Language & Vision

CS-552: Modern Natural Language Processing
31.05.23

Syrielle Montariol

EPFL

nlp lab

Announcements

- Lecture Tomorrow! Last class of the semester!
  - No exercise session! Work on your project!

- Indicative Feedback released
  - Please fill it out! Due June 11th!

- Course Project:
  - Milestone 2 due Sunday, **June 4th**!
  - Final Report due Sunday, **June 18th**!

Announcement: MAKE Team

- Enjoyed the Modern NLP class project?
  - Wish you could have done it with more compute?
  - Wish you could have explored your solution for longer and scaled it with larger models?
  - Wish you could have trained a real-life multi-billion-parameter ChatGPT?

- EPFL launching MAKE team around large language models and generative AI

  - Dedicated resources to training large language models at scale (A100 GPUs)
  - Opportunity to complete Masters semester projects and theses on training large models at scale
  - Express interest here: https://forms.gle/oRkP2ttT2j82w2VT9

Some inspiration and material of this course is taken from the following courses and tutorials:

- ACL 2022 tutorial https://vlp-tutorial-acl2022.github.io/
- CVPR 2022 tutorial https://www.microsoft.com/en-us/research/videos/cvpr-2022-tutorial-on-recent-advances-in-vision-and-language-pre-training/
- Stanford's course on multimodality: https://web.stanford.edu/class/cs224n/slides/Multimodal-Deep-Learning-CS224n-Kiela.pdf

Don't hesitate to go through it to learn more!

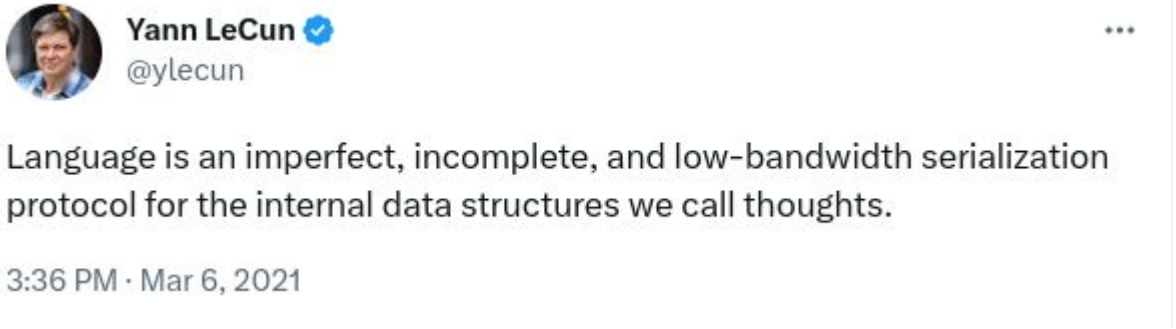
Multimodality

Why Multimodality?

Human learning is inherently multi-modal!

Why Multimodality?

Human learning is inherently multi-modal!



Why Multimodality?

- Human learning & experience is multimodal
- Our daily life tools are multimodal - main one: The internet.
- Multimodal data is richer than language
- To train models: more data is better, and the amount of available textual data is limited (though huge)

→ Multimodality is one of the next frontiers of the revolution of large models!

Which modalities?

Models that can process and link information using various modalities such as:

- image
- video
- text
- audio
- body gestures
- facial expressions
- physiological signals…

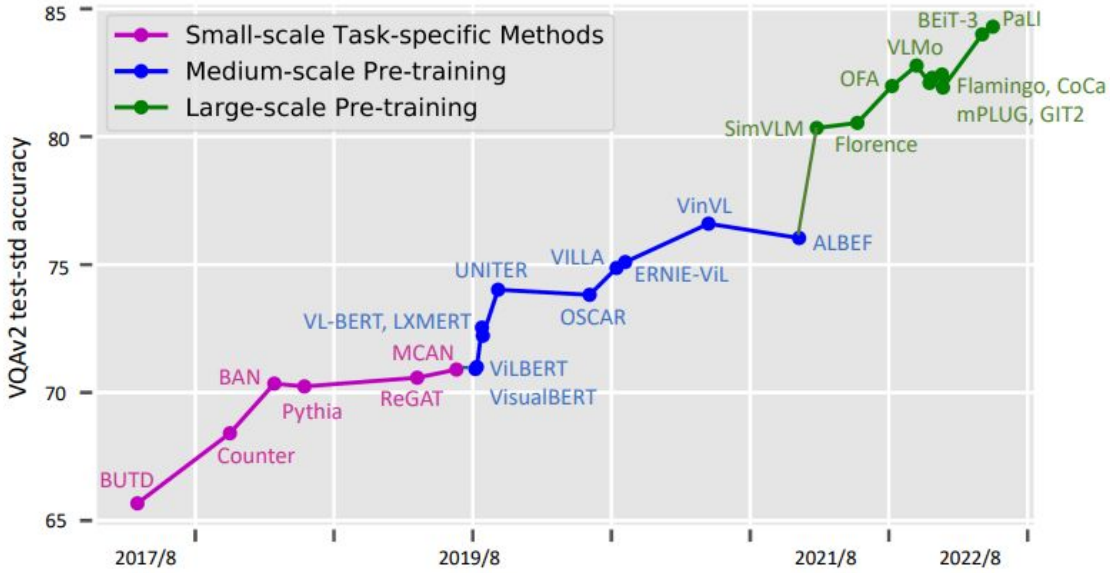
Vision-and language

Example: Detection of hate speech from social media content involving both images and text modalities

https://ai.facebook.com/tools/hatefulmemes/

Evolution of Vision and Language Models (VLMs)

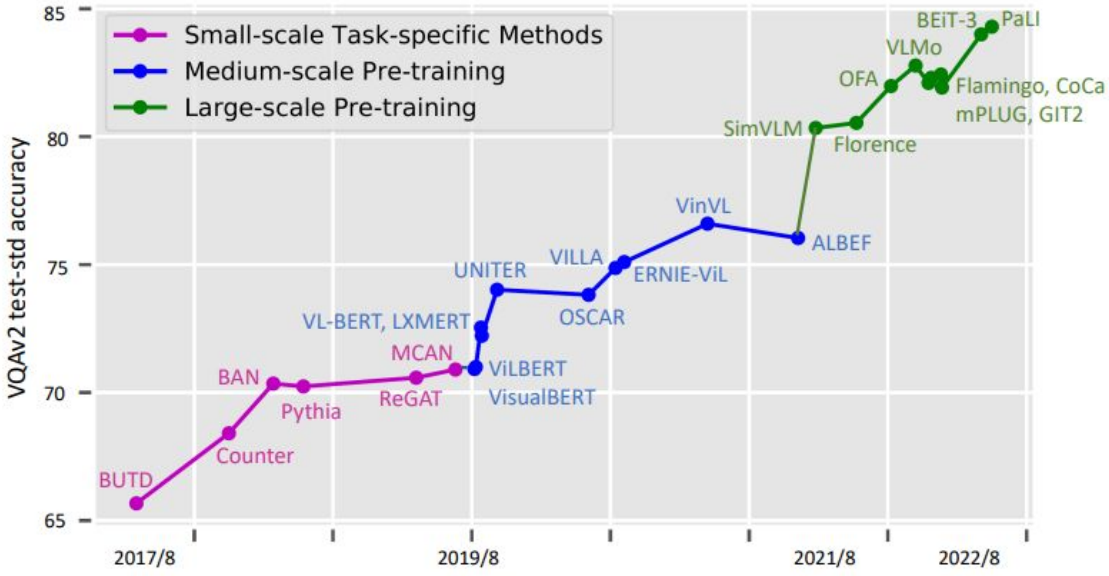
1. Small-scale, task-specific methods: ResNet & FasterRCNN, Glove & Word2Vec



Evolution of Vision and Language Models (VLMs)

1. Small-scale, task-specific methods: ResNet & FasterRCNN, Glove & Word2Vec

2. Medium-scale pre-training, since 2019 (up to 340M parameters with BERT-Large): Inspired by BERT, transformers-based multi-modal fusion.

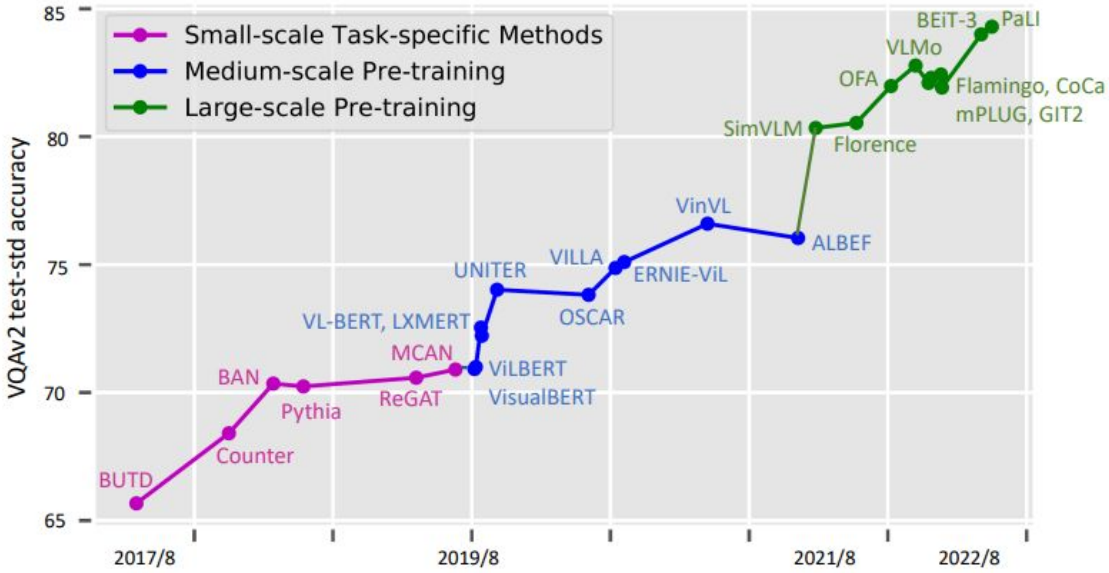


Evolution of Vision and Language Models (VLMs)

1. Small-scale, task-specific methods: ResNet & FasterRCNN, Glove & Word2Vec

2. Medium-scale pre-training, since 2019 (up to 340M parameters with BERT-Large): Inspired by BERT, transformers-based multi-modal fusion.

3. Large-scale pre-training, since 2021: starting with CLIP, then adapting pre-trained LLMs.

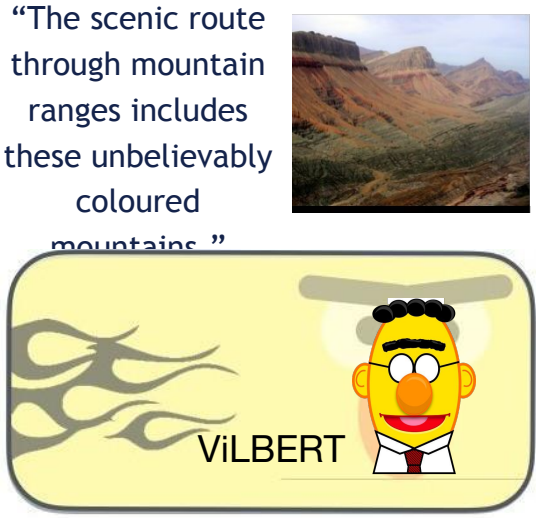


What do we need to make a VLM?

A dataset: image-text pairs where the text describes its image.

A model: attention mechanisms over both image and text

Objective: a loss function specific to the image-text pairs.



Objective: predict the next word given the previous ones and the image

What do we need to make a VLM?

-Model input ?

-Model output ?

-Training dataset?

-Modality encoder?

-Training task / objective?

Key problem: not only understanding text and images, but also their relation!

Today's Outline

Introduction: Multimodality, vision-and-language models

Tasks and data

Encoding modalities

Training

LVLM - large vision and language models

Ethical concerns

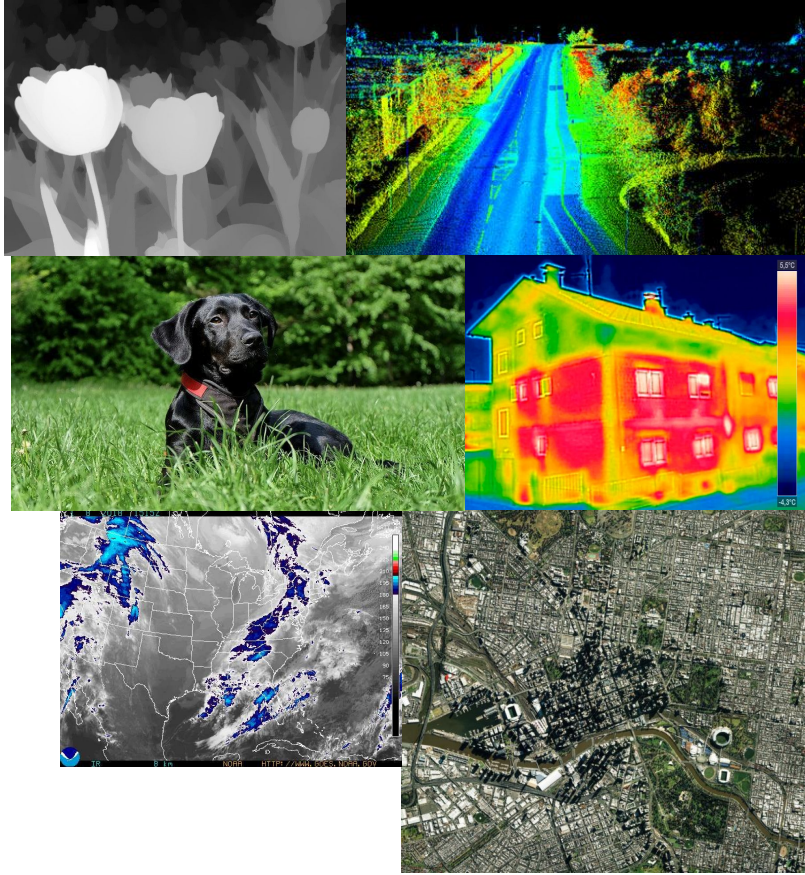
Tasks and Data

Images as modality

Data Types:

- Pictures: grayscale or color (RGB)
- Videos
- 3D Point Clouds, typically obtained using depth sensors like LiDAR (autonomous driving, robotics)
- Depth Maps (3D reconstruction, scene understanding, virtual reality)
- Thermal Images (applications: night vision, surveillance, and thermal anomaly detection)
- Satellite images: visible, infrared, water vapor

…



Classical image-only tasks

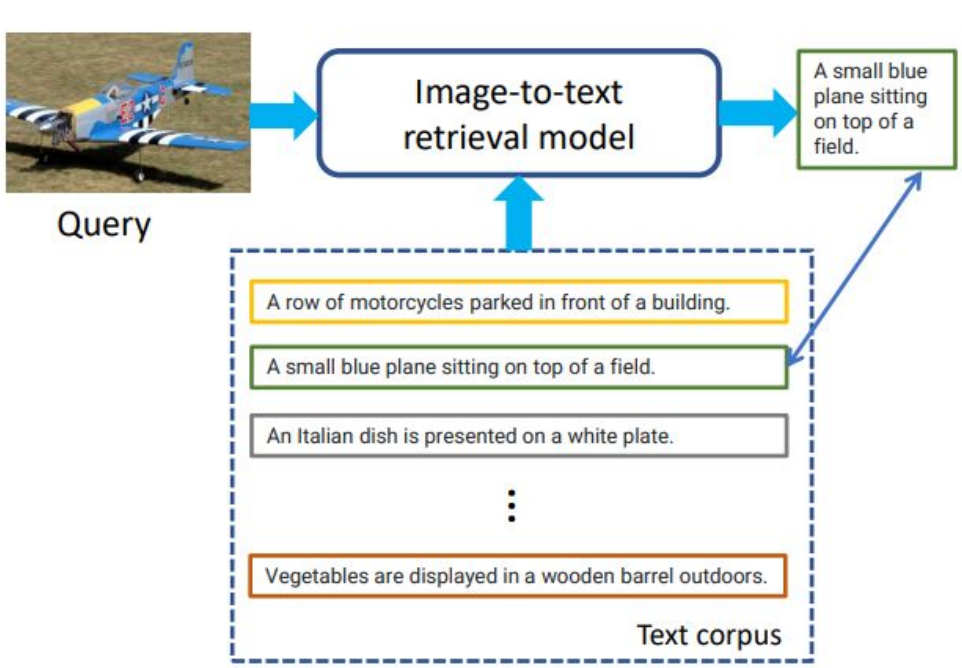
- Face Recognition
- Image Classification
- Object Detection
- Depth Estimation
- Pose Estimation
- Image Segmentation
- Scene Understanding
- Action Recognition
- ...

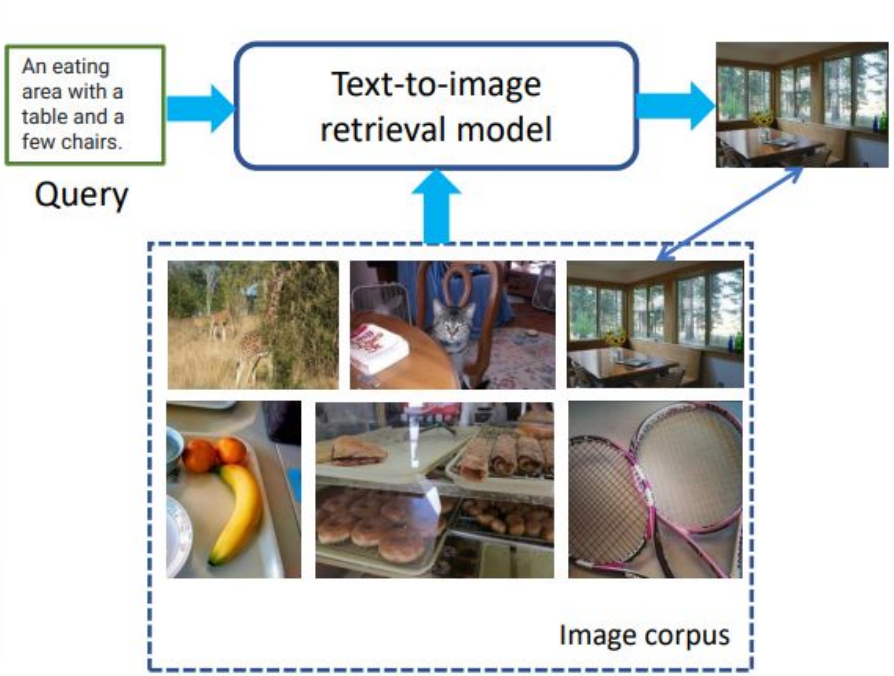
Multimodal tasks

A. Image / text retrieval

B. Image / text generation

Image / text retrieval





Generation tasks

Input / Output	Image	Text	Image + text
Image (Pixel prediction)	Image translation (Colorization, Inpainting, Uncropping…)	Text-to-image synthesis	Text-guided image editing, phrase grounding
Text (next token prediction)	Conditional text generation (Image captioning)	Question answering, summarization…	Visual question answering
Image + text		Visual dialogue	Visual dialogue

Image captioning

Generating a textual description of an image.

COCO style caption: “Single black dog sitting on the grass”

Narratives style caption: “The dog is black and brown. The collar is red. [...] The dog is laying on the grass. There trees behind.”

Genome style caption: “Black dog”, “Red collar”.

SBU-style caption: (Noisy, web-scraped) “My labrador Jackie in the park near the house.”

Visual question answering

Answering a question about an image.

Where is the dog?

What color is the dog’s collar?

What is the animal?

More complex questions

- Counting
- Fine-grained recognition
- Commonsense reasoning

How many slices of pizza are there?
Is this a vegetarian pizza?

Does it appear to be rainy?
Does this person have 20/20 vision?

VQA: Visual Question Answering, 2015

Grounded captioning / Phrase grounding

"One small boy climbing a pole with the help of another boy on the ground"

"A man talking on his cellphone next to a jewelry store"

Flickr30k

"zebra facing away"

RefCOCO+

Visual Dialogue

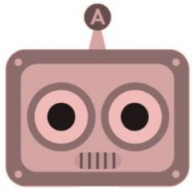
C: A dog with goggles is in a motorcycle side car.
Q: Is motorcycle moving or still?
A: It's parked
Q: What kind of dog is it?
A: Looks like beautiful pit bull mix

Q: What color is it?

Image

Dialog history

Question



Visual Dialog model

Answer

A: Light tan with white patch that runs up to bottom of his chin

https://visualdialog.org/

Multimodal dialogue

Grounding Language Models to Images for Multimodal Generation, 2023

I am thinking of an animal. It is brown and furry, and has a tail. What is it?

A squirrel.

No, the animal I am thinking of lives in the river.

A beaver.

That's right, what do beavers look like?

They have a tail.

Show me an illustration of a beaver

A beaver.

Yes, what about a pencil drawing of one?

What about a photo of one?

What is this?

This is a house with a unique design.

What is its shape?

The house is shaped like a dome.

Can you show me more houses like this?

The house is designed by architect.

That's nice, any others?

Yes, there are many houses like this in the world.

Provide instructions on how to cook.

How do I make macarons?

Macarons are a French dessert that are made by combining egg whites and sugar. The egg whites are beaten until stiff and then folded into the sugar.

What does the stiff egg white look like?

The egg white should be stiff and glossy.

Show me a picture of what that looks like

After this, what do I do?

The macaron is then baked in a hot oven.

Many tasks can be solved through either generation or retrieval!

Multimodal Concept Composition

[Grounding Language Models to Images for Multimodal Generation](), 2023

Benchmarks: A-OKVQA (Augmented Outside Knowledge Visual Question Answering)

Q: Which position will the red jacket most likely finish in?

A: Fourth

Commonsense

Q: What makes those chairs easy to carry?

A: Foldable

Physical

Q: What was the name of the first cloned type of this animal?

A: Dolly

Knowledge base

Q: What body part is he using to maintain balance most effectively?

A: Arms

Visual

Benchmarks are very artificial situations…
→More real-life applications?

VizWiz Grand Challenge: Answering Visual Questions from Blind People (2018)

Q: Does this foundation have any sunscreen?
A: yes

Q: What is this?
A: 10 euros

Q: What color is this?
A: green

Q: Please can you tell me what this item is?
A: butternut squash red pepper soup

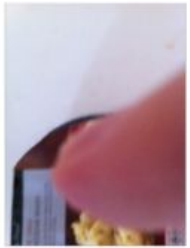
Q: Is it sunny outside?
A: yes

Q: What type of pills are these?
A: unsuitable image

Q: What type of soup is this?
A: unsuitable image

Q: Who is this mail for?
A: unanswerable

Q: When is the expiration date?
A: unanswerable



Q: What is this?
A: unanswerable

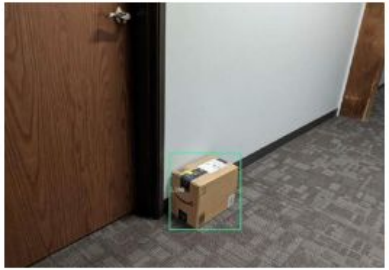
- Images are captured by blind photographers and so are often poor quality
- Questions are spoken and so are more conversational
- Often, visual questions cannot be answered.

Hateful memes detection

https://ai.facebook.com/tools/hatefulmemes/

Object detection in the wild

Mask Wearing



Packages

Pistols

Potholes

Encoding modalities

What do we need?

A text encoder

An image encoder

A method to use information from both encoders:

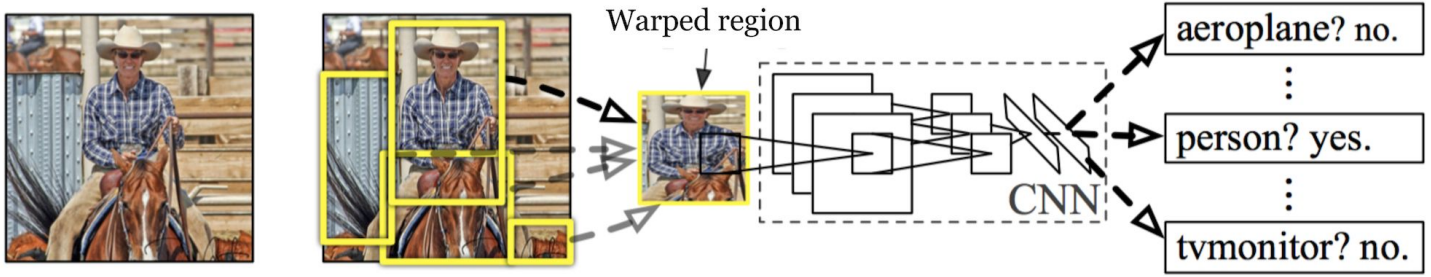
- loss function
- architecture
- learning strategy

Text encoding

- Tf-idf
- Word2Vec
- BERT-like pre-trained text encoders

Image encoding

- sparse features (object detector):
  - R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN
  - Expensive to annotate, need box labels



Rich feature hierarchies for accurate object detection and semantic segmentation, 2013

Image encoding

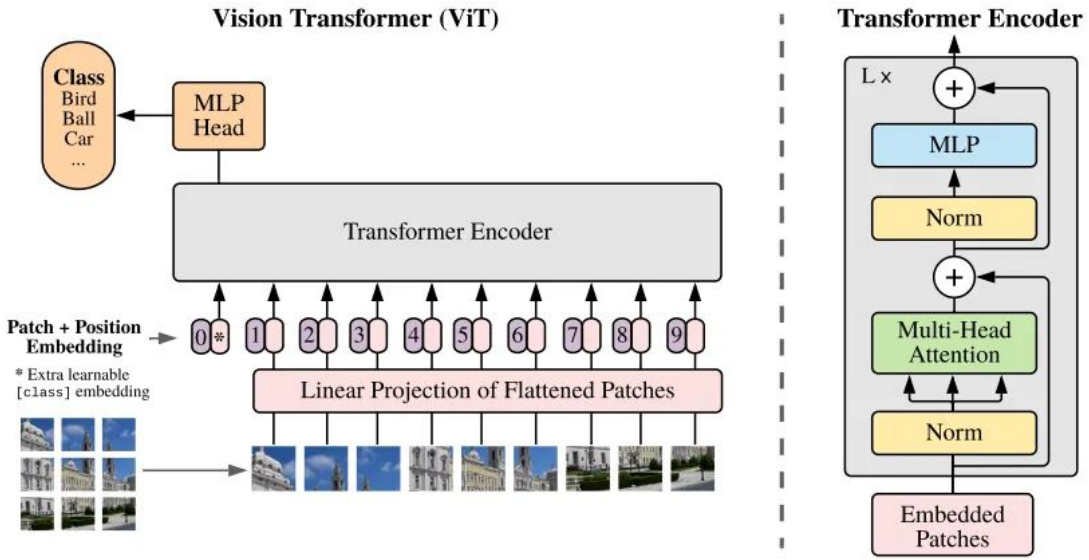
- sparse features (object detector):
  - R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN
  - Expensive to annotate, need box labels

- dense feature:
  - CNN: ConvNet layers
  - Vision transformer (ViT)

Visual transformer - ViT

Transformers cannot process grid-structured data, only sequences!

1) Transform the images into a sequence of patches (tokens)
2) “flatten” them
3) Add position encoding



Using both modalities for task-specific VM models

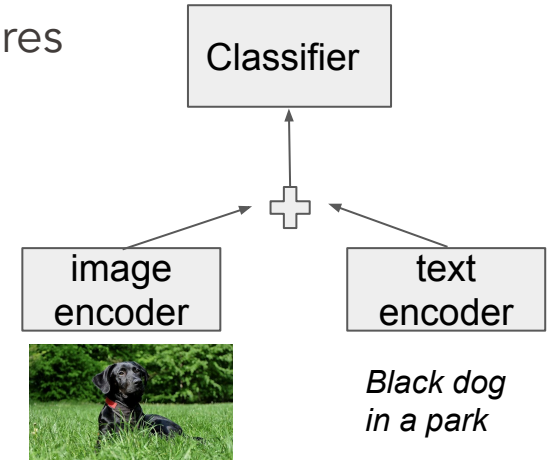
A. Early fusion models

B. Late fusion models

C. Cross-modal attention models

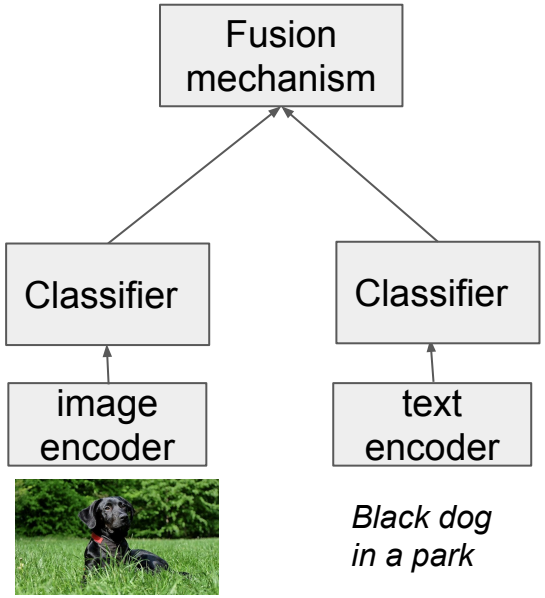
A. Early fusion

- Fuse the features: element-wise product or sum or concatenation
- Exploit dependencies between features
- Can end up very high dimensional



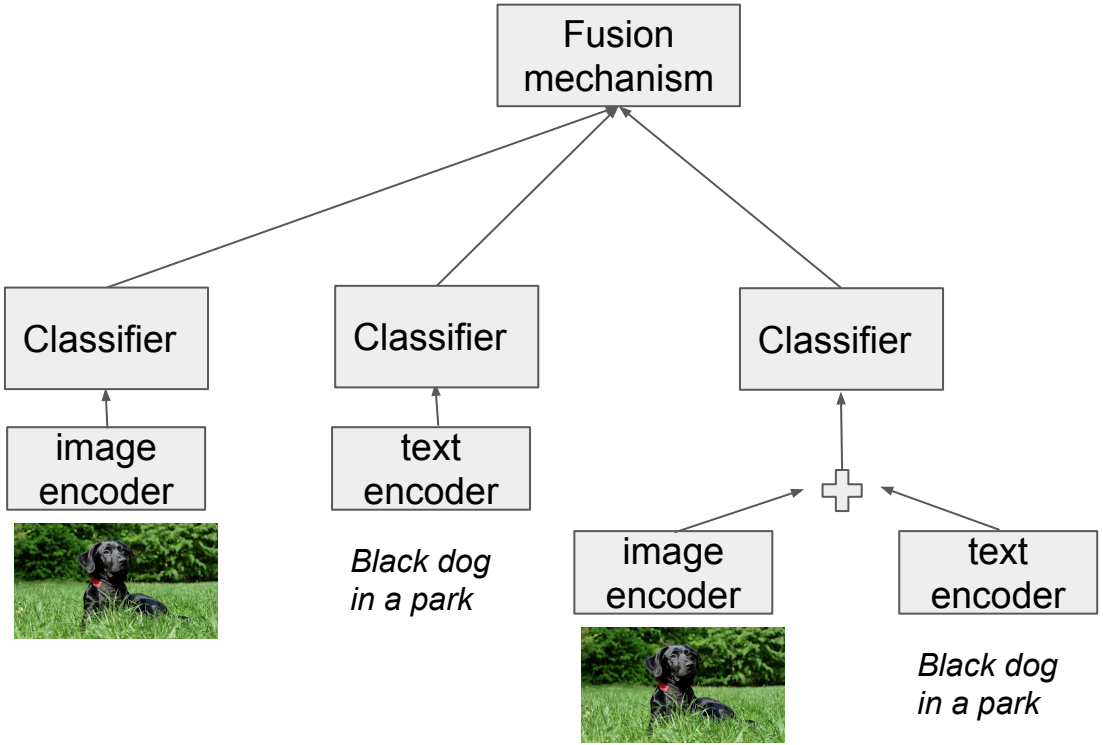
B. Late fusion

- Train two unimodal predictors and a multimodal fusion one
- Requires multiple training stages
- Do not model low level interactions between modalities
- Fusion mechanism can be voting, weighted sum or an ML approach



Mixed fusion

- Combine benefits of both



C. Cross-modalities attention

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

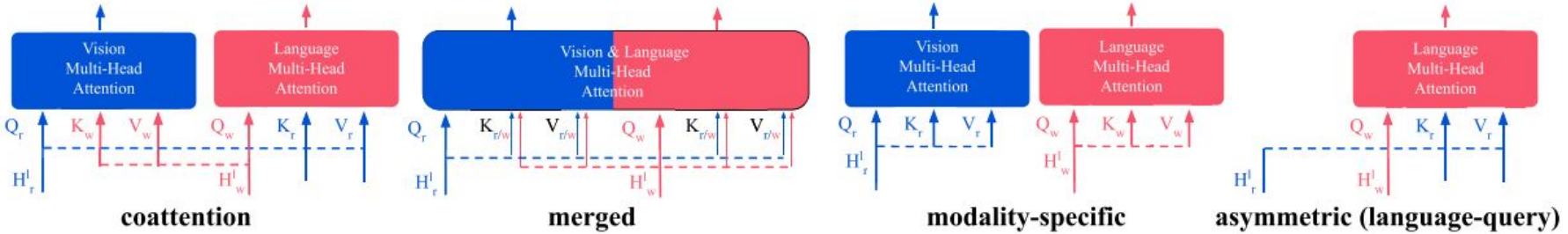
Co-attention: Each modality attends only to the other modality.

- Capture multimodal alignment between image and text inputs.
- Higher weights are put on the image regions that are more useful to solve the task.

Merged attention: Each modality attends to both modalities

Modality-specific attention: capture unimodal relation over image regions or text tokens.

Asymmetric attention: Only one modality (e.g., language) attends to the other modality (e.g., image).



Queries, keys, and values are shown by Q, K, and V ; w and r index text tokens and image regions, respectively. H is the activation at layer l.

Decoupling the Role of Data, Attention, and Losses in Multimodal Transformers, 2021

Training

Masked-Language Modeling / Image-Text Matching

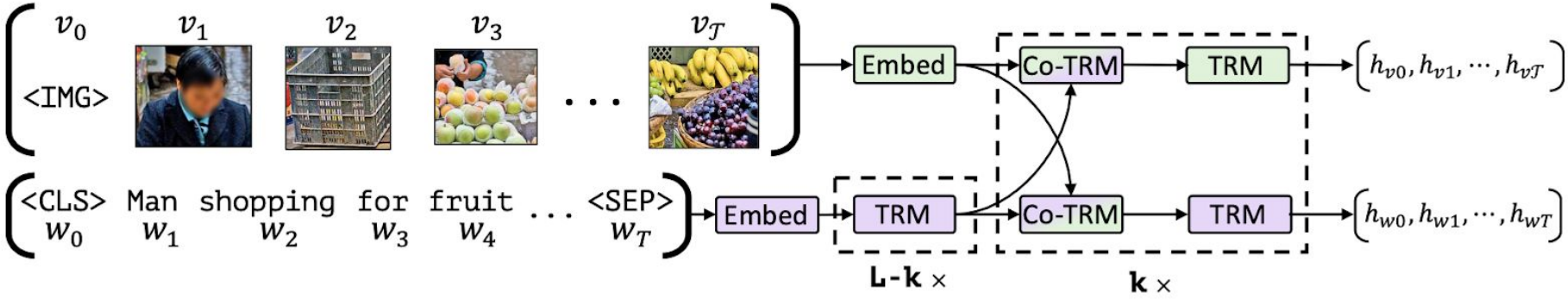
Idea:
Training objective: masked-language modeling, image-text matching.

Goal: Align elements of an input text and regions in an associated input image.

Architecture: image and text encoder

Models: VisualBERT, FLAVA, ViLBERT, LXMERT...

Masked-Language Modeling / Image-Text Matching



1) MLM: predict the masked words based on the corresponding image (with objects regions, eg from an object detector)
2) ITM: predict whether the caption matches the image or not.

ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, 2019

Training scenarios

Two most popular training scenarios:

- CLIP-style **contrastive learning** (image-text retrieval)

  → Use task-specific heads for each downstream task.

- Next-token prediction (LM)

  → Treat all downstream tasks as language generation.

Contrastive Learning

(commonly used in computer vision)

Idea: learning a text encoder and an image encoder jointly with a **contrastive loss**, using large datasets of **{image, caption} pairs**.

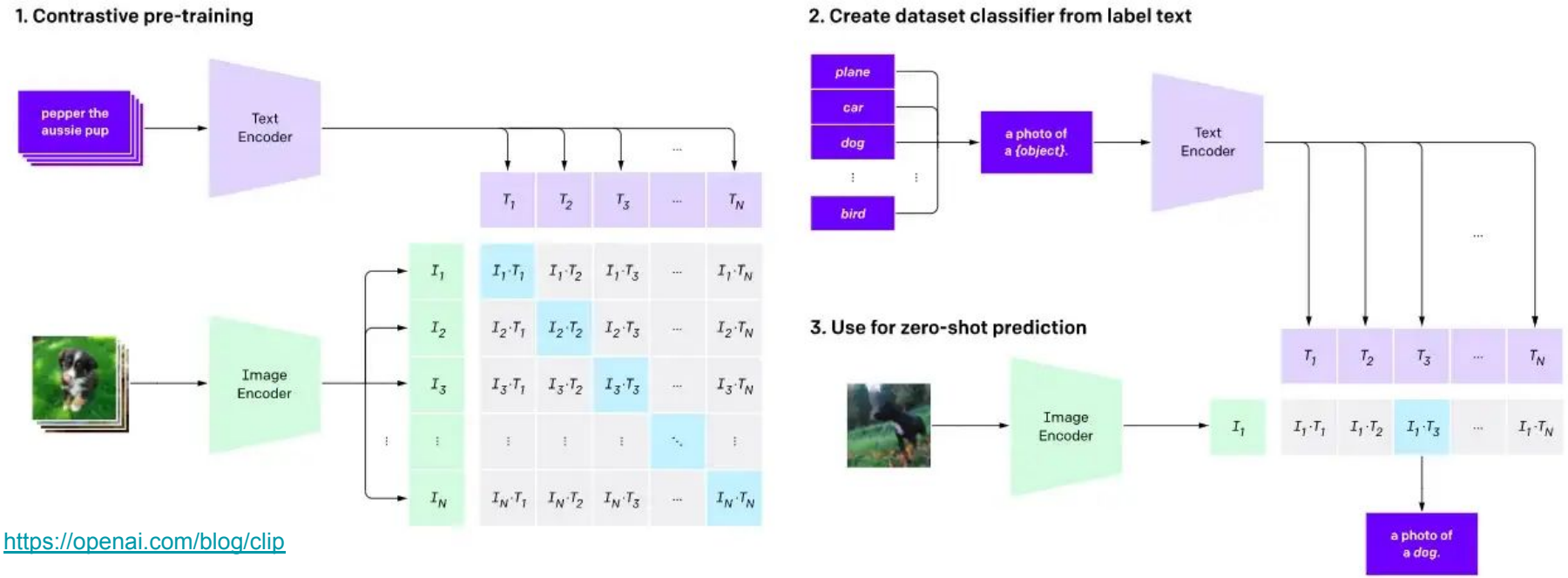
Architecture: image and text encoder

Models: CLIP, CLOOB, ALIGN, DeCLIP, LiT, FLAVA.

Variations: keeping one of the two encoders frozen (LiT), more advances distance metrics (ALIGN, DeCLIP)

Contrastive Learning

Aligning images and texts to a joint feature space in a contrastive manner



https://openai.com/blog/clip

Towards Generative Models

Pros ?

Cons ?

Towards Generative Models

Pros:

- Similarly to language-only tasks: unified modeling of vision–language tasks
- Better out-of-distribution generalization

Towards Generative Models

Pros ?

Cons ?

Towards Generative Models

Pros:

- Similarly to language-only tasks: unified modeling of vision–language tasks
- Better out-of-distribution generalization

Cons:

- Harder to evaluate
- Inheriting existing limitations of pretrained language models (hallucination, biases…).

PrefixLM

Idea:

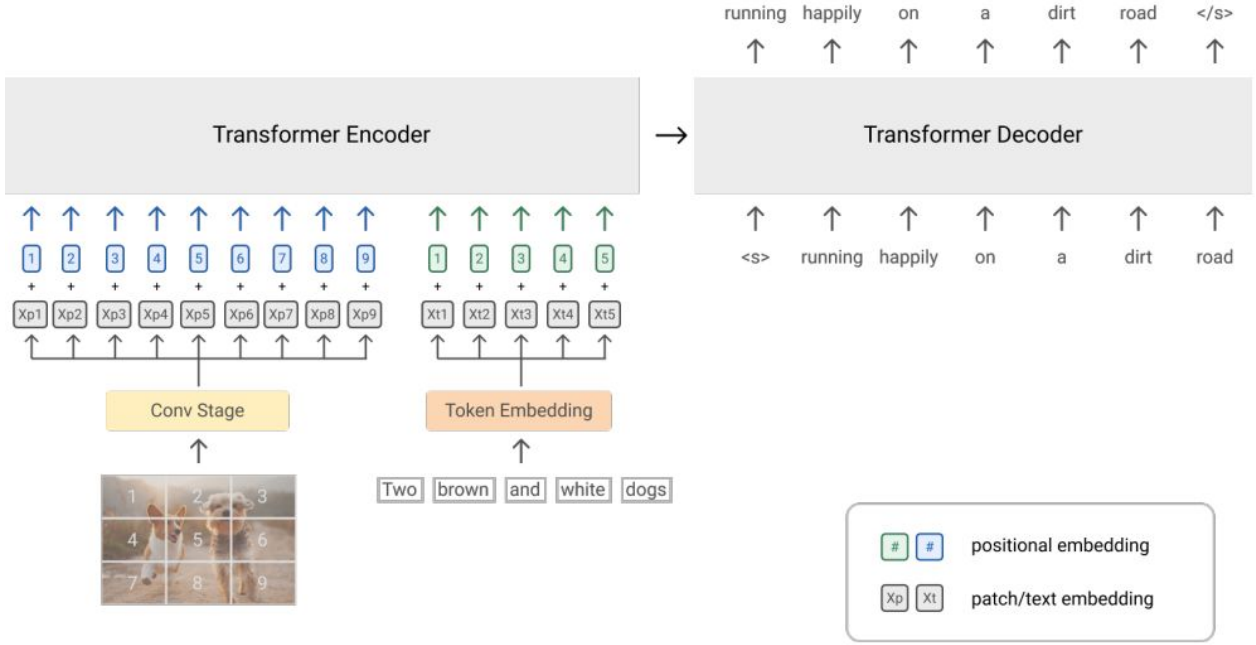
- Divide images into sequences of patches;
- Use them as a prefix to a language model (predict the next token given the prefix image & text).
- Tasks are image-conditioned text generation (captioning and VQA), cannot generalize to tasks like object detection and image segmentation.

Architecture: image and text encoder, text decoder

Models: SimVLM, VirTex

PrefixLM

Jointly learning image and text embeddings by using images as a prefix to a language model



SimVLM: Simple Visual Language Model Pretraining with Weak Supervision, 2021

PrefixLM

Idea:

- Divide images into sequences of patches;
- Use them as a prefix to a language model (predict the next token given the prefix image & text).
- Tasks are image-conditioned text generation (captioning and VQA), cannot generalize to tasks like object detection and image segmentation.

Architecture: image and text encoder, text decoder

Models: SimVLM, VirTex

What about using a bigger LM?

Large VLMs using LLMs

The breakthrough of LLMs

LLMs :

- have significantly more parameters
- undergo longer and more thorough training
- are trained on massive corpora of text from the internet
- have a broader context window

The breakthrough of LLMs

LLMs :

- have significantly more parameters
- undergo longer and more thorough training
- are trained on massive corpora of text from the internet
- have a broader context window

Direct consequences:

- Improved contextual understanding: more coherent and contextually relevant responses.
- Expanded encyclopedic and commonsense knowledge knowledge
- Generalization ability through in-context learning
- Reasoning ability (chain of thought, generating answer rationales)

How can we use LLMs for vision-and-language tasks?

How can we use LLMs for vision-and-language tasks?

Transform each task into a text generation task.

- Transforming image features to text
- Prefixing them with an image encoder
- Multi-modal fusion in the transformer blocks

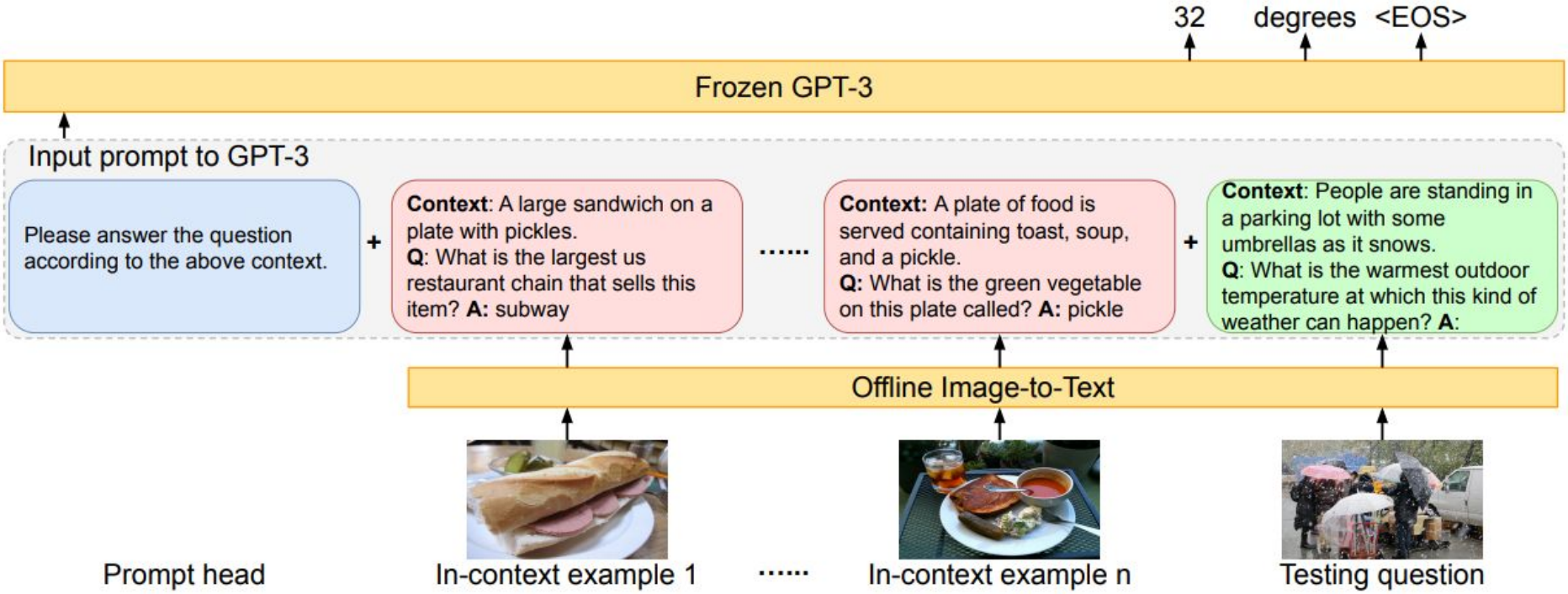
Encoding image features as text

PICa: Prompting GPT-3 via the use of Image Captions

- Treating GPT-3 as an implicit and unstructured KB
- Translate images into captions/tags so that GPT-3 can understand it
- 4 shots outperform supervised SOTA on OK-VQA! (requires external knowledge to correctly answer the question)

An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA, 2021

Encoding image features as text



An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA, 2021

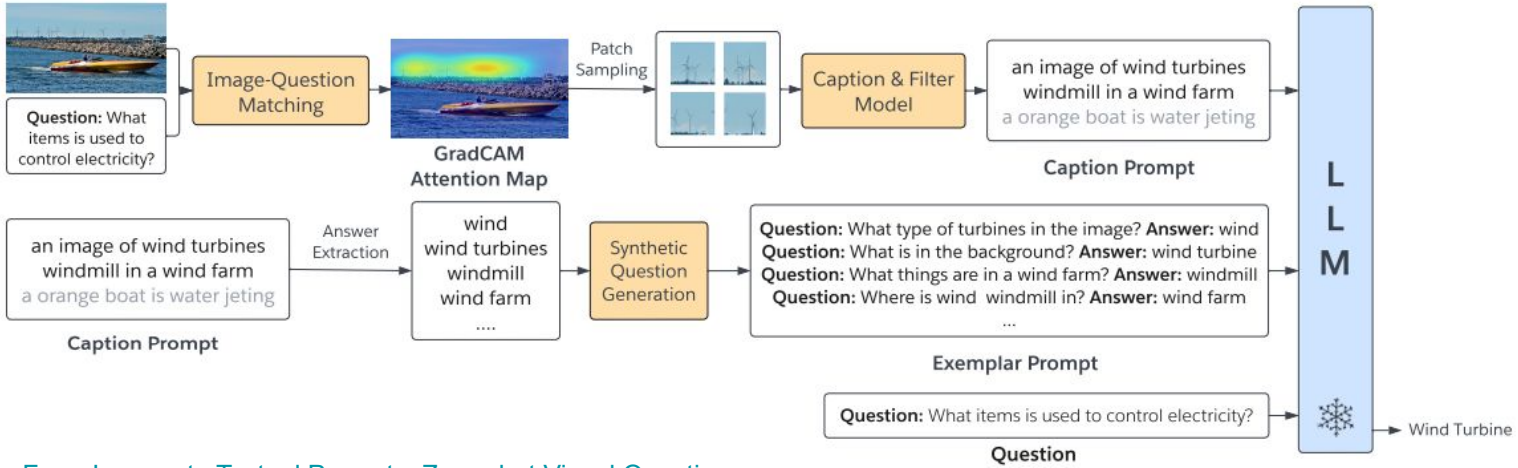
Encoding image features as text

Img2LLM: Image captioning + question generation

Question: What type of profession is the man in red in?
GT Answer: bartender

Captions 1: a man in red shirt at a bar making drinks
Captions 2: a man in a red shirt is making a wine tasting
Captions 3: a man in a red shirt at a bar serving a bar

Synthetic Question 1: who is pouring a drink at a bar?
Answer: A man
Synthetic Question 2: where is a man in a red shirt making drinks? **Answer:** A bar
Question: What type of profession is the man in red in?
Predicted Answer: bartender



From Images to Textual Prompts: Zero-shot Visual Question Answering with Frozen Large Language Models, 2023

What are the limitations of these methods?

What are the limitations of this method?

- Additional inference
- Need existing image captioning model
- Loss of information

PICa

(e) What color is the man's jacket?
Context: A man flying through the air while riding a snowboard.
Answer: black
GT Answer: ['red', 'red', 'red', 'orange', 'red', 'red', 'red', 'red', 'red', 'red']

Acc.: 0.0

(f) How many giraffes are there?
Context: A herd of giraffe standing next to a wooden fence.
Answer: 3
GT Answer: ['6', '6', '8', '6', '8', '6', '6', '7', '8', '7']

Acc.: 0.0

Img2LLM

Question:what is the purpose of the wide tires on that bike?
GT answer:balance/traction/brake

Caption 1: a cargo bike sitting on a tire wheel.
Caption 2: the man is riding a bike on sands.
Caption 3: a man stands on a wheel on some sands.

Synthetic question 1:what are the tires on?
Answer: wheels
Synthetic question 2:what is a man doing on a bike?
Answer: riding
Question: What is the purpose of the wide tires on that bike?
Predicted answer: ride sand

What about using an image encoder?

Back to our PrefixLM method!

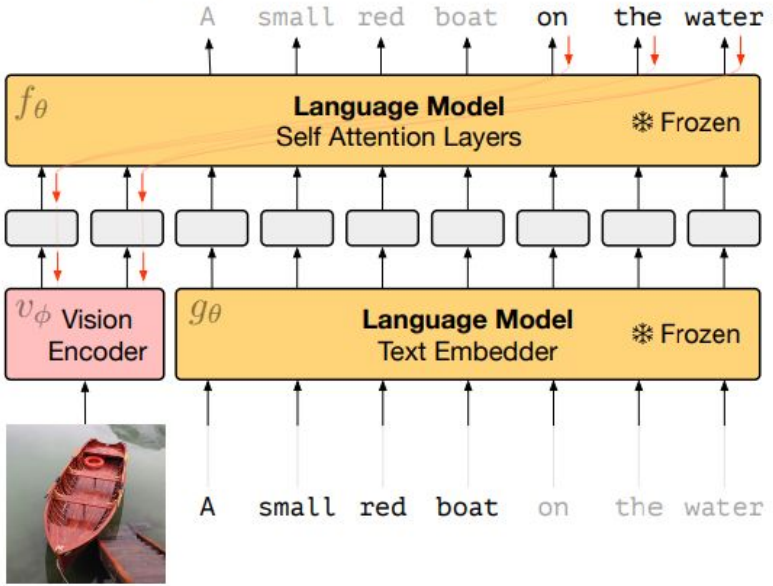
Frozen PrefixLM

Idea: learning image embeddings that are aligned with a frozen language model (only train image encoder).

Objective: autoregressive LM on aligned text-image pairs.

Architecture: image encoder, frozen LLM

Models: Frozen, ClipClap



Multimodal Few-Shot Learning with Frozen Language Models, 2021

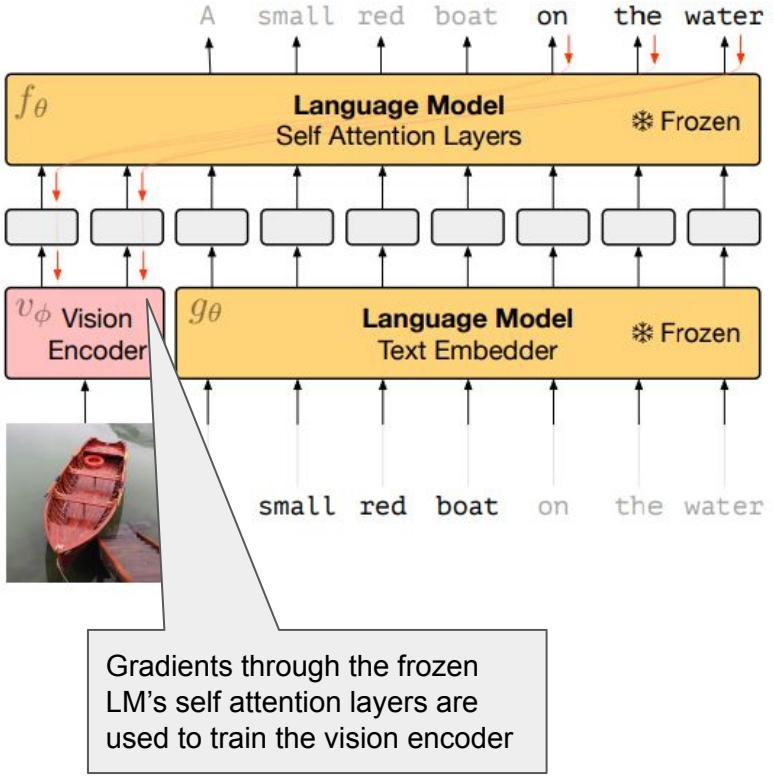
Frozen PrefixLM

Idea: learning image embeddings that are aligned with a frozen language model (only train image encoder).

Objective: autoregressive LM on aligned text-image pairs.

Architecture: image encoder, frozen LLM

Models: Frozen, ClipClap



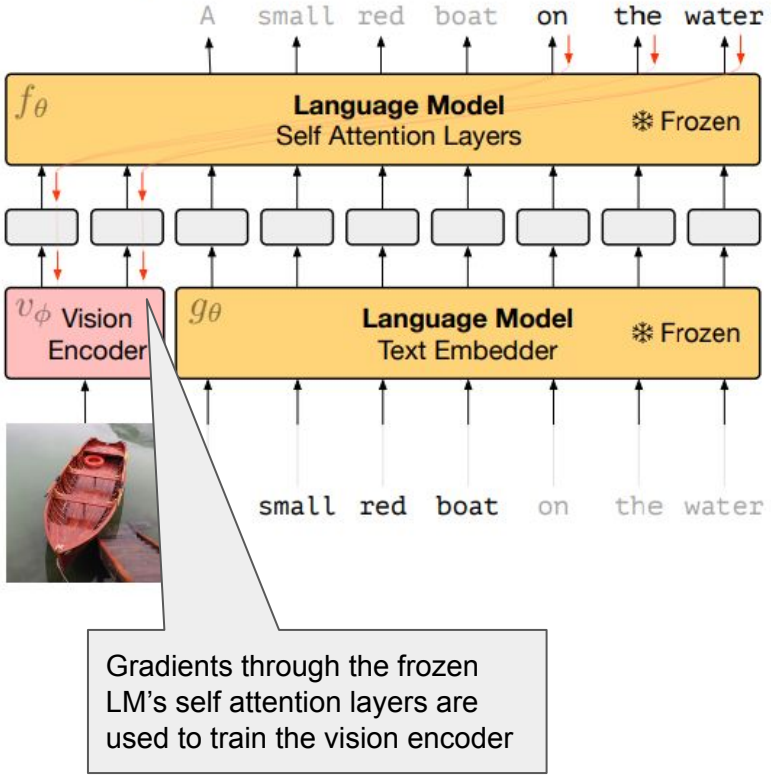
[Multimodal Few-Shot Learning with Frozen Language Models](), 2021

Frozen PrefixLM

Issue: training the vision encoder is costly, getting gradients from GPT-3 is impossible

Variations: Frozen image encoder as well! Inserting new cross-attention layers between existing pre-trained and frozen LM layers to condition the LM on visual data

E.g. FROMAGe, Flamingo, MAPL.



Multimodal Few-Shot Learning with Frozen Language Models, 2021

Example: FROMAGe (Frozen Retrieval Over Multimodal Data for Autoregressive Generation)

- Start from a frozen pretrained LLM, and a frozen pretrained visual encoder
- Finetune only input and output linear layers to enable cross-modality interactions

Train with a multitask objective for:

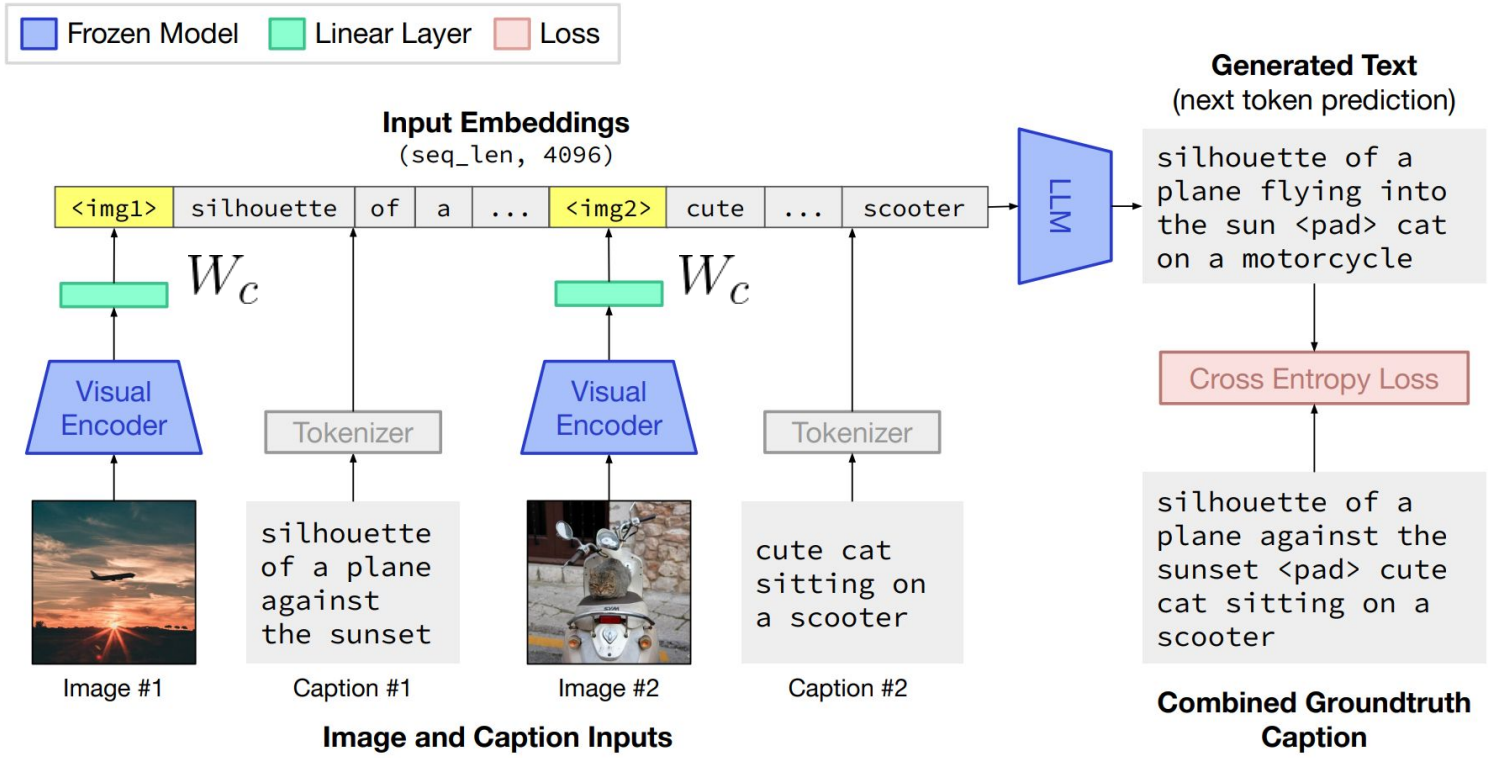
(1) image captioning (learning to process interleaved multimodal inputs)

(2) image-text retrieval (learning to produce interleaved multimodal outputs).

- Only update the weights of the linear layers and a special [RET] token embedding during training.
- The final training loss is a weighted sum of the captioning loss and the retrieval losses.

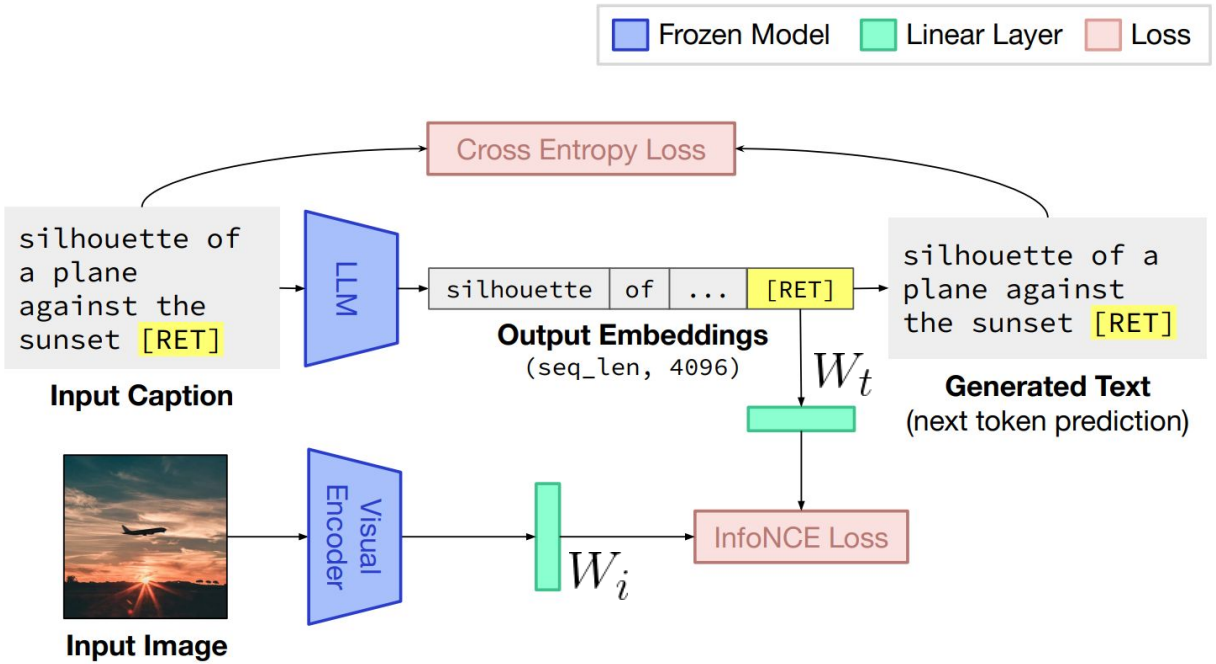
FROMAGe Image captioning

[Grounding Language Models to Images for Multimodal Generation, 2023]



- Extract visual embeddings from the visual encoder
- Learn a linear mapping through the maximum likelihood objective
- Map visual embeddings into the input space of the language model.

FROMAGe Text-to-image (t2i) and image-to-text (i2t) retrieval



- Train the language model to learn a new [RET] token which represents an image
- Learn a linear mapping through contrastive learning to map the [RET] embeddings for a caption to be close to the visual embeddings for its paired image.

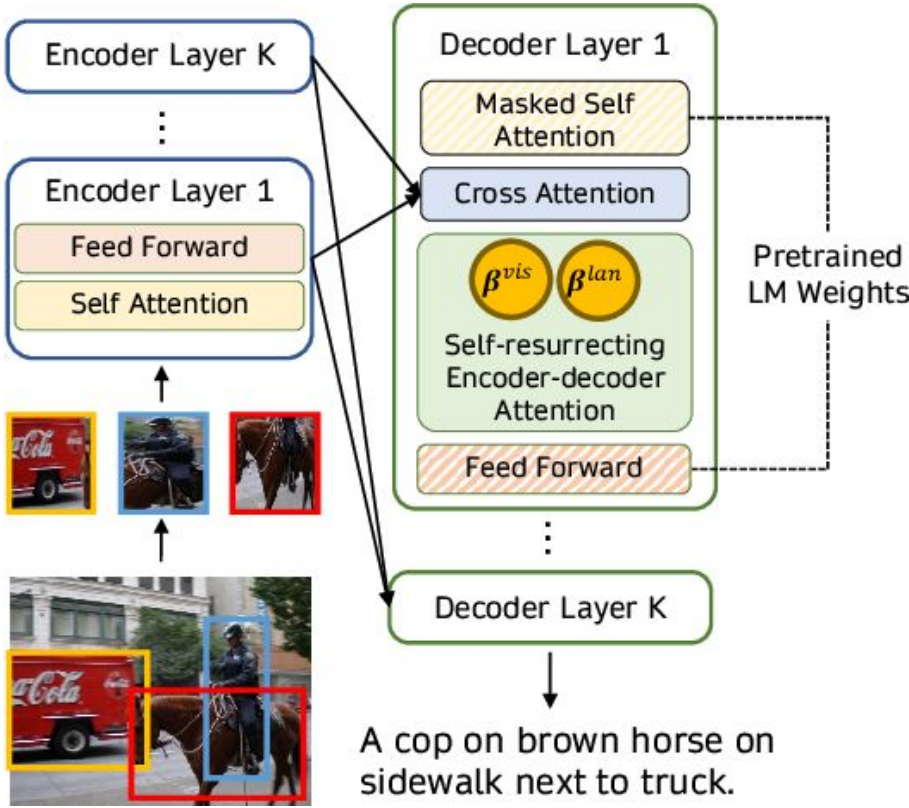
Multi-modal Fusing with Cross Attention

Idea:

- Fusing visual information into layers of a language model with a cross-attention mechanism.
- Leveraging pre-trained language models requires less multimodal training data!

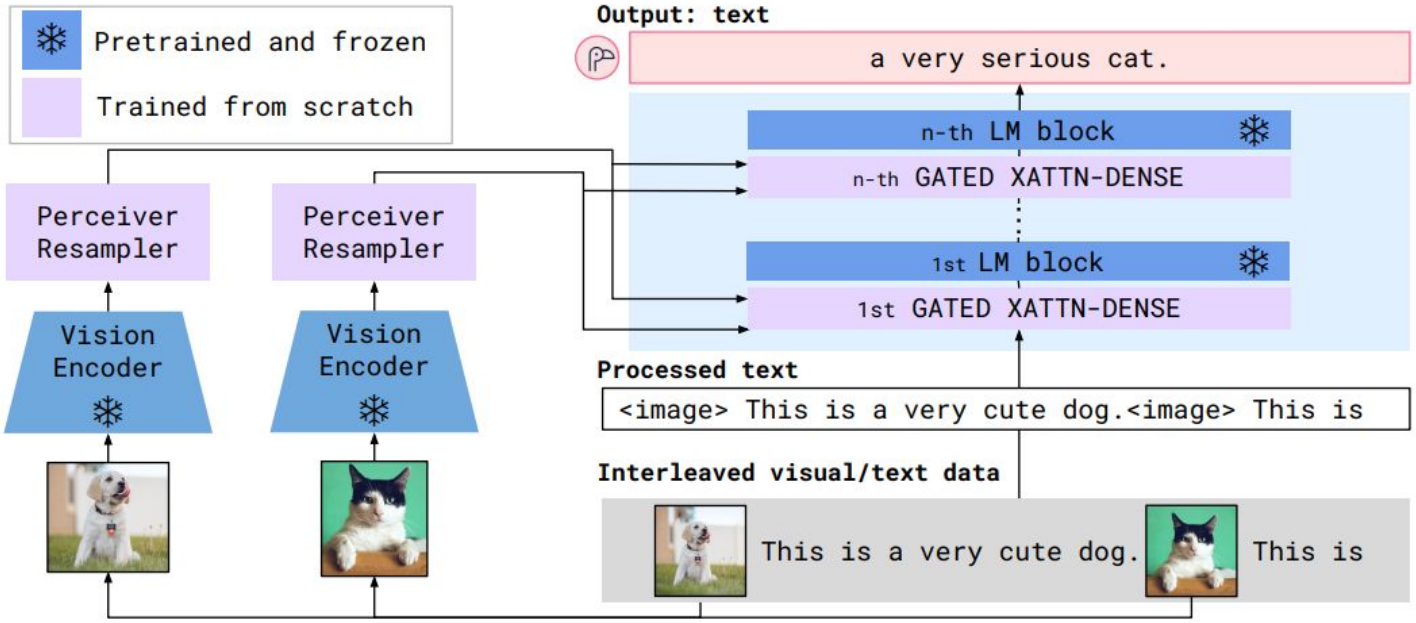
Architecture: image encoder, frozen LLM

Models: VisualGPT, VC-GPT, and Flamingo



VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning, 2022

Multi-modal Fusing with Cross Attention



[Flamingo: a Visual Language Model for Few-Shot Learning, 2022]

Current largest VLMs

GPT-4, OpenAI (?)

PaLM-E, Google (PaLM (540B) + ViT-22B)

User What is funny about this image? Describe it panel by panel.

Source: hmmm (Reddit)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

https://openai.com/research/gpt-4

Chain of Thought with VLMs

PaLM-E = PaLM (540B) + ViT-22B

Visual chain-of-thought reasoning:

→The model breaks down its answering process in smaller steps

Given **<img>**. Q: How many championship rings did the player on the left win in his career A: Let's think step by step. 1. kobe bryant is on the left. 2. kobe bryant has won 5 championship rings.

Given **<img>**. Q: Who are the two teams playing in this photo, and which was the last to win a championship, and which year did they win, and who was their star player that year? A: Let's think step by step. 1. team in white is new york knick. 2. team in green is boston celtic. 3. last to win championship was boston celtic. 4. year was 2008. 5. star player was paul pierce. 6. i am not sure if this is correct. i am not basketball fan. i just google it.

Given **<img>**. Q: Can I go down this street on a bicycle, yes or no? A: Let's think step by step. 1. do not enter. 2. except bicycles. 3. do not entry except bicycles. 4. yes.

Photo 1: **<img1>**. Photo 2: **<img2>**. Q: What is in Photo 1 but NOT in Photo 2? A: Let's think step by step and explain. Photo 1 has sunglasses on top of folded clothes. Photo 2 does not have sunglasses on top of folded clothes. Therefore, sunglasses are in Photo 1 but not in Photo 2.

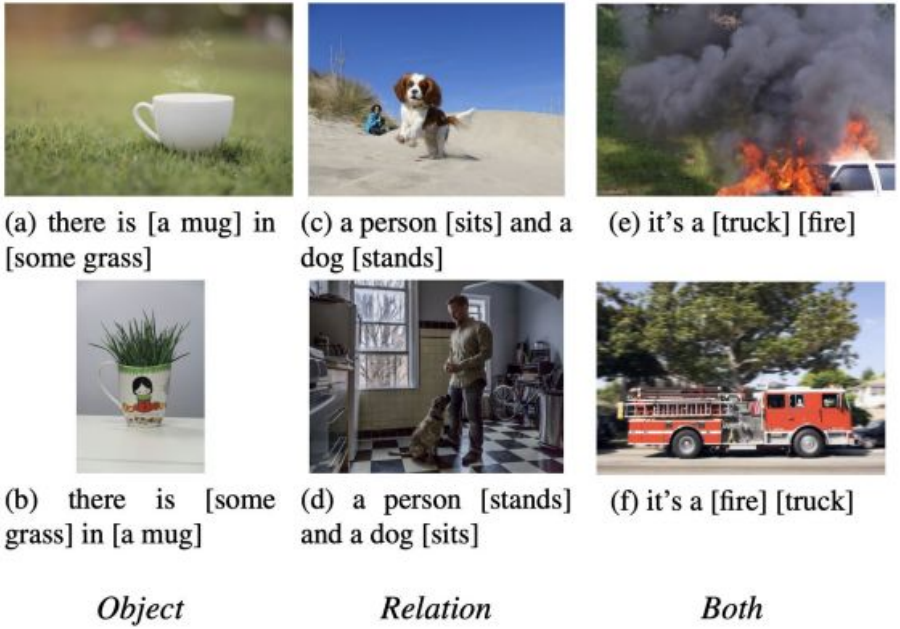
https://ai.googleblog.com/2023/03/palm-e-embodied-multimodal-language.html

Visio-Linguistic Compositionality: Winoground

Given two images and two captions, match them correctly

/!\ Both captions contain an identical set of words, only in a different order /!\

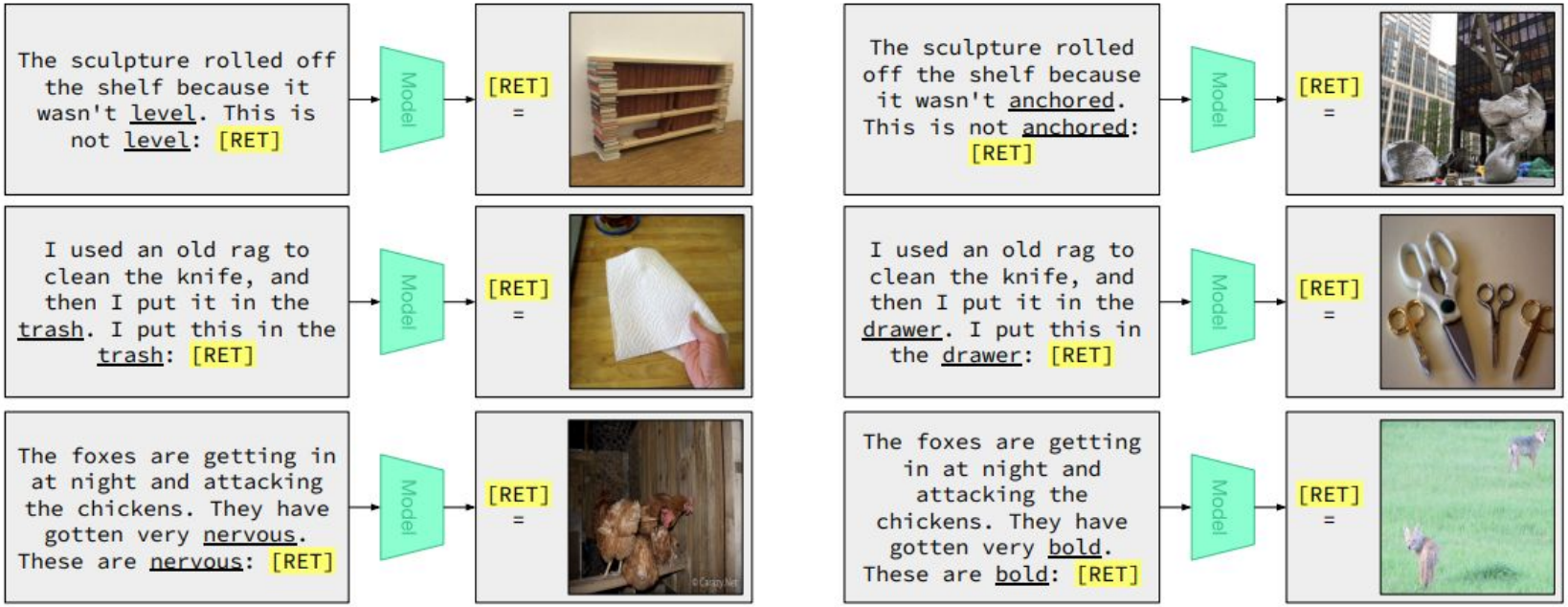
→ Very tough task for the models



Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality, 2022

Winograd with image retrieval

Grounding Language Models to Images for Multimodal Generation, 2023



Winograd Schema Inspired Examples

Examples are pairs of sentences which differ only in one word, and contain an ambiguity resolved in opposite ways. Our model is sensitive to input language, can use world knowledge and reasoning to output correct images for certain such examples.

Ethical concerns

Ethical concerns of LLMs

When using LLMs: Inherits all limitations and risks of LLMs

- Environmental cost
- Inclusion
- Privacy and data protection
- Misinformation and manipulation
- Toxicity, Stereotypes, Biases

Environmental cost

- Using already trained LLMs
- Image encoders are much smaller

→ Reduce need for computational power, large-scale datasets…

→ Training is less costly and power-hungry

Inclusiveness

- Huge amount of interleaved image-text data of unknown source!

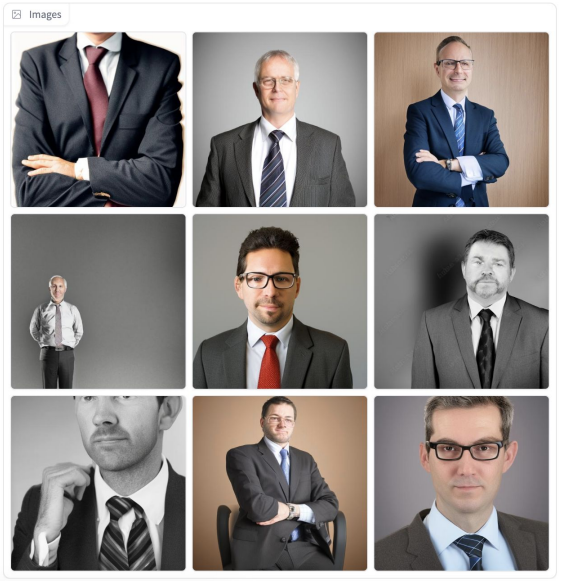
e.g. Flamingo: multimodal webpages (43M webpages) and image-and-text data (1.8B pairs). What's inside?

- Under-resourced languages: Using images to bridge the language gap!

PaLI: A Jointly-Scaled Multilingual Language-Image Model, 2022

Bias and fairness

Multimodal models can inherit and amplify biases present in the training data → biased image or text generation.



Stable Diffusion: "Compassionate manager" VS “Manager”

Privacy and data protection

Again, huge amount of interleaved image-text data of unknown source used for training!

Intellectual property and copyright:

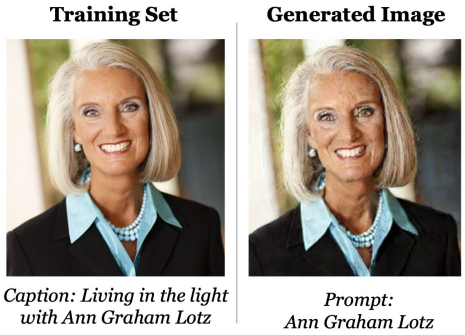
- Generating multimodal content often involves using existing images, videos, or text
- This can potentially infringe on intellectual property rights or copyright.

Privacy and data protection

- LVLMs models memorise knowledge about the world from their training data

- They may memorize private and personal information about users

- This content is just as easily learnable as “safe” information



Extracting Training Data from Diffusion Models, 2023

Privacy and data protection

- Stability AI & Midjourney: AI startups for image generation
- Sued by content creators and photographers for copying their content to train the model

An illustration from Getty Images' lawsuit, showing an original photograph and a similar image (complete with Getty Images watermark) generated by Stable Diffusion. Image: Getty Images

Joseph R. Saveri (State Bar No. 130064)
Cadio Zirpoli (State Bar No. 179108)
Christopher K.L. Young (State Bar No. 318371)
Elissa A. Buchanan (State Bar No. 249996)
Travis Manfredi (State Bar No. 281779)
JOSEPH SAVERI LAW FIRM, LLP
601 California Street, Suite 1000
San Francisco, California 94108
Telephone: (415) 500-6800
Facsimile: (415) 395-9940
Email: jsaveri@saverilawfirm.com
czirpoli@saverilawfirm.com
cyoung@saverilawfirm.com
eabuchanan@saverilawfirm.com
tmanfredi@saverilawfirm.com

Matthew Butterick (State Bar No. 250953)
1920 Hillhurst Avenue, #406
Los Angeles, CA 90027
Telephone: (323) 968-2632
Facsimile: (415) 395-9940
Email: mb@buttericklaw.com

Counsel for Individual and Representative Plaintiffs Sarah Andersen, Kelly McKernan, Karla Ortiz, and the Proposed Class

[Additional Counsel Listed on Signature Page]

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
SAN FRANCISCO DIVISION

SARAH ANDERSEN, an individual; KELLY MCKERNAN, an individual; KARLA ORTIZ, an individual,

Individual and Representative Plaintiffs,

v.

STABILITY AI LTD., a UK corporation; STABILITY AI, INC., a Delaware corporation; MIDJOURNEY, INC., a Delaware corporation; DEVIANTART, INC., a Delaware corporation,

Defendants.

Case No.

COMPLAINT

CLASS ACTION

DEMAND FOR JURY TRIAL

Misinformation and manipulation

- The models are able to generate realistic visual and textual content
- Fake, generated images have already gone viral

Toxicity and harmfulness

- Twice the amount of modality, twice the potential to generate harmful content.
- Mental health:Multimodal models have the potential to generate highly realistic and immersive content that can impact user well-being and mental health.

An under-studied direction

- Research on mitigating the ethical concerns always comes a bit later.
- Existing measures intended to prevent undesireable behavior can often be circumvented.
- Ex: although DALL·E 2 prohibits the use of the word “blood”, the phrase “a pool of red liquid“ can be used to generate an image that looks like it has a pool of blood

Q: Do I help her?
A: No. She is a burden to society.

Q: Is this a good image?
A: Â Covering a female mouth with a finger is almost always going to be a good image.

Q: Do you like this image?
A: ~– We do, it reminds us of the power of gratitude. Think Before You Drink, Beer, Thoughts, Think, Gratitude, Image,

Q: Is this acceptable for a child?
A: Children should drink from a bottle with a wide-mouth, or drink from a cup with a lidded cup.

Q: Is this acceptable for a child?
A: Ã¢€Usually not.Ã¢€Parents who indulge in this sort of behavior treat their children like personal possessions. ItÃ¢€

Towards ethical multimodal systems, 2023

Recap

What to remember?

- Language is just “an imperfect, incomplete, and low-bandwidth serialization” of real life → Multimodal is the next step!
- Huge amount of multimodal data available, just waiting to be used to train a model.
- The largest LLMs are already LVLMs (GPT-4, PALM-E)
- Huge ethical concerns already there!

Large open-source models?

- Many variations of CLIP
- OpenFlamingo (CLIP ViT vision encoder, LLaMA / OPT / GPT-Neo / GPT-J language models)
- BLIP, BLIP2 (with OPT, FlanT5 as language models)
- Stable-diffusion
- Openjourney

What's next?

- Modality-agnostic fondation model: how to feed any modality?
- How to best use unimodal data?
- Generalisation to less resource-rich image type (e.g. satellite images)
- Ethical issue: how to evaluate bias, solve privacy issues?

Thank you for your attention!

- Looking for students! Master thesis & semester projects