Vision and Language - Visual Referring Expression

CSE597: Vision and Language

Huijuan Xu, hkx5063@psu.edu

Outline

- Visual Referring Expression - images
- Visual Referring Expression - videos
  - Video Moment Retrieval
  - Video Corpus Moment Retrieval

Outline

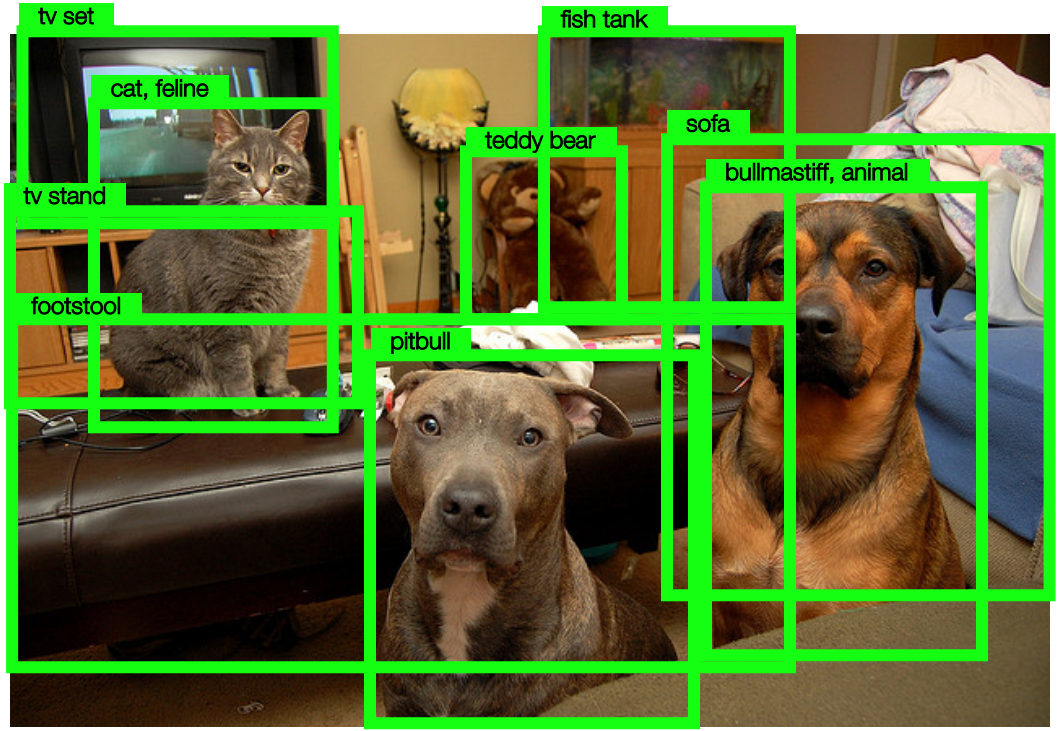
- **Visual Referring Expression - images**
- Visual Referring Expression - videos
  - Video Moment Retrieval
  - Video Corpus Moment Retrieval

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



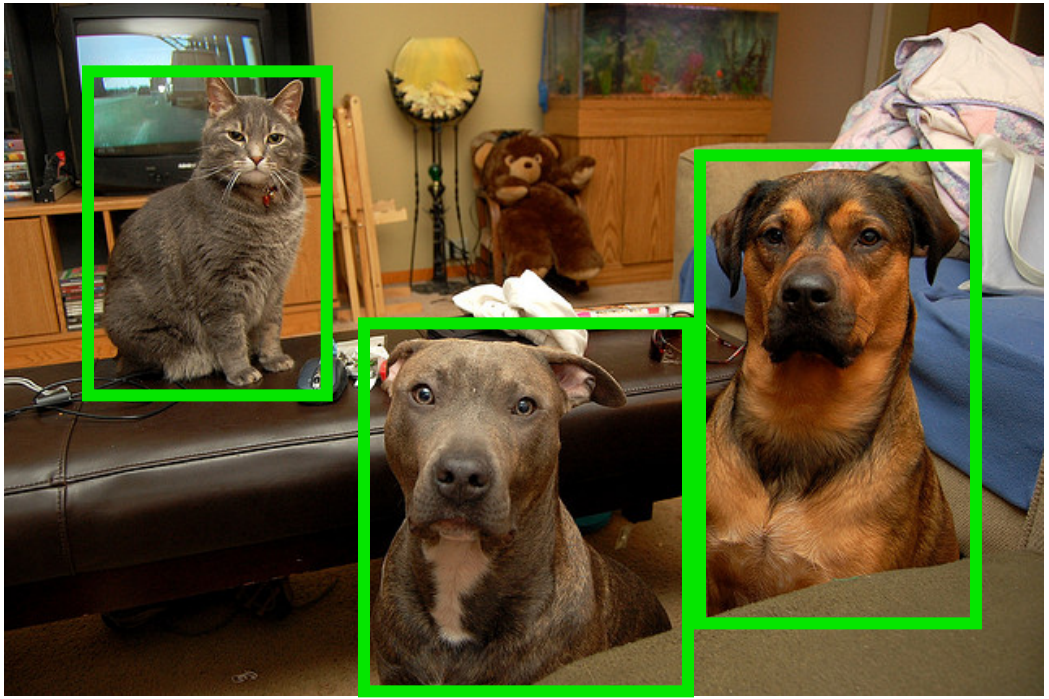
Object Detection

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

Computer Vision

Image Parsing / Image Segmentation

How do we describe images?



Object Importance

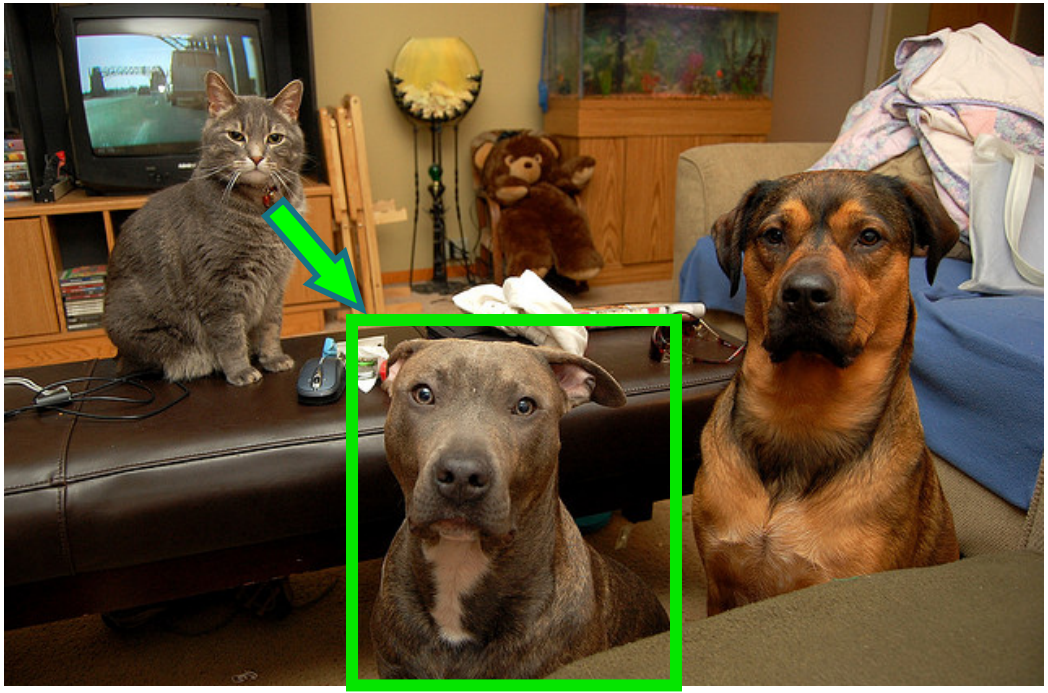
Attribute Importance

Action Importance

World knowledge

A cat and two big dogs staring at the camera

Referring to objects



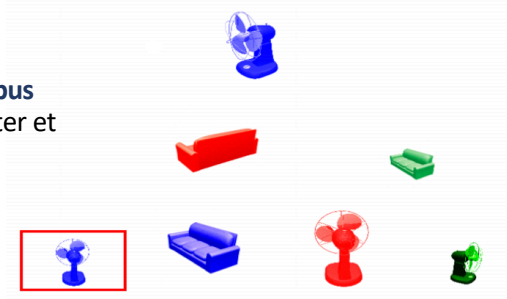
The dog in the middle

The gray dog in the middle

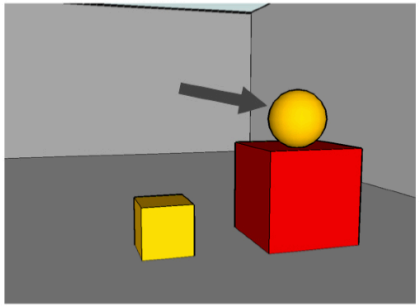
The gray dog

Work on Referring Expression

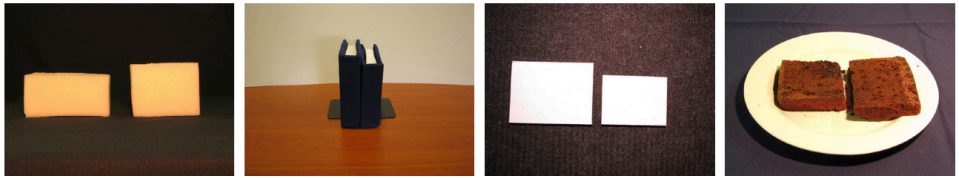
TUNA Corpus
van Deemter et al 2006



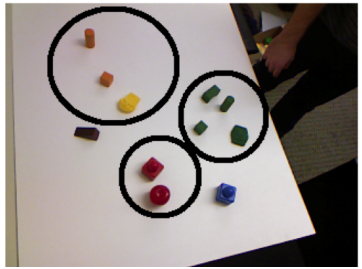
GRE3D3 Corpus
Viethen and Dale 2008
[20 scenes]



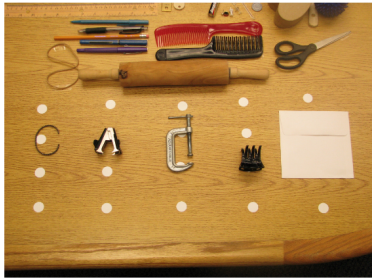
Size Corpus
Mitchell et al 2011
[96 scenes]



GenX Corpus
FitzGerald et al 2013
[269 scenes]



Typicality Corpus
Mitchell et al 2013
[35 scenes]



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

car

P: target object

S: scene

R =

r_1: object name
r_2: color
r_3: size
r_4: absolute location
r_5: relative location
r_6: relative object
r_7: other

Referring Expression Generation Output

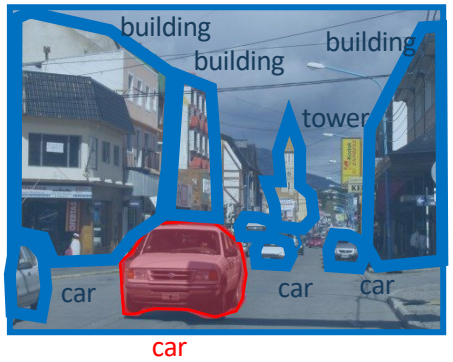
P: target object

S: scene

R = [r_1: truck, r_2: white, r_3: Ø, r_4: front, r_5: Ø, r_6: Ø, r_7: Ø]

“the white truck in front”

Referring Expression Generation



P: target object
S: scene

$R =$

- r_1: object name
- r_2: color
- r_3: size
- r_4: absolute location
- r_5: relative location
- r_6: relative object
- r_7: other

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

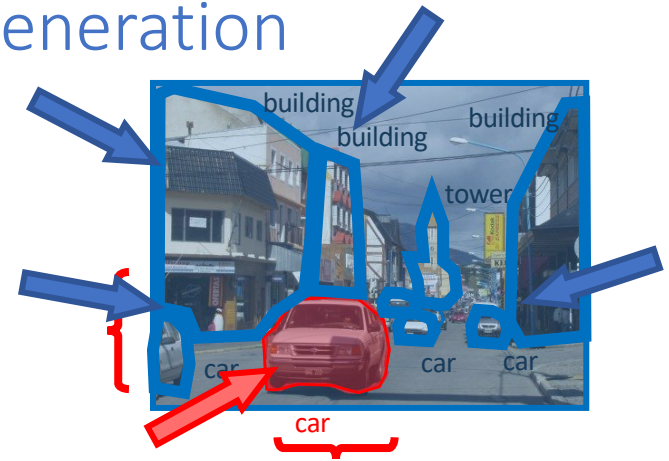
$$s.t. \quad f_i(R) \leq b_i$$

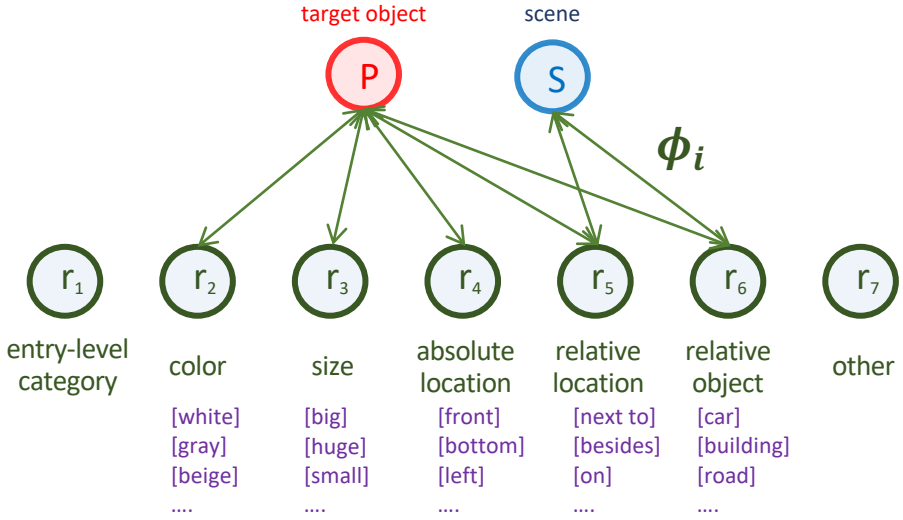
Where the function *F* scores the compatibility between a triple *R, P, S.*
And f_i, b_i impose constraints on the solution.

Referring Expression Generation

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

Content-based potential



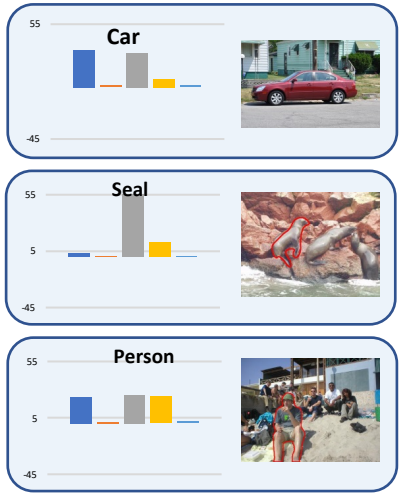


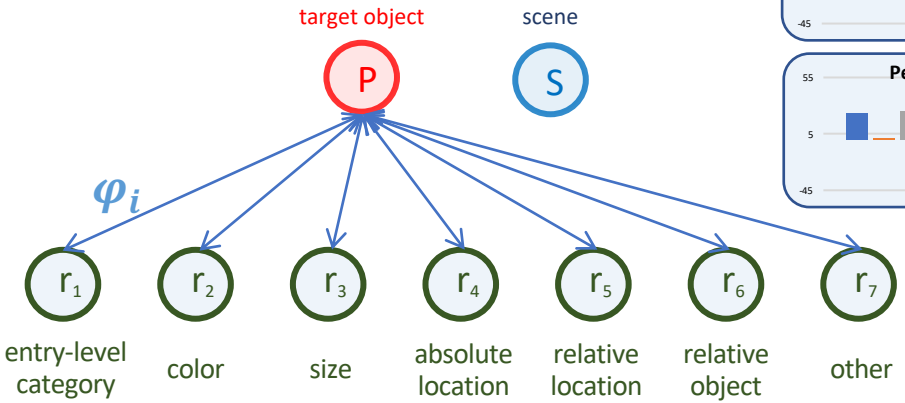
RefExp Generation: Prior-based term

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

Content-based potential

Prior-based potential





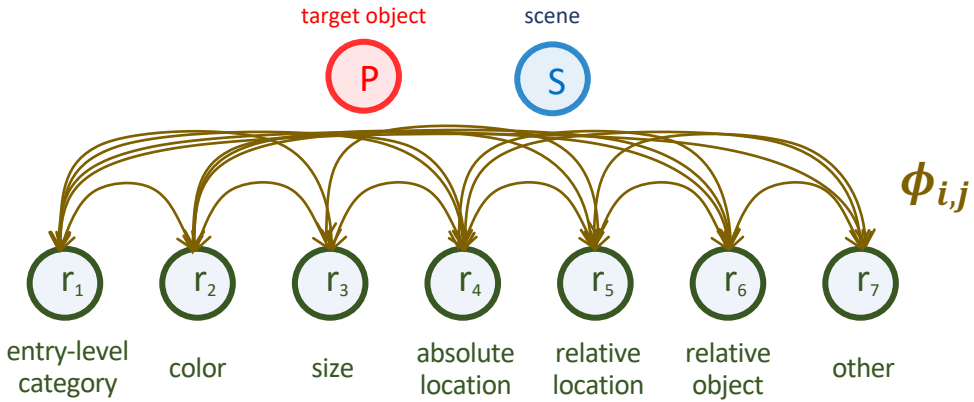
Referring Expression Generation

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

Content-based potential

Prior-based potential

Pairwise prior potential



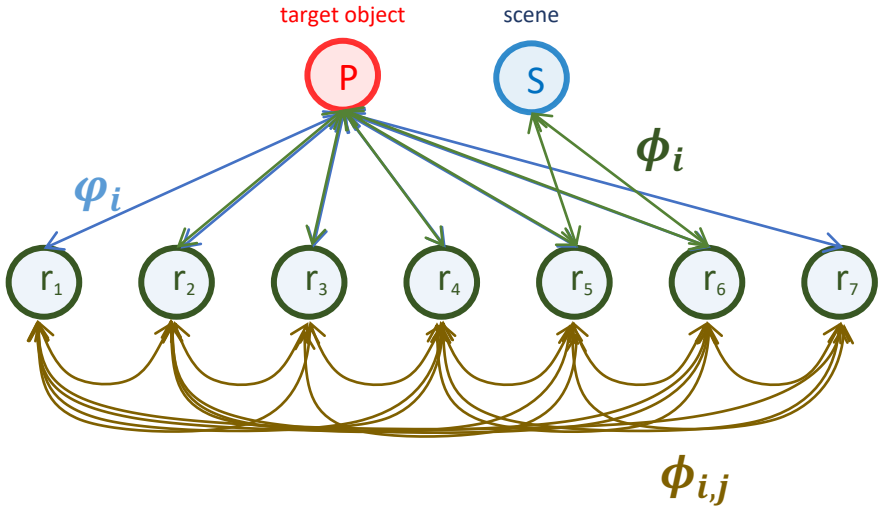
Referring Expression Generation

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

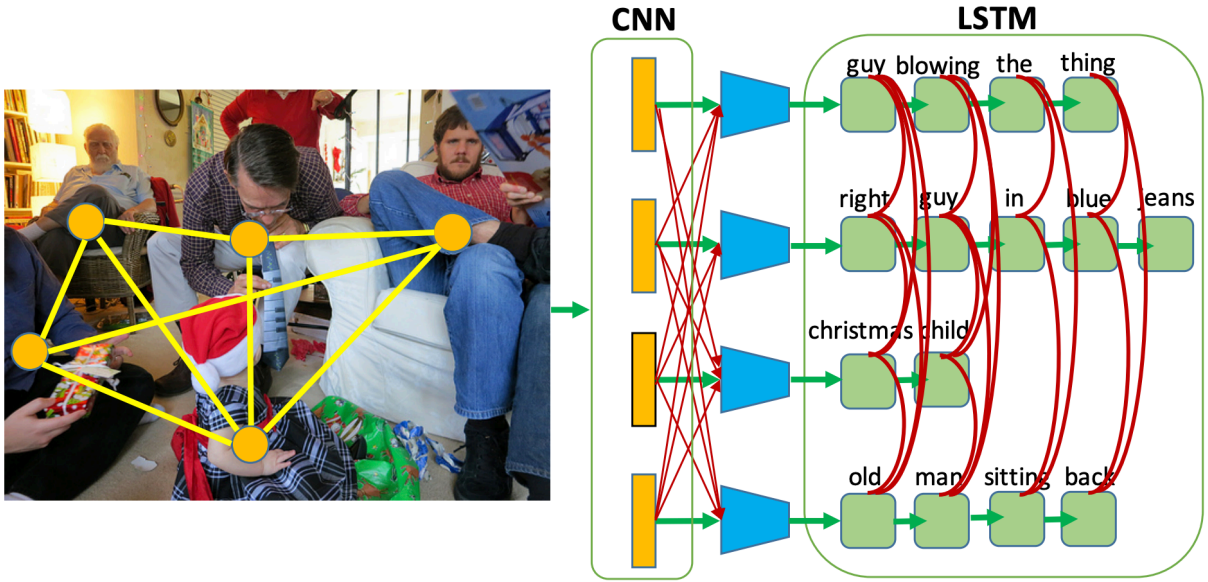
Content-based potential

Prior-based potential

Pairwise prior potential



Deep Generation of Referring Expressions



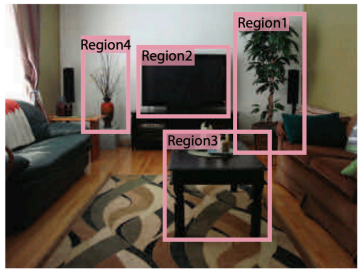
Modeling Context in Referring Expressions

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

Referring Expression Comprehension

The plant on the
right side of the TV

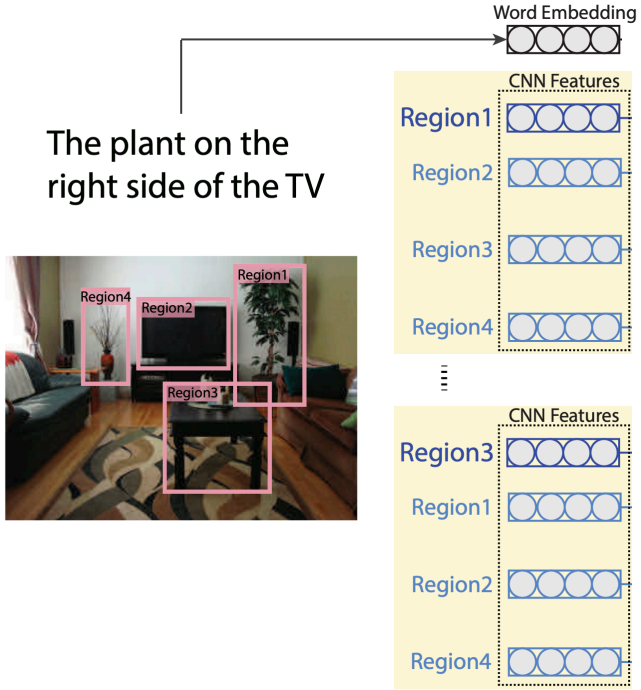


Modeling Context Between Objects for Referring Expression Understanding

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

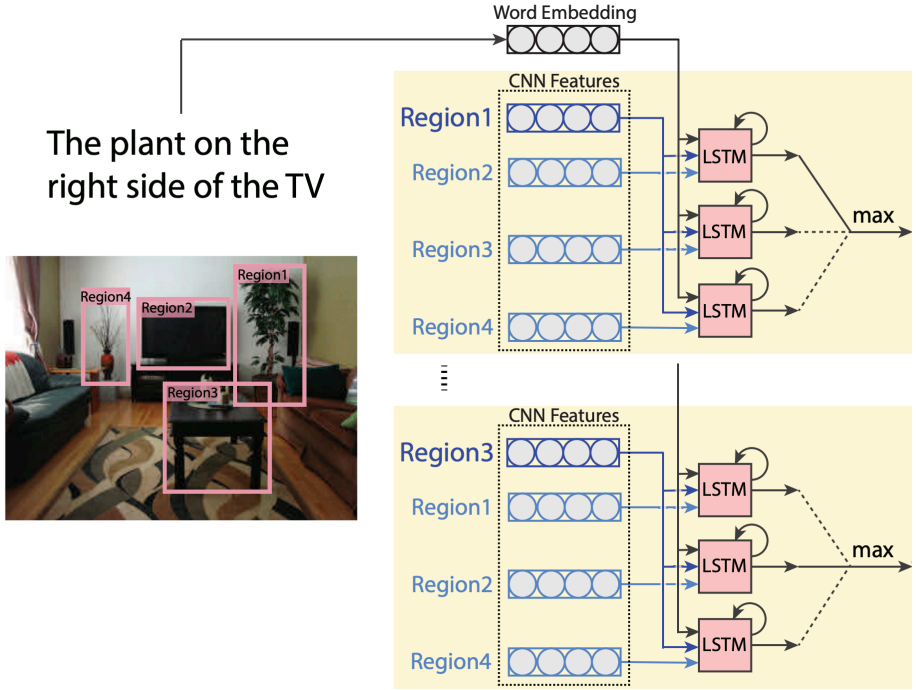


Modeling Context Between Objects for Referring Expression Understanding

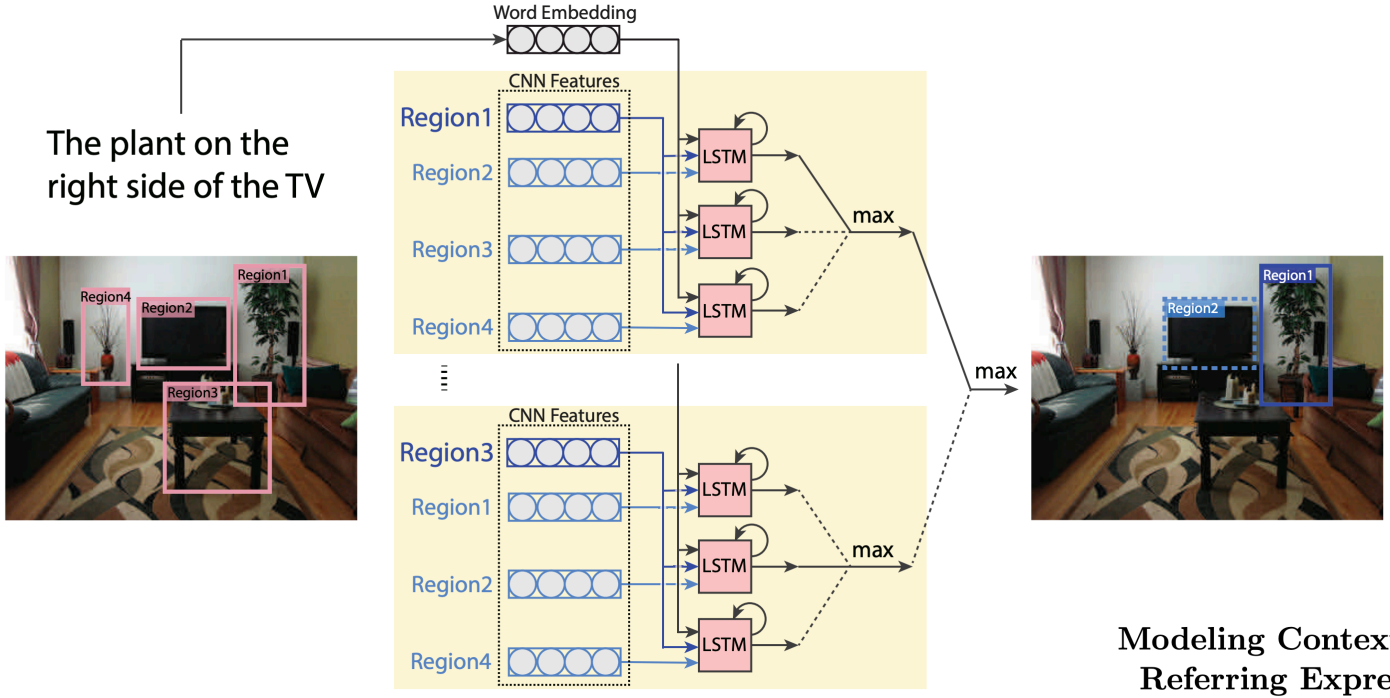
Varun K. Nagaraja Vlad I. Morariu Larry S. Davis

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

Referring Expression Comprehension



Referring Expression Comprehension



Modeling Context Between Objects for Referring Expression Understanding

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

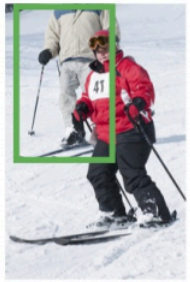
University of Maryland, College Park, MD, USA.
{varun,morariu,lsd}@umiacs.umd.edu

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

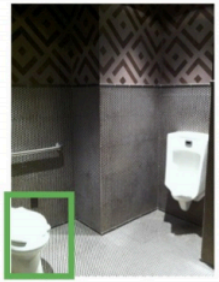
Baseline: man
MMI: man
visdif: man with glasses
visdif+tie: man with glasses



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

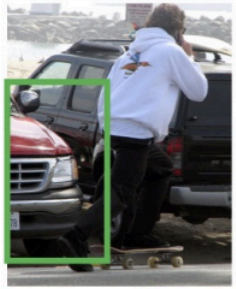
RefCOCO+ testB

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

Outline

- Visual Referring Expression - images
- **Visual Referring Expression - videos**
  - Video Moment Retrieval
  - Video Corpus Moment Retrieval

Single Video Moment Retrieval (SVMR)

a.k.a., temporal sentence grounding in video

Inputs:

An untrimmed video + **a language query**

Outputs:

The target moment

Query: Rachel explains to her dad on the phone why she can't marry her fiancé.

Video:

00:44 00:54

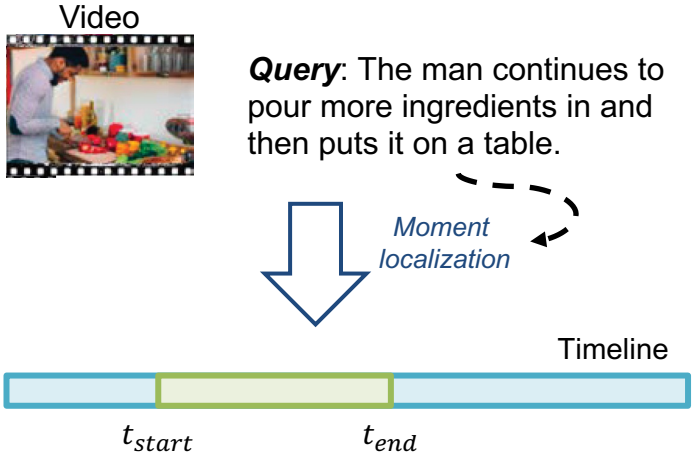
Target Moment

The example from TVRetrieval.

Video Corpus Moment Retrieval (VCMR)

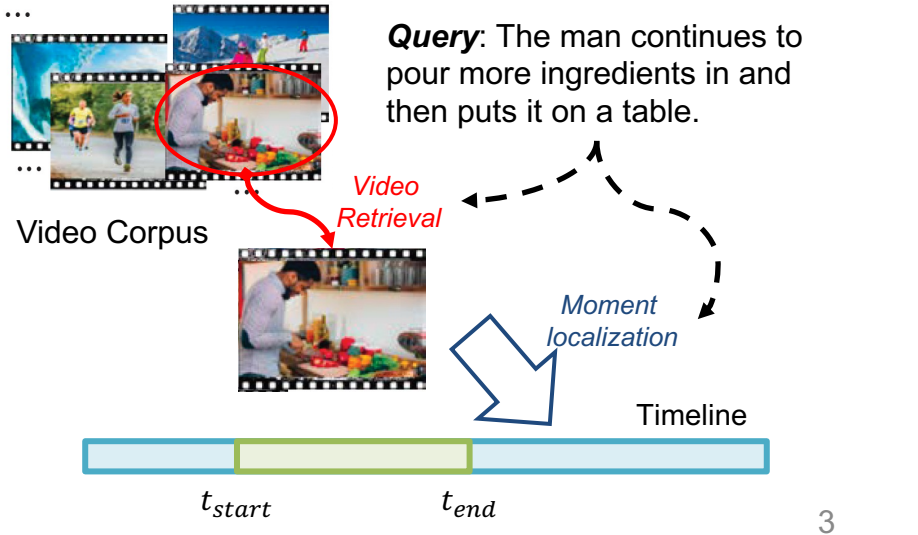
SVMR

Input: ***an untrimmed video***, language query
Output: target moment



VCMR

Input: ***video corpus with multiple videos***, language query
Output: target moment

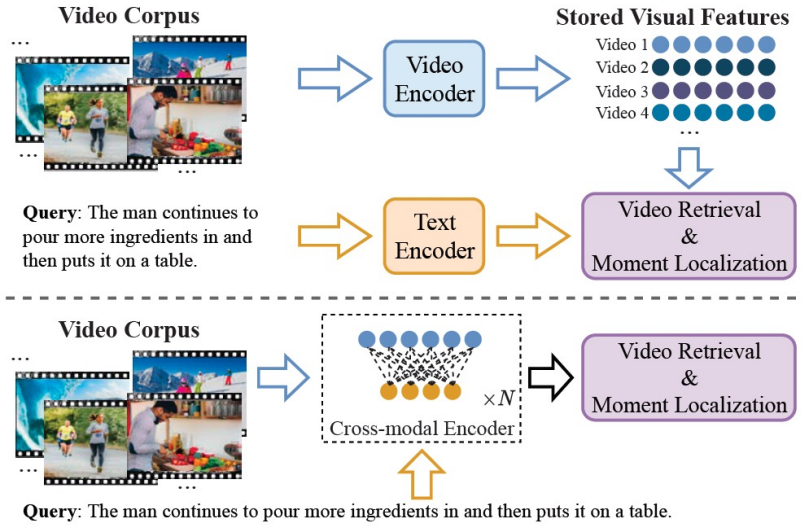


Existing VCMR Approaches

Video retrieval and **moment localization**

$$V^* = \arg\max_{V} p(V|Q) \quad \text{and} \quad m^* \approx \arg\max_{m \in V^*} p(m|V^*, Q)$$

V^* denotes the target video
m^* is the target temporal moment.



Existing VCMR Approaches

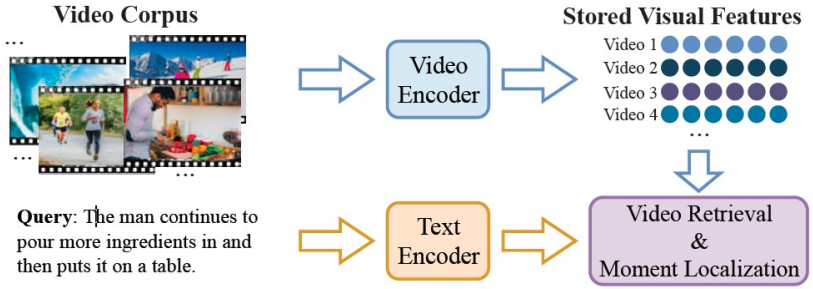
Unimodal Encoding Approach: to encode video and text **separately** and learn the matching through **late feature fusion**.

Pros:

High efficiency

Cons:

Low retrieval accuracy



Cross-modal Encoding Approach: to **jointly** encode query words and video features by **cross-modal reasoning** at fine-grained granularity.

Pros:

High retrieval accuracy

Cons:

Low efficiency