Review for Sequence Level Semantics Aggregation for Video Object Detection

Can Koz

April 25, 2021

1 Summary

In video object detection task, aggregation of features from several frames is required since single frame is not sufficient due to fast motion. Existing methods rely one temporally nearby frames and authors propose a method which aggregates features from full-sequence. They uses a novel module called Sequence Level Semantics Aggregation (SELSA) to utilize full sequence. They experimented the method on the ImageNet VID and EPIC KITCHENS dataset and authors achieve state of the performance on these datasets.

2 Strengths

- In case of fast moving objects, image detectors may fail and it is required to analyze several frames. Optical flow focuses on temporally nearby images which is unsatisfactory. Authors reinterprets video object detection as multi shot detection task which overcomes the disadvantages of sequential detection.
- Although the method has a simple pipeline, it works effectively.
- Experiments demonstrate significant improvement over previous methods and it generalizes well to complex scenes.

3 Weaknesses

- Proposed feature aggregation method works fully global and this introduces a weakness for temporal localization.
- Datasets and comparisons can be extended with model complexity and inference time.

4 Evaluation

The methods are evaluated on EPIC KITCHENS and ImageNet VID datasets.They experimented the performance of the method based on mAP metric. In the ablation studies, they tested the effectiveness of SELSA module, sampling strategies for feature aggregation, semantics aggregation in sequence level and data augmentation. For both datasets their results outperformns the other methods.

5 Final Comments and Future Work

Instead of the optical flow or RNN approach for aggregating semantic features, proposed SELSA module achieves superior performance by utilizing features of full sequences. Also the method does not require sophisticated post processing methods. In the future work performance can be improved by considering joint view of global and local aggregation instead of fully global aggregation.