Introduction

- Fully-supervised object detection requires instance-level annotations, which are labor-expensive
- Weakly-supervised object detection (WSOD) still requires an unnatural, crowdsourced environment
  - It requires only image-level annotations, which alleviates the burden to a certain extent
  - Its use of Multiple Instance Learning (MIL) requires precise labels, but in the wild, some objects in the image may not be mentioned
- Our proposed method utilizes free-form captions; these pose a challenge:
  - Contributions
    - New task: Learning from noisy caption annotations
    - Benchmark and baseline: We show that predicting what truly is in an image (by training a robust text classifier) is a good way to mediate the reporting bias [Misra 2016], as compared to text matching

Method

- Use pseudo labels extracted from the free-form text as supervision
- Label inference module
  - It amplifies the supervision signal that captions provide, and squeezes more accurate information out of them
  - It performs basic reasoning based on the textual context
- Multiple instance detection module
  - It predicts detection / classification scores based on proposal features:
    - Detection score – weight of the i-th proposal for predicting class c
    - Classification score – probability that the proposal belongs to class c
  - It aggregates image-level prediction using an attention mechanism
  - Attention: focus more on the regions with high detection scores
- Online refinement module [Tang 2017]
  - Iterative refining – previous instance predictions are used as ground-truth to supervise learning in the next iteration

Experiments

- Benchmark
  - Training on COCO (118,287 images, 591,435 captions)
  - Training on Flickr30K (31,783 images, 158,915 captions)
  - Evaluate on Pascal VOC and COCO (mAP@0.5)
- Baselines
  - GT-LABELS (upper bound): Using ground-truth labels
  - EXACTMATCH: Lexical matching method
  - EXTENDVOCAB: Using a manually constructed, hence expensive COCO synonym mapping dictionary
  - GLOVEPSEUDO: Assigning pseudo-labels based on word embedding distance
  - LEARNEDGLOVE: Same as the previous one, but we learn the word embedding based on an image-text ranking loss
  - TEXTCLSF: Using the label inference module trained on COCO

<table>
<thead>
<tr>
<th># Training Examples</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>MAP@0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>20K</td>
<td>P 89.4%</td>
<td>R 62.3%</td>
<td>42.5</td>
</tr>
<tr>
<td>40K</td>
<td>P 81.1%</td>
<td>R 60.6%</td>
<td>40.5</td>
</tr>
<tr>
<td>100K</td>
<td>P 84.7%</td>
<td>R 28.9%</td>
<td>45.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th># Training Examples</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>MAP@0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>200K</td>
<td>P 81.1%</td>
<td>R 60.6%</td>
<td>45.3</td>
</tr>
<tr>
<td>60K</td>
<td>P 81.1%</td>
<td>R 60.6%</td>
<td>45.3</td>
</tr>
<tr>
<td>200K</td>
<td>P 81.1%</td>
<td>R 60.6%</td>
<td>45.3</td>
</tr>
</tbody>
</table>