Segmentation from Natural Language Expressions
Ronghang Hu1, Marcus Rohrbach1,2, Trevor Darrell1
1University of California, Berkeley 2ICSI, Berkeley

Overview

Previous work

This work

image segmentation from natural language expressions

Grounding visual entities described by referential expressions via dense pixel-wise segmentation

- input: image, referential expression
- output: segmentation mask for the expression

Project page: http://ronghanghu.com/text_object

Our Model

"people on the right side"

"a dark horse below a woman in a striped shirt"

Main components:
- Embed the image: spatial feature map through CNN
- Embed the expression: final hidden state in LSTM
- Fully convolutional classification: match input expression to every location on the spatial grid and up sample

End-to-end trainable with back-propagation.

Detailed model structure

Dataset:
- ReferItGame [4] - pixel-wise region annotation for referential expressions
- Baseline approaches as comparison:
  - Combine per-word segmentation results (bag-of-words)
  - Foreground segmentation from bounding box prediction (e.g. [1,3])
  - Classification over segmentation proposals (e.g. [5])

Evaluation metric: precision and overall IoU

Experiments

Visualized Results

- Visualized comparison with baseline approaches
- Segmentation examples on object regions
- Segmentation examples on stuff regions
- Speed: over 3 fps using GPU

References