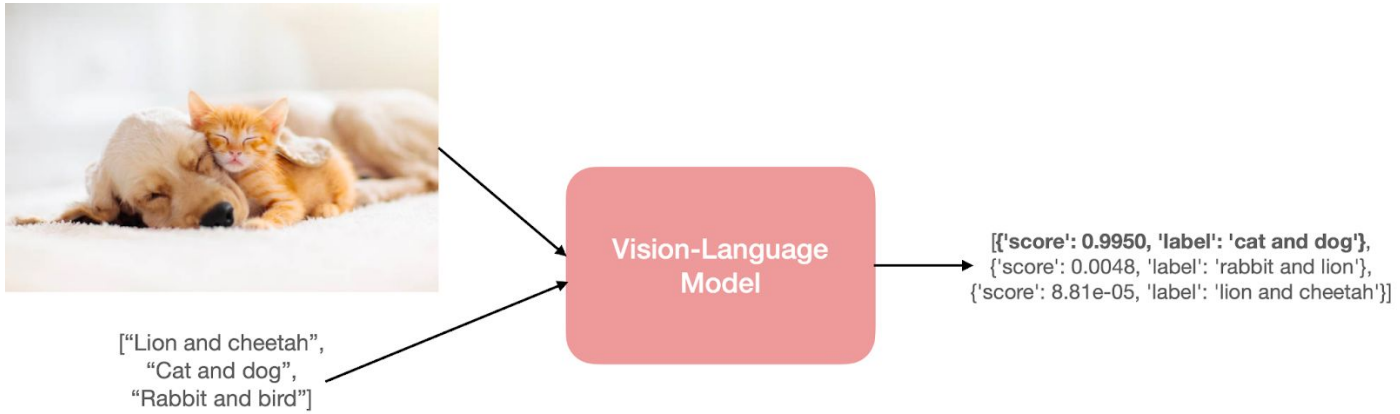
Introduction to Vision-Language Model

Date : 2024-12-10

Table of contents

Introduction

- What is Vision-Language Model?
  - In any case, if the input of a model is an image and a language, it can be called a vision-language model.
  - Very simple model can be vision-language model



https://huggingface.co/blog/vision_language_pretraining

Learning Strategies

- Contrastive Learning
  - Aligning images and texts to a joint feature space in contrastive manner
- PrefixLM
  - Jointly learning image and text embeddings by using images as a prefix to a language model
- Multi-modal Fusing with Cross Attention
  - Fusing visual information into layers of a language model with a cross-attention mechanism
- MLM / ITM
  - Aligning parts of images with text with masked-language modeling and image-text matching objectives

Learning Strategies

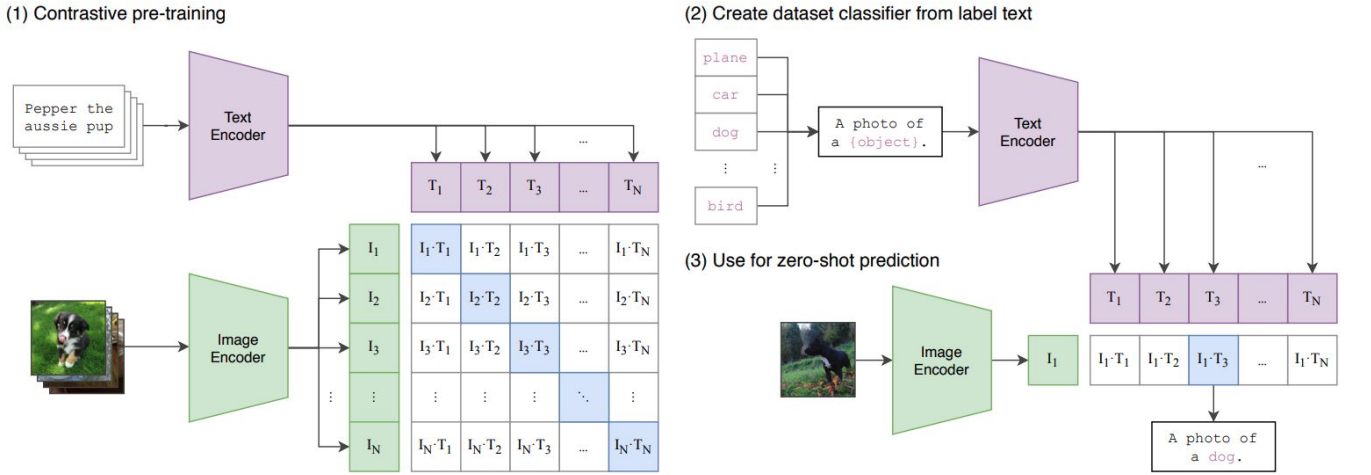
- Contrastive Learning
  - used as pre-training objective for vision encoders and text encoders
    - Mainly reusing vision encoders
  - Representative researches
    - CLIP: Learning Transferable Visual Models From Natural Language Supervision
    - ALIGN: Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision
    - SigLIP: Sigmoid Loss for Language Image Pre-Training

Learning Strategies

- Contrastive Learning
  - CLIP: Learning Transferable Visual Models From Natural Language Supervision
    - Dataset : WebImageText(WIT)
      - 400M image-text pairs
      - Use 50,000 queries to get the images
      - For class balance, each query did not exceed 20,000 image-text pairs.

Learning Strategies

- Contrastive Learning
  - CLIP: Learning Transferable Visual Models From Natural Language Supervision
    - Pretraining
      - Contrastive Loss (= InfoNCE) $L_{NCE} = -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)}{\sum_{k=0}^{N} \exp(\mathbf{z}_i \cdot \mathbf{z}_k / \tau)}$



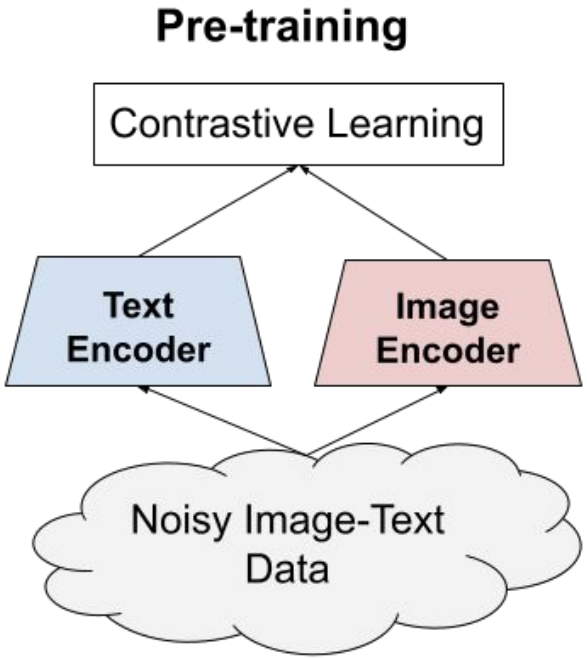

```python
# image_encoder - ResNet or Vision Transformer
# text_encoder - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l] - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t - learned temperature parameter
# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T) #[n, d_t]
# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)
# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)
# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss = (loss_i + loss_t)/2
```

Learning Strategies

- Contrastive Learning
  - ALIGN: Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision
    - Dataset : Less Clean but Large-scale
      - 1.8B image-text pairs
      - Image-based filtering
        - remove
          - pornographic images
        - keep
          - images whose shorter dimension is larger than 200 pixels
          - aspect ratio is smaller than 3
      - Text-based filtering
        - remove
          - alt-texts that are shared by more than 10 images
            e.g. “1920x1090” or “alt_img”
          - outside of 100 million most frequent unigrams and bigrams from the raw dataset
          - too short(<3 unigrams) or too long(>20 unigrams)

Learning Strategies

- Contrastive Learning
  - ALIGN: Scaling Up Visual and Vision-Language Representation Learning
    With Noisy Text Supervision
    - Pretraining
      - Contrastive Loss (= InfoNCE)
      - Two stage Training
        - First stage : 346 x 346
        - Second stage : 289 x 289



Learning Strategies

- Contrastive Learning
  - SigLIP: Sigmoid Loss for Language Image Pre-Training
    - Dataset : WebLI(Web Language Image)
      - multilingual image-text dataset designed to support google’s vision language research

Learning Strategies

- Contrastive Learning
  - SigLIP: Sigmoid Loss for Language Image Pre-Training
    - Pretraining
      - Contrastive Learning using Sigmoid Loss
      - Multi class classification -> binary classification

```
# img_emb : image model embedding (n, dim)
# text_emb : text model embedding (n, dim)
# t_prime, b : learable temperature and bias
# n : mini-batch size

t = exp(t_primt)
zimg = l2_normalize(img_emb)
ztxt = l2_normalize(txt_emb)
logits = dot(zimg, ztxt.T) * t + b
labels = 2 * eye(n) - ones(n) # -1 with diagonal 1
l = -sum(log_sigmoid(labels * logits )) / n
```

$$-\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_i\cdot\mathbf{y}_j}}}^{\text{image}\to\text{text softmax}}+\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_j\cdot\mathbf{y}_i}}}^{\text{text}\to\text{image softmax}}\right)$$

Contrastive Learning

$$-\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}}}_{\mathcal{L}_{ij}}$$

Contrastive Learning with sigmoid

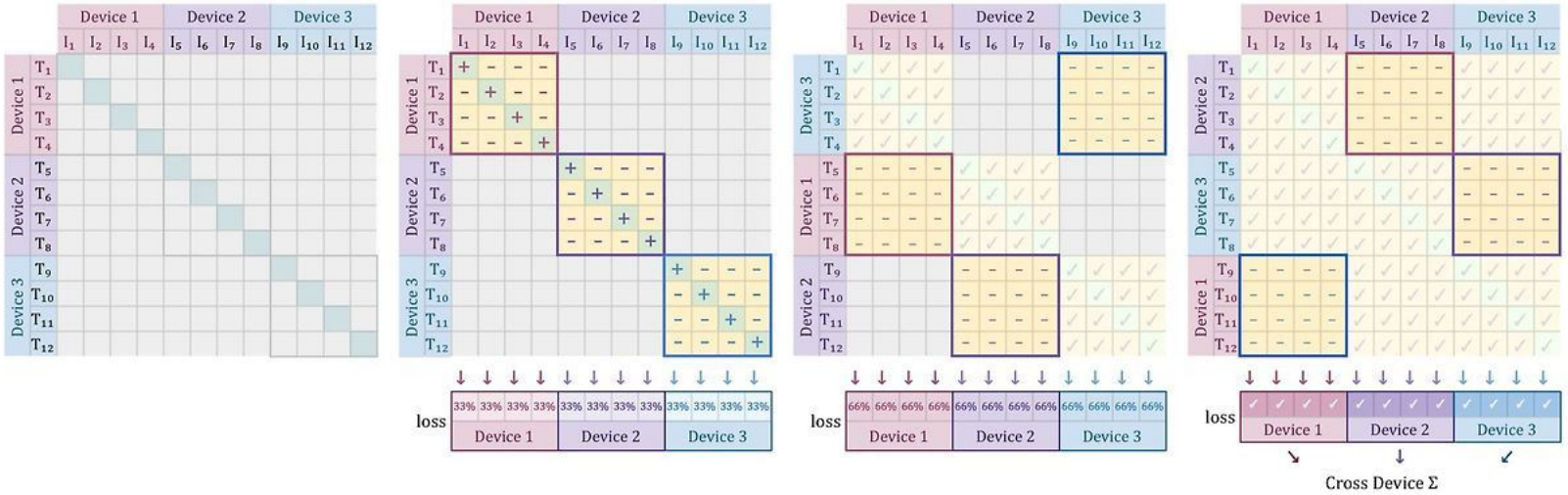
Learning Strategies

$$-\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_i\cdot\mathbf{y}_j}}}^{\text{image}\to\text{text softmax}}+\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_j\cdot\mathbf{y}_i}}}^{\text{text}\to\text{image softmax}}\right)$$

Contrastive Learning

$$-\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}}}_{\mathcal{L}_{ij}}$$

Contrastive Learning with sigmoid



Learning Strategies

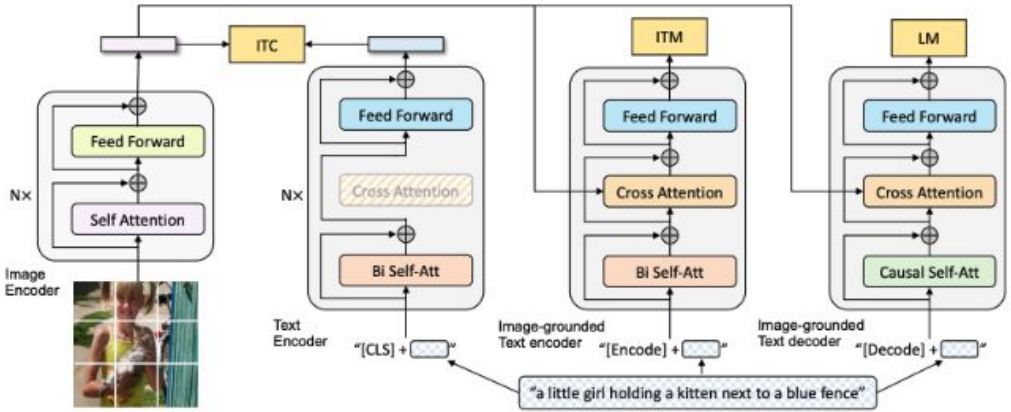
- MLM / ITM (Multi Objective Learning)
  - Masked Language Modeling / Image Text Matching
  - Same with BERT
    - Next Sentence Prediction -> Image Text Matching
  - Representative researches
    - BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation
    - BLIP2: Bootstrapping Language-Image Pretraining with Frozen Image Encoders and Large Language Models

Learning Strategies

- MLM / ITM (Multi Objective Learning)
  - BLIP: Bootstrapping Language-Image Pre-training
    for Unified Vision-Language Understanding and Generation
    - ITC(Image Text Contrastive Learning)
      - Vision Encoder, Text Encoder
    - ITM(Imgae Text Matching)
      - Encoder(Bidirectional) Decoder(Bidirectional)
    - LM(Language Modeling)
      - Encoder(Bidirectional) Decoder(Unidirectional)

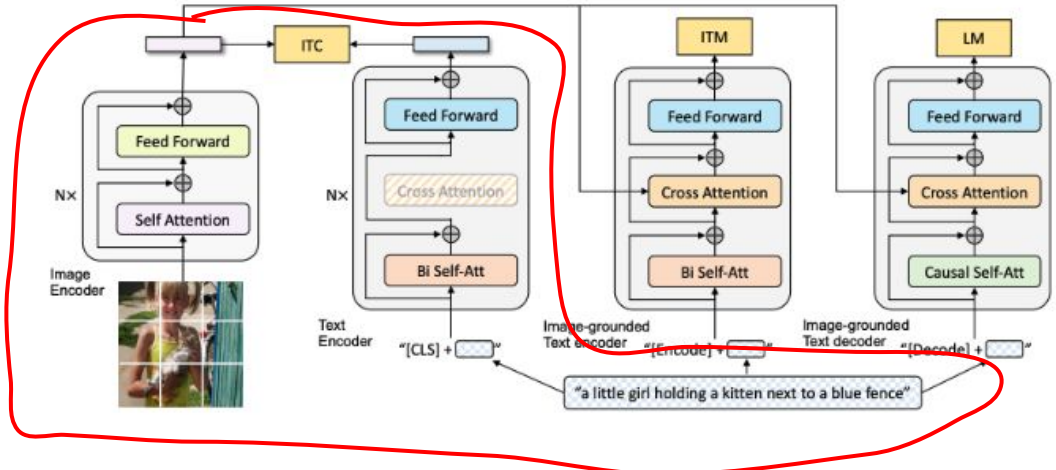
Learning Strategies

- MLM / ITM (Multi Objective Learning)
  - BLIP: Bootstrapping Language-Image Pre-training
    for Unified Vision-Language Understanding and Generation



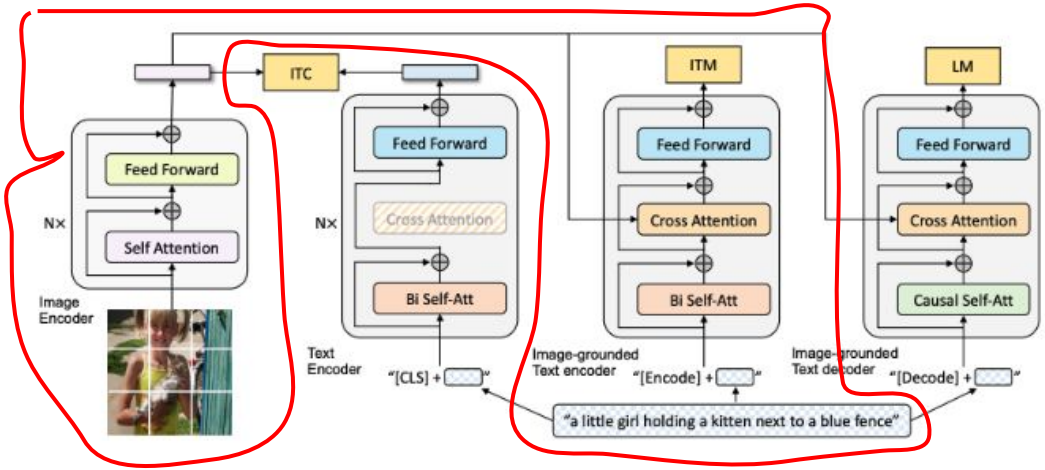
Learning Strategies

- MLM / ITM (Multi Objective Learning)
  - BLIP: Bootstrapping Language-Image Pre-training
    for Unified Vision-Language Understanding and Generation
    - ITC(Image Text Contrastive Learning)
      - Vision Encoder, Text Encoder



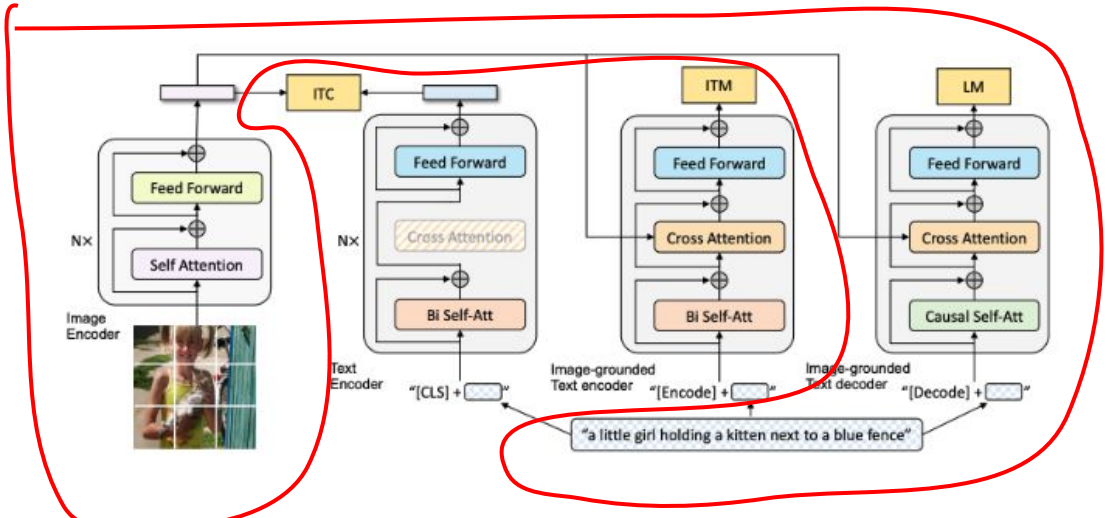
Learning Strategies

- MLM / ITM (Multi Objective Learning)
  - BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation
    - ITM(Imgae Text Matching)
      - Encoder(Bidirectional) Decoder(Bidirectional)



Learning Strategies

- MLM / ITM (Multi Objective Learning)
  - BLIP: Bootstrapping Language-Image Pre-training
    for Unified Vision-Language Understanding and Generation
    - LM(Language Modeling)
      - Encoder(Bidirectional) Decoder(Unidirectional)

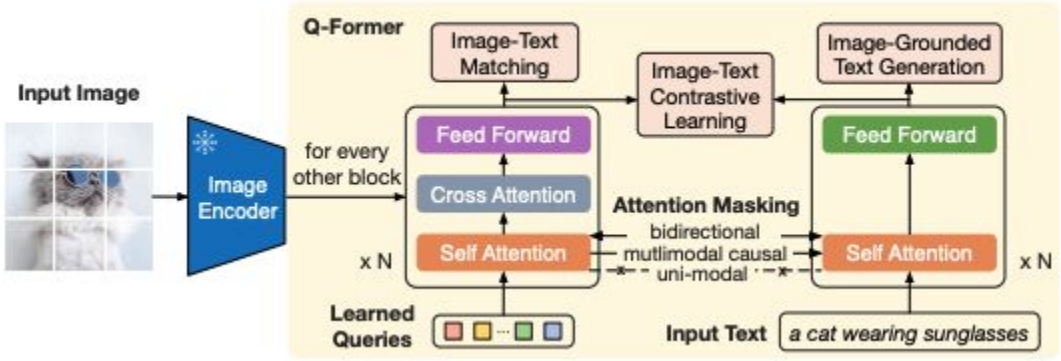


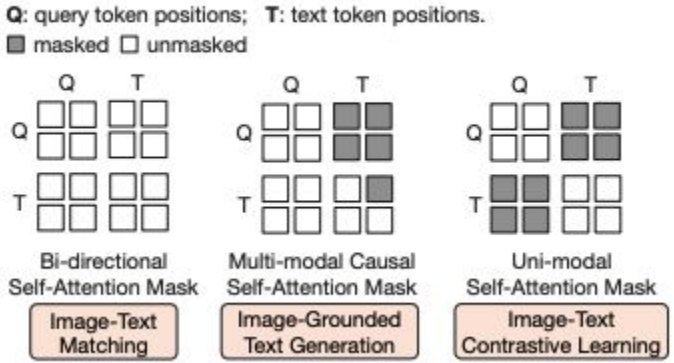
Learning Strategies

- MLM / ITM (Multi Objective Learning)
  - BLIP2: Bootstrapping Language-Image Pretraining with Frozen Image Encoders and Large Language Models

Learning Strategies

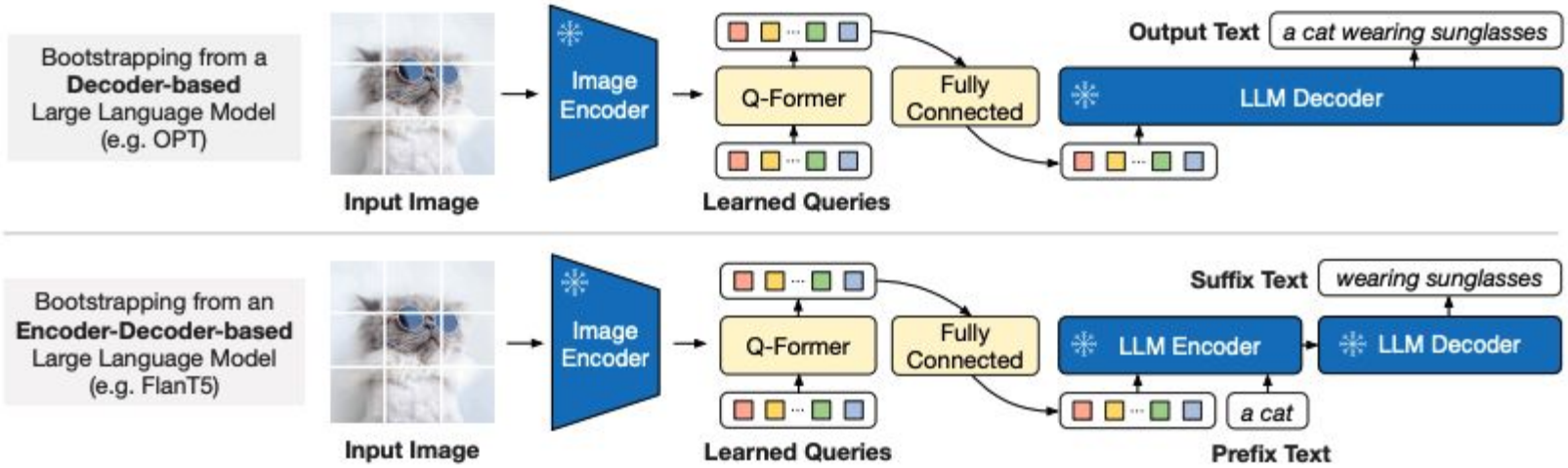
- MLM / ITM (Multi Objective Learning)
  - BLIP2: Bootstrapping Language-Image Pretraining with Frozen Image Encoders and Large Language Models
    - Language Image Pretraining
    - Frozen Image Encoders





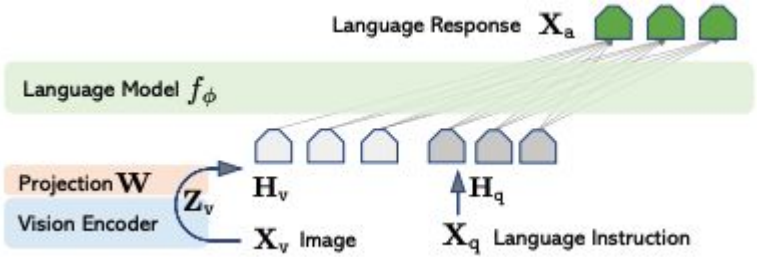
Learning Strategies

- MLM / ITM (Multi Objective Learning)
  - BLIP2: Bootstrapping Language-Image Pretraining with Frozen Image Encoders and Large Language Models
    - Large Language Models



Learning Strategies

- PrefixLM
  - Image features will be used as prefix tokens for the Large Language Model
  - Most of the Vision Language Models currently using LLM use this structure
  - Representative researches
    - LLaVa: Large Language and Vision Assistant
    - LLaVa-CoT: Let Vision Language Models Reason Step-by-Step



Learning Strategies

- PrefixLM
  - LLaVa: Large Language and Vision Assistant
    - Dataset
      - Stage 1 : Pre-training for Feature Alignment
        - filter CC3M to 595K image-text-pairs
      - Stage 2 : Fine-tuning End-to-End
        - conversation_58k.json
        - detail_23k.json
        - complex_reasoning_77k.json

Learning Strategies

- PrefixLM
  - LLaVa: Large Language and Vision Assistant
    - Dataset
      - Use ChatGPT/GPT-4

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], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

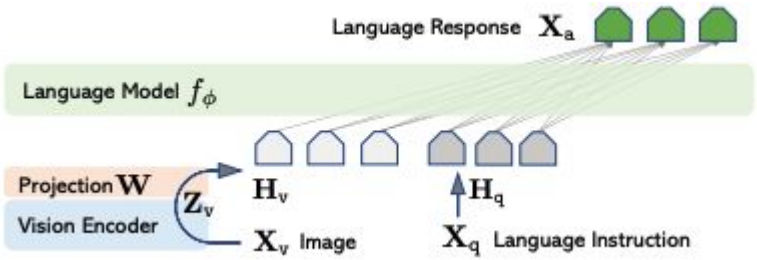
Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

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.
In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.
Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

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 to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.

Learning Strategies

- PrefixLM
  - LLaVa: Large Language and Vision Assistant
    - Pretraining
      - Blue : Freeze, Red : Tuning
      - Stage 1 : Pre-training for Feature Alignment
        - Vision Encoder + Linear Projection + LLM
      - Stage 2 : Fine-tuning End-to-End
        - Vision Encoder + Linear Projection + LLM

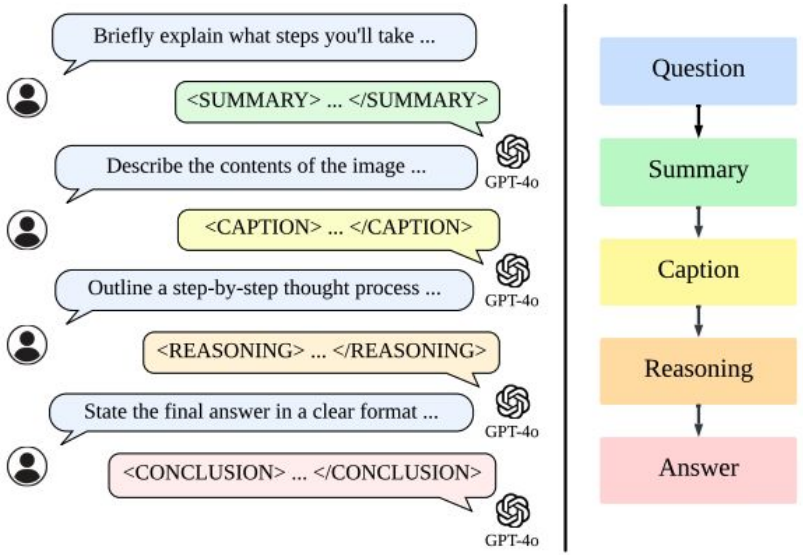


Learning Strategies

- PrefixLM
  - LLaVa-CoT: Let Vision Language Models Reason Step-by-Step
    - Use Chain-of-thought concept

Learning Strategies

- PrefixLM
  - LLaVa-CoT: Let Vision Language Models Reason Step-by-Step
    - Dataset Curation
      - Use GPT-4o



Learning Strategies

- PrefixLM
  - LLaVa-CoT: Let Vision Language Models Reason Step-by-Step
    - Pretraining
      - Llama-3.2-11B-Vision-Instruct
      - Full parameter fine-tuning(SFT)

Learning Strategies

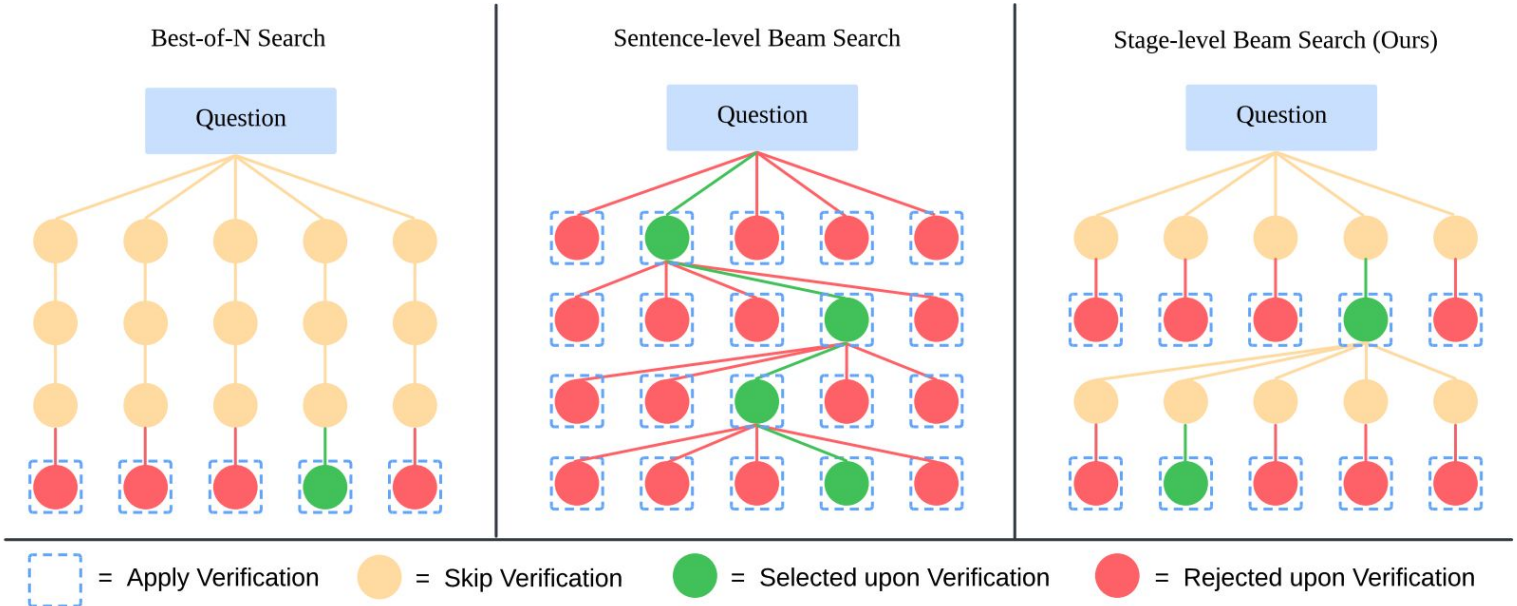
- PrefixLM
  - LLaVa-CoT: Let Vision Language Models Reason Step-by-Step
    - Inference
      - Sample N responses for the first stage in the solution
      - Randomly sample 2 responses and let the model determine which is better, keeping the better response
      - Repeat for N-1 times, retraining the best responses
      - Sample N responses for the next stage, then repeat steps 2-4 until all stages are processed

Learning Strategies

- PrefixLM
  - LLaVa-CoT: Let Vision Language Models Reason Step-by-Step
    - Inference
      - Sample N responses for the first stage in the solution
      - Randomly sample 2 responses and let the model determine which is better, keeping the better response
      - Repeat for N-1 times, retraining the best responses
      - Sample N responses for the next stage, then repeat steps 2-4 until all stages are processed

Learning Strategies

- PrefixLM
  - LLaVa-CoT: Let Vision Language Models Reason Step-by-Step
    - Inference



Learning Strategies

- PrefixLM
  - LLaVa-CoT: Let Vision Language Models Reason Step-by-Step
    - Inference

Model	MMStar	MMBench	MMVet	MathVista	AI2D	Hallusion	Average
Base Model							
Llama-3.2-11B-Vision-Instruct	49.8	65.8	57.6	48.6	77.3	40.3	56.6
Our Models							
LLaVA-CoT	57.6	75.0	60.3	54.8	85.7	47.8	63.5
LLaVA-CoT(BS = 2)	58.1	75.6	61.7	56.1	87.5	48.2	64.5

Table 4. **Experimental results during inference time.** LLaVA-CoT (BS = 2) denotes the model using stage-level beam search with a beam size of 2. The results show that stage-level beam search can achieve further significant performance improvements.

Method	Number of Beam	MMVet Score
No Inference Scaling	1	60.3
Best-of-N Search	10	60.9
Sentence-level Beam Search	2	58.4
Stage-level Beam Search	4	62.9

Table 5. **Comparison to Baseline Methods.** Our stage-level beam search outperforms both best-of-N and sentence-level beam search under **comparable inference time compute**.

Method	Number of Beam	MMVet Score
No Inference Scaling	1	60.3
Stage-level Beam Search	2	61.7
Stage-level Beam Search	3	62.3
Stage-level Beam Search	4	62.9

Table 6. **Scaling Trend of LLaVA-CoT.** As the the number of candidate responses increases, the model's performance consistently improves.

Learning Strategies

- PrefixLM
  - LLaVa-CoT: Let Vision Language Models Reason Step-by-Step
    - Inference

Question: What is the total price for a bottle of Merlot and a bottle of Cabernet shown in the image?

<*SUMMARY*> To solve the problem, I will first identify the prices of the Merlot and Cabernet bottles shown in the image. Then, I will add these prices together to find the total cost. </*SUMMARY*>
<*CAPTION*> The image displays three wine bottles with their respective prices. The first bottle is a Northstar Merlot from Columbia Valley, priced at $139.99. The second bottle is a Moët & Chandon Imperial from France, priced at $114.99. The third bottle is a Chateau St. Michelle Cold Creek Cabernet from Columbia Valley, priced at $109.99. </*CAPTION*>

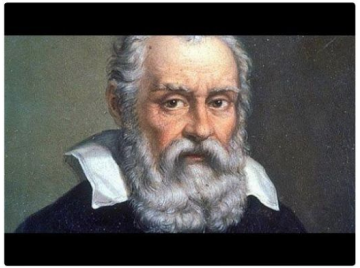
Answer without inference scaling:

<*REASONING*> To find the total price for a bottle of Merlot and a bottle of Cabernet, I will identify the prices of these two specific wines. The Merlot is priced at $139.99, and the Cabernet is priced at $109.99. **Adding these two prices together will give the total cost for one bottle of each.** </*REASONING*>
<*CONCLUSION*> 149.99 </*CONCLUSION*>

Answer from beam search:

<*REASONING*> To find the total price for a bottle of Merlot and a bottle of Cabernet, I will add the price of the Northstar Merlot ($139.99) to the price of the Chateau St Michelle Cold Creek Cabernet ($109.99). **The calculation is as follows: $139.99 + $109.99 = $249.98.** </*REASONING*>
<*CONCLUSION*> The total price for a bottle of Merlot and a bottle of Cabernet shown in the image is $249.98. </*CONCLUSION*>

Example 2: Open-ended Problems



Question: Can you give a short introduction to this person?

Answer from Llama-3.2-11B-Vision-Instruct:

This is a portrait of **an elderly gentleman**, likely a man of great wisdom and dignity, with a full beard and a serious expression.

He is dressed in a dark blue garment with white details around the collar, which suggests a sense of formality and tradition. His hair is mostly gray, and his facial expression is solemn and dignified.

Answer from LLaVA-CoT:

Summary stage: What's the problem? What should I do?

<SUMMARY> I will identify the person in the image and provide a brief introduction based on historical context and known facts about the individual. *</SUMMARY>*

Caption stage: What can I know from the image?

<CAPTION> The image depicts a man with a white beard and hair, wearing a black and white attire, which is typical of historical portraits from the Renaissance era. *</CAPTION>*

Reasoning stage: How to solve the problem step-by-step?

<REASONING> The attire and style of the portrait suggest a historical figure from the Renaissance. The white beard and hair are characteristic of **Galileo Galilei**, an Italian astronomer and physicist known for his contributions to science during that period. By associating these features with historical records, I can identify the person as Galileo Galilei.*</REASONING>*

Conclusion stage: What is the final answer?

<CONCLUSION> Galileo Galilei, a renowned Italian astronomer, physicist, and mathematician, is the subject of this image. He is depicted with a white beard and hair, dressed in black and white attire, a style typical of the Renaissance era. Galileo is celebrated for his groundbreaking contributions to the scientific revolution of the 16th and 17th centuries. He is best known for his work on the principles of motion, his development of the telescope, and his support for the heliocentric model of the solar system. His legacy continues to influence modern science and his name has become synonymous with the pursuit of knowledge and the scientific method. *</CONCLUSION>*

Example 1: Fill-in-the-blank Problems

Question: Which spot is the red car parking on?

Answer from Llama-3.2-11B-Vision-Instruct:

To determine which spot the red car is parked on, let's break down the information step by step:

1. Identify the Red Car's Position: **The red car is parked in a space labeled '17'.**
2. Understanding Parking Space Labeling: Typically, parking spaces are labeled with numbers or letters to indicate their location within the parking lot. The number '17' likely corresponds to a specific parking space designated for a car.
3. Analyzing the Information: Given that the red car is parked in a space labeled '17', this directly indicates the parking spot where the red car is located.
4. Conclusion: Based on the information provided, the red car is parked in the parking space labeled '17'.

Answer from LLaVA-CoT:

Summary stage: What's the problem? What should I do?

<SUMMARY> To solve the problem, I will analyze the image to identify the parking spot where the red car is located. I will focus on the markings and numbers on the ground to determine the correct parking spot. *</SUMMARY>*

Caption stage: What can I know from the image?

<CAPTION> The image shows a parking lot with several cars parked. There are white lines and numbers on the ground, indicating the parking spots. A red car is parked in the lot, and the numbers on the ground are visible, which help identify the specific spot. *</CAPTION>*

Reasoning stage: How to solve the problem step-by-step?

<REASONING> The image displays a grid of parking spots marked with numbers. The red car is positioned within this grid. **By examining the numbers closest to the red car, I can determine which spot it is parked in.** The numbers are clearly visible, and the car's position aligns with one of these numbered spots.*</REASONING>*

Conclusion stage: What is the final answer?

<CONCLUSION> **31** *</CONCLUSION>*

Thank you