ComMU: Datasheet for Dataset

Hyun Lee[*]
Pozalabs
hyun@pozalabs.com

Taehyun Kim[*]
Pozalabs, Yonsei Univ.
taehyun@pozalabs.com
kimth0101@yonsei.ac.kr

Hyolim Kang
Yonsei Univ.
hyolimkang@yonsei.ac.kr

Minjoo Ki
Yonsei Univ.
minjoo@yonsei.ac.kr

Hyeonchan Hwang
Pozalabs
hyeonchan@pozalabs.com

Kwanho Park
Pozalabs
kwanho@pozalabs.com

Sharang Han
Pozalabs
sharang@pozalabs.com

Seon Joo Kim
Pozalabs, Yonsei Univ.
seonjoo@pozalabs.com
seonjookim@yonsei.ac.kr

We use the datasheets for datasets [1] for this documentation.

1 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. This dataset is created to tackle combinatorial music generation task. While commercial adoption of automatic music composition desires sophistication in generating diverse and high quality music suitable for desired context, existing datasets are insufficient in providing a comprehensive package. However, ComMU provides a complete one by being the first to allow 12 different metadata to be shown through MIDI file and to introduce track-role as one of its metadata. ComMU now opens up a new range of studies to be done with conditional music generation.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? This dataset is created by professional composers hired by PozaLabs Inc., on behalf of PozaLabs Inc..

Who funded the creation of the dataset? Creation of the dataset was funded by Pozalabs Inc.

Any other comments? N/A

2 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 is comprised of MIDI files and a csv file that contains metadata information of each MIDI files.

How many instances are there in total (of each type, if appropriate)? ComMU contains 11,114 MIDI samples with 12 different metadata information for each sample, and the total number of notes in the MIDI samples is 526,612.

[*]equal contribution, ordered by first name

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

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). ComMU contains all possible instances and is completely original. The dataset is made not part of any or larger dataset in order to keep the originality of the samples and to ensure the samples accurately reflect metadata and the composers' intentions.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description. Our dataset is consisted of raw MIDI files for each sample and its metadata information in csv file.

Is there a label or target associated with each instance? If so, please provide a description. As included in the aforementioned explanation, our dataset is a set of MIDI data and their matching metadata. The 12 metadata includes: BPM, genre, key, instrument, track-role, time signature, pitch range, number of measures, chord progression, min velocity, max velocity, and rhythm.

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 information is missing from our dataset.

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. Each metadata contains an identification number that links with the MIDI data, written in the csv file.

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. Our dataset is split into training set and validation set. The training set takes up 90 percent of the data, and the remaining 10 percent is for the validation set. We split the dataset in such way to allow effective testing of the training model through validation set, which is consisted of metadata combinations that are different from the training set.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. There are no errors, sources of noises, nor redundancies in the dataset.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? Our dataset is self-contained.

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. ComMU does not contain any 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. ComMU does not contain any data that can be offensive, insulting, threatening, nor otherwise cause anxiety.

Does the dataset relate to people? If not, you may skip the remaining questions in this section. No, ComMU does not relate to people

Any other comments? N/A

3 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. All data were acquired by manual generation of MIDI files by composers. Hence, the process of data collection was manual and original.

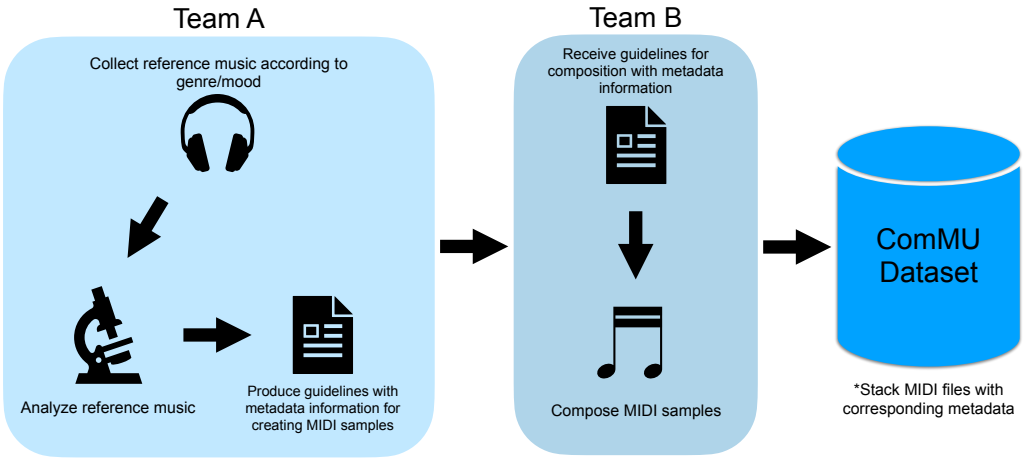

Figure 1: Data pipeline of ComMU dataset.

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? Figure 1. The whole of dataset was manually created by professional composers. Fourteen composers have teamed up to generate all 11,114 samples. The composers were divided into two different teams who (1) create composition guidelines for certain genres and moods by selecting then analyzing reference music according to metadata information and (2) compose MIDI files according to the guidelines. Here, the composition guideline served as an instruction with metadata information such as but not limited to — instruments, track-role, chord progressions, etc. — for composing samples. When the MIDI sample is created, it is validated by the guideline team members. They check whether the created MIDI sample well reflects the guidelines or not, and only validated samples are stored in the ComMU dataset. The dataset generation process took a total of 6 months.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? ComMU is all original and thus is not part 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)? As aforementioned under subsection 3.1, 14 professional composers manually created MIDI samples with metadata in constructing ComMU dataset. As we hired them and obtained consent, ComMU dataset is work made for hire and belongs to the company.

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 MIDI samples with metadata were generated manually by composers from 2021 June to 2021 December, for 6 months. ComMU is self-contained and thus its content will not change over time.

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

Does the dataset relate to people? If not, you may skip the remaining questions in this section. No, ComMU does not relate to people.

Any other comments? N/A

4 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. All instances of the dataset are original. Therefore, there were no preprocessing, cleaning, nor labeling done from any larger nor smaller dataset. However, preprocessing and labeling is done within our dataset to transform the raw MIDI files into trainable representations. First, 12 pieces of metadata information are labeled as follows: BPM, Key, Instrument, Time Signature, Pitch Range, Number of Measures, Min Velocity, Max Velocity, Rhythm. In addition, preprocessing is done before the training process through data augmentation and encoding. Data augmentation to enlarge the data is done by giving variations to BPM and audio key. Encoding is done by mapping each metadata and note sequence to an integer value according to encoding dictionary.

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. MIDI file comes in raw data, which is available here: `https://pozalabs.github.io/ComMU/`. The code for preprocessing the MIDI files is available here: `https://github.com/POZAlabs/ComMU-code`.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point. Code for preprocessing is avialable at `https://github.com/POZAlabs/ComMU-code`.

Any other comments? N/A

5 Uses

Has the dataset been used for any tasks already? If so, please provide a description. No, the dataset has not been used for any tasks yet. This paper is the first to utilize and address the dataset.

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 code needed to obtain the results discussed in the paper is uploaded in the following repository: `https://github.com/POZAlabs/ComMU-code`.

What (other) tasks could the dataset be used for? Apart from combinatorial music generation task, one could use the MIDI files in the dataset to make supposition and back-track its metadata.

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 minimal risk for undesirable harms from ComMU.

Are there tasks for which the dataset should not be used? If so, please provide a description. The dataset nor any music produced using the dataset may not be used in any and all commercial use.

Any other comments? N/A

6 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. Yes, ComMU dataset is publicly available at `https://pozalabs.github.io/ComMU/`.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)? ComMU dataset is distributed via `https://pozalabs.github.io/ComMU/`. We will make the DOI available after finalizing the cite.

When will the dataset be distributed? ComMU dataset is already distributed.

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. ComMU dataset is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International 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, there are no third party restrictions on the data associated with the instances.

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 export controls or other regulatory restrictions apply to the dataset or to individual instances.

Any other comments? N/A

7 Maintenance

Who will be supporting/hosting/maintaining the dataset? The dataset will be supported, hosted, and maintained by Pozalabs Inc..

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? All comments and questions about ComMU can be sent to Pozalabs Inc.: `contact@pozalabs.com`. Other contacts can be found at: `https://pozalabs.github.io/ComMU/`.

Is there an erratum? If so, please provide a link or other access point. All changes to the 200 dataset will be announced on `https://pozalabs.github.io/ComMU/`.

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)? We have ensured there is no error in the dataset. Hence the dataset will not be updated unless a critical error arises.

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. Expect no change in version. However, if any changes are made, all versions of ComMU will be continue to be supported and maintained on `https://pozalabs.github.io/ComMU/`. We will post the updates on the site.

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. Others can contact the authors of this paper describing their proposed extension or contribution. We would discuss their proposed contribution to confirm its validity, and if confirmed, we will release a new version of the dataset on `https://pozalabs.github.io/ComMU/` and will announce it accordingly.

Any other comments? N/A

8 Metadata

Table 1 presents the metadata of ComMU dataset.

Table 1: Metadata of ComMU dataset.

Filenames	dataset/commu_meta.csv dataset/commu_midi.tar
Format	csv, MIDI
URL	`https://pozalabs.github.io/ComMU/`
Domain	Music generation
License	CC BY-NC-SA 4.0

9 Responsibility

Pozalabs Inc. bears all responsibility in case of violation of rights, etc. The license of ComMU dataset is under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

References

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021.