Flickr-tag Prediction using Multi-modal Fusion and Meta Information

Yu-Chuan Su, Tzu-Hsuan Chiu, Guan-Long Wu
Chun-Yen Yeh, Felix Wu, Winston H. Hsu
National Taiwan University, Taipei, Taiwan

ABSTRACT

We present our evaluation and analysis on Yahoo! Large-scale Flickr-tag Image Classification dataset. Our evaluations show that combining multi-features and different classification models, the MAP of tag prediction can be significantly improve over ordinary linear classification. Further analysis shows that some tags are given not because of the visual content but the meta information of images. Our experiments show that we can make more accurate prediction on certain tags using meta information without any training process, compared with visual content based classifiers. Combine the meta information, multi-features and multi-models fusion, we achieve significantly better performance than simple linear classification. We also evaluate the performance of various mid-level feature, and the results suggest that "Concept Bank" feature may be a promising direction for the task.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis

Keywords

Multi-features Fusion; Meta data; Concept Bank

1. INTRODUCTION

Targeting on ACM Multimedia 2013 Yahoo! Large-scale Flickr-tag Image Classification Grand Challenge, we evaluate various classification models and features, both low-level and mid-level ones, on the dataset. The main challenge of the task is the intra-class visual diversity. Unlike existing image classification datasets which contain many classes with small number of visually consistent images, the proposed dataset consists of only 10 classes with 150k images for each class, and the visual content of the 150k images are so diverse that some of them might not share any visual similarity. Therefore, traditional image classification approach may not apply well on the dataset.

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2508117.

To overcome the difficulty, we propose to adopt high dimensional image feature along with multi-features and models fusion. By combining multiple features and models, we are able to capture more diverse visual content. In particular, we believe fusing with classification models that are resistant to intra-class diversity, such as nearest neighbor (NN) classifier, can significantly improve the performance. Beside features and models fusion, we also propose to adopt meta information for tag-prediction. Our observations indicate that some tags are given mainly based on the meta information rather than the visual content. Without meta information, some tags are doomed to be ill predicted by visual content.

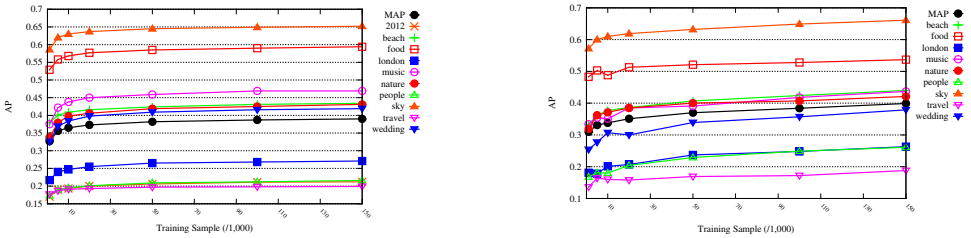
We also evaluate the performance of various mid-level features. With the higher level semantic embedded in mid-level features, they should be more robust to visual content diversity. In particular, we examined human face feature and "Concept Bank" feature on the dataset. Our results suggest that "Concept Bank" feature is a promising direction for the task, but a larger concept space is necessary for competitive performance, where the 136 dimensional concept space in our experiment is not big enough.

The outline of our evaluation on the dataset is as follows. First, we perform linear classification and multi-features fusion (section 3) following state-of-the-art image classification systems [1]. Second, we examine the performance of k-nearest neighbor (kNN) model (section 4). The results motivated the use of meta information for tag prediction, as presented in section 5. Finally, we evaluate the efficacy of mid-level features (section 6).

2. RELATED WORK

Linear support vector machine (SVM) is the most common classifier in large scale image classification. Among all variants of linear SVM solver, LIBLINEAR [4] is the most popular one; the solver, however, does not handle large data that does not fit into memory well. A common approach to overcome the problem is to use stochastic gradient descend (SGD) for optimization; alternatively, Yu et. al. proposed a block optimization approach that solve the SVM problem in batch manner [9]. The approach shows better performance on several datasets, and the training time is comparable with SGD solver.

The most popular dataset for large scale visual recognition is ImageNet [2]. Unlike the Yahoo! Grand Challenge dataset, ImageNet is a dataset with large number of concepts, where each concept contains relatively small number of images with higher visual consistency. Based on the



(a) Linear SVM (b) k-Nearest Neighbor (kNN)

Figure 1: The AP of each tag with different classification models using Dense SIFT. Tags that are ill predicted by linear SVM also have poor AP using kNN classifier (k=10), which indicates that visual content is not discriminative for those tags.

dataset, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [1] is a contest that requires the recognition of 1,000 image categories with a total of 1.2 million training images.

In state-of-the-art large scale image classification systems, high dimensional descriptors based on local features such as Vector of Locally Aggregated Descriptors (VLAD) [5] and Locality onstraint Linear Coding (LLC) [8] have become the standard features. Following [7, 8, 5], we adopt the following four features in our experiments:

- Hessian Affine SIFT + VLAD (HA) [5]
- Grid Color Moment + VLAD (CM) [7, 5]
- Dense SIFT + LLC (Dense) [8]
- LBP + LLC (LBP) [8]

For color moment, we extract the 96 dimensional grid color moment (4×4 grid, 1st and 2nd moment in HSV color space) from dense sampled image patches and use it as local feature to aggregate VLAD feature. For LBP, the 59 dimensional LBP is extracted from 32×32 dense sampled patch as local feature.

3. LINEAR SVM CLASSIFIER

Following state-of-the-art large scale image classification systems [1], we perform 10-classes linear classification and use the normalized decision value to rank the test images. 1-against-all strategy is used for the multi-class classification, and the decision value is first scaled by logistic function and then L1-normalized over the 10-classes. When the training image number from each tag is less than 10k, we use the primal space solver in LIBLINEAR; however, when the training image number is more than 10k, the entire training set is too large to fit into memory, so we use an alternative solver from LIBLINEAR that handles training data by batch [9]. The AP of each tag and overall MAP using Dense SIFT are in fig. 1a, which shows that increasing training data improves the performance.

Following the observation, we examine that given a fixed number of training data, can bagging strategies achieve similar or even better performance than using all training data to train one single classifier. If bagging classifier performs well, the scalability of the training process would become free from the hardware constraints of single machine and more training samples can be used. To answer the question, we perform bagging with 1,000 and 5,000 randomly sampled training image for each base classifier; the output of bagging classifier is the average of 100 and 30 classifiers respectively. To control the exact number of training samples, the random sample is done without replacement The results are shown in table 1, where the significant performance gap between single classifier and bagging classifier with the same number of training samples indicates that we should train the base classifier with as many training sample as possible.

To further improve the performance, we perform multi-features fusion. We use late fusion strategy by simply add the decision score of different classifiers, and the resultant performances are in fig. 2. The performances are significantly improved by multi-features fusion, although ill predicted tags such as "travel" still has AP lower than 0.3. In fact, the relative performance of tags are the same over all features and fusion results, which means that "hard tags" are general for all features. Therefore, we turn the focus to exploiting the power of classification model next.

Table 1: MAP with bagging algorithm. The training sets are randomly sampled without replacement to ensure the total number of training samples. Using the same number of training samples, the performance of bagging classifier is inferior to single classifier trained with all samples.

1k samples Single Classifier	1k samples 100 Classifiers	100k samples Single Classifier
0.304	0.319	0.364
5k samples Single Classifier	5k samples 30 Classifiers	150k samples Single Classifier
0.330	0.345	0.366

4. K-NEAREST NEIGHBOR CLASSIFIER

The main difficulty of the dataset is claimed to be the intra-class diversity, where single classifier might not be able to capture the visual appearance of all the sub-classes under the same tag. Theoretically, NN based classifiers can overcome the problem by nature, because the algorithm requires

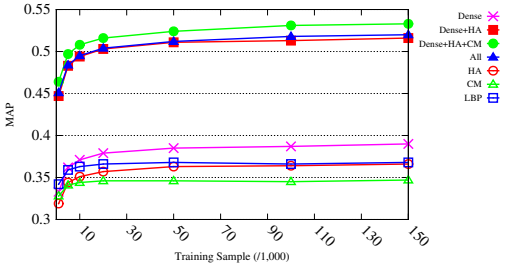


Figure 2: Multi-features fusion. Performance is significantly improved with late fusion which simply average the prediction score of different classifiers.

Table 2: MAP of late fusion on both features and models. By fusing different models, the overall MAP shows significant improvement over single model and feature. Note the MAP is computed without tag "2012."

	Dense	Dense+HA	Dense+HA+CM
SVM	0.409	0.543	0.553
kNN	0.406	0.517	0.568
Both	0.413	0.541	0.582

only the existence of similar samples with correct tag instead of the intra-class similarity. It is also shown that kNN classifier performs well when the dataset is large enough [3]. Therefore, we focus on classification by retrieval in this section.

When we apply kNN classifier directly on the dataset, we found that the retrieval results are dominant by tag "2012." Combined with the poor AP of tag "2012" with linear classification, we believe the visual content of "2012" is so diverse that it is actually a superset of other tags. To overcome the problem, we removed images with tag "2012" from the training set in the following experiments (test set is left unchanged).

The experiment settings are as followed. For the kNN classifier, L2 distance is used for retrieval, and the retrieval depth k is set to 10. The raw decision score of test image i on tag t is defined as

$$\tilde{i}_t = \sum_{k=1}^{10} \frac{\delta_{t_k}^t}{d_{ik}^2}, \quad (1)$$

where t_k is the tag of the kth nearest image and d_{ik} is the distance between the two images. $\tilde{i}$ is then L1-normalized, where the normalized decision score is used to rank test images, following the procedure in linear classification. The results are shown in fig. 3.

We can see from the results that kNN generally performs worse than linear SVM, especially when the number of training samples is small. However, when the training set grows larger, the performance difference becomes smaller; for color moment and LBP, kNN even outperforms linear SVM when the training set is large enough. More importantly, while the performance of linear SVM shows only minor improvement with more than 20k training samples per class, the performance of kNN doesn't seem to saturate even with 150k training images per class. Therefore, we believe that kNN has the potential to outperform linear SVM using all features with more training samples per class.

We also show the AP of each tag in fig. 1b. We can see that the relative performance of each tag is similar to that of linear SVM, which indicates that tags such as "London," "people" and "travel" are indeed less discriminative using visual content regardless of the model.

Finally, we perform late fusion on linear SVM and kNN. Because kNN performs well only with large training set, we only show the result using the entire training set in table 2. By combining both model and three features, the MAP can reach nearly 0.6.

5. PREDICTION WITH META DATA

In previous section, we ignore tag "2012" because of its intra-class visual diversity and visual content overlap with other tags. Combined with the tag name, we postulate that the tag was given based on the context instead of the content of images. More specifically, we argue that the tag "2012" was given because of the taken time of the photo. To justify the argument, we perform experiments as follow.

Table 3: The result of prediction tag "2012" with upload year. Because the prediction does not require any training, the experiment is perform on both training and test set. The AP of purely using upload year predictor is better than the optimal AP using visual content.

	Precision	Recall	AP
Training	0.530	0.966	0.470
Testing	0.529	0.966	0.470

First, we predict tag "2012" using only the upload time. More precisely, we predict positive for tag "2012" if the image is uploaded in 2012 and negative otherwise. The results are shown in table 3. We can see the results on training set and test set are almost the same, and the precision is 5 times higher than random guess; more importantly, the recall is close to 1.0. To further compare the AP, we randomly permute the image order of positive and negative images respectively and take the 10 times average. The AP is higher than the optimal result we can achieve using visual content.

Next, we examine whether visual content can further improve the result. Because the recall of upload year prediction is close to 1.0, we focus only on positive images. For the training set, we only retain images uploaded in 2012, and we keep the ratio of positive and negative to 1:1; for testing, we randomly permute the negative images and use the classifier to rank the positive images. HA feature is used for the experiments. The MAP with different numbers of training images are in fig. 4. By combining meta information and visual content, the originally ill predicted tag "2012" can achieve AP higher than the overall MAP.

The above results indicate that some tags are given mainly due to the context, rather than the visual content of the image. We believe the same argument apply to other tags such as "London" that are not well predicted by pure visual content. Therefore, the meta information of an image is crucial for tag prediction.

6. MID-LEVEL FEATURES

In this section, we present our experiments on two mid-level features. By casting images into a high level semantic space, we believe mid-level features are more resistant to intra-class visual diversity, compared with ordinary low level visual features. In practice, we examined human face feature and "Concept Bank" feature in our experiments.

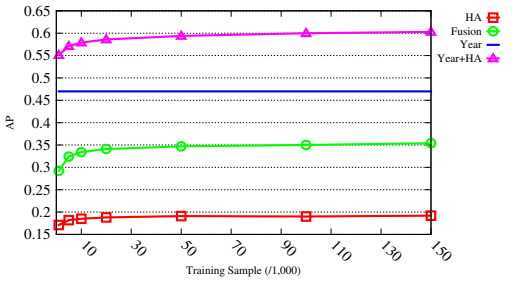


Figure 4: AP for tag "2012". We can see the meta information significantly outperform visual content, while combining both shows further improvement.

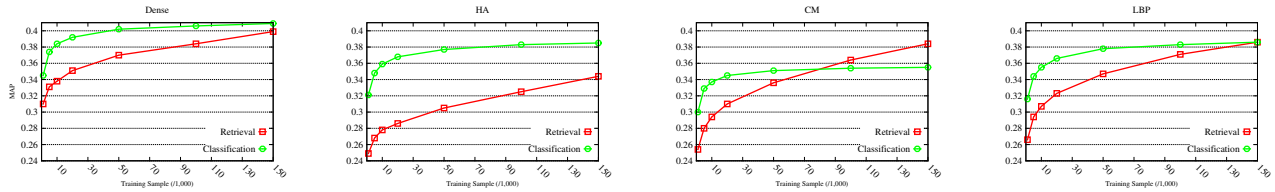


Figure 3: Result of kNN classifier. Although the perform is worse than that of linear SVM under small training set, the performances are comparable when all training data are used. More importantly, kNN classifier shows higher potential of improvement with more training samples, which suggests it would eventually outperform linear SVM.

6.1 Human Face Feature

Motivated by the success of upload year predictor for tag "2012," we try to improve the prediction of tag "people" with face detector. In practice, we use the Omron face detector for face detection. Unlike upload year, however, the existence of human face is not a good indicator for tag "people." In fact, more than half of the training images contain human face except tag "beach" and "sky," and human face are actually more common in tag "music" and "wedding" than in tag "people." Therefore, the existence of human face is not helpful for the prediction of tag "people."

6.2 Concept Bank Feature

Inspired by "Object Bank" [6], we cast each image onto a set of "concepts" and use the response of the concepts as feature for tag prediction. In practice, we randomly select 136 concepts from ILSVRC2012 for the concept space; for each concept, we use all the training data (about 1,000 images) in ILSVRC2012 as positive samples and randomly select the same number of images from the training set of Yahoo! dataset as negative samples, using Hessian affine feature. A logistic regression predictor is trained for each concept, and the normalized response on the predictors is used as the new concept bank feature.

Using the new concept bank feature, we perform logistic regression with the same setups as previous linear SVM classification. For comparison, we perform the same experiment on 136 dimension signatures generated by random projection. The results are shown in fig. 5. The performance of concept bank is significantly better than random projection even if the concepts are selected randomly, which justify the efficacy of concept bank feature. Although concept bank feature does not perform as good as original feature, the training time and feature size is significantly reduced because the feature dimension is reduced from 8,192 to 136. It is worth noting that the performance saturates around 5k of training samples per tag, which is much faster than the original feature. We believe extending the concept space can further improve the performance and is a promising direction for the task.

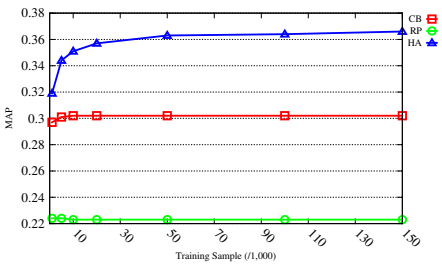


Figure 5: MAP of concept bank (CB) feature. We train 136 concepts randomly selected from ILSVRC2012 and represent each image using the response on the concepts. Hessian affine (HA) feature is used for concept detection. CB significantly outperforms random projection (RP) under same dimension (136). Although the original HA still performs the best, CB significantly reduce feature dimension (8192→136), which in turn reduce data size and training time.

7. CONCLUSIONS

In this paper, we present our evaluation and analysis on the Flickr-tag Image Classification dataset. Our results show that combining multi-features and models can significantly improve the tag prediction performance. However, our analysis also shows that some tags are given due to the meta information of the image instead of its visual content. The most obvious example is the tag "2012," which is better predicted by the upload year than the visual content. For such tags, visual content along can not yield satisfactory performance and meta information is essential for prediction. In fact, we believe the main difficulty of the dataset comes from these visually indiscriminable tags rather than the diverse visual content.

We also examined the "Concept Bank" feature, which cast each image into a concept space representation for classification. Although the new feature does not perform as well as the original feature in our experiments, it significantly reduced the feature dimension and training time while the performance is nearly 50% better than random projection. By increasing the concept space and possibly with selection, the "concept bank" feature has great potential on the task.

8. REFERENCES

[1] A. Berg et al. Large scale visual recognition challenge 2010, 2010.

[2] J. Deng et al. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[3] J. Deng et al. What does classifying more than 10,000 image categories tell us? In *ECCV*, 2010.

[4] R.-E. Fan et al. LIBLINEAR: A library for large linear classification. *JMLR*, 9, 2008.

[5] H. Jégou et al. Aggregating local descriptors into a compact image representation. In *CVPR*, 2010.

[6] L.-J. Li et al. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.

[7] F. Perronnin et al. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[8] J. Wang et al. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[9] H.-F. Yu et al. Large linear classification when data cannot fit in memory. In *KDD*, 2012.