ANALYSIS OF THE INFLUENCE OF TRAINING DATA ON ROAD USER DETECTION

Carlos Guindel, David Martín, José María Armingol, and Christoph Stiller

20th IEEE International Conference on Vehicular Electronics and Safety
Madrid · 12 September 2018
Agenda

• Motivation and goals
• Experimental setup
• Analysis
• Conclusion
Motivation

Object detection

Computer vision

Autonomous driving

Deep Learning

Data

Instance segmentation (e.g., Mask R-CNN)
Motivation

Object detection

Computer vision

Autonomous driving

Deep Learning

Data

**IMAGENET**
450,000+ images
200 categories

**COCO**
200,000+ images
80 categories

**The KITTI Vision Benchmark Suite**
7,481 images
9 categories

**CITYSCAPES DATASET**
2,975 images
10 categories
Motivation

Object detection

Autonomous driving

Deep Learning

Motivation and goals · Experimental setup · Analysis · Conclusions

The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago

7481 images
9 categories

CITYSCAPES DATASET

2975 images
10 categories

Different labels

Data
Goals

Research is often narrow-focused on the development of new architectures and models. R-FCN, SSD; ResNet, Inception, MobileNet,…

Instead, we investigate:

- The improvement provided by introducing additional samples into the training process.
- The possibility of using heterogeneous labels in a multi-task learning method.

Faster R-CNN
State-of-the-art object detection meta-architecture
Datasets

The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago

Training

3,712 images

Validation

3,769 images

The Cityscapes Dataset

Training

2,975 images

Motivation and goals · Experimental setup · Analysis · Conclusions

Adaptation

1. Semantic labeling to bounding boxes

- Instantiable objects
- Minimum enclosing box

Categories:
- Person → Pedestrian
- Rider + Bicycle → Cyclist

The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago

CITYSCAPES DATASET
Occlusion & truncation

Occlusion
Cityscapes labels contain foreground-background ordering
Degree of occlusion = \[
\frac{\text{intersection}}{\text{background}}
\]

Truncation
Whenever any of the sides of the b. box coincides with the image boundaries
Adaptation

Motivation and goals · Experimental setup · Analysis · Conclusions

Resolution/FOV

2048 × 1024

(KITTI: 1224 × 370 aprox.)

2048 × 620

Removes the hood and the Mercedes-Benz emblem
4 Difficulty levels

We ignore samples not meeting the KITTI’s **Hard** level requirements:

- Larger than **25 pixels**
- Max **occlusion**: “Difficult to see”: level 2 (KITTI) or 75% (Cityscapes)
- Maximum **truncation**: 50% (KITTI) or no truncation (Cityscapes)
Object detection method

**Vanilla Faster R-CNN**

- **Feature extractor**
- **Feature maps**
- **RPN** generates proposals

**Network backbone**

- **VGG-16**

- **RPN**
- **Classification (FC layers)**

- **B. box regression**
- **Viewpoint**

**Modified**

- **Class**

Each proposal is classified, refined and given an estimated observation angle.

Object detection method

Motivation and goals · Experimental setup · Analysis · Conclusions


\[ r^k = (r^k_0, \ldots, r^k_{N_b}) \quad \text{for} \quad k = 0, \ldots, K \]

\[ N_b \cdot K \]

\[ N_b \] angle bins

\[ K \] classes

Multinomial logistic loss

\[ \frac{1}{N_{B_2}} \sum_{i \in B_2} [v_i \geq 1] \cdot L_{cls}(r^v_i, w_i) \]
Motivation and goals · Experimental setup · Analysis · Conclusions

## Multi-task loss


\[
\text{Loss} = \text{Objectness} + \text{Proposal regr.} + \text{Class} + \text{B.box regr.} + \text{Viewpoint}
\]

Multinomial logistic loss:

\[
\frac{1}{N_{B_2}} \sum_{i \in B_2} [v_i \geq 1] L_{cls}(r_i^{v_i}, w_i)
\]
Multi-task loss

The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago

Motivation and goals · Experimental setup · Analysis · Conclusions

Viewpoint loss ≠ 0
Multinomial logistic loss

Viewpoint loss = 0

\[ \frac{1}{N_{B_2}} \sum_{i \in B_2} [v_i \geq 1] L_{cls}(r_i^{v_i}, w_i) \]
Assessment method

Evaluation metrics

Average precision (AP)

\[ AP = \frac{1}{N} \sum_r \hat{p}(r) \]

\[ \hat{p}(r) = \max_{\tilde{r}: \tilde{r} > r} p(\tilde{r}) \]

Assess object detection

Average orientation similarity (AOS)

\[ AOS = \frac{1}{N} \sum_r \hat{s}(r) \]

\[ \hat{s}(r) = \max_{\tilde{r}: \tilde{r} > r} s(\tilde{r}) \]

Assess object detection AND orientation

Common Testbed

The KITTI Vision Benchmark Suite

A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago

Validation set

3,769 images

Training Parameters

- 1- image batch
- SGD, initial lr = 0.001
- Step decay schedule: 0.1× every 50k iterations
- 80k iterations
Experiment 1: Combined datasets

- In each training iteration, an image (single batch) is randomly chosen from a mix of both datasets.
- Tests without (a) and with (b) viewpoint estimation branch
Experiment 1: Combined datasets (a)

<table>
<thead>
<tr>
<th>category</th>
<th>tr. data</th>
<th>Easy</th>
<th>Mod.</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car</td>
<td>KITTI</td>
<td>90.05</td>
<td>79.32</td>
<td>70.04</td>
</tr>
<tr>
<td></td>
<td>Cityscapes</td>
<td>81.37</td>
<td>63.66</td>
<td>53.47</td>
</tr>
<tr>
<td></td>
<td>KITTI + CS</td>
<td><strong>90.31</strong></td>
<td><strong>84.94</strong></td>
<td><strong>70.33</strong></td>
</tr>
<tr>
<td>Pedestrian</td>
<td>KITTI</td>
<td>75.80</td>
<td>67.17</td>
<td>58.58</td>
</tr>
<tr>
<td></td>
<td>Cityscapes</td>
<td>72.00</td>
<td>63.92</td>
<td>55.33</td>
</tr>
<tr>
<td></td>
<td>KITTI + CS</td>
<td><strong>77.77</strong></td>
<td><strong>68.72</strong></td>
<td><strong>60.05</strong></td>
</tr>
<tr>
<td>Cyclist</td>
<td>KITTI</td>
<td>77.47</td>
<td>56.96</td>
<td>54.64</td>
</tr>
<tr>
<td></td>
<td>Cityscapes</td>
<td>63.09</td>
<td>50.14</td>
<td>46.85</td>
</tr>
<tr>
<td></td>
<td>KITTI + CS</td>
<td><strong>82.90</strong></td>
<td><strong>62.50</strong></td>
<td><strong>58.05</strong></td>
</tr>
</tbody>
</table>

+5.62 AP

+1.55 AP

+5.54 AP
Experiment 1: Combined datasets (b)

Two alternative strategies:

1. Pick images from a **KITTI+Cityscapes** mix
   
   Viewpoint = 0 when a Cityscapes sample is chosen

2. Pre-train with **Cityscapes**, fine-tune with **KITTI**
   
   Train without viewpoint branch and transfer weights to the complete model
## Experiment 1: Combined datasets (b)

1. Pick images from a KITTI+Cityscapes mix
2. Pre-train with Cityscapes, fine-tune with KITTI

<table>
<thead>
<tr>
<th>Category</th>
<th>tr. data</th>
<th>Detection (AP)</th>
<th>Orientation (AOS)</th>
<th>+5.6 AP</th>
<th>+1.6 AOS</th>
<th>+3.9 AP</th>
<th>+2 AOS</th>
<th>+13.6 AP</th>
<th>+12.4 AOS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Easy</td>
<td>Mod.</td>
<td>Hard</td>
<td>Easy</td>
<td>Mod.</td>
<td>Hard</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Car</td>
<td>KITTI</td>
<td>90.01</td>
<td>79.03</td>
<td>69.67</td>
<td>88.26</td>
<td>77.35</td>
<td>67.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>KITTI + CS</td>
<td><strong>90.39</strong></td>
<td>84.59</td>
<td>70.21</td>
<td><strong>88.68</strong></td>
<td>82.79</td>
<td>68.57</td>
<td><strong>+5.6 AP</strong></td>
<td><strong>+1.6 AOS</strong></td>
</tr>
<tr>
<td></td>
<td>KITTI (w. CS pret.)</td>
<td>90.33</td>
<td><strong>86.16</strong></td>
<td><strong>70.58</strong></td>
<td>88.63</td>
<td><strong>84.43</strong></td>
<td>69.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pedestrian</td>
<td>KITTI</td>
<td>71.19</td>
<td>64.05</td>
<td>55.75</td>
<td>65.31</td>
<td>57.62</td>
<td>50.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>KITTI + CS</td>
<td><strong>76.32</strong></td>
<td><strong>67.98</strong></td>
<td><strong>59.11</strong></td>
<td><strong>67.83</strong></td>
<td><strong>59.65</strong></td>
<td><strong>51.69</strong></td>
<td><strong>+3.9 AP</strong></td>
<td><strong>+2 AOS</strong></td>
</tr>
<tr>
<td></td>
<td>KITTI (w. CS pret.)</td>
<td>74.54</td>
<td>66.01</td>
<td>57.68</td>
<td>67.33</td>
<td>59.01</td>
<td>51.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyclist</td>
<td>KITTI</td>
<td>77.33</td>
<td>54.87</td>
<td>52.89</td>
<td>69.73</td>
<td>48.79</td>
<td>47.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>KITTI + CS</td>
<td><strong>86.11</strong></td>
<td><strong>68.49</strong></td>
<td><strong>63.46</strong></td>
<td><strong>77.66</strong></td>
<td><strong>61.23</strong></td>
<td><strong>56.83</strong></td>
<td><strong>+13.6 AP</strong></td>
<td><strong>+12.4 AOS</strong></td>
</tr>
<tr>
<td></td>
<td>KITTI (w. CS pret.)</td>
<td>83.18</td>
<td>60.37</td>
<td>57.35</td>
<td>75.55</td>
<td>54.36</td>
<td>51.74</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiment 2: Can we get rid of ImageNet?

Pre-training with ImageNet (generalist dataset) generates good initial weights.

<table>
<thead>
<tr>
<th>init.</th>
<th>tr. data</th>
<th>Detection (mAP)</th>
<th>Orientation (mAOS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Easy</td>
<td>Mod.</td>
</tr>
<tr>
<td>Yes</td>
<td>KITTI</td>
<td>79.51</td>
<td>65.98</td>
</tr>
<tr>
<td>No</td>
<td>K. + CS</td>
<td>53.80</td>
<td>42.99</td>
</tr>
</tbody>
</table>

-22.99 mAP  
-22.12 mAOS

Initialization with a large dataset is still an essential requirement to achieve a proper generalization ability.
## Experiment 3: Overfitting

Performance on the validation set vs # of iterations

![Graphs showing mAP and mAOS vs iterations for KITTI and KITTI + Cityscapes datasets.](image)

- **No symptoms of overfitting**
- **Dropout?** \( p = 0.5 \)

### Dropout Analysis

<table>
<thead>
<tr>
<th>dropout</th>
<th>Detection (mAP)</th>
<th>Orientation (mAOS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy</td>
<td>Mod.</td>
</tr>
<tr>
<td>No</td>
<td>79.51</td>
<td>65.98</td>
</tr>
<tr>
<td>Yes</td>
<td>79.20</td>
<td>65.34</td>
</tr>
</tbody>
</table>

- **−0.64 mAP**
- **−0.52 mAOS**

### No apparent benefit
Additional measure

Mean precision in pose estimation (MPPE)

- Discrete viewpoint estimation is a classification problem
- MPPE is the mean of the elements on the main diagonal of the confusion matrix (correctly predicted viewpoint bins)

Experiment 4: Missing labels

Motivation and goals · Experimental setup · Analysis · Conclusions

Observation angle annotations ✓ No observation angle annotations ✗

The KITTI Vision Benchmark Suite + CITYSCAPES DATASET

Detection + Orientation +6.64 mAOS Orientation +1.52 MPPE

<table>
<thead>
<tr>
<th>category</th>
<th>tr. data</th>
<th>Easy</th>
<th>Mod.</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car</td>
<td>KITTI</td>
<td>92.24</td>
<td>80.93</td>
<td>69.29</td>
</tr>
<tr>
<td></td>
<td>KITTI + CS</td>
<td>92.13</td>
<td>83.40</td>
<td>71.72</td>
</tr>
<tr>
<td>Pedestrian</td>
<td>KITTI</td>
<td>59.03</td>
<td>51.02</td>
<td>43.71</td>
</tr>
<tr>
<td></td>
<td>KITTI + CS</td>
<td>57.74</td>
<td>51.35</td>
<td>44.41</td>
</tr>
<tr>
<td>Cyclist</td>
<td>KITTI</td>
<td>70.71</td>
<td>49.84</td>
<td>49.00</td>
</tr>
<tr>
<td></td>
<td>KITTI + CS</td>
<td>64.95</td>
<td>51.60</td>
<td>49.11</td>
</tr>
</tbody>
</table>

+2.47 MPPE +0.33 MPPE +1.76 MPPE

Missing labels does not hurt orientation estimation performance

Analysis of the Influence of Training Data on Road User Detection · Carlos Guindel · ICVES 2018

24
Experiment 5: Mixed labels

Including all categories
Car, Truck, Pedestrian, Cyclist, Train, Traffic Sign

Cityscapes-only
Experiment 6: Data augmentation

Horizontal flip + texture augmentations

Choose a random subset of them, between 0 and 4

Add
$[-40, 40]$

Multiplication
$[0.5, 1.5]$

Gaussian noise
$\mathcal{N}(0, 5.1^2)$

Saturation
$[-20, 20] \ (H, S)$
## Experiment 6: Data augmentation

Two separated experiments

1. **Only KITTI**
   
   To assess the overall effect of the augmentation techniques

2. **Cityscapes + KITTI**
   
   Augmentation could help mitigate the difference between both sets of images

<table>
<thead>
<tr>
<th>tr. data</th>
<th>aug.</th>
<th>Detection (mAP)</th>
<th>Orientation (mAOS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Easy</td>
<td>Mod.</td>
</tr>
<tr>
<td>K.</td>
<td>No</td>
<td>79.51</td>
<td>65.98</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td><strong>80.39</strong></td>
<td>65.87</td>
</tr>
<tr>
<td>K. + CS</td>
<td>No</td>
<td>84.27</td>
<td>73.69</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>83.96</td>
<td><strong>74.14</strong></td>
</tr>
</tbody>
</table>

- **No apparent benefit**
- **Limited benefit**
Comparison

Analysis of the Influence of Training Data on Road User Detection
C. Guindel, D. Martín, J. M. Armingol and C. Stiller

Motivation and goals · Experimental setup · Analysis · Conclusions
Conclusion

Modestly enhancing the training data can lead to notable improvements on the results obtained by a CNN object detector.

The variability introduced by Cityscapes samples can achieve a non-negligible improvement, even when evaluated on the KITTI dataset.

Results pave the way for future works taking advantage of multiple data sources.
THANK YOU