Weakly-supervised Action Localization with Background Modeling Supplementary Materials

Phuc Xuan Nguyen
University of California, Irvine
nguyenpx@ics.uci.edu

Deva Ramanan
Carnegie Mellon University
deva@cs.cmu.edu

Charless C. Fowlkes
University of California, Irvine
fowlkes@ics.uci.edu

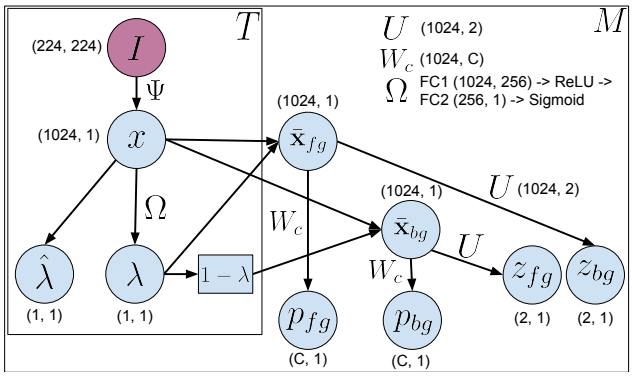

Figure 1: Network architecture for our weakly supervised action localization model.

1. Network Architecture

Figure 1 shows the overall network architecture with the dimensions for the internal components along with dimensions of the parameters.

2. Micro-videos

Micro-videos are obtained from using a third-party Instagram API[1]. Videos are kept untrimmed through the process. No video editing or temporal annotation is involved. We curate user-provided tags into weak video-level labels by removing duplicate and mis-tagged videos. The background fraction in MVs ranges from 10-50%. This cheaply acquired additional dataset was sufficient to bridge the gap between weakly-supervised and fully-supervised methods at lower IoUs. To improve at higher IoUs, models likely need to factor in the domain shift between MVs and THUMOS14 videos, as pointed out in [2].

3. Extra results

Table 1 shows the ablation study for the post-processing procedure. While our method benefits greatly from the new post-processing, STPN [1] does not. The new procedure aggressively generates many more proposals by using a large number of fine-grained thresholds, operating at much higher recall. STPN's precision suffers due to lower-quality intermediate outputs (attention/TCAMs) and doesn't benefit from our post-processing. It is the conjunction of background modeling and postprocessing that yields the large performance gain

Table 1: Results on THUMOS-14.

Methods	0.3	0.4	0.5	0.6	0.7
STPN (re-implementation)	34.2	24.7	15.4	8.2	3.7
STPN w/ our post-process	35.1	25.1	14.9	7.5	3.4
Ours w/ STPN's post-process	37.3	29.9	21.5	14.2	7.6
Ours	46.6	37.5	26.8	17.6	9.0

References

[1] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. *CVPR*, 2018.

[2] Phuc Xuan Nguyen, Gregory Rogez, Charless Fowlkes, and Deva Ramanan. The open world of micro-videos. *CVPR BigVision Workshop*, 2016.

[1] https://github.com/althonos/InstaLooter