

Learning Transferable Visual Models From Natural Language Supervision (CLIP)

Alec Radford * 1 Jong Wook Kim * 1 Chris Hallacy 1 Aditya Ramesh 1 Gabriel Goh 1
Sandhini Agarwal 1 Girish Sastry 1 Amanda Askell 1 Pamela Mishkin 1 Jack Clark 1
Gretchen Krueger 1 Ilya Sutskever 1

OpenAI

April, 15, 2021
구재원

Introduction

>> CLIP: Zero-shot Image Classifier similar to “GPT-2” and “GPT-3”



Combining Text and Image

>> CLIP efficiently learns visual concepts from natural language supervision



CLIP
Predict



My cute dog.



DALL-E
: Image Generative Model

Building Classifiers

>> 일반적인 zero-shot classifier

Efficiency of CLIP

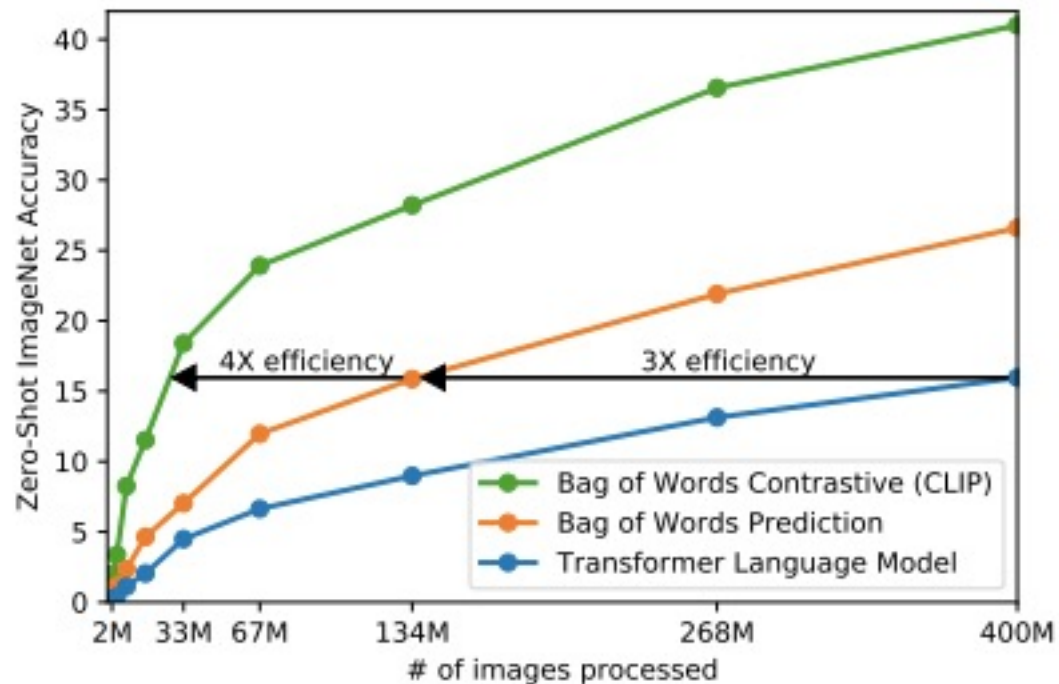


Figure 2. CLIP is much more efficient at zero-shot transfer than our image caption baseline. Although highly expressive, we found that transformer-based language models are relatively weak at zero-shot ImageNet classification. Here, we see that it learns 3x slower than a baseline which predicts a bag-of-words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

>>Transformer Language Model:
used CNN for Image,
trained to predict exact word

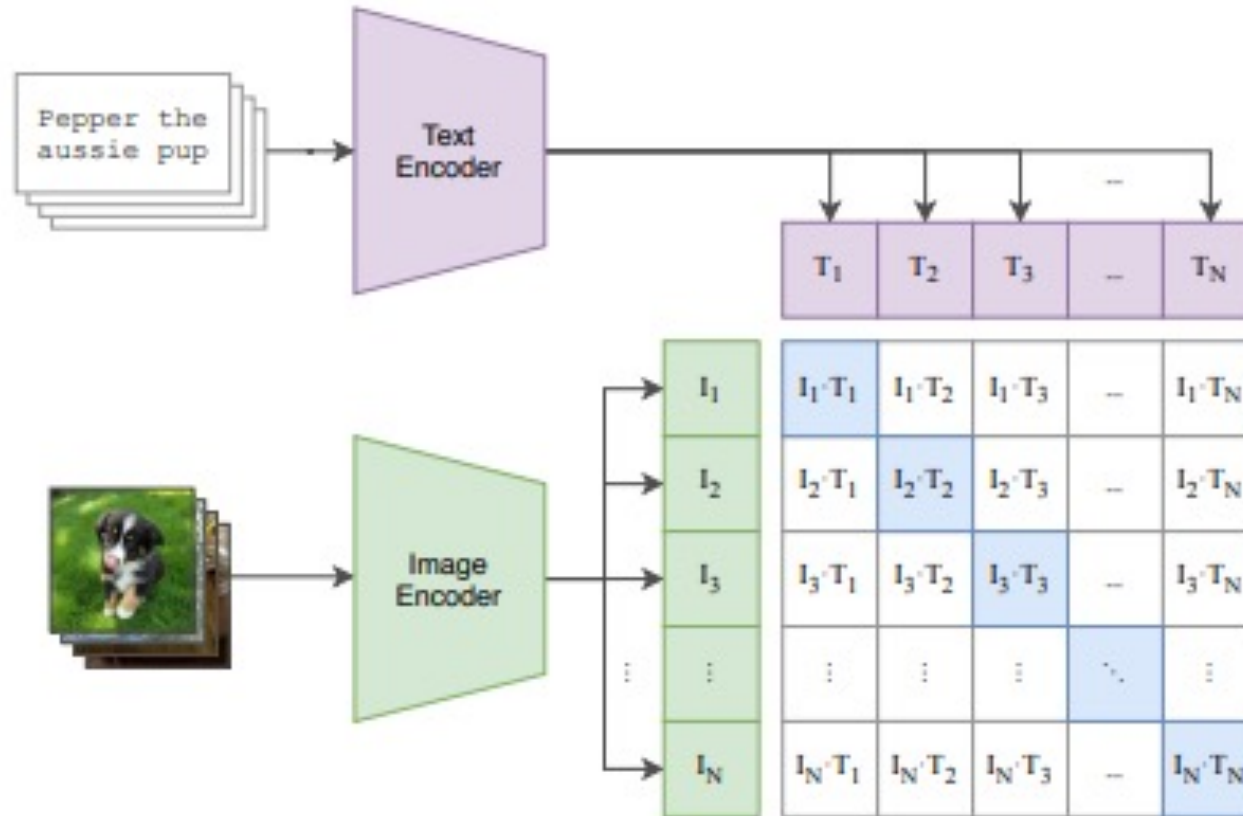
>>Bag of Words Prediction

>>Contrastive Representation Learning

Overall Network: CLIP

[Training Time]

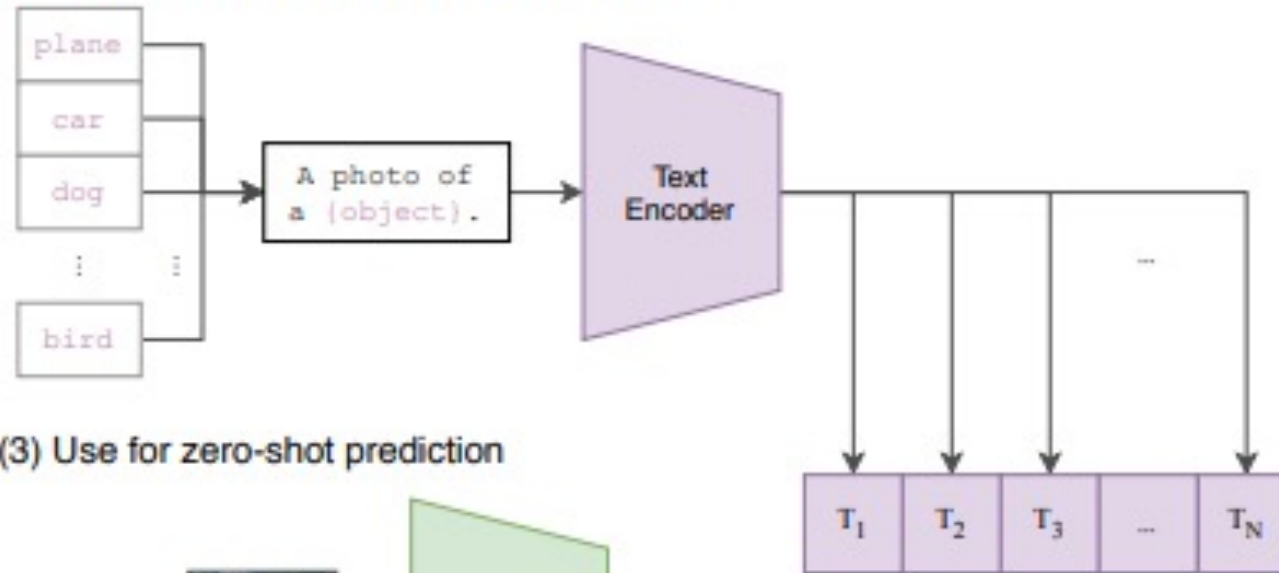
(1) Contrastive pre-training



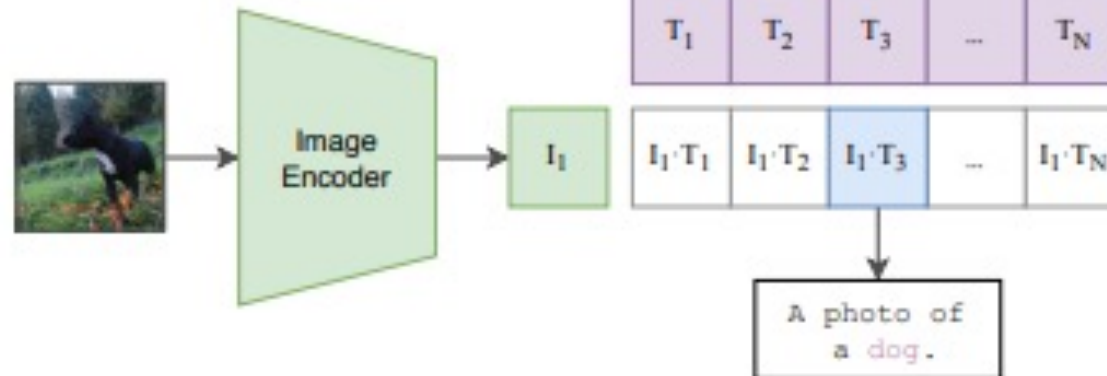
Overall Network: CLIP

[Inference Time] : zero training needed

(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



Prompt Engineering and Ensemble

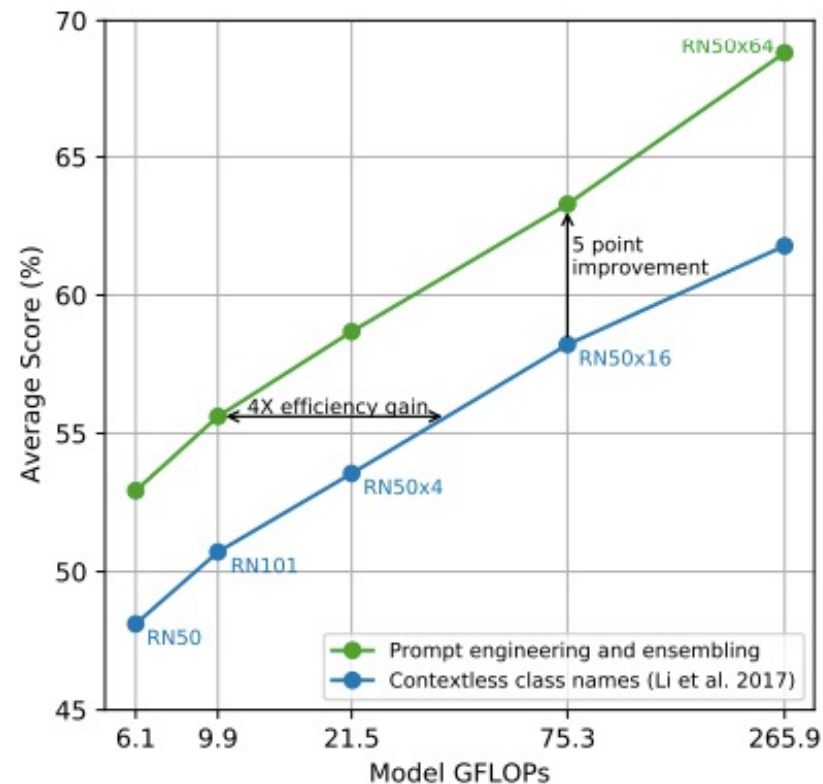


Figure 4. Prompt engineering and ensembling improve zero-shot performance. Compared to the baseline of using contextless class names, prompt engineering and ensembling boost zero-shot classification performance by almost 5 points on average across 36 datasets. This improvement is similar to the gain from using 4 times more compute with the baseline zero-shot method but is “free” when amortized over many predictions.

Analysis of Zero-shot CLIP Performance

>> Linear Probe

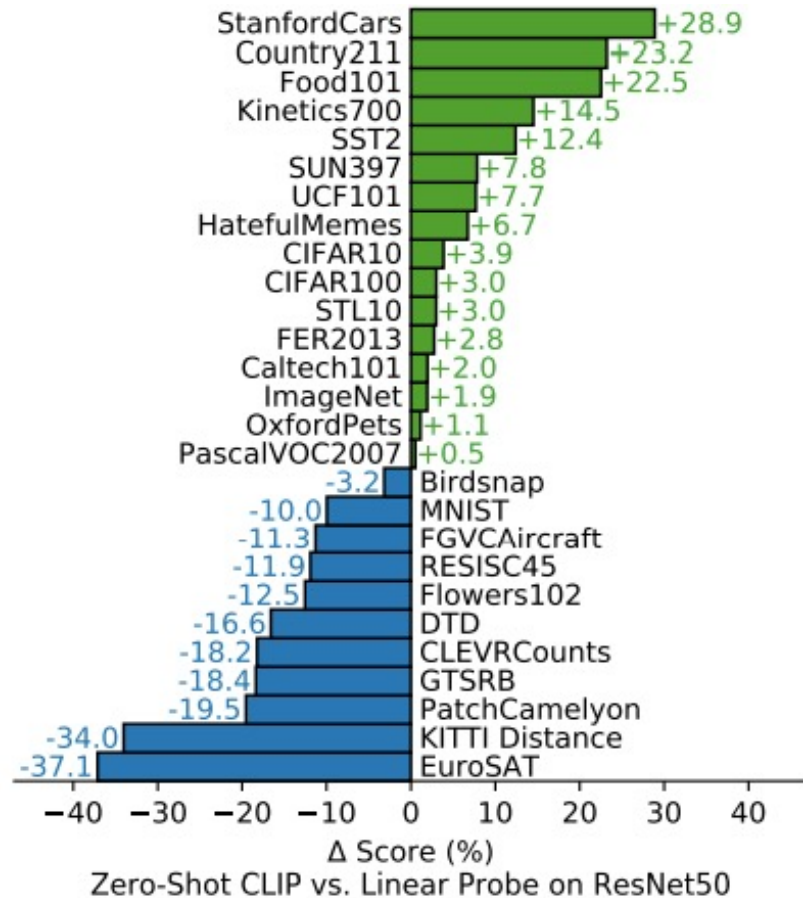


Figure 5. Zero-shot CLIP is competitive with a fully supervised baseline. Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

Analysis of Zero-shot CLIP Performance

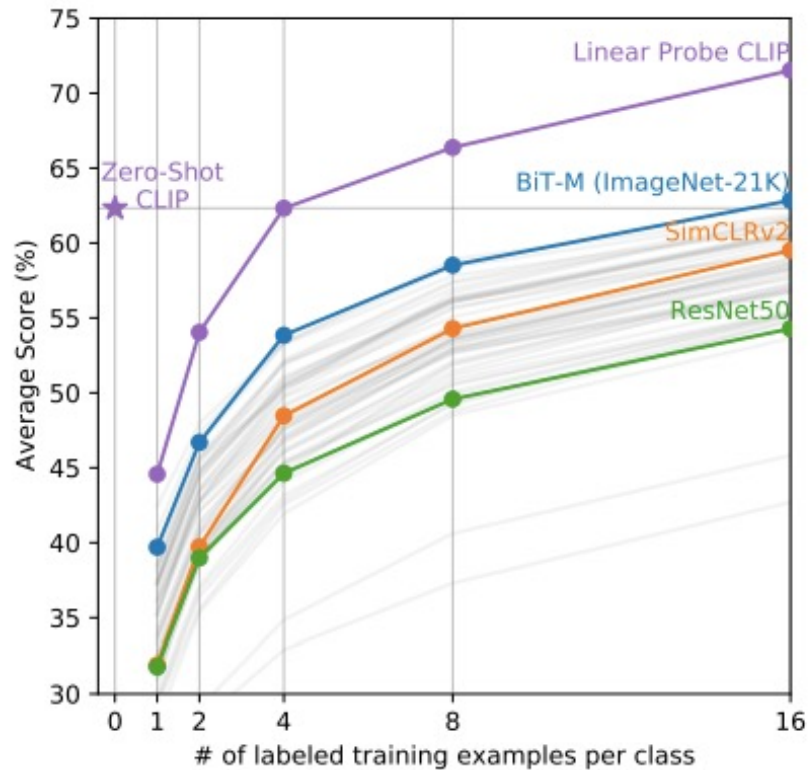


Figure 6. Zero-shot CLIP outperforms few-shot linear probes. Zero-shot CLIP matches the average performance of a 4-shot linear classifier trained on the same feature space and nearly matches the best results of a 16-shot linear classifier across publicly available models. For both BiT-M and SimCLRv2,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

>> Zero-shot CLIP > other few-shot linear probes
>> CLIP에서도 4 shot이상에서 성능이 더 좋아짐

Analysis of Zero-shot CLIP Performance

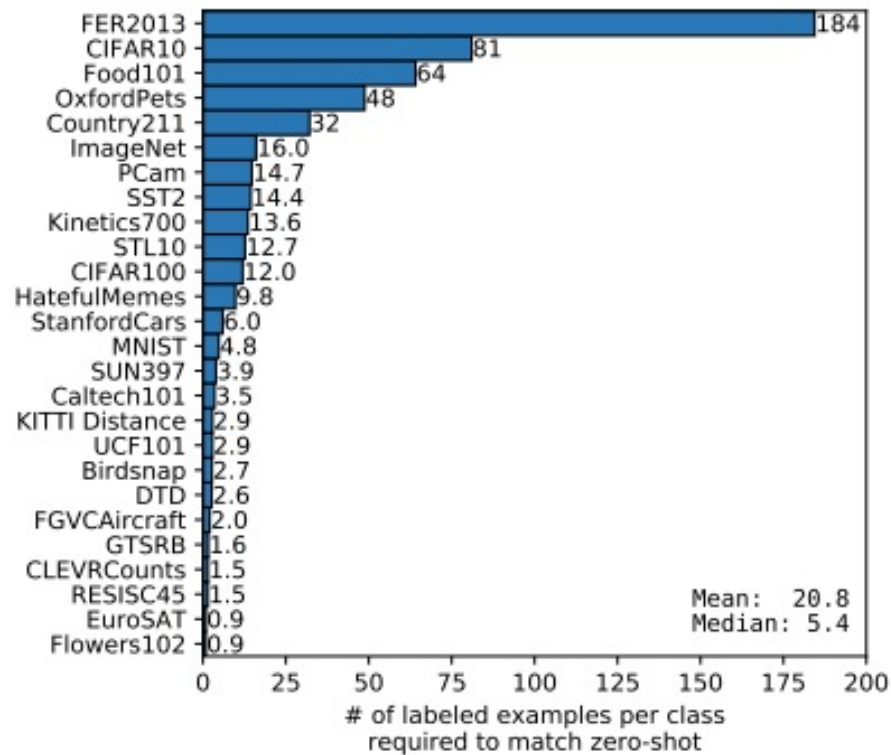


Figure 7. The data efficiency of zero-shot transfer varies widely. Calculating the number of labeled examples per class a linear classifier on the same CLIP feature space requires to match the performance of the zero-shot classifier contextualizes the effectiveness of zero-shot transfer. Values are estimated 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.

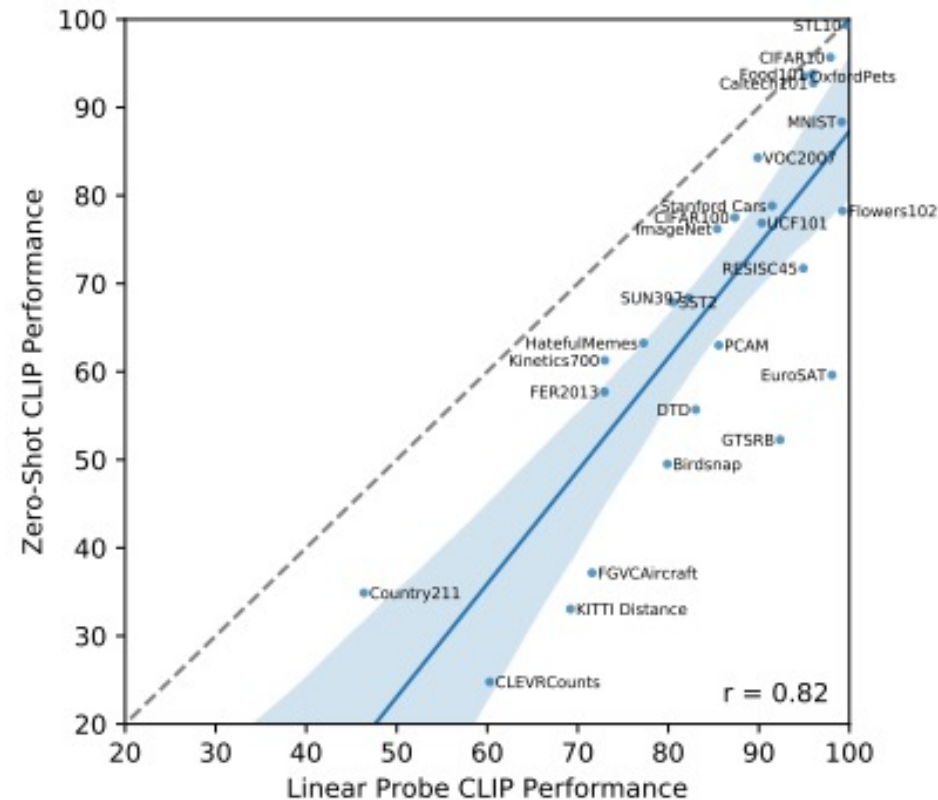


Figure 8. Zero-shot performance is correlated with linear probe performance but still mostly sub-optimal. Comparing zero-shot and linear probe performance across datasets shows a strong correlation with zero-shot performance mostly shifted 10 to 25 points lower. On only 5 datasets does zero-shot performance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

Analysis of Zero-shot CLIP Performance

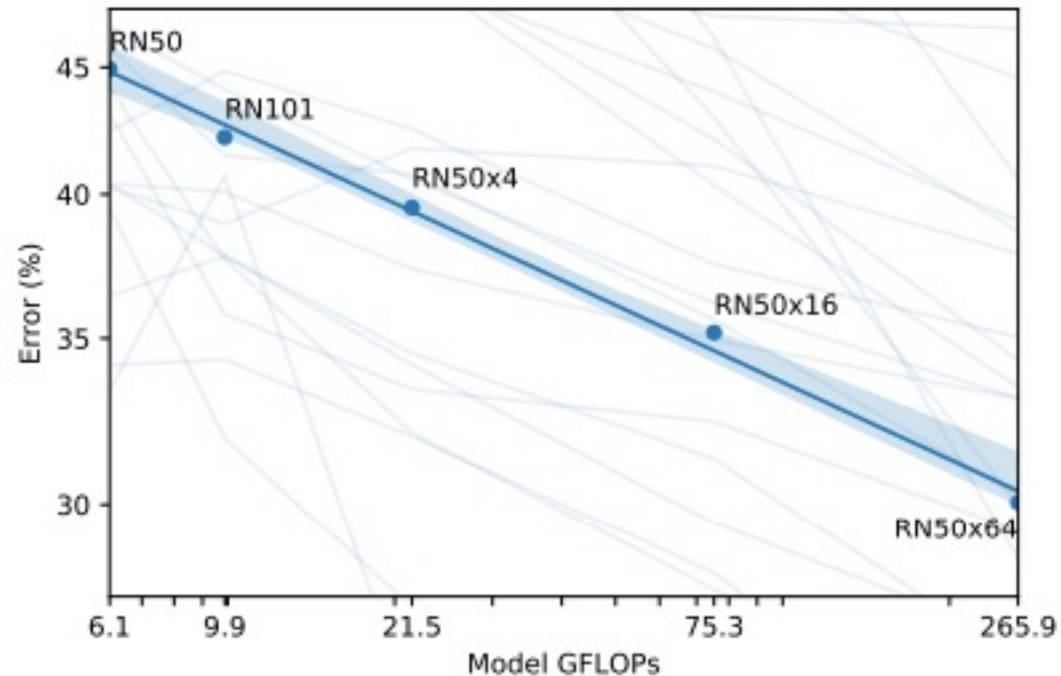
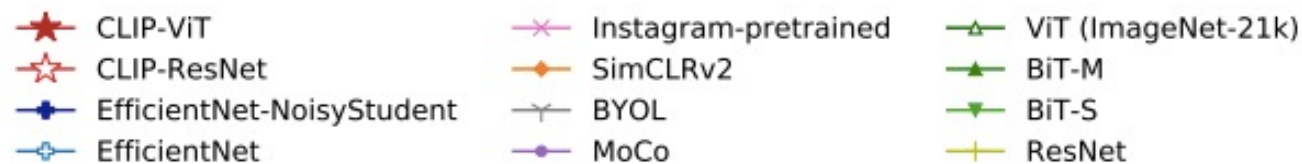


Figure 9. Zero-shot CLIP performance scales smoothly as a function of model compute. Across 39 evals on 36 different datasets, average zero-shot error is well modeled by a log-log linear trend across a 44x range of compute spanning 5 different CLIP models. Lightly shaded lines are performance on individual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

[illegible]

Linear-probe Performance

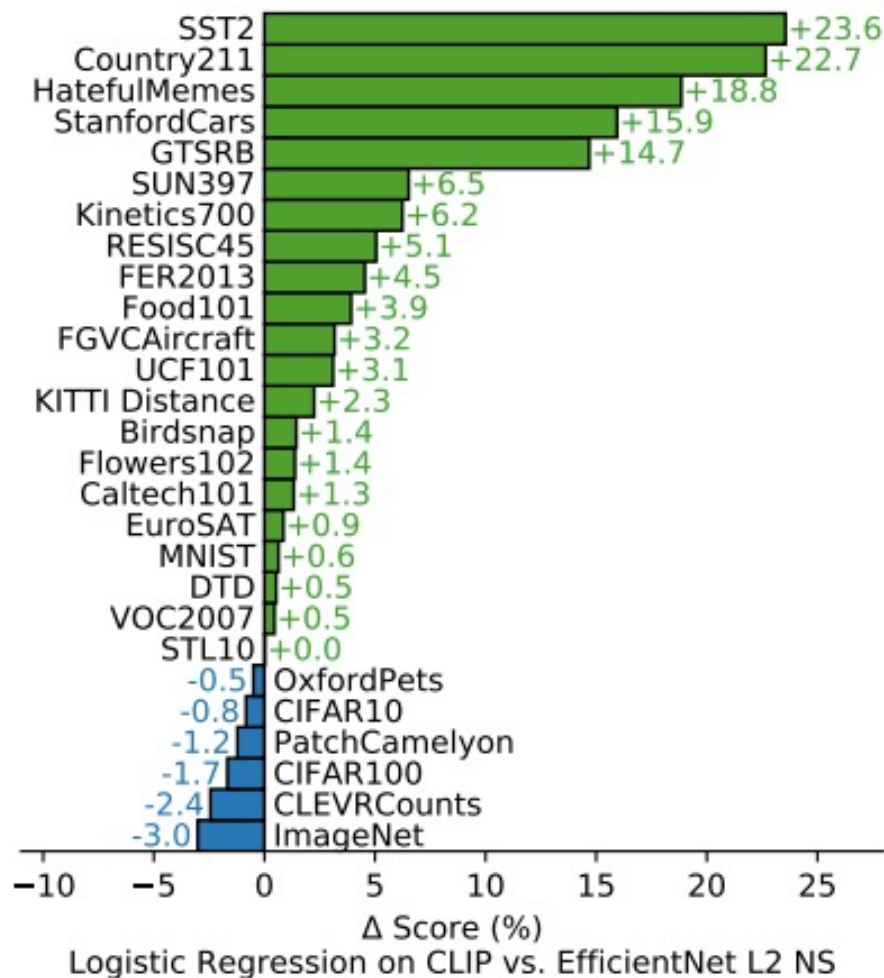
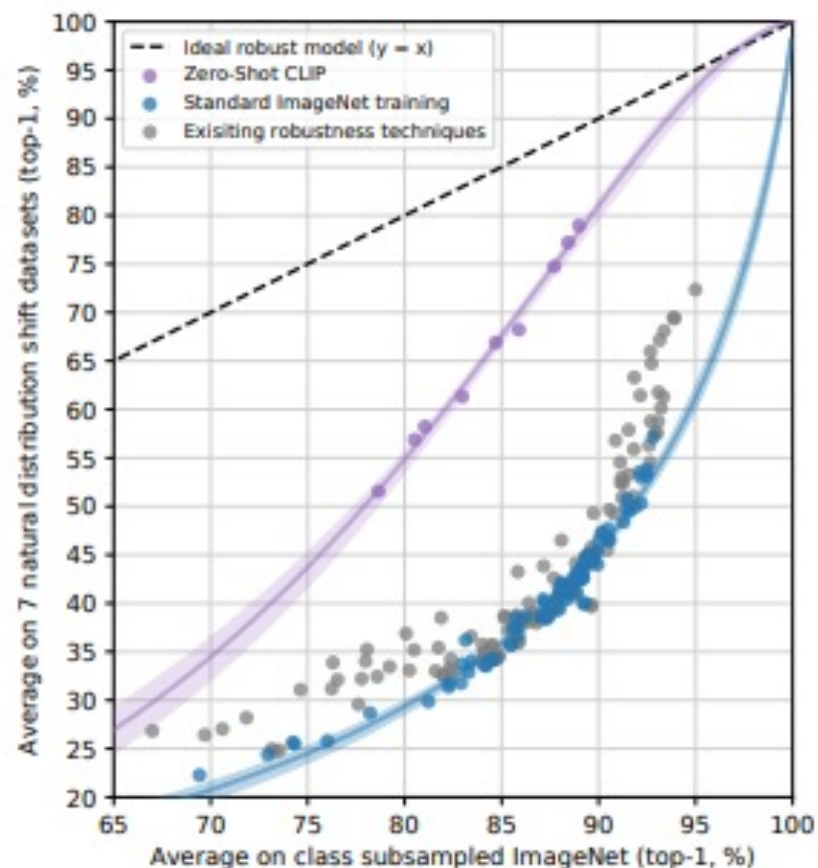
















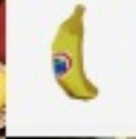


















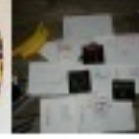


Figure 11. CLIP's features outperform the features of the best ImageNet model on a wide variety of datasets. Fitting a linear classifier on CLIP's features outperforms using the Noisy Student EfficientNet-L2 on 21 out of 27 datasets.

Robustness of CLIP



	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%

Robustness of CLIP

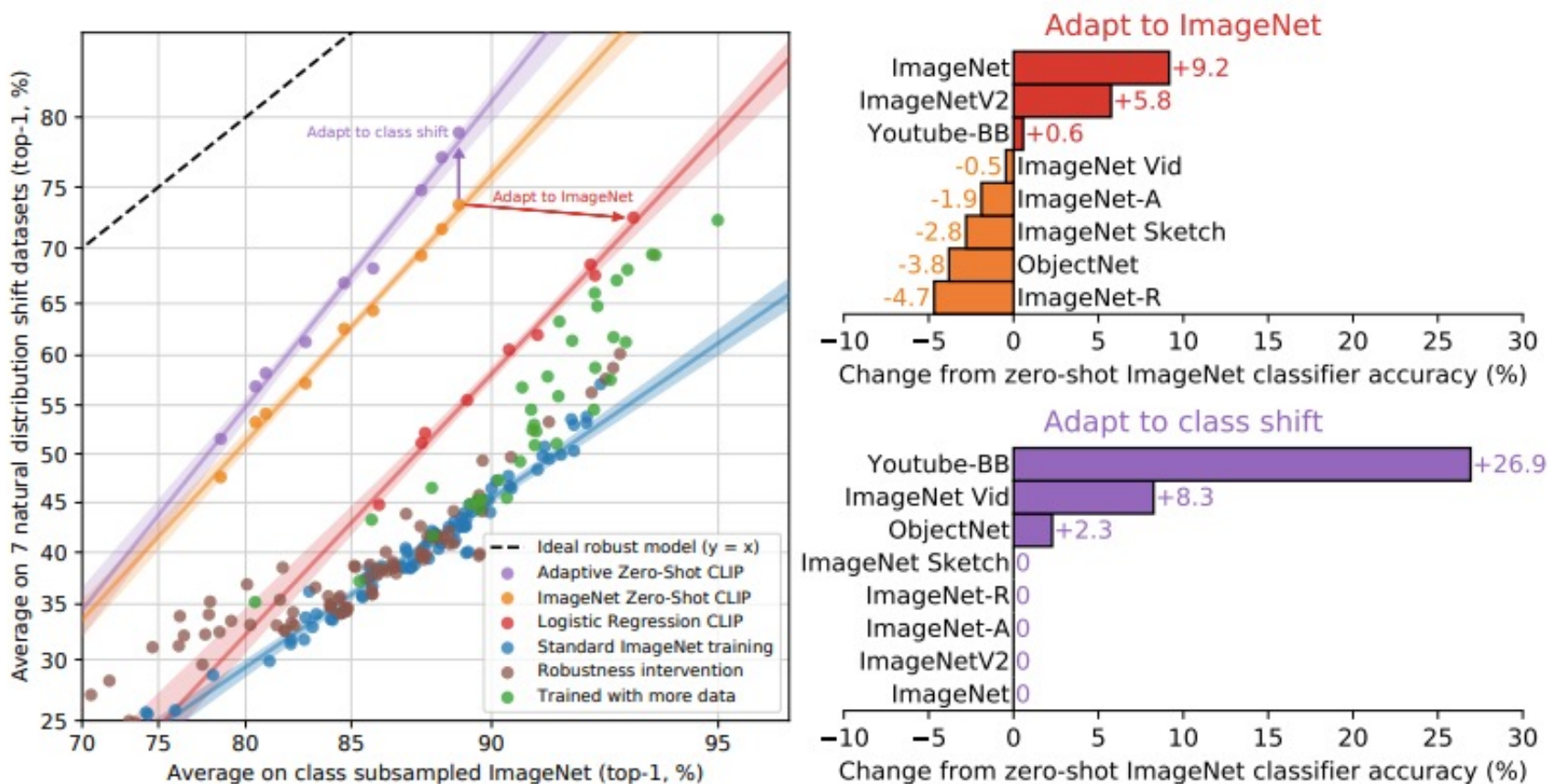


Figure 14. While supervised adaptation to ImageNet increases ImageNet accuracy by 9.2%, it slightly reduces average robustness. (Left) Customizing zero-shot CLIP to each dataset improves robustness compared to using a single static zero-shot ImageNet classifier and pooling predictions across similar 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.

Comparison to Human Performance

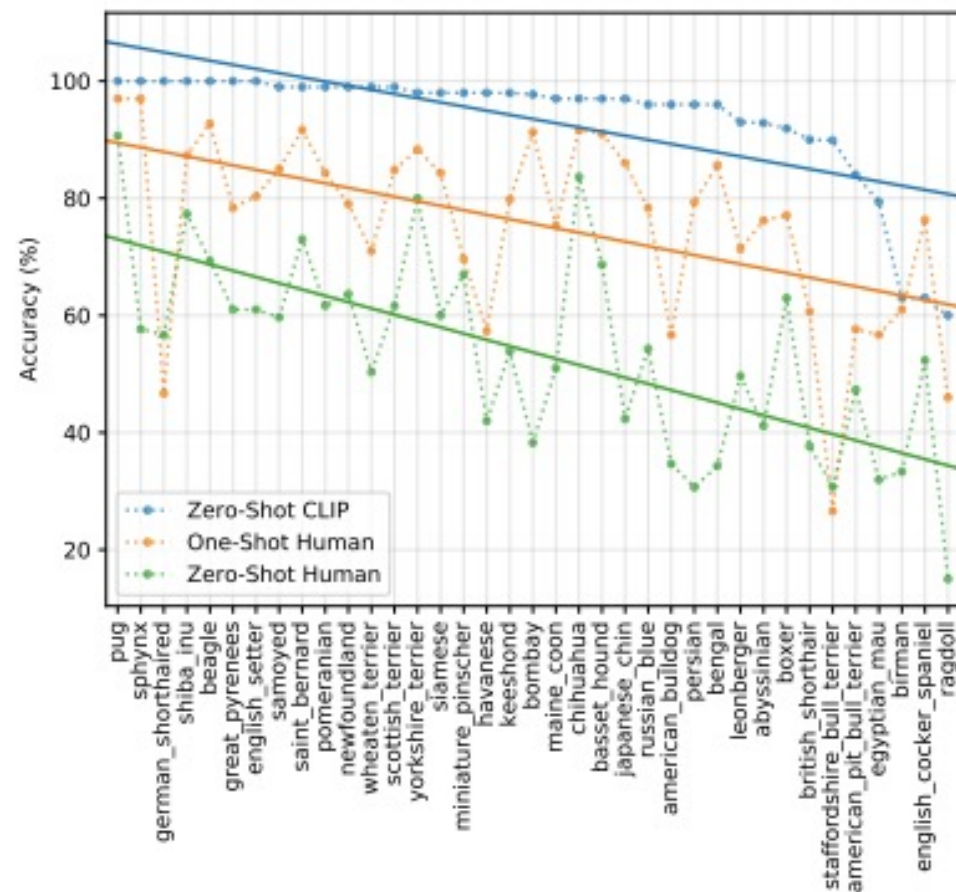


Figure 16. The hardest problems for CLIP also tend to be the hardest problems for humans. Here we rank image categories by difficulty for CLIP as measured as probability of the correct label.

Broader Impacts

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

Table 7. Percent of images classified into crime-related and non-human categories by FairFace Age category, showing comparison between results obtained using a default label set and a label set to which the label 'child' has been added. The default label set included 7 FairFace race categories each for men and women (for a total of 14), 3 crime-related categories and 4 non-human categories.