



Flamingo: a Visual Language Model for Few-Shot Learning

Сергей Лоптев

Рецензия

Авторы

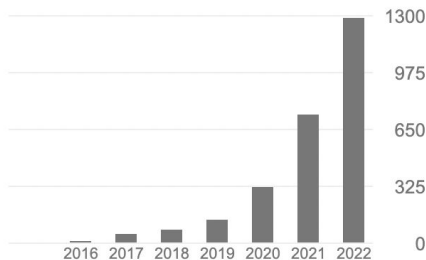


Jean-Baptiste Alayrac

- DeepMind, London
- Coавтор Perceiver IO

Cited by

	All	Since 2017
Citations	2636	2609
h-index	22	22
i10-index	26	26

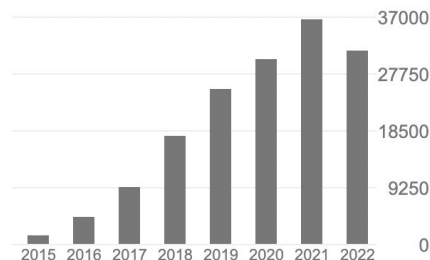


Karen Simonyan

- Co-Founder and Chief Scientist at Inflection AI
- Coавтор NFNet, Chinchilla

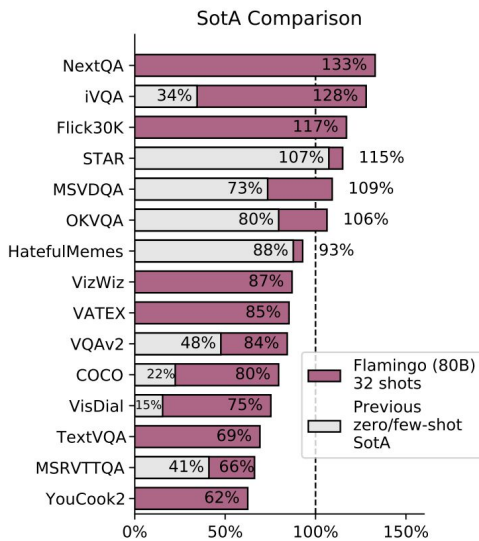
Cited by

	All	Since 2017
Citations	159278	150981
h-index	52	50
i10-index	67	64



+ Успех модели

- Few-shot Фламинго бьёт 6 из 16 анализируемых бенчмарков, становясь в них state-of-the-art моделью.
- Дообученная Фламинго становится state-of-the-art моделью еще на пяти бенчмарках.



Method	FT	Shot	OKVQA	VQAv2	COCO	MSVDQA	VATEX	VizWiz	Flick30K	MSRVTTQA	iVQA	YouTubeCook2	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	OOO	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	-
Flamingo	✗	32	51.0	60.4	106.3	47.2	57.4	44.0	72.8	29.4	40.7	77.3	41.2	OOO	32.6	28.4	63.5	-
	✗	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	-
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)	3

+ Успех модели



This is a picture of a cat.



This is a picture of an elephant's skin.



how do you know?



I have seen many pictures of elephants before.



where did you get the pictures from?



I got them from the internet.



This is a model of a city. It looks like a tiny city with a lot of people and cars.



This is not a model of a city. It is a photo of a city. Is there anything unusual about the photo?



I see a cat in the photo.



Is there anything unusual about the cat?



The cat is very big.

+ Универсальность

- Фламинго обучается с минимальным числом ограничений и имеет обобщающую способность между задачами
- Также Фламинго может быть использована через интерфейс в виде чата, отвечая на открытые вопросы
- Таким образом, конечные пользователи могут использовать Фламинго для задач с недостаточным для дообучения моделей числом данных
- Пример: Фламинго может использоваться для фильтрации выборок от токсичных или оскорбляющих картинок и текстов

+ Хорошо написанная статья

- Статья составляет 66 страниц, идеи объясняются довольно подробно
- Есть обширный ablation study
- Также достаточно подробно описывается техническая сторона обучения, гиперпараметры
- Аккуратно проведено тестирование (в частности, не применялся prompt engineering)
- Помимо анализа few-shot модели, также проанализировали дообученную Фламинго

+ Переиспользование моделей

- Обучение больших моделей (например, языковых) занимает как много времени на вычисления, так и много энергии
- Фламинго использует замороженную языковую модель и таким образом экономит энергию

- Качество классификации

- Фламинго довольно значительно проигрывает моделям, обученным на ContrastiveLoss на классификации
- Это связано с особенностями обучения модели и её заточенностью на универсальность, а также с использованием языковой модели



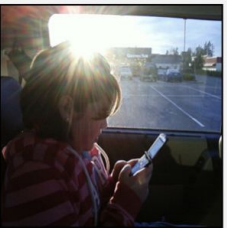
Model	Method	Prompt size	shots/class	ImageNet top 1	Kinetics700 avg top1/5
SotA	Fine-tuned	-	full	91.0 [137]	89.0 [144]
SotA	Contrastive	-	0	85.7 [90]	69.6 [94]
NFNetF6	Our contrastive	-	0	77.9	62.9
<i>Flamingo-3B</i>	RICES	8	1	70.9	55.9
		16	1	71.0	56.9
		16	5	72.7	58.3
<i>Flamingo-9B</i>	RICES	8	1	71.2	58.0
		16	1	71.7	59.4
		16	5	75.2	60.9
	Random	16	≤ 0.02	66.4	51.2
<i>Flamingo-80B</i>	RICES	8	1	71.9	60.4
		16	1	71.7	62.7
		16	5	76.0	63.5
	RICES+ensembling	16	5	77.3	64.2

- Проблемы языковой модели

- Визуальные токены используются в языковой модели однонаправленно, что проигрывает в производительности двунаправленному использованию (causal vs bidirectional)
- Трансформеры плохо генерализуются для текстов сильно длиннее, чем на этапе обучения

- Проблемы языковой модели

- Возможно, языковые модели (точнее, их веса) влекут некоторые галлюцинации и необоснованные ответы модели

Input Prompt	 <p>Question: What is on the phone screen? Answer:</p>	 <p>Question: What can you see out the window? Answer:</p>	 <p>Question: Whom is the person texting? Answer:</p>
Output	<p>A text message from a friend.</p>	<p>A parking lot.</p>	<p>The driver.</p>

- Проблемы метода few-shot обучения

- Фламинго использует in-context learning для few-shot обучения, с этим связаны некоторые проблемы:
- Сложность применения возрастает линейно или квадратично от числа примеров
- Чувствительность к промптам
- Качество метода выходит на плато после 32 примеров
- Показано, что на самом деле in-context learning не учится на примерах непосредственно, а пытается понять формат задачи (task location)

- Дополнительно

- Нет ablation study на выбор языковой модели, взята модель Chinchilla (тоже от DeepMind)
- Не выложены и не будут выкладываться веса модели, данные, код
- В выборке M3W нет видео

Итоги: плюсы и минусы

+	-
Модель получилась успешной	Плохое качество классификации
Модель получилась универсальной	Проблемы из-за языковой модели
Хорошо написанная статья	Проблемы из-за метода few-shot learning
Переиспользование моделей	Нет экспериментов над языковой моделью Все результаты приватные Нет выборки с видео на веб-страницах

Направления дальнейшей работы

- Исследовать и исправить все упомянутые проблемы данной работы
- Расширить спектр задач для Flamingo
 - Добавить другие модальности (например, аудио)
 - Подумать над задачами, включающими пространственные отношения между объектами (bounding box, optical flow)
- Изучить законы масштабирования (scaling laws) для визуальных моделей
- Поисследовать другие способы few-shot learning'a с Flamingo
- Сравнить с CoCa

ИСТОЧНИКИ

- <https://arxiv.org/pdf/2204.14198v1.pdf>
- https://scholar.google.com/citations?user=_VmflIEAAAAJ&hl=en&oi=ao
- <https://scholar.google.com/citations?user=L7IMQkQAAAAJ&hl=en&oi=ao>
- <https://www.youtube.com/watch?v=smUHQndcmOY>