Poster session: TUE-AM-268

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

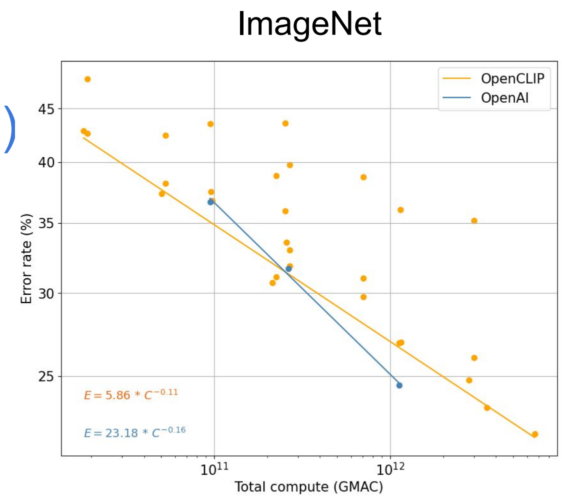
# Reproducible Scaling Laws for Contrastive Language-Image Learning

Mehdi Cherti[1,5] §§ Romain Beaumont[1] §§ Ross Wightman[1,3] §§
Mitchell Wortsman[4] §§ Gabriel Ilharco[4] §§ Cade Gordon[2]
Christoph Schuhmann[1] Ludwig Schmidt[4] °° Jenia Jitsev[1,5] §§°°
LAION[1] UC Berkeley[2] HuggingFace[3] University of Washington[4]
Juelich Supercomputing Center (JSC), Research Center Juelich (FZJ)[5]

# Overview

We study CLIP (Radford et al., 2021) scaling laws:

- Based openly available data (LAION-400M,2B)
- Using open source software (OpenCLIP)




Evaluation:

- Zero-shot classification & retrieval, linear probing, fine-tuning.

Our Findings:

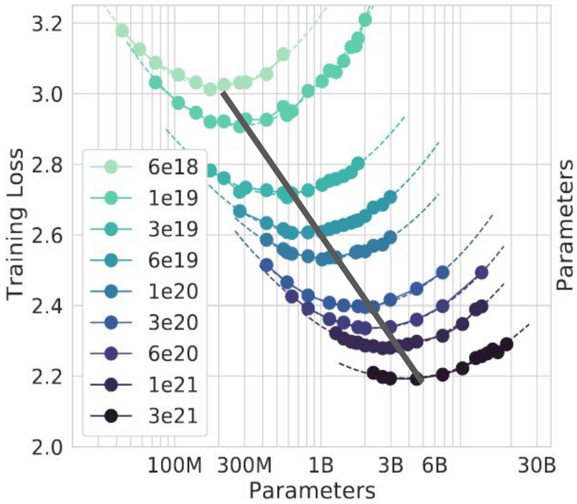
- Improvement on downstream tasks with scale, following a power-law
- Bottlenecks with small data scale/samples seen
- Task specific scaling laws: advantage of OpenCLIP LAION over OpenAI's CLIP on retrieval, advantage of OpenAI's CLIP on classification

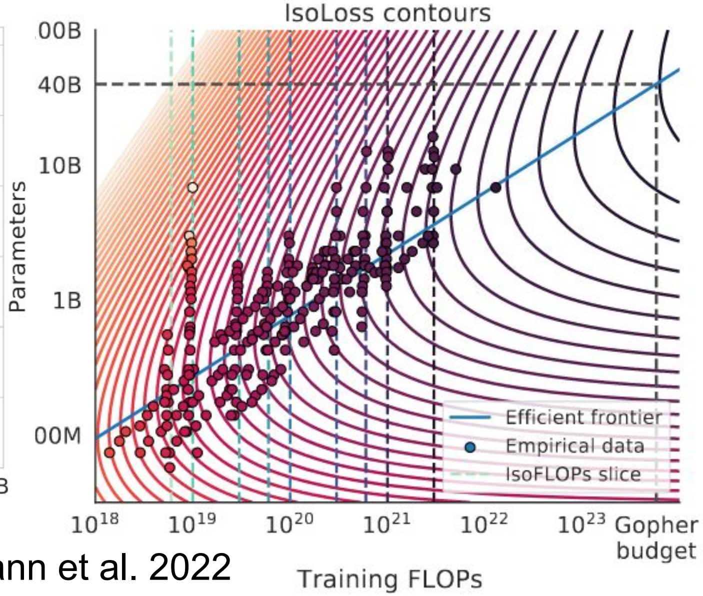
  **We open source code, full checkpoints and training/eval workflow**

# Background: neural scaling laws

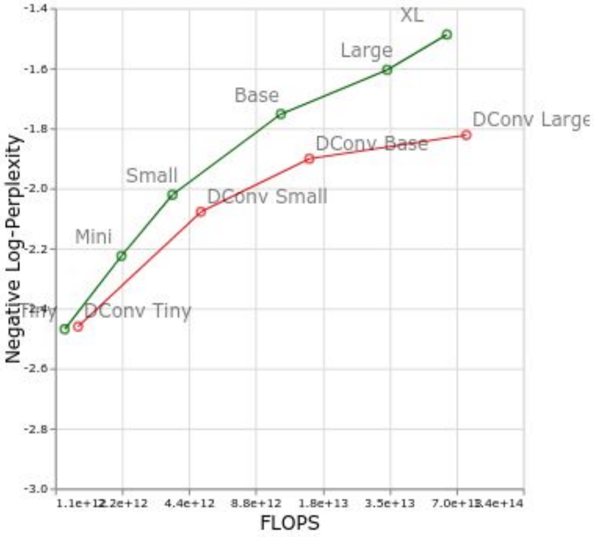
Why scaling laws?

- Extrapolate model performance on larger scale
- Compute optimal model size for a given compute budget
- Compare scaling curves of different architectures/pre-training datasets/losses




Hoffmann et al. 2022




Tay et al. 2022

# Scaling laws for contrastive language-image training

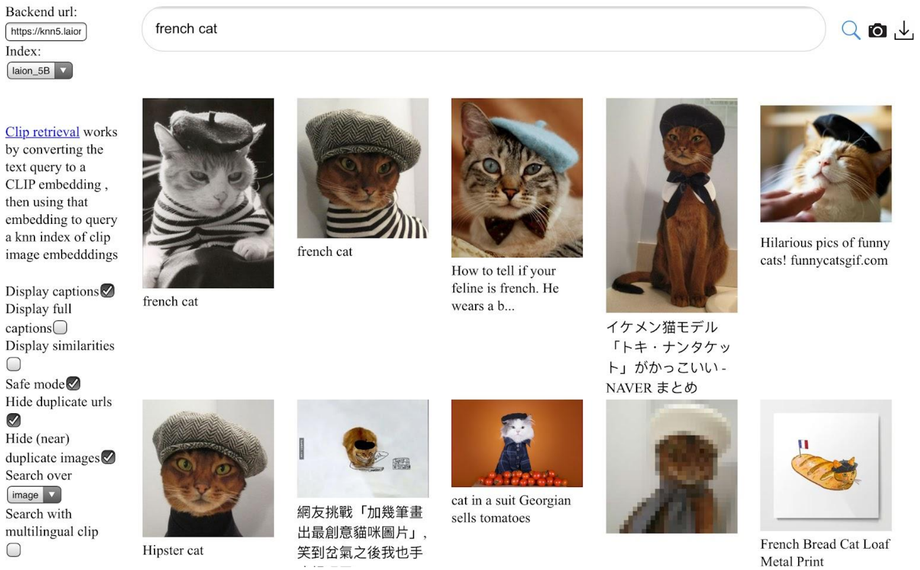
Existing works on contrastive language-image training:

- Show benefit of scaling but do not study it systematically
- Rely on private datasets
- Usually involve a customized training procedure

# Pre-training data

We use the open & large LAION-5B dataset

- Data scales: 80M, 400M, 2B
- Samples seen (compute): 3B, 13B, 34B




| Dataset | # English Img-Txt Pairs |
|---|---|
| **Public Datasets** | |
| LAION-400M | 407M |
| LAION-2B | 2.3B |
| **Private Datasets** | |
| CLIP WIT (OpenAI) | 400M |
| ALIGN | 1.8B |
| BASIC | 6.6B |

# Model sizes

We use models ranging from ~150M to ~1.4B parameters

| Name | Width | Emb. | Depth | Acts. | Params | GMAC |
|---|---|---|---|---|---|---|
| ViT-B/32 | 768 / 512 | 512 | 12 / 12 | 10 M | 151 M | 7.40 |
| ViT-B/16 | 768 / 512 | 512 | 12 / 12 | 29 M | 150 M | 20.57 |
| ViT-L/14 | 1024 / 768 | 768 | 24 / 12 | 97 M | 428 M | 87.73 |
| ViT-H/14 | 1280 / 1024 | 1024 | 32 / 24 | 161 M | 986 M | 190.97 |
| ViT-g/14 | 1408 / 1024 | 1024 | 40 / 24 | 214 M | 1.37 B | 290.74 |

# Downstream datasets

| Dataset | Abbr. | Test size | #Classes |
|---|---|---|---|
| ImageNet | INet | 50,000 | 1,000 |
| ImageNet-v2 | INet-v2 | 10,000 | 1,000 |
| ImageNet-R | INet-R | 30,000 | 200 |
| ImageNet Sketch | INet-S | 50,889 | 1,000 |
| ObjectNet | ObjNet | 18,574 | 113 |
| ImageNet-A | INet-A | 7,500 | 200 |
| CIFAR-10 | - | 10,000 | 10 |
| CIFAR-100 | - | 10,000 | 100 |
| MNIST | - | 10,000 | 10 |
| Oxford Flowers 102 | Flowers102 | 6,149 | 102 |
| Stanford Cars | Cars | 8,041 | 196 |
| SVHN | - | 26,032 | 10 |
| Facial Emotion Recognition 2013 | FER2013 | 7,178 | 7 |
| RenderedSST2 | - | 1,821 | 2 |
| Oxford-IIIT Pets | Pets | 3,669 | 37 |
| Caltech-101 | - | 6,085 | 102 |
| Pascal VOC 2007 Classification | VOC2007-Cl | 14,976 | 20 |
| SUN397 | - | 108,754 | 397 |
| FGVC Aircraft | - | 3,333 | 100 |
| Country211 | - | 21,100 | 211 |
| Describable Textures | DTD | 1,880 | 47 |
| GTSRB | - | 12,630 | 43 |
| STL10 | - | 8,000 | 10 |
| Diabetic Retinopathy | Retino | 42,670 | 5 |
| EuroSAT | - | 5,400 | 10 |
| RESISC45 | - | 6,300 | 45 |
| PatchCamelyon | PCAM | 32,768 | 2 |
| CLEVR Counts | - | 15,000 | 8 |
| CLEVR Object Distance | CLEVR Dist | 15,000 | 6 |
| DSPRITES Orientation | DSPRITES Orient | 73,728 | 40 |
| DSPRITES Position | DSPRITES pos | 73,728 | 32 |
| SmallNORB Elevation | SmallNORB Elv | 12,150 | 9 |
| SmallNORB Azimuth | SmallNORB Azim | 12,150 | 18 |
| DMLAB | - | 22,735 | 6 |
| KITTI closest vehicle distance | KITTI Dist | 711 | 4 |
| MS-COCO | - | 5,000 | - |
| Flickr30K | - | 1,000 | - |

Table 25: Datasets used for evaluating downstream performance. Adapted from [65].

We evaluate the models on :

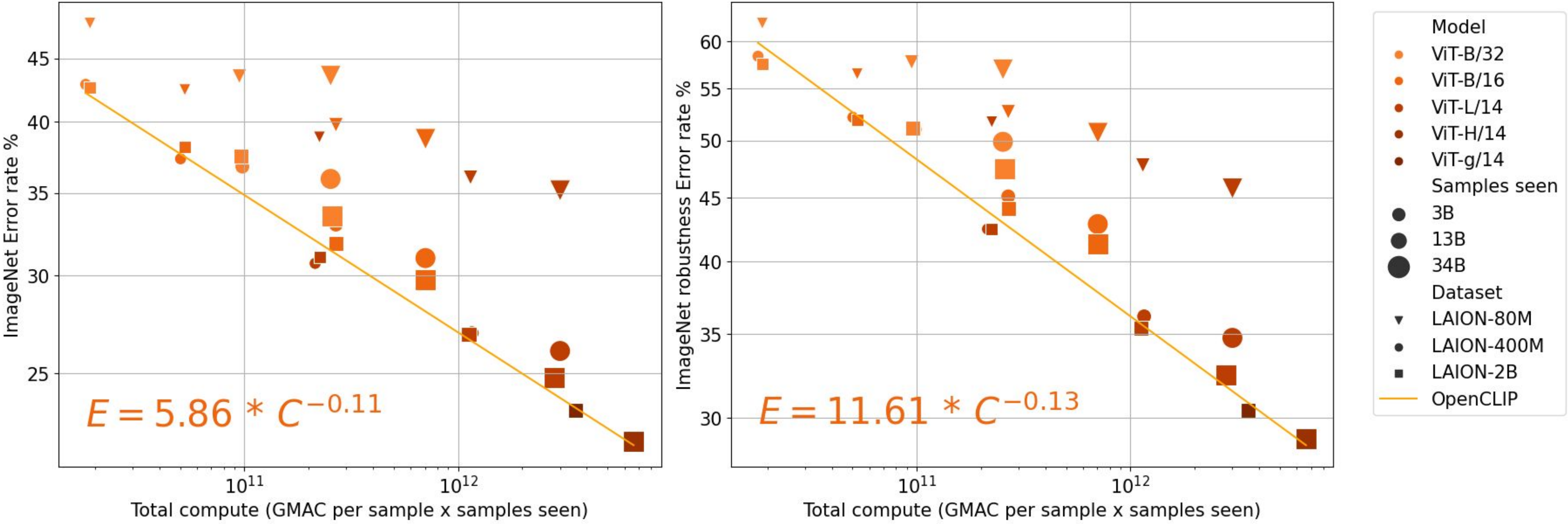
- Zero-shot classification
- Few-shot and full- shot linear probing
- Fine-tuning
- Zero-shot retrieval (COCO, FLickr-30K)

# Evaluation: zero-shot classification

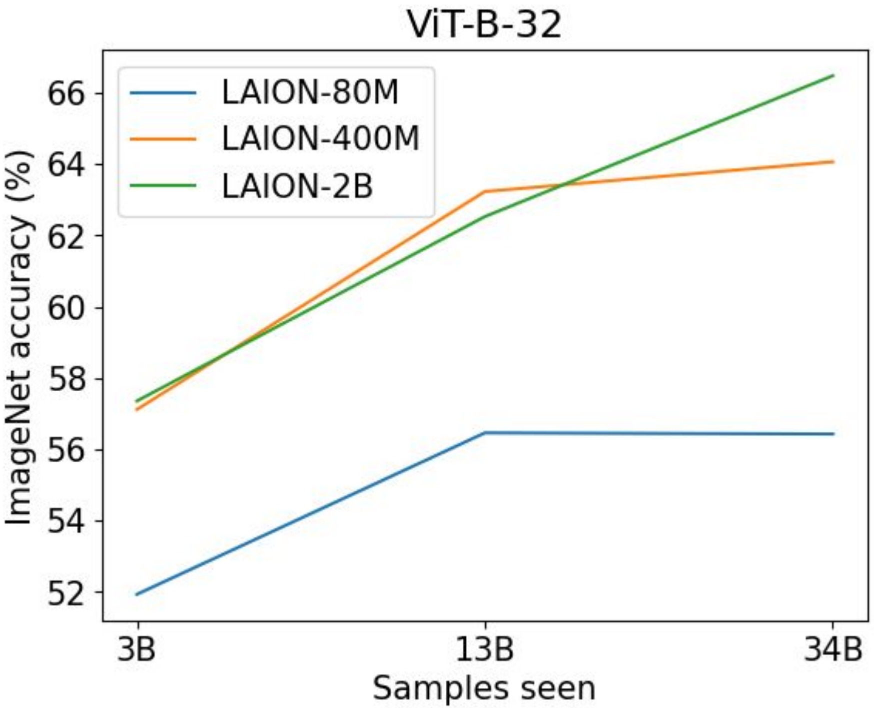
$E = \beta C^{\alpha}$ ,where E is error rate (downstream), C is total compute (GMAC)

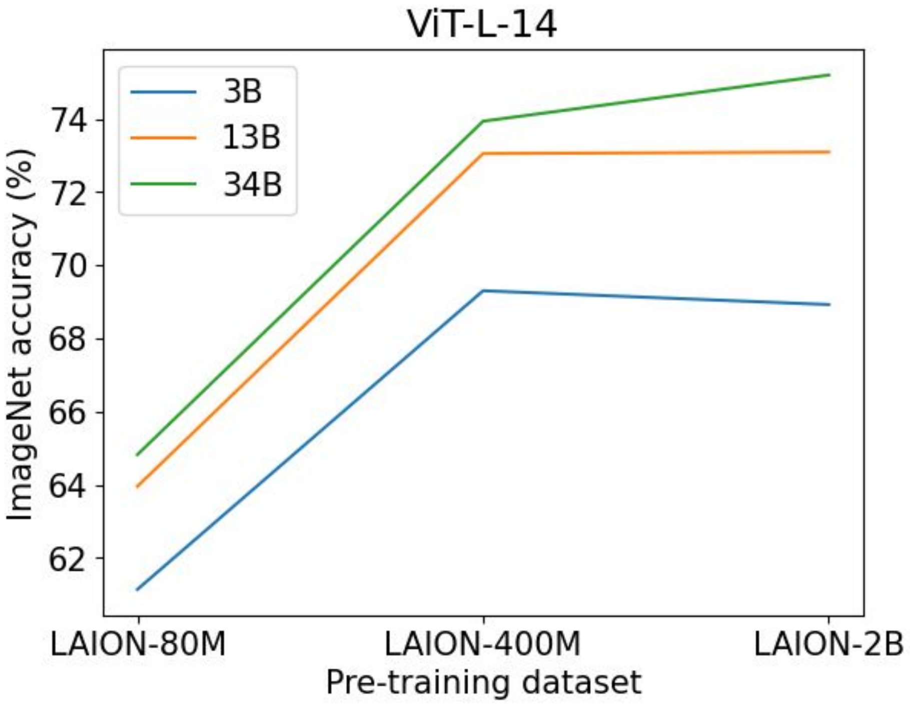
## ImageNet

## ImageNet robustness



# Evaluation: zero-shot classification, bottlenecks



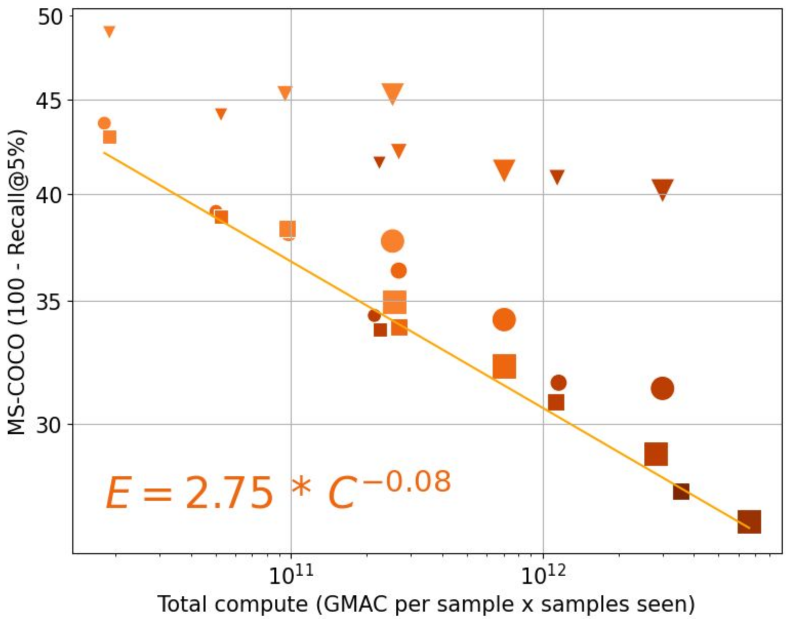

**Data scale bottleneck**

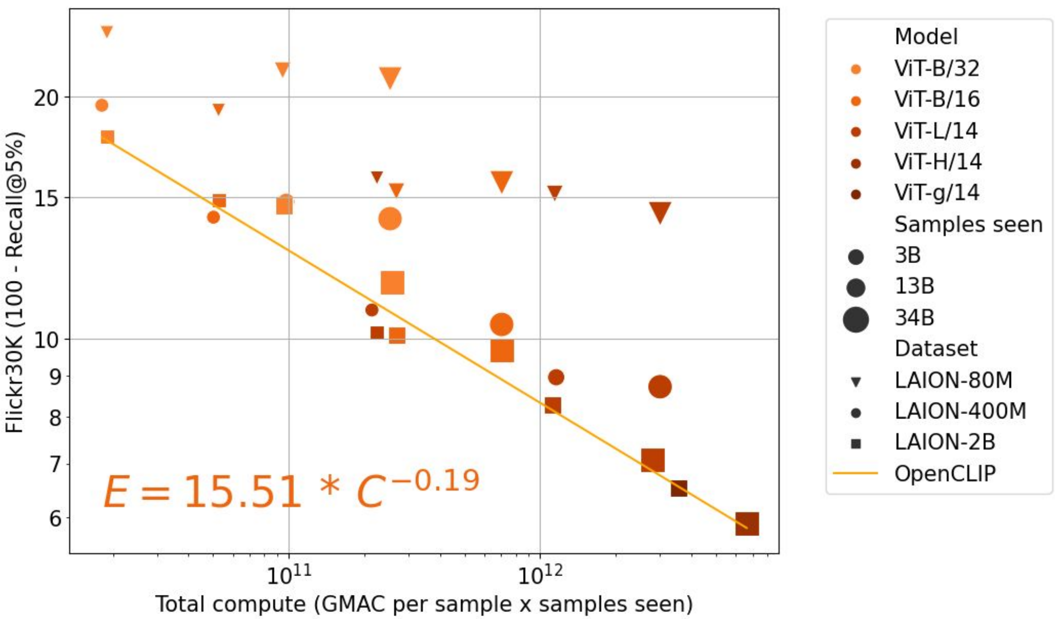
**Samples seen bottleneck**

# Evaluation: zero-shot retrieval

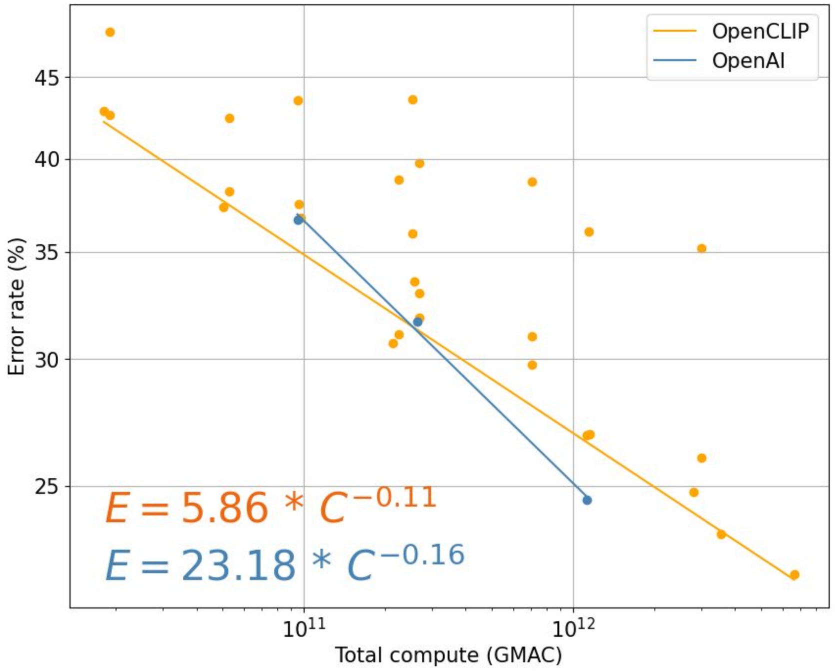
COCO

Flickr-30K

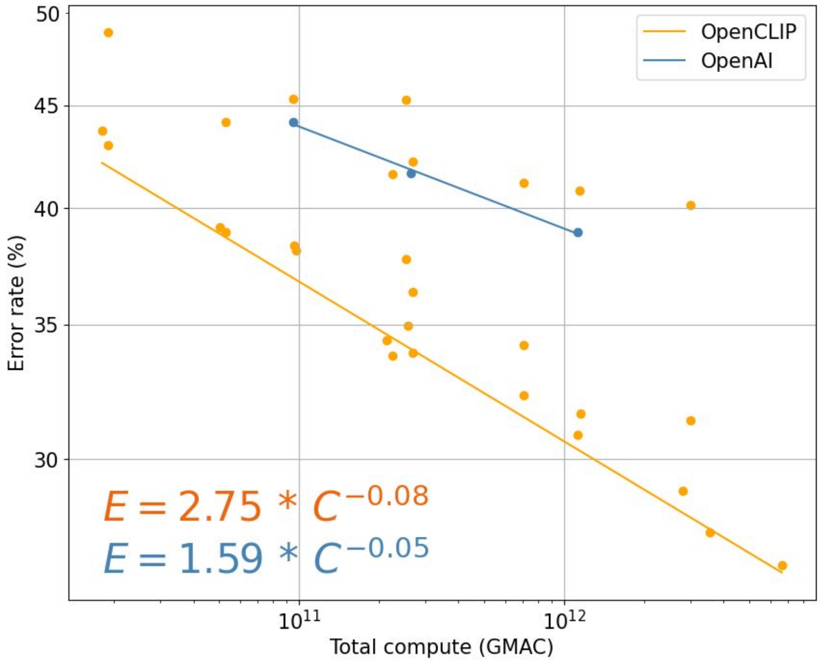


# Task-specific scaling laws



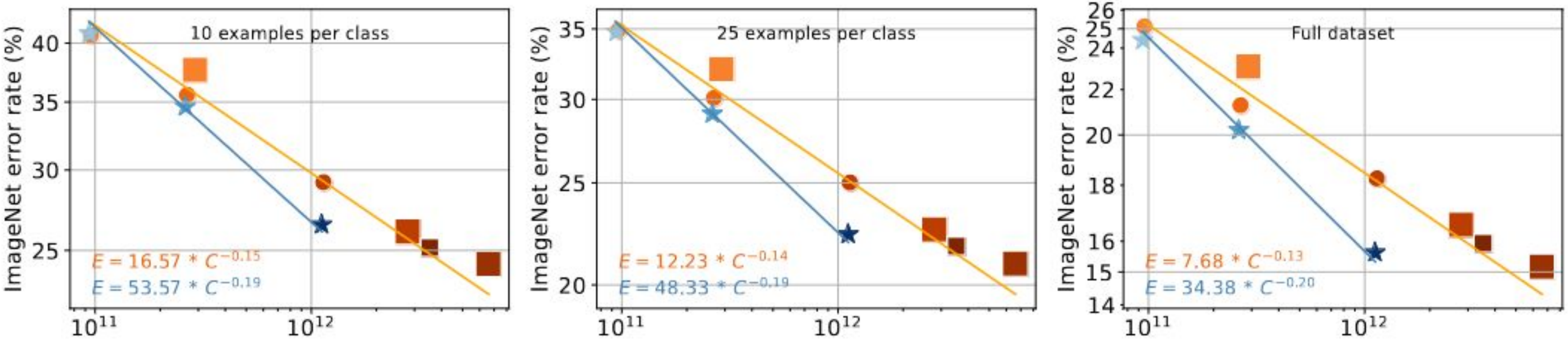

Zero-shot classification (ImageNet)



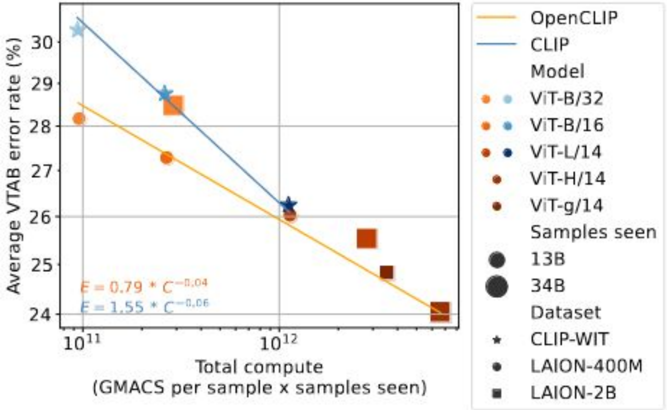

Zero-shot retrieval (COCO)

# Evaluation: linear probing

ImageNet



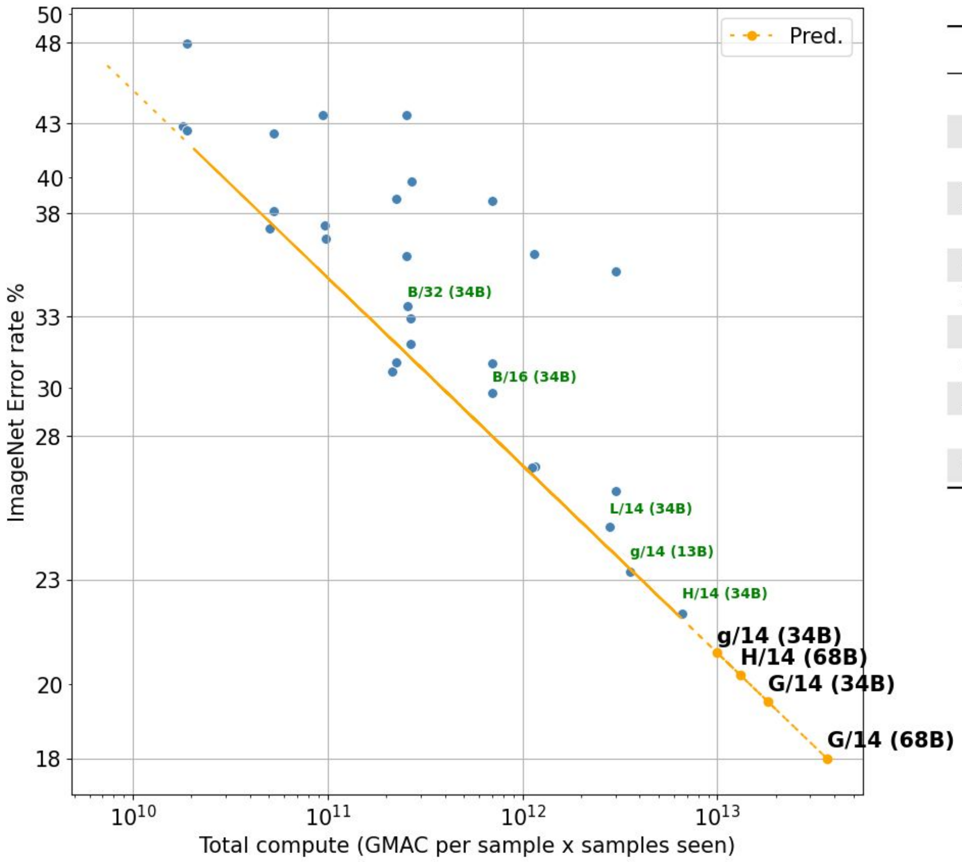

VTAB (19 tasks)



# Evaluation: fine-tuning

| | ImageNet-1k top-1 accuracy (%) | |
|---|---|---|
| Model | No extra FT | Extra FT (ImageNet-12k) |
| ViT-B/32 | 82.58 | 85.11 |
| ViT-B/16 | 86.53 | 87.17 |
| ViT-L/14 | 87.78 | 88.17 |
| ViT-H/14 | 87.59 | **88.50** |

# Scaling curves for performance prediction




| Model | ImageNet top-1 (%) | MS-COCO Recall@5 (%) |
|---|---|---|
| H/14 (3B) | 70.78 | 67.58 |
| g/14 (3B) | 72.11 | 68.65 |
| G/14 (3B) | 73.93 | 70.12 |
| H/14 (13B) | 75.62 | 71.52 |
| g/14 (13B) | 76.66* | 72.40* |
| G/14 (13B) | 78.26 | 73.75 |
| H/14 (34B) | 77.97* | 73.43* |
| g/14 (34B) | 79.11 | 74.48 |
| G/14 (34B) | 80.47 | 75.68 |
| H/14 (68B) | 79.73 | 75.03 |
| g/14 (68B) | 80.66 | 75.85 |
| G/14 (68B) | 81.92 | 76.99 |

(*): actual measured model performance values

# Thank you!

https://github.com/LAION-AI/scaling-laws-openclip