Fraunhofer IOSB

FRAUNHOFER INSTITUTE OF OPTRONICS, SYSTEM TECHNOLOGIES AND IMAGE EXPLOITATION

Search Area Reduction Faster R-CNN for Fast Vehicle Detection in Large Aerial Imagery

Lars Sommer, Nicole Schmidt, Arne Schumann and Jürgen Beyerer

Motivation

- Region proposals are predicted for each feature map location
- Increasing the resolution of the used feature map as required for vehicle detection in aerial imagery results in a strong increase in inference time
- However, many image regions, particularly in rural areas, do not contain any vehicles
- We extend Faster R-CNN [1] by a Search Area Reduction module (SAR) to identify and filter out regions that do not contain any vehicle

Approach

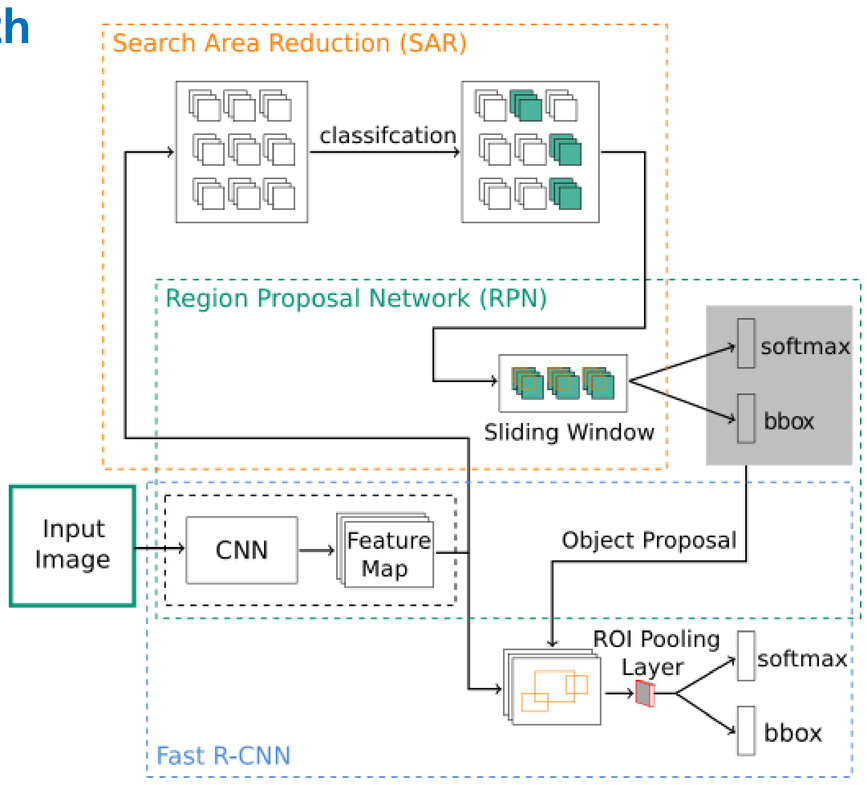

Fig. 1: Schematic structure of our proposed framework comprised of Faster R-CNN and the integrated SAR module to directly reduce the search area. Our custom layers and network architecture will be made available at: https://github.com/vehicledetect/sar-frcnn

- The SAR and Faster R-CNN are merged by sharing the convolutional layers
- The novel *sub-division-SAR* layer, which is applied on the last convolutional layer, is used to divide the input image into sub-regions
- For each sub-region, we predict a confidence score of how likely a region contains at least one object, which is then used to forward an adaptive set of regions to the actual detection module
- The *sub-division-RPN* layer stacks all relevant regions into one batch
- The *merge* layer maps the coordinates of the generated proposals given inside each region, are mapped to their position in the original image
- All mapped proposals are then classified by the Fast R-CNN detector [2]

Adaptation to Aerial Imagery

- conv3_3 of VGG-16 [3] is used as feature map to increase the spatial feature map resolution required to accurately localize small objects
- We set the minimal proposal height/width to 4 and anchor base size to 2

Two-stage Training

- Direct multi-task learning is not possible as Faster R-CNN requires at least one GT object per training image
- First, Faster R-CNN is trained end-to-end for a total of 60,000 iterations
- Second, the SAR classification module is trained on sub-images of size 224×224 pixels that either contain objects or not
- The weights of the shared convolutional layers are kept fixed
- For deployment, the sub-division-SAR layer, sub-division-RPN layer and merge layer are added to the framework

Detection Results

Table 2: Comparison of Average Precision and inference time of Faster R-CNN with and without SAR. All experiments are performed on the publicly available VEDAI dataset [5] with a single NVIDIA Titan X GPU.

Approach	AP (in %)	Time (in ms)	
Faster R-CNN	97.4	Total	321
		RPN	205
		Fast R-CNN	86
SAR + Fast R-CNN	97.3	Total	194
		Subdivision_SAR	3
		SAR	64
		Subdivision_RPN	11
		RPN	42
		Fast R-CNN	26

- Similar AP for Faster R-CNN and SAR + Faster R-CNN while the overall inference time is clearly reduced
- The inference time for the RPN is almost reduced by 80% due to the clearly reduced number of feature map locations used for prediction
- Most image regions are filtered out by the SAR module as on average only about 0.2% of an image are covered by vehicles.
- The slightly worse AP is due to objects at edges of sub-regions, which are partially visible and thus are sometimes misclassified as vehicles

Qualitative Results

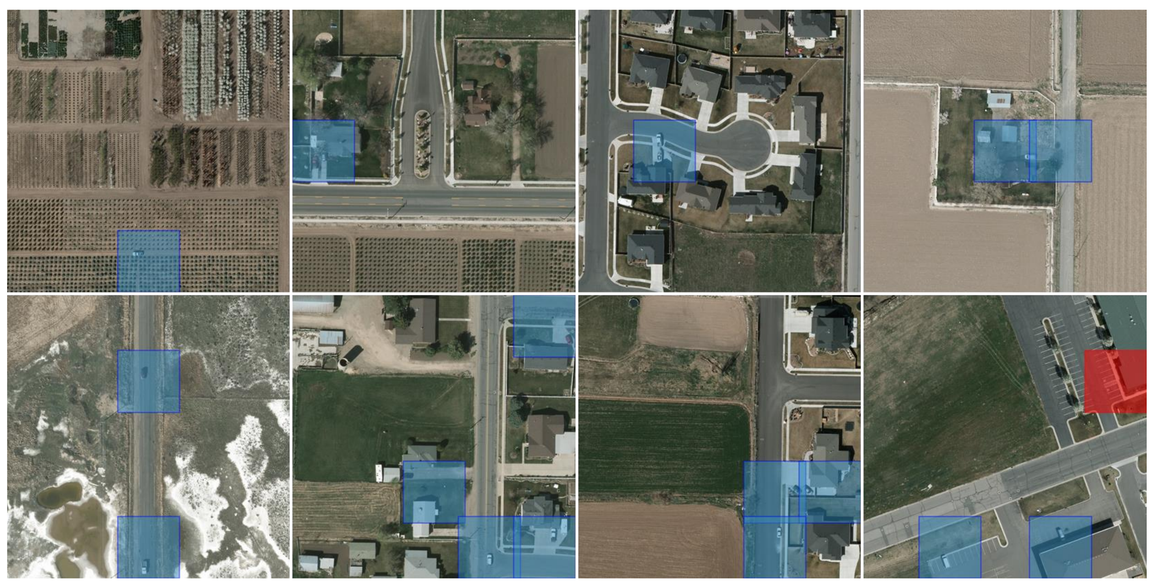

Fig. 3: Qualitative examples of the SAR module exhibit that the search area is clearly reduced. Highlighted regions are considered for detection. Regions labeled in blue contain at least one vehicle whereas the region labeled in red contains no vehicle.

Conclusions

- We propose an extension of Faster R-CNN for vehicle detection in aerial imagery that adaptively reduces the search area and the inference time
- State-of-the-art detection results are achieved on a publicly available dataset while the inference time of the detection module is reduced by more than 75%
- Implementation of all additional layers in CUDA can further reduce the overall inference time

References

[1] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, 2015.

[2] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[3] S. Razakarivony and F. Jurie. Vehicle detection in aerial imagery: a small target detection benchmark. Journal of Visual Communication and Image Representation, 2016.

M. Sc. Lars Wilko Sommer
Video Exploitation Systems
Tel +49 (0) 721 6091-657
lars.sommer@iosb.fraunhofer.de