**Goal**  Produce visual groundings of linguistic phrases at all levels (words, phrases, sentences), by training with weak supervision (img-cap pairs).

**Network architecture**

(1) visual encoder (2) language encoder (3) semantic embedding module (4) loss functions

**Loss functions**

\[ L = L_{PC} + L_{SIB} + L_{disc} \]

**Transferring linguistic structures**

- Sibling exclusivity constraint: Masks for nodes that are siblings should be exclusive:

\[ L_{SIB} = \frac{1}{|S|} \sum_{s \in S} \sum_{p \in \mathcal{P}} \max_{A_{i,n}} \left( \sum_{m=n}^{n} A_{i,m} \right) \]

- Parent-child inclusivity constraint: Mask of parent node should be union of children's masks:

\[ L_{PC} = \frac{1}{|P|} \sum_{k \in P} \left| A_k - \max_{l \in \text{child}(k)} A_l \right|^2 \]

**Results**

- Train models on MS COCO; test on both MS COCO and Visual Genome

  **Quantitative results on Visual Genome**

<table>
<thead>
<tr>
<th></th>
<th>Random</th>
<th>Disc-only</th>
<th>Token</th>
<th>PC</th>
<th>SIB</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.115</td>
<td>0.230</td>
<td>0.222</td>
<td>0.236</td>
<td>0.231</td>
<td>0.244</td>
</tr>
</tbody>
</table>

  **Qualitative results on MS COCO**

  Our results (left) are more accurate (bottle, dog) and cleaner (cat, boat)

  **Quantitative results (Seg mAP) on MS COCO**

<table>
<thead>
<tr>
<th></th>
<th>Disc-only</th>
<th>0.3</th>
<th>Disc-only</th>
<th>0.4</th>
<th>Disc-only</th>
<th>0.5</th>
<th>Avg mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOU@0.3</td>
<td>0.322</td>
<td>0.199</td>
<td>0.110</td>
<td>0.203</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IOU@0.4</td>
<td>0.327</td>
<td>0.213</td>
<td>0.118</td>
<td>0.219</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IOU@0.5</td>
<td>0.316</td>
<td>0.203</td>
<td>0.134</td>
<td>0.211</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Token</td>
<td>0.334</td>
<td>0.240</td>
<td>0.138</td>
<td>0.238</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>0.347</td>
<td>0.246</td>
<td>0.159</td>
<td>0.251</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Previous works**

- Treat the caption as a sequence of word tokens [Xu 2014, Rohrbach 2016, ...].
- Require phrase-region pairs or pretrained detectors [Karpathy 2015, Plummer 2015, ...].

**Motivation & Key ideas**

- Hard, if not impossible, to annotate a large collection of phrase-segment pairs.
- A sentence is not simply a sequence of token. We leverage structure in natural language for weakly supervised visual grounding.

**Benefits of exploiting linguistic structures**

- Avoid grounding nonsensical tokens ("a", "the", etc.)
- Enrich training data, in a linguistically sound way (i.e., words, phrases, sentence)
- Transfer linguistic structure to visual domain