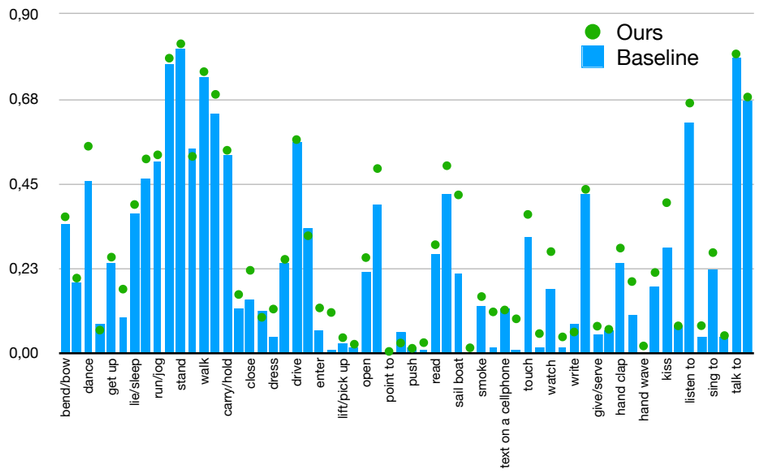
Spatio-Temporal Context for Action Detection

Manuel Sarmiento, David Varas, Elisenda Bou-Balust
CVPR 2021 · Apple

Motivation

Reason on **interactions** between actors and elements of the scene taking into account **spatio-temporal** information.

Baseline Comparison



Label Category	Improvement (mAP)
Person Movement	+2.43
Object Manipulation	+4.04
Person Interaction	+3.35

Mean improvement per label type.

Results

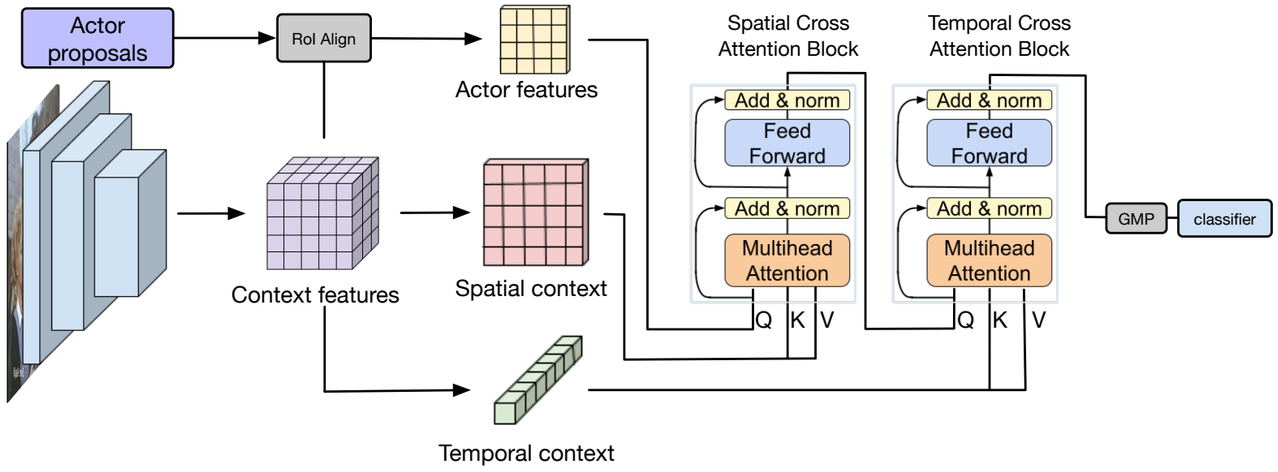
Ablation study with different versions of our architecture using pooled actor features and spatial or spatio-temporal context:

Method	Actors Features	mAP
Baseline	-	24.80
Spatial Context	-	26.50
Spatial Context	Spatial	26.75
Spatial+Temporal Context	-	26.65
Spatial+Temporal Context	Spatial	27.02

Comparison against other methods with similar backbone and first order relations:

Model	Backbone	AVA	mAP
SlowFast [1]	R50 8×8	2.1	24.80
Action Tx [2]	I3D	2.1	25.00
LFB [3]	R50	2.1	25.80
Context-Aware RCNN [4]	R50 16×4	2.1	25.80
AVSlowFast [5]	R50 4×16	2.2	25.90
ACAR-Net [6]	R50 8×8	2.2	26.71
Ours	R50 8×8	2.2	27.02

System Overview



Overview of the proposed system with two cross attention blocks to **enrich actor features**.

Conclusions

- We propose a novel system that leverages **temporal information** from adjacent frames together with **spatial information** to improve the recognition of actor interactions in video clips.
- The architecture uses **two cross attention mechanisms** to extract the relevant information from spatial and temporal features.
- Results **open the door** towards the usage of short-term temporal information in contextualized action detection.

References

[1] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In Proc. ICCV, 2019.

[2] R. Girdhar, J. Joao Carreira, C. Doersch, and A. Zisserman. Video action transformer network. In Proc. CVPR, 2019.

[3] C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krahenbuhl, and R. Girshick. Long-term feature banks for detailed video understanding. In Proc. CVPR, 2019

[4] J. Wu, Z. Kuang, L. Wang, W. Zhang, and G. Wu. Context-aware rcnn: A baseline for action detection in videos. In Proc. ECCV, 2020.

[5] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, and C. Feichtenhofer. Audiovisual slowfast networks for video recognition, 2020

[6] J. Pan, S. Chen, M. Z. Shou, Y. Liu, J. Shao, and H. Li. Actor-context-actor relation network for spatio-temporal action localization. In Proc. CVPR, 2021