Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework

Paper by
L. Li, R. Socher and L. Fei-Fei
Published in CVPR 2009

Presented by Eric Wang
May 8, 2009
Outline

• Introduction and motivation
• Generative model
• Automatic initialization scheme
• Putting it all together: Annotated imagery segmentation
• Experiments and Results
• Summary and Conclusion
Introduction and Motivation

• When humans observe a scene, three simultaneous actions happen
  – High level recognition (classification)
  – Identification of specific items in the scene (annotation)
  – Localization of scene components (segmentation)
Introduction and Motivation

• Current methods
  – Provide a single label to an image.
  – Provide multiple labels to an image without localization.
  – Separate imagery between background clutter and foreground objects.

• The proposed model attempts to capture the simultaneous occurrence of multiple objects and high level scene classes.

• The authors claim that this results in more accurate semantic representations of the observed images.
The authors also mention two further advantages:

- Automatic and highly scalable learning from tagged datasets such as Flickr.
- Ability to account for annotation noise due to miss-annotation or missing annotations.
- Model can be fitted into an automatic learning framework, in which no manual tagging of specific image regions is necessary for training.
Generative Model

Set of image patches (characterized by extracted SIFT features), quantized to 500 codewords.

Observed Features

- **R**: Set of region wide features (shape, color, location, texture), 100, 30, 50, 120 codewords, respectively.
- **X**: Set of image patches (characterized by extracted SIFT features), quantized to 500 codewords.
- **T**: Noisy tags
Generative Process

- Step 0: Oversegment each of the D images into Nr regions, with Nt associated words.
Generative Process

- Step 1: Draw an image class indicator $C$ from a fixed uniform prior.
• Step 2: Based on the chosen class, draw an object \( O \) from the class specific distribution over objects.
Step 3: Having drawn an object, we then sample the region wide features \( R_i \sim \text{Mult}(\alpha_i|O) \) and the quantized SIFT features \( X \sim \text{Mult}(\beta|O) \).
Generative Process

- Step 4: Draw a binary switch variable $S$ which determines if a word is mapped to a specific region or represents an image related non-visual concept.
• Step 5a: If word represents a specific region on the image, draw the word from $T \sim \text{Mult}(\theta_{OZ})$, where $Z$ is an indicator which associates the word with a region, $Z \sim \text{Unif}(N_r)$. 

Generative Process
Based on the diagram and the text provided:

- **Step 5b:** If the word relates to a non-visual concept, then draw it from $T \sim \text{Mult}(\varphi_c)$. 
Model Inference

• The model can be summarized as

\[
p(C, O, R, X, S, T, Z | \eta, \alpha, \beta, \gamma, \theta, \varphi) = p(C) \cdot \left( \prod_{n=1}^{N_r} p(O_n | \eta, C) \right) \\
\times \left( \prod_{n=1}^{N_r} \left( \prod_{i=1}^{N_P} p(R_{ni} | O_n, \alpha_i) \right) \cdot \prod_{r=1}^{A_r} p(X_{nr} | O_n, \beta) \right) \\
\times \left( \prod_{m=1}^{N_t} p(Z_m | N_r)p(S_m | O_{Z_m}, \gamma) p(T_m | O_{Z_m}, S_m, \theta, C, \varphi) \right)
\]

\( \text{Classification} \)

\( \text{Segmentation} \)

\( \text{Annotation} \)

• Parameter inference is performed via collapsed Gibbs sampling.
Discussion

• Intuitively, this model can be thought of as 4 parallel LDA models sharing the class indicator C.

• An important contribution of this paper is the switch variable S which allows association of words that are only conceptually (rather than contextually) related to an image.

Figure 3. Probabilities of different objects. Words such as ‘horse’ or ‘net’ have higher probability because users tend to only tag them when the objects are really present, largely due to their clear visual relevance. On the contrary, words such as ‘island’ and ‘wind’ are usually related to the location or some other visually irrelevant concept, and usually not observable in a normal photograph.
Automatic Initialization

• The proposed model still requires a handful of relatively clean and partially labeled images on which to learn.

• To this end, the authors use the following initialization method
  – Remove words which are “physical entities”.
  – Query Flickr using the cropped word list.
  – Train object models using a previously proposed method by the authors.
  – Apply trained object models to the remainder of the training dataset.
Algorithm 1 Automatic training framework

Step 1: Obtaining Candidate Tags. Reduce the number of tags by keeping words that belong to the ‘physical entity’ group in WordNet. Group synonyms using WordNet synsets.

Step 2: Initialize Object Regions
Obtain initial object models. Apply the automatic learning method of [4] to learn an initial object model.
Annotate scene images. Apply the learned object model to annotate candidate object regions in each scene image.
Select initialization images. Select a small number of initialized images by a ranking metric described by Footnote 2.

Step 3: Automatic Learning. Treat the automatically selected top ranked images as ‘supervised’ data, add more flickr images and their tags to jointly train the model described in section 2.

(a) The original image.
(b) The original tags from Flickr. Visually irrelevant tags are colored in red.
(c) Output of Step 1 of Algorithm 1: Tags after the WordNet pruning.
(d) Output of Step 2 of Algorithm 1.
(e) Output of Step 3 of Algorithm 1.
(f) Final annotation proposed by our approach. Blue tags are predicted by the visual component \((S = \text{visual})\). Green tags are generated from the top down scene information learned by the model \((S = \text{non-visual})\).
Experiments and Results

• Experiments were conducted on a corpus with the following categories
  – badminton, bocce, croquet, polo, rock climbing, rowing, sailing, snowboarding.

• 800 images were collected for each category, with 200 images per category set aside for test, yielding 1256 unique tags from which the authors chose the 30 most frequently appearing for segmentation experiments.
Experiments and Results

• **Classification results**

Figure 5. Comparison of classification results. **Left: Overall performance.** Confusion table for the 8-way scene classification. Rows represent the models for each scene while the columns represent the ground truth classes. The overall classification performance is 54%. **Right: Comparison with different models (Experiment A).** Performance of four methods. Percentage on each bar represents the average scene classification performance. 3rd bar is the modified Corr-LDA model [3].
Experiments and Results

• Effect of noisy annotation

Figure 6. Left: **Influence of unannotated data (Experiment B).** Classification performance as a function of number of unannotated images. The y axis represents the average classification performance. The x axis represents the number of unlabeled images. It shows the unannotated images also contribute to the learning process of our model. Right: **Effect of noise in tags (Experiment C).** Performance of different models as a function of noise percentage in the tags. The y axis is average classification performance. The x axis represents the percentage of noisy tags. While the performance of corr-LDA decreases with the increase of percentage of noise, our model performs robustly by selectively learning the related tags.
Experiments and Results

- **Precision**

\[
\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}
\]

- **Recall**

\[
\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}
\]

- **F-measure**

\[
F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}
\]
Experiments and Results

- Annotation results

<table>
<thead>
<tr>
<th>Object</th>
<th>Alipr Prec</th>
<th>Alipr Rec</th>
<th>Alipr F</th>
<th>Corr LDA Prec</th>
<th>Corr LDA Rec</th>
<th>Corr LDA F</th>
<th>Our Model Prec</th>
<th>Our Model Rec</th>
<th>Our Model F</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>.83</td>
<td>.97</td>
<td>.89</td>
<td>.83</td>
<td>1.00</td>
<td>.91</td>
<td>.85</td>
<td>.98</td>
<td>.91</td>
</tr>
<tr>
<td>horse</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>.17</td>
<td>.91</td>
<td>.28</td>
<td>.17</td>
<td>.91</td>
<td>.29</td>
</tr>
<tr>
<td>grass</td>
<td>.42</td>
<td>.86</td>
<td>.56</td>
<td>.22</td>
<td>1.00</td>
<td>.35</td>
<td>.33</td>
<td>.86</td>
<td>.48</td>
</tr>
<tr>
<td>sky</td>
<td>.59</td>
<td>.33</td>
<td>.43</td>
<td>.55</td>
<td>.17</td>
<td>.26</td>
<td>.44</td>
<td>.92</td>
<td>.59</td>
</tr>
<tr>
<td>tree</td>
<td>.45</td>
<td>.38</td>
<td>.41</td>
<td>.25</td>
<td>.01</td>
<td>.03</td>
<td>.38</td>
<td>.93</td>
<td>.54</td>
</tr>
<tr>
<td>net</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>.27</td>
<td>.85</td>
<td>.41</td>
</tr>
<tr>
<td>sand</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>.24</td>
<td>.46</td>
<td>.32</td>
</tr>
<tr>
<td>Mean</td>
<td>.15</td>
<td>.22</td>
<td>.16</td>
<td>.16</td>
<td>.40</td>
<td>.15</td>
<td>.28</td>
<td>.73</td>
<td>.34</td>
</tr>
</tbody>
</table>

Table 1. Comparison of precision and recall values for annotation with Alipr, corr-LDA and our model. Detailed results are given for seven objects, but means are computed for all 30 object categories (Experiment D).

Alipr = Automatic Linguistic Indexing of Pictures - Real Time
Experiments and Results

• Segmentation results

<table>
<thead>
<tr>
<th>Object</th>
<th>Cao &amp; Fei-Fei, 2007</th>
<th>Our Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
</tr>
<tr>
<td>human</td>
<td>.35</td>
<td>.23</td>
</tr>
<tr>
<td>horse</td>
<td>.13</td>
<td>.49</td>
</tr>
<tr>
<td>grass</td>
<td>.62</td>
<td>.38</td>
</tr>
<tr>
<td>sky</td>
<td>.79</td>
<td>.44</td>
</tr>
<tr>
<td>tree</td>
<td>.40</td>
<td>.48</td>
</tr>
<tr>
<td>net</td>
<td>.04</td>
<td>.09</td>
</tr>
<tr>
<td>sand</td>
<td>.11</td>
<td>.32</td>
</tr>
<tr>
<td>Mean</td>
<td>.22</td>
<td>.34</td>
</tr>
</tbody>
</table>

Table 2. Results of segmentation on seven object categories and mean values for all 30 categories (Experiment E).
Experiments and Results

- **Influence of the top level class influence**

Figure 7. Comparison of object segmentation results with or without the top down scene class influence. Each triplet of images show results of one scene class (Experiment F). The left image shows object segmentation result without the top down contextual information, i.e., by setting the probability distribution of object given scene class to a fixed uniform distribution. The center image shows object segmentation result by using the full model. We observe objects are more accurately recognized and delineated. The right image shows the probability of the 5 most likely objects per scene class. This probability encodes the top down contextual information.
Summary and Conclusion

• In this paper, the authors have proposed a model which simultaneously classifies, segments, and tags a corpus of images in a unified hierarchical framework.

• The authors claim that their model is the first to account for noisy tags often seen in real world data.

• The authors plan on improving future performance by implementing a better image model which accounts for the geometry and appearance information of objects.