Adding Unlabeled Samples to Categories by Learned Attributes

Jonghyun Choi Mohammad Rastegari Ali Farhadi[†] Larry S. Davis

University of Maryland, College Park [†]University of Washington

{jhchoi,mrastega,lsd}@umiacs.umd.edu ali@cs.uw.edu

Abstract

We propose a method to expand the visual coverage of training sets that consist of a small number of labeled examples using learned attributes. Our optimization formulation discovers category specific attributes as well as the images that have high confidence in terms of the attributes. In addition, we propose a method to stably capture example-specific attributes for a small sized training set. Our method adds images to a category from a large unlabeled image pool, and leads to significant improvement in category recognition accuracy evaluated on a large-scale dataset, ImageNet.

1. Introduction

Designing generalizable classifiers for visual categories is an active research area and has led to the development of many sophisticated classifiers in vision and machine learning [20]. Building a good training set with minimal supervision is a core problem in training visual category recognition algorithms [1].

A good training set should span the appearance variability of its category. While the internet provides a nearly boundless set of potentially useful images for training many categories, a challenge is to select the relevant ones – those that help to change the decision boundary of a classifier to be closer to the best achievable. So, given a relatively small initial set of labeled samples from a category, we want to mine a large pool of unlabeled samples to identify *visually different* examples without human intervention.

This problem has been studied by two research communities: active learning and semi-supervised learning. In active learning, the goal is to add visually different samples using human intervention, but to minimize human effort and cost by choosing informative samples for people to label [6, 13, 16]. Even though the amount of human intervention is minimized and its cost is getting cheaper via crowd sourcing, *e.g.*, Amazon Mechanical Turk, it is still preferable to not have humans in the loop because of issues like quality control and time [16].

Semi-supervised learning (SSL) aims at labeling unlabeled images based on their underlying distribution shared with a few labeled samples [5, 17, 21]. In SSL, it is assumed that the unlabeled images that are distributed around the labeled samples are highly likely to be members of the labeled category. However, if we need to dramatically change the decision boundary of a category to achieve good classification performance, it is unlikely that this can be done just by adding samples that are similar in the space in which the original classifier is constructed.

To expand the boundary of a category to an *unseen* region, we propose a method that selects unlabeled samples based on their attributes. The selected unlabeled samples are not always instances from the same category, but they can still improve category recognition accuracy, similar to [7, 10]. We use two types of attributes: category-wide attributes and example-specific attributes. The category-wide attributes find samples that share a large number of discriminative attributes with the preponderance of training data. The example-specific attributes find samples that are highly predictive of the *hard* examples from a category - the ones poorly predicted by a leave one out protocol.

We demonstrate that our augmented training set can significantly improve the recognition accuracy over a very small initial labeled training set, where the unlabeled samples are selected from a very large unlabeled image pool, *e.g.*, ImageNet. Our contributions are summarized as follows:

1. We show the effectiveness of using attributes learned with auxiliary data to label unlabeled images without annotated attributes.
2. We propose a framework that jointly identifies the unlabeled images and category wide attributes through an optimization that seeks high classification accuracy in both the original feature space and the attribute space.
3. We propose a method to learn example specific attributes with a small sized training set, used with the proposed framework. We then combine the category wide and the example specific attributes to further improve the quality of image selection by diversifying the variations of selected images.

The rest of the paper is organized as follows: Section 2 reviews related work. Section 3 presents the overview of our approach. Section 4 describes our optimization framework for discovering category wide attributes and the unlabeled images as well as a method to capture exemplar specific attributes. Section 5 describes the details of the dataset configurations used in our experiments. Experimental results that demonstrate the effectiveness of our method is presented in Section 6. Section 7 concludes the paper.

2. Related Work

Our work is related to active learning, semi-supervised learning, transfer learning and recent work about borrowing examples from other categories.

Active Learning The goal of active learning is to add examples with minimal human supervision [6]. [16] provides a comprehensive survey. Recently, Parkash *et al.* proposed a novel active learning framework based on interactive communication between learners and supervisors (teachers) via attributes [13]. It requires fairly extensive human supervision with rich information.

Semi-Supervised Learning Semi-supervised learning (SSL) adds unlabeled examples to a training set by modeling the distribution of features without supervision. [21] is a detailed review of the SSL literature. Fergus *et al.* proposed a computationally efficient SSL technique for large datasets [5]. Our approach also uses a large dataset and scales linearly in the size of that dataset; it differs from conventional SSL approaches because we do not use the distribution of sample in the original feature space, but in an attribute space. Recently, Shrivastava *et al.* proposed a SSL based scene category recognition framework using attributes, constrained by a category ontology [17]. They leverage the inter-class relationships as constraints for SSL using semantic attributes given by a category ontology as a priori. Our approach is similar to their work in terms of using attributes, but aims to discover attributes without any structured semantic prior.

Transfer Learning and Borrowing Examples Our work is related to recent work on transfer learning [12] and borrowing examples [7, 10, 15].

Ruslan *et al.* [15] proposed building a hierarchical model from categories to borrow images of a useful category for detection and classification. They assume that the images in a category are not diverse and adding all images from some selected category will help to build a better model for the target category. The assumption, however, is bound to be violated by visually diverse categories.

Instead, Lim *et al.* [10] propose a max-margin formulation to borrow some samples from other categories based on a symmetric borrowing constraints.

Kim and Grauman [7] propose a shape sharing method to improve segmentation based on the insight that shapes are often shared between objects of different categories.

Attributes Research on attributes recently has been drawing a lot of attention in the computer vision community because of their robustness to visual variations [4, 8, 9]. Attributes can, in principle, be used to construct models of new objects without training data - zero shot learning [9]. Recently, Rastegari *et al.* [14] propose discovering implicit attributes that are not necessarily semantic for category recognition. The discovered attributes preserve category-specific traits as well as their visual similarity by an iterative algorithm that learns discriminative hyperplanes with max-margin and locality sensitive hashing criteria.

3. Approach Overview

Given a handful of labeled training examples per category, it is difficult to build a generalizable visual model of a category even with sophisticated classifiers [20]. To address the lack of variations of the few labeled examples, we expand the visual boundary of a category by adding unlabeled samples based on their attributes. The attribute description allows us to find examples that are visually different but similar in traits or characteristics [4, 8, 9].

Based on recent work on automatic discovery of attributes [14] and large scale category-labeled image datasets [2], we discover a rich set of attributes. These attributes are leaned using an auxiliary category-labeled dataset to avoid biasing the attribute models towards the few labeled examples. The motivation here is similar to what underlies the successful Classemes representation [18] which achieved good category recognition performance by representing samples by external data that consists of a large number of samples from various categories.

Across the original visual feature space and the attribute space, we propose a framework that jointly selects the unlabeled images to be assigned to each category and the discriminative attribute representations of the categories based on either a category wide or exemplar based ranking criteria. Sec. 4.1 presents the optimization framework for category wide addition of unlabeled samples to categories. This adds samples that share many discriminative attributes amongst themselves and the given labeled training data. The same framework can be applied to identify relevant unlabeled samples based on their attribute similarity to specific instances of the training data. This only involves a simple change to one term of the optimization, and is based on how ranks of unlabeled samples change as labeled samples are left out, one at a time, from the attribute based classifier. So, the optimization runs twice - one to identify samples that share large numbers of discriminative attributes within class and a second to find samples that share strong attribute similarity with specific members of the class, and the two sets of samples are then combined to create the augmented training set for the class. We refer to the first as a categorical analysis and the second as an exemplar analysis.

4. Joint Discovery of Discriminative Attributes and Unlabeled Samples

4.1. Categorical Analysis

We simultaneously discover discriminative attributes and images from the unlabeled data set in a joint optimization framework formulated in both visual feature space and attribute space with a max margin criterion for discriminativity. Unlike [17], we do not require a label taxonomy to find the shared properties. Also unlike [10], we do not need to learn the distributions of the unlabeled images in the original feature space.

For each category c, we will construct a classifier in visual feature space, w_c^v, using the set $X = \{x_i | i \in \{1, \ldots, l, l+1, \ldots, n\}\}$ that consists of the initially given labeled training images $\{x_i | i \in \{1, \ldots, l\}\} \subset X$ and the selected images from the unlabeled image pool $\{x_i | i \in \{l+1, \ldots, n\}\} \subset X$. The subset of images from the unlabeled set is assigned to a category based on identifying discriminative attribute models. Since the problems of determining the discriminative attributes and selecting the subset of unlabeled data to assign to a category are coupled, we learn them jointly. Additionally, we want to mitigate against unlabeled samples being assigned to multiple categories, so a term $M(\cdot)$ is added to the optimization criteria to enforce that. The joint optimization function is:

$$\min_{I_c \in I, w_c^v, w_c^a} \sum_c \Big(\alpha J_c^v(I_c, w_c^v) + \beta J_c^a(I_c, w_c^a) \Big) + M(I)$$

subject to

$$\begin{aligned}
& J_c^v(I_c, w_c^v) = \|w_c^v\|_2^2 + \lambda_v \sum_{i=1}^{n} \xi_{c,i} \\
& I_{c,i} \cdot y_{c,i}(w_c^v x_i) \geq 1 - \xi_{c,i}, \quad \forall i \in \{1, \ldots, n\} \\
& J_c^a(I_c, w_c^a) = \|w_c^a\|_2^2 + \lambda_a \sum_{j=1}^{n} \zeta_{c,j} - \sum_{k=l+1}^{n} I_{c,k} \Big(w_c^a \phi(x_k) \Big) \\
& I_{c,j} \cdot y_{c,j}(w_c^a \phi(x_j)) \geq 1 - \zeta_{c,j}, \quad \forall j \in \{1, \ldots, n\} \\
& \sum_{k=l+1}^{n} I_{c,k} \leq \gamma, \quad I_{c,k} = 1, \; \forall k \in \{1, \ldots, l\} \\
& M(I) = \sum_{c1 \neq c2} \sum I_{c1} \cdot I_{c2},
\end{aligned} \tag{1}$$

$I_c \in \{0, 1\}$ is the sample selection vector for category c, and indicates which unlabeled samples are selected for assignment to the training set of category c. $I_{c,i} = 1$ when the i^{th} sample is selected for category c. $x_i \in \mathbb{R}^D$ is the visual feature vector of image i. $y_{c,i} \in \{+1, -1\}$ indicates whether the label assigned to x_i is c $(+1)$ or not (-1). $\phi(\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}^A$ is a mapping function of visual feature to the attribute space that is learned from auxiliary data, where $\mathbb{R}^D$ and $\mathbb{R}^A$ denote visual feature space and attribute space, respectively. α and β are hyper-parameters for balancing the max margin objective terms for both the visual feature and attribute based classifiers. γ is a hyper-parameter for specifying the number of selected images.

$J_c^v(I_c, w_c^v)$ and the second constraint of Eq. 1 are a max-margin classification terms in visual feature space. $J_c^a(\cdot)$ and the forth constraint of Eq. 1 are a max-margin classifier in the attribute space (T_A) with a selection criterion (T_R); we divide it as follows:

$$J_c^a(I_c, w_c^a) = \underbrace{\|w_c^a\|_2^2 + \sum_{j=1}^{n} \zeta_{c,j}}_{T_A} \underbrace{- \sum_{k=l+1}^{n} I_{c,k} \Big(w_c^a \phi(x_k) \Big)}_{T_R}. \tag{2}$$

T_R essentially chooses the top γ responses of the attribute classifier from the unlabeled set by the fifth constraint of Eq. 1. The term $M(I_c)$ penalizes adding the same sample to multiple categories (sixth constraint of Eq. 1).

The objective function is obviously not convex due to the interconnection of the two spaces by the example selecting indicator vector I and the attribute mapper $\phi(\cdot)$. However, if the I_c's were known and we fix either $J_c^v(I_c, w_c^v)$ or $J_c^a(I_c, w_c^a)$, the function becomes convex and can be solved with an iterative block coordinate descent algorithm. At each iteration we fix one of the terms and the entire objective function becomes an ordinary max margin classification formulation with a selection criterion. Each iteration of the block coordinate descent algorithm updates the set of indicator vectors I. At the first iteration, the initial value of I is determined by training the attribute classifier w_c^a on the given labeled training set. Then, after the two SVM's in both spaces are updated, we update I. Since there is no proof of convergence for the algorithm, we iterate it a fixed number of times - $1 \sim 3$ in practice. The iterations could be controlled using a held out validation set, but since our premise is that labeled samples are rare we do not do that.

4.2. Exemplar Analysis

The discriminative attributes learned in Sec. 4.1 capture commonality among all examples in a category. We refer them as *categorical attributes*. Each example, however, has its own characteristics that may help to expand the visual space of the category by identifying images based on example-specific characteristics. To discover *exemplar attributes*, a straightforward solution would be to learn exemplar-SVMs [11]. The exemplar-SVM, however, requires many negative samples to make the classifier output stable. For our purposes, though, we can accomplish the same thing by analyzing how the ranks of unlabeled samples change when a single sample is eliminated from the training set of the attribute SVM. If an unlabeled sample sees its rank drop sharply from its rank in the full-sample SVM, then the training sample dropped should have strong attribute similarity to the unlabeled sample.

This is illustrated in Figure 1. The top row shows the ten initial labeled orange samples. The leftmost column shows unlabeled samples sorted by their rank in the attribute classifier learned from that set. Then we construct leave one out attribute classifiers, and each column shows the new rankings of unlabeled samples when each image at the top of the column is eliminated from the training set. Eliminating the half orange (second sample, top row) from the training set reduces the rank of the globally best unlabeled sample from 1 to 10.

First, let w_c^a be the attribute classifier for the current training set for category c (while the process is initialized based on the labeled training set, after each iteration we use the additional unlabeled samples added to the category to construct a new attribute classifier). Let $w_{c,\bar{j}}^a$ be the attribute classifier learned when the i^{th} sample is removed from the training set. We next describe how we use the ranks of unlabeled samples in these two classifiers to modify T_R in Eq. 2. Basically, we are going to re-rank the unlabeled samples based on their rank changes from w_c^a to $w_{c,\bar{j}}^a$. We want samples whose ranks are lowered dramatically by the elimination of a single sample from the training set to be highly ranked by the re-ranking function. This can be accomplished by computing the following score based on rank changes, and the sorting the unlabeled samples by this score:

$$e_j(x_i) = \frac{\mu}{r_g(x_i)} - \frac{\nu}{r_j(x_i)}, \tag{3}$$

where x_i is a sample from the an unlabeled pool, $r_g(\cdot)$ and $r_j(\cdot)$ are the rank functions of w_c^a and $w_{c,\bar{j}}^a$ respectively. μ and ν are the balancing hyper-parameters for two ranks. T_R is then simply determined by first selecting the new top ranked sample from each leave one out SVM, then the second ranked, until a fixed number of samples are selected (skipping over duplicates). This set is then used to re-learn the feature and attribute based SVM's and the entire process iterates.

5. Dataset

We construct a dataset from a large scale dataset for category recognition, ImageNet [2] using its standard benchmark subset, ILSVRC 2010 dataset. We will publicly release our dataset for future comparison.[1] It consists of approximately 1 million images of 1,000 categories. The images are downloaded from a photo sharing portal[2]. It provides fine grained category labels such as specific breed of dogs, *e.g.*, Yorkshire Terrier and Australian Terrier.

We randomly choose 11 categories among natural objects such as vegetable and dogs as the categories of interest. Those categories have very large appearance variations due to factors including non-rigid deformation, lighting, camera angle, intra-class appearance variability *etc*. For each category, we randomly choose ten images as an initial labeled training set and 500 images as a testing set. The unlabeled image pool consists of images that are arbitrarily chosen from the entire 1,000 categories in the ILSVRC 2010 benchmark dataset, but includes at least 50 samples from each of the categories to be learned. The size of the image pool varies in the experiments but is much larger (from 5,000 to 50,000) than the initial training set. For learning the attribute space and the mapper, it is expected that the attribute mapper should capture some attribute of the categories of interest. For this purpose, we use 50 labeled samples from 93 categories that are similar to the 11 categories to learn the attribute space.

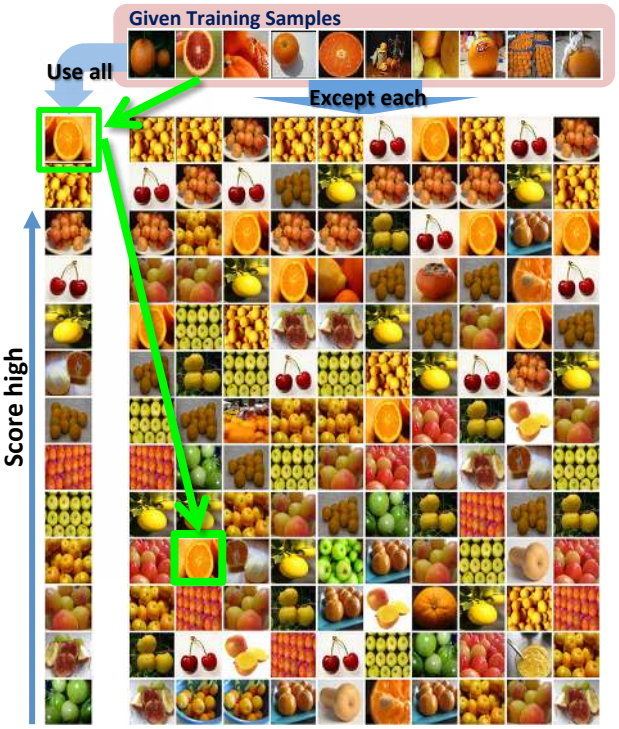

Figure 1. Unlabeled images ordered by confidence score by w_c^a and a set of $w_{c,\bar{i}}^a$'s (column wise). The first row shows the labeled training samples (10 examples). The left most column is a list of unlabeled images ordered by confidence score by w_c^a. Rest of the columns are lists of unlabeled images ordered by each $w_{c,\bar{i}}^a$'s. Note that an image of halved orange in the second column makes the first ranked images in the left most column (by w_c^a) go down because the halved orange was removed in the training set of $w_{c,\bar{i}}^a$.

6. Experiments

The main goal of our method is to add unlabeled images to the initial training set in order to classify more test images correctly. We demonstrate the effectiveness of our

[1] http://umiacs.umd.edu/∼jhchoi/addingbyattr/
[2] http://www.flickr.com

method by improvements in average precision (AP) of category recognition. We also evaluate our approach under various scenarios including the precision of the unlabeled image pool and the size of the learned attribute space and also the effect of parameters including number of selected examples. Moreover, we evaluate the effect of selecting images that are not from the category of interest.

6.1. Experimental Setup

Visual feature descriptors: We use various visual feature descriptors including HOG, GIST and color histograms. Since the feature dimensionality is prohibitively large, we reduce the dimension to 6,416 by PCA.

Attribute discovery: We use the binary attribute discovery method of Rastegari *et al.* [14] as the attribute mapping function, $\phi(\cdot)$ in Eq. 1. We learn the mapper with default hyper-parameter sets as suggested in [14]. We use 400 bits in most of our experiments. We also present performance as a function of the number of bits.

Max margin optimization: We use LibLinear [3] for training all max-margin based objective functions. To address the non-linearity of visual feature space, we use homogeneous kernel mapping [19] on the original features with the linear classifier. For the hinge loss penalty hyper-parameter, we use 0.1.

Parameters: For the parameter in Eq. 1, we use $\alpha = 1, \beta = 1$. For categorical attribute only, we mostly use $\gamma = 50$ except ones in Section 6.4. For combining exemplar and categorical attributes, we mostly use $\gamma = 20$ and $\gamma_i = 3$ except for Section 6.4. We investigate algorithm performance as a function of γ in Section 6.4. For the parameters of the scoring function for exemplar-attributes in Eq. 3, we use $\mu = 1$ and $\nu = 1$.

6.2. Qualitative Results

Our method discovers examples that expand the visual coverage of a category by not only adding the examples from the same category but also examples from other categories. Figure 2 illustrates qualitative results on the category *Dalmatian* for both categorical and exemplar attributes analyses. The selected examples based on categorical attributes exhibit characteristics commonly found in the labeled examples such as dotted, four legged animal. The exemplar attributes, on the other hand, select examples that exhibit the characteristic of individual labeled training examples.

6.3. Comparison with Other Selection Criteria

Given our goal of selecting examples from a large unlabeled data with only a small number of labeled training samples, we do not compare with semi-supervised learning methods because they need more labeled data to model the distribution. Since our method does not involve human intervention, we do not compare to active learning.

Category Name	Init.	NN	ALC	Cat.	E+C
Mashed Potato	45.03	34.02	51.15	61.39	63.92
Orange	29.84	16.29	26.97	40.61	41.05
Lemon	32.21	27.58	32.43	35.37	34.23
Green Onion	25.06	16.50	19.66	38.57	40.20
Acorn	13.09	11.05	15.41	19.35	20.10
Coffee bean	58.29	43.89	56.62	64.65	66.54
Golden Retriever	14.54	15.57	12.61	17.54	18.61
Yorkshire Terrier	29.62	13.62	27.63	41.41	45.65
Greyhound	15.24	15.73	15.64	14.75	15.22
Dalmatian	43.84	27.97	37.91	54.42	57.23
Miniature Poodle	26.10	12.50	21.16	28.87	30.21
Average	30.26	21.34	28.84	37.90	39.36

Table 1. Comparison of average precision (AP) (%) for each category with 50 added examples by various methods. 'Init.' refers to initial labeled training set. 'NN' refers to addition by 'nearest neighbor' in visual feature space, 'ALC' refers to addition by 'active learning criteria (ALC)' that finds the examples close to the current decision hyperplanes [6]. 'Cat.' refers to our method of select examples using categorical attributes only. 'E+C' refers to addition using categorical and exemplar attributes. The size of the unlabeled dataset is roughly 3,000 from randomly chosen categories out of 1,000 categories.

We compare to baseline algorithms which are applicable to the large unlabeled data scenario. The first baseline algorithm is to select nearest neighbors. The second baseline selects images by an active criterion that finds examples close to a learned decision hyperplanes [6]. Both baseline algorithms selects images based on analysis in the visual feature space.

As shown in Table. 1, the two baseline strategies decrease mean average precision (mAP). However, our method identifies useful images in the unlabeled image pool and significantly improves mAP by 7.64%. Except for the category *Greyhound*, we obtain performance gain from 2.77% - 16.36% in all categories. The added examples serve not only as positive samples for each category but also as negative samples for other categories. The quality of the selected set can change the mAP significantly in both ways.

6.4. Number of Selected Examples

As we select more examples, controlled by γ in Eq. 1, the chances of both selecting useful images and harmful images for a category increase simultaneously. We vary the number of selected examples and observe mean average precision as shown in Figure 3. The category wide attributes identify useful unlabeled images. In addition, the exemplar attributes further improve the recognition accuracy.

6.5. Adding Examples from Similar Categories

Among the selected images per category, some examples are true instance of the category. We refer to these as *exact*

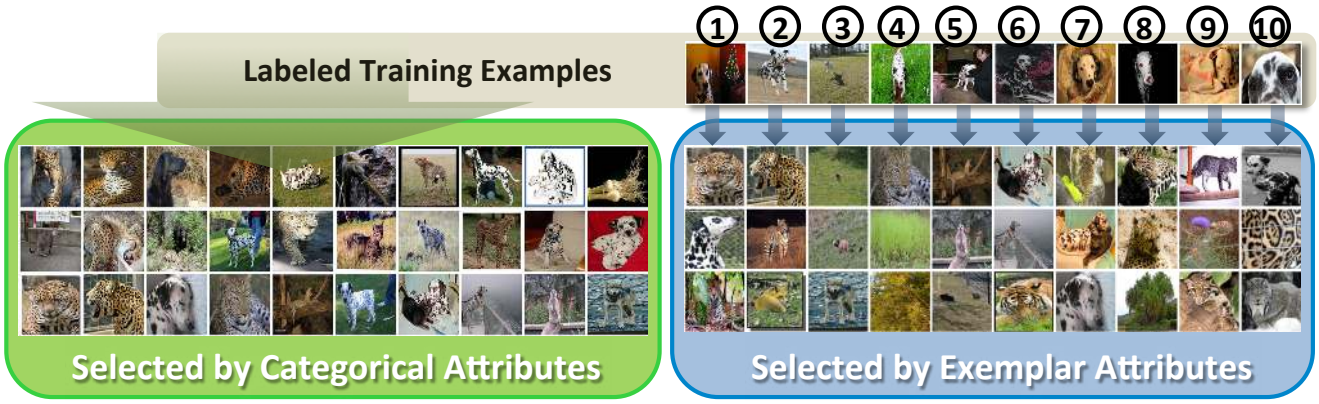

Figure 2. Qualitative results of our method. Note that the selected examples by categorical attributes display characteristics commonly found in the labeled training examples such as 'dotted', 'four legged animal'. In contrast, the exemplar attributes select the examples that display the characteristic of individual example.

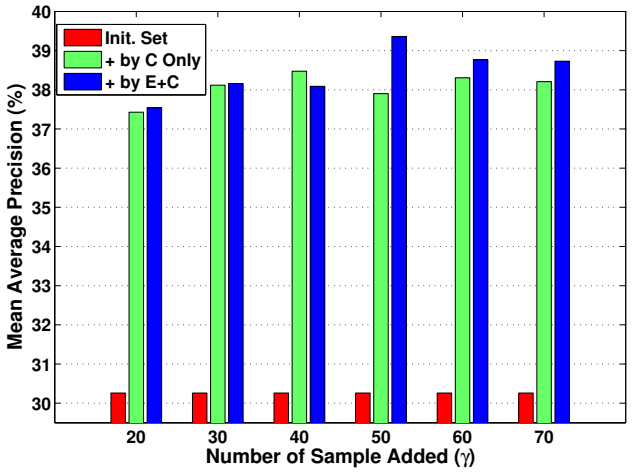

Figure 3. Mean average precision (mAP) of 11 category by our method varying the number of unlabeled images selected. The red, green and blue are the mAP using the initial labeled set (Init. Set), the augmented set by our method using category wide attributes only (+ by C only) and categorical+exemplar attributes respectively. (+ by E+C)

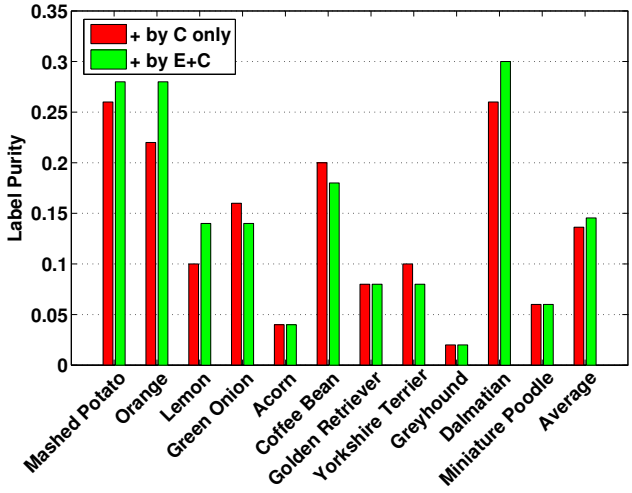

Figure 4. Purity of added examples. Red bars denote the purity of selected images using category wide attributes only (+ by C only) and the green bars are obtained from categorical+exemplar attributes (+ by E+C).

examples and the rest as *similar examples*. We are interested in how much the similar examples improve category recognition. First, we examine the purity of the selected set in Figure 4. The purity is the percentage of exact samples in the set. Surprisingly, even though the purity values seem low, they still improve classification performance.

We now investigate how much the similar examples improve the average precision (AP) by removing the exact examples from the selected set. The blue bars in Figure 5 represent the AP using just the similar examples. It is interesting to note that using only the similar examples still improves the APs over the initial labeled set.

In addition, it is also interesting to observe how the performance changes when we add the same number of similar examples as the size of the initially selected image set (50). This is shown as green bars in Figure 5. All results are obtained using categorical attributes only. (The results using both exemplar and categorical attributes are similar so are omitted).

6.6. Precision of Unlabeled Data

The unlabeled data can be composed of images from many categories. The precision of the unlabeled data is defined as the ratio of size of the unlabeled images from extraneous categories to the size of the entire unlabeled image data. The larger the unlabeled data, the lower we expect its precision to be (imagine running a text based image search

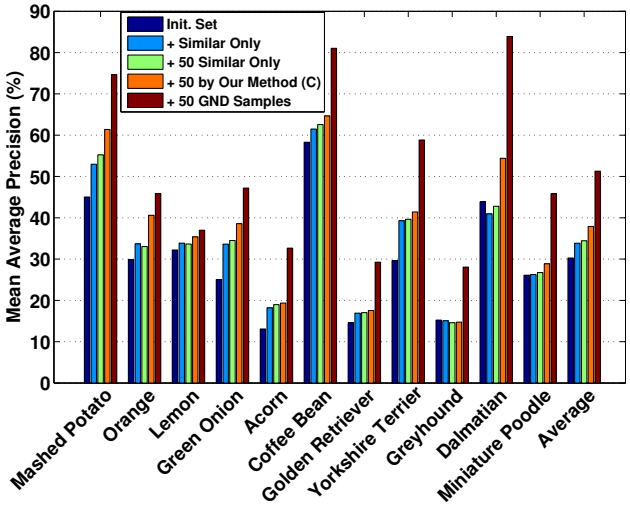

Figure 5. Mean average precision (mAP) as a function of the purity of the selected examples. The navy colored bars are obtained using the initial labeled set (baseline). The blue bars use only similar examples among the selected 50 examples. The green bars use 50 similar examples to compare with the result of our selected 50 examples (orange bars) including both similar and exact examples. The red bars are obtained using a set of 50 ground truth images, which is the best achievable accuracy (upper bound). Even the similar examples alone improve the category recognition accuracy compared to just using the initial labeled set.

using the category name and accepting the first k images returned). It is interesting to observe how robust our method is against the precision of unlabeled data.

We start with an unlabeled set (550 images, 50 from each of the 11 categories) of precision 1.0, and reduce precision by adding images from other categories. The number of the unrelated images ranges from 2,500 to 50,000, which are randomly chosen from the entire 1,000 categories of the ImageNet ILSVRC 2010 dataset.

As shown in Figure 6, we observe that the accuracy improvement by our method using categorical attributes is quite stable even when precision is low.

6.7. Size of Initial Labeled Set

We next explore how the size of the initial labeled set effects accuracy. We systematically vary the size from 5 to 50 and show mAP compared to an SVM learned on the initial training set - see Figure 7. The mAP gain for the smallest initial labeled set (5) is the highest as expected. When the number of samples is larger than 25, our method (+ by C only) does not improve the mAP much, although it still improves by $1.18 - 2.74\%$. Interestingly when there are many samples in the initial training set (*e.g.*, more than 25), the exemplar traits begin to reduce the mAP.

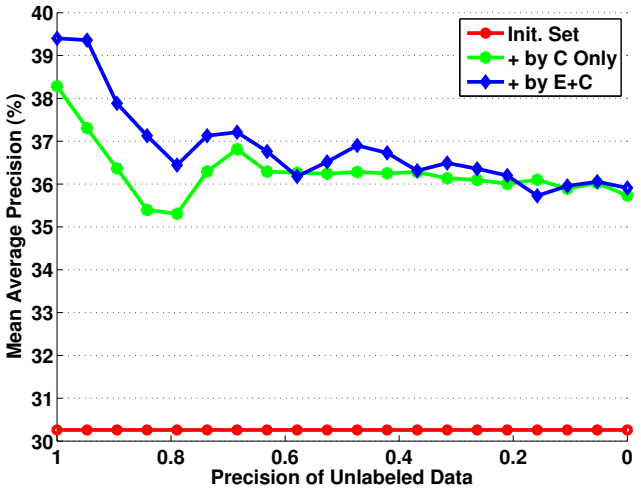

Figure 6. Mean average precision (mAP) as a function of precision of unlabeled data. Precision denotes the ratio of size of the unlabeled images from extraneous categories to the size of the entire unlabeled image data (size = 50,000). Although precision decreases, the mean average precisions (mAP) by our method do not decrease much.

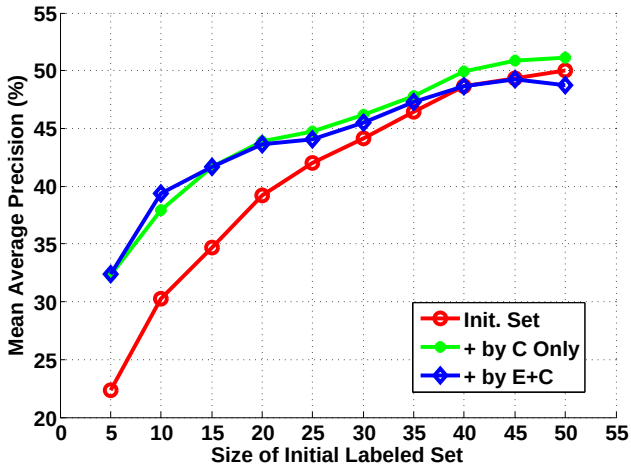

Figure 7. Mean average precision (mAP) as a function of the size of the initial labeled set. The number of added samples is 50 in all experiments.

6.8. Comparison to Exemplar SVM

We also compare the effectiveness of our proposed exemplar attributes discovery method (Sec. 4.2) to a conventional exemplar SVM [11]. It is straightforward to integrate the exemplar SVM into our formulation (Eq. 1): by setting label $y_{c,j}$ to 1 for the j^{th} example, the label coresponding to the examples in the same category to 0 and the rest to 1. To stabilize the exemplar SVM scores, we employ 50,000 external negative samples to learn each exemplar SVM while we use the small original training set for our method. Fig-

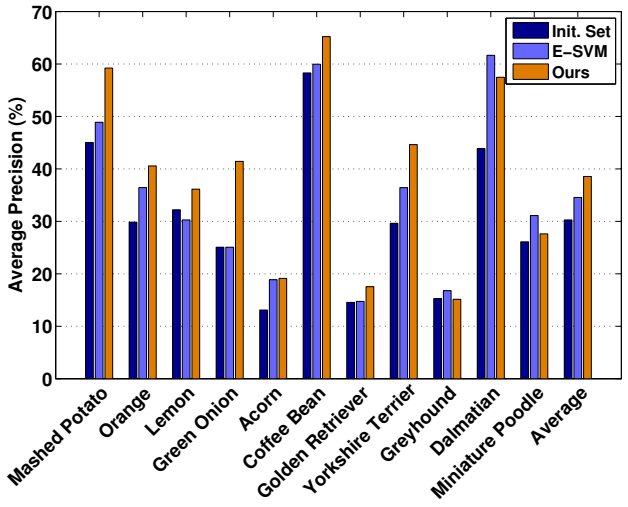

Figure 8. Comparison of our exemplar attribute discovery method (Sec. 4.2) to exemplar SVM. Our method outperforms the exemplar SVM in terms of category recognition accuracy by APs without the extra large negative example set (size = 50,000).

ure 8 shows that our exemplar attribute discovery method outperforms the exemplar SVM by large margins even without the large negative example set.

7. Conclusion

We proposed a method to select unlabeled images to learn classifiers based on learned attributes. The unlabeled images selected by our method do not necessarily belong to the category of interest but are similar in attributes. Our method does not require any annotated attribute set a priori but first builds an automatically learned attribute space. We formulate a joint optimization framework to select both images and the attributes for a category and solve it iteratively. In addition to the category wide attributes, we identify example specific attributes to diversify the selected images. For addressing the problem of small size training set to learn the example specific attributes, we propose a method that can be intuitively regarded as an inverse of exemplar SVM.

From a large unlabeled data pool, the selected images improve category recognition accuracy significantly over accuracy obtained using the initial labeled training set.

Acknowledgements: This work was partially supported by MURI from the Office of Naval Research under the Grant N00014-10-1-0934.

References

[1] T. L. Berg, A. Sorokin, G. Wang, D. A. Forsyth, D. Hoiem, A. Farhadi, and I. Endres. It's All About the Data. In *Proceedings of the IEEE, Special Issue on Internet Vision*, 2010. 1

[2] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 2, 4

[3] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 2008. 5

[4] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing Objects by their Attributes. In *CVPR*, 2009. 2

[5] R. Fergus, Y. Weiss, and A. Torralba. Semi-supervised Learning in Gigantic Image Collections. In *NIPS*, 2009. 1, 2

[6] P. Jain, S. Vijayanarasimhan, and K. Grauman. Hashing Hyperplane Queries to Near Points with Applications to Large-Scale Active Learning. In *NIPS*, 2010. 1, 2, 5

[7] J. Kim and K. Grauman. Shape Sharing for Object Segmentation. In *ECCV*, 2012. 1, 2

[8] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *ICCV*, 2009. 2

[9] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 2

[10] J. J. Lim, R. Salakhutdinov, and A. Torralba. Transfer Learning by Borrowing Examples for Multiclass Object Detection. In *NIPS*, 2011. 1, 2, 3

[11] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of Exemplar-SVMs for Object Detection and Beyond. In *ICCV*, 2011. 3, 7

[12] S. J. Pan and Q. Yang. A Survey on Transfer Learning. *IEEE Trans. on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. 2

[13] A. Parkash and D. Parikh. Attributes for Classifier Feedback. In *ECCV*, 2012. 1, 2

[14] M. Rastegari, A. Farhadi, and D. Forsyth. Attribute Discovery via Predictable Discriminative Binary Codes. In *ECCV*, 2012. 2, 5

[15] R. Salakhutdinov, A. Torralba, and J. Tenenbaum. Learning to Share Visual Appearance for Multiclass Object Detection. In *CVPR*, 2011. 2

[16] B. Settles. Active Learning Literature Survey. Technical report, 2009. 1, 2

[17] A. Shrivastava, S. Singh, and A. Gupta. Constrained semi-supervised learning using attributes and comparative attributes. In *ECCV*, 2012. 1, 2, 3

[18] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient Object Category Recognition Using Classemes. In *ECCV*, 2010. 2

[19] A. Vedaldi and A. Zisserman. Efcient Additive Kernels via Explicit Feature Maps. *IEEE Trans. PAMI*, 2011. 5

[20] W. Zhang, S. X. Yu, and S.-H. Teng. PowerSVM: Generalization with Exemplar Classification Uncertainty. In *CVPR*, 2012. 1, 2

[21] X. Zhu. Semi-Supervised Learning Literature Survey. Technical report, 2008. 1, 2