Image search như Google với MultiModal Model

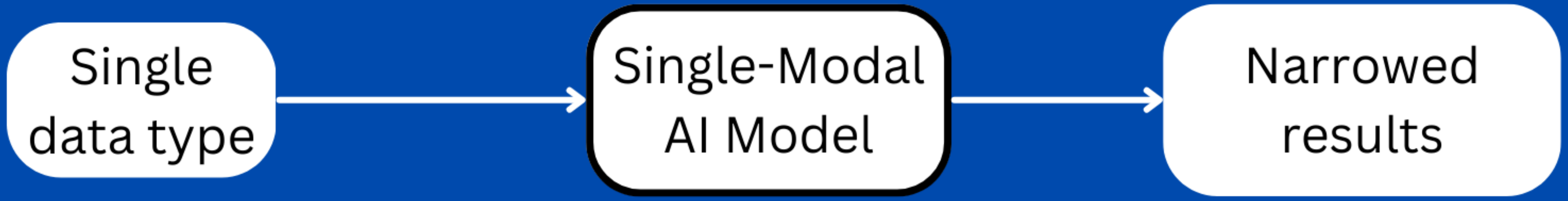
Mì AI

MultiModal Model

Multimodal models are a class of machine learning models that can process and integrate information from **multiple modalities** or types of data. Unlike traditional models that handle a single type of input (e.g., text or image), multimodal models can simultaneously handle and learn from different types of inputs such as text, images, audio, and more. Pretty cool!

Figure 1: The different types of data that multimodal models can handle simultaneously.

Single-modal AI Model

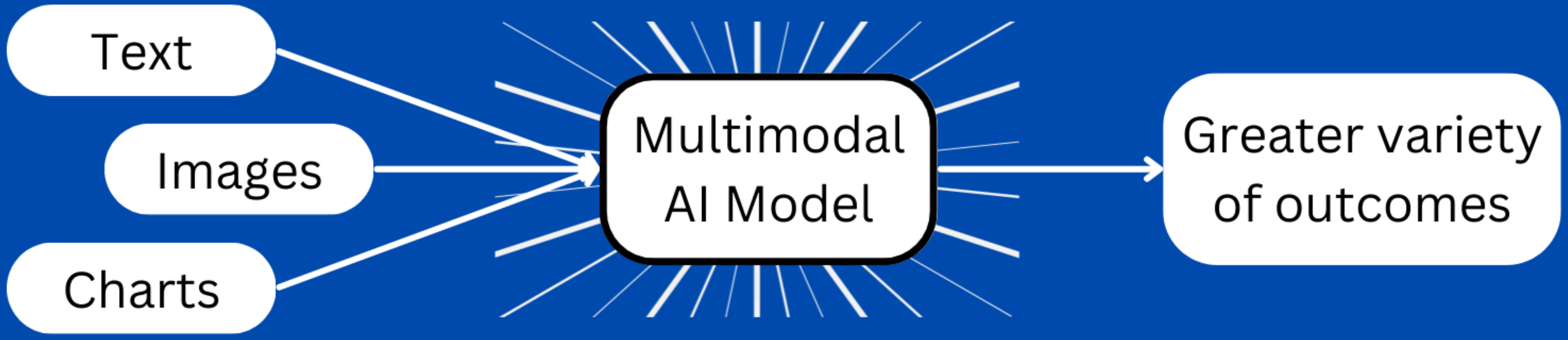


Single data type

Single-Modal AI Model

Narrowed results

Multimodal AI Model?



Text

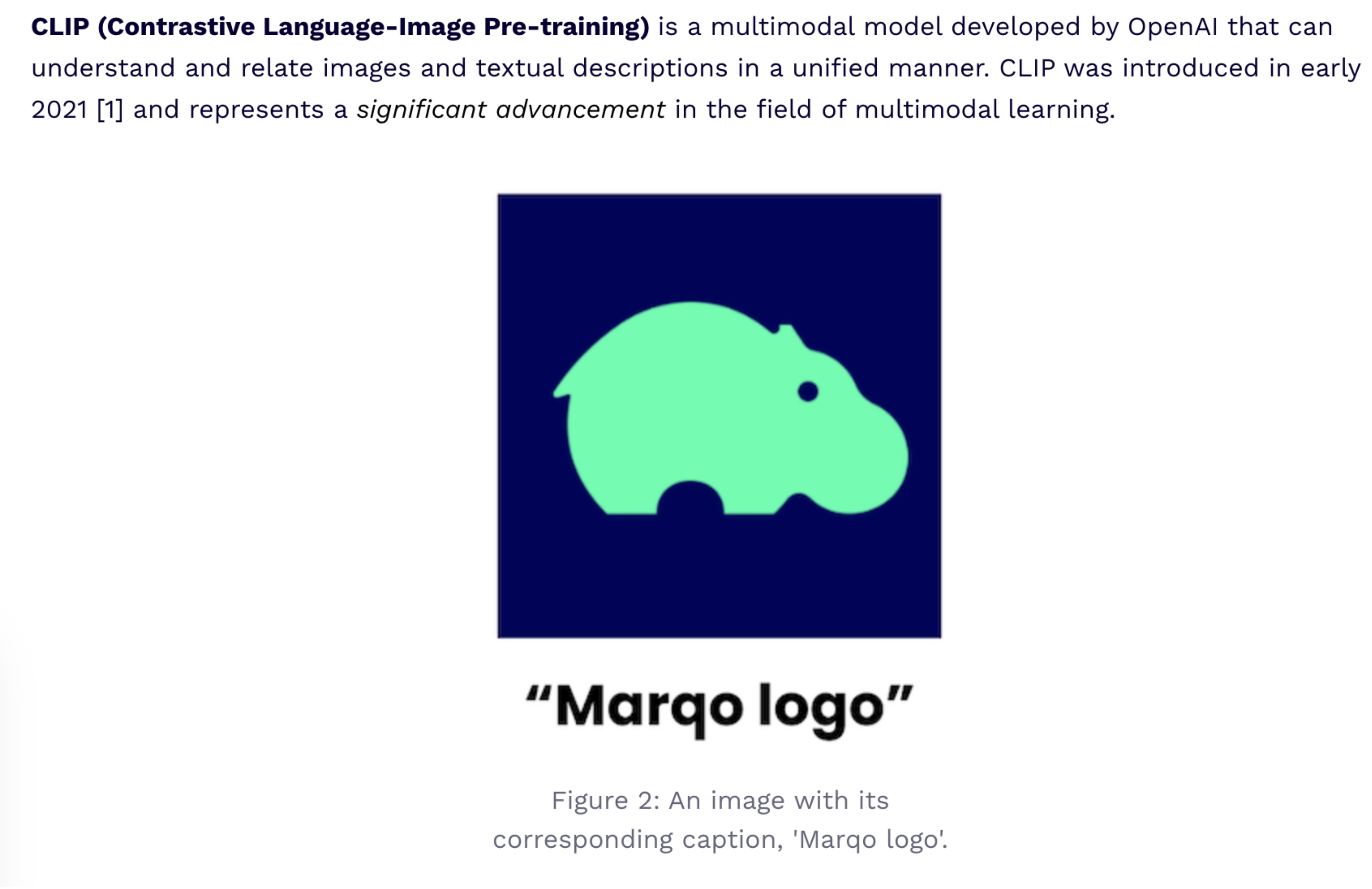
Images

Charts

Multimodal AI Model

Greater variety of outcomes

What is CLIP?

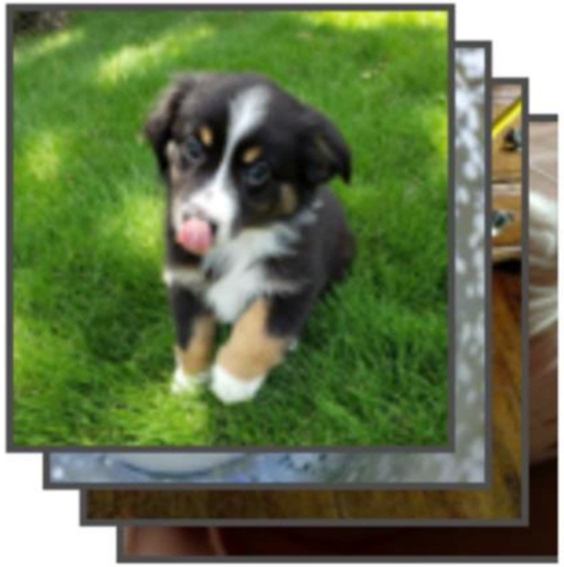


CLIP (Contrastive Language-Image Pre-training) is a multimodal model developed by OpenAI that can understand and relate images and textual descriptions in a unified manner. CLIP was introduced in early 2021 [1] and represents a *significant advancement* in the field of multimodal learning.

“Marqo logo”

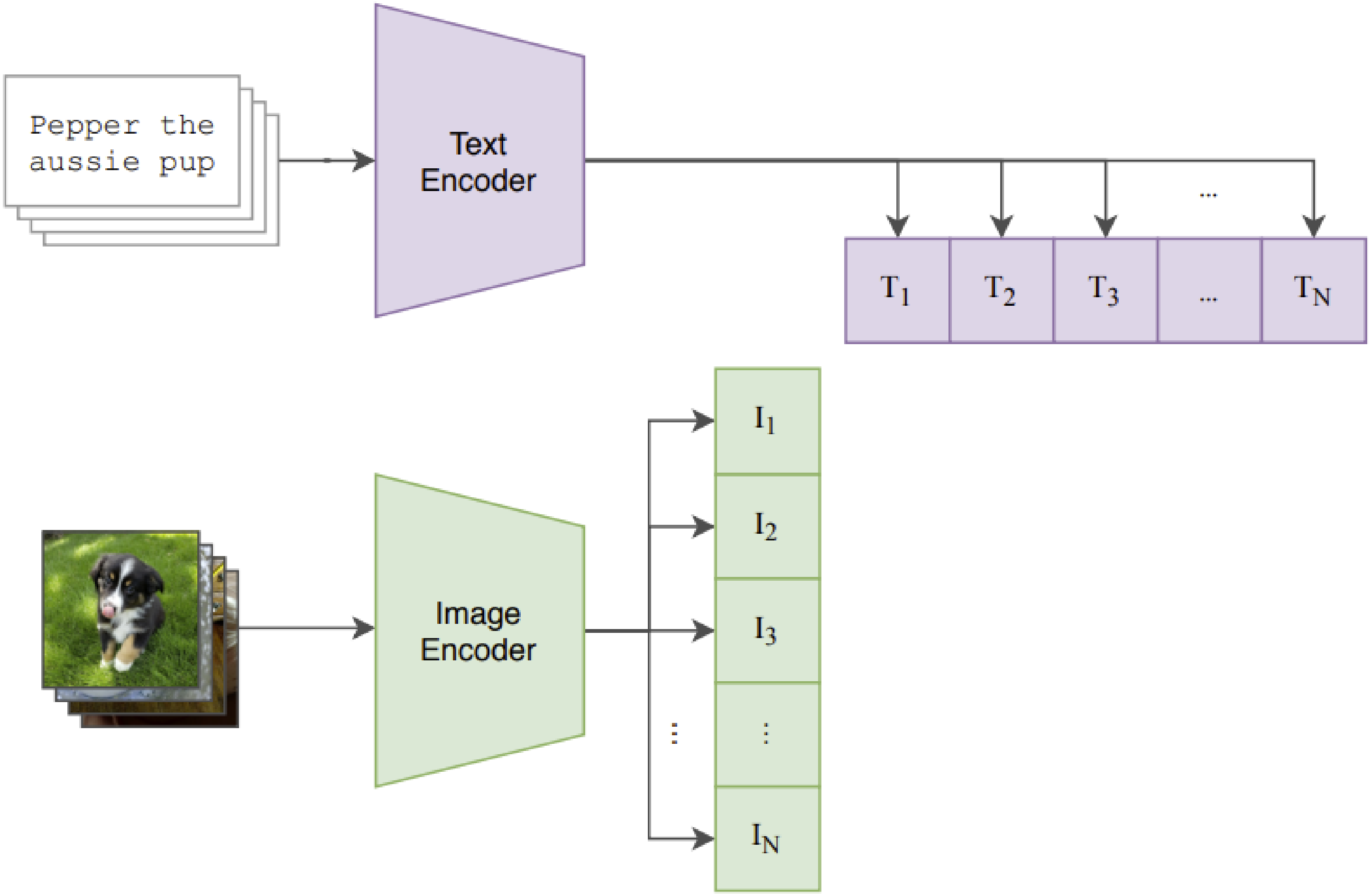
Figure 2: An image with its corresponding caption, 'Marqo logo'.

How to train CLIP?

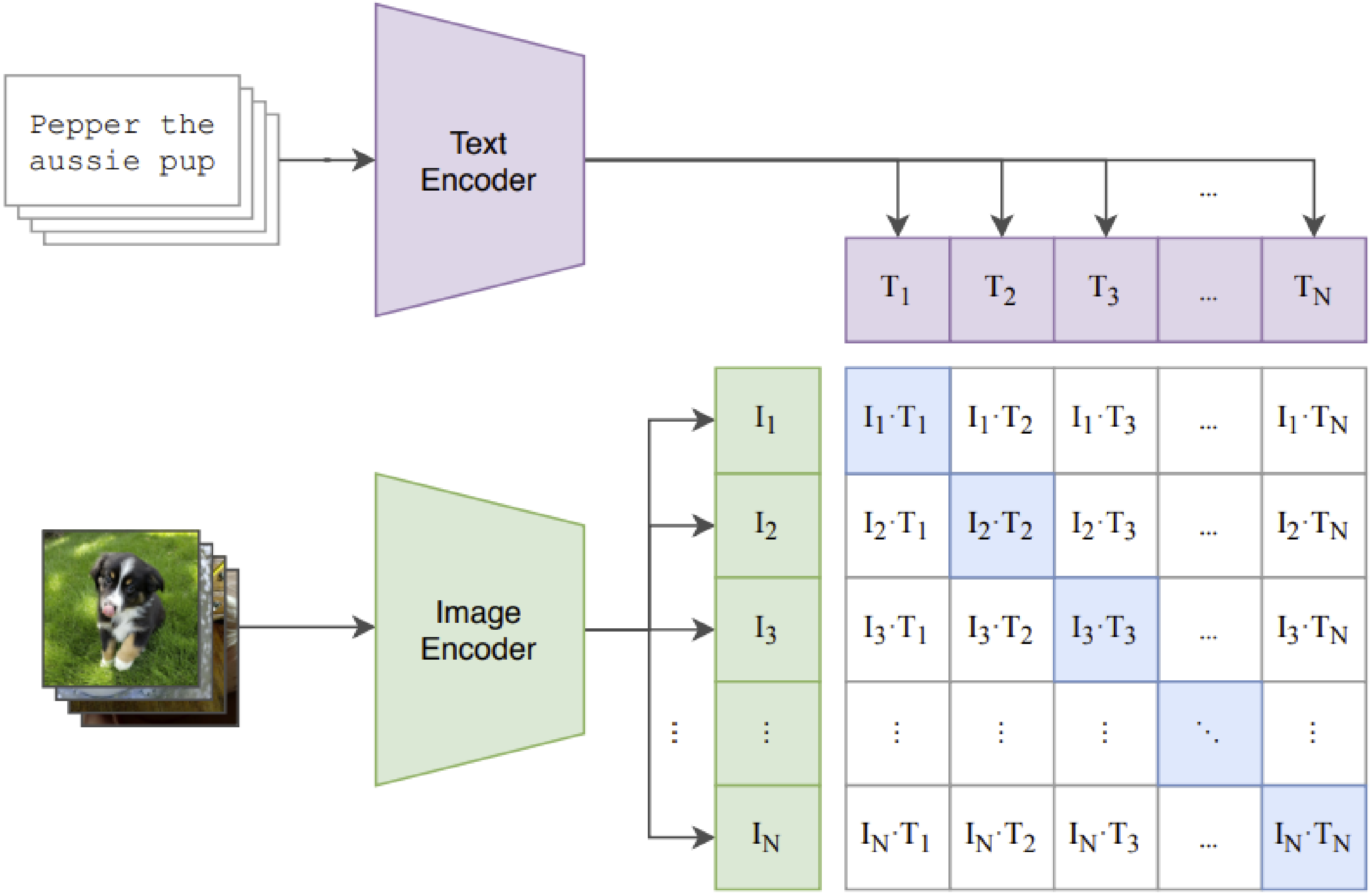


Pepper the aussie pup

How to train CLIP?

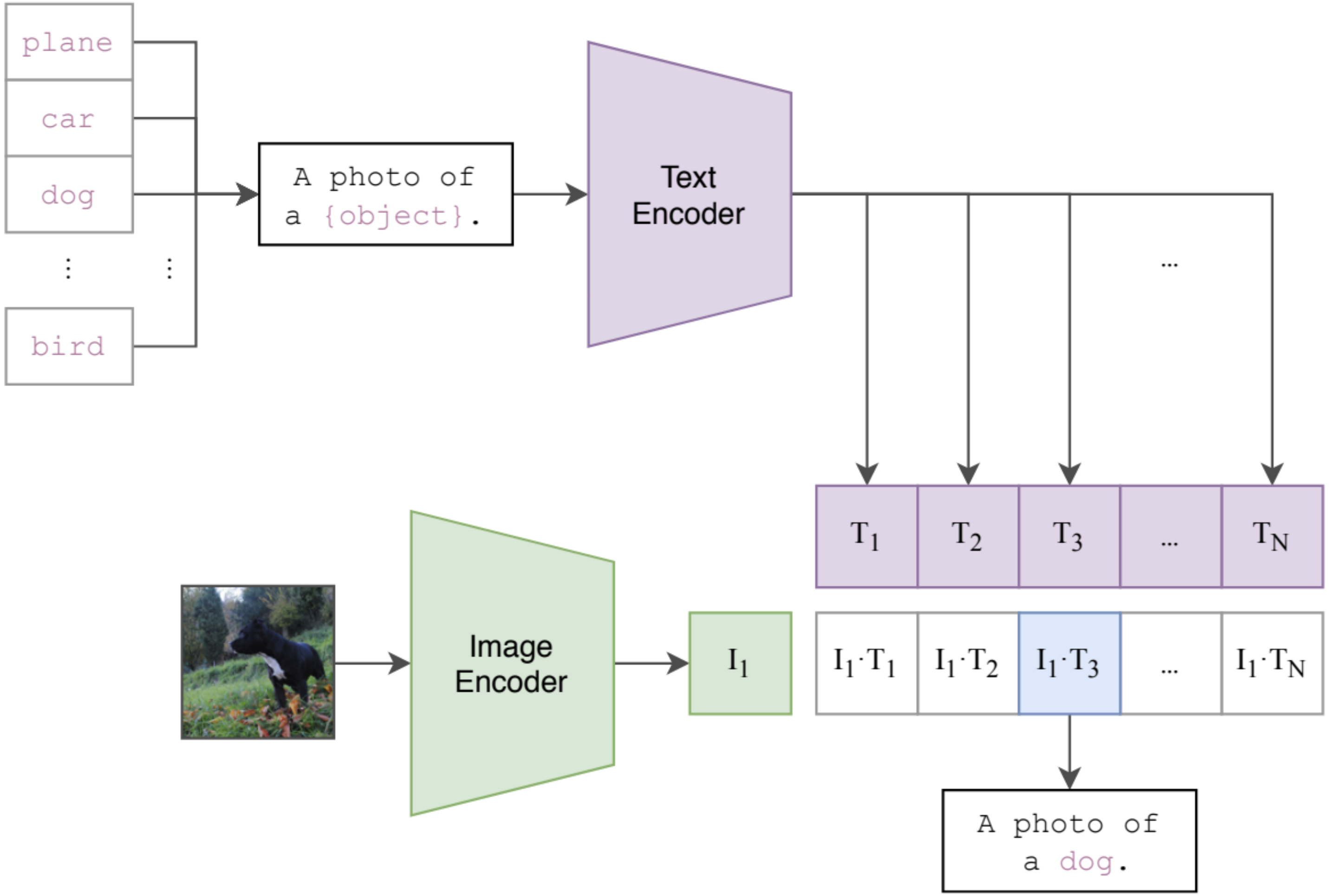


How to train CLIP?

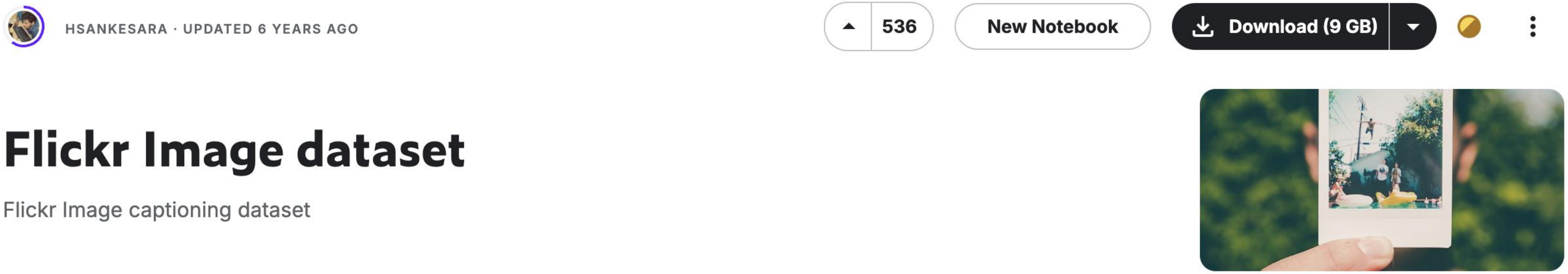


	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

Zero shot Learning



Handson



HSANKESARA · UPDATED 6 YEARS AGO

Flickr Image dataset

Flickr Image captioning dataset

https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset?resource=download