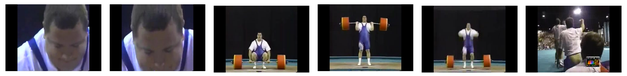
A Hybrid Attention Mechanism for Weakly-Supervised Temporal Action Localization

Ashraful Islam[1], Chengjiang Long[2], Richard J. Radke[1]
[1]Dept. Of Electrical, Computer, and Sytems Engineering, Rensselaer Polytechnic Institute; [2]JD Digits AI Lab

Problem & Background

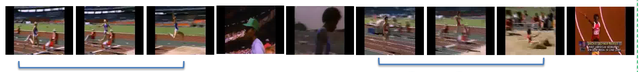
- Most current action localization methods depend on untrimmed videos with full temporal annotations
- Full temporal annotations are expensive and time-consuming
- We propose a weakly supervised temporal action localization method that only requires video-level action instances as supervision during training
- During evaluation, the model predicts both action instances and temporal localizations

Training



{ CleanAndJerk }

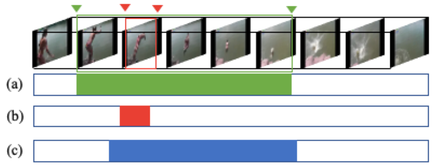
Testing



LongJump LongJump

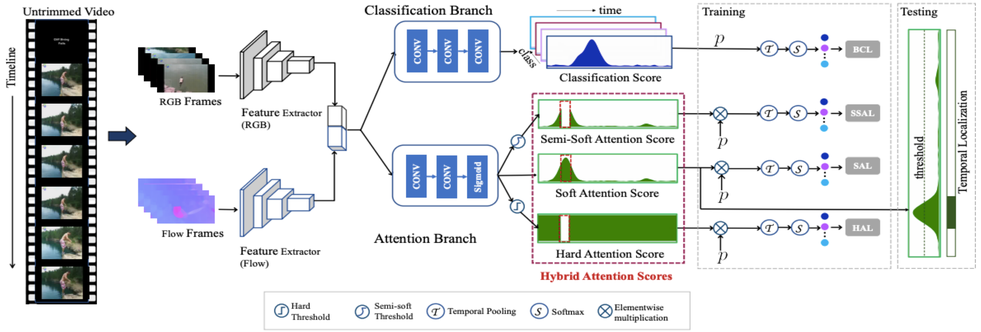
Contributions

- We propose a hybrid attention mechanism to
  - Detect complete actions
  - Model background activity



- We achieve state-of-the-art performance on both the THUMOS14 and ActivityNet-1.2 datasets
- We describe extensive experiments to demonstrate the efficacy of different components of our approach

Proposed Approach



Attention Modules

$$\mathcal{L}_{\text{guide}} = \sum_{i=1}^{T} |1 - a_i - \bar{s}_{c+1}| \qquad \mathcal{L}_{\text{sparse}} = \sum_{i=1}^{T} |a_i|$$

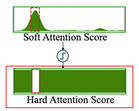
Soft Attention

Remove background frames

$$\mathcal{L}_{\text{SAL}} = -\sum_{j=1}^{c+1} y_j^f \log(p_j)$$

Hard Attention

Remove more confident foreground frames

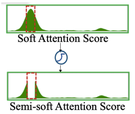


$$a_i^{\text{hard}} = \begin{cases} 1, & \text{if } a_i < \gamma \\ 0, & \text{otherwise} \end{cases}$$

$$\mathcal{L}_{\text{HAL}} = -\sum_{j=1}^{c+1} y_j \log(p_j^{\text{hard}})$$

Semi-Soft Attention

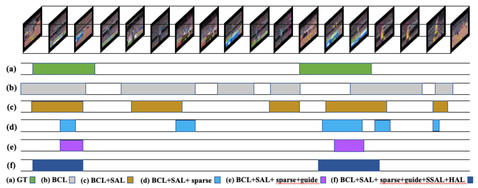
Remove both confident foreground frames and background frames



$$a_i^{\text{semi-soft}} = \begin{cases} a_i, & \text{if } a_i < \gamma \\ 0, & \text{otherwise} \end{cases}$$

$$\mathcal{L}_{\text{SSAL}} = -\sum_{j=1}^{c+1} y_j^f \log(p_j^{\text{semi-soft}})$$

Ablation Studies on THUMOS14



Exp	$\mathcal{L}_{\text{BCL}}$	$\mathcal{L}_{\text{SAL}}$	$\mathcal{L}_{\text{HAL}}$	$\mathcal{L}_{\text{SSAL}}$	$\mathcal{L}_{\text{sparse}}$	$\mathcal{L}_{\text{guide}}$	AVG mAP
1	✓	-	-	-	-	-	24.6
2	✓	✓	-	-	-	-	30.8
3	✓	✓	-	-	-	✓	28.9
4	✓	✓	-	-	✓	-	30.9
5	✓	✓	-	-	✓	✓	**34.8**
6	✓	✓	✓	✓	-	-	30.9
7	✓	✓	✓	✓	-	✓	31.1
8	✓	✓	✓	✓	✓	-	37.9
9	✓	✓	✓	-	✓	✓	36.6
10	✓	✓	-	✓	✓	✓	38.1
11	✓	✓	✓	✓	✓	✓	**39.8**

Quantitative Results

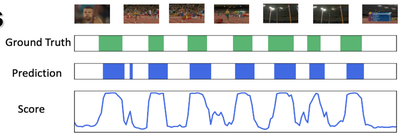
THUMOS14

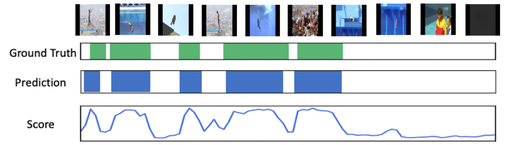
Method	Feature	IoU 0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG
Hide-and-Seek	-	36.4	27.8	19.5	12.7	6.8	-	-	-
UntrimmedNets	-	44.4	37.7	28.2	21.1	13.7	-	-	-
STPN	I3D	52.0	44.7	35.5	25.8	16.9	9.9	4.3	26.4
AutoLoc	UNT	-	-	35.8	29.0	21.2	13.4	5.8	-
W-TALC	I3D	55.2	49.6	40.1	31.1	22.8	-	7.6	-
Liu *et al*	I3D	57.4	50.8	41.2	32.1	23.1	15.0	7.0	32.4
MAAN	I3D	59.8	50.8	41.1	30.6	20.3	12.0	6.9	31.6
CleanNet	UNT	-	-	37.0	30.9	23.9	13.9	7.1	-
3C-Net	I3D	56.8	49.8	40.9	32.3	24.6	-	7.7	-
Nguyen *et al*	I3D	60.4	56.0	46.6	37.5	26.8	17.6	9.0	36.3
Islam *et al*	I3D	62.3	-	46.8	-	29.6	-	9.7	-
BaS-Net	I3D	58.2	52.3	44.6	36.0	27.0	18.6	10.4	35.3
DGAM	I3D	60.0	54.2	46.8	38.2	28.8	19.8	11.4	37.0
HAM-Net (Ours)	I3D	**65.9**	**59.6**	**52.2**	**43.1**	**32.6**	**21.9**	**12.5**	**39.8**

ActivityNet-1.2

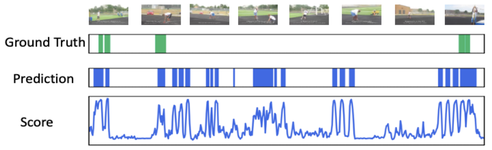
Supervision	Method	IoU 0.5	0.75	0.95	AVG
Full	SSN	41.3	27.0	6.1	26.6
Weak	UntrimmedNets	7.4	3.2	0.7	3.6
	AutoLoc	27.3	15.1	3.3	16.0
	W-TALC	37.0	12.7	1.5	18.0
	Islam *et al*	35.2	-	-	-
	TSM	28.3	17.0	3.5	17.1
	3C-Net	35.4	-	-	21.1
	CleanNet	37.1	20.3	5.0	21.6
	Liu *et al*	36.8	22.0	5.6	22.4
	BaS-Net	34.5	22.5	4.9	22.2
	DGAM	**41.0**	23.5	**5.3**	24.4
	HAM-Net (Ours)	**41.0**	**24.8**	**5.3**	**25.1**

Qualitative Results





Failure Case



person starts to do the high jump activity but stops short without completing the full action

Limitations:

- Cannot detect multiple overlapping actions
- Cannot always solve action ambiguity

Both are partly due to the inherent limitations of the weakly supervised action localization paradigm. Additional supervision is required to solve these issues.

Key References

Nguyen, P. X.; Ramanan, D.; and Fowlkes, C. C. 2019. Weakly-supervised action localization with background modeling. In *Proceedings of the IEEE International Conference on Computer Vision*

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*

Nguyen, P.; Liu, T.; Prasad, G.; and Han, B. 2018. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*,

This material is based upon work supported by the U.S. Department of Homeland Security under Award Number 2013-ST-061-ED0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.