SSA2D: Single Shot Actor-Action Detection in Videos (Student Abstract)

Aayush Jung Rana, Yogesh S Rawat
University of Central Florida
4000 Central Florida Blvd
Orlando, Florida 32816
aayushjr@knights.ucf.edu, yogesh@crcv.ucf.edu

Abstract

We propose a single-shot approach for actor-action detection in videos. The existing approaches use a two-step process, which rely on Region Proposal Network (RPN), where the action is estimated based on the detected proposals followed by post-processing such as non-maximal suppression. While effective in terms of performance, these methods pose limitations in scalability for dense video scenes with a high memory requirement for thousand of proposals, which leads to slow processing time. We propose SSA2D, a unified end-to-end deep network, which performs joint actor-action detection in a single-shot without the need of any proposals and post-processing, making it memory as well as time efficient.

Introduction

Actor-action detection in videos is a challenging problem where the goal is to detect all the actors in the video and determine which different actions they are performing (Xu et al. 2015; Wang et al. 2020). One natural solution to this problem is to perform object detection, filter all the actors and classify those detected actors for corresponding actions. Motivated by this, existing methods proposed to utilize region proposal (RPN) (Ren et al. 2015) based approaches (Dang et al. 2018; Ji et al. 2018), where they first detect proposal for objects and use them for action detection. However, processing thousands of region proposal is memory and time intensive for large scenes with multiple actor-action pairs. Hence, existing methods can only perform detection at single frame at a time and have to be trained in multiple stages.

We propose SSA2D, an encoder-decoder based unified network that can detect multiple actors and actions in a single forward pass. Instead of using thousands of region proposals, we propose a technique to fuse all the object information priors for a scene with the action features and perform selective attention over regions of potential interest. This allows SSA2D to utilize contextual information between objects and their surrounding for joint actor-action detection in single shot while giving more attention to object regions.

The proposed SSA2D method is computationally less intensive and does not have scalability issues for dense scenes. As SSA2D detects actor-action per pixel without using any region proposals, it can predict all actors and related actions in a scene in a single shot. We evaluate the proposed approach on A2D and VidOR dataset and demonstrate the efficiency and effectiveness for joint actor and action detection in videos. The performance achieved by SSA2D is comparable (sometimes better) with existing methods and it is significantly faster (both training and inference time) and requires fewer network parameters when compared to prior works.

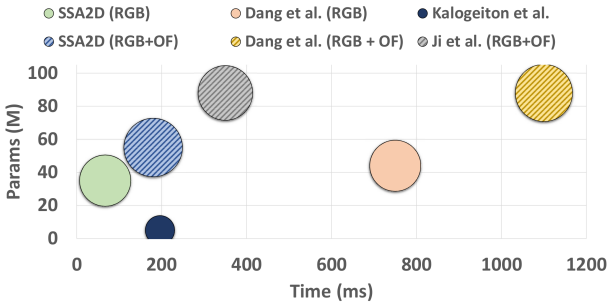

Figure 1: Comparison of SSA2D with existing approaches in terms of speed, number of parameters and performance (size of bubble). We can observe that SSA2D is faster and requires less parameters with comparable performance.

Method

SSA2D consists of a *3D convolution* based encoder E which extracts spatio-temporal features f_{enc} from the input video, and has separate decoder branches (D) for pixel-wise actor ($Actor$), and action ($Action$) detection. The $Actor$ branch predicts the object each pixel belongs to, and in process learns object priors. This object prior is passed to the $Action$ branch which utilizes the spatio-temporal object information and predicts pixel-wise actions. The objective function for each branch is represented by two different loss terms; pixel-wise categorical cross-entropy (L_{CLS}), and volumetric dice loss (L_{DL}). The volumetric dice loss is extended from standard dice loss formulation, which computes loss based on the IoU of each class C over all pixels N of the video given by Eq. 1. The unified end-to-end network is trained using a

Method	Actor	Action	Joint
(Kalogeiton et al. 2017)	42.7	35.5	24.9
(Kalogeiton et al. 2017)	33.3	31.9	19.9
(Ji et al. 2018)	66.4	46.3	36.9
(Dang et al. 2018)	68.1	51.1	38.6
SSA2D(w/o object prior)	65.7	40.9	32.1
SSA2D(w/o attention mask)	67.2	43.6	33.8
SSA2D (RGB)	67.5	46.5	34.6
SSA2D (RGB + OF)	67.5	51.3	39.5

Table 1: Quantitative comparison $(mIoU)$ of SSA2D on A2D dataset with prior state-of-the-art baselines using both RGB and optical flow (OF) as input.

combination of these two loss terms for each branch.

$$L_{DL} = 1 - \frac{2\sum_{c=1}^{C}\sum_{i=1}^{N} p_{ci} * \hat{p_{ci}}}{\sum_{c=1}^{C}(\sum_{i=1}^{N} p_{ci}^2 + \sum_{i=1}^{N} \hat{p_{ci}}^2 + \epsilon)} \quad (1)$$

We propose a selective attention technique to focus on the related activity features for each object. We apply the object mask to the resultant of object priors and action features, which masks out the features from non-object pixels and only retains related features. This object prior and selective attention masking technique removes the need for region proposals, as each pixel feature is individually evaluated for all related actor-action pairs in surrounding region.

Experimental Evaluation

We evaluate our proposed model on the A2D dataset (Xu et al. 2015) and the VidOR dataset (Shang et al. 2019). We measure mean pixel Intersection-over-Union $(mIoU)$ as the evaluation metric following protocols from prior works (Dang et al. 2018; Xu et al. 2015). SSA2D has better action and joint actor-action mIoU score while taking less test time (Table 1, Figure 1). Since we do not rely on region proposal networks, we can perform a single stage training and testing and achieve better results without using segmentation pretrained weights like Dang et al. (Dang et al. 2018). We observe that the proposed method improves the general actor-action relation in a video and the object prior infusion mechanism aids in having a better contextual information for joint actor-action detection(figure 2). On the harder VidOR dataset, SSA2D achieves mIoU of 5.1, 7.9 and 2.1 for actor, action and joint detection.

Conclusion

We demonstrate that the proposed SSA2D method is able to learn actor-action interaction from a video clip and jointly perform pixel-wise actor and action detection. The pixel-wise detection of SSA2D allows multiple actors and actions detection simultaneously without relying on region proposal network or RoI-pooling. As a result the proposed approach is significantly faster and allows scaling on dense video scenes. The proposed object prior infusion mechanism efficiently integrates relevant object features with contextual action features, improving actor and action understanding.

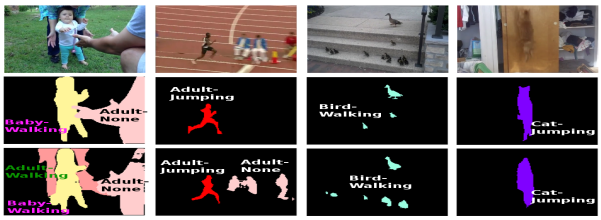

Figure 2: Qualitative results of our approach on A2D dataset. The top, middle and bottom row represents input key frame, ground truth and SSA2D predictions with label respectively.

Acknowledgments

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. D17PC00345. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

Dang, K.; Zhou, C.; Tu, Z.; Hoy, M.; Dauwels, J.; and Yuan, J. 2018. Actor-Action Semantic Segmentation with Region Masks. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Ji, J.; Buch, S.; Soto, A.; and Carlos Niebles, J. 2018. End-to-end joint semantic segmentation of actors and actions in video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 702–717.

Kalogeiton, V.; Weinzaepfel, P.; Ferrari, V.; and Schmid, C. 2017. Joint learning of object and action detectors. In *Proceedings of the IEEE International Conference on Computer Vision*, 4163–4172.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Shang, X.; Di, D.; Xiao, J.; Cao, Y.; Yang, X.; and Chua, T.-S. 2019. Annotating Objects and Relations in User-Generated Videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, 279–287. ACM.

Wang, H.; Deng, C.; Ma, F.; and Yang, Y. 2020. Context Modulated Dynamic Networks for Actor and Action Video Segmentation with Language Queries. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, 12152–12159.

Xu, C.; Hsieh, S.-H.; Xiong, C.; and Corso, J. J. 2015. Can Humans Fly? Action Understanding with Multiple Classes of Actors. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.