Self-Supervised Audio-Visual Representation Learning with Relaxed Cross-Modal Synchronicity

Pritam Sarkar[1, 2] Ali Etemad[1]

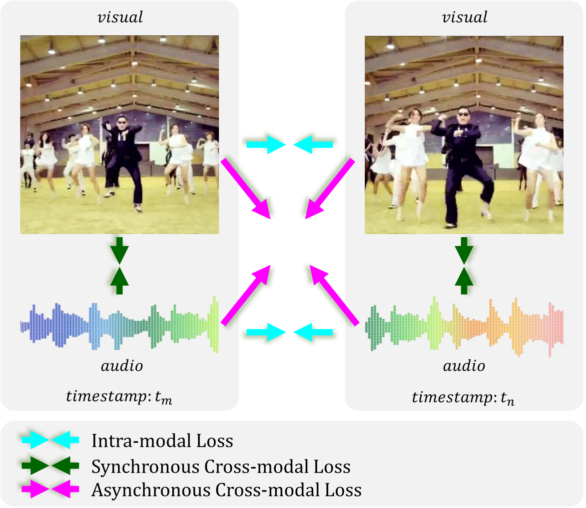
[1] Queen's University, Canada [2] Vector Institute

https://pritamsarkar.com/CrissCross

Overview

The common and standard practice in self-supervised audio-visual representations learning is to learn intra-modal and synchronous cross-modal relationships between the audio and visual streams maintaining a strict frame-wise coupling.

Our **intuition** is that the *temporal synchronicity between audio and visual segments can be relaxed to some extent* to learn more robust representations.



Proposed Framework

We obtain two augmented views of $v=\{v_t\}_{t=0}^{T}$, denoted by v_1 and v_2, defined as $\{v_t\}_{t=t_1}^{t_1+t_v}$ and $\{v_t\}_{t=t_2}^{t_2+t_v}$ respectively. Similarly, two augmented views of $a=\{a_t\}_{t=0}^{T}$ can be obtained as a_1 and a_2 as $\{a_t\}_{t=t_1}^{t_1+t_a}$ and $\{a_t\}_{t=t_2}^{t_2+t_a}$, respectively.
We calculate cosine distance of two embeddings as $\mathcal{D}(p,z)=-\frac{p}{||p||_2}\cdot\frac{z}{||z||_2}$, where p and z are obtained as $h(f(x_1))$ and $\mathtt{S}(f(x_2))$. Here, `predictor` head is denoted by h, `stop-gradient` is denoted by $\mathtt{S}$, f denotes the feature encoder, and augmented views are denoted by x_1 and x_2 which are t seconds apart. The final training objective $\mathcal{L}_{CrissCross}$ calculated as:

$$\mathcal{L}_{intra}=\underbrace{\frac{1}{2}\mathcal{D}(p_{v_1},\mathtt{S}(z_{v_2}))+\frac{1}{2}\mathcal{D}(p_{v_2},\mathtt{S}(z_{v_1}))}_{\mathcal{L}_{v_1,v_2}}+\underbrace{\frac{1}{2}\mathcal{D}(p_{a_1},\mathtt{S}(z_{a_2}))+\frac{1}{2}\mathcal{D}(p_{a_2},\mathtt{S}(z_{a_1}))}_{\mathcal{L}_{a_1,a_2}}$$

$$\mathcal{L}_{sync}=\underbrace{\frac{1}{2}\mathcal{D}(p_{v_1},\mathtt{S}(z_{a_1}))+\frac{1}{2}\mathcal{D}(p_{a_1},\mathtt{S}(z_{v_1}))}_{\mathcal{L}_{a_1,v_1}}+\underbrace{\frac{1}{2}\mathcal{D}(p_{v_2},\mathtt{S}(z_{a_2}))+\frac{1}{2}\mathcal{D}(p_{a_2},\mathtt{S}(z_{v_2}))}_{\mathcal{L}_{a_2,v_2}}$$

$$\mathcal{L}_{async}=\underbrace{\frac{1}{2}\mathcal{D}(p_{v_1},\mathtt{S}(z_{a_2}))+\frac{1}{2}\mathcal{D}(p_{a_2},\mathtt{S}(z_{v_1}))}_{\mathcal{L}_{v_1,a_2}}+\underbrace{\frac{1}{2}\mathcal{D}(p_{v_2},\mathtt{S}(z_{a_1}))+\frac{1}{2}\mathcal{D}(p_{a_1},\mathtt{S}(z_{v_2}))}_{\mathcal{L}_{a_1,v_2}}$$

$$\mathcal{L}_{CrissCross}=\frac{1}{3}(\mathcal{L}_{intra}+\mathcal{L}_{sync}+\mathcal{L}_{async})$$

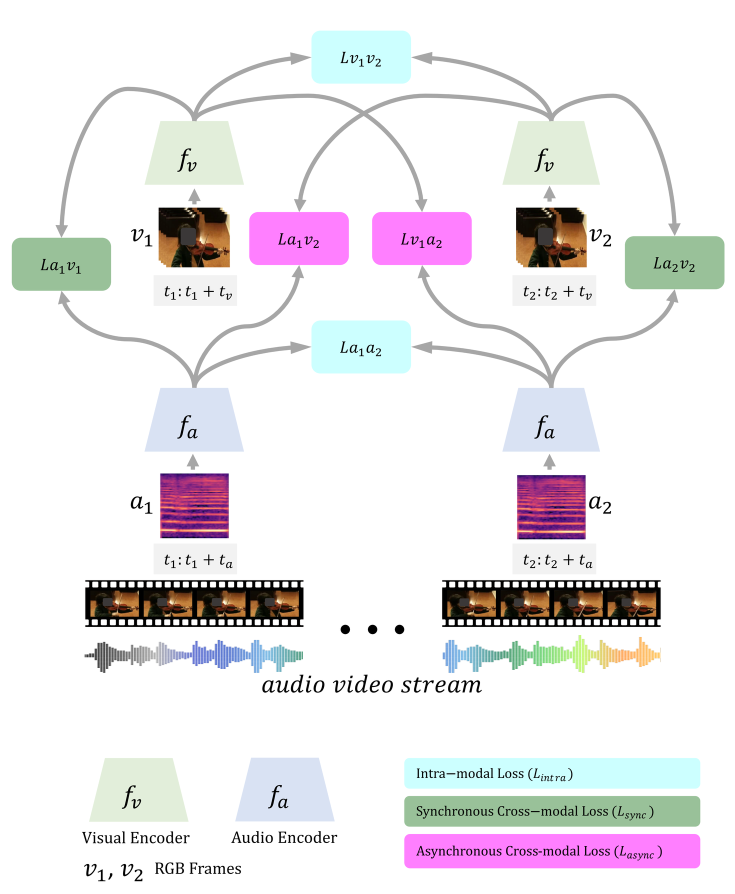


Fig. 1: Our proposed framework.

Temporal Relaxation

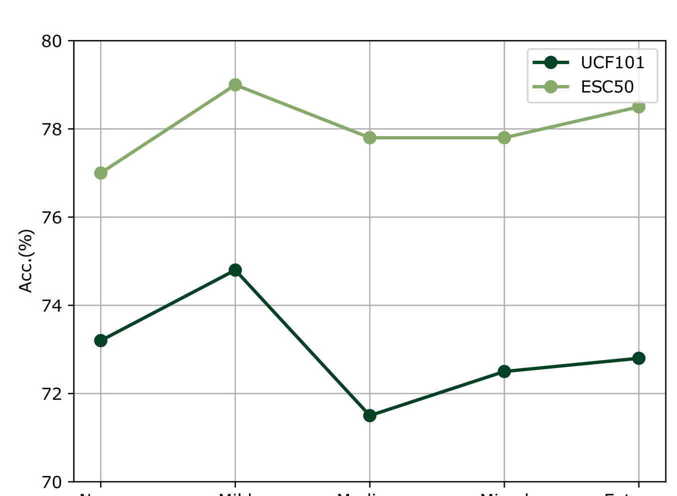


Fig. 2: Exploring temporal relaxation.

None: both the audio and visual segments are sampled from the exact same timestamp.
Mild: the two views of the audio-visual segments share 50% overlap amongst them.
Medium: the adjacent frame sequences and audio segments are sampled.
Mixed: the two audio-visual segments are sampled in a temporally random manner.
Extreme: one view is sampled from the first half of the source clip, and the other view is sampled from the second half of the source clip.

Effect of Learning Asynchronous Cross-modal Relations

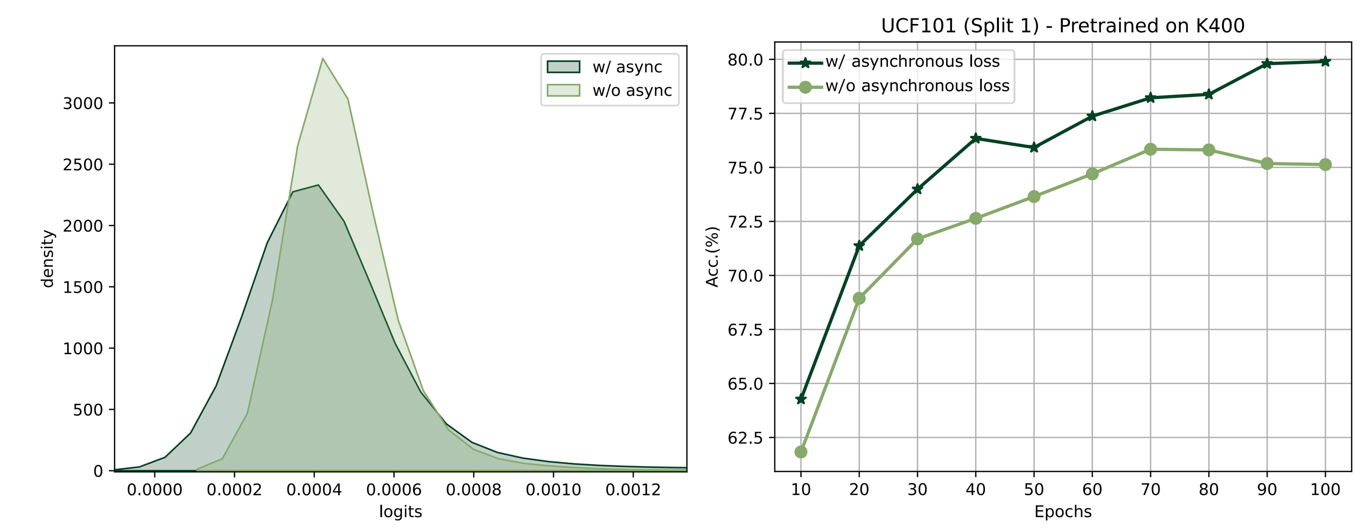


Fig. 3: Left: Distribution of the learned representations; Right: Linear eval. top-1 acc. vs. pretraining epochs; w/ and w/o the asynchronous cross-modal optimization.

Pretrain	Downstream	w/o $\mathcal{L}_{async}$	w/ $\mathcal{L}_{async}$
Kinetics400	UCF101	75.8(↓ 4.1)	79.9
Kinetics400	ESC50	78.5(↓ 3.5)	82.0
Kinetics400	Kinetics-Sound (a)	43.2(↓ 3.9)	47.1
Kinetics400	Kinetics-Sound (v)	53.3(↓ 2.4)	55.7
Kinetics400	Kinetics-Sound (a+v)	65.0(↓ 1.7)	66.7

Tab. 1: Impact of $\mathcal{L}_{async}$ optimization in different pretraining and evaluation setups.

Fig. 4: Visualization of saliency maps while pretrained without (left) and with (right) asynchronous loss.

	Pretraining Dataset		
	Kinetics-Sound (22K)	Kinetics400 (240K)	AudioSet (1.8M)
HMDB51	45.7	50.0	56.2
UCF101	78.1	83.9	87.7
Kinetics400	39.0	44.5	50.1
ESC50	82.8	86.8	90.5
DCASE	93.0	96.0	97.0

Tab. 2: The top-1 acc. of linear evaluation on action recognition and sound classification with varying sizes of pretraining data.

Results

Method	Pretraining Compute	Pretrained Dataset	Backbone (#Params (M))	Finetune #frames	UCF101	HMDB51
CM-ACC	40 GPUs	Kinetics-Sound	3D-ResNet18 (33.4)	32	77.2	40.6
CrissCross	**4 GPUs**	Kinetics-Sound	R(2+1)D-18 (**15.4**)	32	**88.3**	**60.5**
Supervised	-	Kinetics-Sound	3D-ResNet18 (33.4)	32	86.9	53.1
XDC	64 GPUs	Kinetics400	R(2+1)D-18 (31.5)	8	74.2	39.0
AVID	64 GPUs	Kinetics400	R(2+1)D-18 (15.4)	8	83.7	49.5
CrissCross	**8 GPUs**	Kinetics400	R(2+1)D-18 (**15.4**)	8	**86.9**	**54.3**
XDC	64 GPUs	Kinetics400	R(2+1)D-18 (31.5)	32	86.8	52.6
AVID	64 GPUs	Kinetics400	R(2+1)D-18 (15.4)	32	87.5	60.8
CrissCross	**8 GPUs**	Kinetics400	R(2+1)D-18 (**15.4**)	32	**91.5**	**64.7**
Supervised	-	Kinetics400	R(2+1)D-18 (31.5)	32	95.0	74.0
XDC	64 GPUs	AudioSet	R(2+1)D-18 (31.5)	8	84.9	48.8
AVID	64 GPUs	AudioSet	R(2+1)D-18 (15.4)	8	88.6	57.6
CrissCross	**8 GPUs**	AudioSet	R(2+1)D-18 (**15.4**)	8	**89.4**	**58.3**
XDC	64 GPUs	AudioSet	R(2+1)D-18 (31.5)	32	**93.0**	63.7
AVID	64 GPUs	AudioSet	R(2+1)D-18 (15.4)	32	91.5	64.7
CrissCross	**8 GPUs**	AudioSet	R(2+1)D-18 (**15.4**)	32	92.4	**67.4**
Supervised	-	AudioSet	R(2+1)D-18 (31.5)	32	96.8	75.9

Tab. 3: SOTA comparison on action recognition.

Method	ESC50		DCASE	
	Kinetics400	AudioSet	Kinetics400	AudioSet
XDC	78.0	84.8	91	95
AVID	79.1	89.1	93	96
CrissCross	**86.8**	**90.5**	**96**	**97**

Tab. 4: SOTA comparison on sound classification.

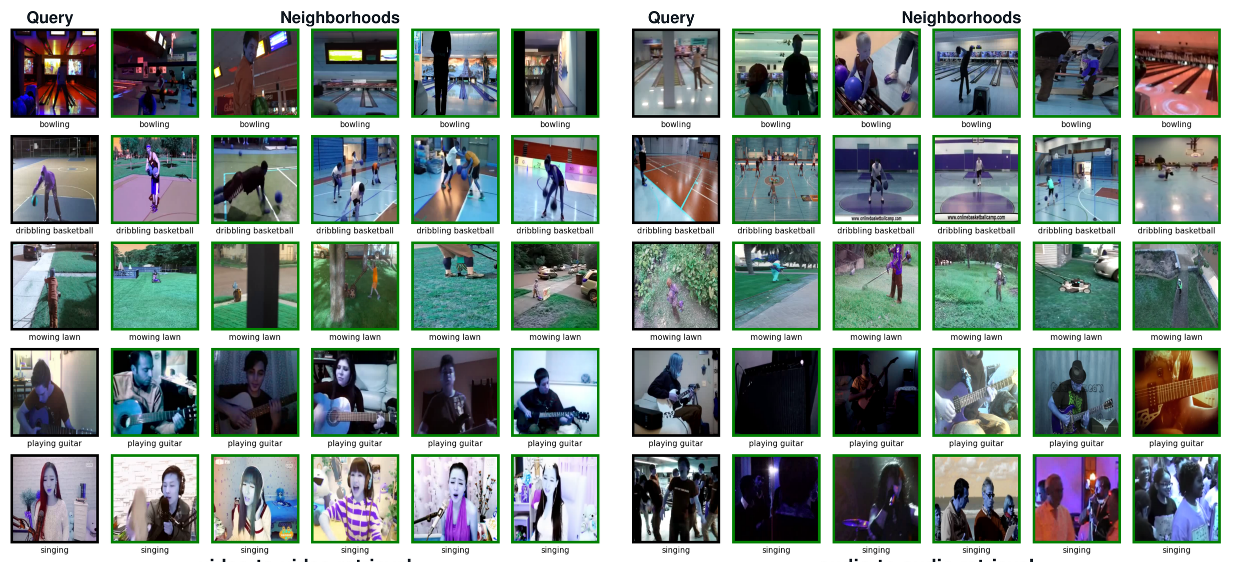


Fig. 5: We present a few randomly selected samples of video-to-video (left) and audio-to-audio (right) retrieval.

Summary

- We propose a novel self-supervised framework CrissCross to learn audio-visual representations by exploiting intra-modal, as well as, synchronous and *asynchronous* cross-modal relationships. Our findings show that the relaxation of cross-modal temporal synchronicity to some extent helps in learning more generalized representations which results in better downstream performance.
- Our experiments show that CrissCross either outperforms or achieves performances on par with the current state-of-the-art self-supervised methods on action recognition and retrieval on UCF101 and HMDB51, as well as sound classification on ESC50 and DCASE.

Acknowledgement

We are grateful to the Bank of Montreal and Mitacs for funding this research. We are also thankful to SciNet HPC Consortium for helping with the computation resources.