何凡, 刘美玲, 于海泉,等. 基于多模态的短视频标签分类模型[J]. 智能计算机与应用,2025,15(5):194-198. DOI:10.20169/j.issn.2095-2163.250527

基于多模态的短视频标签分类模型

何 凡, 刘美玲, 于海泉, 范缤元, 赵柯桥
(东北林业大学 计算机与控制工程学院, 哈尔滨 150040)

摘 要: 目前网络短视频形式多样、内容灵活,传统视频识别方法在对其标签分类上大多效果有限,而视频的内容具有明显的多模态特征,融合多个模态进行视频识别成为了该研究领域的热点问题之一。基于此,提出一种利用 BiLSTM-Attention 网络的多模态融合视频标签分类模型。该模型利用视频的图形和音频特征为时间可对齐的序列特征的特点,通过 BiLSTM 对提取出的图形、音频特征数据进行对齐和处理,并把文本特征融入到图形、音频的时序 Attention 中,以此融合了 3 个模态,接下来在短视频标签数据集上进行训练和测试。结果表明在使用 3 个模态下一级精度和二级精度指标的准确率分别为 72%和 84%,相比使用 2 个模态的准确率有明显提升,尤其在精确度要求较高的一级精度指标中提升最为显著,提升了 9%的准确率,说明该模型引入多模态可以一定程度上提升短视频分类的精确度和准确率。

关键词: 视频标签分类; 多模态; 双向长短期神经网络; 时序; 注意力机制
中图分类号: TP391.1 **文献标志码**: A **文章编号**: 2095-2163(2025)05-0194-05

A multimodal-based model for short video tag classification

HE Fan, LIU Meiling, YU Haiquan, FAN Binyuan, ZHAO Keqiao
(College of Computer and Control Engineering, Northeast Forestry University, Harbin 150040, China)

Abstract: Currently, short videos on the internet have diverse forms and flexible content. Traditional video recognition methods are mostly limited in their ability to classify tags for them. However, the content of videos has clear multimodal features, and the integration of multiple modalities in video recognition has become one of the hot topics in this field. To address this, a multimodal fusion video tag classification model using BiLSTM-Attention networks is proposed. This model takes advantage of the temporal alignment of sequences by utilizing the graphical and audio features of videos. The extracted graphical and audio feature data is processed and aligned by BiLSTM, and text features are integrated into the temporal Attention of graphics and audio, thereby fusing three modalities. The resulting model is trained and tested on a short video tag dataset. The results show that the accuracy of the first-level precision and second-level precision indicators when using three modalities are 72% and 84%, respectively. Compared with using two modes, the accuracy has been significantly improved, especially in the first-level precision indicator where it has increased by 9% accuracy, indicating that introducing multiple modalities can improve the accuracy and precision of short video classification to a certain degree.

Key words: video tag classification; multimodal; BiLSTM; temporal; attention mechanism

0 引 言

随着智能手机和移动互联网的普及,人们越来越倾向于使用短视频来传达信息并分享生活中的点滴。但是,由于短视频存在时间短、内容丰富、质量不一等特点,传统的视频识别方法往往无法准确地对其进行识别与分类。为了解决这一问题,研究者们开始尝试将语音、图像和文本等多模态信息进行融合[1],以提高短视频识别的准确度。

基于多模态的短视频识别方法[2]利用了视频

基金项目: 黑龙江省大学生创新创业训练计划(S202210225334)。
作者简介: 何 凡(2002—),男,本科生,主要研究方向:深度学习,多模态; 于海泉(2002—),男,本科生,主要研究方向:深度学习,多模态; 范缤元(2002—),男,本科生,主要研究方向:深度学习,多模态; 赵柯桥(2002—),男,本科生,主要研究方向:深度学习,多模态。
通信作者: 刘美玲(1981—),女,副教授,主要研究方向:自然语言处理,机器翻译,数据挖掘,地理信息系统。Email:lmling2008@163.com。
收稿日期: 2023-12-25

中的视觉特征、音频特征、文本特征等多个维度的信息，并通过机器学习和深度学习等技术进行综合分析和处理[3]。目前，多模态方法主要包括以下几种：MFM[4]是一种针对多模态数据推荐问题的模型，能够将不同模态之间的相互作用进行建模；DeepCrossing[5]是一种基于神经网络和交叉层的多模态模型，适用于 CTR 预估等应用场景；DAN[6]是一种用于融合文本和图像信息的多模态模型，通过注意力机制从 2 个模态中提取最具代表性的特征，并对其进行加权组合。CNN-RNN[7]联合模型可以结合图像和文本信息，使用 CNN 提取图像特征以及运用 RNN 处理文本特征，再将其结合起来进行分类或预测等任务。

在过去的几年中，多模态相关领域的研究已经得到了快速发展，在许多应用领域取得了广泛的应用，如智能家居、社交媒体、安防监控等。然而，基于多模态的短视频识别仍然存在一些挑战，包括数据获取困难、数据标注不精确、多模态融合难度大等问题。

本文提出了多模态视频分类模型，融合文本、视频图像、音频三种模态进行视频多模态标签分类，采用特征融合和注意力机制，相比纯视频图像特征，有效提升高层语义标签分类效果。

1　多模态特征提取

在短视频分析[8]中，将图像、音频和文本信息进行提取可以获取更全面的多模态特征表示[9]，从而提升短视频相关任务的性能。本文使用 ResNet、VGGish 和 ERNIE 这些经典的模型，在短视频预处理中分别单独提取图像、音频和文本特征信息，为此后多模态特征的融合做准备，操作如下文所示。

ResNet[10]作为一种具有残差连接的深度卷积神经网络，在视频图像分类任务中有着广泛的应用。本研究使用 ResNet 提取短视频中图像信息的特征，以进行后续的多模态融合和分类等任务。

VGGish[11]是一个基于 VGG 网络架构的音频特征提取模型，可以通过卷积和池化层来高效地提取出音频中有用的频谱特征。本研究使用 VGGish 提取短视频中的音频信息特征，此后再与图像特征一同进行融合。

ERNIE[12]是一种基于预训练的语言模型，利用大规模语料库进行预训练，并能够通过微调适应各种下游自然语言处理任务，可以有效地提取出文本数据中的有用特征。本研究使用 ERNIE 模型提取短视频文本信息特征，从而为后续任务提供更全面的信息引导。

2　BiLSTM-Attention 多模特征融合模型

BiLSTM-Attention[13]是深度学习领域的一个经典模型，本文基于该模型通过 BiLSTM 对提取出的图形、音频特征数据进行对齐和处理，并把文本特征融入到图形、音频的时序 Attention 中，以此融合了 3 个模态。该模型的主体结构如图 1 所示。

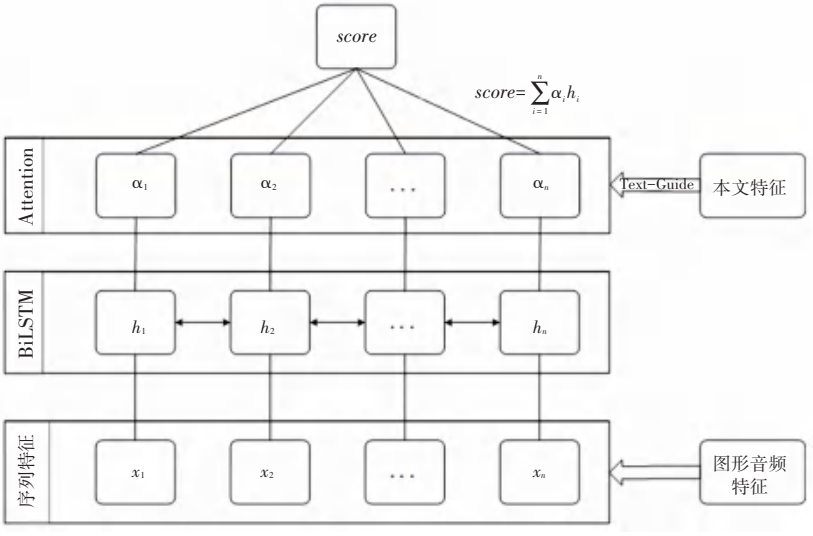

图 1　BiLSTM-Attention 主体结构

Fig. 1　Main structure of BiLSTM-Attention

2.1 BiLTSM 图像音频特征序列学习

BiLSTM 是一种双向长短期记忆神经网络(Bidirectional Long Short-Term Memory),结合了 LSTM 和双向性的思想[14]。LSTM 由于其设计特点,非常适合用于对时序数据的建模,图像和音频特征为时间可对齐的序列特征,双向 LSTM 能够很好地捕捉时间序列数据之间的依赖关系[13],进而更好地整合不同维度的数据。通过双向 LSTM 模型对提取出的多模态特征数据进行对齐和处理,以便进行后续的维度融合和分类[15]。

本文 BiLSTM 的核心是将输入的时序可对齐的图形或音频特征序列同时从左向右和从右向左分别输入到 2 个 LSTM 层中,并且将计算得到的隐状态拼接在一起作为模型最终的输出。这样,在任意时刻 t 处,这个输出就可以根据该序列在 t 之前和 t 之后的全部信息来计算得到,包括远距离时序依赖的信息。该方法能够显著提高模型的表现,因为能够更好地捕捉序列中的内在规律,尤其适用于内部关系比较复杂的问题[15],其网络结构如图 2 所示。

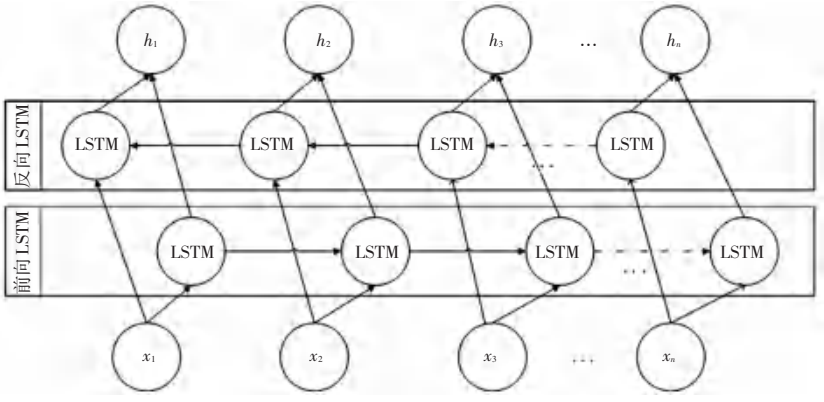

图 2 BiLSTM 双向网络结构

Fig. 2 Structure of BiLSTM bidirectional network

BiLSTM 的具体实现方式与普通的 LSTM 类似,都由 3 个门(遗忘门、输入门、输出门)和 1 个状态单元组成。除此之外,BiLSTM 引入了 1 个前向 LSTM 和 1 个反向 LSTM 分别负责将序列从左到右和从右到左传播,使得每个时刻的输出可以从左右 2 个方向上来求得[16]。首先对于前向 LSTM,其中遗忘门用于控制上一个状态细胞中的信息被丢弃的程度,输入门用于决定当前输入 x_t 的信息有多少可以被加入到细胞状态 c_t^f 中,即:

$$f_t^f = \sigma(W_f^f x_t + U_f^f h_{t-1}^f + b_f^f) \tag{1}$$

$$i_t^f = \sigma(W_i^f x_t + U_i^f h_{t-1}^f + b_i^f) \tag{2}$$

其中 W_f^f,U_f^f 和 b_f^f 是 LSTM 的模型参数,用于实现门控和状态更新,$\sigma(\cdot)$ 表示 Sigmoid 函数。然后根据当前的输入 x_t 和前一个状态 h_{t-1}^f 计算新的细胞状态,将上一时刻的状态细胞信息乘以遗忘门级别,再将当前的新信息通过输入门考虑进来,即:

$$c_t^f = \tanh(W_c^f x_t + U_c^f h_{t-1}^f + b_c^f) \tag{3}$$

$$c_t^f = f_t^f \odot c_{t-1}^f + i_t^f \odot c_t^f \tag{4}$$

其中,$\tanh(\cdot)$ 表示双曲正切激活函数,"$\odot$"表示向量的逐元素乘法(Hadamard 积)。接下来根据当前状态细胞 c_t^f 和前向 LSTM 的上一个状态 h_{t-1}^f 计算出对应时刻的输出 o_t^f,再根据当前状态细胞和输出门计算出最后的隐含状态 h_t^f,即:

$$o_t^f = \sigma(W_o^f x_t + U_o^f h_{t-1}^f + b_o^f) \tag{5}$$

$$h_t^f = o_t^f \odot \tanh(c_t^b) \tag{6}$$

反向 LSTM 与前向 LSTM 类似,只是在计算过程中的输入序列是反向的,得到隐含状态 h_t^b。最后,最终时刻 t 的网络输出由前向 LSTM 和后向 LSTM 的状态拼接而成 $h_t = [h_t^f, h_t^b] \in R^{2m}$。至此完成了图形和音频特征的融合。

2.2 Attention 文本特征指导多模态拼接

在本模型中,Attention 机制是通过文本特征指导[17]实现的。具体来说,给定输入序列的表示矩阵 $\boldsymbol{H} = [h_1, h_2, h_3, \cdots, h_n] \in R^{d \times n}$,其中 d 为 BiLSTM 隐层向量的维度,n 为输入序列的长度。那么,对于每一个时刻 $t \in [1,n]$,可以定义一个查询向量 $\boldsymbol{q}_t = l(h_t)$,其中 $l(\cdot)$ 为线性变换操作。

然后,计算所有位置与当前查询位置之间的相似性得分,可由下式计算求得:

$$e_{ij} = \boldsymbol{q}_i^{\mathrm{T}} \boldsymbol{W}_\alpha h_j + \boldsymbol{H}^{\mathrm{T}} \boldsymbol{v} \tag{7}$$

其中,$\boldsymbol{W}_\alpha$ 表示学习参数的可学习矩阵。得分 e_{ij} 衡量着当前时刻 t 对所有其它时刻之间的重要程度。可以通过使用文本特征来动态调整 Attention 机制中计算的权重矩阵 $\boldsymbol{W}_\alpha$ 的某些元素值,以此引

导模型更好地关注特定的输入部分。其方法是为每个特征类别设置一个权重参数 $\lambda_i(i \in \{1,2,\cdots,k\})$，用于控制其对应的词向量权重的加强或削弱，研发设计结构如图 3 所示。

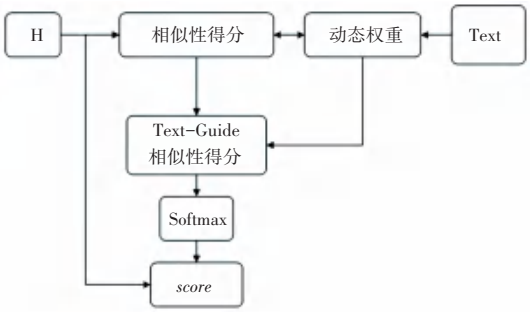

图 3　Text-Guide Attention 结构

Fig. 3　Structure of Text-Guide Attention

具体来说，首先预处理了文本数据，对含有特定特征类别（如名称、行为等）的单词进行标记[18]。那么，在计算 Attention 权重时，可以先将这些含有特定类别单词的词向量，分别加上一个对应的辅助向量 $\beta_i \in R^d$，其中 d 为词向量维度大小，即：

$$v_{ij} = e_{ij} + \beta_i \times e_{ij},\ i = 1,2,\cdots,n;\ j = 1,2,\cdots,d \tag{8}$$

这样，表示矩阵就变成了一个加性表示形式 $\boldsymbol{V} = [v_1, v_2, \cdots, v_n] \in R^{d\times n}$。然后，根据查询向量 $\boldsymbol{q}_t$ 和每个位置 $H_j(j \in [1,n])$ 之间的内积，计算得到相似性向量 $e_t \in R^n$：

$$e_t = \boldsymbol{W}_s \boldsymbol{q}_t + \boldsymbol{H}^{\mathrm{T}} \boldsymbol{V} \tag{9}$$

其中，权重矩阵 $\boldsymbol{W}_s \in R^{d\times d}$ 用来映射查询向量 $\boldsymbol{q}_t$ 的维度以适应和表示矩阵 $\boldsymbol{V}$ 相乘。此外，注意力得分可以通过对相似性向量进行 Softmax 操作，并用以构建每个时刻的注意力权重向量 α_t，即：

$$\alpha_t = Softmax(e_t) \tag{10}$$

将 α_t 和表示矩阵 $\boldsymbol{H}$ 进行加权求和，即可得到 $score_t$：

$$score_t = \sum_{j=1}^{n} \alpha_{tj} h_j \tag{11}$$

最终，可以将 $score_t$ 作为输出进行分类。这样，通过引入注意力机制，就强化了模型中文本与特定视频帧、音频特征的匹配作用，提高了模型在序列对齐及分类任务上的性能表现[19-20]。

3　实验方法

3.1　数据集

数据集使用 AI Studio 提供的视频标签数据集，数据源来自 UGC 视频和视频标题，包含视频特征、以及标题和标签信息。该数据集提供已经抽取好的图像、音频、标题的特征文件，可有效加快训练效率，其中各特征的抽取操作和格式为对视频进行抽帧，获得图像序列；抽取视频的音频 pcm 文件；收集视频标题，简单进行文本长度截断。

3.2　模型训练与测试

本文算法基于深度学习框架 Paddle2.0，实验环境为 Windows 10 操作系统，代码运行环境为 Python3.6，使用 NVIDIA GeForce GTX 1650 的图形处理器（GPU）加速运算。训练时所设置的具体参数如下：学习率（*learning_rate*）为 0.000 7，训练轮数（*epoch*）为 30 次，暖身期为 5 轮，梯度下降优化器进行学习率衰减的 *epoch* 数分别为 5、10、15 和 20，每经过梯度下降优化器中设定次数，则将学习率按 0.2 倍进行衰减。使用数据集上预训练好的模型框架作为本文的初始化网络模型，将训练好的模型在 Pycharm 软件上进行测试，验证本文模型对网络视频多模态识别的准确性。

4　实验结果分析

4.1　评价指标

研究使用多级精度（*Hit*@1，*Hit*@2）指标作为视频识别准确性的数值指标，其计算方法分别是：

$$Hit@1 = \frac{1}{N}\sum_{i=1}^{N}(y_i(1) = y_i) \tag{12}$$

$$Hit@2 = \frac{1}{N}\sum_{i=1}^{N}[(y_i(1) = y_i) \vee (y_i(2) = y_i)] \tag{13}$$

其中，N 表示测试集样本数；$y_i(1)$ 和 $y_i(2)$ 分别表示模型对第 i 个样本的最高概率预测结果和次高概率预测结果的类别标签；y_i 表示该样本的真实类别标签。符号“∨”表示逻辑或运算，即只要 2 个条件中有任意一个满足，整个条件就为 1。

可以看出，*Hit*@1 代表的识别精确度比 *Hit*@2 高，同时两者越高，代表各精度下的识别准确率越高。

4.2　多模态下的识别效果和性能分析

将本文算法模型图像音频模态和图像音频文本模态下检测结果进行比较，结果见表 1。

表 1　模型测试结果比较

Table 1　Comparison of model test results

模型	*Hit*@1	*Hit*@2
图像+音频	0.63	0.78
图像+音频+标题文本	0.72	0.84

由表 1 可以看出，只有图像和音频模态的识别

准确率相对较低,多级精度指标只达到 0.63 和 0.78,使用图像、音频和文本三个模态来识别的准确率有了一定提升,多级精度指标达到了 0.72 和 0.84,尤其在精确度要求较高的指标 *Hit*@1 中提升更为显著,提升了 0.09 的准确率。

经过一系列实验测试,得出结论,所提出的多模态融合模型可以有效地提升短视频分类的准确性,在多级精度上均有显著的提高。

5 结束语

本文提出了一种基于多模态的短视频识别方法,可以结合图形、音频、文本多个模态来识别出短视频的多级标签,以实现视频自动分类标注等功能。该方法首先使用 ResNet、VGGish、ERNIE 等网络实现了各个模态特征的提取,接下来提出使用 BiLSTM 网络实现图形和音频这 2 个时序特征的对齐,Attention 拼接实现文本指导特征的融合。通过在视频标签数据集上进行训练和测试,实验结果表明,多级精度均随着融合模态量的增加有了明显提升,更好地实现了视频多级标签识别,提升了识别精度。后续研究主要考虑 2 个方面。一是改进各模态特征抽取的网络模型,使抽取的特征能更好反映高层语义信息,以更高效提升内容识别的精确度;另一方面是在文本特征融合策略上进行改进,如可考虑在增加字幕文本或视频内含文本等特征的融合等方面进行深入研究。

参考文献

[1] 李祎. 基于多模态深度学习的情感识别研究[D]. 武汉:华中科技大学,2022.

[2] 乾竞元,高伟,滕国伟. 基于多模态特征融合的动态视频摘要算法[J]. 工业控制计算机,2022,35(10):81-84.

[3] 吕凤川. 基于语音与图像信息的端到端多模态情感识别[D]. 天津:天津大学,2021.

[4] SYMEONIDIS P, MALAKOUDIS D. Multi - modal matrix factorization with side information for recommending massive open online courses[J]. Expert Systems with Applications, 2019, 118: 261-271.

[5] WANG Ruoxi, FU Bin, FU Gang, et al. Deep & cross network for Ad click predictions[J]. arXiv preprint arXiv, 1708.05123, 2017.

[6] NAM H, HA J W, KIM J. Dual attention networks for multimodal reasoning and matching [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ:IEEE,2017: 2156-2164.

[7] XU K, BA J, KIROS R. Show, attend and tell: Neural image caption generation with visual attention[J]. arXiv preprint arXiv, 1502.03044,2015.

[8] 李亚鑫. 短视频深度多模态关联表示学习及其应用研究[D]. 天津:天津大学,2021.

[9] 朱康. 基于深度学习和特征融合的多模态情感识别研究[D]. 南京:南京邮电大学,2022.

[10] 张顺,龚怡宏,王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用[J]. 计算机学报,2019,42(3):453-482.

[11] 马银蓉. 基于表情、文本和语音的多模态情感识别[D]. 南京:南京邮电大学,2021.

[12] 黄山成,韩东红,乔百友,等. 基于 ERNIE2.0 - BiLSTM - Attention 的隐式情感分析方法[J]. 小型微型计算机系统,2021,42(12):2485-2489.

[13] 阮进军,杨萍. 基于 Att-CN-BiLSTM 模型的中文新闻文本分类[J]. 通化师范学院学报,2022,43(12):65-70.

[14] ZHOU Peng, SHI Wei, TIAN Jun, et al. Attention - based bidirectional long short - term memory networks for relation classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. ACL, 2016: 207-212.

[15] 谢浩鹏. 基于时序网络与时空信号先验的视频稳定方法研究[D]. 南京:南京理工大学,2021.

[16] 陈巧红, 李妃玉, 孙麒,等. 基于 LSTM 与衰减自注意力的答案选择模型 [J]. 浙江大学学报(工学版), 2022, 56 (12): 2436-2444.

[17] LONG Xiang, GAN Chuang, MELO G D, et al. Attention clusters: Purely attention based local feature integration for video classification [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7834-7843.

[18] 张帅,黄勃,巨家骥. 一种改进的融合文本主题特征的情感分析模型[J]. 数据与计算发展前沿,2022,4(6):118-128.

[19] 肖允鸿. 基于视觉和文本的标注短视频情感分析研究[D]. 南京:南京邮电大学,2023.

[20] 顾晓娜. 基于时空特征融合的多模态情感识别研究[D]. 南京:南京邮电大学,2023.