LAION-5B: An open large-scale dataset for training next generation image-text models

Christoph Schuhmann[1] §§°° **Romain Beaumont**[1] §§°° **Richard Vencu**[1,3,8] §§°°
Cade Gordon[2] §§°° **Ross Wightman**[1]§§ **Mehdi Cherti** [1,10]§§
Theo Coombes[1] **Aarush Katta**[1] **Clayton Mullis**[1] **Mitchell Wortsman**[6]
Patrick Schramowski[1,4,5] **Srivatsa Kundurthy**[1] **Katherine Crowson**[1,8,9]
Ludwig Schmidt[6] °° **Robert Kaczmarczyk**[1,7] °° **Jenia Jitsev**[1,10] °°
LAION[1] UC Berkeley[2] Gentec Data[3] TU Darmstadt[4] Hessian.AI[5]
University of Washington, Seattle[6] Technical University of Munich[7] Stability AI[8]
EleutherAI[9] Juelich Supercomputing Center (JSC), Research Center Juelich (FZJ)[10]
`contact@laion.ai`
§§ Equal first contributions, °° Equal senior contributions

Large datasets are key to recent advances in multimodal learning

CLIP: 400 million image-text pairs

DALL-E: X million image-text pairs

BASIC: 6.6 billion image-text pairs

Imagen: Y million image-text pairs

CLIP: Connecting Text and Images

We're introducing a neural network called CLIP which efficiently learns visual concepts from natural language supervision. CLIP can be applied to any visual classification benchmark by simply providing the names of the visual categories to be recognized, similar to the "zero-shot" capabilities of GPT-2 and GPT-3.

January 5, 2021
15 minute read

However, none of these training sets are publicly available.

What is LAION-5B?

- 5.85B Image - Text - Pairs
- filtered with OpenAI CLIP B/32 & mCLIP
- all en samples cos similarity >0.28 between image & text embeddings (>0.26 with mCLIP for non en samples)
- Source: Common Crawl
- KNN - Index

Community project

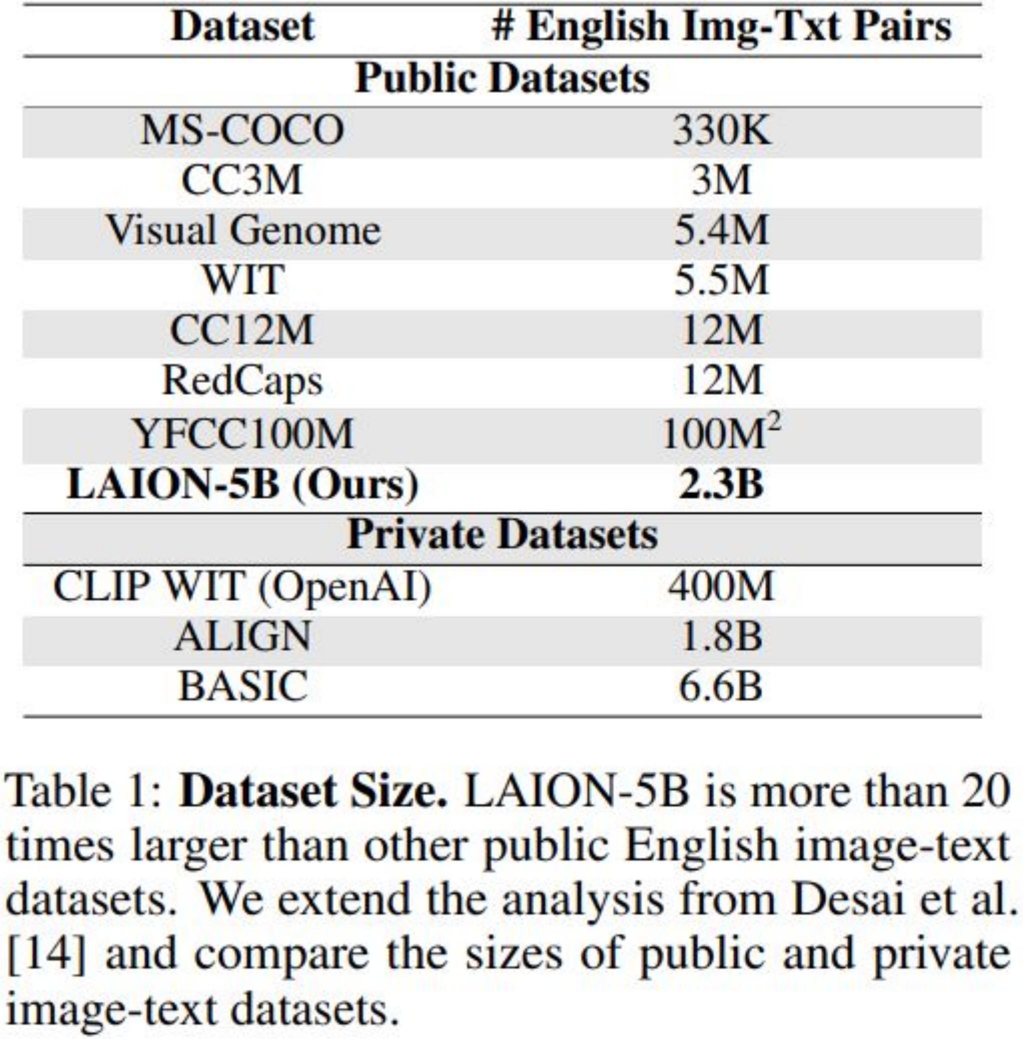
Why

Empower **independent** researchers & ML practitioners to

- Study **training** of large multi-modal models like

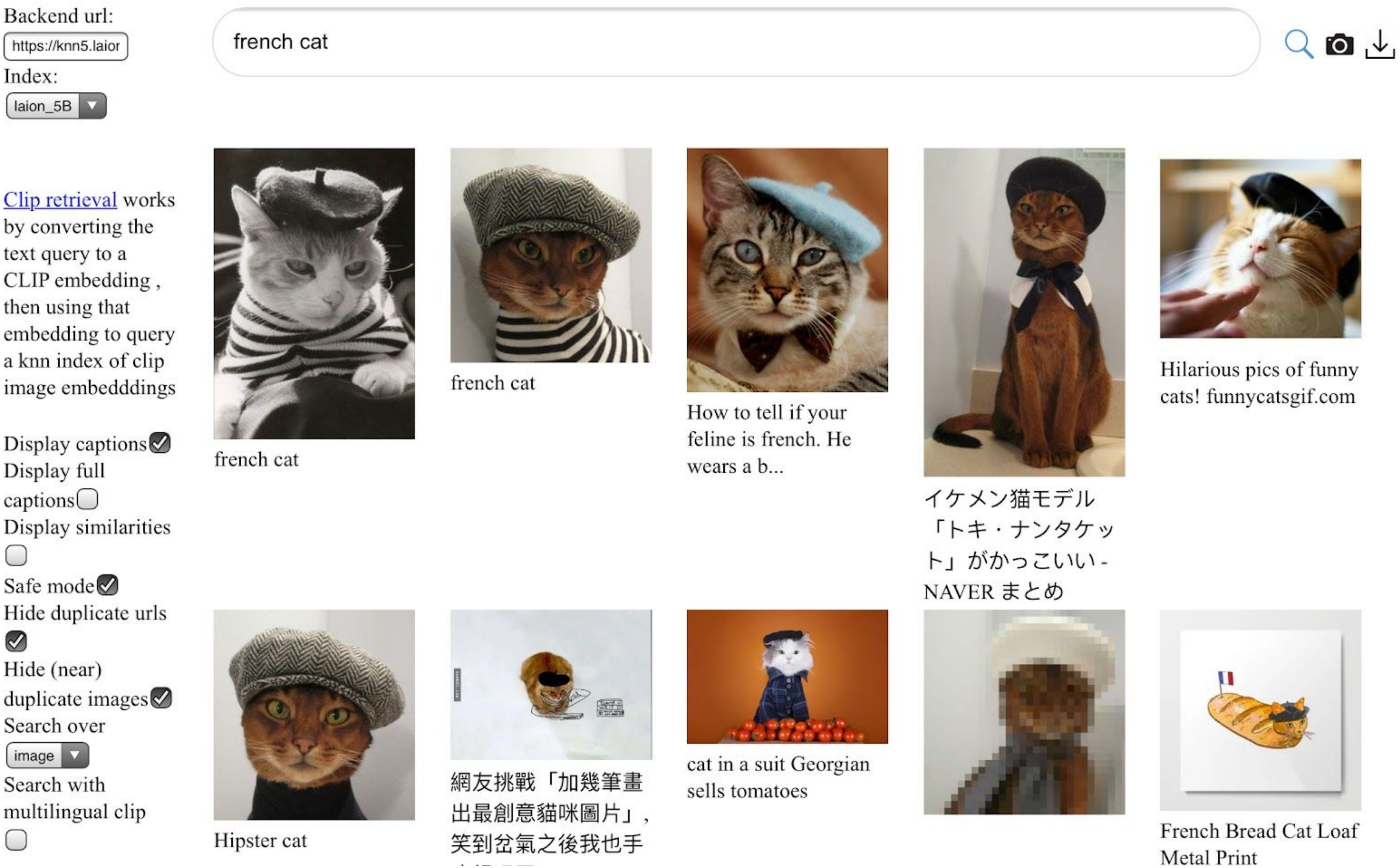
  CLIP, Stable Diffusion, Make-a-Video, …

- easily **create** domain specific **datasets**
- study **potentials** and **pitfalls** of large-scale crawled data



Dataset	# English Img-Txt Pairs
Public Datasets	
MS-COCO	330K
CC3M	3M
Visual Genome	5.4M
WIT	5.5M
CC12M	12M
RedCaps	12M
YFCC100M	100M[2]
LAION-5B (Ours)	**2.3B**
Private Datasets	
CLIP WIT (OpenAI)	400M
ALIGN	1.8B
BASIC	6.6B

Table 1: **Dataset Size.** LAION-5B is more than 20 times larger than other public English image-text datasets. We extend the analysis from Desai et al. [14] and compare the sizes of public and private image-text datasets.



Backend url: https://knn5.laior

Index: laion_5B

french cat

Clip retrieval works by converting the text query to a CLIP embedding , then using that embedding to query a knn index of clip image embedddings

Display captions ☑

Display full captions □

Display similarities □

Safe mode ☑

Hide duplicate urls ☑

Hide (near) duplicate images ☑

Search over image

Search with multilingual clip □

french cat

french cat

How to tell if your feline is french. He wears a b...

イケメン猫モデル「トキ・ナンタケット」がかっこいい - NAVER まとめ

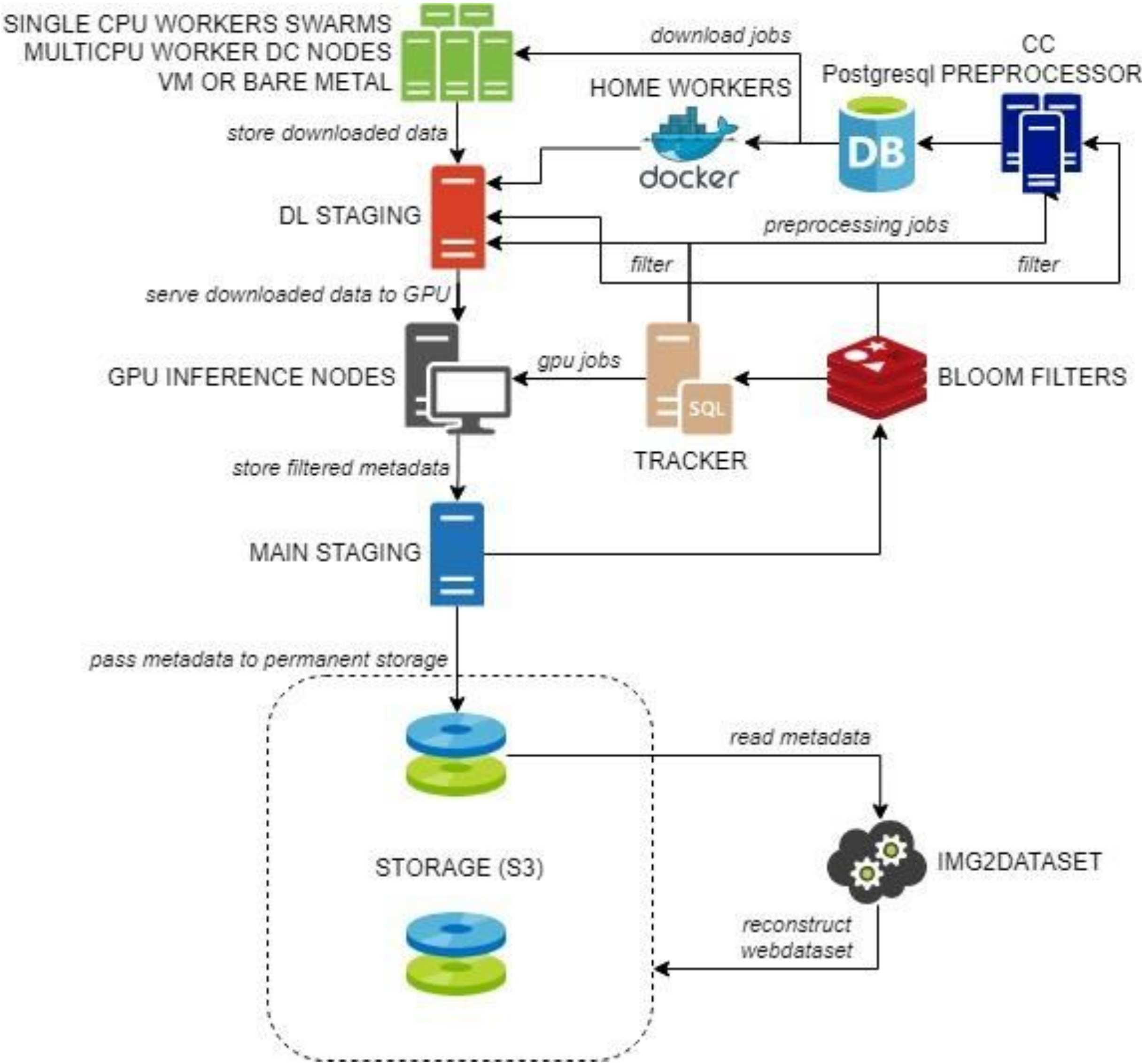
Hilarious pics of funny cats! funnycatsgif.com

Hipster cat

網友挑戰「加幾筆畫出最創意貓咪圖片」，笑到岔氣之後我也手

cat in a suit Georgian sells tomatoes

French Bread Cat Loaf Metal Print


SINGLE CPU WORKERS SWARMS
MULTICPU WORKER DC NODES
VM OR BARE METAL
download jobs
HOME WORKERS
Postgresql
CC
PREPROCESSOR
DB
docker
store downloaded data
DL STAGING
preprocessing jobs
filter
filter
serve downloaded data to GPU
GPU INFERENCE NODES
gpu jobs
SQL
TRACKER
BLOOM FILTERS
store filtered metadata
MAIN STAGING
pass metadata to permanent storage
read metadata
STORAGE (S3)
IMG2DATASET
reconstruct
webdataset

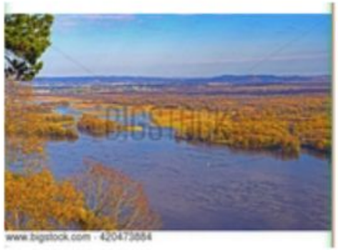
Safety

- NSFW: https://github.com/LAION-AI/CLIP-based-NSFW-Detector
- Offensive Content: https://arxiv.org/abs/2202.06675
- Watermark detection: https://github.com/LAION-AI/watermark-detection

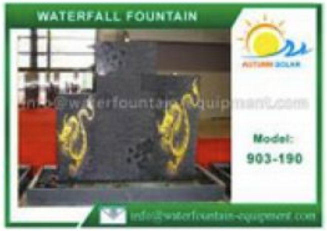
Watermark detection

WATERMARK

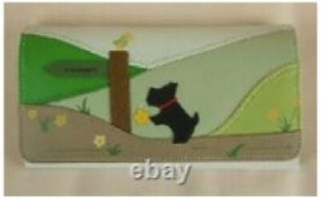
Classic watermark



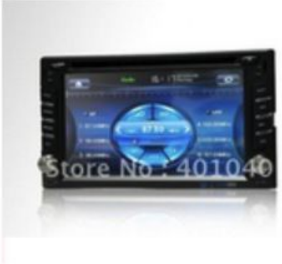
Advertisement **and** watermark


WATERFALL FOUNTAIN
info@waterfountain-equipment.com
Model:
903-190
info@waterfountain-equipment.com

Brand text covering image, watermark looks

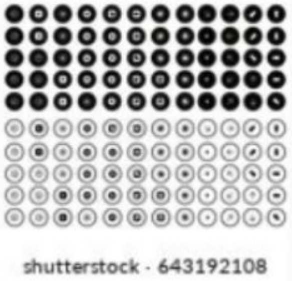


No trademark/brand, but "100%" watermark characteristics

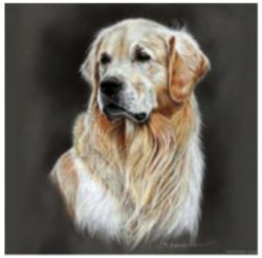


NO WATERMARK

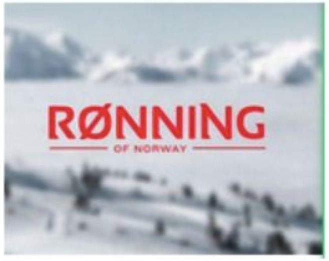
Text not covering image (also containing ®, TM)



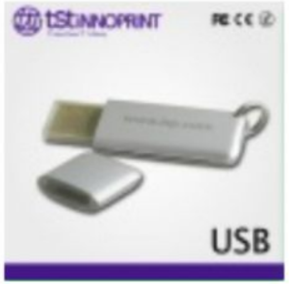
Subtle text not covering main parts of the image



Logo

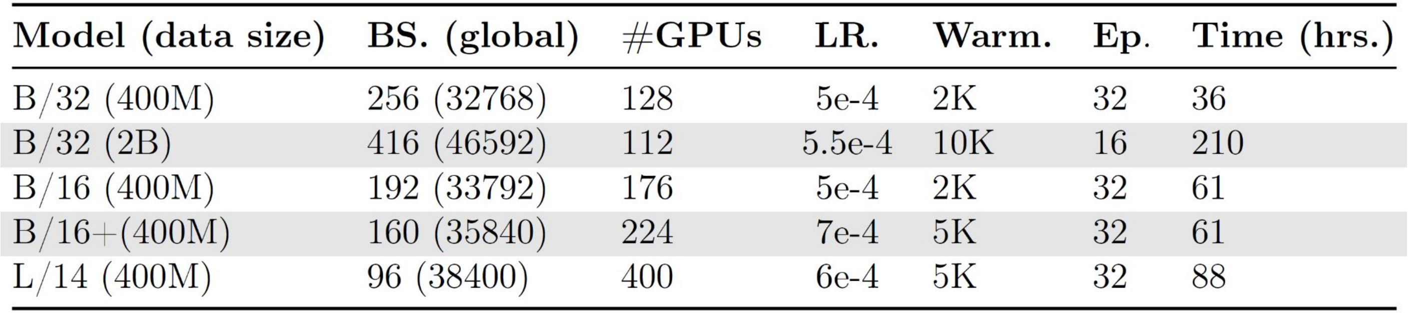


Advertisement



Training on Supercomputers

- JUWELS Booster: Juelich Supercomputing Center, Helmholtz Society, Germany: ca. 4k A100
- Stability AWS Supercomputer: ca. 4k A100



Model (data size)	BS. (global)	#GPUs	LR.	Warm.	Ep.	Time (hrs.)
B/32 (400M)	256 (32768)	128	5e-4	2K	32	36
B/32 (2B)	416 (46592)	112	5.5e-4	10K	16	210
B/16 (400M)	192 (33792)	176	5e-4	2K	32	61
B/16+(400M)	160 (35840)	224	7e-4	5K	32	61
L/14 (400M)	96 (38400)	400	6e-4	5K	32	88

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

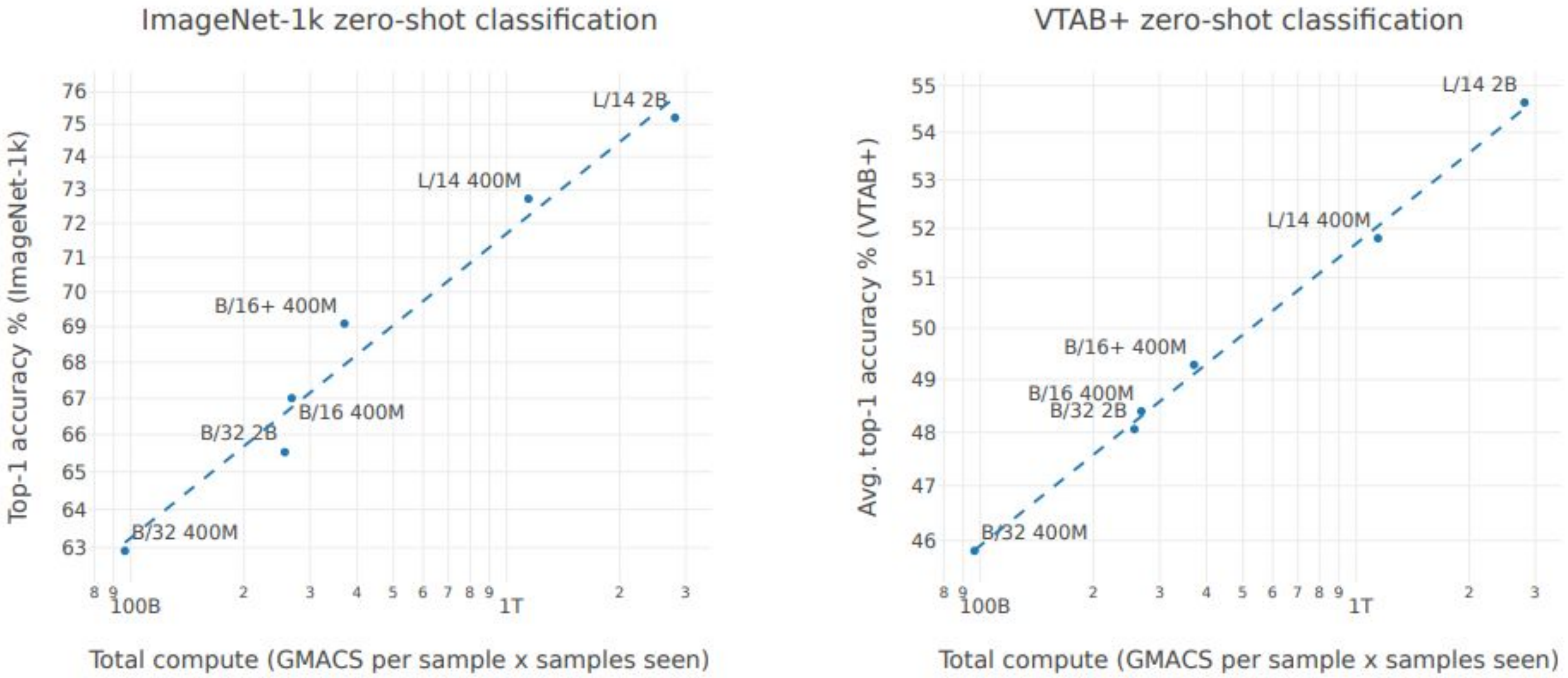


Figure 4: The relationship between total compute (giga multiply–accumulates (GMACS)) and zero-shot top-1 classification accuracy (%) of models trained on LAION (400M, 2B-en). The dashed line in each figure is a linear fit in log-log space. Each point corresponds to a model trained on either the 400M or 2B-en LAION subsets. We show results on ImageNet-1k (left) and VTAB+ (right) where we average the accuracy over 35 tasks (see Appendix E.3 for details). Clear effect of model, data and compute training scale is evident on zero-shot performance that increases following scale power law.

Get it - Use it - Improve it

- https://laion.ai/blog/laion-5b/
- https://github.com/rom1504/img2dataset
- https://github.com/rom1504/clip-retrieval
- Dataset exploration: https://knn5.laion.ai

Connect

Our LAION Discord Server
https://discord.gg/nGuc6rGdqP

Mail
contact@laion.ai

Website
https://laion.ai