SELECT: A Large-Scale Benchmark of Data Curation Strategies for Image Classification

Benjamin Feuer[*1], **Jiawei Xu**[*1], **Niv Cohen**[1],
Patrick Yubeaton[1], **Govind Mittal**[1], **Chinmay Hegde**[1]

[1] NYU

Abstract

Data curation is the problem of how to collect and organize samples into a dataset that supports efficient learning. Despite the centrality of the task, little work has been devoted towards a large-scale, systematic comparison of various curation methods. In this work, we take steps towards a formal evaluation of data curation strategies and introduce SELECT , the first large-scale benchmark of curation strategies for image classification.

In order to generate baseline methods for the SELECT benchmark, we create a new dataset, IMAGENET++ , which constitutes the largest superset of ImageNet-1K to date. Our dataset extends ImageNet with 5 new training-data shifts, each approximately the size of ImageNet-1K itself, and each assembled using a distinct curation strategy. We evaluate our data curation baselines in two ways: (i) using each training-data shift to train identical image classification models from scratch (ii) using it to inspect a fixed pretrained self-supervised representation.

Our findings show interesting trends, particularly pertaining to recent methods for data curation such as synthetic data generation and lookup based on CLIP embeddings. We show that although these strategies are highly competitive for certain tasks, the curation strategy used to assemble the original ImageNet-1K dataset remains the gold standard. We anticipate that our benchmark can illuminate the path for new methods to further reduce the gap. We release our checkpoints, code, documentation, and a link to our dataset at https://github.com/jimmyxu123/SELECT.

1 Introduction

Data curation is the process of collecting and organizing a corpus of data into a dataset that supports efficient learning. Until recently, data curation was an implicit consideration in most of the academic discourse on machine learning, and the vast majority of research works were oriented towards introducing novel methods, theories, or architectures.

However, data curation has begun to gain prominence as a research topic in its own right; several recent works have contended that labeling errors pervade commonly used benchmark datasets, with error rate estimates varying from 3% to 50% on the most popular ones [4, 20, 26, 30]. Group imbalances are often inadvertently introduced during the curation process, biasing model predictions [9, 19]. The work of [31] created a standard, now widely adopted, for reporting on the process for creating new datasets. Unfortunately, despite growing attention of the centrality of the data curation problem to model performance, many works in the literature do not adhere to best practices, reporting very little about the data on which they are trained, or how that data was curated [5, 22, 32]. To address this,

[*]First two authors contributed equally. Correspondence to: bf996@nyu.edu.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

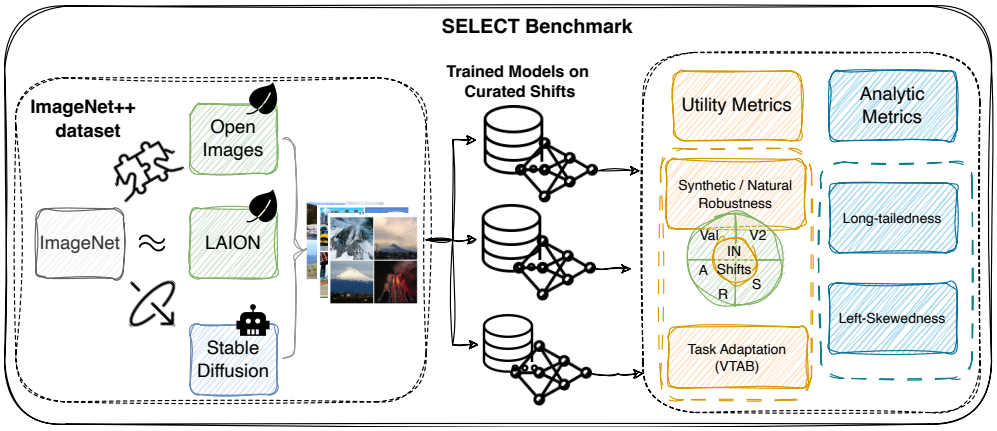

Figure 1: **Overview of SELECT Benchmark.** (**Left**) The ImageNet++ dataset is composed of different shifts of the ImageNet train set. The shifts were generated using different curation strategies and drawn from diverse data sources including OpenImages (natural images), LAION (natural images), and Stable Diffusion (synthetic images). (**Right**) We trained identical models on the sets collected using the different strategies (producing different 'shifts'), and evaluated them in two ways: (i) *Utility metrics*: quantifying the models ability to predict different in-distribution and out-of-distribution test sets, and (ii) *Analytic metrics*: examining various statistics of the distribution of the samples among the various classes.

[13] in NeurIPS 2023 introduced the DataComp competition, where the model architecture and loss (following CLIP [32]) were fixed and the challenge was to filter (subsample) a large pool of images to find high-performant sets of image samples for a suite of zero-shot tasks.

Our goal in this paper is to bring the implicitly studied subject of data curation into sharper focus, broaden the scope of curation beyond data filtration, and introduce it as a topic of research in its own right. In Sec. 2, we use rational choice theory to formalize any data curation strategy as a utility function, where an increment to the marginal cost produces an expected gain in utility. In Sec. 3, we introduce SELECT , a benchmark that serves as a diverse measure of utility of data curation methods in the domain of image classification. In Sec. 4, we introduce IMAGENET++ , which we leverage to produce a large-scale set of baselines for data curation, composed of 5 new training-data shifts of ImageNet-1K. Finally, in Sec. 5, we compare our IMAGENET++ baselines on the SELECT benchmark and derive several useful insights. Specifically, our contributions are as follows.

1. We introduce SELECT , a diverse benchmark for data curation methods for computer vision (in particular, image classification).

2. We introduce IMAGENET++ , the largest, most diverse set of distribution shifts for ImageNet-train to date [10]. This serves as a rich source of data curation baselines on which we train over 130 models (Fig. 2).

3. We analyze our baseline models and derive several novel insights:

   (a) On certain metrics in SELECT (pretraining and fine-tuning), reduced-cost curation methods perform as well as expert-labeled data.
   (b) However, on most metrics, expert labeling continues to outperform the alternatives.
   (c) Image-to-image curation methods generally outperform those which rely on text.
   (d) Both label noise and label imbalance remain important limiting factors on the utility of cost-efficient data-curation.

In order to enable future research and reproducibility, we release our code, our dataset, and a complete enumeration of our results for all models in the study (see supplemental attachments).