Learning Transferable Visual Models From Natural Language Supervision

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever,

OpenAI

Summary

- Introduces **Contrastive Language-Image Pre-training** (CLIP), an efficient learning method from natural language supervision, to computer vision tasks.

- Demonstrates impressive transfer learning/zero-shot capabilities using CLIP.

- Concise summary of concepts: https://openai.com/blog/clip/#rf35

Motivation

- DNNs for image classification still fail when tested on tasks outside of the training domain.
    - Networks are only optimizing for the training/test benchmark
- In contrast, deep learning for natural language processing (NLP) has been successfully deployed.
    - ‘text-to-text’ pretraining on task-agnostic architectures shows great transferability to downstream tasks.
    - Pre-training on big, diverse data = robust understanding.
- Why are NLP training methods much more robust than those in computer vision?
    - NLP uses millions of random text snippets; CV relies on hand-crafted labels in narrow datasets

Natural Language Supervision

- Numerous methods have tried applying NLP-type training to image tasks.
- Uses natural language associated with an image to create a label/task
- General approach: Use internet images to predict a word in the caption.
- Problems:
  - Requires extensive pre-training
  - Still performs poorly on zero-shot image tasks

Improving Natural Language Supervision

- Current metadata for images is varying in quality control.
    - Many tags are not relevant to photo content.
    - Authors created new dataset of 400M image/text pairs from 0.5M different internet queries.

- Predicting exact words in a tag is often ill-posed.
    - Many words can fit an image and context
    - Authors instead use a contrastive learning task

- Ex: Li et al. only reaches 11.5% accuracy on (zero-shot) ImageNet using this.

Key Idea: Use Contrastive Learning

- Minibatch of images and captions are obtained
- Text captions are encoded via transformer model
- Images are encoded via feature extractor
  - Either ResNet50 or vision transformer

- *Contrastive loss* maximizes cosine similarity between image/text pairs.
  - This works much better than past objectives

Key Idea

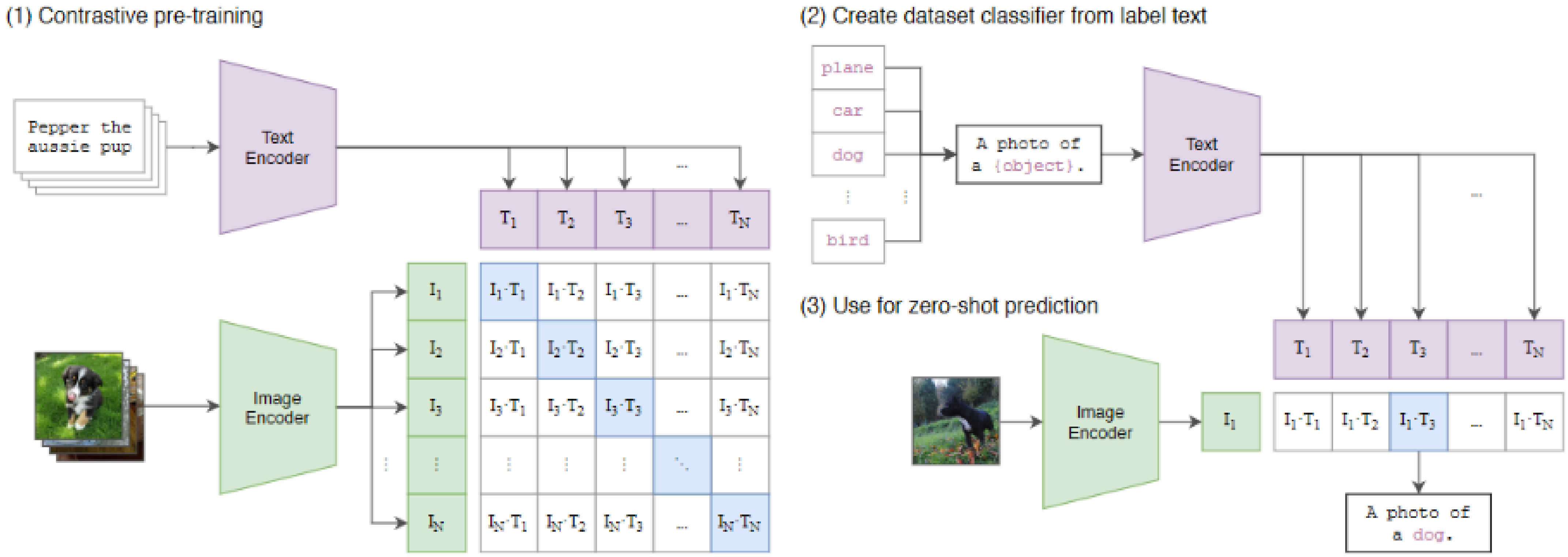
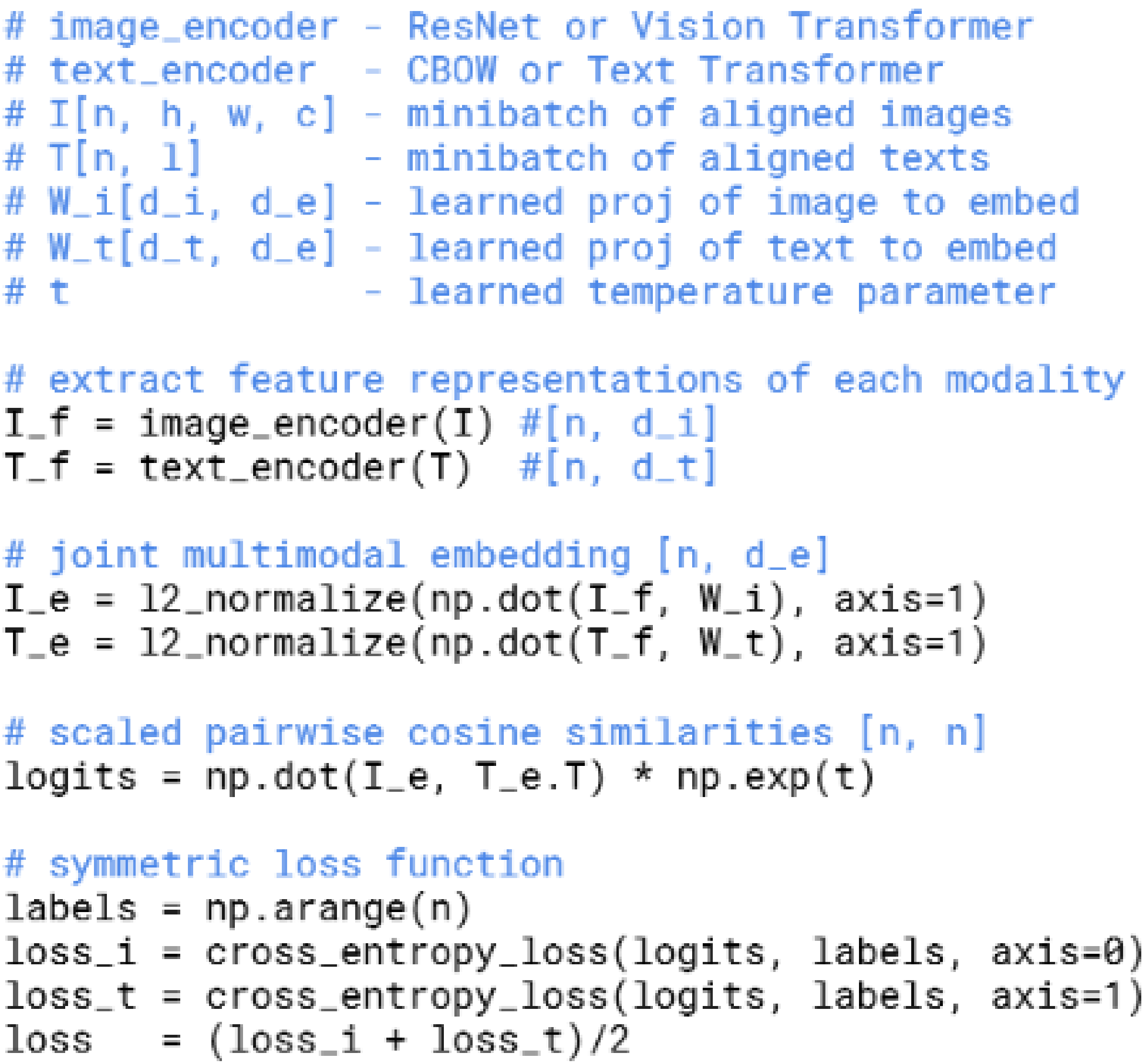

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
T_1 T_2 T_3 ... T_N
Image Encoder
I_1 I_2 I_3 ... I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
T_1 T_2 T_3 ... T_N
(3) Use for zero-shot prediction
Image Encoder
I_1
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

Key Idea

- Since cosine similarity scaled by temperature τ is used, in a sense, comparison is a multiclass logistic regressor:
- Image encodings are (L2 normalized) input vectors X.
- Text encodings are (L2 normalized) weights W for each class.
- Softmax is used to convert sets of logits along both test and image axes.

Key Idea: CLIP Pseudocode Summary



```
# 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
```

Experiments: Zero-Shot Learning

- ‘Zero-shot’ in this paper means prediction on a dataset without prior exposure to it.
    - A test dataset evaluates performance on task-specific distribution
    - First explored in Visual N-Grams (Li et al., 2017)

- CLIP is first trained with natural language supervision, then tested on dataset (e.g. ImageNet)
    - Dataset classes are converted into text tags

- Largest model took 18 training days on 592 V100GPUs

CLIP objective is more training efficient

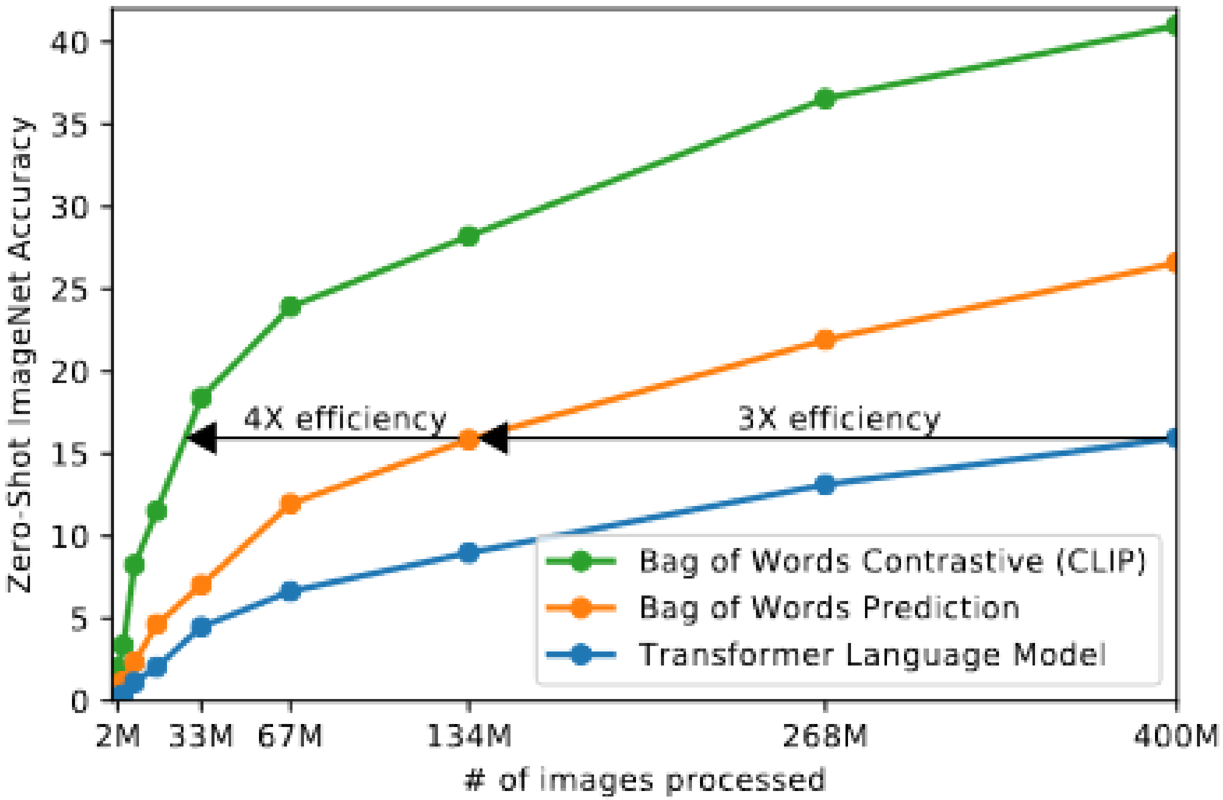
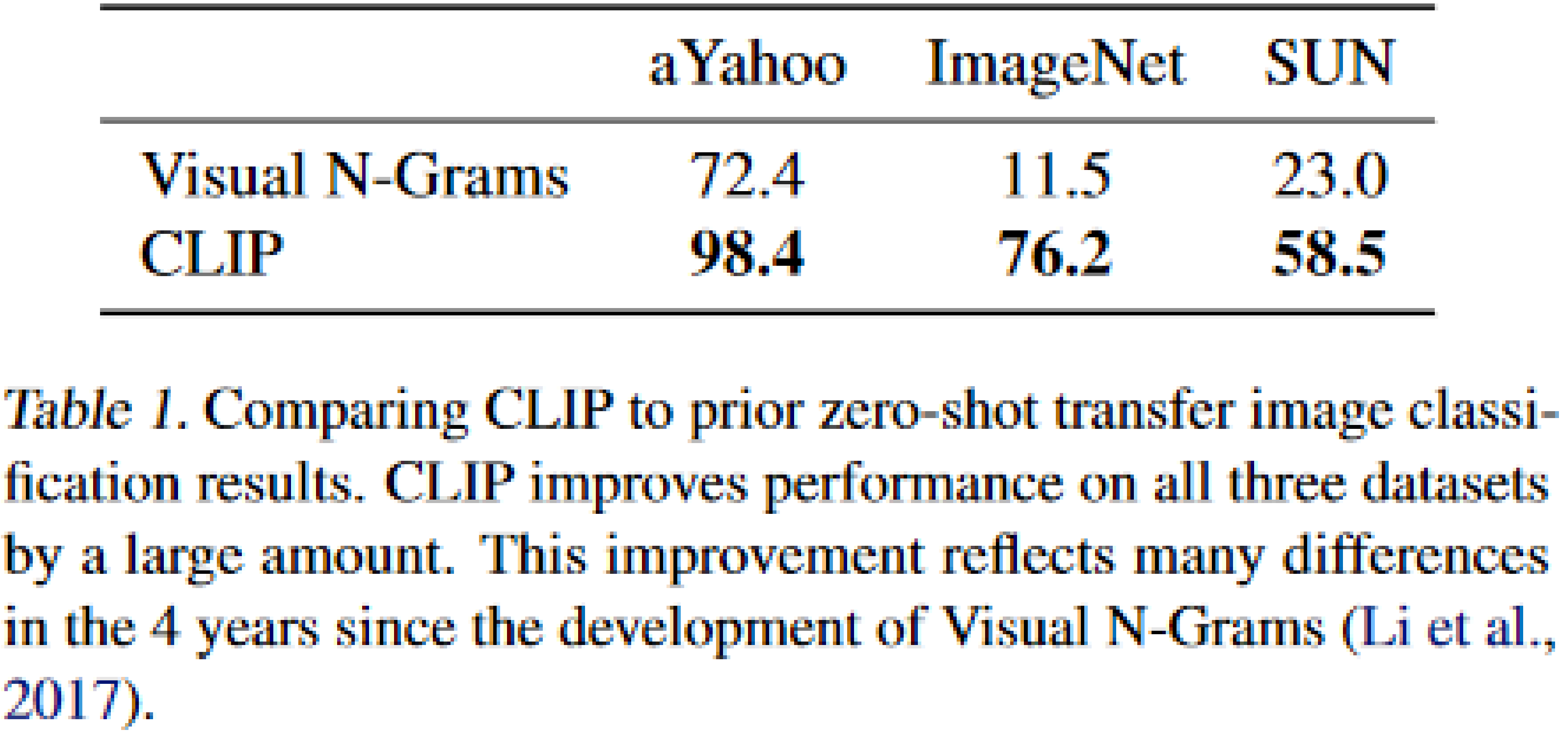


Figure 2. **CLIP is much more efficient at zero-shot transfer than our image caption baseline.** Although highly expressive, we found that transformer-based language models are relatively weak at zero-shot ImageNet classification. Here, we see that it learns 3x slower than a baseline which predicts a bag-of-words (BoW) encoding of the text (Joulin et al., 2016). Swapping the prediction objective for the contrastive objective of CLIP further improves efficiency another 4x.

CLIP greatly outperforms on 0-shot learning



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

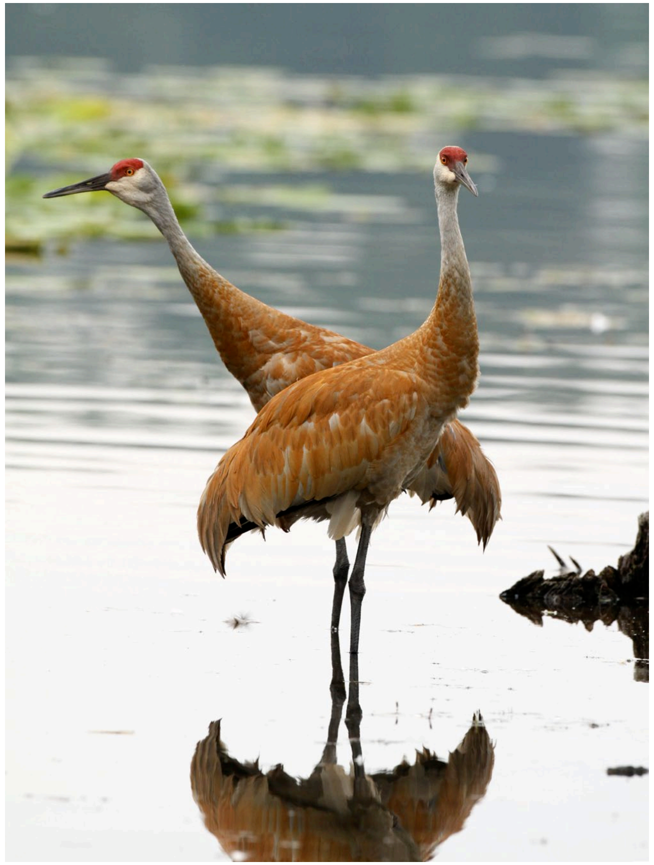
Table 1. Comparing CLIP to prior zero-shot transfer image classification results. CLIP improves performance on all three datasets by a large amount. This improvement reflects many differences in the 4 years since the development of Visual N-Grams (Li et al., 2017).

Prompt Engineering for Labels

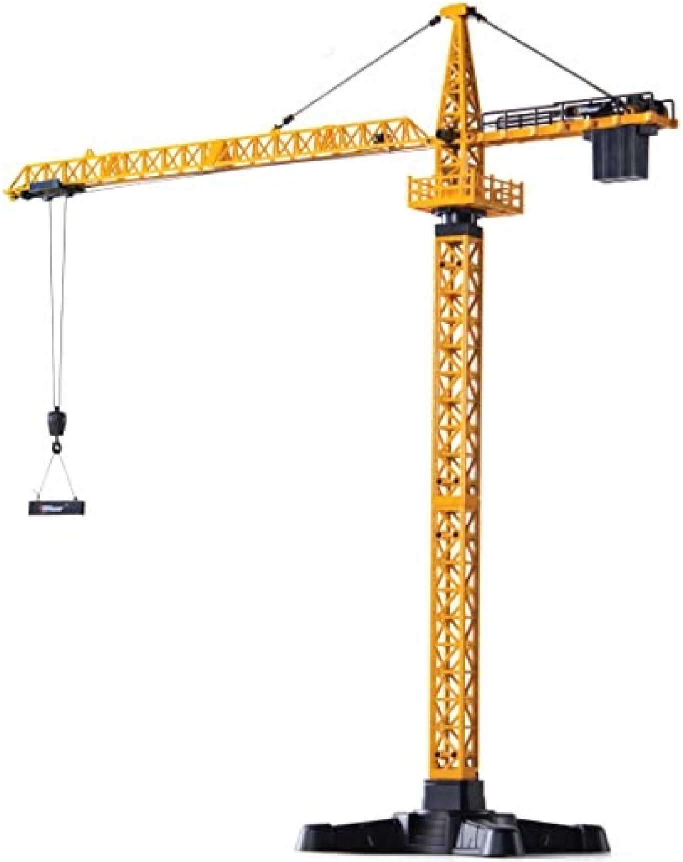
- CLIP chooses for a list of prompt tags, but dataset labels are usually one word (e.g.: ‘dog’)
- Single words are hard to predict as prompts:
  - Provide no context for the task
  - Polysemy issue with many labels (multiple meanings for one word)
  - Most tags encountered on the internet are more than one word
  - Many single words can be associated with one image

Polysemy Issue

Crane



Crane



Solution: Add context to label

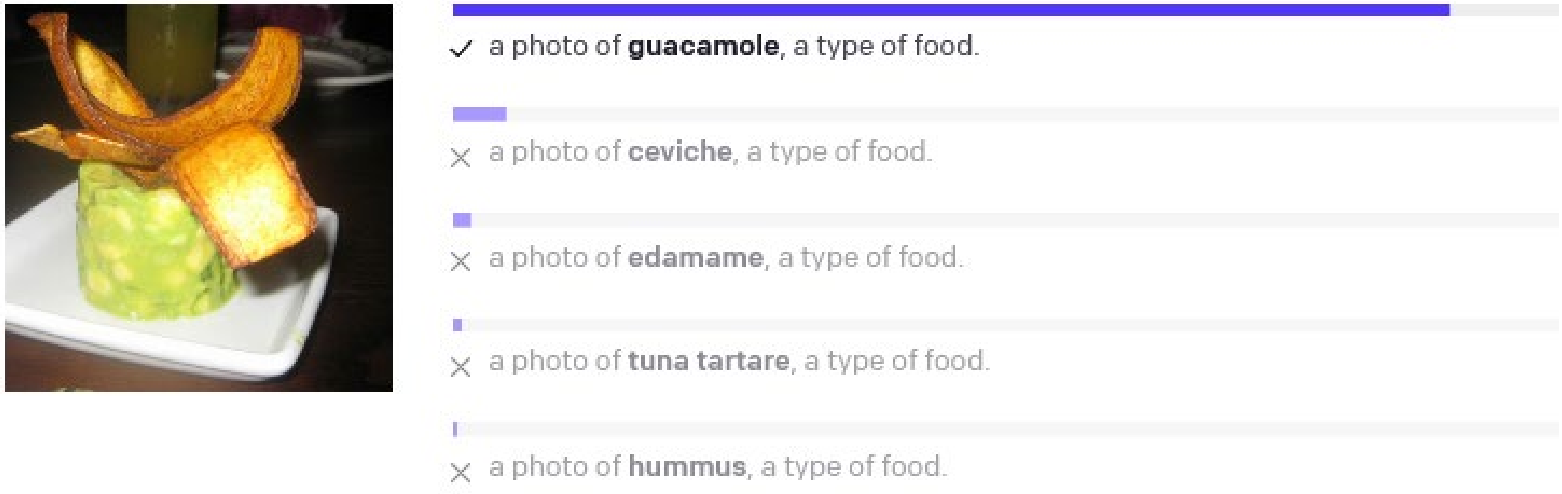
- Generic labels were transformed to: “A photo of {label}”
  - Resulted in better overall performance

- Changing the context also helped guide the task for more specific applications.
  - For Oxford Pets dataset: “A photo of a {label}, a type of pet”
    - Otherwise {Boxer} could be a type of dog or a type of sports athlete.
  - For satellite image classification “a satellite photo of {label}” helped.

- Missing word also implicitly specifies the task.

FOOD101

guacamole (90.1%) Ranked 1 out of 101 labels



✓ a photo of **guacamole**, a type of food.

× a photo of **ceviche**, a type of food.

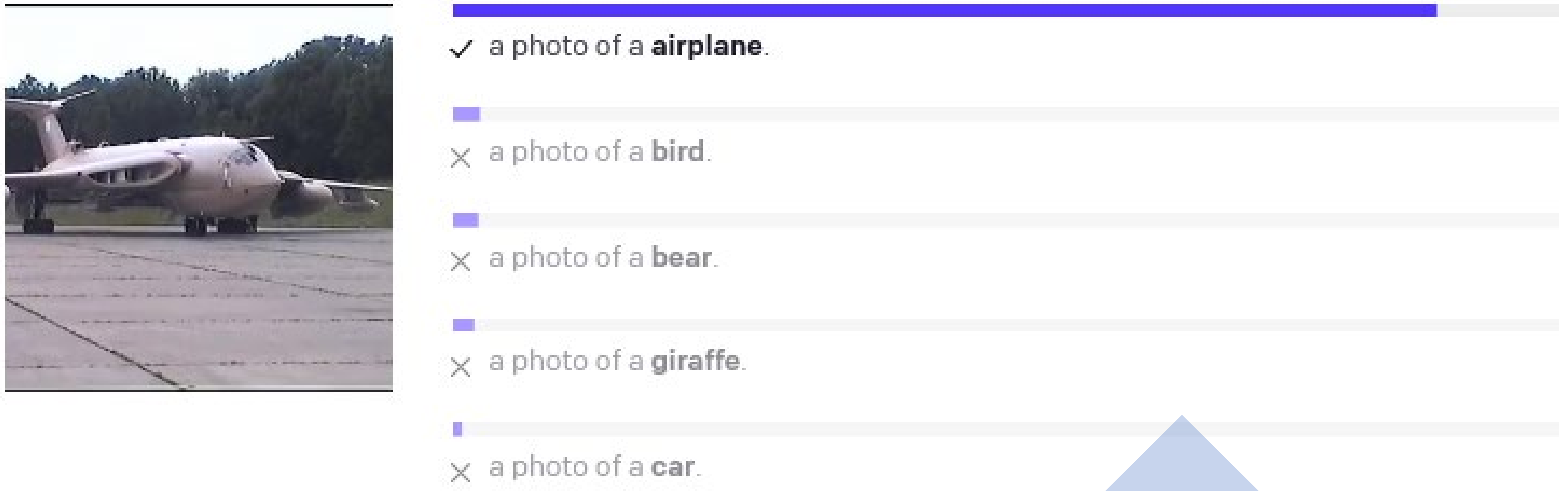
× a photo of **edamame**, a type of food.

× a photo of **tuna tartare**, a type of food.

× a photo of **hummus**, a type of food.

YOUTUBE-BB

airplane, person (89.0%) Ranked 1 out of 23



✓ a photo of a **airplane**.

× a photo of a **bird**.

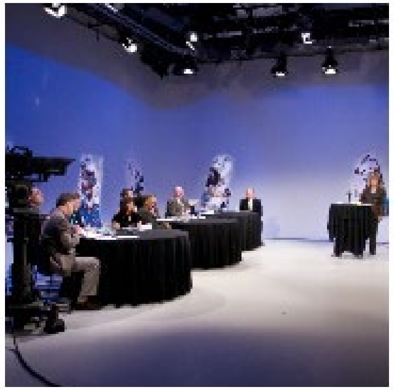
× a photo of a **bear**.

× a photo of a **giraffe**.

× a photo of a **car**.

SUN397

television studio (90.2%) Ranked 1 out of 397



✓ a photo of a **television studio**.

✗ a photo of a **podium indoor**.

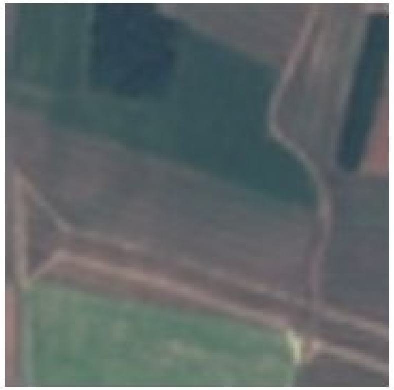
✗ a photo of a **conference room**.

✗ a photo of a **lecture room**.

✗ a photo of a **control room**.

EUROSAT

annual crop land (12.9%) Ranked 4 out of 10



✗ a centered satellite photo of **permanent crop land**.

✗ a centered satellite photo of **pasture land**.

✗ a centered satellite photo of **highway or road**.

✓ a centered satellite photo of **annual crop land**.

✗ a centered satellite photo of **brushland or shrubland**.

Zero-shot Clip vs Fully Supervised Baseline

- CLIP even sometimes outperforms supervised baseline
- Best performance is on more ‘generic’ tasks, which probably have more overlap with internet training set.

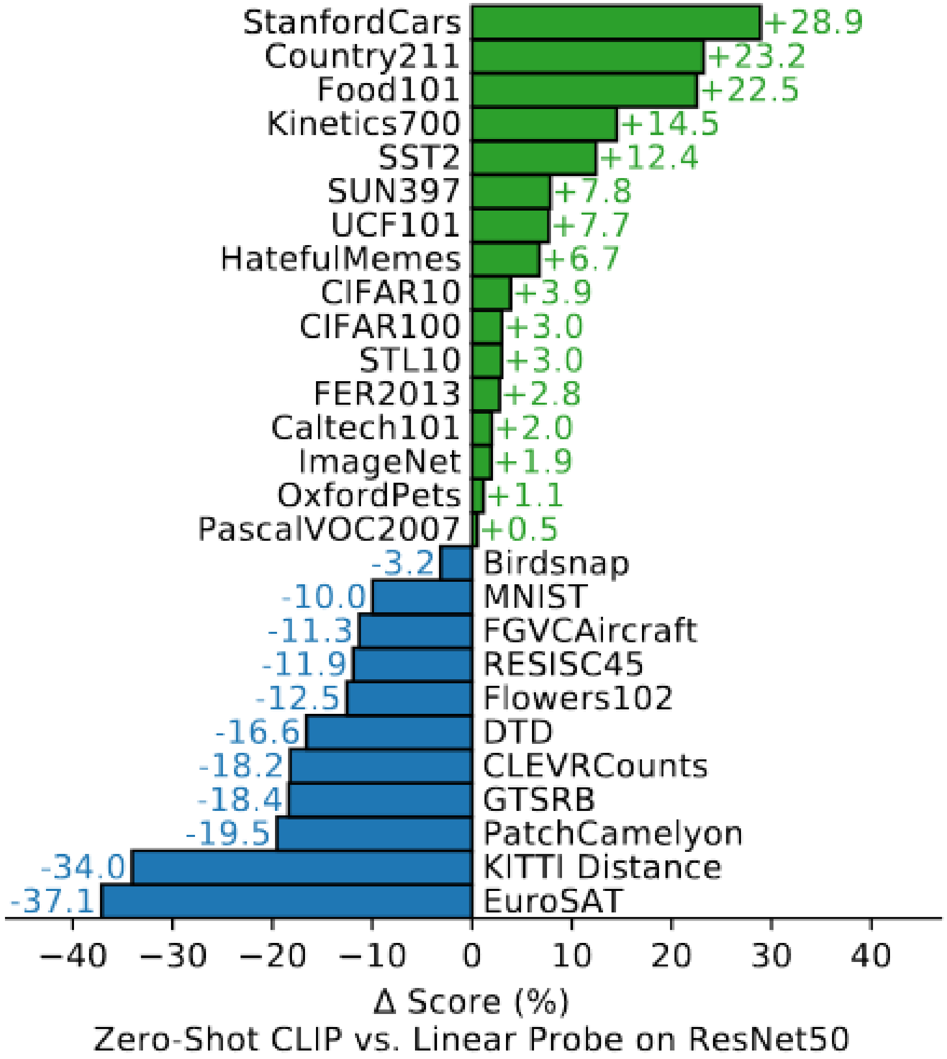
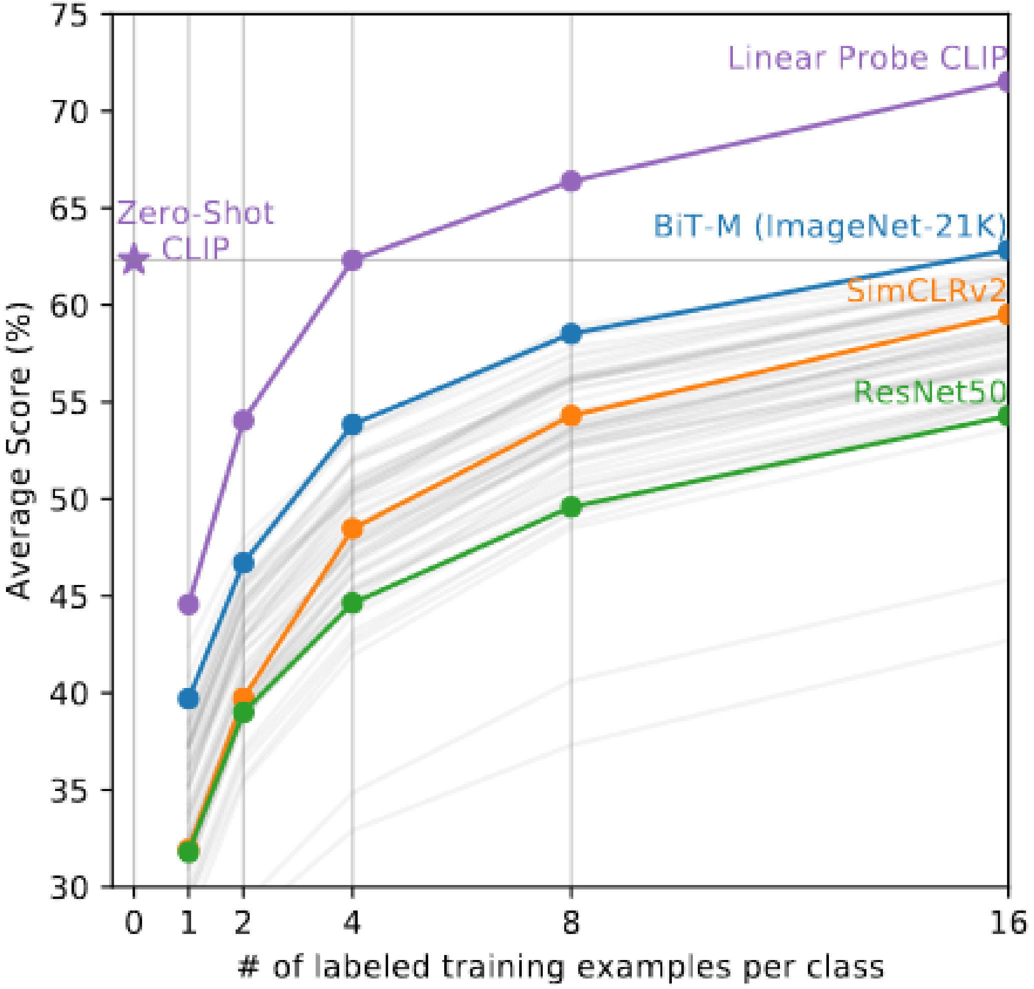


Figure 5. **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

0-shot vs Few-shot linear probes

- Compared CLIP to many few-shot learning networks.
- 0-shot CLIP matches or outperforms few-shot, supervised only networks.



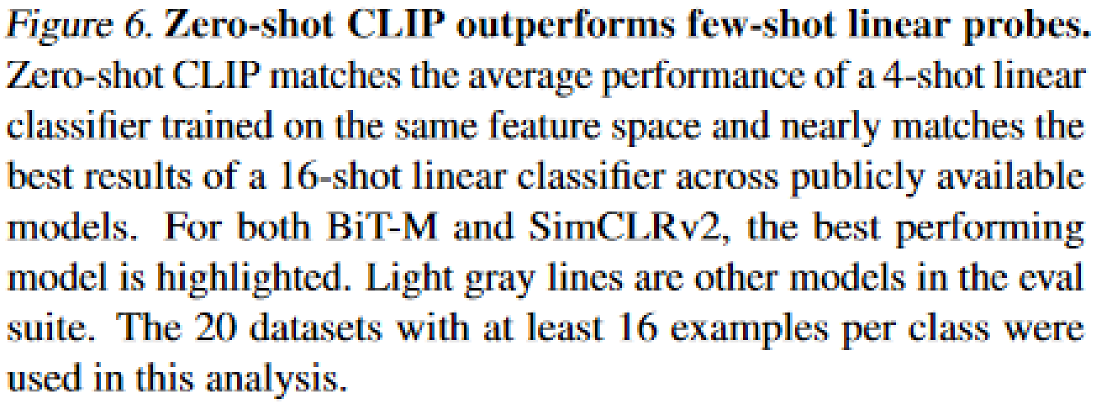


Figure 6. **Zero-shot CLIP outperforms few-shot linear probes.** Zero-shot CLIP matches the average performance of a 4-shot linear classifier trained on the same feature space and nearly matches the best results of a 16-shot linear classifier across publicly available models. For both BiT-M and SimCLRv2, the best performing model is highlighted. Light gray lines are other models in the eval suite. The 20 datasets with at least 16 examples per class were used in this analysis.

Representation Learning

- To evaluate learned visual representations, a linear probe is attached to the end of image-trained encoding models and trained on specific dataset
  - CLIP models compared to various supervised models.

- CLIP models generally learn better representations than fully supervised counterparts.
- Transformer-based CLIP image encoder is more efficient than CNN-based encoder

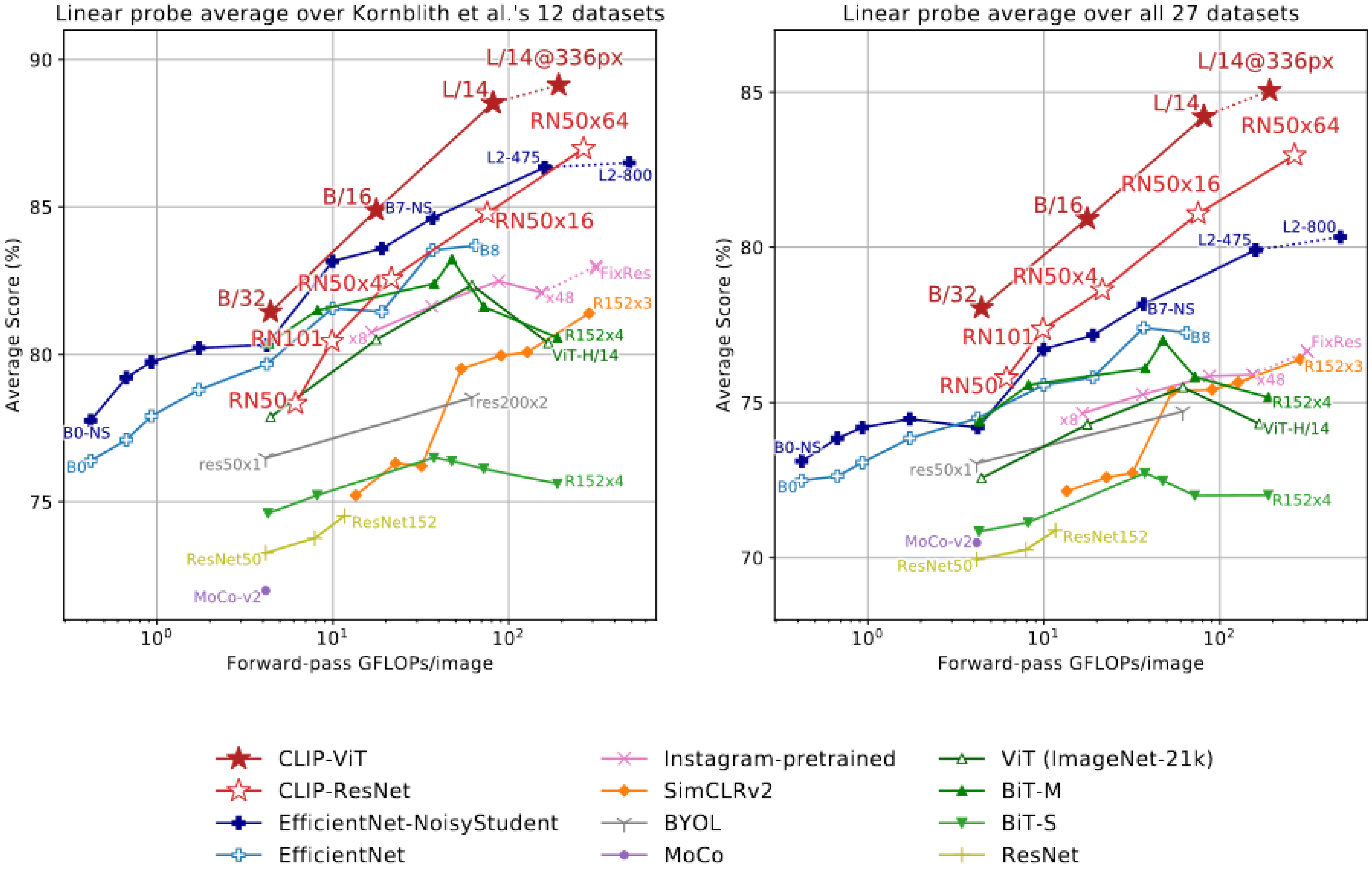


Figure 10. **Linear probe performance of CLIP models in comparison with state-of-the-art computer vision models**, including EfficientNet (Tan & Le, 2019; Xie et al., 2020), MoCo (Chen et al., 2020d), Instagram-pretrained ResNeXt models (Mahajan et al., 2018; Touvron et al., 2019), BiT (Kolesnikov et al., 2019), ViT (Dosovitskiy et al., 2020), SimCLRv2 (Chen et al., 2020c), BYOL (Grill et al., 2020), and the original ResNet models (He et al., 2016b). (Left) Scores are averaged over 12 datasets studied by Kornblith et al. (2019). (Right) Scores are averaged over 27 datasets that contain a wider variety of distributions. Dotted lines indicate models fine-tuned or evaluated on images at a higher-resolution than pre-training. See Table 10 for individual scores and Figure 20 for plots for each dataset.

Per Dataset Comparison

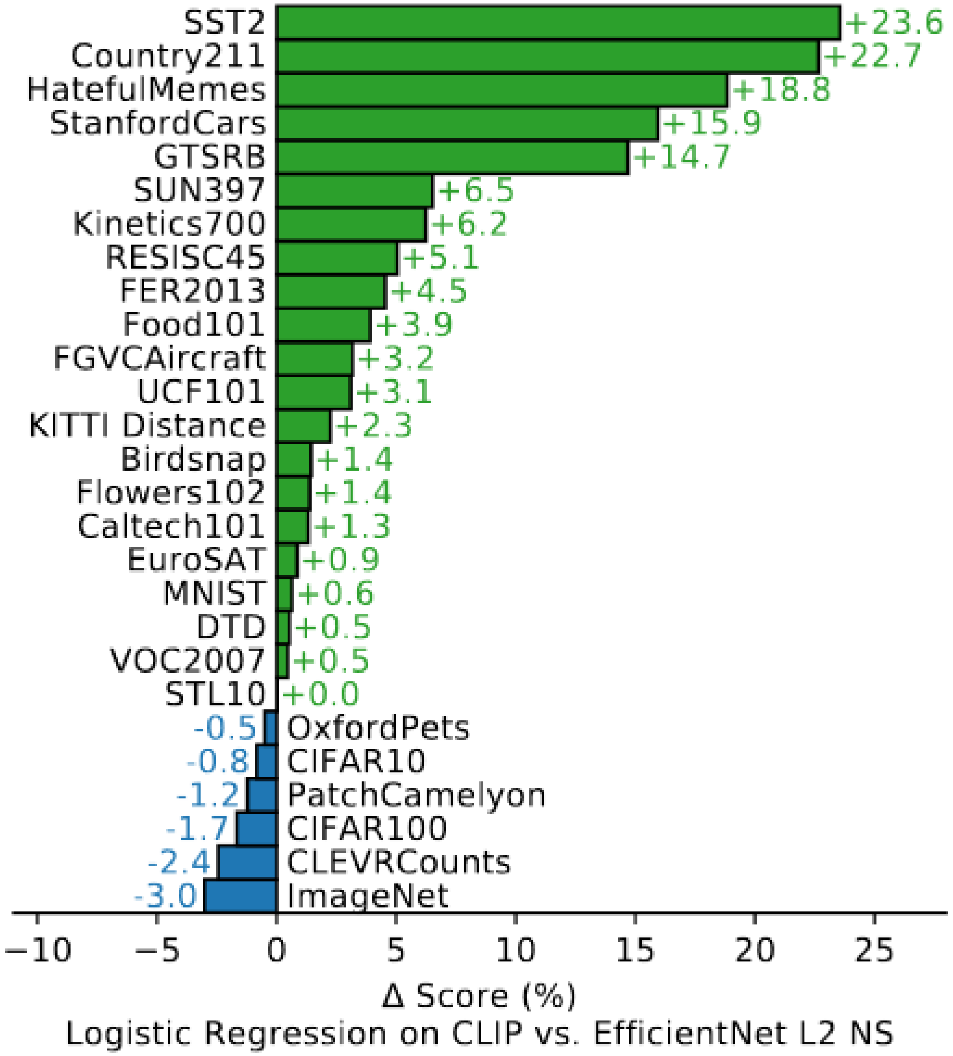


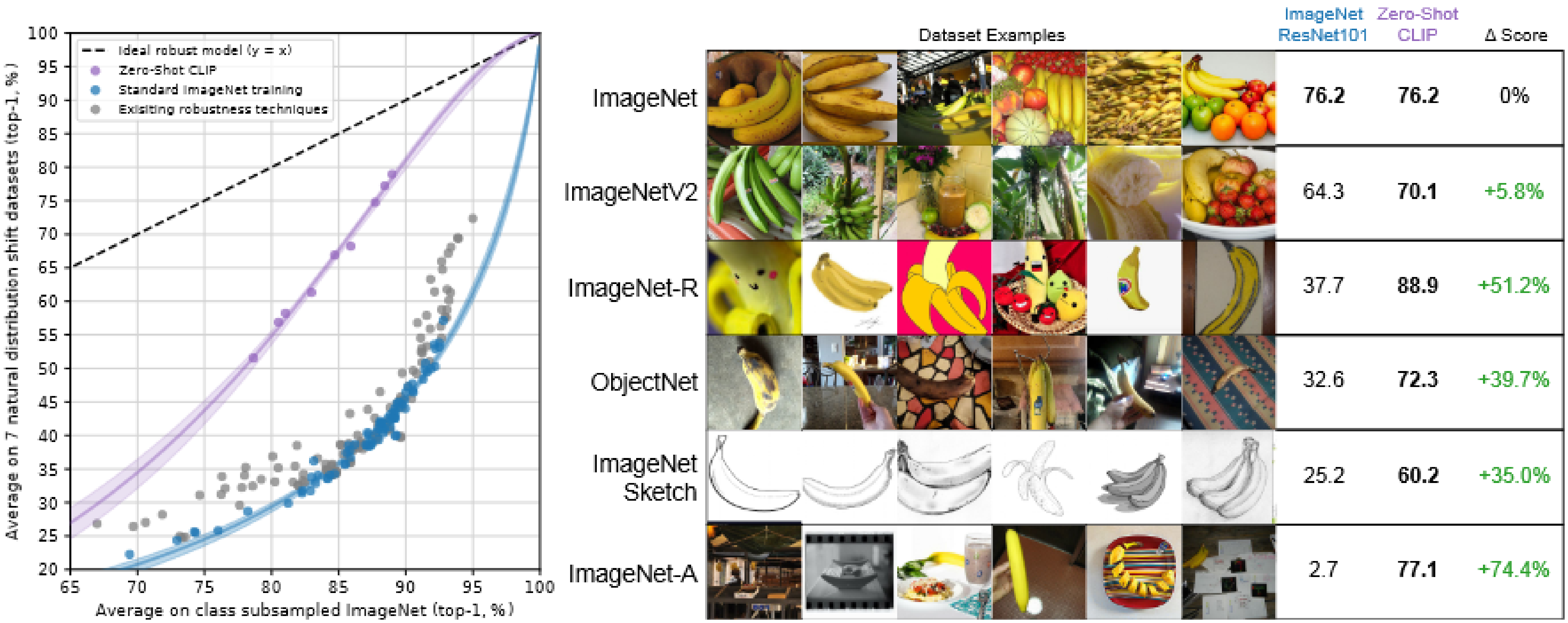
Figure 11. **CLIP's features outperform the features of the best ImageNet model on a wide variety of datasets.** Fitting a linear classifier on CLIP's features outperforms using the Noisy Student EfficientNet-L2 on 21 out of 27 datasets.

Robustness to Natural Distribution Shift

- Supervised models trained on dataset often only perform well on the specific benchmark.
  - Due to optimizing to only one distribution specific to a dataset.

- **Representations leaned with CLIP transfer better to shifted distributions compared to supervision on specific dataset.**

- CLIP’s 0-shot ability shows robust adaptability to various datasets compared to models trained on specific datasets.



	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%

Figure 13. **Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models.** (Left) An ideal robust model (dashed line) performs equally well on the ImageNet distribution and on other natural image distributions. Zero-shot CLIP models shrink this "robustness gap" by up to 75%. Linear fits on logit transformed values are shown with bootstrap estimated 95% confidence intervals. (Right) Visualizing distribution shift for bananas, a class shared across 5 of the 7 natural distribution shift datasets. The performance of the best zero-shot CLIP model, ViT-L/14@336px, is compared with a model that has the same performance on the ImageNet validation set, ResNet-101.

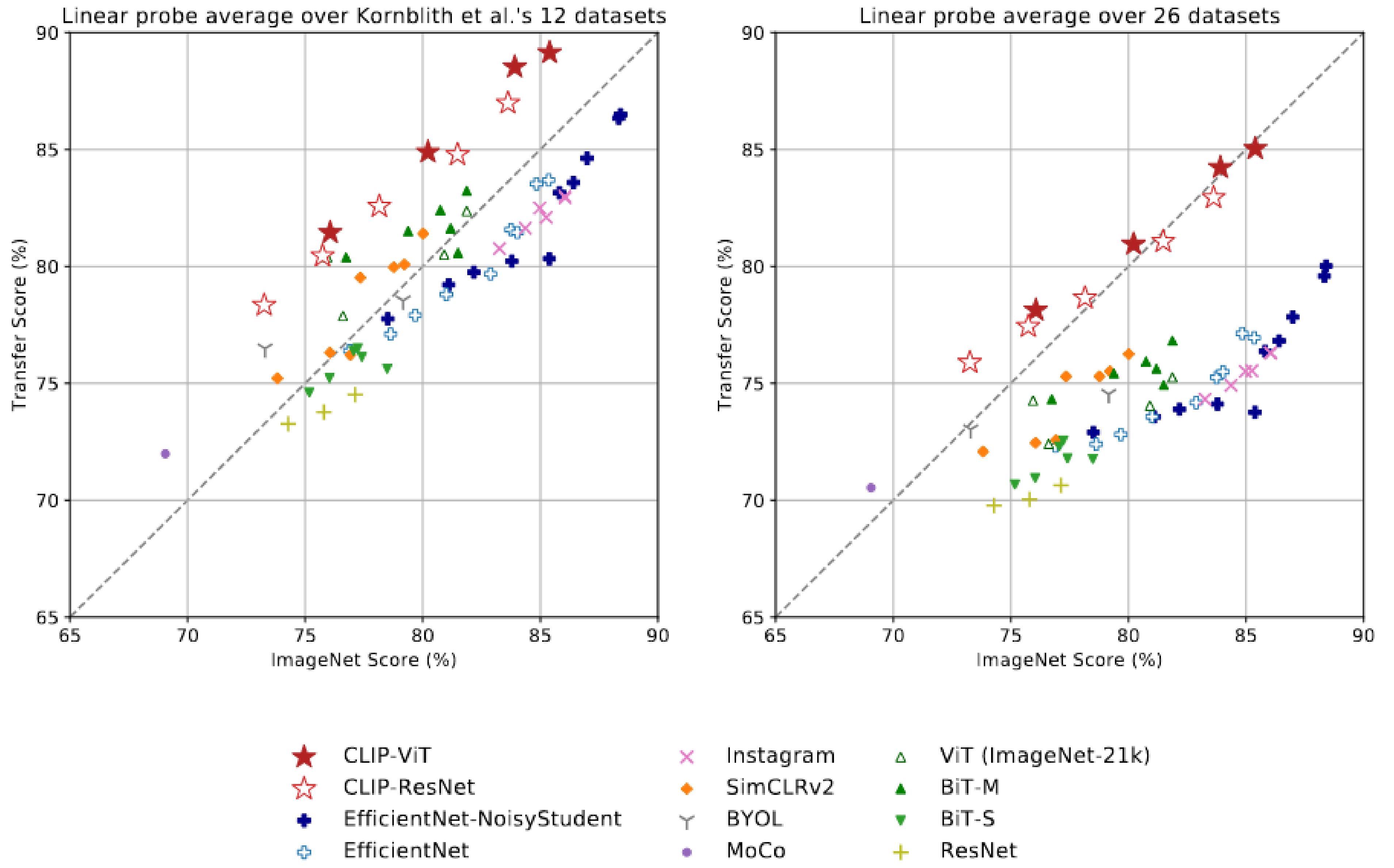


Figure 12. **CLIP's features are more robust to task shift when compared to models pre-trained on ImageNet.** For both dataset splits, the transfer scores of linear probes trained on the representations of CLIP models are higher than other models with similar ImageNet performance. This suggests that the representations of models trained on ImageNet are somewhat overfit to their task.

Robust vs Specific Representations

- Authors found that increased supervised training of CLIP on specific datasets increased specific performance but reduced performance on outside datasets.

- *"Across our experiments, high effective robustness seems to result from minimizing the amount of distribution specific training data a model has access to, but this comes at a cost of reducing dataset-specific performance."*

Overlap with Human Performance

- Human vs CLIP performance on various image tasks correlates strongly.

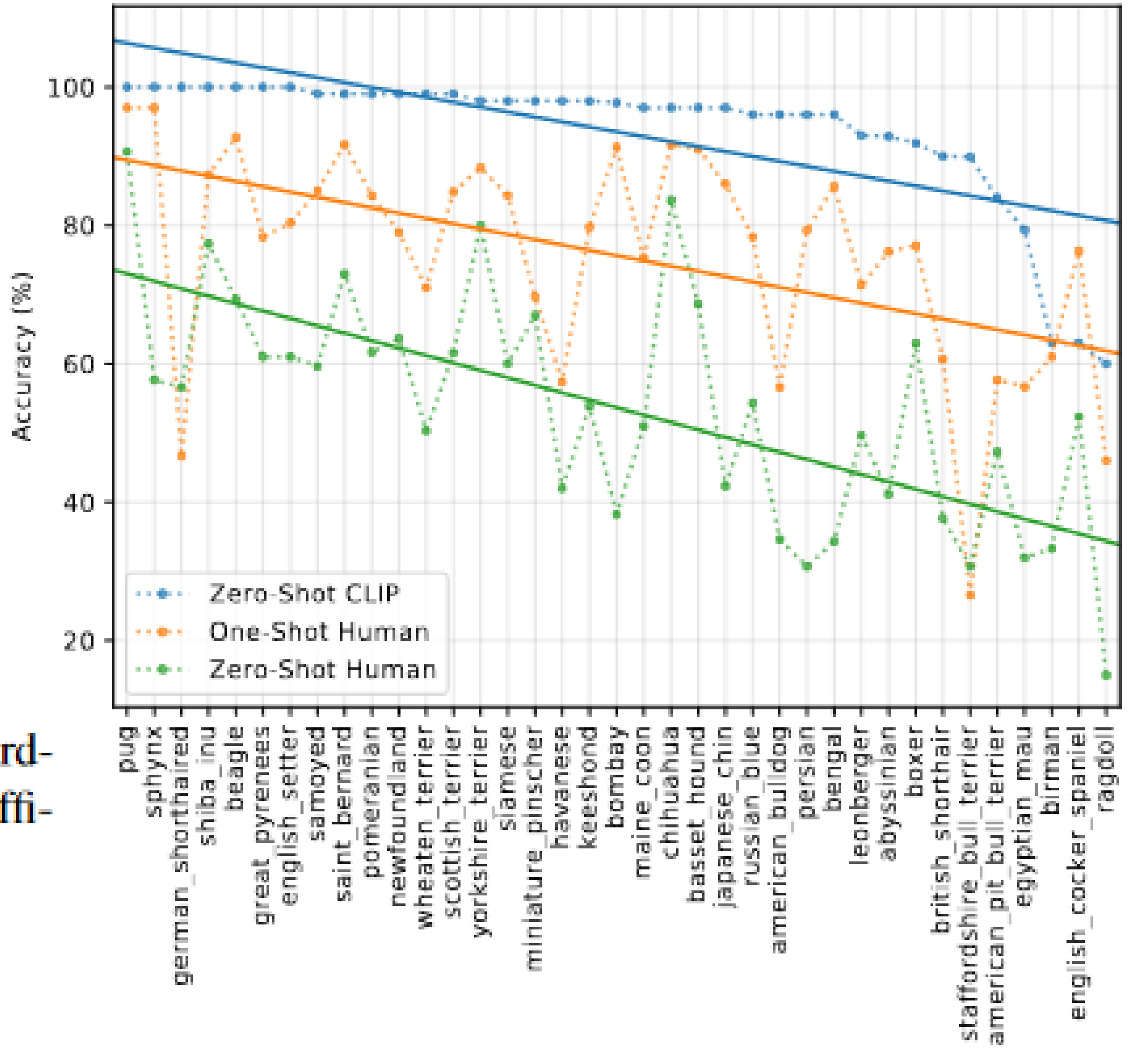


Figure 16. The hardest problems for CLIP also tend to be the hardest problems for humans. Here we rank image categories by difficulty for CLIP as measured as probability of the correct label.

Limitations

- CLIP still must select from list of tags.
    - Cannot generate novel outputs
- 0-shot CLIP still underperforms on more specific tasks or datasets when compared to supervised counterparts trained on that benchmark.
    - Fine-grained classification (e.g.: species of flower), abstract tasks (counting # of objects)
    - Authors estimate 1000x increase in computation to match SOTA.
    - These tasks essentially don’t exist in CLIP’s training set.

Limitations cont.

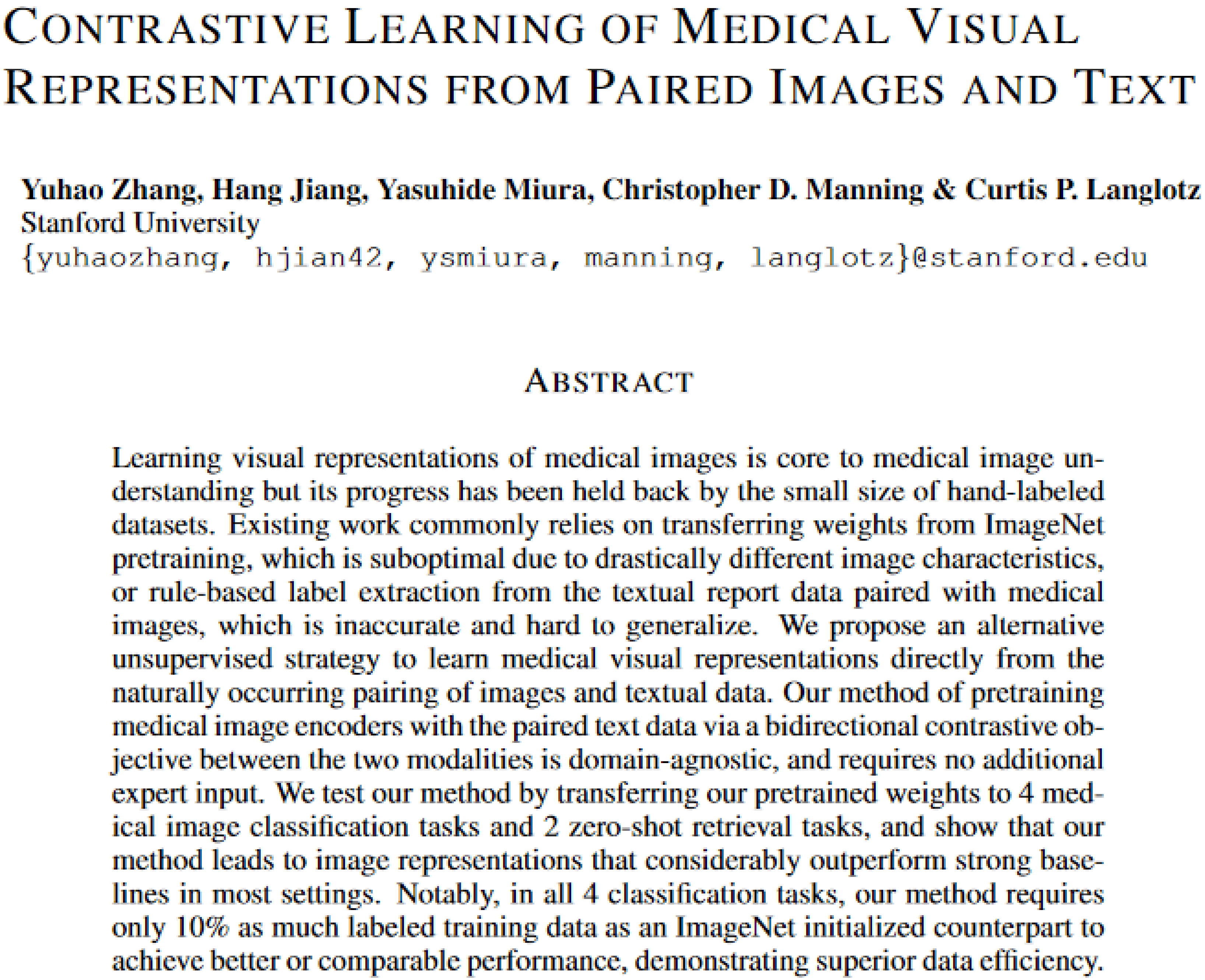
- Still generalized poorly to data that is truly ‘out-of-distribution’.
- Does not address poor data efficiency in deep learning.
- Unfiltered internet training set leads to learned social biases.
- Some complex tasks are difficult to specify via text.
  - Can account for worse performance in few-shot vs 0-shot setting.

Conclusion

- CLIP proposes a powerful framework for natural language supervision for image-based tasks.

- Results show a robust learned representation that performs well in 0-shot setting in many tasks.

- Actual paper is ~50 pages
  - Many more details on results and broader impacts.



Contrastive Learning of Medical Visual Representations from Paired Images and Text

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning & Curtis P. Langlotz
Stanford University
`{yuhaozhang, hjian42, ysmiura, manning, langlotz}@stanford.edu`

Abstract

Learning visual representations of medical images is core to medical image understanding but its progress has been held back by the small size of hand-labeled datasets. Existing work commonly relies on transferring weights from ImageNet pretraining, which is suboptimal due to drastically different image characteristics, or rule-based label extraction from the textual report data paired with medical images, which is inaccurate and hard to generalize. We propose an alternative unsupervised strategy to learn medical visual representations directly from the naturally occurring pairing of images and textual data. Our method of pretraining medical image encoders with the paired text data via a bidirectional contrastive objective between the two modalities is domain-agnostic, and requires no additional expert input. We test our method by transferring our pretrained weights to 4 medical image classification tasks and 2 zero-shot retrieval tasks, and show that our method leads to image representations that considerably outperform strong baselines in most settings. Notably, in all 4 classification tasks, our method requires only 10% as much labeled training data as an ImageNet initialized counterpart to achieve better or comparable performance, demonstrating superior data efficiency.