TRECVID 2022 Activities in Extended Video (ActEV)

Dynamic Interactive Aggregation Network for TRECVID'22 ActEV Task

Xingchao Ye, Ping Li*

School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, P. R. China

Introduction

- In this paper, we present our solutions for the Activities in Extended Video (ActEV) in TRECVID 2022.
- The integration of different types of interaction is beneficial for capturing higher level spatial-temporal features.
- We adopt a Dynamic Interactive Aggregation Network (DIAN) (i.e., AIA [1]), which attempts to (1) integrates different interactive relationships through intensive series connection, and (2) dynamically updates memory features through iterative self-learning.

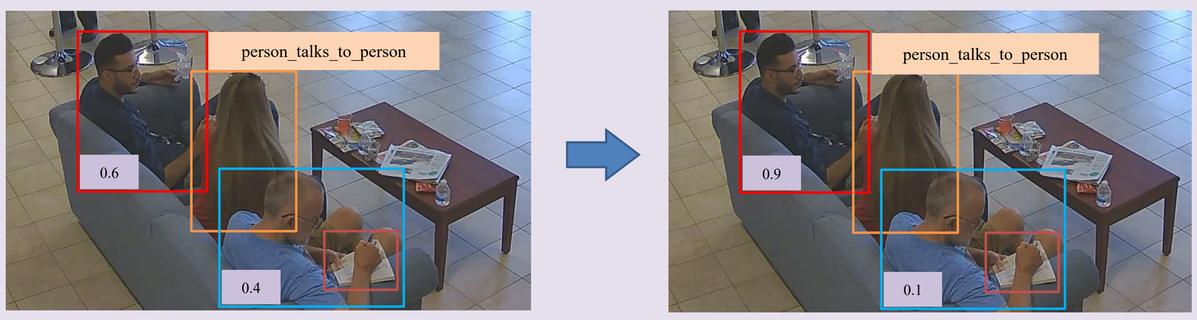


Figure 1: The motivation of the Dynamic Interactive Aggregation Network (DIAN)

Framework

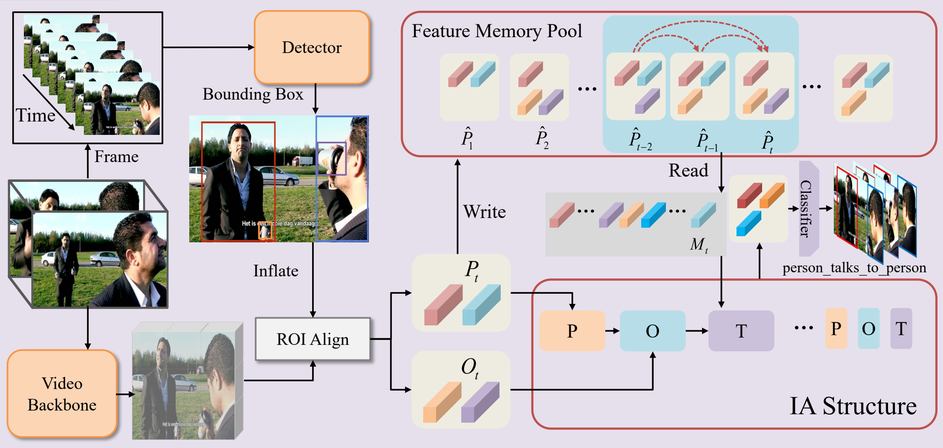
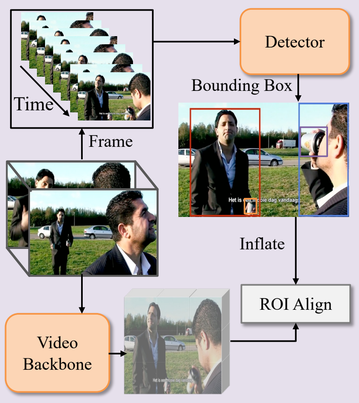


Figure 2: The framework of DIAN

The DIAN contains of (a) Interactive aggregation module: Person features, object features and memory features from the feature pool Ω in c are fed to IA in order to integrate multiple interactions. The output of IA is passed to the final classifier for predictions; (b) Iterative memory update module: read memory features from feature pool and writes fresh person features to it.

Video encoder



- Given an input video V, N frames $\{f_i\}_{i=1}^{N}$ and clips $\{c_i\}_{i=1}^{N}$ are uniformly sampled where each clip v_i consists of consecutive frames around each sampled frame f_i.
- We use Faster-RCNN[1], SlowFast[2] as video encoder to generate iteration features $\mathrm{P}^t, \mathrm{O}^t$

Interactive aggregation module

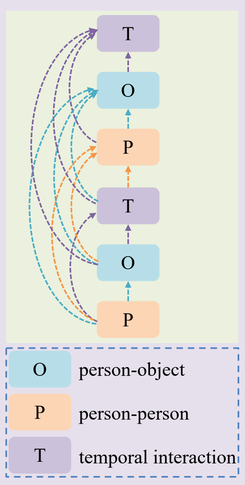
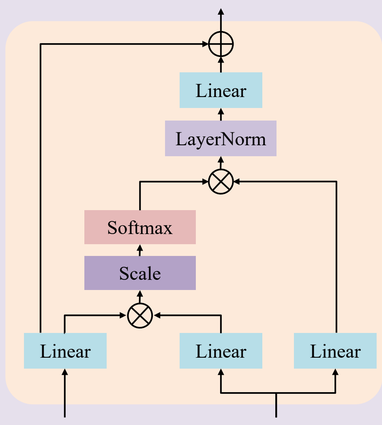


Figure 4: The structure of visual encoder

- In Dense Serial IA, each interaction block takes all the outputs of previous blocks and aggregates them using a learnable weight. Formally, the query of the i^{th} block can be represented as

$$Q_{t,i} = \sum_{j \in \mathbf{C}} W_j \odot E_{t,j}$$

Iterative memory update module



READ: At the beginning of each iteration, given a video clip $V_i^{(t)}$ of the i^{th} video, the memory features are read from the memory pool Ω, which is $[\hat{P}_{\mathrm{t-L}}^{(i)}, \dots, \hat{P}_{\mathrm{t-1}}^{(i)}]$.

WRITE: At the end of each iteration, personal features for the target clip $\mathrm{P}_t^{(i)}$ are written back to the memory pool Ω as estimated memory features $\widehat{P}_{\mathrm{t}}^{(i)}$, tagged with the current loss value.

Competition Results

Organization(s)	AOD mean Pmiss@0.1rfa	AOD mean nMODE@0.1rfa	AOD mean nAUDC@0.2rfa
Beijing University of Posts and Telecommunications	0.6309	0.0538	0.6705
UMD	0.8131	0.1620	0.8300
Ours	0.9921	0.0303	0.9922
Waseda University, Meimei University, SoftBank Corpotation	0.9961	0.1080	0.9964
M4D_team	-	-	-

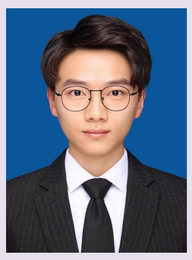
Organization(s)	AD mean Pmiss@0.1rfa	AD mean nAUDC@0.1rfa
Beijing University of Posts and Telecommunications	0.5805	0.6231
UMD	0.7789	0.7995
Ours	0.9728	0.9732
Waseda University, Meimei University, SoftBank Corpotation	0.9829	0.9850
M4D_team	0.9823	0.9819

- Our model ranks the **third** in the overall evaluation of Activities in Extended Video task. Among the five metrics, ours ranks the **first** in terms of AOD mean nMODE@0.1rfa on TRECVID 2022 ActEV task.

Conclusion

In this poster, we adopt a Dynamic Interactive Aggregation Network (DIAN), which integrates different types of interactions in the same segment in a dense serial manner to adjust the weights indicating the object relations. The memory features are dynamically updated to obtain long-term temporal interaction dependency by iterative self-learning. Experimental results on ActEV dataset show that our proposed model has excellent performance.

Member Introduction



Xingchao Ye is currently pursuing his master degree in the School of Computer Science and Technology at Hangzhou Dianzi University. His research interests include computer vision.

Ping Li is currently an Associate Professor in computer science at Hangzhou Dianzi University. His research interests include machine learning, computer vision, and data mining.

Contact

Xingchao Ye
212050155@hdu.edu.cn

Ping Li
lpcs@hdu.edu.cn
*Corresponding author

References: [1] Tang J, Xia J, Mu X, and et al. Asynchronous interaction aggregation for action detection[C]. In Proceedings of the European Conference on Computer Vision (ECCV), 2020:71-87.
[2] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. In IEEE Transactions Pattern Analysis and Machine Intelligence (TPAMI), 2017:1137-1149.
[3] Feichtenhofer C, Fan H and et al. Slowfast networks for video recognition[J]. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019:6201-6210.