Learning Transferable Visual Models From Natural Language Supervision (CLIP)

OpenAI 2021

Presented by Changyuan Qiu

February 10, 2021

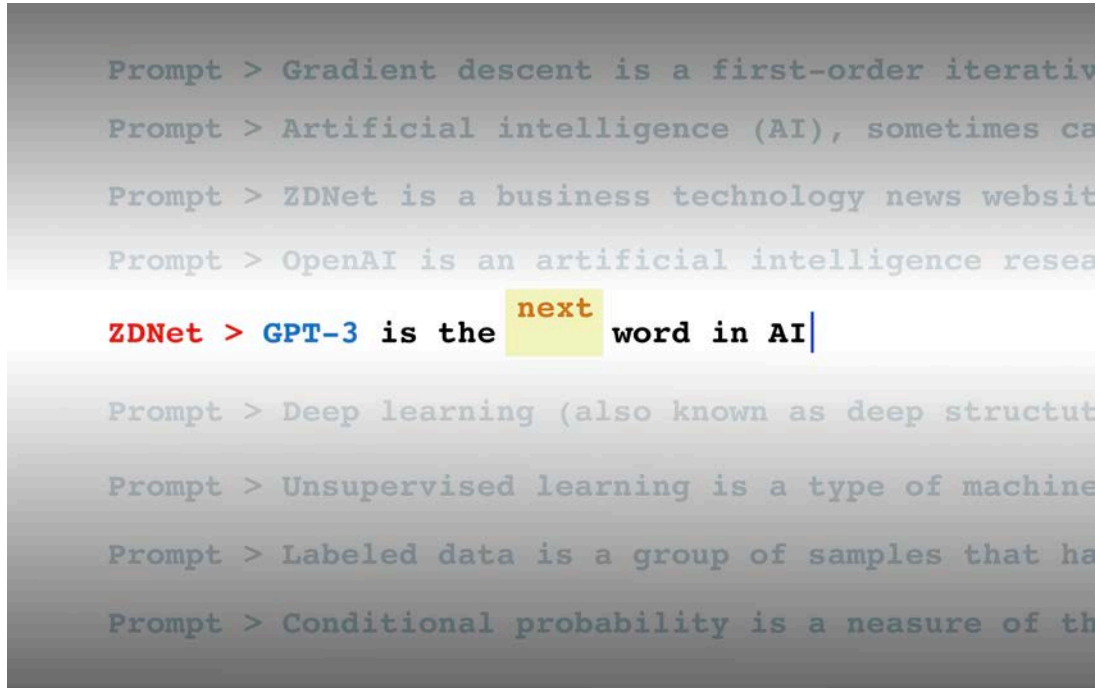
Motivations

- **Costly dataset:** Vision datasets are labor intensive and costly to create
  - ImageNet: 25,000 workers to annotate 14 million images for 22,000 object categories

- **Narrow:** Standard vision models are good at one task and one task only, and require significant effort (fine-tuning) to adapt to a new task
  - like a student who passed an exam by studying only the questions on past years' exams.

Motivations

- Unsupervised Generative Pre-training revolutionized *NLP*
- Flagship systems like *GPT-3 (Brown et al., 2020)* are now competitive across many zero-shot tasks requiring little/no dataset specific training data.
- Pre-training methods of large-scale **web text** data surpasses that of high-quality human-labeled NLP datasets



Motivations

- Could large-scale **web text** helps with pre-training in computer vision?
- *Learning Visual N-Grams From Web Data (Li et al., 2017)*
- Train a CNN to predict n-grams from images
- Matched target classes with n-grams in the dictionary
- For a given image, predict n-grams, score each class, and predict the highest-scoring class for that image

Predicted n-grams
lights
Burning Man
Mardi Gras
parade in progress

Predicted n-grams
GP
Silverstone Classic
Formula 1
race for the

Predicted n-grams
navy yard
construction on the
Port of San Diego
cargo

	aYahoo	Imagenet	SUN
Class mode (in dictionary)	15.3	0.3	13.0
Class mode (all classes)	12.5	0.1	8.6
Jelinek-Mercer (in dictionary)	88.9	35.2	34.7
Jelinek-Mercer (all classes)	72.4	11.5	23.0

Motivations

- Advance in architectures and pre-training approaches in learning visual representations from natural language supervision
- Auto-regressive language modeling: *VirTex (Desai & Johnson, 2020)*
- Masked language modeling: *ICMLM (Bulent Sariyildiz et al., 2020)*
- **Contrastive Objective (used in CLIP)**: *ConVIRT (Zhang et al., 2020)*

Dataset

Scale Difference

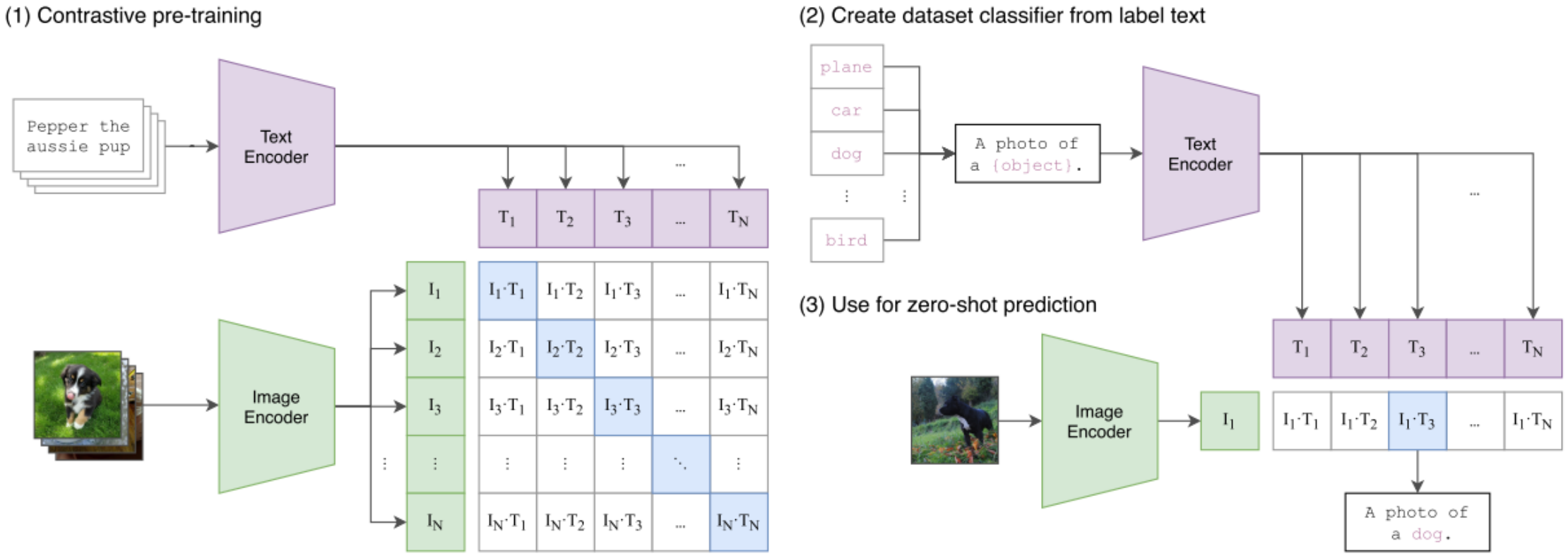
- WebText (dataset for GPT-2)
  - 40GB data, ~10B word count
- MS-COCO
  - 100K
- Visual Genome
  - 100K
- YFCC100M
  - 100M, only 15M after data-cleaning (similar size as ImageNet)

Dataset

- **WebImageText (WIT)**
  - 10B (image, text pair), similar scale as WebText used in GPT-2
  - 500000 queries, 20000 (image, text) pair for each query
  - query list: all words occurring >= 100 times in Wikipedia, augmented with common bi-grams as well as the names of all Wikipedia articles above a certain search volume.

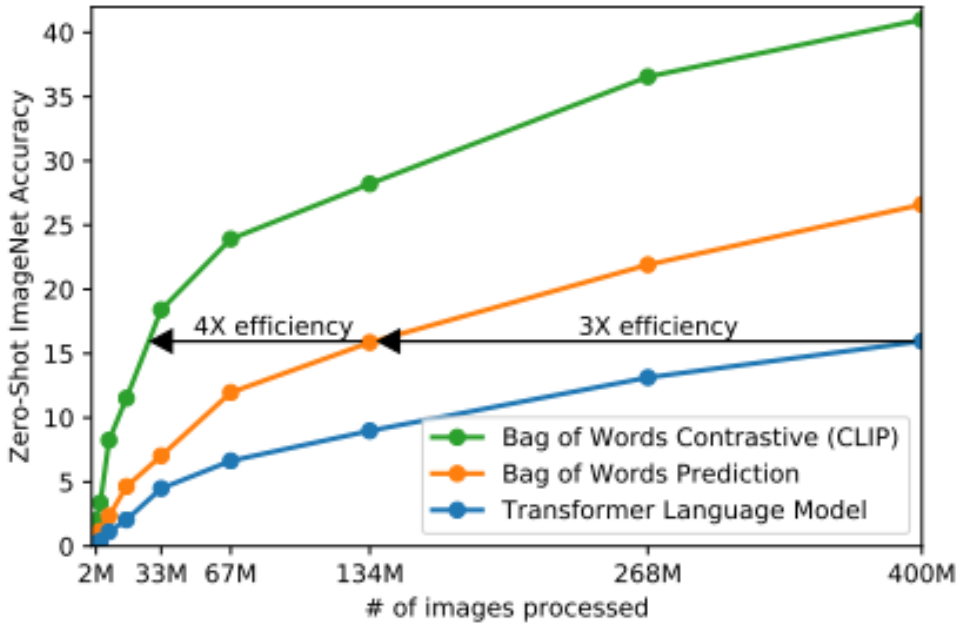
Approach

CLIP (Contrastive Language–Image Pre-training)



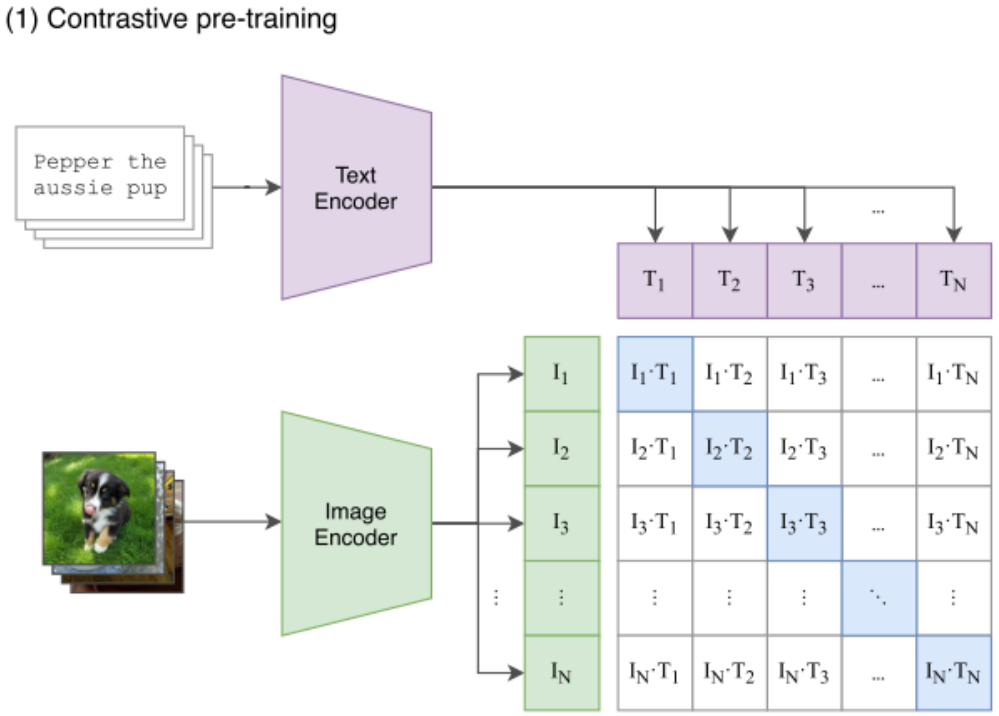
Approach

- **Contrastive Pre-training**
  - Initially try to trained an image CNN and text transformer jointly to **predict the caption** of an image (similar to VirTex)
  - Change objective to **predict Bag of Words** with 3X efficiency.
  - Change predictive objective to **contrastive objective** with 4X efficiency.



Approach

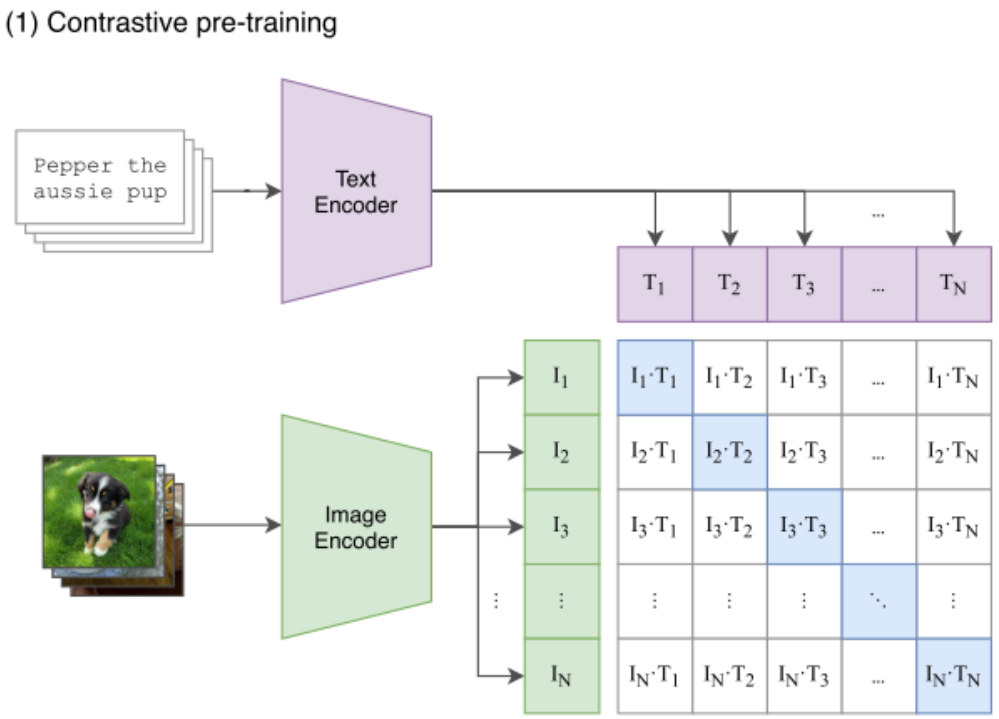
- **Contrastive Pre-training**
  - Given a batch of N (image, text) pairs, trained to predict which N of the $N \times N$ possible (image, text) pairings actually occurred.
  - Maximize the cosine similarity of the image and text embeddings of the N correct pairs while minimizing the cosine similarity of the $N^2 - N$ incorrect pairs.
  - Optimize a symmetric cross entropy loss over the similarity scores.



Approach

- **Contrastive Pre-training (InfoNCE loss)**
  - Suppose that I_i is paired with T_i

$$\mathcal{L}_i = -\log \frac{\exp(\langle I_i, T_i \rangle)/\tau}{\sum_{j=1}^{N} \exp(\langle I_i, T_j \rangle/\tau)}$$

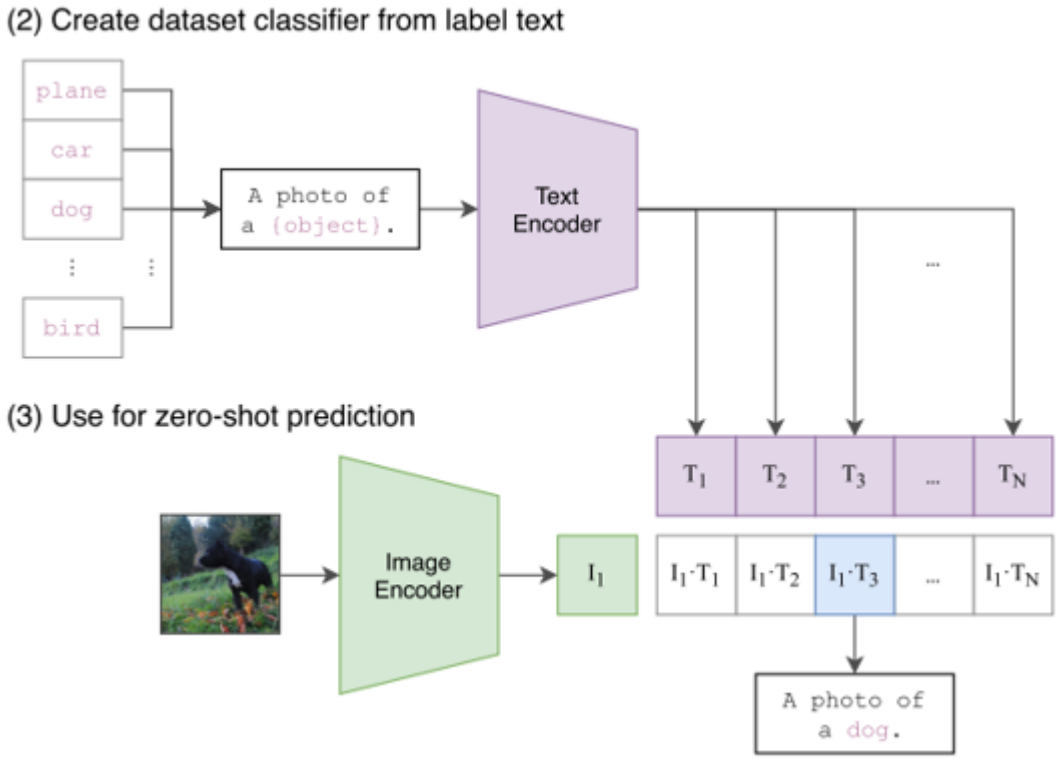


Approach

- **Contrastive Pre-training (More details)**
- Image Encoder
  - 5 ResNets (ResNet-50, ResNet-101, RN50x4, RN50x16, RN50x64)
  - 3 Vision Transformers (ViT-B/32, ViT-B/16, ViT-L/14)
- Text Encoder
  - Text Transformer adapted from *Language Models are Unsupervised Multitask Learners (Radford et al., 2019)*
- Adam Optimizer
- Mini-batch size: 32768, 2^{15}
- Temperature parameter: initialized at 0.07
- Training time: RN50x64 - 18 days on 592 V100 GPUs, ViT-L/14 - 12 days on 256 V100 GPUs

Experiments

- **Zero-Shot Transfer**
  - (2) All the classes in the dataset for evaluation are arranged in a specific format like “a photo of a {classname}” and fed into the text encoder.
  - (3) Feed the image into the image encoder, then CLIP performs a similarity searches and predict the best pair.

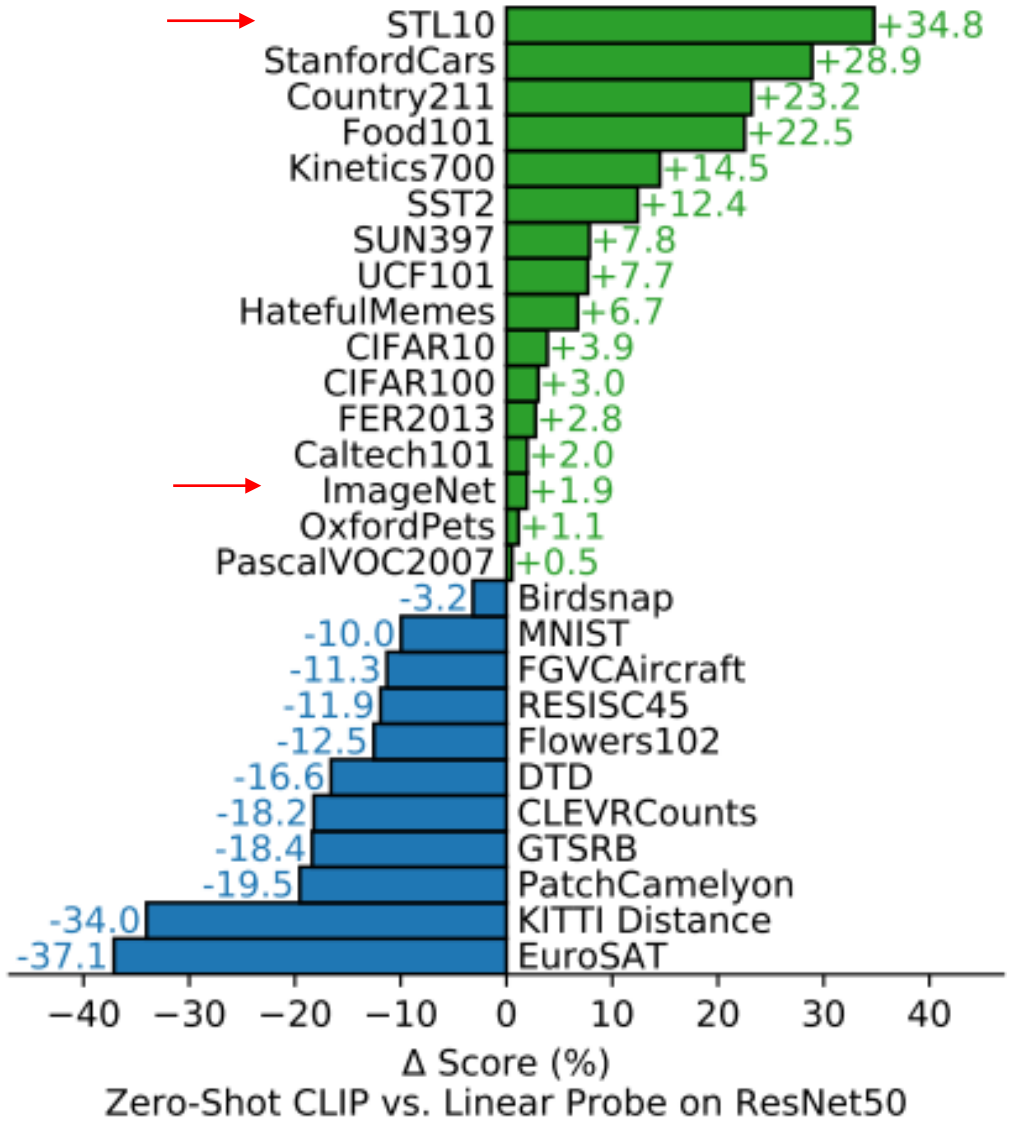


Experiments

Dataset	Classes	Train size	Test size	Evaluation metric
Food-101	102	75,750	25,250	accuracy
CIFAR-10	10	50,000	10,000	accuracy
CIFAR-100	100	50,000	10,000	accuracy
Birdsnap	500	42,283	2,149	accuracy
SUN397	397	19,850	19,850	accuracy
Stanford Cars	196	8,144	8,041	accuracy
FGVC Aircraft	100	6,667	3,333	mean per class
Pascal VOC 2007 Classification	20	5,011	4,952	11-point mAP
Describable Textures	47	3,760	1,880	accuracy
Oxford-IIIT Pets	37	3,680	3,669	mean per class
Caltech-101	102	3,060	6,085	mean-per-class
Oxford Flowers 102	102	2,040	6,149	mean per class
MNIST	10	60,000	10,000	accuracy
Facial Emotion Recognition 2013	8	32,140	3,574	accuracy
STL-10	10	1000	8000	accuracy
EuroSAT	10	10,000	5,000	accuracy
RESISC45	45	3,150	25,200	accuracy
GTSRB	43	26,640	12,630	accuracy
KITTI	4	6,770	711	accuracy
Country211	211	43,200	21,100	accuracy
PatchCamelyon	2	294,912	32,768	accuracy
UCF101	101	9,537	1,794	accuracy
Kinetics700	700	494,801	31,669	mean(top1, top5)
CLEVR Counts	8	2,000	500	accuracy
Hateful Memes	2	8,500	500	ROC AUC
Rendered SST2	2	7,792	1,821	accuracy
ImageNet	1000	1,281,167	50,000	accuracy

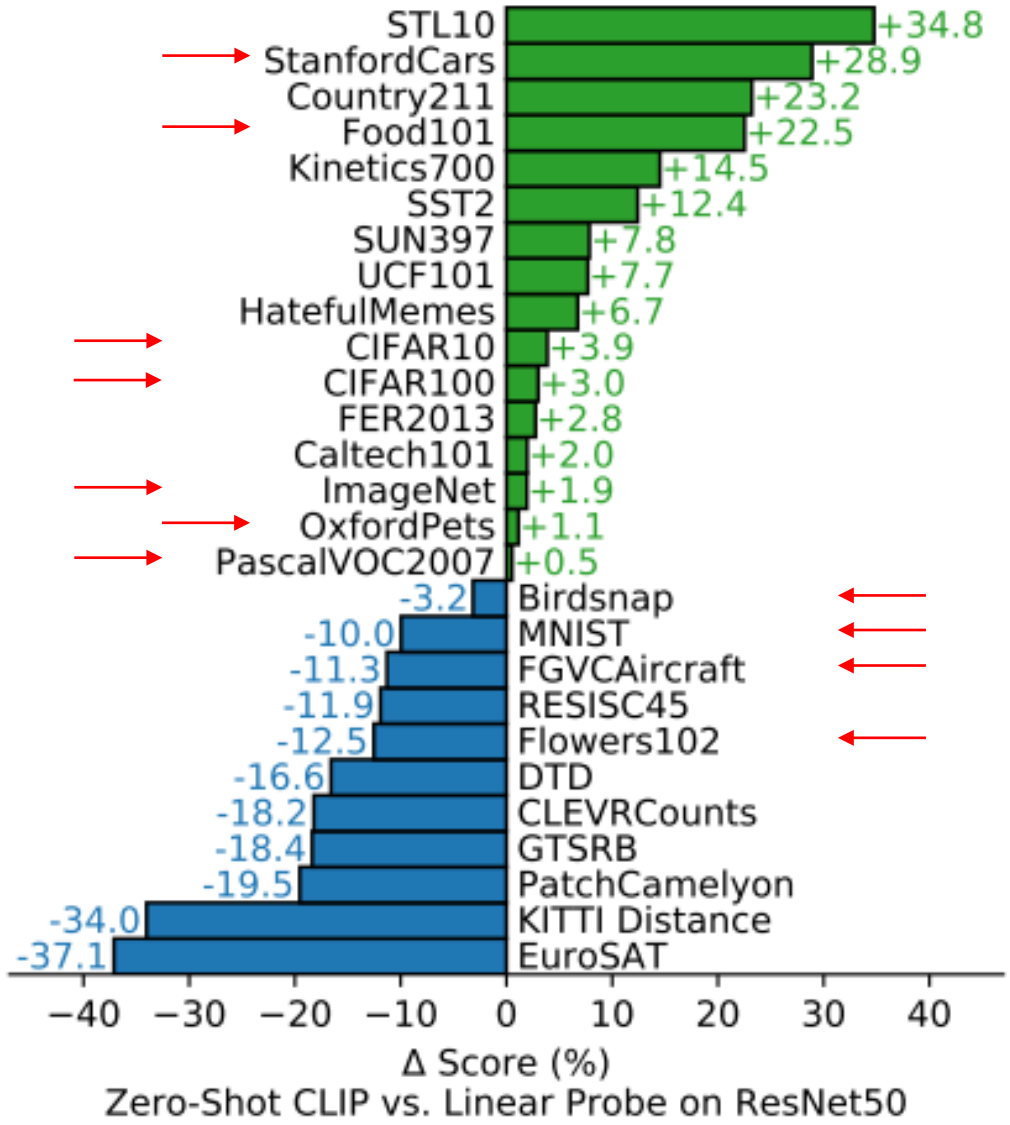
Experiments

- **Zero-Shot Transfer**
  - Compared with a fully supervised baseline (ResNet50, trained on ImageNet) across 27 datasets.
  - Outperforms supervised ResNet50 **without seeing any of the data in ImageNet**
  - Achieve SOTA on STL10, a dataset designed to encourage **efficient learning** by providing only a limited number of labeled examples



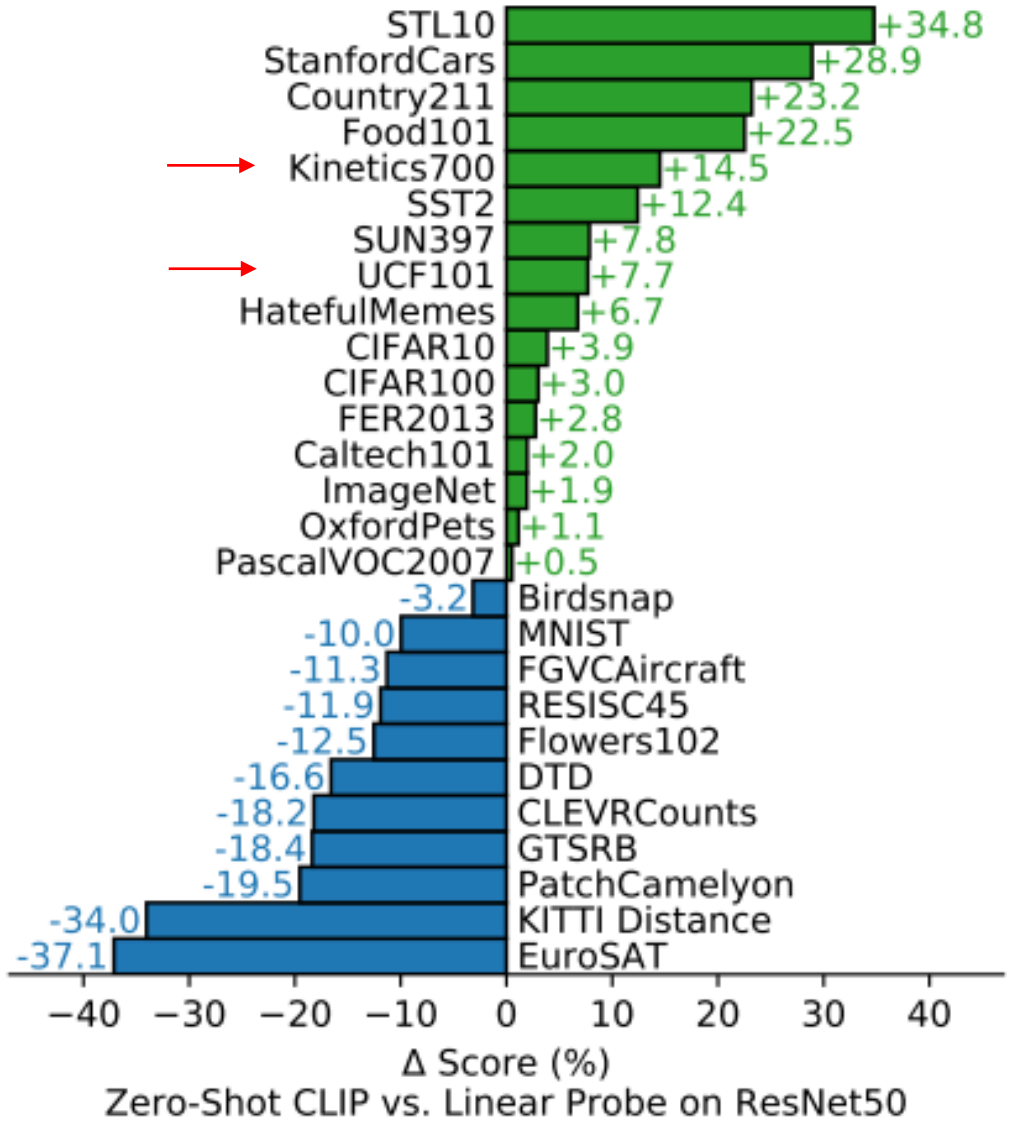
Experiments

- **Zero-Shot Transfer**
  - **Task-Specific Object Detection**
    - Outperforms on StanfordCars and Food101
    - Underperforms on FGVCAirCraft, Flowers102 and MNIST
    - Matches on OxfordPets and Birdsnap
    - Difference comes from varying task-specific supervision in WIT and ImageNet
  - **General Object Detection**
    - Outperforms slightly on CIFAR10,CIFAR100, ImageNet, and PascalVOC2007



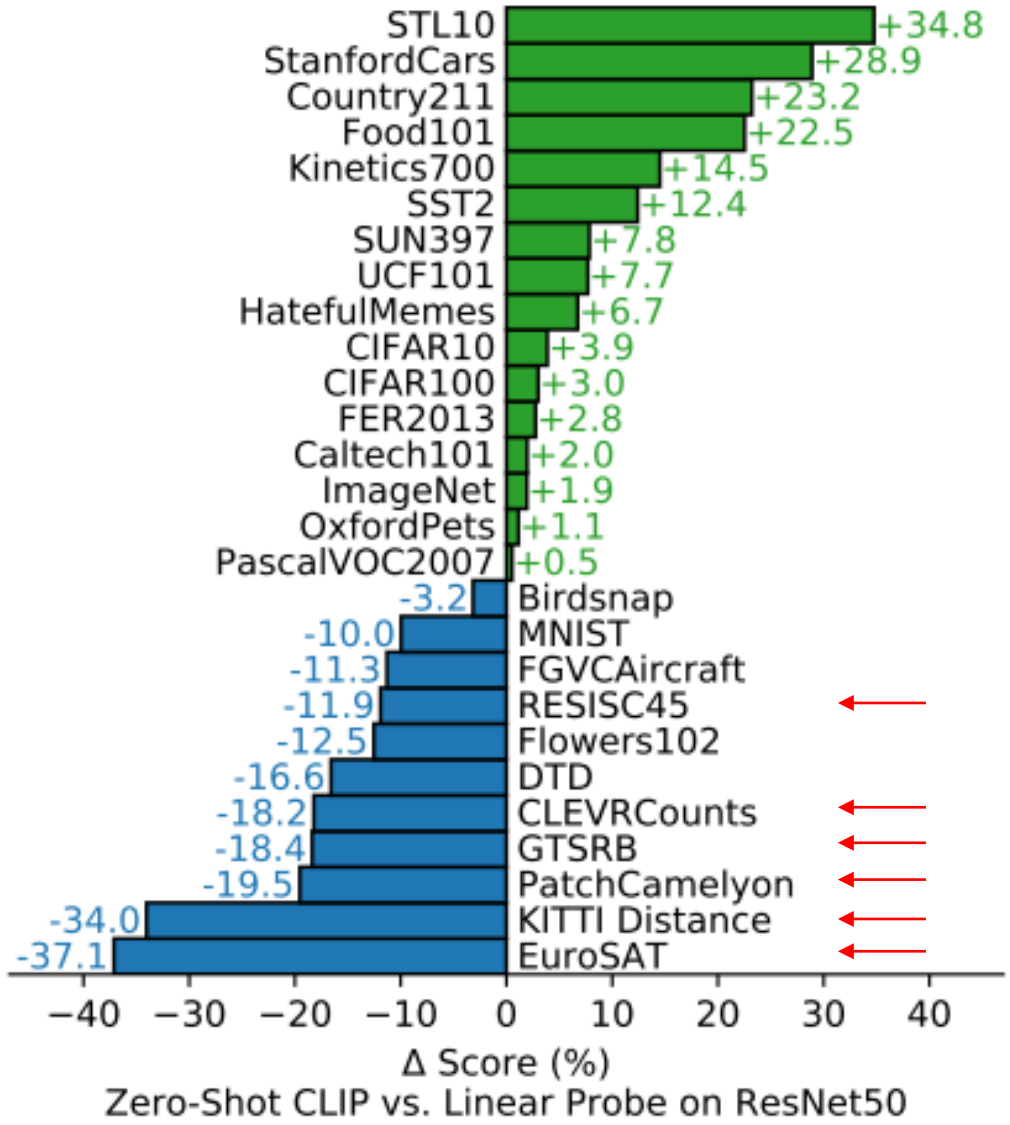
Experiments

- **Zero-Shot Transfer**
  - Outperforms baseline on 2 datasets measuring **action recognition in videos.**
  - Intuition: Natural language provide wider supervision for visual concepts involving **verbs**, compared to the **noun-centric** object supervision in ImageNet



Experiments

- **Zero-Shot Transfer**
  - Underperforms on **specialized, complex, abstract tasks.**
  - Satellite image classification (EuroSAT and RESISC45)
  - Tumor detection (PatchCamelyon)
  - Counting objects in synthetic scenes (CLEVRCounts)
  - Self-driving related tasks (GTSRB, KITTI Distance)
  - **Intuition:** whether zero-shot transfer, as opposed to few-shot or supervised ones, could really performs well on complex tasks without any “prior experience”.



Experiments

IMAGENET

King Charles Spaniel (91.6%) Ranked 1 out of 1000

✓ a photo of a **king charles spaniel**.

× a photo of a **brittany dog**.

× a photo of a **cocker spaniel**.

× a photo of a **papillon**.

× a photo of a **sussex spaniel**.

CLEVR COUNT

4 (17.1%) Ranked 2 out of 8

× a photo of **3** objects.

✓ a photo of **4** objects.

× a photo of **5** objects.

× a photo of **6** objects.

× a photo of **10** objects.

PASCAL VOC 2007

motorcycle (99.7%) Ranked 1 out of 20

✓ a photo of a **motorcycle**.

× a photo of a **bicycle**.

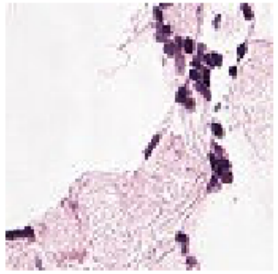
× a photo of a **car**.

× a photo of a **horse**.

× a photo of a **dining table**.

PATCHCAMELYON (PCAM)

healthy lymph node tissue (22.8%) Ranked 2 out of 2

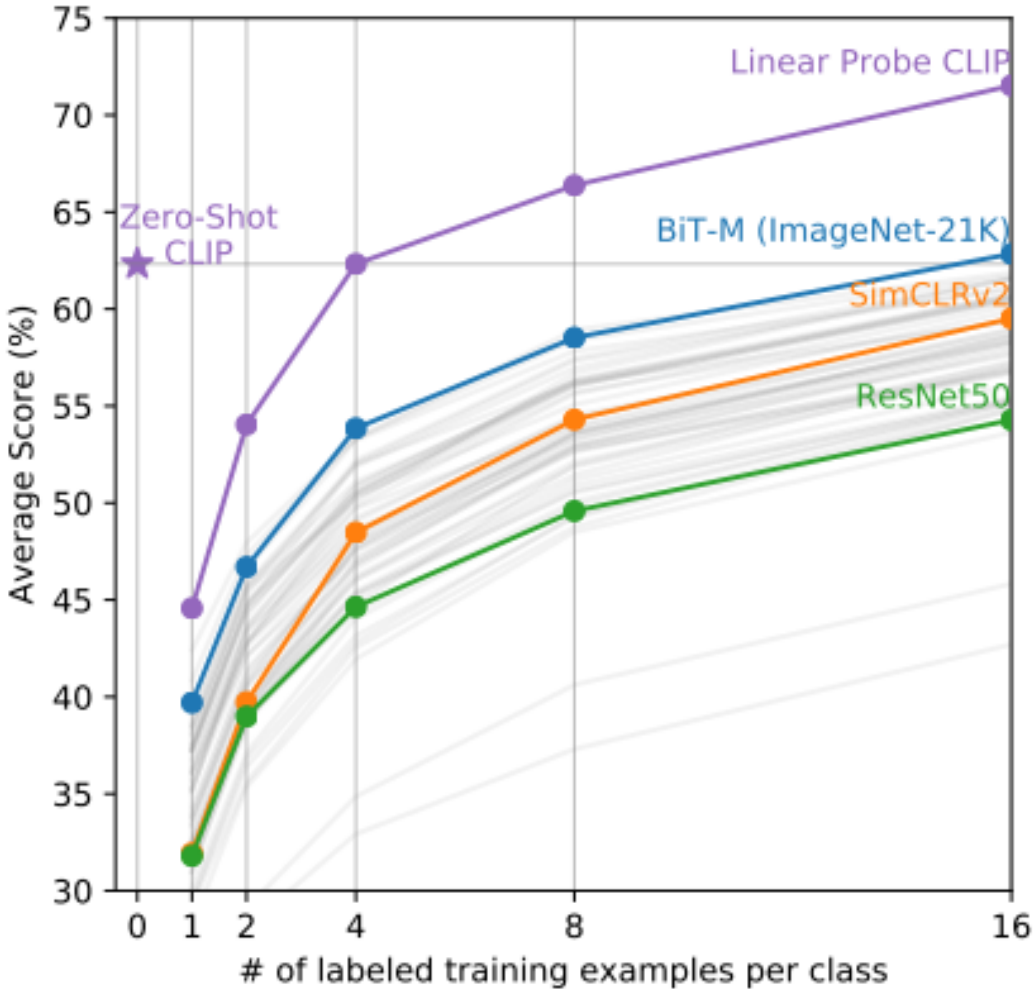


× this is a photo of **lymph node tumor tissue**

✓ this is a photo of **healthy lymph node tissue**

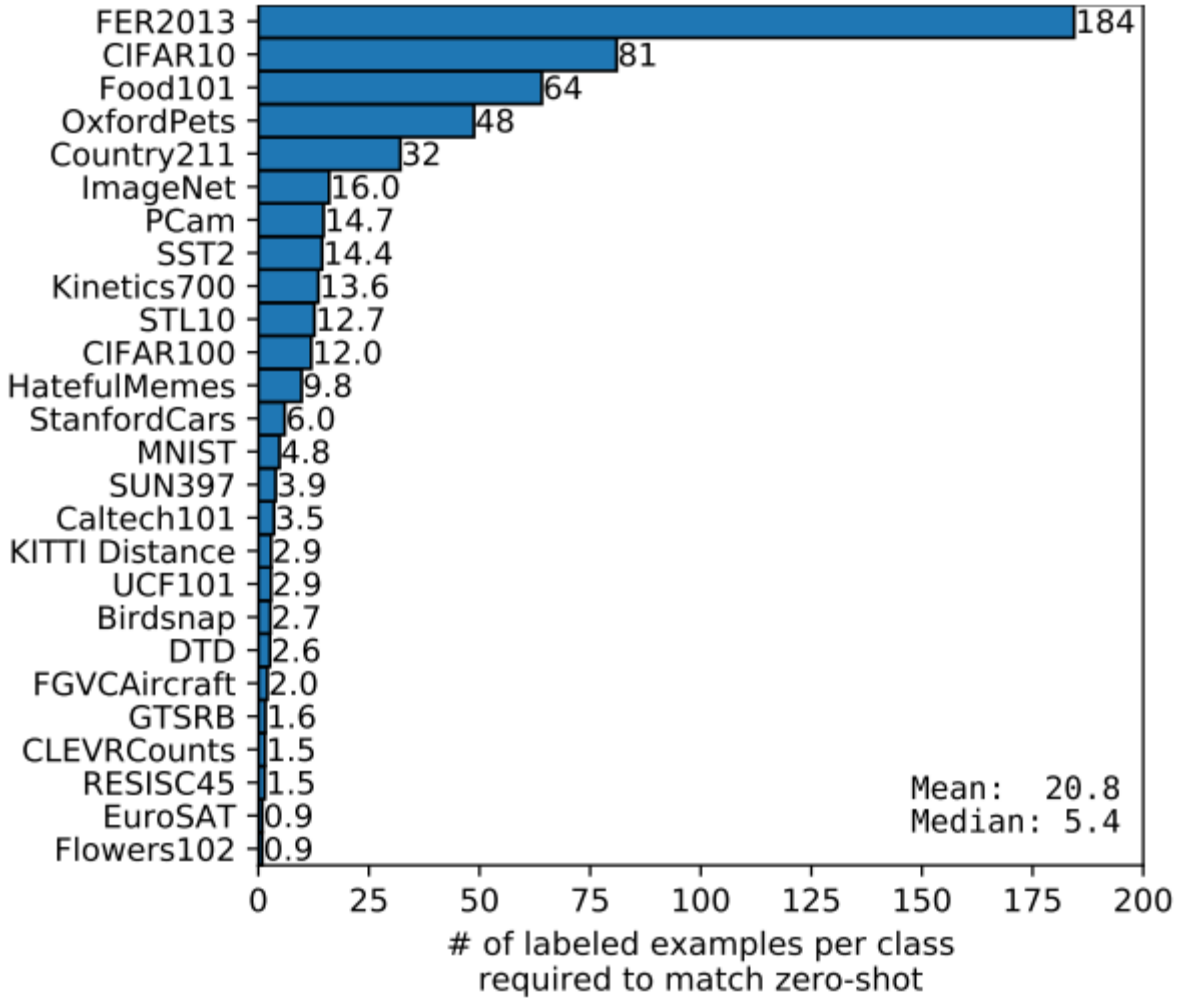
Experiments

- **Zero-Shot Transfer**
  - Matches the best results of a 16-shot linear classifier across publicly available models (BiT-M)
  - **Zero-shot outperforms one-shot!**
  - Matches the average performance of a 4-shot linear classifier trained on the same feature space



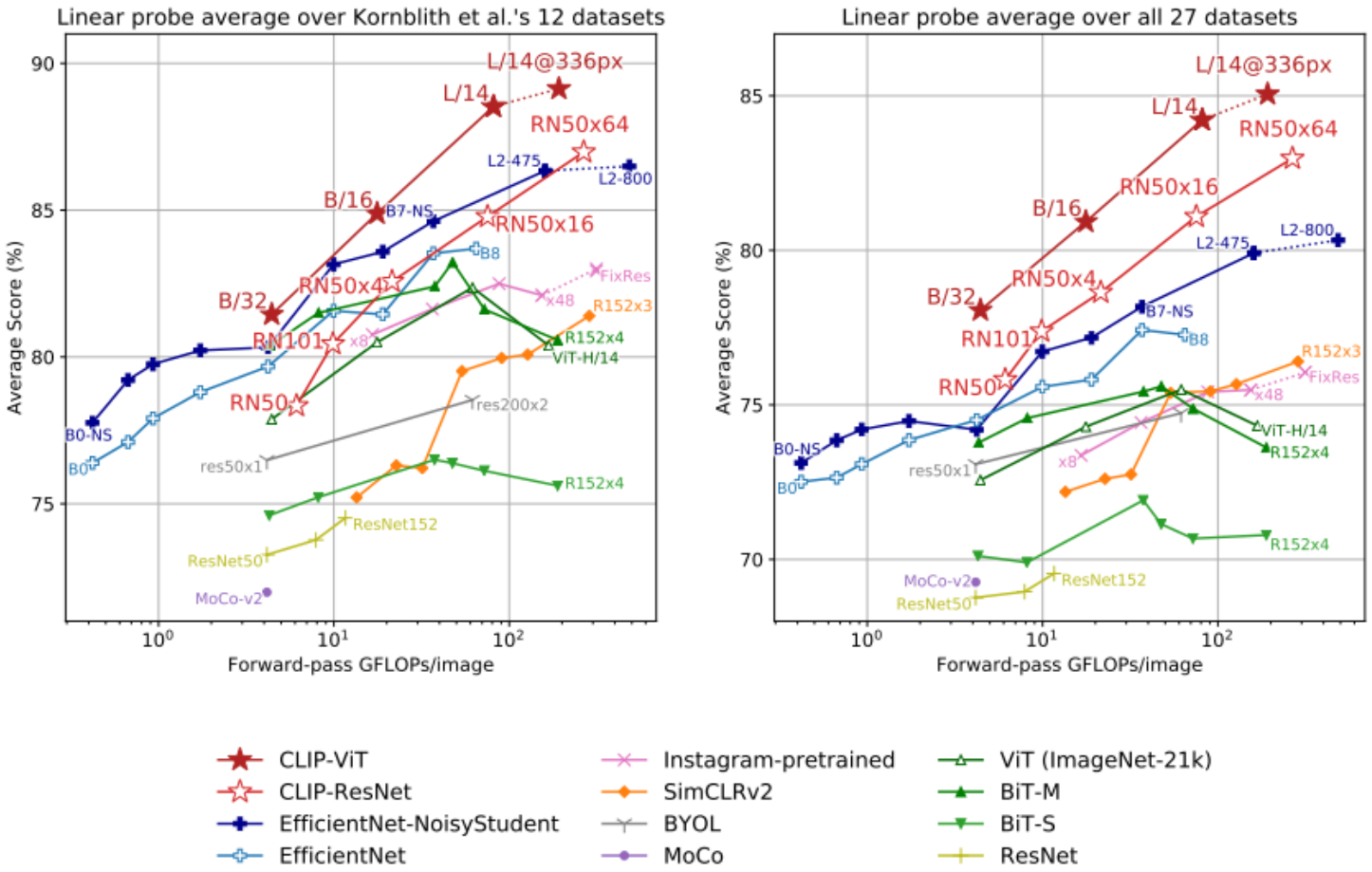
Experiments

- **Zero-Shot Transfer**
  - **Data efficient:**
    - N-shot matches zero-shot
    - Mean: 20.8
    - Median: 5.4
    - ImageNet: 16



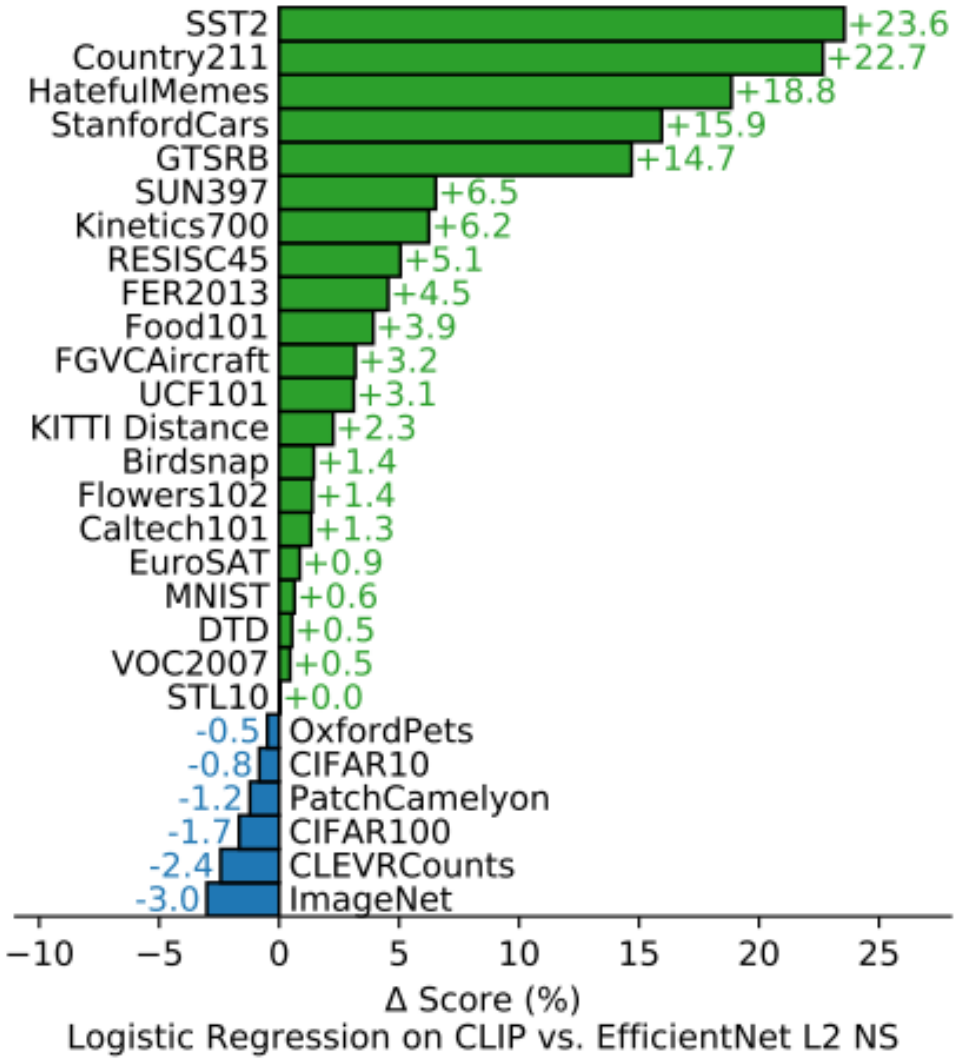
Experiments

- **Representation Learning - Linear Probing**
  - Vision Transformer are about 3x more compute efficient than ResNets, consistent with the findings of *An image is worth 16x16 words: Transformers for image recognition at scale (Dosovitskiy et al., 2020)*
  - Fine-tuned ViT-L/14 at 336x336 px for 1 additional epoch
  - Evaluated across 27 datasets
  - Achieved SOTA on 21 datasets



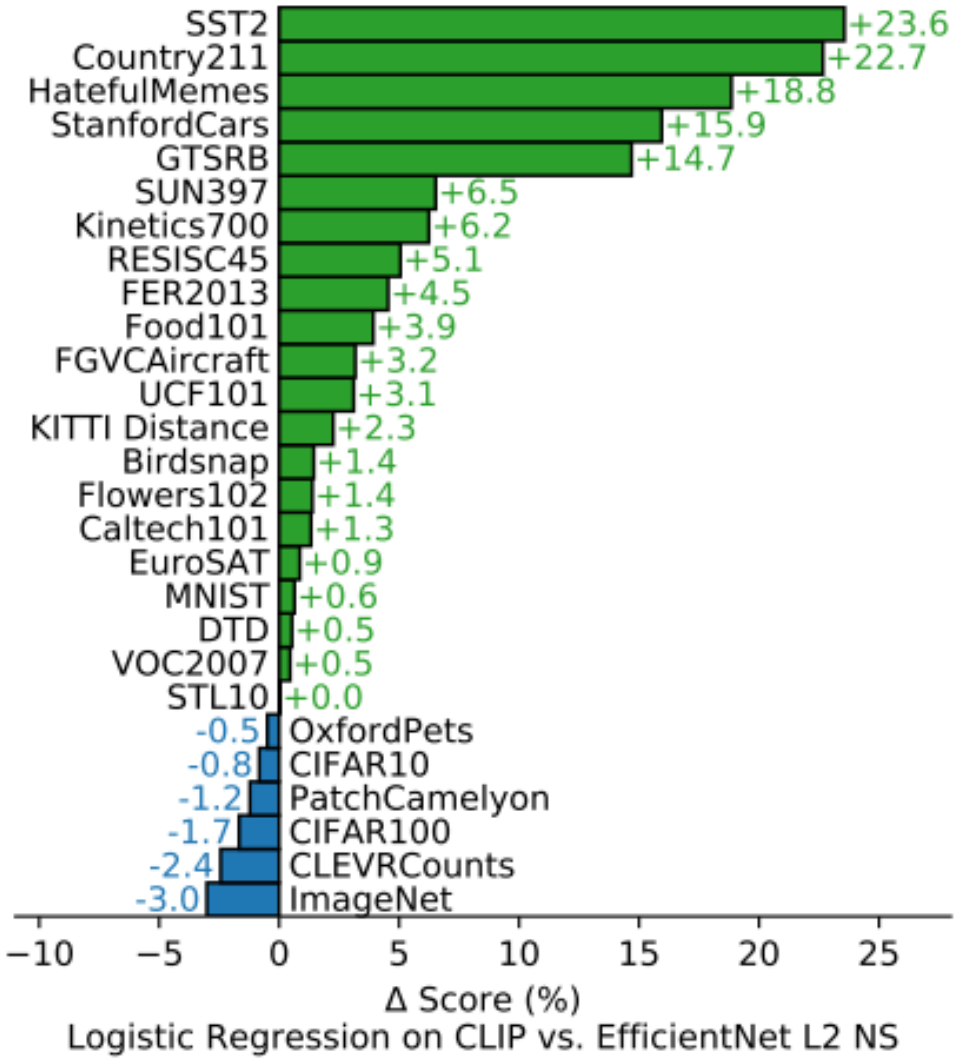
Experiments

- **Representation Learning - Linear Probing**
  - CLIP outperforms the current best ImageNet model (Noisy Student EfficientNet-L2) on 21 out of 27 datasets.
  - Outperforms on
    - OCR (SST2 and HatefulMemes)
    - Geo-localization and scene recognition (Country211, Sun397)
    - Activity recognition in videos (Kinetics 700 and UCF101)
    - **Traffic Sign Recognition (GTSRB)**



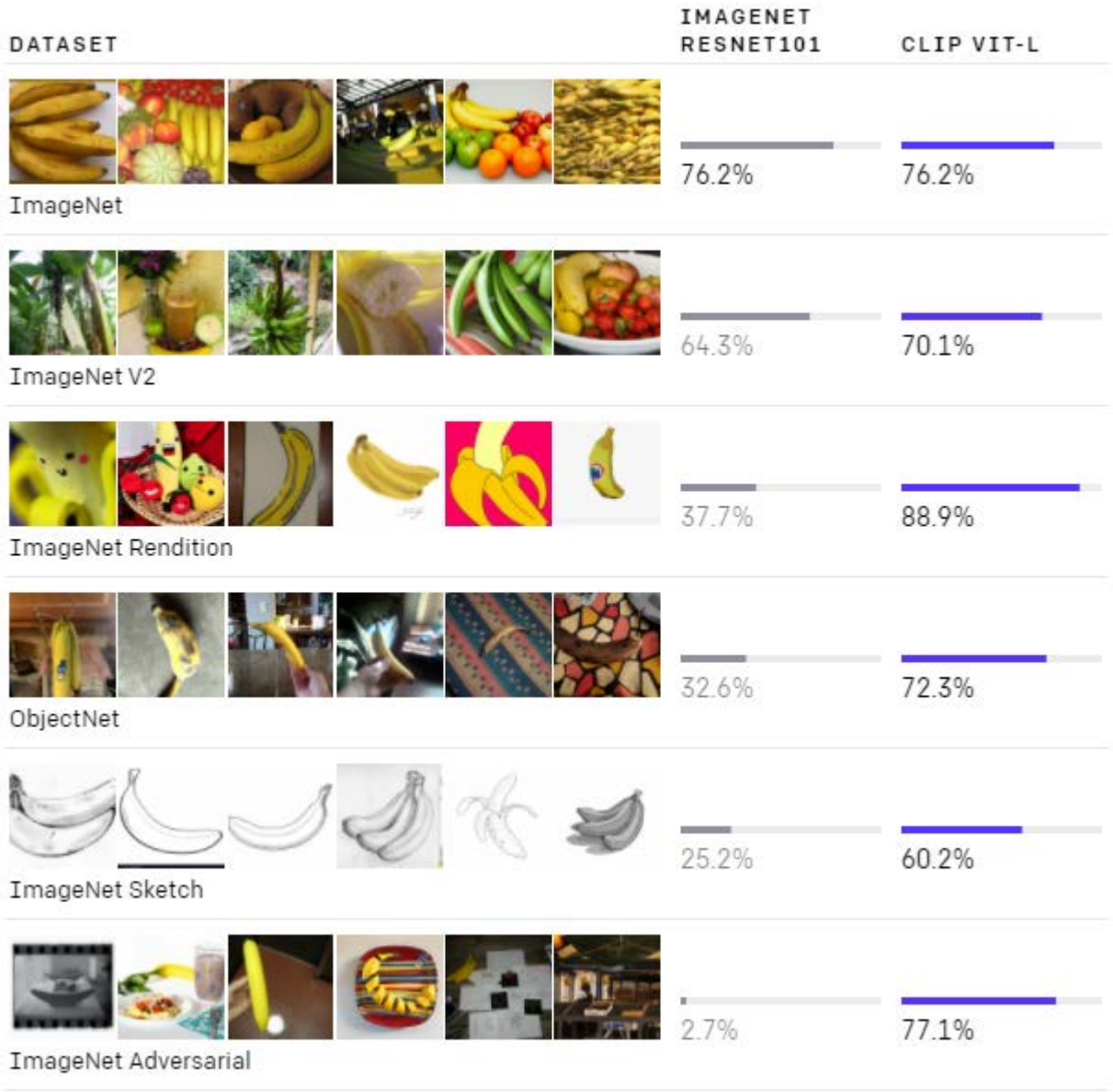
Experiments

- **Representation Learning - Linear Probing**
  - Underperforms on ImageNet
  - Underperforms on low-resolution datasets (CIFAR10, CIFAR100), intuition: lack of scale-based data augmentation



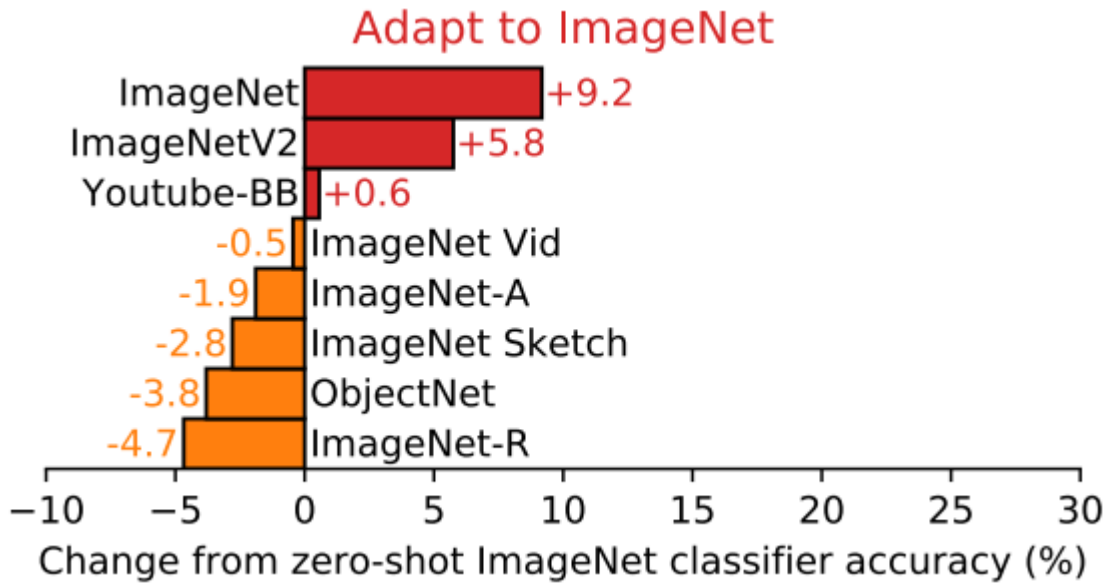
Experiment

- **Robustness**
  - Compare zero-shot ViT-L/14@336px with supervised ResNet101, which has matched ImageNet performance.



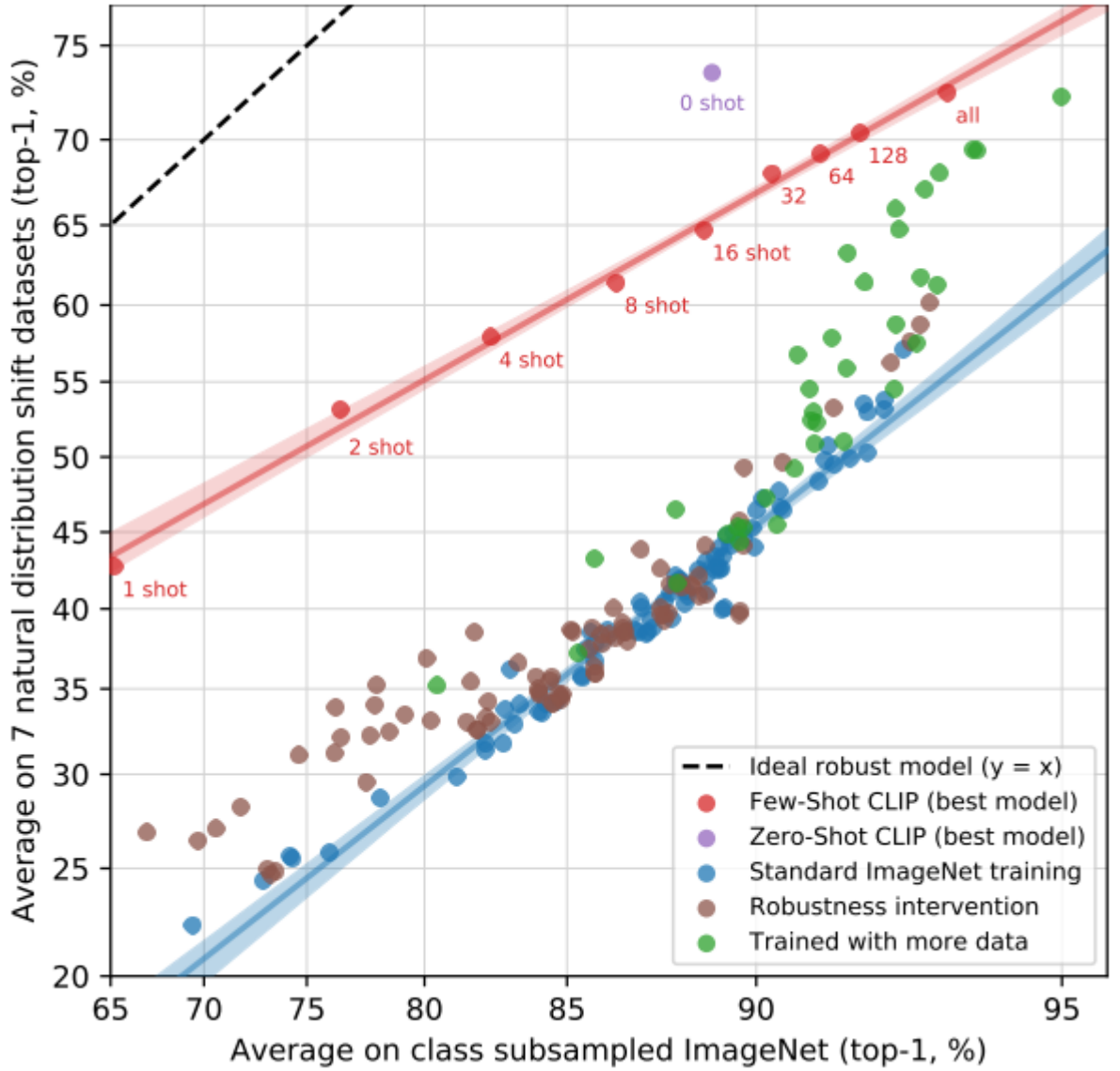
Experiment

- **Robustness**
  - Adapted CLIP to the ImageNet Distribution via a L2 regularized logistic regression classifier fit to CLIP features on the ImageNet training set.
  - Accuracy on ImageNet increases by 9.2% while average accuracy across 7 dataset slightly decreases.
  - Fitting to the ImageNet is not robust to natural distribution shift.



Experiment

- **Robustness**
  - Visualize the performance of zero-shot, 2^n shot and fully supervised logistic regression classifier on CLIP
  - Robustness gap between CLIP and existing models shrinks as getting “more supervised”
  - High effective robustness seems to result from minimizing the amount of distribution specific training data a model has access to (zero-shot gives best robustness), but this comes at a cost of reducing dataset-specific performance



Summary

- Pros:
  - Abundant web image-text data vs. “Golden” human-labeled data
  - Generalize well to a series of tasks
  - Compute Efficient (Bag of words + Contrastive Objective + Vision Transformer)
- Cons:
  - Web data are unfiltered and uncurated, and may exhibit social bias.
  - Performance still well below supervised SOTA
  - Struggles on more abstract or complex tasks (e.g. counting objects, predicting tumor)
  - Poor generalization to images not covered in its pre-training dataset (88% on MNIST, worse than a simple logistic regression)

Discussion

- Should we really consider language-based supervision to be “unsupervised“?
- What’s the main advantage of web data compared with human-labeled data? What about disadvantages?
- Does language supervised visual pre-training really generalize well or is it mainly the power of data?
- What do you think of the future of language supervised visual pre-training?