OpenAI

CLIP

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford, Jong Wook Kim, et al.

January 2021

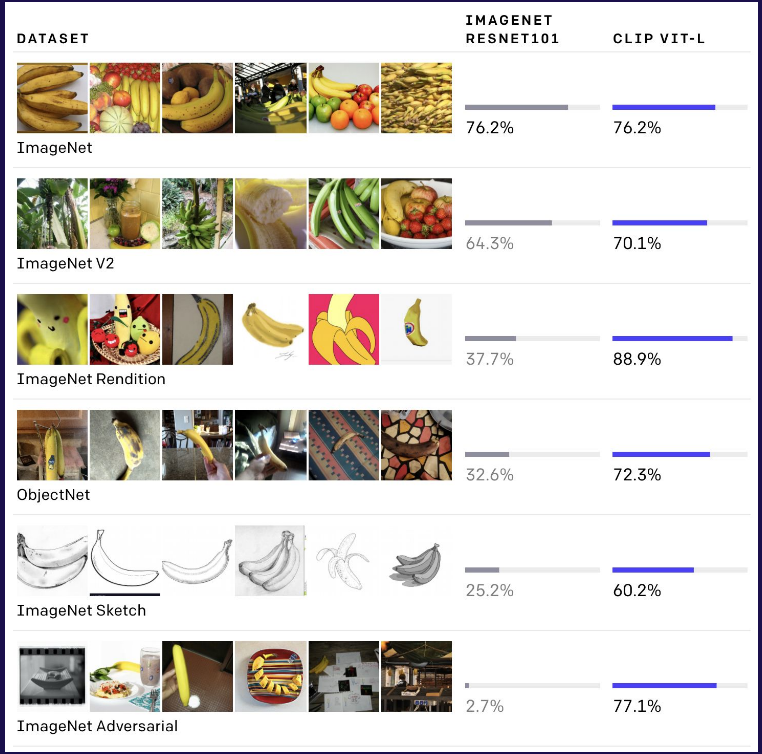
What makes CLIP special?

Motivation:

Instead of using a fixed set of labels,
Get supervision from natural language

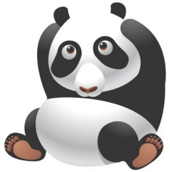
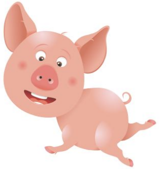
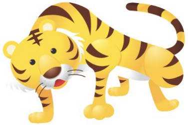
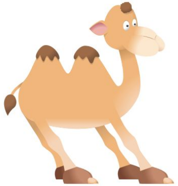
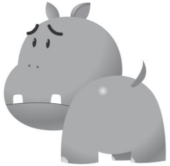
Result:

Robust zero-shot inference
Multimodal embedding space



DATASET	IMAGENET RESNET101	CLIP VIT-L
ImageNet	76.2%	76.2%
ImageNet V2	64.3%	70.1%
ImageNet Rendition	37.7%	88.9%
ObjectNet	32.6%	72.3%
ImageNet Sketch	25.2%	60.2%
ImageNet Adversarial	2.7%	77.1%

How does it work?

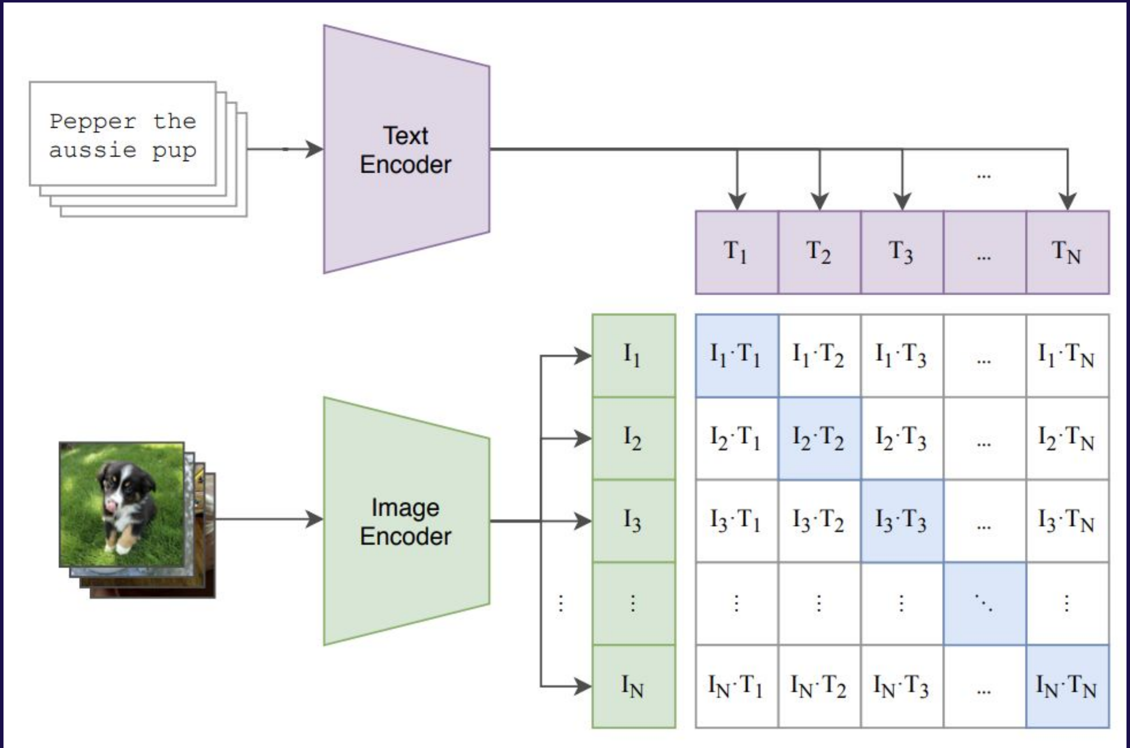






● ● ● ● ●

● ● ● ● ●

Pig Tiger Panda Hippo Camel

CLIP: Contrastive Language-Image Pre-training



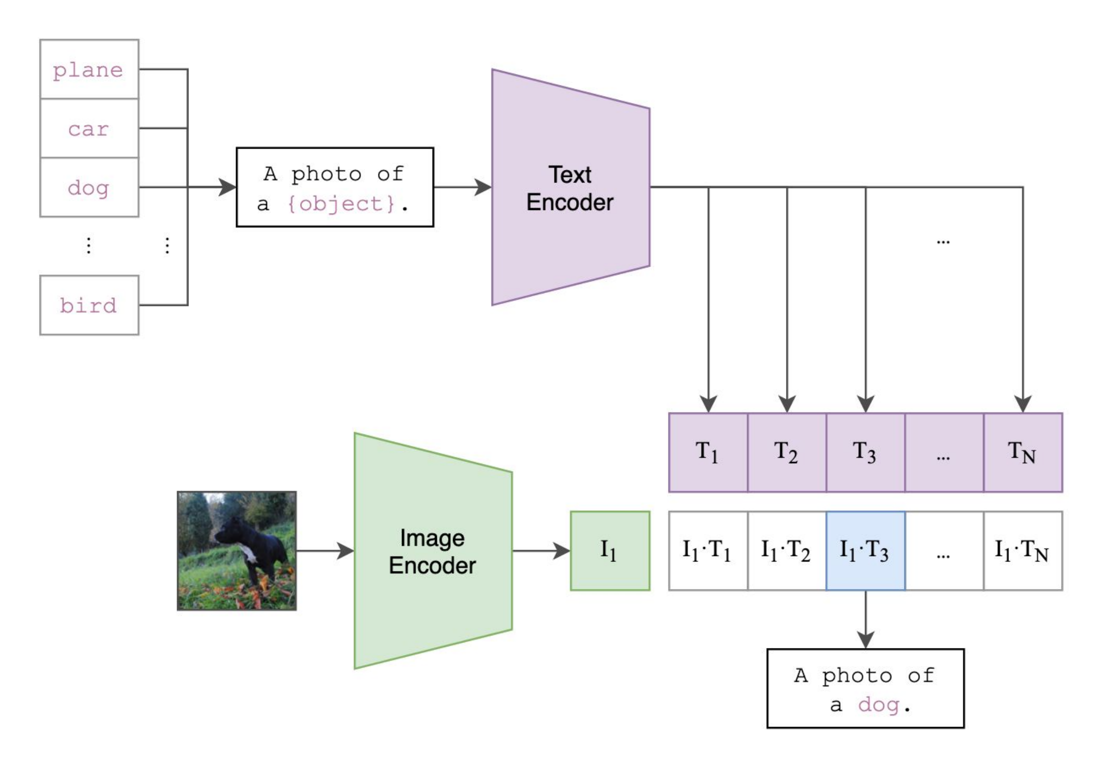
Pepper the aussie pup

Text Encoder

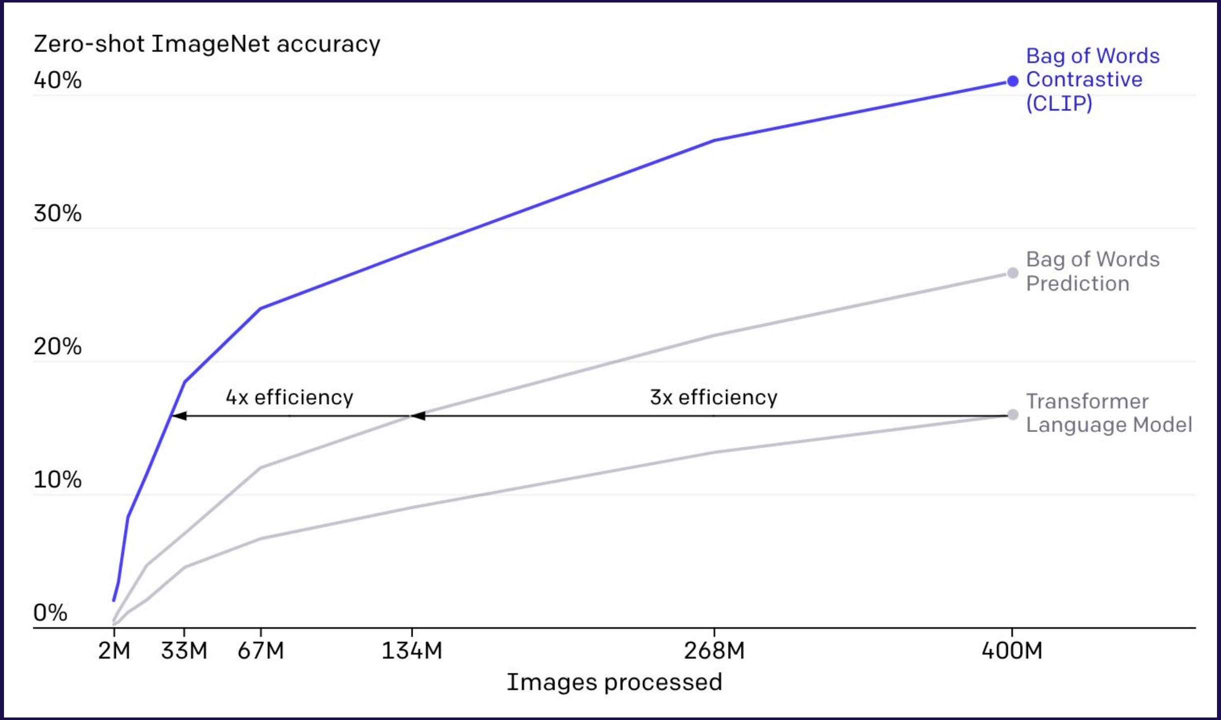
Image Encoder

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

Zero-shot image classification



Why contrastive



Some CLIP details

Training

- Trained on 400M image-text pairs from the internet
- Batch size of 32,768
- 32 epochs over the dataset
- Cosine learning rate decay

Architecture

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

Representation Learning

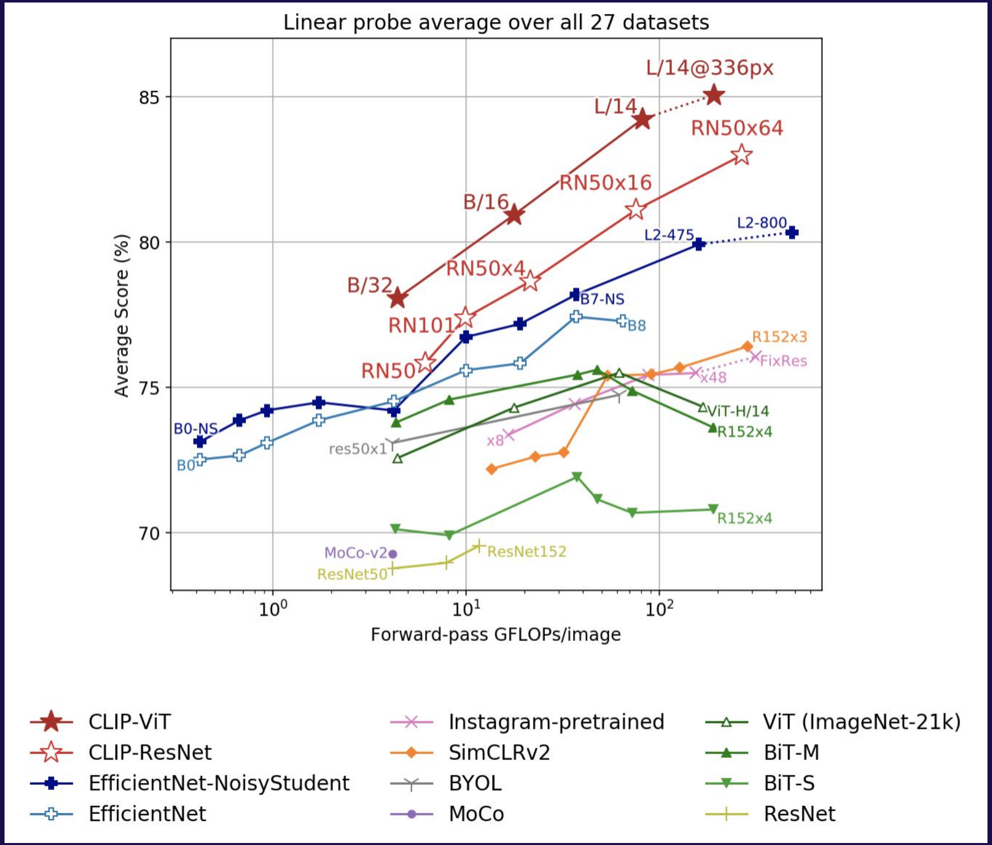
Linear probe

Logistic regression classifier on image features

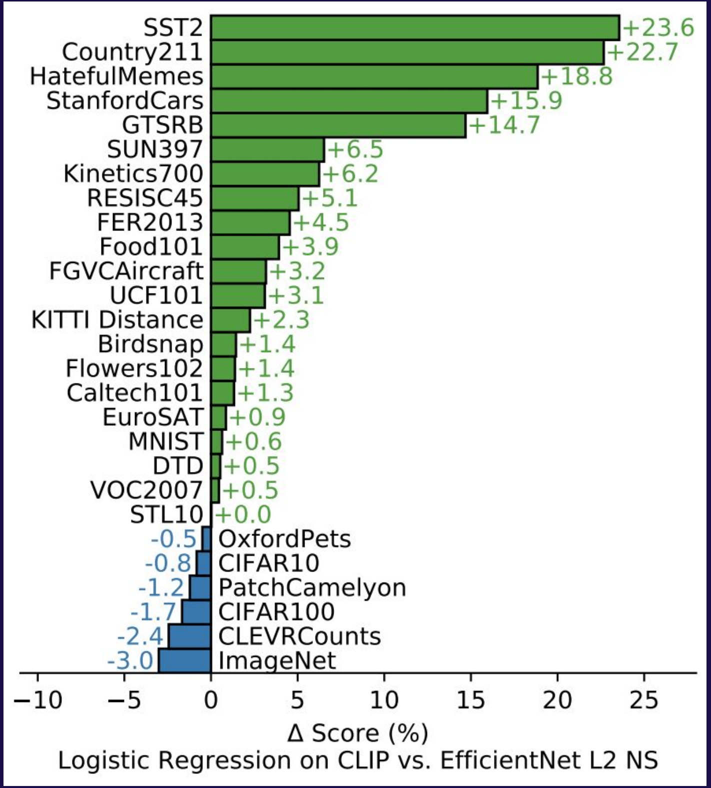
- L-BFGS
- Only one hyperparameter
- Allows “fair” comparisons with other vision models
- Provides lower bound for fine-tuned models

Evaluated on 27 datasets × 65 vision models

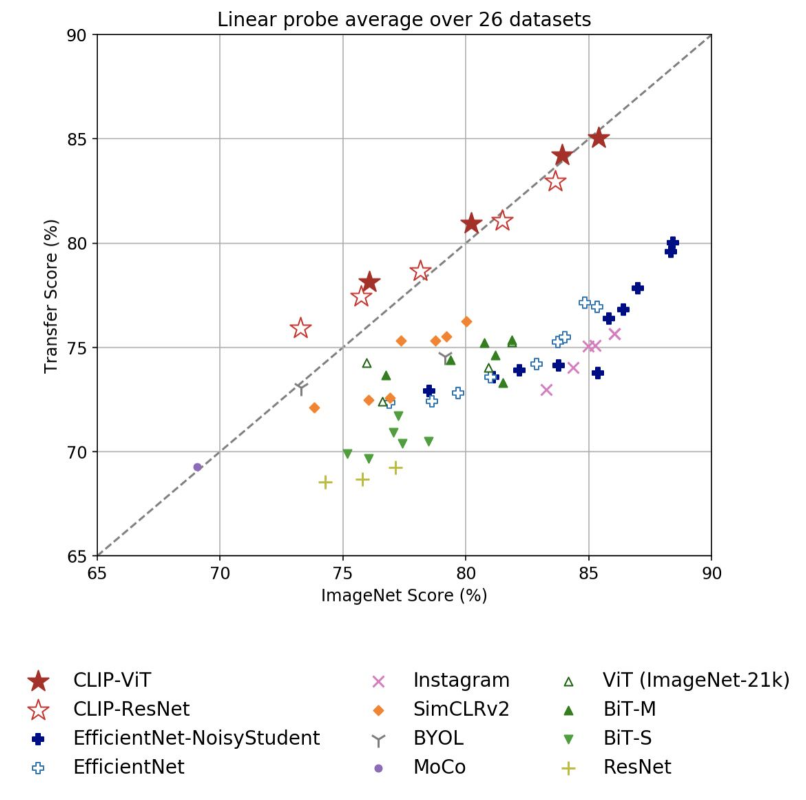
Linear probe performance vs SOTA vision models



Linear-probe CLIP vs Linear-probe EfficientNet-L2

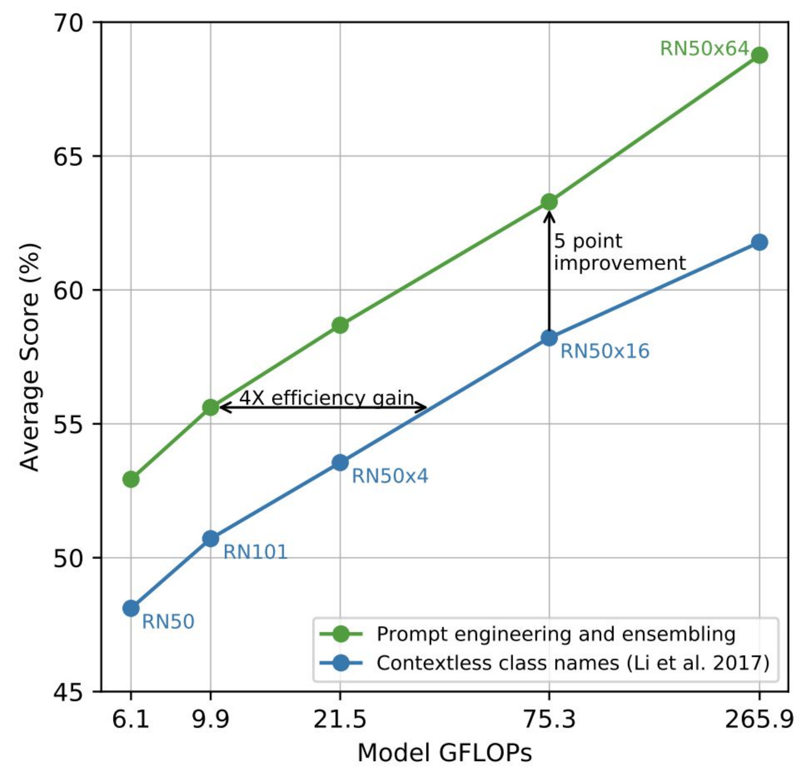


vs ImageNet score



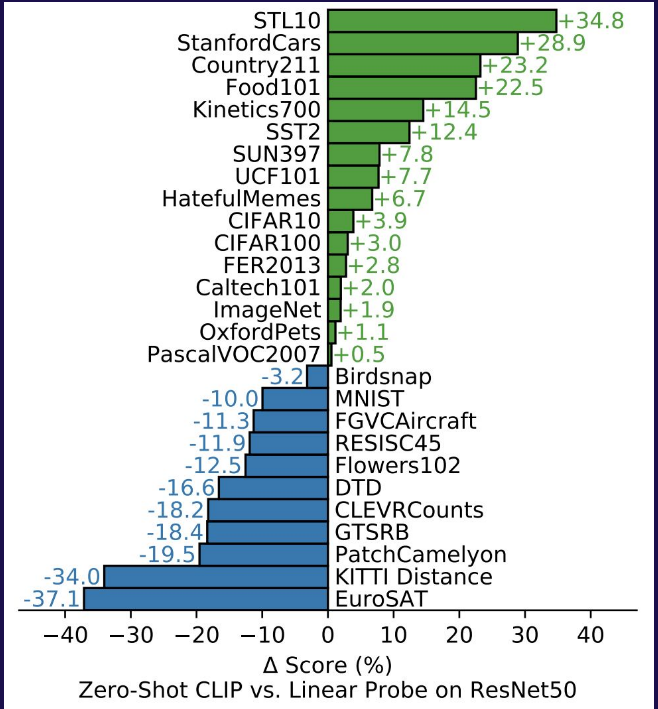
Zero-Shot Transfer

Prompt engineering



Zero-shot vs Linear-probe ResNet-50

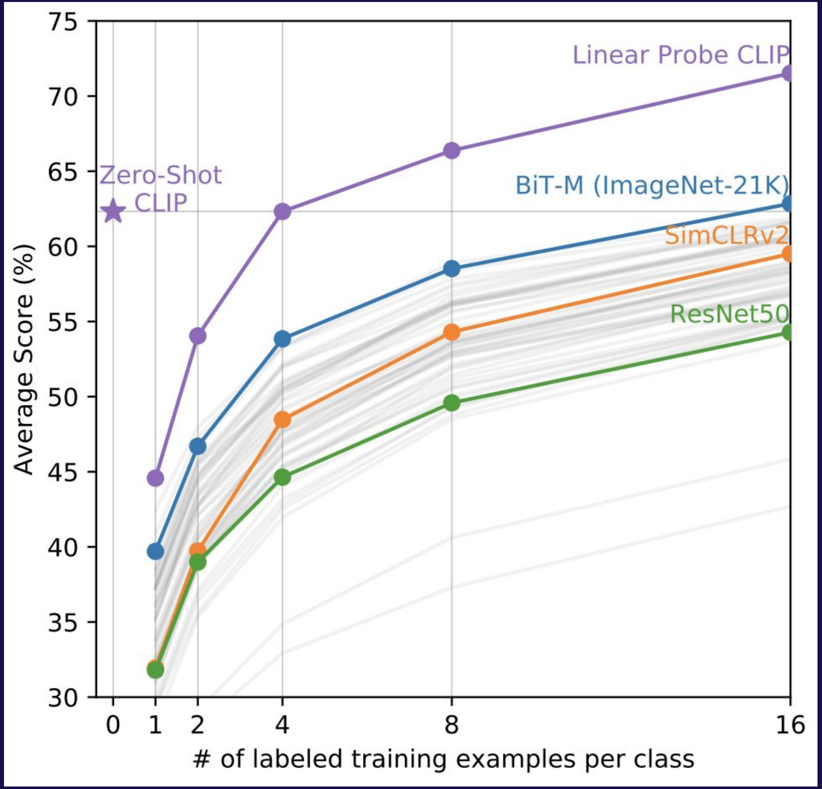
Zero-shot CLIP outperforms ResNet-50 on 16 of 27 datasets



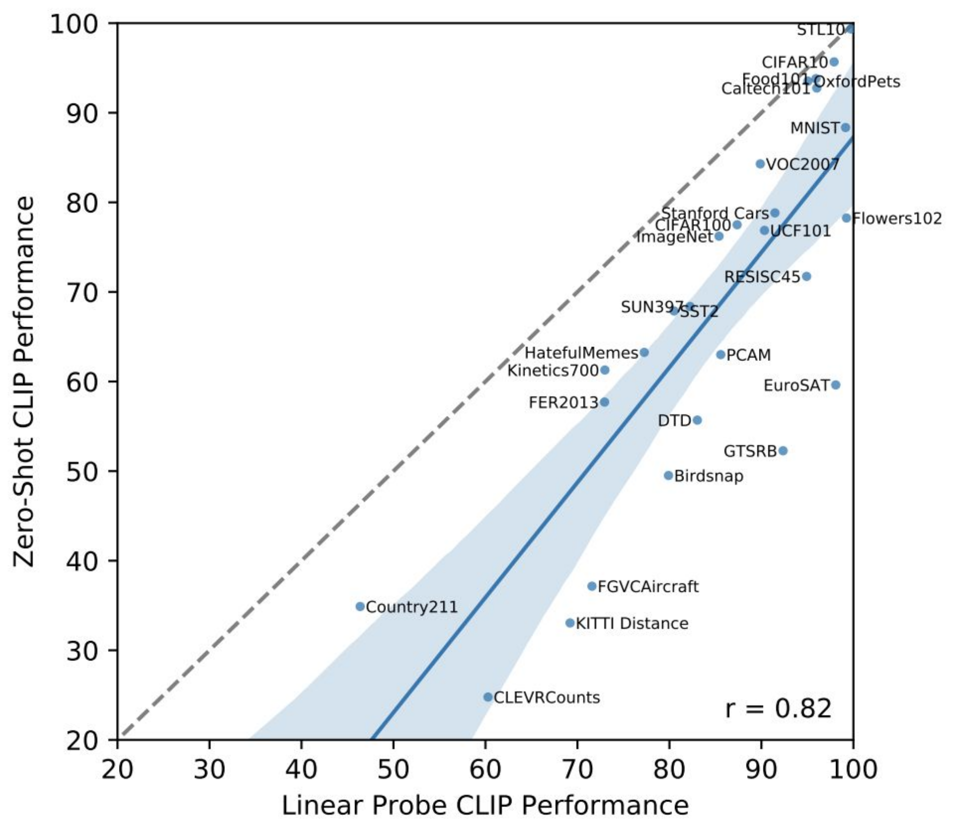
Zero-shot CLIP vs Few-shot linear probes

Zero-shot CLIP is as good as

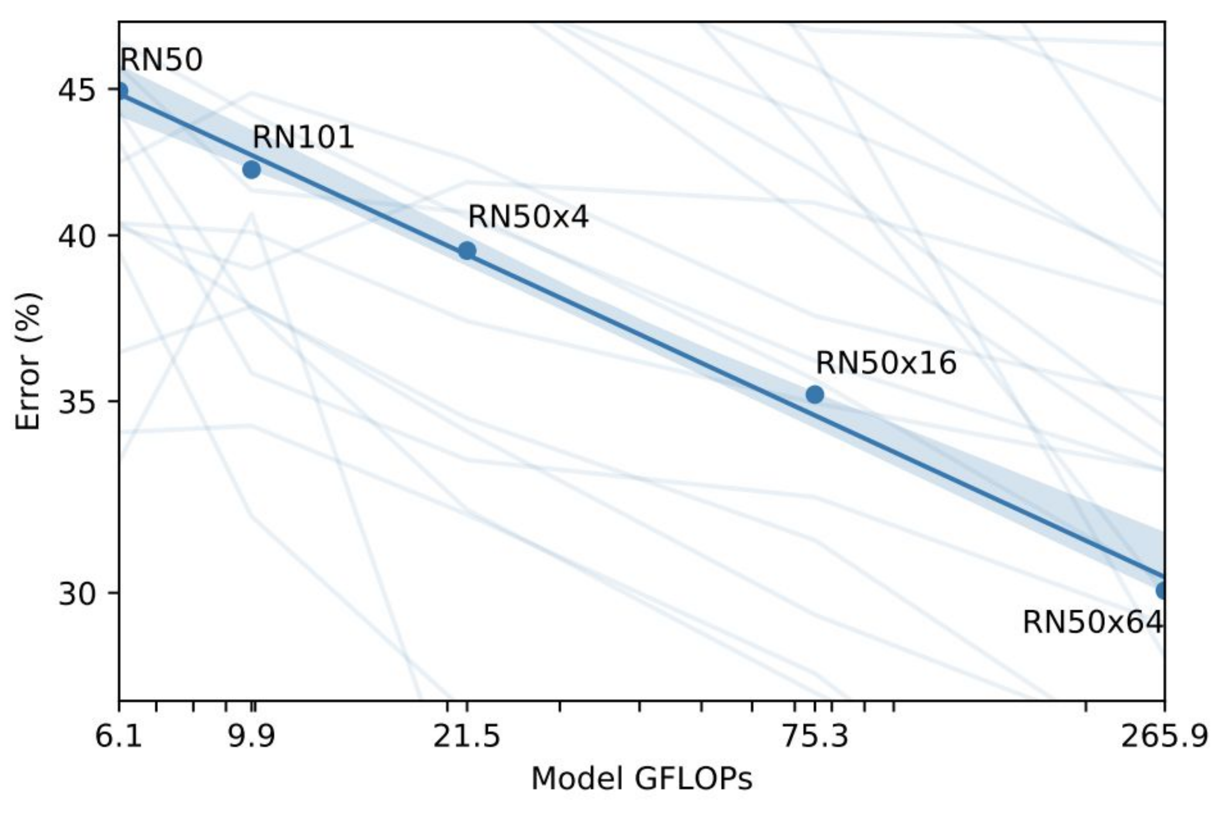
- 4-shot linear-probe CLIP
- 16-shot BiT-M



Zero-shot vs Linear-probe CLIP



Zero-shot performance vs model size



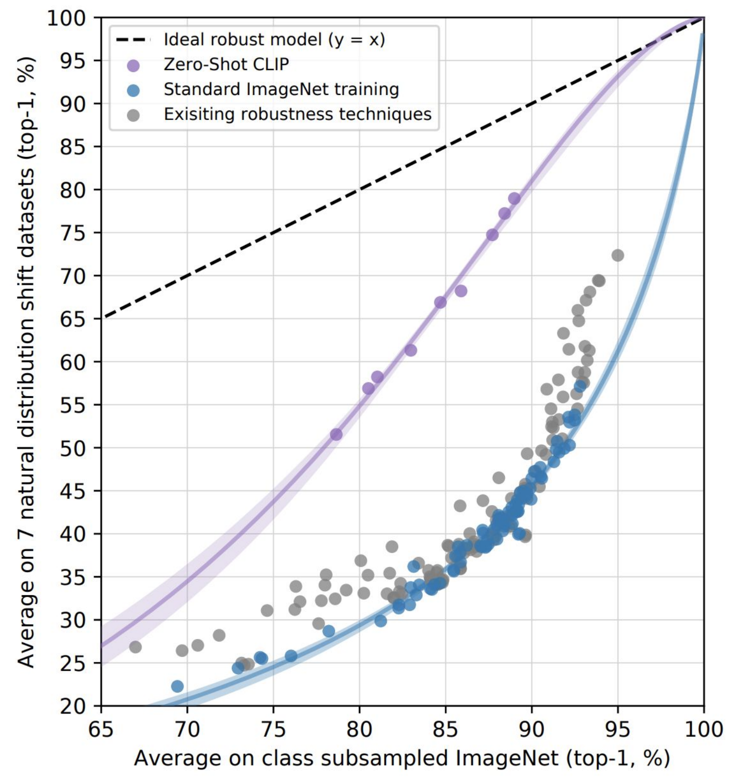
Robustness to Natural Distribution Shift

Robustness to natural distribution shift

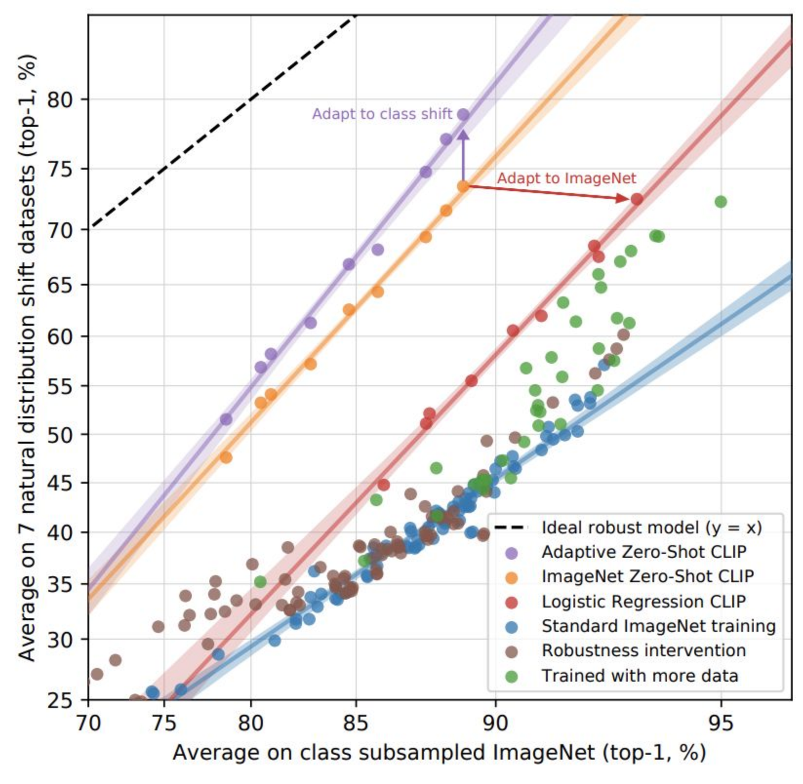
CLIP is significantly more robust!

7 ImageNet-like Datasets (Taori et al.)

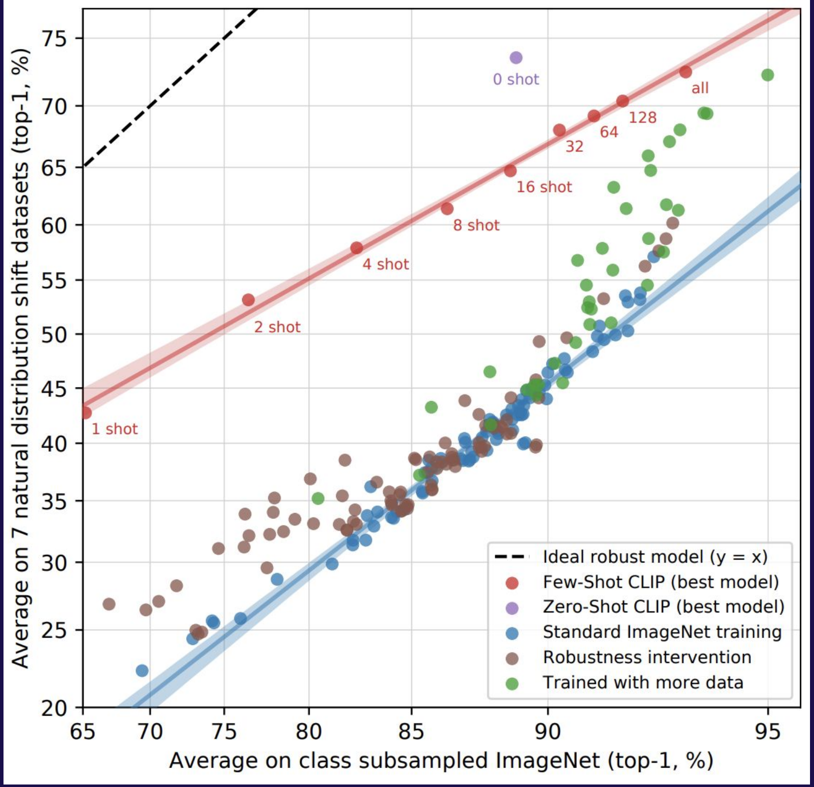
- ImageNetV2
- ImageNet-A
- ImageNet-R
- ImageNet Sketch
- ObjectNet
- ImageNet Vid
- Youtube-BB



Adapting to ImageNet does not help robustness



Robustness of few-shot linear probes



Limitations and Broader Impacts

Limitations of CLIP

- Zero-shot performance is well below the SOTA
- Especially weak on abstract tasks such as counting
- Poor on out-of-distribution data such as MNIST
- Susceptible to adversarial attacks
- Dataset selection bias
- Social biases

Quantifying the (un)safety of CLIP models

CLIP has societal biases

- Race
- Gender
- Age

Surveillance usage

- Zero-shot scene classification
- Zero-shot identification of celebrities

Not comprehensive, will continue research to ensure safety
Model card limits usage of CLIP to research-only

Related Work

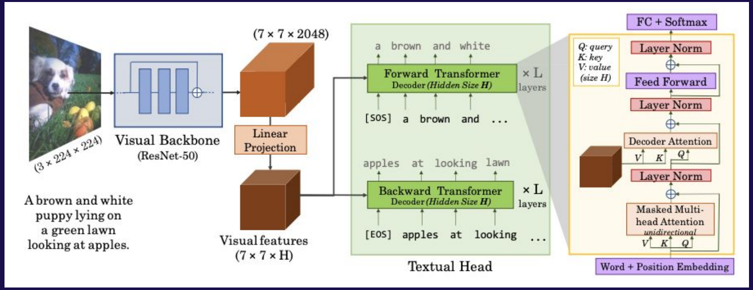
Multimodal learning

- VirTex
- ICMLM
- ConVIRT

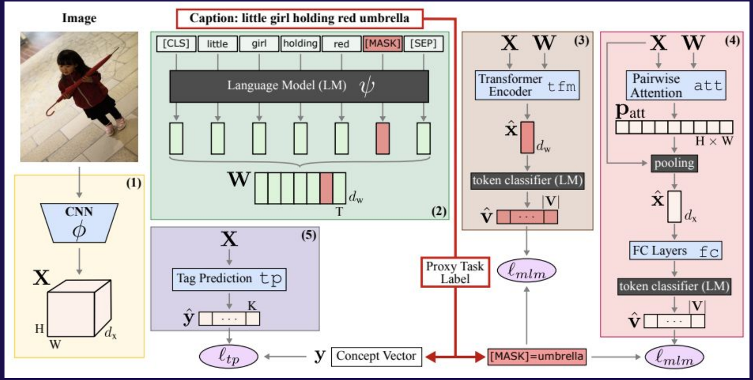
Natural language supervision

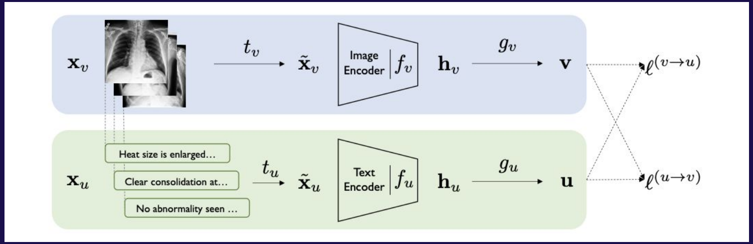
Text-image retrieval

Webly supervised learning



(3 × 224 × 224)
A brown and white puppy lying on a green lawn looking at apples.
Visual Backbone (ResNet-50)
(7 × 7 × 2048)
Linear Projection
Visual features (7 × 7 × H)
a brown and white
Forward Transformer Decoder (Hidden Size H)
× L layers
[SOS] a brown and ...
apples at looking lawn
Backward Transformer Decoder (Hidden Size H)
× L layers
[EOS] apples at looking ...
Textual Head
Q: query K: key V: value (size H)
FC + Softmax
Layer Norm
Feed Forward
Layer Norm
Decoder Attention
V K Q
Layer Norm
Masked Multi-head Attention unidirectional
V K Q
Word + Position Embedding

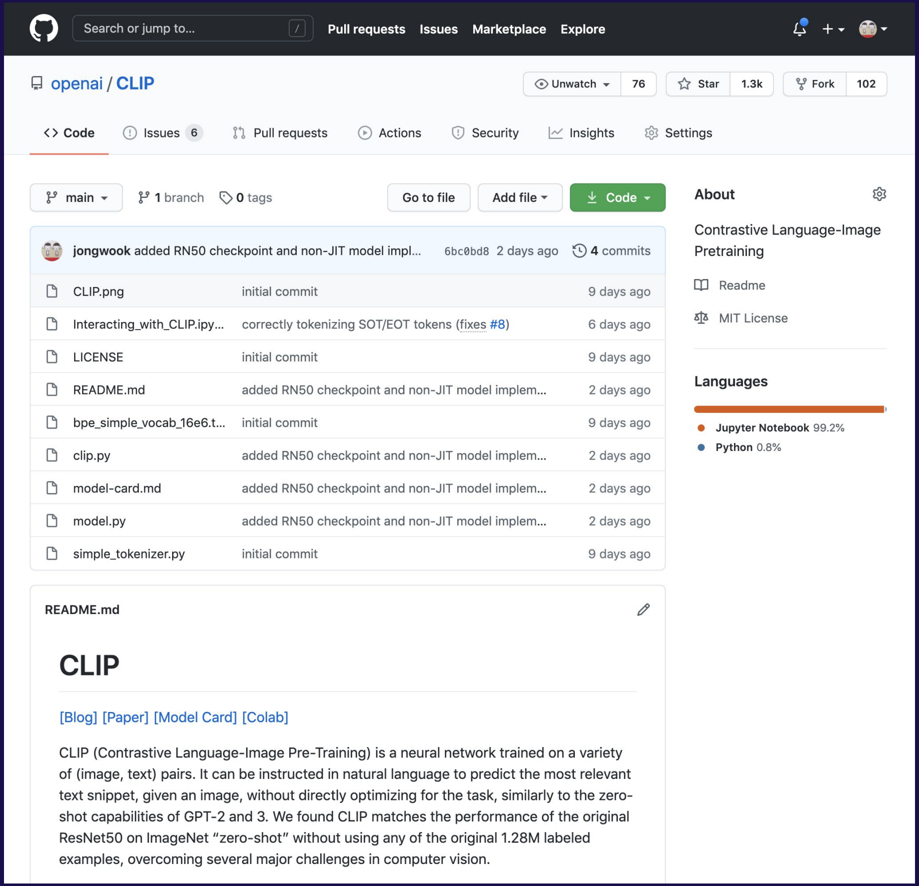




Try CLIP today!

https://github.com/openai/CLIP

- PyTorch implementation
- Colab notebook



jongwook added RN50 checkpoint and non-JIT model impl... 6bc0bd8 2 days ago 4 commits

CLIP.png	initial commit	9 days ago
Interacting_with_CLIP.ipy...	correctly tokenizing SOT/EOT tokens (fixes #8)	6 days ago
LICENSE	initial commit	9 days ago
README.md	added RN50 checkpoint and non-JIT model implem...	2 days ago
bpe_simple_vocab_16e6.t...	initial commit	9 days ago
clip.py	added RN50 checkpoint and non-JIT model implem...	2 days ago
model-card.md	added RN50 checkpoint and non-JIT model implem...	2 days ago
model.py	added RN50 checkpoint and non-JIT model implem...	2 days ago
simple_tokenizer.py	initial commit	9 days ago

About

Contrastive Language-Image Pretraining

Readme

MIT License

Languages

- Jupyter Notebook 99.2%
- Python 0.8%

README.md

CLIP

[Blog] [Paper] [Model Card] [Colab]

CLIP (Contrastive Language-Image Pre-Training) is a neural network trained on a variety of (image, text) pairs. It can be instructed in natural language to predict the most relevant text snippet, given an image, without directly optimizing for the task, similarly to the zero-shot capabilities of GPT-2 and 3. We found CLIP matches the performance of the original ResNet50 on ImageNet “zero-shot” without using any of the original 1.28M labeled examples, overcoming several major challenges in computer vision.

Thank You

Visit openai.com for more information.

FOLLOW @OPENAI ON TWITTER