UCF Center for Research in Computer Vision

Visual-Language Models Introduction Part-II:FLAMINGO, FLAVA, PAINTER, BLIP-2 Lecture-5 CAP6412 Spring 2024

Mubarak Shah

shah@crcv.ucf.edu

Important Concepts

- Supervised, Self-Supervised, Zero-Shot, Few-Shot
- Pretrained Vs Finetuned (snowflakes vs Flame)
- Image-Encoder
  - VIT CLIP,…
- Text-Encoder
  - T5, BERT, …
- LLM
  - GPT, Vicuna, OPT, LaMA, ..
- Datasets
  - Image-Text pairs

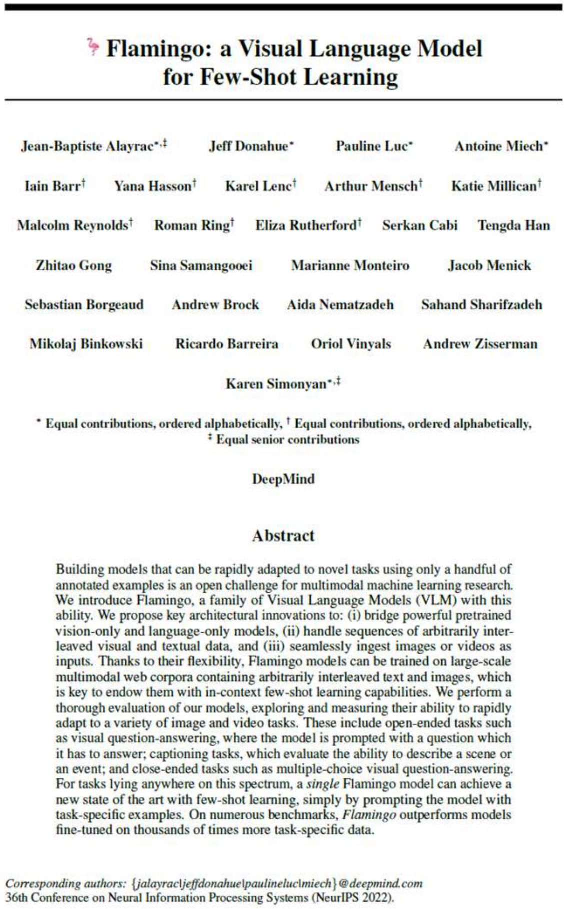
- Loss or Objective Function
  - Contrastive
  - Captioning
  - Image-Text Matching
- Tasks
  - Recognition: Accuracy
  - Retrieval (T2I, I2T): Recall
  - VQA: Accuracy
  - Captioning: BLUE,..
- Open-source vs Proprietary
  - Datasets
  - Models

Contents

- CoCa
- PALI
- FLAMINGO
- FLAVA
- Painter

- BLIP-2
- Image-Bind
- Language-Bind

- LLaVA
- Video ChatGPT



Flamingo: a Visual Language Model for Few-Shot Learning

Jean-Baptiste Alayrac*,‡ Jeff Donahue* Pauline Luc* Antoine Miech*

Iain Barr† Yana Hasson† Karel Lenc† Arthur Mensch† Katie Millican†

Malcolm Reynolds† Roman Ring† Eliza Rutherford† Serkan Cabi Tengda Han

Zhitao Gong Sina Samangooei Marianne Monteiro Jacob Menick

Sebastian Borgeaud Andrew Brock Aida Nematzadeh Sahand Sharifzadeh

Mikolaj Binkowski Ricardo Barreira Oriol Vinyals Andrew Zisserman

Karen Simonyan*,‡

* Equal contributions, ordered alphabetically, † Equal contributions, ordered alphabetically,
‡ Equal senior contributions

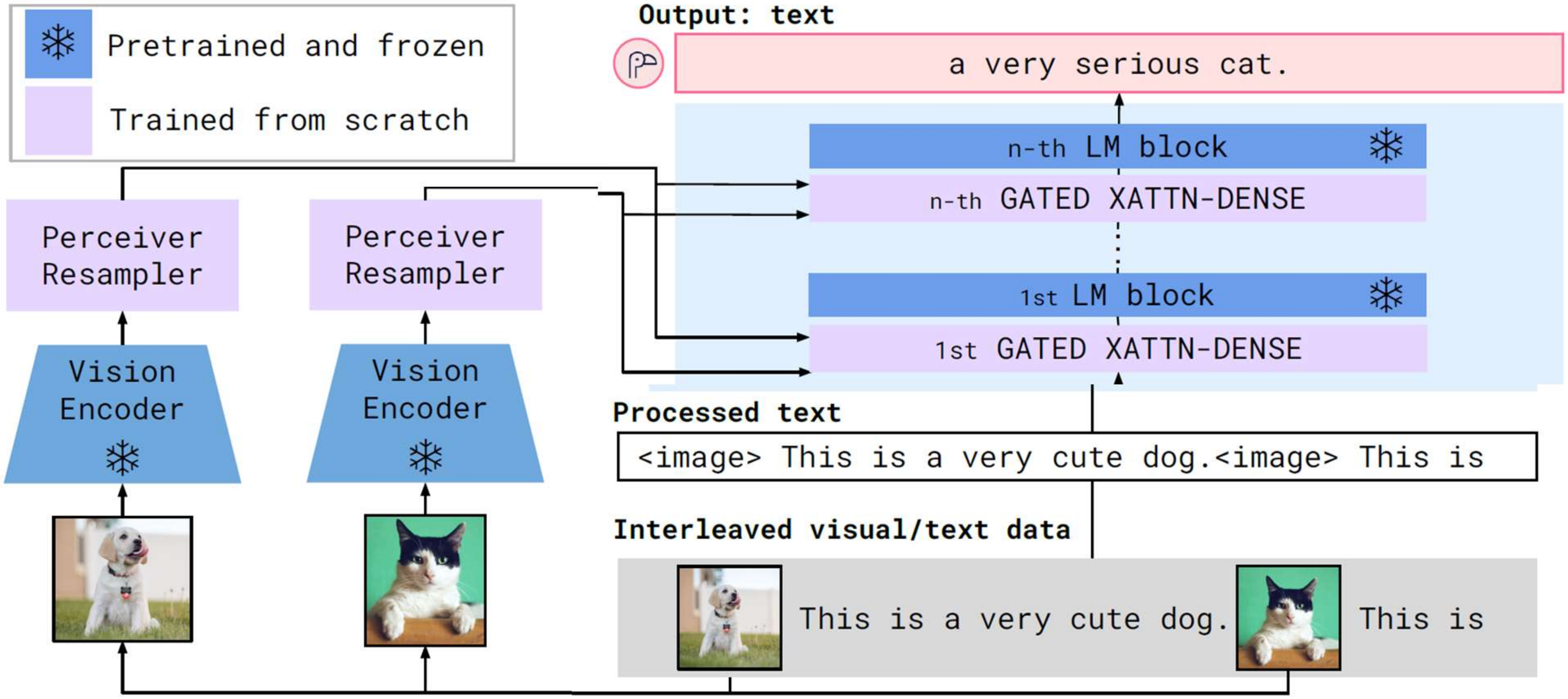
DeepMind

Abstract

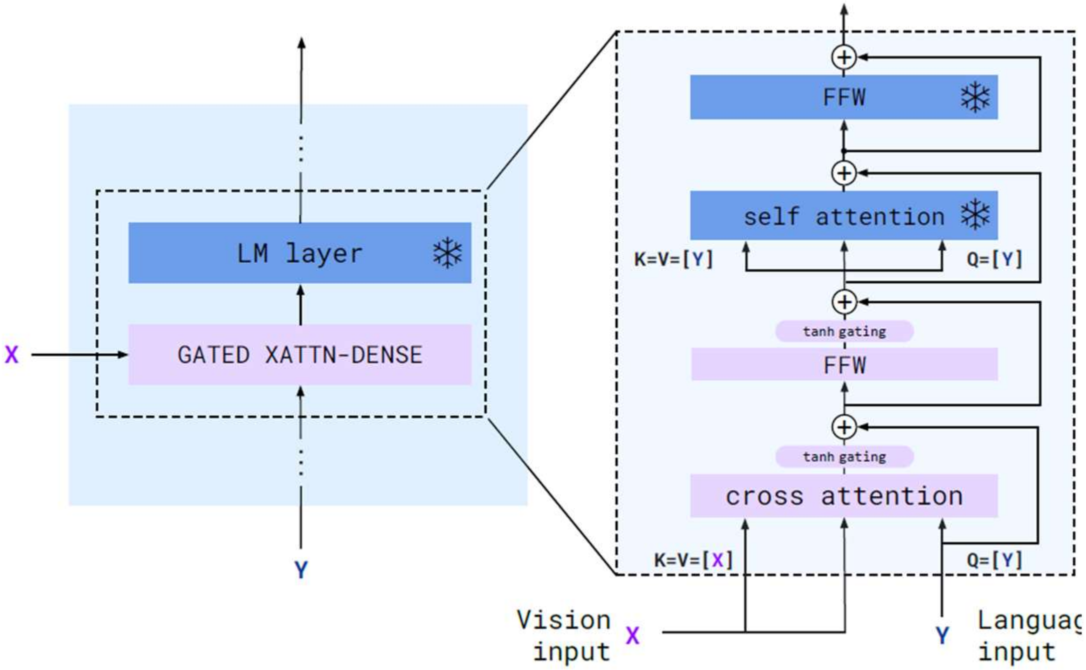
Building models that can be rapidly adapted to novel tasks using only a handful of annotated examples is an open challenge for multimodal machine learning research. We introduce Flamingo, a family of Visual Language Models (VLM) with this ability. We propose key architectural innovations to: (i) bridge powerful pretrained vision-only and language-only models, (ii) handle sequences of arbitrarily interleaved visual and textual data, and (iii) seamlessly ingest images or videos as inputs. Thanks to their flexibility, Flamingo models can be trained on large-scale multimodal web corpora containing arbitrarily interleaved text and images, which is key to endow them with in-context few-shot learning capabilities. We perform a thorough evaluation of our models, exploring and measuring their ability to rapidly adapt to a variety of image and video tasks. These include open-ended tasks such as visual question-answering, where the model is prompted with a question which it has to answer; captioning tasks, which evaluate the ability to describe a scene or an event; and close-ended tasks such as multiple-choice visual question-answering. For tasks lying anywhere on this spectrum, a *single* Flamingo model can achieve a new state of the art with few-shot learning, simply by prompting the model with task-specific examples. On numerous benchmarks, *Flamingo* outperforms models fine-tuned on thousands of times more task-specific data.

Corresponding authors: {jalayrac|jeffdonahue|paulineluc|miech}@deepmind.com
36th Conference on Neural Information Processing Systems (NeurIPS 2022).

FLAMINGO



Perceiver Presentation Video: https://www.youtube.com/watch?v=XyZ9ZaxAhro


LM layer
GATED XATTN-DENSE
X
Y
FFW
self attention
K=V=[Y]
Q=[Y]
tanh gating
FFW
tanh gating
cross attention
K=V=[X]
Q=[Y]
Vision input
X
Y
Languag input

FLAMINGO

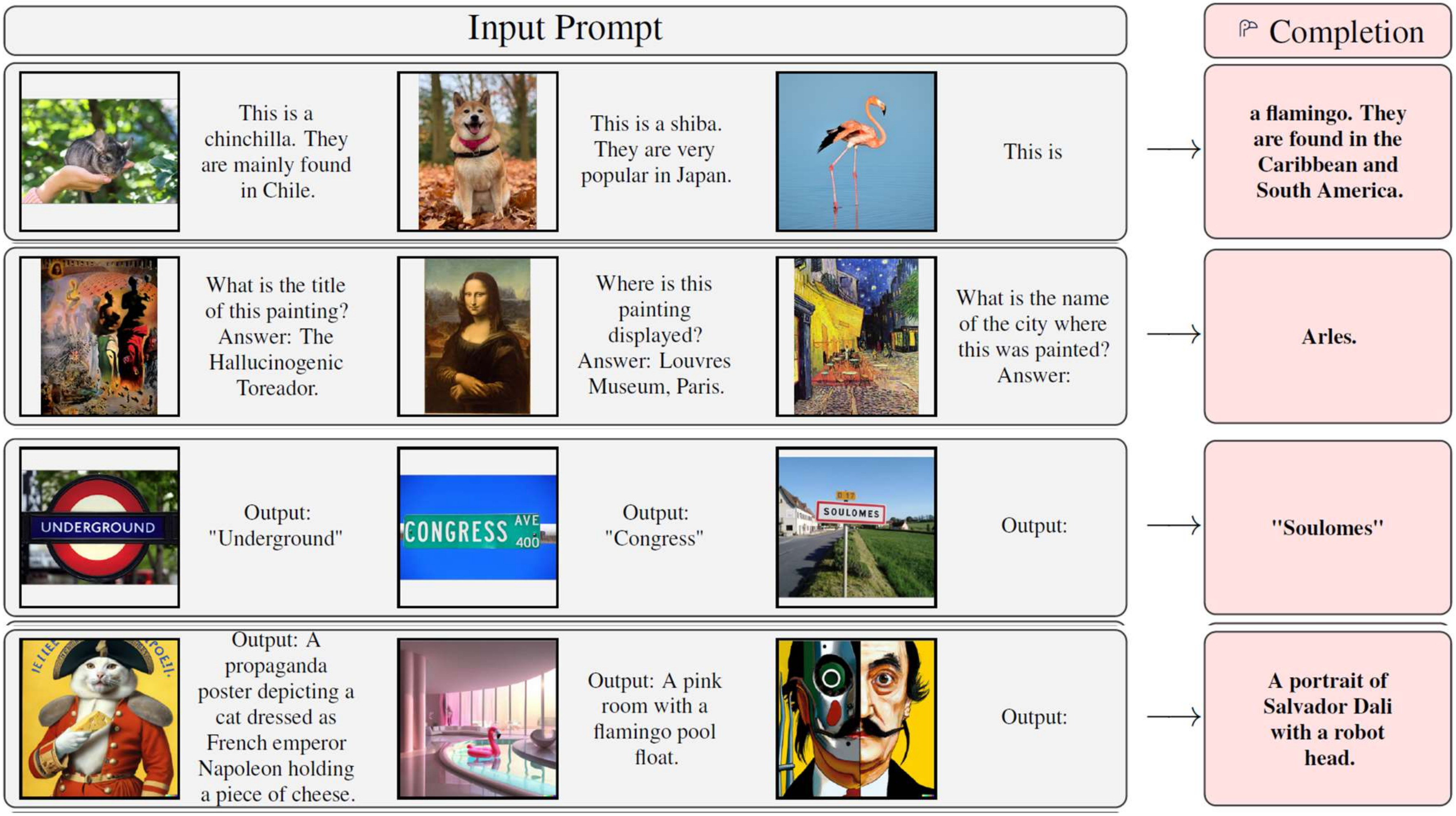
- 70B Chinchilla language model
- ResNet Vision encoder

- Objective

$$p(y|x) = \prod_{\ell=1}^{L} p(y_\ell | y_{<\ell}, x_{\leq\ell}),$$

Dataset

- Training on a mixture of vision and language datasets
  - MultiModal MassiveWeb (M3W) (Interleaved..)
    - 43 million webpages
- Pairs of image/video and text
  - LTIP (Long Text & Image Pairs) which consists of 312 million image and text pairs.
  - VTP (Video & Text Pairs) consists of 27 million short videos (approximately 22 seconds on average) paired with sentence descriptions



Input Prompt

Completion

This is a chinchilla. They are mainly found in Chile.

This is a shiba. They are very popular in Japan.

This is → a flamingo. They are found in the Caribbean and South America.

What is the title of this painting? Answer: The Hallucinogenic Toreador.

Where is this painting displayed? Answer: Louvres Museum, Paris.

What is the name of the city where this was painted? Answer: → Arles.

Output: "Underground"

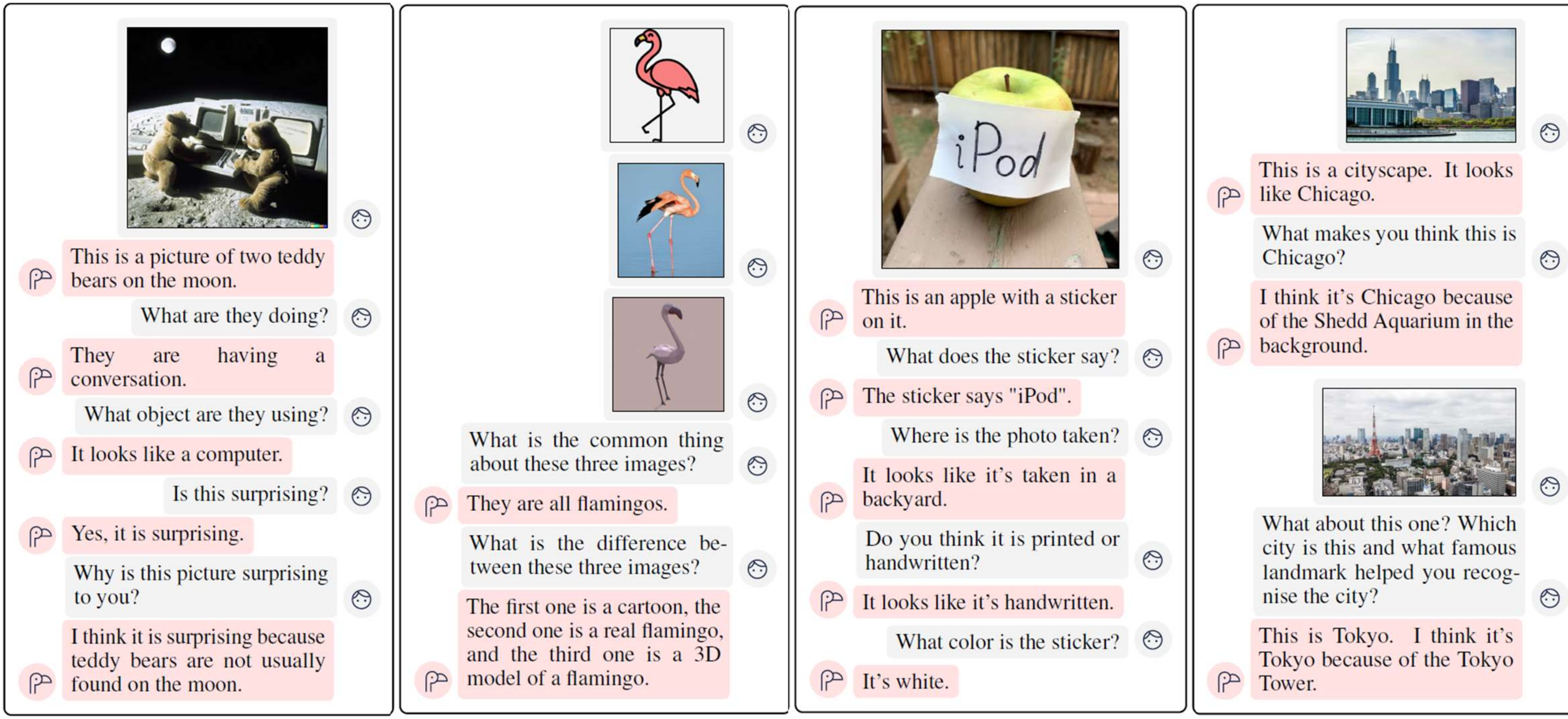
Output: "Congress"

Output: → "Soulomes"

Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese.

Output: A pink room with a flamingo pool float.

Output: → A portrait of Salvador Dali with a robot head.



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

What are they doing?

They are having a conversation.

What object are they using?

It looks like a computer.

Is this surprising?

Yes, it is surprising.

Why is this picture surprising to you?

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

What is the common thing about these three images?

They are all flamingos.

What is the difference between these three images?

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.

What does the sticker say?

The sticker says "iPod".

Where is the photo taken?

It looks like it's taken in a backyard.

Do you think it is printed or handwritten?

It looks like it's handwritten.

What color is the sticker?

It's white.

This is a cityscape. It looks like Chicago.

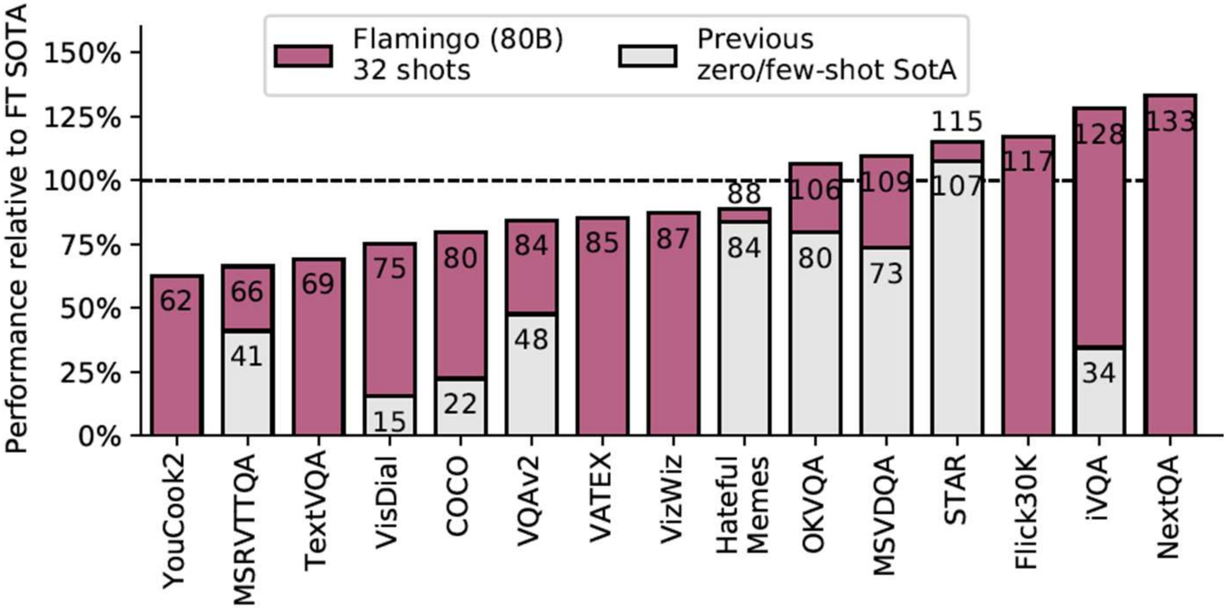
What makes you think this is Chicago?

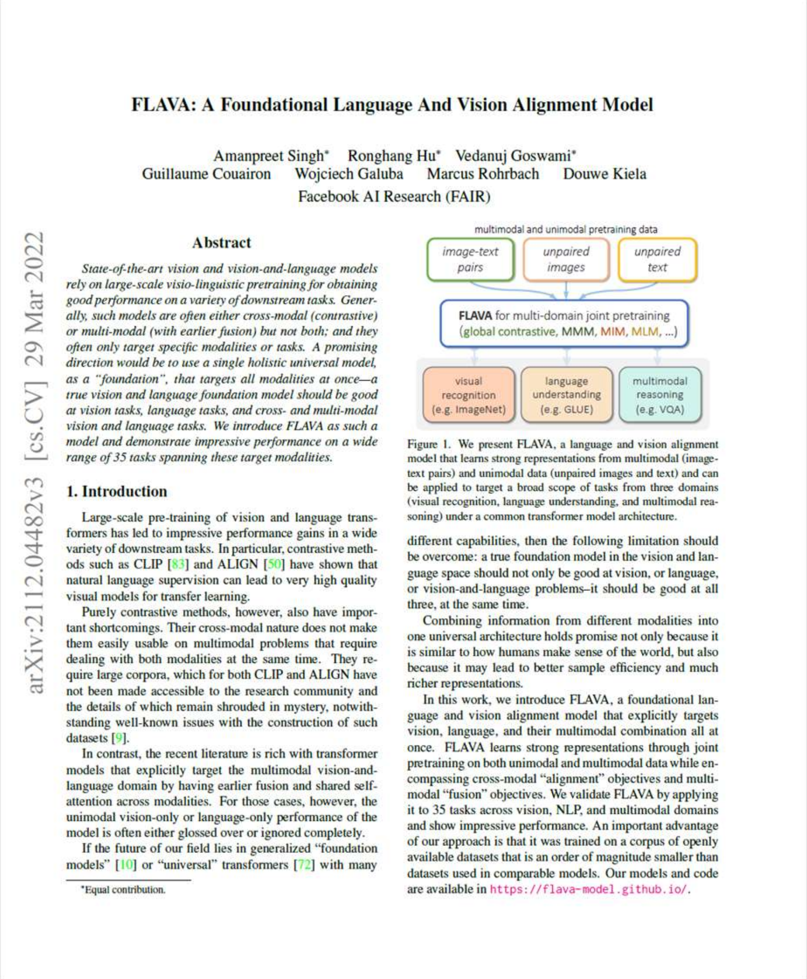
I think it's Chicago because of the Shedd Aquarium in the background.

What about this one? Which city is this and what famous landmark helped you recognise the city?

This is Tokyo. I think it's Tokyo because of the Tokyo Tower.

Results


Flamingo (80B)
32 shots
Previous
zero/few-shot SotA
Performance relative to FT SOTA
150%
125%
100%
75%
50%
25%
0%
62
66
41
69
75
15
80
22
84
48
85
87
88
84
106
80
109
73
115
107
117
128
34
133
YouCook2
MSRVTTQA
TextVQA
VisDial
COCO
VQAv2
VATEX
VizWiz
Hateful
Memes
OKVQA
MSVDQA
STAR
Flick30K
iVQA
NextQA



FLAVA: A Foundational Language And Vision Alignment Model

Amanpreet Singh* Ronghang Hu* Vedanuj Goswami*
Guillaume Couairon Wojciech Galuba Marcus Rohrbach Douwe Kiela
Facebook AI Research (FAIR)

arXiv:2112.04482v3 [cs.CV] 29 Mar 2022

Abstract

State-of-the-art vision and vision-and-language models rely on large-scale visio-linguistic pretraining for obtaining good performance on a variety of downstream tasks. Generally, such models are often either cross-modal (contrastive) or multi-modal (with earlier fusion) but not both; and they often only target specific modalities or tasks. A promising direction would be to use a single holistic universal model, as a "foundation", that targets all modalities at once—a true vision and language foundation model should be good at vision tasks, language tasks, and cross- and multi-modal vision and language tasks. We introduce FLAVA as such a model and demonstrate impressive performance on a wide range of 35 tasks spanning these target modalities.

1. Introduction

Large-scale pre-training of vision and language transformers has led to impressive performance gains in a wide variety of downstream tasks. In particular, contrastive methods such as CLIP [83] and ALIGN [50] have shown that natural language supervision can lead to very high quality visual models for transfer learning.

Purely contrastive methods, however, also have important shortcomings. Their cross-modal nature does not make them easily usable on multimodal problems that require dealing with both modalities at the same time. They require large corpora, which for both CLIP and ALIGN have not been made accessible to the research community and the details of which remain shrouded in mystery, notwithstanding well-known issues with the construction of such datasets [9].

In contrast, the recent literature is rich with transformer models that explicitly target the multimodal vision-and-language domain by having earlier fusion and shared self-attention across modalities. For those cases, however, the unimodal vision-only or language-only performance of the model is often either glossed over or ignored completely.

If the future of our field lies in generalized "foundation models" [10] or "universal" transformers [72] with many different capabilities, then the following limitation should be overcome: a true foundation model in the vision and language space should not only be good at vision, or language, or vision-and-language problems–it should be good at all three, at the same time.

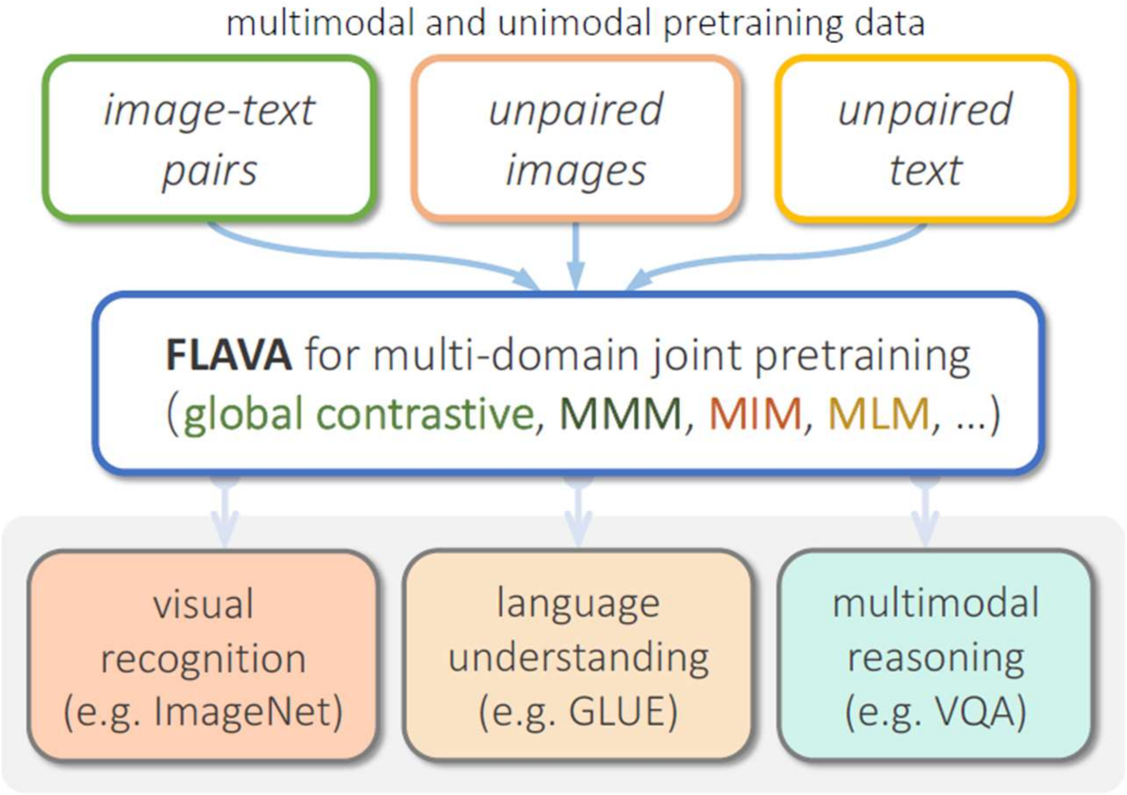
multimodal and unimodal pretraining data
image-text pairs
unpaired images
unpaired text
FLAVA for multi-domain joint pretraining
(global contrastive, MMM, MIM, MLM, ...)
visual recognition (e.g. ImageNet)
language understanding (e.g. GLUE)
multimodal reasoning (e.g. VQA)

Figure 1. We present FLAVA, a language and vision alignment model that learns strong representations from multimodal (image-text pairs) and unimodal data (unpaired images and text) and can be applied to target a broad scope of tasks from three domains (visual recognition, language understanding, and multimodal reasoning) under a common transformer model architecture.

Combining information from different modalities into one universal architecture holds promise not only because it is similar to how humans make sense of the world, but also because it may lead to better sample efficiency and much richer representations.

In this work, we introduce FLAVA, a foundational language and vision alignment model that explicitly targets vision, language, and their multimodal combination all at once. FLAVA learns strong representations through joint pretraining on both unimodal and multimodal data while encompassing cross-modal "alignment" objectives and multimodal "fusion" objectives. We validate FLAVA by applying it to 35 tasks across vision, NLP, and multimodal domains and show impressive performance. An important advantage of our approach is that it was trained on a corpus of openly available datasets that is an order of magnitude smaller than datasets used in comparable models. Our models and code are available in `https://flava-model.github.io/`.

*Equal contribution.



multimodal and unimodal pretraining data

image-text pairs

unpaired images

unpaired text

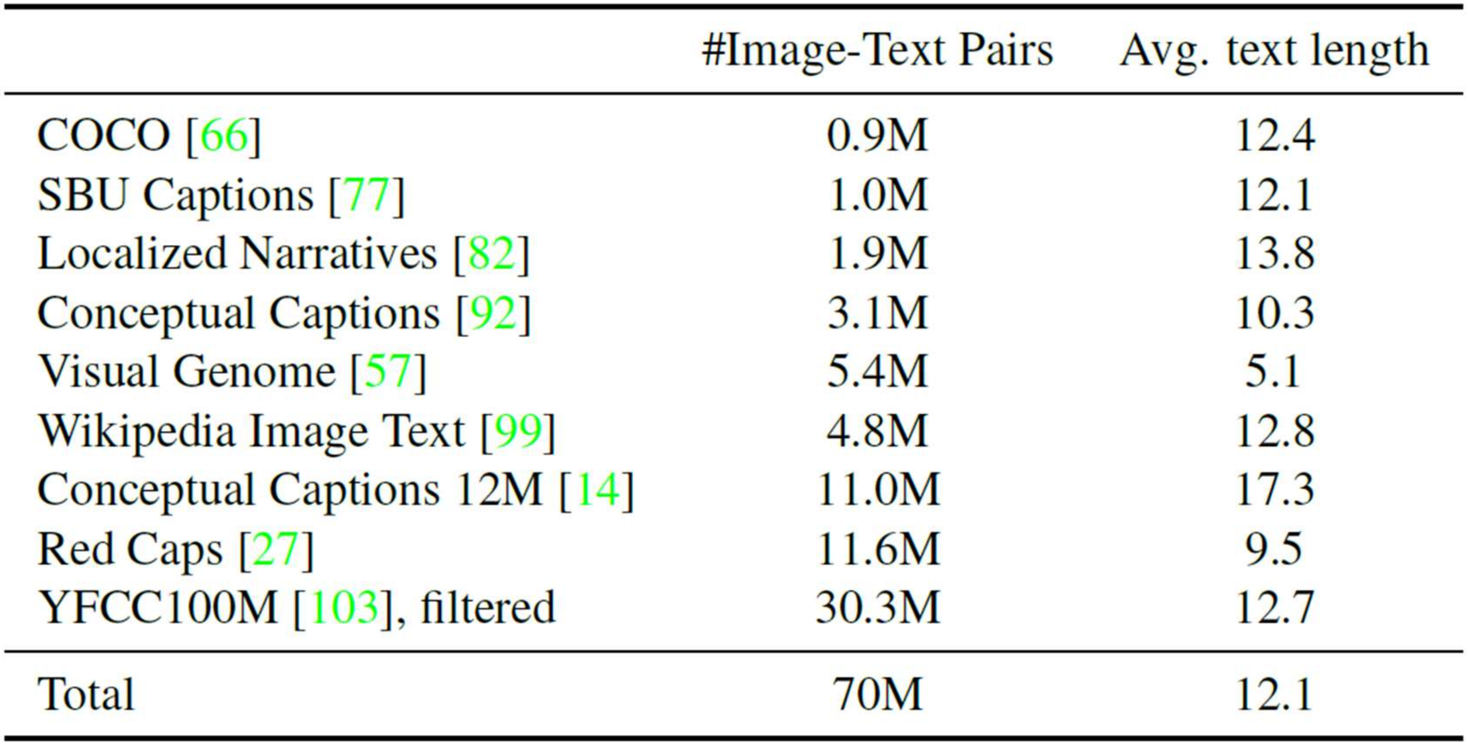
FLAVA for multi-domain joint pretraining (global contrastive, MMM, MIM, MLM, ...)

visual recognition (e.g. ImageNet)

language understanding (e.g. GLUE)

multimodal reasoning (e.g. VQA)

Trained on a corpus of openly available datasets



	#Image-Text Pairs	Avg. text length
COCO [66]	0.9M	12.4
SBU Captions [77]	1.0M	12.1
Localized Narratives [82]	1.9M	13.8
Conceptual Captions [92]	3.1M	10.3
Visual Genome [57]	5.4M	5.1
Wikipedia Image Text [99]	4.8M	12.8
Conceptual Captions 12M [14]	11.0M	17.3
Red Caps [27]	11.6M	9.5
YFCC100M [103], filtered	30.3M	12.7
Total	70M	12.1

FLAVA

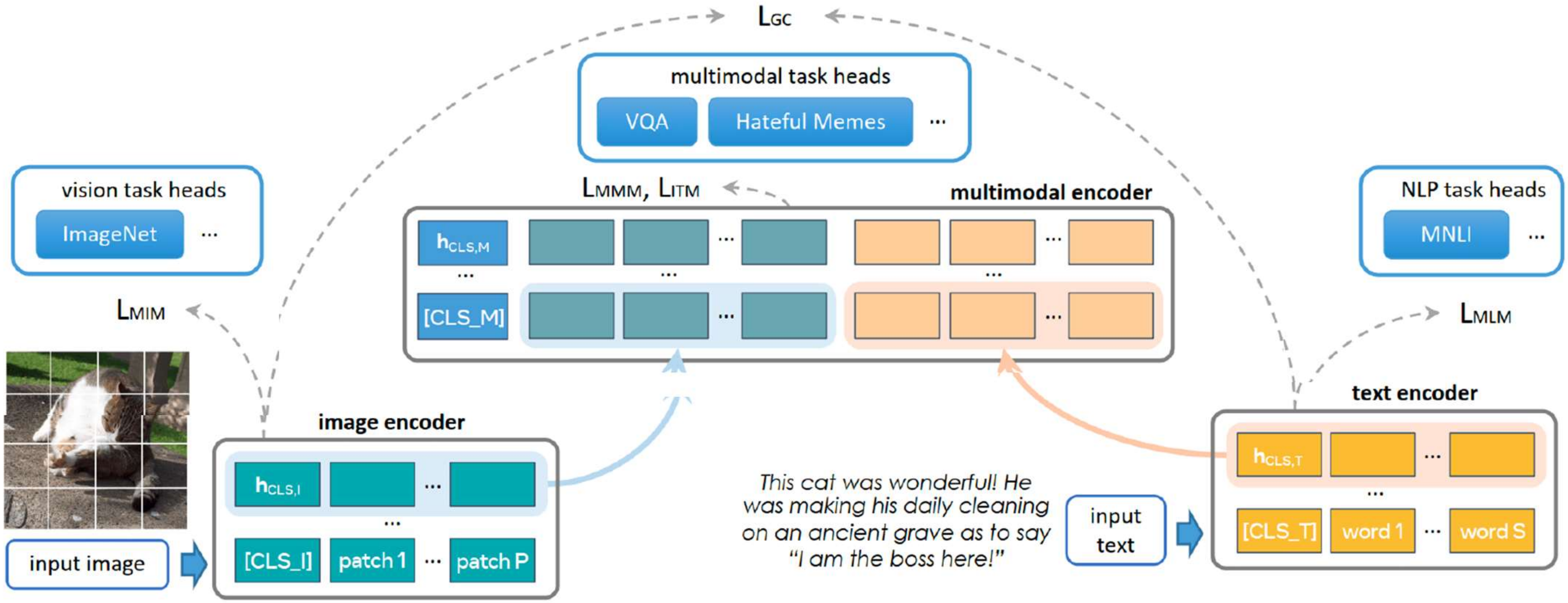

L_{GC}
multimodal task heads
VQA
Hateful Memes
...
vision task heads
ImageNet
...
L_{MMM}, L_{ITM}
multimodal encoder
NLP task heads
MNLI
...
$h_{CLS,M}$
[CLS_M]
L_{MIM}
L_{MLM}
text encoder
image encoder
$h_{CLS,I}$
[CLS_I]
patch 1
patch P
$h_{CLS,T}$
[CLS_T]
word 1
word S
input image
input text
This cat was wonderful! He was making his daily cleaning on an ancient grave as to say "I am the boss here!"

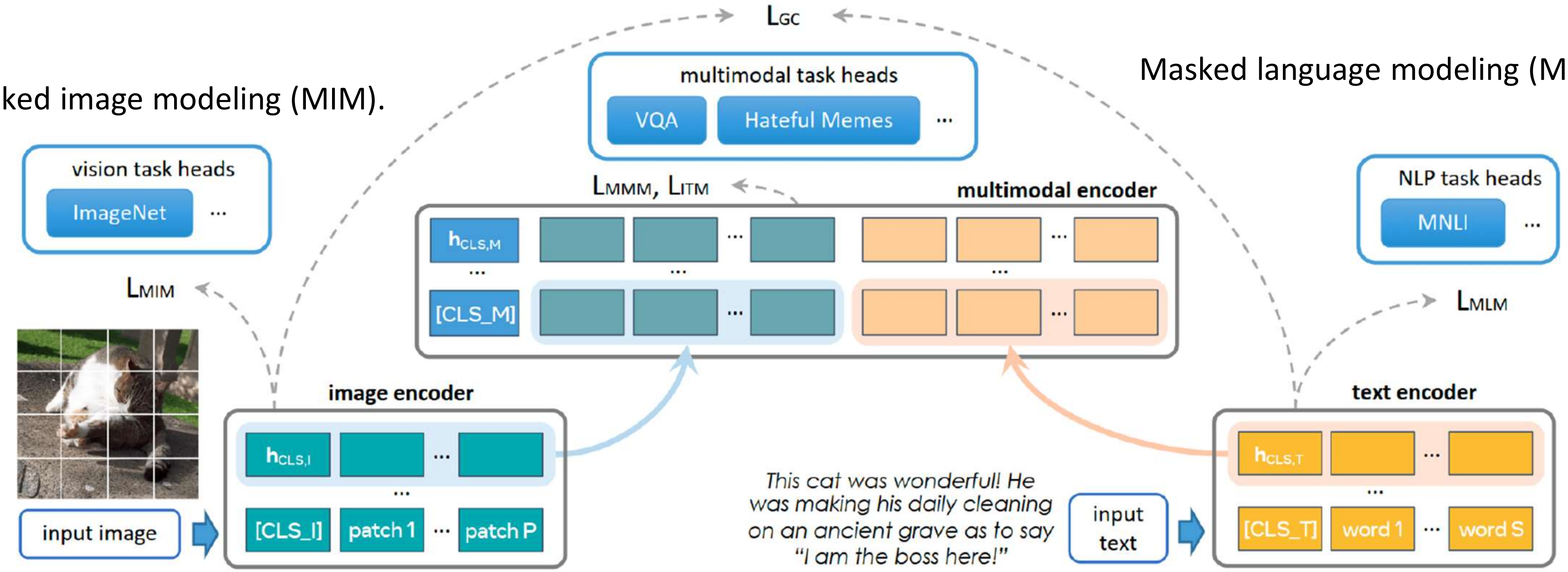
Image-text matching (ITM).

Masked multimodal modeling (MMM).

Global contrastive (GC) loss

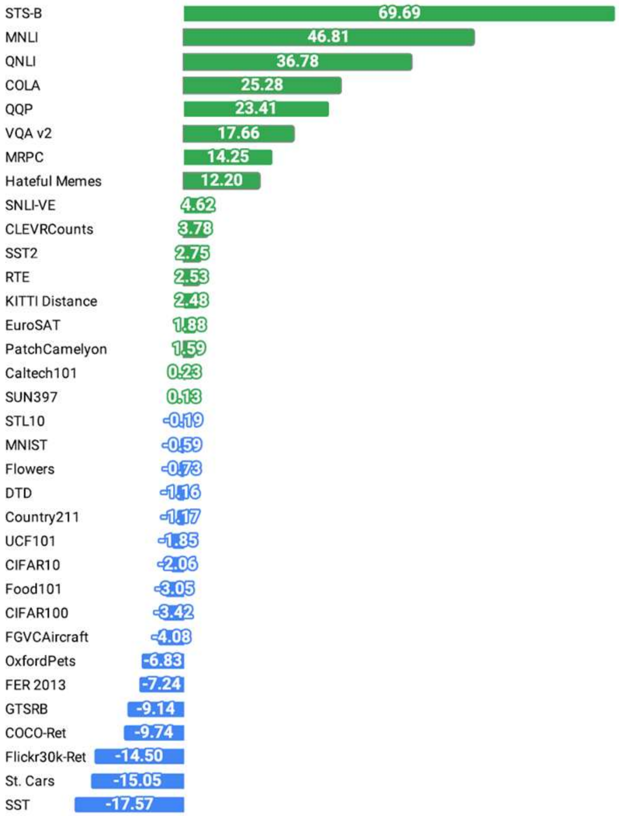
Masked image modeling (MIM).

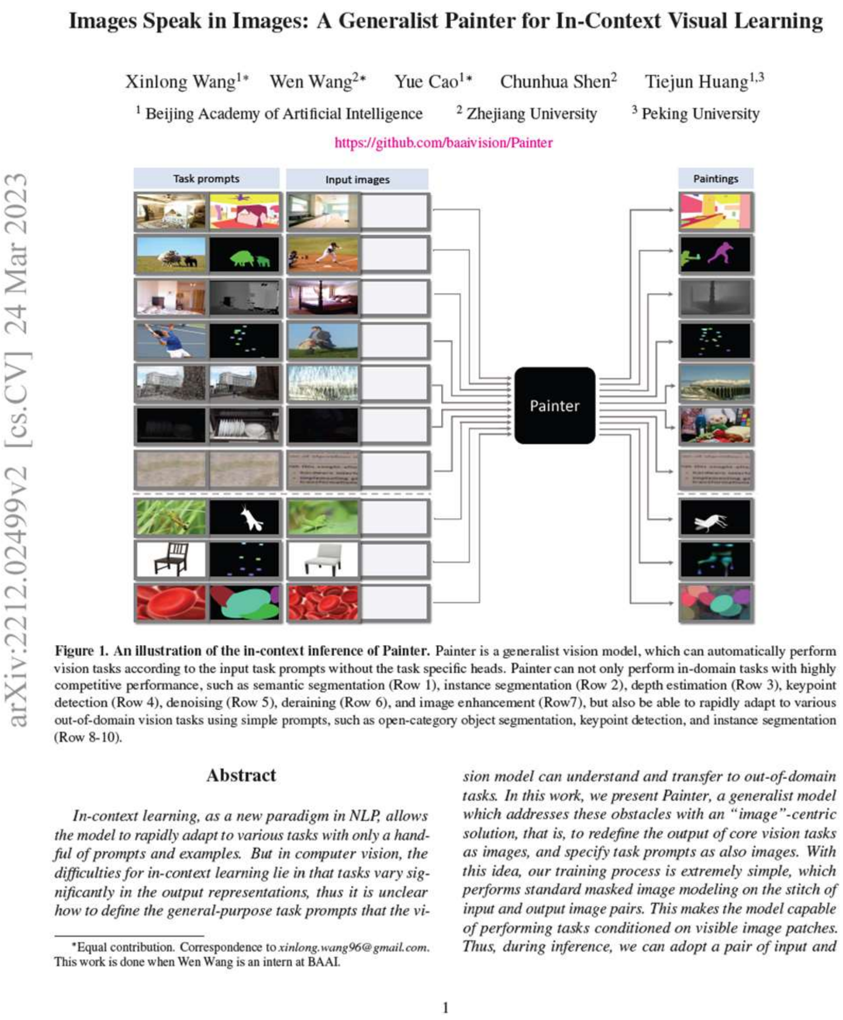
Masked language modeling (MLM).



This cat was wonderful! He was making his daily cleaning on an ancient grave as to say "I am the boss here!"

FLAVA Vs CLIP





Images Speak in Images: A Generalist Painter for In-Context Visual Learning

Xinlong Wang[1*] Wen Wang[2*] Yue Cao[1*] Chunhua Shen[2] Tiejun Huang[1,3]
[1] Beijing Academy of Artificial Intelligence [2] Zhejiang University [3] Peking University
https://github.com/baaivision/Painter

Figure 1. An illustration of the in-context inference of Painter. Painter is a generalist vision model, which can automatically perform vision tasks according to the input task prompts without the task specific heads. Painter can not only perform in-domain tasks with highly competitive performance, such as semantic segmentation (Row 1), instance segmentation (Row 2), depth estimation (Row 3), keypoint detection (Row 4), denoising (Row 5), deraining (Row 6), and image enhancement (Row7), but also be able to rapidly adapt to various out-of-domain vision tasks using simple prompts, such as open-category object segmentation, keypoint detection, and instance segmentation (Row 8-10).

Abstract

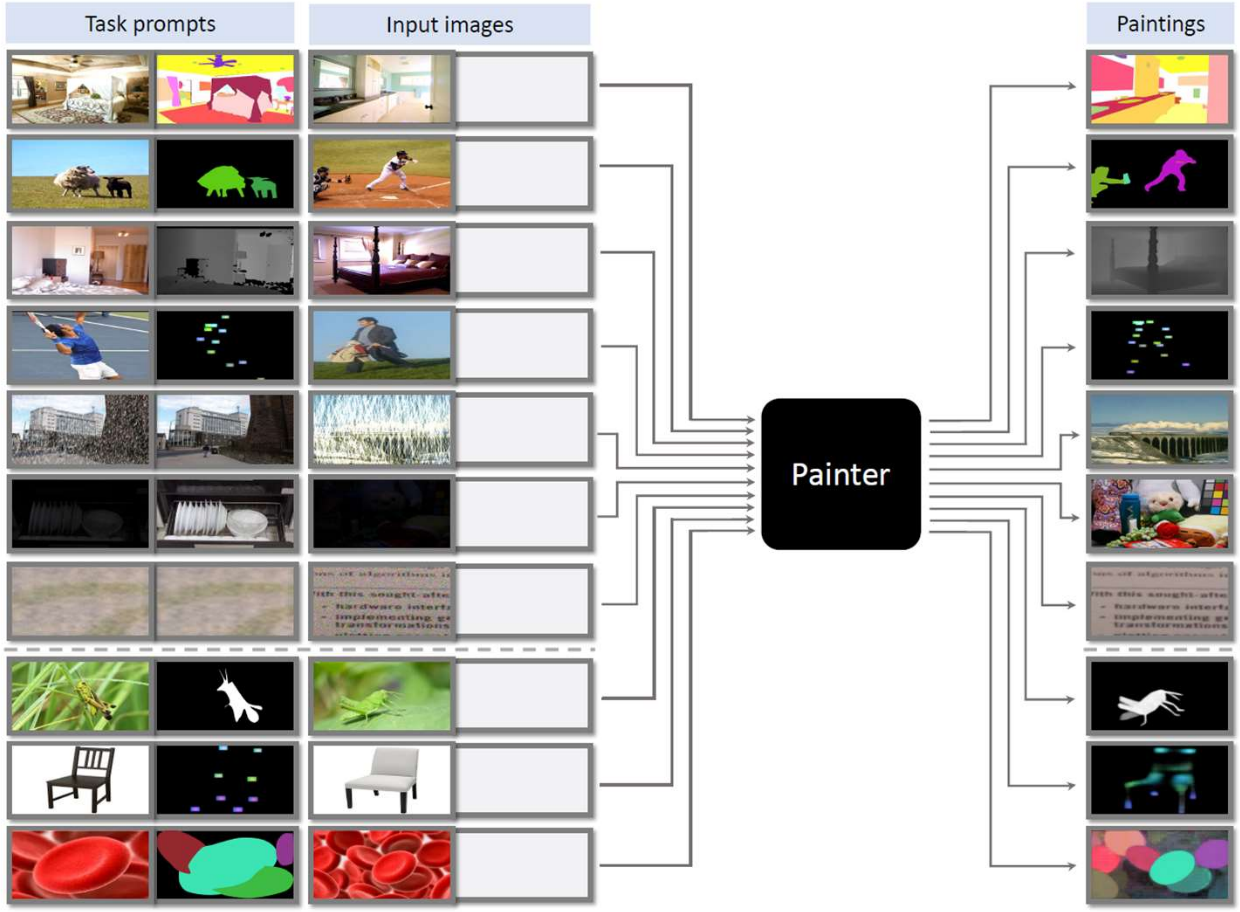
In-context learning, as a new paradigm in NLP, allows the model to rapidly adapt to various tasks with only a handful of prompts and examples. But in computer vision, the difficulties for in-context learning lie in that tasks vary significantly in the output representations, thus it is unclear how to define the general-purpose task prompts that the vision model can understand and transfer to out-of-domain tasks. In this work, we present Painter, a generalist model which addresses these obstacles with an "image"-centric solution, that is, to redefine the output of core vision tasks as images, and specify task prompts as also images. With this idea, our training process is extremely simple, which performs standard masked image modeling on the stitch of input and output image pairs. This makes the model capable of performing tasks conditioned on visible image patches. Thus, during inference, we can adopt a pair of input and

*Equal contribution. Correspondence to *xinlong.wang96@gmail.com*. This work is done when Wen Wang is an intern at BAAI.

1

PAINTER

- In In-context learning the model rapidly adapts to various tasks
- Painter is a generalist model with an “image”-centric solution
- Given an input image inpaint the desired but missing output “image”
- 24 blocks for ViT-large
- Smooth-ℓ1 loss



Semantic segmentation

Instance segmentation

Depth estimation

Keypoint detection

Denoising

Deraining

Image enhancement

open-category

Object segmentation

Keypoint detection

Instance segmentation

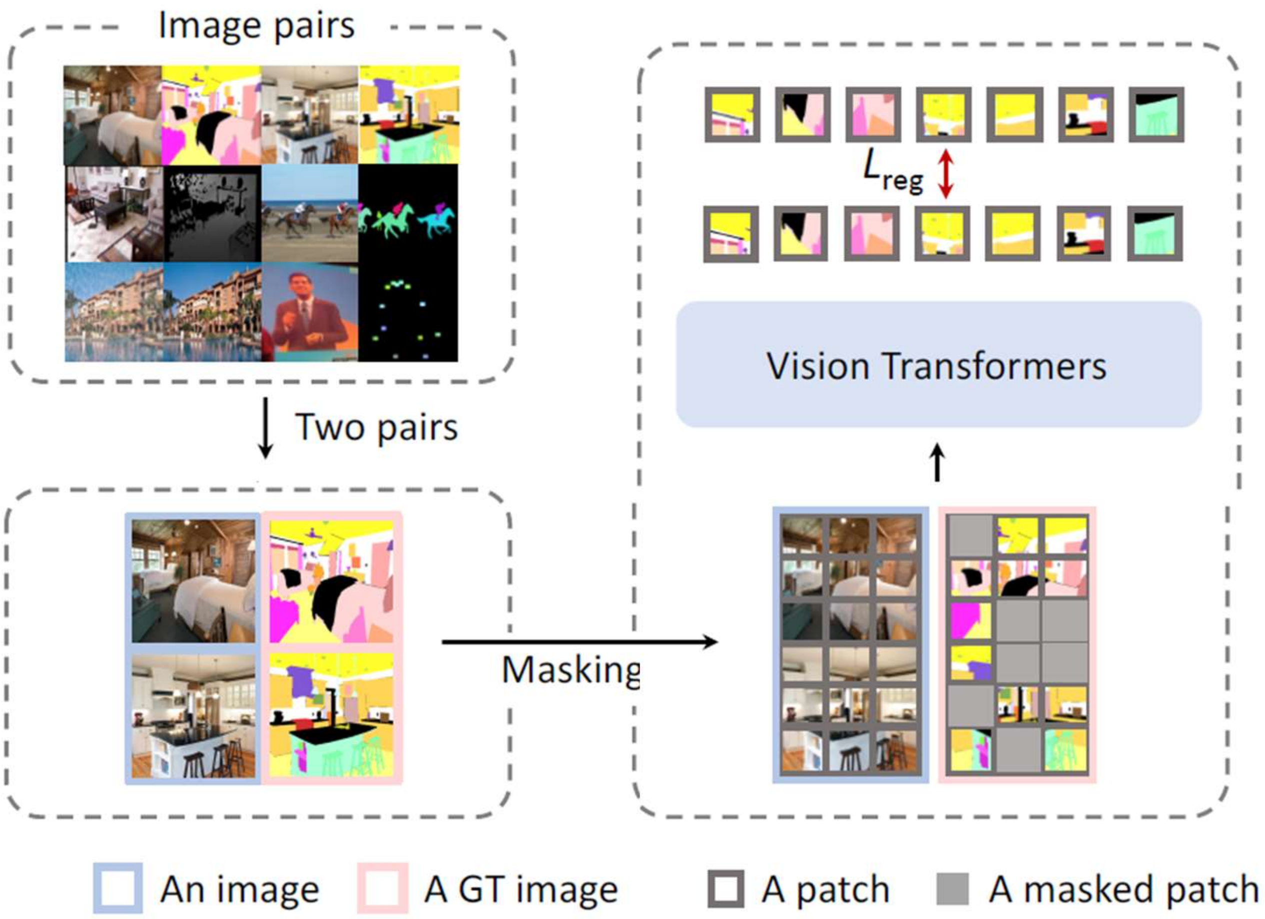
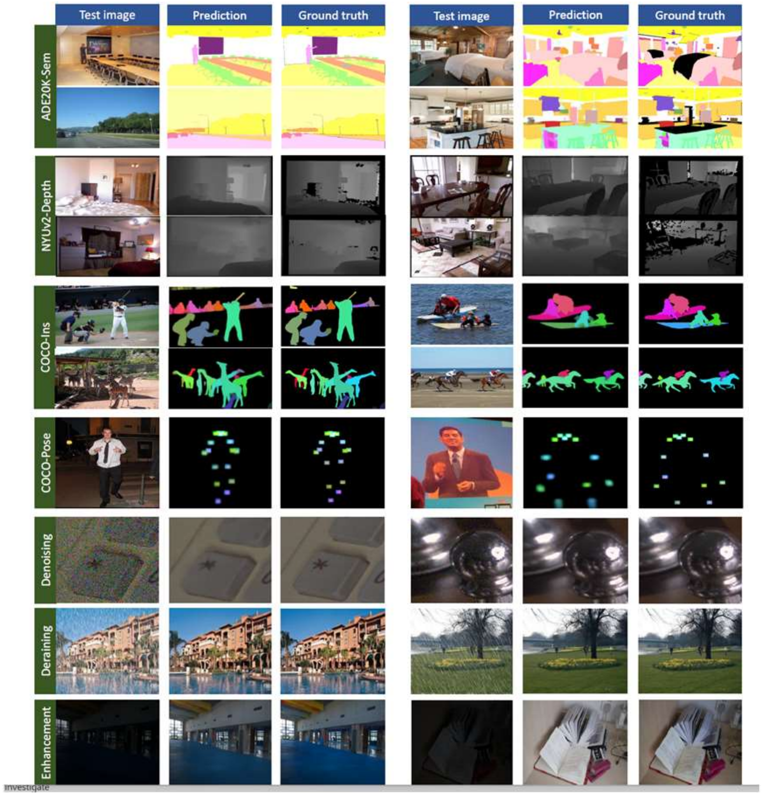
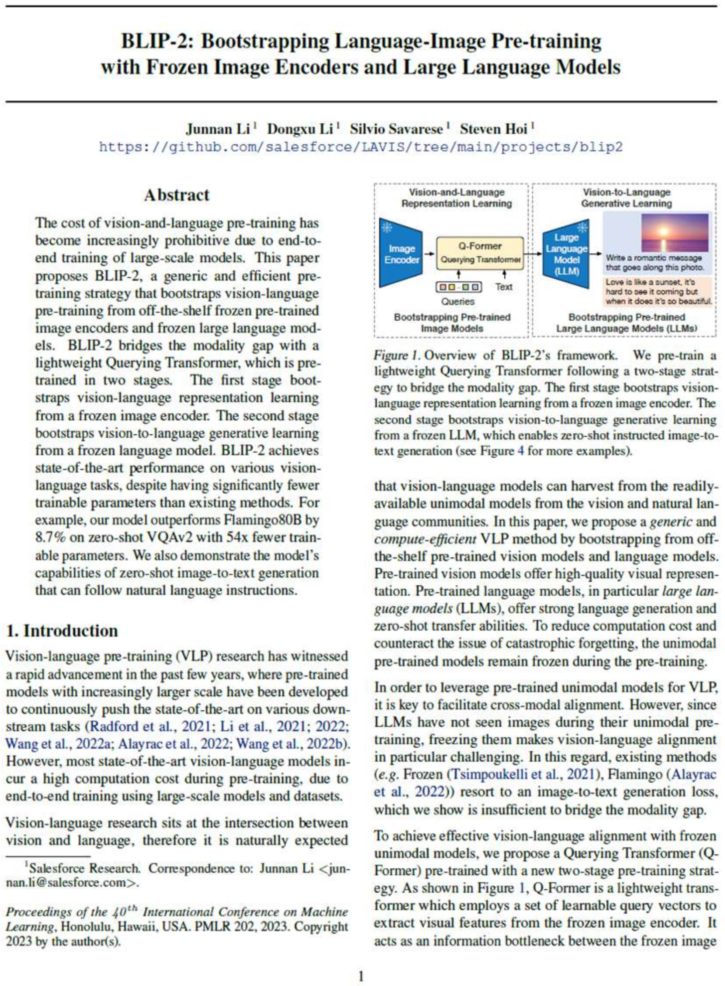


Image pairs
Two pairs
Masking
Vision Transformers
L_{reg}
An image
A GT image
A patch
A masked patch

Results


Test image
Prediction
Ground truth
Test image
Prediction
Ground truth
ADE20K-Sem
NYUv2-Depth
COCO-Ins
COCO-Pose
Denoising
Deraining
Enhancement
investigate



BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Junnan Li[1] Dongxu Li[1] Silvio Savarese[1] Steven Hoi[1]
https://github.com/salesforce/LAVIS/tree/main/projects/blip2

Abstract

The cost of vision-and-language pre-training has become increasingly prohibitive due to end-to-end training of large-scale models. This paper proposes BLIP-2, a generic and efficient pre-training strategy that bootstraps vision-language pre-training from off-the-shelf frozen pre-trained image encoders and frozen large language models. BLIP-2 bridges the modality gap with a lightweight Querying Transformer, which is pre-trained in two stages. The first stage bootstraps vision-language representation learning from a frozen image encoder. The second stage bootstraps vision-to-language generative learning from a frozen language model. BLIP-2 achieves state-of-the-art performance on various vision-language tasks, despite having significantly fewer trainable parameters than existing methods. For example, our model outperforms Flamingo80B by 8.7% on zero-shot VQAv2 with 54x fewer trainable parameters. We also demonstrate the model's capabilities of zero-shot image-to-text generation that can follow natural language instructions.

1. Introduction

Vision-language pre-training (VLP) research has witnessed a rapid advancement in the past few years, where pre-trained models with increasingly larger scale have been developed to continuously push the state-of-the-art on various downstream tasks (Radford et al., 2021; Li et al., 2021; 2022; Wang et al., 2022a; Alayrac et al., 2022; Wang et al., 2022b). However, most state-of-the-art vision-language models incur a high computation cost during pre-training, due to end-to-end training using large-scale models and datasets.

Vision-language research sits at the intersection between vision and language, therefore it is naturally expected that vision-language models can harvest from the readily-available unimodal models from the vision and natural language communities. In this paper, we propose a *generic* and *compute-efficient* VLP method by bootstrapping from off-the-shelf pre-trained vision models and language models. Pre-trained vision models offer high-quality visual representation. Pre-trained language models, in particular *large language models* (LLMs), offer strong language generation and zero-shot transfer abilities. To reduce computation cost and counteract the issue of catastrophic forgetting, the unimodal pre-trained models remain frozen during the pre-training.

Vision-and-Language
Representation Learning
Image Encoder
Q-Former
Querying Transformer
Queries
Text
Bootstrapping Pre-trained
Image Models
Vision-to-Language
Generative Learning
Large Language Model (LLM)
Write a romantic message that goes along this photo.
Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.
Bootstrapping Pre-trained
Large Language Models (LLMs)

Figure 1. Overview of BLIP-2's framework. We pre-train a lightweight Querying Transformer following a two-stage strategy to bridge the modality gap. The first stage bootstraps vision-language representation learning from a frozen image encoder. The second stage bootstraps vision-to-language generative learning from a frozen LLM, which enables zero-shot instructed image-to-text generation (see Figure 4 for more examples).

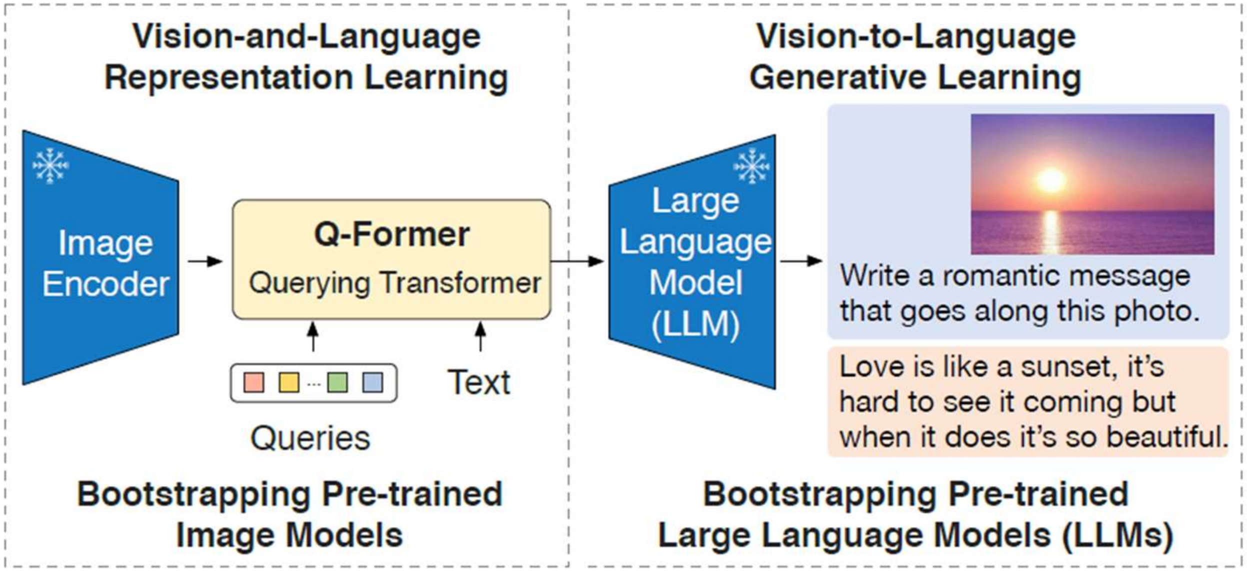
In order to leverage pre-trained unimodal models for VLP, it is key to facilitate cross-modal alignment. However, since LLMs have not seen images during their unimodal pre-training, freezing them makes vision-language alignment in particular challenging. In this regard, existing methods (*e.g.* Frozen (Tsimpoukelli et al., 2021), Flamingo (Alayrac et al., 2022)) resort to an image-to-text generation loss, which we show is insufficient to bridge the modality gap.

To achieve effective vision-language alignment with frozen unimodal models, we propose a Querying Transformer (Q-Former) pre-trained with a new two-stage pre-training strategy. As shown in Figure 1, Q-Former is a lightweight transformer which employs a set of learnable query vectors to extract visual features from the frozen image encoder. It acts as an information bottleneck between the frozen image

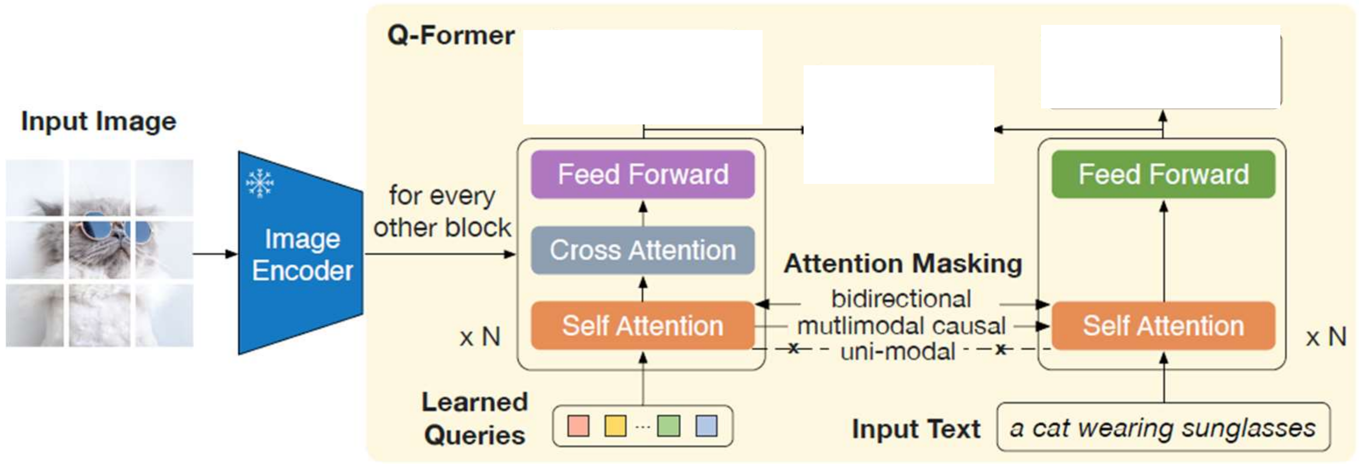
[1] Salesforce Research. Correspondence to: Junnan Li <junnan.li@salesforce.com>.

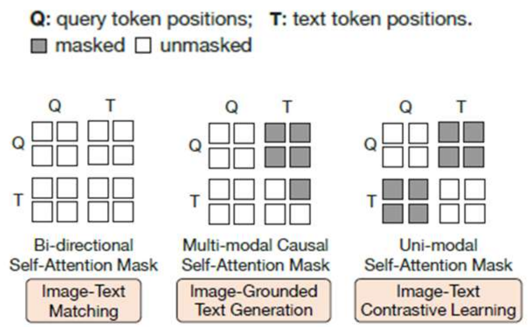
Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models


Vision-and-Language Representation Learning
Image Encoder
Q-Former
Querying Transformer
Queries
Text
Bootstrapping Pre-trained Image Models
Vision-to-Language Generative Learning
Large Language Model (LLM)
Write a romantic message that goes along this photo.
Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.
Bootstrapping Pre-trained Large Language Models (LLMs)

BLIP-2

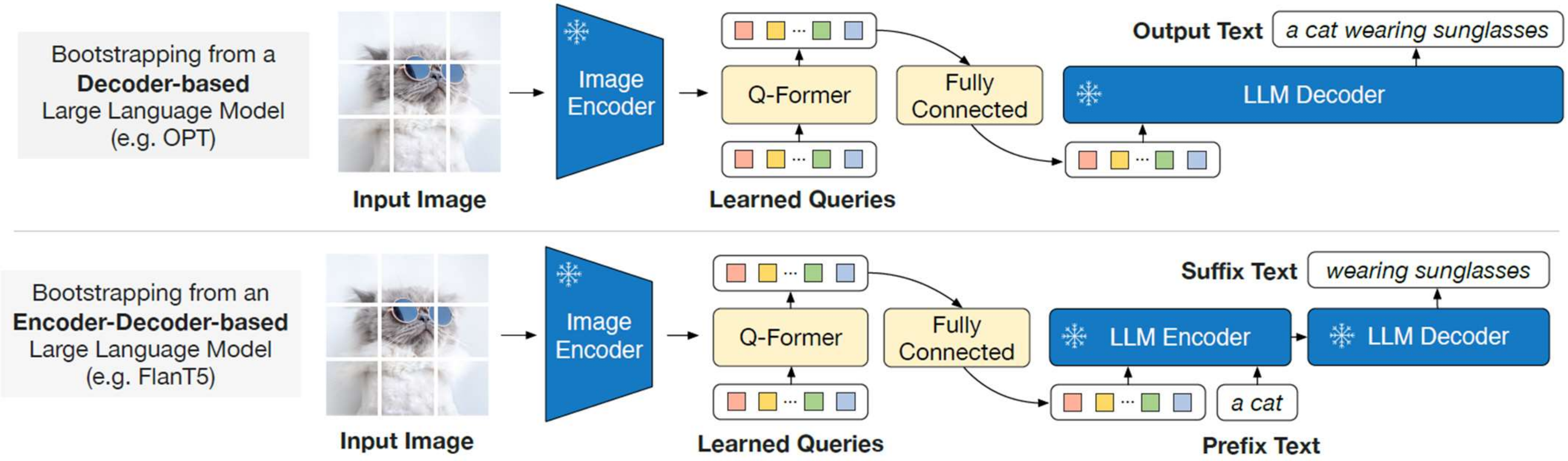

Input Image
Image Encoder
for every other block
Q-Former
Feed Forward
Cross Attention
Self Attention
x N
Attention Masking
bidirectional
mutlimodal causal
uni-modal
Feed Forward
Self Attention
x N
Learned Queries
Input Text
a cat wearing sunglasses

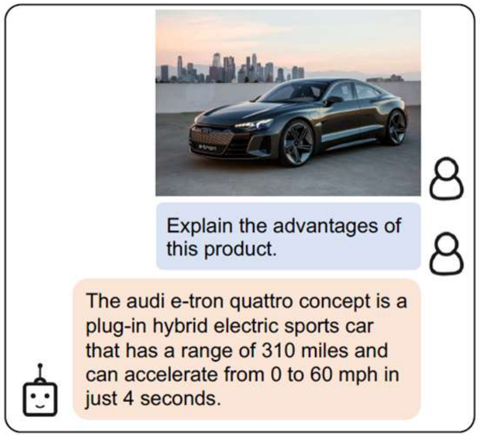

Q: query token positions; T: text token positions.
▨ masked □ unmasked
Q T
Bi-directional Self-Attention Mask
Image-Text Matching
Multi-modal Causal Self-Attention Mask
Image-Grounded Text Generation
Uni-modal Self-Attention Mask
Image-Text Contrastive Learning

BLIP-2

- Image Models:
  - ViT-L/14 from CLIP and
  - ViT-g/14 from EVA-CLIP
- LLMs:
  - OPT for Decoder model and
  - FlanT5 model family for encode
- Dataset 129M images in total, including
  - COCO
  - Visual Genome
  - CC3M
  - CC12M
  - LAION400M
- CapFilt to create synthetic captions

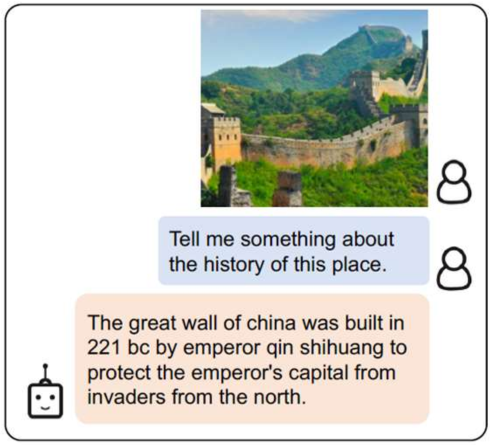
BLIP-2


Bootstrapping from a
Decoder-based
Large Language Model
(e.g. OPT)
Input Image
Image Encoder
Q-Former
Learned Queries
Fully Connected
LLM Decoder
Output Text
a cat wearing sunglasses
Bootstrapping from an
Encoder-Decoder-based
Large Language Model
(e.g. FlanT5)
Input Image
Image Encoder
Q-Former
Learned Queries
Fully Connected
LLM Encoder
LLM Decoder
Suffix Text
wearing sunglasses
a cat
Prefix Text



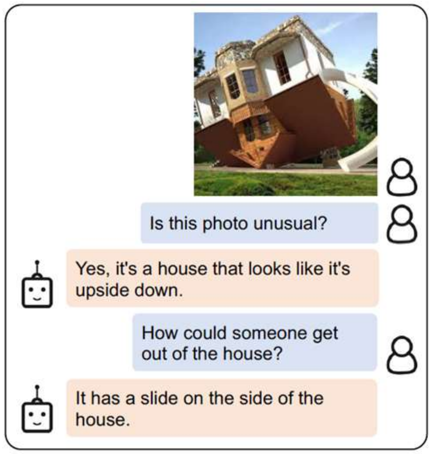
Explain the advantages of this product.

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.



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.

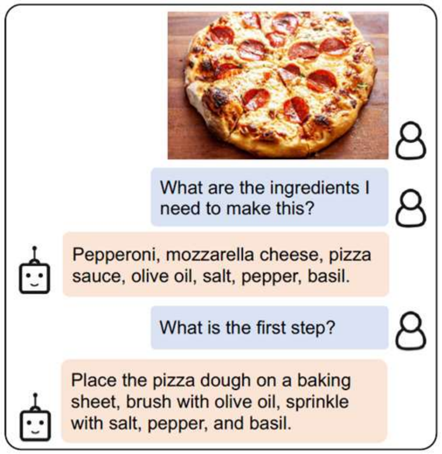


Is this photo unusual?

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.



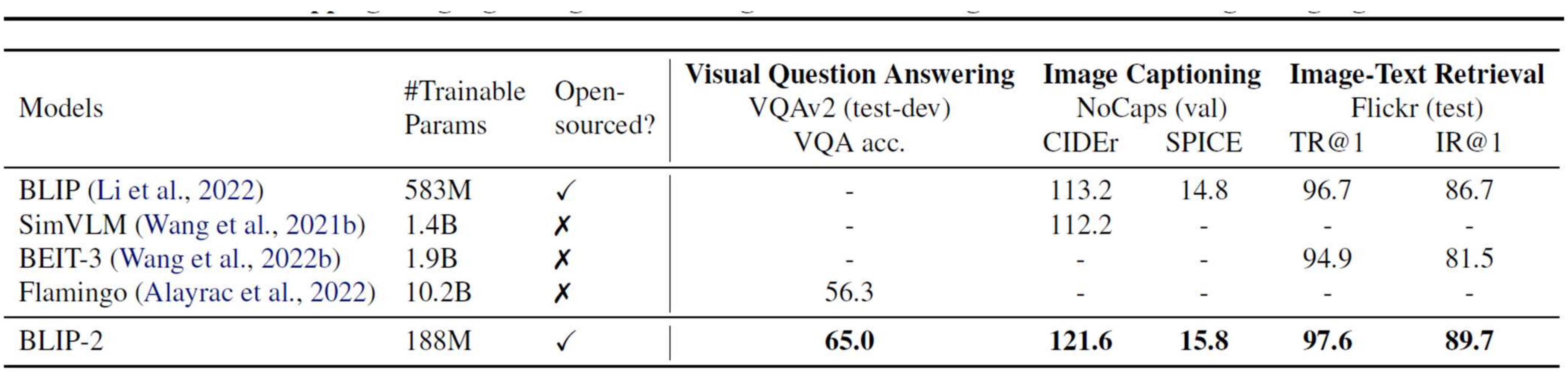
What are the ingredients I need to make this?

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.

Results



Models	#Trainable Params	Open-sourced?	**Visual Question Answering**	**Image Captioning**		**Image-Text Retrieval**	
			VQAv2 (test-dev)	NoCaps (val)		Flickr (test)	
			VQA acc.	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**