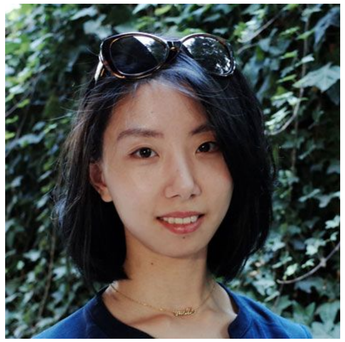
Data Determines Distributional Robustness in CLIP

Alex Fang

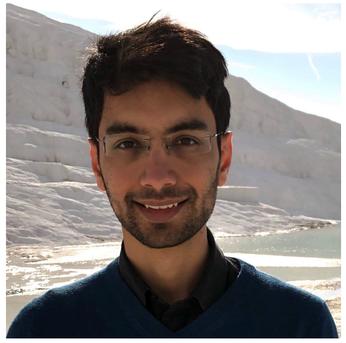
Gabriel Ilharco

Mitchell Wortsman



Yuhao Wan

Vaishaal Shankar



Achal Dave

Ludwig Schmidt

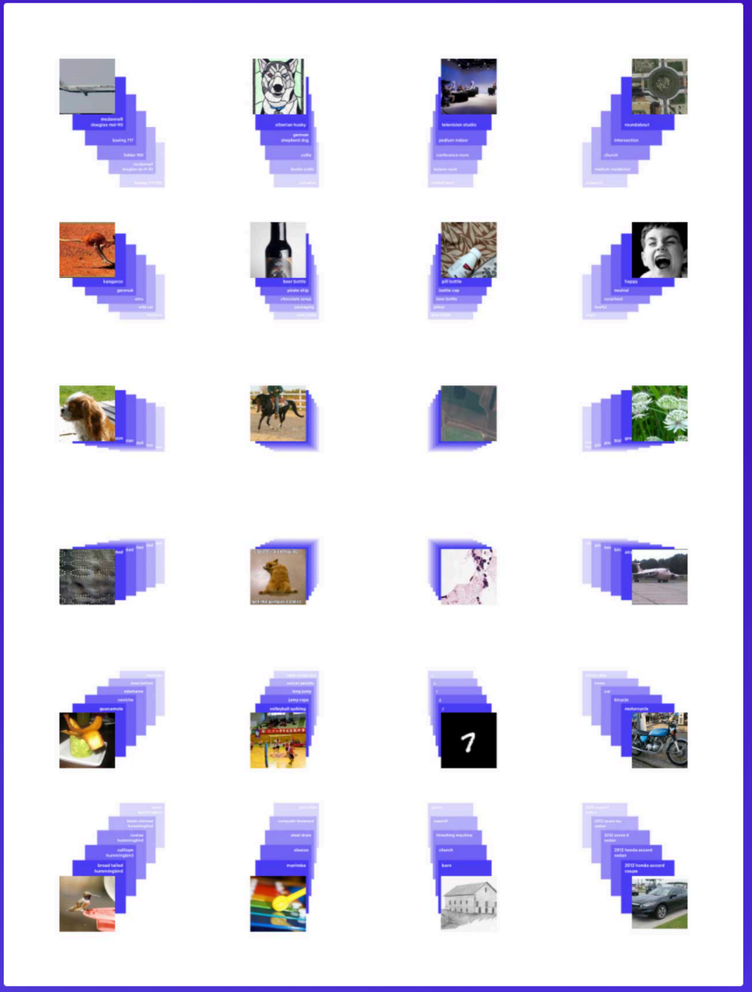
UNIVERSITY of WASHINGTON

amazon

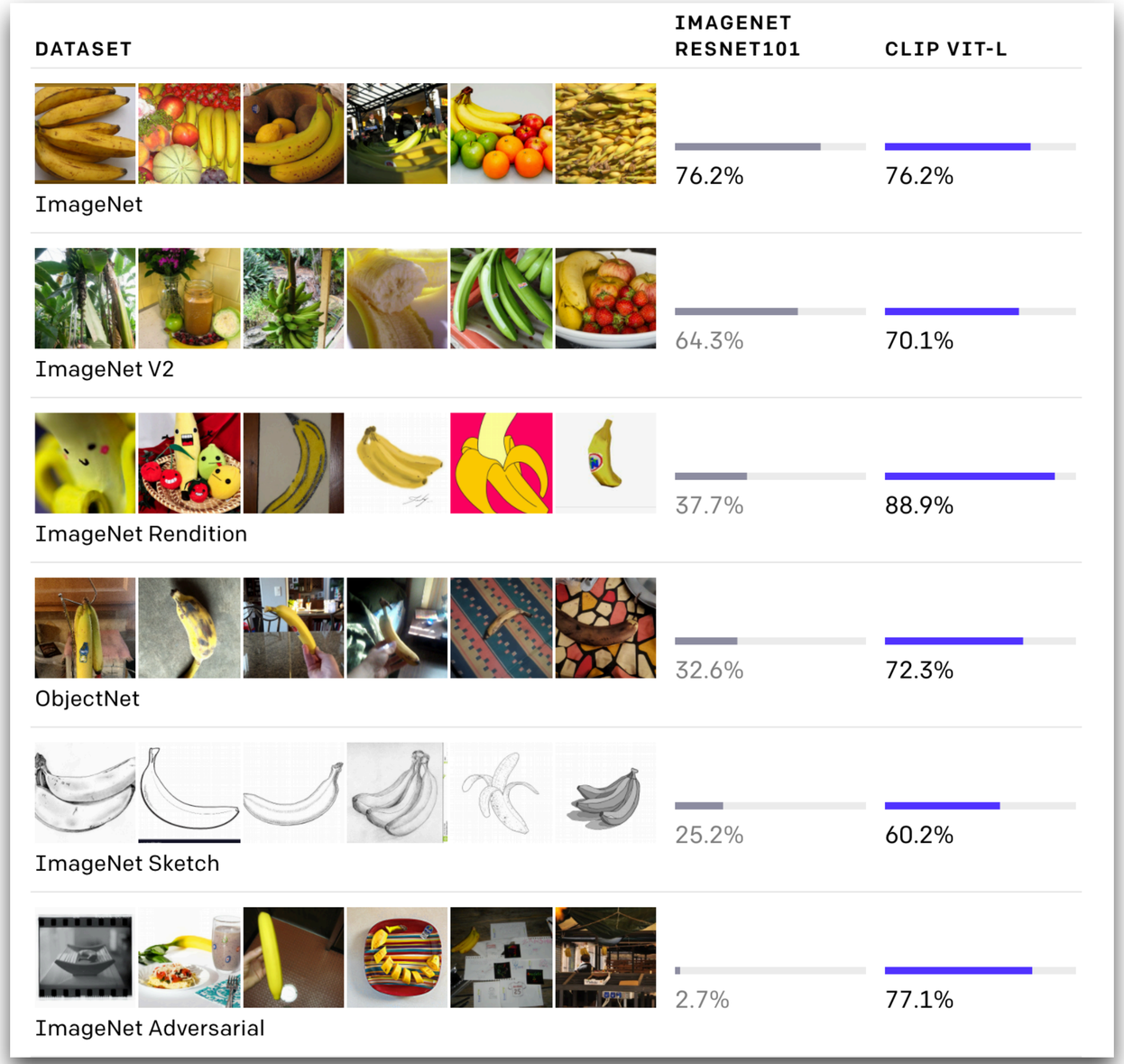
AI2 Allen Institute for AI

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



Effective robustness

+6%

+51%

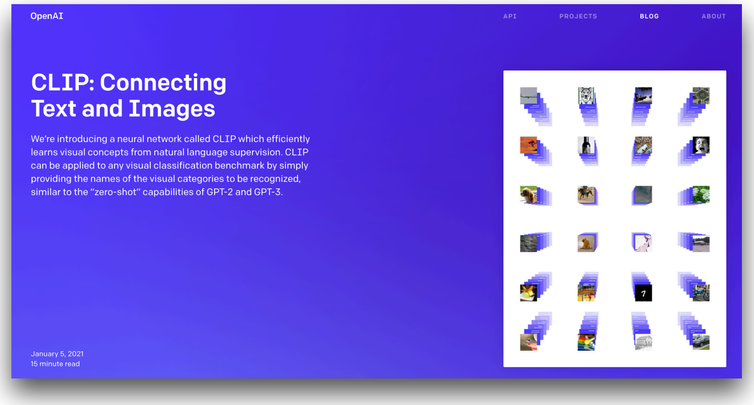
+40%

+35%

+74%

Very large improvements in out-of-distribution robustness.

How to isolate source of robustness?



CLIP

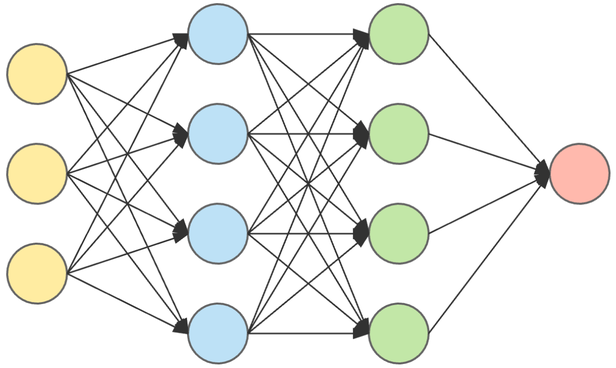
More Robust to ImageNet distribution shifts

Uses Contrastive Loss to optimize

Uses Language Supervision at training time

400M examples

Trained on private image-caption training-set



Standard ImageNet Models (ResNet50 etc..)

Not Robust to ImageNet distribution shifts

Uses cross entropy loss to optimize

No language supervision at training time

1.2 Million training examples

Trained on ILSVRC 2012 training set

Hypotheses for CLIP's Robustness

- Large training set size
- Training distribution
- Language supervision at training time
- Language supervision at test time (prompts)
- Use of contrastive loss functions

Hypotheses for CLIP's Robustness

- Large training set size
- Training distribution
- Language supervision at training time
- Language supervision at test time (prompts)
- Use of contrastive loss functions

Main Findings

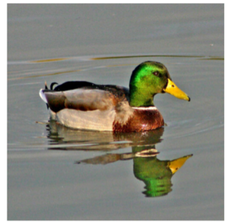
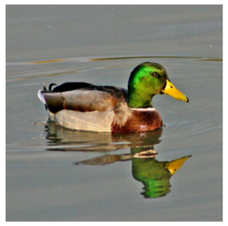
- Changing the training distribution affects robustness
- Presence of language supervision does **not** affect robustness

YFCC-15M

- 15M Subset of publicly available YFCC-100M dataset explicitly used in OpenAI CLIP training

- Dataset of Image-Caption pairs sourced from Flickr

Experimental Setup

	Supervised (Without Language)	Contrastive (With Language)
ImageNet	ImageNet Standard (Baseline) 	ImageNet-Captions + CLIP (Us)  Title: `Reflected Duck` Description: Tags: `lake`, `water`, `bird` [6 tags omitted]
YFCC	YFCC Hardmatch + NoCLIP (Us) 	YFCC-15M CLIP (Baseline)  one of the most dramatic mountain ranges I have seen

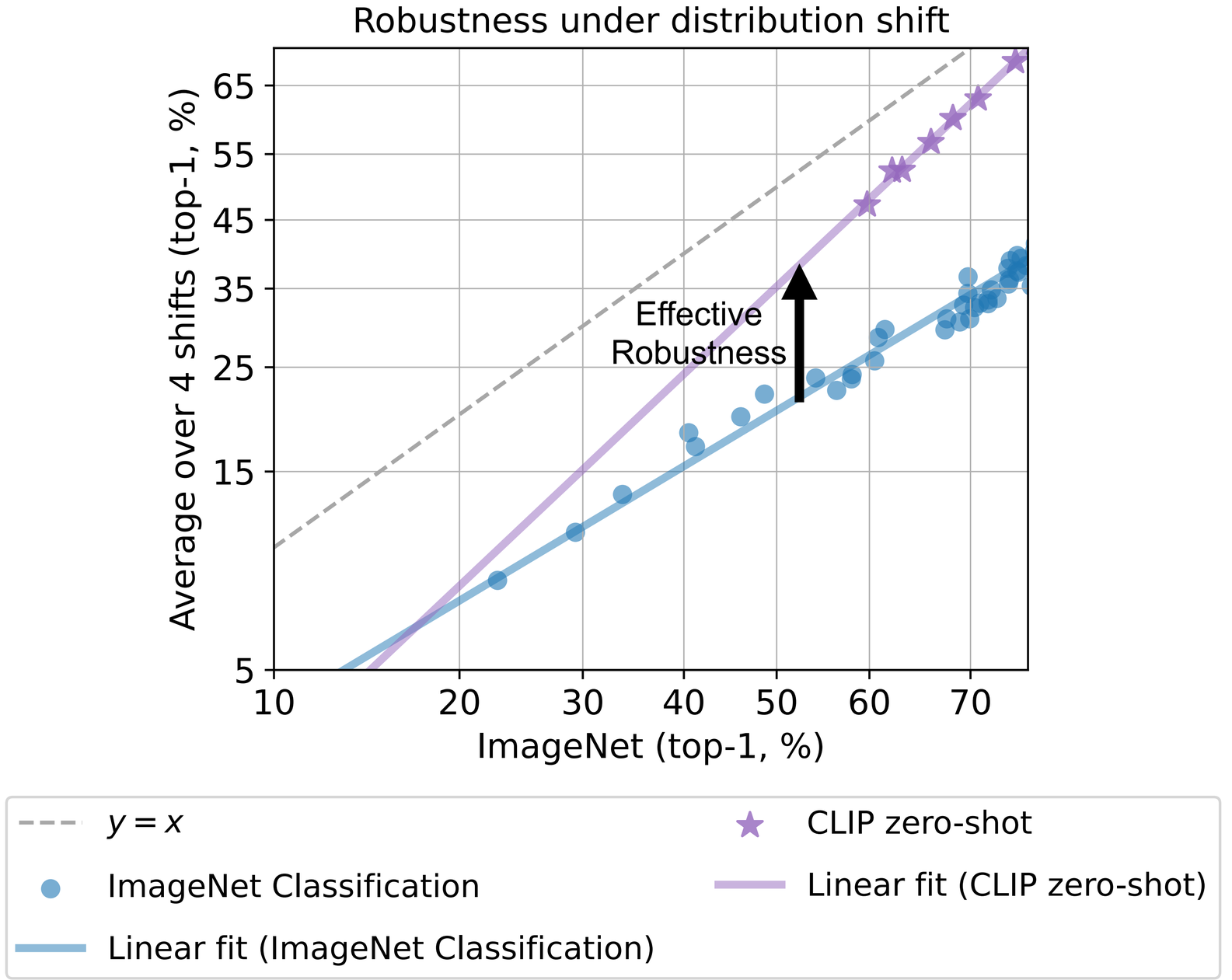
ImageNet-Captions Examples

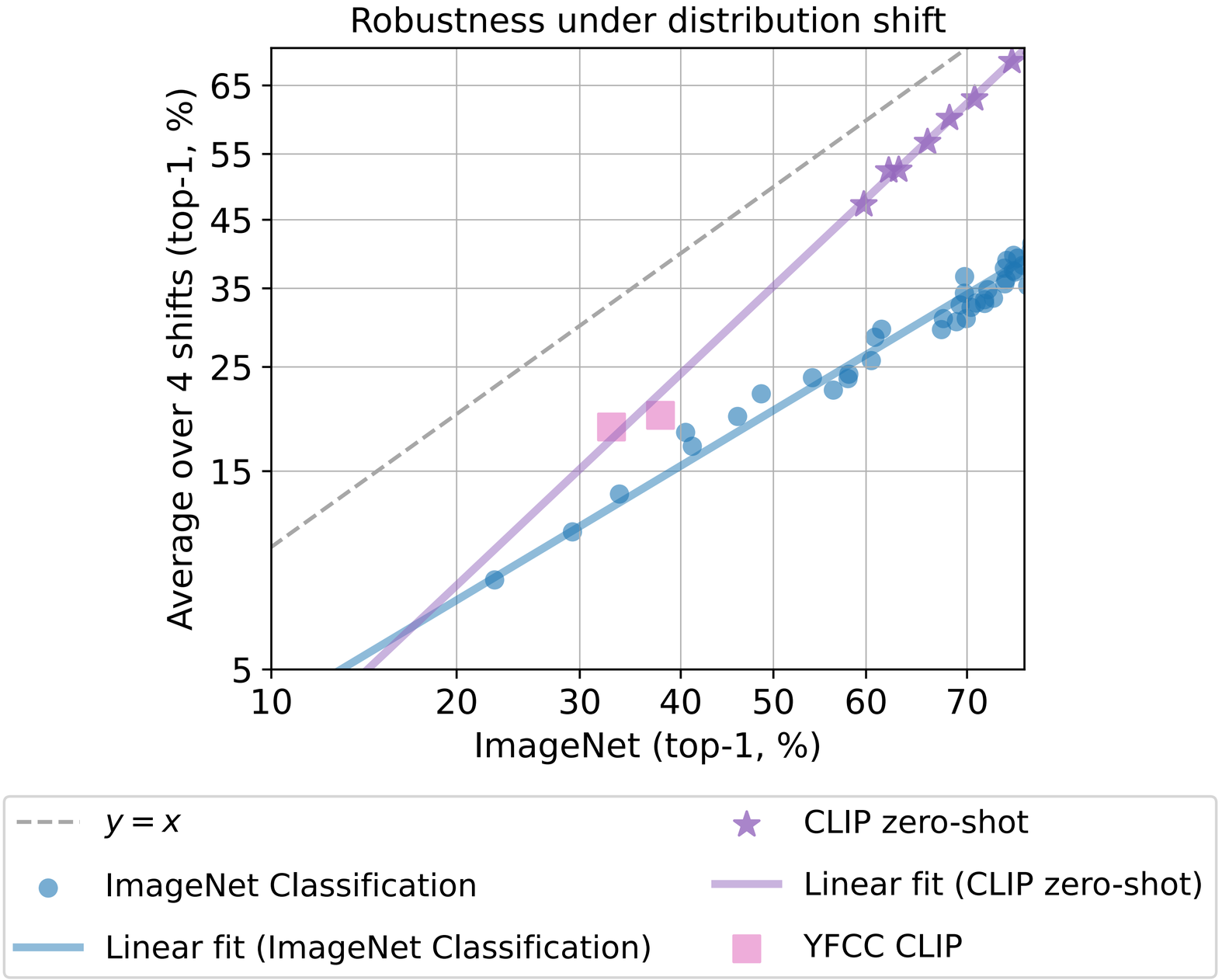
Title: `Reflected Duck`
Description:
Tags: `lake`, `water`, `bird` **[6 tags omitted]**

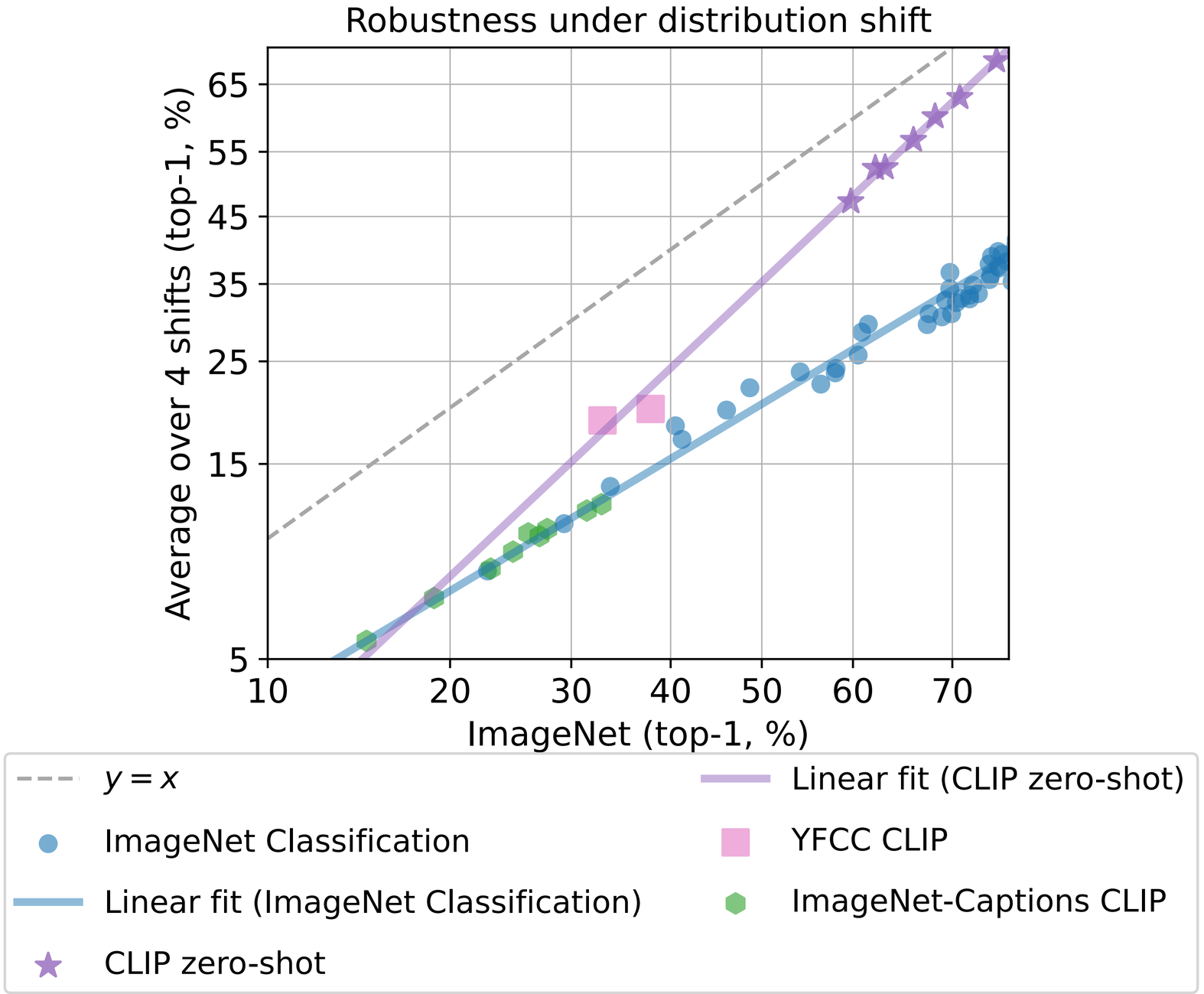
Title: `SILENT ROCKER`
Description: `MOSE'S MOTHER HAS LEFT THE BuILDING` **[10 words omitted]**
Tags: `rockingchair`, `rock`, `chair` **[2 tags omitted]**

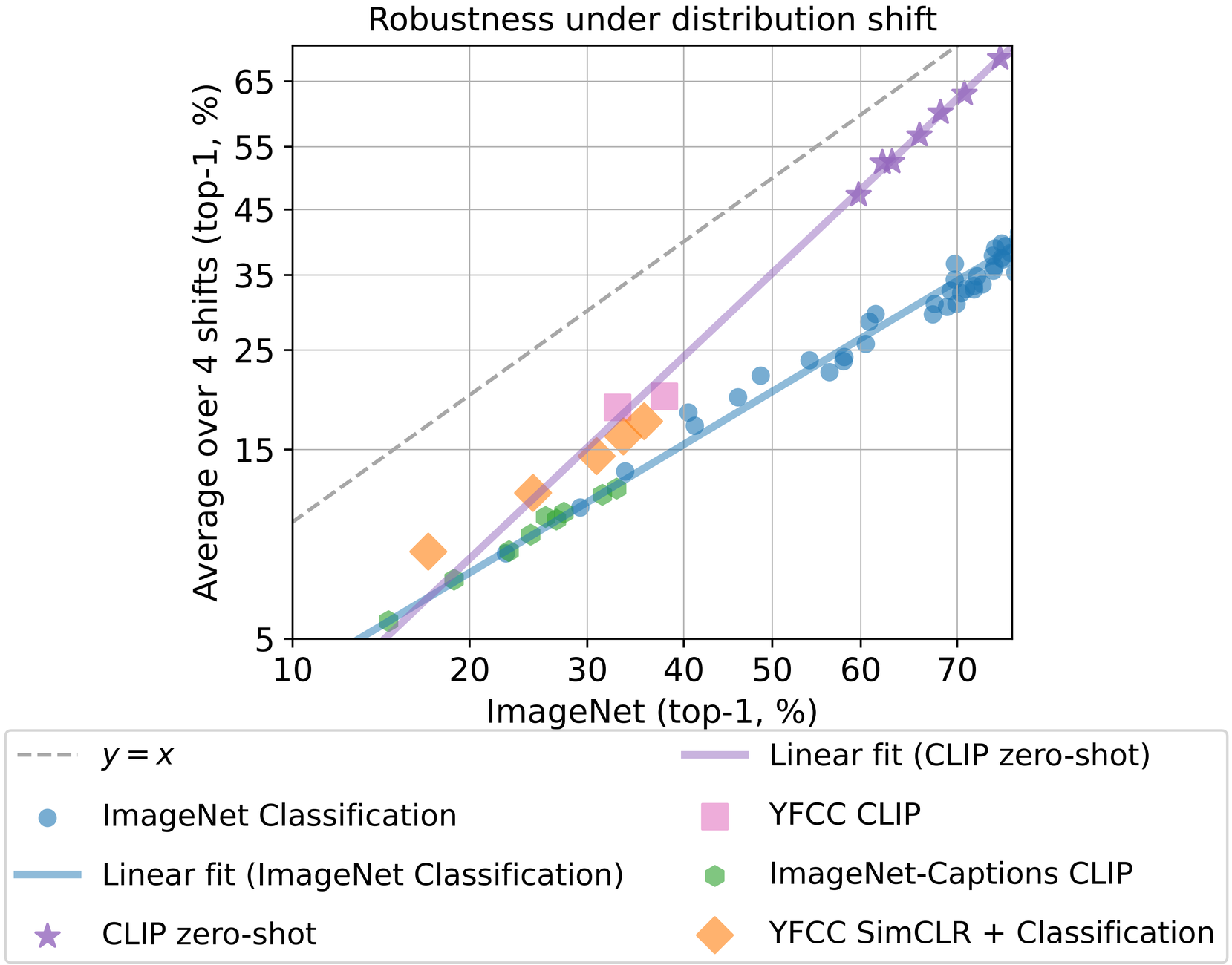
Title: `A Phone Call at Night`
Description: `I might have a thing with telephones` **[174 words omitted]**
Tags: `phone`, `telephone`, `blackandwhite` **[7 tags omitted]**

- Large portions of ImageNet are still on Flickr
- Queried text data from Flickr API, restrict to ILSVRC 2012, run image deduplication (463,622 images)


Robustness under distribution shift
Average over 4 shifts (top-1, %)
65
55
45
35
25
15
5
10
20
30
40
50
60
70
ImageNet (top-1, %)
Effective
Robustness
y = x
ImageNet Classification
Linear fit (ImageNet Classification)
CLIP zero-shot
Linear fit (CLIP zero-shot)

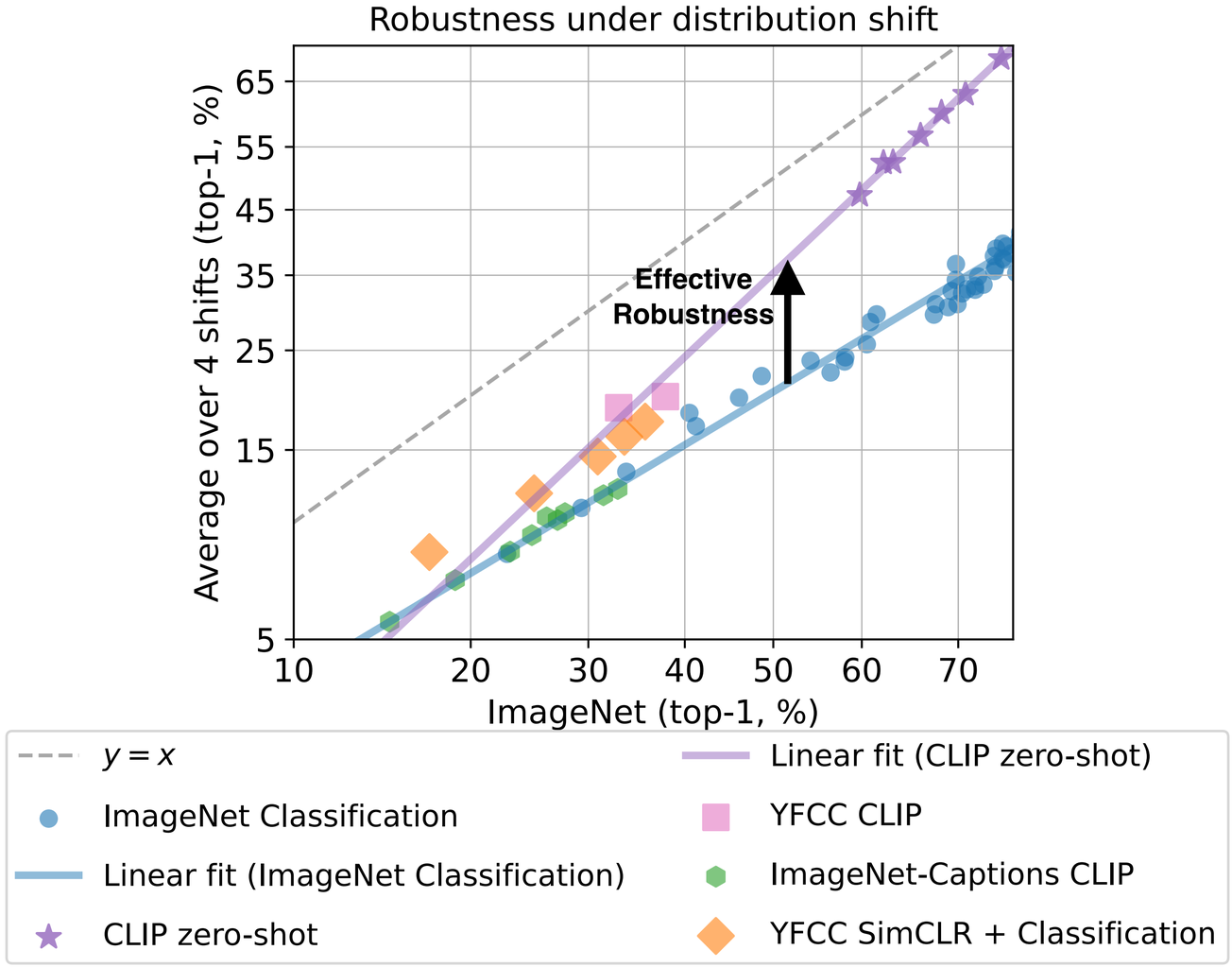

Robustness under distribution shift
Average over 4 shifts (top-1, %)
65
55
45
35
25
15
5
10
20
30
40
50
60
70
ImageNet (top-1, %)
y = x
ImageNet Classification
Linear fit (ImageNet Classification)
CLIP zero-shot
Linear fit (CLIP zero-shot)
YFCC CLIP


Robustness under distribution shift
Average over 4 shifts (top-1, %)
65
55
45
35
25
15
5
10
20
30
40
50
60
70
ImageNet (top-1, %)
y = x
ImageNet Classification
Linear fit (ImageNet Classification)
CLIP zero-shot
Linear fit (CLIP zero-shot)
YFCC CLIP
ImageNet-Captions CLIP

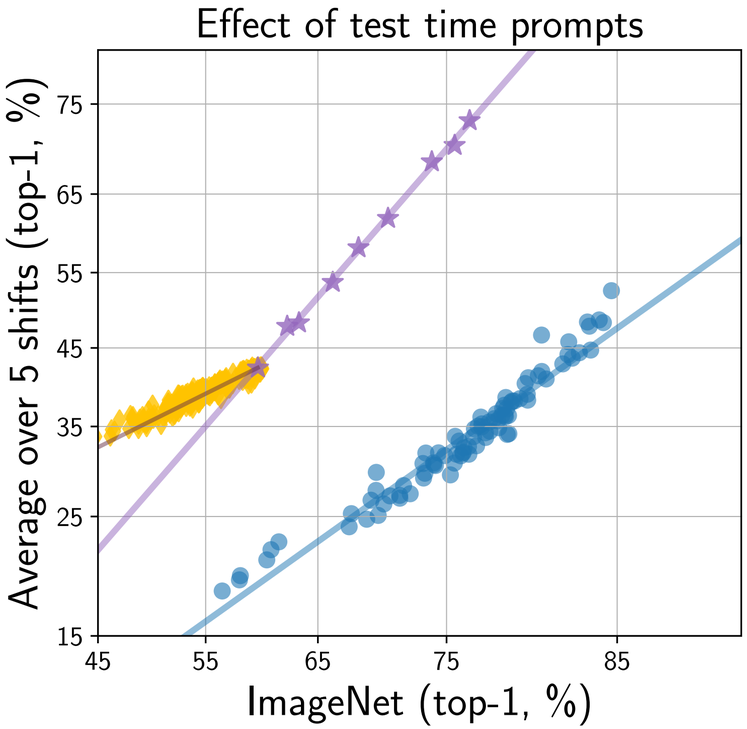

Robustness under distribution shift
Average over 4 shifts (top-1, %)
65
55
45
35
25
15
5
10
20
30
40
50
60
70
ImageNet (top-1, %)
y = x
ImageNet Classification
Linear fit (ImageNet Classification)
CLIP zero-shot
Linear fit (CLIP zero-shot)
YFCC CLIP
ImageNet-Captions CLIP
YFCC SimCLR + Classification

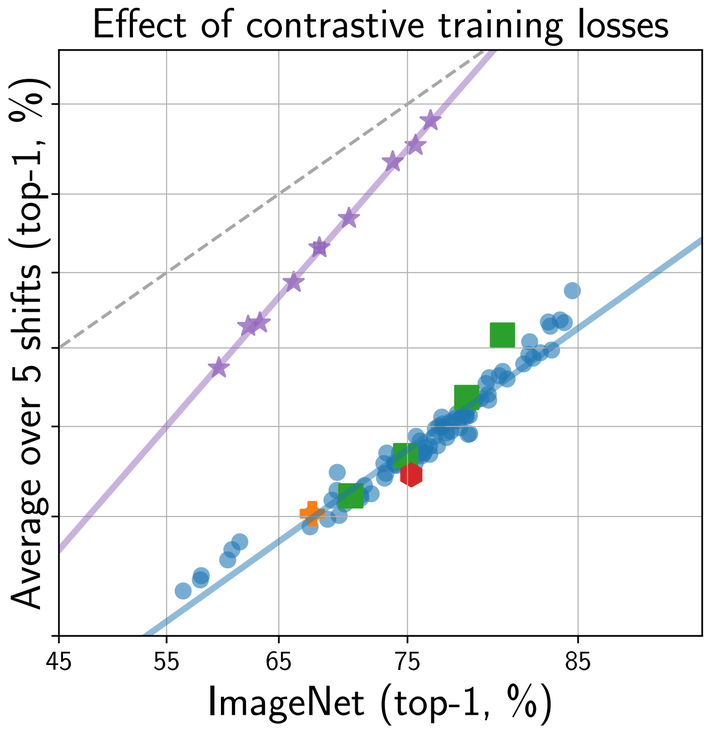
Results

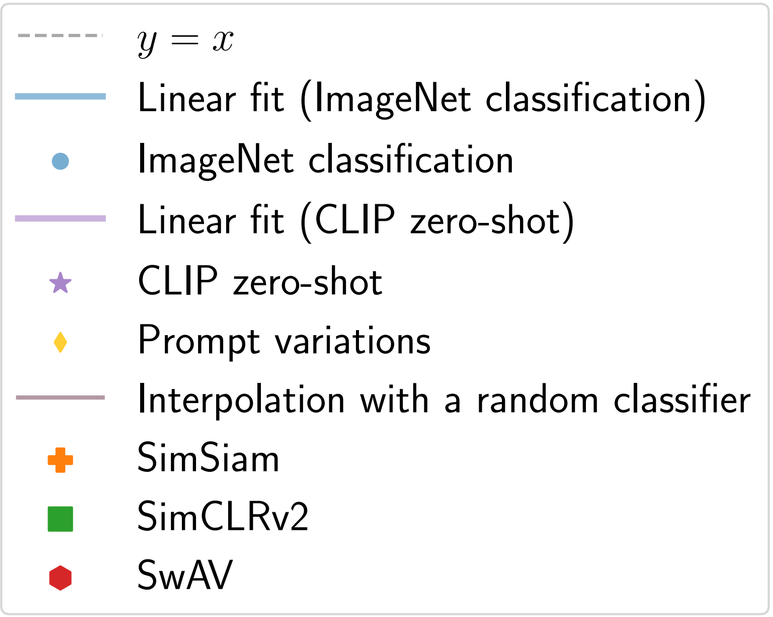
- Contrastively trained model on Imagenet-Captions is **not** robust
- Classification trained model on YFCC-15M **is** robust



Prompts and Contrastive Loss







Conclusion

Hypotheses:

- ~~Large training set size~~
- **Training distribution**
- ~~Language supervision at training time~~
- ~~Language supervision at test time (prompts)~~
- ~~Use of contrastive loss functions~~
- ~~Model architecture and size~~ (Radford et al.)

Promising future research directions in dataset design and analysis