Overview of the ImageCLEF 2012 Scalable Web Image Annotation Task

Mauricio Villegas and Roberto Paredes
{mvillegas, rparedes}@iti.upv.es

CLEF 2012

Rome, 19\textsuperscript{th} of September 2012
Overview of the ImageCLEF 2012 Scalable Web Image Annotation Task

1. Introduction

2. Task Description
   - Subtasks
   - Web Training Dataset

3. Evaluation Results
   - Subtask 1
   - Subtask 2

4. Conclusions and Future Work
Introduction

Image Annotation: Detecting concepts present in an image.

- Dogs
- Wheelchair
- Breeds
- Table
- Rural
- Grass
- Daytime
- Trees
- ...

Overview of the ImageCLEF 2012 Scalable Web Image Annotation Task
The research on image annotation has mostly relied on manually labeled training data, for which crowdsourcing has become a common practice.

Even though crowdsourcing has proved to be very useful, it is expensive and difficult to scale to a large amount of concepts.

Millions of images and corresponding related text can be cheaply crawled from the Internet for practically any topic.

With the aim of exploring possible complements or alternatives to the crowdsourcing approach by using Web data, we proposed this new task for ImageCLEF 2012.
How to effectively use Web data for image annotation?

- The text in websites is noisy and the degree of relationship to the images varies greatly.

- The types of images also varies. Take for example images from a Web search query of “rainbow”:
Outline

1. Introduction

2. Task Description
   - Subtasks
   - Web Training Dataset

3. Evaluation Results
   - Subtask 1
   - Subtask 2

4. Conclusions and Future Work
Subtask 1: Complementing Manually Labeled Data

- **Objective:** Try to use both automatically gathered Web data and labeled data to enhance the performance in comparison to using only the labeled data.

- **Training set:**
  - Web (250,000 images, unlabeled, textual features).
  - Flickr (15,000 images, labeled for 94 concepts).

- **Test set:** Flickr (10,000 images, labeled for same 94 concepts).

- **Submission:** Concept scores and which were annotated per image (max. 5 runs per group).
Subtask 2: Scalable Concept Image Annotation

- **Objective:** Use only automatically gathered Web data and language resources to develop a concept scalable annotation system.

- **Training set:** Web (250,000 images, unlabeled, textual features).

- **Development set:** Web (1,000 images, labeled for 95 concepts).

- **Test set:** Web (2,000 images, labeled for 105 concepts).

- **Submission:** Concept scores and which were annotated per image (max. 5 runs per set per group).
Web training dataset composed of 250,000 images, 7 visual features types and 4 textual feature types.

Images found by querying Google, Bing and Yahoo using the words from the English dictionary.

Precautions taken to avoid “message images”, duplicates and near-duplicates.

To ease data download and handling by participants, the subset of 250,000 images was selected using 158 concepts (including the concepts for the task).
### Visual Features:

<table>
<thead>
<tr>
<th>Feature</th>
<th>BoW</th>
<th>Raw desc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>5k dim. (780M) and 50k dim. (1.3G)</td>
<td>128 dim. (7G)</td>
</tr>
<tr>
<td>C-SIFT</td>
<td>5k dim. (680M) and 50k dim. (1.2G)</td>
<td>384 dim. (20G)</td>
</tr>
<tr>
<td>RGB-SIFT</td>
<td>5k dim. (760M) and 50k dim. (1.3G)</td>
<td>384 dim. (18G)</td>
</tr>
<tr>
<td>OPP-SIFT</td>
<td>5k dim. (630M) and 50k dim. (1.2G)</td>
<td>384 dim. (19G)</td>
</tr>
<tr>
<td>SURF</td>
<td>–</td>
<td>64 dim. (11G)</td>
</tr>
<tr>
<td>GIST</td>
<td></td>
<td>480 dim. (570M)</td>
</tr>
<tr>
<td>Color Hist.</td>
<td></td>
<td>576 dim. (170M)</td>
</tr>
<tr>
<td>Thumbnails</td>
<td></td>
<td>Max. 200 pixels high</td>
</tr>
</tbody>
</table>
Task Description – Web Training Dataset

Textual Features:

1. Words used to find the images (3M).
2. Relative URLs of images in webpages (25M).

Dogs can tell size of another dog by listening to its growls

Washington, Dec 21: A new study has shown that dogs can tell the size of another dog by listening to its growls. Peter Pongracz and his team recruited 96 dogs of various breeds ...

<html>
<head>
<title>Dogs can tell size of another dog by listening to its growls | Science / Technology</title>
</head>
<body>
<h2>Dogs can tell size of another dog by listening to its growls</h2>
<img src="img/dogs.jpg" alt="dogs in the park" />
<p>Washington, Dec 21: A new study has shown that dogs can tell the size of another dog by listening to its growls. </p>
<p>Peter Pongracz and his team recruited 96 dogs of various breeds ...</p>
</body>
</html>

3. Website text 1 (300M):

... to its growls. dogs in the park {X}. Washington. Dec 21. A new study has ...

4. Website text 2 (110M):

dogs 0.09 of 0.0422 by 0.0336 growls 0.33 to 0.0326 dog 0.0321 can 0.0309 size 0.0307 ...
Dogs can tell size of another dog by listening to its growls

Washington, Dec 21: A new study has shown that dogs can tell the size of another dog by listening to its growls. Peter Pongracz and his team recruited 96 dogs of various breeds...

<html>
<head>
<title>Dogs can tell size of another dog by listening to its growls | Science / Technology</title>
</head>
<body>
<h2>Dogs can tell size of another dog by listening to its growls</h2>
<img src="img/dogs.jpg" alt="dogs in the park"/>
<p>Washington, Dec 21: A new study has shown that dogs can tell the size of another dog by listening to its growls. Peter Pongracz and his team recruited 96 dogs of various breeds ...</p>
</body>
</html>

Website text 1 (300M):
... to its growls. dogs in the park {X}. Washington. Dec 21. A new study has ...

dogs 0.09 of 0.0422 by 0.0336 growls 0.33 to 0.0326 dog 0.0321 can 0.0309 size 0.0307 ...

Website text 2 (110M):
Overview of the ImageCLEF 2012 Scalable Web Image Annotation Task
Subtask 1 – Participation

Received 16 runs from 3 groups (useful from only 2 groups).

- **KIDS-NUTN (National University of Tainan):**
  - Proposed a fusion of several visual features and textual.
  - For annotation, they tried Random Forests and Multiple Bernoulli Relevance Models.
  - Unclear how they handled the Web data.

- **ISI (University of Tokyo):**
  - Method focused on scalability.
  - Used a combination of several *SIFT features.
  - For annotation, they used their online learning method Passive-Aggressive with Averaged Pairwise Loss.
  - Tackled the Web data by labeling it using the appearance of concept words in the textual features.
# Subtask 1 – Results

Best results\(^1\) using only Flickr data:

<table>
<thead>
<tr>
<th></th>
<th>MAP (%)</th>
<th>MF(_1) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Baseline</td>
<td>10.3</td>
<td>10.0</td>
</tr>
<tr>
<td>ISI 1424</td>
<td>70.8</td>
<td>55.3</td>
</tr>
<tr>
<td>KIDS-NUTN 1451</td>
<td>57.9</td>
<td>45.4</td>
</tr>
</tbody>
</table>

Results\(^1\) using both Web and Flickr data:

<table>
<thead>
<tr>
<th></th>
<th>MAP (%)</th>
<th>MF(_1) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISI 1393</td>
<td>25.0</td>
<td>18.2</td>
</tr>
<tr>
<td>ISI 1398</td>
<td>24.7</td>
<td>18.1</td>
</tr>
<tr>
<td>ISI 1399</td>
<td>24.5</td>
<td>17.8</td>
</tr>
<tr>
<td>ISI 1400</td>
<td>24.1</td>
<td>17.5</td>
</tr>
<tr>
<td>KIDS-NUTN 1369</td>
<td>52.1</td>
<td>39.9</td>
</tr>
<tr>
<td>KIDS-NUTN 1370</td>
<td>53.8</td>
<td>39.7</td>
</tr>
<tr>
<td>KIDS-NUTN 1371</td>
<td>49.3</td>
<td>33.1</td>
</tr>
<tr>
<td>KIDS-NUTN 1372</td>
<td>52.8</td>
<td>40.0</td>
</tr>
</tbody>
</table>

\(^1\)MAP (mean average precision) and MF\(_1\) (mean F-measure) computed per image.
Received 10 runs from 1 group.

- **ISI (University of Tokyo):**
  - Method focused on scalability.
  - Used a combination of several *SIFT features.
  - For annotation, they used their online learning method Passive-Aggressive with Averaged Pairwise Loss.
  - Tackled the Web data by labeling it using the appearance of concept words in the textual features.
Results\textsuperscript{1} for test set:

<table>
<thead>
<tr>
<th></th>
<th>MAP (%)</th>
<th>MF\textsubscript{1} (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Baseline</td>
<td>6.7</td>
<td>5.5</td>
</tr>
<tr>
<td>Co-occurrence Baseline</td>
<td>22.1</td>
<td>17.1</td>
</tr>
<tr>
<td>ISI 1407</td>
<td>31.5</td>
<td>24.6</td>
</tr>
<tr>
<td>ISI 1408</td>
<td>32.2</td>
<td>25.1</td>
</tr>
<tr>
<td>ISI 1411</td>
<td>32.4</td>
<td>25.2</td>
</tr>
<tr>
<td>ISI 1412</td>
<td>32.3</td>
<td>25.4</td>
</tr>
<tr>
<td>ISI 1415</td>
<td>32.1</td>
<td>24.9</td>
</tr>
</tbody>
</table>

\textsuperscript{1}MAP (mean average precision) and MF\textsubscript{1} (mean F-measure) computed per image.
Subtask 2 – Results

ISI results\(^2\) for Flickr ann. subtask:

<table>
<thead>
<tr>
<th>Concept</th>
<th>$F_1$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>87.2</td>
</tr>
<tr>
<td>noblur</td>
<td>82.7</td>
</tr>
<tr>
<td>dog</td>
<td>72.2</td>
</tr>
<tr>
<td>fireworks</td>
<td>66.7</td>
</tr>
<tr>
<td>flower</td>
<td>66.2</td>
</tr>
<tr>
<td>partialblur</td>
<td>64.8</td>
</tr>
<tr>
<td>fooddrink</td>
<td>62.3</td>
</tr>
<tr>
<td>adult</td>
<td>61.2</td>
</tr>
<tr>
<td>one</td>
<td>59.3</td>
</tr>
<tr>
<td>female</td>
<td>58.9</td>
</tr>
<tr>
<td>outdoor</td>
<td>58.8</td>
</tr>
<tr>
<td>tree</td>
<td>58.1</td>
</tr>
</tbody>
</table>

ISI results\(^2\) for Web subtask 2:

<table>
<thead>
<tr>
<th>Concept</th>
<th>$F_1$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>fireworks</td>
<td>70.3</td>
</tr>
<tr>
<td>pencil</td>
<td>69.2</td>
</tr>
<tr>
<td>stars</td>
<td>64.6</td>
</tr>
<tr>
<td>sunrise/sunset</td>
<td>56.5</td>
</tr>
<tr>
<td>drawing/diagram</td>
<td>56.4</td>
</tr>
<tr>
<td>galaxy</td>
<td>50.0</td>
</tr>
<tr>
<td>space</td>
<td>48.3</td>
</tr>
<tr>
<td>newspaper</td>
<td>46.2</td>
</tr>
<tr>
<td>lightning</td>
<td>45.2</td>
</tr>
<tr>
<td>forest</td>
<td>43.0</td>
</tr>
<tr>
<td>pool</td>
<td>42.6</td>
</tr>
<tr>
<td>fire</td>
<td>42.4</td>
</tr>
</tbody>
</table>

\(^2\) $F_1$ (F-measure) computed per concept.
Subtask 2 – Example annotations

**Annotations:**
- person/people
- drawing/diagram
- child
- pencil
- baby

**Textbfmissed:**
- 

**Annotations:**
- person/people
- garden/park
- horse
- sign
- sports

**Missed:**
- building
- tree

**Annotations:**
- water
- aerial
- garden/park
- road
- grass

**Missed:**
- —
Overview of the ImageCLEF 2012 Scalable Web Image Annotation Task
Conclusions and Future Work

- Participation was disappointingly low, making it hard to draw good conclusions.

- Subtask 1:
  - None of the participants were able to take advantage of the Web data.

- Subtask 2:
  - The ISI system obtained a considerable better performance than the provided baselines.
  - The processing of the textual data of ISI is rather simple, so possibly there is much room for improvement.
  - For some concepts, the performance is relatively good, indicating that the Web data can be quite useful.
Conclusions and Future Work

- Results in subtask 2 were interesting, thus repeating the task with more participants would be desirable.

- However, we would like to know:
  - Why so few groups participated?
  - Right now, is there enough interest to repeat the task?
  - Suggestions?
Thank you for your attention!

Questions?