Supplementary Material for Weakly-Supervised Temporal Action Detection for Fine-Grained Videos with Hierarchical Atomic Actions

Zhi Li[1], Lu He[2], and Huijuan Xu[3]

[1] University of California, Berkeley, USA
zhili@berkeley.edu
[2] Tencent America, Palo Alto, USA
lhluhe@tencent.com
[3] Pennsylvania State University, University Park, USA
hkx5063@psu.edu

1 Additional Ablation Study

We additionally conduct ablation study to understand the performance impact of the number of visual concepts per class. Tab. 1 shows the detection results on FineAction [2] with different numbers of visual concepts per class. Results show that using 5 visual concepts per class achieves the best performance. And in general the HAAN model's performance is stable with respect to the number of visual concepts. The number of visual concepts does not significantly affect the detection results so long as the hierarchical visual concept modeling exists.

Since a different pooling method (Eq. 1 in the main paper) is proposed in our HAAN model's MIL framework, we conduct the ablation study to compare the proposed pooling method with other two pooling methods used in previous works. Results in Tab. 2 shows that the proposed pooling method outperforms other pooling methods in HAAN.

2 Qualitative Visualization of Action Detection Results

We visualize some example action detection results from the proposed HAAN model and compare them with the ground truth. We select correct examples as well as partially correct examples to get a holistic understanding of the HAAN model's prediction.

As shown in Fig. 1, our HAAN model is able to detect complete action instances in (a) and (b). In (c), the number of action instances is predicted correctly but with slightly mismatched action boundaries. In (d), the model combines two consecutive action instances into one action prediction, due to the challenge of recognizing the exact action boundary for fine-grained actions, especially in the weakly-supervised setting.

Table 1. Action detection results of HAAN model on FineAction dataset with different numbers of visual concepts for each fine-level action class. The avg.mAP refers to the average of the mean Average Precision (mAP) at different temporal IoU thresholds ranging from 0.5 to 0.95 with an interval of 0.05

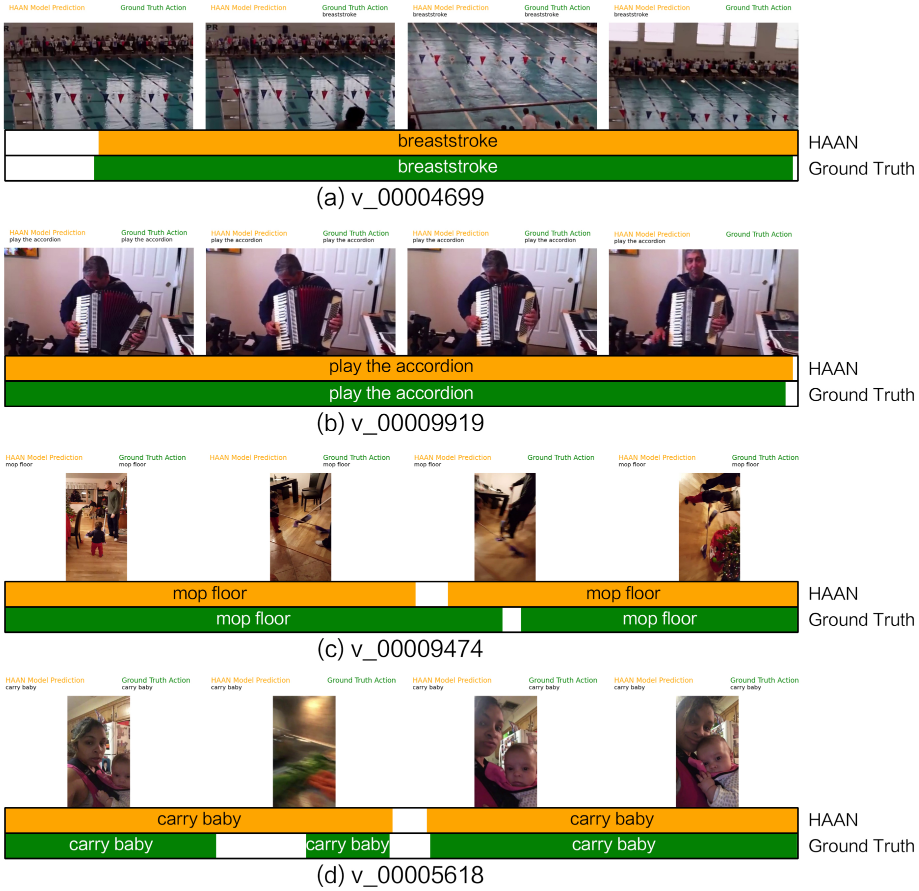
Number of visual concepts per class	mAP@ τ			avg.mAP
	0.5	0.75	0.95	
2	6.91	3.85	1.06	4.01
3	6.93	3.93	1.09	4.07
5	7.05	3.95	1.14	4.10
8	6.87	3.79	1.03	3.94
10	6.94	3.86	1.09	4.01

Table 2. mAP@0.5 results of different pooling methods for HAAN model on FineAction dataset

Pooling methods	$\mathcal{L}_{mil}$	$\mathcal{L}_{mil}+\mathcal{L}_{pseudo}+\mathcal{L}_{concept}+\mathcal{L}_{coarse}$
Our proposed pooling	5.74	7.05
K-max in [8,6,4,7,3]	5.01	6.58
Attention-based in [8,5,9,1]	4.55	6.27

References

1. Liu, D., Jiang, T., Wang, Y.: Completeness modeling and context separation for weakly supervised temporal action localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1298–1307 (2019) 2
2. Liu, Y., Wang, L., Ma, X., Wang, Y., Qiao, Y.: Fineaction: A fine-grained video dataset for temporal action localization. arXiv preprint arXiv:2105.11107 (2021) 1
3. Ma, J., Gorti, S.K., Volkovs, M., Yu, G.: Weakly supervised action selection learning in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7587–7596 (2021) 2
4. Narayan, S., Cholakkal, H., Khan, F.S., Shao, L.: 3c-net: Category count and center loss for weakly-supervised action localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8679–8687 (2019) 2
5. Nguyen, P., Liu, T., Prasad, G., Han, B.: Weakly supervised action localization by sparse temporal pooling network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6752–6761 (2018) 2
6. Paul, S., Roy, S., Roy-Chowdhury, A.K.: W-talc: Weakly-supervised temporal activity localization and classification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 563–579 (2018) 2
7. Shou, Z., Gao, H., Zhang, L., Miyazawa, K., Chang, S.F.: Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 154–171 (2018) 2
8. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 4325–4334 (2017) 2
9. Yuan, Y., Lyu, Y., Shen, X., Tsang, I., Yeung, D.Y.: Marginalized average attentional network for weakly-supervised learning. In: ICLR 2019-Seventh International Conference on Learning Representations (2019) 2



(a) v_00004699

(b) v_00009919

(c) v_00009474

(d) v_00005618

Fig. 1. Visualization of HAAN's predicted actions in FineAction dataset. (a,b) are correct detections. In (c), the HAAN model predicts the correct number of action instances, but there exists a slight mismatch in the detected time interval. In (d), the model combines two consecutive action instances into one