ModEFormer: Modality-Preserving Embedding for Audio-Video Synchronization using Transformers

ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

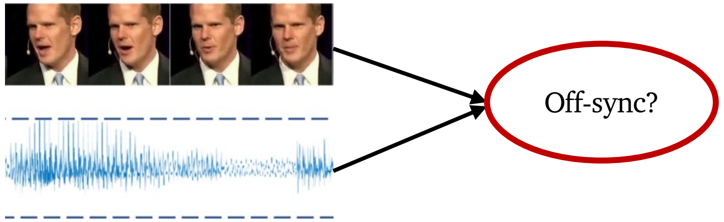
Akash Gupta
New York University
aksg@nyu.edu

Rohun Tripathi
Amazon Studios
rt443@cornell.edu

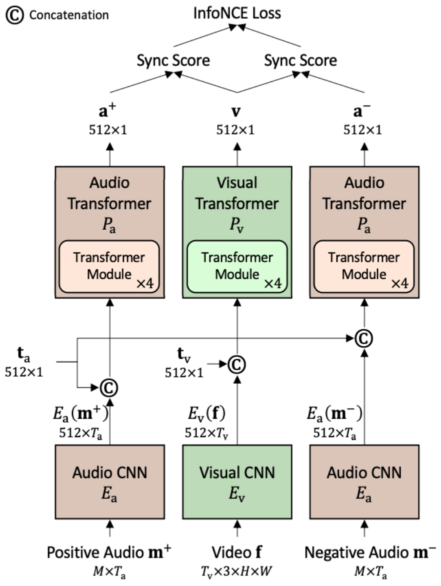
Wondong Jang
Amazon Studios
dotoli216@gmail.com

Introduction

- Our task is to identify audio-video off-sync errors that often occur in TV broadcasts or video conferencing leading to poor viewing experience

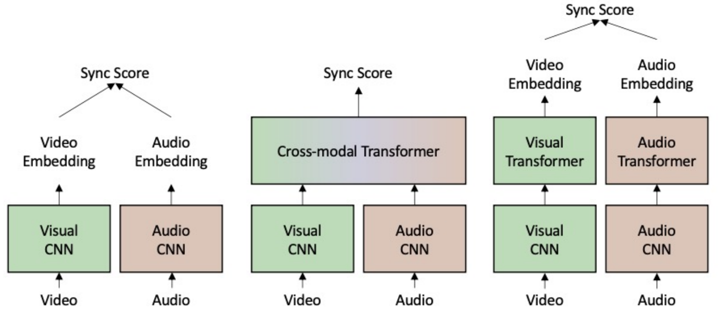


Model Architecture



- We propose **ModEFormer** which is an automated transformer-based detection technique to identify these errors and provide audio-video synchronization.
- **ModEFormer** has separate encoders for audio and video modalities and extracts the corresponding embeddings
- The embeddings are used to calculate a sync score to be used in the InfoNCE loss function for contrastive learning

Comparison with previous approaches



(a) SyncNet / PM (b) AVST / VocaLiST (c) ModEFormer - Ours

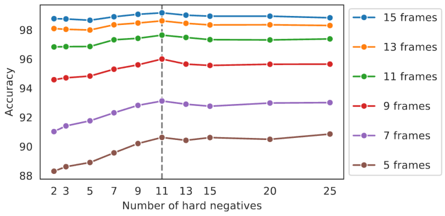
- Unlike previous approaches, **ModEFormer** ensure no mixing between modalities at any step.
- The proposed modality-specific embedding architecture provides the advantage of using large batch size with abundant negative samples useful in contrastive learning.

Results

Dataset	Model	Var	Clip Length in Frames (Seconds) 5 (0.2s)	7 (0.28s)	9 (0.36s)	11 (0.44s)	13 (0.52s)	15 (0.6s)	# of params (M=Millions)
LRS2	AVST[3]	✓	91.9	97.0	98.8	99.6	99.8	99.9	42.4M
	SyncNet[1]		75.8	82.3	87.6	91.8	94.5	96.1	13.6M
	PM[2]		88.1	93.8	96.4	97.9	98.7	99.1	13.6M
	VocaLiST[4]		92.8	96.7	98.4	**99.3**	99.6	**99.8**	80.1M
	ModEFormer - Ours		**94.5**	**97.1**	**98.5**	**99.3**	**99.7**	**99.8**	59.0M
LRS3	AVST[3]	✓	77.3	88.0	93.3	96.4	97.8	98.6	42.4M
	ModEFormer - Ours		**90.9**	**93.1**	**96.0**	**97.7**	**98.7**	**99.2**	59.0M

- **ModEFormer** outperforms all the other approaches using a fixed number of input frames.
- The significant increase is due to the modality-preserving architecture and novel sampling strategy involving multiple hard negatives during training.

Ablation Study

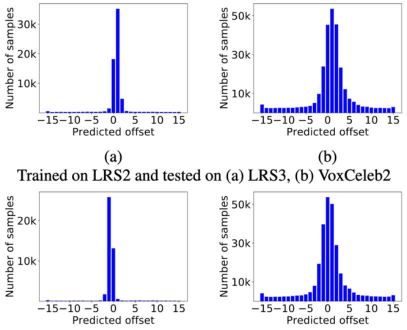


- We do further analysis to find the optimal number of negatives (N) in a training batch and observe highest accuracy at N=11.
- We ablate transformers and infer that it leads to reduction in accuracy of 8.1%

Table 2. Results of 3D-SyncNet and ModEFormer on LRS3 test set

	3D-SyncNet	ModEFormer (1st stage)	ModEFormer (2nd stage)
Accuracy	80.2%	88.3%	90.9%

Application – Offset Detection



(a) (b)
Trained on LRS2 and tested on (a) LRS3, (b) VoxCeleb2

(c) (d)
Trained on LRS3 and tested on (c) LRS2, (d) VoxCeleb2

- We apply a pretrained **ModEFormer** to detect any audio-video lag in a given test clip by measuring the offset from a cosine-similarity histogram.
- We found that LRS2 and LRS3 datasets are out-of-sync by 1 frame.

Conclusions

- We present **ModEFormer**, a modality-preserving embedding architecture for audio-video synchronization
- The proposed architecture and negative sampling strategy gives state-of-the-art performance on lip-reading datasets and benefits from large batch sizes used in contrastive learning.