Reviewer: Ling Zhang, Feburary 6, 2020

Citation: Bo li, Junjie Yan, Wei Wu, Zheng Zhu, Xiaolin Hu, "*High Performance Visual Tracking with Siamese Region Proposal Netwrok*", CVPR2018.

Quick Summary: The authors use a AlexNet which modified that first three conv layers with fixed parameters and last two conv layers with fine-tune parameters. They extract the image pairs from VID and Youtube-BB and prove the model on VOT2015, VOT2016, VOT2017 real-time experiment, OTB2015. The Siamese RPN perform well on all of them in both accuracy and efficiency.

Ideas/Approach/Result:
Idea: The authors find that most of state-of-the-art deep learning based tracker cannot get top performance with real-time speed. So there find the way which is ***Siamese-RPN*** can perform well on recent video tracking challenges.

They introduce: ***Siamese-RPN*** which consists of *Siamese subnetwork* and *RPN subnetwork* . The ***Siamese-RPN*** use region proposal network which first proposed in ***Fast-RCNN*** to replace FC in ***Siamese-FC***. The ***Siamese-RPN*** trained with large scale image pairs off-line then save to the kernel. Because the off-line training, it can use large data sets. More data can help to get even better performance. During online tracking part, proposed framework is formulated as a local one-shot detection task which can refine the proposal to drop the expensive multi-scale test. The most of the efforts on this paper are mentioned above.

Methods: The ***Siamese-RPN*** is divided into two parts: *Siamese subnetwork* for feature extraction and *Region proposal subnetwork* which consists proposal classification branch and regression branch for proposal generation.

In *Siamese subnetwork*, the authors modified AlexNet and remove its groups from conv2 and conv4. One main point here, the fully convolution network must have none padding. The *Siamese subnetwork* consists of template branch and detection branch share parameters in CNN. The authors need two patches are encoded by the same transform. This is why they share the same parameters. The outputs are feature maps of Siamese subnetwork.

In *Region Proposal subnetwork*, the supervision section have two branches, foreground-background classification and proposal regression. The authors use correlation section to separate *Siamese subnetwork* output to two branches; one for classification and one for regression. For here, they use two convolution layers to do this operation, but for output of template frame, the conv layers increase the channel, for detection frame output, they do not. For classification, they cross-entropy loss function that is used in *Faster R-CNN*. For regression, they adopt smooth L_1 loss with normalzied coordinates.

They use the template branches' outputs as kernels for local detection. These kernels are pre-computed on initial frame and will not change during tracking period. Based on feature map convolved by kernels, the detection branch use one-shot detection to get top proposals. The authors also proposed two strategies to choose the proposals.

Result: The authors implement the ***Siamese-RPN*** and test it on state-of-the-art challenges including: VOT2015, VOT2016, VOT2017 real-time experiment, and OTB2015. They compared their tracker with top 10 trackers that few trackers can tracker with real-time speed, however, their trackers can. Also their tracker's accuracy is pretty good.

Relate to what we have been discussing in class: The ***Siamese-RPN*** deal with the proposal selection problem which we mentioned in toady's class. Also, it use CNN in *Siamese subnetwork* to do feature extraction which related to our discussion about the convolutions neural network related problems.

Questions: After reading the paper, I am interested in how does the correlation layer works. Why they

choose correlation layer? Is there any other choice except AlexNet can perform well in pre-trained part? Can the models more deeper? I think I still need more reading to solve these doubts.