# 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

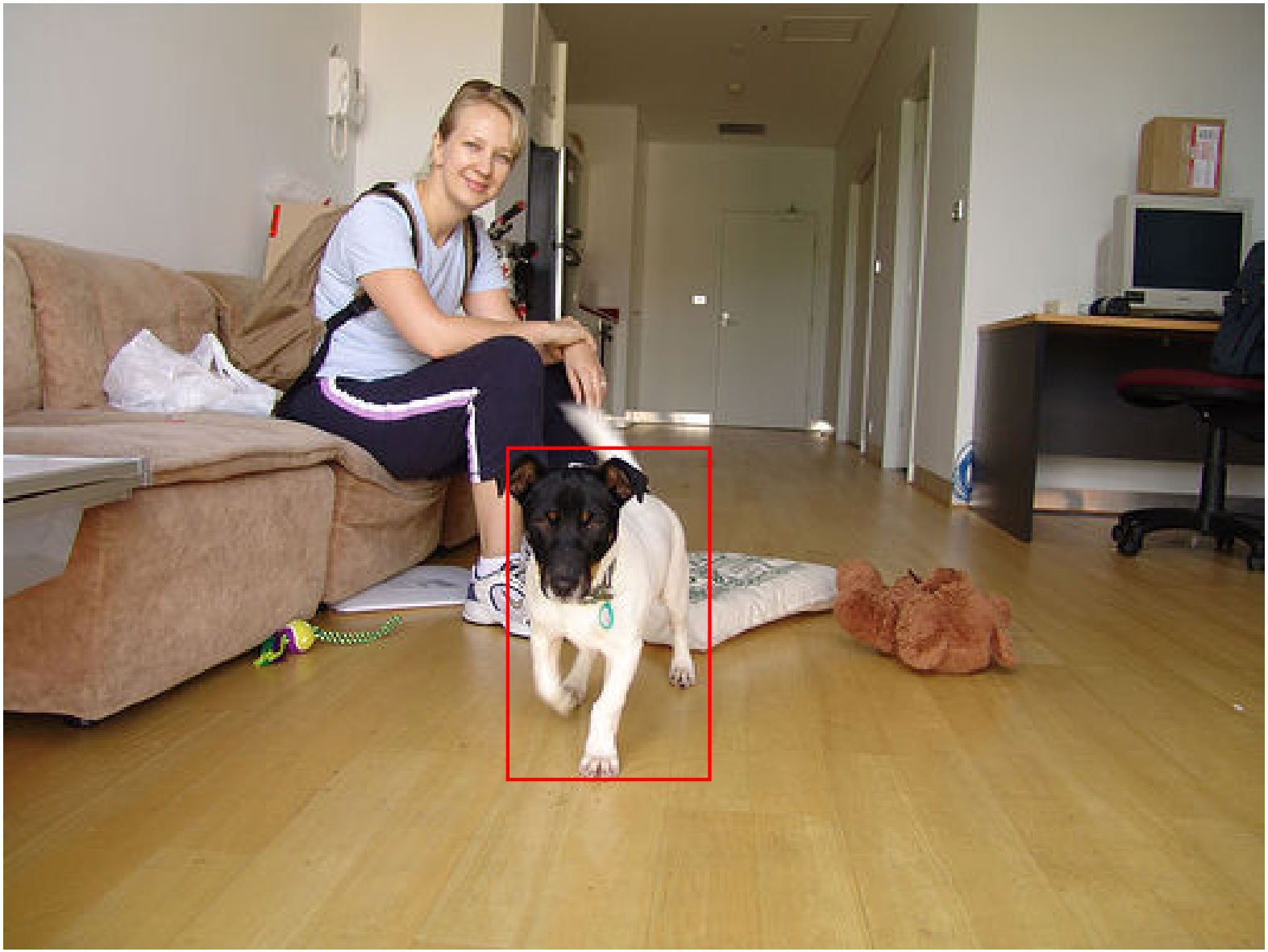
# Natural Language Generation

- Given an image
- Generate english description
- Extremely useful for image indexing and search

# Summarizing Images

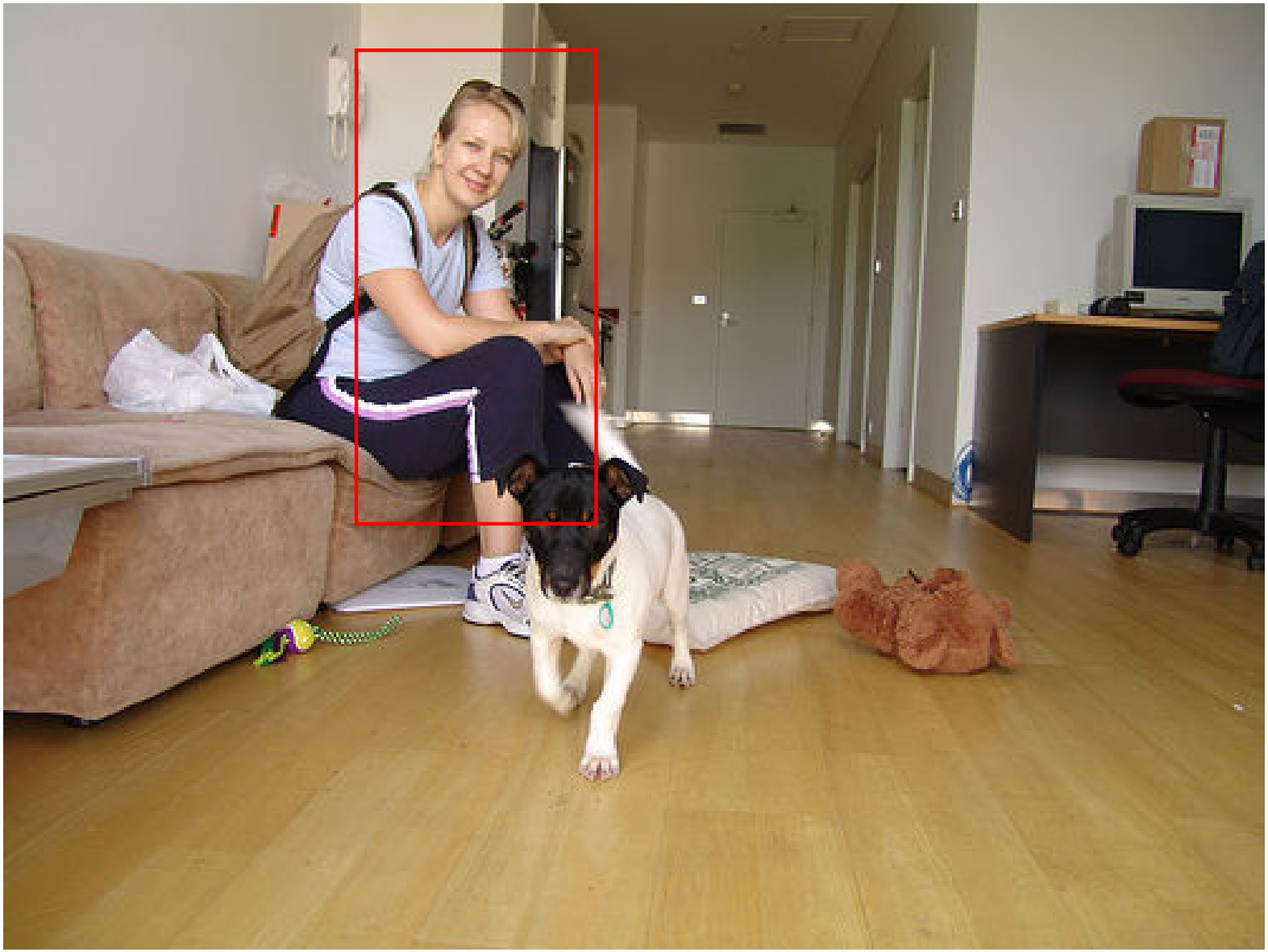
# Summarizing Images

## Object Detection



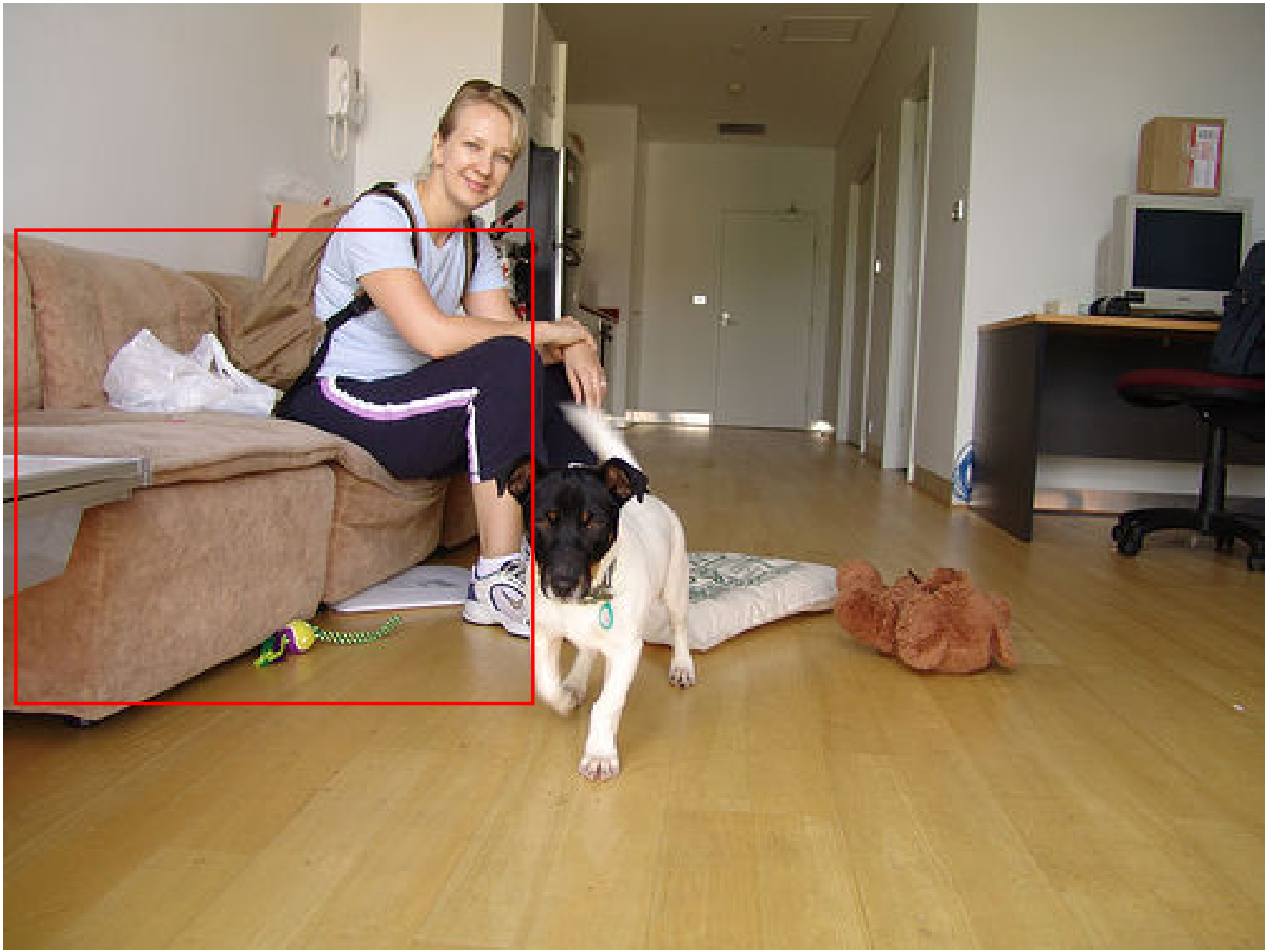
# Summarizing Images

## Object Detection



# Summarizing Images

## Object Detection



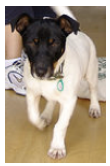
# Summarizing Images

## Object Detection

# Summarizing Images

## Object Detection

1) Objects

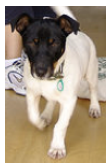


a) dog

# Summarizing Images

## Object Detection

1) Objects



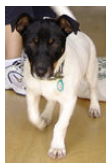
a) dog

b) person

# Summarizing Images

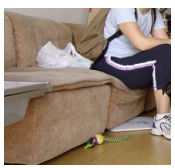
## Object Detection

1) Objects



a) dog

b) person



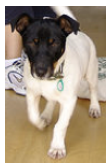
c) sofa

# Summarizing Images

## Attribute Extraction

1) Objects

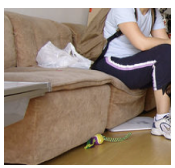
2) Attributes



a) dog

| | |
|---|---|
| brown | 0.01 |
| striped | 0.16 |
| furry | 0.26 |
| wooden | 0.2 |
| feathered | 0.06 |
| ⋮ | ⋮ |

b) person

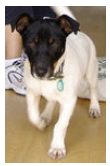


c) sofa

# Summarizing Images

## Attribute Extraction

1) Objects 2) Attributes

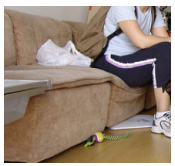


a) dog

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

b) person

| | |
|---|---|
| brown | 0.32 |
| striped | 0.09 |
| furry | .04 |
| wooden | .2 |
| Feathered | .04 |
| . | . |
| . | . |
| . | . |

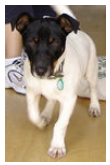


c) sofa

# Summarizing Images

## Attribute Extraction

1) Objects 2) Attributes

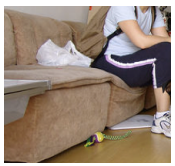


a) dog

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

b) person

| | |
|---|---|
| brown | 0.32 |
| striped | 0.09 |
| furry | .04 |
| wooden | .2 |
| Feathered | .04 |
| . | . |
| . | . |
| . | . |

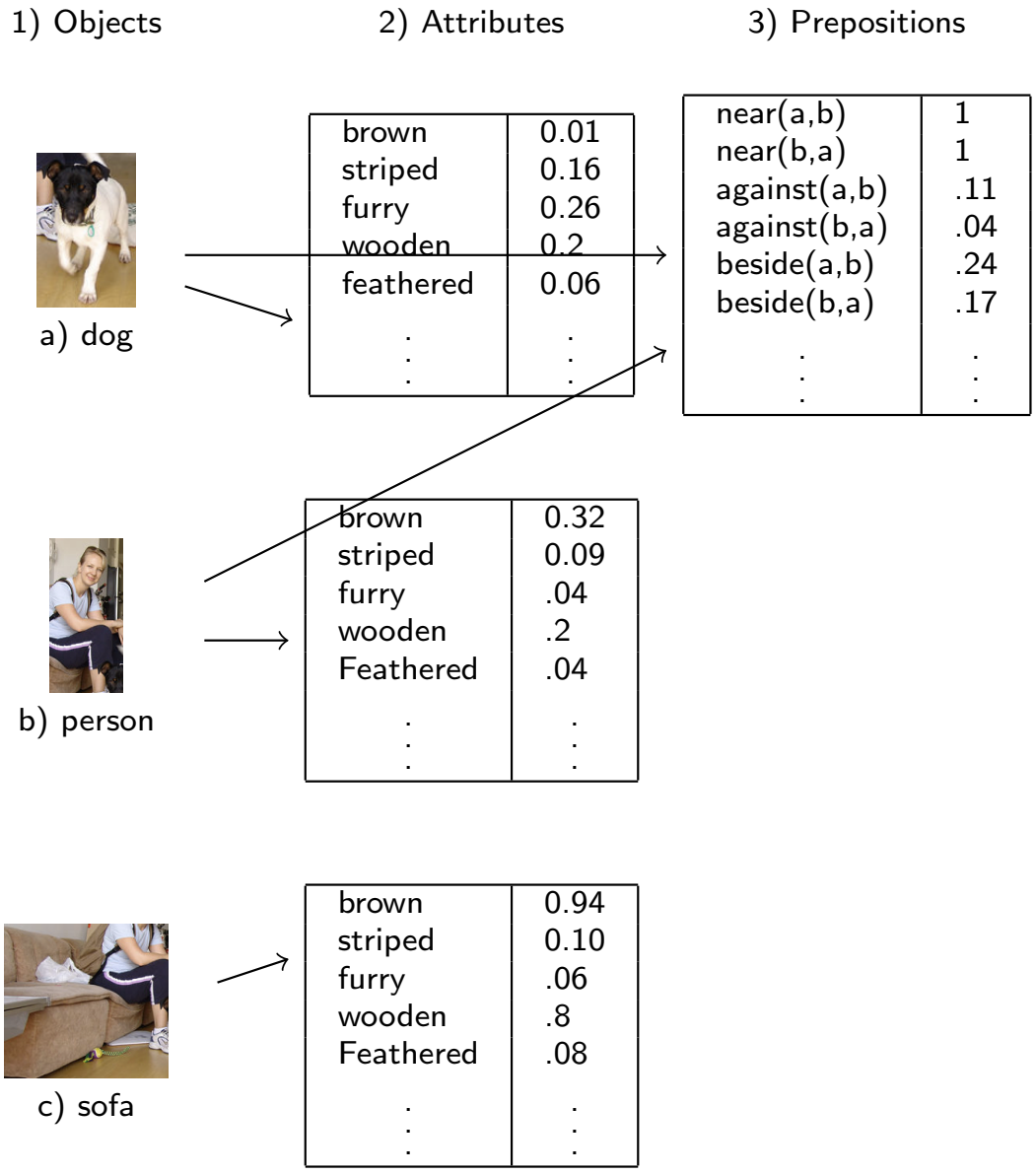


c) sofa

| | |
|---|---|
| brown | 0.94 |
| striped | 0.10 |
| furry | .06 |
| wooden | .8 |
| Feathered | .08 |
| . | . |
| . | . |
| . | . |

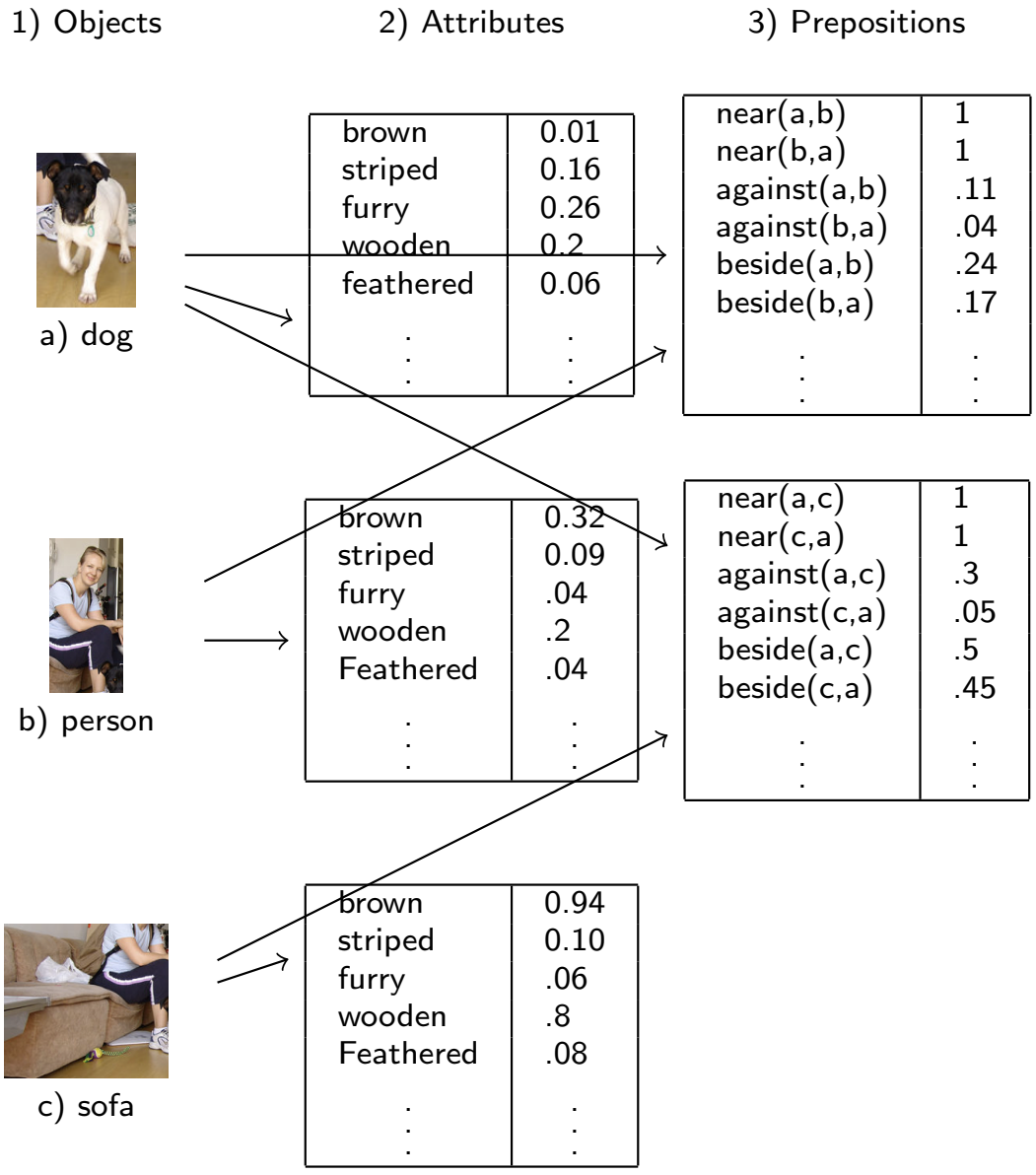
# Summarizing Images

## Relative Positioning



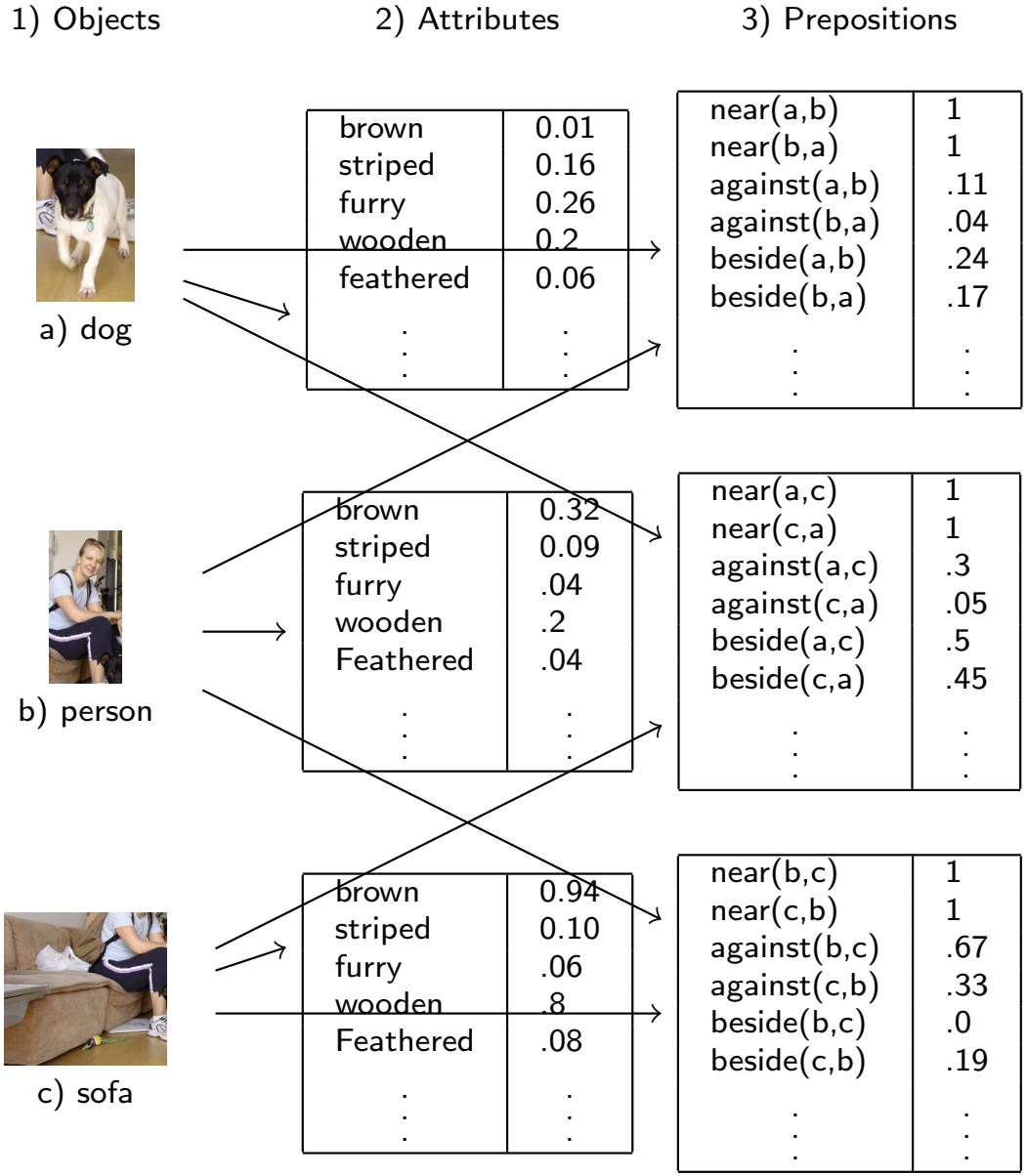
# Summarizing Images

## Relative Positioning



# Summarizing Images

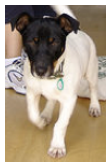
## Relative Positioning



# Summarizing Images

## Sentence generation

1) Objects 2) Attributes 3) Prepositions 4) Construct a CRF



a) dog

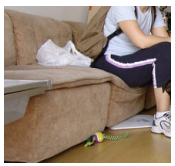
| | |
|---|---|
| brown | 0.01 |
| striped | 0.16 |
| furry | 0.26 |
| wooden | 0.2 |
| feathered | 0.06 |
| ⋮ | ⋮ |

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

b) person

| | |
|---|---|
| brown | 0.32 |
| striped | 0.09 |
| furry | .04 |
| wooden | .2 |
| Feathered | .04 |
| ⋮ | ⋮ |

| | |
|---|---|
| near(a,c) | 1 |
| near(c,a) | 1 |
| against(a,c) | .3 |
| against(c,a) | .05 |
| beside(a,c) | .5 |
| beside(c,a) | .45 |
| ⋮ | ⋮ |



c) sofa

| | |
|---|---|
| brown | 0.94 |
| striped | 0.10 |
| furry | .06 |
| wooden | .8 |
| Feathered | .08 |
| ⋮ | ⋮ |

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

# Summarizing Images

## Sentence generation

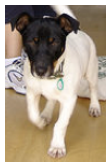
1) Objects

2) Attributes

3) Prepositions

4) Construct a CRF

5) Predicted Labeling



a) dog

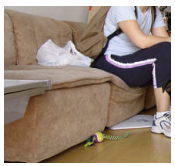
| | |
|---|---|
| brown | 0.01 |
| striped | 0.16 |
| furry | 0.26 |
| wooden | 0.2 |
| feathered | 0.06 |
| ⋮ | ⋮ |

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

b) person

| | |
|---|---|
| brown | 0.32 |
| striped | 0.09 |
| furry | .04 |
| wooden | .2 |
| Feathered | .04 |
| ⋮ | ⋮ |

| | |
|---|---|
| near(a,c) | 1 |
| near(c,a) | 1 |
| against(a,c) | .3 |
| against(c,a) | .05 |
| beside(a,c) | .5 |
| beside(c,a) | .45 |
| ⋮ | ⋮ |



c) sofa

| | |
|---|---|
| brown | 0.94 |
| striped | 0.10 |
| furry | .06 |
| wooden | .8 |
| Feathered | .08 |
| ⋮ | ⋮ |

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

$<< null, person_b >, against, < brown, sofa_c >>$

$<< null, dog_a >, near, < null, person_b >>$

$<< null, dog_a >, beside, < brown, sofa_c >>$

# Summarizing Images

## Sentence generation

1) Objects

2) Attributes

3) Prepositions

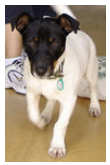
4) Construct a CRF

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.



a) dog

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

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

b) person

| | |
|---|---|
| brown | 0.32 |
| striped | 0.09 |
| furry | .04 |
| wooden | .2 |
| Feathered | .04 |
| . | . |
| . | . |
| . | . |

| | |
|---|---|
| near(a,c) | 1 |
| near(c,a) | 1 |
| against(a,c) | .3 |
| against(c,a) | .05 |
| beside(a,c) | .5 |
| beside(c,a) | .45 |
| . | . |
| . | . |
| . | . |

c) sofa

| | |
|---|---|
| brown | 0.94 |
| striped | 0.10 |
| furry | .06 |
| wooden | .8 |
| Feathered | .08 |
| . | . |
| . | . |
| . | . |

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

# Results

- Given an image:

# Results

- Given an image:

- 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:

# Results

- Given an image:


- Generates a 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."*

# Results

- Given an image:

# Results

- Given an 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:

# Results

- Given an image:


- Generates a description:
  *"Here we see one person and one train. The black person is by the train."*

# Results

| Method | Score |
|---|---|
| Quality of image parsing | 2.85 |
| Language model-based generation | 2.77 |
| Template-based generation | 3.49 |

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

| Method | k=1 | k=2 | k=3 | k=4+ |
|---|---|---|---|---|
| Quality of image parsing | 2.90 | 2.78 | 2.82 | 3.33 |
| Language model-based | 2.27 | 3.00 | 2.76 | 2.95 |
| Template-based generation | 3.83 | 3.50 | 3.43 | 3.61 |

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 pottedplant."*

# Bad Results

- Given an image:

# Bad Results

- Given an image:

- 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?