MindSet: A human-centric approach to debias your data

Ziyao Shang Alexander Kichutkin Senthuran Kalananthan András Strausz

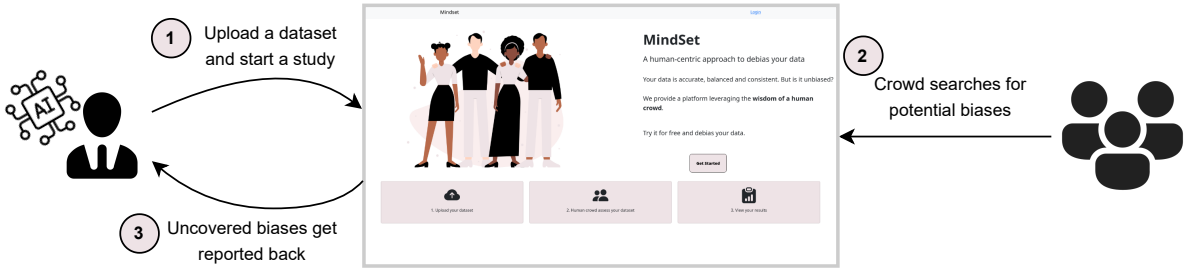


Figure 1: General concept of the MindSet Survey workflow based on the work of Hu et al. [5]

Abstract

The rise of deep learning models in automated decision making has raised great concerns about bias and fairness. A main cause of biased models lies in asymmetries in the training datasets, for instance, due to reporting or selection biases. Discovering such artifacts, however, is especially difficult as for many data types there are no common quantitative measures of bias. Therefore, a common approach is to rely on human feedback. This is based on the assumption that a sufficiently large number of human judgements will correctly detect such anomalies. In a former paper, Hu et al. [5] introduce a three stage study framework to use human feedback to detect biases in visual datasets and ran a small-sample study with encouraging results. Unfortunately, there is no public, dataset-agnostic implementation available that would enable researchers to use this method to clean their self-collected, or even synthetically generated image datasets. To this end, we develop a highly flexible study interface that follows the framework proposed by Hu et al. [5] and present a demo study to highlight the most import features.

Index Terms: H.5.2 [User Interfaces]: ,— [H.3.3]: Dataset Bias—, I.2.6 [Bias in Machine Learning]: —

1 Introduction

Computer vision, particularly visual pattern recognition, has been dominated by deep learning techniques since the appearance of the CNN and, later, the transformer architectures. In recent years we have seen huge improvements in all the different visual tasks, such as image segmentation [8], classification [6], or most recently in image generation [9]. Moreover, these techniques have long left academia and have been deployed in various real-life scenarios, often involving such where an ethical and fair decision is indispensable [3, 7].

A common challenge with deep learning techniques stems from their heavy reliance on the training data [12]. Biases included in the training data may not only be reflected but even enlarged by the models prediction, disallowing their use in any real-world scenario. There are commonly three stages where model bias can be mitigated:

1. *pre-processing*: refactoring, cleaning the training data such that it contains possibly no bias.
2. *training*: most commonly, this is achieved by adversarial training methods where the model is also updated to prevent an adversarial model from recovering a protected concept.
3. *post-processing*: the model output, or the learned representation, is post-processed after training, usually through projection-based methods

Developing fair and trustworthy models in the visual domain is notoriously difficult, however, as there exists no precise quantitative formulation of bias that could be controlled for during training. As a result, either human qualitative judgments are taken into account, or proxy measures are used aiming to grasp some parts of the contained biases.

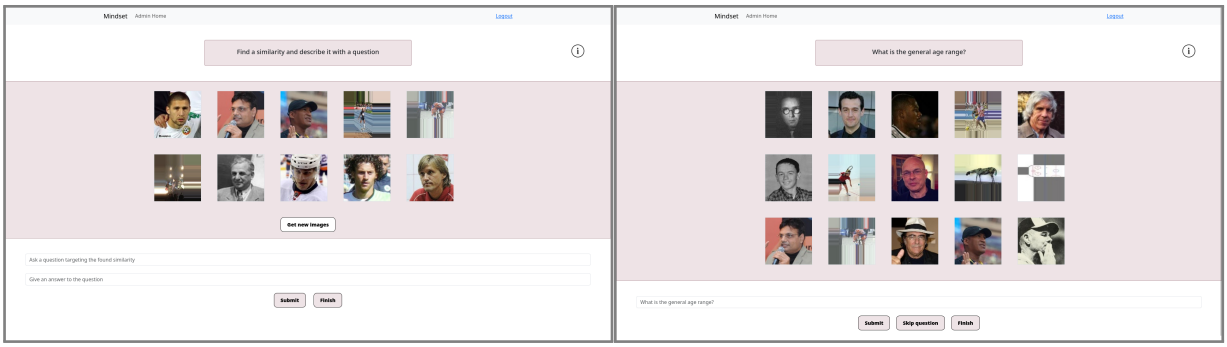
While the latter may be useful to give specific, often technical, fairness guarantees (such as no MLP can recover the protected concept from the learned representation), it falls short of ensuring a generally fair model. On the other hand, humans can naturally detect visual biases, and their assessment can be later included in the Machine Learning pipeline. While humans themselves have prejudices, and a single personal judgment may itself not be representative, this may be counteracted by a larger sample of human opinions. In fact, according to the *wisdom of the crowd* hypotheses [10], if enough diverse opinions are gathered, the aggregation of this sample may lie closer to the ground truth. It is worth noting, that the diversity of the sample is an integral part of this concept.

The most common way to include human judgments in bias mitigation is to filter the training data based on human inputs. Hu et al. [5] propose a three-stage study technique to detect sample biases in visual datasets trough gathering human inputs. In their evaluation, they show that the framework allows finding both commonly known as well as dataset-specific, yet unknown biases among images. However, the authors did not develop any implementation for their study but used static forms that were tailored for a single dataset and was not made available to the public.

In this work, we develop an interactive web interface that follows the framework of Hu et al. [5], but allows the use of any datasets uploaded by the study administrator and is accessible to anybody. We hope that through this work, we could greatly ease the detection of biases in image datasets, and thus contribute to the development of fair and trustworthy visual ML models.

2 *Wisdom of the Crowd* bias detection

In this section, we summarize the study procedure proposed by Hu et al. [5], as this serves as the base of our interface. We only describe the main stages of the study here and defer any specifics or changes to it to Sect. 3. For the detailed arguments for the different steps of the study, we refer the reader to the original paper.



(a) Stage 1 with 10 images shown.

(b) Stage 2 with 15 images shown.

(c) Stage 3

Figure 2: The main interface of the study. Participants are guided through the interface at the start and further aided with hints.

2.1 Question generation

The study starts by asking the participants to enter question-answer pairs that describe a similarity among the set of images that are currently shown. Participants are encouraged to ask questions starting with *What, Where, When* or *How*, and to avoid questions describing common characteristics of objects (e.g., *How many wings does an airplane have?*). These questions are then merged to keep only questions targeting different biases. Answers of this stage are not used anymore, they are only to encourage the participant to ask more human-like questions.

2.2 Answer collection

In the second phase, the collected questions are shown to the users again but with a possibly different sample of images. The user is then asked to enter an answer to the question if at least half of the images share the answer, otherwise skip it. Afterwards, answers are similarly merged to avoid ambiguities from different spellings or synonyms.

2.3 Bias Judgement

Lastly, in Hu et al., questions and their corresponding answers are used to generate universal statements describing a possible bias. Users are then asked whether this statement is true in the real world or is a specific attribute of the dataset. In order not to influence the user's judgment, there are no images shown in this round.

3 The MindSet interface

To better support bias mitigation in visual datasets, we implement a user-friendly interface for the study framework of Hu et al. [5]. The implementation is accessible with a demo study under [1].

[1] `http://a10-bias-assessment-with-human-feedback.course-xai-iml23.isginf.ch/`

3.1 Implementation details

MindSet closely follows the study framework described in [5], and aims to offer flexibility in choosing some of the study parameters.

3.1.1 Technical details

The interface's front-end is implemented in `TypeScript` and uses the `D3` library for data visualizations. The website interfaces are implemented using the React framework. The CSS `Bootstrap` framework is used for webpage styling. The backend is developed using `Flask Restful` and supported by an `SQLite` database. For processing the text inputs we currently use the `all-MiniLM-L12-v2` pre-trained sentence transformer language model, but it may be changed to improve sentence embeddings in the case when server resources allow. Frontend-backend interactions are implemented using `Axios`, and the communications between them are structured according to the RESTful architecture. Additional storage spaces in Azure are used in the deployed version to save the image datasets.

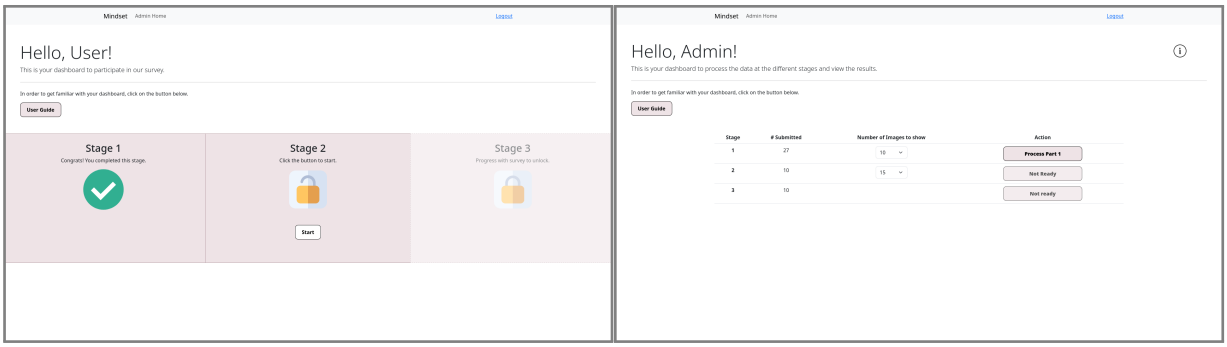
3.1.2 User types

We differentiate between two types of users, *participants* and *study admins*. In order to be able to delete spurious inputs, we ask users to first register with their email accounts. Registrations of participants can be verified by the participant itself trough a code sent by email. Registration of study admins have to be accepted by the developers.

Participants can only access the current stage of the study and should be notified when the study progresses to the next stage. *Study admins* are also *participants* as well; however, they can choose to move their study from one stage to another and see overview statistics as well as detected biases. They are also responsible for setting the number of images the participants see during each step.

3.1.3 Study workflow

The interface for the different steps is depicted in Figure 2. We aim to create a neutral interface with as few text as possible to avoid



(a) User overview at Stage 2.

(b) Admin overview at the beginning of the study.

Figure 3: The overview pages for users and study admins.

influencing the participant. Participants receive a short introduction to every state and are guided through the interface before starting the study. A hint is also available in case the participant loses track. A difference to the original framework is that we do not provide any suggestions at no stage of the study, in order to avoid influencing the participant.

3.1.4 Administrator view

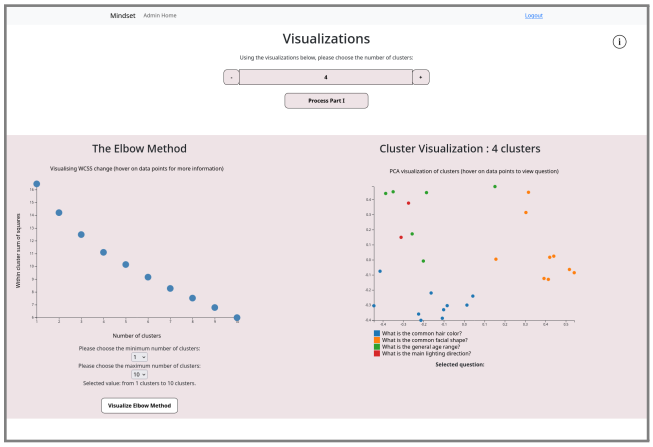


Figure 4: Interface for processing inputs after Step 1

We create a separate overview for study administrators where they can manage their current ongoing study.

One of the main tasks of study administrators is to proceed with the study from one stage to the next one. To process Step 1 (question generation), the administrator needs to extract certain representative questions from existing questions. The extraction is done via clustering. Each question is first embedded into a vector by an NLP model (see Section 3.1.1). Then, the embeddings are clustered using K-means clustering. For each resulting cluster, the question whose embedding is closest to the cluster centroid is chosen to represent that cluster. Thus, the administrator has to choose the number of questions they want to keep. This decision is aided by an interactive visualization of the elbow method [11] and the clustering for the currently chosen setting.

For the elbow method, the administrator specifies a range of centroid numbers. The visualization would be able to plot the Within Clustering Sum of Squares (WCSS) of all centroid numbers within that range. Generally, the elbow point of this graph would be a sound choice for the number of clusters. The administrator can hover over the data points on the visualization to see their actual WCSS values. Next, if the administrator clicks on one of the data points, a preview of the clustering results will be presented.

To create the preview, the embedding for each question is reduced to a 2D vector using Principal Component Analysis [4]. On the preview, each question would be a data point, where the 2D vector will be used as its x,y coordinates. The cluster to which each point belongs would be encoded using the color (hue) of the points. Hovering on the points, the administrator would be able to see the actual question behind the point. The preview also contains a legend containing the questions chosen for each centroid. An example state of the dashboard for processing Step 1 is shown in Figure 4

Once the study administrator decides to finish Stage 2, the answeres given for the questions are processed. For each question, all its answers are embedded, and the answer that is closest to the centroid is chosen. At this step, our implementation also deviates from [5], as they decide to construct statements from the question-answer pairs. While such simple statements may be easier to understand at Stage 3, as statements must be generated with language models, they may not just be slightly imprecise, but a bias could be introduced originating from the language model. To avoid this, we directly present the question-answer pairs to the participants.

During the final stage, the administrator can see an overview (see Figure 5) of the biases detected in the dataset. This overview contains a list view containing the number of agreements, agree ratio, and the text of the biases, where the administrator can label biases, filter biases based on the labels, and save labels into the database. The users are also able to download the table as a .csv file.

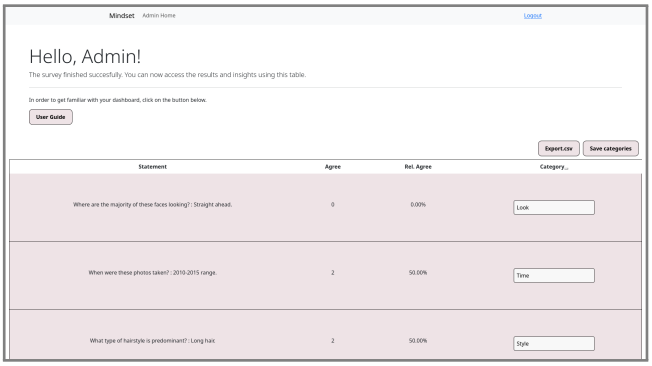


Figure 5: List view of results

3.1.5 Demo Study

To showcase the workflow, we used the CelebFaces Attributes Dataset (CelebA), a comprehensive collection comprising more than 200,000 celebrity images [2]. Additionally, for demonstration purposes, we have prefilled the database with dummy data, allowing for a comprehensive illustration of the system's functionality.

3.2 Design Choices

3.2.1 Color

Color psychology suggests that colors can have an impact on our moods, emotions, and behaviors. However, it's worth noting that individual experiences and cultural influences play a significant role in shaping color associations. Pink is commonly associated with symbols of peace, innocence, and warmth. Different individuals may associate pink with varying emotions and concepts, such as joy, happiness, creativity, and artistry [1]. Our choice of this color scheme was motivated by our intention to create a welcoming and pleasant user experience.

3.2.2 Illustrations & Storytelling

Illustration plays a vital role in visualization as it enhances understanding, engagement, and communication. Within our framework, illustrations serve multiple purposes, including providing context and guiding users through the various stages of the study using visual narratives. Their presence aims to create a more engaging and pleasant user experience. We have carefully selected illustrations that align with our color scheme and depict different scenarios. For instance, on the welcome home page, we incorporated a picture showcasing multiple individuals, symbolizing our intention to leverage the collective wisdom of the crowd. Additionally, on the user home page, we utilized lock icons that progressively unlock as users make progress through the study, encouraging them to complete the entire process. These thoughtful illustrations contribute to an interactive and motivating user journey.

3.2.3 Human-centric Approach

Taking a human-centric approach, our aim was to provide participants with comprehensive information to understand the various steps and processes involved. To achieve this, we have implemented information buttons at each stage, allowing participants to access additional explanations and enabling them to revisit instructions during the survey. Moreover, we have incorporated information buttons for administrators, ensuring they comprehend the underlying processing mechanism. Additionally, user guides have been incorporated to offer a brief introduction to the functionalities of their respective dashboards. The key idea was to keep the user interface clean, while providing sufficient information and minimizing visual clutter. Thus, centering our focus around the user.

Transparency was a priority, as we made certain not to conceal any information. We thus ensured that administrators are aware of whether the numbers presented are absolute or relative and that we utilized PCA to visualize the data points in the administrators' view. We also prioritize providing the study administrator with enough freedom. This includes granting the administrator the ability to determine the number of images to display and the flexibility to choose the desired number of clusters. Our aim is to empower the administrator by allowing them to make these decisions according to their preferences.

3.2.4 Visualization

Visualizing high-dimensional text data poses significant challenges. In order to extract meaningful insights, we adopted a processing technique to represent the questions as vectors. To visualize the resulting clusters, we employed Principal Component Analysis (PCA) to extract essential features and created a straightforward 2D scatterplot.

4 Summary

The fast advancements in computer vision technology have granted its high popularity but also increased the stake of the fairness and ethics of automated visual decision making implementations in real life. The first stage of computer vision pipelines–input training data–is particularly vulnerable to association biases. Inspired by Hu et al. [5]'s three-stage pipeline for obtaining biases in images, we created *MindSet*, a human-centric, web-based, interactive framework for stakeholders to identify possible biases in image datasets through crowdsourcing. We implemented Hu et al. [5]'s pipeline with an emphasis on transparency and clarity. By including interactions and visualizations, *MindSet* strives to narrow the Gulf of Evaluation for all types of users while avoiding visual clutter and engaging the users throughout the pipeline. *MindSet* could be used in various situations and has a large potential for adaptations and extensions.

5 Outline

One present constraint of our framework is the absence of real-world testing. It would be intriguing to explore the biases that may emerge when conducting surveys with a live crowd. Additionally, there are certain missing functionalities, such as the ability to upload datasets to the server, which are crucial for real-world implementation. We believe that additional visualization alternatives (i.e., an interactive dashboard) for the final summary table would also be beneficial.

A good adaptation of our interface would be converting it into a deductive interaction workflow for evaluating synthetic image generation pipelines. Our interface could be used to compare real-world and synthetic data and probe whether the data generation repeats/amplifies/mitigates real-life biases.

Another possible future extension of our framework would be to enable the refinement of a new dataset. Given the detected biases, it would be very convenient if the user could refine the dataset at the same location, preferably through an interactive workplace with certain implemented data refinement methods.

References

[1] The color psychology of pink. `https://www.verywellmind.com/the-color-psychology-of-pink-2795819`.

[2] Large-scale celebfaces attributes (celeba) dataset. `https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`.

[3] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. T. Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128, 2019.

[4] K. P. F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720

[5] X. Hu, H. Wang, A. Vegesana, S. Dube, K. Yu, G. Kao, S.-H. Chen, Y.-H. Lu, G. K. Thiruvathukal, and M. Yin. Crowdsourcing detection of sampling biases in image datasets. In *Proceedings of The Web Conference 2020*, WWW '20, p. 2955–2961. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3366423.3380063

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[7] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021. doi: 10.1145/3457607

[8] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021.

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[10] J. Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.

[11] M. Syakur, B. Khotimah, E. Rochman, and B. D. Satoto. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP conference series: materials science and engineering*, vol. 336, p. 012017. IOP Publishing, 2018.

[12] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars. A deeper look at dataset bias, 2015.