Completeness Modeling and Context Separation for Weakly Supervised Temporal Action Localization

Daochang Liu¹, Tingting Jiang¹, Yizhou Wang¹,²,³
¹NELVT, Cooperative Medianet Innovation Center, School of EECS, Peking University
²Peng Cheng Lab, ³Deepwise AI Lab
{daochang, ttjiang, yizhou.wang}@pku.edu.cn

Abstract

Temporal action localization is crucial for understanding untrimmed videos. In this work, we first identify two underexplored problems posed by the weak supervision for temporal action localization, namely action completeness modeling and action-context separation. Then by presenting a novel network architecture and its training strategy, the two problems are explicitly looked into. Specifically, to model the completeness of actions, we propose a multi-branch neural network in which branches are enforced to discover distinctive action parts. Complete actions can be therefore localized by fusing activations from different branches. And to separate action instances from their surrounding context, we generate hard negative data for training using the prior that motionless video clips are unlikely to be actions. Experiments performed on datasets THUMOS’14 and ActivityNet show that our framework outperforms state-of-the-art methods. In particular, the average mAP on ActivityNet v1.2 is significantly improved from 18.0% to 22.4%. Our code will be released soon.

1. Introduction

Temporal action localization is an important visual task with potential applications in video surveillance [42], video summarization [28], skill assessment [16], and others. The goal is to predict not only the action label but also the start and end times of each action instance from untrimmed videos. Fully supervised temporal action localization has witnessed remarkable progress recently [39, 48, 15, 9, 37, 51, 47, 8, 2, 31]. However, precisely annotating the temporal extent of action instances is labor-intensive and time-consuming, which undermines fully supervised approaches in real-world large-scale scenarios. Therefore the weakly supervised setting, where only video-level category labels are available during training, is more practical and draws increasing attention from the community. This paper works on temporal action localization with such weak labels.

Most of existing weakly supervised methods [45, 33, 38, 36, 52] fall into the framework of Multiple Instance Learning (MIL) [54]. In this framework, a video is treated as a bag of sampled frames or snippets and fed into video-level classification networks. Action instances are then localized using the Class Activation Sequence (CAS) [38], a 1D temporal classification score sequence of each action.

Compared with its fully-supervised counterpart, weakly supervised temporal action localization introduces two new challenges, dubbed as action completeness modeling and action-context separation. The two issues have not been well considered before and have notably limited the per-
formance. The first challenge is how to detect action instances in their entirety without full annotations. An action is intrinsically a temporal composition of elementary sub-actions [20], which are supposed to be wholly included in the prediction without omission. In the fully supervised setting, whether an action is complete is learned directly from the ground truth of temporal boundaries. In contrast, when weakly supervised, the lack of fine-grained annotations complicates the completeness modeling since the localization task is now formulated as the classification on video level. Identifying one fragment of an action is sufficient for video-level classification but not for segment-level localization. For example, the action Soccer Penalty can be roughly divided into two sub-actions, i.e., Player Shooting and Ball Flying. The activation only on the more discriminative Player Shooting part is adequate to classify the video, but leaving the Ball Flying part false negative for localization. The second challenge of action-context separation is how to distinguish action instances from their context with weak labels. Action instances of the same class are usually surrounded by visually similar clips, such as action Billiards commonly enclosed by commentary clips with a static pool table in the screen. Such clips appear together with the true actions in most videos, thus termed as context in this paper. Context clips are different from ordinary background clips in terms of distributions in videos. Context clips co-occur with the true action in most cases and are not involved in videos of other action categories, while background clips are class-independent and distributed randomly. For this reason, context clips can be regarded as hard negatives. Video-level classifiers learn the correlation between videos with the same tag and discover their common contents, which unfortunately include not only the common action (e.g., Billiards) but also the common context (e.g., the static pool table). We argue that action-context separation is inherently difficult with weak supervision, unless employing the prior knowledge about actions.

To tackle the two issues respectively, we propose a multi-branch network architecture and a hard negative data generation scheme. To model action completeness, feature sequences extracted from input videos are fed into a network with multiple classification branches in parallel. A diversity loss is devised to ensure the dissimilarity between the Class Activation Sequences output by different branches, such that each branch is trained to locate distinct fractions of an action. Thereupon, as the example in Fig. 1, complete actions can be retrieved by aggregating activations from multiple branches. The class activation is then pooled over time with temporal attention, producing a video-level category distribution. We compute its cross-entropy with the ground truth, i.e., the standard MIL loss, which is minimized along with the diversity loss to learn the network parameters. As for action-context separation, we develop a simple yet effective strategy for mining hard negatives using the prior that actions should be of motions. We search for stationary clips in the training videos, as illustrated in Fig. 1. Then pseudo videos are generated using static clips and labeled with a new background class. Such a strategy can assist the model in rejecting the common context, as long as some hard negatives are included in the generated pseudo videos.

On two benchmark datasets THUMOS’14 [21] and ActivityNet [6], the proposed method outperforms the state-of-the-art methods, demonstrating the effectiveness of handling the two problems. In summary, our contributions are three-fold: 1) A multi-branch network with diversity loss is proposed to model action completeness. 2) A hard negative video generation scheme is devised to separate common context. 3) Our method achieves superior results on two benchmark datasets.

2. Related Works

Action recognition on trimmed videos has been extensively studied in the past. Early methods were mainly based on hand-crafted features [26, 43, 34]. In recent years, various deep networks have been proposed, such as two-stream networks [40, 46], LSTM [12], 3D ConvNets [41], I3D [7], and others [23, 44, 53]. Please refer to recent surveys [1, 3, 22, 19] for a detailed review.

Fully supervised temporal action localization approaches have been largely based on the proposal-plus-classification paradigm [39, 37, 51, 15, 5, 9, 47, 8, 31], where temporal proposals are generated first and then classified. Other categories of methods have also been studied, such as those based on single-shot detectors [4, 30] or sequential decision-making process [48, 2]. Given full annotations, the proposal-plus-classification methods usually filter out the common context at the proposal stage via a binary actionness classifier. As for the completeness modeling, Zhao et al. [51] used a structural temporal pyramid pooling followed by an explicit binary classifier to determine whether an instance is complete. Hou et al. [20] clustered video segments of an action into different sub-actions and then detected the whole action as an ordered sequence of sub-actions. Yuan et al. [49] structured an action into three components, i.e. the start, middle and end, to model its temporal evolution. But they all require full annotations. Other works on spatial-temporal action detection [17] and video temporal segmentation [27] are beyond our scope.

Weakly supervised temporal action localization algorithms mostly belong to the Multiple Instance Learning (MIL) [54]. Wang et al. [45] proposed a framework called UntrimmedNet composed of a classification module and a selection module, based on which a sparsity regularization was later introduced in [33]. Paul et al. [36] used a co-activity similarity loss to enforce the feature similarity between the localized instances of the same class. Instead of
The proposed multi-branch network consists of a feature extraction module, a feature embedding module, a multi-branch classification module, and a temporal attention module. In the classification module, multiple branches are trained with the diversity loss to discover different action parts.

Regarding the first challenge, there are two prior works that attempt to model action completeness. Hide-and-Seek [25] hid random frame sequences while training to force the network responsive to multiple relevant parts. However, randomly hiding frames does not always guarantee the discovery of new parts and also disrupts the training process. Recently, Zhong et al. [52] trained a series of classifiers iteratively to find complementary pieces, by erasing the predictions of predecessor classifiers from input videos. The major drawback with this approach is the extra time cost and computational expense to train multiple classifiers. The other challenge of action-context separation is inherently tricky and remains unexplored in the literature. The selection module in UntrimmedNet [45] is intended for eliminating irrelevant background clips rather than semantically related context. Researchers have also studied action localization with other types of weak supervision, such as movies scripts [13], ordered action lists [11], and web images [14].

**Diversity Loss** was initially introduced for text embedding [32] to extract different aspects of a sentence. Recently, Li et al. [29] utilized the diversity loss to deal with occlusions in person re-identification. Unlike previous works, we use diversity loss to model the action completeness, which is of different specification and motivation.

### 3. Proposed Method

In this section, we present the proposed methodology for weakly supervised temporal action localization. The input is an untrimmed video with varying frames in length. Let a one-hot vector \( y \in \{0, 1\}^{C+1} \) denote the ground truth video-level category label, where \( C \) is the number of action classes and \( C + 1 \) represents the newly added background class. During the test time, the output for each test video is a set of localized action instances \( \{(s_i, e_i, c_i, q_i)\} \), where \( s_i \) and \( e_i \) denote the start time and end time of the \( i^{th} \) detection, \( c_i \) represents the predicted category, and \( q_i \) denotes the confidence score.

#### 3.1. Hard Negative Video Generation

Weakly supervised models are inclined to confuse the true action with its surrounding context, i.e., the hard negatives, especially when the context appears in a majority of videos of that class. We observe that it is the motion that makes an action different from its context. An action must involve the movement of human or other subjects, while the context clips are allowed to stay static (e.g., the static pool table). Therefore, we generate hard negative training data using stationary video clips, labeling them with a new background class. Concretely, for each video in the training set, we compute its optical flow using the TV-L1 algorithm [50] and average the intensity in every frame. Since the motion magnitude differs among action categories and even some actions exhibit minor motion, a small predefined percentage \( \rho \) of video frames with the lowest optical flow intensity are picked out from each video individually. Frames picked from the same video are then concatenated into a pseudo video, which is labeled with the background class and added to the training set. We expect that the generated videos par-
entially include the hard negatives and drop hints for the proposed network to deal with the challenge of action-context separation. Details and the generated video examples are provided in the supplementary material.

3.2. Multi-branch Network

To model action completeness, a multi-branch network is designed such that each branch focuses on different action parts. As shown in Fig. 2, the proposed multi-branch network consists of a feature extraction module, an embedding module, a multi-branch classification module, and a temporal attention module, which are detailed as follows.

Feature extraction module. Given an input video, a snippet-wise feature sequence \( X \in \mathbb{R}^{T \times D} \) is first extracted by pre-trained deep networks, where \( T \) denotes the number of snippets and \( D \) denotes the feature dimensions. The extracted feature sequence provides a high-level representation of the appearance and motion of the input video and is fed into the next layers in the network. Note that \( T \) and \( D \) depend on the choice of feature extraction network. In experiments, we focus on two off-the-shelf models, namely UntrimmedNet [45] and I3D [7].

Embedding module. The feature extraction module is followed by an embedding module. Since the features may not be originally trained for weakly supervised action localization, a task-specific embedding of the features is desired. We utilize a temporal convolutional layer followed by a ReLU activation layer to embed the features:

\[
\phi(X) = \max (W_{emb} \ast X + b_{emb}, 0) \tag{1}
\]

where \( \ast \) represents the convolution operation, \( W_{emb} \) and \( b_{emb} \) are the weights and biases of temporal filters, \( \phi(X) \in \mathbb{R}^{T \times F} \) denotes the learned embedding, and \( F \) is the number of filters. Temporal convolutions integrate the information from neighboring time locations, enabling the network to capture the temporal structure. The embedded feature sequence is then passed to subsequent layers.

Multi-branch classification module. In this module, \( K \) classification branches are organized in parallel to discover complementary pieces of an action. Each branch inputs the embedded feature sequence into a temporal convolutional layer and outputs a sequence of classification scores:

\[
A^k = W^k_{cls} \ast \phi(X) + b^k_{cls} \tag{2}
\]

where \( A^k \in \mathbb{R}^{T \times (C+1)} \), \( W^k_{cls} \) and \( b^k_{cls} \) are respectively the classification scores, the filter weights, and filter biases in the \( k^{th} \) branch. Then each \( A^k \) is passed through a softmax along the category dimension, yielding class distributions at each time location:

\[
\hat{A}^k = \text{softmax}(A^k) \tag{3}
\]

where \( \hat{A}^k \) is referred to as Class Activation Sequence (CAS). For clarity, we use the bar notation in this paper to indicate it has undergone a softmax. For action completeness modeling, we expect the CASes from multiple branches differ from each other. However, without constraint, the branches could lazily concentrate on a single same action part. To avoid such degenerate cases where branches give identical results, a diversity loss based on cosine similarity is imposed on the CASes:

\[
\mathcal{L}_{div} = \frac{1}{Z} \sum_{c=1}^{C+1} \sum_{i=1}^{K} \sum_{j=i+1}^{K} \frac{\hat{A}^c_i \cdot \hat{A}^c_j}{\|\hat{A}^c_i\| \|\hat{A}^c_j\|} \tag{4}
\]

which is the cosine similarities between the CASes from every two branches, averaged over all branch pairs and action categories. \( \hat{A}^c_i \in \mathbb{R}^T \) means the activation sequence for class \( c \) from the \( i^{th} \) branch and 

\[
Z = \frac{1}{2} K (K - 1) (C + 1)
\]

is a normalization factor. By minimizing such a diversity loss, branches are encouraged to produce activations on different action parts. Then CASes from multiple branches are averaged and passed through a softmax along the category dimension:

\[
A^\text{avg} = \frac{1}{K} \sum_{k=1}^{K} A^k \tag{5}
\]

\[
\hat{A}^\text{avg} = \text{softmax}(A^\text{avg}) \tag{6}
\]

where \( \hat{A}^\text{avg} \in \mathbb{R}^{T \times (C+1)} \) is referred to as average CAS, which combines all part activations and encodes the full action. Moreover, the softmax operation inhibits activations of action classes when the score of the background class is large, thus reducing false positives on context clips.

We empirically notice that \( A^k \) from some branches tend to be nearly all zeros while those from other branches explode, which may corrupt the training process. More importantly, if one branch dominates, the average CAS is effectively responsive to single action part instead of the whole action. From another perspective, these parallel branches can be considered to be in an adversarial relationship, competing with each other to find different discriminative action segments. It is expected that the branches are balanced to have comparable strength. Similar ideas can be seen in the training strategy of Generative Adversarial Networks [18]. Therefore we introduce another regularization term on the norms of the original score sequences without softmax:

\[
\mathcal{L}_{\text{norm}} = \frac{1}{K (C + 1)} \sum_{c=1}^{C+1} \sum_{i=1}^{K} \| A^i_{*,c} \| - \| A^\text{avg}_{*,c} \| \tag{7}
\]

which is the deviations from the norm of \( A^\text{avg} \), averaged over branches and categories. Equipped with the diversity loss and norm regularization, the multi-branch design is capable of discovering diverse action parts without full supervision and therefore modeling the action completeness.

Temporal attention module. Since the input video is untrimmed and contains irrelevant backgrounds, we utilize a temporal attention module to learn the importance of
video snippets. The attention module feeds the embedded feature sequence into a temporal convolutional layer followed by a softmax along the temporal dimension:

$$\bar{U} = \text{softmax}(W_{\text{att}} \ast \phi(X) + b_{\text{att}})$$  \hspace{1cm} (8)

where \(W_{\text{att}}\) and \(b_{\text{att}}\) are the weight parameter and bias of the temporal filter, and \(\bar{U} \in \mathbb{R}^T\) represents the sequence of learned class-agnostic attention. To get the video-level score where the attention:

$$s_c(t, \cdot) = \text{softmax}_T(\mathcal{A}^{avg}_{t, \cdot})$$

is computed:

$$\mathcal{L}_{mil} = - \sum_{c=1}^{C+1} y_c \log \bar{p}_c$$  \hspace{1cm} (9)

Finally, we combine the MIL loss with the diversity loss and norm regularization:

$$\mathcal{L}_{sum} = \mathcal{L}_{mil} + \alpha \mathcal{L}_{div} + \beta \mathcal{L}_{norm}$$  \hspace{1cm} (11)

where \(\alpha\) and \(\beta\) are coefficients. All the three components have sub-gradients at least and can be minimized using gradient descent.

3.3. Action Localization

During the test time, we leverage the trained multi-branch network to classify test videos and localize actions. Since multiple categories of actions can occur in one video, we first threshold on the video-level classification score. Given a test video, we detect action instances for each non-background category \(c\) with \(\bar{p}_c\) larger than 0.1. Then we threshold on the average CAS of class \(c\), i.e., \(\mathcal{A}^{avg}_{s,c}\), to localize action instances. Let \(\{(s_i, e_i, c, q_i)\}\) denotes the corresponding output detections. Similar to the Outer-Inner-Contrastive loss proposed in [38], we score each localized instance using the contrast between the mean activation of the instance itself and its surrounding areas:

$$q_i = m_{inner} - m_{outer} + \gamma p_c$$

$$m_{inner} = \text{mean}(\mathcal{A}^{avg}_{s_i; e_i; c})$$

$$m_{outer} = \text{mean}(\mathcal{A}^{avg}_{s_i-l_i; e_i-c; c}, \mathcal{A}^{avg}_{e_i+l_i; c})$$

where \([\cdot]\) denotes concatenation and \(l_i = (e_i - s_i)/4\) is the inflation length. The video-level score \(\bar{p}_c\) is combined as well with the coefficient \(\gamma\).

4. Experiments

In this section, we first discuss the datasets and our implementation details. Then comparisons between the proposed method and state-of-the-art approaches are presented. At last, we examine the impact of each model component by ablation studies. In the supplementary material, more experiment results are reported.

4.1. Datasets

Extensive experiments are conducted on two large-scale benchmarks: THUMOS’14 [21] and ActivityNet [6]. Videos in both datasets are untrimmed and only video-level category labels are used for training.

THUMOS’14. A subset of THUMOS’14 including 20 action classes is provided with temporal annotations and used for the localization task. Following the previous convention, we use the validation set of 200 videos for training, and the test set of 213 videos for evaluation. From the training data, 152 hard negative videos\(^1\) are generated. This dataset has a large amount of action instances per video and the length of videos varies widely.

ActivityNet. Experiments are performed on both two release versions of ActivityNet. ActivityNet1.3 covers 200 action classes and consists of 10,024 training videos, 4,926 validation videos and 5,044 testing videos, with 7323 hard negative videos generated using the training videos. We train on the training set and report results on the validation set as well as the testing set. To facilitate comparisons, we also evaluate on ActivityNet1.2, a subset of version 1.3, which has 4,819 training videos, 2,383 validation videos, 2,480 testing videos, and 100 classes. 3469 hard negative videos are generated on ActivityNet1.2. We employ the training set for training and the validation set for evaluation.

Evaluation Metrics. We follow the standard evaluation protocol and report mean average precision (mAP) at different thresholds of temporal intersection over union (IoU). The mAP values are computed using the evaluation codes provided by the datasets. All results on THUMOS’14 are averaged over three runs. The performance on ActivityNet1.3 testing set is obtained by submitting results to the evaluation server.

4.2. Implementation Details

Two deep networks with two-stream architecture are tried out for feature extraction, namely UntrimmedNet [45] and 13D [7], which are pre-trained and fixed during training. UntrimmedNet is pre-trained on ImageNet [10] and takes video snippets of 1 RGB frame and 5 stacked optical flow frames as input. 13D is pre-trained on Kinetics [7] and takes non-overlapping 16-frame chunks as input for both two streams. Video snippets are sampled every 15 frames

---

\(^1\)Please refer to the supplementary for details.
### 4.3. Comparisons with the State-of-the-art

Experimental results on THUMOS’14 testing set are shown in Table 1. Our proposed multi-branch network along with hard negative mining is compared to existing methods for weakly supervised temporal action localization, as well as several fully supervised ones. Our model outperforms previous weakly supervised methods at most IoU thresholds regardless of the choice of feature extraction network. The gain is not as substantial at higher IoU due to the observation that our model sometimes produces overly complete instances which lead to false positives. Note that AutoLoc [38] regresses temporal action boundaries for localization and therefore obtain high mAP at higher IoU thresholds, while we simply threshold on the CAS and still achieve comparable results. We argue that their method and ours can promote the performance further if combined.

Table 2 presents the results on ActivityNet1.2 validation set, and results on the validation and testing sets of ActivityNet1.3 are reported in Table 3. On both versions of this large dataset, the proposed method outperforms the state-of-the-art significantly, verifying the effectiveness of handling action completeness modeling and context separation.

#### 4.4. Ablation Studies

To analyze the contribution of each model component, we perform a set of ablation studies, with results on THUMOS’14 testing set shown in Table 4. Our best model is compared to a baseline and other configurations with each of the following components removed: 1) multi-branch design 2) hard negative generation 3) both diversity loss and norm regularization 4) only norm regularization 5) the temp-
Table 4. Ablation study results on THUMOS’14 testing set. ‘Single’ and ‘Multiple’ indicate the number of branches in the classification module, and ‘HN’ denotes that hard negative videos are used when training. ‘TA’ denotes the temporal attention module.

<table>
<thead>
<tr>
<th>Methods</th>
<th>AVG (0.1:0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours (UNT), Single + $L_{mil}$ (Baseline)</td>
<td>28.8</td>
</tr>
<tr>
<td>Ours (UNT), Single + $L_{mil}$ + HN</td>
<td>32.7</td>
</tr>
<tr>
<td>Ours (UNT), Multiple + $L_{sum}$</td>
<td>34.8</td>
</tr>
<tr>
<td>Ours (UNT), Multiple + $L_{mil}$ + HN</td>
<td>34.7</td>
</tr>
<tr>
<td>Ours (UNT), Multiple + $L_{mil}$ + $L_{div}$ + HN</td>
<td>35.6</td>
</tr>
<tr>
<td>Ours (UNT), Multiple + $L_{sum}$ + HN (No TA)</td>
<td>36.3</td>
</tr>
<tr>
<td>Ours (UNT), Multiple + $L_{sum}$ + HN (Full)</td>
<td><strong>37.4</strong></td>
</tr>
</tbody>
</table>

Figure 3. **Left:** Experiments on the branch number. **Right:** Experiments on the diversity loss weight. The average mAP at IoU thresholds from 0.1 to 0.5 is reported.

Figure 4. Class-specific results with different branch number. The optimal branch number depends on the complexity of the action. **Left:** Results of Shotput. **Right:** Results of Cliff Diving. The average mAP at IoU thresholds from 0.1 to 0.5 is reported.

<table>
<thead>
<tr>
<th>Ratio</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>15%</td>
<td>32.4</td>
<td>33.3</td>
<td>32.7</td>
<td>32.8</td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Impact of the selection ratio $\rho$ in hard negative mining. AVG indicates the average mAP at IoU thresholds from 0.1 to 0.5.

4.5. Qualitative Results

We plot several interesting examples of localized actions and corresponding CASes in Fig. 5 to show the effectiveness of tackling the two challenges qualitatively. Examples are from THUMOS’14 testing set using UntrimmedNet features. In the first example of Diving, incomplete actions displayed in red bounding boxes such as the one with only Entry into the Water part and the other one with only Standing on the Platform part are discovered by single branch but excluded from final predictions due to incompleteness. In the second example of Billiards, commentary clips (yellow boxes) semantically similar to the true actions (pink boxes) are effectively filtered out using hard negative generation. In the third example of High Jump, CASes from multiple branches are very diverse, localizing different action parts.

5. Discussion and Future Work

We devise a simple yet effective data generation scheme to separate action context, while the assumption behind might not hold in all cases. We find its effect closely related to the action class, with details in the supplementary. In future, more advanced techniques such as Generative Adversarial Networks can be applied to mine the hard negatives deeper. As for action completeness modeling, distinctive
action parts are discovered automatically by the proposed multi-branch module in an unsupervised manner. In practice, the learned action parts may not exactly correspond to semantically meaningful sub-actions. Instead, the model may capture different action modes, aspects, stages, or other underlying structures, depending on which kind of representation benefits the learning target most. Along the key idea of dividing an action into parts, there can be many potential future directions for weakly supervised temporal action localization, including but not limited to 1) using the learned part representation to understand actions or measure action complexity 2) modeling the temporal configuration of action parts 3) representing actions hierarchically to handle ambiguities or subjective annotation biases.

6. Conclusion

In this work, we identified two challenges posed by the weak supervision for temporal action localization, namely action completeness modeling and action-context separation. To handle the first challenge, a multi-branch network was proposed to find different action parts and therefore locate action instances in their integrity. Meanwhile, we mined hard negatives to handle the second issue of action-context separation. Experiments on two benchmarks showed that our framework tackles the two problems effectively and outperforms state-of-the-art methods.

Acknowledgment. This work was partially supported by National Basic Research Program of China (973 Program) under contract 2015CB351803 and the Natural Science Foundation of China under contracts 61572042, 61527804, 61625201. We acknowledge the high-performance computing platform of Peking University for providing computational resources. Thanks to Jingjia Huang for providing the optical flow data.
References


