Semantic-Aware Multi-modal Grouping for Weakly-Supervised Audio-Visual Video Parsing

Shentong Mo[1] and Yapeng Tian[2,*]

[1] Carnegie Mellon University
[2] University of Texas at Dallas

1 Introduction

Humans understand the surrounding environment by integrating signals from different senses. In our daily life, sound and sight are two of the most commonly used modalities, which have drawn much attention from researchers to explore computational audio-visual scene understanding.

Previous audio-visual work [6, 2] usually assumes audio and visual data are temporally aligned. However, the alignment does not always exist in real-world videos. For example, sounding objects in many videos are outside of the field-of-view (FoV). For these non-aligned cases, audio signals become more reliable in understanding the events of interest. In this work, we address the audio-visual video parsing (AVVP) task [5] that aims to parse a video into temporal event segments and predict the audible, visible, or audio-visible event categories.

Existing approaches [5, 8, 3] usually focus on learning to leverage the unimodal and cross-modal temporal contexts from weak supervisions. HAN [5] introduced a simple Multimodal Multiple Instance Learning framework with cross-modal and self-modal attention layers to utilize the video-level labels. Recent state-of-the-art methods usually use the HAN as the baseline and modify it to further improve parsing performance. Particularly, contrastive learning and label refinement are proposed in Wu and Yang [8], where they adopted a contrastive loss to enforce the temporal alignment between the audio and visual features at the same timestamp and augmented training data with modality-aware event labels generation.

Our key motivation is to learn compact and discriminative audio and visual representations by explicit multi-modal grouping for mitigating the modality and temporal uncertainties in the weakly-supervised audio-visual video parsing problem. Different from past approaches, we propose a new Multi-modal Grouping Network, namely MGN, to explicitly group semantic-aware multi-modal contexts, which enables learning more compact and discriminative audio and visual representations. Specifically, we first extract event-aware unimodal features through unimodal grouping in terms of learnable categorical embedding tokens for each individual modality. Then, we introduce a cross-attention layer with a hard attention mechanism to aggregate cross-modal temporal contexts. Finally, we utilize a cross-modal grouping module to predict the modality category from updated class-aware unimodal embeddings.

[*] Corresponding author.

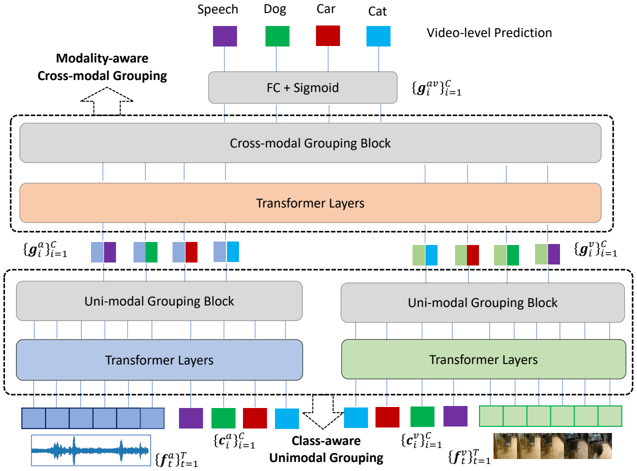

Fig. 1: Illustration of our Multi-modal Grouping Network (MGN).

Experimental results on the LLP [5] dataset validate that our new audio-visual video parsing framework achieves superior results over previous state-of-the-art methods [6, 2, 5, 8]. Empirical results also demonstrate the generalizability of our approach to contrastive learning and label refinement proposed in MA [8]. In addition, we substantially reduce the parameters of previous work by using only 47.2% parameters of baselines (17 MB vs. 36 MB).

2 Method

Given a video with both audio and visual tracks, our goal is to parse the video into temporal segments associated with audible, visible, and audio-visible event categories. We propose a novel baseline: Multi-modal Grouping Network (MGN) to address the task, which mainly consists of two modules: class-aware unimodal grouping and modality-aware cross-modal grouping, as shown in Figure 1.

Problem Setup and Notations. Given a video with T non-overlapping audio and visual segments, our goal is to temporally localize and recognize audio, visual, and audio-visual events that existed in the video. For the multi-label events with C event categories at time t, we have audio, visual, and audio-visual event labels for evaluation, which are denoted as: $\mathbf{y}_t^a, \mathbf{y}_t^v, \mathbf{y}_t^{av} \in \mathbb{R}^{1\times C}$. During training, we do not have the segment- and modality-level annotations. Therefore, we can only use the video-level label $\mathbf{y}^{av} \in \mathbb{R}^{1\times C}$ to perform weakly-supervised learning.

Revisit Multimodal Multiple Instance Learning. To address the weakly-supervised audio-visual video parsing problem, HAN [5] introduced a Multimodal Multiple Instance Learning (MMIL) framework based on transformers [7]. Given a set of audio-visual features $\mathbf{F}^a = \{\mathbf{f}_t^a\}_{t=1}^T, \mathbf{F}^v = \{\mathbf{f}_t^v\}_{t=1}^T$ in T segments, HAN applied self-attention and cross-attention layers to aggregate the unimodal and cross-modal information at each timestamp. Then, the probability of segment-wise categories for audio, visual, and audio-visual events ($\mathbf{p}_t^a, \mathbf{p}_t^v, \mathbf{p}^{av} \in \mathbb{R}^{1\times C}$) is predicted by a shared fully-connected (FC) layer and sigmoid function. Finally, the model is trained to optimize a weakly-supervised loss of $\mathbf{p}^{av}$ and a guided loss of $\mathbf{p}^a, \mathbf{p}^v$ with label smoothing: $\mathcal{L}_{base} = \mathrm{O}(\mathbf{y}^{av}, \mathbf{p}^{av}) + \mathrm{O}(\overline{\mathbf{y}}^a, \mathbf{p}^a) + \mathrm{O}(\overline{\mathbf{y}}^v, \mathbf{p}^v)$,

Table 1: Quantitative results of weakly-supervised audio-visual video parsing. 'C' and 'R' denote the contrastive learning and label refinement proposed in MA [8].

Method	Segment-Level					Event-Level				
	A	V	A-V	Type	Event	A	V	A-V	Type	Event
AVE [6]	47.2	37.1	35.4	39.9	41.6	40.4	34.7	31.6	35.5.	36.5
AVSDN [2]	47.8	52.0	37.1	45.7	50.8	34.1	46.3	26.5	35.6.	37.7
HAN [5]	60.1	52.9	48.9	54.0	55.4	**51.3**	48.9	43.0	47.7	48.0
MGN (ours)	**60.8**	**55.4**	**50.4**	**55.5**	**57.2**	51.1	**52.4**	**44.4**	**49.3**	**49.1**
MA [8] (w C)	**61.9**	53.1	49.7	54.9	56.2	**52.8**	49.4	43.8	48.7	49.0
MGN (w C)	60.6	**56.7**	**52.5**	**56.6**	**57.4**	51.4	**53.2**	**46.4**	**50.3**	**49.4**
MA [8] (w R)	59.8	57.5	52.6	56.6	56.6	**52.1**	54.4	45.8	50.8	**49.4**
MGN (w R)	**60.0**	**60.6**	**54.0**	**58.2**	**58.2**	50.3	**58.4**	**47.9**	**52.2**	49.1
MA [8] (w C+R)	**60.3**	60.0	55.1	58.9	57.9	**53.6**	56.4	49.0	53.0	**50.6**
MGN (w C+R)	60.2	**61.9**	**55.5**	**59.2**	**58.7**	50.9	**59.7**	**49.6**	**53.4**	49.9

where O$(\cdot)$ is a binary cross-entropy function-based loss term, which summarizes binary cross-entropy for all categories and $\mathrm{O}(\mathbf{y}, \mathbf{p}) = \sum_i BCE(y_i, p_i)$. $\overline{\mathbf{y}}^a, \overline{\mathbf{y}}^v$ are video-level audio and visual labels generated by smoothing $\mathbf{y}^{av}$ to decrease the confidence of positive labels.

Class-aware Unimodal Grouping. In order to explicitly group class-aware matching semantics for audio-visual representations, we introduce a novel class-aware unimodal grouping module by incorporating learnable modality-specific class tokens $\{\mathbf{c}_i^a\}_{i=1}^C, \{\mathbf{c}_i^v\}_{i=1}^C$ to help to group raw input unimodal features $\{\mathbf{f}_t^a\}_{t=1}^T, \{\mathbf{f}_t^v\}_{t=1}^T$. We first use self-attention transformers to temporally aggregate unimodal features from audio and visual inputs and align the features with the categorical token embeddings: Then, the unimodal grouping blocks take the learned audio and visual event class tokens and aggregated features as inputs to generate class-aware audio and visual embeddings $\{\mathbf{g}_i^a\}_{i=1}^C, \{\mathbf{g}_i^v\}_{i=1}^C$.

In order to constrain the independence of each class token $\mathbf{c}_i^a, \mathbf{c}_i^v$, we apply FC layers and softmax operation to generate the category probability $\mathbf{e}_i^a, \mathbf{e}_i^v$ of class tokens for audio and visual modalities with a class-constrained loss as:

$$\mathcal{L}_{cls} = \sum_{i=1}^{C} \mathrm{CE}(\mathbf{h}_i, \mathbf{e}_i^a) + \mathrm{CE}(\mathbf{h}_i, \mathbf{e}_i^v) \tag{1}$$

where CE$(\cdot)$ refers to cross-entropy loss; $\mathbf{h}_i$ is an one-hot encoding vector and only its element for the target class entry i is 1. After the class-aware unimodal grouping, the video-level category predictions $\mathbf{p}^a, \mathbf{p}^v$ of audio and visual events is simply computed by a FC layer and sigmoid operator.

Modality-aware Cross-modal Grouping. The modality uncertainty requires us to predict the modality category for matching with the only given video-level target in an explicit way. To achieve this, we propose a modality-aware cross-modal grouping module composed of cross-modal transformers and grouping blocks to aggregate class-aware representations. Based on the audio-visual similarity in the grouping stage, we combine all the audio features with visual features into new cross-modal modality-aware features. Then, we leverage the joint audio-visual representations $\{\mathbf{g}_i^{av}\}_{i=1}^C$ to predict the video-level target of audio and visual events via a FC layer and sigmoid function. The whole model can be optimized

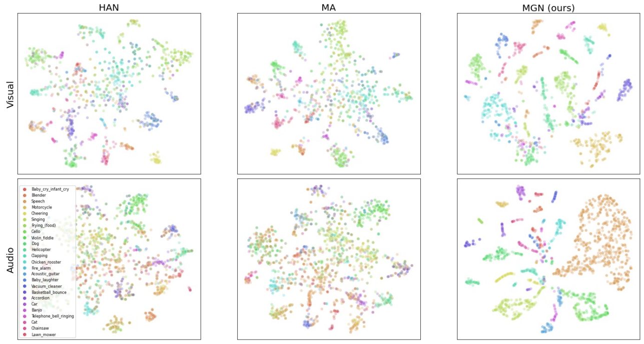

Fig. 2: Qualitative visualizations of learned audio and visual features.

in an end-to-end manner in terms of the objective function:

$$\mathcal{L} = \mathcal{L}_{base} + \mathcal{L}_{cls} \tag{2}$$

At inference time, the unimodal class-aware similarity is used to predict the audio, visual, and audio-visual events for each segment t.

3 Experiments

Dataset. The Look, Listen and Parse (LLP) Dataset [5] contains 11,849 YouTube video clips of 10-seconds long from 25 different event categories. We follow the prior work [5, 8] and use F-scores to evaluate both segment-level and event-level predictions for audio, visual, and audio-visual events. The model is trained with Adam [1] optimizer with β_1=0.9, β_2=0.999 and with an initial learning rate of 3e-4. We train the model with a batch size of 16 for 40 epochs.

Comparison to Prior Work. The quantitative comparisons with previous baselines [6, 2, 5, 8] on the LLP dataset are reported in Table 1. As can be seen, the proposed MGN achieves the overall best results against previous network baselines in terms most of metrics. Furthermore, significant gains can be observed in the setting of using the audio-visual contrastive learning and label refinement. These improvements imply the strong generalizability of the proposed MGN to the audio-visual contrastive learning and the label refinement.

Learned Class-aware Features. The learned class tokens are essential to grouping class-aware semantics from audio and visual features. To better evaluate the quality of those learned class-level features, we visualize the learned audio and visual representations of 25 categories by t-SNE [4], as shown in Figure 2. These meaningful visualizations further demonstrate that our MGN successfully learns compact and discriminative features for each modality.

4 Conclusion

In this work, we present MGN, a fully novel Multi-modal Grouping Network with class-aware unimodal grouping and modality-aware cross-modal grouping, to explicitly group class-aware matching semantics for weakly-supervised audio-visual video parsing. Experimental results demonstrate the effectiveness and superiority of our MGN against previous baselines. We also show the generalizability of our simple framework to the audio-visual contrastive learning and label refinement.

References

1. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
2. Lin, Y.B., Li, Y.J., Wang, Y.C.F.: Dual-modality seq2seq network for audio-visual event localization. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2002–2006 (2019)
3. Lin, Y.B., Tseng, H.Y., Lee, H.Y., Lin, Y.Y., Yang, M.H.: Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2021)
4. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008)
5. Tian, Y., Li, D., Xu, C.: Unified multisensory perception: Weakly-supervised audio-visual video parsing. In: Proceedings of European Conference on Computer Vision (ECCV). p. 436–454 (2020)
6. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: Proceedings of European Conference on Computer Vision (ECCV) (2018)
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS). Curran Associates, Inc. (2017)
8. Wu, Y., Yang, Y.: Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1326–1335 (2021)