Cross-Modal learning for Audio-Visual Video Parsing

Jatin Lamba, Abhishek, Jayaprakash Akula, Rishabh Dabral, Preethi Jyothi, Ganesh Ramakrishnan

Indian Institute of Techonology, Bombay

{jatinl, abhishekthakur, jayaprakash, rdabral, pjyothi, ganesh}@cse.iitb.ac.in

Abstract

In this paper, we present a novel approach to the audio-visual video parsing (AVVP) task that demarcates events from a video separately for audio and visual modalities. The proposed parsing approach simultaneously detects the temporal boundaries in terms of start and end times of such events. We show how AVVP can benefit from the following techniques geared towards effective cross-modal learning: (i) adversarial training and skip connections (ii) global context aware attention and, (iii) self-supervised pretraining using an audio-video grounding objective to obtain cross-modal audio-video representations. We present extensive experimental evaluations on the Look, Listen, and Parse (LLP) dataset and show that we outperform the state-of-the-art Hybrid Attention Network (HAN) on all five metrics proposed for AVVP. We also present several ablations to validate the effect of pretraining, global attention and adversarial training.

Index Terms: audio-visual video parsing, cross-modal learning.

1. Introduction

Audio Visual Video Parsing (AVVP) [1] is a newly introduced multi-modal task that involves detecting and localizing occurrences of events within the audio and visual streams of a video. AVVP has numerous potential applications. It can directly contribute to audio-visual source separation [2], especially when the sources of audio are occluded in the video. AVVP could also feed into downstream video understanding tasks (such as captioning or summarization) that could benefit from both audio and visual cues.

AVVP has been formulated as a Multi-modal Multi-Instance Learning (MMIL) task. The task becomes particularly challenging when there is only weak supervision; the set of events occurring in the audio and visual streams are available as a bag of events, while the start and end times of each individual event are not available. It is also entirely possible to have asynchronous annotations with different start and end times for the same event in different modalities. Furthermore, it is common to observe certain events (such as human speaking, telephone ringing, singing, baby crying, dog barking, violin playing, car, and vacuum cleaning) occurring either only in the audio modality or only in the visual modality.

Existing methods [1] have attempted to solve this task by learning cross-modal audio-visual features wherein representations for each modality are expected to benefit from the contexts surrounding the other modality. While such cross-modal representations are indeed powerful, they may not always be beneficial across modalities and might come at the cost of event detection performance for a specific modality. For the Look, Listen Parse (LLP) dataset [1], we believe the audio modality may not always benefit from the proposed cross-modal architecture for the following reasons. *Firstly,* there exists a dataset skew

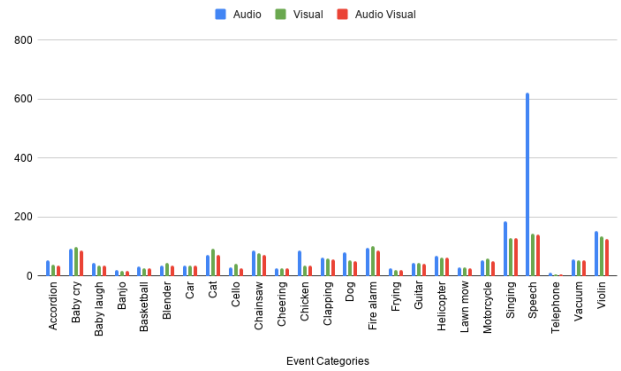

Figure 1: *Distribution of audio, visual and audio-visual events in the LLP dataset [1], across the* 25 *event categories. Note that the three lines blue, green and red are not disjoint sets.*

in favour of the audio modality that renders video sequences inconsequential for detecting audio streams. E.g., there are approximately 600 videos with speech as an audio event, while only 120 videos list it as a video event. In Figure 1 we present the distribution of audio, visual and audio-visual events across all event categories. As another example, visual features may be useless if the *dog bark* audio event occurs in the background of the scene. *Secondly*, it may be argued that input features for audio (VGGish) [3] are already well suited for audio parsing and the inclusion of cross-modal information from visual stream introduces noise into the resulting features. *Finally*, the existing state-of-the-art architecture for AVVP [1] attempts to project audio and visual features to a shared event label space using a common fully connected layer. Such a common projection layer might hurt certain audio events that do not benefit from visual features. For example, *doorbell ringing* might often be an audio event in the background without any supporting visual cues.

To alleviate these issues, we attempt to improve the performance of both visual and audio modalities. To that end, we introduce an adversarial loss to improve the quality of the shared audio-visual features. We alter the architecture to ensure that the audio features are not always influenced by cross-modal attention from the visual stream. We also leverage a global context attention mechanism [4] so that event prediction benefits not only from temporally local features but also from the global video context. Finally, we explore the benefits of using pretrained representations trained using self-supervised objectives (such as audio-video grounding) on a large audio dataset.

With our proposed techniques, we observe significant improvements not only in audio event detection performance but also in video detection performance on the Look, Listen and Parse (LLP) dataset [1]. Specifically, we improve the audio, vi-

sual and audio-visual F1-scores from 60.1% to 61.6%, 52.9% to 54.7% and 48.9% to 50.3%, respectively.

2. Related Work

2.1. Cross-Modal Attention Networks

Most of the work done in audio-visual learning assumes that temporal synchronized audio and visual content convey the same semantic meaning. However, natural unconstrained videos are noisy and tend to have redundant audio visual events that recur many times in the video, both within the same modality [5] as well as across different modalities [6],[7]. Hybrid Attention Network (HAN) [1] tries to jointly model the modalities and any asynchrony between them in a unified manner via self-attention and cross-attention networks. Xuan *et. al.* [8] propose a cross-modal attention network for audio-visual event localization with modality-specific CNNs acting on each modality to produce feature maps. These are subsequently fed into spatial, global context-aware and cross-modal attention modules that respectively learn 'where', 'when' and 'which' modality to attend to for event localization. We adopt the idea of a global context-aware attention mechanism which reflects the contribution of each individual segment towards event localization. Audio-visual speech recognition has also benefitted from cross-modal architectures [9]. In this work, a dual cross-modality attention scheme is proposed for a transformer-based model that combines two cross-modality attentions. DCM attention magnifies the role of visual modality to the same level as that of audio modality thus yielding better performance.

2.2. Multiple Instance Learning (MIL)

Multiple instance learning (MIL)[10] is a form of weakly supervised learning, wherein training instances are arranged in sets, called bags, and a label is provided for the entire bag instead of each element of the bag. Using a collection of labeled bags, the learner tries to induce a concept that will either (i) label individual instances correctly or (ii) label bags. Unlike the previous audio-visual event localization works (e.g., [11]) that are formulated as an MIL problem where an audio-visual snippet pair is regarded as an instance, each audio snippet and the corresponding visual snippet occurred at the same time denote two individual instances in a bag in our task. The multimodal multiple instance learning (MMIL) method in AVVP respects audio-visual temporal asynchrony.

2.3. Self-supervised learning

There have been previous attempts toward self-supervised audio-visual pre-training such as [12, 13, 14]. In [12], in a self supervised manner, the authors learn a temporal, multisensory representation that fuses the visual and audio components of a video signal. They train a neural network on a pretraining task of detecting misalignment between audio and visual streams in synthetically-shifted videos. In order to detect misalignment in a video of human speech, the integration of low level information across modalities is required, for which they propose a 3D multisensory convolutional network (CNN) with early fusion of audio and visual streams for modeling actions that produce a signal in both modalities.

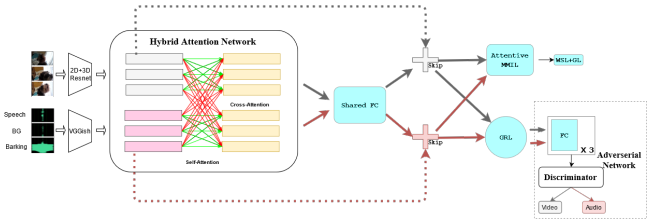

Figure 2: *The figure shows the architectural changes made to the HAN network with respect to [1]. The blocks in the cyan colour highlight modifications to the adversarial network.*

3. Our Approach

AVVP requires temporally grouping (contiguous) video snippets into audio, visual, and audio-visual events, and associating semantic labels with each event. More specifically, let $\mathcal{S}$ be a set of event categories. In the case of LLP, $\mathcal{S}$ has 25 events categories; *e.g.*, *human speaking*, *singing*. Given a video sequence containing both audio and visual tracks, the sequence is divided into T non-overlapping audio and visual snippet pairs $\{V_t, A_t\}_{t=1}^{T}$, where each snippet is 1 second long. V_t and A_t respectively denote the visual and audio content within the video snippet. Let $y_t = \left\{ (y_t^a, y_t^v, y_t^{av}) \,\middle|\, [y_t^a]_s, [y_t^v]_s, [y_t^{av}]_s \in \{0,1\}, s \in \mathcal{S} \right\}$ be the event label set for the video snippet $\{V_t, A_t\}$, where $s \in \mathcal{S}$ refers to an event category and y_t^a, y_t^v, and y_t^{av} denote audio, visual, and audio-visual event labels, respectively. Here, we have a relation: $y_t^{av} = y_t^a * y_t^v$, which means that audio-visual events occur only when there exist both audio and visual events during the same time span and and from the same event categories. We leverage the state-of-the-art Hybrid Attention Network (HAN) [1] as the base for our network architecture. Given separate streams of pre-trained audio (VGGish [3]) and visual (ResNet-152 [15] and 3DResNet [16] features), HAN uses cross-modal attention to learn modality contextualized features. As described in Figure 2 these features are then provided as an input to a shared linear layer that produces joint features for audio and video before finally passing them for Attentive MMIL Pooling.

Dataset Bias in LLP Dataset: We observe in Table 2 of [1] that improvement of visual event parsing degrades the performance of audio event parsing. This may be attributed to the fact that the LLP dataset (*c.f.* Figure 1 for stats) is itself collected from an audio oriented dataset, *viz.*, AudioSet. AudioSet has 1447 events in the video domain, as against a much higher 2090 events in the audio domain. Overall, this leads us to hypothesize that learning joint features through cross-modal attention and a shared linear layer, though beneficial for visual features, could be detrimental to the quality of audio features.

3.1. Our Proposed Model

In order to address the aforementioned issue of bias in the LLP dataset [1] affecting the AVVP task, we propose to (a) use the joint features as residuals to the self-attention outputs through skip connections, (b) introduce an adversarial loss to improve the learnt joint-features and (c) use global context attention for learning higher level features.

Skip connections: Skip connections are used in the attention module to pass self attended signals over the cross attention module. Using skip connections via addition, we try to preserve pure signals from each modality to avoid cluttering of unneces-

sary information from the other modality. Since LLP is rich in audio data (again see Figure 1), audio representation should not require as much support from visual data as the support visual representation might require from audio. The effectiveness of skip connections on audio is evident from table 1.

Adversarial Training: The attention network tries to learn information across both modalities which can be essential for the audio visual parsing task. In order to improve upon learning of joint features across modalities, just prior to the MMIL pooling layer, we add a modality discriminator to learn to discriminate between audio and visual features. During training, the parameters of the underlying attention network are optimized in order to minimize the loss of the video parser classifier and to maximize the binary cross entropy loss of the modality discriminator. In practice, a gradient reversal layer [17] is added over the features which feeds into the modality discriminator. Concretely, for a feature vector $f \in \mathbb{R}^d$, where d is the hidden dimension, let $G_m(f, \theta_m)$ be the modality discriminator with parameters θ_m. With the ground truth modality label y_m, we can optimize the proposed discriminator with a binary cross-entropy loss function:

$$\mathcal{L}_{ad} = -(y_m \log(G_m(f,.) + (1 - y_m) \log(G_m(f,.)) \quad (1)$$

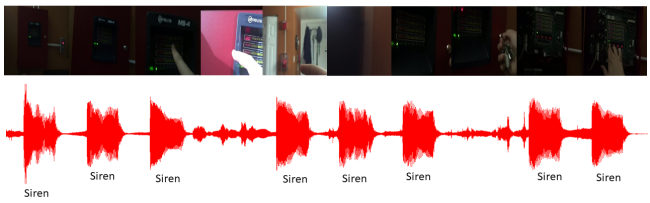

Figure 3: *The video and audio correspond to a fire alarm event. The video frames have no cues relevant to fire alarm. But only from the audio using the* `siren` *sound, the model can understand the actual event*

Global Context Aware Attention: We replace the multi head attention in HAN with global context aware attention [4]. A global context vector is computed, taking average over the features of the video segments to represent the global meaning of the entire sequence. The local context of a segment is modified using the segment level representation and the global context vector. These local context vectors are employed to obtain attention similar to the scaled dot product attention in transformers. We use this modified attention module for both self and cross modal attention network.

Overall Loss Objective: Our overall loss can be expressed as:

$$\mathcal{L} = \mathcal{L}_{wsl} + \lambda_g \mathcal{L}_g - \lambda_{ad} \mathcal{L}_{ad} \quad (2)$$

where $\mathcal{L}_{wsl}$ and $\mathcal{L}_g$ are borrowed from [1] and $\mathcal{L}_{ad}$ is the adversarial loss discussed above. $\mathcal{L}_{wsl}$ is the main weakly supervised loss [1] for event classification while $\mathcal{L}_g$ is the modality specific classification loss that is applied after label smoothing. These two components are also elaborated in our extended version[1].

3.2. Pretraining for Improved Features

In the context of our model proposed in Section 3.1, we investigate if pretraining the audio and visual input features could further enhance the performance of our model on the AVVP task. To this end, we introduce a novel pretraining strategy that jointly trains audio and visual inputs using the task of aligning the audio and visual modalities.

For pretraining through audio-visual alignment, we adapt the architecture of UNITER [18] and accommodate the modifications necessary for the audio and visual modalities. UNITER is a large scale image-text embedding network. It adopts the transformer architecture as the core and leverages its self attention mechanism to learn representations for both modalities in a joint embedding space. These joint representations are learnt by simultaneous application of four different self-supervised pre-training objectives *viz.*, Masked Language Modeling (MLM), Masked Region Modeling (MRM), Image-Text Matching (ITM), and Word-Region Alignment (WRA).

We use the pretrained embeddings for audio and visual modalities extracted from VGGish and Resnet152 networks. The pretrained and temporal features for both the modalities are then passed through a fully-connected (FC) layer, to obtain the embedding vectors. These embeddings are then fed into a multi-layer Transformer to learn a cross-modal contextualized embedding across video and audio clips. Note, that audio-visual alignment can be treated as a problem in itself and here we use the alignment task purely for the purpose of pretraining.

Audio-Visual Grounding (AVG) as a Pretraining Task: The inputs to AVG are a video clip and an audio clip sampled at random from a video segment and an audio segment respectively. A concatenated audio-visual embedding is fed into our model as a fused representation of both modalities. These embeddings are then trained to be discriminative based on the contrastive Loss. We extract the joint representation of the input snippet pairs, used cosine similarity, to predict a score between 0 and 1. The AVG supervision is over the predicted binary labels. During training, we sample an audio-visual snippet pair $\langle a_i, v_i \rangle$ and ground it with respect to to all video-audio snippets. We apply the contrastive loss $\mathcal{L}_{avg}$ in (3) for pretraining, where p and n are the margin parameters for the positive and negative pairs respectively. $[]_+$ refers to max(.,0).

$$\begin{aligned}\mathcal{L}_{avg} = &\sum_j [p - \langle a_i, a_j \rangle]_+ + \sum_j [\langle a_i, a_j \rangle - n]_+ \\ &+ \sum_j [p - \langle v_i, v_j \rangle]_+ + \sum_j [\langle v_i, v_j \rangle - n]_+\end{aligned} \quad (3)$$

Obtaining ground truth for AVG: The binary ground truth for each audio-video snippet pair could have been based on exact match of their temporal spans. However an audio snippet could also loosely (semantically) correspond to several video snippets with spans different from its own. Hence, a one-to-one mapping based on temporal span match might itself not be effective. Hence, we generate the ground truth using the method described next.

We construct a graph where each snippet is represented by a node and nodes are connected only if the corresponding snippets are *semantically similar*. Further, we consider two snippets to be *semantically similar* if either (i) the Resnet152 embeddings of corresponding video snippet or (ii) VGGish representation for the corresponding audio snippet, have similarity above some threshold. Now, all the audio-video snippets across different pairs (nodes) in the same connected component in this graph are considered to have ground truth label 1. Any other pair of audio-video snippets is considered to have ground truth label 0 (see detailed illustration in the extended version).

[1] https://avvp-iitb.github.io

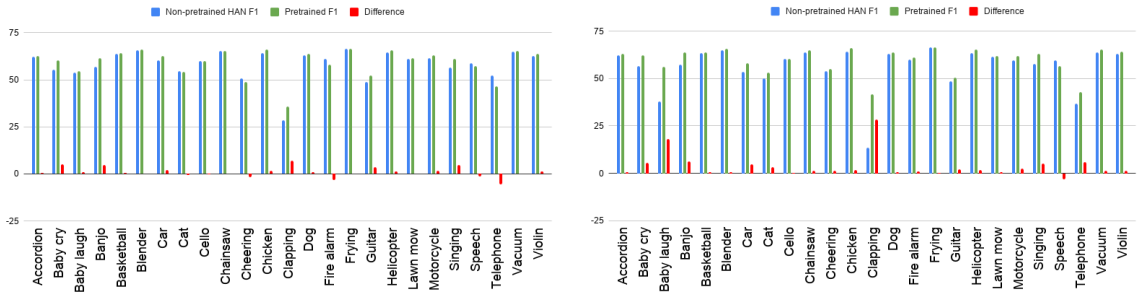

Figure 4: *Category-wise F1 scores from the non pretrained HAN and pretrained models for the audio (left) and visual (right) modalities.*

Method/Event	Audio	Visual	AV	Ty@AV	Ev@AV
HAN [1]	60.1	52.9	48.9	54.0	55.4
Base	60.0	52.6	48.4	53.7	54.8
Base + GCAA	**62.0**	53.4	48.9	54.8	**56.3**
Base + Adv+Skip	61.5	53.3	48.8	54.5	56
Base + Adv+Skip+GCAA	61.6	**54.7**	**50.3**	**55.5**	56.1
Pretrained	58.4	54.2	**50.4**	54.4	54.4
Pretrain E2E	59.8	**54.7**	49.4	54.7	56.2

Table 1: *AVVP segment-level results on the LLP test set.*

Method/Event	Audio	Visual	AV	Ty@AV	Ev@AV
HAN [1]	51.3	48.9	43.0	47.7	48.0
Base	51.9	48.5	41.9	47.4	49.5
Base + GCAA	53.4	48.9	42.0	48.1	50.7
Base + Adv+Skip	53.4	48.8	42.0	48.0	50.5
Base + Adv+Skip+GCAA	**53.7**	**50.5**	**43.5**	**49.2**	**50.9**
Pretrained	50.0	50.9	**45.0**	48.6	47.6
Pretrain E2E	50.8	**51.5**	42.9	48.4	49.3

Table 2: *AVVP event-level results on the LLP test set.*

4. Experiments and Results

Experimental Setup: The LLP dataset consists of videos, each of duration 10 secs. We divide each video into ten segments, each 1 sec long. For a visual segment, ResNet152 [15] features are extracted over frames sampled at 8fps and later fused with 3D ResNet to obtain 512-dimensional segment-level visual features. For an audio segment, we use the VGGish extractor [3] to get 512-dimensional audio features. We set $\lambda_g = 0.6$ and $\lambda_{ad} = 0.4$ in our objective function in eq.(2). We train our model with a learning rate of 3e-4 and decay it by a multiplicative factor of 0.5 every 5 epochs. We train our models for 40 epochs with a batch size of 64. We pretrain on the audio visual grounding task (*c.f.*, Section 3.2), on a subset (29K videos) of the AudioSet dataset [19] and extract representations for both audio and visual streams. We perform pretraining in two different ways. The first variant uses the frozen pretrained embeddings for both modalities with no further fine-tuning using LLP. The second variant uses standalone embeddings for the audio and visual modalities extracted from the ResNet and VGGish networks, respectively. These embeddings are trained with the pretraining objective in (3). The resulting pretrained representations along with r2p1d features [16] are further fine-tuned using the LLP dataset for the AVVP task.

Results and Analysis: We report results using the same evaluation metrics as in [1] (and elaborated in the extended version). Tables 1 and 2 present evaluations at the segment and event levels, respectively. 'Audio', 'Visual' and 'AV' refer respectively to audio, visual, and audio-visual events. Segment-level F-scores are evaluated at the level of segments, while the event-level F-scores are computed by concatenating consecutive positive snippets in the same event categories and then computing an event-level F-score based on mIoU = 0.5 as the threshold. We also compute two aggregated metrics, *viz.*, Type@AV (Macro F1) and Event@AV (Micro F1) which are averaged audio, visual, and audio-visual event evaluation results and the F-scores considering all events for each sample, respectively.

HAN is the current state-of-the-art system [1] for AVVP. "Base" differs from HAN in: (i) the use of separate self attention modules for the two modalities whereas HAN uses a shared module (ii) HAN performs cross modal attention directly over encoded features while Base does modality fusion over self attended features.[2] 'Skip', 'Adv' and 'GCAA' refer to the skip connections, adversarial training and global context-aware attention techniques respectively (*c.f.*, Section 3). 'Pretrained' and 'Pretrain E2E' refers to the two pretraining variants described earlier in this section. We observe that our proposed techniques lead to significant improvements over the state-of-the-art HAN network on *all five* metrics. Inclusion of 'Adv+Skip+GCAA' improves segment-level and event-level F1 scores on three of the five evaluation metrics.

Interestingly, while pretrained features help significantly with visual events, they do not improve performance on audio events. To understand this phenomenon better, we show category-wise F1 scores in Figure 4 for both audio and visual modalities. As per Figure 1, 77% of the 'speech' category events exist in the audio-only modality, possibly therefore 'speech' is not helped by pretraining on audio and visual events. Further, it is the most dominant (comprising more than 28% of all events), possibly explaining the overall drop in Table 1. Telephone ringing (that typically appears in the background) is an audio event that degrades most in performance with the pretrained model. Baby laughter and clapping, which are inherently rich in audio and visual cues, benefit the most from cross-modal pretraining.

5. Conclusions

In this paper, we present improved techniques for the new task of Audio-Visual Video Parsing. We address issues in existing cross-modal feature learning that might benefit the visual streams at the cost of event detection performance in audio modality. We demonstrate significant improvements over the state-of-the-art for AVVP and present comprehensive ablations. We believe our method will further advance the understanding of this new and challenging multi-modal MIL problem.

[2]These modifications in 'Base' were beneficial when used in conjunction with our Adv, Skip techniques.

6. References

[1] Y. Tian, D. Li, and C. Xu, "Unified multisensory perception: Weakly-supervised audio-visual video parsing," in *ECCV*, 2020.

[2] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party," *ACM Transactions on Graphics*, vol. 37, no. 4, p. 1–11, Aug 2018. [Online]. Available: http://dx.doi.org/10.1145/3197517.3201357

[3] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015, pp. 730–734.

[4] H. Xuan, Z. Zhang, S. Chen, J. Yang, and Y. Yan, "Cross-modal attention network for temporal inconsistent audio-visual event localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 279–286.

[5] M. R. Naphade and T. S. Huang, "Discovering recurrent events in video using unsupervised methods," in *Proceedings. International Conference on Image Processing*, vol. 2, 2002, pp. II–II.

[6] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/c4616f5a24a66668f11ca4fa80525dc4-Paper.pdf

[7] J. Vroomen, M. Keetels, B. de Gelder, and P. Bertelson, "Recalibration of temporal order perception by exposure to audio-visual asynchrony," *Cognitive Brain Research*, vol. 22, no. 1, pp. 32–35, 2004, ¿2000 woorden.

[8] H. Xuan, Z. Zhang, S. Chen, J. Yang, and Y. Yan, "Cross-modal attention network for temporal inconsistent audio-visual event localization," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, pp. 279–286, Apr. 2020. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5361

[9] Y.-H. Lee, D.-W. Jang, J.-B. Kim, R.-H. Park, and H.-M. Park, "Audio–visual speech recognition based on dual cross-modality attentions with the transformer model," *Applied Sciences*, vol. 10, no. 20, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/20/7263

[10] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," *Advances in neural information processing systems*, pp. 570–576, 1998.

[11] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," 2018.

[12] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," *CoRR*, vol. abs/1804.03641, 2018. [Online]. Available: http://arxiv.org/abs/1804.03641

[13] P. Morgado, N. Vasconcelos, and I. Misra, "Audio-visual instance discrimination with cross-modal agreement," 2021.

[14] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, and D. Tran, "Self-supervised learning by cross-modal audio-video clustering," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[16] R. Hou, C. Chen, R. Sukthankar, and M. Shah, "An efficient 3d CNN for action/object segmentation in video," *CoRR*, vol. abs/1907.08895, 2019. [Online]. Available: http://arxiv.org/abs/1907.08895

[17] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, pp. 59:1–59:35, 2016. [Online]. Available: http://jmlr.org/papers/v17/15-239.html

[18] Y. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "UNITER: learning universal image-text representations," *CoRR*, vol. abs/1909.11740, 2019. [Online]. Available: http://arxiv.org/abs/1909.11740

[19] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

Cross-Modal learning for Audio-Visual Video Parsing (Appendix)

Jatin Lamba, Jayaprakash Akula, Abhishek, Rishabh Dabral, Ganesh Ramakrishnan, Preethi Jyothi

Indian Institute of Techonology, Bombay

{jatinl, jayaprakash, abhishekthakur,rdabral, ganesh, pjyothi } @cse.iitb.ac.in

1. Obtaining ground truth for AVG

The binary ground truth for each audio-video snippet pair could have been based on exact match of their temporal spans. However an audio snippet could also loosely (semantically) correspond to several video snippets with spans different from its own. Hence, a one-to-one mapping based on temporal span match might itself not be effective. Hence, we generate the ground truth using the method described next. Instead of simple clustering based grouding where-in the snippets are clustered in their representation space, we take a graph based connectivity approach for obtaining the ground truth as described below.

We construct a graph where each snippet is represented by a node and nodes are connected only if the corresponding snippets are *semantically similar*. Further, we consider two snippets to be *semantically similar* if either (i) the Resnet152 embeddings of corresponding video snippet or (ii) VGGish representation for the corresponding audio snippet, have similarity above some threshold. This is a adaptation of agreement score as in [1], where similarity score used to determine the positives and negatives required in their setup. Now, all the audio-video snippets across different pairs (nodes) in the same connected component in this graph are considered to have ground truth label 1. Any other pair of audio-video snippets is considered to have ground truth label 0 see figure 1.

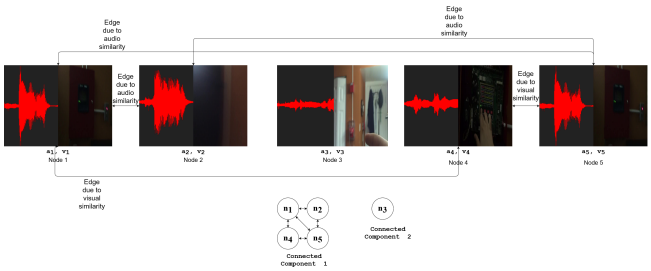

Figure 1: *A detailed illustrative description of the procedure to obtain the grounding labels. Node 1, 2, 4 and 5 belong to same group while node 3 is in different group.*

2. Ablation Studies

Audio-Visual Grounding (AVG) as a Pretraining Task: The inputs to AVG are a video clip and an audio clip sampled at random from a video segment and an audio segment respectively. A concatenated audio-visual embedding is fed into our model as a fused representation of both modalities. These embeddings are then trained to be discriminative based on the contrastive Loss. We extract the joint representation of the input snippet pairs, used cosine similarity, to predict a score between 0 and 1. The AVG supervision is over the predicted binary labels. During training, we sample an audio-visual snippet pair $\langle a_i, v_i \rangle$ and ground it with respect to to all video-audio snippets. We apply

Algorithm 1: Algorithm describing the procedure to obtain the ground truth for AVG task

Input: (V, A) set of video-audio dataset
Output: The ground truth for grounding task
Data: $S_i = (V_i, A_i)$ such that V_i, A_i are the i-th video-audio pair
$s_{ij} = (v_{ij}, a_{ij}), j \in \{1, 2, \ldots, 10\}$ ie 1-sec long video-audio snippet pair embeddings.
initialisation: Graph G_i for the i-th video-audio pair
A node n_{ij} for each video-audio snippet pair.
for *each pairs of snippets* (s_u, s_v) *in* S_i **do**
  $GT(u, v) \leftarrow 0$
for $j \leftarrow 1$ ***to*** 10 **do**
  for $k \leftarrow j + 1$ ***to*** 10 **do**
    if $\langle v_{ij}, v_{ik} \rangle \geq v_t$ ***or*** $\langle a_{ij}, a_{ik} \rangle \geq a_t$ **then**
      connect node n_{ij} and n_{ik};
Obtain the connected components $C_i$in G_i.
for *each component* C *in* C_i **do**
  for *each pairs of nodes* (u, v) *in* C **do**
    $GT(s_u, s_v) \leftarrow 1$

the contrastive loss for pretraining, where p and n are the margin parameters for the positive and negative pairs respectively. $[]_+$ refers to max(.,0). We try several variants of AVG tasks (i) Uni-modal AVG$_{u-avg}$, (ii) Cross-modal AVG$_{x-avg}$ and (iii) Multi-modal grounding AVG$_{m-avg}$. We provide ablation studies to understand the efficiency of the different variants.

$$\mathcal{L}_{u-avg} = \sum_j [p - \langle a_i, a_j \rangle]_+ + \sum_j [\langle a_i, a_j \rangle - n]_+ + \sum_j [p - \langle v_i, v_j \rangle]_+ + \sum_j [\langle v_i, v_j \rangle - n]_+ \quad (1)$$

$$\mathcal{L}_{x-avg} = \sum_j [p - \langle a_i, v_j \rangle]_+ + \sum_j [\langle a_i, v_j \rangle - n]_+ + \sum_j [p - \langle v_i, a_j \rangle]_+ + \sum_j [\langle v_i, a_j \rangle - n]_+ \quad (2)$$

$$\mathcal{L}_{m-avg} = \mathcal{L}_{u-avg} + \mathcal{L}_{x-avg} \quad (3)$$

Due to the fact that pretraining occured on snippet level, we have better results on all the metrics. In hindsight, event is just a contiguous set of snippets. So, essentially snippet level information played out the major factor in the improvement of results.

3. References

[1] P. Morgado, N. Vasconcelos, and I. Misra, "Audio-visual instance discrimination with cross-modal agreement," 2021.

Method/Event	Audio	Visual	AV	Ty@AV	Ev@AV
HAN [2]	60.1	52.9	48.9	54.0	55.4
Base	60.0	52.6	48.4	53.7	54.8
Base + GCAA	**62.0**	53.4	48.9	54.8	**56.3**
Base + Adv+Skip	61.5	53.3	48.8	54.5	56
Base + Adv+Skip+GCAA	61.6	**54.7**	**50.3**	**55.5**	56.1
Pretrain E2E(uni)	59.8	54.7	49.4	54.7	56.2
Pretrain E2E(cross)	60.3	**55.1**	49.9	55.1	**56.4**
Pretrain E2E(multi)	**60.8**	54.8	**50**	**55.2**	56.2

Table 1: *AVVP segment-level results on the LLP test set.*

Method/Event	Audio	Visual	AV	Ty@AV	Ev@AV
HAN [2]	51.3	48.9	43.0	47.7	48.0
Base	51.9	48.5	41.9	47.4	49.5
Base + GCAA	53.4	48.9	42.0	48.1	50.7
Base + Adv+Skip	53.4	48.8	42.0	48.0	50.5
Base + Adv+Skip+GCAA	**53.7**	**50.5**	**43.5**	**49.2**	**50.9**
Pretrain E2E(uni)	50.8	51.5	42.9	48.4	49.3
Pretrain E2E(cross)	52.3	**52.1**	43.9	49.4	50.4
Pretrain E2E(multi)	**53.4**	51.6	**44.3**	**49.8**	**51.2**

Table 2: *AVVP event-level results on the LLP test set.*

[2] Y. Tian, D. Li, and C. Xu, "Unified multisensory perception: Weakly-supervised audio-visual video parsing," in *ECCV*, 2020.