Learning Transferable Visual Models From Natural Language Supervision

https://arxiv.org/pdf/2103.00020.pdf

Under supervision of Dr. Jimson Mathew

Link to code implementation:
https://colab.research.google.com/drive/1owc_9Wp-KUA-pZzpK-1z3U6FkKNXsXqI?usp=sharing

By Mehuli Pal
mehuli_1901cs78@iitp.ac.in

OpenAI
ICML 2021

Contents

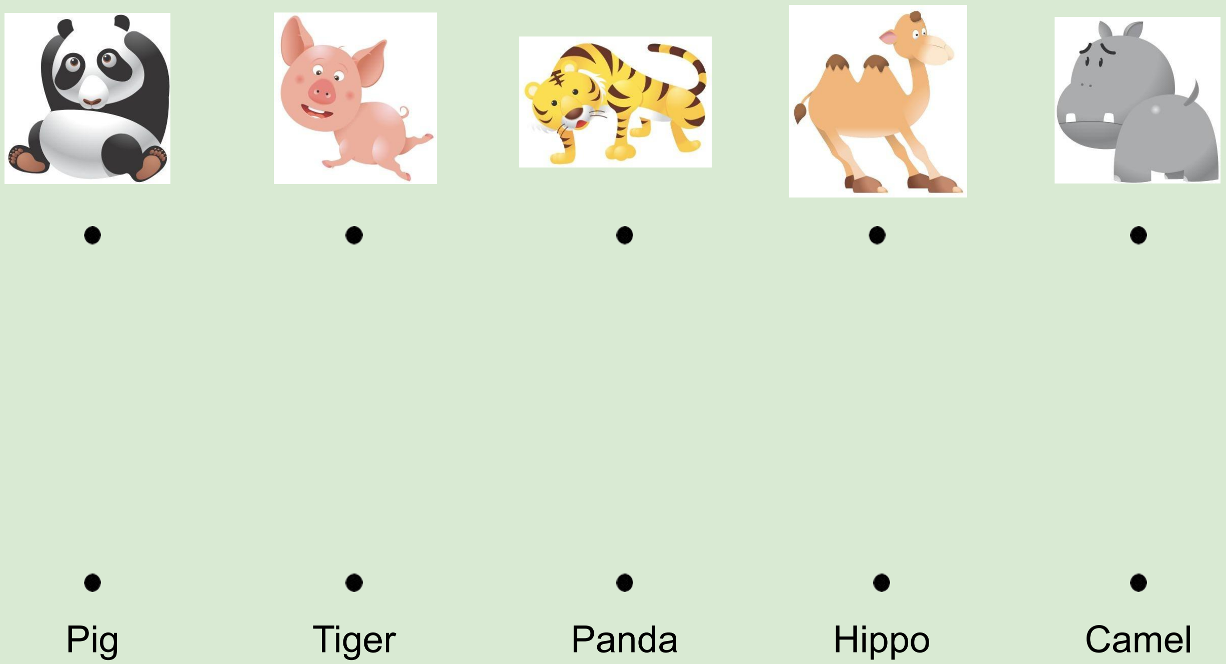
Limitations of Existing Methods

- Standard vision models are good at one task and one task only
- Typical vision datasets are labour intensive and costly to create
- Models that perform well on benchmarks have disappointingly poor performance on stress tests

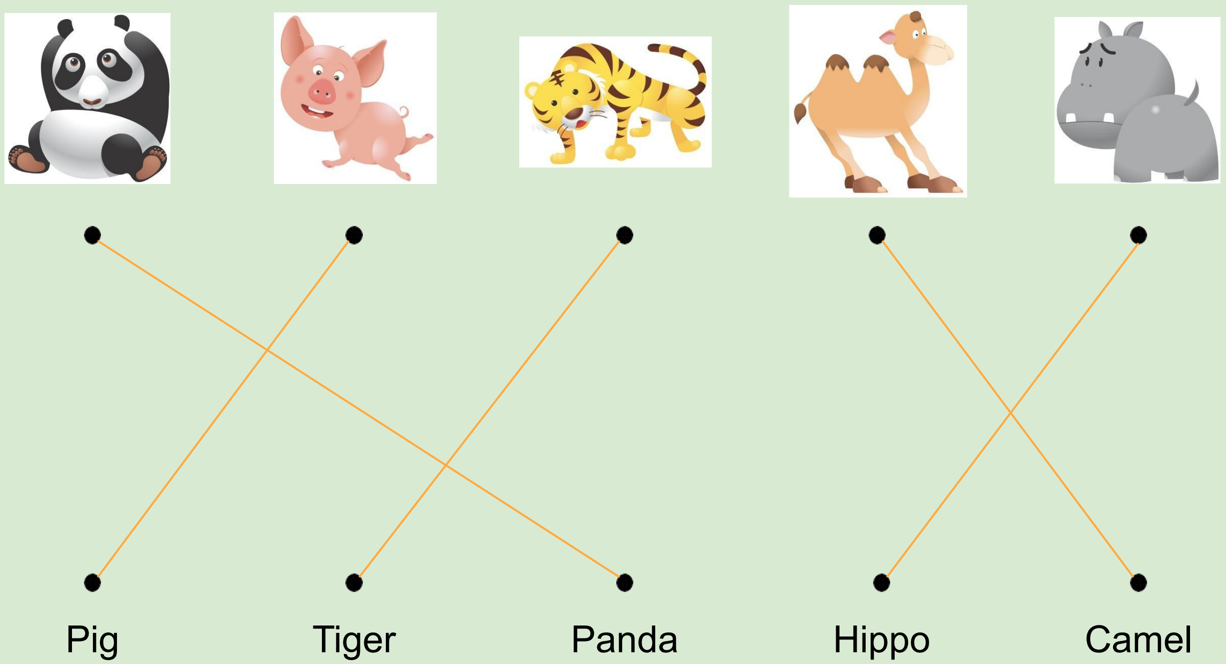
Introduction to CLIP

- Shorthand for **Contrastive Language-Image Pre-training**
- A neural network model trained on a wide variety of images and captions that's abundantly available on the internet
- CLIP expands knowledge of classification models to a wider array of things by leveraging semantic information in text
- Has impressive **zero-shot** capabilities, making it able to accurately predict entire classes it's never seen before!

Contrastive Learning



Contrastive Learning



Predictive Approach

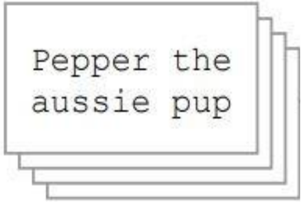
the dog is on the table

are	cat	dog	is	now	on	table	the
0	0	1	1	0	1	1	1

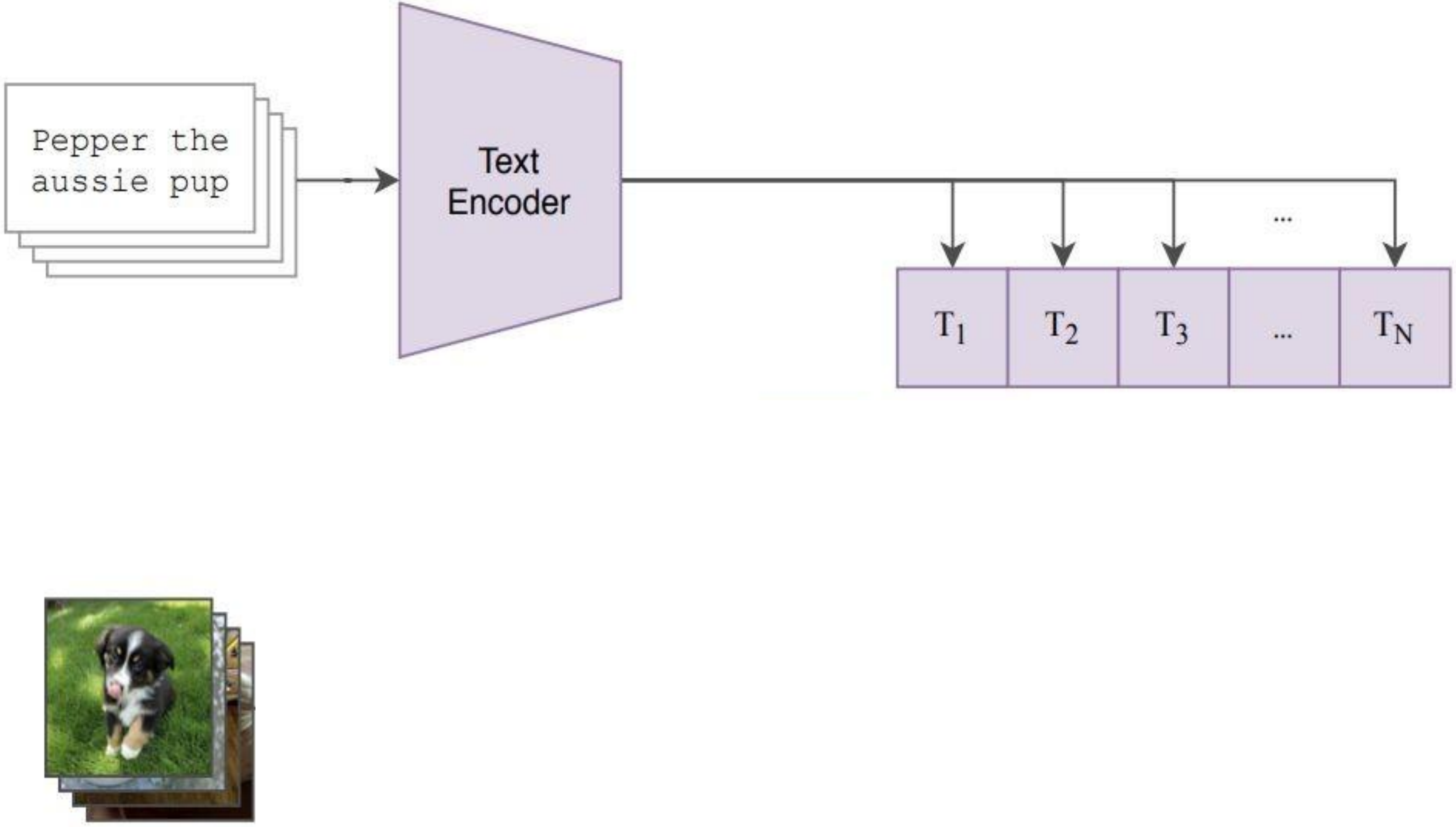
Contrastive Approach

a photo of a siberian husky

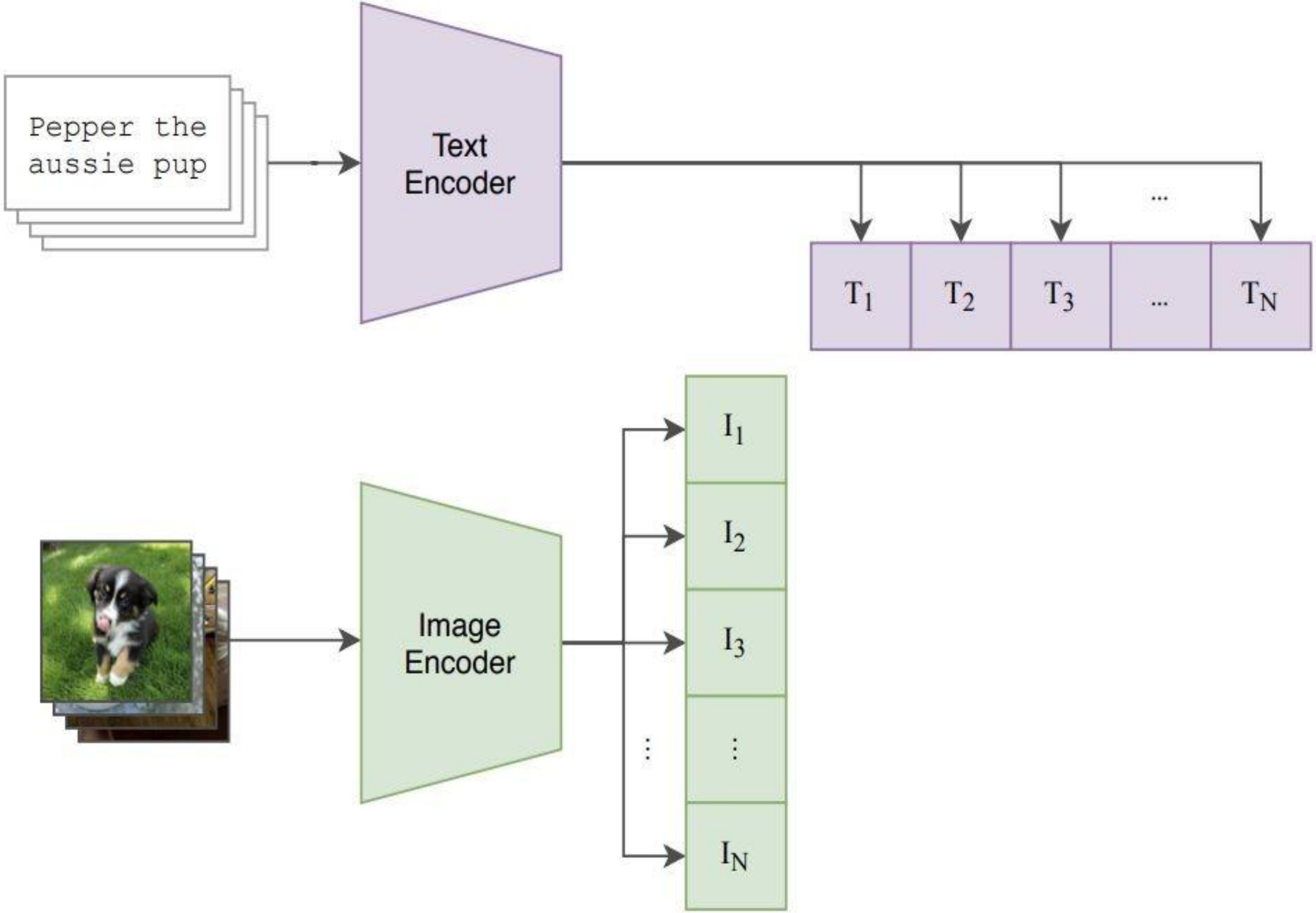
Contrastive pre training



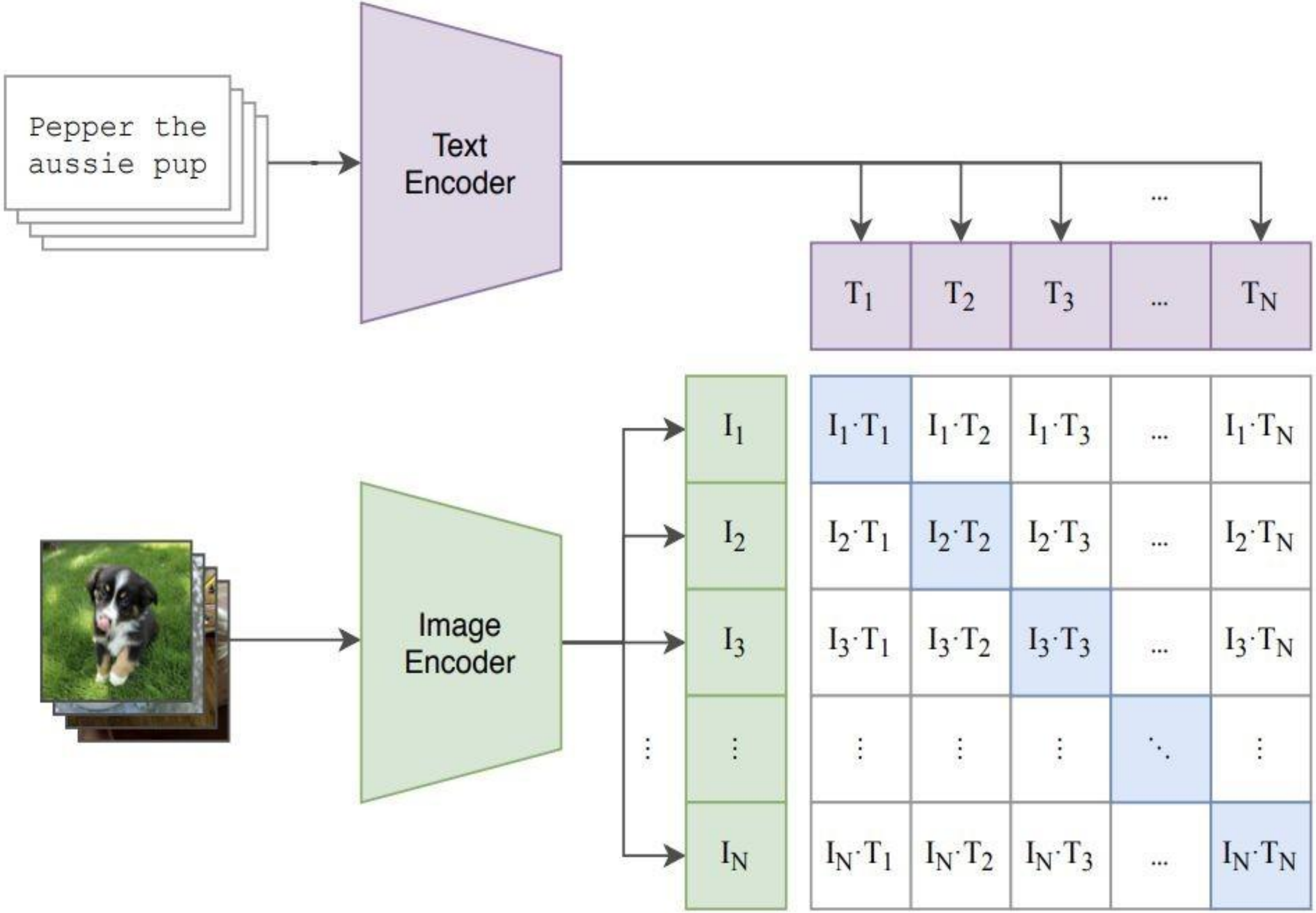
Contrastive pre training



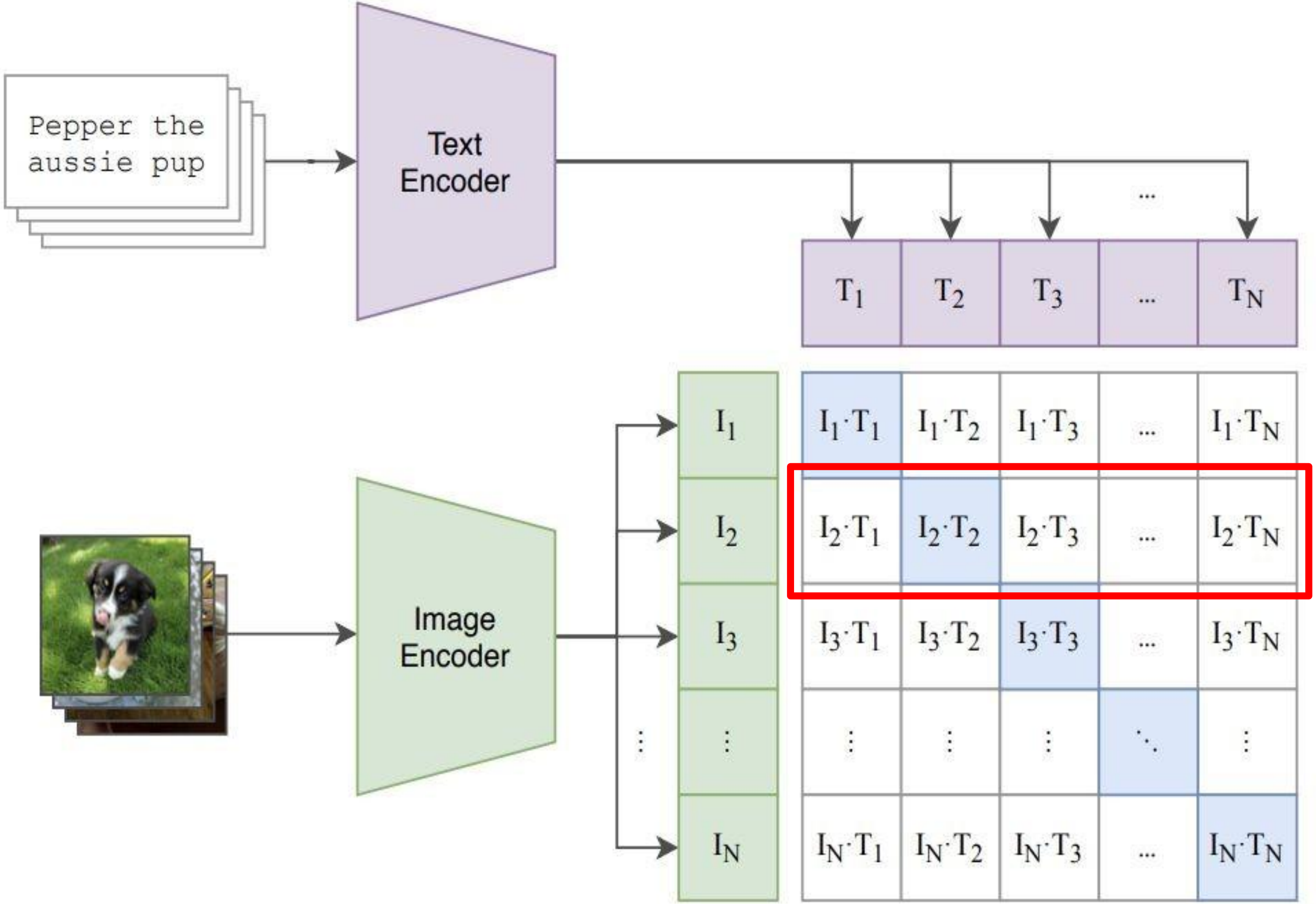
Contrastive pre training



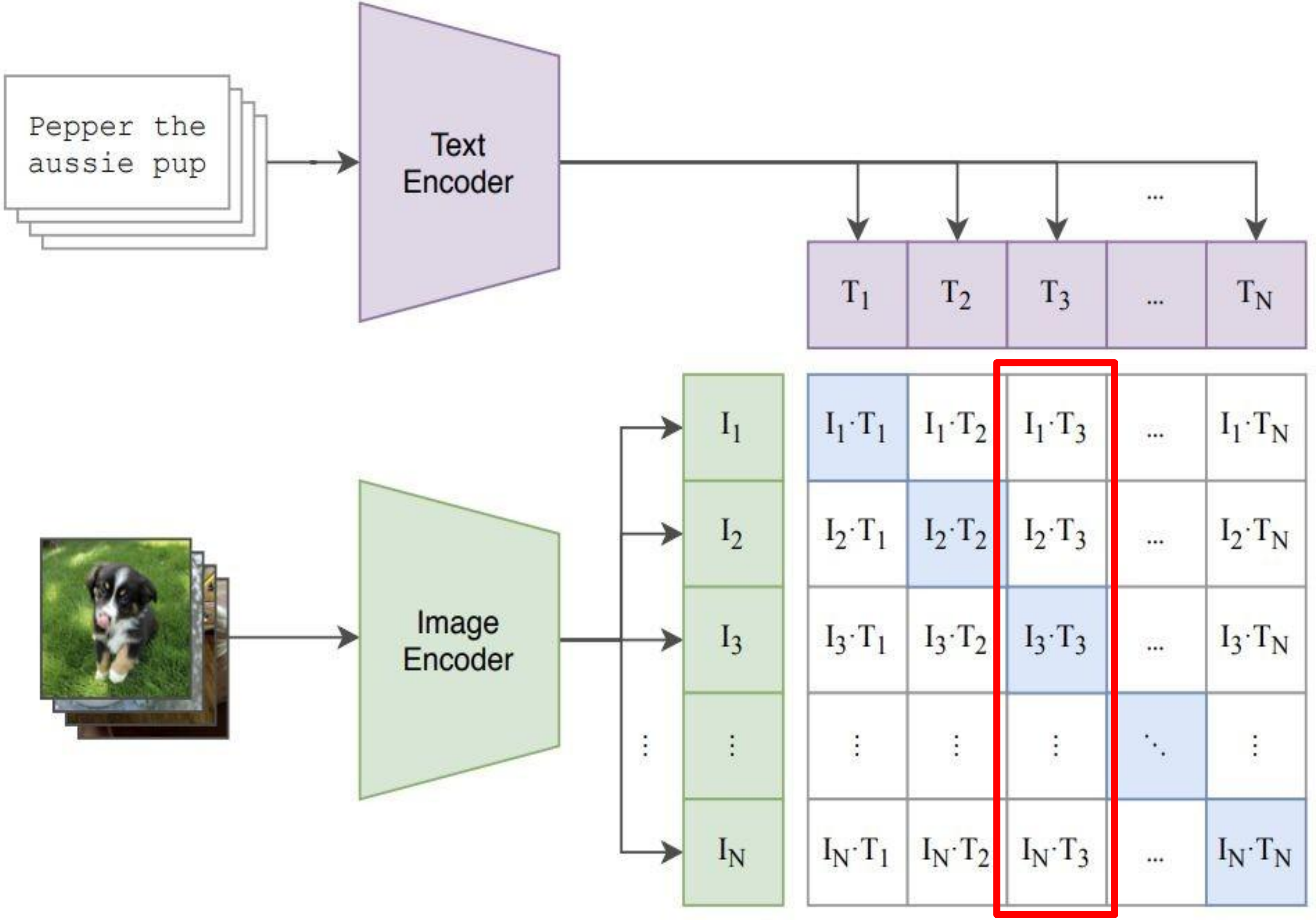
Contrastive pre training



Contrastive pre training



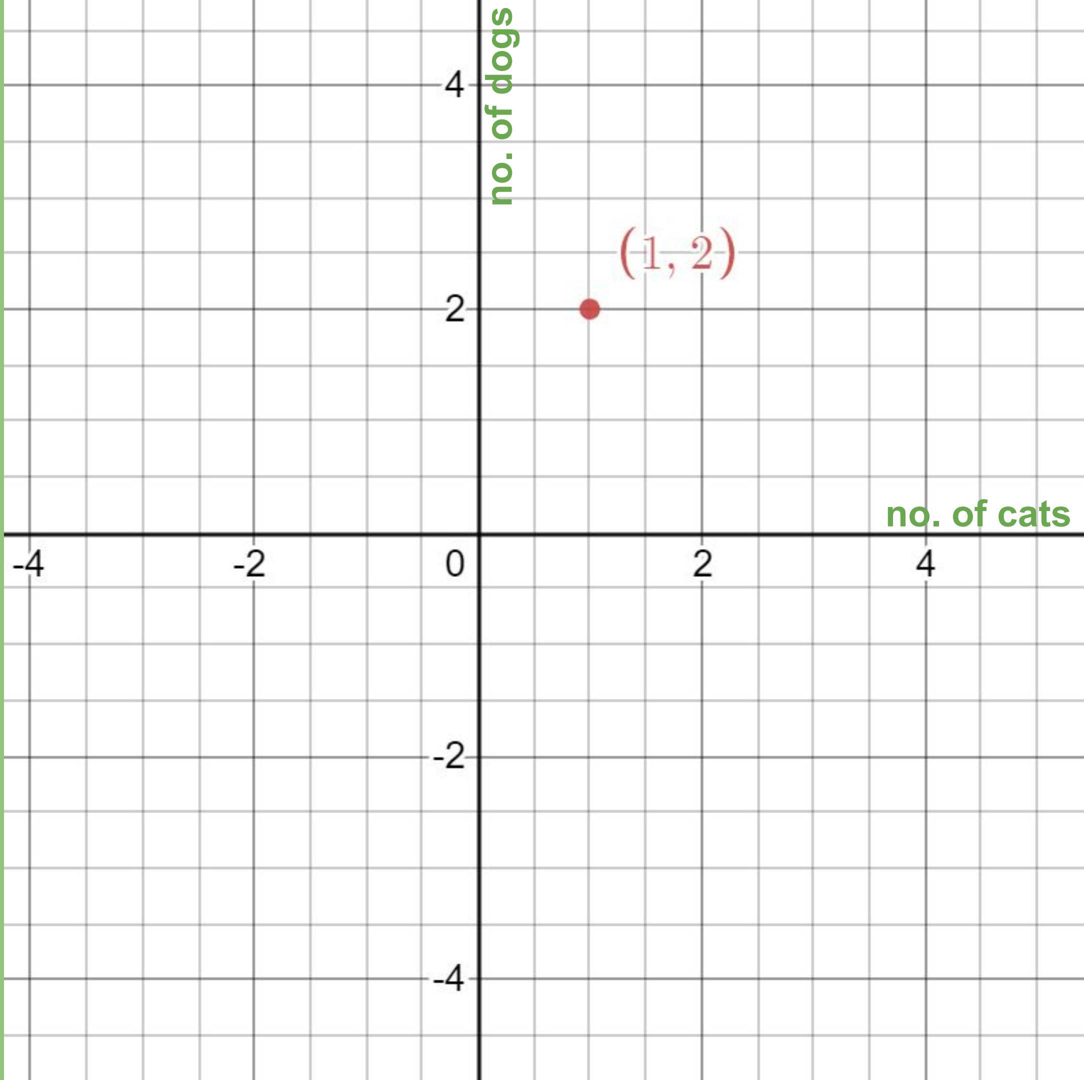
Contrastive pre training



Embedding

Suppose we have one cat and two dogs. This data can be represented as a dot on a graph.

We can do the same thing with text and with images!

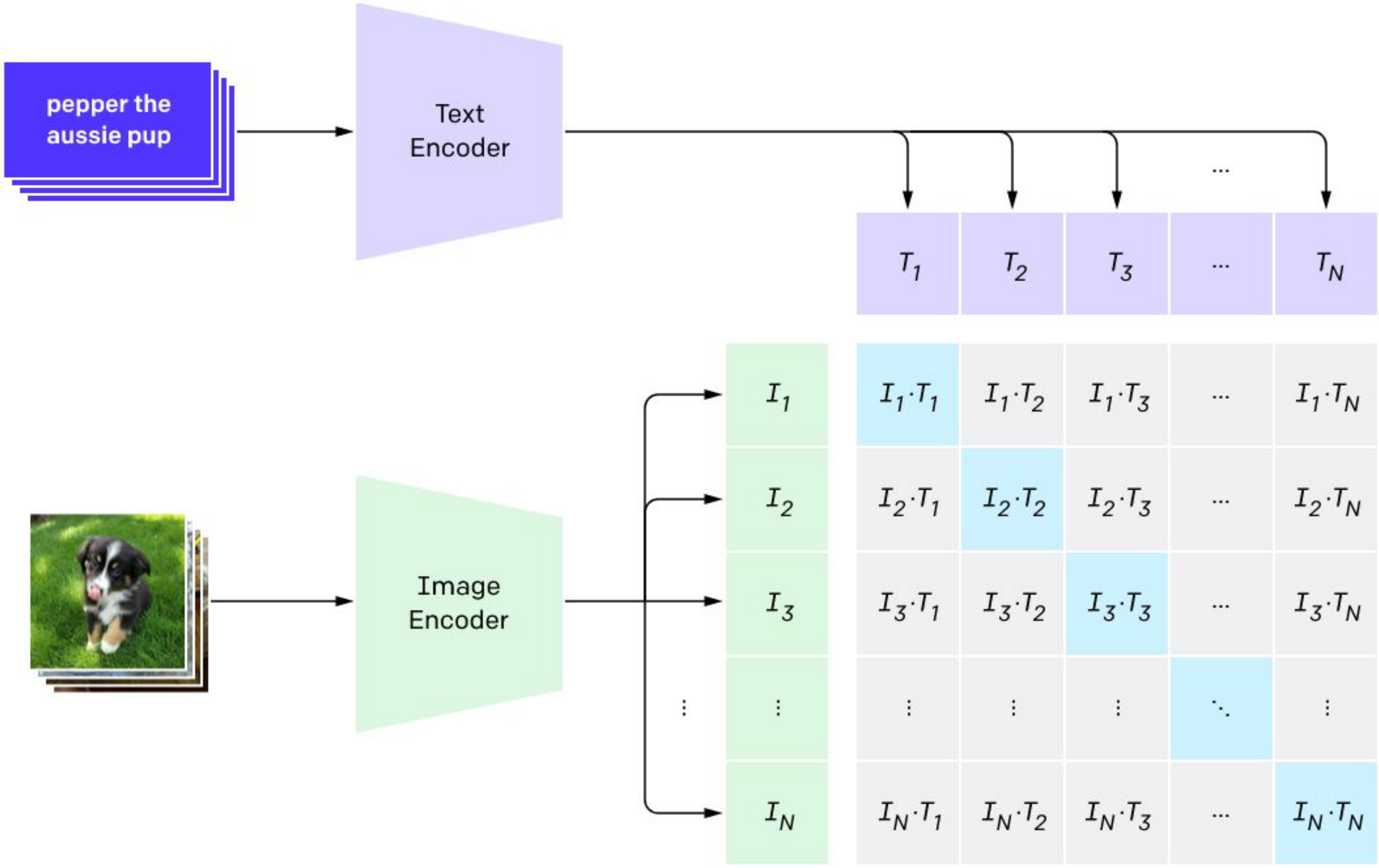


How Embedding Works...

The CLIP model consists of two sub-models called encoders:

- a text encoder that will embed (smash) text into mathematical space.
- an image encoder that will embed (smash) images into mathematical space.

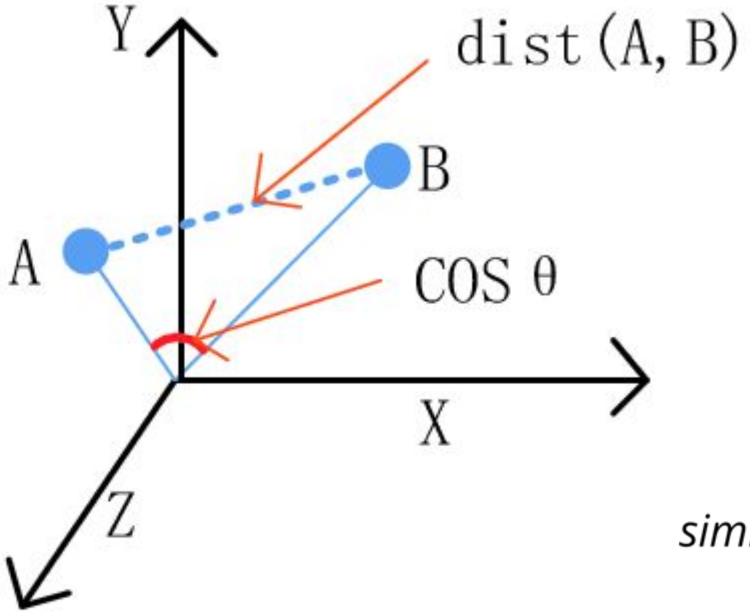
Measuring "Goodness" & "Badness"



The top card, **pepper the aussie pup** would enter the text encoder and come out as a series of numbers like (0, 0.2, 0.8).

The picture of, **pepper the aussie pup**, would enter the image encoder and come out like (0.05, 0.25, 0.7).

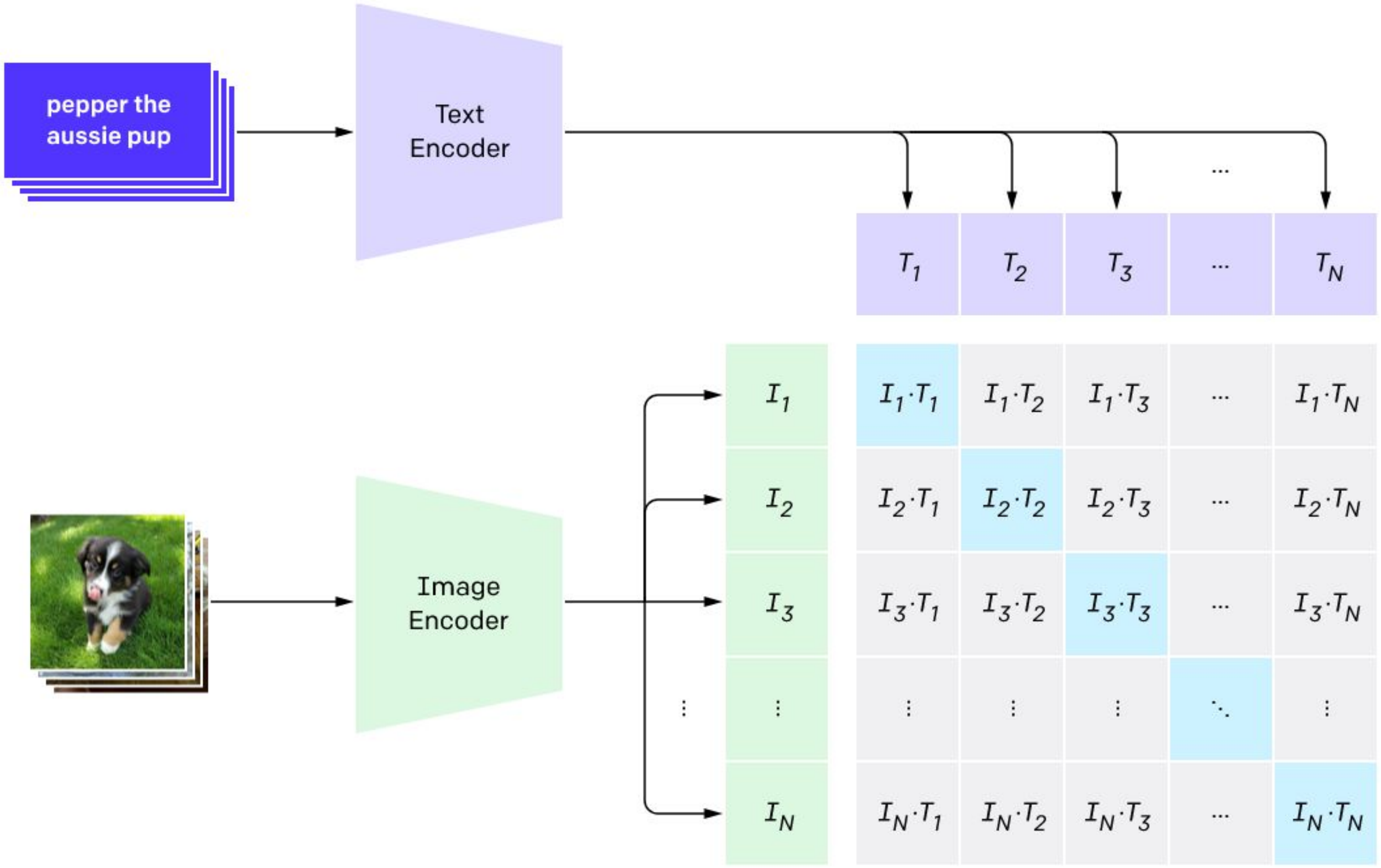
Cosine Similarity



One way for us to measure "goodness" of our model is how close the embedded representation (series of numbers) for each text is to the embedded representation for each image. There is a convenient way to calculate the similarity between two series of numbers: the **cosine similarity**.

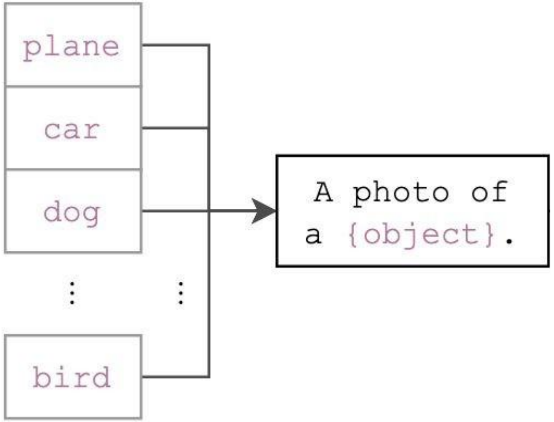
$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Measuring "Goodness" & "Badness"

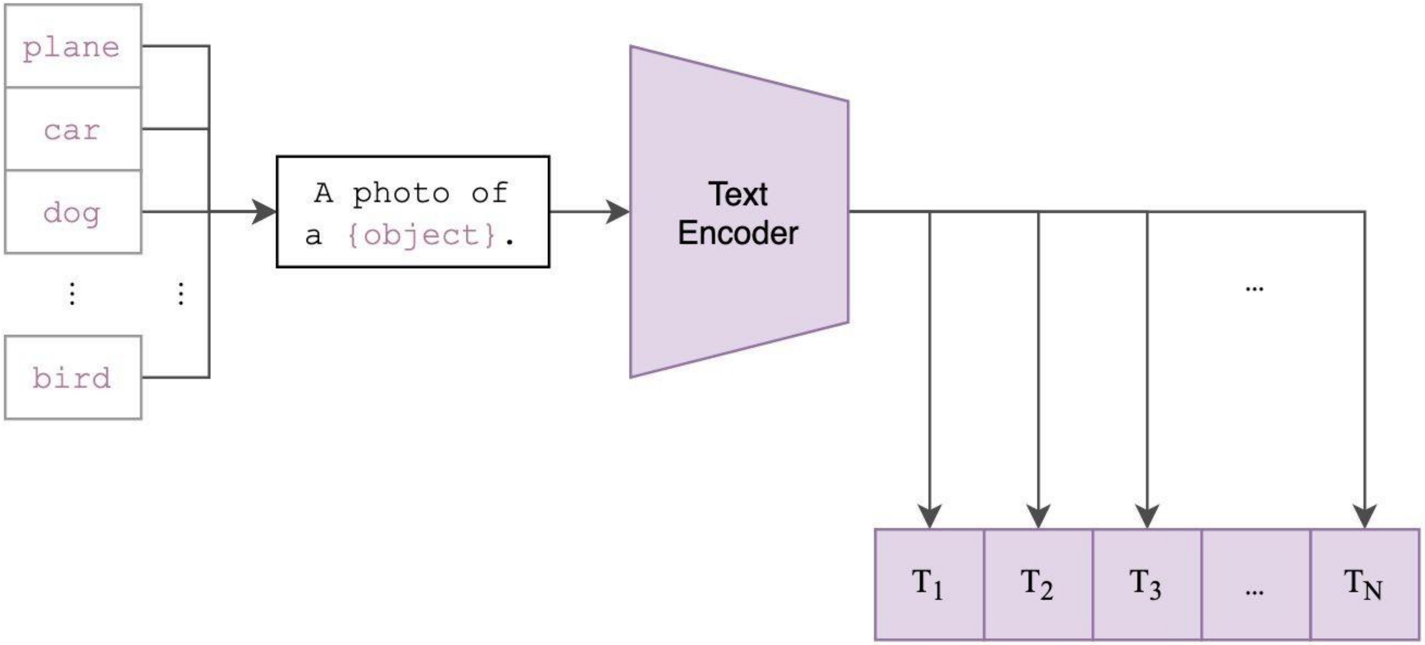


Zero-shot prediction

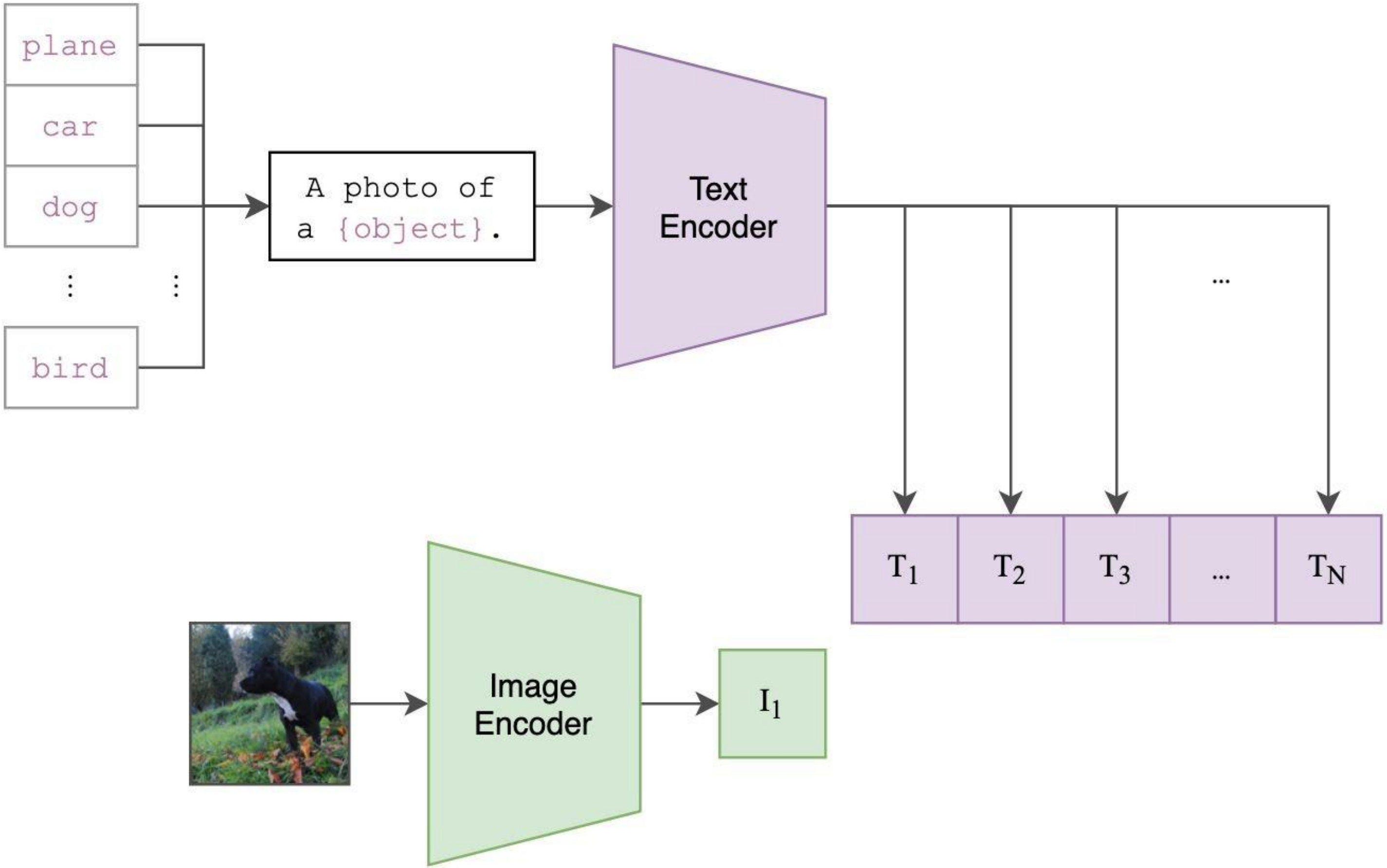
Zero-shot prediction



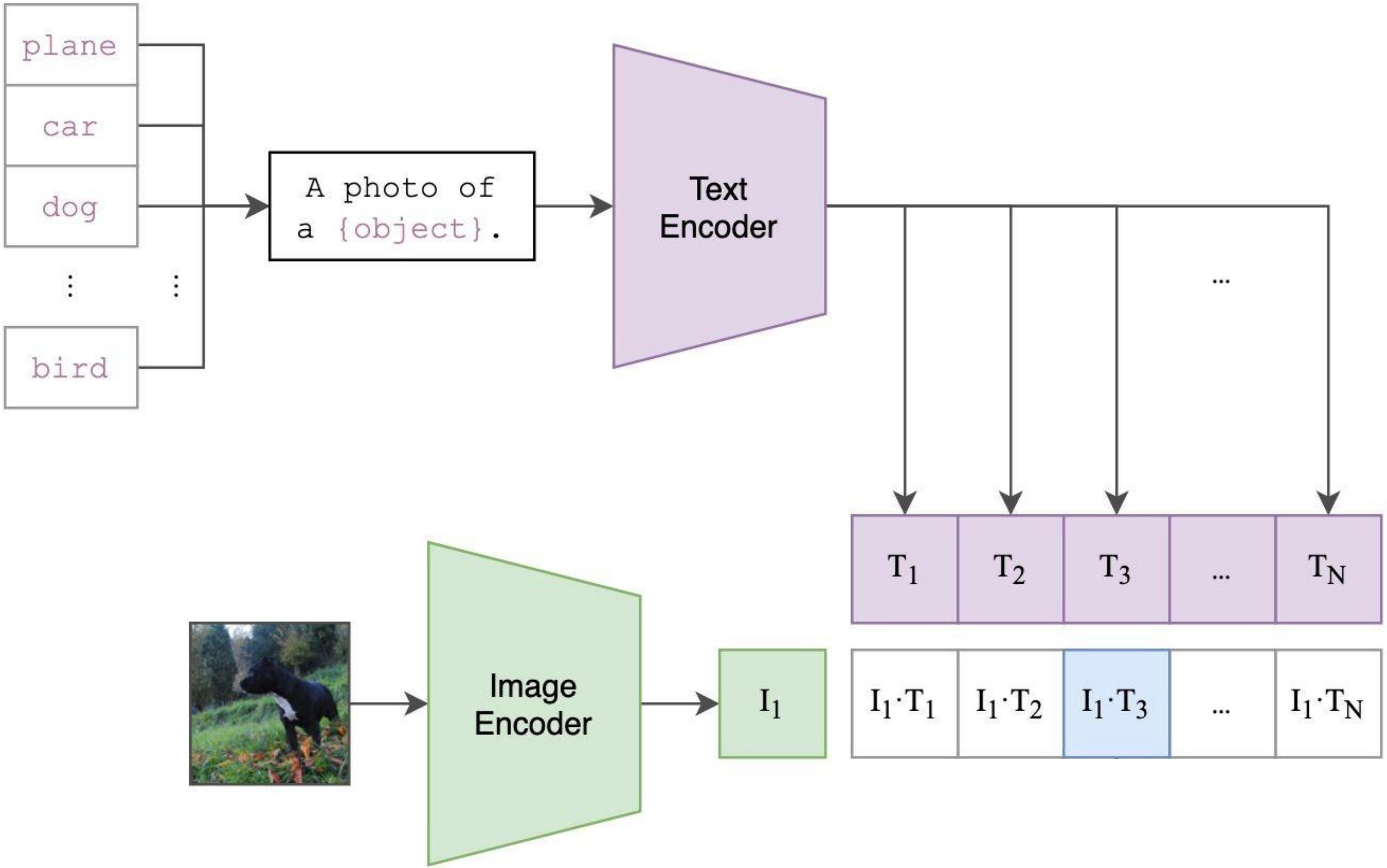
Zero-shot prediction



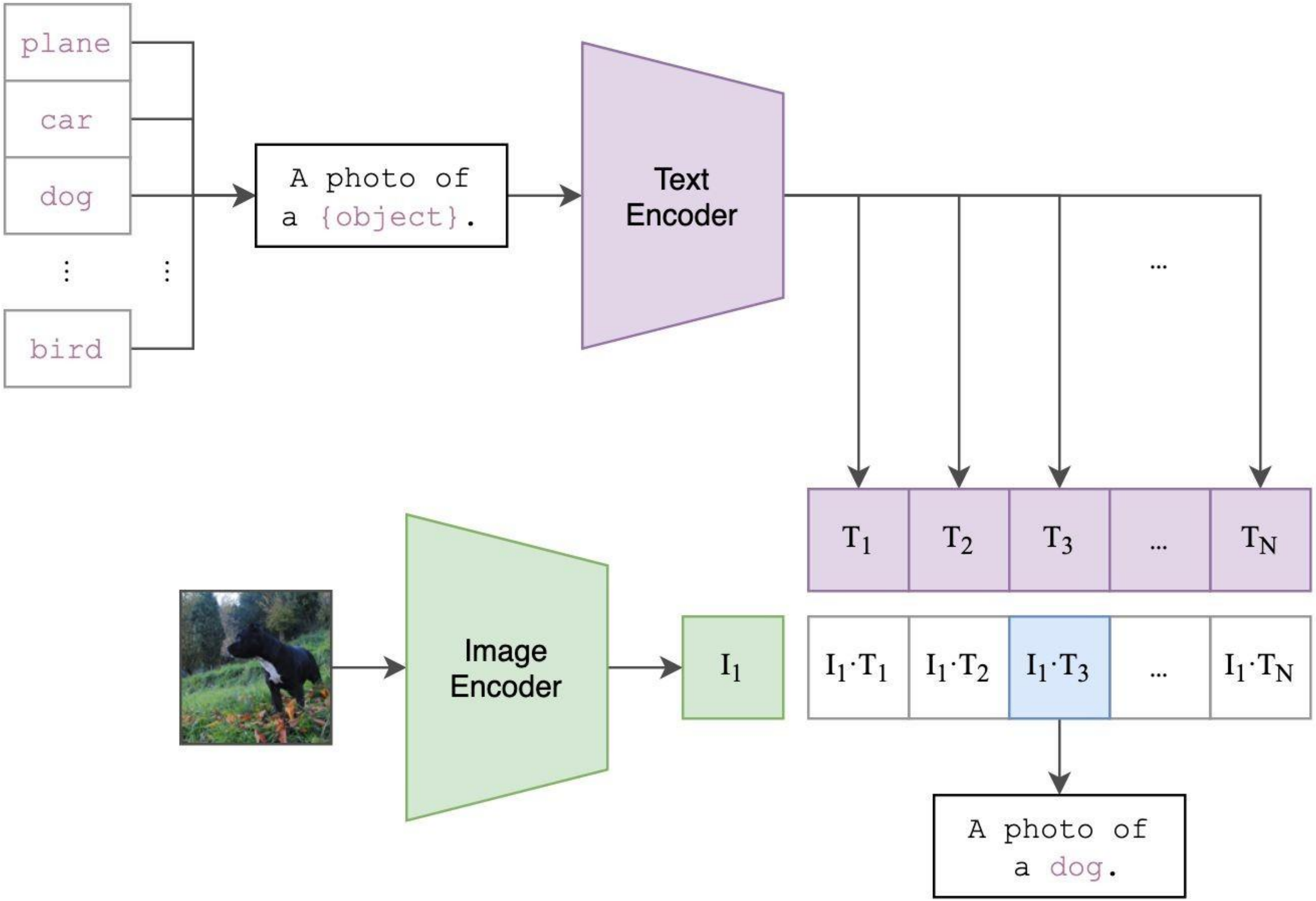
Zero-shot prediction



Zero-shot prediction



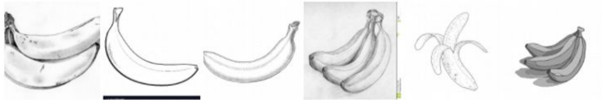
Zero-shot prediction



Zero-shot learning is when a model attempts to predict a class it saw zero times in the training data

CLIP as a bridge between Computer Vision & Natural Language Processing

Results...

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%

Link to code implementation:

https://colab.research.google.com/drive/1owc_9Wp-KUA-pZzpK-1z3U6FkKNXsXqI?usp=sharing

Some CLIP details

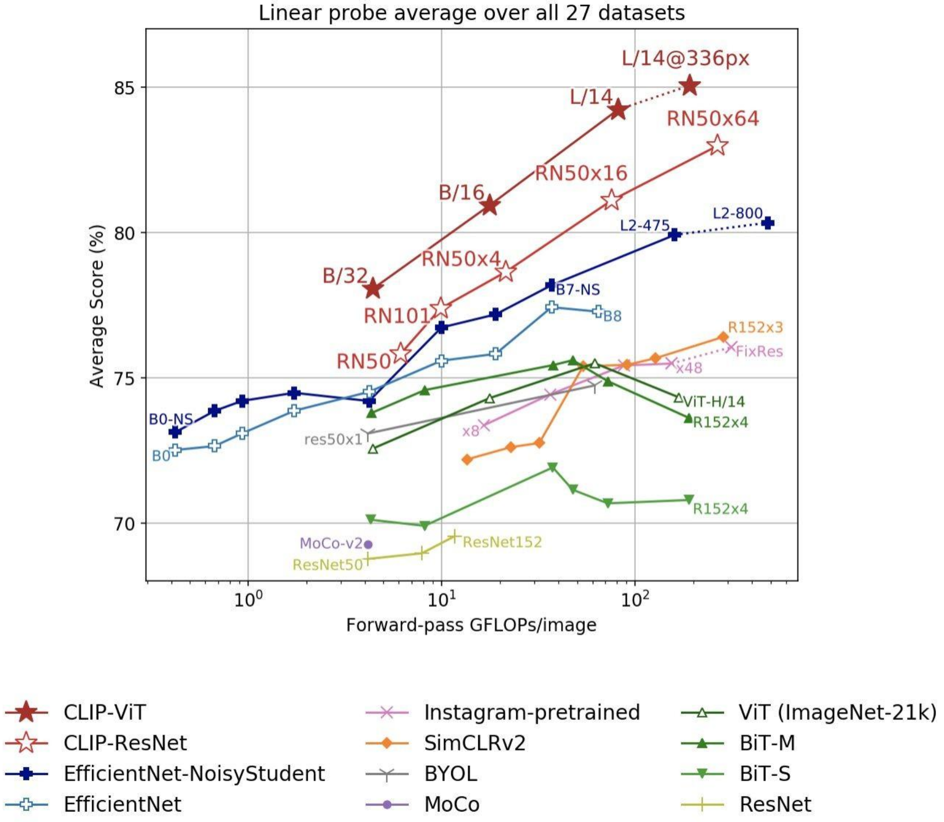
Training

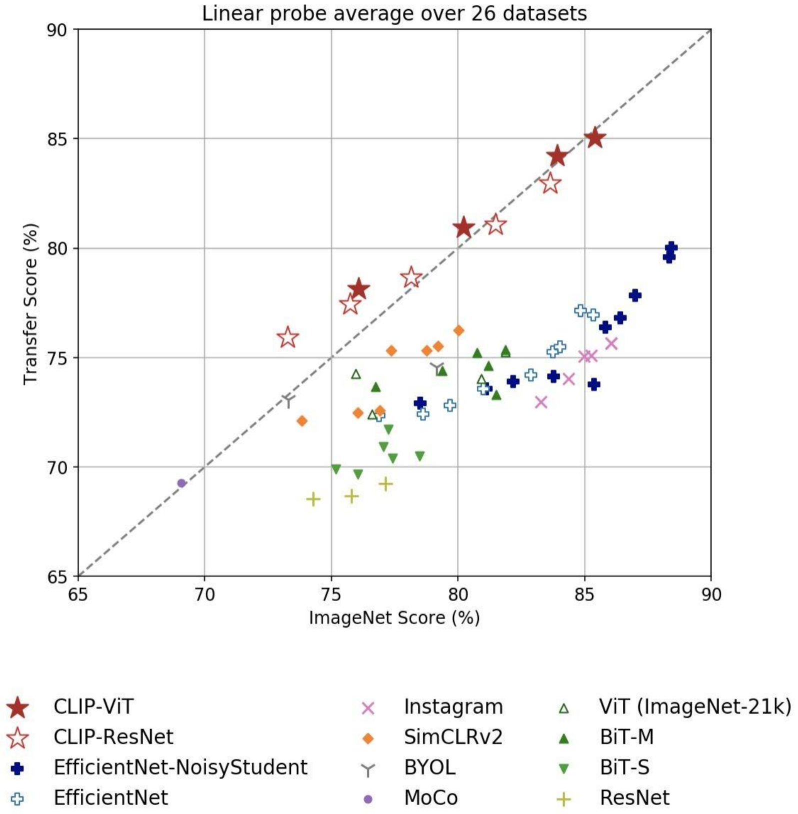
- Trained on 400M image-text pairs from the internet
- Batch size of 32,768
- 32 epochs over the dataset
- Cosine learning rate decay

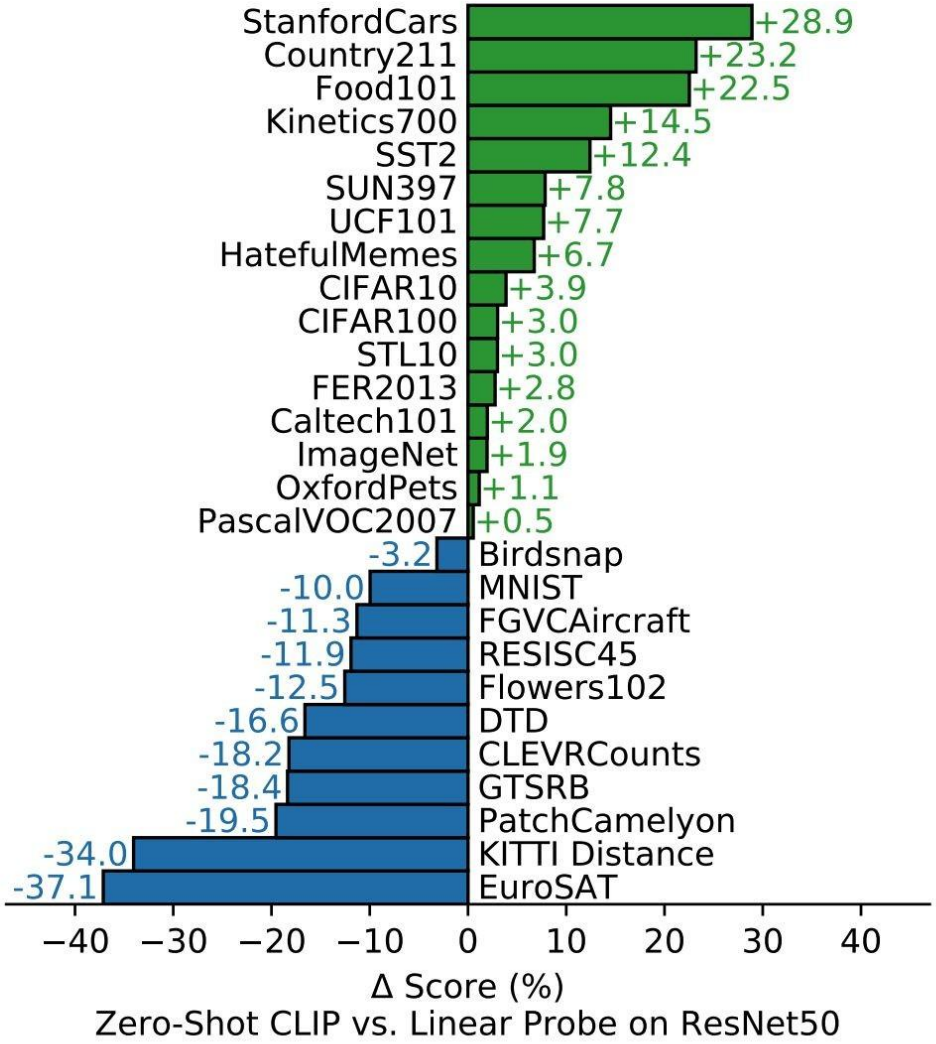
Architecture

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

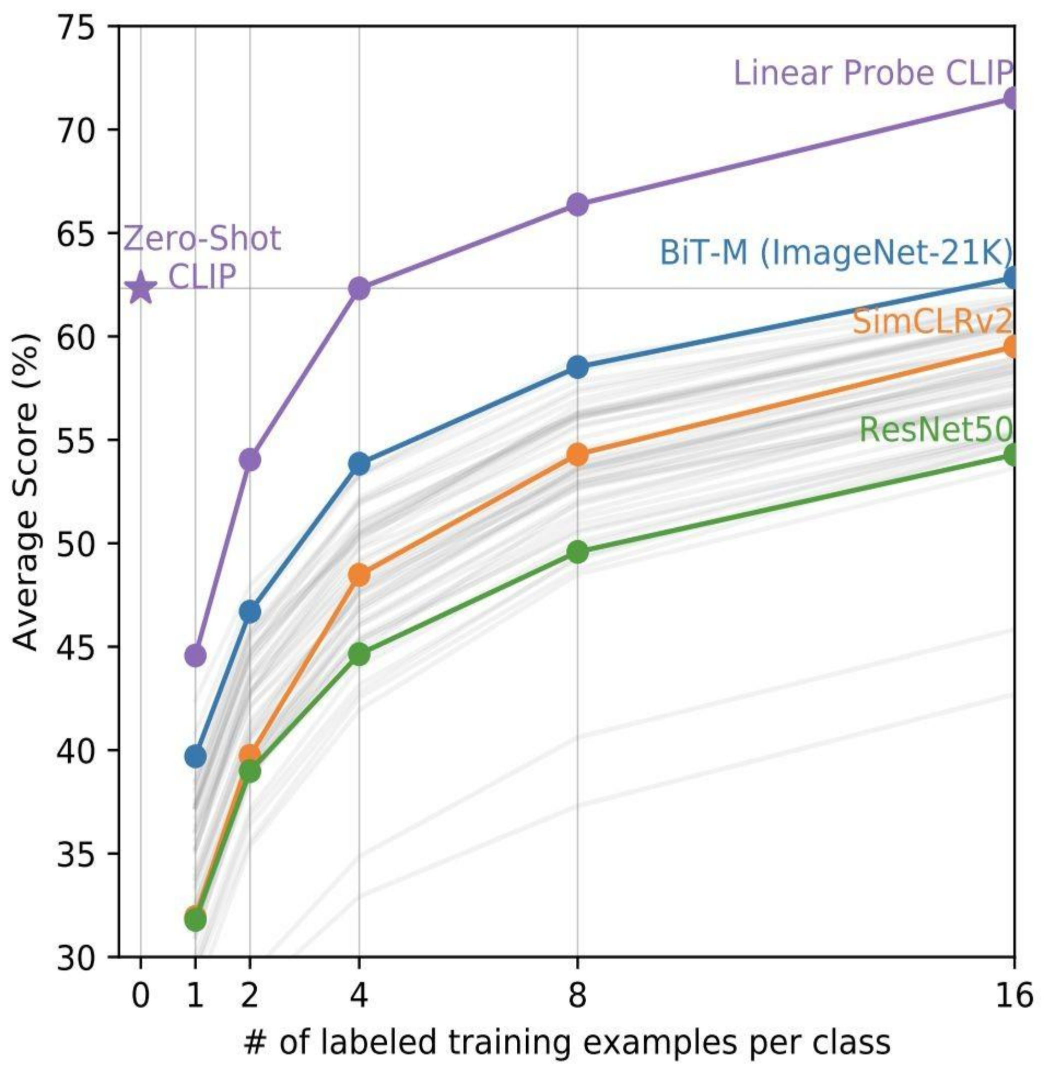
Linear probe performance vs SOTA vision models







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



TyZero-shot CLIP is as good as

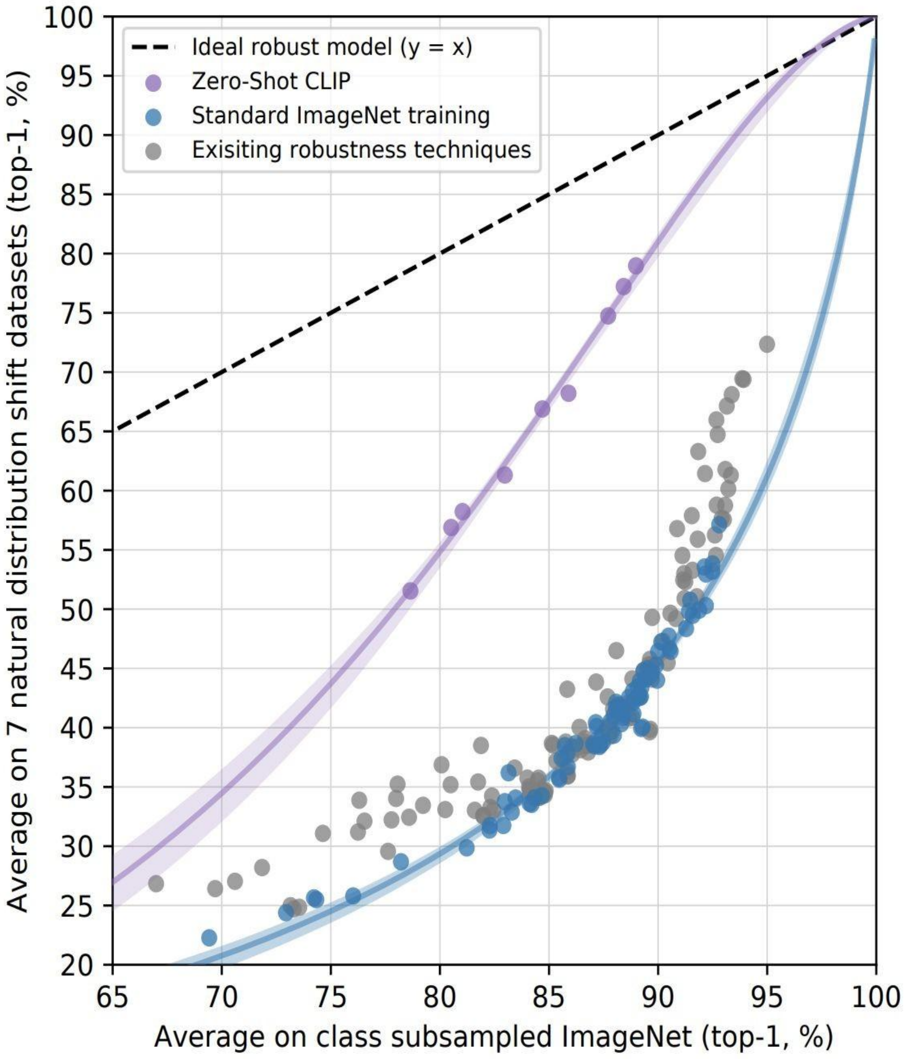
- 4-shot linear-probe CLIP
- 16-shot BiT-M

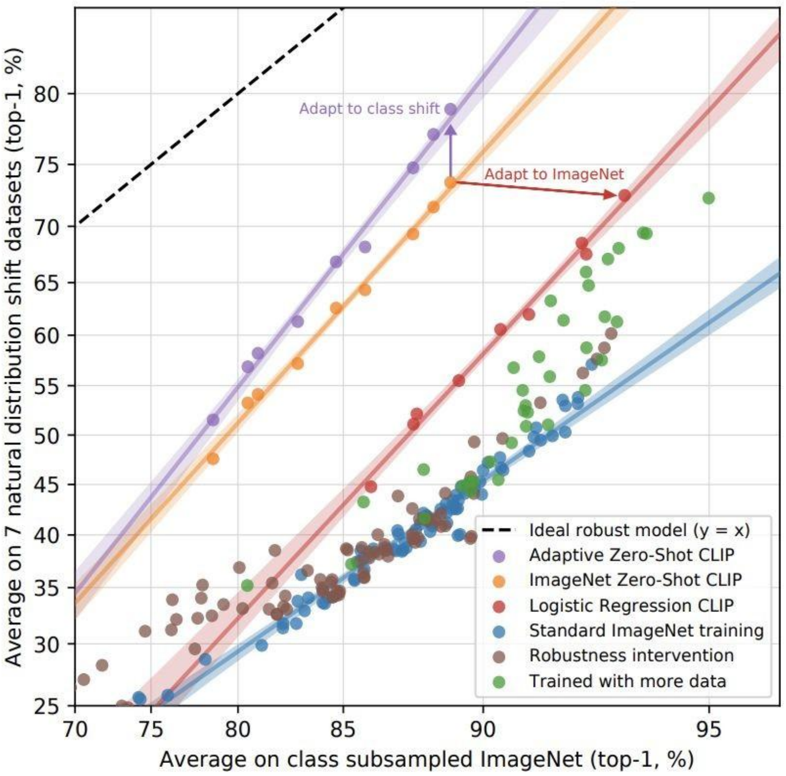
Robustness to natural distribution shift

Zero-Shot CLIP is much more robust!

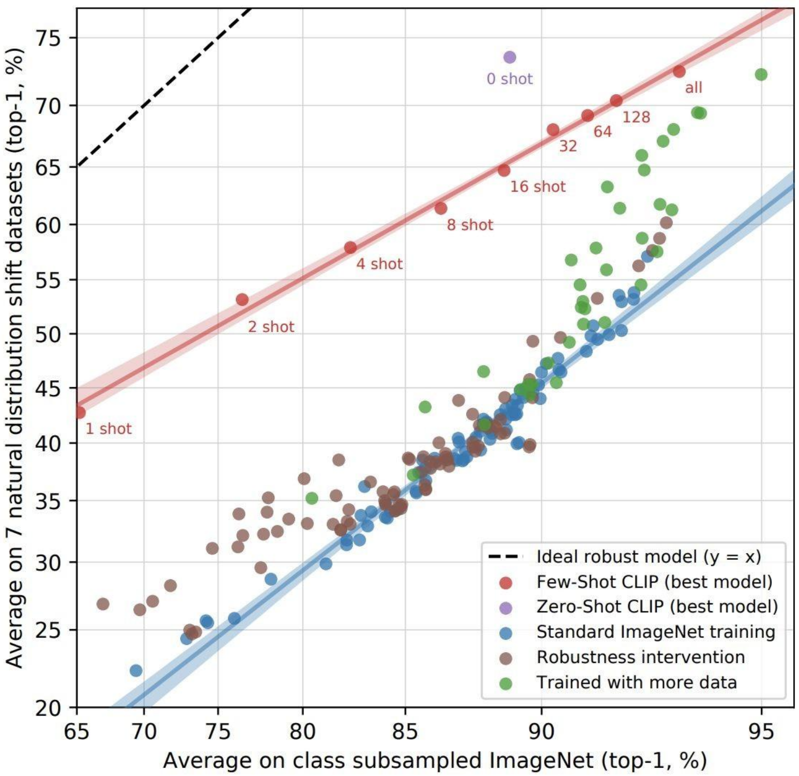
7 ImageNet-like Datasets

- 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

AYAHOO

building (97.7%) Ranked 1 out of 12

- ✓ a photo of a **building**.
- × a photo of a **carriage**.
- × a photo of a **statue**.
- × a photo of a **bag**.
- × a photo of a **mug**.

OBJECTNET IMAGENET OVERLAP

Pill bottle (98.3%) Ranked 1 out of 113

- ✓ a photo of a **pill bottle**.
- × a photo of a **bottle cap**.
- × a photo of a **beer bottle**.
- × a photo of a **pillow**.
- × a photo of a **wine bottle**.

IMAGENET BLURRY

marimba (79.5%) Ranked 1 out of 1000

- ✓ a photo of a **marimba**.
- × a photo of a **abacus**.
- × a photo of a **steel drum**.
- × a photo of a **computer keyboard**.
- × a photo of a **pool table**.

DESCRIBABLE TEXTURES DATASET (DTD)

perforated (20.5%) Ranked 2 out of 47

- × a photo of a **polka-dotted** texture.
- ✓ a photo of a **perforated** texture.
- × a photo of a **dotted** texture.
- × a photo of a **studded** texture.
- × a photo of a **freckled** texture.

KINETICS-700

country line dancing (99.0%) Ranked 1 out of 700

✓ a photo of **country line dancing**.

× a photo of **square dancing**.

× a photo of **swing dancing**.

× a photo of **dancing charleston**.

× a photo of **salsa dancing**.

FLOWERS-102

great masterwort (74.3%) Ranked 1 out of 102

✓ a photo of a **great masterwort**, a type of flower.

× a photo of a **bishop of llandaff**, a type of flower.

× a photo of a **pincushion flower**, a type of flower.

× a photo of a **globe flower**, a type of flower.

× a photo of a **prince of wales feathers**, a type of flower.

IMAGENET

King Charles Spaniel (91.6%) Ranked 1 out of 1000

✓ a photo of a **king charles spaniel**.

× a photo of a **brittany dog**.

× a photo of a **cocker spaniel**.

× a photo of a **papillon**.

× a photo of a **sussex spaniel**.

BIRDSNAP

Black chinned Hummingbird (12.0%) Ranked 4 out of 500

× a photo of a **broad tailed hummingbird**, a type of bird.

× a photo of a **calliope hummingbird**, a type of bird.

× a photo of a **costas hummingbird**, a type of bird.

✓ a photo of a **black chinned hummingbird**, a type of bird.

× a photo of a **annas hummingbird**, a type of bird.

COUNTRY211

Belize (3.9%) Ranked 5 out of 211

× a photo i took in **french guiana**.

× a photo i took in **gabon**.

× a photo i took in **cambodia**.

× a photo i took in **guyana**.

✓ a photo i took in **belize**.

RESISC45

roundabout (96.4%) Ranked 1 out of 45

✓ satellite imagery of **roundabout**.

× satellite imagery of **intersection**.

× satellite imagery of **church**.

× satellite imagery of **medium residential**.

× satellite imagery of **chaparral**.

STANFORD CARS

2012 Honda Accord Coupe (63.3%) Ranked 1 out of 196

✓ a photo of a **2012 honda accord coupe**.

× a photo of a **2012 honda accord sedan**.

× a photo of a **2012 acura tl sedan**.

× a photo of a **2012 acura tsx sedan**.

× a photo of a **2008 acura tl type-s**.

SUN

kennel indoor (98.6%) Ranked 1 out of 723

✓ a photo of a **kennel indoor**.

× a photo of a **kennel outdoor**.

× a photo of a **jail cell**.

× a photo of a **jail indoor**.

× a photo of a **veterinarians office**.

PASCAL VOC 2007

motorcycle (99.7%) Ranked 1 out of 20

- ✓ a photo of a **motorcycle**.
- × a photo of a **bicycle**.
- × a photo of a **car**.
- × a photo of a **horse**.
- × a photo of a **dining table**.

MNIST

7 (85.3%) Ranked 1 out of 10

- ✓ a photo of the number: "**7**".
- × a photo of the number: "**2**".
- × a photo of the number: "**1**".
- × a photo of the number: "**6**".
- × a photo of the number: "**4**".

STREET VIEW HOUSE NUMBERS (SVHN)

158 (0.3%) Ranked 83 out of 2000

- × a street sign of the number: "**1157**".
- × a street sign of the number: "**1165**".
- × a street sign of the number: "**1164**".
- × a street sign of the number: "**1155**".
- × a street sign of the number: "**1364**".

IMAGENET VID

antelope (99.8%) Ranked 1 out of 30

- ✓ a photo of a **antelope**.
- × a photo of a **zebra**.
- × a photo of a **car**.
- × a photo of a **cattle**.
- × a photo of a **elephant**.

Limitations of CLIP

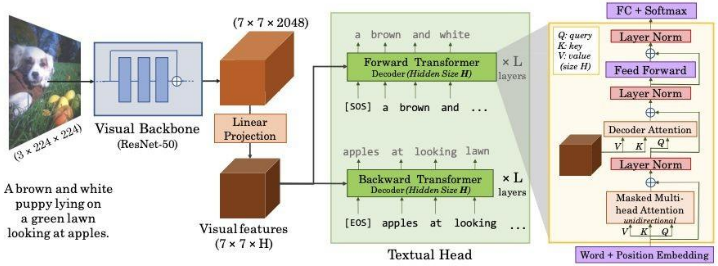
- Zero-shot performance is well below the SOTA
- Especially weak on abstract tasks such as counting
- Poor on out-of-distribution data such as MNIST
- Susceptible to adversarial attacks
- Dataset selection in the eval suite, use of large validation sets for prompt engineering
- Social biases

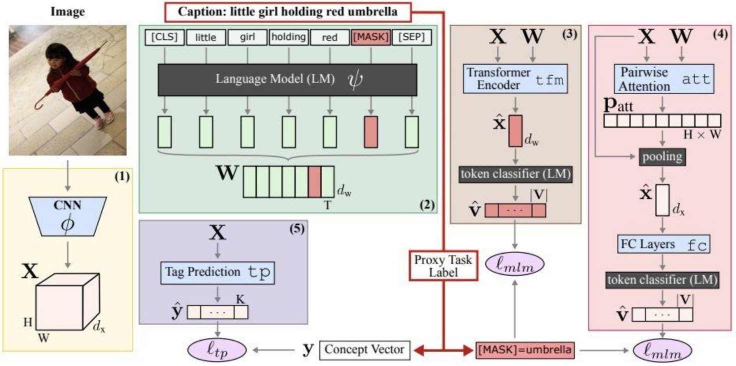
Related Works

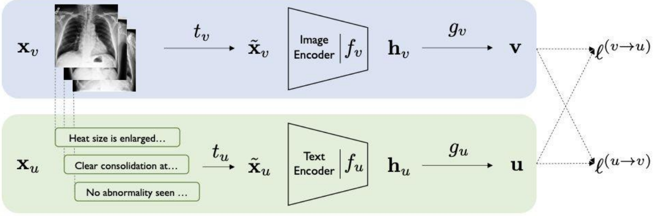
Natural language supervision:
- YFCC100M WSL (Joulin et al.)
- VirTex (Desai and Johnson)
- ICMLM (Sariyildiz et al.)
- ConVIRT (Zhang et al.)

Zero-Shot Transfer:
- Visual N-Grams (Li et al.)

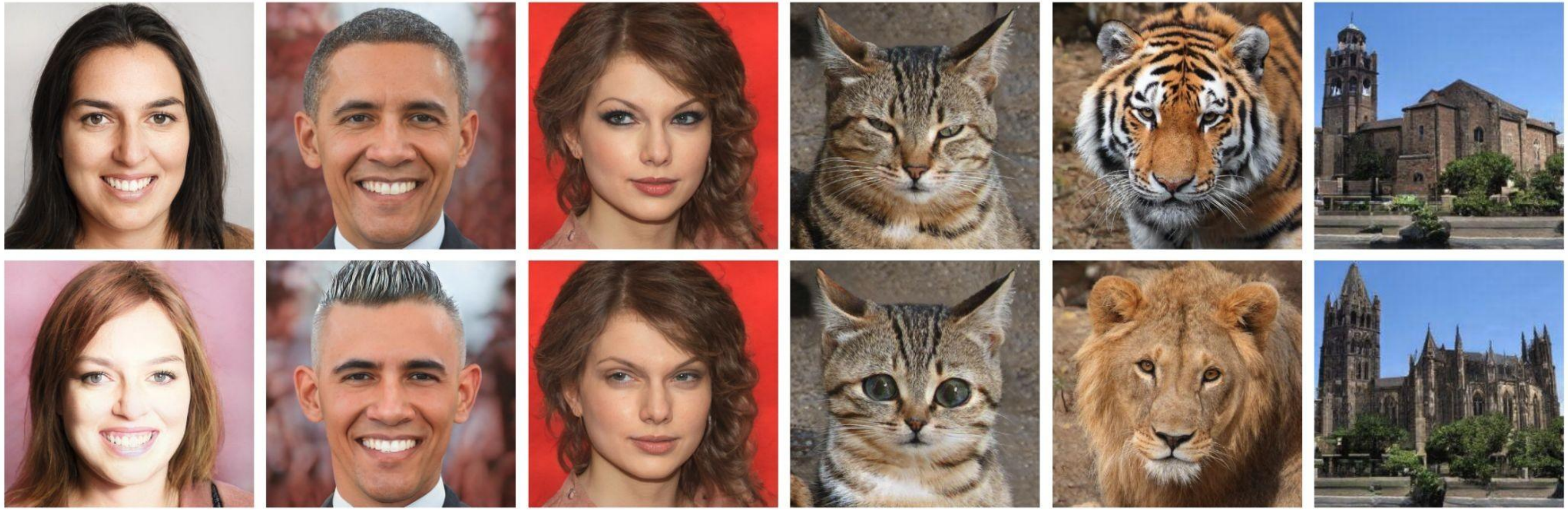
Broad Evaluation and Robustness:
- VTAB (Zhang et al.)
- ImageNet Testbed (Taori et al.)







Applications of CLIP



“Emma Stone” “Mohawk hairstyle” “Without makeup” “Cute cat” “Lion” “Gothic church”

StyleCLIP (Patashnik et al.)
Steering a GAN Using CLIP

Applications of CLIP

A banquet hall

Geoffrey Hinton

Dogs playing poker

Thank You!..