计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200800198

弹幕信息协助下的视频多标签分类

陈洁婷　王维莹　金　琴
中国人民大学信息学院　北京 100872
(jietingchen@ruc.edu.cn)

摘　要　文中探究了弹幕信息协助下的视频多标签分类任务。多标签视频分类任务根据视频内容从不同角度赋予视频多个标签,与视频推荐等应用紧密相关。多标签视频数据集的高标注成本和对视频内容的多角度理解是该研究领域面临的主要问题。弹幕是一种新近出现的用户评论形式,受到了众多用户的欢迎。由于用户参与度高,弹幕视频网站的视频拥有大量用户自发添加的标签,这些标签是天然的多标签数据。文中以此构建了一个多标签视频数据集,并整理出了视频标签间的层级语义关系,该数据集在未来将公开发布。同时,弹幕文本模态包含大量与视频内容相关的细粒度信息,因此在以往视频分类工作融合视觉和音频模态的基础上,引入弹幕文本模态进行视频多标签分类研究。在基于聚类的 NeXtVLAD 模型、注意力 Dbof 模型和基于时序的 GRU 模型上进行实验,在增加弹幕模态后,*GAP* 指标最高提升了 23%,证明了弹幕信息对该任务具有辅助作用。此外,还探索了如何在分类中利用标签层级关系,通过构建标签关系矩阵来改造标签,进而将标签语义融入训练。实验结果表明,加入标签关系后,*Hit*@1 指标提升了 15%,因此其能优化多标签分类的效果。此外,MAP 指标在细粒度小类上提升了 4%,说明标签语义的引入有利于预测样本量较少的类别,具有研究价值。

关键词:分类;多标签;弹幕;视频;标签关系;多模态
中图法分类号　TP399

Multi-label Video Classification Assisted by Danmaku

CHEN Jie-ting, WANG Wei-ying and JIN Qin
School of Information, Renmin University of China, Beijing 100872, China

Abstract　This work explores the multi-label video classification task assisted by danmaku. Multi-label video classification can associate multiple tags to a video from different aspects, which can benefit video understanding tasks such as video recommendation. There are two challenges in this task, one is the high annotation cost of dataset, and the other is how to understand video from multi-aspect and multimodal perspectives. Danmaku is a new trend of online commenting. Danmaku video has lots of manual annotations added by website users for high user engagement. It can be used as classification data directly. This work collects a multi-label danmaku video dataset and builds a hierarchical label correlation structure for the first time on danmaku video data. The dataset will be released in the future. Danmaku contains informative and fine-grained interaction data with the video content. This paper introduces danmaku modality to assist classification based on previous works, most of which only combine the visual and audio modalities. This paper choses cluster-based model NeXtVLAD, attention Dbof and temporal based GRU models as baselines. Experiments show that danmaku data is helpful, which improves *GAP* by 0.23. This paper also explores the use of label correlation, updating the video labels by a relationship matrix to integrate the semantic information into training. Experiments show that the leverage of label correlation improves *Hit*@1 by 0.15. Besides, the *MAP* can be improved by 0.04 in fine-grained labels, which indicates that the label semantic information benefits the prediction of small classes and it is valuable to explore.

Keywords　Classification, Multi-label, Danmaku, Video, Label correlation, Multi-modal

1　引言

近年来,流行的信息载体逐渐转变成了多模态媒体,如短视频、弹幕视频。面对庞大的用户群体和激增的视频数量,我们亟需相关的多媒体视频理解技术来处理多模态的信息。视频分类是视频理解中的经典任务,是视频推荐等应用的基础。

到稿日期:2020-08-29　返修日期:2020-10-05　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61772535);北京市自然科学基金(4192028);国家重点研发计划(2016YFB1001202)
This work was supported by the National Natural Science Foundation of China (61772535), Beijing Municipal Natural Science Foundation (4192028) and National Key Research and Development Plan (2016YFB1001202).
通信作者:金琴(qjin@ruc.edu.cn)

视频承载的信息丰富多元，在现实应用中，因为不同用户群体关注视频中的不同内容，所以一个视频应用有多个标签描述。因此，我们关注到视频的多标签分类任务，期望使用深度学习模型协助视频平台进行预测，补充和纠正视频标签，以提升用户体验感。

对多标签视频数据集进行标注需耗费大量的人力，这在一定程度上限制了相关研究，如今出现了一种新兴的媒体类型——弹幕视频。弹幕是一种短评，以滚动、停留等形式按时序和视频一同显示，为观看者营造了一种共同分享、实时互动的气氛，因此广受欢迎。在众多用户的参与下，弹幕视频拥有多个发布者和观众自发添加的标签，这些标签是天然的多标签数据。对视频内容进行多角度的理解是视频多标签分类的难点之一，尤其在标签含义相近的情况下。而弹幕数据含有与视频内容相关的丰富信息，具有交互性强、时序性好的特点，有助于对视频进行多标签分类。

为便于开展视频多标签分类的相关研究，本文利用 Bilibili 平台的视频制作了一个多标签弹幕视频数据集。该数据集共包含 4951 个视频，总时长 557h，含 5330393 条弹幕评论，视频内容涵盖科教、生活、体育、娱乐、影视、文艺等领域，共涉及 252 个标签，平均每个视频有 6.35 个标签。除此之外，我们还结合视频网站分区，整理了这些标签之间的层级关系，并以树状层次结构呈现，以支持标签关系及标签语义的相关研究。图 1 给出了数据样例。

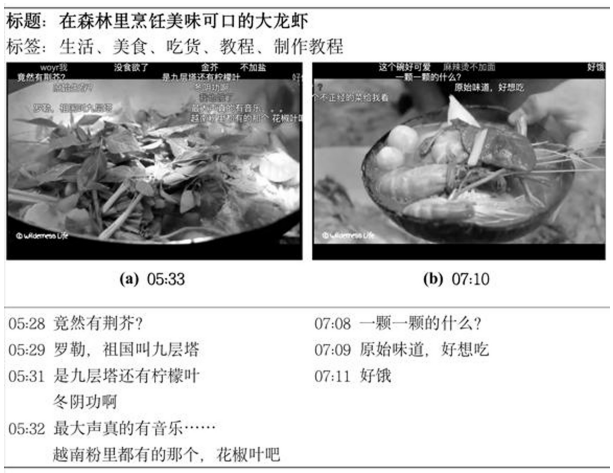


图 1 数据样例

Fig. 1 Data example

本文以基于聚类的 NeXtVLAD 模型[1]、注意力 Dbof 模型[2-3]和基于时序的 GRU 模型[4]这 3 个深度学习模型作为基础进行实验。已有的工作往往仅考虑了音频和视觉模态，因此本文加入了弹幕文本模态，并进行了多模态信息融合。实验结果表明，3 个模型的 GAP 指标[5]在弹幕文本的协助下均有提升，其中提升最明显的是 GRU 模型，其 *GAP* 由 0.30 提升至 0.53，增量高达 23%，证明了弹幕信息对视频理解工作的价值。

此外，本文还尝试利用标签之间的层级语义关系协助视频分类，通过构造标签关系矩阵来改造标签，将标签语义关系融合到训练中。改造后的标签不再呈现非 0 即 1 的形式，而是可以用其他数值表达，能够体现标签语义的分布。多个指标的提升表明，该方案能够提升多标签分类的效果，其中 GRU 模型上的 *Hit*@1 指标[5]提升了 15%，Dbof 模型的 *MAP* 指标[5]在细粒度小类上提升了约 4%。这也说明，标签语义的引入有利于预测样本量较少的类别，具有研究价值。

2 相关工作

本任务涉及视频的多标签分类，且引入了弹幕数据进行多模态信息融合。本节将从弹幕相关任务、视频多标签分类、标签关系利用、多模态分类这 4 个方面展开介绍。

2.1 弹幕的相关研究

近年来，弹幕作为一种新兴媒体逐渐引起了人们的关注，Yang 等[6]从 Acfun 网站上收集了一个包含音乐和搞笑视频的弹幕视频数据集，用于自动标注任务。第一个公开的弹幕数据集 TSCSet[7]包含动漫类型视频和少量标签（共计 42 种），平均每季动漫有 3 个标签。随后，两个单标签弹幕视频数据集 DR_E[8]和 Livebot[9]被发布，前者有 17 个影视类的标签，视频内容均为影视戏剧，后者有 19 个标签，内容为多类型短片。在这些数据集上，利用弹幕视频数据进行的研究有精彩片段预测、视频受欢迎程度评估、视频自动标注生成、视频摘要生成等。这些工作涉及统计分析、情感检测、主题分析等方法。

目前虽然已有 12 个弹幕视频数据集，但是这些数据集存在一些局限。首先，除 Livebot 以外，其他数据集均专注于特定类别的视频，如动漫、电视剧。而包含多种类视频的 Livebot 的总时长仅为 114h，每个视频仅包含一个标签，且其类别涵盖也不够广泛。其次，上述工作大都针对弹幕本身的单个模态进行统计分析和数据挖掘，仅有弹幕生成[9]和视频摘要任务[10-11]涉及到视频的多模态研究。总之，由于弹幕功能自 2014 年流行至今仅 6 年，相关研究工作仍处于初级阶段，与弹幕视频相关的领域还有很大的研究空间。

2.2 视频多标签分类

由于缺乏多标签视频数据集，相关工作多基于文献[3]发布的 Youtube-8M 数据集和多标签分类任务基准方法。文献[3]将多标签视频分类分为两个阶段，即帧级特征的聚合和分类预测，这两个阶段可以通过端到端训练实现。其中，帧级特征聚合方法包括基于聚类的聚合（如 Dbof[3]，Fisher Vector[12]，VLAD[13]）和基于时序的聚合（如 LSTM[14]，GRU[4]）。针对分类阶段，文献[3]给出了两种分类器：Logistic 分类和多专家系统分类。此外，其还提出了适用于视频多标签分类问题的指标 MAP，Hit@1 和 PERR。

在第一届 Youtube-8M 视频多标签分类挑战赛中，获胜者提出了 WILLOW 框架[15]，该框架利用 NetVLAD[16]模型来聚合视频特征，并设计了 Context Gating 模块以捕捉标签间的依赖关系。然而，这一届的参赛模型大多存在着参数规模过大的问题。次年，官方总结了第一届的优秀模型，并提出了一个新的评价指标 GAP[5]，该指标能够同时评估准确率和预测置信度。该年的任务重点是分类模型的压缩，模型大小需限制在 1GB 以内。在单模型上，表现最佳的是基于 WILLOW 框架的改进方案[1]，改进的 WILLOW 框架能在保证性能的同时，显著减小模型规模。

2.3 标签关系的利用

一些文本分类数据集具有与本文数据集类似的标签关系结构,文献[17]利用递归正则化的方法,将标签层级结构关系融入分类模型。同年,文献[18]提出了基于图神经网络的解决方案,利用标签共现条件概率,将原本非 0 即 1 的标签向量转变成元素值在 0~1 之间分布的形式。此外,一些研究者还通过迁移学习[19]、胶囊网络[20]等方法,探索如何利用标签关系协助文本分类。

2.4 多模态融合

融合多模态进行分类的研究已取得了一定的进展。文献[21]利用残差注意力机制,融合视频、文本、音频这 3 个模态,实现了灾害场景分类。第一届 Youtube-8M 挑战赛中,Wang 等[22]自行收集了视频对应的文本信息(如简介、关键词等),分别聚合成视频级别特征后,以串接的方式将其简单融合,投入分类器。虽然同时应用上述 3 种模态的分类工作还较少,但是一些学者对其他模态组合进行了研究。文献[23]提出了一种文本和图片模态融合以进行细粒度分类的方法。该方法将图片模态用于普通分类,语言模态则侧重于细粒度分类,各自预测后再按一定权重进行后期融合,二者起到互补作用。

3 弹幕视频的多标签数据集构建

本文构建了一个多标签弹幕视频数据集,该数据集将被公开发布。数据集包含 4951 个视频,总时长为 557 h,视频标签共 252 种,平均每个视频有 6.35 个标签,且提供了标签的层级结构关系。数据集还包含了超过 500 万条弹幕,平均每个视频有 1077 条弹幕。本节将介绍数据的采集过程、数据的预处理过程和数据集的统计分析结果。

3.1 数据的采集

本文的数据来源于中文 Bilibili 视频网站,该网站在 2019 年的第三季度拥有的活跃用户数达 1.28 亿。我们根据 Bilibili 的导航分区构建了一个主题表,再以主题表中的内容为方向,找到其中热度较高的话题标签,并将其收入关键词表。接着,我们用关键词表中的词进行查询,将查询结果按弹幕数量从高到低排序,按序在每个主题方向下采集了 1000 个热门视频。通过这种方式获取到的数据涵盖类型广、弹幕数量多、用户互动情况好、数据具有代表性。

3.2 数据的处理

我们对数据进行了清洗、层级标签关系构建、数据划分等处理,以更好地支持相关工作。

3.2.1 数据的清洗

由于网络原因,在采集的过程中难免出现损坏的数据,因此需要对数据进行清洗。首先,我们去掉了下载不完全的损坏的视频,以及弹幕下载不完全或弹幕数量少于 20 的视频。对于弹幕文本,我们过滤了弹幕中的动画表情,仅保留了 UTF-8 编码范围内的字符。为了便于后续实验的展开,本文使用 ffmpeg 工具每隔 1 s 对视频进行抽帧,从视频中提取音频,并将其切分为 1 s 的小片段。然而,一部分视频出现了提取出的图片帧和音频片段不对齐的情况,本文也将这类视频一并过滤掉。最终,共整理出 4951 个能顺利进行处理的视频。

此外,在提取的数据中,视频的标签为用户手工标注,没有语法限制,不够规范,还有部分标签较为少见,缺乏代表性。因此,我们过滤了视频数量少于 20 的标签,并将不规范的标签合并到表述相近的规范标签中。最后,通过 3 名工作人员的整理和检查,本文确定了 244 个具有代表性的标签。其中,有 4%的视频原有标签均为不规范标签,在过滤后失去了所有的标签。因此,通过 3 名工作人员来人为地重新进行标注。为了减少偶然性,只有当 3 人中的 2 人都打了同一个标签,我们才采纳此标签为视频的标签标注。

3.2.2 层级标签关系的构建

在处理标签的过程中,我们注意到弹幕视频下的标签是多粒度的。在同一视频下,不同粒度的标签比较杂乱,如一个猫咪视频中可能同时出现“暹罗猫”“猫”“宠物”“生活”这几个标签。而传统的多标签分类数据集中,大多标签粒度统一,比较规整。我们考虑过将标签粒度筛选到同一粒度,但发现筛选后的标签数量变少,不能完整地体现视频主题,也不符合人们在标注时的真实思维过程。近年的几个中文文本多标签数据集,如知乎看山杯数据集[24]、LSHTC 竞赛中的数据集[25],其标签粒度情况与本文类似,它们都提供了标签的层次结构关系。受此启发,我们决定保留多粒度这一特性,通过标签之间的层次结构进行梳理,给出不同粒度标签之间的“父子关系”和“兄弟关系”,以辅助后续任务。

参考 Bilibili 网站的导航分区,本文设计了对应的 3 种标签粒度,即粗粒度、中粒度、细粒度,粒度之间无交叉。中粒度为粗粒度标签的子节点,细粒度为中粒度标签的子节点,同一细粒度标签可以属于多个中粒度标签,如细粒度标签“经典”可以属于中粒度标签“电视剧”“电影”“动画”。图 2 给出了标签结构示例,这里以粗粒度标签“生活”下的树状结构为例。

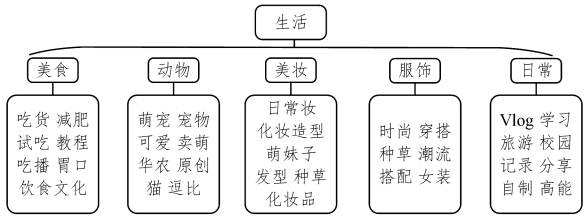


图 2 标签层级结构关系示例

Fig. 2 Example of branch of label hierarchical relationship

根据该结构,我们利用标签的“父子关系”,对视频的标签进行了补充。当视频的原标签作为子标签可以对应到唯一的父标签时,将父标签添加到视频标注中完善标注。最终的标签情况如表 1 所列。

表 1 标签分布及统计

Table 1 Label distribution and statistics

	标签种数	每个视频的平均标签数
粗粒度	10	1.82
中粒度	49	2.08
细粒度	193	2.45
总计	252	6.35

3.2.3 数据的划分

我们将 4951 个视频按照 7:2:1 的比例划分成训练集、验

证集和测试集。为避免某些标签在训练集中缺失，我们对每个划分上的标签分布情况进行了统计和调整，以确保不存在0次学习的情况。最终划分后的视频数量为训练集3466个，验证集990个，测试集495个。

3.3 数据分析

由于已有多位研究者发布了弹幕数据集，我们选择了与本文数据集最为相似的3个数据集进行比较。由表2可知，相比TSCSet和DR_E数据集的单一视频类型，Livebot和本文提出的数据集涵盖了更多种类型的视频，泛化性更强。从标签的角度来看，本文的数据集是多标签的，有252种标签，既包含可视实体也包含抽象概念，内容多样，是弹幕视频中首次给出标签间的层级结构关系的数据集。此外，本文数据集中每个视频的平均标签数为6.35，比同为多标签数据集的TSCSet的视频平均标签数高出两倍多，因此本文的数据集能更完整地描述每个视频。从弹幕的角度来看，本文数据集的弹幕密度最高，其值高达2.66条每秒，远多于TSCSet数据集的0.19条每秒，因此本文的数据集内容更丰富，交互性更强。

表2 弹幕视频数据集比较

Table 2 Comparison of danmaku video datasets

	TSCSet	DR_E	Livebot	本文
视频数量	17870	8156	2361	4951
视频时长/h	47835	—	114	557
弹幕数量/条	32949297	57176457	895929	5330393
平均每秒弹幕数	0.19	—	2.18	2.66
标签形式	多标签	单标签	单标签	多标签
标签种数	42	17	19	252
视频平均标签数	3	1	1	6.35
标签关系	未给出	未给出	未给出	给出
视频内容	动漫	影视戏剧	多类型	多类型
来源网站	Bilibili	Youku	Bilibili	Bilibili

综上，本文数据集具有内容多样、弹幕质量高、标签丰富且具备标签结构关系的优点，对于相关的多媒体研究能起到较好的支持作用。

4 多标签分类模型

本节将介绍选用的视频多标签分类模型和不同阶段融入弹幕模态的设计，以及利用标签层级关系为模型融入语义信息的方法。

4.1 视频多标签分类模型

本文选择了基于聚类的视频多标签分类模型(NeXtVLAD方案、Dbof方案)，以及基于时序的模型GRU方案作为基准模型。

4.1.1 NeXtVLAD方案

NeXtVLAD[1]是一种局部特征聚合模型，它起源于VLAD模型[13]，再经NetVLAD[15-16]模型演化而来，具体如图3所示。经典的VLAD模型能以聚类的方式，将图片的若干局部特征整合成一个全局特征。但由于计算过程中存在不可导符号函数，无法应用于端到端结构。NetVLAD模型则对这一缺点做出了改进，将符号函数变为平滑可微的softmax函数，使其可以进行端到端训练，但这一改进的代价是参数量较多。受到ResNeXt对ResNet模型改进的启发，后来的NeXtVLAD模型通过分组增加VLAD层的非线性参数，减少了VLAD输出层参数，缩小了整体参数规模，同时保持了模型性能。

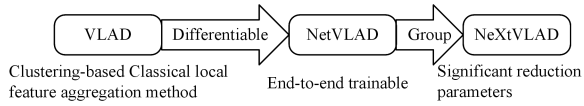


图3 NeXtVLAD模型演进过程

Fig. 3 Development of NeXtVLAD model

图4给出了NeXtVLAD方案的整体情况，该方案借鉴了WILLOW框架[15]。对于视频分类，其所使用的模态包括视觉和音频，将视觉和音频帧级特征分别进行NeXtVLAD聚合，得到视频级特征，再进行串接融合。

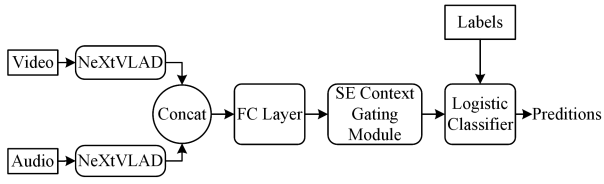


图4 NeXtVLAD方案结构

Fig. 4 NeXtVLAD project structure

4.1.2 Dbof方案

受到词袋模型的启发，文献[3]提出了一种将帧级特征聚合为视频级特征的方法Dbof，即深度帧袋池化。如图5所示，对于一个有M个帧级特征、每个帧特征为N维的视频数据，Dbof模型从M个帧级特征中随机选取出R个分别投入共享参数的全连接层中，而经过这一层后，R个N维的特征被投影扩充到了U维。这一过程相当于把帧级特征差别性地分配到了“词袋”中的U个“单词”中，从而得到一个U维的视频级向量表示。接着，模型将这R个U维特征采用池化的方法聚合。文献[2]将注意力池化用于Dbof模型，认为对于一个视频，每个帧的重要性不同，而常用的平均池化和最大池化不能实现这一点。

完成池化操作后，由于N维特征被扩充到了较高的U维，不利于训练，模型接入了一个单层的全连接层将其降维至H，然后将其投入Logistic分类器，经过端到端的训练后，即可得到预测结果。

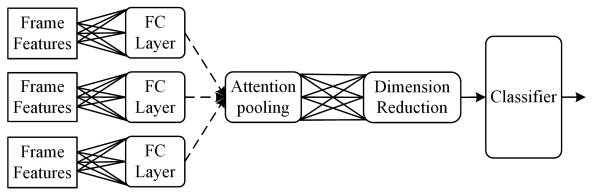


图5 Dbof方案的结构

Fig. 5 Structure of Dbof project

4.1.3 GRU方案

前两个模型均为基于聚类的帧级特征聚合方案，这里我们选用了一个基于时序的聚合方案GRU作为补充，其结构如图6所示。GRU是LSTM的一种变体，参数量更少，但两者产生的效果在多种任务上相当[4]，我们认为其在时序模型中较适合本文数据集的情况。本文采用的是最基础的GRU

模型,激活函数选用 RELU,隐状态及输出的维度为 512,且仅使用单层 GRU 单元,不做叠加。然后,将其接入 Logistic 分类器得到预测结果。

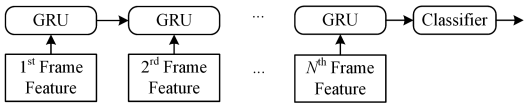


图 6 GRU 方案的结构

Fig. 6 Structure of GRU project

4.2 弹幕信息协助设计

为了利用丰富的弹幕信息,我们需要将弹幕文本模态融合到模型中。文本模态的融合可以在不同阶段进行,阶段的差别对参数规模及最终效果会造成不同的影响。由于本文数据集较基准模型原适配的数据集小,本着减小参数规模的原则,我们分别计算 NeXtVLAD 方案、Dbof 方案和 GRU 方案在模型的不同阶段融合文本模态所需的参数量。

在 NeXtVLAD 方案中,设一个视频有 M 个帧特征,每个帧特征的长度为 N 维,拟获得的视频级特征维度为 H,则其单个模态在聚合及降维操作上所需要的参数量为:

$$\lambda N(N+G+K(G+(H+1)/G)) \tag{1}$$

其中,λ 为模型中第一个全连接层扩充特征维度的倍数,G 为 N 维空间分组降维时的组数。经计算比较,我们选择将 3 个模态的帧级特征分别使用不同的 NeXtVLAD 模型进行聚合,再将生成的视频级特征串接,以实现模态融合,如图 7 所示。

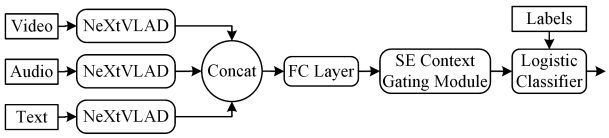


图 7 弹幕信息协助下的 NeXtVLAD 方案

Fig. 7 Structure of NeXtVLAD project assisted by danmaku

对于 Dbof 方案,假设一个视频有 M 个帧特征,每个帧特征的长度为 N 维,拟获得的视频级特征维度为 H。记模型中的"词袋"中有 U 个"单词"。依据模型结构,可以计算出 Dbof 方案所需的参数量为:

$$U(N+1)+H(U+1) \tag{2}$$

经计算比较,对于 Dbof 方案,本文选择先将 3 个模态的特征串接成一个多模态帧特征,再投入 Dbof 模型中进行聚合。

对于 GRU 方案,若记输入特征维度为 J,隐藏状态特征维度为 S,则模型参数量为:

$$3*(S(S+J)+S) \tag{3}$$

经计算,先拼接特征,再将其放入 GRU 中进行时序聚合得到的参数量更小,因此这里也选择先拼接特征的方式。

4.3 标签关系利用设计

2019 年的多标签文本分类工作[18]使用标签改造的方式融合了标签间的关系,其利用每个视频上已有的标签标注,以及标签间在视频上的共现条件概率,对标注外的其他标签维度进行了修正。受此启发,本文决定将标签的层级关系通过标签改造的方式融入训练过程中。

记标签数量为 Q,本文将标签间的关系表示为一个 $Q*Q$ 的关系矩阵 $\boldsymbol{C}$,252 个标签在横轴和纵轴上按相同的顺序占位。在本文数据集中,标签被分为 3 个层级,呈树状结构。对于呈父子关系的结点,我们认为其存在较强的关联,因此在关系矩阵上将其对应位置的值加上一个相关值 ε 作为更新(本文设 ε 为 0.1)。假设我们检索到了父标签 i 和子标签 j,则可按式(4)更新关系矩阵位置 $\boldsymbol{C}[i,j]$ 的值,具体公式如下:

$$\boldsymbol{C}[i,j]=\boldsymbol{C}[i,j]+\varepsilon \tag{4}$$

需要注意的是,最终不同标签之间的关联值不应超过相同标签之间的关联值。为了规范,本文将矩阵的对角线设置为 1,并将其他位置上超过 1 的关联值截取至 1,如此可以构造出一个标签关联矩阵,如图 8 所示。

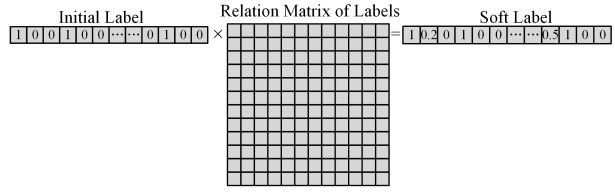


图 8 标签语义关系利用方法

Fig. 8 Method for leveraging label semantic relationship

将每个视频的标签向量与该矩阵相乘,即可得到一个新的标签向量。新标签向量上每一个分量元素的值都融合了原标注为 1 的标签与其他标签的关联程度。新标签向量将标签从过去非黑即白的 0/1 形式,转变成了分布的形式。我们期望这样的设计能够利用标签间的层级关系融入语义信息,在训练过程中起到积极作用。

5 实验与分析

5.1 数据预处理

(1)视觉。我们将视频按照 1 s 的间隔进行帧采样,并将其压缩成统一的尺寸,即 224 * 224,再放入在 ImageNet 上预训练好的 ResNet101 模型中[26]进行特征提取,每张图片可得一个 2048 维的视觉特征。

(2)音频。我们将音频切割成时长为 1 s 的小段,使用在 Youtube 视频数据集上预训练好的 VGG 模型进行特征提取。每段音频可得一个 128 维的特征。

(3)文本。弹幕是用户观看视频时即兴发布的短评,存在许多不常用的表达和杂乱的标点符号噪音。为消除这些噪音对分类任务的影响,我们利用 jieba 分词工具,依据 tf-idf 值,提取出本弹幕语料中最关键的 30 000 个词作为词典,并将所有弹幕去除标点符号后按照词典过滤,筛选掉不常用词汇。此外,由于弹幕文本为句法灵活、用词新颖的短评,少有适用的预训练语料,我们利用 gensim 工具包在本文数据集上训练了词向量模型,最终得到的词向量维度为 512。为了与其他模态的帧特征对齐,我们依照时间,对 1 s 内的多条弹幕进行拼接,经过分词取得词典中词的词向量,最后取平均,便可得到能代表 1 s 内弹幕文本信息的 512 维文本特征向量。

5.2 评价指标

本文选用了3个经典多标签分类指标作为评价指标[5]，取每个视频预测置信度最高的前20个标签作为预测结果进行计算。

(1)GAP。全局平均精度由所有视频预测的所有标签在不同置信度阈值下的精确率和召回率计算而得。GAP不仅考虑到了预测结果的命中情况，还能反映每个预测的置信度，其计算公式如下：

$$P(\tau)=\frac{\sum_{v\in V}\sum_{k=1}^{20}I(f_{v,k}\geqslant\tau)I(e_{v,k}\in G_v)}{\sum_{v\in V}\sum_{k=1}^{20}I(f_{v,k}\geqslant\tau)} \tag{5}$$

$$R(\tau)=\frac{\sum_{v\in V}\sum_{k=1}^{20}I(f_{v,k}\geqslant\tau)I(e_{v,k}\in G_v)}{\sum_{v\in V}|G_v|} \tag{6}$$

$$GAP=\sum_{j=1}^{10000}P(\tau_j)[R(\tau_{j-1})-R(\tau_j)] \tag{7}$$

其中，v为视频集合V中的一个视频，其对应的标注标签集为G_v。$(e_{v,k},f_{v,k})$表示视频v的第k个预测，$e_{v,k}$为预测标签，$f_{v,k}$为置信度，$f_{v,k}\in[0,1]$。式(7)对PR曲线求积分，其中$\tau_j=j/10000$，$j\in\{0,1,\cdots,10000\}$。GAP值越大，则预测效果就越好。

(2)MAP。平均精度先计算所有视频在每类标签下的精确率和召回率，再采用类似于GAP的方法，计算出每类的命中情况及预测置信度的值。该方法考虑到了因数据标签分布不均衡而造成的误差，将所有类别一视同仁分别进行计算，最后对所有标签类别求和取平均，得到整体情况。该指标的计算公式如下(设变量含义与GAP相同)：

$$P_e(\tau)=\frac{\sum_{v\in V}\sum_{k=1}^{20}I(f_{v,k}\geqslant\tau)I(e_{v,k}\in G_v)I(e_{v,k}=e)}{\sum_{v\in V}\sum_{k=1}^{20}I(f_{v,k}\geqslant\tau)I(e_{v,k}=e)} \tag{8}$$

$$R(\tau)=\frac{\sum_{v\in V}\sum_{k=1}^{20}I(f_{v,k}\geqslant\tau)I(e_{v,k}\in G_v)I(e_{v,k}=e)}{|v:e\in G_v|} \tag{9}$$

$$AP_e=\sum_{j=1}^{10000}P_e(\tau_j)[R_e(\tau_{j-1})-R_e(\tau_j)] \tag{10}$$

$$MAP=\frac{1}{|E|}\sum_{e\in E}AP_e \tag{11}$$

(3)Hit@1。该指标表示样本中视频的预测置信度最高的标签$e_{v,1}$属于标注标签集G_v的概率。计算公式如下(设变量含义与GAP相同)：

$$Hit@1=\frac{\sum_{v\in V}I(e_{v,1}\in G_v)}{|V|} \tag{12}$$

5.3 实现细节

在模型训练中，我们选用的优化器为Adam优化器，学习率下降方式均为指数衰减法，损失计算的基础为交叉熵。我们将视频长度及其对应的弹幕规范到5 min，将每秒内弹幕词数量规范到20个，超过规范的则截断，不足则用0补全。

5.4 弹幕协助下的视频多标签分类实验

为探索弹幕模态对多标签视频分类的作用，我们分别在3种模型上进行了仅用视觉和音频模态的实验(V+A)、仅用文本模态的实验(T)和多模态的实验(V+A+T)，具体结果如表3所列。

表3 弹幕协助下的视频多标签分类结果

Table 3 Experiment results of multi-label video classification assited by danmaku

	GAP	*Hit@1*	*MAP*
NeXtVLAD(T)	0.58	**0.81**	**0.50**
NeXtVLAD(V+A)	0.44	0.68	0.37
NeXtVLAD(V+A+T)	**0.60**	0.80	0.47
Dbof(T)	0.51	0.75	0.42
Dbof(V+A)	0.48	0.71	0.36
Dbof(V+A+T)	**0.56**	**0.79**	**0.43**
GRU(T)	0.48	**0.76**	0.35
GRU(V+A)	0.30	0.60	0.16
GRU(V+A+T)	**0.53**	0.64	**0.38**

实验结果显示，相比传统的视觉和音频模态，多模态信息的利用使得大部分指标提升。多模态方案在GRU模型上的效果最为显著，相比视觉加音频的结果在GAP上提升了23%，相比纯文本模态，Dbof多模态模型在GAP上提升了5%。这表明弹幕信息的多模态融合对弹幕视频的多标签分类有着重要的作用。此外，对比3个模型的结果可知，基于聚类的模型表现更好。本文认为，分类问题的重点不在于时序信息，因此时序性模型GRU的优势未能得到体现。

对于同一个模型，纯文本模态也发挥了很好的作用，全面超过了视觉加音频的模态，其在Hit@1和MAP指标上甚至存在超过多模态的情况。这说明视频弹幕与视频内容是密切联系的，这类数据有利于多媒体的研究。

5.5 标签关系利用实验

我们对标签层级关系的利用也进行了设计，并在3个模型上进行了实验(以H为前缀)，结果如表4所列。

表4 标签层级关系利用实验的结果

Table 4 Experiment results of leveraging label hierarchical relationship

	GAP	*Hit@1*	*MAP*
NeXtVLAD(V+A+T)	**0.60**	0.80	0.47
Dbof(V+A+T)	0.56	0.79	0.43
GRU(V+A+T)	**0.53**	0.64	**0.38**
H_NeXtVLAD(V+A+T)	0.59	**0.81**	**0.48**
H_Dbof(V+A+T)	**0.59**	**0.80**	**0.49**
H_GRU(V+A+T)	0.51	**0.79**	0.36

由表4可知，利用标签关系后，Dbof模型的指标全面提升，而在GAP指标上3个模型的整体表现不如预期。本文通过分析认为，标签关系的利用会引导模型同时预测出相关的父子标签，不再将两个标签永远看作独立的。但这样的引导有时是错误的，因为父子类别不一定存在必然的共现，或者虽然其本质上是正确的，但没有被用户标注上。因此，对于反映整体预测情况的GAP来说，这部分的负面效果抵消了其优势。

在Hit@1指标上，3个模型均表现优秀，GRU模型提升了15%，并且MAP指标也基本提高或相似。这说明标签语义关系确实能够提升预测命中效果，且对某些类别有积极作用。为探究各类的情况，我们进一步统计了MAP提升明显的Dbof模型在不同粒度标签上的表现。为了统计所有标签

的情况，我们不再只选取预测置信度前 20 高的标签，而考虑取置信度前 252 高的标签（即所有标签）作为预测结果来计算 *MAP* 值，具体结果如表 5 所列。

表 5 各粒度内的 MAP 增量

Table 5 MAP increments for each grain

（单位：%）

	粗粒度	中粒度	细粒度
MAP 增量	0.04	0.81	3.78

表 5 中的 MAP 增量指 H_Dbof 模型与 Dbof 模型的 MAP 指标的差值。统计显示，标签关系的利用对于样本量较小的细粒度类别效果显著，增量高达 3.78%，远高出有大量样例的中粒度和粗粒度标签。这一结果充分表明，标签语义关系的引入有利于多标签分类中样本量较少的类别的预测，具有研究价值。

5.6 预测实例展示

图 9 给出了实验中 *GAP* 最高的 NeXtVLAD 模型的可视化结果，按置信度由高到低排序，所展示的预测标签数为原标签数加 3。加实线框的内容表示成功命中，加虚线框的内容表示虽未命中，但明显与视频内容语义相关的预测。由图 9 中的实例(1)、实例(2)、实例(4)可见，本文模型在大多数情况下能对各种类型的视频进行准确预测。但在实例(3)的情况下，由于原始标签本身较为笼统，模型预测的命中为 0。然而，观看视频后不难发现，虚框中的预测结果其实已经理解了视频和弹幕内容，是虽未命中但高度相关的预测。结合实例(1)、实例(4)中虚框的表现可以看出，模型还可以对原有标签起到补充、修正的作用。

图 9 预测实例

Fig. 9 Prediction cases

结束语 本文研究了弹幕信息协助下的视频多标签分类，构建了一个新的多标签弹幕视频数据集，同时首次在弹幕视频上整理出了标签层次关系，能较好地支持相关的多媒体研究工作。我们以 NeXtVLAD，Dbof，GRU 这 3 个模型为基础，进行了融合弹幕信息的多模态实验和标签语义关系利用实验。实验结果表明，弹幕数据信息丰富，且与视频内容密切相关，融入该模态的多模态模型能够显著提升分类效果，辅助视频理解。此外，标签语义关系也值得利用，其在细粒度小类上的作用尤为明显。我们也注意到了本文工作的不足，如多模态融合方法简单、对数据不均衡问题的处理欠考虑、标签关系利用方法有待优化等，我们将在未来的工作中逐一改进这些不足。总体来说，弹幕视频是一种新兴的媒体形式，是信息丰富、交互性强且具备时序性的多模态数据，值得更多的探索。

参 考 文 献

[1] LIN R，XIAO J，FAN J. Nextvlad：An efficient neural network to aggregate frame-level features for large-scale video classification[C]// Proceedings of the European Conference on Computer Vision (ECCV). Munich，Germany，2018.

[2] GARG S. Learning video features for multi-label classification [C]// Proceedings of the European Conference on Computer Vision (ECCV). Munich，Germany，2018.

[3] ABU-EL-HAIJA S，KOTHARI N，LEE J，et al. Youtube-8m：A large-scale video classification benchmark[J]. arXiv：1609.086.75.

[4] CHO K，VAN MERRIENBOER B，BAHDANAU D，et al. On the properties of neural machine translation：Encoder-decoder approaches[J]. arXiv：1409.1259.

[5] LEE J，NATSEV A，READE W，et al. The 2nd YouTube-8M Large-Scale Video Understanding Challenge[C]// Proceedings of the European Conference on Computer Vision (ECCV). Munich，Germany，2018：193-205.

[6] YANG W，RUAN N，GAO W，et al. Crowdsourced time-sync video tagging using semantic association graph[C]// 2017 IEEE International Conference on Multimedia and Expo (ICME). Hong Kong，China，2017：547-552.

[7] LIAO Z，XIAN Y，YANG X，et al. TSCSet：A crowdsourced time-sync comment dataset for exploration of user experience improvement[C]// 23rd International Conference on Intelligent User Interfaces. Tokyo，Japan，2018：641-652.

[8] BAI Q，HU Q V，GE L，et al. Stories That Big Danmaku Data Can Tell as a New Media[J]. IEEE Access，2019，7：53509-53519.

[9] MA S，CUI L，DAI D，et al. Livebot：Generating live video comments based on visual and textual contexts[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Hilton Hawaiian Village，Honolulu，Hawaii，USA，2019，33：6810-6817.

[10] OLSEN D R，MOON B. Video summarization based on user interaction[C]// Proceedings of the 9th European Conference on Interactive TV and Video. Lisbon，Portugal，2011：115-122.

[11] WANG X，JIANG Y G，CHAI Z，et al. Real-time summarization

of user-generated videos based on semantic recognition[C]// Proceedings of the 22nd ACM International Conference on Multimedia. Orlando, Florida, USA, 2014: 849-852.

[12] SÁNCHEZ J, PERRONNIN F, MENSINK T, et al. Image classification with the fisher vector: Theory and practice[J]. International Journal ofCcomputer Vision, 2013, 105(3): 222-245.

[13] JÉGOU H, DOUZE M, SCHMID C, et al. Aggregating local descriptors into a compact image representation[C]// 2010 IEEE computer society conference on computer vision and pattern recognition. San Francisco, California, USA, 2010: 3304-3311.

[14] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[15] MIECH A, LAPTEV I, SIVIC J. Learnable pooling with context gating for video classification[J]. arXiv: 1706.06905.

[16] JÉGOU H, DOUZE M, SCHMID C, et al. Aggregating local descriptors into a compact image representation[C]// 2010 IEEE computer society conference on computer vision and pattern recognition. San Francisco, California, USA, 2010: 3304-3311.

[17] PENG H, LI J, HE Y, et al. Large-scale hierarchical text classification with recursively regularized deep graph-cnn[C]// Proceedings of the 2018 World Wide Web Conference. Lyon, France, 2018: 1063-1072.

[18] WANG L, CHEN S, ZHOU H. Boosting Up Segment-level Video Classification Performance with Label Correlation and Reweighting [EB/OL]. https://static.googleusercontent.com/media/research.google.com/zh-CN//youtube8m/workshop2019/c_07.pdf.

[19] BANERJEE S, AKKAYA C, PEREZ-SORROSAL F, et al. Hierarchical Transfer Learning for Multi-label Text Classification [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Fortezza da Basso, Florence, Italy, 2019: 6295-6300.

[20] CHEN B, HUANG X, XIAO L, et al. Hyperbolic Capsule Networks for Multi-Label Classification[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle, Washington, USA, 2020: 3115-3124.

[21] POUYANFAR S, WANG T, CHEN S C. Residual Attention-Based Fusion for Video Classification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, California, USA, 2019.

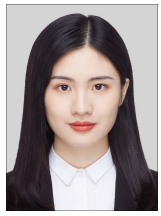
[22] WANG Z, KUAN K, RAVAUT M, et al. Truly multi-modal youtube-8m video classification with video, audio, and text[J]. arXiv: 1706.05461.

[23] HE X, PENG Y. Fine-grained image classification via combining vision and language[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA, 2017: 5994-6002.

[24] 中国人工智能学会, 知乎. 2017 知乎看山杯机器学习挑战赛 [EB/OL]. https://www.biendata.xyz/competition/zhihu/.

[25] PARTALAS I, KOSMOPOULOS A, BASKIOTIS N, et al. Lshtc: A benchmark for large-scale text classification[J]. arXiv: 1503.08581.

[26] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, 2016: 770-778.



CHEN Jie-ting, born in 1997, postgraduate, is a member of China Computer Federation. Her main research interests include multimedia computing and so on.

JIN Qin, born in 1972, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include multimedia computing and human computer interaction.