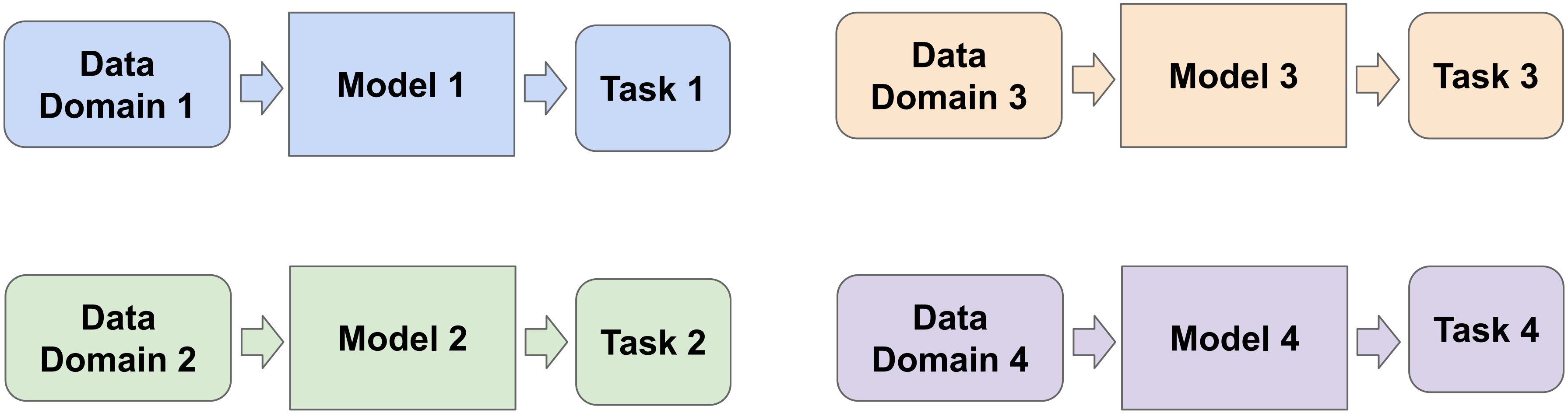
Lecture 16:
Multi-Modal Foundation Models

Administrative

- Milestone due today
- Quiz 4 Friday

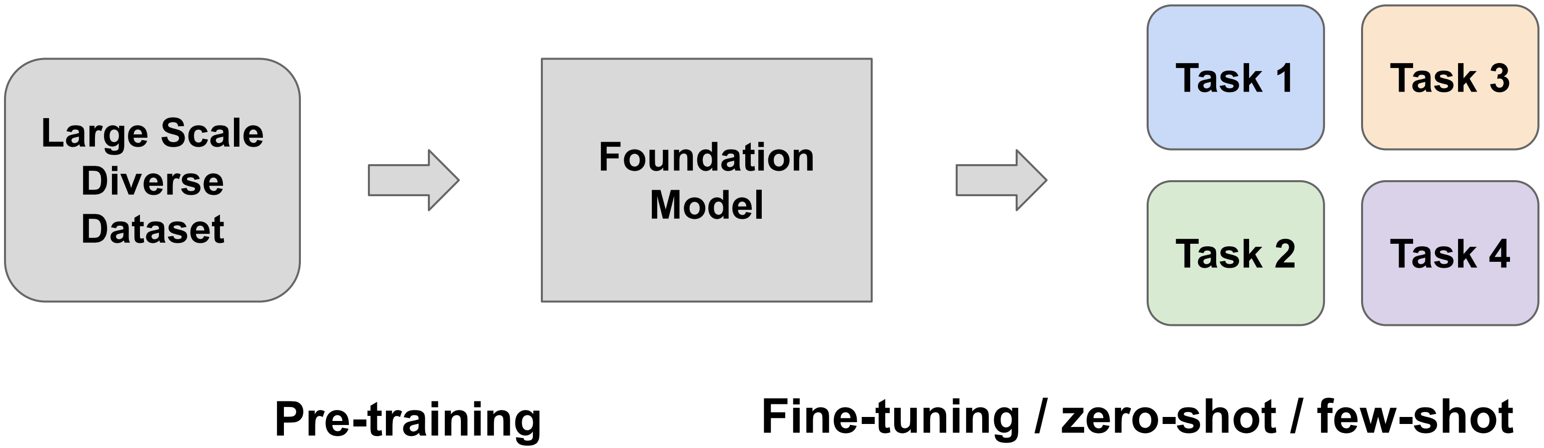
Foundation Models

Old: train specialized models for each task



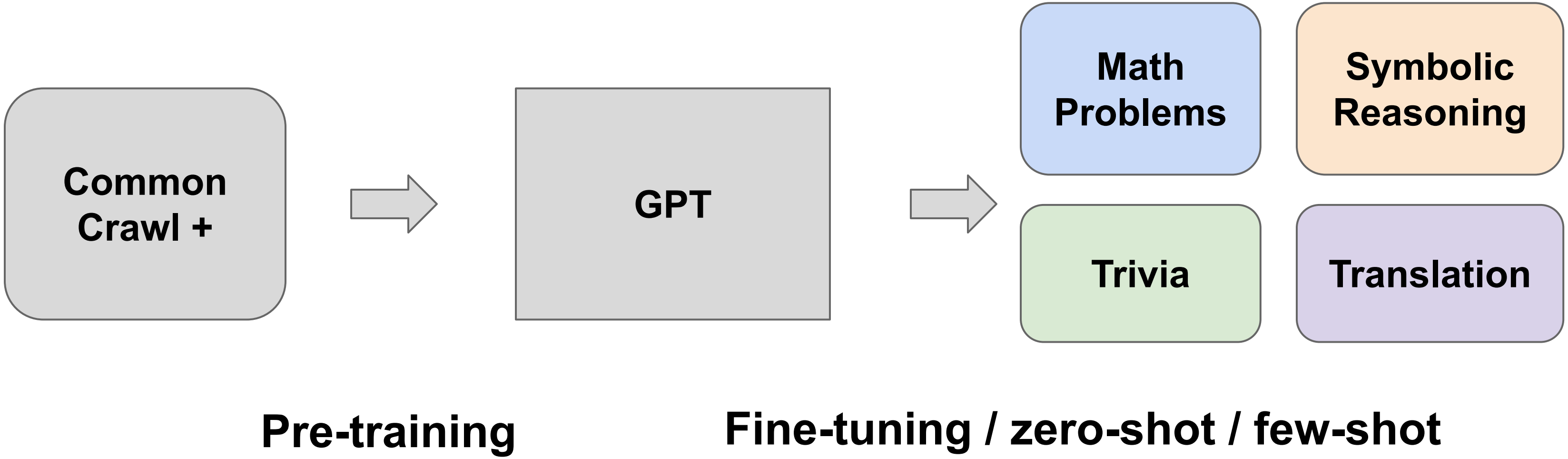
Foundation Models

New: pre-train one model that acts as the foundation for many different tasks



Foundation Models

Language



Foundation Models

Language	Classification	LM + Vision	And More!	Chaining
ELMo	CLIP	Flamingo	Segment Anything	LMs + CLIP
BERT	CoCa	GPT-4V	Whisper	Visual Programming
GPT		Gemini	Dalle	
T5			Stable Diffusion	
			Imagen	

Foundation Models

Always see with foundation models:

- general /robust to many different tasks

Often see with foundation models:

- Large # params
- Large amount of data
- Self-supervised pre-training objective

Foundation Models

Language	Classification	LM + Vision	And More!	Chaining
ELMo	CLIP	Flamingo	Segment Anything	LMs + CLIP
BERT	CoCa	GPT-4V	Whisper	Visual Programming
GPT		Gemini	Dalle	
T5			Stable Diffusion	
			Imagen	

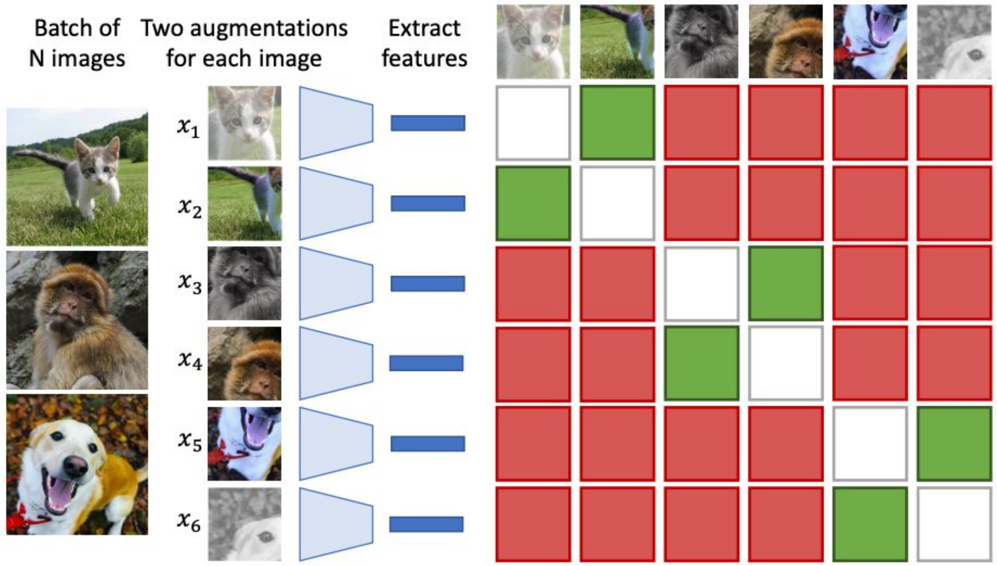
Foundation Models

Language	Classification	LM + Vision	And More!	Chaining
ELMo	**CLIP**	**Flamingo**	**Segment Anything**	**LMs + CLIP**
BERT	**CoCa**	GPT-4V	Whisper	**Visual Programming**
GPT		Gemini	Dalle	
T5			Stable Diffusion	
			Imagen	

Foundation Models

Language	Classification	LM + Vision	And More!	Chaining
ELMo	**CLIP**	**Flamingo**	**Segment Anything**	**LMs + CLIP**
BERT	**CoCa**	GPT-4V	Whisper	**Visual Programming**
GPT		Gemini	Dalle	
T5			Stable Diffusion	
			Imagen	

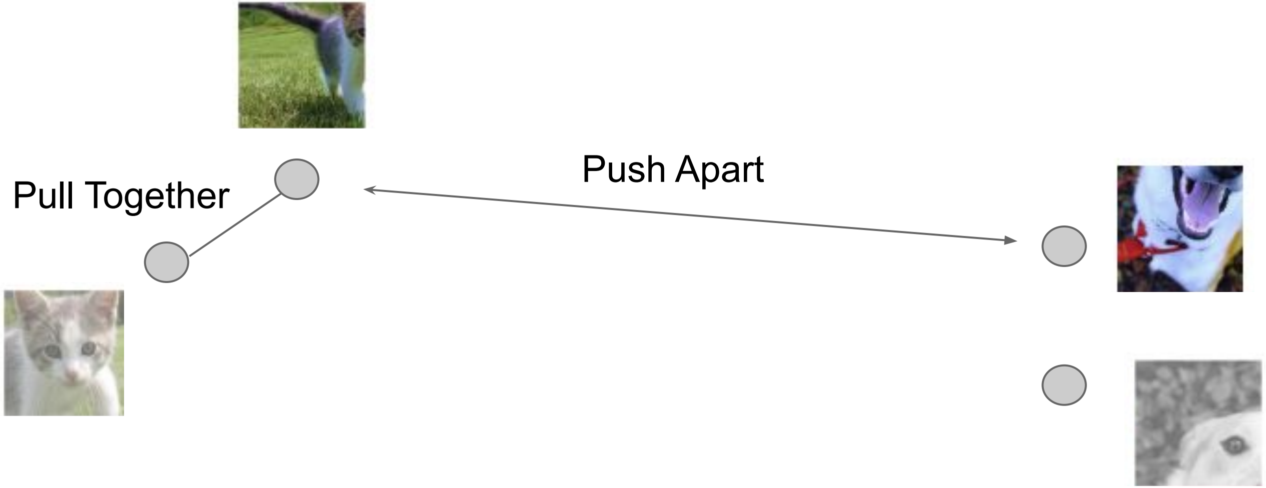
Previously…



Use Self Supervised learning to learn good image features

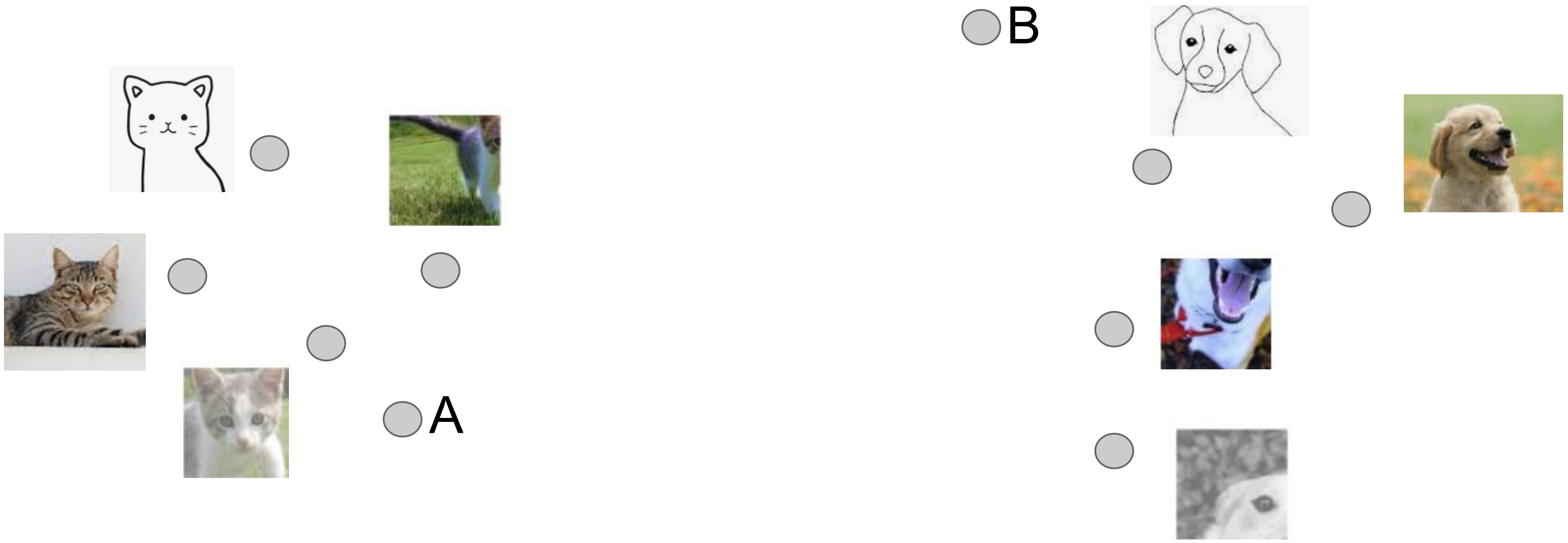
Can train small classifiers on top of these features using supervised learning

Learning Concepts without Labels



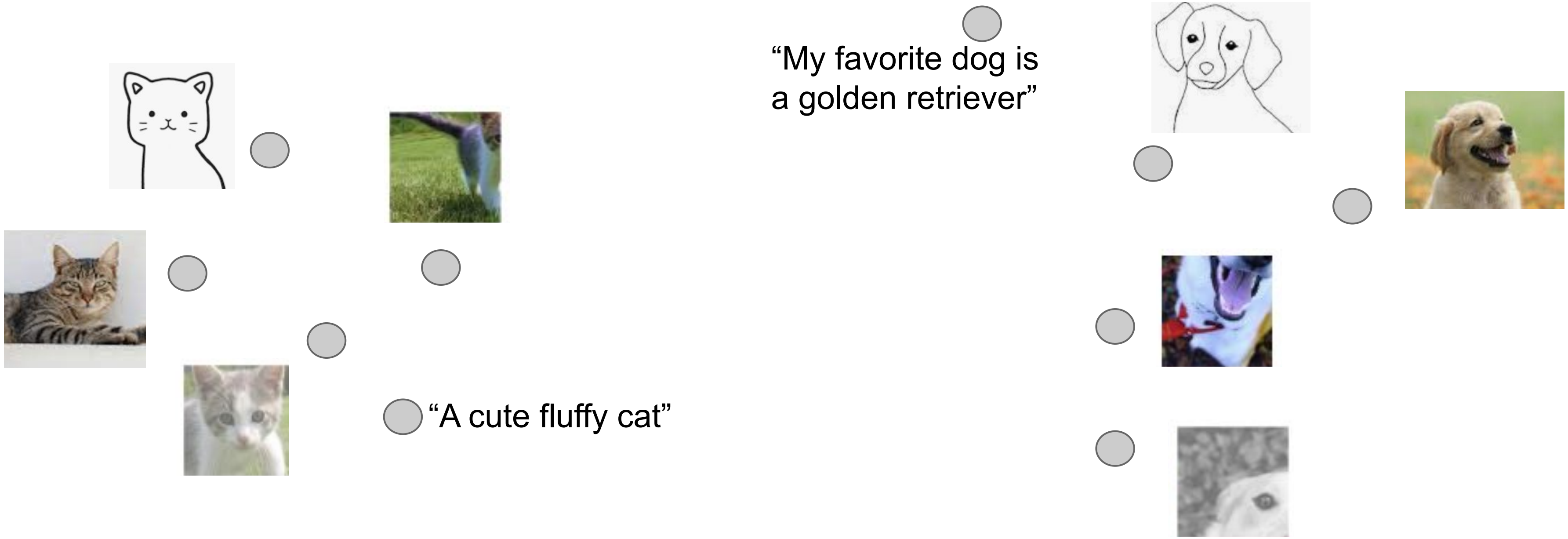
Learning Concepts without Labels

Learning Concepts without Labels

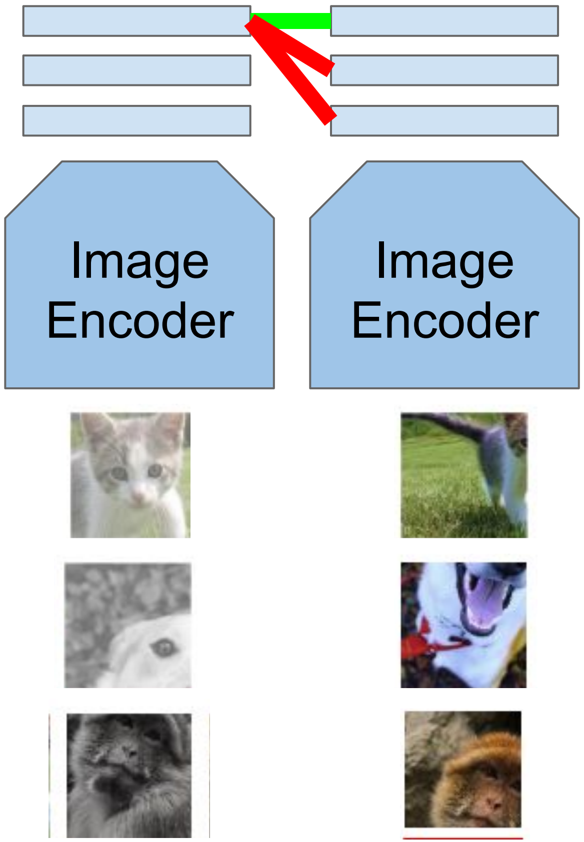


1. "A cute fluffy cat"
2. "My favorite dog is a golden retriever"

Learning Concepts without Labels



SimClr



SimClr

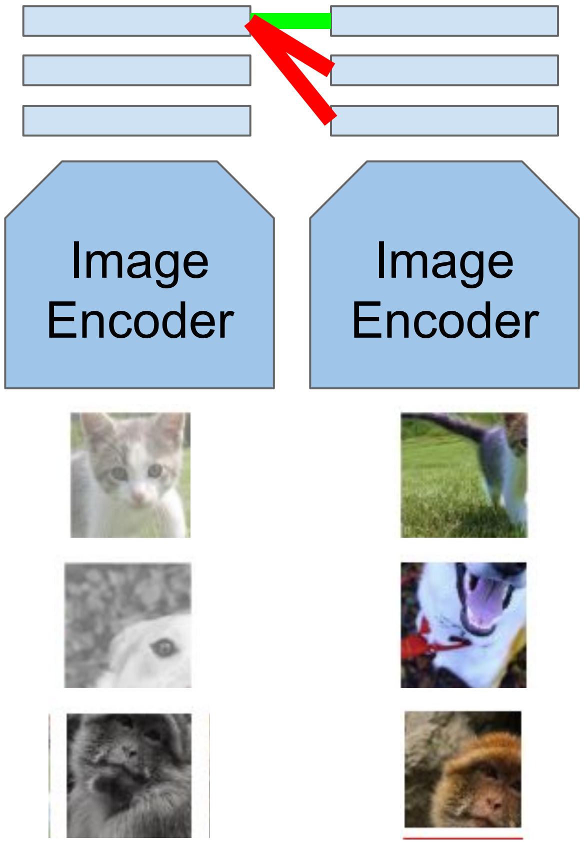


Image Encoder
Image Encoder

CLIP

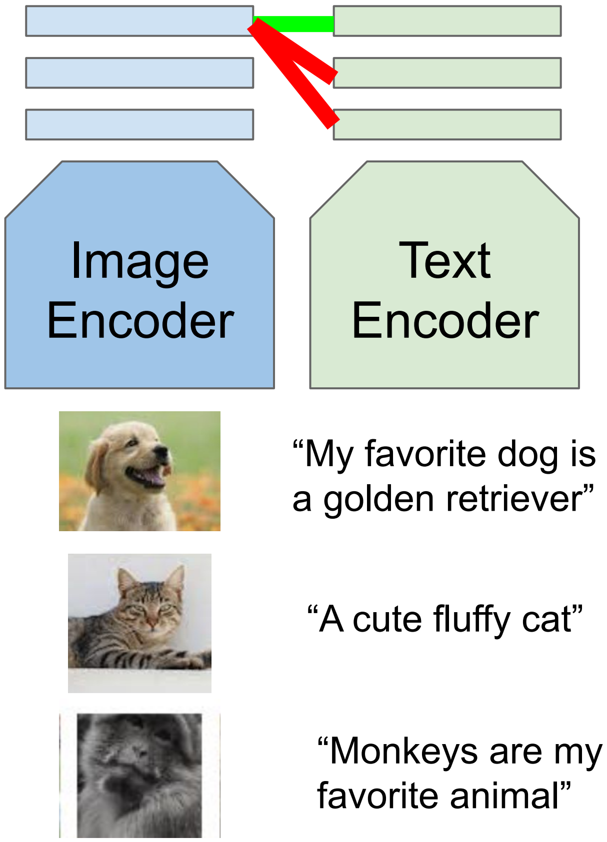
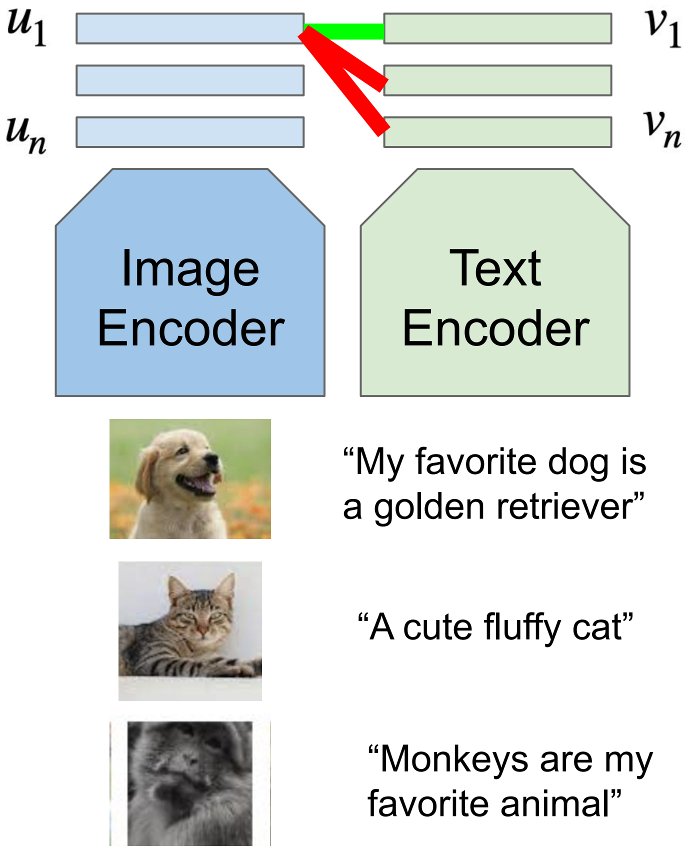


Image Encoder
Text Encoder
“My favorite dog is a golden retriever”
“A cute fluffy cat”
“Monkeys are my favorite animal”

CLIP Training Objective

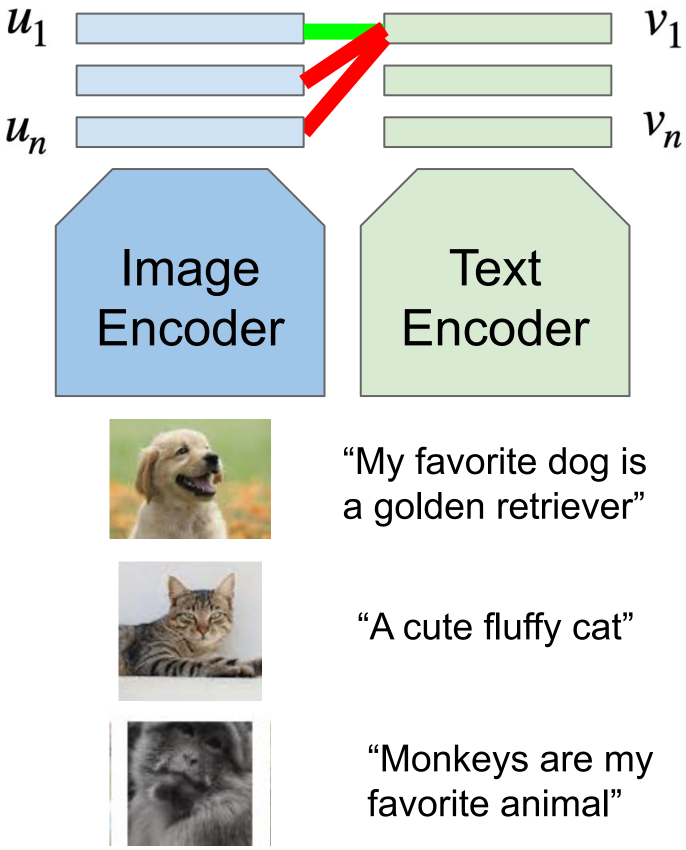
$$\sum_{i=1}^{n} -\log\left(\frac{e^{\langle u_i, v_i\rangle}}{\sum_{j=1}^{n} e^{\langle u_i, v_j\rangle}}\right)$$



CLIP Training Objective

$$\sum_{i=1}^{n} -\log\left(\frac{e^{\langle u_i, v_i\rangle}}{\sum_{j=1}^{n} e^{\langle u_i, v_j\rangle}}\right)$$

$$+\sum_{i=1}^{n} -\log\left(\frac{e^{\langle u_i, v_i\rangle}}{\sum_{j=1}^{n} e^{\langle u_j, v_i\rangle}}\right)$$



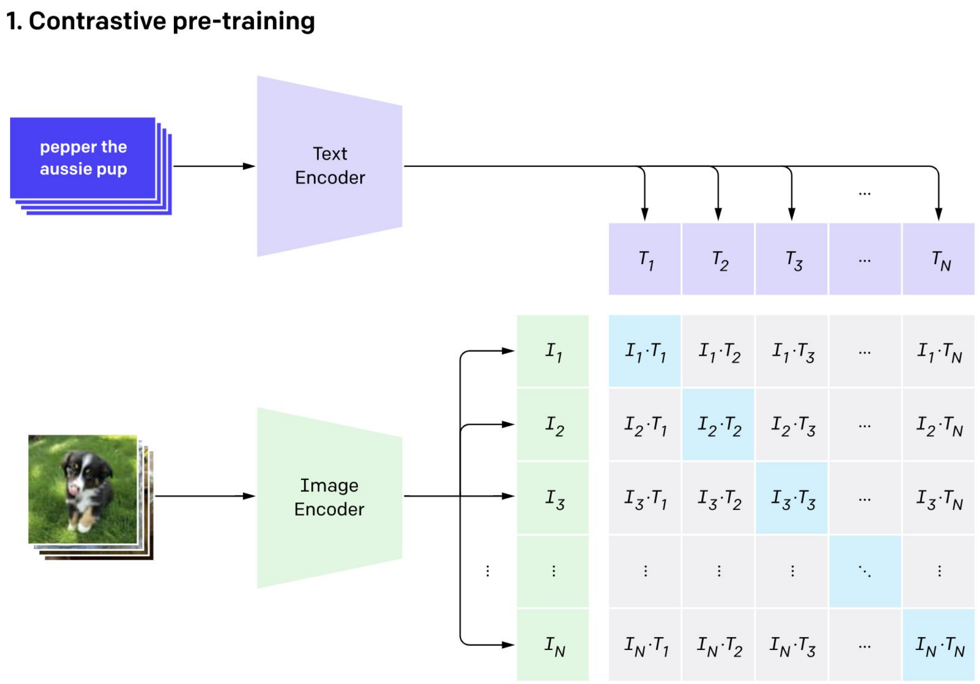
CLIP Training Data

Mount Rainier's northwestern slope viewed aerially just before sunset on September 6, 2020

Image-Text pairs scraped from the internet

https://en.wikipedia.org/wiki/Mount_Rainier

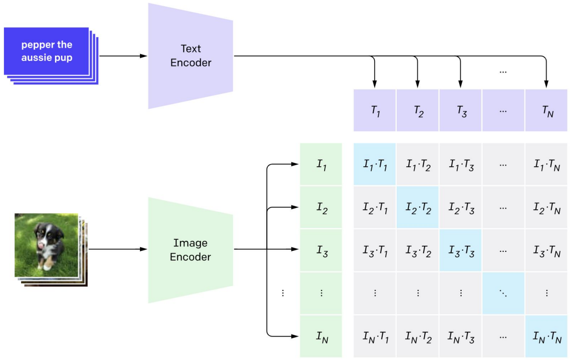
CLIP Training Objective



At the end of training, you have a model that will give you a similarity score between an image and a text

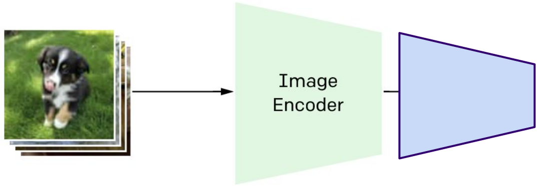
Using pre-trained models out of the box

Step 1: <u>Pretrain</u> a network on a <u>pretext task</u> that doesn't require supervision



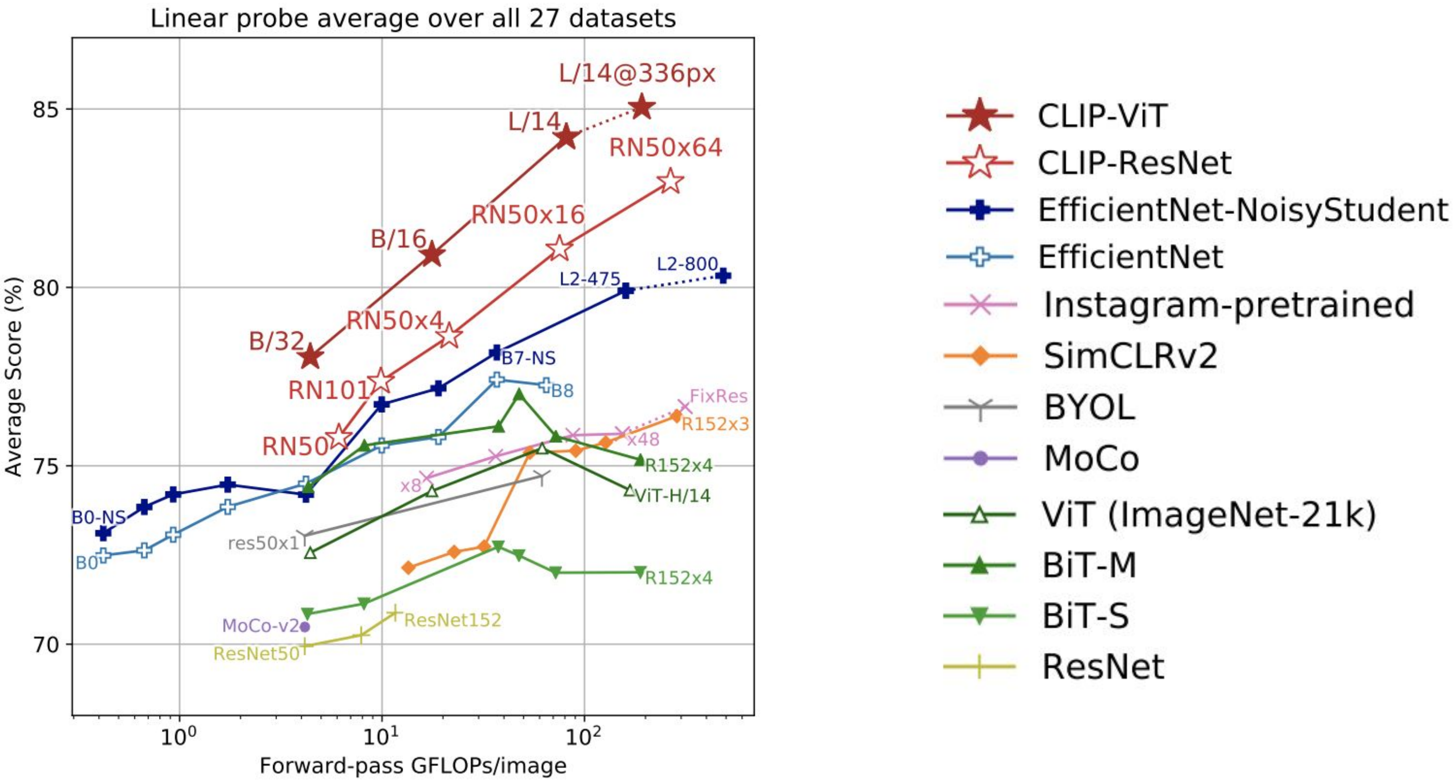
Pre-training tasks:
Contrastive Objective

Step 2: Transfer encoder to <u>downstream tasks</u> via linear classifiers, KNN, finetuning



Downstream tasks:
Image classification, object detection, semantic segmentation

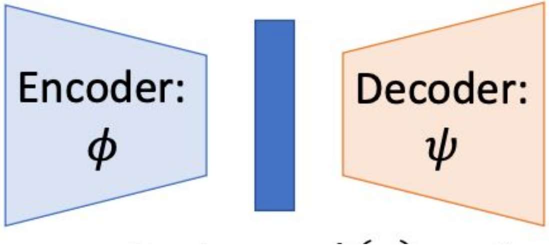
CLIP features w/ linear probe across datasets



Using pre-trained models out of the box

Step 1: <u>Pretrain</u> a network on a <u>pretext task</u> that doesn't require supervision

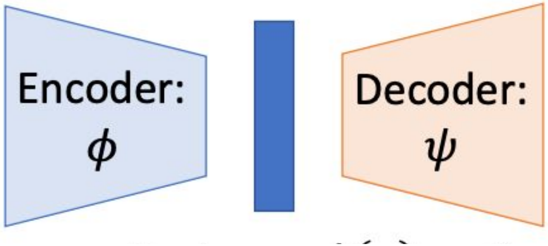
"I love ____"



"cake"

Step 2: Use the model out of the box in a creative way!

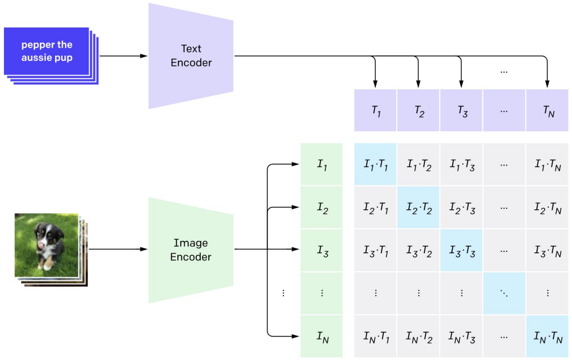
"The movie review 'I hated the movie' is ____"



"negative"

Using pre-trained models out of the box

Step 1: <u>Pretrain</u> a network on a <u>pretext task</u> that doesn't require supervision



Pre-training tasks:
Contrastive Objective

Step 2: Use the model out of the box in a creative way!

Out of the box classification (No fine-tuning)

Creating a classifier

word

Image Encoder

Text Encoder

0.27

Image vector

text vector

matching score

Radford et al “Learning Transferable Visual Models From Natural Language Supervision”

Creating a classifier

plane

dog

bird

Image Encoder

Text Encoder

Image vector

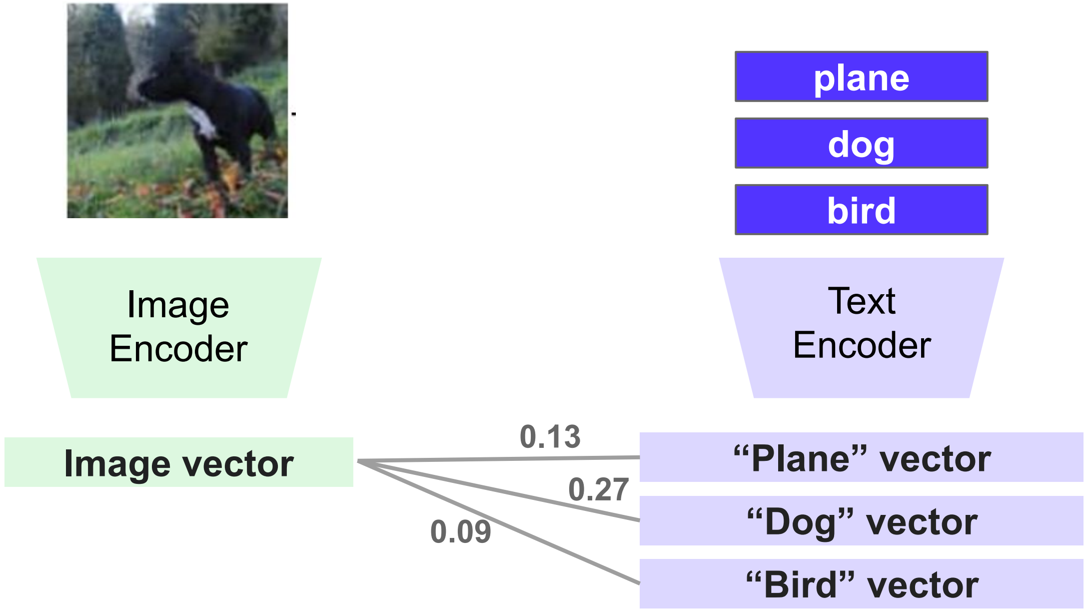
“Plane” vector

“Dog” vector

“Bird” vector

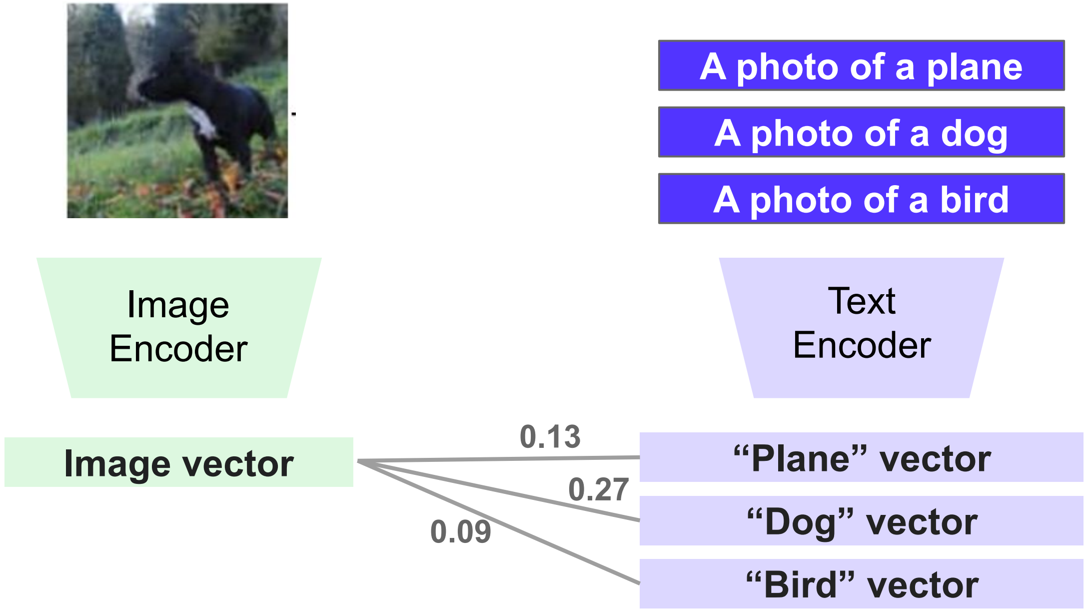
Radford et al “Learning Transferable Visual Models From Natural Language Supervision”

Creating a classifier



Radford et al “Learning Transferable Visual Models From Natural Language Supervision”

Creating a classifier



+1.3% on ImageNet

Radford et al "Learning Transferable Visual Models From Natural Language Supervision"

Creating a classifier

A photo of a plane | A drawing of a plane …
A photo of a dog | A drawing of a dog …
A photo of a bird | A drawing of a bird …

Image Encoder

Text Encoder

Image vector

"Plane" vector 1 | "Plane" vector 2
"Dog" vector 1 | "Dog" vector 2
"Bird" vector 1 | "Bird" vector 2

Radford et al "Learning Transferable Visual Models From Natural Language Supervision"

Creating a classifier

A photo of a plane | A drawing of a plane | …
A photo of a dog | A drawing of a dog | …
A photo of a bird | A drawing of a bird | …

Image Encoder

Text Encoder

+5% on ImageNet

Image vector

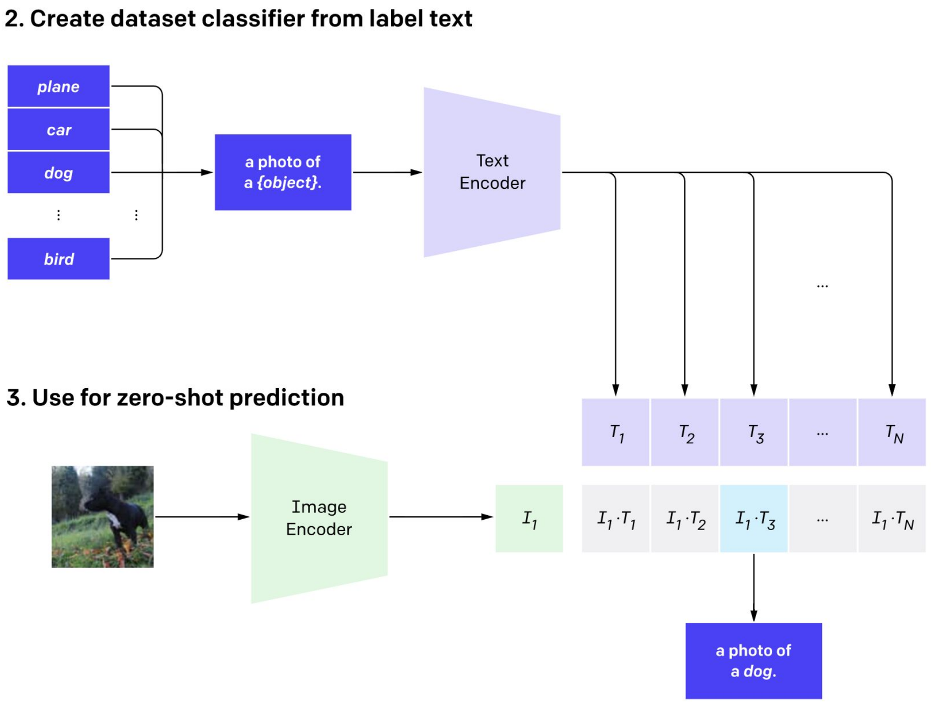
0.13 — **Mean "Plane" vector**

0.27 — **Mean "Dog" vector**

0.09 — **Mean "Bird" vector**

Radford et al "Learning Transferable Visual Models From Natural Language Supervision"

Creating a classifier



Radford et al “Learning Transferable Visual Models From Natural Language Supervision”

Imagenet Accuracy

DATASET	IMAGENET RESNET101	CLIP VIT-L
 ImageNet	76.2%	76.2%

Matches the accuracy of of ResNet 101 that has been trained on human labeled data with no human labels at all!

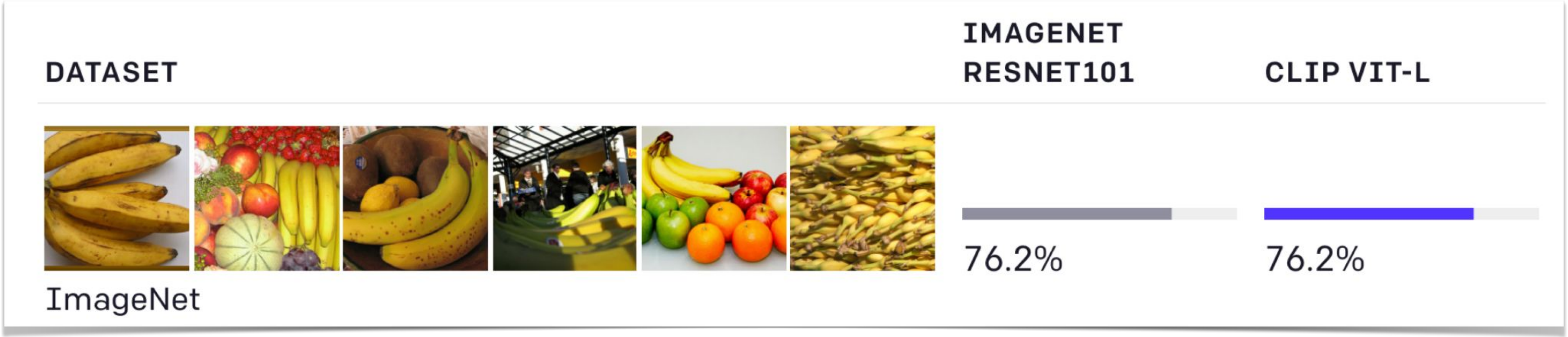
Radford et al "Learning Transferable Visual Models From Natural Language Supervision"

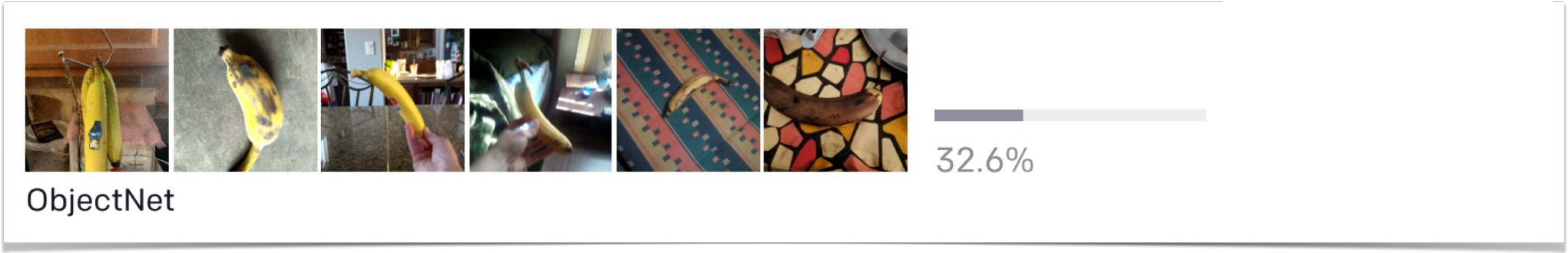
Imagenet Accuracy

DATASET	IMAGENET RESNET101	CLIP VIT-L
 ImageNet	76.2%	76.2%
 ObjectNet		

Radford et al "Learning Transferable Visual Models From Natural Language Supervision"

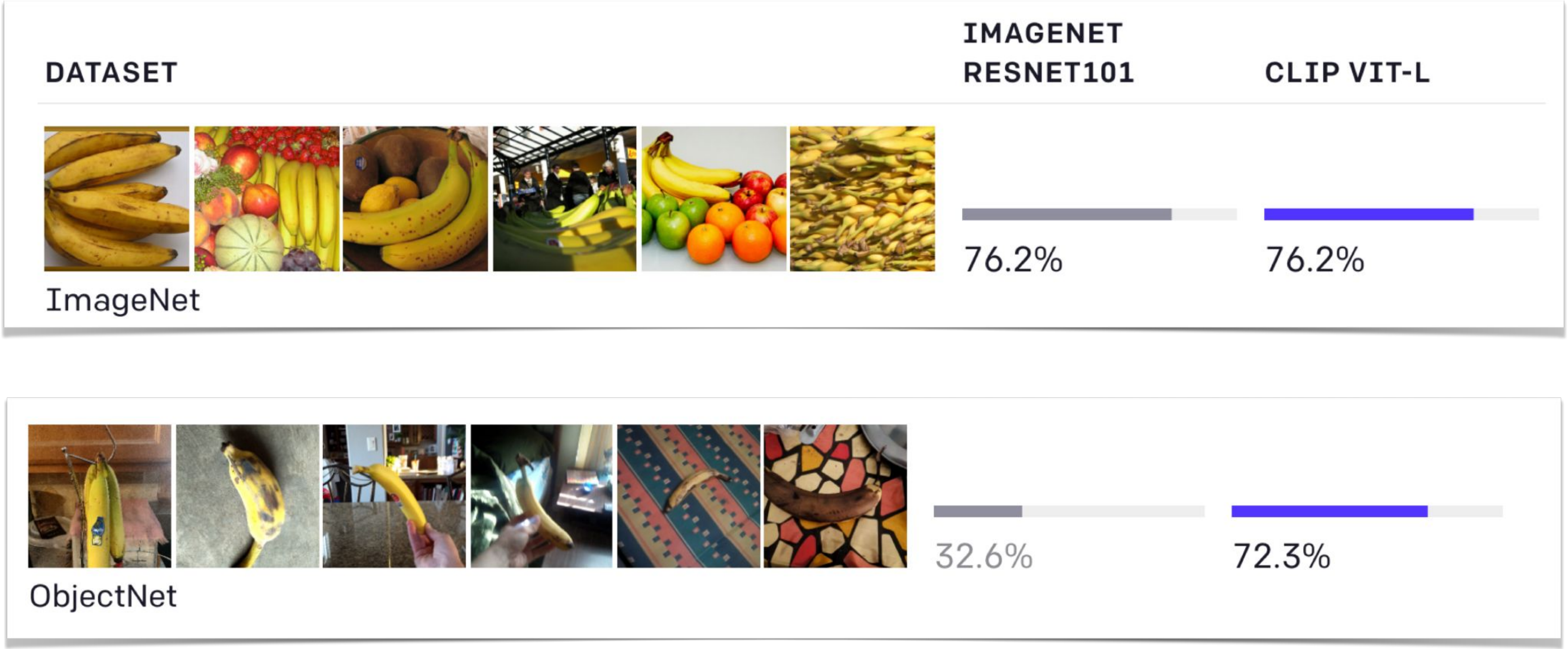
Imagenet Accuracy





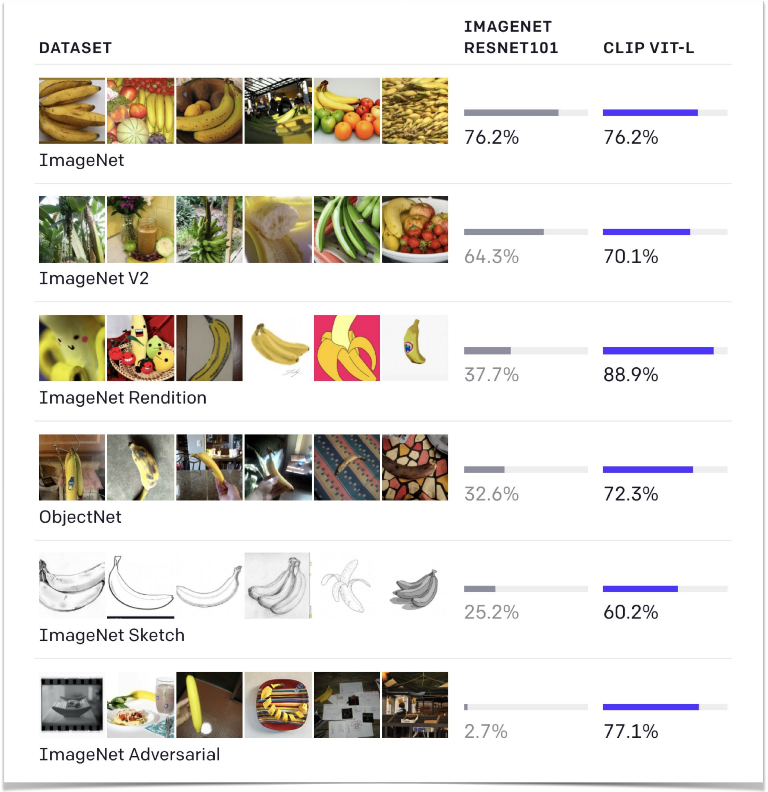
Radford et al “Learning Transferable Visual Models From Natural Language Supervision”

Imagenet Accuracy



Radford et al “Learning Transferable Visual Models From Natural Language Supervision”

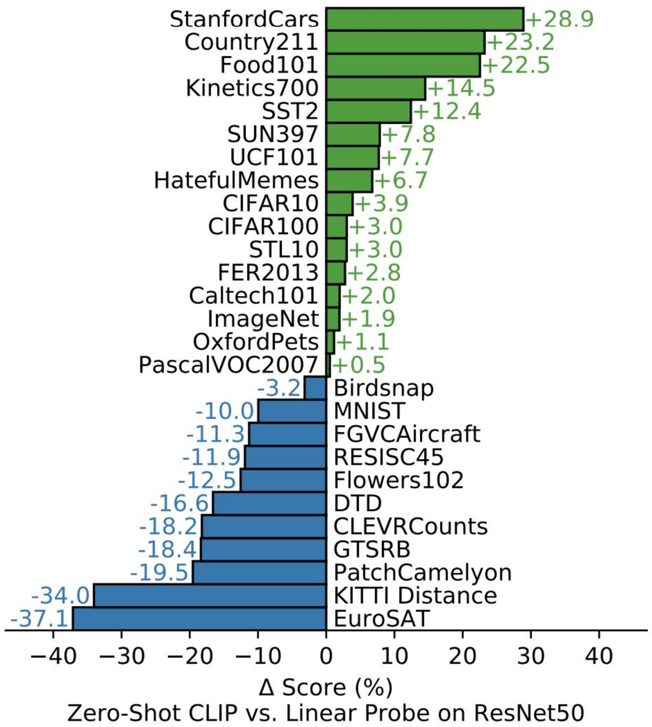
Imagenet Accuracy



DATASET	IMAGENET RESNET101	CLIP VIT-L
ImageNet	76.2%	76.2%
ImageNet V2	64.3%	70.1%
ImageNet Rendition	37.7%	88.9%
ObjectNet	32.6%	72.3%
ImageNet Sketch	25.2%	60.2%
ImageNet Adversarial	2.7%	77.1%

Radford et al “Learning Transferable Visual Models From Natural Language Supervision”

Accuracy on other datasets



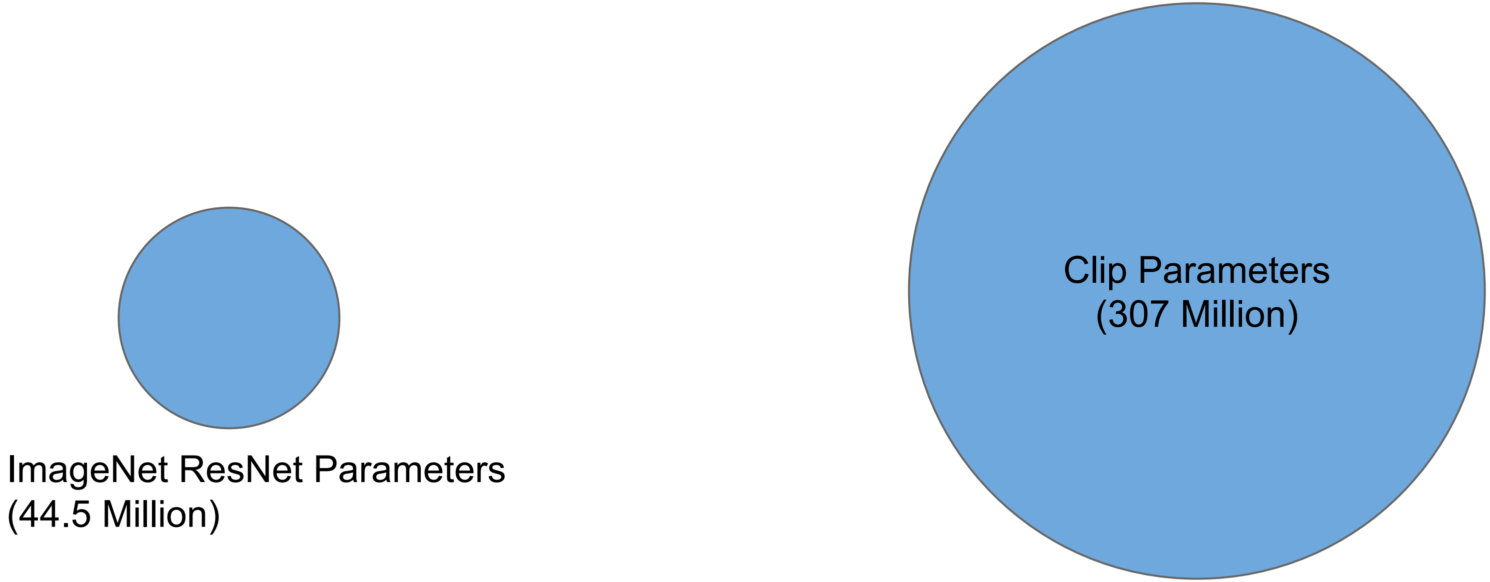
Radford et al “Learning Transferable Visual Models From Natural Language Supervision”

Key to high accuracy

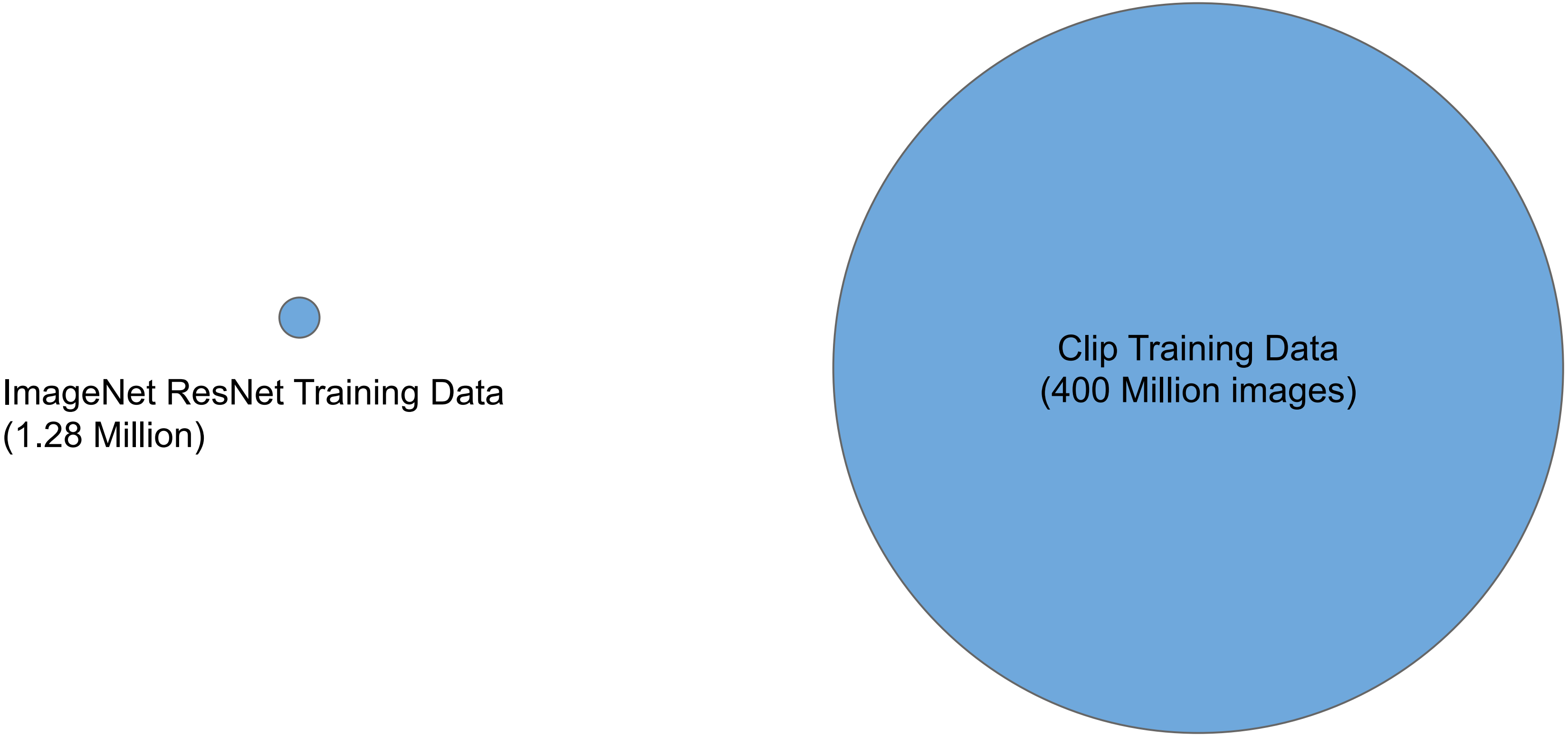
How can no labels beat labels??

Scale!

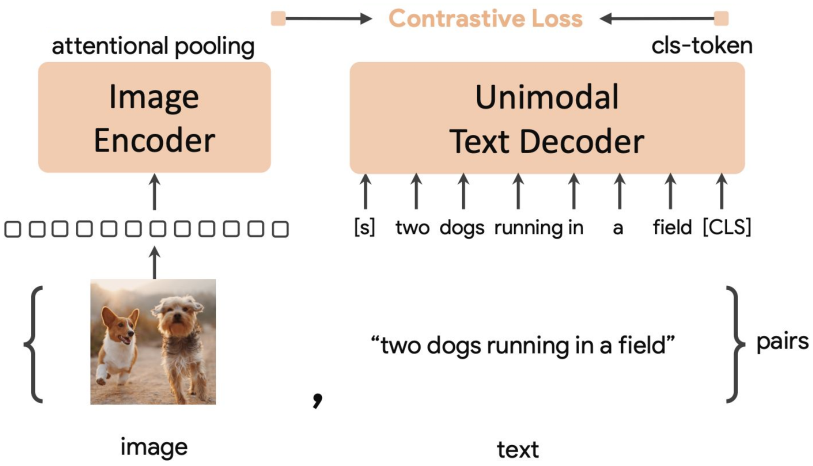
Model Scale


ImageNet ResNet Parameters
(44.5 Million)
Clip Parameters
(307 Million)

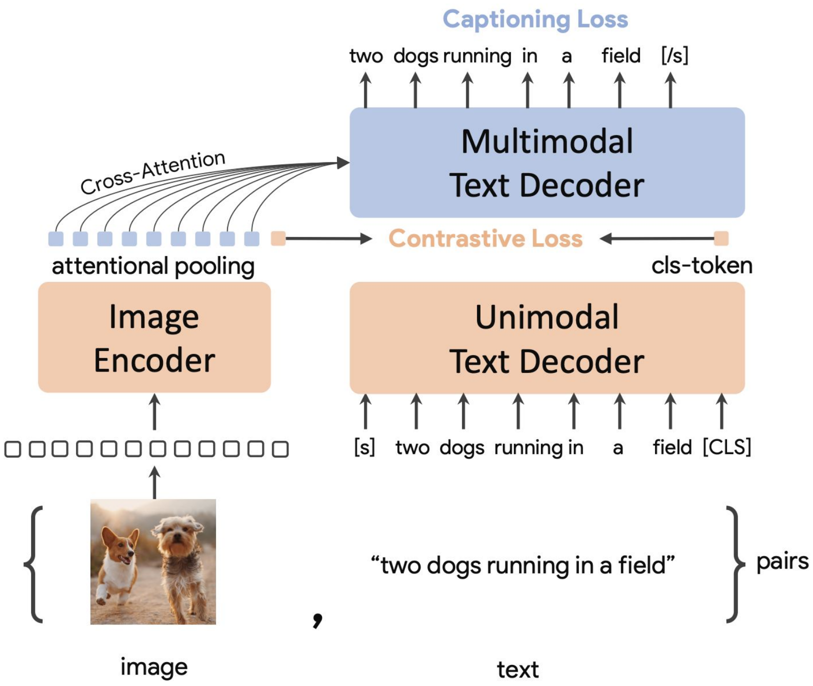
Data Scale


ImageNet ResNet Training Data
(1.28 Million)
Clip Training Data
(400 Million images)

CoCa: Contrastive Captioners are Image-Text Foundation Models



CoCa: Contrastive Captioners are Image-Text Foundation Models



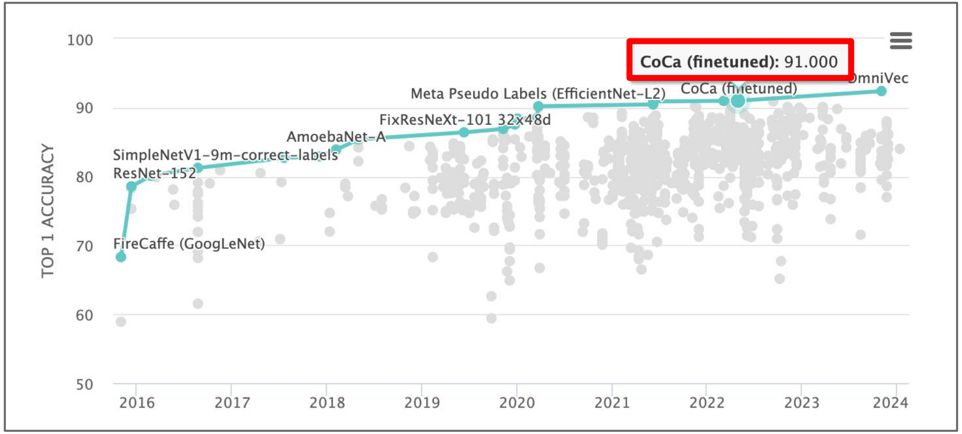
CoCa: Contrastive Captioners are Image-Text Foundation Models

Model	ImageNet	ImageNet-A	ImageNet-R	ImageNet-V2	ImageNet-Sketch	ObjectNet	Average
CLIP [12]	76.2	77.2	88.9	70.1	60.2	72.3	74.3
ALIGN [13]	76.4	75.8	92.2	70.1	64.8	72.2	74.5
FILIP [61]	78.3	-	-	-	-	-	-
Florence [14]	83.7	-	-	-	-	-	-
LiT [32]	84.5	79.4	93.9	78.7	-	81.1	-
BASIC [33]	85.7	85.6	95.7	80.6	76.1	78.9	83.7
CoCa-Base	82.6	76.4	93.2	76.5	71.7	71.6	78.7
CoCa-Large	84.8	85.7	95.6	79.6	75.7	78.6	83.3
CoCa	**86.3**	**90.2**	**96.5**	**80.7**	**77.6**	**82.7**	**85.7**

Table 4: Zero-shot image classification results on ImageNet [9], ImageNet-A [64], ImageNet-R [65], ImageNet-V2 [66], ImageNet-Sketch [67] and ObjectNet [68].

CoCa: Contrastive Captioners are Image-Text Foundation Models

Model	ImageNet
ALIGN [13]	88.6
Florence [14]	90.1
MetaPseudoLabels [51]	90.2
CoAtNet [10]	90.9
ViT-G [21]	90.5
+ Model Soups [52]	90.9
CoCa (frozen)	90.6
CoCa (finetuned)	**91.0**



Foundation Models

Language	Classification	LM + Vision	And More!	Chaining
ELMo	**CLIP**	**Flamingo**	**Segment Anything**	**LMs + CLIP**
BERT	**CoCa**	GPT-4V	Whisper	**Visual Programming**
GPT		Gemini	Dalle	
T5			Stable Diffusion	
			Imagen	

Flamingo

Motivation: CLIP is extremely general in its learned representation, but limited in its out-of-the box applications.

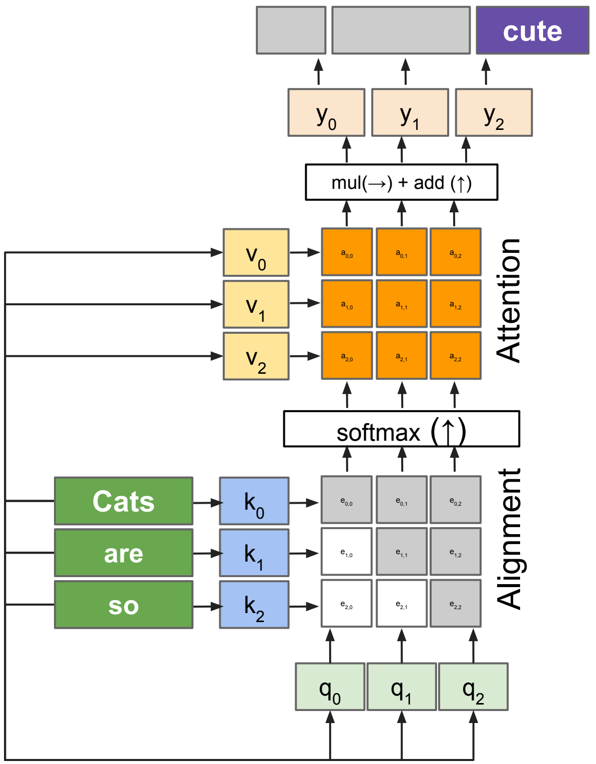
(only can output similarity scores between image and text)

Flamingo

Motivation: Language models which do next token prediction can be applied to a wide variety of tasks at inference (Math, sentiment analysis, symbolic reasoning)

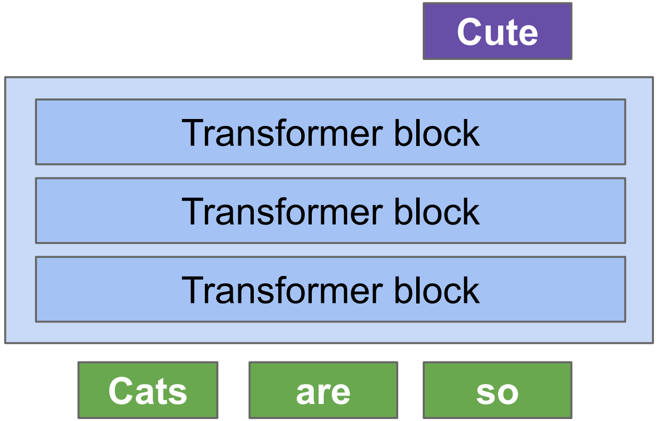
Can we build something like GPT but can accept images and text as input, and then output text?

Flamingo



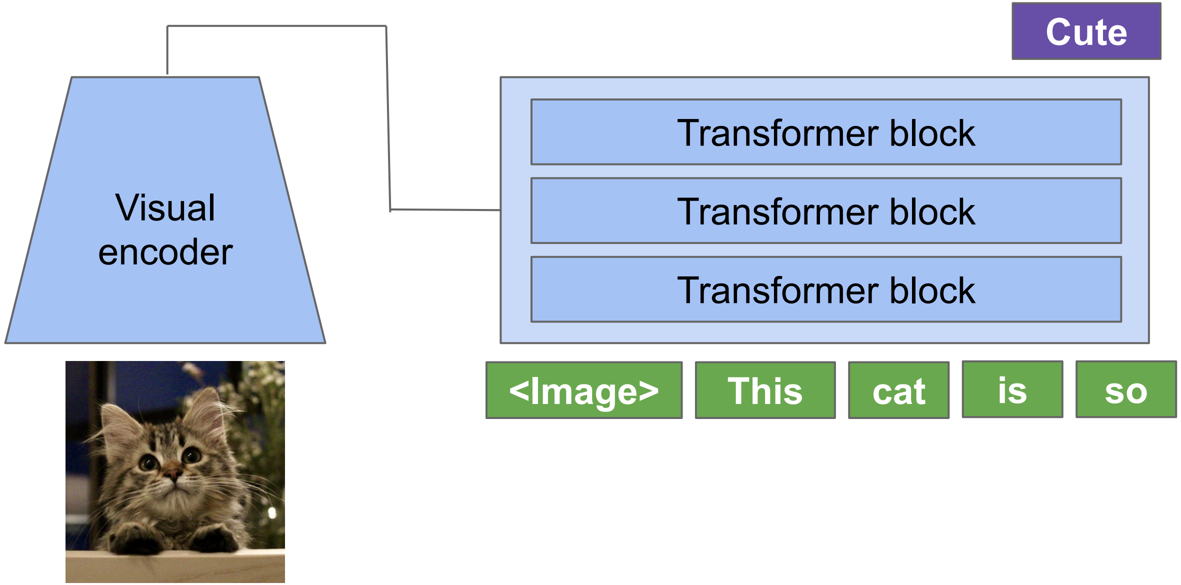
cute
y_0
y_1
y_2
mul(→) + add (↑)
v_0
v_1
v_2
Attention
softmax (↑)
Cats
are
so
k_0
k_1
k_2
Alignment
q_0
q_1
q_2

Flamingo

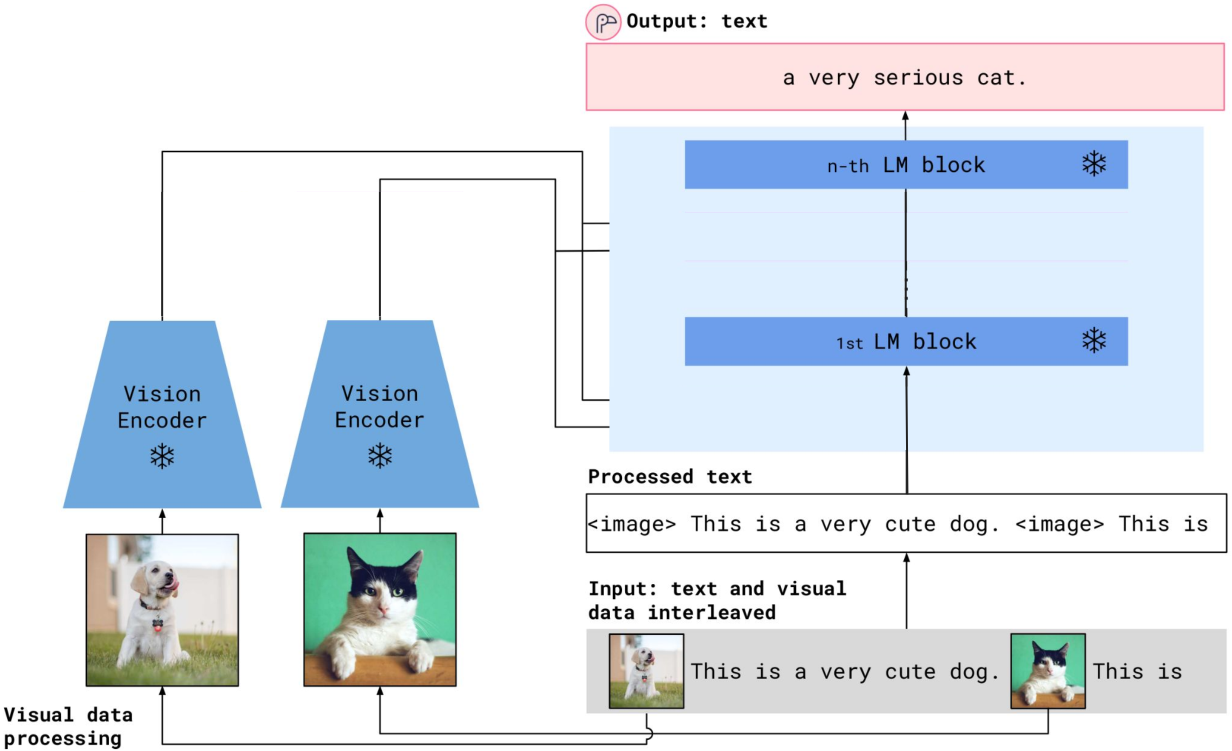

Cute
Transformer block
Transformer block
Transformer block
Cats
are
so

Flamingo

What kind of model is this? (think types of LLMs)

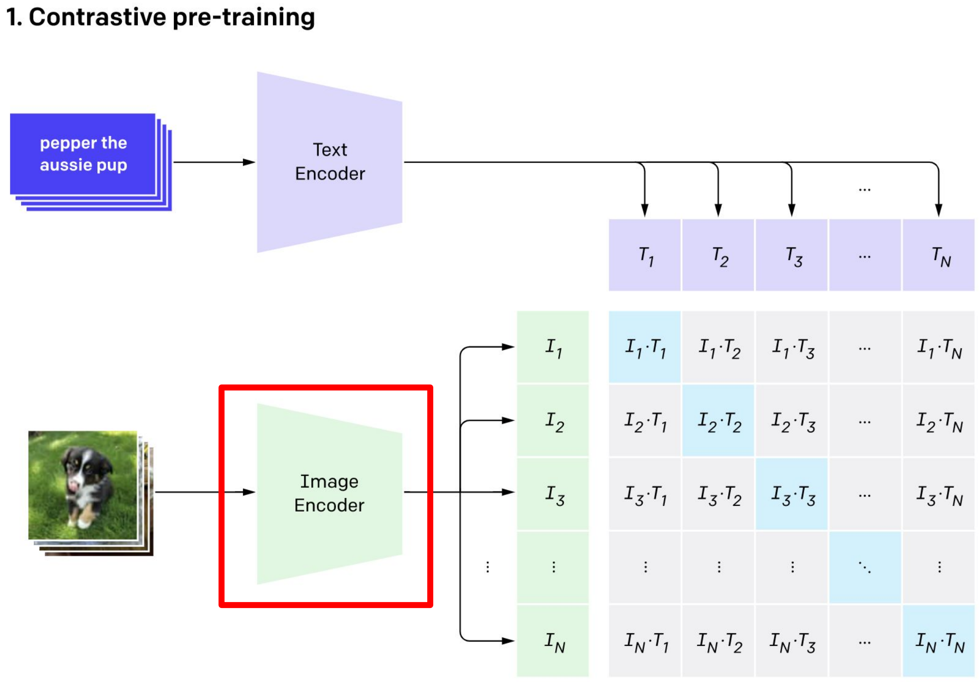


Pre-trained parts of Flamingo



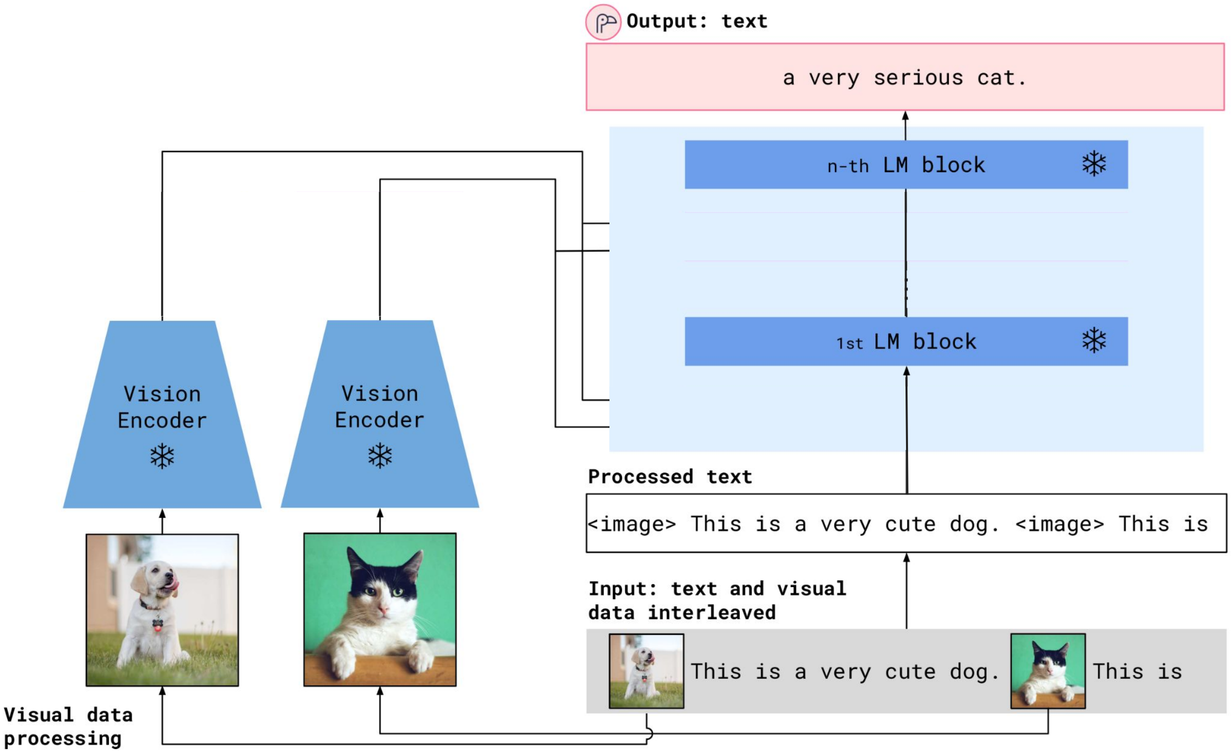
Alayrac et al “Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

CLIP Training Objective



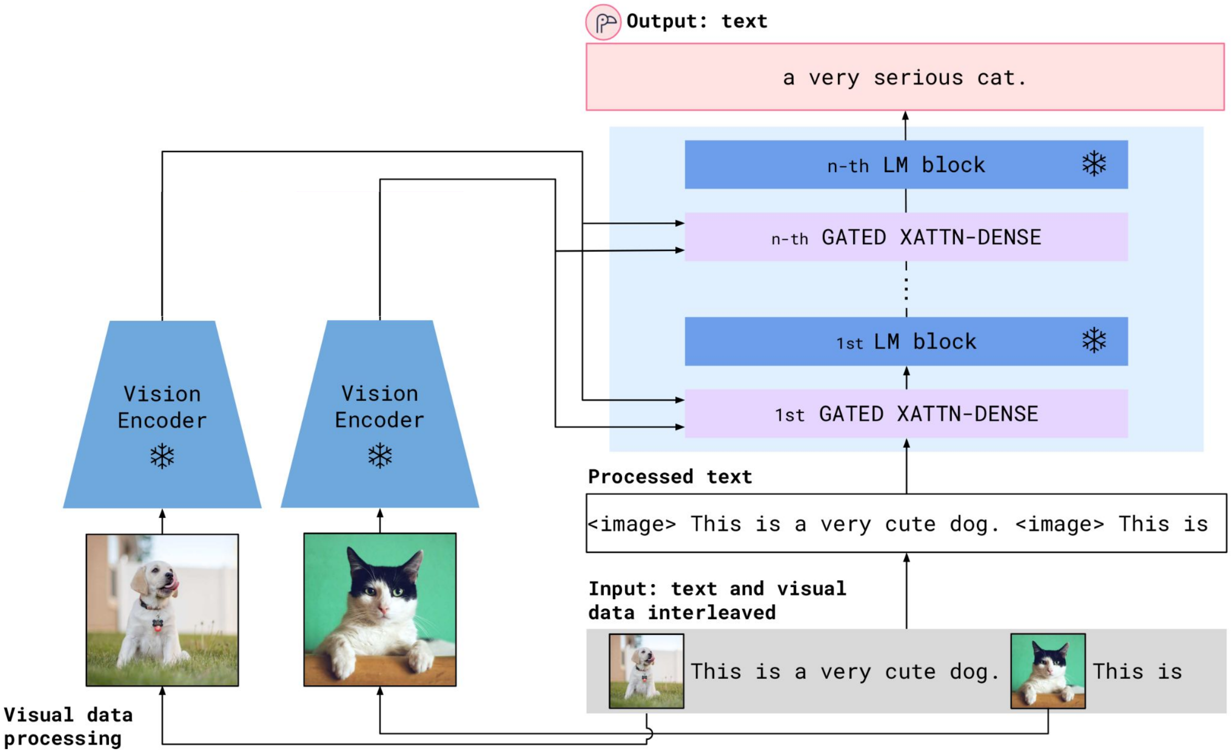
At the end of training, you have a model that will give you a similarity score between an image and a text

Pre-trained parts of Flamingo



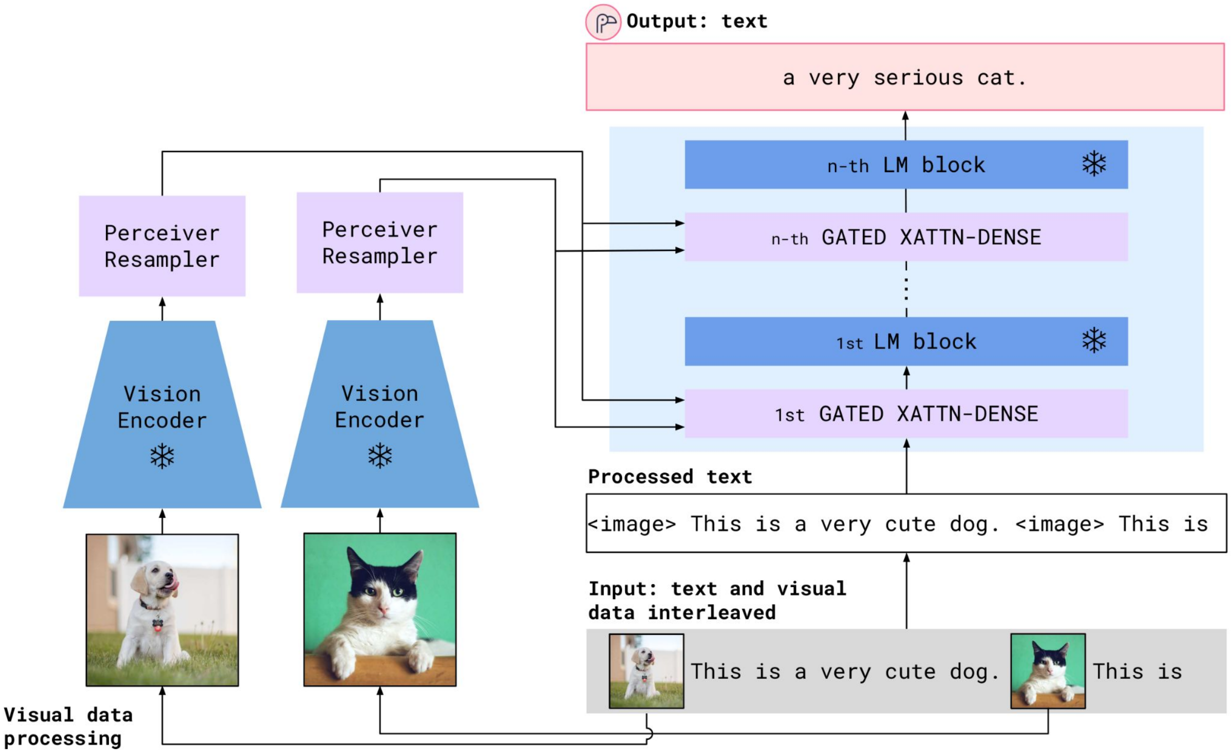
Alayrac et al "Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

Learned parts of Flamingo



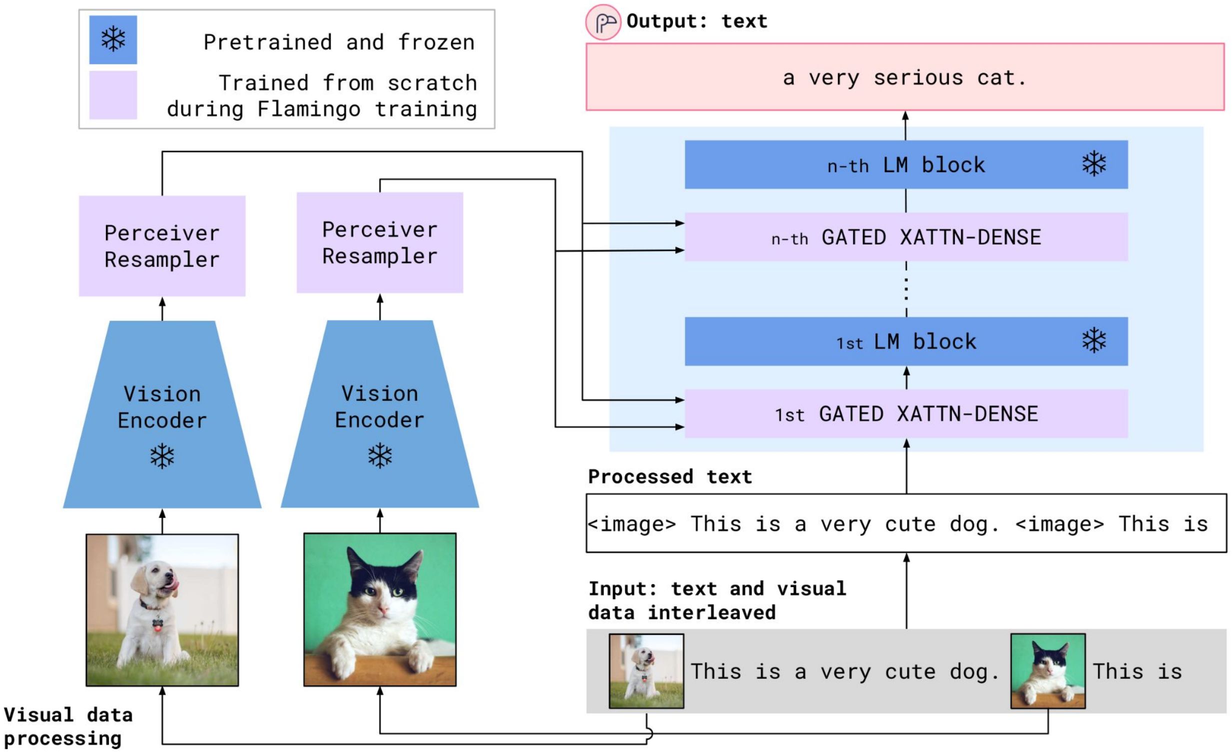
Alayrac et al “Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

Learned parts of Flamingo



Alayrac et al "Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

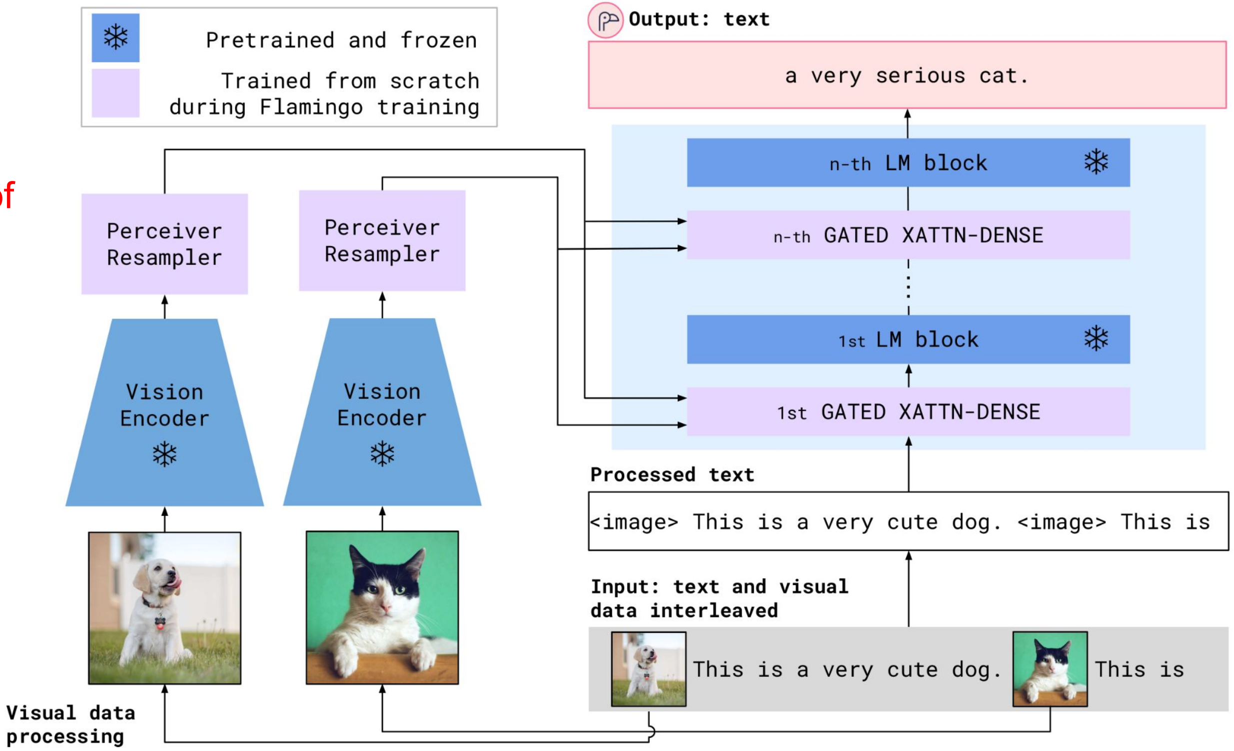
Flamingo full architecture



Alayrac et al "Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

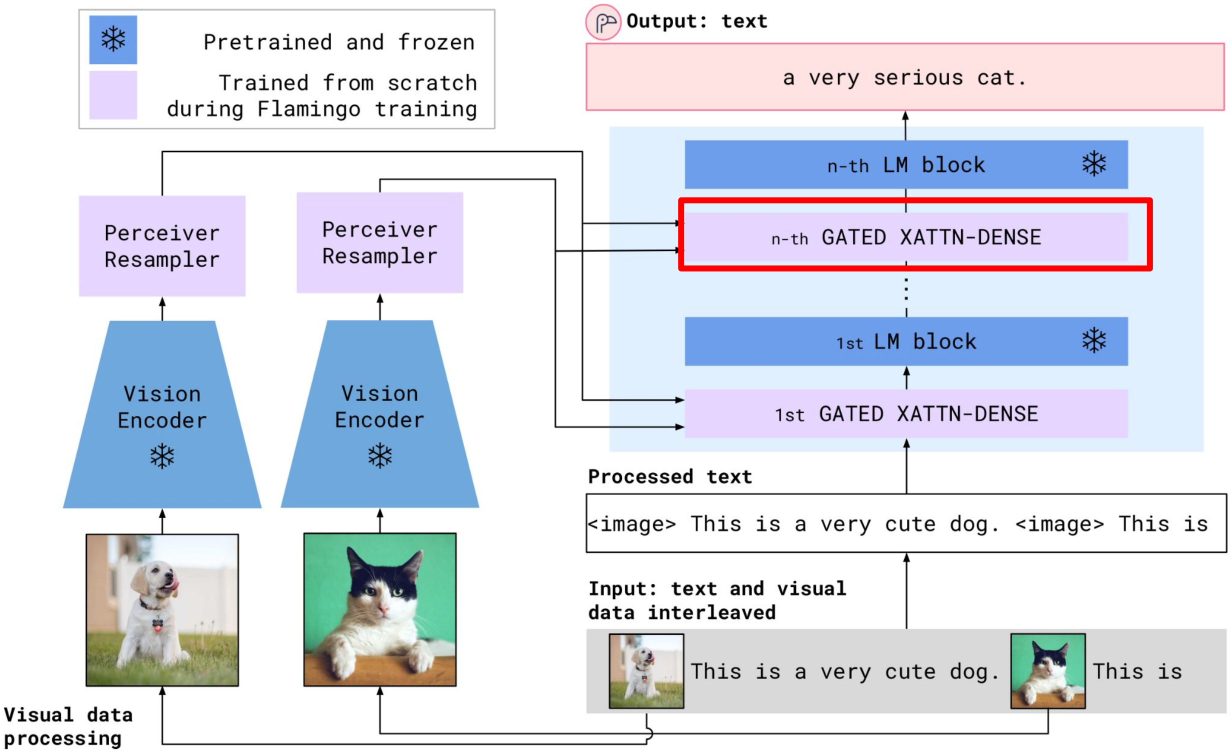
Flamingo full architecture

Learned method of down-sampling image/video representations



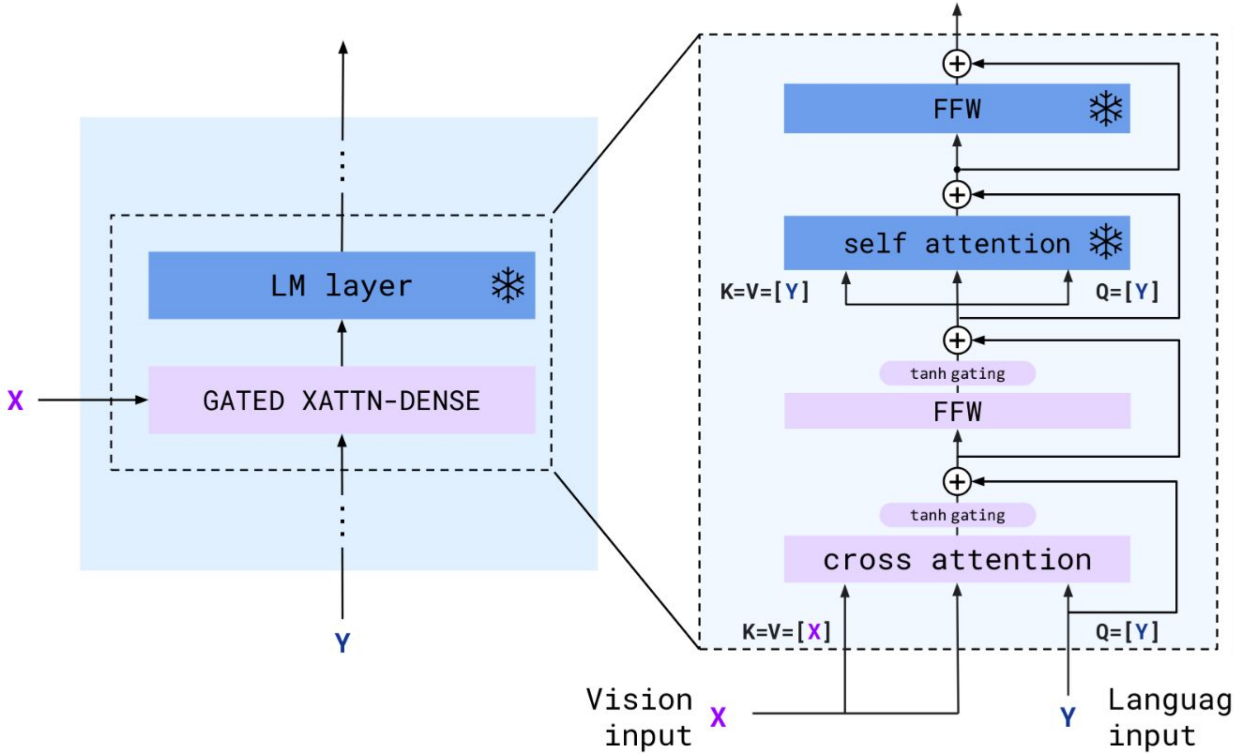
Alayrac et al "Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

Flamingo full architecture



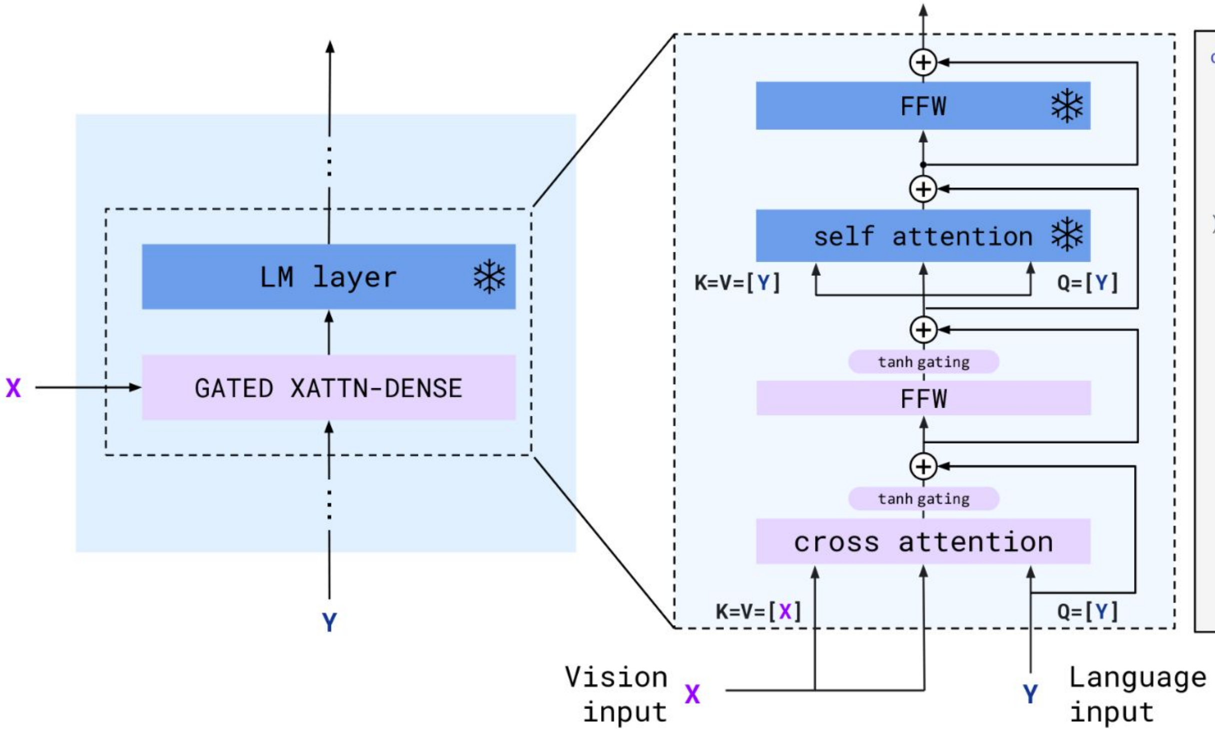
Alayrac et al "Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

Flamingo gated cross-attention



Alayrac et al “Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

Flamingo gated cross-attention



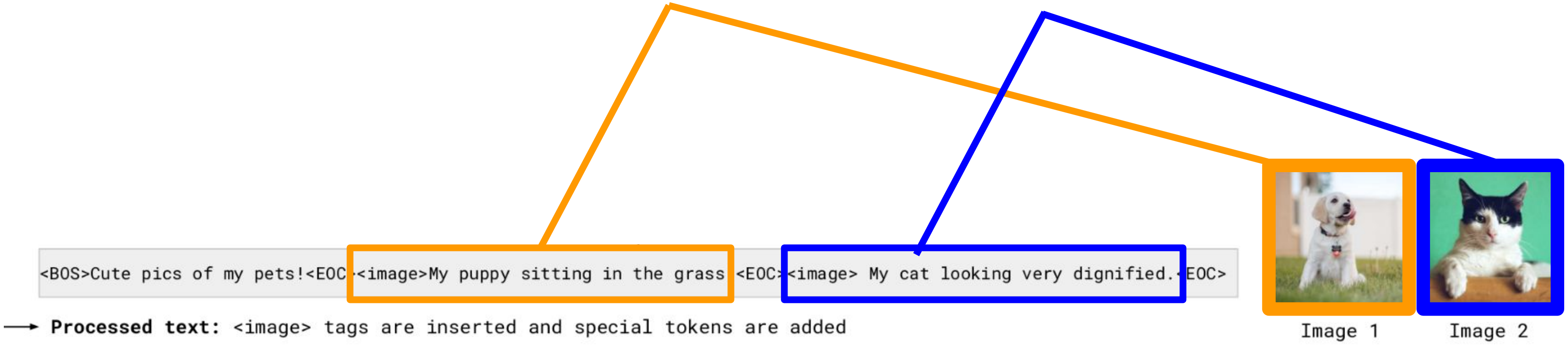

```python
def gated_xattn_dense(
    y,  # input language features
    x,  # input visual features
    alpha_xattn, # xattn gating parameter – init at 0.
    alpha_dense, # ffw gating parameter – init at 0.
):
  """Applies a GATED XATTN-DENSE layer."""

  # 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
```

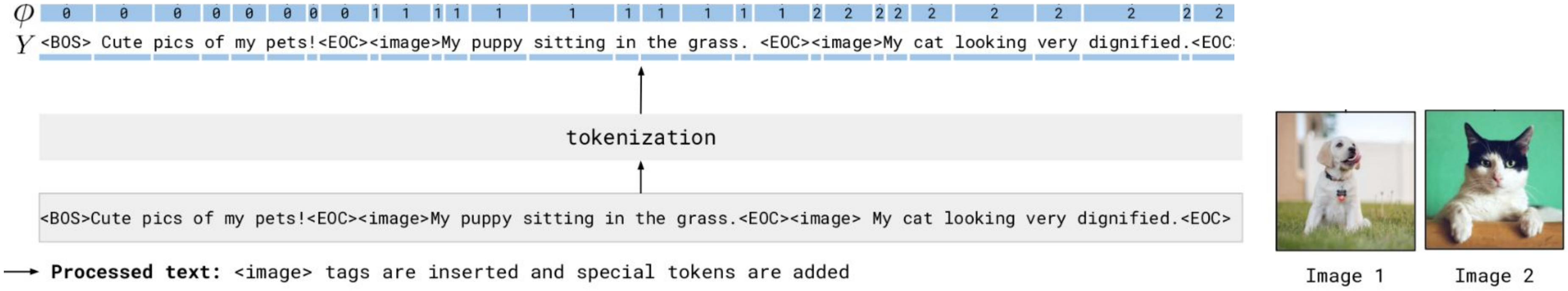
Alayrac et al “Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

Flamingo masked attention



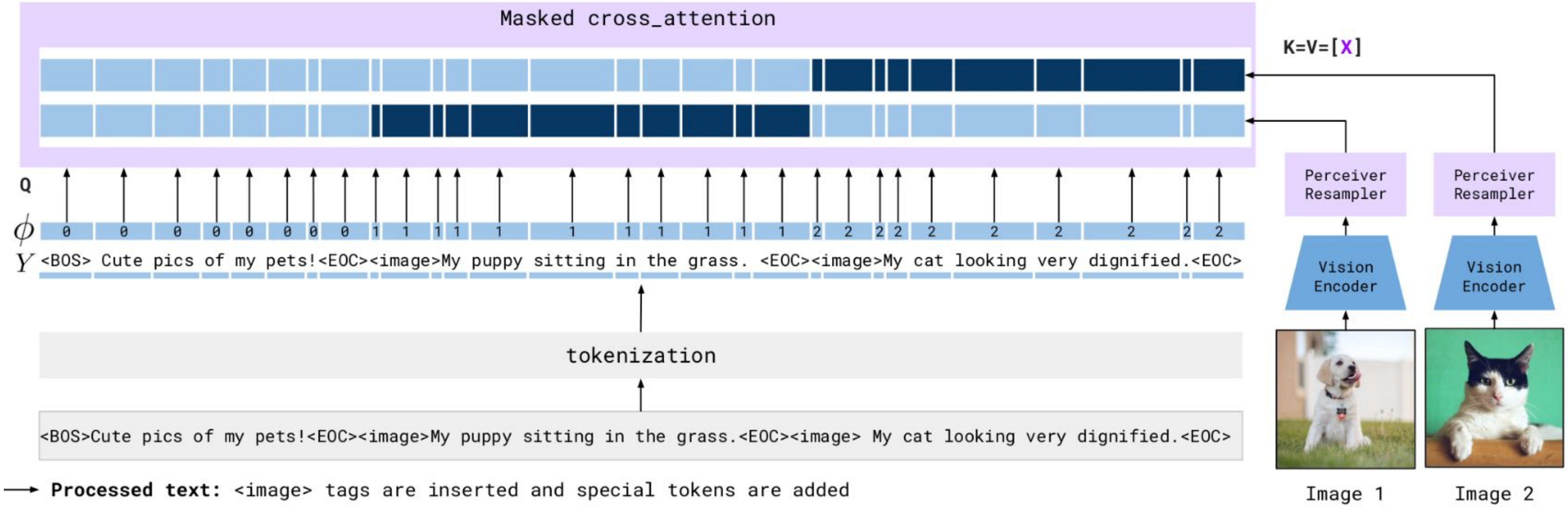
Alayrac et al “Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

Flamingo masked attention



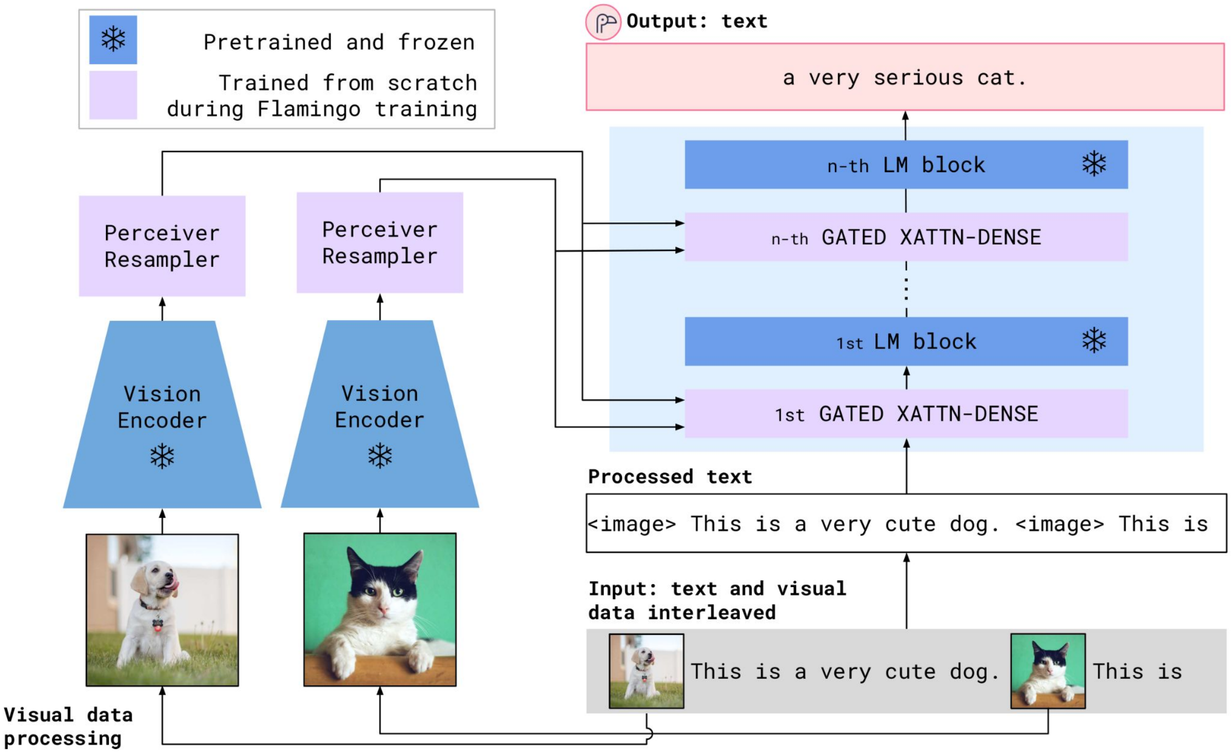
Alayrac et al "Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

Flamingo masked attention



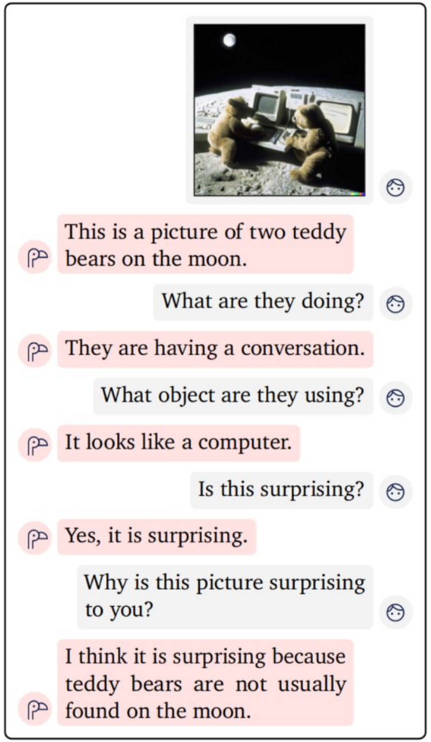
Alayrac et al “Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

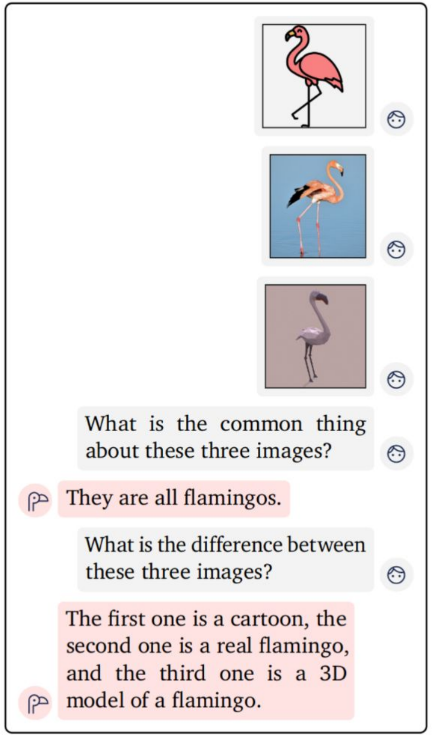
Flamingo full architecture

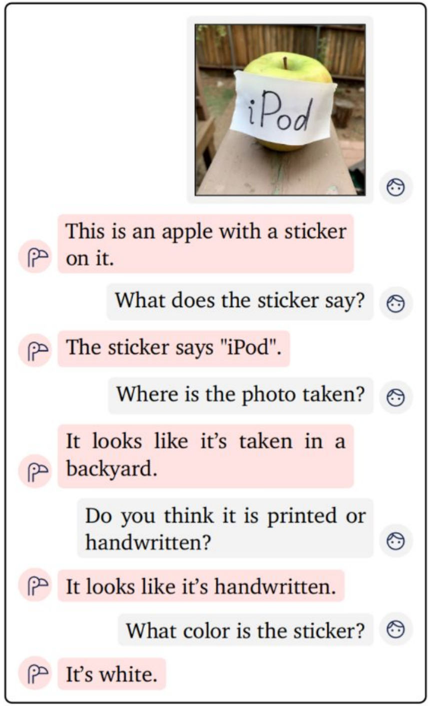


Alayrac et al "Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

Flamingo results



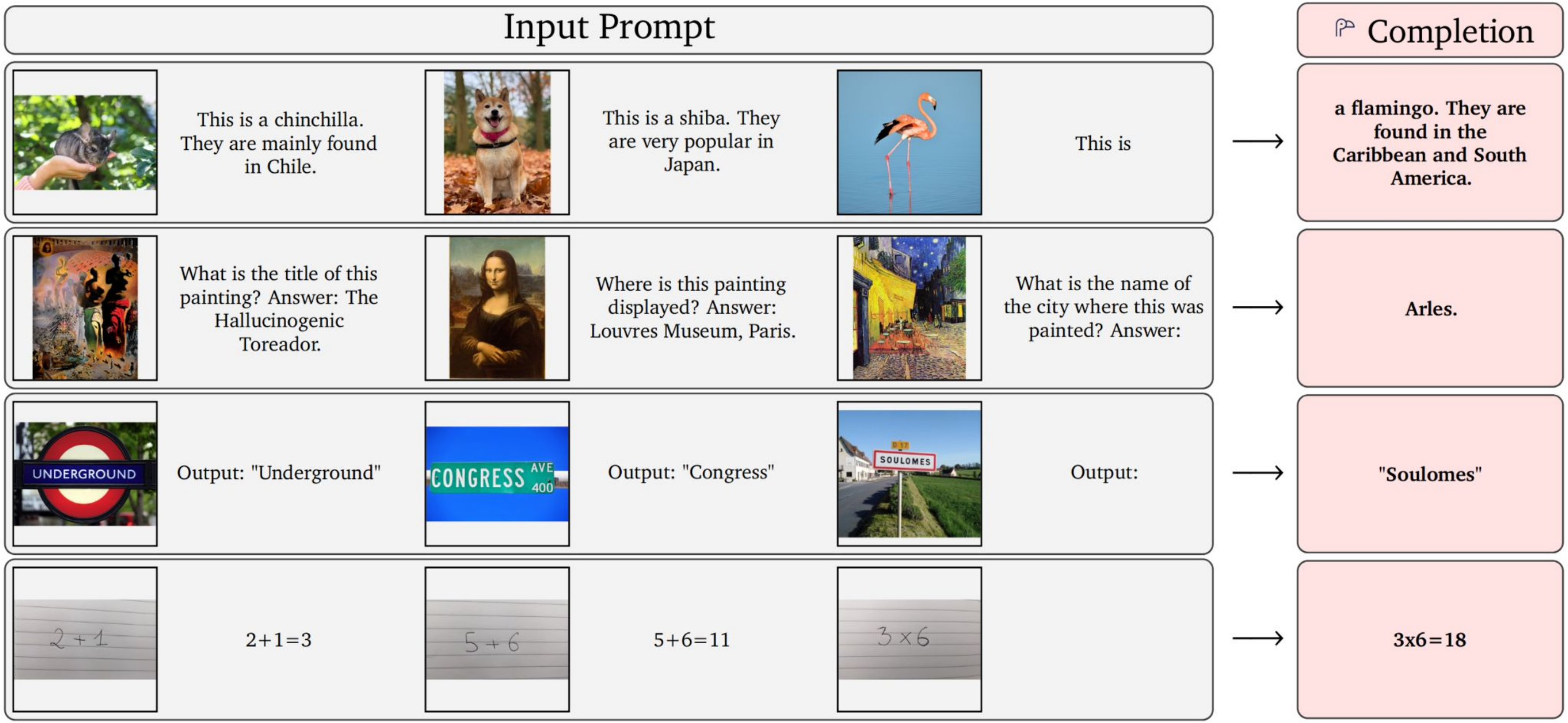




Alayrac et al "Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

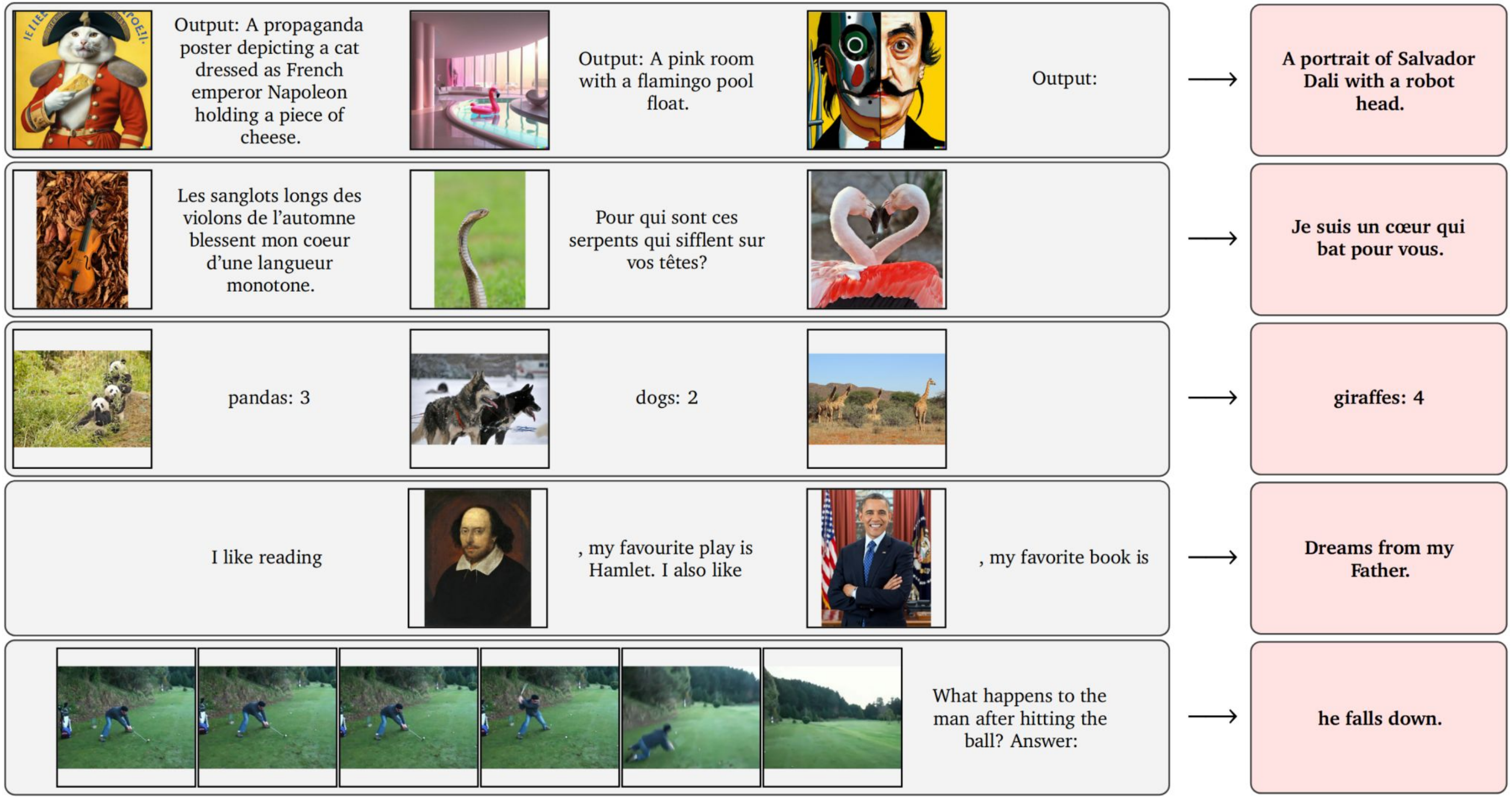
Flamingo results

What is this type of learning called?



Alayrac et al “Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

Flamingo results



Alayrac et al "Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

Results: zero & few shot

Method	FT	Shot	OKVQA	VQAv2	COCO	MSVDQA	VATEX	VizWiz	Flick30K	MSRVTTQA	iVQA	YouCook2	STAR	VisDial	TextVQA	NextQA	HatefulMemes	RareAct
Zero/Few shot SOTA	✗	(X)	[39] 43.3 (16)	[124] 38.2 (4)	[134] 32.2 (0)	[64] 35.2 (0)	-	-	-	[64] 19.2 (0)	[145] 12.2 (0)	-	[153] 39.4 (0)	[87] 11.6 (0)	-	-	[94] 66.1 (0)	[94] 40.7 (0)
Flamingo-3B	✗	0	41.2	49.2	73.0	27.5	40.1	28.9	60.6	11.0	32.7	55.8	39.6	46.1	30.1	21.3	53.7	58.4
	✗	4	43.3	53.2	85.0	33.0	50.0	34.0	72.0	14.9	35.7	64.6	41.3	47.3	32.7	22.4	53.6	-
	✗	8	44.6	55.4	90.6	37.0	54.5	38.4	71.7	19.6	36.8	68.0	40.6	47.6	32.4	23.9	54.7	-
	✗	16	45.6	56.7	95.4	40.2	57.1	43.3	73.4	23.4	37.4	73.2	40.1	47.5	31.8	25.2	55.3	-
	✗	32	45.9	57.1	99.0	42.6	59.2	45.5	71.2	25.6	37.7	76.7	41.6	OOC	30.6	26.1	56.3	-
Flamingo-9B	✗	0	44.7	51.8	79.4	30.2	39.5	28.8	61.5	13.7	35.2	55.0	41.8	48.0	31.8	23.0	57.0	57.9
	✗	4	49.3	56.3	93.1	36.2	51.7	34.9	72.6	18.2	37.7	70.8	42.8	50.4	33.6	24.7	62.7	-
	✗	8	50.0	58.0	99.0	40.8	55.2	39.4	73.4	23.9	40.0	75.0	**43.4**	51.2	33.6	25.8	63.9	-
	✗	16	50.8	59.4	102.2	44.5	58.5	43.0	72.7	27.6	41.5	77.2	42.4	51.3	33.5	27.6	64.5	-
	✗	32	51.0	60.4	106.3	47.2	57.4	44.0	72.8	29.4	40.7	77.3	41.2	OOC	32.6	28.4	63.5	-
Flamingo	✗	0	50.6	56.3	84.3	35.6	46.7	31.6	67.2	17.4	40.7	60.1	39.7	52.0	35.0	26.7	46.4	**60.8**
	✗	4	57.4	63.1	103.2	41.7	56.0	39.6	75.1	23.9	44.1	74.5	42.4	55.6	36.5	30.8	68.6	-
	✗	8	57.5	65.6	108.8	45.5	60.6	44.8	78.2	27.6	44.8	80.7	42.3	56.4	37.3	32.3	**70.0**	-
	✗	16	57.8	66.8	110.5	48.4	62.8	48.4	**78.9**	30.0	45.2	84.2	41.1	**56.8**	37.6	32.9	**70.0**	-
	✗	32	**57.8**	**67.6**	**113.8**	**52.3**	**65.1**	**49.8**	75.4	**31.0**	**45.3**	**86.8**	42.2	OOC	**37.9**	**33.5**	**70.0**	-
Pretrained FT SOTA	✔	(X)	54.4 [39] (10K)	80.2 [150] (444K)	143.3 [134] (500K)	47.9 [32] (27K)	76.3 [165] (500K)	57.2 [70] (20K)	67.4 [162] (30K)	46.8 [57] (130K)	35.4 [145] (6K)	138.7 [142] (10K)	36.7 [138] (46K)	75.2 [87] (123K)	54.7 [147] (20K)	25.2 [139] (38K)	75.4 [60] (9K)	-

Results: zero & few shot

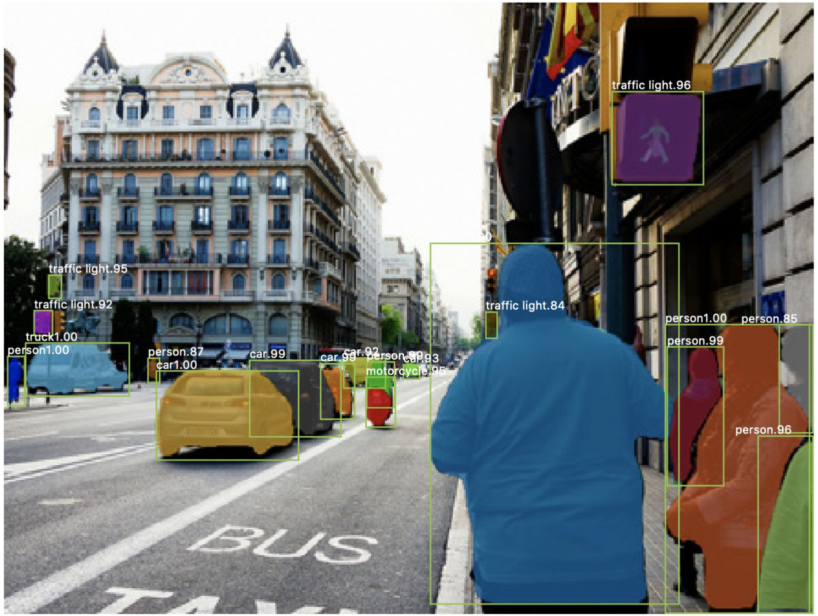
Method	FT	Shot	OKVQA	VQAv2	COCO	MSVDQA	VATEX	VizWiz	Flick30K	MSRVTTQA	iVQA	YouCook2	STAR	VisDial	TextVQA	NextQA	HatefulMemes	RareAct
Zero/Few shot SOTA	✗	(X)	[39] 43.3 (16)	[124] 38.2 (4)	[134] 32.2 (0)	[64] 35.2 (0)	-	-	-	[64] 19.2 (0)	[145] 12.2 (0)	-	[153] 39.4 (0)	[87] 11.6 (0)	-	-	[94] 66.1 (0)	[94] 40.7 (0)
Flamingo-3B	✗	0	41.2	49.2	73.0	27.5	40.1	28.9	60.6	11.0	32.7	55.8	39.6	46.1	30.1	21.3	53.7	58.4
	✗	4	43.3	53.2	85.0	33.0	50.0	34.0	72.0	14.9	35.7	64.6	41.3	47.3	32.7	22.4	53.6	-
	✗	8	44.6	55.4	90.6	37.0	54.5	38.4	71.7	19.6	36.8	68.0	40.6	47.6	32.4	23.9	54.7	-
	✗	16	45.6	56.7	95.4	40.2	57.1	43.3	73.4	23.4	37.4	73.2	40.1	47.5	31.8	25.2	55.3	-
	✗	32	45.9	57.1	99.0	42.6	59.2	45.5	71.2	25.6	37.7	76.7	41.6	OOC	30.6	26.1	56.3	-
Flamingo-9B	✗	0	44.7	51.8	79.4	30.2	39.5	28.8	61.5	13.7	35.2	55.0	41.8	48.0	31.8	23.0	57.0	57.9
	✗	4	49.3	56.3	93.1	36.2	51.7	34.9	72.6	18.2	37.7	70.8	42.8	50.4	33.6	24.7	62.7	-
	✗	8	50.0	58.0	99.0	40.8	55.2	39.4	73.4	23.9	40.0	75.0	**43.4**	51.2	33.6	25.8	63.9	-
	✗	16	50.8	59.4	102.2	44.5	58.5	43.0	72.7	27.6	41.5	77.2	42.4	51.3	33.5	27.6	64.5	-
	✗	32	51.0	60.4	106.3	47.2	57.4	44.0	72.8	29.4	40.7	77.3	41.2	OOC	32.6	28.4	63.5	-
Flamingo	✗	0	50.6	56.3	84.3	35.6	46.7	31.6	67.2	17.4	40.7	60.1	39.7	52.0	35.0	26.7	46.4	**60.8**
	✗	4	57.4	63.1	103.2	41.7	56.0	39.6	75.1	23.9	44.1	74.5	42.4	55.6	36.5	30.8	68.6	-
	✗	8	57.5	65.6	108.8	45.5	60.6	44.8	78.2	27.6	44.8	80.7	42.3	56.4	37.3	32.3	**70.0**	-
	✗	16	57.8	66.8	110.5	48.4	62.8	48.4	**78.9**	30.0	45.2	84.2	41.1	**56.8**	37.6	32.9	**70.0**	-
	✗	32	**57.8**	**67.6**	**113.8**	**52.3**	**65.1**	**49.8**	75.4	**31.0**	**45.3**	**86.8**	42.2	OOC	**37.9**	**33.5**	**70.0**	-
Pretrained FT SOTA	✔	(X)	54.4 [39] (10K)	80.2 [150] (444K)	143.3 [134] (500K)	47.9 [32] (27K)	76.3 [165] (500K)	57.2 [70] (20K)	67.4 [162] (30K)	46.8 [57] (130K)	35.4 [145] (6K)	138.7 [142] (10K)	36.7 [138] (46K)	75.2 [87] (123K)	54.7 [147] (20K)	25.2 [139] (38K)	75.4 [60] (9K)	-

Foundation Models

Language	Classification	LM + Vision	And More!	Chaining
ELMo	**CLIP**	**Flamingo**	**Segment Anything**	**LMs + CLIP**
BERT	**CoCa**	GPT-4V	Whisper	**Visual Programming**
GPT		Gemini	Dalle	
T5			Stable Diffusion	
			Imagen	

Segment Anything Model (SAM)

What does it mean to have a segmentation foundation model?



Masking model trained on dataset of specific number of objects (80 in COCO)

Model outputs masks of all objects in that image that is one of the categories of interest

Images: He et al. Mask R-CNN. 2017

Segment Anything Model (SAM)

What does it mean to have a segmentation foundation model?

Masking model trained on a dataset of a huge number of categories

Model outputs mask of any objects that the user cares about

Images: Kirillov et al. Segment Anything. 2023.

Segment Anything Model (SAM)

What does it mean to have a segmentation foundation model?

Masking model trained on a dataset of a huge number of categories

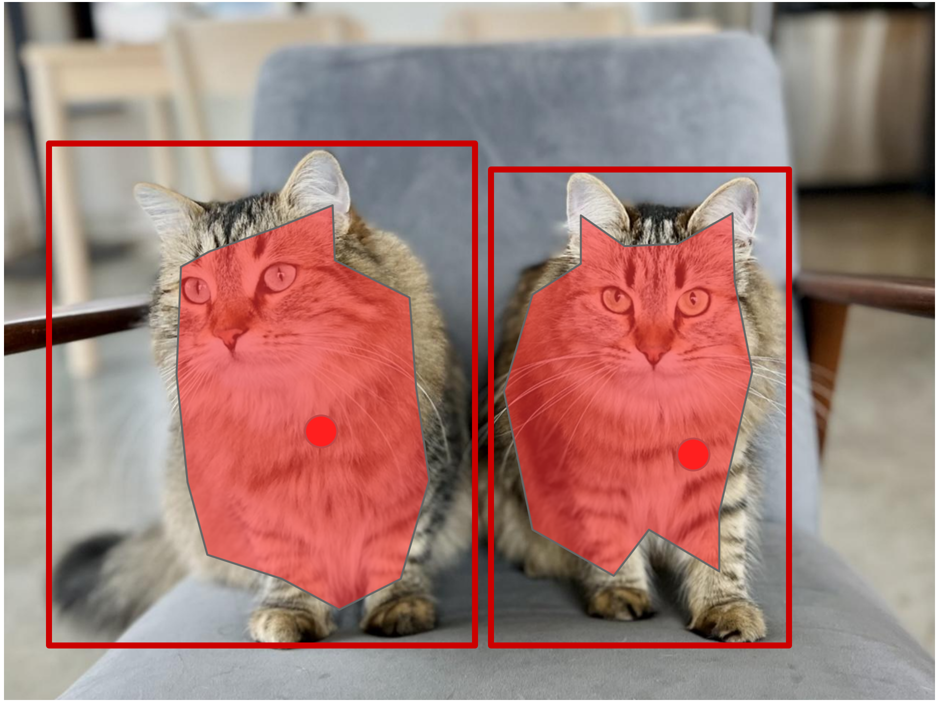
How to get this?

Model outputs mask of any objects that the user cares about

How to know this?

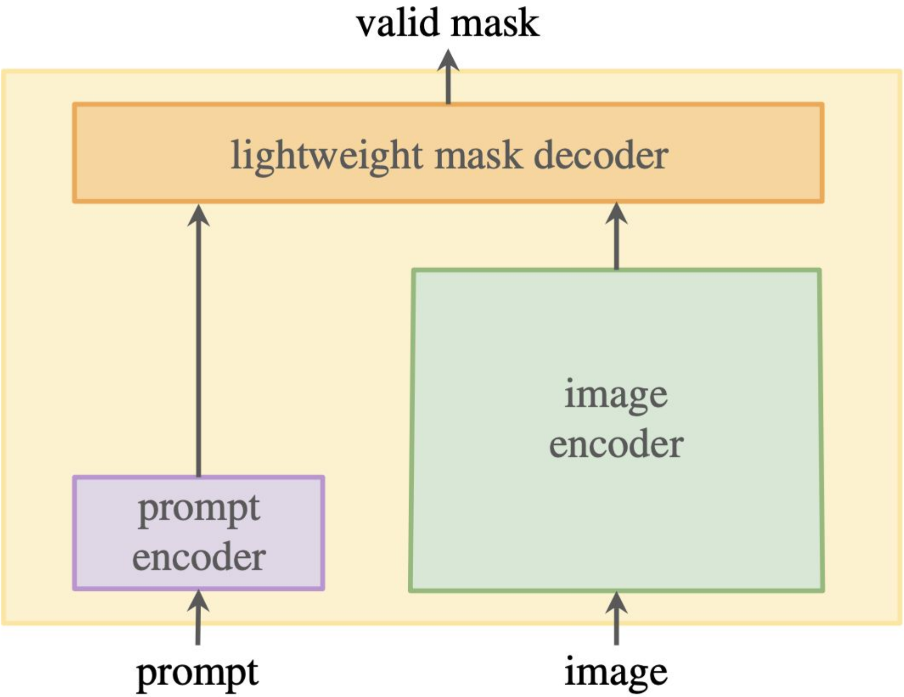
Images: Kirillov et al. Segment Anything. 2023.

How to know what to mask?



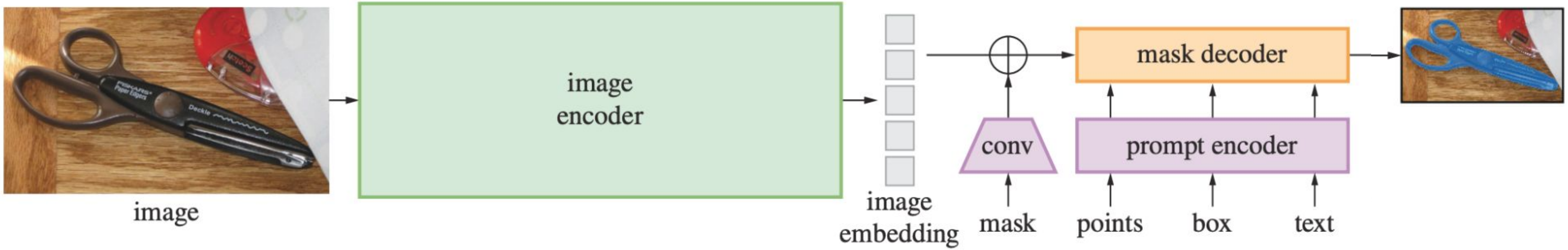
"Cats"

Basic SAM Architecture



Images: Kirillov et al. Segment Anything. 2023.

SAM Architecture



Images: Kirillov et al. Segment Anything. 2023.

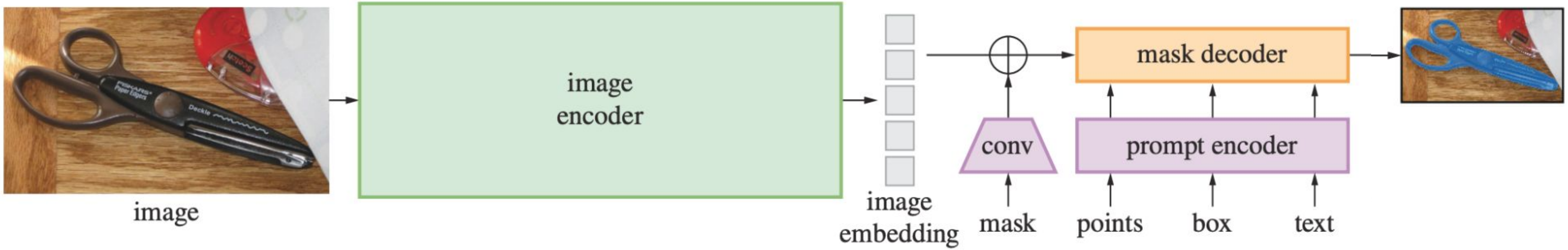
Ambiguity in correct prompt

Images: Kirillov et al. Segment Anything. 2023.

Ambiguity in correct prompt

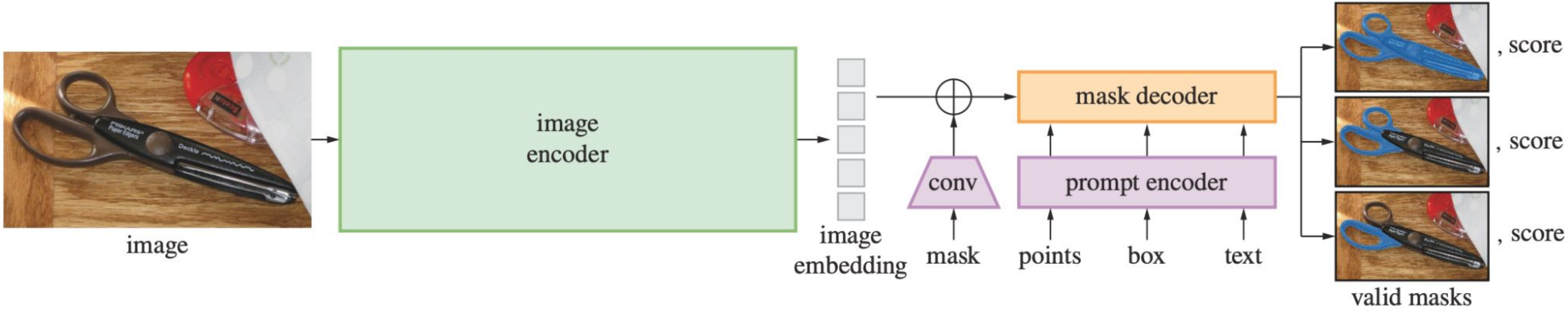
Images: Kirillov et al. Segment Anything. 2023.

SAM Architecture



Images: Kirillov et al. Segment Anything. 2023.

Basic SAM Architecture



1. Loss only calculated with respect to best mask
2. Model also trained to output confidence score for each mask

Images: Kirillov et al. Segment Anything. 2023.

Segment Anything Model (SAM)

What does it mean to have a segmentation foundation model?

Masking model trained on a dataset of a huge number of categories

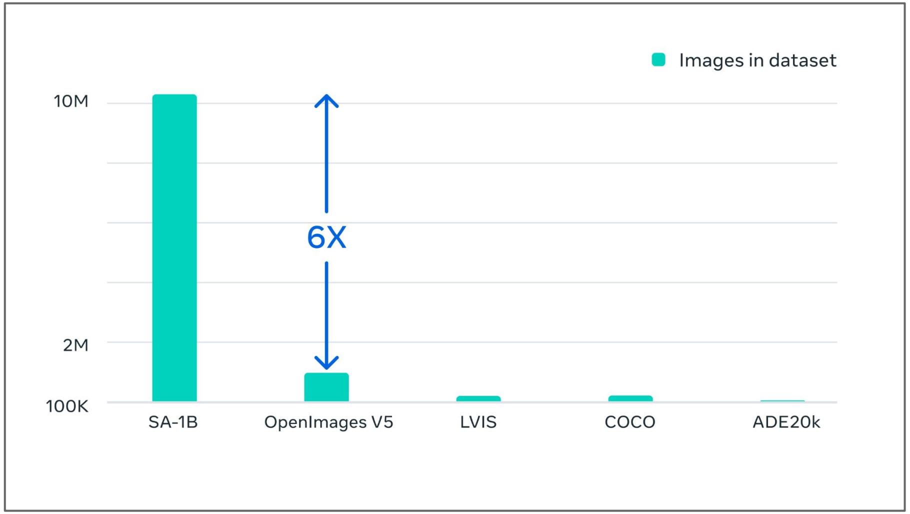
How to get this?

Model outputs mask of any objects that the user cares about

How to know this?

Images: Kirillov et al. Segment Anything. 2023.

Segment Anything Model (SAM)



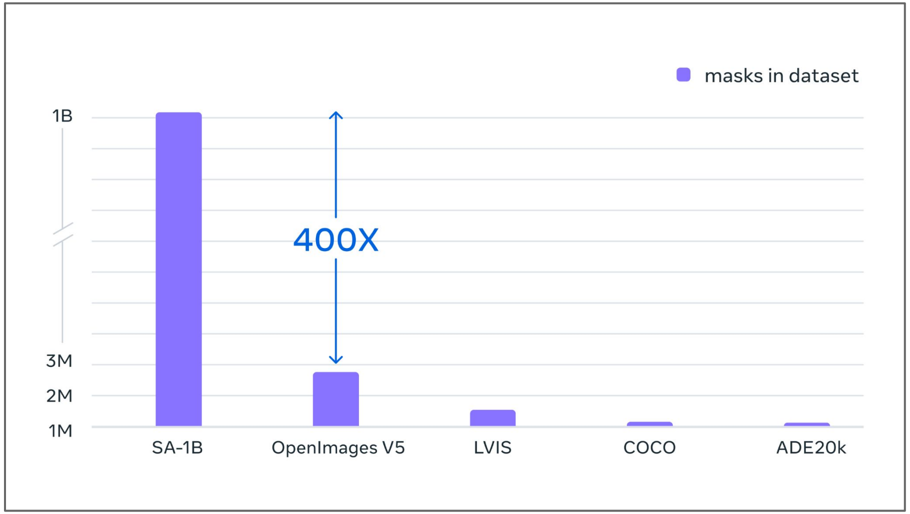
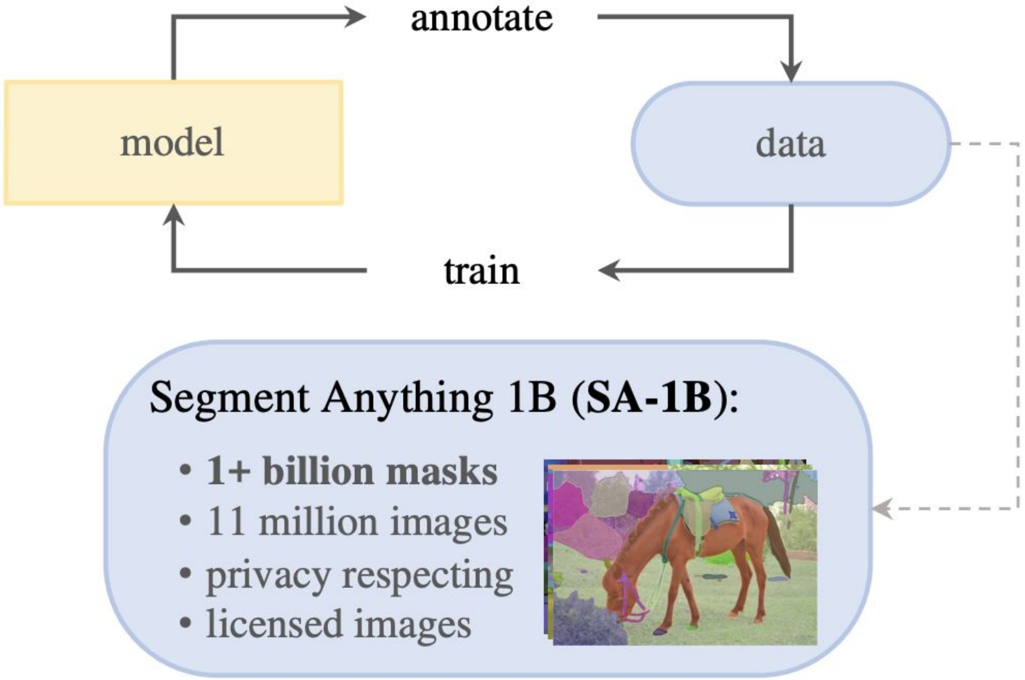


Image Source: https://segment-anything.com/

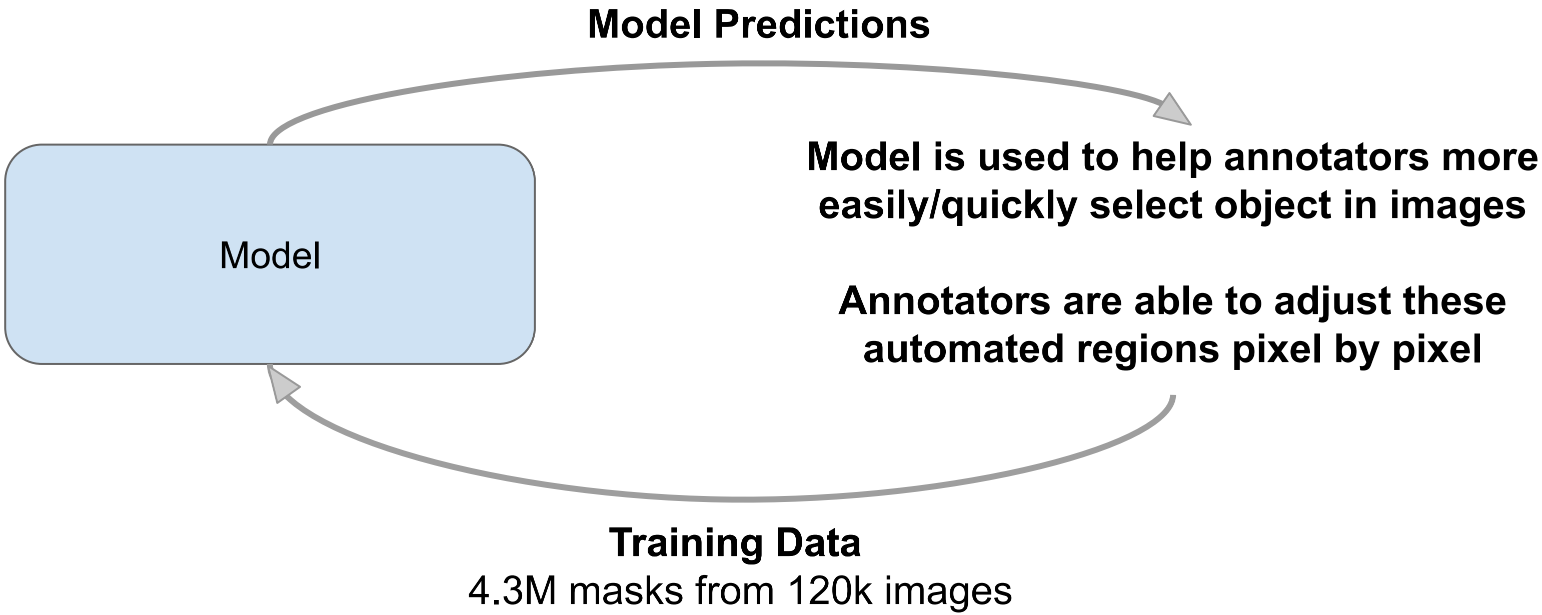
Segment Anything Model (SAM)



Images: Kirillov et al. Segment Anything. 2023.

Segment Anything Model (SAM)

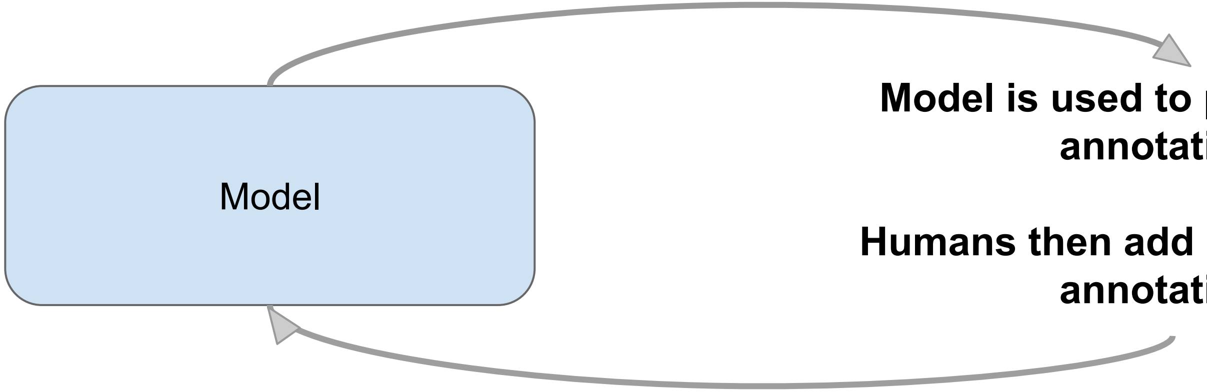
Assisted-manual stage



Segment Anything Model (SAM)

Semi-automatic stage

Model Predictions



Model is used to pre-fill "easy" annotations

Humans then add missing object annotations

Training Data
5.9M masks in 180k images

Segment Anything Model (SAM)

Fully automatic stage

Model Predictions

Model

Model outputs all annotations on a large set of training data.

This is done by getting annotations for objects at every location of the image and then removing unconfident/duplicate annotations

Humans validate random subsamples of annotations

Training Data
1.1B masks in 11M images

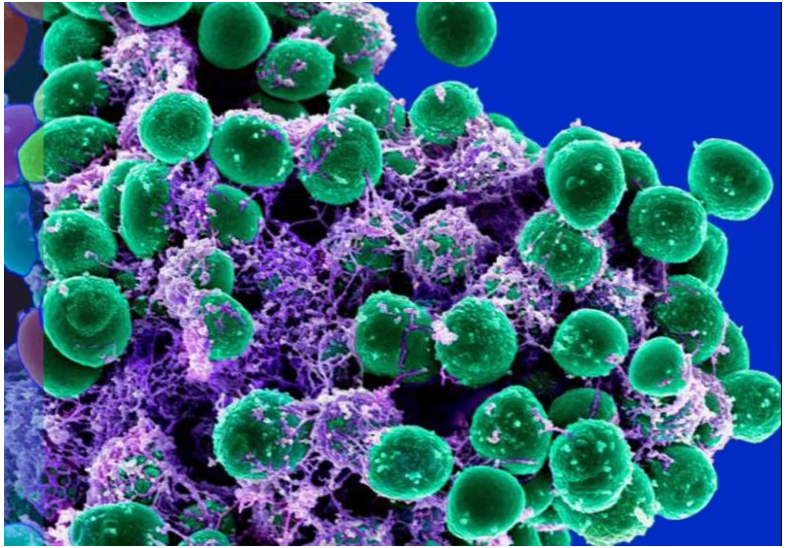
SAM Results

Image Source: Kirillov et al. Segment Anything. 2023

SAM Results

Image Source: Kirillov et al. Segment Anything. 2023

Zero-Shot with SAM



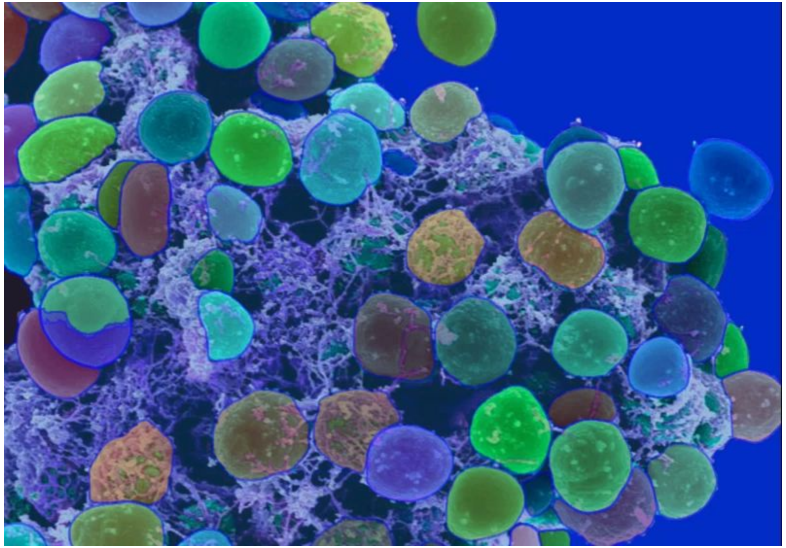


Image Source: https://segment-anything.com/

Zero-Shot with SAM

Image Source: https://segment-anything.com/

Foundation Models

Language	Classification	LM + Vision	And More!	Chaining
ELMo	**CLIP**	**Flamingo**	**Segment Anything**	**LMs + CLIP**
BERT	**CoCa**	GPT-4V	Whisper	**Visual Programming**
GPT		Gemini	Dalle	
T5			Stable Diffusion	
			Imagen	

CuPL (CUstomized Prompts via Language models)

A photo of a marimba
A photo of a viaduct
A photo of a papillon
A photo of a lorikeet




Pratt et al “What does a platypus look like? Generating customized prompts for zero-shot image classification”. 2023.

CuPL (CUstomized Prompts via Language models)

"A marimba is a large wooden percussion instrument that looks like a xylophone."
"A viaduct is a bridge composed of several spans supported by piers or pillars."
"A papillon is a small, spaniel-type dog with a long, silky coat and fringed ears."
"A lorikeet is a small to medium-sized parrot with a brightly colored plumage."




Pratt et al "What does a platypus look like? Generating customized prompts for zero-shot image classification". 2023.

CuPL (CUstomized Prompts via Language models)

"A marimba is a large wooden percussion instrument that looks like a xylophone."
"A viaduct is a bridge composed of several spans supported by piers or pillars."
"A papillon is a small, spaniel-type dog with a long, silky coat and fringed ears."
"A lorikeet is a small to medium-sized parrot with a brightly colored plumage."

Lorikeet

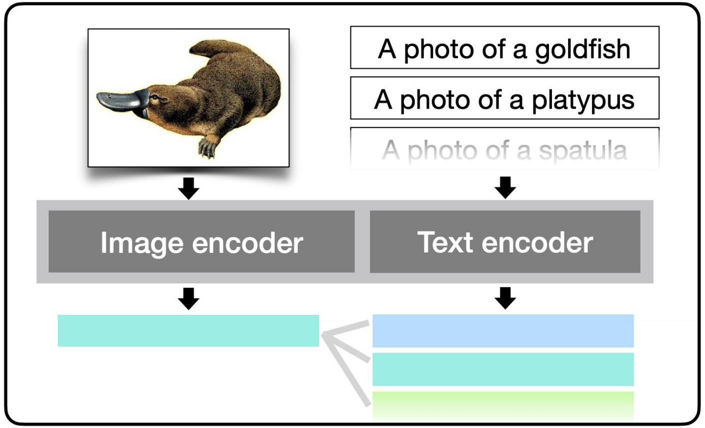
Marimba

Viaduct

Papillon

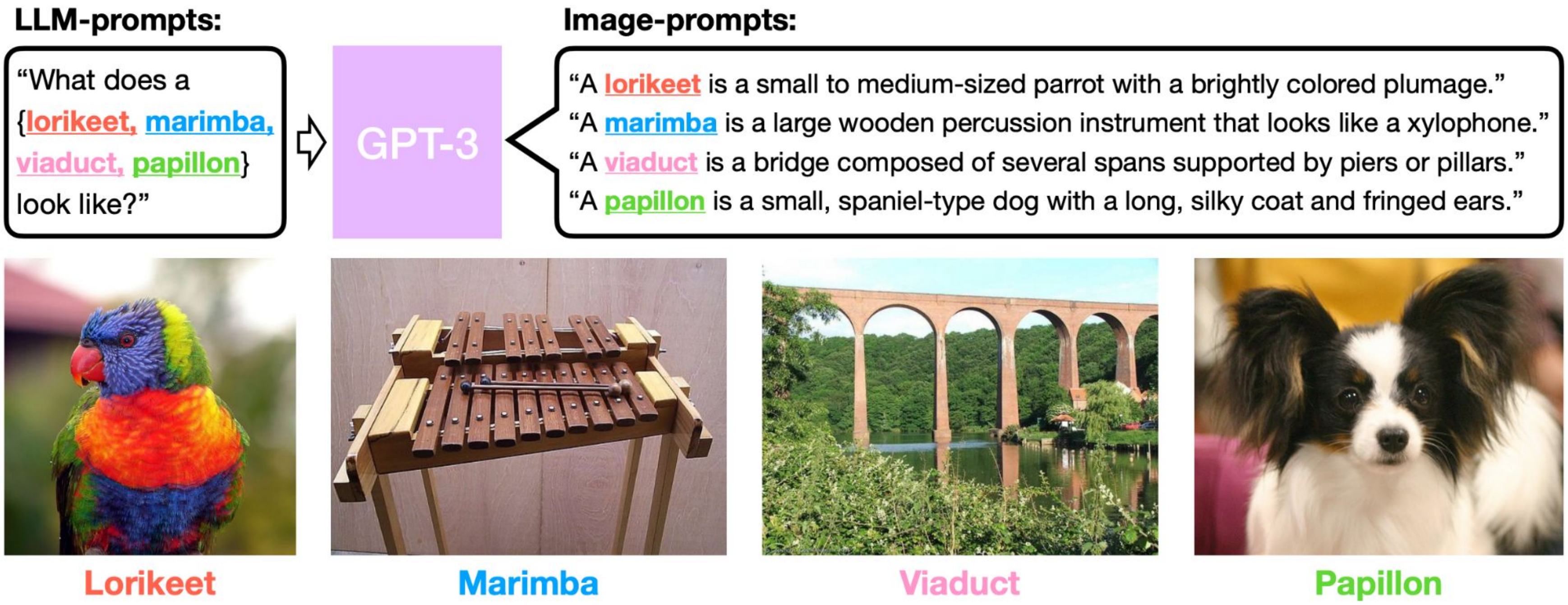
Pratt et al "What does a platypus look like? Generating customized prompts for zero-shot image classification". 2023.

CuPL (CUstomized Prompts via Language models)



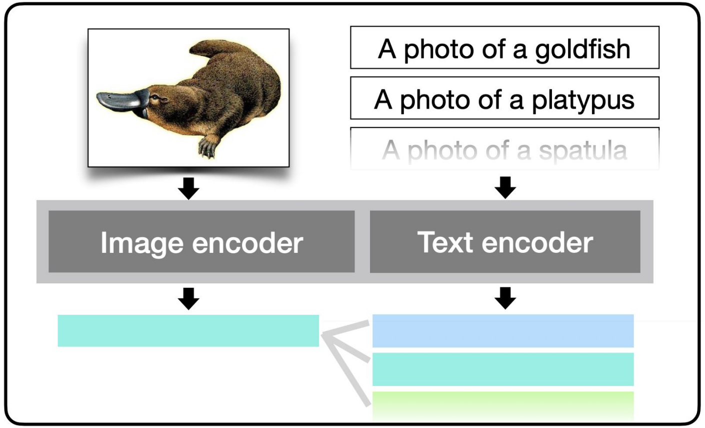
Pratt et al "What does a platypus look like? Generating customized prompts for zero-shot image classification". 2023.

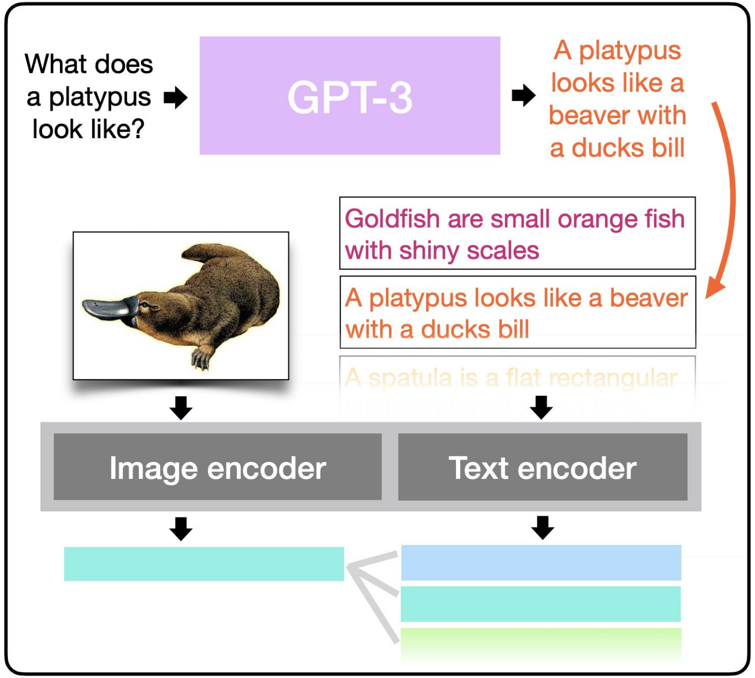
CuPL (CUstomized Prompts via Language models)



Pratt et al “What does a platypus look like? Generating customized prompts for zero-shot image classification”. 2023.

CuPL (CUstomized Prompts via Language models)





Pratt et al "What does a platypus look like? Generating customized prompts for zero-shot image classification". 2023.

CuPL (CUstomized Prompts via Language models)

	ImageNet	DTD	Stanford Cars	SUN397	Food101	FGVC Aircraft	Oxford Pets	Caltech101	Flowers 102	UCF101	Kinetics-700	RESISC45	CIFAR-10	CIFAR-100	Birdsnap
std	75.54	55.20	77.53	69.31	93.08	32.88	93.33	93.24	78.53	77.45	60.07	71.10	95.59	78.26	50.43
# hw	80	8	8	2	1	2	1	34	1	48	28	18	18	18	1
CuPL (base)	76.19	58.90	76.49	72.74	93.33	36.69	93.37	93.45	78.83	77.74	60.24	68.96	95.81	78.47	51.11
Δ std	+0.65	+3.70	-1.04	+3.43	+0.25	+3.81	+0.04	+0.21	+0.30	+0.29	+0.17	-2.14	+0.22	+0.21	+0.63
# hw	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3

Pratt et al “What does a platypus look like? Generating customized prompts for zero-shot image classification”. 2023.

VisProg (visual programming)

Many Visual Question Answering models which have been trained to do this type of task

`Are there 3 people in the boat?`

Gupta et al "Visual Programming: Compositional visual reasoning without training". 2023.

VisProg (visual programming)

LEFT:

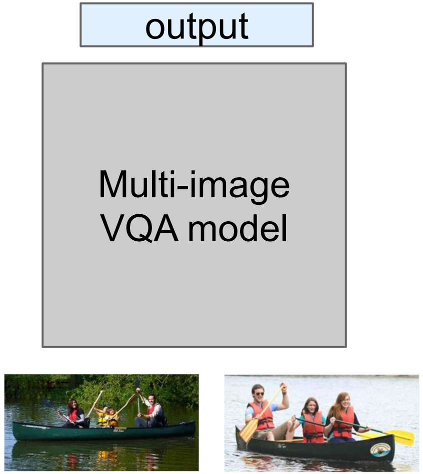
RIGHT:

Statement: The left and right image contains a total of six people and two boats.

Gupta et al "Visual Programming: Compositional visual reasoning without training". 2023.

VisProg (visual programming)

Train a new model for your task



Write a python script with the models you have

```
Class MyMultiImageVQA():

  Def ProcessIms():
    Ans1 = VQA(Image1)
    Ans2 = VQA(Image2)
    Return Ans1 + Ans2
```

General to 2 images now, but not beyond that

Gupta et al "Visual Programming: Compositional visual reasoning without training". 2023.

VisProg (visual programming)

Statement: The left and right image contains a total of six people and two boats.

GPT

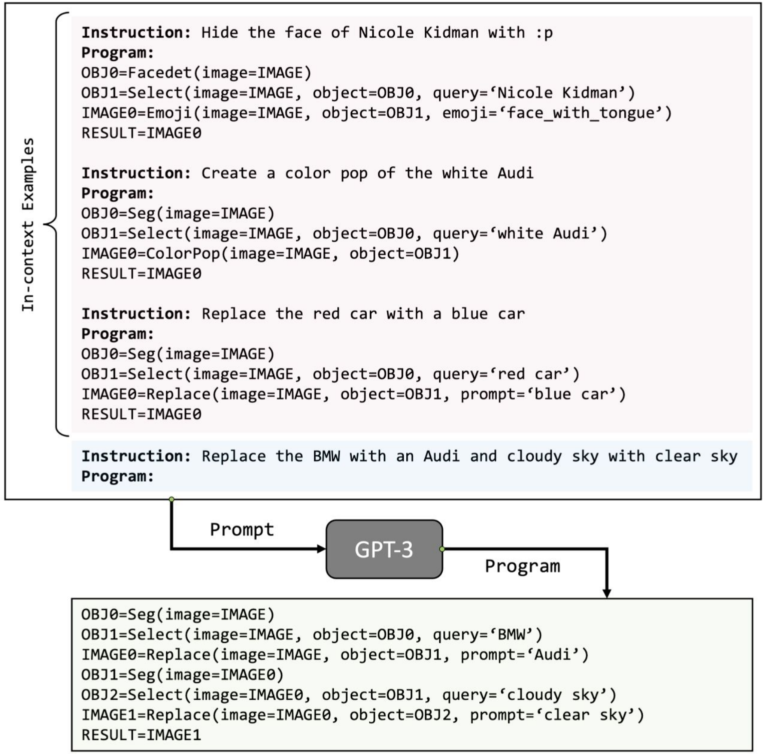
```
Class MyMultiImageVQA():

  Def ProcessIms():
    Ans1 = VQA(Image1)
    Ans2 = VQA(Image2)
    Return Ans1 + Ans2
```

False

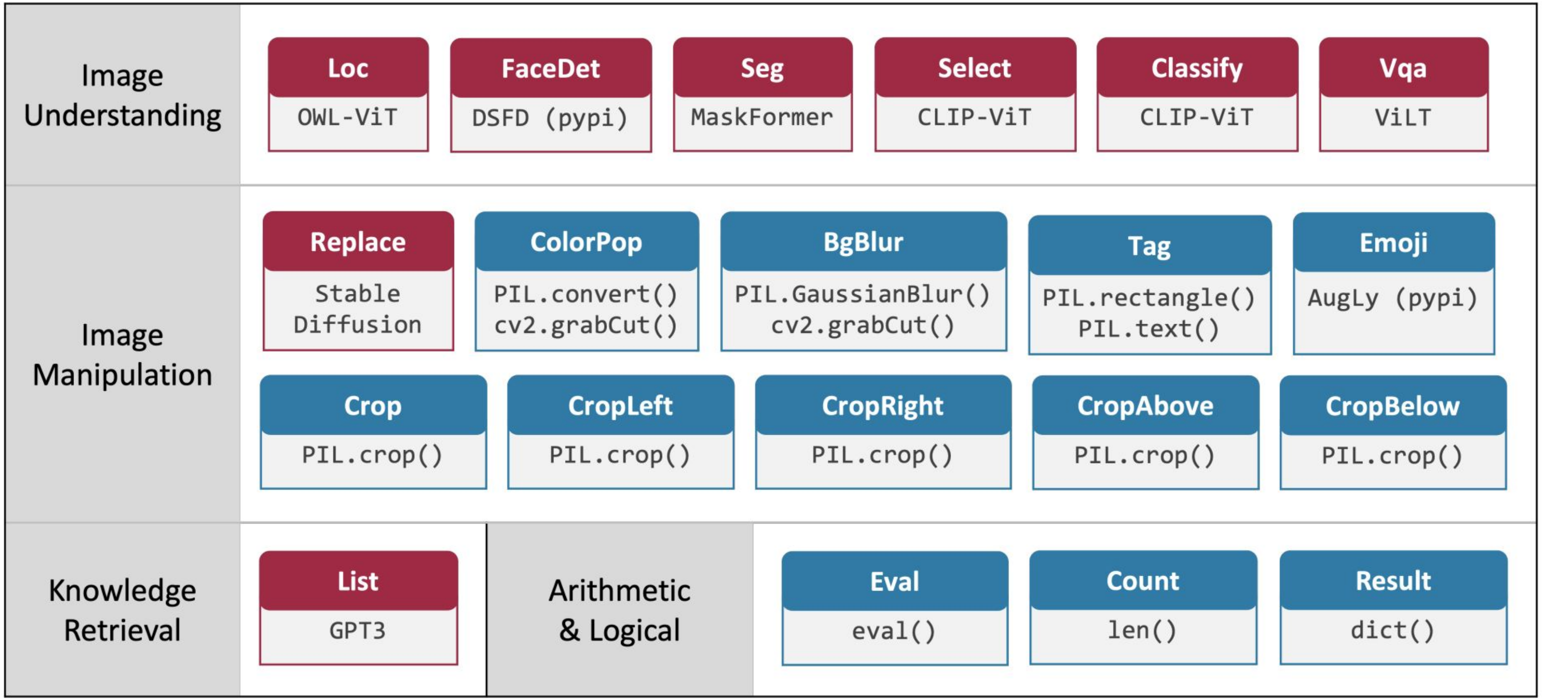
Gupta et al "Visual Programming: Compositional visual reasoning without training". 2023.

VisProg (visual programming)



Gupta et al "Visual Programming: Compositional visual reasoning without training". 2023.

VisProg (visual programming)



Gupta et al “Visual Programming: Compositional visual reasoning without training”. 2023.

VisProg (visual programming)

Natural Language Visual Reasoning

LEFT:

RIGHT:

Statement: The left and right image contains a total of six people and two boats.

Program:

```
ANSWER0=Vqa(image=LEFT, question='How many people are in the image?')
ANSWER1=Vqa(image=RIGHT, question='How many people are in the image?')
ANSWER2=Vqa(image=LEFT, question='How many boats are in the image?')
ANSWER3=Vqa(image=RIGHT, question='How many boats are in the image?')
ANSWER4=Eval('{ANSWER0} + {ANSWER1} == 6 and {ANSWER2} + {ANSWER3} == 2')
RESULT=ANSWER4
```

Prediction: False

Gupta et al "Visual Programming: Compositional visual reasoning without training". 2023.

VisProg (visual programming)

Factual Knowledge Object Tagging

IMAGE:

Prediction: IMAGE0

Instruction: Tag the 7 main characters on the TV show Big Bang Theory

Program:

```
OBJ0=FaceDet(image=IMAGE)
LIST0=List(query='main characters on the TV show Big Bang Theory', max=7)
OBJ1=Classify(image=IMAGE, object=OBJ0, categories=LIST0)
IMAGE0=Tag(image=IMAGE, object=OBJ1)
RESULT=IMAGE0
```

Gupta et al "Visual Programming: Compositional visual reasoning without training". 2023.

VisProg (visual programming)

IMAGE:

Prediction: IMAGE0

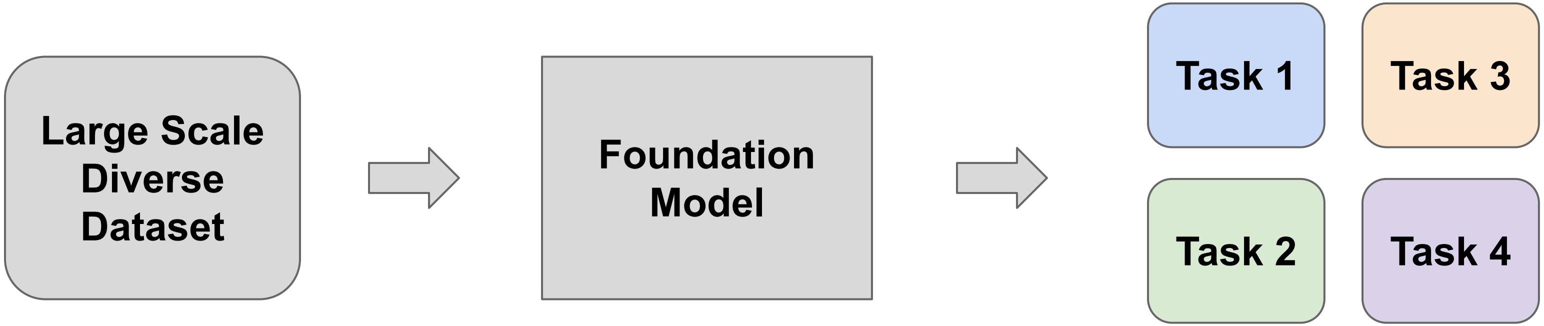
Instruction: Replace desert with lush green grass

Program:

```
OBJ0=Seg(image=IMAGE)
OBJ1=Select(image=IMAGE, object=OBJ0, query='desert', category=None)
IMAGE0=Replace(image=IMAGE, object=OBJ1, prompt='lush green grass')
RESULT=IMAGE0
```

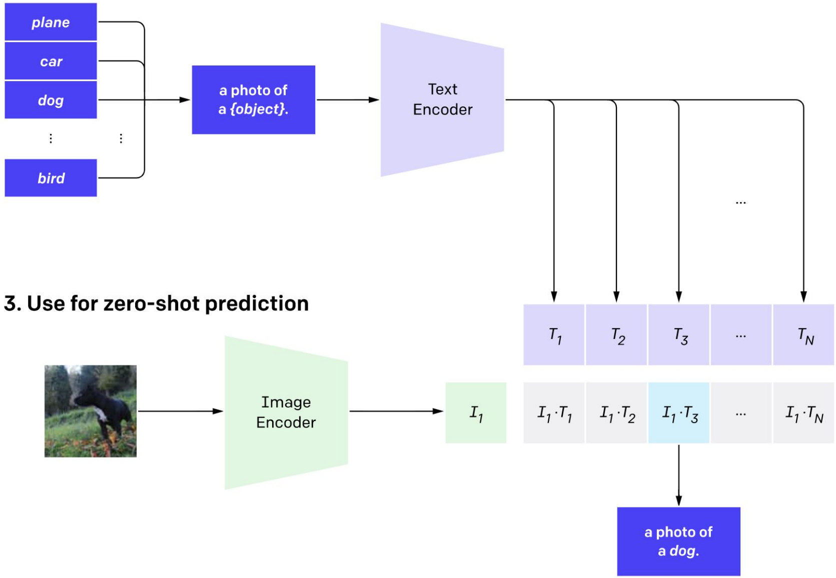
Gupta et al "Visual Programming: Compositional visual reasoning without training". 2023.

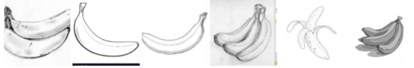
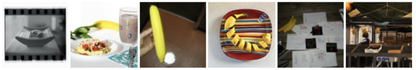
Summary



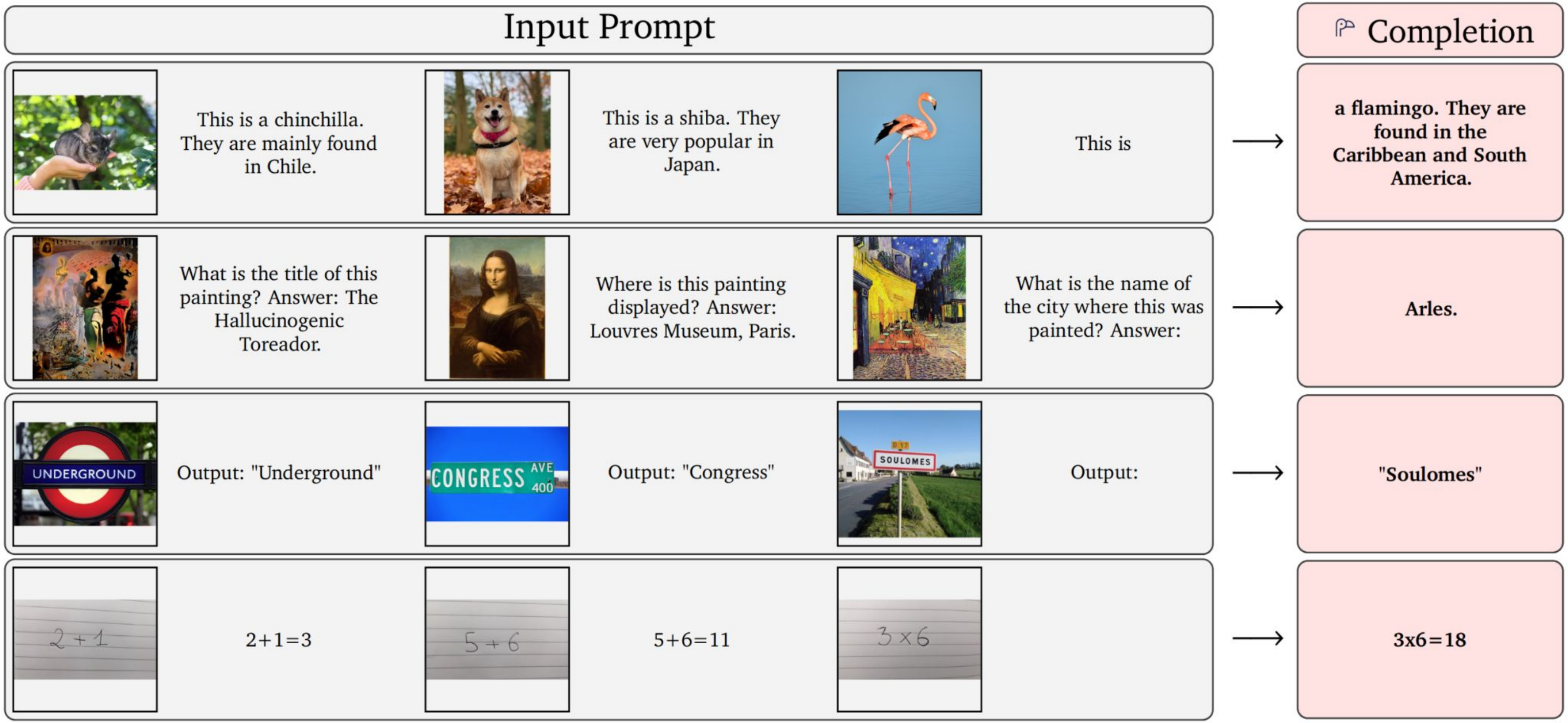
Large Scale Diverse Dataset
Foundation Model
Task 1
Task 3
Task 2
Task 4

Summary



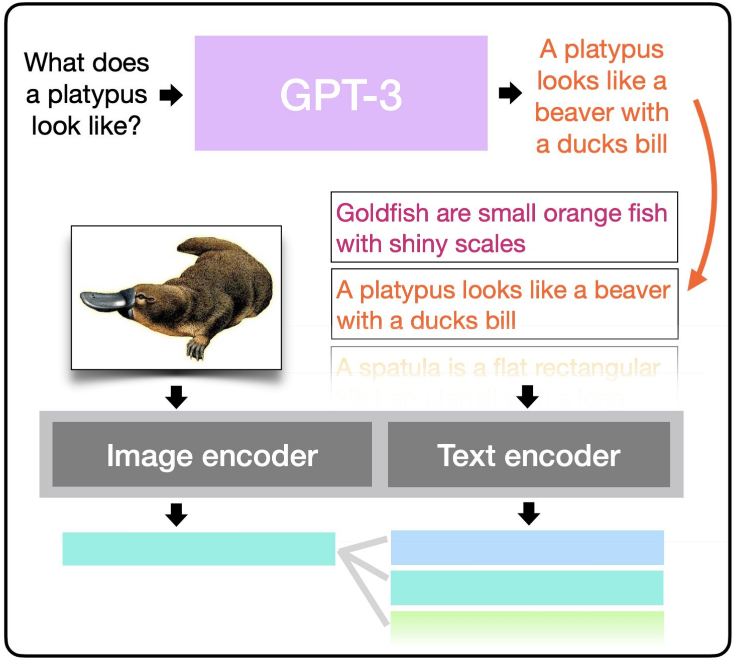
DATASET		IMAGENET RESNET101	CLIP VIT-L
ImageNet		76.2%	76.2%
ImageNet V2		64.3%	70.1%
ImageNet Rendition		37.7%	88.9%
ObjectNet		32.6%	72.3%
ImageNet Sketch		25.2%	60.2%
ImageNet Adversarial		2.7%	77.1%

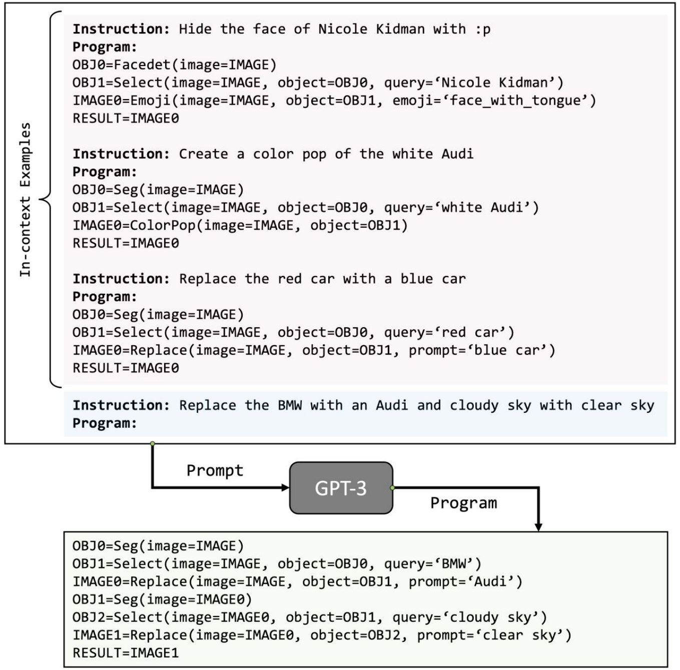
Summary


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 400
Output: "Congress"
SOULOMES
Output:
"Soulomes"
2+1
2+1=3
5+6
5+6=11
3x6
3x6=18

Summary

Summary


What does a platypus look like?
GPT-3
A platypus looks like a beaver with a ducks bill
Goldfish are small orange fish with shiny scales
A platypus looks like a beaver with a ducks bill
A spatula is a flat rectangular
Image encoder
Text encoder


In-context Examples
Instruction: Hide the face of Nicole Kidman with :p
Program:
OBJ0=Facedet(image=IMAGE)
OBJ1=Select(image=IMAGE, object=OBJ0, query='Nicole Kidman')
IMAGE0=Emoji(image=IMAGE, object=OBJ1, emoji='face_with_tongue')
RESULT=IMAGE0
Instruction: Create a color pop of the white Audi
Program:
OBJ0=Seg(image=IMAGE)
OBJ1=Select(image=IMAGE, object=OBJ0, query='white Audi')
IMAGE0=ColorPop(image=IMAGE, object=OBJ1)
RESULT=IMAGE0
Instruction: Replace the red car with a blue car
Program:
OBJ0=Seg(image=IMAGE)
OBJ1=Select(image=IMAGE, object=OBJ0, query='red car')
IMAGE0=Replace(image=IMAGE, object=OBJ1, prompt='blue car')
RESULT=IMAGE0
Instruction: Replace the BMW with an Audi and cloudy sky with clear sky
Program:
Prompt
GPT-3
Program
OBJ0=Seg(image=IMAGE)
OBJ1=Select(image=IMAGE, object=OBJ0, query='BMW')
IMAGE0=Replace(image=IMAGE, object=OBJ1, prompt='Audi')
OBJ1=Seg(image=IMAGE0)
OBJ2=Select(image=IMAGE0, object=OBJ1, query='cloudy sky')
IMAGE1=Replace(image=IMAGE0, object=OBJ2, prompt='clear sky')
RESULT=IMAGE1

Next time: Generative models