Baby Talk: Understanding and Generating Simple Image Descriptions

Girish Kulkarni  Visruth Premraj  Sagnik Dhar  Siming Li
Yejin Choi  Alexander C Berg  Tamara L Berg

Stony Brook University, NY 11794, USA

February 12, 2013

Presented by: Prakhar Banga
Natural Language Generation
Natural Language Generation

- Given an image
Natural Language Generation

- Given an image
- Generate English description

Extremely useful for image indexing and search
Natural Language Generation

- Given an image
- Generate English description
- Extremely useful for image indexing and search
Summarizing Images
Summarizing Images

Object Detection
Summarizing Images

Object Detection
Summarizing Images

Object Detection
Summarizing Images

Object Detection

1) Objects
   a) dog
   b) person
   c) sofa

2) Attributes
   brown
   striped
   furry
   wooden
   feathered

3) Prepositions
   near(a,b)
   near(b,a)
   against(a,b)
   against(b,a)
   beside(a,b)
   beside(b,a)

4) Construct a CRF

5) Predicted Labeling

6) Sentence
   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.
Summarizing Images

Object Detection

1) Objects

a) dog

...
Summarizing Images

Object Detection

1) Objects

   a) dog

   b) person

2) Attributes

   brown 0.01
   striped 0.16
   furry 0.26
   wooden 0.2
   feathered 0.06
   ... 
   brown 0.32
   striped 0.09
   furry 0.04
   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) 0.11
   against(b,a) 0.04
   beside(a,b) 0.24
   beside(b,a) 0.17
   ... 
   near(a,c) 1
   near(c,a) 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) Construct a CRF

5) Predicted Labeling

   <<null, personb, against, <brown, sofaca>>
   <<null, doga, near, <null, personb>>
   <<null, doga, beside, <brown, sofaca>>

6) Sentence

   This is a photograph of one person(97,294),(181,409) 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.
Summarizing Images

Object Detection

1) Objects

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 0.04
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) 0.11
against(b,a) 0.04
beside(a,b) 0.24
beside(b,a) 0.17
... ...

near(a,c) 1
near(c,a) 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) Construct a CRF

5) Predicted Labeling

<<null, personb>, against, <brown, sofac>>

<<null, doga>, near, <null, personb>>

<<null, doga>, beside, <brown, sofac>>

6) Sentence

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.
Summarizing Images

Attribute Extraction

1) Objects

- a) dog
- b) person
- c) sofa

2) Attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>brown</td>
<td>0.01</td>
</tr>
<tr>
<td>striped</td>
<td>0.16</td>
</tr>
<tr>
<td>furry</td>
<td>0.26</td>
</tr>
<tr>
<td>wooden</td>
<td>0.20</td>
</tr>
<tr>
<td>feathered</td>
<td>0.06</td>
</tr>
</tbody>
</table>

3) Prepositions

<table>
<thead>
<tr>
<th>Preposition</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>near(a,b)</td>
<td>1.00</td>
</tr>
<tr>
<td>near(b,a)</td>
<td>1.00</td>
</tr>
<tr>
<td>against(a,b)</td>
<td>0.11</td>
</tr>
<tr>
<td>against(b,a)</td>
<td>0.04</td>
</tr>
<tr>
<td>beside(a,b)</td>
<td>0.24</td>
</tr>
<tr>
<td>beside(b,a)</td>
<td>0.17</td>
</tr>
<tr>
<td>near(a,c)</td>
<td>1.00</td>
</tr>
<tr>
<td>near(c,a)</td>
<td>1.00</td>
</tr>
<tr>
<td>against(a,c)</td>
<td>0.30</td>
</tr>
<tr>
<td>against(c,a)</td>
<td>0.05</td>
</tr>
<tr>
<td>beside(a,c)</td>
<td>0.50</td>
</tr>
<tr>
<td>beside(c,a)</td>
<td>0.45</td>
</tr>
<tr>
<td>near(b,c)</td>
<td>1.00</td>
</tr>
<tr>
<td>near(c,b)</td>
<td>1.00</td>
</tr>
<tr>
<td>against(b,c)</td>
<td>0.67</td>
</tr>
<tr>
<td>against(c,b)</td>
<td>0.33</td>
</tr>
<tr>
<td>beside(b,c)</td>
<td>0.00</td>
</tr>
<tr>
<td>beside(c,b)</td>
<td>0.19</td>
</tr>
</tbody>
</table>

4) Construct a CRF

5) Predicted Labeling

<<null, personb, against, <brown, sofaca>>

<<null, doga, near, <null, personb>>

<<null, doga, beside, <brown, sofaca>>

6) Sentence

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.
Summarizing Images

Attribute Extraction

1) Objects
   a) dog
   b) person
   c) sofa

2) Attributes
   - brown: 0.01
   - striped: 0.16
   - furry: 0.26
   - wooden: 0.2
   - feathered: 0.06

3) Prepositions
   - near(a,b): 1
   - near(b,a): 1
   - against(a,b): 0.11
   - against(b,a): 0.04
   - beside(a,b): 0.24
   - beside(b,a): 0.17

4) Construct a CRF

5) Predicted Labeling
   - null, personb, against, <brown, sofac>
   - null, doga, near, <null, personb>
   - null, doga, beside, <brown, sofac>

6) Sentence
   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.
Summarizing Images

Attribute Extraction

1) Objects

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a) dog</td>
<td></td>
</tr>
<tr>
<td>b) person</td>
<td></td>
</tr>
<tr>
<td>c) sofa</td>
<td></td>
</tr>
</tbody>
</table>

2) Attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>brown</td>
<td>0.01</td>
</tr>
<tr>
<td>striped</td>
<td>0.16</td>
</tr>
<tr>
<td>furry</td>
<td>0.26</td>
</tr>
<tr>
<td>wooden</td>
<td>0.2</td>
</tr>
<tr>
<td>feathered</td>
<td>0.06</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>brown</td>
<td>0.32</td>
</tr>
<tr>
<td>striped</td>
<td>0.09</td>
</tr>
<tr>
<td>furry</td>
<td>0.04</td>
</tr>
<tr>
<td>wooden</td>
<td>0.2</td>
</tr>
<tr>
<td>Feathered</td>
<td>0.04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>brown</td>
<td>0.94</td>
</tr>
<tr>
<td>striped</td>
<td>0.10</td>
</tr>
<tr>
<td>furry</td>
<td>0.06</td>
</tr>
<tr>
<td>wooden</td>
<td>0.8</td>
</tr>
<tr>
<td>Feathered</td>
<td>0.08</td>
</tr>
</tbody>
</table>

3) Prepositions

<table>
<thead>
<tr>
<th>Preposition</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>near(a,b)</td>
<td>1</td>
</tr>
<tr>
<td>near(b,a)</td>
<td>1</td>
</tr>
<tr>
<td>against(a,b)</td>
<td>0.11</td>
</tr>
<tr>
<td>against(b,a)</td>
<td>0.04</td>
</tr>
<tr>
<td>beside(a,b)</td>
<td>0.24</td>
</tr>
<tr>
<td>beside(b,a)</td>
<td>0.17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Preposition</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>near(a,c)</td>
<td>1</td>
</tr>
<tr>
<td>near(c,a)</td>
<td>1</td>
</tr>
<tr>
<td>against(a,c)</td>
<td>0.3</td>
</tr>
<tr>
<td>against(c,a)</td>
<td>0.05</td>
</tr>
<tr>
<td>beside(a,c)</td>
<td>0.5</td>
</tr>
<tr>
<td>beside(c,a)</td>
<td>0.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Preposition</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>near(b,c)</td>
<td>1</td>
</tr>
<tr>
<td>near(c,b)</td>
<td>1</td>
</tr>
<tr>
<td>against(b,c)</td>
<td>0.67</td>
</tr>
<tr>
<td>against(c,b)</td>
<td>0.33</td>
</tr>
<tr>
<td>beside(b,c)</td>
<td>0.0</td>
</tr>
<tr>
<td>beside(c,b)</td>
<td>0.19</td>
</tr>
</tbody>
</table>

4) Construct a CRF

5) Predicted Labeling

<null>, personb, against, <null>, personb>
<null>, doga, near, <null>, personb>
<null>, doga, beside, <null>, personb>

6) Sentence
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.
Summarizing Images
Relative Positioning

1) Objects
2) Attributes
3) Prepositions

<table>
<thead>
<tr>
<th>Objects</th>
<th>Attributes</th>
<th>Prepositions</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) dog</td>
<td>brown 0.01</td>
<td>near(a,b) 1</td>
</tr>
<tr>
<td></td>
<td>striped 0.16</td>
<td>near(b,a) 1</td>
</tr>
<tr>
<td></td>
<td>furry 0.26</td>
<td>against(a,b) 0.11</td>
</tr>
<tr>
<td></td>
<td>wooden 0.2</td>
<td>against(b,a) 0.04</td>
</tr>
<tr>
<td></td>
<td>feathered 0.06</td>
<td>beside(a,b) 0.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>beside(b,a) 0.17</td>
</tr>
<tr>
<td>b) person</td>
<td>brown 0.32</td>
<td>near(a,c) 1</td>
</tr>
<tr>
<td></td>
<td>striped 0.09</td>
<td>near(c,a) 1</td>
</tr>
<tr>
<td></td>
<td>furry 0.04</td>
<td>against(a,c) 0.3</td>
</tr>
<tr>
<td></td>
<td>wooden 0.2</td>
<td>against(c,a) 0.05</td>
</tr>
<tr>
<td></td>
<td>Feathered 0.04</td>
<td>beside(a,c) 0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>beside(c,a) 0.45</td>
</tr>
<tr>
<td>c) sofa</td>
<td>brown 0.94</td>
<td>near(b,c) 1</td>
</tr>
<tr>
<td></td>
<td>striped 0.10</td>
<td>near(c,b) 1</td>
</tr>
<tr>
<td></td>
<td>furry 0.06</td>
<td>against(b,c) 0.67</td>
</tr>
<tr>
<td></td>
<td>wooden 0.8</td>
<td>against(c,b) 0.33</td>
</tr>
<tr>
<td></td>
<td>Feathered 0.08</td>
<td>beside(b,c) 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>beside(c,b) 0.19</td>
</tr>
</tbody>
</table>

4) Construct a CRF
5) Predicted Labeling

<<null, personb, against, brown, sofac>>
<<null, doga, near, null, personb>>
<<null, doga, beside, brown, sofac>>

6) Sentence
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.
Summarizing Images
Relative Positioning

1) Objects
   a) dog
   b) person
   c) sofa

2) Attributes
   - brown
   - striped
   - furry
   - wooden
   - feathered

3) Prepositions
   - near(a,b)
   - near(b,a)
   - against(a,b)
   - against(b,a)
   - beside(a,b)
   - beside(b,a)
   - beside(a,c)
   - beside(c,a)

4) Construct a CRF
5) Predicted Labeling

6) Sentence
   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.
### Summarizing Images

#### Relative Positioning

1) **Objects**

   a) dog
   
   b) person
   
   c) sofa

2) **Attributes**

   - brown
   - striped
   - furry
   - wooden
   - feathered

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>brown</td>
<td>0.01</td>
</tr>
<tr>
<td>striped</td>
<td>0.16</td>
</tr>
<tr>
<td>furry</td>
<td>0.26</td>
</tr>
<tr>
<td>wooden</td>
<td>0.2</td>
</tr>
<tr>
<td>feathered</td>
<td>0.06</td>
</tr>
</tbody>
</table>

3) **Prepositions**

   - near(a,b)
   - against(a,b)
   - beside(a,b)
   - against(b,a)
   - beside(b,a)
   - near(b,a)
   - against(b,c)
   - beside(b,c)
   - against(c,a)
   - beside(c,a)

### Construct a CRF

#### Predicted Labeling

- ([null], person, against, [null])
- ([null], dog, near, [null], person)
- ([null], dog, beside, [null], sofa)

#### Sentence

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.
Summarizing Images

Sentence generation

1) Objects
   a) dog
   b) person
   c) sofa

2) Attributes
   - brown
   - striped
   - furry
   - wooden
   - feathered

3) Prepositions
   - near(a,b)
   - near(b,a)
   - against(a,b)
   - against(b,a)
   - beside(a,b)
   - beside(b,a)

4) Construct a CRF

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.
Summarizing Images

Sentence generation

1) Objects
   a) dog
   b) person
   c) sofa

2) Attributes
   brown
   striped
   furry
   wooden
   feathered

3) Prepositions
   near(a,b)
   near(b,a)
   against(a,b)
   against(b,a)
   beside(a,b)
   beside(b,a)

4) Construct a 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 >>
Summarizing Images

Sentence generation

1) Objects
   a) dog
   b) person
   c) sofa

2) Attributes
<table>
<thead>
<tr>
<th>Objects</th>
<th>Attribute</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>brown</td>
<td>0.01</td>
</tr>
<tr>
<td>dog</td>
<td>striped</td>
<td>0.16</td>
</tr>
<tr>
<td>dog</td>
<td>furry</td>
<td>0.26</td>
</tr>
<tr>
<td>dog</td>
<td>wooden</td>
<td>0.2</td>
</tr>
<tr>
<td>person</td>
<td>feathered</td>
<td>0.06</td>
</tr>
<tr>
<td>sofa</td>
<td>feathered</td>
<td>0.04</td>
</tr>
</tbody>
</table>

3) Prepositions
<table>
<thead>
<tr>
<th>Objects</th>
<th>Preposition</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b</td>
<td>near</td>
<td>1</td>
</tr>
<tr>
<td>b, a</td>
<td>near</td>
<td>1</td>
</tr>
<tr>
<td>a, b</td>
<td>against</td>
<td>0.11</td>
</tr>
<tr>
<td>b, a</td>
<td>against</td>
<td>0.04</td>
</tr>
<tr>
<td>a, b</td>
<td>beside</td>
<td>0.24</td>
</tr>
<tr>
<td>b, a</td>
<td>beside</td>
<td>0.17</td>
</tr>
</tbody>
</table>

4) Construct a 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) Sentence
   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.
Results

Given an 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.
Results

- Given an image:

![Image of a person and a dog next to a brown sofa.]

- Generates a description:

  "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."
Results

- Given an image:

  ![Image Description](image)

  "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."
Results

- Given an image:

  ![Image](image-url)

  "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."
Results

Given an image:

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

- Given an image:

![Two buses](image)

- Generates a description:

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

Given an image:

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

▶ Given an image:

▶ Generates a description:

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

<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)

<table>
<thead>
<tr>
<th>Method</th>
<th>k=1</th>
<th>k=2</th>
<th>k=3</th>
<th>k=4+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality of image parsing</td>
<td>2.90</td>
<td>2.78</td>
<td>2.82</td>
<td>3.33</td>
</tr>
<tr>
<td>Language model-based generation</td>
<td>2.27</td>
<td>3.00</td>
<td>2.76</td>
<td>2.95</td>
</tr>
<tr>
<td>Template-based generation</td>
<td>3.83</td>
<td>3.50</td>
<td>3.43</td>
<td>3.61</td>
</tr>
</tbody>
</table>

Human Evaluation: k refers to the number of objects detected by CRF. Possible scores are 4 (perfect without error), 3 (good with some errors), 2 (many errors), 1 (failure)
Bad Results

- Given an image:
Bad Results

- Given an image:

- Generates a description:
  “Here we see one potted plant.”
Bad Results

- Given an image:

![Image of a cat with a collar]

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

- Given an image:

![Image of a cat]

- Generates a description:

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

- Given an image:
Bad Results

▶ Given an image:

▶ Generates a description:

“This is a photograph of one person and one sky. The white person is by the blue sky.”
Questions?