Boosting Positive Segments for Weakly-Supervised Audio-Visual Video Parsing

Kranthi Kumar Rachavarapu  
Indian Institute of Technology Madras, India  
kranthi.rachavarapu@gmail.com

Rajagopalan A. N.  
Indian Institute of Technology Madras, India  
raju@ee.iitm.ac.in

Abstract

In this paper, we address the problem of weakly supervised Audio-Visual Video Parsing (AVVP), where the goal is to temporally localize events that are audible or visible and simultaneously classify them into known event categories. This is a challenging task, as we only have access to the video-level event labels during training but need to predict event labels at the segment level during evaluation. Existing multiple-instance learning (MIL) based methods use a form of attentive pooling over segment-level predictions. These methods only optimize for a subset of most discriminative segments that satisfy the weak-supervision constraints, which miss identifying positive segments. To address this, we focus on improving the proportion of positive segments detected in a video. To this end, we model the number of positive segments in a video as a latent variable and show that it can be modeled as Poisson binomial distribution over segment-level predictions, which can be computed exactly. Given the absence of fine-grained supervision, we propose an Expectation-Maximization approach to learn the model parameters by maximizing the evidence lower bound (ELBO). We iteratively estimate the minimum positive segments in a video and refine them to capture more positive segments. We conducted extensive experiments on AVVP tasks to evaluate the effectiveness of our proposed approach, and the results clearly demonstrate that it increases the number of positive segments captured compared to existing methods. Additionally, our experiments on Temporal Action Localization (TAL) demonstrate the potential of our method for generalization to similar MIL tasks.

1. Introduction

Event detection and localization in videos is an important task for video understanding. Event localization using visual frames has attracted a lot of attention [49, 50, 36, 35, 22, 18, 4, 41, 28, 33, 23, 13]. These methods rely only on visual cues and ignore a crucial modality - audio, which plays an integral part in human perception. To address this, there has been increased attention on audio-visual event detection [39, 43], where an event could occur in either of the modalities. Furthermore, audio and visual modalities could aid each other in better event localization.

In this paper, we explore the problem of weakly-supervised Audio-Visual Video Parsing (AVVP) task, where the goal is to detect and localize events using only video-level event labels. Such a formulation is attractive as it forgoes the need for expensive and tedious fine-grained labeling. However, this is a challenging problem due to the absence of segment-level labels during training and the requirement to process multi-modal (audio-visual) data in unconstrained videos with varying scene content.

Most previous works [38, 43, 15, 20, 26, 3] use instance-level MIL technique [12] with attentive pooling to model the video-level labels from the segment-level predictions. Such models are then trained to optimize for the video-level labels. These models are then used to identify the positive segments from segment-level predictions. While recent models have shown promising results, they often fail to identify all positive segments of an event due to the MIL-based objective as it implicitly prioritizes the most discriminative segments that satisfy weak supervision constraints. Consequently, the model selects only a subset of the most discriminative incomplete set of positive segments, which are sufficient for correctly classifying the video.

Recent work by Wu et.al [43] focuses on reducing the uncertainty due to the absence of the modality labels by estimating them from a model trained with video-level weak labels. Inspired by this, we focus on improving temporal localization by capturing all of the positive segments. To achieve this, we explicitly model the number of positive segments in a video (z) and optimize this along with the MIL objective. We show that the number of positive segments in a video follows the Poisson binomial distribution over segment-level event probabilities, which can be computed exactly. We can use the weak-supervision constraint that a positive video must contain at least one positive segment (z ≥ 1) to provide supervision for z and train the model. However, this approach also faces issues similar to that of other MIL-based techniques in accurately localizing events temporally. To overcome the challenge of weak la-
belts, we propose an iterative optimization approach using the Expectation-Maximization (EM) algorithm by modeling the number of positive segments \( (z) \) as the latent variable. In E-Step, we employ the trained model to estimate the number of positive segments in a video. In the M-Step, we optimize for the video-level labels through classification. Here, we propose the MIL objective using our Poisson binomial formulation with weakly-supervised constraints. We iteratively optimize for the model parameters and estimate the minimum number of positive segments in a video to improve the performance of the task.

We fix our network architecture to that of the HAN [38] and show that our carefully-designed training strategy yields performance gains. We evaluate our approach on the weakly-supervised AVVP task and show that our proposed method achieves state-of-the-art performance. Here, our method achieves improvement in terms of recall and precision, which indicates that it is able to better capture the positive segments. Moreover, while architectural changes may constrain a method to the particular task at hand, innovative training strategies can generalize to multiple related tasks. To validate this, we evaluate our approach on the Temporal Action Localization (TAL) task, which involves a substantially higher number of instances per video compared to AVVP. We achieve state-of-the-art results, which show that our proposed approach can generalize effectively to other related tasks as well. Our contributions are:

- We propose to explicitly model and maximize the number of positive segments in a video to improve localization under weak supervision. We show that the number of positive segments follows the Poisson binomial distribution and can be computed exactly from segment-level probabilities in a fully differentiable manner.
- Given the absence of explicit supervision on the number of positives within a video, we propose an EM-based optimization and iteratively optimize for the proposed Poisson binomial-based MIL loss to boost the total number of positive segments using a model under weak supervision.
- Our experiments on the AVVP task show that our proposed approach consistently performs favorably over the state-of-the-art methods on various metrics. Additionally, our experimentation on TAL demonstrates its potential for generalization to similar tasks.

2. Related Work

Temporal Action Localization (TAL) aims at localizing actions in visual frames within a video. Many approaches \([49, 50, 36, 35, 22, 18, 4]\) have been proposed to solve this task under full supervision. Recently, many weakly-supervised approaches have been proposed to alleviate the cost of frame-level annotation. Attention-based methods \([41, 28, 33, 23, 13]\) focus on selecting keyframes with high action probability by learning attention mechanisms with additional constraints. Others aim to identify key frames by exploiting the complementary nature of RGB and flow modalities \([47, 45]\), or using the most discriminative parts \([37, 51, 27]\) of a video.

Audio-Visual Event Localization. One way to extend the TAL problem is to require the model to reason about the multiple modalities of information in videos (the visual and audio streams) as well. Audio-Visual Event Localization (AVEL) \([39]\) task aims to localize events that are both audible and visible. Several approaches are proposed \([44, 52, 21, 31, 32]\) to solve this problem in weak supervision. These models implicitly assume that audio and visual modalities are always correlated and temporally aligned. These approaches model temporal dependencies of audio and video segments using LSTMs or attention mechanisms to fuse features from multiple modalities.

Audio-Visual Video Parsing task also aims at localizing events in a multimodal setting \(i.e.\) using both audio and visual streams. In contrast to AVEL, the task of audio-visual video parsing (AVVP) \([38]\) aims at localizing events that are either audible, visible or both, and classifying them into known categories under weak supervision. Tian et al. \([38]\) proposed a Multimodal Multiple Instance learning (MMIL) based hybrid attention network to capture temporal, unimodal, and cross-modal context simultaneously. Lamba et al. \([15]\) efficiently utilize cross-modal information, along with self-supervised and adversarial training to learn better representations for improved event localization. Wu et al. proposed an improvement over \([43]\) by generating reliable modality labels and optimizing for them using MMIL approach. Lin et al. method \([20]\) exploits both the common and diverse event semantics across videos to identify audio or visual events by exploring event co-occurrence across modalities. Recently Mo et al. \([26]\) explicitly modeled semantic-aware grouping to learn discriminative multi-modal subspaces. In contrast to these approaches, we propose to improve the temporal localization of events by modeling the total number of positive segments within a video. Then, an iterative refinement strategy is introduced to boost the total number of positive segments within a video, starting with a model trained on weak-supervision constraints.

3. Weakly-Supervised Event Prediction

We first describe the problem formulation \((\S 3.1)\) and then describe the standard instance-level MIL approach for solving this problem \((\S 3.2)\). We describe our proposed Poisson binomial distribution formulation and our training strategy in \(\S 3.3, \S 3.4, \S 3.5\). We then show in \(\S 3.6\) that our proposed approach is an EM algorithm. Finally, we discuss how our proposed formulation relates to many of the existing MIL techniques in \(\S 3.7\).
Y are then aggregated by MIL pooling. More formally, the each instance is classified first. These classification scores

In weakly-supervised AVVP, for each instance, we only have access to the corresponding video-level event label vector \( Y = [Y_0, Y_1, \ldots, Y_C] \in \{0, 1\}^C \), where \( Y_c = 1 \) if any of the segments in the video contains \( c \)-th event, otherwise \( Y_c = 0 \). These weak labels only indicate whether an event occurred in the given video or not. During evaluation, we need to identify segment-level labels \( \{(y_t^a, y_t^v)\}_{t=1}^N \). Also, note that more than one event can occur in a video, i.e. \( \sum_c Y_c \geq 1 \).

### 3.2. Overall approach and Architecture

We adopt an instance-level MIL approach [12] where each instance is classified first. These classification scores are then aggregated by MIL pooling. More formally, the video level-label \( \hat{P} \in [0, 1]^C \) is generated for a video

\[ X = \{x_t^m\}_{t=1}^N, \forall m \in \{a, v\} \] with \( T \) segments as,

\[ \hat{P} = \sigma_{x \in X} \left[ g_\theta(f_\theta(x)) \right] \] (1)

where, \( f_\theta(\cdot) \) is a feature extractor, \( g_\theta(\cdot) \) is a segment classifier, \( \sigma[\cdot] \) is an MIL-pooling operator, and \( x \) is a audio/visual segment in the video \( X \).

In our work, we adopt the Hybrid Attention Network (HAN) architecture of Tian et al. [38] to implement Eq. 1. Here, the feature extractor \( f_\theta(\cdot) \) consists of Pre-trained audio CNN (\( \Phi_a \)), visual CNN (\( \Phi_v \)) followed by a two-stream cross-modal transformer (\( T_{av} \)). The segment classifier \( g_\theta(\cdot) \) is a linear classifier. And, \( \sigma[\cdot] \) is an attentive Multi-modal MIL (MMIL) pooling operator. An overview of our method is shown in Figure 1. Please refer to the Supplemental (§S1) for architecture details. During training, the binary cross-entropy (CE) loss between the predicted video-level event probability vector \( \hat{P} \) and the weak video-level label \( Y \) is minimized as,

\[ \mathcal{L}_{\text{MIL}} = \text{CE}(\hat{P}, Y). \] (2)

During inference, we obtain segment-level predictions for each segment \( x_t^m \) in a video and threshold them i.e. \( y_t^m = \mathbb{I}[g_\theta(f_\theta(x_t^m)) \geq 0.5] \) where \( \mathbb{I}[\cdot] \) is an indicator function.

Such instance-level MIL approaches enable in finding key segments using the segment classifier. The general weakly-supervised MIL-based formulation for event localization optimizes for a set of most discriminative segments during training and fails to capture all positive segments because of a lack of fine-grained supervision. To address this
limitation, we propose to boost the total number of positive segments captured by the model in a video.

3.3. Modelling the number of positive segments using Poisson Binomial Distribution

In AVVP task, the target $y_t^\tau(c) \in \{0,1\}$ is a binary for each class in AVVP task. Bernoulli distribution is a natural choice for modeling binary classification in ML, where the outcomes are classified as success or failure. This formulation facilitates estimating the success probabilities associated with each label. Therefore, for any given video, we model the segment-level event probabilities as Bernoulli distribution with the success probability of $\hat{p}_t^m(c) \in [0,1]$ $\forall t \in [T]$, $m \in \{a, v\}$, $c \in [C]$ for binary classification problem. Thus, the label distribution of all the segments of an event-$c$ are independent and non-identical Bernoulli random variables, as each segment has a different success probability $\hat{p}_t^m(c)$. To simplify the discussion, we will disregard the event subscript $c$ as we describe our modeling for a class $c$ without any loss of generality. Therefore, for the remainder of the discussion, $\hat{p}_t^m$ will imply $\hat{p}_t^m(c)$ and $\hat{Y}$ will imply $Y_c \in \{0,1\}$ for the event $c$ under consideration.

Let $z$ be a random variable (RV) denoting the number of positive segments with the event in a video i.e. $z = \sum_{l \in t,m} \hat{Y}_l^m$, where $\hat{Y}_l^m \sim \text{Bernoulli}(\hat{p}_t^m)$ indicates whether $t$-th segment in modality-$m$ has the event or not. The RV $z$ follows the Poisson Binomial distribution, which can be computed exactly from the segment probabilities $\{\hat{p}_t^m\}_{i=1}^N$, $m \in \{a, v\}$ as,

$$\hat{P}_z(k; N) = \frac{1}{N^{k+1}} \sum_{l=0}^N e^{-ik\omega} \left[ \prod_{l=0}^N \left( 1 - \hat{p}_l^m + \hat{p}_l^m e^{i\omega} \right) \right]$$

where, $\omega = \frac{2\pi}{NT}$. Here, $\hat{P}_z(k; N)$ is the probability mass function of $z$ and gives the probability of exactly $k$ positive segments and $N-k$ negative segments in a video with $N$ segments. Please refer to Supplementary (§S2) for the derivation. For ease of reference, we define an abbreviated form of Eq. 3 as $\hat{P}_z(k; N) = \text{PoiBin}(\{\hat{p}_l^m\}_{\forall t,m})$. Therefore, the distribution of RV $z$, which indicates the number of positive segments, can be modeled explicitly from the segment-level event probabilities. Eq. 3 can be efficiently implemented using 1D-IDFT. Since Eq. 3 is differentiable, it can be effectively utilized in the loss function to enable end-to-end training of the model. The code for implementing Poisson Binomial distribution from segment probabilities is available on our project page.

3.4. Poisson Binomial based MIL formulation

We propose a novel Poisson binomial MIL-pooling operator by modeling the number of positive segments ($z$) in a video. In contrast to attentive-MIL pooling, our proposed pooling explicitly models the total number of positive segments and thereby aids in improving temporal localization.

In weakly-supervised MIL setup, the only constraint on the segment-level label is that at least one segment (or a portion of segments) in each positive video is positive, and all segments in the negative video are negative. Therefore, the final video-level event probability is computed from the Poisson binomial distribution as,

$$\hat{P} = \sum_{k \geq \tau} \hat{P}_z(k; N)$$

where $\tau$ is a hyper-parameter indicating the minimum number of positive segments in a video. Eq. 4 describes our choice of the pooling operator $\sigma[.]$ in Eq. 1. Unlike the previous methods that use either pseudo-segment labels [1, 47, 23] or only optimize for weak labels [4, 28, 33, 23, 13, 38, 43, 15, 26] through attentive pooling, we propose to use this $z$ as an intermediate level of supervision for MIL-formulation. We hypothesize that $z$ provides a more informative signal than weak-label while being less noisy than pseudo-segment labels.

Under the weak-supervision constraints, a positive video must contain at least one positive segment. Therefore, we initialize $\tau = 1$ for all videos. We minimize the binary cross-entropy loss between predicted video-level event probability vector $\hat{P}$ and weak video-level label $\hat{Y}$, given by,

$$\mathcal{L}_{\text{PoiBin}} = \text{CE}(\hat{P}, \hat{Y})$$

This is equivalent to optimizing the video-level labels when there are at least $\tau$ positive segments. As we do not have access to the ground-truth $\tau$ during training, optimizing for $k \geq \tau$ includes all possible label assignments. For $\tau = 1$, this approach also faces the same issues as other MIL-based approaches in identifying all the positive segments. Therefore, we use the trained model to get a better estimate of the minimum number of positive segments in a video. We estimate this dynamic threshold for each video clip as,

$$\tau^* = \begin{cases} \arg \max_k \hat{P}_z(k; N), & \text{if } \hat{Y} = 1 \\ 1, & \text{otherwise} \end{cases}$$

Using this new threshold $\tau^*$, we recompute the video-level probability (Eq. 4) and optimize loss (Eq. 5) to retrain the model. The threshold $\tau^*$ is computed separately for each event $c$ since the number of positive segments may differ for each event. The complete pseudo-code is given in Algorithm 1. Our proposed approach can be interpreted as an Expectation-Maximization (see §3.6). Consequently, all the convergence guarantees of the EM extend to our proposed iterative approach. We also propose an efficient

\[\text{https://github.com/KranthiKumarR/poiBin}\]
way of adapting this approach for multi-class classification in Supplementary (§S3) and employ it in our experiments on TAL task in §5.

3.5. Regularization through Data Augmentation

The Poisson binomial based MIL formulation provides additional flexibility for better estimating the video-level probability, which may not be possible with the attention-based MIL formulations. Consider two training videos \(X^i\) and \(X^j\) with \(N\) segments each, and their corresponding video-level weak label sets \(Y^i\) and \(Y^j\). If we create a synthetic video \([X^i \; X^j]\) with 2\(N\) segments, we have three possible cases for any event, and each of them offers a better bound on the threshold \(\tau\), as described below.

- \(Y^i \neq \emptyset \) \& \(Y^i \cap Y^j \neq \emptyset\): If both of the videos contain an event, then the event probability for the synthetic video is \(\hat{P} = \sum_{k=2}^{2N} \hat{P}_z(k; 2N)\), where \(\tau = \tau_i^* + \tau_j^*\).
- \(Y^i \neq \emptyset \) \& \(Y^i \cap Y^j = \emptyset\): Here, only one of the two videos is positive. In such cases, we can get a much better estimate of the video-level event probability as \(\hat{P} = \sum_{k=2}^{2N} \hat{P}_z(k; 2N)\) with \(\tau = \max(\tau_i^*, \tau_j^*)\).
- \(Y^i = \emptyset \) \& \(Y^i \cap Y^j = \emptyset\): Here, none of the videos contain the event. Thus, the video level event probability is \(\hat{P} = \sum_{k=2}^{2N} \hat{P}_z(k; 2N)\) with \(\tau = 1\).

Therefore, during training, we train using the synthetic videos \([X^i \; X^j]\) with new thresholds described above to better estimate the video-level event probabilities. Here, we minimize the following loss:

\[
L_{\text{Aug}} = CE(\hat{P}, Y^i + Y^j) \tag{7}
\]

In summary, we train our model following the Algorithm 1 using the following loss function: \(L = L_{\text{MIL}} + L_{\text{PoiBin}} + L_{\text{Aug}}\). During inference, we predict the segment-level probabilities \(\hat{p}_i(c) = q_\theta(f_\theta(x^n_i)), \forall c\), for each segment \(x^n_i\) and threshold it to detect all events as \(y^n_i(c) = 1[\hat{p}_i(c) \geq 0.5]\), where \(1[\cdot]\) is an indicator function.

3.6. Relation with the Expectation-Maximization

The proposed iterative approach can be reinterpreted as an Expectation-Maximization (EM) approach. To see this, consider a video \(X = \{x_t\}_{t=1}^T\) containing \(T\) segments with the video label \(W\). In our proposed approach, we first predict the segment labels \(Y\) and use them to compute the distribution of the number of positive segments \(Z\). Here \(X, W\) are the observed variables, and \(Z\) is the latent variable. Therefore, we have the following graphical model:

\[
P_\theta(X, Y, Z, W) = P(X)P_\theta(Y|X)P_\theta(Z|Y)P_\theta(W|Z) \tag{8}
\]

where, \(P_\theta(Y|X)\) is the segment classifier, and \(P_\theta(W|Y, Z)\) is some MIL pooling operator. Then, to learn the model parameters \(\theta\) from weakly-supervised data, we adopt an EM-based learning strategy. We alternate between the E-step and M-steps by optimizing for the evidence lower bound (ELBO) on the observed data log-likelihood \(\log P(X, W)\) as,

\[
\log P(X, W) \geq \mathbb{E}_{Q(Z|X, W)} \log \frac{P_\theta(X, Z, W)}{Q(Z|X, W)} \tag{9}
\]

where, \(Q(\cdot)\) is any posterior distribution on latents. This is a result of Jensen’s inequality and is tight if and only if \(Q(Z|X, W)\) equals the true posterior \(p(Z|X, W)\).

**E-Step.** The purpose of E-Step is to estimate the posterior on latent \(Z\) given access to the latest model parameters \(\theta^t\). For a given \(X, W\), the latent variable can be estimated by

\[
z^* = \arg \max_z P_\theta^t(Z|X, W) \tag{10}
\]

Here, we can further decompose the posterior using Eq. 8 as \(P_\theta^t(Z|X, W) = P(Z|Y)P_\theta^t(Y|X, W)\). By plugging in our proposed Poisson binomial modeling (§3.3) for \(P(Z|Y)\) and the pre-trained segment-classifier (§3.2) for \(P(Y|X, W)\), we can arrive at our proposed approach of estimating dynamic threshold \(\tau^*\) (Eq. 6) for latent \(z\).

**M-step.** The M-step is to learn the model parameters \(\theta\) by optimizing the ELBO from Eq. 9. By ignoring the terms that do not depend on model parameters, the objective of the M-step is,

\[
\mathcal{L} \simeq \mathbb{E}_{Q(Z|X, W)} \log p_\theta(W|X, Z) \tag{11}
\]
This is equivalent to optimizing the classification performance of the video classifier given $z^*$. Therefore, we can rewrite this as,

$$L = CE(p_\theta(W|X, Z = z^*), W).$$  \hspace{1cm} (12)

In our approach, we implement the video classifier $p_\theta(W|X, Z = z^*)$ using the Poisson binomial-based MIL pooling (Eq. 4). Thus, our proposed loss (Eq. 5) is the M-step ELBO objective under weak-supervision constraints.

3.7. Relation with Previous MIL Methods

The proposed Poisson binomial distribution generalizes many of the existing MIL techniques for obtaining video-level probabilities from segment probabilities. The Noisy-OR (NOR) model \cite{25, 48} estimates the bag-level probability as $P(\sum_{t,m} y_{i} \geq 1) = 1 - \prod_{t,m} (1 - \hat{P}_t^m)$. Using the Poisson binomial distribution, this is exactly $\sum_{k \geq 1} \hat{P}_z(k; N)$. The max pooling based bag-level probability ($\hat{P} = \max_{t,m} \hat{P}_t^m$) is included in $\hat{P}_z(k = 1; N)$. The average-pooling based bag-level event probability ($\hat{P} = \frac{1}{T} \sum_{t,m} \hat{P}_t^m$) is equivalent to optimizing for the expected success of the distribution $\hat{P}_z(k; N)$.

4. Experiments

4.1. Experimental setup

**LLP dataset.** We conduct our experiments on The Look, Listen and Parse (LLP) dataset \cite{38} which consists of 11849 YouTube videos of 10 seconds duration, labelled into 25 event categories. These videos are unconstrained and consist of a wide variety of scene content including daily activities, music performances, vehicle sounds etc. We use 10,000 videos with weak labels (only video-level labels) for training. The rest of the 1,849 fully-annotated videos (with segment-level labels) are used for validation and testing. We use the standard train-val-test split from the dataset.

**Evaluation Metrics.** Following previous works \cite{43, 38, 15}, we use F1-scores on all types of events (audio, visual and audio-visual) as evaluation metrics. These metrics are computed both at the segment level and event level. To compute segment-level metrics, segment-level predictions are evaluated. Event level metrics are computed by computing F1-score on positive consecutive snippets in the same event with mIoU = 0.5 as the threshold. In addition, we also evaluate the overall audio-visual scene parsing performance of our method by computing aggregated results, i.e., “Type@AV” and “Event@AV”. Specifically, Type@AV computes averaged audio, visual, and audio-visual event evaluation results, while Event@AV computes the F1-score considering all audio and visual events for each sample rather than directly averaging results from different events.

**Implementation Details.** For all experiments, we sub-sample the audio stream to 16 KHz and visual frames are processed at 8 fps. We use the same feature extractors as the baselines \cite{43, 38, 15}. We use ResNet-152 \cite{8} pretrained on ImageNet and R(2+1)D-18 \cite{40} pretrained on Kinetics-400 as visual feature extractors to generate 512 dimensional feature. Audio features of 128-dimension are extracted from VGGish model \cite{9} pretrained on AudioSet \cite{6}. Our model is trained using Adam optimizer with a mini-batch of 16 and a learning rate of 3e−4 for 20 epochs.

### 4.2. Quantitative and Qualitative Comparisons

We compare our method with the following baselines on weakly-supervised event detection methods - AVE \cite{39} and AVSDN \cite{19}. These approaches are designed for AVEL \cite{39} task, which only focuses on identifying audio-visual events. To adopt these methods for AVVP, additional audio and visual branches are introduced \cite{38}. We also compare with the following models proposed for weakly supervised AVVP task - HAN \cite{38}, Lambda et al. \cite{15}, MA \cite{42}, Lin et al. \cite{20}, MGN \cite{26}, and JoMoLD \cite{3}. These approaches use cross-modal and self-attention and train with MIL setting along with additional losses. We refer to the results reported in \cite{38, 43, 15, 3} for quantitative comparison. For qualitative comparison, we generate results by retraining the authors’ publicly available code with default hyper-parameter settings.

**Quantitative Results.** For a fair comparison, all of these models are trained on the LLP dataset with the same data split. Table 1 shows the quantitative comparison of our method with the baseline methods. Our approach achieves an average improvement of 0.6 percentage points over the SOTA on the F1 score metric. This indicates that our method is able to localize and detect events more accurately, leading to an increase in both precision and recall. Given that the task under consideration is weakly supervised with no fine-grained supervision coupled with severe label imbalance for most of the events \cite{15}, this improvement is significant for this challenging task. We also show in §5 that our method is more robust and achieves more stable
results across different datasets and experimental settings. 

**Qualitative Analysis.** We present some qualitative results in Figure 2. Here, we compare our method with HAN [38], MA [43], JoMoLD [3] methods on one example and report the final parsing results. This video contains a musical scene with events "Singing" and "Banjo" occurring in both audio and visual modalities. While all three baseline methods (HAN, MA, JoMoLD) fail to localize "Banjo" in the visual modality, our method localizes when Banjo is completely visible (0 – 2 seconds). Even on the Visual-Singing event, our method localizes much better than MA, while HAN and JoMoLD fail to localize this event completely. In audio event localization, our method and JoMoLD perform better than MA and HAN by localizing the audio-singing event completely. On Audio-Banjo event, our method struggles to localize the event in its entirety when compared to JoMoLD. Overall, our proposed approach achieves better parsing results than baseline methods. These results indicate that our proposed approach of boosting positive segments in a video, using a model trained with weak-supervision constraints alone, helps in better scene parsing. We report a few more qualitative results in Supplementary.

**On the convergence of number of positive segments.** The parameter $\tau$ in Eq. 4 indicates the minimum number of positive segments ($z$) in a given video. Since we do not have any supervision on the number of positive segments, we iteratively estimate $\tau$ from a pre-trained model and then retrain the model using the proposed Poisson binomial based MIL loss. We empirically show that the model learns a better estimate of $\tau$ with the proposed iterative training. For this, we start with a model pre-trained with MA [43] setup and refine using our proposed approach. We report the histo-

<table>
<thead>
<tr>
<th>N</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>80</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>0.2</td>
<td>0.21</td>
<td>0.21</td>
<td>0.22</td>
<td>0.33</td>
<td>0.59</td>
<td>2.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Analysis of runtime of our proposed Poisson Binomial pooling. $N$ indicates the number of segments; Time in milliseconds.

![Figure 2: Audio-Visual Video Parsing results of our method with HAN [38], MA [43], JoMoLD [3] on one video.](image)

<table>
<thead>
<tr>
<th>Losses</th>
<th>Audio</th>
<th>Visual</th>
<th>AV</th>
<th>Type@AV</th>
<th>Event@AV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_{\text{MIL}}$</td>
<td>61.4</td>
<td>51.9</td>
<td>55.9</td>
<td>52.2</td>
<td>53.1</td>
</tr>
<tr>
<td>$L_{\text{PoiBin}}$ &amp; 60.0</td>
<td>52.9</td>
<td>60.1</td>
<td>56.3</td>
<td>54.7</td>
<td>46.8</td>
</tr>
<tr>
<td>$L_{\text{AttMIL}}$</td>
<td>61.1</td>
<td>53.3</td>
<td>62.6</td>
<td>58.9</td>
<td>57.7</td>
</tr>
<tr>
<td>$L_{\text{PoiBinMIL}}$ &amp; 60.6</td>
<td>52.5</td>
<td>58.4</td>
<td>54.6</td>
<td>54.0</td>
<td>46.6</td>
</tr>
<tr>
<td>$L_{\text{PoiBin-pooling}}$</td>
<td>61.9</td>
<td>53.6</td>
<td>62.9</td>
<td>59.5</td>
<td>57.7</td>
</tr>
<tr>
<td>$L_{\text{PoiBin-pooling}}$ &amp; 63.1</td>
<td>54.1</td>
<td>63.5</td>
<td>60.4</td>
<td>57.7</td>
<td>51.5</td>
</tr>
</tbody>
</table>

Table 3: Ablation studies to evaluate various components of the loss function.
over the baselines without augmentation (first and second rows). Overall, our proposed approach complements the existing method and improves the performance of the model, indicating the effectiveness of our proposed formulation for a weakly-supervised task.

**Effect of \( \tau \):** We iteratively estimate \( \tau \) from a pre-trained model, and then retrain the model using our proposed loss. We report results in Table 4. By fixing \( \tau = 1 \) (the lowest possible value, as \( \tau = 0 \) indicates the event does not occur), our algorithm performance is comparable to the baseline method. This is expected as our Poisson binomial formulation indicates that there is at least one positive segment in the video, which is exactly similar to the weak-supervision constraints. Therefore our formation is equivalent to the baseline model. A large \( \tau \) implies that most of the segments contain the event, even when they may not. This would be detrimental to the training as such a setup is equivalent to training with many (noisy) false positive labels. Our experiments also show that this generally degrades the performance, as shown for the case when \( \tau = 10 \). On the other hand, in our proposed approach, we initialize \( \tau \) to 1, which adheres to the weak-supervision constraints, and iteratively estimate a new \( \tau \) as described in §3.4. Our experiments also show that this approach improves the performance (§4.2), and our model estimates better after each training epoch (Figure 3).

**Usefulness of modeling \( z \) as latent over \( y \).** Here, we perform ablation experiments by considering different choices for the latent variable in our formulation. Results reported in Table 5 indicate that our proposed modeling with \( z \) as the latent variable is more robust than using segment-level pseudo-label \( y \) based formulation. These results support our initial hypothesis (in §3.3) that modeling the number of positive segments \( z \) as latent variables provides a more informative signal than weak labels while being less noisy than pseudo-segment labels (\( y \)).

## 5. Generalization to other tasks

The proposed Poisson binominal formulation for the weakly supervised AVVP task does not make any task-specific assumptions. Therefore, we investigate the generalizability of our approach to other similar weakly supervised tasks. We perform preliminary experiments on the Temporal Action Localization (TAL) that aims to localize the start and end timestamps of action instances and recognize their categories simultaneously in untrimmed videos. We experiment with THUMOS14 dataset [11], which consists of videos with 100’s of frames belonging to 20 action categories. We adopt Poisson binomial based MIL pooling to model this multi-class classification setup. The training details, along with architecture information, are available in Supplementary (§S3).

We report the results for this setup in Table 6. We evaluate in terms of mean Average Precision (mAP) with different temporal Intersection over Union (IoU) thresholds, which is denoted as mAP@\( \alpha \) where \( \alpha \) is the threshold. Our model, trained with our proposed Poisson binomial-based MIL approach from §3.3, performs better than the current state-of-the-art model DELU [2].

![Figure 3: Effect of multi-stage training in estimating a better \( \tau \), which indicates a minimum number of positive segments in each video.](image)

<table>
<thead>
<tr>
<th>Table 4: Effect of ( \tau ) on performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau )</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>refine</td>
</tr>
</tbody>
</table>

Table 5: Ablation studies to evaluate the modeling of \( z \) as the latent variable (+PL \( y \)) vs as the latent variable (+PL \( z \)).

<table>
<thead>
<tr>
<th>Methods</th>
<th>Audio</th>
<th>Visual</th>
<th>Audio-Visual</th>
<th>Type@AV</th>
<th>Event@AV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>60.0</td>
<td>32.9</td>
<td>50.7</td>
<td>46.8</td>
<td>30.6</td>
</tr>
<tr>
<td>+PL ( y )</td>
<td>58.5</td>
<td>53.3</td>
<td>56.5</td>
<td>53.1</td>
<td>45.2</td>
</tr>
<tr>
<td>+PL ( z )</td>
<td>63.1</td>
<td>54.1</td>
<td>63.4</td>
<td>57.7</td>
<td>51.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 6: Results on Temporal Action Localization task on THUMOS14 dataset [11].</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>Wang et al. [12]</td>
</tr>
<tr>
<td>W-TALC [30]</td>
</tr>
<tr>
<td>EM-MIL [24]</td>
</tr>
<tr>
<td>Nguyen et al. [29]</td>
</tr>
<tr>
<td>HAM-Net [14]</td>
</tr>
<tr>
<td>FTCL [5]</td>
</tr>
<tr>
<td>UGCT [46]</td>
</tr>
<tr>
<td>DCC [16]</td>
</tr>
<tr>
<td>DGCNN [34]</td>
</tr>
<tr>
<td>Li et al. [17]</td>
</tr>
<tr>
<td>Huang et al. [10]</td>
</tr>
<tr>
<td>ASM-Loc [7]</td>
</tr>
<tr>
<td>DELU [3]</td>
</tr>
<tr>
<td>OURS</td>
</tr>
</tbody>
</table>

The proposed Poisson binomial formulation for the weakly supervised AVVP task does not make any task-specific assumptions. Therefore, we investigate the generalizability of our approach to other similar weakly supervised tasks. We perform preliminary experiments on the Temporal Action Localization (TAL) that aims to localize the start and end timestamps of action instances and recognize their categories simultaneously in untrimmed videos. We experiment with THUMOS14 dataset [11], which consists of videos with 100’s of frames belonging to 20 action categories. We adopt Poisson binomial based MIL pooling to model this multi-class classification setup. The training details, along with architecture information, are available in Supplementary (§S3).

We report the results for this setup in Table 6. We evaluate in terms of mean Average Precision (mAP) with different temporal Intersection over Union (IoU) thresholds, which is denoted as mAP@\( \alpha \) where \( \alpha \) is the threshold. Our model, trained with our proposed Poisson binomial-based MIL approach from §3.3, performs better than the current state-of-the-art model DELU [2].
Our model also shows more significant improvements at high threshold metrics tIoU=0.7, which implies that our action proposals are more complete. Note that, to use our proposed approach, we need to have a model that predicts per-frame probabilities only. Therefore, these results further indicate that our proposed approach has the potential to be further optimized and integrated with other MIL-based techniques to achieve even better results.

6. Limitations

Although our proposed approach performs better than state-of-the-art on AVVP and TAL tasks, there are a few limitations, which we discuss here. We can guarantee convergence by reinterpreting our proposed iterative approach as an EM algorithm. But it is known that EM algorithms converge to a locally optimal solution. From Table 1, performance gains in the audio-visual event (when an event co-occurs in both audio and visual modalities) are slightly lower than in audio events and visual events. One reason for this is that we are not modeling audio-visual events explicitly. We can potentially overcome this by modeling AV events, such as exploiting the temporal correlation between audio/visual modalities and using the sequential ordered nature of data. Learning better feature representations by employing explicit constraints on feature similarity may mitigate this issue. The LLP dataset suffers from severe label imbalance and strong label correlations [15], i.e., a set of labels co-occur more often than others. Our proposed strategy is not designed to address this issue.

7. Conclusions

In this paper, we proposed a method for improving temporal localization in a weakly-supervised audio-visual video parsing task. To this end, we proposed to model the total number of positive segments (z) in a video. We showed that this follows Poisson binomial distribution, which can be computed exactly from segment-level event probabilities. Since we do not have explicit supervision on the number of positive segments in a video, we proposed an iterative algorithm. We first estimate the minimum number of positive segments (τ) in a video and then optimize for the model parameters using the proposed Poisson binomial-based MIL loss. We also proposed a data-augmentation method that aided in improving the performance. Our proposed approach can be interpreted as an EM algorithm, which provides convergence guarantees. Experiments on the LLP dataset demonstrate that our proposed approach outperforms the state-of-the-art, validating the efficacy of our approach in improving the localization capacity under weak supervision. Additionally, our experiments on Temporal Action Localization demonstrate its potential for generalization to similar MIL tasks.

Acknowledgement

Support from Institute of Eminence (IoE) project No. SB22231269EEETW0005001 for Research Centre in Computer Vision is gratefully acknowledged. The first author thanks Lokesh Bommisetty for valuable discussions in the early stages of the work.

References


[34] Haichao Shi, Xia-Yu Zhang, Changsheng Li, Lixing Gong, Yong Li, and Yongjun Bao. Dynamic graph modeling for


