CS6501/4501: Vision and Language

Referring Expressions

Last Class

- Overview on
  - Multilingual Image Captioning
  - Multimodal Machine Translation

Today

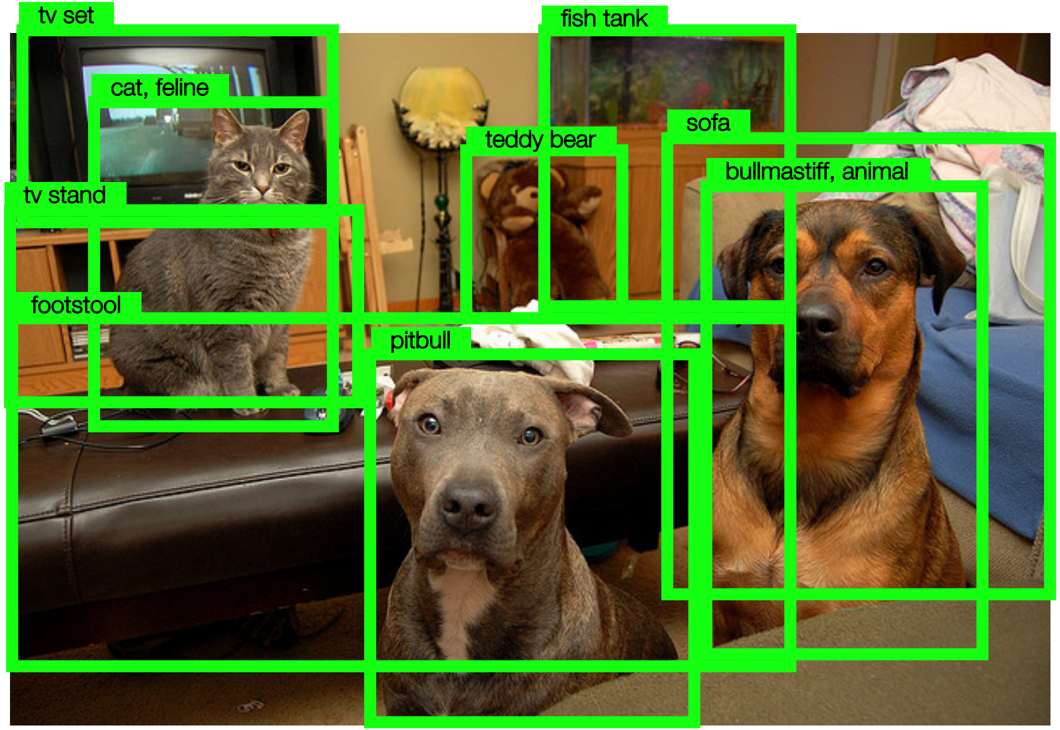
- Referring Expressions
  - Referring Expressions vs Image Captions
  - Generating Referring Expressions
  - Referring Expression Comprehension

Computer Vision

feline
tv set
teddy bear
pitbull
bullmastiff
cat
tv stand
group of dogs
fish tank
room
indoor
man-made
footstool
furniture

Image tagging / Image classification

Computer Vision



Object Detection

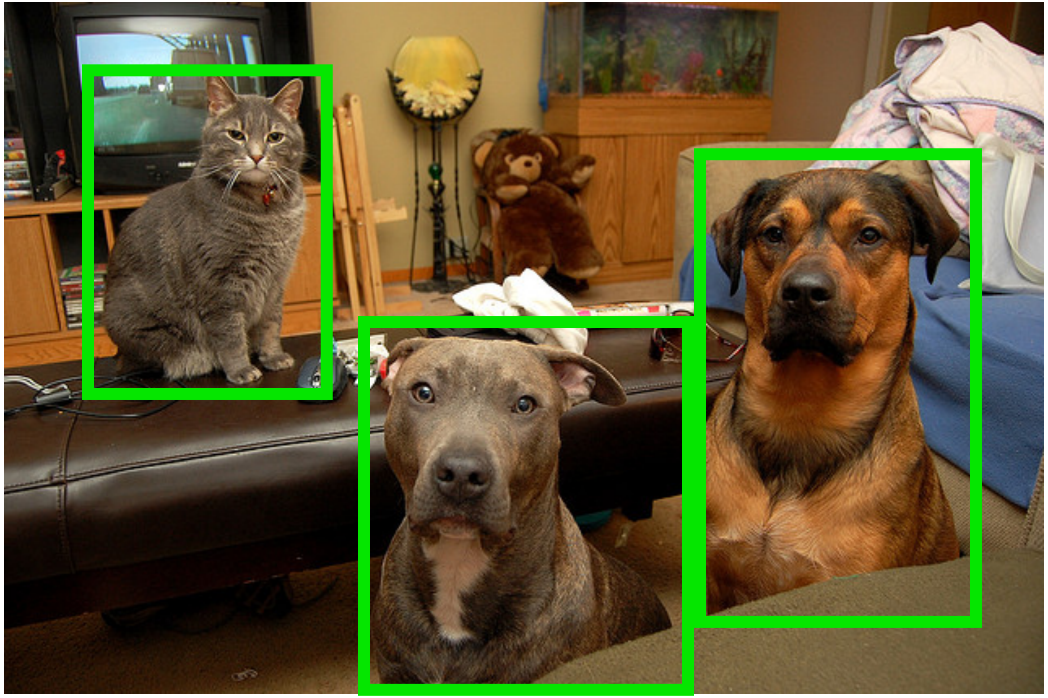
feline
tv set
teddy bear
pitbull
bullmastiff
cat
tv stand
group of dogs
fish tank
room
indoor
man-made
footstool
furniture

Computer Vision

- feline
- tv set
- teddy bear
- pitbull
- dog
- cat
- tv stand
- group of dogs
- fish tank
- room
- indoor
- man-made
- footstool
- furniture

Image Parsing / Image Segmentation

How do we describe images?



Object Importance

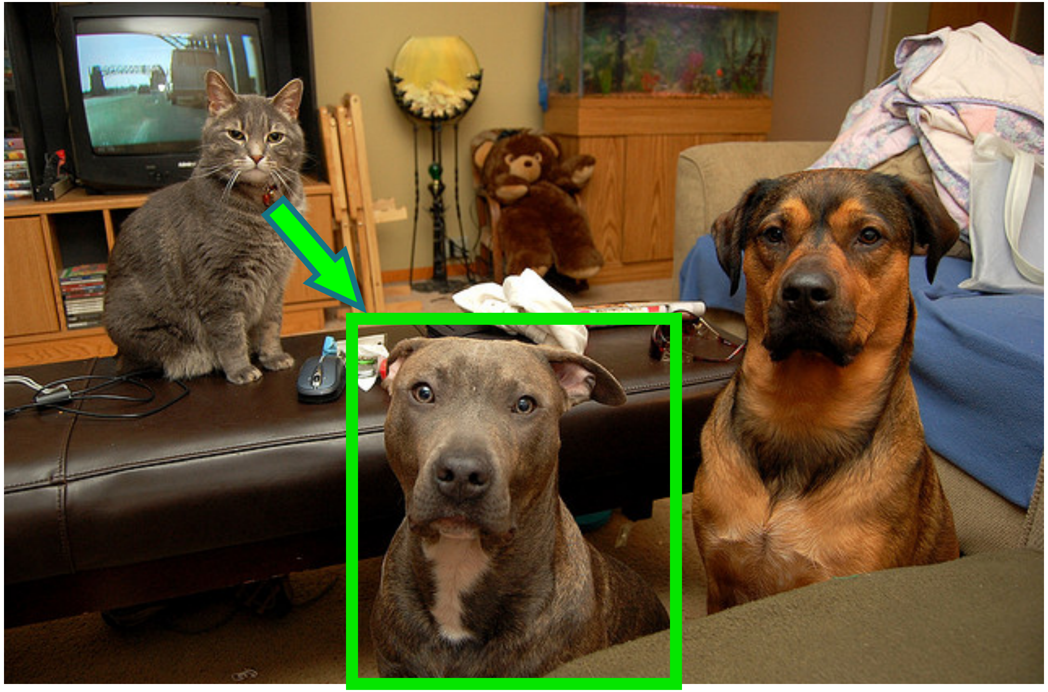
Attribute Importance

Action Importance

World knowledge

A cat and two big dogs staring at the camera

Referring to objects



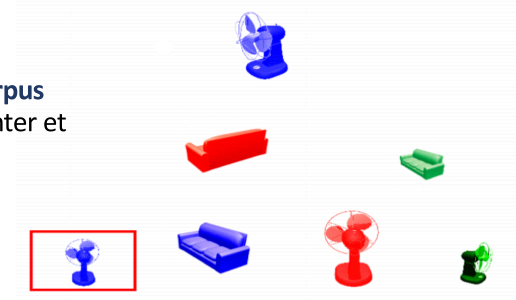
The dog in the middle

The gray dog in the middle

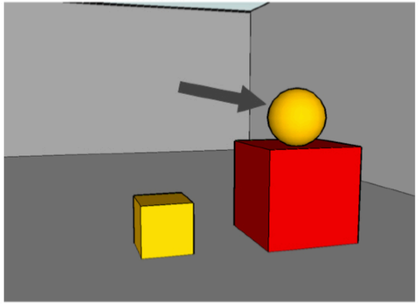
The gray dog

Work on Referring Expression

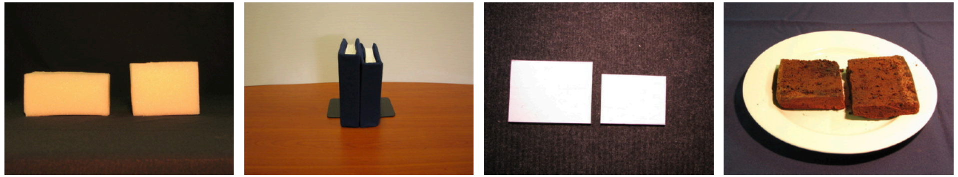
TUNA Corpus
van Deemter et al 2006



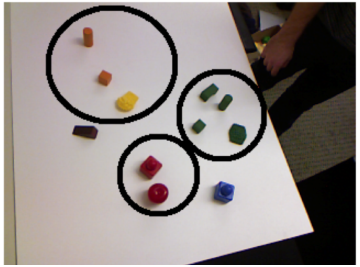
GRE3D3 Corpus
Viethen and Dale 2008
[20 scenes]



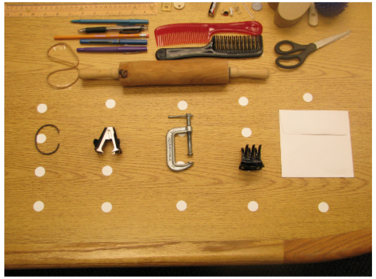
Size Corpus
Mitchell et al 2011
[96 scenes]



GenX Corpus
FitzGerald et al 2013
[269 scenes]

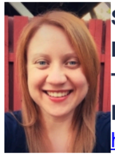


Typicality Corpus
Mitchell et al 2013
[35 scenes]



Tomorrow – Please Attend

SEMINAR ANNOUNCEMENT



Speaker: Margaret Mitchell
Date: Friday, November 6, 2020
Time: 12:00 p.m. ET
Location: Zoom meeting
https://virginia.zoom.us/j/99513114387?pwd=b1BjK3VQd0dLamw5dy9PTlJmWUcvUT09
Meeting ID: 995 1311 4387
Passcode: 966810
(**Please do not share this link on any website/forum.)*

Host: Vicente Ordonez-Roman (vo2m)

Title: *Ethics in the Vision and Language of Artificial Intelligence*

Abstract:
This talk is intended for all audiences, discussing how social inequality is propagated in machine learning systems. I will explain (some of) the role of human cognition in creating and amplifying systemic social issues in AI, the effects of Big Data on system development, and the role that ethics can play in the machine learning lifecycle.

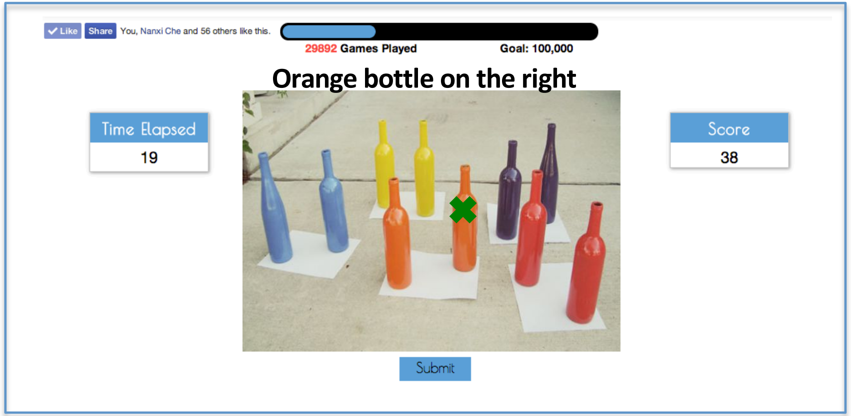
About the speaker:
Margaret Mitchell is a Staff Research Scientist at Google AI. She founded and co-leads Google's Ethical AI group, focused on foundational sociotechnical research and operationalizing AI ethics Google-internally. She has spearheaded a number of workshops and initiatives at the intersections of diversity, inclusion, computer science, and ethics. Prior to Google, Margaret was a researcher at Microsoft Research, where she focused on computer vision-to-language generation research; a postdoctoral researcher at Johns Hopkins, where she focused on Bayesian statistics and Information Extraction in text; a PhD student in Computing Science at the University of Aberdeen (Scotland), focused on generating reference to visible objects; a Master's student in Computational Linguistics at the University of Washington; and simultaneously a Scholar/Associate/etc. for 7+ years working on machine learning, neurological disorders, and assistive technology at CSLU within Oregon Health and Science University. She is both a dog person and a cat person.

ReferItGame: Referring to Objects in Photographs of Natural Scenes
Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, Tamara L. Berg.
Empirical Methods on Natural Language Processing. **EMNLP 2014**. Doha, Qatar. October 2014.

Referit Game

Player 1

Player 2



Referring Expressions for Natural Scenes

Diverse

Many real world objects

Complex

Many object instances

Big

IAPR TC-12 Segmented and Annotated Dataset. Escalante et. al. 2009

Referit Game Dataset

Blue shirt man

Blue guy

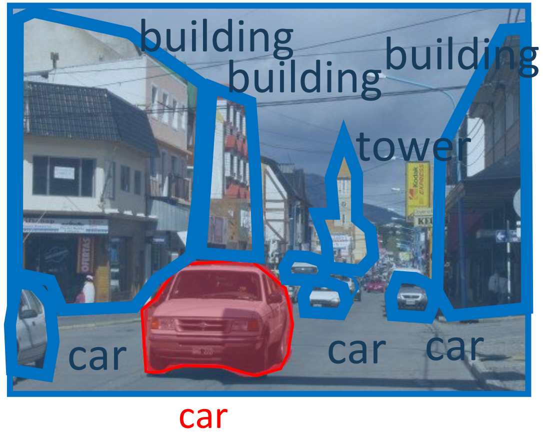
Second guy from left

ReferItGame Dataset

130k Referring expressions for **90k** Objects in **19k** images

ReferItGame: Referring to Objects in Photographs of Natural Scenes
Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, Tamara L. Berg.
Empirical Methods on Natural Language Processing. **EMNLP 2014**.

Referring Expression Generation

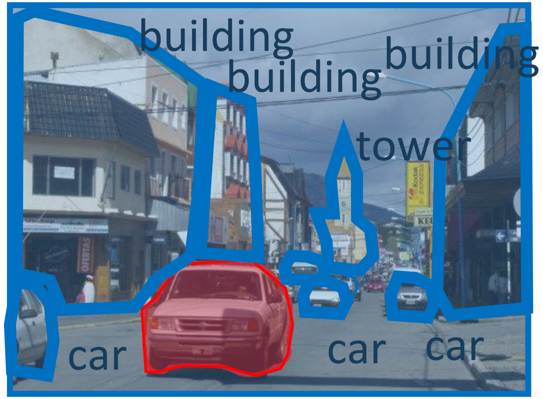


P: target object

S: scene

$$R = \begin{bmatrix} r_1\text{: object name} \\ r_2\text{: color} \\ r_3\text{: size} \\ r_4\text{: absolute location} \\ r_5\text{: relative location} \\ r_6\text{: relative object} \\ r_7\text{: other} \end{bmatrix}$$

Referring Expression Generation Output



car

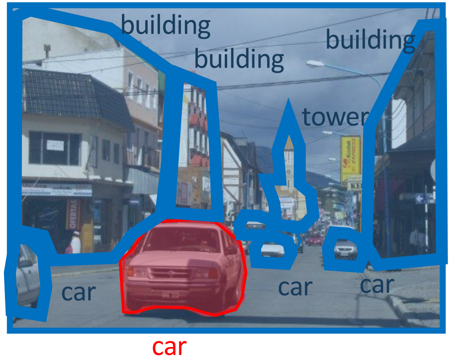
P: target object

S: scene

$$R = \begin{bmatrix} r_1\text{: truck} \\ r_2\text{: white} \\ r_3\text{: } \emptyset \\ r_4\text{: front} \\ r_5\text{: } \emptyset \\ r_6\text{: } \emptyset \\ r_7\text{: } \emptyset \end{bmatrix}$$

"the white truck in front"

Referring Expression Generation



P: target object
S: scene

$$R = \begin{pmatrix} r_1\text{: object name} \\ r_2\text{: color} \\ r_3\text{: size} \\ r_4\text{: absolute location} \\ r_5\text{: relative location} \\ r_6\text{: relative object} \\ r_7\text{: other} \end{pmatrix}$$

$$R^* = \underset{R}{\operatorname{argmax}} \; F(R, P, S)$$

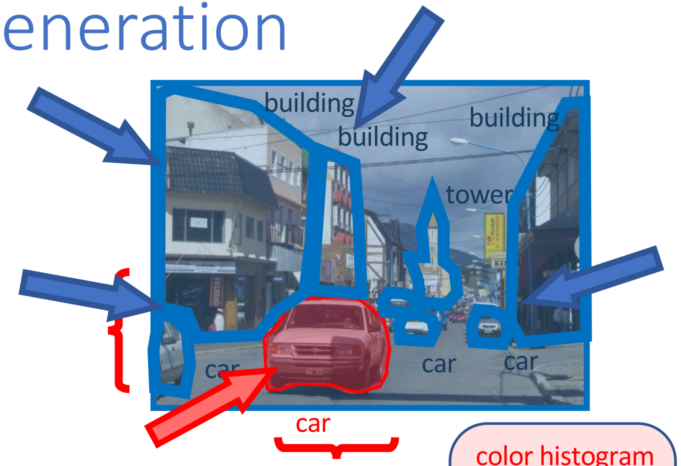
$$s.t. \quad f_i(R) \leq b_i$$

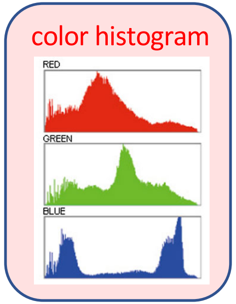
Where the function *F* scores the compatibility between a triple *R, P, S.*
And f_i,b_i impose constraints on the solution.

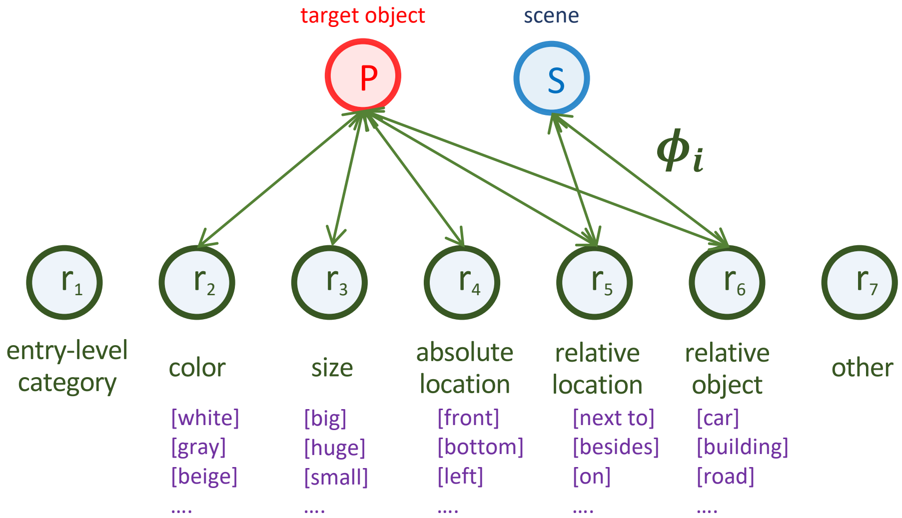
Referring Expression Generation

$$F(R, P, S) = \alpha \sum_{i=2}^{6} \boldsymbol{\phi}_i(r_i, P, S)$$

Content-based potential





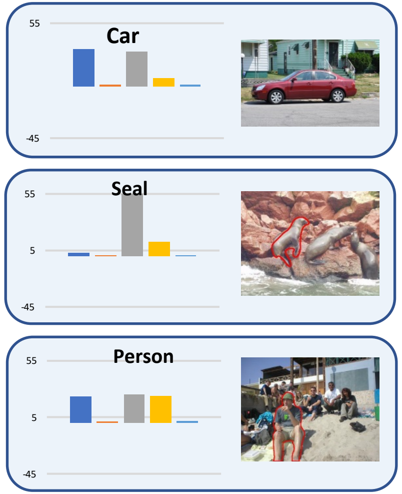


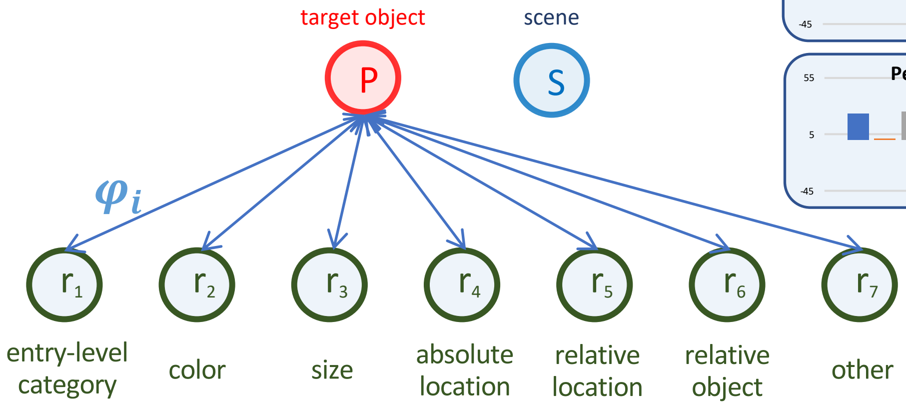
RefExp Generation: Prior-based term

$$F(R,P,S) = \alpha \sum_{i=2}^{6} \boldsymbol{\phi}_i(r_i,P,S) + \beta \sum_{i=1}^{7} \varphi_i(r_i, type(P))$$

Content-based potential

Prior-based potential





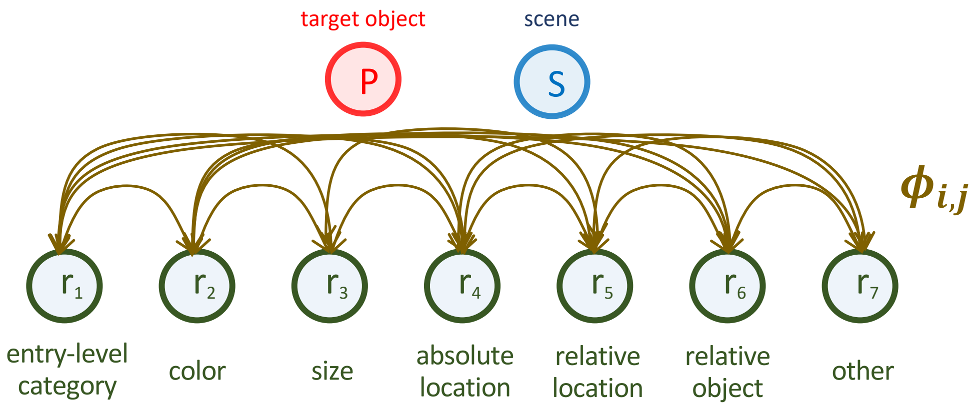
Referring Expression Generation

$$F(R,P,S) = \alpha \sum_{i=2}^{6} \phi_i(r_i,P,S) + \beta \sum_{i=1}^{7} \varphi_i(r_i, type(P)) + \sum_{i>j} \phi_{i,j}(r_i,r_j)$$

Content-based potential

Prior-based potential

Pairwise prior potential



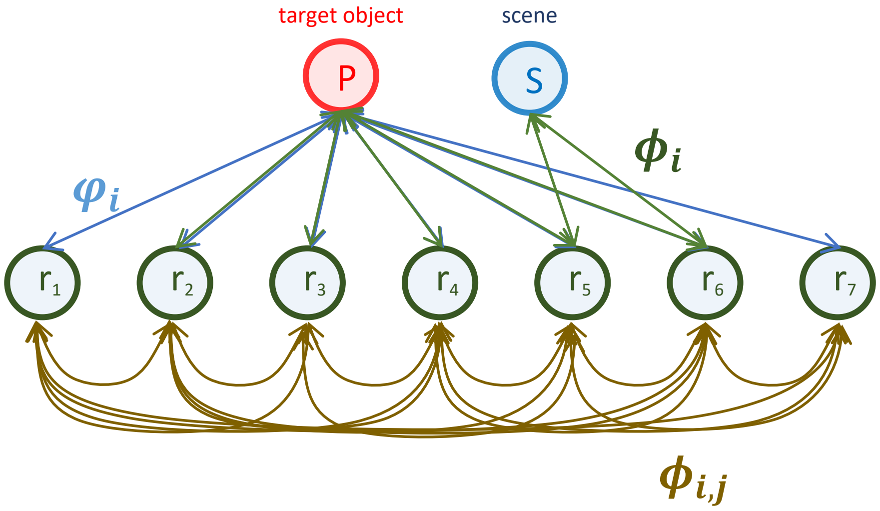
Referring Expression Generation

$$F(R,P,S) = \alpha \sum_{i=2}^{6} \phi_i(r_i,P,S) + \beta \sum_{i=1}^{7} \varphi_i(r_i, type(P)) + \sum_{i>j} \phi_{i,j}(r_i,r_j)$$

Content-based potential

Prior-based potential

Pairwise prior potential

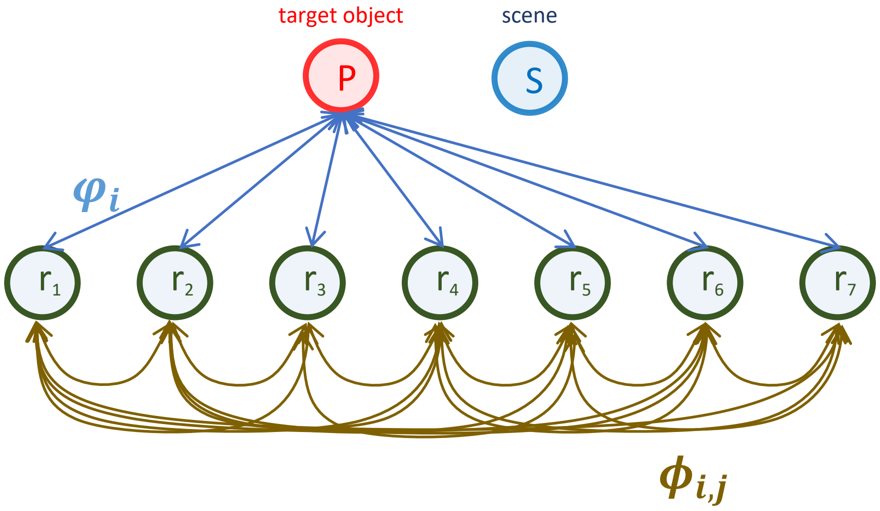


Referring Expression Generation

$$F(R, P, S) = \beta \sum_{i=1}^{7} \varphi_i(r_i, type(P)) + \sum_{i>j} \phi_{i,j}(r_i, r_j)$$

Prior-based potential

Pairwise prior potential



Referring Expression Generation: Results

Baseline: [door, white, , right , , ,]

Full: [door, white, , middle , , ,]

"white door"
"white door in the middle"
"door"

Baseline: [picture, white, , right, , ,]
Full: [picture, , , , prep_on, wall,]

"picture on the wall"
"picture"
"picture"

Referring Expression Generation: Results

Baseline: [building, white , , right , , ,]
Full: [building, brown , , middle , , ,]

"house"
"house"
"red brick house"

Baseline: [man, , , right, prep_in, floor,]
Full: [man, , , left, prep_in, floor,]

"red biker"
"person in red"
"far left person"

Referring Expression Generation: Evaluation

Test set A – 1000 Random Images

	Precision	Recall
Baseline	27.92	43.27
Full Model	**36.28**	**53.44**

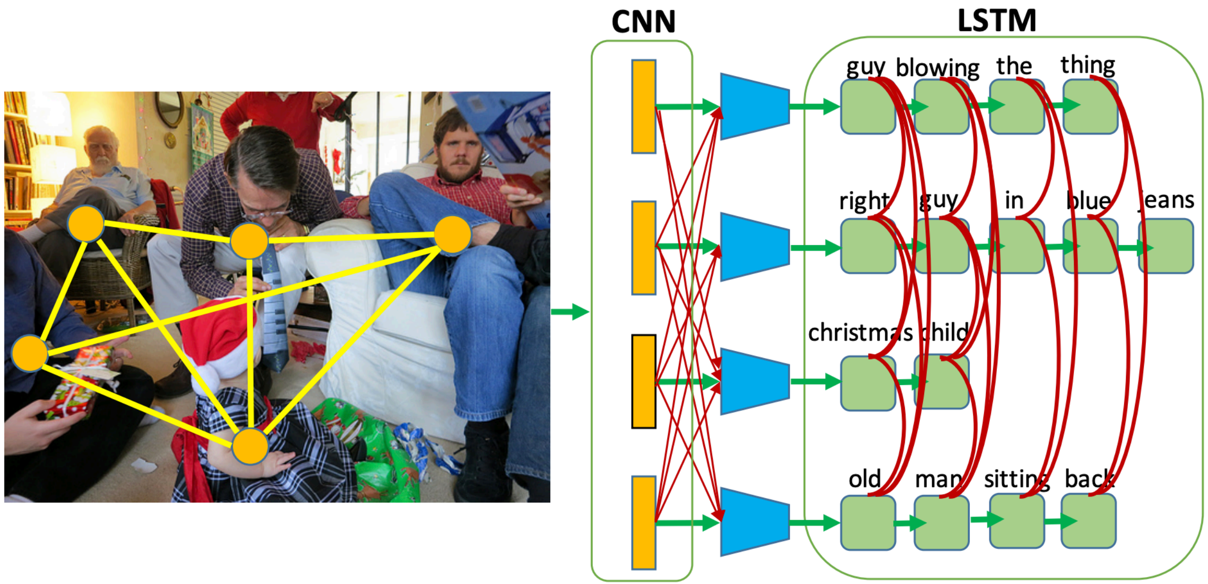
Test set B – 1000 Selected Objects

	Precision	Recall
Baseline	29.87	50.57
Full Model	**36.68**	**59.80**

Test set C – 1000 Images with Many Object Instances

	Precision	Recall
Baseline	28.85	37.41
Full Model	**37.73**	**48.54**

Deep Generation of Referring Expressions



Modeling Context in Referring Expressions

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, Tamara L. Berg

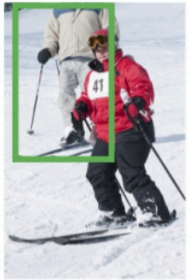
Department of Computer Science,
University of North Carolina at Chapel Hill
{licheng,poirson,alexyang,aberg,tlberg}@cs.unc.edu

RefCOCO+ testA

Baseline: blue shirt
MMI: black shirt
visdif: person in stripped shirt
visdif+tie: arm with stripped shirt

Baseline: tennis player
MMI: girl
visdif: woman in white
visdif+tie: tennis player

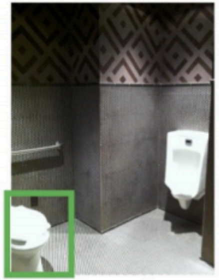
Baseline: man
MMI: man
visdif: man with glasses
visdif+tie: man with glasses



Baseline: red jacket
MMI: red jacket
visdif: skier in white
visdif+tie: man in white

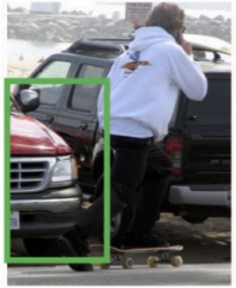
RefCOCO+ testB

Baseline: plant
MMI: plant that is cut off
visdif: tall plant
visdif+tie: plant on screen side



Baseline: toilet
MMI: toilet
visdif: toilet with lid
visdif+tie: toilet with lid

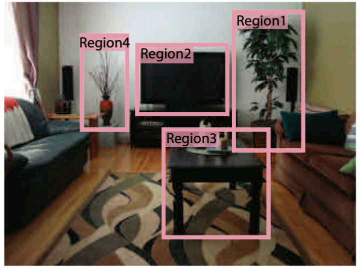
Baseline: donut at 3
MMI: glazed donut
visdif: donut with hole
visdif+tie: donut with hole



Baseline: car with red roof
MMI: car
visdif: car with headlights
visdif+tie: car with headlights

Referring Expression Comprehension

The plant on the right side of the TV

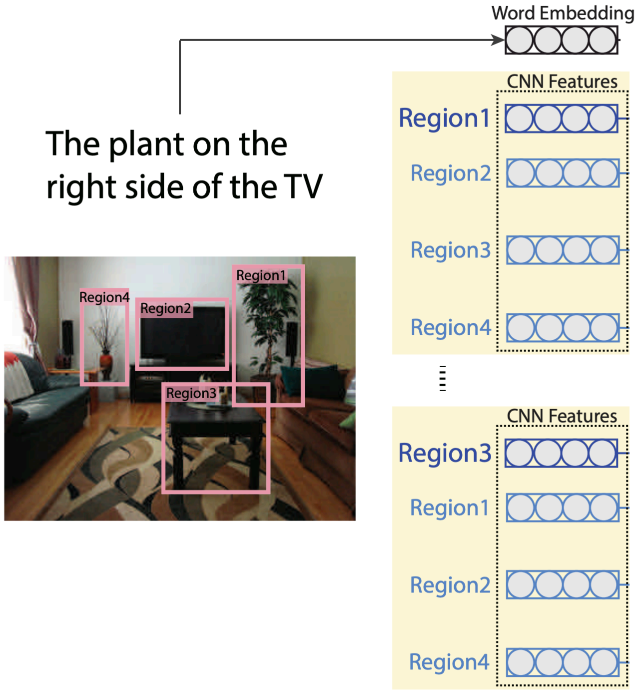


Modeling Context Between Objects for Referring Expression Understanding

Varun K. Nagaraja Vlad I. Morariu Larry S. Davis

University of Maryland, College Park, MD, USA.
{varun,morariu,lsd}@umiacs.umd.edu

Referring Expression Comprehension

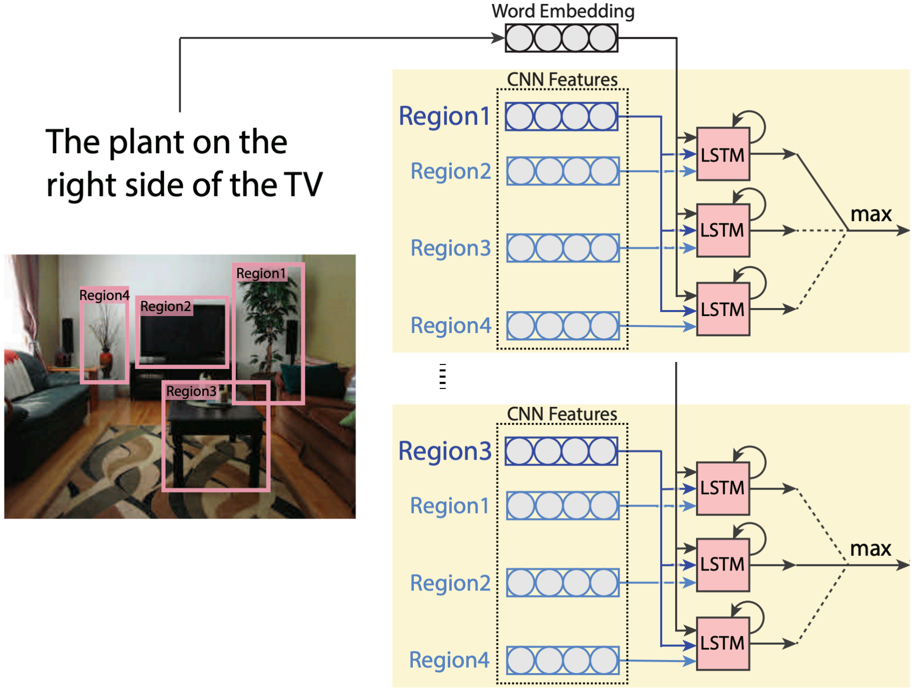


Modeling Context Between Objects for Referring Expression Understanding

Varun K. Nagaraja Vlad I. Morariu Larry S. Davis

University of Maryland, College Park, MD, USA.
{varun,morariu,lsd}@umiacs.umd.edu

Referring Expression Comprehension

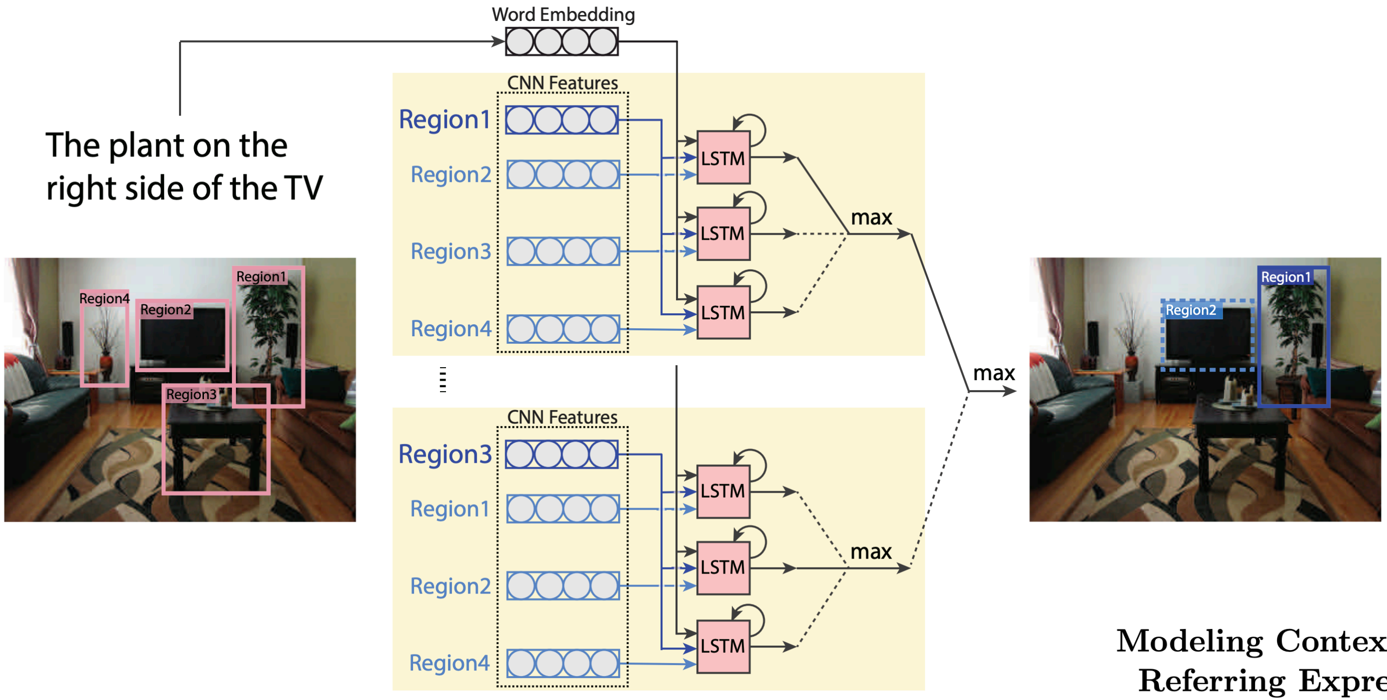


jects for
anding

Varun K. Nagaraja   Vlad I. Morariu   Larry S. Davis

University of Maryland, College Park, MD, USA.
{varun,morariu,lsd}@umiacs.umd.edu

Referring Expression Comprehension



Modeling Context Between Objects for Referring Expression Understanding

Varun K. Nagaraja Vlad I. Morariu Larry S. Davis

University of Maryland, College Park, MD, USA.
{varun,morariu,lsd}@umiacs.umd.edu

Other important work

MattNet: Yu et al. https://arxiv.org/abs/1801.08186

Mao et al. https://arxiv.org/abs/1511.02283

Rohrbach et al. https://arxiv.org/abs/1511.03745

Questions?