考虑视图可信度的用户多模态意图识别方法

杨 颖*①② 杨艳秋①② 余本功①③
①(合肥工业大学管理学院 合肥 230009)
②(合肥工业大学过程优化与智能决策教育部重点实验室 合肥 230009)
③(智能决策与信息系统技术教育部工程研究中心 合肥 230009)

摘 要：在人机交互的闲聊型对话中，准确理解用户多模态意图有助于机器为用户提供智能高效的聊天服务。当前的用户多模态意图识别方法面临着跨模态信息交互性与模型不确定性的挑战。该文提出一种基于Transformer的可信多模态意图识别方法。考虑用户意图表达时的文本、视频和音频等数据的异质性，通过模块特定编码模块，生成单模态特征视图；为了捕捉跨模态间的互补性和长距离依赖性，通过跨模态交互模块，生成跨模态特征视图；为了降低模型的不确定性，设计一个多视图可信融合模块，考虑每个视图的可信度进行主观意见的动态融合，基于主观意见的Dirichlet分布，设计一种组合优化策略进行模型训练。最后在多模态意图识别数据集MIntRec上进行实验。实验结果表明，与基线模型相比，该文方法在准确率和召回率上分别提升了1.73%和1.1%。该方法不仅能够提升多模态意图识别的效果，而且能够对每个视图预测结果的可信度进行度量，提高模型的可解释性。

关键词：意图识别；多模态融合；多视图学习
中图分类号：TN911.7; TP391 **文献标识码**：A **文章编号**：1009-5896(2025)06-1966-10
DOI: 10.11999/JEIT240778 **CSTR**: 32379.14.JEIT240778

1 引言

移动终端和语音识别技术的发展推动了人机交互系统的普及。相比任务型对话系统，开放域闲聊系统更注重情感交流，能满足陪伴、共情等情感需求，而多模态意图识别在其中越来越重要。例如，在智能客服系统或智能座舱场景中，结合语音、文本中用户的信息，可以更精准地理解用户需求并提供个性化服务。研究人员[1]发现在现实世界的社交对话中，除了对话文本外，人们的表情和肢体动作等非语言符号也蕴含着丰富的用户意图信息。仅依赖文本易受信息不足、歧义等问题限制，融合多模态信息可提高意图识别准确率，从而实现更智能的人机交互服务。

近年来，深度学习与多模态融合技术已广泛应用于情感分析[2]和讽刺检测[3]等自然语言理解的任务中。然而，现有多模态意图识别研究往往关注社交媒体平台[4]、报刊新闻平台[5]或教学直播平台[6]等场景下展示出的文档或视频意图，对真实社交对话场景关注不足。随着聊天机器人技术的发展，人们越来越关注社交对话环境中的多模态意图，例如电视剧[1]中真实聊天场景下的人类意图的自动识别方法，有助于提高人机交互的效果。与传统任务型对话不同，社交闲聊场景具有开放性、参与对话者多、短文本及高噪声等特点。在进行社交对话意图识别过程中，不仅要考虑模态间信息的互补性，还要考虑由于噪声信息带来的模态特征视图的可靠性。因此，在闲聊场景下的多模态意图识别主要面临两个挑战：(1)如何在保留模态内部的特定信息的同时有效地融合多模态异质信息以实现互补；(2)多模态数据中的噪声导致不同模态在不同场景中的可靠性不同，如何动态评估每个视图的可靠性并据此进行融合有待进一步研究。

为了解决上述问题，本文提出一种面向社交对话的可信多模态意图识别(Trusted Multimodal Intent Recognition, TMIR) 方法。首先，考虑多模态数据的异质性及其交互关系设计了跨模态交互模块，使不同模态数据充分交互获得模态间的互补信息，同时模态特定编码模块保留了模态内部的特有信息；其次，设计了多视图可信融合模块，考虑视图的可靠性进行主观意见的动态融合。该模块在输出意图识别结果的同时，可以给出决策结果的可信度以及单个视图对于不同样本的可信度。最后在多模态意图识别数据集MIntRec[1]上进行了实验，实验结果与基线模型相比在多个评价指标上都有所提升，验证了本文所提方法的有效性。

2 相关研究

用户意图识别通常是指将用户表达的意愿和需

收稿日期：2024-09-10；改回日期：2025-05-21；网络出版：2025-05-28
*通信作者： 杨颖 yangying@hfut.edu.cn
基金项目：国家自然科学基金(72071061)
Foundation Item: The National Natural Science Foundation of China (72071061)

求映射到预定义的意图类别中。目前的人机对话系统主要包括任务型对话和闲聊型对话等两种类型。任务型对话中的意图识别往往和槽位填充任务一起进行。闲聊型的对话系统中，人机对话主要是在开放域环境中进行情感交流和针对客观问题的讨论，用户意图往往伴随着情绪识别任务[7]。本文将主要聚集闲聊型对话情景下的用户多模态意图识别研究。

当前意图识别的方法主要包括基于规则的方法、传统机器学习方法和深度学习方法等，根据语音或文本等单模态信息对用户意图进行判断。例如，钱岳等人[8]提出了一种卷积长短期记忆网络(Convolutional Long Short-Term Memory, Convolutional-LSTM)，对用户聊天文本中的出行消费意图进行识别。随着多模态语言理解技术在情感分析和讽刺检测等任务上得到越来越多的应用，多模态意图识别的相关研究也相继出现。例如，Kruk等人[4]考虑Instagram平台上图像和文本标题信息，对作者的发布意图进行分类。Zhang等人[5]对报刊新闻媒体平台发布内容中的营销意图进行了分析，提出了一种图卷积网络模型来建模广告中图像和文本之间的交互关系。Maharana等人[6]提出了一种多模态交叉注意力模型，结合了视频和文本识别直播视频中的教学意图。现实世界中人类的社交对话往往具有内容短小、参与的说话人较多且噪声干扰较多等特点。Zhang等人[1]基于电视剧中的人物对话构建了多个真实生活场景下的多模态意图识别基准数据集MIntRec，并分别采用了多模态情感分析中的多模态Transformer (MULtimodal Transformer MULT)[9]、模态不变与模态特定的多模态情感分析表示(Modality-Invariant and Specific representations for multimodal sentiment Analysis, MISA)[10]等方法进行意图识别。在该数据集上Huang等人[11]提出了一种基于注意力门控神经网络的自适应的多模态融合方法进行意图识别。除了数据集MIntRec外，Singh等人[7]基于多种体裁的电影构建了多模态对话数据集EmoInt_MD，并提出一种同时识别说话者的情绪及其意图的多任务多模态情境Transformer网络。

然而，上述的多模态意图识别方法往往假设每个模态的数据质量是一样的，忽略了特征视图中存在的不确定性。在现实的社交聊天环境下，当一些样本的模态数据存在较大噪声时，模型学习到的特征视图对于不同样本来说可信度也不同。而现有的多模态融合方法往往是同等对待这些特征视图，不去区分由于噪声感知不同而导致的特征视图表示上的差异性。传统贝叶斯方法通过概率分布来建模不确定性，但这种方法通常需要大量的训练数据来估计先验概率和似然函数，且计算复杂度较高。与贝叶斯方法不同，基于Dirichlet的方法通过引入主观逻辑理论来预测迪利克雷分布的参数，实现对不确定性的量化，避免了较大的计算负担，基于Dirichlet的方法[12]已成为深度学习不确定性估计的一个重要分支。本文借鉴可信多视图学习的思想，首先对来自不同模态的特征视图及其可信度进行建模，根据不同数据样本的噪声程度，学习得到特征视图的可信度，然后采用不确定信息融合理论构建一个端到端的可信多视图融合网络，从而提高社交聊天对话中的用户多模态意图识别效果。

3　可信多模态对话意图识别方法

本文提出可信多模态意图识别模型TMIR，用于处理社交聊天型对话中的文本、音频和视频等多模态信息，模型结构如图1所示。多模态特征表示层通过预训练模型提取特征向量；多视图特征提取层生成单模态特征视图和跨模态特征视图；可信融合层基于主观逻辑理论，建模可信度并通过Dempster规则融合视图意见，最终输出分类概率和可信度，实现多模态意图识别。下文将详细介绍TMIR模型的建模过程。

3.1　多模态特征表示

多模态特征表示旨在将不同类型的原始数据转化为更适合模型处理、能有效表征数据特征的形式。在传统的多模态特征表示方法中，梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficients, MFCCs)常被用于提取音频模态的特征，词向量(Word to Vector, Word2Vec)、词频-逆文档频率(Term Frequency-Inverse Document Frequency, TF-IDF)等用来进行文本特征表示。然而这些传统的方法往往难以捕捉到模态内深层次的语义信息。由于神经网络具有强大的表征能力，基于深度学习的特征表示方法成为主流。因此在多模态特征表示层，通过预训练语言模型双向编码器表征(Bidirectional Encoder Representations from Transformers, BERT)、波形到向量 2.0模型(Waveform to Vector 2.0, Wav2Vec2.0)[13]和预训练图片模型区域卷积神经网络(Faster Regions with Convolutional Neural Networks, Faster-RCNN)分别将文本、音频和图片等多模态语言数据转换为特征向量。3个模态的特征表示分别为 $\hat{\boldsymbol{X}}_{\mathrm{T}} \in R^{L_{\mathrm{T}} \times d_{\mathrm{T}}}$，$\hat{\boldsymbol{X}}_{\mathrm{V}} \in R^{L_{\mathrm{V}} \times d_{\mathrm{V}}}$ 和 $\hat{\boldsymbol{X}}_{\mathrm{A}} \in R^{L_{\mathrm{A}} \times d_{\mathrm{A}}}$。其中 $L_m \in (\mathrm{T}, \mathrm{V}, \mathrm{A})$ 表示序列长度，$d_s \in (\mathrm{T}, \mathrm{V}, \mathrm{A})$ 表示特征维数。

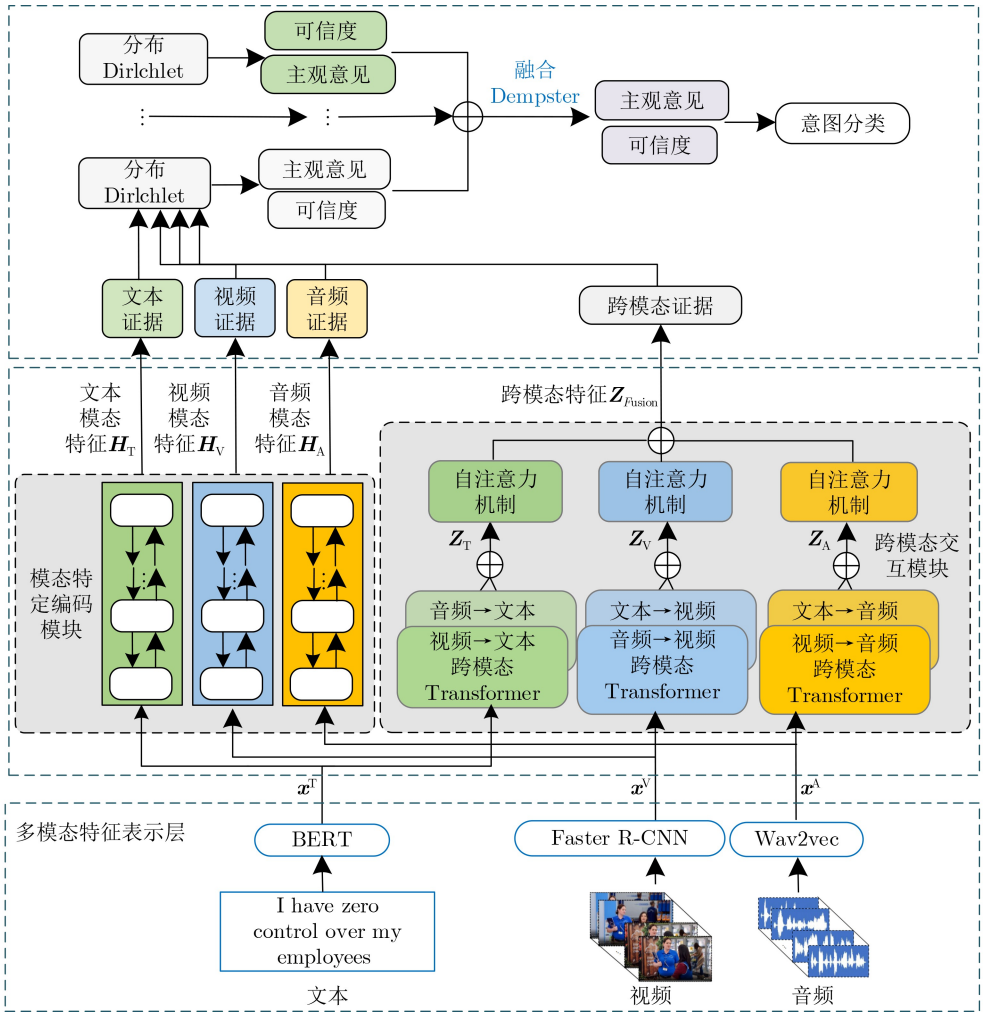


图 1 可信多模态意图识别模型框架

随后本文采用1维卷积神经网络来分别提取3个模态的局部特征，在保留关键特征的同时降低维度，如式(1)所示。其中，$f_{\{\mathrm{T,V,A}\}}$是3个模态卷积核的大小，d是特征向量维数

$$\boldsymbol{X}_{\{\mathrm{T,V,A}\}}=\mathrm{Conv1D}\left(\hat{\boldsymbol{X}}_{\{\mathrm{T,V,A}\}},f_{\{\mathrm{T,V,A}\}}\right)\in R^{L_{\{\mathrm{T,V,A}\}}\times d} \tag{1}$$

3.2 多视图特征提取

为了充分利用多模态数据的互补性和多样性，本文设计了跨模态交互模块和模态特定编码模块。前者是在得到多模态数据的特征表示后，进一步提取跨模态特征来学习模态间的交互关系；后者是提取单模态特征来获得模态内部的特有信息，进而为意图识别提供更丰富的特征表示。

3.2.1 跨模态交互模块

跨模态交互模块通过跨模态注意力机制，使不同模态的数据充分交互，捕捉模态间的互补性和依赖性。具体来说，通过学习辅助模态的特征表示来增强目标模态的特征表示，可以提高多模态数据的融合效果。具体来说，该模块构建跨模态Transformer模型(Cross-Modal Transformer, CMT)生成跨模态特征视图。CMT的核心是跨模态注意力机制，通过学习辅助模态的特征表示来增强目标模态的特征表示。将3种模态进行两两组合，本文共构建了6个CMT模型。其中，以视频-文本CMT模型为例，其接收视频和文本两个模态的特征作为输入，分别用$\boldsymbol{Z}_{\mathrm{V}}^{[0]}$和$\boldsymbol{Z}_{\mathrm{T}}^{[0]}$表示初始的视频模态和文本模态输入。经过多头跨模态注意力后，将得到的融合了视频特征之后的文本特征向量通过残差连接、归一化层和前馈层，得到第i层的输出$\boldsymbol{Z}_{\mathrm{V}\to\mathrm{T}}^{[i]}$。其中前馈层是由两层线性层组成的前馈神经网络。CMT模型将重复进行多次这样的计算，使文本和视频模态的信息充分进行交互融合。经过这个过程，目标文本模态通过视频模态的信息来增强自身的表征能力。将最后一层得到的经过视频模态向量增强后的文本向量记为$\boldsymbol{Z}_{\mathrm{V}\to\mathrm{T}}\in R^{L_{\mathrm{T}}\times d}$。

同理，文本、视频和音频中的任意一种模态均可作为目标模态，与其它模态进行跨模态相互作用。将目标模态相同的CMT的输出向量拼接起

来，分别记为$\boldsymbol{Z}_{\mathrm{A}}$,$\boldsymbol{Z}_{\mathrm{T}}$和$\boldsymbol{Z}_{\mathrm{V}}$，例如$\boldsymbol{Z}_{\mathrm{A}}$= Concat($\boldsymbol{Z}_{\mathrm{V}\to\mathrm{A}}$, $\boldsymbol{Z}_{\mathrm{T}\to\mathrm{A}}$) $\in R^{L_{\mathrm{T,V,A}}\times 2d}$。采用自注意力机制来强化其特征表示，使得模型能够捕捉序列中不同元素之间的依赖关系。最后将自注意力机制后的输出拼接起来作为跨模态视图最终的特征表示，记为$\boldsymbol{Z}_{\mathrm{Fusion}}$。

3.2.2 模态特定编码模块

在模态特定编码模块，通过3个模态特定的双向长短时记忆网络(Bidirectional Long and Short Term Memory, Bi-LSTM)，提取每个模态内部的特有信息，保留模态内部的特有信息，避免模态内的信息缺失。这有利于捕捉每个模态的独特性，提高多模态数据的表示能力。

跨模态视图虽然捕捉了模态间的交互信息，却忽略了单模态内部特有的重要信息，因此我们设计了模态特定编码模块，将多模态特征映射到模态特定的特征空间，捕捉每个模态的独特性。该模块包含3个Bi-LSTM，分别对文本、音频和视频特征在正反两个方向上进行建模，提取各个模态在时间序列上的前向和反向的依赖关系，以提取全局特征。以音频单模态视图的获取过程为例，如式(2)所示

$$\boldsymbol{H}_{\mathrm{A}}=\mathrm{BiLSTM}\left(\boldsymbol{X}_{\mathrm{A}};\theta_{\mathrm{A}}^{\mathrm{BiLSTM}}\right)\in R^{d_{\mathrm{A}}} \tag{2}$$

其中，$\theta_{\mathrm{A}}^{\mathrm{BiLSTM}}$表示网络隐藏层的参数，$\boldsymbol{H}_{\mathrm{A}}$是将正向和反向LSTM网络最后一个隐藏状态的输出拼接后得到的音频单模态视图的特征表示。同理，我们可以得到视频单模态视图$\boldsymbol{H}_{\mathrm{V}}$和文本单模态视图$\boldsymbol{H}_{\mathrm{T}}$。

3.3 多视图可信融合

在得到上述用户意图的单模态和跨模态视图特征表示之后，依据主观逻辑理论，本文将其转化为证据，并对多视图的可信度进行测量，得到每个视图相应的主观意见，最后通过主观意见的融合得到用户意图识别的结果。整个决策过程可以分为如下4个阶段。

(1)视图证据的生成：**定义1**(视图证据)：视图证据是对用户意图识别结果的类概率分布表示。它是从用户多模态数据中获取到的多视图向量通过非线性变换后得到。它是一个非负向量，表示为$\boldsymbol{e}$= [$\boldsymbol{e}^1$,$\cdots$,$\boldsymbol{e}^k$,$\cdots\boldsymbol{e}^K$]。其维度等于意图类别数，每个分量都与意图类别一一对应，且取值越高意味着样本属于该意图类别的概率越大。

以文本视图的特征向量$\boldsymbol{H}_{\mathrm{T}}$为例，为了能将文本视图的特征转化为证据，首先使用全连接神经网络将文本视图的特征映射到视图证据对应的维空间中。然后采用非负激活函数Softplus将文本视图的特征向量$\boldsymbol{H}_{\mathrm{T}}$转化为跨模态证据$\boldsymbol{e}^1$，如式(3)所示

$$\boldsymbol{e}^1=\mathrm{Softplus}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{H}_{\mathrm{T}}+\boldsymbol{b}^{\mathrm{T}}\right)\in R^K \tag{3}$$

其中，$\boldsymbol{W}^{\mathrm{T}}$表示全连接层参数的转置，$\boldsymbol{b}^{\mathrm{T}}$为偏置项。同样，可以分别从视频、音频等单模态视图和跨模态视图获得证据$\boldsymbol{e}^2$, $\boldsymbol{e}^3$和$\boldsymbol{e}^4$，至此我们一共得到四条视图证据。

(2)主观意见的生成：引入Dirichlet分布来描述上述证据中的不确定性。它可以看作是分类概率分布的共轭先验分布。根据Dirichlet分布获得不确定性，可以缓解置信度偏高的情况，从而提供更可信的预测。Dirichlet分布的参数是一个非负的K维向量$\boldsymbol{\alpha}$=[α_1,α_2,$\cdots$,α_K]，其中参数$\boldsymbol{\alpha}$控制了各个类别在总体分布中的权重和比例。为了获得每个视图对应的意图分类概率的Dirichlet分布，并满足参数$\boldsymbol{\alpha}$非负的要求，在形式上采取$\boldsymbol{\alpha}=\boldsymbol{e}+\boldsymbol{a}K$的形式作为Dirichlet分布的参数[14]。其中$\boldsymbol{e}$表示获得的证据。$\boldsymbol{a}=[a_1,a_2,\cdots,a_K]$代表对第$k$类意图的优先偏好，且$\sum_{i=1}^{K}a_i=1, 0\le a_1,a_2,\cdots,a_K\le 1$，因为本文对类别不存在偏好时，因此$\boldsymbol{a}$中的每一个分量都取值为$1/K$。于是在本文中Dirichlet分布的参数$\boldsymbol{\alpha}=\boldsymbol{e}+\boldsymbol{1}$。

定义2(主观意见)：主观意见表示由一个特征视图产生的带有不确定性的决策结果。它是一个包含视图不确定性、每个意图类别的置信度以及类别偏好的组合，表示为$\omega=\{\boldsymbol{b},u,\boldsymbol{a}\}$。主观意见与迪利克雷分布之间存在映射关系$\omega=\{\boldsymbol{b},u,\boldsymbol{a}\}\leftrightarrow \mathrm{Dir}\left(\boldsymbol{P}|\boldsymbol{\alpha}\right)$，其中$\boldsymbol{b}=[b_1,b_2,\cdots,b_K]$表示意图识别的置信度分布，$u$表示识别结果的不确定性。它们和证据、Dirichlet分布参数$\boldsymbol{\alpha}$之间的数学关系如式(4)和式(5)[12]

$$u+\sum_{k=1}^{K}b_k=1 \tag{4}$$

$$S=\sum_{k=1}^{K}(e_k+1)=\sum_{i=1}^{K}\alpha_k,b_k=\frac{e_k}{S}=\frac{\alpha_k-1}{S},u=\frac{K}{S} \tag{5}$$

其中，S表示Dirichlet分布的强度。对于第$\boldsymbol{k}$类意图，其获得的证据越多，分配的信任度b_k越高。反之，整体获得的总证据加和越少，总体的不确定性越高。主观意见的生成使得本文可以更加灵活地融合多个视图，获取了视图的不确定性也可以提高模型的可解释性和可靠程度。下面本文定义了可信度。

(3)视图的可信度计算：**定义3**(视图可信度)：视图可信度t是用于衡量不同视图对用户意图识别结果的可靠程度，可信度取值越大，表明通过该视图得到的识别结果越可靠。第i个视图的可信度t^i和Dirichlet分布参数α_k^i、不确定性u^i以及分类置信度b_k^i之间的关系如式(6)所示

$$t^i = \frac{\sum_{k=1}^{K}(\alpha_k^i - 1)}{\sum_{k=1}^{K}\alpha_k^i} = \sum_{k=1}^{K} b_k^i = 1 - u^i \tag{6}$$

(4)主观意见融合：在得到每个视图对应的主观意见后需要将其进行融合，本文引入了一种简化的Dempster融合规则[12]。假设有两个主观意见分别为 $\boldsymbol{\omega}^1=[b_1^1,b_2^1,\cdots,b_K^1],u^1,[a_1^1,a_2^1,\cdots,a_K^1]$ 和 $\boldsymbol{\omega}^2=[b_1^2,b_2^2,\cdots,b_K^2],u^2,[a_1^2,a_2^2,\cdots,a_K^2]$，且类别偏好均取值为$1/K$，融合后的主观意见 $\boldsymbol{\omega}=[b_1,b_2,\cdots,b_K],u,[a_1,a_2,\cdots,a_K]$ 在形式上和单个主观意见相同。

在主观意见的融合过程中，最终决策结果的可信度t与融合前的单个视图的可信度相比，获得了提高。如式(7)所示

$$\begin{aligned} t &= 1-u = 1 - \frac{u^1u^2}{\sum_{k=1}^{K} b_k^1 b_k^2 + u^1 + u^2 - u^1u^2} \\ &\geq 1 - \frac{u^1u^2}{u^1+u^2-u^1u^2} \geq t^1 \end{aligned} \tag{7}$$

其中，u^1和u^2为两个视图主观意见的不确定性估计，u为这两个意见融合后的观点的不确定性估计，该式表明了融合后的不确定性估计小于融合之前的不确定性估计。本方法构建了多个视图，通过融合这些视图对应的多个主观意见，可以有效降低不确定性，从而提高最终意图预测结果的可靠性。

3.4 优化策略

本文提出一种可信多视图优化策略，在获得某样本第i个视图的证据e后，可以得到Dirichlet分布的参数 $\boldsymbol{\alpha}^i=[\alpha_1^i,\alpha_2^i,\cdots,\alpha_K^i]$，将交叉熵函数调整为交叉熵的Dirichlet分布的期望[12]

$$\begin{aligned} L_{\mathrm{ace}}(\alpha_i) &= \int \left[\sum_{k=1}^{K} -y_{ik}\ln(p_{ik})\right] \frac{1}{B(\alpha_i)} \prod_{k=1}^{K} p_{ik}^{\alpha_{ik}-1} \cdot \mathrm{d}p_i \\ &= \sum_{k=1}^{K} y_{ik}\left(\psi(S_i) - \psi(\alpha_{ik})\right) \end{aligned} \tag{8}$$

p_{ik}表示从某样本第i个视图预测其属于第k类的概率，从视图得到的主观意见$\omega^i=\{\boldsymbol{b}^i,u^i,\boldsymbol{a}^i\}$中得到，即$p_{ik}=b_k^i+a_k^iu^i$。$y_{ik}$是真实标签值，$K$为总的意图类别数。该式由交叉熵损失在参数为$\alpha_i$的Dirichlet分布上积分得到，$\psi$是digamma函数，$B$是Beta函数。式(8)能够约束每个样本的正确类别能够生成更多证据，但我们还期望其它类别的证据越少越好，直至缩小为0，因此引入了KL散度[12]

$$\begin{aligned} D_{\mathrm{KL}}(\tilde{\alpha}_i) = \ln &\left(\frac{\Gamma\left(\sum_{k=1}^{K}\tilde{\alpha}_{ik}\right)}{\Gamma(K)\prod_{k=1}^{K}\Gamma(\tilde{\alpha}_{ik})} \right) + \sum_{k=1}^{K}(\tilde{\alpha}_{ik}-1) \\ &\cdot \left[\psi(\tilde{\alpha}_{ik}) - \psi\left(\sum_{k=1}^{K}\tilde{\alpha}_{ik}\right)\right] \end{aligned} \tag{9}$$

其中，$\tilde{\alpha}_i=y_i+(1-y_i)\odot\alpha_i$，该损失函数Dirichlet分布在真实标签处不做惩罚，在非正确标签对应的参数趋于1，证据趋向0。因此，对于样本i的第v个视图，损失函数为：

$$L(\alpha_i^{\mathrm{v}}) = L_{\mathrm{ace}}(\alpha_i^{\mathrm{v}}) + \lambda D_{\mathrm{KL}}(\tilde{\alpha}_i^{\mathrm{v}}) \tag{10}$$

其中，$\lambda>0$，是平衡两项损失函数系数。该损失函数的目的是最小化非正确标签所对应的分布参数，使该类别收集的证据越少越好。在训练网络时可以采取逐渐增大λ的策略，以防止网络在训练初期过度关注KL散度从而输出近似的均匀分布。

4 实验

4.1 数据

本文使用的数据集MIntRec是由Zhang等人[1]构建的。数据集的原始数据来自于美剧《Superstore》，它包含丰富的角色和场景，这有利于收集不同的社交对话中的用户意图类别和丰富的多模态信息。MIntRec是第一个用于真实世界场景下多模态意图识别的基准数据集，共有2 224个样本，每个样本都包含文本、音频和视频模态的数据。文献[1]在构建MIntRec数据集时，首先从视频网站上获取包含字幕的原始视频，然后提取每个发言者的开始和结束的时间戳，分割原始视频获得视频片段。然后，从视频片断中获取到文本字幕和每一帧的图片，用moviepy工具包提取其中的音频片段。数据集包含“表达情感或态度”和“达成目的”两大意图类别。其中，“表达情感或态度”细分为11类意图。“达成目的”则分为9类。本文按照3:1:1的比例来划分训练集、验证集和测试集。

4.2 实验设计

本文使用准确率(Acc)、宏观平均召回率(m-R)、宏观平均F1值(m-F1)和宏观平均精确率(m-P)作为模型的性能评估指标，这些指标取值越高则表明模型性能越好。本实验是基于Pytorch深度学习框架搭建的，实验中的主要可调参数设置如表1所示。本文采用的BERT版本为BERT-base-uncased。

4.3 实验结果分析

为了验证本模型的分类效果，选取了以下9种基线模型在数据集MIntRec上进行了对比。其中，

方法(1)～(4)为单模态意图识别方法； 有效的多模态表示与融合方法(Effective Multimodal Representation and Fusion Method, EMRFM)和模态感知提示模型(Modality-Aware Prompting, MAP)是多模态意图识别方法。此外，本文还对比了常用于情感分析的多模态融合方法，如MISA和多模式适应门BERT (Multimodal Adaptation Gate BERT, MAG-BERT)。对比方法如下：

(1)ConvLSTM[8]：在LSTM的循环单元中引入了卷积运算，使其可以处理具有空间维度的数据。

(2)多视图深度学习模型(Multiview Deep Learning, MDL)[15]：考虑单词间局部和长期依赖关系，利用CNN, LSTM和平均池化获得多视图特征表示。

(3)Wav2vec+Transformer：将Wav2vec提取的音频特征输入到Transformer中，然后将输出结果用于分类。

(4)Faster-RCNN+Transformer：将Faster-RCNN提取的视频特征输入到Transformer，将输出结果用于分类。

(5)MISA[10]：将每一个模态的信息映射到模态间共享子空间和模态内部私有子空间，并且通过重构损失、相似度损失、差异损失和分类任务损失对模型进行训练，最后进行分类。

(6)MULT[9]: 通过双向跨模态注意力机制扩展了普通的Transformer，捕获了不同模态之间的相互作用。

(7)MAG-BERT[16]：通过一个多模态自适应门(MAG)将音频和视频模态集成到BERT中，MAG可以在语义空间中产生与音频和视频信息相适应的位置偏移，嵌入到BERT层之间，以接受来自音频和视频的输入。

(8)EMRFM[11]：一种基于注意力门控神经网络的多模态融合模型，能够区分不同模态的贡献并降低噪音。

(9)MAP[17]：根据模态间的相似性来对齐不同模态，并利用跨模态注意力机制生成模态感知提示。

本文方法和对比方法实验结果如表2所示。对于单模态模型而言，ConvLSTM和MDL等基于文本的方法的意图识别效果要远高于基于音频和基于视频模态的方法。这表明文本在多模态意图识别任务中发挥了主要作用，文本中包含了比音频和视频更多的意图信息，是意图识别的重要信息来源。多模态模型的效果显著高于单模态模型，这表明通过对多模态数据进行特征提取并实现有效融合，丰富了文本、音频或视频等单模态信息，实现单模态信息之间的相互补充或支持，从而能够更准确地识别人类的意图。与其它多模态模型相比，本文提出的TMIR相较于MISA和MSZ等模型在4个评估指标上均有较大的提升，主要原因可能在于TMIR考虑每个视图对不同样本的可信度进行主观意见的动态融合，从而能够区分不同视图的贡献。整体来看，TMIR与其它表现最好的模型相比Acc提升了1.73个百分点，m-R提升了1.1个百分点，证明了TMIR能更好地捕捉多模态数据的内部重要信息和交互融合信息，并结合了多视图的动态可信融合，从而达到了良好的意图识别效果。

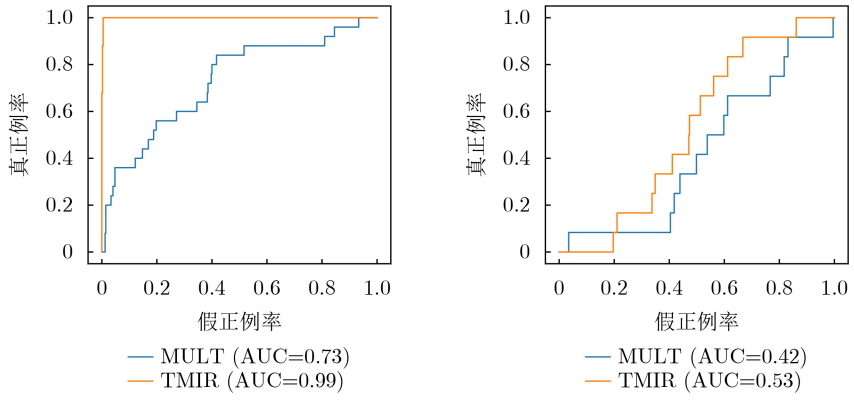
为了更清晰地评估本文方法的性能，本文先后选择了 “感谢”意图和“嘲讽”意图作为正类，将其他类别统一视为负类，进行了二分类问题的ROC曲线绘制，并与MULT模型进行了对比，如图2所示。从图2中可以看出，本文方法在两个类别上的AUC值均高于MULT模型。对于“感谢”意图，两个模型的AUC值均表现优异，其中本文提出的TMIR模型的AUC值高达0.99，这充分证明了其在检测“感谢”意图方面的有效性。而对于“嘲讽”意图，虽然两个模型的AUC值都较低，但本文方法的AUC值仍明显优于MULT模型，后者的

表 1　可调参数设置

参数名称	参数值
向量维度(文本、视频、音频)	768, 256, 768
序列最大长度(文本、视频、音频)	48, 480, 230
学习率	10^{-5}
Batch-size	20
CNN卷积核大小(文本、视频、音频)	5, 10, 10
CNN卷积核通道数(文本、视频、音频)	120, 120, 120
Transformer层数	8
BiLSTM隐藏层大小(文本、视频、音频)	60,60,60

表 2　对比方法实验结果(%)

模态	模型	Acc	m-P	m-R	m-F1
单模态	ConvLSTM	70.11	57.90	60.09	58.38
	MDL	70.79	70.38	66.76	66.89
	Wav2vec+Transformer	24.04	7.31	12.25	8.72
	Faster R-CNN+Transformer	15.28	7.86	8.12	5.77
多模态	MISA	71.91	70.46	67.82	68.07
	MULT	72.52	70.25	69.24	69.25
	MAG-BERT	72.65	69.08	69.28	68.64
	EMRFM	72.58	71.90	70.45	70.46
	MAP	72.13	72.50	68.80	**71.80**
	TMIR	**74.38**	**72.83**	**71.55**	71.36



(a) “感谢”意图的ROC曲线 (b) “嘲讽”意图的ROC曲线

图 2 TMIR与MULT模型的ROC曲线

AUC值仅为0.42，这反映出MULT模型在检测“嘲讽”意图方面的不足。

4.4 消融实验

为了进一步探讨本文构建的可信多模态意图识别模型TMIR中各个组成部分对分类效果的影响，对模型进行了消融实验，表3为实验结果，消融实验说明如下：

(1) w/o多视图特征提取层：在完整的模型上去掉跨模态交互模块和模态特定编码模块，将多模态特征表示层得到特征向量直接输入到多视图可信融合模块。

(2) 在完整的模型中依次去掉跨模态交互模块中的每个跨模态Transformer。

(3) w/o多视图可信融合模块：在完整模型上去掉多视图可信融合模块，直接将4个视图的特征向量拼接后经过全连接层和Softmax激活函数获得预测分类结果，并使用传统交叉熵损失进行模型训练。

实验(1)去掉跨模态交互模块和模态特定编码模块后，Acc下降3.82个百分点，说明捕捉模态内部特有信息及多模态交互信息对意图识别效果至关重要。实验(2)依次去掉跨模态Transformer后，意图识别效果均下降，尤其是去掉文本-音频Transformer时，4个指标大幅下降，表明以文本为辅助模态、音频为目标模态的交互关系对结果影响显著。实验(3)去掉多视图可信融合模块后，准确率和精确率分别下降1.12和1.67个百分点，原因在于固定权重的特征拼接无法动态调整视图权重，导致模型无法有效处理噪声或不一致信息。多视图可信融合模块通过动态分配权重，提升了模型泛化能力。完整TMIR模型在消融实验中表现最优，证明其合理性与优越性，不仅能提高意图识别效果，还能提供预测结果的可信度估计。

4.5 模型复杂度对比

本文采用浮点计算量(FLOPs)和参数量(Params)来评估方法的复杂度与计算资源消耗。将Batch_size设定为4，计算本文提出的方法TMIR以及MISA和MULT等对比模型的FLOPs和Params。表4呈现了实验结果。从结果中可以观察到，本文的TMIR模型与其他模型的复杂度相差不大。在资源消耗方面，与多模态语言理解中常用的MISA模型相比，Params指标性能有所降低。综合来看，本文提出的TMIR模型在复杂度和资源消耗与其他方法相当的前提下，展现出良好的用户意图识别性能。

5 结束语

本文提出一种可信的多模态意图识别方法，捕

表 3 消融实验对比结果(%)

模型	Acc	m-P	m-R	m-F1
w/o 多视图特征提取层	70.56	68.70	66.78	66.94
w/o音频-文本Transformer	72.58	72.23	70.26	70.15
w/o 视频-文本Transformer	72.13	70.98	70.85	70.04
w/o 文本-音频Transformer	71.91	66.87	69.54	67.32
w/o 视频-音频Transformer	72.13	71.85	70.91	70.75
w/o 音频-视频Transformer	73.26	67.76	70.60	68.52
w/o 文本-视频Transformer	72.58	70.99	71.18	70.45
w/o 多视图可信融合模块	73.26	71.16	71.14	70.37
TMIR	**74.38**	**72.83**	**71.55**	**71.36**

表 4 浮点计算量与参数量对比结果

模型	FLOPs(G)	Params(M)
MISA	15.58	115.93
MULT	16.60	105.03
DEAN	14.22	89.88
MAG-BERT	14.04	88.43
TMIR	17.14	108.24

捉多模态特征的互补性，并保留模态内部的特有信息。在多视图可信融合模块中，定义包含了预测的置信度分布、不确定性和类别偏好的主观意见，最后融合多个意见获得意图识别结果。实验表明，本方法不仅提升了意图识别的准确率，还能对预测结果的可信度进行度量。但是本文的模型在部分意图类别上的表现仍有待进一步提高。当意图类别的分布不平衡，模型可能会偏向于多数类，从而导致对少数类的识别性能下降。因此未来将考虑采用数据增强技术来扩充少数类样本数量，或者在模型训练中引入加权损失函数，为少数类样本分配更高的权重，以平衡模型对不同类别的关注。此外，后续也会考虑在将模型应用在医疗或客户服务领域，以充分展示本模型的多领域适用性。

参 考 文 献

[1] ZHANG Hanlei, XU Hua, WANG Xin, *et al*. MIntRec: A new dataset for multimodal intent recognition[C]. The 30th ACM International Conference on Multimedia, Lisboa, Portugal, 2022: 1688–1697. doi: 10.1145/3503161.3547906.

[2] SINGH U, ABHISHEK K, and AZAD H K. A survey of cutting-edge multimodal sentiment analysis[J]. *ACM Computing Surveys*, 2024, 56(9): 227. doi: 10.1145/3652149.

[3] HAO Jiaqi, ZHAO Junfeng, and WANG Zhigang. Multi-modal sarcasm detection via graph convolutional network and dynamic network[C]. The 33rd ACM International Conference on Information and Knowledge Management, Boise, USA, 2024: 789–798. doi: 10.1145/3627673.3679703.

[4] KRUK J, LUBIN J, SIKKA K, *et al*. Integrating text and image: Determining multimodal document intent in Instagram posts[C]. The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 2019: 4622–4632. doi: 10.18653/v1/D19-1469.

[5] ZHANG Lu, SHEN Jialie, ZHANG Jian, *et al*. Multimodal marketing intent analysis for effective targeted advertising[J]. *IEEE Transactions on Multimedia*, 2022, 24: 1830–1843. doi: 10.1109/TMM.2021.3073267.

[6] MAHARANA A, TRAN Q, DERNONCOURT F, *et al*. Multimodal intent discovery from livestream videos[C]. Findings of the Association for Computational Linguistics: NAACL, Seattle, USA, 2022: 476–489. doi: 10.18653/v1/2022.findings-naacl.36.

[7] SINGH G V, FIRDAUS M, EKBAL A, *et al*. EmoInt-trans: A multimodal transformer for identifying emotions and intents in social conversations[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023, 31: 290–300. doi: 10.1109/TASLP.2022.3224287.

[8] 钱岳, 丁效, 刘挺, 等. 聊天机器人中用户出行消费意图识别方法[J]. 中国科学: 信息科学, 2017, 47(8): 997–1007. doi: 10.1360/N112016-00306.
QIAN Yue, DING Xiao, LIU Ting, *et al*. Identification method of user's travel consumption intention in chatting robot[J]. *Scientia Sinica Informationis*, 2017, 47(8): 997–1007. doi: 10.1360/N112016-00306.

[9] TSAI Y H H, BAI Shaojie, LIANG P P, *et al*. Multimodal transformer for unaligned multimodal language sequences[C]. The 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019: 6558–6569. doi: 10.18653/v1/P19-1656.

[10] HAZARIKA D, ZIMMERMANN R, and PORIA S. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis[C]. The 28th ACM International Conference on Multimedia, Seattle, USA, 2020: 1122–1131. doi: 10.1145/3394171.3413678.

[11] HUANG Xuejian, MA Tinghuai, JIA Li, *et al*. An effective multimodal representation and fusion method for multimodal intent recognition[J]. *Neurocomputing*, 2023, 548: 126373. doi: 10.1016/j.neucom.2023.126373.

[12] HAN Zongbo, ZHANG Changqing, FU Huazhu, *et al*. Trusted multi-view classification with dynamic evidential fusion[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(2): 2551–2566. doi: 10.1109/TPAMI.2022.3171983.

[13] BAEVSKI A, ZHOU H, MOHAMED A, *et al*. Wav2vec 2.0: A framework for self-supervised learning of speech representations[C]. Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2020: 1044. doi: 10.5555/3495724.3496768.

[14] LIU Wei, YUE Xiaodong, CHEN Yufei, *et al*. Trusted multi-view deep learning with opinion aggregation[C]. The Thirty-Sixth AAAI Conference on Artificial Intelligence, 2022: 7585–7593. doi: 10.1609/aaai.v36i7.20724.

[15] ZHANG Zhu, WEI Xuan, ZHENG Xiaolong, *et al*. Detecting product adoption intentions via multiview deep learning[J]. *INFORMS Journal on Computing*, 2022, 34(1): 541–556. doi: 10.1287/ijoc.2021.1083.

[16] RAHMAN W, HASAN M K, LEE S, *et al*. Integrating multimodal information in large pretrained transformers[C]. The 58th Annual Meeting of the Association for Computational Linguistics, 2020: 2359–2369. doi: 10.18653/v1/2020.acl-main.214.

[17] ZHOU Qianrui, XU Hua, LI Hao, *et al*. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition[C]. The Thirty-Eighth AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2024: 17114–17122. doi: 10.1609/aaai.v38i15.29656.

杨 颖：女，教授，研究方向为不确定性推理与复杂产品开发工程管理.
杨艳秋：女，硕士生，研究方向为多模态意图识别、不确定性推理.
余本功：男，教授，研究方向为信息系统工程与决策科学与技术.

责任编辑：余 蓉

Multimodal Intent Recognition Method with View Reliability

YANG Ying[①②] YANG Yanqiu[①②] YU Bengong[①③]

[①](*School of Management*, *Hefei University of Technology*, *Hefei* 230009, *China*)

[②](*Key Laboratory of Process Optimization & Intelligent Decision-making*, *Ministry of Education*, *Hefei University of Technology*, *Hefei* 230009, *China*)

[③](*Engineering Research Center for Intelligent Decision-Making & Information System Technologies*, *Ministry of Education*, *Hefei* 230009, *China*)

Abstract:

Objective With the rapid advancement of human-computer interaction technologies, accurately recognizing users' multimodal intentions in social chat dialogue systems has become essential. These systems must process both semantic and affective content to meet users' informational and emotional needs. However, current approaches face two major challenges: ineffective cross-modal interaction and difficulty handling uncertainty. First, the heterogeneity of multimodal data limits the ability to leverage intermodal complementarity. Second, noise affects the reliability of each modality differently, and traditional methods often fail to account for these dynamic variations, leading to suboptimal fusion performance. To address these limitations, this study proposes a Trusted Multimodal Intent Recognition (TMIR) method. TMIR adaptively fuses multimodal information by assessing the credibility of each modality, thereby enhancing intent recognition accuracy and model interpretability. This approach supports intelligent and personalized services in open-domain conversational systems.

Methods The TMIR method is developed to improve the accuracy and reliability of intent recognition in social chat dialogue systems. It consists of three core modules: a multimodal feature representation layer, a multi-view feature extraction layer, and a trusted fusion layer (Fig. 1). In the multimodal feature representation layer, BERT, Wav2Vec 2.0, and Faster R-CNN are used to extract features from text, audio, and video inputs, respectively. The multi-view feature extraction layer comprises a cross-modal interaction module and a modality-specific encoding module. The cross-modal interaction module applies cross-modal Transformers to generate cross-modal feature views, enabling the model to capture complementary information between modalities (e.g., text and audio). This enhances the expressiveness of the overall feature representation. The modality-specific encoding module employs Bi-LSTM to extract unimodal feature views, preserving the distinct characteristics of each modality. In the trusted fusion layer, features from each view are converted into evidence. Subjective opinions are formulated according to subjective logic theory and are fused dynamically using Dempster's combination rules. This process yields the final intent recognition result and provides a measure of credibility. To optimize model training, a combinatorial strategy based on Dirichlet distribution expectation is applied, which reduces uncertainty and enhances recognition reliability.

Results and Discussions The TMIR method is evaluated on the MIntRec dataset, achieving a 1.73% improvement in accuracy and a 1.1% increase in recall compared with the baseline (Table 2). Ablation studies confirm the contribution of each module: removing the cross-modal interaction and modality-specific encoding components results in a 3.82% drop in accuracy, highlighting their roles in capturing intermodal interactions and preserving unimodal features (Table 3). Excluding the multi-view trusted fusion module reduces accuracy by 1.12% and recall by 1.67%, demonstrating the effectiveness of credibility-based dynamic fusion in enhancing generalization (Table 3). Receiver Operating Characteristic (ROC) curve analysis (Fig. 2) shows that TMIR outperforms the MULT model in detecting both "thanks" and "taunt" intents, with higher Area Under the Curve (AUC) values. In terms of computational efficiency, TMIR maintains comparable FLOPs and parameter

counts to existing multimodal models (Table 4), indicating its feasibility for real-world deployment. These results demonstrate that TMIR effectively balances performance and efficiency, offering a promising approach for robust multimodal intent recognition.

Conclusions　This study proposes a TMIR method. By addressing the heterogeneity and uncertainty of multimodal data—specifically text, audio, and video—the method incorporates a cross-modal interaction module, a modality-specific encoding module, and a multi-view trusted fusion module. These components collectively enhance the accuracy and interpretability of intent recognition. Experimental results demonstrate that TMIR outperforms the baseline in both accuracy and recall, and exhibits strong generalization in handling multimodal inputs. Future work will address class imbalance and the dynamic identification of emerging intent categories. The method also holds potential for broader application in domains such as healthcare and customer service, supporting its multi-domain scalability.

Key words: Intent recognition; Multimodal fusion; Multi-view learning