Every Picture Tells a Story: Generating Sentences from Images

-Slides Credited Yukun Zhu
Goal

Auto-annotation: find text annotations for images

- This is a lot of technology.
- Somebodys screensaver of a pumpkin
- A black laptop is connected to a black Dell monitor
- This is a dual monitor setup
- Old school Computer monitor with way to many stickers on it
Goal

Auto-illustration: find pictures suggested by given text

Yellow train on the tracks.
• Map from Image Space to Meaning Space
• Map from Sentence Space to Meaning Space
• Retrieve Sentences for Images via Meaning Space

Farhadi et al, ECCV10
Retrieval through meaning space

- Map from Image Space to Meaning Space
- Map from Sentence Space to Meaning Space
- Retrieve Sentences for Images via Meaning Space

Farhadi et al, ECCV10
Image Space $\rightarrow$ Meaning Space

Predict Image Content using trained classifiers

Farhadi et al, ECCV10
Retrieval through meaning space

- Map from Image Space to Meaning Space
- Map from Sentence Space to Meaning Space
- Retrieve Sentences for Images via Meaning Space

Farhadi et al, ECCV10
Sentence Space → Meaning Space

• Extract **subject**, **verb** and **scene** from sentences in the training data

  black *cat* over pink chair
  A black color *cat sitting* on chair *in a room.*
  *cat sitting* on a chair looking in a mirror.

  Subject: Cat
  Verb: Sitting
  Scene: room

• Use taxonomy trees

  ![Taxonomy Tree Image]

  - Object
    - Animal
      - Cat
      - Dog
      - Horse
    - Human
    - Vehicle
      - Car
      - Bike
      - Train

Farhadi et al, ECCV10
Retrieval through meaning space

- Map from Image Space to Meaning Space
- Map from Sentence Space to Meaning Space
- Retrieve Sentences for Images via Meaning Space

Farhadi et al, ECCV10
Results

- Sentence Generation

see something unexpected.
Cow in the grass field.
Beautiful scenery surrounds a fluffy sheep.
Dog hearding sheep in open terrain.
Cattle feeding at a trough.

Farhadi et al, ECCV10
Results

- Sentence Generation

Refrigerator almost empty.
Foods and utensils.
Eatables in the refrigerator.
The inside of a refrigerator apples,
cottage cheese, tupperwares and lunch bags.
Squash apenny white store with a hand statue,
picnic tables in front of the building.

Farhadi et al, ECCV10
Mistakes

The two girls read to drive big bullet.

Black goat in cage.

Farhadi et al, ECCV10
Data

1,000 images
Rashtchian et al 2010, Farhadi et al 2010
5 descriptions per image
20 object categories

20,000 images
Image-Clef challenge
2 descriptions per image
Select image categories

More data needed?

Large amounts of paired data can help us study the image-language relationship

Berg, Attributes Tutorial CVPR13
Baby Talk: Understanding and Generating Image Descriptions

-Presented by Yingjie Tang
More Nuance than Traditional Recognition…

person

shoe

car

Berg, Attributes Tutorial CVPR13

credit: Tamara
Toward Complex Structured Outputs

Berg, Attributes Tutorial CVPR13

car

credit: Tamara
Toward Complex Structured Outputs

Attributes of objects

Berg, Attributes Tutorial CVPR13

credit: Tamara
Toward Complex Structured Outputs

Relationships between objects
Toward Complex Structured Outputs

Telling the “story of an image”

Little pink smart car parked on the side of a road in a London shopping district.

… Complex structured recognition outputs

Berg, Attributes Tutorial CVPR13
Problem: Generate Natural Language Descriptions for Images

"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."
Problem: Generate Natural Language Descriptions for Images

Descriptive Language:

✓ More information about the visual world
✓ Convey the style how people describe world
Learning from Descriptive Text

“It was an arresting face, pointed of chin, square of jaw. Her eyes were pale green without a touch of hazel, starred with bristly black lashes and slightly tilted at the ends. Above them, her thick black brows slanted upward, cutting a startling oblique line in her magnolia-white skin—that skin so prized by Southern women and so carefully guarded with bonnets, veils and mittens against hot Georgia suns”

Scarlett O’Hara described in Gone with the Wind.

Visually descriptive language provides:
• Information about the world, especially the visual world.
• Information about how people construct natural language for imagery.
• Guidance for visual recognition.

How does the world work?

What should we recognize?

How do people describe the world?

Berg, Attributes Tutorial CVPR13
Problem: Generate Natural Language Descriptions for Images

statistics gleaned from parsing large quantities of text data + recognition algorithms from computer vision
Brief Introduction:

Generating Sentences for images

- Scene
- Objects
- Relative locations
- Detectors
- N-gram
- Mining for statistic models

Generating Sentences
Brief Introduction:

Generating Sentences for images

Most previous work in NLP on automatically generating captions or descriptions for images is based on retrieval and summarization.
Brief Introduction:

Generating Sentences for images

Scene Based generated sentences are not as descriptive enough!
Brief Introduction:

Generating Sentences for images

✓ Small number of instances form large number of scenes

✓ Avoid whole image features recognition and make tight connection between image content and sentence generation.
Related works:

- Individual words with image regions, use of spatial relationships between labeled parts of image.

- Use the attributes in computer vision to estimate modifiers for objects in images.

- Use human loop for hierarchical image parsing
Method Overview:

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

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

3) Prepositions
   - near(a, b) 0.1
   - against(a, b) 0.11
   - against(b, a) 0.04
   - beside(a, b) 0.24
   - beside(b, a) 0.17

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.
Conditional Random Field:

Conditional random fields, a framework for building probabilistic models to segment and label sequence data. Conditional random fields offer several advantages over hidden Markov models and stochastic grammars for such tasks, including the ability to relax strong independence assumptions made in those models.

— Lafferty, J., McCallum, A., Pereira, F.
CRF Labeling:

Figure 3. CRF for an example image with 2 object detections and 1 stuff detection. Left shows original CRF with trinary potentials. Right shows CRF reduced to pairwise potentials by introducing z variables with domains covering all possible triples of the original 3-clique.
CRF Labeling:

Nodes of the CRF:
✓ Objects
✓ Attributes
✓ Prepositions
CRF Labeling:

Nodes of the CRF:
✓ Objects

- A large set of detectors collect a set of high-score detections.
- Merge detections that are highly overlapping into groups.
- Create an object node for each group.
CRF Labeling: —Independent Domain

- Objects nodes: Set of object detectors that fired at that region in the image.
- Attribute nodes: A set of appearance attributes that can modify the objects.
- Preposition nodes: A set of prepositional relations that can occur between two objects.
CRF Labeling: —Energy Function

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

\[
F_i = \alpha_0 \beta_0 \psi(\text{obj}_i; \text{objDet}) + \alpha_0 \beta_1 \psi(\text{attr}_i; \text{attrCl}) \quad (2)
\]
\[
+ \alpha_1 \gamma_0 \psi(\text{attr}_i, \text{obj}_i; \text{textPr}) \quad (3)
\]

\[
G_{ij} = \alpha_0 \beta_2 \psi(\text{prep}_{ij}; \text{prepFuns}) \quad (4)
\]
\[
+ \alpha_1 \gamma_1 \psi(\text{obj}_i, \text{prep}_{ij}, \text{obj}_j; \text{textPr}) \quad (5)
\]
CRF Labeling: —Converting to Pairwise Potentials

\[
\psi(obj_i, prep_{ij}, obj_j; textPr) \tag{6}
\]

\[
\psi(z_{ij}, obj_i) = \begin{cases} 
0 & \text{if } Z_{ij}(1) = O_i \\
-\infty & \text{otherwise}
\end{cases} \tag{7}
\]

\[
\psi(z_{ij}, prep_{ij}) = \begin{cases} 
0 & \text{if } Z_{ij}(2) = P_{ij} \\
-\infty & \text{otherwise}
\end{cases} \tag{8}
\]

\[
\psi(z_{ij}, obj_j) = \begin{cases} 
0 & \text{if } Z_{ij}(3) = O_j \\
-\infty & \text{otherwise}
\end{cases} \tag{9}
\]
CRF Learning: —Scoring

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

Number of Objects

a) the number of true obj labels minus the number of false obj labels normalized by the number of objects

b) the number of true mod-obj label pairs minus the number of false mod-obj pairs

c) the number of true obj-prep-obj triples minus the number of false obj-prep-obj triples normalized by the number of nodes and the number of pairs of objects (N choose 2).
Potential Functions:

Image Based Potentials:
the image potentials come from hand designed detection strategies optimized on external training sets

Text Potential:
the text potentials are based on text statistics collected automatically from various corpora
Image Based Potentials:

ψ(obji; objDet) – Object and Stuff Potential
for Object Detectors:
20 PASCAL 2010 object categories detectors + trained 4 additional non-PASCAL object categories for flower, laptop, tiger, and window.

for Stuff Detectors:
Trained linear SVMs on the low level region features to recognize: sky, road, building, tree, water, and grass stuff categories. SVM outputs are mapped to probabilities.
Image Based Potentials:

\[ \psi(\text{attr}; \text{attrCl}) - \text{Attribute Potential:} \]
Train visual attribute classifiers that are relevant for our object (and stuff) categories.

Mine large text corpus of Flickr descriptions (described in Sec. 5.2) to find attribute terms.

The resulting list consists of 21 visual attribute terms describing color (e.g. blue, gray),
texture (e.g. striped, furry), material (e.g. wooden, feathered), general appearance (e.g. rusty, dirty, shiny), and shape (e.g. rectangular) characteristics.
Image Based Potentials:

$\psi(\text{prepij} ; \text{prepF uns})$ - Preposition Potential

16 preposition terms
Text Based Potentials:

Two potential functions calculated from large corpora.

- pairwise potential on attribute-object label pairs $\psi(\text{attri, obji}; \text{textP r})$
- a trinary potential on object-preposition-object triples $\psi(\text{obji, prepij, objj}; \text{textP r})$

These potentials are the probability of various attributes for each object (given the object) and the probabilities of particular prepositional relationships between object pairs (given the pair of objects).
Text Based Potentials:

Parsing Potential:

➤ Collect a large set of Flickr image descriptions to count object potential \( \psi_p(\text{attri, obji; textP r}) \).

➤ Collect statistics about the occurrence of each attribute and object pair to count \( \text{amod} (\text{attribute, object}) \).

➤ Collect \( \sim 1.4 \text{ million} \) Flickr image descriptions by querying for pairs of object terms for \( \psi_p(\text{obji, prepij, objj; textP r}) \).
Google Potentials:

Reasons:
The counts for some objects can be too sparse.

Collect additional Google Search based potentials: \( \psi_g(\text{attri, obji}; \text{textPr}) \) and \( \psi_g(\text{obji, prepij, objj}; \text{textPr}) \).

Smooth Potential:

Final potentials are computed as a smoothed combination of the parsing based potentials with the Google potentials: \( \alpha \psi_p + (1 - \alpha) \psi_g \).
Generation:

Output of CRF(Triples) \[\langle\langle \text{white, cloud} \rangle, \text{in}, \langle \text{blue, sky} \rangle \rangle\] \[\rightarrow\]  

Our goal \[\rightarrow\] \text{“There is a white cloud in the blue sky”}

Language Models and Templates
Generation:

N-gram Model:

The prediction of the next word depends only on the previous N-1 words.

We want to determine whether to insert a function word $x$ between a pair of words $\alpha$ and $\beta$ in the meaning representation.

Calculating $p(\alpha x \beta) = p(\alpha)p(x|\alpha)p(\beta|x)$ using bigram (2-gram) language models

Weakness:
(1) it is difficult to enforce grammatically correct sentences using language models alone
(2) it is ignorant of discourse structure (coherency among sentences), as each sentence is generated independently.
Generation:

Templates with Linguistic Constraints:

Constructing templates with linguistically motivated constraints.

This approach is based on the assumption that there are a handful of salient syntactic patterns in descriptive language that we can encode as templates.
Generation:

Templates with Linguistic Constraints:
Generation:

Templates with Linguistic Constraints:
Some good results

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.

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

Kulkarni et al, CVPR11
Some bad results

Missed detections:
Here we see one potted plant.

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

Incorrect attributes:
This is a photograph of two sheeps 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.

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.

Kulkarni et al, CVPR11

credit: Tamara
Experiment Results:

Training sets & Test sets:
Training set:
Crawled Wikipedia pages that describe objects our system can recognize to construct the training corpus for language models.

Test set:
Use the UIUC PAS-CAL sentence dataset2, which contains up to five human-generated sentences that describe 1000 images.
Experiment Results:

Automatic Evaluation:

BLEU:
a widely used metric for automatic evaluation of machine translation that measures the n-gram precision of machine generated sentences with respect to human generated sentences.

Weakness:
BLEU will inevitably penalize many correctly generated sentences.

<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
Experiment Results:

Human Evaluation:
Perform human judgment on the entire test set to directly quantify these aspects.

<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 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>

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)

Overall THE template generation method demonstrates a very high average human evaluation score of 3.49 (max 4) for the quality of generated sentences.
Conclusions:

- An effective, fully automatic, system that generates natural language descriptions for images.

- Produce results much more specific to the image content than previous automated methods.

- Human evaluation validates the quality of the generated sentences.

Keys to success:
- Automatically mining and parsing large text collections.

- Taking advantage of state of the art vision systems and combining all of these in a CRF to produce input for language generation methods.
My view:

1. The descriptions of a sentence for the image is always biased.

“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.”

“The boy is happy to standing in front of the shop with his snacks ”

How can we find an unbiased way to describe the image?
My view:

2. Scene based description is in a high level description while the thing & stuff based description is the low level description.

Can we reverse back to the scene based description from the things & stuff information?