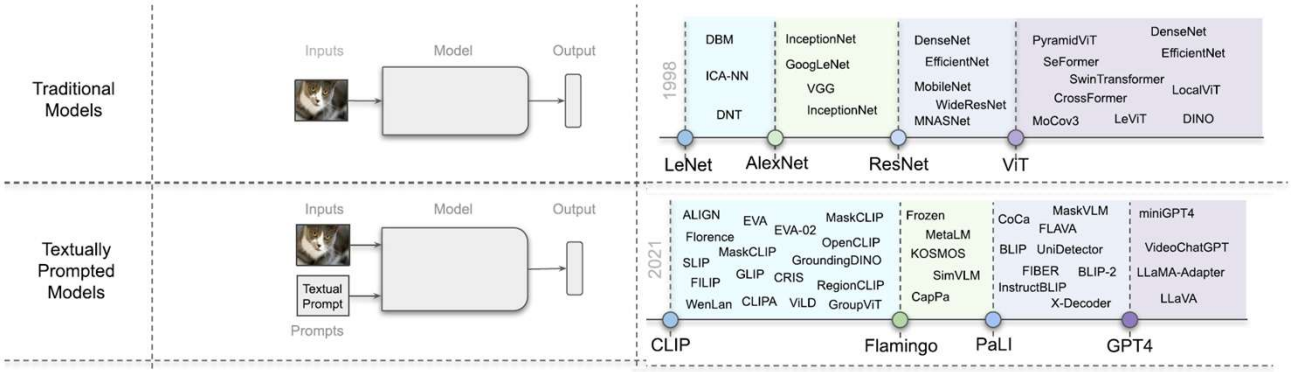
UCF Center for Research in Computer Vision

Visual-Language Models Introduction
Part-I
Lecture-5
CAP6412 Spring 2025

Mubarak Shah

shah@crcv.ucf.edu

Evolution of Computer Vision Models



Foundation Models

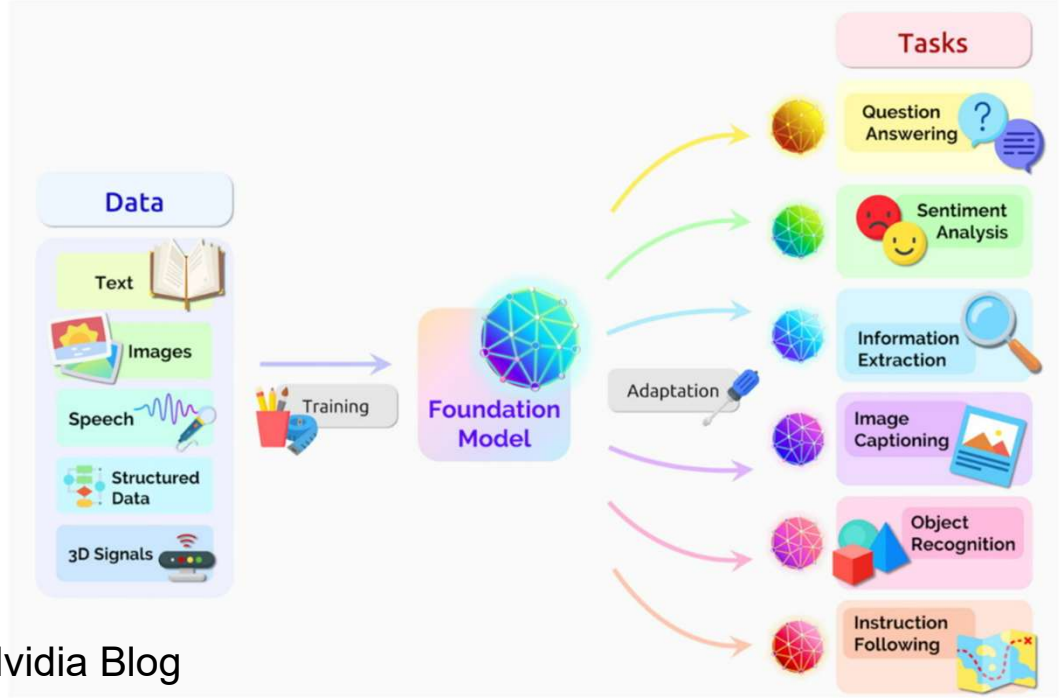


Source: Nvidia Blog

Foundation Models

- They are trained on massive amount of data
- They can be adapted to different downstream tasks

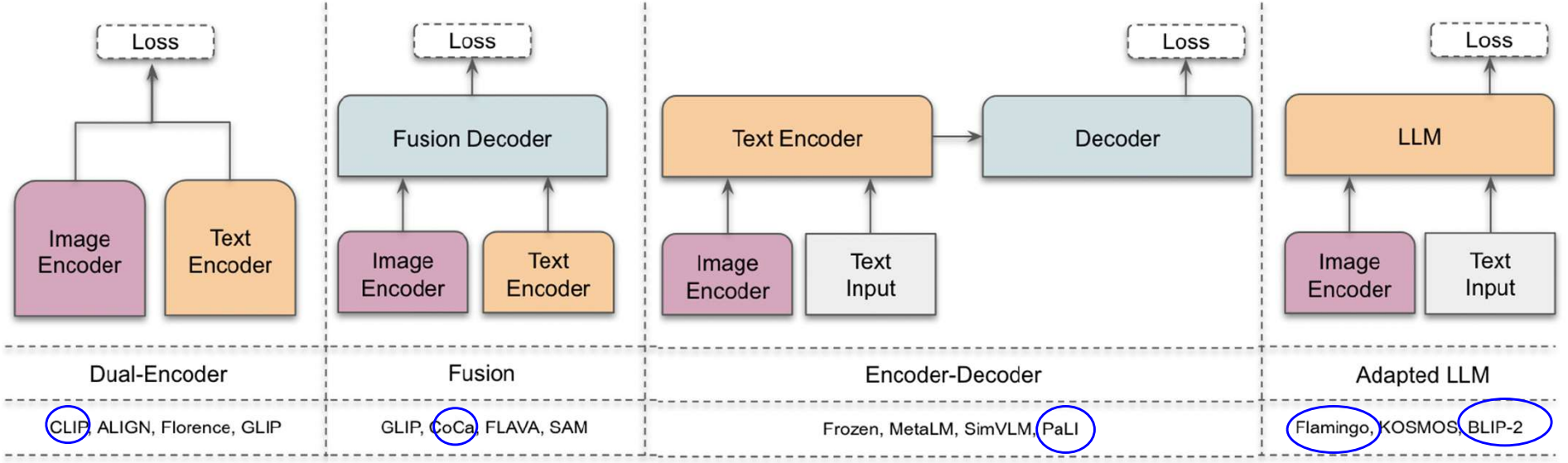
Foundation Models



Source: Nvidia Blog



Different Architecture Styles



UCF Center for Research in Computer Vision

Contents

- CoCa
- PALI
- FLAMINGO
- FLAVA
- Painter
- BLIP-2
- Image-Bind
- Language-Bind
- LLaVA
- Video ChatGPT

UCF Center for Research in Computer Vision



Published in Transactions on Machine Learning Research (08/2022)

CoCa: Contrastive Captioners are Image-Text Foundation Models

Jiahui Yu* *jiahuiyu@google.com*
Zirui Wang* *ziruiw@google.com*
Vijay Vasudevan
Legg Yeung
Mojtaba Seyedhosseini
Yonghui Wu
Google Research
** Equal contribution.*

Reviewed on OpenReview: *https://openreview.net/forum?id=Ee277P3AYC*

Abstract

Exploring large-scale pretrained foundation models is of significant interest in computer vision because these models can be quickly transferred to many downstream tasks. This paper presents **Co**ntrastive **Ca**ptioner (**CoCa**), a minimalist design to pretrain an image-text encoder-decoder foundation model jointly with contrastive loss and captioning loss, thereby subsuming model capabilities from contrastive approaches like CLIP and generative methods like SimVLM. In contrast to standard encoder-decoder transformers where all decoder layers attend to encoder outputs, CoCa omits cross-attention in the first half of decoder layers to encode *unimodal* text representations, and cascades the remaining decoder layers which cross-attend to the image encoder for *multimodal* image-text representations. We apply a contrastive loss between unimodal image and text embeddings, in addition to a captioning loss on the multimodal decoder outputs which predicts text tokens autoregressively. By sharing the same computational graph, the two training objectives are computed efficiently with minimal overhead. CoCa is pretrained end-to-end and from scratch on both web-scale alt-text data and annotated images by treating all labels simply as text, seamlessly unifying natural language supervision for representation learning. Empirically, CoCa achieves state-of-the-art performance with zero-shot transfer or minimal task-specific adaptation on a broad range of downstream tasks, spanning visual recognition (ImageNet, Kinetics-400/600/700, Moments-in-Time), crossmodal retrieval (MSCOCO, Flickr30K, MSR-VTT), multimodal understanding (VQA, SNLI-VE, NLVR2), and image captioning (MSCOCO, NoCaps). Notably on ImageNet classification, CoCa obtains 86.3% *zero-shot* top-1 accuracy, 90.6% with a *frozen encoder* and learned classification head, and 91.0% with a *finetuned encoder*.

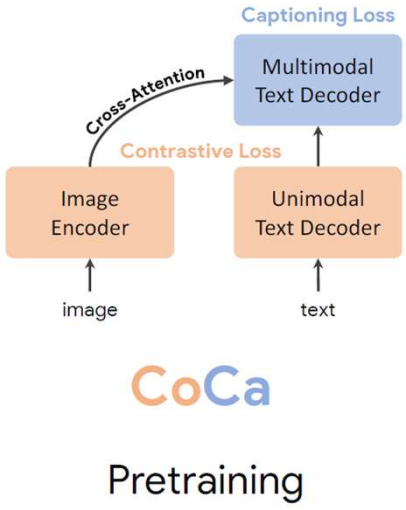
1 Introduction

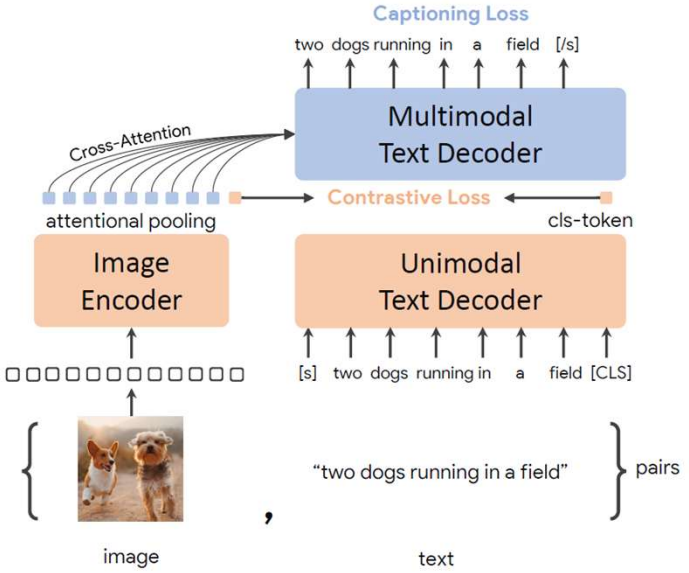
Deep learning has recently witnessed the rise of foundation language models (Bommasani et al., 2021) such as BERT (Devlin et al., 2018), T5 (Raffel et al., 2019), GPT-3 (Brown et al., 2020), where models are pretrained on web-scale data and demonstrate generic multi-tasking capabilities through zero-shot, few-shot or transfer learning. Compared with specialized individual models, pretraining foundation models for massive downstream

1

https://openreview.net/pdf?id=Ee277P3AYC

Co-CA: Contrastive Captioners are Image-Text Foundation Models





Co-CA Losses

Dual-Encoder Contrastive Learning

$$\mathcal{L}_{\text{Con}} = -\frac{1}{N}\Big(\underbrace{\sum_{i}^{N}\log\frac{\exp(x_i^\top y_i/\sigma)}{\sum_{j=1}^{N}\exp(x_i^\top y_j/\sigma)}}_{\text{image-to-text}} + \underbrace{\sum_{i}^{N}\log\frac{\exp(y_i^\top x_i/\sigma)}{\sum_{j=1}^{N}\exp(y_i^\top x_j/\sigma)}}_{\text{text-to-image}}\Big),$$

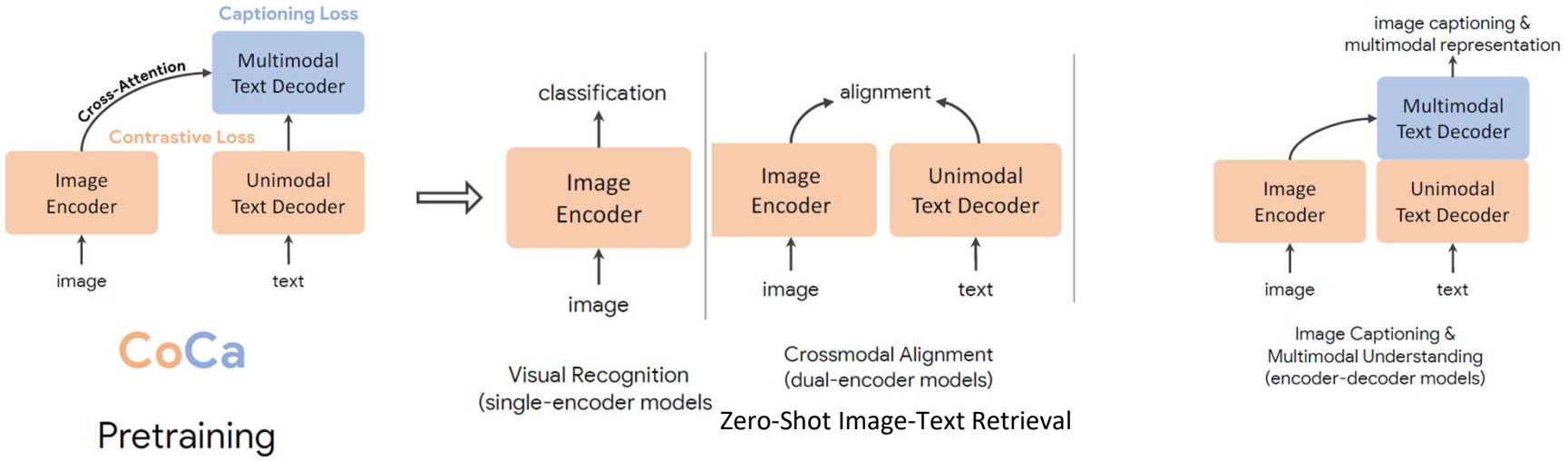
Encoder-Decoder Captioning

$$\mathcal{L}_{\text{Cap}} = -\sum_{t=1}^{T}\log P_\theta(y_t|y_{<t}, x).$$

Contrastive Captioners Pretraining

$$\mathcal{L}_{\text{CoCa}} = \lambda_{\text{Con}}\cdot\mathcal{L}_{\text{Con}} + \lambda_{\text{Cap}}\cdot\mathcal{L}_{\text{Cap}},$$

Co-CA: Contrastive Captioners are Image-Text Foundation Models



Dataset

- JFT-3 Billions
- Larger than JFT-300 Million dataset used in VIT
- 30K Labels and web-scale alt-text data

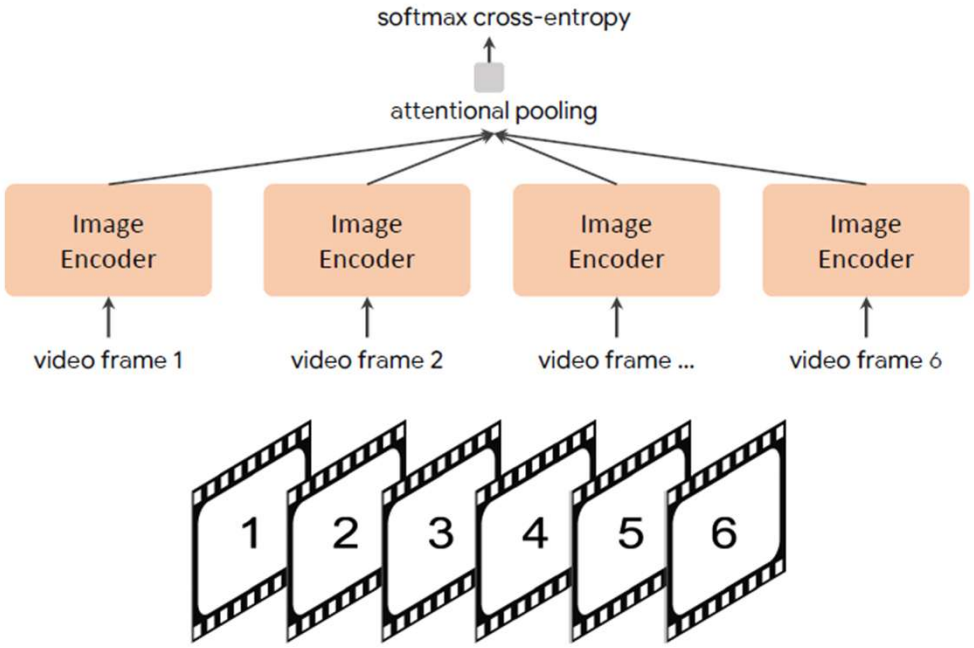
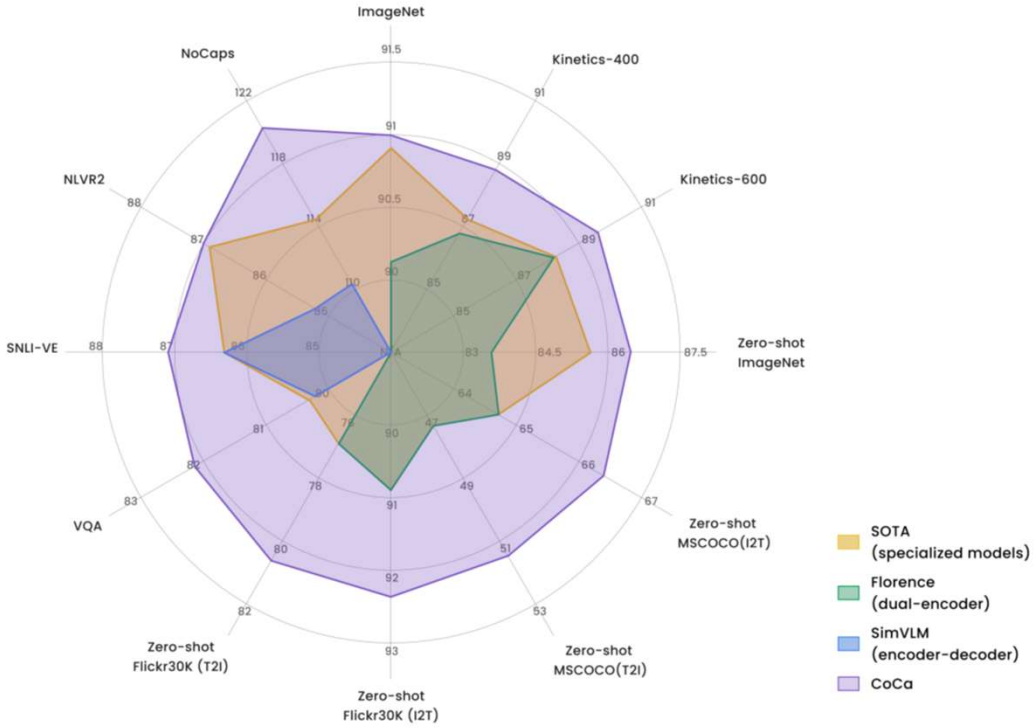


Figure 3: CoCa for video recognition.

Results


ImageNet
Kinetics-400
Kinetics-600
Zero-shot ImageNet
Zero-shot MSCOCO(I2T)
Zero-shot MSCOCO(T2I)
Zero-shot Flickr30K (I2T)
Zero-shot Flickr30K (T2I)
VQA
SNLI-VE
NLVR2
NoCaps
SOTA (specialized models)
Florence (dual-encoder)
SimVLM (encoder-decoder)
CoCa

SNLI-VE: Visual Entailment Dataset



Premise

+

- *Two woman are holding packages.*
- *The sisters are hugging goodbye while holding to go packages after just eating lunch.*
- *The men are fighting outside a deli.*

Hypothesis

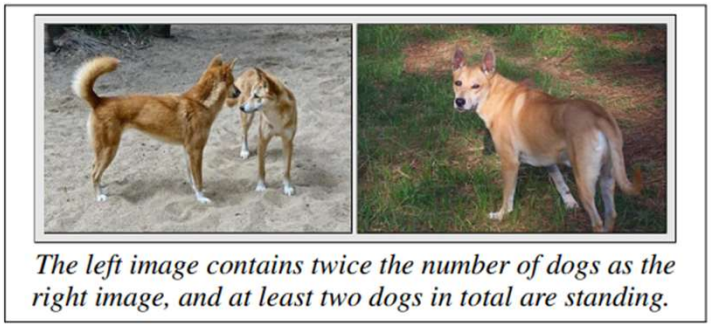
=

- *Entailment*
- *Neutral*
- *Contradiction*

Answer

NLVR2

- Each caption is paired with two images.
- The task is to predict if the caption is True or False



The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.



One image shows exactly two brown acorns in back-to-back caps on green foliage.

arXiv:2209.06794v4 [cs.CV] 5 Jun 2023

Published as a conference paper at ICLR 2023

PaLI: A Jointly-Scaled Multilingual Language-Image Model

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski
Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer
Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari
Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo
Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme
Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, Radu Soricut
Google Research*

Abstract

Effective scaling and a flexible task interface enable large language models to excel at many tasks. We present **PaLI** (**Pa**thways **L**anguage and **I**mage model), a model that extends this approach to the joint modeling of language and vision. PaLI generates text based on visual and textual inputs, and with this interface performs many vision, language, and multimodal tasks, in many languages. To train PaLI, we make use of large pre-trained encoder-decoder language models and Vision Transformers (ViTs). This allows us to capitalize on their existing capabilities and leverage the substantial cost of training them. We find that joint scaling of the vision and language components is important. Since existing Transformers for language are much larger than their vision counterparts, we train a large, 4-billion parameter ViT (ViT-e) to quantify the benefits from even larger-capacity vision models. To train PaLI, we create a large multilingual mix of pre-training tasks, based on a new image-text training set containing 10B images and texts in over 100 languages. PaLI achieves state-of-the-art in multiple vision and language tasks (such as captioning, visual question-answering, scene-text understanding), while retaining a simple, modular, and scalable design.

1 Introduction

Increasing neural network capacity has been a successful trend in the modeling of language and vision tasks. On the language side, models such as T5 (Raffel et al., 2020), GPT-3 (Brown et al., 2020), Megatron-Turing (Shoeybi et al., 2019), GLaM (Du et al., 2022), Chinchilla (Hoffmann et al., 2022), and PaLM (Chowdhery et al., 2022) have shown significant advantages from training large Transformers on large amounts text data. On the vision side, CNNs (Mahajan et al., 2018; Huang et al., 2019; Kolesnikov et al., 2020), Vision Transformers (Dosovitskiy et al., 2021), and other models (Tolstikhin et al., 2021; Riquelme et al., 2021) have seen similar benefits from scale (Zhai et al., 2022a), albeit to a lesser extent than in language. Language-and-vision modeling has followed a similar trend, e.g., SimVLM (Wang et al., 2021), Florence (Yuan et al., 2021), CoCa (Yu et al., 2022), GIT (Wang et al., 2022a), BEiT-3 (Wang et al., 2022c), and Flamingo (Alayrac et al., 2022).

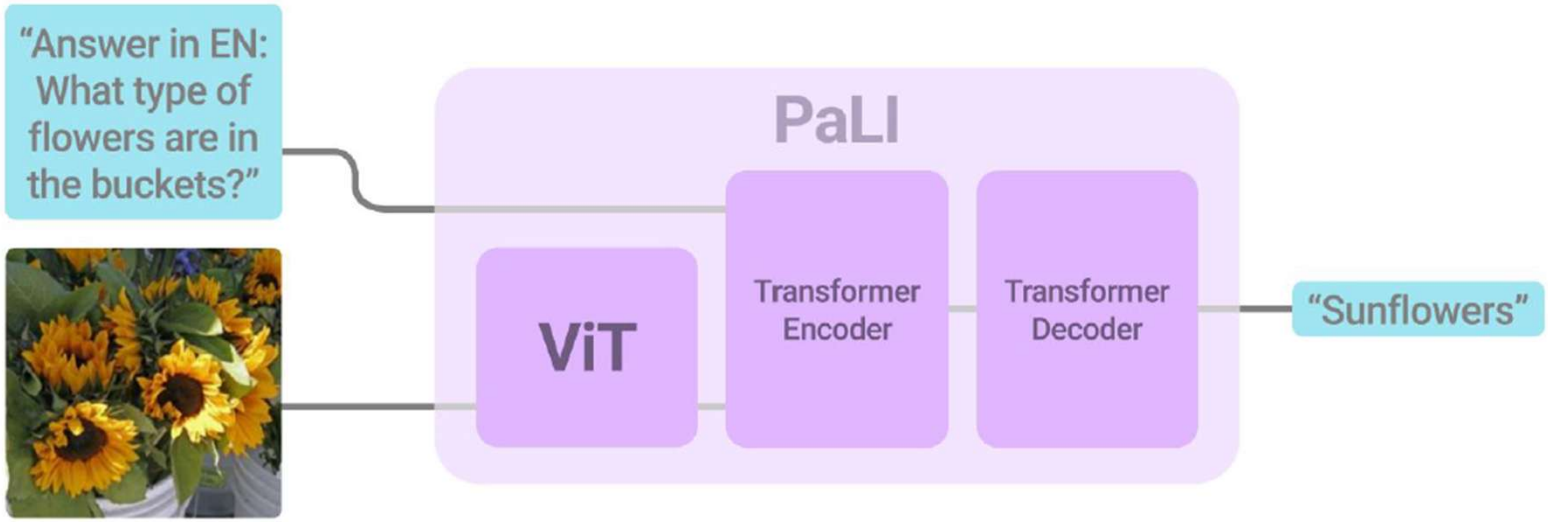
We introduce PaLI, a model that performs image-only, language-only, and image+language tasks across many languages, using a single "image-and-text to text" interface. A key characteristic of PaLI is a more balanced parameter share between the language and vision components, with more capacity to the vision backbone yielding large gains in performance. Another key ingredient to PaLI is the reuse of large unimodal backbones for language and vision modeling, in order to transfer existing capabilities and reduce training cost. On the language side, we reuse the 13B-parameter model mT5-XXL (Xue et al., 2021), which already packages language understanding and generation capabilities. We show that these capabilities are maintained and extended into a multimodal setting. On the vision side, in addition to reusing the 2B-parameter ViT-G model (Zhai et al., 2022a), we

*Correspondence: pali-communications@google.com

PALI (Pathways Language and Image model)

- A model that performs
  - image-only,
  - language-only, and
  - image+language tasks
- Across many languages, using a single “image-and-text to text” interface.

PALI (Pathways Language and Image model)



PALI

- 4-billion parameter ViT (ViT-e)
- 13B-parameter text model mT5-XXL
- WebLI10B Dataset
  - images and texts in over 100 languages
  - 10 billion images and tens of billions of image-text pairs.

Training Mixture

- Span corruption on text-only data
- Split-captioning on WebLI alt-text data
- Captioning on CC3M-35L
- OCR on WebLI OCR-text data
- English and Cross-Lingual VQA
- English and Cross-Lingual visual question generation (VQG)
- English-only Object-Aware (OA) VQA
- Object detection

Results

- Image Captioning
- Multilingual captioning
- VISUAL QUESTION ANSWERING
- Cross-lingual and Multilingual VQA
- ZERO-SHOT IMAGE CLASSIFICATION



Input: Generate the alt_text in EN
Output: A cellar filled with barrels of wine



Input: Generate the alt_text in EN
Output: a clock on a building that says 'lyvania' on it



Input: Generate the alt_text in EN
Output: Two helicopters are flying in the sky and one has a yellow stripe on the tail



Input: Generate the alt_text in EN
Output: A cellar filled with barrels of wine



Input: Generate the alt_text in EN
Output: a clock on a building that says 'lyvania' on it



Input: Generate the alt_text in EN
Output: Two helicopters are flying in the sky and one has a yellow stripe on the tail



Input: Generate the alt_text in FR
Output: Un arbre debout dans un champ avec un ciel violet

(A tree standing in a field with a purple sky)



Input: Generate the alt_text in TH
Output: ลา สี เทา เดิน ไป ตาม ถนน

(A gray donkey walks down the street)



Input: Generate the alt_text in ZH
Output: 一辆 电动 汽车 停 在 充电 桩 上 。

(An electric car parking on a charging station)



Input: Answer in EN: what time is it according to this radio
Output: 1254



Input: Answer in EN: what website is on the wall in back
Output: arsenaldirect.com



Input: Answer in EN: what is the brand of this watch
Output: seiko

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

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

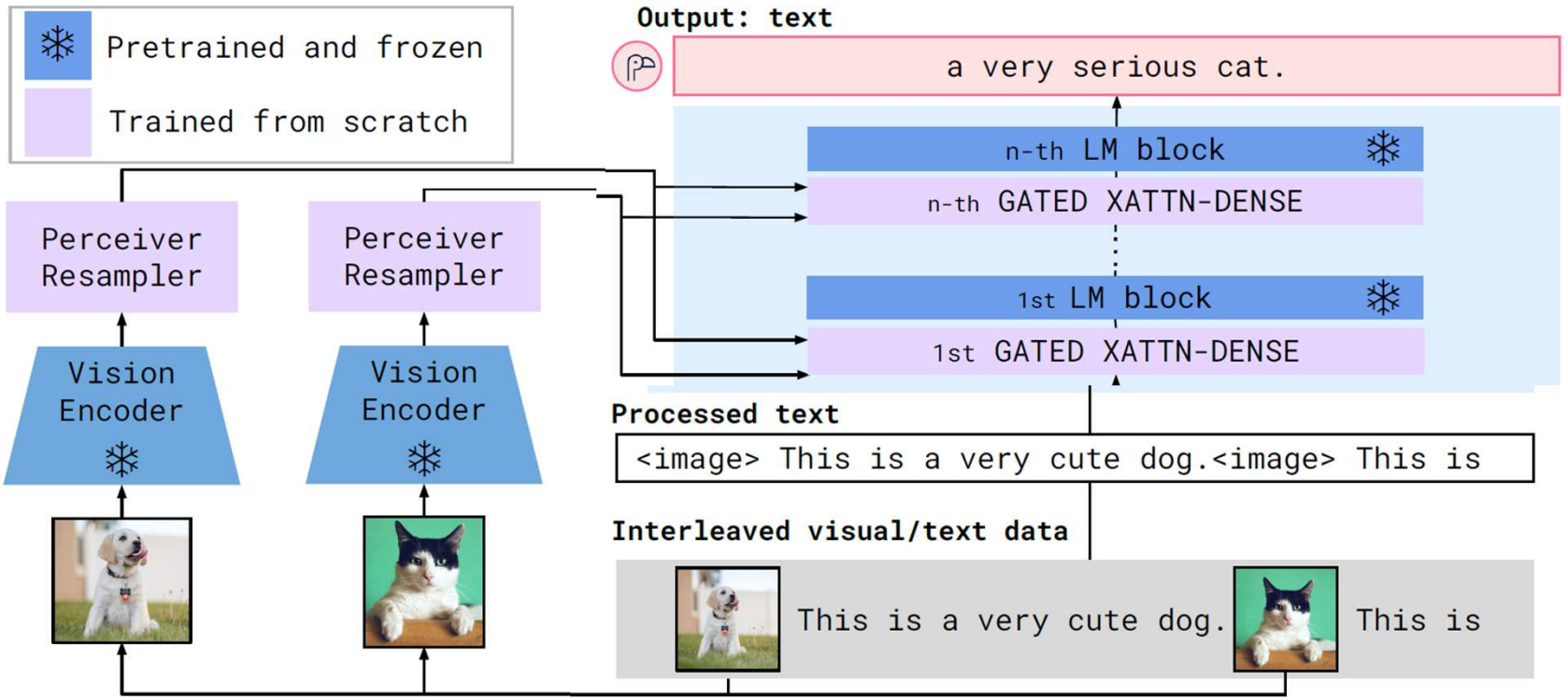
DeepMind

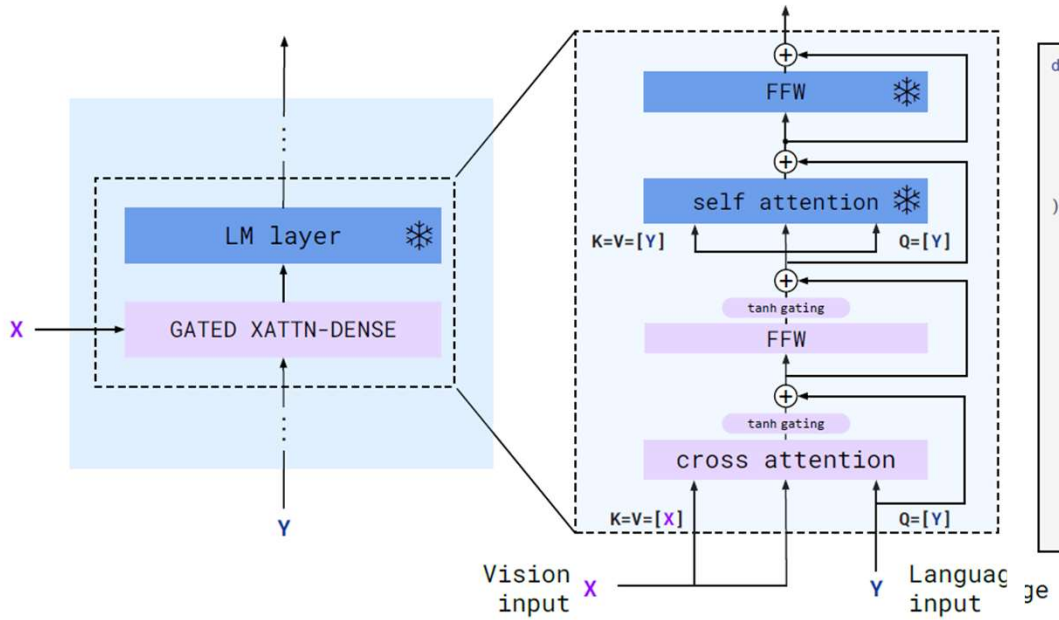
Abstract

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).

FALMINGO


Pretrained and frozen
Trained from scratch
Output: text
a very serious cat.
n-th LM block
n-th GATED XATTN-DENSE
1st LM block
1st GATED XATTN-DENSE
Perceiver Resampler
Perceiver Resampler
Vision Encoder
Vision Encoder
Processed text
<image> This is a very cute dog.<image> This is
Interleaved visual/text data
This is a very cute dog.
This is




```
def gated_xattn_dense(
    y,  # input language features
    x,  # input visual features
    alpha_xattn, # xattn gating parameter – init at 0.
    alpha_dense, # ffw gating parameter – init at 0.
):
  """Applies a GATED XATTN-DENSE layer."""

  # 1. Gated Cross Attention
  y = y + tanh(alpha_xattn) * attention(q=y, kv=x)
  # 2. Gated Feed Forward (dense) Layer
  y = y + tanh(alpha_dense) * ffw(y)

  # Regular self-attention + FFW on language
  y = y + frozen_attention(q=y, kv=y)
  y = y + frozen_ffw(y)
  return y  # output visually informed language features
```

FLAMINGO

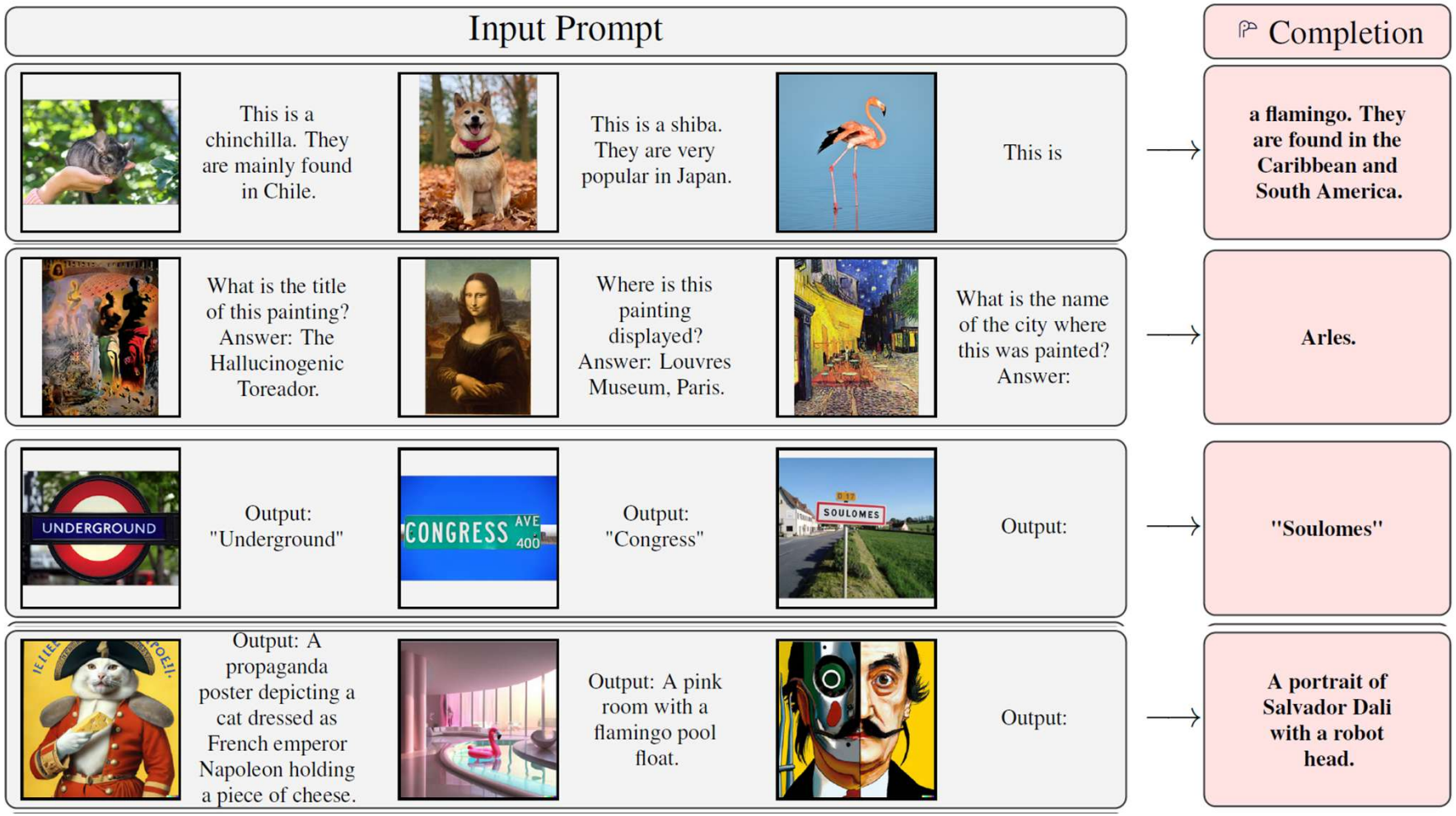
- 70B Chinchilla language model
- ResNet Vision encoder

- 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)
    - 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.
UNDERGROUND
Output: "Underground"
CONGRESS AVE
Output: "Congress"
SOULOMES
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.

UCF Center for Research in Computer Vision



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.

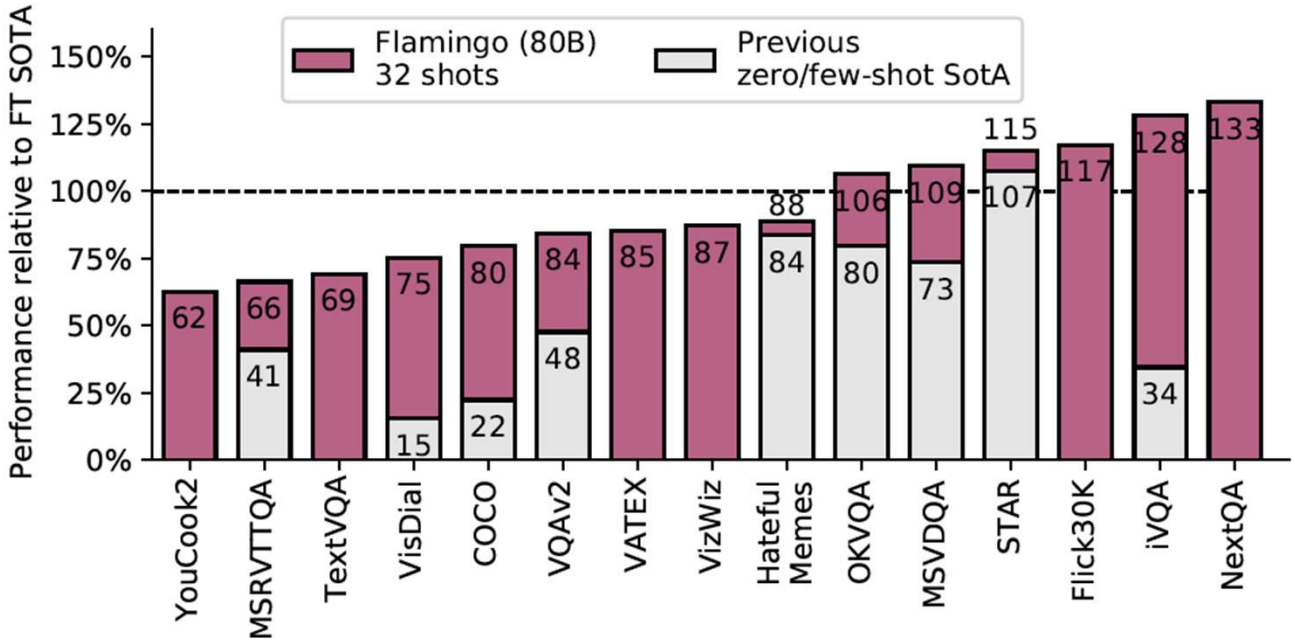
What makes you think this is Chicago?

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.


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

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

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)

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.

*Equal contribution.

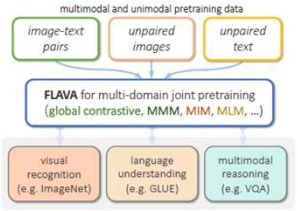
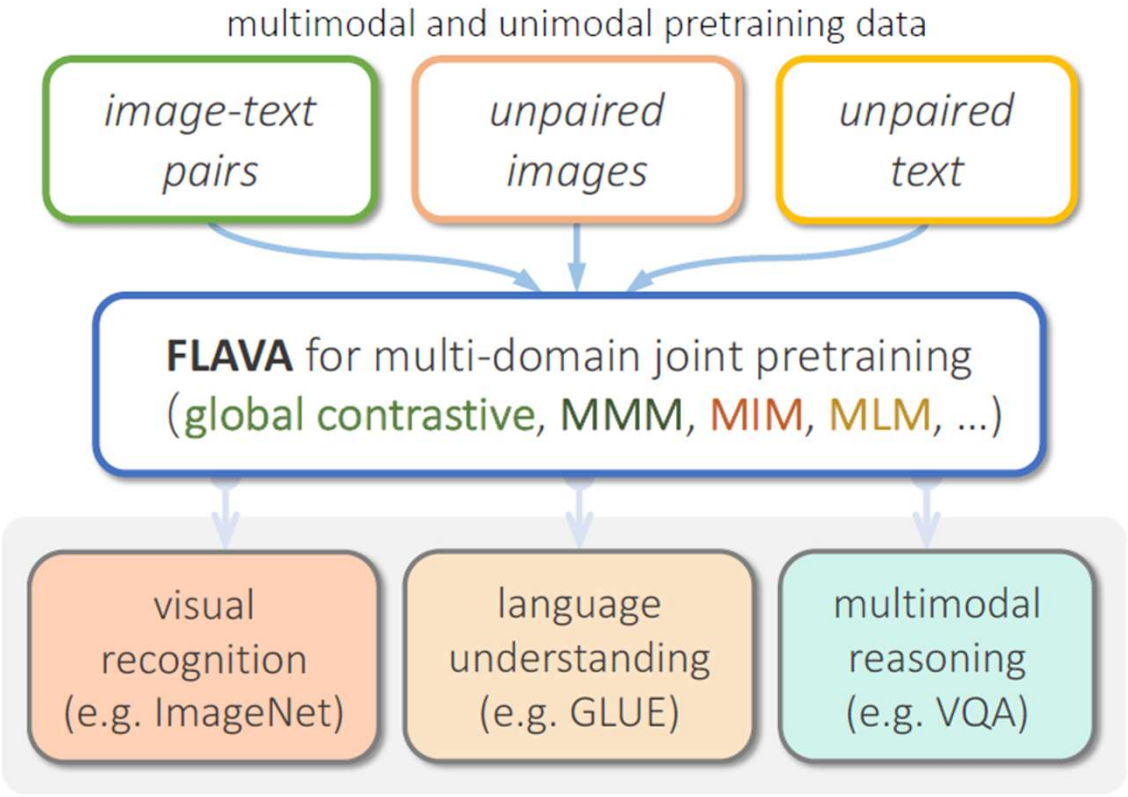


Figure 1. We present FLAVA, a language and vision alignment model that learns strong representations from multimodal (image-text pairs) and unimodal data (unpaired images and text) and can be applied to target a broad scope of tasks from three domains (visual recognition, language understanding, and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


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

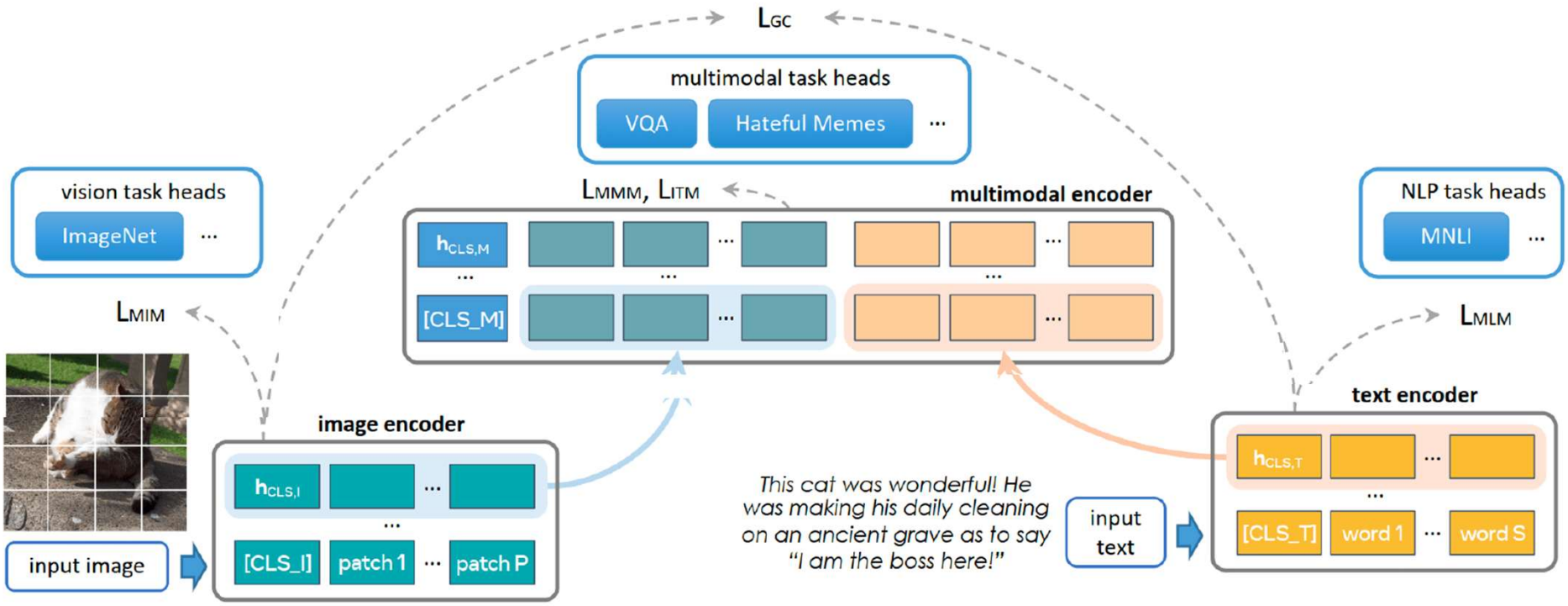

LGC
multimodal task heads
VQA
Hateful Memes
...
vision task heads
ImageNet
...
LMMM, LITM
multimodal encoder
NLP task heads
MNLI
...
LMIM
LMLM
[CLS_M]
text encoder
image encoder
[CLS_I]
patch 1
patch P
This cat was wonderful! He was making his daily cleaning on an ancient grave as to say "I am the boss here!"
input text
[CLS_T]
word 1
word S
input image

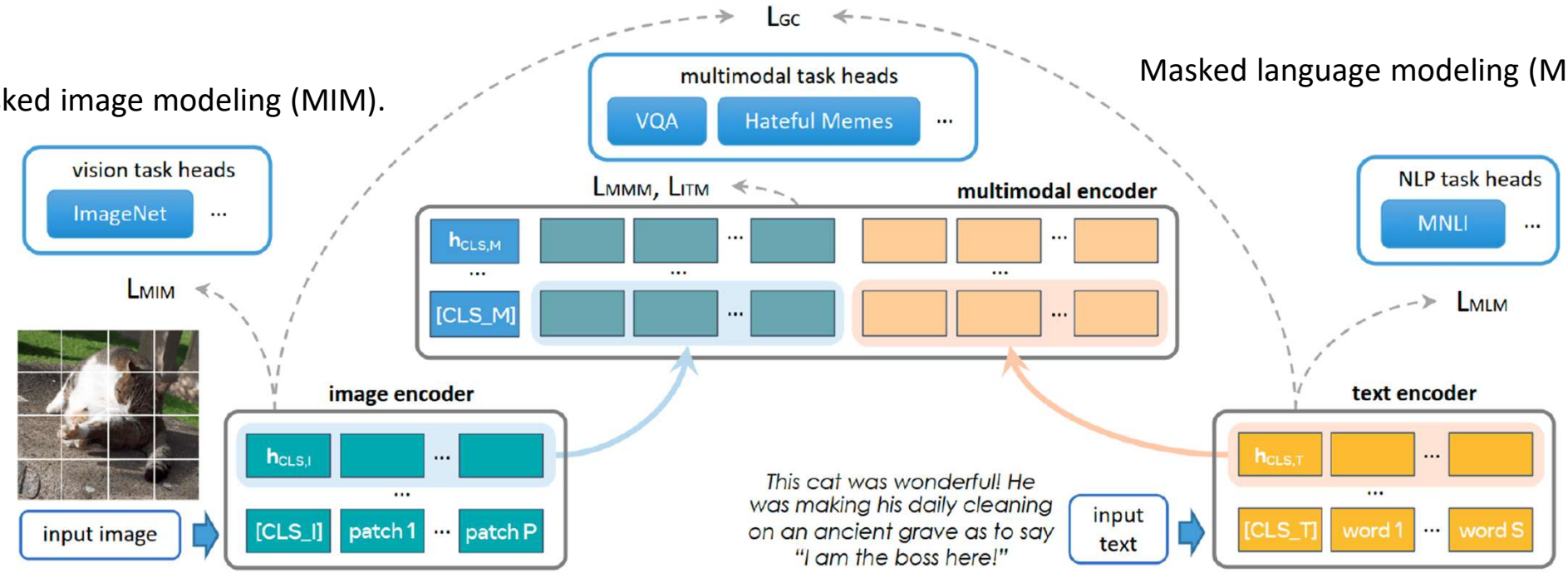
Image-text matching (ITM).

Masked multimodal modeling (MMM).

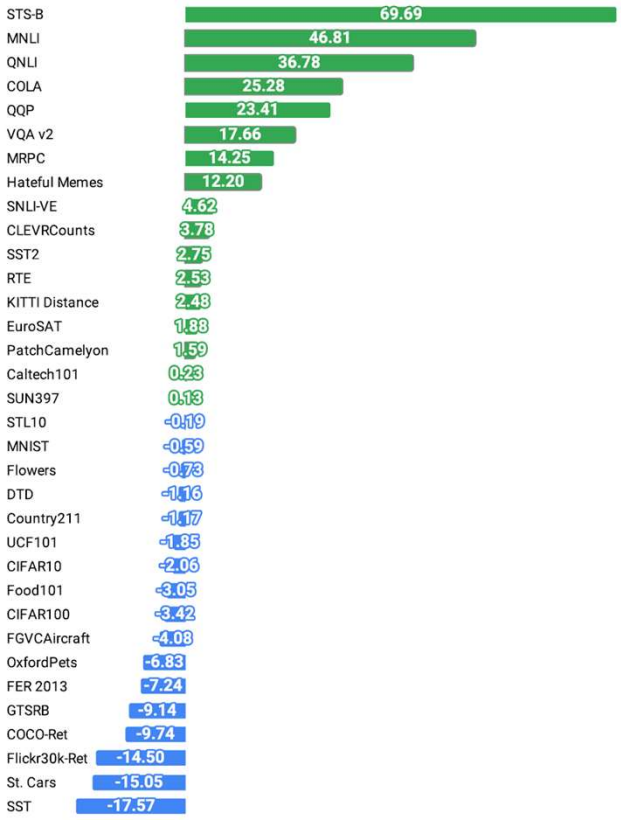
Global contrastive (GC) loss

Masked image modeling (MIM).

Masked language modeling (MLM).



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

arXiv:2212.02499v2 [cs.CV] 24 Mar 2023

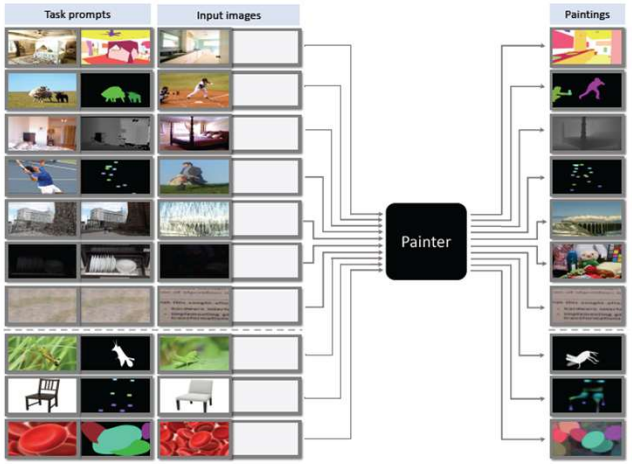


Figure 1. An illustration of the in-context inference of Painter. Painter is a generalist vision model, which can automatically perform vision tasks according to the input task prompts without the task specific heads. Painter can not only perform in-domain tasks with highly competitive performance, such 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).

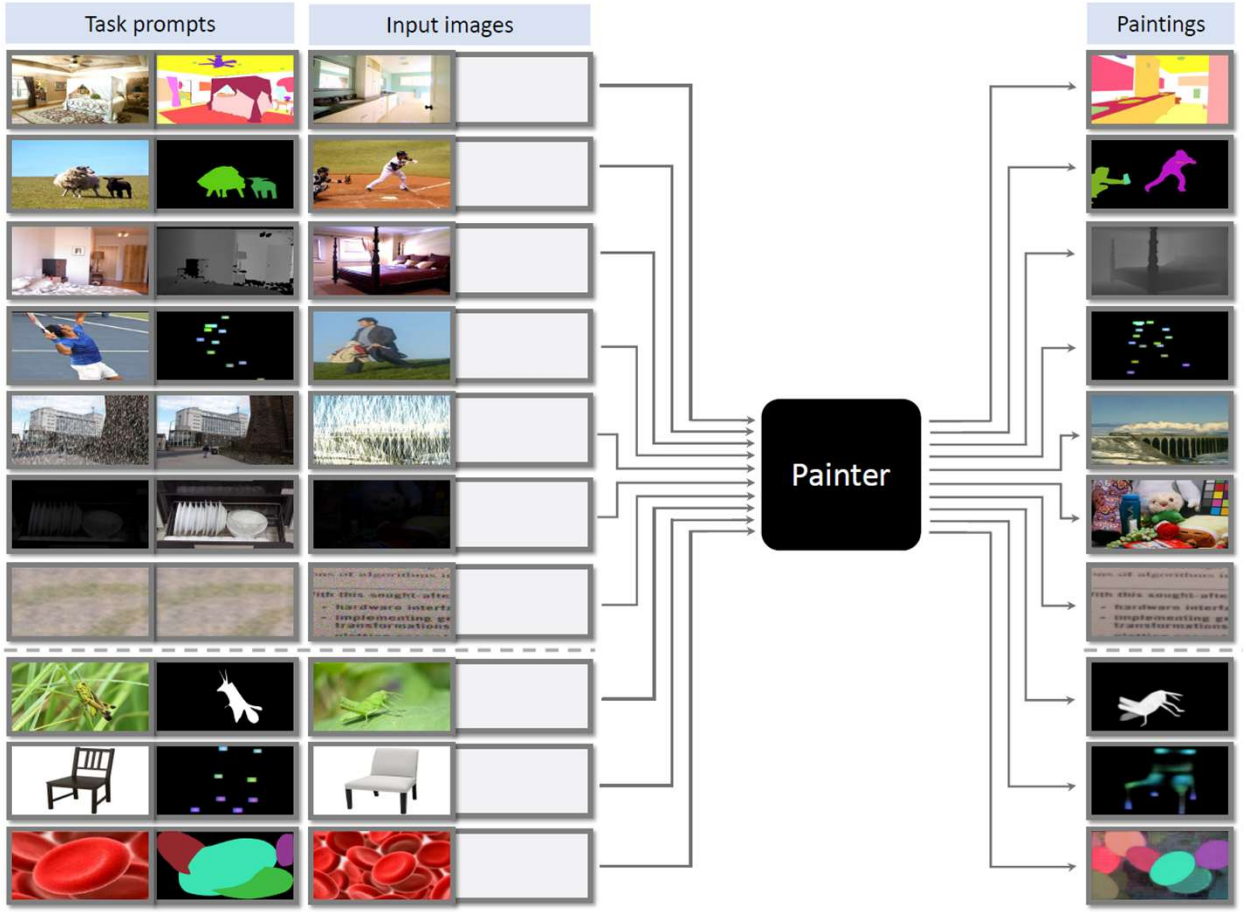
Abstract

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.

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

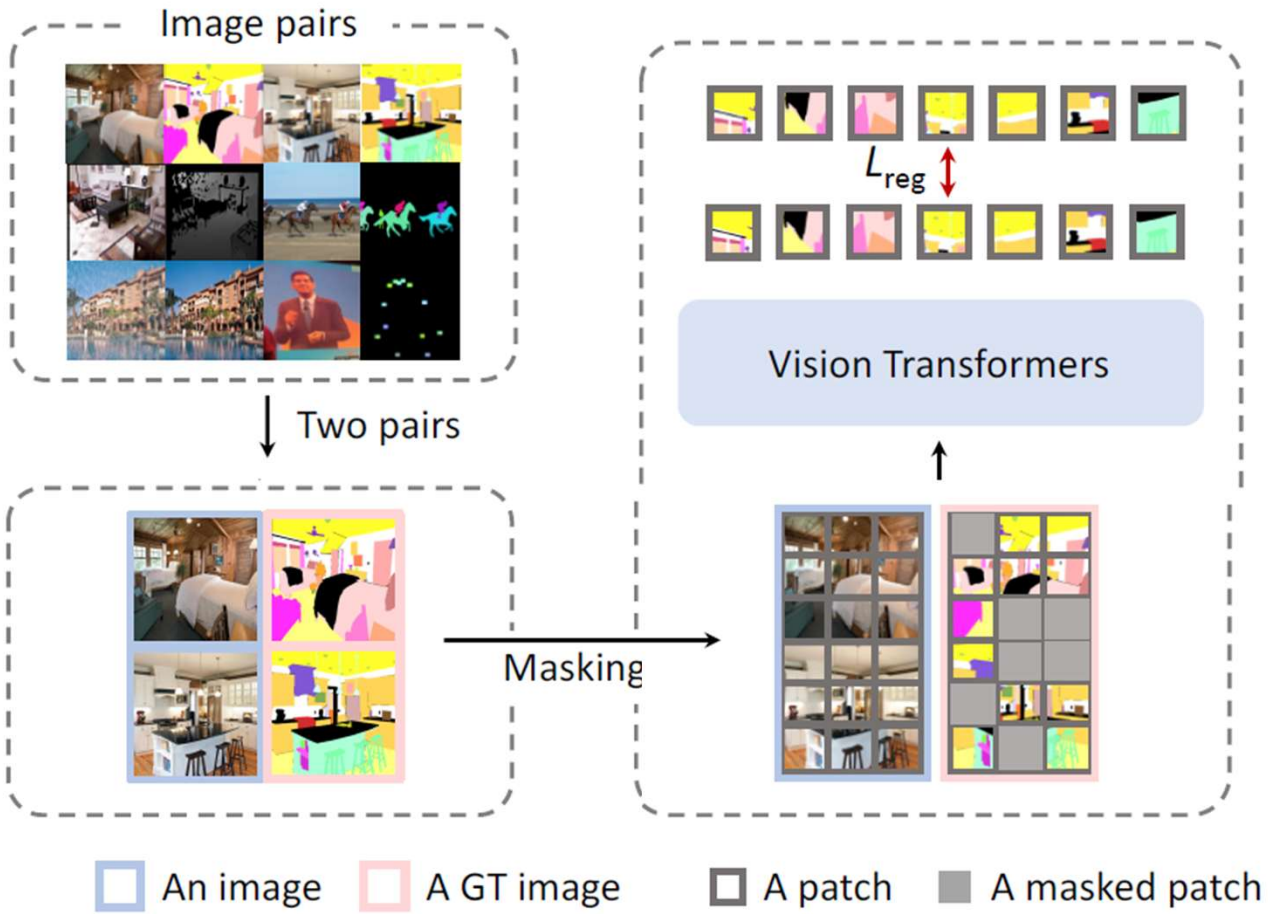
Image enhancement

open-category

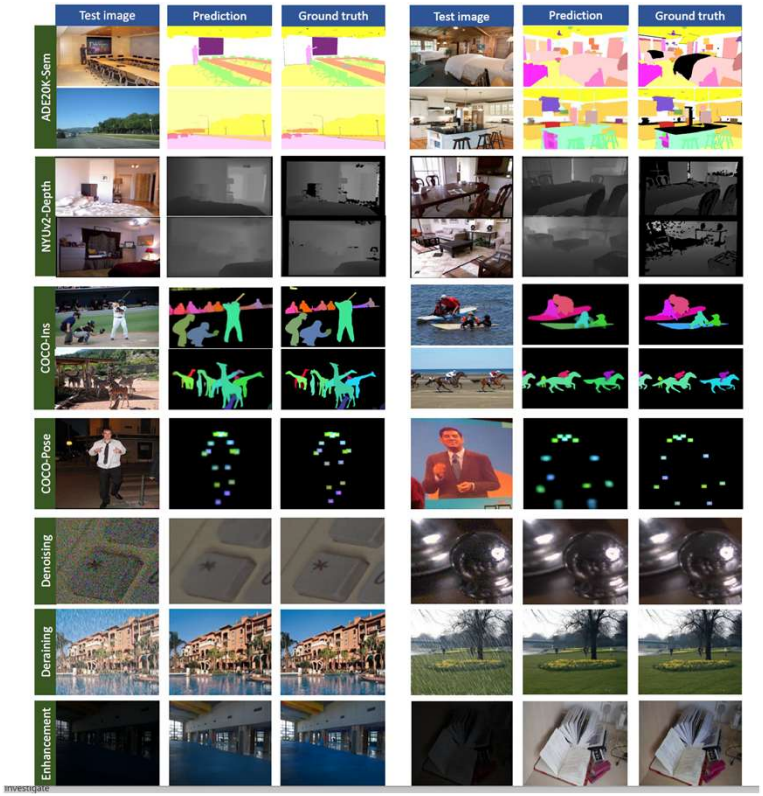
Object segmentation

Keypoint detection

Instance segmentation


Image pairs
Two pairs
Masking
L_{reg}
Vision Transformers
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

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.

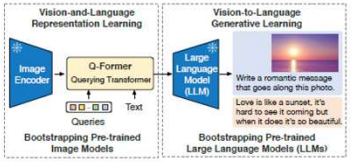


Figure 1. Overview of BLIP-2's framework. We pre-train a lightweight Querying Transformer following a two-stage strategy to bridge the modality gap. The first stage bootstraps vision-language representation learning from a frozen image encoder. The second stage bootstraps vision-to-language generative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

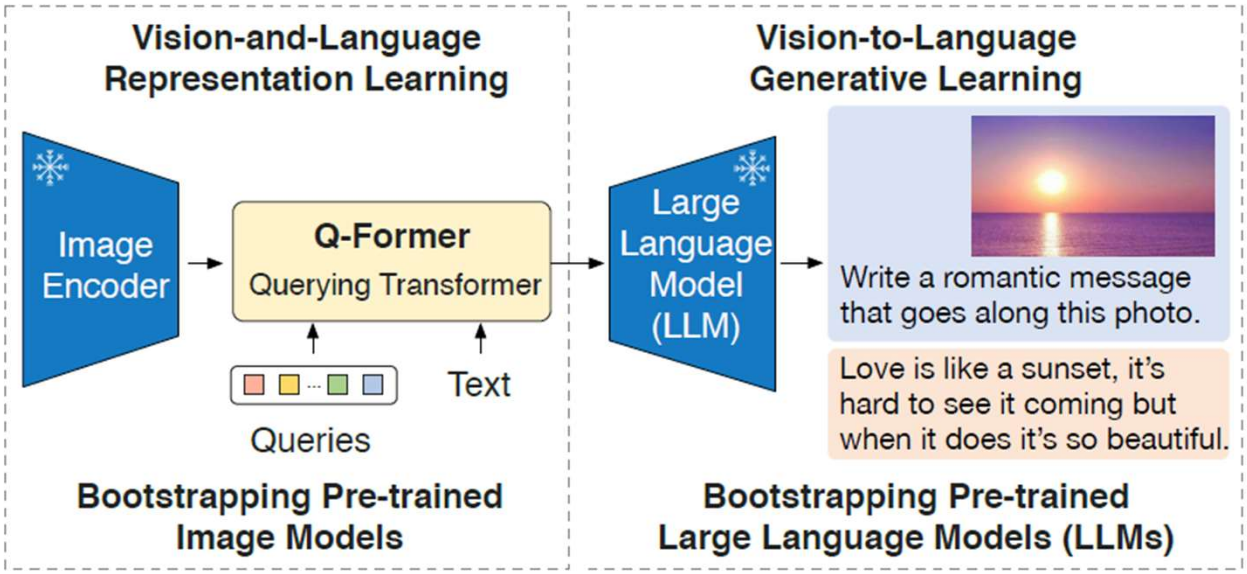
In order to leverage pre-trained unimodal models for VLP, it is key to facilitate cross-modal alignment. However, since LLMs have not seen images during their unimodal pre-training, freezing them makes vision-language alignment in particular challenging. In this regard, existing methods (*e.g.* Frozen (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

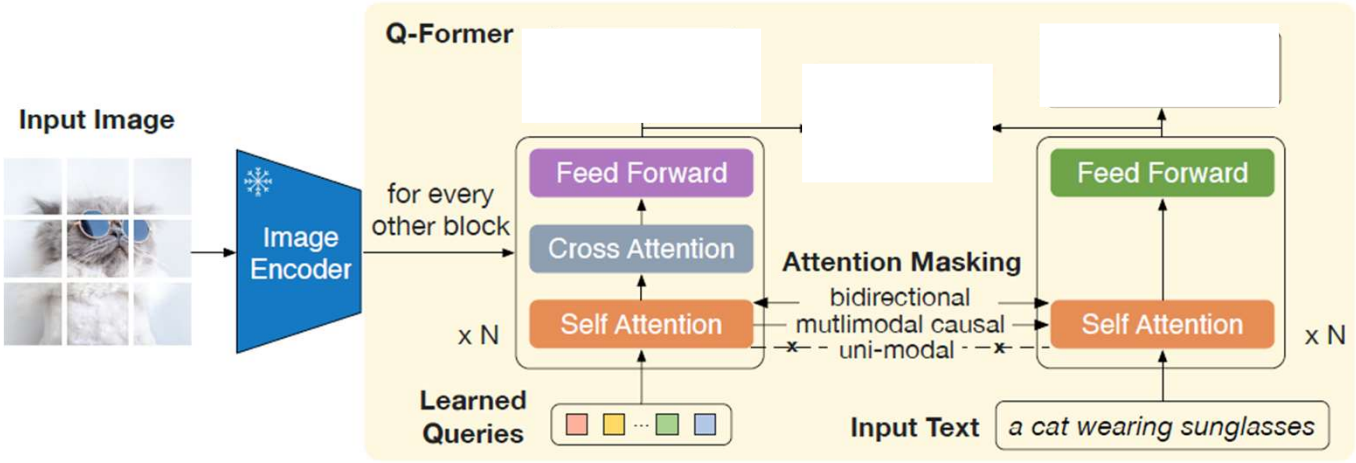
[1]Salesforce Research. Correspondence to: Junnan Li <junnan.li@salesforce.com>.

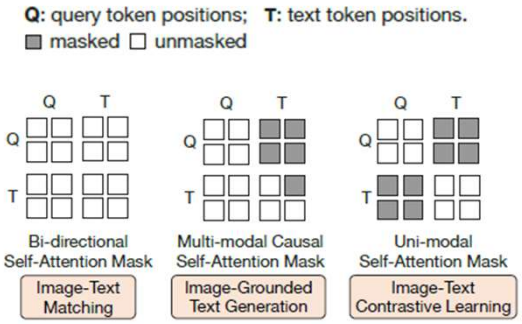
Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

BLIP-2

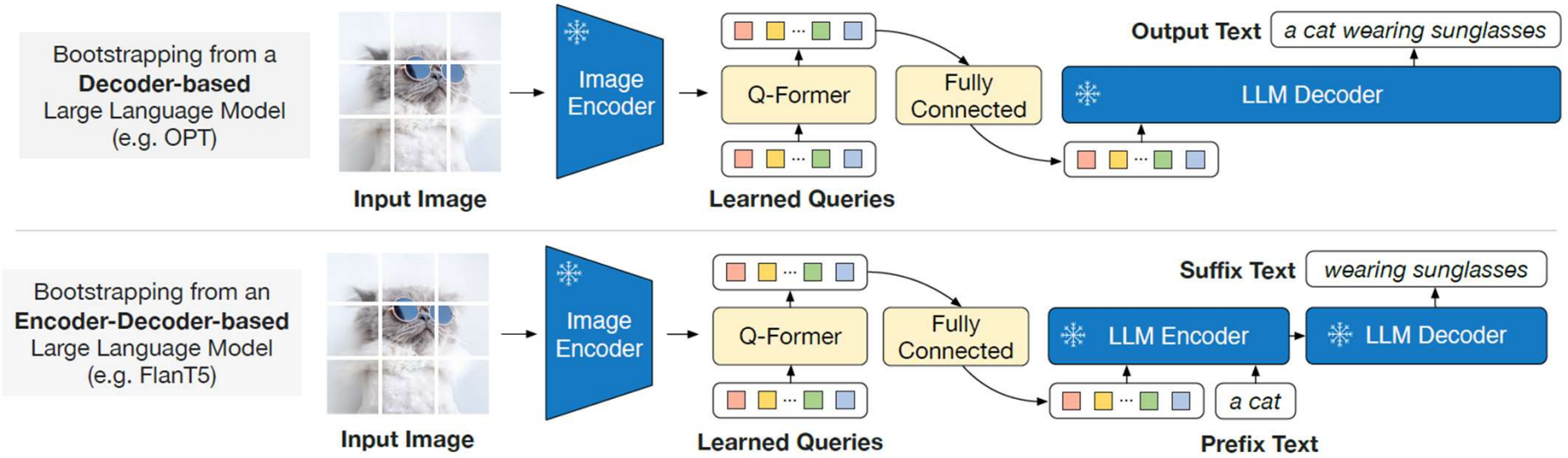

Vision-and-Language Representation Learning
Vision-to-Language Generative Learning
Image Encoder
Q-Former
Querying Transformer
Text
Queries
Large Language Model (LLM)
Write a romantic message that goes along this photo.
Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.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

BLIP-2


Input Image
Image Encoder
Q-Former
for every other block
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

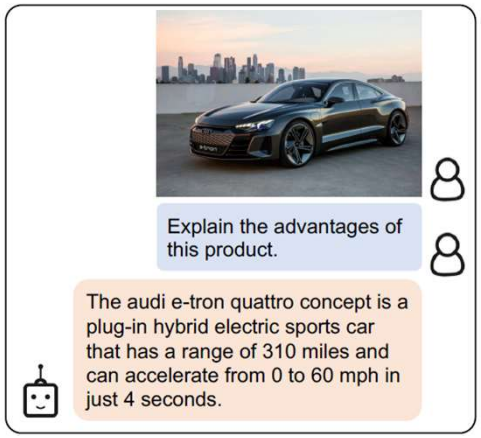

Q: query token positions; T: text token positions.
masked unmasked
Q T
Bi-directional Self-Attention Mask
Image-Text Matching
Multi-modal Causal Self-Attention Mask
Image-Grounded Text Generation
Uni-modal Self-Attention Mask
Image-Text Contrastive Learning

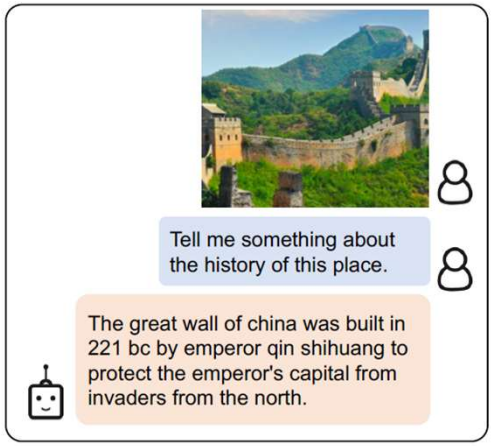
BLIP-2

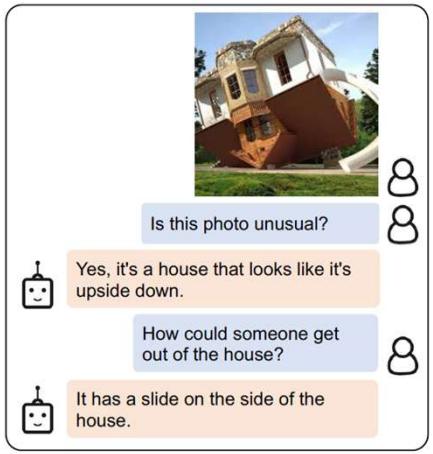

Bootstrapping from a
Decoder-based
Large Language Model
(e.g. OPT)
Input Image
Image Encoder
Q-Former
Learned Queries
Fully Connected
Output Text
a cat wearing sunglasses
LLM Decoder
Bootstrapping from an
Encoder-Decoder-based
Large Language Model
(e.g. FlanT5)
Input Image
Image Encoder
Q-Former
Learned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

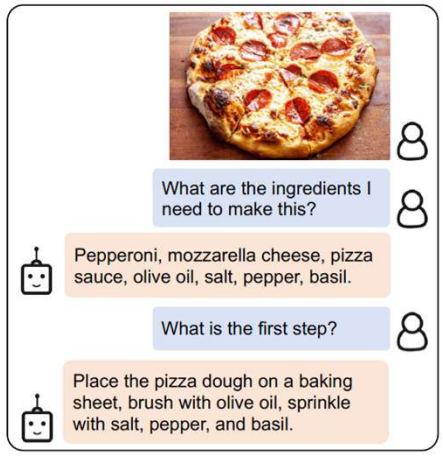
BLIP-2

- 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


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** VQAv2 (test-dev) VQA acc.	**Image Captioning** NoCaps (val) CIDEr	SPICE	**Image-Text Retrieval** Flickr (test) 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**

UCF Center for Research in Computer Vision

X-Former: Unifying Contrastive and Reconstruction Learning for MLLMs

Sirnam Swetha[1], Jinyu Yang[2], Tal Neiman[2], Mamshad Nayeem Rizve[2], Son Tran[2], Benjamin Yao[2], Trishul Chilimbi[2], and Mubarak Shah[1,2]

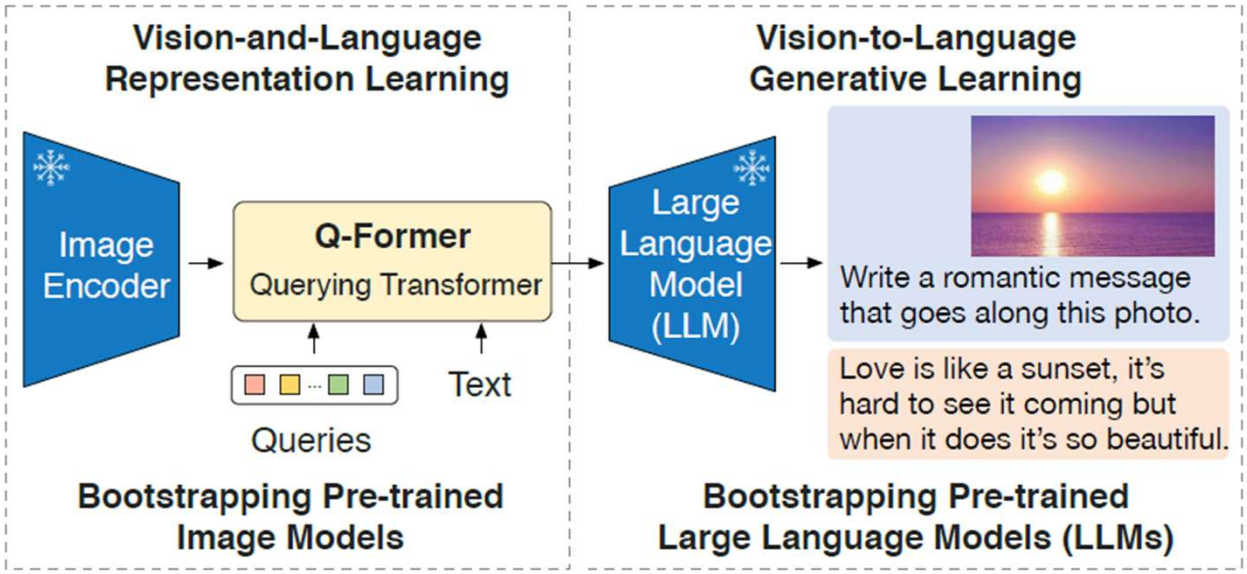
[1] Center for Research in Computer Vision, University of Central Florida
swetha.sirnam@ucf.edu, shah@crcv.ucf.edu
[2] Amazon
{viyjy, taneiman, mnrizve, sontran, benjamy, trishulc}@amazon.com

Abstract. Recent advancements in Multimodal Large Language Models (MLLMs) have revolutionized the field of vision-language understanding by integrating visual perception capabilities into Large Language Models (LLMs). The prevailing trend in this field involves the utilization of a vision encoder derived from vision-language contrastive learning (CL), showing expertise in capturing overall representations while facing difficulties in capturing detailed local patterns. In this work, we focus on enhancing the visual representations for MLLMs by combining high-frequency and detailed visual representations, obtained through masked image modeling (MIM), with semantically-enriched low-frequency representations captured by CL. To achieve this goal, we introduce X-Former which is a lightweight transformer module designed to exploit the complementary strengths of CL and MIM through an innovative interaction mechanism. Specifically, X-Former first bootstraps vision-language representation learning and multimodal-to-multimodal generative learning from two frozen vision encoders, i.e., CLIP-ViT (CL-based) and MAE-ViT (MIM-based). It further bootstraps vision-to-language generative

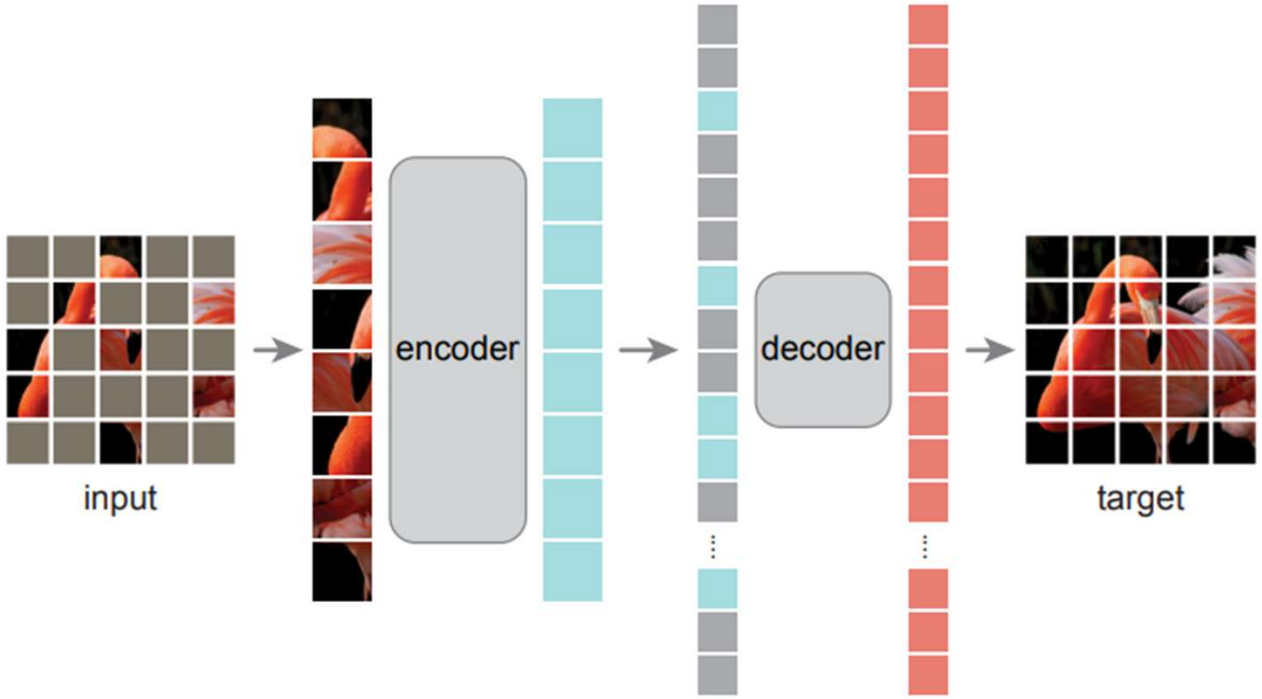
BLIP-2

CLIP Image Encoder

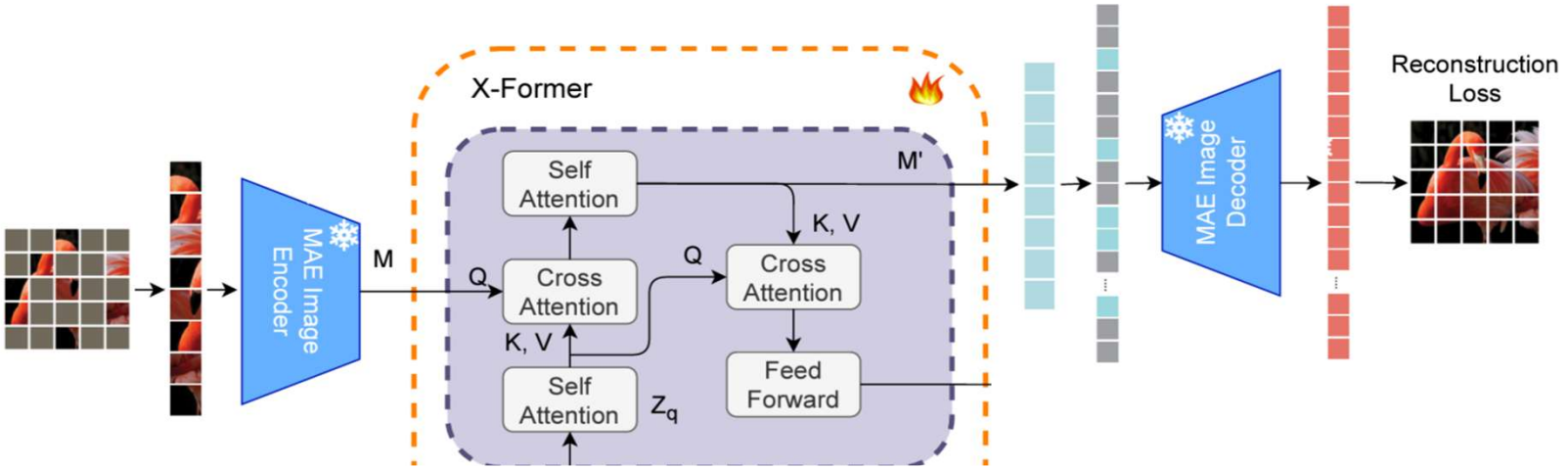
Contrastive Learning



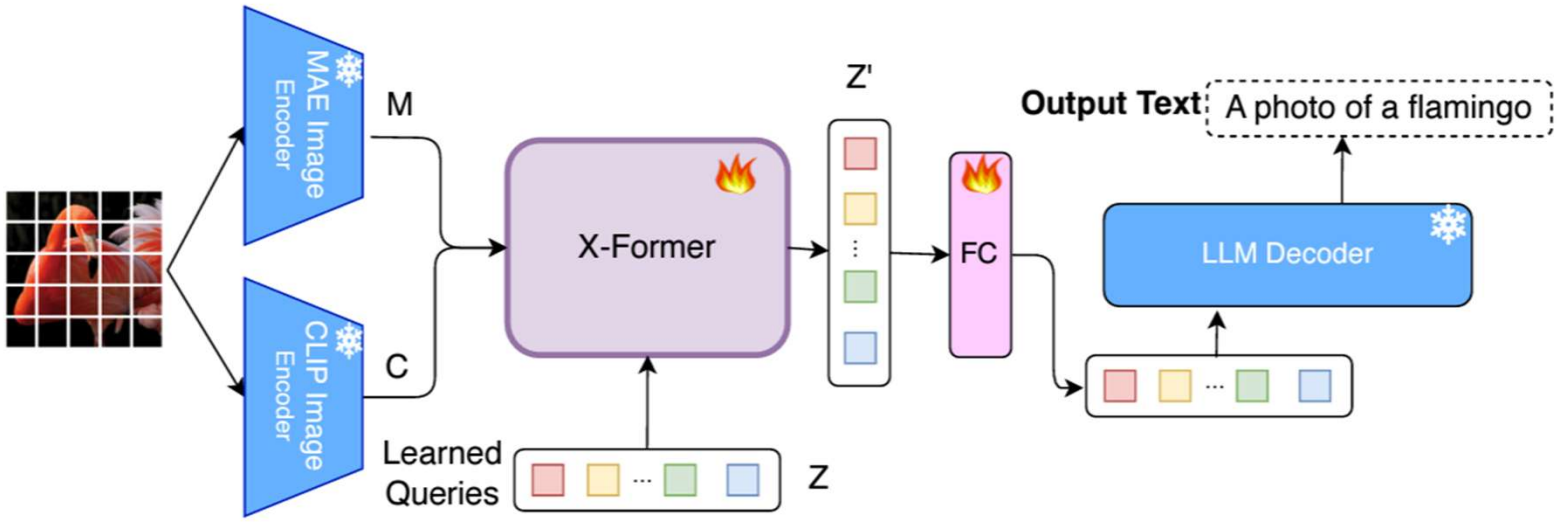
Self-Supervised Learning: MAE (Masked Auto-Encoder)



Stage-I


MAE Image Encoder
M
X-Former
Self Attention
M'
K, V
Q
Cross Attention
Q
Cross Attention
K, V
Self Attention
Z_q
Feed Forward
MAE Image Decoder
Reconstruction Loss

Stage-2


MAE Image Encoder
CLIP Image Encoder
M
C
X-Former
Learned Queries
Z
Z'
FC
Output Text
A photo of a flamingo
LLM Decoder

UCF Center for Research in Computer Vision



Q: How many donuts are there in the image?

BLIP-2: 4 **Ours: 6**



Q: How many airplanes are there in the image?

BLIP-2: 6 **Ours: 4**



Q: How many cups are there in the image?

BLIP-2: 3 **Ours: 2**



Q: How many dogs are there in the image?

BLIP-2: 2 **Ours: 1**



Q: Does fire hydrant exist in the image?

BLIP-2: Yes **Ours: No**



Q: Does potted plant exist in the image?

BLIP-2: No **Ours: Yes**



Q: Does baseball bat exist in the image?

BLIP-2: Yes **Ours: No**



Q: Does umbrella exist in the image?

BLIP-2: No **Ours: Yes**

Table 3: Zero-shot Fine-Grained Visual Perception evaluation of MLLMs on Object Counting (OC) & Multi-class Identification (MCI) tasks. For fair comparison, we compare with models trained only on image-text data. *evaluated using official checkpoint.

		OC		MCI	
Method	Data	COCO	VCR	COCO	VCR
BLIP-2* [23]	129M	34.3	18.9	69.44	74.16
BLIP-2 [23]	14M	25.88	21.12	61.5	65.3
X-Former (Ours)	14M	**39.64**	**27.24**	**69.44**	**69.28**

Method	Data	GQA	OKVQA
FewVLM [19]	9.1M	29.3	16.5
Frozen [35]		-	5.9
VLKD [7]	3.7M	-	13.3
Flamingo3B [2]	>2B	-	41.2
Flamingo9B [2]	>2B	-	44.7
Flamingo80B [2]	>2B	-	50.6
BLIP-2 $OPT_{6.7B}$* [23]	129M	34.2	35.3
BLIP-2 $OPT_{2.7B}$ [23]	14M	33.6	24.2
X-Former (Ours) $\mathbf{OPT_{2.7B}}$	14M	**34.1**	**27.7**
BLIP-2 $OPT_{6.7B}$ [23]	14M	33.1	31.5
X-Former (Ours) $\mathbf{OPT_{6.7B}}$	14M	**34.9**	**34.2**