UCF CENTER FOR RESEARCH IN COMPUTER VISION

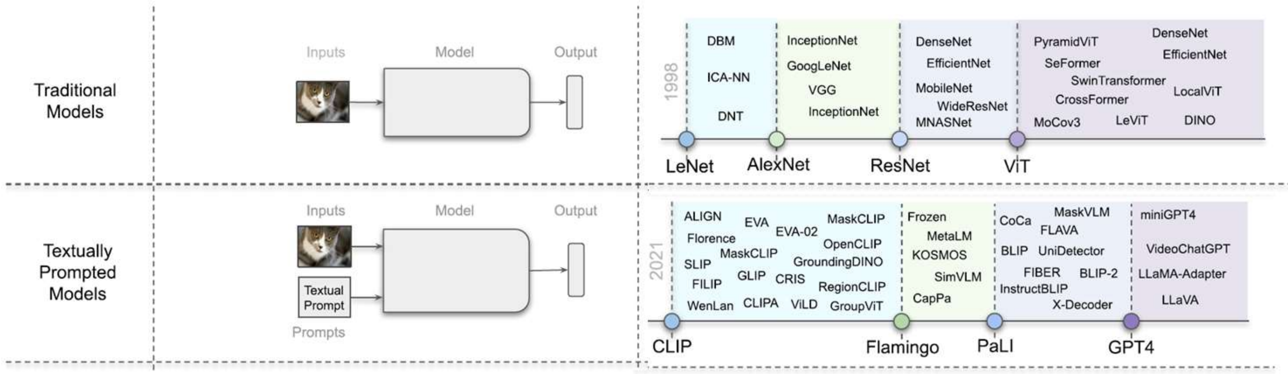
Visual-Language Models Introduction Part-I: CoCA, PALI, Lecture-4 CAP6412 Spring 2024

Mubarak Shah

shah@crcv.ucf.edu

UCF CENTER FOR RESEARCH IN COMPUTER VISION

Evolution of Computer Vision Models



Foundation Models

Source: Nvidia Blog

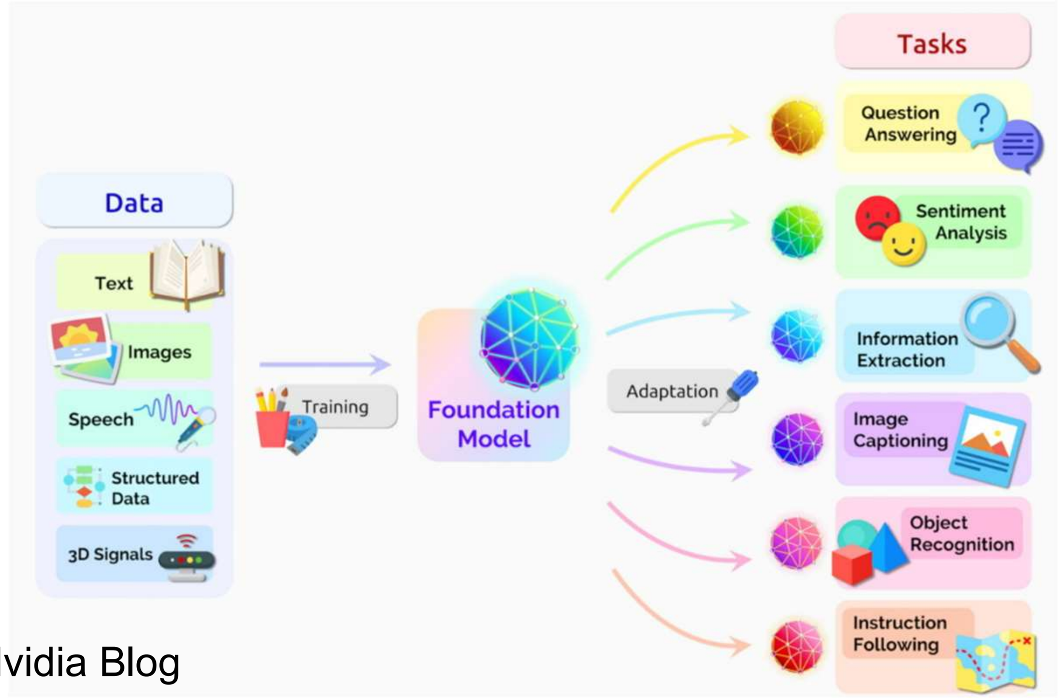
UCF CENTER FOR RESEARCH IN COMPUTER VISION

Foundation Models

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

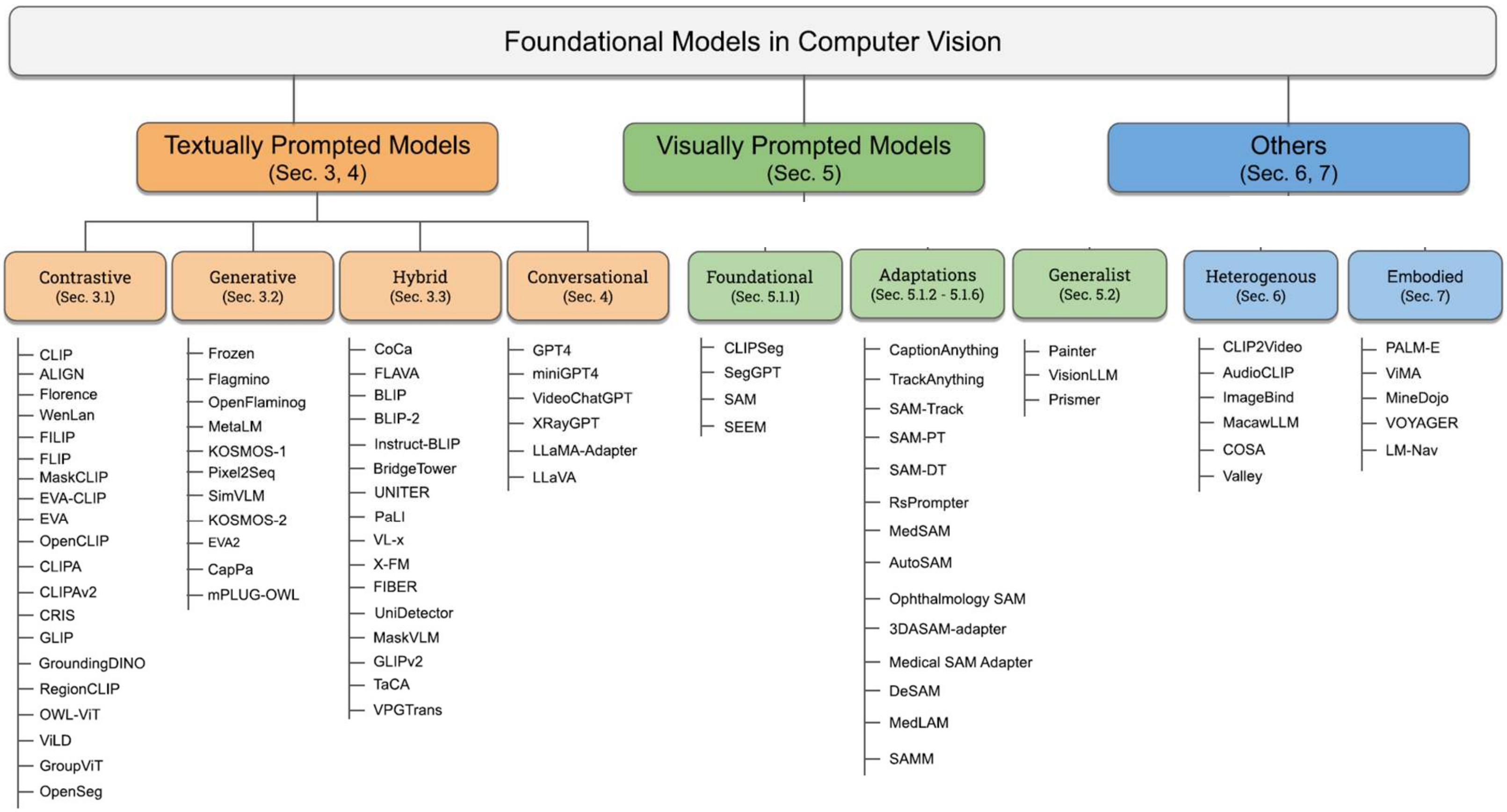
UCF Center for Research in Computer Vision

Foundation Models



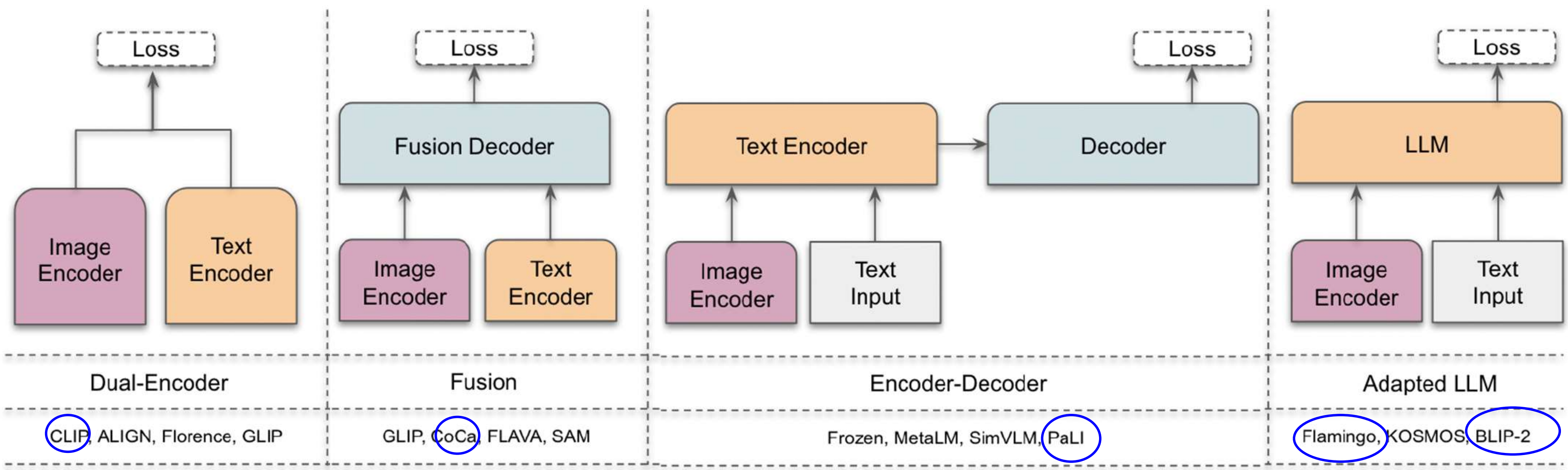
Source: Nvidia Blog

UCF Center for Research in Computer Vision



UCF CENTER FOR RESEARCH IN COMPUTER VISION

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

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*.

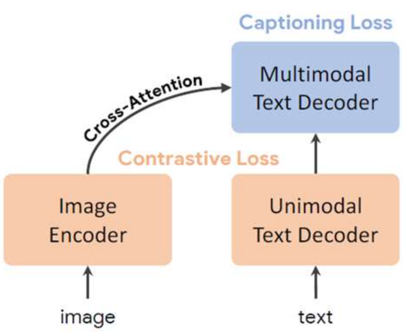
1 Introduction

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

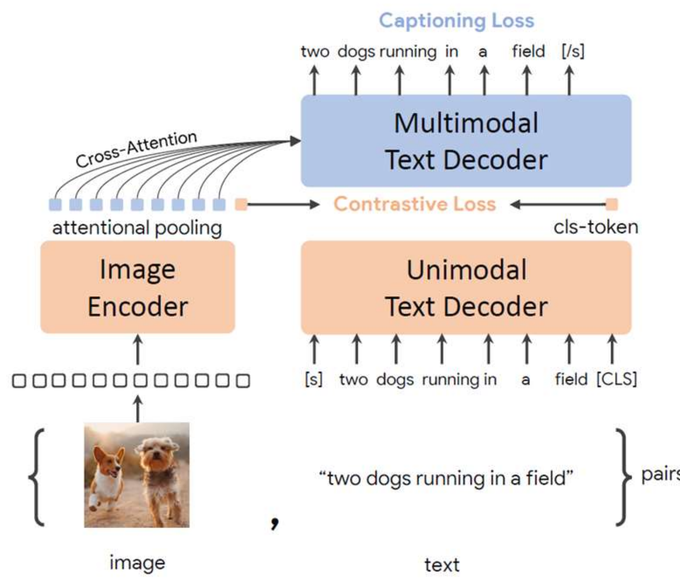
https://openreview.net/pdf?id=Ee277P3AYC

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



CoCa

Pretraining



UCF CENTER FOR RESEARCH IN COMPUTER VISION

Co-CA Losses

Dual-Encoder Contrastive Learning

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

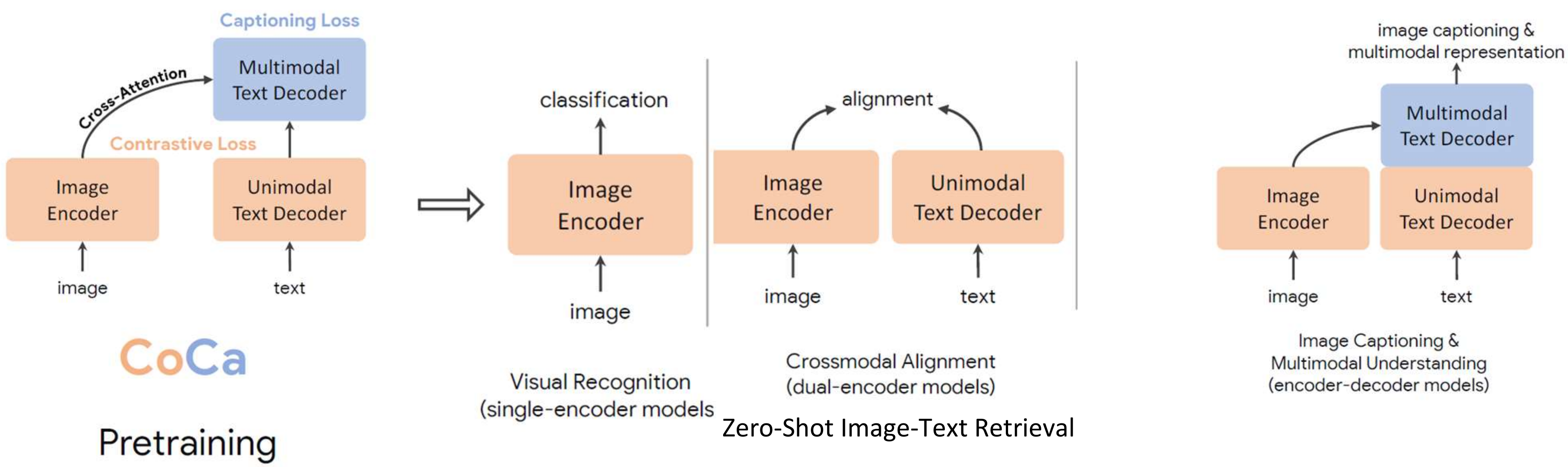
Encoder-Decoder Captioning

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

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

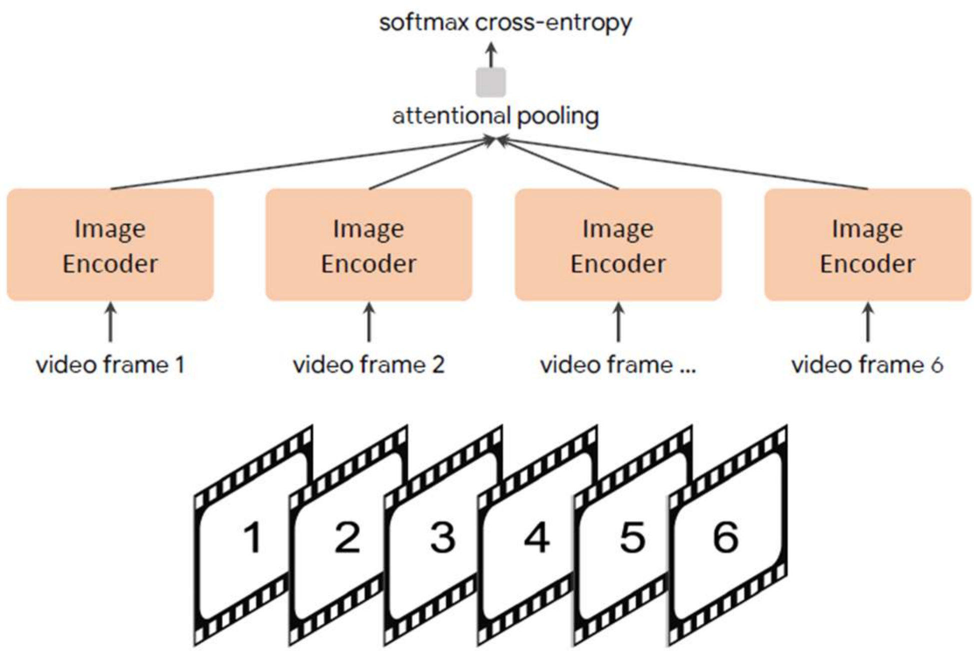
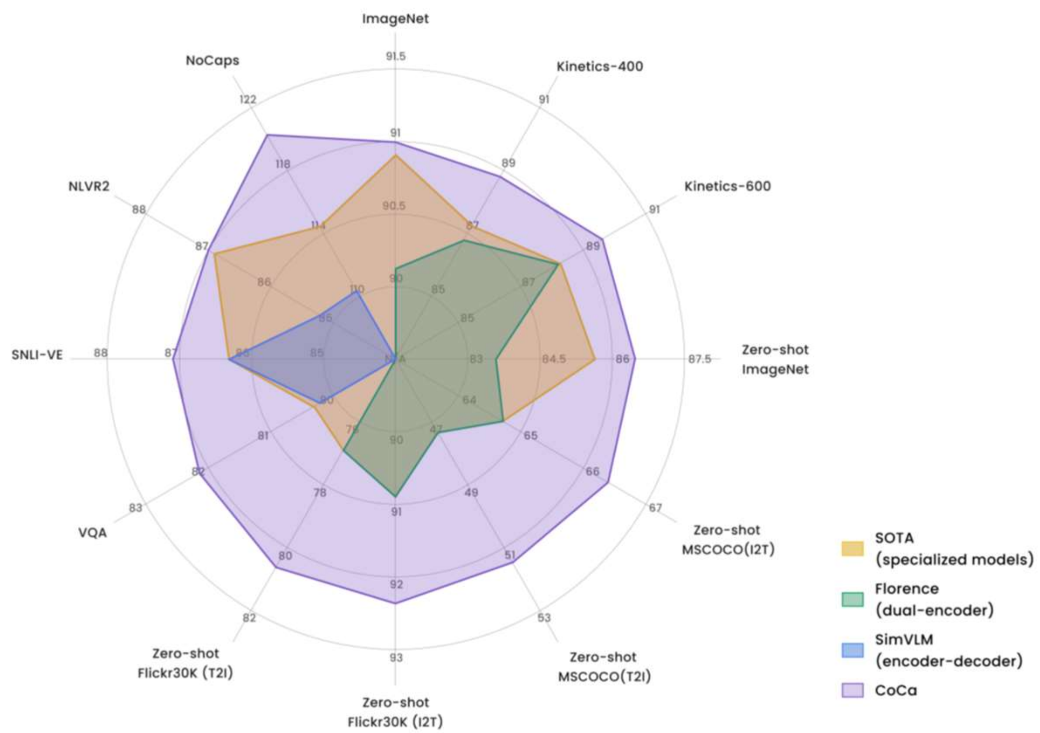


Figure 3: CoCa for video recognition.

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

+

- 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.

=

- Entailment
- Neutral
- Contradiction

Premise

Hypothesis

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.

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).

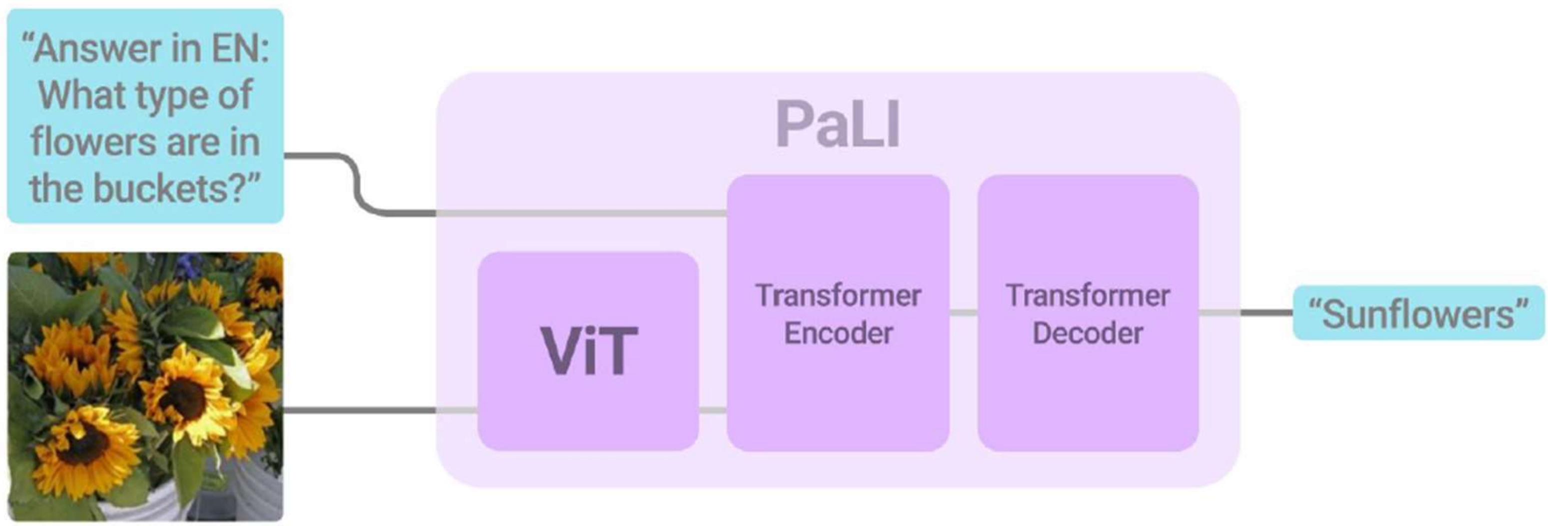
We introduce PaLI, a model that performs image-only, language-only, and image+language tasks across many languages, using a single "image-and-text to text" interface. A key characteristic of PaLI is a more balanced parameter share between the language and vision components, with more capacity to the 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

UCF Center for Research in Computer Vision

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