AU24 CSE 5539 Presentation

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

Jean-Baptiste Alayrac*, Jeff Donahue*, Pauline Luc*, Antoine Miech* et al.

NEURAL INFORMATION PROCESSING SYSTEMS

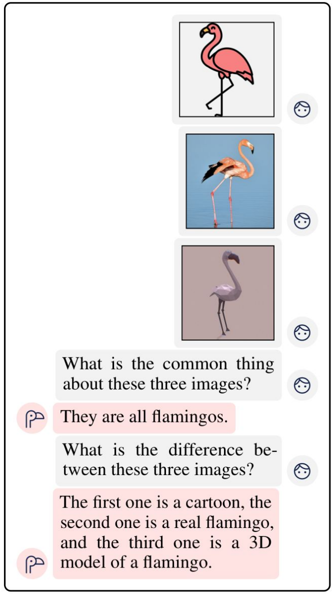
NeurIPS 2022

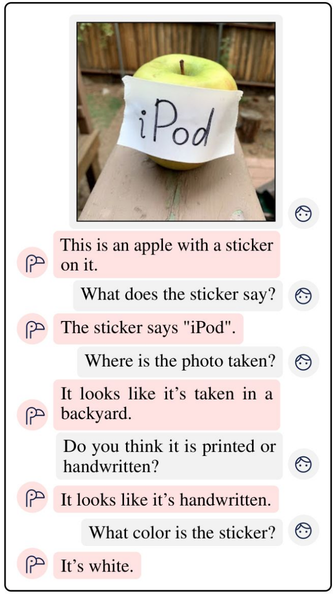
Paper: arxiv.org/abs/2204.14198

✍️ **Sriram Sai Ganesh**

Introduction

- Extending few-shot generalization to multimodality.
- Foundational Visual Language Model (VLM) –
  - 🦩 *Flamingo-80B, 9B, 3B.*
  - Classification, captioning, VQA.
- New SoTA on various benchmarks
  - Sometimes with 1000x lesser data*.

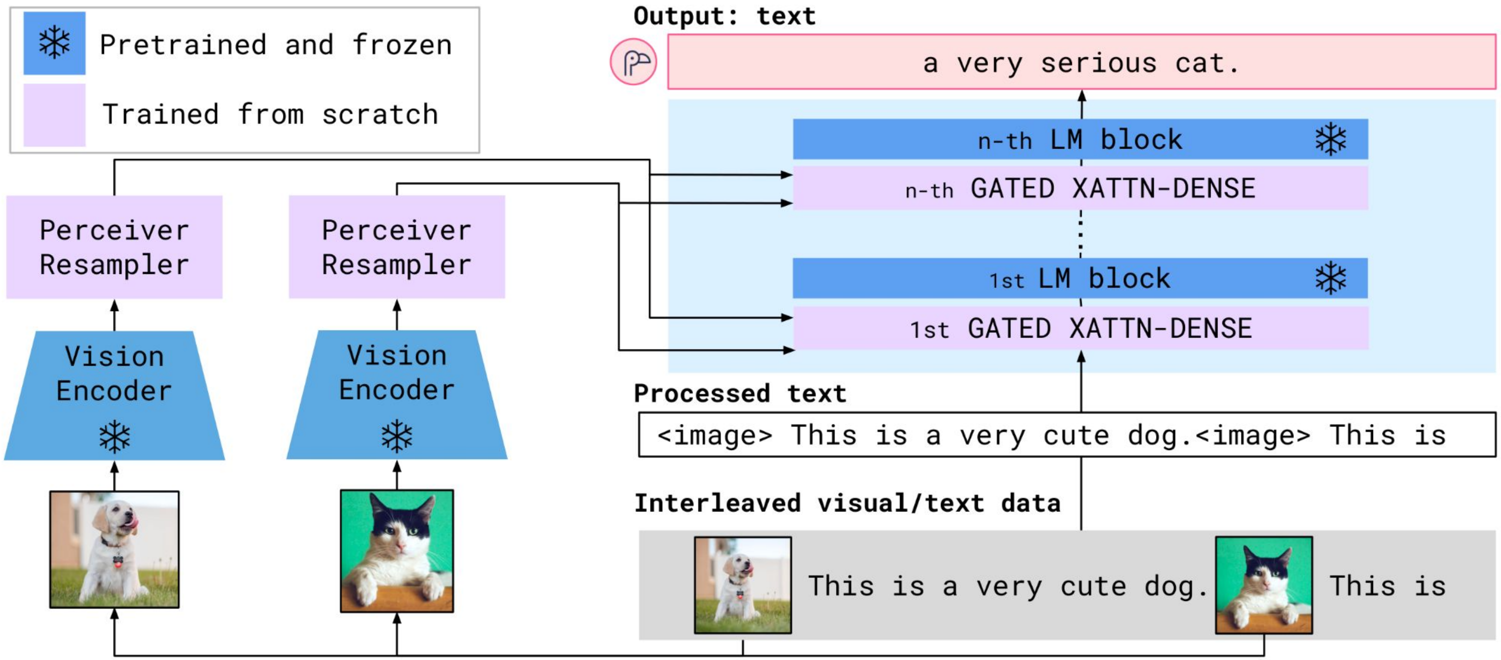




✍️ Sriram Sai Ganesh

Approach

Overview

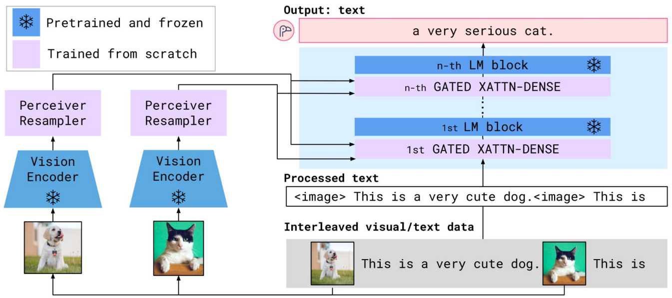


Approach

Overview

Inference steps:

1. **Perceiver Resampler:** Receive spatio-temporal features from an encoder & output visual tokens.
2. **Language Model:** Conditioned with interleaved cross-attention layers.



Likelihood of text *y* on preceding tokens *x* as $p(y|x) = \prod_{\ell=1}^{L} p(y_\ell | y_{<\ell}, x_{\leq\ell})$

where *L* = number of tokens, *p* = *Flamingo*.

✍️ **Sriram Sai Ganesh**

Approach

Visual Processing

- **Vision Encoder:**
  - F6 NFNet pretrained using CLIP objective.
  - Videos – encoded 1FPS, temporally flattened before being processed.
- **Perceiver Resampler:**
  - Receives a *variable* number of visual features, *fixed* number (64) visual outputs.
  - ↓ Complexity of vision/text cross-attention.

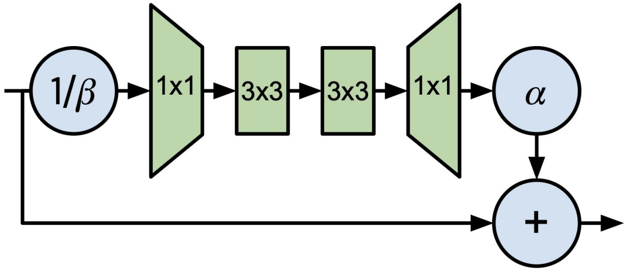


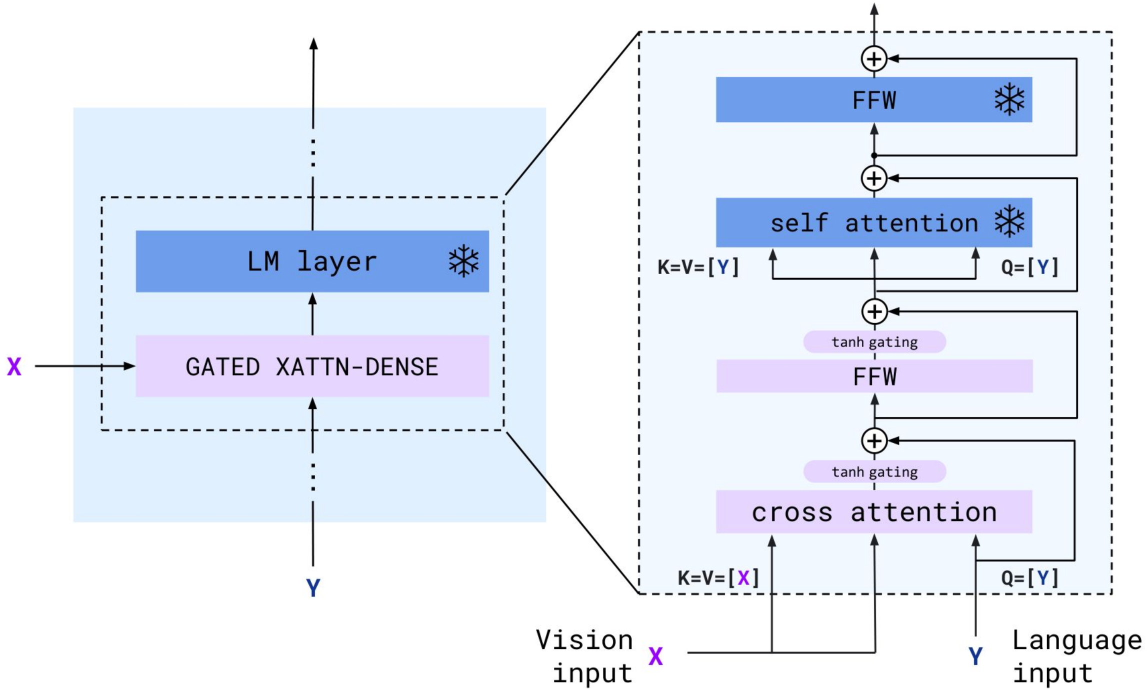
Table 1. NFNet family depths, drop rates, and input resolutions.

Variant	Depth	Dropout	Train	Test
F0	[1, 2, 6, 3]	0.2	192px	256px
F1	[2, 4, 12, 6]	0.3	224px	320px
F2	[3, 6, 18, 9]	0.4	256px	352px
F3	[4, 8, 24, 12]	0.4	320px	416px
F4	[5, 10, 30, 15]	0.5	384px	512px
F5	[6, 12, 36, 18]	0.5	416px	544px
F6	[7, 14, 42, 21]	0.5	448px	576px

NFNet: arxiv.org/abs/2102.06171

Approach

LM + Visual Representations

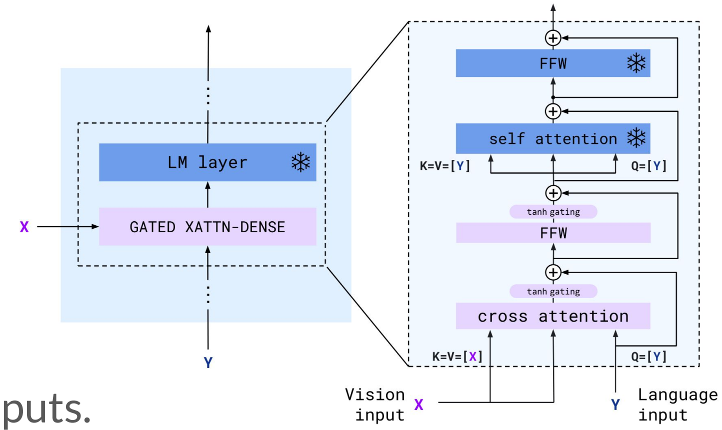


✍️ Sriram Sai Ganesh

Approach

LM + Visual Representations

- **Text generation:** Performed by a transformer decoder, conditioned on text & Perceiver Resampler outputs.
- Interleave frozen pretrained LM blocks with trained **gated cross-attention dense blocks.**
  - K & V from vision features; Q from LM inputs.
  - Enables selective/dynamic attention to inputs from different modalities.



✍️ Sriram Sai Ganesh

Approach

Multi-visual Input

- **Masking cross-attention:** Model directly attends to only *immediately preceding* image/videos.
- Dependency on previously seen visual inputs – LM self-attention.
- 5-shot training for interleaved datasets.
- Generalizes well, up to ~32 shot during testing.

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

Approach

Multimodal Training

Trained on web-scraped datasets of 3 kinds:

- **MultiModalMassiveWeb (M3W)** – 43M webpages. Trained up to first 5 images out of 256 random tokens on a document.
- **Visual/Text pairs** – ALIGN dataset of 1.8B images. Augmented with Long Text Image Pairs (LTIP – 312M) and Video & Text Pairs (VTP – 27M).
- Minimize a weighted **per-dataset negative log likelihood** of text:

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

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

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

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

✍️ Sriram Sai Ganesh

Experiments

Ablations

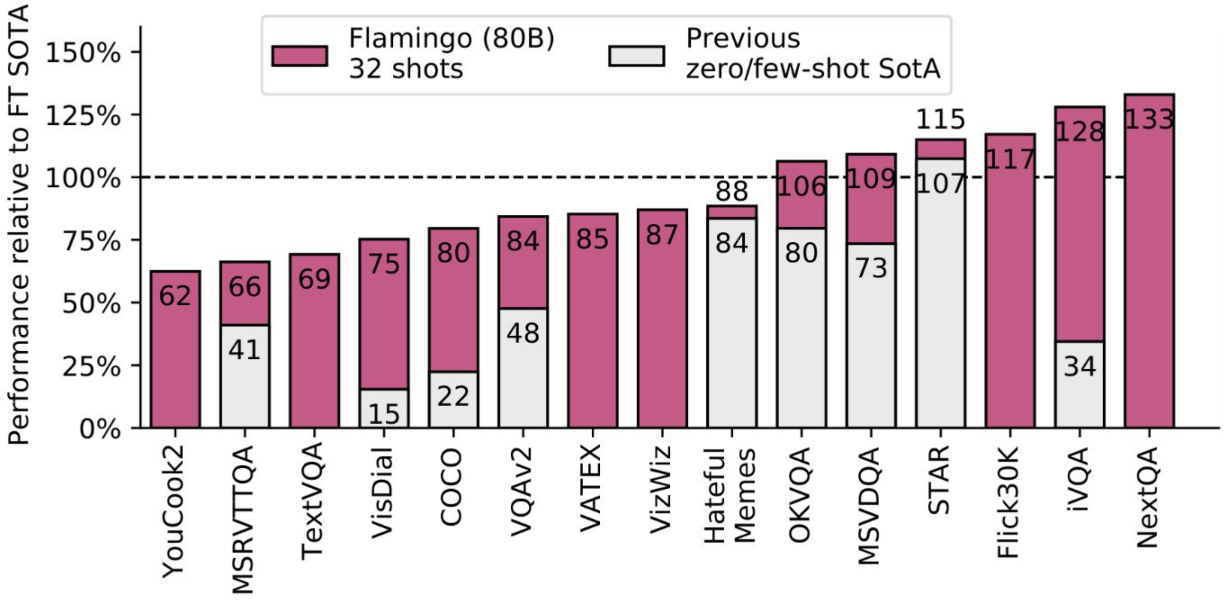
- Comparisons with Image **(I)** and Video **(V)** SoTA. 🦩 > SoTA on 7 or 12* of 16 datasets.

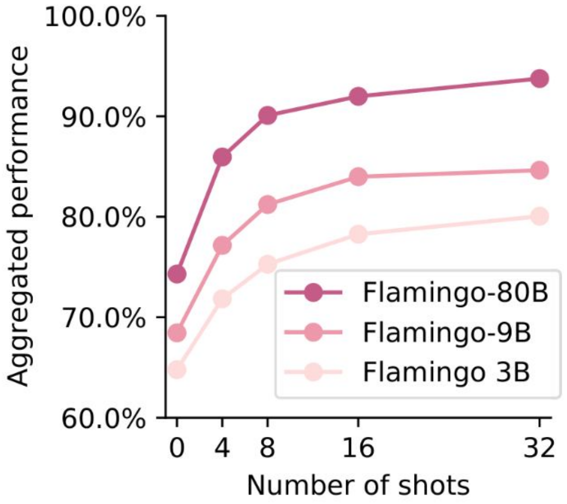
	Ablated setting	*Flamingo*-3B original value	Changed value	Param. count ↓	Step time ↓	COCO CIDEr↑	OKVQA top1↑	VQAv2 top1↑	MSVDQA top1↑	VATEX CIDEr↑	Overall score↑
		***Flamingo*-3B model**		3.2B	1.74s	86.5	42.1	55.8	36.3	53.4	**70.7**
(i)	Training data	All data	w/o Video-Text pairs	3.2B	1.42s	84.2	43.0	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

Experiments

Overview

- Results reported with ***in-context learning***:





✍️ Sriram Sai Ganesh

Experiments

FT-SoTA Comparison

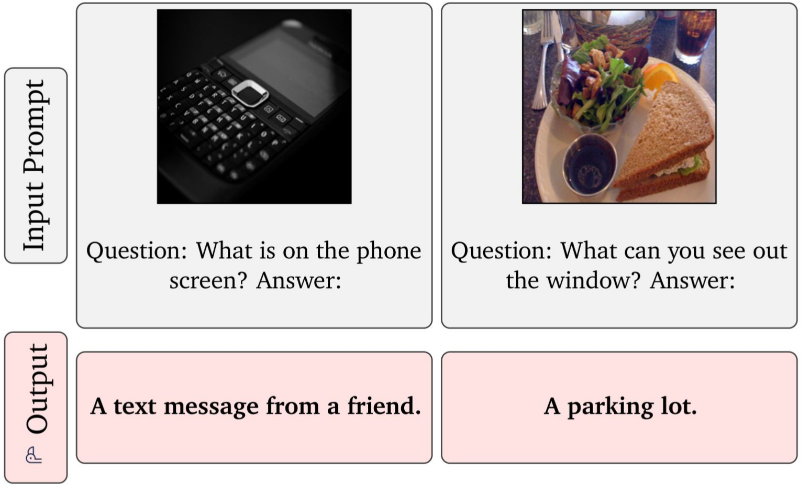
- Comparisons with Image **(I)** and Video **(V)** SoTA. 🦩 > SoTA on 7 or 12* of 16 datasets.

Method	FT	Shot	OKVQA (I)	VQAv2 (I)	COCO (I)	MSVDQA (V)	VATEX (V)	VizWiz (I)	Flick30K (I)	MSRVTTQA (V)	iVQA (V)	YouCook2 (V)	STAR (V)	VisDial (I)	TextVQA (I)	NextQA (I)	HatefulMemes (I)	RareAct (V)
Zero/Few shot SOTA	✗	(X)	[34] 43.3 (16)	[114] 38.2 (4)	[124] 32.2 (0)	[58] 35.2 (0)	-	-	-	[58] 19.2 (0)	[135] 12.2 (0)	-	[143] 39.4 (0)	[79] 11.6 (0)	-	-	[85] 66.1 (0)	[85] 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	-
	✗	32	45.9	57.1	99.0	42.6	59.2	45.5	71.2	25.6	37.7	76.7	41.6	47.3	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	-
	✗	32	51.0	60.4	106.3	47.2	57.4	44.0	72.8	29.4	40.7	77.3	41.2	50.4	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	-
	✗	32	**57.8**	**67.6**	**113.8**	**52.3**	**65.1**	**49.8**	**75.4**	**31.0**	**45.3**	**86.8**	42.2	**55.6**	**37.9**	**33.5**	**70.0**	-
Pretrained FT SOTA	✔	(X)	54.4 [34] (10K)	80.2 [140] (444K)	143.3 [124] (500K)	47.9 [28] (27K)	76.3 [153] (500K)	57.2 [65] (20K)	67.4 [150] (30K)	46.8 [51] (130K)	35.4 [135] (6K)	138.7 [132] (10K)	36.7 [128] (46K)	75.2 [79] (123K)	54.7 [137] (20K)	25.2 [129] (38K)	79.1 [62] (9K)	-

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†

Limitations & Conclusion

- Inherits weaknesses of LMs
  - Hallucination/random guessing.
- Classification performance lags behind contrastive approaches
  - Not optimized for text-image retrieval
- Few-shot inference has several advantages, but is sensitive to in-context examples.



Thank you!

Questions?

Flamingo: a Visual Language Model for Few-Shot Learning

Jean-Baptiste Alayrac*, Jeff Donahue*, Pauline Luc*, Antoine Miech* et al.

Paper: arxiv.org/abs/2204.14198