Multimodal Transformer Distillation for Audio-Visual Synchronization

Xuanjun Chen Haibin Wu Chung-Che Wang Hung-yi Lee Jyh-Shing Roger Jang

National Taiwan University

Introduction

Audio-Visual Synchronization (AVS)

- **Goal:** Determine whether the mouth and speech are synchronized
- **VocaLiST:** A SOTA model as shown in the teacher model in Figure 1
- **Applications:** Most audio-visual applications, such as dubbing
- **Challenges:** Require high computing resources

Contributions

- Proposed an MTDVocaLiST model, which is trained by our proposed Multimodal Transformer Distillation (MTD) loss
- MTD encourages MTDVocaLiST to mimic the cross-attention distribution and value-relation of VocaLiST deeply
- MTDVocaLiST outperforms similar-size models, reducing VocaLiST's size by 83.52% while maintaining similar performance

MTDVocaLiST

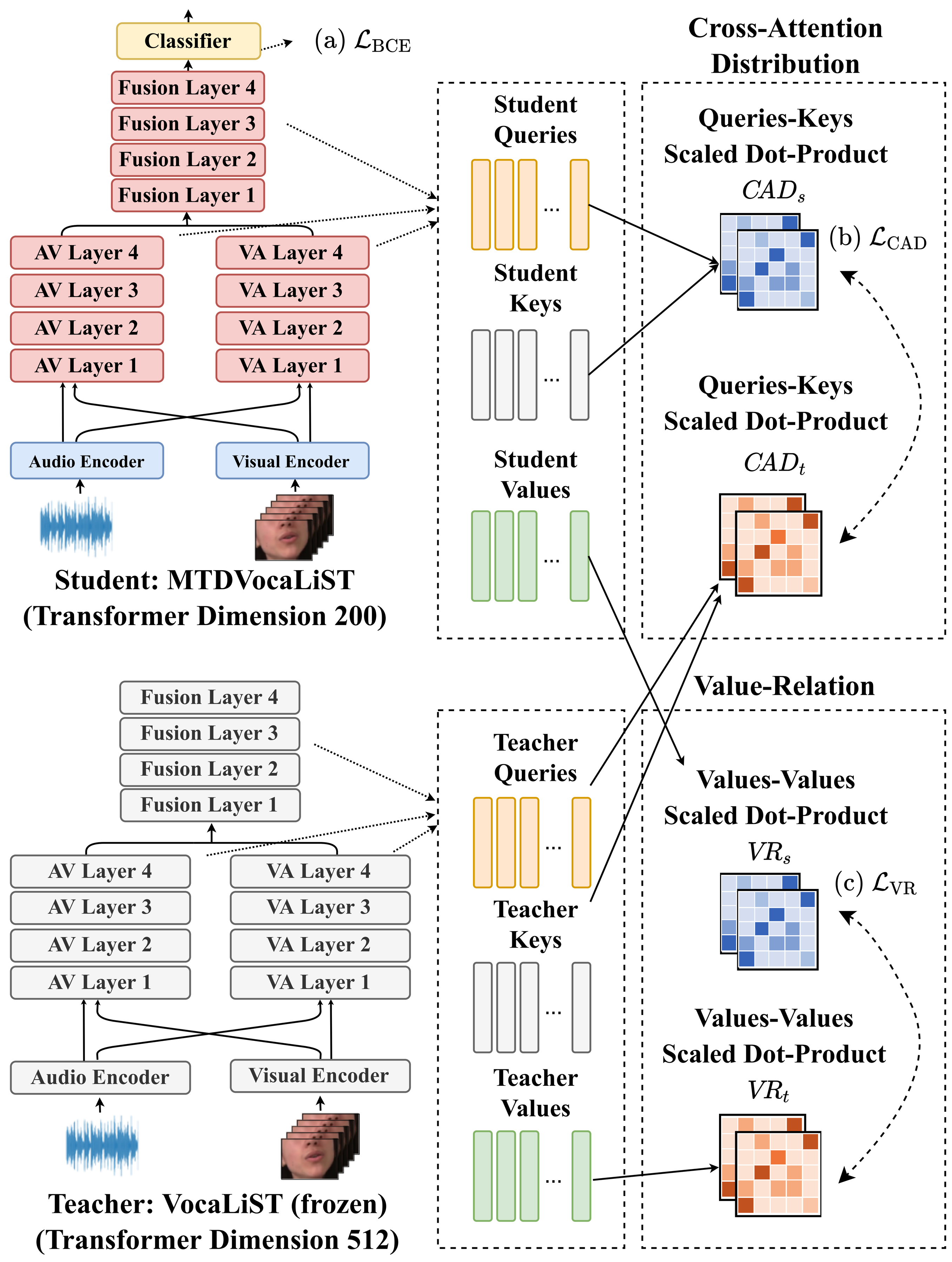


Figure 1. The proposed MTDVocaLiST model. (a) binary cross entropy loss. (b) cross-attention distribution distillation loss. (c) value-relation distillation loss.

Naïve Multimodal Transformer Distillation (NMTD)

$$\mathcal{L}_{NMTD} = w_0 \cdot \mathcal{L}_{BCE} + \sum_{l}^{L} w_{l1} \cdot \mathcal{L}_{CAD_l} + \sum_{l}^{L} w_{l2} \cdot \mathcal{L}_{VR_l}, \quad (1)$$

- w_0, w_{l1}, and w_{l2} represent the weights for $\mathcal{L}_{BCE}$, $\mathcal{L}_{CAD_l}$, and $\mathcal{L}_{VR_l}$
- L denotes a candidate layer set, l-th is the sub-layer in the set

Multimodal Transformer Distillation (MTD)

- After utilizing uncertainty weighting [1], overall MTD is as follows:

$$\mathcal{L}_{MTD} = \sum_{\tau}^{T} \left(\frac{1}{2 \cdot w_\tau^2} \cdot \mathcal{L}_\tau + ln(1 + w_\tau^2) \right), \quad (2)$$

- T represents a task set
- $\mathcal{L}_\tau$ denotes the τ-th loss, which could be $\mathcal{L}_{BCE}$, $\mathcal{L}_{CAD}$ or $\mathcal{L}_{VR}$ loss
- w_τ are learnable parameters. $ln(1 + w_\tau^2)$ serves to enforce positive regularization values

Experiment setup

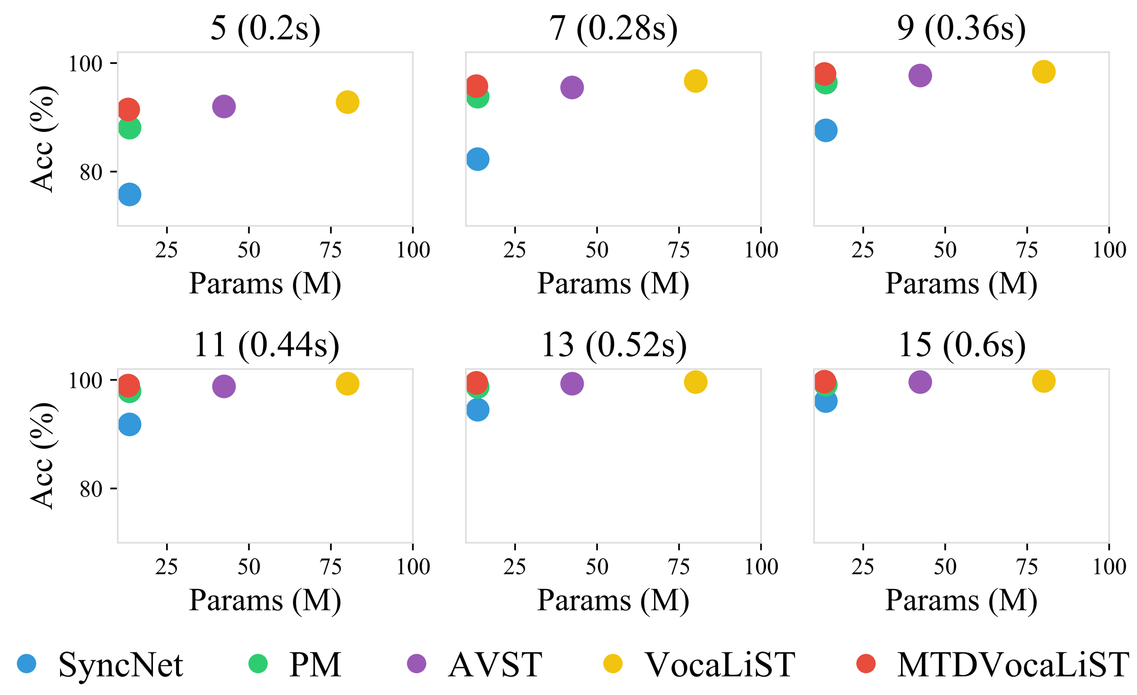
- **Dataset:** Lip Reading Sentences 2 (LRS2) dataset
- **Training:** Positive and negative samples are sampled on the fly
- **Evaluation protocol:** Accuracy of the cross-modal retrieval task

Main results

Table 1. Accuracy of different distillation methods in evaluation.

Distillation method	Input frame length (seconds)					
	5 (0.2s)	7 (0.28s)	9 (0.36s)	11 (0.44s)	13 (0.52s)	15 (0.6s)
$\mathcal{L}_{BCE}$	71.36	81.44	88.84	93.41	96.19	97.69
KD	80.87	88.62	93.48	96.32	97.90	98.82
RKD	86.06	92.42	95.95	97.80	98.75	99.29
MiniLM*	85.60	92.03	95.91	97.72	98.72	99.25
FitNets	90.81	95.48	97.77	98.81	99.42	99.66
MTD	**91.45**	**95.75**	**97.99**	**98.95**	**99.46**	**99.68**

Figure 2. Comparison of model size and accuracy.



Comparison with Different Distillation Methods (Table 1)

- **Length 5:** $\mathcal{L}_{BCE}$ results in the lowest accuracy at 71.36%.
- **Length 5:** MTD significantly improves accuracy, surpassing KD by 10.58%, RKD by 5.39%, MiniLM* by 5.85%, and FitNets by 0.64%.
- Similar trends are observed across different input frame lengths.

Comparison with SOTA models (Figure 2)

- MTDVocaLiST outperforms similar-size SOTA models, SyncNet, and Perfect Match models by 15.65% and 3.35%;
- MTDVocaLiST reduces the model size of VocaLiST by 83.52%, yet still maintaining similar performance.

Ablation study and analysis

Figure 3. Ablation study of NMTD loss.

Loss	Val F1 (%)	Eval Acc (%)
$\mathcal{L}_{BCE}$	87.91	71.36
NMTD w/o $\mathcal{L}_{VR}$	91.78	83.55
NMTD w/o $\mathcal{L}_{CAD}$	91.97	83.53
NMTD	**92.81**	**85.60**

Figure 4. *Different layer selection strategies.*

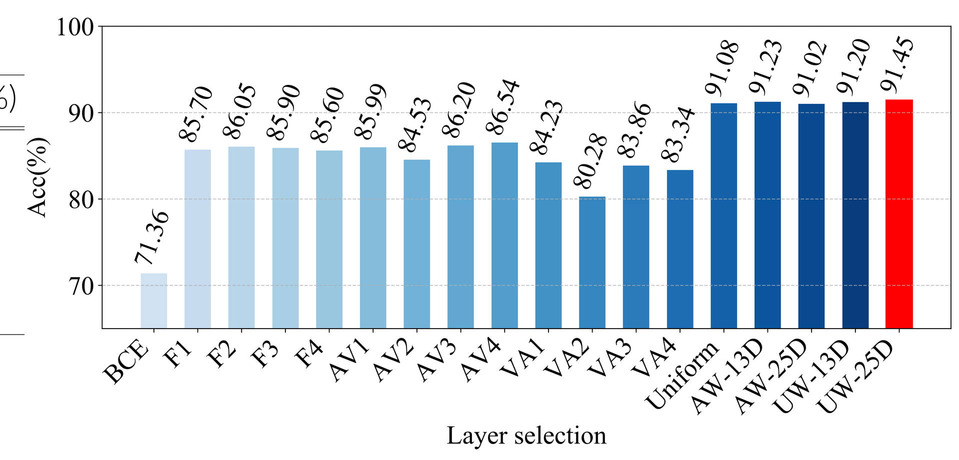
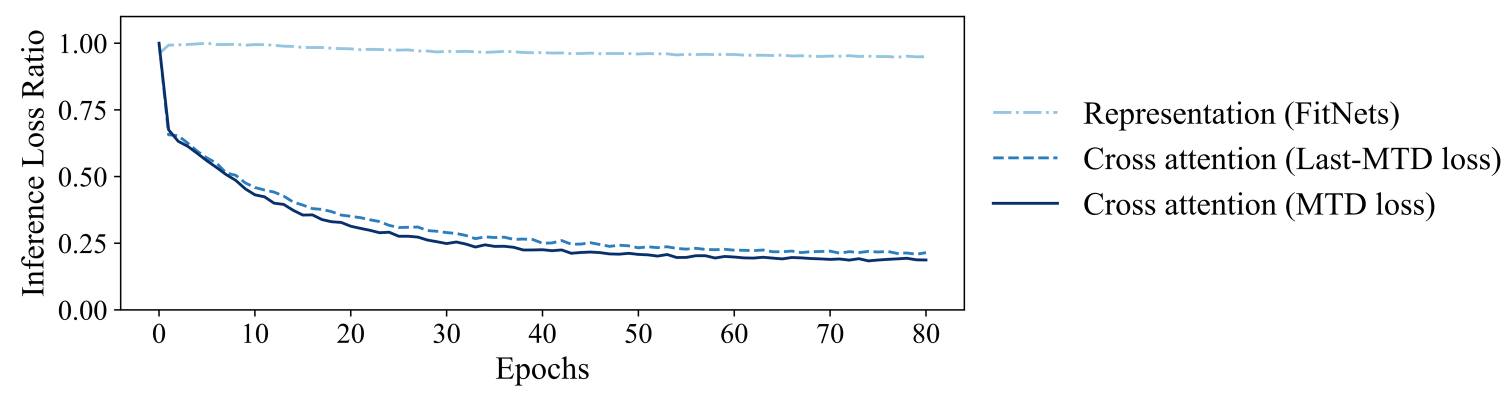


Figure 5. Comparison of Transformer representation and cross-attention loss in **inference**. Note that the MTDVocaLiST only optimizes the MTD loss during **training**.



Indispensability (Figure 3): Both cross-attention distribution and value-relation contribute significantly to NMTD loss

Layer selection (Figure 4)

- Distilling any Transformer layer significantly improves performance.
- VA layers contribute minimally to the student's final performance.
- Single-layer distillation and BCE training perform worse.
- UW-25D layer weighting outperforms Uniform, AW and UW-13D

Transformer behavior and Transformer representation (Figure 5)

- The Transformer representation loss will not decrease along with the cross attention loss in the inference phase of MTDVocaLiST

References

[1] Kendall et al., "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," arXiv:1705.07115, 2017.