Learning Transferable Visual Models from Natural Language Supervision

Anurat Bhattacharya Abhinandan De

IIT Kharagpur

Paper Presentation
March 29, 2022

Presentation Overview

Abstract

- Current SOTA systems predict fixed categories.
- Our model CLIP learns from captions of images.
- Zero shot transfer after pretraining.
- Applied to $>$ 30 CV datasets spanning common tasks.

Fun facts

- CLIP matches accuracy of original ResNet-50 on ImageNet without using any of the 1.28 M examples.
- It was developed in conjunction with DALL-E by Open-AI to evaluate the latter's performance.

Diagram

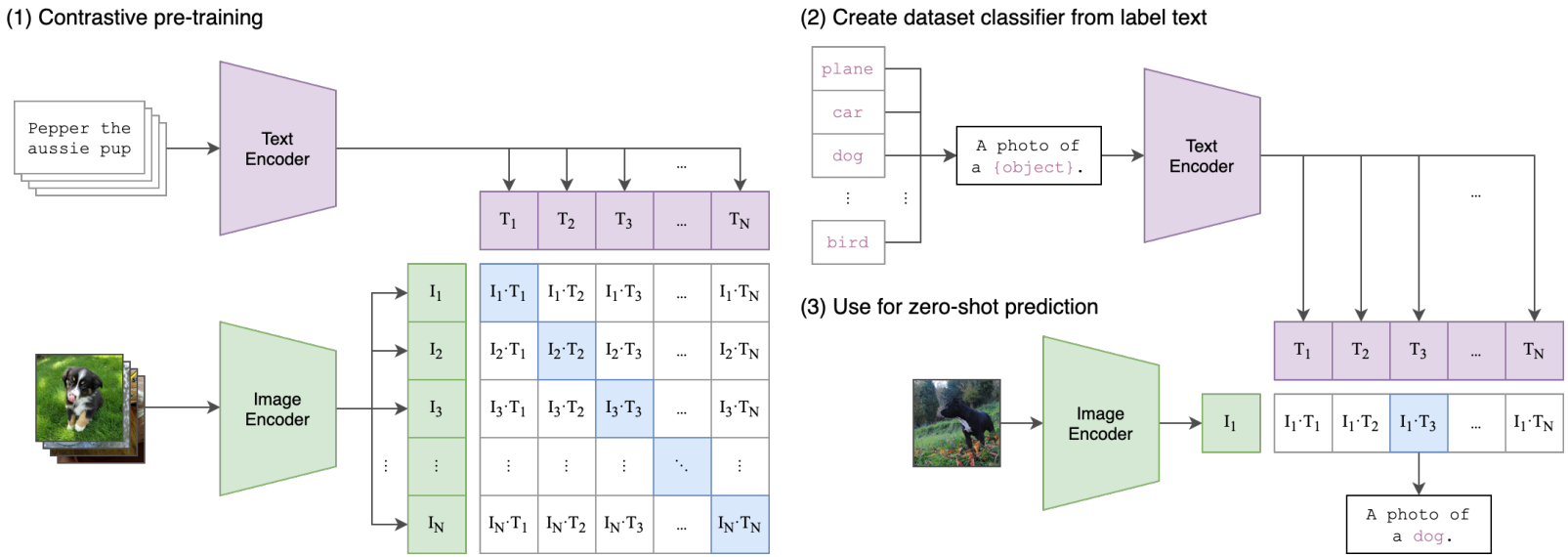
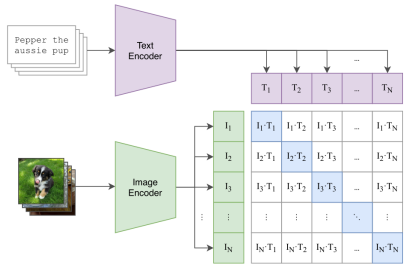


Figure: The CLIP model

Architecture

- Text encoder is a Transformer[4] with modifications[3].
- 63 M parameters, 12-layers, 512-wide model.
- Image encoder had 2 possible architectures: ResNet[2], ViT[1].
- 5 ResNets and 3 ViT models were explored



Pretraining

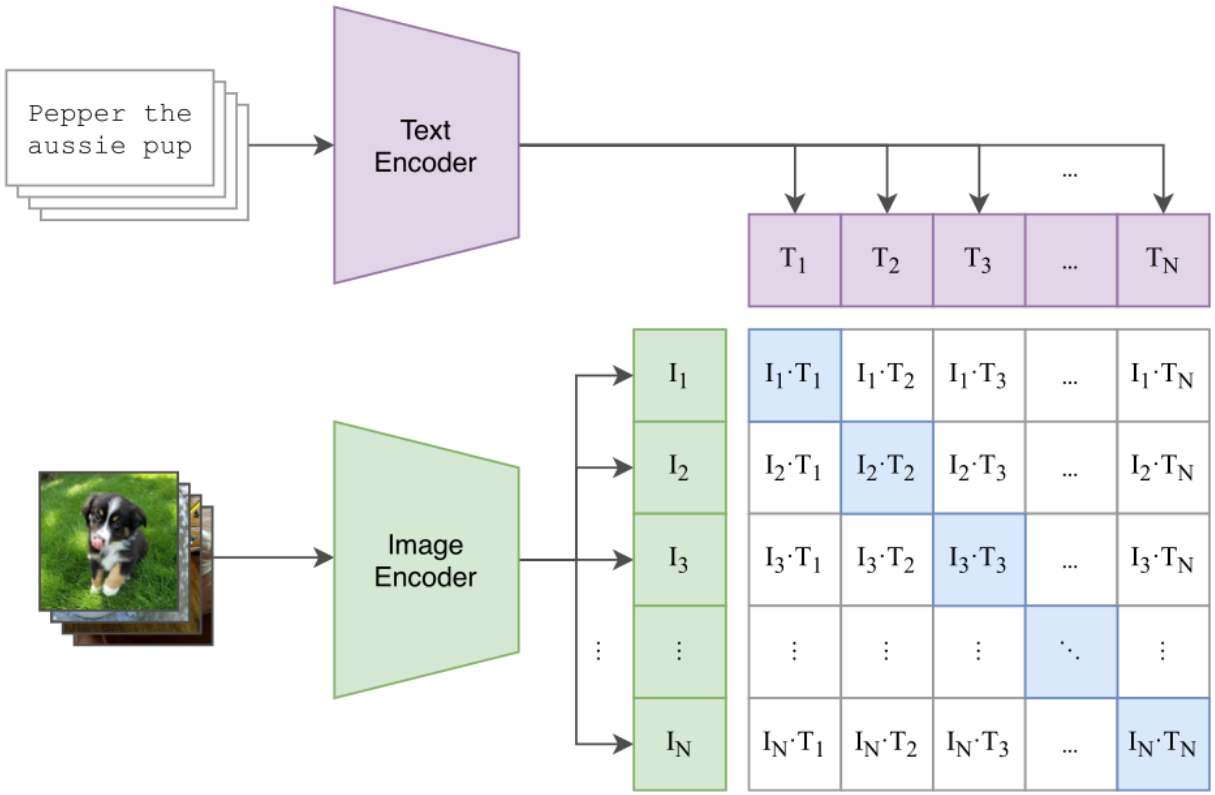


Figure: Contrastive Pretraining.

Pretraining

- Changing the caption from just a number to text
- Enabled by large amounts of publicly available data of this form
- Dataset of 400 M (img,text) pairs and trained using a simplified version of ConVIRT
- Scalability judged by training 8 models spanning 2 orders of magnitude

Testing

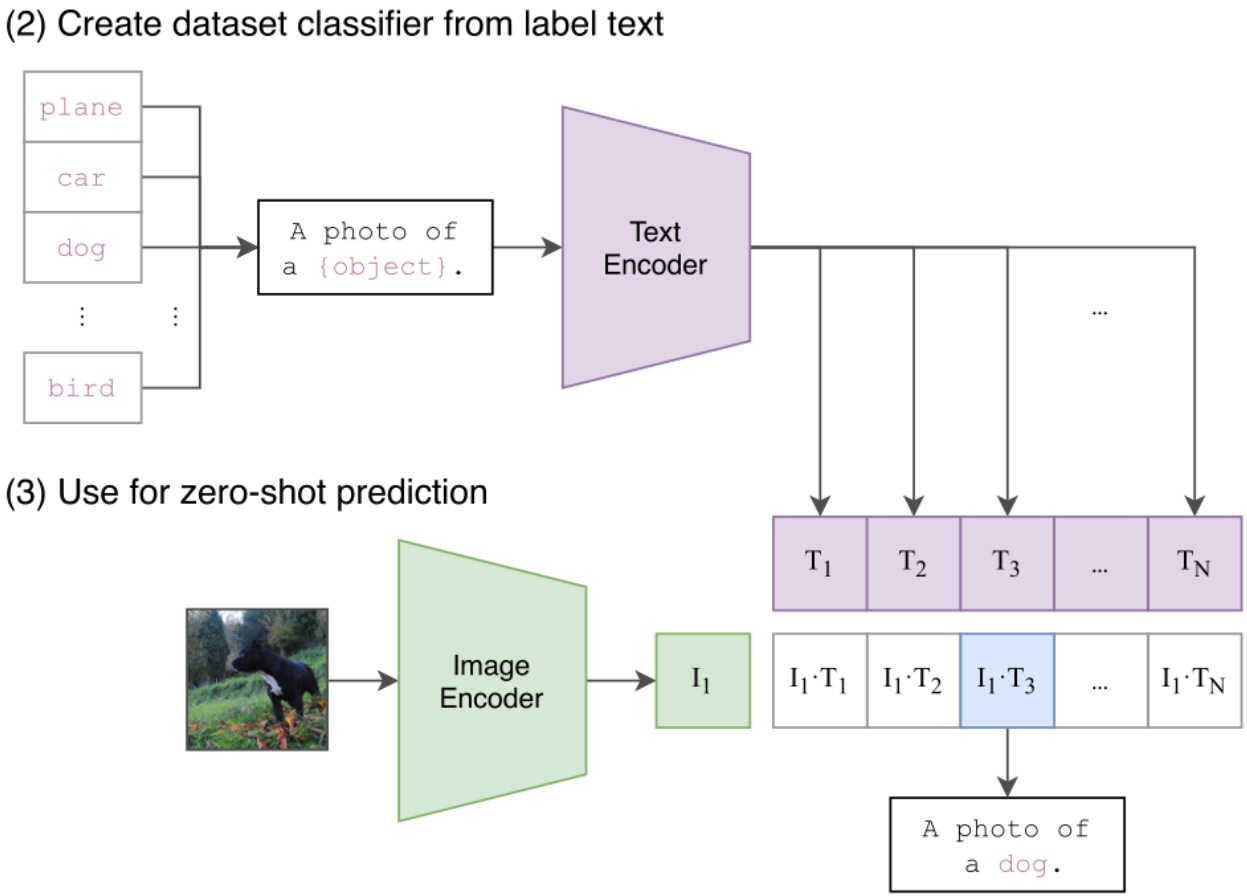


Figure: Testing phase

Zero Shot Transfer

- Comparison with Visual N Grams on aYahoo, ImageNet and SUN
- Comparison with a fully supervised linear classifier fitted on ResNet-50 features on 27 datasets
- A look at where Zero shot CLIP underperforms

Zero Shot Transfer

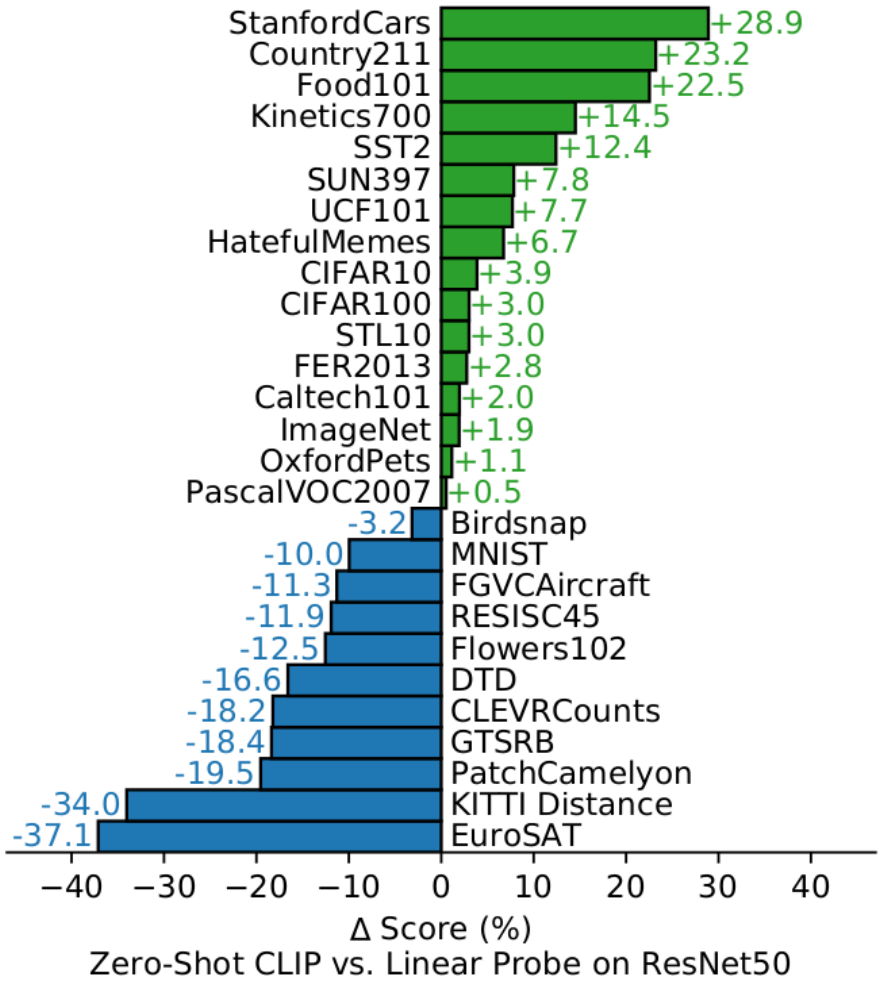


Figure: Evaluation on Different Datasets compared to fully supervised linear classifier fitted on ResNet-50 features.

Few Shot Transfer

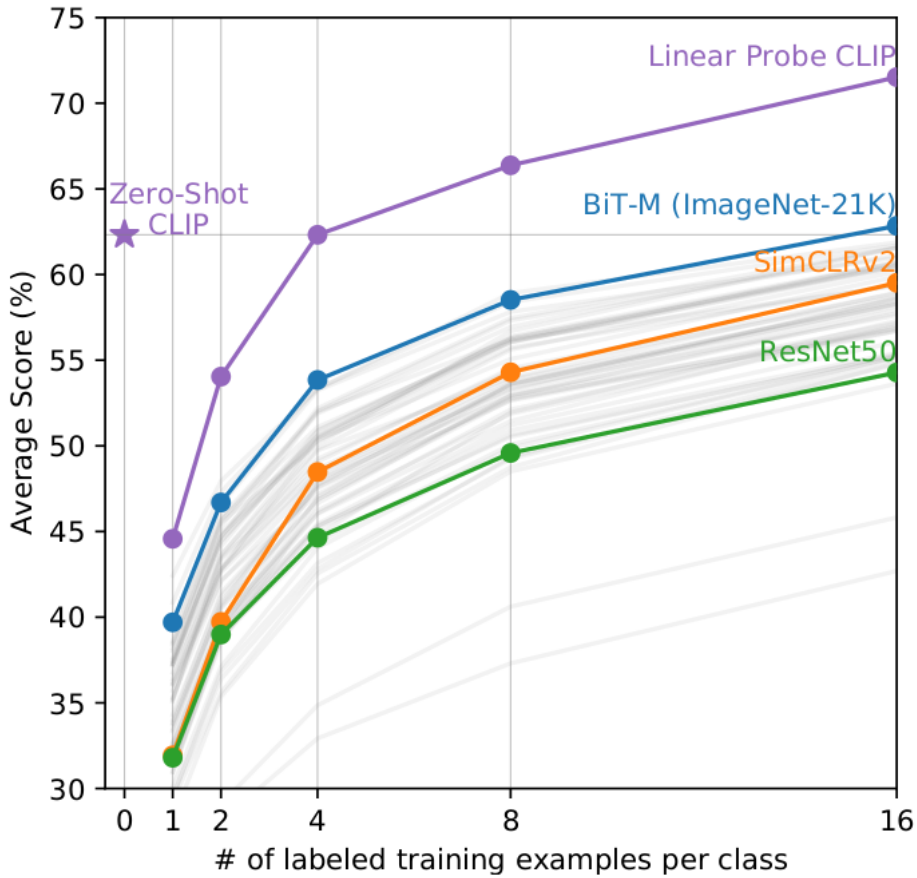


Figure: Few Shot Tranfer Comparison with different models.

Few Shot Transfer

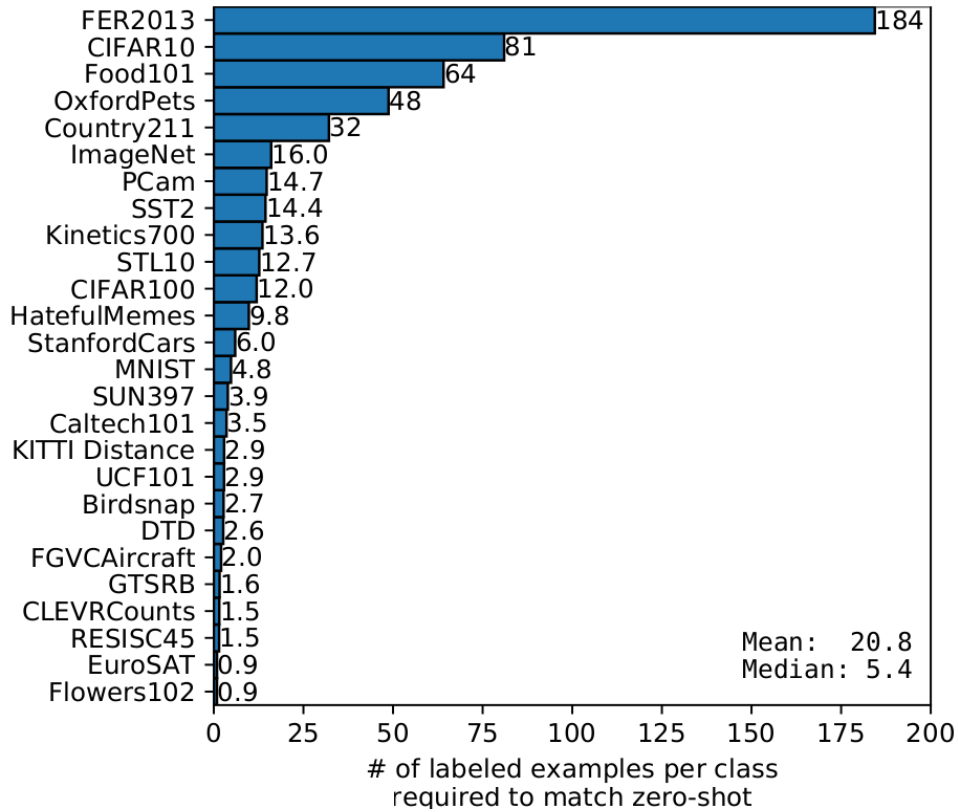


Figure: No. of labelled examples to match zero shot.

Pros

- Wide Range of capabilities
- Significant Benefit for tasks that have low data
- Suitable for tasks like Image Retrieval/Search from a Database

Cons

- Need to improve scalability
- Poor performance on fine grained classification tasks
- Performs poorly on tasks like object detection and semantic segmentation

References

[1] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[2] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[3] Tong He et al. “Bag of tricks for image classification with convolutional neural networks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 558–567.

[4] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

Thank You!

Questions? Comments?