MultiNet: Real-time Joint Semantic Reasoning for Autonomous Driving

Marvin Teichmann, Michael Weber, Marius Zöllner, Roberto Cipolla and Raquel Urtasun

Presented by: Matt Angus

May 10, 2017
Overview

1. New Layers
2. MultiNet Tasks
3. Network Architecture
4. Loss Functions
5. Training
6. Results
7. Limitations
8. Summary
New Layers

Convolution

https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/
New Layers

Transposed Convolution

https://github.com/vdumoulin/conv_arithmetic
ROI-Pooling

Rezoom

1. Apply ROI-Pooling to higher dimension features (e.g. $156 \times 48$)
2. Concatenate with extracted features (e.g. $39 \times 12$)
3. Apply $1 \times 1$ convolution
MultiNet Tasks

- Classification
  - Highway/Minor Road
- Detection
  - Bounding box
- Segmentation
  - Free space
CNN Encoder (VGG16)

https://www.cs.toronto.edu/~frossard/post/vgg16/
Detection Decoder

VGG16 → Encoded Features 39x12x512 → Conv 1 x 1 → Hidden 39 x 12 x 500 → Conv: 1 x 1 → Prediction 39x12x6

Rezoom Layers 39 x 12 x 1524 → Conv: 1 x 1 → Delta Prediction 1248x384x2
Loss Functions

\( p \): prediction
\( q \): ground truth
\( I \): examples in minibatch
\( C \): classes
\( \delta \): 1 if cell has positive confidence, 0 otherwise

**Classification/Segmentation**

\[
\text{loss}_{\text{class}}(p, q) := -\frac{1}{|I|} \sum_{i \in I} \sum_{c \in C} q_i(c) \log(p_i(c))
\]

**Bounding Box**

\[
\text{loss}_{\text{box}}(p, q) := \frac{1}{|I|} \sum_{i \in I} \delta_{q_i} \cdot (|x_{p_i} - x_{q_i}| + |y_{p_i} - y_{q_i}| + |h_{p_i} - h_{q_i}| + |w_{p_i} - w_{q_i}|)
\]
Training Strategy

- merge gradients computed by each loss function, with equal weight
- detection network requires more iterations
  - two updates with just detection, then one update with all
- 0.5 dropout probability on inner $1 \times 1$ conv.
Validation Scores

MaxF1 = \max_{\tau} \left( 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \right)
Results

Joint Training

<table>
<thead>
<tr>
<th>Task: Metric</th>
<th>separate</th>
<th>2 losses</th>
<th>3 losses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segmentation: MaxF1</td>
<td>95.83%</td>
<td>94.98%</td>
<td>95.13%</td>
</tr>
<tr>
<td>Detection: Moderate</td>
<td>83.35%</td>
<td>83.91%</td>
<td>84.39%</td>
</tr>
<tr>
<td>Classification: Accuracy</td>
<td>92.65%</td>
<td>—</td>
<td>94.38%</td>
</tr>
</tbody>
</table>
## Results

### Runtime

<table>
<thead>
<tr>
<th>MultiNet</th>
<th>Segmentation</th>
<th>Detection</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>98.10 ms</td>
<td>94.6 ms</td>
<td>37.5 ms</td>
<td>35.94 ms</td>
</tr>
<tr>
<td>10.2 Hz</td>
<td>10.6 Hz</td>
<td>27.7 Hz</td>
<td>27.8 Hz</td>
</tr>
</tbody>
</table>
Limitations

- Intentionally overfit / Bad generalization
- Not much information gain from classification
- Only detects cars
- Not so good with dark shadows
Limitations

CamVid
Limitations

Cityscapes
Limitations

Shadows
Summary of Contributions

- Rezoom layer: increase performance, little cost
- Joint training: slight increase performance, slower convergence
- Runtime: 10 fps for all tasks
Questions?