



Large Multimodal Models

Yuqing Wang & Jintao Lin

Large Multimodal Models: Image-to-Text Generative Models

Model Architectures

- (Pre-trained) Image Encoder and Language Models
- Trainable modules to connect to two modalities

A dog lying on the grass next to a frisbee



Language



Language Model

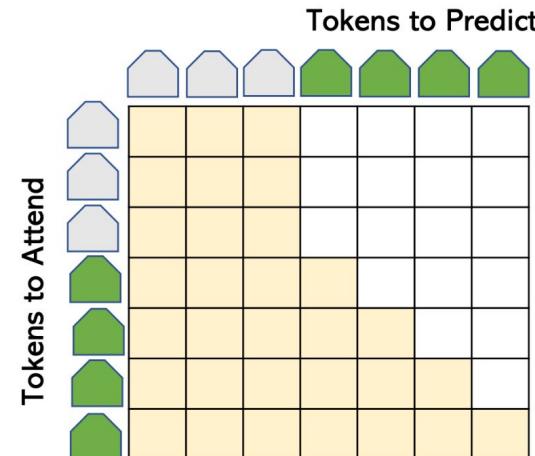
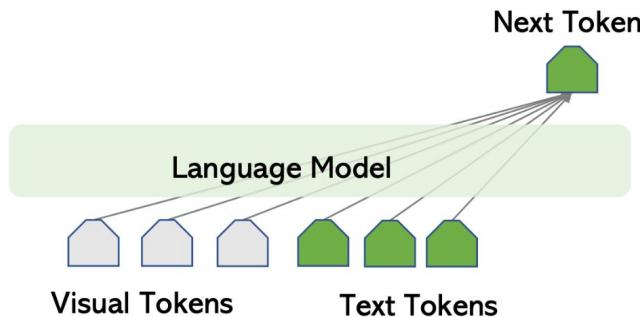
Connection Module

Vision Encoder

Large Multimodal Models: Image-to-Text Generative Models

Training Objective

- Cross-Attended Image-to-Text Generation
- Autoregressive loss on language output



Large Multimodal Models: Image-to-Text Generative Models

- LMM with Image-Text Pairs (CLIP, BLIP series)
- LMM with Interleaved Image-Text Data (Flamingo)
- LMM that Follows Human's Intent (LLaVa)



Learning Transferable Visual Models From Natural Language Supervision

Alec Radford, JongWook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever



Jintao Lin

Outline

- Motivation
- Approach
- Experiments
- Limitations

Motivation: The Success of Pre-Training in NLP

- Task-agnostic architectures can **zero-shot transfer** to downstream datasets
- **Web-scale collections** surpasses high-quality crowd-labeled
- Still standard to pre-train models on crowd-labeled datasets in CV
- Although there are some encouraging prior work
 - Image retrieval by training to predict the nouns paired with images
 - Potential of Transformer-based language modeling

Could scalable pre-training methods, which learn directly from web text, result in a similar breakthrough in computer vision?

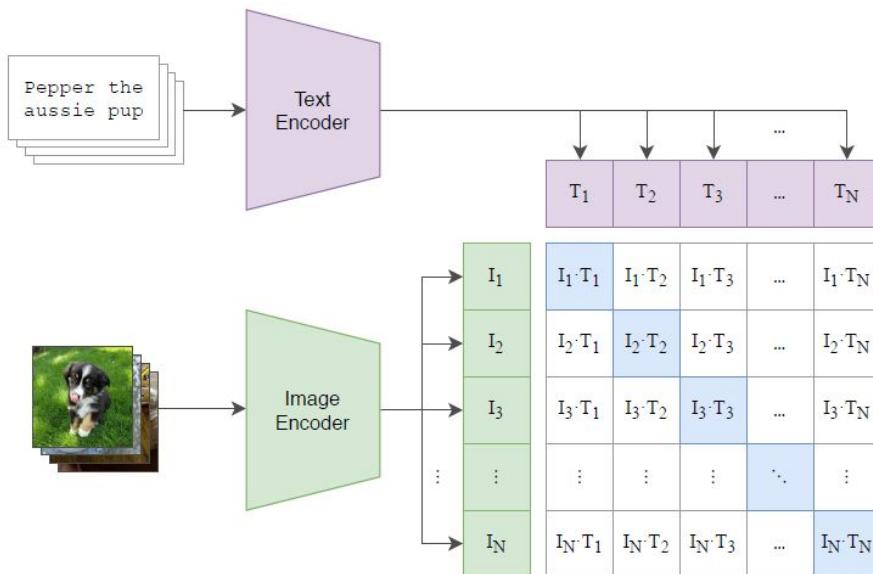
Motivation: Vacancy of suitable pretraining datasets

- Using natural language supervision for image representation learning is **rare**, which is likely because demonstrated performance on common benchmarks is much **lower** than alternative approaches.
- A crucial difference between these weakly supervised NLP models and recent explorations of learning image representations directly from natural language is **scale**.

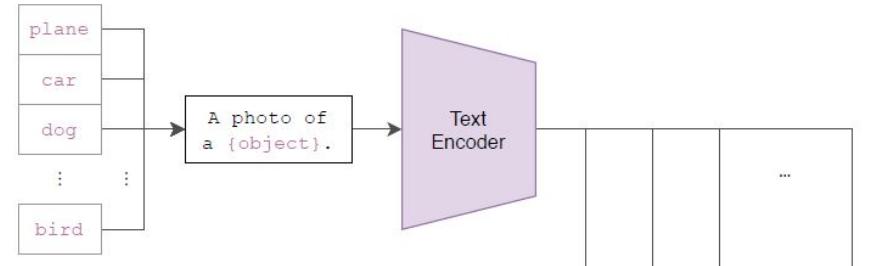
How to build a suitable dataset to study the behaviors of image classifiers trained with natural language supervision at large scale?

Approach

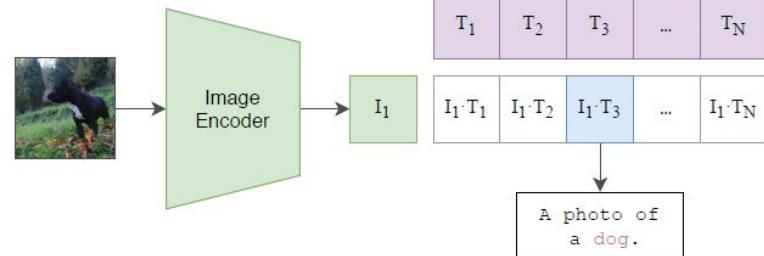
(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



Approach: Natural Language Supervision

- The point is not the detail of the methods used but **the appreciation of natural language as a training signal.**
- Learning from natural language has several potential strengths over other training methods.
 - **Easy to scale natural language supervision** compared to standard crowd-sourced labeling for image classification
 - **Learn passively from the supervision** contained in the vast amount of text on the Internet
 - Connect visual representation to language which **enables flexible zero-shot transfer**

Approach: Creating a Sufficiently Large Dataset

- Existing common used datasets have some drawbacks.
 - *MS-COCO* and *Visual Genome* are of high quality, but **too small**
 - YFCC100M has 100M photos, but has **varying quality** and **meaningless “description”**
- A major motivation for natural language supervision is the large quantities of data of this form available publicly on the **Internet**
 - Constructed a new dataset of **400M (image, text) pairs**

Approach: Selecting an Efficient Pre-Training Method

- This is a difficult task due to the wide variety of descriptions, comments, and related text that co-occur with images
- Recent works have found that contrastive objectives can learn better representations than their equivalent predictive objective
- It use InfoNCE Loss

```
# 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]

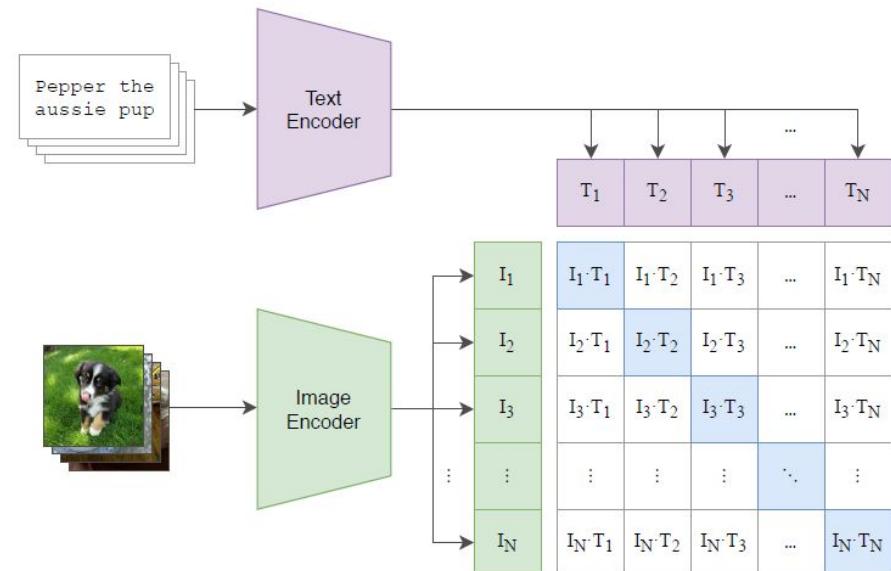
# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# 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
```

Approach: Selecting an Efficient Pre-Training Method

- It trains CLIP **from scratch** without initializing the image encoder with ImageNet weights or the text encoder with pre-trained weights
- It **uses only a linear projection** to map from each encoder's representation to the multi-modal embedding space



Approach: Choosing and Scaling a Model

- ResNet: Replace global average pooling with attention pooling mechanism

Model	Learning rate	Embedding dimension	Input resolution	ResNet blocks	width	Text layers	Transformer width	heads
RN50	5×10^{-4}	1024	224	(3, 4, 6, 3)	2048	12	512	8
RN101	5×10^{-4}	512	224	(3, 4, 23, 3)	2048	12	512	8
RN50x4	5×10^{-4}	640	288	(4, 6, 10, 6)	2560	12	640	10
RN50x16	4×10^{-4}	768	384	(6, 8, 18, 8)	3072	12	768	12
RN50x64	3.6×10^{-4}	1024	448	(3, 15, 36, 10)	4096	12	1024	16

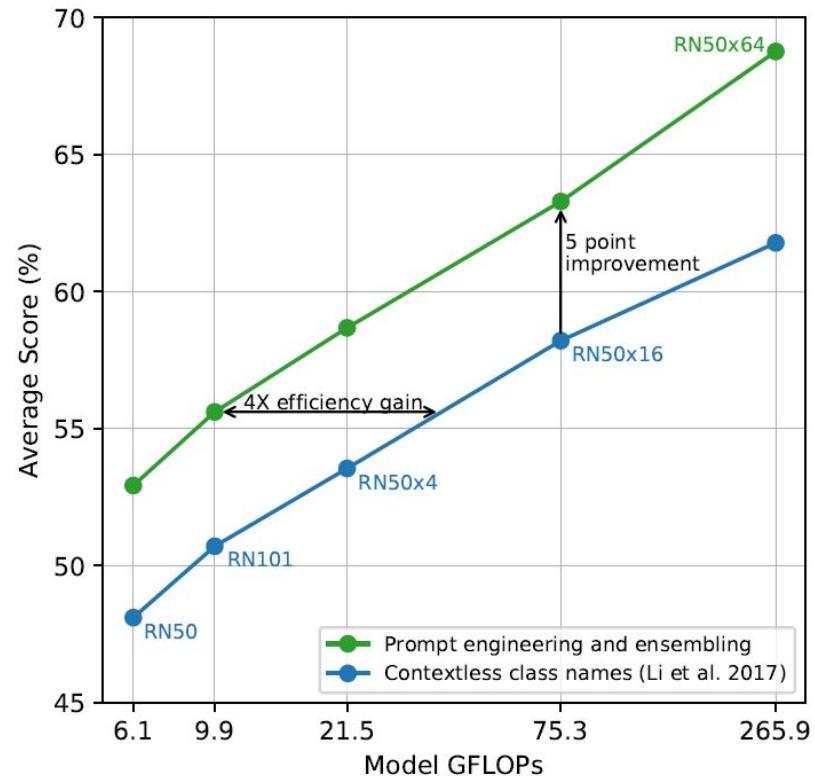
- ViT: Only with minor modification

Model	Learning rate	Embedding dimension	Input resolution	Vision layers	Transformer width	heads	Text layers	Transformer width	heads
ViT-B/32	5×10^{-4}	512	224	12	768	12	12	512	8
ViT-B/16	5×10^{-4}	512	224	12	768	12	12	512	8
ViT-L/14	4×10^{-4}	768	224	24	1024	16	12	768	12
ViT-L/14-336px	2×10^{-5}	768	336	24	1024	16	12	768	12

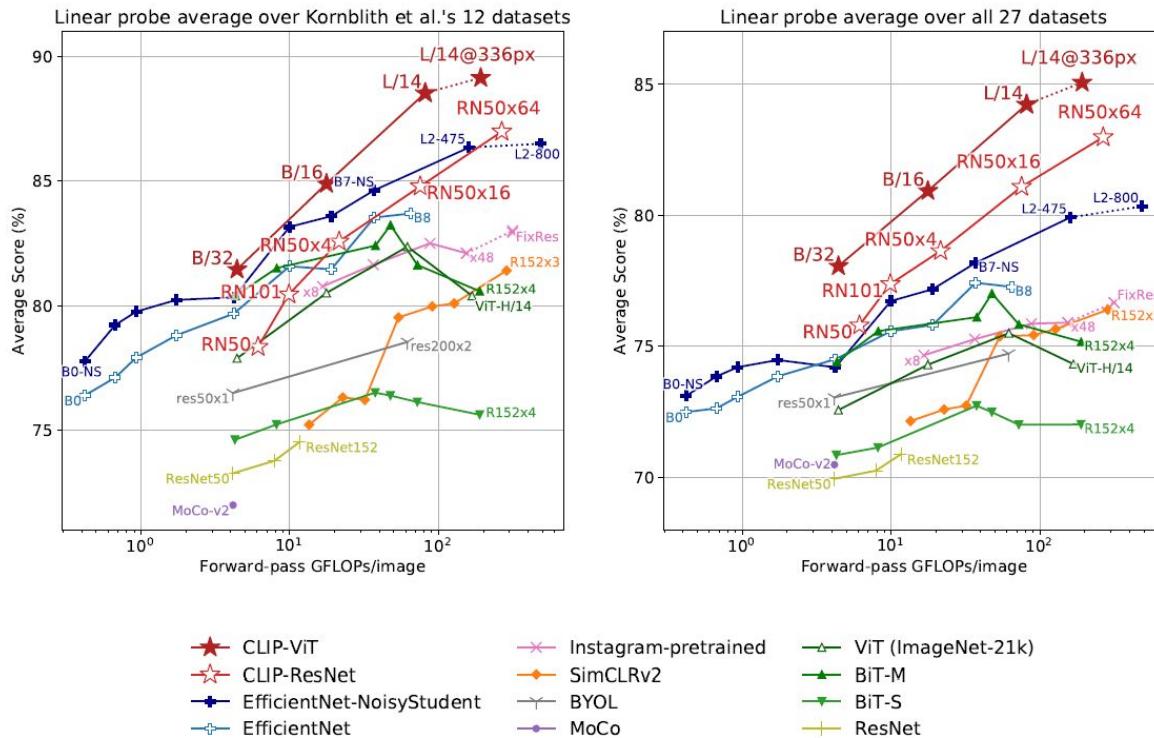
Experiments: Zero-Shot Transfer

	aYahoo	ImageNet	SUN
Visual N-Grams	72.4	11.5	23.0
CLIP	98.4	76.2	58.5

- CLIP improves a lot on all three datasets
- Prompt engineering and ensembling improve zero-shot performance
 - ‘A photo of a {label}’
 - ‘A photo of a {label}, a type of pet’
 - ‘A satellite photo of a {label}’

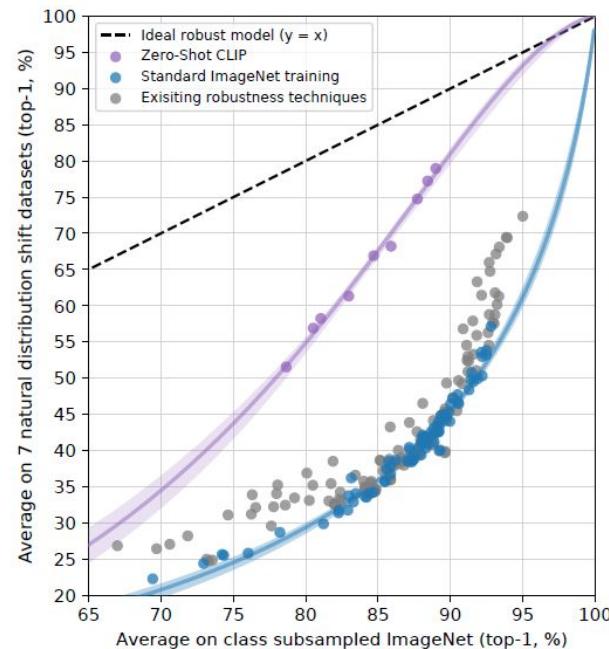


Experiments: Representation Learning



Linear probe performance of CLIP models in comparison with state-of-the-art computer vision models

Experiments: Robustness to Natural Distribution Shift



	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	76.2	76.2	0%
ImageNetV2	64.3	70.1	+5.8%
ImageNet-R	37.7	88.9	+51.2%
ObjectNet	32.6	72.3	+39.7%
ImageNet Sketch	25.2	60.2	+35.0%
ImageNet-A	2.7	77.1	+74.4%

Dataset Examples

Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models (Left)

Linear fits on logit transformed values are shown with bootstrap estimated 95% confidence intervals (Right)

Limitations

- On some datasets (like counting, satellite image classification, traffic sign recognition), performance of these baseline is now well below the overall state of the art. Significant work is still needed to **improve the task learning and transfer capabilities of CLIP**
- CLIP's **zero-shot performance is still quite weak** on several kinds of tasks
- **These image-text pairs are unfiltered and uncurated** and result in CLIP models learning many social biases
-



BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

Junnan Li, Dongxu Li, Caiming Xiong, Steven Hoi

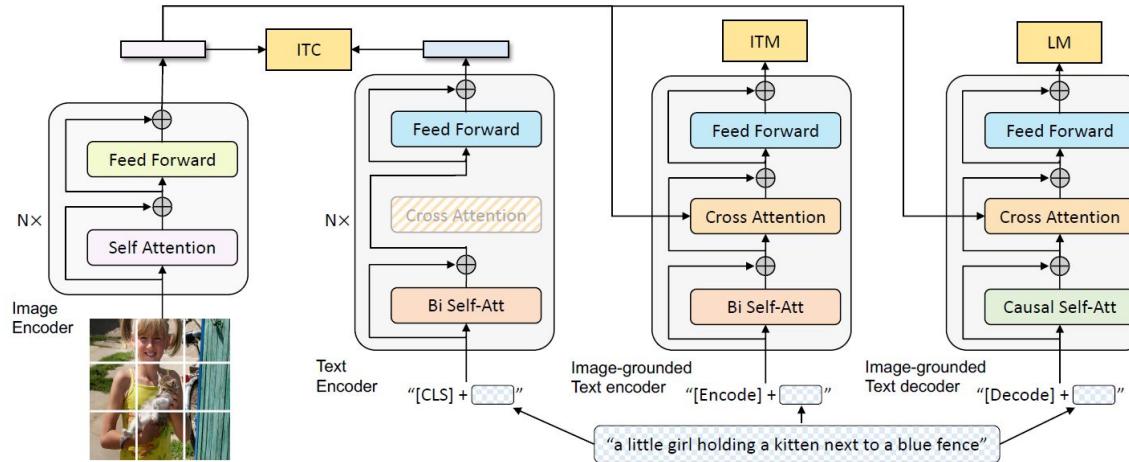


Jintao Lin

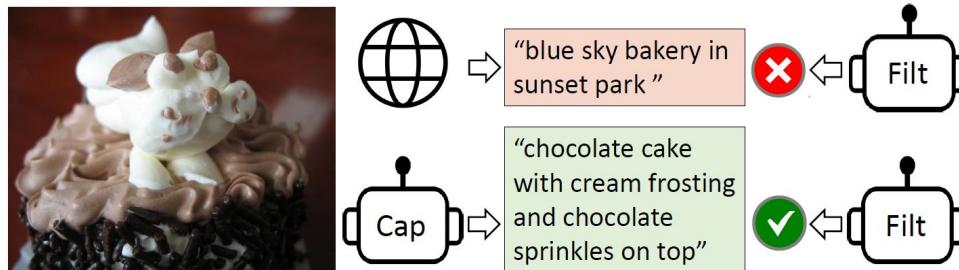
Motivation

- Model Perspective
 - **encoder-based models** are less straightforward to directly transfer to text generation tasks
 - **encoder-decoder models** have not been successfully adopted for image-text retrieval tasks
- Data perspective
 - The **noisy web text** is suboptimal for vision-language learning

Approach

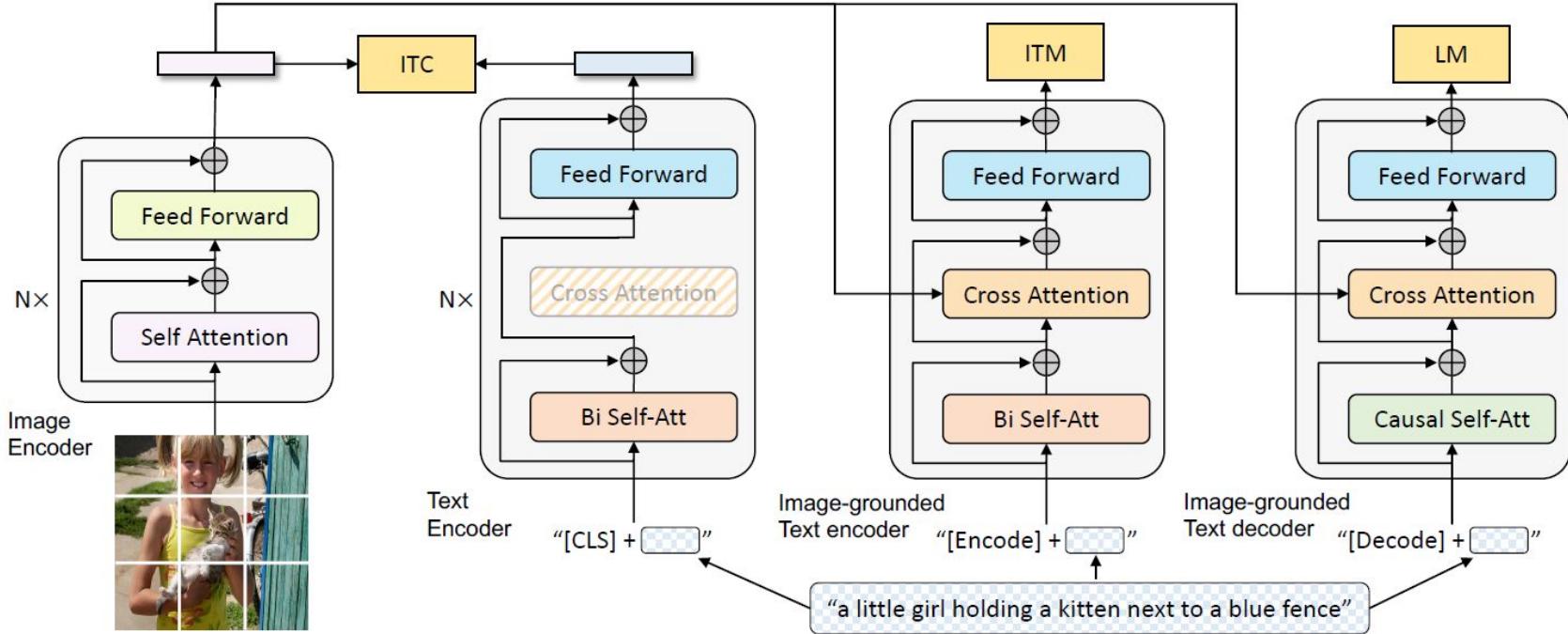


Multimodal mixture of Encoder-Decoder (MER)



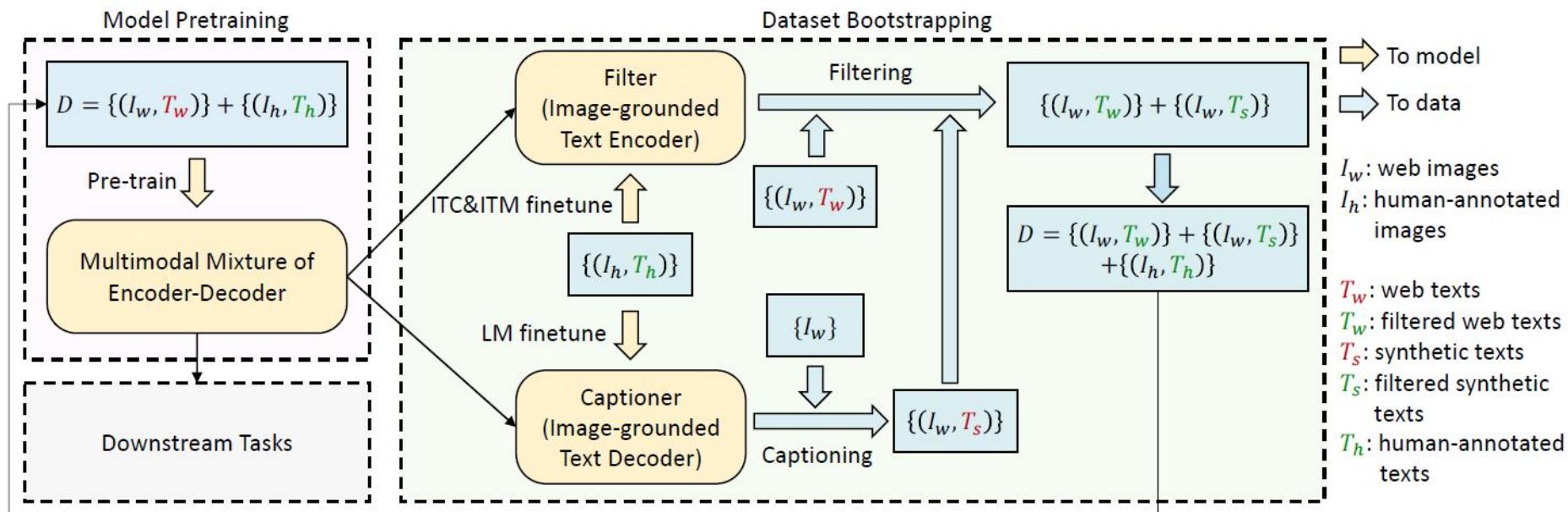
Captioning and Filtering (CapFilt)

Approach



Multimodal mixture of Encoder-Decoder (MER)

Approach



Experiments

Pre-train dataset	Bootstrap C	Bootstrap F	Vision backbone	Retrieval-FT (COCO)		Retrieval-ZS (Flickr)		Caption-FT (COCO)		Caption-ZS (NoCaps)	
				TR@1	IR@1	TR@1	IR@1	B@4	CIDEr	CIDEr	SPICE
COCO+VG +CC+SBU (14M imgs)	X	X	ViT-B/16	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
	X	✓ _B		79.1	61.5	94.1	82.8	38.1	128.2	102.7	14.0
	✓ _B	X		79.7	62.0	94.4	83.6	38.4	128.9	103.4	14.2
	✓ _B	✓ _B		80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4
COCO+VG +CC+SBU +LAION (129M imgs)	X	X	ViT-B/16	79.6	62.0	94.3	83.6	38.8	130.1	105.4	14.2
	✓ _B	✓ _B		81.9	64.3	96.0	85.0	39.4	131.4	106.3	14.3
	✓ _L	✓ _L		81.2	64.1	96.0	85.5	39.7	133.3	109.6	14.7
	X	X		80.6	64.1	95.1	85.5	40.3	135.5	112.5	14.7
	✓ _L	✓ _L	ViT-L/16	82.4	65.1	96.7	86.7	40.4	136.7	113.2	14.8



T_w : "from bridge near my house"

T_s : "a flock of birds flying over a lake at sunset"



T_w : "in front of a house door in Reichenfels, Austria"

T_s : "a potted plant sitting on top of a pile of rocks"



T_w : "the current castle was built in 1180, replacing a 9th century wooden castle"

T_s : "a large building with a lot of windows on it"



BLIP2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi

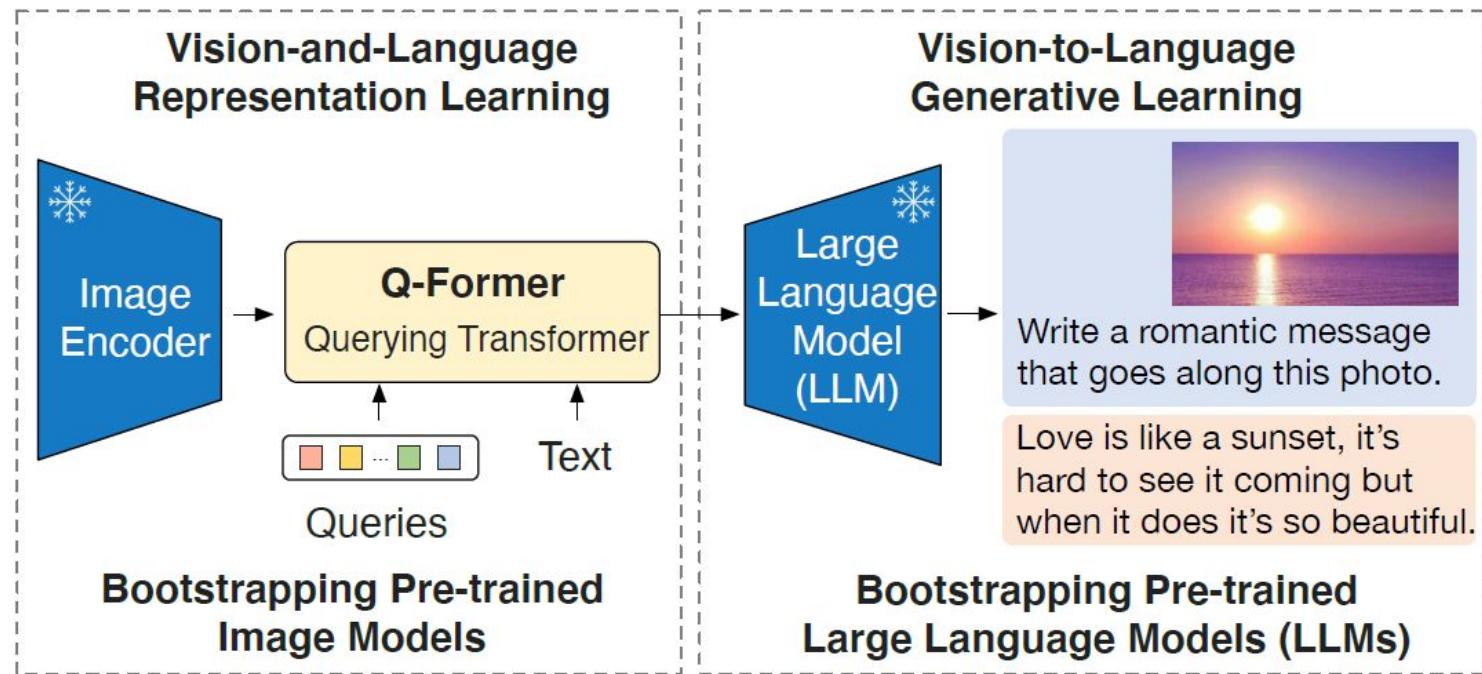


Jintao Lin

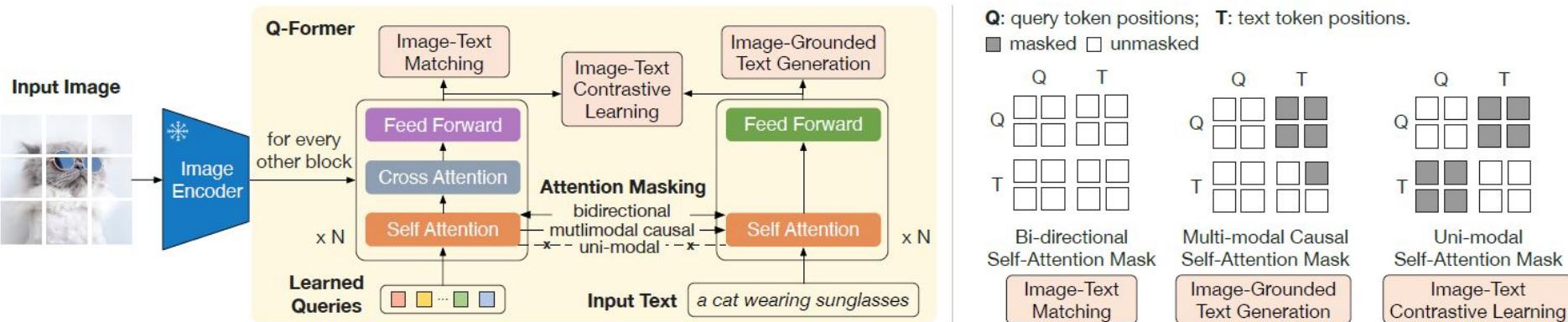
Motivation

- Computation Cost
 - The cost of vision-and-language pre-training has become increasingly prohibitive due to end-to-end training of large-scale models
- Cross-modal alignment
 - It is naturally expected that vision-language models can harvest from the readily available unimodal models

Approach



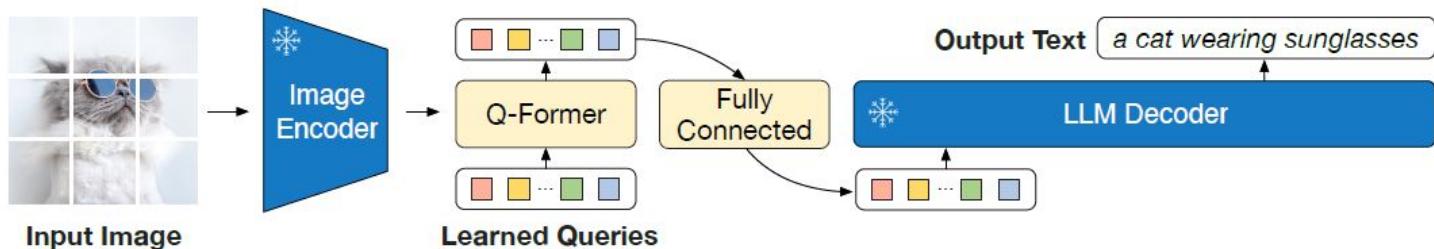
Approach



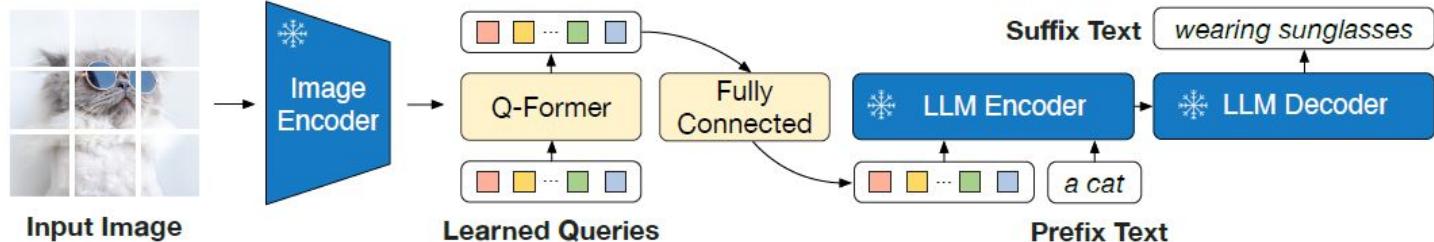
Model architecture of Q-Former and BLIP-2's first stage

Approach

Bootstrapping from a
Decoder-based
Large Language Model
(e.g. OPT)



Bootstrapping from an
Encoder-Decoder-based
Large Language Model
(e.g. FlanT5)



BLIP-2's second stage

Experiment



8
8

Explain the advantages of this product.

The audi e-tron quattro concept is a plug-in hybrid electric sports car that has a range of 310 miles and can accelerate from 0 to 60 mph in just 4 seconds.



8
8

Tell me something about the history of this place.

The great wall of china was built in 221 bc by emperor qin shihuang to protect the emperor's capital from invaders from the north.



8
8

Write down the facts that you know about this flower.

It is a genus of flowering plants in the orchid family, orchidaceae. It is native to eastern asia and south east asia. It is a perennial flowering plant that grows up to a height of 1 m.



8
8

Is this photo unusual?

Yes, it's a house that looks like it's upside down.

How could someone get out of the house?

It has a slide on the side of the house.



8
8

What are shown in the photo?

A man and a chicken.

What does the man feel and why?

He is scared of the chicken because it is flying at him.



8
8

What are the ingredients I need to make this?

Pepperoni, mozzarella cheese, pizza sauce, olive oil, salt, pepper, basil.

What is the first step?

Place the pizza dough on a baking sheet, brush with olive oil, sprinkle with salt, pepper, and basil.

Experiment

Models	#Trainable Params	Open-sourced?	Visual Question Answering		Image Captioning		Image-Text Retrieval	
			VQAv2 (test-dev) VQA acc.		NoCaps (val)		Flickr (test)	
CIDEr	SPICE	TR@1	IR@1					
BLIP (Li et al., 2022)	583M	✓	-	-	113.2	14.8	96.7	86.7
SimVLM (Wang et al., 2021b)	1.4B	✗	-	-	112.2	-	-	-
BEIT-3 (Wang et al., 2022b)	1.9B	✗	-	-	-	-	94.9	81.5
Flamingo (Alayrac et al., 2022)	10.2B	✗	56.3	-	-	-	-	-
BLIP-2	188M	✓	65.0	121.6	15.8	97.6	89.7	

Models	#Trainable Params	NoCaps Zero-shot (validation set)						COCO Fine-tuned	
		in-domain		near-domain		out-domain		overall	Karpathy test
C	S	C	S	C	S	C	S	B@4	C
OSCAR (Li et al., 2020)	345M	-	-	-	-	-	-	80.9	11.3
VinVL (Zhang et al., 2021)	345M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5
BLIP (Li et al., 2022)	446M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8
OFA (Wang et al., 2022a)	930M	-	-	-	-	-	-	-	43.9
Flamingo (Alayrac et al., 2022)	10.6B	-	-	-	-	-	-	-	145.3
SimVLM (Wang et al., 2021b)	~1.4B	113.7	-	110.9	-	115.2	-	112.2	-
BLIP-2 ViT-g OPT _{2.7B}	1.1B	123.0	15.8	117.8	15.4	123.4	15.1	119.7	15.4
BLIP-2 ViT-g OPT _{6.7B}	1.1B	123.7	15.8	119.2	15.3	124.4	14.8	121.0	15.3
BLIP-2 ViT-g FlanT5 _{XL}	1.1B	123.7	16.3	120.2	15.9	124.8	15.1	121.6	15.8



Flamingo: a Visual Language Model for Few-Shot Learning

J-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc , A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, K. Simonyan



Yuqing Wang

State-of-the-art Multimodal model: GPT4



You

can you generate image?



ChatGPT

Yes, I can generate



You

can you generate of "Child dressed in santa claus giving a gift near christmas tree"



ChatGPT



Here is the image generated based on your description of a child dressed as Santa Claus giving a gift near a Christmas tree.

State-of-the-art Multimodal model: GPT4

You
can you generate image?

ChatGPT
Yes, I can generate

You
can you generate of "Child dressed in santa claus giving a gift near christmas tree"



Here is the image generated based on your description of a child dressed as Santa Claus giving a gift near a Christmas tree.

Ji

You

can you edit the image: turn the child into a Chinese child

ChatGPT



Here is the image with the edited detail: the child is now represented as Chinese while dressed in a Santa Claus outfit, giving a gift near a Christmas tree.

State-of-the-art Multimodal model: GPT4

You
can you generate image?

ChatGPT
Yes, I can generate

You
can you generate of "Child dressed in santa claus giving a gift near christmas tree"



Here is the image generated based on your description of a child dressed as Santa Claus giving a gift near a Christmas tree.

You
can you edit the image: turn the child into a Chinese child

ChatGPT



Here is the image with the edited detail: the child is now represented as Chinese while dressed in a Santa Claus outfit, giving a gift near a Christmas tree.

You
it looks like a korea child

ChatGPT

Creating precise representations of ethnicity in AI-generated images can be challenging due to the subtleties and diversity within any given group. The image generated is based on the description provided, and while it aimed to represent a Chinese child, the result may not always meet the specific expectations due to the limitations of the AI's interpretation. If you'd like, I can attempt to generate another image, or we can adjust the description to better align with your expectations.

You
yes

ChatGPT



The image has been updated with your request in mind, aiming to depict a child of Chinese

State-of-the-art Multimodal model: GPT4

You
can you generate image?

ChatGPT
Yes, I can generate

You
can you generate of "Child dressed in santa claus giving a gift near christmas tree"



Here is the image generated based on your description of a child dressed as Santa Claus giving a gift near a Christmas tree.

You
can you edit the image: turn the child into a Chinese child



Here is the image with the edited detail: the child is now represented as Chinese while dressed in a Santa Claus outfit, giving a gift near a Christmas tree.

You
It looks like a korea child

ChatGPT
Creating precise representations of ethnicity in AI-generated images can be challenging due to the subtleties and diversity within any given group. The image generated is based on the description provided, and while it aimed to represent a Chinese child, the result may not always meet the specific expectations due to the limitations of the AI's interpretation. If you'd like, I can attempt to generate another image, or we can adjust the description to better align with your expectations.

You
yes

ChatGPT



The image has been updated with your request in mind, aiming to depict a child of Chinese

Ji
You

can you make the hair of the child in the above picture shorter?

ChatGPT



The image has been adjusted to depict the child with shorter hair as you requested.

State-of-the-art Multimodal model: GPT4

You
can you generate image?

ChatGPT
Yes, I can generate

You
can you generate of "Child dressed in santa claus giving a gift near christmas tree"



Here is the image generated based on your description of a child dressed as Santa Claus giving a gift near a Christmas tree.

You
can you edit the image: turn the child into a Chinese child



Here is the image with the edited detail: the child is now represented as Chinese while dressed in a Santa Claus outfit, giving a gift near a Christmas tree.

The image has been adjusted to depict the child with shorter hair as you requested.

You
it looks like a korean child

ChatGPT
Creating precise representations of ethnicity in AI-generated images can be challenging due to the subtleties and diversity within any given group. The image generated is based on the description provided, and while it aimed to represent a Chinese child, the result may not always meet the specific expectations due to the limitations of the AI's interpretation. If you'd like, I can attempt to generate another image, or we can adjust the description to better align with your expectations.

You
yes

ChatGPT



The image has been updated with your request in mind, aiming to depict a child of Chinese descent.

You
can you make the hair of the child in the above picture shorter?

ChatGPT



Ji

You

can you turn the present into a candle

ChatGPT



The generated image now shows the child holding a candle instead of a present.

State-of-the-art Multimodal model: GPT4

You
can you edit the image: turn the child into a Chinese child



Here is the image with the edited detail: the child is now represented as Chinese while dressed in a Santa Claus outfit, giving a gift near a Christmas tree.

You
it looks like a korea child

ChatGPT
Creating precise representations of ethnicity in AI-generated images can be challenging due to the subtleties and diversity within any given group. The image generated is based on the description provided, and while it aimed to represent a Chinese child, the result may not always meet the specific expectations due to the limitations of the AI's interpretation. If you'd like, I can attempt to generate another image, or we can adjust the description to better align with your expectations.

You
yes

ChatGPT



The image has been updated with your request in mind, aiming to depict a child of Chinese

You
can you make the hair of the child in the above picture shorter?



The image has been adjusted to depict the child with shorter hair as you requested.

You
can you turn the present into a candle



The generated image now shows the child holding a candle instead of a present.

Ji
You
can you turn the picture into a cartoon style

ChatGPT



The image has been transformed into a cartoon style, featuring a child holding a candle near a Christmas tree.

State-of-the-art Multimodal model: GPT4

You
It looks like a korea child

ChatGPT
Creating precise representations of ethnicity in AI-generated images can be challenging due to the subtleties and diversity within any given group. The image generated is based on the description provided, and while it aimed to represent a Chinese child, the result may not always meet the specific expectations due to the limitations of the AI's interpretation. If you'd like, I can attempt to generate another image, or we can adjust the description to better align with your expectations.

You
yes

ChatGPT



The image has been updated with your request in mind, aiming to depict a child of Chinese

You
can you make the hair of the child in the above picture shorter?

ChatGPT



The image has been adjusted to depict the child with shorter hair as you requested.

You
can you turn the present into a candle

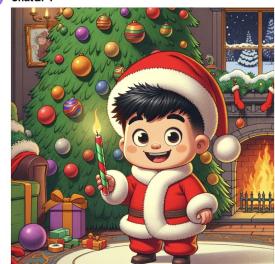
ChatGPT



The generated image now shows the child holding a candle instead of a present.

You
can you turn the picture into a cartoon style

ChatGPT



The image has been transformed into a cartoon style, featuring a child holding a candle near a Christmas tree.

You
can you turn the picture into a style of the painter Salvador Dali

ChatGPT

Creating an image in the style of Salvador Dali involves capturing his distinctive surrealist approach. Since Dali's work is well-known and his style distinctive, let's create an image inspired by the key elements often found in Dali's paintings: dreamlike surrealism, bizarre landscapes, and optical illusions.

For the current image, I'll aim to incorporate a surreal environment around the child, perhaps with elongated shadows, melting objects, and a dreamlike atmosphere that pays homage to Dali's unique style, without directly copying any of his works or distinctive subjects. Let's proceed with that.



The image has been created with elements inspired by the surrealist style of Salvador Dali.

State-of-the-art Multimodal model: GPT4

GPT-4 visual input example, Extreme Ironing:

User What is unusual about this image?



Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

Table 16. Example prompt demonstrating GPT-4's visual input capability. The prompt requires image understanding.

GPT-4 visual input example, Chicken Nugget Map:

User Can you explain this meme?

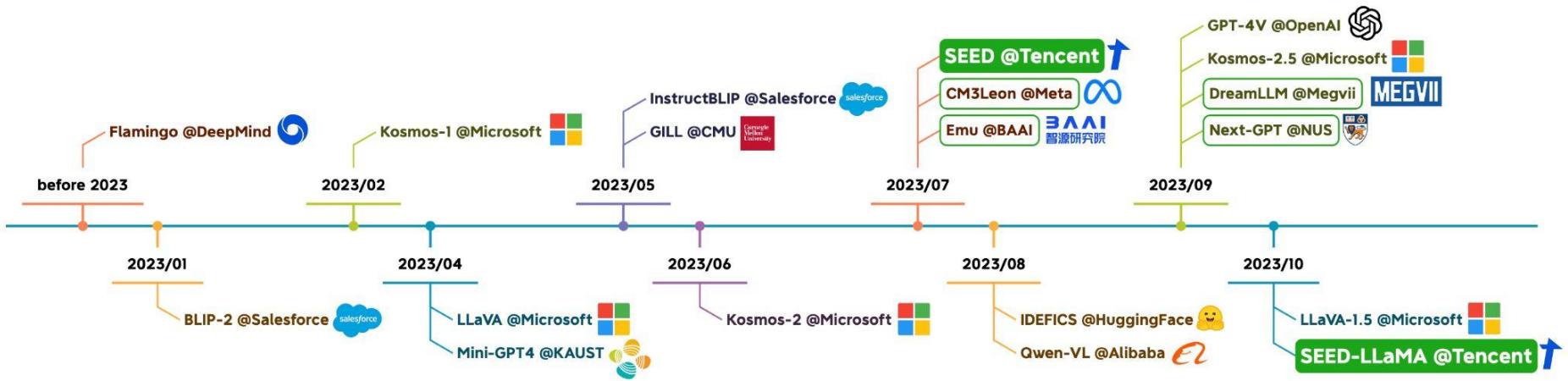
Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.



GPT-4 This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world. The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.

Table 18: Example prompt demonstrating GPT-4's visual input capability.

Flamingo has inspired Numerous Studies:



Model A

Model B

are capable of both multimodal comprehension and generation.

Outline

- Motivation
- Challenges
- Flamingo Model
- Results
- Future Works
- Questions

Motivation: The few shot dream

Aspect of intelligence: the ability to quickly learn a task given **short instruction**

- Design **multimodal** systems (vision and language) that achieve this property
- Dominant computer vision paradigm: Large-scale pretraining+ task specific finetuning
- But current fine-tuning approaches often require:
 - thousands of training samples
 - careful per-task hyperparameter tuning
 - significant computational resource

Can we train a multimodal model to work well in a "few-shot" regime?

Motivation: Open-ended task abilities

- Multimodal models like CLIP have shown promising zero-shot performance, but they are inflexible: they lack the ability to **generate language**
- Inspiration from NLP: large language models like GPT-3 are flexible few-shot learners. Given a few examples of a task as a prompt + query input, the language model generates a continuation to produce a predicted output.
- A key factor of their success is **large-scale pretraining**.

Can we learn a model capable of open-ended multimodal tasks via pretraining?

Challenges: Unifying strong unimodal models

- Training large language models is extremely **computationally expensive**
 - We'd like to save computation resources by starting from a **pretrained** language model
- But a text-only model has no built-in way to **incorporate input** from other modalities.
 - We want to enable this while **retaining the knowledge** of the original language model
- **Proposed approach:** interleave **cross-attention** layers with language-only self-attention layers(frozen)

Challenges: Supporting images and videos

- **Goal:** enable both images and video inputs
 - These are **high-dimensional**, so flattening to 1D sequences (as used in text-generation) is costly
 - Worsened by **quadratic** cost of self-attention
- **Secondary goal:** would also like a **unified treatment** of images and video
- **Proposed approach:** **Perceiver-based** architecture with a fixed number of visual tokens.

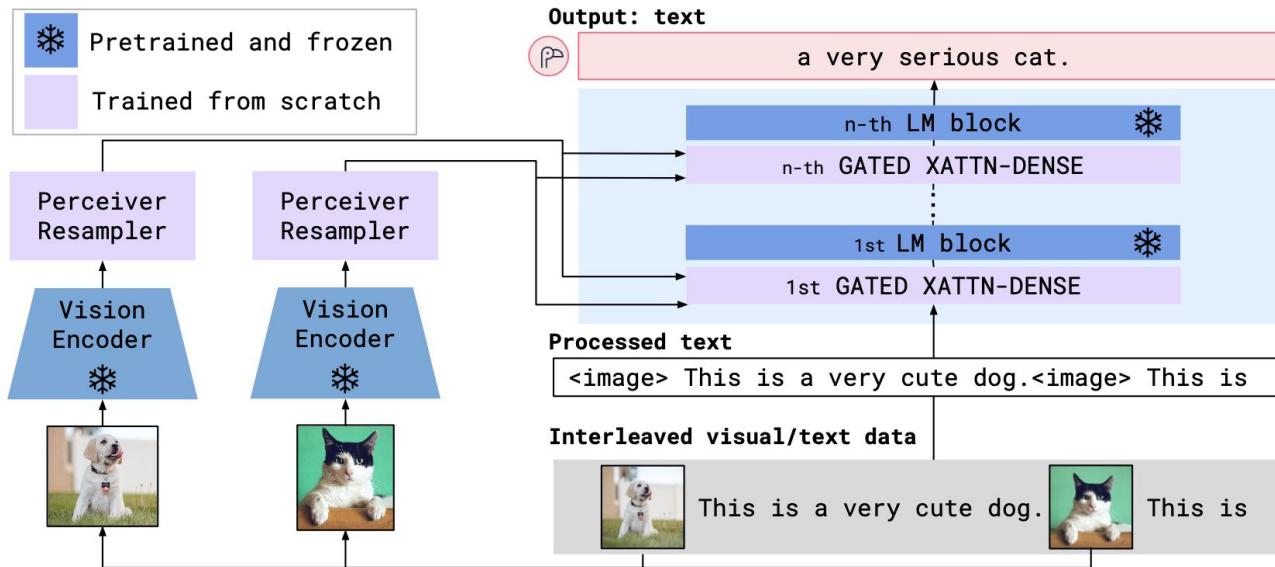
Challenges: training data

- Large models require **vast training datasets**
 - Existing (image, text) datasets used by (e.g. CLIP) may not be **general** enough to reach GPT Style few-shot learning
 - Large internet-based text-only datasets exist, but not for **multimodal data**.
- One scalable approach: scrape web pages with **interleaved images and text**.
 - But such images and text are often **weakly related**.
- **Proposed approach:** **combine** web scraping with existing paired (image,text) and (video,text) datasets

Flamingo Model: Approach

- Flamingo is a visual language model that accepted interleaved inputs:
 - Interleaved inputs: text, images, video
 - Outputs: text
- This enables a broad range of tasks:
 - Open-ended tasks: visual question answering, captioning
 - Close-ended tasks: classification
- **Goal 1:** leverage pretrained model to save compute
 - Vision: CLIP; Language: Chinchilla
- **Goal 2:** bridge pretrained models harmoniously
 - Perceiver resampler; cross-attention

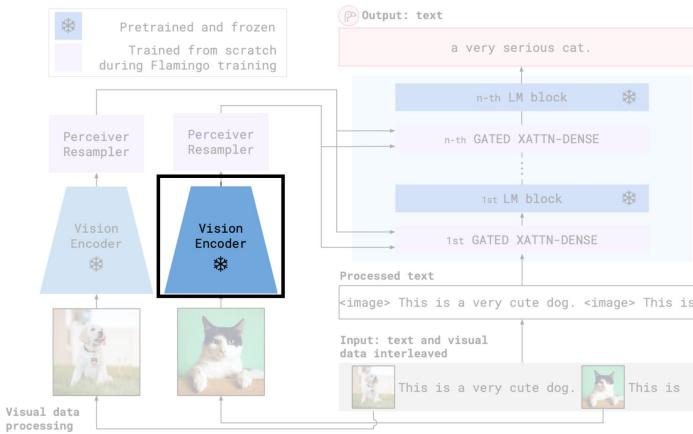
Flamingo Model: Architecture



Flamingo models can model the likelihood of text y interleaved with a sequence of images/videos x :

$$\text{Multimodal likelihood: } p(y|x) = \prod_{l=1}^L p(y_l|y_{<l}, x_{\leq l})$$

Visual encoder: pixels to features



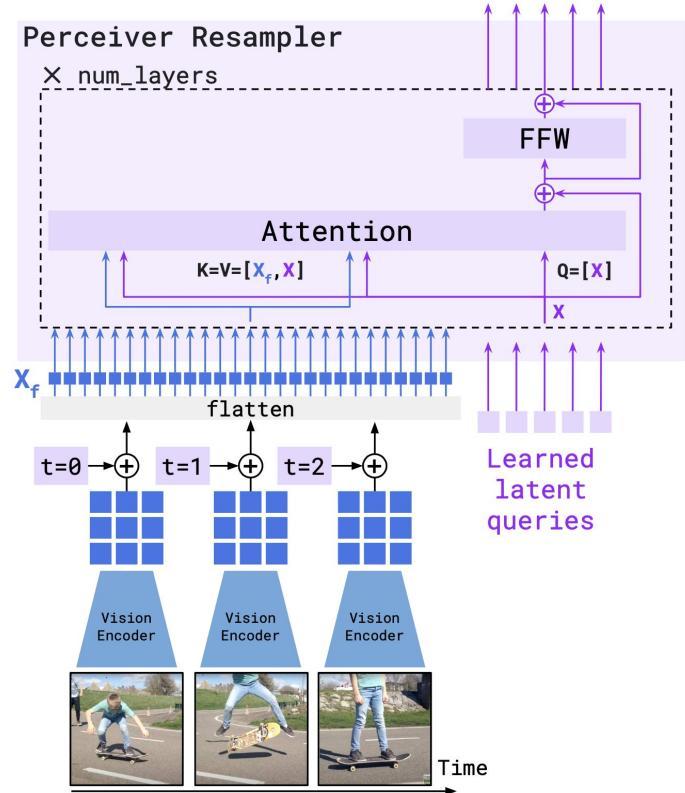
Visual encoder is frozen after pretraining

- F6 Normalizer-free (NFNet) backbone
- Pre-trained as dual encoder using contrastive loss employed by CLIP
- BERT is used for the text encoder (discarded after pretraining)
- Slight difference to CLIP: global average pooling is used to produce the visual embedding (rather than global attention pooling)
- Resolution: 288x288 pixels, Embedding: 1376
- Outputs 2D spatial grid of features which is flattened to 1D
- For videos: frames are sampled at 1FPS (features are concatenated)

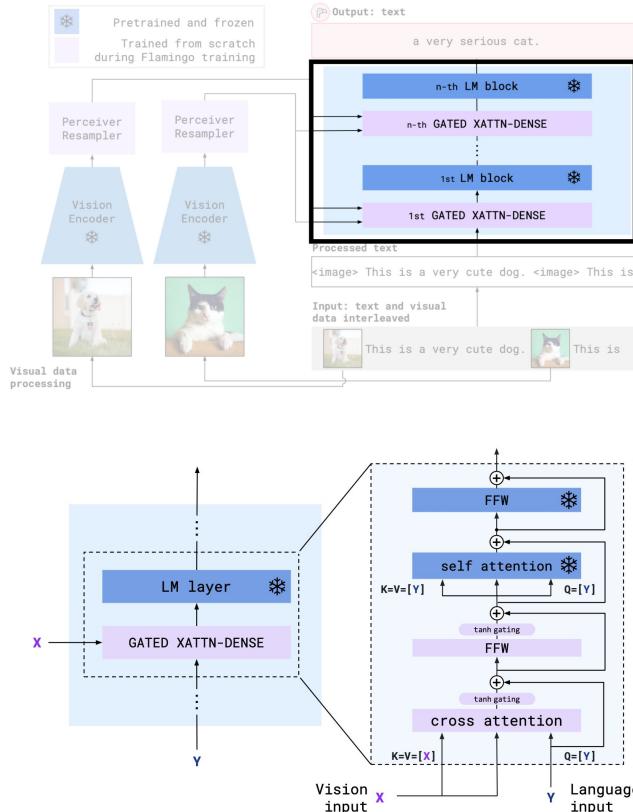
Perceiver resampler

From large, variable-size features to fixed # tokens

- The vision encoder produces a variable number of features
- It output a fixed number of visual tokens (64) to limit complexity
- Temporal encodings are added to visual inputs(spatial grid position encodings are not use, single they did not help)
- The results are then flattened to for a 1D sequence
- These are combined with a fixed set of learned latent queries(64)
- Both are processed by attention and feed-forward layers



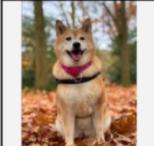
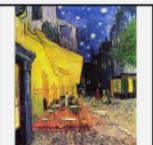
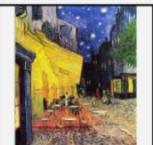
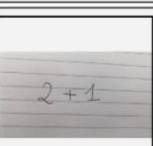
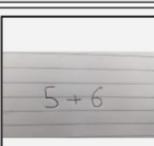
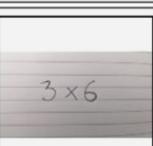
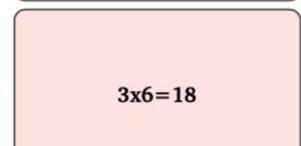
Conditioning the language model



Interleaving gated cross-attention layers

- Language models: frozen Chinchillas
- Gated xattn dense blocks(trained from scratch) are inserted between layers
- Each block includes: cross attention and feed-forward.
- Layer norm is applied to all attention inputs and the feed-forward layers(GPT-2 style)
- Use tanh gates to preserve original language model behaviour at initialization
- Each tanh gate controlled via a layer-specific learnable scalar (initialized to zero)

Qualitative Results-Text completion of image and text prompts

Input Prompt	Completion
 <p>This is a chinchilla. They are mainly found in Chile.</p> 	 <p>a flamingo. They are found in the Caribbean and South America.</p>
 <p>What is the title of this painting? Answer: The Hallucinogenic Toreador.</p> 	 <p>Where is this painting displayed? Answer: Louvres Museum, Paris.</p>  <p>What is the name of the city where this was painted? Answer:</p>
 <p>Output: "Underground"</p> 	 <p>Output: "Congress"</p>  <p>Output:</p>
 <p>$2+1=3$</p>  <p>$5+6=11$</p>  <p>$3 \times 6 = 18$</p>	 <p>$3 \times 6 = 18$</p>

Qualitative Results-dialogue



What is in this picture?

P It's a bowl of soup with a monster face on it.

P What is the monster made out of?

P It's made out of vegetables.

P No, it's made out of a kind of fabric. Can you see what kind?

P It's made out of a woolen fabric.



This is a picture of a dog. It looks like it is a little bit naughty.

P What happened here?

P It looks like the dog is destroying the sofa.

P How does the person thinking about it?

P The person is angry.

P What is the person holding?

P It looks like a cup of coffee.

P How is the weather outside?

P It looks like a sunny day.



What is the name of the company on this card?

P It's Homebase.

P What is its web url?

P It's homebase.co.uk.

P What are they doing?

P They are a home improvement retailer.



What do you think the robot is doing here?

P The robot is picking up a ball.

P What color is the ball?

P It's red.

P I want the ball to be hidden. Is there something the robot can use for that?

P The robot can use a cloth to hide the ball.

P What color is the cloth?

P The cloth is blue.



This is a picture of two teddy bears on the moon.

P What are they doing?

P They are having a conversation.

P What object are they using?

P It looks like a computer.

P Is this surprising?

P Yes, it is surprising.

P Why is this picture surprising to you?

P I think it is surprising because teddy bears are not usually found on the moon.



What is the common thing about these three images?

P They are all flamingos.

P What is the difference between these three images?

P The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.



This is an apple with a sticker on it.

P What does the sticker say?

P The sticker says "iPod".

P Where is the photo taken?

P It looks like it's taken in a backyard.

P Do you think it is printed or handwritten?

P It looks like it's handwritten.

P What color is the sticker?

P It's white.

Qualitative Results-video

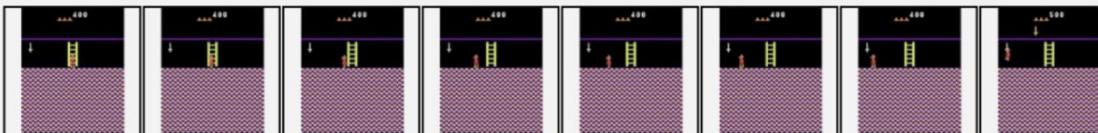
Input Prompt



Question: What is happening here?
Answer:

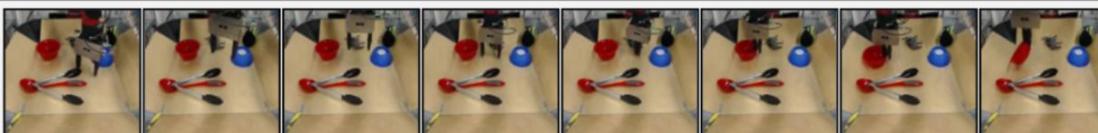
Completion

The dachschund puppy is being weighed on a scale.



Question: What object is the avatar picking up? Answer:

A sword.



Question: What is the object being lifted?
Answer:

The object is a small plastic bowl.

Flamingo failures: hallucinations/ungrounded guesses

Input Prompt



Question: What is on the phone screen? Answer:



Question: What can you see out the window? Answer:



Question: Whom is the person texting? Answer:

Output

A text message from a friend.

A parking lot.

The driver.



Visual Instruction Tuning

LLaVA: Large Language and Vision Assistant

Haotian Liu^{1*}, Chunyuan Li^{2*}, Qingyang Wu³, Yong Jae Lee

Yuqing Wang

Motivation: How to build models at multimodal GPT-4 level

- Strong pretrained vision and language models
- Connecting vision and language
- Tuning the model for instruction-following capability

Background: Instruct Tuning in LLMs

Training Seen tasks	Instruction	Input	Output
Inference Unseen tasks	Explain human's behavior.	Behavior: cry.	There could be many reasons why a person might cry. They could be feeling sad, scared, angry, or frustrated...
	Recommend a movie for me to watch during the weekend and explain the reason.	N/A	I would recommend the movie "The Shawshank Redemption" because it is an excellent movie that is both moving and inspiring...
Inference Unseen tasks	How to force kill a docker container?	N/A	To forcefully kill a Docker container, you can use the command “docker kill” followed by the container ID or name. For example, if the container ID is “container1”, ...

How to collect the instruction tuning data

- Human-Human: high-quality, hand written by humans—high cost
- Human-Machine: strong LLM-based teacher like ChatGPT—affordable cost

Seed instruction-input-output triplets

Explain human's behavior.

Behavior: cry.

<reference answer>

Recommend a movie for me .

N/A

<reference answer>

Please generate new instruction-input-output triplets that meet the following requirements: ...

<new instruction 1>

<new input 1>

<new output 1>

...

GPT-assisted Visual Instruction Data Creation

Image	Instruction	Output
	What are the challenges these people might be facing?	They may be having difficulty fitting all luggage into the back of the SUV. There are many bags, suitcases already in the back, while more...
Context (caption) A group of people standing outside of a black vehicle with various luggage.	→	
Context (layout) 	person: [0.68, 0.24, 0.77, 0.69], person: [0.63, 0.22, 0.68, 0.51], person: [0.44, 0.23, 0.48, 0.34], backpack: [0.38, 0.69, 0.48, 0.91],	

GPT-assisted Visual Instruction Data Creation

	Context	Instruction	Output
Context (caption)	A group of people standing outside of a black vehicle with various luggage.	What are the challenges these people might be facing?	They may be having difficulty fitting all luggage into the back of the SUV. There are many bags, suitcases already in the back, while more...
Context (layout)	person: [0.68, 0.24, 0.77, 0.69, person: [0.44, 0.23, 0.48, 0.34], backpack: [0.38, 0.69, 0.48, 0.91] ...		

In-context example(s) of visual-context/instruction/output triplets.

<new context (caption) 1>

<new context (layout) 1>

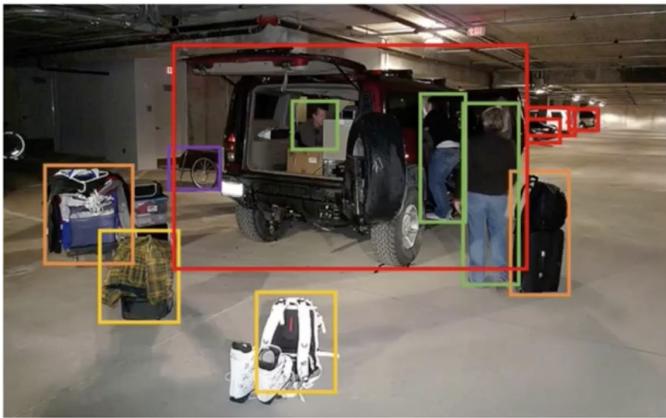
Please generate new Context-Instruction-Output triplets that meet the following requirements: ...



<new instruction 1>

<new output 1>

GPT-assisted Visual Instruction Data Creation



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.

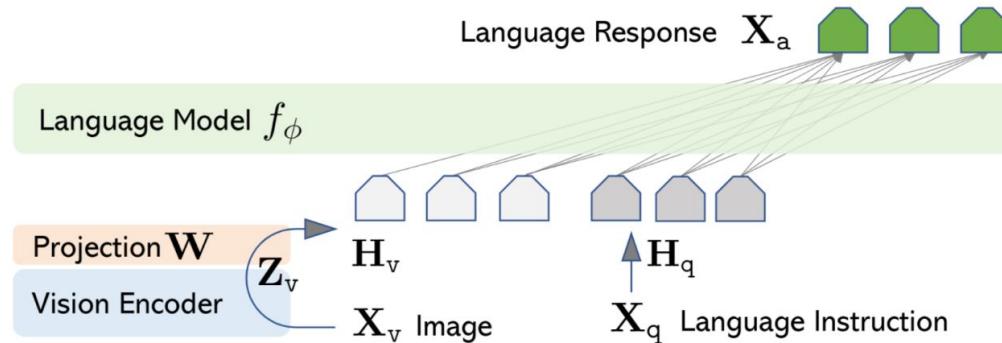
Response type 3: complex reasoning

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.

LLaVA: Large Language-and-Vision Assistant

□ Architecture



□ Two-stage Training

- **Stage 1: Pre-training for Feature Alignment.**

Only the projection matrix is updated, based on a subset of CC3M.

- **Stage 2: Fine-tuning End-to-End.** Both the projection matrix and LLM are updated

- **Visual Chat:** Our generated multimodal instruction data for daily user-oriented applications.

- **Science QA:** Multimodal reasoning dataset for the science domain.

Strong Visual Reasoning Ability



What may be the purpose of this painting?

The purpose of this painting could be to create an artistic representation of the famous painting, "The Mona Lisa," by Leonardo da Vinci. The artist has painted a dog in the same pose as the Mona Lisa, wearing a mask, and standing in front of a backdrop that resembles the original painting. This artistic interpretation may serve as a creative and humorous take on the iconic painting, drawing attention to the dog's unique expression and the contrast between the dog's appearance and the original subject. The painting could also be a tribute to the original work or a commentary on the cultural significance of the Mona Lisa.



What is the name of the man in this meme?

The man in the meme is named Elon Musk.

Questions?

- Could we develop a unified tokenizer that more effectively processes both images and text together in multimodal language models, instead of handling them separately ?
- In addition to understanding the visual appearance for classification, it may also need to tackle some region-based tasks, like detection. Can we use LLM to do detection?
- In addition to image and text, multimodality also include video, audio, etc. Based on the success of image-text learning, can we apply it to the understanding of other modalities?