Bootstrapping Fine-Grained Classifiers: Active Learning with a Crowd in the Loop

Genevieve Patterson[1] Grant Van Horn[2] Serge Belongie[2] Pietro Perona[3] James Hays[1]
[1] Brown University, [2]University of California, San Diego, [3]California Institute of Technology
{gen, hays}@cs.brown.edu gvanhorn@ucsd.edu sjb@cs.ucsd.edu
perona@caltech.edu

Abstract

We propose an iterative crowd-enabled active learning algorithm for building high-precision visual classifiers from unlabeled images. Our method employs domain experts to identify a small number of examples of a specific visual event. These expert-labeled examples seed a classifier, which is then iteratively trained by active querying of a non-expert crowd. These non-experts actively refine the classifiers at every iteration by answering simple binary questions about the classifiers' detections. The advantage of this approach is that experts efficiently shepherd an unsophisticated crowd into training a classifier capable of fine-grained distinctions. This obviates the need to label an entire dataset to obtain high-precision classifiers. We find these classifiers are advantageous for creating a large vocabulary of visual attributes for specialized taxonomies. We demonstrate our crowd active learning pipeline by creating classifiers for attributes related to North American birds and fashion.

1 Introduction

Object detection is a broad sub-field of computer vision. Traditionally, algorithms for object detection are trained using datasets where all the instances of an object have been labeled by human annotators. This annotation process scales both with the number of images in the dataset and with the number of objects the researcher intends to detect. This process becomes even more difficult when the labeling task is crowdsourced. Ideally, members of the crowd will be able to recognize and accurately annotate the objects in question, but in reality workers have different levels of competency and attention. As a result, efficient dataset construction has itself become a topic of research.

Active learning has been investigated as a way to limit the amount of required labeled training data. In previous research, active learning has been used to detect pedestrians and discriminate between object categories[1, 18, 5]. These earlier methods were successful at the time for identifying things that had a distinctive appearance and were distinct from the background of the image, e.g. pedestrians on a street, or an animal or object centered in the frame with a plain background in the fashion of the Caltech-101 and 256 datasets.

As researchers have become more interested in real-world style images, a new class of feature has emerged. Unlike pixel level color and gradient features, larger sized features that capture a part or attribute of a larger object have become popular [11, 9, 16, 2]. Methods for training classifiers to detect these mid-level attribute features without a ground truth labeled dataset have been unsupervised up to this point [6].

Typically, attribute or part classifiers are created using datasets labeled by human annotators. State of the art classifiers for parts and attributes related to vehicles, animals, scenes, and faces have been trained using large sets of ground truth labels for supervision [9, 17, 11]. Unfortunately, creating datasets with detailed annotations is costly and time consuming. Obtaining high quality results from crowdsourced annotators, who are often untrained and of unknown provenance, is open research topic [21, 10, 12].

To obtain classifiers capable of identifying salient visual features, researchers have recently designed unsupervised and weakly supervised training pipelines[6, 19, 14]. These methods can also be costly and time consuming, as they require thousands of CPU hours to train classifiers. After creating large vocabularies of mid-level classifiers, these methods still need a human to review the classifiers' detection behavior in order to understand the contextual meaning of each mid-level feature. The salient visual events identified by these methods may not even be conveniently described with language. While it is possible to use ensembles of these mid-level features for object classification, the discovered, mid-level structures don't necessarily have one to one mapping to high level visual concepts.

While non-nameable mid-level features are useful for K-way classification, we aim to provide models for nameable attributes and parts. Nameable attributes are instrumental in broader problems of high-level context understanding and reasoning [16].

We propose an approach to training nameable mid-level feature classifiers that neither requires large, labeled datasets nor massive computing architectures. We use a combination of experts and untrained crowd members to iteratively improve discriminative classifiers. We aim to make our method cheaper and faster than comparable methods that must first create a labeled dataset. We can also take advantage of adding new images to our training set at any point in the pipeline while other methods would have to label the image or re-initialize the unsupervised pipeline to ensure optimality. Our proposed method allows users to dynamically create new attribute classifiers in parallel and scalably add images from new datasets.

2 Related Work

Our proposed method produces classifiers for fine-grained visual phenomena. This type of visual phenomena is often described as attributes or parts in the literature. The objective of our iterative training method is to create classifiers that detect salient, discriminative phenomena. Existing methods for identifying these types of localized, discriminative visual events are often fully automatic methods that operate on labeled datasets [8, 3]. Other successful approaches have employed humans in the attribute discovery process [16, 17, 14]. These human-in-the-loop attribute discovery methods eventually use ground truth labels to create attribute classifiers.

In comparison, our method directly avails itself of expert knowledge on attributes that are essential for fine grained categorization within a given taxon. We ask the experts to provide a small number of exemplar images to seed the initial stage of classifier training. Using our method, experts in vision or other fields can create classifiers for attributes they themselves have identified. The domain experts can create multiple exemplar models of a given attribute if multiple orientations or poses are required to adequately describe the variability of an attribute or part.

The concept of creating multiple exemplar models for one higher-level visual phenomena was introduced by [15]. Malisiewicz et al. use a set of classifiers to describe one object. Each classifier in the set is trained to detect the object in a specific orientation. The output of several relatively simple exemplar classifiers can then be used in conjunction to identify an object with variable appearance. Related work in the automatic and unsupervised discovery of salient image patches has also shown that a library of classifiers, each describing a particular example of an object or attribute, can be useful in state of the art classification pipelines [6, 19].

Similarly, we propose a method for creating distinct exemplar classifiers that can be used to identify a more complex attribute. Each exemplar classifier is seeded with expert-identified images of the attribute or part in a canonical pose or orientation. At each iteration of the active learning process, untrained crowd workers only have to identify if a set of detections do or do not contain the given attribute. This enables the crowd to accurately respond to active learning queries that would otherwise be prohibitively domain-specific. Our method takes advantage of straightforward binary questions, guided by the expert exemplars, in order to elicit reliable results from the crowd. Similar research into parallelizing crowd micro-tasks has recently been described by [13, 4, 7].

As previously mentioned, active learning has been used to create pedestrian and object classifiers [1, 5]. Collins et al. envisioned their active learning method as a way to scalably construct datasets. Vijayanarasimhan and Grauman evaluated active learning as a way to manage effort and expense for crowd annotation [20]. We believe our method is the first to exploit the combined efforts of expert annotators and the crowd via active learning. In the next section, we describe the details of our method and demonstrate its potential for easy parallelization and cost effective creation of large numbers of attribute classifiers.

3 Bootstrapping Classifiers

We begin the bootstrapping process by identifying a group of experts who would like to create visual classifiers for their data. To demonstrate our pipeline, we consulted experts in two distinct domains, common birds of eastern North America and street style fashion. We consulted experts at Cornell University's Lab of Ornithology to obtain a dataset of bird images and a set of exemplars patches for the heads of different bird species. This dataset contains 4150 images. We crawled images and meta-data from the fashion blog *thesartorialist.com* for street style clothing images and attributes. This dataset contains 882 images. For this initial prototyping, a self-described fashion expert on our team created the exemplar patches for the fashion attributes.

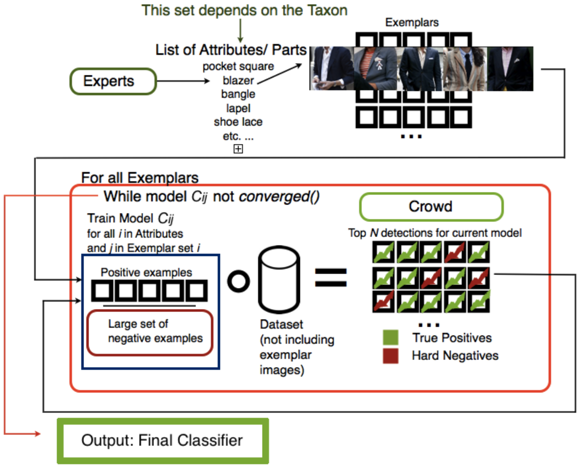

This set depends on the Taxon
Exemplars
List of Attributes/ Parts
Experts
pocket square
blazer
bangle
lapel
shoe lace
etc. ...
...
For all Exemplars
While model C_{ij} not *converged()*
Crowd
Train Model C_{ij}
for all i in Attributes
and j in Exemplar set i
Top N detections for current model
Positive examples
Large set of negative examples
Dataset (not including exemplar images)
...
True Positives
Hard Negatives
Output: Final Classifier

Figure 1: *Crowd Active Learning Pipeline*. This diagram illustrates the expert labeling pipeline in conjunction with Algorithm 1. A group of experts on a given taxon create a list of significant attributes. They provide approximately 5 examples of each attribute in different canonical poses for that attribute. This is the set of seed training examples, or 'exemplars' [15]. The process detailed in Algorithm 1 uses these exemplar sets to construct the final classifiers. Note the ∘ symbol indicates the function `orderDetections()` in Algo. 1.

Equipped with our attributes and exemplar images, we ran the pipeline illustrated in Fig. 1 and formalized in Algorithm 1. We took exemplars for 5 bird attributes and 3 fashion attributes and prompted non-expert members of our labs to answer our active queries for 5-10 iterations of the pipeline. We use a maximum number of iterations (10 in the case of birds and 5 in the case of fashion) if the classifier had not converged.

The interface for the active query component of our pipeline is shown in Fig. 2. In the following initial experiments, only responses from untrained lab members were collected. This interface will be used on Amazon Mechanical Turk without alteration.

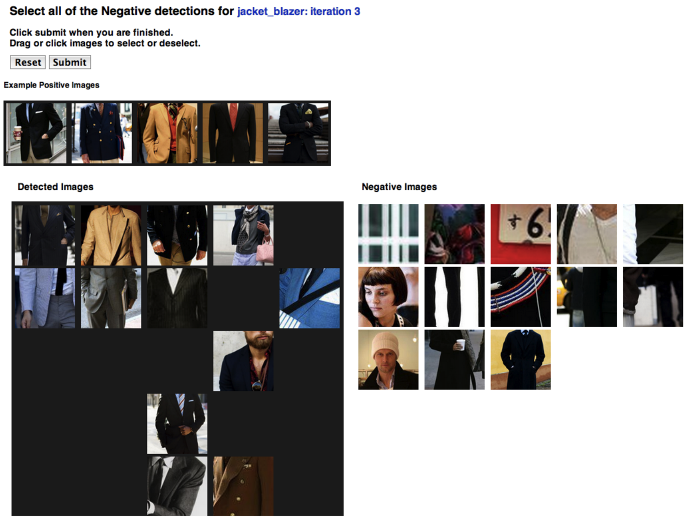
Figure 2 demonstrates how our pipeline simplifies the annotation question posed to the crowd. Instead of asking users to find and annotate a particular attribute in a large set of images where the attribute will likely be rare, we simply ask if a patch does or does not match a set of exemplars. This enables the crowd to accurately respond to queries that would otherwise be prohibitively domain-specific. The classifier and the crowd are able to collaborate to inductively learn the expert-defined attribute.

We chose to ask the crowd user to select negatives for two reasons. Firstly, adding hard negatives, patches that do not contain the attribute but were selected with high confident, to the dataset is one

Input: Dataset $\mathcal{D}$ of image patches, set of negative images from the wild $\mathcal{N}$
Output: Classifiers C for attributes A

```
$A \Leftarrow$ attributes                              ▷ acquired through consultation with experts
for $A_i \in A$ do
    $S_{ij} \Leftarrow$ seed exemplars of $A_i$ in pose $j$
end
for $A_i \in A$ do
    for $S_{ij} \in S_i$ do
        $C_{ij}$ = svmTrain( $S_{ij}, \mathcal{N}$)
        $N_{ij} = \emptyset \quad \forall i, j$                ▷ set of hard negatives is initially empty
        repeat
            $\mathcal{D}$= orderDetections($C_{ij}$, $\mathcal{D} - N_{ij}$)
            $N_{ij}$ = $N_{ij}$ ∪ hasNegatives($\mathcal{D}$)          ▷ crowdsourced method
            $C_{ij}$ = svmTrain( $S_{ij}$, $N_{ij} \cup \mathcal{N}$)
        until convergence()
    end
end
return $C, A$
```

Algorithm 1: Crowd Active Learning for Attribute Patches



Select all of the Negative detections for jacket_blazer: iteration 3

Click submit when you are finished.
Drag or click images to select or deselect.

Figure 2: *Active Query User Interface*. The UI shown in this figure implements the function `hasNegatives()` in Algo. 1. Users are shown the exemplar patches for the attribute in question. They are given a display of the top 200 detections from the validation set $\mathcal{D}$ (not all pictured). Users click all patches they believe are not visually and contextually similar to the exemplar patches.

way to implement the active learning technique of maximizing expected model change [18]. Hard negative examples will by definition change the learned hyperplane of the linear SVM we chose to use as our discriminative classifier. While a variable number of hard negatives are not guaranteed to maximally alter this learned hyperplane, they can qualitatively ensure a large deviation in the learned model at successive iterations of the pipeline. Secondly, we found that the crowd responds more reliably when they record the absence of an exemplar rather than its presence.

In this initial prototyping of our pipeline we demonstrated the method on few attributes, modestly sized datasets, and a local crowd. However, the following section shows promising results on this early venture.

4 Results

Figure 3 shows the exemplars used to train 3 bird and 3 fashion attributes. Beneath each set of exemplar patches are the top 5 most confident detections for the given attribute at successive iterations of the pipeline. Attributes for the Ruby Throated Hummingbird, jackets and blazers show an increasingly improving classifier. The attributes for Blue Jay and shorts seem to plateau with non-optimal results. In future work we plan to correct this behavior by improving the pixel level features used and greatly increasing the size of the training datasets. Because these attributes are rare, classifiers can only improve if the dataset providing patches is large enough for the attribute to occur dozens or more times.

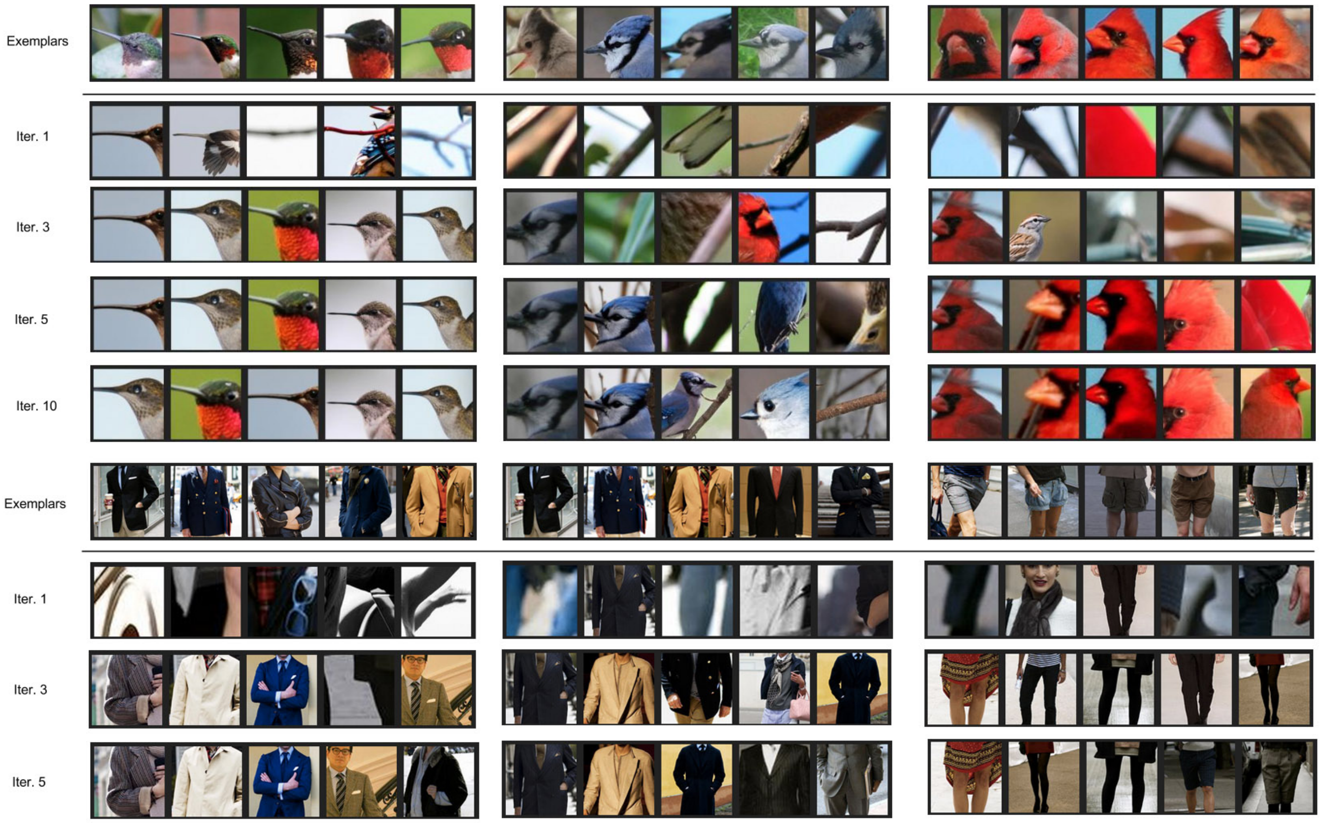


Figure 3: *Example detections at different iterations of the pipeline.* This figure shows the top 5 detections of the attribute classifiers at different iterations of the training pipeline. The attributes include the Ruby Throated Hummingbird Head, Blue Jay Head, Cardinal Head, jackets (general category), blazers, and shorts. The bird attribute classifiers were run for more iterations because the bird image dataset contained 4150 images while the street style fashion dataset only held 882.

To quantify the iterative performance of our crowd-trained classifiers, Fig. 4 plots the number of selected hard negatives against the iteration number. It is important to note that the 'positives' at each iteration may include patches that are visually similar but not actually the same attribute as the exemplar. For example the exemplars for Ruby Throated Hummingbirds show only adult males while the top detections are picking up juvenile males and females. The detector for shorts is also picking up skirts, which may be related to the small number of shorts in the fashion dataset. Given larger datasets and the combination of several exemplars for the same attribute, the top detections of our classifiers would be more likely to contain the correct attribute.

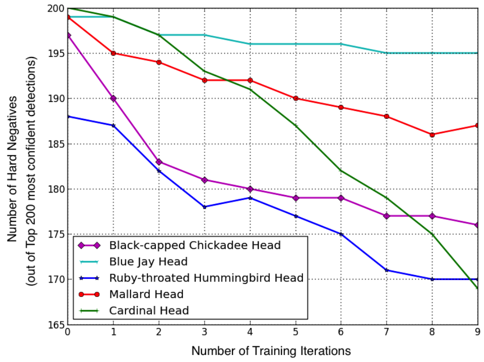


Figure 4: *Number of Hard Negatives versus Number of training iterations.* Results of the `hasNegatives()` in Algo. 1 on 5 different bird part attributes. Respondents were non-experts from the authors' lab. For most of the attributes, the classifiers are increasingly returning patches that either contain the correct attribute or are strongly similar to the exemplar attribute. Note that the validation set $\mathcal{D}$ begins with approx. 1 million images patches, with approx. 25 patches sharing exactly the same attribute and pose as the exemplars. For a visualization of these results, refer to Fig. 3. In the Blue Jay Head trial, the classifier is not finding many acceptable patches, even after 10 iterations. This is due to a combination of the classifier confusing the Blue Jay with birds that have similar head shape and the current prototype state of our color and gradient features, which are not sufficiently representing that attribute.

5 Discussion

In this work we presented an efficient approach to training nameable mid-level feature classifiers. Our approach uses experts to capture and share visual expertise with unskilled crowd sourced annotators via active learning. The pipeline we propose does not require a labeled dataset nor a massive computing architecture and is highly scalable. This makes our approach straight forward and cost-effective to deploy to new categories.

Our future work on this project will involve deploying our pipeline on dozens of bird and fashion attributes. We will expand our fashion dataset to include both runway and street style images to make a dataset containing tens of thousands of unlabeled images. After creating large number of classifiers, we will be able to compare our classifier performance to labels provided by untrained crowd members without any help from experts. These experiments will allow us to demonstrate that our method captures and shares visual expertise with the crowd, improving their ability to contribute to classifier training.

Acknowledgements. The authors would like to thanks Steve Branson and Vansh Kumar at Caltech for their support for these experiments. Genevieve Patterson is supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. Additional support provided by Google Focused Research Award.

References

[1] Y. Abramson and Y. Freund. Active learning for visual object recognition. Technical report, Technical report, UCSD, 2004.

[2] T. Berg and P. N. Belhumeur. Poof: Part-based one-vs-one features for fine-grained categorization, face verification, and attribute estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2013.

[3] T. Berg, A. Berg, and J. Shih. Automatic Attribute Discovery and Characterization from Noisy Web Data. *ECCV*, pages 663–676, 2010.

[4] L. B. Chilton, G. Little, D. Edge, D. S. Weld, and J. A. Landay. Cascade: crowdsourcing taxonomy creation. In *Proceedings of the 2013 ACM annual conference on Human factors in computing systems*, pages 1999–2008. ACM, 2013.

[5] B. Collins, J. Deng, K. Li, and L. Fei-Fei. Towards scalable dataset construction: An active learning approach. In *Computer Vision–ECCV 2008*, pages 86–98. Springer, 2008.

[6] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris? *ACM Transactions on Graphics (TOG)*, 31(4):101, 2012.

[7] S. Dow, A. Kulkarni, S. Klemmer, and B. Hartmann. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1013–1022. ACM, 2012.

[8] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3474–3481. IEEE, 2012.

[9] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *CVPR*, 2010.

[10] R. Gomes, P. Welinder, A. Krause, and P. Perona. Crowdclustering. In *Neural Information Processing Systems (NIPS)*, 2011.

[11] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

[12] G. Little, L. B. Chilton, M. Goldman, and R. C. Miller. Turkit: tools for iterative tasks on mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 29–30. ACM, 2009.

[13] G. Little, L. B. Chilton, M. Goldman, and R. C. Miller. Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 68–76. ACM, 2010.

[14] S. Maji and G. Shakhnarovich. Part discovery from partial correspondence. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012.

[15] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 89–96. IEEE, 2011.

[16] D. Parikh and K. Grauman. Interactively Building a Discriminative Vocabulary of Nameable Attributes. In *CVPR*, 2011.

[17] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2751–2758. IEEE, 2012.

[18] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 2010.

[19] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *Computer Vision–ECCV 2012*, pages 73–86. Springer, 2012.

[20] S. Vijayanarasimhan and K. Grauman. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2262–2269. IEEE, 2009.

[21] P. Welinder, S. Branson, P. Perona, and S. J. Belongie. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*, pages 2424–2432, 2010.