2. Fast online object tracking and segmentation: A unifying approach

URL: 1812.05050.pdf (arxiv.org)

Key terms: Visual object tracking, semi-supervised video object segmentation, fully convolutional Siamese network
Benchmark Datasets: VOT-2018, DAVIS-2016, DAVIS-2017
Review Author: Nimish Magre
Contribution: Real-time object tracking and segmentation at 55 frames ps
Introduced a Fully convolutional 3 branch Siamese network with offline training to produce a pixel-wise binary mask, object bounding box and an object classification score
.
Self ideas:

- Could improve by fine-tuning regularization parameters for the final Siamese loss function
- Could utilize GRUs to make use of information from previous frames for better results

Summary:
The 3 branch Siamese convolutional network (SiamMask) proposed in the paper builds upon earlier work with a single branch Siamese network (SiamFC) and a 2 branch Siamese network (SiamRPN).
SiamFC compares an exemplar image '*z*' against a larger search image '*x'* to obtain dense response map. $g_\theta(z, x) = f_\theta(z) * f_\theta(x)$. The network produces 'n' such maps corresponding to the n-th candidate window in 'x'. Consequently, the highest response map value is supposed to correspond to the target location in search area '*x*'. SiamMask replaces the cross-correlation used in SiamFC with depth-wise cross-correlation and uses logistic loss (L_{sim}) for training.
SiamRPN (Regional Proposal Network) improves on SiamFC performance by estimating a bounding box along with the target location using a 2 branch Siamese network. It uses the L1 loss (L_{box}) for the bounding box branch and the cross-entropy loss (L_{score}) for the classification score.
SiamMask adds a 3rd branch that produces a (w * h) binary mask for each of the 'n' candidate windows in 'x' using a 2 layer neural network. $m_n = h_\Phi(g^n_\theta(z, x))$ where 'm' represents the corresponding (w*h) mask and 'h_Φ' is the network function with learnable parameters 'Φ' . Each of the 'n' windows has a ground-truth binary label '$y_n \in \{\pm1\}$' and a pixel-wise ground-truth mask '$c_n^{i,j} \in \{\pm1\}$' of size (w*h). It uses a binary logistic regression loss $L\text{mask}(\theta, \phi) = \sum_n \frac{1 + y_n}{(2wh)} \sum_{i,j} \log(1 + e^{-c^n_{i,j} m^n_{i,j}}))$
For experimentation, the total loss for the 3-branch network is
$L_{3B} = \lambda_1 \cdot \text{Lmask} + \lambda_2 \cdot \text{Lscore} + \lambda_3 \cdot \text{Lbox}$ where the 'λ' values are prefixed.

Pros:

- Real-time object tracking and segmentation
- Higher speed compared to previous online tracking and segmenting networks (55 frames ps)
- Only requires a single bounding box initialisation during testing
- Use of simple 2 layer CNN for mask representation and a 3 branch Siamese network
- Doesn't require fine-tuning of hyperparameters for the final Siamese loss function

Cons:

- Fails in cases of motion blur and 'non-object' instances
- Offline training requires millions of videos in dataset
- Does not adjust to multi-object tracking