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

Image tagging / Image classification

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

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]
SEMINARY 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 Ordornez-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-externally. 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.
Referit Game

Player 1

Player 2

Orange bottle on the right

Orange bottle on the right
Referring Expressions for Natural Scenes

**Diverse**
Many real world objects

**Complex**
Many object instances

**Big**
20k images

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

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 = 

\[ r_1: \text{object name} \]
\[ r_2: \text{color} \]
\[ r_3: \text{size} \]
\[ r_4: \text{absolute location} \]
\[ r_5: \text{relative location} \]
\[ r_6: \text{relative object} \]
\[ r_7: \text{other} \]
Referring Expression Generation Output

P: target object
S: scene

"the white truck in front"
Referring Expression Generation

\[ R^* = \arg \max_R F(R, P, S) \]

\[ s.t. \quad f_i(R) \leq b_i \]

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

\[ F(R, P, S) = \alpha \sum_{i=2}^{6} \phi_i(r_i, P, S) \]

Content-based potential

- **P**
  - entry-level category
  - color: [white] [gray] [beige] ...
  - size: [big] [huge] [small] ...
  - absolute location: [front] [bottom] [left] ...
  - relative location: [next to] [besides] [on] ...
  - relative object: [car] [building] [road] ...
  - other

- **S**
  - scene
  - target object

- **φ_i**

**color histogram**
RefExp Generation: Prior-based term

\[ F(R, P, S) = \alpha \sum_{i=2}^{6} \phi_i(r_i, P, S) + \beta \sum_{i=1}^{7} \phi_i(r_i, \text{type}(P)) \]

- **Content-based potential**
- **Prior-based potential**

Diagram showing
- Target object
- Scene

**\( \phi_i \) for**
- \( r_1 \): entry-level category
- \( r_2 \): color
- \( r_3 \): size
- \( r_4 \): absolute location
- \( r_5 \): relative location
- \( r_6 \): relative object
- \( r_7 \): other

Comparative charts for:
- **Car**
- **Seal**
- **Person**
Referring Expression Generation

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

Content-based potential
Prior-based potential
Pairwise prior potential

Diagram showing target object and scene with nodes for entry-level category, color, size, absolute location, relative location, relative object, and other.
Referring Expression Generation

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

- Content-based potential
- Prior-based potential
- Pairwise prior potential

Diagram:
- Target object
- Scene
- Relations \( r_1, r_2, r_3, r_4, r_5, r_6, r_7 \)
- Relations \( \phi_{i,j} \)
Referring Expression Generation

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

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>27.92</td>
<td>43.27</td>
</tr>
<tr>
<td>Full Model</td>
<td>36.28</td>
<td>53.44</td>
</tr>
</tbody>
</table>

### Test set B – 1000 Selected Objects

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>29.87</td>
<td>50.57</td>
</tr>
<tr>
<td>Full Model</td>
<td>36.68</td>
<td>59.80</td>
</tr>
</tbody>
</table>

### Test set C – 1000 Images with Many Object Instances

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>28.85</td>
<td>37.41</td>
</tr>
<tr>
<td>Full Model</td>
<td>37.73</td>
<td>48.54</td>
</tr>
</tbody>
</table>
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,tilberg}@cs.unc.edu
RefCOCO+ testA

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

Baseline: tennis player
MMI: girl
visdif: woman in white
visdif+tie: man with glasses
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
The plant on the right side of the TV
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

The plant on the right side of the TV
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
Other important work


Questions?