语音情感识别实验

国科大人工智能学院	2022K8009907013	林佳豪

实验目的

搭建一个具有GUI的、可即时录制音频上传的、可本地运行的纯语音情感识别系统。

实验方法

语音情感识别（Speech Emotion Recognition, SER）是语音信号处理与人工智能结合的一个重要研究领域，旨在通过分析音频数据提取人类情感特征。在实际应用中，SER系统可广泛用于人机交互、心理健康监测、客户服务等领域。随着深度学习的兴起，端到端的情感识别方法已逐渐取代传统的特征工程方法，成为研究热点。本实验中，我们希望尝试几种不同的端到端SER方法，利用神经网络直接从波形提取特征，而不借助于手工设计。我们分别使用卷积神经网络（CNN）与Wav2Vec+池化作为主干网络，先对不同长度的音频提取固定维度的整体特征，最后都用一层全连接网络进行分类。

CNN

卷积神经网络（CNN） 是一种常用于计算机视觉的神经网络结构，但在音频数据处理方面也具有较好表现。我们参考以往的纯卷积提取句子整体特征的思路(Asiful A. et al., 2019)，尝试了近似原文的模型结构，和一个轻量化的结构，分别是：卷积核尺寸21, 19, 17, 15, 13, 11, 9的七层1d卷积块，通道数分别为1, 32, 64, 128, 256, 512, 1024, 1024，最后进行全局最大池化，得到 `(B,1024)` 特征；轻量化后的卷积核尺寸19, 17, 15, 13的四层1d卷积块，通道数分别为1, 32, 64, 128, 256，最后进行全局最大池化，得到 `(B,256)` 特征。一个卷积块由一层1dCNN，后接ReLU激活函数、批次归一化、核尺寸2的最大池化构成。然后对最后的特征进行0.5的dropout处理，减小过拟合风险。

Wav2Vec

本实验使用了Wav2Vec 2.0模型，将情感识别作为一个分类任务，在IEMOCAP数据集上进行微调。

Wav2Vec (Steffen S. et al., 2019)是由Facebook AI开发的一种基于自监督学习的语音特征提取模型。它通过大规模无标签音频数据进行预训练，能够有效提取音频数据中的高级特征，并通过微调适应具体任务。Wav2Vec 2 (Alexei B. et al., 2020)输入是原始音频波形，通过多个卷积层和Transformer层直接对原始音频数据建模，提取在时间上序列性的特征。其借鉴了BERT的训练方式，在时间维度上随机选择一部分潜在表示进行掩码，并要求模型预测这些被掩码的部分，使用对比学习的方式，通过目标表示与掩码表示的对齐，学习音频数据中的上下文依赖关系。

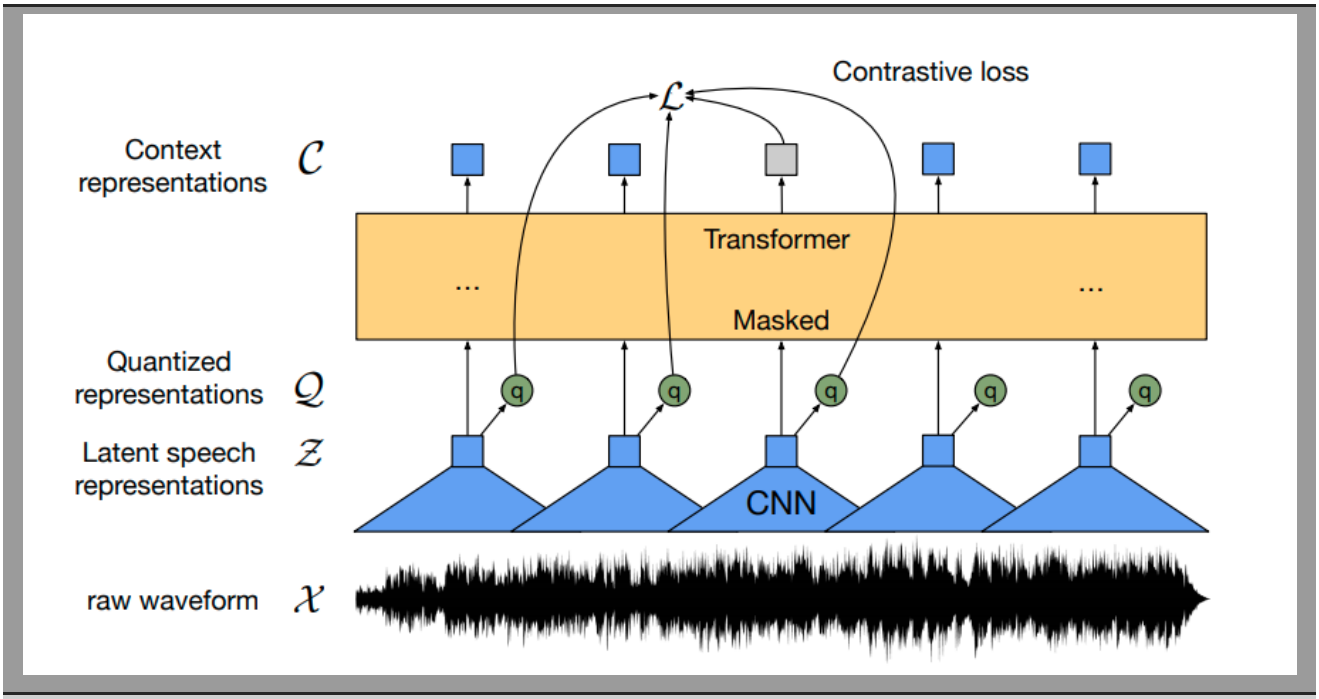


图1 Wav2Vec 2 结构

由于Wav2Vec的输出特征是时间序列，对于不同时常的语音长度不同，我们需要将其转化为维度不变的句子整体特征。我们尝试了在时间维度上进行全局最大池化、注意力池化和卷积-全局最大池化。设Wav2Vec输出的特征序列形状 `(B,D,T)`，分别为批次大小、特征维度、序列长度。全局最大池化即沿着序列长度对每个特征维度取最大值，得到 `(B,D)` 特征；注意力池化含有一个全连接层，对每个特征 `(D,)` 估计一个权重，然后对时间序列维度上的特征加权求和，得到 `(B,D)` 特征；卷积-全局最大池化先对特征序列进行1d卷积得到 `(B,D',T')` 特征，然后对这个特征进行全局最大池化，得到 `(B,D')` 特征。这些特征形状都和时间无关，可以直接通过全连接网络进行分类。

Gradio

Gradio 是一个 快速搭建交互式机器学习模型和演示 GUI 的 Python 库，广泛应用于展示、测试模型及收集用户反馈，特别适合本实验中用于实现可本地运行的、即时录制音频上传的语音情感识别系统。

代码结构

```
.
├── LICENSE                 # 许可证
├── README.md               # 项目说明文件
├── checkpoints/            # 模型检查点
├── config.py               # 模型配置文件
├── data_process.py         # 数据处理和数据集加载脚本
├── logs/                   # 日志文件
├── metrics.py              # 模型评估指标定义
├── models                  # 模型定义目录
│   ├── __init__.py         # 模型包初始化文件
│   ├── backbone.py         # 主干网络的选择
│   ├── cnn.py              # 卷积神经网络主干
│   ├── emotion_classifier.py # 情感分类器定义
│   └── wav2vec.py          # wav2vec2主干
├── train.py                # 训练模型
├── ui                      # Web界面相关代码目录
```

```
│   ├── __init__.py
│   ├── events.py          # 事件处理
│   └── ui.py              # UI界面定义
└── webui.py               # 启动Web界面
```

我们使用 transformers 的 Trainer 框架进行训练，使用 datasets 加载模型。修改 train.py 中的 config ，运行 train.py 进行对应模型训练。运行 webui.py 以打开SER图形界面系统。

实验数据

我们使用了Huggingface上的 IEMOCAP_Audio 数据集，并且使用 datasets.load_dataset 进行加载。IEMOCAP数据集是一个广泛用于语音情感识别的标准数据集，由南加州大学（USC）开发。该数据集包含5个不同的对话会话（Session），每个会话中包含两名演员（男女各一名）进行即兴或脚本化的情绪对话，包含语音音频、视频、文本转录以及身体动作捕捉数据。数据集的情感标签包括愤怒（Angry）、高兴（Happy）、中性（Neutral）、悲伤（Sad）等情绪类别。而 IEMOCAP_Audio 则是只取了其中的音频和标签。各个会话的数据规模如下：

```
DatasetDict({
    session1: Dataset({
        features: ['audio', 'label'],
        num_rows: 1085
    })
    session2: Dataset({
        features: ['audio', 'label'],
        num_rows: 1023
    })
    session3: Dataset({
        features: ['audio', 'label'],
        num_rows: 1151
    })
    session4: Dataset({
        features: ['audio', 'label'],
        num_rows: 1031
    })
    session5: Dataset({
        features: ['audio', 'label'],
        num_rows: 1241
    })
})
```

打印其中一个样本，是如下的字典，包含音频路径、采样数组、采样率和标签：

```
{'audio': {'path': 'Ses01M_impro06_M000.wav', 'array': array([0.00192261,
0.00247192, 0.00256348, ..., 0.00683594, 0.0065918 ,
       0.0055542 ]), 'sampling_rate': 16000}, 'label': 3}
```

为了训练和验证模型，我们将5个会话的数据集合并，随机按9:1的比例划分为训练集与验证集。由于本实验的目的是构建一个可行的语音情感识别系统，而非测试最佳配置性能，因此我们省略了测试集的构建。最终数据集规模如下：

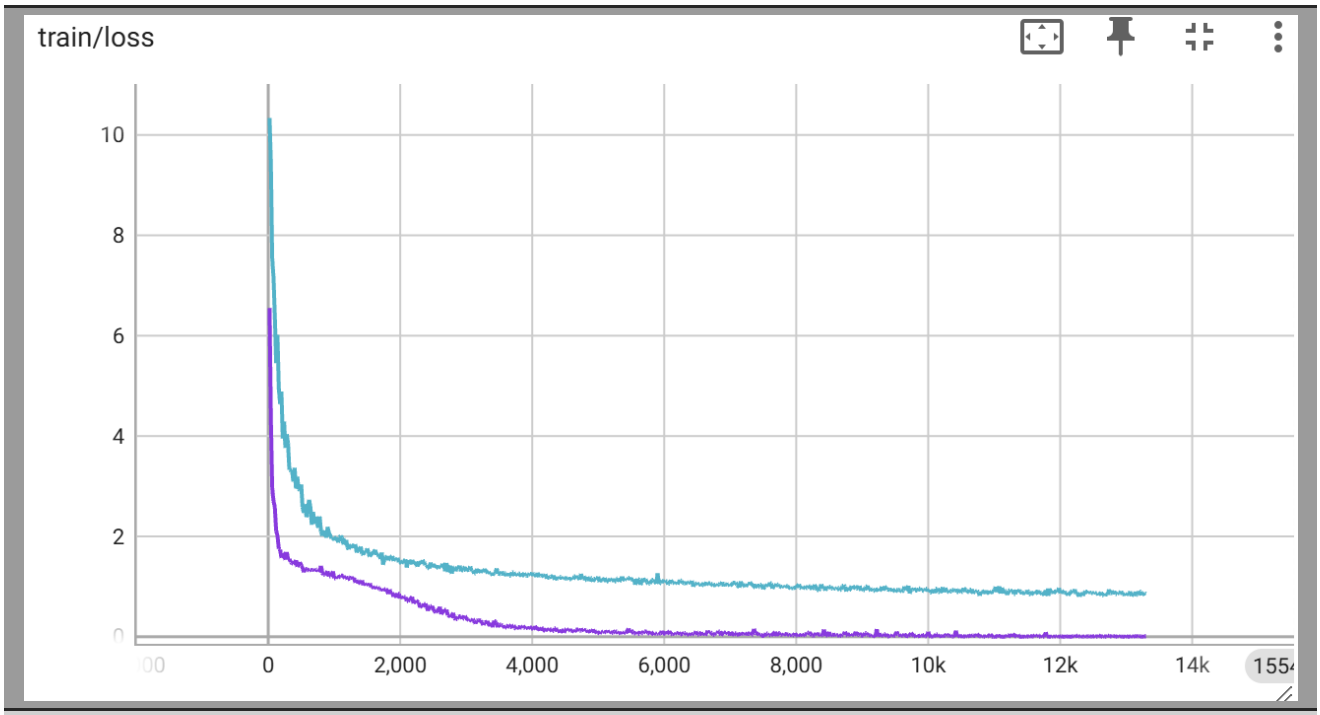
```
Train Size: 4977
```

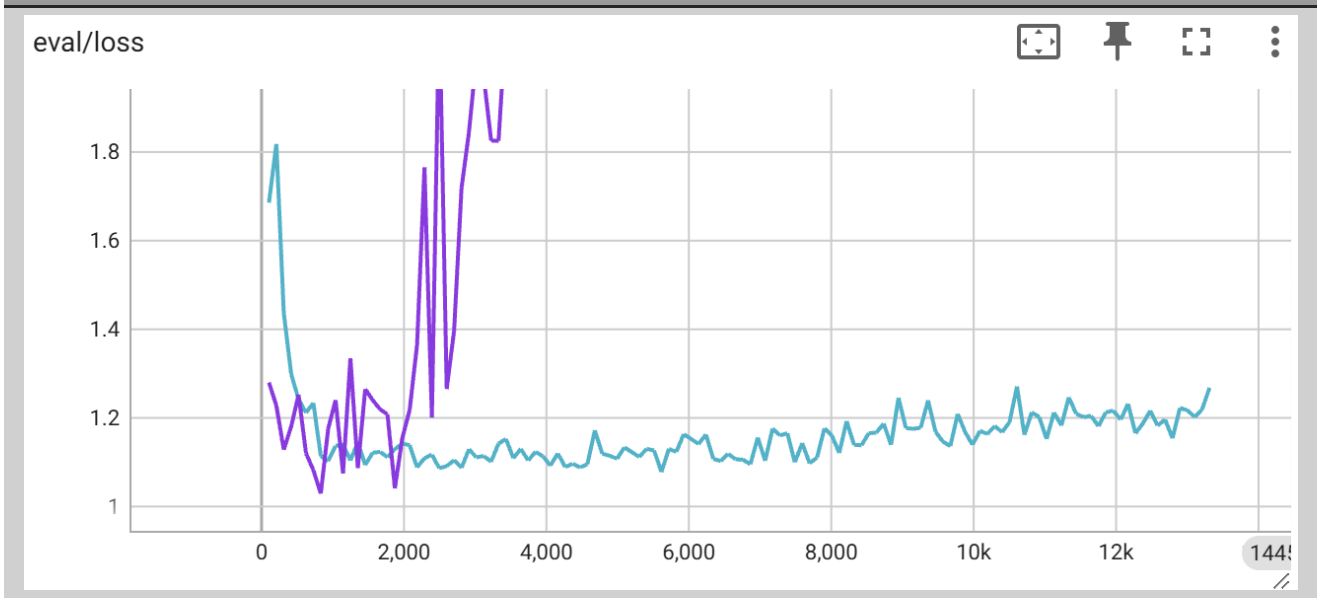
```
Eval Size: 554
```

实验结果

CNN

分别对两个配置的CNN进行训练，结果如下：





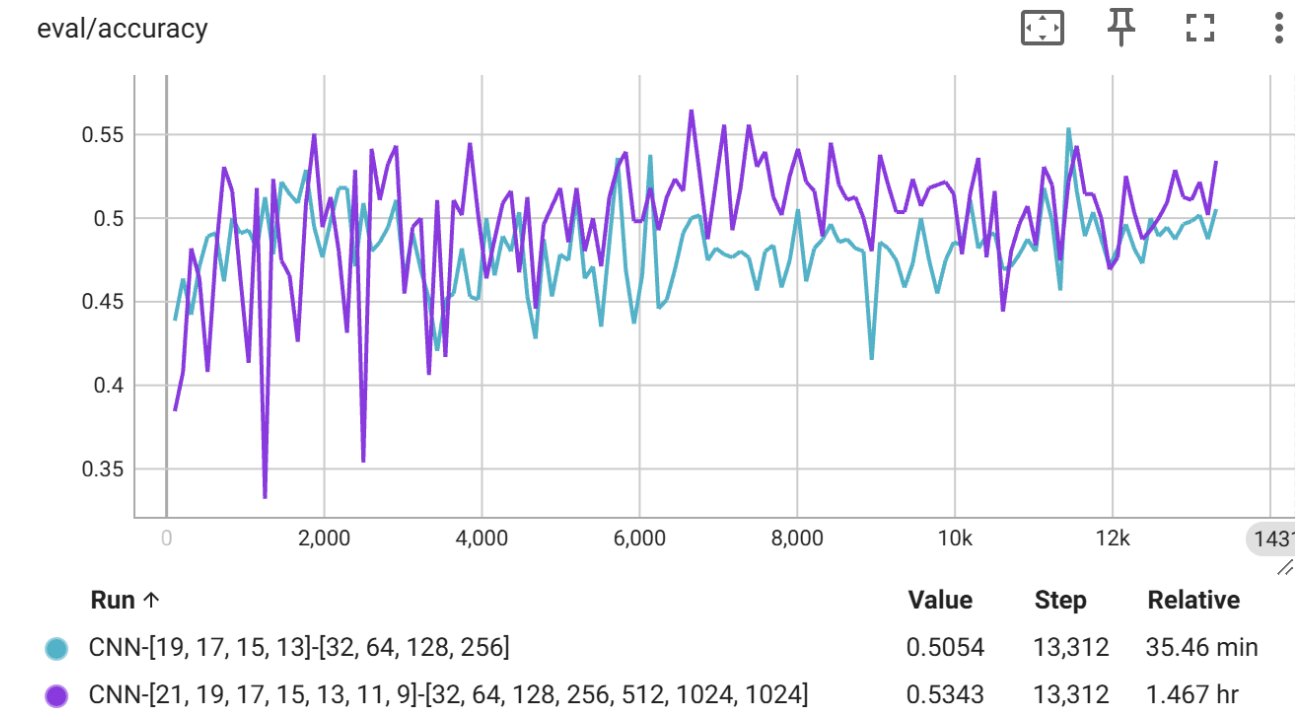
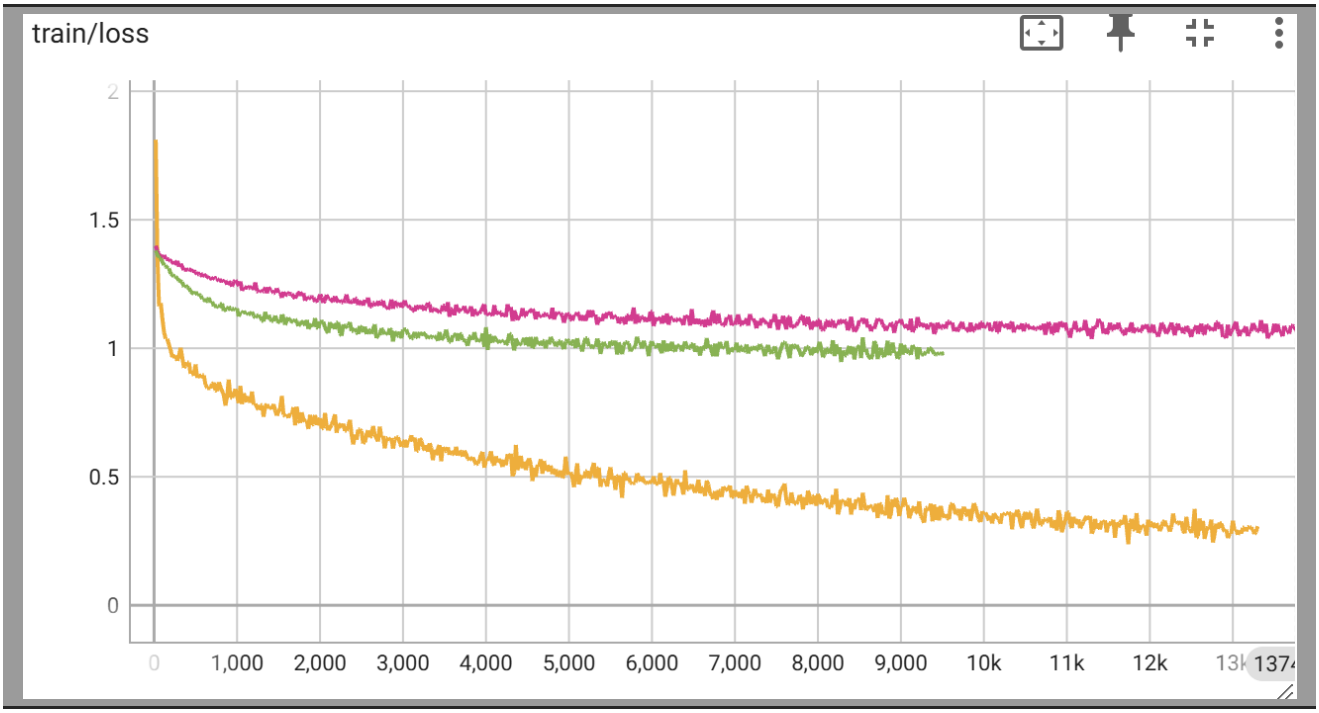


图2 CNN不同配置的的损失与准确率曲线

实验结果表明，4层CNN和7层CNN获得了相当的准确率。虽然4层CNN的训练损失比7层CNN更高，说明4层CNN拟合能力不足。但7层CNN在训练2000步之后出现了梯度爆炸和严重的过拟合现象，验证损失飙升，但是验证准确率受影响不大。最后导致7层的拟合能力没有完全发挥，二者最终4分类准确率都在50%上下。

Wav2Vec

分别用不同池化方式对Wav2Vec进行微调，结果如下：



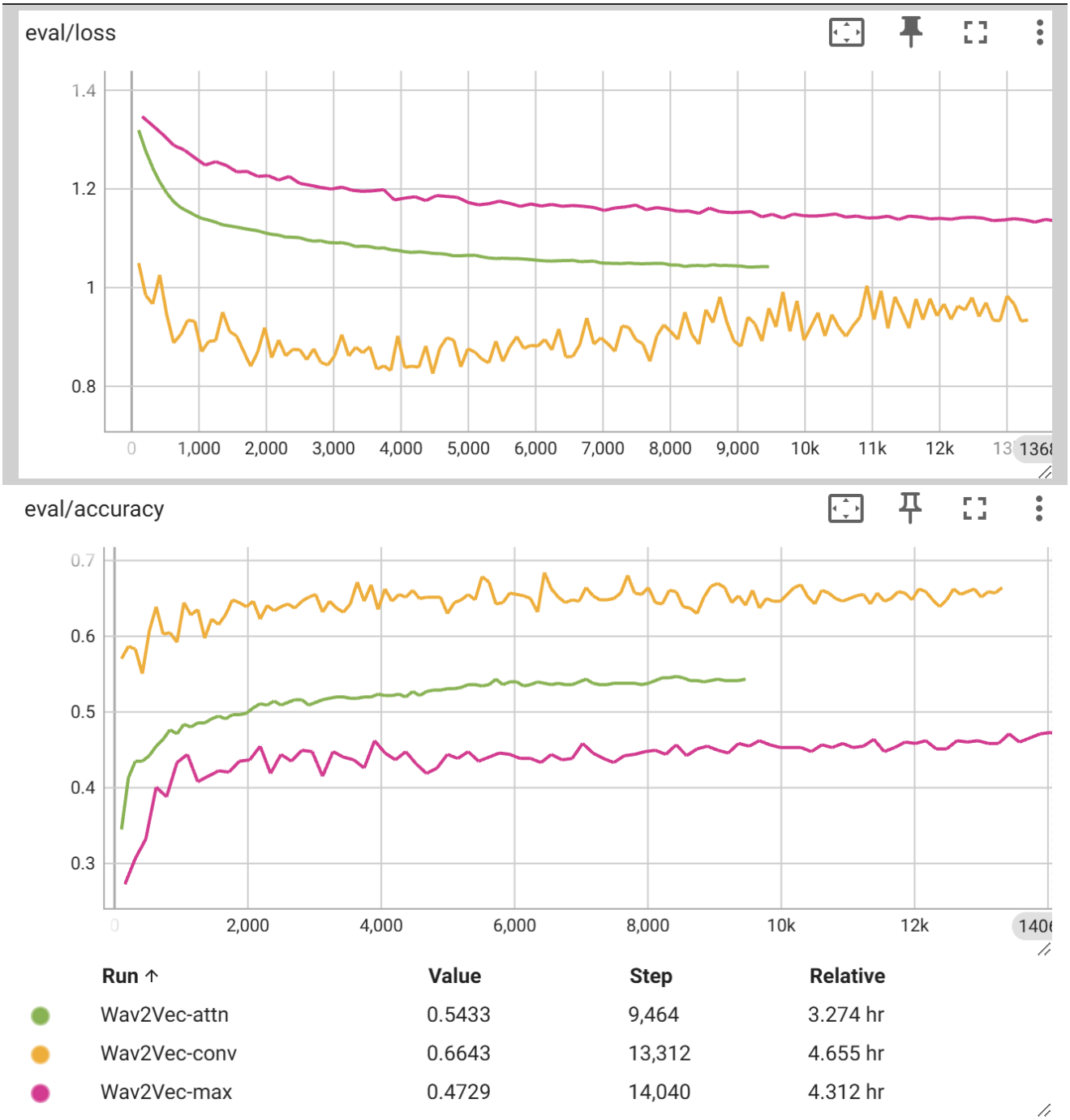


图3 Wav2Vec后接不同池化的损失与准确率曲线

实验结果表明，从全局最大池化、注意力池化到卷积-全局最大池化，拟合能力越来越强，训练损失和验证损失都越来越低，验证准确率越来越高。其中卷积-全局最大池化在5k步后训练损失持续降低的同时验证损失开始升高，出现了过拟合现象。这说明Wav2Vec产生的特征序列直接池化或混合，复杂度太低，并不能很好地预测情感；但经过一层卷积处理后，已经可以很好地对情感分类问题进行拟合。这可能是因为Wav2Vec不仅仅提取了情感相关的特征，还包含了语义、声纹等信息，通过简单的池化或混合不能得到很好的句子级情感特征，Wav2Vec提取的特征序列相当于词向量，还需要经过学习才能提取出其中和情感有关的部分。表现最好的卷积-全局最大池化最终可以达到4分类68%的准确率。

系统展示

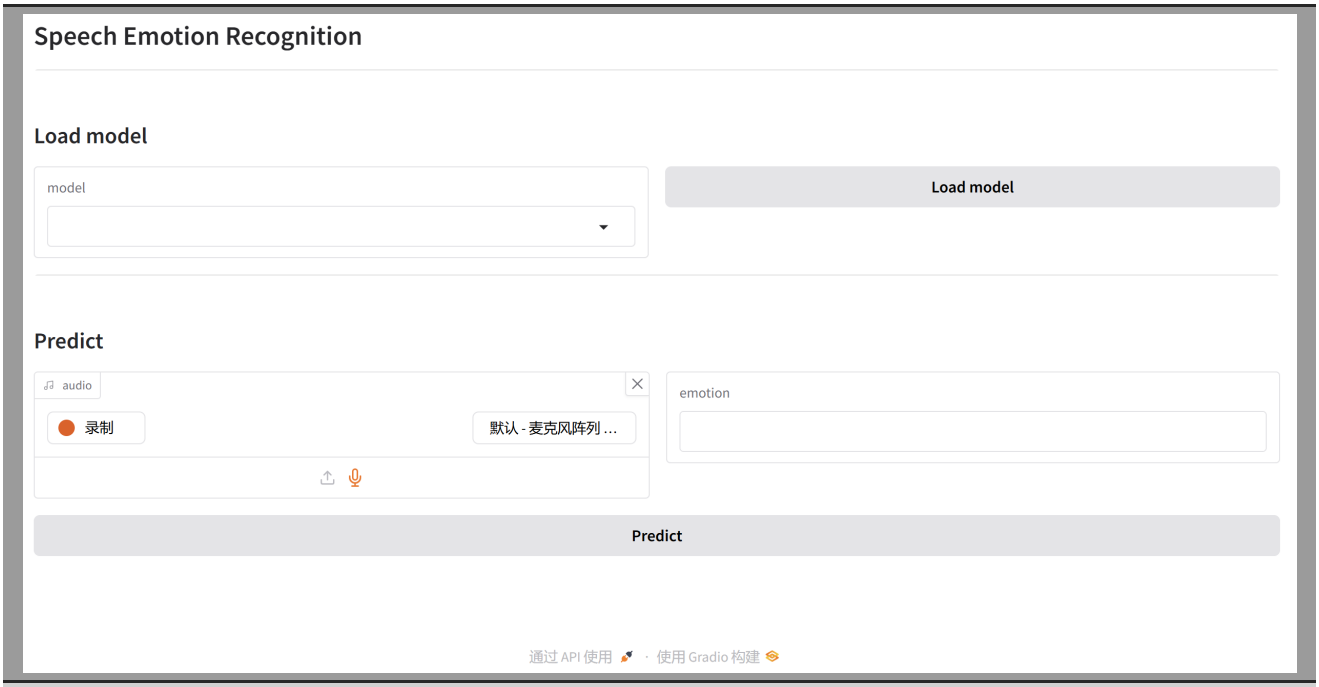


图4 SER系统的图形用户界面

首先可以在左上方选择模型，然后点击按钮加载模型，之后可以现场录制音频或者上传音频文件，点击按钮即可识别语音情感。

改进方向

虽然在验证集上表现较好，但经过实测，模型对现场录音的泛化能力不强，这可能是由于训练集只包含了五个会话的英文语料，导致模型在说话人和语言上泛化能力不强，最终的实际使用效果不佳。可以尝试使用更丰富的数据集，如使用多个语言的数据集拼接，分别处理成统一的格式形成新数据集进行训练。此外，可以使用专门在情感丰富的语料上自监督预训练的模型，如Emotion2Vec、HiCMAE等，这些模型提取的特征与情感更相关，且泛化性更强，效果应该会更好。

实验总结

本实验成功搭建了一个基于深度学习的带有GUI的纯语音情感识别系统，结合CNN与Wav2Vec模型，直接从音频波形中提取特征进行分类。实验结果显示，Wav2Vec+卷积-全局最大池化方法表现最佳，验证集准确率达到68%。然而，模型在现场录音数据上的泛化能力较弱，主要由于训练数据的单一性。未来可通过引入多语言、多说话人数据集及使用情感自监督预训练模型，进一步提升系统的鲁棒性与实际应用效果。