Flamingo 🦩: a Visual Language Model for Few-Shot Learning

Ananya, Mingcheng

Motivation and Introduction

Context of AI Challenges

- Multimodal machine learning seeks to unify the sensory inputs from different modalities
  - Visual and textual
  - Perform tasks that mimic human cognitive abilities
- The challenge lies in rapidly adapting to new tasks with minimal examples.
- To perform few-shot learning, the most widely used paradigm is ***pre-training*** on a large amount of supervised data before ***fine-tuning*** the model on task of interest
  - However, this requires:
    - Many thousands of annotated data points
    - Careful per-task hyperparameter tuning
    - Resource intensive

Introduction to Flamingo

- Context
  - Current AI models struggle with adapting to new tasks efficiently.
  - A significant gap in multimodal research, bridging vision and language.
- Flamingo's Goal:
  - A Visual Language Model (VLM) targeting few-shot learning efficiency across vision and language tasks
  - Bridge pre-trained vision and language models for enhanced multimodal understanding
- Innovation
  - Architectural novelties: Efficient handling of sequences with interleaved visual and textual data.
  - Training on large-scale multimodal web corpora, for robust few-shot learning capabilities.

Flamingo’s Approach to Few Shot Learning

- Architecture
  - Unites pre-trained vision and language models.
  - Processes sequences with interleaved visual-text data.
  - Utilizes 'Perceiver Resampler' for efficient visual information integration.
- Data Handling for Few-Shot Learning
  - Trains on large-scale, diverse web-captured datasets.
  - Merges multiple types of datasets—interleaved images-text, paired image-text, and video-text—to enforce versatile learning
  - Trains to predict text sequences given visual contexts, enabling the model to adapt by example rather than explicit instruction
- Ablation Studies for Enhanced Model Understanding
  - Systematically evaluates the influence of individual model components on overall performance to refine design choices
  - Validates the performance benefits of the Perceiver Resampler and the interleaved data handling strategy.

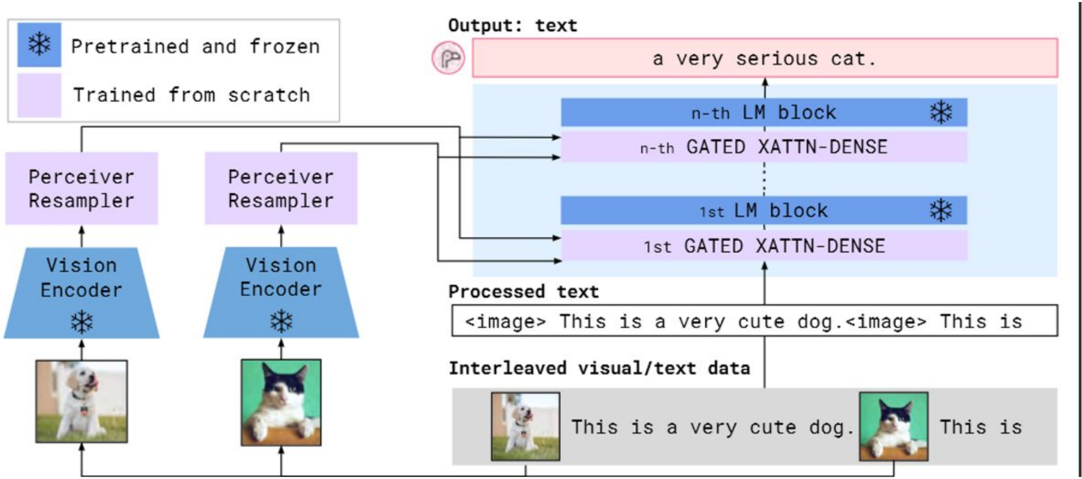
Methods

We propose Flamingo, a family of Visual Language Models that can rapidly adapted to novel tasks using only a handful of annotated examples.

Architecture Overview

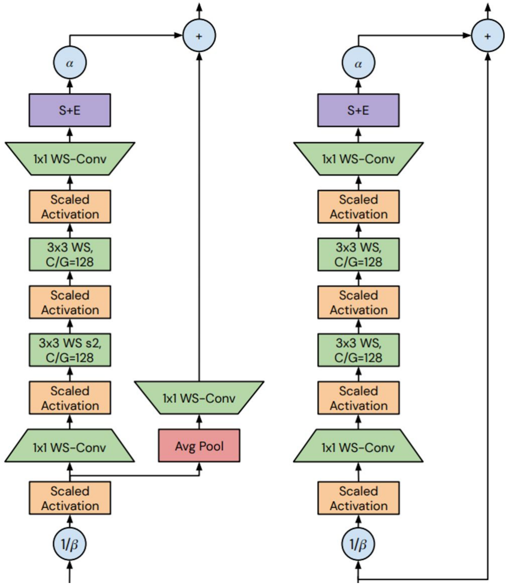
Flamingo models leverage two complementary pre-trained and frozen models:

- A vision model that can perceive visual scenes
- A LLM which performs basic form of reasoning
- Novel architecture components are added to connect these two models



Visual Processing

A pretrained and frozen Normalizer-Free ResNet (NFNet-F6) is used to extract features from raw pixels

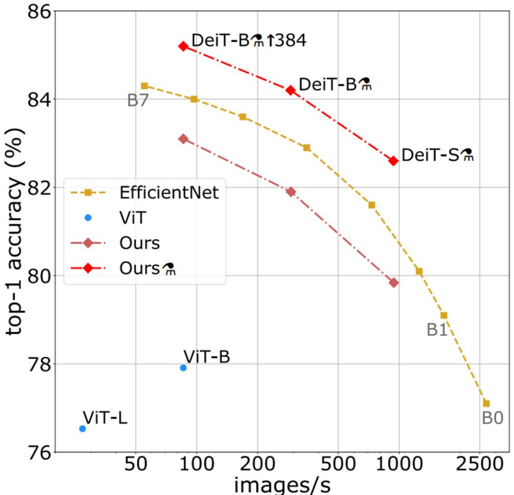


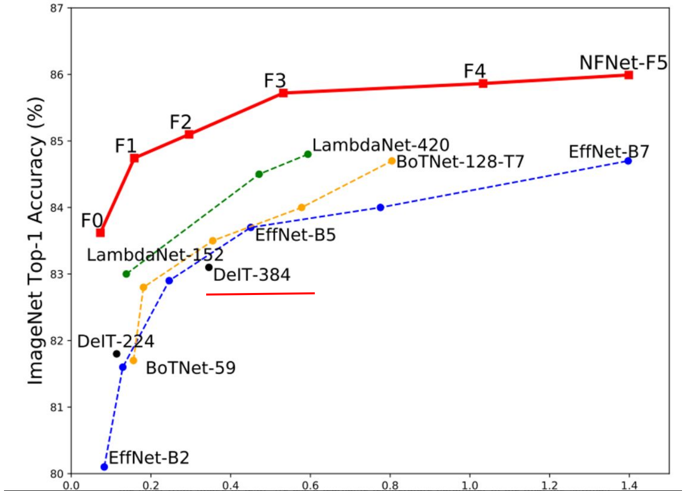
For images inputs, the output of the final stage, a 2D spatial grid of features is flattened to a 1D sequence

Videos inputs are sampled at 1 FPS to obtain a 3D spatio-temporal grid, which is then flattened to 1D before being fed to the Perceiver Resampler

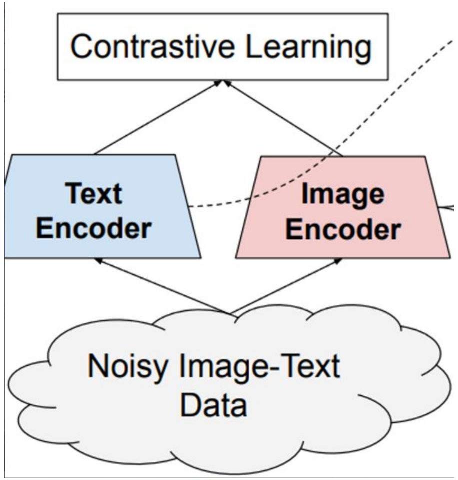
NFNet is a significantly improved class of Normalizer-Free ResNets in terms of accuracies, as batch normalization has many undesirable properties stemming from its dependence on the batch size and interactions between examples.

Despite the significant gap between Convolution and Transformer, the experiment on ImageNet shows that the fine-tuned NFNet can achieve comparable performance with ViT.





The vision encoder is pre-trained using a contrastive objective on datasets of image and text pairs, using the two-term contrastive loss.

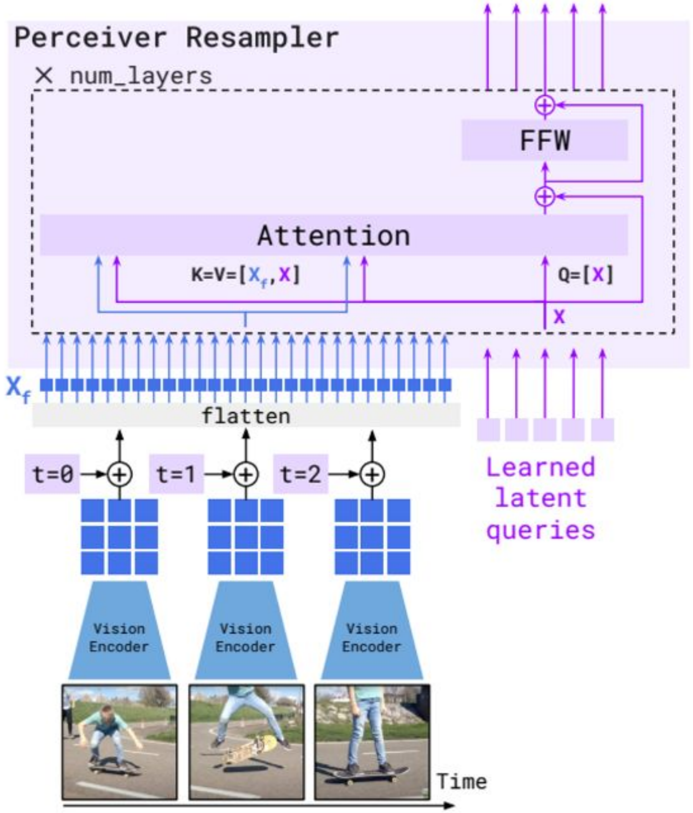


$$L_{contrastive:im2txt} = -\frac{1}{N}\sum_{i}^{N} \log\left(\frac{\exp(V_i^{\intercal} L_i \beta)}{\sum_j^N \exp(V_i^{\intercal} L_j \beta)}\right)$$

$$L_{contrastive:txt2im} = -\frac{1}{N}\sum_{i}^{N} \log\left(\frac{\exp(L_i^{\intercal} V_i \beta)}{\sum_j^N \exp(L_i^{\intercal} V_j \beta)}\right)$$

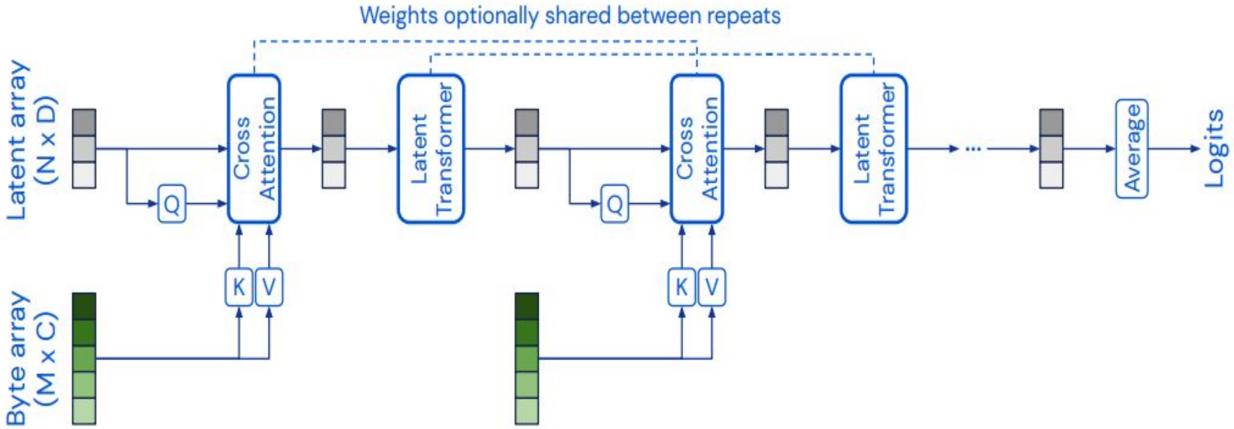
Perceiver Resampler

Perceiver resampler receives spatio-temporal features from Vision Encoder and outputs **a fixed number of visual tokens** regardless of input image resolution, thus reducing the computational complexity of vision-text cross attention.



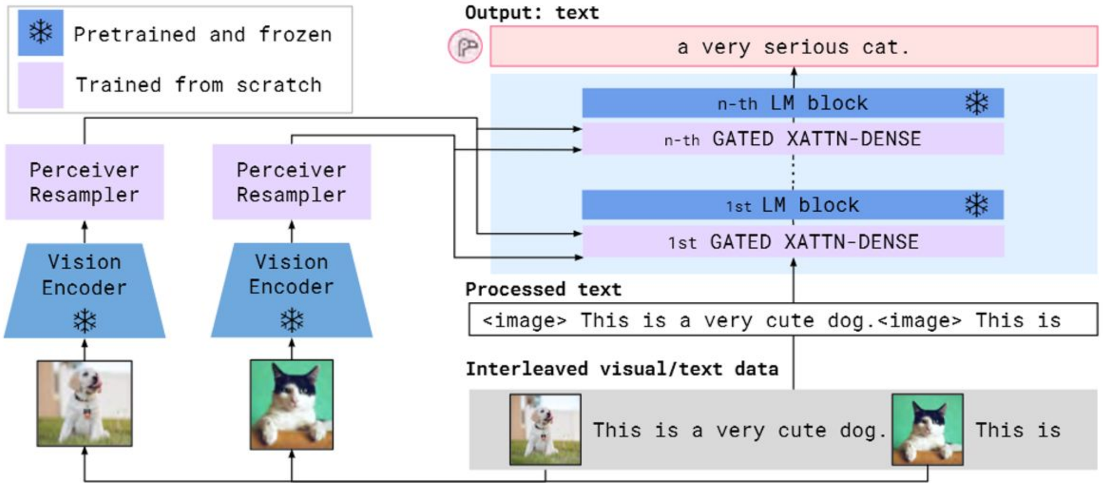
The Perceiver iteratively attends to the input byte array by alternating cross-attention and latent self-attention blocks.

The core idea is to introduce a small set of latent units that **forms an attention bottleneck through which the inputs must pass**, thus eliminate the quadratic scaling problem.



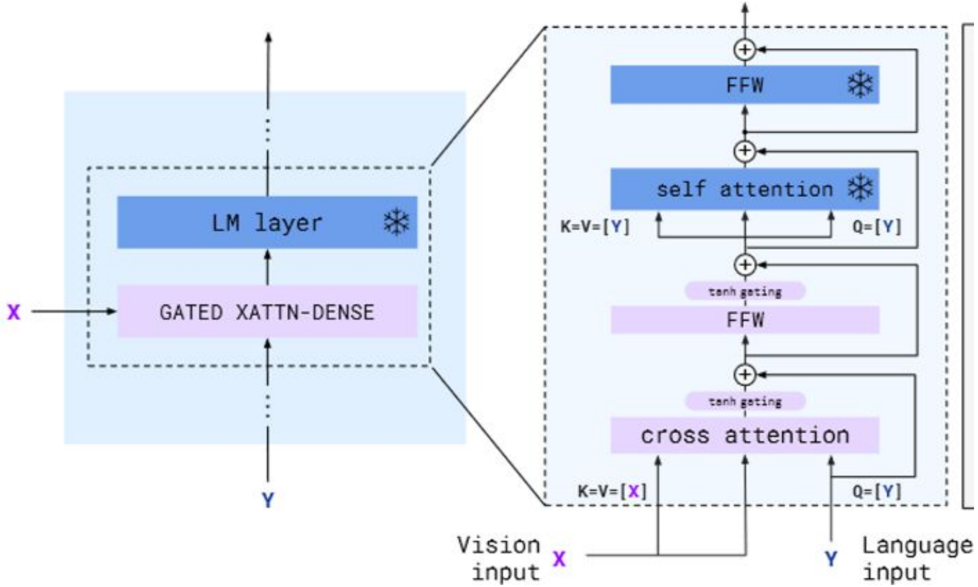
Conditioning Frozen Language Models

The visual tokens from perceiver resampler are then used to condition the frozen LLM using freshly initialized cross-attention layers that are interleaved between the pretrained LLM layers so that text generation can be performed by Transformer decoder.



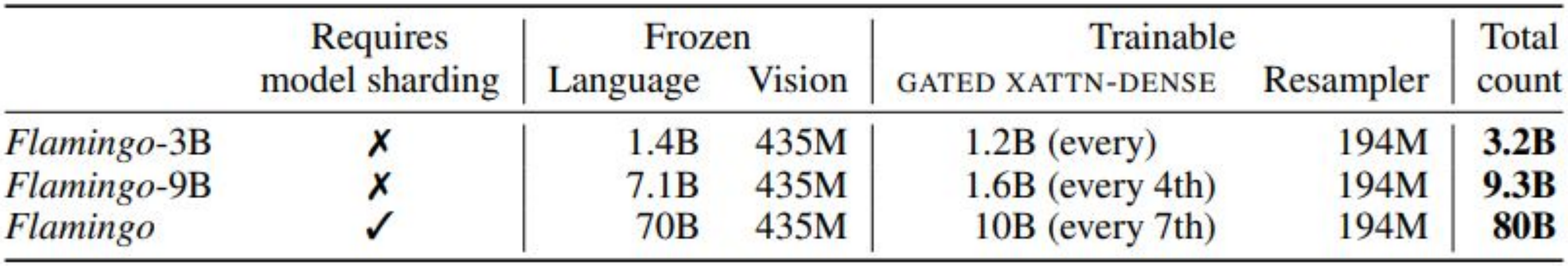
We **freeze** the pre-trained LLM blocks, and insert gated cross-attention dense blocks between the original layers, **trained from scratch**.

To ensure the conditioned model yields the same results as the original language model, we multiply the output of a newly added layer by tanh(ɑ), where ɑ is a layer-specific learnable scalar initialized to 0. Such gating mechanism improves training stability and final performance (which will be shown in ablation study)



We perform experiments across three model sizes, building on the 1.4B, 7B and 70B parameter Chinchilla models; calling them respectively Flamingo-3B, Flamingo-9B and Flamingo-80B. We refer to the last as *Flamingo* throughout the paper.

While increasing the parameter count of frozen LLM and trained vision-text gated cross attention dense modules, we maintain a **fixed-size** frozen token vision encoder and trainable perceiver resampler across different models.



	Requires	Frozen		Trainable		Total
	model sharding	Language	Vision	GATED XATTN-DENSE	Resampler	count
Flamingo-3B	✗	1.4B	435M	1.2B (every)	194M	**3.2B**
Flamingo-9B	✗	7.1B	435M	1.6B (every 4th)	194M	**9.3B**
Flamingo	✓	70B	435M	10B (every 7th)	194M	**80B**

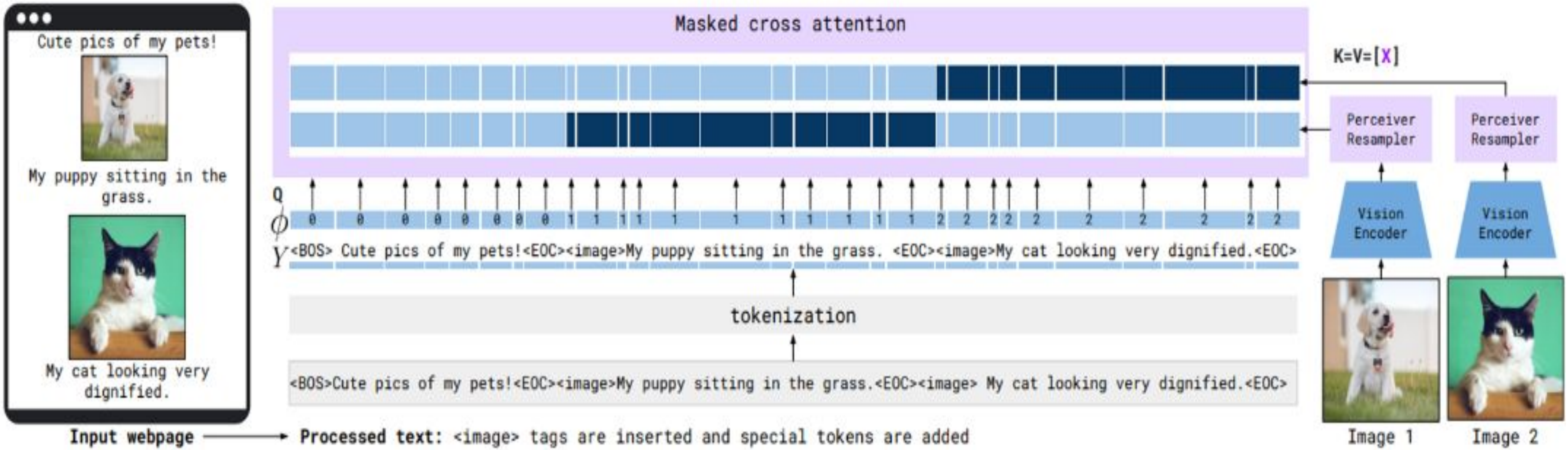
Multi-Visual Input Support

Flamingo models the likelihood of text y conditioned on interleaved images or videos x as

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

At a given text token, the model attends to the visual tokens of image that appeared **just before** it. This single-image cross attention scheme importantly allows the model to seamlessly generalize to any number of visual inputs, regardless of how many are used during training.

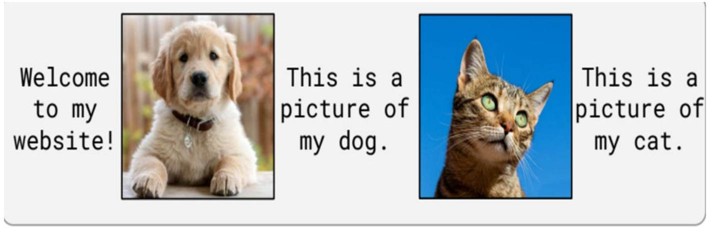
However, the dependency on all previous images remains via self-attention in the LLM.


Cute pics of my pets!
My puppy sitting in the grass.
My cat looking very dignified.
Input webpage
Masked cross attention
K=V=[X]
Perceiver Resampler
Perceiver Resampler
Vision Encoder
Vision Encoder
Q
ϕ
0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2
Y <BOS> Cute pics of my pets!<EOC><image>My puppy sitting in the grass. <EOC><image>My cat looking very dignified.<EOC>
tokenization
<BOS>Cute pics of my pets!<EOC><image>My puppy sitting in the grass.<EOC><image> My cat looking very dignified.<EOC>
Processed text: <image> tags are inserted and special tokens are added
Image 1
Image 2

Training on a Mixture of Vision and Language Datasets

We train Flamingo models on a mixture of three kinds of dataset, all scraped from the web without any annotation:

1. Multimodal Massive Web(M3W) dataset
2. Pairs of image and text: ALIGN, LTIP
3. Pairs of video and text: VTP



Multi-Modal Massive Web (M3W) dataset
[N>1, T=1, H, W, C]

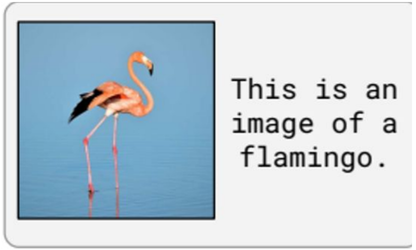
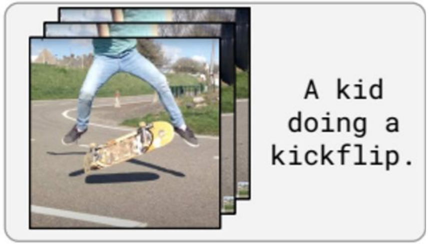


Image-Text Pairs dataset
[N=1, T=1, H, W, C]



Video-Text Pairs dataset
[N=1, T>1, H, W, C]

Multi-objective training and optimisation strategy: We train the model by minimizing a weighted sum of per-dataset expected negative log-likelihood given visual inputs

$$\sum_{m=1}^{M} \lambda_m \cdot \mathbb{E}_{(x,y)\sim\mathcal{D}_m}\left[-\sum_{\ell=1}^{L} \log p(y_\ell | y_{<\ell}, x_{\leq\ell})\right],$$

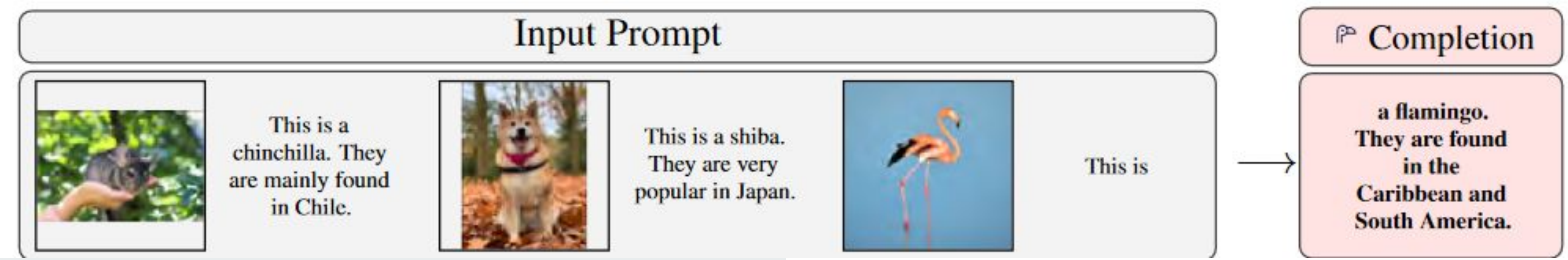
Task Adaption with Few-Shot In-Context Learning

We evaluate the ability of our models to rapidly adapt to new tasks using **in-context learning**, by interleaving supporting example pairs in form of (image, text) or (video, text), followed by the query visual input.

We perform **open-ended** evaluations (visual question answering, captioning task) as well as **close-ended** evaluations (multiple choice visual question answering).

We also explore **zero-shot** generalization by prompting the model with two text-only examples from the task.

Experiments



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

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

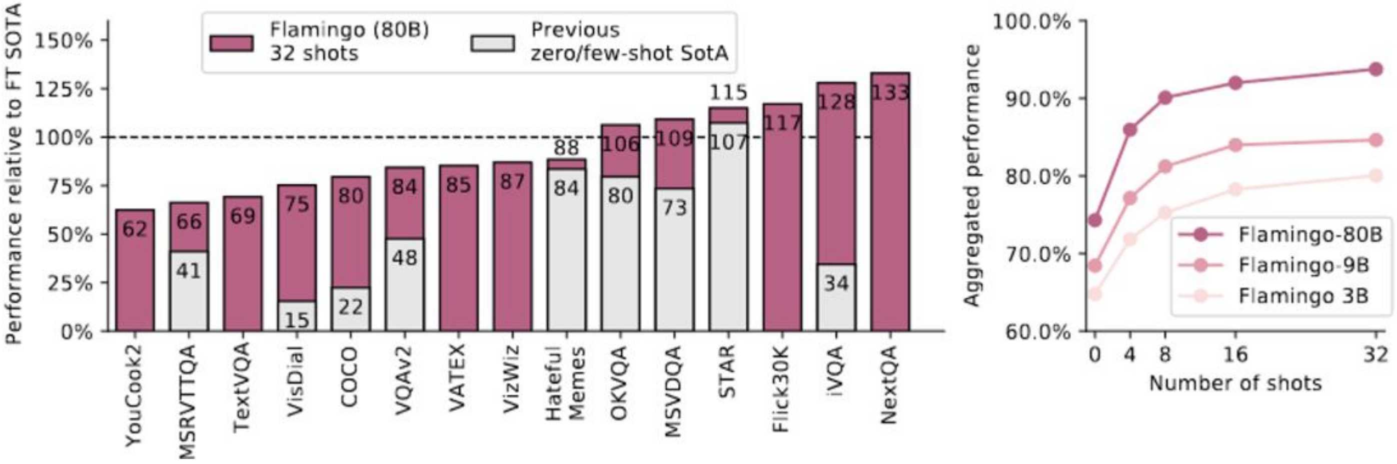
This is

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

Few Shot Learning on Vision-Language Tasks

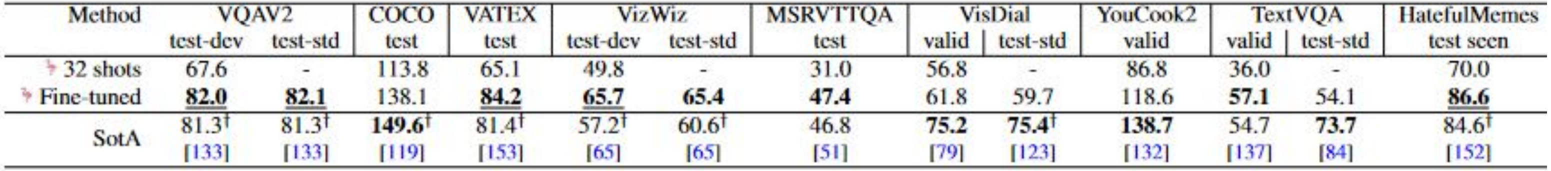
- Evaluated across 16 diverse benchmarks for vision-language tasks.
- Demonstrates the ability to efficiently adapt to new tasks with a minimal number of examples, sometimes as few as four, significantly reducing the data requirement for high performance.
- Outperforms zero-shot and few-shot state-of-the-art (SotA) methods, with significant gains in benchmarks like VQAv2, COCO, and TextVQA. For instance, in VQAv2, it achieves a new few-shot SotA performance.
    - Achieving 57.8% accuracy in VQAv2 with only 32 examples

The larger the model, the better the few-shot performance. The performance also improves with number of shots



Fine-tuning Flamingo as a Pretrained Vision-Language Model

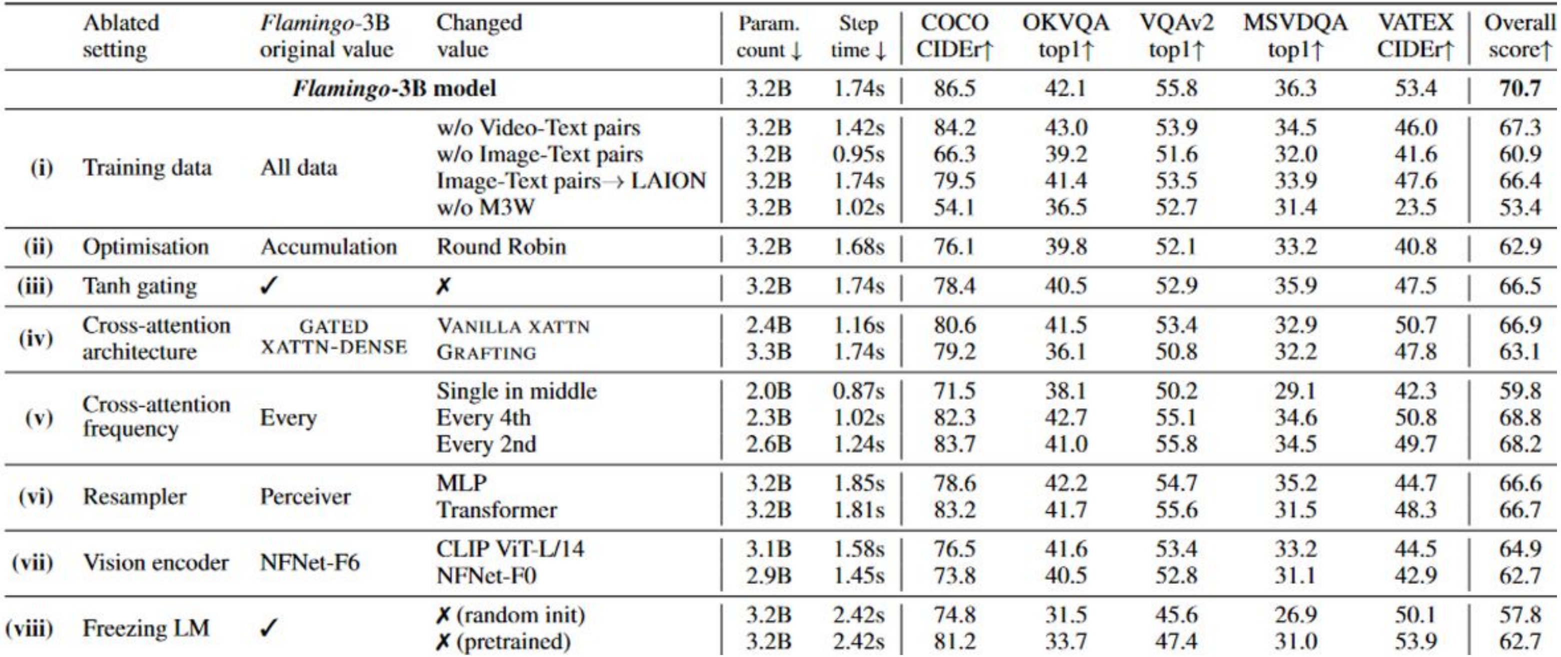
- Post fine-tuning, Flamingo sets new SotA on five additional challenging benchmarks, such as VQAv2 and COCO, indicating substantial improvements over pre fine-tuning performances
- The fine-tuning process employs a higher annotation budget, contrasting the few-shot learning scenario, thereby showcasing the model's scalability and effectiveness in leveraging larger datasets for performance enhancements



Method	VQAV2		COCO	VATEX	VizWiz		MSRVTTQA	VisDial		YouCook2	TextVQA		HatefulMemes
	test-dev	test-std	test	test	test-dev	test-std	test	valid	test-std	valid	valid	test-std	test seen
🦩 32 shots	67.6	-	113.8	65.1	49.8	-	31.0	56.8	-	86.8	36.0	-	70.0
🦩 Fine-tuned	**82.0**	**82.1**	138.1	**84.2**	**65.7**	**65.4**	**47.4**	61.8	59.7	118.6	**57.1**	54.1	**86.6**
SotA	81.3†	81.3†	**149.6**†	81.4†	57.2†	60.6†	46.8	**75.2**	**75.4**†	**138.7**	54.7	**73.7**	84.6†
	[133]	[133]	[119]	[153]	[65]	[65]	[51]	[79]	[123]	[132]	[137]	[84]	[152]

Ablation Studies for Enhanced Model Understanding

- The ablation studies reveal significant findings, such as the importance of the Perceiver Resampler for efficient visual information processing and the effectiveness of training on a diverse web-captured dataset mix.
- Validates the architectural design choices made in Flamingo, with empirical evidence supporting the inclusion of specific components and strategies to achieve high performance in vision-language tasks.



	Ablated	*Flamingo*-3B	Changed	Param.	Step	COCO	OKVQA	VQAv2	MSVDQA	VATEX	Overall
	setting	original value	value	count ↓	time ↓	CIDEr↑	top1↑	top1↑	top1↑	CIDEr↑	score↑
		Flamingo-3B model		3.2B	1.74s	86.5	42.1	55.8	36.3	53.4	**70.7**
(i)	Training data	All data	w/o Video-Text pairs	3.2B	1.42s	84.2	43.0	53.9	34.5	46.0	67.3
			w/o Image-Text pairs	3.2B	0.95s	66.3	39.2	51.6	32.0	41.6	60.9
			Image-Text pairs→ LAION	3.2B	1.74s	79.5	41.4	53.5	33.9	47.6	66.4
			w/o M3W	3.2B	1.02s	54.1	36.5	52.7	31.4	23.5	53.4
(ii)	Optimisation	Accumulation	Round Robin	3.2B	1.68s	76.1	39.8	52.1	33.2	40.8	62.9
(iii)	Tanh gating	✓	✗	3.2B	1.74s	78.4	40.5	52.9	35.9	47.5	66.5
(iv)	Cross-attention architecture	GATED XATTN-DENSE	VANILLA XATTN	2.4B	1.16s	80.6	41.5	53.4	32.9	50.7	66.9
			GRAFTING	3.3B	1.74s	79.2	36.1	50.8	32.2	47.8	63.1
(v)	Cross-attention frequency	Every	Single in middle	2.0B	0.87s	71.5	38.1	50.2	29.1	42.3	59.8
			Every 4th	2.3B	1.02s	82.3	42.7	55.1	34.6	50.8	68.8
			Every 2nd	2.6B	1.24s	83.7	41.0	55.8	34.5	49.7	68.2
(vi)	Resampler	Perceiver	MLP	3.2B	1.85s	78.6	42.2	54.7	35.2	44.7	66.6
			Transformer	3.2B	1.81s	83.2	41.7	55.6	31.5	48.3	66.7
(vii)	Vision encoder	NFNet-F6	CLIP ViT-L/14	3.1B	1.58s	76.5	41.6	53.4	33.2	44.5	64.9
			NFNet-F0	2.9B	1.45s	73.8	40.5	52.8	31.1	42.9	62.7
(viii)	Freezing LM	✓	✗ (random init)	3.2B	2.42s	74.8	31.5	45.6	26.9	50.1	57.8
			✗ (pretrained)	3.2B	2.42s	81.2	33.7	47.4	31.0	53.9	62.7

Remarks and Conclusion

Key Architectural Innovations

1. Bridge powerful visual-only and language-only models
2. Handle sequences of arbitrarily interleaved visual and textual (multimodal) data; Seamlessly ingest images or videos as inputs
3. Can perform various multimodal tasks (such as captioning, visual dialogue and visual question-answering) from only a few input/output examples
4. Set a new state-of-the-art in few shot learning on a wide array of 16 multimodal tasks, and it surpasses fine tuned state-of-the-art model in 6 out of 16 tasks

Limitations

1. LLM play a role in occasional hallucination and ungroundedness guesses.
2. Classification performance lags behind STOA contrastive models as the model handles a wider range of tasks such as open-ended ones.
3. In-context learning is highly sensitive and depend on the characteristics of the application at hand.
4. Exposed to the risk such as outputting offensive language, propagating social biases and stereotypes, leaking private information.