手机屏幕的截图

中度可信度描述已自动生成

情感计算大作业

成员1：李一鸣 202228013229030

成员2：张 兆 202228013229029

成员3：陈国鑫 2022E8013282125

代码开源地址：https://github.com/zhangzhao219/Multimodal-Sentiment-Analysis

2023年05月

目录

[1 引言 1](#_Toc132047245)

[1.1 任务概览 1](#_Toc132047246)

[1.2 数据描述 1](#_Toc132047247)

[1.3 评价指标 1](#_Toc132047248)

[2 基于音频模态的语音情感识别 1](#_Toc132047249)

[2.1 数据处理 1](#_Toc132047250)

[2.2 网络架构 1](#_Toc132047251)

[2.3 性能提升 1](#_Toc132047252)

[2.3.1 预训练模型 1](#_Toc132047253)

[2.3.2 数据增强策略 1](#_Toc132047254)

[3 基于语言模态的语音情感识别 1](#_Toc132047255)

[4 基于多模态的语音情感识别 1](#_Toc132047256)

[4.1 网络架构 2](#_Toc132047257)

[4.2 融合策略 2](#_Toc132047258)

[5 多模态情感识别系统 2](#_Toc132047259)

[6 参考文献 2](#_Toc132047260)

1. 引言
   1. 任务概览

语音作为语言的第一属性，在语言中起决定性的支撑作用，它不仅包含了说话人所要表达的文本内容，也包含说话人所要表达的情感信息。而情感则与人态度中的内向感受、意向具有协调一致性，是态度在生理上一种较复杂而又稳定的评价和体验，是一种综合了人类行为、思想和感觉的现象。

语音情感是指从语音信号中获取相应的情感信息，语音情感识别是计算机对人类上述情感感知和理解过程的模拟，利用计算机分析情感，提取出情感特征值，并利用这些参数进行相应的建模和识别，建立特征值与情感的映射关系，最终对情感进行分类。

语音情感识别是人机情感交互的关键[1]，对语音情感的有效识别能够提升语音的通俗性，使各种智能设备最大限度理解用户意图，提高机器人性化水平，从而更好地为人类服务。

* 1. 数据描述

本文选择南加州（USC）语音分析和解释实验室（SAIL）收集的交互式情感二元运动捕捉数据库IEMOCAP[2]进行语音情感识别探究。该数据集包含大约十二个小时数据，记录了来自 10 位演员的二元会话信息，它们被要求在假设场景中即兴对话，旨在引发特定的情绪（快乐、愤怒、悲伤、沮丧和中性等状态），此外，数据集当中还包含演员的面部、手部动作等模态信息。

考虑到本文致力于探索语音情感识别，并考虑到平衡数据类别分布等因素，本文选择从数据库当中抽取音频信息（wav文件夹下）及转录的文本信息（transcription文件夹下），由于某些类别存在较为严重的长尾现象，因此，本文选择了“neu”（中性）、“sad”（悲伤）、“hap”快乐、“ang”愤怒等四个类别，并将“exc”（激动）类别划分为“hap”类别。划分后的数据集包含5531条数据，数据分布如下图所示，

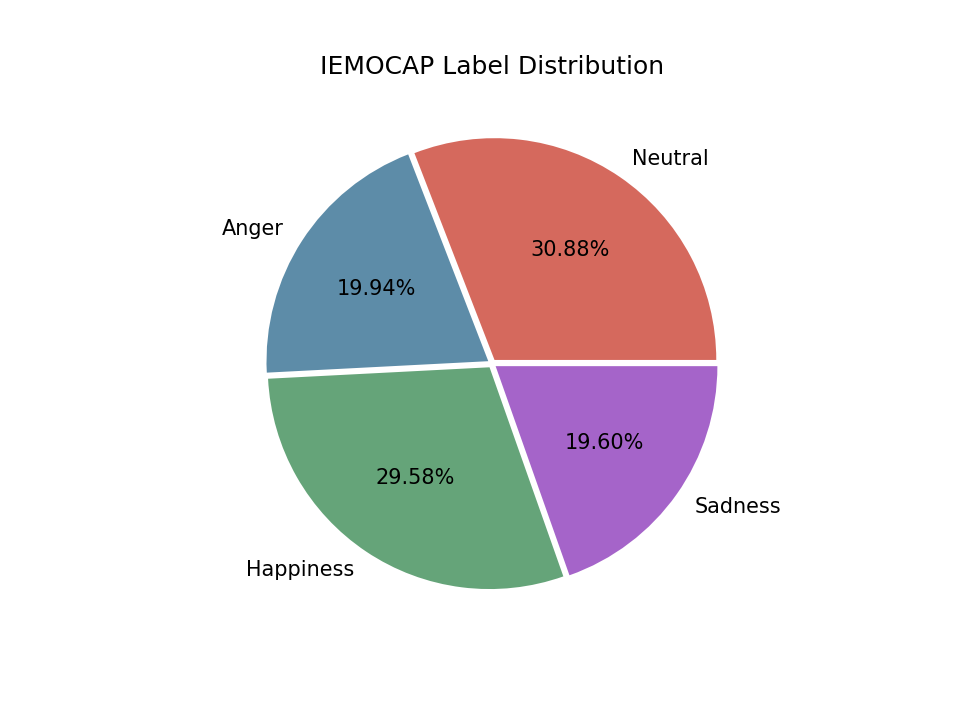


图1 数据集标签分布情况

此外，我们将数据集按照8：1：1划分训练集、验证集与测试集。

* 1. 评价指标

IEMOCAP情感识别任务本质是一类多类别分类任务，本文采用了学界相对常用的加权准确率分数（Weighted Accuracy，WA）及非加权准确率分数（Unweighted Accuracy, UA）作为评价指标，以更加充分地考虑各类情感的识别效果。

1. 基于音频模态的语音情感识别
   1. 数据处理

IEMOCAP数据集音频文件以每一场景下的对话为单位进行录制，每一音频文件对应一份标注文件，标注了该音频以第{start}秒为开始时间，第{end}秒为结束时间的音频片段及其对应的情感类别标签。考虑到神经网络不易处理时间长短不一的音频，本文先按照开始、结束时间戳截取音频片段，并将所有音频的长度统一为10s：对于少于10s的音频，将其重复播放至10s止；对于大于10s的音频，将其截取至10s。

对于音频模态，采用对数梅尔频谱图作为输入特征，经预加重、分帧、加窗、短时傅里叶变换、梅尔滤波、对数运算等操作，得到最终的声学特征。频谱图的计算依赖hop size（决定帧的数目）、window length（决定一帧的窗口内有多少采样点）、mel bins（决定特征维数） 等参数，参数依据模型的选择略有不同，将在下文进行详细说明。此外，计算整个数据集音频在频域维度及时域维度的归一化参数进行特征归一化。

* 1. 网络架构

我们首先设计了一个CRNN网络作为基准网络，其通过2D卷积层提取频谱图的特征，并通过池化操作逐步缩小特征图尺寸，直至将频域维度特征由频谱图的mel bins压缩至1，得到的时序特征经过Bi-GRU进一步加强时序关联后，最后一个隐状态被送入全连接分类头，通过Softmax激活得到音频属于4个类别的概率，CRNN的网络架构如下图所示，在图中，我们标注了网络各个层次输出的张量尺寸，方便读者更好地理解模型设计及音频分类任务处理的大致流程，

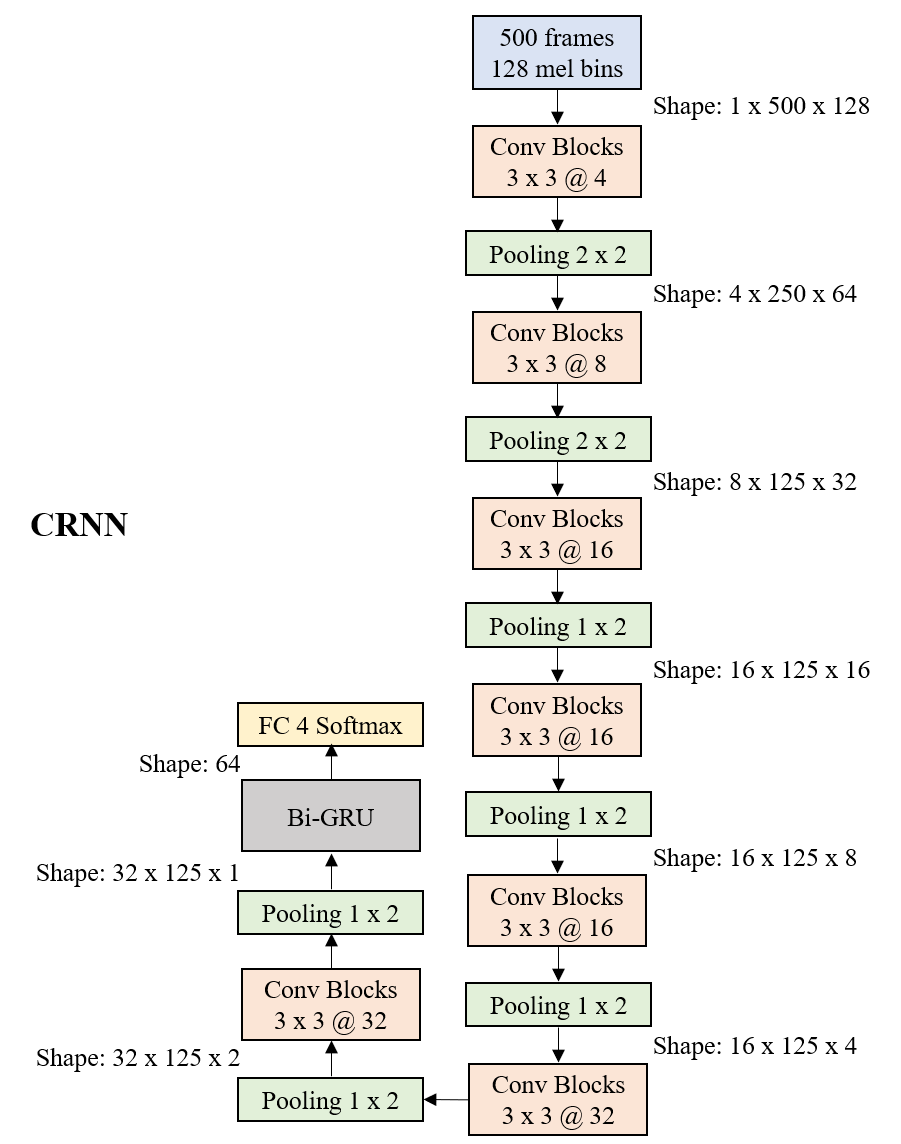


图2 CRNN网络架构示意图

网络输入频谱图的hop size 为320，window length为1024，sample rate 为16k.模型采用交叉熵损失函数进行训练，优化器为AdamW，学习率设置为0.0001，训练迭代周期为20 epochs。我们保存了在验证集上UA + WA 最高的模型并在测试集上进行测试。除输入频谱图外的其它训练设置在2.3小节及第4章均保持一致。最终，基准模型CRNN在验证集上的UA为0.556，WA为0.552，测试集上的UA为0.554，测试集WA为0.571。

* 1. 性能提升
     1. 预训练模型

近年来，在大规模无标签数据集上进行预训练，再在下游数据集上进行微调的策略逐渐成为自然语言处理、计算机视觉等领域的主流学习范式。基于此，本文考虑使用在AudioSet数据集[3]上进行音频预训练的网络模型在IECOMAP数据集上进行相应下游任务的微调，所考虑的两类模型分别为CNN类的网络模型PANNS[4]及Transformer类的网络模型HTSAT[5]，

1. PANNS，为AudioSet预训练的CNN模型，上游堆叠而成的ConvBlocks（Conv + BN + ReLU + Pooling）提取音频特征，并在时域与频域进行下采样，通过Global Temporal Pooling将时域特征汇聚为整个音频的表征向量，再将该表征向量送入分类头进行分类。本文选择了论文提供的CNN10、CNN14两类架构，相较于前者，后者的参数规模更大，二者的网络架构如下图所示，

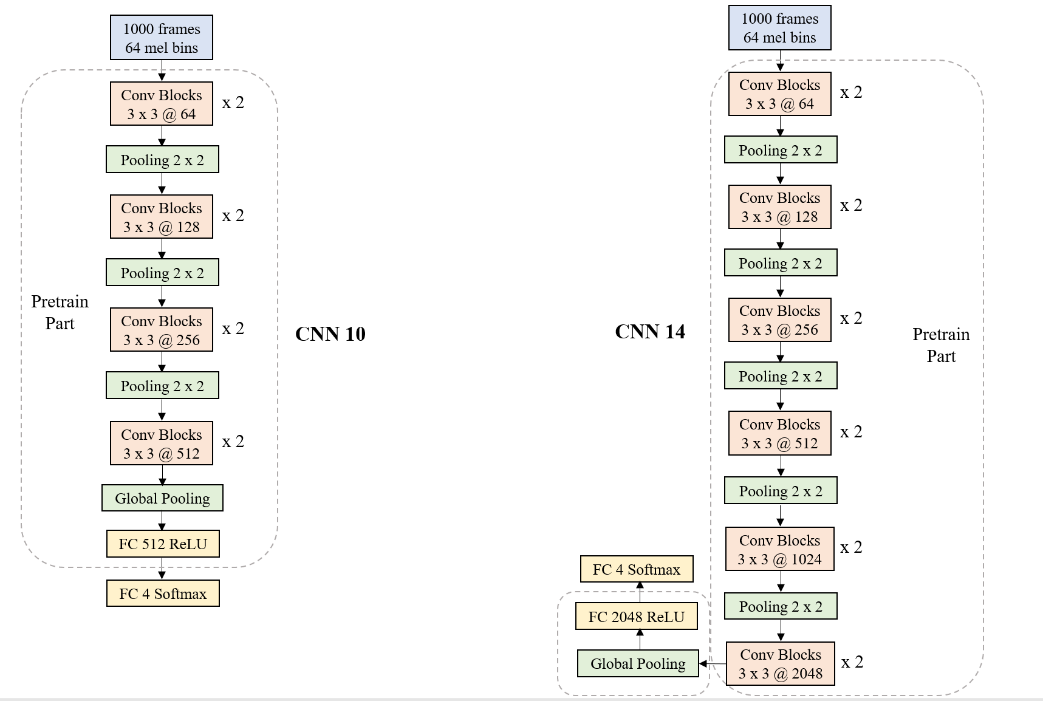


图3 CNN10与CNN14网络架构示意图

其中，CNN10的hop size 为320，window length为1024，sample rate 为32k；CNN14的hop size 为512，window length为160，sample rate 为16k，二者均采用64 个mel bins。

1. HTSAT，为AudioSet预训练的Transformer模型，先通过Patch Embed将频谱图转成Patch序列，上游堆叠而成的SwinBlocks提取层次化的帧级别的音频特征，之后帧级别特征被送入分类器中得到帧级别概率，再由平均池化层得到音频片段级别的概率，HTSAT的特征参数同CNN10，其网络架构示意图如下，其中T、F分别代表输入频谱图在时域、频域的尺寸，P为Patch化时的Patch大小，D为Patch的特征维度，最后输出的Label Prediction即为C个标签的置信分数，HTAST的网络架构示意图如下，

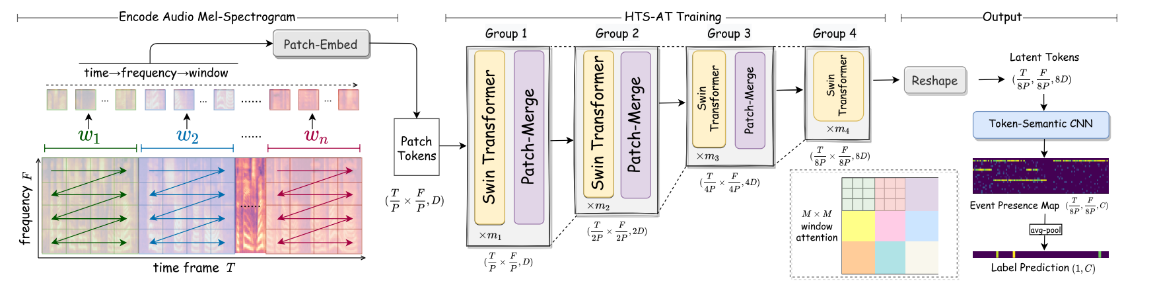


图4 HTAST网络架构示意图

我们分别应用上述预训练模型，比较是否加载预训练权重对于模型性能的影响如下表所示（汇报了模型在验证集及测试集的各项指标，指标计算在十次试验下取平均值），

表1 大规模预训练对于模型性能的影响

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Pre-train | Val UA | Val WA | Test UA | Test WA | Params |
| CRNN |  | 0.556 | 0.552 | 0.554 | 0.571 | 0.4M |
| CNN10 |  | 0.578 | 0.588 | 0.587 | 0.634 | 6.1M |
| CNN10 | √ | 0.654 | 0.651 | 0.628 | 0.659 | 6.1M |
| CNN14 |  | 0.452 | 0.417 | 0.466 | 0.423 | 84.8M |
| CNN14 | √ | 0.685 | 0.682 | 0.636 | 0.661 | 84.8M |
| HTSAT |  | 0.554 | 0.548 | 0.510 | 0.549 | 29.2M |
| HTSAT | √ | 0.659 | 0.664 | 0.631 | 0.658 | 29.2M |

从上表看出，当不加载预训练模型时，参数量较小的CRNN、CNN10均可以在该任务上取得不错的结果，而参数量极大的CNN14则由于过拟合表现极差。当加载预训练参数后，CNN10、CNN14、HTSAT等模型的性能均得到了较大提升，此时，参数量较大的HTSAT、CNN14模型的表现要略胜于CNN10模型，其中，以CNN14模型的分类表现最佳。

* + 1. 数据增强策略

为了进一步提升语音情感识别系统的稳定性并扩充数据规模，本文借鉴了两类音频数据增强策略：SpecAugment[6]与FilterAugment[7]，前者主要针对频谱图的时域与频域信息进行随机掩膜，后者对特定频段的信息进行扰动，增强的超参数选择与原文保持一致。除此之外，本文采用Mixup[8]来鼓励模型做出插值一致性的预测，具体地，对于两样本 、，假设其对应标签为 、，交叉熵损失函数为 ，混合系数 采样于Beta分布，模型映射以 表示，则Mixup损失函数为

我们分别应用上述数据增强策略于上一小节采用的CNN10模型（模型加载预训练参数），比较其对于模型性能的影响如下表所示（汇报了模型在验证集及测试集的各项指标，指标计算在十次试验下取平均值），

表2 数据增强策略对于模型性能的影响

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | SpecAug | FilterAug | Mixup | Val UA | Val WA | Test UA | Test WA |
| CNN10 |  |  |  | 0.654 | 0.651 | 0.628 | 0.659 |
| CNN10 | √ |  |  | 0.672 | 0.657 | 0.635 | 0.661 |
| CNN10 |  | √ |  | 0.651 | 0.652 | 0.608 | 0.640 |
| CNN10 |  |  | √ | 0.654 | 0.654 | 0.630 | 0.671 |
| CNN10 | √ |  | √ | 0.670 | 0.661 | 0.638 | 0.664 |
| CNN10 |  | √ | √ | 0.638 | 0.648 | 0.605 | 0.646 |

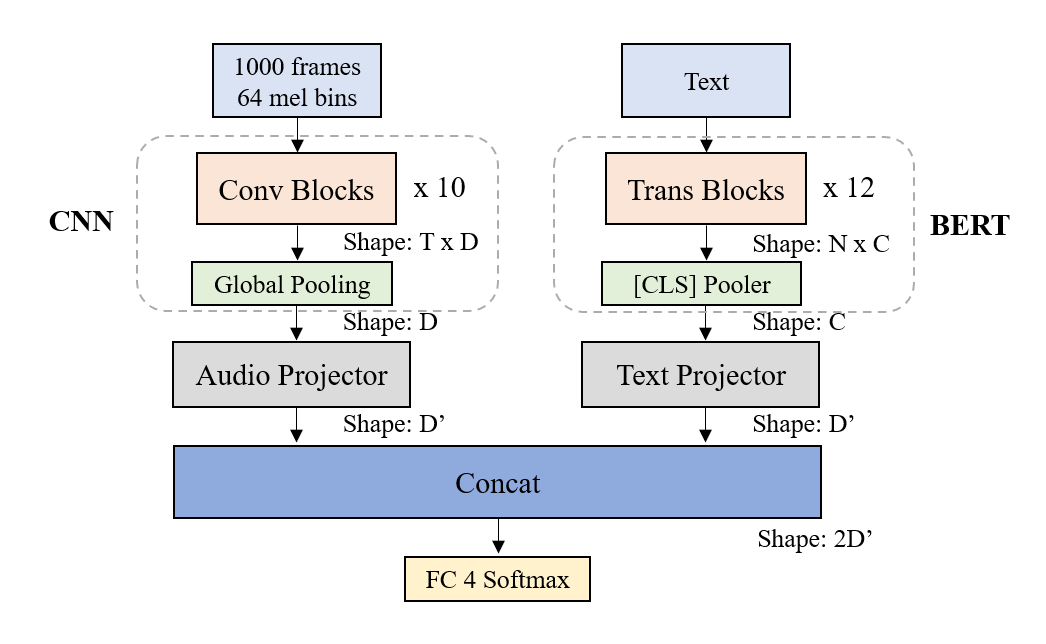
从上表看出，SpecAugment与Mixup两类增强方法均可以提升模型性能，而将FilterAugment应用于模型训练时则会引起模型在测试集上的性能较大幅度的下降。

1. 基于语言模态的语音情感识别
2. 基于多模态的语音情感识别
   1. 网络架构

采用双流网络架构，音频编码器由第2章CNN模型（如CNN10模型）构成，语言编码器由第3章BERT模型（如BERT-Tiny模型）构成，二者提取的特征经过特征投影及融合机制融合后送入分类头进行分类。具体地，考虑前期融合、注意力机制融合两类融合策略。

* 1. 融合策略
     1. 前期融合

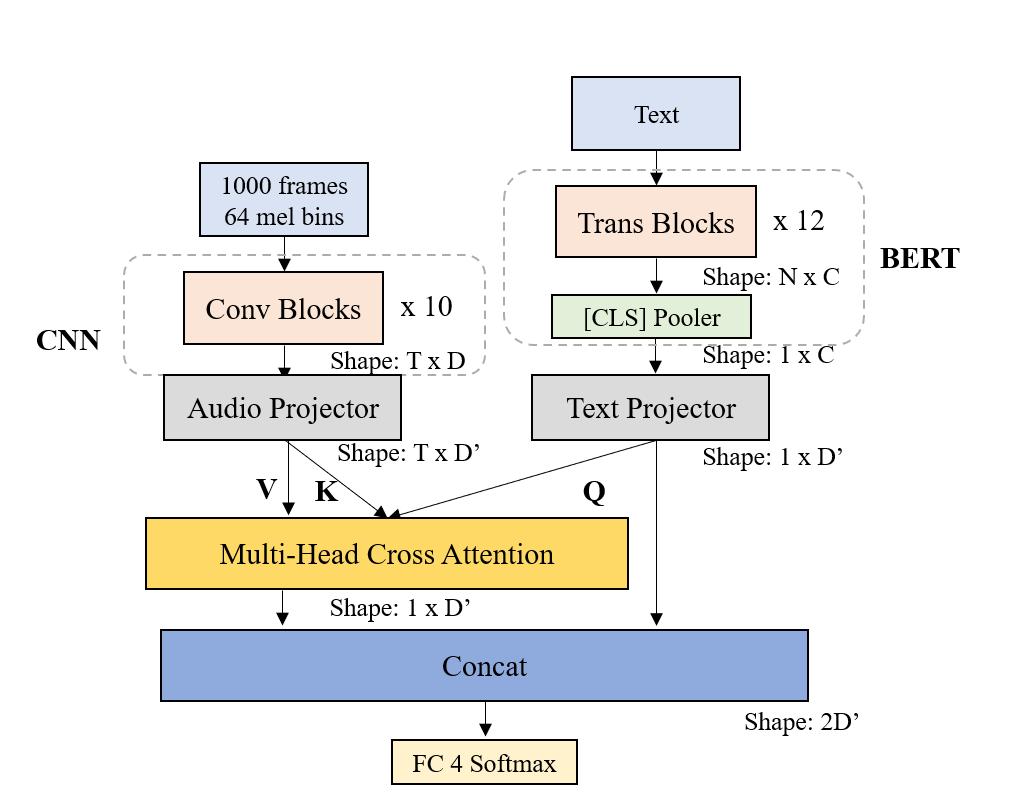
CNN音频编码器通过堆叠的卷积层及最后的Temporal Pooling层得到音频的D维表征，BERT语言编码器的CLS Token Embedding 也可以看作文本的C维表征，二者分别经过各自的Projector（线性层）将表征投影到D’维空间，再将投影后的向量进行拼接，作为多模态表征送入分类头中，前期融合的原理示意图如下，



图X多模态特征前期融合原理示意图

* + 1. 注意力融合

CNN音频编码器通过堆叠的卷积层得到帧级别的特征（不经过Temporal Pooling），BERT语言编码器的CLS Token Embedding 也可以看作文本的C维表征。此时，将BERT输出的CLS Token Embedding作为交叉多头注意力的Q，而将CNN输出的帧级别特征作为K、V，此时相当于利用文本特征对音频模态的帧级别特征进行加权而不是直接取平均或者最大值，注意力融合的原理示意图如下，



图X多模态特征注意力融合原理示意图

利用第2、3章提出的单模态模型进行组合验证，为了匹配各模态的特征提取能力，将CNN10模型与BERT-Tiny搭配使用，比较其对于模型性能的影响如下表所示（汇报了模型在验证集及测试集的各项指标，指标计算在十次试验下取平均值），

表X多模态特征融合策略对于模型性能的影响

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| A-Model | T-Model | Fusion | Val UA | Val WA | Test UA | Test WA |
| CNN10 | - | - | 0.654 | 0.651 | 0.631 | 0.659 |
| - | BERT-Tiny | - |  |  |  |  |
| CNN10 | BERT-Tiny | Early | 0.707 | 0.722 | 0.714 | 0.730 |
| CNN10 | BERT-Tiny | Attention | 0.709 | 0.726 | 0.723 | 0.741 |

从上表可以看出，相较于单模态，多模态模型由于可以同时融合多种来源的信息，在性能上普遍胜过各个单模态。此外，基于注意力的多模态特征融合优于早期融合，原因可能是其更多地考虑了模态之间的交互而不是简单地使用模态特征的拼接。

1. 多模态情感识别系统
2. 参考文献

[1] El Ayadi M, Kamel M S, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases[J]. Pattern recognition, 2011, 44(3): 572-587.

[2] Busso C, Bulut M, Lee C C, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. Language resources and evaluation, 2008, 42: 335-359.

[3] Gemmeke J F, Ellis D P W, Freedman D, et al. Audio set: An ontology and human-labeled dataset for audio events[C]//2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017: 776-780.

[4] 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.

[5] Chen K, Du X, Zhu B, et al. HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 646-650.

[6] Park D S, Chan W, Zhang Y, et al. Specaugment: A simple data augmentation method for automatic speech recognition[J]. arXiv preprint arXiv:1904.08779, 2019.

[7] Nam H, Kim S H, Park Y H. Filteraugment: An acoustic environmental data augmentation method[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 4308-4312.

[8] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization[J]. arXiv preprint arXiv:1710.09412, 2017.