Deep Sensor Fusion for 3D Bounding Box Estimation and Recognition of Objects

Ayush Gupta (ayushg@stanford.edu)  Malavika Bindhi (mbindi@stanford.edu)

Objective

• To use Deep Learning for sensor fusion of camera and LiDAR information for 3D bounding box estimation and object recognition without geometric modelling
• Unlike existing methods that either use multistage pipelines or hold sensor and dataset-specific assumptions, PointFusion is conceptually simple and application agnostic
• Using PointNet to produce point cloud features and a standard CNN to process the corresponding image, it learns to combine and use these features to predict 3D box hypothesis and object identification
• The obtained average IOU score of 0.71 and classification accuracy of 95.62% is state-of-the-art

Dataset

KITTI 3D Object Detection Dataset
Contains recorded traffic scenarios, duly annotated, ranging from freeways, over rural areas, to inner-cities, with many static and dynamic objects
• Image Input: Left color image, Sony ICX267 CCD
• Point Cloud: LiDAR points, Velodyne HDL-64E

<table>
<thead>
<tr>
<th># examples</th>
<th>Train data</th>
<th>Dev.data</th>
<th>Test data</th>
</tr>
</thead>
<tbody>
<tr>
<td>6750</td>
<td>365</td>
<td>366</td>
<td></td>
</tr>
</tbody>
</table>

• Trained through all difficulties: easy, moderate, hard
• Classes: Car, Van, and Pedestrian (Predominant)

Preprocessing

• Filtered cloud points outside camera view angle
• Randomly sampled 2048 point cloud points
• Transformed labels to velodyne coordinates
• Applied Spatial Transformation Net. to canonicalize input space

Model

• PointFusion has three main components:
  1) A PointNet network that extracts point cloud features
  2) A CNN that extracts image appearance features
  3) A fusion network that combines both features
• The PointNet network directly consumes the point cloud, respecting the permutation invariance of points, learning embedding space

• Using Transfer Learning, we obtain image features by ResNet-50 pre-trained on ImageNet
• The fusion layer concatenates the feature vectors and applies some fully connected layers, outputting a 3D box hypothesis and classification output

Results

<table>
<thead>
<tr>
<th>Class accuracy</th>
<th>Training</th>
<th>Dev.</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.</td>
<td>96.27%</td>
<td>96.16%</td>
<td>95.62%</td>
</tr>
<tr>
<td>B. Box Average IOU</td>
<td>0.73</td>
<td>0.73</td>
<td>0.71</td>
</tr>
</tbody>
</table>

For ref., 701119/CVPR 2019: 2616143, gets a best case IOU of 0.75 on UIW-8680D dataset

Discussion & Future Scope

• Strength: Fusing data without lossy input pre-processing
• Drawback: The variance of the regression target is directly dependent on the particular scenario
• Solution: Generate box proposals by sliding windows instead of directly regressing
• Future Work: A single end-to-end 3D detector

Experimentation

• Based on empirical observations across multi-runs,
  1. Batch normalization hampers 3D bounding box estimation performance, and hence is not used
  2. SmoothL1 and mean-square error loss works well for the box-corner predictions and classification, respectively
  3. Adam optimization with a decaying learning rate is used
• Total trainable parameters: 1,808,027

References