**Motivation**

Problem Description:
- Use natural language expressions to segment an image
- Very challenging: label space is free-form natural language descriptions, instead of 20 or 80 pre-selected categories
- Application in interactive image segmentation: selecting image regions of interest by typing or speaking

Motivation:
- Existing methods encode image and sentence independently
- However, people go back-and-forth between image and sentence according to a psychology study, suggesting early fusion
- A more plausible model: sequentially pruning out irrelevant regions as reading the sentence from left to right

Our contribution:
- A novel, more human-interpretable model that captures the motivation above while achieving state-of-the-art

**Code released at** [https://github.com/chenxi116/TF-phrasecut-public](https://github.com/chenxi116/TF-phrasecut-public)

**Baseline Model**

- Slightly adapted from (Hu et al. 2016)
- Encode image with a fully convolutional network
- Encode referring expression with an LSTM

\[
\begin{align*}
\text{LSTM} : (w_t, h_{t-1}, c_{t-1}) & \rightarrow (h_t, c_t) \\
\begin{bmatrix} i \\ f \\ \sigma \\ o \\ g \end{bmatrix} & = \begin{bmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \text{tanh} \end{bmatrix} M_{\text{LSTM}} \begin{bmatrix} w_t \\ h_{t-1} \end{bmatrix} \\
\end{align*}
\]

- Features from two modalities are concatenated
- Two more convolution layers as pixel-wise binary classifier
- Sentence-to-image: Independent encoding of two modalities

**Recurrent Multimodal Interaction**

- Novel two-layer recurrent neural network architecture
- Lower level (LSTM):
  - Model the progression of semantics
  - Same LSTM as the one used in the baseline model
- Upper level (mLSTM):
  - Model the progression of segmentation beliefs
  - Input is the concatenation of image features, spatial coordinates, LSTM hidden states, and word embeddings
  - Same mLSTM cell is shared among all locations

\[
mLSTM : \begin{bmatrix} i_t \\ h_{t-1}^i, c_{t-1}^i \end{bmatrix} \rightarrow \begin{bmatrix} h_t^i \\ c_t^i \end{bmatrix}
\]

- Equivalent to a convolutional LSTM with $1 \times 1$ kernel
- Word-to-image scheme; Early fusion of expression and image

**Analysis & Conclusion**

**Performance Evaluation by Mean IOU:**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Shortest 1/4</th>
<th>Shorter 1/4</th>
<th>Longer 1/4</th>
<th>Longest 1/4</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-Ref</td>
<td>9.44%</td>
<td>12.37%</td>
<td>12.17%</td>
<td>14.81%</td>
</tr>
<tr>
<td>UNC</td>
<td>1.94%</td>
<td>3.10%</td>
<td>3.15%</td>
<td>4.19%</td>
</tr>
<tr>
<td>UNC+</td>
<td>3.84%</td>
<td>5.67%</td>
<td>12.55%</td>
<td>16.85%</td>
</tr>
<tr>
<td>RG</td>
<td>0.69%</td>
<td>0.90%</td>
<td>1.82%</td>
<td>2.10%</td>
</tr>
</tbody>
</table>

**More Robust to Longer Expressions:**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Shortest 1/4</th>
<th>Shorter 1/4</th>
<th>Longer 1/4</th>
<th>Longest 1/4</th>
</tr>
</thead>
<tbody>
<tr>
<td>D+LSTM</td>
<td>34.52%</td>
<td>45.18%</td>
<td>45.69%</td>
<td>45.57%</td>
</tr>
<tr>
<td>D+RMI+DCRF</td>
<td>29.86%</td>
<td>30.34%</td>
<td>29.86%</td>
<td>25.50%</td>
</tr>
<tr>
<td>R+LSTM+DCRF</td>
<td>28.42%</td>
<td>29.86%</td>
<td>29.01%</td>
<td>27.70%</td>
</tr>
<tr>
<td>R+RMI+DCRF</td>
<td>35.31%</td>
<td>36.14%</td>
<td>35.31%</td>
<td>35.31%</td>
</tr>
</tbody>
</table>

**Visualizing Intermediate Segmentation Beliefs:**

- We visualize and interpret the internal segmentation beliefs
- We achieve new SOTA on all large-scale benchmark datasets

**Conclusion:**

- We propose a novel two-layer recurrent neural network architecture that jointly models the progression of semantics and the progression of segmentation beliefs
- We visualize and interpret the internal segmentation beliefs