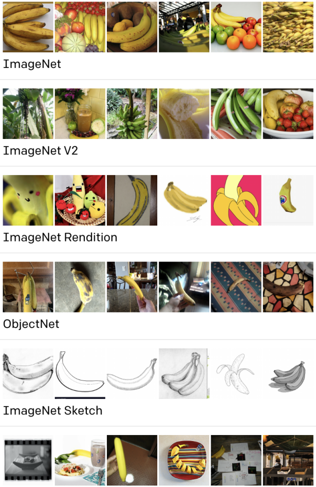
Does CLIP's generalization performance mainly stem from high train-test similarity?

Prasanna Mayilvahanan, Thaddäus Wiedemer, Evgenia Rusak, Matthias Bethge, Wieland Brendel

MPI-IS & University of Tübingen

April 19, 2024

Motivation



DATASET	IMAGENET RESNET101	CLIP VIT-L
ImageNet	76.2%	76.2%
ImageNet V2	64.3%	70.1%
ImageNet Rendition	37.7%	88.9%
ObjectNet	32.6%	72.3%
ImageNet Sketch	25.2%	60.2%
ImageNet Adversarial	2.7%	77.1%

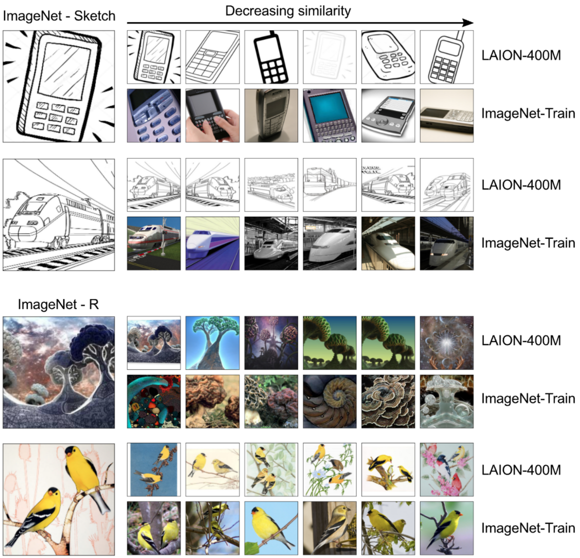
Radford et al. 2019

Why is CLIP so good?

Possible reasons (Fang et al 2022):

- Architecture
- Language supervision
- Zero-shot prediction
- **Data distribution**

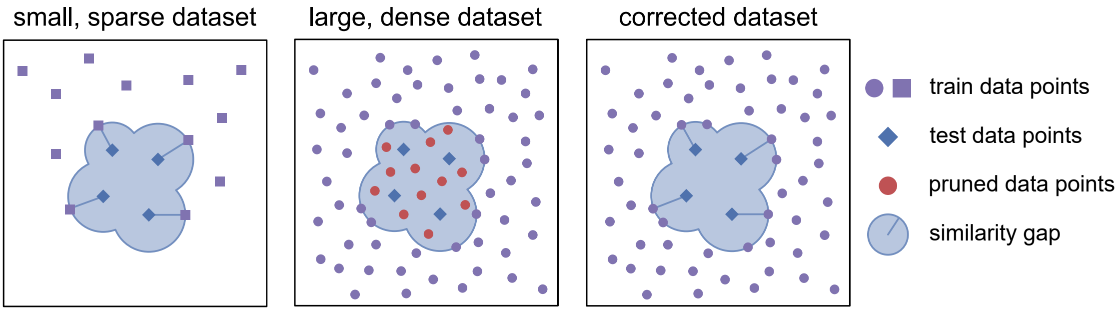
Nearest neighbors visualized



Similarity hypothesis

Is CLIP doing well only because its training set has *highly similar* images to test sets?

Pruning *highly similar* images from LAION



CLIP's accuracy after pruning

Dataset	Size	Top-1 Accuracy					
		Val	Sketch	A	R	V2	ON
OpenAI (Radford et al., 2021)	400 000 000	63.38	42.32	31.44	69.24	55.96	44.14
L-400M (Schuhmann et al., 2021)	413 000 000	62.94	49.39	21.64	73.48	55.14	43.94
L-200M	199 824 274	62.12	48.61	21.68	72.63	54.16	44.80
L-200M + IN-Train	200 966 589	68.66	50.21	23.33	72.9	59.7	43.99
├ val-pruned	*−377 340*	68.62	49.58	23.47	72.74	59.47	45.08
├ sketch-pruned	*−8 342 783*	68.34	44.78	22.7	69.35	59.52	44.12
├ a-pruned	*−138 852*	68.85	50.25	22.99	72.44	60.05	44.43
├ r-pruned	*−5 735 749*	68.71	46.92	23.44	69.48	59.6	45.08
├ v2-pruned	*−274 325*	68.79	50.45	23.19	72.58	59.84	45.33
├ objectnet-pruned	*−266 025*	68.75	50.14	22.70	72.82	59.37	43.73
└ combined-pruned	*−12 352 759*	68.05	44.12	22.15	67.88	58.61	44.39

Summary

Is CLIP doing well only because its training set has *highly similar* images to test sets?

No, the dataset scale and diversity drives CLIP to learn generalizable representations.