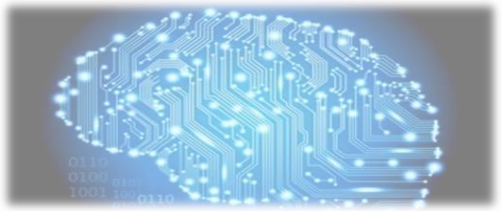
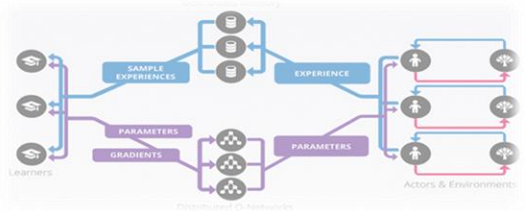
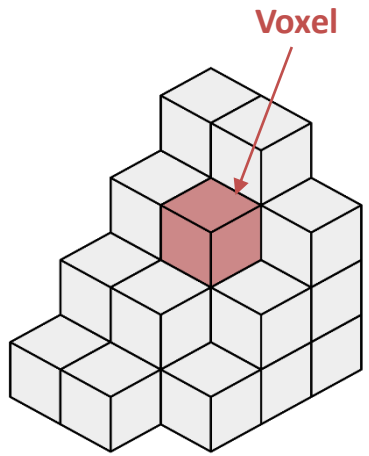



VoxelNet 구현 및 실험 결과 보고

AutoML 연구실
석사과정 문아성

21.02.23



VoxelNet: End-to-end Learning for Point Cloud Based 3D Object Detection

VoxelNet Architecture

1. Feature Learning Network
2. Convolutional Middle Layers
3. Region Proposal Network

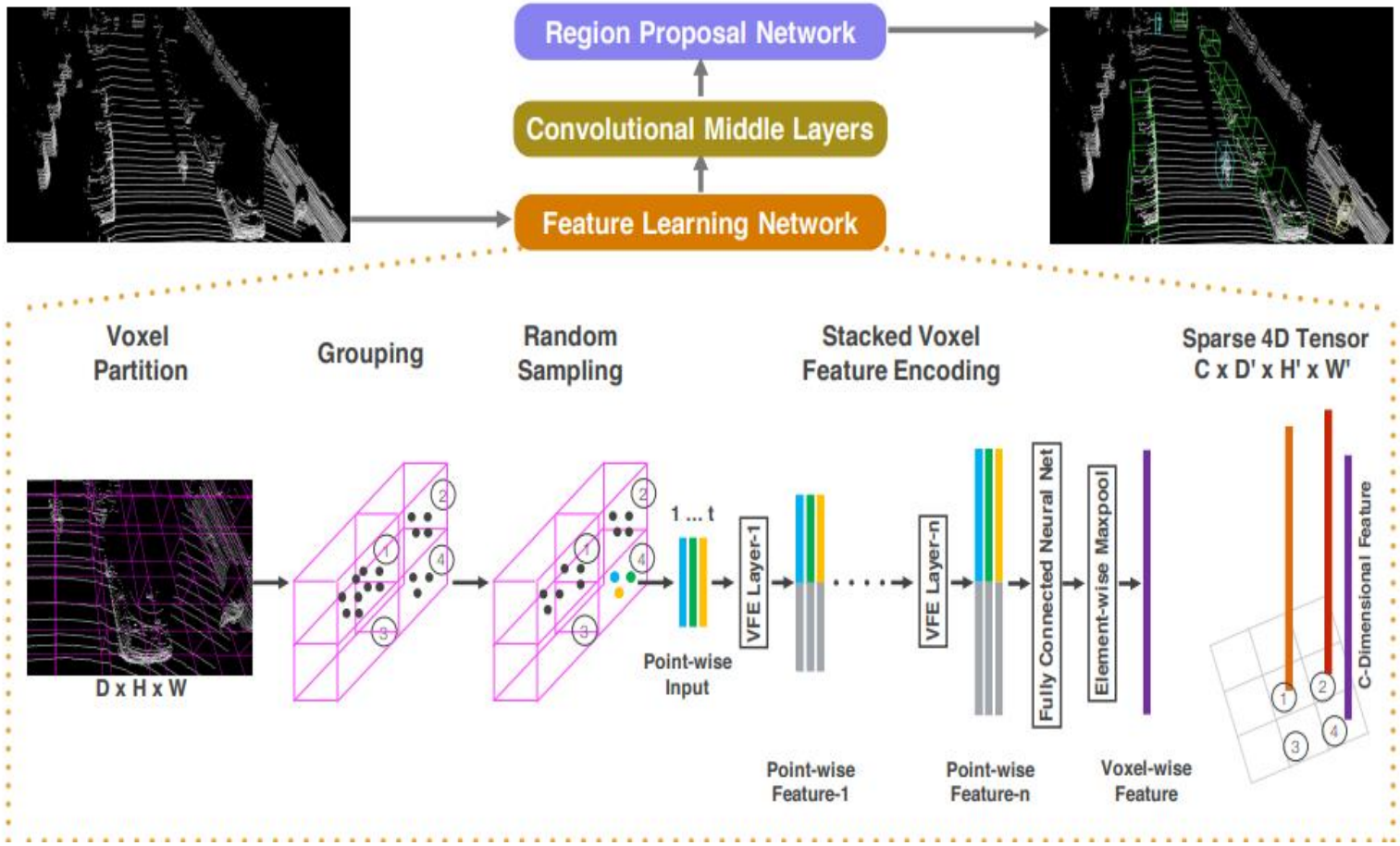


Figure 1. VoxelNet architecture. The feature learning network takes a raw point cloud as input, partitions the space into voxels, and transforms points within each voxel to a vector representation characterizing the shape information. The space is represented as a sparse 4D tensor. The convolutional middle layers processes the 4D tensor to aggregate spatial context. Finally, a RPN generates the 3D detection.

VoxelNet: Experimental Setting

Dataset

KITTI Train set 7481 images

KITTI Validation set 3769 images

Input Image

Bird's eye View Image, 3D RGB Image

Categories

Car Pedestrian Cyclist

| Easy Moderate Hard |

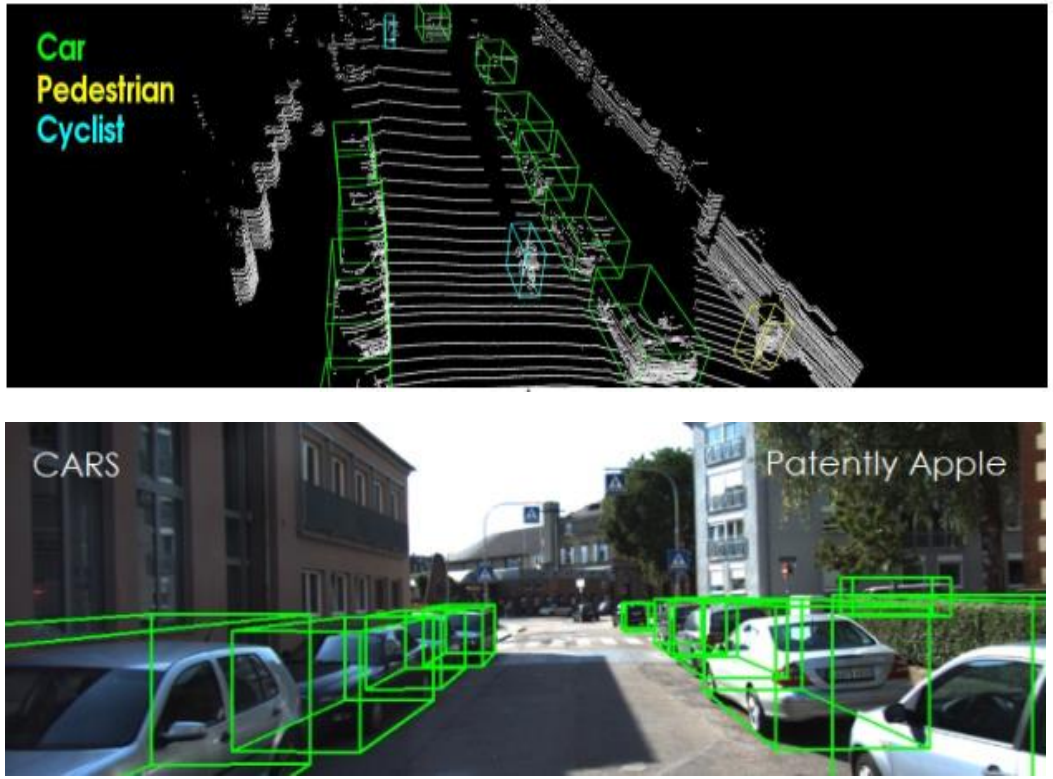


Figure 2. Top: 3D boxes detected in Bird's view image.
Bottom: 3D boxes detected in 3D RGB image.

VoxelNet: Experimental Setting

Data Augmentation

- After rotating by the angle sampled from the $[-\pi/10, \pi/10]$ uniform distribution around the center of the box, sampling from the Gaussian distribution with mean 0 and standard deviation 1 in the (x,y,z) direction.

- [0.95, 1.05] Global scaling is applied to all GT boxes and point clouds by the number of samples sampled from the uniform distribution.

- $[-\pi/4, \pi/4]$ Global rotation is applied to the GT box and the entire point cloud by the angle sampled from the uniform distribution around the origin.

VoxelNet: Experimental Setting

Efficient Implementation

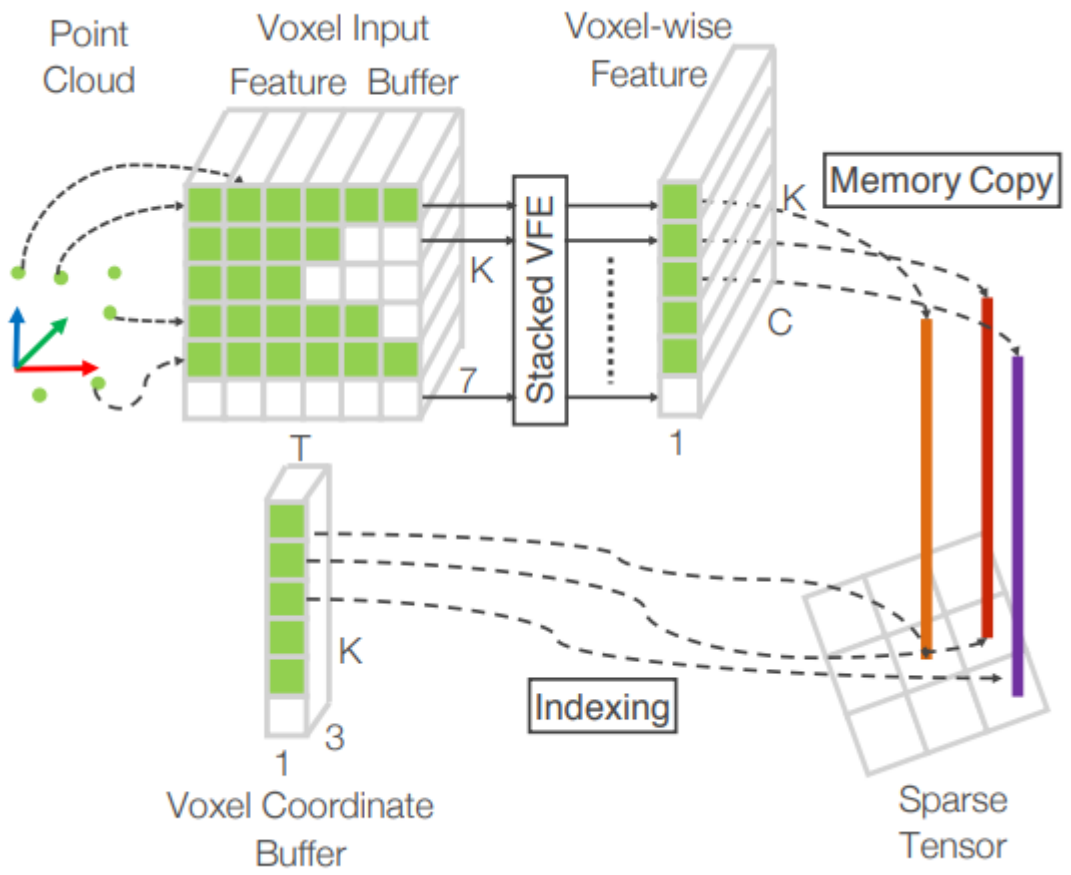


Figure 3. Illustration of efficient implementation.

Voxel Input Feature Buffer (K x T x 7)

Voxel Coordinate Buffer (K x T x 7)

* K: maximum number of non-empty voxels

  T: maximum number of points each voxel

  bounding box vector: $(x, y, z, r, x-vx, y-vy, z-vz)$

3D Object Detection Performance Table

Bird'eye view and 3d object detection performance table provided by VoxelNet (2018)

Model	Car		
	Easy	Moderate	Hard
VeloFCN	40.14	32.08	30.47
HC-baseline	88.26	78.42	77.66
VoxelNet (Reported)	89.60	84.81	78.57
VoxelNet (Me)	84.39	82.94	76.85

Table 1. Performance comparison in bird's eye view detection: average precision (in %) on KITTI validation set.

Model	Car		
	Easy	Moderate	Hard
VeloFCN	15.20	13.66	15.98
HC-baseline	71.73	59.75	55.69
VoxelNet (Reported)	81.97	65.46	62.85
VoxelNet (Me)	52.16	47.74	46.85

Table 2. Performance comparison in in 3D object detection : average precision (in %) on KITTI validation set.

Reference

Zhou, Yin, and Oncel Tuzel. "Voxelnet: End-to-end learning for point cloud based 3d object detection." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

“The KITTI Vision Benchmark Suite”, Karlsruhe Institute of Technology, last modified 2021, http://www.cvlibs.net/datasets/kitti/eval_object.php?obj_benchmark=bev