Action Graphs: Weakly Supervised Action Localization with Graph Convolution Networks

Maheen Rashid[1] Hedvig Kjellström[2] Yong Jae Lee[1]

[1]University of California, Davis. Department of Computer Science

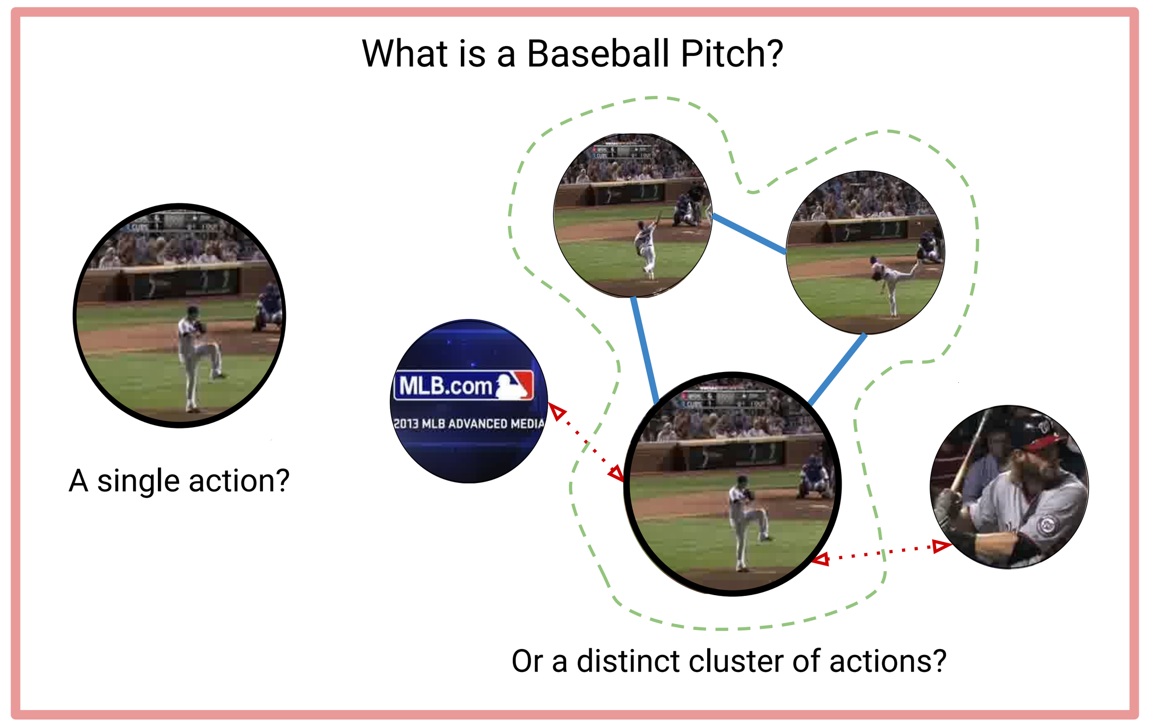
[2]KTH Royal Institute of Technology. Division of Robotics, Perception, and Learning.

WACV 2020

Motivation and Key Idea

Challenge:
Temporally localize actions in videos without frame level annotation, using only weak video level training labels.

Weakly-supervised systems must use similarity between time segments to make predictions.



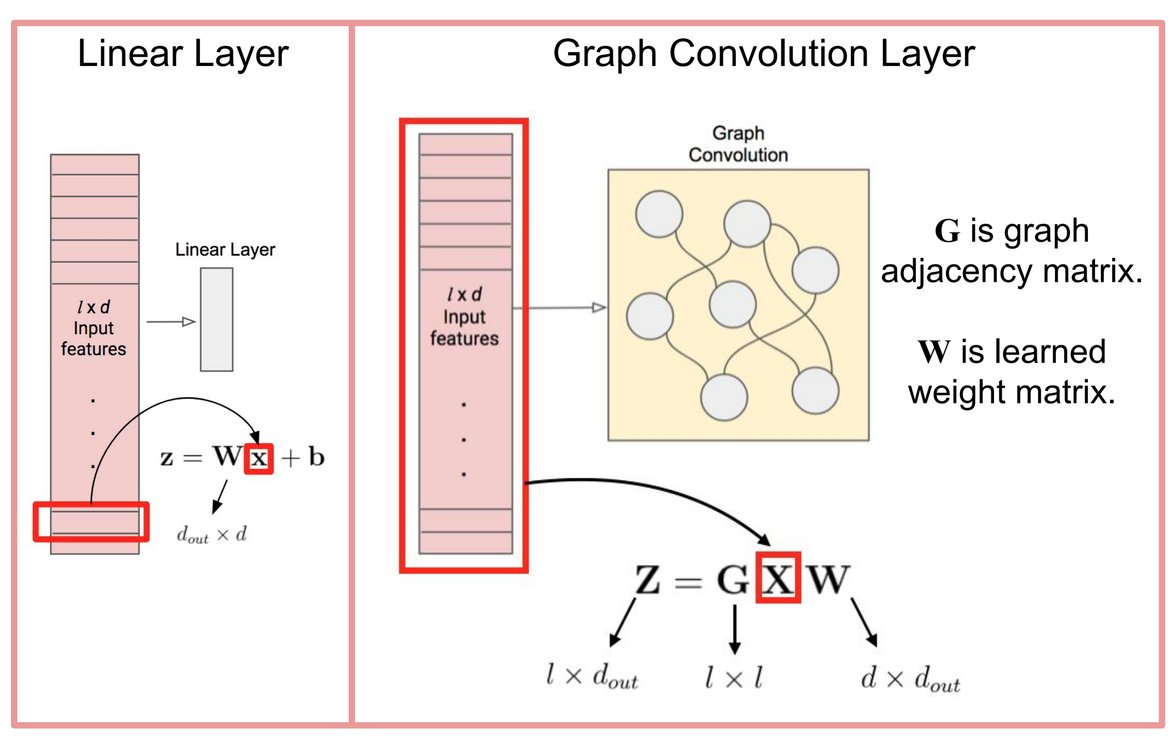
Key Idea:
Train and infer on clusters of similar time segments explicitly using graph convolutions.

Videos as Graphs

Break up each video in to l time segments and extract features per time segment.

Represent each time segment as a node in a graph.

Nodes' edge weights are proportional to their similarity.



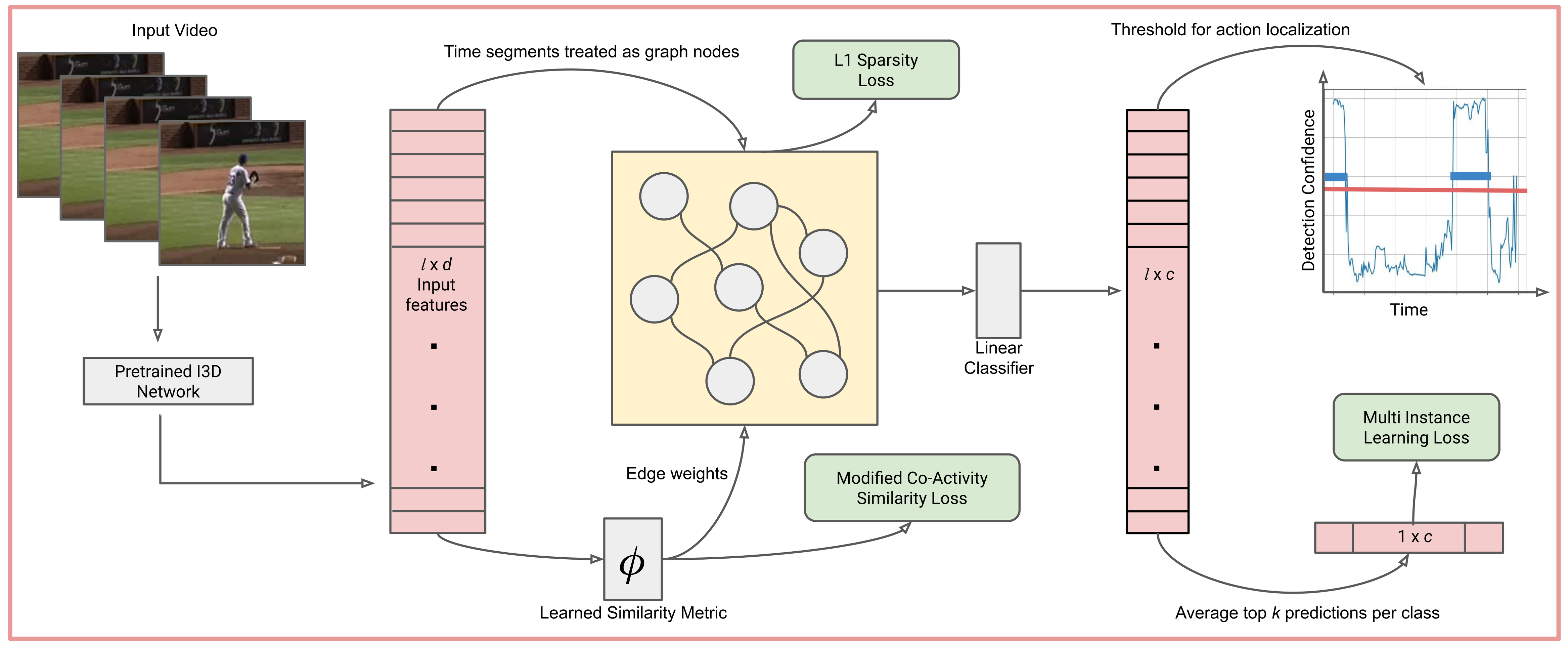
$\mathbf{G}$ transforms each row of $\mathbf{X}$ to a weighted combination of other rows of $\mathbf{X}$

Approach

Extract flow and RGB features from pretrained I3D network [1] for every 16 frames of video to get input features.

Use graph layer to build a graph from each input video. Classify graph output of each time segment in to c classes.

Average top k time segments to get video level score, where $k = \max(1, \lfloor \frac{l}{d} \rfloor)$. Use multi class cross entropy loss.



Use cosine similarity $f(.)$ of ϕ to weigh edges in the graph layer:

$$\mathbf{G}_{ij} = f(\phi(\mathbf{x}_i), \phi(\mathbf{x}_j))$$

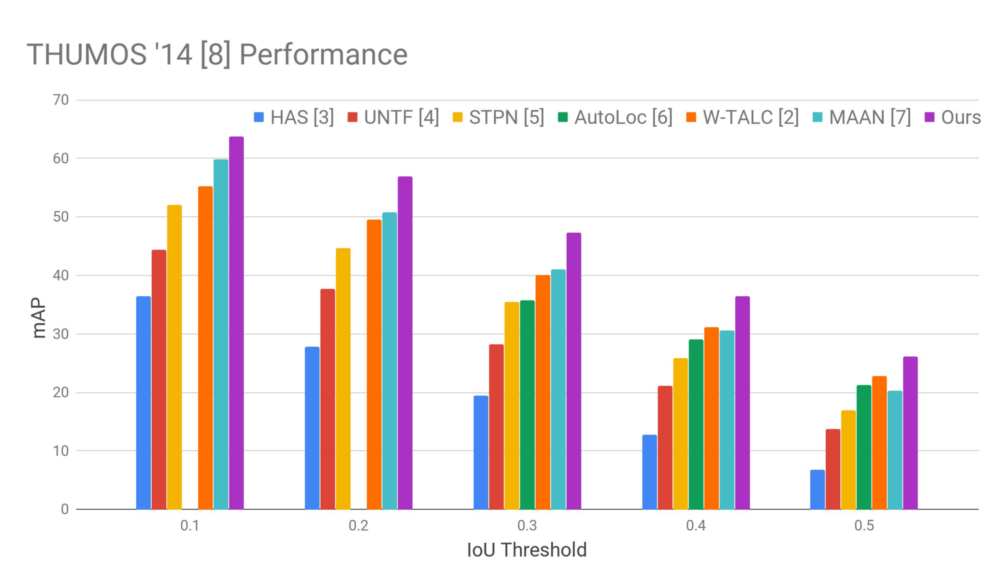
Use an L1 loss to encourage disjoint cliques:

$$L_{L1} = \frac{\sum_{i=1}^{l} \sum_{j=1}^{l} |\mathbf{G}_{ij}|}{l^2}$$

If two cliques have similar classification, encourage them to have high edge weights. Use Co-Activity Similarity Loss [2]:

$$L_{CASL}^{j,k,i} = \max(0, \bar{f}(\mathbf{f}_i^j, \mathbf{f}_i^k) - \bar{f}(\mathbf{b}_i^j, \mathbf{f}_i^k) + 0.5) + \max(0, \bar{f}(\mathbf{f}_i^j, \mathbf{f}_i^k) - \bar{f}(\mathbf{b}_i^k, \mathbf{f}_i^j) + 0.5)$$

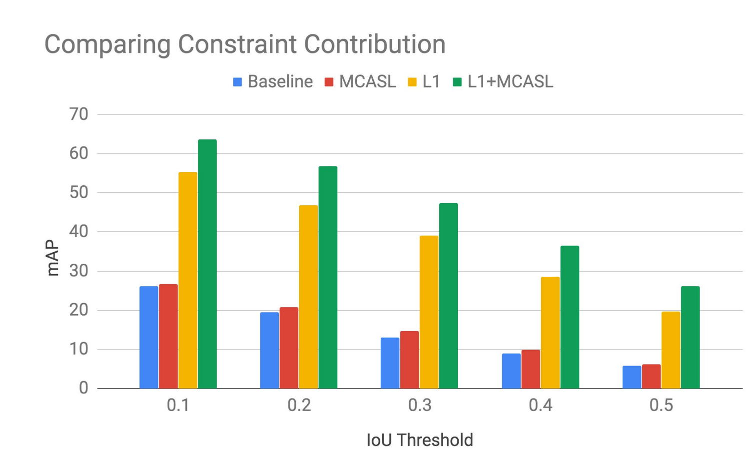
Quantitative Results

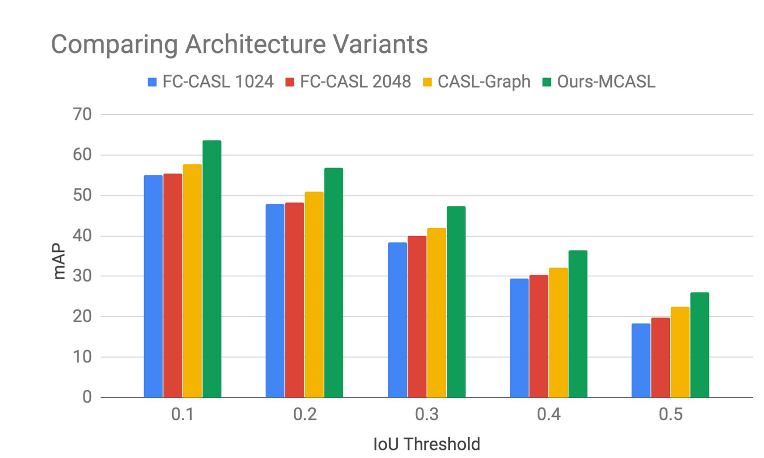


ActivityNet 1.2 [9] Performance

Method	mAP@IoU 0.5	0.7	0.9
UNTF [4]	7.4	3.9	1.2
Auto-Loc [6]	27.3	17.5	6.8
W-TALC [2]	**37.0**	14.6	-
Ours	29.4	**17.5**	**7.5**

On Charades [10] we achieve 15.8 mAP.



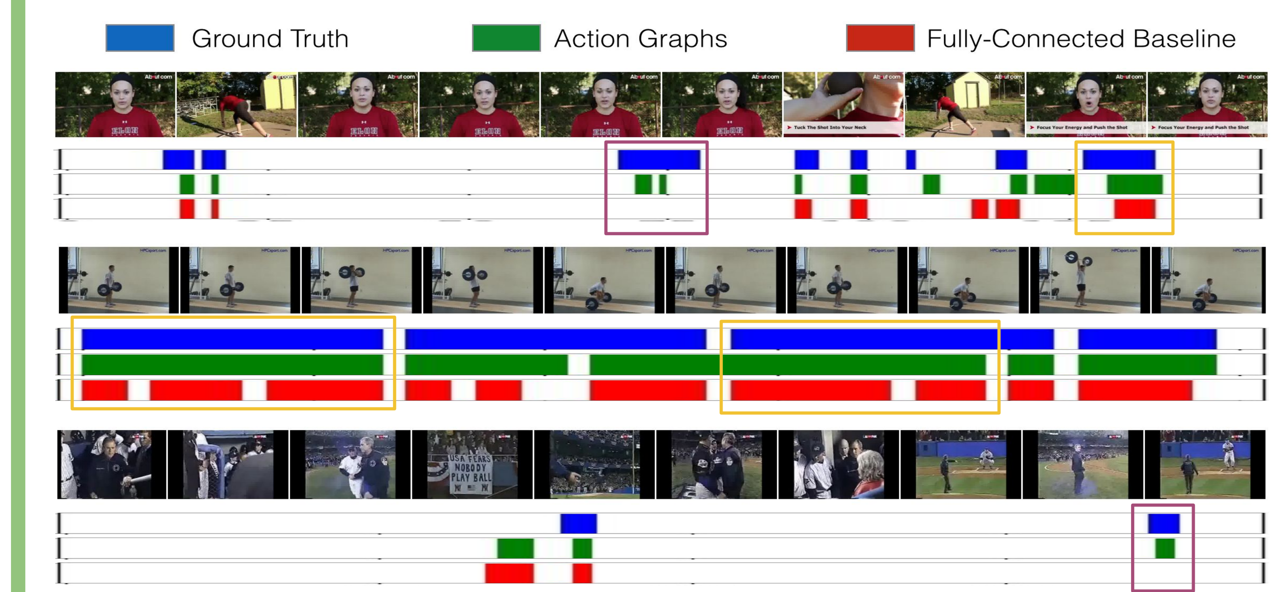


Hyperparameter d determines k used in multi instance learning loss. It affects the temporal length of localization predictions. Varying d randomly during training can help.

d	Video %	THUMOS mAP @ 0.5 IoU	THUMOS Test Data %	ActivityNet mAP @ 0.5 IoU	ActivityNet Test Data %	Charades mAP Per Frame	Charades Test Data %
1	50-100%	18.5	2.8	**29.4**	57.2	14.9	76.9
2	25-50%	44.9	3.8	5.5	19.0	15.4	82.0
4	12.5-25%	58.4	14.1	1.7	14.4	15.2	75.5
8	0-12.5%	**63.7**	93.9	1.4	18.8	13.8	15.4
Random		39.0	-	14.3	-	**15.8**	-

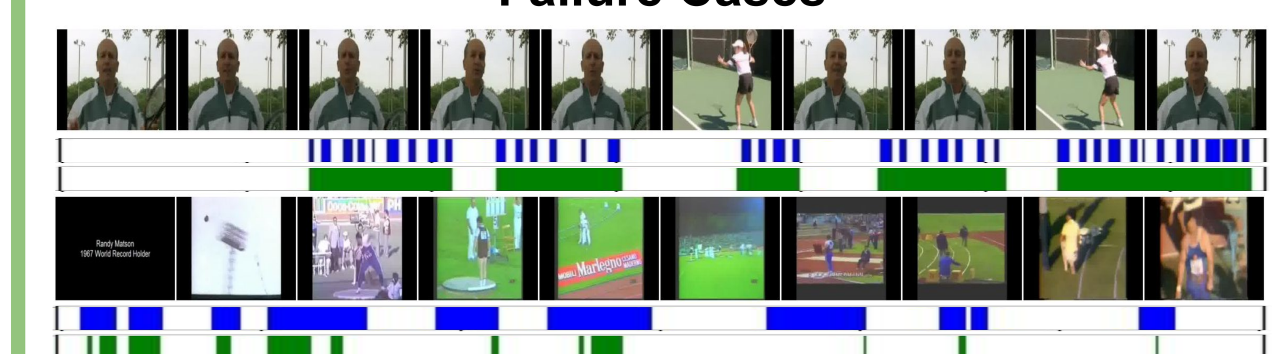
Qualitative Results

Comparison Against FC-CASL

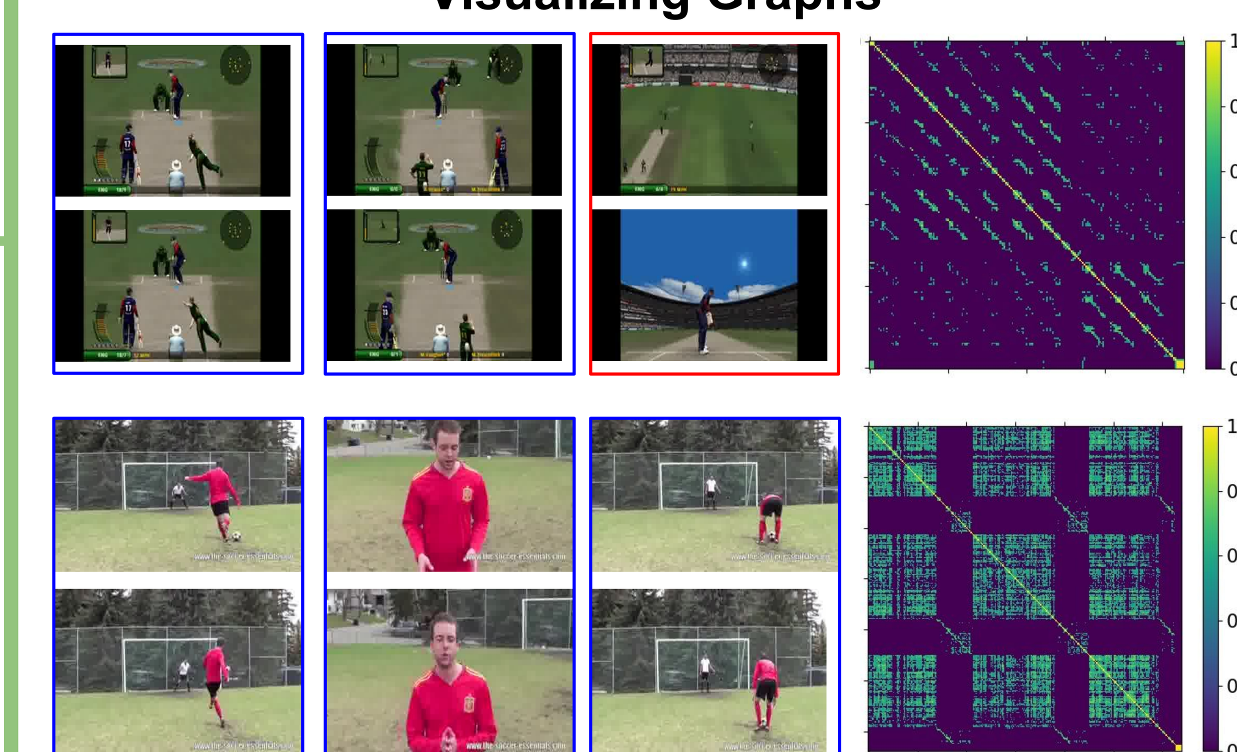


Additional Results

Failure Cases



Visualizing Graphs



Conclusion

Our novel weakly supervised action localization method explicitly uses similarity between video segments during both training and testing by using graph convolutions.

The method pushes the state of the art and outperforms equivalent networks that do not use graphs.

References:
[1] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. CVPR, 2017
[2] S. Paul, S. Roy, and A. K. Roy-Chowdhury. W-TALC: Weakly-supervised temporal activity localization and classification. ECCV, 2018
[3] K. K. Singh and Y. J. Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. ICCV, 2017.
[4] L.Wang, Y. Xiong, D. Lin, and L. Van Gool. Untrimmednets for weakly supervised action recognition and detection. CVPR, 2017.
[5] P. Nguyen, T. Liu, G. Prasad, and B. Han. Weakly supervised action localization by sparse temporal pooling network. CVPR, 2018.
[6] Z. Shou, H. Gao, L. Zhang, K. Miyazawa, and S.-F. Chang. AutoLoc: Weakly-supervised temporal action localization in untrimmed videos. ECCV, 2018
[7] Y. Yuan, Y. Lyu, X. Shen, I. W. Tsang, and D.-Y. Yeung. Marginalized average attentional network for weakly supervised learning. ICLR, 2019.
[8] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev,R. Sukthankar, and M. Shah. The THUMOS challenge on action recognition for videos "in the wild". CVIU, 2017.
[9] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. CVPR, 2015.
[10] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. ECCV, 2016.

Acknowledgements: This work was supported in part by a grant from Swedish Research Council Formas, Hellman Fellowship, and ARO YIP W911NF17-1-0410.