Investigating Modality Bias in Audio Visual Video Parsing

Piyush Singh Pasi, Shubham Nemani, Preethi Jyothi, Ganesh Ramakrishnan

Indian Institute of Technology, Bombay

{piyushsinghpasi, nshubham, pjyothi, ganesh}@cse.iitb.ac.in

Abstract

We focus on the audio-visual video parsing (AVVP) problem that involves detecting audio and visual event labels with temporal boundaries. The task is especially challenging since it is weakly supervised with only event labels available as a bag of labels for each video. An existing state-of-the-art model for AVVP uses a hybrid attention network (HAN) to generate cross-modal features for both audio and visual modalities, and an attentive pooling module that aggregates predicted audio and visual segment-level event probabilities to yield video-level event probabilities. We provide a detailed analysis of modality bias in the existing HAN architecture, where a modality is completely ignored during prediction. We also propose a variant of feature aggregation in HAN that leads to an absolute gain in F-scores of about 2% and 1.6% for visual and audio-visual events at both segment-level and event-level, in comparison to the existing HAN model.

Index Terms: AVVP, weakly-supervised learning, modality bias

1. Introduction

The Audio-Visual Video Parsing (AVVP) [1] task involves the fine-grained parsing of a video to generate temporal audio, video and audio-visual event labels. The additional challenge in this task is that there is only weak supervision during training in the form of event labels for the entire video, whereas the objective is to predict fine-grained audio and visual events for temporal event segments. AVVP is regarded as a Multimodal Multiple Instance Learning (MMIL) problem and it has a number of applications in audio-visual source separation and other video understanding tasks.

In order to utilize the weak labels effectively, Tian *et. al.* [1] proposed a Hybrid Attention Network (HAN) that generates aggregated feature representations for the audio and visual modalities with attention within and across modalities. These aggregated representations are further processed by an attentive MMIL pooling module that combines the representations using attention weights to produce a probability distribution across event labels for each video. This probability distribution can be directly used within a cross-entropy loss with the weak video-level labels serving as ground-truth.

After carefully analyzing the training losses and the assumptions in the model proposed by [1], we revisit their design choice of using cross attention for both modalities when constructing aggregated features. We hypothesize that using cross-modal attention on the audio modality but only self-attention on the visual modality could be more effective. And we indeed empirically verify that this variant outperforms the model in [1] with significant gains on the visual evaluation metrics.

When learning from multiple modalities, the model could establish spurious correlations between the target event and one (or more) modalities and not extract meaningful signals from all modalities. Such spurious correlations could lead to reasonable downstream task performance on the training dataset. However, the resulting models would demonstrate a modality bias and would not generalize well to test instances where the ignored modalities have rich signals. Tian *et. al.* [1] aim at alleviating modality bias by added modality-specific losses and label smoothing. Our experiments reveal that label smoothing could in fact reinforce modality biases depending on which modalities use smoothed labels, and hence should be used carefully.

Our main contributions can be summarized as follows:

1. We propose a simple and well-motivated variant of feature aggregation within the Hybrid Attention Network that yields significant gains on the evaluation metrics compared to [1].
2. We carefully analyze the effect of label smoothing and explain how it causes drastic shifts in the audio-visual attention distributions generated by attentive MMIL pooling. We also examine its impact on the final performance using a detailed ablation study [1].

2. Related Work

2.1. Audio-Visual Representation Learning

Videos contain information in both audio and visual modalities. Due to temporal synchronization in audio and visual data, most works [2, 3, 4, 5, 6, 7, 8] focus on learning joint embedding to exploit information present in both the modalities. Ying *et. al.* [3] employ cross-attention and self-attention for audio-visual representation learning. Ruohan *et. al.* [2] propose an attention based LSTM network for fusing audio-visual information and use it for action recognition. George *et. al.* [4] use cross-modal alignment for robust speech recognition. The Hybrid Attention Network HAN [1] jointly models the audio and visual modalities via self-attention and cross-attention.

2.2. Multiple Instance Learning (MIL)

Multiple instance learning (MIL) [9, 10, 11, 12, 13, 14] is a form of supervised learning in which the dataset consists of bags. Each bag contains multiple training instances and we have labels for a bag; each instance in the bag is itself unlabeled. From a collection of labeled bags, the learner tries to induce a model that will label individual instances correctly. Yapeng *et. al.* in [15] formulate audio-visual event localization as a MIL problem. They use an audio-visual snippet pair as a single instance, in contrast to AVVP in which the audio and visual snippet at the same time instance are considered as two separate instances. Similarly, multiple instance learning has also been used for object detection [16, 17].

[1] More details at: https://www.cse.iitb.ac.in/~malta/mbias

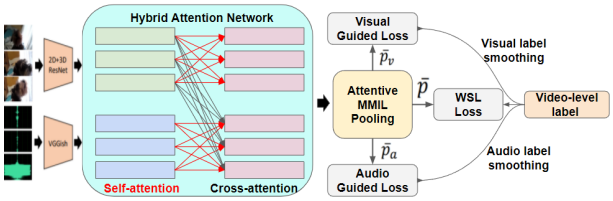

Figure 1: *Depiction of architectural changes made to the HAN network [1]. The block in the aqua colour highlights modifications made to the attention module for A_{cross}, V_{self} (there is no cross attention for visual features).*

3. AVVP

As described in [1], each video in the AVVP task is divided into T segments of audio and visual snippet pairs denoted by $\{A_t, V_t\}_{t=1}^{T}$. Each snippet $\{A_t, V_t\}$ is associated with an event label set $\boldsymbol{y}_t = \{y_t^a, y_t^v, y_t^{av}\}$; here, y_t^a, y_t^v, y_t^{av} are vectors of dimensionality C denoting audio, visual and audio-visual event labels respectively, where C is the size of the event label set. Only video-level labels are assumed to be available during training (resulting from weak supervision), while audio and visual events will be predicted for each video snippet during inference.

[1] proposed the use of a hybrid attention network (HAN) and attentive multimodal multiple instance learning (MMIL) pooling for the AVVP task. First, the audio and visual snippet pairs $\{A_t, V_t\}_{t=1}^{T}$ are passed through pretrained feature extractors to generate audio and visual representations $\{f_a^t\}_{t=1}^{T}$ and $\{f_v^t\}_{t=1}^{T}$, respectively. Next, HAN learns hybrid attention functions from both audio and visual representations at each time-step t, $\boldsymbol{f}_a = [f_a^1, f_a^2, ..., f_a^T]$ and $\boldsymbol{f}_v = [f_v^1, f_v^2, ..., f_v^T]$, to yield the following aggregate audio (A_{cross} or $\hat{f}_a^t$) and video (V_{cross} or $\hat{f}_v^t$) representations:

$$A_{cross} \equiv \hat{f}_a^t = f_a^t + g_{sa}(f_a^t, \boldsymbol{f}_a) + g_{ca}(f_a^t, \boldsymbol{f}_v) \quad (1)$$

$$V_{cross} \equiv \hat{f}_v^t = f_v^t + g_{sa}(f_v^t, \boldsymbol{f}_v) + g_{ca}(f_v^t, \boldsymbol{f}_a) \quad (2)$$

where g_{sa} and g_{ca} are self-attention and cross-modal attention functions, respectively. These attention functions capture intra-modal and inter-modal similarities and are computed using the standard attention formulation [18]: $g_{sa}(f_a^t, \boldsymbol{f}_a) = \text{softmax}(\frac{f_a^t \boldsymbol{f}_a^\intercal}{\sqrt{d}})\boldsymbol{f}_a$ and $g_{ca}(f_a^t, \boldsymbol{f}_v) = \text{softmax}(\frac{f_a^t \boldsymbol{f}_v^\intercal}{\sqrt{d}})\boldsymbol{f}_v$, where d is the dimensionality of the audio/visual features.

Attentive MMIL Pooling. [1] uses attentive MMIL pooling to leverage weakly-supervised video-level labels during training. In [1], the aggregated temporal features $\{\hat{f}_a^t, \hat{f}_v^t\}_{t=1}^{T}$ are passed through a shared fully-connected (FC) layer with sigmoid activation to obtain output probabilities for each individual event category. The predicted audio and visual event probabilities for time-step t are p_a^t and p_v^t, respectively. The audio and visual event probabilities are aggregated using attentive MMIL pooling to predict the video-level event probability $\bar{\boldsymbol{p}}_{wsl}$ as follows:

$$\bar{\boldsymbol{p}}_{wsl} = \sum_{t=1}^{T}\sum_{m=1}^{M}(W_{tp} \odot W_{av} \odot P)[t, m, :] \quad (3)$$

where $P(t, 1, :) = p_a^t$ and $P(t, 2, :) = p_v^t$, $\odot$ denotes element-wise multiplication, M is the total number of modalities (*i.e.*, 2 in AVVP where $m \in \{1, 2\}$ refers to the audio and visual modalities, respectively). W_{tp} and W_{av} are temporal attention and audio-visual attention tensors predicted from the aggregated features $\{\hat{f}_a^t, \hat{f}_v^t\}_{t=1}^{T}$ respectively. W_{tp} and W_{av} are computed as $W_{tp}[:, m, c] = \text{softmax}(F_{tp}[:, m, c])$ and $W_{av}[t, :, c] = \text{softmax}(F_{av}[t, :, c])$ where F_{tp} and F_{av} refer to two different fully-connected layers, $t = 1, 2, \ldots, T$, $m = 1, 2$ and $c = 1, 2, \ldots, C$. The ground truth labels $\bar{\boldsymbol{y}}$ and the predicted video-level event probabilities $\bar{\boldsymbol{p}}_{wsl}$ are used to optimize the binary cross-entropy loss function specified as: $\mathcal{L}_{wsl} = CE(\bar{\boldsymbol{p}}_{wsl}, \bar{\boldsymbol{y}}) = -\sum_{c=1}^{C} \bar{\boldsymbol{y}}[c] log(\bar{\boldsymbol{p}}_{wsl}[c])$.

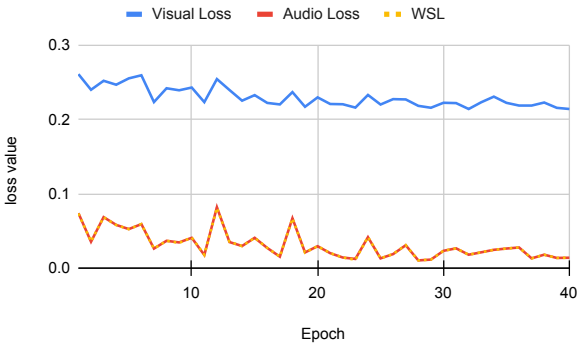

Figure 2: *Comparison between loss values of visual loss $\mathcal{L}_v$, audio loss $\mathcal{L}_a$, and weakly-supervised loss $\mathcal{L}_{wsl}$ for HAN [1]*

.

For better results and to alleviate modality bias, [1] proposes the use of cross-entropy losses specific to each individual modality. This modality-guided loss, $\mathcal{L}_g$ is computed as follows: $\mathcal{L}_g = \mathcal{L}_a + \mathcal{L}_v = CE(\bar{\boldsymbol{p}}_a, \bar{\boldsymbol{y}}_a) + CE(\bar{\boldsymbol{p}}_v, \bar{\boldsymbol{y}}_v)$, where $\bar{\boldsymbol{y}}_a$ and $\bar{\boldsymbol{y}}_v$ are video-level ground truth labels for the audio and visual modalities, respectively (that is simply initialized as $\bar{\boldsymbol{y}}_a = \bar{\boldsymbol{y}}_v = \bar{\boldsymbol{y}}$), $\bar{\boldsymbol{p}}_a$ and $\bar{\boldsymbol{p}}_v$ are video-level audio and visual event probabilities given by:

$$\bar{\boldsymbol{p}}_a = \sum_{t=1}^{T}(W_{tp} \odot P)[t, 1, :] \quad (4)$$

$$\bar{\boldsymbol{p}}_v = \sum_{t=1}^{T}(W_{tp} \odot P)[t, 2, :] \quad (5)$$

During training, the model is optimized over the combined loss $\mathcal{L} = \mathcal{L}_{wsl} + \mathcal{L}_g$.

Label Smoothing. Since we do not have video-level ground truth labels for each modality, $\bar{\boldsymbol{y}}_a$ and $\bar{\boldsymbol{y}}_v$ are set to $\bar{\boldsymbol{y}}$ which lacks modality-specific information. Tian *et. al.* [1] suggest using label smoothing on $\bar{\boldsymbol{y}}_a$ and $\bar{\boldsymbol{y}}_v$, as some events in $\bar{\boldsymbol{y}}$ can be noise for a particular modality. For example, if $\bar{\boldsymbol{y}}$ = {*Telephone_bell_ringing, Cat*}, the telephone bell could be ringing in the background without any accompanying visual cues and conversely, the cat could only be seen in the video and not heard at all. Label smoothing is implemented in [1] as: $\bar{\boldsymbol{y}}_m = (1 - \delta_m)\bar{\boldsymbol{y}} + \frac{\delta_m}{K}$ where $m \in \{a, v\}$ for audio and visual modalities; $\delta_m \in [0, 1)$ is the confidence parameter yielding a convex combination of (i) the event probability distribution and (ii) the uniform distribution $U = \frac{1}{K}$ (where $K > 1$) which helps in smoothing the event probability distribution.

Behaviour of Training Losses. Figure 2 shows the loss values corresponding to $\mathcal{L}_a$, $\mathcal{L}_v$, and $\mathcal{L}_{wsl}$ during training. We observe that the weakly-supervised loss $\mathcal{L}_{wsl}$ and the audio loss functions $\mathcal{L}_a$ nearly coincide. The mean squared difference between $\mathcal{L}_{wsl}$ and $\mathcal{L}_a$ for 40 epochs denoted by $MSE_{wsl,a}$ is

9e-8. In contrast, the mean squared difference between $\mathcal{L}_{wsl}$ and $\mathcal{L}_v$ denoted by $MSE_{wsl,v}$ is 0.104. This curious training artefact of $\mathcal{L}_{wsl}$ and $\mathcal{L}_a$ mirroring each other can be explained as follows.

Given that $\mathcal{L}_{wsl}$ is a binary cross-entropy loss function between $\bar{\boldsymbol{p}}_{wsl}$ and $\bar{\boldsymbol{y}}$ and $\mathcal{L}_a$ is a binary cross-entropy loss function of $\bar{\boldsymbol{p}}_a$ and $\bar{\boldsymbol{y}}$, the only difference appears in the event probability values. For $\mathcal{L}_{wsl}$ to imitate $\mathcal{L}_a$, the video-level event probability values $\bar{\boldsymbol{p}}_{wsl}$ and the audio event probabilities $\bar{\boldsymbol{p}}_a$ should be very close to each other. From equations 3, 4, and 5, $\bar{\boldsymbol{p}}_{wsl}$ can be written as a weighted combination of $\bar{\boldsymbol{p}}_a$ and $\bar{\boldsymbol{p}}_{\boldsymbol{v}}$:

$$\bar{\boldsymbol{p}}_{wsl} = W_a \odot \bar{\boldsymbol{p}}_a + W_v \odot \bar{\boldsymbol{p}}_v \tag{6}$$

where $W_a = W_{av}[:,1,:]$ and $W_v = W_{av}[:,2,:]$ correspond to audio and visual attention tensors, respectively. If $\boldsymbol{p}_{wsl}$ and $\boldsymbol{p}_a$ become close to each, then $W_a \approx \mathbf{1}$ and $W_v \approx \mathbf{0}$. This could be attributed to label smoothing (which will be discussed in detail in Section 5.3). This modality bias also motivates us to explore different variants of the aggregate features that we outline in the next section.

4. Proposed Variants of Aggregate Features

Tian *et. al.* [1] propose using the HAN network to generate aggregate audio and video representations, *viz.*, $\hat{f}_a^t$ and $\hat{f}_v^t$, each containing both self-attention and cross-attention based functions. However, this symmetric treatment of constructing audio and video representations may not be the best choice due to the nature of the modalities involved. Consider the example of $\bar{\boldsymbol{y}}_m = \{$*Telephone_bell_ringing, Fire_alarm*$\}$. These events can occur in the background without any supporting visual clues, and hence will not appear in the ground truth event labels for the visual modality. However, cross-attention from audio in this case could induce false signals of the presence of these events in the visual modality and subsequently hurt performance. Audio events, on the other hand, could benefit from the visual modality in helping disambiguate between audio sounds (*e.g.*, {*Blender, Vacuum_cleaner*}) using visual clues. These observations suggest that *a model with visual features aggregated using only self-attention and audio features aggregated using both self-attention and cross-attention might perform better.*

We remove the cross-modal attention function $g_{ca}()$ while computing aggregated features for the visual modality to get:

$$V_{self} = \tilde{f}_v^t = g(f_v^t, \boldsymbol{f}_a, \boldsymbol{f}_v) = f_v^t + g_{sa}(f_v^t, \boldsymbol{f}_v) \tag{7}$$

For a complete analysis, we can similarly remove cross-attention from the aggregated features for the audio modality to get:

$$A_{self} = \tilde{f}_a^t = g(f_a^t, \boldsymbol{f}_v, \boldsymbol{f}_a) = f_a^t + g_{sa}(f_a^t, \boldsymbol{f}_a) \tag{8}$$

Recall that in Eq. (1) and (2), we referred to the aggregated audio and visual features as A_{cross} and V_{cross}, respectively.

The model $A_{cross}+V_{self}$ will refer to the above-mentioned feature aggregates where the visual features do not use cross-attention from audio. For the sake of completeness, we also compare against the remaining three variants, $A_{self} + V_{self}$, $A_{self}+V_{cross}$ and $A_{cross}+V_{cross}$. Note that $A_{cross}+V_{cross}$ is the same model proposed in [1].

5. Experiments

5.1. Experimental Setup

The LLP dataset [1] contains 11,849 YouTube video clips with 25 event categories for a total of 32.9 hours. Each video is 10-second-long. The training set consists of 10,000 videos with weak labels i.e video-level event labels. The validation and test sets (that are manually annotated with segment-level labels) have 649 and 1200 videos, respectively. Following [1], we sample these videos at 8 frames per second and break into non-overlapping segments of length 1 second. Features extracted from ResNet152 [19] and 3D ResNet [20] are fused to get 512-dimensional segment-level visual features. We use a VGGish extractor [21] to extract 128-dimensional segment-level audio features. We train our model for 40 epochs with batch size 16, Adam as the optimizer, and initial learning rate set to 3e-4. The learning rate is dropped by a factor of 0.1 after every 10 epochs.

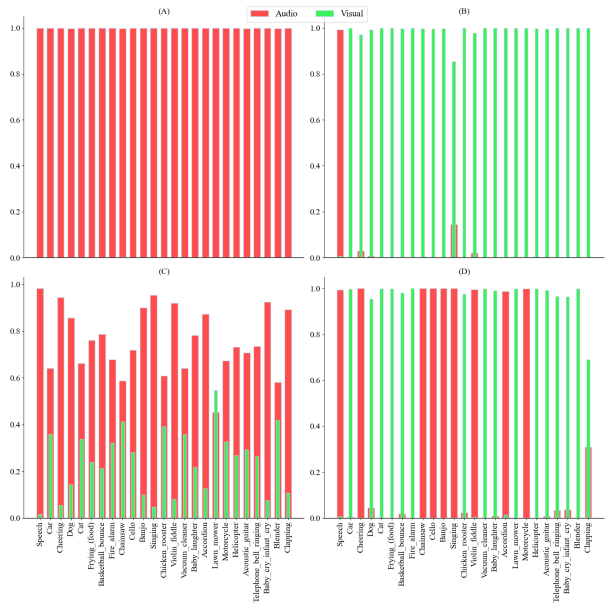

Figure 3: *Category-wise audio-visual attention weights distribution aggregated over the test set for different types of label smoothing. Visual(A), Audio(B), no smoothing(C) and Audio-Visual(D)*

5.2. Evaluation Metrics

All our models are evaluated on the metrics proposed by [1]. We use F-scores for audio, visual, and audio-visual events for segment-level and event-level metrics, indicating segment-level performance and video-level performance, respectively. To calculate event-level metrics, consecutive events of the same category are concatenated, and the F-score is calculated based on *mIoU* = 0.5 as the threshold. The Ty@AV metric refers to an average over audio, visual, and audio-visual evaluation results. The Ev@AV metric calculates the F-score for all audio and visual events of each video.

5.3. Results and Analysis

Table 1 shows the segment and event-level F1-scores for AVVP using the four different variants of aggregate features detailed in Section 4. We make the following two key observations. 1. $A_{cross} + V_{self}$ is the best performing variant across all five evaluation metrics, and consistently outperforms the baseline model [1], *viz.*, $A_{cross} + V_{cross}$. 2. Using A_{self} instead of A_{cross} leads to a large and consistent drop in performance across all metrics. This suggests that the audio modality clearly

Table 1: *Results of AVVP on LLP dataset for different variants of HAN*

Event type	Method	Segment level	Event level
Audio	$A_{self} + V_{self}$	52.1	41.6
	$A_{self} + V_{cross}$	49.9	39.2
	$A_{cross} + V_{self}$	**60.5**	**51.9**
	$A_{cross} + V_{cross}$ [1]	60.1	51.3
Visual	$A_{self} + V_{self}$	53.4	48.7
	$A_{self} + V_{cross}$	53.9	49
	$A_{cross} + V_{self}$	**54.9**	**51.2**
	$A_{cross} + V_{cross}$ [1]	52.9	48.9
Audio & Visual	$A_{self} + V_{self}$	41.2	33.4
	$A_{self} + V_{cross}$	40	30.9
	$A_{cross} + V_{self}$	**50.5**	**44.3**
	$A_{cross} + V_{cross}$ [1]	48.9	43
Ty@AV	$A_{self} + V_{self}$	48.9	41.2
	$A_{self} + V_{cross}$	48	39.7
	$A_{cross} + V_{self}$	**55.3**	**49.1**
	$A_{cross} + V_{cross}$ [1]	54	47.7
Ev@AV	$A_{self} + V_{self}$	52.2	42.7
	$A_{self} + V_{cross}$	51.5	42.3
	$A_{cross} + V_{self}$	**56.5**	**48.9**
	$A_{cross} + V_{cross}$ [1]	55.4	48

benefits from cross-modal attention from the visual modality.

Label Smoothing. Figure 3 shows the audio-visual attention weight distributions aggregated for the test set using four different types of label smoothing, *viz.*, (A) Smoothing of labels of only the audio (LS-A) or (B) only of video (LS-V), (C) no smoothing at all (No-LS) and (D) smoothing the labels of both modalities (LS-AV). It is evident that smoothing only the labels of one modality (*i.e.*, visual in plot A and audio in plot B) leads to the attention weights being completely biased towards the other modality (*i.e.*, audio in plot A and visual in plot B). Removing label smoothing entirely or adding label smoothing to both modalities yields attention distributions without a clear modality bias (shown in (C) and (D)).

This behaviour can be explained by examining the effect of label smoothing on the losses and its subsequent effect on the audio-visual attention weights. Adding smoothing to a particular modality $m \in \{a, v\}$ makes $\bar{\boldsymbol{y}}_m$ a real-valued vector, increasing the terms in $\mathcal{L}_m$ to the total number of events. In the case of multi-hot vectors, only a few terms equal to the number of ground-truth events are present. This poses a challenge to minimizing the loss. In LS-A, $\bar{\boldsymbol{y}}, \bar{\boldsymbol{y}}_v$ are multi-hot vectors and $\bar{\boldsymbol{y}}_a$ is a real-valued vector. Setting $W_a = \mathbf{0}$ and $W_v = \mathbf{1}$ will yield $\bar{\boldsymbol{p}}_{wsl}$ in the form of a multi-hot vector and will thus minimize $\mathcal{L}_{wsl}$. Any other W_a and W_v will not generate a multi-hot vector. Similarly, in LS-V, the model picks $W_a = \mathbf{1}$ and $W_v = \mathbf{0}$ to minimize $\mathcal{L}_{wsl}$. In LS-AV, for $\bar{\boldsymbol{p}}_{wsl}$ to be closest to the multi-hot vector form, in the absence of an event, setting $W_m = \mathbf{1}$ for modality with least probability is the best resort for $\mathcal{L}_{wsl}$. Similarly, in the presence of an event, setting $W_m = \mathbf{1}$ for modality with the highest probability will minimize $\mathcal{L}_{wsl}$ more. When there is no label smoothing, such extreme skewness vanishes.

Table 2 shows the segment-level and event-level results using $A_{cross} + V_{self}$ and $A_{cross} + V_{cross}$ with four types of label smoothing (LS-A, LS-V, No-LS and LS-AV). When compared to the model No-LS, the model LS-V shows a significant increase in the performance of the audio modality. Similarly, label smoothing on audio modality *i.e.*, LS-A shows a drop in audio evaluation metrics and some gain on visual evaluation metrics. On applying label smoothing on both modalities *i.e.*, LS-AV, the model favors the visual modality more. This aligns with the audio-visual attention weights in Fig. 3 (D). Another interesting observation from Table 2 is that $A_{cross} + V_{self}$ is more robust to label smoothing compared to $A_{cross} + V_{cross}$; the averaged variance of F1-scores of audio, visual and audio-visual events across types of label smoothing is 0.35 for $A_{cross} + V_{self}$ compared to 0.95 for $A_{cross} + V_{cross}$. Such empirical stability, particularly in the segment-level evaluation metrics, is owing to our reformulated HAN- $A_{cross} + V_{self}$ and aligns with the intuition stated in Section 4. Audio events can occur in the background with no support for visual cues in the frames. Using such events as ground truth in $\mathcal{L}_v$ reduces the performance, since the ground truth itself is noisy, and thus label smoothing helps.

Table 2: *Effect of Label Smoothing on HAN [1] i.e.* $A_{cross} + V_{cross}$ *and* $A_{cross} + V_{self}$. *No-LS implies no label smoothing; LS-A denotes smoothing only in audio modality; LS-V denotes smoothing only in visual modality; LS-AV denotes smoothing on both audio and visual modalities.*

Event type	Smoothing modality	$A_{cross} + V_{cross}$[1]		$A_{cross} + V_{self}$	
		Segment level	Event level	Segment level	Event level
Audio	No-LS	58	49.7	60.3	52.1
	LS-A	57.9	49.1	60.3	51.8
	LS-V	**60.1**	**51.3**	**60.5**	**51.9**
	LS-AV	57.5	48	59.9	51
Visual	No-LS	52.6	48.6	53.7	50
	LS-A	53.1	48.5	53.7	50.4
	LS-V	52.9	48.9	**54.9**	**51.2**
	LS-AV	**54.3**	**50.3**	49.3	42.8
Audio & Visual	No-LS	47.6	41.4	49.4	43.8
	LS-A	47.7	41	49.4	43.8
	LS-V	**48.9**	**43**	**50.5**	**44.3**
	LS-AV	48.6	42.2	49.3	42.8
Ty@AV	No-LS	52.7	46.6	54.5	48.7
	LS-A	52.9	46.2	54.5	48.7
	LS-V	**54**	**47.7**	**55.3**	**49.1**
	LS-AV	53.4	46.8	54.3	47.7
Ev@AV	No-LS	54.3	47.3	56.3	48.7
	LS-A	54.9	47	56	48.8
	LS-V	**55.4**	**48**	**56.5**	**48.9**
	LS-AV	54.7	46.6	56	47.8

6. Conclusions

In this work, we focus on the AVVP problem and specifically study the issue of modality bias in the main model proposed for this task in [1]. We trace the source of modality bias to label smoothing that was a part of the originally proposed framework for AVVP. We propose a new variant for aggregating features within this framework that is not only more accurate than the baseline but is also more robust to label smoothing. As part of future work, we propose to develop modality-aware techniques that explicitly discourage modality bias in the model objective.

7. References

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

[2] Gao, Ruohan and Oh, Tae-Hyun, and Grauman, Kristen and Torresani, Lorenzo, "Listen to look: Action recognition by previewing audio," in *CVPR*, 2020.

[3] Y. Cheng, R. Wang, Z. Pan, R. Feng, and Y. Zhang, "Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[4] G. Sterpu, C. Saam, and N. Harte, "Attention-based audio-visual fusion for robust automatic speech recognition," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ser. ICMI '18. Association for Computing Machinery, 2018.

[5] A. Rouditchenko, A. Boggust, D. Harwath, B. Chen, D. Joshi, S. Thomas, K. Audhkhasi, H. Kuehne, R. Panda, R. Feris *et al.*, "Avlnet: Learning audio-visual language representations from instructional videos," *arXiv preprint arXiv:2006.09199*, 2020.

[6] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[7] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," *Advances in neural information processing systems*, vol. 29, 2016.

[8] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in *European conference on computer vision*. Springer, 2016, pp. 801–816.

[9] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," *Advances in neural information processing systems*, 1997.

[10] Y. Chen, J. Bi, and J. Z. Wang, "Miles: Multiple-instance learning via embedded instance selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1931–1947, 2006.

[11] Q. Zhang and S. Goldman, "Em-dd: An improved multiple-instance learning technique," *Advances in neural information processing systems*, vol. 14, 2001.

[12] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *2009 IEEE Conference on computer vision and Pattern Recognition*. IEEE, 2009, pp. 983–990.

[13] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*. PMLR, 2018, pp. 2127–2136.

[14] O. Maron and A. L. Ratan, "Multiple-instance learning for natural scene classification." in *ICML*, vol. 98. Citeseer, 1998, pp. 341–349.

[15] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[16] F. Wan, C. Liu, W. Ke, X. Ji, J. Jiao, and Q. Ye, "C-mil: Continuation multiple instance learning for weakly supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[17] N. Gonthier, S. Ladjal, and Y. Gousseau, "Multiple instance learning on deep features for weakly supervised object detection with extreme domain shifts," *Computer Vision and Image Understanding*, 2022.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[20] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018.

[21] 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)*. IEEE, 2015.

Appendix

A. Qualitative Analysis

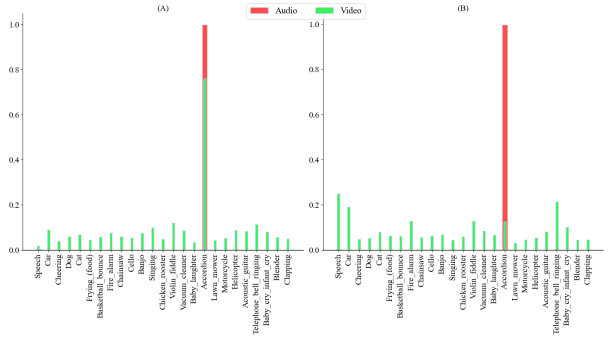

Figure 4: *Example 1: Comparison of audio-visual probabilities of a video segment. (A) $A_{cross} + V_{cross}$ (B) $A_{cross} + V_{self}$. Audio ground truth $\{Accordion\}$, Visual ground truth $\{\}$.*

We saw $A_{cross} + V_{self}$ outperforming the $A_{cross} + V_{cross}$ model quantitatively. To explore how $A_{cross} + V_{self}$ performs well, we analyze some instances qualitatively. In figure 4, no events occur in the visual modality, and the event $Accordion$ occurs in the audio modality. But, due to cross attention from audio modality, visual modality might have received spurious signals. As shown in the figure 4 (A) both audio and visual modalities predict the event $Accordion$. In figure 4 (B), $A_{cross} + V_{self}$ assigns less probability for visual modality producing correct results.

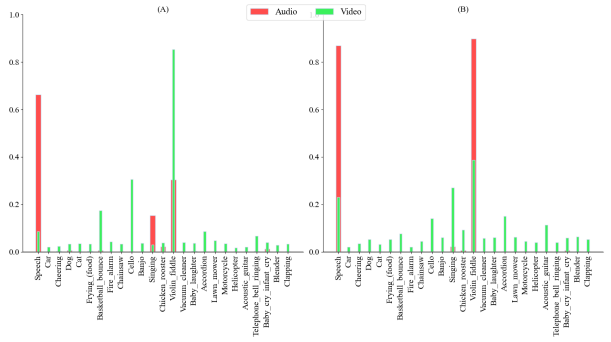

Figure 5: *Example 2: Comparison of audio-visual probabilities of a video segment. (A) $A_{cross} + V_{cross}$ (B) $A_{cross} + V_{self}$. Audio ground truth $\{Speech, Violin\ fiddle\}$, Visual ground truth $\{Speech\}$.*

Figure 5 represents probabilities for a video with audio events $\{Speech, Violin\ fiddle\}$ and visual events $\{Speech\}$. In figure 5 (A), cross attention from audio confuses visual modality that leads to incorrect prediction of the event $Violin\ fiddle$ that occurs only in audio modality. $A_{cross} + V_{self}$(figure 5 (B)) increase audio probability and reduces visual probability for event $Violin\ fiddle$. We can also see in figure 5 (B) that $A_{cross} + V_{self}$ improves the audio and visual probability of event $Speech$ aiming to get closer to the ground truth.

Recall, we said it is plausible that audio events can occur in the background and have no visual cue in the video segment

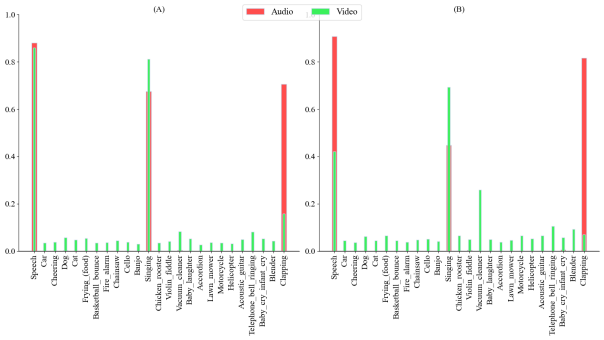

Figure 6: *Example 3: Comparison of audio-visual probabilities of a video segment. (A) $A_{cross} + V_{cross}$ (B) $A_{cross} + V_{self}$. Audio ground truth $\{Speech, Clapping\}$, Visual ground truth $\{\}$.*

(telephone ringing example). One such instance is represented using figure 6. The video is of an interview at some press conference where the interviewer is not in the frame. On the interviewee's response (not in the frame), people clap, and the interviewer proceeds to ask another question leading to no visual events. We have $\{Clapping, Speech\}$ as audio events. In fig. 5 (A), $A_{cross} + V_{cross}$ incorrectly predicts event speech for visual modality with high probability. $A_{cross} + V_{self}$ (fig. 6 (B)) placates this issue as no cross attention from audio is present to misguide. It also increases audio and diminishes visual probability for event $Clapping$. As audio is noisy, the event $Singing$ is also predicted. $A_{cross} + V_{self}$ reduces the probabilities for both the modalities and succeeds to bring audio probability below the $classification\ threshold$ (0.5) for the event $Singing$.

B. Comparison of losses with variants of label smoothing

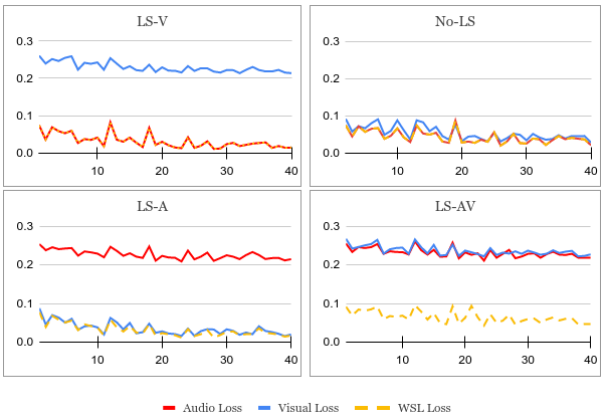

Figure 7: *Comparison of audio loss $\mathcal{L}_a$, visual loss $\mathcal{L}_v$, WSL loss $\mathcal{L}_{wsl}$ with label smoothing variants (No-LS, LS-A, LS-V, LS-AV).*

We compare how losses change with label smoothing. As discussed earlier, label smoothing challenges the model and adds more terms in cross-entropy loss as ground truth is a real-valued vector. This increases the loss noticeably. As seen in

the figure 7, adding label smoothing to a modality causes an increase in the loss for that particular modality. $\mathcal{L}_{wsl}$ is less in all variations of label smoothing since the ground truth is a multi-hot vector. Fig. 7 gives a high-level view of losses and hints at how $\mathcal{L}_{wsl}$ redistributes audio-visual attention weights to be minimal and remain in the same loss range. It aligns with the explanation that label smoothing is the root of the skewness in audio-visual attention distribution, validating our analysis.