Recurrent Multimodal Interaction for Referring Image Segmentation

Chenxi Liu[1] Zhe Lin[2] Xiaohui Shen[2] Jimei Yang[2] Xin Lu[2] Alan Yuille[1]
[1]Johns Hopkins University [2]Adobe Research

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

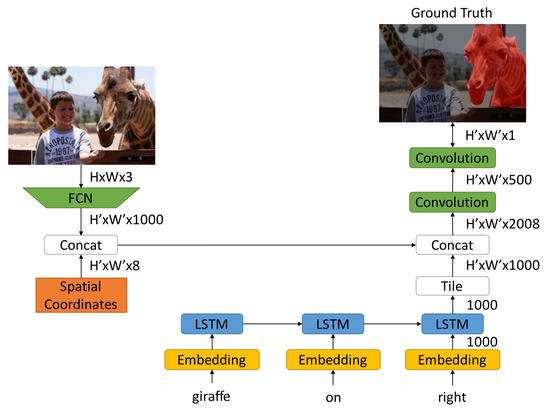
Motivation:

- 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`

Baseline Model



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

$$\text{LSTM} : (\mathbf{w}_t, \mathbf{h}_{t-1}, \mathbf{c}_{t-1}) \rightarrow (\mathbf{h}_t, \mathbf{c}_t)$$

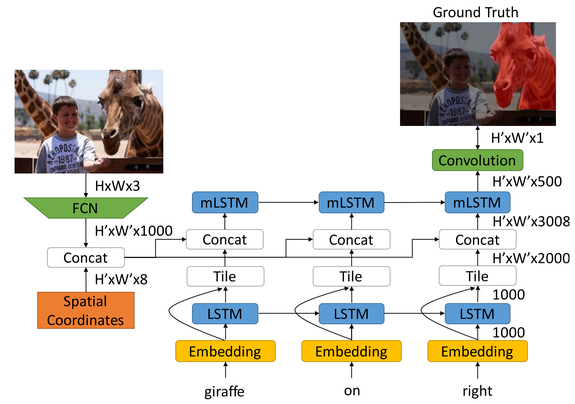
$$\begin{pmatrix} \mathbf{i} \\ \mathbf{f} \\ \mathbf{o} \\ \mathbf{g} \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} M_{4n, D_S+n} \begin{pmatrix} \mathbf{w}_t \\ \mathbf{h}_{t-1} \end{pmatrix}$$

$$\mathbf{c}_t = \mathbf{f} \odot \mathbf{c}_{t-1} + \mathbf{i} \odot \mathbf{g}$$

$$\mathbf{h}_t = \mathbf{o} \odot \tanh(\mathbf{c}_t)$$

- 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

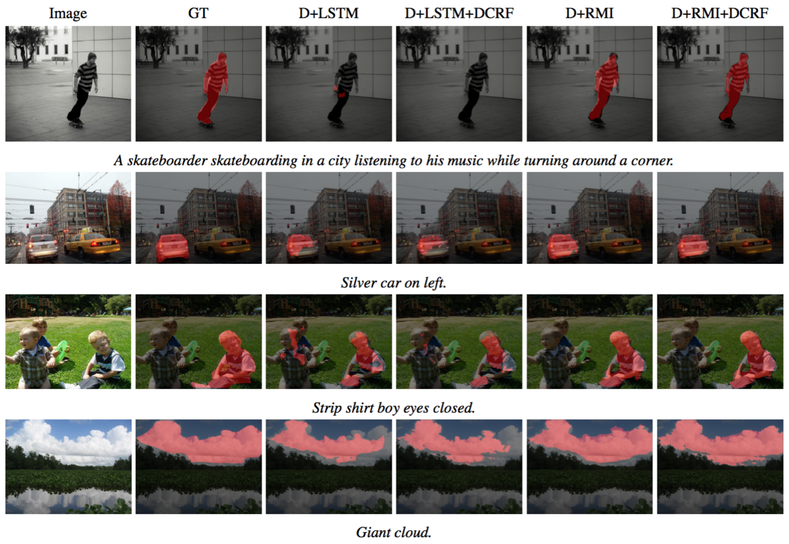


- Novel two-layer recurrent neural network architecture
- Lower level (LSTM):
  - Model the progression of semantics
  - Same LSTM as the one used in the baseline model
- Upper level (mLSTM):
  - Model the progression of segmentation beliefs
  - Input is the concatenation of image features, spatial coordinates,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

$$\text{mLSTM} : (\begin{bmatrix} \mathbf{l}_t \\ \mathbf{v}^{ij} \end{bmatrix}, \mathbf{h}_{t-1}^{ij}, \mathbf{c}_{t-1}^{ij}) \rightarrow (\mathbf{h}_t^{ij}, \mathbf{c}_t^{ij})$$

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

Qualitative Results



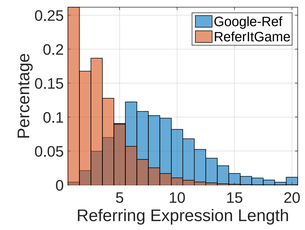
Analysis & Conclusion

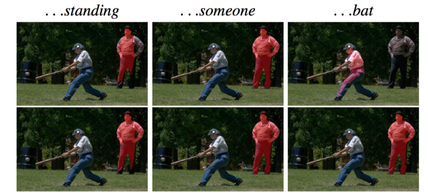
Performance Evaluation by Mean IOU:

	G-Ref	UNC			UNC+			RG
	val	val	testA	testB	val	testA	testB	test
(Hu et al. 2016)	28.14	-	-	-	-	-	-	48.03
R+LSTM	28.60	38.74	39.18	39.01	26.25	26.95	24.57	54.01
R+RMI	**32.06**	**39.74**	**39.99**	**40.44**	**27.85**	**28.69**	**26.65**	**54.55**
R+LSTM+DCRF	28.94	39.88	40.44	40.07	26.29	27.03	24.44	55.90
R+RMI+DCRF	**32.85**	**41.17**	**41.35**	**41.87**	**28.26**	**29.16**	**26.86**	**56.61**
D+LSTM	33.08	43.27	43.60	43.31	28.42	28.57	27.70	56.83
D+RMI	**34.40**	**44.33**	**44.74**	**44.63**	**29.91**	**30.37**	**29.43**	**57.34**
D+LSTM+DCRF	33.11	43.97	44.25	44.07	28.07	28.29	27.44	58.20
D+RMI+DCRF	**34.52**	**45.18**	**45.69**	**45.57**	**29.86**	**30.48**	**29.50**	**58.73**

More Robust to Longer Expressions:

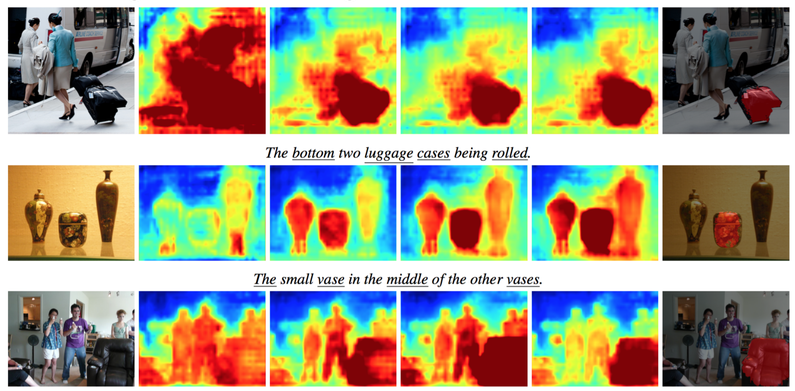
Dataset	Shortest 1/4	Shorter 1/4	Longer 1/4	Longest 1/4
G-Ref	9.44%	12.37%	12.17%	14.81%
UNC	1.94%	3.10%	3.15%	4.19%
UNC+	3.84%	5.67%	12.55%	16.85%
RG	0.69%	0.90%	1.82%	2.10%





Man in a vest and blue jeans standing watching someone swing a bat.

Visualizing Intermediate Segmentation Beliefs:



The bottom two luggage cases being rolled.

The small vase in the middle of the other vases.

An empty leather chair with a cup holder built in.

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