DARPA PPAML Challenge Problem #6: Image Labeling

13 May 2015

# Summary

Multimedia data such as video and pictures are being produced and shared at an unprecedented and accelerating pace in recent years. For example, on YouTube, video data is currently being uploaded at the rate of approximately 30 million hours a year. This drives a strong need to develop automated tools to help users understand, organize, and retrieve images and videos from very large collections.

More concretely, the goal of the proposed challenge problem is to assign labels or tags (e.g. "blue boat", "lake", or "Yosemite State Park") to the query images based on relationships found in a rich social multimedia database. This is comprised of visual features and metadata such as: user information (e.g. username, location, network of contacts), comments, user image gallery, uploader defined groups, and links between shared contents. When all this information is used collectively in a suitable fashion, it may be able to advance the state-of-the-art in image labeling.

|  |  |
| --- | --- |
| Tags: [empty]  Labels: plant\_life, sky, structures, tree | Tags: Paris, Ile-de-France, France, Act Up-Paris  Labels: female, people, plant\_life, sky, structures, tree |
| Tags: Dashboard, steering, wheel, Knox Cruise Night  Labels: car, night, structures, transport | Tags: Tuscon, flower, Om  Labels: people, plant\_life |

#### Sample images, first four associated user-supplied tags, and ground-truth labels, from [Huiskes2008].

# Problem Specification

## Overview

The problem of social multimedia retrieval is to develop the scientific methodology to understand and discover images/videos with particular content from a complex, large, and potentially growing collection of multimedia. Formally, this can be formulated as an image annotation problem to infer content labels, *L*, conditioned on an image, *I*, and other related metadata information, *M*, e.g. *P(L|I,M)*.

In particular, for the current Challenge Problem (CP) we propose to address the automatic image annotation or labeling problem by exploiting the metadata, *M*, in addition to the visual information, *I*. For example, each image is often associated with metadata such as user-provided tags, galleries/groups to which authors shared the multimedia, (optional) timestamp and location, and `likes’ or text comments provided by others. Real-world multimedia, especially as shared on the Internet, can be challenging to retrieve using only visual information, due to complex content, partial occlusion, and diverse styles and quality. This CP will help researchers to explore approaches to utilize such extra information on relationships between images in addition to the visual information.

We will use the MIRFLICKR [Huiskes2008] dataset to supply the ground-truth image labels, image features, and related metadata. The MIRFLICKR data is available under Creative Commons licenses.

In relation to the PPAML taxonomy of challenge problems, this CP is related to the Intelligence Analysis domain; the data structures are a hybrid of discrete (categorical) and continuous (features and feature distances) presented in both relational and vector forms. An undirected, graphical parameterized model over a fixed model with latent variables could be devised to solve the problem. Queries could be formulated as marginal MAP for individual images, or MAP for the entire graph. The query timing could be considered as one-shot with slow tempo and stationary parameters.

## Problem Statement

Some definitions:

* A set of labels is a set of strings representing the "ground truth" of concepts present in an image. MIRFLICKR provides 24 labels, such as "river", "food", and "baby". For the purposes of this CP, is fixed and independent of any particular set of images (although we assume is relevant to whatever set of images we are considering.)
* An image is a matrix of pixel values, which are not available directly in the context of this CP, but instead summarized by various image features such as histograms, texture measures, bag-of-word descriptors, and other standard descriptors, as well as more specialized features such the output of detectors tuned to specific real-world objects such as cars or people.
* Each image is associated with metadata describing, for example, collection parameters (e.g. date and time), camera parameters (e.g. focal length), and whatever text tags the photographer chose to describe the image. **Note in particular** that the text tags are freeform and unrelated to the strings in the label set . Relational metadata includes what other images share the same tags, or were taken by the same photographer.

Given these definitions, the problem is: **Given a set of images and their metadata, infer their labels.** The following sections describe the baseline model, data package and formats, and evaluation methodology.

## Baseline Model

There are a variety of model that can capture relational dependencies, some are probabilistic, while others use potential functions or max-margin optimization [McAuley2012]. These models can jointly learn relationships between the image labels, the image features, and the metadata. The metadata can include: tags, visual categories, user information (username, location, network contacts), groups to which the images are assigned, and user-created collections. We propose to implement a Conditional Random Field (CRF) as one of the baseline models due to its natural ability to model dependencies between pairs of labels while being conditioned on the input image, *I*, and metadata, *M*, . The work in [McAuley2012] can serve as another baseline model, which has existing results for the non-probabilistic models on the same dataset that is being proposed for use here.

Conditional Random field as Baseline P(L|I,M) Model:

The CRF will capture both unary dependencies between image labels, , and the input features (e.g. image features and metadata), as well as the pairwise dependencies between pairs of labels and the input features to produce the conditional probability . The observed input features can include data from three sources:

* raw image features
* classifier outputs
* metadata

The raw image features, , include low-level descriptors such as histogram of oriented gradients and color histograms. The classifier outputs,, are posterior probabilities or scores that represent how well the data matches a set of class models, which are previously trained using image featuresThese classifiers characterize classes such as scene categories (building, grass, road), object categories (person, bicycle, vehicle), or image type (birthday party, nature, dancing). The metadata, *M*, is a binary indicator vector that indicates the occurrence of *words* (derived from titles, descriptions, and comments), *groups,* and *tags* for a single image.

The metadata is also used to define the cliques in the CRF prior to the parameter learning process. The cliques will represent collections of labels that are dependent on each other based on having common properties (i.e. assigned to the same *gallery* or *group*).

The image labels, are treated as hidden nodes in the CRF and the image features, , classifier outputs, f, and metadata, M, are used in the observation nodes. The conditional probability of the CRF is:

, (1)

where Z is the normalization constant, the *A* and B potentials capture the dependencies between the label and the features, and between the pairs of labels and the features, and is the clique neighborhood. For simplicity one binary CRF model can be learned for each category or label (airplane, dog).

The implementation of the CRF can follow the approach used in [Kumar2006], where a fixed feature function is used to calculate the unary potential, *A*:

, (2)

where

, (3)

and

, (4)

where is the classifier output feature vector, but can include the image features, and metadata, while *w* is a vector of learned weights.

The pairwise potential, B, from equation (1) can be modeled using a discriminative model similar to equation (2):

(5)

where is the parameter to be learned for the pairwise potential.. The pairwise potential for image *n* and *m* is denoted as for simplicity, but refers to the features that co-exist for the two images, i.e. , which can be concatenated image features and/or classifier outputs from the two images. The relational metadata can also be added to the pairwise potential by calculated common properties between the two images. These include:

* Number of common tags
* Flag indicating if both images were taken by the same user

Kitware will provide the input features that will be used as observations in the CRF as well as correct labels for each image in the training set as part of a data package.

The number of labels as well as metadata categories can be progressively increased. They may also be changed independently, i.e. increasing the label set while keeping the metadata categories constant, or increasing the complexity of the metadata while keeping the label set constant.

The number of images in the training data can be increased. In the worst case, the number of relationships between images increases quadratically with respect to the size of the dataset, although in practice the increase is only slightly higher than linear.

## Data Source

We will be using data from the MIRFLICKR dataset [Huiskes2008], which contains 25,000 images placed on Flicr under a Creative Commons license by 9,862 unique users; each image has:

1. Label annotations for each image for 24 concepts: *animals, baby, bird, car, clouds, dog, female, flower, food, indoor, lake, male, night, people, plant\_life, portrait, river, sea, sky, structures, sunset, transport, tree, water.* These are organized in a two-level semantic hierarchy described in [Huisekes2008]. When evaluating if two labels match, this hierarchy can be taken into account to estimate semantic precision when estimating labels.
2. The user ID
3. EXIF metadata, providing the time the photo was taken
4. The set of tags the user added to the photo

Additionally, Kitware will provide a suite of features for all the images. The precise set of features is still being developed, but will include standard features such as bag-of-words, color and edge histograms, and/or wavelet textures. Regardless of the precise feature, all of these features will be presented as a vector of integers or floats together with a distance metric such as inner product or chi-squared distance. We will also provide more specialized classifier outputs as described above for e.g. specific object detections or scene classifications. These will be provided as a per-image vector of likelihoods, one for each classifier type.

In round 1 (introduced July 2015), a reduced label set of 10 concepts will be used with a training data set size of 5000 images and a test set of 3000 images. In round 2 (introduced January 2016), the full set of 24 concepts will be used, with an additional 10000 training images (total 15000) and additional 7000 test images (total 10000.) Some of the round 1 training images may acquire additional ground-truth labels as the round 2 concepts are introduced.

## Evaluation

The performance of the image label classifiers will be determined using the sequestered data, which represents about half of the overall dataset. The performance will be measured using both the Mean Average Precision (mAP) as well as the Balanced Error Rate from [McAuley2012].

The mAP is a single value metric that summarizes the quality of a ranked list of classified images based on their associated classifier probability/score. More precisely, the average precision (AP) is the average of the precision values that are calculated at all true positives in a descending ranked list. The AP is calculated for each class (i.e. label) and then averaged over all label experiments to obtain the mAP.

The Balanced Error Rate is designed to assign equal importance to false positives and negatives, which McAuley et al. believe more accurately represents the performance when simultaneously making binary label predictions for the entire dataset. The balanced error rate is calculated as follows:

(6)

where is the number of images with positive labels, is thenumber of negative images, and and are the number of correct positive and negative images, respectively.

# Solution I/O Specification

## Input Data Format

All images (test and training) will reference a list of concepts, e.g. 0 = 'animal', 1 = 'baby', etc. The list will be the same for rounds 1 and 2.

Training data will be provided as a set of files for label, image feature, and metadata, one set per image in the training set.

* Label: one file with two lines: the first line is the image ID; the second is a single line of 24 space-separated "0" or "1" characters, indicating the absence or presence of the corresponding concept in the image. In round 1, the deferred concepts will be represented by 'x' characters.
* Features: one file with the image ID on the first line followed by image features, one vector per line. Precise identity of the feature vectors TBD.
* Metadata: one file with
  + image ID
  + user ID
  + image capture time
  + user tags

Test data will be the same, except without the label file. User IDs will be separated between test and train data; i.e. no individual user ID will have data in both the test and train sets.

## Output Data Format

For test data, your system should return a label file for each image: the image ID on the first line, then a second line with 24 floating point values between 0 and 1, each representing the likelihood that your system estimates the corresponding concept is present in the image. (In round 1, the deferred concepts should be marked with '-1' values.)

# References

* [McAuley2012] Julian McAuley and Jure Leskovec, Image Labeling on a Network: Using Social-Network Metadata for Image Classification, in ECCV 2012.
* [Chua2009] T.S. Chua, J. Ang, R. Hong, H. Li, Z. Luo, Y.T. Zheng: NUS-WIDE: A real-world web image database from the National University of Singapore, in CIVR (2009)
* [Huiskes2008] Huiskes, M., Lew, M. The MIR Flickr retrieval evaluation. In: CIVR. (2008)
* [Nowak2010] S. Nowak, M. Huiskes. New strategies for image annotation: Overview of the photo annotation task at ImageCLEF 2010, in CLEF (2010).
* [Denoyer2010] Denoyer, L., Gallinari, P.: A ranking based model for automatic image annotation in a social network. In: ICWSM. (2010)
* [Lindstaedt2008] Lindstaedt, S., Pammer, V., Morzinger, R., Kern, R., Mulner, H., Wagner, C.: Recommending tags for pictures based on text, visual content and user context. In: Internet and Web Applications and Services 2008.
* [Sigurbjornsson2008] Sigurbjornsson, B., van Zwol, R.: Flickr tag recommendation based on collective knowledge, in WWW, 2008.
* [Sawant2010] Sawant, N., Datta, R., Li, J., Wang, J.: Quest for relevant tags using local interaction networks and visual content. In: MIR. (2010)
* [Stone2008] Stone, Z., Zickler, T., Darrell, T.: Autotagging Facebook: Social network context improves photo annotation. In: CVPR Workshop on Internet Vision. (2008)
* [Kumar2006] S. Kumar and M. Hebet, “Discriminative Random fields,” IJCV, 2006