Datasheet for Dataset Associated with "Before and After Pseudoscience: An Empirical Challenge to Social Classification using Facial Recognition"

I. Motivation For Datasheet Creation

A. Why was the datasheet created?

The dataset was created to be used as a set of control examples for a study which claims to have shown that you can tell a person's political orientation using facial recognition tools. The intention was to test whether superficial presentation details are what drives the predictions in that study. To this end, the dataset contains pairs of pictures of the same person presenting differently, before and after makeovers, haircuts, and other superficial changes in appearance.

B. Has the dataset been used already?

This is the first study to use this dataset.

C. What (other) tasks could the dataset be used for?

This dataset could be used as a control for other facial recognition tasks, where some sense is needed of how reliable a prediction is across different views of the same face.

D. Who created the dataset, and who funded the creation of the dataset?

The dataset was created by members of the Ethics and Technology Lab at Queen's University, in Kingston, Ontario, Canada. The students who created the dataset were funded by 1) a Queen's University graduate stipend, 2) a graduate stipend from the CREATE Cybersecurity program at Queen's University, and 3) a SSHRC Institutional Grant awarded by Queen's University to the PI. None of these funding agencies were involved in the decision to create the dataset, nor place any restrictions on the type of research performed.

II. Datasheet Composition

A. What are the instances?

The instances in the original dataset are images of faces, however, what we have made available are trained features based on these images, in order to preserve the privacy of the people pictured.

B. How many instances are there in total (of each type, if appropriate)?

There are 2584 instances in total, with 1292 representing the before images in the pair and 1292 representing the after images in the pair. There are two sets of features provided to represent the dataset, the first is extracted using ResNet-50 trained on VGGFace2, and the second is extracted using ResNet-18 trained on ImageNet. The last layer of both models (i.e. classification layer) are discarded before the extraction process.

C. What data does each instance consist of?

The first set of features are extracted using a ResNet-50 trained on VGGFace2 and each instance consists of a 2048-feature long tensor. The second set of features is extracted using a ResNet-18 trained on ImageNet, and each instance consists of a 512-feature long tensor. The raw images are currently being stored on a secure hard drive that only authors have access to and will be deleted within a year of publication of the study.

Labels are provided for "before" and "after".

D. Is there a label or target associated with each instance? If so, please provide a description.

Labels are provided for "before" and "after" a certain change in presentation. Subcategories of the change in presentation are recorded but not made public as of yet, as we do not yet have a strong reason to do so.

E. Is any information missing from individual instances?

No.

F. Are relationships between individual instances made explicit

Instances are paired, with one labeled "before" and one "after". The instances are sorted according to their pairing in the before and after files respectively. Hence index n in the "before" file should be paired with index n in the "after" file where n is a number between 0 and 1291.

G. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The entire dataset is included.

H. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

While there are no recommended data splits, it is worth noting that the experiments in the paper used a 80-20 random split between training and test data.

I. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

There may be some cases where the ”before” and ”after” labels are incorrect, as these had to be inferred by a human labeller in some cases.

J. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset is self-contained, and will be available on GitHub indefinitely.//

III. Collection Process

A. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

Images were scraped using Microsoft Bing Image Search API using keywords “ face before after” preceeded by either ”beard”, ”hairdo”, ”makeup”, ”drag”, ”glasses” or ”haircut”. We also redid the search using languages such as japanese, swahili and hindi after the keywords to include more diversity in results. The images were verified by one of the authors who did not perform the scraping, who then ensured each image consisted of before and after faces, and that there was no cosmetic surgery or major age difference. Afterwards the faces were automatically detected and cropped from the images using MTCNN face detector [1]. Manual cropping was done where automatic detection was not possible and a final round of verification was performed.

B. How was the data associated with each instance acquired?

The before and after labels were given according to the following criteria: if the image was already labelled (which was often the case) the given label was used. If not, the image that visually appeared to be before the presentational change was labelled as before. This may include longer hair in the case of haircut, less or no makeup in case of makeovers or drag or longer beard in case of beard. In the case where this distinction was not obvious, the image on the left was labelled as ”before”.

C. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The dataset is in a sense a sample taken from the set of all before and after images on the internet.

D. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

An undergraduate student paid at the standard institutional rate for Research Assistants did the data collection. They were only made aware of the study’s aims after data collection was finished.

E. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data was collected between November 2021 and January 2022.

IV. Data Preprocessing

A. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

The images were passed through a ResNet-50 pre-trained on ImageNet with the final (classification) layer discarded. This results in a 512 length tensor for each image.

B. Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

The raw data will be saved for 1 year after publication on a secure server, but will not be made available publicly to protect the privacy of the people pictured in the images.

C. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

We built a simple GUI tool for labelling images available here: `https://github.com/rohanfaiyazkhan/tkinter-image-labeller-gui`

D. Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?

Yes, the dataset was used to perform experiments in the study “Before and After Pseudoscience: Empirical Challenge to Social Classification using Facial Recognition”.

V. Dataset Distribution

A. How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

The dataset is available through Github as files readable using PyTorch with `torch.load`.

B. When will the dataset be released/first distributed? What license (if any) is it distributed under?

MIT License

C. Are there any copyrights on the data?

Copyright 2022 Rohan Faiyaz Khan, Sam Baranek, Georgia Reed, Catherine Stinson

D. Are there any fees or access/export restrictions?

No.

VI. Dataset Maintenance

A. Who is supporting/hosting/maintaining the dataset?

Github is hosting the dataset, and the first author is maintaining it.

B. Will the dataset be updated? If so, how often and by whom?

There are no plans to update the dataset.

C. How will updates be communicated? (e.g., mailing list, GitHub)

Any updates will be communicated on GitHub.

D. Is there a repository to link to any/all papers/systems that use this dataset?

No. The original code used in the project is available here: `https://github.com/rohanfaiyazkhan/before-and-after-pseudoscience`

E. If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

There is no such mechanism, but parties interested in making extensions may contact the authors.

VII. Legal and Ethical Considerations

A. Were any ethical review processes conducted (e.g., by an institutional review board)? processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

B. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor patient confidentiality, data that includes the content of individuals non-public communications)?

The raw data does, which is why we are only releasing anonymized features to preserve the privacy of those pictured.

C. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No

D. Does the dataset relate to people?

Yes

E. Does the dataset identify any subpopulations (e.g., by age, gender)?

No.

F. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

No.

G. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

No.

H. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Third parties, via web scraping.

I. Were the individuals in question notified about the data collection?

Not directly, however we chose to use before and after pictures because these are typically promotional pictures taken with the intention of being posted publicly.

J. Did the individuals in question consent to the collection and use of their data?

Not directly with the data collectors, however, before and after pictures are generally posted with the expressed consent of the people pictured.

K. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No. As the data is anonymized, and not of a sensitive nature, there is no anticipated risk to the data subjects.

REFERENCES

[1] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, October 2016. arXiv: 1604.02878.