Advanced CV/VLM Workshop

Beyond the Basics: Advanced CV — CNN and VLM Techniques

Ming Hao

Data Scientist @ FWD

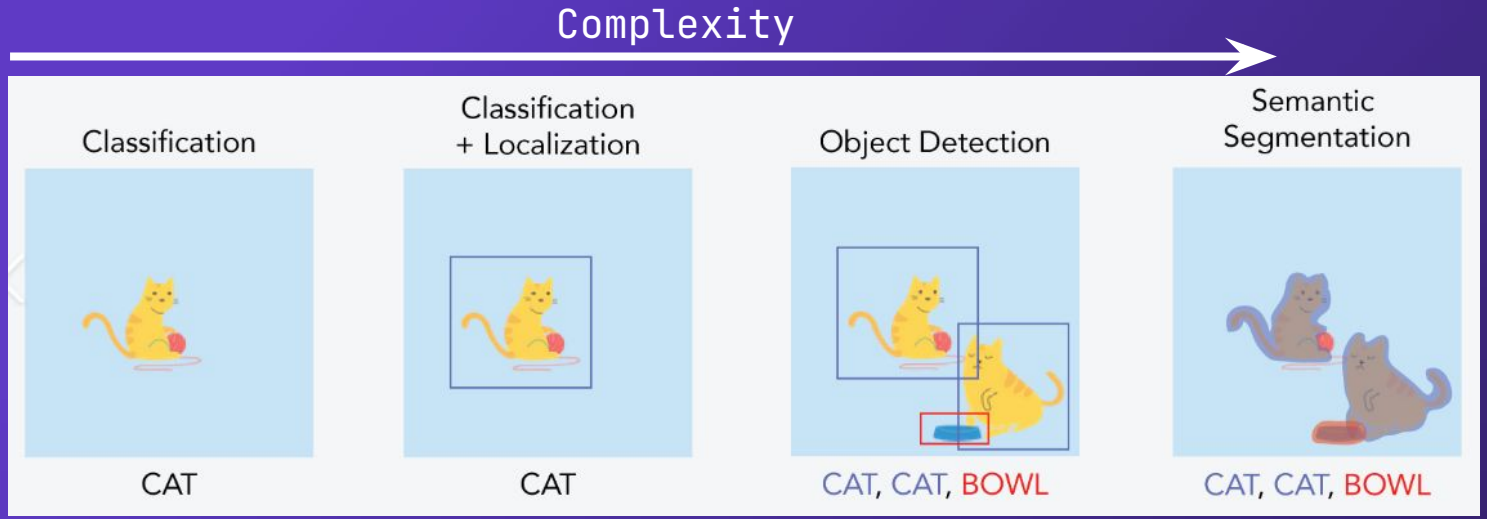
Table of Contents

1

Deep Dive into Object Detection Architectures

TIL-AI
TODAY I LEARNED AI

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

OD Algorithm Examples

There are many object detection algorithms available. We will discuss:

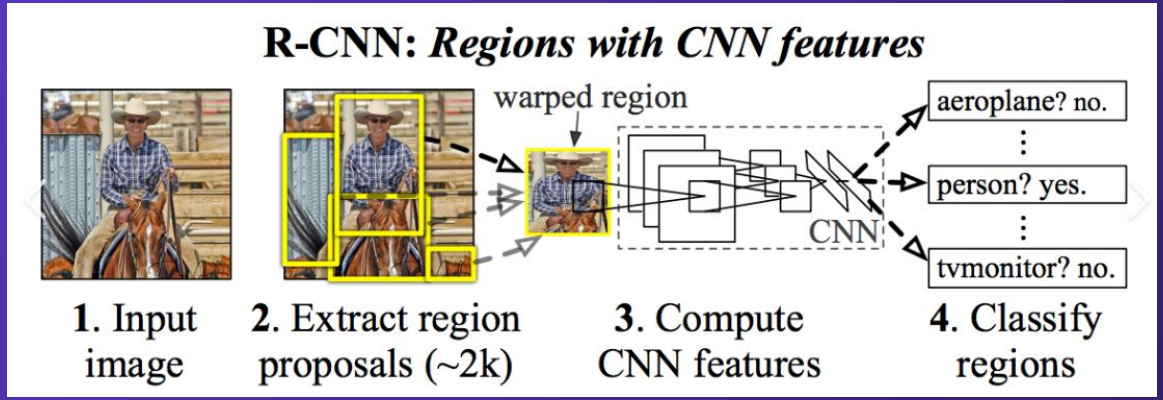
- Region-CNN Family (R-CNN)

- Single-Shot Detectors (SSD)

Intro to Region-CNN (R-CNN)

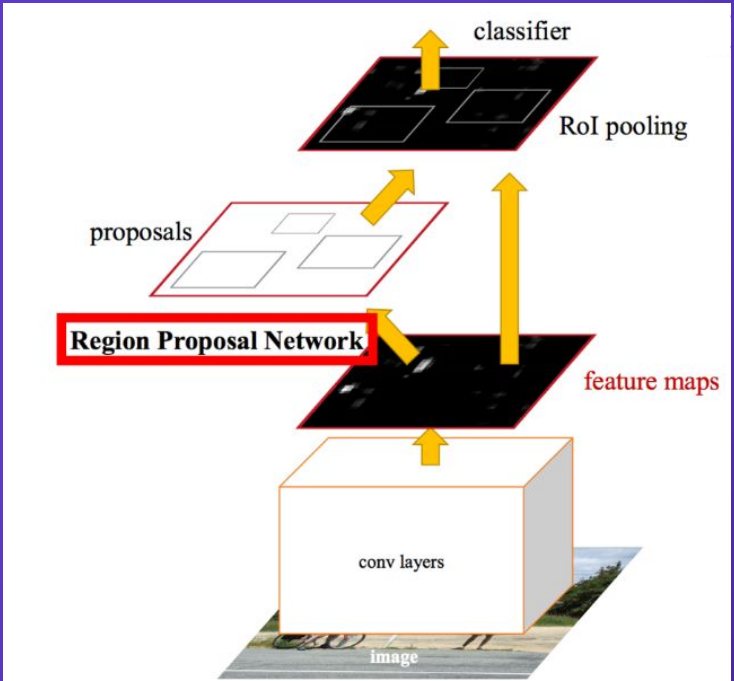
The process of R-CNN detection is as follows:

1. Generate **bottom-up** region proposals (selective search)
2. **Warping & Feature Extraction:** Extract features for each proposal using CNNs
3. **Classification:** SVM to segregating the objects into classes



A Survey on Deep Learning Based Approaches for Scene Understanding in Autonomous Driving Zhiyang Guo- 2021/4

Faster R-CNN - Technique Overview



Steps

1. CNN to extract feature maps.

2. **RPN (RoI Pooling)**: Determine regions of interest from the feature maps with a seperate CNN.

3. Fully connected layers

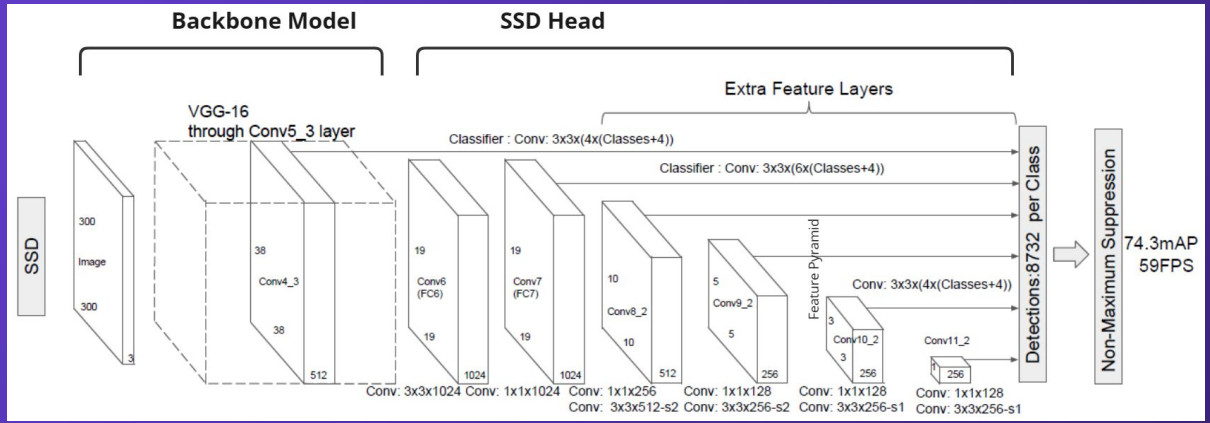
4. Classification with softmax and bounding box regression

TIL-AI
TODAY I LEARNED AI

Single-Shot Detector (SSD)

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.

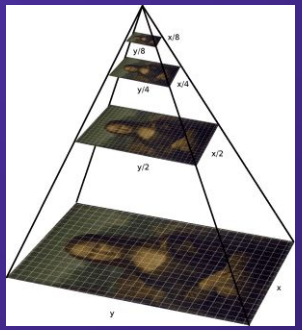


Wei Liu et al. in the paper SSD: Single Shot MultiBox Detector.

Single-Shot Detector (SSD)

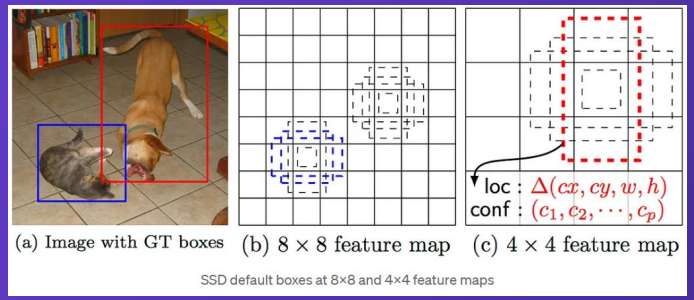
Innovations

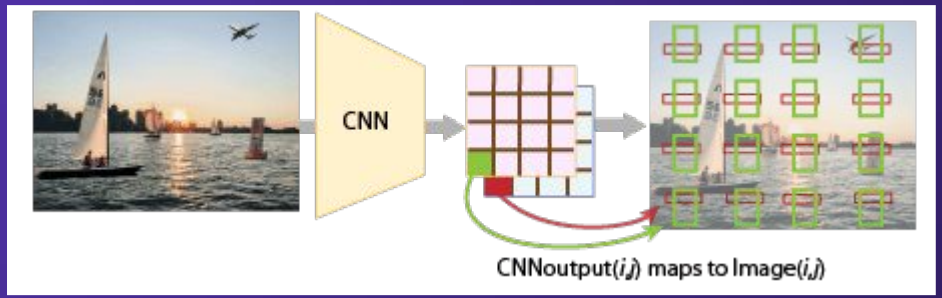
- Applies various feature map grid cell sizes (e.g. 8x8, 6x6, ..., 1x1) to detect objects of different sizes [image pyramid]



- Anchor Boxes

This is all done in the SSD Head network!





2

Object Detection Fine Tuning

Notebook

3

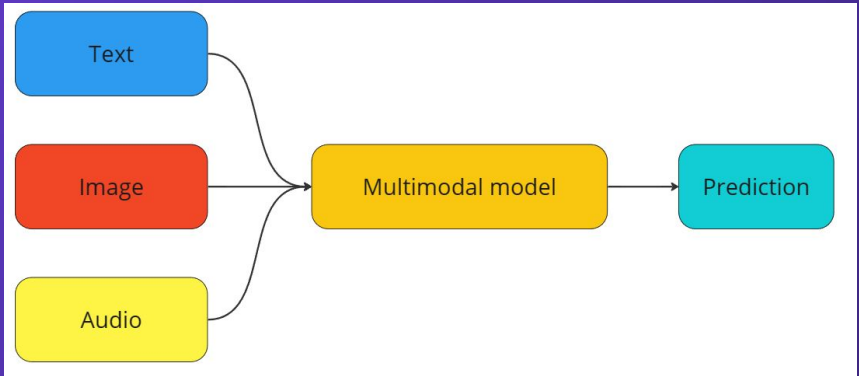
Introduction to Multimodal Models

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.

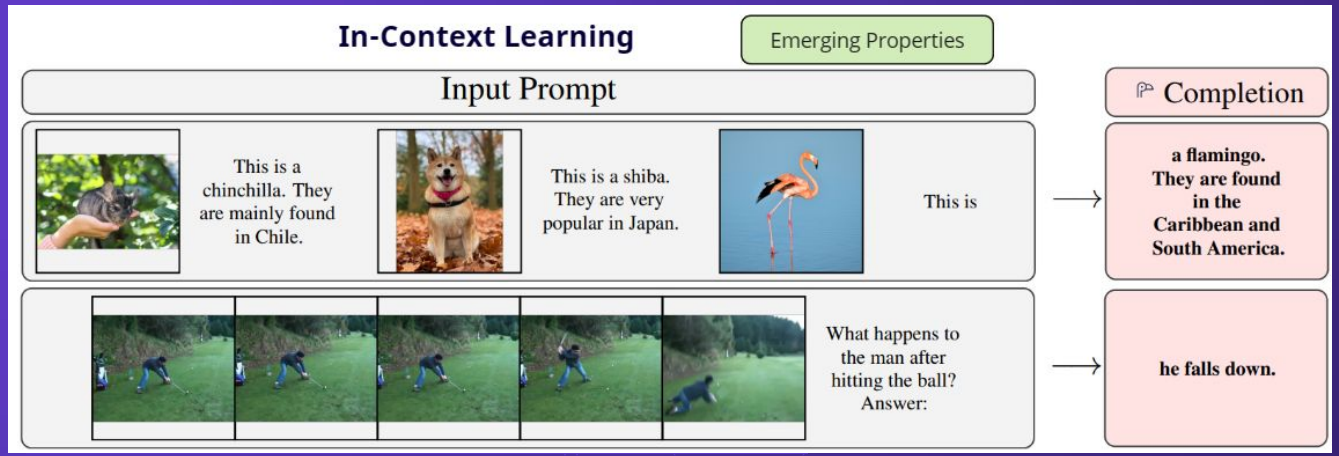
By leveraging different types of data, multimodal models can perform tasks that involve complex interactions between modalities and thus understand and reason about the world in more comprehensive way.



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)

Popular Vision Language Models

There are many Vision-Language models available. We will discuss:

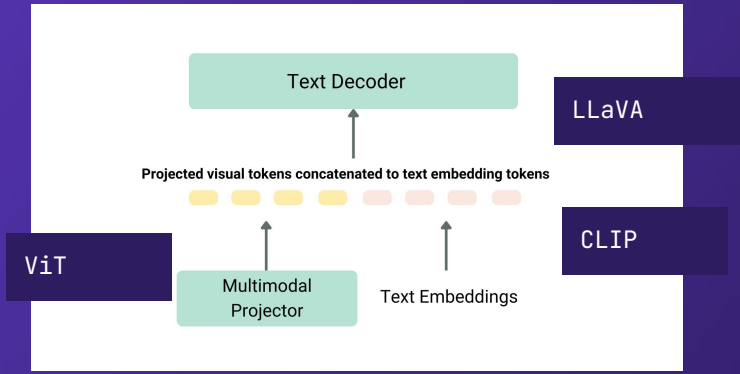
- **CLIP** (*Contrastive Language-Image Pre-training*)
  Image/Text Embedding

- **OWL-ViT** (Vision Transformer for Open-World Localization)
  Image/Text Embedding + Detection

- **LLaVA (Large Language-and-Vision Assistant)**
  Instruct Multi-Modal Models

Which are usually composed of:

- **ViT** (Ie, Vision Transformer)
- **LLM** (Ie, Vicuna)



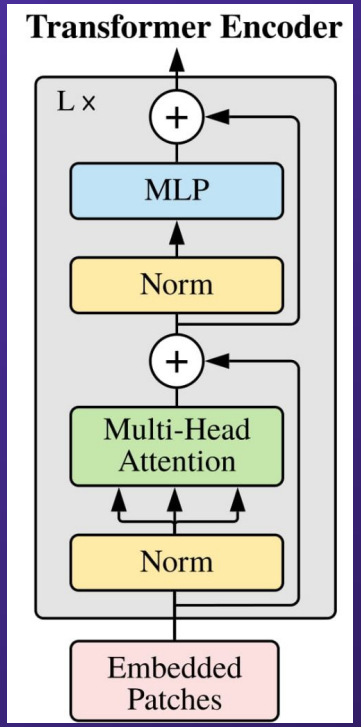
(top) High level diagram on Vision Language models. It consists of 3 major components:

1. A multimodal projector (ViT)
2. Combining embedding text+image pairs (CLIP)
3. A text decoder to facilitate reasoning with instruct (LLaVa)

Vision Transformers (ViT)

Overview

- Transformer architecture is the de-facto standard for NLP tasks and is also applicable to computer vision.
- Unlike CNNs, it does not have a locally restricted receptive field (kernels); instead, it converts everything to sequences and processes it all together.
- It can match or surpass state-of-the-art CNNs when trained on datasets larger than 14 million samples. However, for smaller datasets, ResNets or EfficientNets are more effective.



- Vision Transformer (ViT) (huggingface.co)
- Vision Transformer Explained | Papers With Code
- How the Vision Transformer (ViT) works in 10 minutes: an image is worth 16x16 words | AI Summer (theaisummer.com)

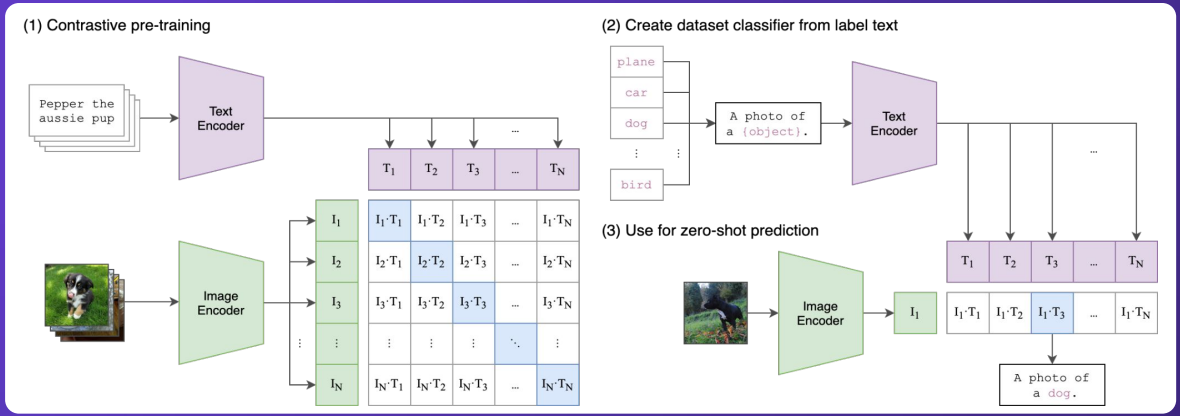
TIL-AI
TODAY I LEARNED-AI

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*)

Comparisons

CLIP generalized better than predictive methods!

Why not CNN + Text Transformers?

The author's initial approach, similar to VirTex, jointly trained an image CNN and text transformer from scratch to predict the caption of an image.

However, the author encountered difficulties with efficiently scaling this method with image/text pairs.

Why a contrastive instead of predictive approach?

The author found that contrastive objectives for images can learn better than their equivalent predictive/similarity objective.

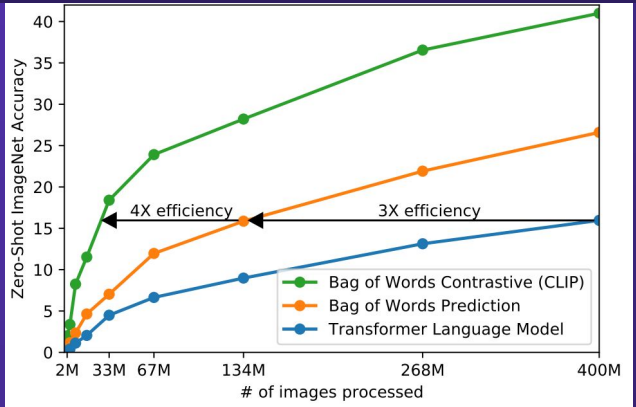


Figure 2. **CLIP is much more efficient at zero-shot transfer than our image caption baseline.** Although highly expressive, we found that transformer-based language models are relatively weak at zero-shot ImageNet classification. Here, we see that it learns 3x slower than a baseline which predicts a bag-of-words (BoW) encoding of the text (Joulin et al., 2016). Swapping the prediction objective for the contrastive objective of CLIP further improves efficiency another 4x.

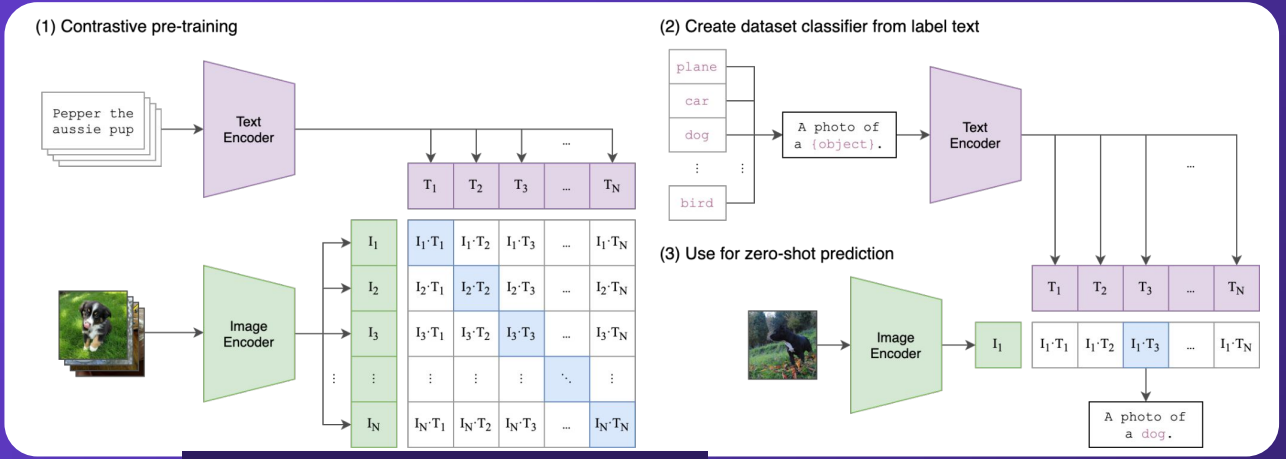
- openai/CLIP: CLIP (Contrastive Language-Image Pretraining), github.com
- CLIP: Connecting text and images (openai.com)
- Learning Transferable Visual Models From Natural Language Supervision

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

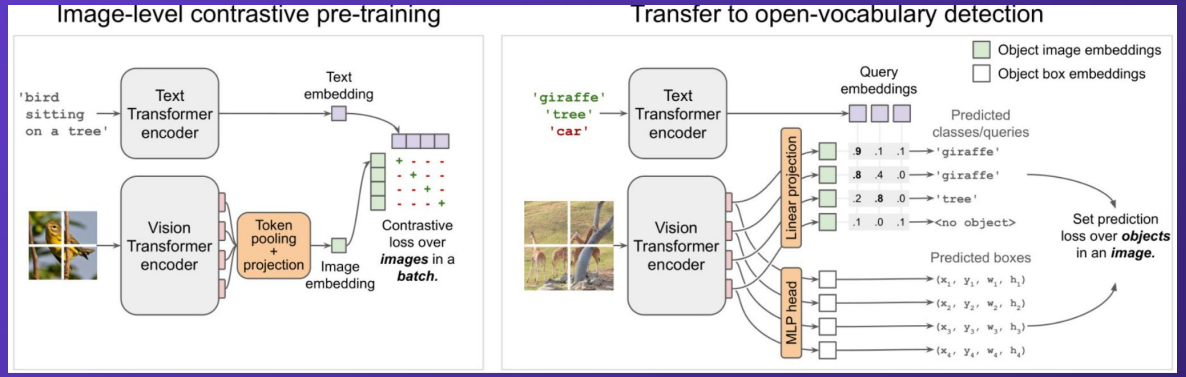
Multi-model embedding space
Maximize the cosine similarity of image vs text

Optimized for a symmetric cross entropy loss.

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

OWL-ViT

OWL-ViT (short for Vision Transformer for Open-World Localization) is an open-vocabulary object detection network trained on a variety of (image, text) pairs.



OWL-ViT's Unique Features:

- **CLIP + Detection:** OWL-ViT builds upon the success of CLIP (Contrastive Language-Image Pretraining). CLIP learns to match text descriptions with corresponding images. OWL-ViT takes this a step further: It removes the final token pooling layer and attaches a lightweight classification and box head to each transformer output token.
- **Fine-tuned End-to-End:** Fine-tuned end-to-end on standard detection datasets using DETR-style bipartite matching loss.

The magic of OWL-ViT lies in its "open vocabulary" capability. OWL-ViT can potentially detect objects described in text, even if those objects weren't explicitly included in its training data.

- Paper:Simple Open-Vocabulary Object Detection with Vision Transformers
- Hugging Face:OWL-ViT (huggingface.co)

4

Vision Language Models

Notebook

5

VLM Fine-tuning

Notebook

6

Deploying VLM using FastAPI

Notebook