Learning Transferable Visual Models From Natural Language Supervision

Seminar-Deep Learning

Abdullah Amawi
Supervised by Dr. Timo Lüddecke

May 24, 2022

University of Göttingen

Authors & Affiliation

- Authors: Alec Radford, JongWook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever.
- Institution: OpenAI.

Table of contents

Introduction

Motivation

- SOTA computer vision systems are trained to predict a fixed set of predetermined object categories.
- This restricted form of supervision limits generality and usability. Additional labeled data is needed.
- Labeling takes time and effort.
- Contrastive learning is the answer to this.
- CLIP can avoid training.

Contrastive learning-Intro

Figure 1: Machine Learning technique-Contrastive learning

1

[1] https://towardsdatascience.com/understanding-contrastive-learning-d5b19fd96607

Contrastive learning-Data Augmentation

Figure 2: Contrastive learning data augmentation

[2]

[2] https://towardsdatascience.com/understanding-contrastive-learning-d5b19fd96607

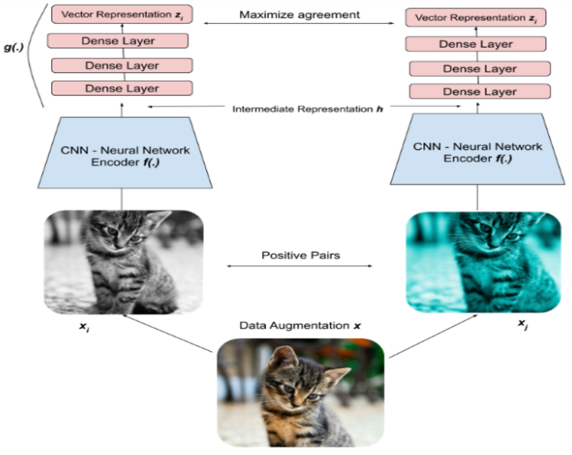

Figure 3: Contrastive learning SimCLRv2 framework

3

[3] https://towardsdatascience.com/understanding-contrastive-learning-d5b19fd96607

Contrastive learning example.

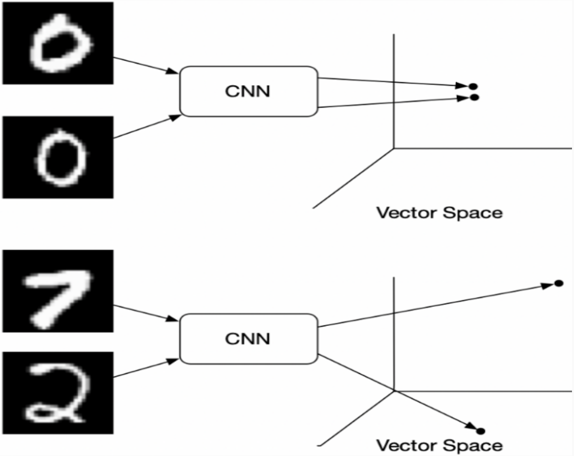

Figure 4: Contrastive learning MNIST example

4

[4] https://towardsdatascience.com/contrastive-loss-explaned-159f2d4a87ec

Methods.

Other Methods

- YFCC100M(Joulin et al.)
    - 100 Million images. Varying quality.
    - Many images has automatic generated file names(Numeric).
    - Only 15 Million after filtering images with natural language titles.
- Mahajan et al.
    - 3.5 Billion Instagram images.
    - Usage of hashtags for weakly supervised pre-training.
    - can be noisy due to the use of hashtags.

[5]

[5] (Joulin et al.), (Mahajan et al.)

- Mahajan et al. main takeaways.

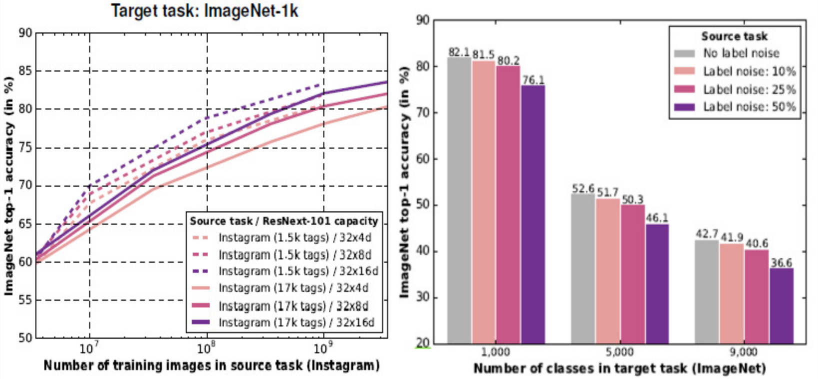

Figure 5: Pros and Cons(Mahajan et al.)

6

[6] (Mahajan et al.)

Related works.

Related Areas	Representative Article	CLIP Advantage
Natural language supervision	YFCC100M (Joulin et al.) VirTex(Desai and Johnson) ICMLM(Sariyildiz et al.) ConVIRT(Zhang et al.)	Better Efficiency(vs YFCC100M). Larger scale(vs VirTex, ICMLM, & ConVIRT). Simplified in comparison to ConVIRT.
Zero-Shot Transfer	Visual N-Grams(Li et al.)	Improves upon, better performance.
Broad Evaluation and Robustness	VTAB(Zhang et al.) ImageNet (Taori et al.)	Adapts VTAB evaluation to counter bias. More robust vs ImageNet. Matches RestNet-50 on Zero-Shot.

Table 1: Related works comparison.

7

[7] (Radford et al.)

CLIP(Contrastive Language-Image Pre-Training) Details

- CLIP is pre-trained on 400M image-text pairs from the internet.
- Batch size of 32,768.
- 32 epochs over the dataset.

- Uses ResNet-based or ViT-based image encoder.
- Uses Transformer-based text encoder.

[8]

[8] (Radford et al.)

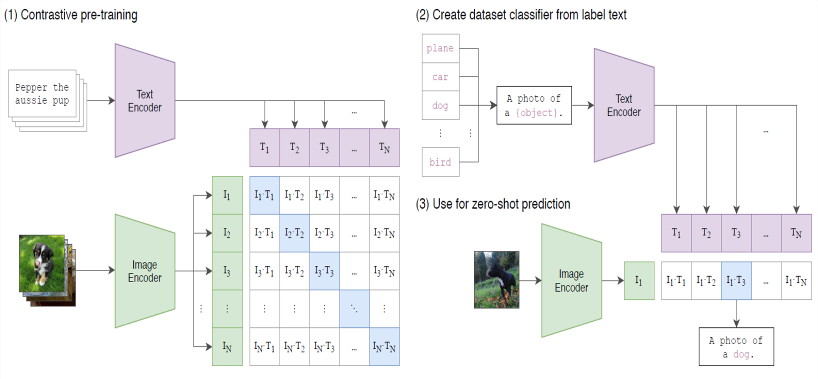

Figure 6: CLIP approach

9

9 (Radford et al.)

Experiments

Zero-Shot Transfer & Prompt engineering

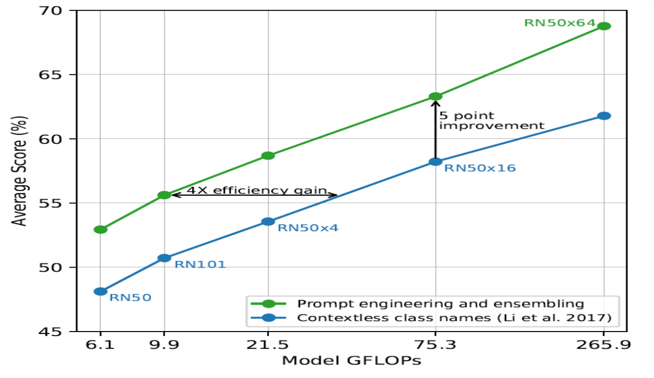

Figure 7: Prompt engineering & Zero-shot performance

10

10 (Radford et al.)

Zero-Shot Transfer CLIP vs. ResNet50

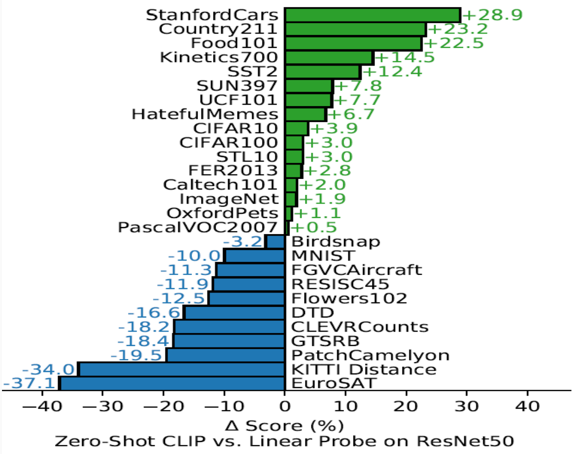

Figure 8: Zero-shot test on 27 datasets.

11

11 (Radford et al.)

Zero-Shot Transfer CLIP vs. Linear probes

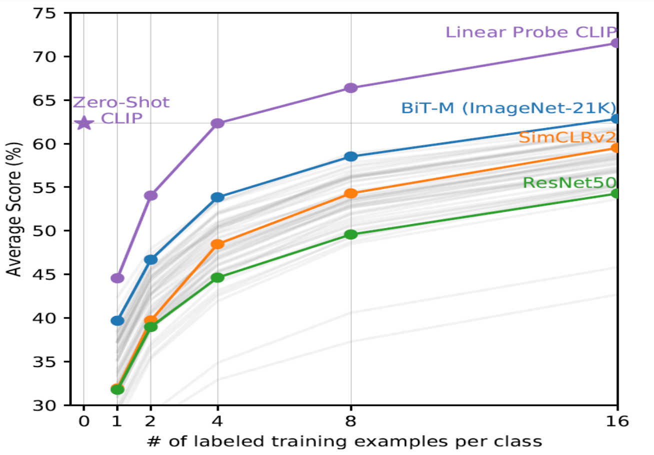

Figure 9: Zero-shot test on 27 datasets.

12

12 (Radford et al.)

Zero-Shot Efficiency

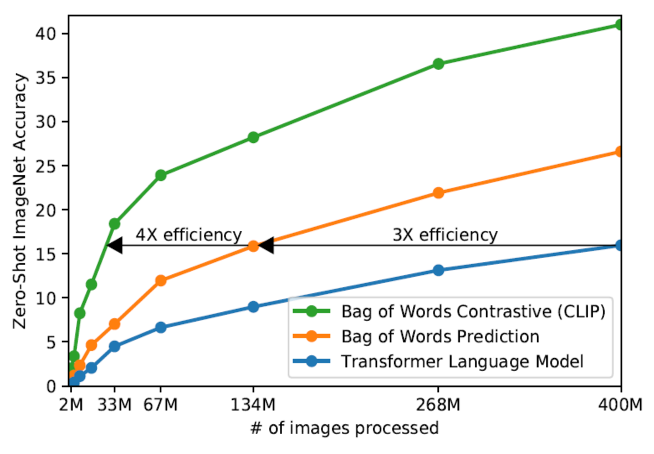

Figure 10: Zero-shot efficiency CLIP vs Joulin et al.

13

13 (Radford et al.) (Joulin et al.)

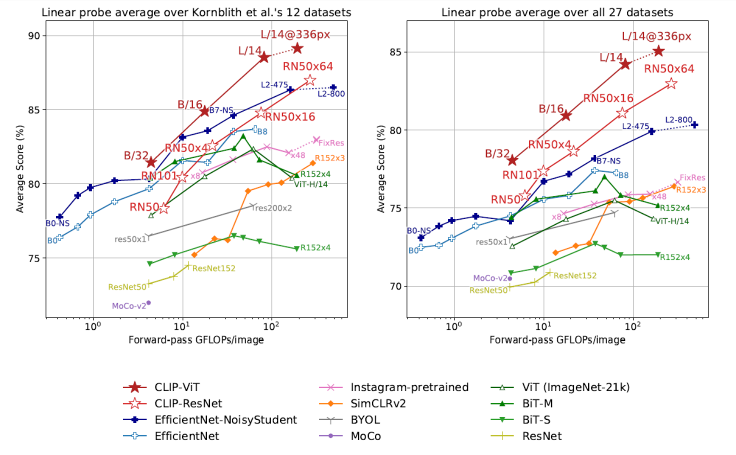

Figure 11: Linear probe performance vs SOTA vision models

14

14 (Radford et al.)

Representation Learning-2

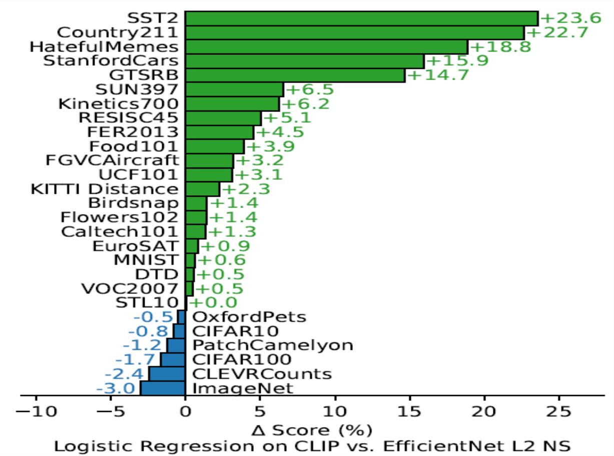

Figure 12: LR-CLIP vs EfficientNet L2 NS

15

[15] (Radford et al.)

Robustness to natural distribution shift-1

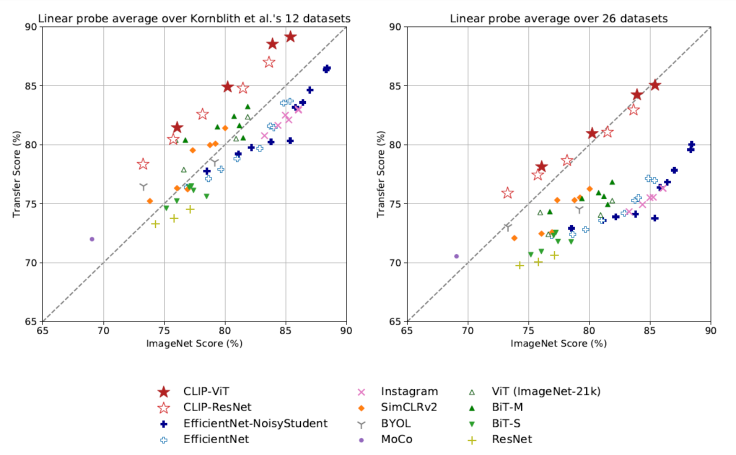

Figure 13: CLIP vs ImageNet on task shift

16 (Radford et al.)

Robustness to natural distribution shift-2

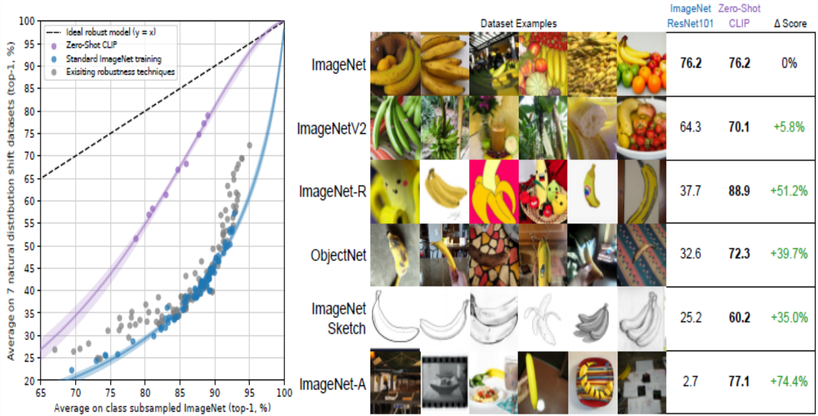

Figure 14: CLIP vs ImageNet on distribution shift

17

17 (Radford et al.)

Conclusion & Opinion

Conclusion.

- CLIP is able to match and outperform ResNet models on zero-shot.
- CLIP zero-shot models are more robust than supervised ImageNet models.
- Weak on some tasks(Ex MNIST, Satellite images datasets).
- CLIP shows social biases
- CLIP demonstrates a lot of potential for future use. But still doesn't match SOTA in many uses.

Thank you!

Questions?

Additional resources

Terms

- CLIP: Contrastive Language-Image Pre-Training.
- ViT: Vision transformers.