基于多任务多模态交互学习的情感分类方法

薛鹏[1]，李旸[2]，王素格[1,3]，廖健[1]，郑建兴[1]，符玉杰[1]，李德玉[1,3]
1.山西大学计算机与信息技术学院，山西省 太原 030006
2.山西财经大学金融学院，山西省 太原 030006
3.山西大学计算智能与中文信息处理教育部重点实验室，山西省 太原 030006
wsg@sxu.edu.cn

摘要

随着社交媒体的快速发展，多模态数据呈爆炸性增长，如何从多模态数据中挖掘和理解情感信息，已经成为一个较为热门的研究方向。而现有的基于文本、视频和音频的多模态情感分析方法往往将不同模态的高级特征与低级特征进行融合，忽视了不同模态特征层次之间的差异。因此，本文采用以文本模态为中心，音频模态和视频模态为补充的方式，提出了多任务多模态交互学习的自监督动态融合模型。通过多层的结构，构建了单模态特征表示与两两模态特征的层次融合表示，使模型将不同层次的特征进行融合，并设计了从低级特征渐变到高级特征的融合策略。为了进一步加强多模态特征融合，使用了分布相似性损失函数和异质损失函数，用于学习模态的共性表征和特性表征。在此基础上，利用多任务学习，获得模态的一致性及差异性特征。通过在CMU-MOSI和CMU-MOSEI数据集上分别实验，实验结果表明本文模型的情感分类性能优于基线模型。

关键词： 多模态融合 ； 多任务学习 ； 情感分析

Sentiment classification method based on multitasking and multimodal interactive learning

Peng Xue[1], Yang Li[2], Suge Wang[1,3], Jian Liao[1], Jianxing Zheng[1], Yujie Fu[1], Deyu Li[1,3]
1.School of Computer and Information Technology, Shanxi University,Shanxi 030006
2.Shanxi University of Finance and Economics,Shanxi 030006
3.Key Laboratory Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University,Shanxi 030006
wsg@sxu.edu.cn

Abstract

With the rapid development of social media, multimodal data has shown explosive growth. How to mine and understand emotional information from multimodal data has become a popular research direction. However, existing multimodal sentiment analysis methods based on text, video, and audio often fuse high-level and low-level features of different modalities, ignoring the differences between different levels of modal features. Therefore, this article proposes a self supervised dynamic fusion model for multitasking and multimodal interactive learning, centered around text modality and supplemented by audio and video modalities. Through a multi-layer structure, a hierarchical fusion representation of single modal feature representation and pairwise modal features was constructed, enabling the model to fuse features from different levels. A fusion strategy was designed to gradually transition from low-level features to high-level features. In order to further strengthen multimodal feature fusion, distributed similarity Loss function and heterogeneous Loss function are used to learn common and specific

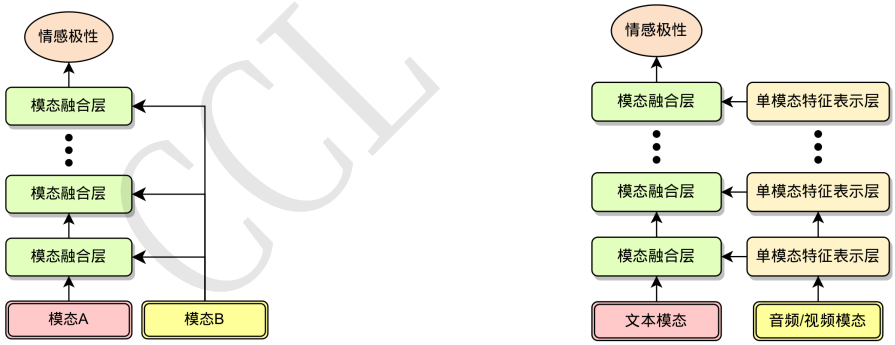
第二十二届中国计算语言学大会论文集，第315页-第327页，哈尔滨，中国，2023年8月3日至5日。

representations of modes. On this basis, multi task learning is utilized to obtain the consistency and difference features of modalities. Through separate experiments on the CMU-MOSI and CMU-MOSEI datasets, the experimental results show that the sentiment classification performance of our model is superior to the baseline model.

Keywords: Multimodal fusion , Multi-task learning , Sentiment analysis

1 引言

随着社交媒体的迅速发展，以及配备高质量摄像头的智能手机的普及，使得多模态数据（如电影、短视频等）呈爆炸式增长。多模态数据通常是由文本、视觉（视频）和声学（语音）组成。在对话系统和虚拟现实等应用领域中，如何让机器在不忽视非语言因素的情况下，能够理解多感官信息，成为保持高质量的用户交互的关键。多模态情感分析(multimodal sentiment analysis, MSA)旨在从音频、视觉和语言特征中预测情感的得分 (Soleymani et al., 2017)。例如，通过产品数据的情感得分，可以获得客户对整体产品反馈信息，通过对选民的评论数据的情感分析，可以预测潜在选民对投票的意图。对于同一时间段的不同模态数据，情感的表达通常是相互补充的，因此，对多模态数据的情感分析，可对语义和情感消歧提供技术支持。MSA的关键问题是如何将多模态数据进行融合，即对所有输入的模态数据，如何提取和整合它们的信息，用于深入理解数据的情感。目前，许多研究者已经提出了利用多模态信息进行情感分析的方法 (Zellinger et al., 2017; Liu et al., 2017; Ruder and Plank, 2018)。其中，跨模态注意力机制是使用较多的多模态融合方法，它可以通过建模不同模态之间的关系，从而强化其中的某一模态。然而，已有的研究工作较少关注将模态的高级特征的融合问题。如图1(a)所示，已有的工作通常是将一个模态的高级特征与另外一个模态的低级特征进行融合，体现了模态融合过程中的不一致性。通过这种方式的模态融合，无法获得不同模态之间的最佳映射，降低了模型的情感分析性能。因此，本文提出了多任务多模态交互学习的自监督动态融合模型(Self-supervised dynamic fusion model for multi-task and multi-modal interactive learning，MMILN)，如图1(b)所示，该模型可在模态的低级特征向高级特征转化时，从多个层次充分融合，同时使用以文本模态为中心，其余模态为辅助的方式，使模型对多个模态的情感信息进行相互关联，并有效表示。



(a)已有工作　　(b)本文工作

Figure 1: 已有工作与本文模态融合图示对比

本文的主要贡献如下:

（1）本文提出了一种多任务多模态交互学习的自监督动态融合方法，采用从低级特征渐变到高级特征的融合策略，建立了单模态特征表示与两两模态特征的层次融合表示。

（2）在模型建立时，构建分布相似性损失函数和异质损失函数，学习不同模态的共性表征以及特性表征，进一步加强多模态特征的有效融合。同时，利用多任务学习策略，获得模态的一致性及差异性特征。

项目基金： 国家自然科学基金项目(62106130, 62076158, 62072294, 62272286)；山西省基础研究计划(20210302124084)；山西省高等学校科技创新项目(2021L284)；CCF-智谱AI大模型基金（CCF-Zhipu202310）

（3）本文的模型在CMU-MOSI和CMU-MOSEI数据集上分别进行实验，实验结果表明本文模型的情感分类性能优于基线模型。

2 相关工作

对于多模态表示学习方面，主要思想在于如何减少单模态表示之间的距离差异，使得不同模态之间的差距尽可能缩小。(Zadeh et al., 2017)利用张量融合网络通过三倍笛卡尔积将单模态分解为张量，然后计算这些张量的外积作为融合结果。(Liu et al., 2018)利用低秩多模态融合将堆叠的高阶张量分解成许多低秩因子，然后基于这些因子进行高效融合。(Tsai et al., 2018)将一个推理网络和一个具有中间模态特定因素的生成网络连接起来，以促进融合过程的重建和判别损失。Yu (2021)等人利用自监督学习策略，设计了一个标签自动生成模块，并将其运用在多模态和单模态的训练任务上，以达到减小模态差异的目的。Han (2021)等人将互信息的概念引入多模态情感分析中，提出了一种最大化互信息学习框架，避免了与任务相关信息的丢失。Hazarika (2020b)等人将模态向量投影至两个不同的空间中，利用正则化组件进行共有模态特征和特有模态特征的表示学习。

在多模态情感分析领域中，更多的研究是针对多模态融合方面的。如何将来自不同模态的数据进行有效融合，是该领域面临的一个挑战性问题。由于Transformer和BERT拥有强大的特征提取能力，许多研究针对Transformer 中的自注意力模块进行改进，使得不同模态的向量可以动态交互，从而达到跨模态融合、互补学习的目的 (Tsai et al., 2019)。Rahman (2020)等人提出了多模态门控组件，使得BERT 模型在不改变结构的基础上能够动态地接受多模态信息。(Sun et al., 2022)引入了一种基于元学习的方法来学习更好的单模态表示，然后将其用于随后的多模态融合。(Sun et al., 2023)提出一个通用的、统一的框架EMT-DLFR来实现鲁棒MSA。

受以上工作的启发，本文使用自注意力机制分别对两个单模态（音频、视频）特征进行抽取，模型采用从低级渐变到高级的特征融合策略，将不同层次的模态特征进行融合，并在多模态特征融合过程中，引入权重系数，用于刻画各个模态的情感贡献度，进一步，利用相似性损失与异质损失学习不同模态间的共性表征与特性表征。

3 多任务多模态交互学习的自监督动态融合方法

3.1 模型描述

为了使不同模态的层次特征充分融合，本文提出了多任务多模态交互学习的自监督动态融合方法(MMILN)，如图2所示。该方法的核心思想是通过对两个单模态特征（音频和视频），按层次进行抽取，并将不同层次的特征与不同层次的文本模态特征进行融合，使得文本模态与另外两个单模态（音频和视频）交互学习，得到更加准确的多模态特征。模型由三个模块组成，分别是不同模态数据的层次融合模块、三种模态表示再融合模块和多任务学习模块。

在多模态情感分析任务中，我们将$X_t \in R^{l_t \times d_t}$，$X_a \in R^{l_a \times d_a}$，$X_v \in R^{l_v \times d_v}$分别作为模型的输入，其中，$l_t$、$l_a$和$l_v$分别是文本、音频和视频的序列长度，$\hat{y}_m \in R$作为情感的预测结果。在训练过程中，采用两个单模态（音频和视频）的模型，辅助多模态的表示学习，这两个单模态模型的输出分别为：$\hat{y}_a, \hat{y}_v$，它们均由自监督学习生成，用于辅助更新多模态融合的输出结果$\hat{y}_m$。

对于音频模态（X_a）和视频模态（X_v），本文使用COVAREP与Openface/Facet提取音频与视频浅层特征后，再分别使用LSTM（Long Short-Term Memory）捕获模态的时序特征，从而获得特征向量M_a和M_v。

$$M_a = LSTM(X_a; \theta_a) \in R^{d_a} \tag{1}$$

$$M_v = LSTM(X_v; \theta_v) \in R^{d_v} \tag{2}$$

对于文本模态，我们使用预训练BERT模型提取句子表示M_t。

$$M_t = BERT(X_t; \theta_t) \in R^{d_t} \tag{3}$$

第二十二届中国计算语言学大会论文集，第315页-第327页，哈尔滨，中国，2023年8月3日至5日。

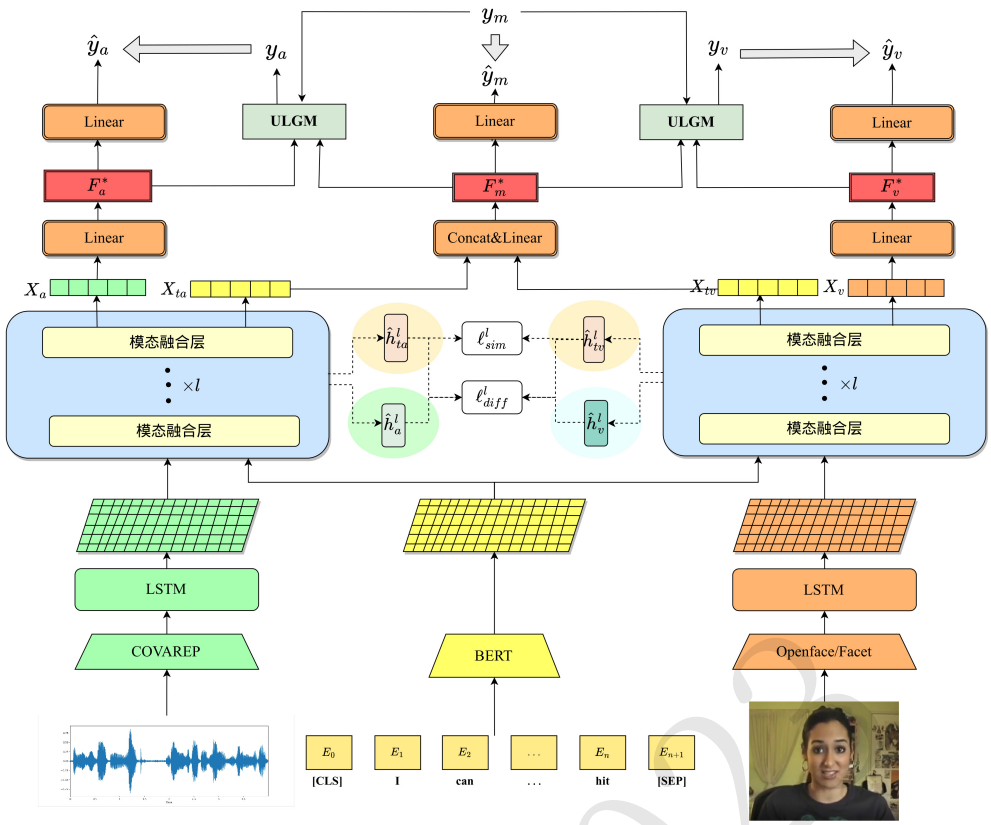


Figure 2: 模型整体结构图

利用公式(1)-(3)，获得了来自三个模态的特征表示:M_t、M_a和M_v，其中，θ_m表示模型的参数。之后我们使用卷积层统一三个模态的维度为R^{d_s} 。

$$X_m = Conv(M_m) \in R^{d_s} \tag{4}$$

其中，$m \in \{t, a, v\}$，其结果序列可以表示为：$X_m^0 = (X_{m,1}^0, X_{m,2}^0, \ldots, X_{m,k}^0)$作为模型的初始输入，k表示序列长度。

3.2 不同模态数据的层次融合模块

为了使不同模态的特征进行有效融合，我们使用多个模态融合层，并且构建了GateTransformer模型和SAG-Transformer模型。其中，前者用于单模态特征表示，建立分层网络结构对单模态特征进行从低级到高级的特征表示；后者用于两两模态特征的层次融合表示。本模块分为两个部分——单模态特征表示与两两模态特征的层次融合表示。模态融合层如图3所示。

3.2.1 单模态特征表示

本文采用以文本模态为中心，音频模态和视频模态为补充的表示方式。在本节中，我们对音频和视频模态分别进行特征表示，用于对文本模态的补充。音频（X_a^{i-1}）和视频（X_v^{i-1}）模态分别输入到各自的GateTransformer中，它们的结构完全一致，仅输入的模态数据不同，输出结果为单模态特征表示（X_n^i）与多头注意力值（$X_{mh\_n}^i$），$n \in \{a, v\}$。其中，单模态特征（X_n^i）作为下一个模态融合层的输入，多头注意力值（$X_{mh\_n}^i$）作为3.2.2节两两模态特征层次融合的输入。在得到前一个模态融合层的输出结果$X_n^{i-1} = (X_{n,1}^{i-1}, X_{n,2}^{i-1}, \ldots, X_{n,k}^{i-1})$之后，将音频模态和视频模态分别输入到GateTransformer中得到单模态特征和多头注意力结果，其公式如

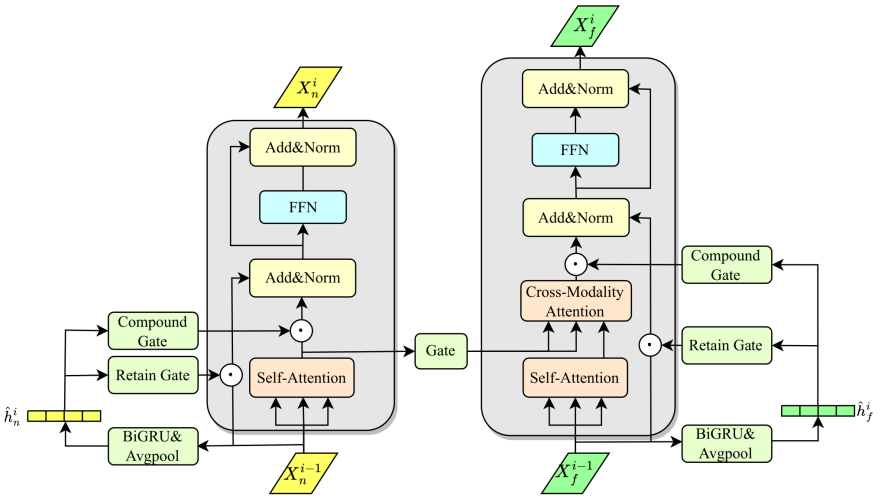


Figure 3: 模态融合层（由GateTransformer模型（左）和SAG-Transformer模型（右）组成）

下。

$$X_a^i, X_{mh\_a}^i = GateTransformer(X_a^{i-1}, X_a^{i-1}, X_a^{i-1}) \tag{5}$$
$$X_v^i, X_{mh\_v}^i = GateTransformer(X_v^{i-1}, X_v^{i-1}, X_v^{i-1}) \tag{6}$$

由于缺乏信息流控制，多头注意的操作表现出次优的性能。为了以细粒度和可控的方式改进它，我们使用GateTransformer，通过两个门——保留门g^r与复合门g^c对信息流控制。其中，保留门g^r决定了残差结构中模态信息的比例，以及复合门g^c决定了目标模态的前向传播比例。其公式如下。

$$\hat{h}_n^i = avgpool(h_n^i) = avgpool(BiGRU(X_n^{i-1}; \theta_n^i)) \tag{7}$$
$$g_r^i = \sigma(W_r^i \hat{h}_n^i) \tag{8}$$
$$g_c^i = \sigma(W_c^i \hat{h}_n^i) \tag{9}$$

其中，θ_n^i是BiGRU在第i层的参数，$W_* \in R^{d_s \times d_s}$，$* \in \{r, c\}$。

query $Q^i = W_Q^i X_n^{i-1}$，key $K^i = W_K^i X_n^{i-1}$，value $V^i = W_V^i X_n^{i-1}$用于多头注意力机制，并且使用门控机制限制单模态特征抽取的残差块与信息流。

$$X_{mh\_n}^i = MH - ATT(Q^i, K^i, V^i) \tag{10}$$
$$\widetilde{X}_n^i = LN(g_c^i \odot X_{mh\_n}^i + g_r^i \odot X_n^i) \tag{11}$$

其中，MH-ATT表示多头注意力、$\odot$表示按元素相乘、LN是层归一化。之后将注意力结果输出到前馈神经网络，并结合残差网络生成当前的模态融合层的最终结果。

$$X_n^i = LN(\widetilde{X}_n^i + FFN(\widetilde{X}_n^i)) \tag{12}$$

此模块生成的结果分别为X_a^i、X_v^i，表示单模态a（音频）、v（视频）在模态融合层第i层的输出，$X_{mh\_a}^i$，$X_{mh_v}^i$表示单模态a（音频）、v（视频）在第i层的多头注意力结果。

3.2.2 两两模态特征的层次融合表示

为了使模态特征从低级特征融合渐变到高级特征融合，我们采用分层的网络结构，将单模态特征从多个层次与文本模态进行融合。在本节中，我们利用3.2.1节单模态特征的多头注意力结果（$X_{mh\_n}^i$）与多模态特征（X_f^{i-1}）输入到SAG-Transformer，其中，$f \in \{ta, tv\}$。在得到

前一个模态融合层的输出结果$X_f^{i-1} = (x_{f,1}^{i-1}, x_{f,2}^{i-1}, \ldots, x_{f,k}^{i-1})$之后，我们将$X_f^{i-1}$与$X_{mh_n}^i$共同输入SAG-Transformer中。其公式分别如下所示。

$$X_{tv}^i = SAG - Transformer(X_{tv}^{i-1}, X_{mh\_v}^i, X_{mh\_v}^i) \tag{13}$$
$$X_{ta}^i = SAG - Transformer(X_{ta}^{i-1}, X_{mh\_a}^i, X_{mh\_a}^i) \tag{14}$$

值得注意的是:$X_{tv}^0 = X_t^0$，$X_{ta}^0 = X_t^0$。

SAG-Transformer是在GateTransformer的基础之上增加了交叉模态注意力机制，SAG-Transformer公式表示如下。

$$H_f^i = Self - Attention(X_f^{i-1}, X_f^{i-1}, X_f^{i-1}) \tag{15}$$
$$\hat{h}_f^i = avgpool(h_f^i) = avgpool(BiGRU(X_f^{i-1}; \theta_f^i)) \tag{16}$$
$$g_s^i = \sigma(W_s^i X_{mh\_n}^i) \tag{17}$$
$$\widetilde{X}_{mh\_n}^i = g_s^i \odot X_{mh\_n}^i \tag{18}$$

其中，θ_f^i是BiGRU在第i层的参数。

我们将query $Q^i = W_Q^i H_f^i$，key $K^i = W_k^i \widetilde{X}_{mh\_n}^i$，value $V^i = W_v^i \widetilde{X}_{mh_n}^i$用于多头注意力机制，并且为了限制双模态融合与残差块的信息流，定义门控机制如下:

$$g_r^i = \sigma(W_r^i \hat{h}_f^i) \tag{19}$$
$$g_c^i = \sigma(W_c^i \hat{h}_f^i) \tag{20}$$
$$r^i = MH - ATT(Q^i, K^i, V^i) \tag{21}$$
$$\widetilde{X}_f^i = LN(g_c^i \odot r^i + g_r^i \odot X_f^{i-1}) \tag{22}$$
$$X_f^i = LN(\widetilde{X}_f^i + FFN(\widetilde{X}_f^i)) \tag{23}$$

其中，$W_* \in R^{d_s \times d_s}$，$* \in \{r, c\}$，$(f, n) \in \{(ta, a), (tv, v)\}$ 也就是说$X_f^i \in \{X_{ta}^i, X_{tv}^i\}$。即，模态融合层在第$i$层的文本模态分别与音频、视频模态融合的输出表示。

3.3 三种模态表示再融合模块

在第3.2节不同模态数据层次融合模块的基础上，本节对文本与音频融合表示和文本与视频融合表示进一步融合，使模型学习到更加全面的模态互补信息。为了减少冗余，并捕获潜在模态表征，本文加入了两种损失函数，相似性损失和异质损失。

3.3.1 相似性损失

相似性损失用于学习不同模态间的共性表征，为此，设计最小化相似性损失，以减少每种模态的共享表征之间的差异。本文使用中心距差异(CMD)作为相似性损失。CMD (Zellinger et al., 2017)是一种先进的距离度量，它通过匹配两个表示的阶数矩差衡量两个表示的分布之间的差异，两个分布越相似，CMD距离会越小。

$$\begin{aligned} CMD_K(X, Y) = & \frac{1}{b-a} \parallel E(X) - E(Y) \parallel_2 \\ & + \sum_{K=2}^{K} \frac{1}{|b-a|^k} \parallel C_k(X) - C_k(Y) \parallel_2 \end{aligned} \tag{24}$$

其中，$E(X) = \frac{1}{|X|} \sum_{x \in X} x$是样本X的经验期望向量。$C_k(X) = E((x - E(X))^k)$是所有$k^{th}$阶样本的中心矩的向量。

在本文中，计算文本音频模态特征序列级隐藏表示（$\hat{h}_{ta}^i$）与文本视频模态特征融合表示（$\hat{h}_{tv}^i$）的CMD损失如下:

$$\ell_{sim}^i = CMD_K(\hat{h}_{ta}^i, \hat{h}_{tv}^i) \tag{25}$$

3.3.2 异质损失

异质损失用于学习不同模态间的特性表征。我们使用软正交约束度量异质损失，获取不同模态的特性表征。软正交约束 (Bousmalis et al., 2016; Liu et al., 2017; Ruder and Plank, 2018)，计算定义如下：

$$\ell^{i}_{diff} = \sum_{(m1,m2)\in\{(ta,a),(tv,v)\}} \| \hat{h}^{i\top}_{m1}\hat{h}^{i}_{m2} \|^2_F \tag{26}$$

3.4 多任务学习模块

表征学习是多模态学习中一个重要而富有挑战性的课题。有效的情感表达应包含一致性和差异性两部分特征。由于统一的多模态标签，现有方法在捕获差异化信息时受到限制。然而，额外标注单模态标签需要耗费大量的时间和人力。在本节中，我们使用了一个基于自监督学习策略的标签生成模块来获取独立的单模态标签。然后，对多模态和单模态任务进行联合训练，分别学习模态的一致性和差异性特征。

利用第3.2.2节两两模态特征的层次融合表示，得到了两个多模态融合表示X^l_f，$f \in \{ta, tv\}$，再利用线性层FFN，得到多模态特征表示F^*_m和预测结果$\hat{y}_m$。

$$\begin{aligned} F^*_m &= FFN(Concat(X^l_{ta}, X^l_{tv})) \\ \hat{y}_m &= FFN(F^*_m) \end{aligned} \tag{27}$$

对于第3.2.1节的单模态特征表示X^l_n,同样利用线性层FFN得到单模态特征表示F^*_n和预测结果$\hat{y}_n$，其中，$n \in \{a, v\}$，l表示模态融合层的最后一层。

$$\begin{aligned} F^*_n &= FFN(X^l_n) \\ \hat{y}_n &= FFN(F^*_n) \end{aligned} \tag{28}$$

与此同时，本文利用Yu (2021)等人设计的单模态标签生成模块(ULGM，Unimodal Label Generation Module)进行单个模态的训练。一般来说，单模态标签与多模态标签高度相关。因此，ULGM根据从模态表示到类中心的相对距离计算偏移量。

对于模态表示，我们使用L2归一化作为F^*_i和类中心之间的距离。

$$D^p_i = \frac{\|F^*_i - C^p_i\|^2_2}{\sqrt{d_i}} \tag{29}$$

$$D^n_i = \frac{\|F^*_i - C^n_i\|^2_2}{\sqrt{d_i}} \tag{30}$$

其中，$i \in \{m, a, v\}$，d_i是表示维度的比例因子，C^p_i和C^n_i分别表示正类中心和负类中心。

然后，对相对距离值进行定义，该值评估了模态表示到正向中心和负向中心的相对距离。

$$\alpha_i = \frac{D^n_i - D^p_i}{D^p_i + \epsilon} \tag{31}$$

其中，$i \in \{m, a, v\}$。ϵ是一个很小的数，为了防止除零异常。

为了得到标签和预测值之间的联系，考虑以下两种关系。

$$\frac{y_s}{y_m} \propto \frac{\hat{y}_s}{\hat{y}_m} \propto \frac{\alpha_s}{\alpha_m} \Rightarrow y_s = \frac{\alpha_s * y_m}{\alpha_m} \tag{32}$$

$$y_s - y_m \propto \hat{y}_s - \hat{y}_m \propto \alpha_s - \alpha_m \Rightarrow y_s = y_m + \alpha_s - \alpha_m \tag{33}$$

其中，$s \in \{a, v\}$。我们可以通过等权求和得到单模态标签。

$$\begin{aligned} y_s &= \frac{y_m * \alpha_s}{2\alpha_m} + \frac{y_m + \alpha_s - \alpha_m}{2} \\ &= y_m + \frac{\alpha_s - \alpha_m}{2} * \frac{y_m + \alpha_m}{\alpha_m} \\ &= y_m + \delta_{sm} \end{aligned} \tag{34}$$

第二十二届中国计算语言学大会论文集，第315页-第327页，哈尔滨，中国，2023年8月3日至5日。

其中，$s \in \{a, v\}$。$\delta_{sm} = \frac{\alpha_s - \alpha_m}{2} * \frac{y_m + \alpha_m}{\alpha_m}$表示单模态标签对多模态标签的偏移值。为了减轻由式(34)计算得到的标签不够稳定的问题，设计了基于动量的更新策略，该策略将新生成的值与历史值相结合，经过多次迭代使生成的标签值稳定。

利用两个单模态的特征表示F_n^*、多模态融合后的特征表示F_m^*和多模态标签y_m，共同输入到ULGM 模块后，生成了两个伪标签y_a,y_v，用于两个单模态任务训练过程。对于单模态子任务，本文使用L1损失，对于多模态子任务，使用均方误差构造MMILN 模型的优化目标函数。对于这两个单模态子任务，伪标签y_n与实际标签y_m的差异作为损失函数的权重，即$W_n^i = tanh(|y_n^i - y_m|)$。这意味着模型将注意力更加集中到模态特征差异性较大的样本上。具体损失函数定义如下。

$$\ell_{task} = \frac{1}{N}\sum_{i=1}^{N}((\hat{y}_m^i - y_m)^2 + \sum_{n\in\{a,v\}} W_n^i * |\hat{y}_n^i - y_n^i|) \tag{35}$$

最终，将ℓ_{sim}^l和ℓ_{diff}^l损失函数与任务损失ℓ_{task}进行组合，作为模型整体的损失函数定义如下:

$$\ell = \ell_{task} + \alpha\ell_{sim}^l + \beta\ell_{diff}^l \tag{36}$$

其中，α、β为两个超参数，决定每个正则化分量对总体损失ℓ的贡献权重。

4 实验与分析

本节包括以下三个部分:实验数据、实验设置、实验结果及分析。

4.1 实验数据

本文采用的实验数据是Zadeh等人发布的CMU-MOSI数据集 (Zadeh et al., 2016)和CMU-MOSEI数据集 (Zadeh et al., 2018b)。它们都是利用社交媒体数据得到的多模态情感分析数据集。CMU-MOSI是从93个Youtube的视频中获取的2199个独白类型的短视频片段。CMU-MOSEI包括来自5000个视频的23453个视频片段。数据的标注由人工完成，为情感的评分，分数值从-3到+3七个等级，其中，负值代表消极情感，正值代表积极情感，0分代表无情感。具体统计信息如表1所示。在实验过程中，为保证结果的公平性，本文采用已有工作 (Yu et al., 2021)对CMU-MOSI和CMU-MOSEI分别按照6:1:3与7:1:2的比例划分训练集、验证集和测试集。

Dataset	#Train	#Valid	#Test	#All
MOSI	1284	229	686	2199
MOSEI	16326	1871	4659	22856

Table 1: CMU-MOSI与CMU-MOSEI数据集的统计信息

4.2 实验设置

本文所提出的MMILN模型是在Yu (2021)等人的Self-MM模型的基础上改进而来。除层次融合模块的参数外，其余参数与Self-MM模型完全一致。我们在CMU-MOSI和CMU-MOSEI数据集上对本文所提出的模型及各基线模型进行了情感分析实验。分别采用分类和回归两种方法进行情感类别判别。对于分类指标，本文分别采用positive/negative和non-negative/negative计算F1-Score和二分类精度（Acc2）。对于回归指标，采用平均绝对误差（MAE）和皮尔逊相关（Corr）。除MAE外，值越高表示这项指标的性能越好。

4.3 实验结果及分析

在这一节，先将本文的模型与基线模型的性能进行对比和分析，之后进行了消融实验、模态融合层的数量对模型效果的影响实验，以验证本文所提出的模型设计的合理性。

4.3.1 与基线模型的对比及分析

为了充分验证本文模型MMILN的性能，我们与多模态情感分析任务中的多个模型进行实验比较。比较模型如下：

- TFN(Zadeh et al., 2017)：计算多维张量（基于外积），获取单模态、双模态和三模态相互作用
- LMF(Liu et al., 2018)：对TFN的改进，采用低秩多模态张量融合技术来提高效率
- MFN(Zadeh et al., 2018a)：持续建模特定视图和交叉视图交互，并通过多视图门控内存对其进行总结
- RAVEN(Wang et al., 2019)：利用基于注意力的模型，根据辅助非言语信号重新调整单词嵌入
- MFM(Tsai et al., 2018)：学习生成表征，获取模态特定的生成特征以及分类的区别表征
- MulT(Tsai et al., 2019)：扩展了具有定向两两交叉关注的多模态Transformer架构，使模态定向两两交互
- MISA(Hazarika et al., 2020b)：结合损失组合，包括分布相似性、正交损失、重建损失和任务预测损失，学习模态不变和模态特定表示
- MAG-BERT(Rahman et al., 2020)：对RAVEN在对齐数据方面改进，在Bert主干的不同层应用了多模态适配门
- SELF-MM(Yu et al., 2021)：为每个模态分配一个带有自动生成标签的单模态训练任务，目的是调整梯度反向传播
- MMIM(Han et al., 2021)：使用一个分层的互信息最大化框架来指导模型从所有模态中学习共享表示
- AMML(Sun et al., 2022)：引入了一种基于元学习的方法来学习更好的单模态表示，然后将其用于随后的多模态融合
- EMT(Sun et al., 2023)：利用统一的框架EMT-DLFR来实现鲁棒MSA，并具有更好的性能

	CMU-MOSI				CMU-MOSEI			
模型	ACC-2	F1-Score	Corr	MAE	ACC-2	F1-Score	Corr	MAE
TFN	-/80.8	-/80.7	0.698	0.901	-/82.5	-/82.1	0.700	0.593
LMF	-/82.5	-/82.4	0.695	0.917	-/82.0	-/82.1	0.677	0.623
MFN	77.4/-	77.3/-	0.632	0.965	76.0/-	76.0/-	-	-
RAVEN	78.0/-	76.6/-	0.691	0.915	79.1/-	79.5/-	0.662	0.614
MFM	-/81.7	-/81.6	0.706	0.877	-/84.4	-/84.3	0.717	0.568
MulT	81.5/84.1	80.6/83.9	0.711	0.861	-/82.5	-/82.3	0.703	0.58
MISA	81.8/83.4	81.7/83.6	0.761	0.783	83.6/85.5	83.8/85.3	0.756	0.555
MAG-BERT	84.2/86.1	84.1/86.0	0.796	0.712	84.7/-	84.5/-	-	-
Self-MM	84.00/85.98	84.42/85.95	0.798	0.713	82.81/85.17	82.53/85.30	0.765	0.530
MMIM	84.14/86.06	84.00/85.98	0.800	0.700	82.24/85.97	82.66/85.94	0.772	0.526
AMML	-/84.9	-/84.8	0.792	0.723	-/85.3	-/85.2	0.776	0.614
EMT	83.3/85.0	83.2/85.0	0.798	0.705	83.4/86.0	83.7/86.0	0.774	0.527
MAG-BERT*	82.54/84.36	82.48/84.37	0.796	0.717	82.10/85.09	82.42/84.95	0.754	0.543
MMIM*	83.67/85.52	83.56/85.47	0.800	**0.694**	77.81/82.55	78.66/82.7	0.700	0.615
EMT*	82.75/84.45	82.67/84.43	0.795	0.709	77.38/83.41	78.31/83.58	0.768	0.532
MMILN（Ours）	**84.55/86.43**	**84.49/86.42**	**0.801**	0.709	**84.67/85.86**	**84.70/85.62**	**0.774**	**0.529**

Table 2: CMU-MOSI数据集与CMU-MOSI数据集的结果

在表2上，分别展示了各模型在CMU-MOSI 和CMU-MOSEI数据集上情感分析的性能，*表示模型在相同条件下重现的结果，“/”左边表示non-negative或negative，右边表示positive或negative。由表2可以看到：

（1）本文的MMILN 模型在两个数据集上的分类性能均优于当前已有模型，且在MOSEI上的回归性能也优于其他模型，但MOSI数据集上的回归表现却并非最佳。其主要原因是MOSEI的数据量远超MOSI，使得测试实验更加稳定，也就是说模型在MOSEI上的性能可能更具有代表性和普适性。

（2）本文的MMILN 模型相较于基线模型Self-MM性能均较为突出。在CMU-MOSI和CMU-MOSEI数据集的ACC-2分别增加了0.55/0.45和1.86/0.69，F1-Score相较于Self-MM分别增加了0.07/0.47和2.17/0.32，Corr相较于Self-MM分别增加了0.003和0.009，MAE相较于Self-MM分别下降了0.004和0.001。这表明，我们所提出的从低级特征渐变到高级特征的融合策略有助于提升模型的整体性能。

4.3.2 消融实验

为了验证本文提出模型各模块的性能，我们使用CMU-MOSEI数据集设计如下的消融实验，其中减号“-”表示在这组实验中删掉的模型结构。

- -MFM：删除单模态特征表示门控机制与两两模态特征的层次融合门控机制（删除本模块会同时删除MRM）
- -MRM：删除三种模态表示再融合模块
- -GateFusion：删除多模态门控融合，仅使用单模态特征进行模态融合

模型/结构	ACC-2	F1-Score	Corr	MAE
-MRM	**85.13**/85.00	84.84/84.52	0.762	0.586
-MFM	84.01/85.75	84.20/85.62	0.771	0.574
-GateFusion	84.98/85.09	**84.78**/84.68	0.764	0.576
MMILN（Ours）	84.67/**85.86**	84.70/**85.62**	**0.774**	**0.529**

Table 3: 消融实验

从表3可以看出，本文所提模型的各个模块均发挥了一定的作用。-MRM模型相比MMILN，CMU-MOSEI数据集在positive/negative (right)方面，ACC-2下降了0.86，F1-Score下降了1.1，Corr下降了0.012，MAE增加了0.057。相应的，在non-negative/negative(left)方面，ACC-2和中F1-Score分别增加了0.46和0.14。说明MRM可以使得模型对于多模态融合特征表示有一定的提升效果，并进一步融合多模态融合特征表示，使模型学习到更加全面的模态互补信息。

-MFM模型相比MMILN，CMU-MOSEI数据集的ACC-2下降了0.66/0.11，F1-Score下降了0.50/0.00，Corr下降了0.003，MAE上升了0.045。说明MFM 与MRM通过使用门控机制可以对信息流进行有效地控制，并且使用两个损失函数可以将多模态表示进一步融合。

-GateFusion模型相比MMILN，在CMU-MOSEI数据集positive/negative (right)方面，ACC-2下降了0.77，F1-Score下降了0.94，Corr下降了0.010，MAE增加了0.047。相应的，在non-negative/negative (left)，ACC-2 增加了0.31和F1-Score 增加了0.08。说明仅使用单模态特征进行模态融合不利于模型学习到更加全面的模态信息，而通过门控机制对单模态信息进行筛选有助于提升模型的整体性能。

4.3.3 实例分析

为了展示CMU-MOSI数据集的示例，如表4所示。基本事实情感标签介于强烈消极(-3)和强烈积极(+3)之间。对于每个例子，均展示了Ground Truth和预测输出。其中，MMILN模型相对Self-MM模型，情感预测的性能更加准确。

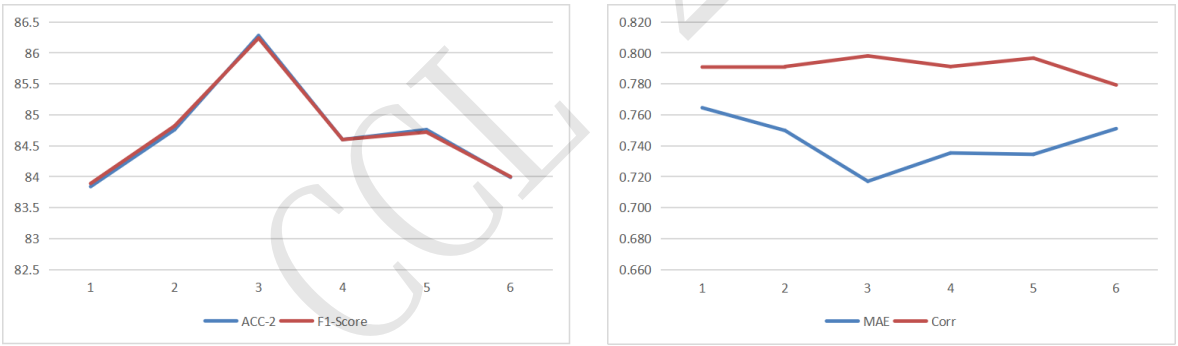
#	Spoken words + acoustic + visual behaviors	Ground Truth	MMILN	Self-MM
1	“I think you will really love this movie if you are 8.” + 强调的语音+嘴巴张大并且表情震惊	2.0	2.0	2.3
2	“It seems like the writers and creators took the easy way out.”+ 强调的语音+ 皱眉、眼角上扬	-1.6	-0.1	0.3
3	“I though that adults would appreciate” + 语调平稳+ 点头、眼睛左右看	0.4	0.7	1.2
4	“and I think its predictable up an to a point” + 语调平稳+ 表情平淡	-1.0	0.3	-0.4

Table 4: 实例分析

在例1和例3中，MMILN均相对准确的预测了情感的强度。在例2中，文本的描述相对中性，但是强调的语音和皱眉、紧张的表情帮助了MMILN模型预测的情感是负面极性。而Self-MM模型预测的情感极性是错误的，标注了正面极性，说明Self-MM模型并没有很好的利用到语音和视频模态。进一步证明了MMILN模型通过多层网络结构，采用从低级特征渐变到高级特征的融合策略是有效的。在例4中，MMILN模型将情感极性预测错误，但是，经过我们查验原始数据集后发现，原始数据的文本没有体现出情感，音频的语调平稳且说话人的表情平淡，并没有体现出情感，再一次体现出我们所提出的MMILN模型的性能比较突出。

4.3.4 模态融合层的数量对模型效果的影响实验

在本节中，我们将通过实验来研究模态融合层的数量对模型效果的影响。我们的实验在CMU-MOSI数据集上进行。其中，ACC-2与F1-Score取positive/negative值，模态融合层的层数选取1-6层。图4(a)展示的是不同层数下ACC-2与F1-Score的性能，图4(b)展示的是不同层数下MAE与Corr的性能。



(a)不同层数下ACC-2与F1-Score的性能展示　(b)不同层数下MAE与Corr的性能展示

Figure 4: 不同层数下MMILN模型性能比较

从图4(a)中可以看到，随着模态融合层的层数增加，ACC-2与F1-Score呈现先上升后下降的趋势，在层数为3时模型效果达到了最优。从图4(b)中可以看到，随着模态融合层的层数增加，Corr指标呈现波动，在层数为3和5时，模型效果较优。而MAE指标呈现出先下降后上升的趋势，总体来看，两者都在层数为3时模型效果达到了最优。这表明，我们所提出的MMILN模型对于模态融合层的层数较为敏感，提高模态融合层的层数一定程度上可以加强模态特征的融合，得到更加丰富的模态表征，并且提高模型的性能。

5 结论

本文提出了多任务多模态交互学习的自监督动态融合方法，该方法通过交互学习方式，采用从低级特征渐变到高级特征的融合策略，获得了模态间的一致性特征，再利用两个单模态特

征抽取子任务，获得模态自身的独特特征，从而对多模态融合特征进行了有效的补充和加强，较好地实现了模态间的相互作用。使用MFM和GateFusion，模型能够动态地、自适应调节不同层次的模态特征融合过程，减少与任务无关的弱模态特征信息对结果预测的干扰。为了进一步提升模型的性能，设计了模态表示再融合模块，加强了模态的融合。最后，通过实验验证，本文提出的方法加入到多任务学习的框架中之后，在测试集上性能优于主流基线模型，提升了模型的整体性能。

参考文献

Md. Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task learning for multi-modal emotion recognition and sentiment analysis. *CoRR*, abs/1905.05812.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. *Advances in neural information processing systems*, 29.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI '17, page 163–171, New York, NY, USA. Association for Computing Machinery.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. CASCADE: contextual sarcasm detection in online discussion forums. *CoRR*, abs/1805.06413.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020a. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1122–1131, New York, NY, USA. Association for Computing Machinery.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020b. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742*.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1044–1054.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.

Zhongkai Sun, Prathusha Kameswara Sarma, William A. Sethares, and Yingyu Liang. 2019. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. *CoRR*, abs/1911.05544.

Y. Sun, S. Mai, and H. Hu. 2022. Learning to learn better unimodal representations via adaptive multimodal meta-learning. *IEEE Transactions on Affective Computing*, (01):1–1, may.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2023. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, pages 1–17.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Yao-Hung Hubert Tsai, Martin Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1823–1833, Online, November. Association for Computational Linguistics.

Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P. Xing. 2017. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 949–954.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10790–10797.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811*.

第二十二届中国计算语言学大会论文集，第315页-第327页，哈尔滨，中国，2023年8月3日至5日。