MIL at ImageCLEF 2013: Scalable System for Image Annotation

Machine Intelligence Laboratory, the University of Tokyo, Japan
Masatoshi Hidaka
Naoyuki Gunji
Tatsuya Harada
Scalable Concept Image Annotation Task

- To make image annotation system from wild web data

- 250,000 Webpages (XML + Image)
- Concepts (labels) to learn
- Test images

Annotating Learning System

Concepts of development set (ground truth given) and test set are different!

Comparison to Ground Truth
Contents

• Scalable Concept Image Annotation Task
  • Image Feature; Fisher Vector, state-of-the-art
  • Textual Feature; our original method which supports concept set change
  • Multilabel Annotation Learning; PAAPL, scalable to the dataset size

Learning Pipeline

1. Training Dataset
   - Image
   - Page Text (XML)

2. Image Feature Extractor
   - Image Feature Vector

3. Textual Feature (Label) Extractor
   - Labels for Image

4. Multilabel Annotation Learning
   - Linear Model Matrix
Image Feature – Fisher Vector [Perronnin et al., 2010]

• Local descriptor
  • SIFT, C-SIFT, GIST, LBP are used separately
  • Using GIST not for global image feature, but for local descriptor

• Statistic calculation
  • Calculate local descriptors \( \{x_1, x_2, \ldots, x_N\} \) statistic using Gaussian Mixture Model \( w_i, \mu_i, \Sigma_i \) calculated by random sample in dataset beforehand

\[
\begin{align*}
\mathbf{u}_i &= \frac{1}{N\sqrt{w_i}} \sum_{n=1}^{N} \gamma_n(i) \Sigma_i^{-\frac{1}{2}} (x_n - \mu_i) \\
\mathbf{v}_i &= \frac{1}{N\sqrt{2w_i}} \sum_{n=1}^{N} \gamma_n(i) [\Sigma_i^{-1} \text{diag}((x_n - \mu_i)(x_n - \mu_i)^T) - 1]
\end{align*}
\]

Image \rightarrow Local Descriptor Extraction \rightarrow Statistic Calculation \rightarrow Normalization \rightarrow Spatial Info \rightarrow Image Feature Vector
Image Feature – Fisher Vector [Perronnin et al., 2010]

- **Normalization**
  - FV representation: \( G = [u_1^T, v_1^T, \ldots, u_K^T, v_K^T]^T \)
  - Power normalization: \( \text{sign}(G)|G|^{1/2} \)

- **Spatial Information**
  - Calculate FVs for divided 8 areas and concatenate them
  \( G = [G_1^T, G_2^T, \ldots, G_8^T]^T \)

- The dimension of our FV is 262144

![Diagram of Image Feature - Fisher Vector process]

Image \( \rightarrow \) Local Descriptor Extraction \( \rightarrow \) Statistic Calculation \( \rightarrow \) Normalization \( \rightarrow \) Spatial Info \( \rightarrow \) Image Feature Vector
Textual Feature – Pipeline

- Supporting concepts of both development and test set is required
- Use WordNet [Fellbaum, 1998] as an external source
- Fast and significantly improves performance
Textual Feature – Text Extraction

- Webpage is NOT concentrating on one image
  - Range of text corresponding to the image is limited
- Parse XML and extract elements
  - Page Title
    - Img tag attributes (filename, alternative text, title)
    - Text displayed near the image
- Select text closely related to the image
- Regard text as a set of words $T$
  - Not considered about grammar

$T = \{\text{swim, with, dolphin, bahama, encounter, …}\}$

Image related words

Text far from image is less related to the image

=> Consider only certain words distant from image
Textual Feature – Label Estimation.

- Simplest method (used in ImageCLEF 2012) [Ushiku et al., 2012]

Silt soil is perfect for certain plants, such as these young celery plants.

http://garden.lovetoknow.com/wiki/Slideshow:Types_of_Soil#7

Word match with label

Image + Estimated Label

Label set: Car, Fish, Plant, Sea, ...

Concept C: {image related words}

Label: Plant
Textual Feature – Label Estimation

• Problem: related word cannot be used

Azaleas in Ilam Botanical Gardens
The Christchurch Ilam Botanical Gardens in flower are a feast of brightly coloured plants and trees. Azaleas and rhododendrons are surrounded by many well-tended shrubs. My first visit was in mid-spring, early for the gardens to be completely in bloom. Paths and lawns are designed for the finest flowery display.

Webpage
Label
Assignment
XML Parsing
Concept C

http://www.mooseyscountrygarden.com/botanical-gardens/ilam-azaleas.html

Azalea is a plant

Label set
Car
Fish
Plant
Sea

Image + Estimated Label

2013/9/25 CLEF
Machine Intelligence Lab. / Univ. of Tokyo
Textual Feature – Label Estimation

- Using related words are important
- [Jin et al., 2005] used semantic distance from WordNet to remove irrelevant keywords from annotation
- [Villegas et al., 2012] used words from definition of concept in English dictionary and constructed probabilistic model
- We try to collect more concept related words simply
Textual Feature – Word Collection

- Collect words $W_C$ related to each concept $C$
- Use synonyms and hyponyms of the concept word
  - Quite simpler than other methods (e.g. Google Distance)
  - Retrieved from WordNet

WordNet Hierarchy (Simplified)

$C = \text{food}$

$W_{\text{food}} = \{\text{food, nutrient, drink, milk, ...}\}$
Textual Feature – Label Assignment

A label is assigned to the image if image related words contains any of concept related words.

From webpage

- T = {pigeon, on, a, tree} (image related words)
- $W_{bird} = \{bird, parrot, \textbf{pigeon}\}$
- $W_{food} = \{food, nutrient, drink\}$
- $W_{plant} = \{plant, \textbf{tree}, rose\}$

Estimated labels

- Bird
- Plant

From WordNet
Online Multilabel Annotation Learning

• To make system scalable, linear model based approach is adopted
  • K-NN based approach: complexity of recognizing is $O(N)$ ($N$ is dataset size)
  • Kernel based approach: complexity of learning is $O(N^2)$

• PAAPL: Passive Aggressive with Averaged Pairwise Loss [Ushiku et al., 2012]

• Passive Aggressive [Crammer et al., 2006] based method
  • Online; requires less RAM
  • Robust to noise of label data

• Converges faster than original PA in multilabel learning
PAAPL – Learning Flow (summarized)

* Update models $\mu^c$ sequentially for each training sample by following
  * Fetch training sample; image feature $f$, assigned labels $Y$, not assigned labels $\bar{Y}$
  * Find a label $r$ in $Y$, a label $s$ in $\bar{Y}$ by follows
    $$r = \text{argmin}_{r \in Y} \mu^r \cdot f$$
    $$s = \text{argmax}_{s \in \bar{Y}} \mu^s \cdot f$$
  * Calculate hinge-loss $l$ and update models according to PA
    $$l = \max(1 - (\mu^r \cdot f - \mu^s \cdot f), 0)$$
    $$\mu^r_{\text{new}} = \mu^r + \frac{l}{(2 |f|^2 + 1/D)} \cdot f$$
    $$\mu^s_{\text{new}} = \mu^s - \frac{l}{(2 |f|^2 + 1/D)} \cdot f$$
  * Repeat above for previously not selected labels
    * This procedure is not in original PA
PAAPL – Learning Flow (summarized)

- Update models $\mathbf{\mu}^C$ sequentially for each training sample by following:
  - Fetch training sample; image feature $\mathbf{f}$, assigned labels $Y$, not assigned labels $\bar{Y}$
  - Find a label $r$ in $Y$, a label $s$ in $\bar{Y}$ by follows
    \[
    r = \operatorname{argmin}_{r \in Y} \mathbf{\mu}^r \cdot \mathbf{f}
    \]
    \[
    s = \operatorname{argmax}_{s \in \bar{Y}} \mathbf{\mu}^s \cdot \mathbf{f}
    \]
  - Mistakenly low scored label

- Calculate hinge-loss $l$ and update models according to PA
  \[
  l = \max(1 - (\mathbf{\mu}^r \cdot \mathbf{f} - \mathbf{\mu}^s \cdot \mathbf{f}), 0)
  \]
  \[
  \mu^r_{\text{new}} = \mu^r + \frac{l}{(2 |\mathbf{f}|^2 + 1/D)} \cdot \mathbf{f}
  \]
  \[
  \mu^s_{\text{new}} = \mu^s - \frac{l}{(2 |\mathbf{f}|^2 + 1/D)} \cdot \mathbf{f}
  \]

- Repeat above for previously not selected labels
  - This procedure is not in original PA
PAAPL – Learning Flow (summarized)

• Update models $\mu^c$ sequentially for each training sample by following
  • Fetch training sample; image feature $f$, assigned labels $Y$, not assigned labels $\bar{Y}$
  • Find a label $r$ in $Y$, a label $s$ in $\bar{Y}$ by follows
    $$r = \arg\min_{r \in Y} \mu^r \cdot f$$
    $$s = \arg\max_{s \in \bar{Y}} \mu^s \cdot f$$

  ![Mistakenly low scored label]

• Calculate hinge-loss $l$ and update models according to PA
  $$l = \max(1 - (\mu^r \cdot f - \mu^s \cdot f), 0)$$
  $$\mu^r_{\text{new}} = \mu^r + \frac{l}{(2 |f|^2 + 1/D)} \cdot f$$
  $$\mu^s_{\text{new}} = \mu^s - \frac{l}{(2 |f|^2 + 1/D)} \cdot f$$

• Repeat above for previously not selected labels
  • This procedure is not in original PA
PAAPL – Learning Flow (summarized)

- Update models $\mu^C$ sequentially for each training sample by following:
  - Fetch training sample; image feature $f$, assigned labels $Y$, not assigned labels $\tilde{Y}$
  - Find a label $r$ in $Y$, a label $s$ in $\tilde{Y}$ by follows
    \[
    r = \arg\min_{r \in Y} \mu^r \cdot f \\
    s = \arg\max_{s \in \tilde{Y}} \mu^s \cdot f
    \]
  - Calculate hinge-loss $l$ and update models according to PA
    \[
    l = \max(1 - (\mu^r \cdot f - \mu^s \cdot f), 0)
    \]
    \[
    \mu^r_{new} = \mu^r + l/(2 |f|^2 + 1/D) \cdot f \\
    \mu^s_{new} = \mu^s - l/(2 |f|^2 + 1/D) \cdot f
    \]

- Repeat above for previously not selected labels
  - This procedure is not in original PA
PAAPL – Advantages

• Score computation process is heavy part of PA
  • PAAPL updates all pairs of models by one score computation
  • It makes convergence faster

• To make faster, random sampling is adopted
  • Only scores of some portion of models are computed

Hinge-loss in PA

Averaged Pairwise Loss

full sampling

random sampling
Multiple Feature Score Combination

- Scores of models which were learned by different image features are summed in annotation step
  - Which combination is best is evaluated by experiment

Annotation Pipeline

- Test Image
- FV of SIFT
- FV of LBP
- Classifiers learned by SIFT
- Classifiers learned by C-SIFT
- Classifiers learned by GIST
- Classifiers learned by LBP
- Scores for labels
- Equally weighted sum
- Output top 5 scored labels

Labels
Experiment Condition

• We applied these methods to ImageCLEF 2013 dataset
• Experiment order
  1. Label estimation condition (whether to use synonyms and hyponyms)
  2. Text extraction condition (whether to use page title etc.)
  3. Comparison of image local descriptors and their score combination
• Image feature for first two experiment is provided C-SIFT + BoVW
• Evaluation was done by F-measure for development set
• Submitted runs are computed with best parameters for development set
Experiment Results – Label Estimation

- Whether we should use synonyms and hyponyms

$C = \text{food}$  \quad W_{\text{food}} = \{\text{food, nutrient, drink, milk, ...}\}

<table>
<thead>
<tr>
<th>Synonym</th>
<th>Hyponym</th>
<th>MF-samples [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✔</td>
<td>23.4</td>
</tr>
<tr>
<td>✔</td>
<td></td>
<td>23.2</td>
</tr>
<tr>
<td></td>
<td>✔</td>
<td>26.1</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>26.6</td>
</tr>
</tbody>
</table>

Using both synonyms and hyponyms is the best.
**Experiment Results – Text Extraction**

- What elements of webpages we should use (best 3 & baseline shown)

<table>
<thead>
<tr>
<th>Text around image [max word distance]</th>
<th>Img tag attributes</th>
<th>Page title</th>
<th>MF-samples [%]</th>
<th>Number of images with label</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>✓</td>
<td></td>
<td>27.6</td>
<td>80009 [lowest]</td>
</tr>
<tr>
<td>10</td>
<td>✓</td>
<td>✓</td>
<td>26.6</td>
<td>129050</td>
</tr>
<tr>
<td>10</td>
<td>✓</td>
<td>✓</td>
<td>26.1</td>
<td>140448</td>
</tr>
<tr>
<td>1000</td>
<td>✓</td>
<td>✓</td>
<td>20.7</td>
<td>193971</td>
</tr>
</tbody>
</table>

People use filename to manage photos

Baseline

- <h1>Swim With Dolphins Bahamas</h1>
  <img src="bahamas-dolphin-encounters.jpg" alt="Swim With Dolphins Bahamas">
  
  The popular Bahamas Dolphin Encounters specializes in creating opportunities for humans to interact safely with dolphins.

10 words after image
**Experiment Results – Image Local Descriptor**

- Best 5 combinations and 4 single features (Fisher Vector applied)

<table>
<thead>
<tr>
<th>C-SIFT</th>
<th>GIST</th>
<th>LBP</th>
<th>SIFT</th>
<th>MF-samples [%]</th>
<th>Test set MF-samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td><strong>34.6</strong> [ISI-1]</td>
<td><strong>33.2</strong></td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>34.3 [ISI-2]</td>
<td>32.7</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>34.2 [ISI-3]</td>
<td>31.8</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>34.0 [ISI-4]</td>
<td>32.4</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>33.9 [ISI-5]</td>
<td>31.7</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td>31.2</td>
<td></td>
</tr>
<tr>
<td>✔</td>
<td></td>
<td>✔</td>
<td></td>
<td><strong>32.4</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td>27.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✔</td>
<td></td>
<td>31.1</td>
<td></td>
</tr>
</tbody>
</table>

Provided C-SIFT + BoVW 27.6

- **Submitted runs**
- **GIST is the best among single descriptor**

+ 7pts
Conclusion

• **Visual Feature**
  - Fisher Vector with four local descriptors was used and the combination of C-SIFT, GIST and SIFT showed superior performance than provided C-SIFT + BoVW

• **Textual Feature**
  - Using synonyms and hyponyms for label estimation improved performance
  - Selecting text related to image also highly improved performance
    - Img tag attributes were the most important
  - Worked well in concepts of both development set and test set

• **Learning**
  - The method which is scalable to the size of dataset was adopted
## Experiment Results – Text Extraction (All)

<table>
<thead>
<tr>
<th>Text around image [max word distance]</th>
<th>Img tag attributes</th>
<th>Page title</th>
<th>MF-samples [%]</th>
<th>Number of images with label</th>
<th>Average number of labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td></td>
<td></td>
<td>25.4</td>
<td>113802</td>
<td>0.7</td>
</tr>
<tr>
<td>100</td>
<td>✔</td>
<td></td>
<td>23.1</td>
<td>183545</td>
<td>2.5</td>
</tr>
<tr>
<td>1000</td>
<td>✔</td>
<td></td>
<td>20.2</td>
<td>192210</td>
<td>5.2</td>
</tr>
<tr>
<td>-</td>
<td>✔</td>
<td></td>
<td>27.6</td>
<td>80009</td>
<td>0.4</td>
</tr>
<tr>
<td>10</td>
<td>✔</td>
<td></td>
<td>26.6</td>
<td>129050</td>
<td>0.8</td>
</tr>
<tr>
<td>100</td>
<td>✔</td>
<td></td>
<td>23.8</td>
<td>185471</td>
<td>2.5</td>
</tr>
<tr>
<td>1000</td>
<td>✔</td>
<td></td>
<td>21.3</td>
<td>193170</td>
<td>5.3</td>
</tr>
<tr>
<td>-</td>
<td>✔</td>
<td></td>
<td>24.6</td>
<td>92254</td>
<td>0.5</td>
</tr>
<tr>
<td>10</td>
<td>✔</td>
<td></td>
<td>25.5</td>
<td>134318</td>
<td>0.9</td>
</tr>
<tr>
<td>100</td>
<td>✔</td>
<td></td>
<td>22.9</td>
<td>185471</td>
<td>2.5</td>
</tr>
<tr>
<td>1000</td>
<td>✔</td>
<td></td>
<td>20.5</td>
<td>193497</td>
<td>5.3</td>
</tr>
<tr>
<td>-</td>
<td>✔</td>
<td>✔</td>
<td>26.0</td>
<td>111247</td>
<td>0.6</td>
</tr>
<tr>
<td>10</td>
<td>✔</td>
<td>✔</td>
<td>26.1</td>
<td>140448</td>
<td>0.9</td>
</tr>
<tr>
<td>100</td>
<td>✔</td>
<td>✔</td>
<td>23.0</td>
<td>186394</td>
<td>2.6</td>
</tr>
<tr>
<td>1000</td>
<td>✔</td>
<td>✔</td>
<td>20.7</td>
<td>193971</td>
<td>5.3</td>
</tr>
</tbody>
</table>
Textual Feature – Implementation Detail

• Text Extraction
  • Words are singularized by ActiveSupport library

• Word Collection
  • Used synset of synset id specified in the concept list
  • Ambiguous words (words of multiple meaning) are not used as related words
    • The word which appears in multiple synset in WordNet is judged to be ambiguous
  • Hyponyms are gathered from all depths from the synset of concept