Supplement of Actionness Inconsistency-guided Contrastive Learning for Weakly-supervised Temporal Action Localization

Anonymous submission

Implementation Details

Following existing methods, we use I3D (Carreira and Zisserman 2017) model pretrained on Kinetics (Carreira and Zisserman 2017) dataset as the RGB and optical flow feature extractors. We sample continuous non-overlapping 16 frames from video as a snippet, where the features for each modal of each snippet are 1024-dimension. In the training and testing stage, we randomly sample 750 snippets for the THUMOS14 dataset, 50 snippets for the ActivityNet1.2 dataset, and 50 snippets for the ActivityNet1.3 dataset. For fair comparisons, we do not finetune the feature extractor, i.e., I3D. In the structure diagram of AICL, the f, f_R, and f_F are constructed with convolution layer and RELU activations. The output dimensions of the convolution layer are 512 for THUMOS14 and ActivityNet1.2, and 1024 for ActivityNet1.3, while the kernel size is 3. The cls is constructed with a convolution layer whose output dimension is 20 for THUMOS14, 100 for ActivityNet1.2, and 200 for ActivityNet1.3, while the kernel size is 1. The cls_R and cls_F are constructed with a convolution layer whose output dimensions are 1 for THUMOS14, ActivityNet1.2, and ActivityNet1.3, while the kernel size is 1. We set noise tolerance q = 0.7 for both datasets, and use instance selection parameters $k = T/8$ for THUMOS14, $k = T/2$ for ActivityNet1.2 and ActivityNet1.3. We set γ to be 1/3 for the A_{score} in THUMOS14 and 1/2 in ActivityNet1.2 and ActivityNet1.3. We set $K = k/20$ to choose the number of inconsistent action and background segments. We set γ_1 = 0.01, γ_2 = 5, γ_3 = 0.1 for THUMOS14, ActivityNet1.2 and ActivityNet1.3.

Visual Results

To better illustrate our method, Figure 1 and Table 1 show DETAD(Alwassel et al. 2018) analysis on THUMOS14 dataset. The false positive profiles show that AICL significantly reduces the background error compared to baseline, i.e., the percentage of background errors dropped from 40.8% to 35.03%.

In Figures 2 and 3, we visualize more qualitative results. The baseline is a model that only uses action loss and class-agnostic loss. The dotted boxes indicate areas where the baseline is clearly wrong. It can be clearly seen that compared with the baseline, our method can avoid false action activation caused by scene information, and can also suppress the interference of label-category-independent actions. Therefore, scene-related errors and action-related errors are effectively controlled by AICL.

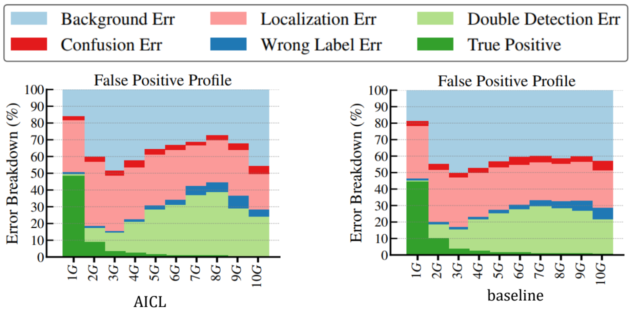


Figure 1: THUMOS-14 error analysis following DETAD(Alwassel et al. 2018). Top row breaks down false positives errors while bottom row shows false negatives by segment lengths.

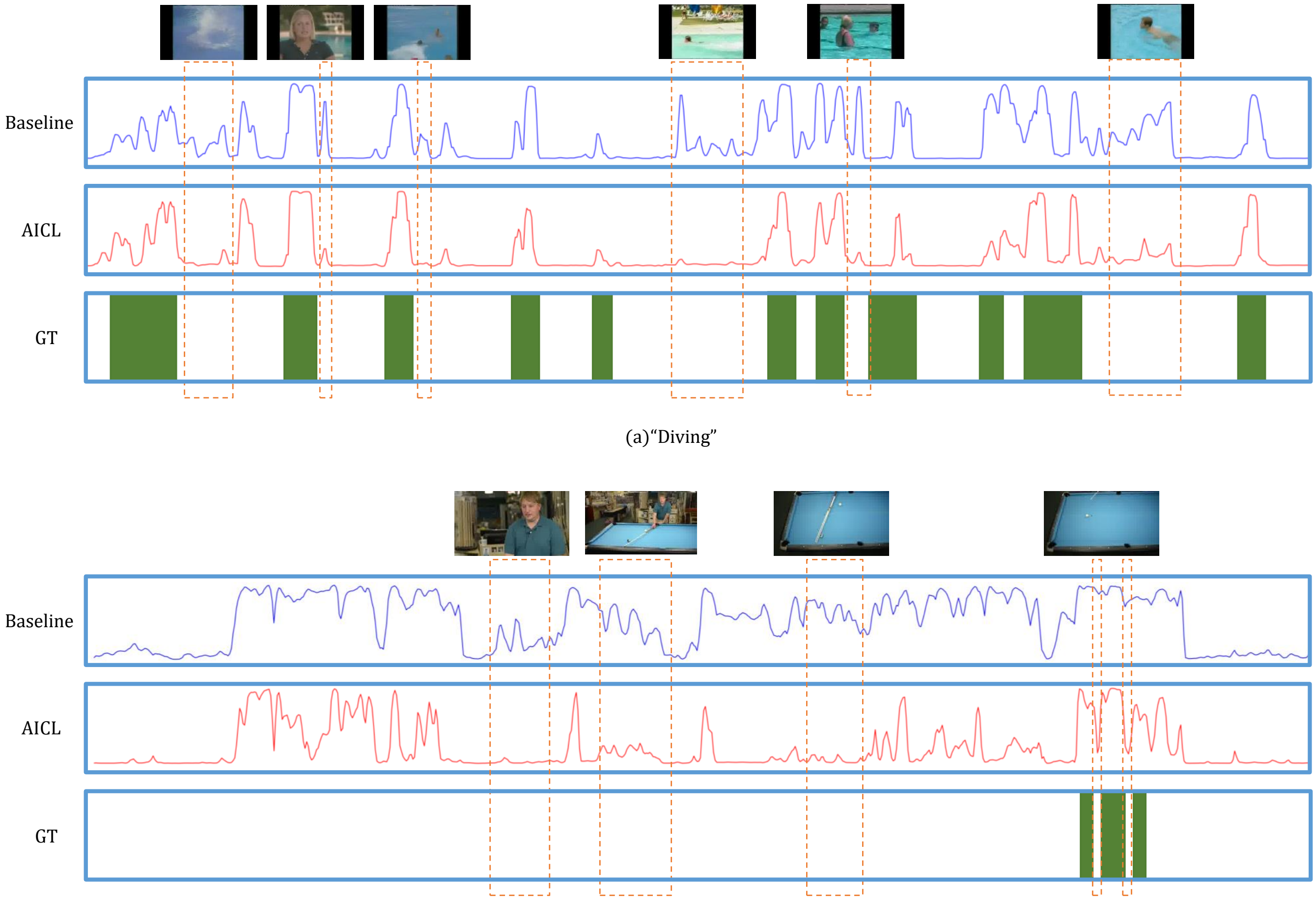
Error Type(%)	AICL	Baseline
Background Err	35.03	40.83
Confusion Err	3.21	3.63
Double Dection Err	21.40	18.57
Localization Err	29.70	26.57
Wrong Label Err	3.14	2.89
True Positive	7.56	7.50

Table 1: Percentage of various types of errors in the prediction results of AICL and baseline.

References

Alwassel, H.; Heilbron, F. C.; Escorcia, V.; and Ghanem, B. 2018. Diagnosing error in temporal action detectors. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

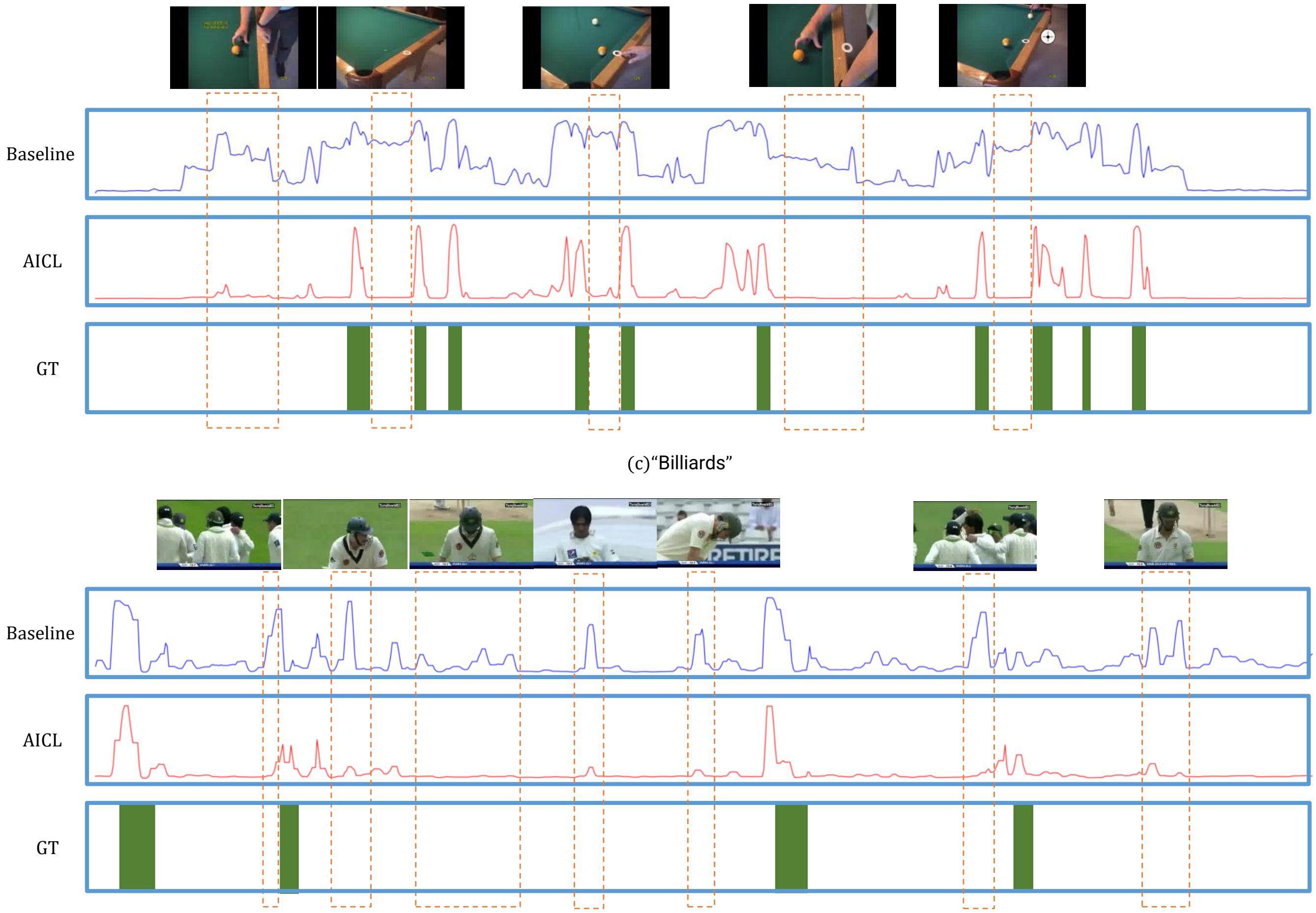
Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 6299–6308.



(a)“Diving”

(b)“Billiards”

Figure 2: Qualitative results of AICL with a baseline on THUMOS14. The curves represent the detection activation scores, and the GT is ground-truth.



(c)“Billiards”

(d)“CricketShot”

Figure 3: Qualitative results of AICL with a baseline on THUMOS14. The curves represent the detection activation scores, and the GT is ground-truth.