Segmentation from Natural Language Expressions

Ronghang Hu[1], Marcus Rohrbach[1,2], Trevor Darrell[1]

[1]UC Berkeley [2]ICSI, Berkeley

Traditional Category-based Segmentation

Semantic segmentation for class ***horse***

Instance segmentation for class ***horse***

Image Segmentation from Referential Expressions

Localize the following entity with bounding box:

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

bounding box localization from language (Mao et al. 2016; Hu et al. 2016)

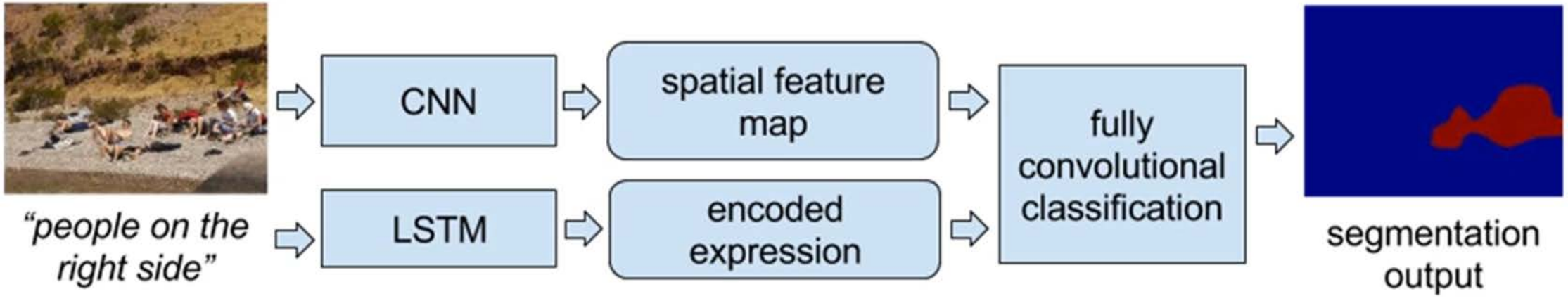
pixel-wise segmentation for expression:

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

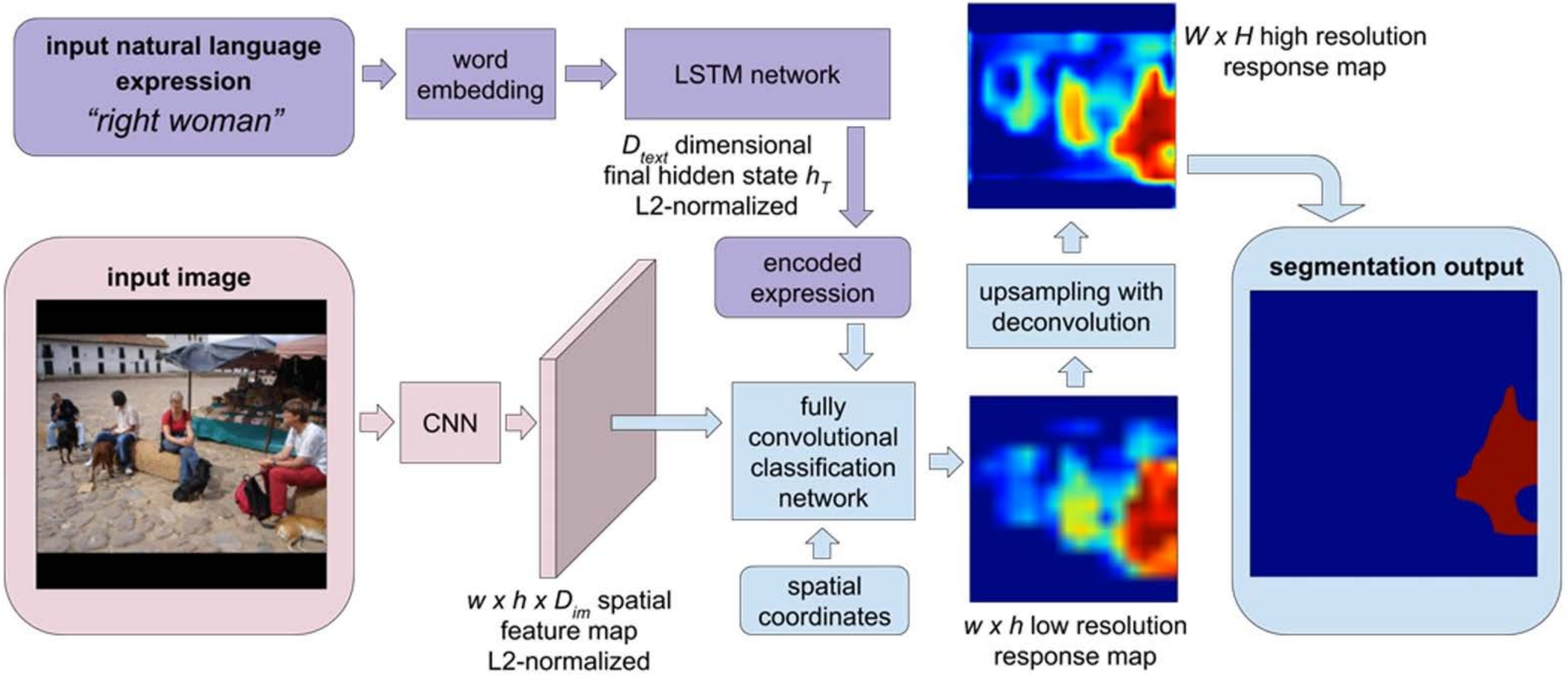
Our work: **segmentation** from natural language expressions

Our Model – Intuition

- **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



Our Model – Details



Experiments

Kazemzadeh et al. 2014

Dataset:

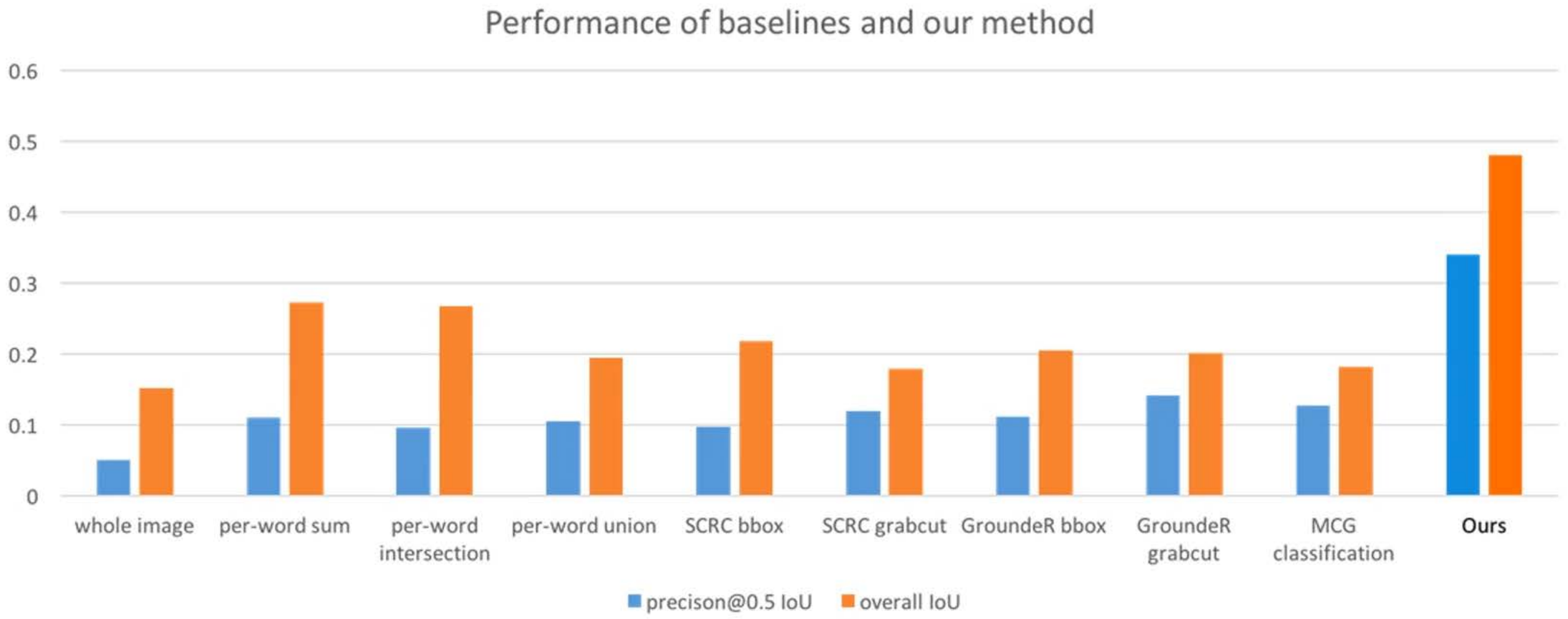
- ReferItGame (Kazemzadeh et al. 2014)
  - Pixel-wise annotation for referential expressions

Baseline approaches as comparison:

1. Combine per-word segmentation results (bag-of-words)
2. Foreground segmentation over bounding box localization methods (e.g. GroundeR)
3. Classification over segmentation proposals (e.g. MCG)

precison@0.5 IoU: percentage of test samples with at least 0.5 IoU
overall IoU: intersection-area-sum / union-area-sum on dataset

Results

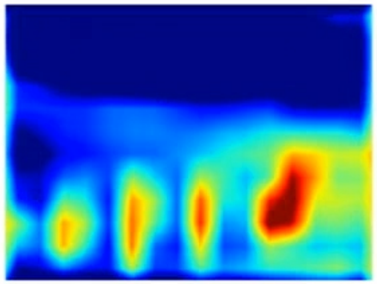


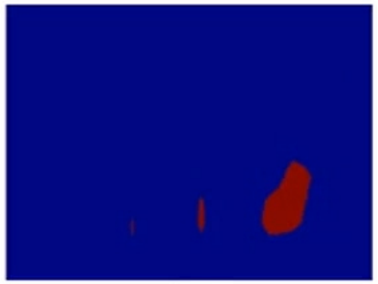
Results

Project page and code release:
http://ronghanghu.com/text_objseg

query='*3 people on right*'

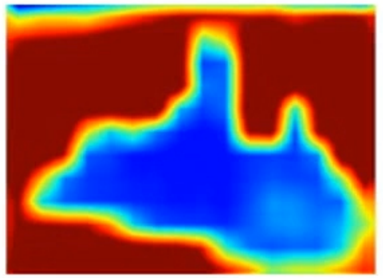

input image

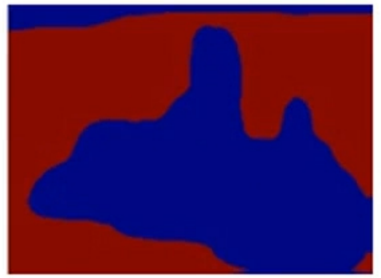

score map


prediction

query='*water*'

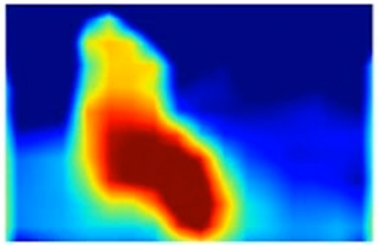

input image

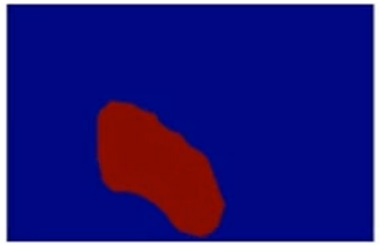

score map


prediction

query='*bike*'

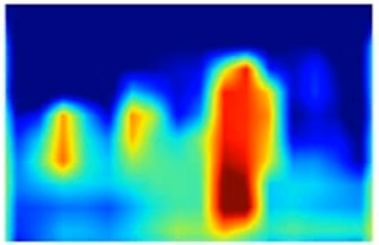

input image


score map


prediction

query='*guy in front*'


input image


score map

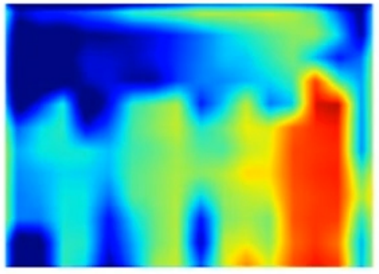

prediction

Results

Project page and code release:
http://ronghanghu.com/text_objseg

query='*man far right*'

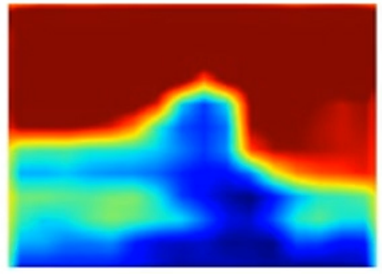

input image

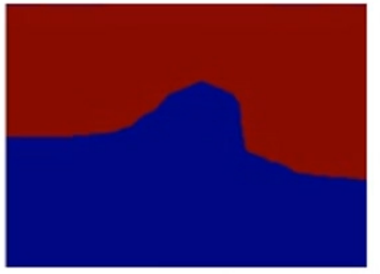

score map


prediction

query='*sky above the bridge*'

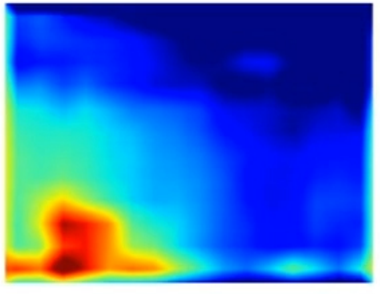

input image


score map


prediction

query='*big black suitcase bottom left*'

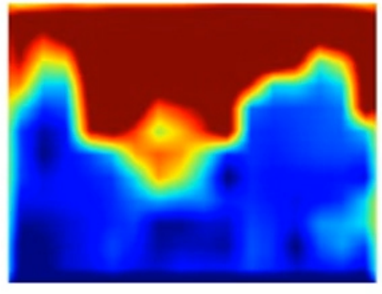

input image


score map


prediction

query='*wall above the people*'


input image


score map


prediction