Мультимодальные модели
Преподаватель: Филатов Андрей

Андрей Филатов

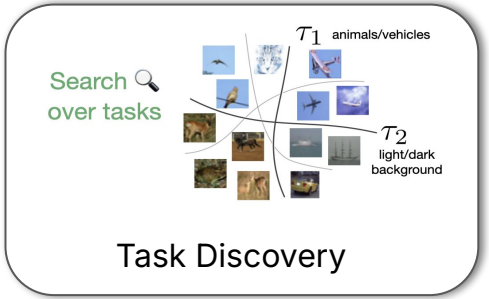
Опыт работы

- AIRI
  Октябрь 2022 - ...
  Руководитель направления по исследованию данных
  Диффузионные модели Kandinsky + языковые модели для транзакций
- МФТИ
  Октябрь 2022 - ...
  Лектор курса по Deep Learning
- Samsung
  Июнь 2022 - Июль 2023
  Стажер
- Ecole Polytechnique Federale de Lausaunne (EPFL)
  Март 2021 - Январь 2023
  Стажер

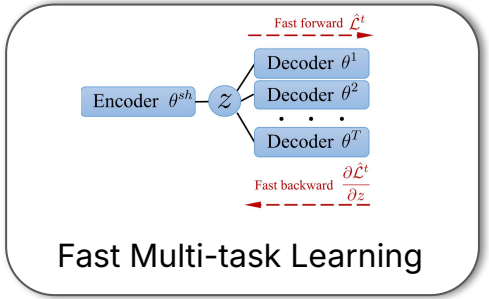
Образование

- Skoltech
  Сентябрь 2023
- МФТИ ФПМИ + Skoltech с отличием
  Сентябрь 2021 - июнь 2023
- МФТИ ФПМИ с отличием
  Сентябрь 2017 - июнь 2021

Избранные проекты

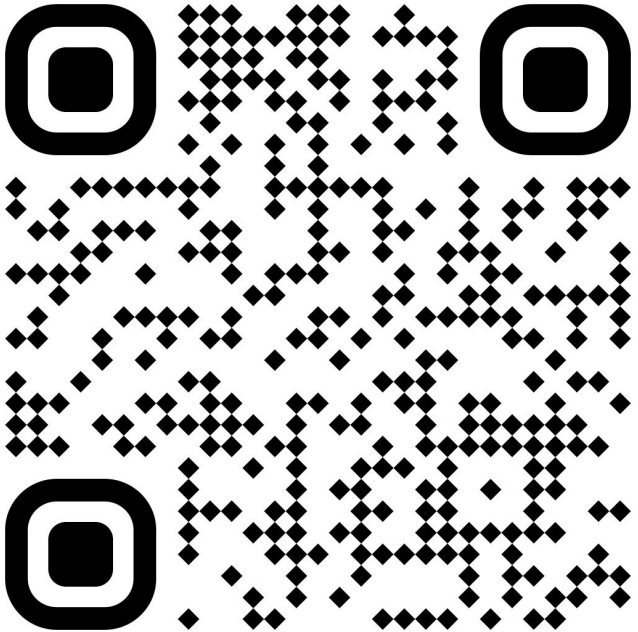


Task Discovery



Fast Multi-task Learning

Писать сюда —>



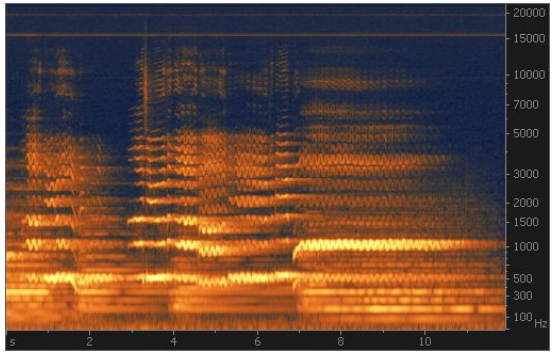
План занятия

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

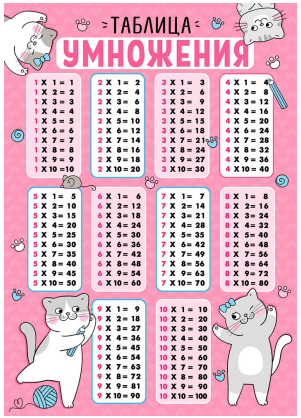
Мультимодальный домен

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

Изображения



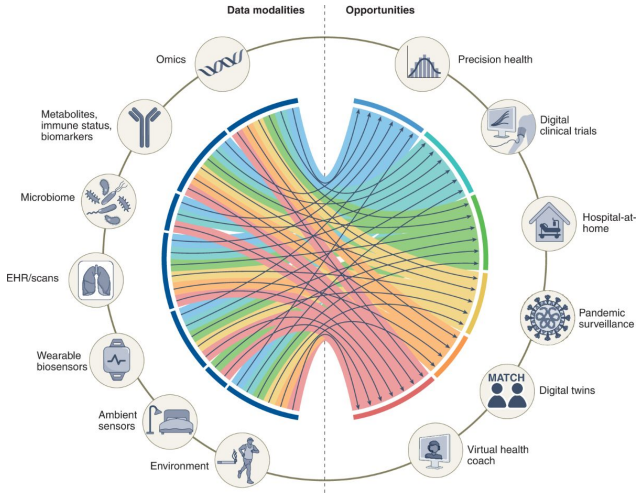
Аудио



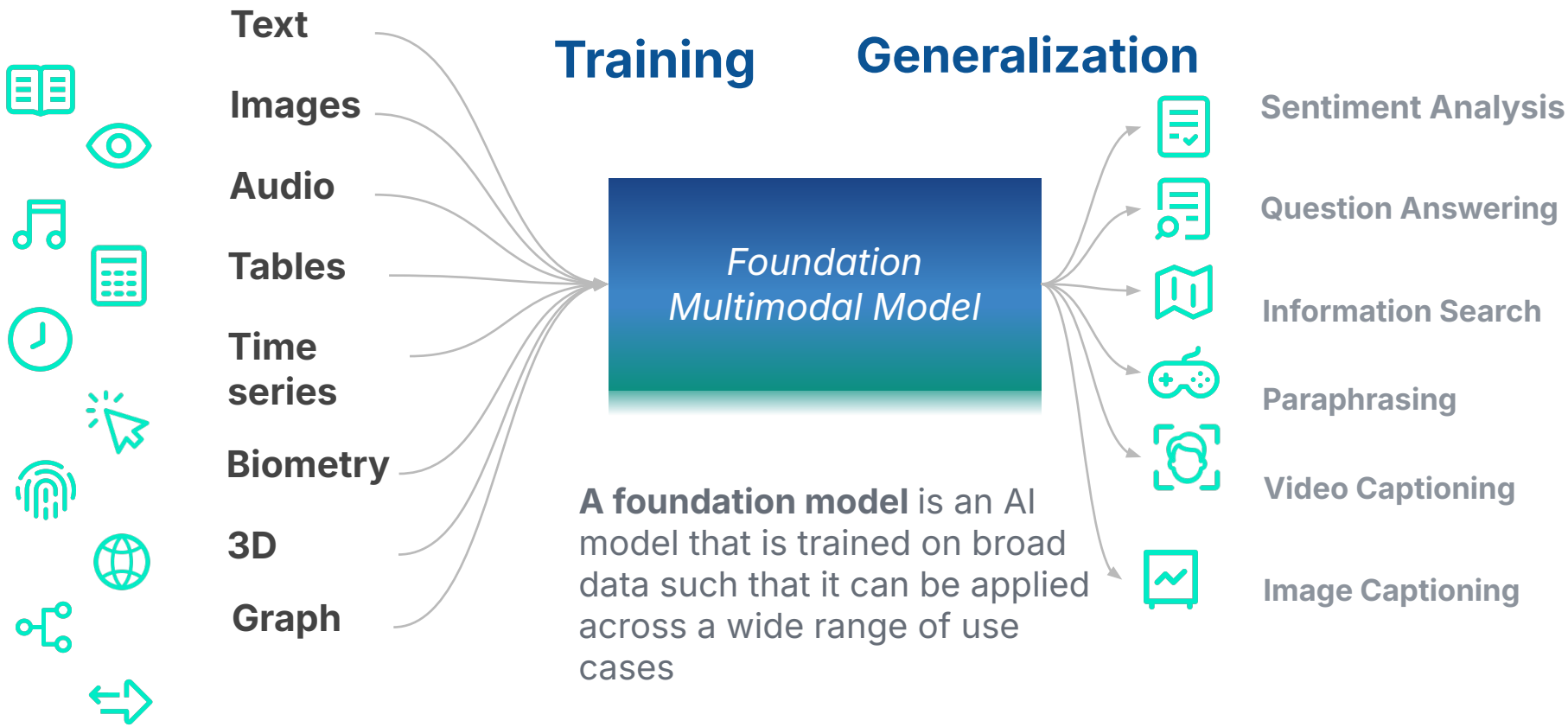
Таблицы/временные ряды

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



A foundation model is an AI model that is trained on broad data such that it can be applied across a wide range of use cases

При чем 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.

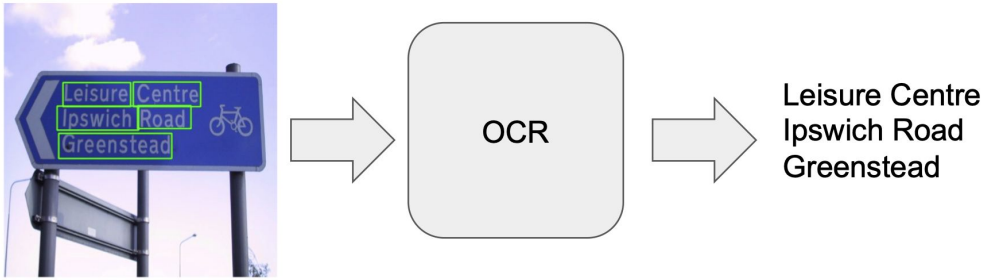
Вопросы?

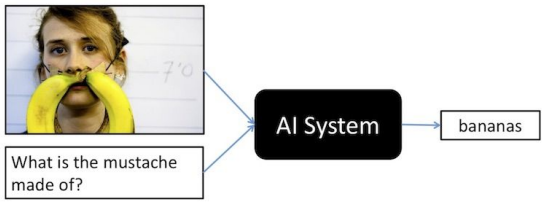
Визуально-текстовая мультимодальность

Одна из самых мультимодальностей – картиночно-текстовая

Текущие задачи:

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

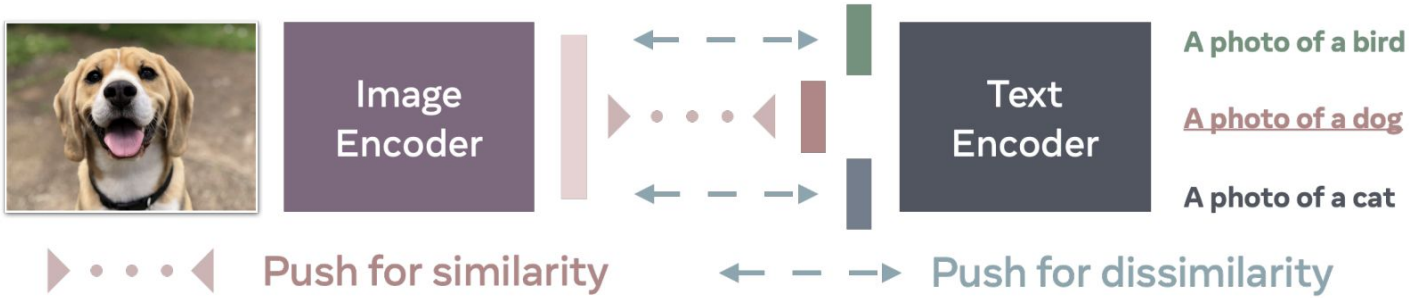




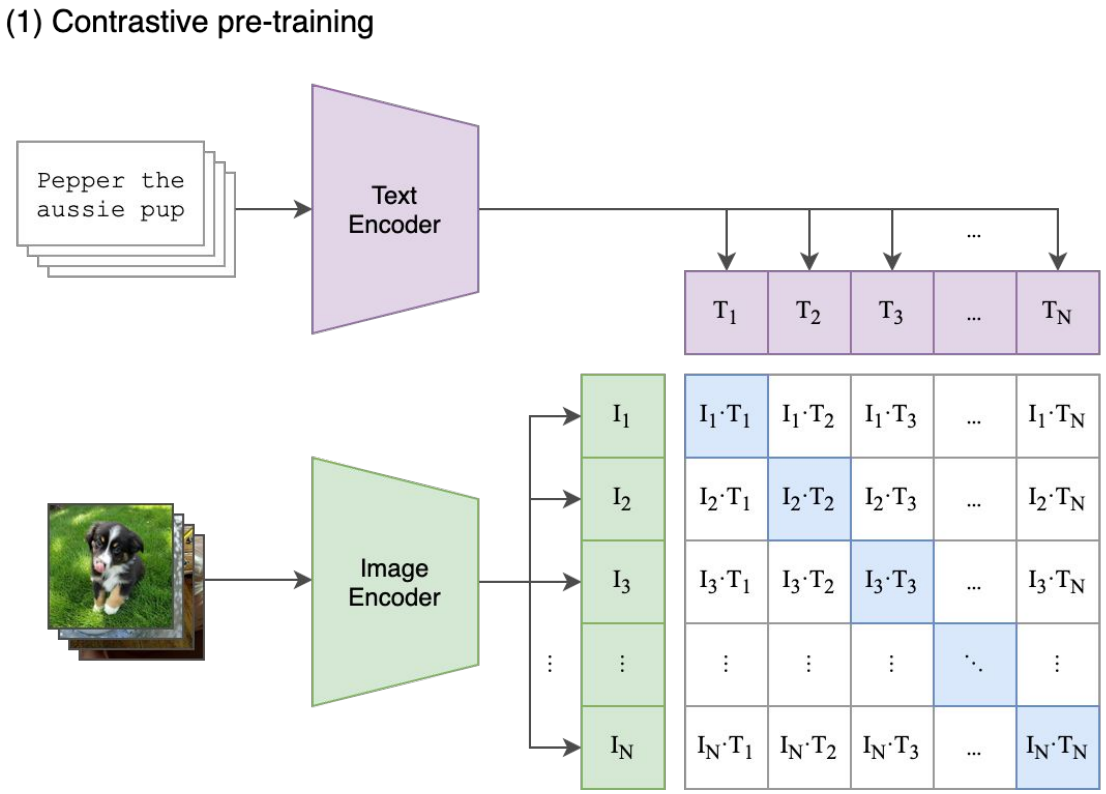
Идея CLIP

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

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



CLIP



Learning Transferable Visual Models From Natural Language Supervision, 2021

https://openai.com/research/clip

https://github.com/openai/CLIP

CLIP: датасет

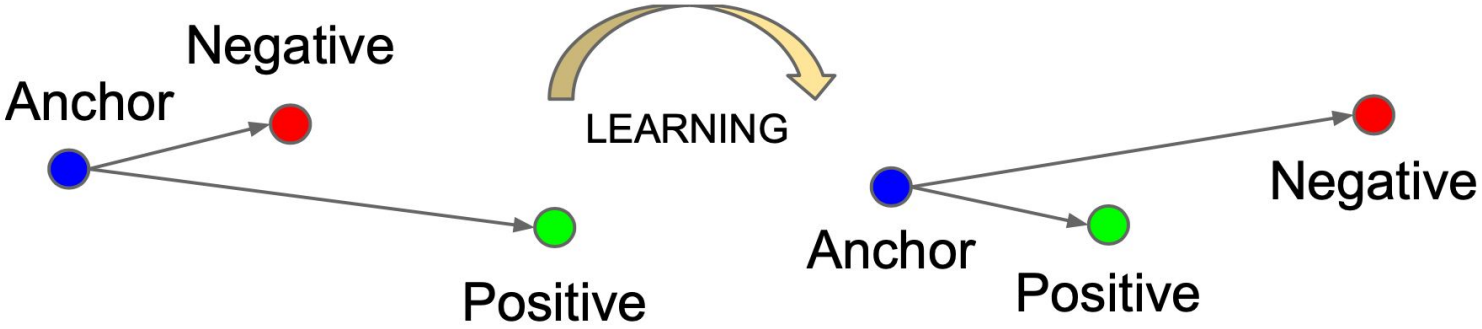
Рыжий кот прыгнул за цветными нитками растопырив лапы

Кот-толстяк пытается слезть с дерева

Просто кот

400 миллионов пар <картинка>:<текст>

CLIP: сетап


Negative
Anchor
Positive
LEARNING
Negative
Anchor
Positive

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;

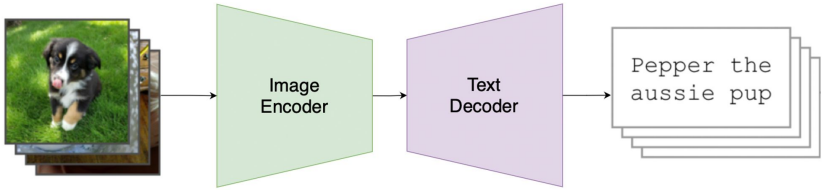
4. складываем эти 2 лосса и делим сумму на 2

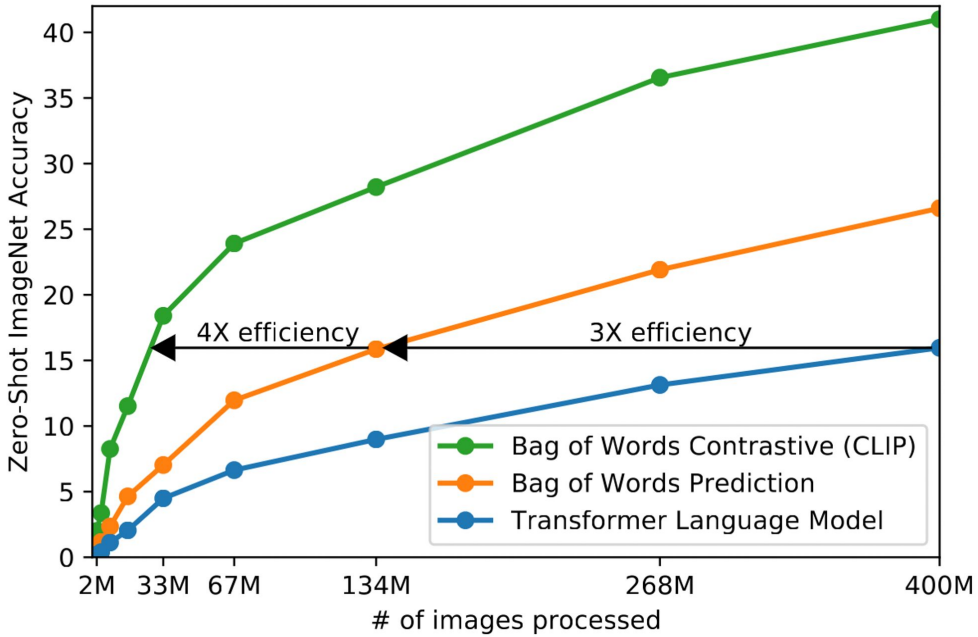
Сетап обучения:

- 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

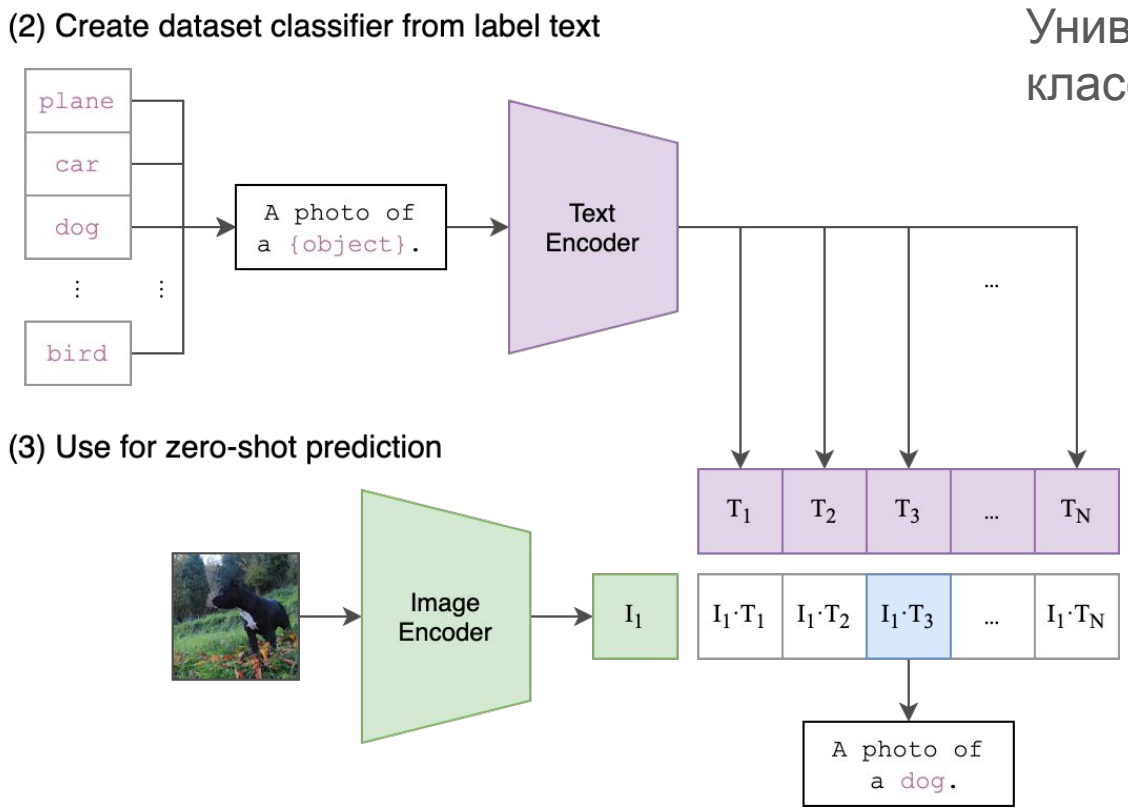




Можно посмотреть на это с такой стороны: **описания картинок из интернета** иногда очень сложно восстановить при помощи одной картинки. Восстановить только **ключевые слова без предлогов** — задача более простая. Но еще проще было бы **соотнести описания и картинки** и найти между ними соответствие.

CLIP

CLIP: zero-shot

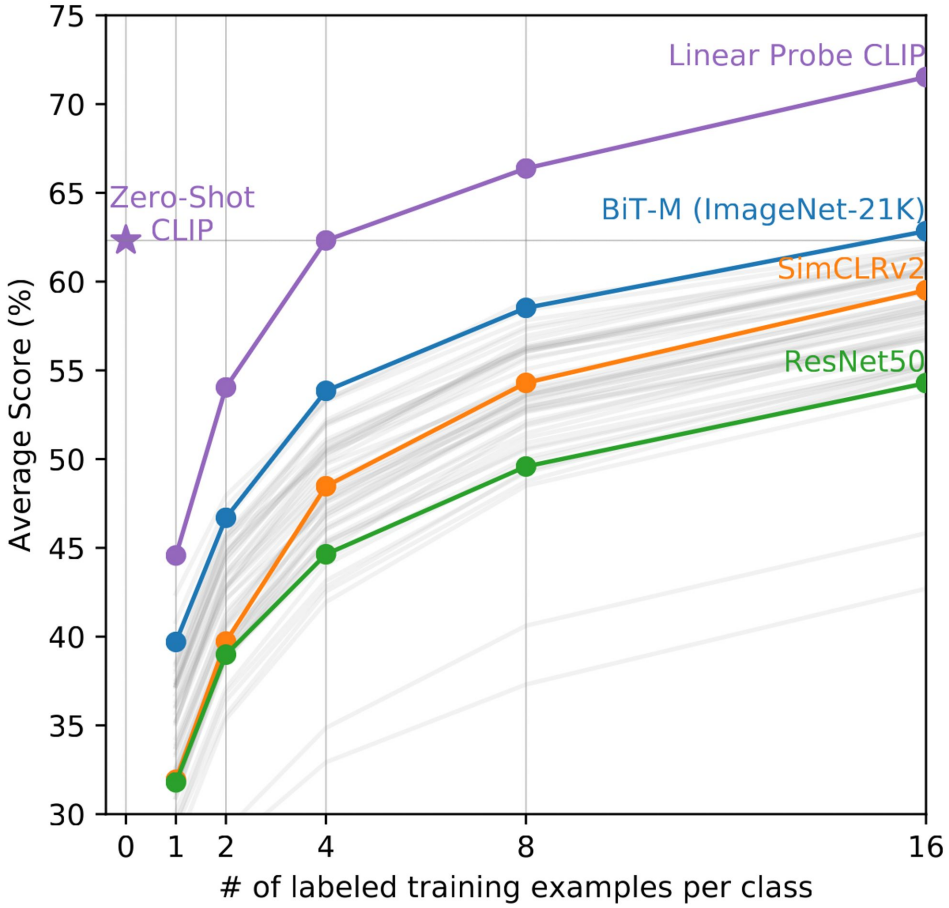


Универсальный классификатор

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.



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	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-G/14	LAION-2B	224px	34B	80.1%
ViT-L/14	OpenAI's WIT	224px	13B	75.5%

Learning Transferable Visual Models From Natural Language Supervision, 2021
https://openai.com/research/clip
https://github.com/openai/CLIP

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.

Вопросы?

Как решали задачу Image Captioning раньше?

Deep Visual-Semantic Alignments for Generating Image Descriptions

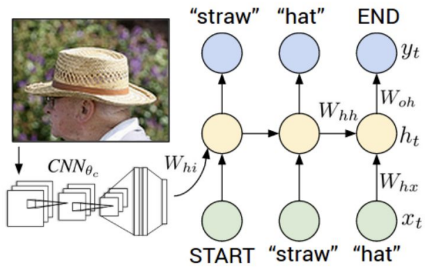


Figure 4. Diagram of our multimodal Recurrent Neural Network generative model. The RNN takes a word, the context from previous time steps and defines a distribution over the next word in the sentence. The RNN is conditioned on the image information at the first time step. START and END are special tokens.

Show-attend and tell

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.)

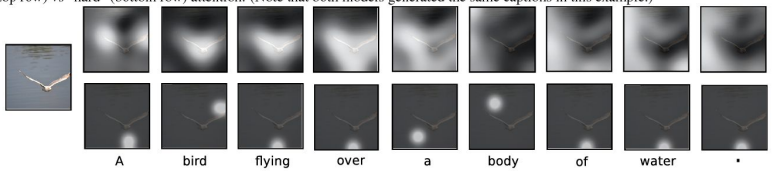
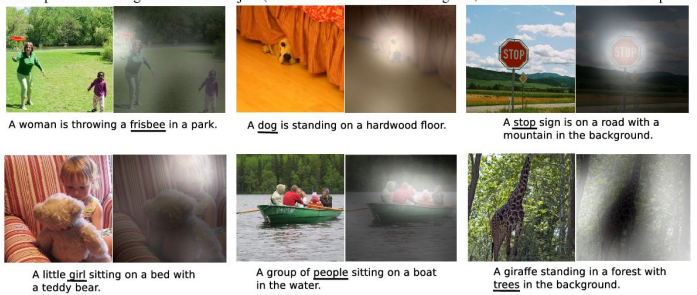


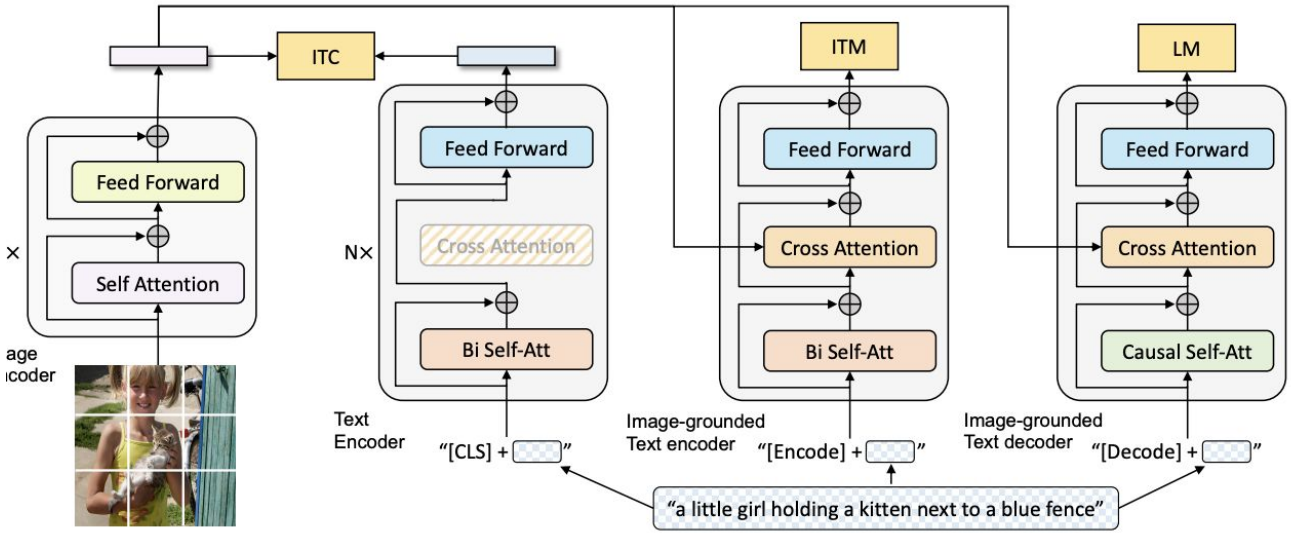
Figure 3. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)



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

Идея BLIP

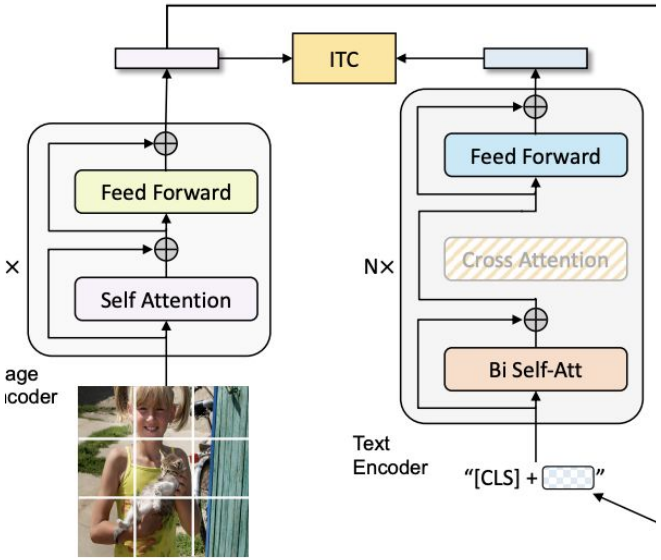
Новый взгляд на идею того, как делать image captioning, используя механизм внимания.



BLIP

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

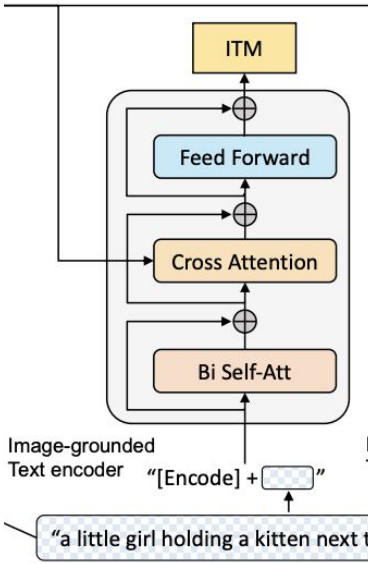
Таким образом, модель выучивает базовое представление между картинками и текстом.



BLIP

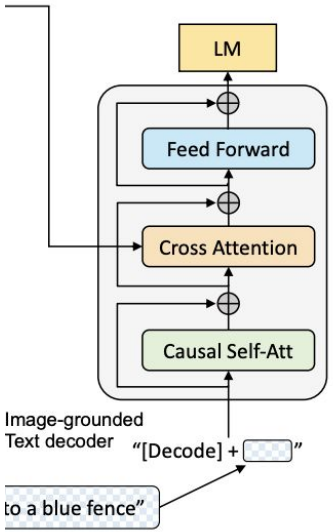
Второй этап – научить модель использовать через cross-attention эти признаки как в старых моделях

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



BLIP

Финальный этап – делаем captioning учим классическую авторегрессию как в LLM



Вопросы?

BLIP: данные

Самое главное всегда и везде будут данные.

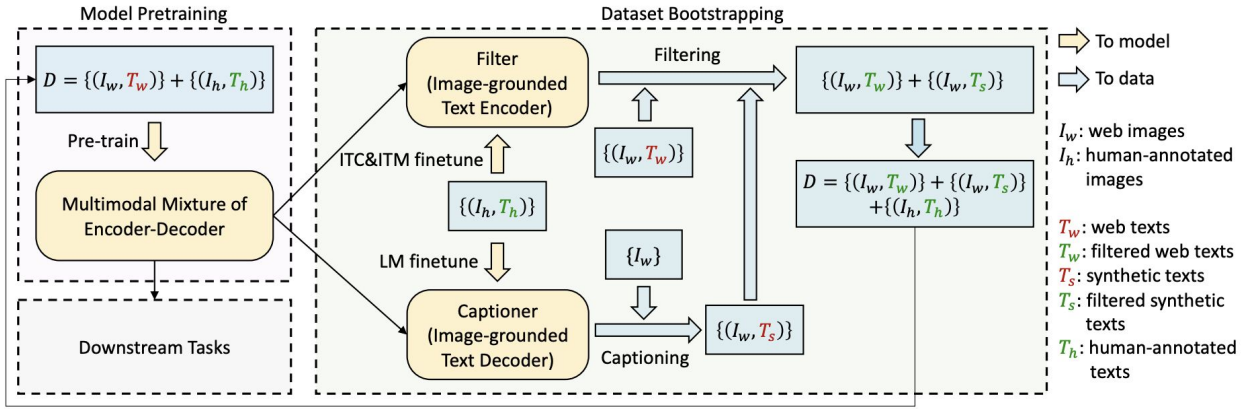


Figure 3. Learning framework of BLIP. We introduce a captioner to produce synthetic captions for web images, and a filter to remove noisy image-text pairs. The captioner and filter are initialized from the same pre-trained model and finetuned individually on a small-scale human-annotated dataset. The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

BLIP: summary

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

Вопросы?

FRoMAGE

Одного линейного слоя достаточно, чтобы научить модель в текст

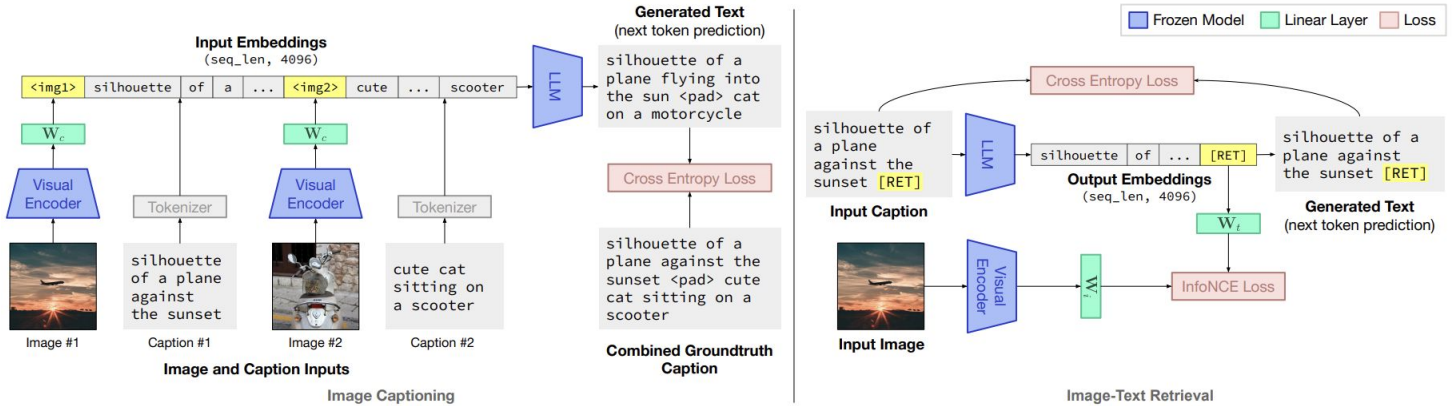


Figure 2. Overview of the FROMAGe architecture. FROMAGe is a model trained on image-text pairs for image captioning and image-text retrieval. It is capable of processing arbitrarily interleaved image and text inputs, and producing interleaved images and text as outputs.

FRoMAGE: возможности

FRoMAGE

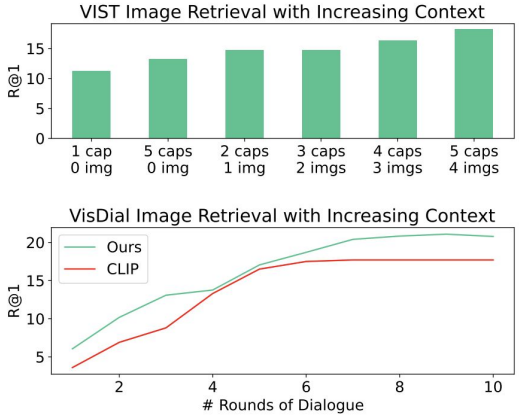
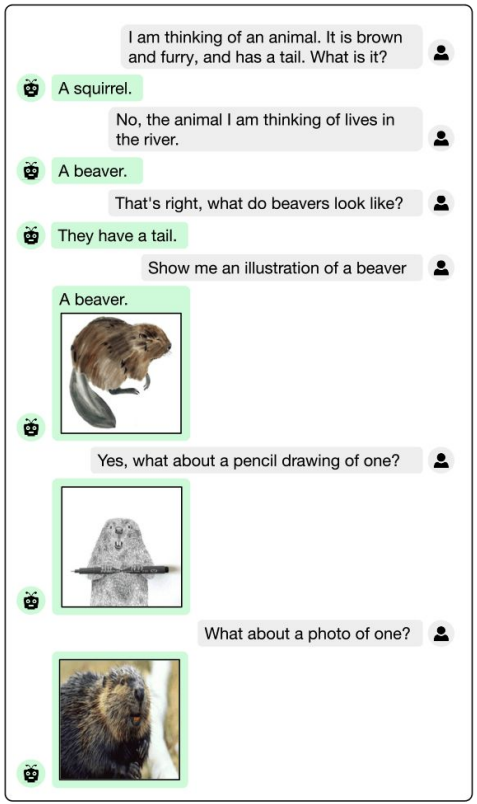
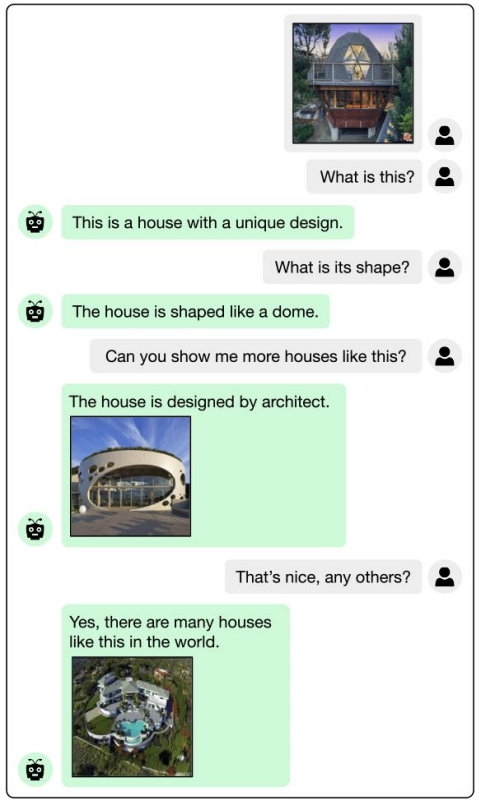


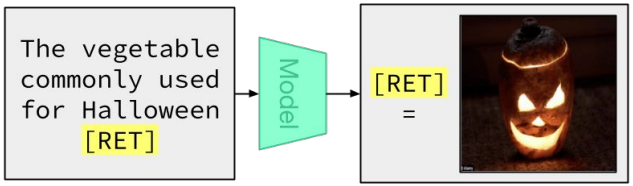
Figure 5. Increasing input context generally improves performance. Shown are results for image retrieval for VIST (Huang et al., 2016) (top) and image retrieval on VisDial (Das et al., 2017) (bottom).

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.


The vegetable commonly used for Halloween
[RET]
Model
[RET]
=

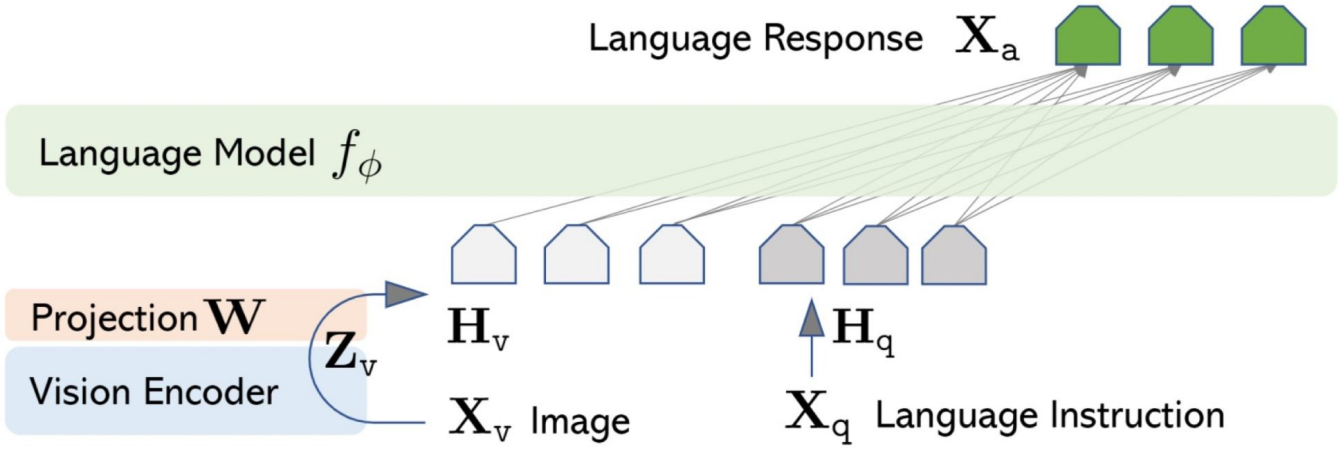
FRoMAGE: summary

1) Простого линейного слоя достаточно для captioning! -> очень быстро обучается
2) Модель может обобщаться на полноценные диалоги и делать сложные задачи

Вопросы?

Идея LLaVA

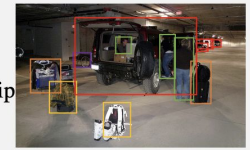
LlaVa



Идея 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. Chicora 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.", ...]

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.", ...]

Сбор датасета

Fine-tuning End-to-End

- We develop a Chatbot by fine-tuning on the 158K language-image instruction-following data in Section 3
- Science QA. We study our method on the ScienceQA benchmark [34], the first large-scale multimodal science question dataset that annotates the answers with detailed lectures and explanations

LLaVA: example

Visual input example, Extreme Ironing:	
	
	Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg
User	What is unusual about this image?
LLaVA	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*]
User	What's happening in the scene?
LLaVA	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	What is unusual about this image?
GPT-4 [36]	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	What is unusual about this image?
BLIP-2	a man is sitting on the back of a yellow cab
User	What is unusual about this image?
OpenFlamingo	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

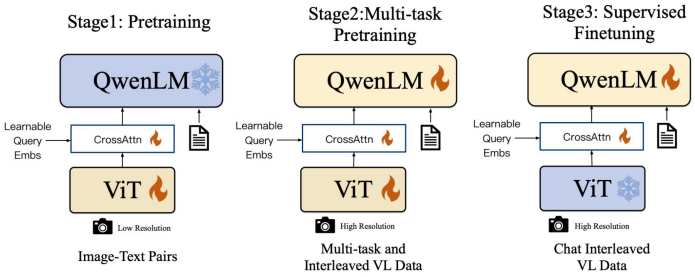
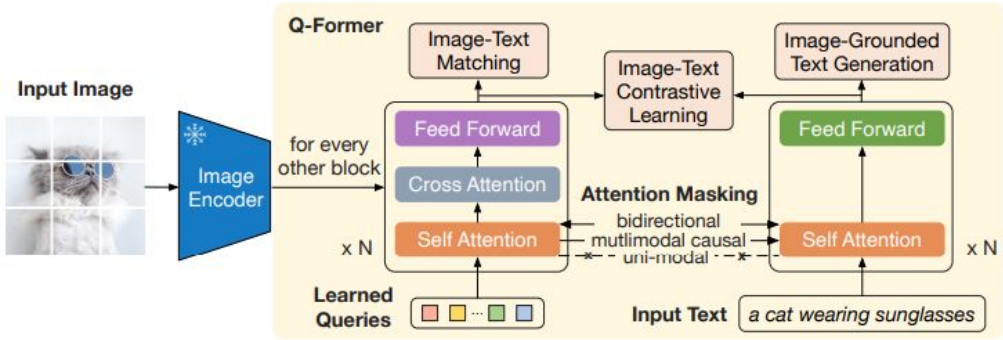


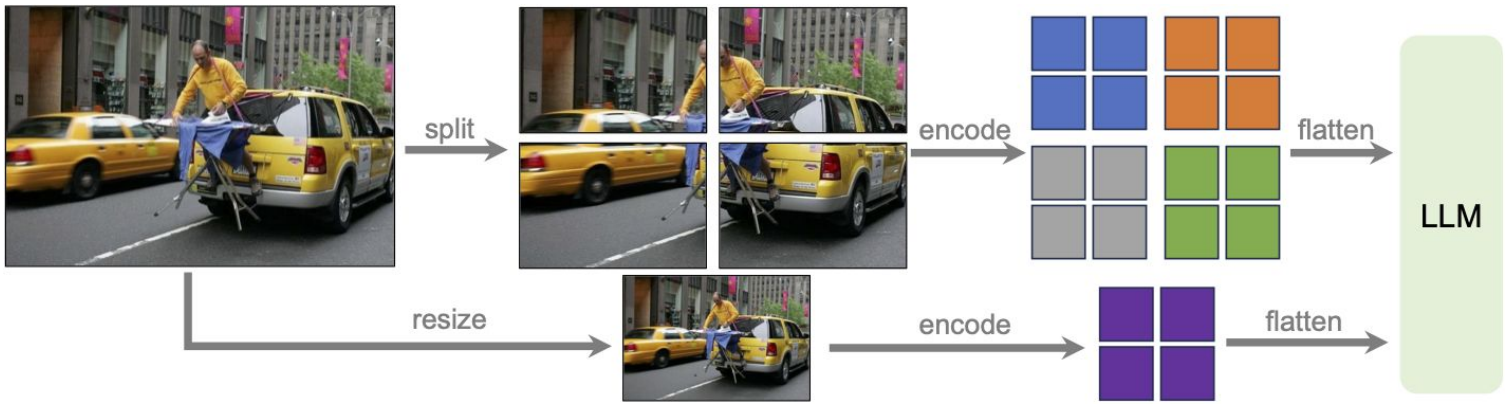
Figure 3: The training pipeline of the Qwen-VL series.

QWEN-VL



BLIP-2 architecture

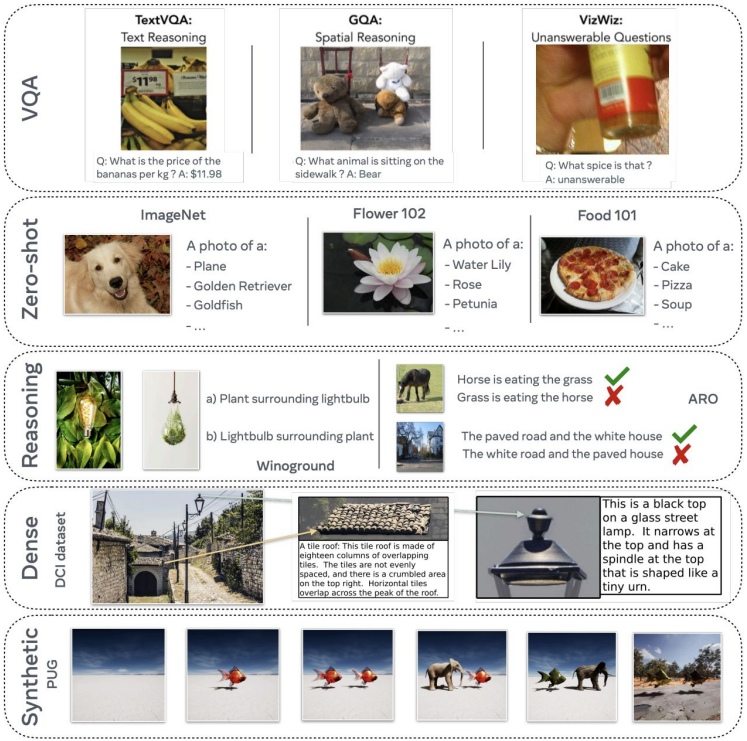
Обработка изображений (LLaVA-Next)



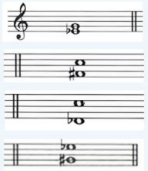
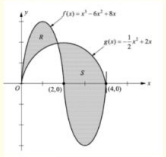
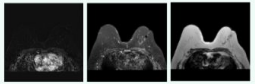
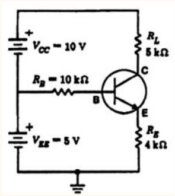
Вопросы?

Как оценивать работу VLM

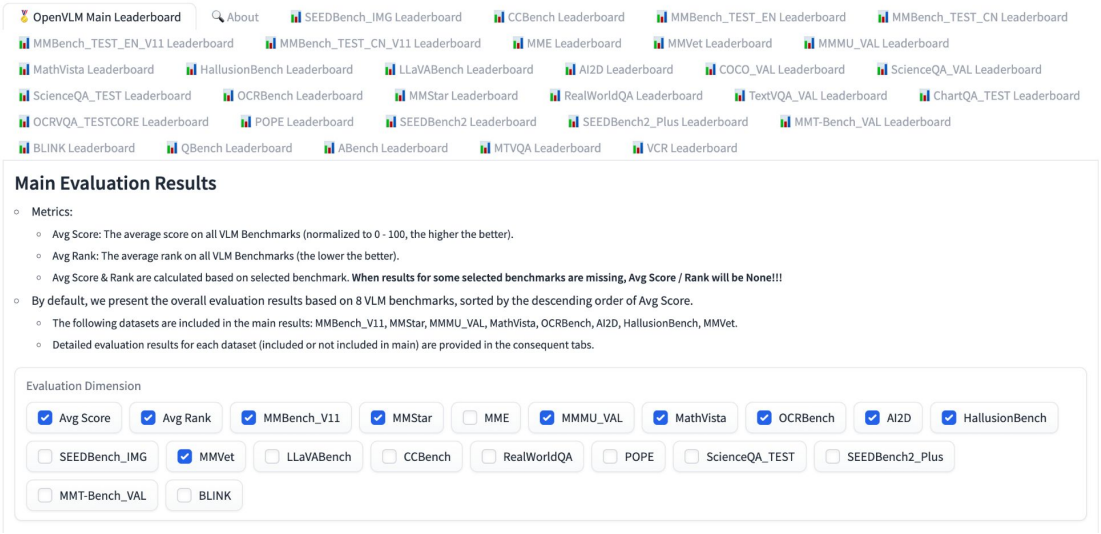
1) BLEU (CiDER) в Image captioning
2) VQAScore, MMMU, GQA
3) Retrieval (Recall@1), Zero-shot classification
4) Object detection, segmentation



MMMU Benchmark

Art & Design	Business	Science
Question: Among the following harmonic intervals, which one is constructed incorrectly? **Options:** (A) Major third <*image 1*> (B) Diminished fifth <*image 2*> **(C) Minor seventh <*image 3*>** (D) Diminished sixth <*image 4*> 	**Question:** ...The graph shown is compiled from data collected by Gallup <*image 1*>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82? **Options:** (A) 0 (B) 0.2142 **(C) 0.3571** (D) 0.5 	**Question:** <*image 1*> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R. **Options:** **(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) $\int_0^{2}[g(x)-x(x)]dx$ 
Subject: Music; **Subfield:** Music; **Image Type:** Sheet Music; **Difficulty:** Medium	**Subject:** Marketing; **Subfield:** Market Research; **Image Type:** Plots and Charts; **Difficulty:** Medium	**Subject:** Math; **Subfield:** Calculus; **Image Type:** Mathematical Notations; **Difficulty:** Easy
Health & Medicine	**Humanities & Social Science**	**Tech & Engineering**
Question: You are shown subtraction <*image 1*>, T2 weighted <*image 2*> and T1 weighted axial <*image 3*> from a screening breast MRI. What is the etiology of the finding in the left breast? **Options:** (A) Susceptibility artifact (B) Hematoma **(C) Fat necrosis** (D) Silicone granuloma 	**Question:** In the political cartoon, the United States is seen as fulfilling which of the following roles? <*image* 1> **Option**: (A) Oppressor (B) Imperialist **(C) Savior** (D) Isolationist 	**Question:** Find the VCE for the circuit shown in <*image 1*>. Neglect VBE **Answer:** 3.75 **Explanation**: ...IE = [(VEE) / (RE)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - IERL = 10 V - (1.25 mA) 5 k-ohm; VCE = 10 V - 6.25 V = 3.75 V 
Subject: Clinical Medicine; **Subfield:** Clinical Radiology; **Image Type:** Body Scans: MRI, CT.; **Difficulty:** Hard	**Subject:** History; **Subfield:** Modern History; **Image Type:** Comics and Cartoons; **Difficulty:** Easy	**Subject:** Electronics; **Subfield:** Analog electronics; **Image Type:** Diagrams; **Difficulty:** Hard

Как выбрать модель?



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
- Что-то быстро потестить – ChatGPT-4o
- Нужна модель модель поменьше – PaliGemma, 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 счёт ее способности к обобщению и объему знаний.
- Построение VLM состоит из двух этапов: pretraining (image captioning, interleavead) и SFT (дообучение на конкретные задачи)
- Для выбора модели смотрите на задачу и её вводные и смотрите на арену.