Baby Talk: Understanding and Generating Image Descriptions

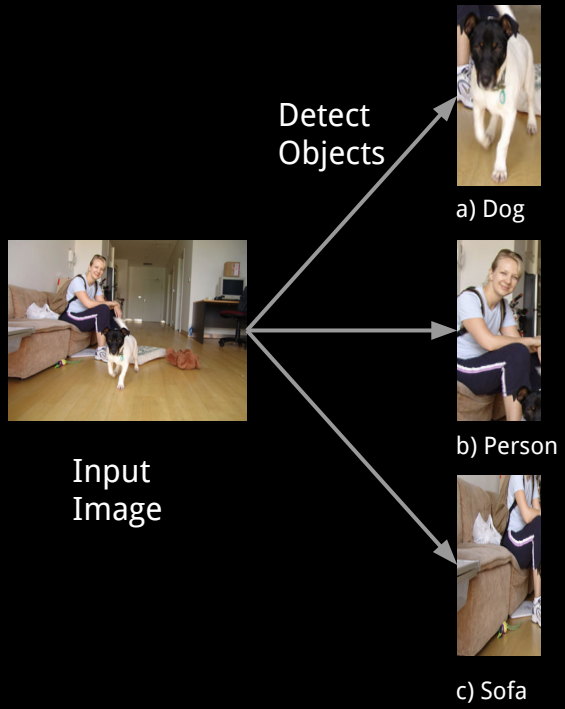
Paper by Kulkarni et al.

Slides by Saheel
(behold my copy-paste skills!)

What?

"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."

How?



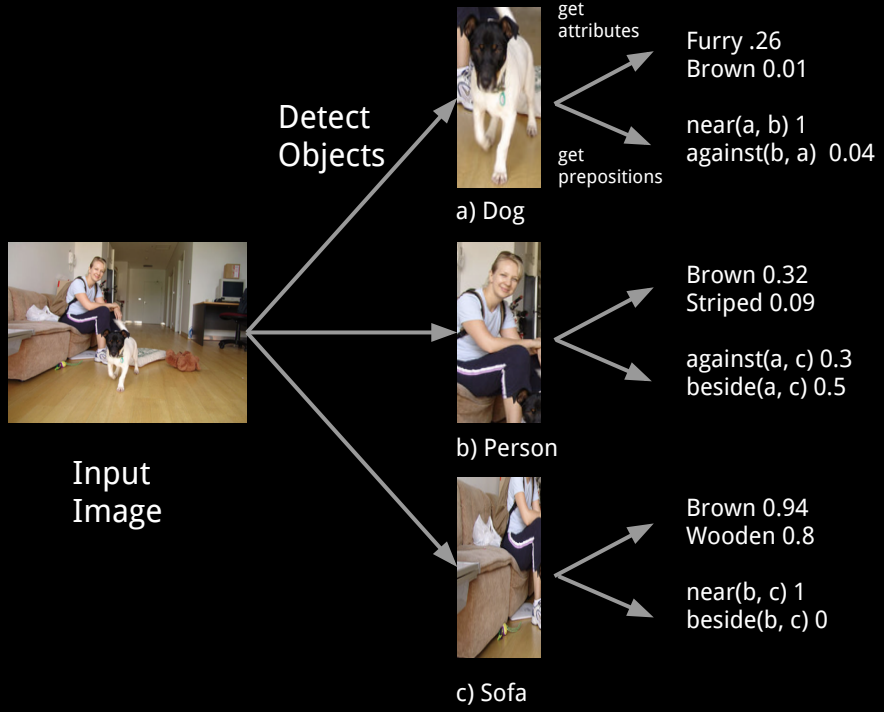
For detecting “things”:

- Object detection system based on mixtures of multi-scale deformable part models (Felzenszwalk et al.)
- 4 additional detectors trained using Imagenet (2009) data

For detecting “stuff”:

- Train linear SVMs on low-level features by Farhadi et al. (2009)

How?



For attributes:

- Find attribute terms commonly used with each object using Flickr descriptions
- For each of 21 such attributes, a classifier is trained using RBF kernel SVM

For prepositions:

- Use spatial relationships to score prepositions like above (a, b)
- Add preposition synonyms to taste

How?

Input Image

Detect Objects

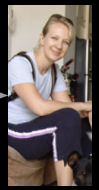
a) Dog

get attributes

Furry .26
Brown 0.01

get prepositions

near(a, b) 1
against(b, a) 0.04



b) Person

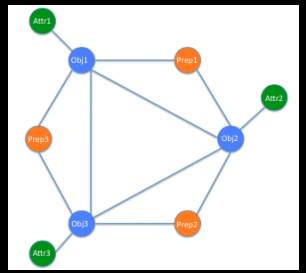
Brown 0.32
Striped 0.09

against(a, c) 0.3
beside(a, c) 0.5

c) Sofa

Brown 0.94
Wooden 0.8

near(b, c) 1
beside(b, c) 0



Construct CRF

How?

Input Image

Detect Objects

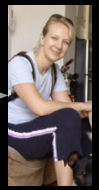
a) Dog

get attributes

Furry .26
Brown 0.01

get prepositions

near(a, b) 1
against(b, a) 0.04



b) Person

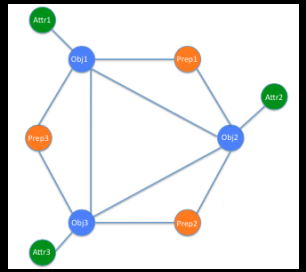
Brown 0.32
Striped 0.09

against(a, c) 0.3
beside(a, c) 0.5

c) Sofa

Brown 0.94
Wooden 0.8

near(b, c) 1
beside(b, c) 0



Construct CRF

Predict labeling

<obj', prep, obj'>

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

Ohhh...

Input Image

Detect Objects

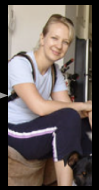
a) Dog

get attributes

Furry .26
Brown 0.01

get prepositions

near(a, b) 1
against(b, a) 0.04



b) Person

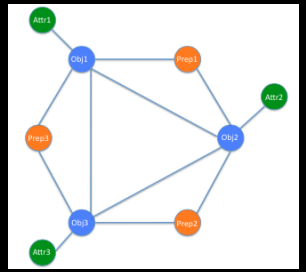
Brown 0.32
Striped 0.09

against(a, c) 0.3
beside(a, c) 0.5

c) Sofa

Brown 0.94
Wooden 0.8

near(b, c) 1
beside(b, c) 0



Construct CRF

Predict labeling

<obj', prep, obj'>

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

Generate 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."

Related Work

Individual words have been associated with image regions

- summarization and retrieval, not generation

Spatial relationships have been used before

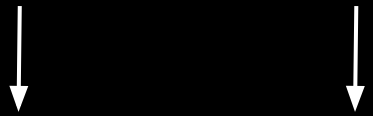
- for labeling, not as outputs by themselves

Closest work by Yao et al. (2010)

- used hierarchical knowledge ontologies
- used human-in-the-loop, not automatic

CRF: Conditional Random Field

- A discriminative graphical model



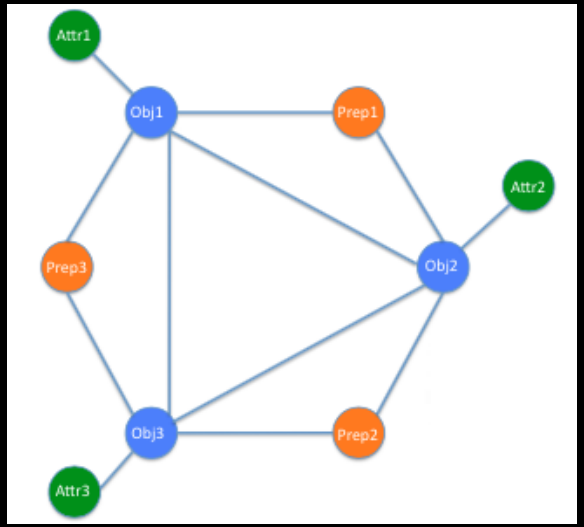
uses conditional dependences

learns structured objects

- Undirected and Probabilistic (duh!)

CRF: Conditional Random Field

- Nodes: objects, attributes, prepositions
- Edges: <obj, attr> pairs and <obj, prep, obj> cliques



What *is* CRF - POS tagging example

Goal: tag the words in a sentence by ADJ, NOUN, PREP, VERB, etc.

"I went fishing for some sea bass." -- noun

"The bass line of the song is too weak." -- adj

Found here: http://blog.echen.me/2012/01/03/introduction-to-conditional-random-fields/

What *is* CRF - POS tagging example

Goal: tag the words in a sentence by ADJ, NOUN, PREP, VERB, etc.

"I went fishing for some sea bass." -- noun

"The bass line of the song is too weak." -- adj

- Conditional dependences help!
  - Given a Noun, next word can be Verb

CRF modeling in POS ex.

- Feature or Potential Function
  - input: sentence, word position, word label, prev-word label
  - output: a real number

CRF modeling in POS ex.

- Feature or Potential Function
  - input: sentence, word position, word label, prev-word label
  - output: a real number

$f_1(s, i, l_i, l_{i-1}) = 1$ if $l_i =$ ADVERB and the ith word ends in "-ly"; 0 otherwise.

$f_3(s, i, l_i, l_{i-1}) = 1$ if $l_{i-1} =$ ADJECTIVE and $l_i =$ NOUN; 0 otherwise.

CRF modeling in POS ex.

weights for the features

$$score(l|s) = \sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s, i, l_i, l_{i-1})$$

$$p(l|s) = \frac{exp[score(l|s)]}{\sum_{l'} exp[score(l'|s)]}$$

Learning the weights

- Gradient Ascent (other approaches are possible, of course)
  - iterative approach
  - Maximize $log(p(l|s))$

Wait, this sounds like HMM -_-

Ummm… yeah, sort-of

Wait, this sounds like HMM -_-

Ummm… yeah, sort-of

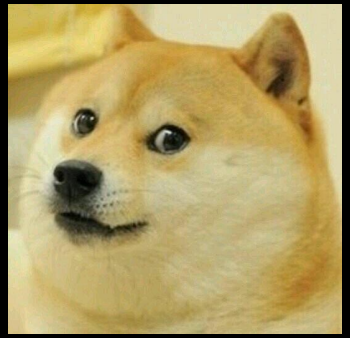
- CRF can have complex features
  - long-distance dependences in POS example
- The weights can be anything

Doge approves

Ummm… yeah, sort-of

- CRF can have complex features
  - long-distance dependences in POS example
- The weights can be anything

Cunning CRF. Such Power. Much Wow.



Back to the paper

We have

- 3 image-based features (scores of image-based detectors)

$$\psi(prep_{ij}; prepFuns)$$ $$\psi(obj_i; objDet)$$ $$\psi(attr_i; attrCl)$$

Back to the paper

And

- 2 text-based features

(which are just counts found in the image descriptions text data)

$$\psi(attr_i, obj_i; textPr)$$

$$\psi(obj_i, prep_{ij}, obj_j; textPr)$$

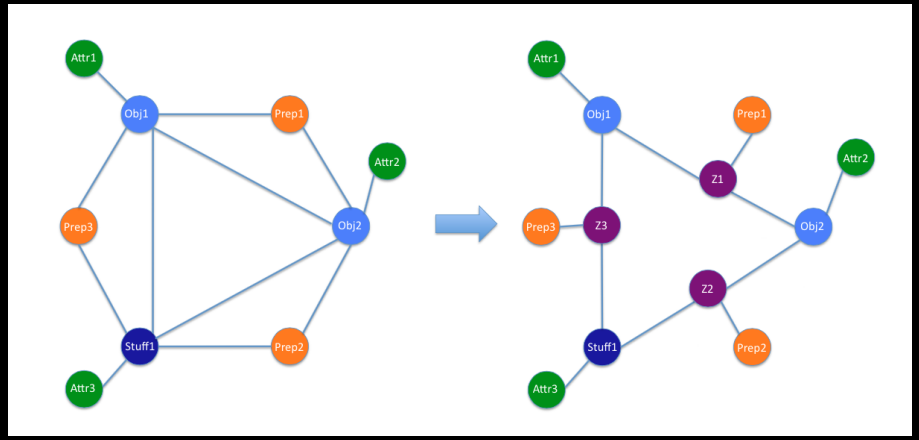
Weighting the features

$$
\begin{aligned}
F_i &= \alpha_0\beta_0\psi(obj_i; objDet) + \alpha_0\beta_1\psi(attr_i; attrCl) \\
&\quad +\alpha_1\gamma_0\psi(attr_i, obj_i; textPr) \\
G_{ij} &= \alpha_0\beta_2\psi(prep_{ij}; prepFuns) \\
&\quad +\alpha_1\gamma_1\psi(obj_i, prep_{ij}, obj_j; textPr)
\end{aligned}
$$

$$
E(L; I, T) = -\sum_{i \in objs} F_i - \frac{2}{N-1} \sum_{ij \in objPairs} G_{ij}
$$

Some implementation details

- Feature transformation
- Score Normalization



$$\frac{obj_{t-f}}{N} + \frac{(mod, obj)_{t-f}}{N} + \frac{2}{N-1}\frac{(obj, prep, obj)_{t-f}}{N}$$

Learning the weights

- Factored learning
  - factored = hierarchical approach
  - fix one parameter, use grid* search to learn others
  - now fix the learned ones and recurse

* ‘grid’ is fancy name for ‘exhaustive’

Pipeline review

Input Image

Detect Objects

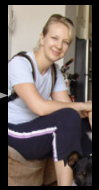
a) Dog

get attributes

Furry .26
Brown 0.01

get prepositions

near(a, b) 1
against(b, a) 0.04



b) Person

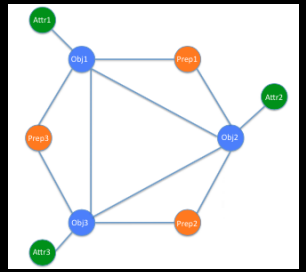
Brown 0.32
Striped 0.09

against(a, c) 0.3
beside(a, c) 0.5

c) Sofa

Brown 0.94
Wooden 0.8

near(b, c) 1
beside(b, c) 0



Construct CRF

Predict labeling

<obj', prep, obj'>

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

Generate 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."

CRF Labeling and Sentence Generation

- UIUC PASCAL sentence dataset
- TRW-S algorithm (I am not gonna explain everything, okay!)
  - to predict the labeling for each test image CRF

- N-gram model
  - to generate sentence from the labeling
  - crawled Wikipedia pages to learn N-gram model

CRF Labeling and Sentence Generation

- Template-based generation
  - beautifying the N-gram-based sentences (How?!)

Evaluation

- Automatic evaluation
  - BLEU metric (compares machine-generated sentences with human-generated ones)
- Human evaluation
  - humans judged the quality of image descriptions
  - does not correlate with BLEU

Tables, yay!

Method	w/o	w/ synonym
Human	0.50	0.51
Language model-based generation	0.25	0.30
Template-based generation	0.15	0.18
Meaning representation (triples)	0.20	0.30

Table 1. Automatic Evaluation: BLEU score measured at 1

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

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

To err is ~~human~~ funny

Incorrect detections:

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

Just all wrong!

This is a photograph of one person and one sky. The white person is by the blue sky.

Comments and such

Conclusion:

- We don’t need to worry about Matrix

Extensions:

- Object priority?
- Action and scene detection?
- More natural sentences?

No questions, right?

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.

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 person and one train. The black person is by the train.

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.

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.

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.

Figure 4. Results of sentence generation using our method with template based sentence generation. These are "*good*" results as judged by human annotators.