OpenAI

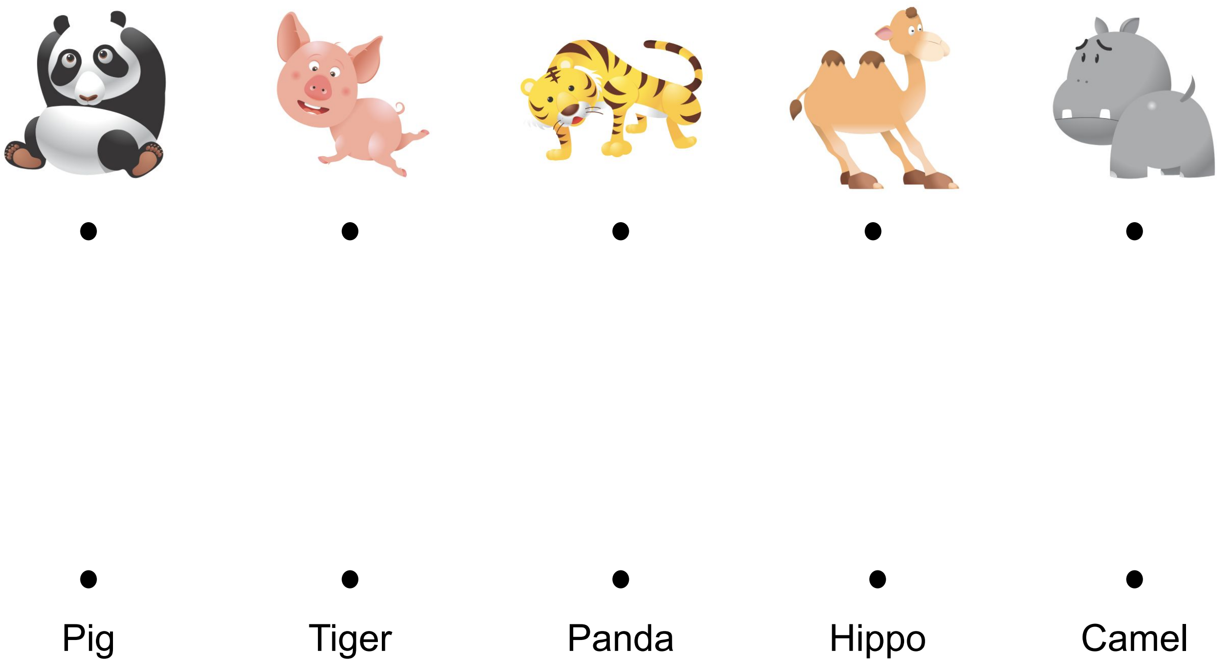
Learning Transferable Visual Models From Natural Language Supervision

ICML 2021

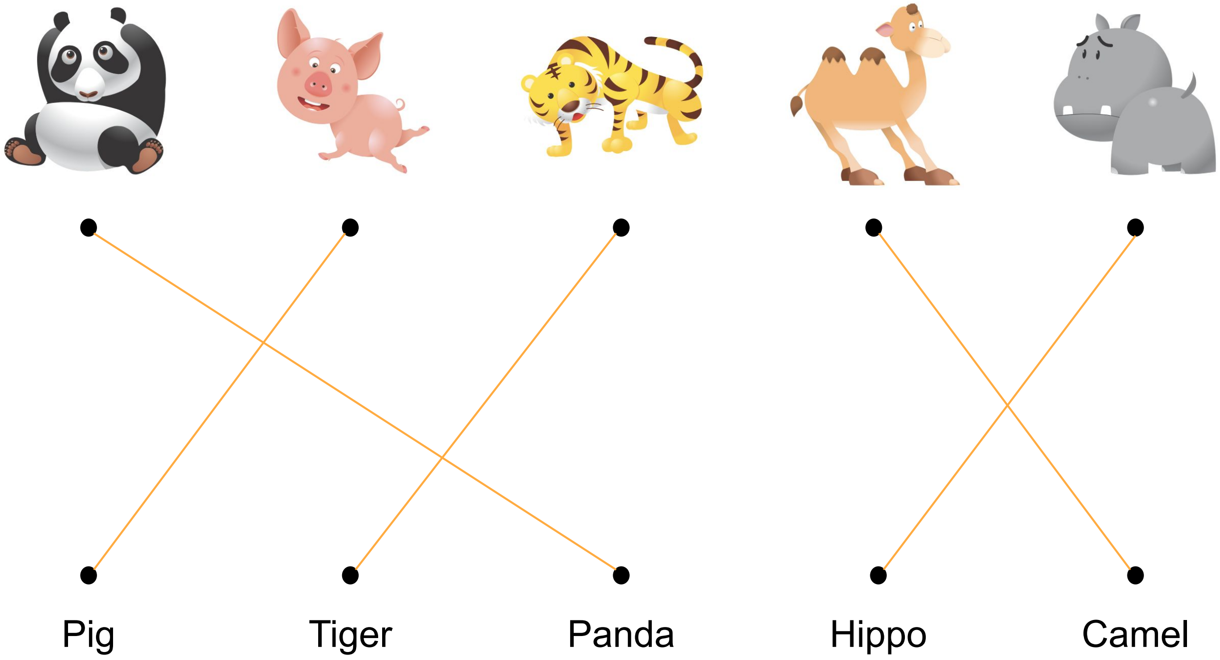
Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever
OpenAI

Contrastive learning

Contrastive learning

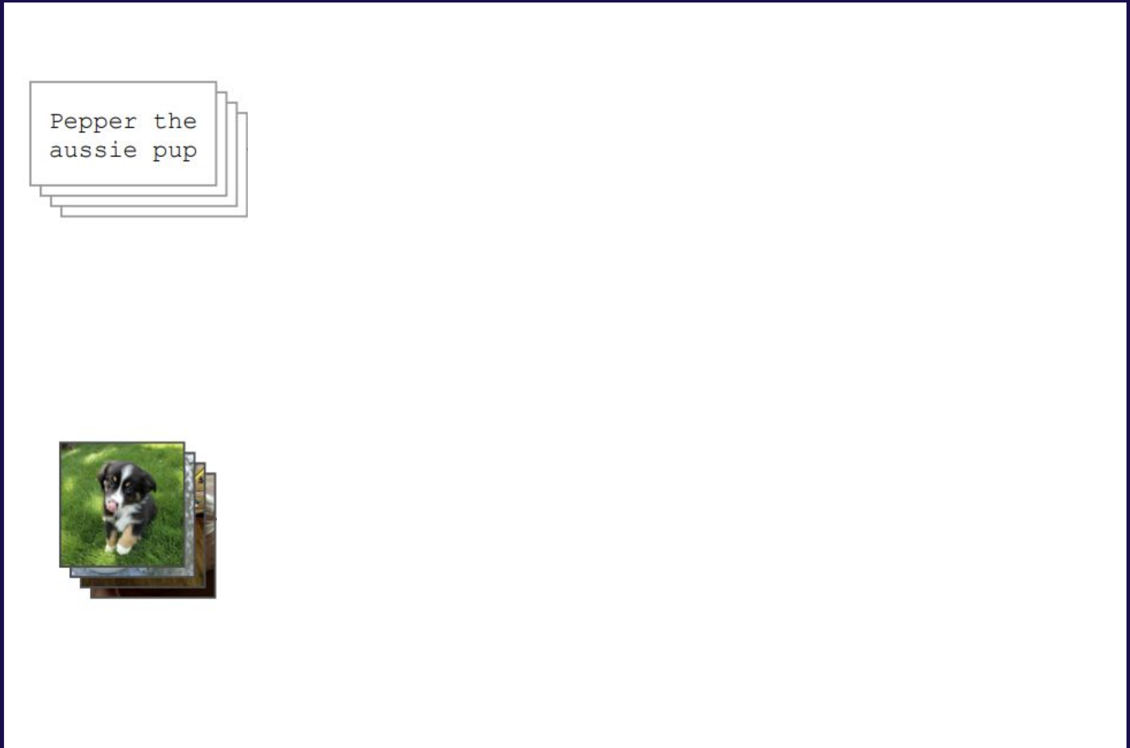


Contrastive learning

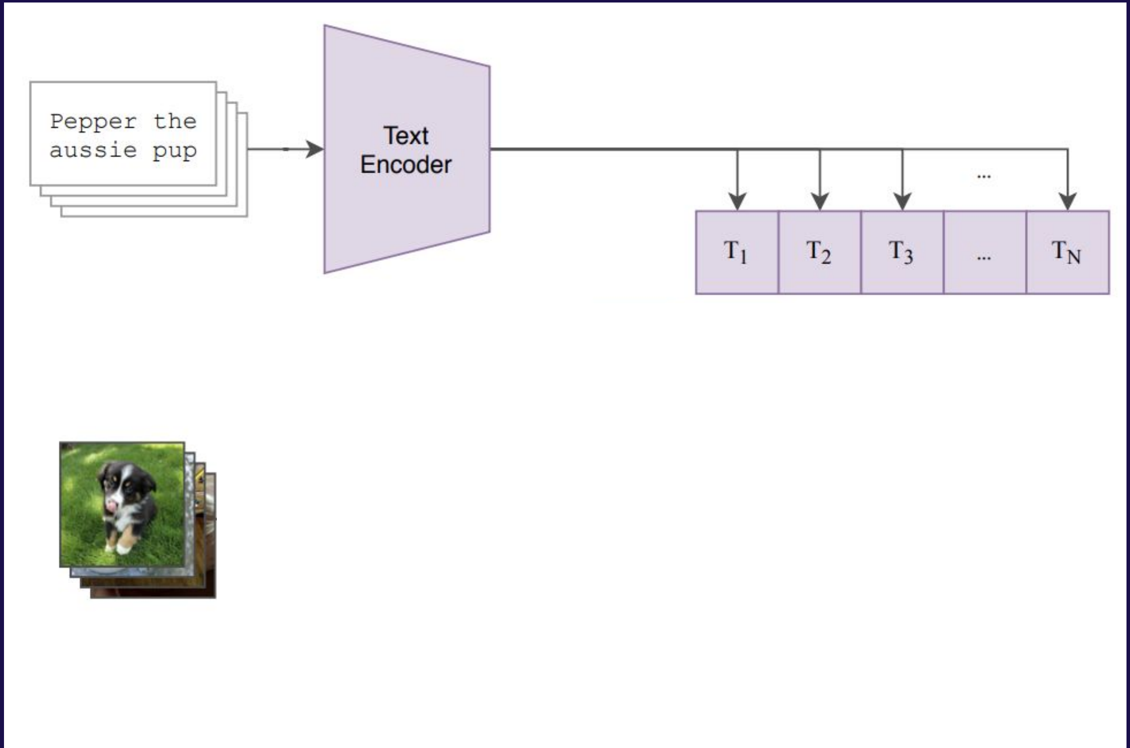


CLIP: Contrastive Language-Image Pre-training

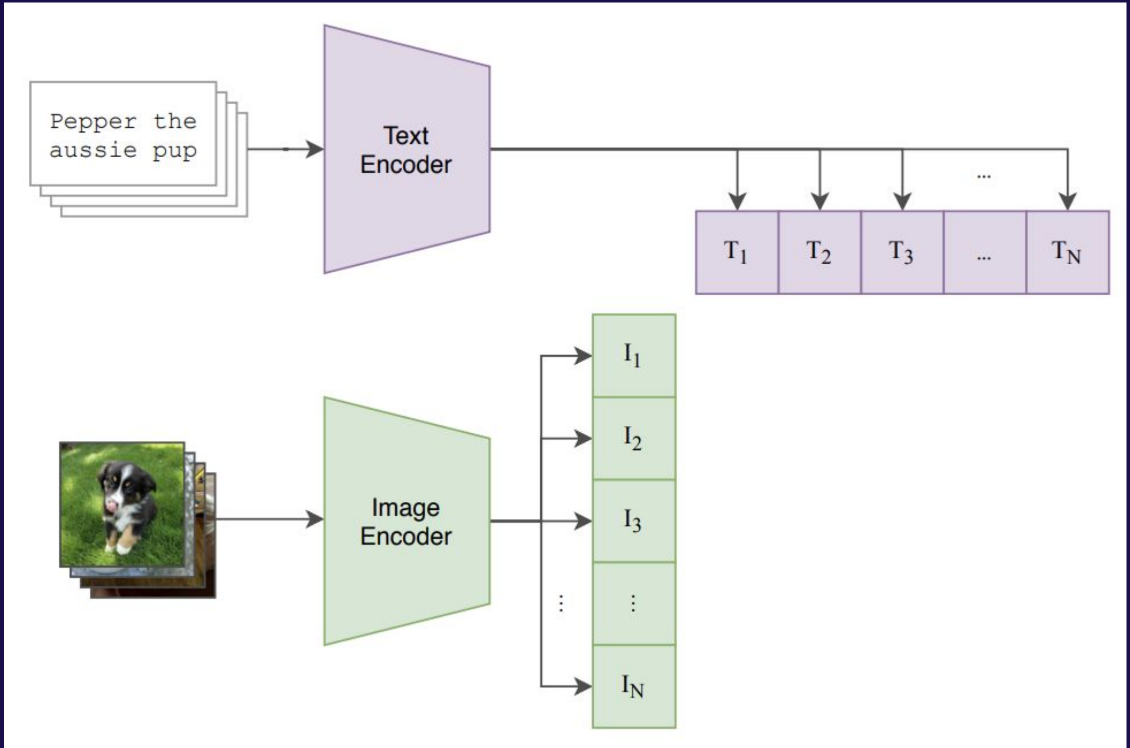
CLIP: Contrastive Language-Image Pre-training



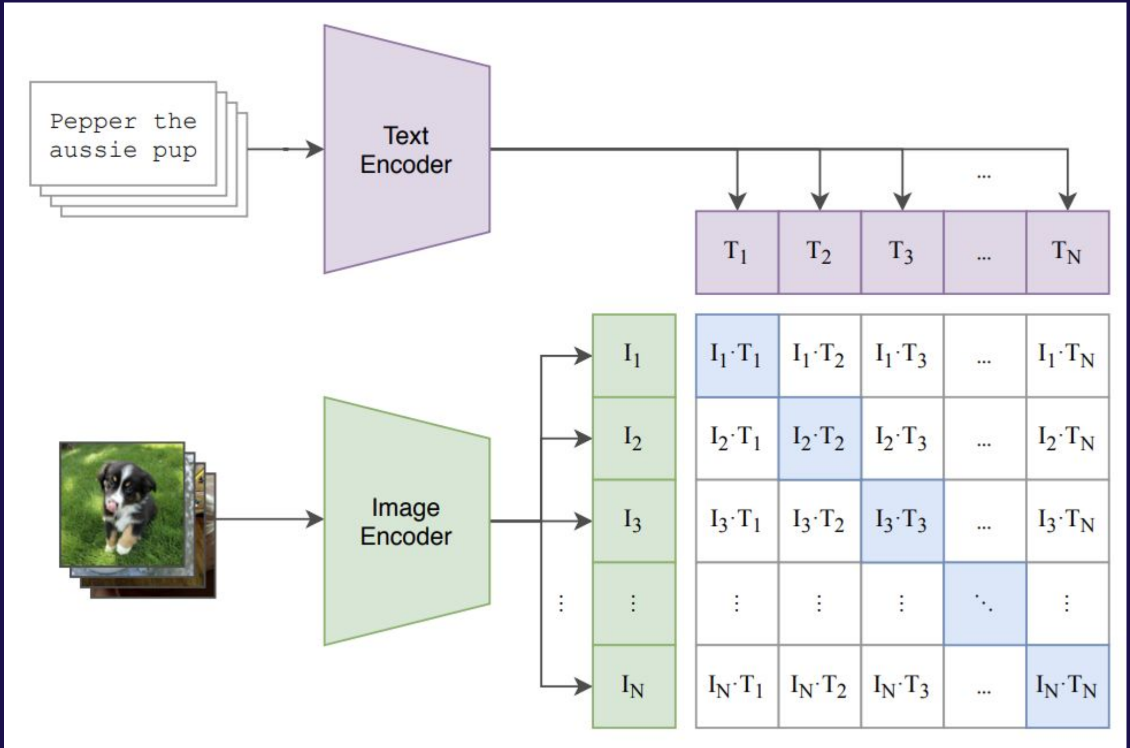
CLIP: Contrastive Language-Image Pre-training



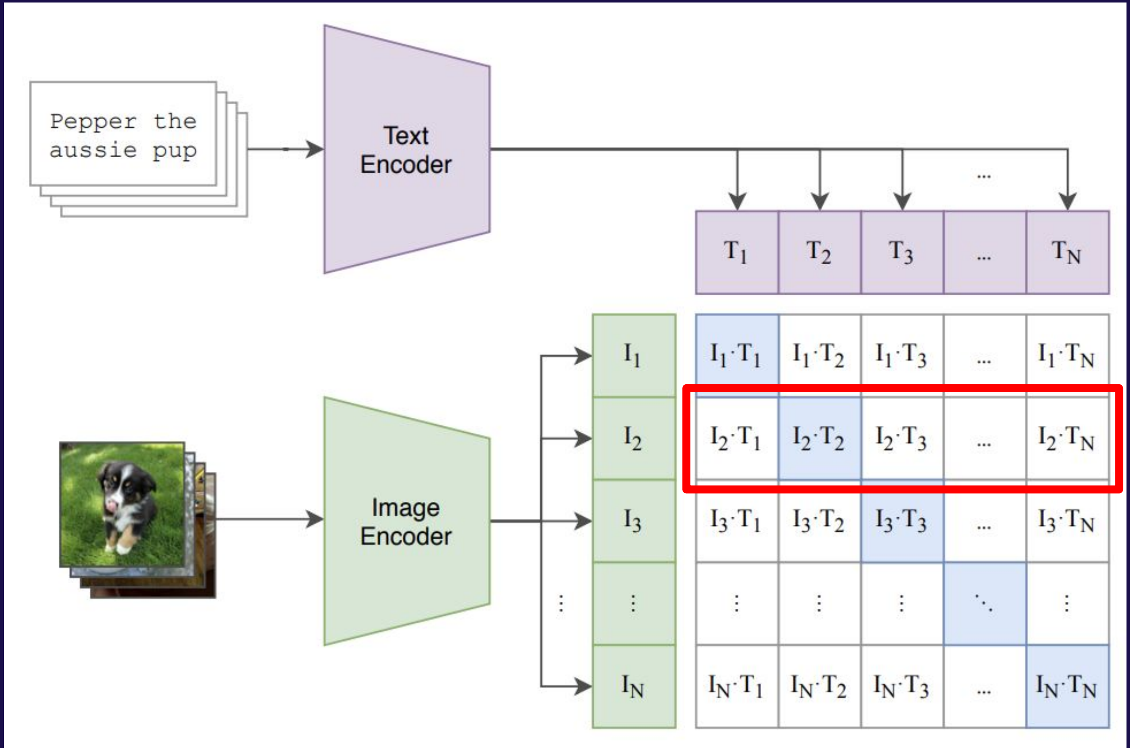
CLIP: Contrastive Language-Image Pre-training



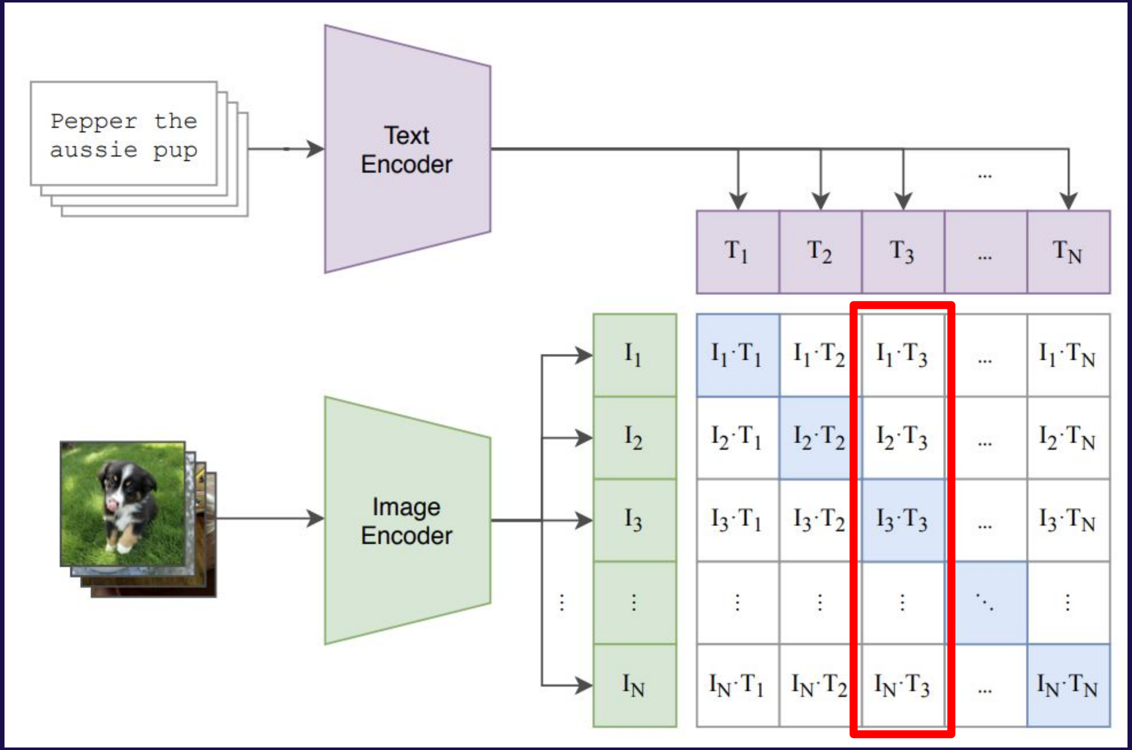
CLIP: Contrastive Language-Image Pre-training



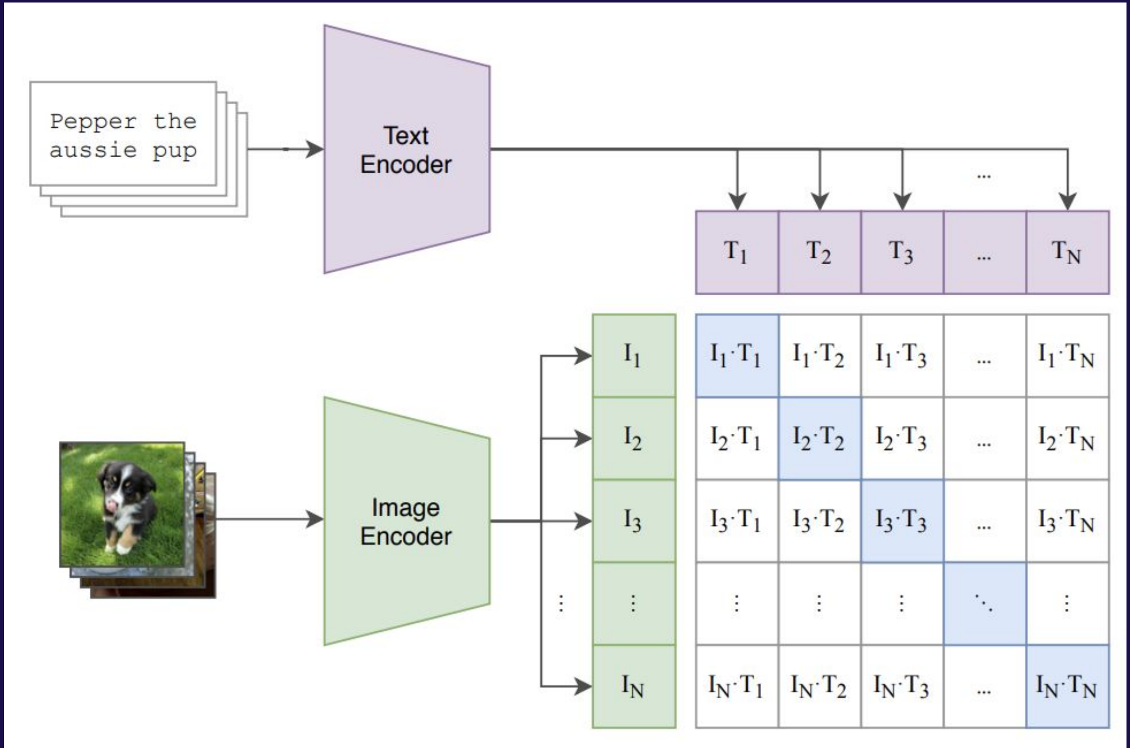
CLIP: Contrastive Language-Image Pre-training



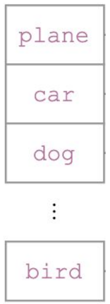
CLIP: Contrastive Language-Image Pre-training



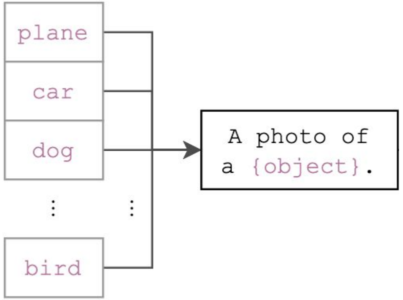
CLIP: Contrastive Language-Image Pre-training



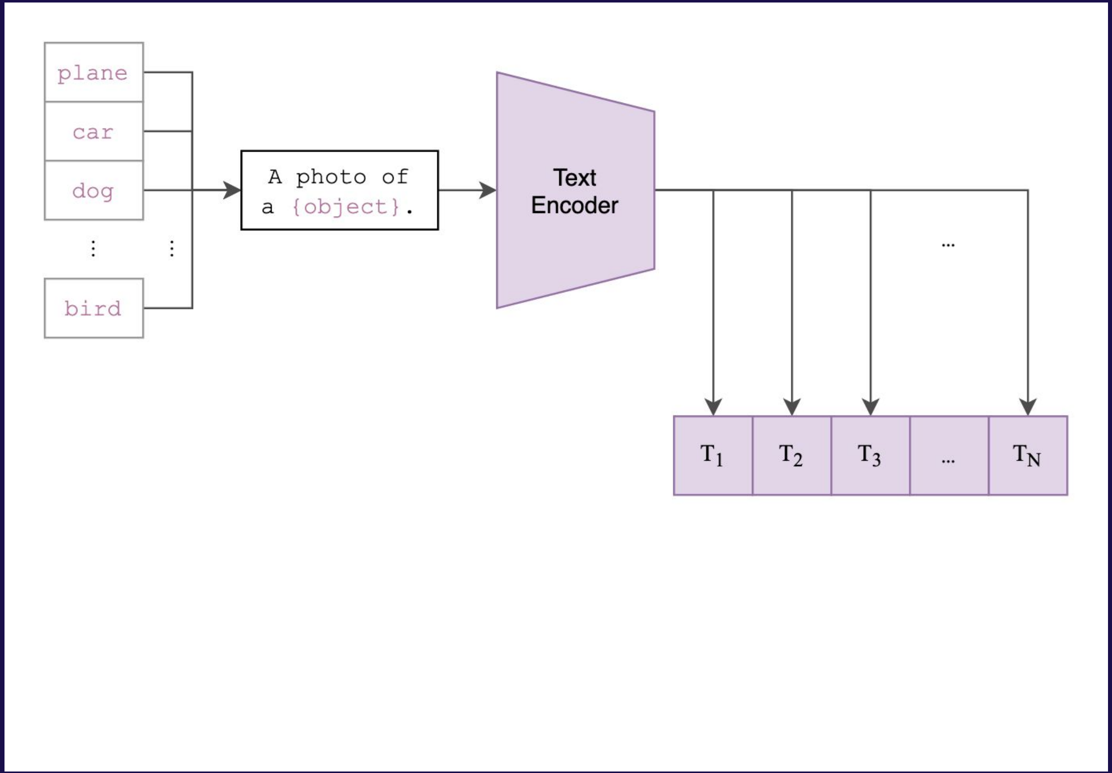
Zero-shot image classification



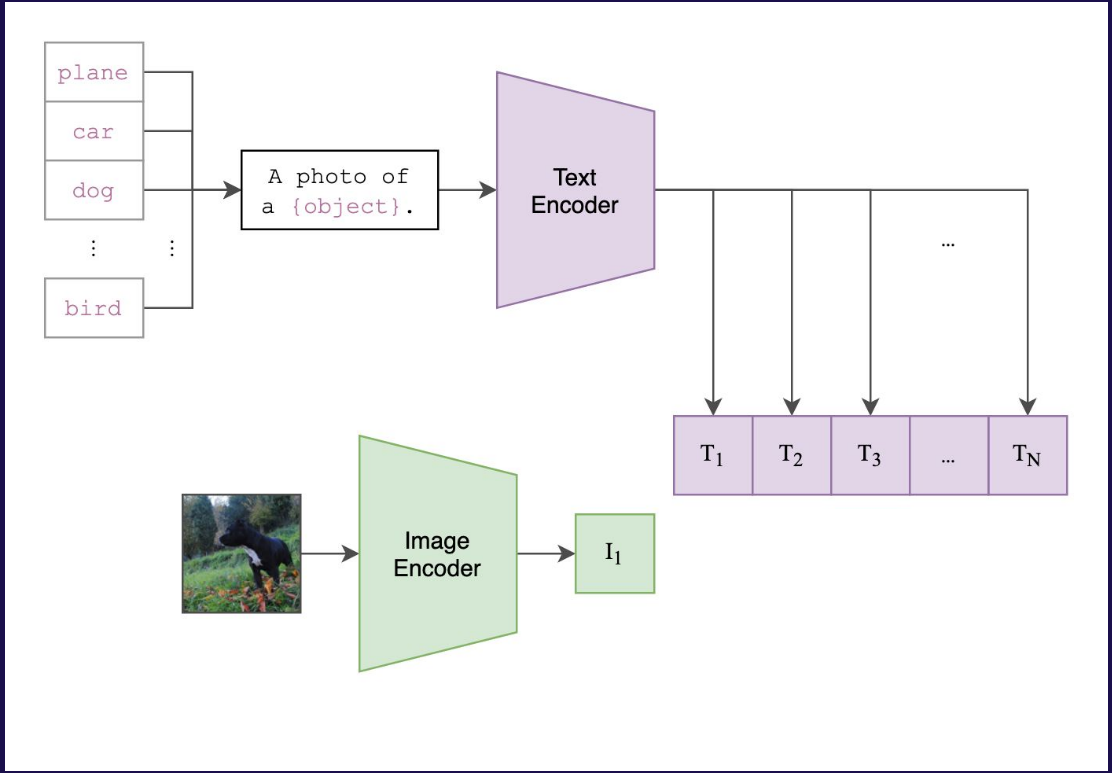
Zero-shot image classification



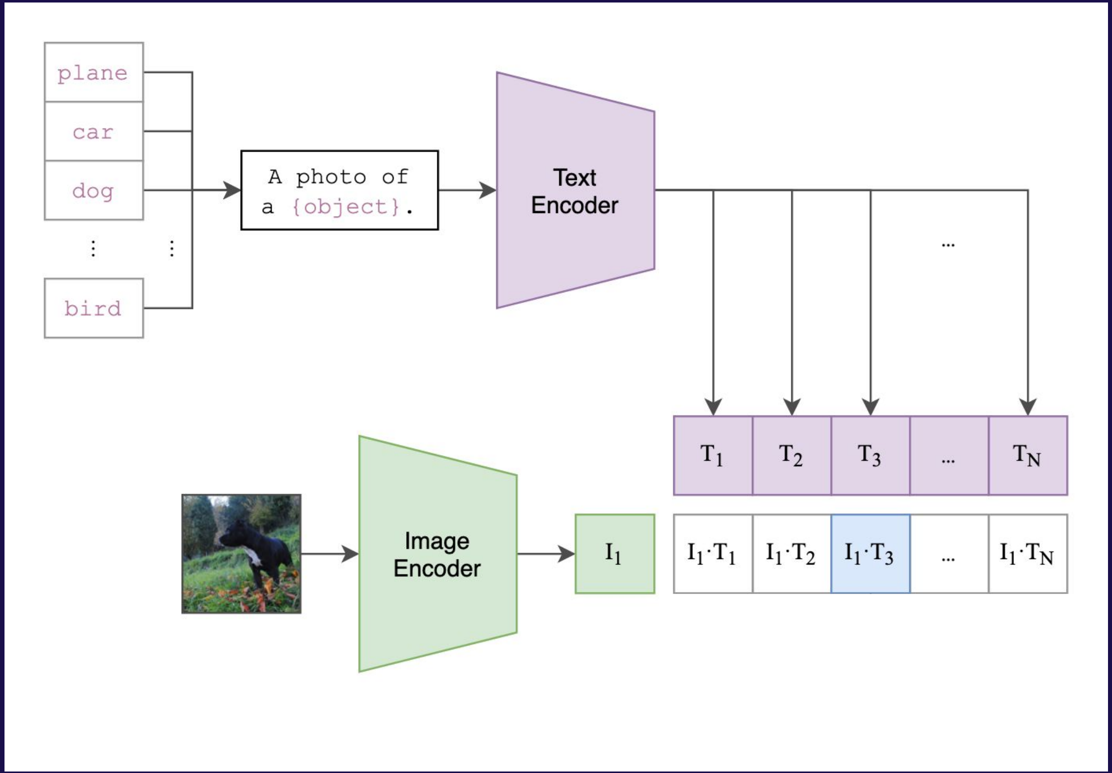
Zero-shot image classification



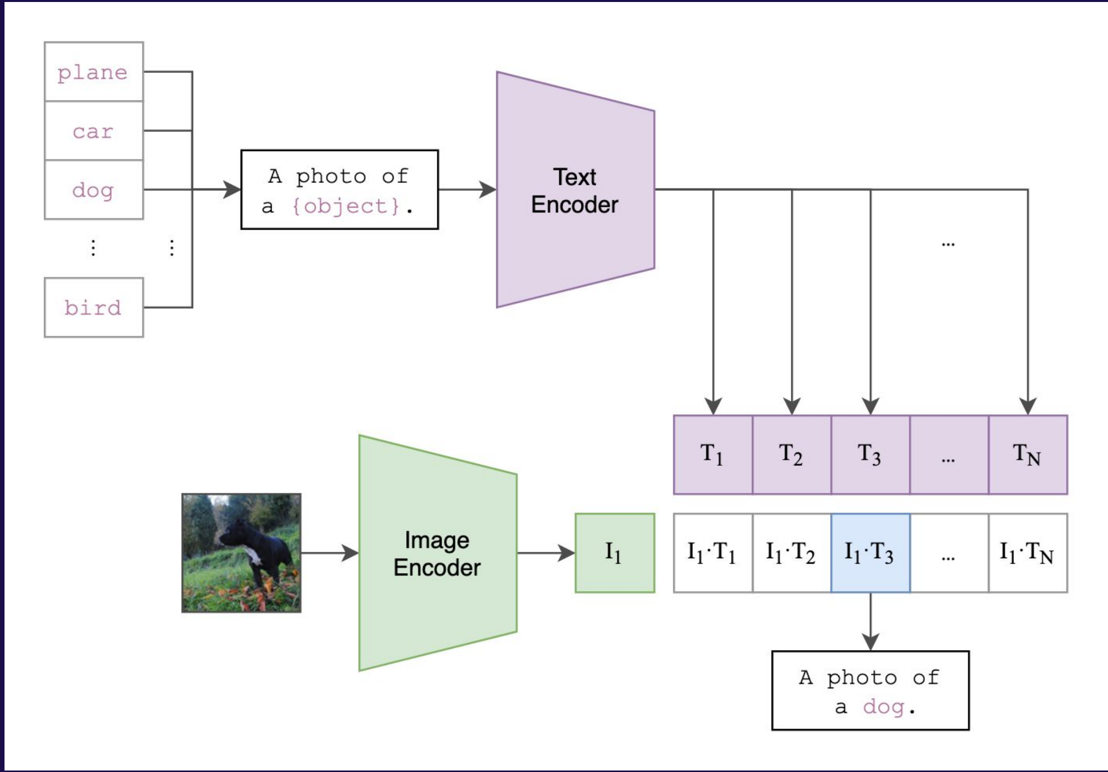
Zero-shot image classification



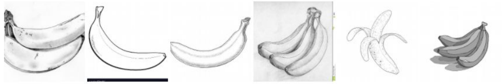
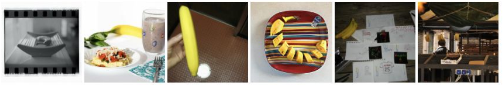
Zero-shot image classification



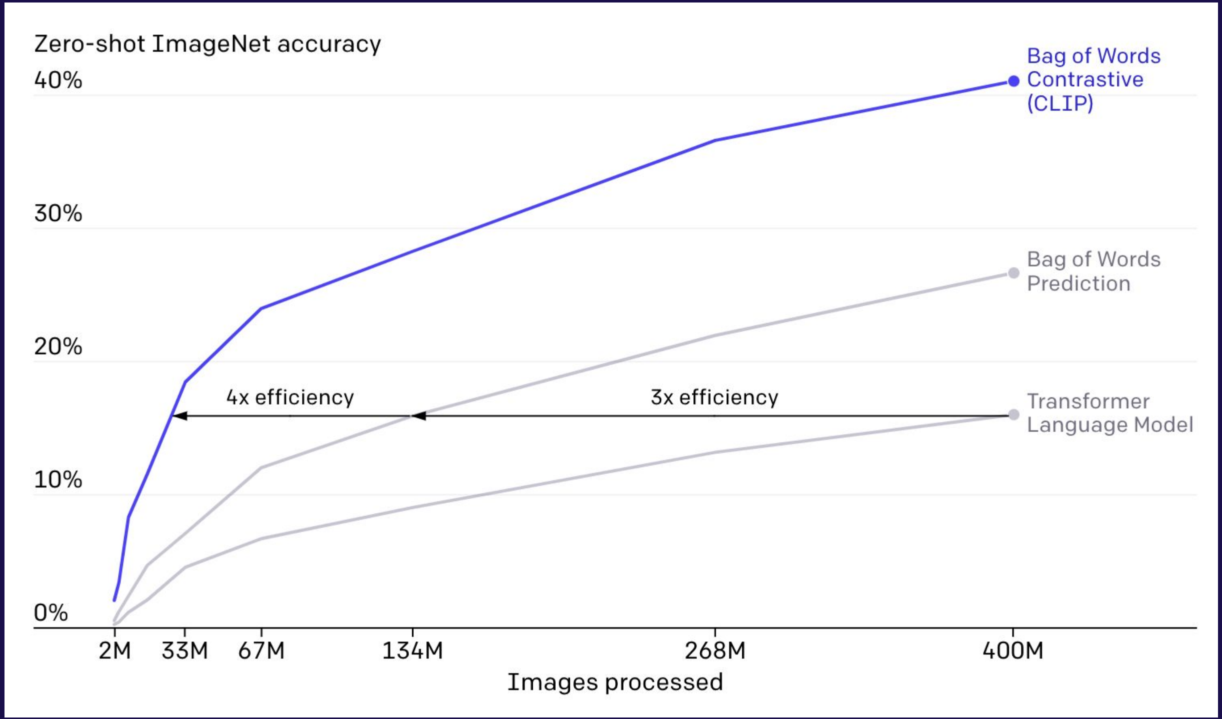
Zero-shot image classification



Zero-shot CLIP is much more robust

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%

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

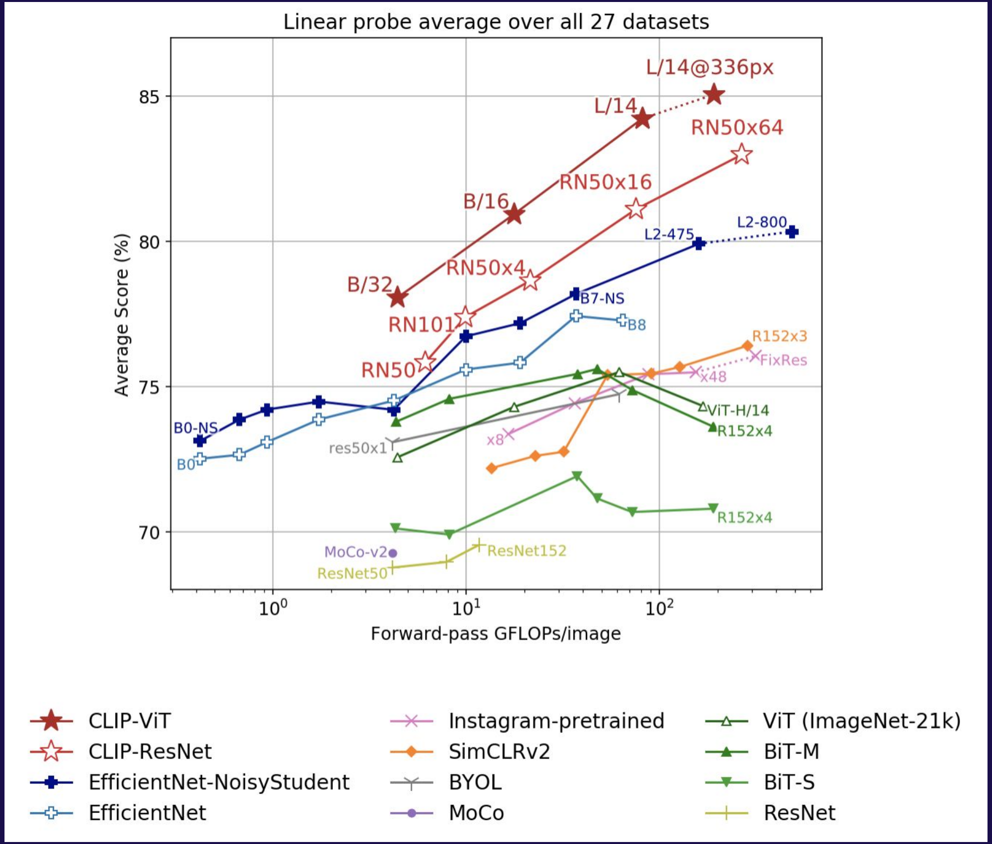
Logistic regression classifier on image features

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

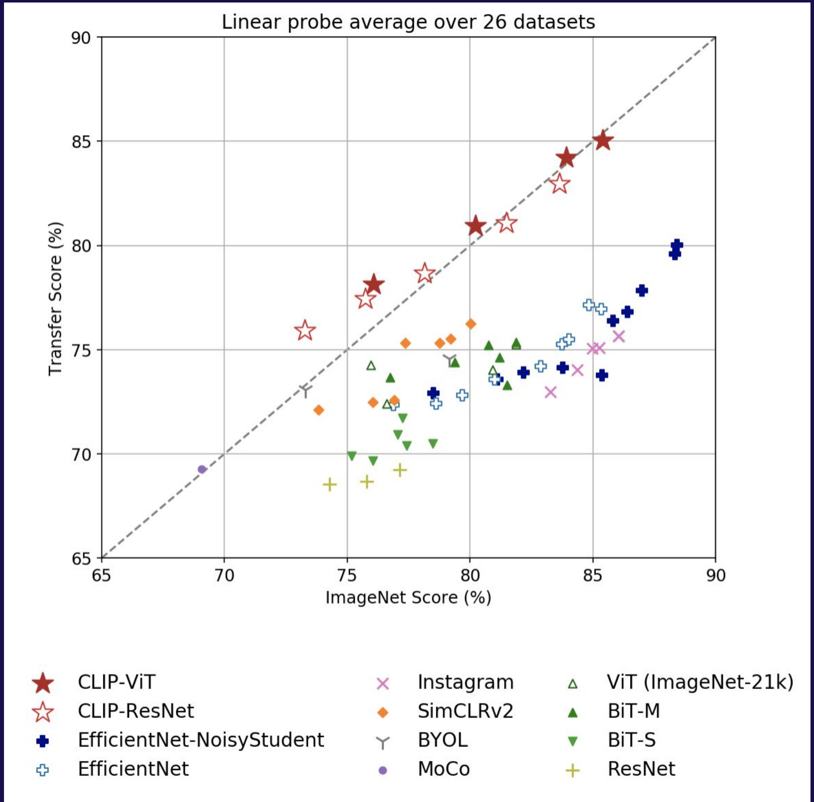
Evaluated on 27 image datasets × 65 vision models

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

Linear probe performance vs SOTA vision models



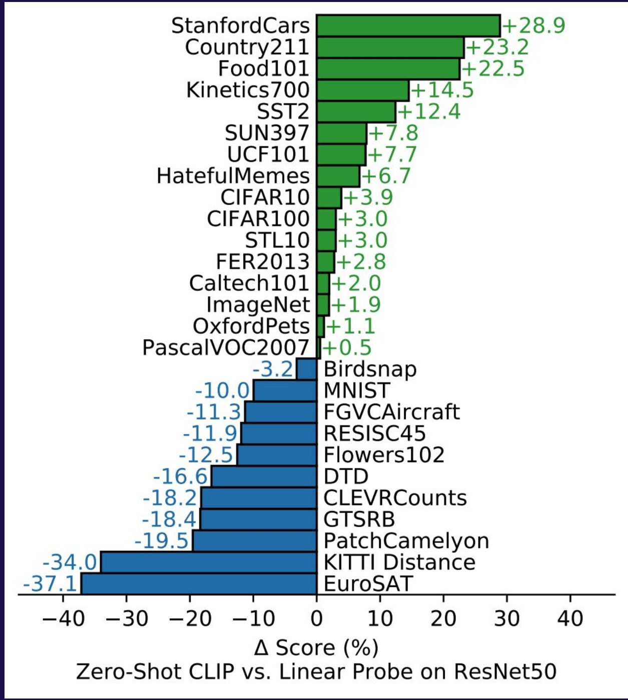
vs ImageNet score



Zero-Shot Transfer

Zero-shot vs Linear-probe ResNet-50

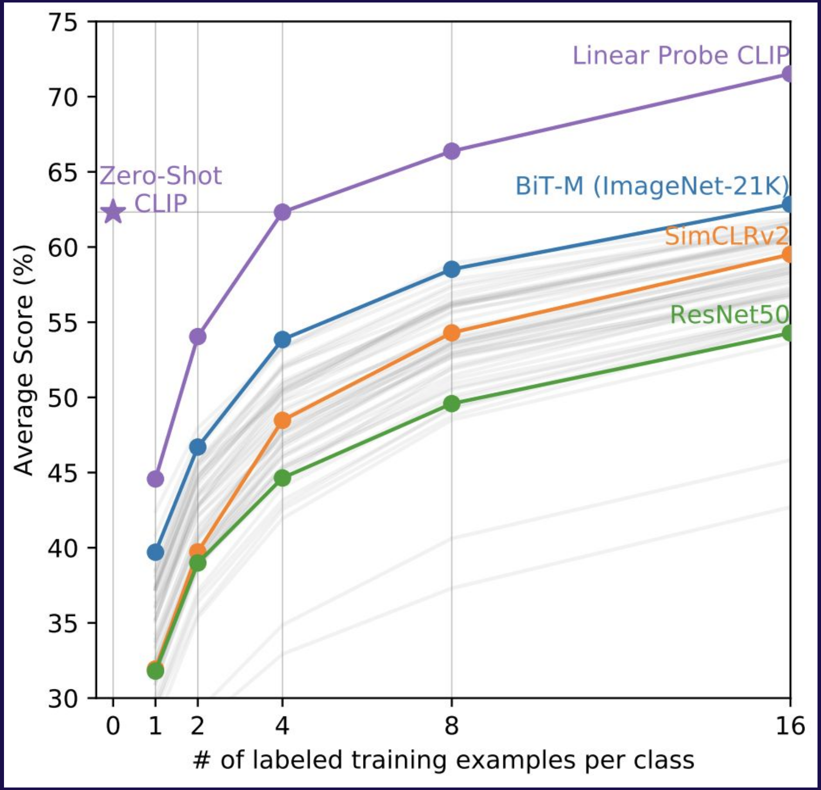
Zero-shot CLIP matches fully supervised ResNet-50 across eval suite



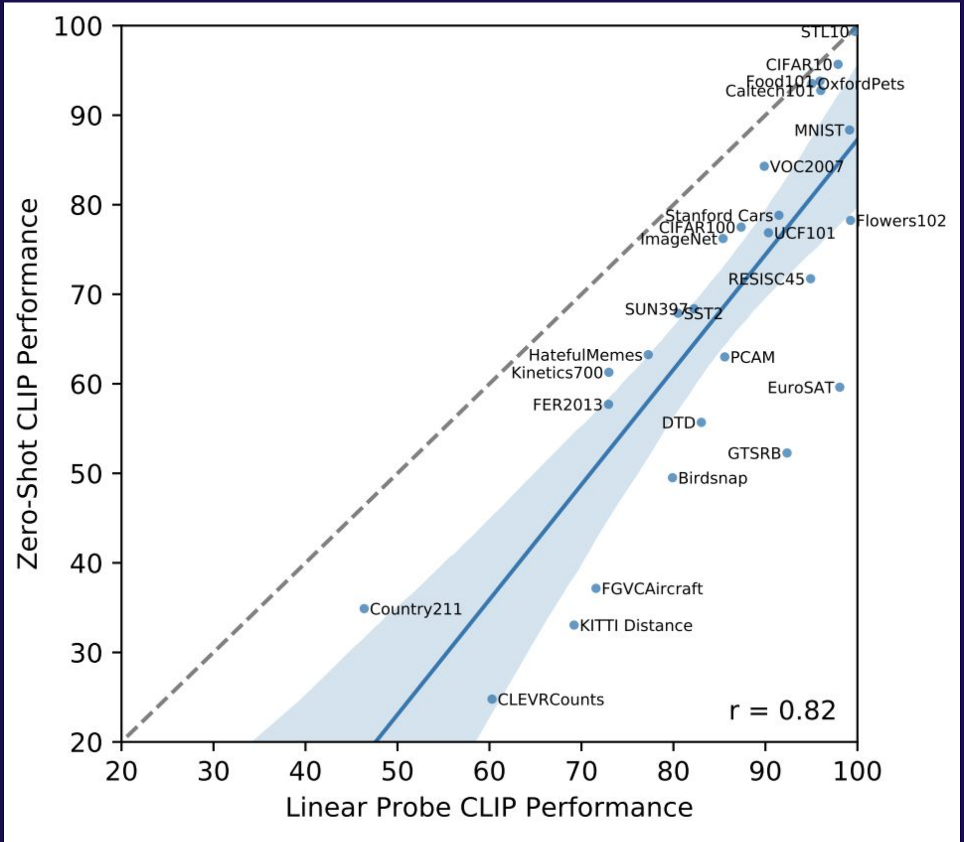
Zero-shot CLIP vs Few-shot linear probes

Zero-shot CLIP is as good as

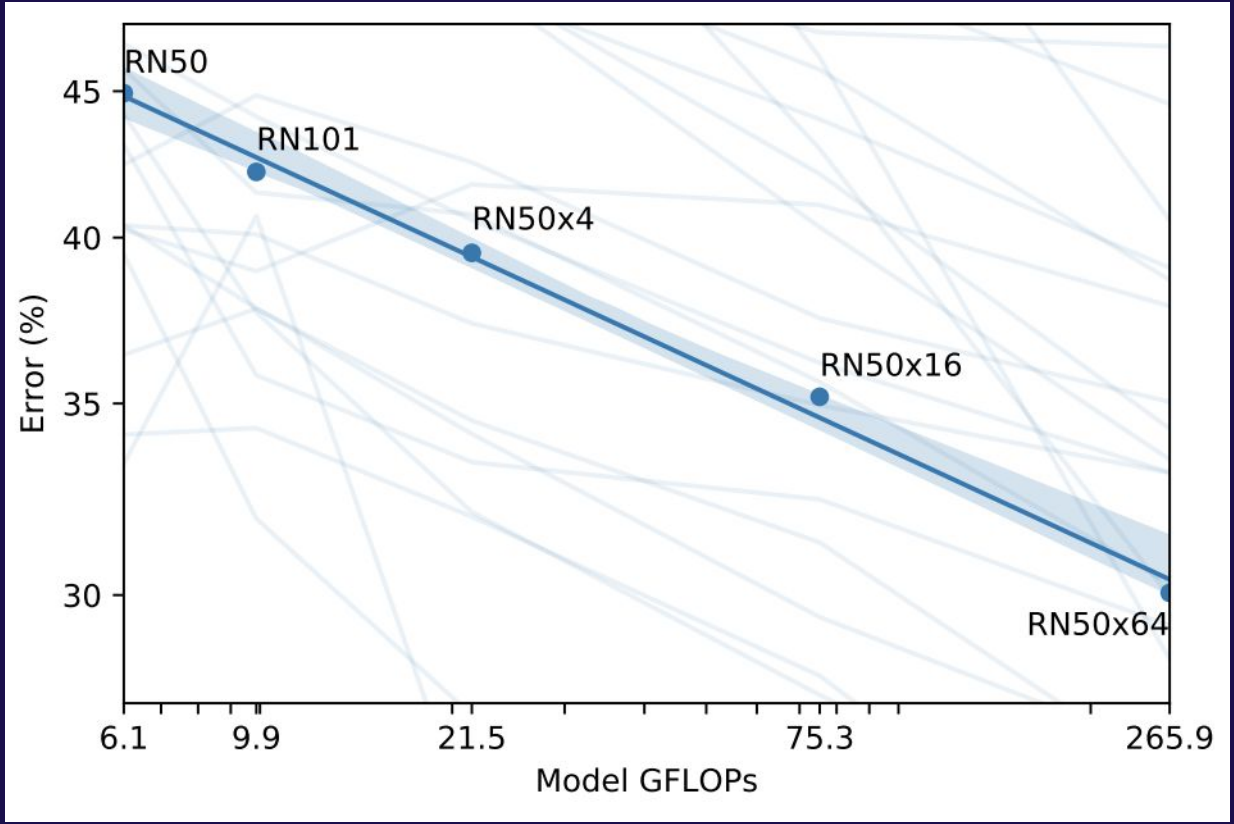
- 4-shot linear-probe CLIP
- 16-shot BiT-M



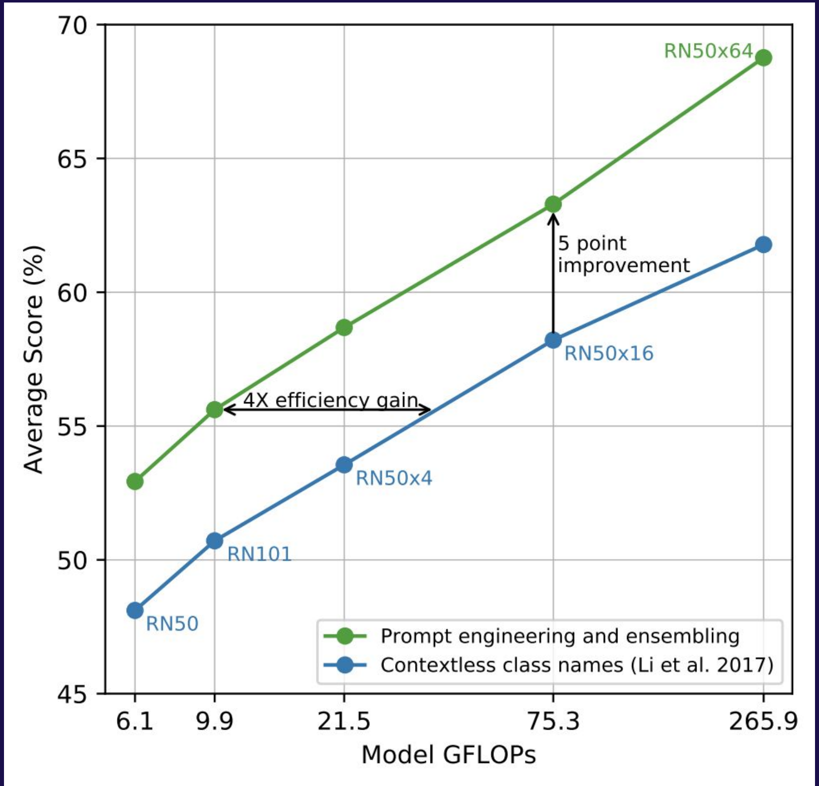
Zero-shot vs Linear-probe CLIP



Zero-shot performance vs model size



Prompt engineering



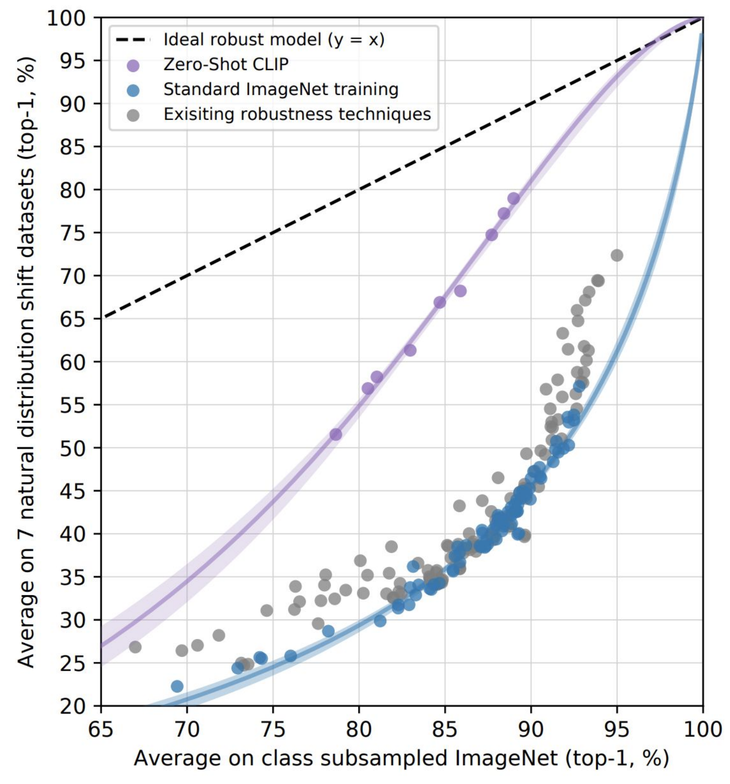
Robustness to Natural Distribution Shift

Robustness to natural distribution shift

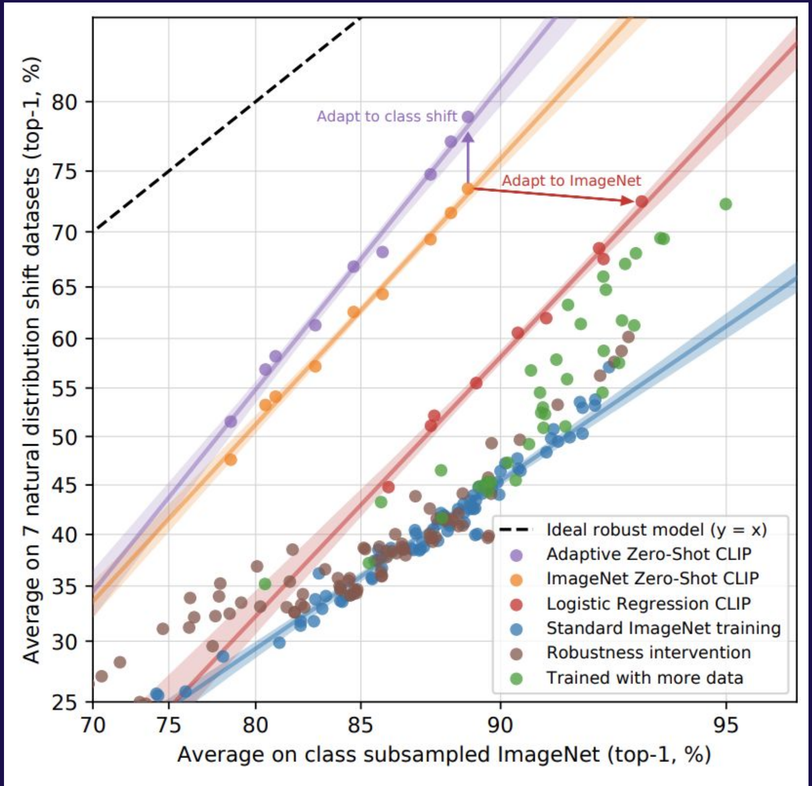
Zero-Shot CLIP is much more robust!

7 ImageNet-like Datasets (Taori et al.)

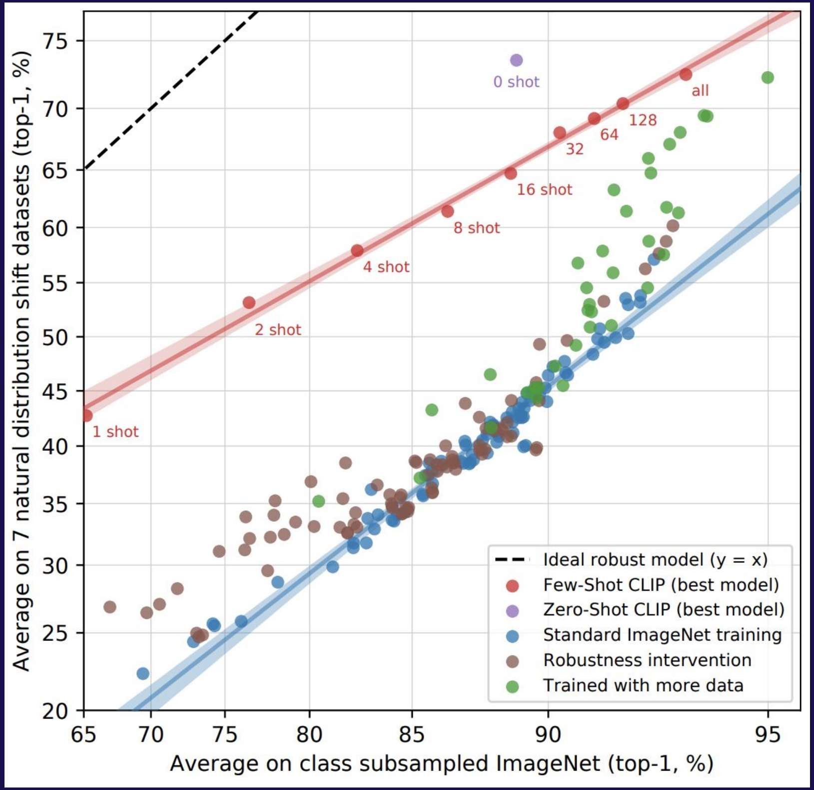
- ImageNetV2
- ImageNet-A
- ImageNet-R
- ImageNet Sketch
- ObjectNet
- ImageNet Vid
- Youtube-BB



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, use of large validation sets for prompt engineering
- Social biases

Quantifying the (un)safety of CLIP models

- Class design can heavily influence bias

Category Label Set	0-2	3-9	10-19	20-29	30-39	40-49	50-59	60-69
Default Label Set	30.3	35.0	29.5	16.3	13.9	18.5	19.1	16.2
Default Label Set + ‘child’	2.3	4.3	14.7	15.0	13.4	18.2	18.6	15.5

Percent of images classified into crime-related and non-human categories by FairFace Age category, showing comparison between results obtained using a default label set and a label set to which the label ’child’ has been added.

Quantifying the (un)safety of CLIP models

- Enables niche tasks which lack training data

CelebA Zero-Shot Top 1 Identity Recognition Results

Model	100 Classes	1k Classes	2k Classes
CLIP L/14	59.2	43.3	42.2
CLIP RN50x62	56.4	39.5	38.4
CLIP RN50x62	52.7	37.4	36.3
CLIP RN50x62	52.8	38.1	37.3

Not comprehensive, continuing to research to ensure safety

Related Work

Prior Related Work

Natural language supervision:

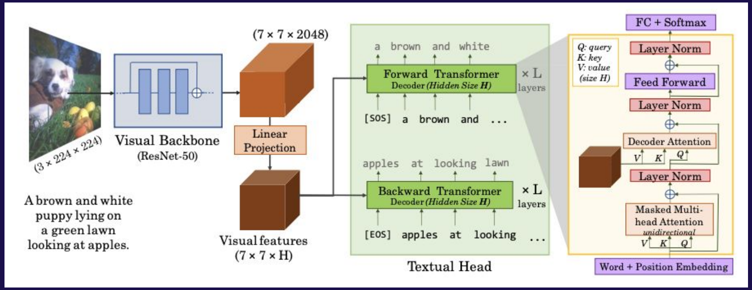
- YFCC100M WSL (Joulin et al.)
- VirTex (Desai and Johnson)
- ICMLM (Sariyildiz et al.)
- ConVIRT (Zhang et al.)

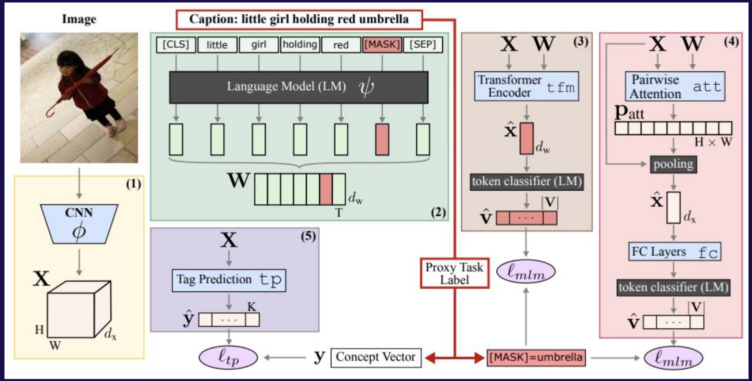
Zero-Shot Transfer:

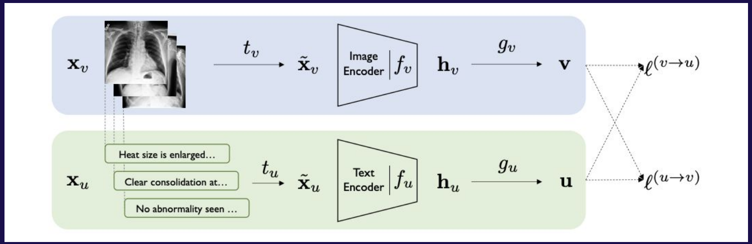
- Visual N-Grams (Li et al.)

Broad Evaluation and Robustness:

- VTAB (Zhang et al.)
- ImageNet Testbed (Taori et al.)







Multimodal Neurons in CLIP (Goh et al. Distill)

BIOLOGICAL NEURON Probed via depth electrodes		**CLIP NEURON** Neuron 244 from penultimate layer in CLIP RN50x4		**PREVIOUS ARTIFICIAL NEURON** Neuron 483, generic person detector from Inception v1		
Halle Berry		Spider-Man		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

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`

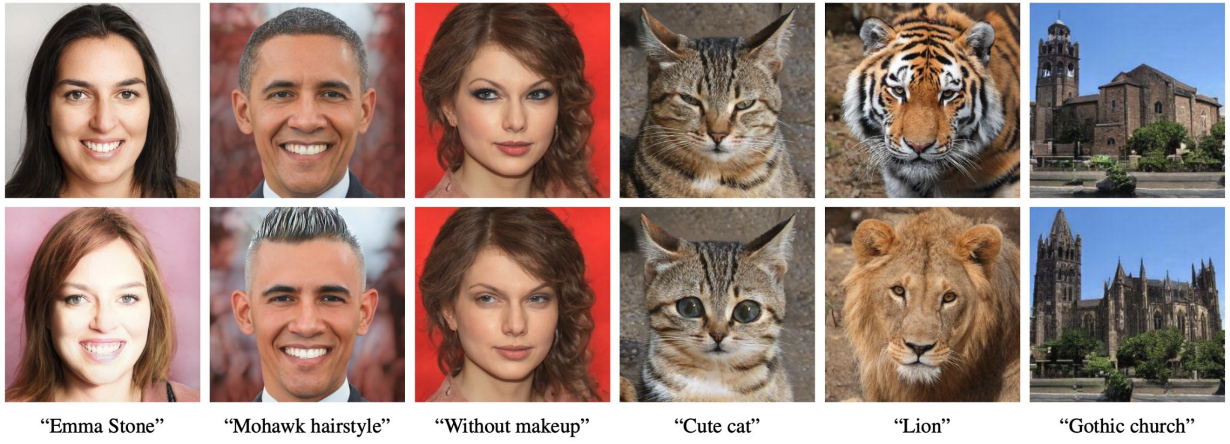
Attack text: *pizza*

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

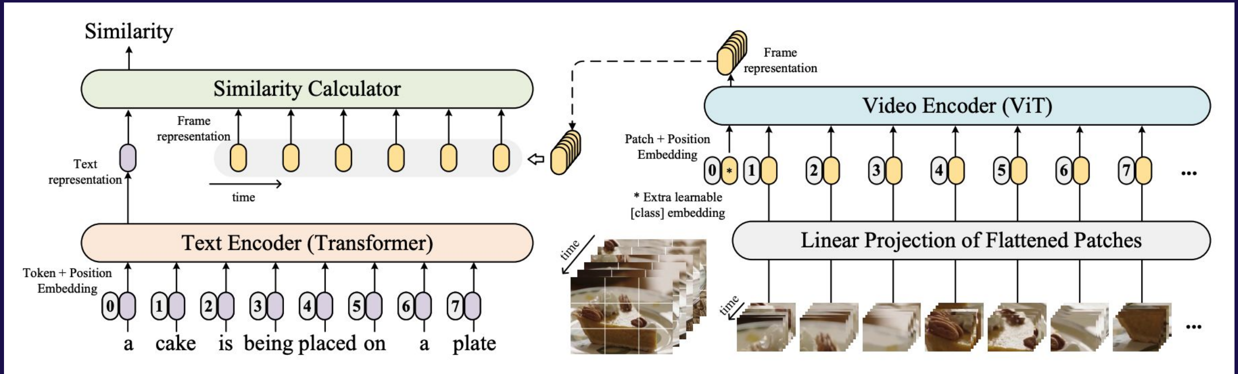
StyleCLIP
(Patashnik et al.)

Steering a GAN Using CLIP



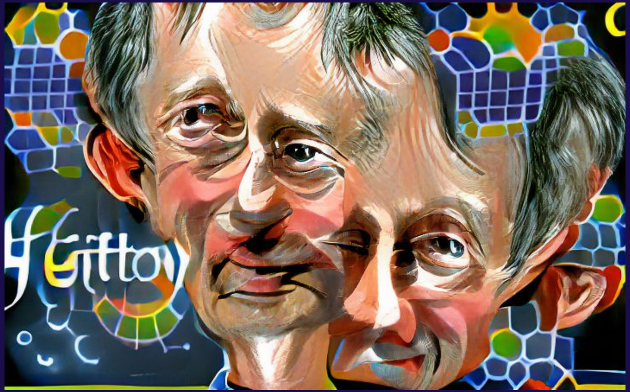
CLIP4Clip
(Luo & Ji, et al.)

Video retrieval using
CLIP features



More text-based image generations using CLIP

“A banquet hall”



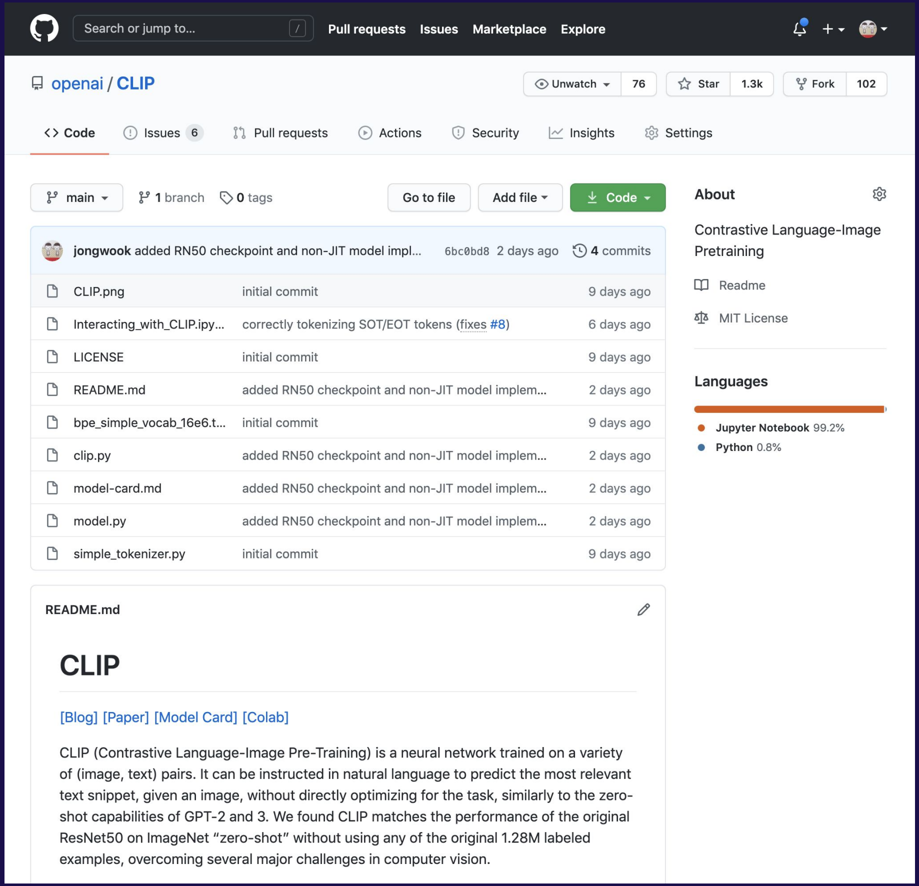
“Geoffrey Hinton”

“Dogs playing poker”

Try CLIP today!

https://github.com/openai/CLIP

- PyTorch implementation
- Colab notebook
- Zero-Shot prediction reference
- Linear probe reference
- YFCC100M dataset
- Released models



Thank You

Visit openai.com for more information.

FOLLOW @OPENAI ON TWITTER
WE ARE HIRING!