Overview of the ImageCLEF 2014 Scalable Concept Image Annotation Task

Mauricio Villegas and Roberto Paredes
mauvilsa@upv.es  rparedes@prhl.t.upv.es

Slides available at:
http://mvillegas.info/pub/Villegas14_CLEF_Annotation-Overview_presentation.pdf
Outline

1. Introduction
   - Motivation

2. Task Description
   - Lines of work
   - Training dataset

3. Evaluation
   - Participation
   - Results

4. Conclusions and Future Work
**Introduction**

- **Automatic image annotation** is the process by which a computer assigns to an image, metadata that describes its content.

- In this work the metadata considered is only the presence or absence of concepts in the images, e.g.

![Image of dogs and picnic table]

- Dog
- Table
- Rural
- Grass
- Daytime
- Tree
- …
Image annotation research has mostly relied on manually labeled training data. Examples of available datasets are:

- **ImageNet**: ≈1.2M images, 1000 concepts, but only one concept per image.
- **NUS-WIDE**: ≈269k images, multiple concepts per image, but only 81 concepts.

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

Are there alternatives that do scale concept-wise?

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

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 search query of “sun”:
Outline

1. Introduction
   - Motivation

2. Task Description
   - Lines of work
   - Training dataset

3. Evaluation
   - Participation
   - Results

4. Conclusions and Future Work
Objective: To use only automatically gathered data for developing concept scalable image annotation systems.

- Any data could be used as training, except for hand labeled images, e.g. crawled data, WordNet, dictionaries, stemmers, etc.

Participants were provided with:

- Crawled dataset (500,000 images and respective webpages).
- Development set (1,940 samples, labeled for 107 concepts).
- Implementation of a baseline system and code for computing the performance measures.
Test set: 7,291 samples, the participants had to label them for 207 concepts, 100 unseen in development (max. 10 runs could be submitted per group).
Divided into 4 subsets with different concept lists:
- Previous ImageCLEF (116 concepts).
- Related to animals (52 concepts).
- Related to foods (41 concepts).
- Complete list (207 concepts).

Concepts: Defined as WordNet synsets and for most of them also a Wikipedia article.
Task description – Lines of work

In contrast to traditional image annotation tasks, the proposed one involves more lines of work:

- Which representation to use for the images (visual features).

- How to use unsupervised web data as training.
  - Automatically assign concepts to the images using the textual data?
  - How to pre-process and clean the textual data?
  - Use other resources:
    - Ontologies
    - Language dictionaries
    - Automatic translation

- Which method to use for modelling the concepts.

- What strategy to use for deciding how many and which concepts are assigned to an image.
Task description – Training dataset

- Web training dataset\(^1\) composed of 500,000 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.

- Subset of images selected using only the used concepts to ease data download and handling by participants.

\(^1\)Dataset available at [http://risenet.prhlt.upv.es/webupv-datasets](http://risenet.prhlt.upv.es/webupv-datasets)
Task description – Training dataset

Visual Features:

<table>
<thead>
<tr>
<th>Feature</th>
<th>Dimensionality</th>
<th>Training data size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thumbnails</td>
<td>Max. 200 pixels high</td>
<td>30 GB</td>
</tr>
<tr>
<td>GIST</td>
<td>480</td>
<td>1.6 GB</td>
</tr>
<tr>
<td>Color Hist.</td>
<td>576</td>
<td>330 MB</td>
</tr>
<tr>
<td>GETLF</td>
<td>256</td>
<td>60 MB</td>
</tr>
<tr>
<td>SIFT</td>
<td>5,000 BoW</td>
<td>1.5 GB</td>
</tr>
<tr>
<td>C-SIFT</td>
<td>5,000 BoW</td>
<td>1.3 GB</td>
</tr>
<tr>
<td>RGB-SIFT</td>
<td>5,000 BoW</td>
<td>1.5 GB</td>
</tr>
<tr>
<td>OPP-SIFT</td>
<td>5,000 BoW</td>
<td>1.4 GB</td>
</tr>
</tbody>
</table>
Task description – Training dataset

Textual Features:

1. Words used to find the images (5MB).
2. Relative URLs of images in webpages (50MB).

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 ...

3. Image webpages as valid XML (4.7GB).

4. Webpage text (218M):

```
<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>
```
Task description – Training dataset

Textual Features:

1. Words used to find the images (5MB).
2. Relative URLs of images in webpages (50MB).

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

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. Image webpages as valid XML (4.7GB).

4. Webpage text (218M):

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

Overview of the ImageCLEF 2014 Scalable Concept Image Annotation Task
1 Introduction
   • Motivation

2 Task Description
   • Lines of work
   • Training dataset

3 Evaluation
   • Participation
   • Results

4 Conclusions and Future Work
### Evaluation – Participation

<table>
<thead>
<tr>
<th>Groups that registered</th>
<th>43</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total submitted runs</td>
<td>58</td>
</tr>
<tr>
<td>Groups that participated</td>
<td>11</td>
</tr>
<tr>
<td>Groups that submitted working notes paper</td>
<td>9</td>
</tr>
<tr>
<td>Data downloads</td>
<td>&gt; 100</td>
</tr>
</tbody>
</table>

### Participants:

- **DISA**: Laboratory of Data Intensive Systems and Applications of the Masaryk University (Brno, Czech Republic).
- **IPL**: Information Processing Laboraroty of the Athens University of Economics and Business (Athens, Greece).
- **KDEVIR**: Computer Science and Engineering department of the Toyohashi University of Technology (Aichi, Japan).
- **MIL**: Machine Intelligence Lab of the University of Tokyo (Tokyo, Japan).
- **MindLab**: Machine learning, perception and discovery Lab from the Universidad Nacional de Colombia (Bogotá, Colombia).
- **MLIA**: Department of Advanced Information Technology of the Kyushu University (Fukuoka, Japan).
- **RUC**: School of Information of the Renmin University of China (Beijing, China).
- **FINKI**: Faculty of Computer Science and Engineering of the Ss. Cyril and Methodius University (Skopje, Republic of Macedonia).
- **IMC**: Institute of Media Computing of the Fudan University (Shanghai, China).
- **INAOE**: Instituto Nacional de Astrofísica, Óptica y Electrónica (Puebla, Mexico).
- **NII**: National Institute of Informatics (Tokyo, Japan).
Evaluation – Results (complete test set)
Evaluation – Results (subsets)

- **MAP-samples (%)**
  - Foods
  - Animals
  - 207

- **MFI-concepts (%)**
  - Various datasets (KDEVIR, MindLab, MLIA, DISA, RUC, IPL, IMC, INAOE, NII, FINKI)
Evaluation – Results (comparison with 2013)

Overview of the ImageCLEF 2014 Scalable Concept Image Annotation Task
# Evaluation – Details of some submitted systems

<table>
<thead>
<tr>
<th>System</th>
<th>Visual Feat.</th>
<th>Training Data Processing</th>
<th>Annotation Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>KDEVIR run #9</td>
<td>Provided by organizers</td>
<td>- Ontology built per concept using WordNet and Wikipedia</td>
<td>- Multiple SVMs per concept with context dependent kernel</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Training positive and negative samples selected by exploiting ontologies</td>
<td>- Annotation of top-k concepts exploiting ontologies</td>
</tr>
<tr>
<td>MIL run #3</td>
<td>Fisher Vectors &amp; ImageNet CNN</td>
<td>- Extract webpage title, image tag attributes and singularize nouns</td>
<td>- Linear multilabel classifier learned by PAAPL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Label training images by appearance of concept (synonyms and hyponyms)</td>
<td>- Annotation of the 4% top scored concepts</td>
</tr>
<tr>
<td>MindLab run #1</td>
<td>ImageNet CNN</td>
<td>- Extract webpage words, stopword removal and stemming</td>
<td>- A logistic regression (soft-max) model</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Annotation based on threshold</td>
</tr>
<tr>
<td>MLIA run #9</td>
<td>Provided by organizers</td>
<td>- Provided webpage features, stopword removal and stemming</td>
<td>- One SVM per concept, F-measure cross-validation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Concepts assigned to training images by appearance of synonyms, filtered using Overfeat</td>
<td>- Annotation based on SVM classification decision</td>
</tr>
<tr>
<td>System</td>
<td>Visual Feat.</td>
<td>Training Data Processing</td>
<td>Annotation Technique</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------------------</td>
<td>------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>KDEVIR</td>
<td>Provided by organizers</td>
<td>- Ontology built per concept using WordNet and Wikipedia</td>
<td>- Multiple SVMs per concept with <strong>context dependent kernel</strong></td>
</tr>
<tr>
<td>run #9</td>
<td></td>
<td>- Training positive and negative samples selected by exploiting ontologies</td>
<td>- Annotation of top-k concepts exploiting ontologies</td>
</tr>
<tr>
<td>MIL</td>
<td>Fisher Vectors &amp; ImageNet CNN</td>
<td>- Extract webpage title, image tag attributes and singularize nouns</td>
<td>- Linear multilabel classifier learned by PAAPL</td>
</tr>
<tr>
<td>run #3</td>
<td></td>
<td>- Label training images by appearance of concept (synonyms and hyponyms)</td>
<td>- Annotation of the 4% top scored concepts</td>
</tr>
<tr>
<td>MindLab</td>
<td>ImageNet CNN</td>
<td>- Extract webpage words, stopword removal and stemming</td>
<td>- A logistic regression (soft-max) model</td>
</tr>
<tr>
<td>run #1</td>
<td></td>
<td></td>
<td>- Annotation based on threshold</td>
</tr>
<tr>
<td>MLIA</td>
<td>Provided by organizers</td>
<td>- Provided webpage features, stopword removal and stemming</td>
<td>- One SVM per concept, F-measure cross-validation</td>
</tr>
<tr>
<td>run #9</td>
<td></td>
<td>- Concepts assigned to training images by appearance of synonyms, filtered using Overfeat</td>
<td>- Annotation based on SVM classification decision</td>
</tr>
</tbody>
</table>
## Evaluation – Details of some submitted systems

<table>
<thead>
<tr>
<th>System</th>
<th>Visual Feat.</th>
<th>Training Data Processing</th>
<th>Annotation Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>KDEVIR</td>
<td>Provided by organizers</td>
<td>- <strong>Ontology built per concept</strong> using WordNet and Wikipedia</td>
<td>- Multiple SVMs per concept with context dependent kernel</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Training positive and negative samples selected by exploiting ontologies</td>
<td>- <strong>Annotation</strong> of top-(k) concepts exploiting ontologies</td>
</tr>
<tr>
<td>MIL</td>
<td>Fisher Vectors &amp; ImageNet CNN</td>
<td>- Extract webpage title, image tag attributes and singularize nouns</td>
<td>- Linear multilabel classifier learned by PAAPL</td>
</tr>
<tr>
<td>run #3</td>
<td></td>
<td>- Label training images by appearance of concept (synonyms and hyponyms)</td>
<td>- Annotation of the 4% top scored concepts</td>
</tr>
<tr>
<td>MindLab</td>
<td>ImageNet CNN</td>
<td>- Extract webpage words, stopword removal and stemming</td>
<td>- A logistic regression (soft-max) model</td>
</tr>
<tr>
<td>run #1</td>
<td></td>
<td></td>
<td>- Annotation based on threshold</td>
</tr>
<tr>
<td>MLIA</td>
<td>Provided by organizers</td>
<td>- Provided webpage features, stopword removal and stemming</td>
<td>- One SVM per concept, F-measure cross-validation</td>
</tr>
<tr>
<td>run #9</td>
<td></td>
<td>- Concepts assigned to training images by appearance of synonyms, filtered using Overfeat</td>
<td>- Annotation based on SVM classification decision</td>
</tr>
<tr>
<td>System</td>
<td>VisualFeat.</td>
<td>Training Data Processing</td>
<td>Annotation Technique</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------</td>
<td>------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| KDEVIR run #9 | Provided by organizers | - Ontology built per concept using WordNet and Wikipedia  
- Training positive and negative samples selected by exploiting ontologies | - Multiple SVMs per concept with context dependent kernel  
- Annotation of top-k concepts exploiting ontologies |
| MIL run #3   | Fisher Vectors & ImageNet CNN | - Extract webpage title, image tag attributes and singularize nouns  
- Label training images by appearance of concept (synonyms and hyponyms) | - Linear multilabel classifier learned by PAAPL  
- Annotation of the 4% top scored concepts |
| MindLab run #1 | ImageNet CNN         | - Extract webpage words, stopword removal and stemming                                   | - A logistic regression (soft-max) model  
- Annotation based on threshold |
| MLIA run #9  | Provided by organizers | - Provided webpage features, stopword removal and stemming  
- Concepts assigned to training images by appearance of synonyms, filtered using Overfeat | - One SVM per concept, F-measure cross-validation  
- Annotation based on SVM classification decision |
Overview of the ImageCLEF 2014 Scalable Concept Image Annotation Task
Conclusions

- Participation was excellent, and the teams presented diverse approaches to address the proposed challenge.
- The results indicate that the web data can be effectively used for training practical and scalable annotation systems.
- The performance of the systems improved with respect to last year.
- Due to the larger number of unseen concepts, results had narrower confidence intervals, so it made the comparison the systems more conclusive.
- The winning team was KDEVIR. Its success is possibly due to classifier that considers contextual information and usage of concept ontologies both in training and test.
The task will hopefully continue for CLEF 2015, pending the notification of acceptance of ImageCLEF 2015 lab due the 19th of September.

Several modifications to the task:
- Localisation within the images.
- Description sentence generation.

New organisers:
- Andrew Gilbert
- Luca Piras
- The ViSen consortium
Thank you for your attention!

Questions? Comments?