Baby Talk: Understanding and Generating Image Descriptions

Paper by Kulkarni et al.

Slides by Saheel

(behind 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 (Felzenszwalb 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?

Detect Objects

a) Dog
- Furry: 0.26
- Brown: 0.01
- 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): 0.0

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?

Detect Objects

get attributes
Furry .26
Brown 0.01
near(a, b) 1
against(b, a) 0.04

get prepositions
Brown 0.32
Striped 0.09
against(a, c) 0.3
beside(a, c) 0.5

Input Image

Construct CRF

a) Dog
Furry .26
Brown 0.32
Striped 0.09
against(a, c) 0.3
beside(a, c) 0.5

b) Person
near(a, b) 1
against(b, a) 0.04
Brown 0.94
Wooden 0.8
near(b, c) 1
beside(b, c) 0

b) Person

near(a, b) 1
against(b, a) 0.04
Brown 0.94
Wooden 0.8
near(b, c) 1
beside(b, c) 0

C) Sofa

How?
How?

Detect Objects

- a) Dog
- b) Person
- c) Sofa

get attributes
- Furry 0.26
- Brown 0.01
- near(a, b) 1
- against(b, a) 0.04
- Brown 0.32
- Striped 0.09
- against(a, c) 0.3
- beside(a, c) 0.5
- Brown 0.94
- Wooden 0.8
- near(b, c) 1
- beside(b, c) 0.0

get prepositions

Predict labeling

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

Construct 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.”
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

\[
f_1(s, i, l_i, l_{i-1}) = 1 \text{ if } l_i = \text{ADVERB and the ith word ends in “-ly”}; 0 \text{ otherwise.}
\]

\[
f_3(s, i, l_i, l_{i-1}) = 1 \text{ if } l_{i-1} = \text{ADJECTIVE and } l_i = \text{NOUN}; 0 \text{ otherwise.}
\]
CRF modeling in POS ex.

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

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

weights for the features
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.
We have
- 3 image-based features (scores of image-based detectors)

\[ \psi(prep_{ij}; prepFuns) \quad \psi(obj_i; objDet) \quad \psi(attr_i; attrCl) \]
Back to the paper

And

- 2 text-based features

\[ \psi(\text{attr}_i, \text{obj}_i; \text{textPr}) \]

\[ \psi(\text{obj}_i, \text{prep}_{ij}, \text{obj}_j; \text{textPr}) \]

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

\[ F_i = \alpha_0 \beta_0 \psi(obj_i; objDet) + \alpha_0 \beta_1 \psi(attr_i; attrCl) \]
\[ + \alpha_1 \gamma_0 \psi(attr_i, obj_i; textPr) \]

\[ G_{ij} = \alpha_0 \beta_2 \psi(prepi_j; prepFuns) \]
\[ + \alpha_1 \gamma_1 \psi(obj_i, prepi_j, obj_j; textPr) \]

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

Detect Objects

- a) Dog
  - Furry: 0.26
  - Brown: 0.01
  - 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

Get attributes

Get prepositions

Construct CRF

Predict labeling

<obj', prep, obj'>

Generate sentence

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

“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 \(\text{ (I am not gonna explain everything, okay!)}\)
  - 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!

<table>
<thead>
<tr>
<th>Method</th>
<th>w/o</th>
<th>w/ synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.50</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>

Table 1. Automatic Evaluation: BLEU score measured at 1

<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>

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?

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