Multimodal Learning

Léopold Maytié

Artificial and Natural Intelligence Toulouse Institute (ANITI)
25th March, 2024

I- Introduction

II- Recall

- A- MLP
- B- CNN
- C- RNN
- D- Transformer
- E- How to train a model

II- How to learn from Multimodality ?

III- Multimodal Tasks

- A- Image Captioning
- B- VQA
- C- Multimodal Dialogue
- D- Language, Vision and Navigation

IV- Examples of Models

V- Conclusion

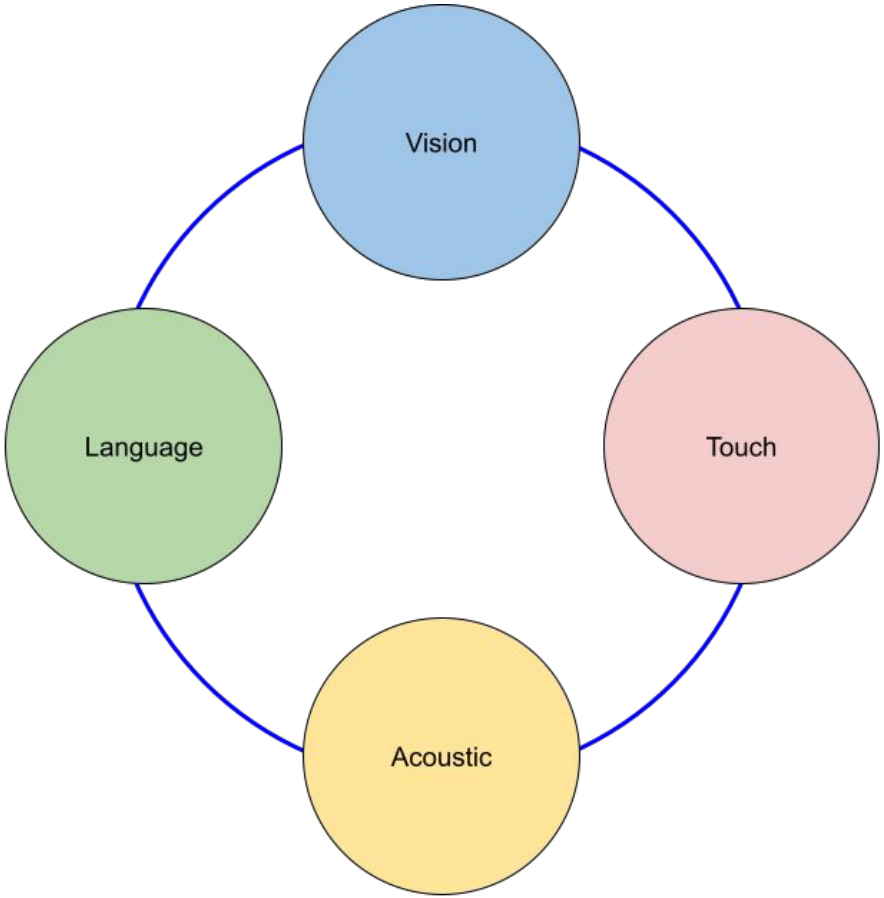
Introduction

I- Introduction

Multimodal Machine Learning is the study of computer algorithms that learn and improve through the use and experience of data from multiple modalities

Multimodal Artificial Intelligence studies computer agents able to demonstrate intelligence capabilities such as understanding, reasoning and planning, through multimodal experiences, and data

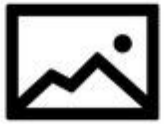
I- Introduction



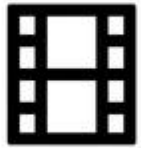
I- Introduction

Why do we need multimodal data ?

I- Introduction

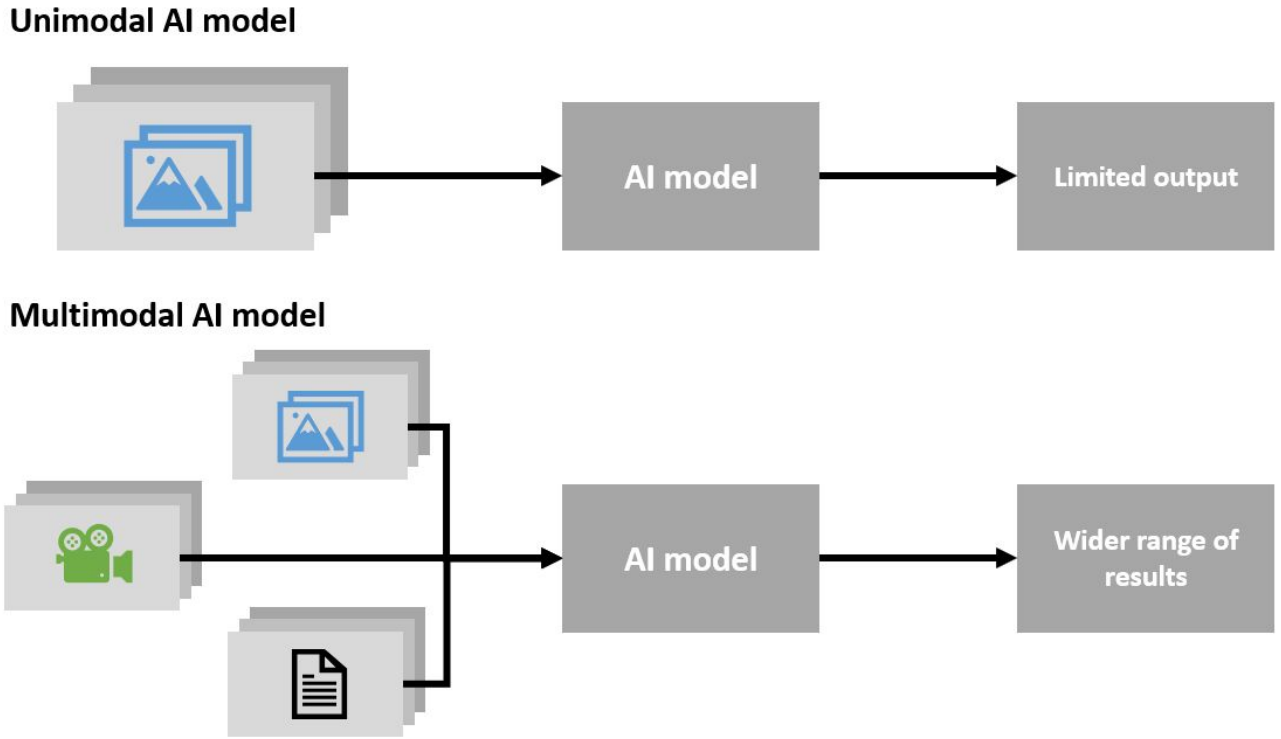


Limited World Model and Knowledge



Better World Model and Knowledge
→Robustness

I- Introduction

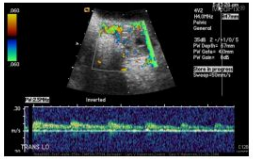
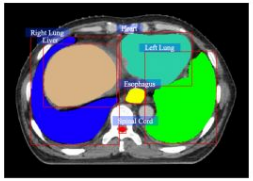
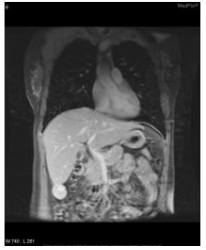
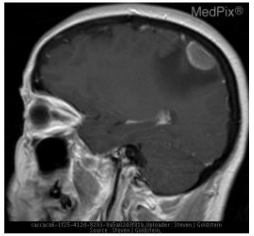


autonomous driving, image generation, robot navigation, …

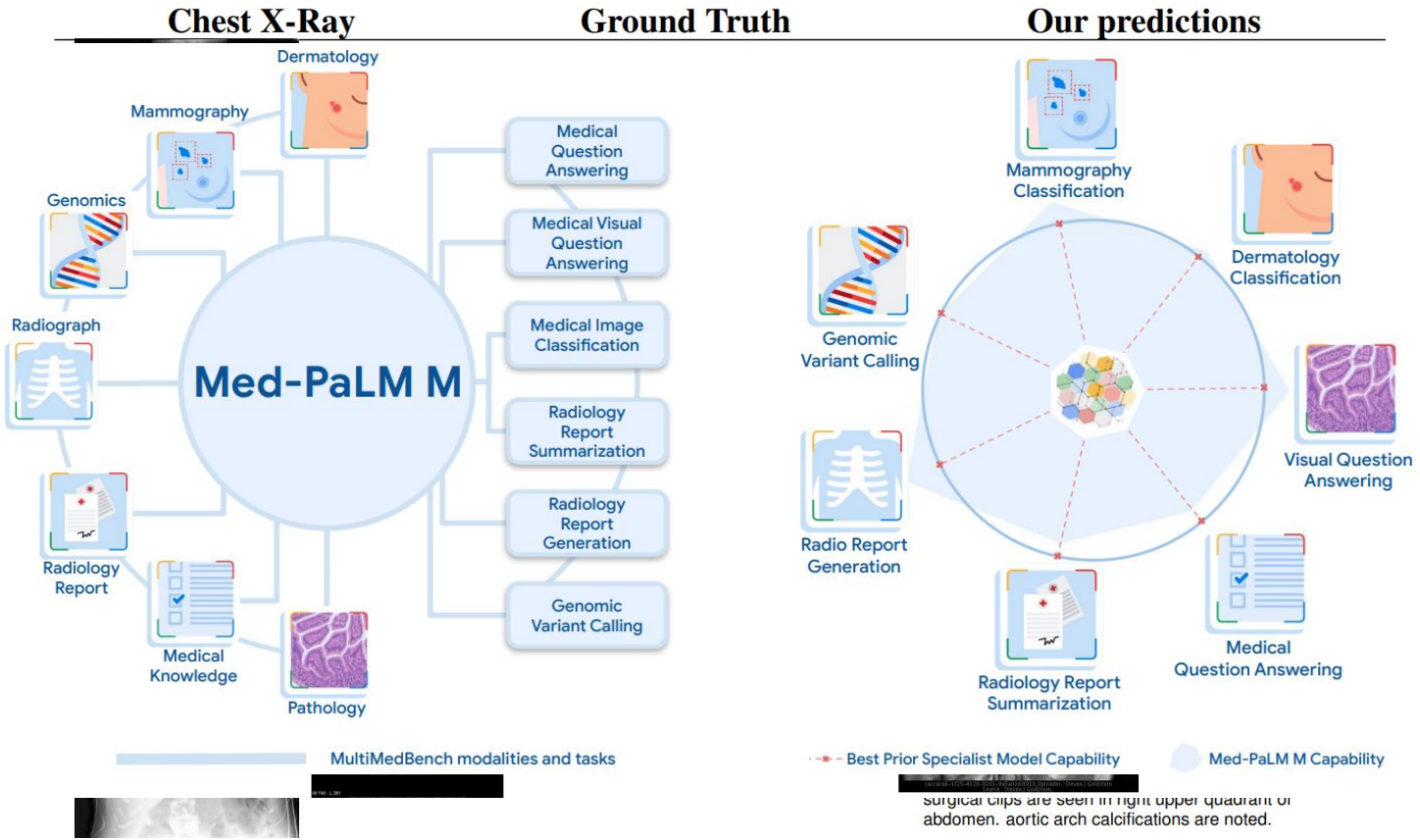
I- Introduction

Chest X-Ray	Ground Truth	Our predictions
	Lungs remain well inflated without evidence of focal airspace consolidation, pleural effusions, pulmonary edema or pneumothorax.Irregularity in the right humeral neck is related to a known healing fracture secondary to recent fall. PA and lateral views of the chest ____ at 09:55 are submitted.	no findings. no pneumonia. no pleural effusion. no edema. there is little change and no evidence of acute cardiopulmonary disease. no pneumonia, vascular congestion, pleural effusion.of incidental note is an azygos fissure, of no clinical significance. this raises possibility of a normal variant.
	1. Stable bilateral small pleural effusions and atelectasis. 2. Enlarged pulmonary artery, suggesting pulmonary hypertension. No significant interval change. Bilateral small pleural effusions and adjacent atelectasis are overall unchanged. The heart is top-normal in size, unchanged. The pulmonary artery is enlarged, suggesting pulmonary hypertension. No demand, focal consolidation to suggest pneumonia, or pneumothorax.	pleural effusion present. lung opacity present. no edema. cardiomegaly present. atelectasis present. as compared to previous radiograph, there is an increase in extent of a pre existing small left pleural effusion with subsequent atelectasis at left lung bases. otherwise, radiograph is unchanged. moderate cardiomegaly. mild fluid overload no overt pulmonary edema. no new focal parenchymal opacities suggesting pneumonia. unchanged position of right pectoral port a cath.
	There is decrease in now small right pleural effusion. There is no pneumothorax. There is a new right pacer pigtail catheter. Cardiomediastinal contours are unchanged. Lines and tubes are in standard position. Left lower lobe opacities, a combination of pleural effusion and atelectasis, are unchanged.	uncertain pneumonia. pleural effusion present. lung opacity present. atelectasis present. bilateral pleural effusions, left greater than right. bibasilar opacities potentially atelectasis in setting of low lung volumes. infection be excluded. frontal and lateral views of chest demonstrate low lung volumes, which accentuate bronchovascular markings. there are small bilateral pleural effusions, right greater than left, with adjacent atelectasis. there is no focal consolidation pneumothorax. cardiomediastinal silhouette is within normal limits. surgical clips are seen in right upper quadrant of abdomen. aortic arch calcifications are noted.

I- Introduction

Chest X-Ray	Ground Truth	Our predictions
		
VQA-Med-2020 [13]		**Q:** what abnormality is seen in the image? **A:** ovarian torsion
	**Q:** what is abnormal in the ct scan? **A:** partial anomalous pulmonary venous return
SLAKE [57]		**Q:** Does the image contain left lung? **A:** Yes
Q: What is the function of the rightmost organ in this picture? **A:** Breathe	
VQA-Med-2021 [15]		**Q:** What is most alarming about this mri? **A:** focal nodular hyperplasia
	**Q:** What abnormality is seen in the image? **A:** Enhancing lesion right parietal lobe with surrounding edema
	surgical clips are seen in right upper quadrant of abdomen. aortic arch calcifications are noted.	

I- Introduction

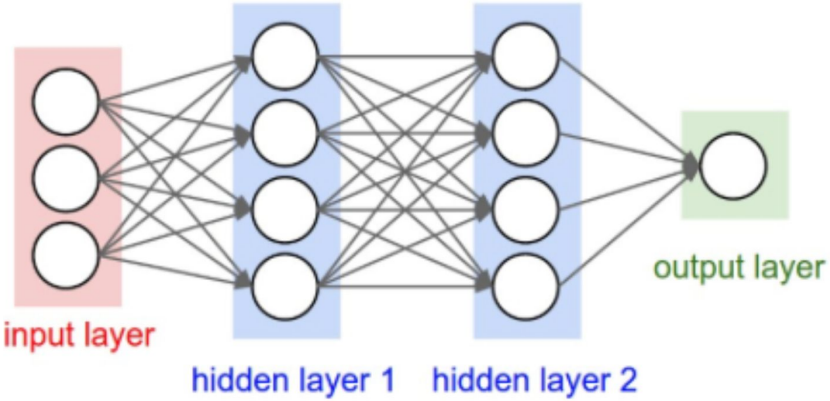


Recall

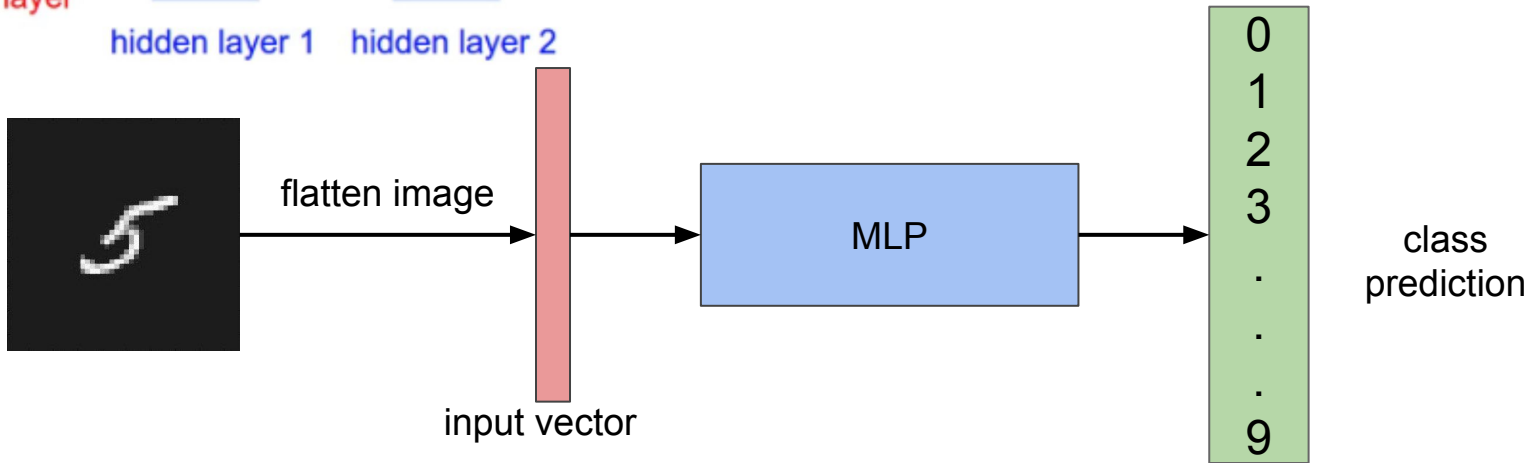
II- Recall

MLP :

- Most basic Deep Learning Model that can handle only **vector** as input



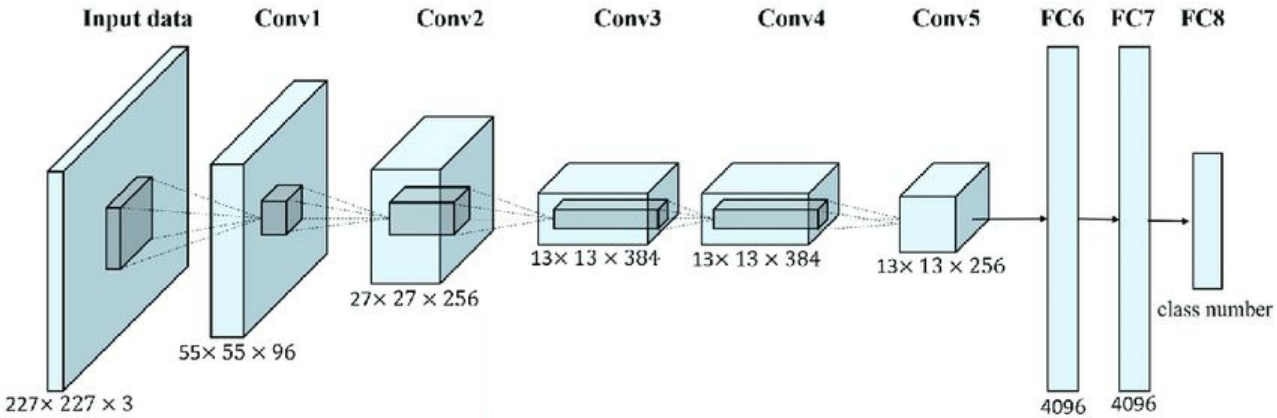
Linear Layer : $\hat{y} = f(Wx + b)$



II- Recall

CNN :

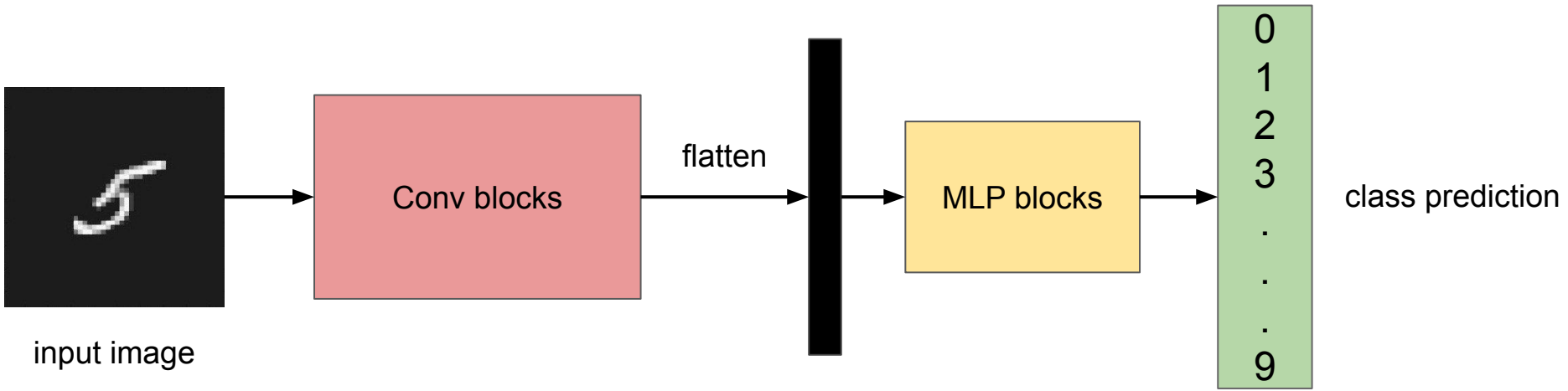
- Very useful to treat **images** or input which have spatial relation thanks to Convolutional Layers



II- Recall

CNN :

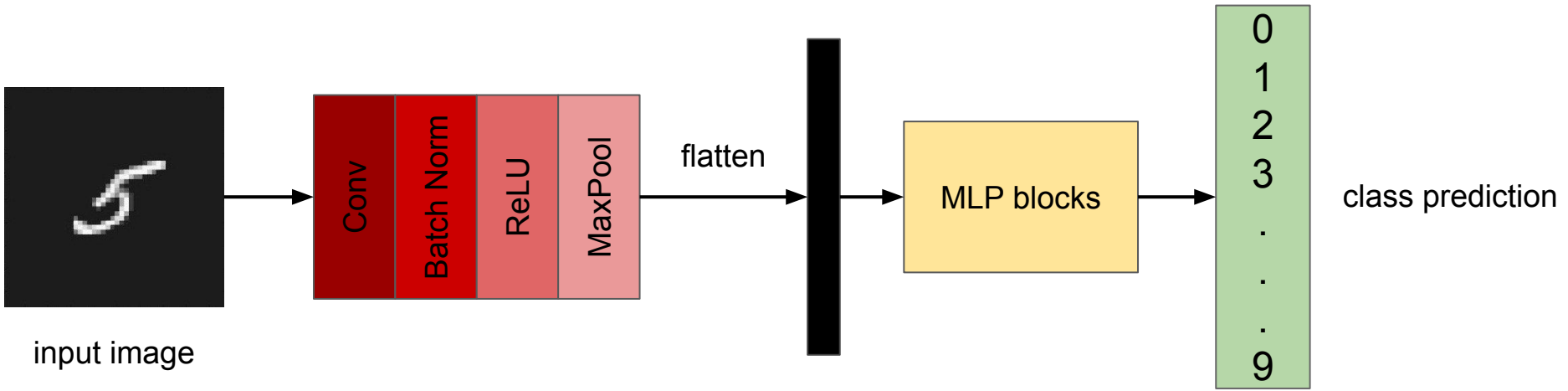
- Very useful to treat **images** or input which have spatial relation thanks to Convolutional Layers



II- Recall

CNN :

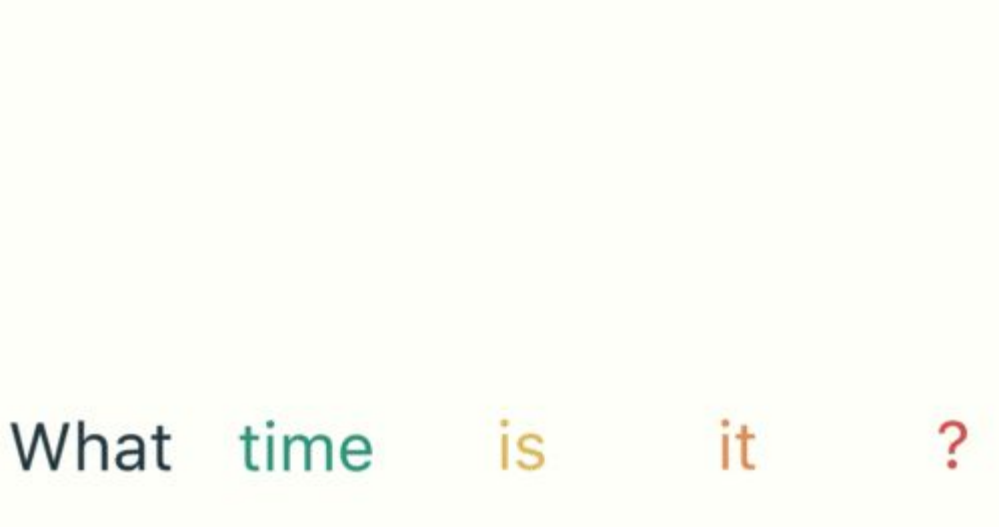
- Very useful to treat **images** or input which have spatial relation thanks to Convolutional Layers



II- Recall

RNN :

- Very good to treat **sequential data** thanks to its hidden state acting like a memory



II- Recall

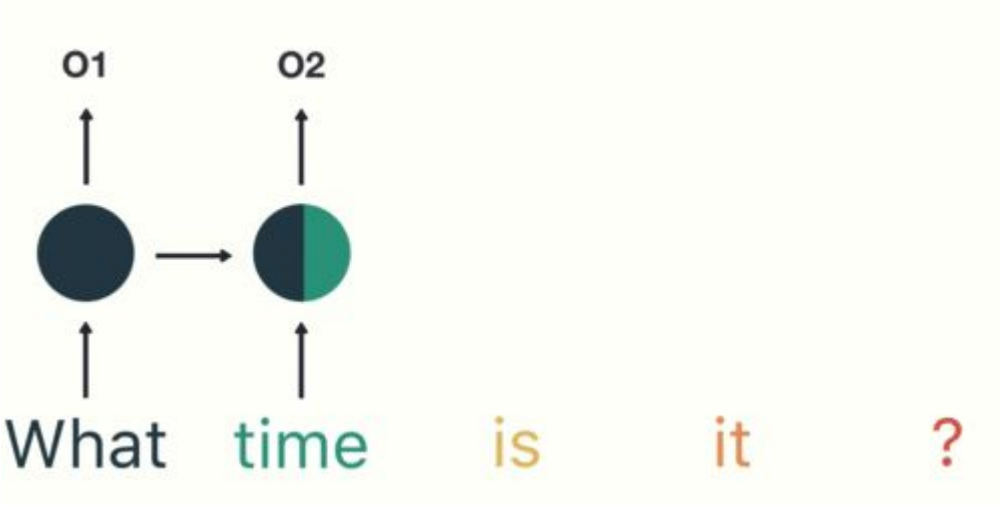
RNN :

- Very good to treat **sequential data** thanks to its hidden state acting like a memory

II- Recall

RNN :

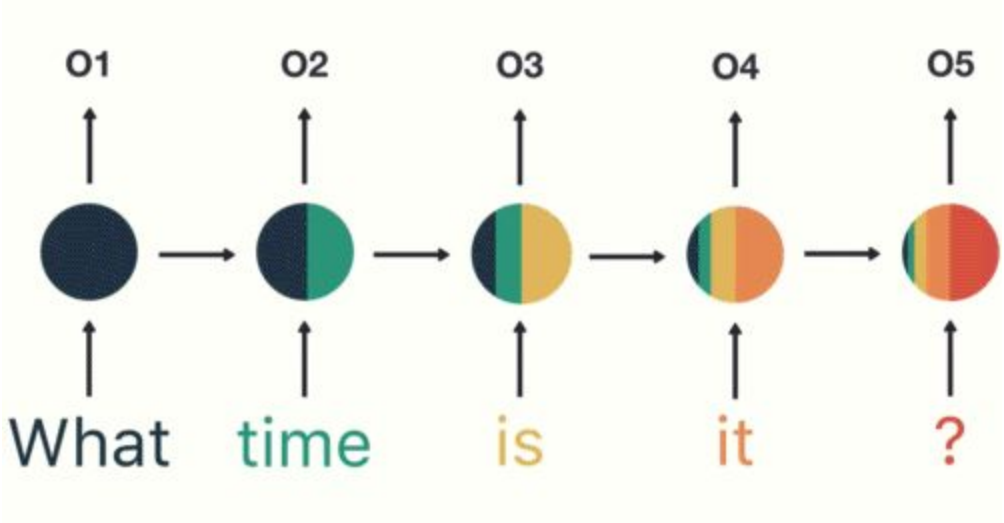
- Very good to treat **sequential data** thanks to its hidden state acting like a memory



II- Recall

RNN :

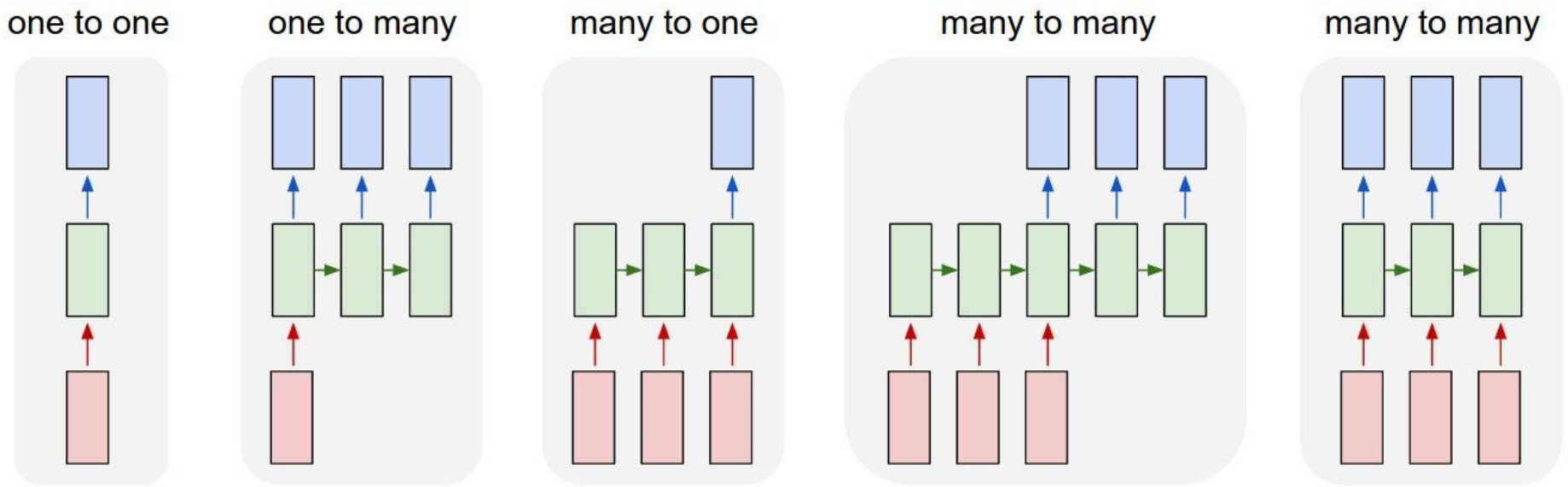
- Very good to treat **sequential data** thanks to its hidden state acting like a memory



II- Recall

RNN :

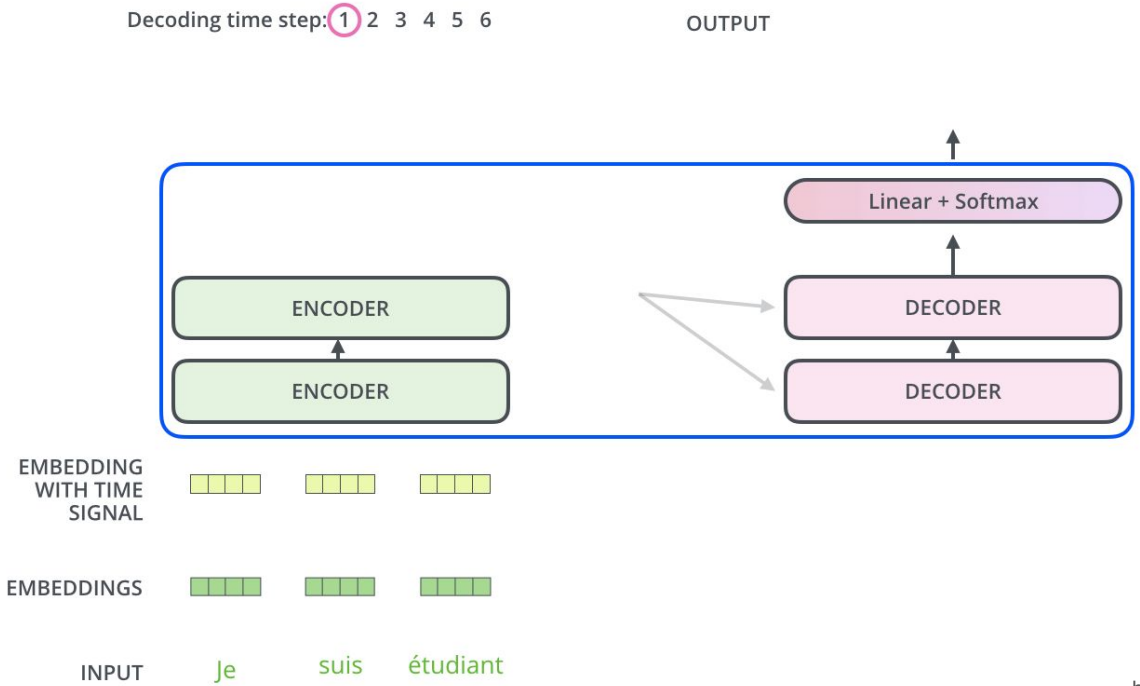
- Different type of RNN depending on the task you want



II- Recall

Transformers :

- Very good to treat **sequential data** (ex : text)
- Able to learn interactions between different element of the sentence

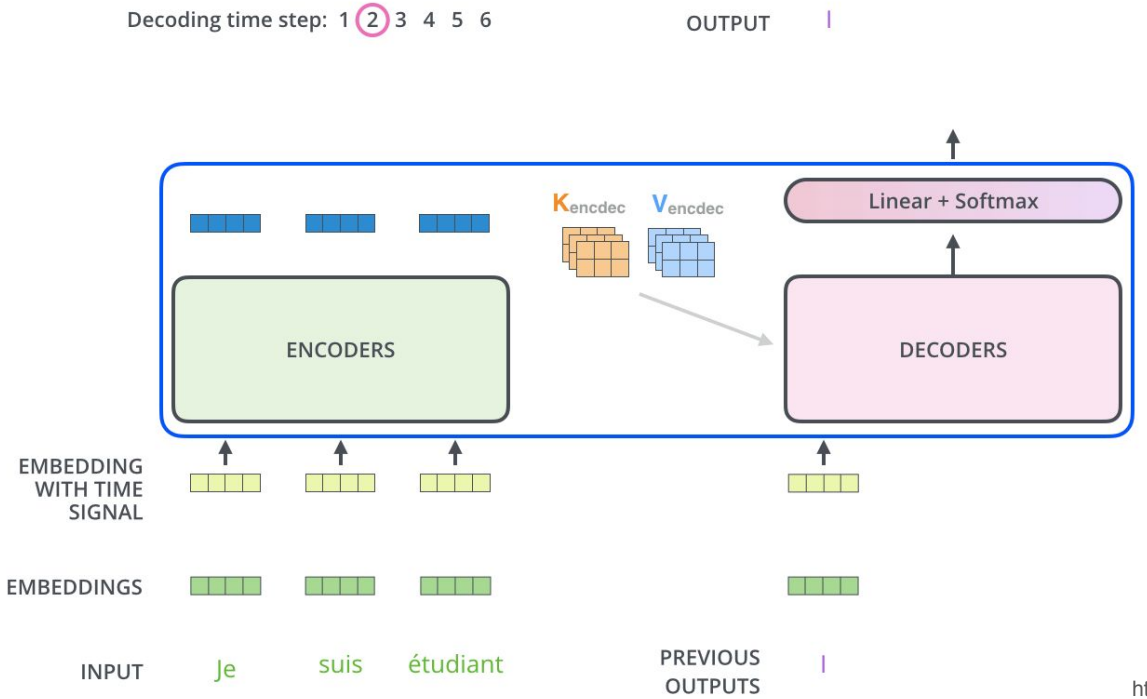


https://jalammar.github.io/illustrated-transformer/

II- Recall

Transformers :

- Very good to treat **sequential data** (ex : text)
- Able to learn interactions between different element of the sentence

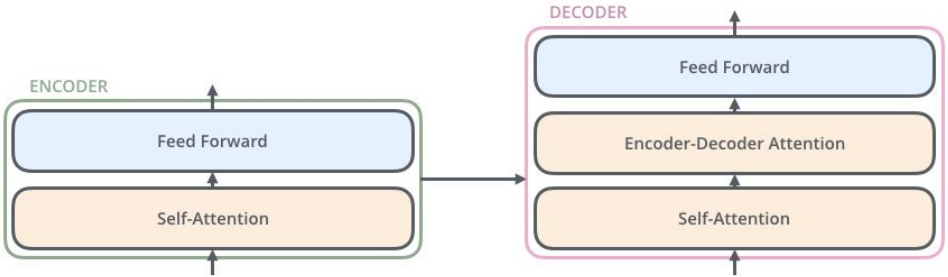


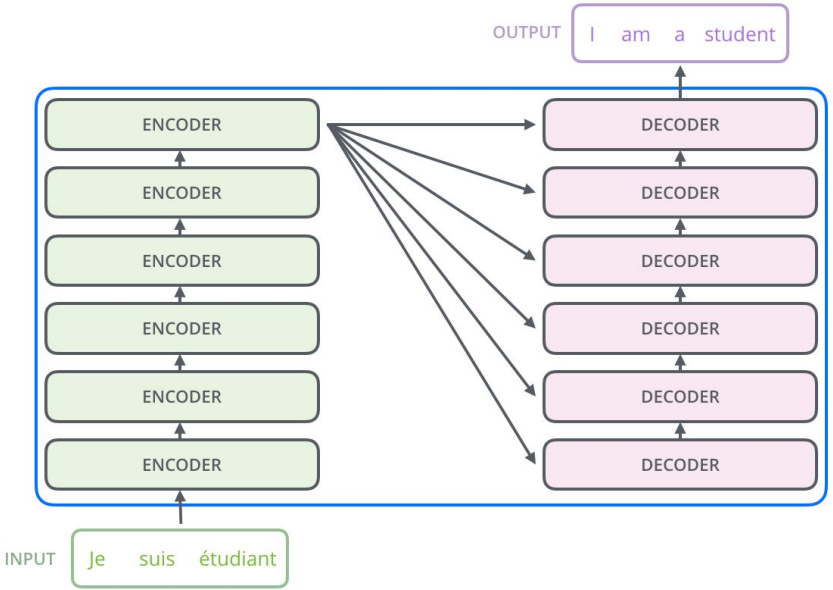
https://jalammar.github.io/illustrated-transformer/

II- Recall

Transformers :

what's inside encoder and decoder ?





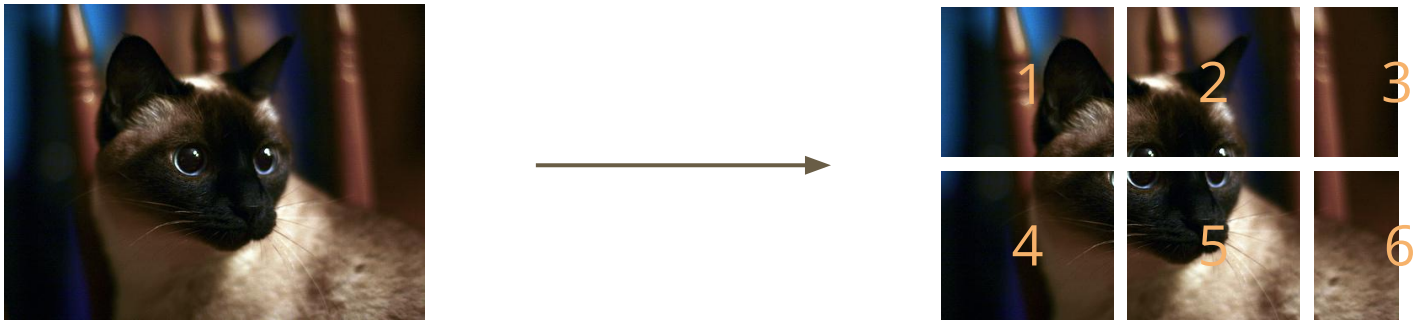
Global architecture Transformer

https://jalammar.github.io/illustrated-transformer/

II- Recall

Transformers :

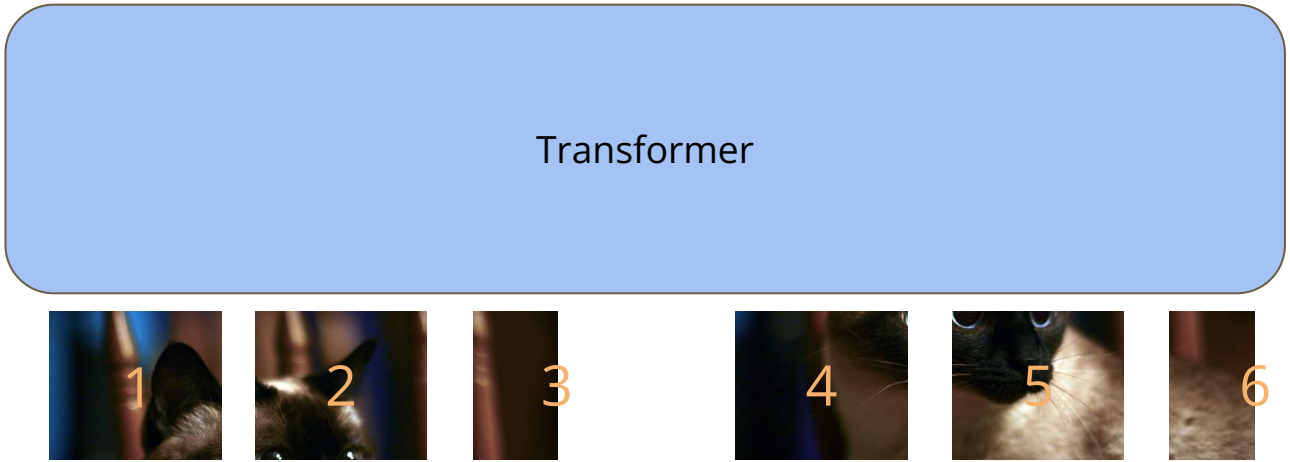
- Now also used for images → need to convert image to sequence



II- Recall

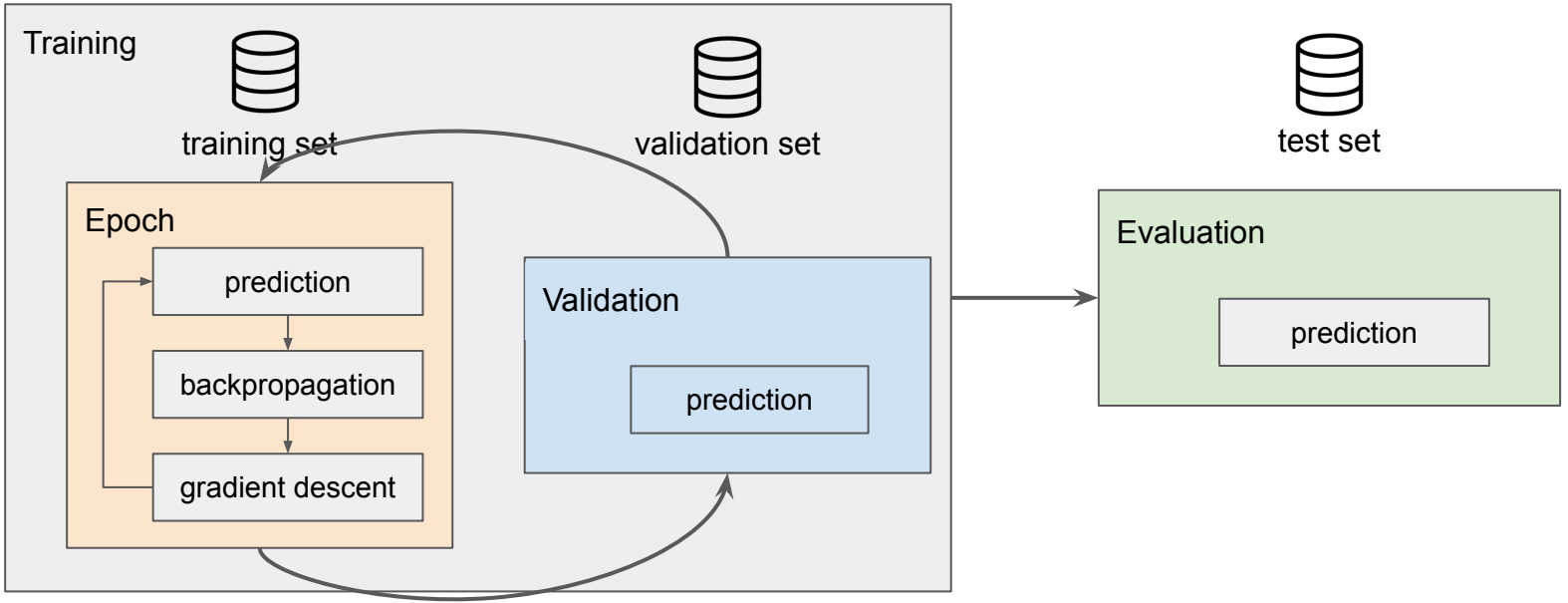
Transformers :

- Now also used for images → need to convert image to sequence
- Image divide into patches passed to the Transformer



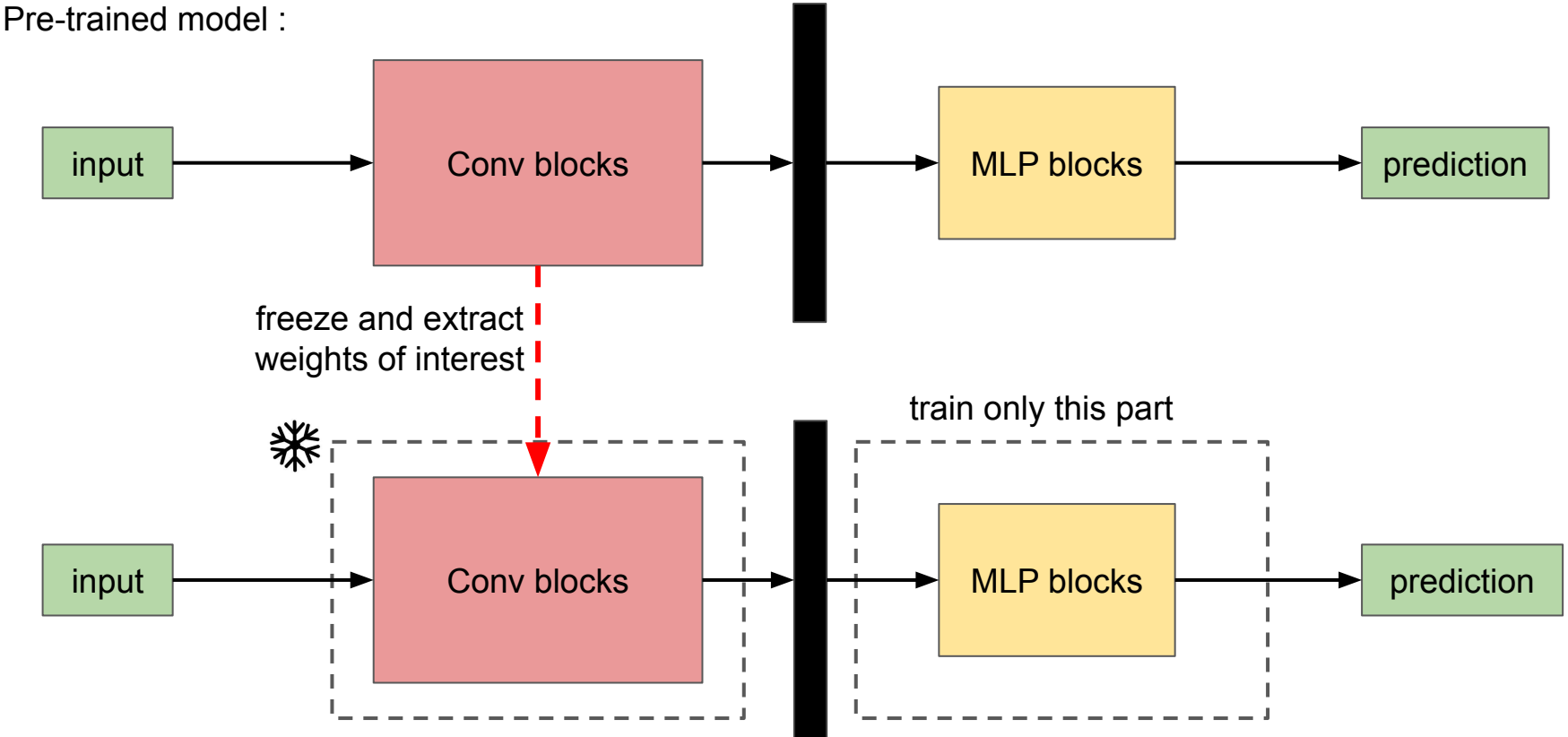
II- Recall

Train a model :



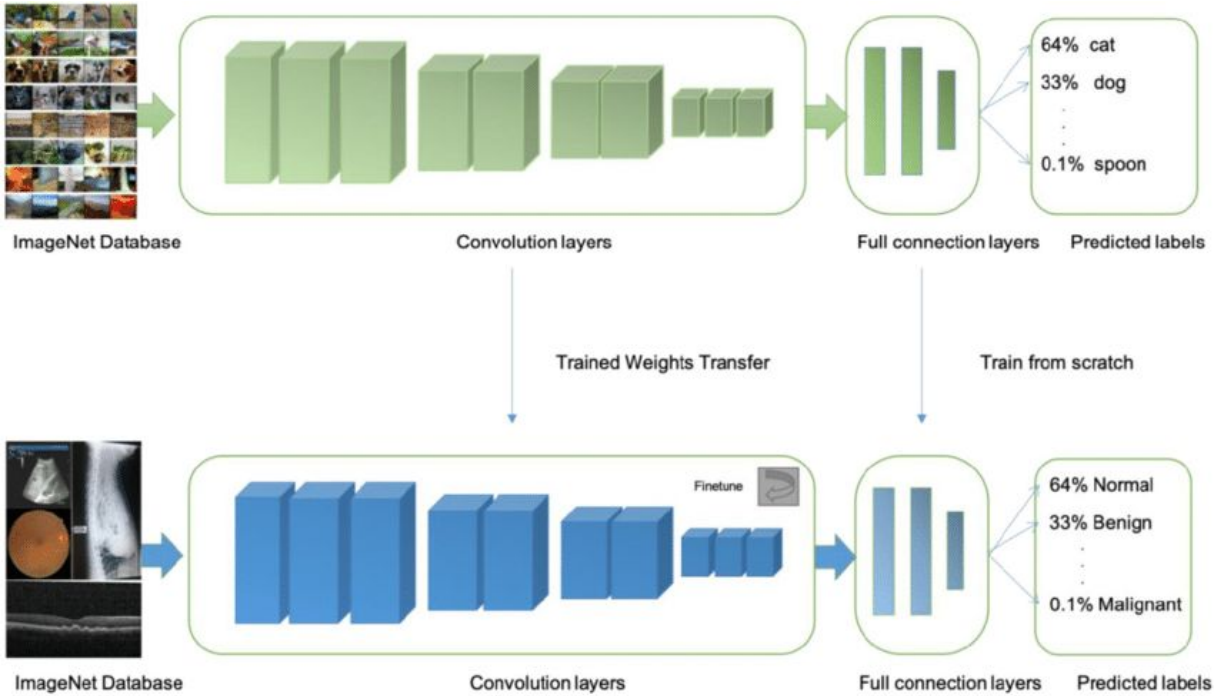
II- Recall

Use pre-trained model : **transfer learning**



II- Recall

Use pre-trained model :

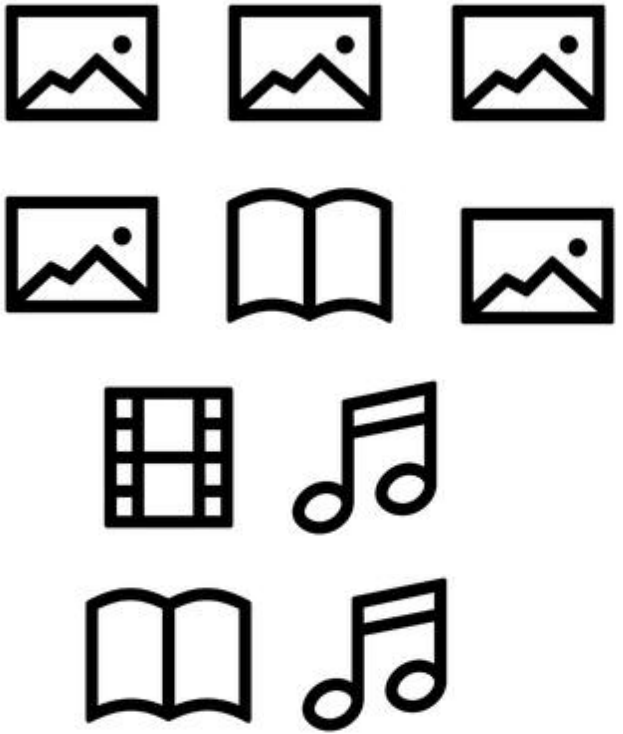


How to learn from Multimodality ?

How to learn from Multimodality ?

Unimodal Data

Multimodal Data



How to learn from Multimodality ?

Raw modalities :
close to the sensor

Abstract modalities :
far from the sensor

Speech signal

Image

Language

Detected objects

Object categories

Sentiment intensity

How to learn from Multimodality ?

the underside of a large airplane in the sky.

a plane is taking off from the airport.

an airplane flying in the air with a sky background

an airplane in the air seen from below at twilight

a grey jet airliner passing across a hazy blue sky.

image + text : multimodal

How to learn from Multimodality ?

image + segmented image : multimodal

How to learn from Multimodality ?

the underside of a large airplane in the sky.

a plane is taking off from the airport.

an airplane flying in the air with a sky background

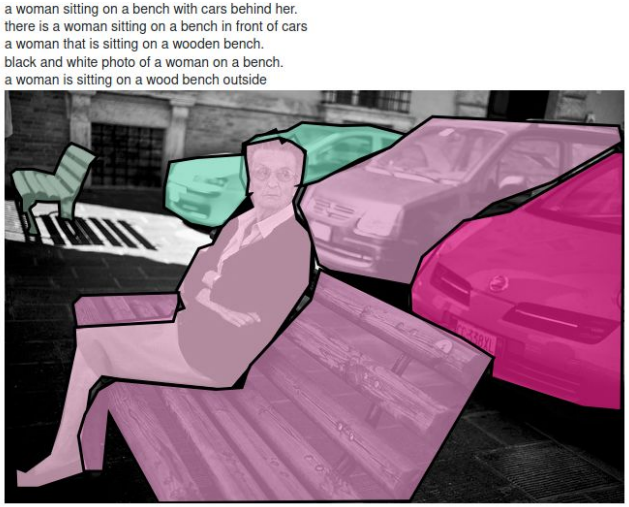
an airplane in the air seen from below at twilight

a grey jet airliner passing across a hazy blue sky.

image + segmented image + text : multimodal

How to learn from Multimodality ?

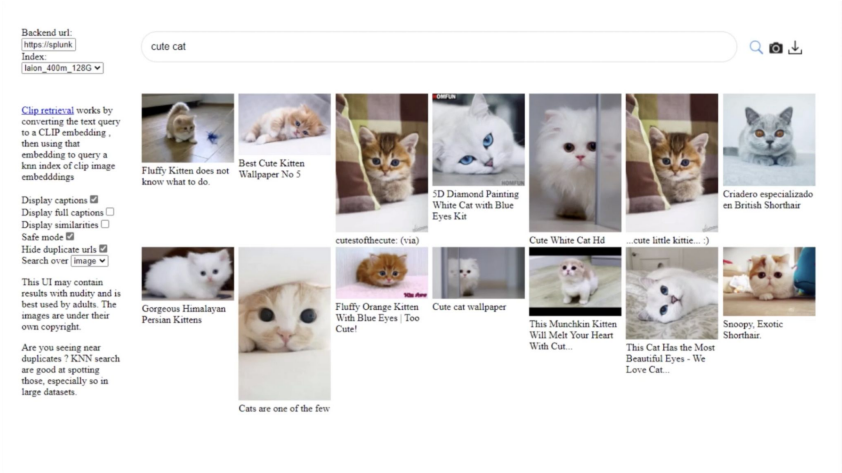
Datasets



- 330K images
- Images segmentation
- Simple descriptive sentences

How to learn from Multimodality ?

Datasets

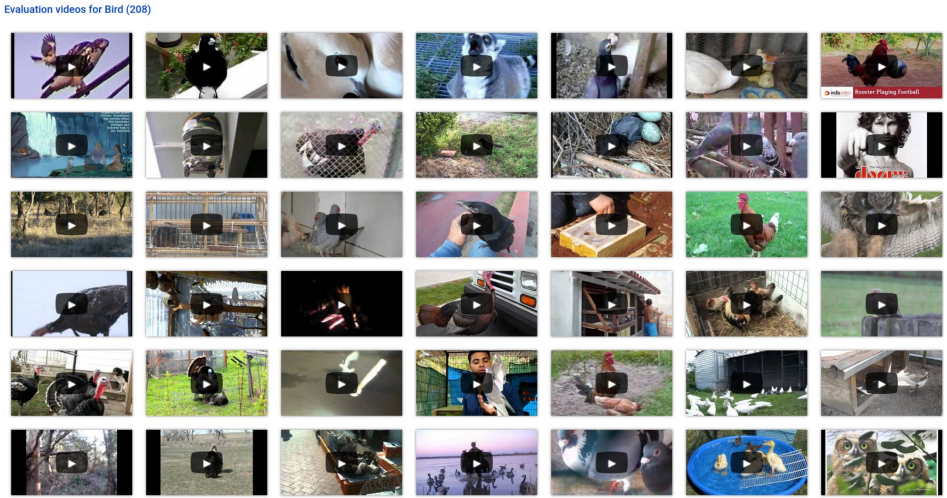


- Image text pairs
- 400M or 5B pairs
- Lot of different specialized version

How to learn from Multimodality ?

Datasets

AudioSet

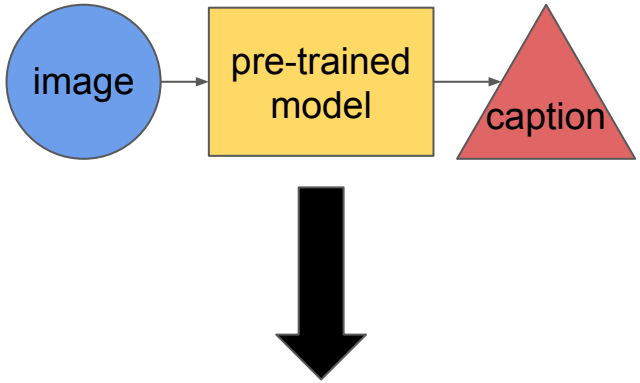


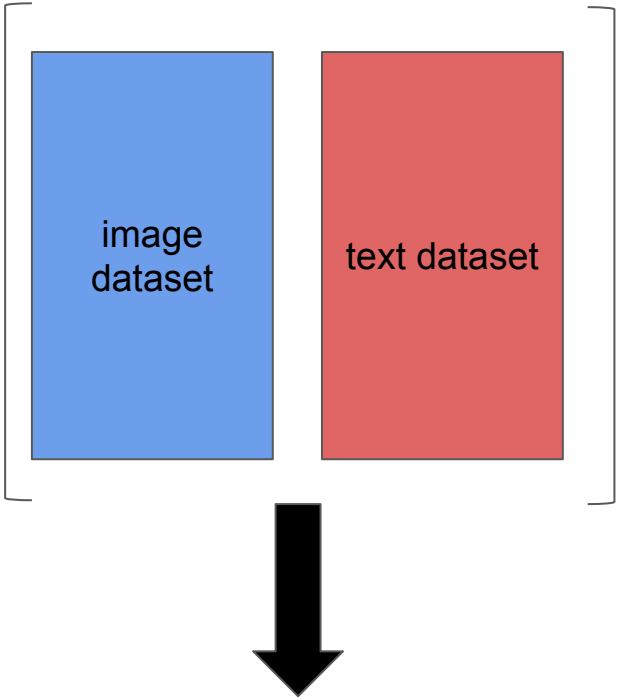
- 2.1 million annotated videos
- 5.8 thousand hours of audio
- 527 classes of annotated sounds

How to learn from Multimodality ?

Datasets :

- create own multimodal dataset





How to learn from Multimodality ?

How to train a model that take as input different modalities ?

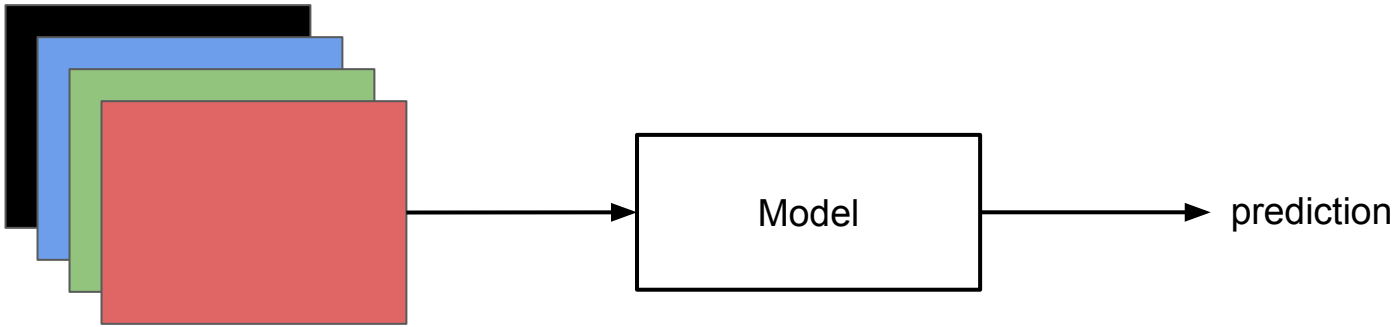
How to learn from Multimodality ?

Concatenation of raw modalities :

- Need to be very close (ex : 2 images : RGB + Depth)
- Increase the information given to the model to do better prediction

RGB

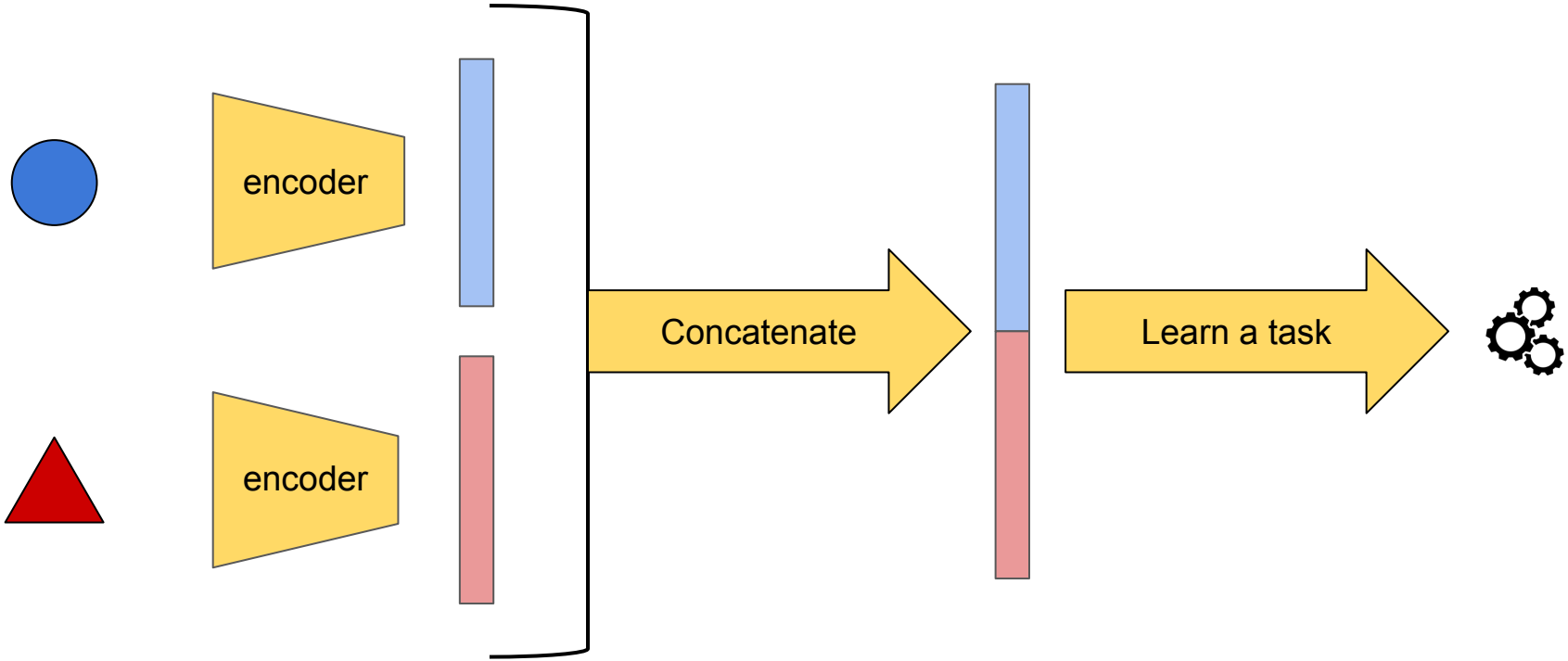
Depth



concatenation →4 channels

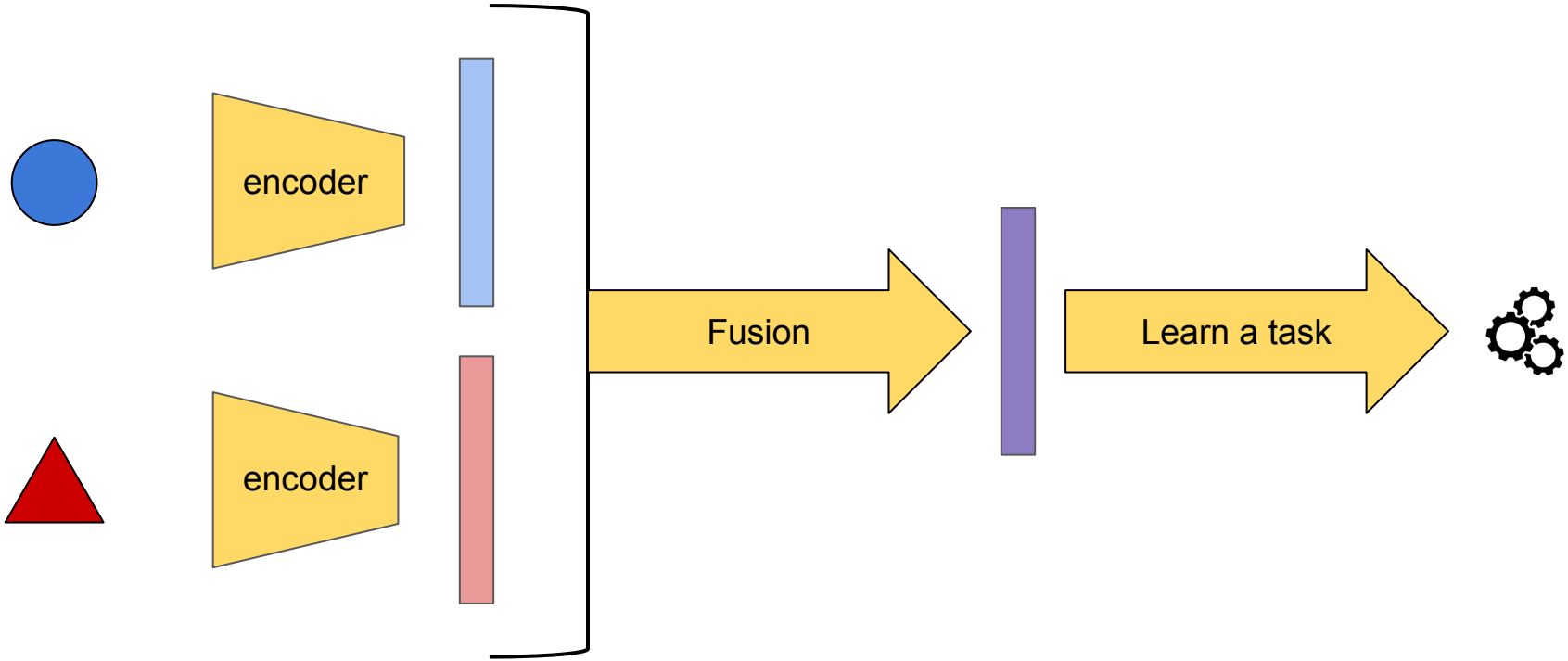
How to learn from Multimodality ?

Common approach : Encode modalities in latent space and merge them



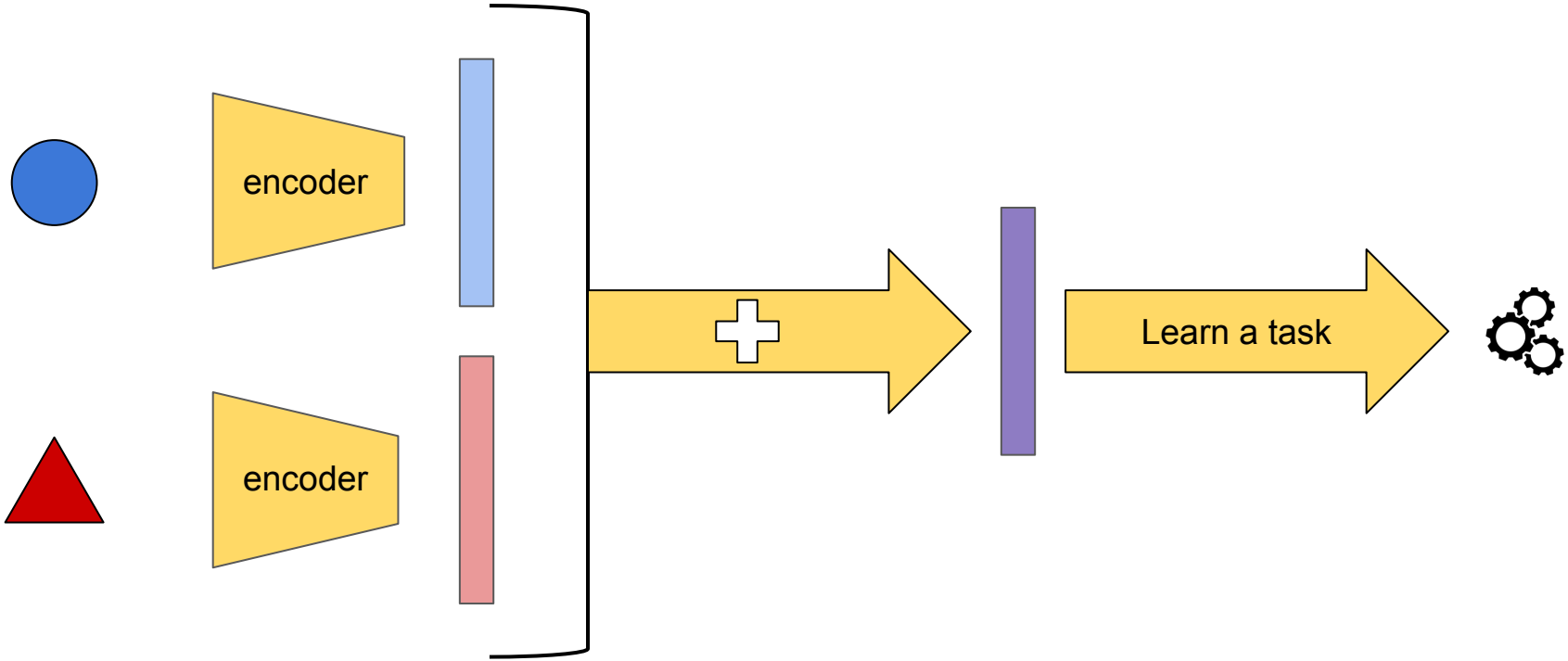
How to learn from Multimodality ?

Fusion of the latent vectors



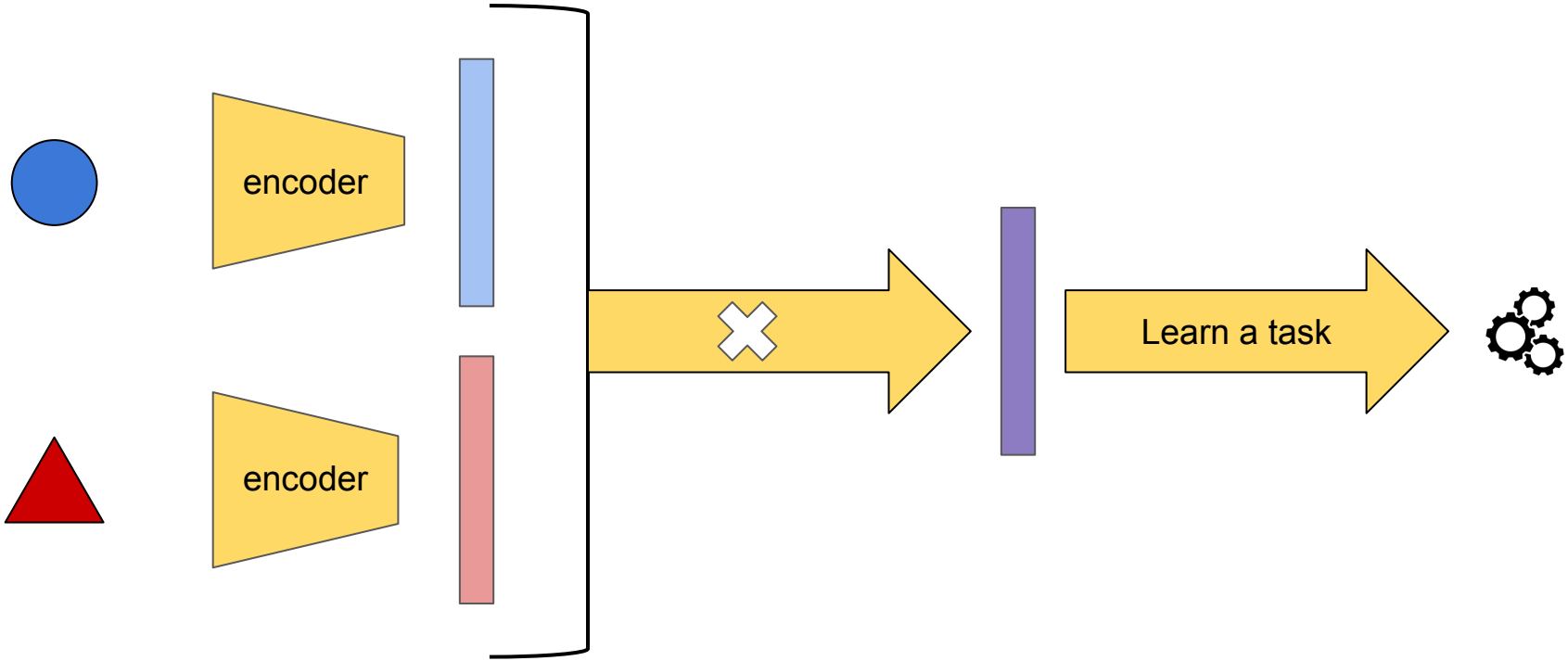
How to learn from Multimodality ?

Fusion of the latent vectors



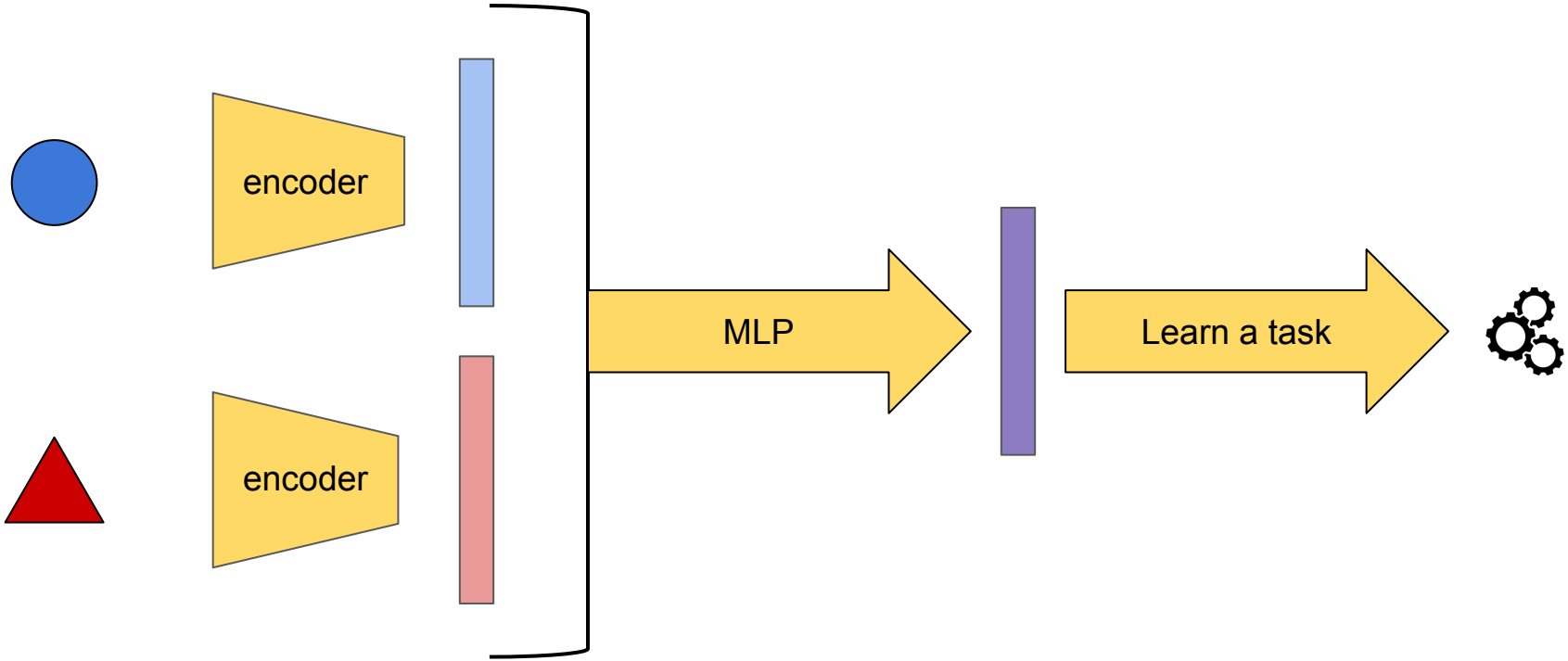
How to learn from Multimodality ?

Fusion of the latent vectors



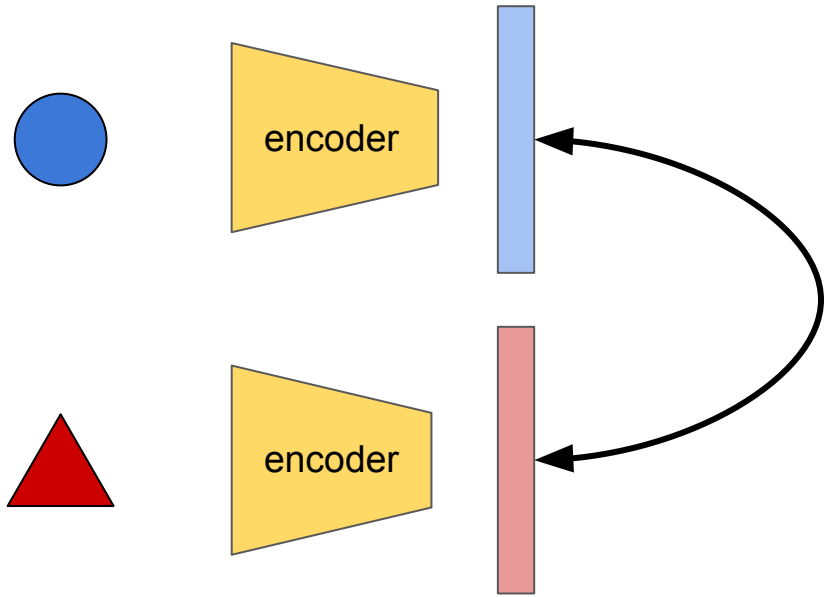
How to learn from Multimodality ?

Fusion of the latent vectors



How to learn from Multimodality ?

Coordinate the latent vectors



Loss to force the coordination :
contrastive loss

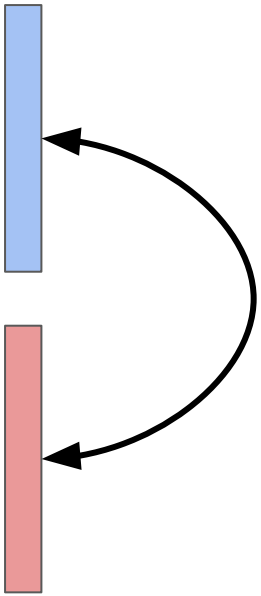
How to learn from Multimodality ?

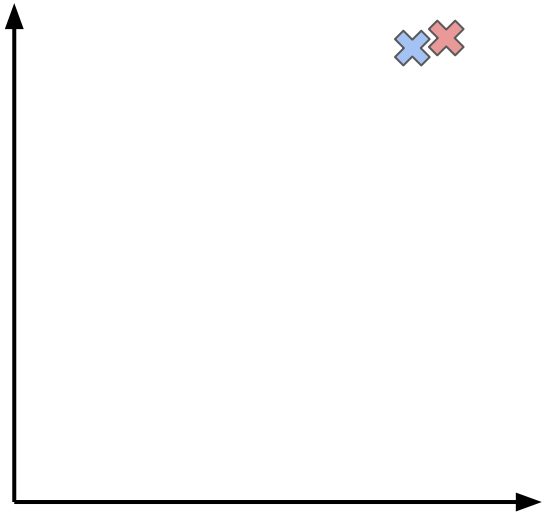
Coordinate the latent vectors

a plane

encoder

encoder

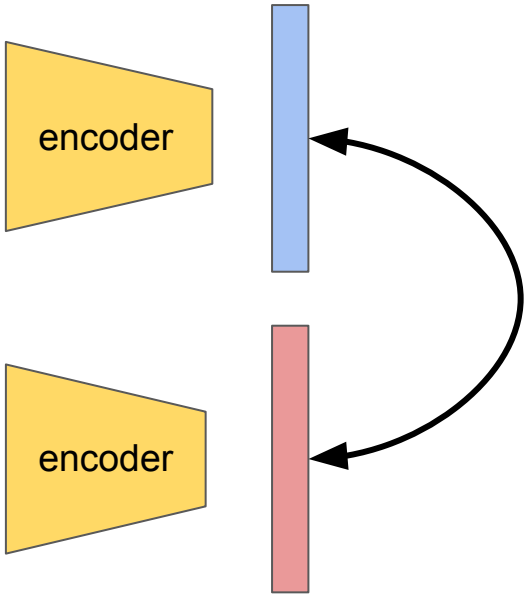


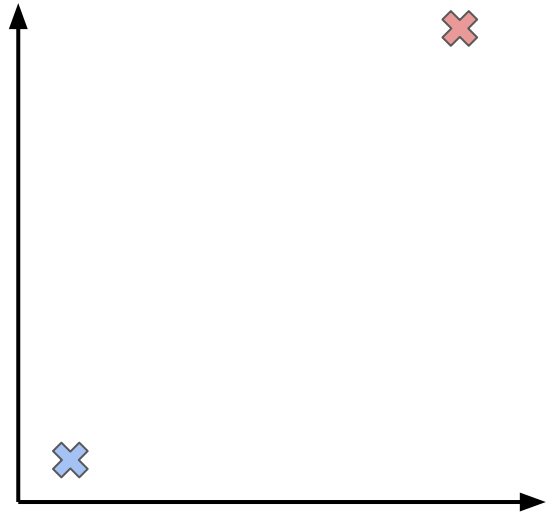


How to learn from Multimodality ?

Coordinate the latent vectors

a plane

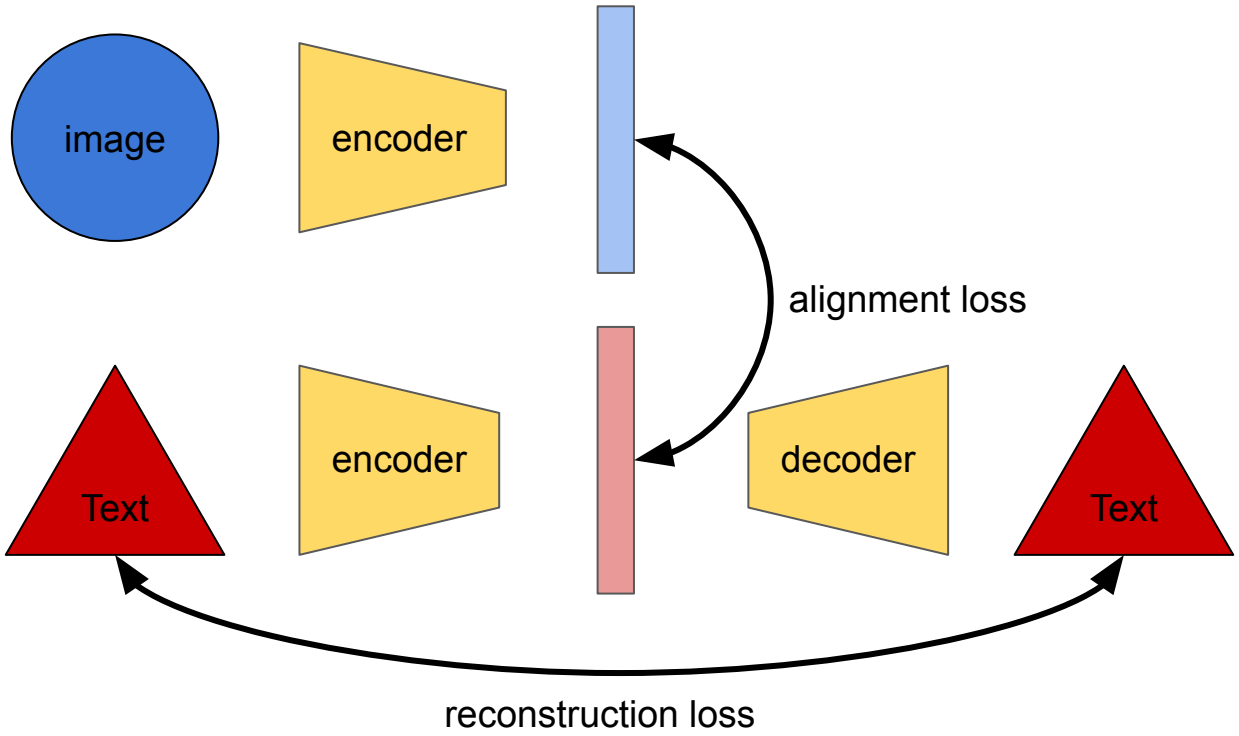




How to learn from Multimodality ?

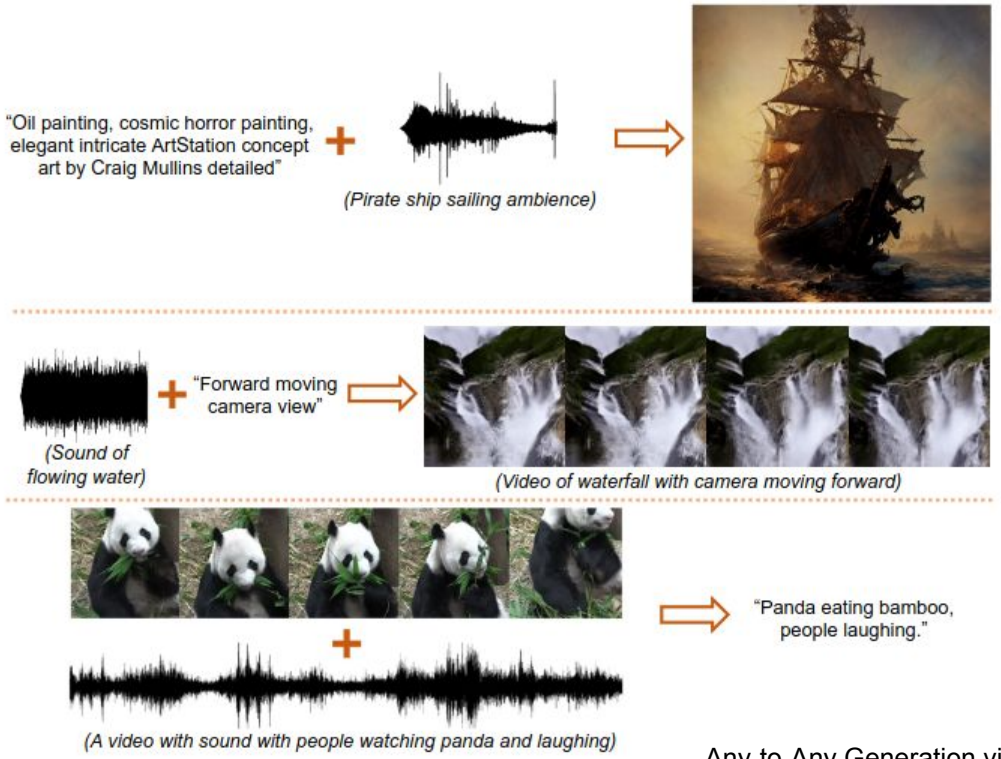
Coordinate the latent vectors

- example : learn a captioning model (create description of images)



How to learn from Multimodality ?

Coordinate the latent vectors

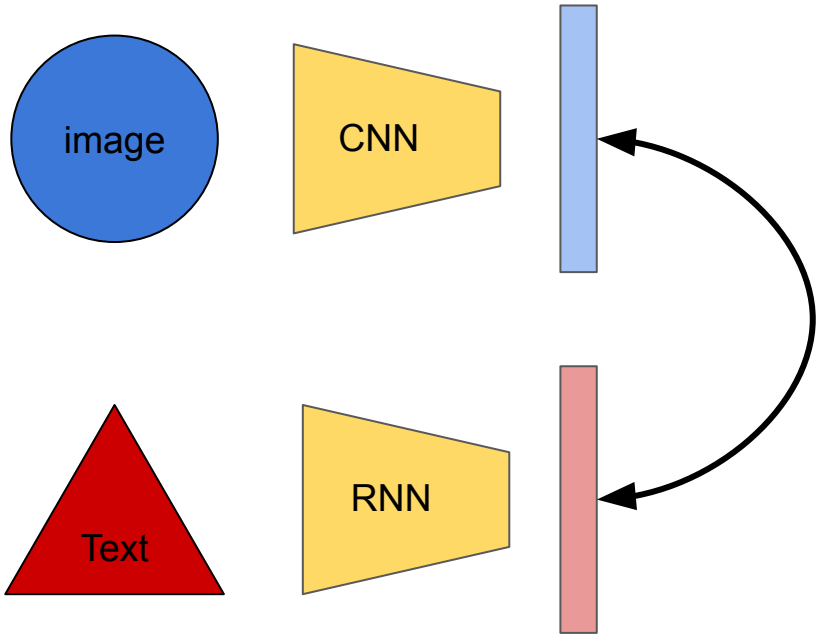


Any-to-Any Generation via Composable Diffusion; Tang & al.

How to learn from Multimodality ?

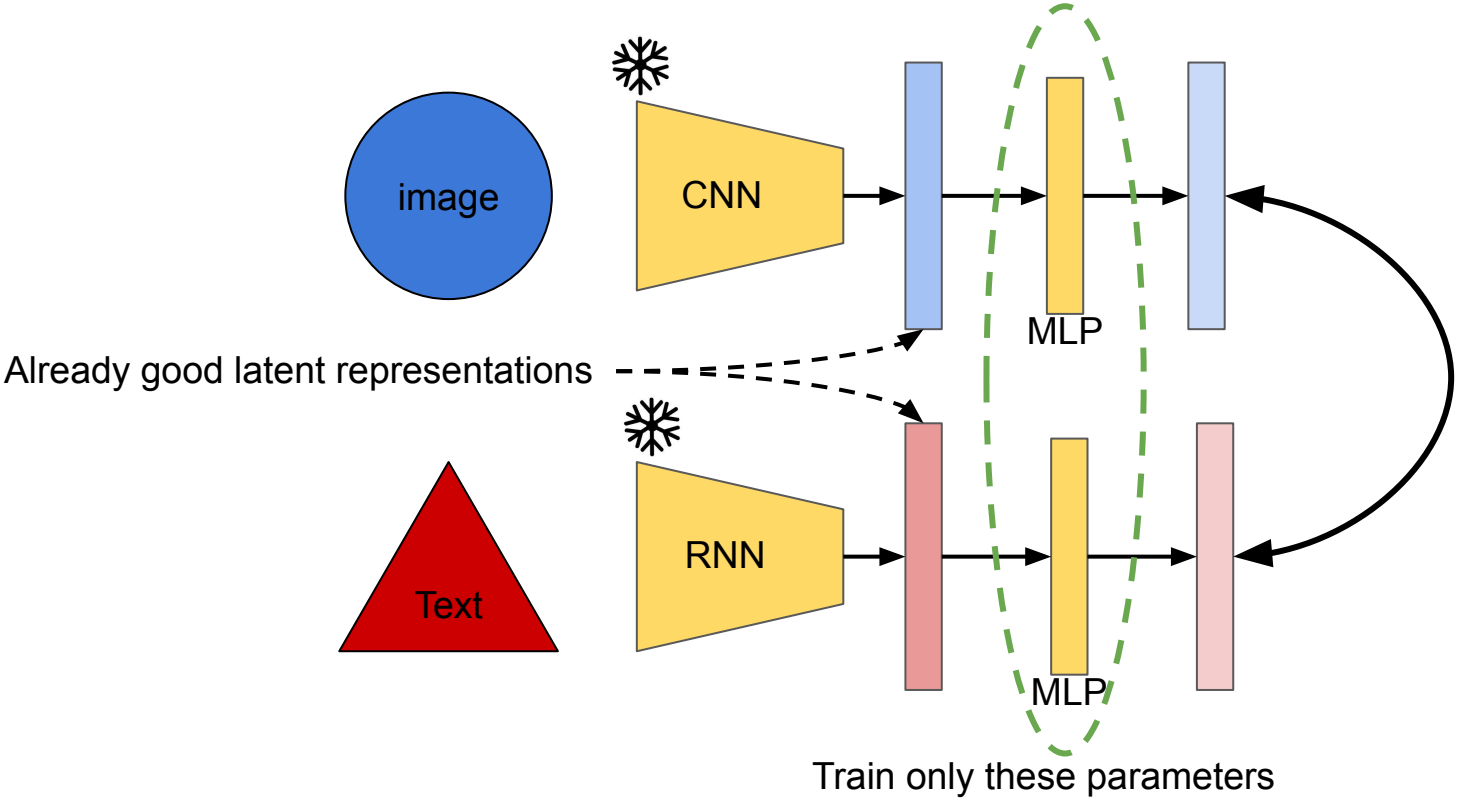
Learn directly the fusion or the coordination with the encoders:

- complex with a lot of parameters to train



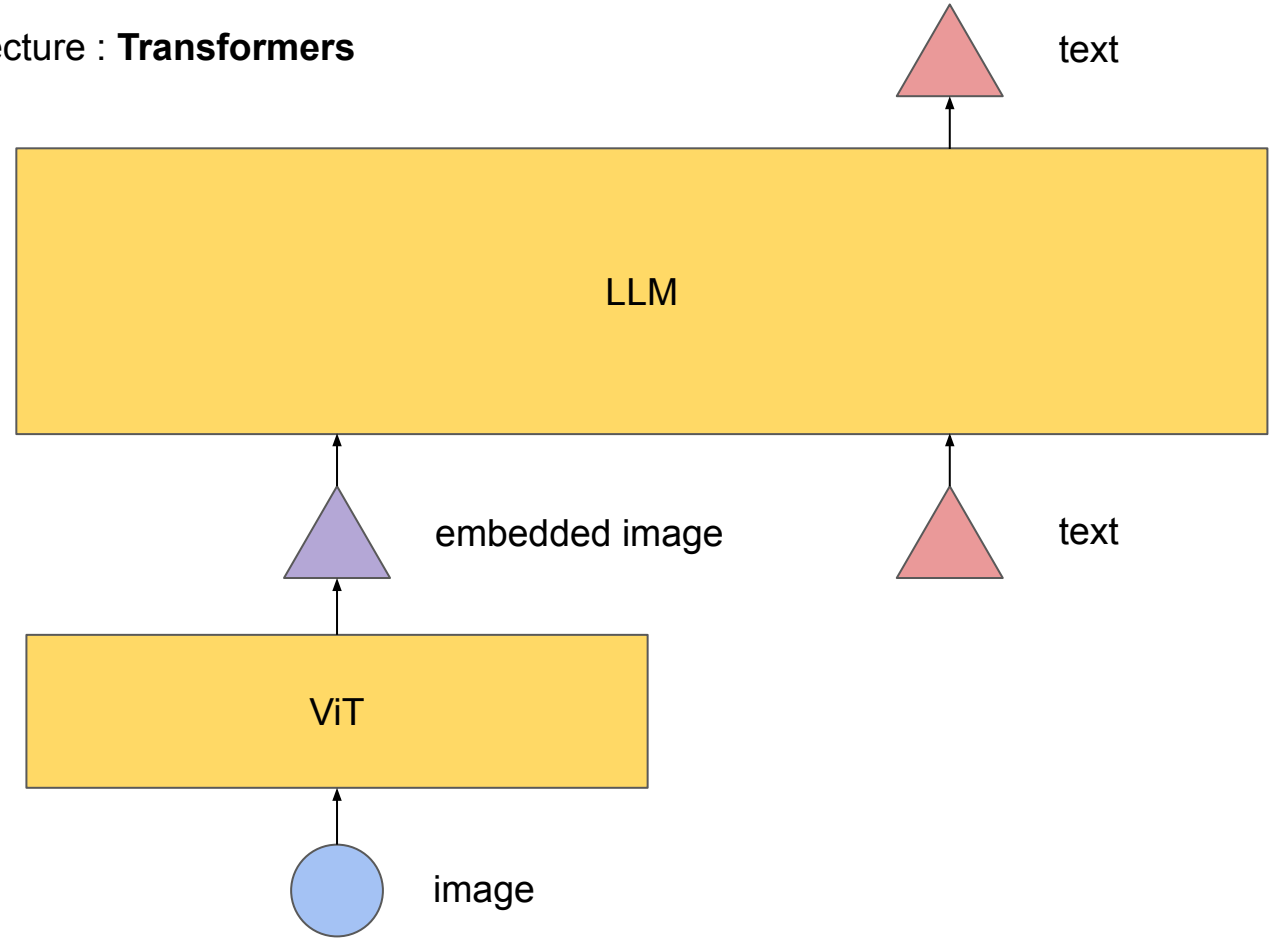
How to learn from Multimodality ?

Use pre-trained models to encode different modalities :



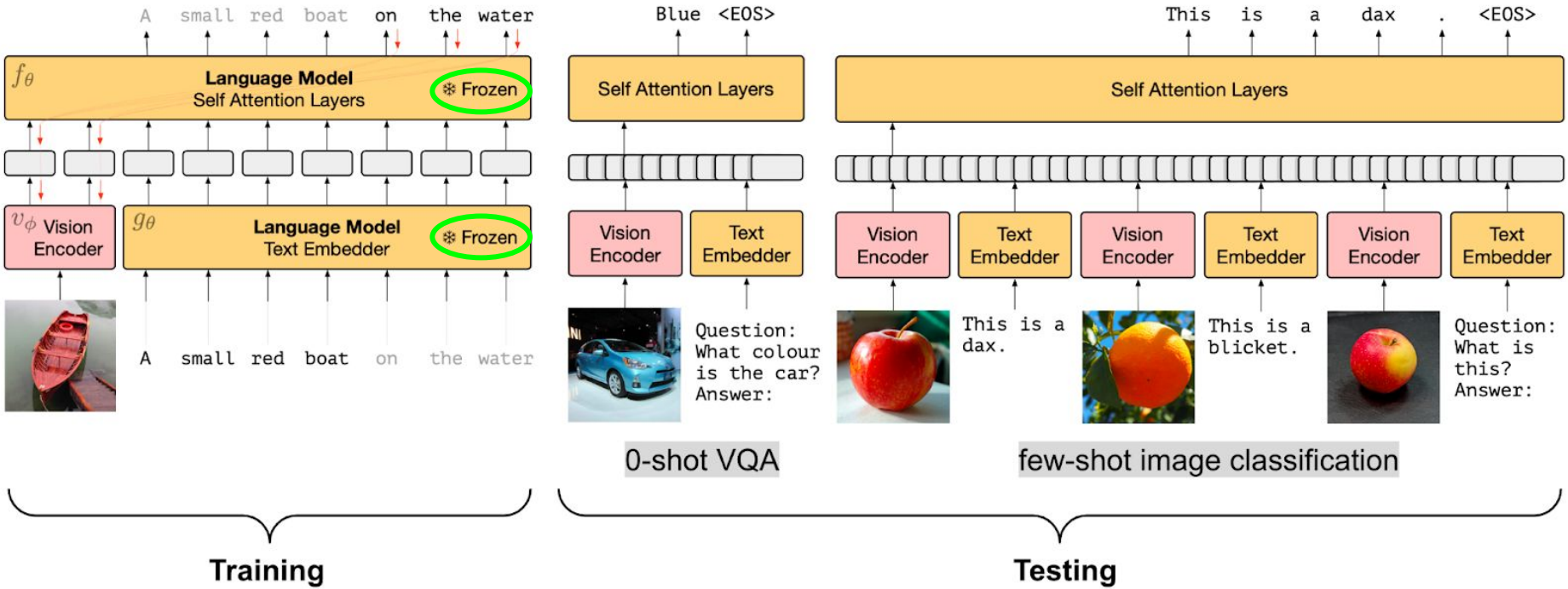
How to learn from Multimodality ?

Popular architecture : **Transformers**



How to learn from Multimodality ?

Popular architecture : **Transformers**

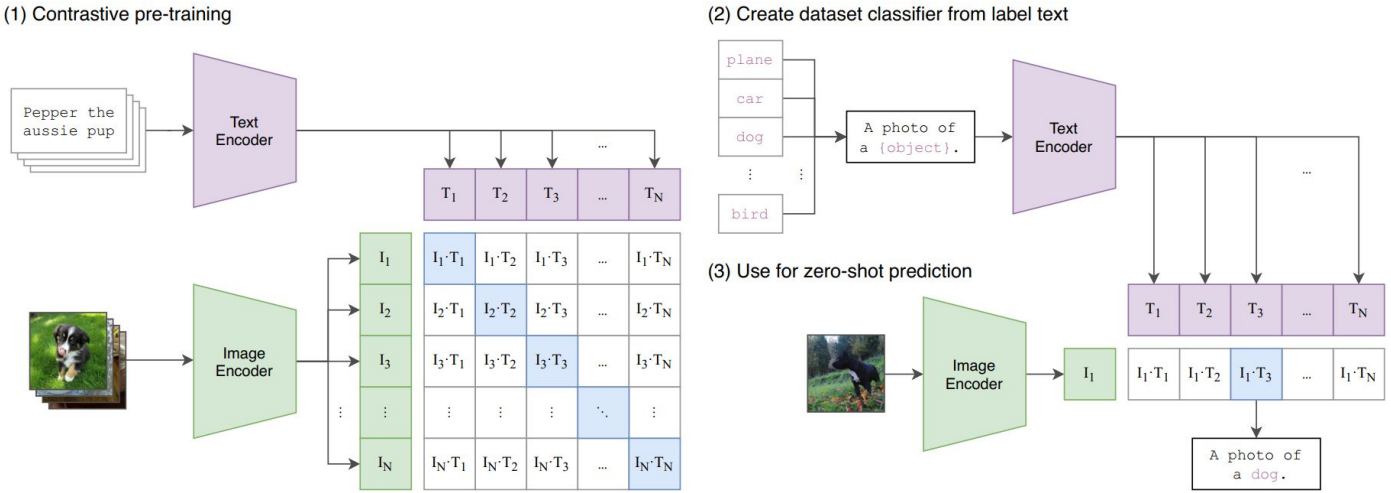


Multimodal Few Shot Learning with Frozen Language Model, Tsimpoukelli & al.

How to learn from Multimodality ?

Popular architecture : **Foundation model**

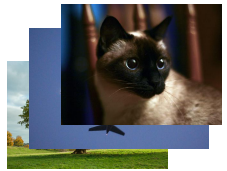
CLIP : trained to encode image and text in aligned latent space



How to learn from Multimodality ?

Popular architecture : **Foundation model**

example : image retrieval



CLIP images

set of images encoded by CLIP

the underside of a large airplane in the sky.

CLIP text

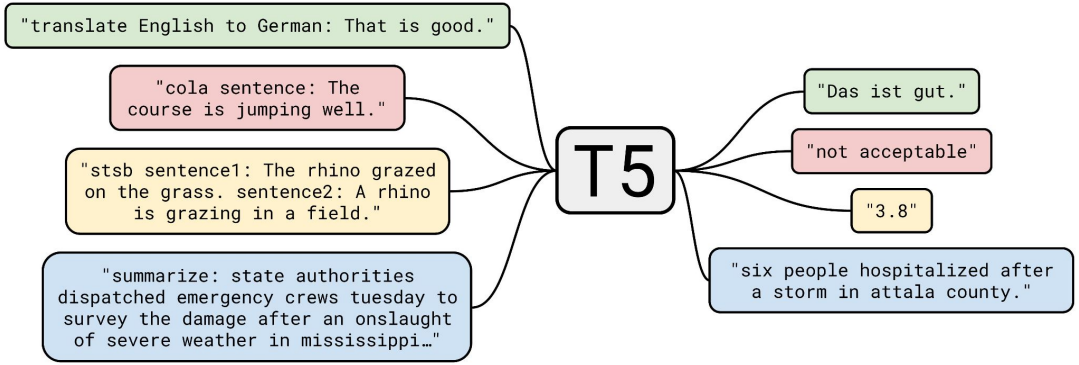
KNN

extract nearest neighbor

How to learn from Multimodality ?

Popular architecture : **Foundation model**

Large Language Model (LLM) : have a huge knowledge of the world and capacity of reasoning due to the quantity of data it saw during training, ex : BERT, GPT, T5, Palm, …

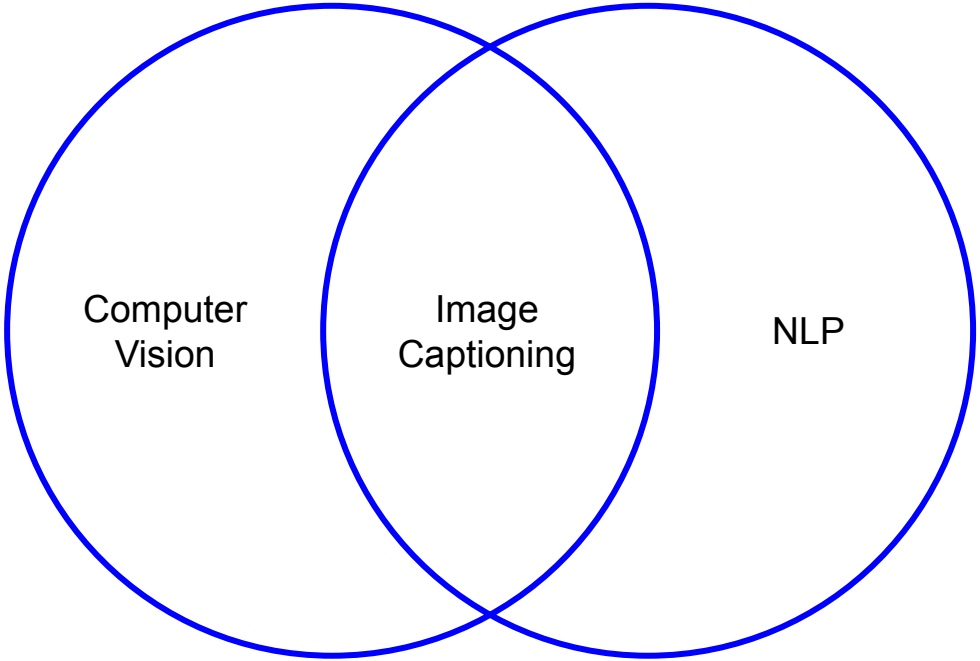


Multimodal Tasks

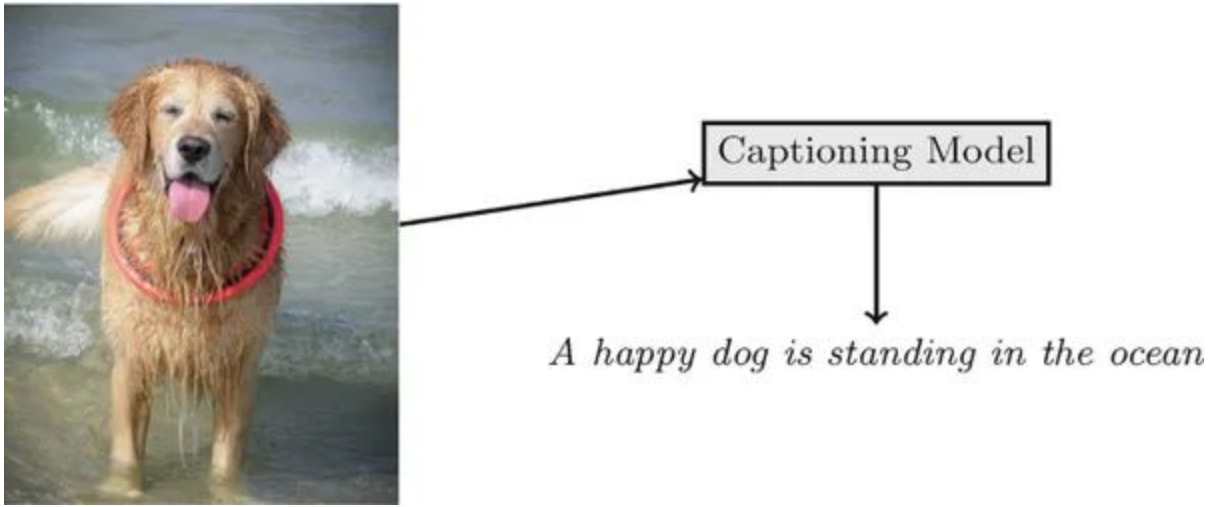
Multimodal Tasks - Image Captioning

Image Captioning is the process of generating textual description of an image

- Input : image
- Output : text



Multimodal Tasks - Image Captioning



Multimodal Tasks - Image Captioning

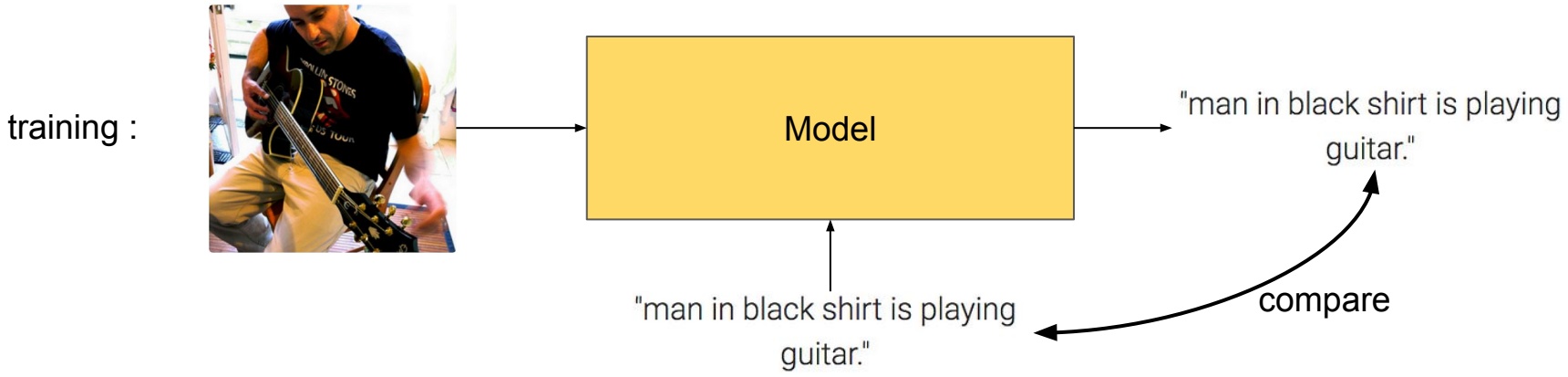
Need textual and visual data for training

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

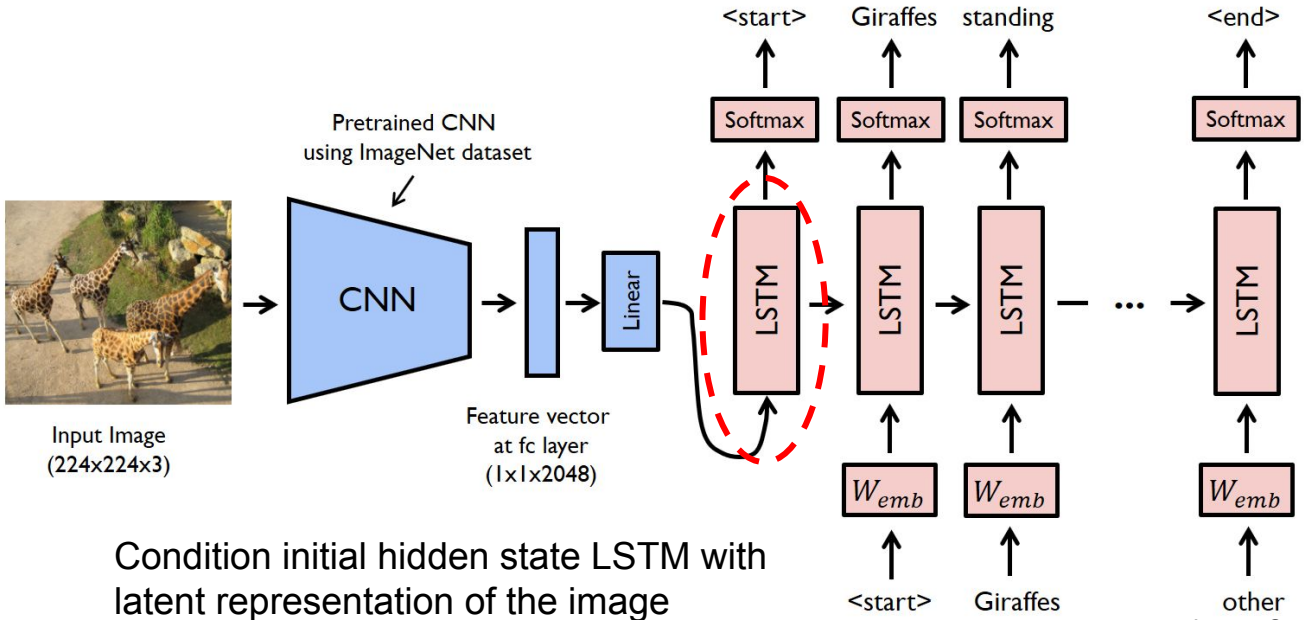
"two young girls are playing with lego toy."

Multimodal Tasks - Image Captioning



Multimodal Tasks - Image Captioning

Example of old architecture :



Condition initial hidden state LSTM with latent representation of the image

Image Captioning Using Deep Convolutional Neural Networks (CNNs), Gunasekar & al.

Turn each architecture to good account by using CNN for images (spatial data) and LSTM for text (sequential data)

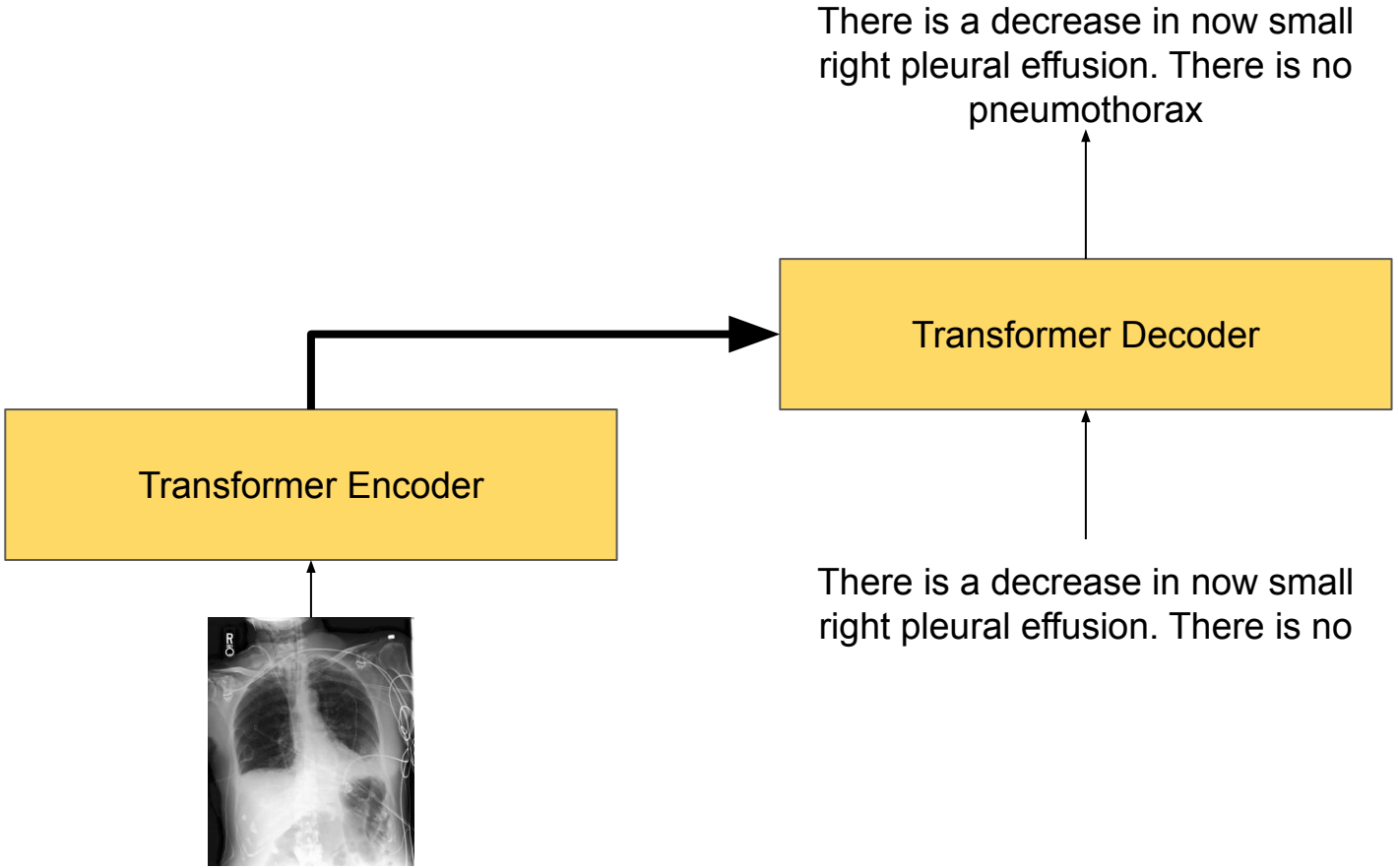
Multimodal Tasks - Image Captioning

Chest X-Ray	Ground Truth	Our predictions
	Lungs remain well inflated without evidence of focal airspace consolidation, pleural effusions, pulmonary edema or pneumothorax.Irregularity in the right humeral neck is related to a known healing fracture secondary to recent fall. PA and lateral views of the chest ____ at 09:55 are submitted.	no findings. no pneumonia. no pleural effusion. no edema. there is little change and no evidence of acute cardiopulmonary disease. no pneumonia, vascular congestion, pleural effusion.of incidental note is an azygos fissure, of no clinical significance. this raises possibility of a normal variant.
	1. Stable bilateral small pleural effusions and atelectasis. 2. Enlarged pulmonary artery, suggesting pulmonary hypertension. No significant interval change. Bilateral small pleural effusions and adjacent atelectasis are overall unchanged. The heart is top-normal in size, unchanged. The pulmonary artery is enlarged, suggesting pulmonary hypertension. No demand, focal consolidation to suggest pneumonia, or pneumothorax.	pleural effusion present. lung opacity present. no edema. cardiomegaly present. atelectasis present. as compared to previous radiograph, there is an increase in extent of a pre existing small left pleural effusion with subsequent atelectasis at left lung bases. otherwise, radiograph is unchanged. moderate cardiomegaly. mild fluid overload no overt pulmonary edema. no new focal parenchymal opacities suggesting pneumonia. unchanged position of right pectoral port a cath.
	There is decrease in now small right pleural effusion. There is no pneumothorax. There is a new right pacer pigtail catheter. Cardiomediastinal contours are unchanged. Lines and tubes are in standard position. Left lower lobe opacities, a combination of pleural effusion and atelectasis, are unchanged.	uncertain pneumonia. pleural effusion present. lung opacity present. atelectasis present. bilateral pleural effusions, left greater than right. bibasilar opacities potentially atelectasis in setting of low lung volumes. infection be excluded. frontal and lateral views of chest demonstrate low lung volumes, which accentuate bronchovascular markings. there are small bilateral pleural effusions, right greater than left, with adjacent atelectasis. there is no focal consolidation pneumothorax. cardiomediastinal silhouette is within normal limits. surgical clips are seen in right upper quadrant of abdomen. aortic arch calcifications are noted.

Medical Image Captioning via Generative Pretrained Transformers, Selivanov & al.

Multimodal Tasks - Image Captioning

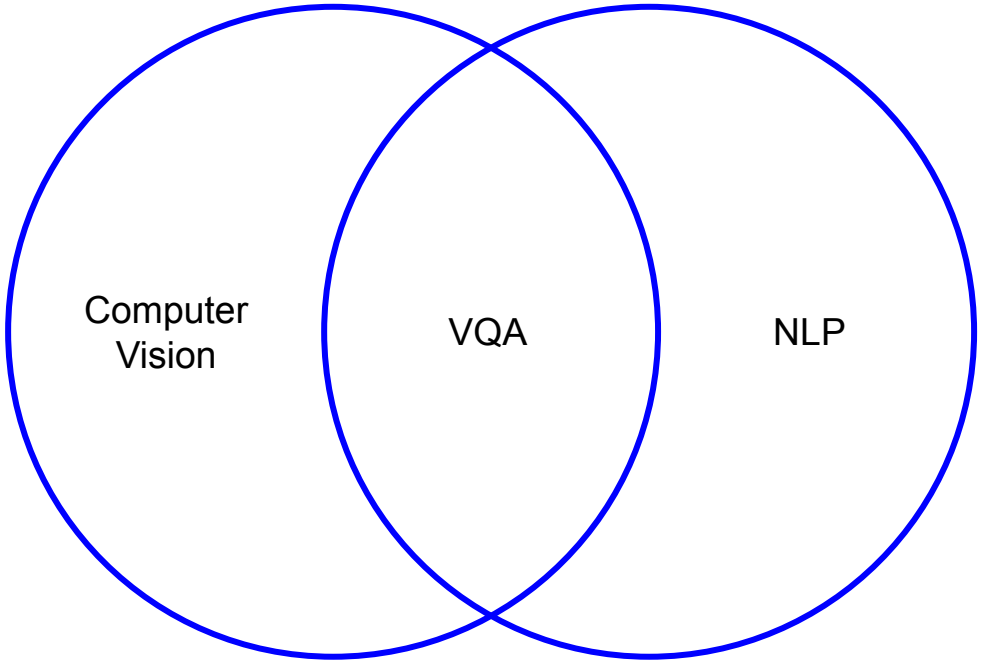
Architecture mostly used : **Transformers**



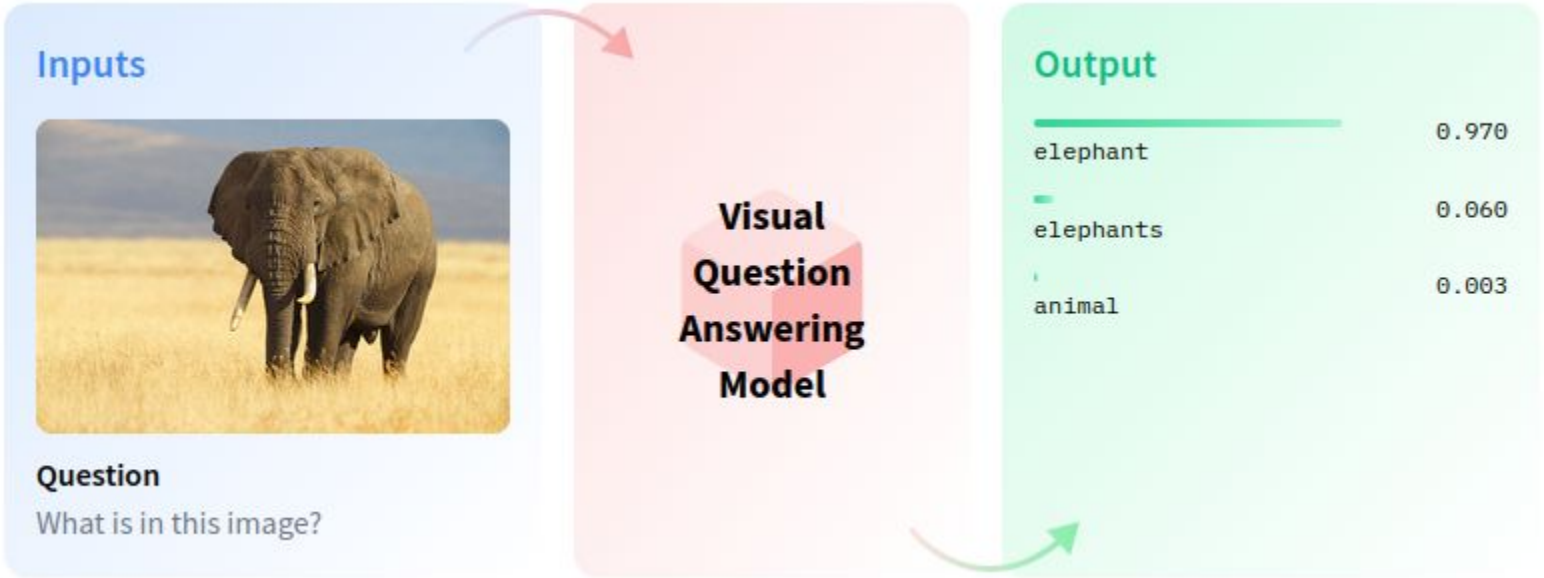
Multimodal Tasks - VQA

Visual Question Answering is the task of answering open-ended questions based on an image

- Input : text + image
- Output : text



Multimodal Tasks - VQA



Multimodal Tasks - VQA

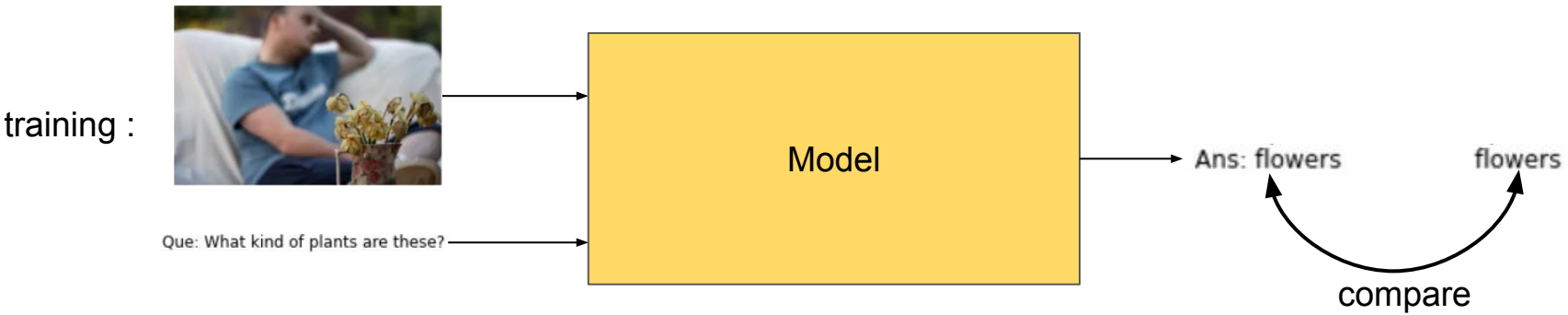
Need textual and visual data for training

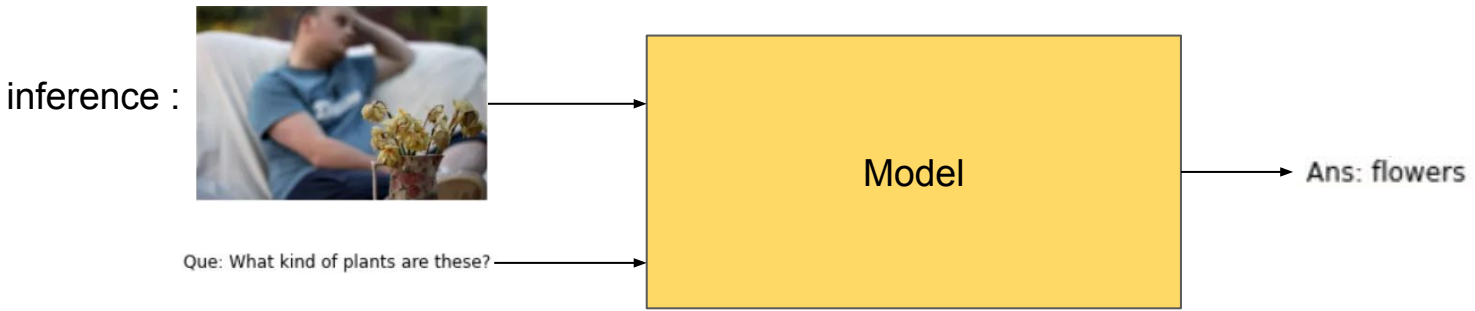
Que: What kind of plants are these?
Ans: flowers

Que: Is the kite high in the air?
Ans: yes

Que: Where is the light coming from?
Ans: window

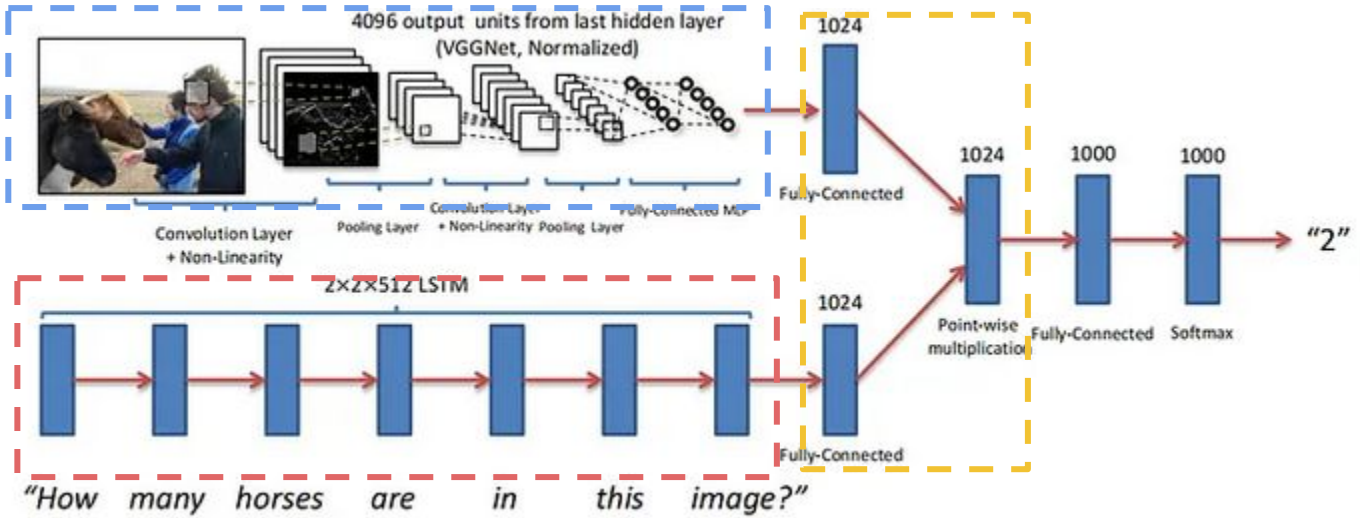
Multimodal Tasks - Image Captioning





Multimodal Tasks - VQA

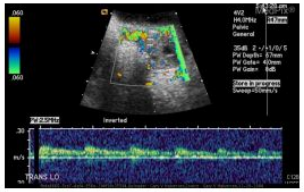
Example of old architecture :



Fusion approach to treat multimodal data : Specialized architecture to encode text and images with linear layers and pointwise multiplication to merge the latent representations

Multimodal Tasks - VQA

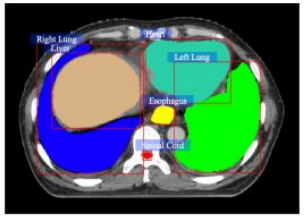
VQA-Med-2020 [13]



Q: what abnormality is seen in the image?
A: ovarian torsion

Q: what is abnormal in the ct scan?
A: partial anomalous pulmonary venous return

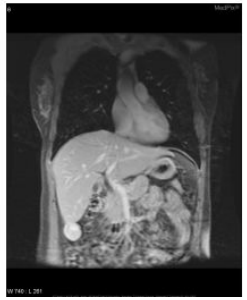
SLAKE [57]



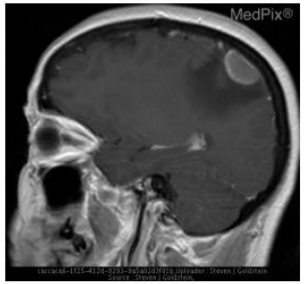
Q: Does the image contain left lung?
A: Yes

Q: What is the function of the rightmost organ in this picture?
A: Breathe

VQA-Med-2021 [15]



Q: What is most alarming about this mri?
A: focal nodular hyperplasia

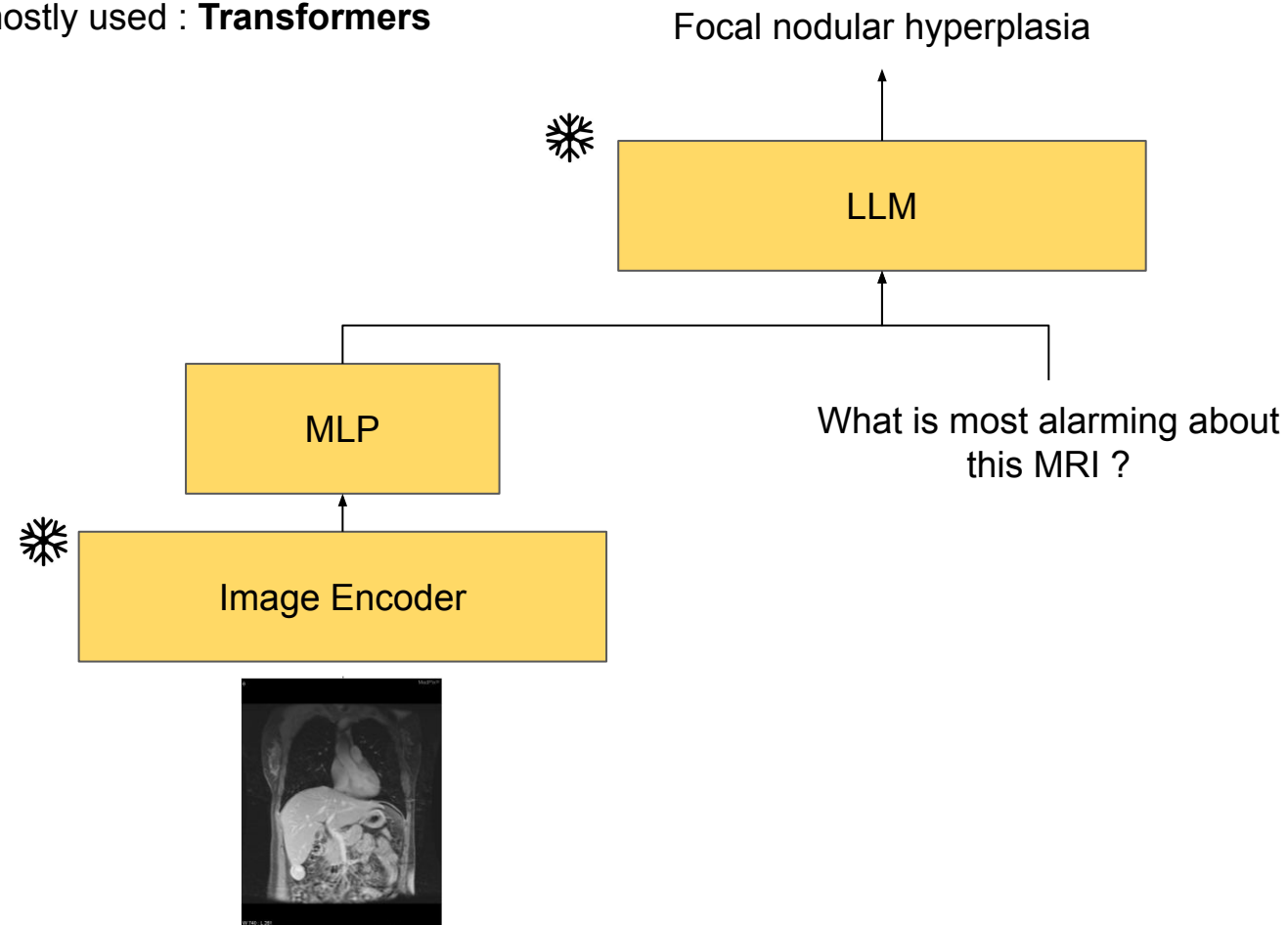


Q: What abnormality is seen in the image?
A: Enhancing lesion right parietal lobe with surrounding edema

Medical Visual Question Answering: A Survey, Lin & al.

Multimodal Tasks - VQA

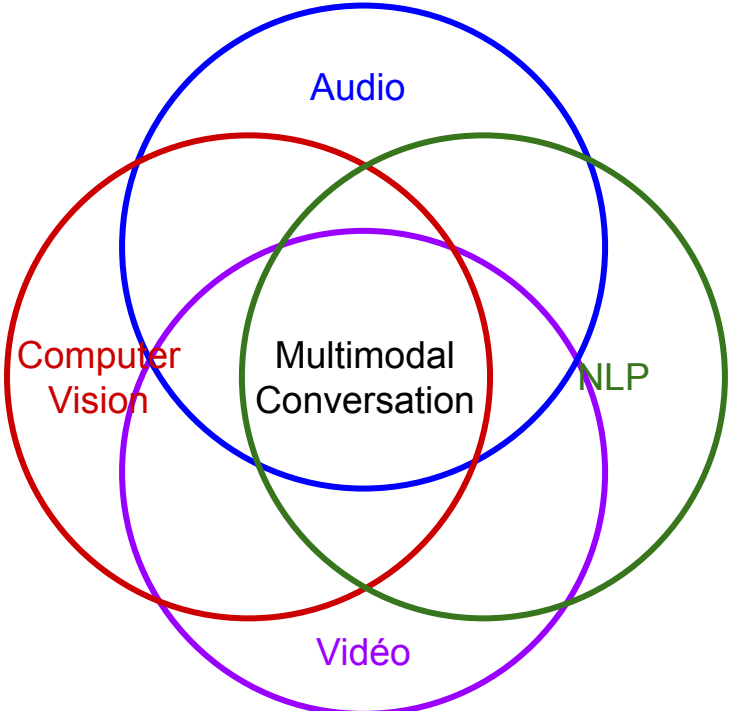
Architecture mostly used : **Transformers**



Multimodal Tasks - Multimodal Conversational System

A **multimodal conversational AI system** answers questions, fulfills tasks, and emulates human conversations by understanding and expressing itself via multiple modalities.

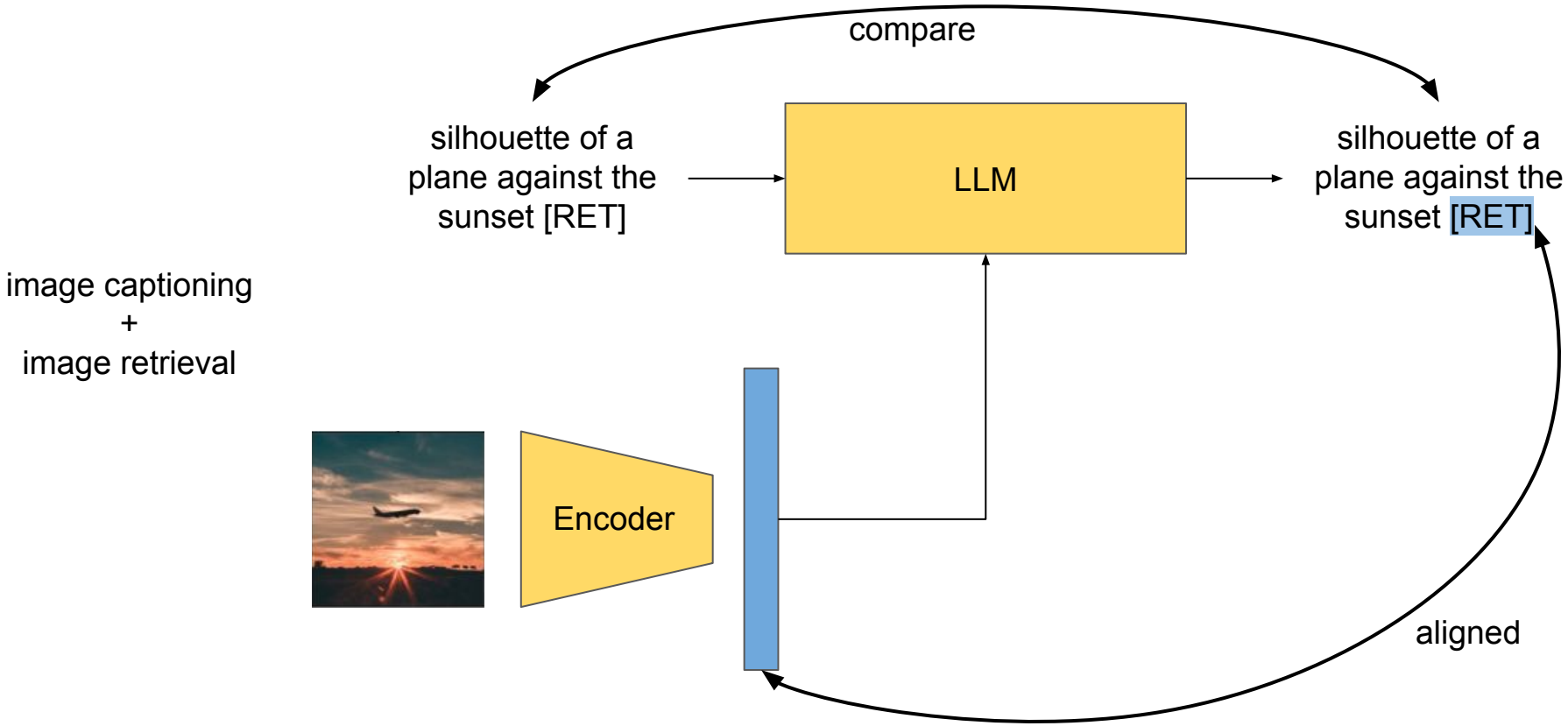
- Input : text, image, video, audio, …
- Output : text, image, video, audio, …



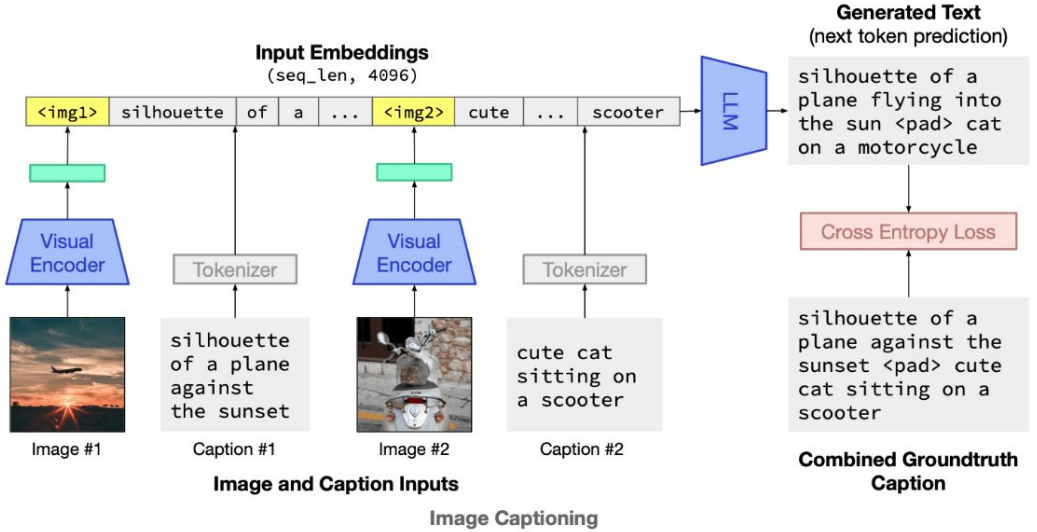
Multimodal Tasks - Multimodal Conversational System

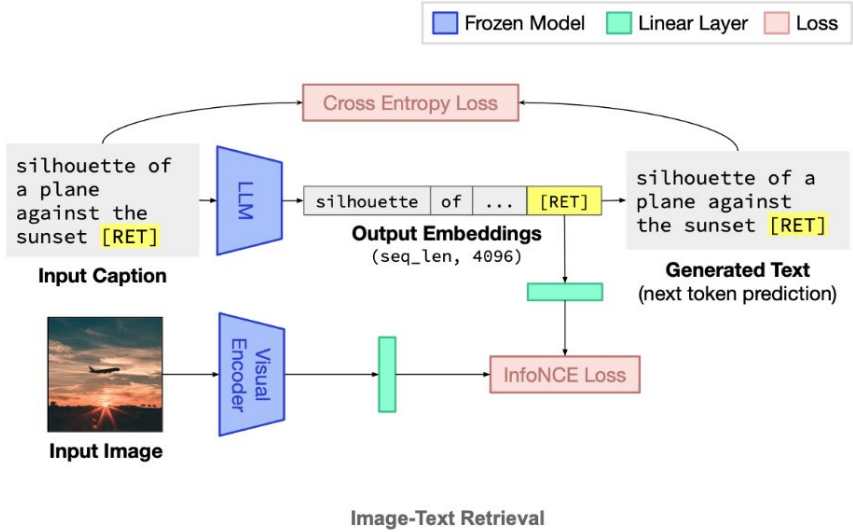
Grounding Language Models to Images for Multimodal Inputs and Outputs; Koh, Jing Yu; Salakhutdinov; Ruslan and Fried, Daniel

Multimodal Tasks - Multimodal Conversational System



Multimodal Tasks - Multimodal Conversational System





Grounding Language Models to Images for Multimodal Inputs and Outputs; Koh, Jing Yu; Salakhutdinov; Ruslan and Fried, Daniel

Multimodal Tasks - Multimodal Conversational System

Instructions: You are a helpful radiology assistant. Describe what lines, tubes, and devices are present and each of their locations. Describe if pneumothorax is present; if present, describe size on each side. Describe if pleural effusion is present; if present, describe amount on each side. Describe if lung opacity (atelectasis, fibrosis, consolidation, infiltrate, lung mass, pneumonia, pulmonary edema) is present; if present, describe kinds and locations. Describe the cardiac silhouette size. Describe the width and contours of the mediastinum. Describe if hilar enlargement is present; if enlarged, describe side. Describe what fractures or other skeletal abnormalities are present.

Given the LATERAL view X-ray image <img>. **Reason for the study:** Amiodarone routine surveillance.
Q: Describe the findings in the image following the instructions.
A: As compared to the previous radiograph there is no relevant change. Normal lung volumes. Mild bilateral apical scarring. Normal size of the cardiac silhouette and tortuosity of the thoracic aorta. No pathologic findings in the lung parenchyma notably no evidence of fibrotic lung parenchymal changes. A faint 2 mm rounded opacity projecting over the lower aspect of the fourth right rib and internally to the upper border of the second right rib is seen on the frontal radiograph only and likely reflects structure on the skin.

Given the PA view X-ray image <img>. **Reason for the study:** History m with malaise pneumonia.
Q: Describe the findings in the image following the instructions.
A:

Instructions: You are a helpful dermatology assistant. The following are questions about skin lesions. Categorize the skin lesions into the most likely class given the patient history.

Given <img>. **Patient History:** Age: 51, Gender: female, Smoke: false, Drink: false, Family skin cancer history: true, Family any cancer history: false, Lesion region: back, Lesion itch: false, Lesion grew: false, Lesion bled: false, Lesion elevation: false, Fitzpatrick scale: 1.0, Diameters (mm): [12.0, 8.0].
Q: Which of the following is the most likely diagnosis of the patient's skin lesion?
(A) Nevus (B) Basal Cell Carcinoma (C) Squamous Cell Carcinoma (D) Actinic Keratosis (E) Seborrheic Keratosis (F) Melanoma
A: Basal Cell Carcinoma.

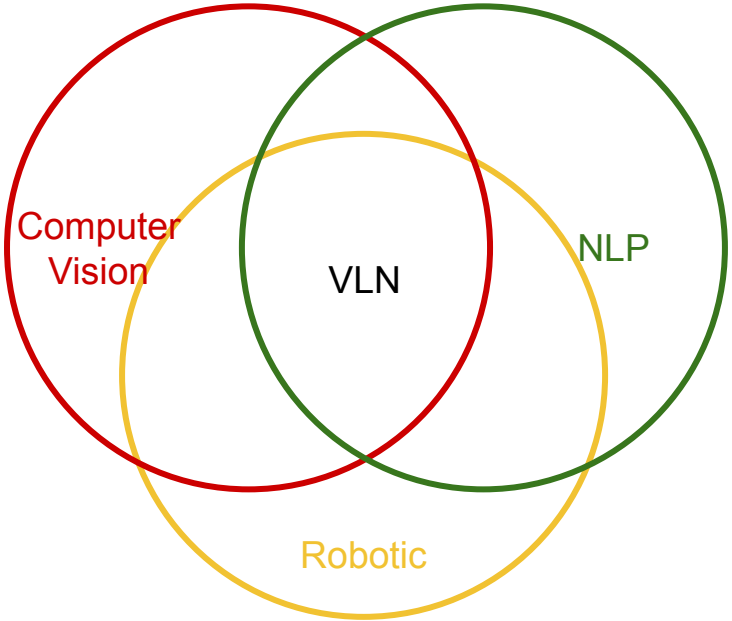
Given <img>. **Patient History:** Age: 39, Gender: unknown, Smoke: unknown, Drink: unknown, Family skin cancer history: unknown, Family any cancer history: unknown, Lesion region: neck, Lesion itch: false, Lesion grew: true, Lesion bled: false, Lesion elevation: true, Fitzpatrick scale: unknown, Diameters (mm): [unknown, unknown].
Q: Which of the following is the most likely diagnosis of the patient's skin lesion?
(A) Nevus (B) Basal Cell Carcinoma (C) Squamous Cell Carcinoma (D) Actinic Keratosis (E) Seborrheic Keratosis (F) Melanoma
A:

Towards Generalist Biomedical AI, Tu & al.

Multimodal Tasks

Vision-and-Language Navigation aims to build an embodied agent that can communicate with humans in natural language and navigate in real 3D environments

- Input : text, image
- Output : text, image, actions



Multimodal Tasks

Pass the pool and go indoors using the double glass doors. Pass the large table with chairs and turn left and wait by the wine bottles that have grapes by them.

Walk straight through the room and exit out the door on the left. Keep going past the large table and turn left. Walk down the hallway and stop when you reach the 2 entry ways. One in front of you and one to your right. The bar area is to your left.

Enter house through double doors, continue straight across dining room, turn left into bar and stop on the circle on the ground.

Standing in front of the family picture, turn left and walk straight through the bathroom past the tub and mirrors. Go through the doorway and stop when the door to the bathroom is on your right and the door to the closet is to your left.

Walk with the family photo on your right. Continue straight into the bathroom. Walk past the bathtub. Stop in the hall between the bathroom and toilet doorways.

Walk straight passed bathtub and stop with closet on the left and toilet on the right.

Exit the office then turn left and then turn left in the hallway and head down the hallway until you get to a door on your left and go into office 359 then stop.

Go out of the room and take a left. Go into the first room on your left.

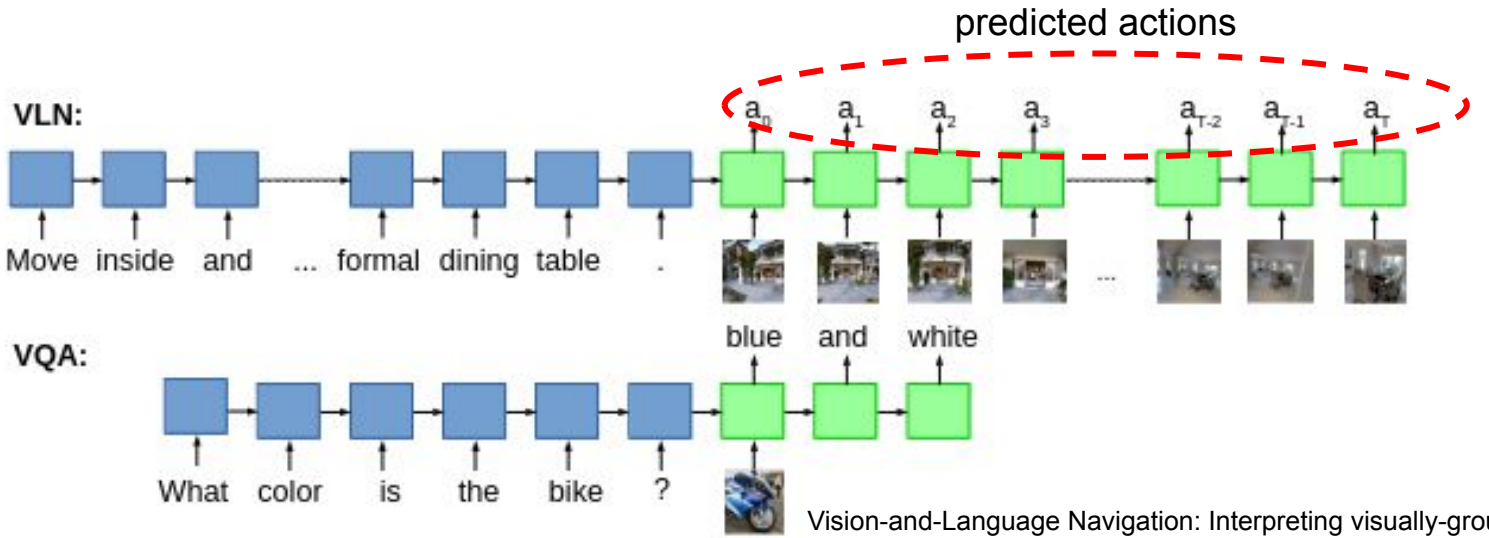
Leave the office and take a left. Take the next left at the hallway. Walk down the hall and enter the first office on the left. Stop next to the door to office 359.

Go up the stairs and turn right. Go past the bathroom and stop next to the bed.

Walk all the way up the stairs, and immediately turn right. Pass the bathroom on the left, and enter the bedroom that is right there, and stop there.

Walk up the stairs turn right at the top and walk through the doorway continue straight and stop inside the bedroom.

Multimodal Tasks



Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments; Anderson & al.

Use LSTM for text → Normal

Use LSTM for images →????? : Normal because images are correlated in time for this case : it's like a video

Multimodal Tasks

- Catch the coke can
- Go to the operator
- Place the can on the table

Controler

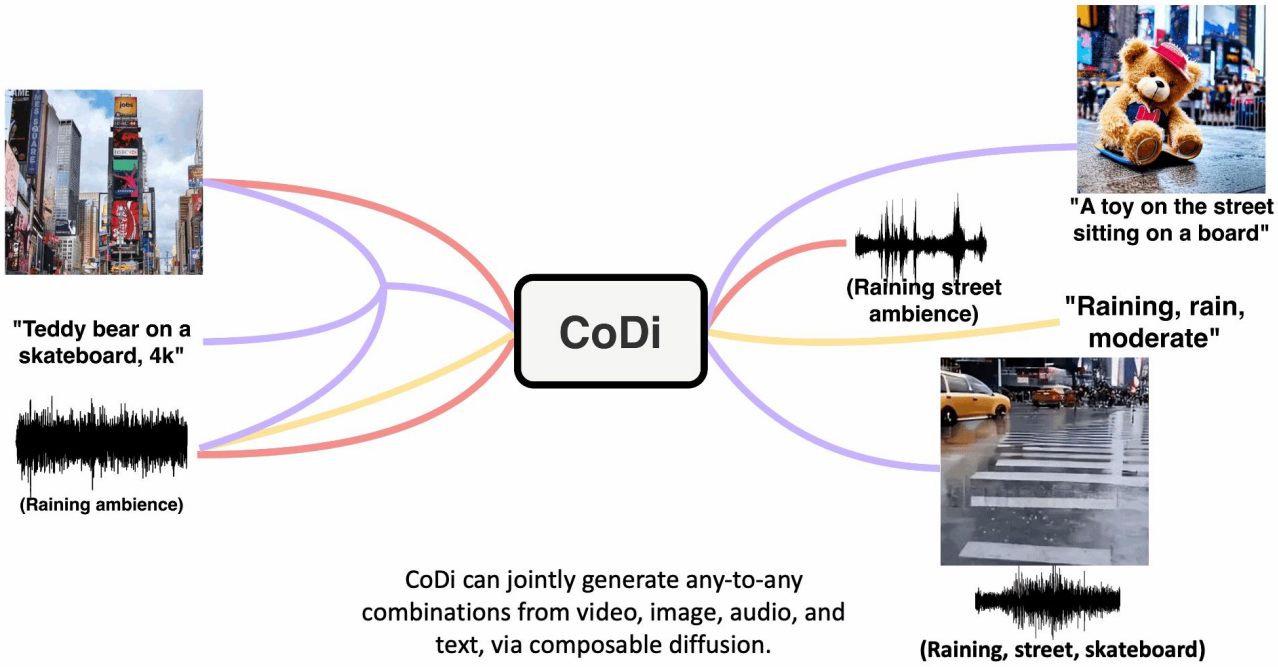
Planer

Bring me a coke from the table

Examples of Models

Examples of Models

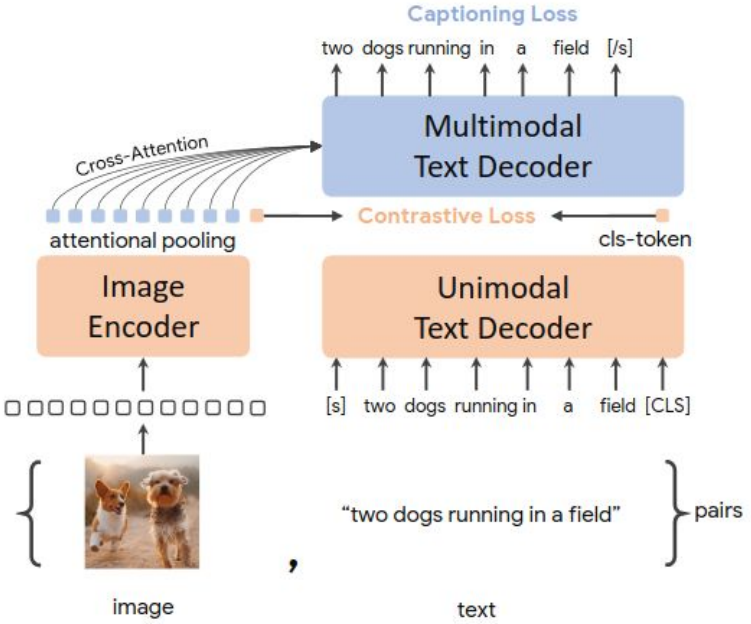
CoDi : Any to any generation via composable diffusion



idea : aligned all modalities in a latent space and use it to condition diffusion models (one for each modality)

Examples of Models

CoCa : Learn a contrastive captioner to be a image-text foundation model



Algorithm 1 Pseudocode of Contrastive Captioners architecture.

```
# image, text.ids, text.labels, text.mask: paired {image, text} data
# con_query: 1 query token for contrastive embedding
# cap_query: N query tokens for captioning embedding
# cls_token_id: a special cls_token_id in vocabulary

def attentional_pooling(features, query):
  out = multihead_attention(features, query)
  return layer_norm(out)

img_feature = vit_encoder(image) # [batch, seq_len, dim]
con_feature = attentional_pooling(img_feature, con_query) # [batch, 1, dim]
cap_feature = attentional_pooling(img_feature, cap_query) # [batch, N, dim]

ids = concat(text.ids, cls_token_id)
mask = concat(text.mask, zeros_like(cls_token_id)) # unpad cls_token_id
txt_embs = embedding_lookup(ids)
unimodal_out = lm_transformers(txt_embs, mask, cross_attn=None)
multimodal_out = lm_transformers(
    unimodal_out[:, :-1, :], mask, cross_attn=cap_feature)
cls_token_feature = layer_norm(unimodal_out)[:, -1:, :] # [batch, 1, dim]

con_loss = contrastive_loss(con_feature, cls_token_feature)
cap_loss = softmax_cross_entropy_loss(
    multimodal_out, labels=text.labels, mask=text.mask)
```

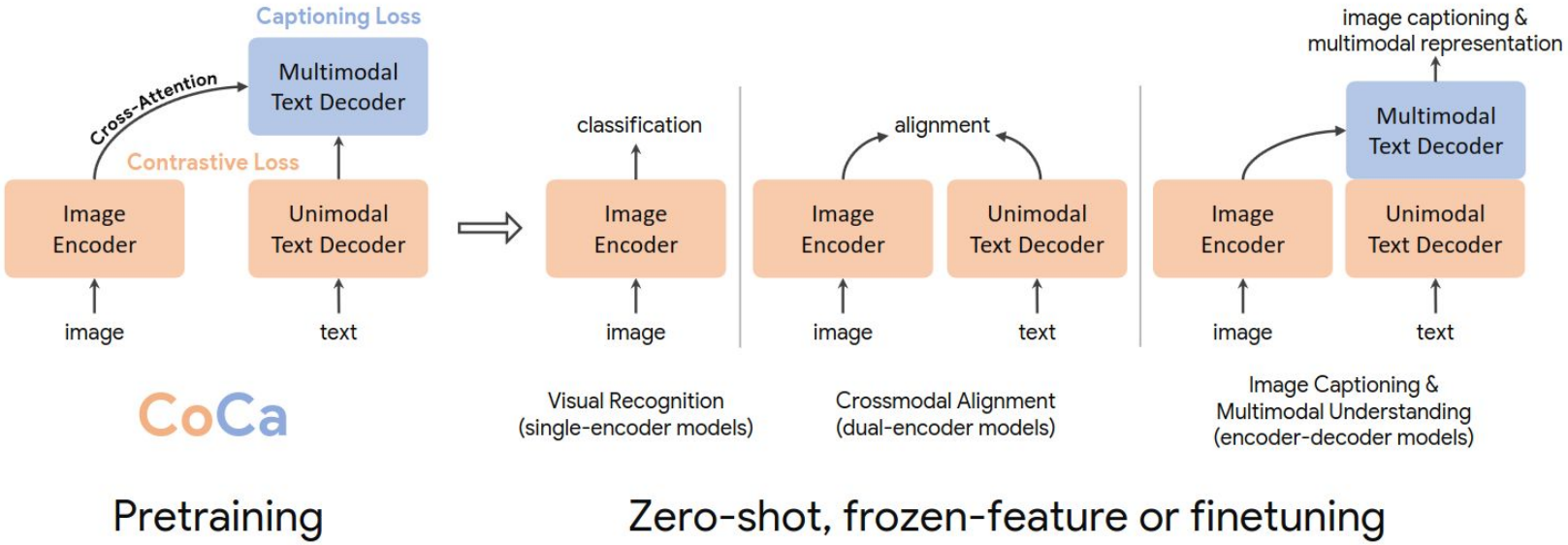
`vit_encoder`: vision transformer based encoder; `lm_transformer`: language-model transformers.

Figure 2: Detailed illustration of CoCa architecture and training objectives.

idea : Build a Foundation Model that do captioning and image-text alignment so that it keeps relevant information in the latent space.

Examples of Models

CoCa : Learn a contrastive captioner to be a image-text foundation model

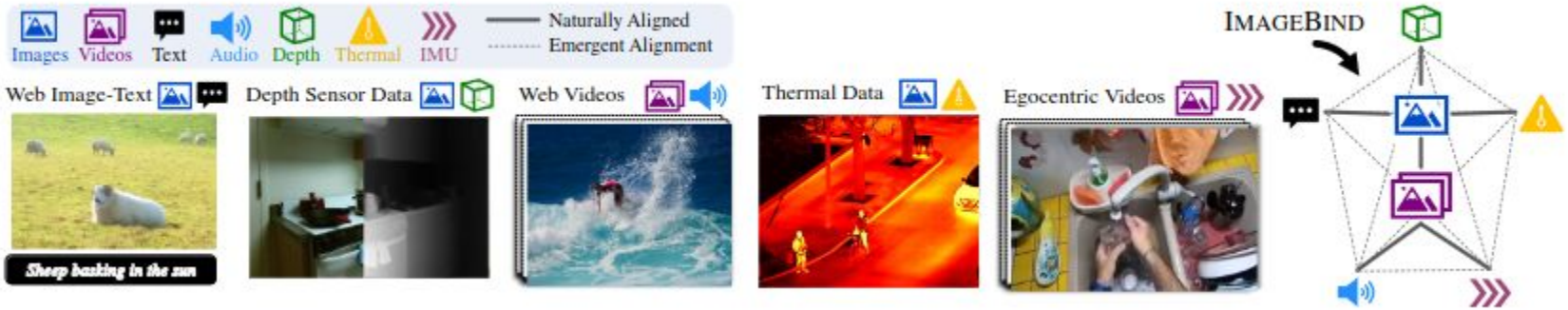


The goal is to use CoCa on :

- zero-shot : use CoCa on unseen tasks without adding any training
- frozen-feature : use CoCa on seen tasks but with unseen dataset
- finetuning : train a bit CoCa on a task or data that it has never seen

Examples of Models

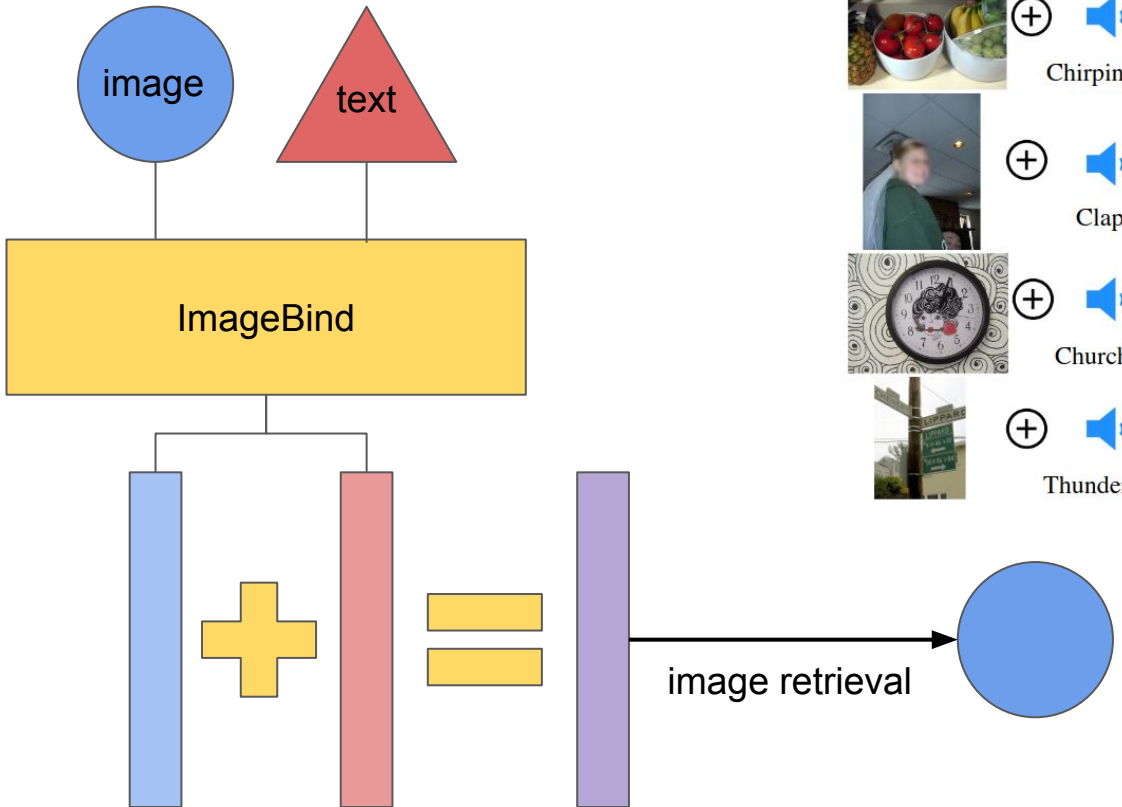
ImageBind : Multimodal alignment using transformer encoders

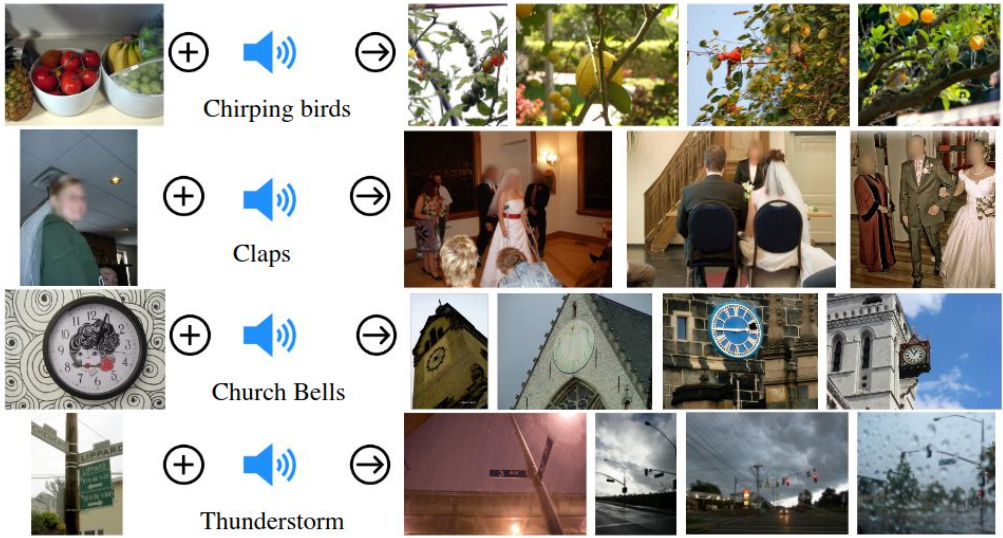


idea : use naturally aligned modalities (image-video, image-text, image-depth, video-audio, …) to align all the modalities

Examples of Models

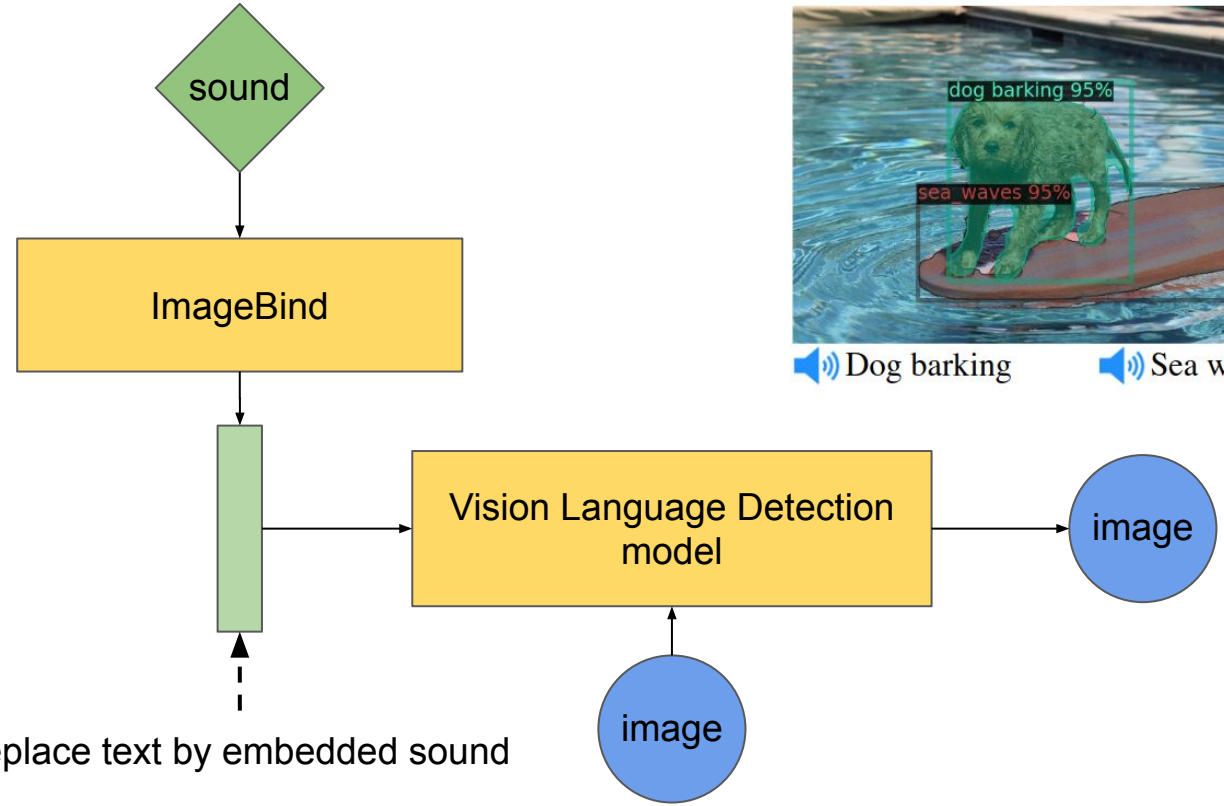
ImageBind : Multimodal alignment using transformer encoders

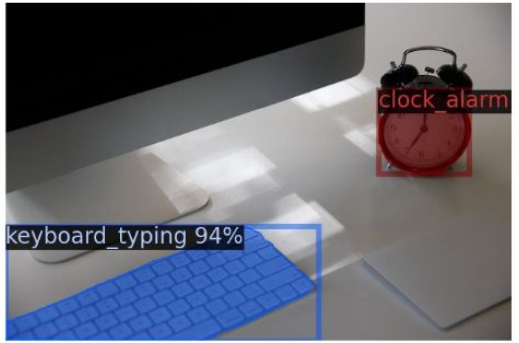




Examples of Models

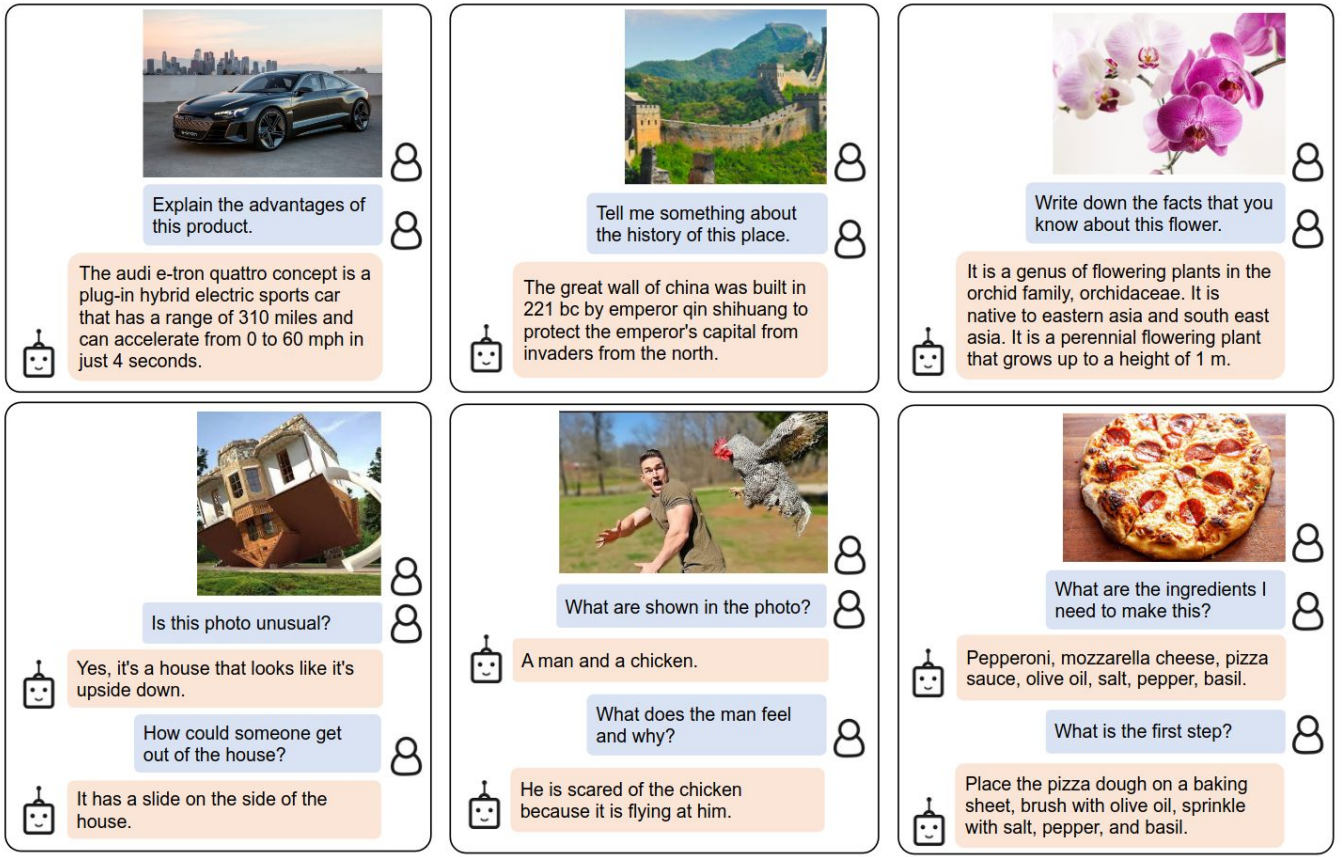
ImageBind : Multimodal alignment using transformer encoders





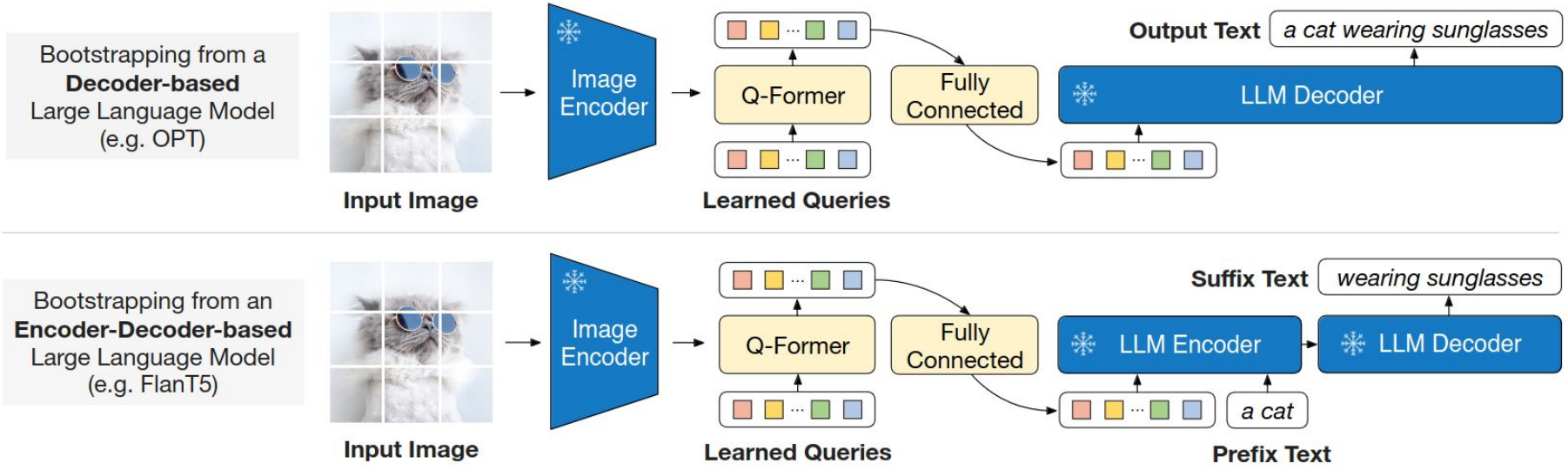
Examples of Models

BLIP-2 : Vision-Language Model using pre-trained models



Examples of Models

BLIP-2 : Vision-Language Model using pre-trained models

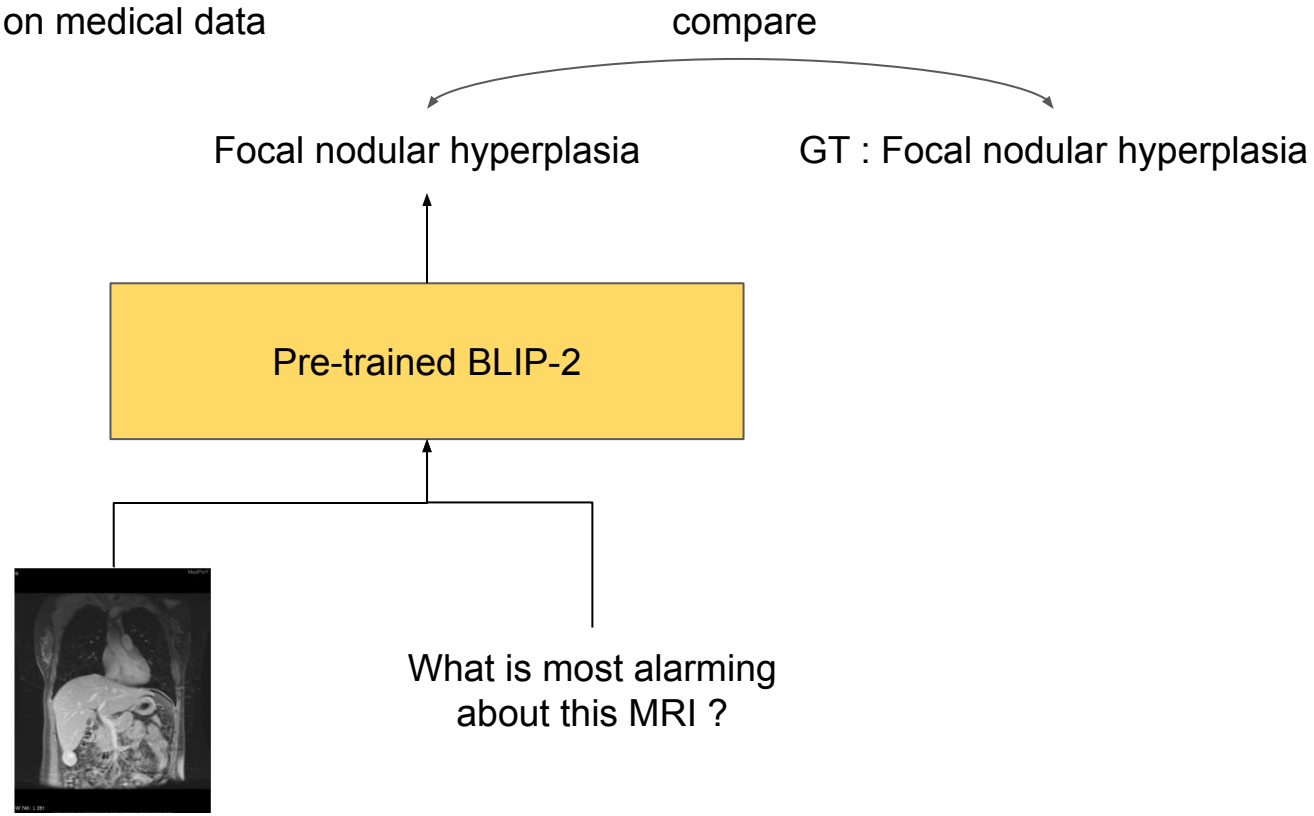


Q-Former : module train to bridge gap between frozen image encoder and a frozen LLM

Examples of Models

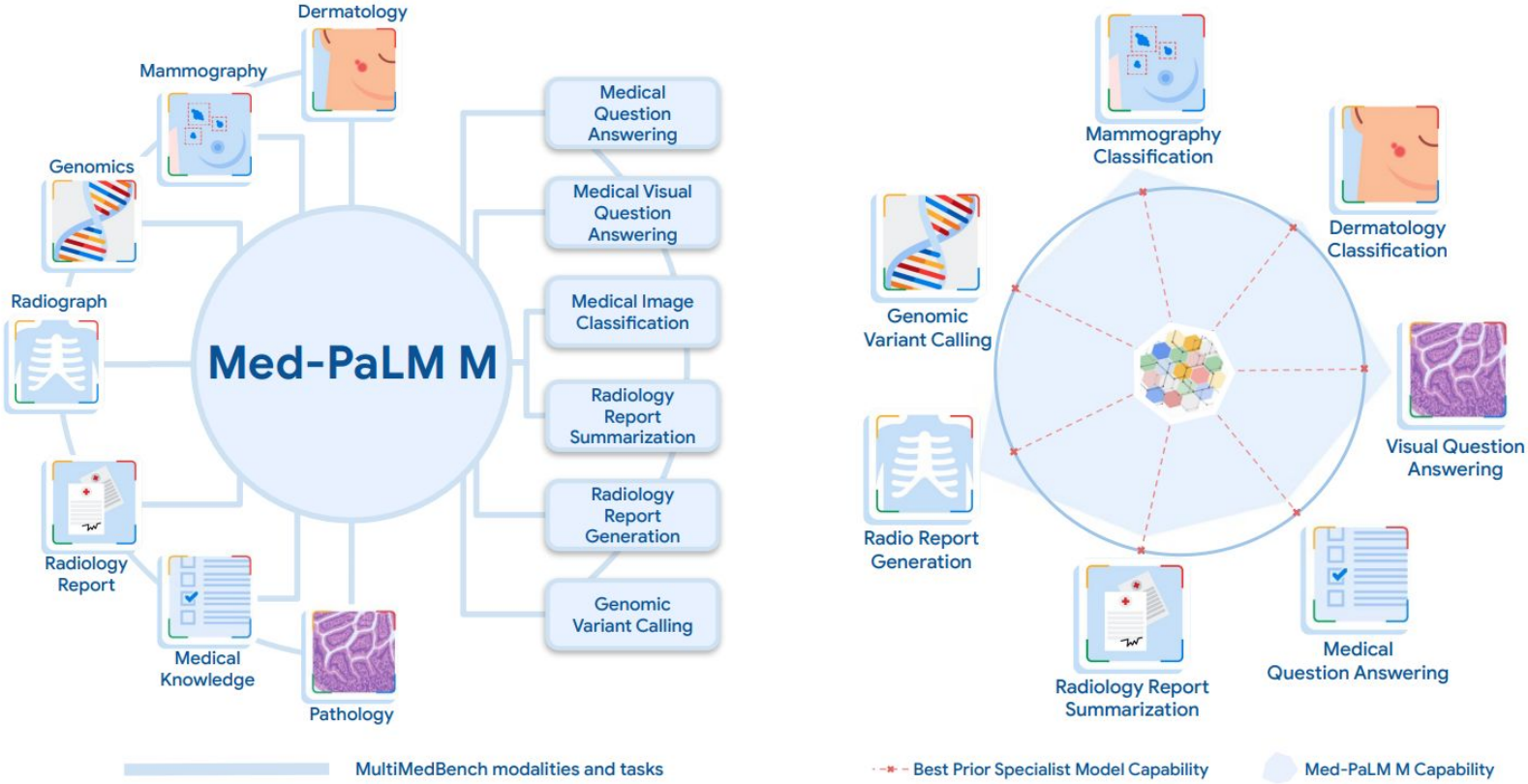
BLIP-2 : Vision-Language Model using pre-trained models

Fine-Tune the model on medical data
(done here)



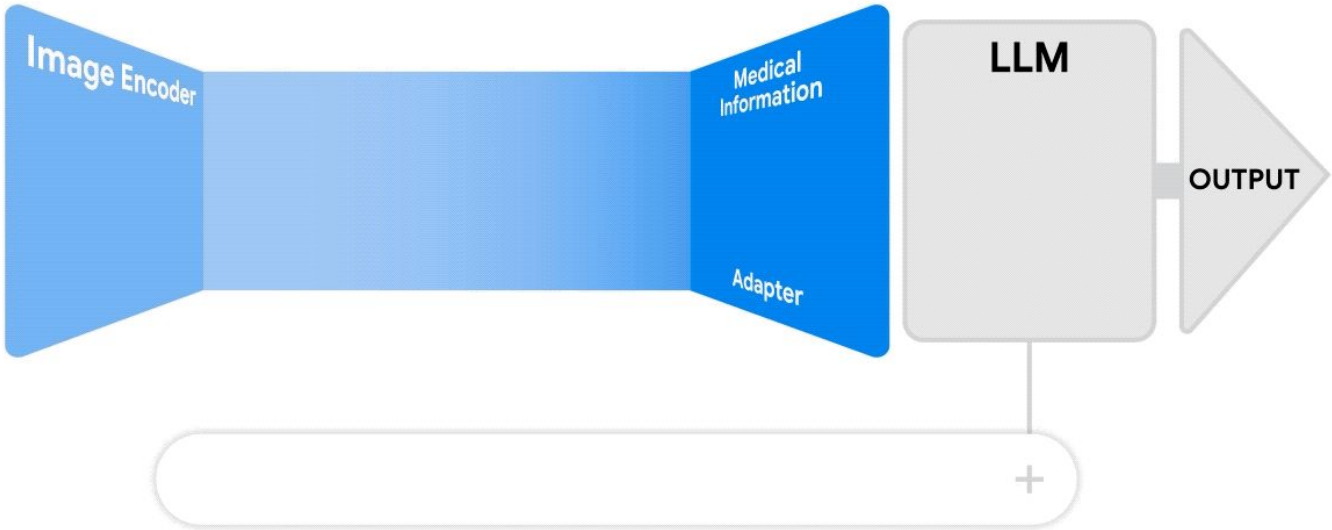
Examples of Models

Med-PaLM : A Medical Large Language Model



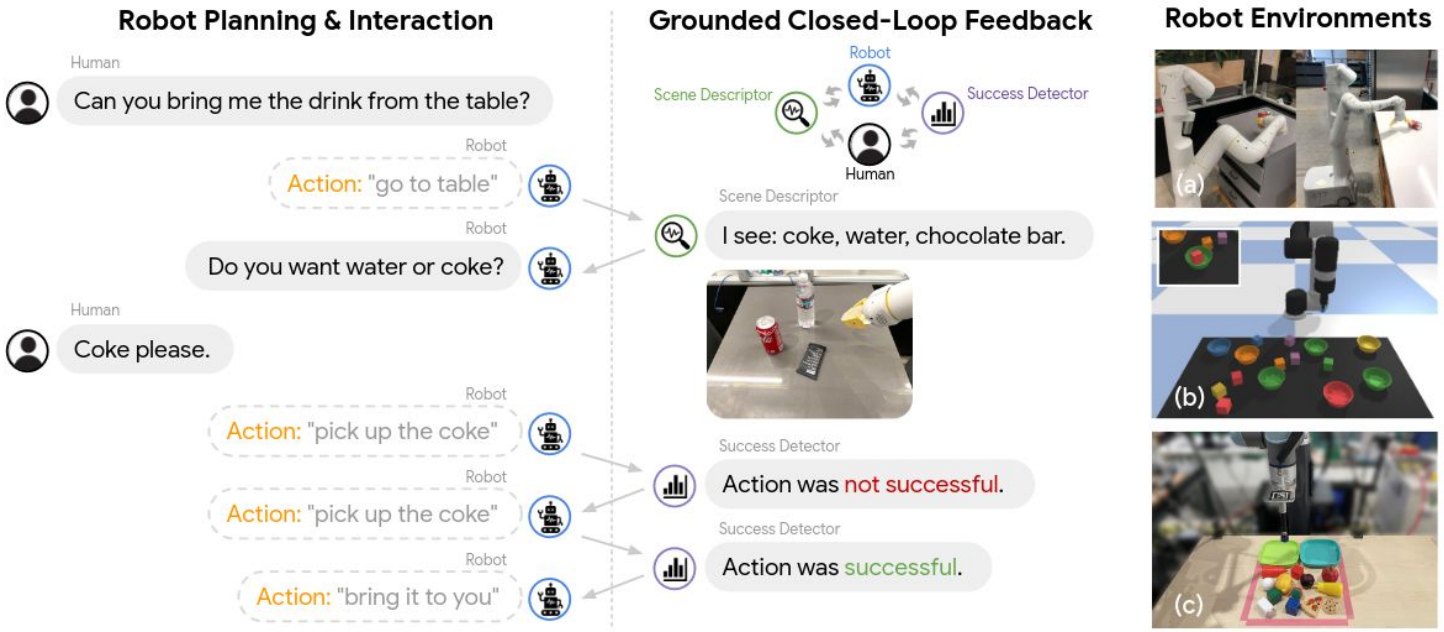
Examples of Models

Med-PaLM : A Medical Large Language Model



Examples of Models

InnerMonologue : Embodied Reasoning through Planning with Language Models



idea : use the knowledge of the LLM to plan a sequence of action and ground this LLM with modules that describe the scene with language

Examples of Models

InnerMonologue : Embodied Reasoning through Planning with Language Models

Go to the table

human : Bring me a drink from the table

Examples of Models

InnerMonologue : Embodied Reasoning through Planning with Language Models

What kind of drink would you like ?

LLM

I see a coke and lime soda

object recognition module

human : Bring me a drink from the table
Robot : Go to the table
Scene : I see a coke and lime soda

Examples of Models

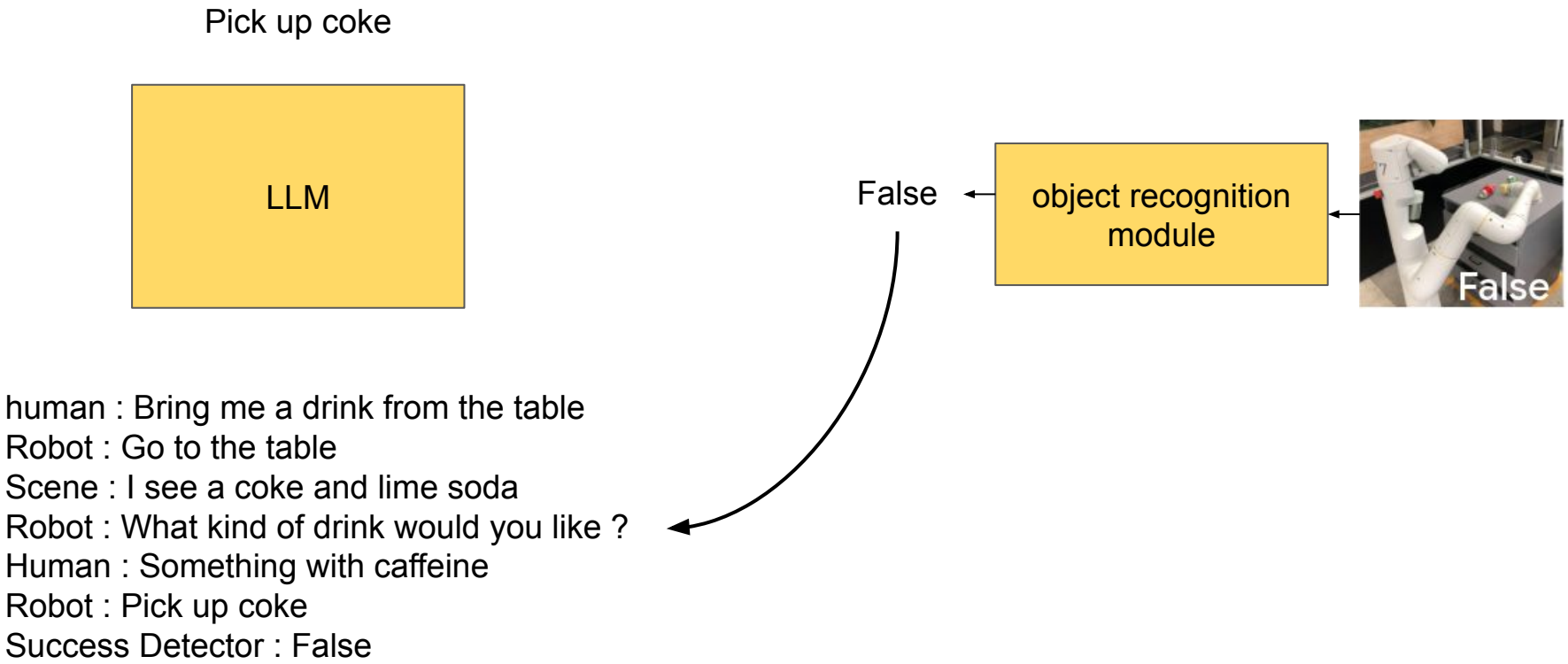
InnerMonologue : Embodied Reasoning through Planning with Language Models

Pick up coke

human : Bring me a drink from the table
Robot : Go to the table
Scene : I see a coke and lime soda
Robot : What kind of drink would you like ?
Human : Something with caffeine

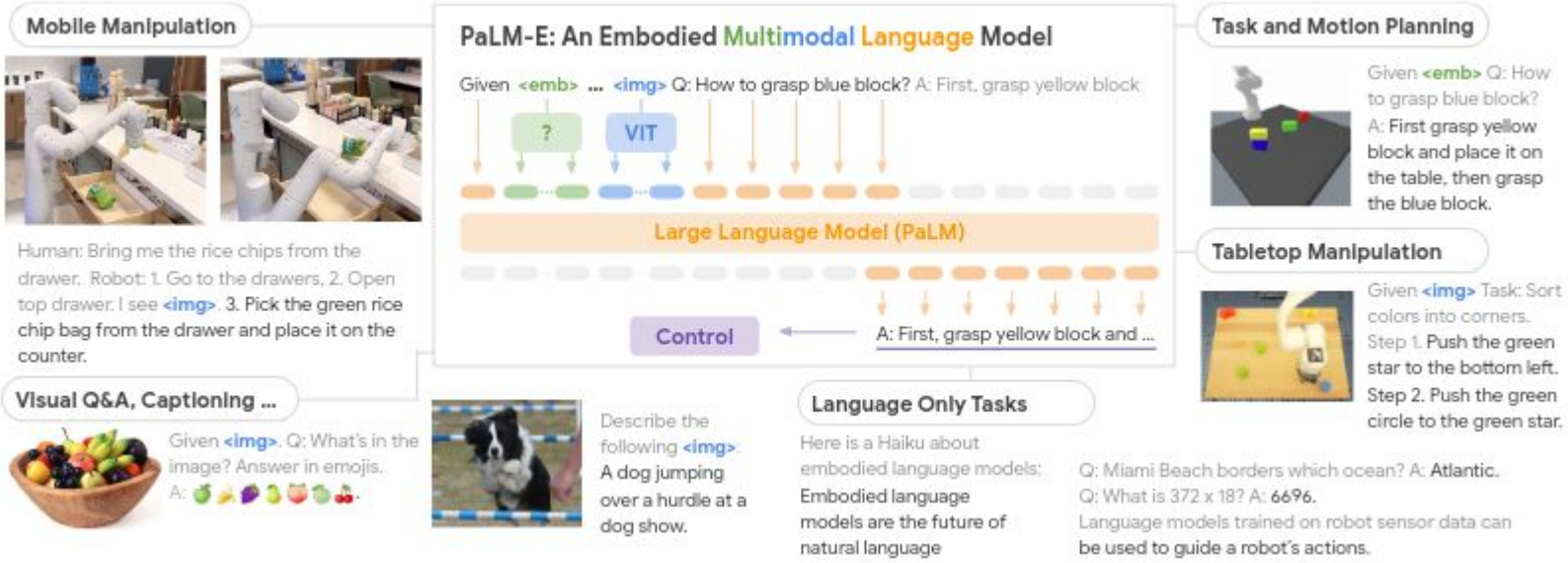
Examples of Models

InnerMonologue : Embodied Reasoning through Planning with Language Models



Examples of Models

Palm-E : An Embodied Multimodal Language Model



idea : Use the LLM to generate text to answer the task (VQA, Captioning, …) and as a high-level policy that sequences and controls the low-level policie

Examples of Models

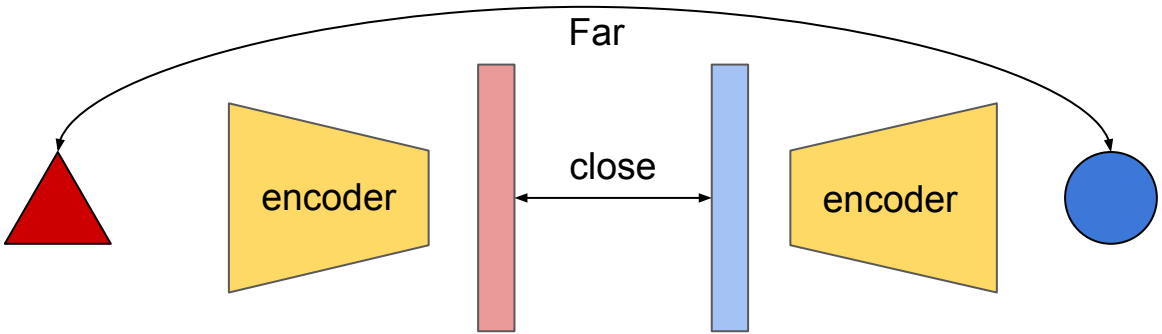
Palm-E : An Embodied Multimodal Language Model

Conclusion

- Use good architecture to encode your data

images	CNN
text	LSTM
video	LSTM / CNN
vectors	MLP

- Encode your data allow you to end up with closer representation of very different modalities



Conclusion

- Learn the task from the multimodal latent space is easier : Captioning, VQA, …
- You can use pre-existing dataset or create your own to train your model

- Nowaday the most popular architecture is the Transformer : can handle multiple modality and is very powerful to learn links between tokens
  - BUT : hard to train → lot of data
- A good practice is also to use foundation models to turn the knowledge they acquired during training to good account