Digest

Contrastive Language-Image Pre-training (CLIP)

Paper: Learning transferable visual models from natural language supervision

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever

ICML (2021)

Digest by Samuel Albanie, April 2022

Outline

- Motivation
- CLIP: Data and Method
- Experiments
- Data Overlap Analysis
- Limitations
- Broader Impacts
- Related Work
- Summary

Motivation

Flexibility

Traditional computer vision systems are trained with a fixed set of predetermined object categories.

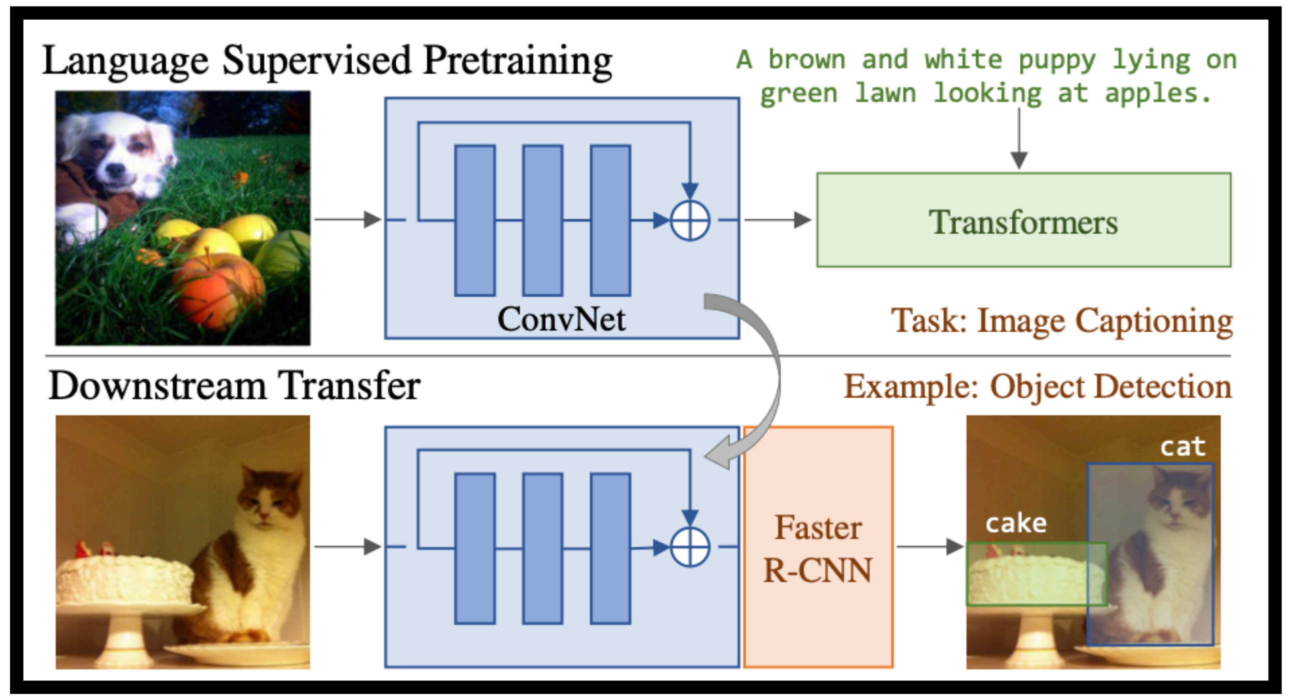
This limits their flexibility: each time we encounter a new visual concept, we need to retrain the model with labelled examples of this concept.

Can we train a vision model to work "zero-shot"?

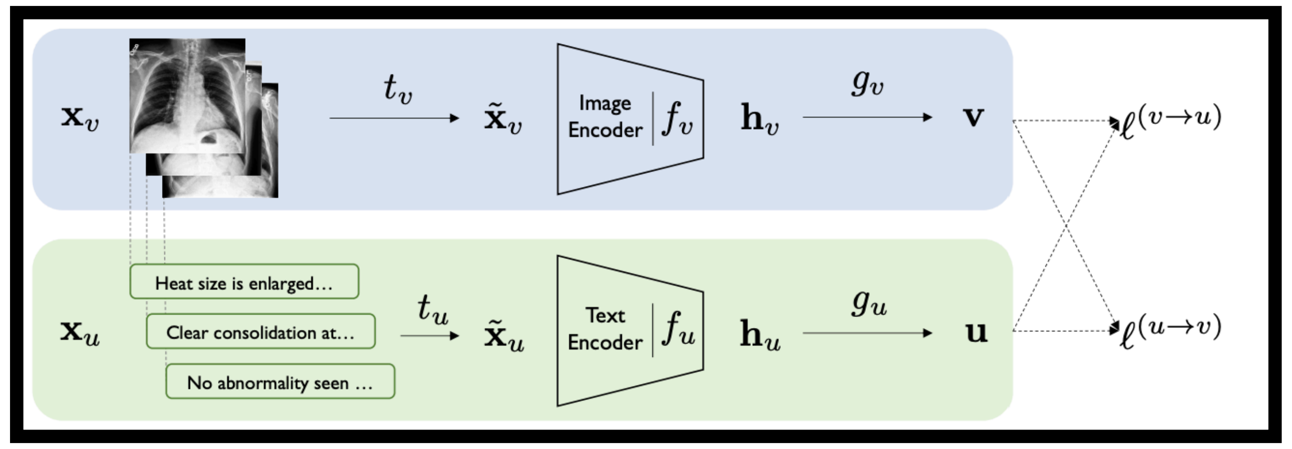
Natural language supervision

Prior works have shown that learning from descriptions rather than fixed labels can be very data efficient.

VirTex demonstrated data efficiency of captioning.



ConVIRT showed data efficiency of contrastive training.



Can we leverage data efficiency of natural language?

Scale

NLP systems have benefited tremendously from scale.

T5 (Raffel et al., 2019), GPT-3 (Brown et al., 2020) etc. showed zero-shot transfer scale benefits.

Web scale supervision seems to surpass manual curation for NLP datasets.

Scaling up manual annotation of images is expensive.

Thanks to alt-text, there are large quantities of images with text descriptions online.

Can we scale up vision training with web text?

Reference/Image credits: A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)
(VirTex) K. Desai and J. Johnson, "Virtex: Learning visual representations from textual annotations", CVPR (2021)
(ConVIRT) Y. Zhang, H. Jiang, Y. Miura, C. Manning and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text", arXiv (2020)
C. Raffel, et al., "Exploring the limits of transfer learning with a unified text-to-text transformer", JMLR (2019)
T. Brown et al., "Language models are few-shot learners", NeurIPS (2020)

Outline

- Motivation
- CLIP: Data and Method
- Experiments
- Data Overlap Analysis
- Limitations
- Broader Impacts
- Related Work
- Summary

Building blocks

Creating a large enough dataset

Prior work learning with natural language has used datasets of limited scale

- MS COCO and Visual Genome (both $\mathcal{O}(100K)$ images)
- YFC100M ($\mathcal{O}(100M)$ images with noisy metadata, so $\mathcal{O}(15M)$ after filtering)

By contrast, strong vision classifiers (Mahajan et al., 2018) have benefited from training on $\mathcal{O}(3B)$ images.

To assess whether natural language works at scale, a new dataset is collected.

The dataset is built by searching for (image, text) pairs with 500K queries.

The queries are formed from:

- words occurring at least 100 times in English Wikipedia
- bi-grams (with high mutual information) augment the initial queries
- names of wikipedia articles above a search volume threshold
- WordNet sysnets

Approximate class balancing: include up to 20K (image, text pairs) per query.

The resulting WebImageText (WIT) dataset contains 400M (image, text) pairs.

Choosing an efficient pre-training method

The strongest computer vision systems use significant computation to train:

- Mahajan et al. (2018) use 19 years of GPU time to train on instagram
- Xie et al. (2020) use 33 years of TPUv3 time to train Noisy Student

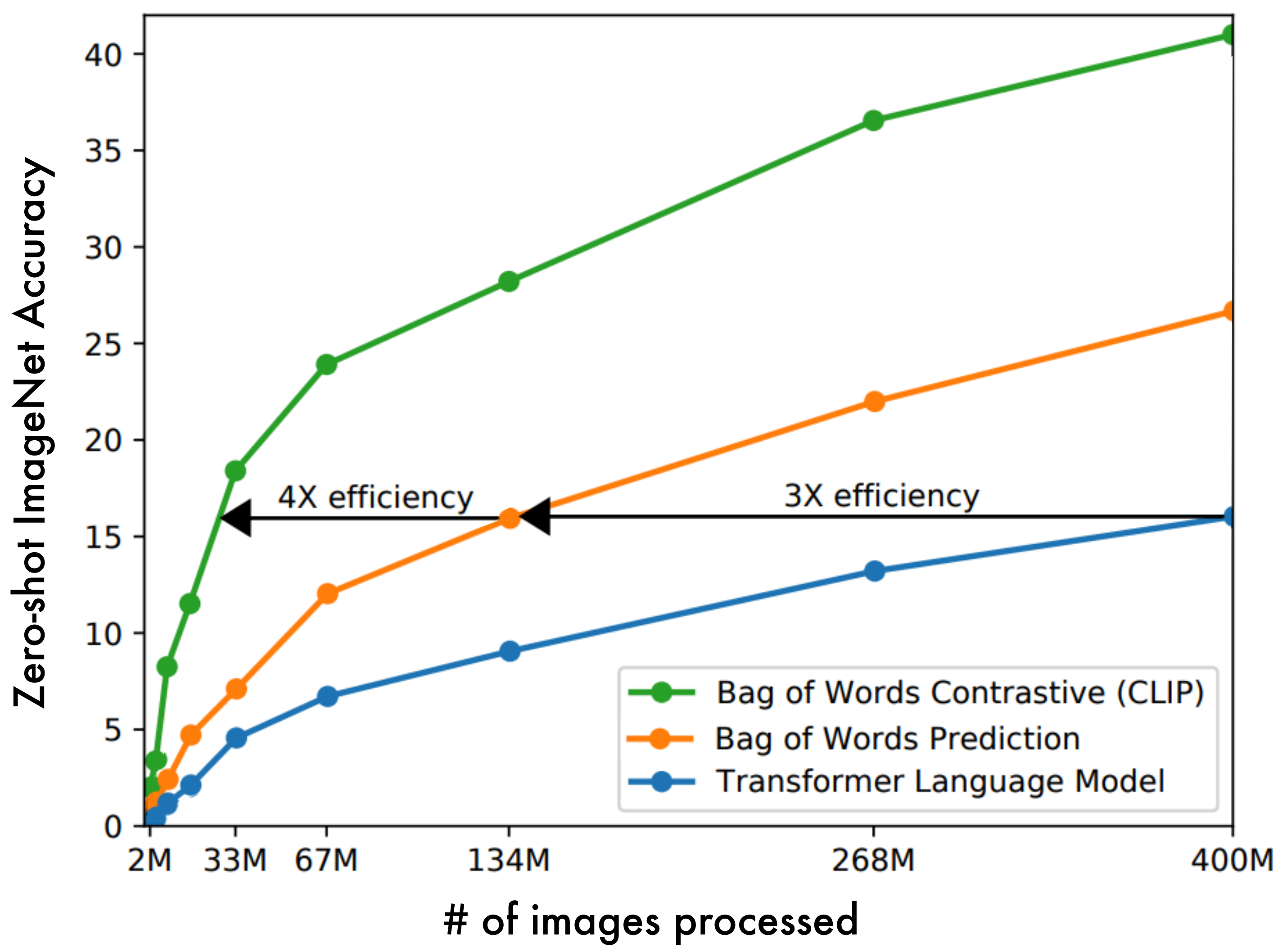
For large-scale pre-training, efficiency is a key consideration.

Baseline: captioning system inspired by VirTex

Improvement: bag of words prediction

CLIP: contrastive image-text matching

Given N image-text pairs, CLIP predicts which of $N \times N$ possible pairs is valid.



Reference/Image credits: T. Lin et al., "Microsoft coco: Common objects in context", ECCV (2014)
R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations", IJCV (2017)
B. Thomee et al., "YFCC100M: The new data in multimedia research", *Communications of the ACM* (2016)
D. Mahajan et al., "Exploring the limits of weakly supervised pretraining", ECCV (2018)
G. A. Miller, "WordNet: a lexical database for English", *Communications of the ACM* (1995)
Q. Xie, et al., "Self-training with noisy student improves imagenet classification", CVPR (2020)
(VirTex) K. Desai, J. Johnson, "Virtex: Learning visual representations from textual annotations", CVPR (2021)

Contrastive Pre-training

Multi-modal embedding

CLIP trains an image and text encoders to maximise cosine similarities of the N valid pairs within each batch (and minimises those of invalid pairings).

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Pseudocode

Training details

Since WIT is large (low risk of overfitting) both encoders are trained from scratch.

Linear projections (rather than non-linear) used between the representations and the shared embedding space, since no difference was observed during training.

Simple image data augmentation: use a random square crop from resized images.

The (log-parameterised) softmax temperature, τ, is learned during training.

Models

Image encoders:

Scaling: equal compute budget to width, depth, resolution

1. ResNet-50 (He et al., 2015, He et al. 2019, Zhang 2019)

Replace Global Average Pooling with attention pooling (in style of Transformer layer) where query is conditioned on the global average pooled feature.

2. Vision Transformer (Dosovitskiy et al., 2020) with additional layer norm

Text encoder:

Scaling: only scale up width proportional to ResNet

Text transformer (Vaswani et al., 2017) trained on BPE text with 49K vocab size

Sentences were capped to 76 tokens and bracketed with [SOS] and [EOS] tokens.

[EOS] embedding at the last transformer layer is used as the text representation.

Reference/Image credits:

A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)

K. He et al., "Deep residual learning for image recognition", CVPR (2016)

T. He et al., "Bag of tricks for image classification with convolutional neural networks", CVPR (2019)

R. Zhang, "Making convolutional networks shift-invariant again", ICML (2019)

A. Vaswani et al., "Attention is all you need", NeurIPS (2017)

A. Dosovitskiy, et al. "An image is worth 16x16 words: Transformers for image recognition at scale", ICLR (2021)

(Scaling) M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks", ICML (2019)

Training - nuts and bolts

CLIP model details

Images	RN50	RN101	RN50x4	RN50x16	RN50x64	ViT-B/32	ViT-B/16	ViT-L/14	ViT-L/14-336px
Resolution	224	224	228	384	448	224	224	224	336
Embedding	1024	512	640	768	1024	512	512	768	768
Text	Transformer	Transformer	Transformer	Transformer	Transformer	Transformer	Transformer	Transformer	Transformer
Width	512	512	640	768	1024	512	512	768	768
Heads	8	8	10	12	16	8	8	12	12

Trained with FixRes (Touvron et al., 2019)

All text transformers have 12 layers.

Best performing CLIP model

CLIP optimisation details

Models were trained for 32 epochs with AdamW (Kingma and Ba, 2014; Loshchilov and Hutter, 2017)

Learnable temperature initialised to the equivalent of 0.07 (Wu et al., 2018) and clipped to prevent logit scaling more than x100 for stability.

A large minibatch size of 32,768 was used in combination with mixed-precision training (Micikevicius et al. 2018) for efficiency.

Gradient checkpointing (Griewank and Walther, 2000) was also used to reduce memory consumption.

The largest ResNet, RN50x64, took 18 days to train on 592 V100 GPUs

The largest Vision Transformer, ViT-L/14, took 12 days on 256 V100 GPUs.

References
H. Touvron et al., "Fixing the train-test resolution discrepancy: FixEfficientNet", arxiv (2020)
D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *ICLR* (2015)
I. Loshchilov and F. Hutter, "Decoupled weight decay regularization", arXiv (2017)
Z. Wu et al., "Unsupervised feature learning via non-parametric instance discrimination", CVPR (2018)
P. Micikevicius et al., "Mixed precision training", ICLR (2018)
A. Griewank and A. Walther, "Algorithm 799: revolve: an implementation of checkpointing for the reverse or adjoint mode of computational differentiation", *TOMS (2000)*

Outline

- Motivation
- CLIP: Data and Method
- Experiments
- Data Overlap Analysis
- Limitations
- Broader Impacts
- Related Work
- Summary

Experiments

1. Zero-shot transfer

Zero-shot learning in computer vision typically refers to the task of generalising to unseen object categories (Lampert et al., 2009).

In this work, the term is used to mean generalisation to unseen datasets (a proxy for unseen tasks).

Rationale: zero-shot transfer can be thought of as assessing the task learning ability of a model:

A dataset evaluates performance on a task on a specific distribution

The zero-shot transfer focus is inspired by works illustrating task learning in NLP.

Notable example: the Wikipedia article generation model of Liu et al. (2018), which learned to reliably transliterate names between languages as an "unexpected side-effect".

rohit viswanath (hindi : रोहित विशानाथ) is an indian politician and a member of the 16th

Note: the authors note that this metaphor of *datasets-as-tasks* is not always clear cut.

Many vision datasets were introduced as benchmarks for generic image classifiers, not specific tasks:

SVHN (task: street number transcription, distribution: Google Street View photos)

CIFAR-10 (task: ?, distribution: TinyImages)

Zero-shot transfer has had limited attention in computer vision - an exception is Visual N-Grams (Li et al., 2017), compared to in the experiments.

2. Representation learning

Evaluate visual representation quality via linear probes:

Linear (rather than non-linear) probes are used to avoid the introduction of additional hyperparameters and cost.

3. Robustness

Assess robustness to "natural distribution shifts" studied by Taori et al. (2020).

References/Image credits

C. H. Lampert et al., "Learning to detect unseen object classes by between-class attribute transfer", CVPR (2009)

P. J. Liu et al., "Generating wikipedia by summarizing long sequences", *ICLR (2018)*

(SVHN) Y. Netzer et al., "Reading Digits in Natural Images with Unsupervised Feature Learning", (2011)

(CIFAR-10) A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images", (2009)

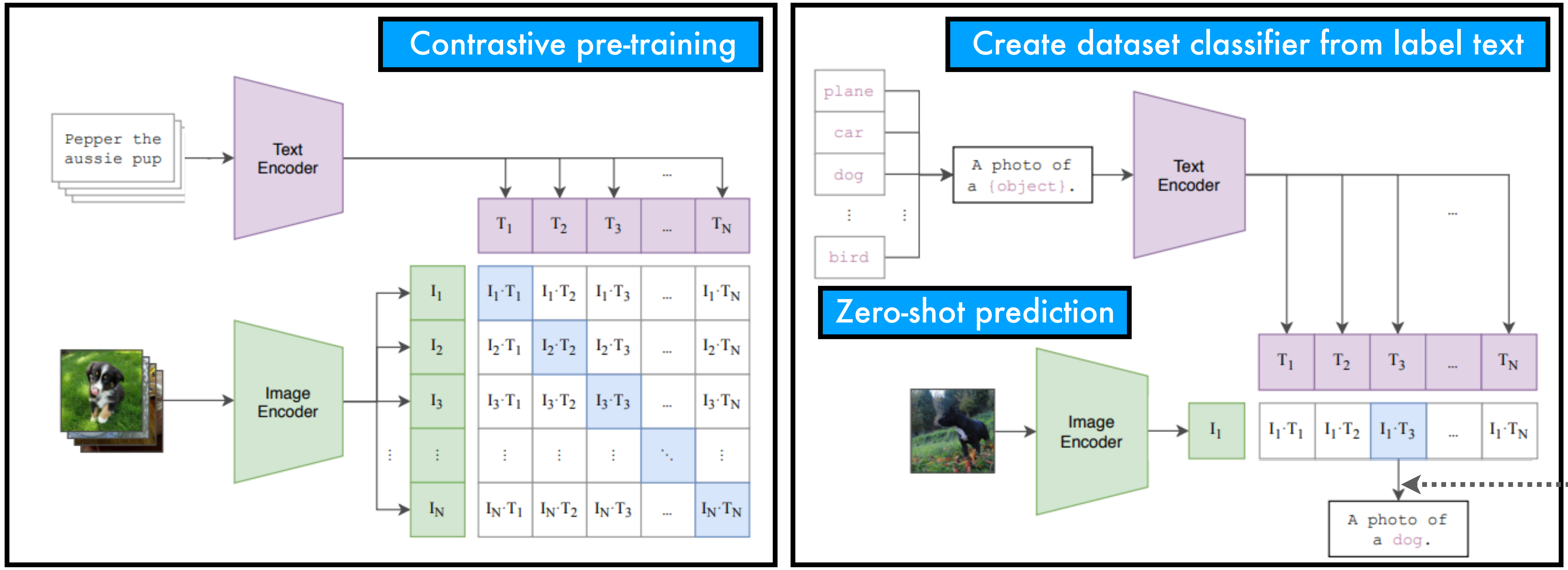
(TinyImages) A. Torralba et al., "80 million tiny images: A large data set for nonparametric object and scene recognition", TPAMI *(2008)*

A. Li, A. Jabri, A. Joulin, and L. Van Der Maaten, "Learning visual n-grams from web data", ICCV (2017)

R. Taori et al., "Measuring robustness to natural distribution shifts in image classification", NeurIPS (2020)

Using CLIP for Zero-shot Transfer

Zero-shot transfer with CLIP



Classification

1. Compute cosine similarities
2. Scale similarities by temperature τ
3. Normalise into probabilities via softmax

(effectively logistic regression)

Notes:

We can interpret the text encoder as a hypernetwork (Ha et al., 2016) that generates the weights of a linear classifier.

The text features for each class are cached, so the cost is amortised over all predictions for a dataset.

References/Image credits

A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)

D. Ha, A. Dai, Q. V. Le, "Hypernetworks", ICLR (2017)

Initial zero-shot transfer experiments/prompting

Comparison to Visual N-grams

Compare zero-shot transfer against Visual N-grams (Li et al., 2017) on three datasets.

Not controlled experiments (in compute, model capacity or data), but useful context for the magnitude of gains.

	aYahoo	ImageNet	SUN
Visual N-Grams	72.4	11.5	23.0
CLIP	**98.4**	**76.2**	**58.5**

Prompt Engineering

In zero-shot transfer, using text class labels can present challenges:

Some datasets only provide integer class id labels (these cannot be used).

One issue is polysemy - the word sense is ambiguous without context.

E.g. in ImageNet there are two "crane" classes (bird and construction)!

Prompt Templates: since images are rarely paired with single words during training, templates like "A photo of a {label}." are useful.

On ImageNet, just using this prompt over raw labels brings a gain of 1.3%.

Customised templates are also useful for fine-grained classification:

- (Oxford-IIIT Pets) "A photo of a {label}, a type of pet."
- (Satellite imagery) "A satellite photo of a {label}"

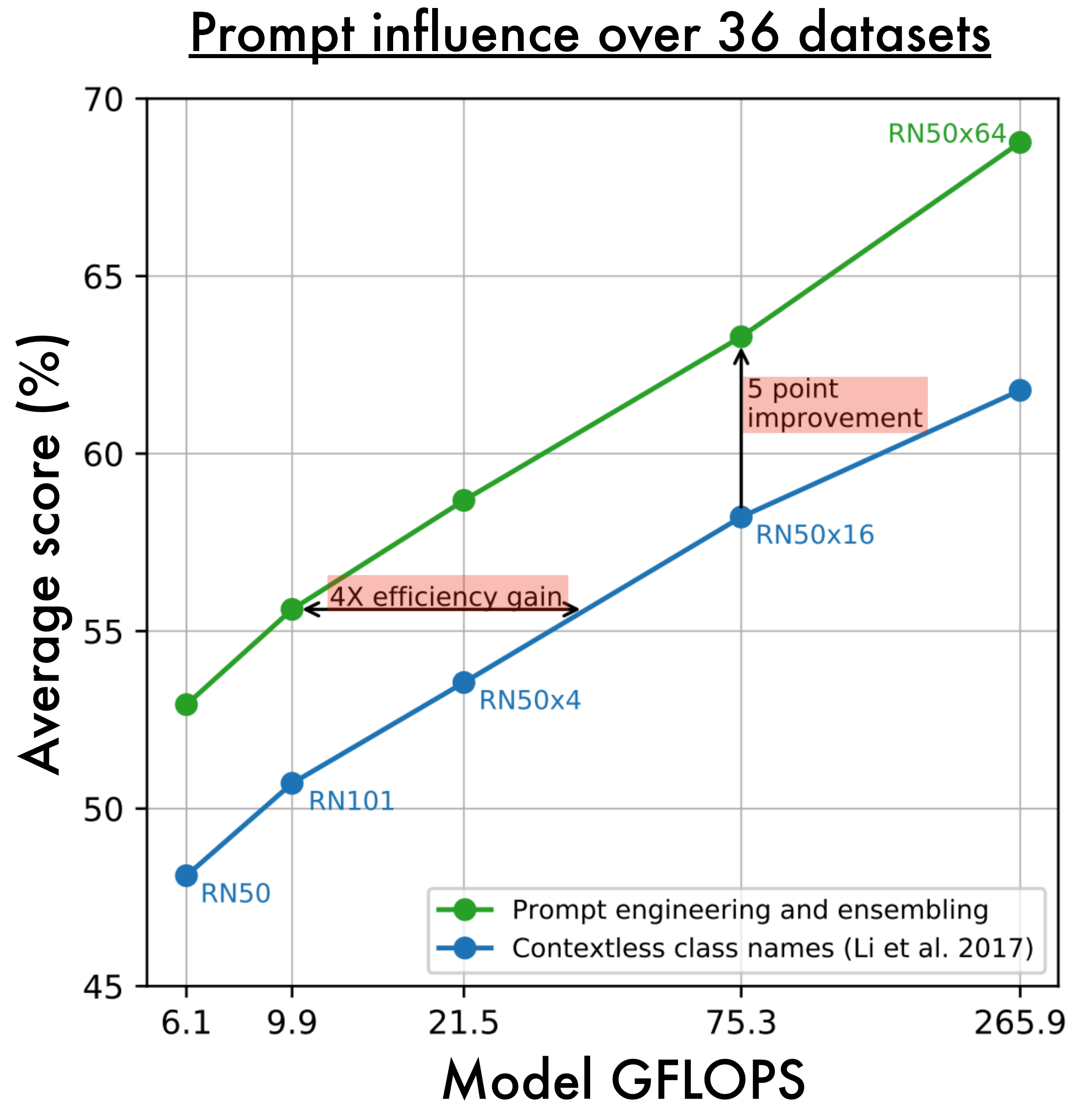
Prompt Ensembling

Ensembling over zero-shot classifiers can further boost performance.

- "A photo of a big {label}."
- "A photo of a small {label}."

Note: Ensembling is performed over the embeddings, rather than predicted probabilities to enable caching so that the cost is amortised over predictions.

On ImageNet, ensembling over 80 different prompts yields a 3.5% gain.



References/Image credits

A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)

A. Li et al., "Learning visual n-grams from web data", ICCV (2017)

(aYahoo dataset) A. Farhadi et al., "Describing objects by their attributes", CVPR (2009)

(ImageNet dataset) J. Deng et al., "Imagenet: A large-scale hierarchical image database", CVPR (2009)

(SUN dataset) J. Xiao et al., "Sun database: Large-scale scene recognition from abbey to zoo", CVPR (2010)

(Oxford-IIIT Pets) O. Parkhi et al., "Cats and dogs", CVPR (2012)

Zero-shot analysis

Datasets

A suite of 27 datasets are used for analysis:

Food-101 | CIFAR-10 | CIFAR-100 | Birdsnap | SUN397
Stanford Cars | FGVC Aircraft | Pascal VOC 2007 Classification
Describable Textures | Oxford-IIIT Pets | Caltech-101
Oxford Flowers 102

12 datasets of Kornblith et al. (2019)

MNIST | FER 2013 | STL-10 | EuroSAT | RESISC45 | GTSRB
KITTI | PatchCamelyon | UCF101 | Kinetics700 | CLEVR Counts
Country211 | Rendered SST2 | Hateful Memes | ImageNet

Video datasets: use frames as input images

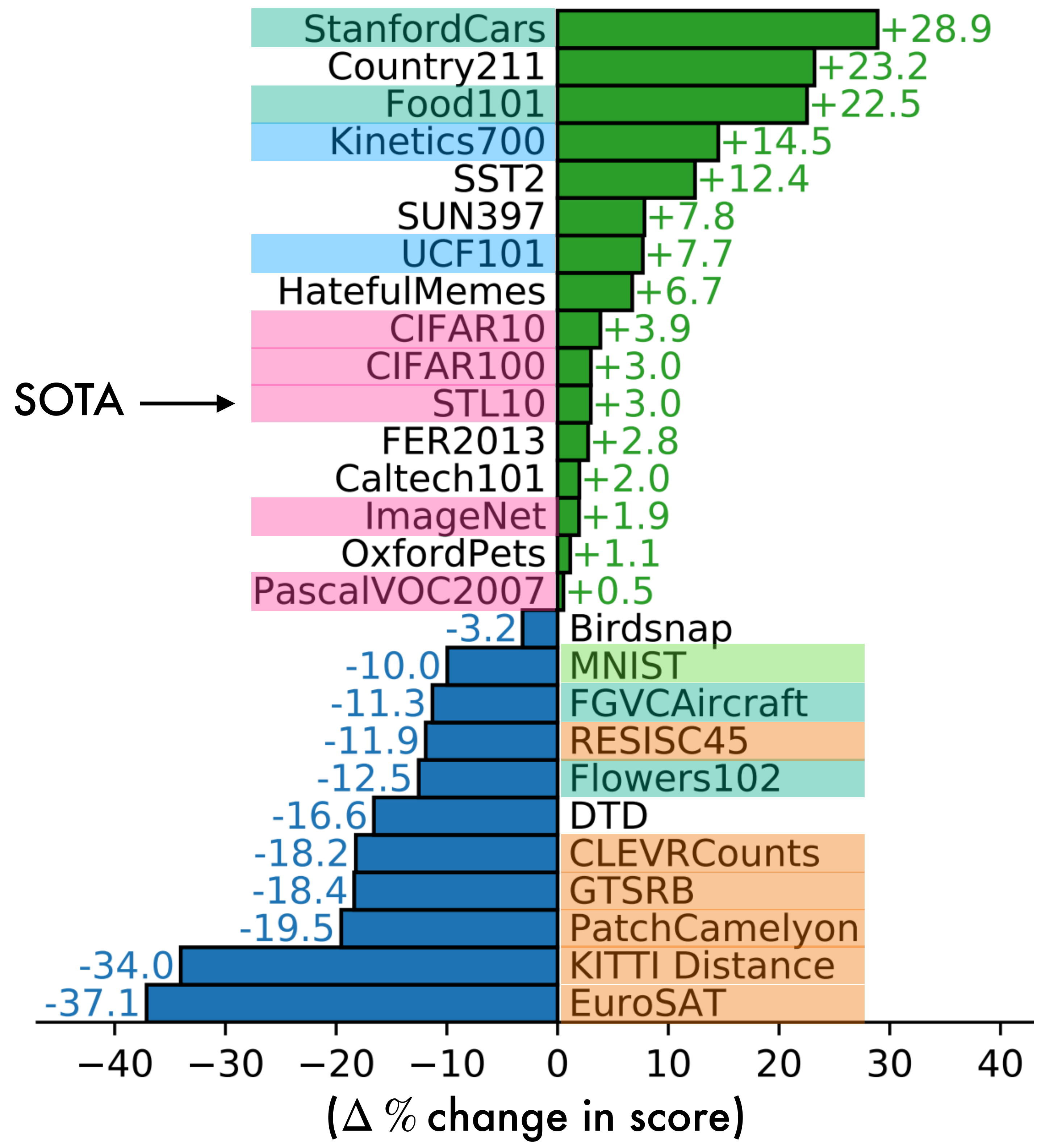
Country211: geolocalisation (photos across 211 countries)

Rendered SST2: OCR evaluation

Additional 15 datasets

Zero-shot evaluation

Baseline: (fully-supervised) linear probe on (ImageNet) ResNet-50 features.



Zero-shot CLIP vs linear probe on ResNet 50

Fine-grained
Different per-task supervision in WIT and ImageNet?

General objects
Similar performance, slightly in favour of CLIP.

Action recognition
Language supervision may help with verbs.

Complex/abstract
CLIP struggles with complexity
Few-shot eval may be better

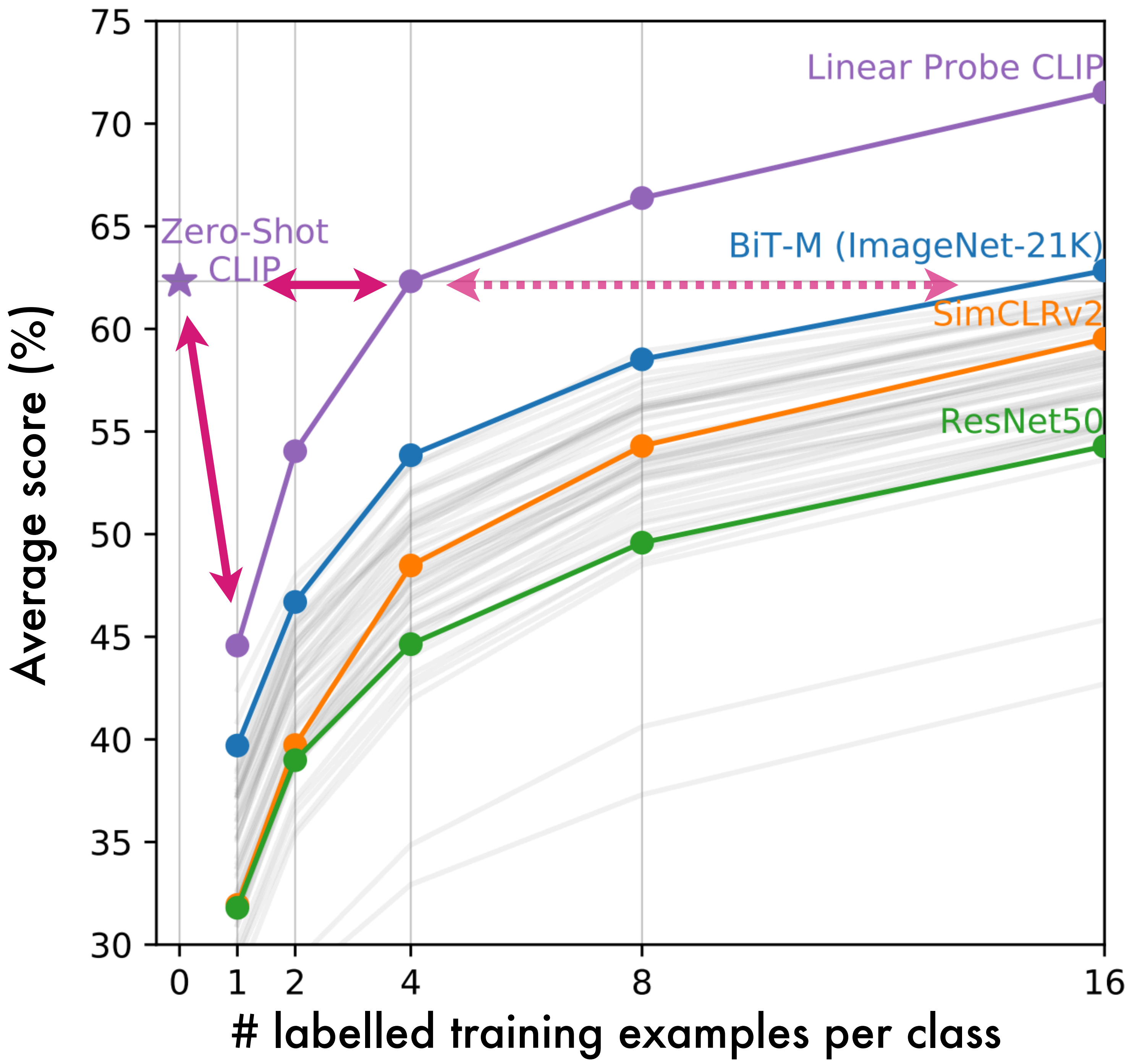
References/Image credits
S. Kornblith, J. Shlens and Q. V. Le, "Do better imagenet models transfer better?", CVPR (2019)
(VTAB) X. Zhai et al., "The visual task adaptation benchmark", openreview.net, (2019)
(ResNet-50) K. He et al., "Deep residual learning for image recognition", CVPR (2016)
A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)

Zero-shot vs few-shot

Comparison to few-shot linear probes

Comparison: zero-shot CLIP vs few-shot linear probes on various features

Data: the 20 datasets with at least 16 examples per class.

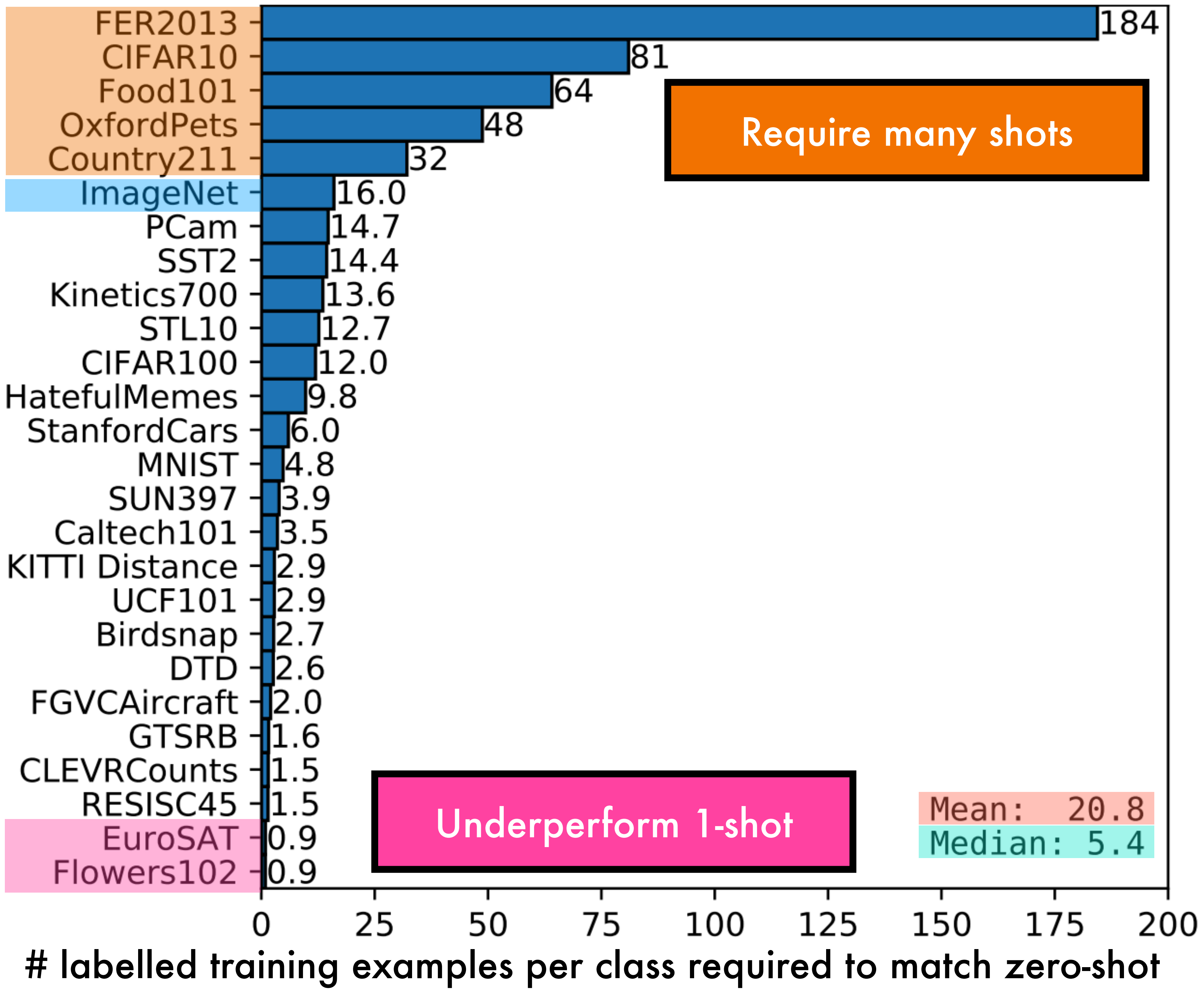


Combining zero-shot & one-shot was non-trivial (but see Zhang et al. (2021))

Individual dataset analysis

Aim: Estimate data efficiency of zero-shot CLIP across datasets.

For efficiency, estimate few-shot score for each #shots via interpolation.



References/Image credits

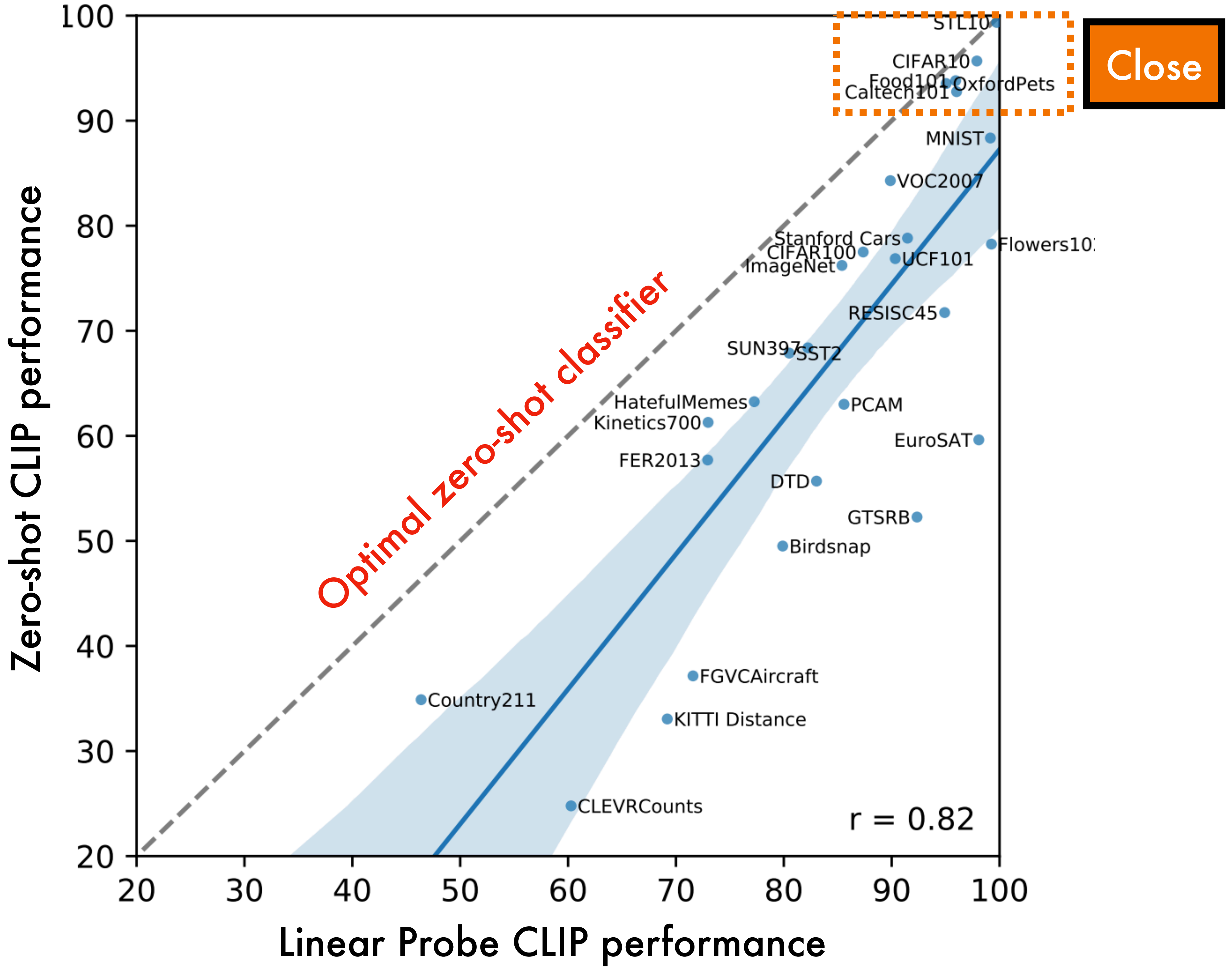
A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)

R. Zhang et al., "Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling", *arXiv* (2021)

Zero-shot optimality and model scaling

Zero-shot vs supervised linear classifier

Since zero-shot classifier is a linear classifier, we can use fully-supervised linear probes as an approximate upper bound for zero-shot transfer.

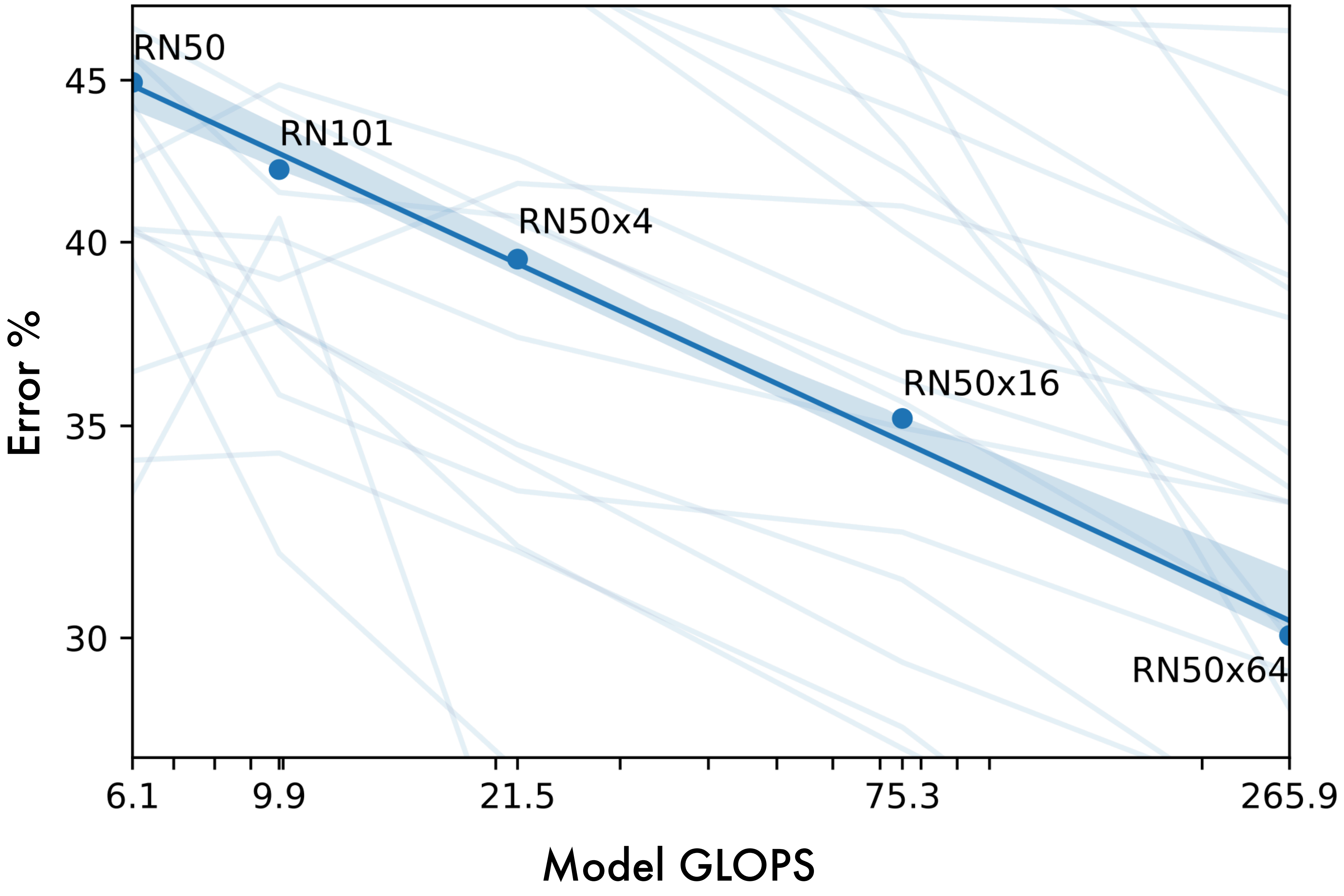


Zero-shot performance is correlated with fully supervised performance.

Model scaling

Empirical studies have shown deep learning performance can scale smoothly with model capacity, dataset size etc. (Kaplan et al. 2020)

Study: compare CLIP across 36 datasets over 44x compute scaling



Similar to prior studies, a log-linear trend is observed across compute.

References/Image credits

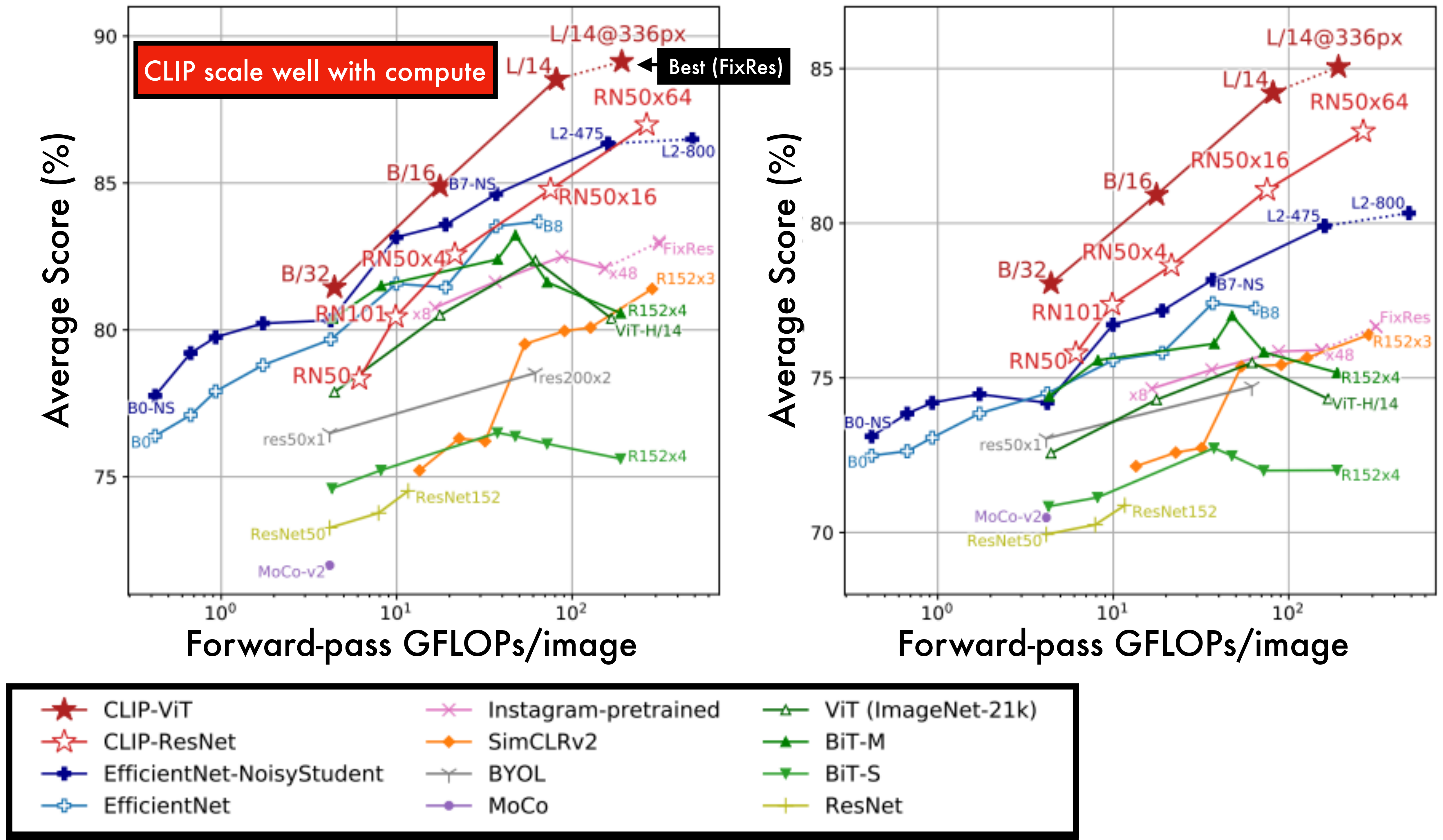
S.Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)

J. Kaplan et al., "Scaling laws for neural language models", arXiv (2020)

Representation Learning

Linear probes

Probe avg. over 12 datasets (Kornblith et al. 2019)   Probe avg. over 27 datasets

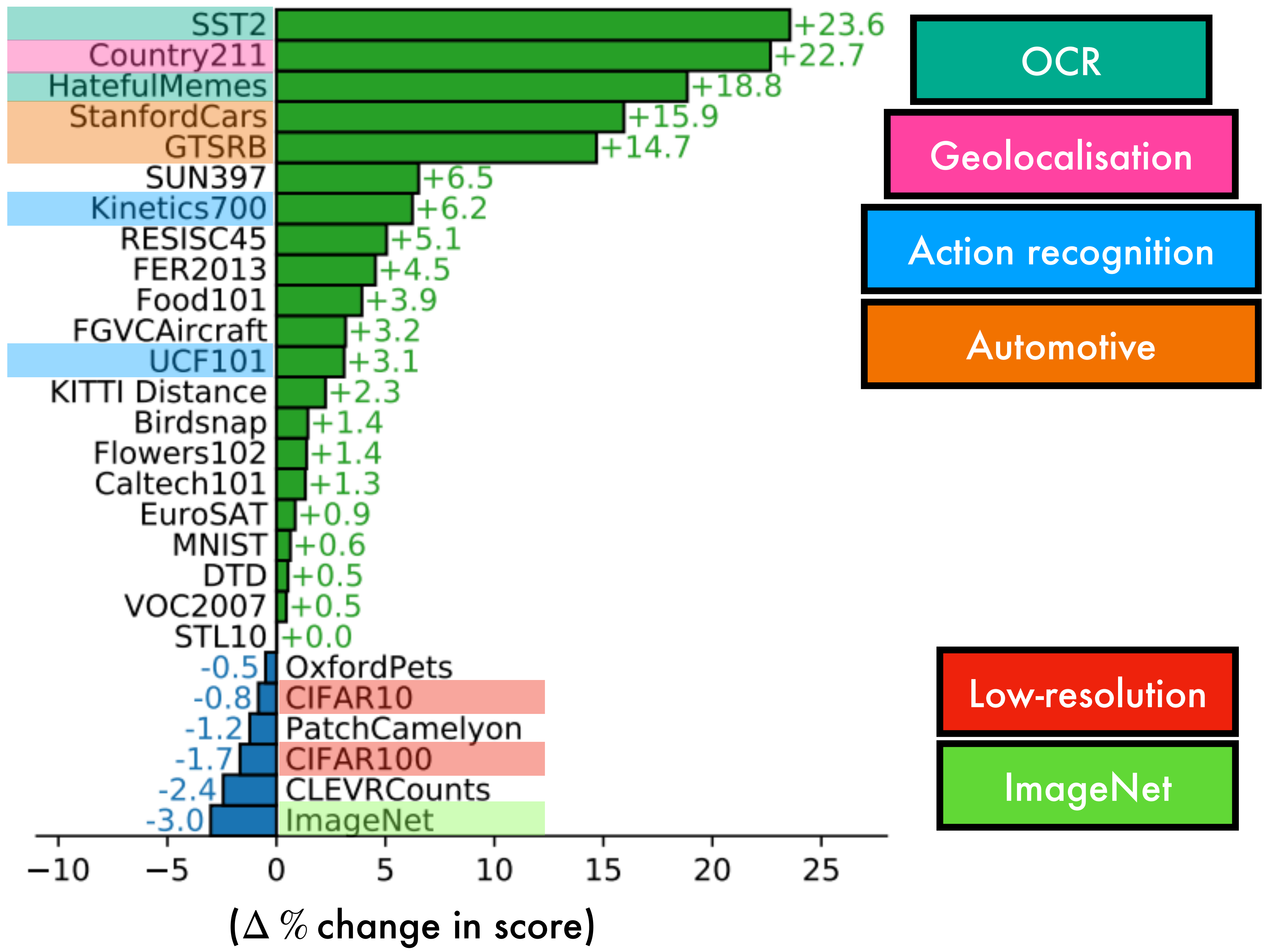


Dotted lines indicate models fine-tuned/evaluated at higher resolution than during pre-training.

CLIP ViTs 3x more compute efficient than CLIP ResNets - similar finding to Dosovitskiy et al. (2021)

Comparison to best model - breakdown

Compare CLIP to best model - Noisy Student EfficientNet-L2 (Xie et al., 2020)



Logistic Regression on CLIP vs EfficientNet L2 NS

References/Image credits

A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)

S. Kornblith, J. Shlens and Q. V. Le, "Do better imagenet models transfer better?", CVPR (2019)

A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale", ICLR (2021)

H. Touvron et al., "Fixing the train-test resolution discrepancy: FixEfficientNet", arxiv (2020)

Q. Xie et al., "Self-training with noisy student improves imagenet classification", CVPR (2020)

Robustness to natural distribution shifts

Motivation

Since 2015, deep learning models have exceeded human performance (as courageously estimated by A. Karpathy) estimate (He et al., 2015)

But later studies have found these systems still make simple mistakes (Dodge et al., 2017) and fall below human performance on other benchmarks (Recht et al. 2019)

Common explanation: deep learning finds both useful and spurious correlations

However, most studies have examined models trained on ImageNet.

To what extent are failures attributable to ImageNet training, deep learning or both?

CLIP models (trained with natural language supervision on very large training dataset - not ImageNet, good zero-shot performance) enable a fresh analysis of this question.

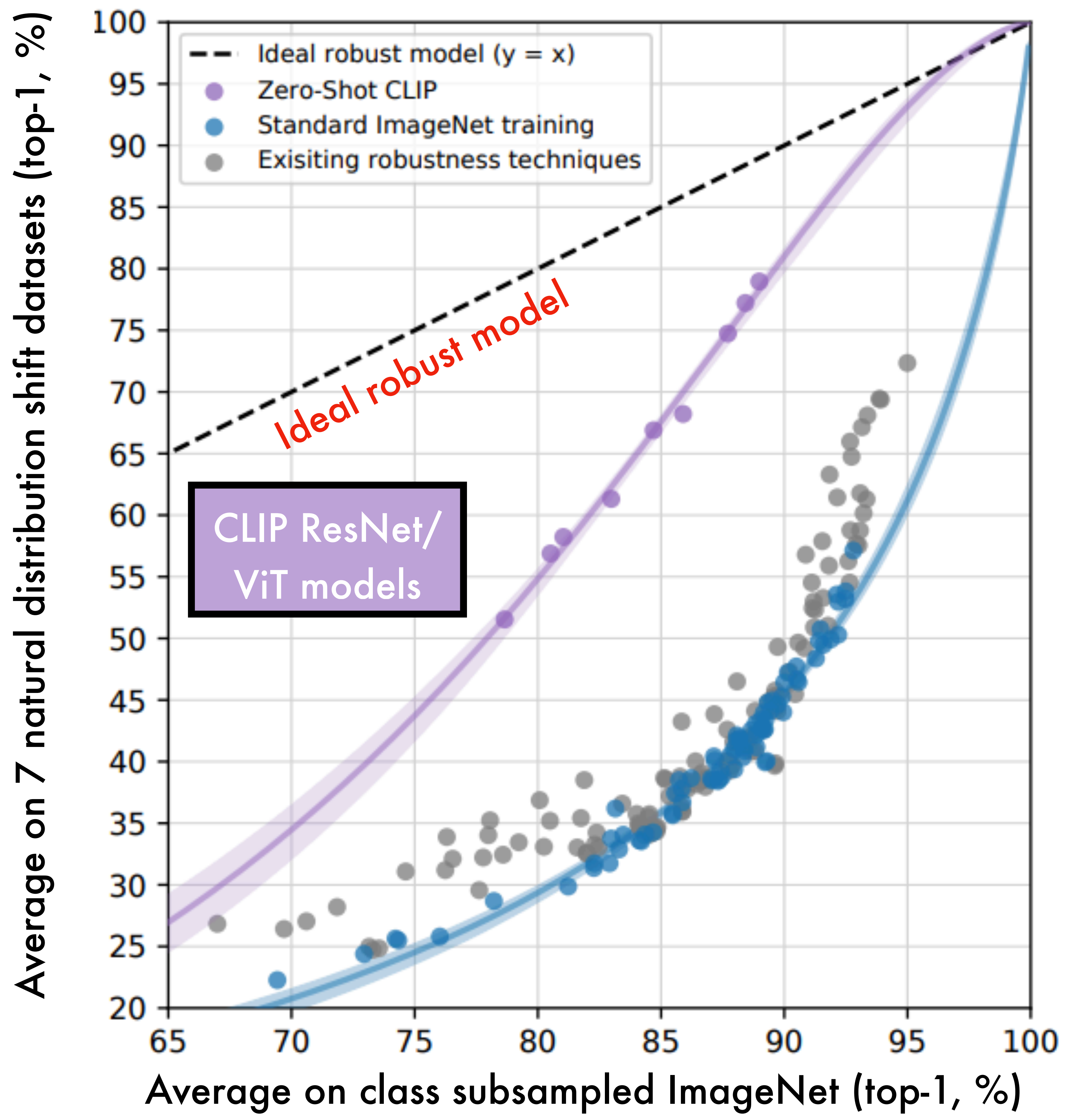
Datasets

Evaluate robustness to seven "natural distribution shifts" investigated by Taori et al. (2020).

ImageNetV2 | ImageNet-Vid | ImageNet Sketch | Youtube-BB

ObjectNet | ImageNet Adversarial | ImageNet Renditions

Robustness to seven natural distribution shifts



References/Image credits

K. He et al., "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification", CVPR (2015)

(Karpathy human estimate) https://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/

S. Dodge et al., "A study and comparison of human and deep learning recognition performance under visual distortions", ICCCN (2017)

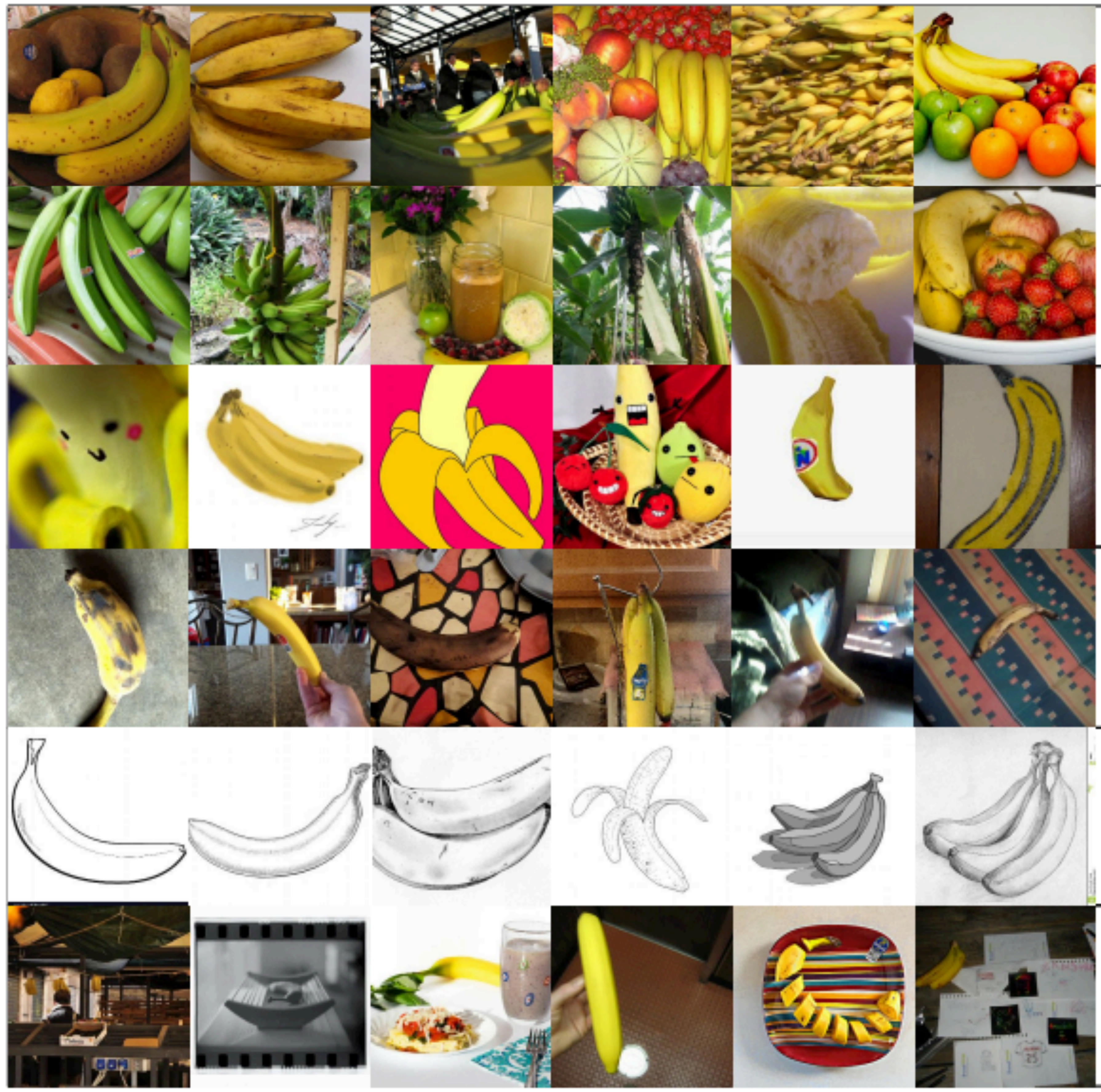
B. Recht et al., "Do imagenet classifiers generalize to imagenet?", ICML (2019)

A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)

Robustness to natural distribution shifts (qualitative)

Banana Visualisation

Dataset Examples



	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	**76.2**	**76.2**	0%
ImageNetV2	64.3	**70.1**	+5.8%
ImageNet-R	37.7	**88.9**	+51.2%
ObjectNet	32.6	**72.3**	+39.7%
ImageNet Sketch	25.2	**60.2**	+35.0%
ImageNet-A	2.7	**77.1**	+74.4%

References/Image credits

A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)

(ImageNet dataset) J. Deng et al., "Imagenet: A large-scale hierarchical image database", CVPR (2009)

(ImageNetV2) B. Recht et al., "Do imagenet classifiers generalize to imagenet?", ICML (2019)

(ImageNet-R) D. Hendrycks et al., "The many faces of robustness: A critical analysis of out-of-distribution generalization", ICCV (2021)

(ObjectNet) A. Barbu et al., "Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models", *NeurIPS* (2019)

(ImageNet Sketch) R. Geirhos et al., "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness", arXiv (2018)

(ImageNet-A) D. Hendrycks et al., "Natural adversarial examples", CVPR (2021)

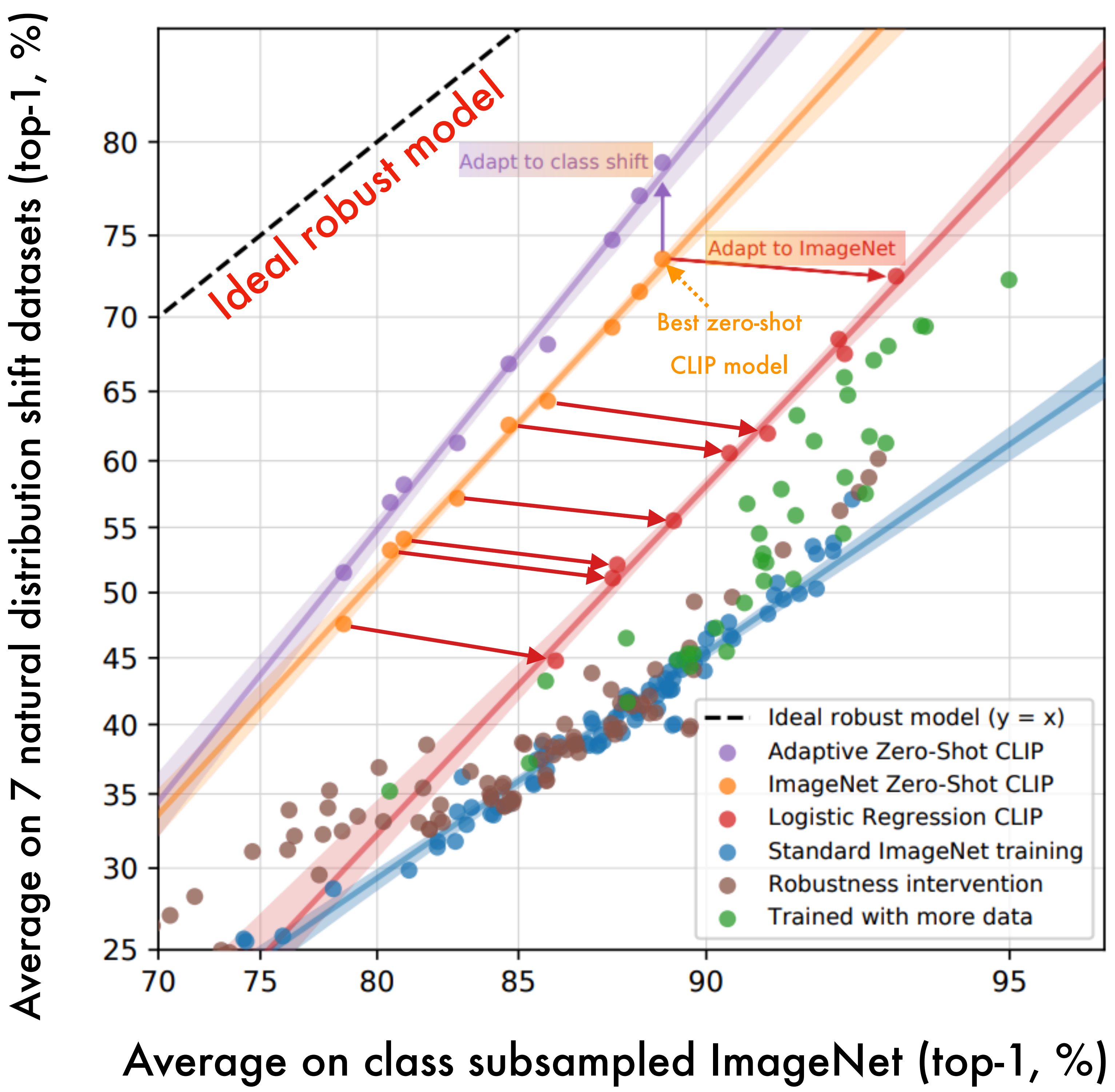
How does ImageNet adaptation affect robustness?

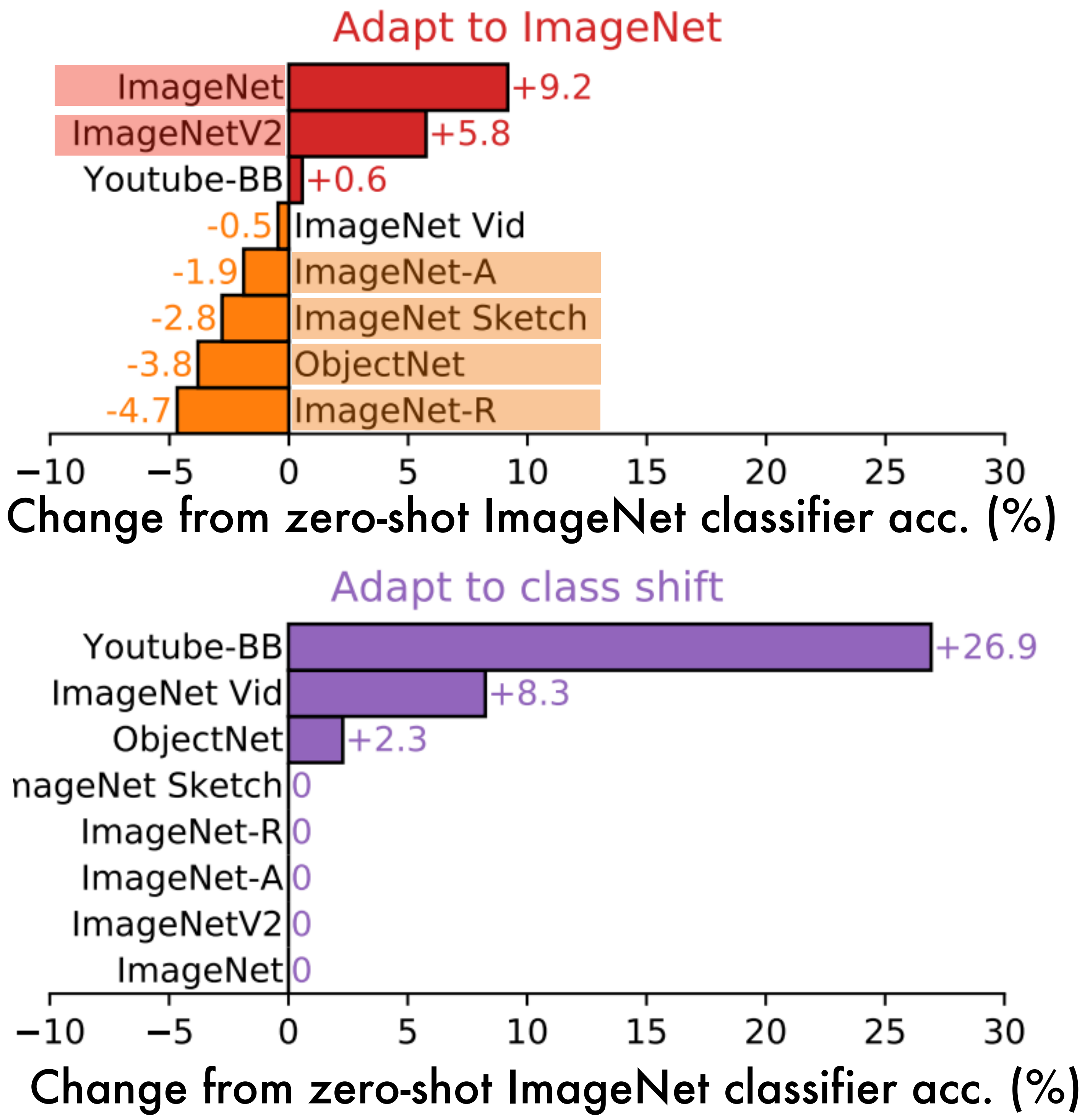
ImageNet adaptation

The strong zero-shot robustness of CLIP need not imply that ImageNet training causes the robustness gap.

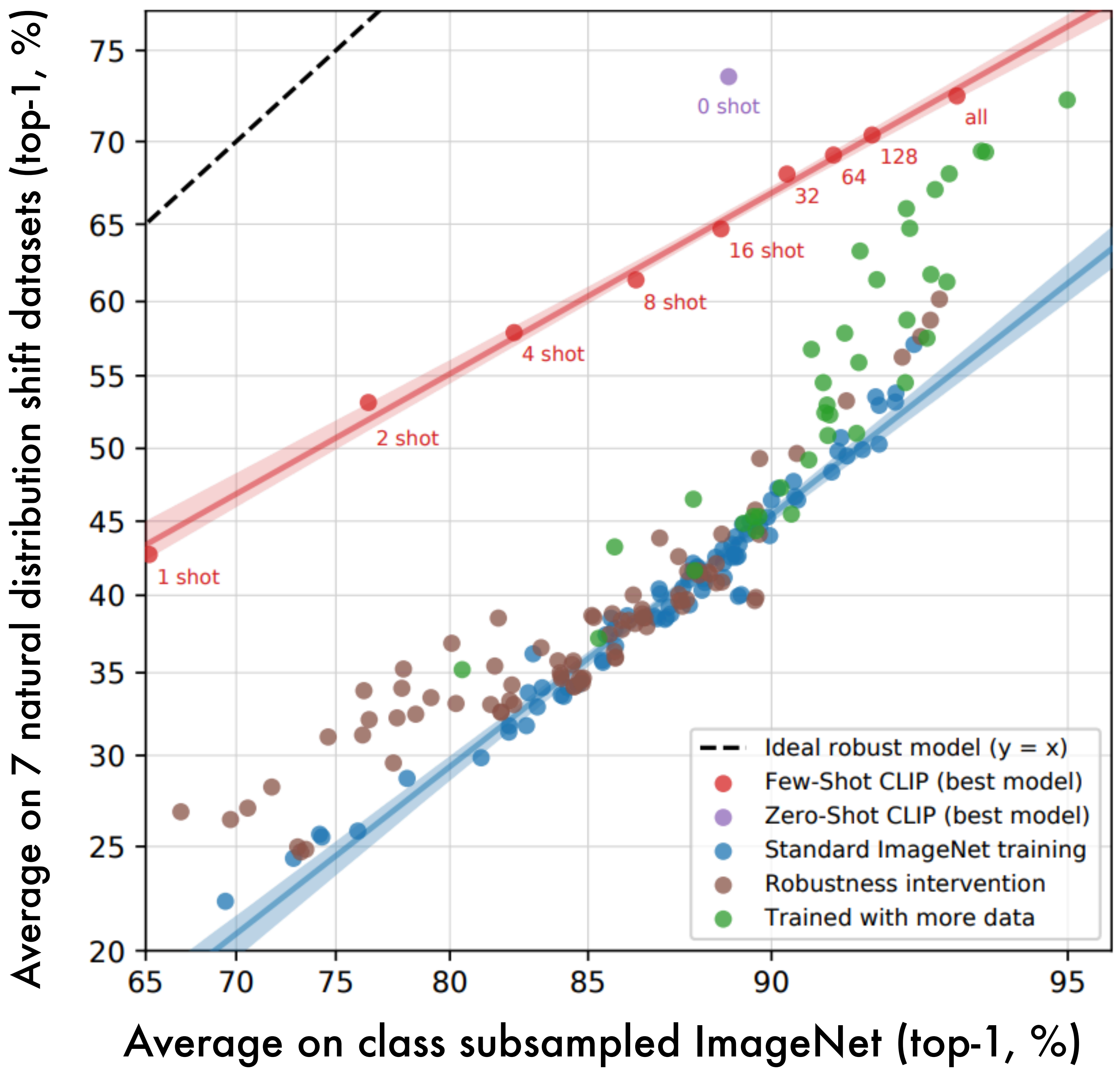
Other elements of CLIP (pretraining dataset size or natural language supervision) could explain its robustness.

Experiment: first fit CLIP features to ImageNet via logistic regression, then re-evaluate robustness.





Few-shot robustness



Takeaway: large-scale task and dataset agnostic pre-training with zero/few-shot evaluation on diverse benchmarks (Yogatama et al., 2019) promotes robustness.

References/Image credits

A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)

D. Yogatama et al., "Learning and Evaluating General Linguistic Intelligence", (2019)

Comparison to Human Performance

Human study

To assess how CLIP compares to humans, 5 humans predicted labels the Oxford IIT Pets dataset (Parkhi et al., 2012), a 37-way dog/cat breed classification task. Humans were evaluated in zero-shot, one-shot and two-shot settings.

	Accuracy	Majority Vote on Full Dataset	Accuracy on Guesses	Majority Vote Accuracy on Guesses
Zero-shot human	53.7	57.0	69.7	63.9
Zero-shot CLIP	**93.5**	**93.5**	**93.5**	**93.5**
One-shot human	75.7	80.3	78.5	81.2
Two-shot human	75.7	85.0	79.2	86.1

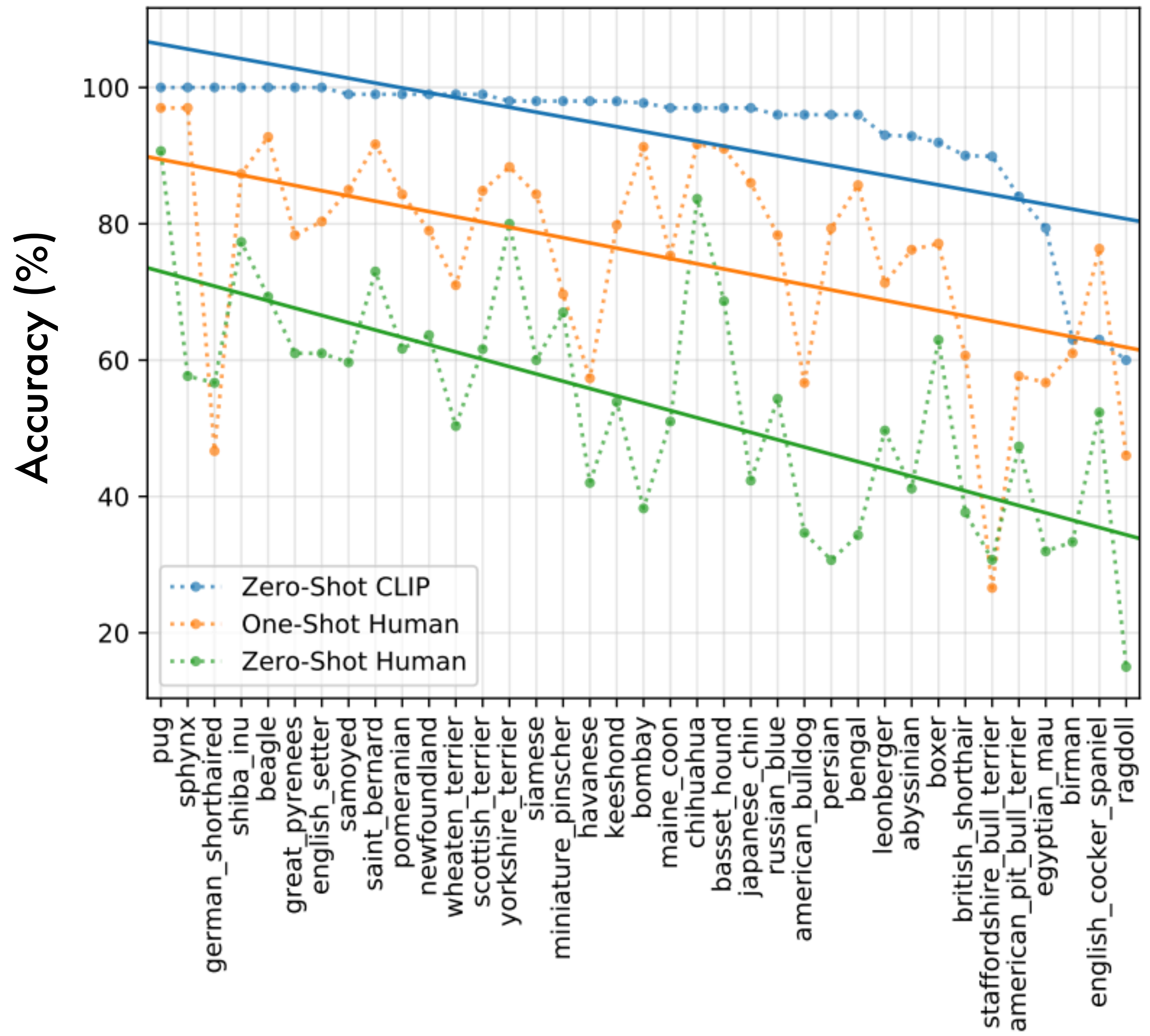
Major gain from zero-shot to one-shot. **No gain** from one-shot to two-shot.

The gain from zero-shot to one-shot is almost entirely on images that humans were uncertain about (i.e. they have a sense of what they don't know).

There are likely opportunities for improvements for machine sample efficiency.

Integrating prior knowledge (like humans) seems a promising direction.

Error analysis



Problems that are hard for CLIP are also hard for humans.

Likely due to label noise and difficulty of out-of-distribution images.

References/Image credits

O. Parkhi et al., "Cats and dogs", CVPR (2012)

A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)

Downstream applications

Text and Image Retrieval

Retrieval Tasks:

Image retrieval - rank images according to how well they fit a query

Text retrieval - rank captions according to how well they describe an image

Datasets: Flickr30K | MSCOCO

Results: Strong zero-shot retrieval results on both datasets vs prior work.

A little behind SOTA among methods fine-tuned on MSCOCO.

Optical Character Recognition (OCR)

Assess performance on tasks requiring direct/indirect use of OCR:

Low-level character/word recognition: MNIST | SVHN | IIIT5K

Semantic tasks: Hateful Memes | SST -2

Results: Strongly dependent on domain (rendered vs natural images)

Strongly dependent on type of text (numbers vs words)

Good: Hateful Memes | SST -2 — OK: IIIT5K — Poor: MNIST | SVHN

Action Recognition

Assess CLIP (both linear probes and zero-shot) for action recognition.

For linear probe, the middle frame of each video is used (to reduce cost)

For zero-shot all frames are used (scores are averaged)

Datasets: UCF-101 | Kinetics-700 | RareAct

Results: Encouraging linear probe/zero-shot on UCF-101 & Kinetics-700

SOTA on zero-shot recognition on RareAct.

Geolocalisation

It was observed during development that CLIP could recognise many locations.

This ability was quantified on two tasks.

Datasets: Country211(new) | IM2GPS

To perform location regression for IMG2GPS, GPS coordinates are estimated via nearest neighbours in a set of 1M reference images with CLIP embeddings.

Results: solid results on IM2GPS (though not SOTA)

References:

(Flickr30K) P. Young et al., "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions", ACL (2014)
(MSCOCO) X. Chen et al. "Microsoft coco captions: Data collection and evaluation server", arXiv (2015)
(MNIST) Y. LeCun et al., "Gradient-based learning applied to document recognition", Proceedings of the IEEE (1998)
(SVHN) Y. Netzer et al., "Reading Digits in Natural Images with Unsupervised Feature Learning", (2011)
(IIIT5K) A. Mishra et al., "Scene text recognition using higher order language priors", BMVC (2012)
(Hateful Memes) D. Kiela et al., "The hateful memes challenge: Detecting hate speech in multimodal memes", NeurIPS (2020)
(SST-2) R. Socher et al., "Recursive deep models for semantic compositionality over a sentiment treebank", EMNLP (2013)
(UCF-101) K. Soomro et al., "UCF101: A dataset of 101 human actions classes from videos in the wild", arXiv (2012)
(Kinetics-700) J. Carreira et al., "A short note on the kinetics-700 human action dataset", arXiv (2019)
(RareAct) A. Miech et al., "RareAct: A video dataset of unusual interactions", arXiv (2020)
(IM2GPS) J. Hays and A. Efros, "IM2GPS: estimating geographic information from a single image", CVPR (2008)

Outline

- Motivation
- CLIP: Data and Method
- Experiments
- Data Overlap Analysis
- Limitations
- Broader Impacts
- Related Work
- Summary

Data Overlap Analysis: Approach

Overview

A key issue with large internet dataset pre-training is unintentional overlap with downstream evaluation datasets (invalidating results).

One solution: remove all duplicates before training a model

Pros: guarantees true downstream hold-out performance

Cons: requires knowing all possible test data ahead of time (limits analysis)

Alternative approach (taken in this paper) is to document:

- how much overlap occurs?
- how much performance changes due to these overlaps?

Near-duplicate Detector

CLIP embeddings do not work well for duplicate detection (too semantic)

Train a ResNet-50 with InfoNCE loss to discriminate augmented versions of images from other images.

Training set: 30 million image subset of 400 million dataset.

At the end of training, it achieves nearly 100% accuracy on proxy training task.

Dataset overlap analysis pipeline

For *each evaluation dataset:*

1. Estimate contamination:

- Run near-duplicate detector
- Use manual inspection to set per-dataset threshold (for high precision & recall)
- Split dataset into `Clean (below thr)` `Overlap (above thr)` `All`
- Report data contamination as the ratio |`Overlap`| / |`All`|

2. Estimate performance change due to contamination:

- Compute zero-shot accuracy of CLIP RN50x64 on `Overlap`, `Clean`, `All`.
- Report acc(`All`) - acc(`Clean`) as metric for performance change

3. Assess significance

- Since overlap is typically small, run binomial significance test (using accuracy on `Clean` as null hypothesis, compute one-tailed p-value for `Overlap` subset)
- Also compute 99.5% Clopper-Pearson confidence intervals on `Overlap`.

Data Overlap Analysis: Results

Visualisation of overlap and contamination influence

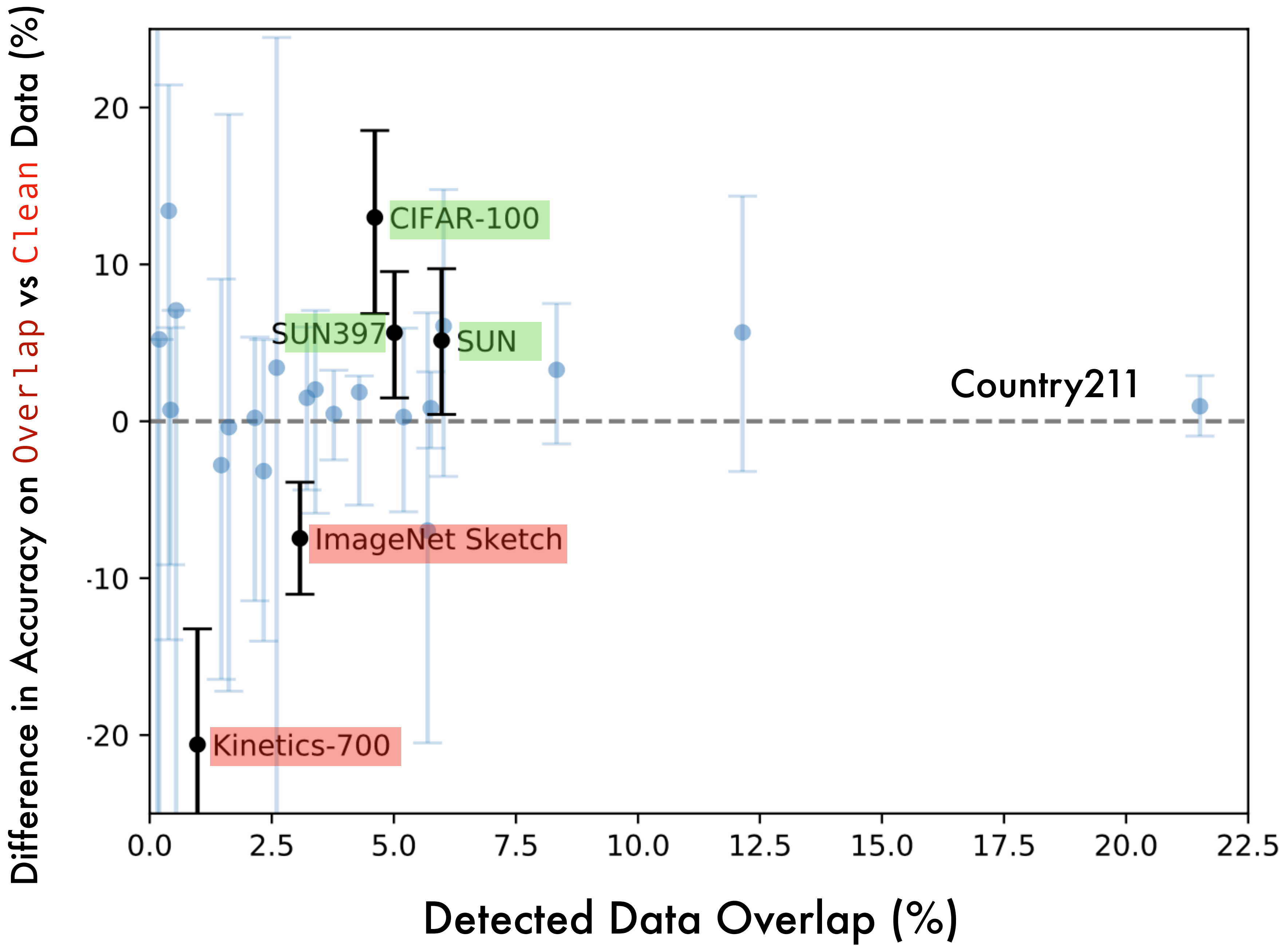
Overlap statistics across the 35 evaluation datasets considered in this work

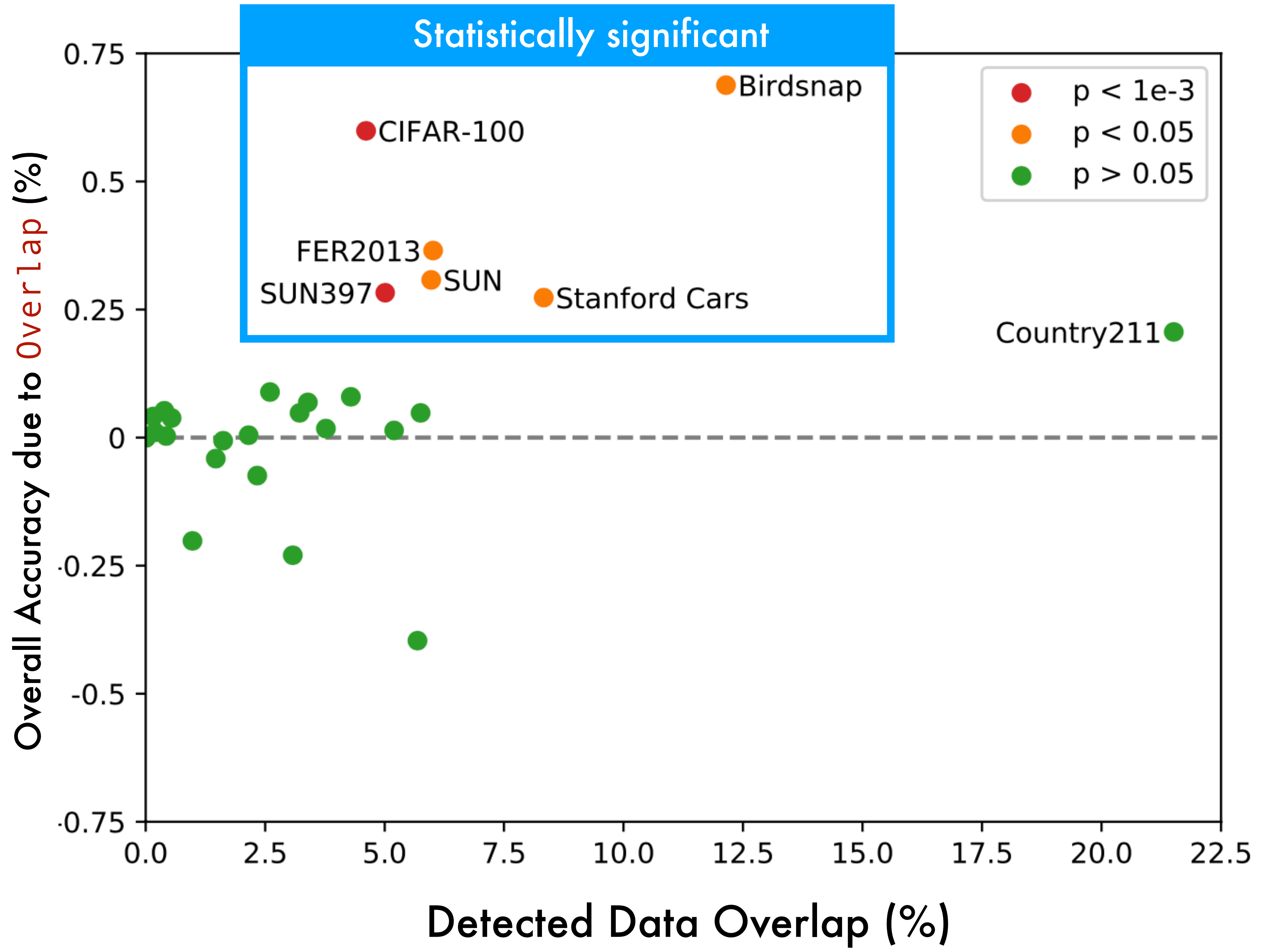
Median overlap: 2.2% with pre-training

Mean overlap: 3.2% with pre-training

Among these datasets, 9 have no detected overlap with the pre-training dataset:

- Some are specialised/synthetic (e.g. MNIST, CLEVR, GTSRB), making them unlikely to posted online as normal images.
- Others contain data created after the pre-training dataset was curated (ObjectNet and Hateful Memes)





Limitations: (1) imperfect duplicate detection (hard to validate); (2) distribution shift (e.g. all "overlaps" in Kinetics are black transition frames)

Summary: data contamination does not appear to have a major effect on results

References/Image credits

(MNIST) Y. LeCun et al., "Gradient-based learning applied to document recognition", Proceedings of the IEEE (1998)

(CLEVR) J. Johnson et al., "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning", *ICCV* (2017)

(GTSRB) J. Stallkamp et al., "The German traffic sign recognition benchmark: a multi-class classification competition", IJCNN (2011)

A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)

(ObjectNet) A. Barbu et al., "Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models", NeurIPS (2019)

(Hateful Memes) D. Kiela et al., "The hateful memes challenge: Detecting hate speech in multimodal memes", NeurIPS (2020)

Outline

- Motivation
- CLIP: Data and Method
- Experiments
- Data Overlap Analysis
- Limitations
- Broader Impacts
- Related Work
- Summary

Limitations

Zero-shot performance

Zero-shot CLIP is competitive against a supervised linear probe on ResNet-50 features, but well behind SOTA on most datasets.

Estimate: 1000x more compute is required for zero-shot CLIP to reach SOTA

Research is required to improve the computational/data efficiency of CLIP.

CLIP struggles on abstract tasks like counting objects in an image, certain fine-grained classification tasks, and tasks likely outside the pre-training data.

On truly out-of-distribution data, such as MNIST, CLIP achieves only 88%, underperforming logistic regression on raw pixels.

Given its good performance on other OCR evaluations, this suggest CLIP does not address the brittle generalisation of deep learning models.

Instead, it hopes all test data will be effectively in-distribution from pre-training.

As MNIST demonstrates, this assumption is easily violated in practice.

Flexibility

CLIP is limited to choosing among concepts in a given zero-shot classifier.

Less flexible than image captioning.

Future work could combine the efficiency of CLIP with flexible captioning.

Data efficiency

CLIP inherits the poor data efficiency of deep learning

It aims to compensate by using a scalable pre-training data source.

Fun fact: if each image seen by CLIP was shown at 1 fps, it would take 405 years to iterate through the 32 epochs of training (12.8 billion images).

Methodology

Repeated querying of validation sets to guide CLIP development.

While 12 datasets used follow Kornblith et al., (2019), the broader suite of 27 datasets is co-adapted with development and capabilities of CLIP.

A benchmark of tasks for broad zero-shot transfer could help address this.

Uncurated data

By training on unfiltered internet image/text CLIP learns many social biases.

Room for few-shot improvement

Few-shot performance often falls below zero-shot: more research is required.

References

(ResNet-50) K. He et al., "Deep residual learning for image recognition", CVPR (2016)

S. Kornblith, J. Shlens, & Q. V. Le, Q "Do better imagenet models transfer better?", CVPR (2019)

Outline

- Motivation
- CLIP: Data and Method
- Experiments
- Data Overlap Analysis
- Limitations
- Broader Impacts
- Related Work
- Summary

Broader Impacts

Overview

Thanks to zero-shot performance, CLIP has a broad range of applications.

Since it allows creating classes for categorisation ("roll your own classifier") it is challenging to characterise - capabilities become clear only after testing for them.

Applications: CLIP shows significant promise for tasks like retrieval, and possibly also for novel applications enabled by its limited need for specialised task data.

Analysis: FairFace bias benchmark, bias probes, surveillance performance.

Limitation: bias tests are limited in scope. Analysis required in deployment context.

Note on class design: Algorithmic design, training data and class definitions/ taxonomies (or "class design") have implications for social biases.

Class design is particularly important for CLIP (anyone can define their own class).

FairFace - classification analysis

Fairface is a dataset of 106K images that are approximately balanced across 7 race categories, annotated with (est.) age, race and gender.

Linear probe CLIP tends to outperform existing baselines race, gender and age classification - zero-shot achieves more mixed results.

Gender classification

Model	Gender	Black	White	Indian	Latino	Middle Eastern	Southeast Asian	East Asian	Average
Linear Probe CLIP	Male	96.9	96.4	98.7	96.5	98.9	96.2	96.9	97.2
	Female	97.9	96.7	97.9	99.2	97.2	98.5	97.3	97.8
		97.4	96.5	98.3	97.8	98.4	97.3	97.1	97.5
Zero-Shot CLIP	Male	96.3	96.4	97.7	97.2	98.3	95.5	96.8	96.9
	Female	97.1	95.3	98.3	97.8	97.5	97.2	96.4	97.0
		96.7	95.9	98.0	97.5	98.0	96.3	96.6	
Linear Probe Instagram	Male	92.5	94.8	96.2	93.1	96.0	92.7	93.4	94.1
	Female	90.1	91.4	95.0	94.8	95.0	94.1	94.3	93.4
		91.3	93.2	95.6	94.0	95.6	93.4	93.9	

Note: probes offer only one approximation of algorithmic fairness.

References/Image credits:
K. Karkkainen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation", WACV (2021)
G. Bowker and S. L Star, "Sorting things out - Classification and its consequences", (1999)
A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)

Broader Impacts - analysis

FairFace - denigration harm terms

Zero-shot CLIP model was required to classify 10,000 images from FairFace dataset.

FairFace classes were augmented with {"animal", "gorilla", "chimpanzee" "orangutan"} (non-human), {"thief", "criminal", "suspicious person"} (crime-related).

Question: are these terms disproportionately assigned to demographic subgroups?

Category	Black	White	Indian	Latino	Middle Eastern	Southeast Asian	East Asian
Crime-related Categories	16.4	24.9	24.4	10.8	19.7	4.4	1.3
Non-human Categories	14.4	5.5	7.6	3.7	2.0	1.9	0.0

% of images classified into crime-related and non-human categories

Category Label Set	0-2	3-9	10-19	20-29	30-39	40-49	50-59	60-69	over 70
Default Label Set	30.3	35.0	29.5	16.3	13.9	18.5	19.1	16.2	10.4
Default Label Set + 'child' category	2.3	4.3	14.7	15.0	13.4	18.2	18.6	15.5	9.4

% of images classified into crime-related or non-human categories

Takeaway: class design can play an important role.

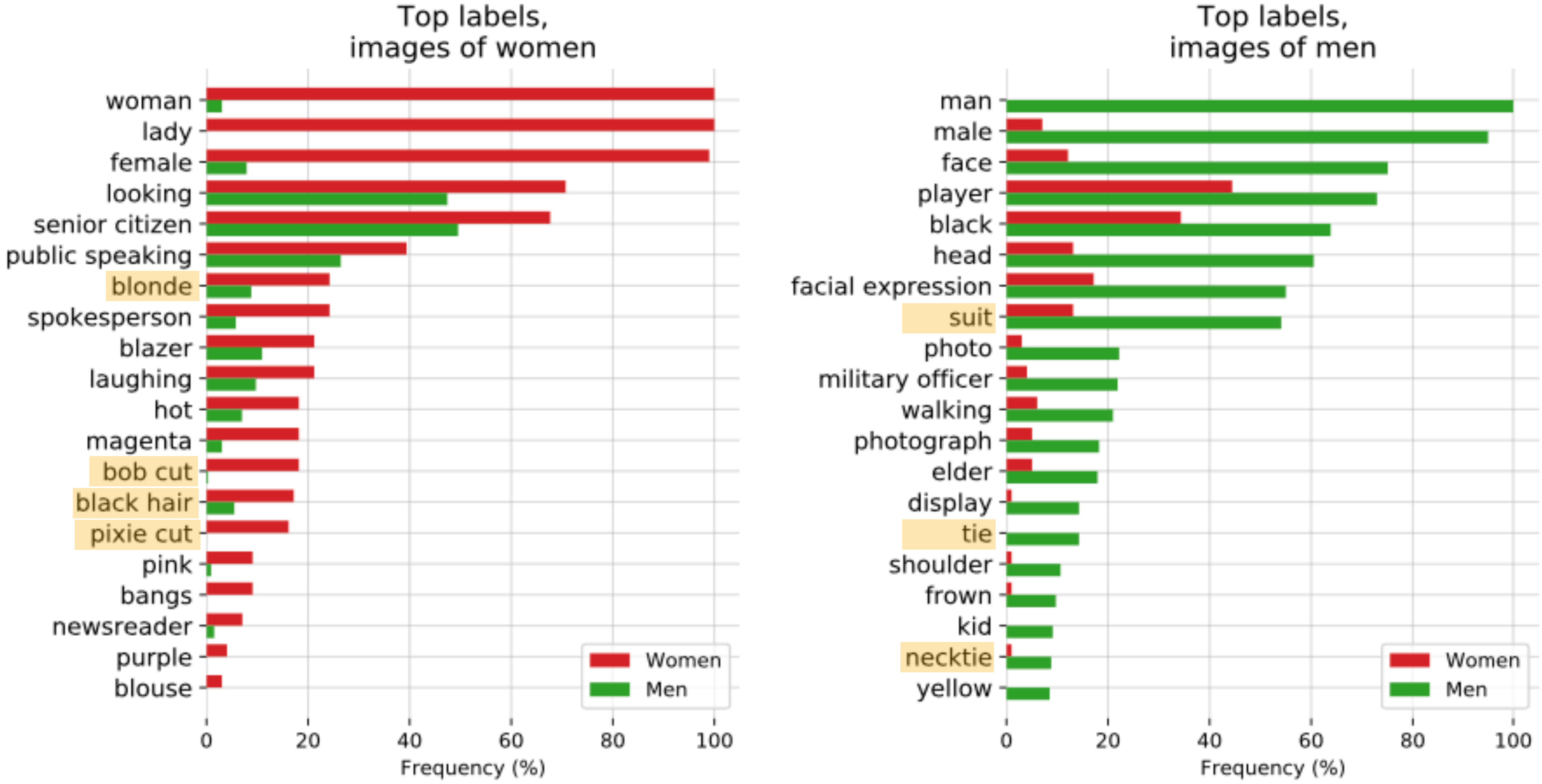
Gender study on congress

Construct label sets: (1) 300 occupations; (2) labels predicted by cloud vision services

Experiment: gender prediction with CLIP on members of congress (100% accuracy)

Influence of thresholds: At 4% probability threshold, highest probability occupation labels across genders were "lawmaker", "legislator", "congressman".

At 0.5% threshold: "nanny", "housekeeper" (women), "prisoner", "mobster" (men)



Label distribution on cloud vision label set at 0.5% threshold

Observation: analysis depends on thresholds

References/Image credits:

K. Karkkainen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation", WACV (2021)

A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)

Broader Impacts - surveillance

Surveillance

Experiment: Measure zero-shot classification on footage from CCTV cameras: VIRAT dataset (Oh et al., 2011) and video from Varadarajan et al. (2009).

Model tasked with predicting coarse-grained and fine-grained labels for images.

Coarse-grained labels: main subject of the image, such as "empty parking lot"

Fine-grained labels: smaller features, e.g. "person standing in the corner"

Coarse-grained accuracy across six labels (including hard negatives) was 51.1%

Fine-grained accuracy was near random.

Takeaway: CLIP is not outstanding on CCTV surveillance footage.

Celebrity Recognition

Zero-shot celebrity recognition: CelebA 8K images

Model	100 Classes	1k Classes	2k Classes
CLIP L/14	59.2	43.3	42.2
CLIP RN50x64	56.4	39.5	38.4
CLIP RN50x16	52.7	37.4	36.3
CLIP RN50x4	52.8	38.1	37.3

CelebA zero-shot Top-1 Identity Recognition

While far from SOTA, the results are notable since the names inferred solely from pre-training data.

Summary

Given existing specialised systems for surveillance, CLIP appeal for such tasks may be relatively low.

By removing the need for training data, it could enable bespoke surveillance systems for which there are no existing models/training data.

It could also lower the skill required to build these applications.

References/Image credits:

S. Oh et al., "A large-scale benchmark dataset for event recognition in surveillance video", CVPR, (2011)

J. Varadarajan and J-M. Odobez, "Topic models for scene analysis and abnormality detection", ICCVW (2009)

(CelebA) Z. Liu et al., "Deep learning face attributes in the wild", ICCV (2015)

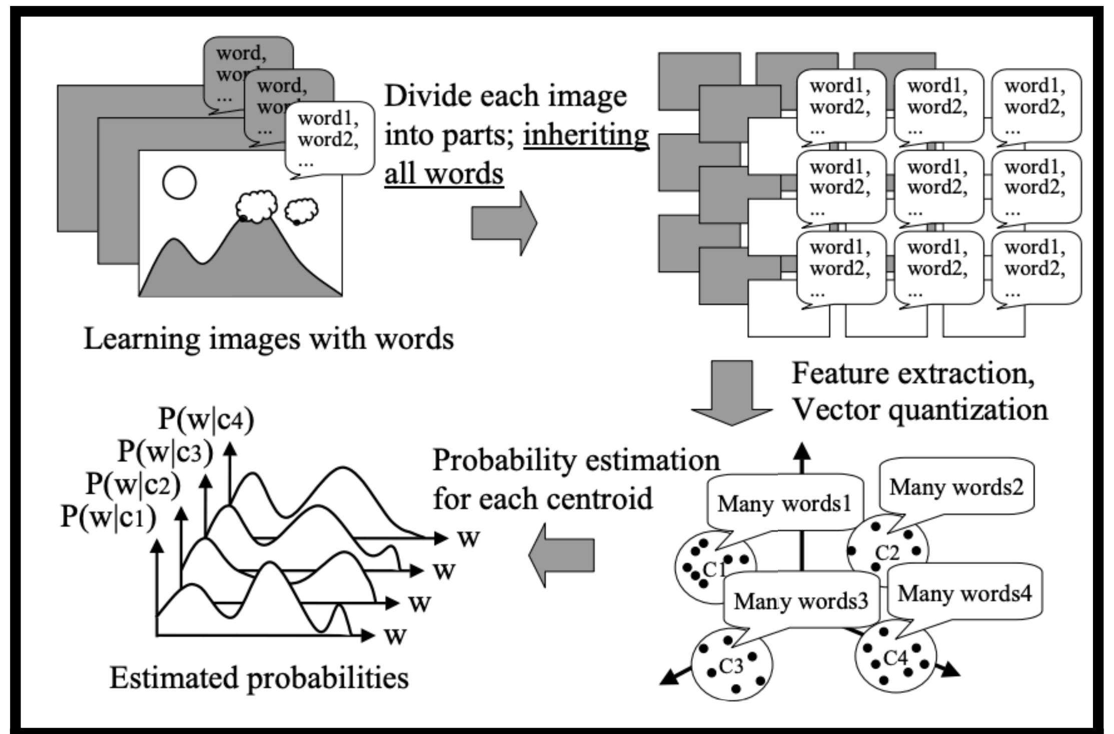
A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)

Outline

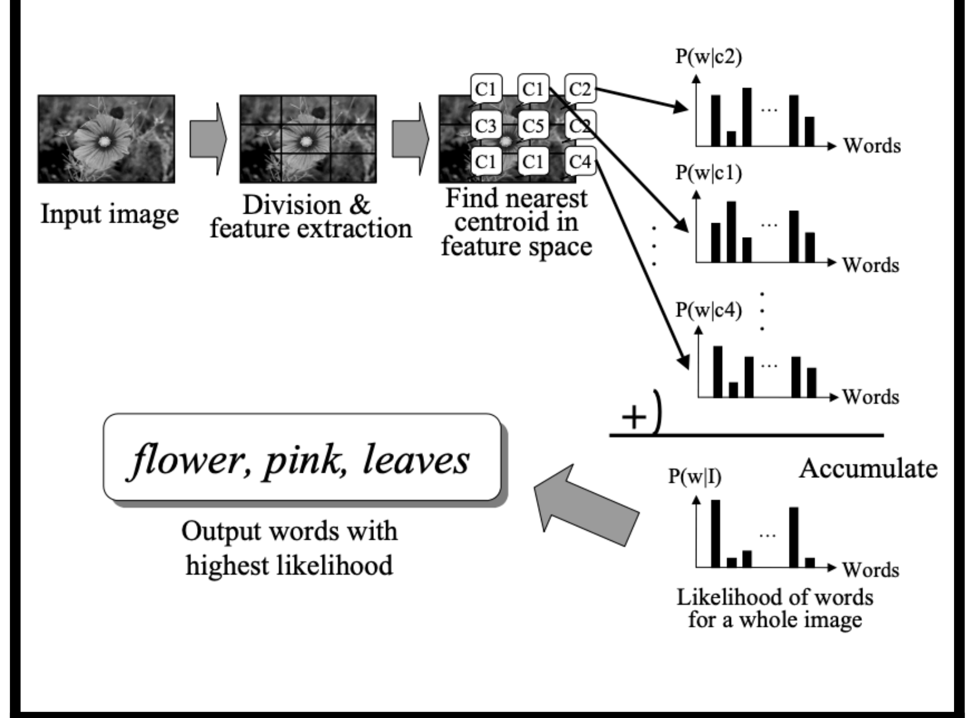
- Motivation
- CLIP: Data and Method
- Experiments
- Data Overlap Analysis
- Limitations
- Broader Impacts
- Related Work
- Summary

Related Work

Mori et al., (1999)



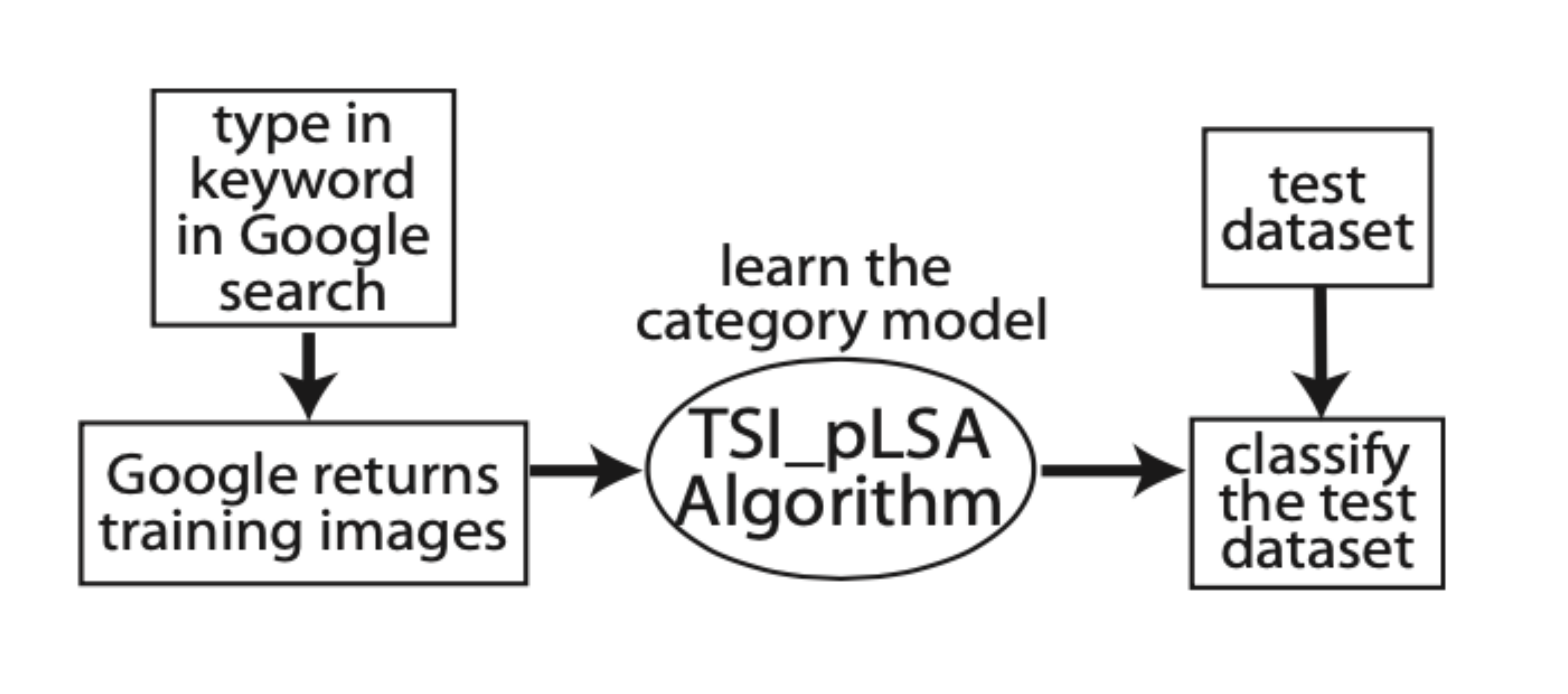
Learning



Inference

Image-to-word transformation

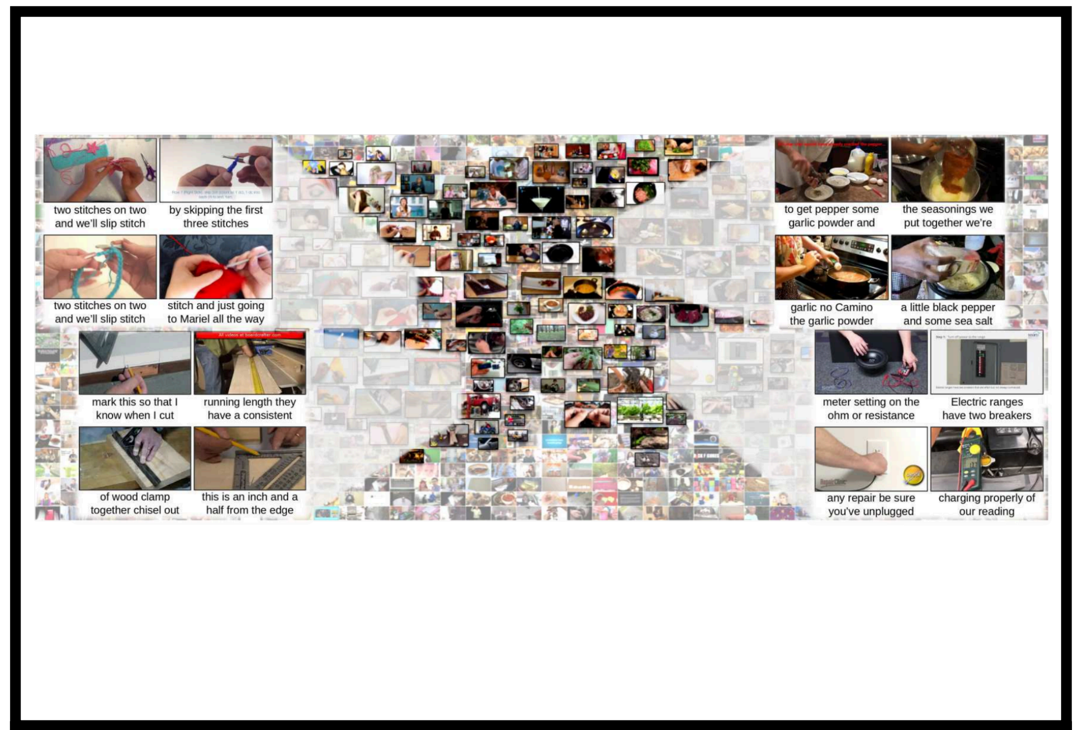
Fergus et al., (2005)



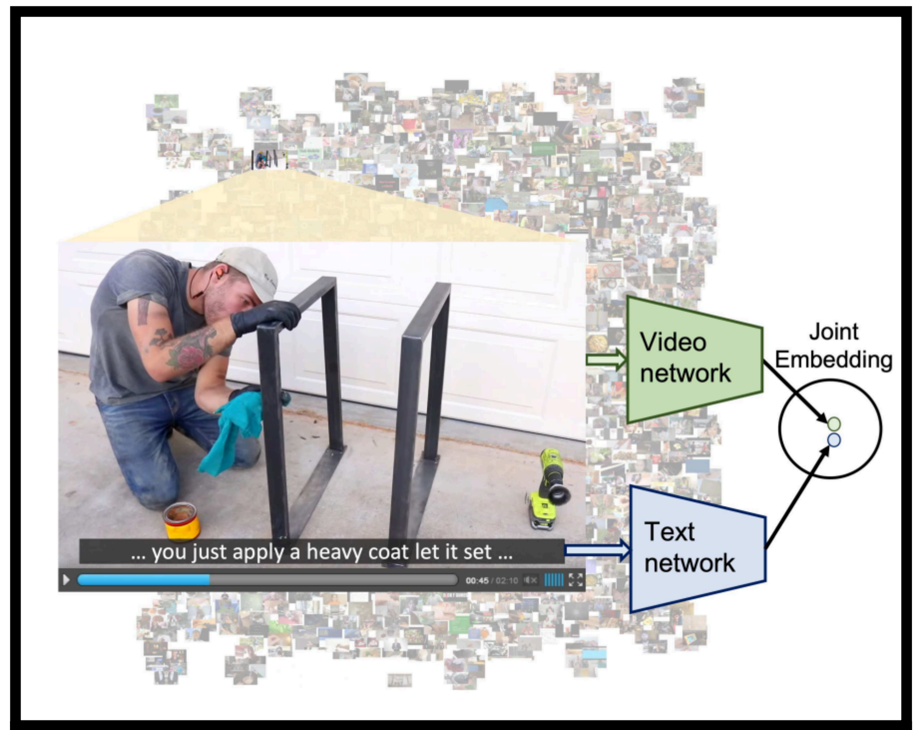
Framework

Webly-supervised learning

Miech et al., (2019)



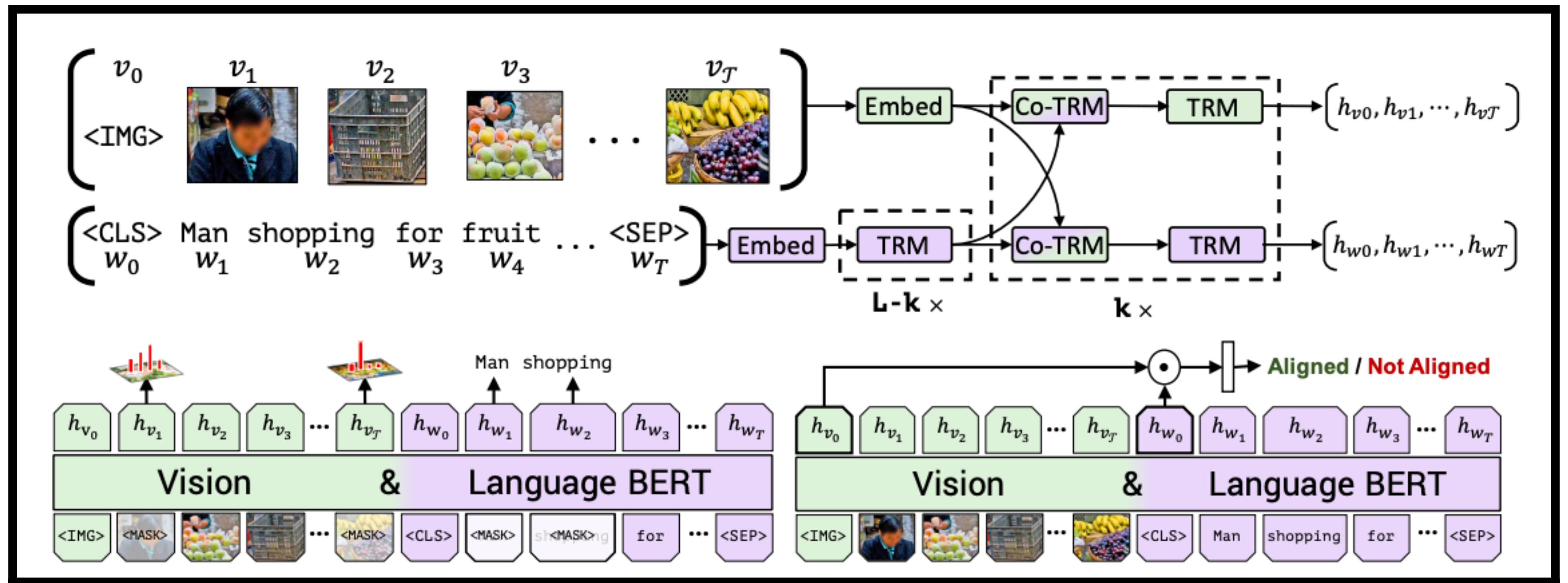
136 million video clips



Joint embedding

Vision/language pre-training

Lu et al., (2019)



Vision and language with interaction

Shared vision and language

References/Image credits:

Y. Mori et al., "Image-to-word transformation based on dividing and vector quantizing images with words", MISRM (1999)

A. Miech et al., "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips", ICCV (2019)

R. Fergus et al., "Learning object categories from google's image search", ICCV (2005)

J. Lu et al., "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks", NeurIPS (2019)

Outline

- Motivation
- CLIP: Data and Method
- Experiments
- Data Overlap Analysis
- Limitations
- Broader Impacts
- Related Work
- Summary

Summary

Takeaway

This work has investigated the feasibility of task-agnostic web-scale pre-training (shown to be effective in NLP) to computer vision.

It has shown computer vision also benefits from such an approach.

During pre-training, CLIP models learn a wide range of tasks.

This pre-training enables non-trivial zero-shot transfer to many datasets.

Reference:
A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)