

Мультимодальные модели

Лекция 13

План занятия

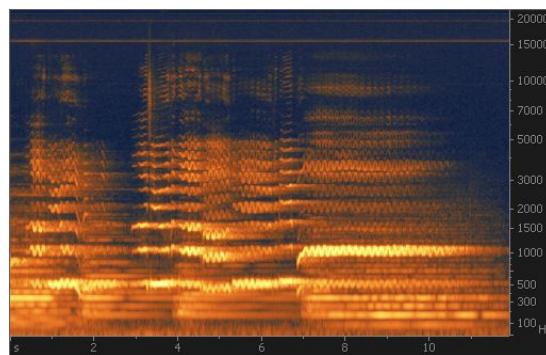
- Что такое мультимодальность? При чем тут языковые модели?
- История моделей: CLIP, BLIP, LLaVA
- Современные модели и бенчмарки

Мультимодальный домен

Помимо текста, есть множество различных модальностей (типов данных/источников информации). Которые по-разному обрабатываются, но могут быть полезны друг другу



Изображения



Аудио

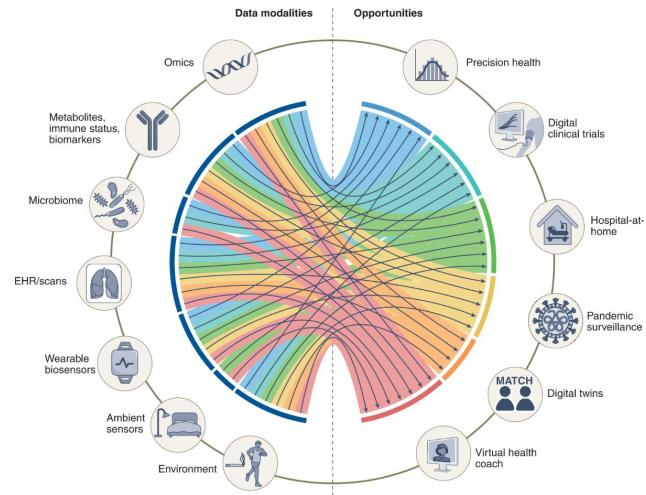
ТАБЛИЦА УМНОЖЕНИЯ

$1 \times 1 = 1$	$2 \times 1 = 2$	$3 \times 1 = 3$	$4 \times 1 = 4$
$1 \times 2 = 2$	$2 \times 2 = 4$	$3 \times 2 = 6$	$4 \times 2 = 8$
$1 \times 3 = 3$	$2 \times 3 = 6$	$3 \times 3 = 9$	$4 \times 3 = 12$
$1 \times 4 = 4$	$2 \times 4 = 8$	$3 \times 4 = 12$	$4 \times 4 = 16$
$1 \times 5 = 5$	$2 \times 5 = 10$	$3 \times 5 = 15$	$4 \times 5 = 20$
$1 \times 6 = 6$	$2 \times 6 = 12$	$3 \times 6 = 18$	$4 \times 6 = 24$
$1 \times 7 = 7$	$2 \times 7 = 14$	$3 \times 7 = 21$	$4 \times 7 = 28$
$1 \times 8 = 8$	$2 \times 8 = 16$	$3 \times 8 = 24$	$4 \times 8 = 32$
$1 \times 9 = 9$	$2 \times 9 = 18$	$3 \times 9 = 27$	$4 \times 9 = 36$
$1 \times 10 = 10$	$2 \times 10 = 20$	$3 \times 10 = 30$	$4 \times 10 = 40$
$5 \times 1 = 5$	$6 \times 1 = 6$	$7 \times 1 = 7$	$8 \times 1 = 8$
$5 \times 2 = 10$	$6 \times 2 = 12$	$7 \times 2 = 14$	$8 \times 2 = 16$
$5 \times 3 = 15$	$6 \times 3 = 18$	$7 \times 3 = 21$	$8 \times 3 = 24$
$5 \times 4 = 20$	$6 \times 4 = 24$	$7 \times 4 = 28$	$8 \times 4 = 32$
$5 \times 5 = 25$	$6 \times 5 = 30$	$7 \times 5 = 35$	$8 \times 5 = 40$
$5 \times 6 = 30$	$6 \times 6 = 36$	$7 \times 6 = 42$	$8 \times 6 = 48$
$5 \times 7 = 35$	$6 \times 7 = 42$	$7 \times 7 = 49$	$8 \times 7 = 56$
$5 \times 8 = 40$	$6 \times 8 = 48$	$7 \times 8 = 56$	$8 \times 8 = 64$
$5 \times 9 = 45$	$6 \times 9 = 54$	$7 \times 9 = 63$	$8 \times 9 = 72$
$5 \times 10 = 50$	$6 \times 10 = 60$	$7 \times 10 = 70$	$8 \times 10 = 80$
$9 \times 1 = 9$	$10 \times 1 = 10$	$11 \times 1 = 11$	$12 \times 1 = 12$
$9 \times 2 = 18$	$10 \times 2 = 20$	$11 \times 2 = 22$	$12 \times 2 = 24$
$9 \times 3 = 27$	$10 \times 3 = 30$	$11 \times 3 = 33$	$12 \times 3 = 36$
$9 \times 4 = 36$	$10 \times 4 = 40$	$11 \times 4 = 44$	$12 \times 4 = 48$
$9 \times 5 = 45$	$10 \times 5 = 50$	$11 \times 5 = 55$	$12 \times 5 = 60$
$9 \times 6 = 54$	$10 \times 6 = 60$	$11 \times 6 = 66$	$12 \times 6 = 72$
$9 \times 7 = 63$	$10 \times 7 = 70$	$11 \times 7 = 77$	$12 \times 7 = 84$
$9 \times 8 = 72$	$10 \times 8 = 80$	$11 \times 8 = 88$	$12 \times 8 = 96$
$9 \times 9 = 81$	$10 \times 9 = 90$	$11 \times 9 = 99$	$12 \times 9 = 108$
$9 \times 10 = 90$			$12 \times 10 = 120$

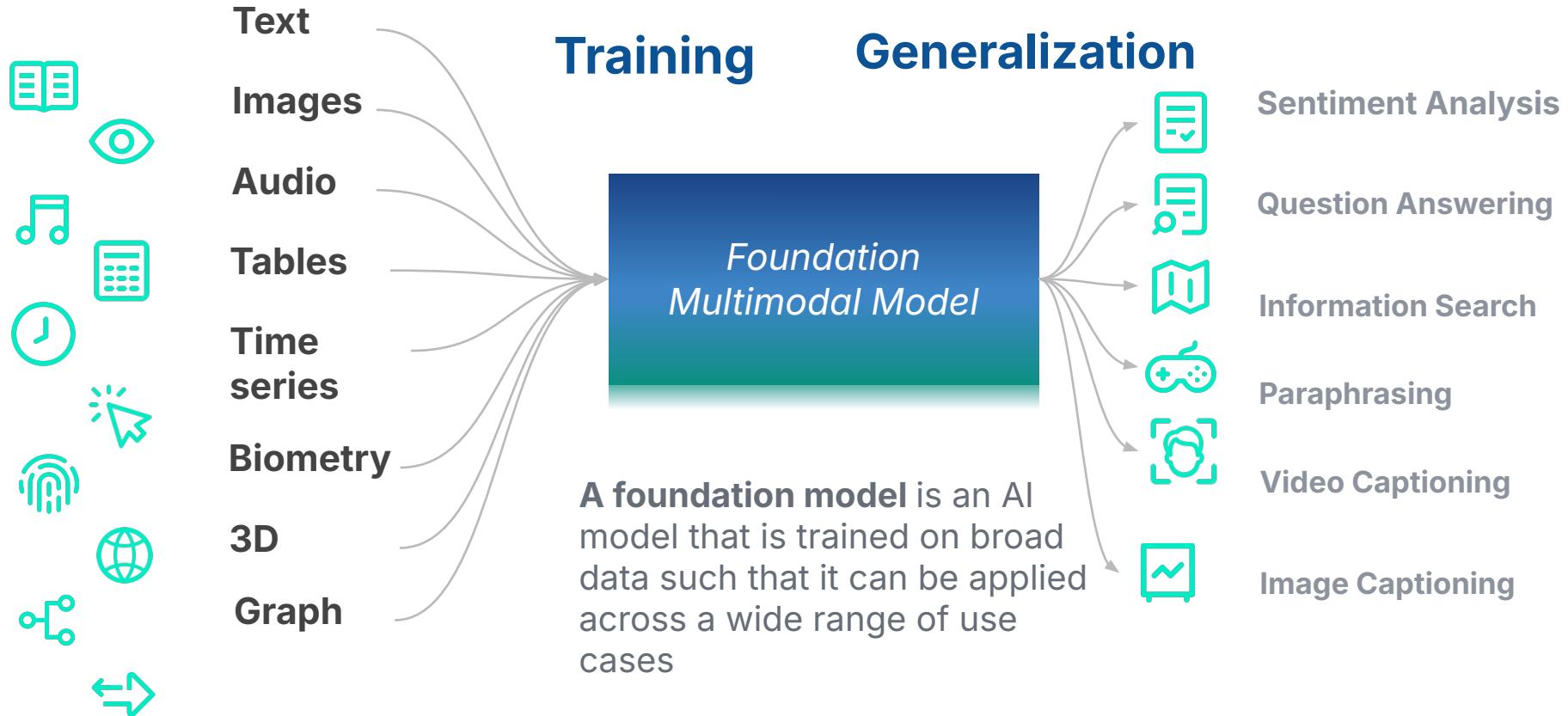
Таблицы, временные ряды, графы

Зачем нужна мультимодальность?

- Информация об объекте может быть получена различным образом: – по словам песни, мелодии мы можем узнать клип, артиста
- Информация из разных модальностей может дополнять друг друга – например, медицинские анализы



Global trend towards multimodality



При чём тут LLM?

LLM – может выступать в качестве foundational model!

- 1) Очень много знаний из различных сфер содержится в LLM
- 2) LLM показала свою обобщаемость и многие задачи решаются в zero-shot формате
- 3) LLM может помочь достичь результата, когда задача полностью описывается текстом.

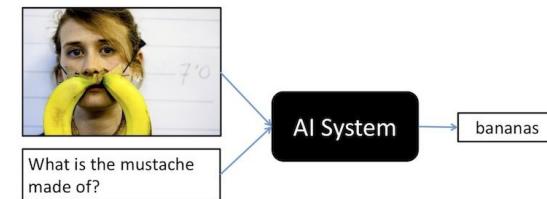
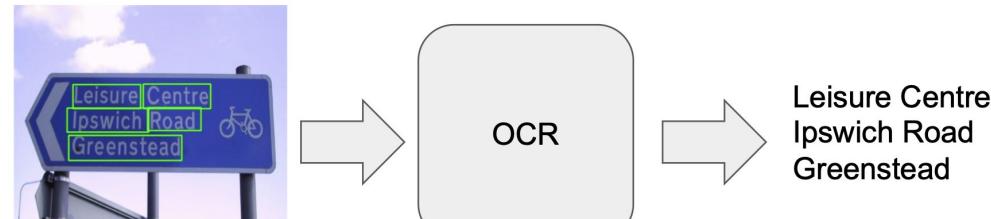
Вопросы?

Визуально-текстовая мультимодальность

Одна из самых популярных мультимодальностей – картиночно-текстовая

Текущие задачи:

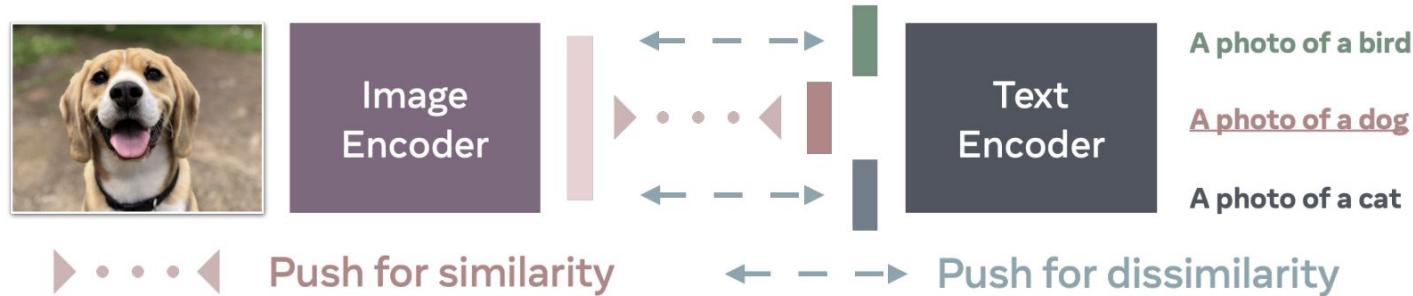
- Image retrieval
- Image captioning
- Visual Questioning Answering
- Classical computer vision tasks
- Text-to-image
- OCR



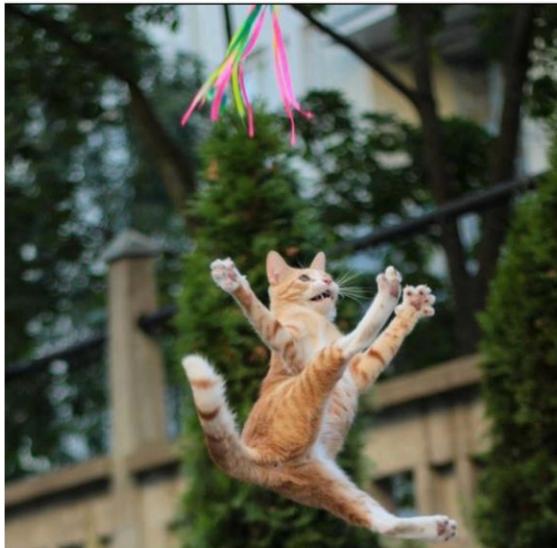
CLIP (Contrastive Language-Image Pre-training 2021)

Основная из работ, которая по сути из запустила мультимодальность – CLIP.

Ее идея заключается в том, чтобы обучить модель сопоставлять изображения и картинки вместе.



CLIP: датасет



Рыжий кот прыгнул за цветными нитками растопырив лапы



Кот-толстяк пытается слезть с дерева

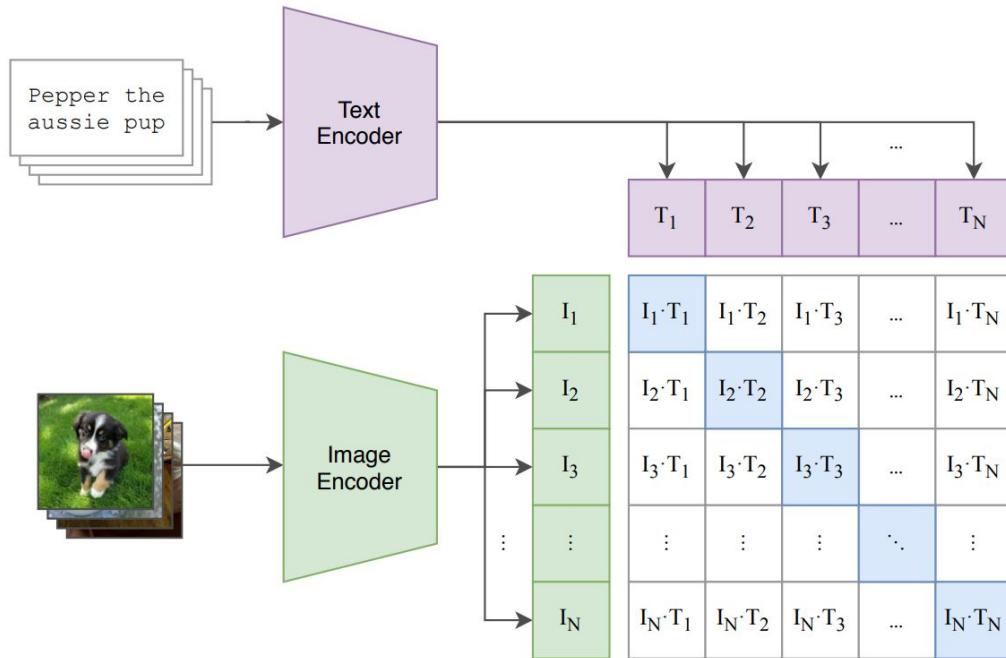


Просто кот

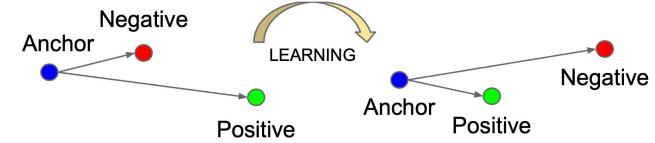
400 миллионов пар <картина>:<текст>

CLIP: сетап

(1) Contrastive pre-training



Учимся контрастивно приближать положительные примеры и отдалять отрицательные



CLIP: loss

```
# image_encoder - ResNet or Vision Transformer
# text_encoder - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l] - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T) #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss = (loss_i + loss_t)/2
```

1. Нормализуем эмбеддинги;
2. Считаем similarity-матрицу скалярных произведений эмбеддингов и умножаем каждый элемент матрицы на температуру;
3. Применяем кросс-энтропию по столбцам и строкам similarity-матрицы;
4. Складываем эти 2 лосса и делим сумму на 2

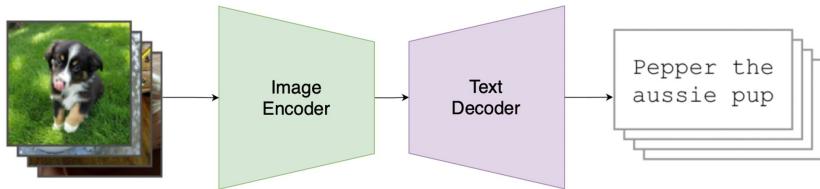
Сетап обучения:

- from scratch
- 32 эпохи
- Adam с weight decay
- cosine scheduler для learning rate.
- большие батчи размера 32,768!

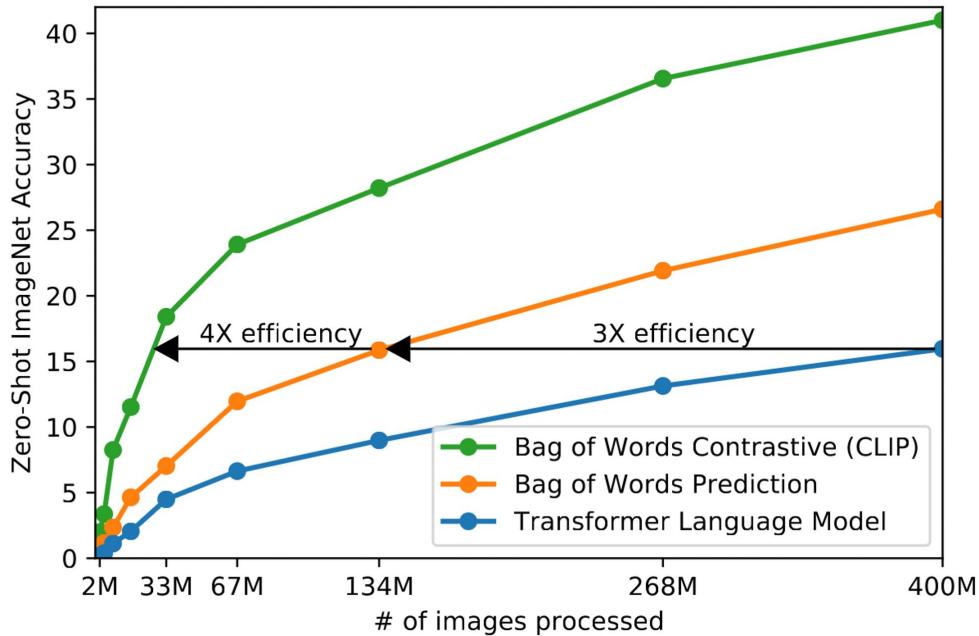
CLIP с самой большой ResNet моделью учился **18 дней на 592 V100 GPU**.

CLIP с самой большой ViT моделью учился **12 дней на 256 V100 GPU**.

CLIP: contrastive vs captioning

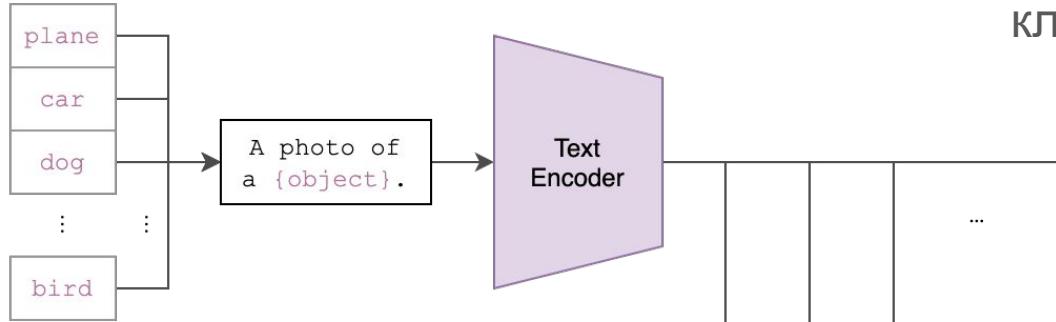


Можно посмотреть на это с такой стороны:
описания картинок из интернета иногда
очень сложно восстановить при помощи
одной картинки. Восстановить только
ключевые слова без предлогов — задача
более простая. Но еще проще было бы
соотнести описание и картинки и найти
между ними соответствие.



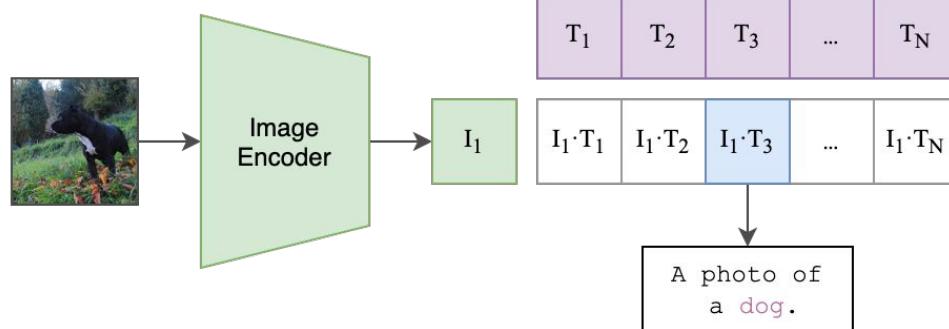
CLIP: zero-shot

(2) Create dataset classifier from label text



Универсальный
классификатор

(3) Use for zero-shot prediction



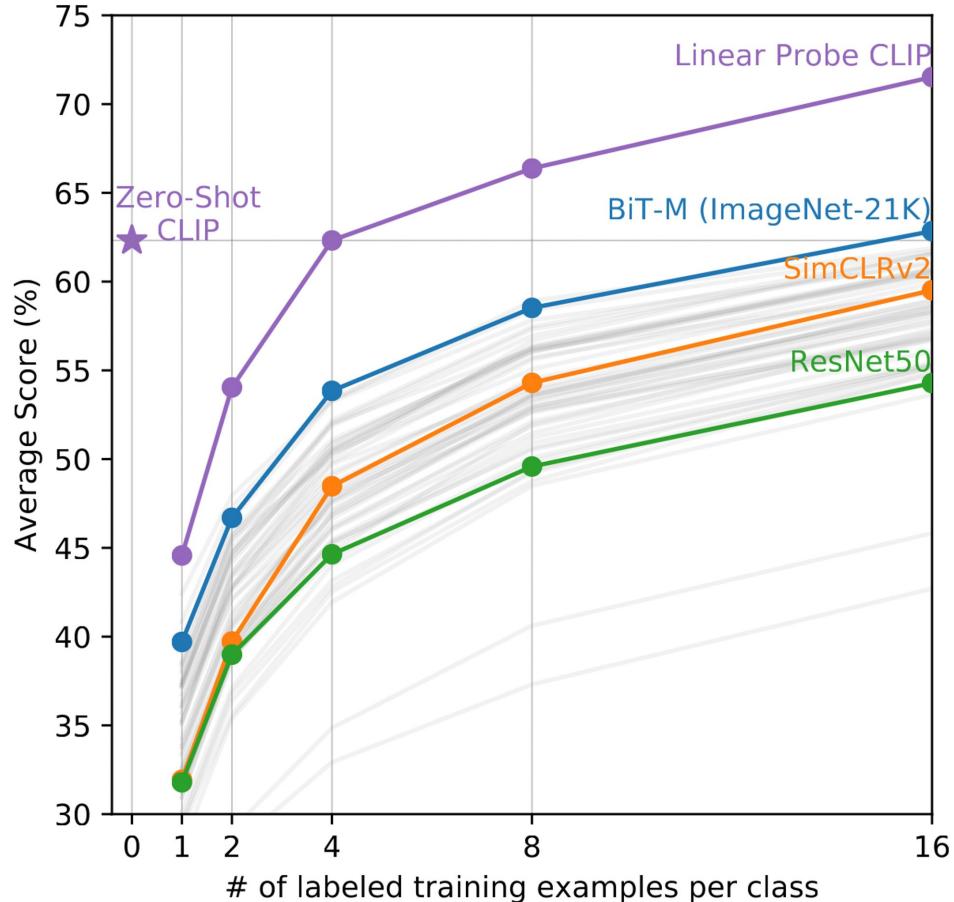
CLIP: linear probing

Суть метода:

мы добавляем линейный классификатор к выходам модели, замораживаем все веса модели и учим веса только этого линейного классификатора.

С ним можно выполнять 1-shot learning, 2-shot learning, 3-shot learning и так далее.

Это подходы, когда для обучения модели используются только по 1, 2, 3 и так далее изображений на класс.



CLIP

Model	Training data	Resolution	# of samples seen	ImageNet zero-shot acc.
ConvNext-Base	LAION-2B	256px	13B	71.5%
ConvNext-Large	LAION-2B	320px	29B	76.9%
ConvNext-XXLarge	LAION-2B	256px	34B	79.5%
ViT-B-32-256	DataComp-1B	256px	34B	72.8%
ViT-B-16	DataComp-1B	224px	13B	73.5%
ViT-L-14	LAION-2B	224px	32B	75.3%
ViT-H-14	LAION-2B	224px	32B	78.0%
ViT-L-14	DataComp-1B	224px	13B	79.2%
ViT-bigG-14	LAION-2B	224px	34B	80.1%

CLIP: overview

1. Это набор очень сильных бэкбонов, которые выдают качественные эмбеддинги. Модели используются в бэкбонах DALLE и Stable-diffusion.
2. Это классная модель для получения бесплатной разметки на ваш датасет, если домен не слишком специфичный. Но CLIP плохо понимает мелкие отличия и плохо работает с маленькими изображениями.
3. На своих задачах можно дотюнить CV-часть при помощи небольшого числа промптов и картинок и качество будет уже неплохим.
4. С CLIP в целом начался бум text2image моделей.
5. Минус: CLIP не генерирует текст

Вопросы?

Как решали задачу Image Captioning раньше?

Deep Visual-Semantic Alignments for Generating Image Descriptions 2015

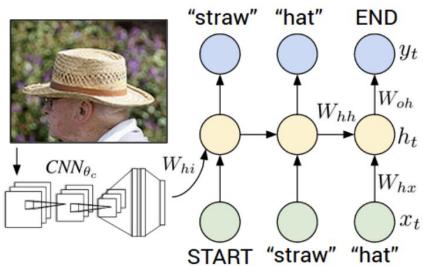


Figure 4. Diagram of our multimodal Recurrent Neural Network generative model. The RNN takes a word, the context from previous time steps and defines a distribution over the next word in the sentence. The RNN is conditioned on the image information at the first time step. START and END are special tokens.

Show-attend and tell 2016

Figure 2. Attention over time. As the model generates each word, its attention changes to reflect the relevant parts of the image. "soft" (top row) vs "hard" (bottom row) attention. (Note that both models generated the same captions in this example.)

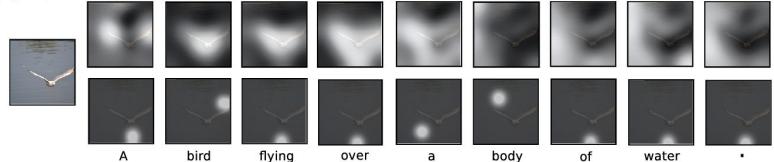
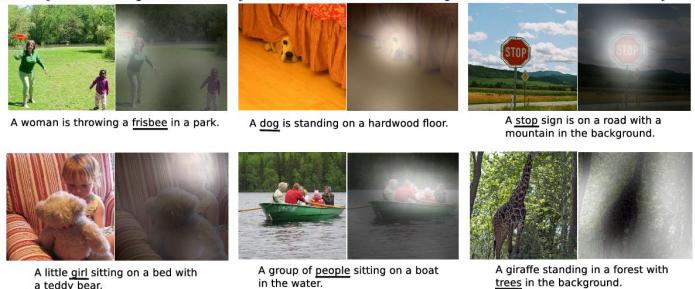


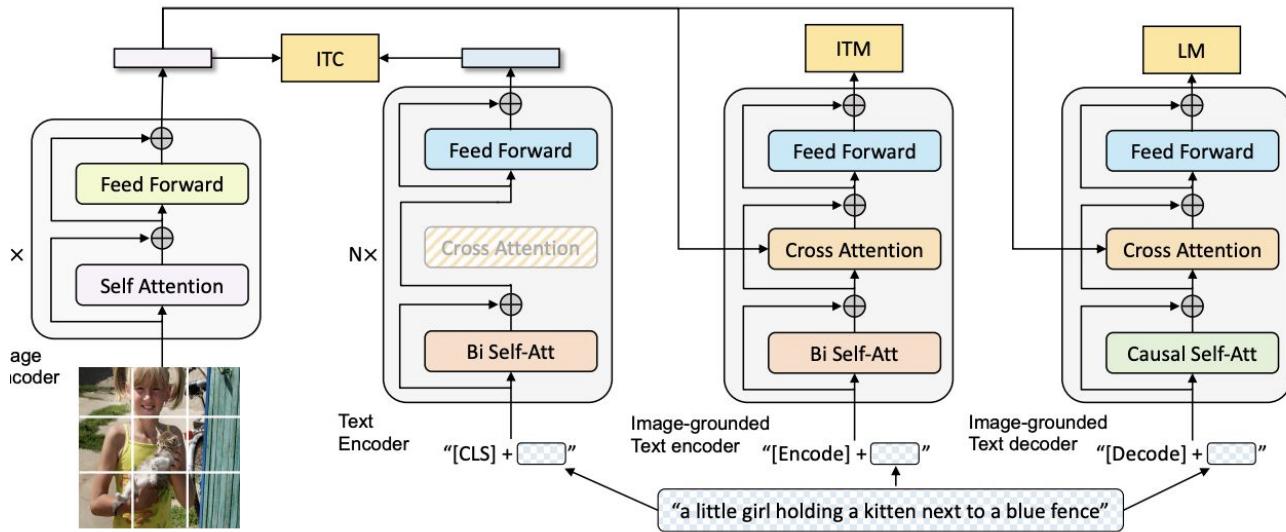
Figure 3. Examples of attending to the correct object (white indicates the attended regions, underlines indicated the corresponding word)



Естественным решением является дать возможность языковой “смотреть на картинку во время генерации”

BLIP (Bootstrapping Language-Image Pre-training 2022)

Новый взгляд на идею того, как делать image captioning, используя механизм внимания.



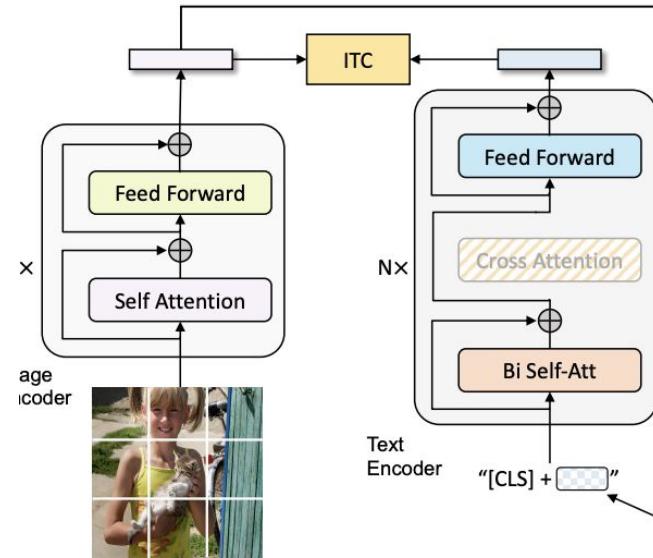
1. **ITC (Image-Text Contrastive):** Контрастивное обучение как в CLIP. ITM (Image-Text Matching): **Бинарная классификация** (ITM). Модель получает пару (изображение, текст) и должна предсказать, соответствуют ли они друг другу (positive) или нет (negative). Это помогает модели глубже понять семантическую связь.
2. **LM (Language Modeling):** Это и есть Image Captioning.

$$\text{Loss} = \text{ITC} + \text{ITM} + \text{LM}$$

BLIP

Первый этап это CLIP по сути – выравнивание картиночных и текстовых представлений.

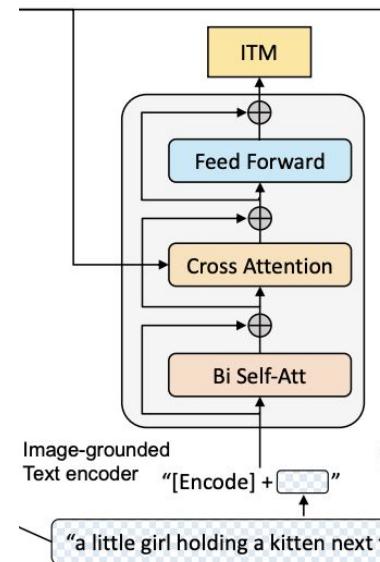
Таким образом, модель выучивает базовое представление между картинками и текстом.



BLIP

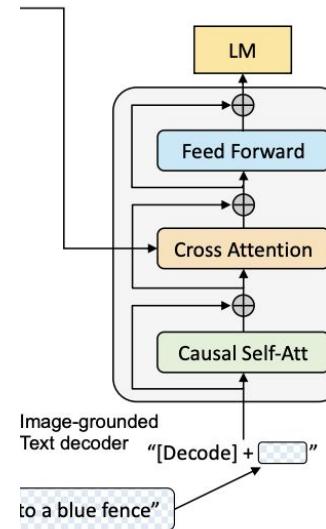
Второй этап – научить модель использовать через cross-attention эти признаки как в старых моделях

Решаем такую задачу, что и раньше на contrastive, но тут уже как бинарная классификация



BLIP

Финальный этап – делаем captioning учим классическую авторегрессию как в LLM



BLIP: данные

Самое главное всегда и везде будут данные.

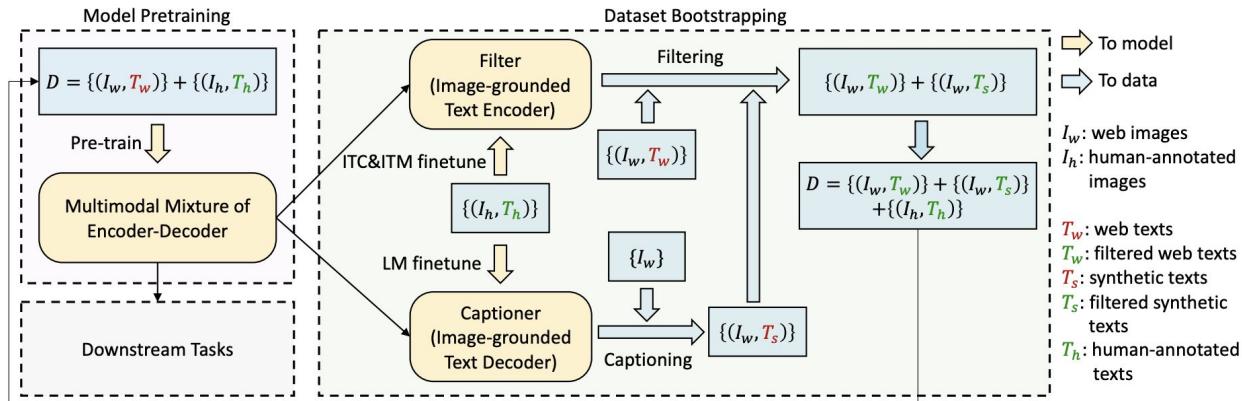


Figure 3. Learning framework of BLIP. We introduce a captioner to produce synthetic captions for web images, and a filter to remove noisy image-text pairs. The captioner and filter are initialized from the same pre-trained model and finetuned individually on a small-scale human-annotated dataset. The bootstrapped dataset is used to pre-train a new model.

Ключевое нововведение BLIP
«Бутстрэппинг данных»

Проблема: данные из интернета (картинка-текст) часто шумные.

Решение:

1. Тюним кэшнер и фильтр на качественно аннотированной небольшой выборке.
2. Новый кэшнер создает чистые описания картинок.
3. Фильтр удаляет мусорные пары.

BLIP: summary

C (Captioner), F (Filter)

	Pre-train dataset	Bootstrap C F	Vision backbone	Retrieval-FT (COCO)		Retrieval-ZS (Flickr)		Caption-FT (COCO)		Caption-ZS (NoCaps)	
				TR@1	IR@1	TR@1	IR@1	B@4	CIDEr	CIDEr	SPICE
Более чистый датасет	COCO+VG	X X	ViT-B/16	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
	+CC+SBU (14M imgs)	X ✓ _B		79.1	61.5	94.1	82.8	38.1	128.2	102.7	14.0
		✓ _B X		79.7	62.0	94.4	83.6	38.4	128.9	103.4	14.2
		✓ _B ✓ _B		80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4
Добавление шумного LAION	COCO+VG	X X	ViT-B/16	79.6	62.0	94.3	83.6	38.8	130.1	105.4	14.2
	+CC+SBU	✓ _B ✓ _B		81.9	64.3	96.0	85.0	39.4	131.4	106.3	14.3
	+LAION (129M imgs)	✓ _L ✓ _L		81.2	64.1	96.0	85.5	39.7	133.3	109.6	14.7
		X X		80.6	64.1	95.1	85.5	40.3	135.5	112.5	14.7
		✓ _L ✓ _L	ViT-L/16	82.4	65.1	96.7	86.7	40.4	136.7	113.2	14.8

Вопросы?

FRoMAGE (2023)

Одного линейного слоя достаточно, чтобы научить модель в текст

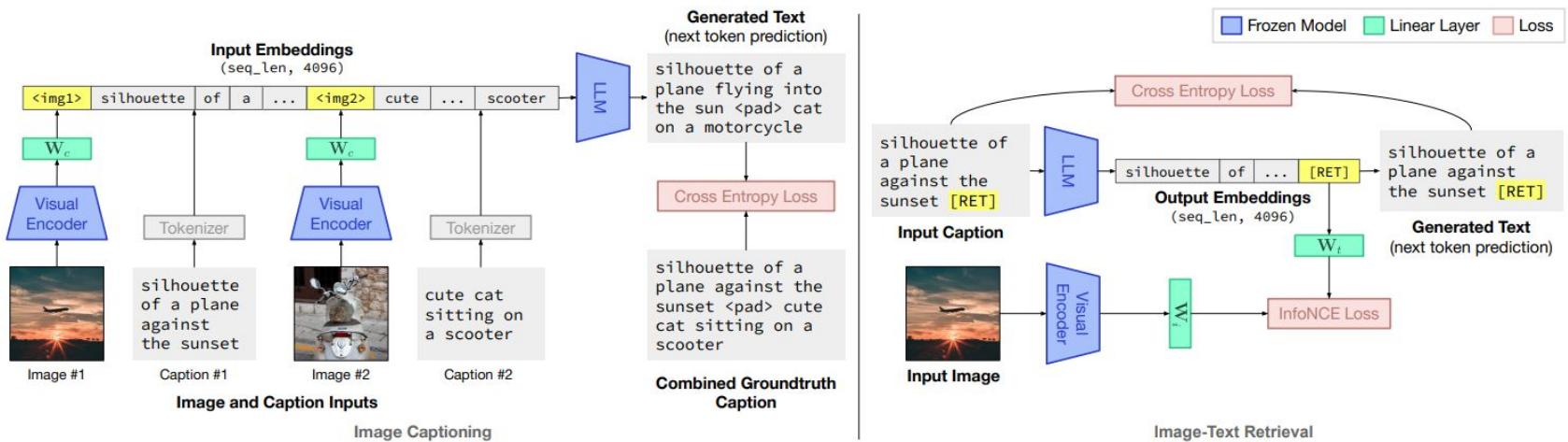
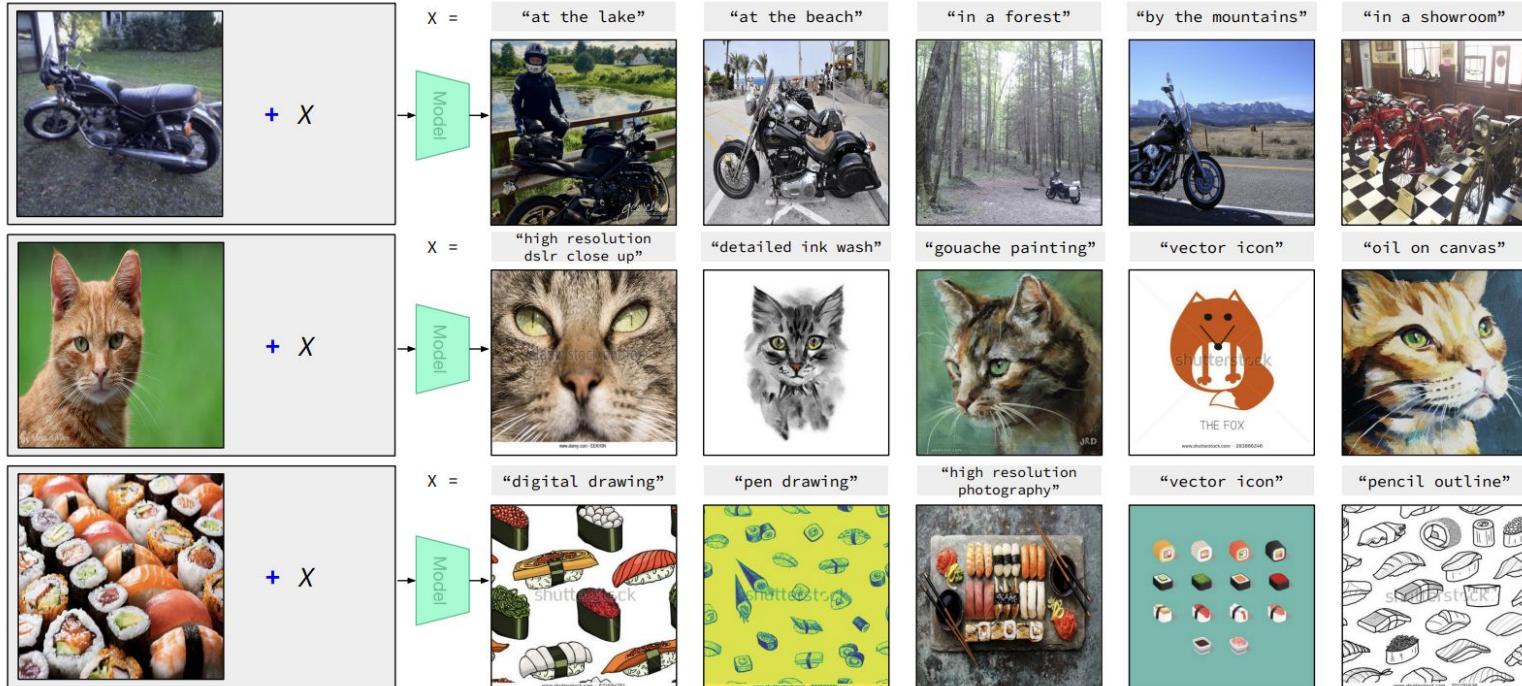


Figure 2. Overview of the FROMAGE architecture. FROMAGE is a model trained on image-text pairs for image captioning and image-text retrieval. It is capable of processing arbitrarily interleaved image and text inputs, and producing interleaved images and text as outputs.

$$\text{Loss} = \text{Caption_loss } (W_c) + \text{Retrieval_loss } (W_i, W_t)$$

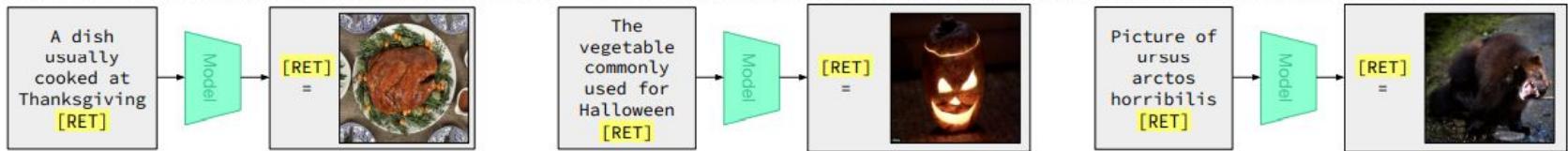
FRoMAGE: ВОЗМОЖНОСТИ



FRoMAGE: ВОЗМОЖНОСТИ

Multimodal Concept Composition

FROMAGe can seamlessly composite image and text data to produce images with the desired style or content.



World Knowledge

Our method can draw upon knowledge learnt outside of CC3M (through large scale text pretraining of the frozen LLM) to return valid image outputs.



Visual Storytelling

FROMAGe can learn from in-context examples to generate story-like image and text outputs.



FRoMAGE

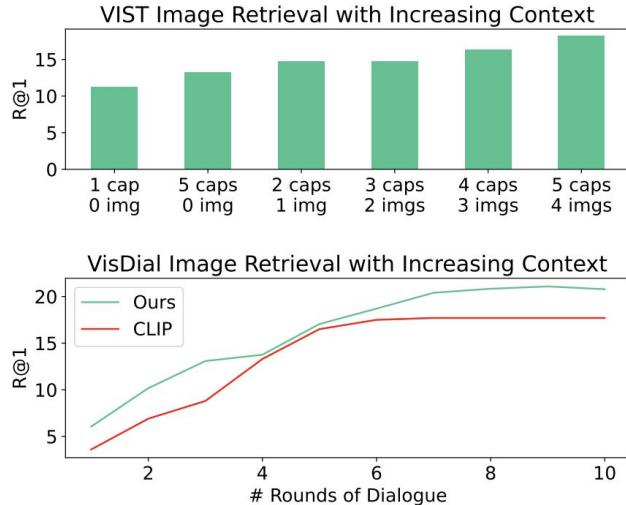


Figure 5. Increasing input context generally improves performance. Shown are results for image retrieval for VIST (Huang et al., 2016) (top) and image retrieval on VisDial (Das et al., 2017) (bottom).

Поиск последнего изображения в истории (story) по предыдущим описаниям и изображениям.

Поиск изображения по истории диалога (вопросов и ответов) о нём.

FROMAGE, благодаря архитектуре, гораздо лучше понимает длинные и сложные текстовые описания, чем CLIP, который был обучен на коротких кэшенах.

Our models are trained with a batch size of 180 for 1 epoch (18000 iterations) on 1 A6000 GPU (clock time of 24 hours).

FRoMAGE

I am thinking of an animal. It is brown and furry, and has a tail. What is it?

A squirrel.

No, the animal I am thinking of lives in the river.

A beaver.

That's right, what do beavers look like?

They have a tail.

Show me an illustration of a beaver

A beaver.


Yes, what about a pencil drawing of one?



What about a photo of one?



What is this?

This is a house with a unique design.

What is its shape?

The house is shaped like a dome.

Can you show me more houses like this?

The house is designed by architect.



That's nice, any others?

Yes, there are many houses like this in the world.



- LLM генерирует авторегрессивно, как обычно.
- В любой момент она может сгенерировать токен [RET].
- Как только [RET] сгенерирован:
 - a. Берётся его hidden state
 - b. Проецируется с помощью W_t
 - c. Ищется изображение, максимизирующее sim
 - d. Это изображение вставляется в вывод

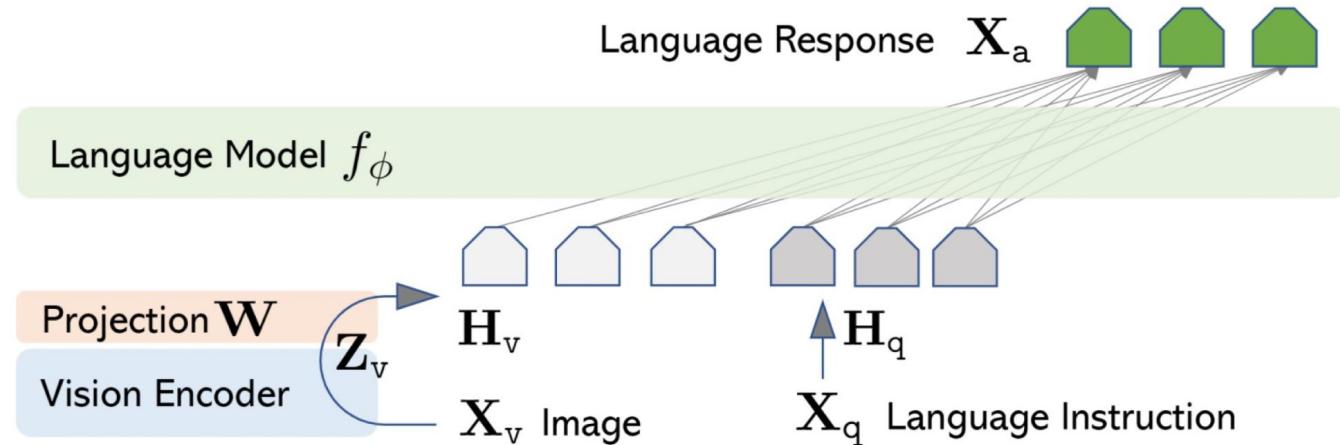
FRoMAGE: summary

- 1) **Поразительная простота:** Отсутствие сложных cross-attention слоев между модальностями. Вся "мультимодальная логика" делегируется LLM.
- 2) **Быстрое обучение:** Благодаря простоте, модель обучается очень быстро (например, 1 эпоха за 24 часа на одном GPU).
- 3) **Сложное поведение:** Несмотря на простоту, модель демонстрирует способность вести диалог с изображениями, отвечать на вопросы, следовать инструкциям ("покажи фото...", "а теперь рисунок...") и даже выполнять рассуждения, основанные на визуальном контексте. Это показывает мощь LLM как универсального процессора информации.

Вопросы?

LLaVA (Visual Instruction Tuning 2023)

LlaVa



Сбор датасета

1. База — Conceptual Captions 3M (CC3M): ~3.3 млн пар «изображение–подпись».
2. Фильтрация — оставлено 595K пар, для которых $\text{CLIP}(\text{image}, \text{caption}) > 0.23$
3. Генерация инструкций — для каждой подписи использовался GPT-4 для синтеза 3–5 вариантов инструкций (VQA, captioning, dialogue). Получены 158K примеров.
4. Дополнение — примеры из ScienceQA (вопросы + лекции + изображения схем/графиков).

Примеры шаблонов LLaVA

Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.

Luggage surrounds a vehicle in an underground parking area

People try to fit all of their luggage in an SUV.

The sport utility vehicle is parked in the public garage, being packed for a trip

Some people with luggage near a van that is transporting it.



Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

Response type 3: complex reasoning

Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

Table 1: One example to illustrate the instruction-following data. The top block shows the contexts such as captions and boxes used to prompt GPT, and the bottom block shows the three types of responses. Note that the visual image is not used to prompt GPT, we only show it here as a reference.

LLaVA: pretraining

To strike a balance between concept coverage and training efficiency, we filter CC3M to 595K image-text pairs

Example#1: Interleaving the image *before* each corresponding text

[..., "Check out Shane Driscoll's take on sustainable communities and how his photograph fits this year's Green Cities theme.", ..., , "Man-made platforms like the one pictured here allow these fish-eating birds of prey to thrive in developed coastal areas.", "A city surrounded by mountains.", "I took this photo in October on a hike in New Hampshire.", , "It is looking at Mt. Chocorua from the middle sister mountain.", "Getting people out into beautiful places like this is becoming more and more popular, and each time we bring a little piece of nature back with us that inspires us to make our cities better.", ...]

Что учится: Линейный проектор

Example#2: Interleaving the image *after* each corresponding text

["This Walnut and Blue Cheese Stuffed Mushrooms recipe is sponsored by Fisher Nuts.", , "Stuffed mushrooms are an appetizer that always grabs my attention at a party.", , "If you are a mushroom lover, like me, you probably feel the same.", "The ideas for stuffing mushrooms are endless, so many combinations to play with, a couple of my personal favorites are these Mediterranean Stuffed Mushrooms and these Spinach and Toasted Pine Nut Stuffed Mushrooms.", , "Well, you can officially add these Walnut and Blue Cheese Stuffed Mushrooms to my favorites list.", "The ingredients for the stuffing are simple, which is always best.", ...]

Заморожен: Vision encoder (CLIP ViT-L/14) и веса LLM (Vicuna)

Fine-tuning End-to-End

Два сценария дообучения:

- Мультимодальный чатбот — дообучение на 158К инструкций.
- ScienceQA — дообучение на научных вопросах с объяснениями (включая изображения).

Что учится: Веса проектора и все веса LLM (Vicuna)

Заморожен: Энкодер CLIP

LLaVA: example

Visual input example, Extreme Ironing:	
	
User LLaVA	Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg What is unusual about this image? The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment. [Start a new conversation, and clear the history] What's happening in the scene? The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.
User GPT-4 [36]	What is unusual about this image? The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User BLIP-2	What is unusual about this image? a man is sitting on the back of a yellow cab
User OpenFlamingo	What is unusual about this image? The man is drying his clothes on the hood of his car.

Table 3: Example prompt from GPT-4 paper [36] to compare visual reasoning and chat capabilities. Compared to BLIP-2 [28] and OpenFlamingo [5], LLaVA accurately follows the user’s instructions, instead of simply describing the scene. LLaVA offers a more comprehensive response than GPT-4. Even when merely asked to describe the image, LLaVA identifies atypical aspects of the image.

Image description

• "Describe the image concisely." •

"Provide a brief description of the given image." •

"Offer a succinct explanation of the picture presented."

"Summarize the visual content of the image."

"Give a short and clear explanation of the subsequent image."

"Share a concise interpretation of the image provided."

"Present a compact description of the photo's key features."

"Relay a brief, clear account of the picture shown."

"Render a clear and concise summary of the photo."

"Write a terse but informative summary of the picture."

"Create a compact narrative representing the image presented."

Вопросы?

QwenVL & BLIP-2

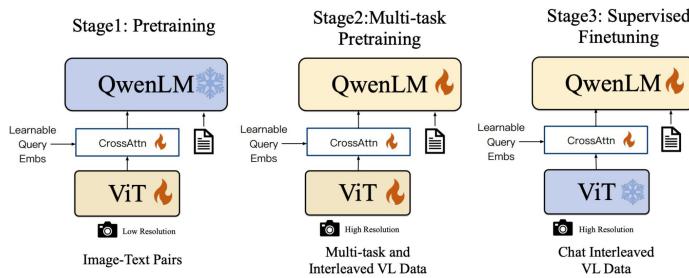
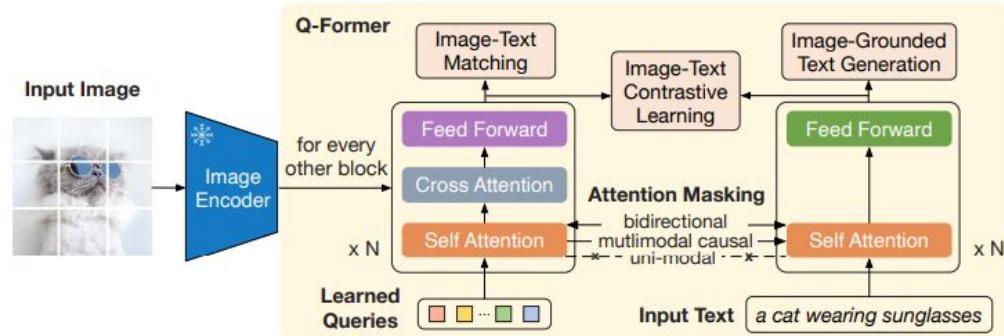


Figure 3: The training pipeline of the Qwen-VL series.



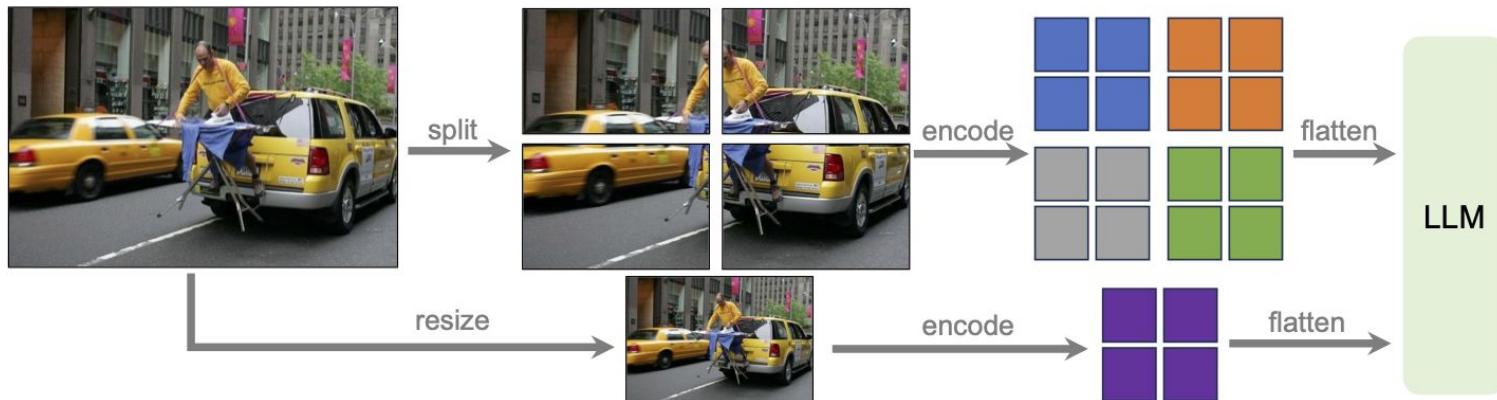
QWEN-VL

- Обучалась на 9 задач (в том числе OCR)
- Базовое разрешение 448x448 (в 4 раза больше пикселей). Для очень больших изображений — скользящее окно.

BLIP-2 (Мост между CLIP/BLIP и VLM)

- Обучается только небольшой Q-Former
- Используется 32 обучаемых эмбеддинги (image query)
- Учится на те же 3 задачи

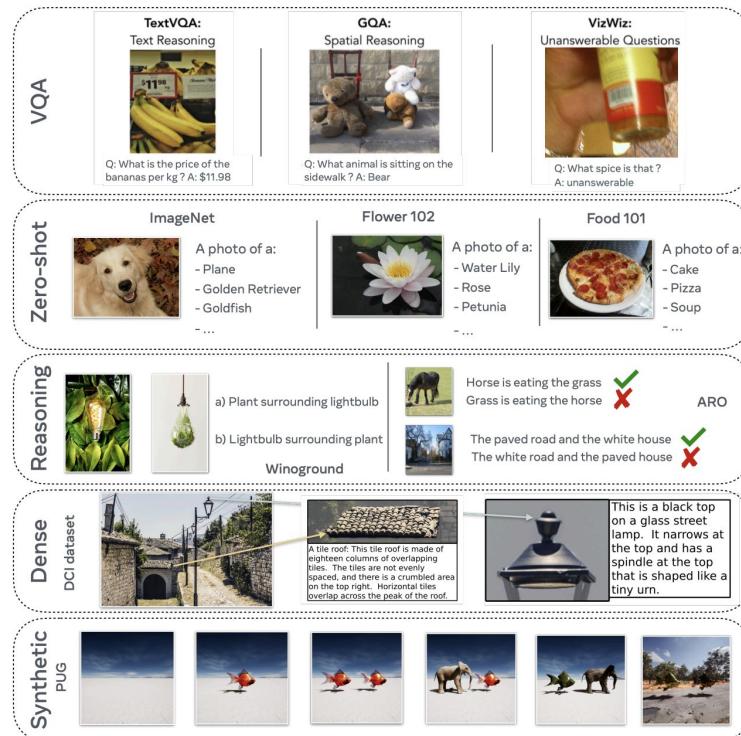
Обработка изображений (LLaVA-Next)



Вопросы?

Как оценивать работу VLM

- 1) BLEU (CiDER) в Image captioning
- 2) VQAScore, MMMU, GQA
- 3) Retrieval (Recall@1), Zero-shot classification
- 4) Object detection, segmentation



MMMU Benchmark Multi-discipline Multimodal Understanding and Reasoning

Art & Design	Business	Science																														
<p>Question: Among the following harmonic intervals, which one is constructed incorrectly?</p> <p>Options:</p> <ul style="list-style-type: none"> (A) Major third (B) Diminished fifth (C) Minor seventh (D) Diminished sixth 	<p>Question: ...The graph shown is compiled from data collected by Gallup . Find the probability that the selected Emotional Health Index Score is between 80.5 and 82?</p> <p>Options:</p> <ul style="list-style-type: none"> (A) 0 (B) 0.2142 (C) 0.3571 (D) 0.5 <table border="1"> <caption>Data from Emotional Health Index Score chart</caption> <thead> <tr> <th>Occupation Group</th> <th>Score</th> </tr> </thead> <tbody> <tr><td>Service</td><td>77.5</td></tr> <tr><td>Manufacturing or production</td><td>78.0</td></tr> <tr><td>Sales</td><td>78.5</td></tr> <tr><td>Clerical or office</td><td>79.0</td></tr> <tr><td>Information and media</td><td>79.5</td></tr> <tr><td>Construction or mining</td><td>80.0</td></tr> <tr><td>Manager</td><td>80.5</td></tr> <tr><td>Business owner</td><td>81.0</td></tr> <tr><td>Farmer</td><td>81.5</td></tr> <tr><td>Other</td><td>82.0</td></tr> <tr><td>Physician</td><td>82.5</td></tr> <tr><td>Farming, fishing, or forestry</td><td>83.0</td></tr> <tr><td>Teacher</td><td>83.5</td></tr> <tr><td>Physician</td><td>84.0</td></tr> </tbody> </table>	Occupation Group	Score	Service	77.5	Manufacturing or production	78.0	Sales	78.5	Clerical or office	79.0	Information and media	79.5	Construction or mining	80.0	Manager	80.5	Business owner	81.0	Farmer	81.5	Other	82.0	Physician	82.5	Farming, fishing, or forestry	83.0	Teacher	83.5	Physician	84.0	<p>Question: </p> <p>The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.</p> <p>Options:</p> <ul style="list-style-type: none"> (A) $\int_0^{1.5} [f(x) - g(x)]dx$ (B) $\int_0^{1.5} [g(x) - f(x)]dx$ (C) $\int_0^2 [f(x) - g(x)]dx$ (D) $\int_0^2 [g(x) - x(x)]dx$
Occupation Group	Score																															
Service	77.5																															
Manufacturing or production	78.0																															
Sales	78.5																															
Clerical or office	79.0																															
Information and media	79.5																															
Construction or mining	80.0																															
Manager	80.5																															
Business owner	81.0																															
Farmer	81.5																															
Other	82.0																															
Physician	82.5																															
Farming, fishing, or forestry	83.0																															
Teacher	83.5																															
Physician	84.0																															
<p>Subject: Music; Subfield: Music; Image Type: Sheet Music; Difficulty: Medium</p>	<p>Subject: Marketing; Subfield: Market Research; Image Type: Plots and Charts; Difficulty: Medium</p>	<p>Subject: Math; Subfield: Calculus; Image Type: Mathematical Notations; Difficulty: Easy</p>																														
Health & Medicine	Humanities & Social Science	Tech & Engineering																														
<p>Question: You are shown subtraction , T2 weighted and T1 weighted axial from a screening breast MRI. What is the etiology of the finding in the left breast?</p> <p>Options:</p> <ul style="list-style-type: none"> (A) Susceptibility artifact (B) Hematoma (C) Fat necrosis (D) Silicone granuloma 	<p>Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? </p> <p>Option:</p> <ul style="list-style-type: none"> (A) Oppressor (B) Imperialist (C) Savior (D) Isolationist 	<p>Question: Find the VCE for the circuit shown in </p> <p>Answer: <u>3.75</u></p> <p>Explanation: ...IE = [(V_{EE}) / (R_E)] = [(5 V) / (4 k-ohm)] = 1.25 mA; V_{CE} = V_{CC} - I_ER_L = 10 V - (1.25 mA) 5 k-ohm; V_{CE} = 10 V - 6.25 V = 3.75 V</p>																														
<p>Subject: Clinical Medicine; Subfield: Clinical Radiology; Image Type: Body Scans: MRI, CT,; Difficulty: Hard</p>	<p>Subject: History; Subfield: Modern History; Image Type: Comics and Cartoons; Difficulty: Easy</p>	<p>Subject: Electronics; Subfield: Analog electronics; Image Type: Diagrams; Difficulty: Hard</p>																														

Как выбрать модель?

The screenshot shows the 'Main Evaluation Results' section of the OpenVLM Main Leaderboard. It includes a list of metrics and a detailed description of the evaluation process, followed by a section for selecting evaluation dimensions.

Main Evaluation Results

- Metrics:
 - Avg Score: The average score on all VLM Benchmarks (normalized to 0 - 100, the higher the better).
 - Avg Rank: The average rank on all VLM Benchmarks (the lower the better).
 - Avg Score & Rank are calculated based on selected benchmark. When results for some selected benchmarks are missing, Avg Score / Rank will be None!!!
- By default, we present the overall evaluation results based on 8 VLM benchmarks, sorted by the descending order of Avg Score.
 - The following datasets are included in the main results: MMBench_V11, MMStar, MMMU_VAL, MathVista, OCRBench, AI2D, HallusionBench, MMVet.
 - Detailed evaluation results for each dataset (included or not included in main) are provided in the consequent tabs.

Evaluation Dimension

Selected Dimensions (checked): Avg Score, Avg Rank, MMBench_V11, MMStar, MMMU_VAL, MathVista, OCRBench, AI2D, HallusionBench.

Available Dimensions (unchecked): SEEDBench_IMG, MMVet, LLavaBench, CCBench, RealWorldQA, POPE, ScienceQA_TEST, SEEDBench2_Plus, MMT-Bench_VAL, BLINK.

https://huggingface.co/spaces/opencompass/open_vlm_leaderboard

<https://lmarena.ai/?leaderboard>

Как выбрать модель?

Выбираем модель в соответствии решаемой задачей и ограничениями:

- Хорошая модель, но иногда китайская – Qwen-VL
- Что-то быстро потестить – ChatGPT
- Нужна модель поменьше – Florence

https://huggingface.co/spaces/opencompass/open_vlm_leaderboard

<https://lmarena.ai/?leaderboard>

Summary

- Мультимодальность возникает позволяет более качественное описание объекта, а также решать задачи в различных доменах
- LLM – основа для любой foundational model за счет ее способности к обобщению и объему знаний.
- Построение VLM состоит из двух этапов: pretraining (image captioning, interleaved) и SFT (дообучение на конкретные задачи)
- Для выбора модели смотрите на задачу и её вводные и смотрите на арену.