Action-Decision Networks for Visual Tracking with Deep Reinforcement Learning cvpr17

Monday, July 10, 2017 10:57 AM

1. The network predicts one of several movements of the object BB as actions and these are repeated in an iterative manner till the final location is found.
2. Network – called ADNet – is pretrained for appearance models using supervised training and then fine tuned using RL for dynamics modelling.
3. There are 11 different actions → 4 for left, right, up and down by an amount proportional to the w and h of the bb and 4 more for twice this amount; 2 are for scaling where the aspect ratio is maintained by increasing or decreasing the bb size by a proportion of the w/h too; one action is for stopping which is taken when the final state has been reached for the frame.
4. Rewards are only given in each frame at the end of the iterative process when the final position has been reached → it is a binary 1/-1 reward based on whether the IoU is > 0.7.
5. The state is made up of two components → the image patch under the bb which is resized to be $112 \times 112 \times 3$ and denoted as p_t and a history of the past actions → this is 110 dimensional since there are 10 actions and each is represented by a vector of size 10 in one-hot encoding
6. The initial network trained by the two step process is a VGG-M one since, apparently comparatively simple/shallow networks perform better with visual tracking than very deep ones.
7. Apart from the two step offline training, there is also an online fine tuning or adaptation of the network for the specific object being tracked.
8. SL → training sample is made up of patch (p), class label (c) and action label (a). The action label is chosen as the action that maximizes the IoU of the moved bb wrt the GT; the class label is 1 if this is > 0.7 and 0 otherwise

7/10/2017 9:52 PM

9. RL → can work in a semi supervised mode by assigning rewards to unlabeled frames on the basis of the result of the overall tracking simulation (or maybe even the nearest labeled frame)
10. Online adaptation is only performed for the fc layers as the conv layers are supposed to retain generic tracking information while the fc layers have object specific info.

7/11/2017 10:39 AM

11. Training samples for online training are generated from those frames where the tracking confidence > 0.5 and if it becomes < 0.5, redetection is performed by evaluating a bunch of samples around the last known location and selecting the one with the maximum confidence.

7/11/2017 12:48 PM

12. Training samples are generated by adding Gaussian noise to the GT box with variance proportional to the w and h for all of $[x, y, w, h]$ where the factor chosen was 0.3 for x, y and 0.1 for w, h
13. While pretraining 250 samples are drawn in each frame while for the online training 3000 samples are used in the first frame and 250 on each one thereafter
14. Online training is done once every 10 frames using samples from the last

20 frames (presumably only those that meet the confidence score criterion)

7/11/2017 3:34 PM

15 The ablation tests shows that the performance improvement using SL and RL over the base VGG-Net is very slight so that most of the learning is probably being done during online adaptation itself.

16 The overall performance of MD Net is roughly the same as the SOT though is speed - 3 fps - is apparently 3 times faster. The faster version with less dense sampling during online training is about 15 fps with only 3% lower performance.