## Online Multi-Object Tracking Using CNN-based Single Object Tracker with Spatial-Temporal Attention Mechanism 1708.02843 iccv17

2019-11-23 7:15:37 PM

Seems to be heavily inspired by the MDP tracker

the main idea is to have a CNN with some shared layers that are trained off-line as well as several target specific layers that are trained online for each new target

the target specific part is composed of two main parts called the visibility map and the spatial attention map where the former is implemented as a two layer network with a single convolutional and fully connected layer each while the latter has something called locally connected layer which is probably some version of fully connected

output of the spatial attention layer is used for ten channel wise multiplication with the ROI features whose result then goes into a binary classifier which again has two layers similar to the visibility map part but outputs a single number which is supposed to be the probability of this target being successfully tracked

In each frame, a motion model is first used to sample a set of candidate locations based on the last known location and these are then combined with all the nearby detections to obtain the set of candidate locations

ROI pooling is used to extract features for each one of these candidates which are then passed through the target specific part of the CNN to get a probability value for each Candidate whereupon the candidate with the maximum probability is chosen as the rough estimated location

If this probably is below a hard threshold, weighted average with the nearest detection is used as the refined location otherwise it is used by itself
this part is suspiciously similar to the lost versus tracked state decision that is made in MDP except that here a low probability target score is called untracked instead of lost

Training of the visibility map is done using synthetic data generated mostly in the first frame where the target is added by using Gaussian perturbation to generate negative samples and also using positive samples from other targets as well as random samples from the background as additional negative samples

There is also something called temporal attention module which seems to be a single sigmoid layer that takes as input the result of some heuristics based function of things like the average visibility to predict the probability of the target being occluded

if this probably is high, then the corresponding feature is used for training with a low weight while if the probably the is very low then it is added to the historical samples set for this target

The criterion for adding new targets as well as for removing existing one seem to be pretty much the same as MDP

Motion model seems to be riddled with heuristics but the main idea seems to be to maintain some sort of Gaussian distribution of locations and velocities and use these to sample possible locations which is usually quite dubious on real videos

Performance improves by about 4% points in terms of MOTA but is actually lower for MOTP, MT, ML as well as FN while being marginally better in terms of FP as well as ID switches