TelME: Teacher-leading Multimodal Fusion Network for Emotion Recognition in Conversation

Peizhen Li

Faculty of Science and Engineering
Macquarie University

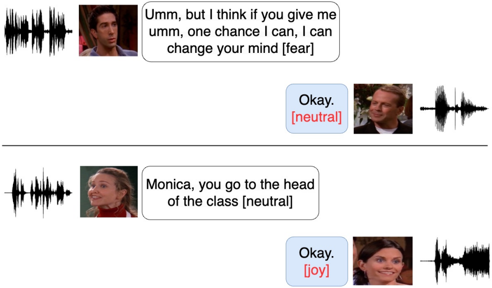
Aug 30, 2024

MACQUARIE University

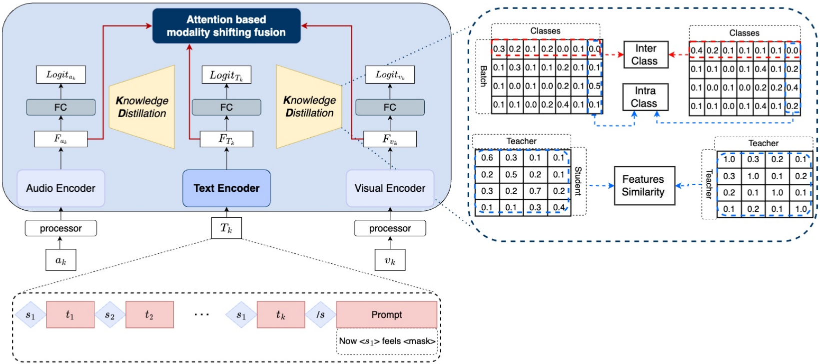
Outline

Introduction

- Task: emotion recognition in conversation
- Motivation: improve the efficacy of weak non-verbal modalities
- Solution and contributions:
  - cross-modal knowledge distillation
  - attention-based modality shifting fusion



Model Overview



Feature Extraction

- Text (encoder: modified Roberta)

$$\begin{aligned} C_k &= [< s_i >, t_1, < s_j >, t2, \cdots, < s_i >, t_k] \\ P_k &= \text{Now} < s_i > \text{feels} < mask > \\ F_{T_k} &= TextEncoder(C_k < /s > P_k) \end{aligned} \tag{1}$$

- Audio (encoder: data2vec)

$$F_{a_k} = AudioEncoder(a_k) \tag{2}$$

- Vision (encoder: Timesformer)

$$F_{v_k} = VisualEncoder(v_k) \tag{3}$$

$$F_{T_k}, F_{a_k}, F_{v_k} \in \mathbb{R}^{1\times d}$$

Knowledge Distillation

$$
\begin{aligned}
L_{student} &= L_{cls} + \alpha L_{response} + \beta L_{feature} \\
L_{response} &= L_{inter} + L_{intra} \\
L_{feature} &= \frac{1}{B}\sum_{i=1}^{B} KL(P_i||Q_i)
\end{aligned}
\tag{4}
$$

$$
L_{inter} = \frac{\tau^2}{B}\sum_{i=1}^{B} d(Y^s_{i,:}, Y^t_{i,:}), \quad L_{intra} = \frac{\tau^2}{C}\sum_{j=1}^{C} d(Y^s_{:,j}, Y^t_{:,j})
$$

$$
Y^t_{i,:} = softmax(Z^t_{i,:}/\tau), \quad Y^s_{i,:} = softmax(Z^s_{i,:}/\tau), \quad d(\mu, v) = 1 - \rho(\mu, v)
$$

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

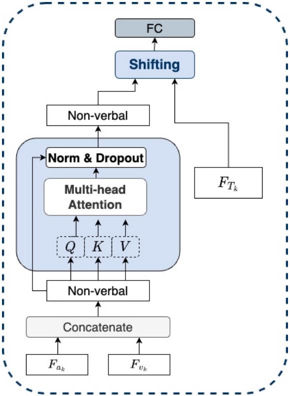
$$
Z^s, Z^t \in \mathbb{R}^{B\times C}, \quad M, M^{'} \in \mathbb{R}^{B\times B}
$$

Knowledge Distillation

```
class Logit_Loss(nn.Module):
    def __init__(self, beta=1.0, gamma=1.0, tau=4.0):⋯
    def forward(self, z_s, z_t):
        y_s = (z_s / self.tau).softmax(dim=1)
        y_t = (z_t / self.tau).softmax(dim=1)
        inter_loss = self.tau**2 * inter_class_relation(y_s, y_t)
        intra_loss = self.tau**2 * intra_class_relation(y_s, y_t)
        kd_loss = self.beta * inter_loss + self.gamma * intra_loss
        return kd_loss

class Feature_Loss(nn.Module):
    def __init__(self, temp=1.0):⋯
    def forward(self, other_embd, text_embd):
        text_embd = F.normalize(text_embd, p=2, dim=1)
        other_embd = F.normalize(other_embd, p=2, dim=1)
        target = torch.matmul(text_embd, text_embd.transpose(0,1))
        x = torch.matmul(text_embd, other_embd.transpose(0,1))
        log_q = torch.log_softmax(x / self.t, dim=1)
        p = torch.softmax(target / self.t, dim=1)
        return F.kl_div(log_q, p, reduction='batchmean')
```

Attention-based Modality Shifting Fusion



$$Z_k = F_{T_k} + \lambda \cdot H_k \tag{5}$$

$$H_k = g_{AV}^k \cdot (W_2 \cdot F_{attention}^k + b_2), \;\; g_{AV}^k = R(W_1 \cdot < F_{T_k}, F_{attention}^k > + b_1)$$

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

Experimental Results

Models	MELD: Emotion Categories								IEMOCAP
	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	F1	F1
DialogueRNN (Majumder et al., 2019)	73.50	49.40	1.20	23.80	50.70	1.70	41.50	57.03	62.75
ConGCN (Zhang et al., 2019)	76.70	50.30	8.70	28.50	53.10	10.60	46.80	59.40	64.18
MMGCN (Hu et al., 2021b)	-	-	-	-	-	-	-	58.65	66.22
DialogueTRM (Mao et al., 2021)	-	-	-	-	-	-	-	63.50	69.23
DAG-ERC (Shen et al., 2021)	-	-	-	-	-	-	-	63.65	68.03
MM-DFN (Hu et al., 2022a)	77.76	50.69	-	22.94	54.78	-	47.82	59.46	68.18
M2FNet (Chudasama et al., 2022)	-	-	-	-	-	-	-	66.71	69.86
EmoCaps (Li et al., 2022)	77.12	**63.19**	3.03	42.52	57.50	7.69	**57.54**	64.00	**71.77**
UniMSE (Hu et al., 2022b)	-	-	-	-	-	-	-	65.51	70.66
GA2MIF (Li et al., 2023a)	76.92	49.08	-	27.18	51.87	-	48.52	58.94	70.00
FacialMMT (Zheng et al., 2023)	80.13	59.63	19.18	41.99	64.88	18.18	56.00	66.58	-
TelME	**80.22**	60.33	**26.97**	**43.45**	**65.67**	**26.42**	56.70	**67.37**	70.48

Leaderboard of Emotion Recognition in Conversation on MELD

Thank you very much!
Q&A