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

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

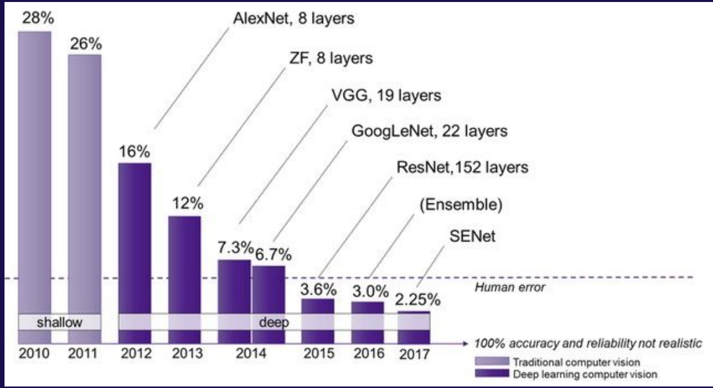
Vision models led the deep learning boom

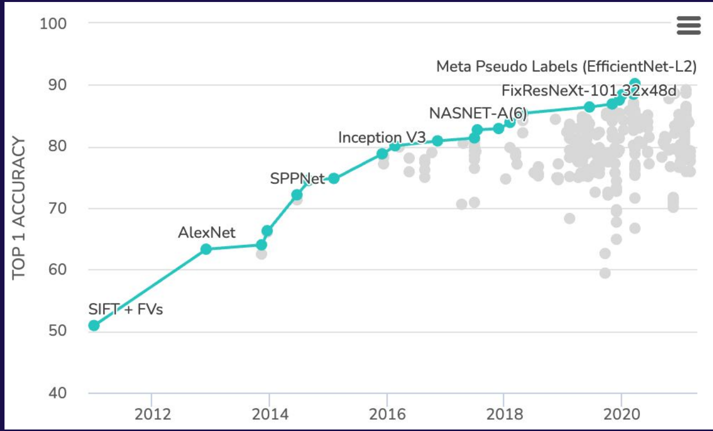
ImageNet competition

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

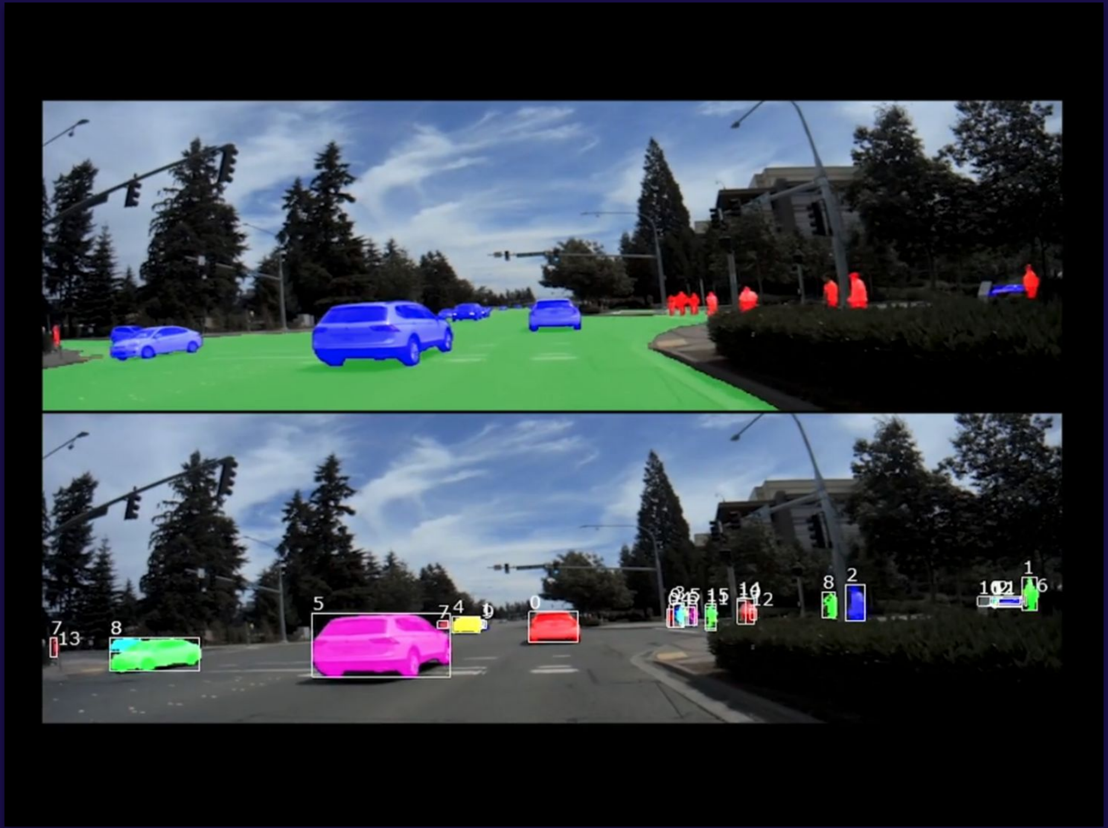
Human top-5 accuracy: 5%

Top-1 as an ongoing benchmark





Vision models today



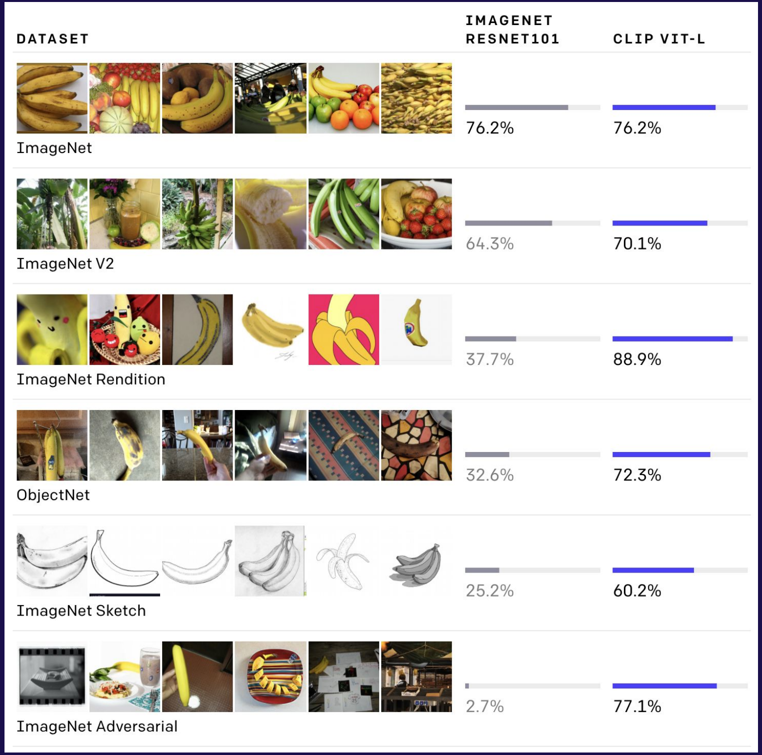
What makes CLIP special?

Motivation:

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

Result:

Robust zero-shot inference
Multimodal feature space

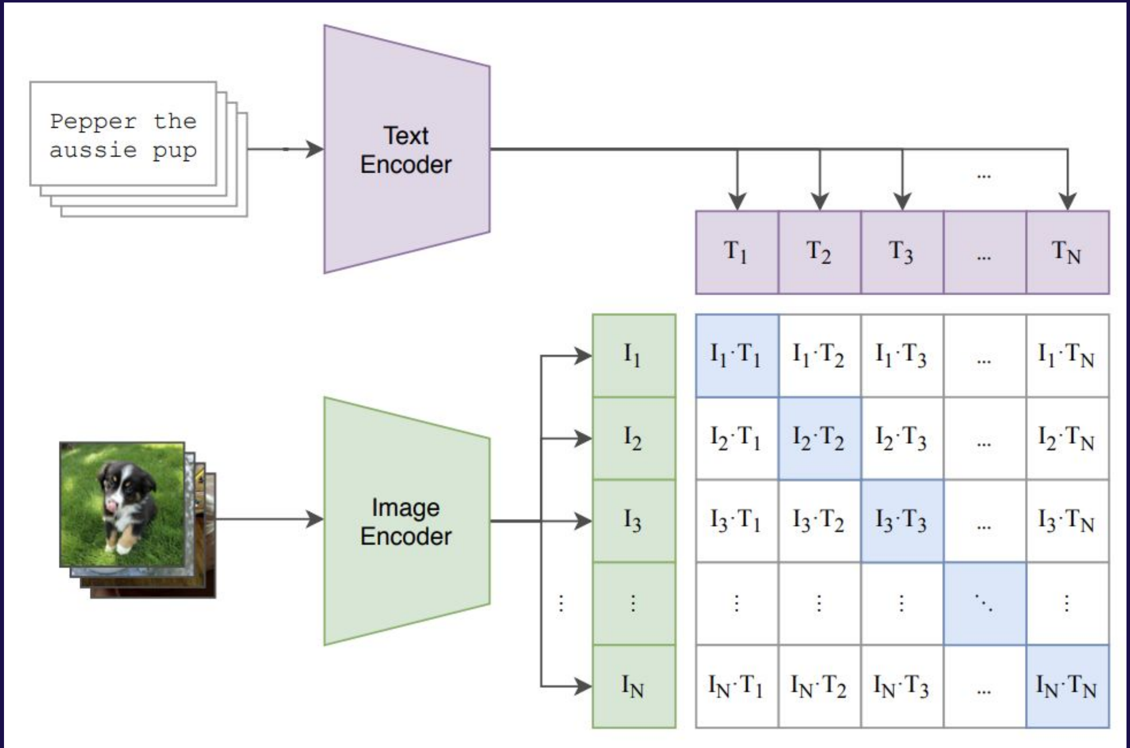


How does it work?

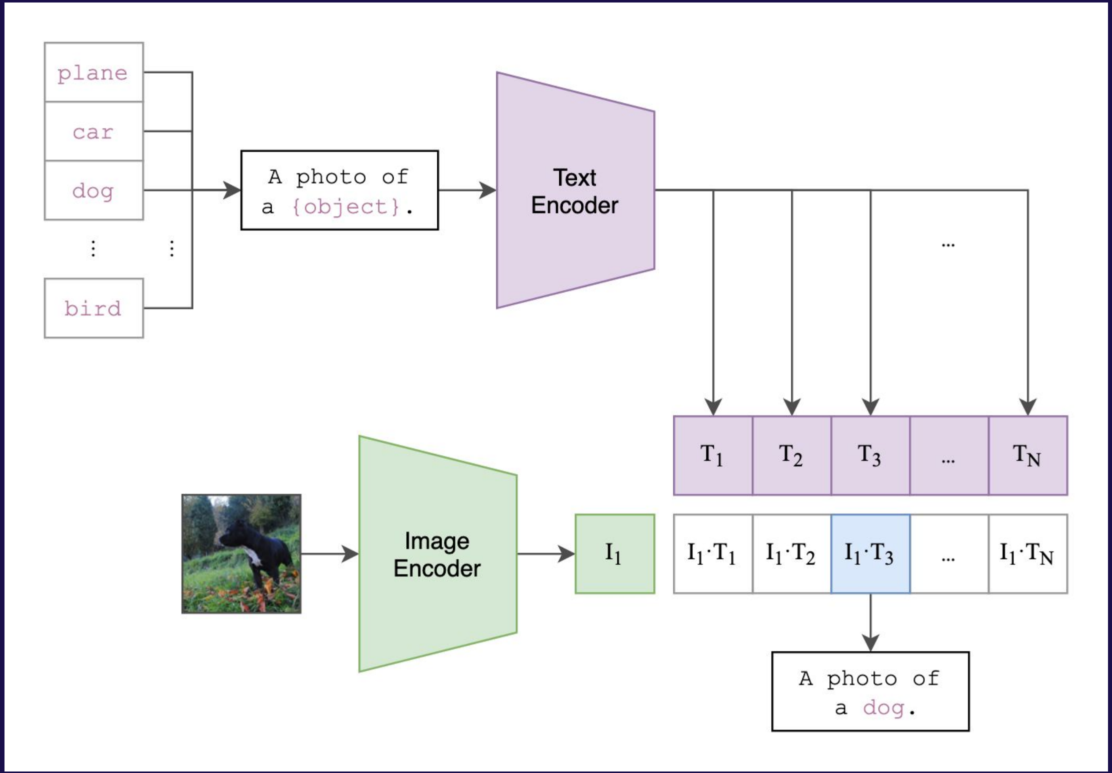
    

Pig Tiger Panda Hippo Camel

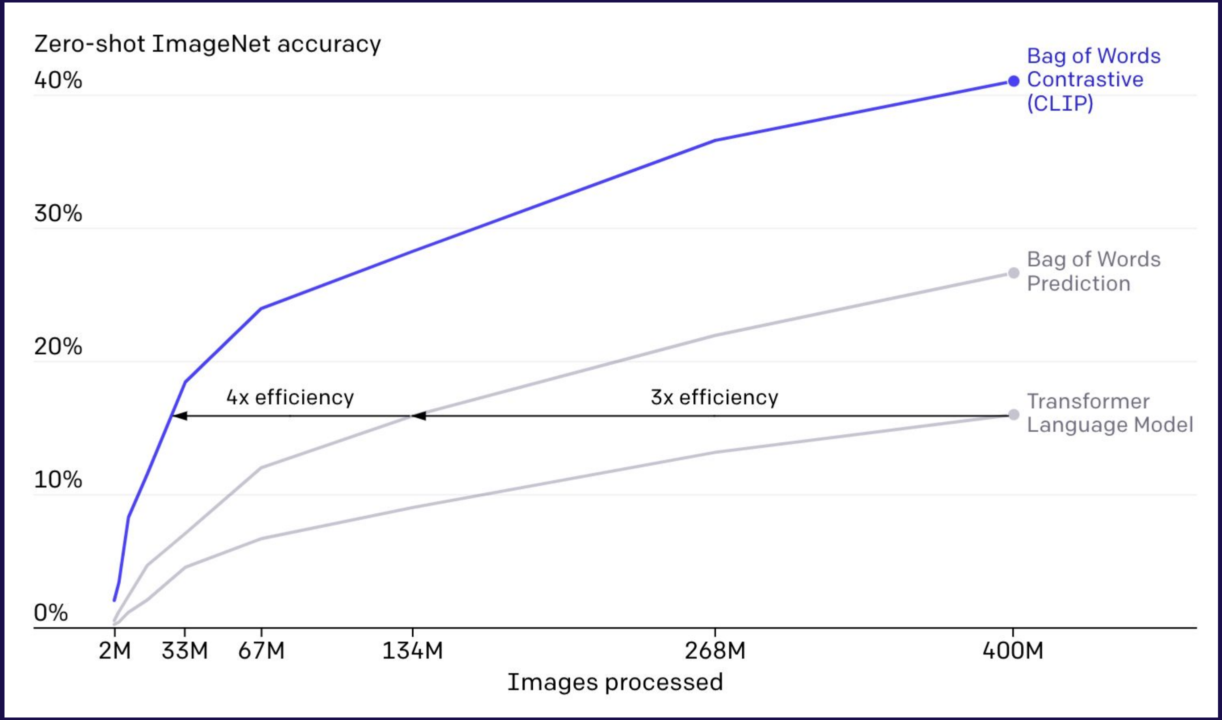
CLIP: Contrastive Language-Image Pre-training



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

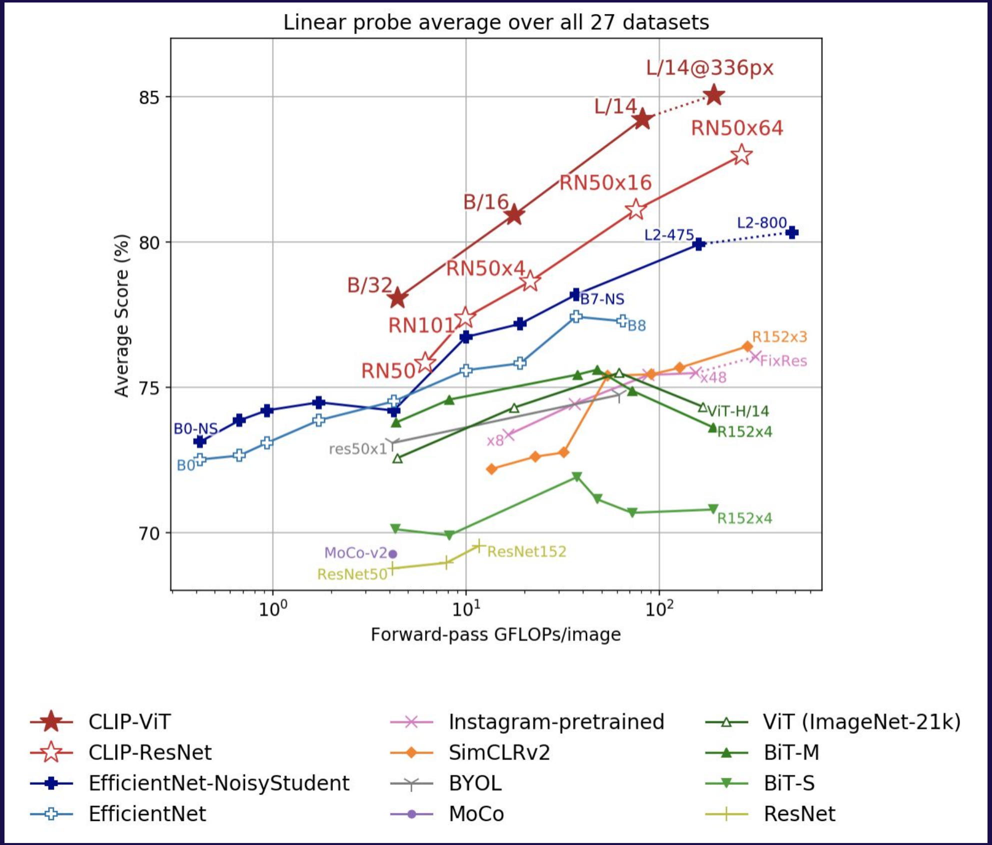
Logistic regression classifier on image features

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

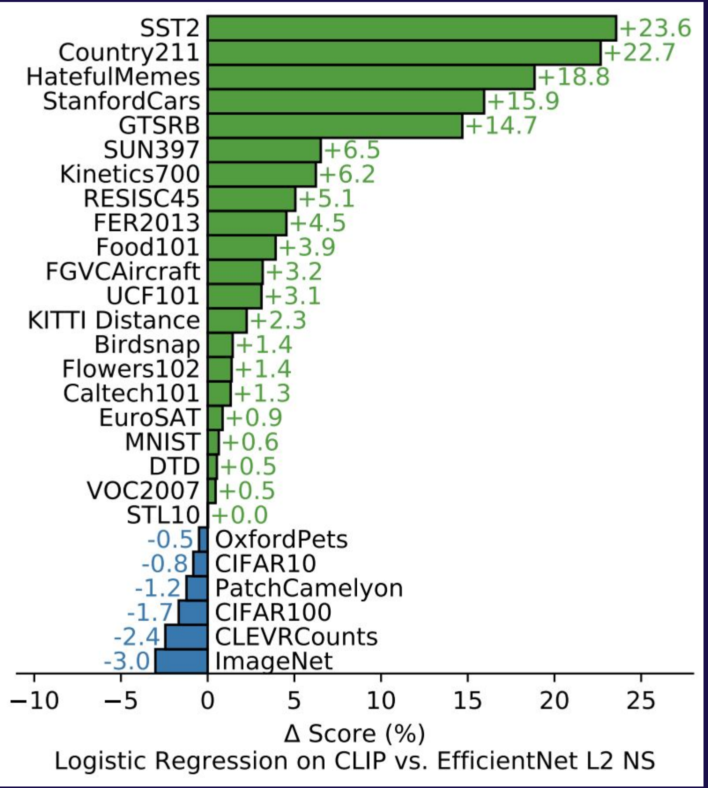
Evaluated on 27 image datasets × 65 vision models

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

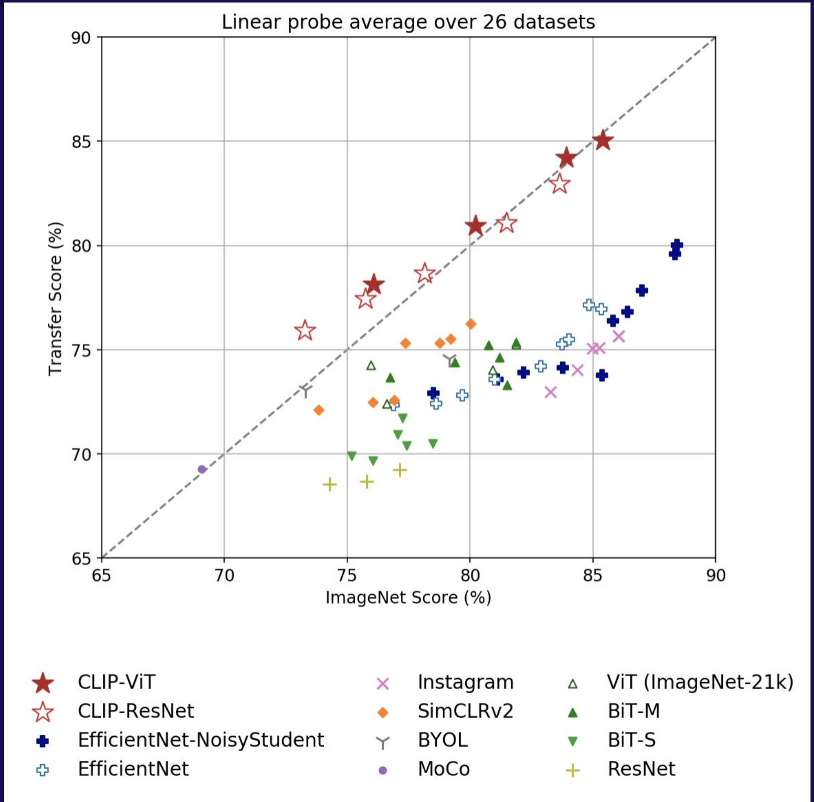
Linear probe performance vs SOTA vision models



Linear-probe CLIP vs Linear-probe EfficientNet-L2

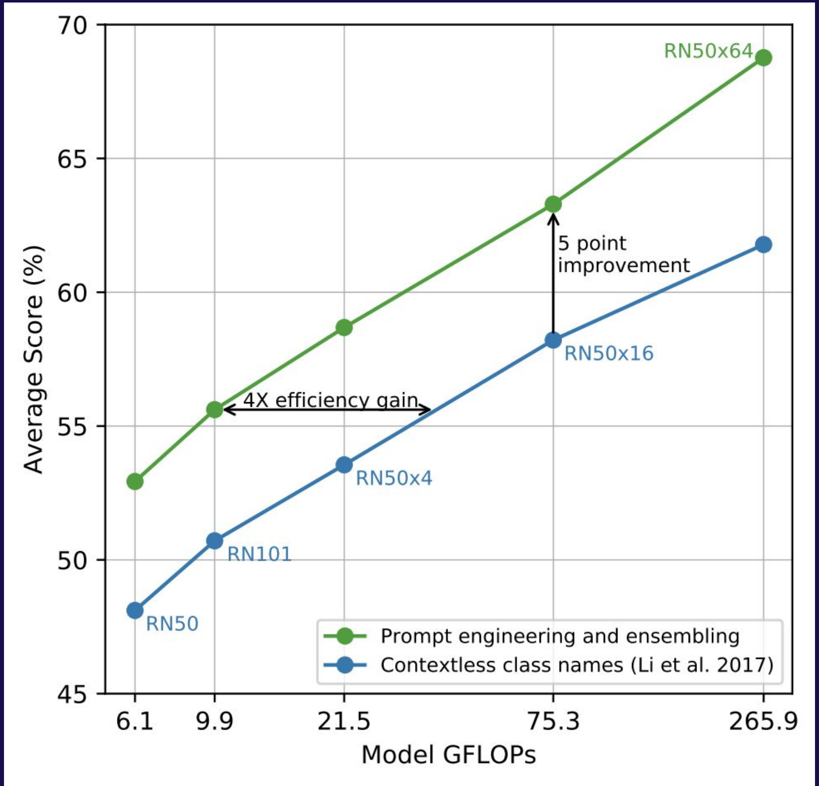


vs ImageNet score



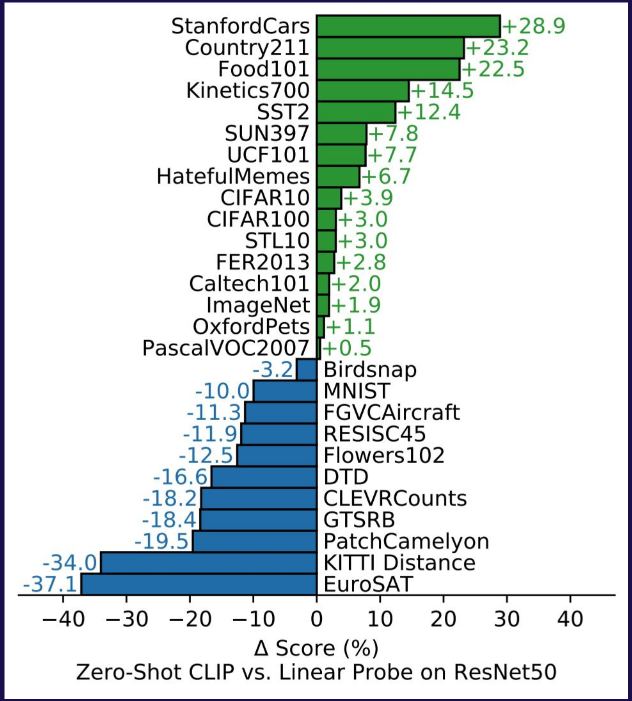
Zero-Shot Transfer

Prompt engineering



Zero-shot vs Linear-probe ResNet-50

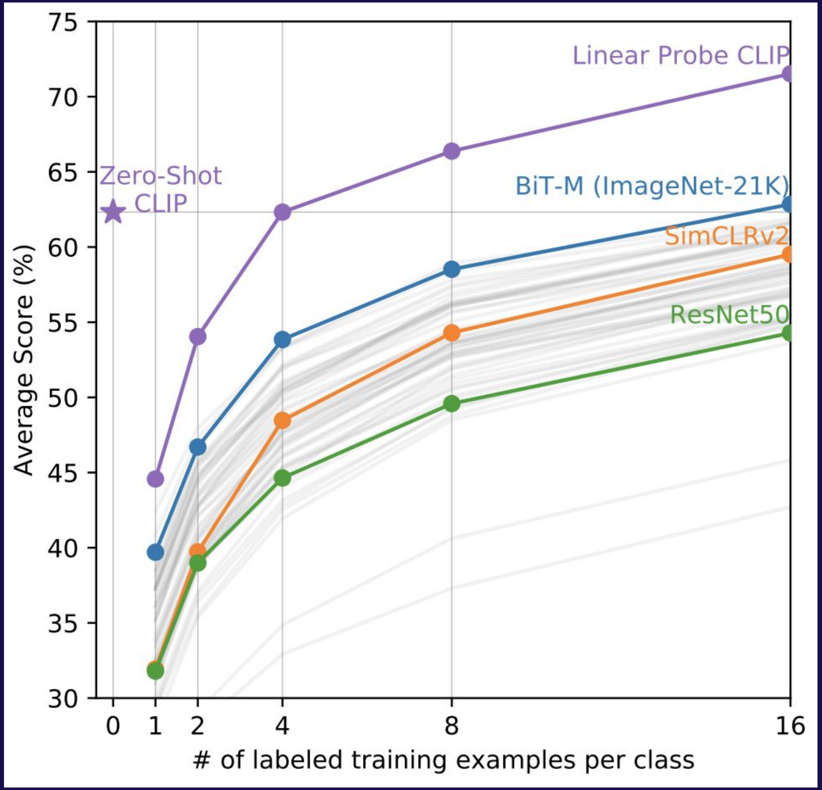
Zero-shot CLIP outperforms ResNet-50 on 16 of 27 datasets



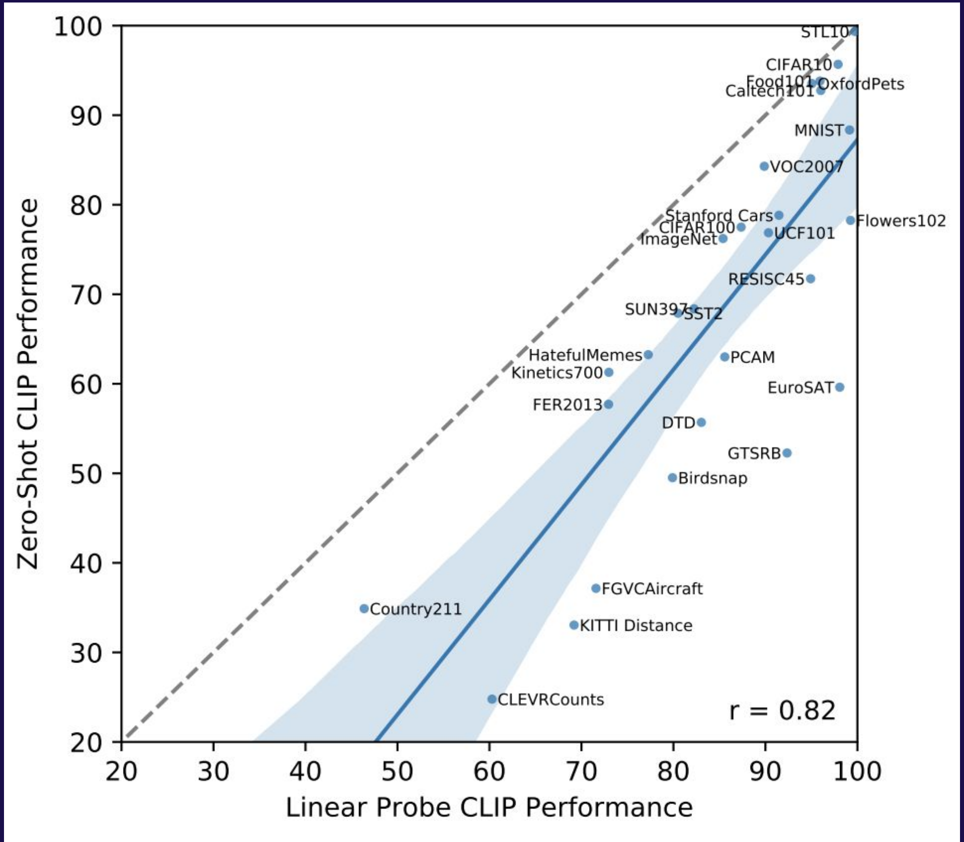
Zero-shot CLIP vs Few-shot linear probes

Zero-shot CLIP is as good as

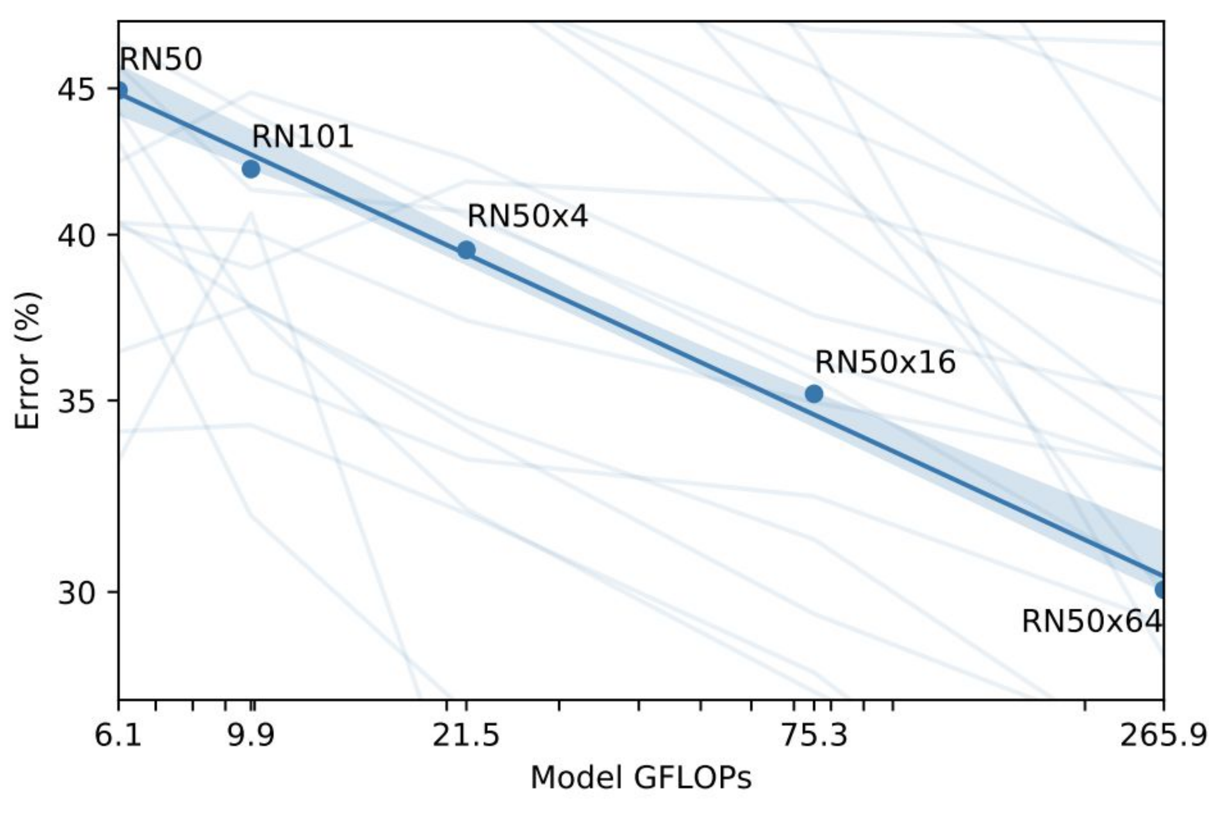
- 4-shot linear-probe CLIP
- 16-shot BiT-M



Zero-shot vs Linear-probe CLIP



Zero-shot performance vs model size



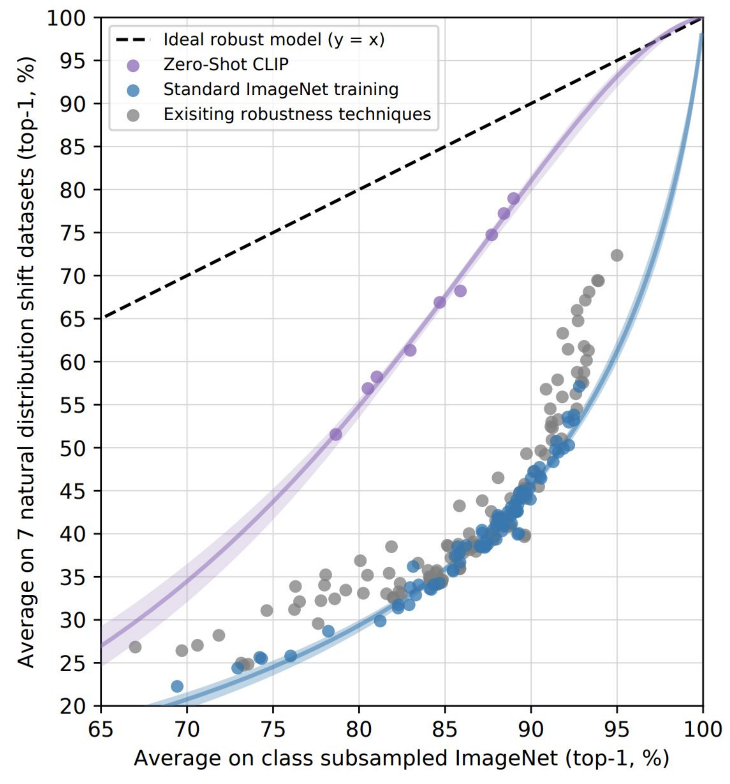
Robustness to Natural Distribution Shift

Robustness to natural distribution shift

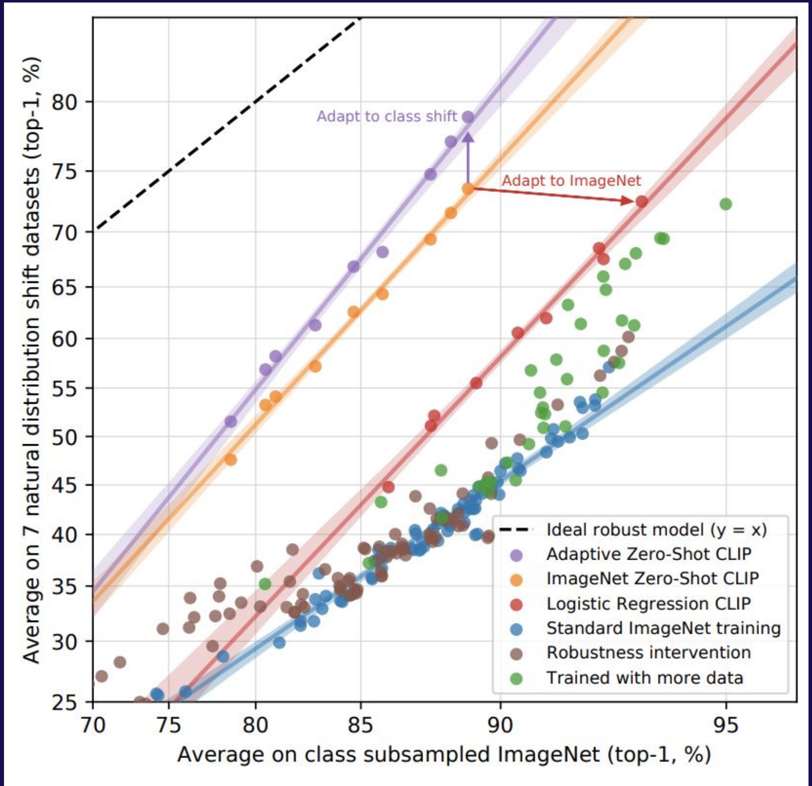
CLIP is significantly more robust!

7 ImageNet-like Datasets (Taori et al.)

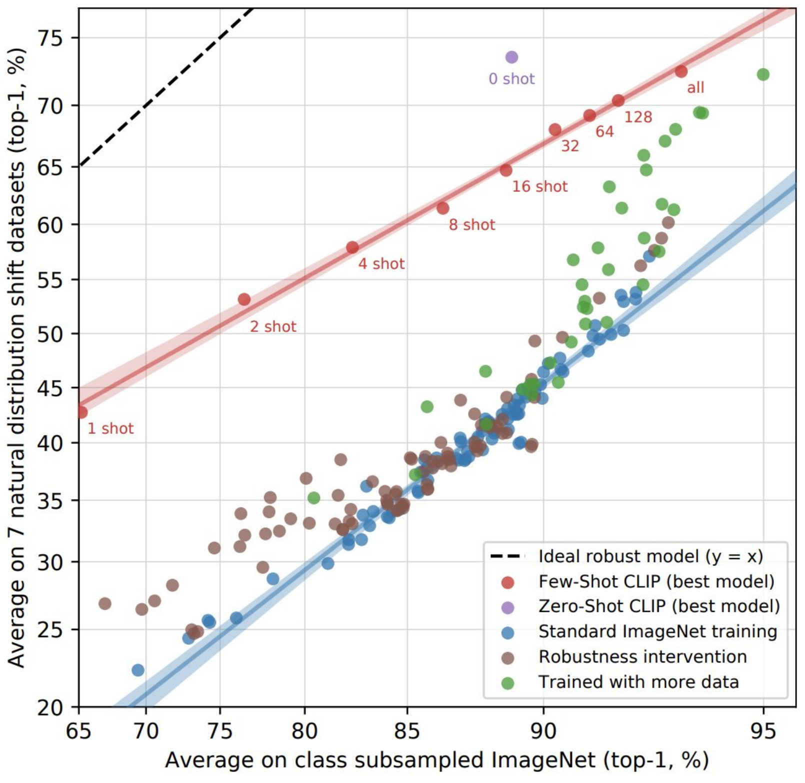
- ImageNetV2
- ImageNet-A
- ImageNet-R
- ImageNet Sketch
- ObjectNet
- ImageNet Vid
- Youtube-BB



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

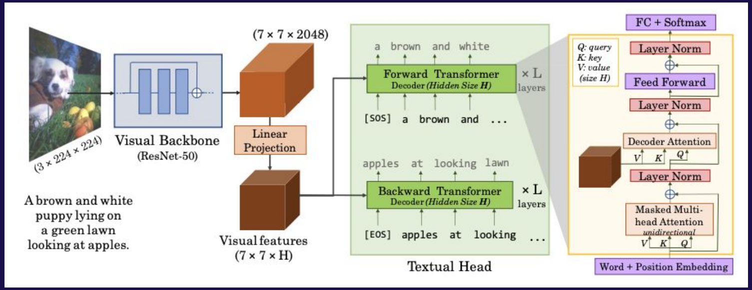
Not comprehensive, continuing to research to ensure safety
Model card limits usage of CLIP to research-only

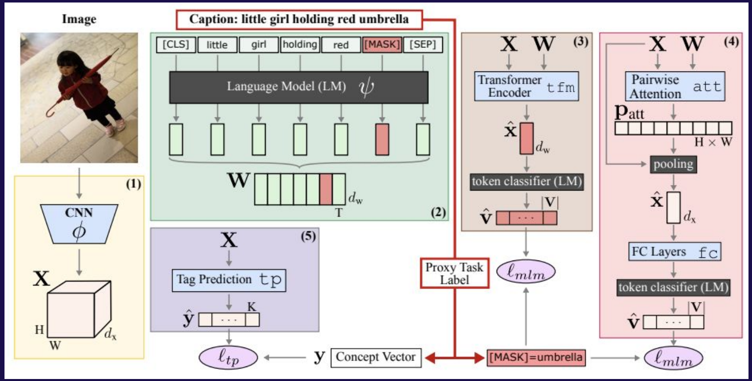
Related Work

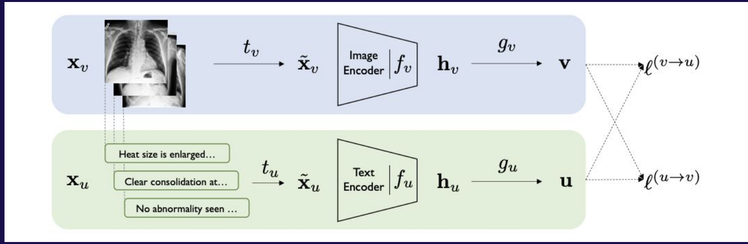
Prior Related Work

Multimodal learning
- VirTex
- ICMLM
- ConVIRT

Natural language supervision
Text-image retrieval
Webly supervised learning



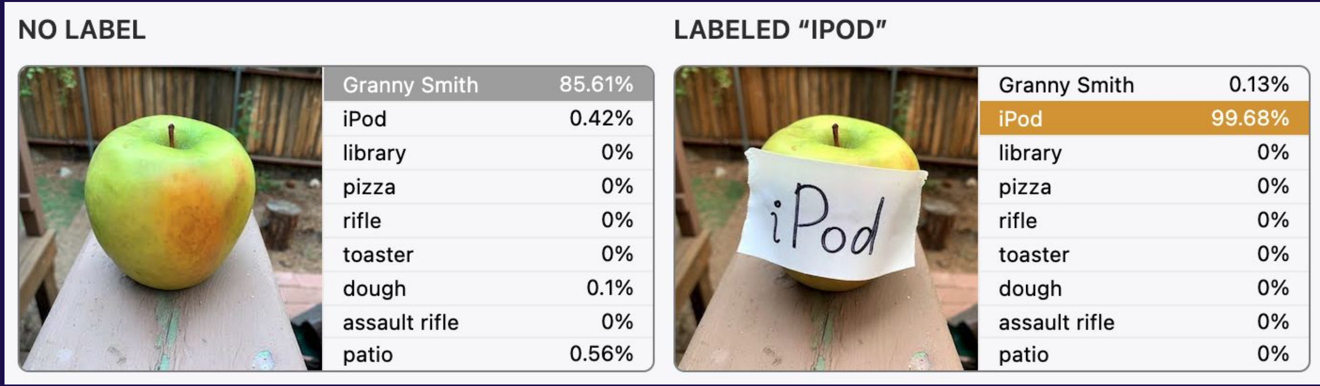


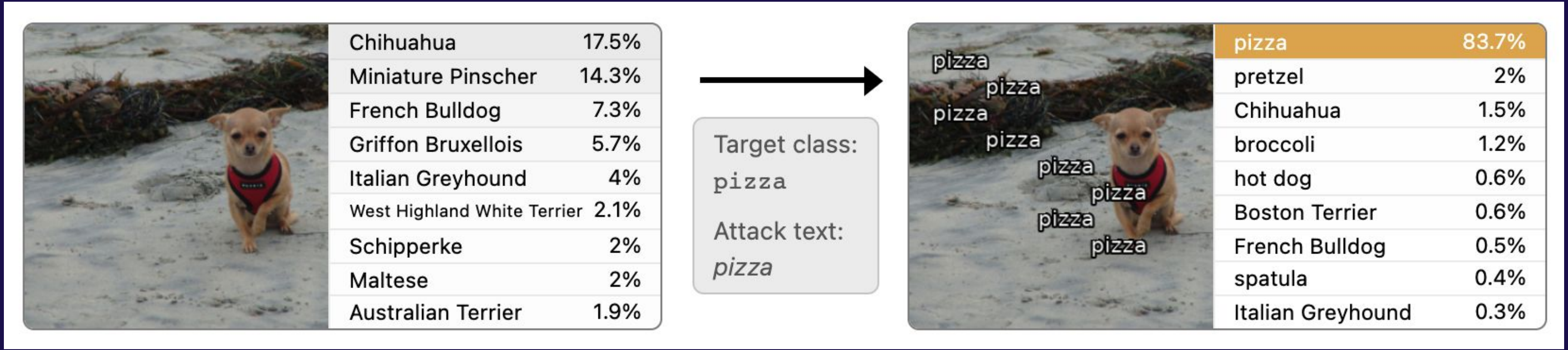


Multimodal Neurons in CLIP

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

Typographic Attacks

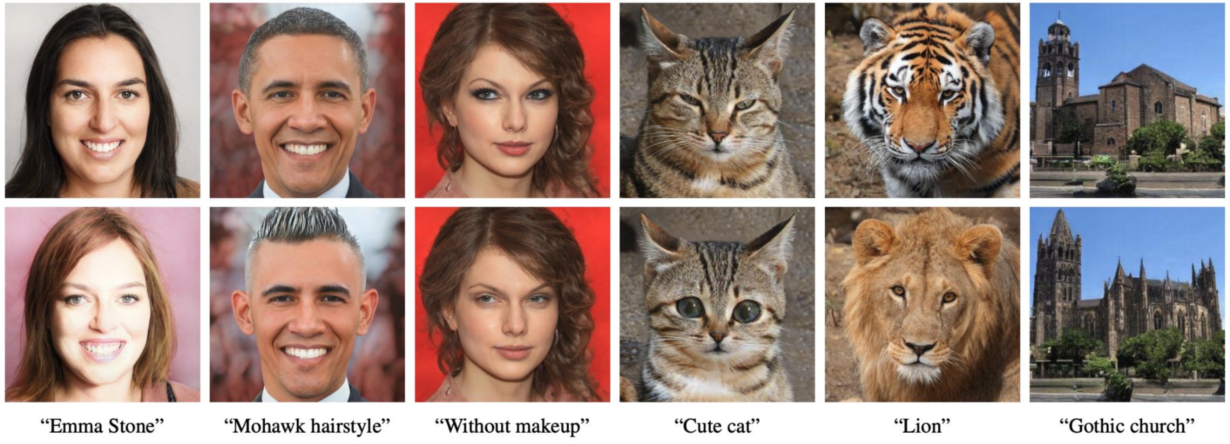




Applications of CLIP

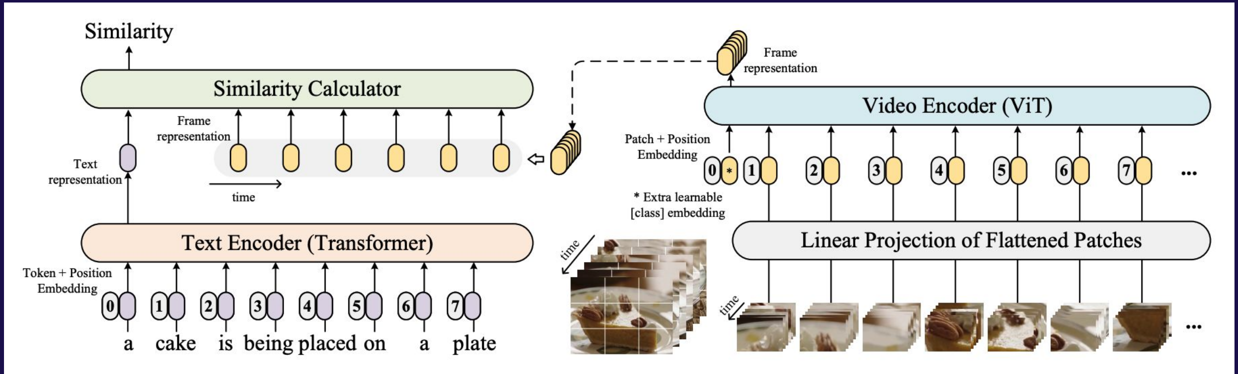
StyleCLIP
(Patashnik et al.)

Steering a GAN Using CLIP



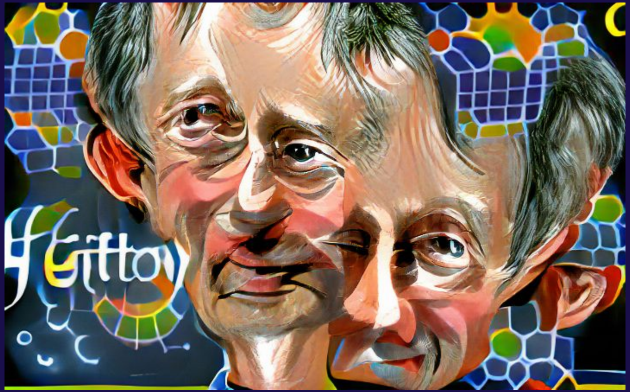
CLIP4Clip
(Luo & Ji, et al.)

Video retrieval using
CLIP features



More text-based image generations using CLIP

“A banquet hall”



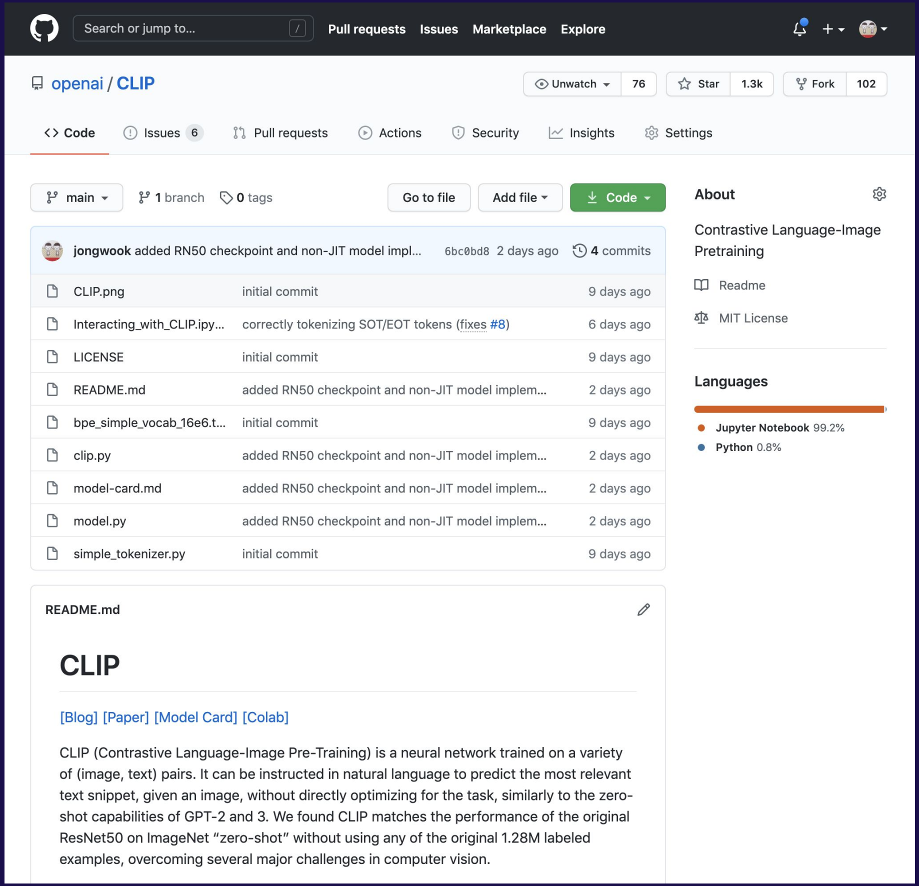
“Geoffrey Hinton”

“Dogs playing poker”

Try CLIP today!

https://github.com/openai/CLIP

- PyTorch implementation
- Colab notebook



Thank You

Visit openai.com for more information.

FOLLOW @OPENAI ON TWITTER
WE ARE HIRING!