



Flamingo: a Visual Language Model for Few-Shot Learning

Петров Олег
Поклонская Мария
Орлов Александр

Введение

Few-shot learning (FSL)

Support Set

Armadillo



Pangolin



Query



Armadillo or Pangolin?

Few-Shot Learning

Query:



Support Set:

Fox



Squirrel



Rabbit



Hamster



Otter



Beaver



Что такое Flamingo?






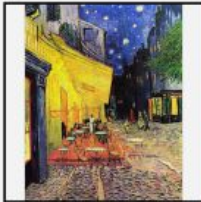



Flamingo — это новая зрительно-языковая модель (VLM) от DeepMind

VLM должна уметь принимать *мультимодальный* вход, а выдавать — текст

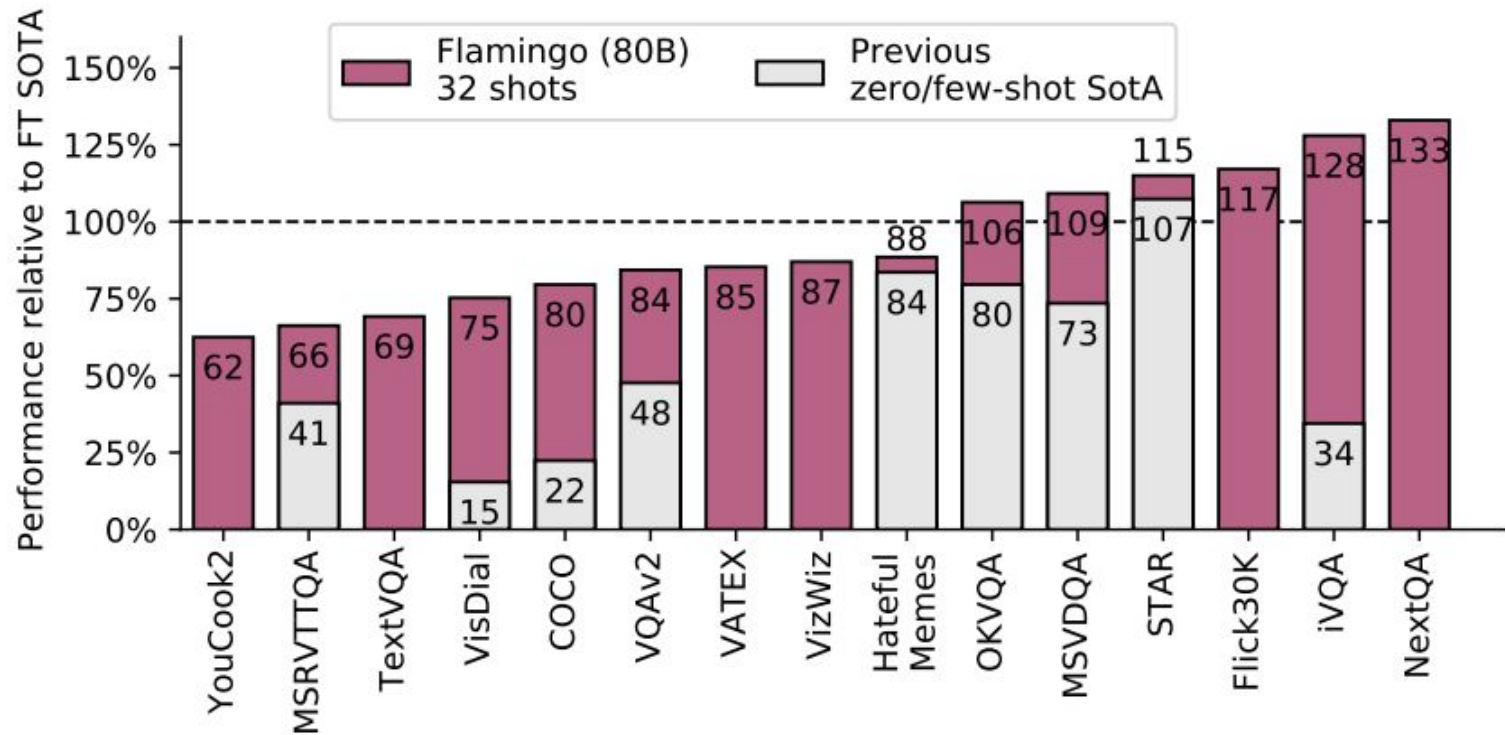
Может решать различные задачи: классификацию, captioning, visual dialogue, visual question answering и т.д., при этом она может *быстро адаптироваться под FSL задачи**

* всего из нескольких примеров выдавать хорошие результаты

Пример

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:	
	Output: "Underground"		Output: "Congress"		Output:	

SotA модель



Вдохновение

Большие LM (good few-shot learners):

1. Вход: примеры + подсказки (prompt), запрос
2. Выход: текст, продолжение запроса

Flamingo: аннотированные визуальные данные на входе:

1. Классификация
2. Субтитры
3. Ответы на вопросы

Архитектура

Две предобученные frozen-модели:

1. CV-модель: “восприятие” визуальных данных
2. Большая LM: базовая форма рассуждения

Инновация: связующая архитектура

Цель: сохранить “знания” моделей

Perceiver-based архитектура: доступ к HD

Подход

В общем

Flamingo моделирует следующее правдоподобие:

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

– вероятность появления текста y при условии изображения x .

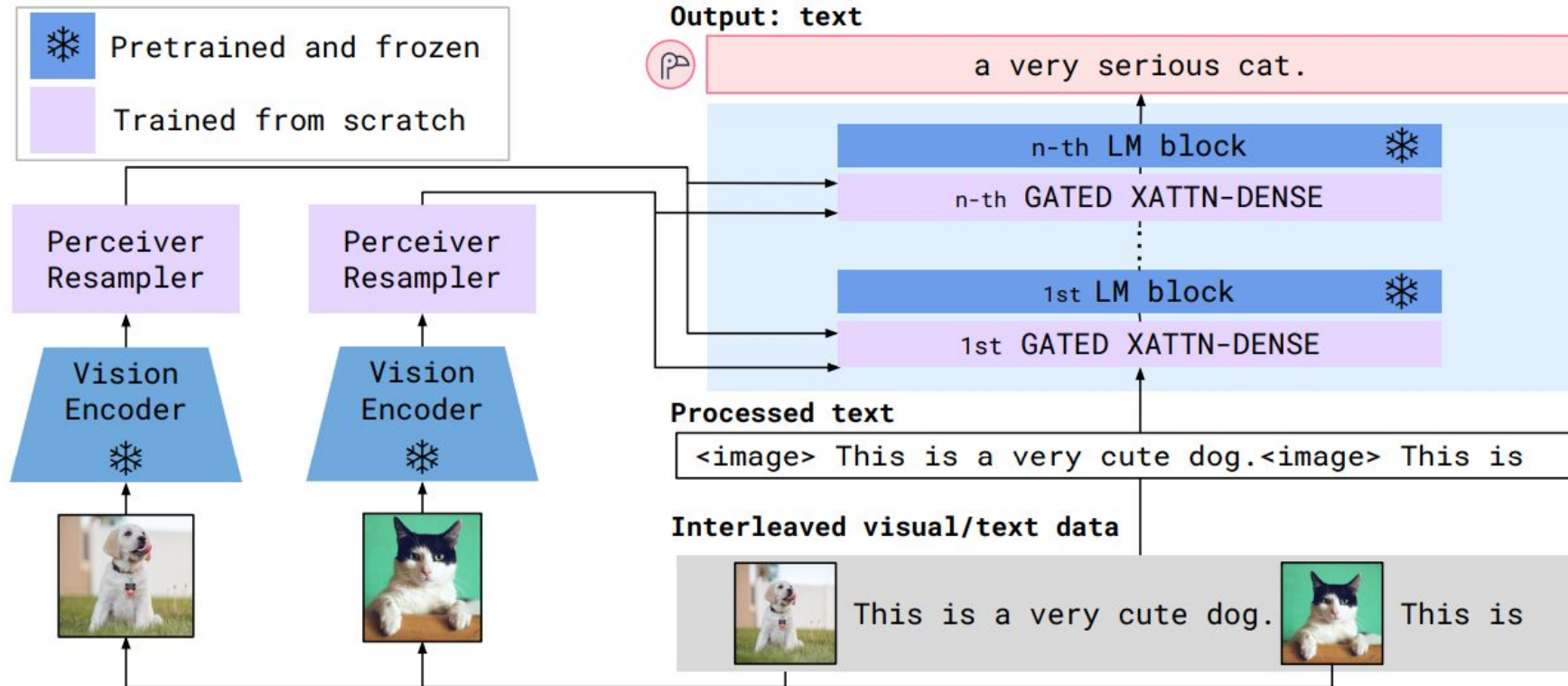
Здесь:

y_{ℓ} – ℓ -ый языковой токен

$y_{<\ell}$ – набор предшествующих токенов

$x_{\leq\ell}$ – набор визуальных токенов в мультимодальной последовательности

В общем



Vision Encoder

Vision Encoder: Normalizer-Free ResNet (NFNet) F6

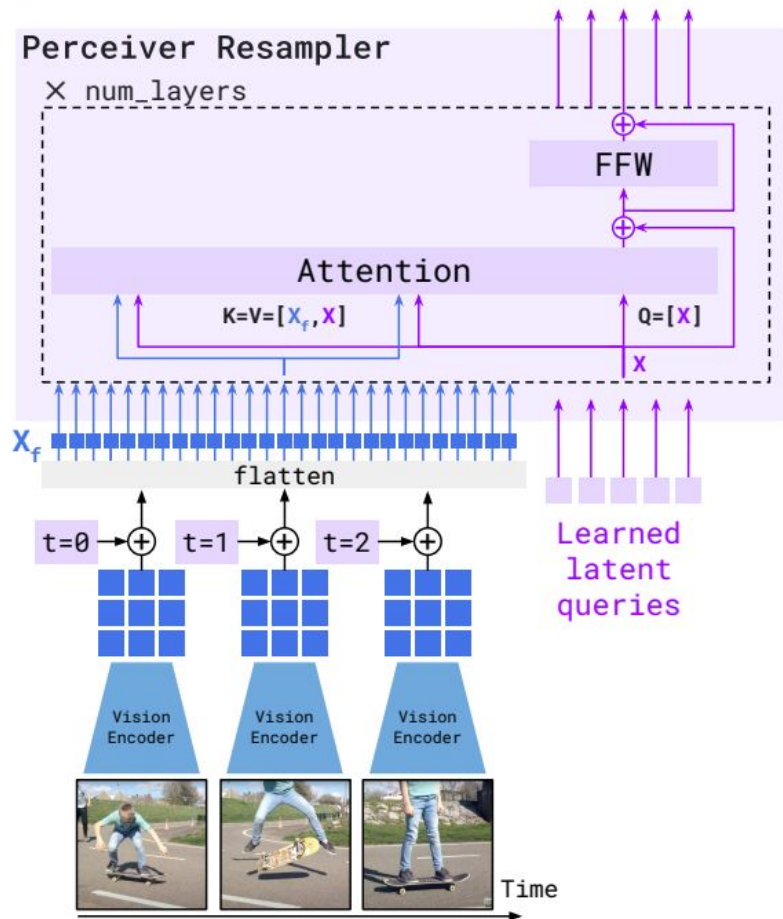
- Contrastive learning
- Two-term contrastive loss (Radford)
- Эмбединги из последней стадии
- Для видео: независимое покадровое кодирование

Все признаки переводятся в 1D на вход Perceiver Resampler

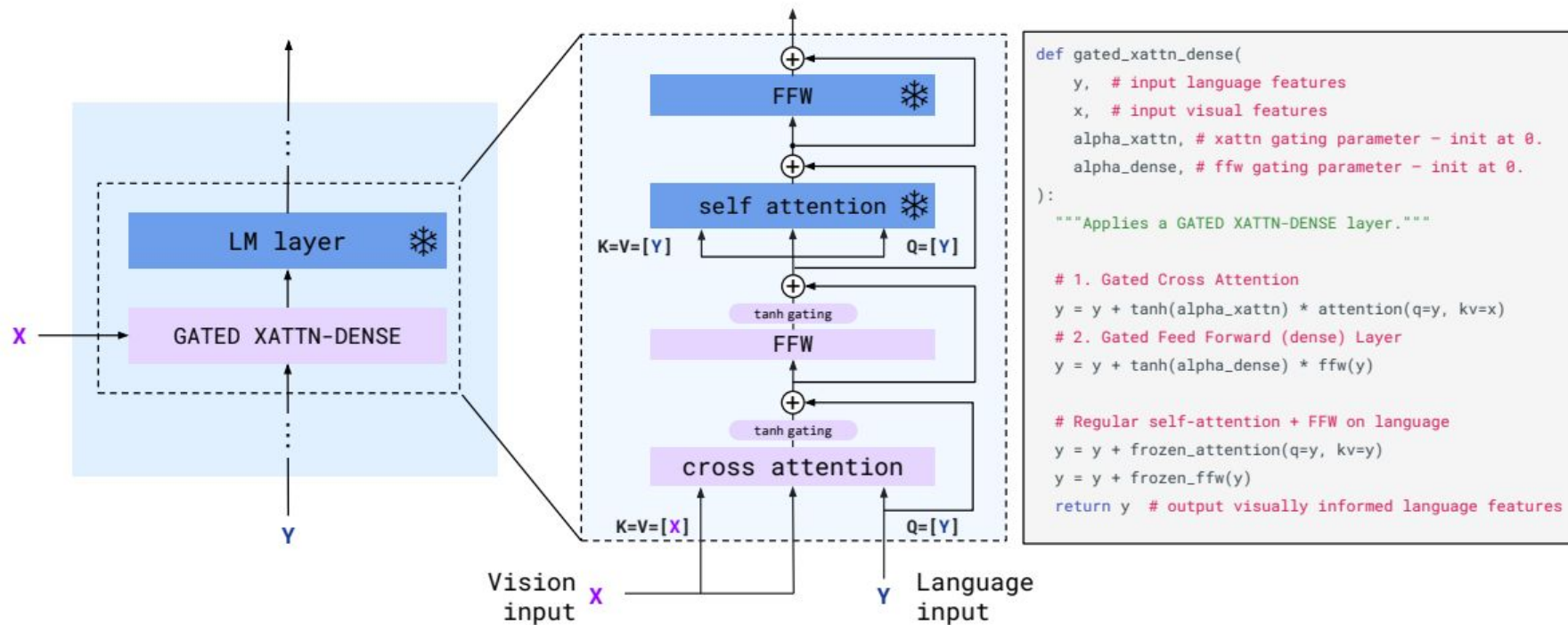
Perceiver Resampler

Perceiver Resampler: трансформер

- Переменный размер признаков в фиксированный
- Число queries предопределяется
- Число токенов на выходе равно числу queries
- K и V не обучаются



Перекрестное внимание



Обучение

Обучение и данные

Large LMs:

1. Большой объем данных
2. General-purpose генерация текста под запрос

Flamingo: обучение критически важно:

1. Много тщательно подобранных мультимодальных данных
2. Никаких данных, аннотированных для ML-целей (?)
3. Адаптация через FSL без подстройки под задачу

Данные

- Смесь из трех типов наборов данных
 - M3W dataset
 - image-text пары
 - video-text пары
- Все из Интернета

M3W (MultiModal Massive Web)

Обеспечивает few-shot возможности модели

- Парсинг HTML из 43M страниц
- Расположение изображений относительно текста
 - <image> токен на позиции изображения в тексте
 - <EOC> токен – выучен из словаря
- Случайные 256 токенов из документа
 - Первые 5 изображений

image/video-text

- ALIGN dataset (1.8M)
 - 12.4 токенов на изображение
- Собственное LTIP-дополнение (312M)
 - 20.5 токенов на изображение
- VTP (27M) с короткими видео
- Препроцессинг по аналогии с M3W

Оптимизация

Weighted sum of per-dataset expected NLL of text, given the visual inputs:

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

Здесь: \mathcal{D}_m , λ_m – набор данных и его вес

Значения λ_m важны

Градиенты накапливаются: лучше, чем “round-robin” подход

Адаптация

In-context learning

Vision to Text tasks (input=vision, output=text)

Support examples

Query



A cat wearing
sunglasses.



Elephants
walking in
the savanna.



<BOS><image>Output: A cat wearing sunglasses.<EOC><image>Output: Elephants walking in the savanna.<EOC><image>Output:

Processed prompt

Visual Question Answering Task (input=vision+text, output=text)

Support examples

Query



What's
the cat wearing?
sunglasses



How many
animals? 3



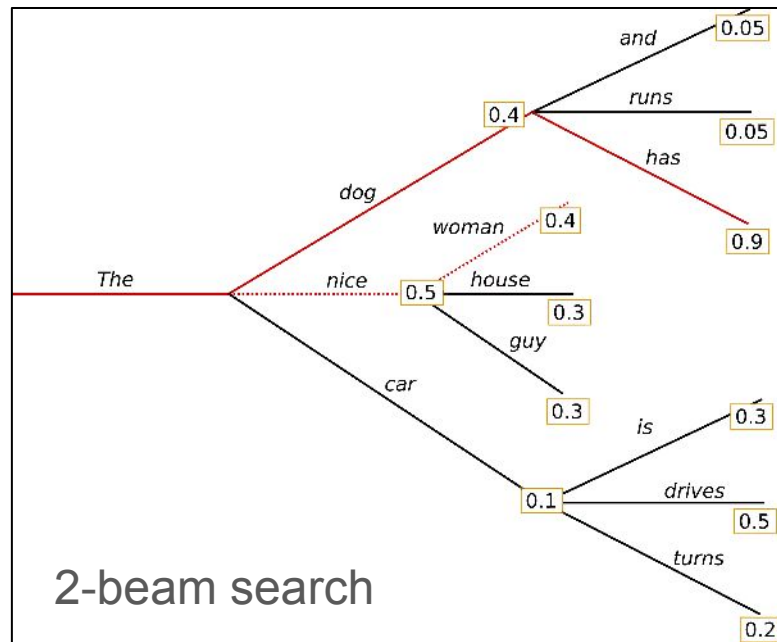
What is on
the water?

<BOS><image>Question: What's the cat wearing? Answer: sunglasses<EOC><image>Question: How many animals? Answer: 3<EOC><image>
Question: What is on the water? Answer:

Processed prompt

Open/close-ended evaluations

- Open-ended:
 - 3-beam search
- Close-ended:
 - Получаем всевозможные ответы
 - Создаем пары ($image, answer_i$)
 - Ранжируем ответы по правдоподобию



Эксперименты

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	(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)
<i>Flamingo-3B</i>	X	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
	X	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	-
	X	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	-
<i>Flamingo-9B</i>	X	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
	X	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	-
	X	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	-
<i>Flamingo</i>	X	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
	X	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	-
	X	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 [†]
	[133]	[133]	[119]	[153]	[65]	[65]	[51]	[79]	[123]	[132]	[137]	[84]	[152]

