Worksheet 5 - outline solutions

Conor Houghton, copying from Martha Lewis

Ethical regards

Context

Neglect in communication between data subjects, data curators and data consumers can lead to a variety of ethical transgressions; amplified in anthropic fields. A method to improve communication is to encourage the widespread use of elucidatory dataset descriptives. There are several proposed standardized formats that may enhance the informational through-put of this process. Outlined in the lectures: Checklists, Codebooks, Datasheets, Model Cards. An influential proponent of these methods is Timnit Gebru who pithily outlines how data descriptives can be seen analogous to the datasheets that accompany all electronic devices, instructing of intended use and safety precautions. Gebru et al, 2017 propose 'Datasheets for Datasets', in which a prescient set of recommended reflections are outlined:

1. Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.

For what purpose was the dataset created?

To provide a large scale, accurate and diverse image data set for computer vision modelling.

- Extending the WordNet dataset by providing 500-1000 clean and full resolution images to 80000 synsets in the WordNet dataset

Was there a specific task in mind?

The general task is computer vision; in particular the utility of the hierarchical dataset was demonstrated in three applications: object recognition, image classification and automatic object clustering.

Was there a specific gap that needed to be filled?

ImageNet is larger in scale and diversity and more accurate than image datasets that were available.

- Larger and more challenging dataset is needed.

- Lower noise level and higher resolution images is needed – making the dataset more suitable for general purpose algorithm development, training, and evaluation.
- More balanced distribution of images across the semantic hierarchy.
- Sense disambiguation needs to be mitigated.
- Freely available.

Who created the dataset and on behalf of which entity?

The dataset was originally created by Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei on behalf of Princeton university.

The senior research team is now Li Fei-Fei, PI, Stanford University, Jia Deng, Princeton University, Olga Russakovsky, Princeton University, Alex Berg, University of North Carolina, Chapel Hill, Kai Li, Princeton University

Who funded the creation of the dataset?

Wei Dong is supported by Gordon Wu fellowship. Richard Socher is supported by the ERP and Upton fellowships. Kai Li is funded by NSF grant CNS-0509447 and by research grants from Google, Intel, Microsoft and Yahoo!. Li Fei-Fei is funded by research grants from Microsoft and Google.

2. Composition

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

What do the instances that comprise the dataset represent? Are there multiple types of instances?

All instances contain images that traverse a range of themes from animals to geoforms, from furniture to people.

- ImageNet follows the hierarchical tree structure of the WordNet database covering a range of themes ("subtrees" as described by the author). Each "subtree" consists of semantically related synsets. A synset is a concept described by multiple words or word phrases, which is illustrated by on average 500-1000 images.

How many instances are there in total?

ImageNet aims to contain in the order of 50 million cleanly labelled full resolution images (500-1000 per synset). It presently contains 14,197,122 images. At the time of publish, 2009, the dataset contained 12 subtrees consist of a total of 3.2 million annotated images.

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

The dataset is a sampled through internet search engines. To obtain a diverse sample for each synset, researchers measure the compression of the average image in each synset.

- By construction, the dataset does not contain all possible instances.
- Not all synsets from the WordNet dataset was illustrated by images in ImageNet at the time of construction of the dataset. This is subject to be improved in the future.
- Candidate images are pulled from search engines where there is a limit to how many images can be retrieved per search. To improve the diversity and the size of the candidate pool, synsets are translated to other languages and used to query the search engine. The candidate pool is then manually "cleaned"" by humans, which is subject to false negative errors.

If the dataset is a sample, then what is the larger set?

It is most sensible to consider the larger set to be all 'Fair use' images.

- Providing images to the whole WordNet dataset.
- All images associated with each synset from 'Fair use' images.

Is the sample representative of the larger set?

No, the use of the internet is not culturally or geographically uniform. Therefore there is likely bias in image samples towards countries and populational sections that have more access and interest in internet use.

What data does each instance consist of?

Image, label and hierarchical relations (Hyponymy).

Is there a label or target associated with each instance?

Each image is assigned to the appropriate Wordnet synset and label.

Is any information missing from individual instances?

Extra content captured in images has not been accounted. Neither actual or upload location, or date are recorded.

Are relationships between individual instances made explicit?

Yes, this is key to the hierarchical structure of the dataset.

Are there recommended data splits?

ILSVRC2012: ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 utilised a 67-33train-test split.

Are there any errors, sources of noise, or redundancies in the dataset?

Extra-focal content may be considered noise, the dataset labelling was conducted using human majority vote. Errors may have occurred to a larger degree in cases where an image consist of several objects.

- Candidate images are screened by humans provided with definitions of each synset. Two issues arise: human errors and disagreement in judging images.

Is the dataset self-contained, or does it link to or otherwise rely on external resources?

The dataset relies on images uploaded to the internet and accessible by search engines. Therefore, collated data could be making biased use of particular media sites.

Additionally, the publicly available dataset relies on URLs to the original images, plus links to WordNet. Since the URLs are from 2009, many of them may be broken. An initial random search shows that some are indeed broken. The original images can be downloaded for non-commercial/educational use.

Does the dataset contain data that might be considered confidential?

The 'Fair use' legislation allows search engine image results to be used for research that is in line with general public interest without explicit permission from owners.

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

Recent work by K. Yang et al, 2019 outlined how ImageNet has been filtered to eradicate offensive synsets and inappropriate labels. It also outlines efforts to balance representation.

Does the dataset relate to people?

Yes, but the people category is one of many categories.

Does the dataset identify any subpopulations?

The nature of the hierarchical structure does by nature discriminate between groups of people; however, insensitive or harmful classifications have been removed.

Is it possible to identify individuals, either directly or indirectly from the dataset?

No information other than image is explicitly recorded. It is possible that images include implicit exposing information e.g. location may be inferred. However recent attempts have been made to safeguard vulnerable groups and filter images especially the people category.

Does the dataset contain data that might be considered sensitive in any way?

No

4. Collection Process

As with 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 of the prior section, the answers to questions here may provide information that allow others to reconstruct the dataset without access to it.

How was the data associated with each instance acquired?

Due to the scale of the project, an automated system (web scraper) was used to collate a set of images associated with wordnet synsets. Only 10further labelling was required, to this end, a paid for service, Amazon Mechanical Turk was utilised. This service pays users to fulfil a variety of online jobs. Web scraped images were distributed to users and majority vote criterion were used to label. A variety of search engines and languages e.g. Chinese, Spanish, Dutch and Italian were used to obtain a better populational representation.

Was the data directly observable, reported by subjects or indirectly inferred/derived from other data?

The data images were obtained indirectly through a search engine; labels identified by human vote.

If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified?

Majority vote-based algorithms. More recently sensitive synsets have been 'cleaned' by researchers.

What mechanisms or procedures were used to collect the data?

For each synset, the ImageNet team in 2009 automatically downloaded images by querying search engines. Information regarding the software behind this automated process was not provided.

How were these mechanisms or procedures validated?

Diversity measures were validated by analysing covariance between images within synsets with an objective of low correlation. They could have also used qualitative assessment akin to Datasheets for Datasets.

If the dataset is a sample from a larger set, what was the sampling strategy?

Images more relevant to the researcher search tags were sampled with higher weight, as were images with a higher weighted majority vote amongst AMT users. This sampling was not explicit.

Who was involved in the data collection process?

Researchers were involved in the data collection whilst crowd workers were heavily involved in the data labelling process. Crowd workers were paid via Amazon Mechanical Turk pay schemes.

Over what timeframe was the data collected?

The project started in early 2007, the dataset took 2 and a half years to complete, and the paper was published in 2009.

Does this timeframe match the creation timeframe of the data associated with the instances?

As these are images scraped from the internet during that timeframe associated to synsets, this should persist overtime.

Were any ethical review processes conducted

? At the time of publication, the paper had no information on whether the dataset was reviewed in an ethical lens. In 2019, a report was submitted for peer review by the original authors and other associates on the diversity of the dataset under the people subtree of the dataset.

Does the dataset relate to people?

Yes, a "subtree" exists for "people".

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources?

The data was collected indirectly via search engines.

Were the individuals in question notified about the data collection?

No.

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

No.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

No.

Has an analysis of the potential impact of the dataset and its use on data subjects been conducted?

Information not available in the original paper.

5. Preprocessing/cleaning/labelling

Dataset creators should read through these questions prior to any preprocessing, cleaning, or labeling and then provide answers once these tasks are complete. The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

Was any preprocessing/cleaning/labeling of the data done?

Authors relied on humans to verify candidate images downloaded from search engines. Amazon Mechanical Turk service was used during this "cleaning" process, where people were paid to do a task for money. Candidate images and definition of associated synset was provided to the human who will go through each image to validate the image. This validation process is performed independently, and a voting system was put in place, the majority vote is accepted as the final judgment on the images.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?

Information not available from original paper.

Is the software used to preprocess/clean/label the instances available?

A simple algorithm was developed to dynamically determine the number of agreements needed for different category of images which produces a confidence score table, indicating the probability of an image being a good image given the user votes.

6. Uses

These questions 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?

At the time of publication, the dataset was used in three experiments to show the advantages of the dataset, underlining its potential and advantages over other datasets. Since the construction of this dataset, it is used in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC).

Is there a repository that links to any or all papers or systems that use the dataset?

Available on ImageNet's official website (http://image-net.org/about).

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

Image classification and visual recognition.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

Issues were identified after the publication of the dataset concerning diversity in images under the "people" subtree. Investigatory report was published for peer review in 2019 outlining potential unsafe uses and its discriminatory range.

Are there tasks for which the dataset should not be used?

Facial recognition. The dataset can be used to train classifiers of people, this use is problematic as the dataset is demonstrated to have disproportionate error rates across race groups due to underrepresentation.

7. Distribution

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created?

ImageNet does not own the copyright of the images. ImageNet only compiles an accurate list of web images for each synset of WordNet. The dataset is publicly available for research and educational use.

How will the dataset be distributed? Does the dataset have a digital object identifier (DOI)?

This dataset can be downloaded from their official website. URLs to the images are publicly available at http://image-net.org/download.php.

When will the dataset be distributed?

The dataset was made available upon publication of the paper in 2009 and is continuously being updated.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

The images are distributed under terms of use as described in http://image-net.org/download-faq.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

Yes, see http://image-net.org/download-faq.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

No information about this is available from the original publication. Since January 2019, authors have disabled downloads of the full ImageNet dataset, except for the small subset of 1000 categories used in the ImageNet Challenge.

8. Maintenance

As with the previous section, dataset creators should provide answers to these questions prior to distributing the dataset. These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.

Who is supporting/hosting/maintaining the dataset?

Stanford Vision Lab, Stanford University, Princeton University

How can the owner/curator/manager of the dataset be contacted?

Fei-Fei Li feifeili@stanford.edu

Is there an erratum?

Information not available.

Will the dataset be updated?

There are updates on the website: http://image-net.org/update-sep-17-2019.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?

No.

Will older versions of the dataset continue to be supported/hosted/maintained?

No, it is continually updated.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

It is possible for people to sign up to be part of AMT to validate candidate image pools.