Supplementary Material for
Towards Transparent Systems: Semantic Characterization of Failure Modes

Aayush Bansal
Carnegie Mellon University
aayushb@cs.cmu.edu

Ali Farhadi
University of Washington
ali@cs.uw.edu

Devi Parikh
Virginia Tech
parikh@vt.edu

In this document, we give a more detailed explanation of the accuracy (ACC) vs. frequency-of-use (FOU) metric introduced in the main paper. We show the ACC vs. FOU curves for a human user[1] as well as automatic failure prediction using our specification sheets. The area under these curves were reported in Tables 2 and 3 in the main paper (and are replicated here in Tables 1 and 2).

Recall that one use case for our proposed specification sheets is the following: a user of a system can use the specification sheet to decide when to trust the system, and when not to trust it. If an input image falls in one of the scenarios listed in the specification sheet, the user would not trust the system. In other words, the user would not use the system for that input image, because according to the specification sheet, the system is like to fail on this input image anyway. Clearly, a specification sheet that lists many scenarios on it will capture many of the failure modes and perhaps (incorrectly capture) several scenarios that are not failures. As a result, the user will not be able to use the system very often (i.e. low frequency-of-use) which may be annoying or counter-productive for the user. But whenever he does use the system, it will likely be very accurate. This trade-off between the accuracy of the system (ACC) and frequency-of-use (FOU) is thus very relevant to the user – perhaps more so than the traditional precision-recall trade-off (results for the latter are shown in the paper, and may be more relevant for the use case where researchers are using our specification sheets to better understand their systems).

If the input pre-trained classification system whose mistakes we are characterizing has a classification accuracy of 70%, a perfect specification sheet would ensure that all the mistakes (30% of the images) are detected by it, while the remaining 70% of the images are left un-flagged. Hence, a user using the specification sheet would ignore the vision system's response (i.e. not use it) 30% of the time. Therefore, anywhere from 0 to 0.7 FOU, the vision system would have 100% accuracy. If the user chooses to work with a specification sheet that flags only 15% of the images (i.e. user uses the system 85% of the time), the accuracy of the vision system would be lower, somewhere between 70% and 100% at 0.85 FOU. If the user insists on ignoring the specification sheets and uses the system 100% of the time, the accuracy of the system will naturally be 70% (at FOU = 1.0). To capture this trade-off between accuracy of the system when used (ACC) and frequency-of-use (FOU), we plot ACC vs. FOU curves as shown in Figures 1 and 2.

[1]For exhaustive quantitative evaluation, the results we report here use simulated users who can identify the presence/absence of attributes correctly, and hence will not make a mistake while following the specification sheet. This allows us to evaluate the quality of the specification sheets themselves while avoiding confounding factors such as human error. Note that this does not result in a (even nearly) perfect failure prediction system. This is because the scenarios listed in the specification sheet are *learnt* summaries of the attributes incorrectly classified images tend to share in common. See Section 4.6 in main paper for real user studies on Amazon Mechanical Turk.

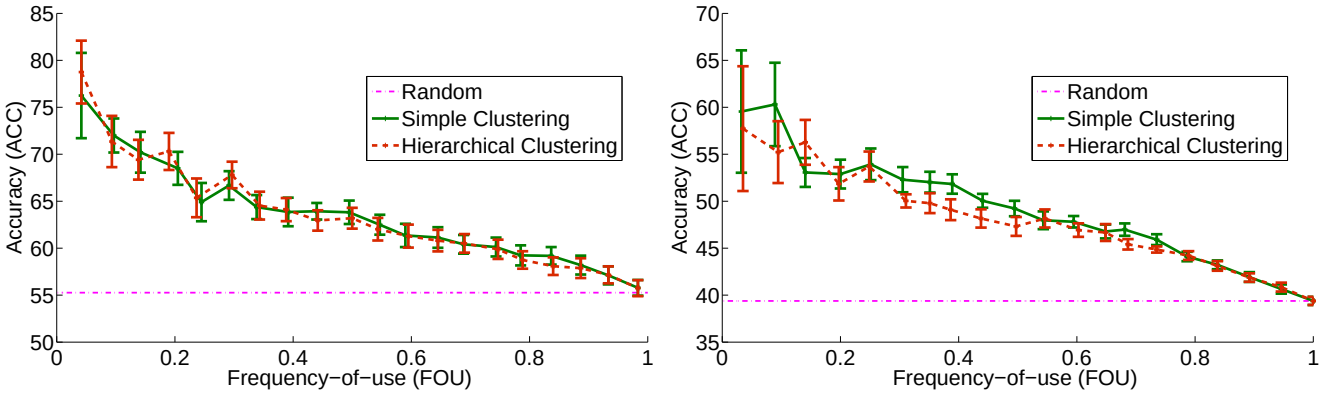

Figure 1: Human users using our specification sheets. Left: Pubfig, Right: AwA. See Table 1 for area under these curves (also provided in Table 2 in main paper).

	Random	SC - all	SC - sel	HC - all	HC - sel
Pubfig	0.5517	0.6181	0.6067	0.6157	0.5997
AwA	0.3929	0.4777	0.4734	0.4636	0.4606

Table 1: Area under the accuracy vs. frequency-of-use (ACC vs. FOU) curves for human users using specification sheets generated using different approaches. SC: simple clustering, HC: hierarchical clustering, all: using all attributes, sel: using a subset of attributes that are easy for lay people to understand. We show only a subset of these methods in Figure 1 for sake of clarity.

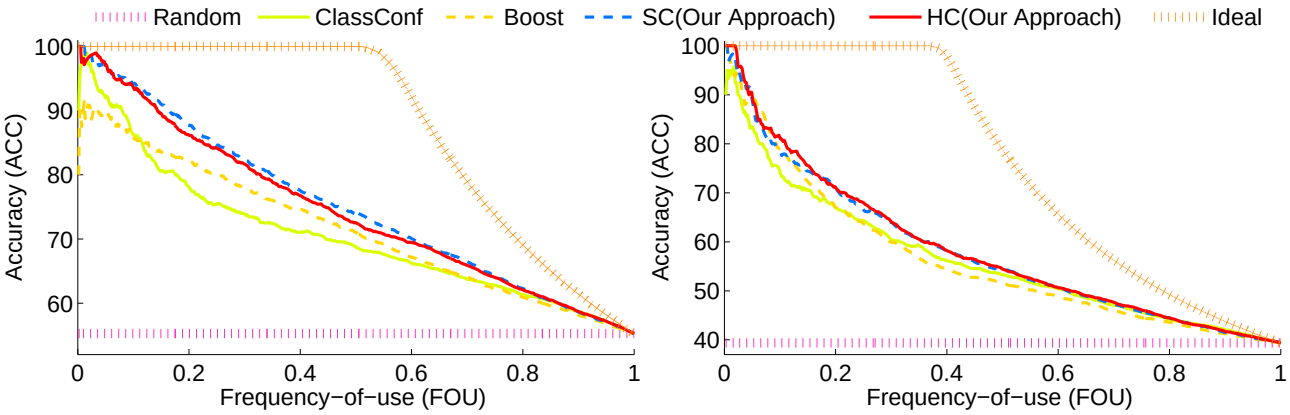

Figure 2: Automatic failure prediction using our specification sheets. Left: Pubfig, Right: AwA. See Table 2 for area under these curves (also provided in Table 3 in main paper). SC: Simple Clustering, HC: Hierarchical Clustering, Boost: Boosting, ClassConf: Confidence of Classifier. See main paper for detailed descriptions of the baselines.

	CC	Boost	SC	HC	CC+HC	Boost+CC	Boost+HC	HC+Boost+CC	GC	GC+CC	Rand	Ideal
Pubfig	0.7033	0.7130	0.7423	0.7316	0.7117	0.7390	0.7409	0.7387	0.6430	0.7293	0.5517	0.8782
AwA	0.5594	0.5573	0.5752	0.5789	0.5640	0.5807	0.5821	0.5809	0.5297	0.5600	0.3929	0.7582

Table 2: Area under the ACC vs. FOU curves corresponding to various methods for automatic failure prediction. CC: ClassConf, SC: simple (discriminative) clustering, HC: hierarchical (discriminative) clustering, GC: generative clustering, Boost: Boosting. Only a subset of these methods are shown in Figure 2 for sake of clarity.