PhD

Object Detection from Video Tubelets with Convolutional Neural Networks CVPR16

2018/05/17 4:51 PM

- Use selective search to generate proposals (for each frame ? -not quite clear about this);
- choose the detection with the highest score as an anchor and track both forwards and backwards in time starting from this box;
- stop tracking when the tracking confidence goes below a threshold which in this case is 0.1;
- remove all proposals whose overlap with the tracklet thus generated exceeds some threshold;
- choose the remaining detection with the highest score and use this as another anchor – not quite sure but apparently this process is repeated for all of the classes;
- augment each tracklet box in each frame by a combination of random perturbations and choosing the original proposals with overlap exceeding some threshold;
- obtain detection score on the original tracking box as well as all of the augmented boxes and retain only the one with the maximum score;
- TCN: 1D four layer fully connected CNN whose input is a timeseries made up of the detection scores, tracking scores and the anchor offsets (whatever those might be) and its output is the probability of this tracklet belonging to a particular class so that we need one TCN trained for each class;
- for supervision while training this network apparently of probability of one was used the box had been overlap exceeding 0.5 with the ground truth and zero otherwise this would hopefully give consistently high probabilities for all of the ground truth tracking its corresponding to each class in the corresponding TCN;
- the performance of the system seems to be quite a bit disappointing;
- using only the tracklet boxes results in around 37% mAP which is quite a bit lower than the 45.3% mAP obtained by simple still image detection;
- using tricks like suppressing nearby proposals and augmentations and other such stuff eventually increases the mAP to be around 45.2% which is a still slightly lower than the straightforward approach but now apparently only uses about 1/38 of the proposals which in some way supposed to be an advantage;
- incorporating the TCN increases the performance by 2% for all of the different methods tried out so that the best performing method gives around

47.5% mAP with the incorporation of the TCN;

- this is still quite disappointing especially the 1D convolution which seems to be kind of dodgy – these are supposed to help to enforce the temporal consistency of detection scores because it turns out that even on ground truth boxes, the detection scores show quite a bit of variation across frames and same for tracking scores so neither of these is reliable by themselves;