M Availability

Our entire dataset, including Croissant metadata record and our trained model checkpoints, are currently available on HuggingFace. All shifts are made available in WebDataset or HuggingFace Datasets format. The links can be accessed at our GitHub repository, https://github.com/jimmyxu123/SELECT. Our hosting and maintenance plan is to preserve the work via the HuggingFace repository, which has proven to be a reliable exchange for large datasets in recent years.

N Not safe for work (NSFW) filtering

The images included in ImageNet++ are sourced from the LAION-5B dataset ([36]), the OpenImages dataset ([25]), and synthetic img2img inversion transformations from the ImageNet-1k dataset. Although these datasets are generally regarded as safe and publicly available, we employ a variety of NSFW content filtering techniques to identify and exclude any potentially problematic images and captions.

Firstly, we filter captions using Detoxify ([17]), a robust language model designed to detect toxic comments. Specifically, we employ the multilingual XLM-roBERTa ([9]) variant. This model generates scores ranging from zero to one for the following categories: toxicity, severe toxicity, obscenity, identity attack, insult, threat, and sexually explicit content. Based on the prior work in image filtering by DataComp ([14]), we heuristically set a threshold of 0.1. This threshold effectively filters NSFW text while minimizing false positives. If any of the Detoxify category scores exceed this threshold, the sample is discarded. Next, we apply a filtering process to the visual data. We utilize a modified version of LAION-5B's CLIP-based binary NSFW classification model by [36], which employs CLIP ViT-L/14 visual embeddings as input. Further information about the training data is provided in Appendix C.5 of the LAION-5B paper. In summary, the dataset comprises 682,000 images, with a roughly equal distribution between Safe for Work (SFW) and NSFW categories.

After applying this filtering to the three subsets of ImageNet++, no toxic images were found, indicating that the dataset's captions are safe. However, after applying this filtering to the three subsets of ImageNet++, no toxic images were found, indicating that the dataset's captions are safe. This result isn't surprising given that the source data has been previously vetted by machine or human experts.

O Datasheet

Motivation

For what purpose was the dataset created?
ImageNet++ aims to facilitate the training of models robust against natural distribution shifts, efficiently utilizing data. Including three datasets, OI1000, Laion-1k, and SD1000, each introducing natural distribution shifts relative to ImageNet-1k, it is the largest and most diverse superset of ImageNet-1k. Moreover, we use ImageNet++ to derive novel insights into scaling factors in this paper.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The dataset was created by researchers in the DICE Lab at New York University.

Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?
The dataset was used for experiments in this paper.

What (other) tasks could the dataset be used for?

The dataset could also be used for model pretraining. The method could also be applied to generate the same-size shifts to other datasets.

Any other comments? None.

Dataset Composition

What do the instances that comprise the dataset represent?

ImageNet++ consists of 5 distinct datasets, each representing a variation of the ImageNet-1k dataset:
1.OpenImages-1000(OI1000): A subset of the Open Image dataset[25], where samples are aligned with ImageNet-1k class names based on human-labeled annotations.
2.Laion-1000(LAION1000): A subset of the unlabeled LAION dataset[36], selected through nearest neighbors search against the ImageNet-1k training set.
3.Stable Diffusion-1000(SD1000): A set generated from the ImageNet-1k dataset using Stable Diffusion, where images are transformed via an inversion process.

How many instances are there in total?

See Table 6 for reference of our dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances?

Instances in OI1000 and LAION1000 are images each associated with labels and captions. SD1000 contains AI-generated features based on the images from ImageNet-1k, also with associated labels. All the included data are filtered for NSFW content (see Appendix N)

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. There is no missing information in the dataset.

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).

Instances in OI1000 and LAION1000 are raw images, while SD1000 comprises AI-generated features derived from ImageNet-1k images. All instances are labeled. The datasets, particularly OI1000 and LAION1000, are subsets of larger sets and are intentionally curated to introduce specific feature shifts relative to ImageNet-1k, rather than to serve as comprehensive representations of their parent datasets.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

There are no known errors, noise, or redundancies in the dataset.

Any other comments?

None.

Collection Process

What mechanisms or procedures were used to collect the data? (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)
All the data of OI1000 and Laion-1k are collected from larger public sets. Data in SD1000 is generated by AI.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
1.OI1000 (OpenImages-1000): The sampling strategy was deterministic, based on a direct mapping of human-labeled class names to the corresponding classes in ImageNet-1k.
2.LAION1000: The sampling was semi-probabilistic. Samples were selected using a nearest neighbors search based on the ImageNet-1k training set. While this approach is guided by the proximity of LAION images to the ImageNet-1k feature space, it inherently introduces a probabilistic element due to the variability in nearest-neighbor results.
3.SD1000 (Stable Diffusion-1000): This subset encompasses all possible instances generated from the ImageNet-1k dataset using Stable Diffusion, hence it's not a sample but a complete set derived from the original dataset through a generative process.

Who was involved in the data collection process (e.g., students, crowd workers, contractors), and how were they compensated (e.g., how much were crowd workers paid)?
The creation of ImageNet++ is done by the author of this work.

Over what timeframe was the data collected?
The timeframe for creating the ImageNet++ is from 12/2023 to 1/2024.

Any other comments? None.

Data Preprocessing

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 remainder of the questions in this section.
As our images are collected either from public data sources or synthetic generation, we did an NSFW filtering on all the images and the captions (see Appendix N).

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.
Yes, the "raw data" was also public.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
The details can be found in Appendix N.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?
We hope that the release of this benchmark suite will achieve our goal of accelerating research in models' robustness to natural shifts, as well as making it easier for researchers and practitioners to

generate data augmentations via our benchmark.

Any other comments? None.

Dataset Distribution

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?
The dataset will be public soon. All researchers and practitioners can access it if they are interested in the dataset.

How will the dataset be distributed (e.g., tarball on website, API, GitHub)?
We will publish all the format of the data.

When will the dataset be released/first distributed? What license (if any) is it distributed under?
The dataset is public as of 6/2024.

Are there any copyrights on the data?
There are no copyrights on the data.

Are there any fees or access/export restrictions?
There are no fees or restrictions.

Any other comments?
None.

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?
The authors of this work are supporting/hosting/maintaining the dataset.

Will the dataset be updated? If so, how often and by whom? We welcome updates from the community.

How will updates be communicated? (e.g., mailing list, GitHub)
Updates will be communicated by the mailing list of the authors.

If the dataset becomes obsolete how will this be communicated?
If the dataset becomes obsolete, it can be communicated by the mailing list of the authors.

If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions? What is the process for communicating/distributing these contributions to users?
Others can publish their extends/augmentation on the benchmark to any open-source website (eg. HuggingFace, Github, etc.)

Any other comments?
None.

Legal and Ethical Considerations

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.
There was no ethical review process. However, we did filtering for NSFW information before publishing the dataset.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.
All the data are either collected from public source or generated by AI. There is no confidential data.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
We did NSFW filtering to prevent this problem. As we believe, none of the data might be offensive, insulting, threatening, or otherwise cause anxiety.

Any other comments?
None.