What Makes CLIP More Robust to Long-Tailed Pre-Training Data?

A Controlled Study for Transferable Insights

Xin Wen[1] Bingchen Zhao[2] Yilun Chen[3] Jiangmiao Pang[3] Xiaojuan Qi[1]

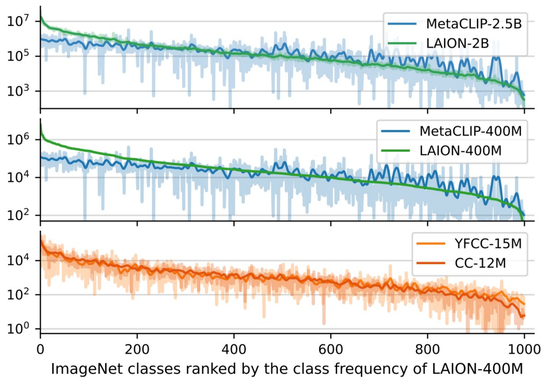
[1]The University of Hong Kong [2]University of Edinburgh [3]Shanghai AI Laboratory

Code

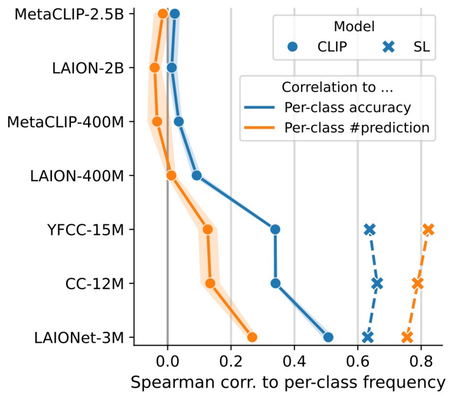
Contact

1. Motivation

Image-text datasets share a long-tailed class distribution



But CLIP is not much affected, especially when data scales up



2. Study on effective factors

Green dots: (Descriptive) language as supervision signal

Blue crosses: Dynamic classification (using subsampled vocabulary) as pretext task

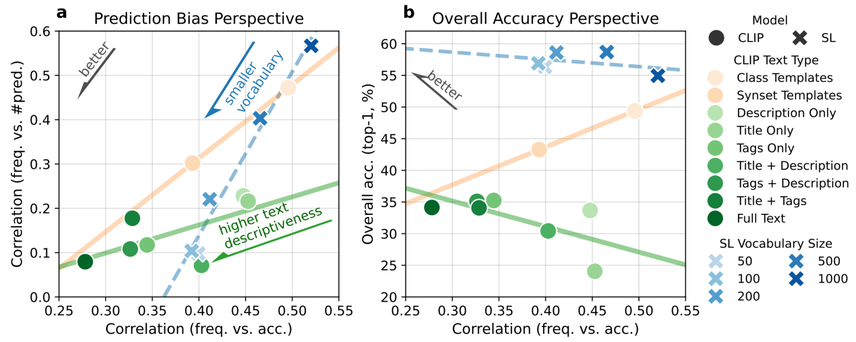
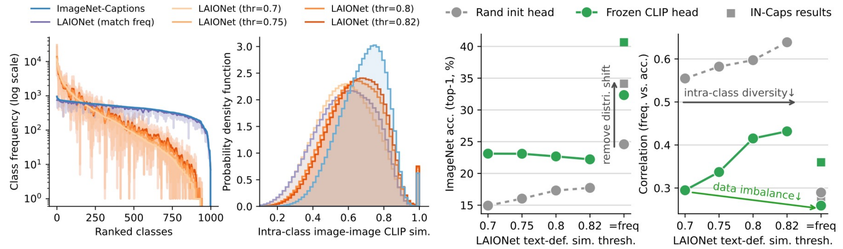


Figure 3: Results on IN-Caps about ● text descriptiveness and ✖ vocabulary size. 1) Increasing ● text descriptiveness improves both robustness (a) and discriminability (b) of CLIP, but the tendency varies if using ● less descriptive (template-based) supervision. 2) The gap between SL and CLIP (a) implies CLIP re-balances predictions, which is replicable by ✖ subsampling the vocabulary SL trains with.

Data factors (imbalance👎, diversity👍, distribution shift 👍)



(a) Distrib. of LAIONet variants (same scale as IN-Caps). (b) Results of CLIP trained on LAIONet variants.

Figure 4: Results on LAIONet about data distribution (level of data imbalance, distribution shift, and data diversity). 1) Extreme data imbalance makes models more prone to bias (last column *vs*. others). 2) Distribution shift (●● *vs*. ■■, last column) harms discriminability but could improve robustness if pre-trained text head is used. 3) Higher data diversity (smaller threshold) also improves robustness.

3. Applications to SL and SSL

SL under extreme long-tail (or open-world recognition)

Sub. Voc is necessary to acquire CLIP knowledge

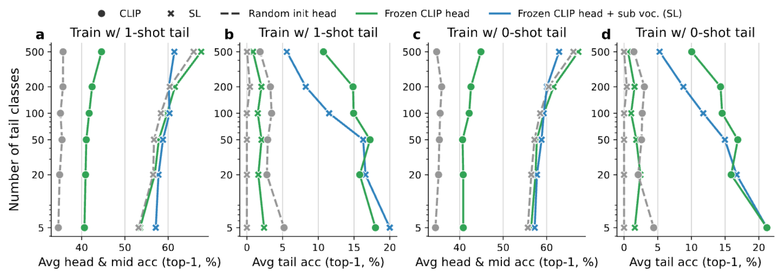


Figure 8: An extreme case: we train SL models on IN-Caps variants that have tail classes trimmed to only one shot (a & b) or even zero shot (c & d), and evaluate the accuracy on the tail and other classes. ● CLIP with a frozen pre-trained text encoder shows superior generalization, which can be acquired by a ✖ SL model with ✖ fixed class prototypes from CLIP and ✖ vocabulary subsampling.

SSL (DINO) on uncurated web data *vs* ImageNet

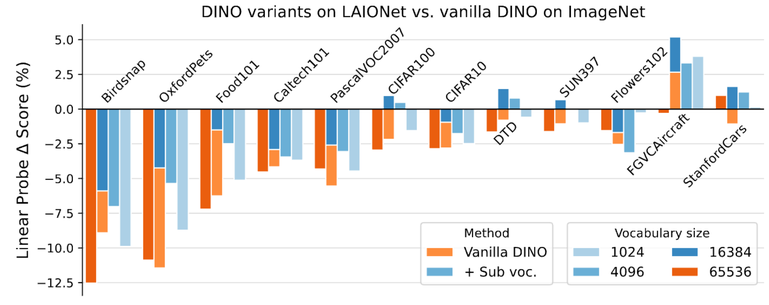
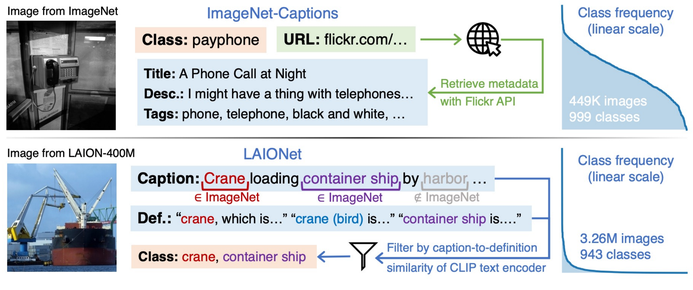


Figure 10: Transfer learning results of DINO variants pre-trained on LAIONet *vs*. vanilla DINO trained on ImageNet. Extreme data imbalance makes LAIONet much harder for DINO to learn transferrable representations. The ■ vocabulary subsampling strategy effectively helps ■ DINO alleviate such defects and generally match ImageNet-pretrained performance.

We study the reason on controllable datasets: ImageNet-Captions and LAIONet



Data scaling (also achievable via CLIP language pre-training)

Pre-trained knowledge help preserve intra-class variation while not harming inter-class margins