- Previous trackers used online model to learn appearance of arbitrary object from video which limits model's power. Training offline also offers speeds beyond real-time.
- Due to data deficiency only smaller models can be learnt and deep learning cannot be applied. Deep learning was used as shallow method or with fine-tuning. Shallow methods lose perks of end-to-end and fine-tuned ones are slow.
- Siam FCs find exemplar image(initial BB) within search img using FC for efficient cross correlation by sliding windows. Similarity func learnt by training on vids with obj paths. It is cross-correlation + bias. FC => commutative wrt translation.
- Displacement of target = stride*dist b/n prev and current pos. Conv net is trained using logistic loss over many +ve -ve pairs. Exemplar and search img pairs are extracted from 2 frames of a vid utmost T frames apart. Scale of obj is normalized retaining aspect ratio, empty space is filled with mean RGB value. Ex = +ve if dist from prev centre <= R.
- Losses for +ve and -ve are weighted to tackle class imbalance. Dense sliding windows => no bias toward centre. Search img near centre as generally obj is adjacent to prev fr. Uniform size is assumed for both imgs for simplicity.
- ILSVRC has targets diff from other datasets, so tracker can be trained on it to set benchmarks on other ones without overfitting.
- (BB + context) is scaled to a constant area image. Padding is not used as it violates the FC property. ReLU, BN and max pooling used. To prove efficiency of Siam, simple algo was used.
- Search space is 4 times size of object in prev frame. Cosine window is applied to penalize large displacements. Scaled versions of search image are processed.
- Siam net can be interpreted as unrolled RNN of len 2.It can be used to initialise RNNs. SO-DLT and MDNet fine tune pretrained net (learnt offline) as detector but very slow as many samples have to be processed. Shallow methods are slow due to high dim embed.
- GOTURN is trained to directly find BB of obj in 1st img in the 2nd img. It has advantage of no need of exhaustive scale and aspect ratio search. But it does not have translational invariance. So it is fed with obj in diff positions by data augmentation.
- Other nets have similar approach but not translationally invariant due to linear layers.
- Exemplar is fixed as no benefits were gained by changing it by simple strategies. As 17x17 score map is coarse, it is upsampled to 272x272 for accurate localization. Scale change is addressed by linear interpolation with damping.
- Success plots for OPE (one pass evaluation), TRE (temporal robustness evaluation) and SRE (spatial robustness evaluation) were calculated for OBT-13 for which 25% of training image pairs are converted to grayscale.
- In VOT, trackers are re-initialised 5 frames after IoU=0. Eval by acc Vs robustness in VOT-14; expected avg overlap (avg IoU before failure) in VOT-15.