Description of ALTO for VOT 2019

June 6, 2019

The key intuition behind the proposed approach is that the learning based frameworks such as Fully Convolutional Siamese Network (SiamFC) [1] can be greatly improved by an additional adversarial learning through proper discriminators. Such an adversarial learning can bring in more precise localization or better fine tuning of the target location and scale. Based on the formulation of discriminator it can be used for information fusion in learning, and with additional training strategies to the generator, learning can be made more robust and generalized. In the proposed framework, unlike the conventional GANs, the generator ($\mathbf{G}$) is a learning based tracker (specifically we adapt SiamFC) which shall provide a rough estimate of the target location and the adversarial learning is used to discriminate the generated track and predicted target wrt ground truth.

1 Training

The exemplar patch ($\mathbf{Z}$) and the search image ($\mathbf{X}$) are passed through same architecture Θ (CNN) and are finally correlated to obtain the response map . The final response map is a scalar valued function in which each value indicates the similarity score for a positive sample or for a negative sample. Then using the indices of the maximum value in response map (x_p, y_p) the possible location of the target is found in the search image $(x_p^{'}, y_p^{'})$ and with this location we crop a 127x127 patch (fake sample) from the search image and feed it to the discriminator ($\mathbf{D}$). During the training phase the exact location of the target is known using which again we crop a 127x127 patch (real sample) ($\mathbf{R}$) from the search image and feed it to the discriminator ($\mathbf{D}$). We also compute mean square error (L_{mse}) between predicted location and actual ground truth. Finally, adversarial learning is initiated to optimize the cost function (Eq. 1) in order to train the generator ($\mathbf{G}$) so that it predicts the accurate location of the target while tracking. Our abalation study shows that in either case i.e. ALTO without (L_{mse}) or ALTO with (L_{mse}) outperform the baseline approach and ALTO with (L_{mse}) achieves the best performance on benchmark datasets. Note that $\mathbf{D}_\Theta$ and $\mathbf{G}_\Theta$ are parameters of discriminator and generator respectively. Our tracker is trained on ImageNet VID dataset.

$$\min_{\mathbf{G}_\Theta} \max_{\mathbf{D}_\Theta} \left\{ \mathbf{E}_{Z,X,R} \left(\log\Big(\mathbf{D}(R; \mathbf{D}_\Theta)\Big) + \log\Big(1 - \mathbf{D}\Big(\mathbf{G}(Z, X; \mathbf{G}_\Theta); \mathbf{D}_\Theta\Big)\Big) + L_{mse}(\mathbf{G}_\Theta) \right) \right\} \tag{1}$$

2 Tracking

During tracking we retain only the generator and a search image centered around previous target location is used for prediction of target's location in the current frame. The position of the maximum value in the score map relative to the centre of the score map multiplied by the stride of the network gives the displacement of the target.

References

[1] Luca Bertinetto et al. *Fully-convolutional siamese networks for object tracking.* Springer, 2016.