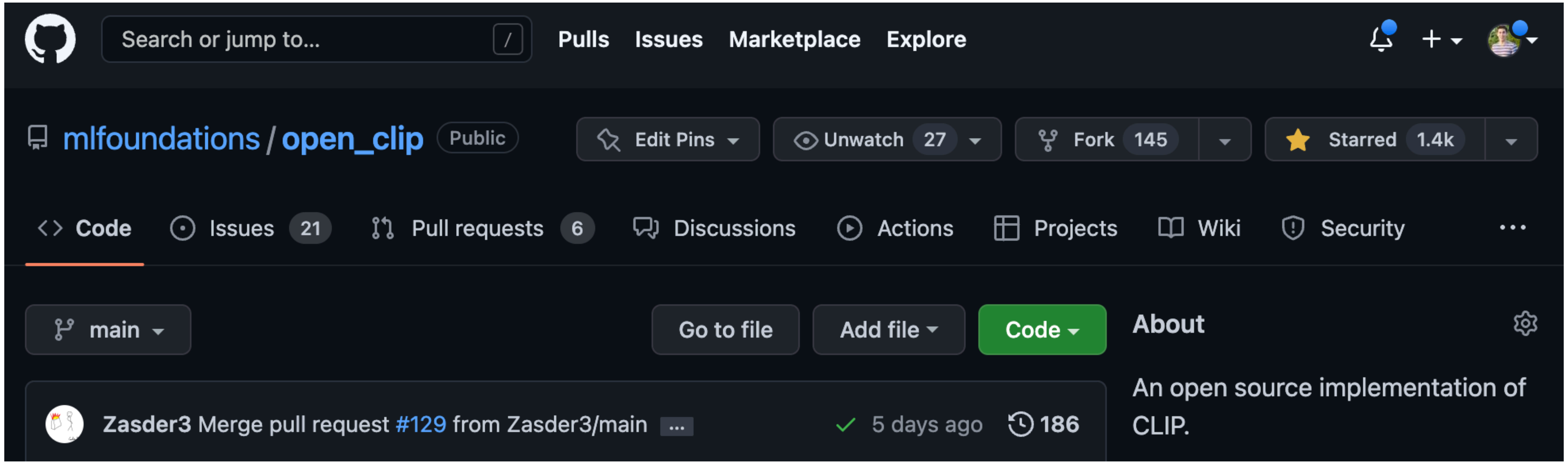
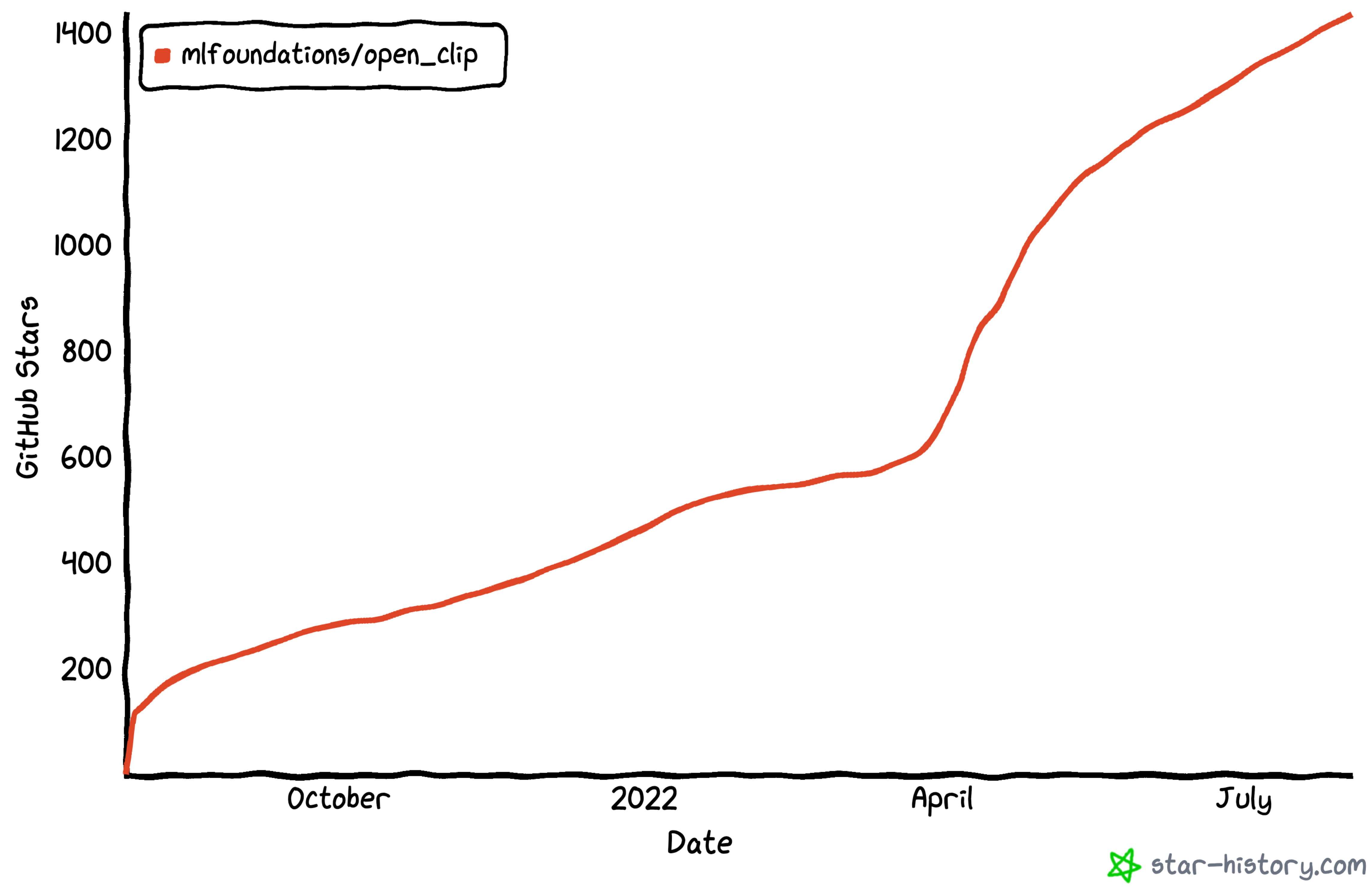
OpenCLIP:
an open source implementation of CLIP

Gabriel Ilharco*, **Mitchell Wortsman***, Cade Gordon*, Ross Wightman*, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, John Miller, Hongseok Namkoong, Hannaneh Hajishirzi, Ali Farhadi, Ludwig Schmidt


Search or jump to...
Pulls
Issues
Marketplace
Explore
mlfoundations / open_clip
Public
Edit Pins
Unwatch 27
Fork 145
Starred 1.4k
Code
Issues 21
Pull requests 6
Discussions
Actions
Projects
Wiki
Security
main
Go to file
Add file
Code
About
Zasder3 Merge pull request #129 from Zasder3/main
5 days ago
186
An open source implementation of CLIP.


mlfoundations/open_clip
GitHub Stars
1400
1200
1000
800
600
400
200
October
2022
April
July
Date
star-history.com

Interacting with CLIP.ipynb

File Edit View Insert Runtime Tools Help

+ Code + Text Copy to Drive

Loading the model

`clip.available_models()` will list the names of available CLIP models.

```
import open_clip
open_clip.list_pretrained()
```

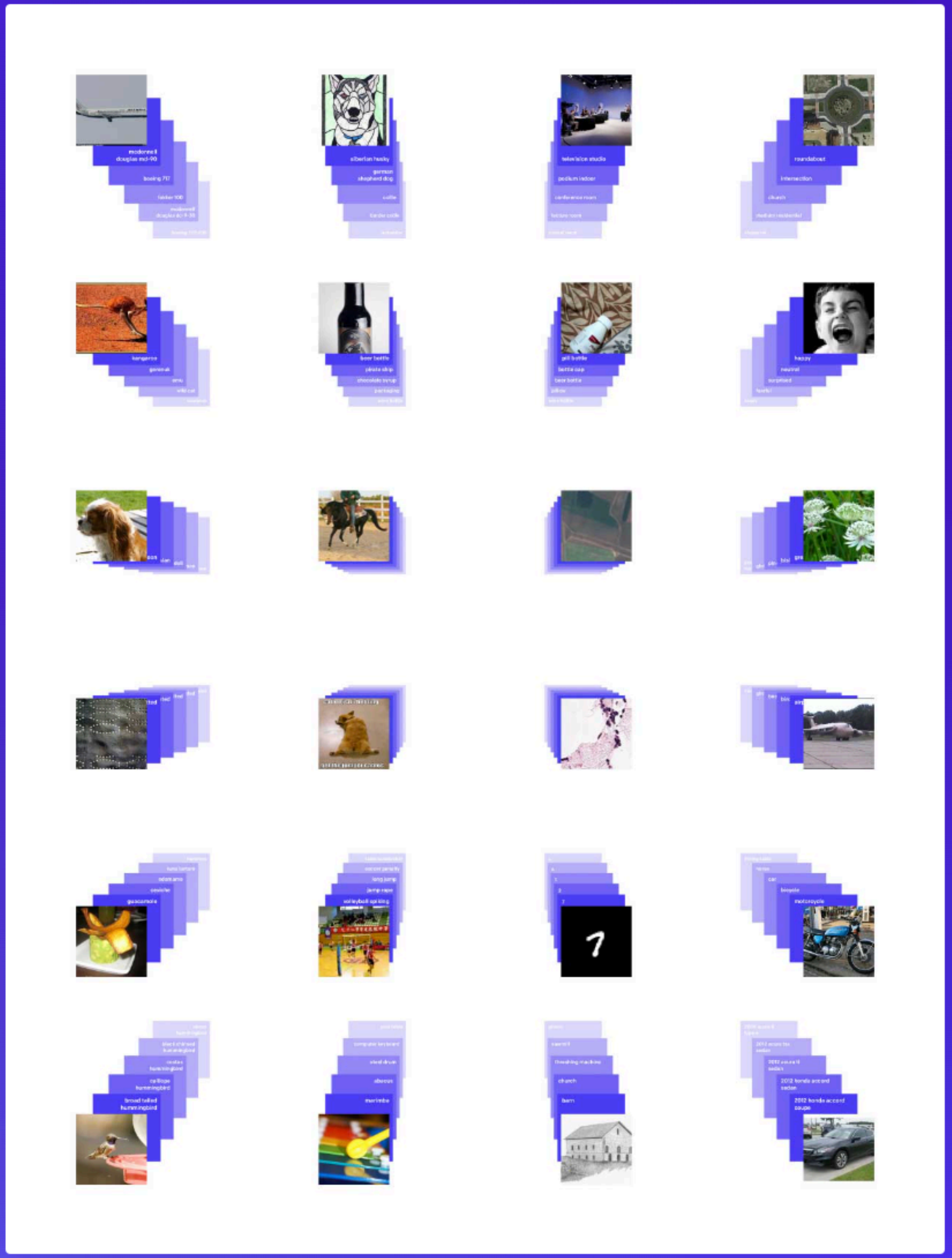
```
[('RN50', 'openai'),
 ('RN50', 'yfcc15m'),
 ('RN50', 'cc12m'),
 ('RN50-quickgelu', 'openai'),
 ('RN50-quickgelu', 'yfcc15m'),
 ('RN50-quickgelu', 'cc12m'),
 ('RN101', 'openai'),
 ('RN101', 'yfcc15m'),
 ('RN101-quickgelu', 'openai'),
 ('RN101-quickgelu', 'yfcc15m'),
 ('RN50x4', 'openai'),
 ('RN50x16', 'openai'),
 ('ViT-B-32', 'openai'),
 ('ViT-B-32', 'laion400m_e31'),
 ('ViT-B-32', 'laion400m_e32'),
 ('ViT-B-32', 'laion400m_avg'),
 ('ViT-B-32-quickgelu', 'openai'),
 ('ViT-B-32-quickgelu', 'laion400m_e31'),
 ('ViT-B-32-quickgelu', 'laion400m_e32'),
 ('ViT-B-32-quickgelu', 'laion400m_avg'),
 ('ViT-B-16', 'openai'),
 ('ViT-L-14', 'openai')]
```

```
[ ] model, _, preprocess = open_clip.create_model_and_transforms('ViT-B-32-quickgelu', pretrained='laion400m_e32')
```

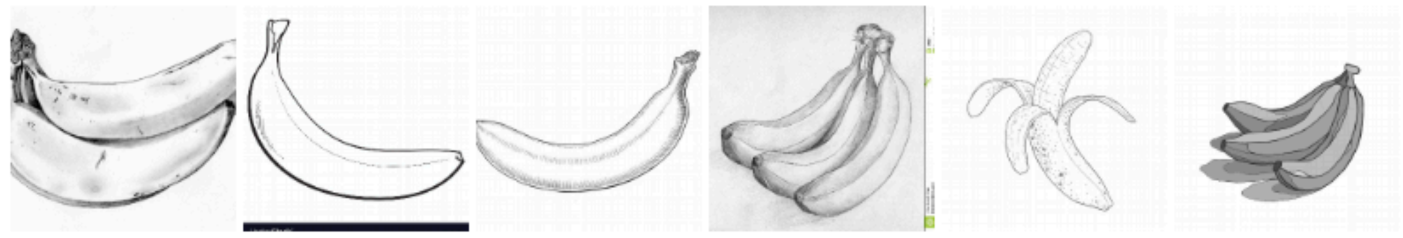
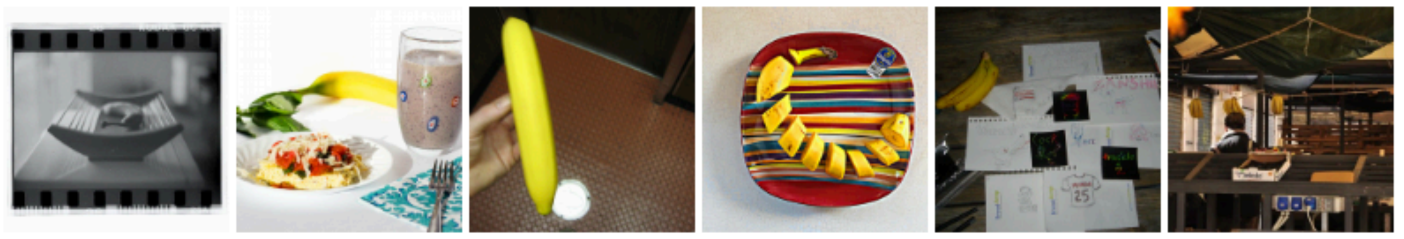
API RESEARCH BLOG ABOUT

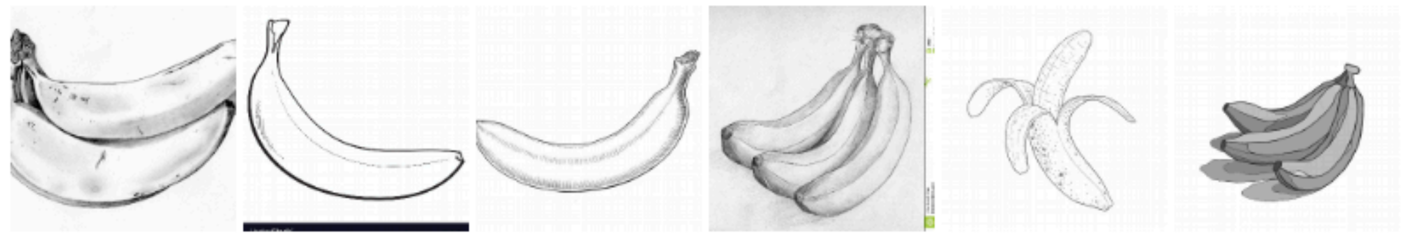
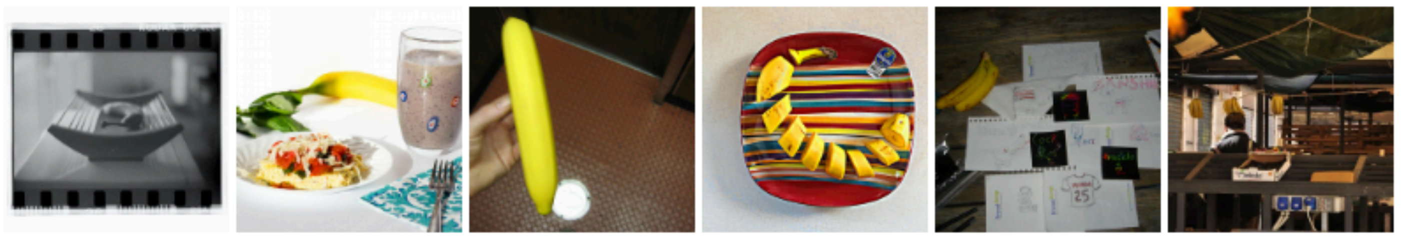
CLIP: Connecting Text and Images

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.

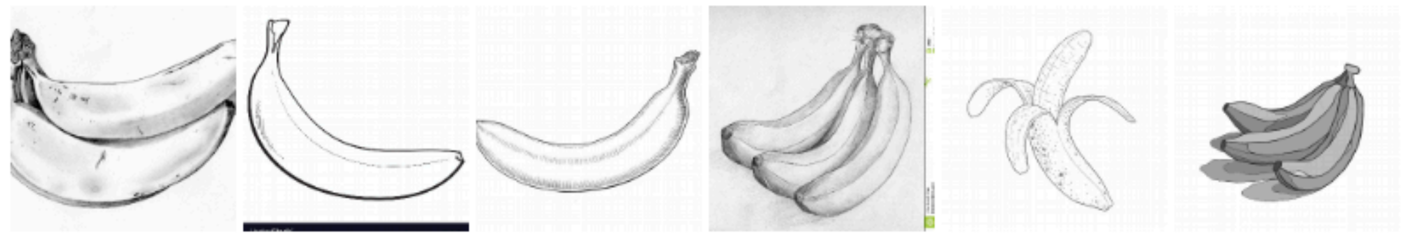
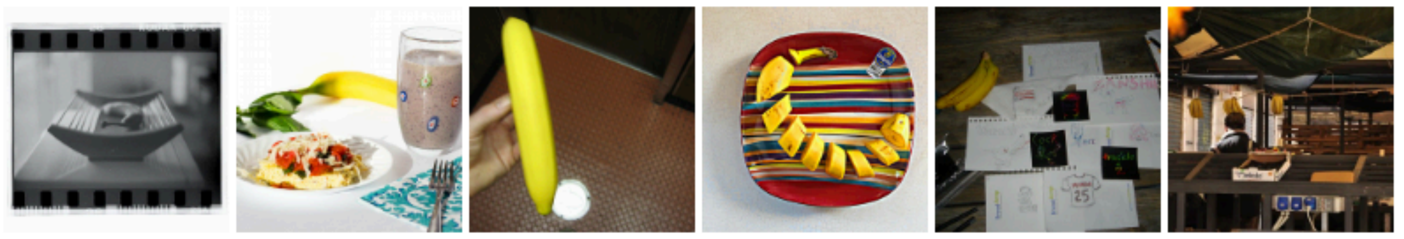


January 5, 2021
15 minute read

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%

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%

+0

DATASET	IMAGENET RESNET101	CLIP VIT-L	
 ImageNet	76.2%	76.2%	+0
 ImageNet V2	64.3%	70.1%	+6
 ImageNet Rendition	37.7%	88.9%	+51
 ObjectNet	32.6%	72.3%	+40
 ImageNet Sketch	25.2%	60.2%	+35
 ImageNet Adversarial	2.7%	77.1%	+74

WILDS

iWildCam

Without unlabeled data

Rank	Algorithm	Model	Test ID Macro F1	Test ID Avg Acc	Test OOD Macro F1 ▼	Test OOD Avg Acc	Contact
1	**Model Soups (CLIP ViT-L)**	ViT-L	57.6 (1.9) *	79.1 (0.4) *	43.3 (1.0) *	79.3 (0.3) *	Mitchell Wortsmar
2	**ERM (CLIP ViT-L)**	ViT-L	55.8 (1.9) *	77.0 (0.7) *	41.4 (0.5) *	78.3 (1.1) *	Mitchell Wortsmar

FMoW

Without unlabeled data

Rank	Algorithm	Model	Val Avg Acc	Test Avg Acc	Val Worst-region Acc	Test Worst-region Acc ▼	Contact
1	**Model Soups (CLIP ViT-L)**	ViT-L	75.7 (0.07) *	69.5 (0.08) *	59.8 (0.43) *	47.6 (0.33) *	Mitchell Wortsm
2	**ERM (CLIP ViT-L)**	ViT-L	73.6 (0.23) *	66.9 (0.17) *	59.5 (1.31) *	46.1 (0.59) *	Mitchell Wortsm

Motivation

CLIP demonstrates unprecedented performance on robustness benchmarks

We would like to understand this, but we can't train our own CLIP as the code is not public

Building OpenCLIP

Building OpenCLIP

- Went fairly smoothly thanks to lots of guidance from Jong Wook Kim and Alec Radford

Building OpenCLIP

- Went fairly smoothly thanks to lots of guidance from Jong Wook Kim and Alec Radford
- Ran into problems with:

Building OpenCLIP

- Went fairly smoothly thanks to lots of guidance from Jong Wook Kim and Alec Radford
- Ran into problems with:
  - Can’t store lots of small files. Solution: WebDataSet.

Building OpenCLIP

- Went fairly smoothly thanks to lots of guidance from Jong Wook Kim and Alec Radford
- Ran into problems with:
  - Can’t store lots of small files. Solution: WebDataSet.
  - PyTorch distributed gradient. Solution: Full contrastive matrix on each GPU.

Building OpenCLIP

- Went fairly smoothly thanks to lots of guidance from Jong Wook Kim and Alec Radford
- Ran into problems with:
  - Can’t store lots of small files. Solution: WebDataSet.
  - PyTorch distributed gradient. Solution: Full contrastive matrix on each GPU.
  - Scaling from 15m to 2b images. Solution: Ross Wightman & Cade Gordon.

What have people done with OpenCLIP?

What have people done with OpenCLIP?

- Match the accuracy of CLIP on YFCC-15m (OpenCLIP v0)

What have people done with OpenCLIP?

- Match the accuracy of CLIP on YFCC-15m (OpenCLIP v0)
- Scale OpenCLIP to LAION 400m and 2b and release all models so that they can be used and analyzed (OpenCLIP v1)

What have people done with OpenCLIP?

- Match the accuracy of CLIP on YFCC-15m (OpenCLIP v0)
- Scale OpenCLIP to LAION 400m and 2b and release all models so that they can be used and analyzed (OpenCLIP v1)
- Investigate the robustness properties of CLIP, and determine that they come from the data distribution and not language (Fang et al., 2022)

What have people done with OpenCLIP?

- Match the accuracy of CLIP on YFCC-15m (OpenCLIP v0)
- Scale OpenCLIP to LAION 400m and 2b and release all models so that they can be used and analyzed (OpenCLIP v1)
- Investigate the robustness properties of CLIP, and determine that they come from the data distribution and not language (Fang et al., 2022)
- Show that different pre-training datasets lead to different robustness trends (Nguyen et al., 2022)

What have people done with OpenCLIP?

- Match the accuracy of CLIP on YFCC-15m (OpenCLIP v0)
- Scale OpenCLIP to LAION 400m and 2b and release all models so that they can be used and analyzed (OpenCLIP v1)
- Investigate the robustness properties of CLIP, and determine that they come from the data distribution and not language (Fang et al., 2022)
- Show that different pre-training datasets lead to different robustness trends (Nguyen et al., 2022)
- <your project here>

YFCC-15m

YFCC-15m

- OpenCLIP trained on YFCC-15m **matches the accuracy of CLIP** trained on YFCC-15m and can be trained in a few days on one node

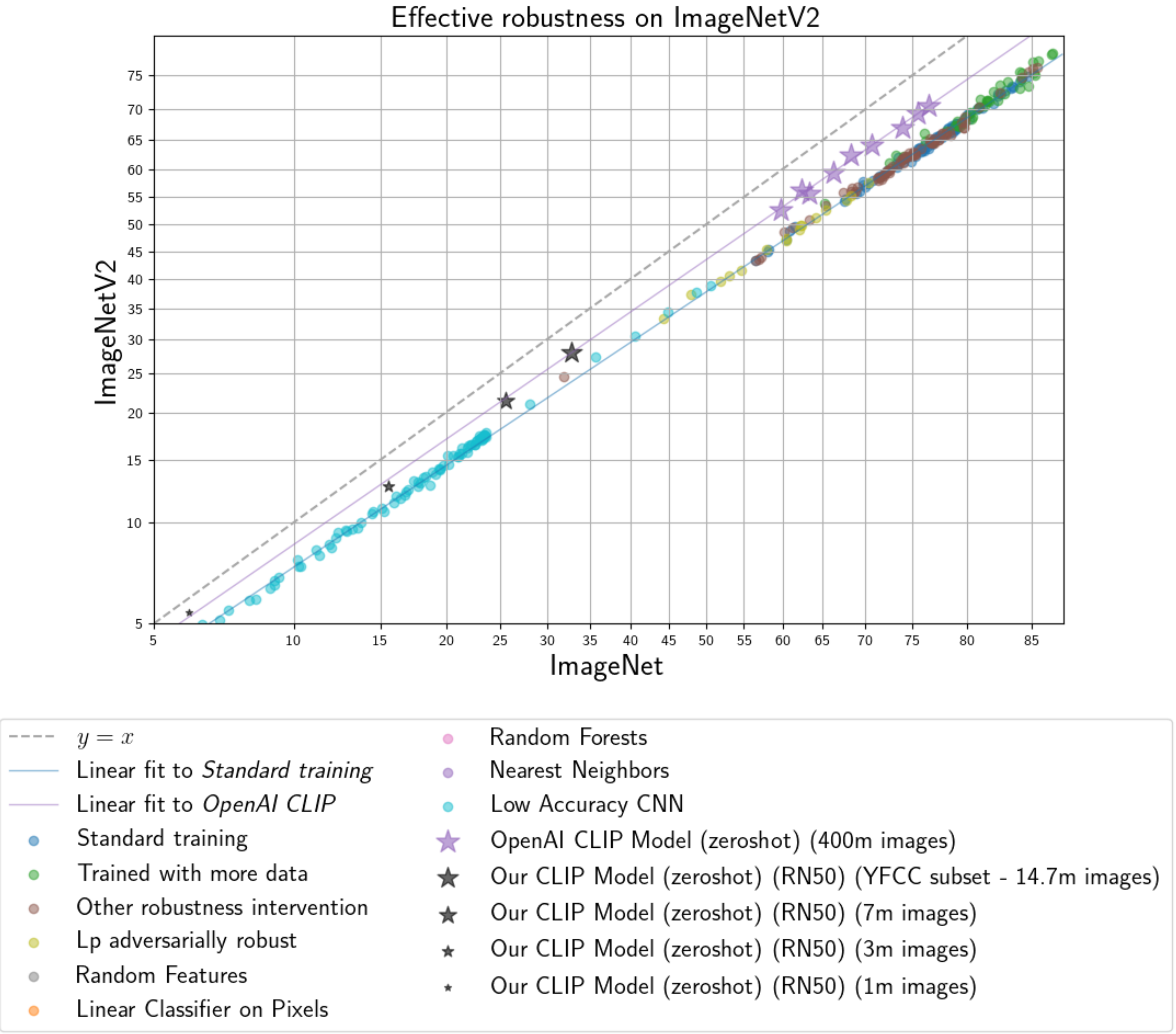
YFCC-15m

- OpenCLIP trained on YFCC-15m **matches the accuracy of CLIP** trained on YFCC-15m and can be trained in a few days on one node
- OpenCLIP trained on YFCC-15m has **similar robustness properties** to full CLIP

YFCC-15m

- OpenCLIP trained on YFCC-15m **matches the accuracy of CLIP** trained on YFCC-15m and can be trained in a few days on one node

- OpenCLIP trained on YFCC-15m has **similar robustness properties** to full CLIP



LAION

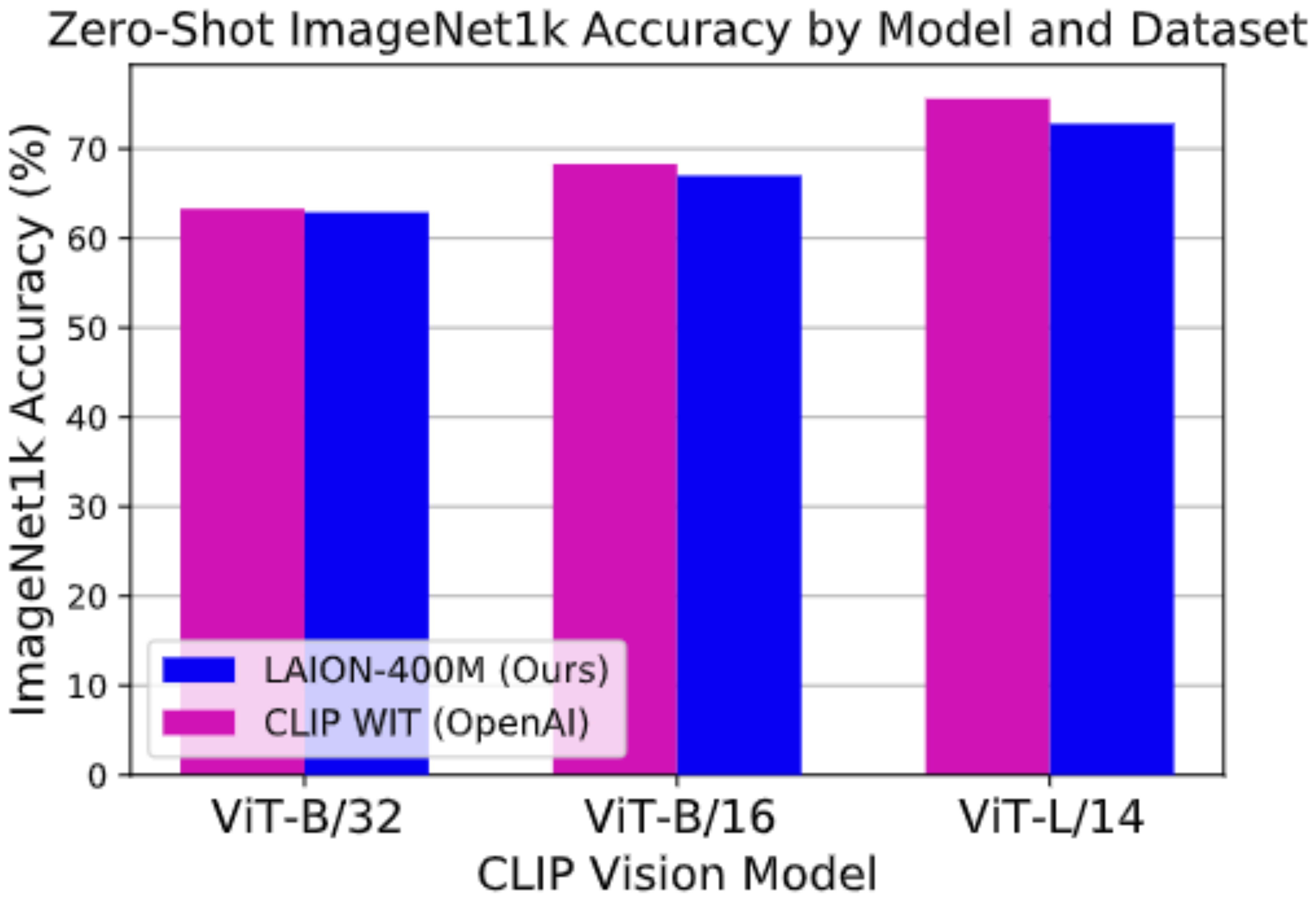
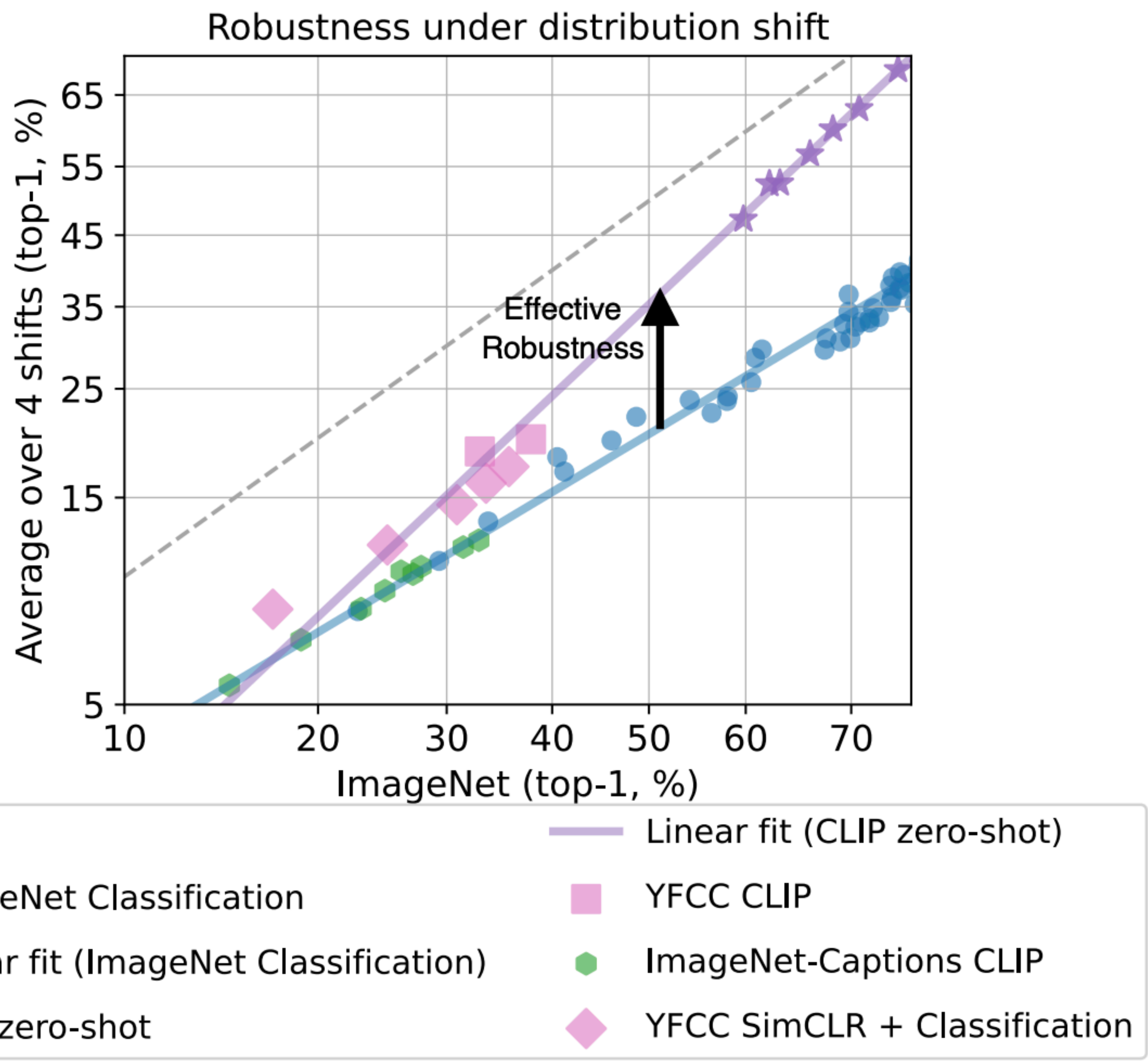


Figure 1: **Zero-Shot Accuracy.** CLIP models trained on LAION-400M (Ours) [52], a previously released preliminary subset of LAION-5B, show competitive zero-shot accuracy compared to CLIP models trained on OpenAI's original training set WIT when evaluated on ImageNet1k.

Data Determines Distributional Robustness in Contrastive Language-Image Pre-training (CLIP)

Alex Fang [1] Gabriel Ilharco [1] Mitchell Wortsman [1] Yuhao Wan [1]
Vaishaal Shankar [2] Achal Dave [2] Ludwig Schmidt [1 3]

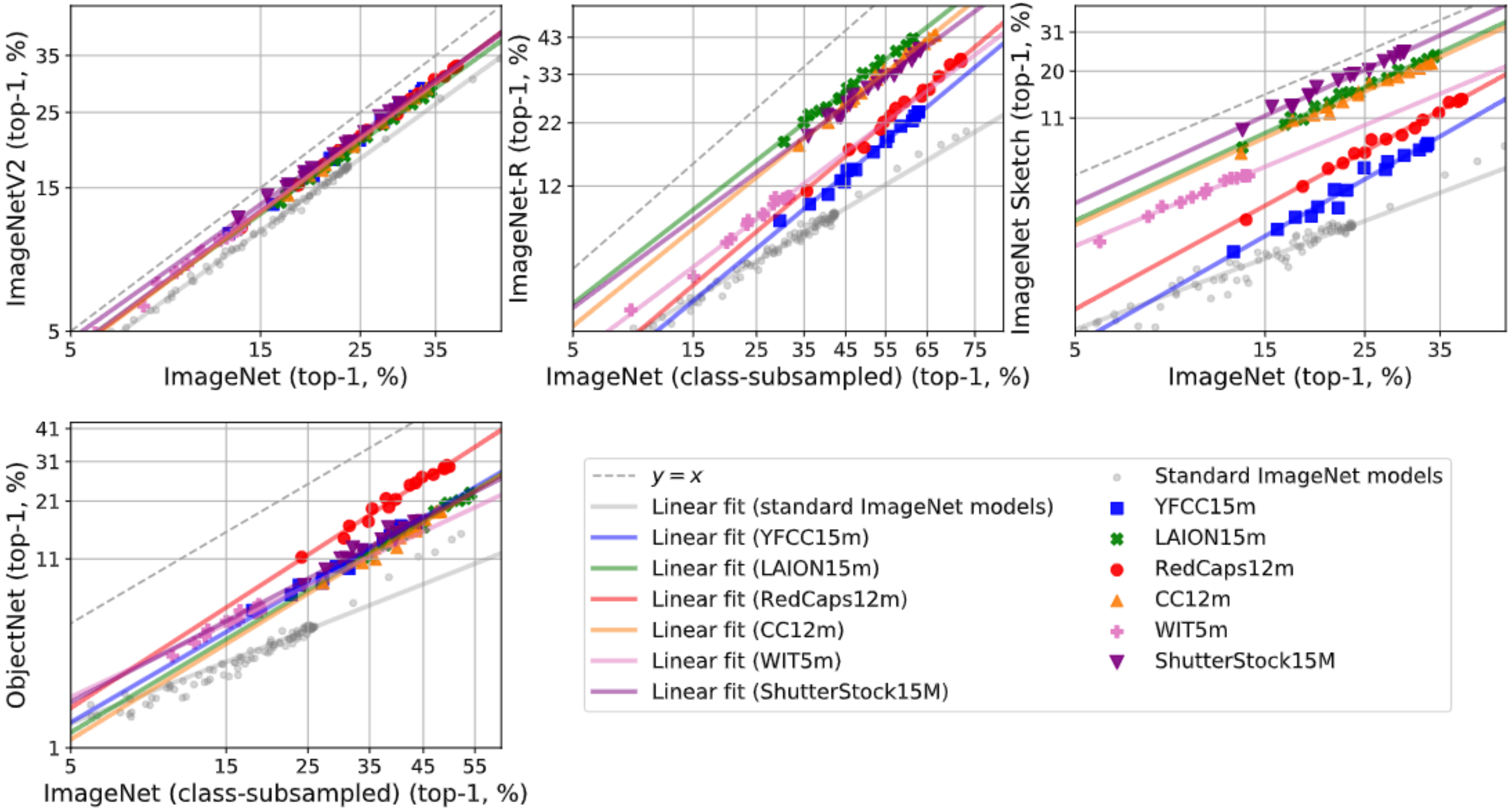


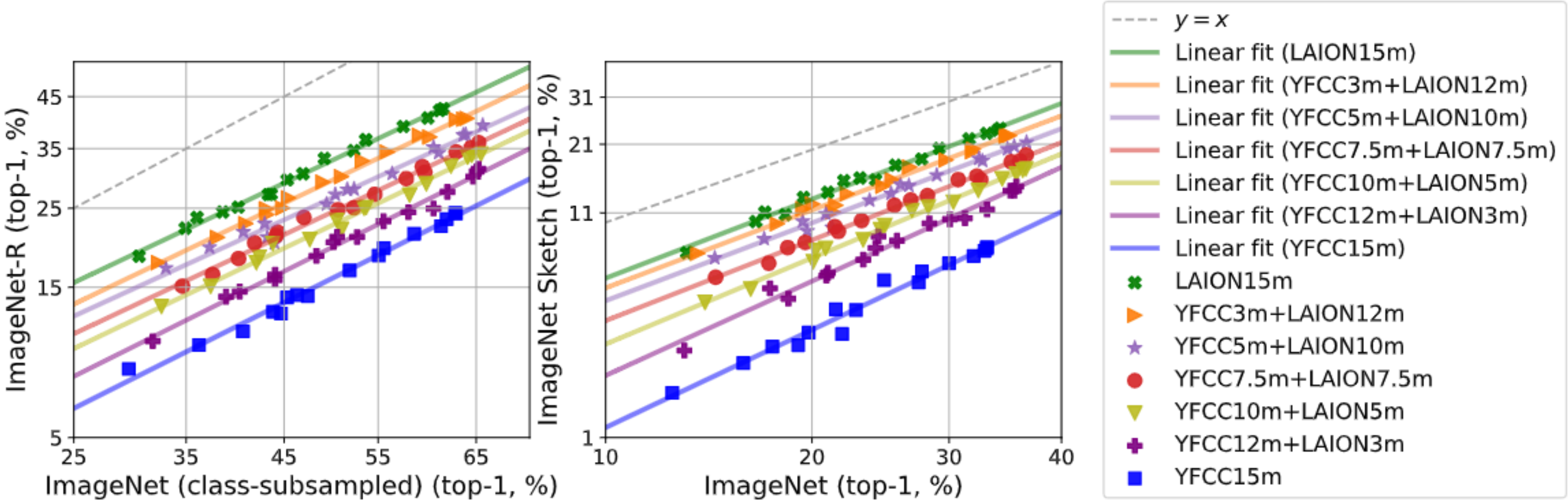
Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP

Thao Nguyen[1] Gabriel Ilharco[1] Mitchell Wortsman[1]
Sewoong Oh[1] Ludwig Schmidt[12]

Abstract

Web-crawled datasets have enabled remarkable generalization capabilities in recent image-text models such as CLIP (Contrastive Language-Image pre-training) or Flamingo, but little is known about the dataset creation processes. In this work, we introduce a testbed of six publicly available data sources—YFCC, LAION, Conceptual Captions, WIT, RedCaps, Shutterstock—to investigate how pre-training distributions induce robustness in CLIP. We find that the performance of the pre-training data varies substantially across distribution shifts, with no single data source dominating. Moreover, we systematically study the interactions between these data sources and find that combining multiple sources does not necessarily yield better models, but rather dilutes the robustness of the best individual data source. We complement our empirical findings with theoretical insights from a simple setting, where combining the training data also results in diluted robustness. In addition, our theoretical model provides a candidate explanation for the success of the CLIP-based data filtering technique recently employed in the LAION dataset. Overall our results demonstrate that simply gathering a large amount of data from the web is not the most effective way to build a pre-training dataset for robust generalization, necessitating further study into dataset design.


ImageNetV2 (top-1, %)
ImageNet (top-1, %)
ImageNet-R (top-1, %)
ImageNet (class-subsampled) (top-1, %)
ImageNet Sketch (top-1, %)
ImageNet (top-1, %)
ObjectNet (top-1, %)
ImageNet (class-subsampled) (top-1, %)
$y = x$
Linear fit (standard ImageNet models)
Linear fit (YFCC15m)
Linear fit (LAION15m)
Linear fit (RedCaps12m)
Linear fit (CC12m)
Linear fit (WIT5m)
Linear fit (ShutterStock15M)
Standard ImageNet models
YFCC15m
LAION15m
RedCaps12m
CC12m
WIT5m
ShutterStock15M



ImageNet-R (top-1, %)
5
15
25
35
45
25
35
45
55
65
ImageNet (class-subsampled) (top-1, %)
ImageNet Sketch (top-1, %)
1
11
21
31
10
20
30
40
ImageNet (top-1, %)
$y = x$
Linear fit (LAION15m)
Linear fit (YFCC3m+LAION12m)
Linear fit (YFCC5m+LAION10m)
Linear fit (YFCC7.5m+LAION7.5m)
Linear fit (YFCC10m+LAION5m)
Linear fit (YFCC12m+LAION3m)
Linear fit (YFCC15m)
LAION15m
YFCC3m+LAION12m
YFCC5m+LAION10m
YFCC7.5m+LAION7.5m
YFCC10m+LAION5m
YFCC12m+LAION3m
YFCC15m

OpenCLIP:
an open source implementation of CLIP

Gabriel Ilharco*, **Mitchell Wortsman***, Cade Gordon*, Ross Wightman*, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, John Miller, Hongseok Namkoong, Hannaneh Hajishirzi, Ali Farhadi, Ludwig Schmidt