# See the reasons why '*I*' fail!: Interactively Discovering Failure Modes

Aayush Bansal

Carnegie Mellon University

## 1 Introduction

Our computer vision community has come a long way in making our vision system work with better features [1–3] for description, better classifiers for learning models [4] and bigger datasets [5–7] for training and testing. Despite all these achievements, we still have a long way to go before actually making our systems reliable for usage. While we work harder to improve our systems, we need some principled approach to understand when and where our systems often fail. Understanding these failure modes can help the community to streamline the focus on figuring out the desired changes required to make the system work. One attempt at this is the work of Hoiem et.al. [8] where they analyzed the impact of different object characteristics such as size, aspect ratio, occlusion, etc. on object detection performance. In this work, we propose a method such that machines/systems can interactively present their problems to a human supervisor like a student-advisor discussion. The most related to ours is the recent work of [9] where they find 'specification sheets' to describe failure modes. In this work, we aim at finding the vocabulary of attributes which is meaningful for characterizing failures, and is expected to be task- and system-dependent. The approach of [10] is also particularly relevant to our work as they also focus on task-based attribute discovery.

**Overview:** We propose an approach that automatically identifies patterns in failures, which is then presented to a human-in-loop to name it. Given a trained classification system and a labeled set of training images, we identify images that are correctly classified (not-mistake images), and those that are misclassified (mistake images). The mistake images are discriminatively clustered. In our work, we assume that for a particular 'cluster' to qualify for nameability, it should have three characteristics: 1. **Uniqueness:** It should have some attributes which are representative of that particular 'mistake' cluster; 2. **Discriminative:** Those attributes should be different from remaining 'mistake' clusters and 'not-mistake' images; and 3. **Sufficiency:** There should be atleast 5 examples in each cluster.

In the remaining sections of this paper, we describe our clustering approach and show some initial qualitative results obtained from this approach for tasks such as recognizing faces [11], and animals [12]. Finally there are some initial quantitative results for failure prediction.

## 2 Approach

While our approach can be applied to any vision system, we use image classification as a case study in this paper. We are given a set of $N$ images along with their corresponding class labels $\{(\boldsymbol{x}_i, y'_i)\}, i \in \{1, \ldots, N\}, y' \in \{1, \ldots, C\}$, where $C$ is the number of classes. We are also given a pre-trained classification system $H(\boldsymbol{x})$. We wish to discover clusters of mistake images i.e. failure modes.

**Discriminative Clustering:** We use a slight variant of clustering approaches used in [9, 13]. In our approach, each image is a data point and is represented by the scores of the classifiers trained to predict attributes. We use the attribute scores provided by [11, 12]. The predicted attributes is used as another feature space. We use this as the feature space because that's what [9] uses, but other features can also be used instead. The mistake images are partitioned into two parts $D1$ and $D2$. The images in $D1$ are initially clustered using k-means algorithm. The clusters which have less than 5 examples are pruned. For each remaining cluster, a discriminative classifier (RBF SVM) is trained using the samples of cluster as positive examples while remaining data in $D1$ and not-mistake images as negative examples. This step ensures 'discriminative' part. These trained classifiers are then used to classify data in $D2$. The detection scores are used to cluster the data points in $D2$. The data point is assigned to a particular cluster for which it has maximum detection score. And then classifiers for that cluster are re-trained. As a proxy of purity measure, we compute the average detection score of samples in each cluster. The clusters having average detection score less than 0.25 are pruned. This step ensures 'uniqueness' part. The clusters having less than 5 examples are pruned. $D1$ and $D2$ are swapped. This process is repeated until convergence. See Figure 1 for qualitative results.

## 3 Experiments

**Datasets:** We experiment with two domains: face (celebrity) and animal species recognition. For faces, 2400 images from 60 categories (40 images per category) from the development set of the Public Figures Face Database (Pubfig) of Kumar *et al.* [11] are used. For animals, 1887 images from 37 categories (51 images per category) from the Animals with Attributes dataset (AwA) of Lampert *et al.* [12] containing 85 (annotated) attributes are used.

**Failure Prediction:** Our approach is separating mistakes from not-mistakes, and hence has the potential to be used as a classifier confidence measure of sorts, to automatically predict oncoming failures. To this end, we use the following approach. We run an image through each of our S clusters. Recall that each cluster is formed by a RBF SVM – one for each cluster each of which produces a probability of the image being a mistake. We build a feature vector for an image by concatenating these output probabilities along with the entropy of the main classifier whose mistakes we are characterizing. We train an SVM on this new representation to classify mistake images from not-mistake images. We have S such classifiers, one for each specification sheet. We average their responses on a test image to estimate the likelihood of that image being a mistake.

(a) More than one animal or things

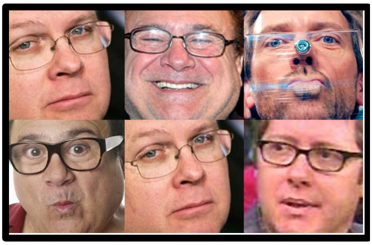
(b) All animals in picture 'appear' to be 'bulbous'

(c) 'small' animals

(d) 'blue' background

(e) smiling young females

(f) hair falling forward OR covered head



(g) (apparently) wearing spectacles

(h) 'yellowish' lighting on images

**Fig. 1.** (a)-(d) are example clusters from Animals with Attribute dataset [12]. (e)-(h) are example clusters from Pubfig dataset [11]. Currently, we have assigned the 'names' to each cluster.

**Metric:** We evaluate the ability of our specification sheets to predict failure using precision and recall (PR), where we evaluate how often an image predicted by the specification sheet to be a failure truly is a failure (precision), and what percentage of the true failures are detected by the specification sheet (recall).

**Baselines:** We compare our automatic failure prediction approach to other non-semantic baselines. **ClassConf (CC):** The conventional approach to estimating the confidence of a classifier is computing the entropy of the probabilistic output of the classifier across the class labels (*e.g.* computed using Platts' method [14]) to a given test instance. **Boost:** Our approach to automatic failure prediction employs multiple classifiers. This is related to boosting approaches [15]. We use Adaboost [16] to learn the weights of 2000 decision trees to separate mistake and not-mistake images. **Random:** We also compare to a baseline that assigns each image a random score between [0,1] as a likelihood of failure. As seen in Table 1 and Figure 2, our approach outperforms these baselines.

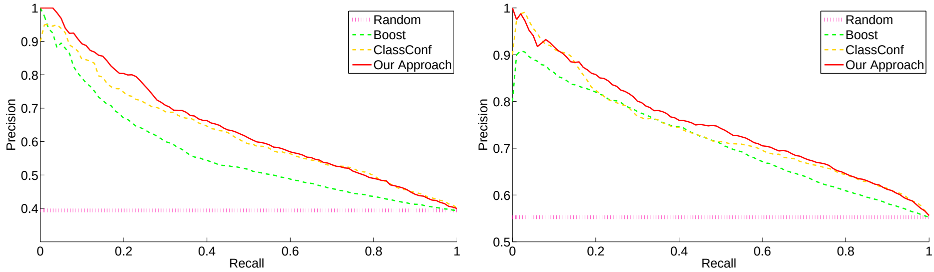



**Fig. 2.** Performance of our specification sheets automatically predicting oncoming failure. Left: AwA, Right: Pubfig.

| | Random | Boost | ClassConf | Our Approach |
|---|---|---|---|---|
| AwA | 0.3939 | 0.5598 | 0.6259 | 0.6435 |
| Pubfig | 0.5527 | 0.7142 | 0.7375 | 0.7496 |

**Table 1.** Area under the precision recall (PR) curve (left) for different approaches.

## 4 Discussion & Future Work

Our initial experiments show that our approach yields somewhat convincing results both qualitatively and quantitatively. Currently, we ourselves 'named' the clusters. In the future, we will use Amazon Mechanical Turk for naming the clusters. We will use our approach to understand the failure modes of state-of-the-art image classification system. This would help give better insights to the community and would be helpful in demonstrating how this approach could be used. Further, we will study the influence of different steps of clustering approach in our work. As a part of later work, we aim to use these 'failure' examples to create better classifiers like the work of [17].

## References


1. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005) 886–893
3. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
4. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(9) (2010) 1627–1645
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR. (2009)
6. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html
7. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR, IEEE (2010)
8. Hoiem, D., Chodpathumwan, Y., Dai, Q.: Diagnosing error in object detectors. In: ECCV. (2012)
9. Bansal, A., Farhadi, A., Parikh, D.: Towards transparent systems: Semantic characterization of failure modes. In: ECCV. (2014)
10. Parikh, D., Grauman, K.: Interactively building a discriminative vocabulary of nameable attributes. In: CVPR. (2011)
11. Kumar, N., Berg, A., Belhumeur, P., Nayar, S.: Attribute and simile classifiers for face verification. In: ICCV. (2009)
12. Lampert, C., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR. (2009)
13. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: European Conference on Computer Vision. (2012)
14. Platt, J.: Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Advances in Large Margin Classiers. (2000)
15. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Machine Learning International Workshop. (1996)
16. Appel, R., Fuchs, T., Dollár, P., Perona, P.: Quickly boosting decision trees - pruning underachieving features early. In: ICML. (2013)
17. Parkash, A., Parikh, D.: Attributes for classifier feedback. In: ECCV. (2012)