Title: The Pile (dataset)

URL: https://en.wikipedia.org/wiki/The_Pile_(dataset)

PageID: 73281585

Categories: Category:Datasets in machine learning, Category:Large language models, Category:Statistical data sets

---

The Pile is an 886.03 GB diverse, open-source dataset of English text created as a training dataset for large language models (LLMs). It was constructed by EleutherAI in 2020 and publicly released on December 31 of that year. [1] [2] It is composed of 22 smaller datasets, including 14 new ones. [1]

Creation

Training LLMs requires sufficiently vast amounts of data that, before the introduction of the Pile, most data used for training LLMs was taken from the Common Crawl . [3] However, LLMs trained on more diverse datasets are better able to handle a wider range of situations after training. [4] The creation of the Pile was motivated by the need for a large enough dataset that contained data from a wide variety of sources and styles of writing. [1] [5] Compared to other datasets, the Pile's main distinguishing features are that it is a curated selection of data chosen by researchers at EleutherAI to contain information they thought language models should learn and that it is the only such dataset that is thoroughly documented by the researchers who developed it. [6]

Contents and filtering

Artificial intelligences do not learn all they can from data on the first pass, so it is common practice to train an AI on the same data more than once with each pass through the entire dataset referred to as an "epoch". [7] Each of the 22 sub-datasets that make up the Pile was assigned a different number of epochs according to the perceived quality of the data. [1] The table below shows the relative size of each of the 22 sub-datasets before and after being multiplied by the number of epochs. Numbers have been converted to GB , and asterisks are used to indicate the newly introduced datasets.

EleutherAI chose the datasets to try to cover a wide range of topics and styles of writing, including academic writing, which models trained on other datasets were found to struggle with. [1]

All data used in the Pile was taken from publicly accessible sources. EleutherAI then filtered the dataset as a whole to remove duplicates. Some sub-datasets were also filtered for quality control. Most notably, the Pile-CC is a modified version of the Common Crawl in which the data was filtered to remove parts that are not text, such as HTML formatting and links. [1]

Some potential sub-datasets were excluded for various reasons, such as the US Congressional Record , which was excluded due to its racist content. [1]

Within the sub-datasets that were included, individual documents were not filtered to remove non-English, biased, or profane text. It was also not filtered on the basis of consent, meaning that, for example, the Pile-CC has all of the same ethical issues as the Common Crawl itself. However, EleutherAI has documented the amount of bias (on the basis of gender, religion, and race) and profanity as well as the level of consent given for each of the sub-datasets, allowing an ethics-concerned researcher to use only those parts of the Pile that meet their own standards. [1]

Use

The Pile was originally developed to train EleutherAI's GPT-Neo models [8] [9] [10] but has become widely used to train other models, including Microsoft 's Megatron-Turing Natural Language Generation, [11] [12] Meta AI 's Open

Pre-trained Transformers, [13] LLaMA , [14] and Galactica, [15] Stanford University 's BioMedLM 2.7B, [16] the Beijing Academy of Artificial Intelligence 's

Chinese-Transformer-XL, [17] Yandex 's YaLM 100B, [18] and Apple 's OpenELM. [19]

In addition to being used as a training dataset, the Pile can also be used as a benchmark to test models and score how well they perform on a variety of writing styles. [2] [20] [21]

DMCA takedown

The Books3 component of the dataset contains copyrighted material compiled from Bibliotik, a pirate website. [22] In July 2023, the Rights Alliance took copies of The Pile down through DMCA notices. [23] [24] Users responded by creating copies of The Pile with the offending content removed. [25]

See also

List of chatbots

List of datasets for machine-learning research

References