Baby Talk: Understanding and Generating Image Descriptions

- Kulkarni et. al.

Prateek Kolhar
Objective

Given an image generate a description that is descriptive of the objects present, their attributes and mutual spatial relationships.

“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.”
Model Objective

Given an image and textual prior produce list of triplets in the format:

\[\langle \langle \text{modifier}_1, \text{obj}_1 \rangle, \text{preposition}, \langle \text{modifier}_2, \text{obj}_2 \rangle \rangle\]
System Flow

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

2) Attributes
   - brown 0.01
   - striped 0.16
   - furry 0.28
   - wooden 0.2
   - feathered 0.06

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

3) Prepositions
   - near(a,c) 1
   - against(a,c) 0.8
   - against(c,a) 0.05
   - beside(a,c) 0.5
   - beside(c,a) 0.45

4) Constructed CRF

5) Predicted Labeling
   - null_person_b
   - against(brown_sofa_c)
   - near(null_person_b)
   - 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

1. Object Detection
2. Attribute classification
3. Prepositional relationship classification
4. A CRF to combine 1-3 potentials and text based potentials
5. A labeling of the graph is predicted
6. Sentences are generated based on the labeling
Detection

For detecting “things”:
- Object detection system based on mixtures of multiscale 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)
Attribute Classification

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
Prepositional Relations

For prepositions:
- Percentage based handcrafted spatial features are used.
- Use spatial relationships to score prepositions like “above (a, b)”
CRF

- Discriminative undirected and probabilistic graphical model
- Nodes can be divided into exactly two disjoint sets $X$ and $Y$ observed variables and output variables, respectively
- The conditional distribution $P(Y|X)$ is modeled.

$$p(Y_v \mid X, Y_w, w \neq v) = p(Y_v \mid X, Y_w, w \sim v)$$

where $w \sim v$ means that $w$ and $v$ are neighbors in $G$. 
CRF

- Compute “potential functions” or “features”
- The product of the potential functions is a representation of the Conditional Probability
- Training is done using sequential tree re-weighted message passing algorithm (TRW -S) - similar to forward backward algorithm.

- Nodes: Objects, Attributes and Prepositions
- Edges: <obj, attr> pairs and <obj_1, prep, obj_2> cliques
Unary Potential Functions

\( \psi(\text{obj}_i; \text{obj Det}) \) - Object and Stuff Potential

Score from the detector models are used

\( \psi(\text{attr}_i; \text{attr Cl}) \) - Attribute Potential

RBF kernel SVM classifier Score used

\( \psi(\text{prep}_{ij}; \text{prep Funcs}) \) - Preposition Potential

SVM scores trained on hand crafted spatial feature based on percentages are used.
Higher Order Potential Functions

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

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

Calculated based on counts of occurrences in a large corpus of data.
Learning

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

\[ F_i = \alpha_0 \beta_0 \psi(\text{obj}_i; \text{objDet}) + \alpha_0 \beta_1 \psi(\text{attr}_i; \text{attrCl}) \]
\[ + \alpha_1 \gamma_0 \psi(\text{attr}_i, \text{obj}_i; \text{textPr}) \]

\[ G_{ij} = \alpha_0 \beta_2 \psi(\text{prep}_{ij}; \text{prepFuns}) \]
\[ + \alpha_1 \gamma_1 \psi(\text{obj}_i, \text{prep}_{ij}, \text{obj}_j; \text{textPr}) \]
Implementation Details

- Feature transformation
- TRW-S algorithm to train CRF
- Factored Hyperparameter tuning
  - hierarchical approach
  - fix one param and grid search for others
  - fix learned ones and recurse
- Normalized score for tuning

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

- N-gram based
  - Generates sentences by looking at sequence probabilities of triple words.
  - can use DP to optimize
  - hard to guarantee coherency
- Template based
  - much simpler to enforce grammar and coherency
System Flow

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

2) Attributes
   - brown 0.01
   - striped 0.16
   - furry 0.25
   - wooden 0.2
   - feathered 0.06

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

3) Prepositions
   - brown 0.92
   - striped 0.09
   - furry 0.04
   - wooden 0.2
   - feathered 0.04

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.
Experimental Setup

- Language model training: wikipedia crawled data, google data
- Image detector training: Flickr
- CRF training: authors built the training set
- Evaluation dataset: UIUC Pascal dataset
- BLEU and Human rating evaluation
Results

<table>
<thead>
<tr>
<th>Method</th>
<th>w/o</th>
<th>w/ syno</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

<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)

<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</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>

Table 3. 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)
Conclusion

- Demonstrated an effective, fully automatic, system that generates natural language descriptions for images.
- Better results than previous automated methods.
- Human evaluation validates the quality.
- Automatically mining and parsing large text collections to obtain statistical models for visually descriptive language.
- The other is taking advantage of state of the art vision systems and combining all of these in a CRF to produce input for language generation methods.
This is a photograph of two buses. The first rectangular bus is near the second rectangular bus.

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

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

This is a picture of one sky, one road and one bicycle. 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 aeroplane. 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.

Here we see one road, one sky and one aeroplane. The first shiny aeroplane is near the second shiny aeroplane.

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

There is 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 window.

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

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.

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 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.

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.

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

- Model is limited to specific hand crafted features. Doesn’t allow for fuzzy/nebulus features
- Doesn’t model anything beyond spacial relationships - action, scene detection, etc.
- The sentences don’t all seem natural
- Object priority not considered
- Hard to scale to new features without redesigning the graph
- They could have added metrics to intermediate steps for root cause analysis.
Future Work

- This has become a rather well studied problem now
- We could use more modern techniques in Deep Learning to capture more nuanced features of the image
- We could use better language models to produce coherent and grammatically correct sentences - neural CRFs
- Build end-to-end models that encode the image to a vector and decode it to a sentence by additionally using attentions over parts of the image.
Thank You!