# EFFICIENT SELECTION OF INFORMATIVE AND DIVERSE TRAINING SAMPLES WITH APPLICATIONS IN SCENE CLASSIFICATION

*Sujoy Paul, Jawadul H. Bappy and Amit K. Roy-Chowdhury*

Department of Electrical and Computer Engineering, University of California, Riverside, CA 92521

## ABSTRACT

The huge amount of time required to construct a set of labeled images to train a classifier has led researchers to develop algorithms which can identify the most informative training images, such that labelling those will be sufficient to achieve a considerable classification accuracy. In this paper we focus on choosing a subset of the most informative and diverse images based on which the classification model can be learned efficiently. The size of the subset to be chosen is determined by the available budget for manual labeling. Although the problem of identifying the informative images can be solved by active learning algorithms, it will require a set of labeled images for initial model construction, which is not required in our method as we identify the best samples at one shot. We incorporate the concepts of strong and weak teacher to help the learner to learn the model efficiently with limited budget for manual labeling. We perform rigorous experiments on two challenging scene classification datasets to demonstrate the effectiveness of our algorithm.

***Index Terms***— scene classification, informative sample selection

## 1. INTRODUCTION

Scene classification is one of the most fundamental problems in computer vision, gaining interest of many researchers over the last decade. Unlike object classification, scenes are generally composed of multiple entities, with different shape, size, color, exposure and interactions between them. The learner in a traditional scene classification algorithm [1, 2, 3] learns over a lot of labeled images. With the huge corpus of images available today, it becomes unrealistic to have all the data labeled beforehand. Moreover, the manual process of labeling all the samples is a tedious job. In this paper, we concentrate on solving this problem, by choosing the most informative samples for manual labeling, thus making efficient use of the available budget for labeling.

Recently, there have been some works [4, 5, 6] addressing the problem involved in manually labeling all data for training a learner. There has been some success in solving this problem because, not all training examples are equally informative [7]. Most of these active learning algorithms begin with a set of labeled images and then update the model based on the initial model through manually labeling the most informative unlabeled samples. This approach is effective in scenarios where the input data is not available at the outset e.g. streaming inputs. On the other hand, there may be scenarios where all the images is available at once and we need to identify the best examples to label.

The proposed method addresses this problem. Instead of labeling all the images, we choose only the most informative and diverse subset of images to train the learner to achieve considerable classification accuracy on the test set. The amount of images to be chosen for manual labeling is dependent on the available budget for manual labeling. Our proposed method complements active learning approaches. Most active learning methods assume that there is a set of labeled instances for initial model construction. The proposed method avoids this assumption and chooses the best subset of images at one shot, thus saving the computational time involved in the iterative process of selecting the informative images in most active learning frameworks.

**Related Work:** An overview of the commonly used techniques for active learning, i.e., efficiently choosing the training instances, may be found in [8]. The idea of selecting the best training instances have been successfully implemented for computer vision problems like tracking [9], activity recognition [10], object detection [11], etc. Recently, an algorithm based on entropy and KL-Divergence [12], was proposed for batch mode active learning. But, most of these works assume that there is a set of labeled images for initial training.

Various confidence measures of the learner on unlabeled instances have been used for image classification [13]. In [14], an active learning framework for scene classification was proposed, which have query for labeling at the scene as well as the object level. A framework combining information density measure and uncertainty measure for query selection was proposed in [5], which requires a labeled set of images for initial training. Batch mode incremental learning coupled with efficient image selection for training was proposed in [15]. The concept of best-vs-second best as an uncertainty measure was used in [16] to choose the best training instances for image classification. Recently, the problem of image classification was addressed using Laplacian Sparse Coding [17]

**Proposed Framework:** A pictorial representation of the flow of the proposed framework is presented in Fig. 1. Given a pool of unlabeled images, the framework starts by extract-

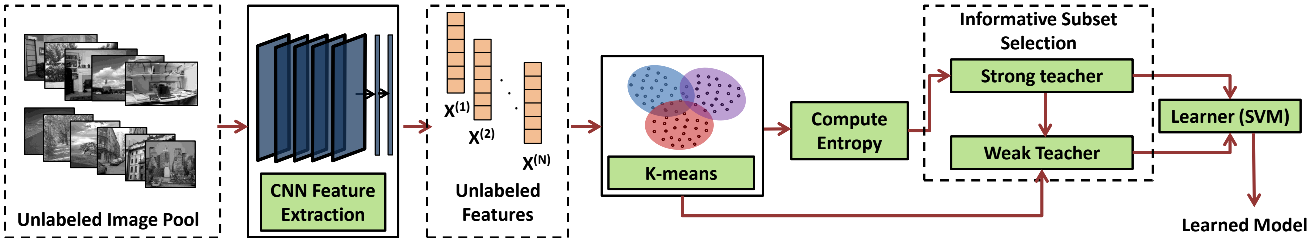


**Fig. 1**: This figure presents our proposed framework, which is comprised of the following stages: CNN feature is extracted from the pool of unlabeled images, k-means is applied on the extracted image features, entropy is used as a measure to identify the most informative images for manual labeling, i.e., strong teacher, thereafter the weak teacher is invoked and finally the model is learned from the set of images labeled by strong and weak teacher.

ing the deep features from them. These unlabeled features are fed to the k-means algorithm, thereby obtaining an estimate of the cluster centroids along with labels of each scene class. Then, using the distance of each image feature from the centroids, we obtain the probability of each image belonging to each scene class. Thereafter, using these probabilities, the uncertainty of each image label obtained from k-means is computed. Considering that each image is independent, the total entropy of all the images in the unlabeled pool boils down to the summation of the individual image entropy. Using this result and to ensure diversity in the chosen set of images, those images having the highest entropy are called for manual labeling, which is the strong teacher. Moreover, we exploit the confidence of the strong teacher and k-means to teach the learner with a weak teacher as discussed in Section 2. We use the SVM with linear kernel as the learner.

## 2. PROPOSED ALGORITHM

Our scene classification algorithm involves two main steps. The first step is the extraction of features from images. The second step is identifying the best training samples and invoking human to label the same. Thereafter training the learner with those samples. The rest of the section discusses these steps in detail.

**Feature Extraction:** We use convolutional neural network (CNN) to extract the features from the images. In our deep network, five convolutional layers and two fully-connected layers are employed. We use pre-trained model *'Places205 Alexnet'* [18] to extract the features from deep network. Finally, we obtain the feature vectors with $4096$ dimension from the last fully-connected ($fc7$) layer.

**Selecting informative images for manual labeling:** After extracting the features from the images, we have a pool of $N$ unlabeled images having feature vectors $\{\underline{x}^{(i)}\}_{i=1}^{N}$. In our algorithm, we incorporate the concept of strong and weak teacher, to choose the best training images for labeling and help the learner to learn the model efficiently. Generally, the strong teacher is human, whose accuracy in labeling is considered to be perfect. The weak teacher on the other hand provides tentative labels. The cost of invoking strong teacher is generally much higher compared to weak teacher and often its budget is specified, i.e., the affordable number of labels that can be provided by the strong teacher.

The goal of our algorithm is to maximize the efficiency of the learner using this limited budget. In the literature, generally a classifier is trained initially, with a small batch of labeled instances. In our method, we don't require any initial batch of labeled images. We start by applying k-means on $\{\underline{x}^{(i)}\}_{i=1}^{N}$, to get an estimate of the boundaries of the classes. Let us consider $\{\underline{C}^{(j)}\}_{j=1}^{k}$ to be the $k$ centroids of the clusters $\{\mathbf{C}_j\}_{j=1}^{k}$ obtained after applying k-means. The probability of $\underline{x}^{(i)}$ belonging to each cluster may be expressed as,

$$\underline{P}^{(i)} = [p(x^{(i)} \in \mathbf{C}_1), p(x^{(i)} \in \mathbf{C}_2), \dots, p(x^{(i)} \in \mathbf{C}_k)] \quad (1)$$

where,

$$p(x^{(i)} \in \mathbf{C}_j) = \frac{\exp(-\alpha||x^{(i)} - \underline{C}^{(j)}||_2^2)}{\sum_{m=1}^{k} \exp(-\alpha||x^{(i)} - \underline{C}^{(m)}||_2^2)} \quad (2)$$

$\alpha$ is a scalar parameter. Using (1) and (2), the entropy for each image can be expressed as,

$$H(\underline{P}^{(i)}) = -\sum_{i=1}^{k} p(x^{(i)} \in \mathbf{C}_j) \log_2(p(x^{(i)} \in \mathbf{C}_j)) \quad (3)$$

which is a measure of the uncertainty in the cluster label assigned by k-means. Assuming that the individual samples are independent of each other, the pairwise mutual information [19] between the samples $I(\underline{x}^{(i)}, \underline{x}^{(j)}) = 0$. Thus, the total entropy of all the samples can be represented as,

$$\begin{aligned} H(\underline{P}^{(1)}, \underline{P}^{(2)}, \dots, \underline{P}^{(N)}) &= \sum_{i=1}^{N} H(\underline{P}^{(i)}) \\ &= \sum_{\underline{x}^{(i)} \in S, |S|=M} H(\underline{P}^{(i)}) + \sum_{\underline{x}^{(j)} \notin S} H(\underline{P}^{(j)}) \end{aligned} \quad (4)$$

where $S \subset \{\underline{x}^{(i)}\}_{i=1}^{N}$, and $M$ is determined by the query budget to be discussed subsequently.

**Diverse subset selection.** It may be noted from (4), that one possible solution of choosing the subset $S$ is such that after labeling those samples by the strong teacher, the total entropy will be minimized. In other words, $S$ may be chosen such that the entropy of those samples in $S$ are among the highest entropy in the entire unlabeled pool of images. But, it may happen that the samples having the highest entropy are cluttered within a small area in the feature space, thus lacking diversity. This will be problematic for the learner, as it will not be able to learn about all the classes efficiently. To avoid this problem, we choose the samples having the highest entropy from each cluster obtained from k-means to constitute the optimal subset $S^*$ which may be expressed as,

$$\begin{aligned} S_j^* &= \underset{S_j, |S_j|=k_j}{\arg\max} \sum_{\underline{x}^{(i)} \in \mathbf{C}_j} H(\underline{P}^{(i)}) \\ S^* &= \{S_1^*, S_2^*, \dots, S_k^*\} \end{aligned} \qquad (5)$$

where $k_j = \frac{(k+M)*L_j}{N} - 1$, $L_j$ being the number of images belonging to the $j^{th}$ class as obtained from k-means. The value of $k_j$ ensures that the number of samples chosen from each cluster of k-means are proportional to the number of samples present in the respective clusters. It may be noted that although this subset of images will not minimize the total entropy in (4) maximally, it will help the learner learn diversified samples. This optimal subset are queried to the human, the strong teacher, to get the correct labels. Intuitively, this approximately boils down to the idea of querying the images near the boundary of dissimilar classes (as shown in Fig. 2), which are in fact the most uncertain images.

**Subset selection for weak teacher.** We can exploit the confidence of the k-means algorithm and the labels provided by strong teacher. There is a high probability that the image features very near to a cluster centroid belong to the same class as that of the centroid. But the labels of the centroids are unknown. We resort to the strong teacher to query the centroids along with the subset chosen in 5. Thus, we need to query $k$ more images to the strong teacher, leading to a total of $k + M$ queries to the strong teacher. Therefore, the total budget of query, i.e. the affordable number of queries to the strong teacher, determines the values of $M$ in (5). After the strong teacher labels the queried images, those having metric $d_{\underline{x}^{(i)}} = \exp(-\beta||\underline{x}^i - \underline{C}^{(i)}||_2) \geq \delta$ are assigned the same label as that of the centroid $\underline{C}^{(i)}$. This is the weak teacher. $\delta$ is generally set to a value high enough to diminish the probability of assigning a wrong label by the weak teacher. Finally, all the images labeled by the strong and weak teacher are used as training images. An example of the data points labeled by strong and weak teacher is presented is in Fig 2.

## 3. EXPERIMENTS

**Objective:** The main objective of the experiments is to analyze to how the proposed method performs with stipulated amount of budget for manual labeling. We use SVM [20] with linear kernel as the learner. We compare our results with three subset selection techniques - entropy based *Uncertainty sampling* [13], *Best v. Second Best* (Bv2B) [15], *Batch Rank*[12]. We have implemented all the algorithms using the same image features and learner used by us. We also compare with three methods which consider the entire dataset to be manually labeled - *Holistic Scene understanding* (HSC) [21], *Optimized Laplacian Sparse Coding* (OLSC) [17] and a baseline *Places* [18] for Scene15 dataset. The sensitivity of the proposed method on the parameter $\delta$ (mentioned in subset selection for weak teacher) is also analyzed.

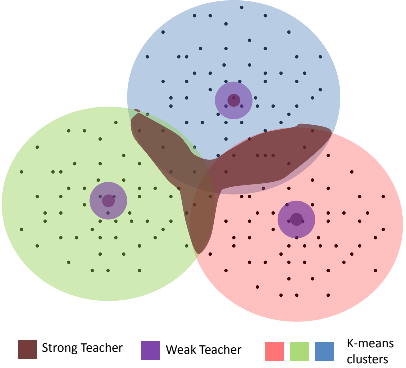
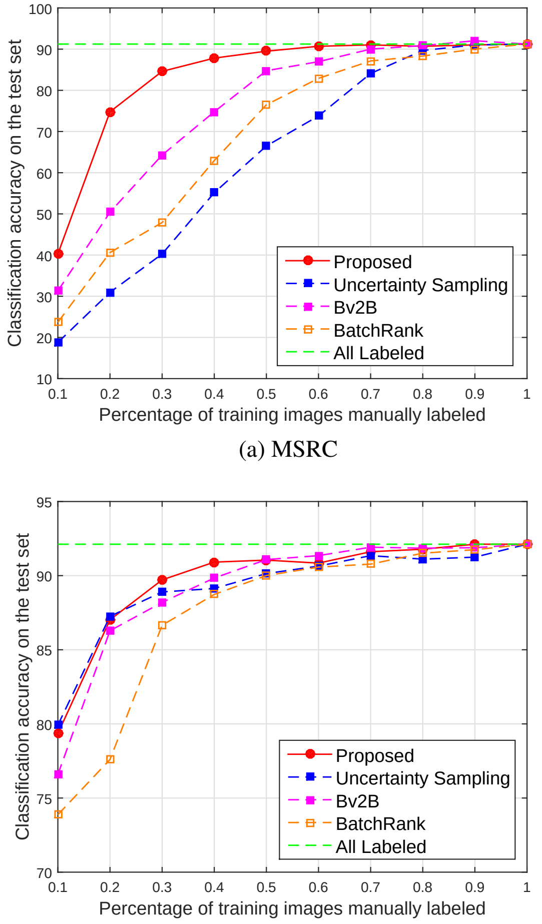


**Fig. 2**: This figure presents an example of the probable area in feature space labeled by strong and weak teacher. Different colors denote the class of each data point as obtained from k-means as well as those labeled by the strong and weak teacher. The other areas may not be labeled.

**Datasets:** In this section, we validate the proposed algorithm by testing it on two widely used scene classification datasets - MSRC [21] and Scene15 [22]. The MSRC dataset consists of 591 images from 21 scene classes. The Scene15 dataset is comprised of 15 scene classes, with 200 to 400 images for each scene category. We randomly choose 100 images per category for training set and the rest for testing.

**Accuracy analysis:** For both the datasets, we have divided the entire dataset into training and testing sets. Then, we consider that the training set is not labeled. Thereafter, we have applied our algorithm to compute the informative images and acquire only their labels. Using these labels, we train the SVM classifier to build a model and obtain the classification accuracy on the test set. For the MSRC dataset, we have considered 5 Fold Cross Validation (5-FCV) to compare with subset selection methods and standard partition as in [21] to compare with the same. In case of Scene15 dataset, we randomly choose 100 images per category for training set and the remaining images for testing. By varying $M$ (mentioned just after (5)), the budget $(k + M)$ varies, and we obtain different sets of informative images. Thus, we obtain different learned models and classification accuracy on test set. These classification accuracy are plotted in Fig. 3. Table 1 presents the comparison with other scene classification methods which consider the entire dataset is manually labeled.

**Sensitivity analysis:** We analyze the sensitivity of the



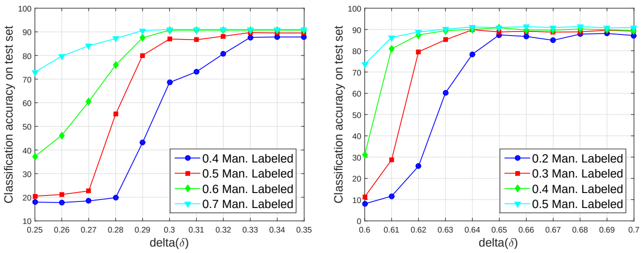
(a) MSRC

(b) Scene15

**Fig. 3**: (a) and (b) present plots for 5-FCV accuracy vs. percentage of training images manually labeled. ’All Labeled’ corresponds to the accuracy of the proposed framework when all the images in the training set is labeled.

proposed algorithm on the parameter $\delta$. For different percentage of manual labeling, we vary $\delta$ from 0.25 to 0.35 for MSRC and 0.6 to 0.7 for Scene15 dataset. After learning the model for each of these values of $\delta$, we obtain the classification accuracy over the test set. These are plotted in Fig 4 for the two datasets.

**Observations and Discussions:** It may be observed from Fig 3 that the proposed method requires much lesser labeled images than the entire training set, to achieve almost similar classification accuracy. Moreover, the proposed method performs considerably well compared to the other subset selection methods, especially for the MSRC dataset. In Table 1, it may be observed that the proposed method requires less than $40\%$ of the entire training dataset to obtain similar classification as HSC, which labels and uses the entire training set to train (Note: standard partition [21] is used for this comparison). It may also be noted from Table 1 that for the Scene15 dataset, the proposed method requires only 40% manual labeling to achieve comparable and better accuracy than [18] and OLSC [17] respectively.



(a) MSRC

(b) Scene15

**Fig. 4**: This figure presents the variation in classification accuracy on the parameter $\delta$ for the two dataset.

The parameter $\delta$ sets a boundary around the centroid, within which all feature vectors can be labeled to have the same label as that of the centroid, i.e. weak teaching. As $\delta$ is decreased, the bound increases, thus labeling higher number of data points to belong to the same class as that of the centroid. But this increases the risk in the label provided because as the bound increases, feature vectors from dissimilar classes may be labeled to belong to similar class. This explains Fig. 4, where the classification accuracy decreases as $\delta$ is decreased. Thus, $\delta$ should be high value such that weak teaching is offered only when the teacher is confident.

**Table 1**: Comparison with scene classification methods which consider the entire dataset is manually labeled

| Dataset | Proposed<br>Acc. (% Man. Lab.) | Other Algo.<br>Acc. (Name) |
|---|---|---|
| MSRC | 84.71 (40)<br>87.84 (70)<br>90.19 (100) | 80.6 (HSC [21]) |
| Scene15 | 90.91 (40)<br>91.61 (70)<br>92.12 (100) | 91.59 (Places [18])<br>90.67 (OLSC [17]) |

## 4. CONCLUSION

In this work, we proposed a framework to choose the most informative images, such that only labeling those will help the learner learn the classification model efficiently with limited budget for manual labeling. Future works will be directed towards exploiting the interrelationships between images which will help the unlabeled images to gain information from the labeled images. Moreover, the proposed method can be framed to select only the informative images to label for initial model construction in most active learning algorithms.

**Acknowledgments.** This work has been partially supported by NSF grant IIS-1316934 and ONR N00014-15-C-5113.

## 5. REFERENCES

[1] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification via plsa," in *ECCV*, pp. 517–530. Springer, 2006.

[2] D.Song and D.Tao, "Biologically inspired feature manifold for scene classification," *Image Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 174–184, Jan 2010.

[3] J.Shi, X.Li, and Y.Dong, "How to represent scenes for classification?," in *Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit and International Conference on*, July 2015, pp. 191–195.

[4] A.J. Joshi, F. Porikli, and N.P. Papanikolopoulos, "Scalable active learning for multiclass image classification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2259–2273, 2012.

[5] X. Li and Y. Guo, "Adaptive active learning for image classification," in *CVPR*. IEEE, 2013, pp. 859–866.

[6] M. Wigness, B.A. Draper, and J.R. Beveridge, "Efficient label collection for unlabeled image datasets," in *CVPR*, 2015, pp. 4594–4602.

[7] A. Lapedriza, H. Pirsiavash, Z. Bylinskii, and A. Torralba, "Are all training examples equally valuable?," *arXiv preprint arXiv:1311.6510*, 2013.

[8] B. Settles, "Active learning literature survey," *Univ. of Wisconsin, Madison*, vol. 52, no. 55-66, pp. 11, 2010.

[9] C. Vondrick and D. Ramanan, "Video annotation and tracking with active learning," in *Advances in Neural Information Processing Systems*, 2011, pp. 28–36.

[10] M. Hasan and A.K. Roy-Chowdhury, "Context aware active learning of activity recognition models," in *ICCV*, 2015, pp. 4543–4551.

[11] S. Vijayanarasimhan and K. Grauman, "Large-scale live active learning: Training object detectors with crawled data and crowds," *International Journal of Computer Vision*, vol. 108, no. 1-2, pp. 97–114, 2014.

[12] S. Chakraborty, V. Balasubramanian, Q. Sun, S. Panchanathan, and J. Ye, "Active batch selection via convex relaxations with guaranteed solution bounds," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 10, pp. 1945–1958, 2015.

[13] D.Tuia, F.Ratle, F.Pacifici, M.F.Kanevski, and W.J.Emery, "Active learning methods for remote sensing image classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 47, no. 7, pp. 2218–2232, 2009.

[14] X. Li and Y. Guo, "Multi-level adaptive active learning for scene classification," in *ECCV*, pp. 234–249. Springer, 2014.

[15] X. Li, R. Guo, and J. Cheng, "Incorporating incremental and active learning for scene classification," in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*. IEEE, 2012, vol. 1, pp. 256–261.

[16] A.J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *CVPR*. IEEE, 2009, pp. 2372–2379.

[17] L. Chen, S. Gao, B. Yuan, Z. Qi, Y. Liu, and F. Wang, "Optimized laplacian sparse coding for image classification," in *Image and Graphics*, pp. 636–645. Springer, 2015.

[18] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *NIPS*, 2014, pp. 487–495.

[19] T. M. Cover and J.A. Thomas, *Elements of information theory*, John Wiley & Sons, 2012.

[20] C.C. Chang and C.J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[21] J. Yao, S. Fidler, and R. Urtasun, "Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation," in *CVPR*, 2012, pp. 702–709.

[22] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 2169–2178.