Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1] Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

Image Manipulation/Generation with CLIP
With
Only 1 / 0 Image & 1 Text

marriage in the mountains

BigSleep

StyleCLIP

CLIPDraw

VQ-GAN+CLIP

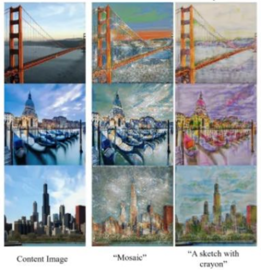
StyleGAN-NADA

Aliens destroying NYC skyline with lasers. #pixelart

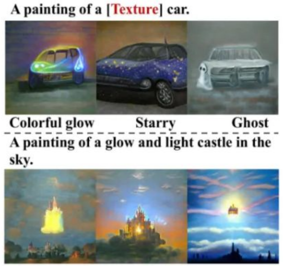
Pixary

Cheese Cake

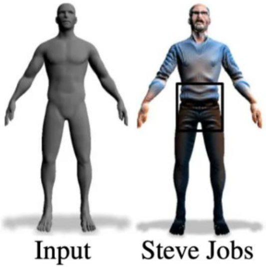
myStyleTransferCLIP



CLIPstyler



FuseDream



Text2Mesh

Motivation

1. Concurent CV system are trained to predict a fixed set of predetermined object categories (image classification)
2. Learning directly from raw text about images provides a broader source of supervision
3. NLP can solve above problems(BERT, GPT)

Approach

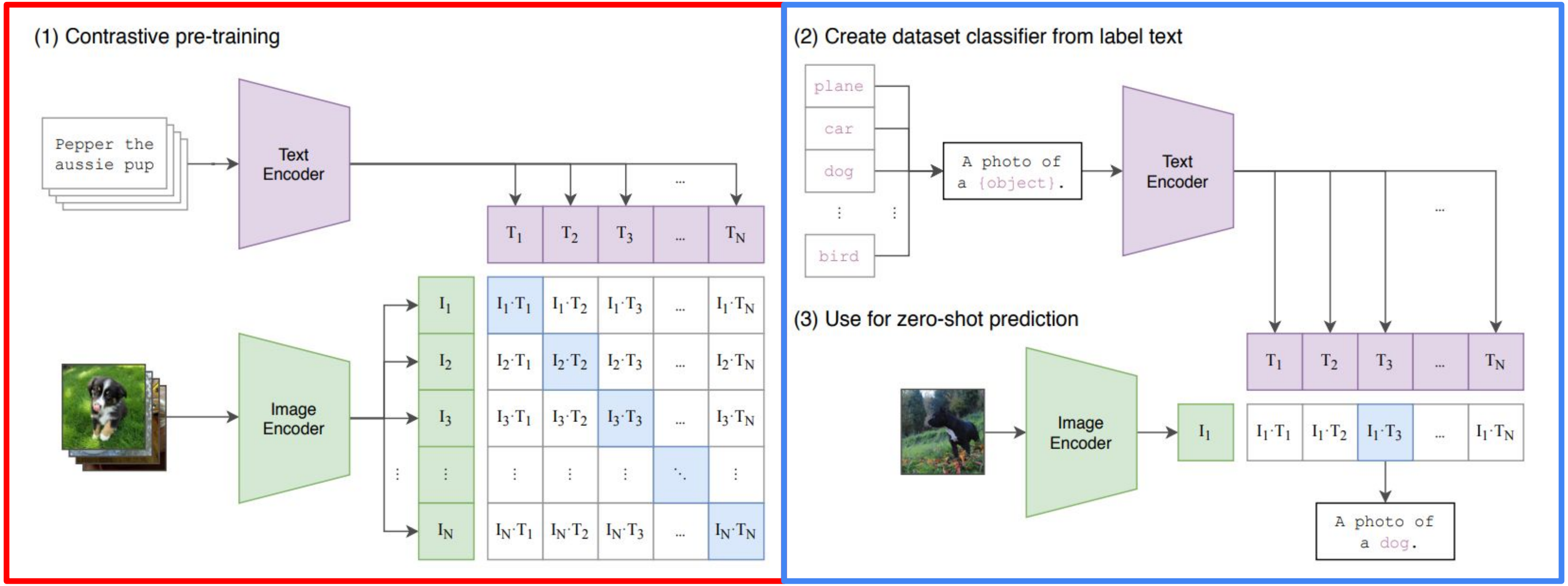


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the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

Approach

1. WebImageText : Large dataset with 400 million (image,text) pairs
    a. existing datasets, such as MS-COCO and Visual Genom are too small
    b. large dataset is necessary to capture the full rank if visual concepts and textual descriptions that exist in the real world
2. Efficient pre-trained method
    a. training model on such dataset would be computationally expensive
    b. masked language model (MLM), where a subset of tokens in an input sequence is masked and the model is trained to predict them based on the remaining token
    c. image masking, where some of the input tokens corresponding to image regions rather than text

Approach

3. Natural Language Supervision
    a. easier to scale natural language supervision and does not require annotations to be in a classic “machine learning compatible format”
    b. can learn passively from the supervision contained in the vast amount of text on the internet
4. Choosing and scaling a model
    a. image encoder : ResNet-50, ViT
    b. text encoder : Transformer

Experiments

1. Zero-shot transfer
    a. ability of a model perform well on a task it has not been explicitly trained on
    b. use zero-shot transfer as an evaluation metric
2. CLIP for zero-shot transfer
    a. CLIP: pre-trained to predict if an image and a text snippet are paired together in its dataset
        i. compute the feature embedding of the the image and feature embedding of the set of possible text
        ii. product them -> similarity score for each (image,text) pair

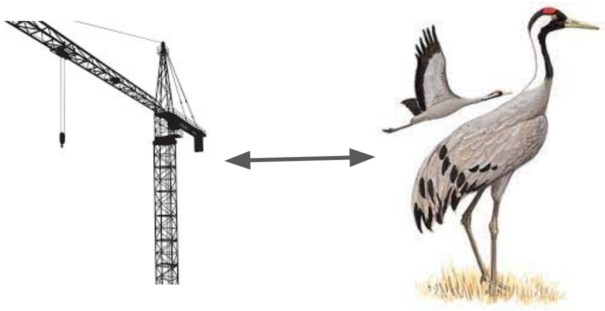
Experiments

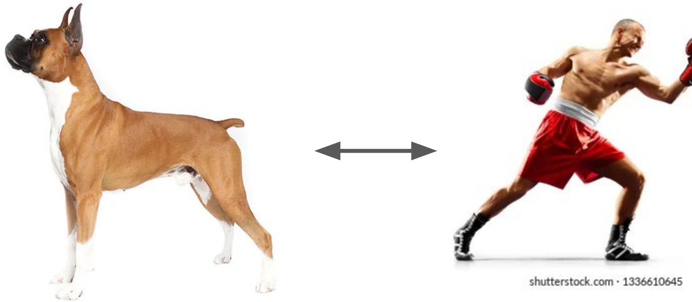
1. CLIP outperforms Visual N-Grams on ImageNet and performs well on task it has not been explicitly trained
    a.

	aYahoo	ImageNet	SUN
Visual N-Grams	72.4	11.5	23.0
CLIP	**98.4**	**76.2**	**58.5**

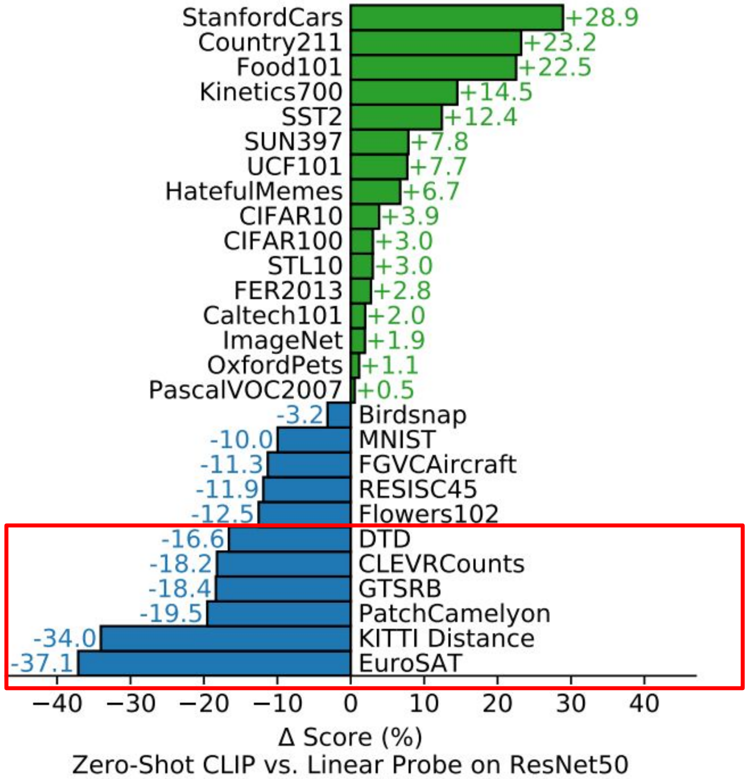
Experiments

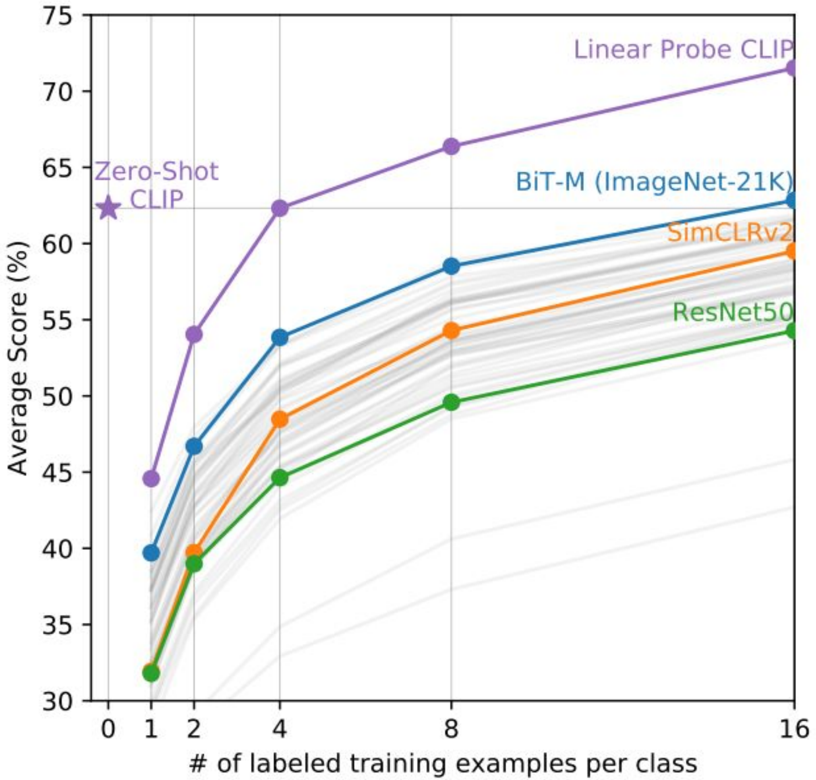
1. prompt engineering
    a. standard image classification datasets treat the information naming or describing classes as an afterthought and annotate images with just a numeric id of the label (then mapping id to their names)
    b. Prompt: construct natural language prompts that can be used to guid the model’s prediction
        i. use a template-based approach where they construct prompts by filling in placeholders with relevant information about the task at hand “A photo of a {label}.”





Experiments





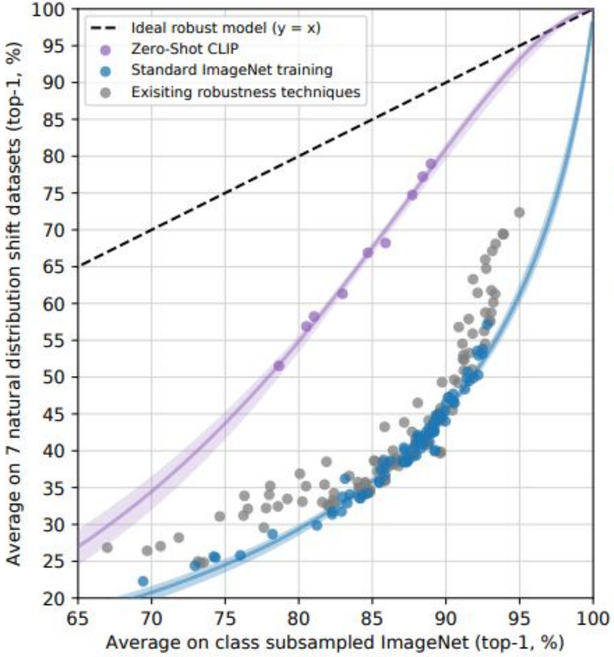
Experiments

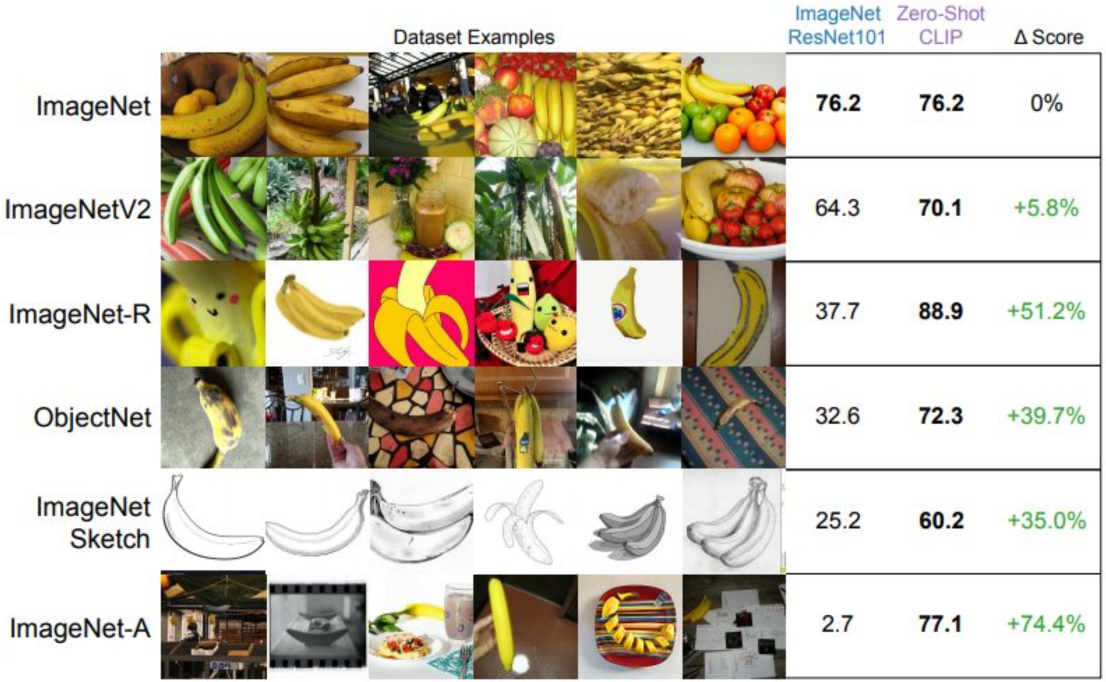
1. zero shot performance of CLIP
    a. impressice on some task but it still quite weak on several kinds of tasks such as
        i. specialized (e.g. satellite image classification)
        ii. abstract and systematic tasks (e.g. counting the number of objects)
        iii. self-driving related tasks (e.g. classifying the distance of the nearest car)
    b. its performance is not yet perfect

Representation Learning Capabilities of CLIP

1. discovering and extracting useful features or representations from raw data
2. common way to evaluating the quality of representation
    a. fit a linear classification on a representation extracted from the model and measures its performance on varios dataset
    b. linear probe, fine-tune

Experiments





Comparison to Human Performance

- Dataset : Oxford IIT Pets dataset select which of the 37 cat or dog breed best matched the image
- Zero-shot : the humans were given no examples of the breeds and asked to label.
- One-shot : one sample image of each breed and in the two-shot experiment they were given two sample images of each breed

	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