

Rethinking the Faster R-CNN Architecture for Temporal Action Localization

Yu-Wei Chao¹, Sudheendra Vijayanarasimhan², Bryan Seybold², David A. Ross², Jia Deng¹, Rahul Sukthankar²



¹University of Michigan, Ann Arbor ²Google Research

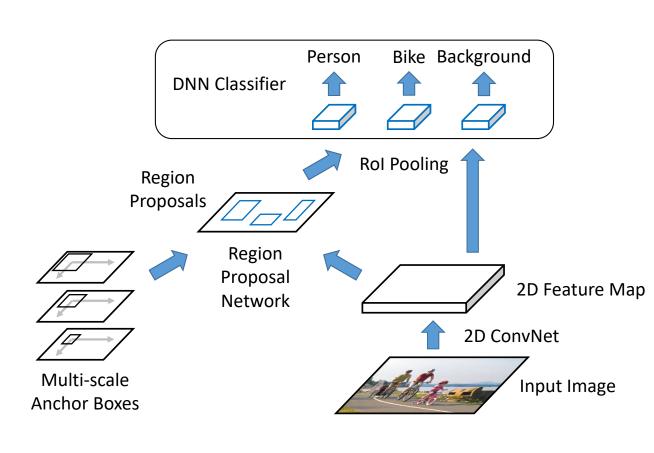
Temporal Action Localization using Faster R-CNN

Temporal Action Localization

Given a long, untrimmed video, identify the start and end times as well as the action label for each action instance in the video.

Faster R-CNN (Ren et al. 2015)

- Use DNN to power the two processes of *proposal generation* and *classification*.
- Originally used for *object detection*, and has been adapted for *temporal action localization*.



Object detection in images

1D Feature Map $\qquad \longrightarrow \qquad$ 2D or 3D ConvNet

Temporal action localization in video (Dai et al. 2017, Gao et al. 2017, Xu et al. 2017)

Contributions

- 1. We address three key issues in applying Faster R-CNN for temporal action localization.
- 2. Our approach, TAL-Net, achieves state-of-the-art performance on THUMOS'14 and competitive performance on ActivityNet challenge.

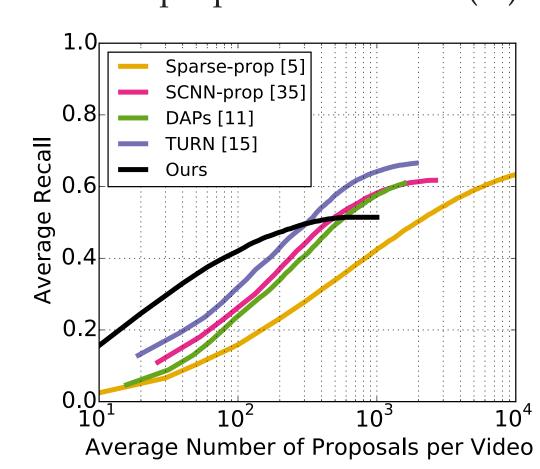
State-of-the-Art Comparisons

Karaman et al. [23]

THUMOS'14

- Outperformed other methods in the low AN region.
- Achieved the highest mAP when the tIoU threshold is over 0.2.

Action proposal in AR-AN (%)



Oneata et al. [31] 36.6 33.6 27.0 20.8 14.4 Wang et al. [45] 47.7 43.5 36.3 28.7 19.0 10.3 5.3 Shou et al. [35] 48.9 44.0 36.0 26.4 17.1 Yeung et al. [50] Yuan et al. [52] Escorcia et al. [11] Buch et al. [2] - - 40.1 29.4 23.3 13.1 7.9 Shou et al. [34] Buch et al. [1] Gao et al. [14] Hou et al. [19] Dai et al. [8] Gao et al. [15] Xu et al. [49] 54.5 51.5 44.8 35.6 28.9 **66.0 59.4** 51.9 41.0 29.8 Zhao et al. [54] 59.8 57.1 **53.2 48.5 42.8 33.8 20.8** Ours

Action localization in mAP (%)

4.6 3.4 2.4 1.4 0.9

ActivityNet v1.3 (validation set)

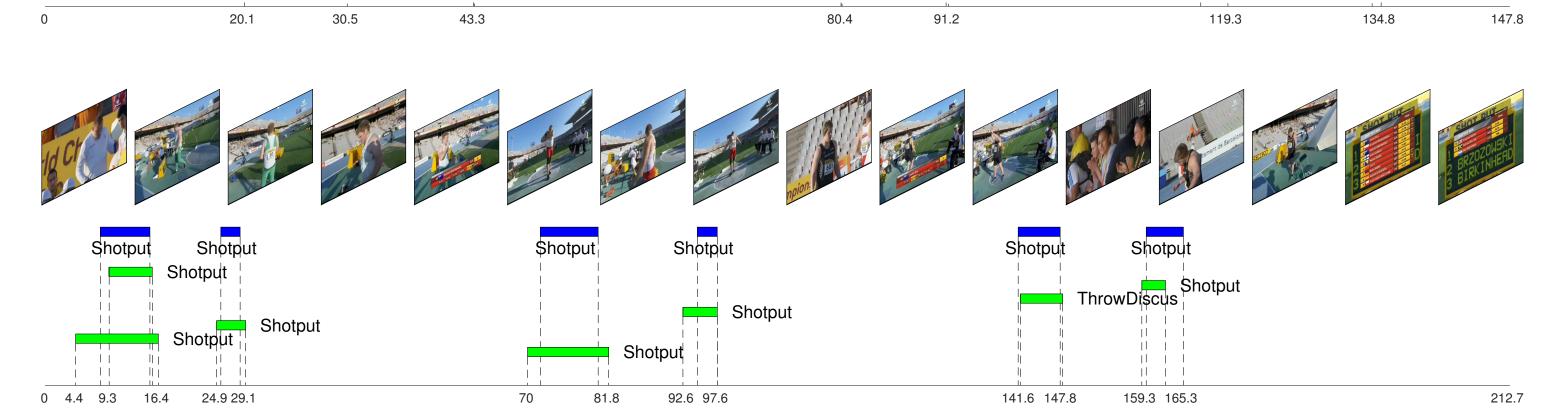
 Outperformed other Faster R-CNN based methods at tIoU threshold 0.5.

Action localization in mAP (%) 0.95 Average Singh and Cuzzolin [38] 34.47 43.65 Wang and Tao [48] 0.20 Shou et al. [34] Dai et al. [8] Xu et al. [49] 38.23 18.30 1.30

Qualitative Results

Below are three examples on THUMOS'14, each consists of a sequence of frames sampled from a full video, and the ground-truth (blue) and predicted (green) action segments and class labels.

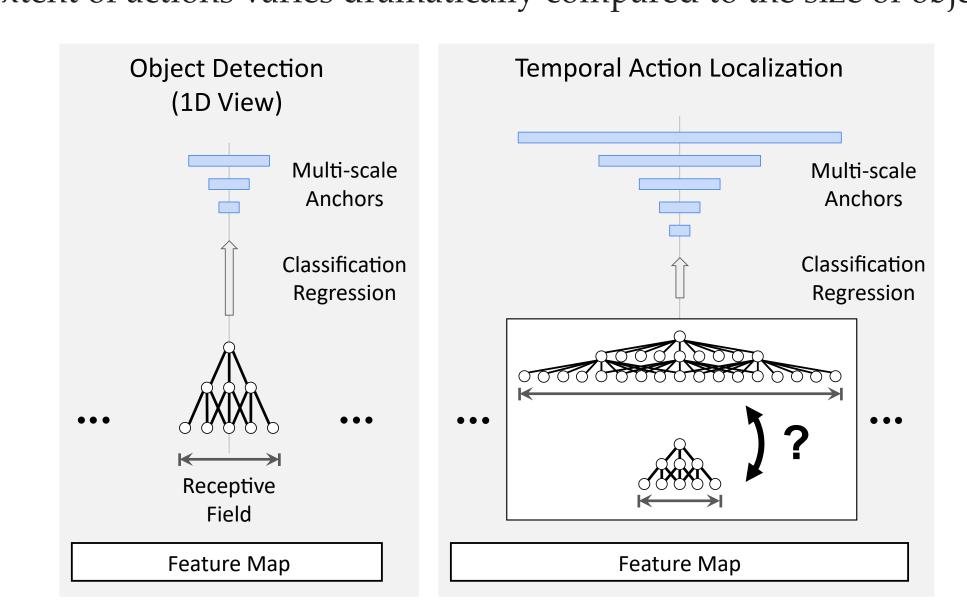




TAL-Net

Issue 1: How to handle large variations in action durations?

The temporal extent of actions varies dramatically compared to the size of objects in an image.

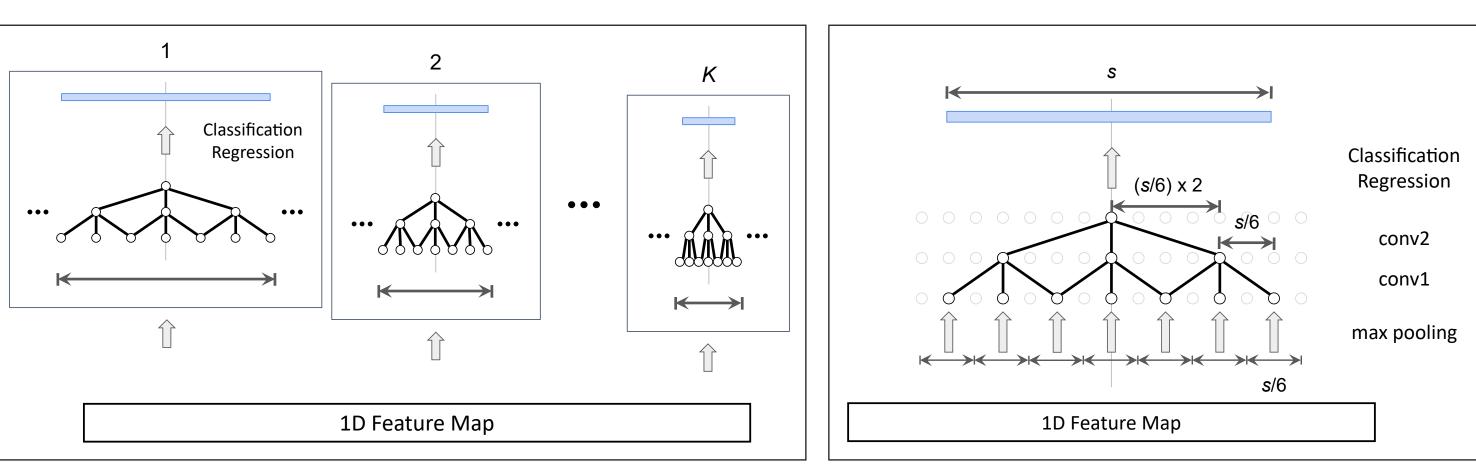


Solution: Receptive field alignment in proposal generation

Align each anchor's receptive field with its temporal span using:

1. Multi-tower network

2. Dilated convolutions



<u>Issue 2</u>: How to utilize temporal context?

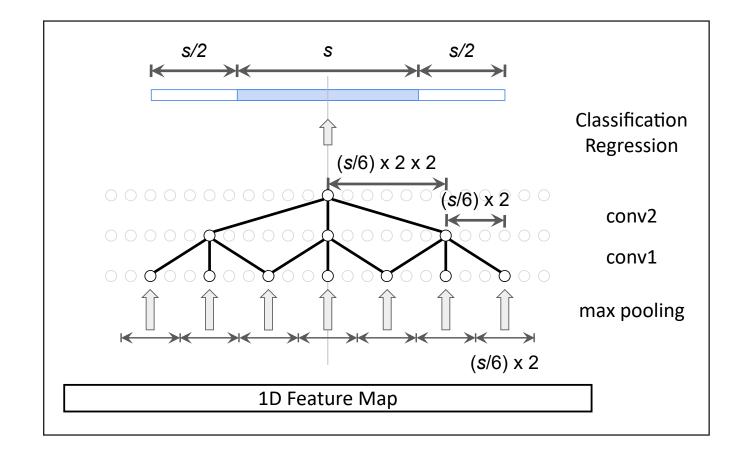
The Moments preceding and following an action instance contain critical information for:

- 1. Accurate localization of action boundaries
- 2. Identify the action class within the boundaries

Solution: Context feature extraction

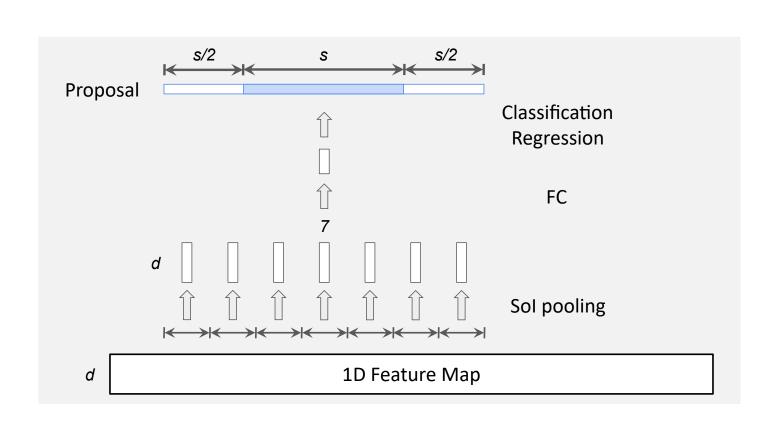
Proposal generation:

Increasing the dilation rate



Action classification:

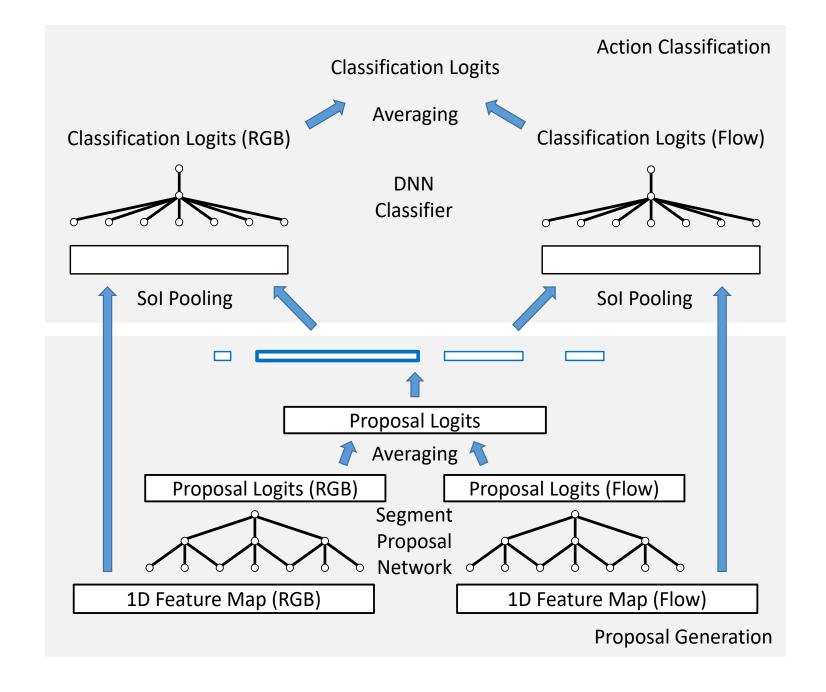
Increasing the extent of RoI pooling



Issue 3: How best to fuse multi-stream features?

There has been limited work in exploring different fusion schemes for Faster R-CNN.

Solution: Late feature fusion



Ablation Study on THUMOS'14

1. Receptive field alignment

	AN	10	20	50	100	200	
oposal eration AR (%)	Single	9.4	15.3	25.3	33.9	41.3	RGB
	Single + TConv	12.9	20.0	30.3	37.6	44.0	
	Multi + TConv	13.4	20.6	31.1	38.1	43.7	
	Multi + Dilated	14.0	21.7	31.9	38.8	44.7	
	Single	11.0	18.0	28.9	36.8	43.6	
	Single + TConv	15.1	23.2	33.7	40.0	44.7	Flow
	Multi + TConv	15.7	24.0	35.0	41.1	46.2	11000
	Multi Dilated	162	25.4	25 0	12.2	17 5	

•	Multi+	Dilated	achieves	the	highes	st AR
	IVICIT .	Dilacca	acine veb			

Proposal generation in AR (%) Action localization in mAP (%)

SoI Pooling + Context **54.3 48.8 38.2** 18.6 **0.9** Multi + Dilated + Context | 17.4 26.5 36.5 43.3 48.6 SoI Pooling + Context

2. Context feature extraction

• Including context features improves both AR and mAP.

3. Late feature fusion

Action localization in mAP (%)

tIoU	0.1	0.3	0.5	0.7	0.9
RGB	49.3	42.6	31.9	14.2	0.6
Flow	54.3	48.8	31.9 38.2	18.6	0.9
Early Fusion Late Fusion	60.5	52.8	40.8	19.3	0.8
Late Fusion	59.8	53.2	42.8	20.8	0.9
L	ı				

Late fusion beats early fusion.