基于知识蒸馏和多模态融合网络的会话情感识别

摘要

在自然语言处理和人机交互领域，对话中的情感情绪识别对于提升对话系统性能、更好地满足用户需求具有至关重要的意义。对话中的情感信息能够通过语音、视觉和文本等多样化的模态进行呈现和识别。然而，在多模态情感识别的过程中，由于非语言模态（如语音和视觉）在情感识别方面所做出的贡献相对较为薄弱，这使得多模态情感识别成为一项充满挑战的复杂任务。为了解决这一难题，本文提出了一种名为“教师引导多模态融合网络（TelME 框架）”的方法。在该框架中，充分利用了跨模态知识蒸馏技术，将具有强大语言理解能力的语言模型作为教师网络，通过知识蒸馏的方式向非语言模态（即学生网络）传递关键信息。这种从教师网络到学生网络的知识传递过程，能够显著增强非语言模态在情感识别中的能力，弥补其在情感识别上的不足。此外，在多模态特征的整合环节，TelME 框架采用了一种动态融合策略。在这一过程中，学生网络不仅仅是知识的接收者，还能够反过来进一步支持教师模型的功能，促进多模态信息的有效融合与协同。通过这种方式，TelME 框架实现了多模态信息的高效利用，充分发挥了各个模态在情感识别中的优势。实验结果表明，TelME 框架在两个数据集上均取得了令人较为满意的效果。在 MELD 数据集中，模型在不同情绪类别上展现出了不同程度的预测准确性，尤其在某些情绪类别上表现突出。在 IEMOCAP 数据集上，同样能够对各类情绪进行较为有效的识别。

关键词：多模态；知识蒸馏；对话情感分析；自然语言处理

1 引言

在社交互动中，人们往往通过对话表达和传递自己的情感与情绪。对自然语言处理和人机交互领域的研究者而言，理解和分析对话中的情感是一项重要且复杂的挑战。对话情感分析（Emotion Recognition in Conversation，简称 ERC）作为一个专门领域，致力于研究对话中的情感表达，帮助计算机理解人类语言互动中的情感变化及其规律。

在实际应用中，对话情感分析在客户服务、在线教育和心理健康等领域展现了广泛的价值。在客户服务中，情感分析能够帮助企业深入了解客户的需求与情绪，从而更精准地提供服务，提升客户满意度。例如，当客户表现出困惑、不满或负面情绪时，情感分析技术可以帮助客服人员迅速捕捉这些情绪并采取有效的应对措施。在在线教育领域，通过情感分析，教师可以实时感知学生的情感状态，了解他们在学习中的情绪变化，从而优化教学策略，提高教育效果。心理健康领域同样受益于情感分析技术，它能够通过分析患者的情感表达，帮助心理医生更好地了解患者的情感状态和心理变化，为诊断和治疗提供重要依据。

对话情感分析的核心目标是通过多模态信息的解析，如文本内容、语音特征等，判断对话中每个参与者的情感状态，并揭示情感变化的趋势与规律。这项研究的意义不仅在于提升

人际沟通的质量，还能为多种场景提供实用价值。通过情感分析，人们能够更好地理解彼此的情感与意图，避免因情绪误解引发的不必要冲突。对于企业来说，情感分析是提升客户服务质量的关键工具，通过对客户需求的精准识别来构建更高效的服务体系。与此同时，情感分析在心理健康领域也扮演了重要角色，能够协助心理医生监测患者的情绪波动，提供更具针对性的治疗方案。

此外，对话情感分析在人工智能领域的应用也备受关注。在人机交互场景中，情感分析技术能够帮助智能机器人理解用户的情感需求，从而实现更自然、更具人性化的互动体验。例如，当用户表现出困惑或挫败感时，智能机器人能够通过情感分析给予适当的反馈，从而增强交互的效果和满意度。

综上所述，对话情感分析在多个领域都展现了重要的实际价值和应用潜力。如何更加全面地挖掘对话中隐含的情感信息，提高情感分析模型的性能，使预测结果更贴近实际情感状态，是当前人工智能发展中的重要课题。这一研究不仅推动了人机交互的智能化进程，也为商业、教育和医疗等多个行业带来了新的机遇和变革。

2 相关工作

2.1 情感分析方法国内外研究现状

在自然语言处理领域中，情感分析是一个备受关注和研究的重要方向，其旨在从文本中提取情感和情感的极性，在近年来国内外都受到广泛关注。以下将从情感词典、机器学习 [1] 和深度学习 [2] 三个方面来分别介绍情感分析的国内外研究现状。

（1）基于情感词典的情感分析方法

情感词典是一种包含大量词汇及其对应情感倾向的工具，同时涵盖程度副词、否定词和同义词等多种语言元素。基于情感词典的情感分析方法主要依赖将输入文本中的词语与情感词典中的词汇进行匹配，通过这些匹配的词汇所反映的情感强度，推断整篇文章的情感走向。这种方法的一大优势在于无需对文本进行人工标注，只需利用现成的情感词汇和规则即可实现情感的自动化分析。

目前，全球范围内已经存在大量广泛使用的情感词典。例如，在美国，常用的英文情感词典包括 SentiWordNet 和 General Inquirer；而在中国大陆，较为知名的中文情感词典有知网（HowNet）情感词典、大连理工大学情感词典、清华大学的褒贬义词典以及台湾大学 NTUSD 简体中文情感词典等。这些词典为情感分析提供了高质量的数据资源。然而，人工创建并持续更新这些情感词典是一项耗时耗力的工作，尤其是在面对社交媒体中大量新颖表达形式时，这种挑战更加显著。国外对于情感词典的研究起步较早。例如，Riloff 等人提出了通过语料库构建特定语言意义词典的方法 [3]；Taj 及其团队在新闻领域测试了一种依赖情感词典的情绪识别技术，并证明其准确性和实用性 [4]。此外，Islam 等人通过建立特定领域的情感词典，提高了情感分析的精准度。在国内，刘慧慧等人利用支持向量机（SVM）技术实现网页文本情感识别，显著提升了分析效率与精度 [5]；李永帅、王黎明等研究人员提出使用双向长短时记忆神经网络（BiLSTM）构建新的情感词典，取得了良好效果 [6]。汪韬等研究团队针对汉语中的多种表达形式制定了改良方案，进一步优化了词典构建质量 [7]。

情感词典的创建经历了从手动标注到部分监督模式，再到基于神经网络技术的逐步发展。尽管技术的进步提高了词典的质量与完整性，但在跨领域情感分析中，如何使词典与文本主

题适配，仍是一个亟待解决的难题。

（2）基于机器学习的情感分析方法

基于机器学习的情感分析方法通过训练模型从文本中学习情感表达模式，以此实现情感分类。这种方法依赖大量标注数据，用于模型学习不同情感类别间的特征和模式。

国外研究者在该领域取得了许多突破。例如，Pang 等人提出利用机器学习算法对文本内容进行情感分类，通过概率统计、支持向量机（SVM）等技术自动提取情感特征点并应用于决策场景 [8]。他们的研究为情感分析领域奠定了基础。此外，Socher 等人利用新型人工智能模型捕获文本语义信息，提高了情感分类的准确性 [9]。

国内研究者也在情感分析领域取得了显著进展。例如，尹臣琼等人提出基于 SVM 的中文情感分类方法，并在大规模中文数据集上取得良好效果。何颖刚等人则采用深度学习模型，如卷积神经网络（CNN），实现了高效的中文情感分析 [10]。熊佳俣等团队提出结合文本和图像信息的多模态深度学习模型，在情感分析任务中表现优异，为多领域情感分析提供了新的思路。

随着深度学习和自然语言处理技术的持续发展，基于机器学习的情感分析方法将进一步优化，为实际应用提供更加可靠的情感识别服务。

（3）基于深度学习的情感分析方法

深度学习作为近年来自然语言处理领域的重要突破，其在情感分析中的应用为提升情感识别的精准度和稳健性提供了技术支持。

在国外，研究者们提出了一系列基于深度学习的情感分析模型。例如，Tai 等人提出基于树形结构的长短时记忆网络（Tree-LSTM），通过捕获文本构造信息，提升情感分析能力 [11]。Kim 等人利用卷积神经网络（CNN）进行情感分类，通过卷积操作提取文本局部特征并加以整合，在情感数据集上取得了优异表现 [12]。此外，Vaswani 等人提出的 Transformer 模型引入自注意力机制，能够捕捉文本中长距离依赖关系，大幅提高情感分类的性能 [13]。

国内的研究者同样在这一领域表现突出。史远航等人提出基于 CNN 的中文情感分类模型，并成功应用于中文文本分析任务 [14]。刘飞生等人通过结合长短时记忆网络（LSTM）和注意力机制，显著提升了文本情感分类的精度。李锦等研究团队引入预训练语言模型（如 BERT），结合注意力机制，在情感分析任务上取得了显著性能提升 [15]。

深度学习技术的不断进步，为情感分析提供了更加强大的工具和方法。未来，随着深度学习与自然语言处理的进一步结合，情感分析技术有望更加精确，为多领域的实际需求提供更高效的解决方案。

2.2 知识蒸馏方法国内外研究现状

在近年来，随着人工智能技术的迅速发展，大规模机器学习和深度学习模型在自然语言处理领域得到了广泛应用。作为一种预训练语言模型，BERT 采用了基于 Transformer 的架构，并通过海量文本数据集进行了训练。凭借强大的语义理解能力，BERT 在多种自然语言处理任务中表现出色。然而，由于其庞大的参数量和高计算需求，使得模型在实际部署和推理时面临巨大挑战，尤其是在计算资源受限的环境中。为了解决这一问题，研究人员提出了知识蒸馏技术。

知识蒸馏 [16] 是一种用于模型压缩的技术，其核心目标是将大型复杂模型（教师模型）中的知识迁移到小型模型（学生模型）中，从而在保持性能的同时显著降低模型的参数量和

计算成本。这项技术在自然语言处理领域被广泛应用于文本分类、情感分析、机器翻译等任务，通过轻量化模型的高效推理来满足实际需求。例如，石佳来和郭卫斌 [17] 等研究人员针对 BERT 提出了多种知识蒸馏策略，包括基于样本权重的蒸馏和基于中间层输出的蒸馏，这些方法在减小模型体积的同时提高了推理效率。此外，黄玉娇、詹李超和范兴刚 [18] 等研究者通过知识蒸馏将 ELECTRA 模型的知识转移到更小的 BiLSTM 模型上，实现了资源效率与性能的平衡。

知识蒸馏的概念最初由 Hinton [19] 提出，但其雏形早在 Bucilă 等人的研究中已被探讨。他们的研究设想是先利用大型集成模型在原始数据集上完成训练，然后通过让学生模型学习教师模型重新标注的数据，从而实现小型模型对大型模型的模仿学习。这一方法显著提高了学生模型的性能，并首次证明了模型间知识传递的可行性。随后，Ba 和 Caruana [20] 进一步改进了这一方法，提出通过模拟教师模型的输出分布和 logits 值来优化学生模型的学习效果。他们发现，拟合 logits 输出的效果尤为显著，因为 Softmax 层的归一化将复杂的特征空间映射为类别概率分布，从而简化了学生模型的学习过程。

Hinton 的研究在此基础上提出了进一步的改进。他认为学生模型不仅需要模拟教师模型的输出概率分布，还应结合原始数据集进行学习。同时，他引入了 Temperature 超参数，使得教师模型输出的概率分布更具灵活性。当 Temperature 值较高时，模拟教师模型输出与拟合 logits 输出效果趋于一致，从而为不同任务的需求提供了更多调整空间。这一改进显著提升了知识蒸馏的适应性和实用性。

在具体应用方面，知识蒸馏技术已经在多种自然语言处理任务中取得了显著成果。例如，研究人员通过知识蒸馏将大型预训练语言模型的知识迁移至轻量级模型，实现了在资源受限场景中的高效推理。此外，许多研究还集中在改进知识蒸馏的技术方法，例如优化蒸馏损失函数、增强中间层特征传递以及结合多任务学习策略等，以进一步提升模型性能。

综上所述，知识蒸馏作为一种有效的模型压缩技术，为解决深度学习模型在实际应用中的计算资源需求提供了重要支持。国内外研究者在这一领域的探索推动了自然语言处理模型的轻量化和高效性。未来，随着对深度学习模型性能与效率需求的不断增加，知识蒸馏技术将在更多应用场景中发挥重要作用，为智能系统的可持续发展提供更多解决方案。

3 本文方法

3.1 问题陈述

给定一组对话参与者 S，话语 U 和情感标签 Y，由 k 个话语组成的对话表示为 $[(s_i, u_1, y_1), (s_j, u_2, y_2), ..., (s_i, u_k, y_k)]$，其中 $s_i, s_j \in S$ 为会话参与者。如果 $i = j$，那么 s_i 和 s_j 指的是同一说话者。

$y_k \in Y$ 是对话中第 k 个话语的情感，属于预定义的情感类别之一。另外，$u_k \in U$ 是第 k 个话语。U_k 以视频剪辑、演讲片段和文本文本的形式提供。即 $U_k = t_k, a_k, v_k$，其中 t, a, v 分别表示文本文本、语音片段和视频片段。ERC 的目标是预测 y_k，即对话中第 k 个话语所对应的情绪。

3.2 TelME 框架

3.2.1 模型概述

本文设计了一种教师主导的 ERC 多模态融合网络（TelME），如图 1 所示。该框架的设计基于以下假设：通过有效利用不同模态在情感识别任务中所起的不同层次的作用，可以提升 ERC 系统的整体性能。因此，我们引入了一种具有针对性的策略，强调优势模态的作用，同时补充和增强弱势模态的表现。

具体而言，我们首先通过文本模态中的情境建模提取强大的情感表征，同时从非语言模态中捕获当前说话者的音频和视觉特征。然而，由于音频和视觉模态在情感识别中的能力较为有限，加之模态间存在显著的异质性，这使得有效的多模态交互难以保证。为了解决这一问题，我们采用了一种知识蒸馏策略，将教师模型中与情感相关的重要知识提炼并传递给非语言学生模型。这一过程不仅缓解了模态间的异质性，还最大限度地提升了非语言模态的有效性。

此外，我们设计了一种融合方法，其中教师模型编码器所提取的强烈情感特征，通过参考学生模型的反向强化表征而得以优化和转移。这种融合机制确保了模态间的信息互补与协调，进一步增强了整体系统的情感识别能力。

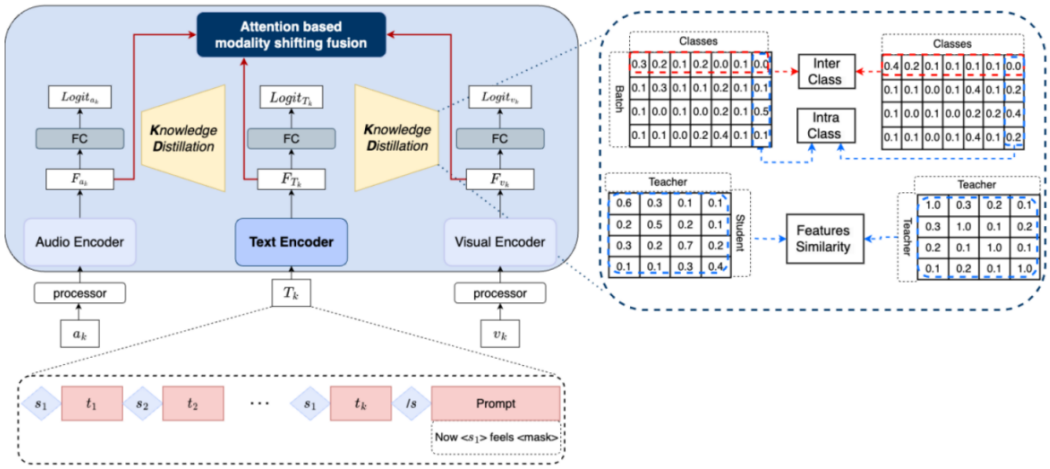

图 1. TelME 的整体框架

接下来将详细探讨 TelME 的三个核心组成部分：特征提取、知识蒸馏和基于注意力机制的模态转换融合，以展示该框架在情感对话识别任务中的有效性和创新性。

3.2.2 特征提取

图 1 直观展示了每个模态编码器如何接收相应的输入信号并提取情感特征。在本节中，我们将详细解释生成不同模态输入信号对应情感特征的方法。在文本模态的研究中，我们根据先前的研究 [21,22] 对文本进行上下文建模，将从对话开始到第 k 次回合的所有话语视为上下文信息。为了处理说话人之间的依赖关系并明确区分不同的说话人，我们引入特殊令牌 <si> 来表示说话者身份。此外，为了突出当前说话者的情感状态，我们构建了一个提示语句：“现在 <si> 感觉 <mask>”，以强调最近发言者的情感信息。情感特征来源于特殊标记 <mask>

的嵌入表示。在文本编码器的选择上，我们采用了改进版的 RoBERTa，情感特征的提取过程如下所示：

$$C_k = [< s_i >, t_1, < s_j >, t_2, ..., < s_i >, t_k] \tag{1}$$

$$P_k = Now < s_i > feels < mask > \tag{2}$$

$$F_{T_k} = TextEncoder(C_k < /s > P_k) \tag{3}$$

其中，$< s_i >$ 为表示当前说话者的特殊令牌，$< /s >$ 是 ROBERTa 的分隔符令牌。$F_{T_k} \in R^{1\times d}$ 为掩码令牌的嵌入表示，d 是编码器的维度。

在音频模态这方面，近年来，Transformer 架构的自监督学习在自然语言处理、音频和视频领域都取得了显著的成果。基于这一趋势，我们采用 data2vec 初始化音频编码器的状态。为了聚焦当前说话者的声音信息，我们仅使用第 k 次回合的音频片段，记为 a_k。音频片段经过预训练的处理器进行预处理，随后音频编码器从处理后的输入中提取情感特征，过程如下：

$$F_{a_k} = AudioEncoder(a_k) \tag{4}$$

其中，$F_{a_k} \in R^{1\times d}$ 为音频片段 a_k 的嵌入表示，d 是编码器的维度。

与音频模态类似，我们使用 TimeSformer 配置视觉编码器的初始状态。为了捕捉当前说话者的面部表情特征，我们仅选取第 k 次回合的视频片段，记为 v_k。通过图像处理技术，我们从视频中提取与第 k 次话语对应的帧，并构建视觉输入 v_k。随后，视觉编码器从处理后的输入中提取情感特征，具体过程如下：

$$F_{v_k} = VisualEncoder(v_k) \tag{5}$$

其中，$F_{v_k} \in R^{1\times d}$ 为视频片段 v_k 的嵌入表示，d 是编码器的维度。

3.2.3 知识蒸馏

解决模式之间异质性的挑战和非语言模式的低情绪识别贡献在促进令人满意的多模式互动方面具有巨大的潜力。因此，我们从理解语言语境的语言模型中提取出强烈的情感相关知识，从而以相对较低的贡献增加了从其他两种模式中提取的情感特征。我们同时使用两种不同类型的知识蒸馏：响应和基于特征的蒸馏。学生的总体损失可以由分类损失、基于响应的蒸馏损失和基于特征的蒸馏损失组成，即:

$$L_{student} = L_{cls} + \alpha L_{response} + \beta L_{feature} \tag{6}$$

其中，α 和 β 是平衡损失的因子。

$L_{response}$ 利用 DIST 作为 ERC 的跨模态蒸馏，该技术最初用于图像网络。由于文本模态和其他两种模态之间的显著差异，有效的知识蒸馏可能具有挑战性。因此，与传统的 KD 方法不同，我们使用的 KD 方法（$L_{response}$）利用 Pearson 相关系数而不是 KL 散度，如下所示。

$$d(\mu, v) = 1 - \rho(\mu, v) \tag{7}$$

其中，$\rho(\mu, v)$ 是两个概率向量 μ 和 v 之间的 Pearson 相关系数。

具体来说，$L_{response}$ 旨在通过教师和学生预测之间的相关性来提炼教师的偏好（预测的相对排名），即使在教师和学生之间存在极端差异的情况下，这也可以有效地进行知识提炼。收集一批内所有实例的预测概率分布，并计算教师和学生之间的班级间和班级内关系的 Pearson 相关系数。将班级间和班级内关系转移给学生。基于反应的蒸馏公式可以描述如下：

$$Y_{i,:}^{t} = softmax(Z_{i,:}^{t}/\tau) \tag{8}$$

$$Y_{i,:}^{S} = softmax(Z_{i,:}^{S}/\tau) \tag{9}$$

$$L_{inter} = \frac{\tau^2}{B}\sum_{i=1}^{B} d(Y_{i,:}^{S}, Y_{i,:}^{t}) \tag{10}$$

$$L_{intra} = \frac{\tau^2}{C}\sum_{j=1}^{C} d(Y_{i,:}^{S}, Y_{i,:}^{t}) \tag{11}$$

$$L_{response} = L_{inter} + L_{intra} \tag{12}$$

给定训练批 B 和情感类别 C，$Z^s \in R^{B\times C}$ 为学生的预测矩阵，$Z^t \in R^{B\times C}$ 为教师的预测矩阵。$\tau > 0$ 是控制 logits 柔软度的温度参数。

然而，引入 $L_{feature}$ 作为额外的蒸馏损失，而不是仅仅依赖 $L_{response}$，以更好地利用教师网络中的嵌入式信息。$L_{feature}$ 旨在减轻教师和学生模型表示之间的异质性，能够从教师那里提取出比仅使用 $L_{response}$ 更丰富的知识。通过这种方式，在多模态融合阶段，学生可以忠实地支持老师。$L_{feature}$ 利用了批处理中教师和学生的归一化表示向量之间的相似性。通过在教师的表示矩阵与其转置矩阵之间执行点积来构建目标相似性矩阵。通过将 softmax 函数应用于该矩阵，得出目标概率分布如下：

$$P_i = \frac{exp(M_{i,j}/\tau)}{\sum_{l=1}^{B} exp(M_{i,l}/\tau)}, \forall i, j \in B \tag{13}$$

其中 B 是训练批，$M \in R^{B\times B}$ 是目标相似度矩阵。$\tau > 0$ 是控制分布平滑度的温度参数。P_i 是目标概率分布。

同样通过计算教师和学生表示的点积来计算他们之间的相似性矩阵。可以计算相似性概率分布如下：

$$Q_i = \frac{exp(M'_{i,j}/\tau)}{\sum_{l=1}^{B} exp(M'_{i,l}/\tau)}, \forall i, j \in B \tag{14}$$

其中 $MR^{B\times B}$ 是学生和教师的相似度矩阵。Q_i 是教师和学生的相似概率分布。利用这两个概率分布，计算 KL 散度作为基于特征的蒸馏的损失:

$$L_{feature} = \frac{1}{B}\sum_{i=1}^{B} KL(P_i||Q_i) \tag{15}$$

其中 KL 是 Kullback-Leibler 散度。

3.2.4 基于注意力机制的模态转换融合

增强后的学生网络所具备的情感特征，能够对教师模型的情感相关表征产生影响，进而提供那些可能无法单纯从文本当中所捕获到的信息。在整个过程中，为了能够充分地利用这些独特的特征，采用一种多模态融合的方法。在这种方法里，源自学生模型的特征向量会对来自教师的表示向量进行操纵，能有效地把非语言信息整合进表示向量之中。

为了更加突出非言语特征的重要性与独特性，将学生模型所生成的向量进行连接操作，并且在此基础上开展多头自我注意的处理。经过多头自我注意这一复杂过程所生成的非言语信息向量，会同教师编码器的情感特征共同进入到移位步骤，成为其重要的输入部分。在移位步骤里，通过对教师模型的向量和非言语信息的向量进行连接与变换等一系列操作，最终生成门控向量，这一向量在整个多模态融合与信息整合的体系中发挥着极为关键的作用，它能够精准地调控信息的流动与整合方向，进一步提升整个模型对于情感信息的理解与处理能力。

$$g_{AV}^{k} = R(W_1 \cdot < F_{T_k}, F_{attention}^{k} + b_1 > \tag{16}$$

其中<,>是向量级联运算，R(x) 是非线性激活函数，W_1 是线性变换的权重矩阵，b_1 是标量偏差。$F_attention$ 是非言语信息的情感表征载体。g_{AV} 是门控向量。门控向量根据教师模型的表示突出非言语向量中的相关信息。通过应用门控向量来定义位移向量，如下所示:

$$H_k = g_{AV}^{k} \cdot (W_2 \cdot F_{attention}^{k} + b_2 > \tag{17}$$

其中 W_2 是线性变换的权重矩阵，b_2 是标量偏差。H 是基于非言语信息的位移向量。利用教师的表示向量和位移向量之间的加权和来生成多模态向量。最后使用多模态向量预测情绪:

$$Z_k = F_{T_k} + \lambda \cdot H_k \tag{18}$$

$$\lambda = min(\frac{||F_k||_2}{||H_k||_2} \cdot \theta, 1) \tag{19}$$

其中 Z 是多峰向量。应用缩放因子 λ 来控制位移矢量的大小,θ 作为阈值超参数。$||F_k||_2$,$||H_k||_2$ 分别表示 F_k 和 H_k 向量的 $L2$ 范数。

4 复现细节

4.1 与已有开源代码对比

在本研究中，我们在 student.py 中的损失函数加入了新的损失项，以适应知识蒸馏框架。传统的损失函数通常只关注交叉熵损失，而在我们的方法中，我们引入了一个新的损失函数 Logit_Loss，用于度量学生模型（Student）输出的 logits 与教师模型（Teacher）输出的 logits 之间的差异。这一修改的核心目的是通过知识蒸馏机制加强学生模型的训练，特别是在音频和视觉模态的情感识别中。

具体来说,我们定义的 CE_Loss 损失函数不仅计算了标准的交叉熵损失,还通过 Logit_Loss 强化了学生模型与教师模型之间的知识传递。通过这种方式，教师模型的强大情感识别能力能够更好地传递给学生模型，尤其是在模态之间存在较大差异时。我还在训练过程中使用了

自适应梯度缩放（GradScaler）和梯度裁剪技术，确保模型训练的稳定性，尤其是在大规模数据集上训练时。此外，我们针对不同的学生模型（如音频和视觉模型）分别进行了训练，确保每个模态的特征能够充分利用并有效提升情感识别能力。

4.2 实验环境搭建

本实验使用数据集上的加权平均 F1 分数来评估所有实验，使用 Huggingface 的 Transformers 中预训练模型的初始权重。所有编码器的输出尺寸统一为 768。优化器为 AdamW 函数，初始学习率为 1e-5。所有实验均在单个 NVIDIA GeForce RTX 4060 上进行。

4.3 创新点

本文提出一个新的多模态情感识别框架——教师引导的多模态融合网络（TelME）。与传统的情感识别方法不同，TelME 框架考虑了文本与非语言模态（如音频和视觉模态）在情感识别中的不同贡献，旨在通过跨模态知识蒸馏技术提高弱模态的效果。在 TelME 中，教师模型是一个强大的语言模型，通过知识蒸馏将其识别到的情感信息传递给学生模型，这样可以增强非语言模态在情感识别中的表现。

其次，TelME 框架使用了跨模态知识蒸馏（Cross-Modal Distillation），这是一个重要创新。通过跨模态蒸馏，教师模型将其在情感识别中获得的丰富信息传递给非语言模态的学生模型（如音频和视觉模态）。这一创新可以显著增强这些弱模态的情感识别能力，尤其是在多模态情感分析任务中，传统方法往往依赖文本模态，而忽视了非语言模态的潜力。通过这一方法，TelME 能够有效弥补弱模态的不足，提高整体情感识别效果。

另外提出的“基于位移的融合方法”（shifting fusion approach）也是一个重要创新。该方法不仅优化了模态之间的信息传递，还使得学生网络能够在支持教师模型的过程中，促进多模态信息的融合。这一方法增强了模态间的协同作用，使得每种模态能够最大化地贡献其优势特征，从而提升情感识别的准确性。

5 实验结果分析

TelME 模型在 MELD 数据集上的测试效果呈现出多方面的特点。如图 2，在不同情绪类别中，中性（neutral）情绪的预测表现最为出色，其精确率达到 0.77506，意味着在所有被模型预测为中性的样本中，真正是中性的样本占比高达 77.506%；召回率为 0.83210，即在所有实际为中性的样本中，被模型正确预测为中性的样本占比为 83.210%；F1 - score 为 0.80215，综合反映了模型对中性情绪的高预测性能，且该情绪类别的样本数量多达 1256 个。

快乐（joy）情绪的表现也较好，精确率、召回率和 F1 - score 均为 0.65672，样本数量为 402 个。愤怒（anger）情绪的精确率为 0.60526，召回率为 0.53333，F1 - score 为 0.56703，样本数量有 345 个。惊讶（surprise）情绪的精确率为 0.55927，召回率为 0.65480，F1 - score 为 0.60328，样本数量为 281 个。悲伤（sadness）情绪的精确率为 0.51656，召回率为 0.37500，F1 - score 为 0.43454，样本数量为 208 个。

然而，模型在厌恶（disgust）和恐惧（fear）情绪上的表现较差。厌恶情绪的精确率为 0.36842，召回率为 0.20588，F1 - score 为 0.26415，样本数量仅 68 个；恐惧情绪的精确率为 0.30769，召回率为 0.24000，F1 - score 为 0.26966，样本数量为 50 个。从平均指标来看，模型的

整体准确率为 0.68199，表明在所有样本中被正确分类的样本占比为 68.199%。宏平均（macro avg）的精确率为 0.54128，召回率为 0.49956，F1 - score 为 0.51393；加权平均（weighted avg）的精确率为 0.67100，召回率为 0.68199，F1 - score 为 0.67375。加权平均指标高于宏平均指标。

TelME 在 MELD 数据集上的性能比之前最先进的方法（M2FNet）高出 0.66%。与 EmoCaps 相比，MELD 数据集上的性能提升了 3.37%。TelME 在除惊讶和愤怒之外的所有情绪类别上的表现均优于其他模型。然而，倘若将惊讶与恐惧、厌恶与愤怒视为相似情绪，EmoCaps 在推理过程中对惊讶和愤怒存在偏向性，其对恐惧和厌恶的 F1 得分分别仅为 3.03% 和 7.69%。而 TelME 能更好地区分这些相似情绪，使恐惧和厌恶的得分分别达到 26.97% 和 26.42%。我们推测，本框架能够更准确地预测少数类别的情绪，原因在于通过知识蒸馏（KD）策略强化的非言语情态信息（例如话语的强度和音高）能更好地辅助教师模型判断容易混淆的情绪。

	precision	recall	f1-score	support
anger	0.60526	0.53333	0.56703	345
disgust	0.36842	0.20588	0.26415	68
fear	0.30769	0.24000	0.26966	50
joy	0.65672	0.65672	0.65672	402
neutral	0.77506	0.83121	0.80215	1256
sadness	0.51656	0.37500	0.43454	208
surprise	0.55927	0.65480	0.60328	281
accuracy			0.68199	2610
macro avg	0.54128	0.49956	0.51393	2610
weighted avg	0.67100	0.68199	0.67375	2610

图 2. TELME 在 MELD 数据集上训练的结果

TelME 在 IEMOCAP 数据集上的训练结果呈现出多方面的特点。如图 3，在各个情绪类别中，愤怒（Anger）情绪的预测表现最为出色，其精确率达到 0.7738，这意味着在所有被模型预测为愤怒的样本中，真正属于愤怒情绪的样本占比为 77.38 %；召回率为 0.7839，即在所有实际为愤怒的样本中，能够被模型正确预测出来的占比为 78.39%；F1 - score 为 0.7788，综合反映了模型对愤怒情绪有着较高的预测性能，且该情绪类别的样本数量为 792 个。

中性（Neutral）情绪的表现也较好，精确率为 0.6742，召回率为 0.6849，F1 - score 为 0.6795，样本数量有 1487 个。沮丧（Frustrated）情绪的精确率是 0.6863，召回率为 0.6839，F1 - score 为 0.6851，样本数量为 914 个。兴奋（Excited）情绪的精确率为 0.6839，召回率为 0.6637，F1 - score 为 0.6737，样本数量为 689 个。

然而，模型在快乐（Happy）和悲伤（Sad）情绪上的表现存在一定的不平衡。快乐情绪的精确率为 0.4946，相对较低，但召回率高达 0.8348；悲伤情绪则是精确率较高，为 0.8348，但召回率较低，为 0.4946。这两种情绪的 F1 - score 相同，均为 0.6248，快乐情绪的样本数量为 1237 个，悲伤情绪的样本数量为 923 个。

从平均指标来看，模型的整体准确率为 0.7048，表明在所有样本中被正确分类的样本占比

为 70.48%。宏平均 (macro avg) 的精确率为 0.6915，召回率为 0.6694，F1 - score 为 0.6780；加权平均 (weighted avg) 的精确率为 0.6792，召回率为 0.6901，F1 - score 为 0.6835。加权平均指标相对较为均衡，宏平均指标则反映了模型在不同情绪类别上的综合表现。总体而言，TelME 在 IEMOCAP 数据集上有一定的有效性，但在某些情绪类别上还有进一步提升的空间。

emotion	precision	recall	f1 - score	support
Happy	0.4946	0.8348	0.6248	1237
Sad	0.8348	0.4946	0.6248	923
Neutral	0.6742	0.6849	0.6795	1487
Anger	0.7738	0.7839	0.7788	792
Excited	0.6839	0.6637	0.6737	689
Frustrated	0.6863	0.6839	0.6851	914
Accuracy			0.7048	6042
macro avg	0.6915	0.6694	0.6780	6042
weighted avg	0.6792	0.6901	0.6835	6042

图 3. TELME 在 IEMOCAP 数据集上训练的结果

6 总结与展望

本研究提出了一种基于知识蒸馏和多模态融合网络（TelME）的框架，旨在提升对话情感识别的性能。通过结合文本、音频和视觉信息，TelME 框架采用知识蒸馏技术，利用强大的语言模型（教师模型）引导非语言模态（学生模型）的学习，从而增强音频和视觉模态在情感识别中的贡献。实验结果表明，TelME 在 MELD 和 IEMOCAP 两个数据集上均取得了优异的成绩，特别是在少数类别情绪（如恐惧和厌恶）的识别上表现突出。通过知识蒸馏增强非语言模态的情感信息，使得 TelME 能够有效区分相似情绪，提升了整体模型的情感识别能力。

尽管 TelME 在多个数据集上展示了强大的性能，仍然存在进一步提升的空间。例如，模型在一些情绪类别的识别上存在不平衡问题，未来可以进一步优化多模态特征融合的机制，并改进少数类别情绪的识别效果。未来的研究可以探索更为精细的模态间交互方式，或采用自监督学习等技术，进一步提升模型对复杂情感的理解能力。此外，TelME 框架的跨领域适应性也值得关注，尤其是在不同任务和数据集间的迁移能力，这将有助于提高其在实际应用中的普适性。

随着人工智能和情感计算技术的不断发展，TelME 框架有望在多个领域中得到广泛应用，例如在线客服、心理健康以及教育领域。未来可以进一步研究如何使模型更具实时性，并在计算资源有限的情况下保持高效性。通过模型压缩和优化技术，如剪枝和量化，可以进一步提高 TelME 的推理效率，为实际应用场景提供更强的支持。总的来说，TelME 为多模态情感识别提供了一种有效的新思路，未来有望在智能交互和情感计算的领域中发挥更加重要的作用。

参考文献

[1] 张倩男. 基于情感分析和机器学习的用户评论信息挖掘. 科技和产业, 23(23):121–127, 2023.

[2] 刘培玉, 卢强, 张殿元, and 朱振方. 基于深度学习的方面级情感分析方法研究进展. 山东师范大学学报：自然科学版, (037-001), 2022.

[3] Ellen Riloff, Janyce Wiebe, and William Phillips. Exploiting subjectivity classification to improve information extraction. In *AAAI Conference on Artificial Intelligence*, 2005.

[4] William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. Inducing domain-specific sentiment lexicons from unlabeled corpora, 2016.

[5] 刘慧慧, 王爱银, and 刘禹彤. 基于 svm 的文本情感分析 – 以新冠疫情事件为例. 信息技术与信息化, (1):37–40, 2023.

[6] 李永帅. 基于双向 *LSTM* 的动态情感词典构建方法研究及文本情感分析. PhD thesis, 郑州大学.

[7] 汪韬 and 张再跃. 面向电影评论的情感词典构建方法研究. 计算机与数字工程, (004):050, 2022.

[8] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques, 2002.

[9] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. *Association for Computational Linguistics*, 2013.

[10] 何颖刚, 王宇, 夏丽丽, 郭静, and 郑新旺. 基于 fasttext 和多尺度深层金字塔卷积神经网络的中文文本情感分类模型. 宁德师范学院学报: 自然科学版, 34(4):382–388, 2022.

[11] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks, 2015.

[12] Yoon Kim. Convolutional neural networks for sentence classification, 2014.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[14] 史远航. 基于图卷积神经网络的方面级文本情感分析方法研究. PhD thesis, 内蒙古农业大学, 2023.

[15] 李锦, 夏鸿斌, and 刘渊. 基于 bert 的双特征融合注意力的方面情感分析模型. 计算机科学与探索, 18(1):205–216, 2024.

[16] 司兆峰 and 齐洪钢. 知识蒸馏方法研究与应用综述. 中国图象图形学报, 28(9):2817–2832, 9 2023.

[17] 石佳来 and 郭卫斌. 一种针对 bert 模型的多教师蒸馏方案. 华东理工大学学报（自然科学版）, 50(2):293–300, 4 2024.

[18] 黄玉娇, 詹李超, 范兴刚, 肖杰, and 龙海霞. 基于知识蒸馏模型 electra-base-bilstm 的文本分类. 计算机科学, 49(S02):6, 2022.

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[20] Lei Jimmy Ba and Rich Caruana. Do deep nets really need to be deep?, 2014.

[21] Joosung Lee and Wooin Lee. Compm: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation, 2022.

[22] Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. Supervised prototypical contrastive learning for emotion recognition in conversation, 2022.