Language-Based Image Editing with Recurrent Attentive Models

Jianbo Chen*, Yelong Shen†, Jianfeng Gao†, Jingjing Liu†, Xiaodong Liu†

University of California, Berkeley*, Microsoft Research†

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.

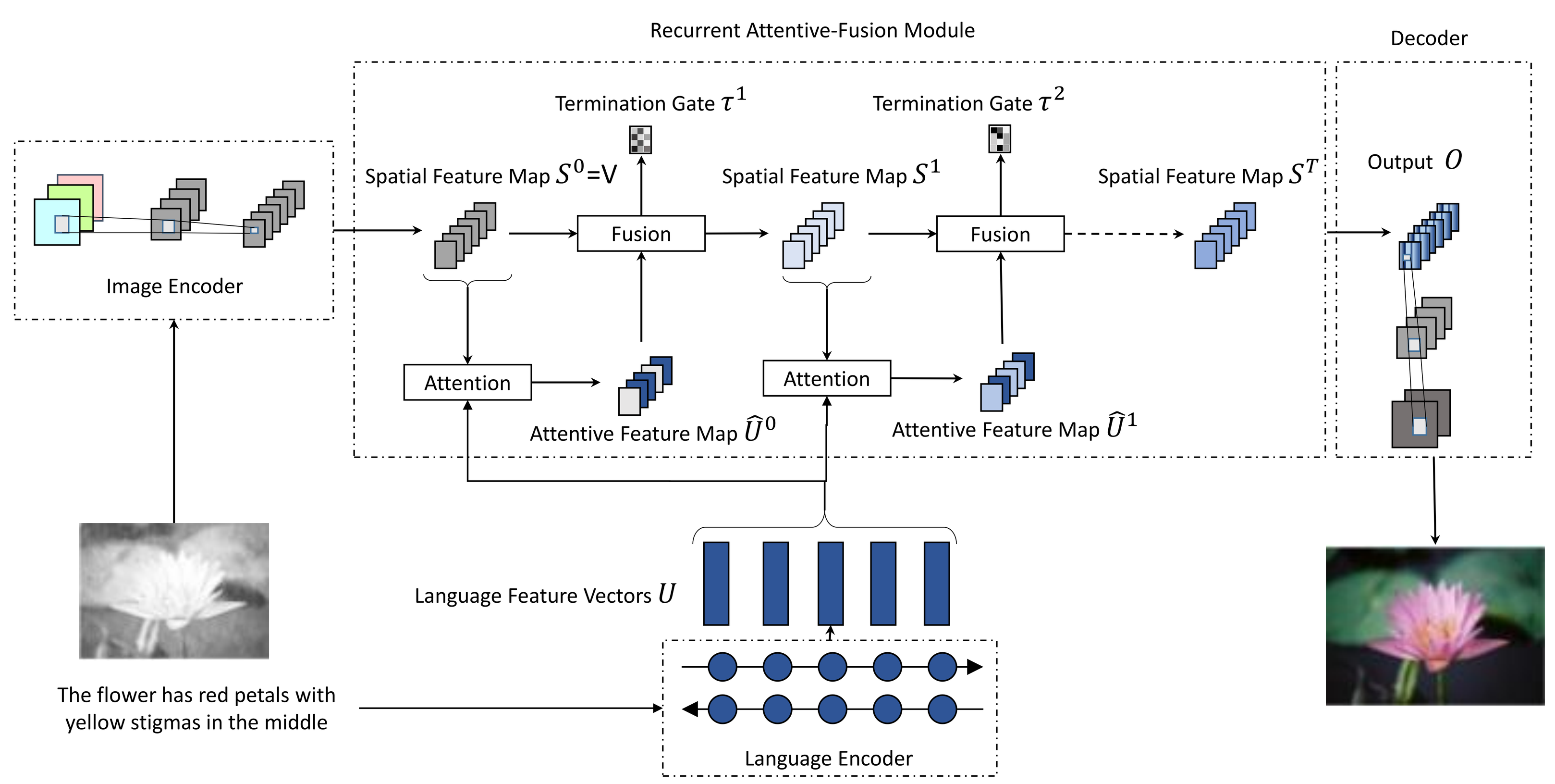
Problem

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.

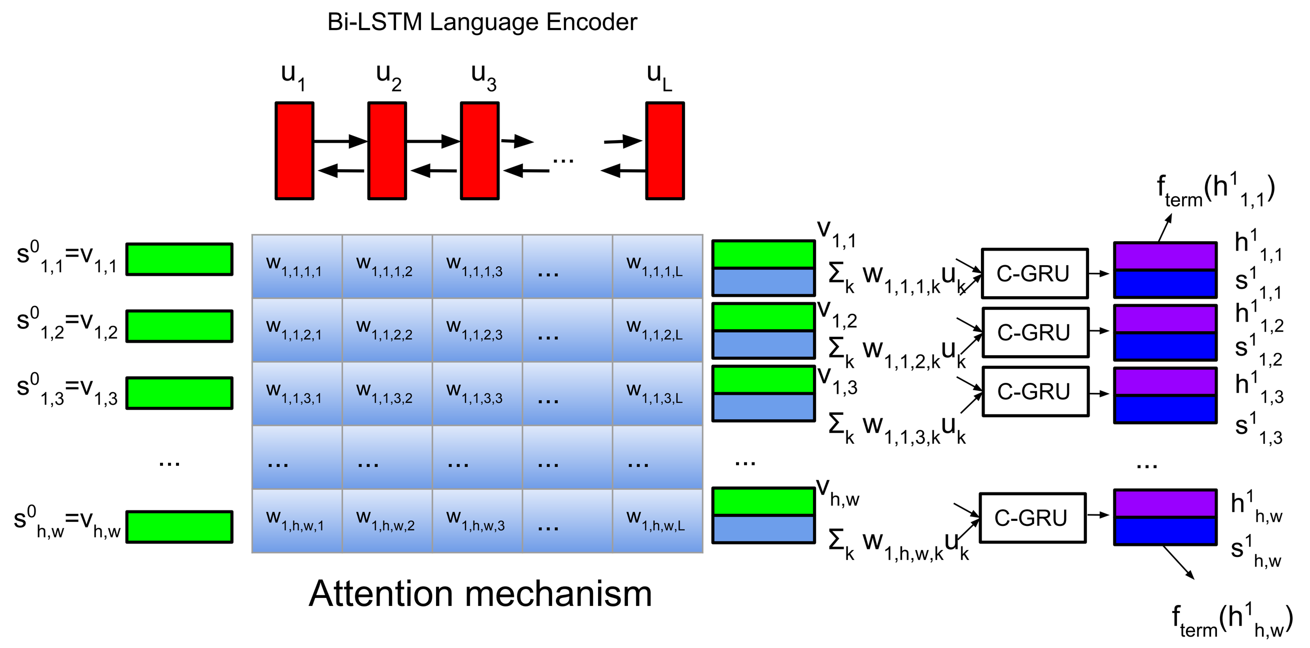
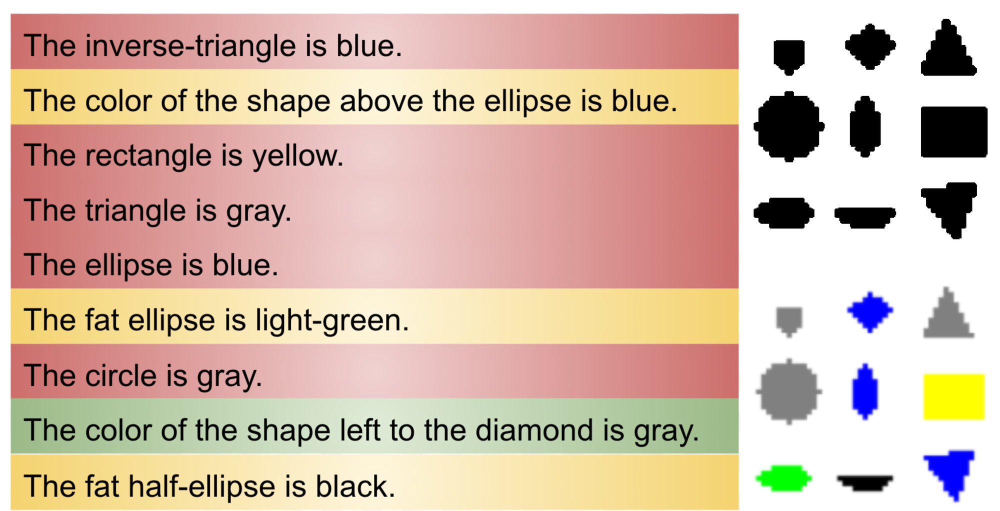


Image decoder: Deconvolutional neural networks.
Loss: Cross-entropy for segmentation; GAN + L1 for colorization.
Training: The Gumbel trick.

CoSaL

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



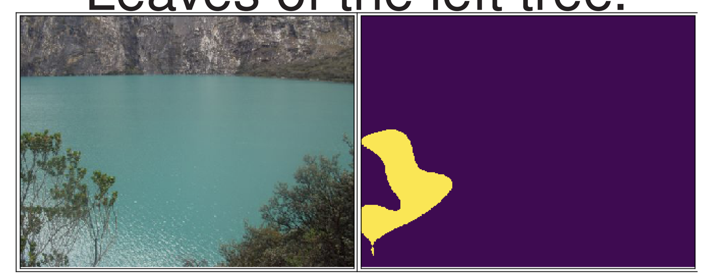
Task: Given a black-white image and its textual description, colorize the nine shapes correspondingly.
Results:

		# direct descriptions		
# Steps	Attention	4	6	8
1	No	0.2107	0.2499	0.3186
1	Yes	0.4030	0.5220	**0.7097**
4	Yes	**0.5033**	**0.5313**	0.7017

Average IoU over nine shapes and the background.

ReferIt

Leaves of the left tree.



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:

Model	Precision@0.5	Precision@0.6	Precision@0.7	Precision@0.8	Precision@0.9	IoU
SCRC bbox	9.73%	4.43%	1.51%	0.27%	0.03%	21.72%
GroundeR bbox	11.08%	6.20%	2.74%	0.78%	0.20%	20.50%
Hu, etc.	**34.02%**	26.71%	**19.32%**	11.63%	3.92%	48.03%
Our model	32.53%	**27.9%**	18.76%	**12.37%**	**4.37%**	**50.09%**

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:

	Our Model	BaseLine	Truth
Consistency	**0.849**	0.27	N/A
Quality	0.598	0.404	**0.856**

First row: Original images. Second row: Baseline. Third row: Our model.

First row: Original. Remaining rows: Results generated with arbitrary textual descriptions: *"The flower is white/red/orange/yellow/blue/purple in color"*.