Review for Sequence Level Semantics Aggregation for Video Object Detection

Irem ARPAG

April 25, 2021

With the usage of deep convolutional networks, significant progress was performed in object detection in recent years, especially in still image object detection and slow-moving objects in video object detection (VID) but fast motion is still and outstanding challenge and problem that cause appearance degradation like motion blur, camera defocus and pose variation. Since single image is not a remedy, aggregating features from different frames by using optical flow or recurrent neural network (RNN) seem to be the natural solution of the problem however, since these methods use temporarily nearby frames, the results are mostly unsatisfactory in such cases. The authors of the article argue that the obtain more robust features for video object detection is carrying out by taking the full-sequence level feature aggregation, and they propose SELSA module for aggregating semantic features across frames. Instead of using fixed time window, to utilize the rich information coming from the whole video frames, they exploit multi-shot view that consists of clusters of objects in which each cluster has hundreds even thousand of shots that is the inspiration of the SALSA spectral clustering.

The simple and clean structure of the pipeline, because of not needing complex pose processing methods such as Seq-NMS and Tubelet Rescoring, is one of the strength of the model. Treating video detection as a multi-shot detection problem rather than sequencial detection task and present a global clustering approach for the first time is a novelty in the research area. The proposed method is tested on the large scale of ImageNet VID that indicates great improvement over previous methods as prevailing 82.7 mAP with Faster-RCNN detector. Suplementary experiments on EPIC KITCHEN dataset proves its generalization capacity to more complex scenes. The motivation of the modeule that is aggregating features from the semantic neighborhood rather than utilizing features from a short temporal window is pretty ingenious.

For training the model mageNet VID and DET datasets and for evaluating the method ImageNet VID dataset are used. As ablation study, the researchers compare their different design choice; the first one is their own single-frame baseline, and the second choice is the semantic aggregation within a single frame and the third one is SELSA model itself. If we compare the first and the last choics in terms of mAP, it proviides a large 6.63 mAP improvement compared

with the baseline method. The comparison of the second and third choices confirm the validity of aggregating sequence level features with its better results. As another ablation study they use different sampling strategies, the fisrt one of which demonstrates that by using more frames for testing increase the performance such as 21 frames instead of 5 contribute a 1.04 mAP improvement. The second experiment that increase the sampling stride from 1 to 10 supplies a gain of 2.34 mAP which supportes their motivation. As the third step, they sample semantic neighbors uniformly from the full video sequence regardless of the temporal orders, they achieve a good performance through the whole video sequence (21 frames) SELSA module performs better results then state of art methods without using post-processing methods which supports performance largely. Additional Experiment on Epic Kitchen dataset which is a large scale, and challenging one, confirms its adaptabillity to more complex video detection tasks.

With its challenges such as declaring with problematic fastmotion area, having simple and clean pipeline, no need for complicated post-processing methods, the artice actually present a novel and promising approach.