计算机科学
COMPUTER SCIENCE

基于多模态视频分类任务的模态融合策略研究

王一帆, 张雪芳

引用本文

王一帆, 张雪芳. 基于多模态视频分类任务的模态融合策略研究[J]. 计算机科学, 2024, 51(6A): 230300212-5.

WANG Yifan, ZHANG Xuefang. Modality Fusion Strategy Research Based on Multimodal Video Classification Task [J]. Computer Science, 2024, 51(6A): 230300212-5.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300212

基于多模态视频分类任务的模态融合策略研究

王一帆　张雪芳
武汉邮电科学研究院　武汉 430070
(wyf1519uir@163.com)

摘　要　尽管过往人工智能相关技术在众多领域取得了成功,但是通常只是模拟了人类的某一种感知能力,也就意味着被限制在处理单个模态的信息之中。从多个模态信息中提取特征并进行有效融合对于从弱/限制领域人工智能向强/通用人工智能的发展迈进具有重要意义。本研究基于编码器-解码器结构,在视频分类任务上对多模态信息的特征编码进行早期特征融合、对各模态信息的预测结果进行后期决策融合以及对两者相结合的不同多模态信息融合策略进行了对比研究;同时对音频模态信息参与模态融合的两种方式进行了对比,即直接将音频进行特征编码进而参与模态融合或音频通过语音转文本进而以文本的形式参与模态融合。实验结果表明,将文本和音频模态单独的预测结果与另外两种模态的融合特征的预测结果进行决策融合能够进一步提高分类预测准确率;此外,通过语音识别将语音转换成文本模态信息,能够更加充分利用其中包含的语义信息。

关键词:多模态;模态融合;语音识别;视频分类
中图分类号　TP181

Modality Fusion Strategy Research Based on Multimodal Video Classification Task

WANG Yifan and ZHANG Xuefang
Wuhan Research Institute of Posts and Telecommunications,Wuhan 430070,China

Abstract　Despite the success of AI-related technologies in many fields,they usually simulate only one type of human perception,which means that they are limited to process information from a single modality. Extracting features from multiple modal information and fusing them effectively is important for developing general AI. In this paper,a comparative study of different multimodal information fusion strategies based on an encoder-decoder architecture with early feature fusion for feature encoding of multimodal information,late decision fusion for prediction results of each modal information,and a combination of both is conducted on a video classification task. This paper also compares two ways to involve audio modal information in modal fusion,i.e.,directly encoding audio with features and then participating in modal fusion or audio by speech-to-text and then participating in modal fusion in the form of text. Experiments show that decision fusion of the prediction results of text and audio modalities alone with those of the fused features of the other two modalities can further improve the classification prediction accuracy under the experimental approach of this study. Moreover,converting speech into text modal information by ASR(Automatic Speech Recognition) can make fuller use of the semantic information contained in it.

Keywords　Multimodality,Modality fusion,Speech recognition,Video classification

1　引言

当一个研究问题涉及多种模态的数据时,它就被定性为多模态问题。多模态机器学习的目的是建立能够处理和联系多种模态信息的模型。这是一个充满活力的多学科交叉研究领域,其重要性会随着弱/限制领域人工智能(Weak/Narrow AI)到强/通用人工智能(Strong/General AI)的过渡而日益彰显。在多模态的早期研究中,研究人员根据在多模态任务中不同模态数据之间融合时期的差异将多模态方法分为早期融合和晚期融合两种。Baltrušaitis 等根据多模态机器学习面临的挑战重新定义了更为细致和通用的分类方式,即表征、翻译、对齐、融合和协同学习,这种新的分类方式有助于研究人员更好地了解该领域的状况[1]。

随着智能手机的广泛应用和安防摄像头的不断普及,视频类数据量呈现爆炸式增长,并已成为当前最具体量的大数据类型之一[2]。根据 CNNIC 发布的第 50 次《中国互联网络发展状况统计报告》,截至 2022 年 6 月,我国短视频用户规模已达到 9.62 亿,占据网民总数的 91.5%[3]。随着知名短视频平台活跃用户数量不断攀升,思科公司预测,到 2023 年,视频数据将占互联网总流量的 80%[4]。这些来源于网络的视频通常是以多模态的形式存在的,即除了视频模态信息之外,还包含了音频和文本(字幕/弹幕)模态的信息。因此,如何利用人工智能相关技术对海量多模态视频中包含的信息进行理解与识别成为亟待解决的问题,这对于通用人工智能的发展和网络信息安全的保障都有着重要意义。

通用的视频理解面临着数据集数据匮乏、视频语义复杂

基金项目:国家重点研发计划(2019YFB1803600)
This work was supported by the National Key R&D Program of China(2019YFB1803600).
通信作者:张雪芳(zhangxuefang@fhxy.net.cn)

和视频多模态语义融合困难等诸多挑战。应对这些挑战的思路可以分为两种，其一是专注于具体应用领域需要解决的具体问题，例如为提高对某一类视频事件的准确率而专门针对其提高数据集质量或改进模型结构，这类工作追求的是模型在某一垂直领域的精通，Kaggle 等平台发布的算法竞赛所追求的精确度就是这类研究思想的代表，这类工作的尽头是在特定域的数据集上取得卓越的性能表现，本质上是基于深度学习的专家系统；其二则是随着多模态预训练的兴起而提出的多模态大一统模型的概念，这类模型的训练不局限于特定模态，它试图打通模态间的壁垒，把各个模态的数据联合起来进行建模学得一个通用型模型。同时，由于训练使用多模态数据的量级足够大，并且数据直接来自互联网，如用户评论、视频弹幕等，因此它不仅节省了人工标注的成本，而且能够很好地适配开域的识别任务。当需要将该模型应用于特定数据域上的任务时，只需采用“预训练＋微调”的迁移使用范式，如果预训练模型足够强大，甚至能够在零样本学习（Zero-shot Learning）中取得不亚于有监督学习模型的成绩。大规模图文预训练模型 CLIP 和 GPT-4 等便是这类工作的优秀代表作品[5-6]。这种多模态大一统模型的优点主要体现在泛化能力强，更接近于通用人工智能；缺点则在于前期预训练阶段需要足够的算力来支撑模型在互联网级别的训练数据上进行训练。

鉴于多模态视频理解任务的巨大算力需求，本文从多模态融合的合理结构的选择多模态视频分类任务展开了三模态融合策略研究。具体而言，本研究的创新性工作可以概括为以下两个方面：1）基于三模态视频涉及的文本、音频、视频 3 种模态信息，使用模态编码后的特征融合、解码后的决策融合构建出三模态融合模型结构设计的基础搜索空间；2）在以上基础上，将音频模态信息通过 ASR 转文本后编码并参与到模态融合，从而进一步扩大三模态融合的模型结构设计的搜索空间。然后基于多模态影视短剧多分类数据集 MM-IMDB，通过控制变量法在该搜索空间上展开了不同的模态融合策略的效果对比[7]，从而对多模态视频中的文本、音频、视频等三模态信息的有效融合策略产生了新的见解。

2 编码器-解码器结构及各模态特征编码器选择

2.1 编码器-解码器结构

编码器-解码器（Encoder-Decoder）结构是深度学习中常见的模型框架，编码器接受的输入数据和解码器的输出数据类型可以是文本、语音、图像、视频等任意模态的数据，同时，编码器和解码器使用的模型可以根据具体任务要求选择 CNN，RNN，Transformer 等任意符合输入输出规范的模型。

所谓编码器，就是用来将输入数据转化成一个固定长度的特征向量，解码器相应地则是用来将编码器生成的特征向量解码成目标输出。因此，编码器-解码器也是一种典型的端到端（End-to-End）的模型结构，在机器翻译、图像/视频描述等领域被广泛使用。基于这类结构思想的设计在生活中并不罕见，电话通信就是在一端将声音信号编码成电信号，然后再在另一端将接收到的电信号解码成声音信号。使用编码器-解码器结构的原因主要在于难以直接完成输入到目标输出的映射，需要经过一种中间数据的过渡才能更好地达到目标，就如同电话将声音信号转化为电信号是为了实现长距离传播[8]。图 2 为编码器-解码器结构示意图。

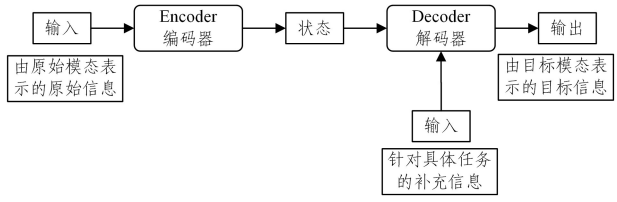

图 1 编码器-解码器结构

Fig. 1 Encoder-Decoder architecture

2.2 文本、音频、视频的编码器选择

不同模态的数据根据其数据特征都有各自适合的编码器模型用来提取特征。在多模态视频理解任务中，编码器输入为多模态数据，解码器输出则为视频分类标签或文本描述等信息。可以将每个模态数据单独用不同的模型编码，进而进行特征向量层面的特征融合，然后再由解码器进行处理。针对不同视频理解任务，解码器也需要进行相应的选择和设置。图 3 给出了以文本、音频、视频三模态为例的编码-解码过程。

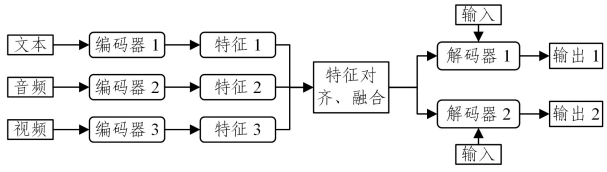

图 2 使用编码器-解码器结构解决多模态视频任务示例

Fig. 2 Example diagram for solving multimodal video tasks using encoder-decoder architecture

视频理解任务可以采用端到端的解决方案，在深度学习的背景下，即为使用编码器-解码器结构对多模态数据进行编码输入，经历中间步骤的特征融合，再针对具体视频理解任务进行解码输出。

文本数据的各类序列化编码表示方法由来已久，语音数据是按照一定采样率对声音进行采样，视频数据可以理解成图像帧序列。因此，以上 3 种模态的数据作为序列天然可以使用 RNN，LSTM，Transformer 等经典序列模型，但针对不同模态数据的细粒度特征已有诸多改进模型[9-10]。

用于文本模态编码的模型从独热编码（One-Hot Embedding）、词向量（Word2Vec）、Glove 和 Cove 等经典模型发展至今，BERT 和 GPT 系列模型凭借其卓越性能表现已占据自然语言处理领域的半壁江山[11-15]。用于音频模态的编码模型按照使用方式可以分为直接提取音频特征的预训练模型和通过模型训练提取特征两种，前者的代表性模型有 VGGish 和 PANNs，后者则可以使用一般性的序列特征提取模型，如 DNN，RNN，1-D CNN，Transformer 等，也可以将音频文件转换为声谱图再用 CNN 等二维卷积神经网络进行处理[16-17]。常用于视频模态编码的代表性模型有 I3D，SlowFast 以及用于视频的 Transformer 的各类变体（TimeSformer 等）[18-20]。

在各模态编码器的选择上，使用预训练的 I3D 和 SlowFast 对视频进行特征抽取；使用预训练的 VGGish 抽取音频特征；文本模态数据应当首选预训练的 BERT，但是考虑到 BERT 模型参数规模过大，采用更为轻量的 ALBERT 作为替代[21]。

在进行特征融合之前，分别使用独立的 LSTM 对图像特

征和音频特征进行序列学习。若选用早期特征融合的策略，在进行特征融合时，考虑到文本具有显式的高层语义信息，因此将其引入 LSTM 池化过程指导图像和音频时序权重分配，进行交叉融合，最后将文本、音频、视频特征进行拼接。在需要进行决策融合的模型组合中，使用集成学习中加权投票的方式进行决策融合。

3 模型结构设计的搜索空间

在编码器-解码器的模型结构下，针对视频分类任务的模型结构设计可以围绕各模态数据编码器的选择和特征融合的时间来进行差异化展开。

3.1 三模态特征融合的基础结构

各模态分别编码，然后将编码得到的特征向量进行特征融合，最后将融合特征输入解码器进行分类，可得多模态数据融合进行视频分类的基础模型，如图 4 所示。

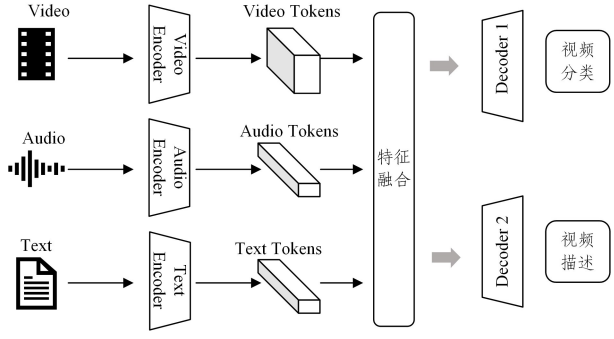

图 3 多模态数据特征融合完成视频任务的基础模型结构

Fig. 3 Basic model structure of multimodal data feature fusion to accomplish video tasks

特征融合的方法是将来自不同模态的特征向量直接使用拼接(Concatenation)操作，该操作是将输入特征结合起来的最常用和基础的操作。为不同下游任务设计的解码器中的网络层会自动对该操作进行自适应，使得各模态信息都能够得到应用[22-23]。

3.2 音频通过 ASR 转文本参与到特征融合的方法

考虑到音频数据与文本数据都具有自然语言的属性，可以在多模态数据进入编码器之前利用自动语音识别技术(Automatic Speech Recognition)将语音识别成文本再和原始文本一起作为文本编码器的输入，将三模态转换为双模态，再使用文本编码器和视频编码器进行特征提取，进行特征融合后解码输出。模型结构如图 5 所示。

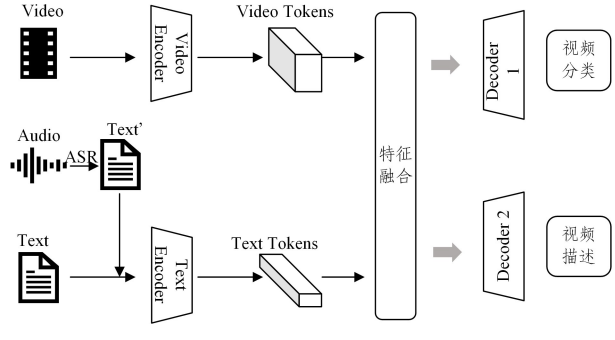

图 4 语音通过 ASR 转文本后双模态特征融合

Fig. 4 Bimodal feature fusion after speech is converted to text by ASR

3.3 早期特征融合结合后期决策融合

借鉴集成学习的思想，将各模态数据后编码分别输入解码器进行结果预测，进而使用后期决策融合的方法得到最终预测结果，或将早期特征融合与后期决策融合结合使用，即模态进行两两早期特征融合，第三种模态单独经过编码器-解码器进行预测，最后再进行决策融合，得到最终预测结果，如图 6 所示。决策融合采用 Boosting 集成学习方法。

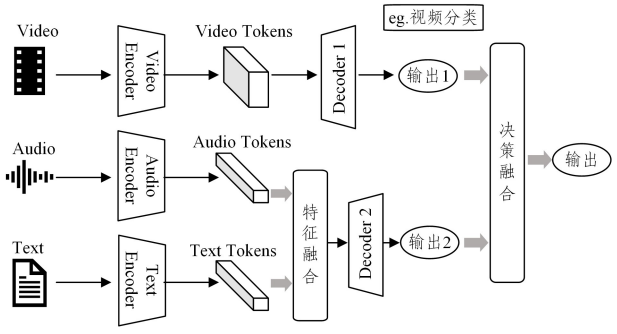

图 5 特征融合＋决策融合的多模态融合方案

Fig. 5 Multimodal fusion scheme of feature fusion＋decision fusion

综上所述，各模态数据可以选择多种不同的编码器，与上述 3 点差异性模型设计思路相结合，进一步扩大模型设计的搜索空间。表 1 列出了组成模型设计搜索空间的各个变量，其中特征融合以及决策融合栏目为各模态提供的方案集合{1，1/2，1/3}表示参与该类型融合的模态数量，“1”表示不参与融合，“1/2”表示参与融合的数量为 2，“1/3”表示参与融合的数量为 3。

表 1 模型设计的搜索空间

Table 1 Search space for model design

	编码器模型选项	语音转文本	特征融合	决策融合
文本	One-Hot Embedding Word2Vec GPT BERT …	—	{1，1/2，1/3}	{1，1/2，1/3}
音频	RNN CNN Transformer VGGish …	是/否	{1，1/2，1/3}	{1，1/2，1/3}
视频	I3D SlowFast TimeSformer …	—	{1，1/2，1/3}	{1，1/2，1/3}

4 实验与分析

实验首先对视频、音频和文本经各自编码器编码所得到的特征表示使用进行早期特征融合，然后将其融合并作为分类器的输入来验证不同模态信息对视频分类性能的贡献。PyTorchVideo 是一个开源的视频理解库，视频模态使用的 I3D 和 SlowFast 模型可以直接从中调用[24]。表 2 列出了不同模态信息及其组合参与特征融合时对应的视频分类表现，使用的数据集为 MM-IMDB 影视短剧多分类数据集。由表 2 中的实验结果可以看出，在该数据集上视频信号采用 SlowFast 模型取得了比 I3D 更好的效果，同时通过模态消融实验验证了多模态中每个模态的收益。单个模态作为特征进行分类预测时，视频的准确率最高，音频的准确率

最低，文本接近视频；双模态时，视频＋文本的预测准确率有明显提升，再加上音频后，提升有限。

表 2 三模态信息融合(特征融合)方案对应的表现

Table 2 Performance corresponding to trimodal information fusion(feature fusion) scheme

(%)

模态	模型	Top-1	Top-5
视频	I3D R50，8×8	50.50	56.22
视频	SlowFast R101	62.07	67.45
音频	VGGish	35.37	43.72
文本	BERT	60.08	68.83
视频＋音频	SlowFast＋VGGish	65.11	70.36
视频＋文本	SlowFast＋BERT	70.16	77.32
音频＋文本	VGGish＋BERT	61.85	69.48
视频＋音频＋文本	SlowFast＋VGGish＋ BERT	74.33	83.95

然后，将音频信号通过 ASR 转成文本并与原有文本信号以文本的形式参与到模态融合中，采用与上面实验相同的模态融合方法和分类器，得到的实验结果如表 3 所列。

表 3 音频通过 ASR 转文本后参与模态融合的效果对比

Table 3 Effect comparison of audio participation in modal fusion after speech is converted to text by ASR

(%)

模态	模型	Top-1	Top-5
ASR 文本	BERT	48.55	58.21
视频＋ASR 文本	SlowFast＋BERT	65.59	72.32
文本＋ASR 文本	BERT	64.43	73.36
视频＋文本＋ASR 文本	SlowFast＋BERT	79.91	88.50

对比表 4 与表 5 中的实验结果可以看出，当语音模态信号通过 ASR 转换成文本并以文本模态的形式参与到模态融合时，相比以音频形式直接编码并参与到模态融合时的分类预测准确率有所提升。其中，当单独使用音频信号进行预测时，转成文本相对于直接使用的 Top-1 准确率从 35.37%提高至 48.55%，效果显著；当三模态同时参与预测时，音频信号提前转换成文本相对于直接融合的 Top-1 准确率从 74.33%提高到 79.91%，Top-5 准确率同样有所提高。这表明在本文的实验条件下，音频信号先转换成文本比直接参与模态融合具有更好的分类效果，也从一定程度上说明了前者相比后者其语义信息能得到更好的利用。

最后，引入后期决策融合的方案，对以下两类特征融合和决策融合的组合分别进行了实验：1)三模态各自编码并各自进行分类预测，然后对它们的预测结果进行决策融合；2)两两模态各自编码并进行早期特征融合进而预测分类，剩余的一个模态单独编码和预测，最终将两个预测结果进行决策融合。实验结果如表 4 所列。

表 4 特征融合结合决策融合组成的不同模态融合方案效果对比

Table 4 Effect comparison of different modal fusion schemes composed of feature fusion combined with decision fusion

(%)

融合策略	Top-1	Top-5
视频＊文本＊音频	67.52	71.30
(视频＋文本)＊音频	65.24	69.20
(视频＋音频)＊文本	72.58	84.15
(文本＋音频)＊视频	78.70	90.95

注：＋”代表特征融合，“＊”代表决策融合。

根据表 4 的实验结果可以看出，将早期特征融合与后期决策融合相结合的多模态融合方案的确对预测准确率产生了正面或负面的影响。其中，直接将三模态分别进行特征编码和解码预测，并将预测结果进行决策融合时的 Top-1 和 Top-5 准确率均不及三模态特征融合时的准确率；将视频和文本进行早期特征融合并与音频的预测结果进行决策融合时两项准确率同样远不如直接对三模态进行特征融合；当使用视频和音频做特征融合，并与文本单独预测的结果进行决策融合时，Top-1 准确率略低于三模态特征融合，但 Top-5 准确率达到了 84.15%，略高于三模态特征融合的 83.95%；当文本与音频进行特征融合，并与视频预测的结果进行决策融合时，结果迎来了改观，其两项准确率均超过了三模态早期特征融合。这表明了使用文本和音频单独的预测结果与另外两种模态的融合特征的预测结果进行决策融合，可进一步提高分类预测准确率，同时也意味着这种融合策略在更好地联合利用多模态所包含的语义信息方面更具潜力。

结束语 本文基于编码器-解码器结构，对多模态信息早期特征融合、后期决策融合以及两者相结合的不同多模态信息融合策略进行了对比研究。实验表明，使用文本和音频单独的预测结果与另外两种模态的融合特征的预测结果进行决策融合可进一步提高分类预测准确率；此外，通过 ASR 将语音转换成文本模态信息，能够更加充分利用其中包含的语义信息。

出于节省算力的考虑，本研究在视频模态的编码器模型的选择上直接从 PytorchVideo 中挑选了其包含的 SlowFast 和 I3D 模型，虽然这两个模型原理直观，但并非目前最先进的视频编码模型。音频和文本分别使用的预训练的 VGGish 和 ALBERT 也是在模型性能与算力需求之间做出的平衡。若计算资源足够充裕，之后的研究则可以针对具体任务使用大规模数据集和 SOTA 模型进行预训练，进而将这些最先进的预训练模型作为编码器，对各模态信息进行更有效的提取，从而实现更好的模型性能表现。此外，特征融合阶段直接选用了常用的特征向量拼接的方法，在后续的研究中可以尝试引入注意力机制；决策融合阶段使用的基础集成学习的方法也可以尝试替换为更加丰富的决策融合算法，如贝叶斯估计、专家系统等。

参 考 文 献

[1] 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.

[2] KEEGAN M. The Most Surveilled Cities in the World[EB/OL] https://www.usnews.com/news/cities/articles/2020-08-14/the-top-10-most-surveilled-cities-in-the-world.

[3] 中国互联网网络信息中心. 第 50 次中国互联网络发展状况统计报告[R/OL]. (2022-08-31)[2022-09-10]. http://www3.cnnic.cn/NMediaFile/2022/1020/MAIN16662586615125EJOL1VKDF.pdf.

[4] Cisco. Cisco Annual Internet Report (2018－2023) White Paper [R]. 2020.

[5] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]// International Conference on Machine Learning. PMLR, 2021:

8748-8763.
[6] OpenAI,GPT-4 Technical Report[R]. 2023.
[7] AREVALO J,SOLORIO T,MONTES-Y-GÓMEZ M,et al. Gated multimodal units for information fusion[J]. arXiv: 1702.01992,2017.
[8] CHO K,VAN MERRIËNBOER B,BAHDANAU D,et al. On the properties of neural machine translation: Encoder-decoder approaches[J]. arXiv:1409.1259,2014.
[9] ELMAN J L. Finding Structure in Time[J]. Cognitive Science,1990,14(2):179-211.
[10] HOCHREITER S,SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation,1997,9(8):1735-1780.
[11] MIKOLOV T,CHEN K,CORRADOG,et al. Efficient estimation of word representations in vector space[J]. arXiv: 1301.3781,2013.
[12] PENNINGTON J,SOCHER R,MANNING C D. Glove: Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP). 2014:1532-1543.
[13] MCCANN B,BRADBURY J,XIONG C,et al. Learned in translation: Contextualized word vectors[J/OL]. https://proceedings. neurips. cc/paper_files/paper/2017/hash/20c86a628232a67e7bd46f76fba7ce12-Abstract. html.
[14] RADFORD A,NARASIMHAN K,SALIMANS T,et al. Improving language understanding by generative pre-training[J/OL]. https://www. semanticscholar. org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035.
[15] DEVLIN J,CHANG M W,LEE K,et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv:1810.04805,2018.
[16] ARANDJELOVIC R,ZISSERMAN A. Look,listen and learn[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:609-617.
[17] KONG Q,CAO Y,IQBAL T,et al. Panns: Large-scale pretrained audio neural networks for audio pattern recognition[J]. IEEE/ACM Transactions on Audio,Speech,and Language Processing,2020,28:2880-2894.
[18] CARREIRA J,ZISSERMAN A. Quo vadis,action recognition? a new model and the kinetics dataset[C]// proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:6299-6308.
[19] FEICHTENHOFER C,FAN H,MALIK J,et al. Slowfast networks for video recognition[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:6202-6211.
[20] BERTASIUS G,WANG H,TORRESANI L. Is space-time attention all you need for video understanding? [C]// ICML. 2021,2(3):4
[21] LAN Z,CHEN M,GOODMAN S,et al. Albert: A lite bert for self-supervised learning of language representations[J]. arXiv: 1909.11942,2019.
[22] NOJAVANASGHARI B,GOPINATH D,KOUSHIKJ,et al. Deep multimodal fusion for persuasiveness prediction[C]// Proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016:284-288.
[23] WANG H,MEGHAWAT A,MORENCYL P,et al. Select-additive learning: Improving generalization in multimodal sentiment analysis[C]// 2017 IEEE International Conference on Multimedia and Expo(ICME). IEEE,2017:949-954.
[24] FAN H,MURRELL T,WANG H,et al. PyTorchVideo: A deep learning library for video understanding[C]// Proceedings of the 29th ACM International Conference on Multimedia. 2021:3783-3786.

WANG Yifan, born in 1997, master. His main research interests include graph neural networks and multimodal machine learning.

ZHANG Xuefang, born in 1989, master, senior engineer. Her main research interests include intelligent optical networks and AI.