CV / VLMs

Unit 4: Vision-Language Models (VLMs)

4.2.1

Pre-trained Vision Language Models

Popular vision language models

Vision Language Models Examples

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

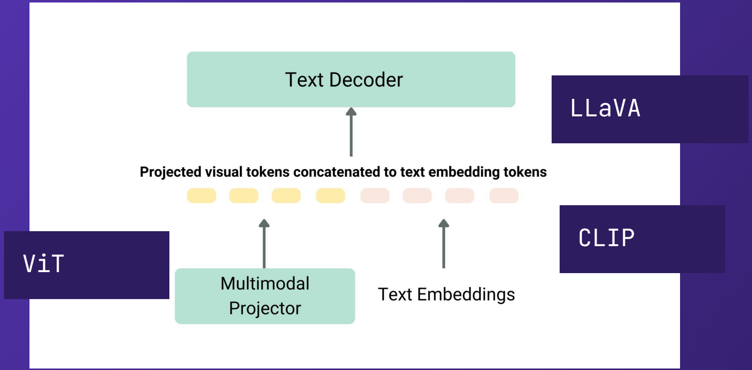
- **CLIP** (*Contrastive Language-Image Pre-training*)
  Image/Text Embedding

- **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)

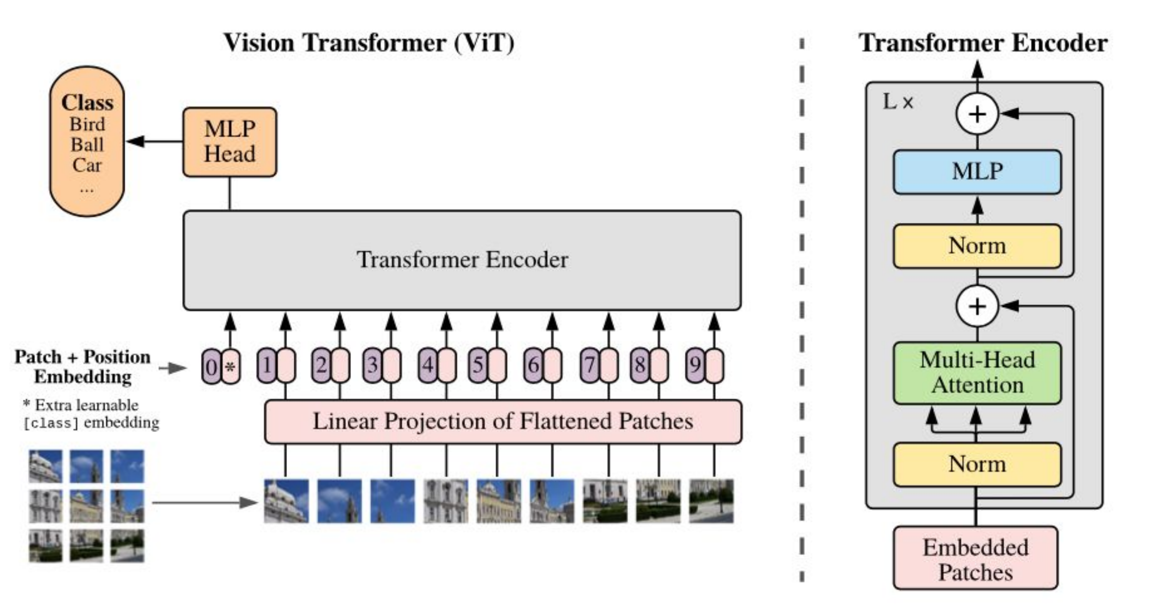


(top) High level diagram on Vision Language models. It consists of 3 major components:

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

ViT (*Vision Transformer*)
Encoding Image with Transformers

To be explored in more depth in Unit 6!



Overview of the model:
i) We split an image into fixed-size patches, linearly embed each of them and add position embeddings.

ii) The resulting sequence is fed to a standard Transformer encoder.

iii) In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence.

Adapted from [google-research/vision_transformer (github.com)]

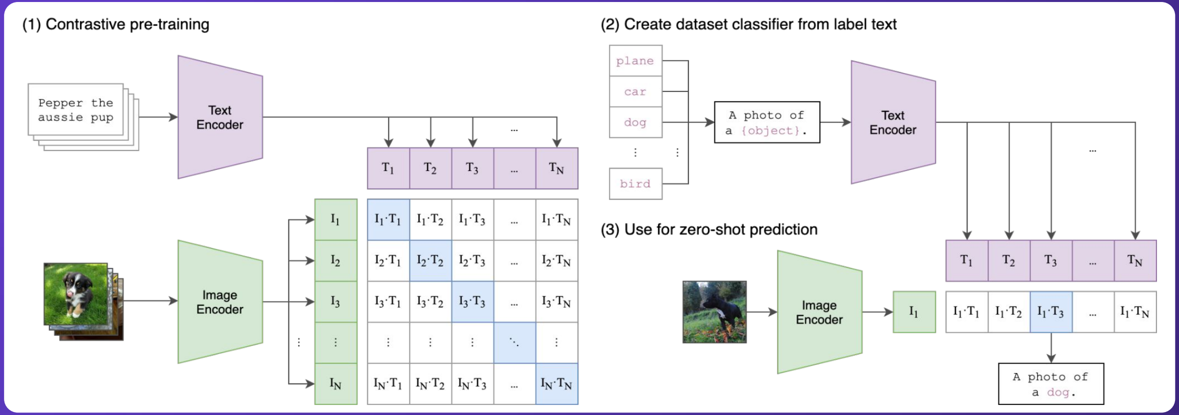
TIL-AI
TODAY I LEARNED AI

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)]

TIL-AI
TODAY I LEARNED-AI

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.

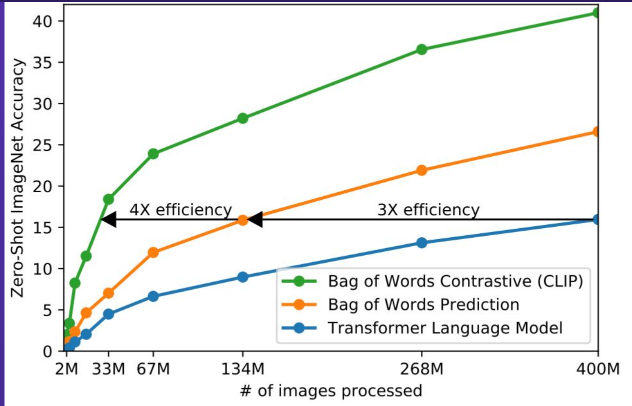


Figure 2. **CLIP is much more efficient at zero-shot transfer than our image caption baseline.** Although highly expressive, we found that transformer-based language models are relatively weak at zero-shot ImageNet classification. Here, we see that it learns 3x slower than a baseline which predicts a 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.

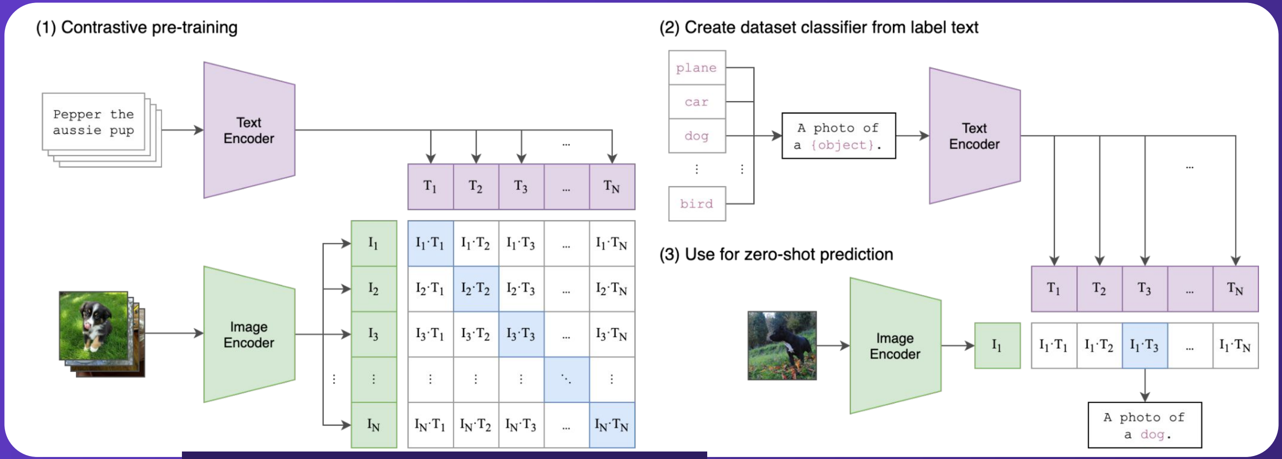
- openai/CLIP: CLIP (Contrastive Language-Image Pretraining), github.com
- CLIP: Connecting text and images (openai.com)
- Learning Transferable Visual Models From Natural Language Supervision

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

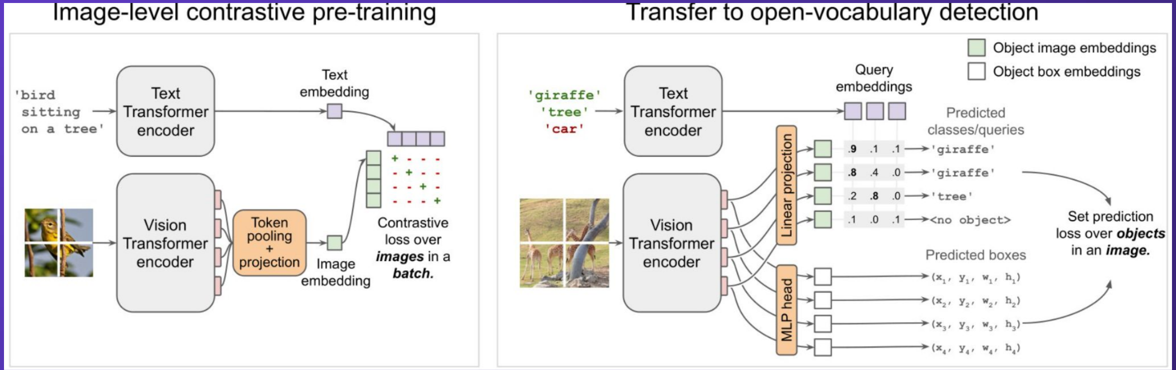
Multi-model embedding space
Maximize the cosine similarity of image vs text

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.

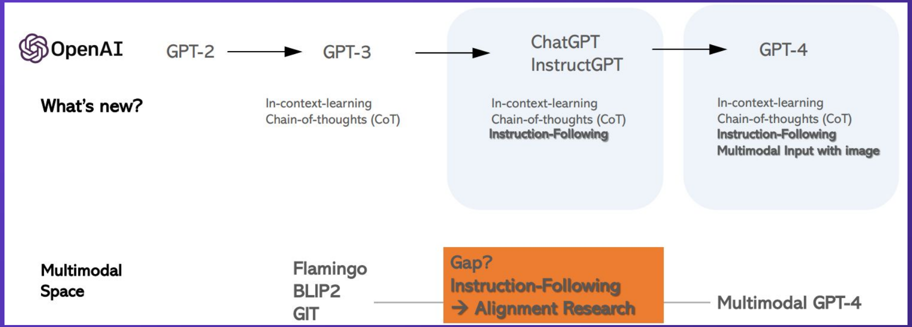
The magic of OWL-ViT lies in its "open vocabulary" capability. OWL-ViT can potentially detect objects described in text, even if those objects weren't explicitly included in its training data.

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

TIL-AI TODAY I LEARNED AI

LLaVA (Large Language-and-Vision Assistant)
Introduction

- LLaVA represents a novel end-to-end trained large multimodal model that combines a vision encoder and Vicuna for general-purpose visual and language understanding.
- Achieving impressive chat capabilities mimicking spirits of the multimodal GPT-4.
- It is an instruction-tuning Large Language Model (LLM), adapting instruction-following advancement in LLMs.



(top) Advancement of LLM capacities, with instruction-following capacity being the one of the recent breakthrough.

[LLaVA (llava-vl.github.io)](llava-vl.github.io)

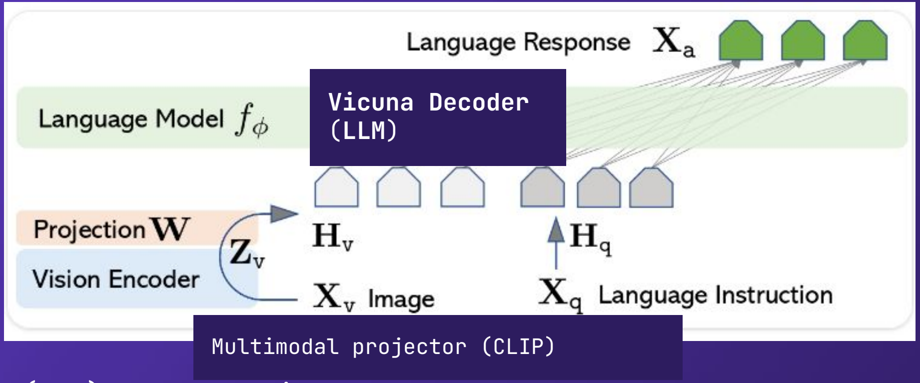
LLaVA (Large Language-and-Vision Assistant)

Unique improvement

1. **Instruction-following LLM:**
   Utilized a vision backbone to encode visual features, a LLM to comprehend the user instructions, and a vision-language cross-modal connector to align.

2. **Multimodal Instruction-following data**
   For visual instruction tuning, LLava pioneer to leverage text-only GPT-4 to expand the existing COCO bounding box and caption datasets to a multimodal instruction following datasets.

3. **Response formatting prompts**
   LLava proposed to use a single response formatting prompt that clearly indicates the output format.



(top) **LLaVa** architecture.
(bottom) Response formatting Prompt

Visual input example, Different Format Prompts:

Formatting prompt	What is the color of the shirt that the man is wearing? **Answer the question using a single word or phrase.**
Response	Yellow.

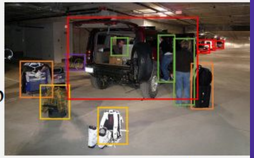
<u>Paper</u>: Improved Baselines with Visual Instruction Tuning

LLaVA (Large Language-and-Vision Assistant) Example Data

An example to illustrate the instruction-following data. The top block shows the contexts such as captions and boxes used to prompt GPT, and the bottom block shows the three types of responses.

Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.



Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation

Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

Response type 3: complex reasoning

Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>