DATASHEET

Lakshya Borra

April 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.

The dataset was created with the intention of developing a machine/deep learning model capable of analyzing images and audio simultaneously. Initially, the goal was to imitate filters commonly found in applications that suggest various associations based on facial features and voice characteristics. Specifically, the aim was to design a model that, given an individual's image and voice, could suggest which "Straw Hat" member from the popular anime series "One Piece" the person resembles the most.

The creation of this dataset was motivated by a notable gap in existing solutions for simultaneously analyzing both images and audio inputs. While

there are numerous applications focusing solely on image recognition or speech processing, there was a distinct lack of integrated models capable of processing both modalities concurrently. This gap prompted the development of a dataset that could facilitate the training of such models, addressing the need for comprehensive solutions that leverage both visual and auditory cues in tandem.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The dataset was created independently by myself as part of a mini-project undertaken during the Multi-Modal Data Processing and Learning-II course at IIT Guwahati. This project was completed as an individual task within the guidelines of the course syllabus.

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.
The creation of the dataset was not funded. It was part of a project assigned by Prof. Neeraj Kumar Sharma and Prof. Prashant Wagambar Patil of IIT Guwahati as part of academic coursework. Therefore, there was no associated grant or external funding source involved.

Any other comments?

While the primary goal of creating this dataset was to develop a model capable of identifying similarities between individuals and characters from 'One Piece' based on facial features and voice, the dataset's versatility extends beyond this initial objective. Once the model is trained using this dataset, it can be leveraged for a myriad of applications across various industries and sectors.

One notable application is in law enforcement, where the developed model could assist in identifying and locating individuals involved in criminal activities after being trained on criminal's dataset (which can be easily accessed by the government people). By comparing the voice and photo of a suspect to those in the dataset, law enforcement agencies could potentially identify and apprehend criminals. For instance, if a known criminal were to visit a location where the system is deployed, it could automatically match the individual to the database and alert authorities if there is a match, facilitating swift action by law enforcement personnel.

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 within the dataset represent two distinct modalities: images and audio. The images correspond to pictures of characters from the "Straw Hat" pirate crew, while the audio samples correspond to their voice recordings.

In terms of multiple types of instances, the dataset encompasses two main categories: images and audio recordings. Each instance consists of a pair, with an image representing a character's visual appearance and an

audio recording capturing their voice sample. These modalities can be considered as distinct types of instances within the dataset, each contributing to the overall goal of character recognition and similarity assessment.

How many instances are there in total (of each type, if appropriate)?
The dataset comprises a total of 10 categories, each corresponding to a member of the Straw Hat pirate crew. For each category, there are approximately 150 images representing visual appearances and approximately 5 minutes of audio recordings capturing voice samples. Therefore, the total number of instances in the dataset is approximately:

Images: 10 categories × 150 images per category = 1500 images

Audio: 10 categories × 5 minutes of audio per category = 50 minutes of audio

So, in total, there are approximately 1500 images and 50 minutes of audio in the 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).
The dataset contains all possible instances for its intended purpose, which is to classify individuals into one of the 10 categories representing members of the Straw Hat pirate crew. As there are currently 10 known members in the crew, the dataset encompasses all possible instances within this classification framework.

Since the classification task is based on identifying similarities between individuals and the existing members of the Straw Hat crew, there is no larger set from which this dataset is sampled. Therefore, the dataset is not a sample but rather a complete representation of the target classification problem.

It's important to note that while the dataset covers all current members of the Straw Hat crew, it may need to be updated in the future if new members join the crew in the ongoing anime series. However, for the current scope of the dataset, it contains all possible instances relevant to its classification task.

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 in the dataset consists of raw data in the form of images and audio recordings:

1.**Images**: Raw image files representing the visual appearance of each member of the Straw Hat crew. These images were collected through web scraping from Google and some were manually collected. They are stored in formats such as JPEG or PNG.

2.**Audio**: Raw audio recordings capturing the voice samples of each member. These audio files are primarily in MP3 format and were manually collected.

Since no feature extraction has been performed on the data, it remains in its original raw format. The images and audio recordings serve as direct in-

put for any subsequent processing or analysis conducted using the dataset.

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

Yes, there is a label associated with each instance in the dataset. The labels correspond to the ten characters of the Straw Hat pirate crew, namely Luffy, Zoro, Nami, Usopp, Sanji, Chopper, Robin, Franky, Brook, and Jimbei (Figure 1 at last page of this datasheet). Each image and audio recording has been manually labeled according to the character it represents.

These labels serve as the target variable for supervised learning tasks, such as classification or similarity assessment. They indicate which member of the Straw Hat crew the instance (image or audio) is associated with, enabling the development and evaluation of machine/deep learning models for character recognition or similarity detection purposes.

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, there is no missing information from individual instances in the dataset. Each instance, whether it is an image or an audio recording, contains complete data without anything being removed. All necessary information, including the visual appearance and voice sample of each character from the Straw Hat crew, is present in the dataset.

There are no instances where information is missing or incomplete due to unavailability or redaction. This ensures that the dataset is comprehensive and suitable for analysis and modeling purposes without any gaps in the data.

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.

In this dataset, there are no explicit relationships between individual instances. Each image and audio recording represents a standalone instance associated with a specific character from the Straw Hat crew. The dataset focuses on character recognition and similarity assessment based on visual and auditory features, rather than on explicit relationships between instances such as social network links or user interactions.

Therefore, there are no interconnections or dependencies between instances that need to be considered during analysis or modeling. Each instance can be treated independently when performing tasks such as classification or similarity comparison.

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 specific recommendations provided for data splits in this dataset. However, practitioners can follow the basic train-validation-test split rule commonly used in machine learning tasks. The dataset can be divided into a training set, used for

model training, a validation set, used for hyperparameter tuning and model selection, and a testing set, used for final model evaluation. This approach helps ensure that the model is trained on a subset of the data, validated on a separate subset, and finally tested on unseen data to assess its performance. The proportions for the train-validation-test split can be adjusted based on the size of the dataset and the specific requirements of the modeling task.

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

The dataset contains certain errors, sources of noise, and redundancies that require attention:

1. **Image Variability**: Images in the dataset exhibit variability in terms of size and aspect ratio. Some images may be larger or smaller than others, and their shapes may vary. This variability necessitates preprocessing steps such as resizing and normalization to ensure consistency across the dataset.

2. **Audio Noise**: Audio samples in the dataset contain background noise, which can introduce variability and affect model performance. To prevent overfitting and improve generalization, it's essential to address this noise through preprocessing techniques such as noise reduction or augmentation. Additionally, while some samples contain background noise, others may be relatively clean. This variation provides the model with exposure to different acoustic environments and helps improve its robustness to real-world conditions.

By addressing these errors, sources of noise, and redundancies through appropriate preprocessing techniques, the dataset can be made more suitable for training robust machine learning models for character recognition and similarity assessment tasks.

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 and does not rely on external resources. It consists of images and audio recordings scraped from Google and collected directly from episodes of the anime series "One Piece." Therefore, there are no guarantees required for external resources to exist or remain constant over time.Additionally, there are no restrictions such as licenses or fees associated with any external resources, as all data was collected from publicly available sources.

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, the dataset does not contain any data that might be considered confidential. As it solely comprises information about fictional characters from the anime series "One Piece," there is no content that is protected by legal privilege, doctor-patient confidentiality, or any other forms of confidentiality.

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

The dataset does not contain any data that might be offensive, insulting, threatening, or otherwise cause anxiety. All images and audio recordings are safe for viewing and do not contain any content that could potentially be harmful or distressing to individuals. The dataset is designed to facilitate machine learning research and applications in a safe and appropriate manner.

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

The dataset is particularly relatable to anime viewers.

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.

The dataset does not identify any subpopulations based on age, gender, or any other demographic factors. The focus of the dataset is character recognition and similarity assessment within the context of anime characters from the Straw Hat pirate crew. While anime viewers may indeed encompass diverse age groups and genders, the dataset itself does not categorize individuals based on these demographic characteristics. As such, there are no specific subpopulations identified or characterized within the dataset. This inclusive approach ensures that the dataset remains relevant and applicable to a broad audience of anime enthusiasts without excluding or targeting specific demographic groups.

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.

No, it is not possible to identify individuals directly or indirectly from the dataset. The dataset is focused on character recognition and similarity assessment within the context of anime characters from the Straw Hat pirate crew. It does not contain personal or identifying information about individuals outside of the characters represented in the images and audio recordings.

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.

It solely consists of images and audio recordings related to characters from the Straw Hat pirate crew in the anime series "One Piece." There is no information related to racial or ethnic origins, sexual orientations, religious be-

liefs, political opinions, union memberships, locations, financial or health data, biometric or genetic data, forms of government identification, criminal history, or any other sensitive information. The dataset is limited to visual and auditory representations of fictional characters and does not contain any personal or sensitive data about real individuals.

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.

The data associated with each instance was acquired through a combination of direct observation and manual collection:

Images: The images were acquired through web scraping from Google and manually collected from various sources. Each image represents the visual appearance of a character from the Straw Hat pirate crew in the anime series "One Piece." The images were directly observable and did not require further validation or verification.

Audio: The audio recordings were manually collected from episodes of the anime series "One Piece." Each audio recording captures the voice sample of a character from the Straw Hat crew. While the audio data was reported by subjects (i.e., voice actors), it was not subjected to formal validation or verification processes. However, the audio recordings were carefully selected to ensure their relevance and authenticity to the characters they represent.

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?

The data collection procedures involved manual human curation for both images and audio recordings:

Images: The images were collected through a combination of web scraping from Google and manual selection from various sources. This process involved manually curating images that accurately represent the visual appearance of characters from the Straw Hat pirate crew in the anime series "One Piece."

Audio: The audio recordings were manually collected from episodes of the anime series "One Piece." This process involved selecting relevant voice samples that accurately represent the characters from the Straw Hat crew. Additionally, an online tool, specifically the Audio Cutter Online from Clideo, was utilized for editing and refining audio recordings as needed.

Validation of the data collection procedures primarily focused on ensuring the relevance and accuracy of the collected data. I manually checked the scraped Images and they are relevant. And I manually collected some Images, and they are relevant. For the audio recordings, validation was achieved by ensuring that only relevant voice

samples were selected and that they accurately represented the characters from the Straw Hat crew. This validation was carried out through manual inspection and verification during the data collection process. Additionally, the use of an online tool for audio editing provided further validation by enabling precise adjustments to the audio recordings to ensure their suitability for the dataset.

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 was created from scratch and is not a sample from a larger set, there was no sampling strategy involved.

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 data collection process was carried out by individuals as part of a course project. Since it was a project undertaken within the framework of a course, there were no external compensations provided for the data collection efforts.

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 data associated with the instances was collected recently, and it was randomly collected from a total of 1097 episodes of the anime series "One Piece." As such, the timeframe for data collection aligns with the creation timeframe of the instances themselves. There were no specific constraints or limitations on the timeframe for data collection, and the data was gathered from episodes spanning the entire series of "One Piece."

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 such review processes has been conducted.

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

Yes, this dataset does relate to people, especially if you are an anime viewer.

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

Collected via third parties (Anime episodes)

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.

There were no individuals involved in this process.

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.

There were no individuals involved in this process.

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).

There were no individuals involved in this process.

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.

Since the dataset consists of images and audio recordings related to fictional characters from the anime series "One Piece," there are no real data subjects involved whose privacy or rights might be impacted by the dataset's creation or use. As such, there has been no formal analysis conducted regarding the potential impact on data subjects.

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 image data, following preprocessing techniques are implemented:

Data Augmentation - Additional training samples has been generated by using transformations like rotation, flipping, shearing etc. as this helps in increasing the diversity of the dataset and improve model generalization.

Histogram Equalization - It helps in increasing the image contrast by redistributing pixel intensities

Grayscale Conversion - to reduce computational complexity

Resizing and **Reshaping**

For the audio data, following preprocessing techniques are implemented:

Resampling- So that the model can process and analyze the audio patterns easily

Filtering - to remove unwanted frequencies from the audio signals.

Normalization - to prevent numerical instabilities during training process

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.

Yes, the raw data has been saved and can be accessed through Github

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

available? If so, please provide a link or other access point.

No software was used for preprocessing, but the steps followed has already been discussed in this Section.

Any other comments?

Uses

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

No, this dataset has been created recently and not yet made public.

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, this dataset has been created recently and not yet made public.

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

This data can also be used for training Auto-Generative models, like generating new character based on existing images of characters and also generating them a new voice based on existing audio of characters.

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?

No, there are no such restrictions. All the data corresponds to fictional characters.

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

If this dataset suits your model or the thing you want to achieve, you can use it. There are no such tasks for which the dataset should not be used

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.

This dataset is created as a part of our course project, so it is not distributed to any third-parties outside the entity.

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

Dataset is shared through Github. As of now, it is not citated. So anyone can use if its suitable for your task.

When will the dataset be distributed?

Dataset has already been uploaded on Github, link provided in the earlier section.

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.

No copyright implementation. Its for public use.

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 are no third parties involved.

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.

There are no restrictions , anyone from anywhere can use this dataset.

Any other comments?

Maintenance

Who will be supporting/ hosting/ maintaining the dataset?

Currently, I am hosting this dataset through my Github account.

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

Its not having any restrictions/citations for downloading. So no need to contact anyone. Feel free to use this dataset.

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

No printing mistakes.

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)?

Yes, I am planning on increasing the size of the dataset by adding more samples (new instances). Since I am working individually on this dataset, It will be updated by me. Last updated dates and updation details like what updates have been made during a given span will be communicated through Github ReadMe file.

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.

No individual is involved in data collection, data is collected based on fictional characters. And there will be no retention of data.

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.

Dataset will be the same. new instances will be added to this data itself. Older versions will also be stored in the same Github. All detais will be communicated through ReadMe file.

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.

At present, there isn't a mechanism in place for others to extend, augment, or contribute to the dataset. However, establishing such a mechanism could greatly benefit the dataset's growth and utility. Further details regarding this will we be communicated.

Any other comments?

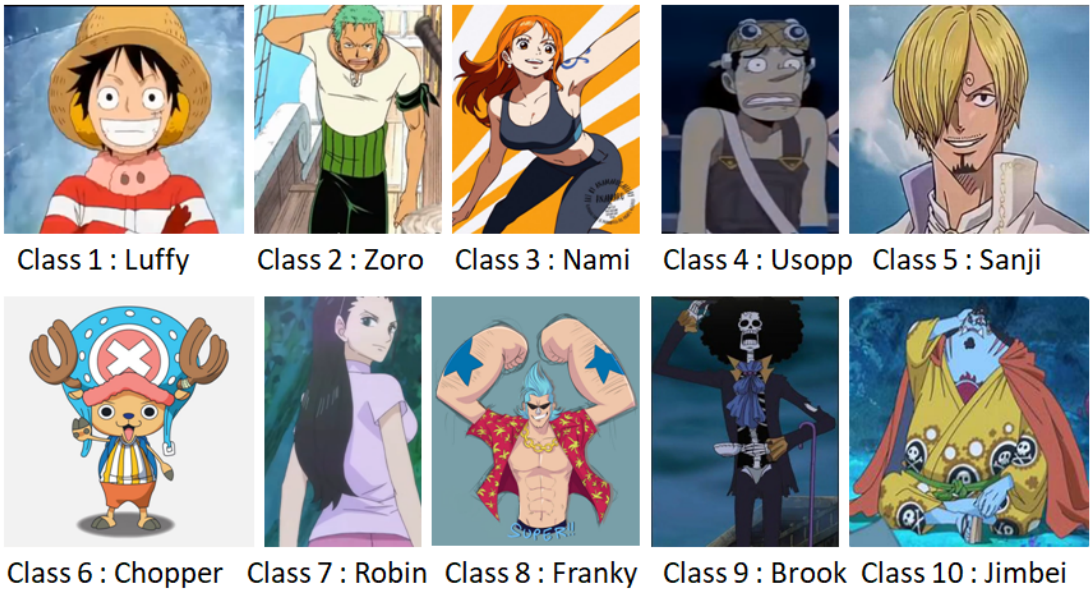

Figure 1: 10 classes of my dataset