Active Learning Based Image Annotation

Priyam Bakliwal C. V. Jawahar
IIIT-Hyderabad

***Abstract*—Automatic image annotation is the computer vision task of assigning a set of appropriate textual tags to a novel image. The aim is to eventually bridge the semantic gap of visual and textual representations with the help of these tags. This also has applications in designing scalable image retrieval systems and providing multilingual interfaces. Though a wide varieties of powerful machine learning algorithms have been explored for the image annotation problem in the recent past, nearest neighbor techniques still yield superior results to them. A challenge ahead of the present day annotation schemes is the lack of sufficient training data. In this paper, an active Learning based image annotation model is proposed. We leverage the image-to-image and image-to-tag similarities to decide the best set of tags describing the semantics of an image. The advantages of the proposed model includes: (a). It is able to output the variable number of tags for images which improves the accuracy. (b). It is effectively able to choose the difficult samples that needs to be manually annotated and thereby reducing the human annotation efforts. Studies on Corel and IAPR TC-12 datasets validate the effectiveness of this model.**

I. Introduction

With the outburst of social media there is a tremendous increase in unannotated raw image data getting archived everywhere. One often needs textual descriptions for these images to build scalable and semantically meaningful access methods. This needs new approaches for scalable and automatic image annotation. Automatic image annotation (here after referred to as simply image annotation) aims at assigning a set of appropriate textual tags to a new test image without explicitly understanding (eg. object detection, image categorization) the images. (See Figure 1 for an example.) Since there are many possible tags for a single image, the problem is very different from that of image classification/categorization. Annotation is essentially a multilabel classification problem. Usually, the annotation data sets have a large vocabulary of tags/labels and the objective is to pick and predict the most appropriate subset.

Many powerful methods (eg, based on CRF, HMM, SVM) were tried for this task in the past. However, in 2008, Makadia *et al.* [6] demonstrated that a simple nearest neighbor method can yield superior results on the popular data sets. In later years, Guillaumin *et al.* [2] as well as Verma and Jawahar [1] extend this method. They used metric learning [1], [2] and also refined the nearest neighbor computation process [1]. Even many of the later attempts with modern machine learning schemes [7], [8] did not yield results that are superior to the two pass nearest neighbor (2-PKNN) [1] over the popular databases such as Corel and IAPR TC-12.

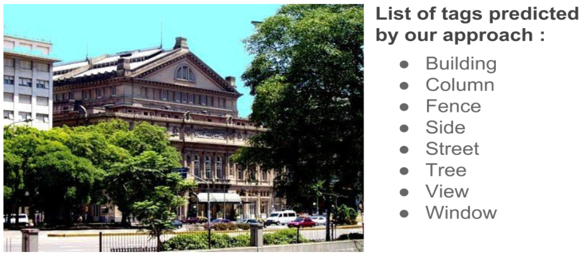


Fig. 1: Automatic image annotation task. Result of our approach on an image from IAPR-TC-12 data set. Our method predicts a set of appropriate tags with minimal amount of training data.

In this work, our objective is to obtain higher performance with minimal amount of training data. We achieve this with the help of active learning and automatically selecting images that are worthy of human labeling. As we demonstrate in the next section, performance of the previous algorithms [1], [6] heavily depends on the number of labeled examples available for training. However practically, it is extremely difficult to get labeled examples. This can be overcome using active learning, where only a selected set of images is manually labeled to improve the performance. A general active learning process consists of two stages: (i) Learning algorithm and (ii) Sample selection algorithm. The performance of an active learning model highly depends on the sample selection scheme. As pointed out in [5] the most common criteria for sample selection are uncertainty, diversity, density and relevance.

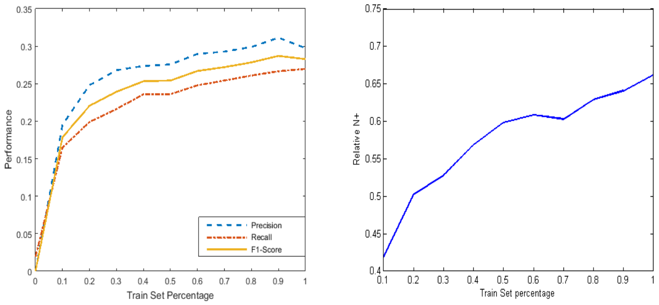
In this paper, we select the harder examples for manual annotation automatically. We use uncertainty for the sample selection. We handle the class imbalance problem by selecting a subset of training examples with similar frequency of labels rather than complete training set. This ensures that the most labeled classes do not dominate the results. We use a KNN classifier for its simplicity. We refine the 2-PKNN [1] scheme further for the active learning by enabling this scheme to predict variable number of tags for each of the images. For each image, we retain all the labels that satisfy a minimum threshold. An image is then considered for active learning based on the prediction score.

We perform our experiments on two benchmark datasets, Coral and IPRTC-12 and show that even with only 10% of train data and a proper selection of the additional examples to annotate, we can achieve a performances that is typically obtained by 80% of the training data. We report experimental results demonstrating the utility of our approach.

II. Active learning for Image Annotation

The focus of this work is to reduce the amount of training data used in annotation task with simultaneous increase of performance. We achieve this by formulating the problem in

978-1-4673-8564-0/15/$31.00 ©2015 IEEE



(a)

(b)

Fig. 2: Dependency of performance on number of train images. Note that the curves are not saturating and the annotation data needs many more examples.

an active learning framework. The performance of a typical annotation algorithm depends on two factors: (a) feature representations and the underlying similarity metric, (b) size and quality of the training data set. It is self evident from Fig: 2 that performance of the annotation is directly proportional with size of the training set, i.e., larger the training data, better the performance. However, many of these training examples are "redundant" and do not add any value to the learning task. Identifying these examples can save the manual efforts which is among the primary goals of the current work.

Consider the training set $\mathcal{T} = \{(I_1, L_1), \ldots, (I_t, L_t)\}$ where, $I_1, \ldots, I_t$ are the images and $L_1, \ldots, L_t$ are its respective label sets. Let $\mathcal{Y} = \{y_1, \ldots, y_l\}$ denotes the vocabulary of labels. Given an unannotated image J, our aim is to assign multiple labels L_j associated with it.

Joint Equal Contribution (JEC) [6] treats image annotation as a retrieval problem. The technique uses a greedy algorithm to find image labels, from its nearest neighbors, found using low-level image features. The method is quite simple and intuitive which strongly claims that a simple combination of basic distance measures defined over commonly used image features can effectively serve as a baseline method for multilabel image annotation tasks. Despite its simplicity, at the time of its proposal, JEC held the state of the art on all benchmark annotation datasets [6]. However, it fails to consider class imbalance and weak labeling issues explicitly. The former problem is tackled by 2-PKNN algorithm [1] by using annotation performance in terms of mean recall. It uses 2-phase nearest neighbor model to predict image annotations. Given an unseen image, the algorithm identifies its semantic neighbors, in the first phase, for all the labels. And in the second phase the selected samples are used to predict the tags.

In our proposed method, for identification of semantic neighbors, we pick K images for each semantic label in the vocabulary that are most similar to J. In this way we ensure that each label appears at least K times in the training data. Let T_{Jx} be the set of K images, that are most useful in predicting the score of label y_x for image J. These neighbors incorporate image-to-label similarities. Once all T_{Jx} are determined, we merge them to form a final subtrain set specific to image J. It can be easily seen that this setting addresses the class imbalance issue by choosing each label to appear at least K times in train data. With this train data, we apply a weighted nearest neighbor algorithm to assign importance to the labels based on image similarity. In this way, we determine the scores for each label y_x for the image J.

Most of the previous works [1], [2] keep a static size of tags to be predicted (generally 5). This helps to retain higher precision. However, in this algorithm we dynamically determine this number. For each test image our algorithm predicts the score for every label in the dictionary. We assign all the labels that satisfy the 'score threshold'. For calculating the score threshold, we randomly sample a set of train images and use our algorithm to find the scores for each label. Later, scores for all the labels present in train ground truth is used to calculate the mean score (S_m), this is then used to calculate the score threshold, to determine the presence of label in an image. The final threshold can be calculated as:

$$\tau = S_m - \epsilon \tag{1}$$

where, ϵ is the tolerance parameter that decides tradeoff between precision and recall. Large value of ϵ leads to smaller τ and thus more number of labels will be assigned to the image leading to higher recall.

For an unseen test image J, the 2PKNN algorithm is used to determine score for each of the labels. All labels L_i with score S_{Ji} greater than τ are kept as predicted labels for the image J.

For active learning we have to decide the images that need to sampled from the test set. As mentioned earlier we use uncertainty principle to select these images to maximize the performance. We take mean of the scores of the predicted tags to decide the prediction confidence for the image.

$$\theta_J = \frac{1}{n} \sum S_{Ji}, \forall S_{Ji} \geq \tau \tag{2}$$

We greedily select $X\%$ of images with the lowest mean score from the total test set. These are the images used for active learning which are combined with existing training set. We recalculate the scores for remaining test images and assign its tags based on the updated scores. This summarizes one iteration of active learning. Thus, with every iteration we improve the model and predict the tags with higher accuracy. Algorithm 1 explains the proposed method algorithmically.

III. Experiments and Results

A. *Data sets, representation and evaluation measures*

We have used two popular image annotation data sets namely, Corel and IAPRTC-12. Corel data set was first used in [3] and since then it has been one of the benchmark data sets in image annotations. IAPRTC-12 data set was first used for cross-lingual retrieval in [4]. In this data set each image is described in detail; but for image annotation task, only nouns are extracted as labels. Table I describes the basic characteristics of the data sets used for experimentation.

For performance analysis of this algorithm, we have used 15 distinct descriptors as used in [2]. These include both global and local features. SIFT and Hue based descriptors covers local features of an image whereas GIST and histogram based

Algorithm 1 Active learning algorithm

```
Input: Trainannotations, DistanceMatrix, $\tau$, $X$
Output: AssignedTestLabels
for $i = 1$ to numOfTestImages do
    Select Train Subset.
    Calculate the scores using KNN and Train Subset.
    for $j = 1$ to numOfLabels do
        if $(S_{ij} \geq \tau)$ then
            Add j to AssignedTestLabels and
        end if
    end for
    $meanScore_i$ = $mean(Score(AssignedLabels))$
end for
Choose X images with lowest meanScore
Ask for user annotations for these images.
Add them to train set.
for $i = 1$ to numOfRemainingTestImages do
    Recalculate Train Subset and AssignedTestLabels
end for
```

Data set	No. of Labels	No. of Train Img	No. of Test Img
IAPRTC-1	291	17665	1962
Corel	268	4500	499

TABLE I: Details about the data sets used for the experiments

descriptors encode the overall characteristics of the image. Distances between the features are calculated following the earlier research [2]. L1 measure is used for colored histograms, L2 for GIST and χ^2 for the SIFT and Hue descriptors.

To compare the result with earlier methods, we have used similar evaluation measures as in [1]. Given an unseen image we predict the label set using the score and the threshold. The evaluation measures include 1) Average precision per label 2) Average recall per label 3) Average F1-score per label and 4) Normalized N+ Score. If for a label l_x there are i_g images in the ground truth and i_p images are predicted for it then, for label l_x, precision P_x and recall R_x are defined as:

$$P_x = \frac{i_g \cap i_p}{i_p} \qquad \text{and} \qquad R_x = \frac{i_g \cap i_p}{i_g} \tag{3}$$

The average of these precision and recall values over $\mathcal{Y}$ gives us mean precision and recall for the data set. To analyze the trade-off between precision and recall we also calculate mean F1-Score as F1 = 2PR / (P + R).

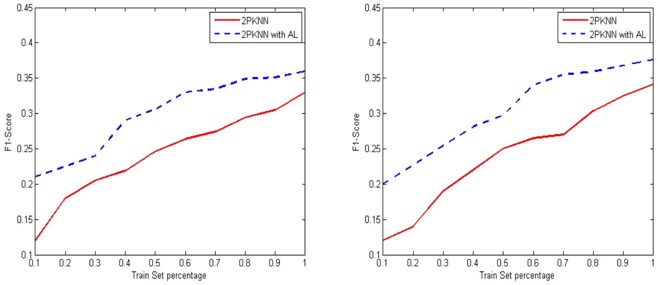
N+ is the number of labels that are assigned correctly to at least one image. But, instead of using absolute value for N+, we are using relative N+, i.e. R_{N+} = N+ Score / # of Labels.

B. Empirical Results

To compare the image annotation performance with other methods, we use the predefined training and test sets used in the past [6], [2], [1]. The summary or our results as well results for previous models are summarized in Table II. We have used active learning along with 2PKNN algorithm to show that there is a significant performance gain even with using only 10% of data for active learning.

	Corel 5K				IAPR TC-12			
Method	P	R	F1	R-N+	P	R	F1	R-N+
JEC[6]	27	32	29.3	53	28	29	28.5	85
TagProp[2]	31	37	33.7	56	48	25	32.9	78
KSVM[8]	32	42	36	68	47	29	36	92
2PKNN [1]	39	40	39.5	68	49	32	38.7	94
2PKNN with AL	**45**	**46**	**45.5**	**80**	**56**	**32**	**41**	**97**
2PKNN[1]	35	32	33	63	43	29	34	93
2PKNN with AL	**36**	**34**	**35**	**67**	**47**	**31**	**37**	**95**

TABLE II: Performance comparison among different image annotation methods. 2PKNN with AL (this work) uses 10% of test data for active learning. The top section shows performance calculated on full test data and the bottom section shows performance for only 90% test data, excluding 10% of actively learned test data.



(a)

(b)

Fig. 3: Change of F1-Score with increase in number of train images for both Corel and IPAR TC-12 datasets.

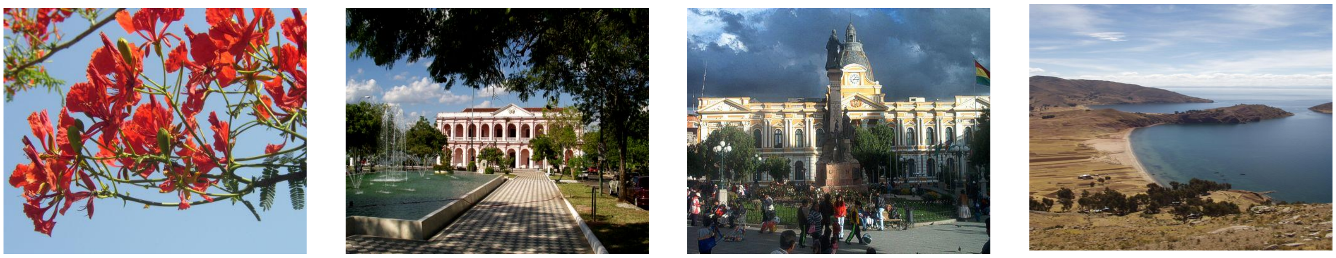
Now we study the performance variation with the size of the training set. It is a known fact that the performance of nearest neighbor algorithms increases with increase in training data. We have calculated the performance change with gradually increasing the size of training data for both 2PKNN algorithm and active learning algorithm. The active learning percentage was kept as 10. It is clear from Fig 3 that learning rate is comparatively high for Active learning method.

In this experiment we gradually increased the quantity of test images picked for active learning. Also, to show that we effectively pick up most uncertain samples, we randomly sample the same quantity of test images and added to the train set to calculate performance. Fig: 5 clearly depicts that the sample selection strategy of our proposed algorithm outperformed with extremely better results.

Use of fixed length annotation method faces disadvantages in the cases where annotation length is either comparatively less or more than the mean annotation length. We have shown in Fig 3 that we are able to predict better tags for images with as low as 1 tag as well as for images with 11 tags. It is also evident from $Image_1$ that we are handling missing label issue.

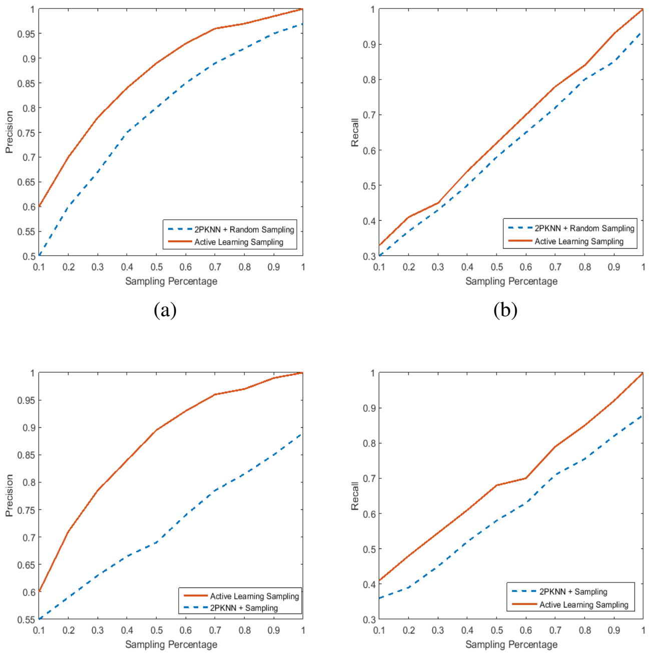
C. Discussions

We have clearly shown with our experiments that the active learning algorithm we propose has multiple advantages in terms of performance gain as well as reduction in requirement of train data. On one hand with active learning we reduce the



Tree	Tree, Fountain	Building, Center, Clock, Flag, Column, Fence, Lot, Statue, People, Square, Window	Bay, Beach, Cloud, Slope, Stone, Tree, View
Tree, Flower, Fruit, Sky, Life	Building,Fountain, Tree, Footpath, Paving	Building, Carpet, Lot, Stair, Tree	Stone, Jetty, Lake, Summit, Mountain
Tree, Flower, Sky	Tree, Building, Fountain	Building, Column, Carpet, Lot, People, Stair, Tree, Fence	Bay, Jetty, Mountain, Tree, Beach, Cloud, Lake, Stone, Summit

Fig. 4: Qualitative Results are shown on IAPR TC-12 dataset. The second row respresents tags present in ground truth. Third and fourth rows represents tags predicted by 2-PKNN and our algorithm respectively. Prediction of 'Sky' and 'Flower' in image 1 shows that we handle weak labeling. Also the prediction of 3, 3, 8 and 9 tags for first, second, third and fourth image respectively shows the effectiveness of varying length annotations.



(a)

(b)

(c)

(d)

Fig. 5: Performance graphs to depict effective selection of active learning samples. The first two images are for Corel dataset and later once are for IPARTC dataset. We have used random sample selection for 2PKNN algorithm and the selected samples are added to the train set vs our selection strategy to compare the performances.

time and cost required for getting train data and on the other hand with dynamic decision of tag length and greedy selection of most uncertain images we improve the performance.

IV. Conclusion

We have proposed an active learning based image annotation model. This model combines nearest neighbor approach along with active learning to annotate an unseen image. One of the most important issues in auto image annotation is deciding the annotation length, and our algorithm gives a simple and effective solution to this issue . Also, it is clear from the results that the algorithm is effectively able to pick hard samples from the test data, so as to dramatically improve the overall accuracy of the test samples even with extremely less actively annotated images. Thus with minimal user efforts, we are able to outperform extremely well. Currently, we have used fixed thresholds for deciding the annotation length. In future, we are planning to learn the thresholds from training data. Also, we can use weighted thresholds based on the properties of the predicted annotation labels.

References

[1] Y. Verma and C. V. Jawahar, *Image Annotation Using Metric Learning in Semantic Neighborhoods*. In: ECCV 2012.

[2] M. Guillaumin, T. Mensink, J. Verbeek and C. Schmid, *TagProp: Discriminative Metric Learning in Nearest Neighbor Models for Image Auto-Annotation*. In: ICCV, 2009.

[3] P. Duygulu, K. Barnard, N. de Freitas and D. Forsyth, *Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary*. In: ECCV, 2004.

[4] M. Grubinger. *Analysis and Evaluation of Visual Information Systems Performance. .* PhD thesis, Victoria University, Melbourne, Australia, 2007.

[5] M. Wang and X.S. Hua. *Active learning in multimedia annotation and retrieval: A survey..* In: ACM Trans, 2011.

[6] A. Makadia, V. Pavlovic and S. Kumar. *A new baseline for Image Annotation*. In: ECCV, 2008.

[7] M. M. Kalayeh, H. Idrees and M. Shah. *NMF-KNN: Image Annotation using Weighted Multi-view Non-negative Matrix Factorization*. In: CVPR, 2014.

[8] Y. Verma and C. V. Jawahar *Exploring SVM for Image Annotation in Presence of Confusing Labels*. In: BMVC, 2013.