Learning Transferable Visual Models From Natural Language Supervision

Liwei Che

03/27/2023

Background

Although deep learning has revolutionized computer vision, current approaches have several major problems:

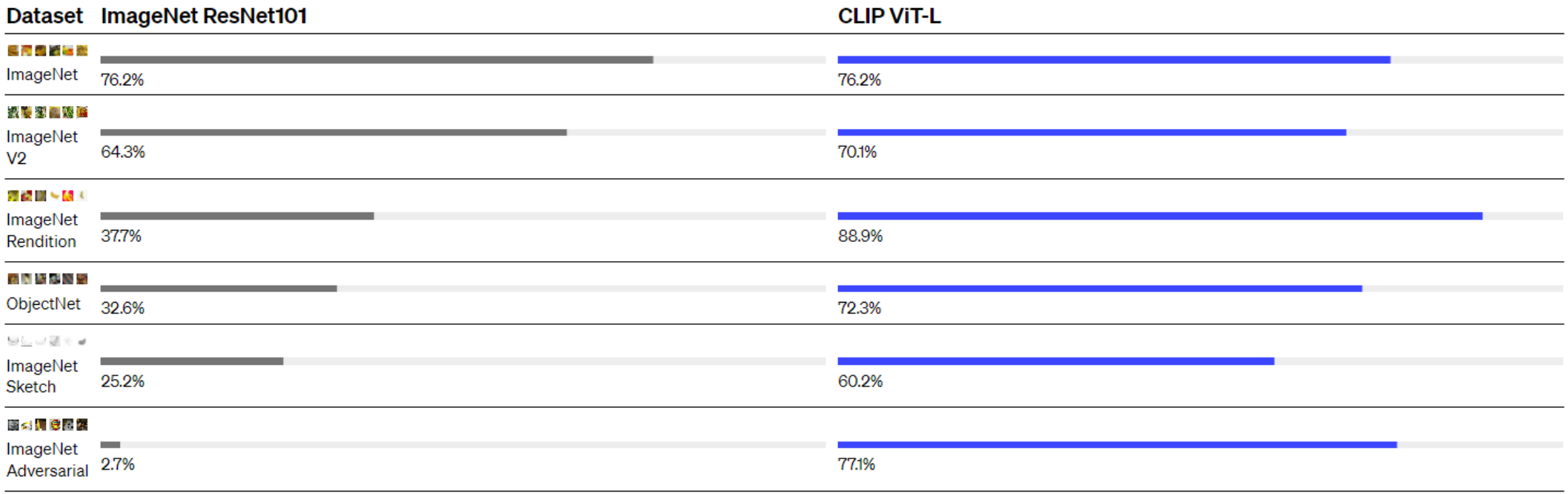
- Typical vision datasets are labor intensive and costly to create while teaching only a narrow set of visual concepts (Expensive);
- Standard vision models are good at one task and one task only, and require significant effort to adapt to a new task (Hard to transfer);
- Models that perform well on benchmarks have disappointingly poor performance on stress tests (Bad at generalization).

Contribution

CLIP (*Contrastive Language–Image Pre-training*) mitigates the gap between source domain and other domains, benchmarking performance and wild test with a cheap and computational efficient way. --- Amazing zero-shot performance!

1. Collected 400 million image-text pairs from Internet as training dataset
2. Get the logits via the cosine similarities between the image-text pairs (metric learning)

Zero-shot Performance of CLIP



Although both models have the same accuracy on the ImageNet test set, CLIP's performance is much more representative of how it will fare on datasets that measure accuracy in different, non-ImageNet settings. For instance, ObjectNet checks a model's ability to recognize objects in many different poses and with many different backgrounds inside homes while ImageNet Rendition and ImageNet Sketch check a model's ability to recognize more abstract depictions of objects.

CLIP - Overview

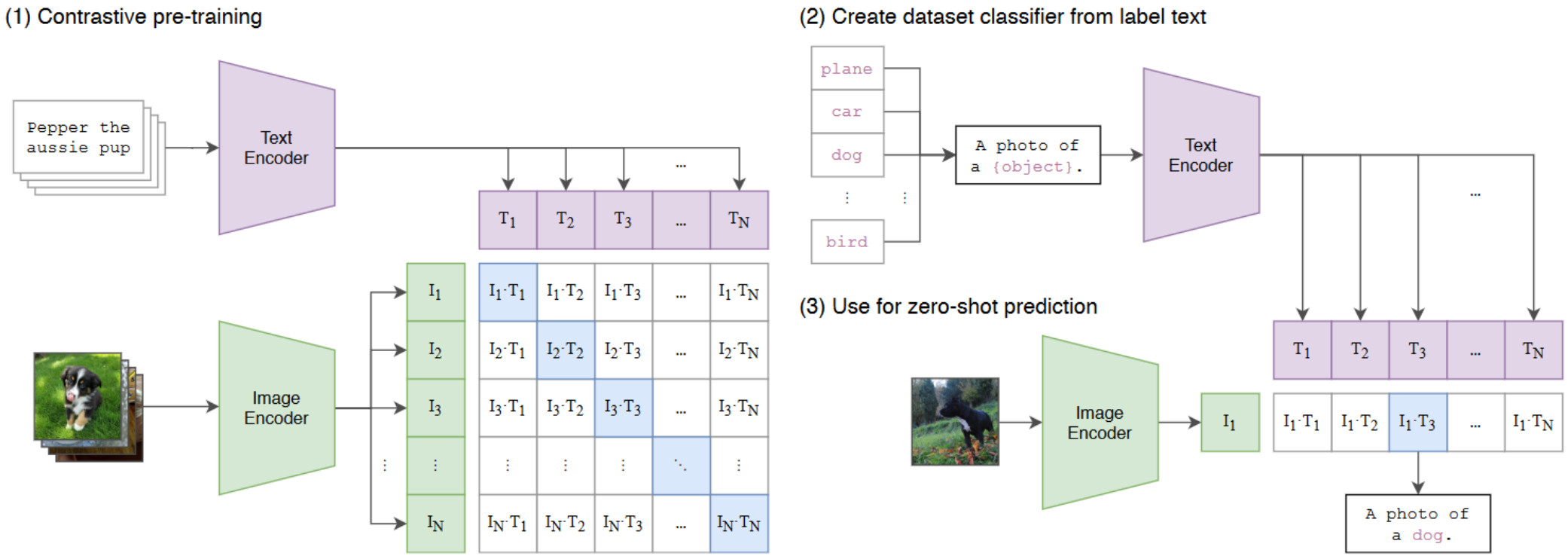


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the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

CLIP – Approach: Natural Language Supervision

Learning perception from supervision contained in natural language, a.k.a, which text is paired with the image input.

Strengths:

1. Needless of crowd-sourced labeling for image classification.
2. Not only learn a representation but also connects that representation to language which enables flexible zero-shot transfer.

CLIP – Approach: Creating a Sufficiently Large Dataset

Instead of using the existing well-labeled datasets, the authors collected over 400 million data pair that are available publicly on the internet. Make it a dataset named **WebImageText (WIT).**

To attempt to cover as broad a set of visual concepts as possible, we search for (image, text) pairs as part of the construction process whose text includes one of a set of 500,000 queries. Class balance the results by including up to 20,000 (image, text) pairs per query.

CLIP – Approach: Selecting an Efficient Pre-Training Method

Training Task: Given N image-text pairs, CLIP is trained to predict which of the N × N possible (image, text) pairings across a batch actually occurred.

Method: Learn a multi-modal embedding space by jointly training an image encoder and text encoder to

maximize the cosine similarity of the image and text embeddings of the N real pairs in the batch

while minimizing the cosine similarity of the embeddings of the $N^2 - N$ incorrect pairings.

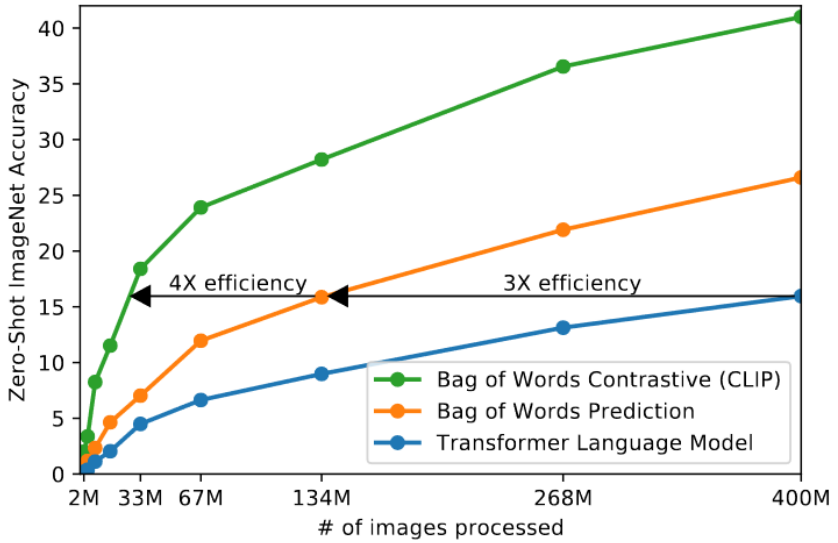


Figure 2. **CLIP is much more efficient at zero-shot transfer than our image caption baseline.** Although highly expressive, we found that transformer-based language models are relatively weak at zero-shot ImageNet classification. Here, we see that it learns 3x slower than a baseline which predicts a 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 – Approach: Selecting an Efficient Pre-Training Method

- Pseudo code

```
# 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
```

CLIP – Approach: Choosing and Scaling a Model

Architectures for the image encoder:

1. ResNet-50
2. Vision Transformer

Architectures for the text encoder:

Basic Transformer

CLIP's performance to be less sensitive to the capacity of the text encoder.

CLIP – Approach: Training

Image encoder:

ResNet-50x4, ResNet-50x16, ResNet-50x64, ResNet-101, ViT-B/32, ViT-L/14.

Epochs: 32

Optimizer: AdamW

Initialization: Grid search

Mini-batch: 32768 (the larger the better for contrastive learning)

Training time:

ResNet-50x64: 592 V100 GPUs for 18 days

ViT-L/14: 256 V100 GPUs for 12 days

CLIP – Experiment: Zero-shot Transfer

Prompt Engineering:

1. Mismatch between the name and the image.
2. Polysemy
3. the prompt template “A photo of a {label}.” improves the baseline of only using the label text.

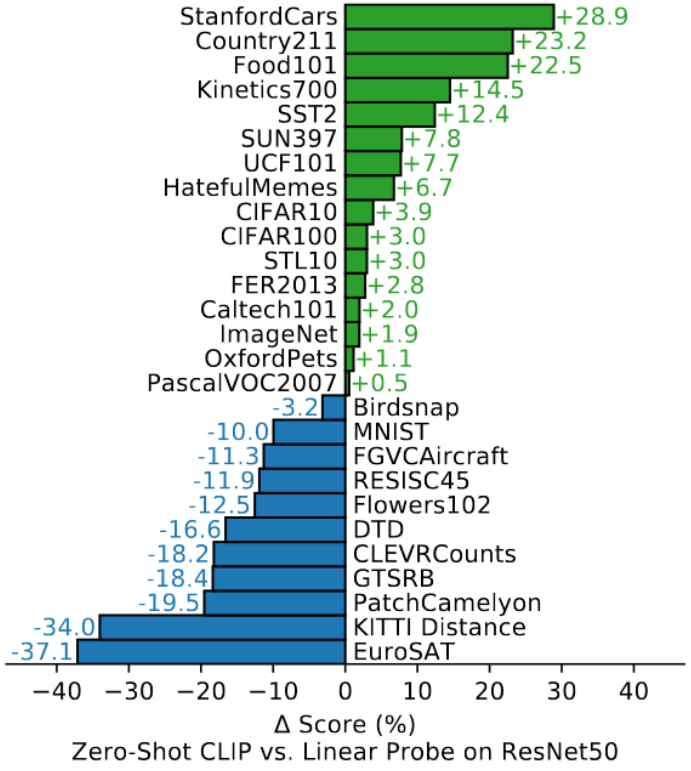


Figure 5. **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

CLIP – Experiment: Zero-shot Transfer

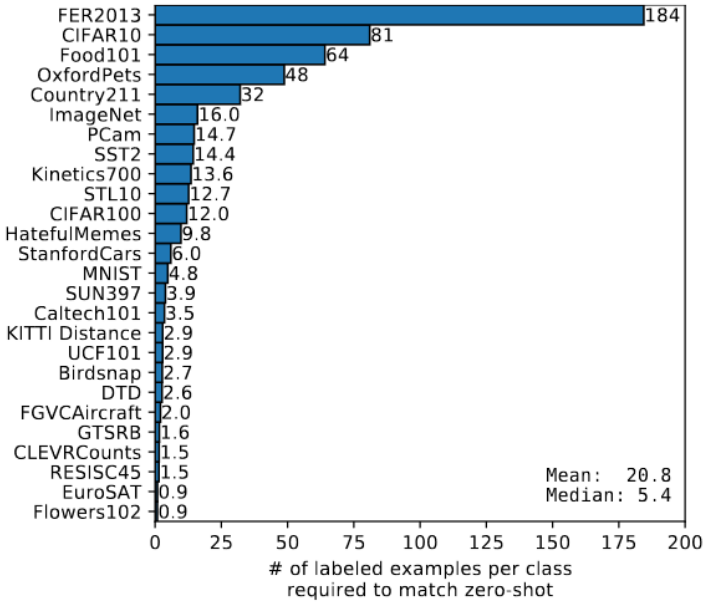


Figure 7. **The data efficiency of zero-shot transfer varies widely.** Calculating the number of labeled examples per class a linear classifier on the same CLIP feature space requires to match the performance of the zero-shot classifier contextualizes the effectiveness of zero-shot transfer. Values are estimated based on log-linear interpolation of 1, 2, 4, 8, 16-shot and fully supervised results. Performance varies widely from still underperforming a one-shot classifier on two datasets to matching an estimated 184 labeled examples per class.

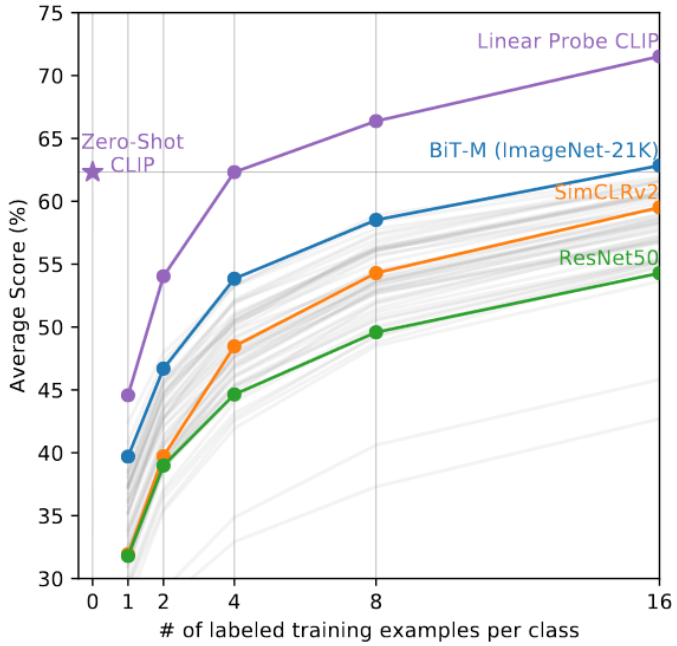


Figure 6. **Zero-shot CLIP outperforms few-shot linear probes.** Zero-shot CLIP matches the average performance of a 4-shot linear classifier trained on the same feature space and nearly matches the best results of a 16-shot linear classifier across publicly available models. For both BiT-M and SimCLRv2, 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.

CLIP – Experiment: Zero-shot Transfer

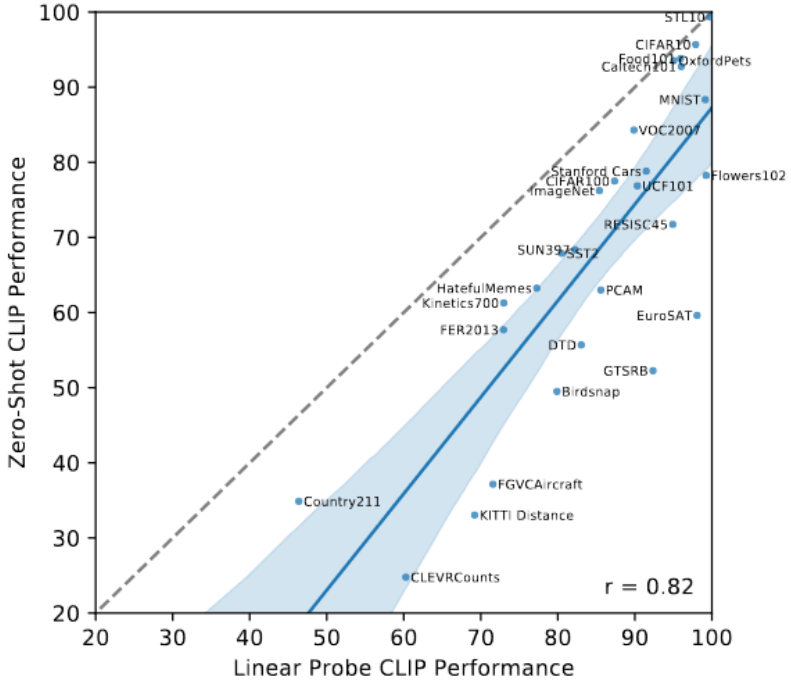


Figure 8. **Zero-shot performance is correlated with linear probe performance but still mostly sub-optimal.** Comparing zero-shot and linear probe performance across datasets shows a strong correlation with zero-shot performance mostly shifted 10 to 25 points lower. On only 5 datasets does zero-shot performance approach linear probe performance ($\leq$3 point difference).

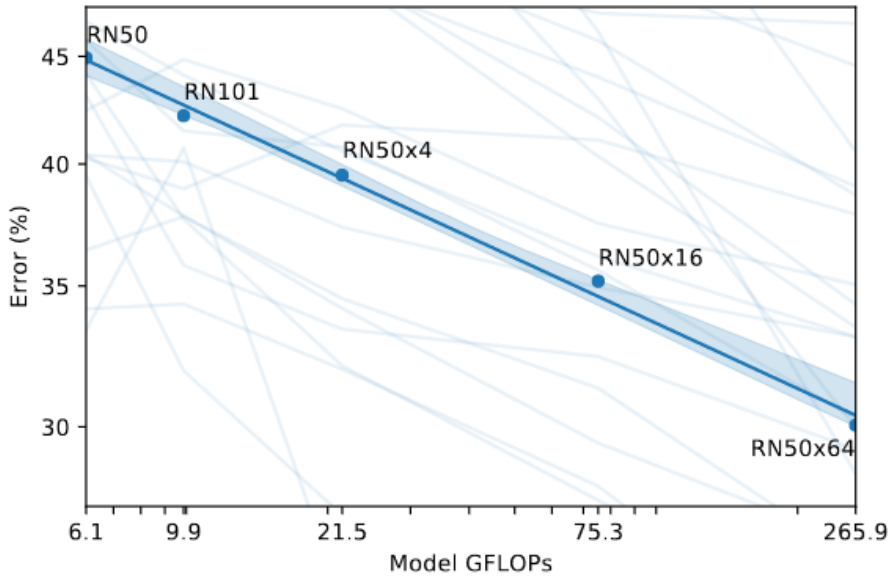
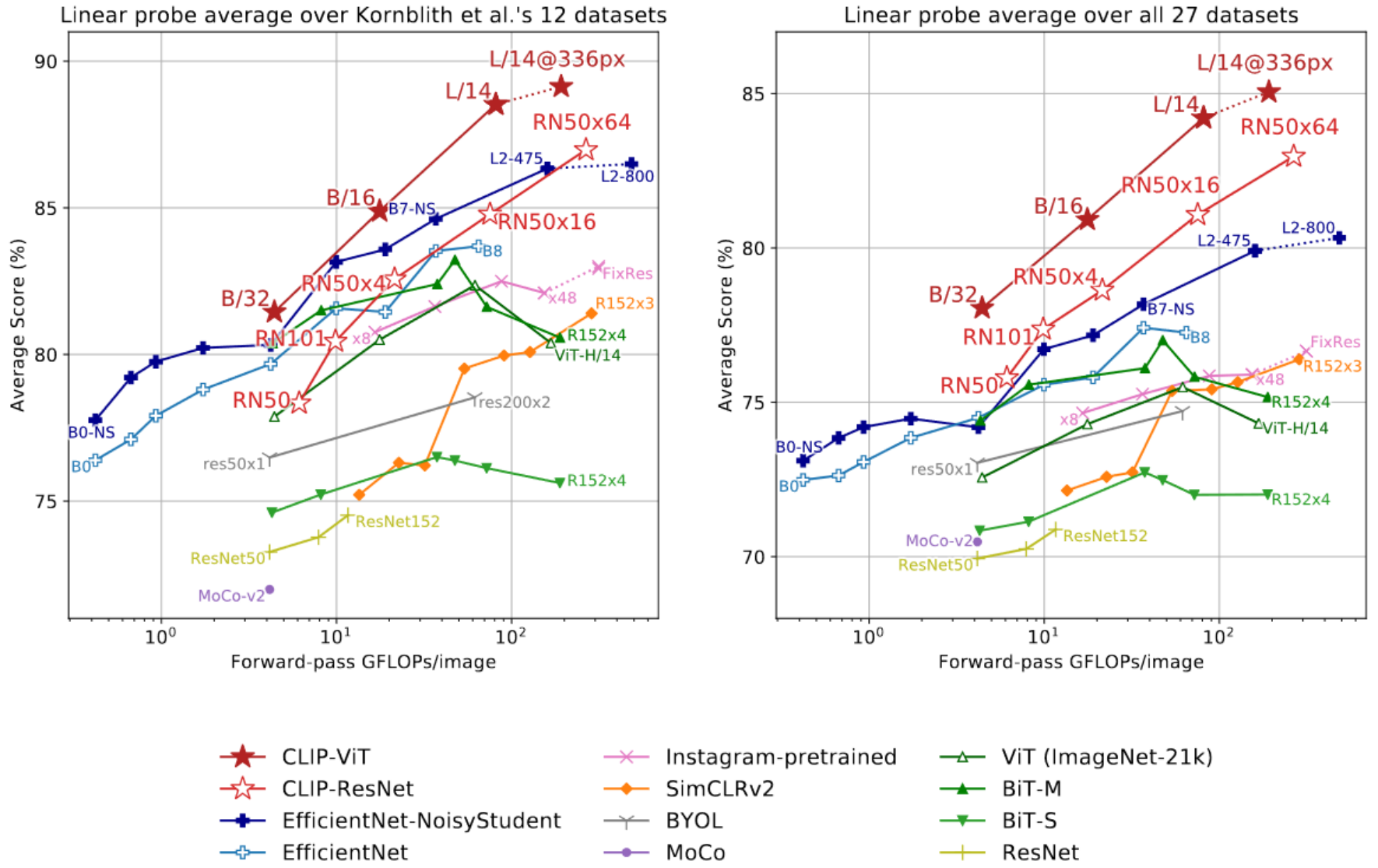
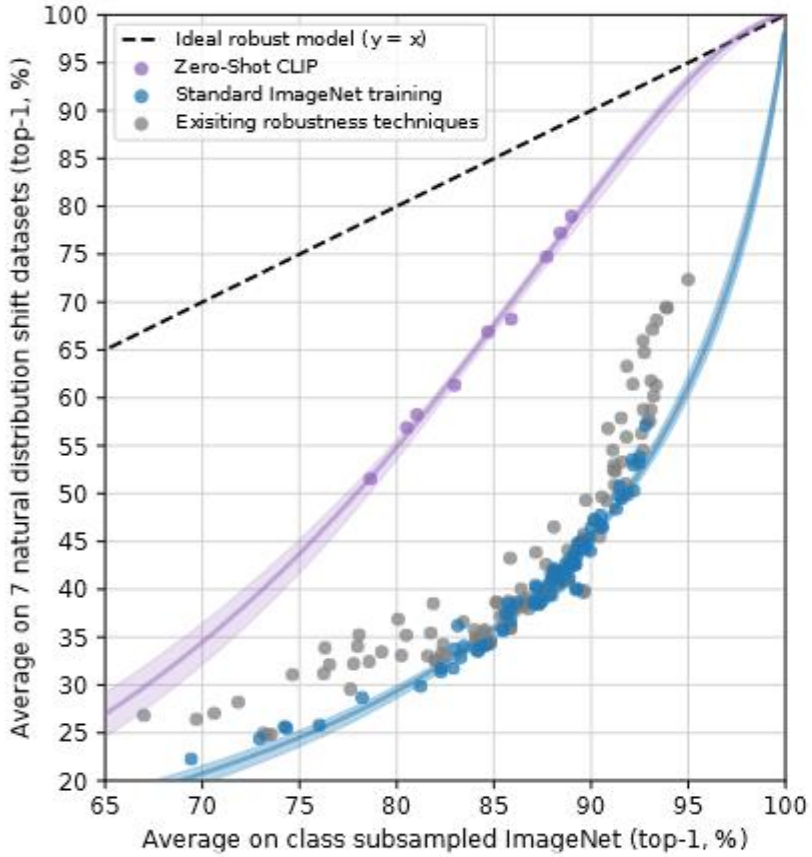


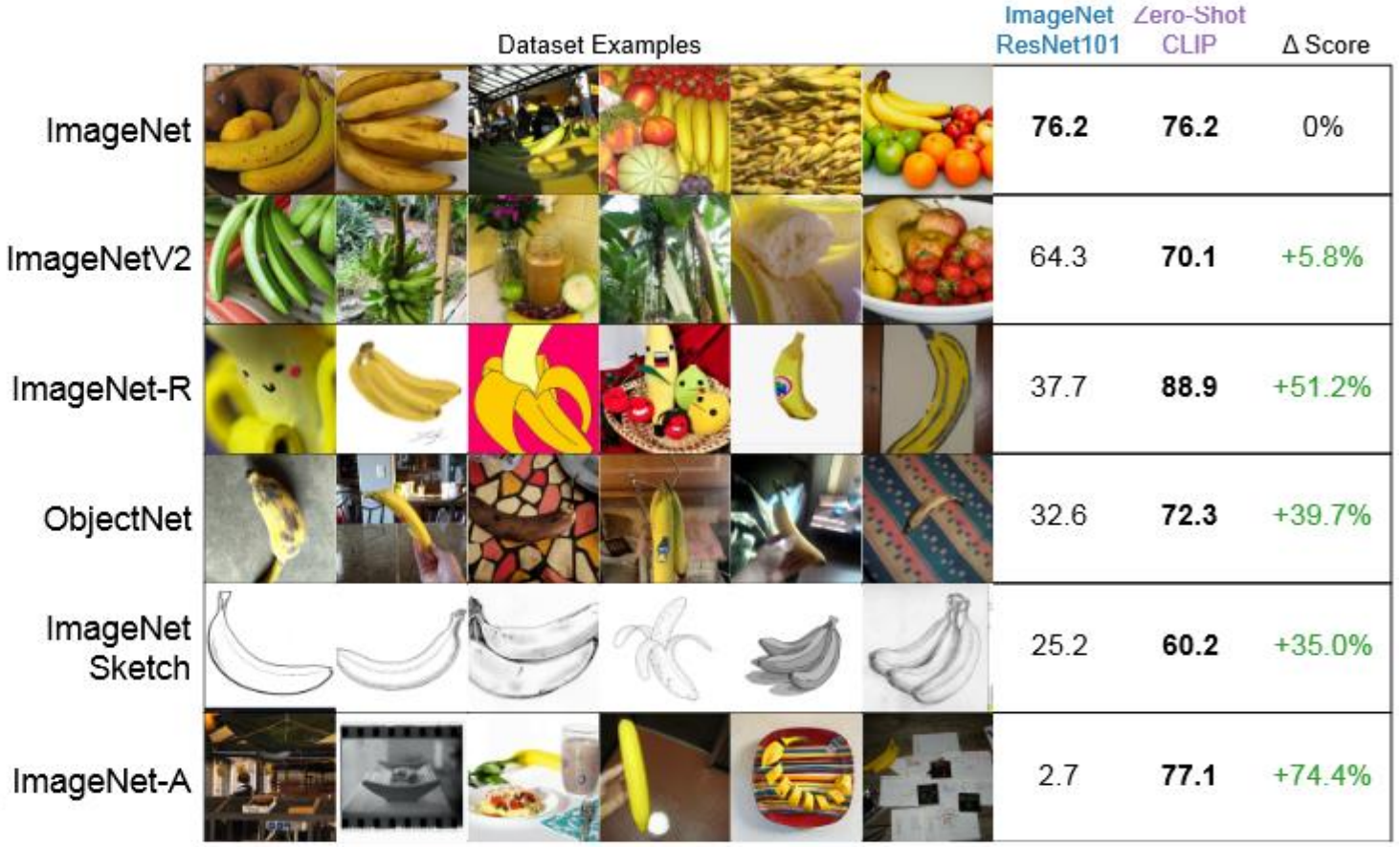
Figure 9. **Zero-shot CLIP performance scales smoothly as a function of model compute.** Across 39 evals on 36 different datasets, average zero-shot error is well modeled by a log-log linear trend across a 44x range of compute spanning 5 different CLIP models. Lightly shaded lines are performance on individual evals, showing that performance is much more varied despite the smooth overall trend.

CLIP – Experiment: Representation Learning



Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models.





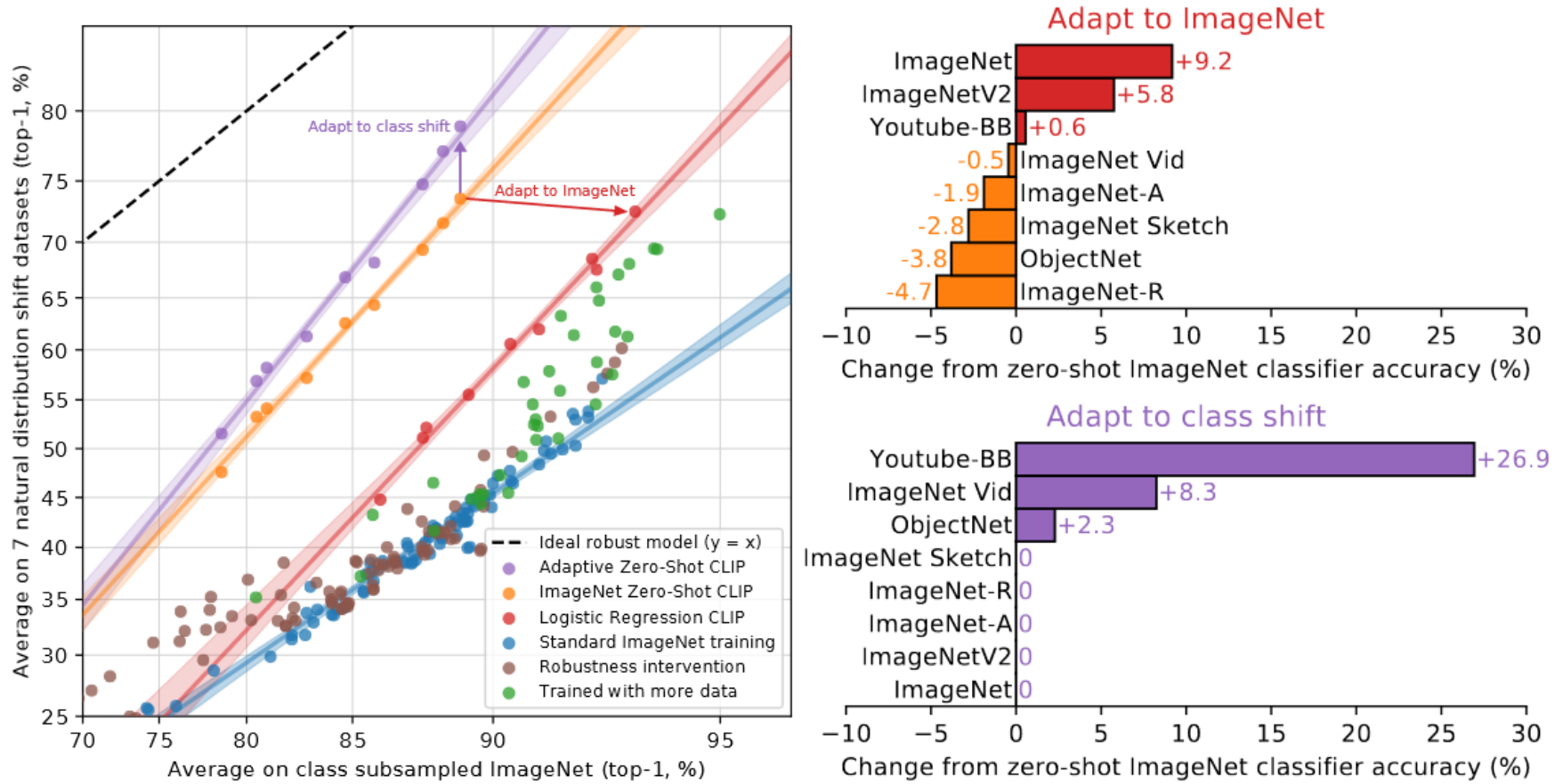


Figure 14. **While supervised adaptation to ImageNet increases ImageNet accuracy by 9.2%, it slightly reduces average robustness.** (Left) Customizing zero-shot CLIP to each dataset improves robustness compared to using a single static zero-shot ImageNet classifier and pooling predictions across similar classes as in Taori et al. (2020). CLIP models adapted to ImageNet have similar effective robustness as the best prior ImageNet models. (Right) Details of per dataset changes in accuracy for the two robustness interventions. Adapting to ImageNet increases accuracy on ImageNetV2 noticeably but trades off accuracy on several other distributions. Dataset specific zero-shot classifiers can improve accuracy by a large amount but are limited to only a few datasets that include classes which don't perfectly align with ImageNet categories.

CLIP - Limitations

1. CLIP struggles on more abstract or systematic tasks such as counting the number of objects in an image and on more complex tasks such as predicting how close the nearest car is in a photo.
2. CLIP also still has poor generalization to images not covered in its pre-training dataset.
3. CLIP's zero-shot classifiers can be sensitive to wording or phrasing and sometimes require trial and error "prompt engineering" to perform well.

Broader impacts

1 . Potential social impacts, such as privacy violations and biases in decision-making.

2. Future application on a wide range of tasks.

Reflection

1. Nature Language Supervision is a powerful high-dimensional supervision signal, which can reduce ambiguity and be easily transferable.
2. Larger queries set will further enhance the model performance.
3. CLIP transforms the classification task into retrieval task, which makes it be efficient.

PennState

Thanks for listening

Any Questions?