Discovering Biases in Image Datasets with the Crowd

Xiao Hu, Haobo Wang, Somesh Dube, Anirudh Vegesana, Kaiwen Yu, Yung-Hsiang Lu, Ming Yin
Purdue University
{hu440, wang2940, dube1, avegesan, yu872, yunglu, mingyin}@purdue.edu

Introduction

Computer vision technologies have been applied to an increasingly wide range of applications from autonomous car navigation, to medical image analysis, to precision agriculture. Despite many of these exciting innovations, recent studies reveal a number of risks in using existing computer vision systems, suggesting results of such systems may be unfair or untrustworthy. For example, major commercial facial analysis tools were shown to have substantial accuracy disparities for people of different gender or with different skin colors (Buolamwini and Gebru 2018). Visual semantic role labeling models were found to exhibit societal biases and stereotypes (Zhao et al. 2017) such as frequently associating certain activity labels with specific gender (e.g., associate "cooking" with woman). Even worse, seemly accurate image classifiers may in fact made the predictions by picking up spurious correlations between objects and irrelevant background information rather than identifying meaningful features of the objects (Ribeiro, Singh, and Guestrin 2016).

Many of the risks embedded in modern computer vision systems can be attributed to the use of a training dataset that is *biased*. Indeed, the computer vision community has long recognized that many visual datasets present varying degrees of build-in bias due to factors such as photographic style of photographers and selection from dataset curators (Torralba, Efros, and others 2011). Using these biased datasets to train machine learning models for addressing different computer vision tasks naturally leads to the phenomenon of "bias in, bias out" and results in undesirable performance. Thus, to mitigate the fairness, accountability, and transparency concerns in computer vision, a crucial step is to start the entire pipeline with high-quality visual datasets that, at least, are authentic representations of the visual world. In other words, being able to detect potential biases hidden in the datasets prior to model development is a key step in guarding against unfair or untrustworthy outcomes in computer vision.

While a few techniques have been developed to automatically detect dataset biases (Tramer et al. 2017), the non-structured nature of visual data makes bias discovery in image datasets particularly challenging. This is because no human-comprehensive attributes can be directly leveraged from the dataset to reason about the statistical associations between different features of the data. Inspired by recent efforts in learning semantic attributes from the crowd (Tian, Chen, and Zhu 2017; Patterson and Hays 2012), we propose a *human-in-the-loop* approach to facilitate bias discovery in image datasets.

More specifically, this paper presents a crowdsourcing workflow for bias detection in image datasets with three steps (Figure 1): (1) inspect random samples of images from the dataset and describe their similarity using a question-answer pair, (2) review separate random samples of images from the dataset and provide answers to questions solicited from the previous step, and (3) judge whether statements of the image dataset that are automatically generated using the questions and answers collected accurately reflect the real world. This workflow is further augmented by back-end text processing techniques to deal with the noisy inputs from the crowd. Our preliminary results suggest that this workflow is promising in uncovering potential biases in visual datasets.

Crowdsourcing Workflow for Bias Discovery

Previous research have shown the success of decomposing complex tasks into small "micro-tasks" and engaging different crowds in working on different subtasks and collectively solving the grand problem (Bernstein et al. 2010; Chilton et al. 2013; Kim et al. 2014). Following this spirit, we decompose the bias detection task into three interconnected steps as described below.

Step 1: Question Generation

In the first step, crowd workers will be presented with n images randomly sampled from the dataset, and different worker may get different samples. Workers are asked to carefully inspect the images and find similarities between them. In an early design of the workflow, we ask workers to name the attributes on which they find similarities. This design, however, leads to inputs from workers that can be hard to interpret. Inspired by recent efforts in collecting visual questions from the crowd (Antol et al. 2015), we redesign the first step and ask workers to describe the similarity using a *question-answer pair*. In this way, we can obtain more contexts of the similarities that crowd workers detect. We further restrict workers to start their questions with "What,"

Presented in the Work in Progress and Demo track, HCOMP 2019.

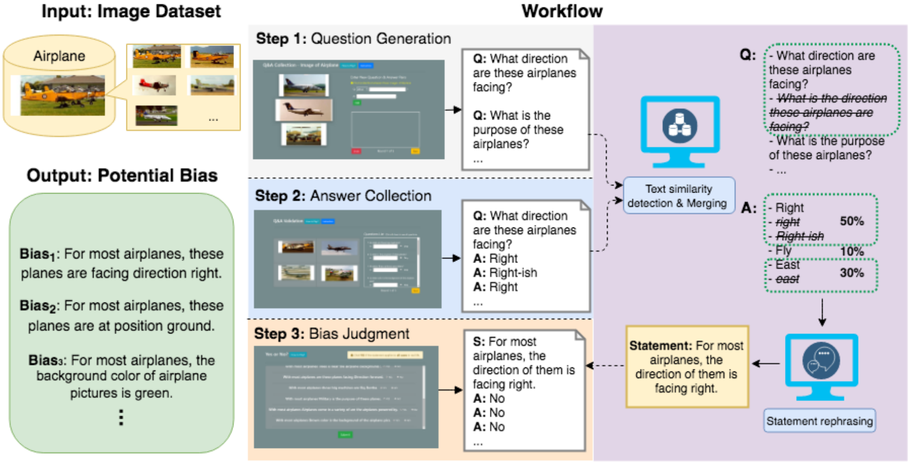
Input: Image Dataset
Airplane
...
Output: Potential Bias
Bias$_1$: For most airplanes, these planes are facing direction right.
Bias$_2$: For most airplanes, these planes are at position ground.
Bias$_3$: For most airplanes, the background color of airplane pictures is green.
⋮
Workflow
Step 1: Question Generation
Q: What direction are these airplanes facing?
Q: What is the purpose of these airplanes?
...
Step 2: Answer Collection
Q: What direction are these airplanes facing?
A: Right
A: Right-ish
A: Right
...
Step 3: Bias Judgment
S: For most airplanes, the direction of them is facing right.
A: No
A: No
A: No
...
Text similarity detection & Merging
Q:
- What direction are these airplanes facing?
- ~~What is the direction these airplanes are facing?~~
- What is the purpose of these airplanes?
- ...
A:
- Right
- ~~right~~
- ~~Right-ish~~
50%
- Fly
10%
- East
- ~~east~~
30%
Statement: For most airplanes, the direction of them is facing right.
Statement rephrasing

Figure 1: A crowdsourcing workflow to discover potential biases presented in image datasets.

"Where," "When," or "How." For example, if workers reviewing images of airplanes find all the airplanes shown to them point to the right-hand side of the image, they can create a question like "What direction are these airplanes facing?" and answer "Right." Workers are told that they are free to find similarities for any part of the images, including main objects and background. In addition, workers are instructed to not ask questions regarding the name or common characteristics of the main objects (e.g., "How many wheels does a car have?"). We encourage workers to generate as many unique questions as they can to describe the similarities among images shown to them.

Post-processing: Crowd workers may easily describe the same kind of similarity using questions with different wordings. To reduce redundancy, we use spaCy, an open source natural language processing tool, to detect similar questions produced by workers and merge them. Specifically, two questions will be combined if their similarity score is above a threshold, and the question with "higher quality"—quantified by having more noun phrases and dependent clauses—will be used to represent this group of questions.

Step 2: Answer Collection

The output of Step 1 is a list of candidate biased attributes represented as questions. As similarities identified among n randomly-sampled images may only capture "biases" within a particular sample, further validation is needed to verify whether such pattern exists outside the specific sample. Thus, in the second step, we use questions generated in the first step as inputs and we collect answers to each of these questions based on separate visual data samples.

In particular, crowd workers will be presented with m images that are, again, randomly sampled from the dataset, along with the list of unique questions produced in Step 1. Workers are asked to carefully review the images and then answer all of the questions. If the majority of m images share the same answer to a question, workers are asked to enter that answer; otherwise, they can click on a button to "skip" the question. We ask workers to answer each question with a simple word or phrase when possible.

Post-processing: Similar as that in Step 1, for each question, we again use spaCy to identify similar answers and merge them. Then, we will compute the "weight" of each unique answer to a question by counting the fraction of workers who provide that answer. Given a question, if the majority of workers choose to skip it, we consider this question as not characterizing actual bias. However, if the weight for the most popular answer to a question is above a threshold, we will use a customized algorithm to rephrase the combination of that question and the answer with highest weight into a statement of the dataset (e.g., "For most airplanes, the direction of them is facing right."). The threshold can be tuned to reflect the degree of biases that we are targeted at.

Step 3: Bias Judgment

Step 3 takes the set of statements about the image dataset as inputs. Crowd workers are asked to review each statement and indicate whether they believe that statement accurately reflects the real world based on their common sense knowledge and subjective belief. The output of this step is a ranked list of statements, sorted in decreasing order of the fraction of workers who indicate the statement does *not* accurately reflect the real world and thus can potentially be a bias of the dataset. This list can then be returned to dataset curators for further investigation.

Preliminary Results

As a proof of concept, we conducted a preliminary experiment with a set of 120 airplane images taken from Caltech 101 (Fei-Fei, Fergus, and Perona 2007) to examine the effectiveness of the proposed workflow. We intentionally injected two biases into this dataset—all planes in the images are facing right, and 80% of images have brownish/greenish background. Workers were recruited from Amazon Mechanical Turk to find biases in this image dataset following our workflow. We set $n = 3$ and $m = 4$ in our experiment. We find that MTurk workers are able to "recover" the injected biases through our workflow, and they also discover some additional biases of this dataset that are not due to our design, like the planes are mostly sitting on the ground (83% of the images in the dataset actually have planes on the ground).

References

[Antol et al. 2015] Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

[Bernstein et al. 2010] Bernstein, M. S.; Little, G.; Miller, R. C.; Hartmann, B.; Ackerman, M. S.; Karger, D. R.; Crowell, D.; and Panovich, K. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 313–322. ACM.

[Buolamwini and Gebru 2018] Buolamwini, J., and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91.

[Chilton et al. 2013] Chilton, L. B.; Little, G.; Edge, D.; Weld, D. S.; and Landay, J. A. 2013. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1999–2008. ACM.

[Fei-Fei, Fergus, and Perona 2007] Fei-Fei, L.; Fergus, R.; and Perona, P. 2007. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer vision and Image understanding* 106(1):59–70.

[Kim et al. 2014] Kim, J.; Nguyen, P. T.; Weir, S.; Guo, P. J.; Miller, R. C.; and Gajos, K. Z. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 4017–4026. ACM.

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

[Ribeiro, Singh, and Guestrin 2016] Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144. ACM.

[Tian, Chen, and Zhu 2017] Tian, T.; Chen, N.; and Zhu, J. 2017. Learning attributes from the crowdsourced relative labels. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[Torralba, Efros, and others 2011] Torralba, A.; Efros, A. A.; et al. 2011. Unbiased look at dataset bias. In *CVPR*, volume 1, 7. Citeseer.

[Tramer et al. 2017] Tramer, F.; Atlidakis, V.; Geambasu, R.; Hsu, D.; Hubaux, J.-P.; Humbert, M.; Juels, A.; and Lin, H. 2017. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, 401–416. IEEE.

[Zhao et al. 2017] Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.