Supplementary Materials

CoLeaF: Contrastive-Collaborative Learning Framework for Weakly Supervised Audio-Visual Video Parsing

Faegheh Sardari[1], Armin Mustafa[1], Philip J. B. Jackson[1], and Adrian Hilton[1]

Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
{f.sardari,armin.mustafa,p.jackson,a.hilton}@surrey.ac.uk

Here we provide:

- More framework generalization of CoLeaF
- Additional ablation studies
- Temporal event-aware contrastive learning
- Computational complexity comparison
- More implementation details for the weakly supervised Audio-Visual Video Parsing (AVVP) task
- Implementation details for the weakly supervised Dense Audio-Visual Event Localization (DAVE) task

1 Framework Genelarization

CoLeaF is a general learning framework, which means that any existing AVVP method can be embedded in our framework. In Table 1, we embed VALOR [6] in CoLeaF. The results show that CoLeaF improves the performance of VALOR across all metrics. Specifically, it significantly increases VALOR's performance in detecting audible-only events (Ao) and audible-visible events (AV) by **1.2%** and **1.0%** F-score at the segment level and event level, respectively. It's worth noting that in VALOR, the networks are additionally trained with modality-specific labels. This indicates that CoLeaF not only facilitates cross-modal learning when we lack modality-specific supervision but also helps approaches that benefit from modality-aware pseudo labels.

2 Ablation Studies

In Table 2, we ablate the term of $\mathcal{L}_{CLS}^{Ref} = BCE(Y, CLS^{\phi})$ from the video level loss $\mathcal{L}_{video}^{Ref}$ on the LLP dataset. Tabel 2 shows that by removing this term, on average across all event types, the CoLeaF's performance decreases.

Table 1: Framework generalization on the LLP dataset. ⊛ indicates that the network has been embedded in CoLeaF as the Anchor branch.

	Segment-level			Event-level		
	Ao	Vo	AV	Ao	Vo	AV
VALOR [6]	49.0	66.7	58.4	44.2	65.0	52.2
VALOR⊛	**50.2** (+1.2)	**66.9** (+0.2)	**58.7** (+0.3)	**44.9** (+0.7)	**65.8** (+0.8)	**53.2** (+1.0)

Table 2: Ablation studies on $\mathcal{L}_{CLS}^{Ref} = BCE(Y, CLS^{\phi})$.

$\mathcal{L}_{CLS}^{Ref}$	Segment-level			Event-level		
	Ao	Vo	AV	Ao	Vo	AV
✗	48.9	62.3	58.4	43.6	**62.2**	52.0
✓	**49.3**	**62.4**	**58.6**	**44.1**	**62.2**	**52.1**

3 Temporal Event-Aware Contrastive Learning

In our proposed event-aware contrastive learning, the degree of encouragement is fixed for all segments of the input video since it is computed from the video-level output probabilities $\ddot{\mathcal{P}}^{\phi}$. However, our approach is capable of adapting such that the strength of the encouragement is computed separately for each segment of the video. To achieve this, instead of utilizing the video-level output probabilities to distil pseudo labels for the event-aware NCE loss, we employ segment-level output probabilities $\{\ddot{p}_t^{\phi}\}_{t=1}^{T}$ as

$$\ddot{g}_{i,t}^{\phi} = \begin{cases} 1 & \text{if } \ddot{p}_t^{\phi}(i) > \theta \\ 0 & \text{else} \end{cases}, \tag{1}$$

$$N_t^a = \sum_{c=1}^{C} \ddot{g}_{i,t}^{a} \odot (1 - \ddot{g}_{i,t}^{v}), N_t^v = \sum_{c=1}^{C} \ddot{g}_{i,t}^{v} \odot (1 - \ddot{g}_{i,t}^{a}), N_t^{av} = \sum_{c=1}^{C} \ddot{g}_{i,t}^{a} \odot \ddot{g}_{i,t}^{v}, \tag{2}$$

$$\vartheta^a = \frac{N_t^a}{N_t^a + N_t^{av}}, \text{ and } \vartheta^v = \frac{N_t^v}{N_t^v + N_t^{av}}. \tag{3}$$

Subsequently, the temporal event-aware NCE loss is computed as

$$\mathcal{L}_{T\text{-}Evt}^{Anch} = -\frac{1}{T} * \sum_{\phi \in \{a,v\}} \sum_{t=1}^{T} \vartheta_t^{\phi} \log \frac{\exp\left(\hat{f}_t^{\phi\intercal} . \ddot{x}_t^{\phi} / \tau\right)}{\sum_{n=1, n \neq t}^{T} \exp\left(\hat{f}_t^{\phi\intercal} . \ddot{x}_n^{\phi} / \tau\right)}. \tag{4}$$

In Table 3, we compare the performance of our proposed method CoLeaF on the LLP dataset when using either $\mathcal{L}_{T\text{-}Evt}^{Anch}$ or $\mathcal{L}_{Evt}^{Anch}$ for training. The results show that while CoLeaF achieves similar performance on most metrics for both losses, it achieves a 0.6% higher F-score overall when trained with $\mathcal{L}_{Evt}^{Anch}$.

4 Computational Complexity Comparison

In Table 4, we compare the computational complexity of the state-of-the-art weakly supervised AVVP approaches. It is shown that while CoLeaF benefits from the Reference branch and class tokens during training, its inference time is the same as its backbone CMPAE [3].

Table 3: CoLeaF's performance on the LLP dataset when using either $\mathcal{L}_{T\text{-}Evt}^{Anch}$ or $\mathcal{L}_{Evt}^{Anch}$ for training.

Loss	Segment-level (%)					Event-level (%)				
	Ao	Vo	AV	T@o	E@o	Ao	Vo	AV	T@o	E@o
$\mathcal{L}_{Evt}^{Anch}$	49.3	**62.4**	**58.6**	56.7	**37.7**	44.1	**62.2**	**52.1**	**52.7**	33.4
$\mathcal{L}_{T\text{-}Evt}^{Anch}$	**49.5**	62.1	**58.6**	**56.8**	37.5	**44.2**	61.9	51.9	52.6	**33.5**

Table 4: State-of-the-art computational complexity comparison. * indicates that the Reference branch in CoLeaF does not leverage the class token during training.

	HAN [7]	MA [8]	JoMoLD [2]	CMPAE [3]	CoLeaF*		CoLeaF	
					Train	Inference	Train	Inference
GFlops	37.29	37.29	37.29	48.15	60.85	48.15	66.15	48.15

5 More Details about Inference 'selected thresholds' Setting

In the standard setting for the AVVP task, during inference (applying a trained model to test (new) data), state-of-the-art approaches typically employ a single, fixed threshold (*e.g.* , 0.5) for all event classes. This threshold determines which network outputs are considered detected events. However, in CMPAE [4], the inference stage is performed under a 'selected thresholds' setting. To implement this, during training, two models are saved: (i) the Early-Stage model, representing the network at a specific point in training (*e.g.* , epoch 5), and (ii) the Best model, which achieves the highest performance on the validation set throughout the training process. After training, the Early-Stage model is utilized on the test data. Here, a range of thresholds (*i.e.* , 0.25 to 0.7) is applied to each event class in the LLP dataset. The thresholds that yield the highest F-score are then saved as the 'best thresholds' for each class. Finally, to evaluate the Best model on the test set, the Best model is employed on the test data again. However, this time, event proposals are obtained using the previously saved 'best thresholds' on the estimated output probabilities for each event class. For more details, please refer to their code available at here.

6 Implementation Details on UnAV-100 Dataset

Following [4], audio input tokens are extracted through pre-trained VGGish [5], and visual tokens are obtained through the pre-trained two-stream I3D [1]. The feature dimension for all input tokens, including the class tokens, is set to 512. To train CoLeaF, we utilized the Adam optimizer with an initial learning rate of 5×10^{-4} and a batch size of 5 for 12 epochs. The learning rate was decayed by a factor of 0.25 every 6 epochs. Our experiments were performed using PyTorch on an NVIDIA GeForce RTX 3090 GPU.

References

1. Carreira, J., Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
2. Cheng, H., Liu, Z., Zhou, H., Qian, C., Wu, W., Wang, L.: Joint-Modal Label Denoising for Weakly-Supervised Audio-Visual Video Parsing. In: European Conference on Computer Vision. pp. 431–448. Springer (2022)
3. Gao, J., Chen, M., Xu, C.: Collecting Cross-Modal Presence-Absence Evidence for Weakly-Supervised Audio-Visual Event Perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18827–18836 (2023)
4. Geng, T., Wang, T., Duan, J., Cong, R., Zheng, F.: Dense-Localizing Audio-Visual Events in Untrimmed Videos: A Large-Scale Benchmark and Baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22942–22951 (2023)
5. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: CNN Architectures for Large-Scale Audio Classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 131–135. IEEE (2017)
6. Lai, Y.H., Chen, Y.C., Wang, Y.C.F.: Modality-Independent Teachers Meet Weakly-Supervised Audio-Visual Event Parser. Advances in Neural Information Processing systems (2023)
7. Tian, Y., Li, D., Xu, C.: Unified Multisensory Perception: Weakly-Supervised Audio-Visual Video Parsing. In: European Conference in Computer Vision. pp. 436–454. Springer (2020)
8. Wu, Y., Yang, Y.: Exploring Heterogeneous Clues for Weakly-Supervised Audio-Visual Video Parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1326–1335 (2021)