基于 DialogueRNN 多折集成模型的 CCAC2023-MERC 分类方法

赵志龙[1] 胥卜凡[1] 徐冰[1*]

（1.哈尔滨工业大学 计算学部 语言技术研究中心，黑龙江省 哈尔滨市）

摘要：该文描述了我们在第三届中国情感计算大会(CCAC2023)多模态对话情感识别任务评测中构建的模型系统。多模态对话中的情感识别任务是一个具有挑战性的研究问题，对话情感受多模态中上下文、对话者刺激、自身情绪惰性等因素影响，因此模型构建较为困难。该文的模型框架基于 DialogueRNN 模型和百度 ERNIE3.0-xbase 预训练模型，在数据方面，为不同类别添加损失权重解决数据不均衡问题，性能提升技术上采用对抗训练、k-fold、等权投票等技术，在测试集上宏平均 F1 达到 50.38，在最终提交的 16 个模型中排名第 3。

关键词：多模态对话；情感识别；集成学习

A CCAC2023-MERC Classification Approach based on DialogueRNN Multi-fold Integration Model

Zhao Zhilong[1], Xu Bufan[1], and Xu Bing[1,*]

(1. Language Technology Research Center of Computing Department of Harbin Institute of Technology ,Harbin city, Heilongjiang province)

Abstract : The presented paper outlines a model system developed for the task of multimodal dialogue emotion recognition at the Third Chinese Conference on Affective Computing (CCAC2023). This task poses significant challenges due to the influence of various factors, such as contextual cues, interlocutor stimuli, and self-emotion inertia. Constructing a model that effectively addresses these complexities is a difficult undertaking. To tackle this problem, the authors propose a model framework based on the DialogueRNN model and the Baidu ERNIE-xbase pretrained model. Additionally, to address the issue of data imbalance, training weights are incorporated into the data. Several performance improvement techniques, including adversarial training, k-fold cross-validation, and equal-weight voting, are employed. The model's performance is evaluated on a test set, and the results show a macro-averaged F1 score of 50.38. Among the 16 models submitted for final evaluation, the proposed model ranks 3rd. These findings suggest that the developed model system demonstrates promising capabilities for multimodal dialogue emotion recognition.

Key words: multi-modal dialogue; emotion recognition; ensemble learning

0 引言

CCAC2023-MERC[1] 是第三届中国情感计算大会多模态对话中的情感识别任务评测。多模态对话中的情感识别旨在自动识别和跟踪对话中说话者的情绪状态，在自然人机交互、教育、医疗等场景具有广泛的应用。与传统的单句多模态情感识别不同，多模态对话场景下的多模态情感识别是一个更具挑战性的问题，因为对话中影响说话者情绪状态的影响因素很多，包括多模态的上下文、对话者刺激、自身情绪惰性、对话场景、人格特征等。对话情感识别任务已经被广泛的研究，但多模态对话情感识别任务研究还在起步阶段，最主要原因在于多模态数据集的不足，特别是中文语料的缺乏。本次评测采用中国人民大学 AI·M³ 多

[1] http://www.ccac2023.cn/

媒体计算实验室提出的 M3ED 中文多模态对话情绪数据集。[1]

本次评测任务旨在识别 M3ED 数据集中不同语句的情感。输入是含有文本、语音以及图像信息的对话，输出是对话中每条语句对应的情感。在本次评测中，将对话中蕴含的情感分为以下七个类别之一：平静（Neutral）、开心（Happy）、惊讶（Surprise）、难过（Sad）、厌恶（Disgust）、生气（Anger）和害怕（Fear）。值得注意的是，数据集原始标注是多标签的，为了简化任务难度，采用主情感标签进行情感识别。

图 1 展示了多模态数据的数据样例，表 1 展示了数据集的标注信息。

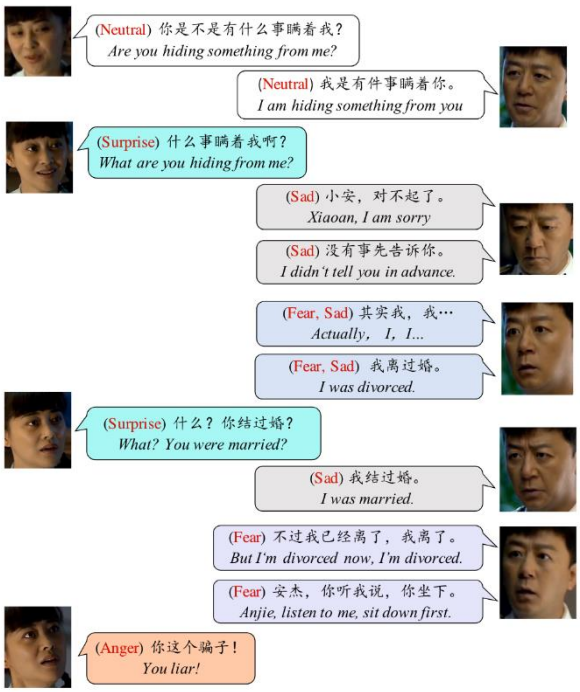

图 1 对话多模态数据示例

表 1 数据标注样例

话语编号	文本	说话者	情感
fumuaiqing_1_1	你是不是有什么事瞒着我?	A	平静
fumuaiqing_1_2	我是有件事瞒着你	B	平静
fumuaiqing_1_3	什么事瞒着我啊？	A	惊讶
fumuaiqing_1_4	小安	B	难过

本文接下来将从数据集的分析、模型的整体结构、实验结果与分析以及总结完整地介绍我们使用的方法。

1 数据分析

表 2 数据集分布

Statistics	Train	Val	Total
TV Series	38	7	45
Dialogs	685	126	811
Turns	6505	1016	7521
Utts	17427	2821	20248
speakers	421	87	508

表 2 列出了本次评测提供的训练集和验证集数据分布，训练集和验证集分别来自 38 部和 7 部中文影视剧，每段对话都有 A，B 两个说话者。

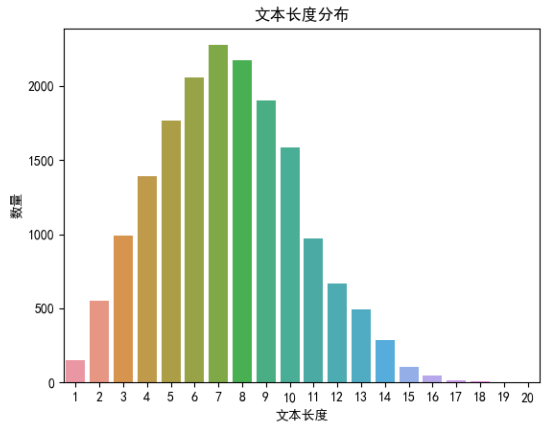

图 2 训练集文本长度分布

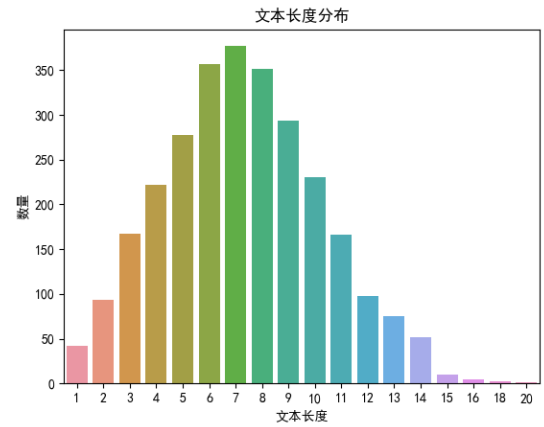

图 3 验证集文本长度分布

图 2 和图 3 展示了数据集中对话文本长度分布，可以看出，对话文本长度主要集中在 7 个字左右，比较简短。所以模型在处理数据时，可以将最大文本长度适度调小。

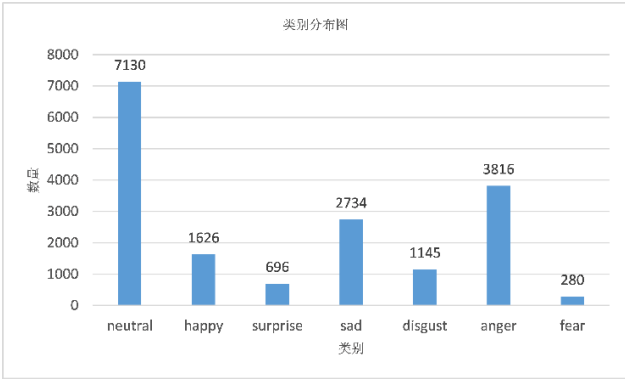

图 4 训练集类别分布图

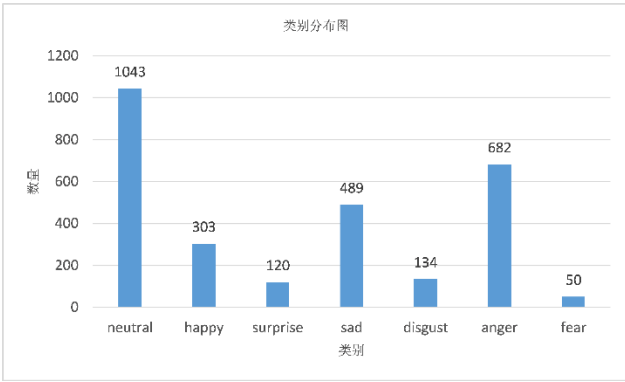

图 5 测试集类别分布图

图 4 和图 5 展示了数据集情感类别分布，可以看出，数量最多的情感类别(neutral)相较于数量最少的情感类别(fear)相差一个数量级，即训练集和验证集存在严重的数据不平衡问题，在构建模型过程中通过对不同类别的 loss 加权解决了该问题，每个类别权重默认设置为类别数所占比例的倒数。

2 多模态特征提取

数据集包含文本、音频和视频三种模态信息，按照 AI·M³ 多媒体计算实验室提供的开源特征文件。文本模态主要采用 RoBERTa-base[2]预训练模型编码，在 M3ED 数据集上进行 finetune。音频模态主要采用 Wav2vec[3]和 OpenSmile 提取特征。视频模态采用 DenseNet[4]模型，首先检测说话者面部信息，基于检测到的说话人的面部，为每个话语提取面部级别的特征。

在本次评测中，我们采用相同的方法，但只实现了文本和音频模态的特征抽取。具体来说，文本模态采用百度的 Ernie3.0-xbase[5]预训练模型，基于此模型在 M3ED 数据集上 finetune，取编码后[CLS]代表的隐藏层向量作为对话句的文本特征向量。音频模态采用 OpenSmile 工具提取特征，在此之前，需要先采用 FFempg 工具将视频数据转为音频数据，进一步按照句子对应的音频时值进行切割。抽取出的音频特征需要进一步进行 z-norm。

3 模型方法

3.1 模型架构

我们采用 DialogueRNN 模型[6]，模型整体架构如下图。该模型结构很好地捕获了对话中的说话者信息和说话主题信息，因此能取得较好的效果。

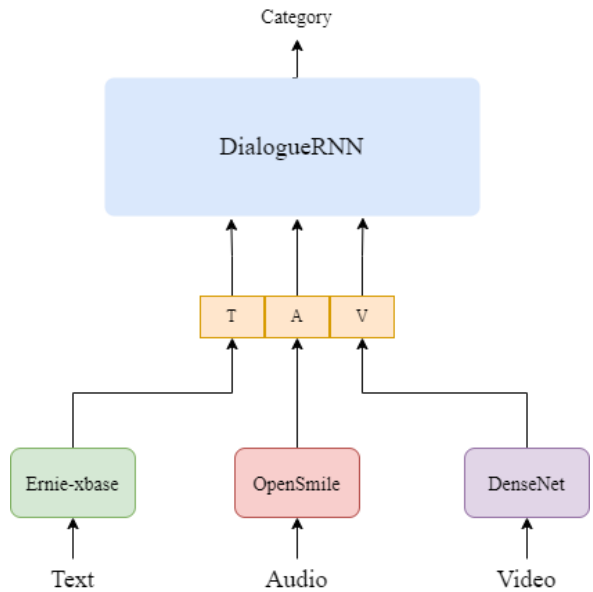

图 6 模型架构图

3.2 文本特征编码优化

一般而言，采用效果更好的预训练模型能取得明显效果的提升。百度提出字词混合的自监督对比学习预训练技术和字词混合数据增强自对抗微调技术，使得 Ernie3.0 模型在 8 个中文主流下游任务上获得了明显的效果提升。因此，我们选择采用的 Ernie3.0-xbase 模型，该模型隐藏层维度为 1024 维，后续实验也证明了采用该模型的有效性。

3.3 对抗学习

FGM 对抗学习[7]，通过对输入文本的

embedding vector 添加扰动，提高模型的泛化能力，如下面公式所示，ϵ为 1，g是 embedding vector 的梯度。

$$r_{adv} = \epsilon \cdot \frac{g}{\|g\|_2} \tag{1}$$

$$g = \nabla_x L(\theta, x, y) \tag{2}$$

PGD 对抗学习[8]是对 FGM 对抗学习的优化，FGM 直接通过ϵ参数一下子算出了对抗扰动，这样得到的可能不是最优的。因此 PGD 进行了改进，多迭代几次，慢慢找到最优的扰动。如下面公式所示，g_t是第 t 步 embedding vector 的梯度。

$$r_{adv|t+1} = \alpha \cdot \frac{g_t}{\|g_t\|_2} \tag{3}$$

$$g = \nabla_x L(\theta, x, y) \tag{4}$$

$$\|r\|_2 \leq \epsilon \tag{5}$$

3.4 集成学习

集成学习方法，我们采用单模多折等权投票集成，简单来说，对于 K 折交叉验证表现最好的 k 个模型，分别对测试集进行预测，预测结果取均值，得到最终的情绪类别。

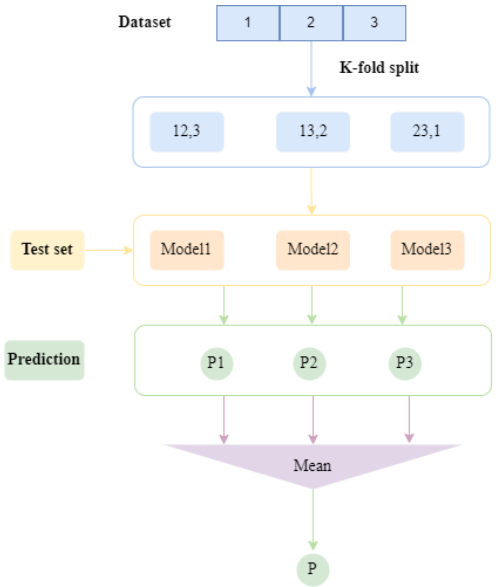

图 7 集成学习框架

4 实验结果及分析

在这一部分，主要介绍三个实验，分别是特征文件选择实验，文本特征优化实验和对抗学习方法实验。最后一张表展示了采用 3 折交叉验证的模型最终结果，由于时间关系，并没有做 k 折交叉验证调参实验。

4.1 特征文件选择

因为 M3ED 数据集官方 github 库提供了 4 个特征文件，分别表示采用不同策略抽取的文本、音频和视频三种模态特征。表现最好的特征文件音频模态采用了 OpenSmile 提取 IS10 音频特征；文本模态采用 RoBERTa-base 在 M3ED 数据集上 finetune；视频模态采用 DenseNet 识别面部表情特征。该特征文件在原始测试集的 Macro-F1 值达到 41.08。

4.2 文本特征优化

文本特征抽取我们抛弃了原始特征文件采用的 RoBERTa-base 预训练模型，改用百度提出的 Ernie3.0-xbase 预训练模型，在 M3ED 数据集上进行 finetune 后对对话文本进行编码，下表展示了融合三种模态特征后在原始测试集中的表现。

表 3 文本特征优化实验

Method	P(%)	R(%)	F1(%)
Roberta-base	47.46	39.43	41.08
Ernie3.0-xbase	45.69	41.42	42.88

可以看出，在采用文本编码能力更强的模型后，整体效果提升明显。

4.3 对抗学习方法

本次评测中，采用了 PGD 和 FGM 两种对抗学习的方法。下表展示了在文本和音频两个模态融合下，最终对抗学习方法的表现。

表 4 对抗学习方法实验

Method	P(%)	R(%)	F1(%)
baseline	44.24	42.77	42.69
PGD	43.74	44.11	43.27
FGM	46.37	43.04	43.98

在本次评测数据集中，FGM 对抗学习方法要略优于 PGD 方法，因此，最后本队采用 FGM 对抗

学习方法。

4.4 实验参数设置

本小组提交的最终结果基于以下实验参数设置，参数 K 代表采用 K 折交叉验证折数。

表 5 实验参数设置

超参数	值
K	3
Batch size	30
Learning rate	5e-4
Epoches	60
Drop_out	0.1

4.5 集成学习方法

我们采用多折交叉验证的模型进行等权投票，来提升整体的模型性能。出于训练时间的考虑，采用三折交叉验证的方式，下表展示了在文本和音频两个模态融合下的三个模型表现。

表 6 交叉验证模型表现

Method	P(%)	R(%)	F1(%)
Model1	57.22	53.94	54.91
Model2	61.87	56.40	56.49
Model3	60.59	56.31	57.42

5 总结

此次评测，我们借鉴经典的对话情绪分析模型 DialogueRNN，通过文本特征编码优化，对抗训练，集成学习等优化策略，取得了不错的成绩。但遗憾的是，本次评测最终提交结果只采用了文本和音频两种模态信息，没有实现 DenseNet 提取视频模态特征信息。除此之外，k 折交叉验证并未做进一步的实验，并且调参工作也做得比较粗糙，这些部分都有待后续优化。

参考文献

[1] ZHAO J, ZHANG T, HU J, et al. M3ED: Multi-modal Multi-scene Multi-label Emotional Dialogue Database[J]. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2022, 1: 5699–5710. DOI:10.18653/v1/2022.acl-long.391.

[2] LIU Y, OTT M, GOYAL N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[J/OL]. 2019(1). . http://arxiv.org/abs/1907.11692.

[3] BAEVSKI A, ZHOU H, MOHAMED A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[J]. Advances in Neural Information Processing Systems, 2020, 2020-December(Figure 1): 1–19.

[4] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[J]. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, 2017-January: 2261–2269. DOI:10.1109/CVPR.2017.243.

[5] WANG S, SUN Y, XIANG Y, et al. ERNIE 3.0: LARGE-SCALE KNOWLEDGE ENHANCED PRE-TRAINING FOR LANGUAGE UNDERSTANDING AND GENERATION[J/OL]. 2021. . http://arxiv.org/abs/2112.12731.

[6] MAJUMDER N, PORIA S, HAZARIKA D, et al. DialogueRNN: An attentive RNN for emotion detection in conversations[J]. 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, 2019: 6818–6825. DOI:10.1609/aaai.v33i01.33016818.

[7] MIYATO T, DAI A M, GOODFELLOW I. Adversarial training methods for semi-supervised text classification[J]. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings, 2017: 1–11.

[8] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, 2018: 1–28.