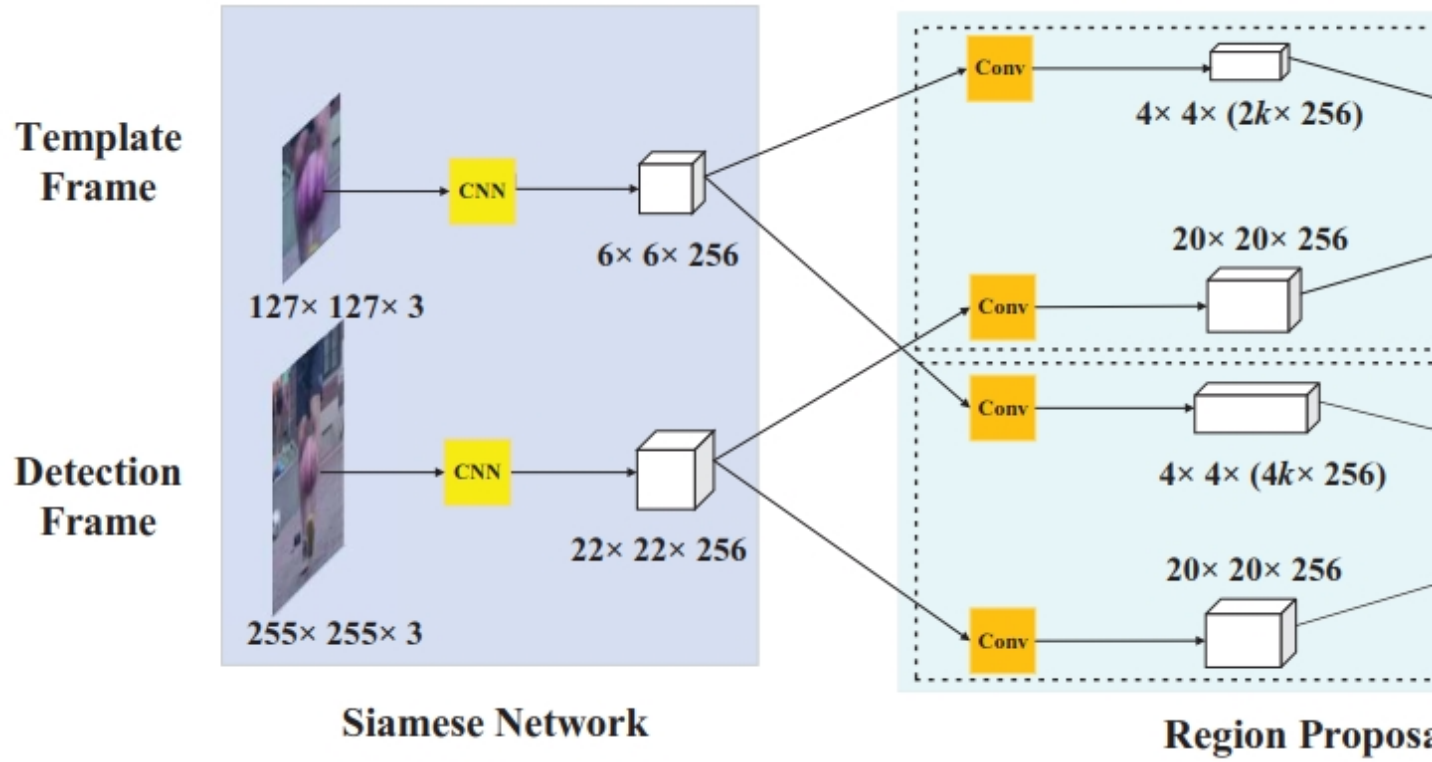
2019/05/19 2:48:12 PM



combines the region proposal network of faster RCNN with the Siamese architecture to be able to do away with some computationally expensive parts of the latter during inference to speed it up – up to 160 PS it seems

output feature maps produced by both branches of the Siamese network are passed through both the classification and regression subnetworks of the RPN

output of the template branch is reduced to N different groups of 4x4x256 filters where N=2k for the classification branch and for the N=4k regression branch

these filters are respectively used as convolutional/correlation kernels for the output produced from the detection branch feature map which is 20 x 20 x 256

during inference, they perform what they call *one shot detection*, using what they call *meta-learning,* though in practice all they are doing is to pass the template patch extracted from the 1st frame 2 the template Siamese branch and use its output as kernel representing the general appearance of all the object in its category there was seen during training and therefore be able to track novel views of the object;

Training is done end to end by using the patch from the history frame for the template branch and patch from the current frame or the detection branch; training is done on general large data sets like the YouTube bounding box and ImageNet video while the Siamese network is also pre-trained on the standard ImageNet;

A bunch of heuristics are used to prune and rank all of the proposals produced by the RPN and finally choose the best one as the new object location after doing something they call size updating by linear interpolation to ensure smooth shape change, though this is not quite clear