Datasheet for the Released Datasets

June 16, 2022

1 Datasheets for datasets

This document contain the motivation, composition, collection process, schema, releasing details, maintenance etc. of the released multilingual pretraining dataset, *SignCorpus* and the fingerspelling dataset, *MultiSign-FS*. The motivation behind the dataset was the lack of large size labelled datasets for sign languages due to which sign language recognition is significantly lacking in research than other AI problems. Hence, this is an effort for advancing the progress in sign language recognition of low resource sign languages.

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.

Sign language is basically a low resource problem. There are more than 300+ sign languages used worldwide but most of them doesn't have large-scale labelled datasets. *SignCorpus* is thus prepared to boost up the progress using various methods like pretraining and multilinguality. For *MultiSign-FS* the main aim was to help learn model the fingerspelling signs of alphabets as well as numerals and to the best of our knowledge, there is no such existing multilingual dataset.

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

The dataset is programmatically created by cleaning, preprocessing the video data by the researchers at AI4Bharat group.

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 work was funded by Microsoft Philanthropies India through Microsoft AI4Accessibility program, via AI4Bharat.

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.

For *SignCorpus*, an instance in our dataset is a combination of sequence of pose keypoints, confidence value of the keypoints and the metadata related to the corresponding video in hdf5 format. The original videos are not part of the dataset.
For *MultiSign-FS*, we are releasing pose pickle files as well as videos. An instance is simply pose keypoints value correspoding to a particular character.

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

SignCorpus: There are different number of instances for each sign language, where each instance is a playlist-specific HDF5 file. For example, for American SL, the videos were taken from 26 YouTube playlists, but for Greek SL, there were 8 playlists.

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

SignCorpus: Each instance (HDF5 file) contains three groups – pose keypoints of videos from that channel, the confidence values associated with the pose keypoints and the metadata for the corresponding videos. The pose keypoints have been obtained from the MediaPipe Holistic library which gives keypoints of the joints present in the human skeleton along with confidence values of the keypoints. The metadata of the video has been obtained from YouTube. It contains useful information like title of the video, video URL, etc. Majority of the data comes from educational and news domain.
MultiSign-FS: Each instance is an pose pickle file corresponding to a video of a character and it contains pose keypoint value as well as confidence values corresponding to the keypoints.

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

SignCorpus: No. The whole multilingual pretraining dataset is unlabelled and is created with the purpose of self supervised learning.
MultiSign-FS: The fingerspelling dataset is labeled with alphabet or numeral as the label of the pose file/video.

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.

For *SignCorpus*, we are releasing only pose keypoints (not the RGB videos) along with some of the metadata for the purpose of backtracking of the video. On the other hand, *MultiSign-FS* is relatively very small dataset, so we are releasing videos as well as pose pickle file for it and none of the information is missing.

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.

All the instances are independent of each other.

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.

SignCorpus: No. We recommend to use the entire data for pretraining purpose and then use some standard dataset for the validation purpose.
MultiSign-FS: We have already done the 4:1 train-test split for every character in each of the language.

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

We tried our best to remove any form of discrepancies by downloading the videos with unique names. Also we also ran sanity checks on the extracted poses to ensure the poses are correct and belongs to signer only.

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.

For *SignCorpus*, the dataset is self contained as we are releasing pose keypoints and they can be used for the pretraining task. We are not giving the original video from which those poses are extracted. Instead of them we are giving the url as a part of metadata. With the help of url, one can backtrack to the original source but it is not guaranteed that video will be present there indefinitely. Its totally upto the owner/creator of the video to keep it hosted or not.
For *MultiSign-FS*also, the dataset is self contained as we are releasing both videos as well as pose pickle file and only poses need to be used for training purposes.

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.

For *SignCorpus*, the whole dataset is created from the videos present on the youtube. All the videos are public and have standard licensing. Also we are releasing poses instead of the videos. The poses will by itself maintain the confidentiality of the person as only keypoints are there instead of the actual person.
For *MultiSign-FS*, almost all the videos are curated from YouTube with standard licensing videos and some videos are curated from opensource websites.

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

might otherwise cause anxiety? If so, please describe why.

For *SignCorpus*, there is no such issue as we are releasing the poses only. Also, even we backtrack to the original videos, majority of the videos we considered belongs to either educational/teaching domain or News channels.

For *MultiSign-FS*, all the videos belong to either alphabet or numeral. So, there isn't any possibility that data is offensive, insulting, or threatening to anyone.

Does the dataset relate to people?

Yes, but the dataset has only the skeleton information of signers without any Personally identifiable information.

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. We are releasing only poses of the persons. They are just human skeletons keypoints.

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.

No, it does not identify any subpopulations.

Collection Process

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 curation of the sign language channels in YouTube was manual. After that, we used the pipeline (explained in "OpenHands: Making Sign Language Recognition Accessible with Pose-based Pretrained Models across Languages") for downloading the video and then preprocessing it. The downloading of the video was done using YT-DLP library. As a part of validation, we performed sanity checks on the final pose data.

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 code for crawling videos and preprocessing them has been written by the researchers at AI4Bharat group.

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 curation of the sign language channels for all the 9 languages in *SignCorpus* and then downloading the videos took almost 3-4 months.

For *MultiSign-FS*, it was almost 2-3 week to collect videos of all the sign languages.

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

SignCorpus: All the pose data released across 9 languages has been generated from the videos taken from the

youtube. They have been crawled programmatically.
MultiSign-FS: Most of these are also crawled from YouTube while for those which does not have adequate resources in YouTube, some external websites are used.

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.

The videos have been taken from the youtube channels with standard licensing and no restriction to use. Also, for *SignCorpus* we are releasing pose data instead of the actual video.

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.

Some of the bad videos (videos containing multiple signer or no signer at all) from the sign language channels have been removed manually. Apart from that, for some of the videos, cropping has been done to take only the signer region.

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

Most of the preprocessing on the data has been done manually.

Uses

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

The experiments shown in the paper are the only results available for the datasets.

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

Apart from the sign language recognition, the dataset could be used for action recognition task.

Dataset Distribution

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

The dataset will be released as zipped HDF5 files, one zip for each YouTube channel, and available via Zenodo hosting platform. A mirror link to the dataset can be availed on request (incase the platform is down or other issues).

When will the dataset be released/first distributed? What license (if any) is it distributed under?

The dataset is released along with the camera-ready version of the paper submitted finally to the conference. It is licensed under the Creative Commons Attribution 4.0 International license.

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, it will be released under the MIT 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.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The dataset is being hosted at Zenodo, an open-access repository to store and distribute scientific artifacts. The dataset is being maintained by AI4Bharat, a research lab in the CSE department of IIT Madras, India.

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)?

If there are any errors found or if any required data is missing, we take responsibility to update/rectify the same. Also in future, we might add datasets for more languages.

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.

Yes, Zenodo maintains the log of old versions as well, hence at any point of time, one may download the earlier versions (if any).

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.

Yes, researchers are welcome to extend this dataset by adding more pretraining data for any sign language, or adding content for new sign languages not studied in this work. This can be communicated / discussed with us by contacting us through the "Issues" section on the GitHub repository, or emailing the first author directly.