PIXOR: Real-time 3D Object Detection from Point Clouds

CVPR 2019

Summary

In this paper, PIXOR is put forward as a fast single-stage 3D object detector. The architecture is specifically designed for autonomous driving applications as it leverages the fact that all objects are located on the ground plane to project the input to a BEV representation that lends itself to the processing by conventional 2D-CNN architectures. Moreover, PIXOR manages to predict highly accurate bounding boxes without additional hyper-parameters inside the network while allowing for real-time detections at 10 FPS.

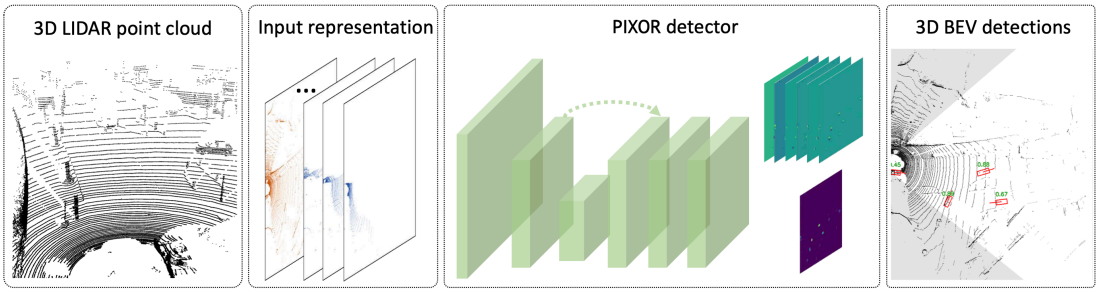

Figure 1: Illustration of the PIXOR pipeline. The raw input point cloud is transformed into a voxelized BEV representation which is fed to the CNN architecture. Based on a rich feature representation, dense bounding box predictions are computed along with class probability scores.

Main Contributions

- **BEV Projection** In order to be able to apply discrete convolutions, the input point cloud is voxelized. To reduce the computational complexity, the resulting grid is treated as a 2D image using the height dimension as feature channels similar to RGB images. Thus, a conventional 2D-CNN architecture can be applied. The decision to drop the third dimension is reasonable since for autonomous driving applications all objects are assumed to be located on the ground plane. The voxelization is performed at a resolution of 0.1m, taking into account [0, 70m] x [-40m, 40m] x [-2.5m, 1m], resulting in an input tensor of dimension [700, 800, 36]. The third dimension includes an additional reflectance channel.
- **Simplified Pipeline** The authors present a simplified detection pipeline that produces accurate bounding boxes while allowing for real-time point cloud processing at 10 FPS. PIXOR is a single-stage dense object detector with no need for tuning hyper-parameters such as anchor boxes or number of generated box proposals. The architecture consists of two main components: 1) a backbone CNN based on residual units; 2) a 2-headed prediction network for object classification and bounding box prediction.
- **SOTA on KITTI BEV Object Detection Benchmark** PIXOR achieves state-of-the-art performance on the KITTI BEV Object Detection Benchmark, outperforming previous methods by large margins at certain categories while being considerably faster.

Implementation Details

- **Backbone Network** The backbone network follows standard design principles for residual networks. The authors stack four residual blocks of down-sampling factor 2. However, problems arise w.r.t. the size of cars at a large distance to the sensor. Due to a low number of pixels in the down-sampled feature maps, the network has to deal with considerable information loss regarding location and size of the car. Consequently, a *Feature Pyramid Module* is employed, that fuses information from different scales to preserve the information contained in the original representation.

- **Header Network** The header network is a simple architecture consisting of four convolutional layers, followed by one classification branch predicting class probabilities using sigmoid activation and a second regression branch which computes bounding box encodings without non-linearity. The classification problem is supervised using cross-entropy while smooth l1-loss is used for the box regression.

- **Bounding Box Encoding** The bounding box regression is factored into 6 values. Since all objects are assumed to be on the ground plane, center location and size along the z-axis are omitted. The center location in the xy-plane is encoded using the logarithm of the offset between the prediction voxel and the ground truth center in real-world metric space. Similarly, width and length are encoded as the logarithmic offset from the average values across the training set. In order to keep the orientation of the bounding box within $[-\pi, \pi)$, the orientation is factored into $\sin(\theta)$ and $\cos(\theta)$.

Evaluation

- **BEV Object Detection** The authors compare PIXOR to several other point cloud-based methods, including an extensive comparison with MV3D. For an objective assessment, an evaluation based on the distance of the detections to the ego-car is put forward. The results of the comparison with MV3D are displayed in Figure 2. It can be seen that PIXOR outperforms MV3D by large margins across the entire detection range. Moreover, PIXOR runs about three times faster than MV3D.

- **Regression Loss Analysis** A second regression loss is put forward based on the decoding of the bounding box during training and the calculation of the loss based on the corner locations of the bounding box. An ablation study shows that pre-training with smooth l1-loss and fine-tuning using the so-called *decoding loss* results in considerable performance gains.

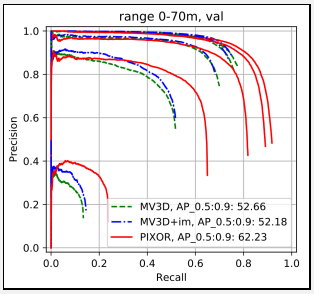

Figure 2: Results of the MV3D comparison.

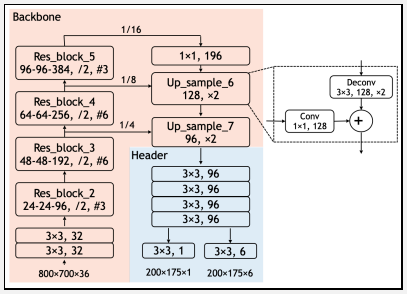

Figure 3: Overview of the PIXOR architecture.

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/1902.06326.pdf`