Attributes for Classifier Feedback

Amar Parkash[1] and Devi Parikh[2]

[1] Indraprastha Institute of Information Technology (Delhi, India)
[2] Toyota Technological Institute (Chicago, US)

Abstract. Traditional active learning allows a (machine) learner to query the (human) teacher for labels on examples it finds confusing. The teacher then provides a label for only that instance. This is quite restrictive. In this paper, we propose a learning paradigm in which the learner communicates its belief (i.e. predicted label) about the actively chosen example to the teacher. The teacher then confirms or rejects the predicted label. More importantly, if rejected, the teacher communicates an explanation for why the learner's belief was wrong. This explanation allows the learner to propagate the feedback provided by the teacher to many unlabeled images. This allows a classifier to better learn from its mistakes, leading to accelerated discriminative learning of visual concepts even with few labeled images. In order for such communication to be feasible, it is crucial to have a language that both the human supervisor and the machine learner understand. *Attributes* provide precisely this channel. They are human-interpretable mid-level visual concepts shareable across categories *e.g*. "furry", "spacious", etc. We advocate the use of attributes for a supervisor to provide feedback to a classifier and directly communicate his knowledge of the world. We employ a straightforward approach to incorporate this feedback in the classifier, and demonstrate its power on a variety of visual recognition scenarios such as image classification and annotation. This application of attributes for providing classifiers feedback is very powerful, and has not been explored in the community. It introduces a new mode of supervision, and opens up several avenues for future research.

1 Introduction

Consider the scenario where a teacher is trying to teach a child how to recognize giraffes. The teacher can show example photographs of giraffes to the child, as well as many pictures of other animals that are not giraffes. The teacher can then only hope that the child effectively utilizes all these examples to decipher what makes a giraffe a giraffe. In computer vision (and machine learning in general), this is the traditional way in which we train a classifier to learn novel concepts.

It is clear that throwing many labeled examples at a passively learning child is very time consuming. Instead, it might be much more efficient if the child actively engages in the learning process. After seeing a few initial examples of giraffes and non-giraffes, the child can identify animals that it finds most confusing (say a camel), and request the teacher for a label for these examples where the labels are likely to be most informative. This corresponds to the now well studied active learning paradigm.

While this paradigm is more natural, there are several aspects of this set-up that seem artificially restrictive. For instance, the learner simply poses a query to the teacher,

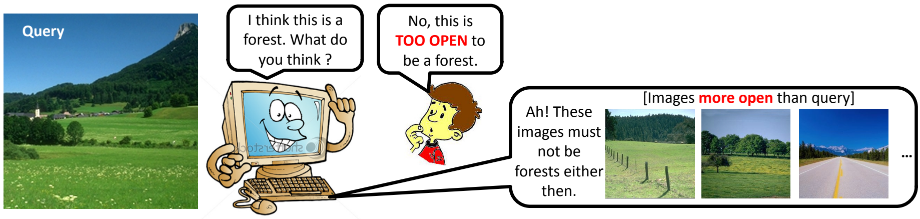

Query
I think this is a forest. What do you think ?
No, this is TOO OPEN to be a forest.
Ah! These images must not be forests either then.
[Images more open than query]
...

Fig. 1: We propose a novel use of attributes as a mode for a human supervisor to provide feedback to a machine active learner allowing it to better learn from its mistakes.

but does not communicate to the teacher what its current model of the world (or giraffes!) is. It simply asks "What is this?". Instead, it could just as easily say "I think this is a giraffe, what do you think?". The teacher can then encouragingly confirm or gently reject this belief. More importantly, if the teacher rejects the belief, he can now proactively provide a focussed explanation for *why* the learners' current belief is inaccurate. If the learner calls a tiger a giraffe, the teacher can say "Unfortunately, you're wrong. This is not a giraffe, because its neck is not long enough." Not only does this give the learner information about this particular example, it also gives the learner information that can be *transferred* to many other examples: all animals with necks shorter than this query animal (tiger) must not be giraffes. With this mode of supervision, the learner can learn what giraffes are (not) with significantly fewer labeled examples of (not) giraffes.

To allow for such a rich mode of communication between the learner and the teacher, we need a language that they both can understand. Attributes provide precisely this mode of communication. They are mid-level concepts such as "furry" and "chubby" that are shareable across categories. They are visual and hence machine detectable, as well as human interpretable. Attributes have been used for a variety of tasks such as multimedia retrieval [1–6], zero-shot learning of novel concepts [7, 8], describing unusual aspects of objects [9], face verification [10] and others.

In this paper, we advocate the novel use of attributes as a mode of communication for a human supervisor to provide feedback to a machine classifier. This allows for a natural learning environment where the learner can learn much more from each mistake it makes. The supervisor can convey his domain knowledge about categories, which can be propagated to unlabeled images in the dataset. Leveraging these images allows for effective discriminative learning even with a few labeled examples. We use a straightforward approach to incorporate this feedback in order to update the learnt classifiers. In the above example, all images that have shorter necks than the tiger are labeled as "not-giraffe", and the giraffe classifier is updated with these new negative examples. We demonstrate the power of this feedback in two different domains (scenes and faces) on four visual recognition scenarios including image classification and image annotation.

2 Related Work

We discuss our work relative to existing work on exploiting attributes, collecting richer annotation, active learning, providing a classifier feedback and discriminative guidance.

Attributes: Human-nameable visual concepts or attributes are often used in the multimedia community to build intermediate representations for images [1–6]. Attributes have also been gaining a lot of attention in the computer vision community over the past few years [7–24]. The most relevant to our work would be the use of attributes for zero-shot learning. Once a machine has learnt to detect attributes, a supervisor can teach it a novel concept simply by describing which attributes the concept does or does not have [7] or how the concept relates to other concepts the machine already knows [8]. The novel concept can thus be learnt without any training examples. However zero-shot learning is restricted to being generative in nature and the category models have to be built in the attribute space. Given a test image, its attribute values need to be predicted, and then the image can be classified. Hence, while very light on supervision, zero-shot learning often lags in classification performance. Our work allows for the marriage of using attributes to alleviate supervision effort, while still allowing for discriminative learning. Moreover, the feature space is not restricted to being the attributes space. This is because we use attributes during training to transfer category labels to other unlabeled image instances, as opposed to directly carving out portions of the attributes space where a category can live. At test time, attribute detectors need not be used.

Detailed annotation: With the advent of crowd-sourcing services, efforts are being made at getting many more images labeled, and more importantly, gathering more detailed annotations such as segmentation masks [25], parts [26], attributes [10, 20], pose [27], etc. In contrast, the goal of our work is not to collect deeper annotations (e.g. attribute annotations) for the images. Instead, we propose the use of attributes as an efficient means for the supervisor to broadly transfer category-level information to unlabeled images in the dataset, that the learner can exploit.

Active learning: There is a large body of work, in computer vision and otherwise, that utilizes active learning to efficiently collect data annotations. Several works consider the problem of actively interleaving requests for different forms of annotation: object and attribute labels [20], image-level annotation, bounding boxes or segmentations [28], part annotations and attribute presence labels [21], etc. To re-iterate, our work focuses only on gathering category-label annotations, but we use attributes as a mode for the supervisor to convey more information about the categories, leading to propagation of category labels to unlabeled images. The system proposed in [22] recognizes a bird species with the help of a human expert answering actively selected questions pertaining to visual attributes of the bird. Note the involvement of the user at test time. Our approach uses a human in the loop only during training. In [29], the learner actively asks the supervisor linguistic questions involving prepositions and attributes. In contrast, in our work, an informative attribute is selected by the supervisor that allows the classifier to better learn from its mistakes. Our work can be naturally extended to leveraging the provided feedback to simultaneously update the attribute models as in [20], or for discovering the attribute vocabulary in the first place as in [23].

Rationales: In natural language processing, works have explored human feature selection to highlight words relevant for document classification [30–32]. A recent work in vision [33] similarly solicits rationales from the annotator in the form of spatial information or attributes for assigning a certain label to the image. This helps in identifying relevant image features or attributes for that image. But similar to zero-shot learning, this can be exploited only if the feature space is the attributes space and thus directly maps to the mode of feedback used by the annotator. Our work also uses attributes to allow the supervisor to interactively communicate an explanation, but this explanation is propagated as category labels to many unlabeled images in the dataset. The feature space where category models are built is thus not constrained by the mode of communication i.e. attributes used by the annotator and can hence be non-semantic and quite complex (e.g. bag-of-words, gist, etc.).

Focussed discrimination: The role of the supervisor in our work can be viewed as that of providing a discriminative direction that helps eliminate current confusions of the learner. Such a discriminative direction is often enforced by mining hard negative examples that violate the margin in large-margin classifiers [34], or is determined after-the-fact to better understand the learnt classifier [35]. In our work, a human supervisor provides this direction by simply verbalizing his semantic knowledge about the domain to steer the learner away from its inaccurate beliefs.[3]

3 Proposed Approach

We consider a scenario where a supervisor is trying to teach a machine visual concepts. The machine has already learnt a vocabulary of attributes relevant to the domain. As learning aid, there is a pool of unlabeled images that the supervisor will label over the course of the learning process for the classifier to learn from. The learning process starts with the learner querying the supervisor for a label for a random example in the pool of unlabeled images. At each subsequent iteration, the learner picks an image from the unlabeled pool that it finds most confusing (e.g. a camel image in the giraffe learning example). It communicates its own belief about this image to the supervisor in the form of a predicted label for this image. The supervisor either confirms or rejects this label. If rejected, the supervisor provides a correct label for the query. He also communicates an explanation using attributes for why the learner's belief was wrong. Note that the feedback can be relative to the category depicted in the query image ("necks of tigers are not long enough to be giraffes") or relative to the specific image instance ("this image is too open to be a forest image", see Fig. 1). The learner incorporates both the label-feedback (whether accepted or rejected with correction) and the attributes-based explanation (when applicable) to update its models. And the process continues.

We train a binary classifier $h_k(\boldsymbol{x}), k \in \{1 \ldots K\}$ for each of the K categories to be learnt. At any point during the learning process, there is an unlabeled pool of images and a training set $T_k = \{(\boldsymbol{x}^k, y^k)\}$ of labeled images for each classifier. The labeled images $\boldsymbol{x}^k$ can lie in any feature space, including the space of predicted attribute values $\mathbb{R}^M$. The class labels are binary $y^k \in \{+1, -1\}$. In our implementation we use RBF SVMs as the binary classifier with probability estimates as output. If $p_k(\boldsymbol{x})$ is the probability

[3] In similar spirit, contemporary work at this conference [36] uses attributes to prevent a semi-supervised approach from augmenting its training data with irrelevant candidates from an unlabeled pool of images, thus avoiding semantic-drift.

the classifier assigns to image $\boldsymbol{x}$ belonging to class k, the confusion in $\boldsymbol{x}$ is computed as the entropy H of the distribution $\tilde{p}(\boldsymbol{x})$, where

$$\tilde{p}_k(\boldsymbol{x}) = \frac{p_k(\boldsymbol{x})}{\sum_{k=1}^{K} p_k(\boldsymbol{x})}. \tag{1}$$

At each iteration in the learning process, the learner actively selects an unlabeled image $\boldsymbol{x}^*$ with maximum entropy.

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, H(\tilde{p}(\boldsymbol{x})) \tag{2}$$

Note that there are two parts to the supervisor's response to the query image. The first is label-based feedback where the supervisor confirms or rejects and corrects the learners predicted label for the actively chosen image instance $\boldsymbol{x}^*$. And the second is an attributes-based explanation that is provided if the learners predicted label was incorrect and thus rejected. We now discuss how the attributes-based feedback is incorporated. The label-based feedback can be interpreted differently based on the application at hand, and we discuss that in Section 4.

3.1 Incorporating Attribute-based Explanation

Let's say the learner incorrectly predicts the label of actively chosen image $\boldsymbol{x}^*$ to be l. The supervisor identifies an attribute a_m that he deems most appropriate to explain to the learner why $\boldsymbol{x}^*$ does not belong to l. In this work we consider two simple forms of explanation. The supervisor can either say "$\boldsymbol{x}^*$ is too a_m to be l" or "$\boldsymbol{x}^*$ is not a_m enough to be l", whichever be the case.

In the former case, the learner computes the strength of a_m in $\boldsymbol{x}^*$ as $r_m(\boldsymbol{x}^*)$, where r_m is a pre-trained attribute strength predictor for attribute a_m (Section 3.2). The learner identifies all images in the currently unlabeled pool of images $\tilde{U}$ with attribute strength of a_m more than $r_m(\boldsymbol{x}^*)$. Clearly, if $\boldsymbol{x}^*$ is too a_m to be l, all images depicting a higher strength of a_m must not be l either (Fig. 1). Hence the training data T_l is updated to be $\hat{T_l} = T_l \cup \{(\boldsymbol{x}, -1)\}\ \forall \boldsymbol{x} \in \tilde{U}\ s.t.\ r_m(\boldsymbol{x}) \geq r_m(\boldsymbol{x}^*)$. Similarly, for the latter form of feedback, $\hat{T_l} = T_l \cup \{(\boldsymbol{x}, -1)\}\ \forall \boldsymbol{x} \in \tilde{U}\ s.t.\ r_m(\boldsymbol{x}) \leq r_m(\boldsymbol{x}^*)$. The category-label information is thus propagated to other images in $\tilde{U}$ aside from just $\boldsymbol{x}^*$.

As is evident, the supervisor provides an explanation only when the classifier is wrong. Arguably, it seems wasteful to request an explanation when the classifier is already right. In addition, the form of feedback we consider only explains why an image is *not* a certain class. As a result only negative labels are propagated to images. Our approach can easily also incorporate explanations that explain why an image does belong to a certain class. We argue that more often than not, several factors come together to make an image a certain concept *e.g.* an image is a coast image if it is open, natural, displays the horizon and a body of water. Hence explaining to the system why an image is a certain concept can be quite cumbersome. On the other hand, the supervisor needs to identify only one reason why an image is not a certain class. Our formulation can be easily extended to incorporate multi-attribute explanations if the supervisor so chooses.

Influence of imperfect attribute predictors: The updated $\hat{T}_l$ may contradict the underlying ground truth labels (not available to the system). For instance, when the supervisor says "$\boldsymbol{x}^*$ is too open to be a forest", there may be images in the unlabeled pool $\tilde{U}$ that are more open than $\boldsymbol{x}^*$ but are in fact forest images. Or, due to faulty attribute predictors, a forest image that is less open than the query image is predicted to be more open, and is hence labeled to be not forest. Hence, when in conflict, the label-based feedback from the supervisor (which is the ground truth by definition) overrides the label of an image that was or will be deduced from any past or future attributes-based feedback. In this way, if the entire pool of unlabeled images U were sequentially presented to the supervisor, even with inaccurate attributes-based feedback or faulty attribute predictors, it will be correctly labeled. Of course, the scenario we are proposing is one where the supervisor need not label all images in the dataset, and employ attributes-based feedback instead to propagate labels to unlabeled images. Hence, in any realistic setting, some of these labels are bound to be incorrectly propagated. Our approach relies on the assumption that the benefit obtained by propagating the labels to many more images outweighs the harm done by inaccuracies in the propagation caused by faulty attribute predictors. This is reasonable, especially since we use SVM classifiers to build our category models, which involve maximizing a *soft* margin making the classifier robust to outliers in the labeled training data. Of course, if the attribute predictors are severely flawed (in the extreme case: predicting the opposite attribute than what they were meant to predict), this would not be the case. But as we see in our results in Section 6, in practice the attribute predictors are reliable enough allowing for improved performance. Since the attribute predictors are pre-trained, their effectiveness is easy to evaluate (perhaps on held-out set from the data used to train the attributes in the first place). Presumably one would not utilize severely faulty attribute predictors in real systems. Recall, attributes have been used in literature for zero-shot learning tasks [7, 8] which also hinge on reasonable attribute predictors. We now describe our choice of attribute predictors.

3.2 Attributes Predictors

The feedback provided by the supervisor relates the query image actively selected by the learner to the concept the learner believes the image depicts. Hence relative attributes [8] are a natural choice. They were shown to respect relative judgements from humans more faithfully than the score of a binary classifier, which is desirable in our approach. We provide a brief overview of relative attributes below.

Suppose we have a vocabulary of M attributes $A = \{a_m\}$. These attributes are mid-level concepts that can be shared across the categories the supervisor is trying to teach the machine. For celebrities, relevant attributes may be "age" , "chubby", etc. while for scenes they may be "open", "natural", etc. This vocabulary of attributes is learnt offline only once, using a set of training images $I = \{i\}$ represented in $\mathbb{R}^n$ by feature-vectors $\{\boldsymbol{x_i}\}$. For each attribute, we are given two forms of supervision: a set of ordered pairs of images $O_m = \{(i, j)\}$ and a set of un-ordered pairs $S_m = \{(i, j)\}$ such that $(i, j) \in O_m \implies i \succ j$, *i.e.* image i has a stronger presence of attribute a_m than j, and $(i, j) \in S_m \implies i \sim j$, *i.e.* i and j have similar relative strengths of a_m. Either O_m or S_m, but not both, can be empty. We wish to learn a ranking function

$r_m(\boldsymbol{x_i}) = \boldsymbol{w_m^T x_i}$ for $m = 1, \ldots, M$, such that the maximum number of the following constraints is satisfied:

$$\forall (i, j) \in O_m : \boldsymbol{w_m^T x_i} > \boldsymbol{w_m^T x_j}, \forall (i, j) \in S_m : \boldsymbol{w_m^T x_i} = \boldsymbol{w_m^T x_j}. \tag{3}$$

This being an NP hard problem a relaxed version is solved using a large margin learning to rank formulation similar to that of Joachims [37] and adapted in [8]. With this, given any image $\boldsymbol{x}$ in our pool of unlabeled images U, one can compute the relative strength of each of the M attributes as $r_m(\boldsymbol{x}) = \boldsymbol{w_m^T x}$.

4 Applications

We consider four different applications to demonstrate our approach. For each of these scenarios, the label-based feedback is interpreted differently, but the attributes-based feedback has the same interpretation as discussed in Section 3.1. The label-based feedback carries different amounts of information, and thus influences the classifiers to varying degrees. Recall that via the label-based feedback, the supervisor confirms or rejects and corrects the classifier's predicted label for the actively chosen image $\boldsymbol{x}^*$.

Classification: This is the classical scenario where an image is to be classified into only one of several pre-determined number of classes. In this application, the label-based feedback is very informative. It not only specifies what the image is, it also indicates what the image is *not* (all other categories in the vocabulary). Let's say the label predicted by the classifier for the actively chosen image $\boldsymbol{x}^*$ is l. If the classifier is correct, the supervisor can confirm the prediction. This would result in $\hat{T}_l = T_l \cup \{(\boldsymbol{x}^*, 1)\}$, and $\hat{T}_n = T_n \cup \{(\boldsymbol{x}^*, -1)\} \; \forall \; n \neq l$. On the other hand, if the prediction was incorrect, the supervisor would indicate so, and provide the correct label q. In this case, $\hat{T}_q = T_q \cup \{(\boldsymbol{x}^*, 1)\}$ and $\hat{T}_n = T_n \cup \{(\boldsymbol{x}^*, -1)\} \; \forall \; n \neq q$ (note that this includes $\hat{T}_l = T_l \cup \{(\boldsymbol{x}^*, -1)\}$). In this way, every label-based feedback impacts *all* the K classifiers, and is thus very informative. Recall that the attributes-based explanation can only affect (the negative side) of one classifier at a time. We would thus expect that attributes-based feedback in a classification scenario would not improve the performance of the learner by much, since the label-based feedback is already very informative.

Large Vocabulary Classification: Next we consider is one that is receiving a lot of attention in the community today: a classification scenario where the number of classes is very large (and perhaps evolving over time); for example (celebrity) face recognition [10] or specialized fine-grained classification of animal species [16] or classifying thousands of object categories [38]. In this scenario, the supervisor can verify if the classifier's prediction of $\boldsymbol{x}^*$ is correct or not. But when incorrect, it would be very time consuming for or beyond the expertise of the supervisor to seek out the correct label from the large vocabulary of categories to provide as feedback. Hence the supervisor only confirms or rejects the prediction, and does not provide a correction if rejecting it. In this case, if the classifier's prediction l is confirmed, similar to classification, $\hat{T}_l = T_l \cup \{(\boldsymbol{x}^*, 1)\}$, and $\hat{T}_n = T_n \cup \{(\boldsymbol{x}^*, -1)\} \; \forall \; n \neq l$. However, if the classifier's prediction is rejected, we only have $\hat{T}_l = T_l \cup \{(\boldsymbol{x}^*, -1)\}$. In this scenario, the label feedback has less information when the classifier is wrong, and hence the classifier would have to rely more on the attributes-based explanation to learn from its mistakes.

Annotation: In this application, each image can be tagged with more than one class. Hence, specifying (one of the things) the image is, does not imply anything about what else it may or may not be. Hence when the classifier predicts that $\boldsymbol{x}^*$ has tag l (as one of its tags), if the supervisor confirms it we have $\hat{T}_l = T_l \cup \{(\boldsymbol{x}^*, 1)\}$ and if the classifier's prediction is rejected, we have $\hat{T}_l = T_l \cup \{(\boldsymbol{x}^*, -1)\}$. Note that only the l^{th} classifier is affected by this feedback, and all other classifiers remain unaffected. To be consistent across the different applications, we assume that at each iteration in the learning process, the classifier makes a single prediction, and the supervisor provides a single response to this prediction. In the classification scenario, this single response can affect all classifiers as we saw above. But in annotation, to affect all classifiers, the supervisor would have to comment on each tag individually (as being relevant to image $\boldsymbol{x}^*$ or not). This would be time consuming and does not constitute a single response. Since the label-based feedback is relatively uninformative in this scenario, we expect the attributes-based feedback to have a large impact on the classifier's ability to learn.

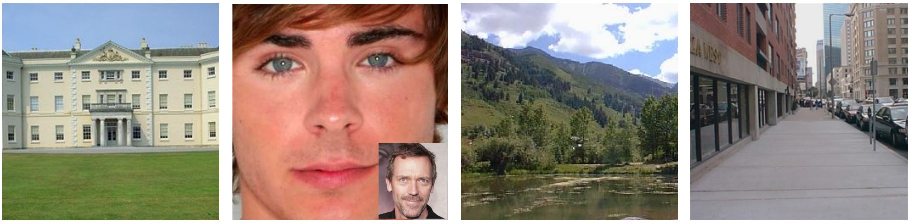
Biased Binary Classification: The last scenario we consider is that of a binary classification problem, where the negative set of images is much larger than the positive set. This is frequently encountered when training an object detector for instance, where there are millions of negative image windows not containing the object, but only a few hundreds or thousands of positive windows containing the object of interest. In this case, an example can belong to either the positive or negative class (similar to classification), and the label-based feedback is informative. However, being able to propagate the negative labels to many more instances via our attributes-based explanation can potentially accelerate the learning process. Approaches that mine hard negative examples [34] are motivated by related observations. However, what we describe above is concerned more with the bias in the class distribution, and not so much with the volume of data itself.

For all three classification-based applications, at test time, an image is assigned to the class that receives the highest probability. For annotation, all classes that have a probability greater than 0.5 are predicted as tags for the image.

5 Experimental Set-up

Datasets: We experimented with the two datasets used in [8]. The first is the outdoor scene recognition dataset [39] (Scenes) containing 2688 images from 8 categories: coast, forest, highway, inside-city, mountain, open-country, street and tall-building described via gist features [39]. The second dataset is a subset of the PubFig: Public Figures Face Database [10] (Faces) containing 772 images of 8 celebrities: Alex Rodriguez, Clive Owen, Hugh Laurie, Jared Leto, Miley Cyrus, Scarlett Johansson, Viggo Mortensen and Zac Efron described via gist and color features [8]. For both our datasets, we used the pre-trained relative attributes made publicly available by Parikh *et al.* [8]. For Scenes they include natural, open, perspective, large-objects, diagonal-plane and close-depth. For Faces: Masculine-looking, White, Young, Smiling, Chubby, Visible-Forehead, Bushy-Eyebrows, Narrow-Eyes, Pointy-Nose, Big-Lips and Round-Face.

Applications: For Scenes, we show results on all four applications described above. The binary classification problems are set up as learning only one of the 8 categories



(a) “This image is not perspective enough to be a street scene.”

(b) “Zac Effon is too young to be Hugh Laurie (bottom right).”

(c) open-country, mountain, forest

(d) street, inside-city, tall-building

Fig. 2: (a), (b) Example feedback collected from subjects (c), (d) Example images from Scenes dataset annotated with multiple tags by subjects.

through the entire learning process, which is selected at random in each trial. Images from the remaining 7 categories are considered to be negative examples. This results in more negative than positive examples. Many of the images in this dataset can be tagged with more than one category. For instance, many of the street images can just as easily be tagged as inside-city, many of the coast images have mountains in the background, etc. Hence, this dataset is quite suitable for studying the annotation scenario. We collect annotation labels for the entire dataset as we will describe next. For Faces, we show results for classification, binary classification and large vocabulary classification (annotation is not applicable since each image displays only one person’s face). For evaluation purposes, we had to collect exhaustive real human data as attributes-based explanations (more details below). This restricted the vocabulary of our dataset to be small. However the domain of celebrity classification lends itself well to large vocabulary classification.

Collecting attributes-based feedback: To gather attributes-based feedback from real users, we conducted human studies on Amazon Mechanical Turk. We gather all the data offline, allowing us to run automatic experiments without a live user in the loop, while still using *real* data from users. Note that the attribute feedback depends on the image being classified, and the label predicted by the classifier for that image. Since we can not anticipate ahead of time what the learner will predict as the label of an image if actively chosen as the query at some point during the learning process, we collect feedback for each image being classified into each one of the classes. As stated earlier, this restricted the number of categories our datasets can have. Note that the restriction on number of classes was only to systematically *evaluate* our approach, and is not a reflection of any limitations of our approach to scale to a large number of classes.

Collecting ground truth annotation labels: We had 5 subjects provide feedback for a total of 240 images (30 per category) from the Scenes dataset. Six attributes allow for 12 possible feedback statements for each category: “this image is too open to be a forest” or “this image is not natural enough to be a forest” and so on. The study asked subjects to select one of the 12 statements that was “most true”. An image was annotated with a statement for that category if the statement was chosen by at least 2 subjects. In 18% of the cases, we found that no such statement existed i.e. all 5 subjects selected a different statement. These cases were left unannotated. In our experiments, no attributes-based

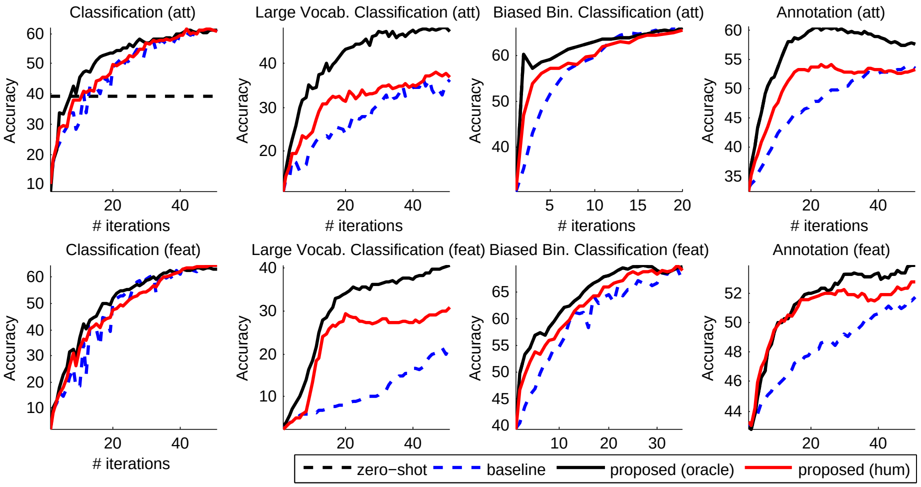


Fig. 3: Results for four different applications on the Scenes Dataset. See text for details.

explanation was provided for those particular image-category combinations. Hence the performance reported in our results is an underestimate.

For the Faces dataset, since the categories correspond to individuals, the attributes we use can be expected to be the same across all images in the category. Hence we collect feedback at the category-level. For each pair of categories, our 11 attributes allow for 22 statements such as "Miley Cyrus is too young to be Scarlett Johansson" or "Miley Cyrus is too chubby to be Scarlett Johansson" and so on. We showed subjects example images for the celebrities. Again, 5 subjects selected one of the statements that was "most true". We found that all category pairs had at least two subjects agreeing on the feedback. Note that during the learning process, unlike Scenes, every time an image from class q in the Faces dataset is classified as class l, the same feedback will be used. Example responses collected from subjects can be seen in Fig. 2 (a), (b).

We annotate all 2688 images in the Scenes dataset with multiple tags. For each image, 5 subjects on Amazon Mechanical Turk were asked to select any of the 8 categories that they think can be aptly used to describe the image. A tag was retained if at least 2 subjects selected it. For sake of completeness, for each image, we append its newly collected tags with its ground truth tag from the original dataset (although in most cases the ground truth label was already provided by subjects). On average, we have 1.6 tags for every image in the dataset. Example annotations can be seen in Fig. 2 (c), (d).

6 Results

We now present our experimental results on the Scenes and Faces datasets for the applications described above. We will compare the performance of our approach that uses both the label-based feedback (Section 4) as well as attributes-based explanations (Section 3.1) for each mistake the classifier makes, to the baseline approach of only using the label-based feedback (Section 4). Hence the difference in information in the label-

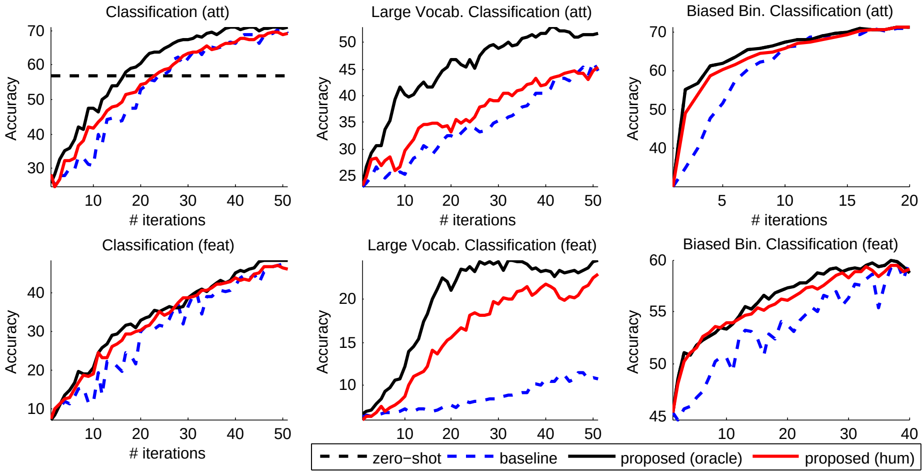


Fig. 4: Results for three different applications on the Faces Dataset. See text for details.

based feedback for the different applications affects *both* our approach and the baseline. Any improvement in performance is due to the use of attributes for providing feedback.

Our results are shown in Figs. 3 and 4[4]. We show results in the raw feature space (feat) *i.e.* gist for Scenes and gist+color for Faces, as well as in the relative attributes space (att). The results shown are average accuracies of 20 random trials. For each trial, we randomly select 50 images as the initial unlabeled training pool of images. Note that for Scenes, these 50 images come from the 240 images we had subjects annotate with feedback. The remaining images (aside from the 240 images) in the datasets are used as test images. On the x-axis are the number of iterations in the active learning process. Each iteration corresponds to one response from the supervisor as described earlier. For classification-based applications, accuracy is computed as the proportion of correctly classified images. For annotation, the accuracy of each of the tags (i.e. class) in the vocabulary is computed separately and then averaged.

As expected, for classification where the label-based feedback is already informative, attributes-based feedback provides little improvement. On the other hand, for annotation where label-based feedback is quite uninformative, attributes-based feedback provides significant improvements. This demonstrates the power of attributes-based feedback in propagating information (i.e. class labels), even if weak, to a large set of unlabeled images. This results in faster learning of concepts. Particularly with large vocabulary classification, we see that the attributes-based feedback boosts performance in the raw feature space more than the attributes space. This is because learning in high-dimensional space with few examples is challenging. The additional feedback of our approach can better tame the process. Attributes-based feedback also boosts performance in the biased binary classification scenario more than in classical classification. In datasets with a larger bias towards negative examples, the improvement may be even larger. Note that in many cases, the same classification accuracy can be reached via our approach using only a fifth of the user's time as with the baseline approach. Clearly,

[4] Number of iterations for binary classification was truncated because performance leveled off

attributes-based feedback allows classifiers to learn more effectively from their mistakes, leading to accelerated learning even with fewer labeled examples.

In Figs. 3 and 4 we also show an oracle-based accuracy for our approach. This is computed by selecting the feedback at each iteration that maximizes the performance on a validation set of images. This demonstrates the true potential our approach holds. The gap between the oracle and the performance using human feedback is primarily due to lack of motivation of mechanical turk workers to identify the best feedback to provide. We expect real users of the system to perform between the two solid (red and black) curves. In spite of the noisy responses from workers, the red curve performs favorably in most scenarios. This speaks to the robustness of the proposed straightforward idea of using attributes for classifier feedback.

In Figs. 3 and 4 we also show the accuracy of zero-shot learning using the direct attribute prediction model of Lampert *et al.* [7]. Briefly, each category is simply described in terms of which attributes are present and which ones are not. These binary attributes are predicted for a test image, and the image is assigned to the class with the most similar signature. We use the binary attributes and binary attributes-based descriptions of categories from [8]. Note that in this case, the binary attributes were trained using images from all categories, and hence the categories are *not* “unseen”. More importantly, the zero-shot learning descriptions of the categories was exactly what was used to train the binary attribute classifiers in the first place. We expect performance to be significantly worse if real human subjects provided the descriptions of these categories. We see that even this optimistic zero-shot learning performance compares poorly to our approach. While our approach also alleviates supervision burden (but not down to zero, of course) by using attributes-based feedback, it still allows access to discriminative learning, making the proposed form of supervision quite powerful.

7 Discussion and Conclusion

The proposed learning paradigm raises many intriguing questions. It would be interesting to consider the different strategies a user may use to identify what feedback should be provided to the classifier for its mistake. One strategy may be to provide the most accurate feedback. That is, when saying “this image is too open to be a forest”, ensuring that there is very little chance any image that is more open than this one could be a forest. This ensures that incorrect labels are not propagated across the unlabeled images, but there may be very few images to transfer this information to in the first place. Another strategy may be to provide very precise feedback. That is, when saying “the person in this image (Miley Cyrus) is too young to be Scarlett Johansson”, ensuring that Scarlett Johansson is very close in age to Miley Cyrus. This ensures that this feedback can be propagated to many celebrities (all that are older than Scarlett Johansson). Yet another strategy may be to simply pick the most obvious feedback. Of the many attributes celebrities can have, some are likely to stand out to users more than others. For example, if a picture of Miley Cyrus is classified as Zac Efron, most of us would probably react to the fact that the genders are not consistent. Analyzing the behavior of users, and perceptual sensitivity of users along the different attributes is part of future work. Developing active learning strategies for the proposed learning paradigm is also part of future work. An image should be actively selected with considerations that the supervisor will provide an explanation that would propagate to many other images relative to the selected image. One can also envision accounting for distance between images in the attributes space when propagating labels. For instance, if a particular image is too open to be a forest, an image that is significantly more open is extremely unlikely to be a forest. This leads to interesting calibration questions: when a human views an image as being “significantly” more open, what is the corresponding difference in the machine prediction of the relative attributes open? One could also explore alternative interfaces for the supervisor to provide feedback e.g. providing an automatically selected short-list of relevant attributes to select from. It would be worth exploring the connections between using attributes-based feedback to transfer information to unlabeled images and semi-supervised learning. Finally, exploring the potential of this novel learning paradigm for other tasks such as object detection is part of future work.

Conclusion: In this work we advocate the novel use of attributes as a mode of communication for the human supervisor to provide an actively learning machine classifier feedback when it predicts an incorrect label for an image. This feedback allows the classifier to propagate category labels to many more images in the unlabeled dataset besides just the individual query images that are actively selected. This in turn allows it to learn visual concepts faster, saving significant user time and effort. We employ a straight forward approach to incorporate this attributes-based feedback into discriminatively trained classifiers. We demonstrate the power of this feedback on a variety of visual recognition applications including image classification and annotation on scenes and faces. What is most exciting about attributes is their ability to allow for communication between humans and machines. This work takes a step towards exploiting this channel to build smarter machines more efficiently.

References

1. Smith, J., Naphade, M., Natsev, A.: Multimedia semantic indexing using model vectors. In: ICME. (2003)
2. Rasiwasia, N., Moreno, P., Vasconcelos, N.: Bridging the gap: Query by semantic example. IEEE Transactions on Multimedia (2007)
3. Naphade, M., Smith, J., Tesic, J., Chang, S., Hsu, W., Kennedy, L., Hauptmann, A., Curtis, J.: Large-scale concept ontology for multimedia. IEEE Multimedia (2006)
4. Zavesky, E., Chang, S.F.: Cuzero: Embracing the frontier of interactive visual search for informed users. In: Proceedings of ACM Multimedia Information Retrieval. (2008)
5. Douze, M., Ramisa, A., Schmid, C.: Combining attributes and fisher vectors for efficient image retrieval. In: CVPR. (2011)
6. Wang, X., Liu, K., Tang, X.: Query-specific visual semantic spaces for web image re-ranking. In: CVPR. (2011)
7. Lampert, C., 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. (2011)
9. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR. (2009)
10. Kumar, N., Berg, A., Belhumeur, P., Nayar, S.: Attribute and simile classifiers for face verification. In: ICCV. (2009)
11. Berg, T., Berg, A., Shih, J.: Automatic attribute discovery and characterization from noisy web data. In: ECCV. (2010)

12. Wang, J., Markert, K., Everingham, M.: Learning models for object recognition from natural language descriptions. In: BMVC. (2009)
13. Wang, G., Forsyth, D.: Joint learning of visual attributes, object classes and visual saliency. In: ICCV. (2009)
14. Wang, Y., Mori, G.: A discriminative latent model of object classes and attributes. In: ECCV. (2010)
15. Ferrari, V., Zisserman, A.: Learning visual attributes. In: NIPS. (2007)
16. Branson, S., Wah, C., Babenko, B., Schroff, F., Welinder, P., Perona, P., Belongie, S.: Visual recognition with humans in the loop. In: ECCV. (2010)
17. Fergus, R., Bernal, H., Weiss, Y., Torralba, A.: Semantic label sharing for learning with many categories. In: ECCV. (2010)
18. Wang, G., Forsyth, D., Hoiem, D.: Comparative object similarity for improved recognition with few or no examples. In: CVPR. (2010)
19. Mahajan, D., Sellamanickam, S., Nair, V.: A joint learning framework for attribute models and object descriptions. In: ICCV. (2011)
20. Kovashka, A., Vijayanarasimhan, S., Grauman, K.: Actively selecting annotations among objects and attributes. In: ICCV. (2011)
21. Wah, C., Branson, S., Perona, P., Belongie, S.: Multiclass recognition and part localization with humans in the loop. In: ICCV. (2011)
22. Branson, S., Wah, C., Babenko, B., Schroff, F., Welinder, P., Perona, P., Belongie, S.: Visual recognition with humans in the loop. In: ECCV. (2010)
23. Parikh, D., Grauman, K.: Interactively building a discriminative vocabulary of nameable attributes. In: CVPR. (2011)
24. Bourdev, L., Maji, S., Malik, J.: Describing people: A poselet-based approach to attribute classification. In: ICCV. (2011)
25. Russel, B., Torralba, A., Murphy, K., Freeman, W.: Labelme: a database and web-based tool for image annotation. IJCV (2008)
26. Farhadi, A., Endres, I., Hoiem, D.: Attribute-centric recognition for cross-category generalization. In: CVPR. (2010)
27. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: ICCV. (2009)
28. Vijayanarasimhan, S., Grauman, K.: Multi-level active prediction of useful image annotations for recognition. In: NIPS. (2008)
29. Siddiquie, B., Gupta, A.: Beyond active noun tagging: Modeling contextual interactions for multi-class active learning. In: CVPR. (2010)
30. Raghavan, H., Madani, O., Jones, R.: Interactive feature selection. In: IJCAI. (2005)
31. Druck, G., Settles, B., McCallum, A.: Active learning by labeling features. In: EMNLP. (2009)
32. Zaidan, O., Eisner, J., Piatko, C.: Using annotator rationales to improve machine learning for text categorization. In: NAACL - HLT. (2007)
33. Donahue, J., Grauman, K.: Annotator rationales for visual recognition. In: ICCV. (2011)
34. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. PAMI (2010)
35. Golland, P.: Discriminative direction for kernel classifiers. In: NIPS. (2001)
36. Shrivastava, A., Singh, S., Gupta, A.: Constrained semi-supervised learning using attributes and comparative attributes. In: ECCV. (2012)
37. Joachims, T.: Optimizing search engines using clickthrough data. In: KDD. (2002)
38. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09. (2009)
39. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV (2001)