PreCog Research Task

On -NLP and Safe AI
By- Agyeya Negi

Introduction:

State-of-the-art vision models rely on manually labeled datasets.

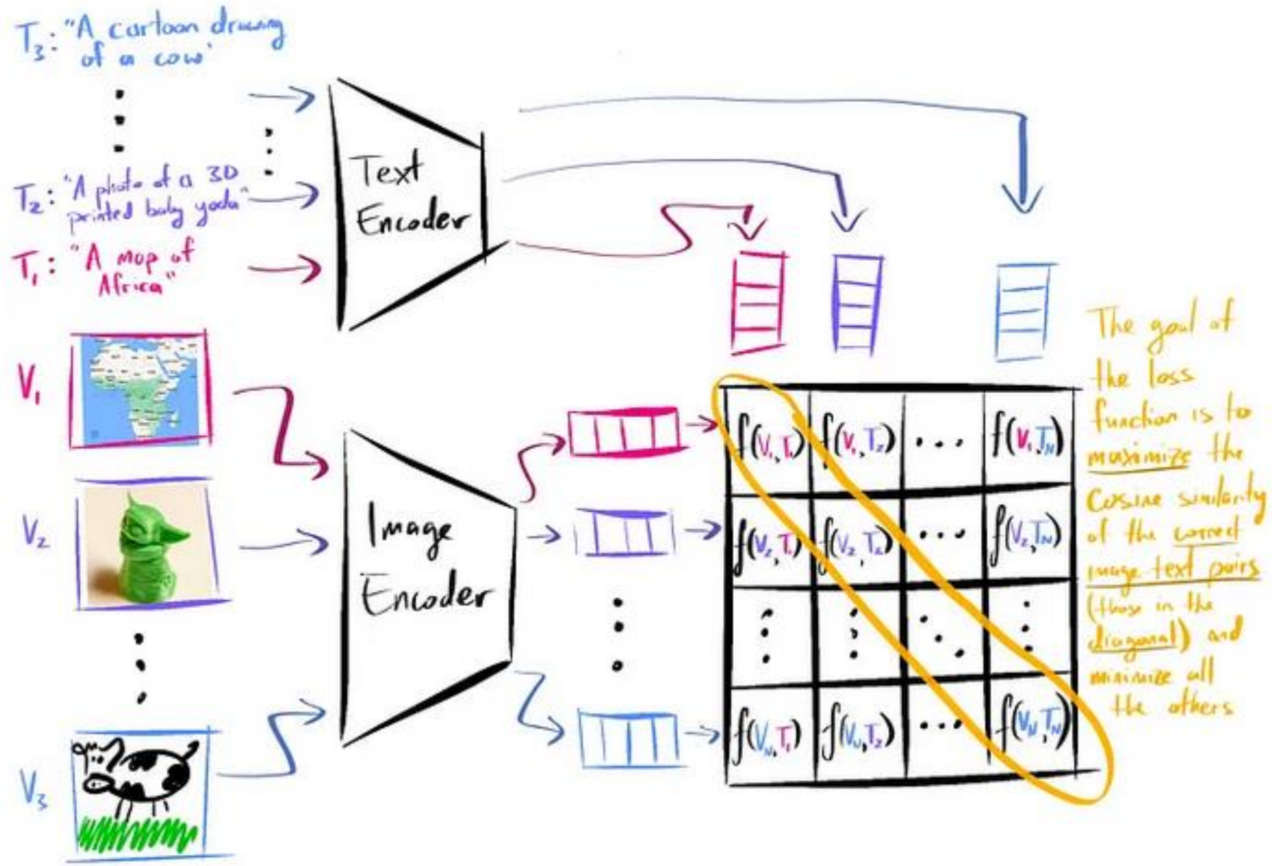
Manually labeling datasets is expensive and limits scalability.

CLIP (Contrastive Language-Image Pre-training) is proposed as an alternative.

CLIP learns from image-text pairs found on the internet.

Goal: Enable zero-shot transfer learning to unseen datasets without retraining.

Objective of the Paper


T3: "A cartoon drawing of a cow"
T2: "A photo of a 3D printed baby yoda"
T1: "A map of Africa"
Text Encoder
V1
V2
V3
Image Encoder
f(V1,T1) f(V1,T2) ... f(V1,TN)
f(V2,T1) f(V2,T2) ... f(V2,TN)
f(VN,T1) f(VN,T2) ... f(VN,TN)
The goal of the loss function is to maximize the cosine similarity of the correct image-text pairs (those in the diagonal) and minimize all the others

Overview of how CLIP works during training.

Train

- Train vision models using natural language supervision.

Address

- Address the challenge of requiring large labeled datasets for training.

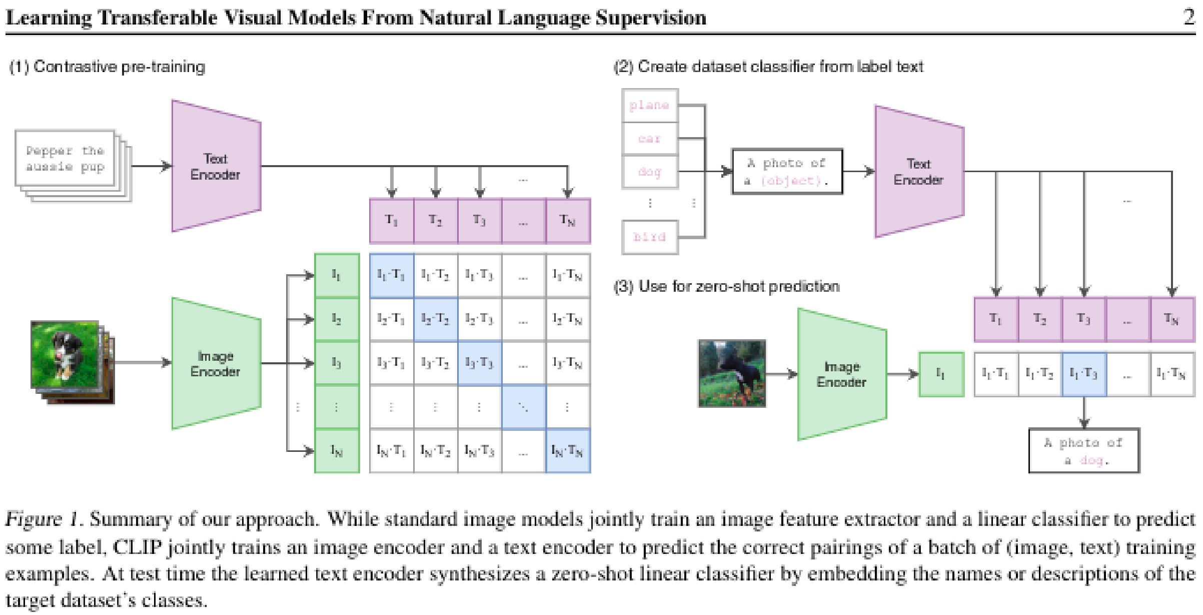
Reduce

- Reduce dependence on dataset-specific labels by leveraging text-image pairs.

Achieve

- Achieve strong generalization across various vision tasks.

Methodology – CLIP Architecture



Learning Transferable Visual Models From Natural Language Supervision 2

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T₁ T₂ T₃ ... T_N
I₁ I₂ I₃ ⋮ I_N
I₁·T₁ I₁·T₂ I₁·T₃ ... I₁·T_N
I₂·T₁ I₂·T₂ I₂·T₃ ... I₂·T_N
I₃·T₁ I₃·T₂ I₃·T₃ ... I₃·T_N
I_N·T₁ I_N·T₂ I_N·T₃ ... I_N·T_N
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I₁
I₁·T₁ I₁·T₂ I₁·T₃ ... I₁·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.

CLIP consists of:

Vision Encoder: A Vision Transformer (ViT) or ResNet model.

Text Encoder: A Transformer-based language model.

Contrastive Learning Approach:

Pairs of images and captions are used to create an embedding space.

The model learns which caption matches which image in a batch.

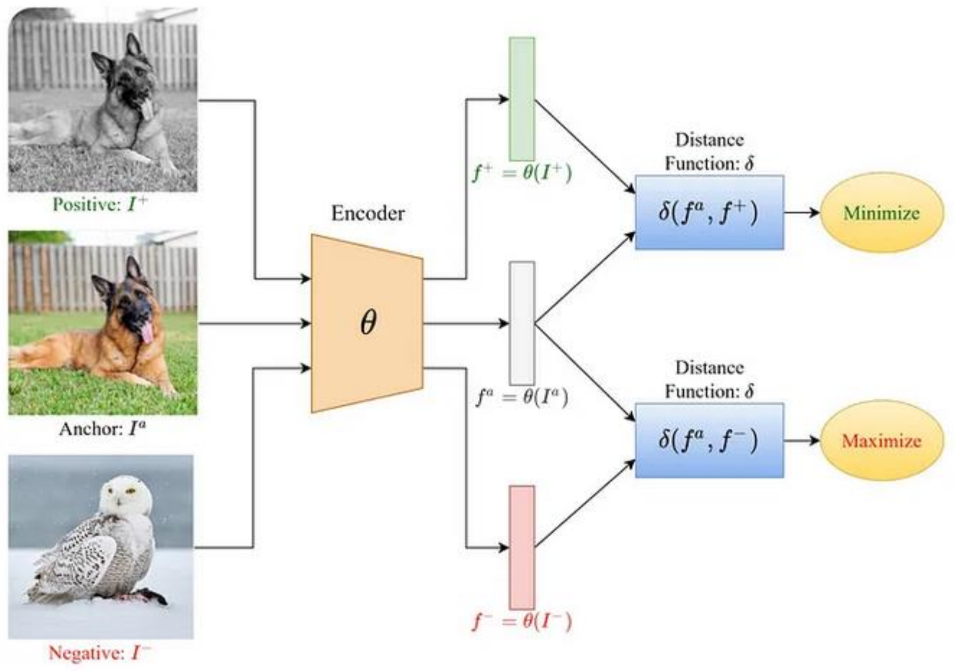
Instead of classifying objects directly, CLIP learns semantic relationships.

Dataset & Training

CLIP is trained on 400 million image-text pairs from the internet.

Uses a contrastive loss function to align text and images in a shared embedding space.

Zero-shot transfer: The model can classify images without task-specific fine-tuning.



A visualisation of **contrastive learning**. Source: https://www.v7labs.com/blog/contrastive-learning-guide

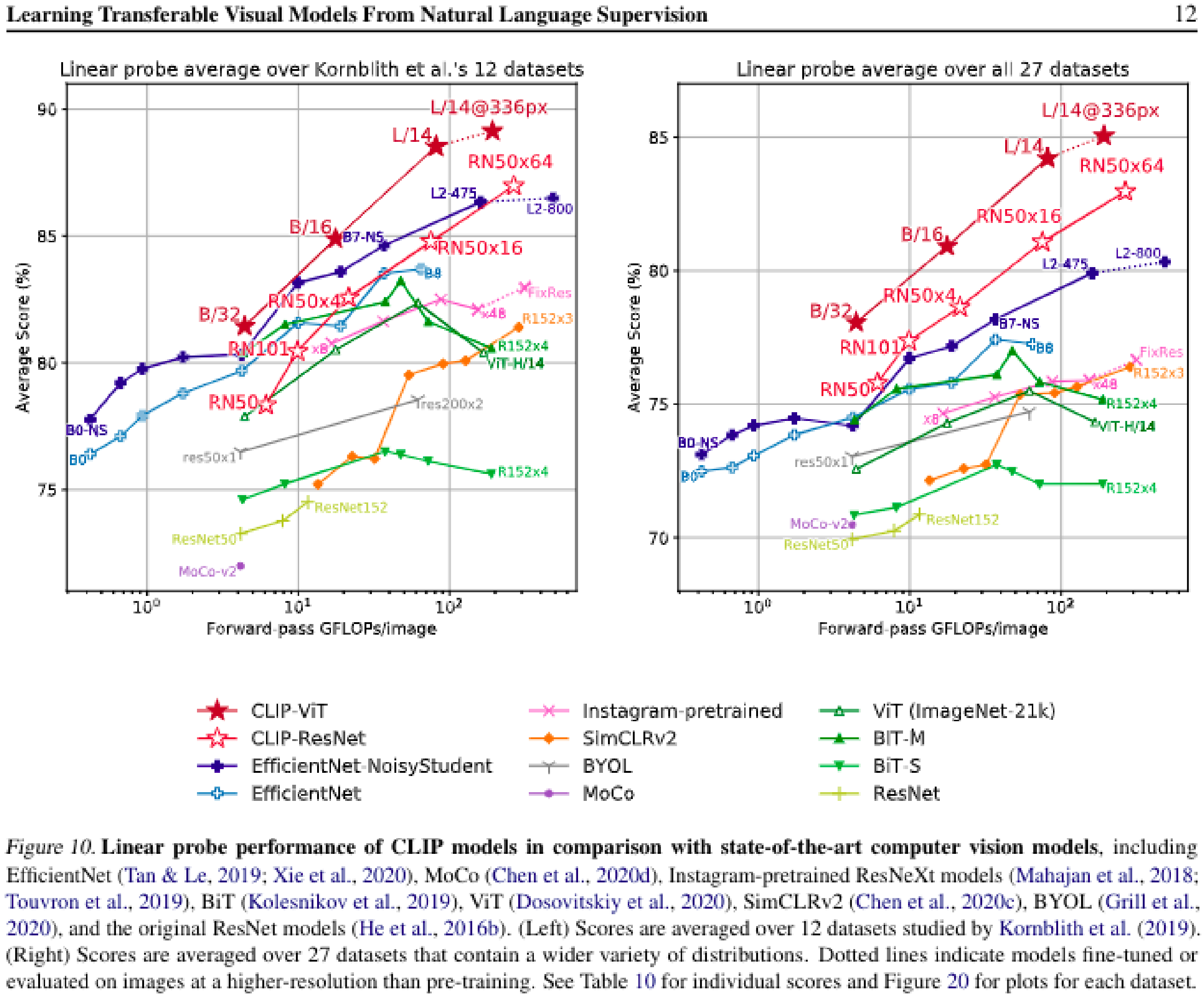
Key Findings & Results

Zero-shot learning capabilities: CLIP generalizes across many vision benchmarks.

Achieves ResNet-50 level accuracy on ImageNet without using ImageNet labels.

CLIP outperforms traditional models in diverse datasets without additional training.

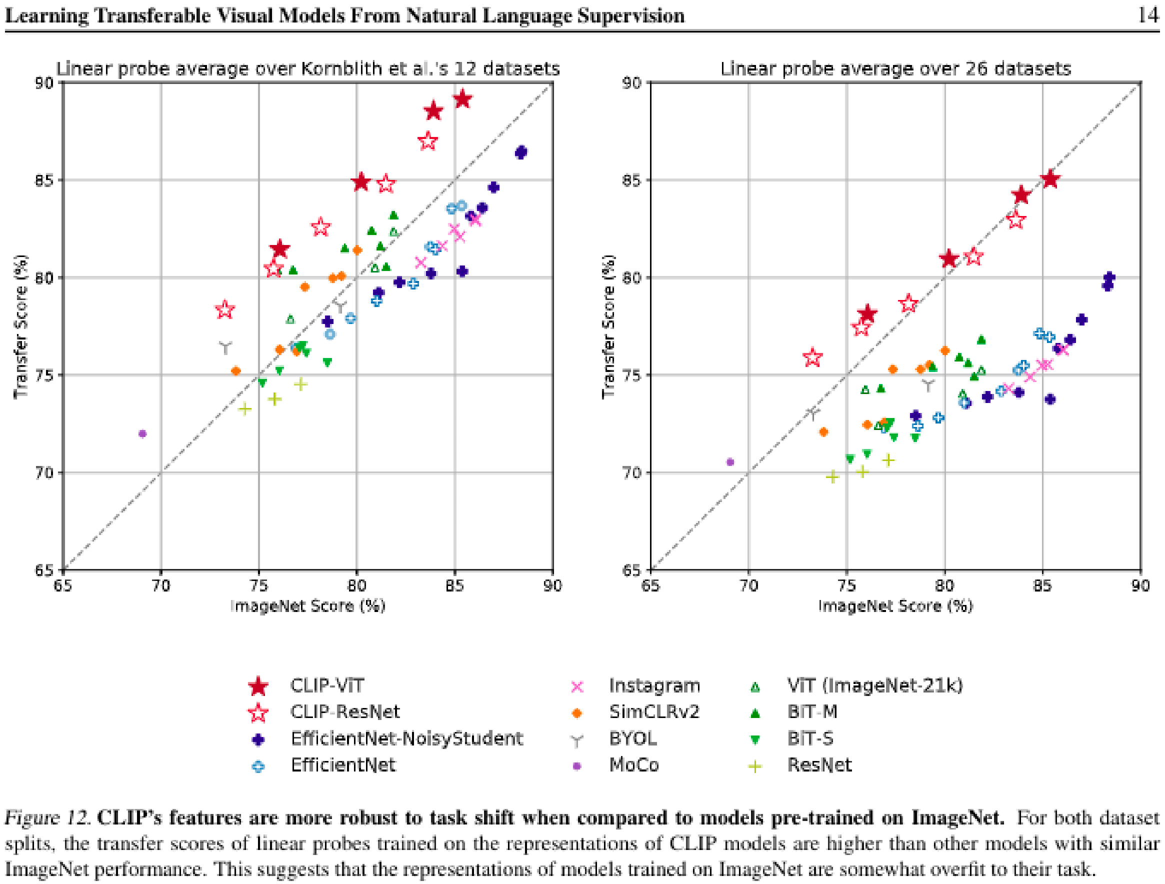
Demonstrates robustness to natural distribution shifts.



Learning Transferable Visual Models From Natural Language Supervision 12

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.

Strengths of the Paper



Learning Transferable Visual Models From Natural Language Supervision 14

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.

Innovative approach: Uses natural language as a scalable form of supervision.

Comprehensive analysis: Evaluates on over 30 vision benchmarks.

Impactful research: Enables zero-shot learning across multiple vision tasks.

Efficient training: Uses contrastive loss for better feature learning.

Weaknesses of the CLIP model

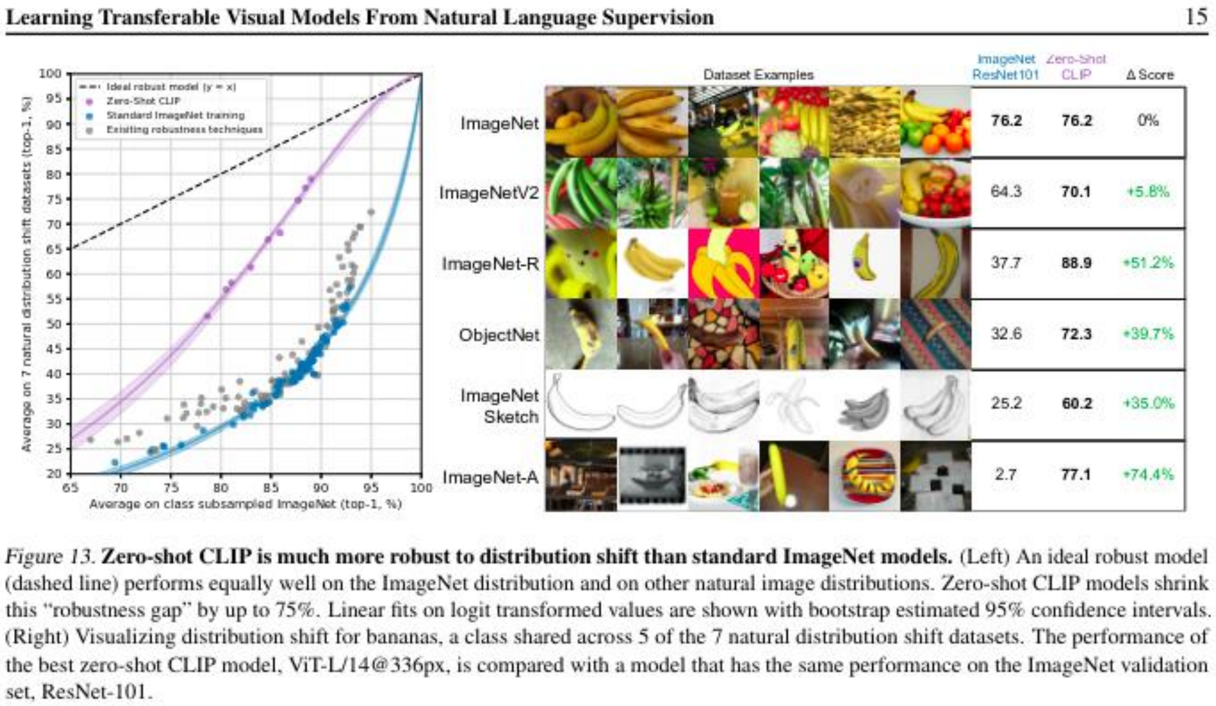
Data dependency: Requires a large-scale dataset with diverse image-text pairs.

Limited scope: Focused mainly on image classification, less exploration of video or multimodal applications.

Computational cost: Large-scale pretraining requires significant computational resources.

Unrealistic assumptions: Not all domains have rich image-text data for training.

Generalisability of Techniques



Learning Transferable Visual Models From Natural Language Supervision 15

Ideal robust model (y = x)
Zero-Shot CLIP
Standard ImageNet training
Existing robustness techniques
Average on 7 natural distribution shift datasets (top-1, %)
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
65
70
75
80
85
90
95
100
Average on class subsampled ImageNet (top-1, %)

	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%

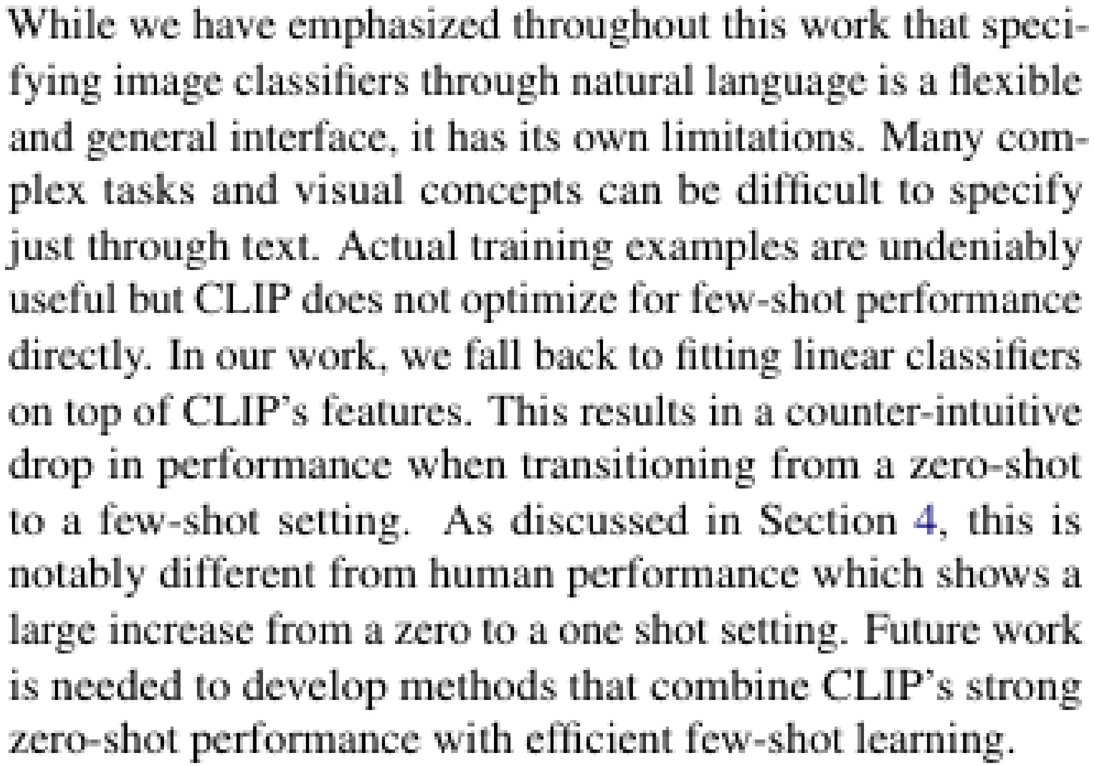
Figure 13. **Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models.** (Left) An ideal robust model (dashed line) performs equally well on the ImageNet distribution and on other natural image distributions. Zero-shot CLIP models shrink this "robustness gap" by up to 75%.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

Applicability: CLIP can be used for OCR, action recognition, geo-localization, and fine-grained object classification.

Scalability: Works well with large datasets, but performance on low-resource tasks is uncertain.

Transferability: Can be adapted to different industries like autonomous driving, healthcare, and e-commerce.

Limitations



While we have emphasized throughout this work that specifying image classifiers through natural language is a flexible and general interface, it has its own limitations. Many complex tasks and visual concepts can be difficult to specify just through text. Actual training examples are undeniably useful but CLIP does not optimize for few-shot performance directly. In our work, we fall back to fitting linear classifiers on top of CLIP's features. This results in a counter-intuitive drop in performance when transitioning from a zero-shot to a few-shot setting. As discussed in Section 4, this is notably different from human performance which shows a large increase from a zero to a one shot setting. Future work is needed to develop methods that combine CLIP's strong zero-shot performance with efficient few-shot learning.

High computational cost: Training requires hundreds of GPUs.

Bias in dataset: Internet images may contain unwanted biases.

Not perfect for all vision tasks: Struggles with some specialized datasets (e.g., medical imaging).

Future Research Directions

Improving efficiency: Can CLIP be made lighter and more resource-efficient?

Expanding multimodal learning: Can CLIP be extended to videos, audio, and 3D models?

Better dataset curation: How to reduce biases in the training data?

Improving zero-shot performance: Can CLIP be enhanced with semi-supervised learning?

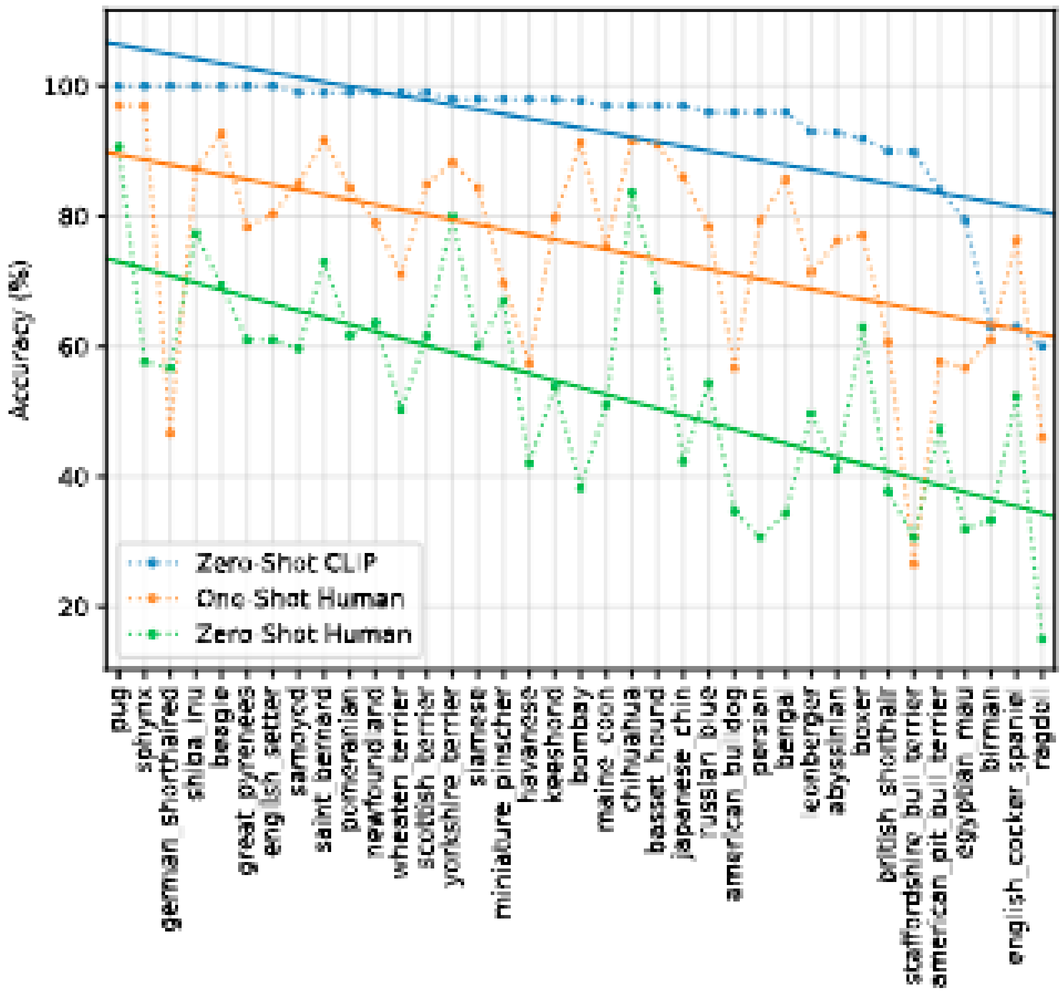


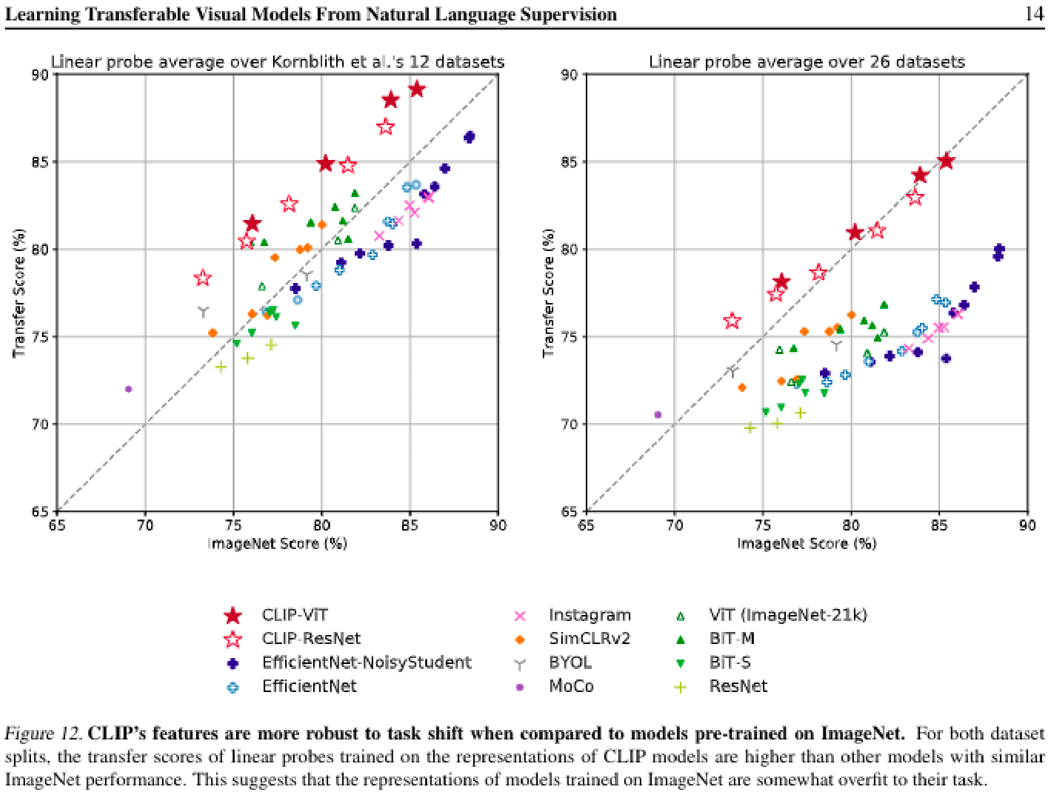
Figure 16. The hardest problems for CLIP also tend to be the hardest problems for humans. Here we rank image categories by difficulty for CLIP as measured as probability of the correct label.

Methodological Insights

Contrastive Learning is key: Learning to match images with texts is more powerful than traditional classification.

Scalability matters: Larger models and datasets improve zero-shot performance.

Evaluation on diverse datasets is crucial: Standard benchmarks may not fully test model generalization.



Learning Transferable Visual Models From Natural Language Supervision 14

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.

Conclusion

CLIP introduces a new paradigm in vision-language learning.

Enables zero-shot learning, reducing the need for dataset-specific training.

Challenges traditional supervised learning models with a new contrastive approach.

Future research should explore multimodal applications and efficiency improvements.

References

- https://medium.com/@paluchasz/understanding-openais-clip-model-6b52bade3fa3
- https://medium.com/one-minute-machine-learning/clip-paper-explained-easily-in-3-levels-of-detail-61959814ad13
- https://arxiv.org/abs/2103.00020
- https://openai.com/index/clip/