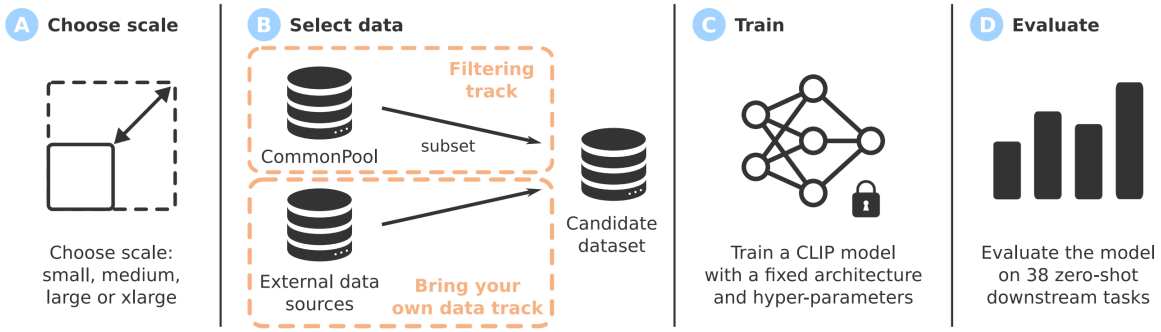
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

DataComp Challenge

The DataComp challenge presents a machine learning benchmark where the model and training procedure is fixed, and participants are tasked with optimizing the training dataset selection. Models trained on a smaller but higher quality subset can outperform models trained on the whole dataset. The challenge is to find filtering strategies with which such high-quality datasets can be created.



The DataComp challenge uses OpenCLIP as the model and training procedure and evaluates on 38 downstream datasets. The OpenCLIP model learns the similarity between image-text pairs from a web-scraped dataset. Contrastive Language-Image Pre-training (CLIP) can be applied to open vocabulary classification by computing the similarity for all images between the image and the class names and classifying the image as the most similar category.

In this project, you will explore how to filter the CommonPool dataset to create the ideal OpenCLIP training set.

Requirements: Strong motivation, knowledge in deep learning, or a solid background in machine learning. Previous experience with Python and libraries such as TensorFlow or PyTorch is an advantage. Reading the DataComp and CLIP papers before the first meeting is recommended. We will have weekly meetings to discuss open questions and determine the next steps.

Interested? Please contact us for more details!

Contact

- Benjamin Estermann: besterma@ethz.ch, ETZ G60.1
- Till Aczel: taczel@ethz.ch, ETZ G60.1