Provenance

Dataset name: Name of the dataset under consideration
Dataset version: Which particular version the information pertains to, if applicable
Origin date: The time when a dataset first came about, e.g., the time of publication of CIFAR-10's technical report
Origin: The place where a dataset first came about, e.g., CIFAR-10's official website
Description of dataset: Brief overview given by the creators of the dataset, e.g., "The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images."
Description of the data collection process: Details of how the data were collected, as reported by the dataset creators, e.g., "CIFAR-10 and CIFAR-100 are labeled subsets of the 80 million tiny images dataset. They were collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton."
Downloaded outlet: If the dataset was downloaded from a different outlet than origin, we record the location
Is outlet licensed?: Whether the downloaded outlet is licensed
Is the dataset publicly available?: Whether the dataset is available to the public or is behind a paywall
Where license was found: Short description of where license was found, e.g., on official website, with the downloaded dataset, etc.
License location: URL or relative path of the license
License content: Record the contents of the license
Hash code: Any hash code(s) supplied by the dataset creators
Size: Size of the dataset download
Format: File type of the dataset download
Additional notes: Any further details that are relevant to this dataset's provenance

Lineage

Dataset/data source name: Name of the dataset or data source under consideration
Is there another data source?: Whether this dataset or data source has a data source
License: Hyperlink(s) to any relevant license information (e.g., information provided on official dataset website, Terms of Use, Creative Commons, etc.)
License range: The time period during which license information is collected; calculated to be the year preceding the origin date

Enhanced MDL

Dataset/data source name: Name of the dataset/data source under consideration
Dataset version: Which particular version the information pertains to, if applicable
Licensor: The owner of the dataset/data source who assigned the license
License name: Name of the license if it is a standard license (e.g., CC-SA); fill in “custom” otherwise
Credit/Attribution notice: When using the dataset if credit/attribution needs to be given, who should it be given to
License validity period: Records the expiry date of the license

Liability/Warranty: Records the warranty/liability specified in the license (we specify N/A if there are no warranties specified in the license)
Designated third parties: The designated third parties that the license specifies as entities that can modify the license (N/A if none specified)
Access rights (data): Right to view or download the data for the purpose of evaluation
Access obligations (data): Obligations to fulfill when viewing or downloading the data for the purpose of evaluation
Tagging rights (data): Right to add tags or labels to the data (or any portion thereof)
Tagging obligations (data): Obligations to fulfill when adding tags or labels to the data (or any portion thereof)
Distribute rights (data): Right to make the data (or any portion thereof) available to third parties
Distribute obligations (data): Obligations to fulfill when making the data (or any portion thereof) available to third parties
Re-represent rights (data): Right to alter the data such that there exists an alternative version that can used in lieu of the original
Re-represent obligations (data): Obligations to fulfill when altering the data such that there exists an alternative version that can used in lieu of the original
Benchmark rights (data with model): Right to use the data to assess the performance of different untrained models; the trained model, however, can only be used to demonstrate the results of the training
Benchmark obligations (data with model): Obligations to fulfill when using the data to assess the performance of different untrained models; the trained model, however, can only be used to demonstrate the results of the training
Research rights (data with model): Right to use the data to train a model; the trained model and its output, however, are restricted to research use only
Research obligations (data with model): Obligations to fulfill when using the data to train a model; the trained model and its output, however, are restricted to research use only
Publish rights (data with model): Right to train a model that is available to third parties for research or publication only
Publish obligations (data with model): Obligations to fulfill when training a model that is available to third parties for research or publication only
Internal use rights (data with model): Right to use the data to train a model and improve its output; however, neither the trained model nor its output can be commercialized or made available for third party use or benefit
Internal use obligations (data with model): Obligations to fulfill when using the data to train a model and improve its output; however, neither the trained model nor its output can be commercialized or made available for third party use or benefit
Output commercialization rights (data with model): Right to use the data to train a model and improve its output. The output can then be commercialized and made available for third party use or benefit; however, the trained model cannot be made available
Output commercialization obligations (data with model): Obligations to fulfill when using the data to train a model and improve its output. The output can then be commercialized and made available for third party use or benefit; however, the trained model cannot be made available

Model commercialization rights (data with model): Right to make a trained model or any product or service utilizing a trained model available to third parties. The model's output may or may not be available to the same third party
Model commercialization obligations (data with model): Obligations to fulfill when making a trained model or any product or service utilizing a trained model available to third parties. The model's output may or may not be available to the same third party
Model reverse engineer rights (data with model): Right to reverse engineer a trained model to reconstruct the dataset used for its training
Model reverse engineer obligations (data with model): Obligations to fulfill when reverse engineering a trained model to reconstruct the dataset used for its training
Additional conditions: Records any additional conditions if specified in the license of the given dataset