Machine Learning Engineer Nanodegree - Capstone Proposal

Micha Bruns

January 31st, 2018

Proposal

Domain Background

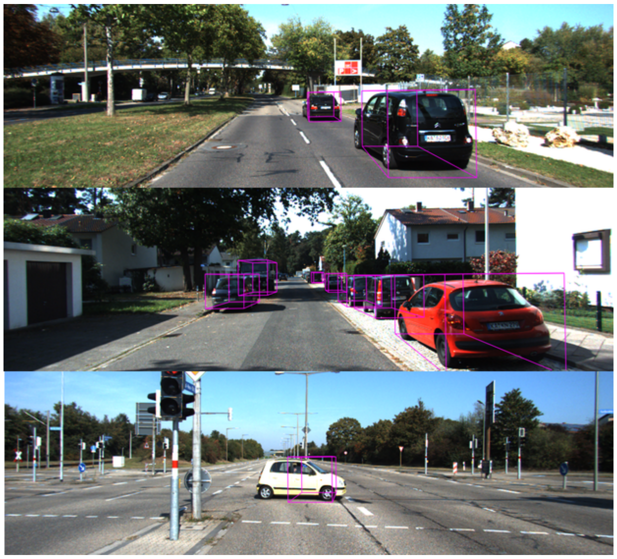

Figure 1: Example results of the "Hand-crafted" baseline

For autonomous driving the perception of other traffic participants, such as cars or pedestrians, is of immense importance. Examples of prominent approaches that tackle this problem in the image space are the *Single Shot MultiBox Detector* (Liu et al. 2015) and *Faster R-CNN* (Ren et al. 2015). While the resulting 2d bounding boxes are sufficient for most tasks, safety relevant applications need precise 3d detections. This is true especially for autonomous driving where a small error may lead to collisions and injuries.

While 2d approaches have grown in maturity for the last years and ready-to-use pretrained detectors are available online, for example Huang et al. (2016), 3d detection is still a point of research. Recent publications are trying to close this gap and show promising results. The most successful approaches are either fusing laser scanner data with camera data (see Ku et al. (2017), Chen et al. (2016)) or solely rely on laser data and make use of powerful feature extraction techniques (see Zhou and Tuzel (2017)).

Problem Statement

In the automotive context, **3d object detection** refers to the detection of *cars, trucks, vans, pedestrians*, and other classes that are relevant for an advanced driver assistance system or for a self driving car. In contrast to a 2d detection task, 3d detection tasks includes the prediction of 3d coordinates of the objects.

The goal of this capstone project is to create a *deep neural network* that detects *cars* in the 3d space with a high accuracy based on light detection and ranging (Lidar) and RGB data.

Datasets and Inputs

The most prominent dataset that offers lidar and rgb data in the automotive context is the kitti dataset(see Geiger, Lenz, and Urtasun (2012)). This dataset was recorded end of 2011 in south of Germany with a Volkswagen Passat B6 that was equiped with 2 color cameras and one Velodyne HDL-64E laserscanner.

Groundtruth for cars and many other object classes are available.

Since I want to focus on the actual deep learning part of this project and not on the data reading and preprocessing, I will use existing code from a team that perticipates in the Udacity didi challenge (BostonDidiTeam 2017). This includes the parsing of the kitti raw dataset and projection functionalities - such as top view projections. Even though the existing repository mentions that they want to implement the MV3D approach (Chen et al. 2016), there are notable difference between their implementation and the publication: the lack auf data augmentation, a different fusion scheme, no default front view projection, and interchangeable feature feature encoders (VGG16, Resnet50, and more).

Solution Statement

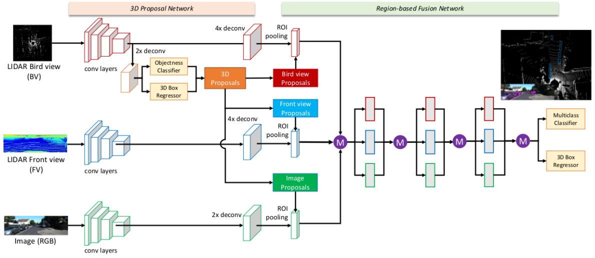

Figure 2: MV3D Architecture

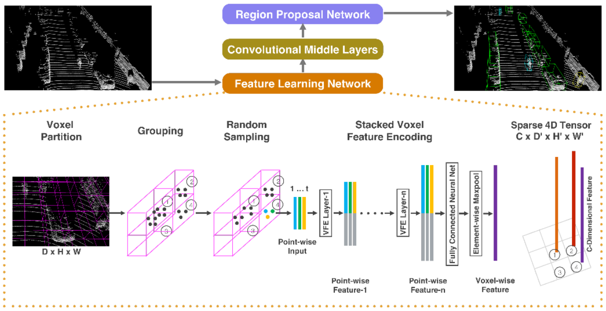

Figure 3: Voxelnet Architecture

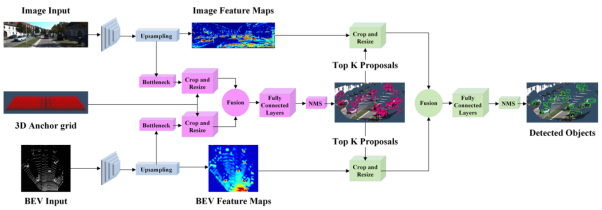

Figure 4: AVOD Architecture

Three different and very successful solutions are (Chen et al. 2016), (Zhou and Tuzel 2017), and (Ku et al. 2017). The respective architectures are depict in the figures 2, 3, and 4. While the MV3D and AVOD architectures use a two stage detection approach where candidate bounding boxes are generated either based on combined features or solely on lidar information and the final classification and regression is only applied on these candidates, the Voxelnet architecture is simpler and follows the approach of a single shot detector.

I want to focus on the AVOD approach because it has the cleanest architecture, uses a combination of rgb and lidar features to generate bounding boxes, and offers the possibility of further extensions. The AVOD approach works as follow:

Features are extracted independently from the rgb image and the top view projections of the corresponding lidar point cloud using VGG16 encoders (see (Simonyan and Zisserman 2014)). For each anchor (see Ren et al. (2015) for a definition of anchors), compressed versions of these features are fused and fed into a fully connected layer followed by a non maximum suppression. The top scoring anchors are kept as proposals and the original features from both data sources are fused as well and used for the final classification and regression output.

Benchmark Model

As a benchmark model the "hand-crafted baseline" network from Zhou and Tuzel (2017) has been implemented and trained with good results. The architecture consists of a simple convolutional part that extracts features from top view projections of lidar point clouds, and a region proposal network that regresses and classifies the resulting 3d bounding boxes. Example detectoins from the validation set can be found in figure 1.

Evaluation Metrics

For evaluating the detection performance I will follow the procedure of the kitti benchmark and use the average precision as the main metric. This is defined similar to the Pascal VOC average precision metric, as the mean precision sampled at equal distant values of the ROC curve. Details on this evaluation metric can be found in (Geiger, Lenz, and Urtasun 2012). Tools for the evaluation are available on the website of the kitti benchmark.

Project Design

At the time of this proposal, the Voxelnet baseline, which will be used at the main benchmark model, has been trained for about 100 epochs (roughly 3 days on a nVidia GTX 970Ti). While the results are already good, they are not on a par with the published results. This may be due to the missing data augmentation, a shorter training time, or an error in the implementation. It is important to rule out the latter because parts of the code will also be used in the final architecture. However, for a benchmark model it is sufficient to archieve satisfying results without the need of reproducing the published results.

Assuming that the existing code is error free, the AVOD architecture can be implemented. It uses no customized layers and the implementation should be straight forward. The creation of anchors as well as several projections methods can be reused from the benchmark model. This includes the projection of 3d bounding boxes to the image plane.

New components will be the bottlneck and the crop-and-resize modules. They focus on compressing the extracted features in order to enable realtime capabilities. Since this is not the goal of this project, changing the compression rate is a possiblity to increase the average precision of the model. The representation of bounding boxes also differs from the benchmark model. This needs to be evaluated and may result in a performance increase of the benchmark model.

Another missing component is the data augmentation. Since the kitti dataset is rather small (about 12000 images for the raw dataset), overfitting may be a problem. A possible solution is the addition of noise to the rgb and the lidar data. More suffisticated approaches can be studied during the project phase.

Once the model has been probably trained, changes in the architecture can be applied. For instance, using a different feature extractor or changing the number of feature layers may yield higher average precision. Another possibility is the use of pretrained encoders.

The evaluation shall consists of at least a comparison of the benchmark model and the AVOD architecture. Additionally, different configurations of the AVOD approach that are not discussed in the original publication shall be evaluated.

References

BostonDidiTeam. 2017. “Multi-View 3D Object Detection Network for Autonomous Driving.” https://github.com/bostondiditeam/MV3D.

Chen, Xiaozhi, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. 2016. “Multi-View 3D Object Detection Network for Autonomous Driving.” *CoRR* abs/1611.07759. http://arxiv.org/abs/1611.07759.

Geiger, Andreas, Philip Lenz, and Raquel Urtasun. 2012. “Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite.” In *Conference on Computer Vision and Pattern Recognition (CVPR).*

Huang, Jonathan, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, et al. 2016. “Speed/accuracy Trade-Offs for Modern Convolutional Object Detectors.” *CoRR* abs/1611.10012. http://arxiv.org/abs/1611.10012.

Ku, J., M. Mozifian, J. Lee, A. Harakeh, and S.

Waslander. 2017. “Joint 3D Proposal Generation and Object Detection from View Aggregation.” *ArXiv E-Prints*, December.

Liu, Wei, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. 2015. “SSD: Single Shot MultiBox Detector.” *CoRR* abs/1512.02325. http://arxiv.org/abs/1512.02325.

Ren, Shaoqing, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” *CoRR* abs/1506.01497. http://arxiv.org/abs/1506.01497.

Simonyan, Karen, and Andrew Zisserman. 2014. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” *CoRR* abs/1409.1556. http://arxiv.org/abs/1409.1556.

Zhou, Yin, and Oncel Tuzel. 2017. “VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection.” *CoRR* abs/1711.06396. http://arxiv.org/abs/1711.06396.