# MultiEMO: An Attention-Based Correlation-Aware Multimodal Fusion Framework for Emotion Recognition in Conversations

**Tao Shi**
Tsinghua Shenzhen International Graduate School,
Tsinghua University
shitao21@mails.tsinghua.edu.cn

**Shao-Lun Huang***
Tsinghua Shenzhen International Graduate School,
Tsinghua University
shaolun.huang@sz.tsinghua.edu.cn

Code:None

—— ACL 2023

2023.8.27 • ChongQing

Reported by JiaWei Cheng

# Motivation

(1) The complementarity of multimodal information has not been well exploited.

**Utterance**: *"Chandler is a great name!"*

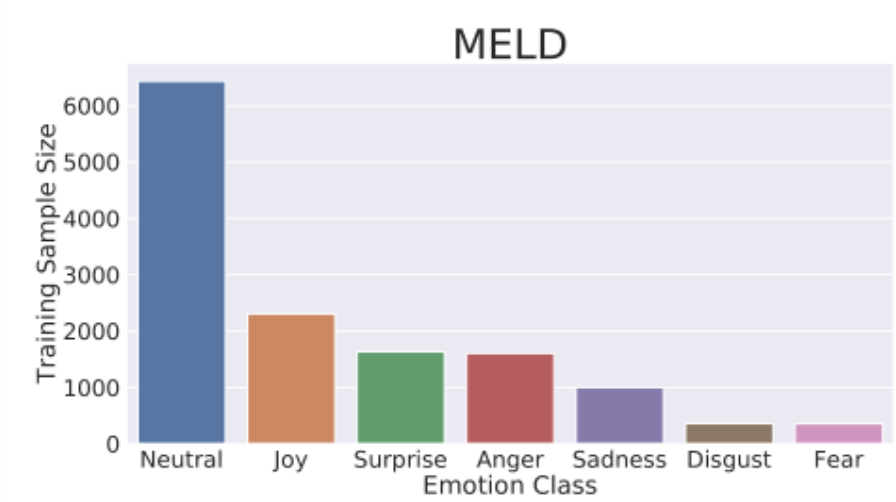
**Speaker**: Phoebe **Emotion**: *Anger*

| Text | Audio | Visual |
|---|---|---|
| Positive | Angry tone | Frown |

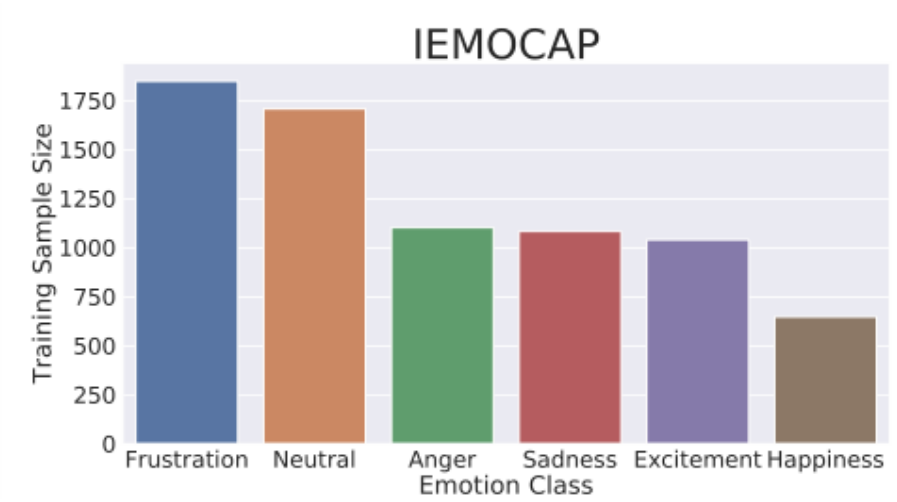
Figure 1: Illustration of the significance of multimodal cues for an accurate prediction, with blue indicating key modalities responsible for the emotion of the utterance.

# Motivation

(2) Unsatisfactory performances in minority emotion classes




(a) Class distribution in MELD.

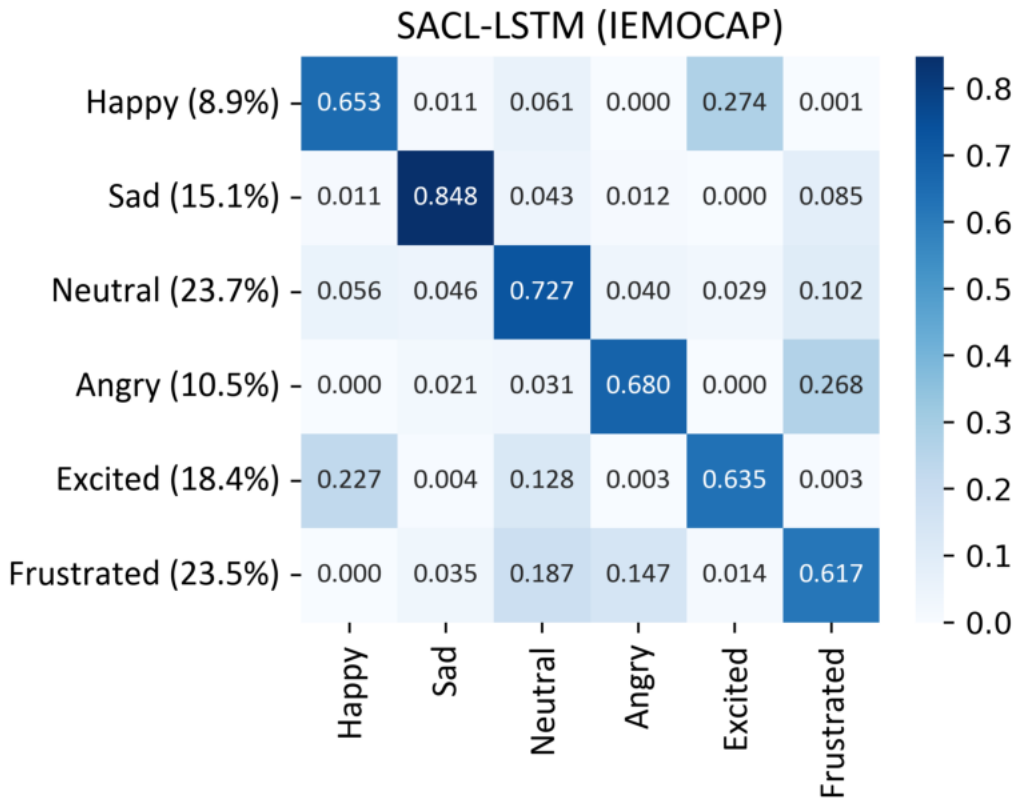



(b) Class distribution in IEMOCAP.

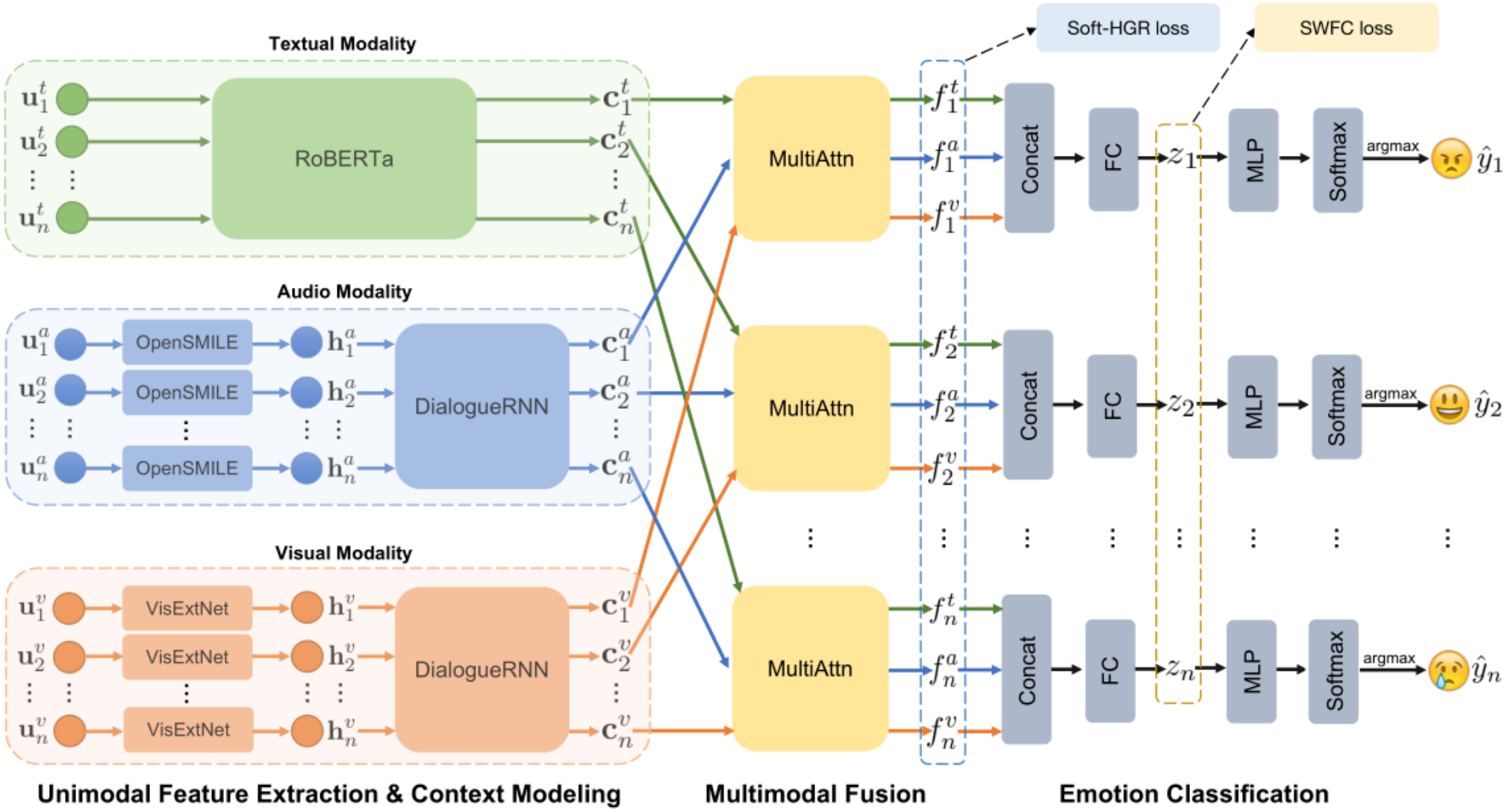
Figure 2: Illustration of the class imbalance problem in MELD and IEMOCAP.

# Motivation

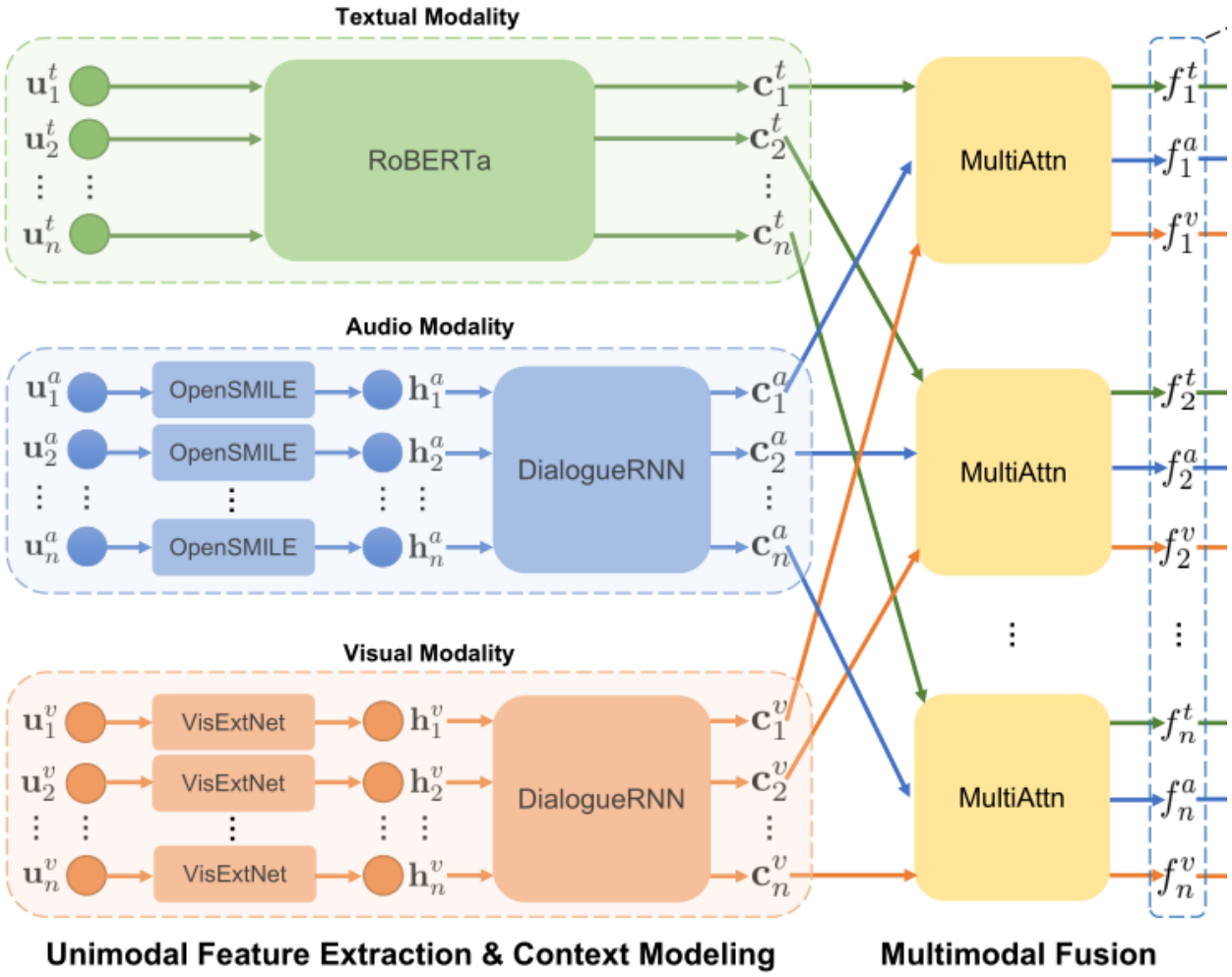
(3) The difficulty of distinguishing between semantically similar emotions.



# Overview


Textual Modality
RoBERTa
Audio Modality
OpenSMILE
DialogueRNN
Visual Modality
VisExtNet
DialogueRNN
MultiAttn
MultiAttn
MultiAttn
Soft-HGR loss
SWFC loss
Concat
FC
MLP
Softmax
argmax
Unimodal Feature Extraction & Context Modeling
Multimodal Fusion
Emotion Classification

# Method

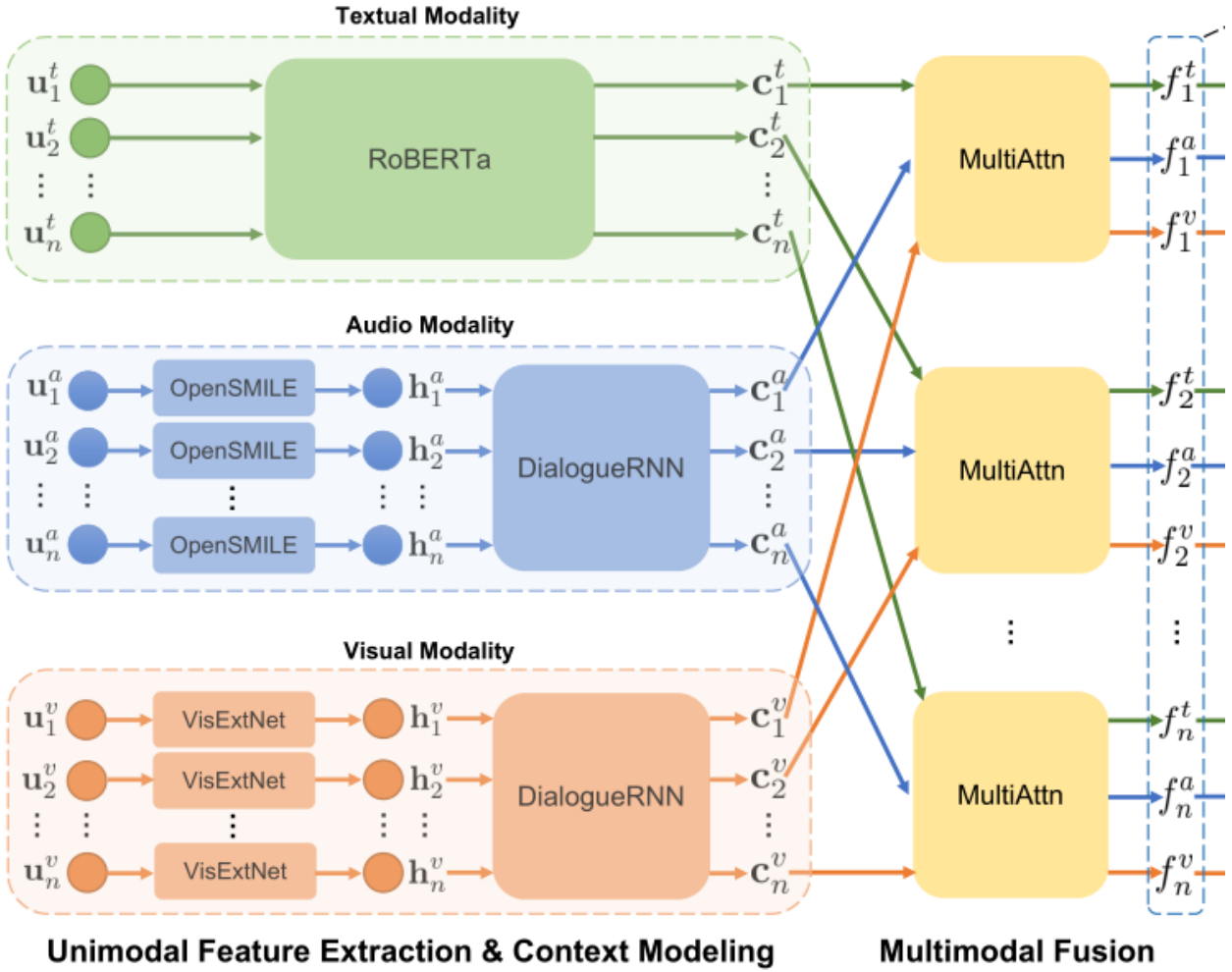



$$\mathbf{u}_i = \{\mathbf{u}_i^t, \mathbf{u}_i^a, \mathbf{u}_i^v\}, i \in \{1, \ldots, n\} \quad (1)$$

$$[\mathbf{Q}_h^{ta(j)}, \mathbf{K}_h^{ta(j)}, \mathbf{V}_h^{ta(j)}] = [\mathbf{F}^{t(j-1)}\mathbf{W}^{\mathbf{Q}_h^{ta(j)}}, \mathbf{C}^a\mathbf{W}^{\mathbf{K}_h^{ta(j)}}, \mathbf{C}^a\mathbf{W}^{\mathbf{V}_h^{ta(j)}}], h \in \{1, \ldots, H\} \quad (2)$$

$$\mathbf{A}_h^{ta(j)} = \text{Softmax}(\frac{\mathbf{Q}_h^{ta(j)}\mathbf{K}_h^{ta(j)\mathrm{T}}}{\sqrt{d_{\mathbf{K}_h^{ta(j)}}}})\mathbf{V}_h^{ta(j)}, \quad h \in \{1, \ldots, H\} \quad (3)$$

# Method




$$\mathbf{MH}^{ta(j)} = \mathrm{Cat}(\mathbf{A}_1^{ta(j)}, \ldots, \mathbf{A}_H^{ta(j)})\mathbf{W}^{O^{ta(j)}} \quad (4)$$

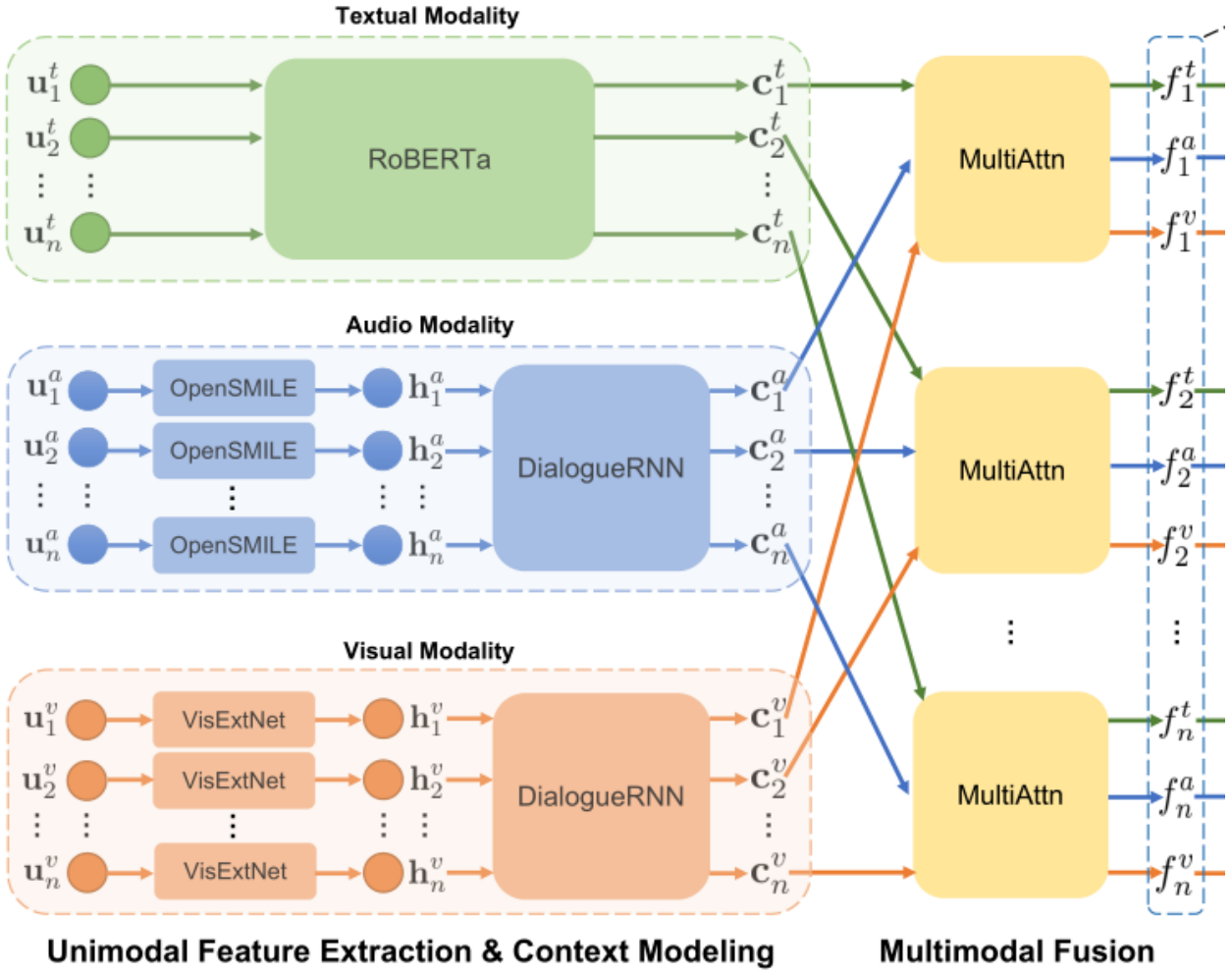
$$\mathbf{F}^{ta(j)} = \mathrm{LayerNorm}(\mathbf{F}^{t(j-1)} + \mathbf{MH}^{ta(j)}) \quad (5)$$

$$[\mathbf{Q}_h^{tav(j)}, \mathbf{K}_h^{tav(j)}, \mathbf{V}_h^{tav(j)}] = [\mathbf{F}^{ta(j)}\mathbf{W}^{\mathbf{Q}_h^{tav(j)}}, \mathbf{C}^v\mathbf{W}^{\mathbf{K}_h^{tav(j)}}, \mathbf{C}^v\mathbf{W}^{\mathbf{V}_h^{tav(j)}}], h \in \{1, \ldots, H\} \quad (6)$$

$$\mathbf{A}_h^{tav(j)} = \mathrm{Softmax}(\frac{\mathbf{Q}_h^{tav(j)}\mathbf{K}_h^{tav(j)\mathrm{T}}}{\sqrt{d_{\mathbf{K}_h^{tav(j)}}}})\mathbf{V}_h^{tav(j)},$$

$$h \in \{1, \ldots, H\} \quad (7)$$

# Method




$$\mathbf{MH}^{tav^{(j)}} = \mathrm{Cat}(\mathbf{A}_1^{tav^{(j)}}, \ldots, \mathbf{A}_H^{tav^{(j)}})\mathbf{W}^{O^{tav^{(j)}}} \tag{8}$$

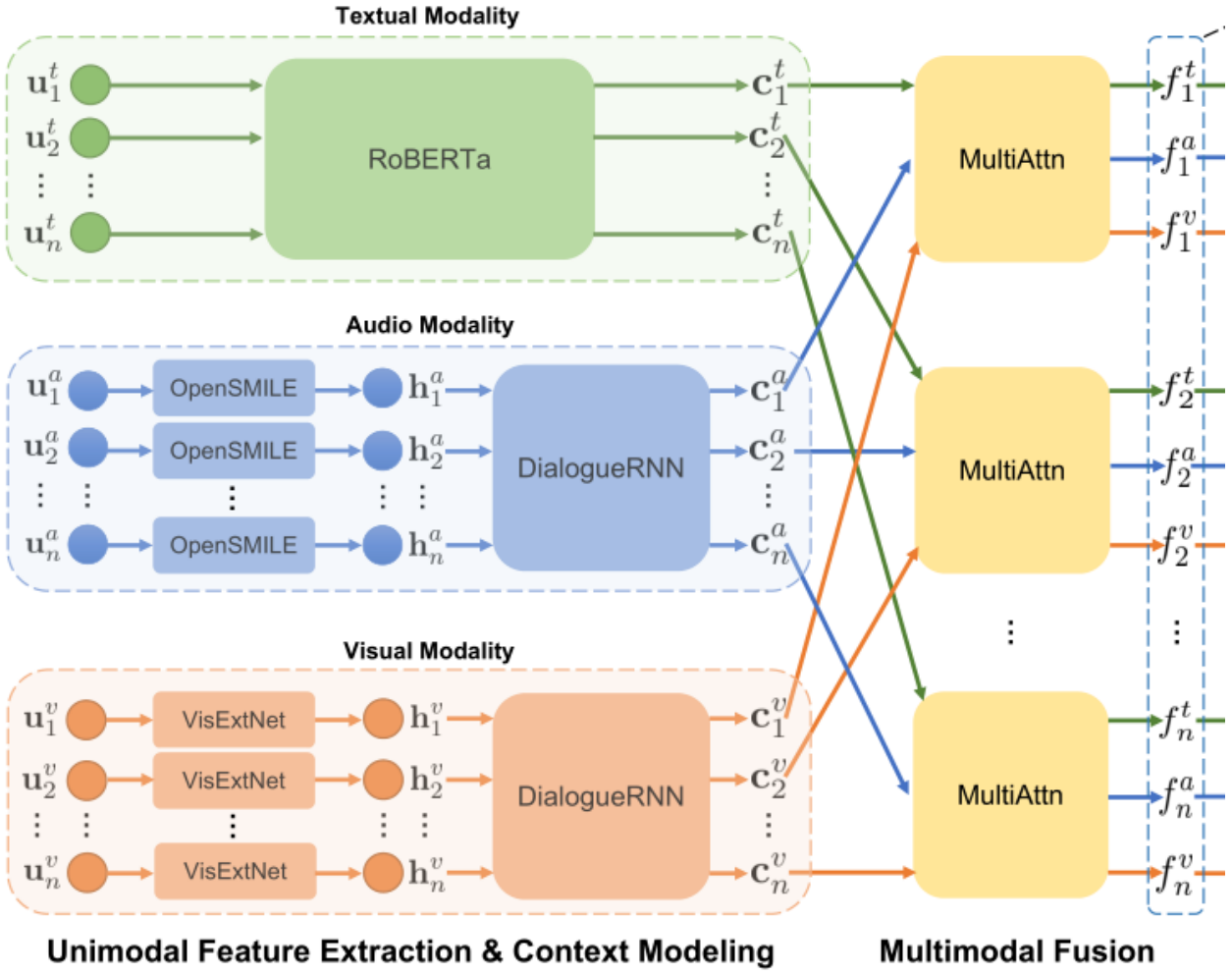
$$\mathbf{F}^{tav^{(j)}} = \mathrm{LayerNorm}(\mathbf{F}^{ta^{(j)}} + \mathbf{MH}^{tav^{(j)}}) \tag{9}$$

$$\mathbf{FFN}_1^{t^{(j)}} = \max(0, \mathbf{F}^{tav^{(j)}}\mathbf{W}^{\mathbf{FFN}_1^{t^{(j)}}} + \mathbf{b}_{\mathbf{FFN}_1^{t^{(j)}}}) \tag{10}$$

$$\mathbf{FFN}_2^{t^{(j)}} = \mathbf{FFN}_1^{t^{(j)}}\mathbf{W}^{\mathbf{FFN}_2^{t^{(j)}}} + \mathbf{b}_{\mathbf{FFN}_2^{t^{(j)}}} \tag{11}$$

$$\mathbf{F}^{t^{(j)}} = \mathrm{LayerNorm}(\mathbf{F}^{tav^{(j)}} + \mathbf{FFN}_2^{t^{(j)}}) \tag{12}$$

# Method




$$\mathbf{MH}^{tav^{(j)}} = \mathrm{Cat}(\mathbf{A}_1^{tav^{(j)}}, \ldots, \mathbf{A}_H^{tav^{(j)}})\mathbf{W}^{O^{tav^{(j)}}} \tag{8}$$

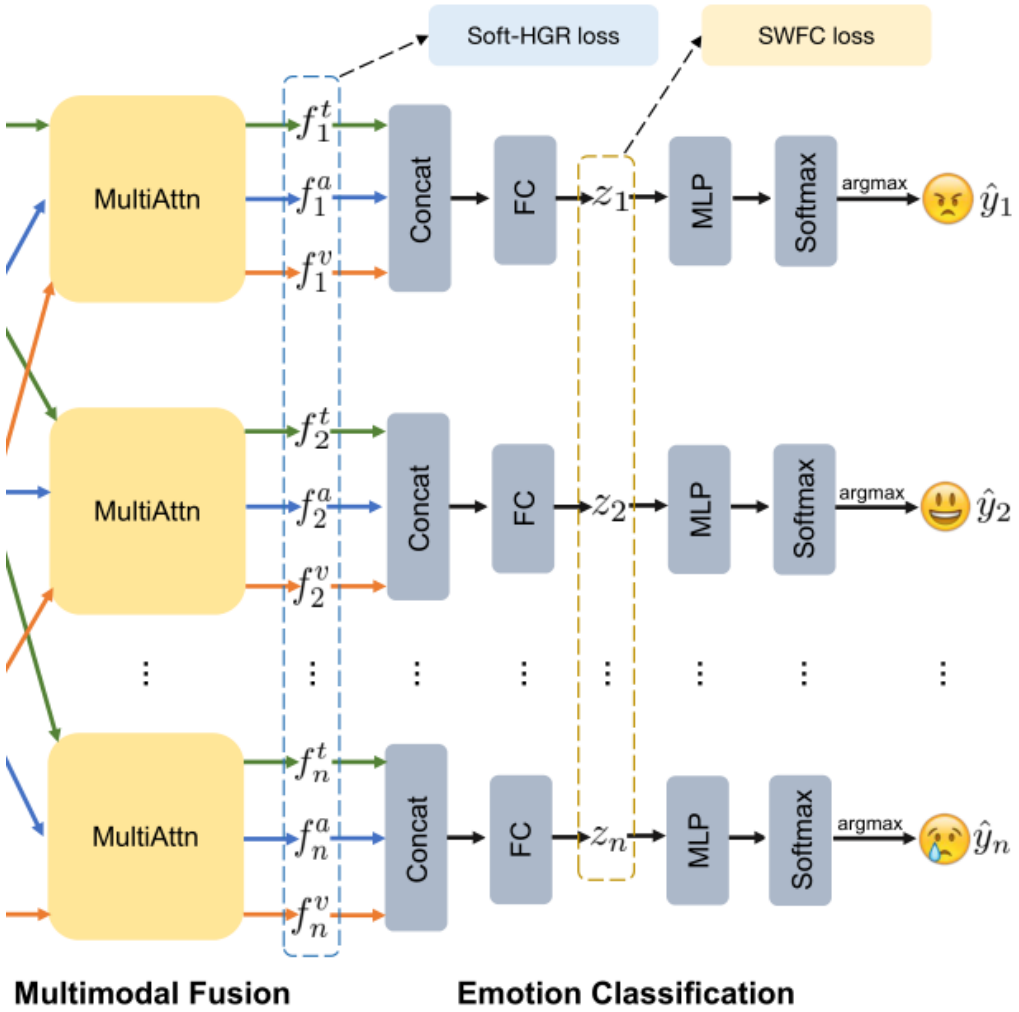
$$\mathbf{F}^{tav^{(j)}} = \mathrm{LayerNorm}(\mathbf{F}^{ta^{(j)}} + \mathbf{MH}^{tav^{(j)}}) \tag{9}$$

$$\mathbf{FFN}_1^{t^{(j)}} = \max(0, \mathbf{F}^{tav^{(j)}}\mathbf{W}^{\mathbf{FFN}_1^{t^{(j)}}} + \mathbf{b}_{\mathbf{FFN}_1^{t^{(j)}}}) \tag{10}$$

$$\mathbf{FFN}_2^{t^{(j)}} = \mathbf{FFN}_1^{t^{(j)}}\mathbf{W}^{\mathbf{FFN}_2^{t^{(j)}}} + \mathbf{b}_{\mathbf{FFN}_2^{t^{(j)}}} \tag{11}$$

$$\mathbf{F}^{t^{(j)}} = \mathrm{LayerNorm}(\mathbf{F}^{tav^{(j)}} + \mathbf{FFN}_2^{t^{(j)}}) \tag{12}$$

# Method




$$\mathbf{f}_i = \mathbf{f}_i^t \oplus \mathbf{f}_i^a \oplus \mathbf{f}_i^v \tag{13}$$

$$\mathbf{z}_i = \mathbf{W}^z \mathbf{f}_i + \mathbf{b}_z \tag{14}$$

$$\mathbf{l}_i = \max(0, \mathbf{W}^l \mathbf{z}_i + \mathbf{b}_l) \tag{15}$$

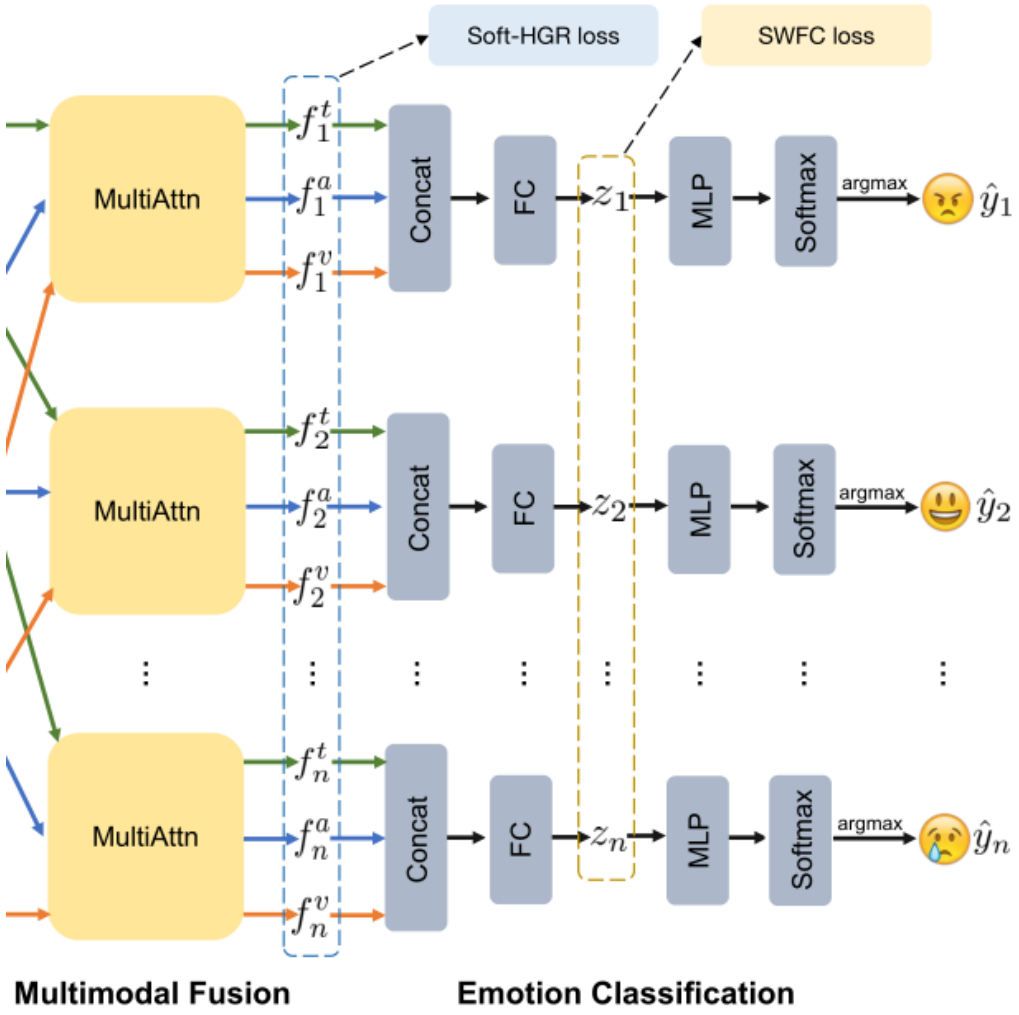
$$\mathbf{p}_i = \text{Softmax}(\mathbf{W}^{smax} \mathbf{l}_i + \mathbf{b}_{smax}) \tag{16}$$

$$\hat{y}_i = \text{argmax}(\mathbf{p}_i[t]) \tag{17}$$

$$s_{j,g}^{(i)} = \frac{\exp\left(\mathbf{z}_{i,j}^{\mathrm{T}} \mathbf{z}_{i,g} / \tau\right)}{\sum_{\mathbf{z}_{i,s} \in A_{i,j}} \exp\left(\mathbf{z}_{i,j}^{\mathrm{T}} \mathbf{z}_{i,s} / \tau\right)} \tag{18}$$

$$L_{\text{SWFC}} = -\sum_{i=1}^{M} \sum_{j=1}^{C(i)} \left(\frac{N}{n_{y_{i,j}}}\right)^{\alpha} \frac{1}{|R_{i,j}|} \sum_{\mathbf{z}_{i,g} \in R_{i,j}} (1 - s_{j,g}^{(i)})^{\gamma} \log s_{j,g}^{(i)} \tag{19}$$

# Method




$$L_{\text{Soft-HGR}} = -\sum_{\mathbf{Q}\neq\mathbf{V},\mathbf{Q},\mathbf{V}\in F}(\mathbb{E}[\mathbf{Q}^{\mathrm{T}}\mathbf{V}] - \frac{1}{2}\text{tr}(\text{cov}(\mathbf{Q})\text{cov}(\mathbf{V}))) \quad (20)$$

$$\text{s.t.} \quad \mathbb{E}[\mathbf{Q}] = 0, \ \forall \mathbf{Q} \in F.$$

$$L_{\text{CE}} = -\sum_{i=1}^{M}\sum_{j=1}^{C(i)} \log \mathbf{p}_{i,j}[y_{i,j}] \quad (21)$$

$$L_{\text{Train}} = \frac{1}{N}(\mu_1 L_{\text{SWFC}} + \mu_2 L_{\text{Soft-HGR}} + (1-\mu_1-\mu_2)L_{\text{CE}}) + \lambda||\theta||_2^2, \ \mu_1, \mu_2 \in [0,1] \quad (22)$$

# Experiments

| Models | MELD | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Neutral | Surprise | Fear | Sadness | Joy | Disgust | Angry | **Weighted-F1** |
| BC-LSTM | 73.80 | 47.70 | 5.40 | 25.10 | 51.30 | 5.20 | 38.40 | 55.90 |
| DialogueRNN | 76.23 | 49.59 | 0.00 | 26.33 | 54.55 | 0.81 | 46.76 | 58.73 |
| DialogueGCN | 76.02 | 46.37 | 0.98 | 24.32 | 53.62 | 1.22 | 43.03 | 57.52 |
| IterativeERC | 77.52 | 53.65 | 3.31 | 23.62 | 56.63 | 19.38 | 48.88 | 60.72 |
| QMNN | 77.00 | 49.76 | 0.00 | 16.50 | 52.08 | 0.00 | 43.17 | 58.00 |
| MMGCN | - | - | - | - | - | - | - | 58.65 |
| MVN | 76.65 | 53.18 | 11.70 | 21.82 | 53.62 | 21.86 | 42.55 | 59.03 |
| UniMSE | - | - | - | - | - | - | - | 65.51 |
| MultiEMO$_{\text{w/o VisExtNet}}$ | 79.16 | 58.22 | 24.80 | 37.61 | 60.65 | 31.73 | 52.08 | 64.89 |
| MultiEMO$_{\text{w/o MultiAttn}}$ | 77.72 | 54.05 | 21.76 | 33.10 | 58.28 | 24.80 | 49.98 | 62.50 |
| MultiEMO$_{\text{w/o SWFC loss}}$ | 79.51 | 56.54 | 20.59 | 32.96 | 58.52 | 25.81 | 51.23 | 63.83 |
| MultiEMO | **79.95** | **60.98** | **29.67** | **41.51** | **62.82** | **36.75** | **54.41** | **66.74** |

66.55

Table 2: Experimental results on MELD. The best results are highlighted in bold. "-" means that the results are unavailable from the original paper.

# Experiments

| Models | IEMOCAP | | | | | | |
|---|---|---|---|---|---|---|---|
| | Happiness | Sadness | Neutral | Anger | Excitement | Frustration | **Weighted-F1** |
| BC-LSTM | 34.43 | 60.87 | 51.81 | 56.73 | 57.95 | 58.92 | 54.95 |
| DialogueRNN | 33.18 | 78.80 | 59.21 | 65.28 | 71.86 | 58.91 | 62.75 |
| DialogueGCN | 51.87 | 76.76 | 56.76 | 62.26 | 72.71 | 58.04 | 63.16 |
| IterativeERC | 53.17 | 77.19 | 61.31 | 61.45 | 69.23 | 60.92 | 64.37 |
| QMNN | 39.71 | 68.30 | 55.29 | 62.58 | 66.71 | 62.19 | 59.88 |
| MMGCN | 42.34 | 78.67 | 61.73 | 69.00 | 74.33 | 62.32 | 66.22 |
| MVN | 55.75 | 73.30 | 61.88 | 65.96 | 69.50 | 64.21 | 65.44 |
| UniMSE | - | - | - | - | - | - | 70.66 |
| $\text{MultiEMO}_{\text{w/o VisExtNet}}$ | 65.06 | 84.80 | 66.13 | 67.98 | 76.16 | 69.66 | 71.72 |
| $\text{MultiEMO}_{\text{w/o MultiAttn}}$ | 55.18 | 78.29 | 62.06 | 63.84 | 73.11 | 63.98 | 66.57 |
| $\text{MultiEMO}_{\text{w/o SWFC loss}}$ | 59.88 | 83.96 | 66.57 | 67.03 | 75.35 | 70.04 | 71.08 |
| MultiEMO | **65.77** | **85.49** | **67.08** | **69.88** | **77.31** | **70.98** | **72.84** |

73.14

Table 1: Experimental results on IEMOCAP. The best results are highlighted in bold. "-" means that the results are unavailable from the original paper.

# Experiments

| Modality | IEMOCAP | MELD |
|---|---|---|
| Text | 64.48 | 61.23 |
| Audio | 38.89 | 33.55 |
| Visual | 35.37 | 33.16 |
| Text + Audio | 69.18 | 64.21 |
| Text + Visual | 67.86 | 63.78 |
| Text + Audio + Visual | **72.84** | **66.74** |

Table 3: Experimental results of MultiEMO with different modality settings on IEMOCAP and MELD.

# Experiments

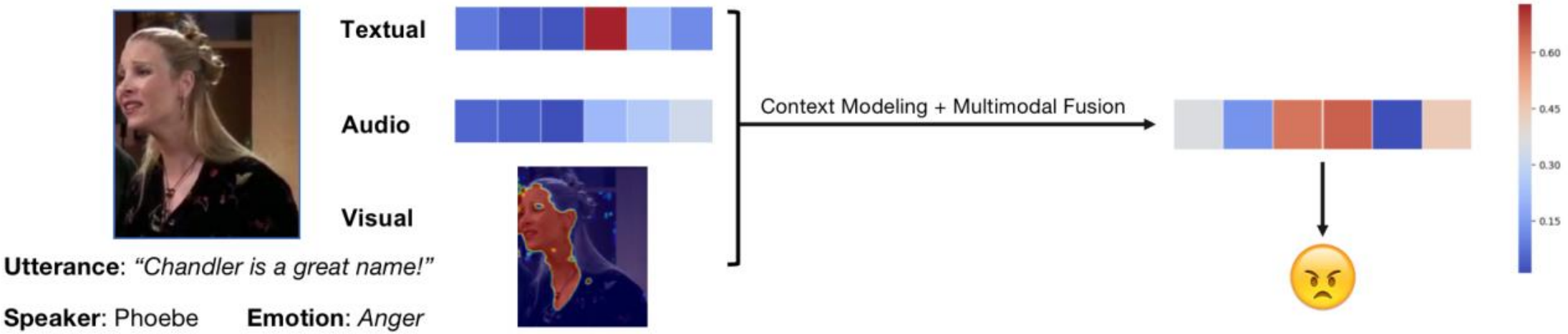



Figure 6: Visualization of the heatmaps of a prone-to-misclassification utterance in MELD.

# Thanks!