Datasheet for STYLE-TRANS-FAIR 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 to serve as one of the meta-learning datasets for the “Creation of AI Challenge class” in T3 (2022-2023) and will be developed further for international competitions. The dataset is intended (but not limited to) to be used for the classical N-way-K-shot few-shots classification with controlled bias, so that the contestants would find a way to overcome with bias-invariant solutions.

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

Khuong Thanh Gia Hieu and Ahmad Nasser created the dataset, under the supervision of Professor Isabelle Guyon. The original datasets are created by Ihsan Ullah. The work was performed at Université Paris-Saclay, France, as part of the HUMANIA project. ChaLearn also supported the creation of the dataset.

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.

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 512×512 RGB artificially created images by performing Neural Style Transfer on Meta Album Dataset. Each task in the dataset includes 9 groups which is the combination of 3 classes and 3 styles. The intended use of the Dataset is to undersample 6 out of 9 groups to create a bias in styles.

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

The dataset consists of 5 tasks with 3 classes/task. There are 120 instances/images per class. Total count of instances is 5 x 3 x 120 = 1,800.

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 is sampled from a large insects dataset which consists of 290,000 images. The extracted dataset is representative in terms of classes, i.e. it represents all classes in original dataset which have at least 40 instances per class.

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 instances is 512x512 RGB image. The instances are preprocessed i.e. resized into 512x512 with anti-aliasing filter.

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

Yes, each instance has a category and style which is provided with the images in meta-data. The category describes the class of the instance. The categories are acquired from Meta Album Dataset.
The style describes the class of the style image in the Neural Style Transfer procedure.

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 information 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 styles.

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 should be splited 50-50 with 20 images per class for training and 20 images per class for testing.

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

No, there are no suspected errors, sources of noise or redundancies.

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 linked to Meta Album.

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. However it will be publicly released after the challenge on Codabench platform (https://www.codabench.org/).

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

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 stylized data(images) is artificially created by the Neural Style Transfer procedure. The original images are taken from Meta Album. The style images are scrapped from the internet.

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 is sampled from a large dataset with the strategy that all categories/classes are taken which have at least 40 instances per category. The categories which have more than 40 instances per category, 40 instances are selected randomly.

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 HUMANIA team were involved in the data collection process.

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 was generated between October 2022 and April 2023.

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.

Yes, the data is preprocessed i.e the images are resized into 512x512 with anti-aliasing filter.

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 sampled raw data is included as reference.

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

Yes, the software of pre-processing is available in the Meta-Album Github repository (https://github.com/ktgiahieu/TER_StyleTransFair).

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, the github repository (https://github.com/ktgiahieu/TER_StyleTransFair) will be active once the dataset is publicly released, it 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.

{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?

The dataset can be used without further considerations.

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

Not that we know of.

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 via the Codabench Platform (https://www.codabench.org/)

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 via Codabench Platform (https://www.codabench.org/) after the international challenge. The access information and any necessary updates will be announced via the github repository (https://github.com/ktgiahieu/TER_StyleTransFair). 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 international challenge.

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 will be distributed via the Codabench Platform (https://www.codabench.org/). It will be licensed under a Creative Commons license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). This comes with the following guarantee disclaimer: Unless otherwise separately undertaken by the Licensor, to the extent possible, the Licensor offers the Licensed Material as-is and as-available, and makes no representations or warranties of any kind concerning the Licensed Material, whether express, implied, statutory, or other. This includes, without limitation, warranties of title, merchantability, fitness for a particular purpose, non-infringement, absence of latent or other defects, accuracy, or the presence or absence of errors, whether or not known or discoverable. Where disclaimers of warranties are not allowed in full or in part, this disclaimer may not apply to You. To the extent possible, in no event will the Licensor be liable to You on any legal theory (including, without limitation, negligence) or otherwise for any direct, special, indirect, incidental, consequential, punitive, exemplary, or other losses, costs, expenses, or damages arising out of this Public License or use of the Licensed Material, even if the Licensor has been advised of the possibility of such losses, costs, expenses, or damages. Where a limitation of liability is not allowed in full or in part, this limitation may not apply to You.

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 this 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 (https://github.com/ktgiahieu/TER_StyleTransFair). In case of emergency, the authors of this 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 github repository (https://github.com/ktgiahieu/TER_StyleTransFair).

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 intentions to update the dataset unless required. In any case updates will be available at the github repository (https://github.com/ktgiahieu/TER_StyleTransFair).

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 github repository (https://github.com/ktgiahieu/TER_StyleTransFair).

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 this paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (https://github.com/ktgiahieu/TER_StyleTransFair). Newly produced datasets can be validated on the Codabench Platform (https://www.codabench.org/). All updates will be available at the given repo and the authors can be contacted via email: meta-album@chalearn.org

Any other comments?