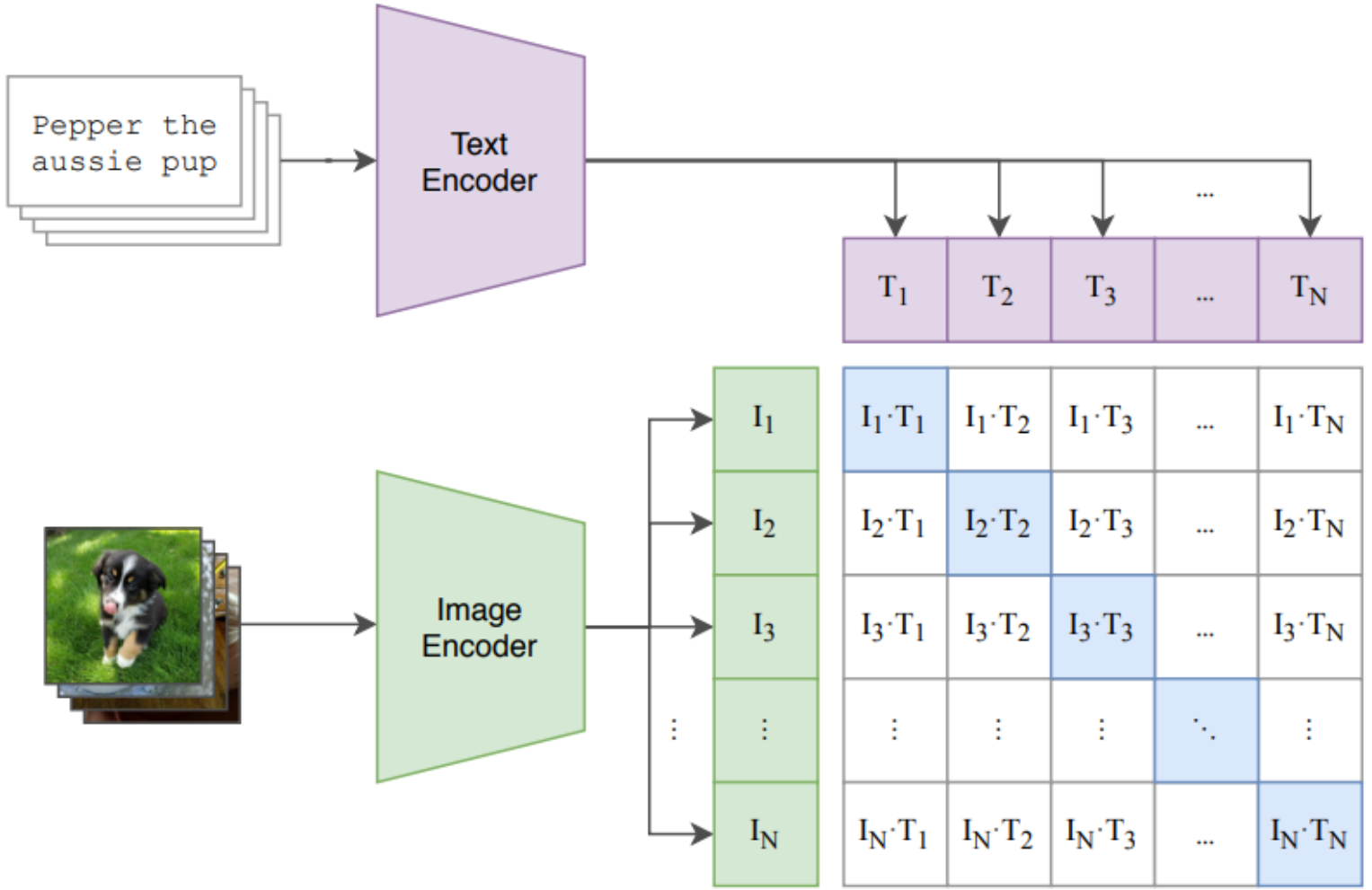
Learning Transferable Visual Models From Natural Language Supervision

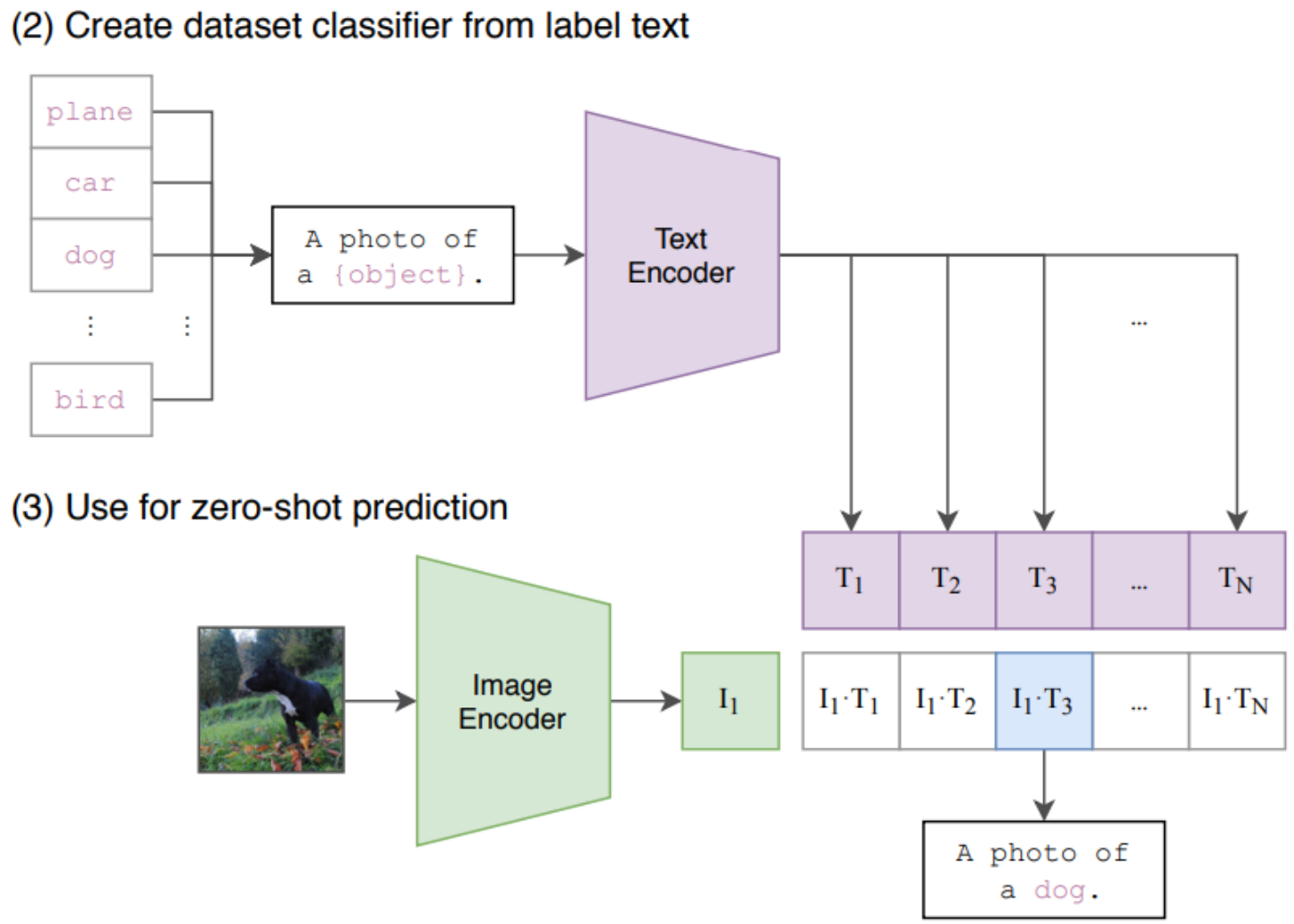
Motivating Work

- "The development of “text-to-text” as a standardized input-output interface has enabled task-agnostic architectures to zero-shot transfer to downstream" (Radford et al., 2021)
- Example: GPT-3
- Web-scale collections of text > High-quality crowd-labeled NLP datasets?
  - Crowd-labeled datasets still standard practice in computer vision
  - Could using web-text results lead to similar breakthrough?

Approach of CLIP - (1) Contrastive pre-training



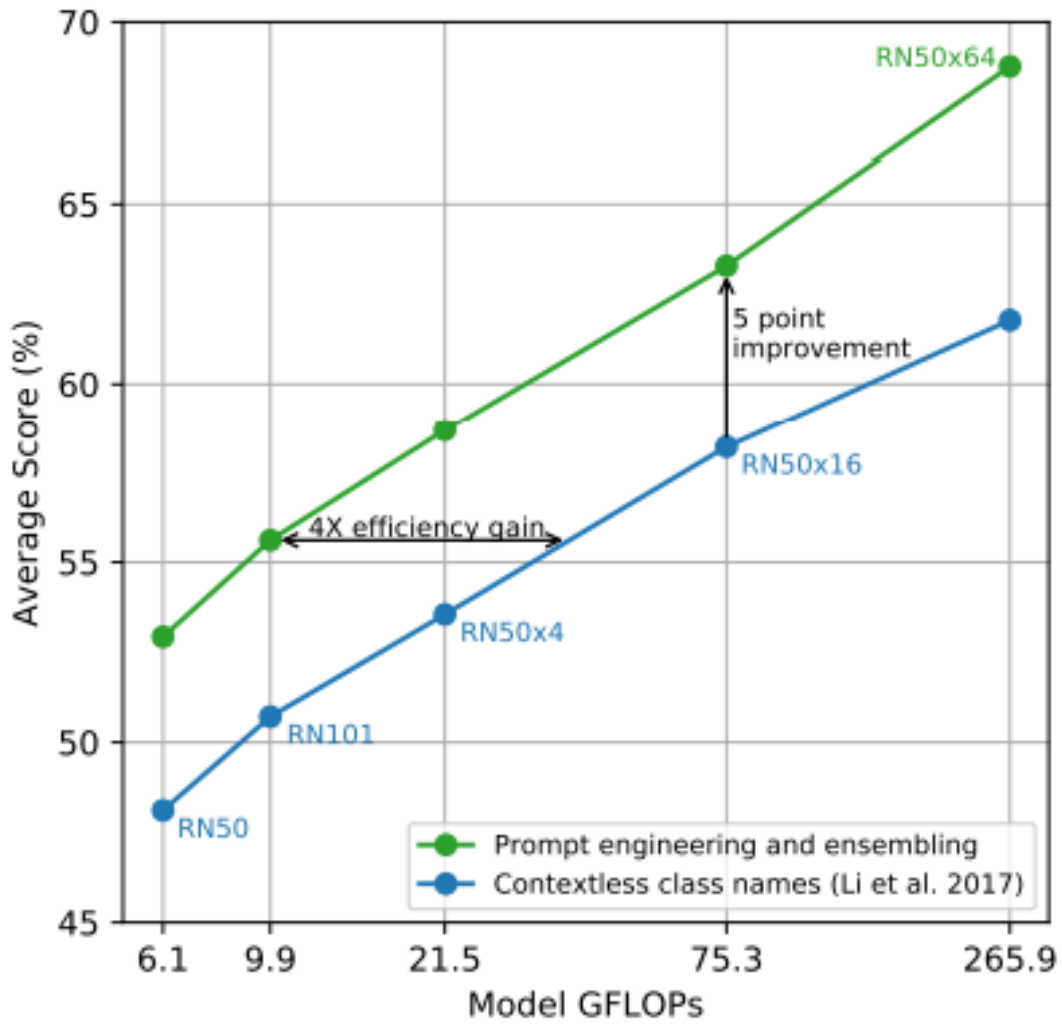
Approach of CLIP - (2) & (3)



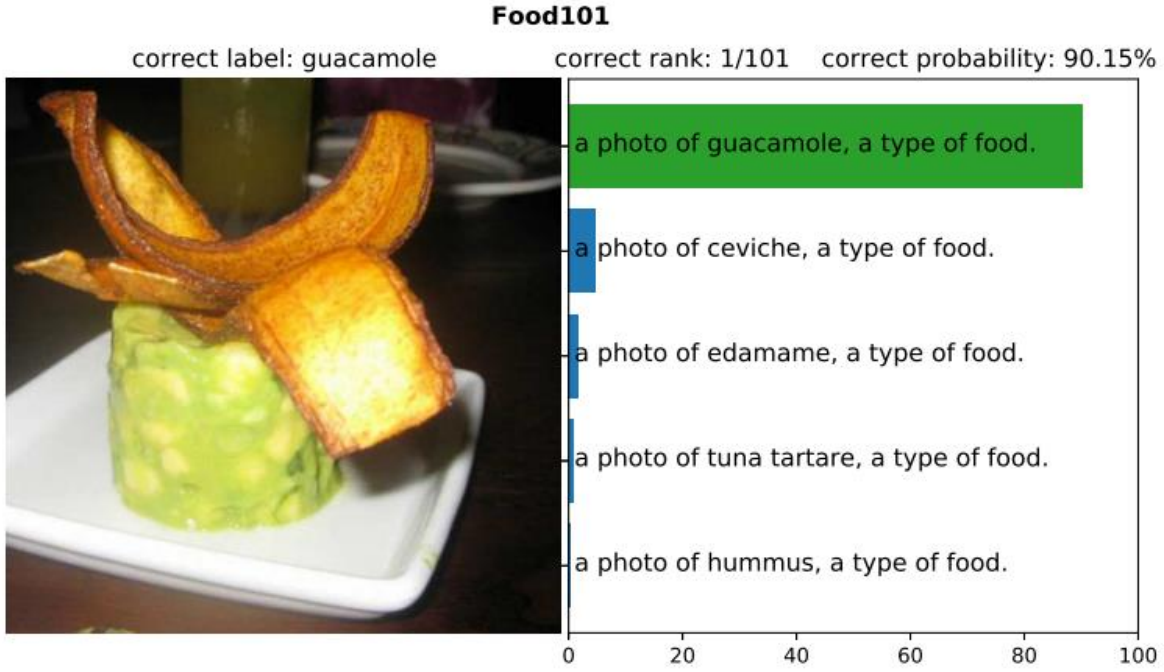
Prompt engineering and Ensembling

- On several fine-grained image classification datasets, prompt engineering improved accuracy.
- Increase in performance likely due to where they have gotten their data from, the internet.
- Examples:
  - Oxford-IIIT Pets, using “A photo of a {label}, a type of pet.” to help provide context worked well.
  - Food101 specifying "a type of food" or on FGVC Aircraft "a type of aircraft" at the end of a label helped too
  - OCR datasets, they found that putting quotes around the text or number to be recognized improved performance.
  - On satellite image datasets it helped to specify that the images were of this form and they use variants of “a satellite photo of a {label}.
- Ensembling also improved accuracy – combining shared context
  - Example: "A photo of a big {label}” and “A photo of a small {label}”.

Prompt engineering and ensembling improve zero-shot performance

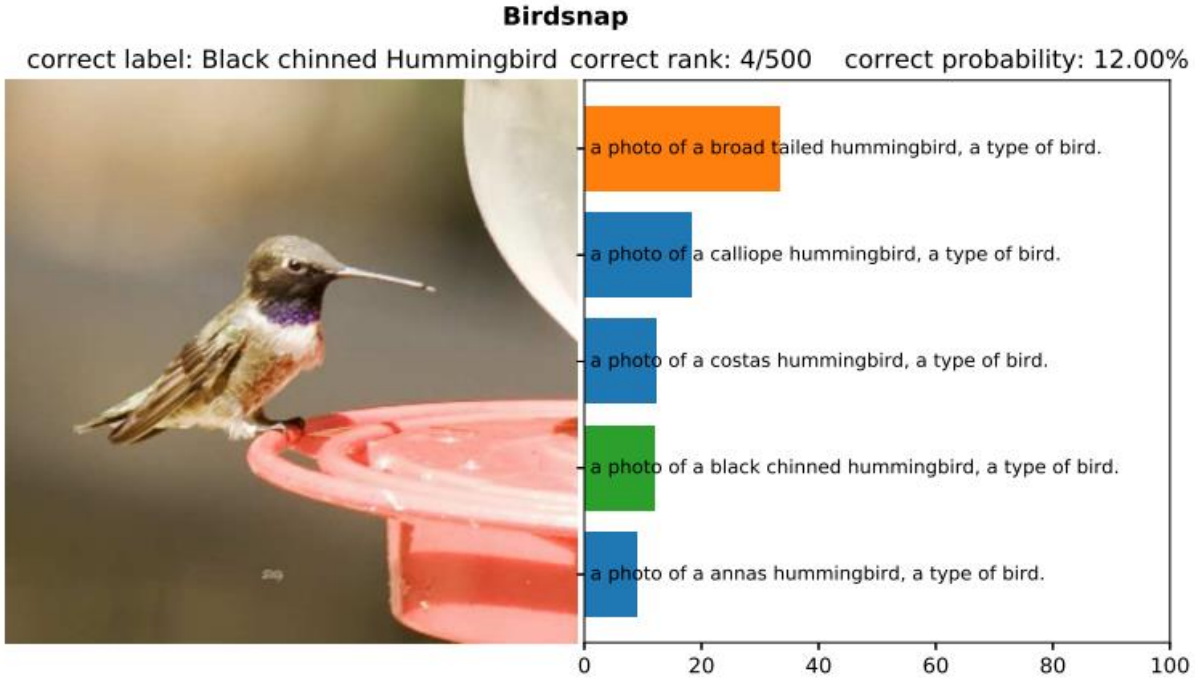


Approach of CLIP - Result

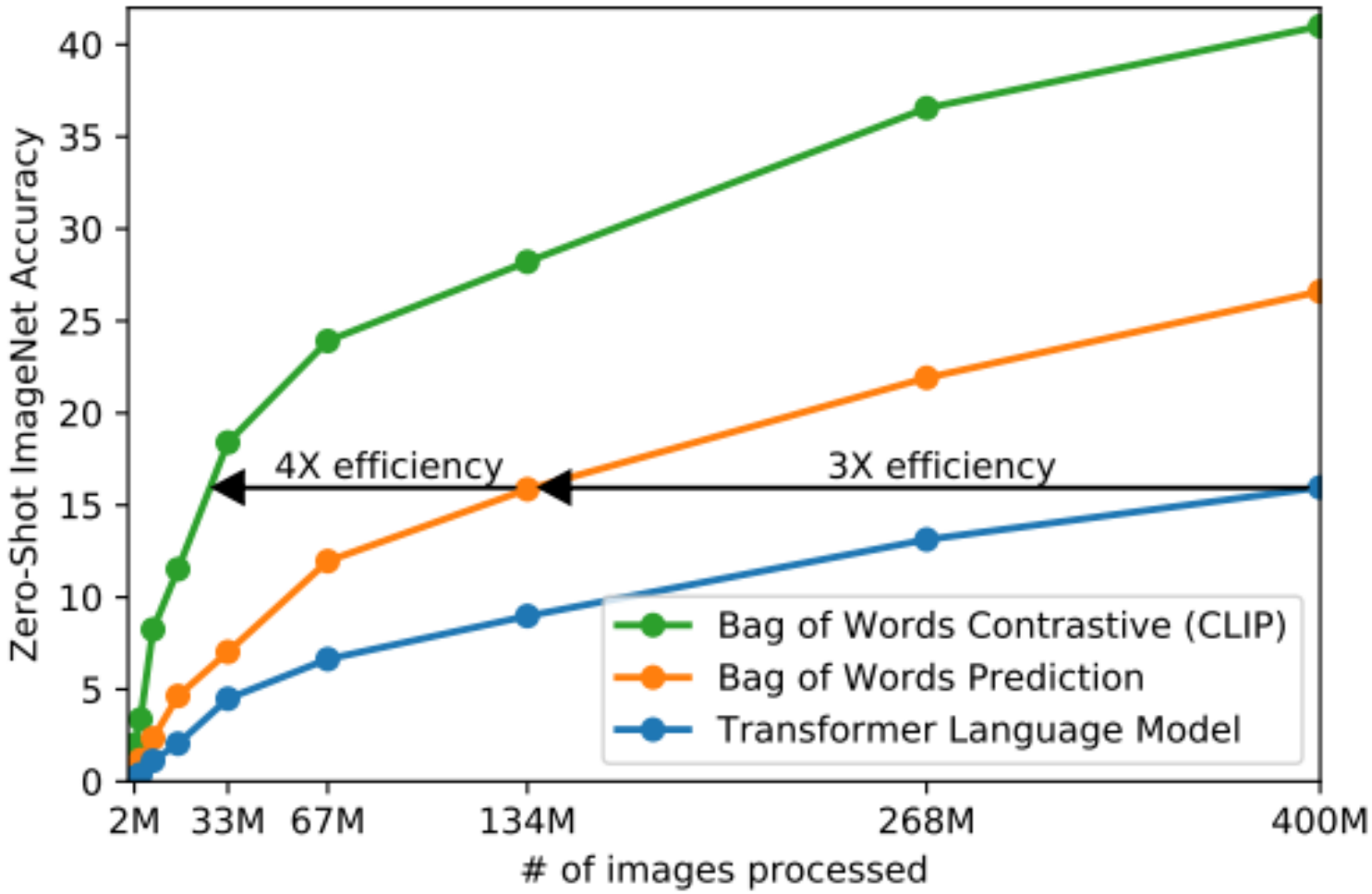


Green – Right option

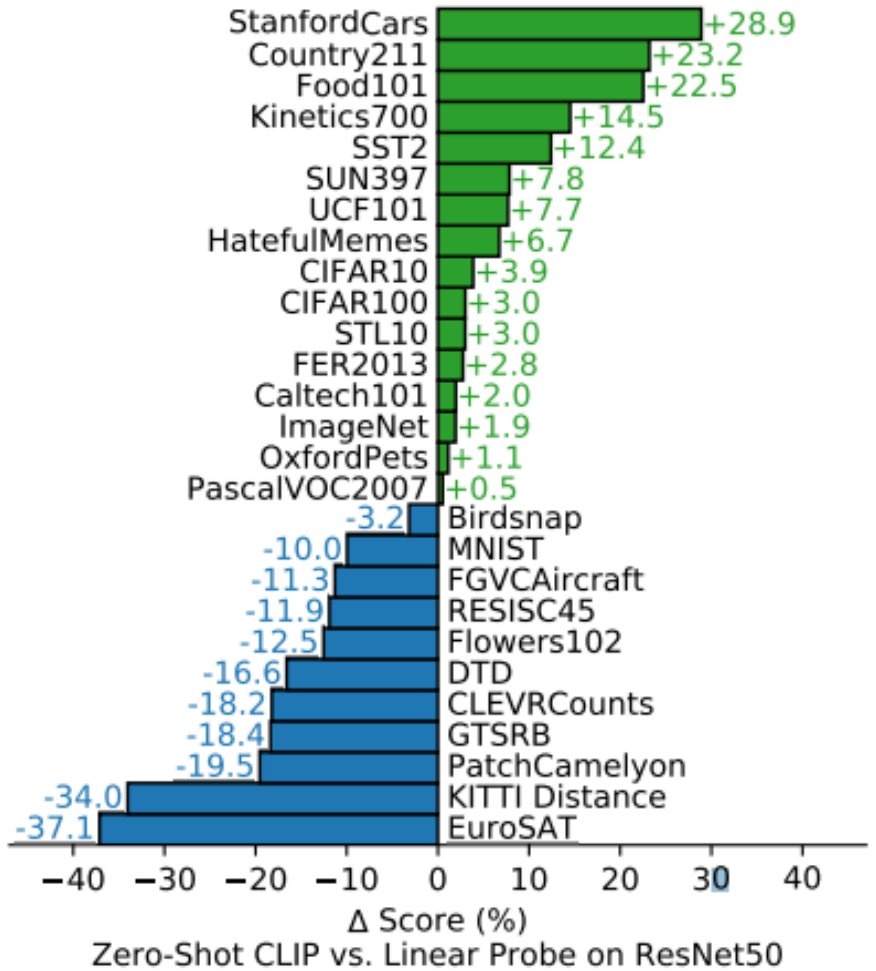
Orange – Picked option, however wrong



CLIP is much more efficient at zero-shot transfer than their image caption baseline.



ANALYSIS OF ZERO-SHOT CLIP PERFORMANCE

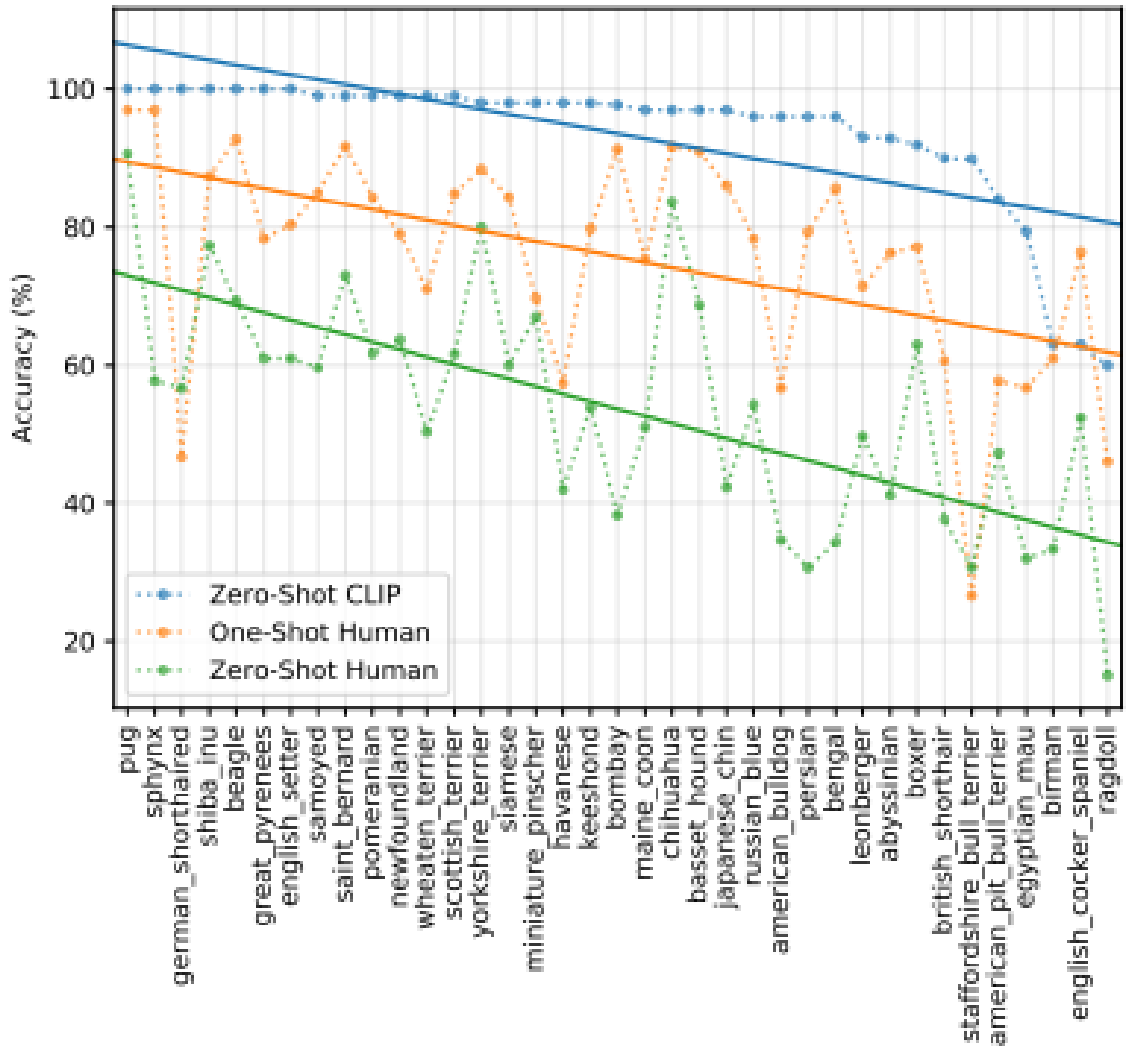


Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models.

Comparison to Human Performance

They had five different humans look at each of 3669 images in the test split of Oxford IIT Pets dataset and select which of the 37 cat or dog breeds best matched the image

	Accuracy	Majority Vote on Full Dataset	Accuracy on Guesses	Majority Vote Accuracy on Guesses
Zero-shot human	53.7	57.0	69.7	63.9
Zero-shot CLIP	**93.5**	**93.5**	**93.5**	**93.5**
One-shot human	75.7	80.3	78.5	81.2
Two-shot human	75.7	85.0	79.2	86.1



Limitations

- Analysis of CLIP's zero-shot performance found that, it is still quite weak on several kinds of tasks. When compared to task-specific models, the performance of CLIP is poor on several types of fine-grained classification such as differentiating models of cars, species of flowers, and variants of aircraft.
- CLIP also struggles with more abstract and systematic tasks such as counting the number of objects or classifying the distance to the nearest car in a photo.
- Although CLIP can flexibly generate zero-shot classifiers for a wide variety of tasks and datasets, CLIP is still limited to choosing from only those concepts in a given zero-shot classifier.
- On most of these datasets, the performance of this baseline is now well below the overall state of the art. They estimate that around 1000x increase in compute is required for zero-shot CLIP to reach SOTA performance. Which is infeasible with the current hardware.