ViPR: datasheets for datasets

Evžen Wybitul, Mikhail Seleznyov, Evan Ryan Gunter, David Lindner

June 2024

1 Datasheets for datasets

Datasheets for Datasets "document [the dataset] motivation, composition, collection process, recommended uses, and so on. [They] have the potential to increase transparency and accountability within the machine learning community, mitigate unwanted biases in machine learning systems, facilitate greater reproducibility of machine learning results, and help researchers and practitioners select more appropriate datasets for their chosen tasks."

The motivation behind the proposal was the electronics industry, where every component has a datasheet that describes its operating characteristics and recommended uses. In machine learning, data is the input for model training. Using the wrong dataset, or using a dataset outside of its original intent, or even not understanding well enough the limitations of a dataset, has dire consequences for the model. However, "[d]espite the importance of data to machine learning, there is no standardized process for documenting machine learning datasets. To address this gap, we propose datasheets for datasets."

2 Template

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.

Reinforcement learning (RL) requires either manually specifying a reward function, which is often infeasible, or learning a reward model from a large amount of human feedback, which is often very expensive. An interesting alternative is to use Vision-Language Models: either as zero-shot/few-shot reward models, or as a substitute for human feedback. However, those approaches are not robust yet. This dataset is created to benchmark capabilities of Vision-Language Models, necessary to use them for reward modeling in Reinforcement Learning. It should be used solely for evaluation purposes.

Who created this dataset (e.g.,

which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

We will disclose this information if the paper will be accepted.

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.

We will disclose this information if the paper will be accepted.

Any other comments?

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 dataset can be thought of as a collection of multiclass classification tasks. In each classification task, instances are pairs (video, class label).

In the Minecraft subset of the dataset, the same video can be reused with different labels.

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

There are 3164 videos in virtual home environment, 53 videos in Minecraft environment and 1267 real life videos, which totals to 4484.

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 aims to represent both simulated RL environments (via virtual home and Minecraft environments) and realistic setting (via real life videos).

While our dataset captures only a subset of infinite possible behaviors of the agents in those environments, we believe it is still representative, since the tasks/behaviors we test are motivated by the typical objectives in each environment.

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

Each instance is a pair (video, class label). The video is usually stored in 'mp4' format, the class label in written in natural language.

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

By design, each instance has a label, which is written in natural language.

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.

The are no instances with missing information.

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.

The relationships between instances are not represented in explicit form (except for the fact that instances are grouped into separate multi-class classification tasks).

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.

This dataset is created for evaluation purposes only, so all instances should be treated for testing.

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

Minecraft and some real life videos were annotated manually, so potentially the labels can contains errors due to the human factor.

Virtual home videos and some real life videos were labeled automatically using the metadata from existing benchmarks (ALFRED, Kinectics-700), so it might inherit their errors.

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.

We will provide a self-contained version of the dataset, including all videos and all labels.

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.

To the best of our knowledge, the dataset does not contain confidential data.

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

To the best of our empathy, the dataset does not contain any offensive, insulting, threatening or otherwise anxiety-causing data.

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

This dataset does not relate to people.

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?

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.

Minecraft and some real life videos were annotated manually.

Virtual home videos and some real life videos were labeled automatically using the metadata from existing benchmarks (ALFRED, Kinectics-700).

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?

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

The videos are not directly sampled from ALFRED but rather created by editing (oftentimes multiple) ALFRED videos. The whole data creation protocol is detailed in the documentation.pdf document.

With Minecraft environment, part of the videos are from BASALT Benchmark. They were chosen at random from a subset of Demonstration dataset, loaded using instruction from the BASALT Benchmark repository with the optional argument for maximum download size in MB set to 1000.

For real life part, we collect 100 videos of a door opening and 100 videos of a door closing from the Kinetics-700 dataset. These videos were selected to contain either a door opening or closing, but not both.

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)?

The only people involved in the data collection process are the authors of the paper.

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.

Virtual home videos were generated using ALFRED Benchmark, which was released in 2020.

Minecraft subset includes videos from Demostration part of BASALT Benchmark, which were collected between April 2022 and July 2022, and videos created specifically for the ViPR between April 2024 and June 2024.

Real life videos were partly taken from Kinetics-700 (released in 2019) and partly manually recorded between April 2024 and June 2024.

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 reiew process was conducted.

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

This dataset does not relate to people.

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?

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.

For the virtual home environment, videos are short clips or concatenations of short clips, cut from the original trajectories. The labels were built automatically based on the labels of the original trajectories.

For the Minecraft environment, videos are short clips from longer footages of the gameplay.

For the real life videos, ...

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.

”Raw” data was not saved.

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

To cut short clips out of longer videos, LossLessCut was used, which is a cross platform GUI over FFmpeg.

To record videos of Minecraft gameplay, OBS Studio was used, which is a free and open source software for video recording and live streaming.

Some other processing was done with custom Python scripts, provided in the code.

Any other comments?

Uses

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

This dataset has been only used for evaluations, described in the paper.

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.

The is no such repository yet.

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

We only intend this dataset for evaluation purposes.

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?

To the best of our knowledge, there are no pitfalls which might impact future uses and cause undesirable harms.

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

One should not use this dataset for training and/or finetuning.

Any other comments?

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 be open-sourced.

How will the dataset will be distributed (e.g., tarball on website,

API, GitHub) Does the dataset have a digital object identifier (DOI)?

We plan to distribute the dataset via GitHub and as an archive on Google Drive or Google Cloud.

When will the dataset be distributed?

We plan to distribute the dataset in the summer of 2024.

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.

We plan to distribute the dataset under CC BY 4.0 license.

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.

No.

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.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of the paper.

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

The owners of the dataset can be contacted via email, mentioned in the camera-ready version of the paper.

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

No.

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)?

The dataset might be updated during summer 2024. The updates will be communicated via Github.

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.

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 will likely not be maintained. Their obsolescence will be communicated via GitHub repository page.

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.

When the dataset GitHub page will be released, contributors might create pull requests.

Any other comments?