



OpenAI

CLIP

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford, Jong Wook Kim, et al.
April 28 2021

What?

Recognizes things
in a visual scene

The diagram features a central title, 'Learning Transferable Visual Models From Natural Language Supervision', which is bracketed at both ends. Above the title, a bracket connects it to the text 'Recognizes things in a visual scene'. Below the title, two brackets extend outwards to the phrases 'One model can be adapted to a variety of tasks' on the left and 'Learns about images from free-form text' on the right.

Learning Transferable Visual Models From Natural Language Supervision

One model can be adapted
to a variety of tasks

Learns about images
from free-form text

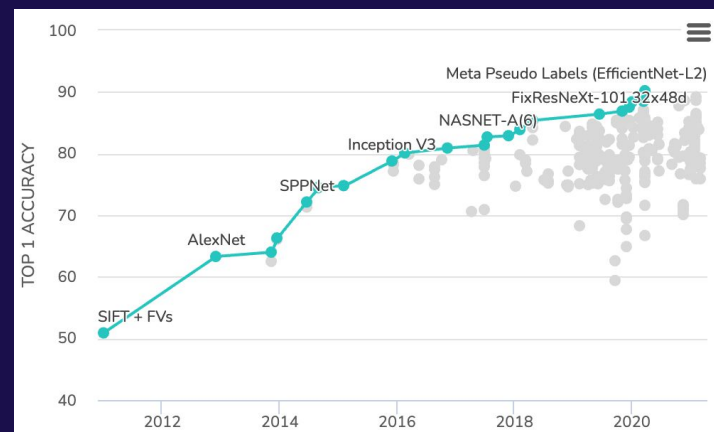
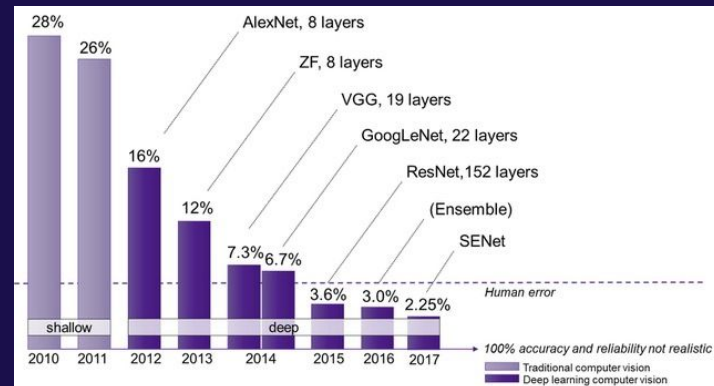
Vision models led the deep learning boom

ImageNet competition

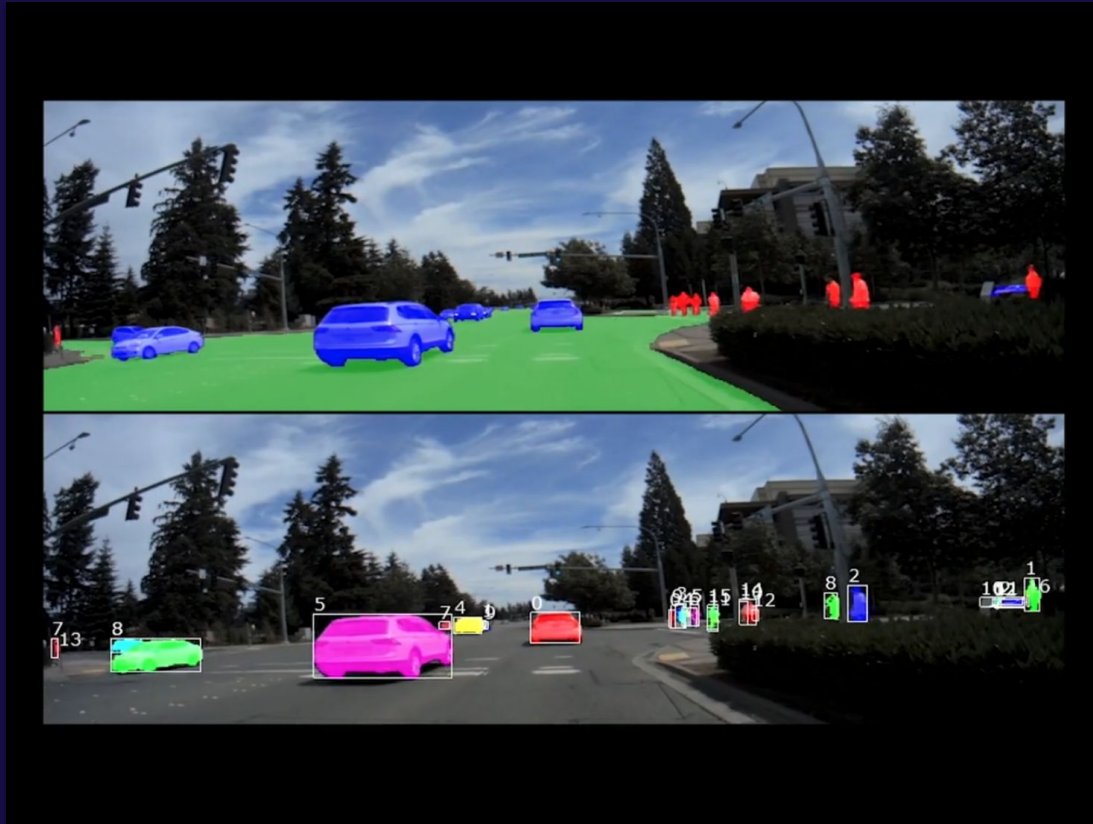
- AlexNet (2012)
- VGG (2014)
- GoogLeNet (2014)
- ResNet (2015)
- SENet (2017)

Human top-5 accuracy: 5%

Top-1 as an ongoing benchmark



Vision models today









What makes CLIP special?

Motivation:

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

Result:

Robust zero-shot inference
Multimodal feature 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

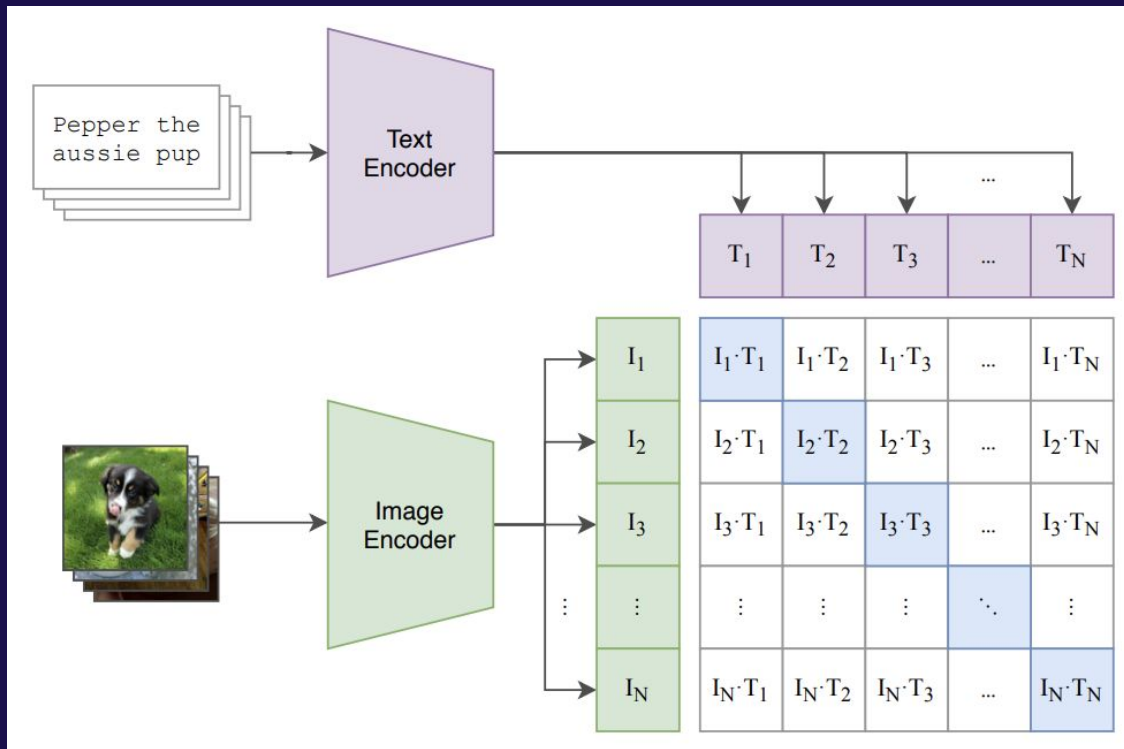


Hippo

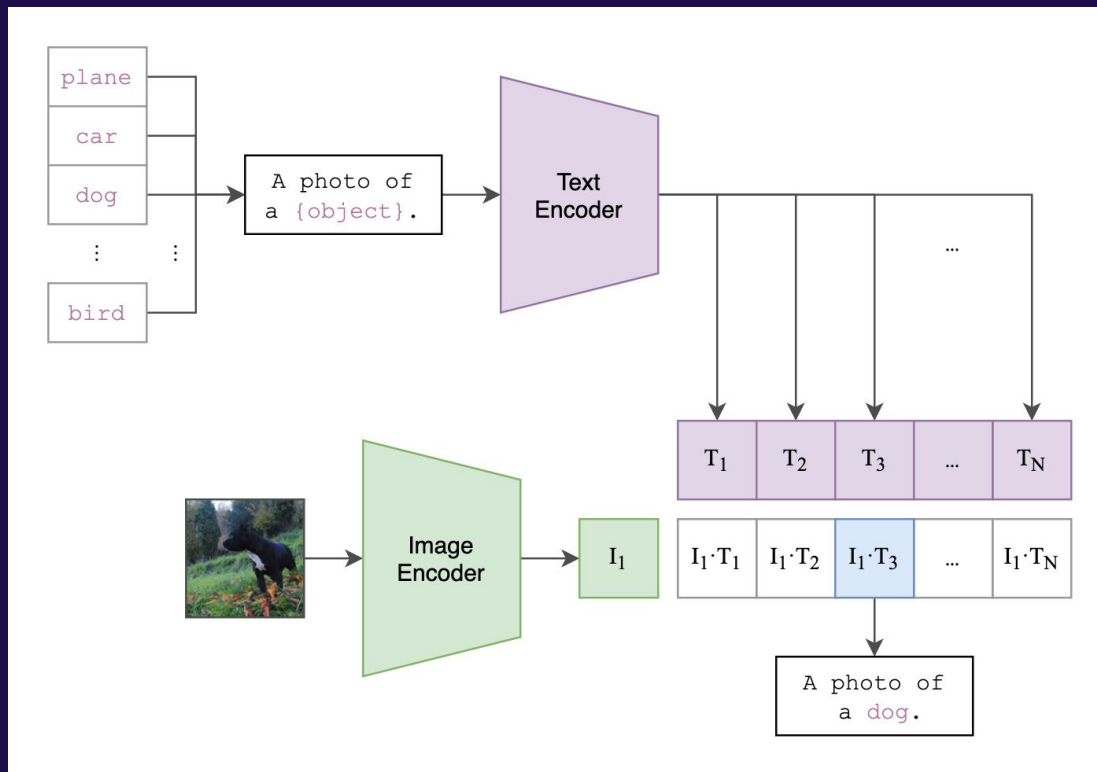


Camel

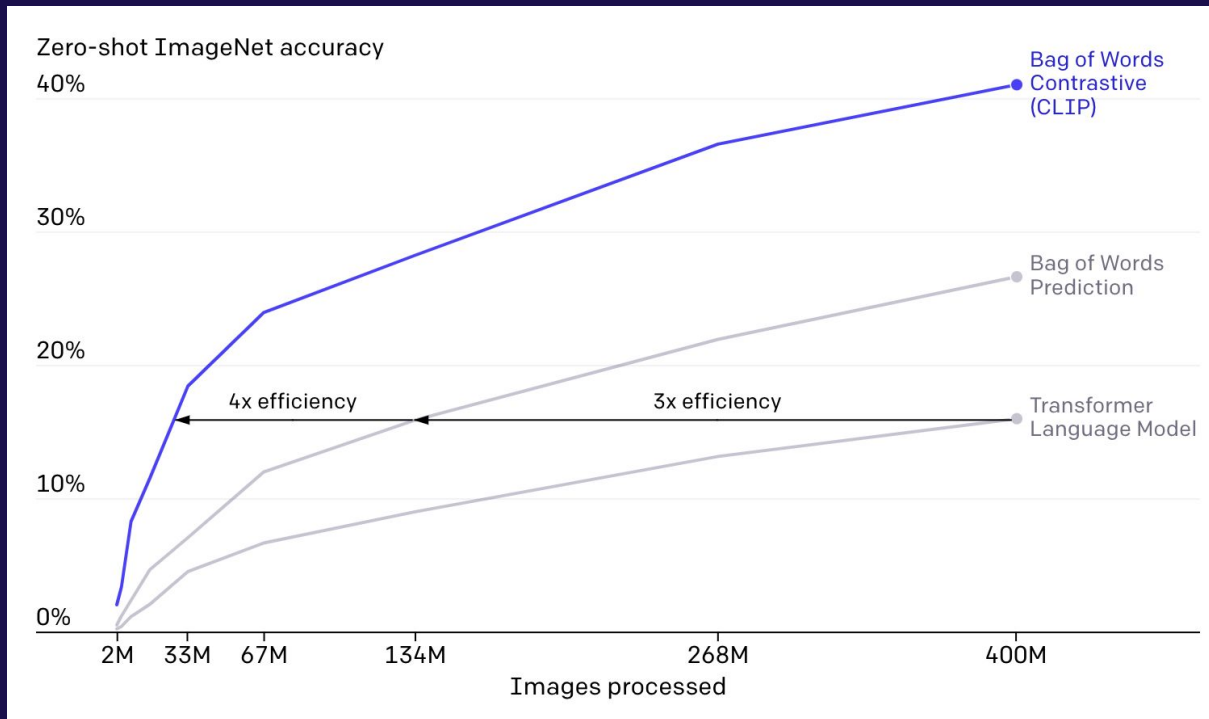
CLIP: Contrastive Language-Image Pre-training



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

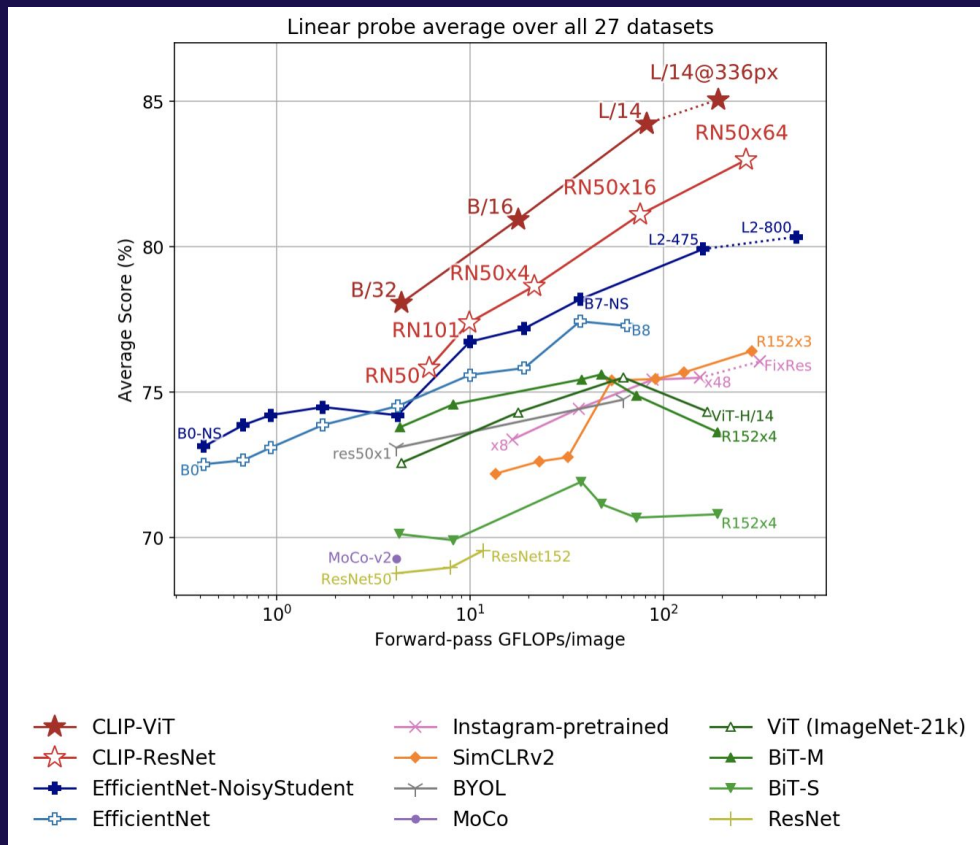
- L-BFGS
- Only one hyperparameter
- Allows “fair” comparisons with other vision models
- Provides lower bound for fine-tuned models

Evaluated on 27 image datasets × 65 vision models

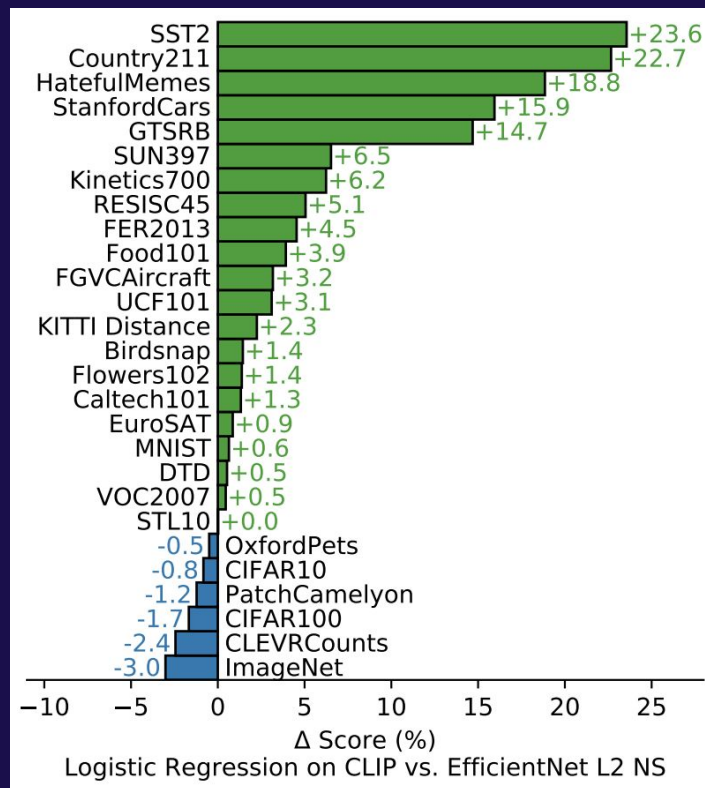


satellite images, car models, medical images, city classification, rendered texts, aircrafts, birds, memes, ...

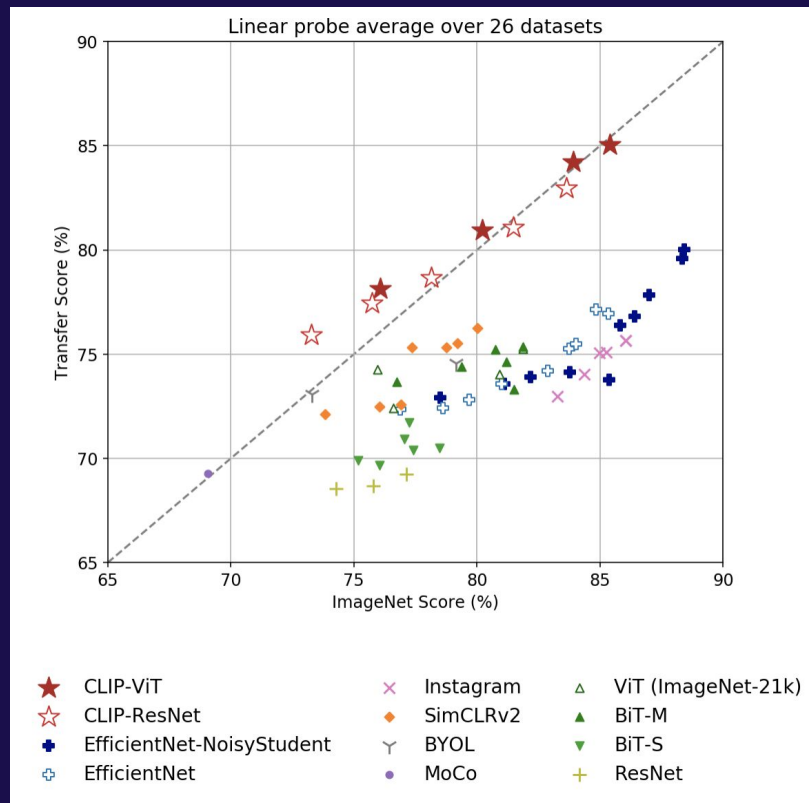
Linear probe performance vs SOTA vision models



Linear-probe CLIP vs Linear-probe EfficientNet-L2

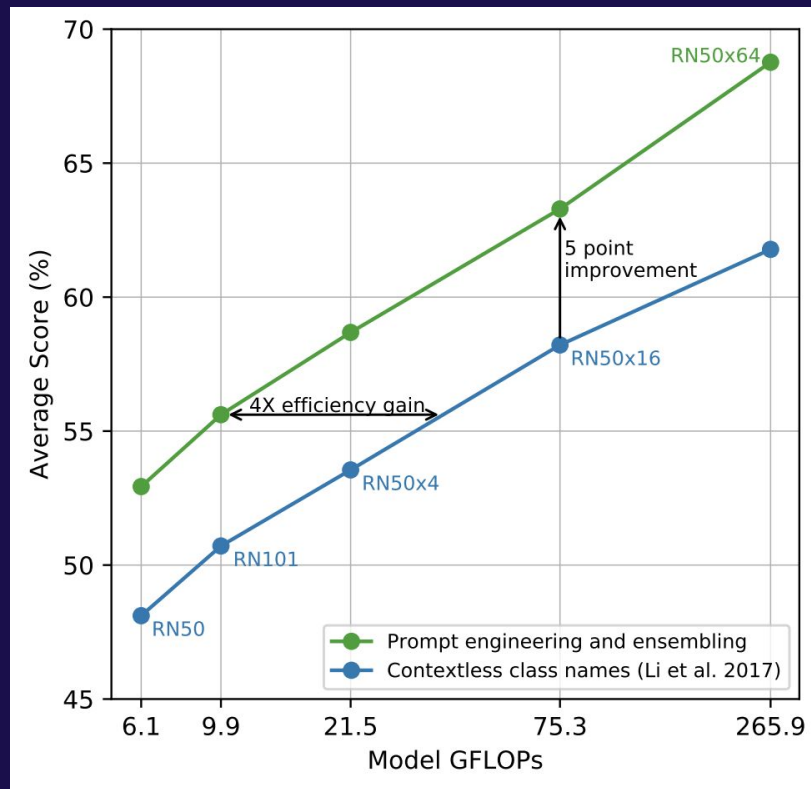


vs ImageNet score



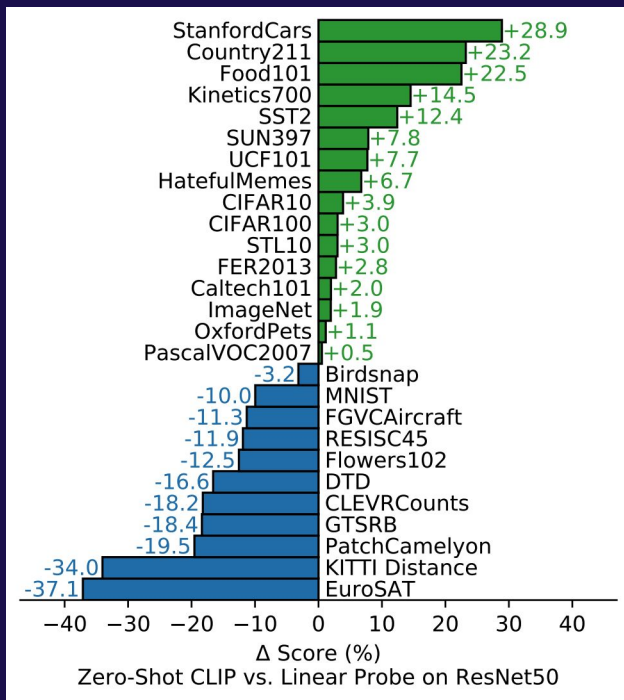
Zero-Shot Transfer

Prompt engineering



Zero-shot vs Linear-probe ResNet-50

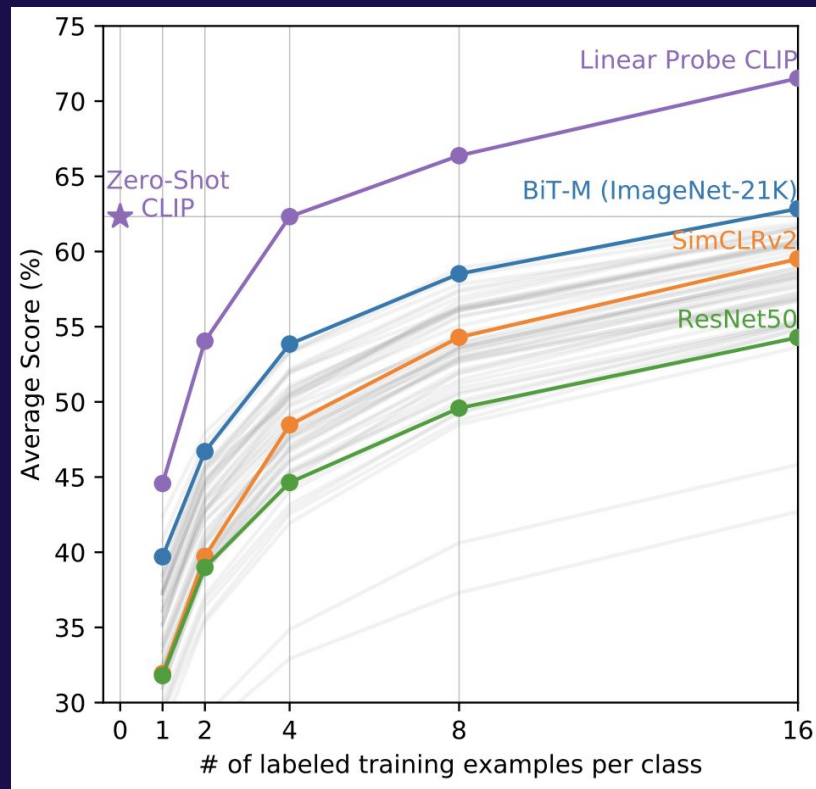
Zero-shot CLIP outperforms ResNet-50 on 16 of 27 datasets



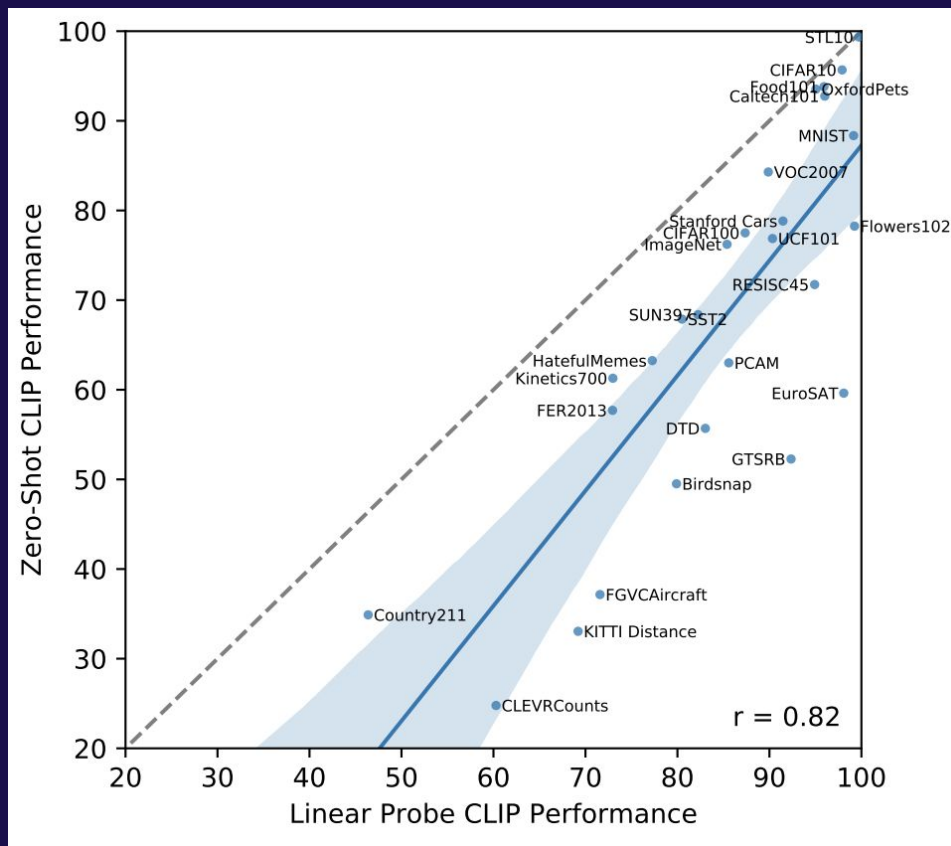
Zero-shot CLIP vs Few-shot linear probes

Zero-shot CLIP is as good as

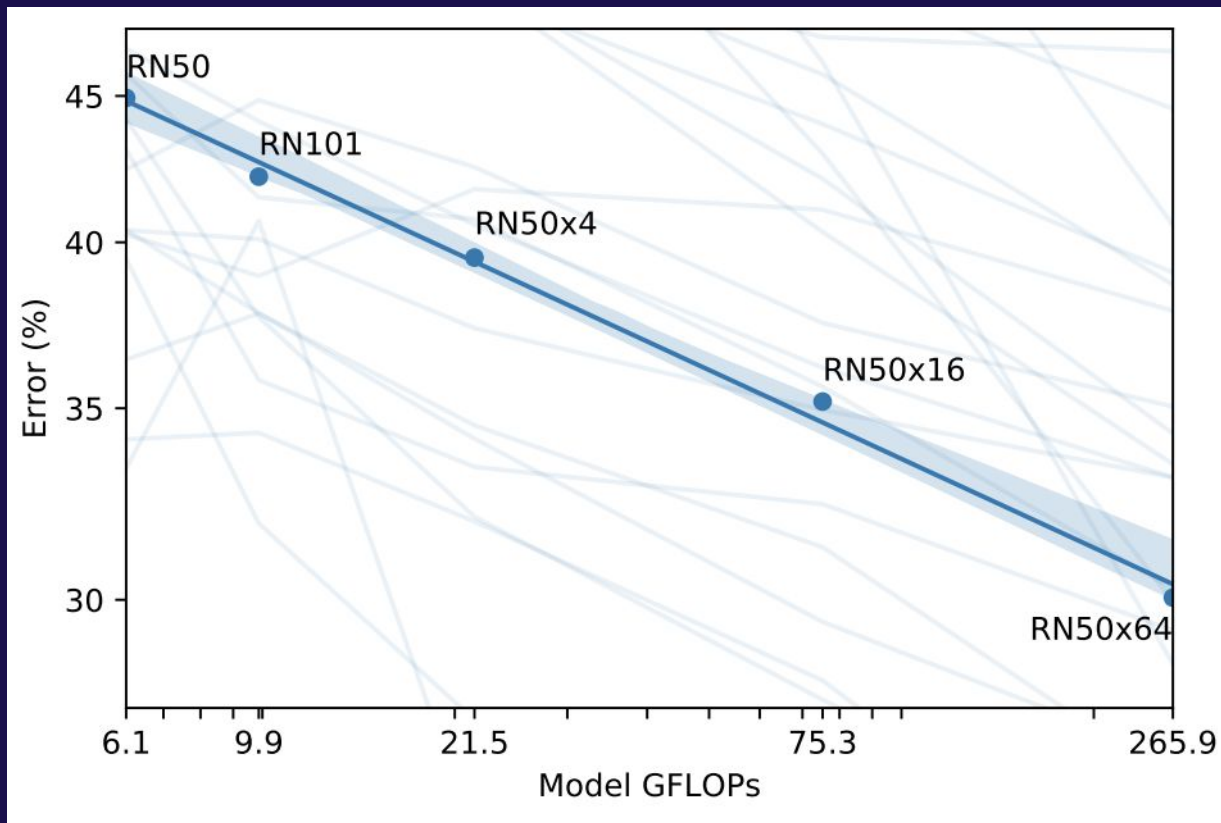
- 4-shot linear-probe CLIP
- 16-shot BiT-M



Zero-shot vs Linear-probe CLIP



Zero-shot performance vs model size



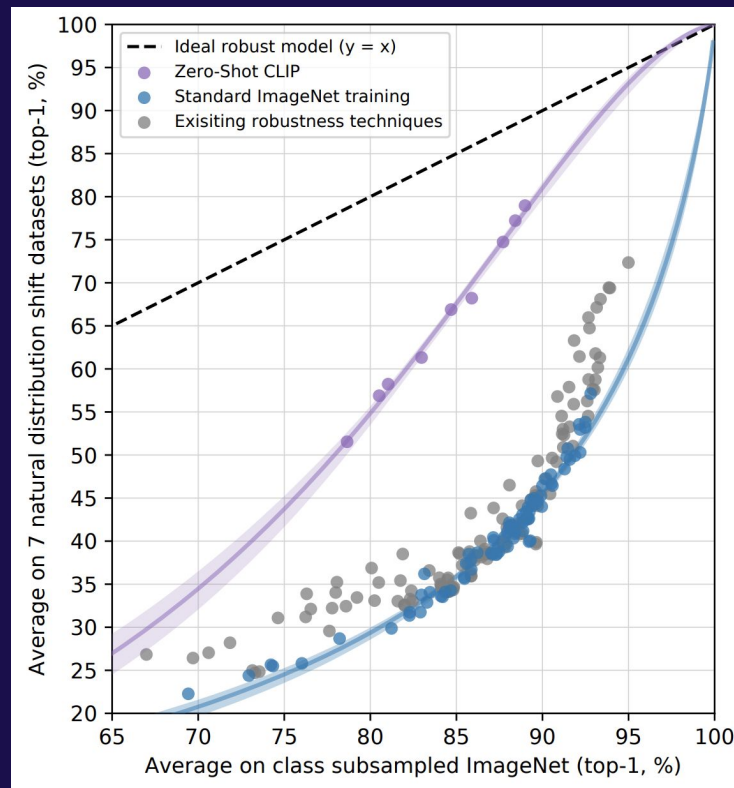
Robustness to Natural Distribution Shift

Robustness to natural distribution shift

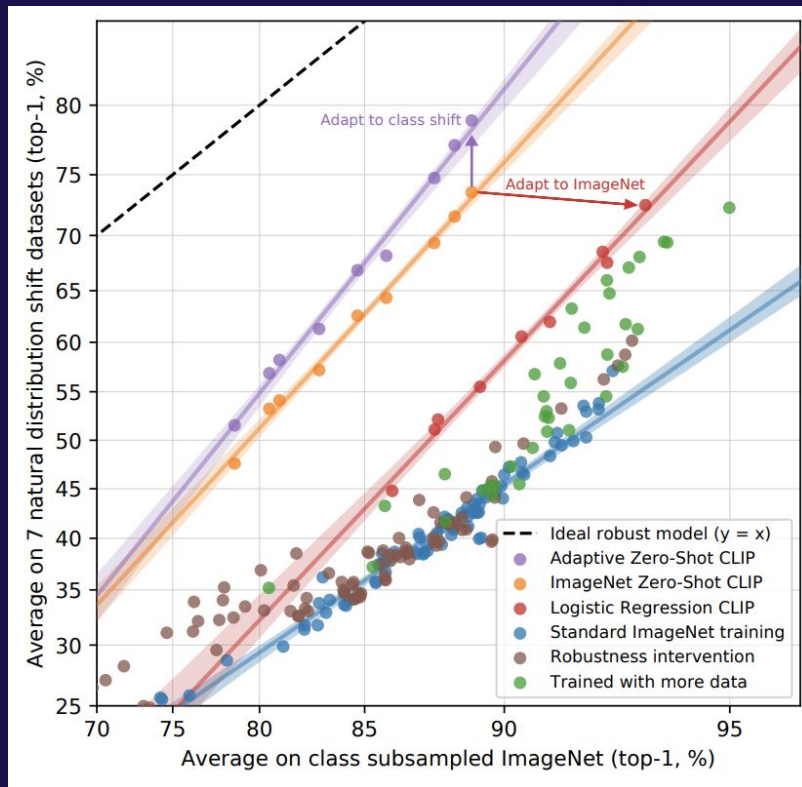
CLIP is significantly more robust!

7 ImageNet-like Datasets (Taori et al.)

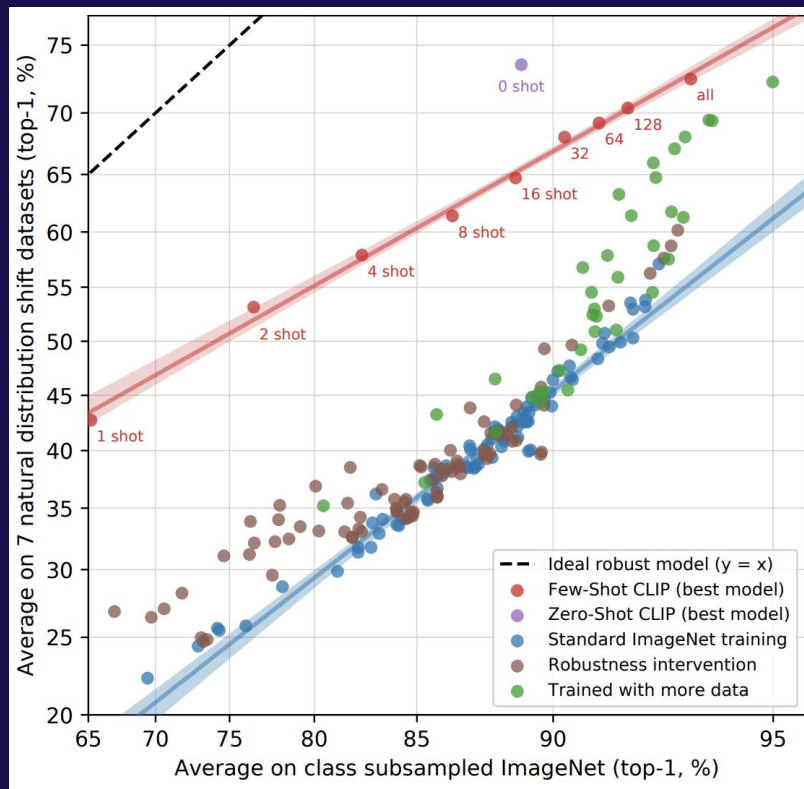
- ImageNetV2
- ImageNet-A
- ImageNet-R
- ImageNet Sketch
- ObjectNet
- ImageNet Vid
- Youtube-BB



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 in the eval suite
- Social biases

Quantifying the (un)safety of CLIP models

Social Biases

- Race
- Gender
- Age

Surveillance usage

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

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

Related Work

Prior Related Work

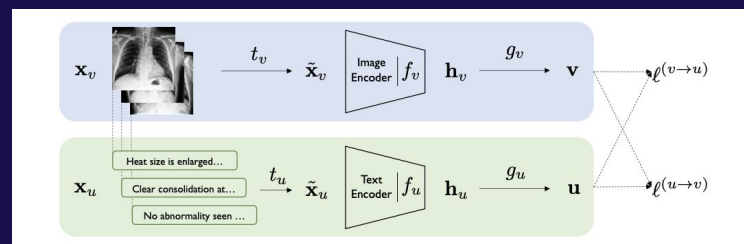
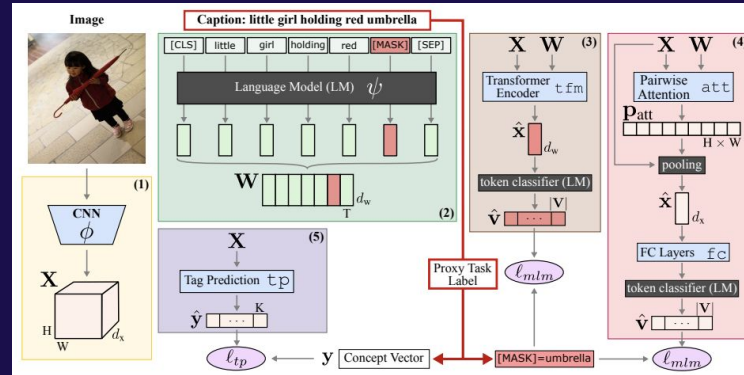
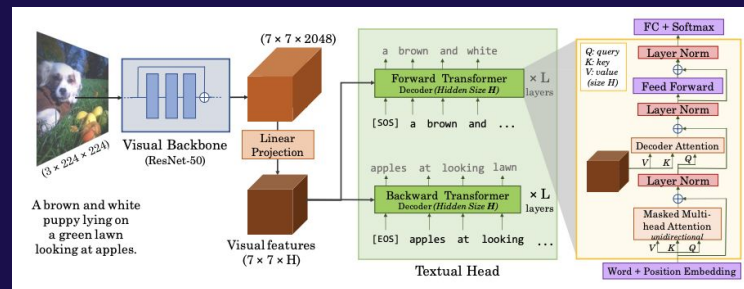
Multimodal learning

- VirTex
- ICMLM
- ConVIRT









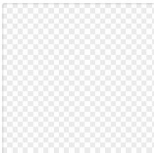
Natural language supervision

Text-image retrieval

Webly supervised learning



Multimodal Neurons in CLIP

BIOLOGICAL NEURON	CLIP NEURON	PREVIOUS ARTIFICIAL NEURON	
Probed via depth electrodes	Neuron 244 from penultimate layer in CLIP RN50x4	Neuron 483, generic person detector from Inception v1	
Halle Berry	Spider-Man	human face	
 <p>Responds to photos of Halle Berry and Halle Berry in costume</p> <p>✓</p>	 <p>Responds to photos of Spider-Man in costume and spiders</p> <p>✓</p>	 <p>Responds to photos of human faces</p> <p>✓</p>	Photorealistic images
 <p>Responds to sketches of Halle Berry</p> <p>✓</p>	 <p>Responds to comics or drawings of Spider-Man and spider-themed icons</p> <p>✓</p>	 <p>Does not respond significantly to drawings of faces</p> <p>✗</p>	Conceptual drawings
 <p>Responds to the text "Halle Berry"</p> <p>✓</p>	 <p>Responds to the text "spider" and others</p> <p>✓</p>	 <p>Does not respond significantly to text</p> <p>✗</p>	Images of text

Typographic Attacks

NO LABEL

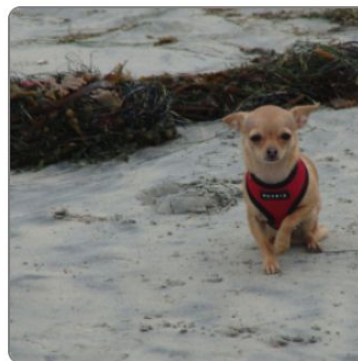


Granny Smith	85.61%
iPod	0.42%
library	0%
pizza	0%
rifle	0%
toaster	0%
dough	0.1%
assault rifle	0%
patio	0.56%

LABELED "IPOD"



Granny Smith	0.13%
iPod	99.68%
library	0%
pizza	0%
rifle	0%
toaster	0%
dough	0%
assault rifle	0%
patio	0%



Chihuahua	17.5%
Miniature Pinscher	14.3%
French Bulldog	7.3%
Griffon Bruxellois	5.7%
Italian Greyhound	4%
West Highland White Terrier	2.1%
Schipperke	2%
Maltese	2%
Australian Terrier	1.9%



Target class:
pizza

Attack text:
pizza



pizza	83.7%
pretzel	2%
Chihuahua	1.5%
broccoli	1.2%
hot dog	0.6%
Boston Terrier	0.6%
French Bulldog	0.5%
spatula	0.4%
Italian Greyhound	0.3%

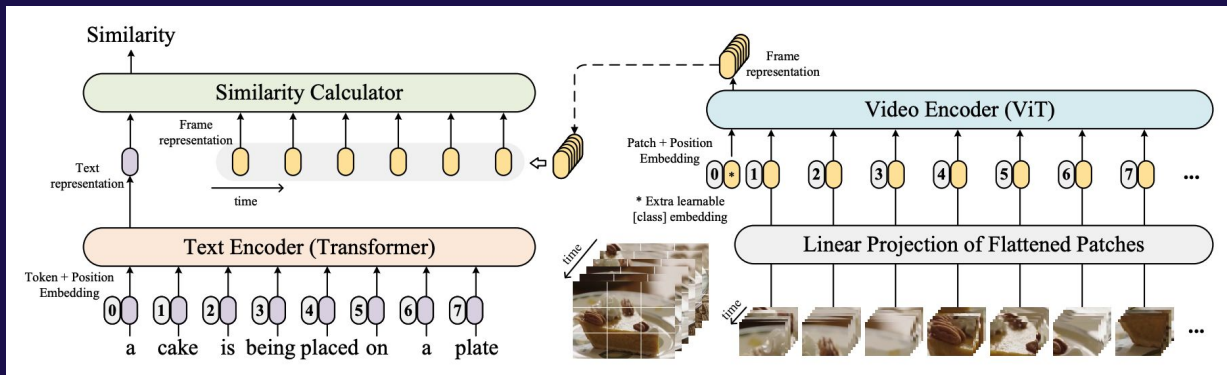
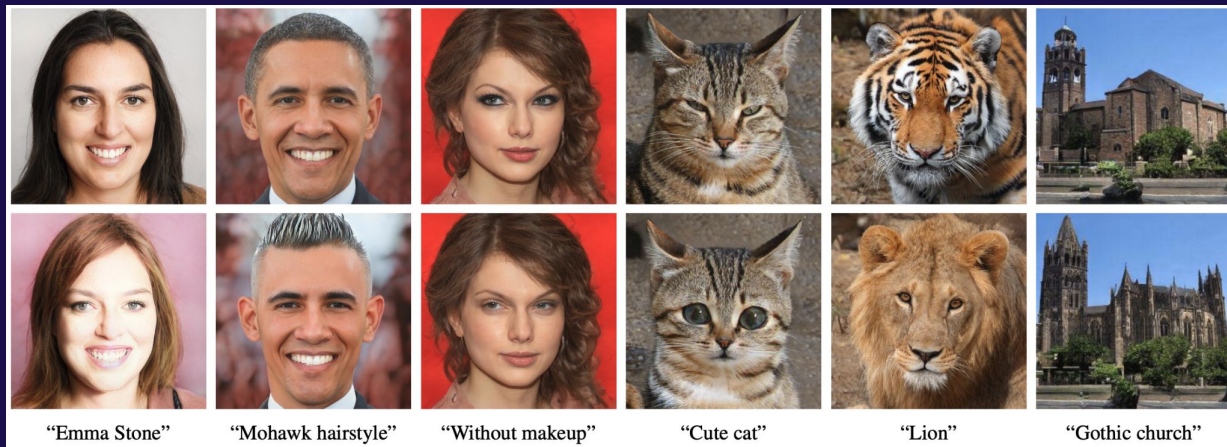
Applications of CLIP

StyleCLIP
(Patashnik et al.)

Steering a GAN Using CLIP

CLIP4Clip
(Luo & Ji, et al.)

Video retrieval using
CLIP features



More text-based image generations using CLIP



“A banquet hall”



“Geoffrey Hinton”

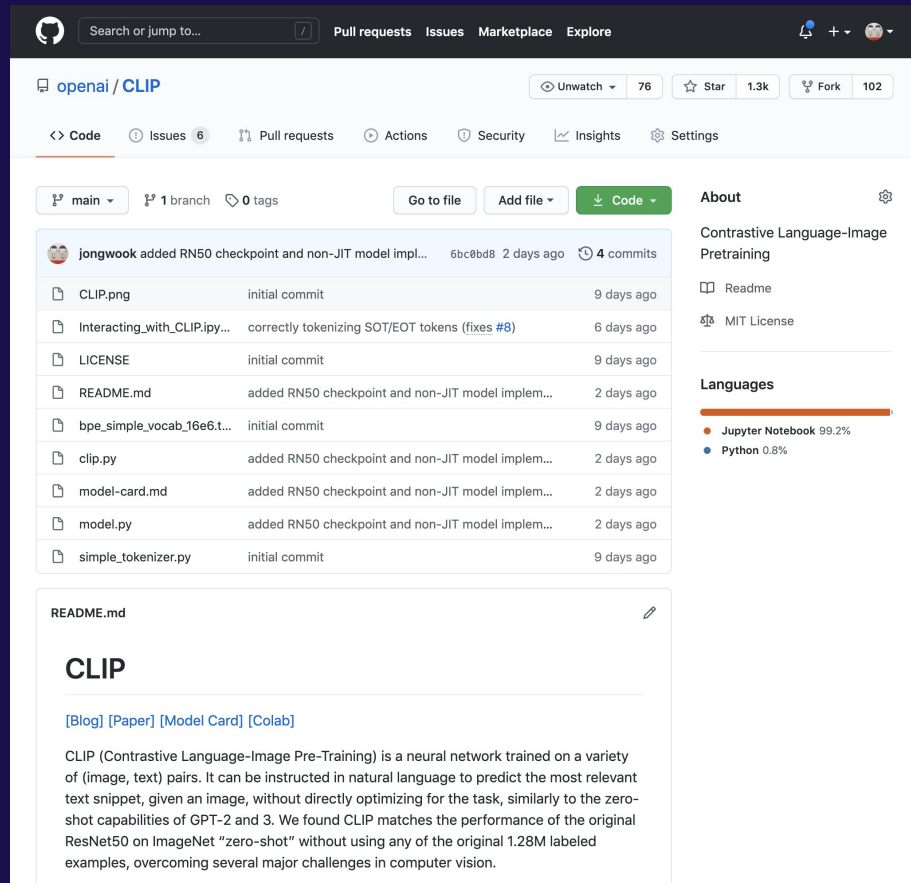


“Dogs playing poker”

Try CLIP today!

<https://github.com/openai/CLIP>

- PyTorch implementation
- Colab notebook



The screenshot shows the GitHub repository page for `openai/CLIP`. The repository has 76 stars, 1.3k forks, and 102 issues. The main branch is `main` with 1 branch and 0 tags. The repository contains several files, including `CLIP.png`, `Interacting_with_CLIP.ipynb`, `LICENSE`, `README.md`, `bpe_simple_vocab_16e6.txt`, `clip.py`, `model-card.md`, `model.py`, and `simple_tokenizer.py`. The `README.md` file is selected, showing the title `CLIP` and links to the [Blog], [Paper], [Model Card], and [Colab]. The text describes CLIP as a neural network trained on a variety of (image, text) pairs, capable of predicting the most relevant text snippet given an image, without directly optimizing for the task. It mentions that 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.

openai / CLIP

Unwatch 76 Star 1.3k Fork 102

<> Code Issues 6 Pull requests Actions Security Insights Settings

main 1 branch 0 tags Go to file Add file Code

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

File	Commit	Time
CLIP.png	initial commit	9 days ago
Interacting_with_CLIP.ipynb	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 impl...	2 days ago
bpe_simple_vocab_16e6.txt	initial commit	9 days ago
clip.py	added RN50 checkpoint and non-JIT model impl...	2 days ago
model-card.md	added RN50 checkpoint and non-JIT model impl...	2 days ago
model.py	added RN50 checkpoint and non-JIT model impl...	2 days ago
simple_tokenizer.py	initial commit	9 days ago

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.

About

Contrastive Language-Image Pretraining

Readme

MIT License

Languages

Jupyter Notebook 99.2%

Python 0.8%

Thank You

Visit openai.com for more information.

FOLLOW @OPENAI ON TWITTER
WE ARE HIRING!