# Self-supervised Cross-modal Pretraining for Speech Emotion and Sentiment Analysis

**Iek-Heng Chu[1]*, Ziyi Chen[1]*, Xinlu Yu[1], Mei Han[1], Jing Xiao[2] and Peng Chang[1]**

[1]PAII Inc., Palo Alto, USA
[2]Ping An Technology, Shenzhen, China
{zhuyixing276,chenziyi253,yuxinlu698,hanmei613,changpeng805}@paii-labs.com
xiaojing661@pingan.com.cn

2023.3.30 • ChongQing

—— EMNLP 2022

**Reported by JiaWei Cheng**

# Motivation

The first challenge is the limited availability of annotated data.

The second challenge is the learning of a multimodal feature space that can well distinguish among different emotions or sentiments, especially in the case of multimodal modeling.

# Overview

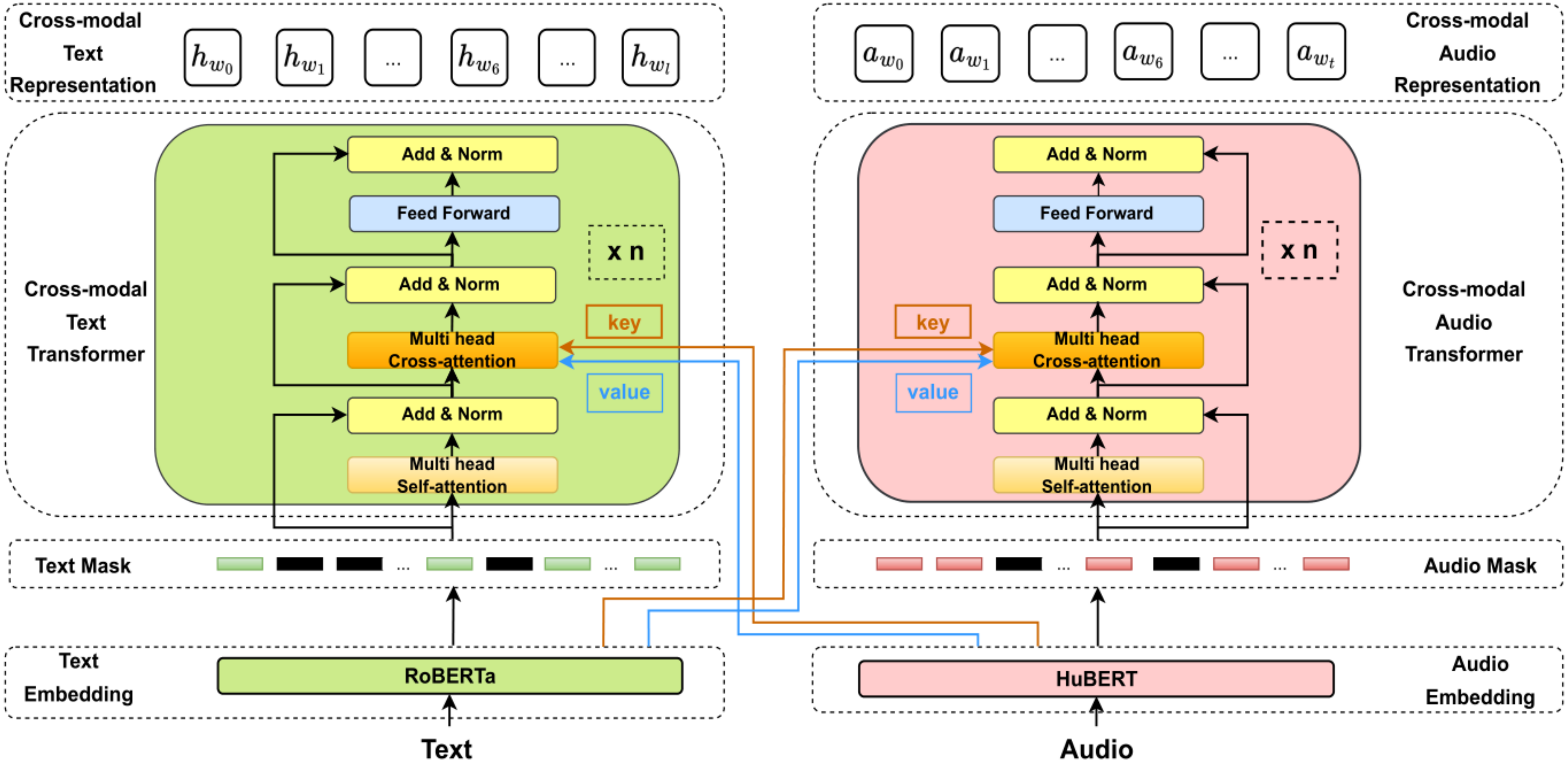
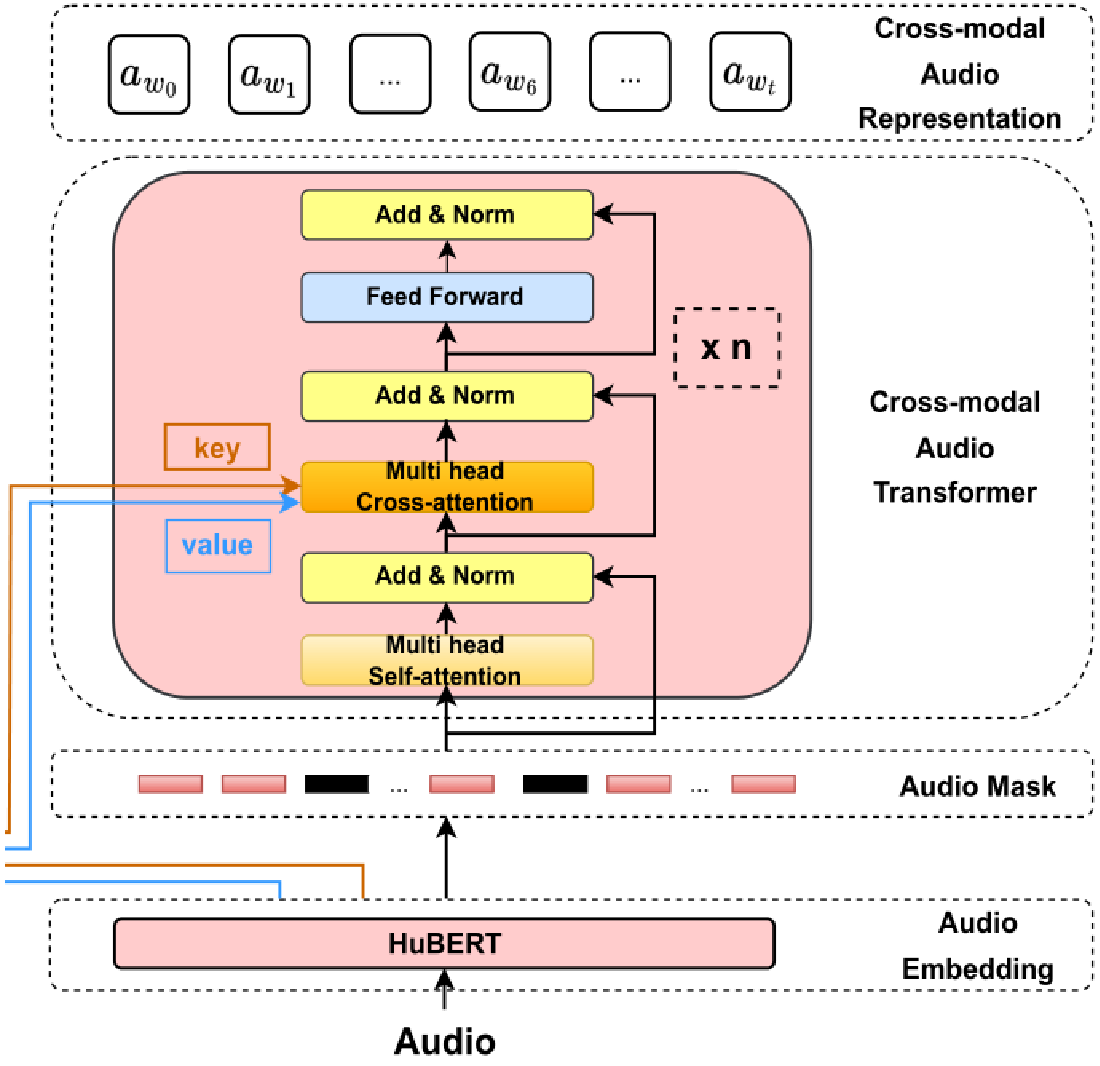



Figure 1: The architecture of the proposed cross-modal pretrained transformer model.

# Method




$$\acute{h}_a^{[l+1]} = Attn(Q = h_a^{[l]}, K = h_a^{[l]}, V = h_a^{[l]}), \quad (1)$$

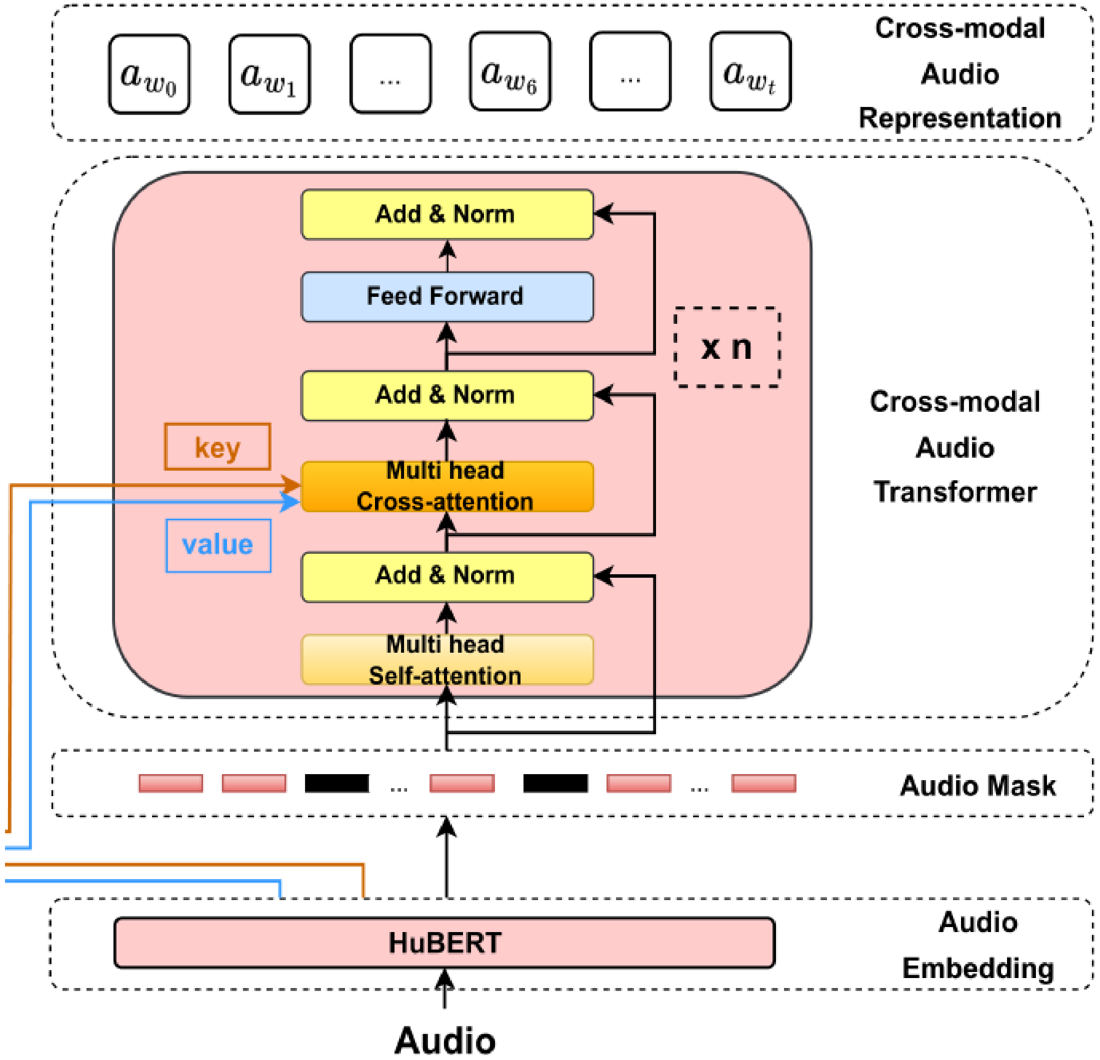
$$\dot{h}_a^{[l+1]} = LN(h_a^{[l]} + \acute{h}_a^{[l+1]}) \quad (2)$$

$$\ddot{h}_a^{[l+1]} = Attn(Q = \dot{h}_a^{[l+1]}, K = e_w, V = e_w) \quad (3)$$

$$\tilde{h}_a^{[l+1]} = LN(\ddot{h}_a^{[l+1]} + \dot{h}_a^{[l+1]}) \quad (4)$$

$$h_a^{[l+1]} = LN(FFN(\tilde{h}_a^{[l+1]}) + \tilde{h}_a^{[l+1]}) \quad (5)$$

# Method




$$h_{\text{fuse}} = \bar{h}_a \oplus h_{w_0}, \quad (6)$$

$$L_{\text{ortho}} = \frac{|\bar{h}_a^{\mathrm{T}} \cdot \bar{h}_w|}{\|\bar{h}_a\| \cdot \|\bar{h}_w\|}. \quad (7)$$

$$L_{\text{FT}} = L_{\text{task}} + \alpha \cdot L_{\text{ortho}}, \quad (8)$$

# Experiments

| Methods | Angry ↑ | Happy ↑ | Neutral ↑ | Sad ↑ | WA ↑ | UA ↑ |
|---|---|---|---|---|---|---|
| MulT (Tsai et al., 2019) | 0.739 | 0.848 | 0.625 | 0.777 | - | - |
| JBLS (Siriwardhana et al., 2020) | **0.920** | 0.870 | 0.809 | 0.908 | - | 0.734 |
| CTAL (Li et al., 2021) | - | - | - | - | 0.740 | 0.746 |
| HuBERT | 0.908 | 0.825 | 0.785 | 0.885 | 0.703 | 0.711 |
| RoBERTa | 0.902 | 0.850 | 0.782 | 0.869 | 0.702 | 0.709 |
| Shallow-Fusion | 0.901 | 0.849 | 0.789 | 0.895 | 0.717 | 0.728 |
| CMT BASE | 0.907 | 0.869 | 0.815 | 0.912 | **0.751** | **0.763** |
| CMT LARGE | 0.898 | **0.872** | **0.817** | **0.913** | 0.750 | 0.761 |

Table 1: Main experimental results on IEMOCAP emotion dataset, where emotion-wise (angry/happy/neutral/sad) binary accuracy, weighted accuracy (WA) and unweighted accuracy (UA) are presented.

# Experiments

| Methods | $Acc_7$ ↑ | $Acc_2$ ↑ | F1-score ↑ | MAE ↓ |
|---|---|---|---|---|
| MulT (Tsai et al., 2019) | 0.507 | 0.816 | 0.816 | 0.591 |
| JBLS (Siriwardhana et al., 2020) | 0.521 | 0.878 | - | 0.518 |
| CTAL (Li et al., 2021) | - | 0.808 | 0.810 | 0.603 |
| HuBERT | 0.486 | 0.796 | 0.799 | 0.634 |
| RoBERTa | 0.521 | 0.876 | 0.877 | 0.523 |
| Shallow-Fusion | 0.538 | 0.861 | 0.860 | 0.518 |
| CMT BASE | **0.546** | 0.880 | 0.878 | 0.501 |
| CMT LARGE | 0.545 | **0.885** | **0.885** | **0.500** |

Table 2: Main experimental results on CMU-MOSEI sentiment dataset, where 7-class accuracy ($Acc_7$), 2-class accuracy ($Acc_2$), F1 score, and mean absolute error (MAE) are presented.

# Experiments

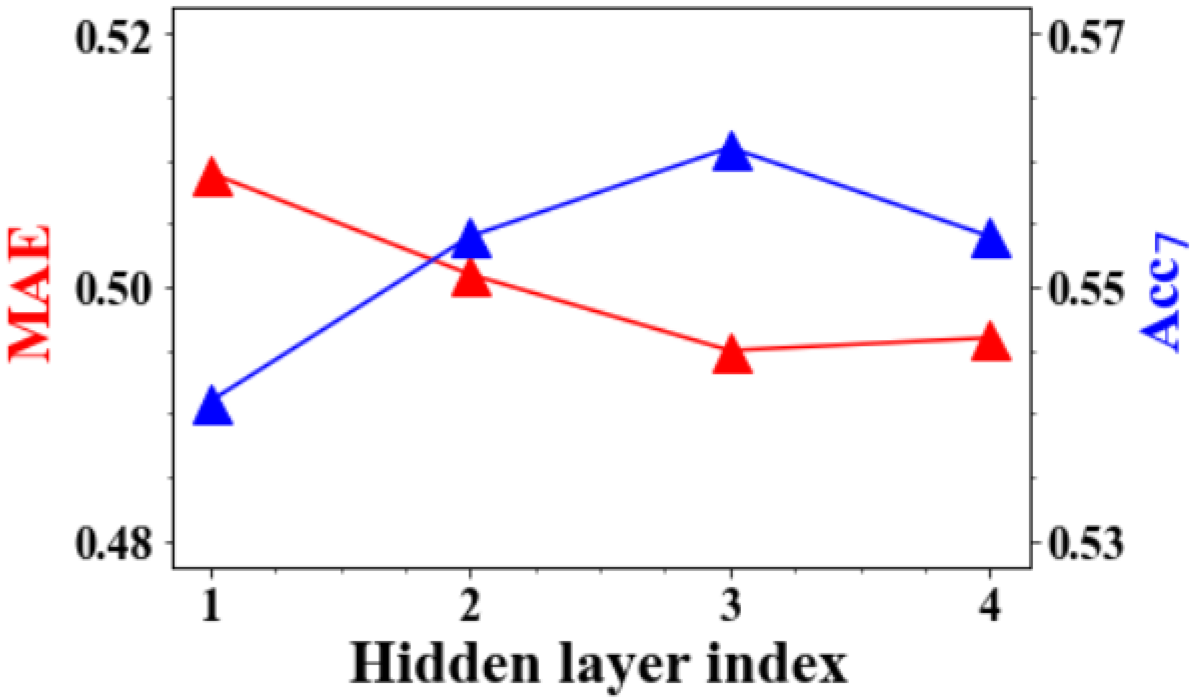



Figure 2: Impact of using different hidden states as the CMT-4 audio transformer representation on CMU-MOSEI metrics MAE (red) and $Acc_7$ (blue).

# Experiments

| Layer index | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| Weight | 0.175 | 0.438 | 0.277 | 0.109 |

Table 3: Learned weights of the layer pooler associated with different audio cross-modal transformer layers in the pretrained CMT-4 model.

# Experiments

| Methods | $Acc_7 \uparrow$ | $Acc_2 \uparrow$ | F1-score $\uparrow$ | MAE $\downarrow$ |
|---|---|---|---|---|
| CMT-4 w/o PT | 0.545 | 0.874 | 0.875 | 0.508 |
| CMT-4 PT | 0.554 | 0.870 | 0.871 | 0.496 |
| CMT-4 PT + TAPT | **0.559** | 0.866 | 0.869 | 0.502 |
| CMT-4 PT + Layer pooler | 0.545 | 0.864 | 0.863 | 0.509 |
| CMT-4 PT + Ortho | 0.554 | 0.879 | 0.878 | **0.493** |
| CMT BASE | 0.546 | 0.880 | 0.878 | 0.501 |
| CMT LARGE | 0.545 | **0.885** | **0.885** | 0.500 |

Table 4: The ablation analysis of our proposed CMT-4 model using CMU-MOSEI. The terms PT, TAPT, Layer pooler and Ortho refer to pretrained, task adaptive pretraining, weighted average layer of cross-modal audio transformer hidden states, and the orthogonality regularization term.

# Experiments

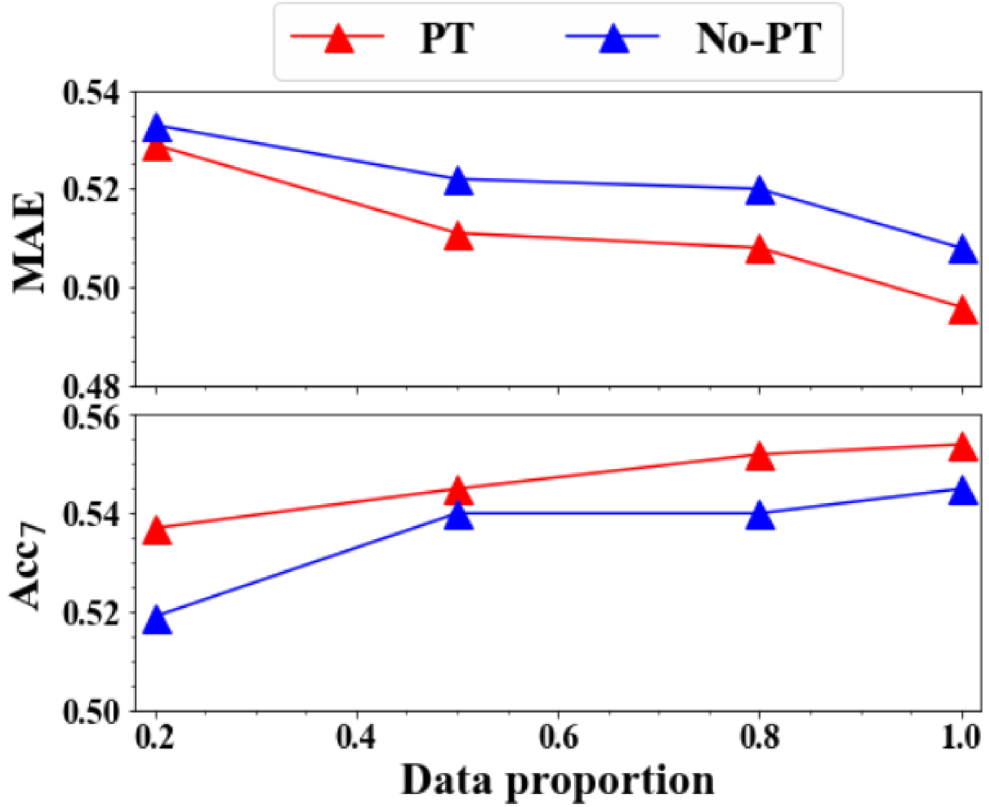



Figure 3: The performance of pretrained (PT) and non-pretrained (No-PT) CMT-4 models with different proportions of CMU-MOSEI training set.

# Thanks!