文章编号：1003-0077（2023）00-0000-00

基于投票方式的多模态情感识别方法

李启飞 王聪 任一鸣 王栋 高迎明 李雅*

（北京邮电大学 人工智能学院，北京 100876）

摘要：该文将呈现针对中国情感计算大会多模态对话中的情感识别挑战赛的解决方法。该文设计了三个模型方案，最终通过三个模型投票的方法得出最终的预测结果。首先，微调基于预训练模型 Hubert 的语音情感识别模型，得到模型一。其次微调基于预训练模型 Macbert 的文本情感识别模型得到模型二。再次，通过抽取 Hubert、Macbert、ResNet-FER2013 三个预训练模型的隐藏层作为三个模态句子级别的表征，再使用注意力机制融合得到最后的情感识别模型三。最后通过投票的方式，融合三个模型的决策得到最终的预测情感类别。本文的方法在情感识别挑战赛数据集（M^3ED）的权重 F1 值达到 0.5272，测试集结果评估中获得了第 2 名的成绩。

关键词：情感识别；预训练模型；多模态

中图分类号：TP391 **文献标识码：**A

Multimodal emotion recognition method based on voting method

Qifei Li, Cong Wang, Yiming Ren, Dong Wang, Yingming Gao, Ya Li

(School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, 100876, China)

Abstract: In this paper, we present a solution for the Emotion Recognition Challenge in Multimodal Dialogue at the China Emotion Computing Conference. Three model solutions are designed in this paper, and the final prediction results are finally obtained by the method of voting among the three models. Firstly, we fine-tune the speech emotion recognition model based on the pre-trained model Hubert to obtain model one. Secondly, the text emotion recognition model based on the pre-training model Macbert was fine-tuned to get model two. Again, the third emotion recognition model is obtained by extracting the hidden layers of the three pre-trained models Hubert, Macbert, and ResNet-FER2013 as the tri-modal sentence-level representations, and then fusing them using the attention mechanism. Finally, the final predicted emotion category is obtained by fusing the decisions of the three models by voting. Our method in this paper has a weight F1 value of 0.5272 in the emotion recognition challenge data set (M^3ED), and won the second place in the test set result evaluation.

Key words: emotion recognition; pretrain-model; multimodal

0 引言

情感计算旨在通过算法来识别、解释、处理和模拟人类的情感，以便更好地实现人类与机器进行人机交互[1]。情感计算在各个领域中都发挥着重要作用，如在医疗领域，情感计算可以帮助改善心理健康治疗和辅助机器人护理；在教育领域，情感计算可以用于个性化教学和学习支持系统；在市场营销领域，情感计算可以用于情感分析和消费者行为预测。这些应用的发展将进一步提升人们的生活质量和社会效益。而情感识别侧重于通过自动化方法来识别和分类人类的情感状

收稿收稿日期：2023-06-14；**定稿日期**：2023-06-14

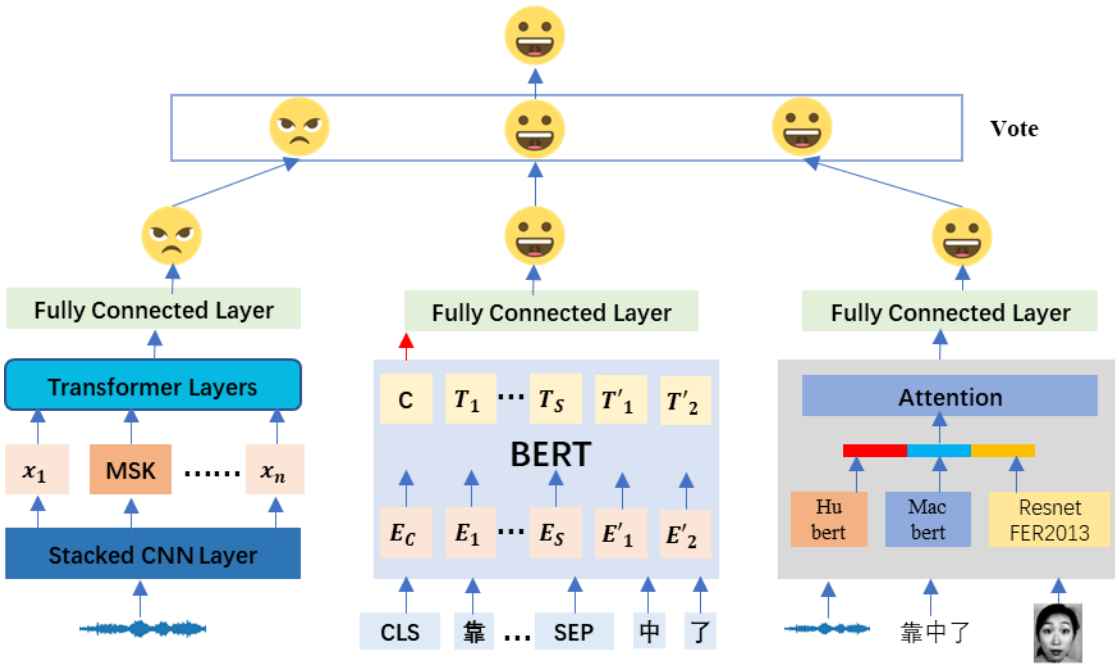

图 1 模型结构图

态，是情感计算的重要组成部分。情感识别关注于从人类产生的文本、语音、图像等数据中，自动地检测和分类情感的能力。

深度学习的发展使得情感识别技术蓬勃发展。相较于早先的基于特征集方法需要手动设计和提取特征，深度学习模型能够通过多层神经网络学习到更高级、更抽象的特征，从而更好地捕获情感信息。此外深度学习模型通常具有较强的鲁棒性和泛化能力，能够从大规模数据集中学习到丰富的模式，从而更好地适应不同的数据分布和情感识别任务，能够处理不同领域和任务的情感识别需求。研究者们采取深度学习方法在情感识别任务中取得了显著的成果，使用卷积神经网络、循环神经网络和注意力机制等深度学习模型，有效捕获了文本、语音和视频中的情感特征，提高了情感识别的准确性和鲁棒性。然而随着研究的不断深入，数据资源和计算资源稀缺的问题也随着显现出来，为了克服这些技术瓶颈，预训练模型策略被引入到情感计算领域中。预训练模型通过在大规模未标注的数据上进行预训练，可以学习到更丰富的语音表示和模式，从而在有限标注数据的情况下，利用先前学习到的知识和表示进行更好的泛化和表达。在情感识别任务中，采取预训练模型进行特征表示学习，能够学习到数据中丰富的语义表示，其中包含了大量的情感信息相关特征，提供了丰富准确的情感信息，从而在相关任务中取得了显著的效果和性能提升。

针对本次多模态情感识别挑战赛，本文使用了一种投票融合多个模型决策的方式。实验模型由三个子模型组成，分别基于 Hubert（Hidden-Unit Bert）微调的语音模态情感识别模型[2]、基于 Macbert（MLM as correction BERT）[3]微调的文本模态情感识别模型和基于注意力机制融合的语音、文本和视频多模态融合的情感识别模型。考虑到不同的模型对样本情感的理解不一致，本文采取了决策投票的融合的方式，获得多个模型中最贴近真实标签的预测结果，实现更好的情感识别性能。本文模型结构图如图 1 所示。

1 模型方法

近年来，语音和文本自监督预训练大模型的语音和文本的各种下游任务大放异彩，取得良好的性能，例如语音识别[4]、说话人识别[5]、语音情感识别[6]、文本情感识别[7]、文本主题分类[8]等。鉴于自监督预训练模型的性能优势，本文使用微调技术在自监督预训练大模型进行下游任务，即情感识别中进行验证。本小节介绍模型的实验细节。

针对子模型 1，本文是在腾讯天籁实验室开源的中文语音自监督预训练大模型 HuBert 进行微

调。本项目的实现方法只是在 Hubert 预训练模型后加一层全连接层，并冻结所有卷积神经网络层，然后直接在竞赛数据集中进行了微调情感识别任务，所用的特征维度为 1024 维。此方法是一个完全端到端的方法，所以不需要对音频提取特征和变换。

针对子模型 2，子模型 2 和子模型 1 的实现方法基本一致。子模型 2 是在 Macbert 预训练模型的基础下进行微调。MacBert 的结构和 Bert 的结构完全一致。它的创新点主要是引入了纠错型掩码语言模型（MLM as correction，Mac）预训练任务，缓解了“预训练-下游任务”不一致的问题。在此模型实现的过程中只需要在 Bert 模型的 pooling 层后添加一层全连接层，然后使用文本情感数据直接进行情感任务的微调即可，特征维度同样选定为 1024。

针对子模型 3，考虑到微调视频模态需要的时间过长和计算资源所需甚大，为了在有限资源的情况下充分利用视频模型的信息，本文实现了抽取视频预训练模型最后的隐藏层信息作为视频特征编码。模型 3 中的 Hubert、Macbert 和 Resnet-FER2013[9]模型均未进行下游任务微调，Resnet-FER2013 维度为 512。在本模型中抽取三个预训练模型的隐藏层特征（句子级别）进行拼接，然后再借助注意力机制融合三个模型实现多模态情感识别方法。具体计算方法如公式（1）和（2）所示：

$$h_i = Concat\left(h_i^a, h_i^l, h_i^v\right) \quad (1)$$

$$c_i = softmax(h_i^T W_a + b_a) \quad (2)$$

其中 h_i^a, h_i^l, h_i^v 分别表示语音、文本和视频模态第 i 维的特征。W_a表示可学习的参数。c_i表示注意力机制的输出。

2 数据处理

本文使用数据为中国人民大学提供的 M³ED 多模态数据集[10]，数据描述如表 1 所示。

表 1 数据分布

	训练集	验证集	测试集
对话数量	685	126	-
对话轮数	6505	1016	-
语句数量	17427	2821	1191
说话人数量	421	87	-

数据集里的样本均是视频片段，所以首先借助 FFmpeg 工具将视频按照样本标注的时间戳切分成视频段，然后再采用 FFmpeg 工具将视频段中的语音转换出语音模态样本，采样率为 16KHz。切分出来的很多音频样本时间极短，无法进行任务微调。文本模态有很多样本极短，所包含的含义无法与标签对应，所以在本项目中的子模型 1 和子模型 2 需要做特殊的样本处理。样本会按照主题-说话人-标签进行拼接，数据预处理的方法如下：

算法 1：数据预处理方法

输入：标注 Json 文件
输出：拼接完成的语音样本和文本
1: If 当前样本的说话人、话题编号和标签与文件中下一条的样本一致 满足 do
2: 将音频名、文本和标签按顺序保存在序列中
3: Else
4: 加载现存的序列，读取音频和文件进行拼接，将多个样本拼接为一个样本。保存格式为:说话人_剧名_话题 ID_1_2_3.wav(表示 3 个样本拼接在一起)
5: End

针对子模型 3 无需做特征的输入处理，在提取 Hubert 隐藏层表示时只取最后一层，提取 MacBert 隐藏层表示时需要取最后四层最均值，在取 Resnet-FER2013 隐藏层表示时只取最后一层。

3 实验

3.1 实验设置

子模型 1 和 2 的训练设置一样，训练 100 个 Epoch，保存验证集权重 F1 值最高的 epoch 的模型。学习率为固定值 1e-5,优化器为 AdamW，损失函数为交叉熵损失函数。为了解决情感类别样本不平衡问题，本文采取采样器对数据进行加权随机采样，提高少类样本出现的频率。本文实验中采用四个 RTX 3090 GPU 进行模型训练，每个 GPU 的 batch size 为 1，梯度累计步数为 2。

子模型 3 的网络结构为每种模态先经过两层全连接层，神经元个数分别为 256 和 128，在将三个模态的全连接层输出频接做注意力机制融合，得到最后的情感识别结果。Batch size 为 64，单卡训练，学习率为 1e-4，优化器为 Adam。。保存验证集性能最好的 Epoch 模型。

3.2 实验结果

大赛规定每个参数队伍可以提交三次实验结果，本文分别提交了子模型 1 的结果、子模型 3 的结果和最终三个子模型投票的结果，在验证集和测试集的结果如表 1 所示：

表 1 不同模型的权重 F1 值

	验证集	测试集
子模型 1	0.6068	0.5127
子模型 2	0.5192	-
子模型 3	0.6013	0.4937
投票集成	-	0.5272

投票的方式为，三个子模型的预测结果取众数，如果三个模型预测的结果均不一致，则此样本的结果取子模型 1 的结果。

2.2 实验分析

通过对比子模型 1 和 2 的在验证集发现，语音模态相较于文本模型更重要。多模态的结果在验证集上的性能和子模型 1 的性能相近，但是在测试集上表现较差。最好的性能是，投票后的模型性能，权重 F1 达到 0.5272，表明有些样本预测结果需要依赖文本模态纠正。换句话说，语音模态是此次比赛的主要性能贡献模态，而文本模态和视频模态能纠正音频模态的误判，从而提升最终的系统性能。

4 总结

本文介提出了三种模型解决方案，包括基于预训练模型 Hubert 的语音情感识别模型，基于预训练模型 Macbert 的文本情感识别模型，以及一种提取了三个预训练模型 Hubert、Macbert 和 ResNet-FER2013 隐藏层的情感识别模型，通过投票的方法对三个模型的输出进行整合，以实现更好的性能表现。本文提出方法在多模态对话中的情感识别中取得第二名，证实了本文方法的有效性。

参考文献

[1] Tao J, Tan T. Affective computing: A review [C]. proceedings of the Affective Computing and Intelligent Interaction: First International Conference, ACII 2005, Beijing, China, October 22-24, 2005. Proceedings 1. Springer Berlin Heidelberg, 2005: 981-995.

[2] Hsu W-N, Bolte B, Tsai Y-H H, et al. Hubert: Self-supervised Speech Representation Learning by Masked Prediction of Hidden Units [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451-60.

[3] Cui Y, Che W, Liu T, et al. Revisiting Pre-trained Models for Chinese Natural Language Processing. arXiv preprint arXiv:200413922, 2020.

[4] Yi C, Wang J, Cheng N, et al. Applying Wav2vec2.0 to Speech Recognition in Various Low-resource Languages [J]. arXiv preprint arXiv:201212121, 2020.

[5] Vaessen N, Van Leeuwen D A. Fine-tuning Wav2vec2 for Speaker Recognition [C]. proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.

[6] Pepino L, Riera P, Ferrer L. Emotion Recognition from Speech using Wav2vec 2.0 Embeddings. arXiv preprint arXiv:210403502, 2021.

[7] Adoma A F, Henry N-M, Chen W. Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition [C]. proceedings of the 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP).

[8] Ma T, Pan Q, Rong H, et al. T-bertsum: Topic-aware Text Summarization Based on Bert [J]. IEEE Transactions on Computational Social Systems, 2021, 9(3): 879-90.

[9] 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, 2016.

[10] Zhao J, Zhang T, Hu J, et al. M3ED: Multi-modal Multi-scene Multi-label Emotional Dialogue Database. arXiv preprint arXiv:220510237, 2022

作者介绍

李启飞（1994-），博士在读，主要研究领域为情感识别、抑郁症检测。
E-mail：liqifei@bupt.edu.cn

王聪（2001-），学士在读，主要研究情感识别。
E-mail：congwang@bupt.edu.cn

任一鸣（2001-）学士在读，主要研究领域为情感识别、抑郁症检测。
E-mail：rym@bupt.edu.cn

王栋（1998-），硕士在读，主要研究领域为情感识别、抑郁症检测。
Email：dong1024mail@163.com

高迎明（1989-），博士，讲师，主要研究领域为语音信息处理、深度学习、计算机辅助语言学习、声学语音学等。
E-mail：yingming.gao@bupt.edu.cn

李雅（1984-），博士，副教授，主要研究领域为语音交互、多模态情感计算。
E-mail：yli01@bupt.edu.cn