## Abstract
We investigate the problem of Language-Based Image Editing (LBIE). Given a source image and a language description, we want to generate a target image by editing the source image based on the description. We propose a generic modeling framework for two sub-tasks of LBIE: language-based image segmentation and image colorization. The framework uses recurrent attentive models to fuse image and language features. Instead of using a fixed step size, we introduce for each region of the image a termination gate to dynamically determine after each inference step whether to continue extrapolating additional information from the textual description. The effectiveness of the framework is validated on three datasets.

## Problem
### Language-based image editing:
Given a source image (a sketch, a grayscale image or a natural image), generate a target image based on natural language instructions.

### Potential applications:
- Computer-Aided Design (CAD)
- Virtual Reality (VR)

## Framework (Overview)

### Framework (Details)

#### Image encoder:
Convolutional neural networks.

#### Language encoder:
Bidirectional long short-term memory.

#### Recurrent attentive fusion module:
Attention; termination.

Use spatial attention mechanism to extract language features. Use termination gates to dynamically control whether to stop.

#### Image decoder:
Deconvolutional neural networks.

#### Loss:
Cross-entropy for segmentation; GAN + L1 for colorization.

#### Training:
The Gumbel trick.

## ReferIt

### Data:
20k photos; 130k textual descriptions; 100k objects.

### Task:
Image segmentation of the referred object based on texts.

### Metrics:
- **Precision@threshold:** % data such that IoU > threshold.
- **IoU:** IoU computed over the entire dataset.

### Results:

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision@0.5</th>
<th>Precision@0.6</th>
<th>Precision@0.7</th>
<th>Precision@0.8</th>
<th>Precision@0.9</th>
<th>IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our model</td>
<td>32.53%</td>
<td>27.9%</td>
<td>18.76%</td>
<td>12.37%</td>
<td>4.37%</td>
<td>50.09%</td>
</tr>
</tbody>
</table>

### Oxford-102 Flower

### Data:
8k images, each equipped with five textual descriptions.

### Task:
Colorize a grayscale flower image based on one of its textual descriptions.

### Metrics:
- **Consistency:** Humans rate consistency of images and captions.
- **Quality:** Humans rate the quality of images.

### Results:

<table>
<thead>
<tr>
<th>Consistency</th>
<th>Our Model</th>
<th>Baseline</th>
<th>Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.849</td>
<td>0.27</td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>

### Qualities:
- Our Model: 0.508
- Baseline: 0.404
- Truth: 0.856

## CoSaL

### Data:
50k images, each equipped with direct and relational descriptions.

### Task:
Given a black-white image and its textual description, colorize the nine shapes correspondingly.

### Results:

<table>
<thead>
<tr>
<th># direct descriptions</th>
<th># Steps</th>
<th>Attention</th>
<th>IOU</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No</td>
<td>0.2107</td>
<td>0.2499 0.3186</td>
</tr>
<tr>
<td>1</td>
<td>Yes</td>
<td>0.4030</td>
<td>0.5220 0.7097</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>0.5033</td>
<td>0.5313 0.7017</td>
</tr>
</tbody>
</table>

Average IoU over nine shapes and the background.