A Hybrid Attention Mechanism for Weakly-Supervised Temporal Action Localization (Supplementary Materials)

Ashraful Islam[1], **Chengjiang Long** [2], **Richard J. Radke** [1]

[1] Rensselaer Polytechnic Institute
[2] JD Digits AI Lab
islama6@rpi.edu, chengjiang.long@jd.com, rjradke@ecse.rpi.edu

Abstract

The supplemental material contains additional experiments, visualization and ablation studies.

Experiments

Action Classification

Table. 1 shows action classification performance of our approach in comaprison with other state-of-the-arts in THUMOS14 and ActivityNet1.2 dataset. We use classification mean average precision (mAP) for evaluation. We see that the classification performance of our approach is very competitive with the SOTAs, specially in THUMOS14 we achieve 7.2% mAP improvement over 3C-Net (Narayan et al. 2019). We also achieve very competitive performance in ActivityNet dataset. Although our approach has not been designed for video action recognition task, it's high performance in action classification reveals the robustness of our method.

Detailed Performance on ActivityNet1.2

Table. 2 shows detailed performance of our approach on ActivityNet1.2 dataset in terms of localization mAP for different IoU thresholds.

More Ablation

Fir. 1 shows ablation studies on the hyper-parameters α, β, and drop threshold γ on THUMOS14 dataset. AVG mAP is the mean mAP value from IoU threshold 0.1 to 0.7 incremented by 0.1. Fig. 1a shows the performance for different weights on sparsity loss. Without sparsity loss, the model hardly learns any localization. As α increases, localization performance increases as well, and we get the best score for $\alpha = 0.8$. Fig. 1b reveals the performance improvement for different weights on guide loss. We empirically find that $\beta = 0.8$ gives the best performance. In Fig. 1c, we see the mAP performance for different values of dropping threshold γ. Fig. 1d and Fig. 1e show the effect of video length during training for THUMOS14 and ActivityNet respectively. Note that THUMOS14 contains more denser videos with a large number of activities per video. Hence we observe that the performance increase for larger video length for THUMOS14, whereas, ActivityNet performs best for 80 length segments. Also, note that the number of segments are chosen randomly only during training. We use all segments during evaluation.

More Qualitative Examples

We show more qualitative examples in Fig. 2. In Fig. 2a, there are several occurrences of Pole Vault activity, and our method can capture most of them. We show some failure examples in Figure. 2b and Fig. 2c. In Fig. 2b, our model erroneously captures some activities as high jump. In those erroneous segments, we observe that the person tends to do a high jump activity but restrain in the end without completing the full action. The same goes for Fig. 2c. Previous WTAL approaches (Islam and Radke 2020; Paul, Roy, and Roy-Chowdhury 2018) have also shown similar issues as an inherent limitation of WTAL methods. Because of the weakly-supervised nature, we infer that some errors related to incomplete activities are inevitable.

References

Islam, A.; and Radke, R. 2020. Weakly Supervised Temporal Action Localization Using Deep Metric Learning. In *The IEEE Winter Conference on Applications of Computer Vision*, 547–556.

Lee, P.; Uh, Y.; and Byun, H. 2020. Background Suppression Network for Weakly-supervised Temporal Action Localization. In *AAAI*.

Liu, D.; Jiang, T.; and Wang, Y. 2019. Completeness Modeling and Context Separation for Weakly Supervised Temporal Action Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1298–1307.

Liu, Z.; Wang, L.; Zhang, Q.; Gao, Z.; Niu, Z.; Zheng, N.; and Hua, G. 2019. Weakly supervised temporal action localization through contrast based evaluation networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 3899–3908.

Narayan, S.; Cholakkal, H.; Khan, F. S.; and Shao, L. 2019. 3c-net: Category count and center loss for weakly-supervised action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 8679–8687.

Methods	THUMOS14	ActivityNet1.2
iDT+FV (Wang and Schmid 2013)	63.1	66.5
C3D (Tran et al. 2015)	-	74.1
TSN (Wang et al. 2016)	67.7	88.8
W-TALC (Paul, Roy, and Roy-Chowdhury 2018)	85.6	**93.2**
3C-Net (Narayan et al. 2019)	86.9	92.4
Ours	**94.1**	90.3

Table 1: Action Classification performance of our method with state-of-the-arts methods on THUMOS14 and ActivityNet1.2 dataset in terms of classification mAP.

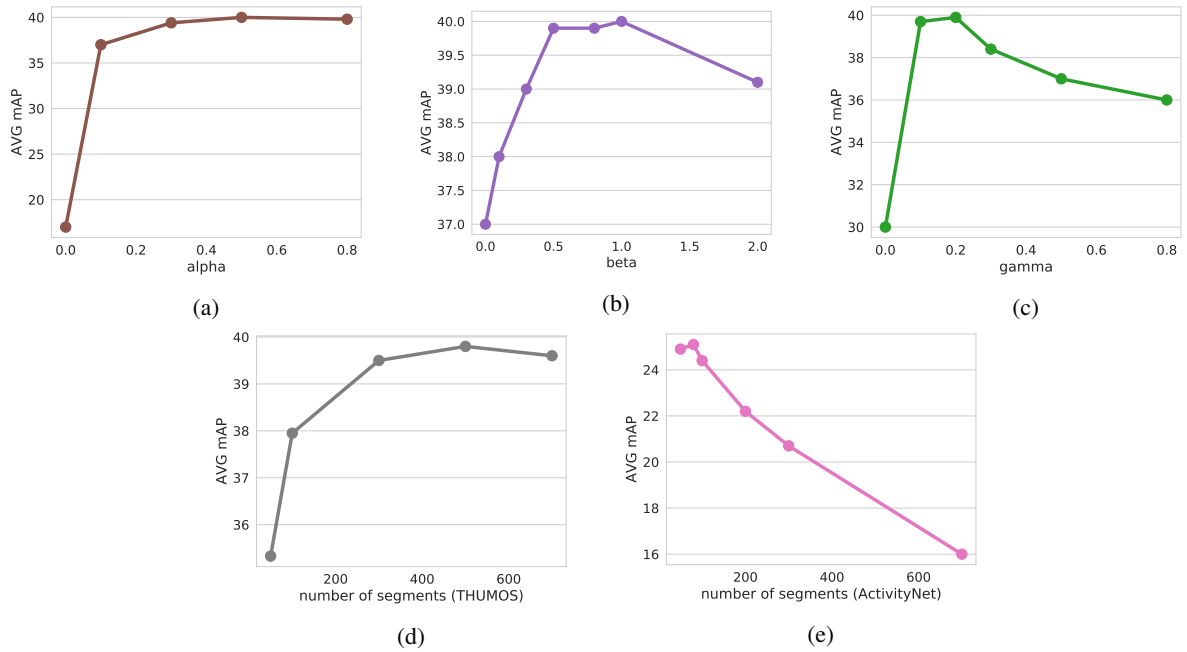


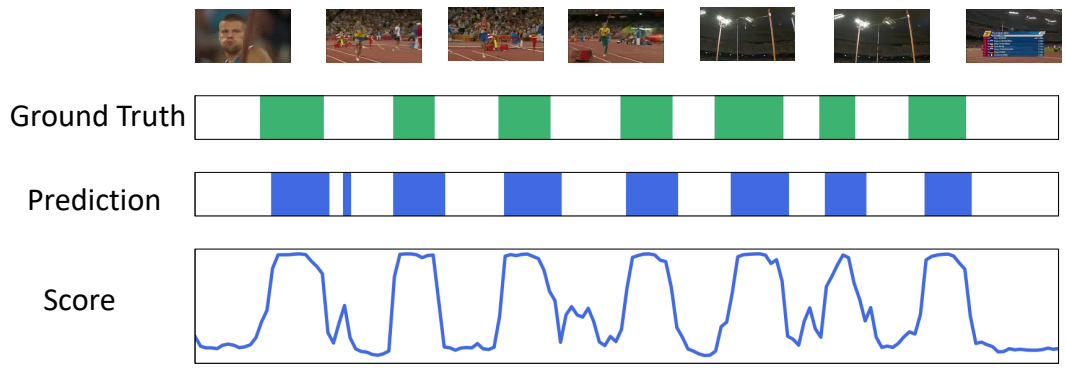
Figure 1: (a) Ablation on the weight of sparsity loss. (b) Ablation on the weight of guide loss. (c) Ablation on the drop-threshold for dropping snippets in the HAD module. (d) and (e) Ablation on the number of segments for a video during training.

Table 2: Comparison of our algorithm with other state-of-the-art methods on the ActivityNet1.2 validation set for temporal action localization.

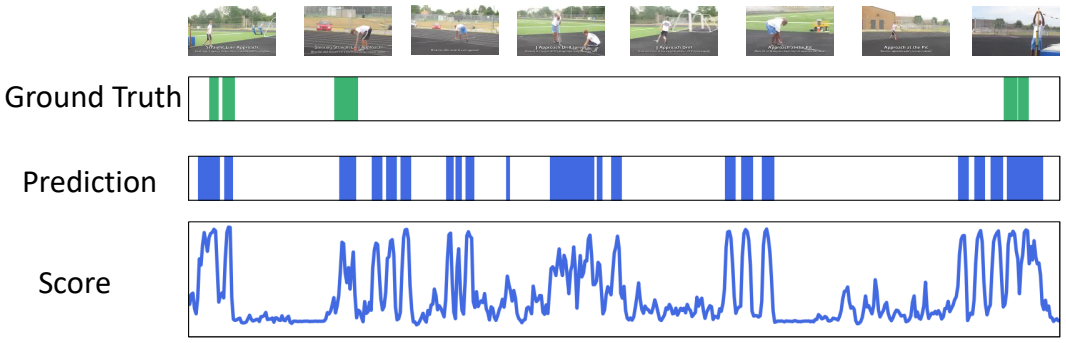
Supervision	Method	IoU										
		0.5	**0.55**	**0.6**	**0.65**	**0.7**	**0.75**	**0.8**	**0.85**	**0.9**	**0.95**	**AVG**
Full	SSN (Zhao et al. 2017)	41.3	38.8	35.9	32.9	30.4	27.0	22.2	18.2	13.2	6.1	26.6
Weak	UntrimmedNets (Wang et al. 2017)	7.4	6.1	5.2	4.5	3.9	3.2	2.5	1.8	1.2	0.7	3.6
	AutoLoc (Shou et al. 2018)	27.3	24.9	22.5	19.9	17.5	15.1	13.0	10.0	6.8	3.3	16.0
	W-TALC (Paul, Roy, and Roy-Chowdhury 2018)	37.0	33.5	30.4	25.7	14.6	12.7	10.0	7.0	4.2	1.5	18.0
	TSM (Yu et al. 2019)	28.3	26.0	23.6	21.2	18.9	17.0	14.0	11.1	7.5	3.5	17.1
	3C-Net (Narayan et al. 2019)	35.4	-	-	-	22.9	-	-	-	8.5	-	21.1
	CleanNet (Liu et al. 2019)	37.1	33.4	29.9	26.7	23.4	20.3	17.2	13.9	9.2	5.0	21.6
	Liu *et al* (Liu, Jiang, and Wang 2019)	36.8	-	-	-	-	22.0	-	-	-	5.6	22.4
	Islam *et al* (Islam and Radke 2020)	35.2	-	-	-	16.3	-	-	-	-	-	-
	BaS-Net (Lee, Uh, and Byun 2020)	34.5	-	-	-	-	22.5	-	-	-	4.9	22.2
	DGAM (Shi et al. 2020)	**41.0**	37.5	33.5	30.1	26.9	23.5	19.8	15.5	**10.8**	**5.3**	24.4
	Ours	**41.0**	**37.9**	**34.6**	**31.3**	**28.1**	**24.8**	**21.1**	**16.0**	**10.8**	**5.3**	**25.1**

Paul, S.; Roy, S.; and Roy-Chowdhury, A. K. 2018. W-TALC: Weakly-supervised Temporal Activity Localization and Classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 563–579.

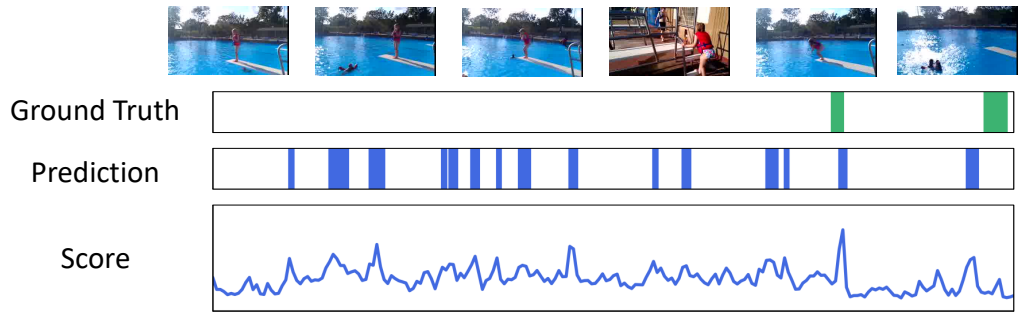
Shi, B.; Dai, Q.; Mu, Y.; and Wang, J. 2020. Weakly-Supervised Action Localization by Generative Attention



(a) Pole Vault



(b) High Jump



(c) Diving

Figure 2: Qualitative results on THUMOS14. The horizontal axis denotes time. On the vertical axis, we sequentially plot the ground truth, our predicted localization, and our prediction score. (b) and (c) represent failure examples of our approach.

Modeling.

Shou, Z.; Gao, H.; Zhang, L.; Miyazawa, K.; and Chang, S.-F. 2018. AutoLoc: Weakly-supervised temporal action localization in untrimmed videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 154–171.

Tran, D.; Bourdev, L. D.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. *2015 IEEE International Conference on Computer Vision (ICCV)* 4489–4497.

Wang, H.; and Schmid, C. 2013. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, 3551–3558.

Wang, L.; Xiong, Y.; Lin, D.; and Van Gool, L. 2017. UntrimmedNets for weakly supervised action recognition and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4325–4334.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Gool, L. V. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *ECCV*.

Yu, T.; Ren, Z.; Li, Y.; Yan, E.; Xu, N.; and Yuan, J. 2019. Temporal structure mining for weakly supervised action detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 5522–5531.

Zhao, Y. S.; Xiong, Y.; Wang, L.; Wu, Z.; Lin, D.; and Tang, X. 2017. Temporal Action Detection with Structured Segment Networks. *2017 IEEE International Conference on Computer Vision (ICCV)* 2933–2942.