Joint 3D Proposal Generation and Object Detection from View Aggregation

IROS 2018

Summary

In this paper, AVOD is put forward as an **A**ggregate **V**iew **O**bject **D**etection architecture specifically designed for autonomous driving purposes. The network leverages LiDAR measurements and RGB images to predict accurate 3D bounding boxes. Feature maps are extracted from both input views which are then used in the process of proposal generation by the Region Proposal Network (RPN) as well as the prediction of the final bounding boxes. The RPN is shown to produce high-recall proposals and AVOD achieves state-of-the-art performance on the car class of the KITTI 3D Object Detection Benchmark.

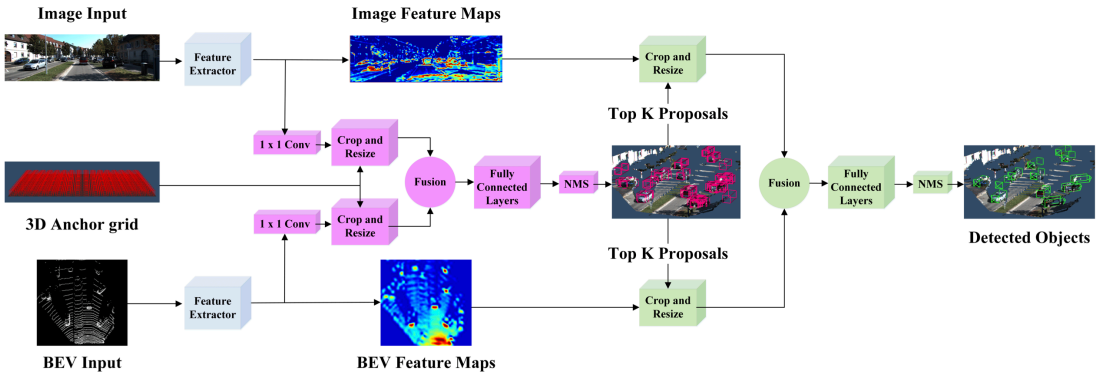

Figure 1: Illustration of the AVOD architecture. Proposals and final detections are obtained by projecting regions onto the feature maps extracted from two input views.

Main Contributions

- **Novel Feature Extraction Architecture** For both input streams, BEV and RGB, AVOD relies on a feature extraction network inspired by VGG-16. However, the encoder down-samples the input image by a factor of 8 which results in object sizes too small for accurate detection. Consequently, a decoder network modeled after the Feature Pyramid Network (FPN) is used to up-sample the feature maps back to the input dimensions.
- **Multi-modal Region Proposal Network** The extracted features from both views are used in a high-recall region proposal network. By cropping and resizing the feature maps corresponding to 3D bounding box anchors, same-length features can be obtained for all anchors which are fused using element-wise mean before being passed to a regression head that outputs axis-aligned bounding box proposals along with an objectness score.
- **Corner Encoding** Contrary to the previously proposed bounding box encoding using eight corner points, an encoding scheme is put forward based on four corner points and two height values (see Figure 3). Thus, the geometric constraints of the bounding box can be accounted for and the regression vector can be reduced from 24 values to 10 values. The orientation of the bounding box is computed by regressing $\cos(\theta)$ and $\sin(\theta)$ which results in an unambiguous representation of the orientation.

Implementation Details

- **Anchor Generation** Axis-aligned anchor boxes are generated by sampling center coordinates at an interval of 0.5m. The size of the boxes is computed based on class-dependent statistics extracted from the training set. By removing empty anchor boxes, a total of 80-100k boxes are obtained for each scene.
- **Crop and Resize** In order to be able to fuse two input representations, fixed-length features are required. For each anchor box, a 3x3 feature grid is obtained for each input view by projecting the region of interest onto the feature maps and applying bi-linear interpolation.
- **Bounding Box Refinement** For the prediction of the final bounding boxes, an approach similar to the proposal generation is employed. The proposals are projected onto the respective feature maps for each input view. However, due to the smaller number of proposals compared to the number of anchors, feature maps of size 7x7 are extracted and more channels are used. These feature maps are fused using element-wise mean and passed through several fully-connected layers to regress the final bounding boxes.

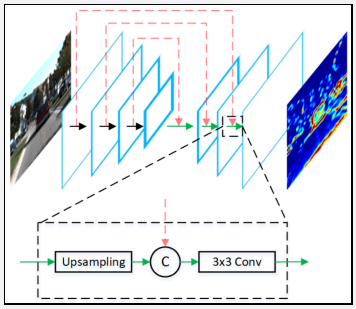

Figure 2: Illustration of the feature extraction pipeline. Skip-connections are employed to ensure highly accurate up-sampling of the feature maps.

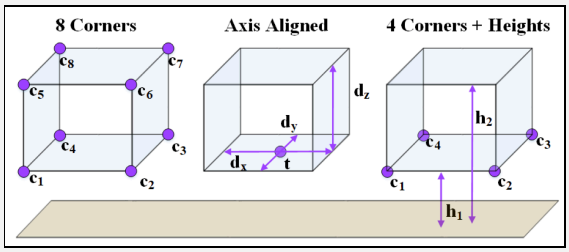

Figure 3: Illustration of the proposed bounding box encoding scheme using four corners and two height values.

Evaluation

- **3D Proposal Recall** A high recall of region proposal networks is required since missed detections cannot be recovered at a later stage. Recall is evaluated at an IoU threshold of 0.5 and the performance is compared to 3DOP, Mono3D and MV3D. It is shown that AVOD outperforms all other methods by large margins. For example, a recall rate of 86% on the car class is already achieved with 10 proposals per frame. Using 50 proposals, a recall of 91% can be achieved. For comparison, MV3D obtains the same recall rate at 300 proposals per frame.
- **KITTI 3D and BEV Object Detection** The performance of AVOD is evaluated on the car, pedestrian and cyclist classes of the KITTI Benchmark. The authors compare AVOD to MV3D, VoxelNet as well as Frustum-PointNet. AVOD achieves state-of-the-art performance on 3D detection of the car class while running at 10 fps which is almost twice as fast as the second fastest method. On the pedestrian class, AVOD's performance is on par with Frustum-PointNet's. However, AVOD struggles on the cyclist class, consistently scoring 4-8% AP less than Frustum-PointNet.

References

This summary is solely based on my understanding of the original paper. All images used here are taken from the original paper as well. The paper can be found under the following link:
`https://arxiv.org/pdf/1712.02294.pdf`