Detective: An Attentive Recurrent Model for Sparse Object Detection

Amine Kechaou, Manuel Martinez, Monica Haurilet, Rainer Stiefelhagen
Motivation

Example of a dense object detector: YOLO v1 (Redmon et al. 2016)

Detective: an attentive sparse object detector
Advantages

- Raw detections are final detections
  - No post-processing required

- Objects are predicted one by one
  - Enable better understanding at the instance level

- Predictions are not made densely over the image
  - No more foreground / background class imbalance

- Possibility for Ordered Object Detection
  - Objects are detected in the right order in one pass
Challenges

- Object detection is framed as an unordered problem
  - Target objects’ set is unordered
  - How to assign target objects to predictions during training?

- Sparse localization is challenging
  - Limited related work

- RNNs have been rarely used for the combination of localization and classification
  - Balancing localization and classification is challenging
Decoder Architecture

- Feature maps: \( [h \times w \times c] \)
- Positional embeddings: \( [h \times w \times 2] \)
- Conv LSTM
- Attention: \([d]\)
- cls head
  - Probability scores over object classes, background class and the EOS token
- loc head
  - Bounding box offsets

- ResNet
- Conv LSTM
- Conv LSTM
- Conv LSTM
- Conv LSTM

- Probability scores: \( p_{cls} \) \( [\mid C \mid+2] \)
- Bounding box offsets: \( p_{loc} \) \( [4] \)

- Feature maps:
  - \( I \) \([h \times w \times c+2]\)
  - \( H \times C \) \([h \times w \times d]\)

- Attention:
  - \( H \times C \) \([h \times w \times d]\)
  - \( I \) \([h \times w \times c+2]\)

- Positional embeddings:
  - \([h \times w \times 2]\)

- Conv LSTM:
  - \( H \) \([h \times w \times d]\)

- 1x1 conv
  - ReLU
  - softmax
  - \([h \times w \times 1]\)
Dynamic Matching

- How to assign target objects to predictions during training?
- Dynamic matching: match target objects \((T)\) to closest predictions \((P)\)
  - Cost function: \(f(t, p) = \mu_{cls} \mathcal{L}_{cls}(t_{cls}, p_{cls}) + \mu_{loc} \mathcal{L}_{loc}(t_{loc}, p_{loc})\)
  - Find the matching function \(g: T \rightarrow P\) such that \(\sum_{t \in T} f(t, g(t))\) is minimal
- Hungarian algorithm efficiently finds the optimal solution
Loss Function

- Targets / predictions mapping: \( M = \{(t, p) \in T \times P \mid p = g(t) \forall t \in T\} \)

- Loss: \( L = \lambda_{cls}L_{cls} + \lambda_{loc}L_{loc} \) where:

\[
L_{cls} = \sum_{(t,p) \in M_{FG}} \mathcal{L}_{cls} (t_{cls}, p_{cls}) + \sum_{(t,p) \in M_{BG}} \mathcal{L}_{cls} (BG, p_{cls}) + \mathcal{L}_{cls} \left( EoS, p_{cls}^{(last)} \right)
\]

and

\[
L_{loc} = \sum_{(t,p) \in M} \mathcal{L}_{loc} (t_{loc}, p_{loc})
\]
Evaluation
Ablation Study

- Dataset: PASCAL VOC
  - Training set: trainval07+12
  - Test set: test07
- Metric: mean Average Precision (mAP)

### TABLE I
RESULTS OF DETECTIVE UNDER DIFFERENT DECODER SETTINGS

<table>
<thead>
<tr>
<th>Model</th>
<th>Conv-LSTM</th>
<th>Attention</th>
<th>Positional embeddings</th>
<th>Background classification</th>
<th>mAP (%) on VOC07</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>40.7</td>
</tr>
<tr>
<td></td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>49.4</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>46.5</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>47.9</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>49.9</td>
</tr>
<tr>
<td>Detective</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>52.0</td>
</tr>
</tbody>
</table>
Effects of Attention
Conclusion

- Detective: sequential detection of object instances in an image
  - Architecture: ConvLSTM + Attention
  - Post-processing (e.g., NMS) **not required**

- Evaluation
  - Promising results on PASCAL VOC
  - ConvLSTM + Attention add spatial awareness
  - Ability to reason at the **instance** level