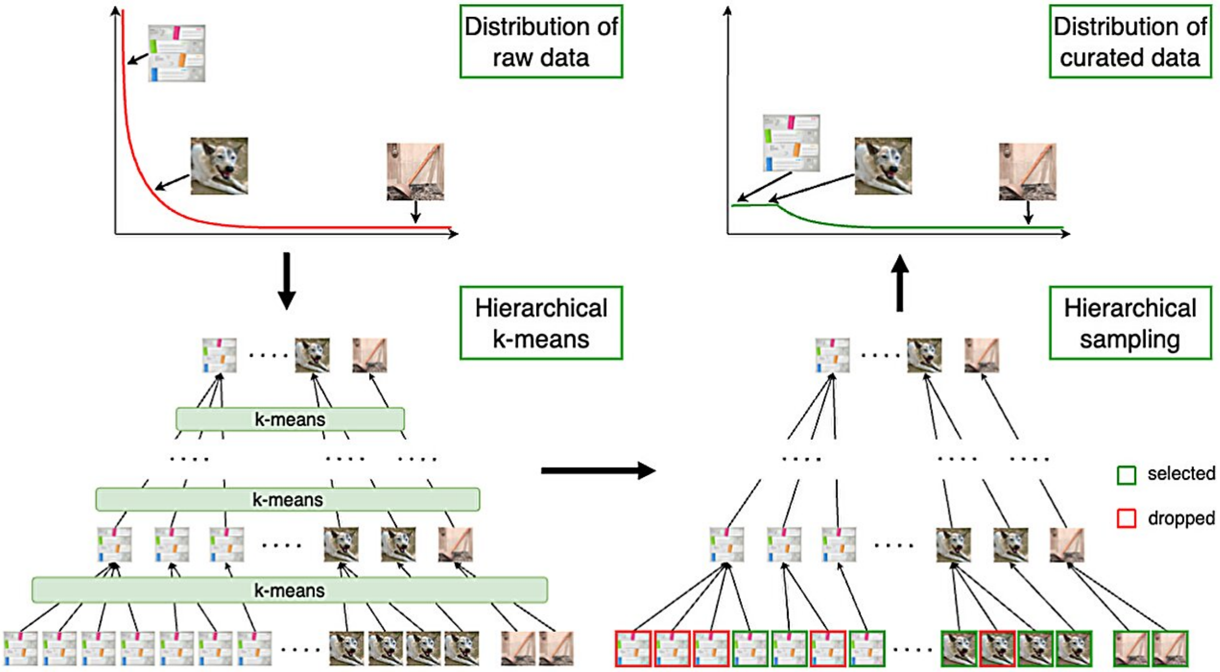
TechXplore

New technique can automate data curation for self-supervised pre-training of AI datasets

June 3 2024, by Bob Yirka



An overview of the data curation pipeline. Large data pool often exhibits a long-tailed distribution of concepts. We apply hierarchical k-means to obtain clusters that spread uniformly over the concepts. Data points are then sampled from the clusters to form a curated dataset that has a better balance of concepts. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2405.15613

A team of computer scientists and AI researchers from FAIR at Meta, INRIA, Université Paris Saclay and Google, has developed a possible means for automating data curation for self-supervised pre-training of AI datasets.

The group has written a paper describing their development process, the technique they developed and how well it has worked thus far during testing. It is posted on the *arXiv* preprint server.

As developers and users alike have been learning over the past year, the quality of the data that is used to train AI systems is tied very closely to the accuracy of results. Currently, the best results are obtained with systems that use manually curated data and the worst are obtained from systems that are uncurated.

Unfortunately, manually curating data takes a lot of time and effort. Therefore, computer scientists have been looking for ways to automate the process. In this new study, the research team has developed a technique that does just that, and that does it in a way that is on a par with manual curation.

The new technique starts with a large dataset, and then carries out a three-step process that results in data that is both more diverse and more balanced.

The first step involves using a feature-extraction model that calculates high-quality places to embed data points. In their approach, the things that are embedded are numbers that represent features of different types of data, such as text, audio, or images.

The second step involves the use of successive k-means clustering, where data points are assigned to a group based on their similarity to other data points.

The third step involves the use of multi-step hierarchical k-means clustering to ensure that data clusters are balanced. It is achieved via building data-cluster trees in a bottom-up fashion.

The research team tested their technique using vision models that had been trained on various types of datasets. They found that models using their technique outperformed those using uncurated data and were as good as or sometimes better than those trained on data that was curated manually.

More testing will have to be done to find out how well their technique works on real-world data and different kinds of AI systems.

More information: Huy V. Vo et al, Automatic Data Curation for Self-Supervised Learning: A Clustering-Based Approach, *arXiv* (2024). DOI: 10.48550/arxiv.2405.15613

Citation: New technique can automate data curation for self-supervised pre-training of AI datasets (2024, June 3) retrieved 21 June 2024 from https://techxplore.com/news/2024-06-technique-automate-curation-pre-ai.html