A Datasheets for FLAIR Dataset

A.1 Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests. The latter may be particularly relevant for datasets created for research purposes.

- **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.
  FLAIR dataset was created for the purpose of providing the community a benchmark in the vision domain to accelerate federated learning research. FLAIR is suitable for multi-label image classification tasks, where the input is an image and output is a set of objects presented in the image.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
  Apple ML privacy team and ML research team created dataset on behalf of Apple Inc.
- **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.
  Apple Inc.

A.2 Composition

Dataset creators should read through these questions prior to any data collection and then provide answers once data collection is complete. Most of the questions in this section are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for their chosen tasks. Some of the questions are designed to elicit information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

Questions that apply only to datasets that relate to people are grouped together at the end of the section. We recommend taking a broad interpretation of whether a dataset relates to people. For example, any dataset containing text that was written by people relates to people.

- **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 Flickr images with annotation and metadata.
- **How many instances are there in total (of each type, if appropriate)?**
  There are 429,078 images in total.
- **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 instances in FLAIR dataset is a subset of the larger set, which is all Flickr images. Not all Flickr images are suitable for research use, i.e. images with personal identifiable information and images without permissive license were excluded.
- **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 consists of an image.
- **Is there a label or target associated with each instance?** If so, please provide a description.
  Each image has two sets of annotated labels from two taxonomies. Each image also has the associated Flickr user ID and image ID.
- **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.

- **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.
Yes, individual images from the same Flickr user have the same Flickr user ID.
- **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.
Yes. FLAIR data is partitioned based on Flickr user IDs, such that the data of a particular user is present in only one of three splits. Out of 51,414 Flickr users, 80% are in the training set, 10% in the validation set and 10% in the test set. There are 345,879 images in total in the training set, 39,239 in the validation set and 43,960 in the test set.
- **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.
N/A.
- **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 dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
FLAIR 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.
No.
- **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. Images with offensive and other inappropriate materials have been removed from FLAIR.

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

- **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.
FLAIR data is only annotated with the Flickr user id and does not explicitly identify any traits.
- **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. Images with personal identifiable information have been removed from FLAIR.
- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race 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.

A.3 Collection Process

As with the questions in the previous section, dataset creators should read through these questions prior to any data collection to flag potential issues and then provide answers once collection is complete. In addition to the goals outlined in the previous section, the questions in this section are designed to elicit information that may help researchers and practitioners to create alternative datasets

with similar characteristics. Again, questions that apply only to datasets that relate to people are grouped together at the end of the section.

- **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 the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
  Images and associated image ID and user ID were acquired from the Flickr website. The data was directly observable.
- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** How were these mechanisms or procedures validated?
  Software API provided by Flickr.
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
  N/A.
- **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)?**
  N/A.
- **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 images were collected from late 2017 to early 2018.
- **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.

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

- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
  The data was collected from the Flickr website.
- **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.
  Each image has an associated license chosen by the Flickr user. FLAIR only contain images with one of the following permissive licenses:
  - Attribution 2.0 Generic (CC BY 2.0)[4]
  - Attribution-ShareAlike 2.0 Generic (CC BY-SA 2.0)[5]
  - Attribution-NoDerivs 2.0 Generic (CC BY-ND 2.0)[6]
  - U.S. Government Works[7]
  - CC0 1.0 Universal (CC0 1.0) Public Domain Dedication[8]

[4] https://creativecommons.org/licenses/by/2.0/
[5] https://creativecommons.org/licenses/by-sa/2.0/
[6] https://creativecommons.org/licenses/by-nd/2.0/
[7] http://www.usa.gov/copyright.shtml
[8] https://creativecommons.org/publicdomain/zero/1.0/

  - Public Domain Mark 1.0[9]

- **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?**
  After initial collection, we applied a two-stage filtering approach to remove images with personal identifiable information and sensitive materials. In the first stage, we used a face detector to automatically remove images with faces. In the second stage, we asked human annotators to filter out images with identifiable human and sensitive materials. Specifically, images with any of the following will be removed from FLAIR:
  - Visible faces or part of visible faces.
  - Visible facial features or part of visible facial features, such as hair, eye, eyebrow, mouth, nose, ear, etc.
  - Human body or part of body has identifiable feature, such as tatto, disabilities, injuries, scars, birthmarks, unique moles, etc.
  - Rude statements and expressions.
  - Profanity, racial, gender, ethnic, or religious slurs.
  - Sexually explicit or pornographic materials.
  - Violent, obscene, graphic or disturbing materials.

A.4 Preprocessing/cleaning/labeling

Dataset creators should read through these questions prior to any preprocessing, cleaning, or labeling and then provide answers once these tasks are complete. The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

- **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 remaining questions in this section.
  Yes. Labeling was done by human annotators where one annotator labeled the objects presented in an image and another annotator validate the labeling. The taxonomy of the labels were constructed as following:
  1. Retrieve all keywords from ShutterStock[10] attached to 1000 images or more.
  2. Remove keywords that are illicit substances, sexual content, negative connotations, adjectives, proper names, places, organizations, occupations, abstract concepts, references to ethnicity, culture, religion, skin color, all body parts, and most animal parts.
  3. Remove plurals, alternative spellings and synonyms.
  4. Leverage WordNet[11] to construct coarse-grained labels.

  Unqualified images are removed as described in Appendix A.3.
- **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.
  No.

[9] https://creativecommons.org/publicdomain/mark/1.0/
[10] https://www.shutterstock.com/
[11] https://wordnet.princeton.edu/

- **Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.
  The script to process data for training is provided at https://github.com/apple/ml-flair.

A.5 Uses

The questions in this section are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

- **Has the dataset been used for any tasks already?** If so, please provide a description.
  FLAIR has been used to benchmark federated learning and differential privacy on multi-label classification task, in this current paper.
- **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 current paper and the code used for experiments are available at https://github.com/apple/ml-flair
- **What (other) tasks could the dataset be used for?**
  FLAIR could be used for other image 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 dataset consumer 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 risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
  This dataset contains a limited number of object classes and is intended to create a benchmark to evaluate and compare algorithms for (private) federated learning.
- **Are there tasks for which the dataset should not be used?** If so, please provide a description.
  It being a subset of images from Flickr, it is not expected to be representative of all images in the world.

A.6 Distribution

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

- **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.
- **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 distributed on AWS S3.
- **When will the dataset be distributed?**
  The dataset will be distributed on June 16th, 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.
  Please see license for FLAIR at https://github.com/apple/ml-flair/blob/master/LICENSE.md
- **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.
N/A.

A.7 Maintenance

As with the questions in the previous section, dataset creators should provide answers to these questions prior to distributing the dataset. The questions in this section are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan to dataset consumers.

- **Who will be supporting/hosting/maintaining the dataset?**
Apple ML Privacy team.
- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
pfl-dev@group.apple.com
- **Is there an erratum?** If so, please provide a link or other access point.
N/A.
- **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 dataset consumers (e.g., mailing list, GitHub)?
N/A.
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the 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 dataset consumers.
N/A.
- **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 dataset consumers? If so, please provide a description.
N/A.
- **Any other comments?**
The annotations and Apple's other rights in the dataset are licensed under CC-BY-NC 4.0 license. The images are copyright of the respective owners, the license terms of which can be found using the links provided in https://github.com/apple/ml-flair/blob/master/ATTRIBUTION.txt (by matching the Image ID). Apple makes no representations or warranties regarding the license status of each image and you should verify the license for each image yourself.

B Benchmark Setup Details

B.1 Computational resources

All experiments are conducted on a cluster with 32 CPU cores and 4 NVIDIA Tesla V100 GPUs.

B.2 Hyper-parameters grids

Below are the hyper-parameter grids that we searched on for benchmarking FLAIR:

- Server learning rate $\in \{0.01, 0.02, 0.05, 0.1\}$
- Server number of rounds $\in \{2000, 5000\}$
- Client local learning rate $\in \{0.01, 0.1\}$
- Client number of epochs $\in \{1, 2, 3, 4, 5\}$
- Target unclipped quantile for adaptive clipping $\in \{0.1, 0.2\}$

B.3 Additional benchmark results

Table 2: FLAIR binary classification benchmark results on test set for *structure* label. AP stands for averaged precision. All experiments are run for 5 times with different random seed, and both mean and standard deviation of metrics are reported.

Setting	Initialization	AP	Precision	Recall	F1
Central	Random	$87.6_{\pm 0.2}$	$78.9_{\pm 0.6}$	$79.0_{\pm 0.9}$	$78.9_{\pm 0.3}$
Federated	Random	$84.5_{\pm 0.3}$	$75.8_{\pm 0.3}$	$77.4_{\pm 1.4}$	$76.6_{\pm 0.6}$
Private Federated	Random	$67.3_{\pm 1.2}$	$63.7_{\pm 0.5}$	$74.9_{\pm 2.0}$	$68.8_{\pm 0.9}$
Central	ImageNet	$92.8_{\pm 0.0}$	$85.1_{\pm 0.5}$	$83.5_{\pm 0.8}$	$84.3_{\pm 0.1}$
Federated	ImageNet	$90.5_{\pm 0.2}$	$82.3_{\pm 0.9}$	$81.1_{\pm 1.1}$	$81.7_{\pm 0.1}$
Private Federated	ImageNet	$84.1_{\pm 0.5}$	$77.5_{\pm 0.7}$	$73.3_{\pm 1.1}$	$75.3_{\pm 0.5}$

We provide additional binary classification benchmark on the most common *structure* label, using the same hyperparameters as in Section 5.1. Table 2 summarizes the results. The performance of models trained in federated setting and private federated setting are much closer to the centralized setting, especially when the models were pretrained on ImageNet. We believe that this simple binary classification baseline could help researchers to quickly verify their proposed algorithms and methods in (private) federated learning setting.