计算机应用研究
Application Research of Computers
ISSN 1001-3695,CN 51-1196/TP

《计算机应用研究》网络首发论文

题目：基于交互注意力机制的多模态情感识别算法
作者：姚懿秦，郭薇
DOI：10.19734/j.issn.1001-3695.2020.09.0230
收稿日期：2020-09-04
网络首发日期：2021-03-05
引用格式：姚懿秦，郭薇．基于交互注意力机制的多模态情感识别算法[J/OL]．计算机应用研究. https://doi.org/10.19734/j.issn.1001-3695.2020.09.0230

基于交互注意力机制的多模态情感识别算法

姚懿秦，郭　薇†

(上海交通大学 电子信息与电气工程学院，上海 200240)

摘　要：在多模态语音情感识别中，现有的研究通过提取大量特征来识别情感，但过多的特征会导致关键特征被“淹没”在相对不重要特征里，造成关键信息遗漏。提出了一种模型融合方法，通过两种注意力机制来寻找可能被遗漏的关键特征。本方法在 IEMOCAP 数据集上的四类情感识别准确率相比现有文献有明显提升；在注意力机制可视化下，两种注意力机制分别找到了互补且对人类情感识别重要的关键信息，从而证明了所提出方法相比传统方法的优越性。

关键词：多模态情感识别；注意力机制；信息交互

中图分类号：TP183　　doi: 10.19734/j.issn.1001-3695.2020.09.0230

Multimodal emotion recognition algorithm based on interactive attention mechanism

Yao Yiqin, Guo Wei†

(*School of Electronic Information & Electrical Engineering*, *Shanghai Jiao Tong University*, *Shanghai* 200240, *China*)

Abstract: In multimodal emotion recognition, the proposed researches focus on extracting more and more features to classify emotion. But too many features which is relatively not important may “flood” the important feature and due to key information lost. This paper proposed a model fusion algorithm (GATASA) which based on multi-attention mechanism. 2 attention mechanisms are used to find out important feature which may be neglect. The proposed interactive attention algorithm effectively can improve recognition accuracy in the 4 emotions classification on IEMOCAP dataset. By using attention visualization those 2 attention mechanisms find complementary and important information from human perspective which proves the proposed GATASA is advanced than traditional algorithm.

Key words: multimodal emotion recognition; attention mechanism; information interaction

0　引言

在许多监督学习的场景或人机交互软件中，人类的情感总是起到至关重要的作用，但是又很难识别。比如在驾驶员安全驾驶监控中，对驾驶员情绪的判断则显得至关重要：如果他十分愤怒，那大概不会安全驾驶。通常可通过很多人类的特征来识别情感，比如表情，声音，说话句子的内容等。结合更多数据源(即多模态情感识别)，往往会对情感的识别带来更加准确的结果。

多模态情感识别的基础是的每一类型特征的单独情感识别，常见的有语音、文本、视频情感识别，本文是基于语音和文本的多模态情感识别。任何一种情感识别都包括三个模块：数据预处理、特征提取和分类器。语音情感识别中，基于深度学习和传统方法的系统区别较大：传统的数据预处理基于 MFCC(Mel-Frequency cepstral coefficients) [1] 或者 LPCC(linear predictive coding coefficients) [2]，基于深度学习的数据预处理则常用 MFCC 或者声谱图(spectrogram)。在传统方法中数据预处理和特征提取可能会成为同一个模块，而在深度学习中由于卷积神经网络(convolutional neural network) [3] 和循环神经网络(recurrent neural network) [4] 的强大性能，可以直接作为特征提取器，以获得更高层更抽象的高级特征。在传统方法中语音情感识别的分类器可以是文献[5]中的高斯混合模型或者文献[6]中的支持向量机，而深度学习的分类器往往是全连接层(fully-connected layer)。

对于文本情感分析(sentiment analysis)，其分类器和其他的情感识别的分类器较为接近，但数据预处理和特征提取区别较大：深度学习中文本常见的数据预处理为去掉不需要的停止词(stop word)，然后对单词做词嵌入(embedding)。词嵌入通常基于现有的词向量(word vector)：基于预训练的 Glove [7] 或者 BERT [8]。对于多个数据源的特征，文献[9]中加入了注意力机制，获得了进一步的性能提升。

由于人类说话时候的随意性和语音文本内容的复杂性，提取出来的特征重要程度并不一致(比如人类判断“I am angry”时，重点是“angry”,而“I am”对判断句子情感帮助较少)，不同文本内容对应的文本特征和不同语音片段的音频特征重要程度需要被强调，强调对判断句子关键的特征，弱化不重要的特征。现有的研究缺乏对人类直观容易理解的关键特征的强调。大多数研究只是将所有特征直接送入神经网络，期待神经网络能够自动完成剩下的所有工作，文献[10]利用全连接层，文献[11]利用循环神经网络。因为没有强调重要特征，准确率不高且可解释性较差(不强调重要特征的缺点)。

针对前文提出的对重要特征的强调程度不够的问题，本文提出了利用多种交互的注意力机制设计的关键特征选择网络 GATASA(Global Acoustic-to-text and Acoustic-to-Self-Acoustic to Text)。GATASA 由两部分组成：GATA 和 ASATA，这两部分由两种不同的注意力机制在文本和音频特征之间交互计算注意力分数。GATA 关注尽可能多的强调所有的文本和音频信息，即从全局角度强调特征，而 ASATA 注重强调音频和文本特征里重要的局部特征，即从局部角度强调特征。GATA 利用经过多次经过激活函数的音频特征在文本特征里寻找对于音频来说重要的特征，再利用找到的文本特征回到音频里寻找重要的音频特征。ASATA 则先是在音频特征内部

收稿日期：2020-09-04；**修回日期：**2020-10-19

作者简介：姚懿秦(1996-)，男，云南红河人，硕士，主要研究方向为机器学习，情感识别，多模态机器学习；郭薇(1963-)，女(通信作者)，湖北武汉人，教授，博士，主要研究方向为光通信网络，卫星通信网络，机器学习(wguo@sjtu.edu.cn).

寻找重要的特征，再利用寻找到的音频特征去寻找重要的文本特征。寻找出来的 4 组重要特征作为文本和语音的最终表示被用于分类。通过实验证明，分类结果达到了 77.0%的加权准确率(WAR) 和 77.7%的未加权准确率(UAR)，相比近期论文有最大 5%的性能提升。并且在注意力机制可视化分析中证明 GATASA 可以有效突出对人类来说容易理解且对分类至关重要的特征，并弱化不重要的特征。

1　交互注意力机制网络 GATASA 和具体算法

1.1　注意力机制

解释 GATASA 前先对注意力机制进行介绍。注意力机制最早被 Bahdanau 等人[12]应用于机器翻译任务上。该技术的思路为通过对特征向量计算权重分数并加权求和，通过不同的权重分数体现特征的重要性，对重要的特征就赋予较大的权重。通常由三部分组成：Query，Key 和 Value。Query 通常是输入的单独向量，而 Key 是多个特征向量。通过点积或可学习参数投影等方法计算出来 Key 和 Query 的相互关系，即注意力机制分数。通过注意力机制分数对 Key 加权求和，得到 Value,公式描述如下：

$$e_i = f(Q, K_i) \tag{1}$$

$$\alpha_i = \frac{\exp(e_i)}{\sum \exp(e_i)} \tag{2}$$

$$Value = \sum \alpha_i * K_i \tag{3}$$

其中式(1)为 Query 和 Key 之间的相似度的计算函数，不同的注意力机制该计算函数不一致，计算出来的相似程度也不一致。

1.2　GATASA

GATASA 由两个部分组成：GATA(global acoustic text acoustic)和 ASATA(acoustic-to-self-acoustic to text)。

GATA 由 Yoon 等人[13]提出，又叫 Multi-hop attention.通过直接将文本和语音的特征交互性地使用注意力机制，让文本和音频直接“选择”对自己来说重要的特征。第一级先以音频特征的最后一个 LSTM 时间步为注意力机制中的 Query,该特征包含了最丰富的，经过激活函数后的高级特征。利用该 Query 向文本特征中寻找重要的文本特征，对得到的文本特征加权求和。第二级用上一级输出的文本特征向量到音频里寻找重要的音频特征。在尽可能保留所有的特征的情况下相互寻找对于语音和文本数据源相互之间重要的部分。GATA 的注意力机制分数计算方法如下

$$e_i = \tanh(Q * K_i) \tag{4}$$

$$\alpha_i = \frac{\exp(e_i)}{\sum \exp(e_i)} \tag{5}$$

$$output = \sum \alpha_i * K_i \tag{6}$$

GATA 并不完美，为此本文用 ASATA 补充。原因有二：

a) 通过直接观察注意力机制的分数，GATA 的注意力机制算法中计算相似度为直接向量做内积，其相似度分数经常很多时候完全一样，并不能有效突出重要特征或者弱化冗余特征。

b) GATA 的第一个 Query 为 LSTM 的最后输出，虽然包含了最丰富的信息量，但是当序列特别长(比如 100 以上)即便是 LSTM 也容易出现梯度消失，重要信息淹没在许多无关词的洪流里。而长度为 100 个词的句子在情感识别中是很常见的。

ASATA 注意于提取局部的关键信息。ASATA 也分为两级，第一级用音频特征的最后一个特征向量在音频信息里利用注意力机制寻找关键的音频信息，对重要特征赋予较大的权重。第二级用加权得到的音频特征向量作为 Query 到文本特征里寻找关键的文本特征。与 GATA 不同,ASATA 通过利用新的权重加权得到的音频特征，可以寻找出来文本特征中不一样的关键信息，这些关键信息为 GATA 所遗漏，但对人类判定情感来说则至关重要，将在后文注意力机制分析中证明。ASATA 的注意力机制分数计算方法如下

$$e_i = \tanh(W * (Q \oplus K_i) + b) \tag{7}$$

$$\alpha_i = \frac{\exp(e_i)}{\sum \exp(e_i)} \tag{8}$$

$$output = \sum \alpha_i * K_i \tag{9}$$

其中的 $\oplus$ 代表向量的拼接。W 和 b 为可学习参数，参与梯度下降。通过加入可学习参数，可以有效缓解 GATA 里注意力机制分数过于相似的问题。

1.3　整个识别系统算法

整个识别系统算法如下：

a) 对音频进行预处理，提取声谱图并归一化处理。将文本送入 BERT 提取所需的词向量。

b) 将提取到的声谱图送入卷积-循环神经网络提取音频的高级特征，将提取到的词向量结合文本内容送入两层循环神经网络提取文本的高级特征。

c) 将得到的音频和文本高级特征送入 GATASA，分别用 GATA 和 ASATA 的第一级计算注意力机制分数。

d) 利用第一级得到的注意力机制分数加权对应的特征，并用加权得到的特征向量去计算第二级的注意力机制的分数和输出向量。

e) 保留第一级和第二级输出的所有特征向量，拼接到一起作为 GATASA 的最终输出。

f) 利用全连接层和 Softmax 函数进行分类，得到输出结果。

2　数据预处理、卷积-循环神经网络

2.1　数据预处理

本文中对音频特征选择的是频谱图。预处理的每一个音频文件为经过降噪，端点分割，无多个说话者重叠且质量较好的，大小约为 100kb 的音频文件。根据文献[3]，利用频谱图可以同时得到音频的时域和频域信息，相比起 MFCC 来省去了人工设计、选择特征的繁琐，且基于合适的后续特征提取可以获得比人工设计特征更好的结果。

音频预处理获得频谱图流程如下：

a) 将音频文件读取进内存，转换为 32 位大小的浮点数并得到采样率。

b) 将音频内容在时间轴上分割为相等长度，前后两段的重叠部分为 50%。

c) 对每一个分割后的片段加上汉宁窗，并取傅里叶变换。

d) 将得到的每一段傅里叶变换的结果按时间顺序合并，得到时频图。其纵轴为时间轴，横轴为频率轴。

对文本特征，本文选择的是利用 Google 发布的 BERT 的隐层输出作为文本特征。BERT 基于大量文本训练得到，其训练所用数据量是前所未有的。通过学习大量文本获得结构化的文本特征，相比起传统词向量 Glove 能有效解决词向量缺失(out of vocabulary)的同时给出较为合理的替代的词向量。因为 BERT 输出的词向量包含了上下文的信息，相比起传统的固定不变的词向量能够一定程度上解决一词多义的问题。

文本预处理如下：

a) 将每个音频文件对应的文本内容记录下来，删除掉其中的标点符号和错误拼写的单词。利用深度学习框架 Keras 自带的 Tokenizer 对文本进行分词和转换成单词序列。

b) 下载 Google 预训练的 BERT 模型并加载到 Python 里。读取模型的后四层 Transfomer 模块的输出，其每一层输出尺寸为 1024 维的向量。

c) 将每一个句子送入 BERT，抽取最后四层 Transfomer 模块的输出并拼接成到一起,得到每个单词对应长度为 4096 的词向量。

2.2　卷积—循环神经网络

本文在音频端使用的特征提取器为卷积-循环神经网络，

而文本端特征提取器为两层循环神经网络。音频端用卷积神经网络提取出来的不同激活区域的特征作为循环神经网络的输入。本文使用较小的卷积核，卷积核尺寸为 5*3。二维卷积操作如下所示。

$$C(s,t)=\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}A(m,n)*B(s-m,t-n) \tag{10}$$

本文使用的循环神经网络为双向的长短期记忆神经网络(long-short term memory，LSTM)，隐层节点数为 100。在 LSTM 中包括遗忘门、输入门、输出门。相比起传统的循环神经网络来 LSTM 能有效的解决梯度消失问题。其工作原理如下所示。遗忘门原理：

$$f_t=\sigma(W_f*[h_{t-1},x_t]+b_f) \tag{11}$$

输入门原理：

$$i_t=\sigma(W_i*[h_{t-1},x_t]+b_i) \tag{12}$$

$$\tilde{C}_t=\tanh(W_c*[h_{t-1},x_t]+b_c) \tag{13}$$

$$C_t=f_t*C_{t-1}+i_t*\tilde{C}_t \tag{14}$$

输出门原理：

$$o_t=\sigma(W_o[h_{t-1},x_t]+b_o) \tag{15}$$

$$h_t=o_t*\tanh(C_t) \tag{16}$$

在循环神经网络后加入了自注意力机制。通过把输出特征投影到不同的空间，让重要的特征被赋予较大权重。其注意力分数计算方法与式(4)相同，只是其中的 Q 和 K 是相同的输入特征。

在 GATA 和 ASATA 输出拼接得到的向量后接全连接层对特征进行投影，得到 256 维大小的最终表示，然后用 Softmax 进行分类。整个神经网络的实现流程如图 1 所示。

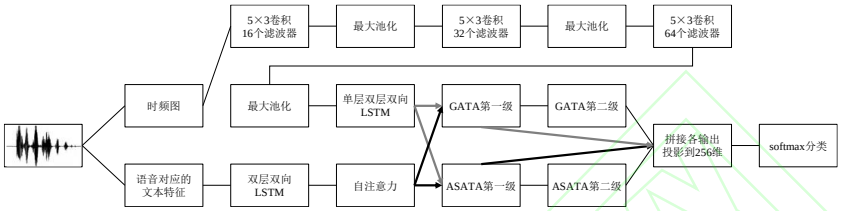

图 1　实验流程图

Fig. 1　A schematic diagram of experiment

情感识别的实现过程：

a) 在现实中获得说话者的录音文件。可以来自现场录音，也可节选自录像。

b) 检查录音质量，如音频采样率太高则对音频降采样，以减少数据预处理的运算量。随后对音频做去噪处理，以降低环境噪声带来的影响，并对音频进行端点分割，将大段的录音分割为一个个长度接近的句子。确保每段录音都只有一个说话者的声音；如存在多个说话者则需要对音频做声源分离。将分段后的句子通过语音识别或人工听写得到文本内容。

b) 将录音文件和文本内容送到数据预处理提取对应的特征。

c) 将得到的特征送入整个卷积-循环神经网络和 GATASA，得到对应的情感分类结果。

3　实验结果与分析

本文选取的数据集为 IEMOCAP(Interactive Emotional Dyadic Motion Capture)数据集,该数据集由南加州大学收集并公开提供。通过预设文本和对应的情景，寻找专业演员录制而成。该数据集总共包含 5 段对话，每段对话为 1 男 1 女演员完成，总共 10 名不同的演员。按照现有文献数据集划分方法，混合了 happiness 分类和 excitement 分类。本文使用主要的 4 个情感分类(1084 句 sad，1103 句 angry，1636 句 happiness 和 1708 句 neutral，总共 5531 句)。

本文的评价指标为加权准确率(WAR)和未加权准确率(UAR)。现有分类任务为 4 分类。WAR 计算方法如下：

$$WAR=\frac{\text{所有分类正确的样本数量}}{\text{总的样本数量}} \tag{17}$$

UAR 计算方法如下，N 为分类数：

$$UAR=\frac{1}{N}*\sum\frac{\text{每个分类正确分类样本数}}{\text{每个分类总样本数}} \tag{18}$$

数据预处理的参数选择如表 1 所示。

表 1　数据预处理参数选择

Tab. 1　Parameter choices of data pre-process

参数类型	参数选择
声谱图尺寸	512*128
声谱图截止频率	4000HZ
词向量尺寸	拼接最后 4 层 BERT 输出 长度为 4096 维的向量

神经网络的参数选择如表 2 所示。

表 2　神经网络参数选择

Tab. 2　Parameter choices of neural network

参数类型	参数选择
循环神经网络节点数	100
卷积层数量	3
滤波器尺寸	5*3
滤波器个数	16，32，64
激活函数	循环神经网络为 Tanh，其余都是 Relu

数据集划分和模型训练：

由于数据集并未规定划分，依据二人对话特性，本文使用 8 个说话者的音频作为训练集，剩余 1 说话者的音频作为测试集，1 说话者的音频作为验证集。由于 5 段对话，采取五折交叉验证。

本文模型利用 Keras 框架基于一块 RTX2080 训练得到。利用验证集在 200 轮训练间来寻找最佳的超参数。在大量实验后发现 Adam 优化器的拟合速度较快，性能也较好；使用较小的学习率(1e-5)可以避免模型在梯度下降时在极小值点附近振荡。详细的超参数如表 3 所示。

表 3　训练参数选择

Tab. 3　Parameter choices of training

参数类型	参数选择
Batch size	40
Learning rate	1.00E-05
Optimizer	Adam
Dropout	0.2
Epoch	200

实验结果如表 4 所示。首先分开对比 GATASA 前网络单独分类的性能。

表 4　单独分类网络结果

Tab. 4　Classification results of backends

分类算法	WAR	UAR
本文的卷积-循环网络	**60.1**	**62**
BN [10]	59.7	61.4
openAIR feature[14]		61.3
本文双层循环网络	**70.8**	**71**
Text-BRE[13]	69.8	70.3

声明“本文”的实验结果为自主实现，其余的实验结果

来自参考文献。从实验结果看来本文选择的声谱图和卷积-循环神经网络的组合，在分类准确率上要优于传统的人工设计特征，大约有 0.4%-0.7%的提升。而本文使用的基于 BERT 提取的词向量比参考文献中实验结果有 0.7%~1%的提升。证明了本文使用新的特征提取前端的优越性。

不同的多模态情感识别的结果如表 5 所示。

表 5　GATASA 性能对比

Tab. 5　Comparison between GATASA and results of literatures

分类算法	WAR	UAR
MHA-2 [13]	76.5	77.6
GATA	75.5	76
ASATA	76	76.8
GATASA	**77**	**77.7**
Proposed in [10]	73.7	75.5
FAF [6]	72.7	72.7

GATA、ASATA 和 GATASA 均为本文实现，其余实验结果来自参考文献。从实验结果来看，本文提出的 GATASA 实验网络的准确率最多可相较参考文献中的结果提升为 4.3%-5.0%，有明显的性能提升。相较于 GATA 融合方法提出的文章，性能也有所提升。而 ASATA 融合方法和 GATA 的融合方法结果相近。需要说明的是 GATA 为 MHA-2 在不同特征提取下的复现，可看到基于不同的前端特征提取，GATA 的实验结果要比 MHA-2 低 1%-1.6%，原因可能来自于卷积神经网络的卷积-池化操作损失了部分信息。由于实验器材和计算能力限制(本文的实验器材仅一块 RTX2080)，并不能像 MHA-2 一样使用不经降维的原始特征(该参考文献里作者使用了不经降维的 MFCC 特征并直接送入循环神经网络)。对比在本文的特征提取后的 GATA 和 ASATA，GATASA 都有明显的性能提升(WAR 上最大提升 1.5%，UAR 上最大提升 1.7%)，证明了 GATASA 融合网络的性能超越了单独的 GATA 或者 ASATA 融合方法。

图 2 为 GATA、ASATA、GATASA 在四种分类上的单独的准确率。可以看到，GATASA 在四种分类上单独的准确率都能超过 GATA 和 ASATA。相比 GATA，GATASA 在 angry，happiness 和 sad 上提升较为明显，而相比 ASATA，GATASA 在 happiness，neutral 和 sad 上提升较为明显。而 GATA 和 ASATA 在准确率上各有高低：在 happiness 和 neutral 上 GATA 的准确率较高，但 angry 和 sad 上 ASATA 的准确率稍好。由于 angry 和 sad 两种情绪的特征较为明显(angry 可能有高昂的语调，较大的声音，而 sad 可能会伴随着抽泣)，这两种分类的情绪相较于 happiness 和 neutral 准确率稍高。

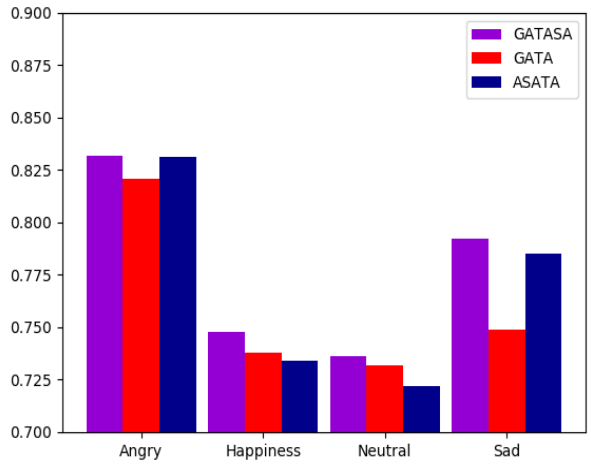

图 2　四种分类的准确率

Fig. 2　Accuracies of 4 categories

为了模拟现实生活中效果，表 6 中实验不使用语音对应的完全正确的文本，而是使用 Google 发布的翻译 API 将对话音频识别为实际文字，在完全相同的神经网络和音频特征下再次进行实验。如表 6 所示，GATASA 对语音识别出来的文本依然有性能上的提升，最大提升为 6.9%。除 GATASA 外的实验结果来自参考文献。

表 6　语音识别文本下 GATASA 性能对比

Tab. 6　Comparison between ASR-GATASA and ASR results of literatures

分类算法	WAR	UAR
GATASA(ASR)	**73.5**	**75**
proposed in [10] (ASR)	66.6	68.7
MHA-2-ASR[13]	73	73.9

本文也对比了基于 BERT 的词向量和基于 Glove 的词向量在不同注意力机制下的性能区别，分别在 GATA，ASATA,GATASA 下进行了实验。以下实验结果均为本文实现，在同一个神经网络架构下，区别仅是使用的词向量。Glove[7] 来自斯坦福大学公开发布的 Glove 词向量数据集，本文选用其中有最大训练数据量和维度的词向量。如表 7 所示。

表 7　BERT 与 Glove 词向量性能对比

Tab. 7　Comparison between BERT and Glove word vector

分类算法	WAR	UAR
GATASA-BERT	77	77.7
GATASA-Glove	76.6	77.3
GATA-BERT	75.5	76
GATA-Glove	75	75.8
ASATA-BERT	76	76.8
ASATA-Glove	76.2	76.4

从结果上看 BERT 提取的词向量的确相比于 Glove 能够有所提升，在 GATASA 下有 0.4%的提升。而在 GATA 和 ASATA 的融合方式下，BERT 和 Glove 性能区别不大：GATA 下 BERT 略好于 Glove,有最大 0.5%的准确度区别；而 ASATA 下 BERT 在 WAR 上略弱于 Glove，为 0.2%。原因是 Glove 仅包含常用单词，导致训练文本中有一些单词无法寻找到对应的词向量表示(out of vocabulary)，性能有所下降；而 BERT 可以动态提取需要的词向量，基于其庞大的训练量，对未曾见过的单词也可给出较为合理的表示。不论是 BERT 词向量还是 Glove 词向量，GATASA 的准确度都明显优于 GATA 和 ASATA，证明了 GATASA 可以带来有效的性能提升。

除词向量外本文还对比了不同的音频特征和卷积神经网络设计的结果。卷积-循环神经网络与表 4 中本文的单独分类网络完全一致，仅有输入数据上的区别。结果如表 8 所示。

表 8　不同音频特征分类结果

Tab. 8　Classification results based on different speech features

分类算法	WAR	UAR
512*128 时频图	**60.1**	**62**
256*64 时频图	59.5	59.8
128*32 时频图	59.3	60.5
64*64 时频图	58.8	59.1
1024*256 时频图	60	61.5
512*128 MFCC	60	60.5

可以看到在相同的分类网络下，当时频图尺寸在一定范围内增大，分类的准确率也随之提升，因为更大的时频图带来了更多的局部细节；但时频图也不是越大越好：可以看到当时频图尺寸增大到 1024*256 后，准确率有所下降。其原因是时频图增大后带来了更多的局部细节，也带来了噪声，影响了最后的分类准确率；尺寸增大也带来了更多的计算开销，增加了训练时长。而同样尺寸下，时频图有比 MFCC 更好的准确度。

基于部分表 8 中的不同尺寸的音频特征，在 GATASA 上做了对比实验，结果如表 9，仅有输入的音频特征尺寸上的区别，模型均如图 1。可以看到用 512*128 尺寸的时频图可以获得最好的 GATASA 分类结果。同样尺寸的 MFCC 特征图效果略微差一些。随着时频图尺寸变小，准确率也随着有所下降。

表 9　不同音频特征在 GATASA 下分类结果

Tab. 9　GATASA classification results based on different speech features

分类算法	WAR	UAR
512*128 时频图	**77**	**77.7**
256*64 时频图	76.5	76.9
128*32 时频图	76.5	76.8
512*128 MFCC	76.8	77

4　对关键信息提取的分析

本实验利用每个词对应的词向量和注意力分数分析关键信息的提取。相比起作为音频特征的时频图，文本内容容易被人类理解，也包括了更多的对话背景信息，用来解释情感较为直观，故本文用注意力机制对每个文本特征(词向量)的重要性打分来解释关键信息的提取。

情感识别中，如何定义、量化人类的情感一直是一个没有定论的问题。传统的文本情感分类做法上多基于情感词典对文本进行分类：利用预训练或人工标定的每个词的感情分数来对整个句子打分。

基于情感词典的做法主要存在一些问题：每个词只有有限个分数值，对一词多义、不同上下文内容、对每个词内含的背景知识这些问题的解决都不够完善。本文以常用的文本处理工具 NLTK 中的情感词典 SentiWordNet 对应的情感分数和注意力机制分数的分布对关键信息的标注进行对比。SentiWordNet 提供的是情感分数，只有两个分类：积极和消极。对于积极则打分为正，对于消极则打分为负，中性则是 0。对于关键特征，不论是积极还是消极，都应当对分辨句子情感产生贡献，即情感分数不为 0。而注意力机制的分数则直接显示了对应特征的重要性：本文注意力机制可视化中颜色越深，分数越大，则对应特征越重要。

对句子“I would never marry you again I would rather die in torment”，该句子的情感分类为 angry。图 3 是 GATA 和 ASATA 的注意力机制分数在该句子的文本特征上的分布，表 10 为 SentiWordNet 对该句子的情感分数的打分。在该句子下，文本内容里的重要信息应当是 never、die 这样的词，而其中的 marry、两次出现的 would 根据人类的认识也可以起到强调的作用。在 SentiWordNet 的打分里，如表 10 所示，两次出现的 would 和 marry 贡献的情感分数为 0。但根据图 3 里的 GATA，两次出现的 would 的显示了不一样的重要性，第一次的可视化颜色较深，分数较大，第二次较浅，分数较小；而 marry 颜色也较深，说明也为区分情感提供了帮助。从语义分析第一次的 would 是强调“never marry again”这个意向，第二次只是补充说明，所以第一次对情感的强调比第二次要重要，符合注意力机制打分的结果。而 marry 则是指代 never 对应的事件，反映了事件的重要性，起到强调强烈的拒绝愿望。基于情感词典 SentiWordNet 的打分则让很多有强调意义和背景事件、先验知识意义的词汇失去了贡献，无法利用上下文信息，每个词都是孤立的。GATASA 则通过利用完全的上下文信息合理的找到了确定情感需要的关键特征。

为了说明 ASATA 和 GATA 是互补的，本文直接对比 ASATA 和 GATA 的注意力机制分数。从图 3 可以看到只有 GATA 的情况下，注意力机制只重点关注了 never、again，而对于 rather、torment 这样具有重要意义的词，GATA 没有给予足够的重视。与之相对，ASATA 的注意力机制分数在 rather、torment 里给出了更大的权重，让整个句子里 rather、torment 的重要性更突出。即两种注意力机制可以互相找到对区分情感重要，互补的关键特征。

表 10　angry 句子的 SentiWordNet 情感分数

Tab. 10　Emotion score of an ‘angry’ sentence based on sentiwordnet

情感单词	词频	分数
Would	2	0
Never	1	-0.4583
Marry	1	0
Rather	1	-0.055
Die	1	0.1172
Torment	1	0.5391

对句子“It ‘s a beautiful campus It’s gonna be really nice um and everyone”，该句子的情感类别为 happiness。图 4 为该句子的注意力机制分数，表 11 是利用 SentiWordNet 对该句子内容进行情感打分的结果。由表 11 可知，根据情感词典的方法对 gonna 的情感分数是 0，但 gonna 这样在口语中具有强调“将要，一定会”的单词能对情感的识别起到辅助，强调某件事的作用。而 campus 往往让人联想到美好的大学生活，且 campus 用 beautiful 修饰，则 campus 也应当对情感识别起到辅助作用，但 SentiWordNet 对 campus 的情感分数是 0，即没有贡献。这些结果都和人类对情感的认知和判别不相符合。从图 4 里，在 GATASA 的 ASATA 里，ASATA 强调了 gonna be 和 campus 这样的词，符合人类的对情感的理解和背景条件的认知。

表 11　happiness 句子的 SentiWordNet 情感分数

Tab. 11　Emotion score of an ‘happiness sentence based on sentiwordnet

情感单词	词频	分数
beautiful	1	0.7083
campus	1	0
gonna	1	0
really	1	0.47
nice	1	0.7098
um	1	0
everyone	1	0

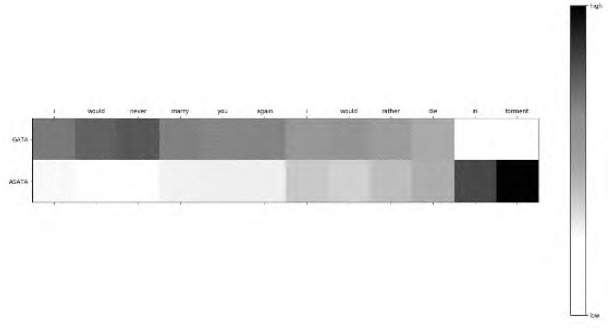

图 3　对应 angry 情感的句子的注意力分数

Fig. 3　Attention scores of an “angry” sentence

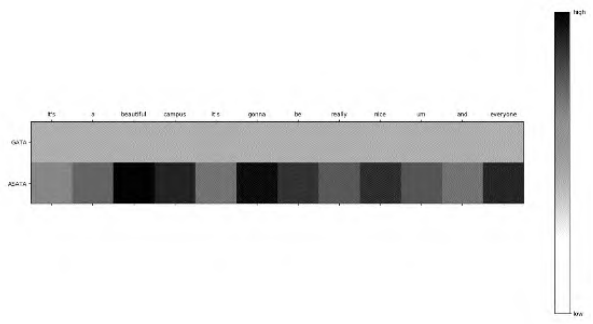

图 4　对应 happiness 情感的句子的注意力分数

Fig. 4　Attention scores of an “happiness” sentence

在 GATASA 内部的对比中，图 4 中可以看到，对于 GATA，它的注意力分数对所有文本几乎一样，并没有突出关键的文本信息；而 ASATA 通过可学习参数的帮助学习到了关键的文本信息：给予了 beautiful 和 nice 很高的权重，对 campus 和 gonna 这样的辅助、背景含义单词也给了较高的重要性，有效突出了关键文本信息。即 ASATA 可以找到 GATA 遗漏的特征，从而提高可解释性和准确率。

综上所述，GATA 在多模态情感识别里寻找关键的文本信息时候并不能完全提取到所有需要的文本信息。而 ASATA 能找到部分被 GATA 遗漏的信息，二者互补，相辅相成。

5 结束语

本文提出了一种基于多注意力机制融合的多模态情感识别模型融合方法 GATASA。通过设置两种不同的注意力机制，一种关注于尽可能多的保留特征信息，而另一种关注于尽可能多的寻找出关键的局部信息，互补地提取对于情感识别而言至关重要的特征部分。通过搭建卷积-循环神经网络提取重要特征，在音频和文本单独分类上都获得了提升。实验证明该多模态情感识别融合方法在 IEMOCAP 数据集上能够有效提升识别准确率，并通过注意力机制分析证明该融合方法能够直观的找出来对于情感识别至关重要的信息，相比起传统基于情感词典的方法，能够更好的利用上下文信息确定情感。未来将会关注于进一步改进关键特征提取方法；利用提取到的关键特征对情感进行更深入的分析，以求对情感进行定义。

参考文献：

[1] Hazarika D, Poria S, Mihalcea R, *et al.* ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection [C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

[2] Wang Fei, Ye Xiaofeng, Sun Zhaoyu, *et al.* Research on speech emotion recognition based on deep auto-encoder [C]// IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems, 2016, pp. 308-312.

[3] Satt A, Rozenberg S, Hoory R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms [C]// Conference of the international speech communication association, 2017: 1089-1093.

[4] Verkholyak O, Fedotov D, Kaya H, *et al.* Hierarchical Two-level Modelling of Emotional States in Spoken Dialog Systems [C]// IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 6700-6704.

[5] Reddy B S, Kumar T K. Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC [C]// The 8th International Conference on Computing, Communication and Networking Technologies, 2017: 1-5.

[6] Fu Liqiu, Mao xia, Chen Lijiang. Speaker independent emotion recognition based on SVM/HMMS fusion system [C]// International Conference on Audio. IEEE, 2008.

[7] Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation [C]// Conference on Empirical Methods in Natural Language Processing. 2014: 1532-1543.

[8] Devlin J, Chang M W, Lee K, *et al.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. arXiv: Computation and Language, 2018. https://arxiv. org/abs/1810. 04805.

[9] Ghosal D, Akhtar M S, Chauhan D, *et al.* Contextual Inter-modal Attention for Multi-modal Sentiment Analysis [C]// Conference on Empirical Methods in Natural Language Processing. 2018: 3454-3466.

[10] Kim E, Shin J W. DNN-BASED EMOTION RECOGNITION BASED ON BOTTLENECK ACOUSTIC FEATURES AND LEXICAL FEATURES [C]// IEEE International Conference on Acoustics, Speech and Signal Processing, 2019: 6720-6724.

[11] Poria S, Cambria E, Hazarika D, *et al.* Context-dependent sentiment analysis in user-generated videos [C]// The 55th annual meeting of the association for computational linguistics, 2017: 873-883.

[12] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate [J]. arXiv: Computation and Language, 2014. https://arxiv. org/abs/1409. 0473.

[13] Yoon S, Byun S, Dey S, *et al.* Speech emotion recognition using multi-hop attention mechanism [C]// IEEE International Conference on Acoustics, Speech and Signal Processing, 2019: 2822-2826.

[14] Poria S, Chaturvedi I, Cambria E, *et al.* Convolutional MKL based multimodal emotion recognition and sentiment analysis [C]// IEEE International Conference on Data Mining (ICDM) , 2016: 439-448