Meta-Album: Multi-domain Meta-Dataset for Few-Shot Image Classification – Datasheet for Dataset

Datasheet for OCR.MD_6 dataset

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 was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

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

Haozhe Sun created the dataset, under the supervision of Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR. ChaLearn also supported the development of the software.

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.

ANR (Agence Nationale de la Recherche, National Agency for Research, https://anr.fr/), grant number 20HR0134 and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

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 instances are 128×128 RGB images of synthetic printed characters.

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

OmniPrint-MD-6 has 28120 images from 703 classes (each class has 40 images).

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

Preprint. Under review.

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).

These datasets are synthesized from the data synthesizer OmniPrint, thus they can be viewed as a sample of instances from all the possible images given the nuisance parameters (fonts, styles, noises, etc.). These datasets are representative of such images because the synthesis parameters of each instance were uniformly sampled, no further selection was performed.

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 128×128 RGB image. Each image contains one single character from a certain script, rendered in a particular way (background, foreground, distortions, noises).

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

Yes, each instance has a category/label which is provided with the images in meta-data. The category describes the character identity. Along with the category, the meta-data also has super-category (alphabet or subset of alphabet) and various synthesis parameters.

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, all of the metadata is provided for each instance.

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 relationships are contained in categories and super-categories, all provided.

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.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

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

We intentionally introduced various transformations and noises to each image instance. The transformation parameter space is large so there is little chance that two instances are identical.

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. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

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.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

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.

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

No.

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.

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? I If so, please describe how.

No.

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.

No.

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

Each instance is synthesized by OmniPrint. Each instance is an image and is directly observable.

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 is synthesized using the data synthesizer OmniPrint. The involved Unicode characters were manually selected from the Unicode standard, which constitutes a set of characters from several languages around the world. The involved fonts were downloaded from a manually-defined list of URLs, the downloaded fonts were then filtered by a python program in order to filter corrupted fonts. Several distortions and noises were involved, including affine and perspective transformations, random elastic transformations, natural background, foreground text filling, etc.

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

The data is synthesized by a data synthesizer OmniPrint. The sampling is uniformly random in the given transformation parameter space.

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 is synthesized by a computer software. However the design and implementation of the software, the choice of characters and fonts involve the authors of OmniPrint [1].

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 datasets were synthesized on June 24, 2021.

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.

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.

No preprocessing/cleaning/labeling was performed. The datasets are made available as they were synthesized. No feature extraction or removal of instances was done.

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.

N/A

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

Yes, the preprocessing software is available in the Meta-Album Github repository and OmniPrint repository (`https://github.com/SunHaozhe/OmniPrint`). Details are provided on the Meta-Album website: `https://meta-album.github.io/`.

Any other comments?

Uses

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

No.

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.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (`https://meta-album.github.io/`). This website will also be used to announce any necessary information related to the dataset.

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

Besides few-shot learning classification tasks, this datasets can be used for classification tasks of a large number of characters, and for domain adaptation tasks. Furthermore, as the meta-data can serve as labels, other kinds of classification or regression problems can also be considered e.g. classification of fonts, classification of languages, regression of rotation angle, regression of horizontal shear, etc. Finally, the datasets can be used to study disentangling the label (class character) from the nuisance variables (font, style, distortions).

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?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

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

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

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 made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

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 after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

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.

The dataset is public for research and is released with its original license : CC BY 4.0 (https://creativecommons.org/licenses/by-nc/4.0/). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

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 Meta-Album paper will be responsible for supporting the dataset.

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

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

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

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

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 no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

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.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

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.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

References

[1] H. Sun, W.-W. Tu, and I. M. Guyon. "OmniPrint: A Configurable Printed Character Synthesizer". In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. 2021.