第 59 卷 第 8 期/2022 年 4 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

# 基于双向深度编码网络的短视频流行度预测

井佩光[1]，叶徐清[1*]，刘昱[2]，苏育挺[1]

[1]天津大学电气自动化与信息工程学院，天津 300072；

[2]天津大学微电子学院，天津 300072

**摘要** 针对短视频流行度预测问题，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关联性解决原始特征之间的数据缺失和维度差异等问题，以获取更全面的特征表示。单模态监督模块利用模态差异性监督多模态特征融合。通过联合训练多模态融合和单模态监督任务，充分学习多模态信息的一致性和差异性以提高算法的泛化能力。在公开NUS数据集上的实验表明所提模型的有效性和优越性。

**关键词** 成像系统；短视频；模态关联性；特征表示；多模态融合；流行度预测

**中图分类号** TP391.4 **文献标志码** A **doi**：10.3788/LOP202259.0811009

# Micro-Video Popularity Prediction with Bidirectional Deep Encoding Network

Jing Peiguang[1], Ye Xuqing[1*], Liu Yu[2], Su Yuting[1]

[1]*School of Electrical and Information Engineering, Tianjin University, Tianjin* 300072, *China*;

[2]*School of Microelectronics, Tianjin University, Tianjin* 300072, *China*

**Abstract** Aiming at the micro-video popularity prediction, we propose a micro-video popularity prediction model with a bidirectional deep encoding network. The model considers both multi-modal fusion and unimodal supervision modeling, and integrates them into a bidirectional deep encoding network. The multi-modal fusion module uses modal relevance to solve problems such as data missing and dimensional differences among original features to obtain a more comprehensive feature representation. The unimodal supervision module uses modal differences to supervise multi-modal feature fusion. Via joint training of multi-modal fusion and unimodal supervision tasks, the consistency and difference of multi-modal information are fully learned to improve the generalization ability of the algorithm. The experiments on the public NUS dataset have proved the effectiveness and superiority of our proposed algorithm.

**Key words** imaging systems; micro-video; modal relevance; feature representation; multi-modal fusion; popularity prediction

## 1 引 言

近年来，短视频已经成为用户生成内容(UGCs)[1]的一种新趋势，并在各种社交平台上广泛传播。《2020年中国网络视听发展研究报告》指出，在抖音、快手等短视频平台的加速渗透下，我国短

**收稿日期**：2021-08-09；**修回日期**：2021-08-31；**录用日期**：2021-09-10

**基金项目**：国家自然科学基金(61802277)、天津市自然科学基金(20JCQNJC01210)、博士后科学基金(2019M651038)、天津市科技重大专项与工程(18ZXJMTG00020)

**通信作者**：*yxq@tju.edu.cn

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网络平台上，但只有很少一部分短视频因大量用户的观看、喜欢、评论和转载被广泛传播而流行起来，大多数短视频很快就会被遗忘。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。

短视频可以由视觉、音频、文本和社会属性等多个模态特征描述，不同类型的模态特征之间存在异构性，同时也有很强的关联性和互补性。因此，在充分挖掘单模态特征表示的同时，越来越多研究者开始从多模态关联性分析的角度探究多媒体数据的特征表示和学习问题。Jiang等[2]提出结合视频的音频信息、低层视觉信息和高层语义信息，通过核映射的方式实现多层特征的融合。Wu等[3]提出一种基于多流多分类的深度网络框架，该框架构建了三种卷积神经网络，针对视频的空间信息、短期运动信息和音频信息进行建模，自动挖掘类间关联性信息并将其作为先验，指导多流多分类模型进行决策分析。杨晓莉等[4]提出一种基于对抗生成网络的多模态图像融合方法，通过网络的自适应学习生成融合图像。Jiang等[5]提出利用三种卷积神经网络提取视频在外观、运动和音频方面的特征，并利用一个特征融合网络获取视频的多种信息的共同表征。Hazarika等[6]针对多模态情感分析中的模态异构问题，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。刘天宝等[7]通过引入注意力机制，为每个视频帧所包含的情感信息分配权重，最后利用加权决策融合方法融合表情和语音信号。陈湟康等[8]提出一种基于深度门的多模态长短记忆网络，利用深度门连接记忆储存单元的上下层来增强上下层之间的关系，同时学习每一层模型之间的联系。Zhang等[9]提出了一种新的多任务多模态算法来解决短视频的场地类别估计问题。Nguyen等[10]构建了一个开放的短视频数据库，用于支撑短视频的分类及标注等一系列问题的研究。Chen等[11]提出一种基于直推式多模态学习的算法以解决短视频的语义流行度预测。Jing等[12]提出一种基于低秩多视角嵌入的直推式的学习方法以解决短视频流行度预测问题。Xie等[13]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，以解决短视频的流行度预测问题。上述方法大多利用多模态之间的一致性和互补性来进行多模态特征融合，对于不同模态之间的差异性考虑得不够充分。

针对上述问题，本文提出基于双向深度编码的短视频流行度预测算法，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。多模态特征融合模块设计了一个自注意力机制网络，充分解决了不同模态特征之间的数据缺失、维度差异明显和模态干扰等情况，同时还引入不同模态之间的交互信息以获取更全面的特征表示。单模态监督利用生成的单模态标签对重要的模态信息的学习进行强化监督，联合学习多模态和单模态双向任务，充分学习多模态信息之间的一致性和差异性。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。

## 2 算法模型

本节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任务和单模态监督任务的双向深度网络。图1为本文提出的算法模型。图1左侧部分为多模态融合模块，$\boldsymbol{f}_{\mathrm{v}}$、$\boldsymbol{f}_{\mathrm{a}}$、$\boldsymbol{f}_{\mathrm{t}}$和$\boldsymbol{f}_{\mathrm{s}}$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；图1右上部分为单模态监督模块，其中单模态标签生成模块（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。

### 2.1 多模态关联性学习

尽管短视频的多模态信息之间具有比较强的一致性和互补性，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，首先使原始特征向量通过三个独立的自注意力机制网络，在进行每一个特征向量的映射时，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卷积的基础上，加入ReLu激活函数层和Dropout层，以提高所提取特征的泛化性。降维或升维是为了使特征

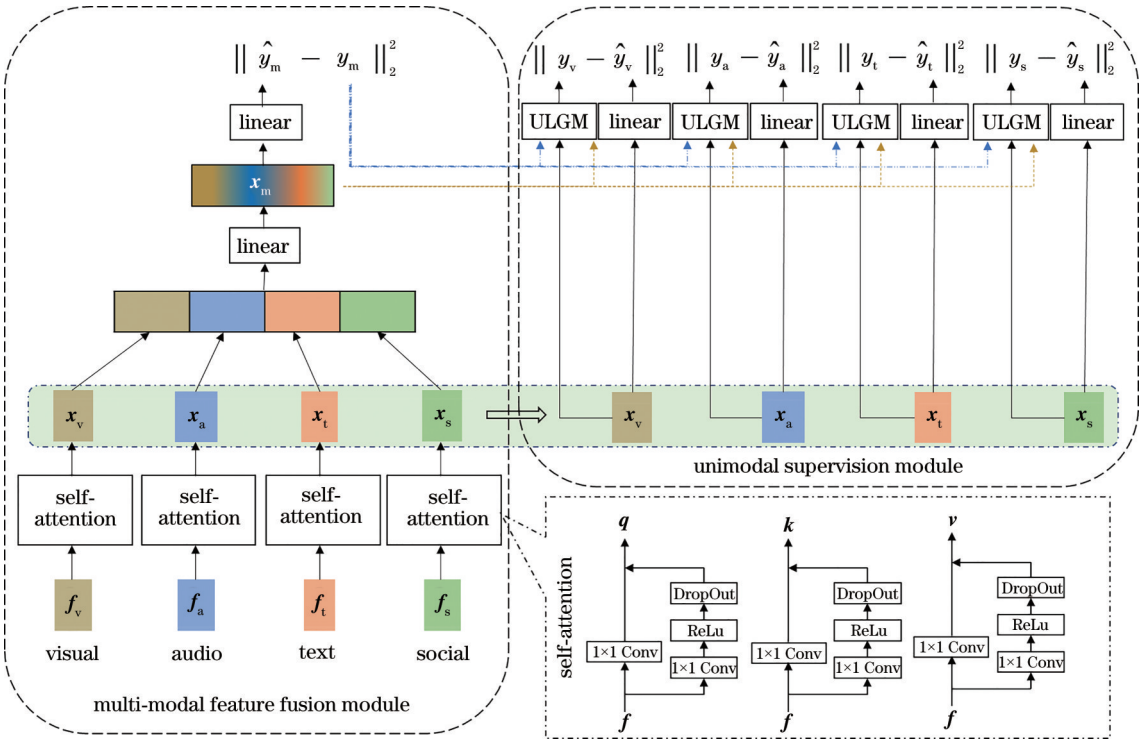


图 1　本文模型示意图

Fig. 1　Illustration of proposed model

对齐，从而解决特征维度不统一的问题。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。原始特征向量经过自注意力机制网络后，得到三个向量，它们分别是查询向量、键向量和值向量，并将其分别简略表示为 $\boldsymbol{q}$、$\boldsymbol{k}$、$\boldsymbol{v}$。

由于4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结构相同的网络模块，将特征 $\boldsymbol{f}$ 映射为对应的查询向量 $\boldsymbol{q}$、键向量 $\boldsymbol{k}$ 和值向量 $\boldsymbol{v}$。然后使用点积注意力的方式计算模态之间的相关度，利用当前模态的查询向量探寻与所有模态特征的键向量和的相关性，将得到的相关性得分作为各个模态特征的值向量的权重。该方法的目的是使得短视频的4个模态在语义内容相似的部分经过自注意力网络之后的特征差异性尽可能小。这样通过键向量 $\boldsymbol{k}$ 的交互能够让各个模态在语义内容上的相似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 Attention($\boldsymbol{f}_u$)为

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

式中：$u\in\{\text{v, a, t, s}\}$；$\boldsymbol{k}=\boldsymbol{k}_{\text{v}}+\boldsymbol{k}_{\text{a}}+\boldsymbol{k}_{\text{t}}+\boldsymbol{k}_{\text{s}}$；$\boldsymbol{q}_u$、$\boldsymbol{v}_u$ 分别是 $\boldsymbol{f}_u$ 经过自注意力机制网络后的查询向量和值向量；$\boldsymbol{k}_{\text{v}}$、$\boldsymbol{k}_{\text{a}}$、$\boldsymbol{k}_{\text{t}}$ 和 $\boldsymbol{k}_{\text{s}}$ 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，进行更新：

$$\boldsymbol{x}_{n+1}=\boldsymbol{x}_n+\text{Attention}(\boldsymbol{x}_n)。 \tag{2}$$

本次实验使用8层自注意力机制网络进行特征编码，最终得到具有由统一维度表示的各个模态特征向量，即 $\{\boldsymbol{x}_{\text{v}},\boldsymbol{x}_{\text{a}},\boldsymbol{x}_{\text{t}},\boldsymbol{x}_{\text{s}}\}\in\mathbb{R}^{d_u}$，$d_u$ 是特征向量统一后的维度。

将上述得到的4个模态特征向量级联起来并将其投影到低维特征向量中，得到 $\boldsymbol{x}_{\text{m}}$：

$$\boldsymbol{x}_{\text{m}}=\text{linear}([\boldsymbol{x}_{\text{v}};\boldsymbol{x}_{\text{a}};\boldsymbol{x}_{\text{t}};\boldsymbol{x}_{\text{s}}];\Theta_1), \tag{3}$$

式中：$\Theta_1$ 为学习参数。最后用多模态融合后的特征向量 $\boldsymbol{x}_{\text{m}}$ 预测短视频的流行度得分 $\hat{y}_{\text{m}}$：

$$\hat{y}_{\text{m}}=\text{linear}(\boldsymbol{x}_{\text{m}};\Theta_2), \tag{4}$$

式中：$\Theta_2$ 为学习参数。

## 2.2　单模态监督模块

对于4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，将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$\{\boldsymbol{x}_{\mathrm{v}},\boldsymbol{x}_{\mathrm{a}},\boldsymbol{x}_{\mathrm{t}},\boldsymbol{x}_{\mathrm{s}}\}\in\mathbb{R}^{d_u}$，然后通过线性回归得到单模态直接预测结果$\hat{y}_u$：

$$\hat{y}_u=\mathrm{linear}(\boldsymbol{x}_u;\Theta_3),\tag{5}$$

式中：$\Theta_3$为学习参数。

为了指导单模态监督任务的训练过程，本文设计了ULGM来获取单模态的标签。ULGM将在2.3节详细展开论述。

$$y_u=\mathrm{ULGM}(y_{\mathrm{m}},\boldsymbol{x}_{\mathrm{m}},\boldsymbol{x}_u),\tag{6}$$

式中：$y_u$表示4种模态中任一模态经过ULGM后生成的单模态标签结果；$y_{\mathrm{m}}$为真实的短视频流行度得分；$\boldsymbol{x}_{\mathrm{m}}$为多模态融合后的输出向量；$\boldsymbol{x}_u$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4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。

### 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，将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，不同模态的中心特征$\boldsymbol{c}_i$可表示为

$$\boldsymbol{c}_i=\frac{\sum_{j=1}^{N}y_i(j)\cdot\boldsymbol{x}_{ij}}{\sum_{j=1}^{N}y_i(j)},\tag{7}$$

式中：$i\in\{\mathrm{m},\mathrm{v},\mathrm{a},\mathrm{t},\mathrm{s}\}$；$N$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；$\boldsymbol{x}_{ij}$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L2范数计算$\boldsymbol{x}_i$和$\boldsymbol{c}_i$之间的距离：

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

式中：$d_i$是$\boldsymbol{x}_i$的维度。监督值和预测值之间的关系为

$$\frac{y_u}{y_{\mathrm{m}}}\propto\frac{\hat{y}_u}{\hat{y}_{\mathrm{m}}}\propto\frac{D_u}{D_{\mathrm{m}}}\Rightarrow y_u=\frac{D_u}{D_{\mathrm{m}}}y_{\mathrm{m}},\tag{9}$$

式中：符号$\propto$表示正比于，即$\frac{y_u}{y_{\mathrm{m}}}$与$\frac{D_u}{D_{\mathrm{m}}}$正相关。由于模态表示的动态变化，由(9)式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：

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

式中：$y_u^i$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。

为了突出模态差异大的样本，将多模态标签和单模态标签的差作为损失函数的权重，最终损失函数为

$$L=\frac{1}{N}\sum_{i=1}^{N}\left[(\hat{y}_{\mathrm{m}}^i-y_{\mathrm{m}}^i)^2+\lambda(y_{\mathrm{m}}^i-\sum_{u\in\{\mathrm{v},\mathrm{a},\mathrm{t},\mathrm{s}\}}y_u^{(i)})^2\sum_{u\in\{\mathrm{v},\mathrm{a},\mathrm{t},\mathrm{s}\}}(\hat{y}_u^i-y_u^{(i)})^2\right],\tag{11}$$

式中：$N$为样本的个数；$\hat{y}_{\mathrm{m}}^i$是第$i$个短视频样本流行度得分的预测值；$y_{\mathrm{m}}^i$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；$\lambda$是平衡参数。

## 3　实验和结果分析

### 3.1　实验数据及设置

本文使用的短视频流行度预测数据集(参考网址 http://acmmm2016.wixsite.com/micro-videos)是由新加坡国立大学多媒体实验室构建。这个数据集共包含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。所有短视频的长度都不超过8 s，其中约75%的视频长度为6～7 s。由于短视频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4类统计数据的平均值，并将其进行归一化，使其最终分数在0到1之间。本文实验只下载到其中的9720条数据，对数据集进行了10轮随机实验，每一轮实验用90%的样本进行训练，剩下的样本用于测试，最终得到10次测试的平均结果。训练和测试都是在GPU上完成，显卡配置是GeForce RTX 3090。模型框架的运行环境具体如下：python为3.6；pytorch为1.7.1；numpy为1.19.5。模型采用随机梯度下降(SGD)算法，学习率设为0.01。

### 3.2 特征提取

#### 3.2.1 视觉特征

1) 颜色直方图(Color)。如文献[14]中所述,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,本文直接使用所提模型提取了50维的颜色直方图特征向量。

2) 目标特征(Object)。深度卷积神经网络在视觉理解任务中具有强大性能[15],本文直接用训练好的ImageNet中AlexNet模型来提取1000维的目标特征。

3) 情感特征(SentiBank)。Chen等[16]训练了一个DeepSentiBank的深度CNN模型,用于视觉情感概念的分类。本文直接用该模型提取2089维的情感特征。

4) 美学特征(Aesthetic)。在Bhattacharya等[17]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4个类型的视觉特征级联起来,形成一个3288维的特征向量,并将其作为最终的视觉模态特征表示。

#### 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[11]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。

#### 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(参考网址https://github.com/klb3713/sentence2vec)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(参考网址http://stanfordnlp.github.io/CoreNLP/)工具提供了一个文本情感分析工具。利用情绪分析工具为每个短视频分配一个情绪分数,该分数是0~4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,并将其作为最终文本模态的特征表示。

#### 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4种类型的社交线索。

1) 跟随者计数:给定发布者的关注者和被关注者的数量。

2) 播放计数:短视频上传后的播放次数和发布者发布的所有短视频的总播放次数。

3) 发帖数:每个发布者的总发帖数。

4) 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,并将其作为最终社交属性模态的特征表示。

本文对数据集进行了统计。分析播放量(NP)、关注量(NF)和总发帖数(TNV)与真实流行度得分(TPS)之间的联系。本文实验样本的发布者都没有经过验证,所以未对其进行统计。

由表1可以看出,在播放量和关注量的统计上,短视频流行度得分(Top50、Top100、Top200、Bottom200、Bottom100、Bottom50的得分分别为$S_{\mathrm{Top50}}$、$S_{\mathrm{Top100}}$、$S_{\mathrm{Top200}}$、$S_{\mathrm{Bottom200}}$、$S_{\mathrm{Bottom100}}$和$S_{\mathrm{Bottom50}}$)满足$S_{\mathrm{Top50}}>S_{\mathrm{Top100}}>S_{\mathrm{Top200}}>S_{\mathrm{Bottom200}}>S_{\mathrm{Bottom100}}\geqslant S_{\mathrm{Bottom50}}$。这说明播放量越高,流行度得分越高;关注者越多,流行度的得分越高。表中关注者Top50的短视频流行度平均得分为0.344,大于播放量Top50的短视频流行度平均得分0.312,这是由于:关注者一般都是短视频发布者的忠实粉丝,他们在贡献播放量的同时,还会贡献评论量、转发量等,从而提高短视频流行得分。对于总发帖数而言,并不是发帖越多的用户的短视频得分越高,但是发帖数很少的作者发

表1 社交线索对流行度得分的影响

Table 1 Impact of social cues on popularity scores

| Parameter | Top50 | Top100 | Top200 | Bottom200 | Bottom100 | Bottom50 |
|---|---|---|---|---|---|---|
| NP | 0.312 | 0.261 | 0.222 | 0.152 | 0.151 | 0.150 |
| NF | 0.344 | 0.260 | 0.224 | 0.151 | 0.150 | 0.150 |
| TNV | 0.256 | 0.272 | 0.222 | 0.151 | 0.150 | 0.150 |
| TPS | 0.402 | 0.295 | 0.226 | 0.151 | 0.150 | 0.150 |

布的视频流行的得分都很低。从Top200这一列可以看出，流行度得分Top200的短视频都是关注者多、播放量高的短视频，所以关注者数量和播放量直接影响短视频流行度得分。

### 3.3　评价指标

本文用归一化均方误差(nMSE)[18]来衡量预测值和真实值之间的一致性：

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

式中：$M$是测试样本的个数；$\sigma$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。nMSE的值越小，说明所提算法的性能越好。

### 3.4　结果和讨论

实验中从以下5个方面对所提算法进行了验证：

1) 收敛性分析。基于本文的算法模型，测试算法的收敛性。

2) 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。

3) 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。

4) 参数分析。在nMSE指标下分析$\lambda$对于本文所提模型的性能影响。

5) 与现有的方法进行比较。通过将本文方法与几种主流算法进行比较，验证了所提方法的有效性。

#### 3.4.1　收敛性分析

实验的算法模型采用SGD进行参数的更新，最小批大小设为512。nMSE随迭代次数的变化如图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训练而收敛，从而验证了算法的可行性。

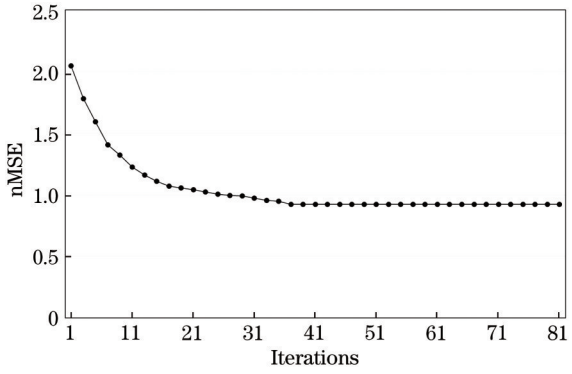


图2　nMSE随模型迭代次数的变化

Fig. 2　nMSE changes with model iterations

#### 3.4.2　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，删除相关的模块，对预测性能进行分析。表2中“√”表示使用该模块，“×”表示删除该模块。进行三种删除操作：1) 删除自注意力机制网络，用一个简单的线性回归函数代替自注意力机制网络，不考虑模态特征的数据缺失、干扰和维度差异性等问题；2) 删除单模态监督模块(USM)，通过将$\lambda$设置为0来消除单模态监督模块的作用；3) 删除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。

表2　框架中所涉及模块的性能比较

Table 2　Performance comparison of modules involved in framework

| Experiment | Experiment 1 | Experiment 2 | Experiment 3 | Experiment 4 | Experiment 5 |
|---|---|---|---|---|---|
| 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2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：1) 第一组实验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4个模态特征拼接起来，其预测性能最差；2) 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地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；3) 在第二组的基础上加入单模态监督模块，此时算法性能提升不明显；4) 第五组实验的性

能最好。这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。

### 3.4.3　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：1）不同视觉特征之间的性能比较；2）不同模态之间的性能比较。对于不同视觉特征之间的性能比较，本文首先从组成视觉模态特征的4个特征中依次选取一个代表视觉模态特征，即分别用颜色直方图（Color）、目标特征（Object）、情感特征（SentiBank）和美学特征（Aesthetic）代替视觉模态特征。对于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4个模态特征中选取其中3个进行实验。视觉、音频、文本和社会属性特征分别表示为V、A、T和S。

表3显示了不同方案的预测结果，分别选取了流行度得分前50、100、200和后50、100、200的样本，然后根据其预测得分算出了平均得分结果。如表3所示，不同范围的预测流行度平均得分为$S_{\mathrm{Top50}}>S_{\mathrm{Top100}}>S_{\mathrm{Top200}}>S_{\mathrm{Bottom200}}>S_{\mathrm{Bottom100}}>S_{\mathrm{Bottom50}}$，这说明预测结果是合理的。因此，从表3中可以得出以下结论：1）用情感特征代替视觉特征的算法性能最好，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；2）目标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；3）用颜色直方图代替视觉特征的算法性能最差，用颜色直方图代替视觉特征时的nMSE为0.951，而从表4中可以看到，不结合视觉特征时的nMSE是0.948，这说明只使用颜色直方图存在严重的干扰；4）将所有视觉特征组合在一起能获得最佳的算法性能，这说明不同的视觉特征表示提供的互补信息是有效的。

表3　不同视觉特征之间的性能比较

Table 3　Performance comparison among different visual features

| Visual feature | 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** |

表4　不同模态特征之间的性能比较

Table 4　Performance comparison among different modal characteristics

| Visual feature | 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** |

从表4中可以得出以下结论：1）4个模态中缺失社会属性模态后的算法预测性能最差，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；2）在缺失视觉模态后算法性能下降很多，这说明视觉模态在短视频流行度预测上至关重要；3）在缺失文本模态的情况下，算法性能下降最少，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相当数量的短视频没有文本描述，且文本描述还存在和短视频无关的信息；4）将所有视图的模态特征结合在一起时，获得最佳的算法性能，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，所有模态信息的贡献程度从大到小依次为：社会属性模态、视觉模态、音频模态和文本模态。

### 3.4.4　参数分析

分析$\lambda$对本文所提模型性能的影响。由图3可以看出：当$\lambda$为0时，模型退化为删除单模态监督模

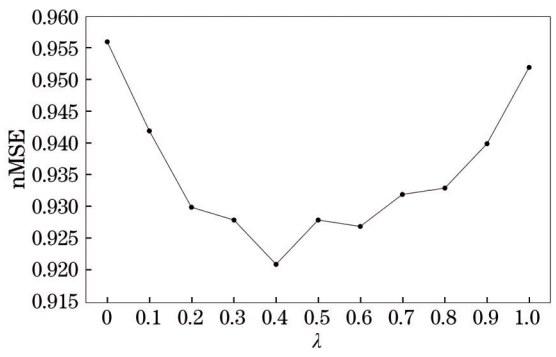


图3　$\lambda$对所提模型性能的影响

Fig. 3　Influence of $\lambda$ on performance of proposed model

块后的框架。框架性能随λ的变化先提升，之后弱化。当λ比较小时，框架监督力度比较小，提升模型泛化能力的作用比较小；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，模型过度放大模态差异性，导致模型性能下降。

3.4.5 与现有方法的比较

将本文提出的算法框架与现有方法进行了比较，对比算法包括通过层次回归的多特征学习(MLHR)[19]。该模型是从多特征融合的角度探索数据中嵌入的结构信息。多社交网络学习(MSNL)[20]通过同时建模源可信度和源一致性来解决源可信度和源一致性方面的不完整数据。多视图判别分析(MvDA)[21]是一种多视图学习模型，通过加强多线性变换的视图一致性来搜索潜在公共空间。直推式多模态学习(TMALL)[11]是一种用于预测微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。极限学习机(ELM)[22]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。基于低秩多视点嵌入学习(TLRMVR)[12]在模型中对学习的短视频设置了一个新的低秩约束，使得模型在最终的特征表示中仅保留特征空间中的主要成分。以上方法多是从多模态信息的互补性出发，解决单个模态信息不足的问题，以获取一个公共表示。

表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：

1) 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。

表5 本文提出的方法和其他方法的性能比较

Table 5 Performance comparison between our proposed method and several methods

| Method | nMSE |
|---|---|
| MLHR | 1.167 |
| MSNL | 1.098 |
| MvDA | 0.982 |
| ELM | 0.982 |
| TMALL | 0.979 |
| TLRMVR | 0.934 |
| Ours | **0.921** |

2) MLHR和MSNL算法的性能稍差一点，这是由于这些算法仅考虑多模态融合的问题。

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4种模态的异构特征。TLRMVR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，TLRMVR性能更好。

## 4 结 论

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，并将其整合为一个多任务学习的统一框架。多模态融合模块利用模态关联性解决原始特征之间的数据缺失和维度差异等问题，以获取更全面的特征表示；单模态监督模块充分利用不同模态之间的差异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一致性和差异性。实验结果表明，所提算法提高了短视频流行度预测的准确性。

## 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] Jiang Y G, Xu B, Xue X. Predicting emotions in user-generated videos[C]//Proceedings of AAAI Conference on Artificial Intelligence, July 27-31, 2014, Quebec City, Canada. Reston: AAAI Press, 2014: 73-79.

[3] Wu Z X, Jiang Y G, Wang X, et al. Multi-stream multi-class fusion of deep networks for video classification[C]//Proceedings of the 24th ACM international conference on Multimedia, Amsterdam The Netherlands. New York: ACM, 2016: 791-800.

[4] Yang X L, Lin S Z, Lu X F, et al. Multimodal image fusion based on generative adversarial networks [J]. Laser & Optoelectronics Progress, 2019, 56(16): 161004.
杨晓莉, 蔺素珍, 禄晓飞, 等. 基于生成对抗网络的多模态图像融合[J]. 激光与光电子学进展, 2019, 56(16): 161004.

[5] Jiang Y G, Wu Z X, Tang J H, et al. Modeling multimodal clues in a hybrid deep learning framework for video classification[J]. IEEE Transactions on

Multimedia, 2018, 20(11): 3137-3147.

[6] Hazarika D, Zimmermann R, Poria S. MISA: modality-invariant and -specific representations for multimodal sentiment analysis[C]//MM'20: Proceedings of the 28th ACM International Conference on Multimedia, October 12-16, 2020, Seattle, WA, USA. New York: ACM, 2020: 1122-1131.

[7] Liu T B, Zhang L T, Yu W T, et al. Hierarchical LSTM-based audio and video emotion recognition with embedded attention mechanism[J]. Laser & Optoelectronics Progress, 2021, 58(2): 0210017.
刘天宝, 张凌涛, 于文涛, 等. 基于嵌入注意力机制层级LSTM的音视频情感识别[J]. 激光与光电子学进展, 2021, 58(2): 0210017.

[8] Chen H K, Chen Y. Speaker identification based on multimodal long short-term memory with depth-gate [J]. Laser & Optoelectronics Progress, 2019, 56(3): 031007.
陈湟康, 陈莹. 基于具有深度门的多模态长短期记忆网络的说话人识别[J]. 激光与光电子学进展, 2019, 56(3): 031007.

[9] Zhang J L, Nie L Q, Wang X, et al. Shorter-is-better: venue category estimation from micro-video [C]//Proceedings of the 24th ACM International Conference on Multimedia, October 15-19, 2016, Amsterdam, The Netherlands. New York: ACM, 2016: 1415-1424.

[10] Nguyen P X, Rogez G, Fowlkes C, et al. The open world of micro-videos[EB/OL]. (2016-03-31)[2021-04-09]. https://arxiv.org/abs/1603.09439.

[11] Chen J Y, Song X M, Nie L Q, et al. Micro tells macro: predicting the popularity of micro-videos via a transductive model[C]//Proceedings of the 24th ACM International Conference on Multimedia, October 15-19, 2016, Amsterdam, The Netherlands. New York: ACM, 2016: 898-907.

[12] Jing P G, Su Y T, Nie L Q, et al. Low-rank multi-view embedding learning for micro-video popularity prediction[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(8): 1519-1532.

[13] Xie J Y, Zhu Y C, Zhang Z B, et al. A multimodal variational encoder-decoder framework for micro-video popularity prediction[C]//Proceedings of the Web Conference 2020, April 20-24, Taipei, Taiwan, China. New York: ACM, 2020: 2542-2548.

[14] Khosla A, Das S A, Hamid R. What makes an image popular?[C]//Proceedings of the 23rd International Conference on World wide web-WWW'14, April 7-11, 2014. Seoul, Korea. New York: ACM Press, 2014: 867-876.

[15] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]//Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012, December 3-6, 2012, Lake Tahoe, Nevada, United States. [S.l.: s.n.], 2012: 1106-1114.

[16] Chen T, Borth D, Darrell T, et al. Deepsentibank: visual sentiment concept classification with deep convolutional neural networks[EB/OL]. (2014-10-30)[2021-04-09]. https://arxiv.org/abs/1410.8586.

[17] 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-MM'13, October 21-25, 2013. Barcelona, Spain. New York: ACM Press, 2013: 361-364.

[18] Nie L Q, Zhang L M, Yang Y, et al. Beyond doctors: future health prediction from multimedia and multimodal observations[C]//Proceedings of the 23rd ACM International Conference on Multimedia, October 26-30, Brisbane, Australia. New York: ACM, 2015: 591-600.

[19] Yang Y, Song J K, Huang Z, et al. Multi-feature fusion via hierarchical regression for multimedia analysis[J]. IEEE Transactions on Multimedia, 2013, 15(3): 572-581.

[20] Song X M, Nie L Q, Zhang L M, et al. Multiple social network learning and its application in volunteerism tendency prediction[C]//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, August 9-13, 2015, Santiago, Chile. New York: ACM, 2015: 213-222.

[21] Kan M N, Shan S G, Zhang H H, et al. Multi-view discriminant analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 188-194.

[22] Huang G B, Zhou H M, Ding X J, et al. Extreme learning machine for regression and multiclass classification[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2012, 42(2): 513-529.