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](#_Toc132461366)

[1.1 任务概览 1](#_Toc132461367)

[1.2 数据描述 1](#_Toc132461368)

[1.3 评价指标 2](#_Toc132461369)

[2 基于音频模态的语音情感识别 2](#_Toc132461370)

[2.1 数据处理 2](#_Toc132461371)

[2.2 网络架构 3](#_Toc132461372)

[2.3 性能提升 4](#_Toc132461373)

[2.3.1 预训练模型 4](#_Toc132461374)

[2.3.2 数据增强策略 6](#_Toc132461375)

[3 基于文本模态的语音情感识别 7](#_Toc132461376)

[3.1 数据处理 7](#_Toc132461377)

[3.2 传统方法 7](#_Toc132461378)

[3.2.1 线性回归 7](#_Toc132461379)

[3.2.2 Logistic回归 8](#_Toc132461380)

[3.2.3 朴素贝叶斯 9](#_Toc132461381)

[3.2.4 决策树 10](#_Toc132461382)

[3.2.5 随机森林 10](#_Toc132461383)

[3.2.6 支持向量机 11](#_Toc132461384)

[3.3 深度学习方法 11](#_Toc132461385)

[3.3.1 TextCNN & BiLSTM 12](#_Toc132461386)

[3.3.2 预训练模型 13](#_Toc132461387)

[3.3.3 Prompt Tuning 14](#_Toc132461388)

[3.4 文本情感分析实验 15](#_Toc132461389)

[4 基于多模态的语音情感识别 15](#_Toc132461390)

[4.1 网络架构 15](#_Toc132461391)

[4.2 融合策略 15](#_Toc132461392)

[4.2.1 前期融合 15](#_Toc132461393)

[4.2.2 注意力融合 16](#_Toc132461394)

[5 多模态情感识别系统 17](#_Toc132461395)

[6 参考文献 18](#_Toc132461396)

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，

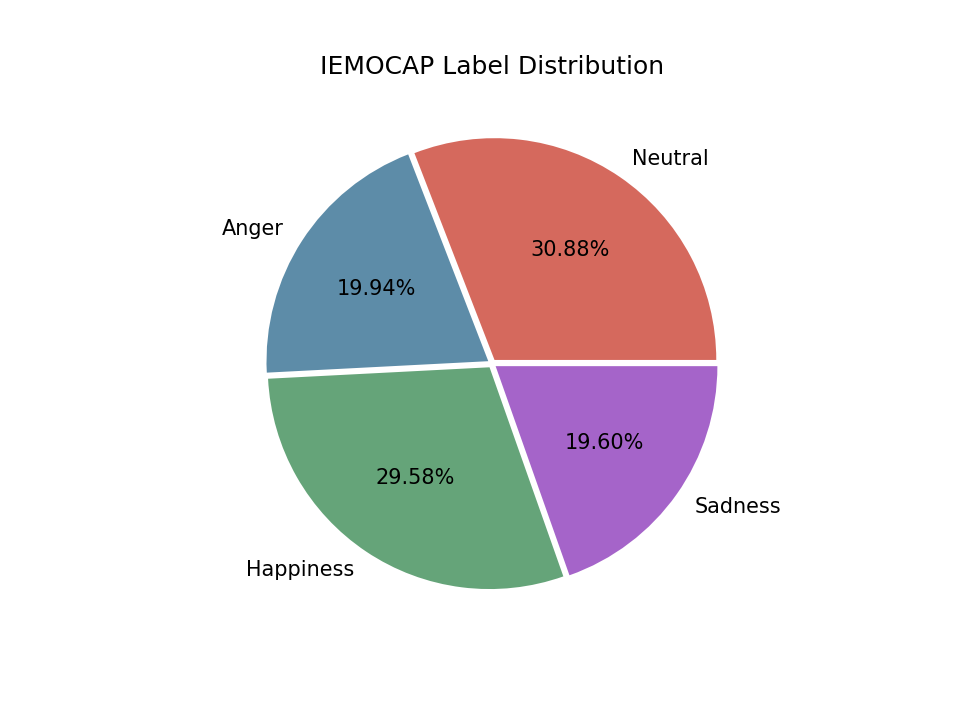


图1 数据集标签分布情况

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，

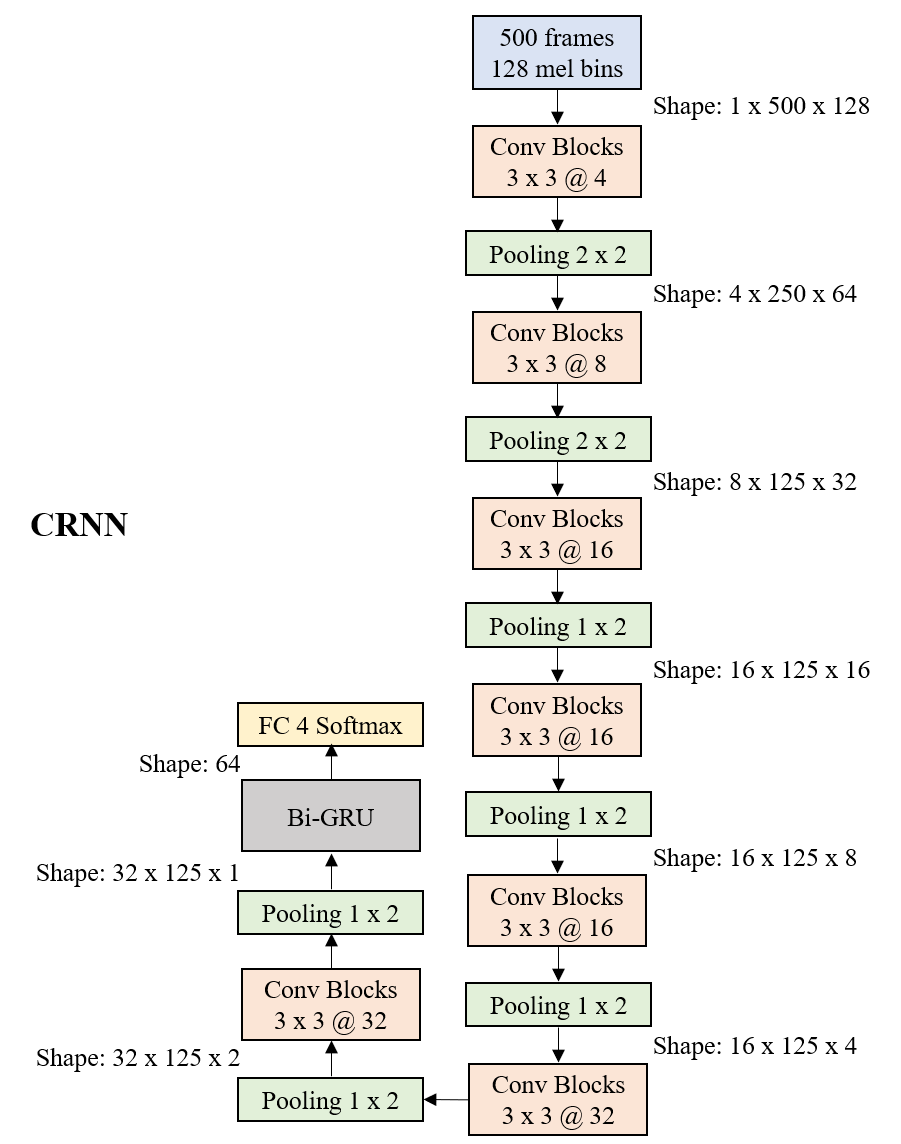


图2 CRNN网络架构示意图

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，

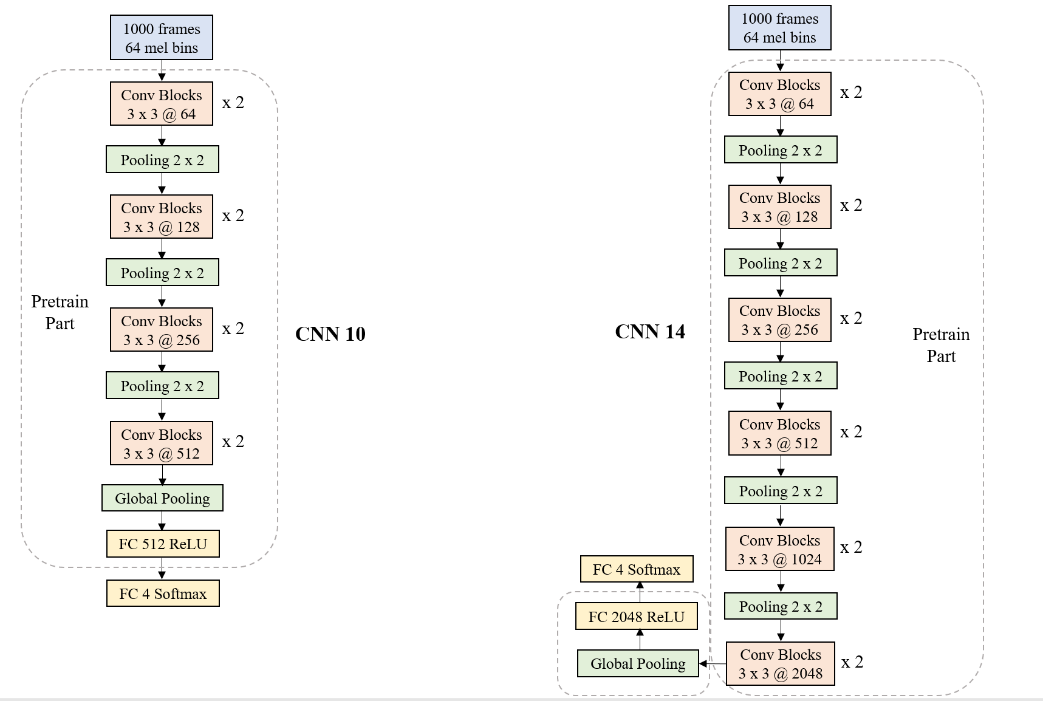


图3 CNN10与CNN14网络架构示意图

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，

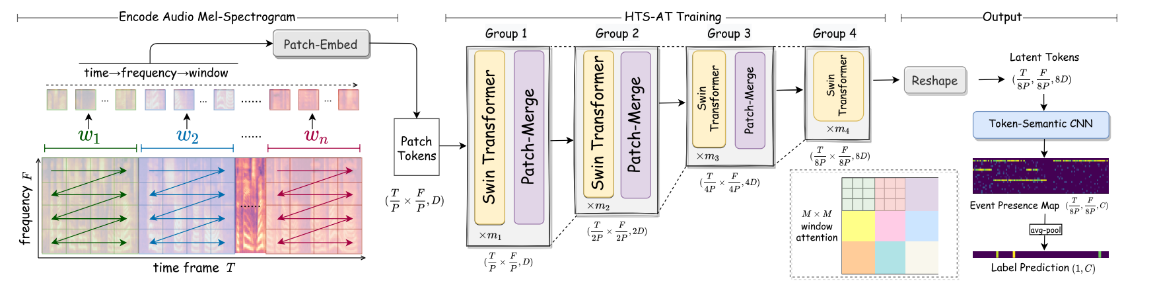


图4 HTAST网络架构示意图

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. 基于文本模态的语音情感识别
   1. 数据处理

IEMOCAP数据集文本数据是在每一场景下音频的内容进行转录而得到的，包括两个人物之间的对话内容，包括语音转写、标点符号、发音注释等信息。这些文本数据通常包含了演员的愤怒、快乐、悲伤等情感状态。摘取第二章中音频文件处理后对应的文本片段作为文本数据。

对于文本模态（此处写数据预处理，bert不咋需要留给兆哥了）

* 1. 传统方法

1. 将文本特征向量化:将文本转换为一组数字特征,如词频特征、TF-IDF特征等。每个特征对应文本的一个属性。

2. 构建标签:对训练集中的每个文本,手动标注其情感极性,如正面、负面、中性。这些标注作为训练数据集的标签。

3. 训练线性回归模型:使用特征向量和标签训练一个线性回归模型。模型会学习每个特征与情感标签的线性关系,得到一组权重。

4. 预测新文本情感:当有新文本需要分类时,先提取其特征向量。然后将特征向量输入线性回归模型,模型输出的预测结果就是该文本的情感分类结果。

5. 改进模型:如果模型的预测效果不佳,可以改进特征工程以更好代表文本属性,调整模型超参数等来提高模型性能。

线性回归之所以可以用于情感分类,是因为它假定情感与文本特征之间存在线性关系。如果增加正面情感的特征,如积极词的使用,那么文本的正面概率也会相应增加。线性回归可以学习这种线性关系,并在新文本上作出比较准确的情感预测。

* + 1. 线性回归

线性回归是一种最简单的机器学习算法，假设特征与标签之间存在线性关系。线性回归的模型表达为：

其中，到是个特征，到是每个特征对应的权重，是模型的偏置。对任意一个特征，如果其权重为正，那么对有正相关影响；如果为负，那么对有负相关影响。权重的大小表示该特征对模型预测的重要性。对于本次的任务来说，特征即为通过TF-IDF提取的文本特征，标签即为各类别的实际标签。

线性回归旨在找到一组权重和偏置，可以使得模型对训练数据的预测与真实标签尽量接近。这是一个优化问题，常用的损失函数是均方误差，通过梯度下降等方法不断调整和的值以最小化损失函数，从而得到最优的模型。

线性回归方法比较简单，容易理解，且计算效率高；缺点是假定了线性关系，因此无法处理非线性问题，学习能力有限，且容易过拟合，对异常值比较敏感。对于本次任务来说，由于标签是离散的值，因此只能将经过模型后输出的值与标签绝对距离进行比较，实际上线性回归并不适用于本次的任务。

* + 1. Logistic回归

逻辑回归是一种广泛使用的机器学习算法，用于解决二分类问题，也可以进行扩展从而解决多分类问题。

逻辑回归模型表达为：

其中，到是n个特征，到是每个特征对应的权重，是模型的偏置。

逻辑回归训练的目的是最大化训练数据被正确分类的概率。它采用最大似然估计法，通过梯度上升等优化方法不断更新权重和的值，使模型对训练数据的分类结果与真实标签尽可能吻合。

多类逻辑回归模型的原理如下：

1. 假设类别之间是互斥的，即每个实例只能属于一个类别。对每个类别,引入表示该实例属于类别的概率。

2. 为每个类别构建一个逻辑回归模型,预测属于该类别的概率。具体而言:

这里是类别对应的权重向量。

3. 对新输入，在每个逻辑回归模型中计算属于对应类别的概率，然后选择概率最大的类别作为预测类别。

多类逻辑回归的优点在于原理简单，易于理解和实现，并且可以在不加修改的情况下，扩展到更多类别。缺点是无法学习类别之间的相关性，且计算量比较大，训练代价高。

* + 1. 朴素贝叶斯

朴素贝叶斯是一种简单实用的机器学习算法，用于解决分类问题。原理如下：

1. 假设特征之间相互独立。这是朴素贝叶斯算法的核心假设，该假设简化了模型使其易于计算。

2. 根据贝叶斯定理计算后验概率。对于某个类别和特征向量，贝叶斯定理为：

其中是后验概率，是联合概率，是先验概率

3. 构建分类器。对于新输入，计算它属于每个类别的后验概率，然后预测类别为先验最大的那个类别。

朴素贝叶斯模型表达为：

其中到是个特征，是类别的先验概率，是特征在类别下的条件概率。

朴素贝叶斯的“朴素”是因为它假定特征相互独立。这个假设简化了模型,使其易计算,但也限制了其表达能力。然而即便特征之间存在依赖,朴素贝叶斯仍然表现良好。这是因为贝叶斯定理在分类效果上更为重要。朴素贝叶斯通过训练统计特征-类别共现频次，获得先验概率和条件概率。预测阶段直接应用贝叶斯定理,以获得每个类别的后验概率。

朴素贝叶斯的优点是计算简单，易理解且易实现。缺点是其对属性独立的假设限制了模型的表达能力，且对小训练集表现不佳。

* + 1. 决策树

决策树是一种流行的机器学习算法，用于解决分类和回归问题。其原理如下：

1. 以树形结构表示决策过程。决策树由根节点、内部节点(非叶子节点)和叶子节点组成。

2. 根节点代表完整训练数据集。内部节点表示数据的某个特征，每个子节点代表该特征的一个取值。

3. 叶子节点代表数据集的分类结果或回归结果。分类树的叶子节点包含类别标签，回归树的叶子节点包含连续值。

4. 决策树通过递归地对数据进行切分，直到达到停止条件，逐步生成。

5. 对新数据实例，从根节点开始，依据其特征取值递归地向下移动，最终到达叶子节点，获得预测结果。

决策树主要学习算法有：ID3、C4.5、CART。其中CART构建决策树通过贪心算法，按照特征选择和数据切分的方式生成二叉树。其具体步骤为：

1. 选择最优特征和切分点：计算各个特征和各个切分点的代价，选择代价最小的特征和切分点进行切分。

2. 数据切分：将节点的数据切分成子节点，根据最优切分点的取值将数据划分到左子节点或右子节点。

3. 生成子节点：为左右子节点重复以上步骤，直到达到停止条件。

4. 剪枝：采用CART算法可以剪去一些子节点，防止决策树模型过拟合。

决策树算法结果易于理解，对异常值不太敏感。但可能导致过拟合，学习结果随训练数据的变化而变化大。决策树是最简单,使用最广的机器学习算法之一。

* + 1. 随机森林

随机森林是一种流行的机器学习算法，用于解决分类和回归问题。它的原理如下：

1. 随机森林由多个决策树组成。每个决策树独立进行训练，最后将多个决策树的结果综合得到最终输出。

2. 随机森林使用bagging的思想，通过有放回地从训练数据中抽样获得子数据集来训练每个决策树。

3. 构建每个决策树时，随机选取特征子集。这增加了随机森林的多样性，避免过拟合。

4. 对新输入实例，将其输入到所有的决策树，并统计各类结果出现的次数，最终预测出现次数最多的类。

由于数据、特征的随机抽取，每棵决策树不同，随机森林具有较强的多样性，从而减少了过拟合的风险。随机森林准确性高、对异常值鲁棒、易于理解和实现。缺点是计算开销大、训练时间长。

* + 1. 支持向量机

支持向量机（SVM）是一种用于解决分类和回归问题的机器学习算法。它的原理如下：

1. SVM通过寻找超平面将不同类的训练数据分开，并且使得该超平面与最近的训练数据点之间的距离尽可能大。

2. SVM使用核技巧隐式地将数据映射到高维空间，使数据在高维空间线性可分，而在原始空间不可分。常用有线性核、多项式核和RBF核等。

3. SVM使用最大间隔分类器，寻找能将不同类的训练数据分开的超平面，使得该超平面与最近的数据点的距离最大。

4. SVM因此得到一个分类器，能够对新数据判断类别。对于非线性可分数据集，SVM可以通过核技巧映射得到其在高维空间的表现，仍然使用最大间隔分类器构建超平面，从而实现非线性分类。

SVM理论基础坚实、泛化能力强，通过核技巧可处理非线性问题，对中间空间数据不敏感。缺点是核的选择和超参数的调优比较困难。

* 1. 深度学习方法
     1. TextCNN & BiLSTM

文本情感分类是一项重要的自然语言处理任务，常常需要利用深度学习技术进行处理。在这些深度学习技术中，TextCNN和双向长短期记忆网络（BiLSTM）已被证明可以有效地应用于文本情感分类任务。

1. TextCNN：TextCNN 是一种具有多通道输入、多尺度卷积和全局最大池化等特征的卷积神经网络模型。它能够充分利用文本的局部和全局信息，并且能够高效地捕捉文本中的关键特征。因此，在情感分类任务中，TextCNN 能够有效地提高分类准确率和鲁棒性，并且具有良好的效率和可扩展性。TextCNN的网络架构示意图如下：

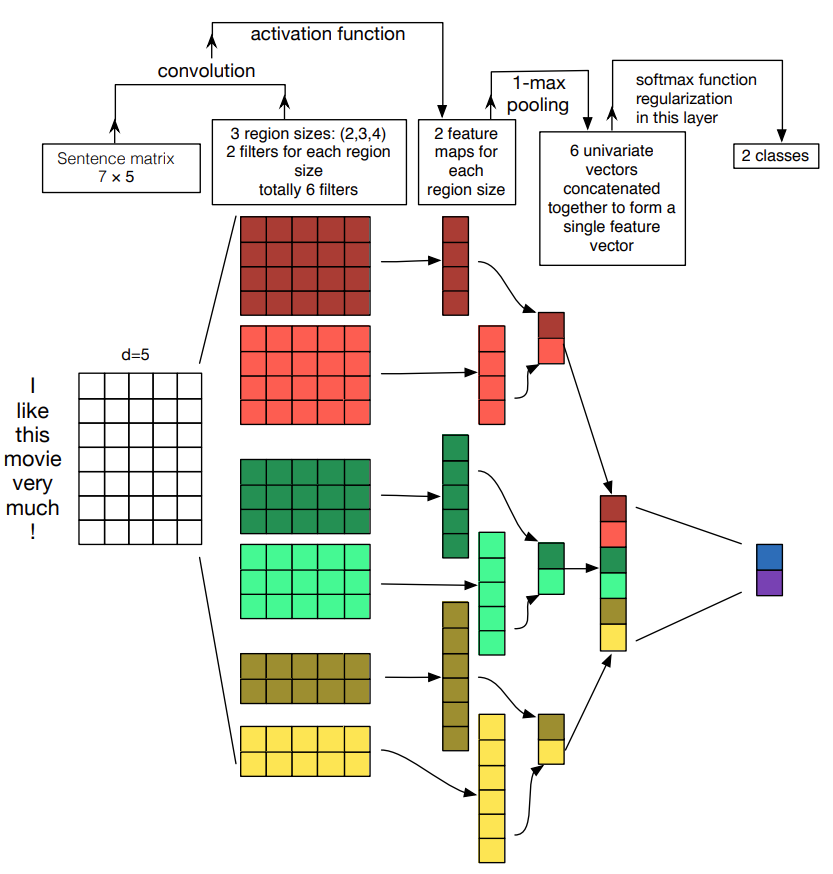


图5 TextCNN模型示意图

1. BiLSTM：BiLSTM 是一种基于循环神经网络（RNN）的模型，在处理时序数据上表现出色。BiLSTM 能够通过正向和反向传递信息，充分利用文本的上下文信息，并且具有优秀的建模能力。在情感分类任务中，BiLSTM 能够更好地处理长文本序列，并且对上下文信息的利用更加精细和全面，因此可以取得更好的分类效果。BiLSTM的网络架构示意图如下：

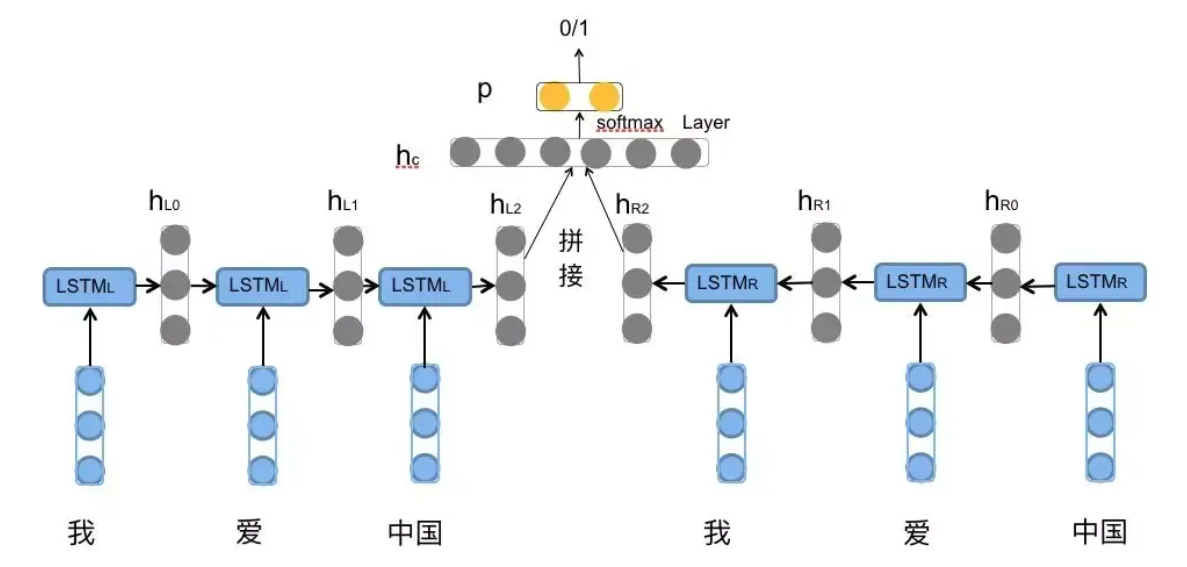


图6 BiLSTM模型示意图

TextCNN通过预训练的词向量模型得到文本数据的词向量矩阵，在本文中包含中文和英文两种语言的数据集，分别采用“tencent-ailab-embedding-zh-d200”和“glove.42B.300d”作为预训练词向量模型。

* + 1. 预训练模型

预训练模型是指在大规模无标签语料库上进行预训练，学习到通用的语言知识和表征能力，然后针对具体的下游任务（如文本分类、命名实体识别等），进行有监督的微调。

Bert 是一种基于 Transformer 架构的预训练语言模型，被广泛应用于自然语言处理任务中。在情感分类任务中，Bert 可以利用大规模语料库的上下文信息和语义特征，并且能够针对不同类型的情感分类任务进行微调，从而取得较好的分类效果。具体来说，Bert 首先使用无监督的方式进行预训练，通过掩码语言模型和下一句预测等任务，学习到了丰富的语言知识和表征能力。在情感分类任务中，Bert 能够充分利用文本的上下文和语义信息，对于长文本和复杂情感分类任务有很好的应用效果。此外，还可以通过多层抽取和特征融合等技术，进一步提高分类效果。Bert的网络架构示意图如下：

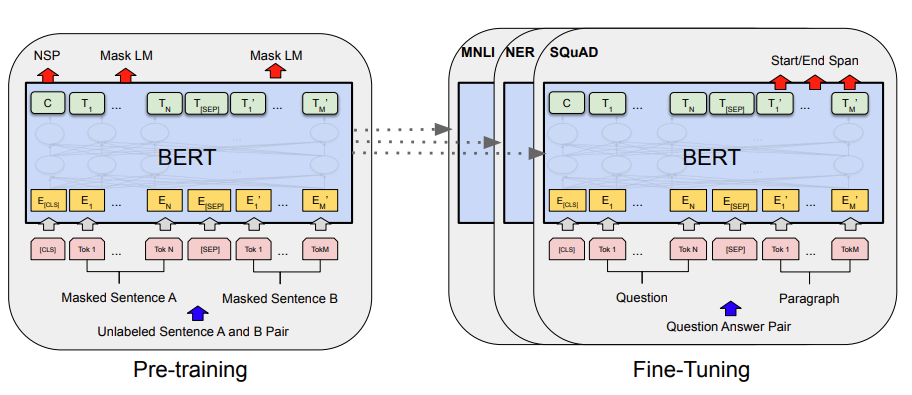


图7 Bert模型示意图

在 Bert 中，每一个文本序列都会被添加特殊标记 [CLS] 和 [SEP]，其中 [CLS] 标记表示整个文本序列的“类别”。在句子级别的情感分类任务中，可以将文本序列作为输入，同时在序列的开头添加 [CLS] 标记，并将该位置的输出向量作为整个句子的表示向量。在本文实验中，我们尝试对不同规模的Bert进行微调，在多个数据集上测试其效果。

* + 1. Prompt Tuning

Prompt 技术是一种基于语言模型的自然语言处理技术，它通过在输入文本中插入提示语（Prompt），来引导模型生成与指定任务相关的输出。Prompt 技术可以帮助解决少样本学习、零样本学习等问题，并且具有较好的可解释性和易用性。其中，Soft Prompt 是一种特殊的 Prompt 技术，它采用软提示方式对模型进行微调。与硬提示不同，Soft Prompt 不会强制指定任务的确切格式和内容，而是使用一组关键词或短语来引导模型生成相关的输出。具体来说，在训练时，Soft Prompt 可以将预定义的提示词或短语加入到输入序列中，作为模型学习的一部分。在推理时，模型将这些提示词或短语视为一种信号或信息，从而更好地完成特定的任务。Soft Prompt 的主要优点是其灵活性和可扩展性。相比硬提示，它能够更准确地匹配输入数据的多样性，从而提高模型的泛化能力。其中较为经典的Soft Prompt方法为P-Tuning v2。与之前的Soft Prompt方法相比，P-Tuning v2 引入了多层Prompt的机制，同时采用了强化学习算法来优化Prompt的生成，提高了模型的适应性和泛化能力。P-Tuning v2的模型示意图如下：

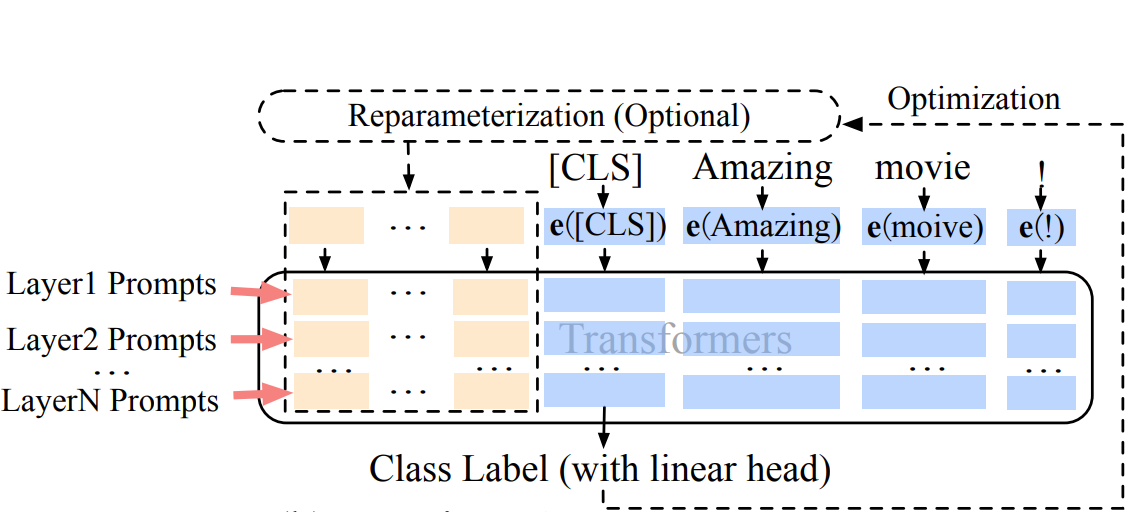


图8 P-Tuning v2的模型示意图

在文本实验中，我们设置Prompt token数为20，并且采用多层感知机（MLP）进行重参数化，提高模型效果。

* 1. 文本情感分析实验

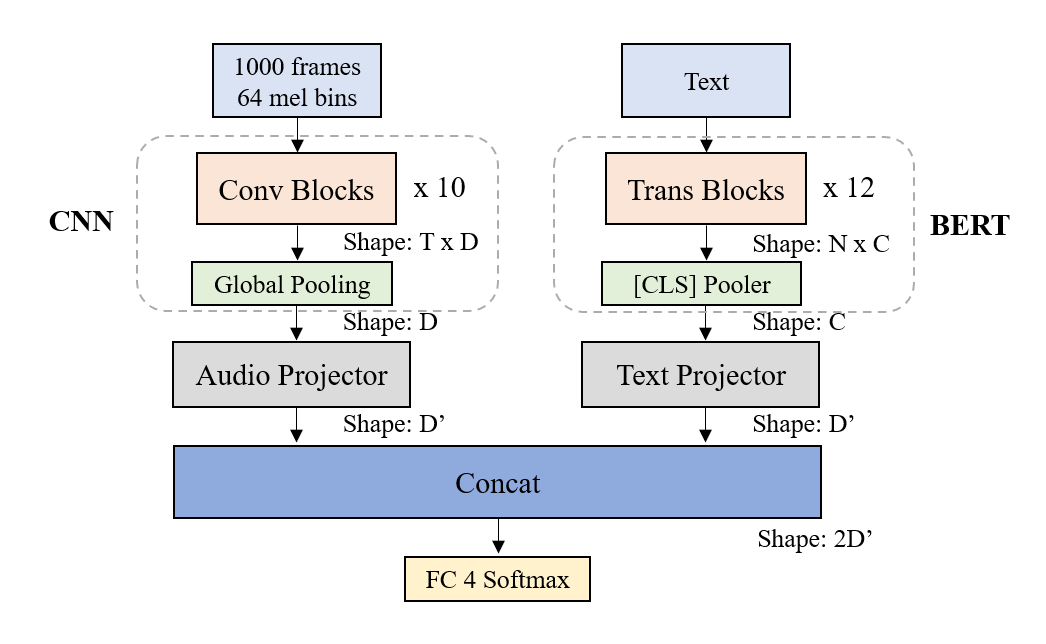
（缺一堆叙述）

1. 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

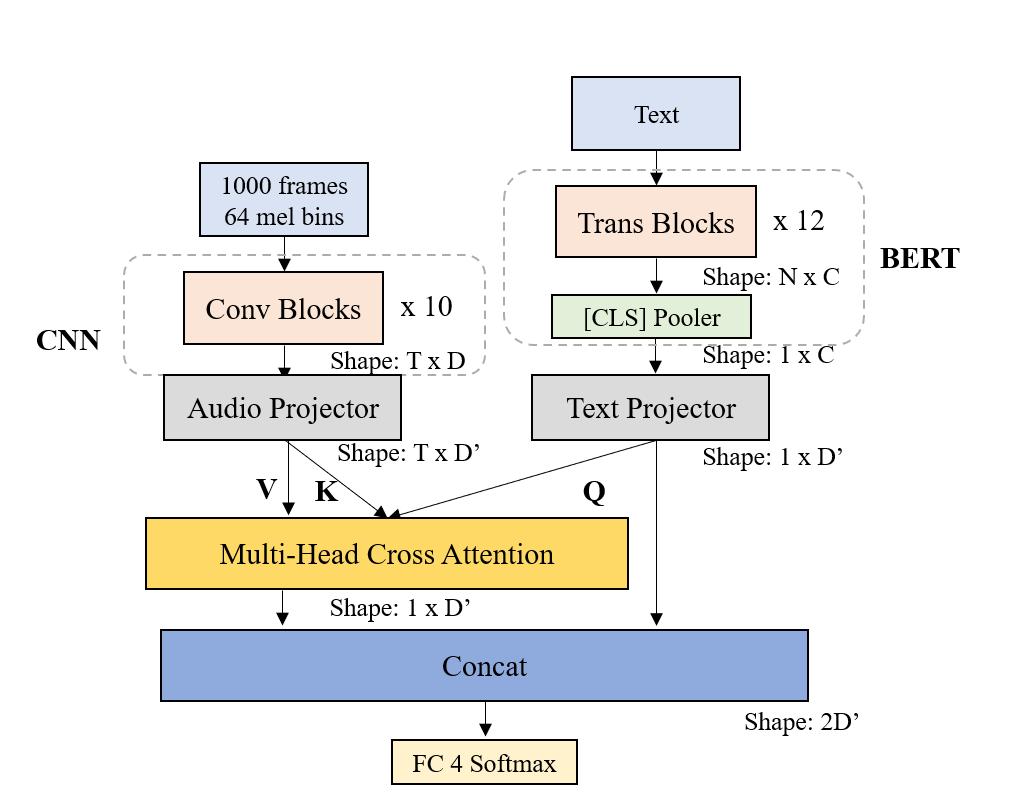
CNN音频编码器通过堆叠的卷积层及最后的Temporal Pooling层得到音频的D维表征，BERT语言编码器的CLS Token Embedding 也可以看作文本的C维表征，二者分别经过各自的Projector（线性层）将表征投影到D’维空间，再将投影后的向量进行拼接，作为多模态表征送入分类头中，前期融合的原理示意图如下，



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.