Datasheets for Datasets – Lucy O’Connor AI4M

This template contains a set of questions covering the information that a datasheet for a dataset might contain, as well as a workflow for dataset creators to use when answering these questions.

The questions are grouped into seven sections that roughly match the key stages of the dataset creation, maintenance, and distribution process. By grouping the questions in this way, we encourage dataset creators to reflect on the process of creating, distributing, and maintaining datasets, and even to modify this process in response to that reflection. We recommend that dataset creators read through the questions in all sections prior to any data collection so as to flag potential issues early on, and then provide answers to the questions in each section during the relevant stage of the process.

We emphasize that the questions are intended to be used as a starting point for dataset creators to customize. Not all questions will be applicable to all datasets, and dataset creators will likely need to add, revise, or remove questions to better fit their specific circumstances and needs.

To prompt dataset creators to provide sufficient information, all questions are worded so as to discourage yes/no answers. The questions are not intended to serve as a checklist, and dataset creators must be as transparent and forthcoming as possible for datasheets to be useful to dataset consumers.

Questions

Motivation

- **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to use in training with a pretrained [unspecified] AI model(s). The dataset could be used in alternative cases such as training to produce artificial intelligence-generated content as a source of inspiration for those working in the aerial field as well as for storytelling purposes. Many aerial shots are divided by different labelling such as ‘aerial shots’; ‘bird’s eye view’; and ‘landscape photography’; this dataset holds all those labels within one inclusive dataset.

- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

For self-imposed research, I, as the dataset creator built the dataset. However, this was prelabelled and is created on a secondary research basis through the social media site ‘Pinterest’.

- **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

N/A

- **Any other comments?**

The dataset comprises of images found through ‘boards’ on the social media-based network ‘Pinterest’ meaning that the images in the dataset, of a secondary nature, are obtained from a number of different users to combine the images into one larger dataset from multiple sources on the site.

For more information about these questions and about datasheets for datasets in general, please see T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford. *Datasheets for Datasets.* The latest version of this paper can be found online at https://arxiv.org/abs/1803.09010

Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances (images) are aerial photos. They are not defined by a specific region or place, yet simply capture from the same camera angle – birds eye view to produce an aerial image. The dataset was visually looked upon in the cleaning process to discount any images which had, in my opinion, an overstatement of people or objects that overrides a natural essence of the type of landscape in each image (for example, lots of people on a boat). Some of these images have a more cultural essence than others (a cityscape against a natural display of land), but those which that are represented in the dataset remain visually stimulating.

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

The dataset includes 567 instances. These files are all secondary images.

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset does not contain all possible instances. The dataset was visually assessed in such a way with the aim for it to be used in AI model training. This meant that a key feature identified was if images were too visually similar, the latter(s) would be deleted to avoid any over representation against shapes and colours in these images.

Some images collected had overlays for advertising purposes and watermarks. These were deleted from the dataset to abide to copyright laws. Equally, any duplication meant the latter was removed.

The larger version of the dataset (567 files) is good representation of the 100mb sample. This includes aerial photos, consisting primarily of landscapes, some of which are more abstract in nature than others and are still visually wide ranging.

- **What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

The images are all PNG files meaning that they are in an uncompressed raster image format. This is a lossless data compression format with no copyright limitations, ideal for building a dataset of images.

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

In the dataset there is no set labelling. They all follow a numbered format saved within the dataset. Prior to being held in a dataset they were collected by searching for labels as stated previously.

- **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

There is no specific information missing however the images were reduced in size so may have decreased in image quality depending on the previous size they were downloaded in. Most if not all images are now 3mb each.

- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links**)? If so, please describe how these relationships are made explicit.

These relationships are not explicit. However, in respects they are all visually similar as they use the same, if not similar, camera angle.

- **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.

There are no recommended data splits. I will likely use the first 10% of the dataset as there is no specific ordering of the images.

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

There are no errors, source of noise or redundancies. These, if any, were removed in the cleaning process of the dataset. For example: duplicates.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g.,**

websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained, however due to the nature of the images, they are likely to be found within other datasets also. So, they are not, in those respects, unique to the dataset. It was relied upon to be obtained from an external site. The 'boards' created by users for individual inspiration sources are essentially like other datasets in themselves. The dataset I created obtained images from a number of these boards but should those be deleted from the site, they are not relied upon and the images in my dataset will remain constant. There are no restrictions besides obtaining permission from the copyright owner where necessary for use of the images (i.e. publishing).

- **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)?** If so, please provide a description.

No, therefore n/a.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

No, therefore n/a.

- **Does the dataset relate to people?** If not, you may skip the remaining questions in this section.

No, therefore n/a.

- **Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

N/A.

- **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?** If so, please describe how.

N/A.

- **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)?** If so, please provide a description.

N/A.

- **Any other comments?**

Some images may contain people, however their faces are not seen and are in no way offensive or considered sensitive.

Collection Process

- **How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The image dataset was indirectly derived from other data. The images were found within other boards on the social media-based network Pinterest and were cleaned and refined respectively to create the new dataset. The images only need to be verified for use by seeking copyright permission from the image owner where necessary.

- **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?

PinDown, a browser extension, was used to scrap full-size images from Pinterest. It is a free extension which allows you to download up to 250 images per board.

- **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 not a sample from a larger set, therefore no sampling strategy to provide.

- **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)?**

One individual, a student collected the data for further research purposes.

- **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 images were collected in an instant. I imagine the images themselves were timely to create. Some *may have used drones to capture the photos which needed to take the time of day into consideration as well as weather suitability to capture the photo at the given point in time.*

- **Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No ethical review processes were conducted. If, however, there was; the institution – University of the Arts, London, would have conducted this process. The one main factor in association with ethical considerations is copyright laws to be dealt with where necessary.

- **Does the dataset relate to people?** If not, you may skip the remaining questions in this section.

No, therefore n/a.

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

N/A.

- **Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A.

- **Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A.

- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A.

- **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.

N/A.

- **Any other comments?**

Individual users create the 'boards' on Pinterest, however there is no action required to download or use photos. They are merely representational of mood boards for inspiration purposes.

Preprocessing/cleaning/labeling

- **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)?** If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Cleaning was manually handled due to the nature of the images. I needed to primarily see if there were any images that were visually very similar (as I didn't wasn't any overcompensating or for training to be biased towards a certain image) or if any duplicates were present. Duplicates occurred as some images appeared on different boards that I had downloaded.
The images were all renamed on a numerical bases for ease of handling.
All of the images were 'cropped to square' so the was a linearity between the images and to ensure future training was made optimal.

- **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.

Images were saved as full-size PNG files from a website. They may not have necessarily been 'raw' versions.

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

None available as this needed to be done individually at a personal discretion and no specific software was used, however commands via the command prompt were used to form a linearity between the images once downloaded.

- **Any other comments?**

The dataset tools library as used in previous UAL session was used for the collection and cleaning process of the dataset.

Uses

- **Has the dataset been used for any tasks already?** If so, please provide a description.

No, it has not. However, I do not have awareness if individual images have been used previously on any form of training.

- **Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.

No, n/a.

- **What (other) tasks could the dataset be used for?**

Image tagging (AI to learn the angle of an aerial shot), content creation (AI to generate what it thinks to be an aerial shot and therefore new associations to what they look like can be identified) or scientific discoveries (in the same fashion, if images input is of experimental nature, output could be of beneficial use in research).

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

If any of the images are to be used in a widespread context (e.g., publishing) the images will need to be sourced back to their owner for copyright laws to be correctly implemented and if use can be allowed. This dataset should not be used towards a third party and the images should be obtained directly for commercial use.

The images have also been cropped to square for linearity for AI training, so the images have a loss of data against their full aspect ratio versions.

- **Are there tasks for which the dataset should not be used?** If so, please provide a description.

Commercial uses of images where copyright may be infringed. For example: advertising and publishing.

- **Any other comments?**

The dataset does not result in any unfair treatment of individuals or groups or cause any undesirable harm.

Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.

The dataset will not be distributed beyond the institution. Only those with institutional access will be able to access the dataset.

- **How will the dataset be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?

As the dataset will not be publicly distributed, it does not have a DOI. This will be shared within the institution via a cloud link.

- **When will the dataset be distributed?**

Upon reading this, the dataset it already distributed to date through the institution previously stated.

- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset would most likely operate under a ToU protocol identifying the ideal uses of the dataset where copyright laws do not intervene and identifying what the dataset should not be used for commercially.

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
-

There is no third part imposed on this dataset. The images downloaded do no go against any copyright fraud due to their nature being help on a social media based platform.

- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No export controls apply to the dataset or regulatory restrictions. They are maintained as PNG files to handle the graphic quality of the images.

- **Any other comments?**

N/a.

Maintenance

- **Who will be supporting/hosting/maintaining the dataset?**

I, as the dataset curator will be maintaining the dataset where required. The parameters may be altered as required for AI training purposes. However, the dataset is dormant on the institution platform so others may download and alter the dataset as they wish.

- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

I, as the dataset curator can be contacted by email address. This is my best point of contact.
would further provide an ideal email address

- **Is there an erratum?** If so, please provide a link or other access point.

There is no erratum and therefore there is not updated version of the dataset to be acquired for institutional use.

- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

If the dataset were to be updated, I would update it with corresponding labels so that the images could be grouped or pooled as required. No images will be deleted however, images may be added to the dataset should they be suitable and unique against other images already in the dataset.

- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.

N/a, not related to people.

- **Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Older versions of the dataset would not be supported as the dataset will only continue to built upon and obtain more image files with the dataset.

- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

I would advise, that as the legal owner of a dataset of images, that it must be requested to add to the dataset. I would not advise freely adding images to the dataset if they are not of similar content and so, this should be regulated. Therefore, requests to contribute will be a way of validating the images so that the content is of a similar nature to those in the dataset.

- **Any other comments?**

Any attempts to contribute with zero consideration for the content of the dataset may be acknowledged and reported.