**Problem Definition:**
For a given question-image pair ✓ answer the question ✓ localize the evidence with VQA supervision alone.

**Motivation:**
✓ VQA systems have high VQA accuracy but poor grounding ability ✓ Grounded VQA systems rely on object detections o Limits scope to only known object classes o Noisy detections

**Solution:** Capsules for Weak Grounding!

**Quantitative Results:**
- Benchmarks: GQA and CLEVR-Answers
- Evaluation metrics:
  - VQA ➔ Accuracy
  - Grounding ➔ Overlap, IOU, Grounding

- Integrated into two VQA systems: MAC (Hudson et al. ICLR 2018), and SNMN (Hu et al. ECCV 2018)
- Training parameters and loss functions are same for baselines and ours.
- Capsules with query-focused soft masking ✓ improves grounding performance ✓ Comparable VQA accuracy

**Table 2:** Comparison with baselines on CLEVR-Answers val split.

<table>
<thead>
<tr>
<th>Method</th>
<th>T param</th>
<th>Acc.</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAC</td>
<td>12.72M</td>
<td>98.00</td>
<td>30.10</td>
<td>52.41</td>
<td>38.24</td>
<td>12.59</td>
<td>23.62</td>
<td>16.42</td>
</tr>
<tr>
<td>MAC-Caps</td>
<td>12.76M</td>
<td>98.02</td>
<td>48.49</td>
<td>79.75</td>
<td>60.31</td>
<td>29.03</td>
<td>47.63</td>
<td>36.87</td>
</tr>
<tr>
<td>SNMN</td>
<td>7.32M</td>
<td>96.18</td>
<td>52.87</td>
<td>67.03</td>
<td>59.12</td>
<td>37.81</td>
<td>47.50</td>
<td>42.11</td>
</tr>
<tr>
<td>SNMN-Caps</td>
<td>6.94M</td>
<td>96.66</td>
<td>73.81</td>
<td>78.13</td>
<td>75.91</td>
<td>50.58</td>
<td>51.89</td>
<td>51.18</td>
</tr>
</tbody>
</table>

**Method:**
- Use capsule representation as image features
- Textual query generator ➔ Outputs T query embeddings (reasoning op)
- Use query to select individual capsules
- Output classifier

**Take away:**
✓ Capsules perform better than CNN features to capture visual entities for weak grounding (Table 1, 2)
✓ Weakly supervised soft masking is helpful even for convolutional features

**SNMN:** most attended question words are on top of attention map at each reasoning step (bottom right).