Template for Coursework Question 2: Creating a datasheet

3.2 Composition

Dataset creators should read through these questions prior to any data collection and then provide answers once data collection is complete. Most of the questions in this section are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for their chosen tasks. Some of the questions are designed to elicit information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

Questions that apply only to datasets that relate to people are grouped together at the end of the section. We recommend taking a broad interpretation of whether a dataset relates to people. For example, any dataset containing text that was written by people relates to people.

- **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 be- tween them; nodes and edges)? Please provide a description.

  Answer: This dataset contains a large number of images. All the images are of items in different states. For example, caramelised apples, sliced apples, core apples, old houses, big houses, etc. There are no dependencies or interactions between all the images, they are all independent data. In addition, there is a csv file called "adj_ant.csv", which contains the list of antonyms for each adjective and independent from the images.

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

  Answer: 63,440 images depicting 245 nouns modified by a total of 115 adjectives. Each individual noun is only modified by ~9 adjectives it affords.

- **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).

  Answer: This is a sample of all existing items and states. As there are so many different kinds of items in the world, there are also many different kinds of states of items. It is not possible to record all objects and states in the form of pictures. This dataset is also not representative of all datasets. Because:

  - There are a large number of items with different properties in the world.
  - Each different item may have a different state.
  - Each different item may have a different state transition process.

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

Answer: The raw data for this dataset consisted of images. No processing of the images was done, but some unclear and mislabelled images were cleaned up manually through the crowdsourcing service platform.

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

  Answer: No, each instance (image) corresponds to just one label.

- **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.

  Answer: This data is not recommended for random splitting. In the dataset there are folders with names starting with "adj" where the items have multiple states. This data can be used for the test set. Random splitting is also not recommended for the training and validation sets. Try to split a portion of each label. This will ensure that each label has data available for training and validation.

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

  Answer: The data has been filtered manually, but there are still some unclear and mislabelled images in the data. For example(Not real), the data for apples labelled as caramel has images of cut apples.

If the dataset does not relate to people, you may skip the remaining questions in this section.

- **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.
- **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.
- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race 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.

3.3 Collection Process

As with the questions in the previous section, dataset creators should read through these questions prior to any data collection to flag potential issues and then provide answers once collection is complete. In addition to the goals outlined in the previous section, the questions in this section are designed to elicit information that may help researchers and practitioners to create alternative datasets with similar characteristics. Again, questions that apply only to datasets that relate to people are grouped together at the end of the section.

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** How were these mechanisms or procedures validated?

Answer: The data is collected by using the search engine Bing to search for data on different labels. Validation of the data is done manually

- **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)?**

  Answer: The researchers and thesis writers for the project. They are: Phillip Isola, Joseph J. Lim and Edward H. Adelson

- **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.

  Answer: No.

If the dataset does not relate to people, you may skip the remaining questions in this section.

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

3.5 Uses

The questions in this section are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

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

  Answer: Yes. This dateset used to discovering states and transformations in image collections

- **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.

  Answer: No, but there is a link which contains a paper using data and dataset source: http://web.mit.edu/phillipi/Public/states_and_transformations/index.html.

- **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 dataset consumer 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 risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

Answer: The composition of the data set is balanced and there are no factors contributing to the imbalance. However, users need to be aware that there may be errors or noise in the dataset. Therefore pay attention to the data before using it.

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

  Answer: No