Flamingo: a Visual Language Model for Few-Shot Learning

(Alayrac et al. 2022)

Zhangir Azerbayev, CPSC670 Spring 2023

Motivation

- Vision and language have been the main domains where applied deep learning has made a lot of progress
  - Language: Machine translation, BERT, GPT-n
  - Vision: AlexNet, inception, resnet, ViT
- In the long run, we don’t want different modalities to be silos
  - AGI should be multimodal
  - Lots of language problems might benefit from vision (e.g Euclidean geometry)
  - Lots of vision problems might benefit from language (e.g modelling physics)

Andrej Karpathy's CV Challenge (2012)

The picture above is funny.

But for me it is also one of those examples that make me sad about the outlook for AI and for Computer Vision. What would it take for a computer to understand this image as you or I do? I challenge you to think explicitly of all the pieces of knowledge that have to fall in place for it to make sense. Here is my short attempt:

1. *You recognize it is an image of a bunch of people and you understand they are in a hallway*
2. *You recognize that there are 3 mirrors in the scene so some of those people are "fake" replicas from different viewpoints.*
3. …

The Problem with CV in 2022

- Despite rapid progress in the field, most computer vision systems in 2022 still require supervised training
- Vision Transformer (2021):
  - Use transformers for computer vision, train on ImageNet
  - While in NLP, all the training was already self-supervised (BERT etc.) and systems could do few-shot inference (GPT-3)
- Labelling is unscalable
- How do we achieve self-supervised training and few-shot inference for vision?

Self-supervised vision

- “Generative Pretraining from Pixels” (Chen et al. 2020)
  - Pretrain a transformer visual backbone using a BERT objective, then finetune on a downstream task
  - Never really caught on, autoregressive objective not very good for denoising images

Motivation for Flamingo

- Multimodal model: jointly reason across vision and language, to make progress towards tasks like Karpathy's challenge
- Leverage pre-training approaches that have been successful in vision and language
- Few-shot learning for diverse vision/language tasks

The Flamingo Architecture

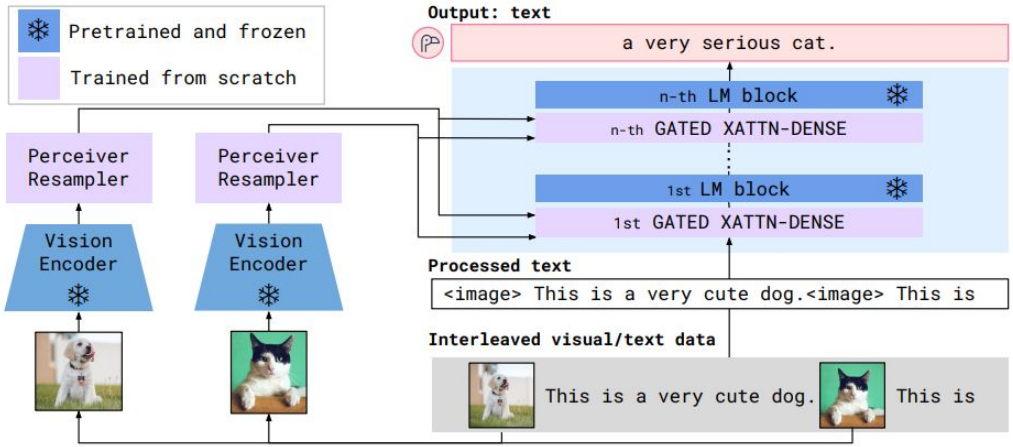
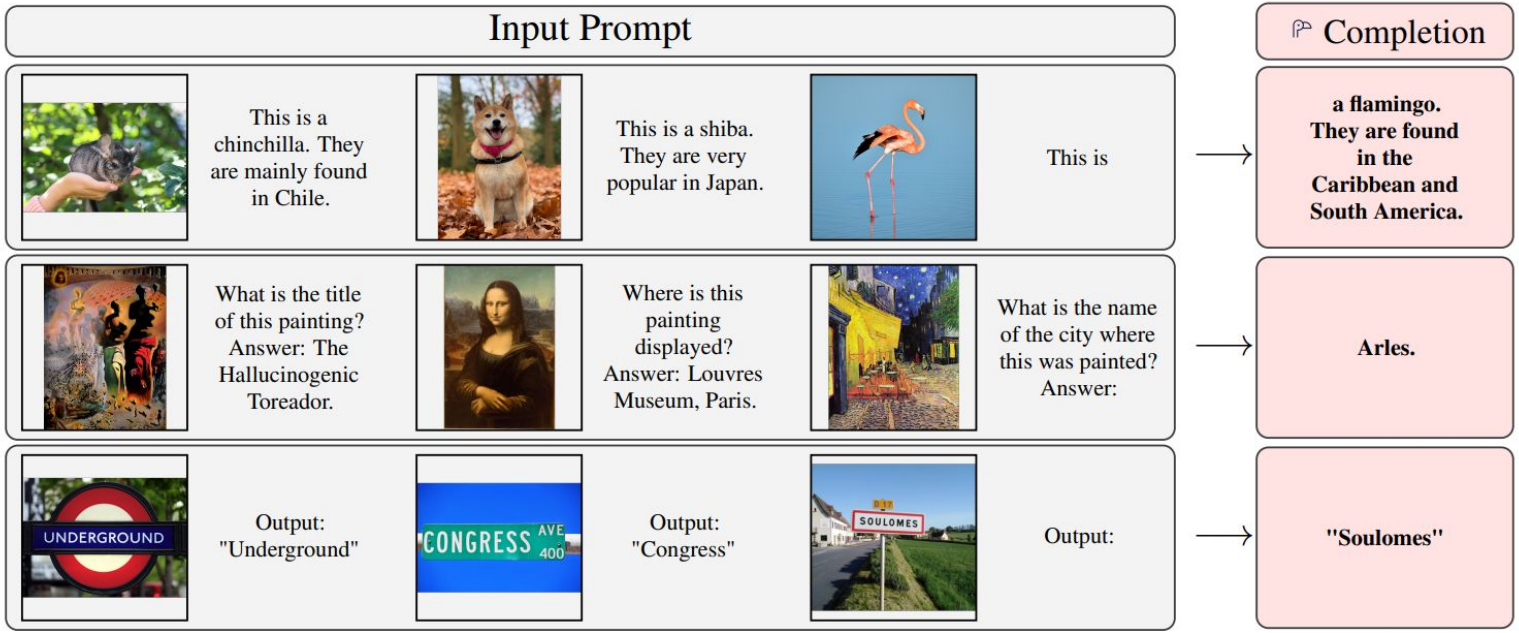


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

- **Frozen** pre-trained language decoder
- **Frozen** pre-trained vision encoder
- Trained multimodal “adapters” between vision and language representations

Accepts input that arbitrarily interleaves text and images

Flamingo capabilities

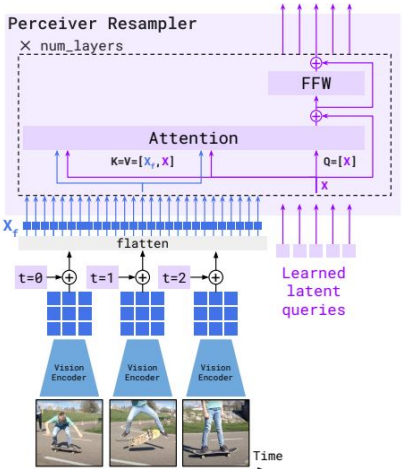


Language and Vision backbones

- Frozen language model: Chinchilla
  - At sizes of 1.4B, 7B, and 70B
  - Flamingo model built on 70B Chinchilla has 80 total parameters
- Frozen vision model:
  - ResNet image encoder from a CLIP model
  - Trained with contrastive CLIP objective on text-image pairs
  - Output: 2D spatial grid of features for images, 3D grid for videos. Flattened to a 1D sequence
- Pre-training of backbones is completely self-supervised, in contrast to most vision papers

Perceiver Resampler

- Eventually, we want to do cross-attention between the vision encoder and the language decoder.
- But the vision encoder may have a very large number of features, especially for videos. What do we do?
- Perceiver Resampler: lightweight attention with variable size input, fixed size output

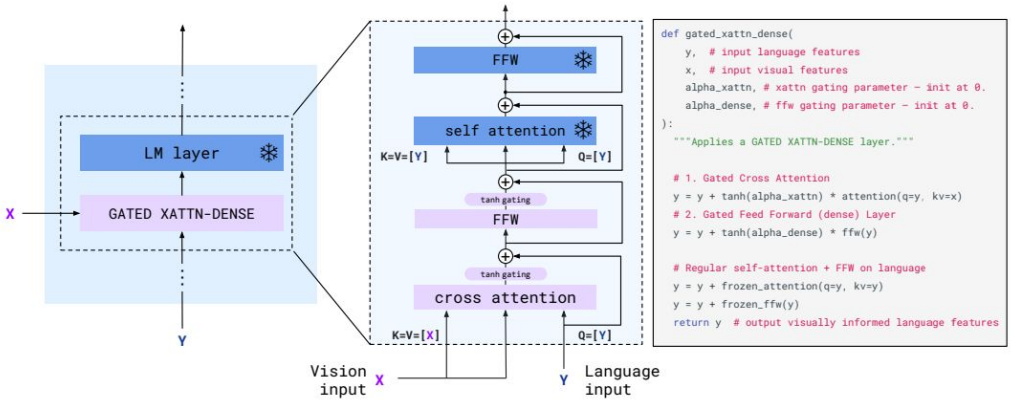



```
def perceiver_resampler(
    x_f,  # The [T, S, d] visual features (T=time, S=space)
    time_embeddings,  # The [T, 1, d] time pos embeddings.
    x,  # R learned latents of shape [R, d]
    num_layers,  # Number of layers
):
  """The Perceiver Resampler model."""

  # Add the time position embeddings and flatten.
  x_f = x_f + time_embeddings
  x_f = flatten(x_f)  # [T, S, d] -> [T * S, d]
  # Apply the Perceiver Resampler layers.
  for i in range(num_layers):
    # Attention.
    x = x + attention_i(q=x, kv=concat([x_f, x]))
    # Feed forward.
    x = x + ffw_i(x)
  return x
```

Vision-language cross-attention

- Use layer called GATED X-ATTN-DENSE
  - Like regular cross attention,
  - Except the output of the multi-head attention layer goes through a tanh
  - This stabilizes the distribution of activations before the frozen self-attention
  - Allows stable training with frozen self-attention



Multi-modal pre-training

Three datasets

- Interleaved image and text
  - MultiModal MassiveWeb (MW3), interleaved image/text web-scrape
- Image + caption
  - 1.8 billion image/video + caption pairs
- Video + caption
  - 27 million examples (short videos, mostly under 20 sec)

Objective is negative log-likelihood over text

Experimental Results

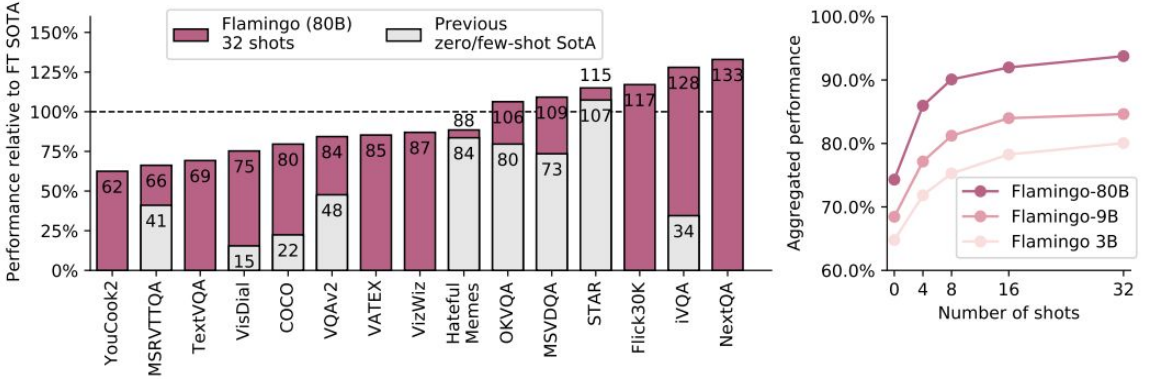


Figure 2: **Flamingo results overview.** *Left*: Our largest model, dubbed *Flamingo*, outperforms state-of-the-art fine-tuned models on 6 of the 16 tasks we consider with no fine-tuning. For the 9 tasks with published few-shot results, *Flamingo* sets the new few-shot state of the art. *Note:* We omit RareAct, our 16th benchmark, as it is a zero-shot benchmark with no available fine-tuned results to compare to. *Right*: Flamingo performance improves with model size and number of shots.

Back to Karpathy's Challenge…

This is a picture of Barack Obama. He is a former president of the United States.

How many people are there in this picture?

There are at least 5 people in this picture.

Where was this picture taken?

It was taken in a school.

How many mirrors are there?

There are at least 2 mirrors in this picture.

What is the person standing on?

The person is standing on a rug.

He is standing on a scale.

I think you are right

What is he doing?

He is looking at the scale.

Where is Obama's foot positioned?

Obama's foot is positioned on the right side of the scale.

What happens as a result?

The scale shows a higher weight.

Is the person on the scale aware of it?

I think he is not aware of it.

Do you think that's why people are laughing?

I think so.

Discussion Questions

- Is gluing together separate vision and text models a hack, or a deep insight?
- There are 70B language parameters and 10B vision/multimodal parameters. Do you think this limits the model's capability, or does vision just require fewer parameters?
- Do you think the standard suite of visual QA benchmarks fully captures the capabilities of the model? What capability do you want to understand that's not captured by the evaluations?