Constrained Semi-Supervised Learning Using Attributes and Comparative Attributes

Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta

The Robotics Institute, Carnegie Mellon University, Pittsburgh (PA), USA
http://graphics.cs.cmu.edu/projects/constrainedSSL/

Abstract. We consider the problem of semi-supervised bootstrap learning for scene categorization. Existing semi-supervised approaches are typically unreliable and face semantic drift because the learning task is under-constrained. This is primarily because they ignore the strong interactions that often exist between scene categories, such as the common attributes shared across categories as well as the attributes which make one scene different from another. The goal of this paper is to exploit these relationships and constrain the semi-supervised learning problem. For example, the knowledge that an image is an auditorium can improve labeling of amphitheaters by enforcing constraint that an amphitheater image should have more circular structures than an auditorium image. We propose constraints based on mutual exclusion, binary attributes and comparative attributes and show that they help us to constrain the learning problem and avoid semantic drift. We demonstrate the effectiveness of our approach through extensive experiments, including results on a very large dataset of one million images.

1 Introduction

How do we exploit the sea of visual data available online? Most supervised computer vision approaches are still impeded by their dependence on manual labeling, which, for rapidly growing datasets, requires an incredible amount of manpower. The popularity of Amazon Mechanical Turk and other online collaborative annotation efforts [1, 2] has eased the process of gathering more labeled data, but it is still unclear whether such an approach can scale up with the available data. This is exacerbated by the heavy-tailed distribution of objects in the natural world [3]: a large number of objects occur so sparsely that it would require a significant amount of labeling to build reliable models. In addition, human labeling has a practical limitation in that it suffers from semantic and functional bias. For example, humans might label an image of *Christ*/*Cross* as *Church* due to high-level semantic connections between the two concepts.

An alternative way to exploit a large amount of unlabeled data is semi-supervised learning (SSL). A classic example is the "bootstrapping" method: start with a small number of labeled examples, train initial models using those examples, then use the initial models to label the unlabeled data. The model is retrained using the confident self-labeled examples in addition to original examples. However, most semi-supervised approaches, including bootstrapping, have often exhibited low and unacceptable accuracy because the limited number of initially labeled examples are insufficient to constrain the learning process. This often leads to the well known problem of "semantic

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part III, LNCS 7574, pp. 369–383, 2012.

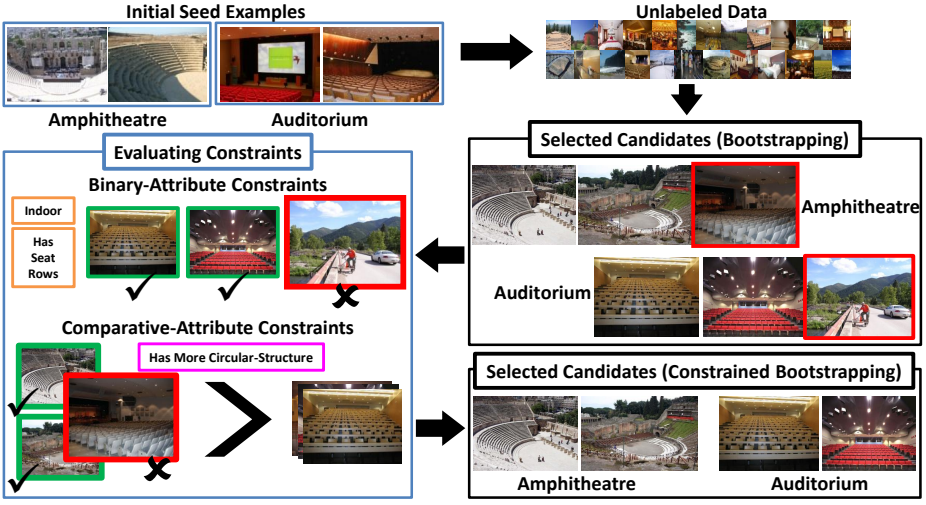


Fig. 1. Standard Bootstrapping vs. Constrained Bootstrapping: We propose to learn multiple classifiers (auditorium and amphitheater) jointly by exploiting similarities (e.g., both are indoor and have seating) and dissimilarities (e.g, amphitheater has more circular structures than auditorium) between the two. We show that joint learning constrains the SSL, thus avoiding semantic drift.

drift" [4], where newly added examples tend to stray away from the original meaning of the concept. The problem of semantic drift is more evident in the field of visual categorization because intra-class variation is often greater than inter-class variation. For example, "electric trains" resemble "buses" more than "steam engines" and many "auditoriums" appear very similar to "amphitheaters".

This paper shows that we can avoid semantic drift and significantly improve performance of bootstrapping approach by imposing additional constraints. We build upon recent work in information extraction [5], and propose a novel semi-supervised image classification framework where instead of each classifier selecting its own set of images, we jointly select images for each classifier by enforcing different types of constraints. We show that coupling scene categories via attributes and comparative attributes[1] provides us with the constraints necessary for a robust bootstrapping framework. For example, consider the case shown in Figure 1. In the case of the naïve bootstrapping approach, the initial "amphitheater" and "auditorium" classifiers select self-labeled images independently which leads to incorrect instances being selected (outlined in red). However, if we couple the scene categories and jointly label all the images in the dataset, then we can use the auditorium images to clean the amphitheater images, since the latter should have more circular structures compared to the former. We demonstrate that the joint labeling indeed makes the data selection robust and improves the performance

[1] Comparative attributes are special forms of attributes that are used to compare and express relationships between two nouns. For example, "banquet halls" are **bigger** than "bedrooms"; "amphitheaters" are **more circular** than "auditoriums".

significantly. While we only explore the application of these new constraints to bootstrapping approaches, we believe they are generic and can be applied to other semi-supervised approaches as well.

Contributions: We present a framework for coupled bootstrap learning and explore its application to the field of image classification. The input to our system is an ontology which defines the set of target categories to be learned, the relationships between those categories and a handful of initial labeled examples. We show that given these initial labeled examples and millions of unlabeled images, our approach can obtain much higher accuracy by coupling the labeling of all the images using multiple classifiers. The key contributions of our paper are: (a) a semi-supervised image classification framework which jointly learns multiple classifiers, (b) demonstrating that sharing information across categories via attributes [6, 7] is crucial to constrain the semi-supervised learning problem, (c) extending the notion of sharing across categories and showing that information sharing can also be achieved by capturing dissimilarities between categories. Here we build upon the recent work on relative attributes [8] and comparative adjectives [9] to capture differences across categories. In this paper, the relationships between attributes and scene categories are defined by a human annotator. As opposed to image-level labeling, attribute and comparative attribute relationships are significantly cheaper to annotate as they scale with the number of categories (not with the number of images). Note that the attributes need not be semantic at all [10]. However, the general benefit of using semantic attributes is that they are human-communicable [11] and we can obtain them automatically using other sources of data such as text [12].

2 Prior Work

During the past decade, computer vision has seen some major successes due to the increasing amount of data on the web. While using big data is a promising direction, it is still unclear how we should exploit such a large amount of data. There is a spectrum of approaches based on the amount of human labeling required to use this data. On one end of the spectrum are supervised approaches that use as much hand-labeled data as possible. These approaches have focused on using the power of crowds to generate hand-labeled training data [1, 2]. Recent works have also focused on active learning [13–17, 11], to minimize human effort by selecting label requests that are most informative. On the other end of the spectrum are completely unsupervised approaches, which use no human supervision and rely on clustering techniques to discover image categories [18, 19].

In this work, we explore the intermediate range of the spectrum; the domain of semi-supervised approaches. Semi-supervised learning (SSL) techniques use a small amount of labeled data in conjunction with a large amount of unlabeled data to learn reliable and robust visual models. There is a large literature on semi-supervised techniques. For brevity, we only discuss closely related works and refer the reader to recent survey on the subject [20]. The most commonly used semi-supervised approach is the "bootstrapping" method, also known as self-training. However, bootstrapping typically suffers from semantic drift [4] – that is, after many iterations, errors in labeling tend to accumulate. To avoid semantic drift, researchers have focused on several approaches such as

using multi-class classifiers [21] or using co-training methods to exploit conditionally independent feature spaces [22]. Another alternative is to use graph-based methods, such as the graph Laplacian, for SSL. These methods capture the manifold structure of the data and encourage similar points to share labels [23]. In computer vision, efficient graph based methods have been used for labeling of images as well [24]. The biggest limitation with graph based approaches is the need for similarity measures that create graphs with no inter-class connections. In the visual world, it is very difficult to learn such a good visual similarity metric. Often, intra-class variations are larger than inter-class variations, which make pair-wise similarity based methods of little utility. To overcome this difficulty, researchers have focused on text based features for better estimation of visual similarity [25].

In this work, we argue that there exists a richer set of constraints in the visual world that can help us constrain the SSL-based approaches. We present an approach to combine a variety of such constraints in a standard bootstrapping framework. Our work is inspired by works from the textual domain [5] that try to couple learning of category and relation classifiers. However, in our case, we build upon recent advances in the field of visual attributes [6, 7] and comparative attributes [8, 9] and propose a set of domain-specific visual constraints to model the coupling between scene categories.

3 Constrained Bootstrapping Framework

Our goal is to use the initial set of labeled seed examples ($\mathcal{L}$) and a large unlabeled dataset ($\mathcal{U}$) to learn robust image classifiers. Our method iteratively trains classifiers in a self-supervised manner. It starts by training classifiers using a small amount of labeled data and then uses these classifiers to label unlabeled data. The most confident new labels are "promoted" and added to the pool of data used to train the models, and the process repeats. The key difference from the standard bootstrapping approach is the set of constraints that restrict which data points are promoted to the pool of labeled data.

In this work, we focus on learning scene classifiers for image classification. We represent these classifiers as functions ($f : X \rightarrow Y$) which, given input image features x, predict some label y. Instead of learning these classifiers separately, we propose an approach which learns these classifiers jointly. Our central contribution is the formulation of constraints in the domain of image classification. Specifically, we exploit the recently proposed attribute-based approaches [6, 7] to provide another view of the same data and enforce multi-view agreement constraints. We also build upon the recent framework of comparative adjectives [9] to formulate pair-wise labeling constraints. Finally, we use *introspection* to perform an additional step of self-refinement to weed out false positives included in the training set. We describe all the constraints below.

3.1 Output Constraint: Mutual Exclusion (ME)

Classification of a single datapoint by multiple scene classifiers is not an independent process. We can use this knowledge to enforce certain constraints on the functions learned for the classifiers. Mathematically, if we know some constraint on output values of two classifiers $f_1 : X \rightarrow Y_1$ and $f_2 : X \rightarrow Y_2$ for an input x, then we can require the

learned functions to satisfy these. One such output constraint is the **mutual exclusion constraint** (ME). In mutual exclusion, positive classification by one classifier immediately implies negative classification for the other classifiers. For example, an image classified as "restaurant" can be used as an negative example for "barn", "bridge" etc.

Current semi-supervised approaches enforce mutual exclusion by learning a multi-class classifier where a positive example of one class is automatically treated as a negative example for all other classes. However, the multi-class classifier formulation is too rigid for a learning algorithm. Consider, for example, "banquet hall" and "restaurant", which are very similar and likely to be confused by the classifier. For such classes, the initial classifier learned from a few seed examples is not reliable enough; hence, adding the mutual exclusion constraint causes the classifier to overfit.

We propose an adaptive mutual exclusion constraint. The basic idea is that during initial iterations, we do not want to enforce mutual exclusion between similar classes (y_1 and y_2), since this is likely to confuse the classifier. Therefore, we relax the ME constraint for manually annotated similar classes – a candidate added to the pool of one is not used as a negative example for the other. However, after a few iterations, we adapt our mutual exclusion constraints and enforce these constraints across similar classes as well.

3.2 Sharing Commonalities: Binary-Attribute Constraint (BA)

For the second constraint, we exploit the commonalities shared by scene categories. For example, both "amphitheaters" and "auditoriums" have large seating capacity; "bedrooms" and "conference rooms" are indoors and man-made. We propose to model these shared properties via attributes [6, 7]. Modeling visual attributes helps us enforce a constraint that the promoted instances must also share these properties.

Formally, we model this constraint in a multi-view framework [22]. For a function $f : X \rightarrow Y$, we partition X into views (X_a, X_b) and learn two classifiers f_a and f_b which can both predict Y. In our case, $f_a : X_a \rightarrow Y$ is the original classifier which uses low-level features to predict the scene classes. We model the sharing between multiple classes via f_b. f_b is a compositional function ($f_b : X_b \rightarrow A \rightarrow Y$) which uses low-level features to predict attributes A and then uses them to predict scene classes. It should be noted that even though we use multi-view framework to model sharing, it is quite a powerful constraint. In case of sharing, the function f_b updates at each iteration by learning a new attribute classifier, $X_b \rightarrow A$, which collects large amounts of data from multiple scene classes (e.g., the indoor attribute classifier picks up training instances from restaurant, bedroom, conference-room etc.). Also, note that the human annotated mapping from attribute space to scene class, $A \rightarrow Y$, remains fixed in our case.

3.3 Pairwise Constraint: Comparative Attributes (CA)

The above two constraints are unary in nature: these constraints still assume that the labeling procedures for two instances X_1 and X_2 should be completely independent of each other. Graph-based approaches [20, 24] have focussed on constraining labels of instances X_1, X_2 based on similarity – that is, if two images are similar they should have same labels. However, learning semantic similarity using image features is an extremely

difficult problem specifically because of high intra-class variations. In this paper, we model stronger and richer pairwise constraints on labeling of unlabeled images using comparative attributes. For example, if an image X_1 is labeled as "auditorium", then another image X_2 can be labeled as "amphitheater" if and only if it has more circular structures than X_1.

Formally, for a given pair of scene classes, $f_1 : X_1 \rightarrow Y_1$ and $f_2 : X_2 \rightarrow Y_2$, we model the pairwise constraints using a function $f^c : (X_1, X_2) \rightarrow Y_c$ and enforce the constraint that (f_1, f_2, f^c) should produce a consistent triplet (y_1, y_2, y_c) for a given pair of images (x_1, x_2) . Some examples of consistent triplets in our case would include (field, barn, more open space) and (church, cemetery, has larger structures) which mean 'field has more open space than barn' and 'church has larger structures than cemetery'.

3.4 Introspection or Self-cleaning

In iterative semi-supervised approaches, a classifier should ideally improve with each iteration. Empirically, these classifiers tend to make more mistakes in the earlier iterations as they are trained on very small amount of data. Based on these two observations, we introduce an additional step of *introspection* where after every five iterations, starting at fifteen, we use the full framework to score already included training data (instead of the unlabeled data) and drop positives that receive very low scores. This results in further performance improvement of the learned classifiers.

4 Mathematical Formulation: Putting It Together

We now describe how we incorporate the constraints described above in a bootstrapping semi-supervised approach. Figure 2 shows the outline of our approach. We have a set of binary scene classifiers $f_1 ... f_N$, attribute classifiers $f_1^a ... f_K^a$ and comparative attribute classifiers $f_1^c ... f_M^c$. Initially, these classifiers are learned using seed examples but are updated at each iteration using new labeled data. At each iteration, we would like to label the large unlabeled corpus ($\mathcal{U}$) and obtain the confidence of each labeling. Instead of labeling all images separately, we label them jointly using our constraints. We represent all images in $\mathcal{U}$ as nodes in a fully connected graph. The most likely assignment of each image (node) in the graph can be posed as minimizing the following energy function $E(y)$ over class labels assignments $y = \{y_1, .., y_{|\mathcal{U}|}\}$:

$$E(y) = -\left[\sum_{x_i \in \mathcal{U}} \Phi(x_i, y_i) + \lambda \sum_{(x_i, x_j) \in \mathcal{U}^2} \Psi(x_i, x_j, y_i, y_j) \right] \quad (1)$$

where $\Phi(x_i, y_i)$ is the unary node potential for image i with features x_i and its candidate label y_i and $\Psi(x_i, x_j, y_i, y_j)$ is the edge potential for labels y_i and y_j of pair of images i and j. It should be noted that y_i denotes assigned label to image i which can take label assignments in $\{l_1 .... l_n\}$.

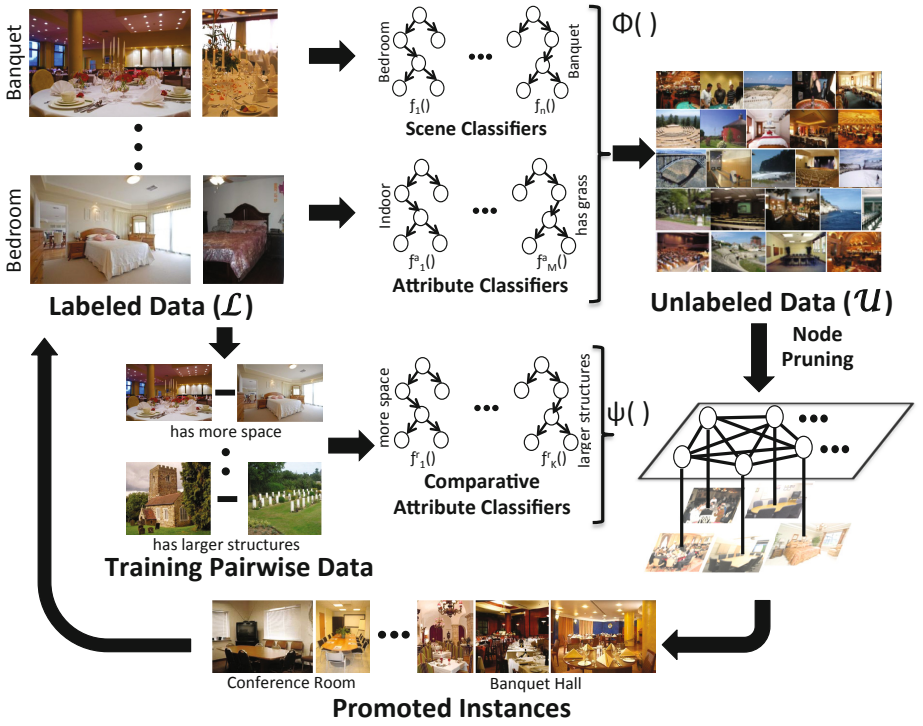


Fig. 2. Overview of our approach

The unary term $\Phi(x_i, y_i)$ is the confidence in assigning label y_i to image i by the combination of scene and attribute classifier scores. Specifically, if $f_j(x_i)$ is the raw score of j^{th} scene classifier on x_i and $f^{a_k}(x_i)$ is the raw score of k^{th} attribute classifier on x_i, then the unary potential is given by:

$$\Phi(x_i, y_i = l_j) = \sigma(f_j(x_i)) + \lambda_1 \sum_{a_k \in \mathcal{A}} (-1)^{\mathbf{1}_{x_i, y_i}(a_k)} \rho(f^{a_k}(x_i)) \tag{2}$$

The first term takes in the raw scene classifier scores and converts these scores into potentials using the sigmoid function ($\sigma(t) = \exp(\gamma t)/(1 + \exp(\gamma t))$) [26]. The second term uses a weighted voting scheme for agreement between scene and attribute classifiers (λ_1 is the normalization factor). Here, $\mathcal{A}$ is the set of binary attributes, $\mathbf{1}_{x_i, y_i}(a_k)$ is an indicator function which is 1 if the attribute a_k is detected in the image i but $\neg a_k$ is a property of scene class y_i (and vice versa). $\rho()$ denotes the confidence in prediction for attribute classifier[2]. Intuitively, this second term votes positive if both the attribute classifier and scene classifier agree in terms of class-attribute relationships. Otherwise, it votes negative where the vote is weighted in terms of the confidence of prediction.

[2] The confidence in prediction is defined as: $\rho(f^{a_k}(x_i)) = max(\sigma(f^{a_k}(x_i)), 1 - \sigma(f^{a_k}(x_i)))$.

The binary term $\Psi(x_i, x_j, y_i, y_j)$ between images i and j with labels y_i and y_j represents comparative-attribute relations between labeled classes.

$$\Psi(x_i, x_j, y_i, y_j) = \sum_{c_k \in \mathcal{C}} 1_{c_k}(y_i, y_j) \log(\sigma(f^{c_k}(x_i, x_j))) \tag{3}$$

where $\mathcal{C}$ denotes the set of comparative attributes, $\mathbf{1}_{c_k}(y_i, y_j)$ denotes if a given comparative attribute c_k exists between pairs of classes y_i and y_j and $f^{c_k}(x_i, x_j)$ is the score of comparative-attribute classifier for the pair of images x_i, x_j. Intuitively, this term boosts the labels y_i and y_j if a comparative attribute c_k scores high on pairwise features. For example, if instances i, j are labeled "conference-room" and "bedroom", their scores get boosted if the comparative attribute "has more space" scores high on pairwise features (since "conference rooms" have more space than "bedrooms").

Promoting Instances: Typically in the semi-supervised problem, $|\mathcal{U}|$ varies from tens of thousand to million images. Estimating the most likely label for each image in $\mathcal{U}$ necessitates minimizing Eq.(1) which is computationally intractable in general. Since our goal is to find a few very confident images to add to the labeled set $\mathcal{L}$ we do not need to minimize Eq.(1) over the entire $\mathcal{U}$. Instead, we follow the standard practice of pruning the image nodes which have low probability of being classified as one of the n scene classes. Specifically, we evaluate the unary term (Φ), that represents the confidence of assigning label to an image, for the entire $\mathcal{U}$ and use it to prune out and keep only the top-N candidate images for each class. In our experiments, we set N to 3 times the number of instances to be promoted.

While pruning image nodes reduces the search space, exact inference still remains intractable. However, approximate inference techniques like loopy belief propagation or Gibbs sampling can be used to find the most likely assignments. In this work, we compute the marginals at each node by running one iteration of loopy belief propagation on the reduced graph. This approximate inference gives us the confidence of candidate class labeling for each image incorporating scene, attribute and comparative-attribute constraints. Now we select C most confidently labeled images for each class ($\mathcal{U}^l$), add them to $(\mathcal{L} \cup \mathcal{U}^l) \rightarrow \mathcal{L}$ (and remove from $(\mathcal{U} \setminus \mathcal{U}^l) \rightarrow \mathcal{U}$) and re-train our classifiers.

4.1 Scene, Attribute and Comparative Attribute Classifiers

We now describe the classifiers used for scenes, attributes and comparative attributes.

Scene and Attribute classifier: Our scene category classifiers as well as attribute classifier are trained using boosted decision trees [27]. We use 20 boosted trees with 8 internal nodes for scene classifier and 40 boosted trees with 8 internal nodes for training our attributes. These classifiers were trained on the 2049 dimensional feature vector from [28]. Our image feature includes 960D GIST [29] features, 75D RGB features (image is resized to 5×5) [3], 30D histogram of line lengths, 200D histogram of orientation of lines and 784D 3D-histogram Lab color space ($14 \times 14 \times 4$).

Comparative Attribute Classifier: Given a pair of images (x_i, x_j), the goal of comparative attributes classifier is to predict whether the pair satisfies comparative relationships such as "more circular" and "has more indoor space". To model and incorporate

comparative attributes, we follow the approach proposed in [9] and train classifiers over differential features ($x_i - x_j$). We train the comparative attribute classifiers using ground truth pair of images that follow such relationships and for the negative data we use random pair of images and inverse relationships. We used a boosted decision tree classifier with 20 trees and 4 internal nodes.

5 Experiments

We now present experimental results to demonstrate the effectiveness of constraints in bootstrapping based approach. We first present a detailed experimental analysis of our approach using the fully labeled SUN dataset [30]. Using a completely labeled dataset allows us to evaluate the quality of unlabeled images being added to our classifiers and how it affects the performance of our system. Finally, we evaluate our complete system on a large scale dataset which uses approximately 1 million unlabeled images to improve the performance of scene classifiers. For all the experiments we use a fixed vocabulary of scene classes, attributes and comparative attributes as described below.

Vocabulary: We evaluate the performance of our coupled bootstrapping approach in learning 15 scene categories[3] randomly chosen from from SUN dataset . These classes are: auditorium, amphitheater, banquet hall, barn, bedroom, bowling alley, bridge, casino indoor, cemetery, church outdoor, coast, conference room, desert sand, field cultivated and restaurant. We used 19 attribute classes: horizon visible, indoor, has water, has building, has seat rows, has people, has grass, has clutter, has chairs, is man-made, eating place, fun place, made of stone, meeting place, livable, part of house, relaxing, animal-related, crowd-related. We used 10 comparative attributes: is more open, had more space (indoor), had more space (outdoor), has more seating space, has larger structures, has horizontally longer structures, has taller structures, has more water, has more sand and has more greenery. The relationship between scene category and attributes were defined using a human annotator (see the website for list of relationships).

Baselines: The goal of this paper is to show the importance of additional constraints in semi-supervised domain. We believe these constraints should improve the performance irrespective of the choice of a particular approach. Therefore, as a baseline, we compare the performance of our constrained bootstrapping approach with two versions of standard bootstrapping approach: one uses independent multiple binary classifiers and the other uses multi-class scene classifiers.

For our first set of experiments, we also compare our constrained bootstrapping framework to the state-of-the-art SSL technique for image classification based on *eigen functions* [24]. Following experimental settings from [24], we map the GIST descriptor for each image down to a 32D space using PCA, use $k = 64$ eigen functions with $\lambda = 50$ and $\epsilon = 0.2$ for computing the Laplacian (see [24] for details).

Evaluation Metric: We evaluate the performance of our approach using two metrics. Firstly, we evaluate the performance of our trained classifier in terms of Average

[3] We expect our approach to scale and perform better with more categories: increasing the number of categories will add more constraints and enforce extensive sharing, which is exactly what our approach exploits.

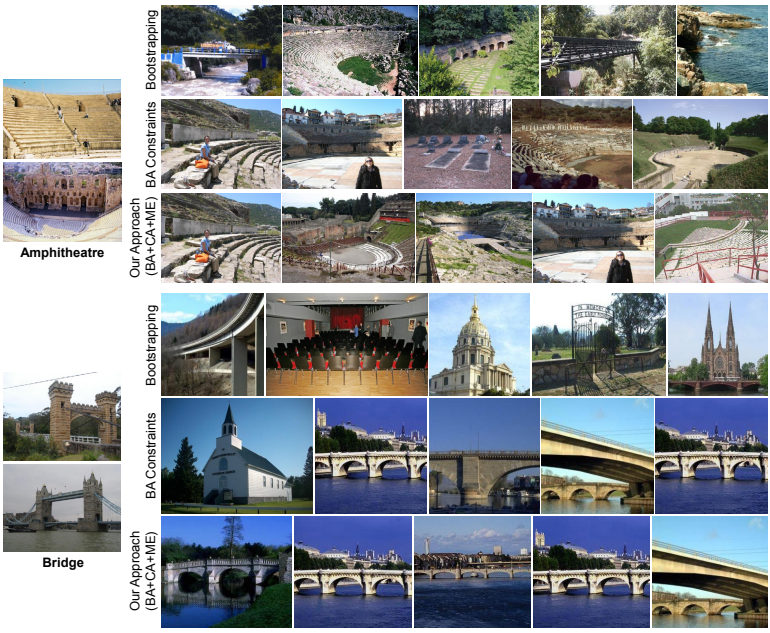


Fig. 3. Qualitative Results: We demonstrate how Binary Attribute (middle row) constraints and Comparative Attribute (bottom row) constraints help us promote better instances as compared to naïve Bootstrapping (top row)

-Precision (AP) at each iteration on a held-out test dataset. Secondly, for the small-scale experiments (sections 5.1 and 5.2), we also evaluate purity of the promoted instances in terms of the fraction of correctly labeled images.

5.1 Experiment 1: Using Pre-trained Attribute Classifiers

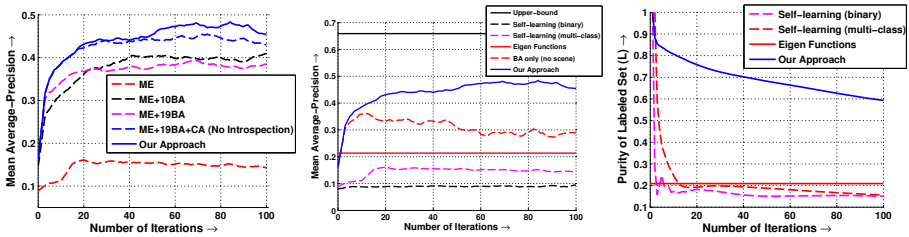
We first evaluate the performance of our approach on SUN dataset. We train the 15 scene classifiers using 2 images each (labeled set). We want to observe the effect of these constraints when the attribute and comparative attribute classifiers are not re-trained at each iteration. Therefore, in this case, we used fixed pre-trained attribute classifiers and relative attribute classifiers. These classifiers were trained on 25 examples each (from a held-out dataset). Our unlabeled dataset consists of 18,000 images from SUN dataset. Out of these 18K images, 8.5K images are from these 15 categories and the remainder are randomly sampled from the rest of the dataset. At each iteration, we add 5 images per category from the unlabeled dataset to the classifier.

Figure 3 shows examples of how each constraint helps to select better instances that should be added to the classifier. The bootstrapping approach clearly faces semantic drift, as it adds "bridge", "coastal" and "park" images to the "amphitheater" classifier. It is the presence of binary attributes such as 'has water' and 'has greenery' that help us to reject these bad candidates. While binary attributes do help to prune lot

Fig. 4. Qualitative results showing selected candidates for our approach at iteration 1, 10, 40 and 90. Notice that during the final iterations, there are errors such as bedroom images being added to conference rooms etc.

of bad instances, they sometimes promote bad instances like the "cemetery" image (3^{rd} in 2^{nd} row). However, comparative attributes help us clean such instances. For example, the "cemetery" image is rejected since it has less circular structure. Similarly, the "church" image is rejected since it does not have the long horizontal structures compared to other bridge images. Interestingly, our approach does not overfit to the seed examples and can indeed cover a greater diversity, thus increasing recall. For example, in Figure 4, the seed examples for banquet hall include close-view of tables but as iterations proceed we incorporate distant views of banquet hall (eg., 3rd image in iteration 1 and 40).

Next, we quantitatively evaluate the importance of each constraint in terms of performance on held-out test data. Figure 5(a) shows the performance of our approach with different combinations of constraints. Our system shows significant improvement in performance by adding attribute-based constraints. In fact, using a randomly chosen set of ten attributes (ME+10BA) seems to provide enough constraints to avoid semantic drift. Adding another nine attributes to the system does provide some improvement during the initial iterations (ME+19BA). However, at later stages, the effect of these attributes saturate. Finally, we evaluate the importance of comparative attributes by comparing the performance of our system with and without CA constraint. Adding CA constraint does provides significant boost (6-7%) in performance. Figure 5(b) shows the comparison of our full system with other baseline approaches. Our approach shows significant improvement over all the baselines which include self-learning approaches based on independent binary classifiers, self-learning based on multi-class classifier and Eigen-functions [24]. We also show the upper-bound on the performance of our approach, which is achieved if all the unlabeled data is manually labeled and used to train the scene classifiers. Since the binary attribute classifiers are pre-trained on some other data, it would be interesting to analyze the performance of these classifiers alone



(a) Evaluating Constraints (b) Comparisons with Baselines (c) Purity based Comparison

Fig. 5. Quantitative Evaluations: (a) We first evaluate importance of each constraint in our system using control experiments. (b) and (c) Show the comparison of our approach against standard baselines in terms of performance on held-out test data and purity of added instances respectively.

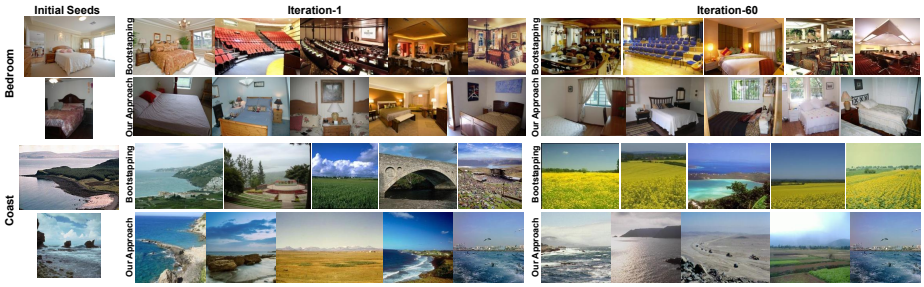


Fig. 6. Selected candidates for baseline and our approach at iterations 1 and 60

(and without scene classifiers). Our approach performs significantly better than just using attributes alone. This indicates that coupling does provide constraints and help in better labeling of unlabeled data. We also compare the performance of our full system in terms of purity of added instances (See Figure 5(c)).

5.2 Experiment 2: Co-training Attributes and Comparative Attributes

In the previous experiment we showed how each constraint is useful in improving the performance of the semi-supervised approach. To isolate the reasons behind the performance boost, we used fixed pre-trained attribute and comparative attribute classifiers. In this experiment, we train our own attribute and comparative attribute classifiers. These classifiers are trained using the **same 30 images (15 categories** $\times 2$ **images)** which were used to train the initial scene classifiers. Now, we use the co-training setup where we retrain these attribute and comparative attribute classifiers at each iteration using the new images from unlabeled dataset.

Figure 6 shows qualitative results of image instances which are added to the classifiers at iterations 1 and 60. The qualitative results show how the baseline approach suffers from semantic drift and adds "auditorium" to "bedrooms" and "field" to "coast". On the other hand, our approach is more robust and even at 60^{th} iteration adds good instances to the classifier. Figure 7 shows the comparison of our approach against

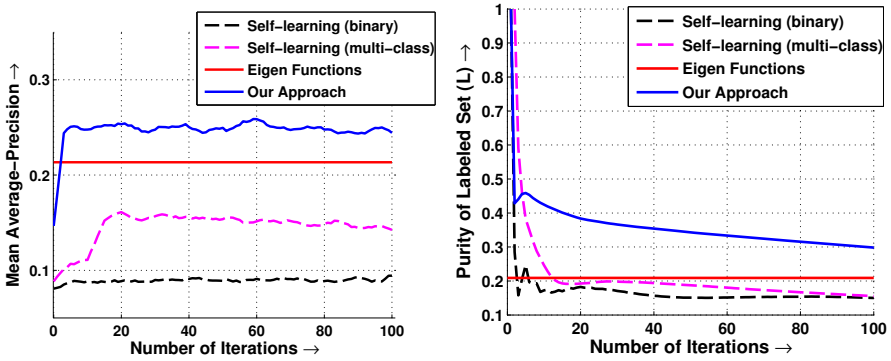


Fig. 7. Quantitative Evaluations: We evaluate our approach against standard baselines in terms of (a) mean AP over all scene classes, (b) purity of added instances

Table 1. Quantitative Results on Large Scale Semi-Supervised Learning (AP Scores)

	Amphitheater	Auditorium	Banquet Hall	Barn	Bedroom	Bowling Alley	Bridge	Casino indoor	Cemetery	Church outdoor	Coast	Conference Room	Desert Sand	Field Cultivated	Restaurant	Mean
Iteration-0	0.517	**0.317**	0.260	0.309	0.481	0.602	0.144	0.647	0.449	0.482	0.539	0.384	0.716	0.700	0.270	0.455
Self (Binary)	0.557	0.269	**0.324**	0.255	0.458	0.590	0.156	0.644	0.453	0.499	0.466	0.317	0.690	0.572	0.241	0.433
Self (Multi-Class)	0.488	0.254	0.290	0.261	0.443	0.601	0.162	**0.655**	**0.509**	0.475	0.548	0.322	0.733	0.657	0.303	0.447
Our Approach	**0.571**	0.298	0.302	**0.352**	**0.521**	**0.627**	**0.209**	0.650	0.506	**0.506**	**0.571**	**0.391**	**0.786**	**0.702**	**0.311**	**0.487**

baselines. Notice that our approach outperforms all the baselines significantly even though in this case we used the same seed examples to train the attribute classifier. This shows that attribute classifiers can pool information from multiple classes to help avoid semantic drift.

5.3 Experiment 3: Large Scale Semi-Supervised Learning

In the two experiments discussed above, we demonstrated the importance and effectiveness of adding different constraints to the bootstrapping framework. As a final experiment, we now demonstrate the utility of such constraints for large scale learning. We start with 25 seed examples from SUN dataset for each of the 15 categories. Our unlabeled dataset consists of one million images selected from the imagenet dataset [31]. At each iteration, we add 10 images per category from unlabeled dataset to the classifier. Table 1 shows the performance of learned scene classifiers after 100 iterations. The constrained bootstrapping approach not only improves the performance by 3.2% but also outperforms all the baselines significantly.

Acknowledgments. The authors would like to thank Tom Mitchell, Martial Hebert, Varun Ramakrishna and Debadeepta Dey for many helpful discussions. This work was supported by Google, ONR-MURI N000141010934 and ONR Grant N000141010766.

References

1. von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: SIGCHI Conference on Human Factors in Computing Systems (2004)
2. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: A database and web-based tool for image annotation. IJCV (2009)
3. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: a large database for non-parametric object and scene recognition. PAMI (2008)
4. Curran, J.R., Murphy, T., Scholz, B.: Minimising semantic drift with mutual exclusion bootstrapping. In: Conference of the Pacific Association for Computational Linguistics (2007)
5. Carlson, A., Betteridge, J., Hruschka Jr., E.R., Mitchell, T.M.: Coupling semi-supervised learning of categories and relations. In: NAACL HLT Workskop on SSL for NLP (2009)
6. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR (2009)
7. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
8. Parikh, D., Grauman, K.: Relative attributes. In: ICCV (November 2011)
9. Gupta, A., Davis, L.S.: Beyond Nouns: Exploiting Prepositions and Comparative Adjectives for Learning Visual Classifiers. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 16–29. Springer, Heidelberg (2008)
10. Rastegari, M., Farhadi, A., Forsyth, D.: Attribute Discovery via Predictable Discriminative Binary Codes. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 876–889. Springer, Heidelberg (2012)
11. Parkash, A., Parikh, D.: Attributes for Classifier Feedback. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 355–369. Springer, Heidelberg (2012)
12. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr., E.R., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: AAAI (2010)
13. Vijayanarasimhan, S., Grauman, K.: Large-scale live active learning: Training object detectors with crawled data and crowds. In: CVPR (2011)
14. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Active learning with gaussian processes for object categorization. In: ICCV (2007)
15. Qi, G.J., Hua, X.S., Rui, Y., Tang, J., Zhang, H.J.: Two-dimensional active learning for image classification. In: CVPR (2008)
16. Joshi, A., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: CVPR (2009)
17. Siddiquie, B., Gupta, A.: Beyond active noun tagging: Modeling contextual interactions for multi-class active learning. In: CVPR (2010)
18. Russell, B., Freeman, W., Efros, A., Sivic, J., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: CVPR (2006)
19. Kang, H., Hebert, M., Kanade, T.: Discovering object instances from scenes of daily living. In: ICCV (2011)
20. Zhu, X.: Semi-supervised learning literature survey. Technical report, CS, UW-Madison (2005)
21. Riloff, E., Jones, R.: Learning dictionaries for information extraction by multi-level bootstrapping. In: AAAI (1999)
22. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: COLT (1998)
23. Ebert, S., Larlus, D., Schiele, B.: Extracting Structures in Image Collections for Object Recognition. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 720–733. Springer, Heidelberg (2010)

24. Fergus, R., Weiss, Y., Torralba, A.: Semi-supervised learning in gigantic image collections. In: NIPS (2009)
25. Guillaumin, M., Verbeek, J., Schmid, C.: Multimodal semi-supervised learning for image classification. In: CVPR (2010)
26. Tighe, J., Lazebnik, S.: Understanding scenes on many levels. In: ICCV (2011)
27. Hoiem, D., Efros, A.A., Hebert, M.: Recovering surface layout from an image. IJCV (2007)
28. Hays, J., Efros, A.A.: Scene completion using millions of photographs. ACM Transactions on Graphics, SIGGRAPH (2007)
29. Oliva, A., Torralba, A.: Building the gist of a scene: the role of global image features in recognition. Progress in Brain Research (2006)
30. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: Sun database: Large scale scene recognition from abbey to zoo. In: CVPR (2010)
31. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)