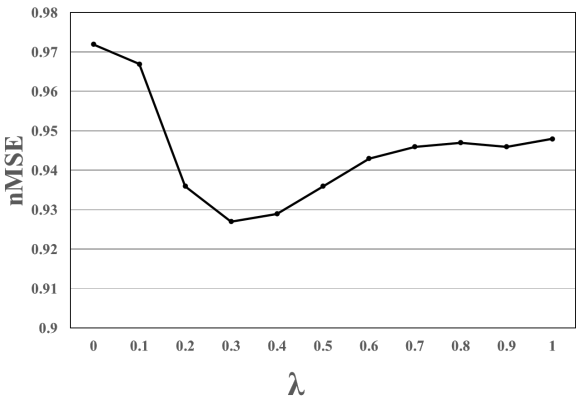

图 3-6 不同的参数λ对短视频流行度预测性能的影响

从表 3-3 中可以看出当θ取 0.05 的时候，模型取得最好的性能表现。但是可以看到，当θ的取值比较大的时候，模型的预测性能非常差，甚至比不用深度信道交换的模型性能还要差 0.014，这是因为当信道交换时，当尺度因子比较大的信道也被替换了，这就导致了大量有用的信息被抛弃了，导致了预测性能的大幅度下降。同时，θ的不同取值对模型性能的影响比较大，这也说明多模态统一特征表示算法的稳定性需要进一步加强。

表 3-3 不同参数θ对短视频流行度预测性能的影响

	0.5	0.2	0.1	0.05	0.02	0.01
Top50	0.302	0.332	0.344	0.346	0.354	0.343
Top100	0.261	0.276	0.287	0.297	0.282	0.279
Top200	0.236	0.247	0.254	0.266	0.242	0.249
Bottom200	0.192	0.186	0.196	0.217	0.186	0.191
Bottom100	0.184	0.181	0.187	0.210	0.183	0.183
Bottom50	0.178	0.176	0.181	0.196	0.181	0.176
nMSE	0.998	0.970	0.948	**0.927**	0.946	0.953

3.4.5 对比实验分析

在本小节中，将本章节提出的多模态统一特征表示算法和几种经典的短视频流行度算法进行比较。包括多视图数据的归纳多任务学习[83](Multi-task learning with multiple view data, RegMVMT)、通过分层回归[84](Multi-feature learning via hierarchical regression, MLHR)、多社交网络学习[85](Multiple social network learning, MSNL)、多视角判别分析[86](Multi-view discriminant analysis, MvDA)、多模态变分编码器[19] (Multimodal variational encoder-decoder, MMVED)、直推式多模态学习[18](Transductive multi-modal learning, TMALL)和极限学习机

[87](Extreme learning machine, ELM)。

- RegMVMT：RegMVMT 是一个归纳学习框架，用于解决一般的多视图学习问题，其中使用了协同正则化技术来强制与未标记样本上的其他视图达成一致。
- MLHR：通过分层回归进行的多特征融合是一种半监督学习方法，其目的是从多特征融合的角度探索数据中嵌入的结构信息。
- MSNL：多社交网络学习提出通过同时建模源可信度和源相关性来解决源可信度和源相关性方面的不完整数据。
- MvDA：多视图判别分析是一种多视图学习模型，通过加强多线性变换的视图来搜索潜在公共空间。
- ELM：提出了一种统一的学习机制，具有更高的可扩展性和更低的计算复杂性。
- TMALL：提出了一种用于预测短视频流行程度的多模态学习模型，该模型将不同的模态特征统一并保存在一个潜在的公共空间中，以解决信息不足的问题。
- MMVED：首先将多模态特征编码为潜在表示，并基于变分推理学习它们的概率分布，其中只有输入模态中的相关特征才能被提取到潜在表示中。然后，通过贝叶斯推理融合特定于模态的隐藏表示，以便充分利用来自所有模态的补充信息。
- TLRMVR：该模型是对 TMALL 模型的一个扩展，中对学习的短视频嵌入设置了一个新的低秩约束，使得在最终的特征表示中仅保留特征空间中的主要成分。

表 3-4 本章节提出的方法和其他方法的性能比较

Methods	nMSE
MLHR	1.167
MSNL	1.098
MvDA	0.982
ELM	0.982
TMALL	0.979
MMVED	0.941
TLRMVR	0.934
OURS	**0.927**

表 3-4 为本章提出的方法和其他算法的预测性能指标，可以得出以下结论：1、本文提出的算法模型在所有方法中表现最好。因为本章提出的算法模型有效解决了原始特征之间的数据缺失、模态干扰等现象。在多模态特征融合的时候得到更加全面的多模态统一特征表示，提高了模型的泛化能力。2、MLHR 和 MSNL 算法性能稍差一点，这是由于这些算法多模态融合的时候并没有考虑冗余的模态信息。3、与将不同模态的特征简单连接在一起的 ELM 相比，TMALL 利用多视图方法融合了受一致性约束的四种模态的异构特征。TRLMVR 通过向多视图学习目标添加隐藏空间的低秩约束来进一步改进 TMALL，从而去除特征空间的无关紧要的组件，与 TMALL 相比，这导致了实质性的改进，因此性能更好，但是整体结果仍然略低于本章所提算法，这是因为没有处理原始特征中的数据缺失、模态干扰等问题。

3.5 本章小结

本章提出了一种基于深度多模态融合的短视频流行度预测算法模型，该模型利用注意力机制网络探寻各个模态之间的关联性，不同模态之信息间的交互解决原始特征中的维度差异、数据缺失和模态干扰等问题，使得不同模态的特征的表征能力更好。同时利用算法网络层中的批量归一化层进行深度信道交换，通过批量归一化的尺度因子引导不同模态的子网络之间动态交换模态信息，融合了各模态信息，获得一个更好的多模态统一特征表示。在这个过程中，尺度因子小的信道的信息得到其他模态信息的替换，增强了模型的可训练性的同时，减少了冗余信息对预测性能的影响。通过在公开数据集上的实验结果证明了深度多模态统一特征表示算法的有效性。

第 4 章 基于双向深度编码网络的短视频流行度预测

4.1 问题描述

在针对短视频流行度的预测中，由于短视频的时间短、内容丰富和多模态的特性，多模态是指短视频的视觉模态、音频模态、文本模态和社交属性模态。这四个模态共同描述一个短视频，所以，不同模态特征在语义层面上必然有一定的相似性，同时，由于不同模态特征是从不同的角度描述短视频，导致这四个模态信息之间也有一定的独立性。因此如何利用不用模态之间的相关性和独立性是获取短视频多模态统一表示的关键。现有的多模态统一表示学习多是利用各个模态间的相关性来进行多模态的融合，往往忽略了各个模态间的独立性。基于以上考虑，需要设计一种能充分利用多模态信息之间的相关性和独立性的模型框架，来提高预测的准确性。

4.2 算法模型

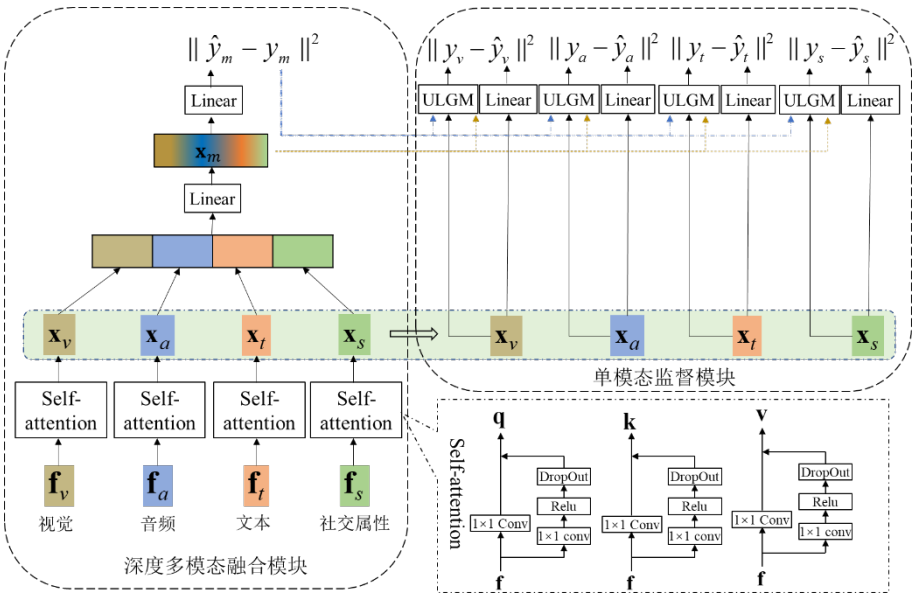

图 4-1 基于深度双向编码网络的短视频流行度预测算法模型

在本节中，将详细介绍基于双向深度编码网络的短视频流行度预测算法模型。该模型是联合训练一个多模态融合和单模态监督任务的双向深度网络。图 4-1 为本章节提出的算法模型，左侧部分为多模态融合模块，$\mathbf{f}_v$、$\mathbf{f}_a$、$\mathbf{f}_t$ 和 $\mathbf{f}_s$ 为输入，

分别代表视觉、音频、文本和社交属性模态特征，其中 self-attention 表示自注意力机制网络；右上部分为单模态监督模块，其中 ULGM 表示单模态标签生成模块。Linear 表示线性回归函数，Conv 表示卷积函数。

4.2.1 多模态融合模块

尽管短视频的多模态信息之间具有比较强的相关性，但原始特征之间存在数据缺失、特征维度不统一等问题，导致获取的原始特征并不适合直接应用到短视频流行度预测任务中。针对这一问题，采用自注意力(Self-attention)机制对原始特征进行编码以获取更有效的特征表示。具体地，首先将原始特征向量通过三个独立的自注意力机制网络，每一个特征向量在映射时综合考虑其线性映射和非线性映射，针对线性映射，使用1×1的卷积实现对特征向量的降维（或升维）和线性映射；针对非线性映射，在保留1×1的卷积的基础上还加入了 Relu 激活函数层和 Dropout 层，增加所提取特征的泛化性。降维或升维是为了特征对齐，解决特征维度不统一问题。然后将线性映射的特征向量与非线性映射的特征向量相加得到输出向量。经过自注意力机制后将得到三个向量，它们分别是查询(Query)向量、键(Key)向量和值(Value)向量，并将其简略表示为 $\mathbf{q}$ 、 $\mathbf{k}$ 、 $\mathbf{v}$ 。

由于四个模态都是对于同一个短视频的描述，所以经过提取的特征在语义内容上具有一定的相似性。针对这种情况，本文使用残差映射的方式进行特征映射，通过构建三个相互独立但是结构相同的网络模块将特征 $\mathbf{f}$ 映射为对应的查询向量 $\mathbf{q}$ 、键向量 $\mathbf{k}$ 和值向量 $\mathbf{v}$ ，然后使用点积注意力的方式去计算模态之间的相关度，通过利用当前模态的查询向量去探寻与所有模态特征的键向量的和的相关性，然后将得到的相关性得分作为各个模态特征的值向量的权重。这种方法的目的是使得短视频的四个模态在语义内容相似的部分在经过自注意力网络之后特征差异性尽可能的小。这样通过键向量 $\mathbf{k}$ 的交互能够让各个模态在语义内容上相似的部分相互补充的同时，还保持了不同模态信息的独立性。增加权重后的特征向量 $\mathrm{Attention}(\mathbf{f}_u)$ ：

$$\mathrm{Attention}(\mathbf{f}_u)=\frac{\mathbf{q}_u\mathbf{k}^T}{\sqrt{d_k}}\mathbf{v}_u, \tag{4-1}$$

其中 $u\in\{v,a,t,s\}$ ， $\mathbf{k}=(\mathbf{k}_v+\mathbf{k}_a+\mathbf{k}_t+\mathbf{k}_s)$ ， $\mathbf{q}_u$ 、 $\mathbf{v}_u$ 分别是 $\mathbf{f}_u$ 经过自注意力机制网络后的查询向量和值向量， $\mathbf{k}_v$ 、 $\mathbf{k}_a$ 、 $\mathbf{k}_t$ 和 $\mathbf{k}_s$ 分别为视觉模态特征、音频模态特征、文本模态特征和社交属性模态特征经过自注意力机制网络后的键向量， d_k 是查询向量的维度。借鉴残差网络的结构，构建由多层自注意力机制模块组成的仿残差网络模块，按照公式(4-2)规则进行更新：

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \text{Attention}\left(\mathbf{x}_n\right). \tag{4-2}$$

经过自注意力机制网络进行特征编码，最终得到具有统一维度表示的各个模态特征向量$\{\mathbf{x}_v, \mathbf{x}_a, \mathbf{x}_t, \mathbf{x}_s\} \in \mathbb{R}^D$， D是特征向量统一后的维度。

然后，将上述得到的四个模态特征向量级联起来投影到低维特征向量中，得到$\mathbf{x}_m$：

$$\mathbf{x}_m = \text{Linear}([\mathbf{x}_v; \mathbf{x}_a; \mathbf{x}_t; \mathbf{x}_s]; \Theta_1), \tag{4-3}$$

其中 Linear 表示线性回归函数，Θ_1为待学习参数。最后，将多模态融合后的特征向量$\mathbf{x}_m$用来预测短视频的流行度得分$\hat{y}_m$：

$$\hat{y}_m = \text{Linear}(\mathbf{x}_m; \Theta_2), \tag{4-4}$$

其中，Θ_2为待学习参数。

4.2.2 单模态监督模块

对于四个不同模态的监督任务，为了减少不同模态之间的维度差异和数据缺失问题，它们与多模态融合任务共享经过自注意力机制模块后的模态表示，即$\{\mathbf{x}_v, \mathbf{x}_a, \mathbf{x}_t, \mathbf{x}_s\} \in \mathbb{R}^D$，然后通过线性回归得到单模态直接预测得到结果$\hat{y}_u$：

$$\hat{y}_u = \text{Linear}(\mathbf{x}_u; \Theta_3), \tag{4-5}$$

其中$u \in \{v, a, t, s\}$， Θ_3为待学习参数。

为了指导单模态监督任务的训练过程，本文设计了一个单模态标签生成模块(Unimodal Label Generation Module, ULGM)来获取单模态的标签。关于单模态标签生成模块将在 4.2.3 小节展开详细论述。

$$y_u = \text{ULGM}(y_m, \mathbf{x}_m, \mathbf{x}_u), \tag{4-6}$$

其中$u \in \{v, a, t, s\}$， y_m为真实的短视频流行度得分，$\mathbf{x}_m$为多模态融合后的输出向量。$\mathbf{x}_u$为单模态的特征向量。

在多模态标签和单模态标签的共同监督下，联合训练了多模态融合任务和单模态监督任务。由于模型的最终输出是流行度预测得分，所以只需要多模态融合部分将四个模态融合好的特征向量进行流行度预测，得出最终的流行度预测得分。因此，单模态监督任务只存在于训练阶段，用于监督多模态特征的融合，在测试阶段只使用多模态融合模块得到流行度预测得分。

4.2.3 单模态标签生成模块

单模态标签生成模块主要用来监督多模态特征的融合，为了避免对网络参数

更新产生不必要的干扰，单模态标签生成模块设计为非参数模块。由于单模态标签与多模态标签高度相关，因此单模态标签生成模块根据模态特征表示到样本中心特征表示的距离来计算偏移量。

在训练过程中，$\mathbf{c}_i$为不同模态的中心特征表示：

$$\mathbf{c}_i = \frac{\sum_{j=1}^{N} y_i(j) \bullet \mathbf{x}_{ij}}{\sum_{j=1}^{N} y_i(j)}, \tag{4-7}$$

其中$i \in \{m,v,a,t,s\}$，N为训练样本的个数，$y_i(j)$是第i个模态下第j个样本的流行度得分，$\mathbf{x}_{ij}$表示第i个模态下第j个样本的特征表示。对于模态表示，使用$\mathcal{L}_2$范数计算$\mathbf{x}_i$和$\mathbf{c}_i$之间的距离：

$$D_i = \frac{\| \mathbf{x}_i - \mathbf{c}_i \|_2^2}{\sqrt{d_i}}, \tag{4-8}$$

其中$i \in \{m,v,a,t,s\}$，d_i是$\mathbf{x}_i$的维度。为了得到监督值和预测值之间的关系，由公式(4-9)给出：

$$\frac{y_u}{y_m} \propto \frac{\hat{y}_u}{\hat{y}_m} \propto \frac{D_u}{D_m} \Rightarrow y_u = \frac{D_u}{D_m} y_m, \tag{4-9}$$

其中$u \in \{v,a,t,s\}$，符号$\propto$表示正比于，即$\frac{y_u}{y_m}$与$\frac{D_u}{D_m}$正相关。由于模态表示的动态变化，由等式(4-9)计算得出的单模态标签不够稳定，因此，为了减小这种不利影响，本文采用一种基于动量的更新策略，将新生成值与历史值相结合。

$$y_u^{(i)} = \begin{cases} y_m & i = 1 \\ \frac{i-1}{i+1} y_u^{(i-1)} + \frac{2}{i+1} y_u^i & i > 1 \end{cases}, \tag{4-10}$$

其中$u \in \{v,a,t,s\}$，y_u^i是第i次迭代新生成的单模态标签，$y_u^{(i)}$是第i次迭代之后最终的单模态标签。

为了突出模态差异化大的样本，将多模态标签和单模态标签之间的差作为损失函数的权重，最终损失函数为：

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left((\hat{y}_m^i - y_m^i)^2 + \lambda (y_m^i - \sum_{u \in \{v,a,t,s\}} y_u^{(i)})^2 \sum_{u \in \{v,a,t,s\}} (\hat{y}_u^i - y_u^{(i)})^2 \right), \tag{4-11}$$

其中N为样本的个数，$\hat{y}_m^i$是第i个短视频样本流行度得分的预测值，y_m^i是第i个短视频样本流行度得分的真实值，$y_u^{(i)}$是第i个样本最终的单模态标签，$\hat{y}_u^i$是线性回归得到的单模态预测值，λ是平衡参数。

4.3 实验设置

4.3.1 实验数据及设置

本文中使用的短视频流行度预测数据集是由新加坡国立大学媒体实验室构建[3]。这个数据集总共包含 303242 个从在线短视频分享网站 Vine 收集的用户生成的短视频，这些视频由 98166 个用户上传。其中约 75%的视频长度为 6 至 7 秒。由于受欢迎程度与在线社交互动高度相关，因此在计算短视频的最终受欢迎程度分数时，需要考虑评论数、转发数、喜欢数和浏览循环数四类统计数据的平均值，并将其归一化，最终分数在 0 到 1 之间，对数据集做了 10 轮随机实验，在每一轮中，实验使用 90%的样本进行训练，剩下的用于测试，最终取 10 次测试的平均结果。训练和测试都是在 GPU 上完成的，显卡配置是 GeForce RTX 3090。模型框架运行环境具体如下：python3.6；pytorch1.7.1；numpy1.19.5。模型采用随机梯度下降(stochastic gradient descent, SGD)，学习率设为 0.01。

4.3.2 特征提取

4.3.2.1 视觉特征

- 颜色直方图(Color Histogram)：如文献[75]所述，由于颜色直方图可以通过显示醒目的颜色来吸引更多的注意力，因此颜色被分组为 50 种不同的颜色，本文直接使用其模型提取了 50 维的颜色直方图特征向量。
- 目标特征(Object Features)：由于深度卷积神经网络在视觉理解任务中的强大性能[76]，本文直接用训练好的 AlexNet 中 ImageNet 模型来提取 1000 维的目标特征。
- 情感特征(SentiBank Features)：Chen 等人[78]训练了一个 DeepSentiBank 的深度 CNN 模型，用于视觉情感概念的分类。本文直接用其模型提取的 2089 维的情感特征。
- 美学特征(Aesthetic Features)：在 Bhattacharya 等人[80]的视频美学评估之后，提取了包括暗通道、锐度、眼睛敏感度、低景深、白平衡、色彩统计的 149 维视觉统计特征作为美学特征。

将上述四个类型的视觉特征级联起来，形成一个 3288 维的特征向量，并作为最终的视觉模态特征表示。

4.3.2.2 音频特征

音频对于各种任务是必不可少的，声学信息可以为视觉内容提供补充线索。按照 Chen 等人[3]的设置，本文使用从音频通道中提取的 522 维特征来表示短视频的音频模态特征。

4.3.2.3 文本特征

附加的文本信息为从不同方面理解短视频内容提供了新的机会。Sentence2Vector[4] 是一种经典的文本特征提取工具，用于生成用于短视频主题表示的 100 维特征。斯坦福 CoreNLP tools[5] 工具提供了一个文本情感分析工具。通过利用情绪分析工具，每个短视频被分配一个情绪分数，该分数是 0 到 4 之间的整数，本文将两种类型的特征级联起来，形成了 101 维的向量，并作为最终的文本模态的特征表示。

4.3.2.4 社交属性特征

虽然一些相关工作已经证明低级视觉特征和高级语义特征能够在一定程度上预测流行度，但社交线索是决定短视频传播范围的重要因素。因此，本文编码了四种类型的社交线索：

- 跟随者计数：给定发布者的关注者和被关注者的数量；
- 播放计数：短视频上传后的播放次数和发布者的所有短视频总播放次数；
- 发帖数：每个发布者的总发贴数；
- 推特验证：反映发布者是否为经过验证的用户的二进制值。

将上述所有数据级联，形成一个由总播放量、当前样本的播放量、关注者数量，被关注者数量、总发帖数量和是否验证构成的 6 维的向量，并作为最终社交属性模态的特征表示。

4.3.3 评价指标

本文用归一化均方误差(normalized Mean Squared Error，nMSE)[82]来衡量预测值和真实值之间的一致性：

$$\mathrm{nMSE} = \frac{1}{M\sigma^2}\sum_{i=1}^{M}(y_i - \hat{y}_i)^2, \tag{4-12}$$

其中 M 是测试样本的个数，σ 是短视频流行度得分的真实值的标准差，y_i 是短视频流行度得分的真实值，$\hat{y}_i$ 是短视频流行度得分的预测值。

4.4 实验结果和讨论

在实验中，从以下五个方面对所提出的算法进行了验证：

- 收敛性分析：基于本文的算法模型，测试算法的收敛性。
- 模块分析：为验证该算法模型中不同模块的有效性，通过删除相关模块来比较预测性能。
- 特征分析：为评估模态特征对短视频流行度预测得分的贡献，本文考虑两种评估方法：一个是不同视觉特征之间的性能比较；另一个是不同模态之间的性能比较。
- 参数分析：在 nMSE 指标下分析 λ 对于本文提出的模型的性能影响。
- 与现有的方法比较：通过与几种主流算法的性能比较，验证了该方法的有效性和优越性。

4.4.1 收敛性分析

实验的算法模型采用 SGD 进行参数的更新，最小批设为 512。nMSE 随着迭代次数的变化如图 4-2 所示。从图中不难发现 nMSE 随着迭代次数的增加而逐渐减小，且在迭代到 40 次左右达到稳定。这证明了本文提出的算法模型能够经过训练而收敛，验证了算法的可行性。

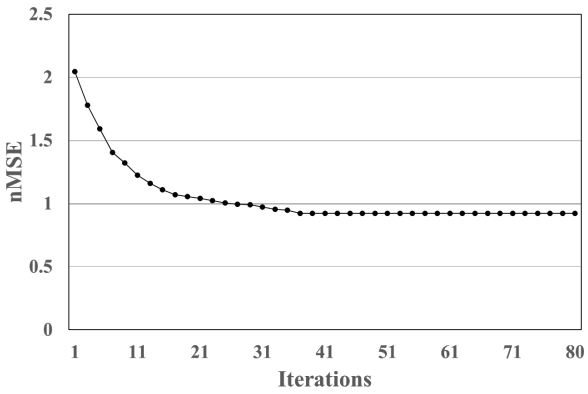

图 4-2 收敛曲线：nMSE 随模型迭代次数的变化

4.4.2 模块分析

为了验证本文提出的算法框架中每个模块的有效性，通过删除相关的模块然后比较预测性能。表 4-1 中“√”表示使用该模块，“×”表示删除该模块。

- 删除自注意力机制网络(Self-attention)：用一个简单的线性回归函数代替自注意力机制网络，不考虑模态特征的数据缺失、维度差异性等问题。
- 删除单模态监督模块(USM)：通过将 λ 设置为 0 来消除单模态监督模块的

作用。此时损失函数变为：$L=\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_m^i-y_m^i)^2$。

- 删除单模态标签生成模块(ULGM)：将生成的单模态标签与使用单模态特征预测的标签之间的损失去掉。此时的损失函数变为：$L=\frac{1}{N}\sum_{i=1}^{N}\left((\hat{y}_m^i-y_m^i)^2+\lambda(y_m^i-\sum_{u\in\{v,a,t,s\}}\hat{y}_u^i)^2\right)$。

表 4-1 框架中所涉及模块的性能比较

	第一组	第二组	第三组	第四组	第五组
Self-attention	×	√	√	×	√
USM	×	×	√	√	√
ULGM	×	×	×	√	√
nMSE	0.984	0.956	0.952	0.944	**0.921**

不同方案的比较结果如表 4-1 所示，从表中可以看出在删除自注意力机制网络、单模态监督模块或单模态标签生成模块后，nMSE 指标均有所上升，即模型性能变差。因此可以得出如下结论：在第一组的情形下，既没有考虑模态特征的数据缺失、干扰和维度差异性等问题，也没有考虑模态差异性问题，直接将四个模态特征拼接起来然后预测的性能最差。第二组在加入自注意力机制网络后，算法性能有所提升，这是由于自注意力机制能够很好解决原始模态特征中的数据缺失、模态干扰和维度差异性等问题，获得全面的特征表示。第三组在第二组的基础上加入单模态监督模块后算法性能提升很小，而第五组实验性能最好。说明简单的单模态监督模式作用有限，单模态标签生成模块非常有效，其主要是用来监督多模态的特征融合，使算法框架对于模态差异性比较大的样本有更好的泛化能力。

4.4.3 特征分析

为了评估模态特征对短视频流行度预测得分的贡献，本文考虑两种评估方法：一个是不同视觉特征之间的性能比较；另一个是不同模态之间的性能比较。

- 不同视觉特征之间的性能比较：本文中首先从组成视觉模态特征中的四个特征依次选取一个代表视觉模态特征，即分别用颜色直方图(Color)、对象特征(Object)、情感特征(SentiBank)和美学特征(Aesthetic)代替视觉模态特征，然后和剩下的三个模态特征进行融合。
- 不同模态之间的性能比较：为了验证不同模态信息对预测性能的贡献，在固定其他情况下，在四个模态特征中选取其中三个然后实验。视觉、音频、

文本和社会属性特征分别表示为“V”、“A”、“T”和“S”。

在本次实验中，将测试集中的短视频样本根据其流行度得分的真实值从大到小进行排名，本次实验中重点关注前 50 名、前 100 名和前 200 名以及后 50 名、后 100 名和后 200 名的短视频流行度得分的预测结果。在本章中，Top50 表示短视频流行度得分真实值为前 50 名的短视频流行度得分的预测结果均值，Bottom50 表示短视频流行度得分真实值为后 50 名的短视频流行度得分的预测结果均值，其他表达方式与此类似。

通过对表 4-2 中数据的分析，可以得到如下结论：用情感特征代替视觉特征的性能最好，这表明视觉情感有使短视频流行的重要信息。这符合日常现象，积极和向上的短视频更容易受到观众的追捧；对象特征和美学特征都有助于短视频流行度的预测，因为好看的对象和完美的搭配会使人愉悦，因此能够被大家喜欢；用颜色直方图代替视觉特征性能最差，用颜色直方图代替视觉特征时 nMSE 为 0.951，而从表 4-3 中可以看到，不结合视觉特征时的性能是 0.948，说明只使用颜色直方图存在严重的干扰；当所有视觉特征组合在一起时获得最佳性能，这说明利用不同的视觉特征表示提供的互补信息是有效的。

表 4-2 不同视觉特征对短视频流行度预测的影响

	Color	Object	SentiBank	Aesthetic	All
Top50	0.326	0.341	0.346	0.332	0.352
Top100	0.231	0.267	0.262	0.265	0.292
Top200	0.228	0.249	0.246	0.241	0.266
Bottom200	0.209	0.192	0.195	0.201	0.213
Bottom100	0.196	0.190	0.190	0.197	0.208
Bottom50	0.190	0.185	0.183	0.181	0.204
nMSE	0.951	0.946	0.944	0.947	**0.921**

从表 4-3 中可以得出以下结论：四个模态中缺失社会属性模态后预测性能最差，这说明在短视频的流行度预测上社交属性这一模态最重要，这是由于社会属性模态中有非常重要的信息，如关注者的数量；在缺失视觉模态后的性能下降很多，这说明视觉模态在短视频流行度预测上至关重要；在缺失文本模态的情况下，性能下降最少，这说明文本模态的信息对于短视频流行度预测的影响很小，造成这种现象的原因是由于相当数量的短视频没有文本描述，此外，文本描述还存在和短视频无关的信息；将所有视图的模态特征结合在一起时，获得最佳的性能，这说明利用不同模态特征表示提供的互补信息是有效的。此外，按照所有模态信息的贡献程度从大到小排列，可以得到：社会属性模态>视觉模态>音频模态>文

本模态。

表 4-3 不同模态对短视频流行度预测的影响

	V+A+T	V+A+S	V+T+S	A+T+S	V+A+T+S
Top50	0.389	0.340	0.352	0.365	0.352
Top100	0.273	0.288	0.297	0.271	0.292
Top200	0.223	0.261	0.264	0.245	0.266
Bottom200	0.200	0.196	0.194	0.194	0.213
Bottom100	0.205	0.192	0.189	0.180	0.208
Bottom50	0.204	0.187	0.177	0.172	0.204
nMSE	0.951	0.930	0.936	0.948	**0.921**

4.4.4 参数分析

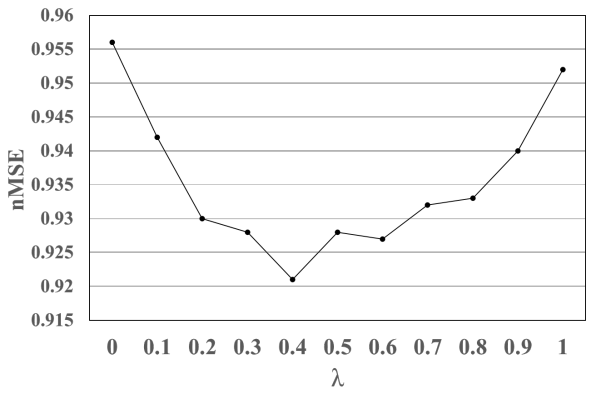

图 4-3 在 nMSE 指标下分析 λ 对于本文提出的模型的性能影响

在 nMSE 指标下分析 λ 对于本文提出的模型的性能影响。由图 4-3 可以看出当 λ 为 0 时，模型退化为删除单模态监督模块后的框架。框架性能随 λ 的变化先减小再增大，当 λ 比较小时，此时监督力度比较小，提升模型的泛化能力作用比较小，当 λ 取值为 0.4 时，模型性能达到最优，当 λ 大于 0.4 时，此时的单模态监督作用较大，使得模型过度放大模态差异性，导致模型性能下降。

4.4.5 与现有的方法比较

在本小节中，将本章节提出的多模态统一特征表示算法和几种经典的短视频流行度算法进行比较。包括多视图数据的归纳多任务学习[83](Multi-task learning with multiple view data，RegMVMT)、通过分层回归[84](Multi-feature learning via hierarchical regression，MLHR)、多社交网络学习[85](Multiple social

network learning，MSNL)、多视角判别分析[86](Multi-view discriminant analysis，MvDA)、多模态变分编码器[19](Multimodal variational encoder-decoder, MMVED)、直推式多模态学习[18](Transductive multi-modal learning，TMALL)和极限学习机[87](Extreme learning machine，ELM)。

- RegMVMT：RegMVMT 是一个归纳学习框架，用于解决一般的多视图学习问题，其中使用了协同正则化技术来强制与未标记样本上的其他视图达成一致。
- MLHR：通过分层回归进行的多特征融合是一种半监督学习方法，其目的是从多特征融合的角度探索数据中嵌入的结构信息。
- MSNL：多社交网络学习提出通过同时建模源可信度和源相关性来解决源可信度和源相关性方面的不完整数据。
- MvDA：多视图判别分析是一种多视图学习模型，通过加强多线性变换的视图相关性来搜索潜在公共空间。
- ELM：提出了一种统一的学习机制，具有更高的可扩展性和更低的计算复杂性。
- TMALL：提出了一种用于预测短视频流行程度的多模态学习模型，该模型将不同的模态特征统一并保存在一个潜在的公共空间中，以解决信息不足的问题。
- MMVED：将多模态的特征编码为潜在表示，并基于变分推理学习它们的概率分布，其中只有输入模态中的相关特征才能被提取到潜在表示中。然后，通过贝叶斯推理融合特定于模态的隐藏表示，以便充分利用来自所有模态的补充信息。
- TLRMVR：该模型是对 TMALL 模型的一个扩展，中对学习的短视频嵌入设置了一个新的低秩约束，使得在最终的特征表示中仅保留特征空间中的主要成分。

表 4-5 为本文提出的方法和其他算法的预测性能指标，可以得出以下结论：本文提出的算法模型在所有方法中表现最好，因为本文的模型不仅从特征编码和模态关联性的角度解决了多模态的融合问题，还利用单模态标签生成模块生成的单模态标签来监督多模态的融合，提高了模型对模态差异比较大的样本的泛化能力。MLHR 和 MSNL 算法性能稍差一点，这是由于这些算法仅仅考虑多模态融合的问题；与将不同模态的特征简单连接在一起的 ELM 相比，TMALL 利用多视图方法融合了受一致性约束的四种模态的异构特征；TRLMVR 通过向多视图学习目标添加隐藏空间的低秩约束来进一步改进 TMALL，从而去除特征空间的无关紧要的组件，与 TMALL 相比，这导致了实质性的改进，因此性能更好，但

是仍然比本章所提出的算法性能差一点，这说明引入单模态差异性监督对于提高模型泛化能力有很大作用。

表 4-5 本文提出的方法和其他方法的性能比较

Methods	nMSE
MLHR	1.167
MSNL	1.098
MvDA	0.982
ELM	0.982
TMALL	0.979
MMVED	0.941
TLRMVR	0.934
OURS	**0.921**

4.5 本章小结

本章提出了一个基于双向深度编码网络的短视频流行度预测模型，该模型同时考虑多模态特征融合和单模态监督的建模并将其整合为一个双向深度编码的统一框架。首先，利用自注意力机制网络内探寻模态间的关联性信息，解决原始特征中的数据缺失、维度差异明显和模态干扰等情况，提高了模态特征的表征能力。在单模态监督模块，充分利用不同模态之间的独立性监督多模态特征的融合，使模型对模态差异性较大的短视频样本也有很好的泛化能力。通过联合学习多模态融合和单模态监督任务，充分学习多模态之间的相关性和独立性。实验结果表明，本文提出的算法模型提高了短视频流行度预测的准确性。

第 5 章 总结与展望

5.1 本文工作总结

随着互联网的迅速发展智能手机的普及，短视频相关产业正在蓬勃发展。短视频流行度旨在衡量一个短视频发布后的一段时间内能够被流传的热度。由于短视频流行度预测在推荐、广告、提高网络舆情预测能力，加深用户群体行为理解等方面有重要的应用价值，越来越多的研究人员投入到短视频流行度预测的研究中。探寻短视频多模态信息之间的相关性和独立性来融合多模态特征是该方向上的研究重点。

本文针对短视频流行度预测，主要从短视频的多模态信息之间的相关性和独立性性出发，本文基于对短视频多模态特征表示和预测模型的创新，完成了以下工作：

本文的前两章主要介绍了短视频流行度预测领域的研究背景、研究现状和相关技术。通过该领域上的文献调研，分析和总结了目前该方向上各种模型算法的优缺点，提出了本文的改进方案。

第三章提出了一种基于深度多模态融合的短视频流行度预测算法模型，该模型利用注意力机制网络探寻各个模态之间的关联性，在不同模态之信息间的交互解决原始特征中的维度差异、数据缺失和模态干扰等问题，使得不同模态的特征的表征能力更好。同时利用算法网络层中的批量归一化层进行深度信道交换，通过批量归一化的尺度因子引导不同模态的子网络之间动态交换模态信息，利用多模态的融合来探寻个各模态信息之间的相关性以获得一个更好的多模态统一特征表示。在这个过程中，尺度因子小的信道的信息得到其他模态信息的替换，增强了模型的可训练性的同时，减少了冗余信息对预测性能的影响。通过在公开数据集上的实验结果证明了深度多模态统一特征表示算法的有效性。

第四章提出了一种基于双向深度编码网络的流行度预测算法模型。该算法模型同时考虑多模态特征融合和单模态监督的建模并将其整合为一个双向深度编码网络的统一框架。利用自注意力机制网络解决了原始特征中的数据缺失、维度差异明显和模态干扰等情况；在单模态监督模块，充分利用不同模态之间的独立性监督多模态特征的融合，使模型对模态差异性较大的样本也有很好的泛化能力。通过联合学习多模态融合和单模态监督任务，充分学习多模态之间的相关性和独立性。实验结果表明，本文提出的算法模型提高了短视频流行度预测的准确性。

5.2 未来工作展望

本文在短视频流行度预测的研究过程中，针对多模态信息之间的相关性和独立性问题，分别提出来基于多模态融合的短视频流行度预测算法和深度双向编码网络的短视频流行度预测算法，虽然这两种算法在预测性能上均有所提升，但是依然具有比较大的改进空间，依旧有需要进一步的研究和验证的地方：

本文虽然提出了两种短视频流行度算法，算法中能够有效的解决原始特征之间的数据缺失和模态干扰等显现，同时实现的多模态的特征融合，但是如何在利用模态相关性进行模态信息补全的同时，在不丢失各个模态信息之间的独立性是一个难题。只有保留了个各模态之间的独立性，才能更好的使用其差异性做好监督作用，使得融合的多模态统一特征表示更加全面，进一步提升算法的泛化能力。

影响短视频流行度得分的因素还有很多，比如时序性，短视频在短时间内收到大量的关注时，这就意味着在接下来的时间可能会有更强的热度，这是一种积极的响应；还有时效性，紧跟当下社会的热门话题的短视频肯定更容易被用户选择。所以在接下来的短视频流行度预测可以将时间这一特性考虑在其中。

参考文献

[1] Saura J R, Bennett D R. A three-stage method for data text mining: Using UGC in business intelligence analysis[J]. Symmetry, 2019, 11(4): 519.

[2] 周煜媛.《2020 中国网络视听发展研究报告》发布，解读行业现状及趋势[J]. 中国广播影视, 2020(21):47-51.

[3] Chen J, Song X, Nie L, et al. Micro tells macro: Predicting the popularity of micro-videos via a transductive model[C]. Proceedings of ACM International Conference on Multimedia Conference. 2016: 898-907.

[4] Redi M, OHare N, Schifanella R, et al. 6 seconds of sound and vision: Creativity in micro-videos[C]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2014: 4272-4279.

[5] Zhang J, Nie L, Wang L, et al. Shorter-is-better: Venue category estimation from micro-video[C]. Proceedings of ACM International Conference on Multimedia. 2016: 1415-1424.

[6] Wang M, Kang D. Research on semantic representation to promote the correlation of instructional micro video[C]. International Conference on Computational Intelligence and Security. 2015: 470-473.

[7] Nguyen P X, Rogez G, Fowlkes C, et al. The open world of micro-videos[J]. arXiv preprint arXiv:1603.09439, 2016.

[8] Konstan J A, Miller B N, Maltz D, et al. Grouplens: Applying collaborative filtering to usenet news[J]. Communications of the ACM, 1997, 40(3): 77-87.

[9] Wei Y, Wang X, Nie L, et al. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video[C]. Proceedings of the 27th ACM International Conference on Multimedia. 2019: 1437-1445.

[10] Ma J, Li G, Zhong M, et al. LGA: latent genre aware micro-video recommendation on social media[J]. Multimedia Tools and Applications, 2018, 77(3): 2991-3008.

[11] Liu S, Chen Z, Liu H, et al. User-video co-attention network for personalized micro-video recommendation[C]. The World Wide Web Conference. 2019: 3020-3026.

[12] Hong L, Dan O, Davison B D. Predicting popular messages in twitter[C]. Proceedings of the 20th International Conference Companion on World Wide Web. 2011: 57-58.

[13] McParlane P J, Moshfeghi Y, Jose J M. " Nobody comes here anymore, it's too crowded"; Predicting Image Popularity on Flickr[C]. Proceedings of International

Conference on Multimedia Retrieval. 2014: 385-391.

[14] Li H, Ma X, Wang F, et al. On popularity prediction of videos shared in online social networks[C]. Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. 2013: 169-178.

[15] Ma Z, Sun A, Cong G. On predicting the popularity of newly emerging hashtags in twitter[J]. Journal of the American Society for Information Science and Technology, 2013, 64(7): 1399-1410.

[16] Yamaguchi K, Berg T L, Ortiz L E. Chic or social: Visual popularity analysis in online fashion networks[C]. Proceedings of the 22nd ACM International Conference on Multimedia. 2014: 773-776.

[17] Trzciński T, Rokita P. Predicting popularity of online videos using support vector regression[J]. IEEE Transactions on Multimedia, 2017, 19(11): 2561-2570.

[18] Jing P, Su Y, Nie L, et al. Low-rank multi-view embedding learning for micro-video popularity prediction[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 30(8): 1519-1532.

[19] Xie J, Zhu Y, Zhang Z, et al. A multimodal variational encoder-decoder framework for micro-video popularity prediction[C]. Proceedings of The Web Conference. 2020: 2542-2548.

[20] Shih T K, Huang J Y, Wang C S, et al. An intelligent content-based image retrieval system based on color, shape and spatial relations[J]. Proceedings-national Science Council Republic of China Part a Physical Science and Engineering, 2001, 25(4): 232-243.

[21] Yang N C, Chang W H, Kuo C M, et al. A fast MPEG-7 dominant color extraction with new similarity measure for image retrieval[J]. Journal of Visual Communication and Image Representation, 2008, 19(2): 92-105.

[22] Jain A K, Vailaya A. Image retrieval using color and shape[J]. Pattern Recognition, 1996, 29(8): 1233-1244.

[23] Flickner M, Sawhney H, Niblack W, et al. Query by image and video content: The QBIC system[J]. Computer, 1995, 28(9): 23-32.

[24] Pass G, Zabih R. Histogram refinement for content-based image retrieval[C]. Proceedings Third IEEE Workshop on Applications of Computer Vision. 1996: 96-102.

[25] Huang J, Kumar S R, Mitra M, et al. Image indexing using color correlograms[C]. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 1997: 762-768.

[26] Manjunath B S, Ma W Y. Texture features for browsing and retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 11(6): 703-715.

[27] Grigorescu S E, Petkov N, Kruizinga P. Comparison of texture features based on

Gabor filters[J]. IEEE Transactions on Image processing, 2002, 11(10): 1160-1167.

[28] Zhang D, Lu G. Review of shape representation and description techniques[J]. Pattern Recognition, 2004, 37(1): 1-19.

[29] Boone M M, Verheijen E N G, Van Tol P F. Spatial sound-field reproduction by wave-field synthesis[J]. Journal of the Audio Engineering Society, 1995, 43(12): 1003-1012.

[30] Pearce A, Brookes T, Mason R. Timbral attributes for sound effect library searching[C]. International Conference on Semantic Audio. Audio Engineering Society, 2017.

[31] Seyerlehner K, Schedl M. Block-level audio feature for music genre classification[J]. 2009.

[32] Moffat D, Ronan D, Reiss J D. An evaluation of audio feature extraction toolboxes[J]. 2015.

[33] Peeters G. A large set of audio features for sound description (similarity and classification) in the CUIDADO project[J]. CUIDADO Ist Project Report, 2004, 54(0): 1-25.

[34] Singh V, Kumar B, Patnaik T. Feature extraction techniques for handwritten text in various scripts: a survey[J]. International Journal of Soft Computing and Engineering, 2013, 3(1): 238-241.

[35] Wang Z, Cui X, Gao L, et al. A hybrid model of sentimental entity recognition on mobile social media[J]. EURASIP Journal on Wireless Communications and Networking, 2016, 2016(1): 1-12.

[36] Trier Ø D, Jain A K, Taxt T. Feature extraction methods for character recognition-a survey[J]. Pattern Recognition, 1996, 29(4): 641-662.

[37] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12: 2493− 2537.

[38] Bordes A, Chopra S, Weston J. Question answering with subgraph embeddings[J]. arXiv preprint arXiv:1406.3676, 2014.

[39] Jean S, Cho K, Memisevic R, et al. On using very large target vocabulary for neural machine translation[J]. arXiv preprint arXiv:1412.2007, 2014.

[40] Peng Y, Qi J. CM-GANs: Cross-modal generative adversarial networks for common representation learning[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2019, 15(1): 1-24.

[41] Xu H, Chua T S. Fusion of AV features and external information sources for event detection in team sports video[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2006, 2(1): 44-67.

[42] Snoek C G M, Worring M, Smeulders A W M. Early versus late fusion in semantic video analysis[C]. Proceedings of the 13th Annual ACM International Conference

on Multimedia. 2005: 399-402.

[43] Llinas J, Hall D L. An introduction to multi-sensor data fusion[C]. Proceedings of the 1998 IEEE International Symposium on Circuits and Systems, 1998, 6: 537-540.

[44] Atrey P K, Kankanhalli M S, Oommen J B. Goal-oriented optimal subset selection of correlated multimedia streams[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2007, 3(1).

[45] Wu Z, Cai L, Meng H. Multi-level fusion of audio and visual features for speaker identification[C]. International Conference on Biometrics. 2006: 493-499.

[46] Ni J, Ma X, Xu L, et al. An image recognition method based on multiple bp neural networks fusion[C]. International Conference on Information Acquisition. 2004: 323-326.

[47] Xu Y, Lu Y. Adaptive weighted fusion: a novel fusion approach for image classification[J]. Neurocomputing, 2015, 168: 566-574.

[48] Jain A, Nandakumar K, Ross A. Score normalization in multimodal biometric systems[J]. Pattern Recognition, 2005, 38(12): 2270-2285.

[49] Somvanshi M, Chavan P, Tambade S, et al. A review of machine learning techniques using decision tree and support vector machine[C]. International Conference on Computing Communication Control and Automation. 2016: 1-7.

[50] Aczel B, Hoekstra R, Gelman A, et al. Discussion points for Bayesian inference[J]. Nature Human Behaviour, 2020, 4(6): 561-563.

[51] Wang Y, Ma F, Jin Z, et al. Eann: Event adversarial neural networks for multi-modal fake news detection[C]. Proceedings of the 24th ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. 2018: 849-857.

[52] Beltrán J, Guindel C, Cortés I, et al. Towards autonomous driving: a multi-modal 360° perception proposal[C]. International Conference on Intelligent Transportation Systems. 2020: 1-6.

[53] Hausman K, Chebotar Y, Schaal S, et al. Multi-modal imitation learning from unstructured demonstrations using generative adversarial nets[J]. arXiv preprint arXiv:1705.10479, 2017.

[54] Zhao H, Cheng W, Yang N, et al. Smartphone-based 3D indoor pedestrian positioning through multi-modal data fusion[J]. Sensors, 2019, 19(20): 4554.

[55] Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(2): 423-443.

[56] Hazarika D, Zimmermann R, Poria S. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis[C]. Proceedings of the 28th ACM International Conference on Multimedia. 2020: 1122-1131.

[57]井佩光. 基于多特征表征学习的多媒体数据预测方法研究[D].天津大学,2017.

[58]Jiang Y G, Wu Z, Wang J, et al. Exploiting feature and class relationships in video categorization with regularized deep neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(2): 352-364.

[59]Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning[C]. International Conference on Machine Learning. 2011.

[60]Aytar Y, Castrejon L, Vondrick C, et al. Cross-modal scene networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(10): 2303-2314.

[61]Peng Y, Qi J, Yuan Y. Modality-specific cross-modal similarity measurement with recurrent attention network[J]. IEEE Transactions on Image Processing, 2018, 27(11): 5585-5599.

[62]He Y, Xiang S, Kang C, et al. Cross-modal retrieval via deep and bidirectional representation learning[J]. IEEE Transactions on Multimedia, 2016, 18(7): 1363-1377.

[63]Rasiwasia N, Costa Pereira J, Coviello E, et al. A new approach to cross-modal multimedia retrieval[C]. Proceedings of the 18th ACM International Conference on Multimedia. 2010: 251-260.

[64]Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]. International Conference on Machine Learning, 2015: 2048-2057.

[65]Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2625-2634.

[66]Reed S, Akata Z, Yan X, et al. Generative adversarial text to image synthesis[C]. International Conference on Machine Learning, 2016: 1060-1069.

[67]Hori C, Hori T, Lee T Y, et al. Attention-based multimodal fusion for video description[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 4193-4202.

[68]Lu J, Xiong C, Parikh D, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 375-383.

[69]Fan D P, Wang W, Cheng M M, et al. Shifting more attention to video salient object detection[C]. Proceedings of the Conference on Computer Vision and Pattern Recognition. 2019: 8554-8564.

[70]Chu Q, Ouyang W, Li H, et al. Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism[C]. Proceedings of

the IEEE International Conference on Computer Vision. 2017: 4836-4845.

[71] Sharma S, Kiros R, Salakhutdinov R. Action recognition using visual attention[J]. arXiv preprint arXiv:1511.04119, 2015.

[72] Girdhar R, Ramanan D. Attentional pooling for action recognition[J]. arXiv preprint arXiv:1711.01467, 2017.

[73] Yao L, Torabi A, Cho K, et al. Video description generation incorporating spatio-temporal features and a soft-attention mechanism[J]. arXiv preprint arXiv:1502.08029, 2015.

[74] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]. International Conference on Machine Learning. 2015: 448-456.

[75] Khosla A, Das Sarma A, Hamid R. What makes an image popular?[C]. Proceedings of the 23rd International Conference on World Wide Web. 2014: 867-876.

[76] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25: 1097-1105.

[77] Gelli F, Uricchio T, Bertini M, et al. Image popularity prediction in social media using sentiment and context features[C]. Proceedings of the 23rd ACM International Conference on Multimedia. 2015: 907-910.

[78] Chen T, Borth D, Darrell T, et al. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks[J]. arXiv preprint arXiv:1410.8586, 2014.

[79] Zhang L, Song M, Yang Y, et al. Weakly supervised photo cropping[J]. IEEE Transactions on Multimedia, 2013, 16(1): 94-107.

[80] Bhattacharya S, Nojavanasghari B, Chen T, et al. Towards a comprehensive computational model foraesthetic assessment of videos[C]. Proceedings of the 21st ACM International Conference on Multimedia. 2013: 361-364.

[81] Mishne G, Glance N S. Predicting movie sales from blogger sentiment[C]. AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. 2006: 155-158.

[82] Nie L, Zhang L, Yang Y, et al. Beyond doctors: Future health prediction from multimedia and multimodal observations[C]. In Proceedings of ACM International Conference on Multimedia. 2015:591-600.

[83] Zhang J, Huan J. Inductive multi-task learning with multiple view data[C]. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2012: 543-551.

[84] Yang Y, Song J, Huang Z, et al. Multi-feature fusion via hierarchical regression for multimedia analysis[J]. IEEE Transactions on Multimedia, 2012, 15(3):572-581.

[85] Song X, Nie L, Zhang L, et al. Multiple social network learning and its application in volunteerism tendency prediction[C]. Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval. 2015:213-222.

[86] Kan M, Shan S, Zhang H, et al. Multi-view discriminant analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(1):188-94.

[87] Huang G B, Zhou H, Ding X, et al. Extreme learning machine for regression and multiclass classification[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 2011, 42(2):513-29.