Learning Transferable Visual Models From Natural Language Supervision

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

Under supervision of Dr. Jimson Mathew

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!

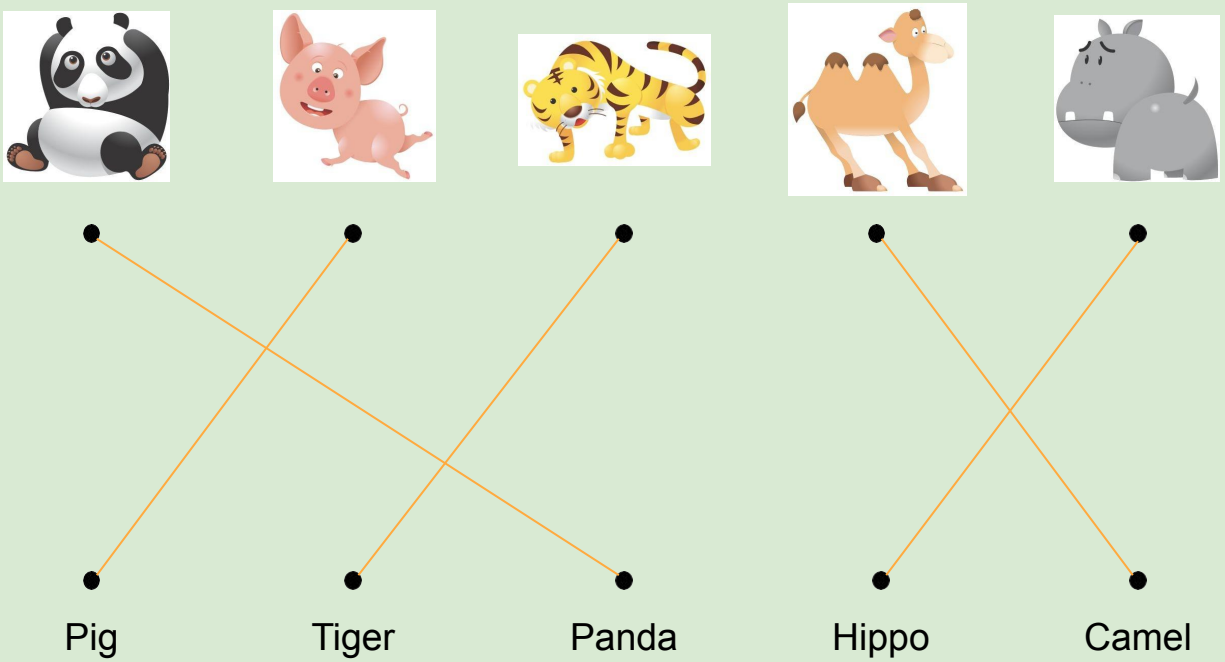
Contrastive Learning





Pig Tiger Panda Hippo Camel

Contrastive Learning



Predictive Approach

the dog is on the table

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

Contrastive Approach

Siberian Husky (76.0%) Ranked 1 out of 200

✓ a photo of a **siberian husky**.

× a photo of a **german shepherd dog**.

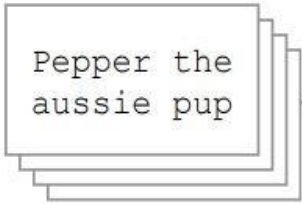
× a photo of a **collie**.

× a photo of a **border collie**.

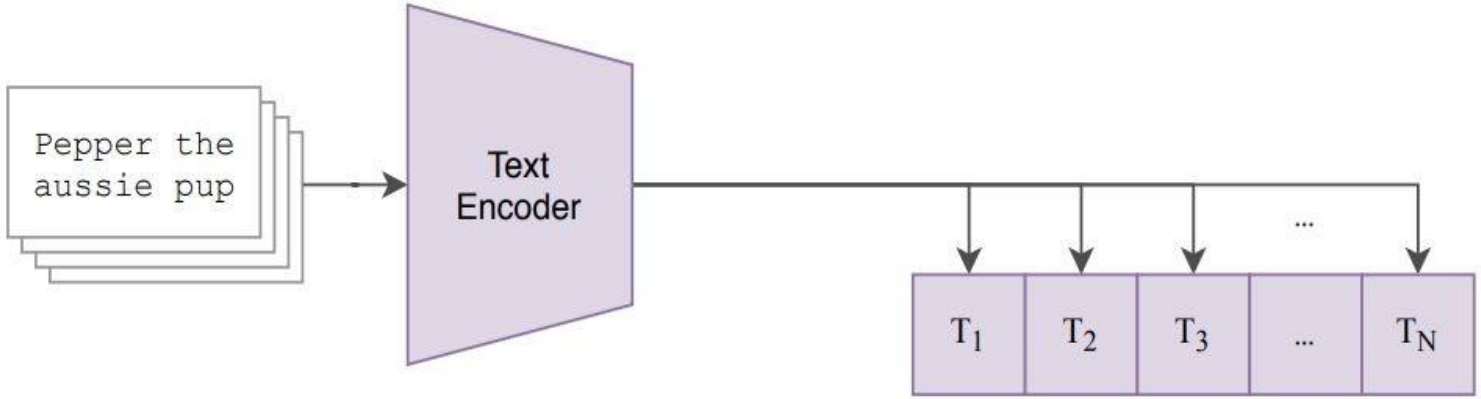
× a photo of a **rottweiler**.

a photo of a siberian husky

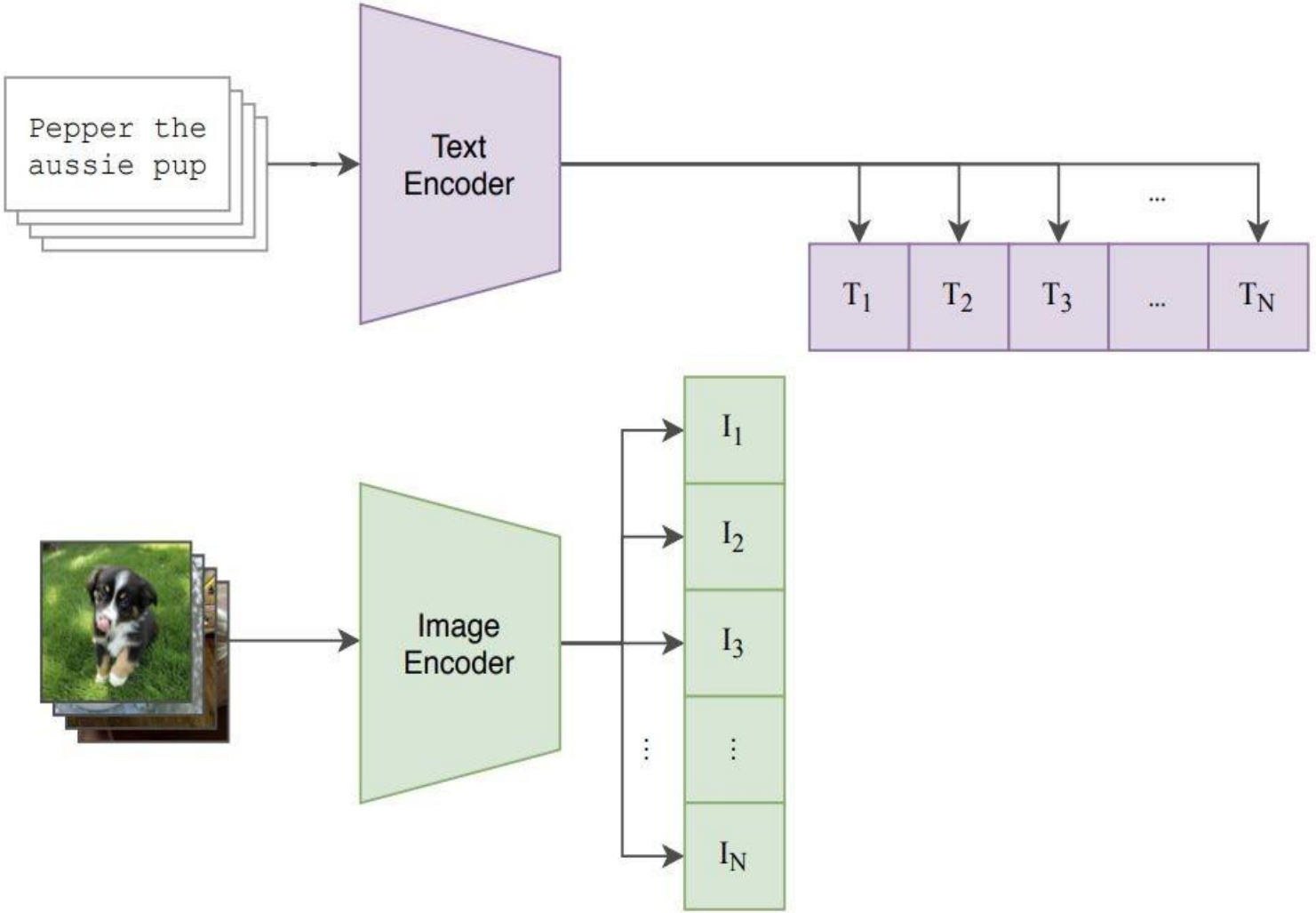
Contrastive pre training



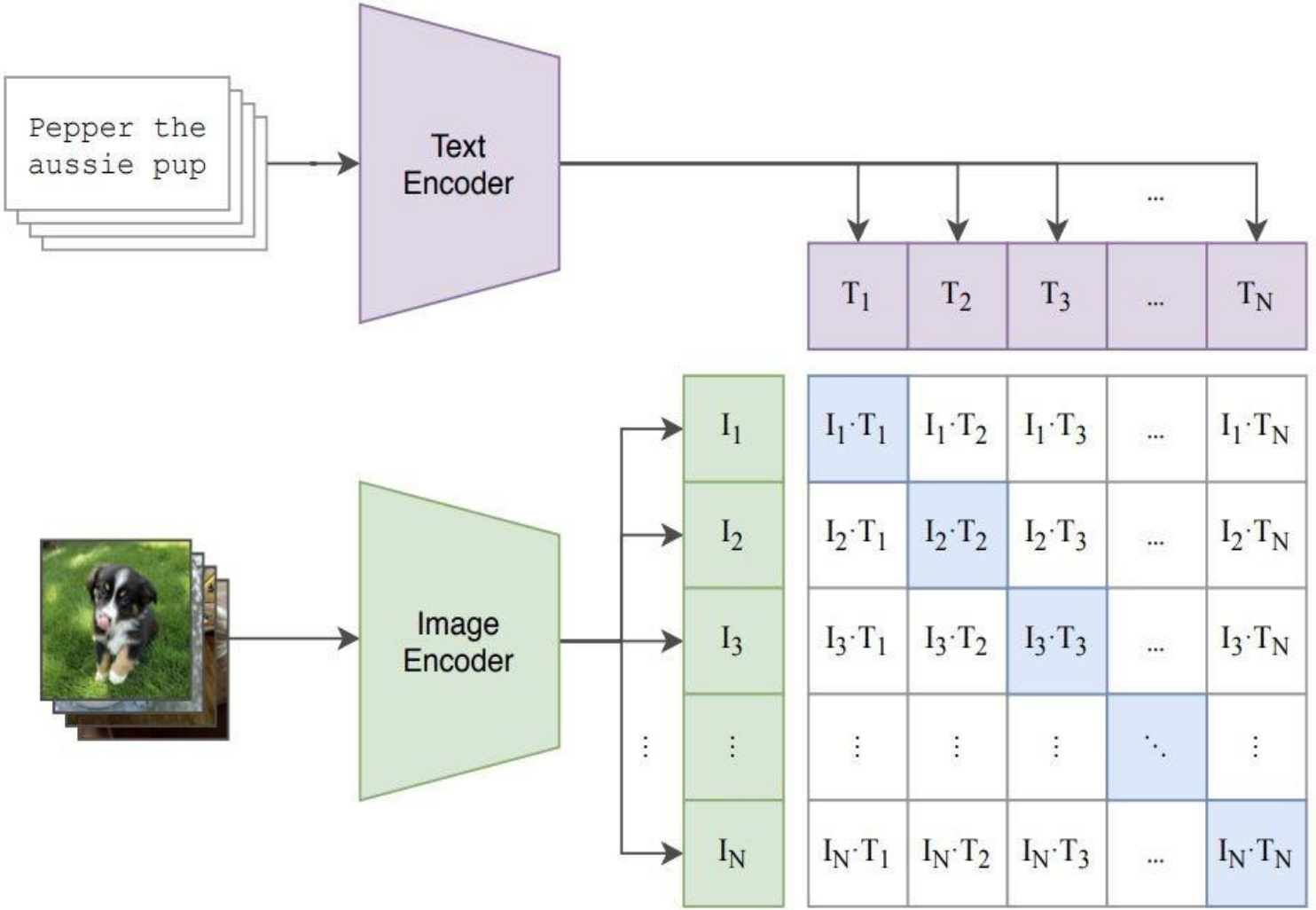
Contrastive pre training



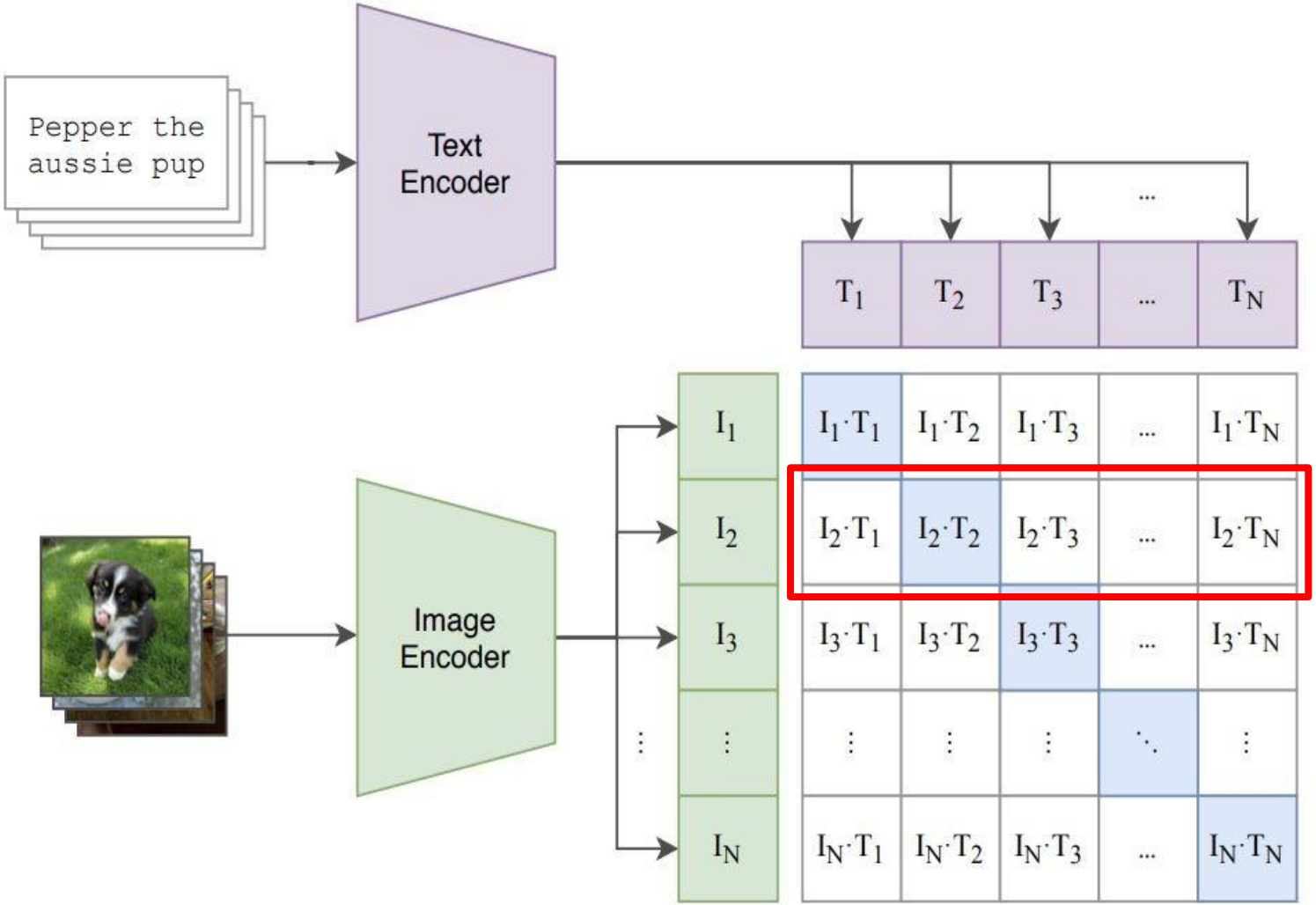
Contrastive pre training



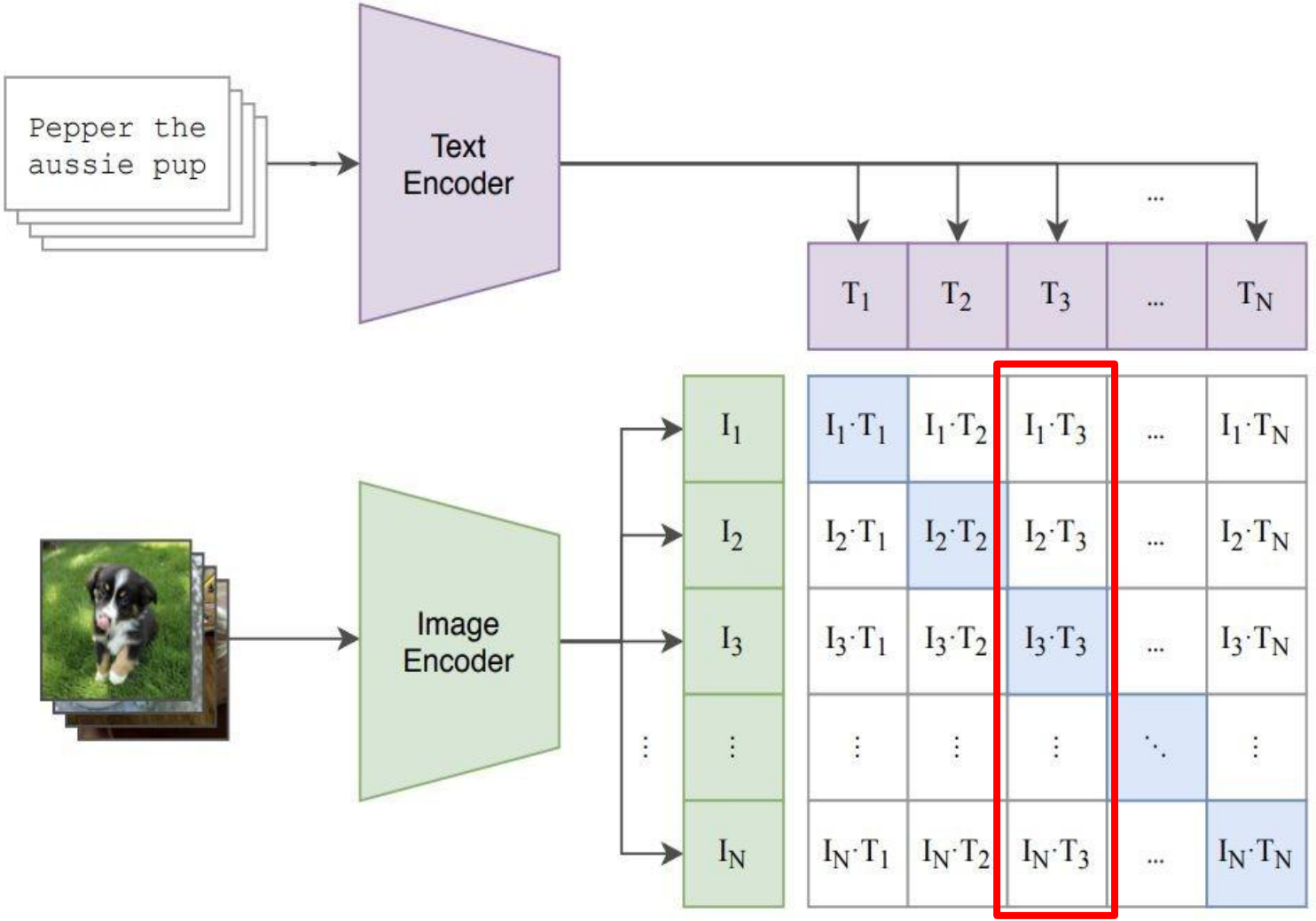
Contrastive pre training



Contrastive pre training



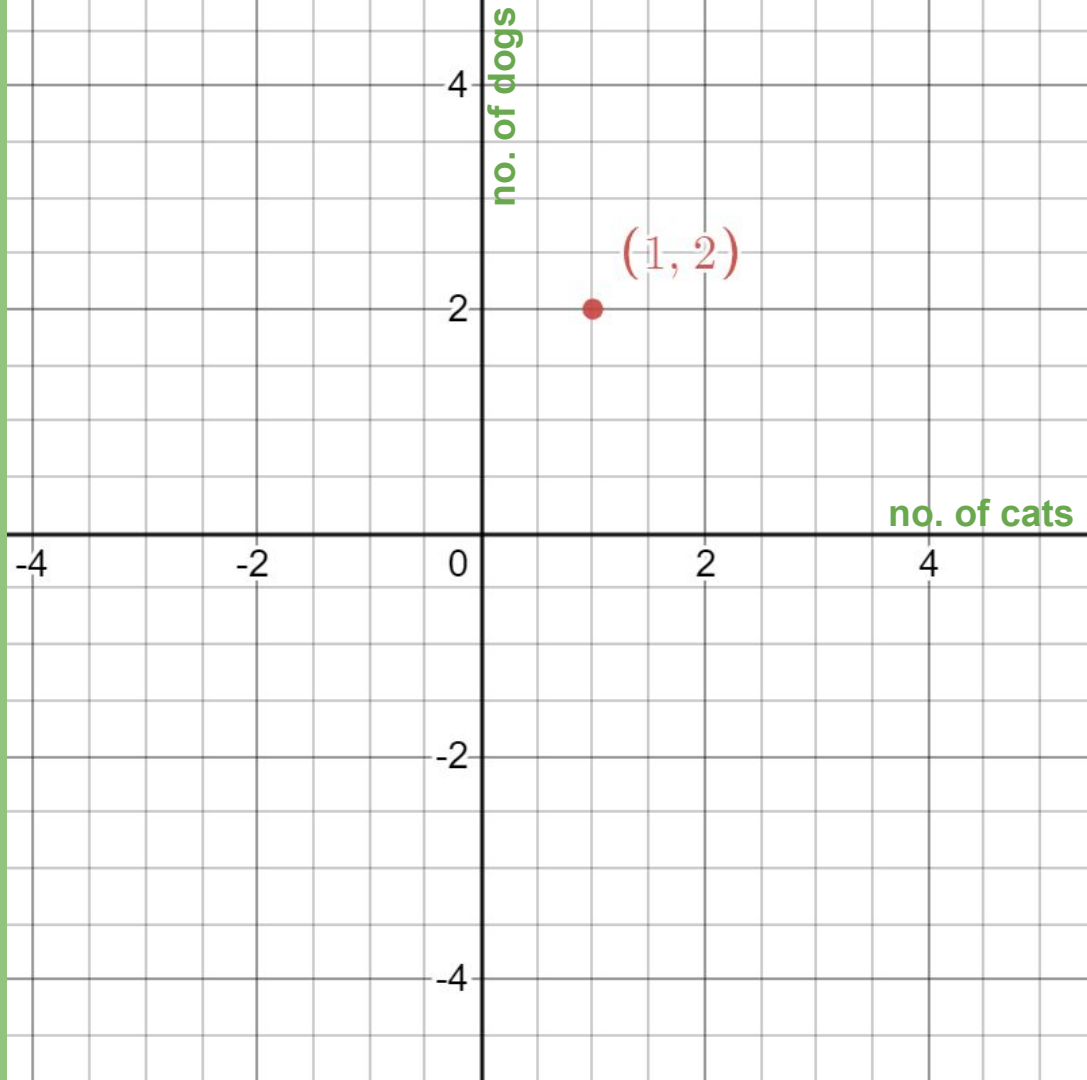
Contrastive pre training



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!

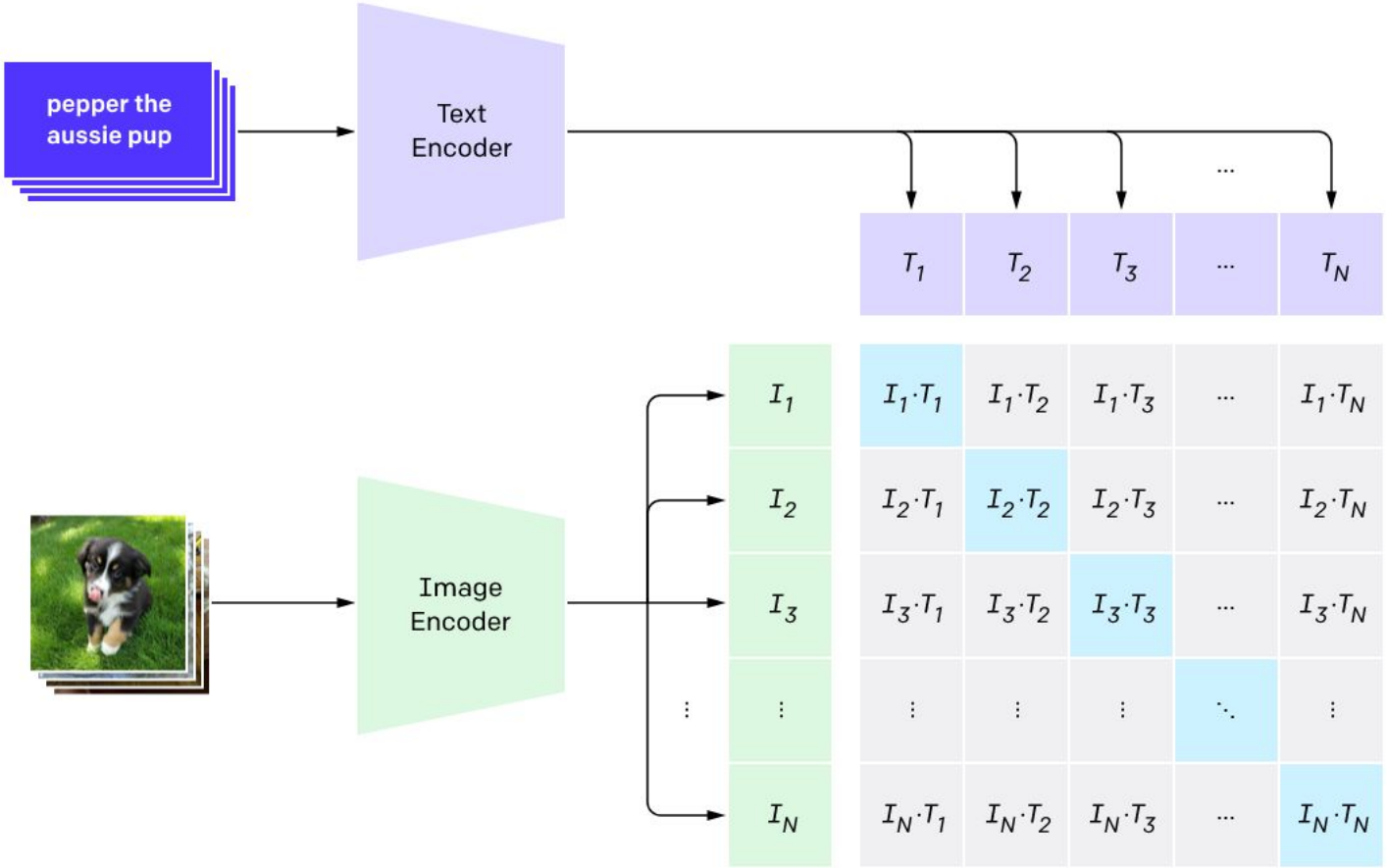


How Embedding Works...

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.

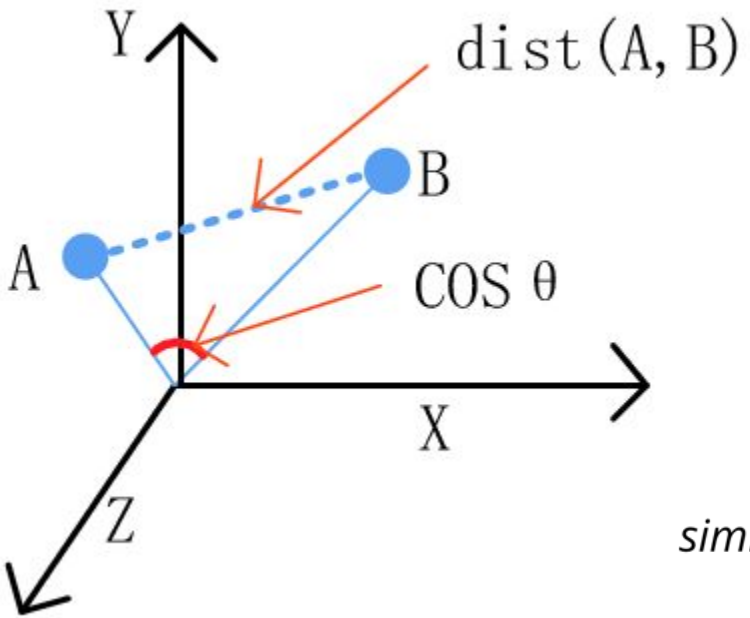
Measuring “Goodness” & “Badness”



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).

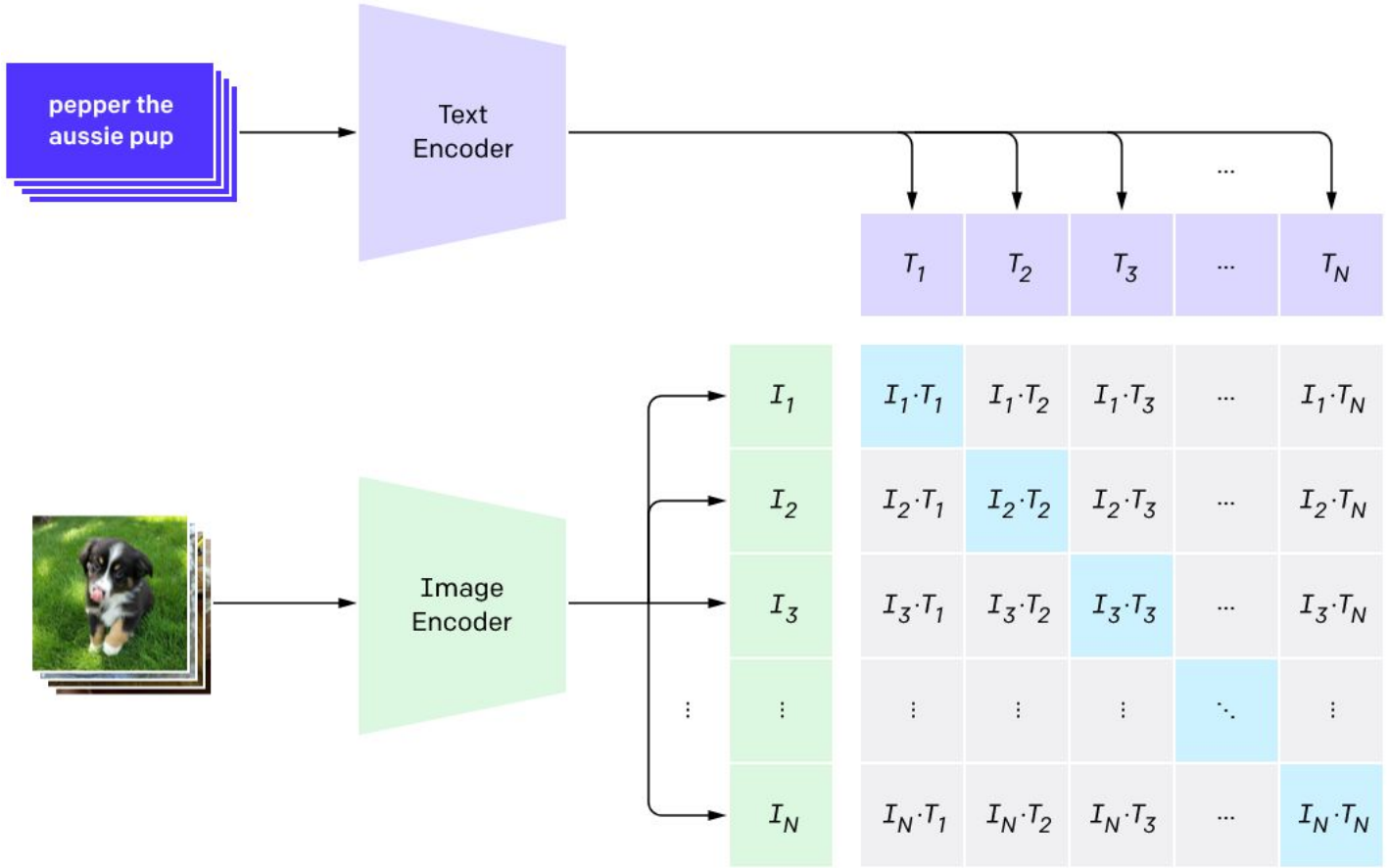
Cosine Similarity



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**.

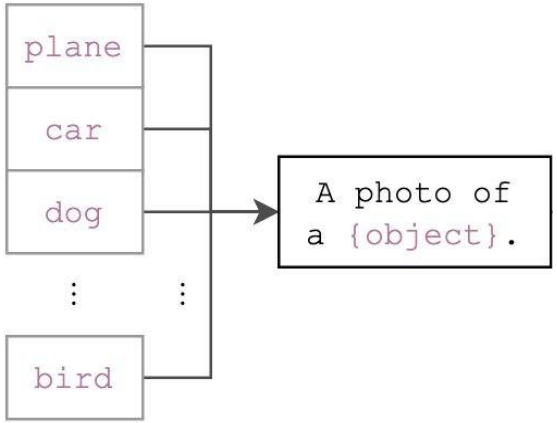
$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Measuring “Goodness” & “Badness”

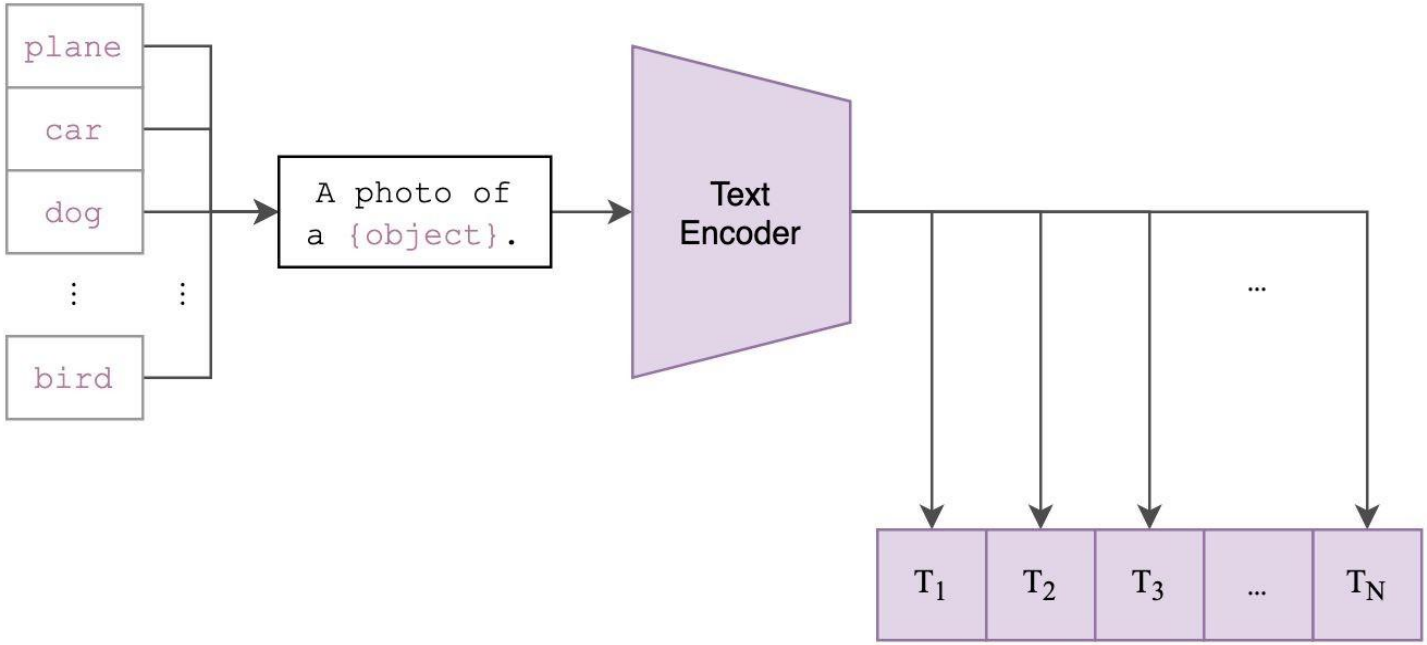


Zero-shot prediction

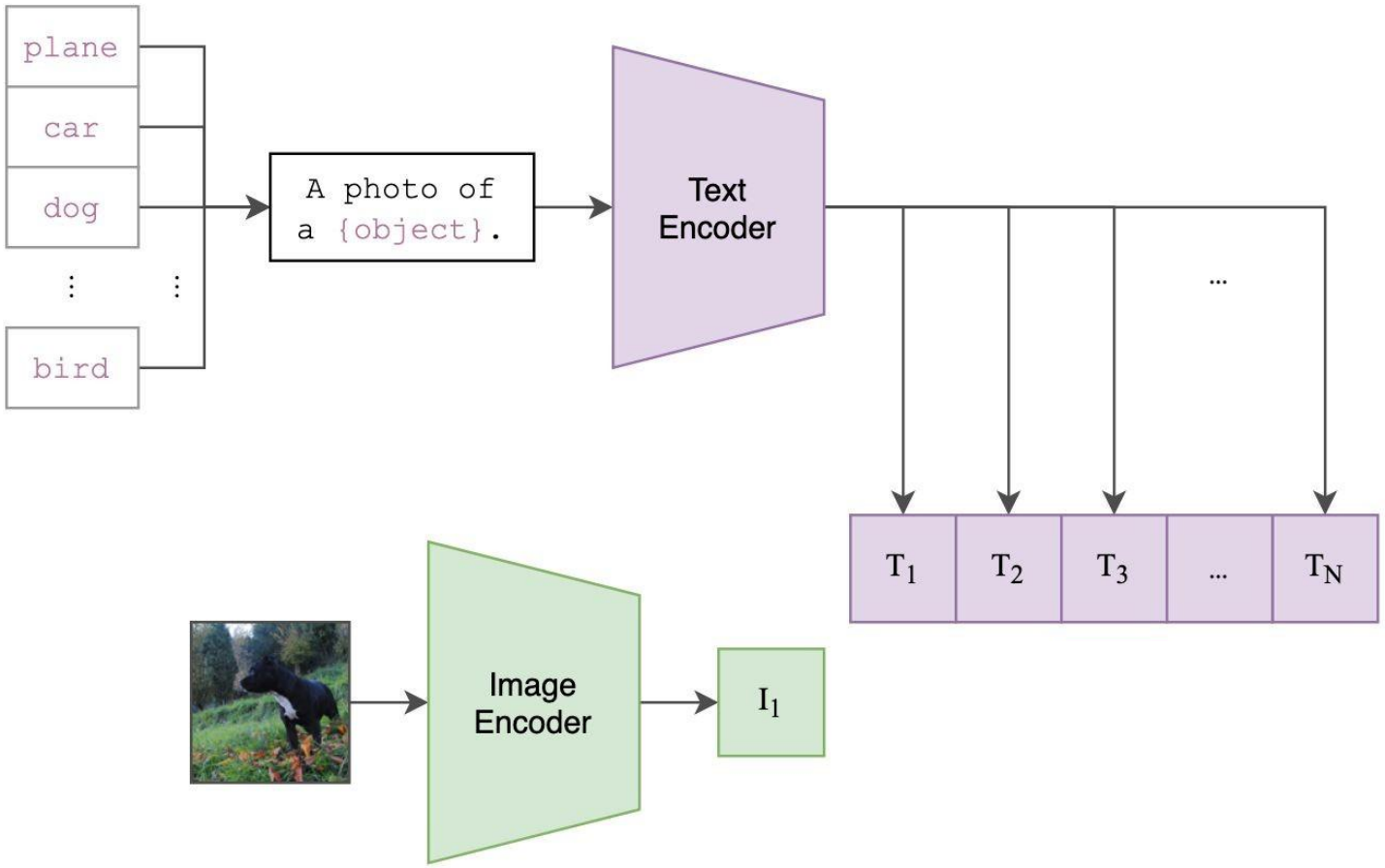
Zero-shot prediction



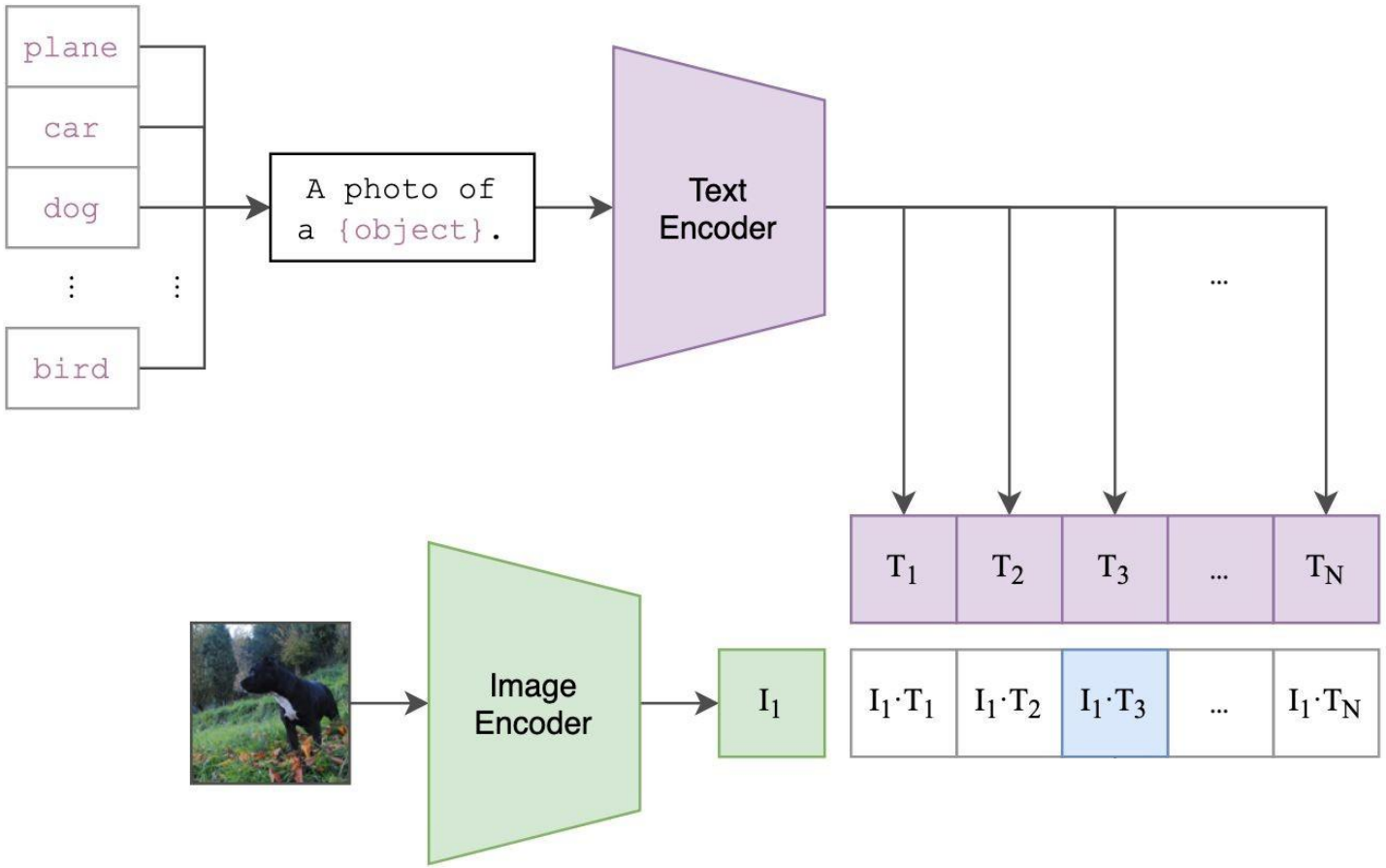
Zero-shot prediction



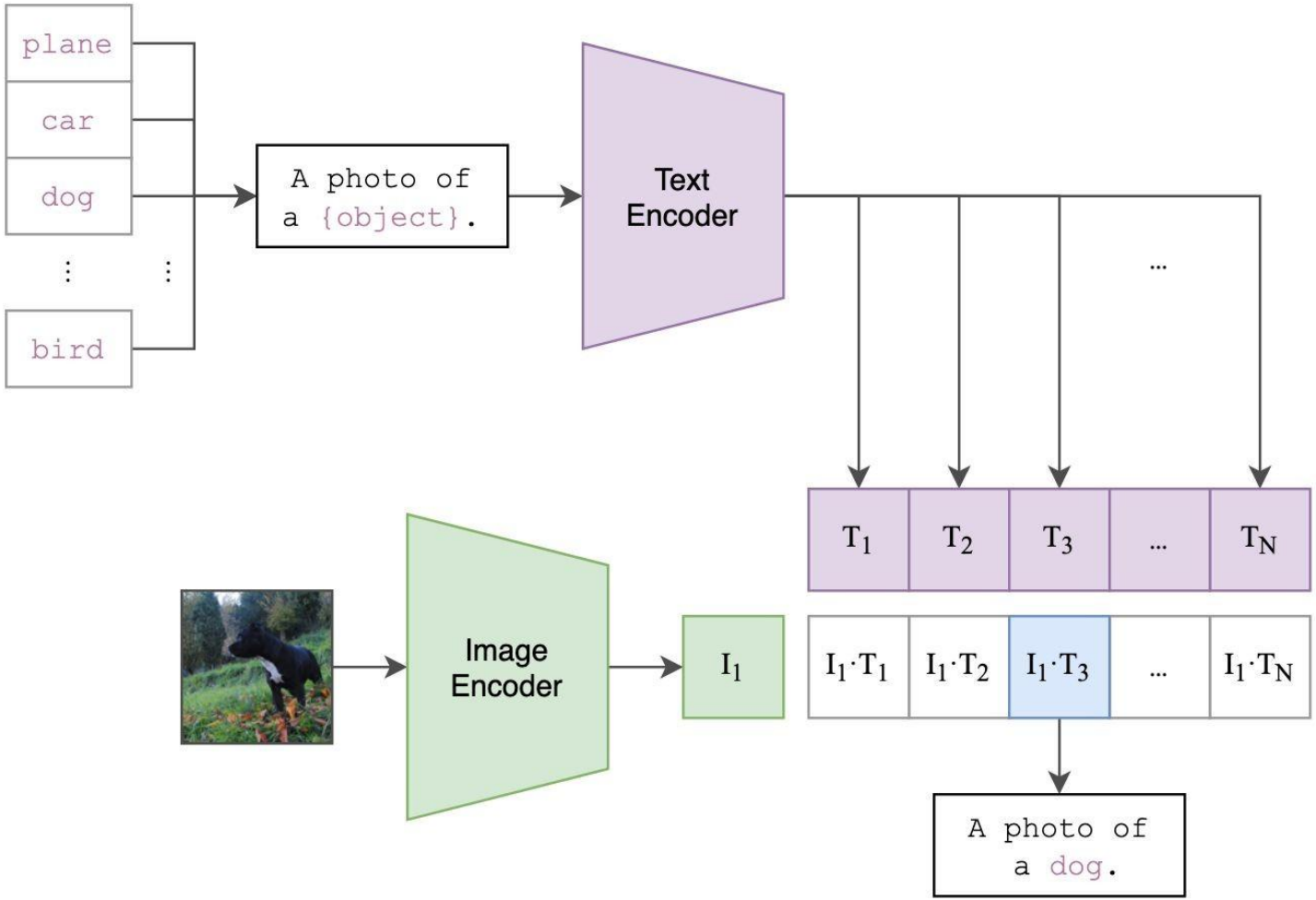
Zero-shot prediction



Zero-shot prediction



Zero-shot prediction

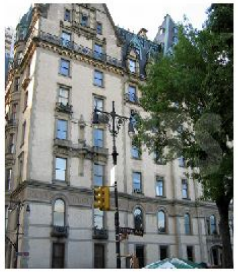


Zero-shot learning is when a model attempts to predict a class it saw zero times in the training data

CLIP as a bridge between Computer Vision & Natural Language Processing

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.

Thank You!..