Words and Pictures

Tamara L. Berg

Advanced Multimedia
Announcements

HW 4 due today!

Project Proposals (Thurs)
  5 mins – idea, challenges, timeline
  come to office hours or send email today
“It was an arresting face, pointed of chin, square of jaw. Her eyes were pale green without a touch of hazel, starred with bristly black lashes and slightly tilted at the ends. Above them, her thick black brows slanted upward, cutting a startling oblique line in her magnolia-white skin—that skin so prized by Southern women and so carefully guarded with bonnets, veils and mittens against hot Georgia suns” – Scarlett O’Hara, Gone with the Wind.

Visually descriptive language tells us:

1. how people construct natural language to describe imagery.
2. information about the world, especially the visual world.
More Nuance than Traditional Recognition…

- person
- shoe
- car
Where Can We Find Descriptive Text and Images?

- Web Pages
- Tagged Images
- Captioned Images
Giraffe

Learn about giraffe teeth on our video.

Giraffe Treat Jar Enrichment

Test your knowledge - take our quiz.

SWAHILI NAME: Twiga

PHYSICAL CHARACTERISTICS:

The Giraffe is the tallest animal in the world. Males may be 16-18 feet tall and weigh up to 2,000 pounds. Females are usually lighter and about two feet shorter. The giraffe's front legs are only slightly longer than the back ones, the height of the fore part of the body being largely due to the heavy muscular development of the base of the neck.

The long neck has the usual seven vertebrae of most mammals, although each is greatly elongated. The giraffe's soup-plate-sized hooves are used as offensive weapons, usually in the defense of the calves. The powerful kick from the front feet can kill a lion.

The giraffe is one of the few ruminant born with horns. Both sexes have horns which are covered with skin. The horns of males are thicker and heavier and are used in fights between males.

The giraffe's canine teeth are splayed out in two or three lobes to comb the leaves off shoots. Their long black tongue, which can be extended 18 inches, is used to gather food into the mouth. To compensate for the sudden increase in blood pressure when the head is lowered, the giraffe has very elastic blood vessels and valves in the venous system of the neck.

The Nubian giraffe's color scheme varies considerably in pattern, but consists of dark-reddish to chestnut brown blotches of various shapes and sizes on a buff ground color. The coloration darkens with age.

The reticulated giraffe has large, four-sided, liver-colored spots separated by a sharply defined network of narrow white lines.

DISTRIBUTION:

Giraffes inhabit savanna country where there are Acacia and climbing leguminous plants. They are found in Africa south of the Sahara to Cape Town. The reticulated giraffe is found in East Africa.
Tags or keywords + images

Tags: canon, eos, macro, japan, vacation, frog, animal, toad, amphibian, pet, eye, feet, mouth, finger, hand, prince, photo, art, light, photo, flickr, blurry, favorite, nice.

Flickr
>5 billion photos
3-5 million new/day
President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops.
President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops.
President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops.

It's the perfect party dress. With distinctly feminine details such as a wide sash bow around an empire waist and a deep scoopneck, this linen dress will keep you comfortable and feeling elegant all evening long.
* Measures 38" from center back, hits at the knee.
* Scoopneck, full skirt.
* Hidden side zip, fully lined.
* 100% Linen. Dry clean.
President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops.

Captioned images

It's the perfect party dress. With distinctly feminine details such as a wide sash bow around an empire waist and a deep scoopneck, this linen dress will keep you comfortable and feeling elegant all evening long.

* Measures 38” from center back, hits at the knee.
* Scoopneck, full skirt.
* Hidden side zip, fully lined.
* 100% Linen. Dry clean.
President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops.

Learn how people describe the world!
Beyond Object Categories

Attributes

- Black satin with clear Austrian crystals. Silver-plated brass...
- Tri-color bull's-eye design features really big French blue dots...
- Add a touch of regal color with these purple bead chandelier...
- Black patent leather slingbacks. Christian Louboutin shoes have …
- red (Color)
- ankle strap (Shape)
- high heel (Shape)

“high, heel” attribute localization

Predicted attributes and characterizations

Relationships

The car is on the street

Learn how the world works!
Themes

- Use large pools of descriptive text to:
  - Learn how people describe the visual world
    -- what text indicates about depiction.
  - Learn how the world works
    -- attributes of, and relationships between, objects.
  - Apply this knowledge to multi-modal collections.
Talk Outline

Faces in the Wild

Photographs with stylized captions.
Build on success of face *detection*.
Incorporate contextual information.

Sentence Generation

Use recognition + text priors to generate descriptions.
Face Recognition Datasets

PIE database

Yale Face Database

M2VTS
What the world really looks like
(scenes from an airport)
Producer and director Bruce Paltrow has died at the age of 58 in Rome, Italy, the U.S. Consulate said on October 3, 2002. Paltrow had suffered from throat cancer for several years, but the cause of his death was not immediately known. He is seen with his daughter actress Gwyneth Paltrow after the Academy Awards in Los Angeles in March 21, 1999 file photo. (Fred Prouser/Reuters)

Actress Winona Ryder (news) reacts to remarks by prosecutor Ann Rundle during the sentencing hearing in her felony shoplifting case Friday, Dec. 6, 2002 at the Beverly Hills, Calif., courthouse. At right is Ryder’s attorney Mark Geragos. Ryder was sentenced to three years of probation and was ordered to perform 480 hours of community service. (AP Photo/Steve Grayson, POOL)
President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops. Photo by Larry Downing/Reuters
President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops. Photo by Larry Downing/Reuters
President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops. Photo by Larry Downing/Reuters
President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops. Photo by Larry Downing/Reuters

**Detected Names:**
President George W. Bush, Defense Donald Rumsfeld, Saddam Hussein.
Image Model

Learn what each individual looks like:

Face Model: $P(\text{Person } i \mid \text{ Face Image})$
What cues do people use to indicate depiction?

President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops. Photo by Larry Downing/Reuters.
Cues - POS tags before and after name, location in caption, distance to closest:
( ) (L) (C) (R) left right center shown pictured above

President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops. Photo by Larry Downing/Reuters
1.) Learn image model for each person.

2.) Learn depiction model across names.

Achieve both of these by considering a big assignment problem -- 30,281 face detections with multiple associated names.
President and Chief Operating Officer of the Vancouver, British Columbia 2010 Bid Corporation John Furlong (rear) smiles while celebrating with compatriots their victory in obtaining the 2010 Winter Olympics bid on late July 2, 2003 in Prague. Vancouver won with 56 votes against 53 votes for Pyeongchang in the second round of balloting at an IOC gathering in Prague. REUTERS/Petr Josek
Assignment Problem

Use clustering to find best assignments
Each distinct name in the dataset is a cluster
Assign each face to one of the detected names
Assignment Problem

Repeat until convergence:
1) Update Appearance Models & Depiction Model
2) Update Assignments
Assignment Problem

Repeat until convergence:
1) Update Appearance Models & Depiction Model
2) Update Assignments
Results

Without – CEO Summit
With – Martha Stewart

Without – James Bond
With – Pierce Brosnan

Without – Dick Cheney
With – George W. Bush

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy of labeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vision model, No Lang model</td>
<td>67%</td>
</tr>
<tr>
<td>Vision model + Lang model</td>
<td>78%</td>
</tr>
</tbody>
</table>
Results - Depiction

<table>
<thead>
<tr>
<th>IN Pete Sampras IN of the U.S. celebrates his victory over Denmark’s OUT Kristian Pless OUT at the OUT U.S. Open OUT at Flushing Meadows August 30, 2002. Sampras won the match 6-3 7- 5 6-4. REUTERS/Kevin Lamarque</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germany’s IN Chancellor Gerhard Schroeder IN, left, in discussion with France’s IN President Jacques Chirac IN on the second day of the EU summit at the European Council headquarters in Brussels, Friday Oct. 25, 2002. EU leaders are to close a deal Friday on finalizing entry talks with 10 candidate countries after a surprise breakthrough agreement on Thursday between France and Germany regarding farm spending.(AP Photo/European Commission/HO)</td>
</tr>
<tr>
<td>‘The Right Stuff’ cast members IN Pamela Reed IN, (L) poses with fellow cast member IN Veronica Cartwright IN at the 20th anniversary of the film in Hollywood, June 9, 2003. The women played wives of astronauts in the film about early United States test pilots and the space program. The film directed by OUT Philip Kaufman OUT, is celebrating its 20th anniversary and is being released on DVD. REUTERS/Fred Prouser</td>
</tr>
<tr>
<td>Kraft Foods Inc., the largest U.S. food company, on July 1, 2003 said it would take steps, like capping portion sizes and providing more nutrition information, as it and other companies face growing concern and even lawsuits due to rising obesity rates. In May of this year, San Francisco attorney OUT Stephen Joseph OUT, shown above, sought to ban Oreo cookies in California – a suit that was withdrawn less than two weeks later. Photo by Tim Wimborne/Reuters REUTERS/Tim Wimborne</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classifier</th>
<th>% correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (all pictured)</td>
<td>67%</td>
</tr>
<tr>
<td>Learned Lang Model</td>
<td>86%</td>
</tr>
</tbody>
</table>
Talk Outline

Faces in the Wild

Photographs with stylized captions.
Build on success of face detection.
Incorporate contextual information.

Sentence Generation

Use recognition + text priors to generate descriptions.

There are one cow and one sky.
The golden cow is by the blue sky.
Baby Talk: Understanding and Generating Simple Image Descriptions

Automatically generated description:

“This picture shows one person, one grass, one chair, and one potted plant. The person is near the green grass, and in the chair. The green grass is by the chair, and near the potted plant.”
Methodology

1) computer vision recognition predictions

2) text statistics - parsed descriptive collections & Google

3) predict image labeling from vision & text input (CRF)

4) generation to produce descriptions
Components

- Objects in the world (objects and stuff)

- Appearances of objects (attributes/modifiers)

- Relationships between objects (spatial/prepositional relationships)
System Flow

1) Object(s)/Stuff
   - a) dog
   - b) person
   - c) sofa

2) Attributes
   - brown 0.01
   - striped 0.16
   - furry 0.26
   - wooden 0.2
   - feathered 0.06
   - brown 0.32
   - striped 0.09
   - furry 0.4
   - wooden 0.2
   - Feathered 0.04
   - brown 0.94
   - striped 0.10
   - furry 0.06
   - wooden 0.8
   - Feathered 0.08

3) Prepositions
   - near(a,b) 1
   - near(b,a) 1
   - against(a,b) 11
   - against(b,a) 0.04
   - beside(a,b) 0.24
   - beside(b,a) 0.17
   - near(a,c) 1
   - against(a,c) 0.3
   - against(c,a) 0.05
   - beside(a,c) 0.5
   - beside(c,a) 0.45
   - near(b,c) 1
   - near(c,b) 1
   - against(b,c) 0.67
   - against(c,b) 0.33
   - beside(b,c) 0.0
   - beside(c,b) 0.19

4) Constructed CRF

5) Predicted Labeling
   - <<null, person_b>, against, <<brown, sofa_c>>
   - <<null, dog_a>, near, <<null, person_b>>
   - <<null, dog_a>, beside, <<brown, sofa_c>>

6) Generated Sentences
   - This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.
System Flow

**Input Image**

This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.
System Flow

Input Image

a) dog

b) person

c) sofa

Extract Objects/stuff

2) Attributes

- brown 0.01
- striped 0.16
- furry 0.26
- wooden 0.2
- feathered 0.06
...

3) Prepositions

- near(a,b) 1
- against(a,b) 11
- against(b,a) 0.04
- beside(a,b) 0.24
- beside(b,a) 0.17
...

4) Constructed CRF

5) Predicted Labeling

6) Generated Sentences

This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.
System Flow

Input Image

a) dog
brown 0.01
striped 0.16
furry .26
wooden .2
feathered .06
...

b) person
brown 0.32
striped 0.09
furry .04
wooden .2
Feathered .04
...

c) sofa
brown 0.94
striped 0.10
furry .06
wooden .8
Feathered .08
...

Predict attributes

4) Constructed CRF

5) Predicted Labeling

6) Generated Sentences
This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.
System Flow

Predict prepositions

a) dog

b) person

c) sofa

near(a,b) 1
near(b,a) 1
against(a,b) .11
against(b,a) .04
beside(a,b) .24
beside(b,a) .17

near(b,c) 1
near(c,b) 1
against(b,c) .67
against(c,b) .33
beside(b,c) .0
beside(c,b) .19

...
System Flow

Input Image

1) Object(s)/Stuff
   - a) dog
   - b) person
   - c) sofa

2) Attribute
   - brown 0.01
   - striped 0.16
   - furry 0.28
   - wooden 0.12
   - feathered 0.06
   - ...

Predicted Labeling:

- <<null,person_b>>, against, <brown, sofa_c>>
- <<null,dog_a>>, near, <null,person_b>>
- <<null,dog_a>>, beside, <brown, sofa_c>>

Predict best (CRF) labeling – vision predictions smoothed with text potentials

ranted Sentences
photograph of one
and one brown sofa
dog. The person is
e the brown sofa. And
is near the person,
de the brown sofa.
This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.
System Flow

Input Image

1) Object(s)/Stuff
   a) dog
   b) person
   c) sofa

2) Attributes
   - brown 0.01
   - striped 0.16
   - furry 0.26
   - wooden_2 0.06
   - feathered 0.06

   - brown 0.32
   - striped 0.09
   - furry 0.04
   - wooden_2 0.04
   - feathered 0.04

   - brown 0.94
   - striped 0.10
   - furry 0.06
   - wooden_8 0.06
   - feathered 0.08

3) Prepositions
   - near(a,b) 1
   - against(a,b) 1
   - against(b,a) 11
   - against(b,a) 0.04
   - beside(a,b) 24
   - beside(b,a) 17
   - near(a,c) 1
   - against(a,c) 0.3
   - against(c,a) 0.05
   - beside(a,c) 5
   - beside(c,a) 4.5
   - near(b,c) 1
   - near(c,b) 1
   - against(b,c) 0.67
   - against(c,b) 0.33
   - beside(b,c) 4.0
   - beside(c,b) 1.9

4) Constructed CRF
   - person
   - sofa
   - dog

5) Predicted Labeling
   - null_person_b, against, brown_sofa_c
   - null_dog_a, near, null_person_b
   - null_dog_a, beside, brown_sofa_c

6) Generated Sentences
   This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.
Some good results

This is a photograph of one sky, one road and one bus. The blue sky is above the gray road. The gray road is near the shiny bus. The shiny bus is near the blue sky.

There are two aeroplanes. The first shiny aeroplane is near the second shiny aeroplane.

There are one cow and one sky. The golden cow is by the blue sky.

This is a picture of one sky, one road and one sheep. The gray sky is over the gray road. The gray sheep is by the gray road.

There are one dining table, one chair and two windows. The wooden dining table is by the wooden chair, and against the first window, and against the second white window. The wooden chair is by the first window, and by the second white window. The first window is by the second white window.

Here we see one road, one sky and one bicycle. The road is near the blue sky, and near the colorful bicycle. The colorful bicycle is within the blue sky.
More good results

Here we see two persons, one sky and one aeroplane. The first black person is by the blue sky. The blue sky is near the shiny aeroplane. The second black person is by the blue sky. The shiny aeroplane is by the first black person, and by the second black person.

This is a picture of two dogs. The first dog is near the second furry dog.

This is a photograph of two buses. The first rectangular bus is near the second rectangular bus.

In this picture there is one cat.

Here we see one person and one train. The black person is by the train.

This is a photograph of one horse and one person. The horse is near the person.
Some bad results

Missing detections:

Here we see one potted plant.

Incorrect detections:

There are one road and one cat. The furry road is in the furry cat.

Incorrect attributes:

This is a photograph of two sheeps and one grass. The first black sheep is by the green grass, and by the second black sheep. The second black sheep is by the green grass.

This is a picture of one tree, one road and one person. The rusty tree is under the red road. The colorful person is near the rusty tree, and under the red road.

This is a photograph of two horses and one grass. The first feathered horse is within the green grass, and by the second feathered horse. The second feathered horse is within the green grass.
Some bad results

Counting is hard!

There are two cows and one person. The first brown cow is against the brown person, and near the second cow. The brown person is beside the second cow.

This is a picture of four persons. The first colorful person is by the second pink person, and by the third colorful person. The second pink person is by the third colorful person, and by the fourth person.

Just all wrong!

There are one potted plant, one tree, one dog and one road. The gray potted plant is beneath the tree. The tree is near the black dog. The road is near the black dog. The black dog is near the gray potted plant.

This is a photograph of one person and one sky. The white person is by the blue sky.
Template vs Language Model

**Template-Based Generation:** This is a photograph of one furry sheep.

**Language Model-Based Generation:** the furry sheep.

**Template-Based Generation:** Here we see three persons, one sky, one grass and one train. The first colorful person is underneath the clear sky, and beside the second colorful person, and within the shiny train. The second colorful person is underneath the clear sky, and by the shiny train. The green grass is near the clear sky. The third black person is underneath the clear sky, and by the green grass, and within the shiny train. The shiny train is by the clear sky, and beside the green grass.

**Language Model-Based Generation:** the colorful person is underneath the clear sky. the colorful person who beside the colorful person. the colorful person is underneath the clear sky. the green grass and near the clear sky. the colorful person is within the shiny train. the black person is underneath the clear sky. the black person and by the green grass. the shiny train and by the clear sky. the colorful person and by the shiny train. the shiny train and beside the green grass. the black person is within the shiny train.
Evaluation

“This picture shows one person, one grass, one chair, and one potted plant. The person is near the green grass, and in the chair. The green grass is by the chair, and near the potted plant.”

Translation between images & text
Evaluation

“This picture shows one person, one grass, one chair, and one potted plant. The person is near the green grass, and in the chair. The green grass is by the chair, and near the potted plant.”

H1: A Lemonade stand is manned by a blonde child with a cookie.
H2: A small child at a lemonade and cookie stand on a city corner.
H3: Young child behind lemonade stand eating a cookie.

BLEU score – measure n-gram precision of machine generated sentences vs human generated sentences.
## Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>w/o synonyms</th>
<th>w/ synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-Human</td>
<td>0.5</td>
<td>0.51</td>
</tr>
<tr>
<td>Language model-based generation</td>
<td>0.25</td>
<td>0.30</td>
</tr>
<tr>
<td>Template-based generation</td>
<td>0.15</td>
<td>0.18</td>
</tr>
<tr>
<td>Meaning representation (triples)</td>
<td>0.20</td>
<td>0.30</td>
</tr>
</tbody>
</table>

**Automatic Evaluation: BLEU score**
Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>w/o synonyms</th>
<th>w/ synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-Human</td>
<td>0.5</td>
<td>0.51</td>
</tr>
<tr>
<td>Language model-based generation</td>
<td>0.25</td>
<td>0.30</td>
</tr>
<tr>
<td>Template-based generation</td>
<td>0.15</td>
<td>0.18</td>
</tr>
<tr>
<td>Meaning representation (triples)</td>
<td>0.20</td>
<td>0.30</td>
</tr>
</tbody>
</table>

**Automatic Evaluation: BLEU score**

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality of Image Parsing</td>
<td>2.85</td>
</tr>
<tr>
<td>Language model-based generation</td>
<td>2.77</td>
</tr>
<tr>
<td>Template-based generation</td>
<td>3.49</td>
</tr>
</tbody>
</table>

**Human Evaluation: possible scores are 4 (perfect without error), 3 (good with some errors), 2 (many errors), 1 (failure)**