F Datasheet for ImageNet-Hard

F.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. The latter may be particularly relevant for datasets created for research purposes.

- **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 ImageNet-Hard is a new benchmark to test the robustness of state-of-the-art image classifiers. It comprises an array of challenging images collected from *six* validation datasets of ImageNet. This dataset challenges state-of-the-art image classification models because even by perfectly localizing the key objects, the state-of-the-art classifiers still fail to correctly recognize.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

  The dataset was created in collaboration efforts between the University of Alberta, Canada, and Auburn University, USA; mostly by, Mohammad Reza Taesiri and Anh Nguyen.
- **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.

  Anh Nguyen was supported by NSF Grant No. 2145767, and donations from NaphCare Foundation, and Adobe Research
- **Any other comments?**

  No.

F.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 between them; nodes and edges)? Please provide a description.

  Each instance of the ImageNet-Hard dataset corresponds to an image and at least one groundtruth label that will be used to assess image classifiers.
- **How many instances are there in total (of each type, if appropriate)?**

  There are 10,980 images in this dataset.
- **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).

  ImageNet-Hard is a combination of various publicly available datasets. We tried multiple refinement steps to make sure to get the best possible samples for the intended purpose. Then, it is not representative of any larger sets but a selective combination of multiple sets.

- **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 dataset contains both raw and processed images. The processed images come from ImageNet-C. Details can be found in Sec. 4.4.

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

  Yes. Each sample has the label that is the folder name the image belongs to. Basically, we follow the structure of the ImageNet paper [56].

- **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.

  No.

- **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.

  No. The individual instances has no relationships.

- **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.

  No. This dataset is created for the testing purposes.

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

  To the best of our knowledge, No. We tried our best efforts to filter any errors, sources of noise, or redundancies to create the ImageNet-Hard dataset.

- **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 *dataset consumer*? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

  Yes. It does link and inherits from existing image datasets and was detailed in Sec. 4.4.

- **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.

- **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.

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.

  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 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.

N/A.

- **Any other comments?**

  No.

F.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.*

- **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 the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

  The dataset is linked from other 6 datasets. Please find the contribution of original daatasets in Appendix E.1)

- **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?

  We used both algorithm and human efforts to collect the data. Algorithms were used to choose hard samples from various datasets. We then used two human groups and their agreement to make sure the high quality of the process. Details for the human validation in Sec. 4.4. Finally, we removed samples that have debatable labels (e.g. sunglass vs. sunglasses).

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

  No, the dataset was not a subset of a larger set.

- **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)?**

  In the data collection process, we involved students who voluntarily participated.

- **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.

  Feedback data was collected from April 20 2023 – May 4 2023.

- **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.

  N/A. In this study, humans are not the subjects. Their voluntary feedback, however, is used to filter out incorrectly labelled samples from the original 6 datasets.

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)?**

  We only involved individuals in the label verification step (i.e. 3133 samples). The answers from individuals directly affect if one of those 3133 samples will be kept or not.