2024「中技社科技獎學金」
2024 CTCI Foundation Science and Technology Scholarship
境外生生活助學金
Living Grant for International Graduate Students

Multimodal Transformer Distillation for Audio-Visual Synchronization

Ph.D. Student: Xuanjun Chen **Advisors:** 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

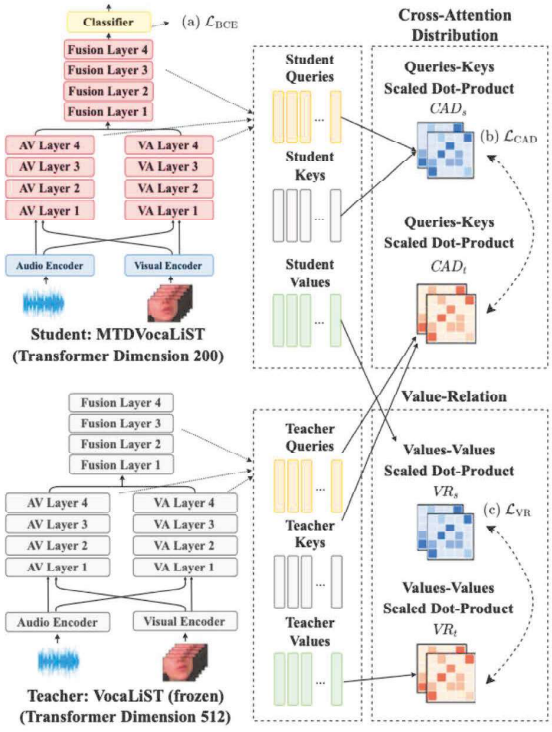


Figure 1. The proposed MTDVocaLiST model. (a) binary cross entropy loss. (b) cross-attention distribution distillation loss. (c) value-relation distillation loss.

Experiment setup

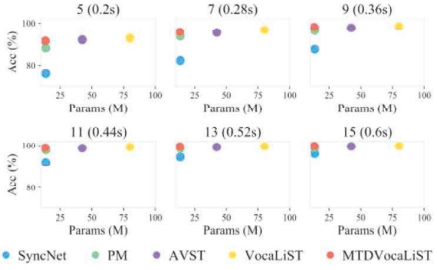
- **Dataset:** Lip Reading Sentences 2 (LRS2) dataset
- **Training:** Positive and negative samples are sampled on the fly
- **Evaluation protocol:** Accuracy of the cross-modal retrieval task

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.



Ablation study and analysis

Figure 3. Ablation study of NMTD loss.

Loss	Val Γ1 (%)	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.*

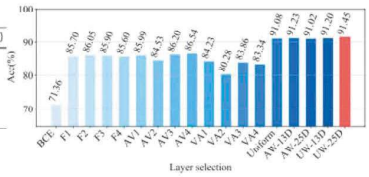
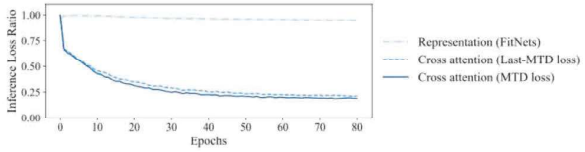


Figure 5. Comparison of Transformer representation and cross-attention loss in **inference**. Note that the MTDVocaLiST only optimizes the MTD loss during **training**.



Xuanjun's Homepage: www.xjchen.tech
CTCI Foundation Bursary Award for Overseas Students
Email: D12942018@ntu.edu.tw

財團法人中技社 CTCI FOUNDATION