Learning Transferable Visual Models From Natural Language Supervision

Alec Radford, Jong Wook Kim et al.

OpenAI

Presenter: Minho Park

Zero-shot Classification

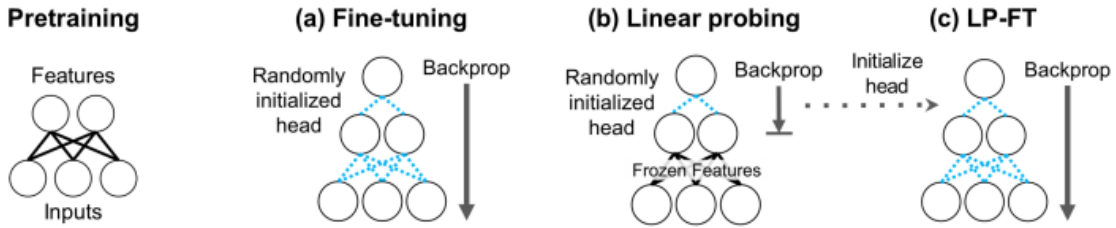
- Learning visual n-grams from web data.
- However, the most cases we use fine-tuning or linear probing

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

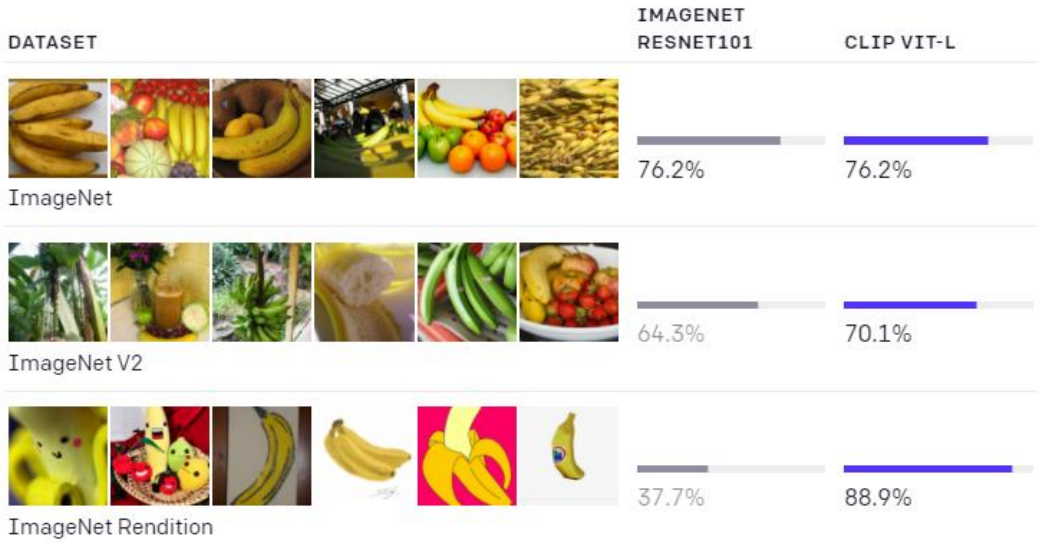
Figure 1. Four high-scoring visual n-grams for three images in our test set according to our visual n-gram model, which was trained *solely* on *unsupervised* web data. We selected the n-grams that are displayed in the figure from the five highest scoring n-grams according to our model, in such a way as to minimize word overlap between the n-grams. For all figures in the paper, we refer the reader to the supplementary material for license information.

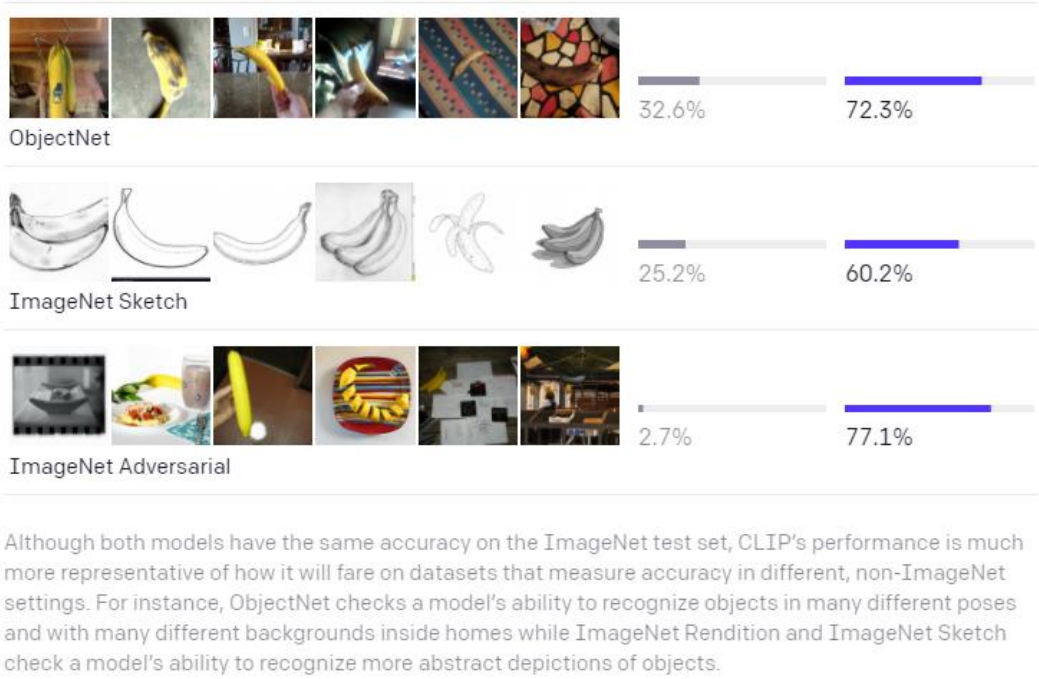


Li, Ang, et al. "Learning visual n-grams from web data." *Proceedings of the IEEE International Conference on Computer Vision*. 2017.
Anonymous authors, “FINE-TUNING DISTORTS PRETRAINED FEATURES AND UNDERPERFORMS OUT-OF-DISTRIBUTION”

“Zero-shot” capabilities of GPT-2 and GPT-3

- This is a key change: by not directly optimizing for the benchmark, we show that it becomes much more representative





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.

Examples of Zero-shot Classification

- FOOD101, ImageNet, ImageNet-A, CIFAR-10, EUROSAT, …
- https://openai.com/blog/clip/

Natural Language Supervision

- Learning perception from supervision contained in natural language.
- It is much easier to scale.
  - It does not require annotations such as the canonical 1-of-N majority vote “gold label”.
- It enables flexible zero-shot transfer.
  - It does not “just” learn a representation but also connects that representation to language.

Creating a Sufficiently Large Dataset

- Existing datasets (e.g., MS-COCO, Visual Genome, and YFCC100M) would underestimate the potential of this line of research.
- Therefore, they constructed a new dataset of 400 million (image, text) pairs collected from a variety of publicly available on the Internet.
- WebImageText (WIT)
  - It has a similar total word count as the WebText dataset used to train GPT-2.

Selecting an Efficient Pre-Training Method

- Blue: Predicts the caption of an image.
- Orange: Predicts a bag-of-words encoding of the same text.
- Why are they bad? **They try to predict the exact words.**
  - Wide variety of descriptions

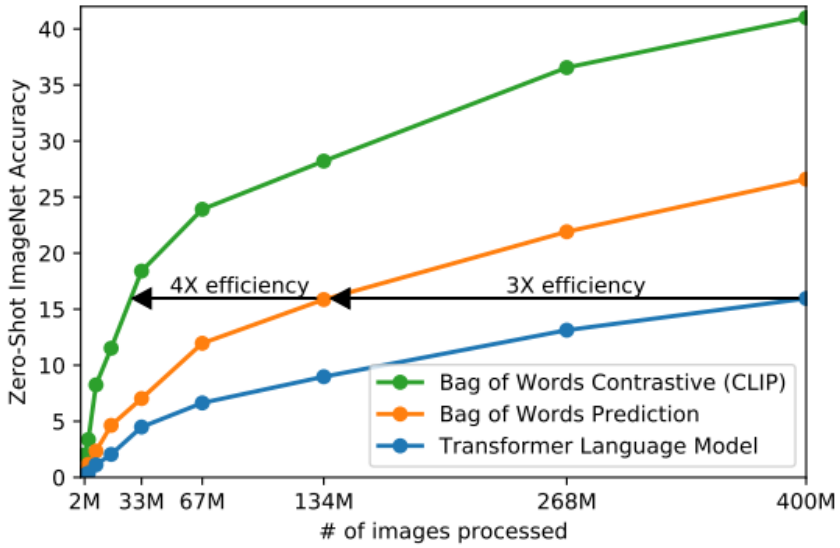


Figure 2. **CLIP is much more efficient at zero-shot transfer than our image caption baseline.** Although highly expressive, we found that transformer-based language models are relatively weak at zero-shot ImageNet classification. Here, we see that it learns 3x slower than a baseline which predicts a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

Selecting an Efficient Pre-Training Method

- Contrastive Language-Image Pre-training

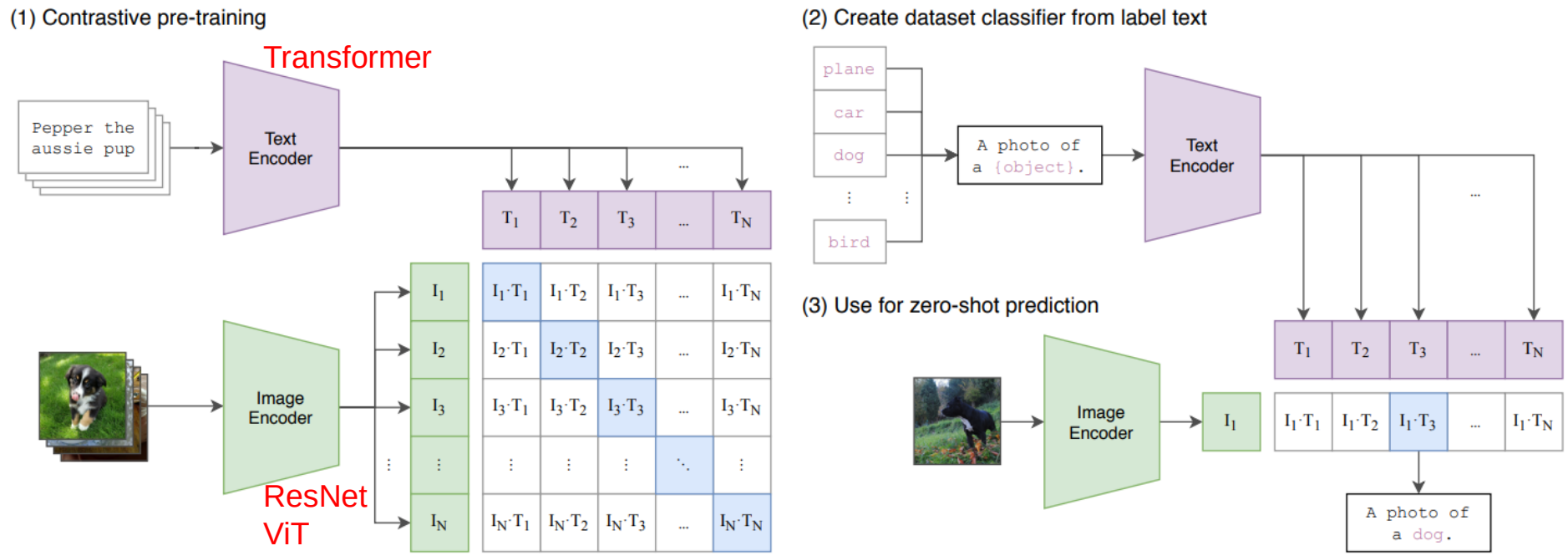


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the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

Implementation Details

- Prompt engineering and ensemble

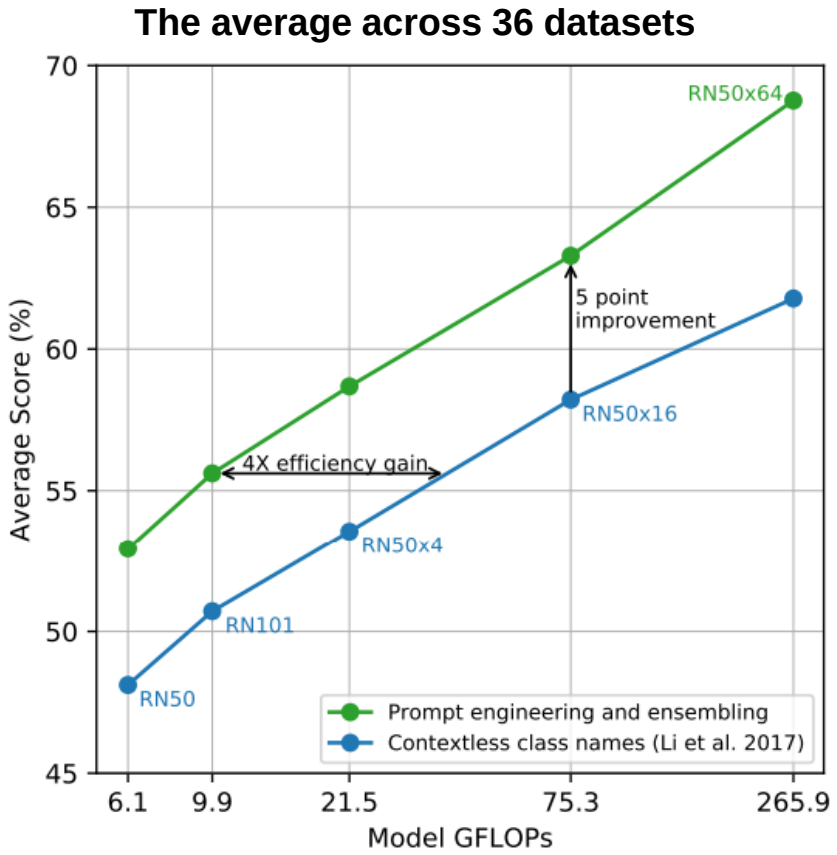


Figure 4. Prompt engineering and ensembling improve zero-shot performance.

Experiments

- Zero-shot CLIP vs. fully supervised baseline (linear probe on ResNet50).

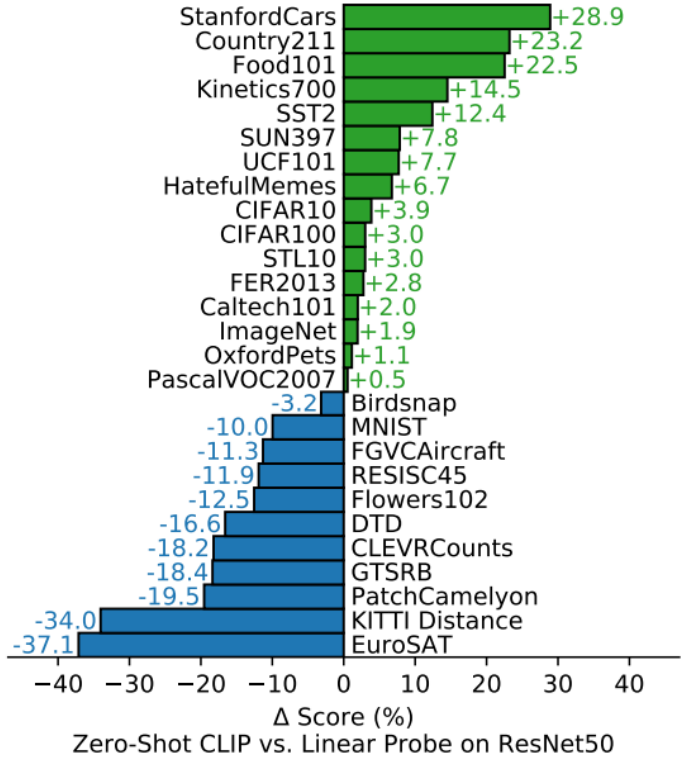


Figure 5. **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

Experiments

- Zero-shot CLIP vs. few-shot linear probes (CLIP, BiT, SimCLRv2, ResNet50)

Unlike traditional classification models,
the last layer of CLIP gives plausible representation.

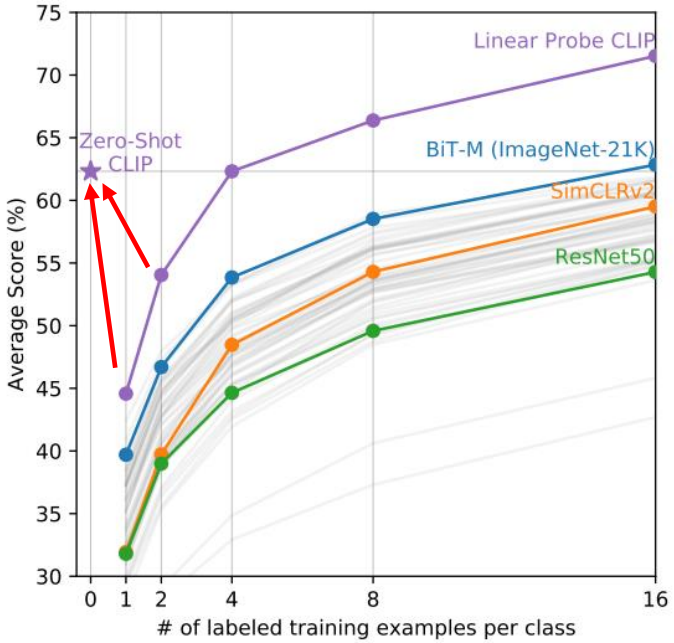


Figure 6. **Zero-shot CLIP outperforms few-shot linear probes.** Zero-shot CLIP matches the average performance of a 4-shot linear classifier trained on the same feature space and nearly matches the best results of a 16-shot linear classifier across publicly available models. For both BiT-M and SimCLRv2,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

Experiments

- Linear probe performance of CLIP models.

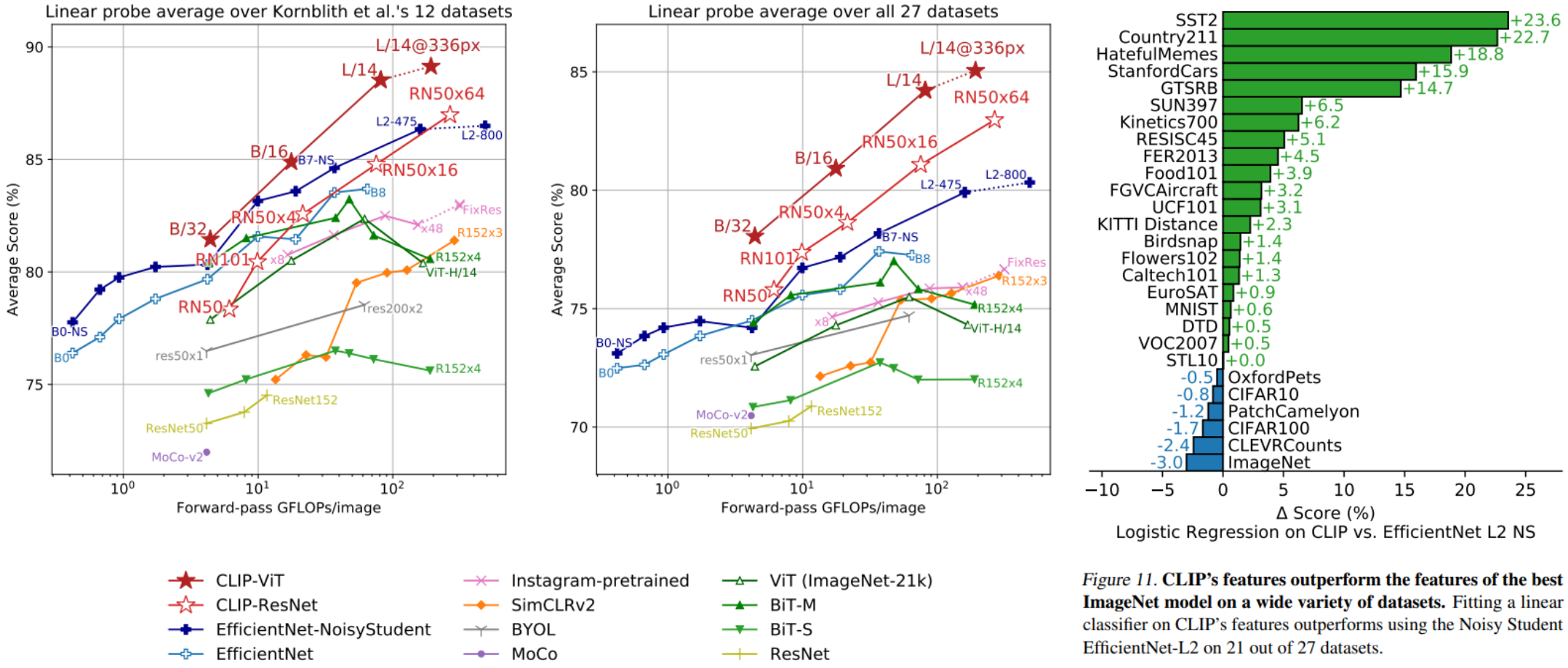


Figure 11. **CLIP's features outperform the features of the best ImageNet model on a wide variety of datasets.** Fitting a linear classifier on CLIP's features outperforms using the Noisy Student EfficientNet-L2 on 21 out of 27 datasets.

Experiments

- Robustness

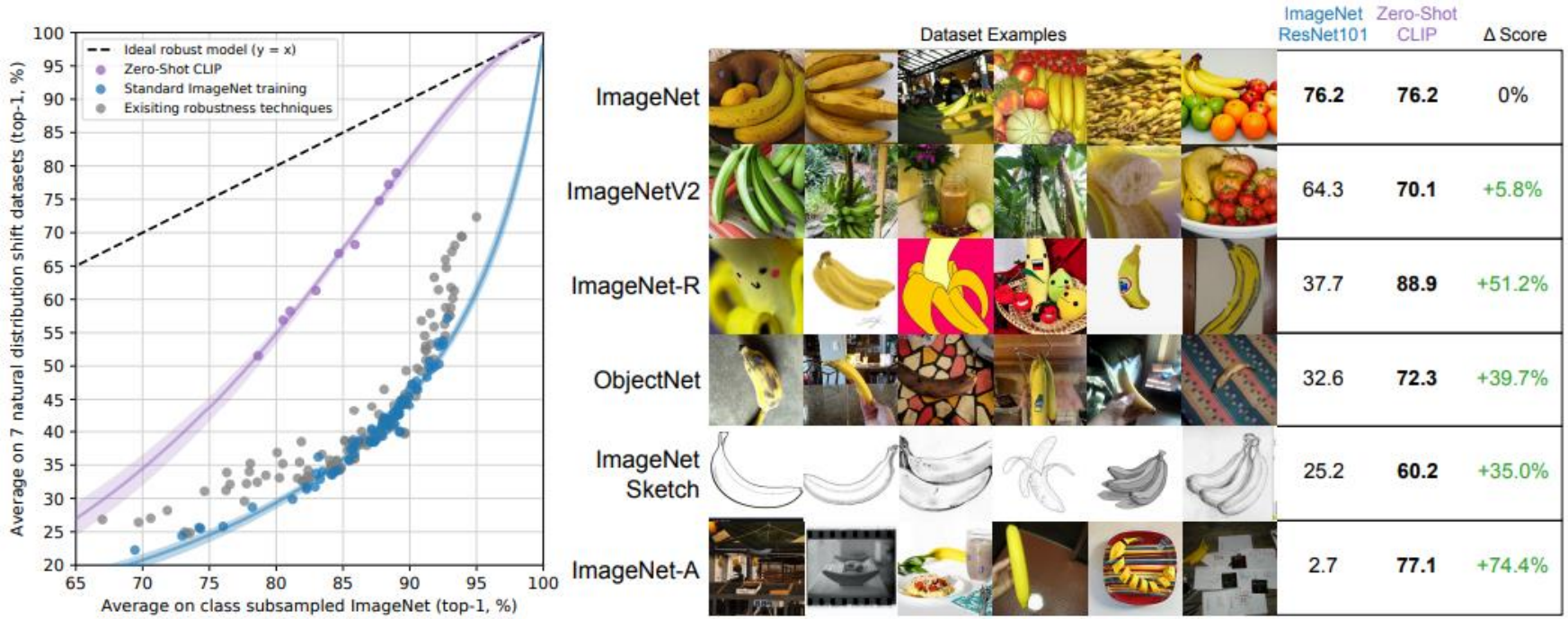


Figure 13. **Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models.** (Left) An ideal robust model (dashed line) performs equally well on the ImageNet distribution and on other natural image distributions. Zero-shot CLIP models shrink this "robustness gap" by up to 75%. 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.

Limitations

- ResNet-50 is now well below the overall state-of-the-art.
- They estimate around a 1000x increase in compute is required for zero-shot CLIP to reach overall state-of-the-art performance (infeasible).
- Zero-shot CLIP generalizes poorly to data truly out-of-distribution for it.
- CLIP is still limited to choosing from only those concepts in a given zero-shot classifier.

Implementation

- Usage and API: https://github.com/openai/CLIP

```python
import torch
import clip
from PIL import Image

device = "cuda" if torch.cuda.is_available() else "cpu"
model, preprocess = clip.load("ViT-B/32", device=device)

image = preprocess(Image.open("CLIP.png")).unsqueeze(0).to(device)
text = clip.tokenize(["a diagram", "a dog", "a cat"]).to(device)

with torch.no_grad():
    image_features = model.encode_image(image)
    text_features = model.encode_text(text)

    logits_per_image, logits_per_text = model(image, text)
    probs = logits_per_image.softmax(dim=-1).cpu().numpy()

print("Label probs:", probs)  # prints: [[0.9927937  0.00421068 0.00299572]]
```

```python
_MODELS = {
    "RN50": "https://openaipublic.azureedge.net/clip/models/afeb0e10f9e5a86da6080e35cf09123aca3b358a0c3e3b6c78a7b63bc04b6762/RN50.pt",
    "RN101": "https://openaipublic.azureedge.net/clip/models/8fa8567bab74a42d41c5915025a8e4538c3bdbe8804a470a72f30b0d94fab599/RN101.pt",
    "RN50x4": "https://openaipublic.azureedge.net/clip/models/7e526bd135e493cef0776de27d5f42653e6b4c8bf9e0f653bb11773263205fdd/RN50x4.pt",
    "RN50x16": "https://openaipublic.azureedge.net/clip/models/52378b407f34354e150460fe41077663dd5b39c54cd0bfd2b27167a4a06ec9aa/RN50x16.pt",
    "ViT-B/32": "https://openaipublic.azureedge.net/clip/models/40d365715913c9da98579312b702a82c18be219cc2a73407c4526f58eba950af/ViT-B-32.pt",
    "ViT-B/16": "https://openaipublic.azureedge.net/clip/models/5806e77cd80f8b59890b7e101eabd078d9fb84e6937f9e85e4ecb61988df416f/ViT-B-16.pt",
}
```

References

- Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *arXiv preprint arXiv:2103.00020* (2021).
- Li, Ang, et al. "Learning visual n-grams from web data." *Proceedings of the IEEE International Conference on Computer Vision*. 2017.