Novice CV/VLM Workshop

Image Processing, Computer Vision, and Your First Vision Language Model

Ming Hao
Data Scientist @ FWD

Table of Contents

TIL-AI
TODAY I LEARNED AI

1

Introduction to Computer Vision

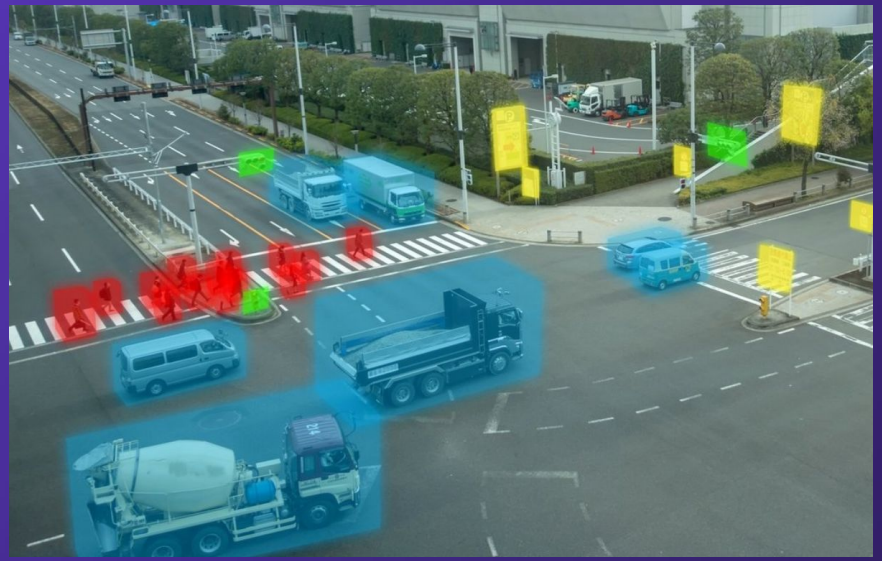
Fundamentals of Computer Vision

TIL-AI
TODAY I LEARNED AI

What is Computer Vision?

- A branch of AI that teaches computers to see and understand visual data in a way that mimics human vision

- Utilizes algorithms to analyse and interpret visual inputs

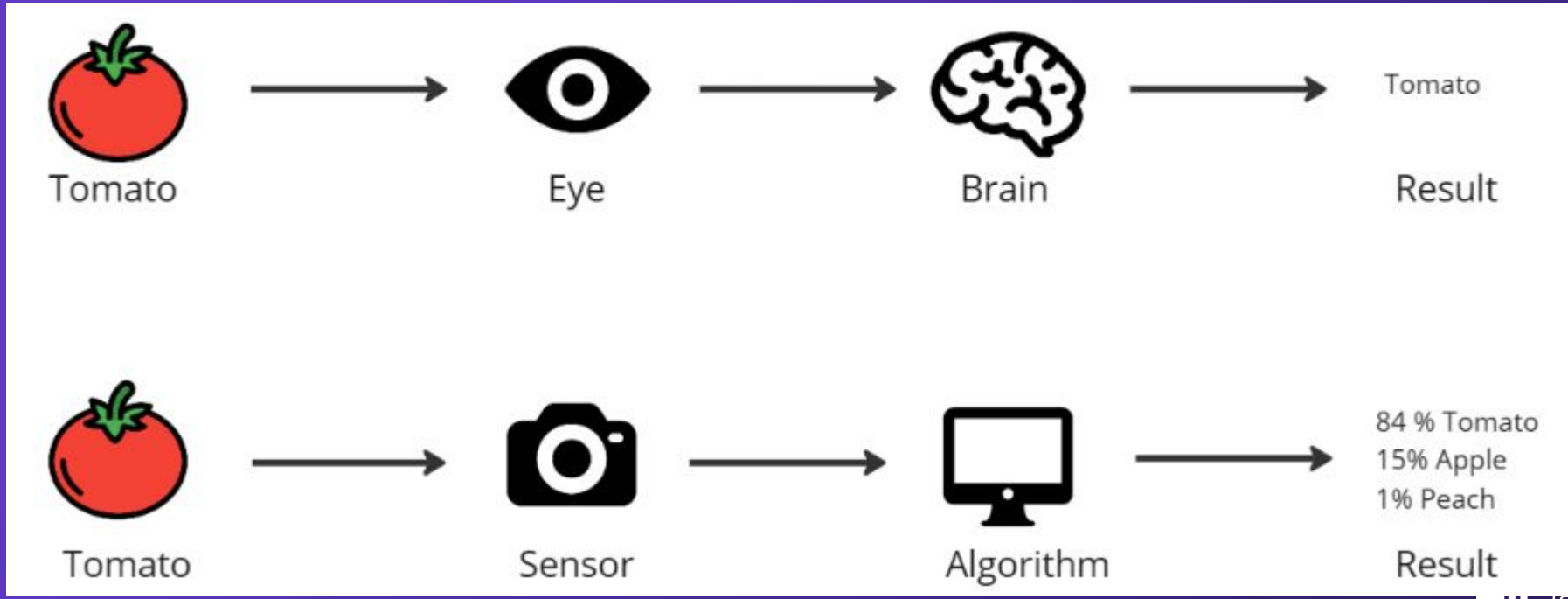


Adapted from viso.ai

TIL-AI
TODAY I LEARNED AI

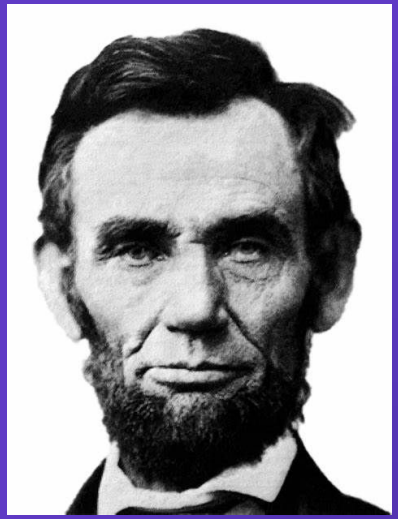
Analog Comparison

Human Vision vs Computer Vision



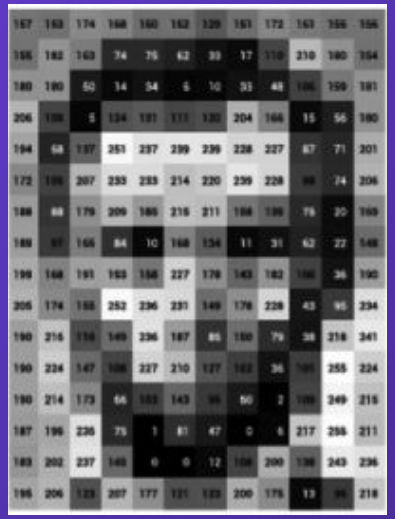
What Computers See - Pixel Values

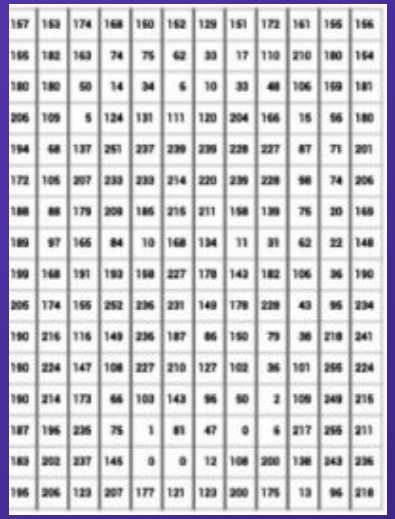
What we see



Grayscale
(Black and White)

What computers see





Array of Pixel values

Colored Images could be coded in RGB (Red, Green, Blue), HSL (Hue, Saturation, Lightness), etc, which would be channels (depth) of pixel value arrays.

TIL-AI
TODAY I LEARNED-AI

2

Image Processing with OpenCV

Hands On

TIL-AI
TODAY I LEARNED AI

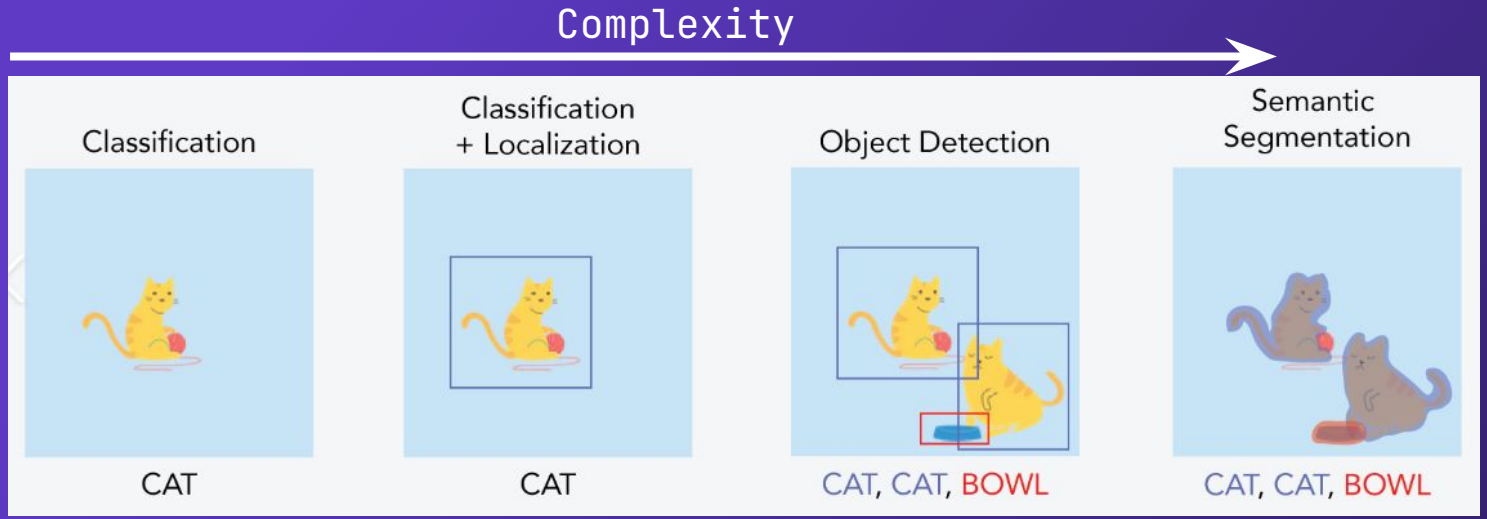
Notebook

3

Using Object Detection Models

Hands On

Common Problem Types in Computer Vision



Features	Classification	Bounding Box	Multiple objects	Segment Mask
Applications	Image Tagging, Scene Understanding, Content moderation		Robot navigation, Facial Recognition Self Driving Cars	Medical Image Diagnosis Aerial Image processing

TIL-AI
TODAY I LEARNED-AI

What is Object Detection?

Localizing and identifying objects within an image.

Tasks:

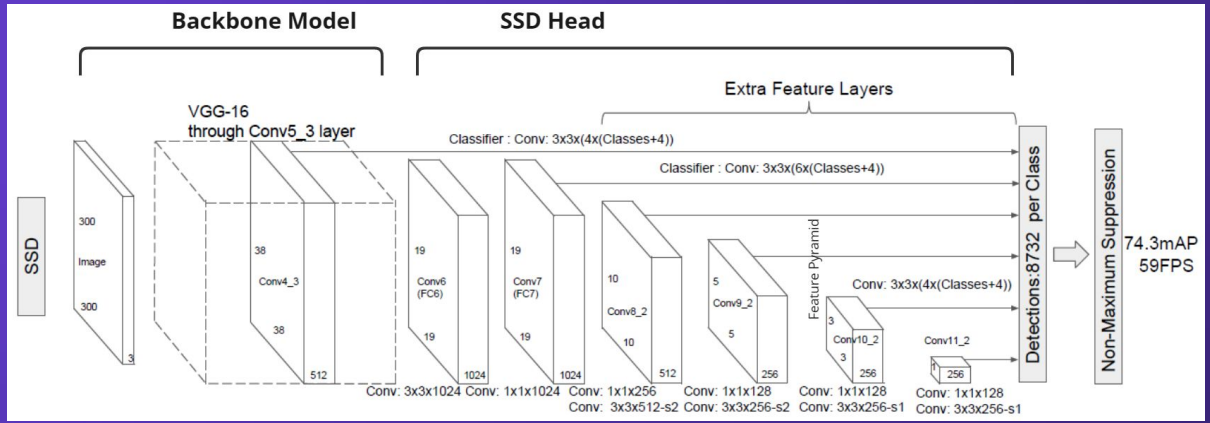
Where	**Localization**: Pinpointing the bounding box (rectangular frame) around the detected object.
What	**Classification**: Assigning a class label to the object (e.g. bus, person, dog).
Score	**Confidence Score**: Indicating the model's certainty in the detection and classification.

Single-Shot Detector (SSD)

SSD is just one technique; explore the CV/VLM educational content for more!

Two Components

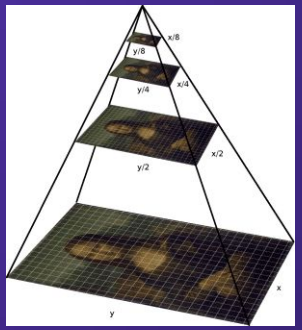
1. **Backbone model**: Pre-trained image CNN (Resnet, VGG)
2. **SSD Head**: More convolutional layers added to the backbone, whereby the outputs are bounding boxes and classes of the objects in the spatial locations.



Wei Liu et al. in the paper SSD: Single Shot MultiBox Detector.

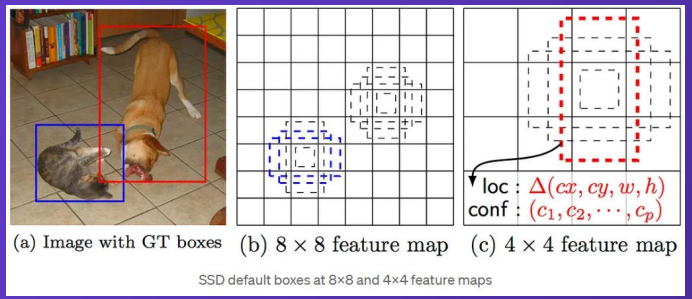
Single-Shot Detector (SSD)
Innovations

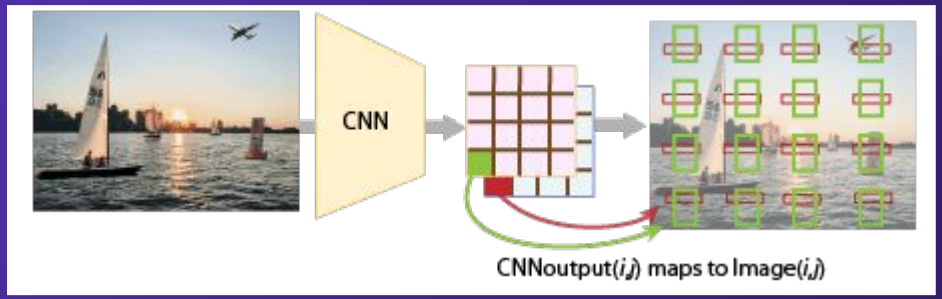
- Applies various feature map grid cell sizes (e.g. 8x8, 6x6, ..., 1x1) to detect objects of different sizes [image pyramid]



- Anchor Boxes

This is all done in the SSD Head network!



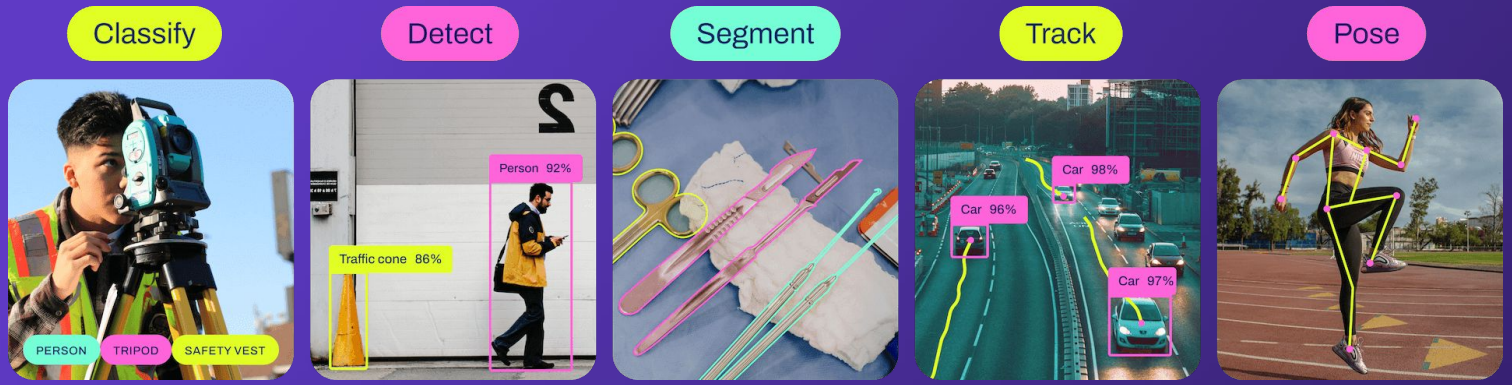


YOLO v8 Architecture

This is another popular Object Detection algorithm.

YOLOv8 has also integrated other submodules

- Classify models pretrained on the ImageNet dataset.
- Detect, Segment and Pose models pretrained on the COCO dataset (with track mode).



- Latest version: YOLO v8
- Adapted from Brief summary of YOLOv8 model structure · GitHub

Notebook

4

Introduction to Multimodal Models

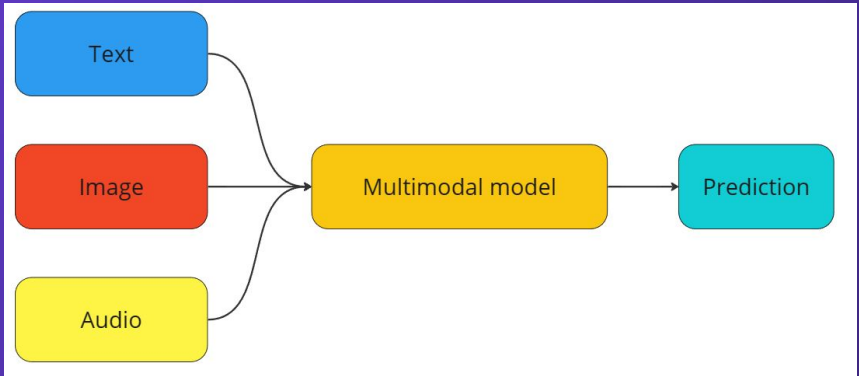
What are multimodal models?

Multimodal models

Definition

Multimodal models are models that integrates information from multiple modalities (e.g. text, image, videos, audio, gestures) to create a unified representation.

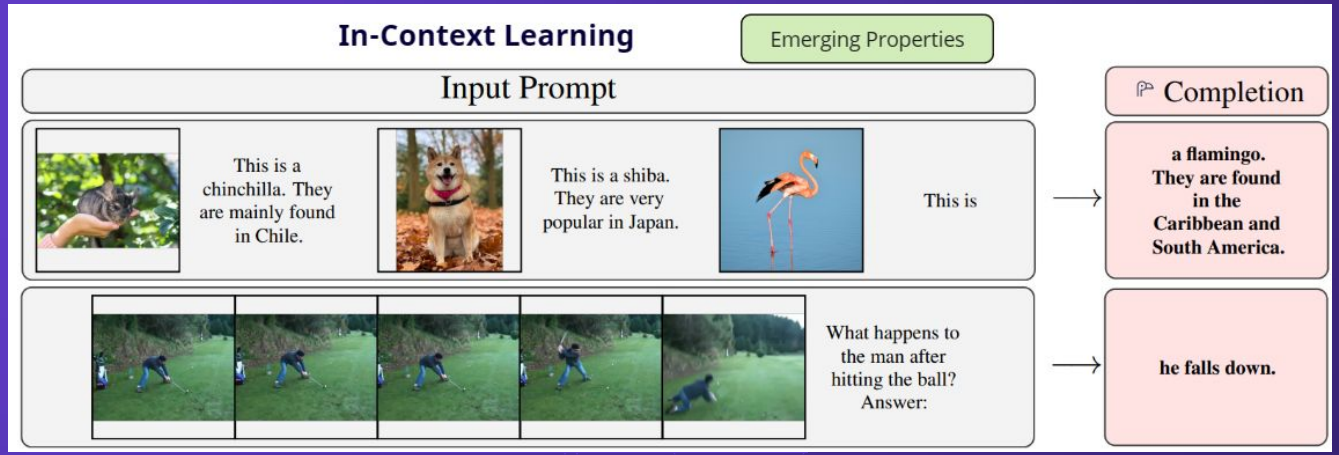
By leveraging different types of data, multimodal models can perform tasks that involve complex interactions between modalities and thus understand and reason about the world in more comprehensive way.



Multimodal models and their importance

The key importance of multimodal models in computer vision are:

1. Richer Understanding
2. Better Performance
3. Facilitate Emerging Properties



(left) The emerging properties of pre-training on web-scale interleaved image-text data: multimodal in-context-learning.VLM (Flamingo)

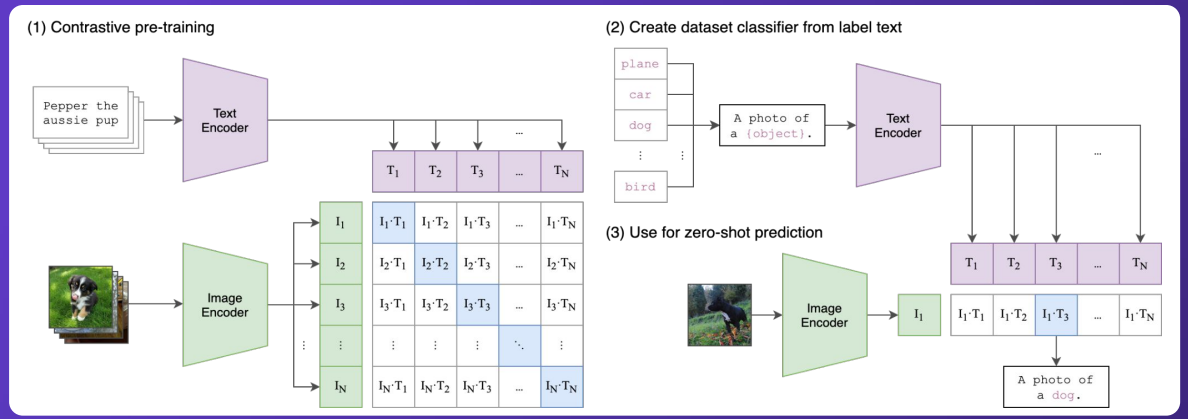
Adapted from Alayrac et al. (2022) via

CLIP (*Contrastive Language-Image Pre-training*)

Connecting Text and Image

"CLIP is a neural network trained on a variety of (image, text) pairs. It can be instructed in natural language to predict the most relevant text snippet, given an image, without directly optimizing for the task, similar to the zero-shot capabilities of GPT-2 and 3. We found CLIP matches the performance of the original ResNet50 on ImageNet "zero-shot" without using any of the original 1.28M labeled examples, overcoming several major challenges in computer vision."

Approach



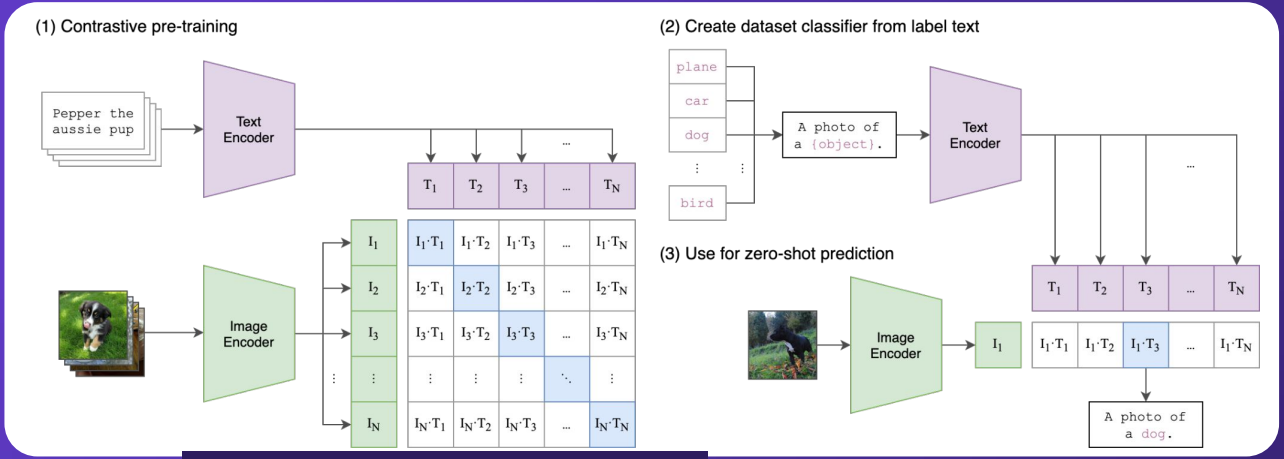
- openai/CLIP: CLIP (Contrastive Language-Image Pretraining), Predict the most relevant text snippet given an image (github.com)
- CLIP: Connecting text and images (openai.com)

CLIP (*Contrastive Language-Image Pre-training*)

Details

Given a batch of N (image, text) pairs, predict which of the N × N possible (image, text) pairings across a batch actually occurred.

(Image, Text) Pairs



Author constructed a new dataset of 400 million (image, text) pairs collected from a variety of publicly available sources on the Internet

Multi-model embedding space
Maximize the cosine similarity of image vs text

Optimized for a symmetric cross entropy loss.

- openai/CLIP: CLIP (Contrastive Language-Image Pretraining), github.com
- CLIP: Connecting text and images (openai.com)

Notebook

5

Deploying pretrained VLM to FastAPI

Hands On

Notebook