Weakly Supervised Temporal Action Localization Using Deep Metric Learning

Ashraful Islam
Rensselaer Polytechnic Institute
islama6@rpi.edu

Richard J. Radke
Rensselaer Polytechnic Institute
rjradke@ecse.rpi.edu

Abstract

The supplementary material provides more ablation studies and results on the benchmark datasets. These are not included in the main paper due to the space limit.

1. More Ablation Studies

1.1. Ablation on Loss Weight

In the main paper, we set $\lambda = 1$ in the final loss function,

$$\mathcal{L} = \mathcal{L}_{\text{BBCE}} + \lambda \mathcal{L}_{\text{metric}} \tag{1}$$

In Table 1, we provide the performance of our model for different values of λ on the THUMOS14 dataset.

λ	IoU			
	0.1	0.3	0.5	0.7
0.2	54.9	39.3	23.0	8.2
0.6	60.6	44.2	27.1	9.3
1	**62.3**	**46.8**	**29.6**	**9.7**
1.4	61.1	45.7	27.5	9.0
1.8	60.0	43.6	24.7	8.4

Table 1: Experiments on loss weight λ

1.2. Ablation on Loss Margin α

In Table 2, we experiment with different values of the loss margin α in the triplet loss function $\mathcal{L}^c_{\text{triplet}} = [d^{+,c} - d^{-,c} + \alpha]_+$. We get the best results when $\alpha = 3$.

1.3. Ablation on Number of Segments

To show how the number of video segments affects the final accuracy, we provide results in Table 3 for different values of the maximum number of segments on THUMOS14. We find the best result when we have maximum segment length 300.

α	IoU			
	0.1	0.3	0.5	0.7
1	53.8	36.0	18.7	4.9
2	60.2	42.5	25.3	7.5
3	**62.3**	**46.8**	**29.6**	**9.7**
4	61.9	45.6	28.8	9.6
5	60.7	45.4	28.4	**9.7**
6	59.8	43.9	27.1	9.1

Table 2: Experiments on the loss margin α

Number of segments	IoU			
	0.1	0.3	0.5	0.7
50	59.9	40.5	21.5	6.5
100	57.3	38.1	20.0	6.3
200	62.0	46.5	28.6	**10.2**
300	**62.3**	**46.8**	**29.6**	9.7
500	61.1	45.7	27.9	9.1
700	60.1	43.7	26.0	8.9

Table 3: Experiments on number of segments

1.4. Ablation on Minimum Number of Videos with Similar Activity Instances per Batch

In the main paper, we use a batch size of 20 with 4 different activity instances per batch such that at least 5 videos have the same activity. In Table 4, we change the number of videos with similar activity per batch, and provide the results. Note that we keep the batch size at 20. Hence, if the number of videos with similar activity is 2, then there will be at least 10 videos with the same activity instance.

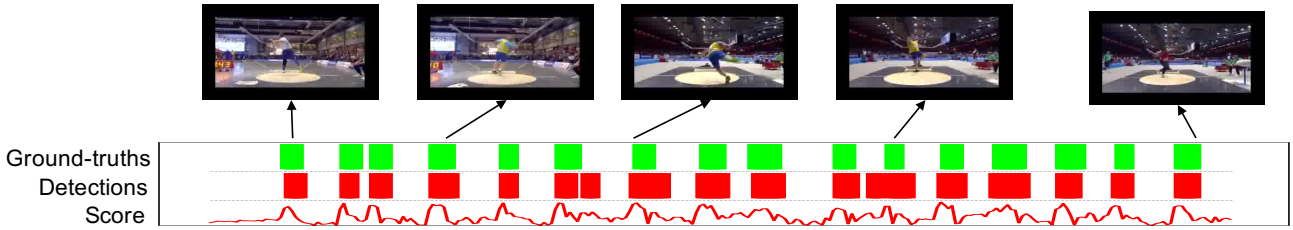
2. Additional Qualitative Results

We provide additional qualitative results in Figure 1. In Fig. 1c, our algorithm cannot differentiate Cricket Shot and Cricket Bowling as two separate activities; hence it localizes the time stamps where there is either activity involved. This is not surprising, as Cricket Shot and Cricket Bowling occur together most of the time, and there are no videos

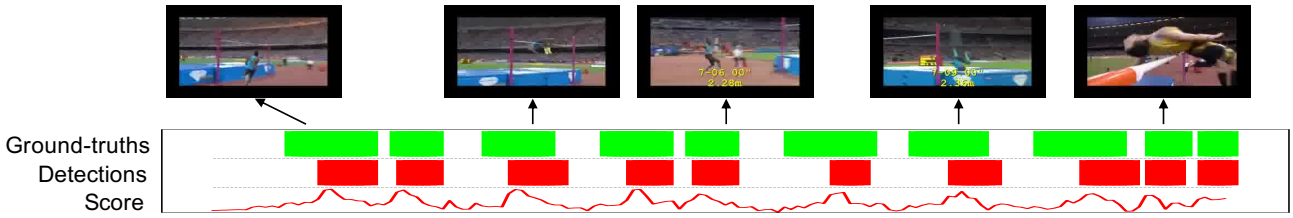
Number of videos with similar activity	IoU			
	0.1	0.3	0.5	0.7
2	57.7	41.1	23.9	7.9
4	60.5	45.5	25.9	9.0
5	**62.3**	**46.8**	**29.6**	**9.7**
10	59.9	43.7	27.4	8.5

Table 4: Experiments on number of videos with similar activity per batch

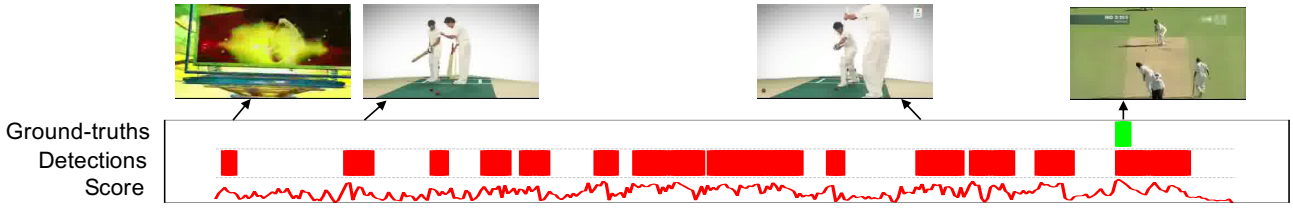
in the dataset where only the Cricket Bowling activity happens.



(a) Shotput



(b) High Jump



(c) Cricket Bowling

Figure 1: Qualitative results on THUMOS14. The horizontal axis denotes time. On the vertical axis, we sequentially plot the ground truth detection, detection score after post-processing, and class activation score for a particular activity. (d) represents a failure case for our method. In (d), we see that the model cannot differentiate Cricket Bowling from Cricket Shot, as both of these activities occur together in general. Without the ground truth localization, it is hard for a model to detect only the Cricket Bowling activity, but not Cricket Shot.