DARPA PPAML Challenge Problem #6: Image Labeling

Version 5, 28 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.

The goal of the proposed challenge problem is to assign labels (e.g. "tree", "people", or "baby") to images based on relationships found in a rich social multimedia database containing 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 content. 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 multimedia retrieval is to develop the scientific methodology to understand and discover images/videos with particular content from a complex, large, and 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 Challenge Problem 6 (CP6), we will solve the task of automatic image annotation or labeling by exploiting the metadata, M, in addition to the visual information, I. Some types of metadata (i.e. EXIF tags) are generated by the camera when the image is taken; others (i.e. user-provided tags, comments from viewers) are generated after the image is uploaded to an image-sharing service such as Flickr. Generation of associative metadata (such as inclusion in themed "galleries") is an continuous process for as long as the image is available; thus datasets collected at a particular point in time may not be precisely reproducible in the future. 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 allow researchers to explore approaches to exploiting information about relationships between images in addition to visual information.

We will use a subset of 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. The subset is the one used by [McAuley2012], which we will henceforth refer to as the MIR14k dataset.

Within 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. The basic parametric probabilistic model is an undirected graphical model over a fixed model structure with latent variables. Queries are formulated as marginal *maximum a posteriori* MAP for individual images, or joint MAP for the entire graph. The query timing is one-shot with slow tempo and stationary parameters.

Problem Statement

Some definitions:

- A set of labels $L = \{l_1, l_2, \ldots, l_N\}$ is a set of strings representing the "ground truth" of concepts present in an image. MIR14k provides 24 labels, such as "river", "dog", and "baby". For the purposes of CP6, L is fixed and independent of any particular set of images (although we assume L is relevant to whatever set of images we are considering.)
- An image I is a matrix of pixel values. However, instead of requiring the teams to work with raw images, we will summarize the images in terms of 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 M, which may be divided into two types: **intrinsic**, based on the EXIF data collected from the camera, and **relational**, describing the context of the image on Flickr. Intrinsic data may include:
    - Date and time the picture was taken
    - Whether or not the flash fired
    - Focal length
    - Geo-location (uncommon)

  Relational metadata may include:
    - User information
    - Photo title
    - The free-form text tags associated with the image, and who provided the tag. These text tags are unrelated to the strings in the label set L.
    - The Flickr **groups** to which the photo belongs (only the uploader may add photos to groups). In the context of this dataset, groups provide information via the title of the group. For example, a photo of a palm tree may be in two groups, one titled "Vacation photos" and another titled "Hawaii". Other photos in the group will not generally be available.
    - The Flickr **galleries** to which the photo belongs (Flickr users create galleries from other users' photos). Galleries provide both a title and a description, as ancillary information such as the number of photos in the gallery, the number of times it has been viewed, etc.

  Not every photo will have every item of metadata.

Given these definitions, the problem can be stated as follows:

Given:

- A database of images
- Metadata for those images
- Labels for a subset of those images (the training set)

Find:

- The labels of all other images in the database

The following sections describe the baseline model, data package and formats, and evaluation methodology.

Baseline Model

Many models 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. As described above, the metadata can include time-of-day, tags, user information, and groups and galleries to which the images belong. We will 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*, $P(L|I,M)$. The work in [McAuley2012] can serve as another baseline model. That paper also provides results for non-probabilistic models applied to the same dataset we are using here.

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

The CRF will capture both unary dependencies between image labels, $L = \{l_1, l_2, \dots, l_N\}$, 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 $P(L|I,M) = P(L|\boldsymbol{x}, f, M)$, where $\boldsymbol{x}$ and f are features derived from the image set, I. The observed input features can include data from three sources:

- raw image features
- outputs from specific object detectors or classifiers
- metadata

The raw image features, $\boldsymbol{x}$, include low-level descriptors such as histogram of oriented gradients and color histograms. The classifier outputs, $f_c(\boldsymbol{x}); c = \{1,2, \dots, C\}$, are posterior probabilities or scores that represent how well the data matches a set of class models, which have been previously trained using image features. These classifiers characterize classes such as scene categories (building, grass, road), object categories (person, bicycle, vehicle), and 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 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, $L = \{l_1, \dots, l_n, \dots l_N\}$, for each of the *N* images are binary values indicating the image has this class label, $l_n = 1$, or not, $l_n = 0$. The labels are treated as binary hidden nodes in the CRF and the image features, $\boldsymbol{x}$, classifier outputs, f, and metadata, M, are used in the observation nodes. The conditional probability of the CRF is:

$$P(L|I,M) = p(L|\boldsymbol{x}, f, M) = \cdots$$

$$\frac{1}{Z}\exp\left(\sum_{n=1}^{N} A_n(l_n, \boldsymbol{x}, f, M) + \sum_{n=1}^{N}\sum_{m\in\mathcal{N}(n)} B_{n,m}(l_n, l_m, \boldsymbol{x}, f, M)\right), \quad (1)$$

where Z is the normalization constant that depends on $\boldsymbol{x}$, *f*, and *M*, while *A* and B are the unary and pairwise potential functions, and $\mathcal{N}(n)$ is the clique neighborhood. The unary potentials are single image potentials, while the pairwise potentials are between pairs of images. For simplicity, a separate binary CRF model can be learned for each label (e.g., airplane).

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*:

$$A_n(l_n, \boldsymbol{x}, f, M) = \log(\sigma(l_n w^T h_n(\boldsymbol{x}))), \quad (2)$$

where

$$\sigma(Y) = \frac{1}{1+e^{-Y}}, \quad (3)$$

and

$$h_n(\boldsymbol{x}) = [1, \mathrm{f}_1(\boldsymbol{x}), \ldots, \mathrm{f}_C(\boldsymbol{x})], \quad (4)$$

where $\mathrm{f}_c(\boldsymbol{x})$ is the classifier output feature vector, but can include the image features, and metadata, while *w* is a vector of learned weights. The feature vector can be normalized for faster convergence and possibly more accurate results.

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

$$B_{n,m}(l_n, l_m, \boldsymbol{x}) = l_n l_m v^T \mu_{n,m}(x), \quad (5)$$

where v is the parameter to be learned for the pairwise potential.. The pairwise potential for image *n* and *m* is denoted as $\mu_{n,m}(x)$ for simplicity, but refers to the features that co-exist for the two images, i.e. $[h_n(\boldsymbol{x})\ h_m(\boldsymbol{x})]$, 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 calculating common properties between the two images. These include:

- Number of common tags
- Flag indicating if both images were taken by the same user
- Temporal separation between two images

Kitware will provide the input features and metatdata 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.

Data Source

CP6 will use a subset of the MIRFLICKR dataset [Huiskes2008] as defined by McAuley [McAuley2012], which contains 14,460 images uploaded to Flickr and licensed under a Creative Commons license. Various logistical constraints for CP6 reduce this to 12,690 images.

Each image has:

1. One or more label annotations for each image drawn from a vocabulary of 24 concepts. See Figure 1 for label frequency and per-image co-occurrence data.
2. The photo's title
3. The owner's user ID
4. EXIF metadata, providing the time the photo was taken, whether the flash was used, etc.

5. Text tags associated with the photo (supplied by the owner or other Flickr users)
6. Information on the Flickr group and galleries the photo belongs to.
7. The comments associated with the photo

(Note that not every image will have all of 5, 6, or 7.)

Additionally, Kitware will provide a suite of features for all the images. The precise set of features is still being developed, and will include standard features such as bag-of-words, color and edge histograms, and/or wavelet textures. Regardless of the precise feature set, 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 specific object detections or scene classifications. These will be provided as a per-image vector of likelihoods, one for each classifier type.

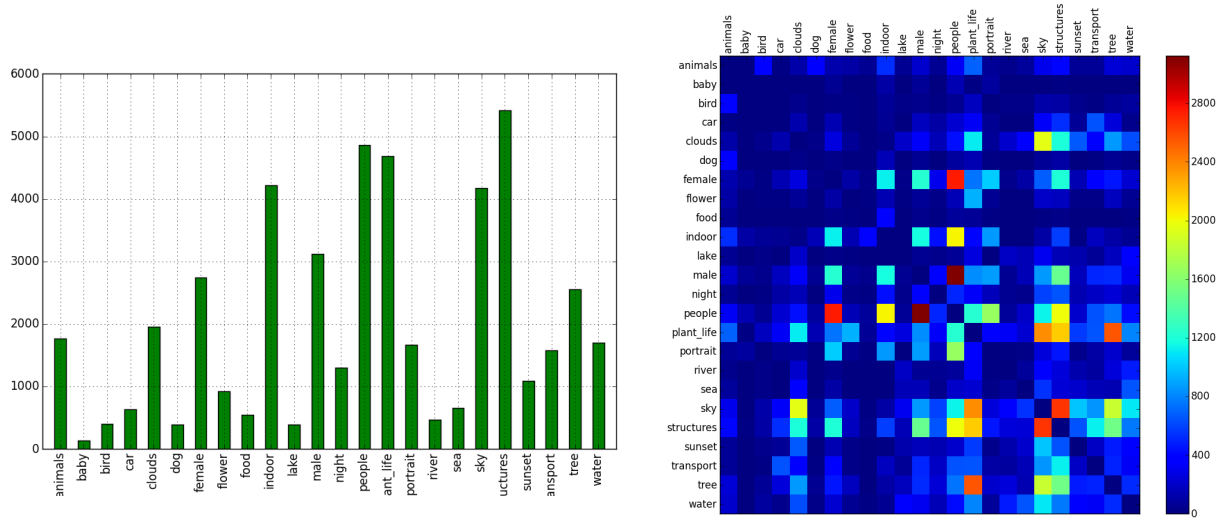

Figure 1; Left, label frequency histogram; right: per-image label co-occurrence heatmap.

The dataset will scale in size from Round 1 to Round 2. The intent is that each round will define training (available) and testing (sequestered) subsets of the MIR14k data, scheduled as follows:

- Round 1 (introduced July 2015, evaluated January 2016): This data drop will focus on images which do **not** include the label "structure", which results in a set of 7269 images.

- Round 2 (introduced January 2016, evaluated July 2016): This data drop will add in the 5421 images with the label "structure".

- For both rounds, data will be partitioned into test and train sets based on EXIF timestamps: data taken before December 2007 will be training, data after December 2007 will be test. Data with no EXIF timestamp (approximately 2000 images) will be split

evenly between testing and training. Figure 2 shows the distribution of Round 1 and Round 2 data versus timestamp.

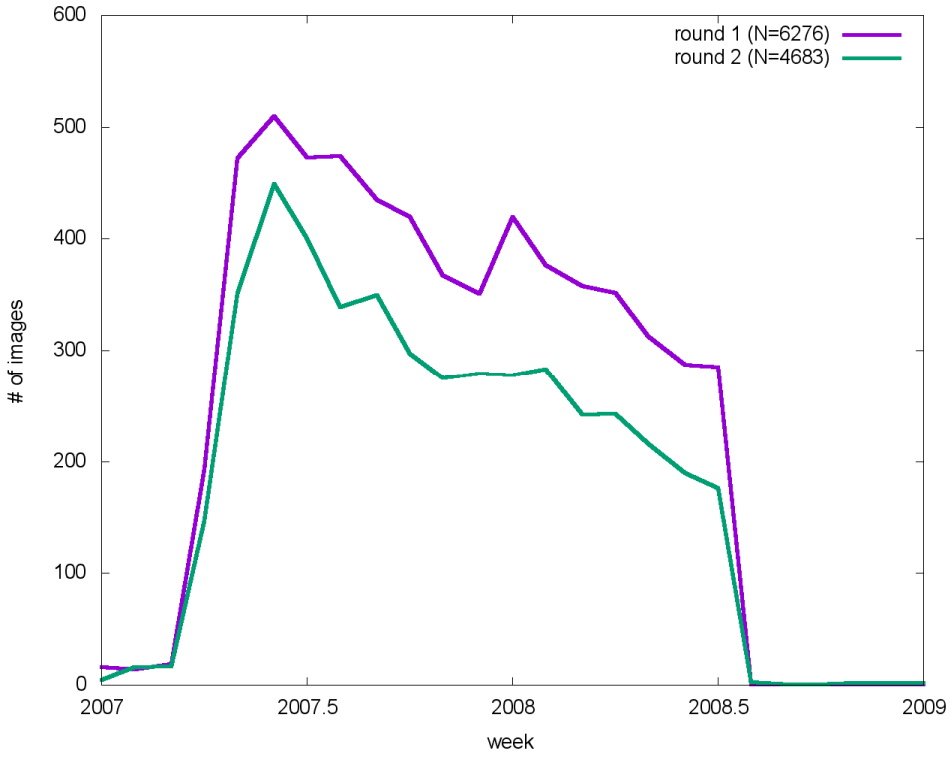

Figure 2; EXIF-timestamp distribution of round 1 and round 2 data

Evaluation

The performance of the image label classifiers will be determined using the test data, which represents about half of the overall dataset. The performance will be measured using both Mean Average Precision (mAP) and Balanced Error Rate [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:

$$\Delta(Y, Y_c) = \frac{1}{2}\left[\frac{|Y^{Pos} \setminus Y_c^{Pos}|}{|Y_c^{Pos}|} + \frac{|Y^{Neg} \setminus Y_c^{Neg}|}{|Y_c^{Neg}|}\right] \quad (6)$$

where Y^{Pos} is the number of images with positive labels, Y^{Neg} is the number of negative images, and Y_c^{Pos} and Y_c^{Neg} 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 24 labels, e.g. 0 = 'animal', 1 = 'baby', etc. 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 label in the image. In round 1, the deferred labels 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 reported as '-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