Motivation

We perform object segmentation in the CVPR 2018 Kaggle Competition for autonomous driving using images taken from a car camera. In our approach, we have adopted the Mask R-CNN model for this task. Assisted by transfer learning via pre-trained weights from the COCO dataset, we have trained the Mask R-CNN model and successfully predicted objects in the image.

Data

- **Images (provided by Baidu, Inc.)**
  - 39,222 RGB images
  - Their corresponding labels (i.e. 39,222 masks)
  - 7 target classes, 6 of which overlap with COCO dataset classes

- **Preprocessing**
  - Downsampled from 3384x2710 to 1024x1024 (same aspect ratio)
  - Mapped the object classes from the labelled images to the number of classes
  - For images without classes, an empty mask is applied

<table>
<thead>
<tr>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>99%</td>
<td>1%</td>
<td>1917 images</td>
</tr>
</tbody>
</table>

Table 1. Train, Dev, and Test Split

References


Model

- Mask R-CNN is the most recent addition to the class of R-CNN models, Mask R-CNN inherits Faster R-CNN (for classification and detection) and performs a pixel-wise segmentation by adding a mask proposal generation with the ROIAlign
- The backbone of the Mask R-CNN is ResNet50-FPN (Feature Pyramid Network)

![Figure 1. Mask R-CNN Overview](image)

![Figure 2. Missing “bicycle” inference](image)

Results

The following results are obtained by training the head layers

**Baseline:** results with pre-trained coco weights
**Model 1:** ResNet50  8500 steps
**Model 2:** ResNet101 3000 steps

<table>
<thead>
<tr>
<th></th>
<th>Training Loss</th>
<th>Dev Loss</th>
<th>Submission Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>—</td>
<td>—</td>
<td>0.02538</td>
</tr>
<tr>
<td>Model 1</td>
<td>1.4182</td>
<td>1.4644</td>
<td>0.03607</td>
</tr>
<tr>
<td>Model 2</td>
<td>1.5459</td>
<td>1.4794</td>
<td>0.03098</td>
</tr>
<tr>
<td>Model 2</td>
<td>1.9589</td>
<td>2.1058</td>
<td>0.02763</td>
</tr>
</tbody>
</table>

Table 2. Training Loss, Validation Loss, and Kaggle Submission Score

![Figure 3. ResNet50 results on test set (top), ResNet101 results (bottom)](image)

Discussion

- Compare losses of different models
- Detection of images are successful
- Some images do not detect instances of target classes, particularly if far away (example in Fig. 2)

Future Work

- Train sequential images using LSTM layers alongside the existing Mask R-CNN model to improve the prediction performance
- Train with more images
- Use the average precision (AP) metric for comparison with other works