Curator: A No-Code Self-Supervised Learning and Active Labeling Tool to Create Labeled Image Datasets from Petabyte-Scale Imagery

Rudy Venguswamy\textsuperscript{1,3}, Mike Levy\textsuperscript{1}, Anirudh Koul\textsuperscript{1,3,2}, Satyarth Praveen\textsuperscript{1}, Tarun Narayanan\textsuperscript{1}, Ajay Krishnan\textsuperscript{1}, Jenessa Peterson\textsuperscript{1}, Siddha Ganju\textsuperscript{1,2,4}, and Meher Kasam\textsuperscript{1,2,5}

\textsuperscript{1}SpaceML
\textsuperscript{2}Frontier Development Lab
\textsuperscript{3}Pinterest
\textsuperscript{4}Nvidia
\textsuperscript{5}Square

Machine learning modeling for Earth events at NASA is often limited by the availability of labeled examples. For example, training classifiers for forest fires or oil spills from satellite imagery requires curating a massive and diverse dataset of example forest fires, a tedious multi-month effort requiring careful review of over 196.9 million square miles of data per day for 20 years. While such images might exist in abundance within 40 petabytes of unlabeled satellite data, finding these positive examples to include in a training dataset for a machine learning model is extremely time-consuming and requires researchers to "hunt" for positive examples, like finding a needle in a haystack.

We present a no-code open-source tool, Curator, whose goal is to minimize the amount of human manual image labeling needed to achieve a state of the art classifier. The pipeline, purpose-built to take advantage of the massive amount of unlabeled images, consists of (1) self-supervision training to convert unlabeled images into meaningful representations, (2) search-by-example to collect a seed set of images, (3) human-in-the-loop active learning to iteratively ask for labels on uncertain examples and train on them.

In step 1, a model capable of representing unlabeled images meaningfully is trained with a self-supervised algorithm (like SimCLR) on a random subset of the dataset (that conforms to researchers' specified "training budget."). Since real-world datasets are often imbalanced leading to suboptimal models, the initial model is used to generate embeddings on the entire dataset. Then, images with equidistant embeddings are sampled. This iterative training and resampling strategy improves both balanced training data and models every iteration. In step 2, researchers supply an example image of interest, and the output embeddings generated from this image are used to find other images with embeddings near the reference image's embedding in euclidean space (hence similar looking images to the query image). These proposed candidate images contain a higher density of positive examples and are annotated manually as a seed set. In step 3, the seed labels are used to train a classifier to identify more candidate images for human
inspection with active learning. Each classification training loop, candidate images for labeling are sampled from the larger unlabeled dataset based on the images that the model is most uncertain about (p = 0.5).

Curator is released as an open-source package built on PyTorch-Lightning. The pipeline uses GPU-based transforms from the NVIDIA-Dali package for augmentation, leading to a 5-10x speed up in self-supervised training and is run from the command line.

By iteratively training a self-supervised model and a classifier in tandem with human manual annotation, this pipeline is able to unearth more positive examples from severely imbalanced datasets which were previously untrainable with self-supervision algorithms. In applications such as detecting wildfires, atmospheric dust, or turning outward with telescopic surveys, increasing the number of positive candidates presented to humans for manual inspection increases the efficacy of classifiers and multiplies the efficiency of researchers’ data curation efforts.