Online Resource 4

June 7, 2021

1 Implementation Details

Visual Siamese Tracking module uses pretrained SiamMask [3] weights to generate the proposals without further fine-tuning. The language-guided module use the binary mask for the training of DMN [2]. In order to maintain the heterogeneity of the language and visual branches, they are separately trained and not combined until the confidence score is obtained by each branch.

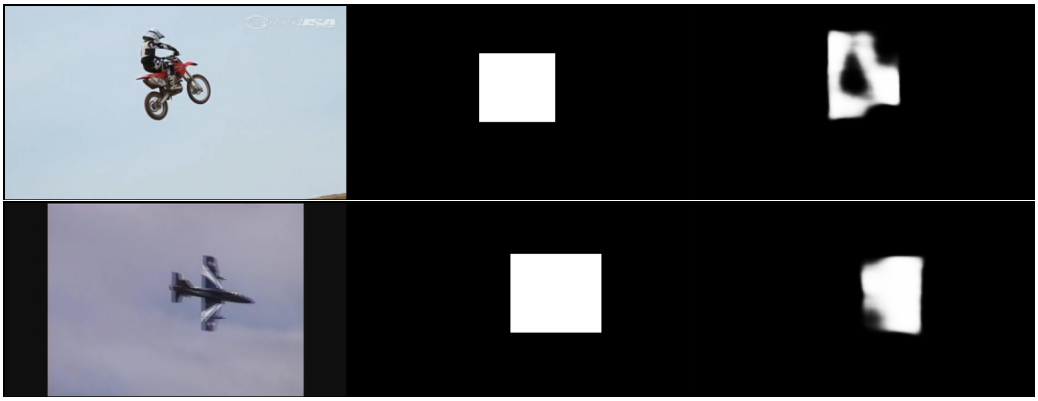

Figure 1: Left is the original frame, middle is the binary mask we convert from Ground Truth and right is the predicted result. Upper one of the referring expression is *motocycle in the middle* while the lower one is *the airplane in the middle*.

We use following steps for training and inference to output the language-guided bounding box.

1)We first turn the ground-truth of mixture of both part of OTB and LaSOT data into the binary mask as ground-truth for DMN training. As the tracking dataset provides only the bounding box as the ground truth, we convert the coordinates of [cx,cy,w,h] into the binary masks as we showed in Fig.1.

2)Based on the pretrained weights from ReferIt, we train the module with mixed dataset to fine-tune the weights, in order to make the model to fit in the tracking tasks.

3)In inference stage, the network output the predicted mask. Due to the value of each pixel in the mask range from 0 to 255, as showed in Fig.2, we regards the value over 128 pixel as positive

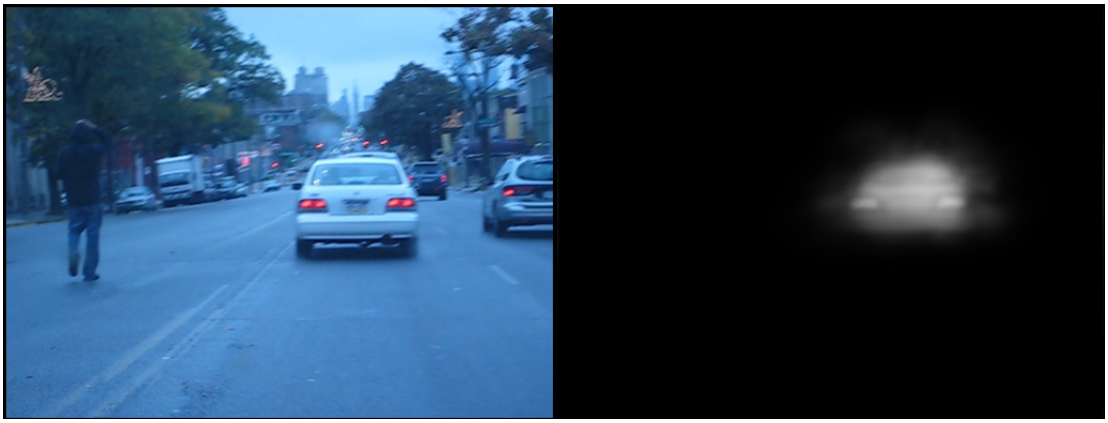

Figure 2: Left:The car in the frame to be tracked. Right: The output of LGM module with referring expression **the white car in the middle**.

and else as negative samples to calculate the location of the mask. As a result,a bounding box is generated by calculating Minimum Enclosing Rectangle as a result based on the mask, as showed in Fig.3.

Figure 3: Left:Output of the mask with red Mininum Enclosing Rectangle. Right: The output in the frame.

In terms of training dataset for Language-guided module, we employ a mixture of data containing 36 training sequences from 22 randomly selected with OTB and 14 subsets with LaSOT , with a total of 39,891 frames. Training of Language-guided module has two stages, first training with low-resolution scale without upsampling and then with high-resolution. Both word embedding size and hidden state size are set to 1000, and the number of dynamic filters is set to 10. The training is done with Adam optimizer [1] with an initial learning rate of 1×10^{-5}, and batch size of 1 image-query pair. We fine-tune the model in low resolution with 5 epochs and in high resolution with

10 epochs. For generalization improvement, we choose one video from each category in LaSOT to further training in high resolution. We actually tried several ways to fine tune the module and use the best weights based on our evaluation results for LangTrack.

All the experiments are implemented based on PyTorch 1.0 and CUDA 10, I7-7700, Nvidia GeForce RTX 2080Ti with 8G RAM.

References

[1] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014.

[2] Edgar A. Margffoy-Tuay, Juan C. Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. 2018.

[3] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H. S. Torr. Fast online object tracking and segmentation: A unifying approach. 2018.