



Краткое введение в Большие языковые модели

Евгений Борисов

Большие языковые модели

LLM - большая языковая модель, содержит большое количество параметров

GPT-3 (OpenAI, 2020) - 175 миллиардов параметров

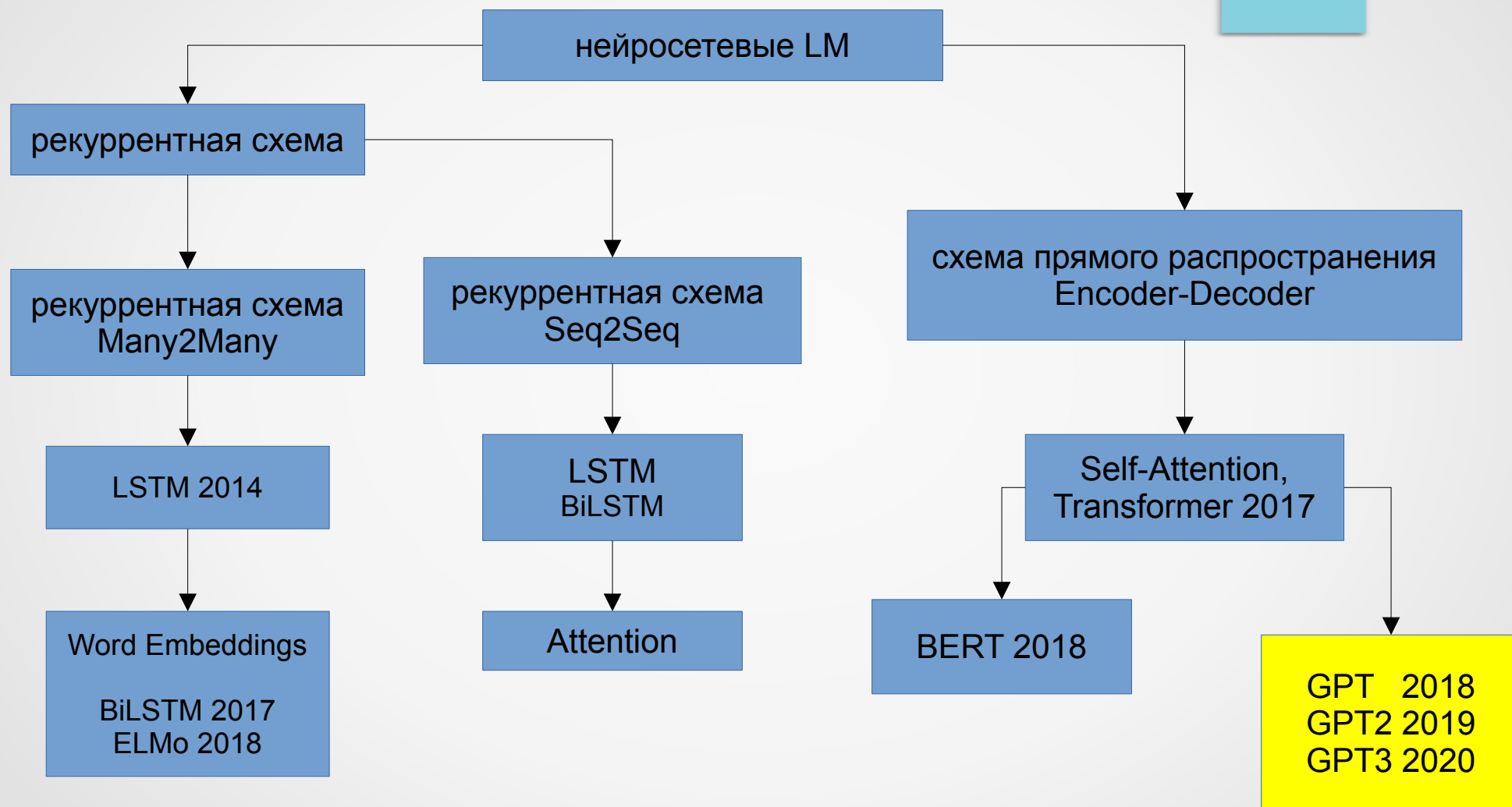
LlaMA (Meta AI, 2023) - 7 до 65 миллиардов параметров (140 GB)

LMSYS Chatbot Arena Leaderboard

<https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard>

Rank ▲	Model ▲	★ Arena Elo ▲	▢ 95% CI ▲	▢ Votes ▲	Organization
1	Claude 3 Opus	1253	+5/-5	33250	Anthropic
1	GPT-4-1106-preview	1251	+4/-4	54141	OpenAI
1	GPT-4-0125-preview	1248	+4/-4	34825	OpenAI
4	Bard (Gemini Pro)	1203	+5/-7	12476	Google
4	Claude 3 Sonnet	1198	+5/-5	32761	Anthropic
6	GPT-4-0314	1185	+5/-4	33499	OpenAI
6	Claude 3 Haiku	1179	+5/-5	18776	Anthropic
8	GPT-4-0613	1158	+4/-5	51860	OpenAI
8	Mistral-Large-2402	1157	+5/-4	26734	Mistral
9	Qwen1.5-72B-Chat	1148	+5/-5	20211	Alibaba
10	Claude-1	1146	+6/-6	21908	Anthropic
10	Mistral-Medium	1145	+5/-4	26196	Mistral

Нейросетевые языковые модели

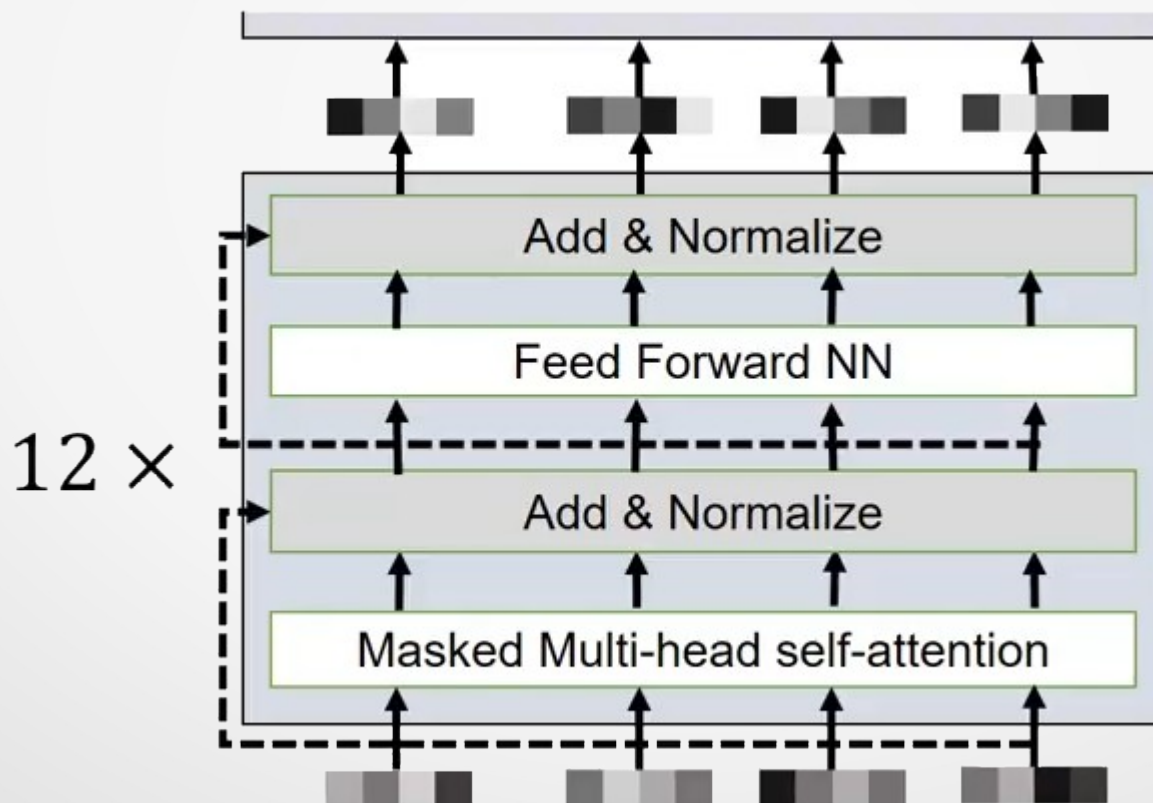


Нейросетевые языковые модели

Модели GPT (Generative Pre-Training)

Radford A. et al. Improving language understanding by generative pre-training. – 2018

Языковая модель, основанная на блоке Decoder модели Transformer, блок Encoder-Decoder-Attention выкидываем.



Нейросетевые языковые модели

Модели GPT (Generative Pre-Training)

Radford A. et al. Improving language understanding by generative pre-training. – 2018

Схема обучения модели на основе GPT

- Предобучение на корпусе \mathcal{U} (без учителя)

$$\begin{aligned}h_0 &= UW_e + W_p; \\h_i &= \text{transformer}(h_{i-1}), i = \overline{1, n} \\P(u) &= \text{softmax}(h_n W_e^T)\end{aligned}$$

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \dots, u_{i-1}; \Theta)$$

- Донастройка под целевую задачу на корпусе \mathcal{C} (с учителем)

$$P(y | x^1, \dots, x^m) = \text{softmax}(h_l^m W_y)$$

$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y | x^1, \dots, x^m)$$

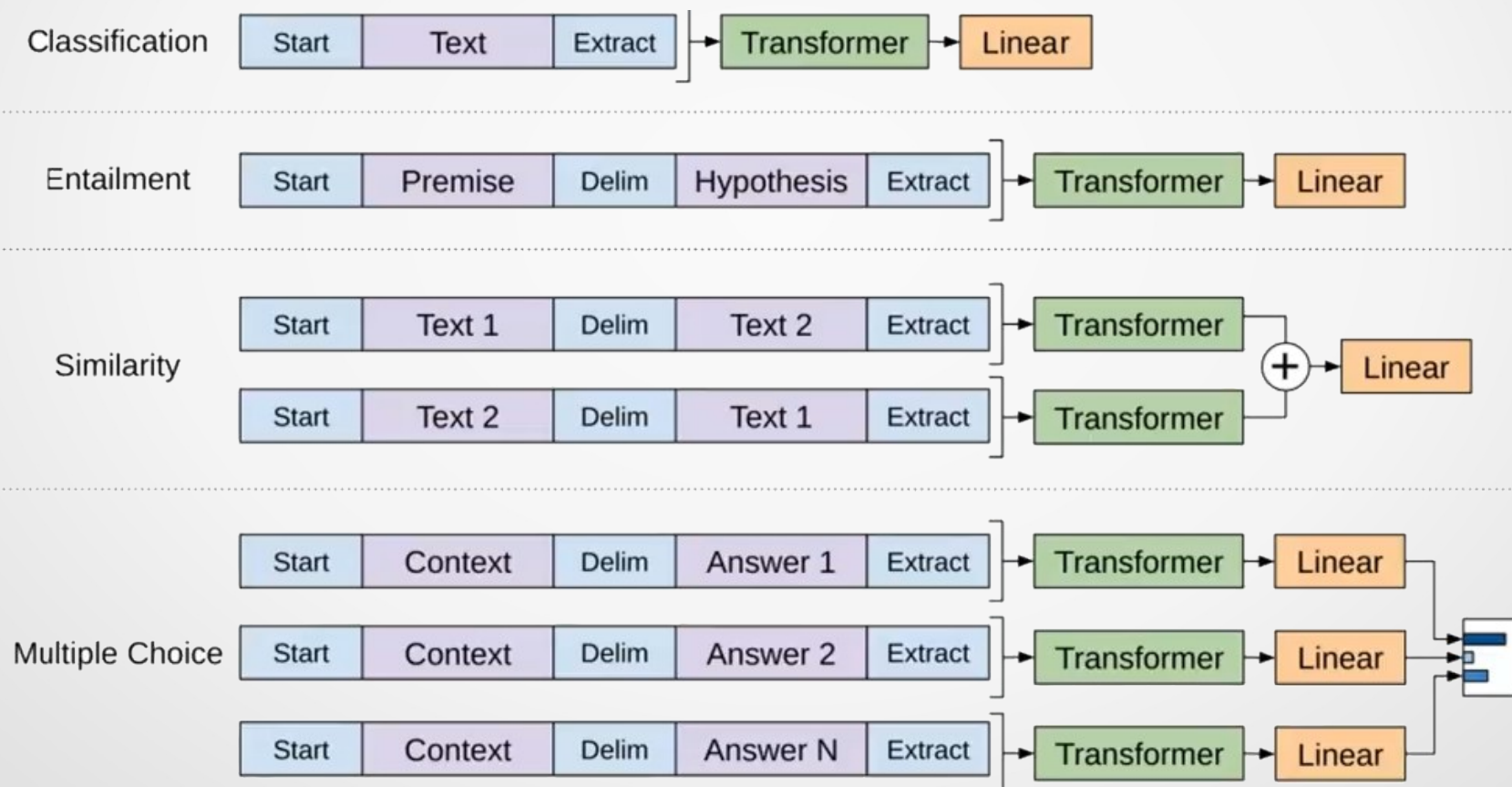
$$L(\mathcal{C}) = L_2(\mathcal{C}) + \lambda * L_1(\mathcal{C});$$

Нейросетевые языковые модели

Модели GPT (Generative Pre-Training)

Radford A. et al. Improving language understanding by generative pre-training. – 2018

Различные варианты применения GPT



Нейросетевые языковые модели

Модели GPT (Generative Pre-Training)

GPT: Radford A. et al. Improving language understanding by generative pre-training. – 2018
- 12 слоев трансформера

GPT-2: Radford A. et al. Language models are unsupervised multitask learners – 2019
- 48 слоев трансформера
- Больше корпус для обучения (40 GB текста)
- Другая токенизация: BPE по байтам, а не по символам
- передвинут Layer normalization
- изменена инициализация

GPT-3: Brown T. B. et al. Language models are few-shot learners – 2020.
(почти не отличается от GPT-2)
- 96 слоев трансформера
- Еще больше корпус для обучения (570 GB текста)
- Еще больше контекст (2048 токенов)

Большие языковые модели

для обучения LLM с нуля требуются большие ресурсы

датасет из текстов 3ТВ

кластер из 200 GPU

процесс может занимать несколько недель

Большие языковые модели

платные API и предобученные модели в открытом доступе

HuggingFace <https://huggingface.co>

The screenshot displays the Hugging Face homepage. At the top, there's a navigation bar with links to Models, Datasets, Spaces, Posts, Docs, Solutions, Pricing, and Log In. A search bar is also present. Below the navigation bar, the left sidebar shows various task categories like Multimodal, Computer Vision, and Natural Language Processing. The main content area is titled 'Models 70,402' and features a grid of model cards. Each card includes the model name, its capabilities, and statistics like the number of downloads and likes. The models listed include databricks/dbrx-instruct, xai-org/grok-1, mistralai/Mistral-7B-Instruct-v0.2, google/gemma-7b, meta-llama/Llama-2-7b-chat-hf, stabilityai/stable-code-instruct-3b, Qwen/Qwen1.5-MoE-A2.7B, CohereForAI/c4ai-command-r-v01, mistralai/Mistral-7B-v0.1, ai21labs/Jamba-v0.1, databricks/dbrx-base, Nexusflow/Starling-LM-7B-beta, alpindale/Mistral-7B-v0.2-hf, mistralai/Mixtral-8x7B-Instruct-v0.1, NousResearch/Hermes-2-Pro-Mistral-7B, hpcai-tech/grok-1, meta-llama/Llama-2-7b, and mlabonne/Beyonder-4x7B-v3.

Hugging Face Search models, datasets, users...

Models 70,402 Filter by name

Tasks Libraries Datasets Languages Licenses Other

Filter Tasks by name Reset Tasks

Multimodal

- Image-Text-to-Text
- Visual Question Answering
- Document Question Answering

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Text-to-Image
- Image-to-Text
- Image-to-Image
- Image-to-Video
- Unconditional Image Generation
- Video Classification
- Text-to-Video
- Zero-Shot Image Classification
- Mask Generation
- Zero-Shot Object Detection
- Text-to-3D
- Image-to-3D
- Image Feature Extraction

Natural Language Processing

- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Feature Extraction
- Text Generation
- Text2Text Generation
- Fill-Mask
- Sentence Similarity

Models

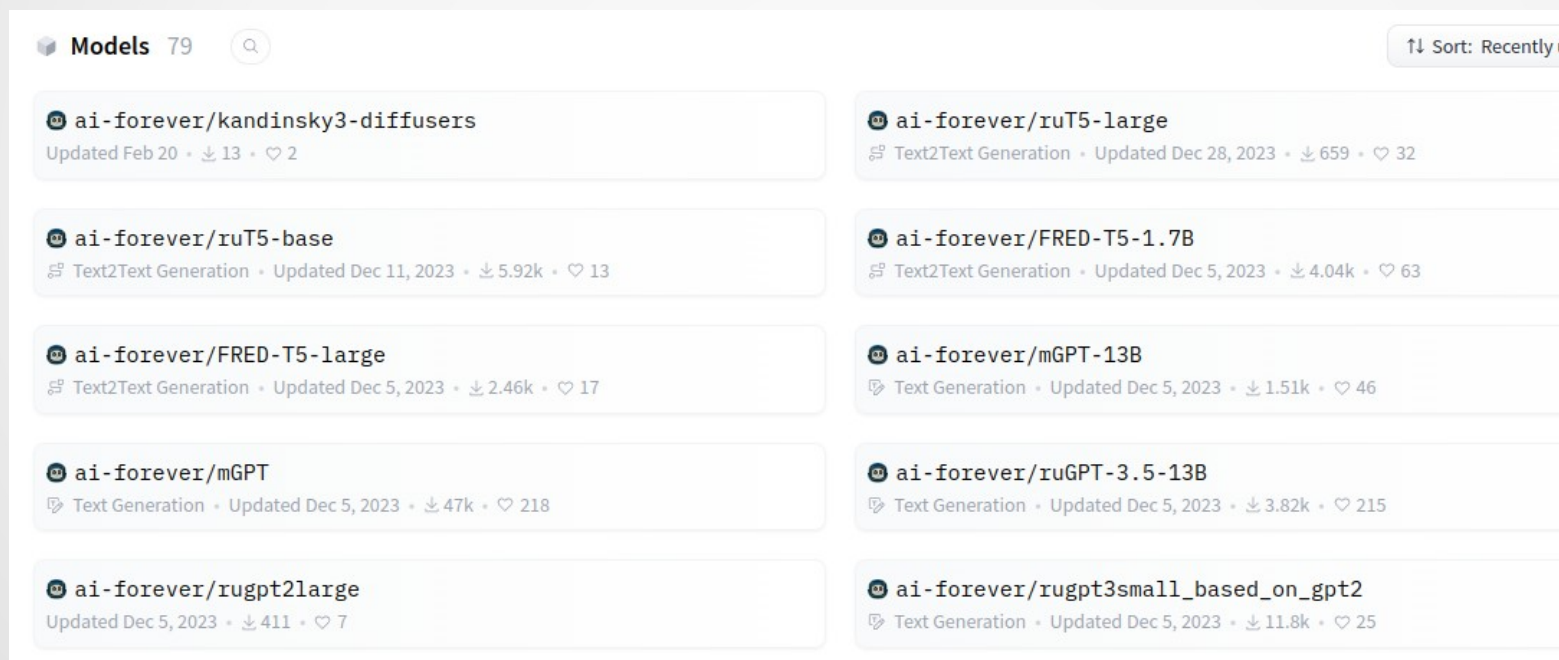
- databricks/dbrx-instruct**
Text Generation • Updated about 16 hours ago • 9.36k • 529
- xai-org/grok-1**
Text Generation • Updated about 18 hours ago • 31.3k • 1.8k
- mistralai/Mistral-7B-Instruct-v0.2**
Text Generation • Updated 5 days ago • 2.09M • 1.49k
- google/gemma-7b**
Text Generation • Updated 30 days ago • 163k • 2.65k
- meta-llama/Llama-2-7b-chat-hf**
Text Generation • Updated 10 days ago • 1.36M • 3.22k
- stabilityai/stable-code-instruct-3b**
Text Generation • Updated 1 day ago • 1.76k • 82
- Qwen/Qwen1.5-MoE-A2.7B**
Text Generation • Updated about 20 hours ago • 124 • 60
- CohereForAI/c4ai-command-r-v01**
Text Generation • Updated about 20 hours ago • 27.1k • 750
- mistralai/Mistral-7B-v0.1**
Text Generation • Updated Dec 11, 2023 • 2.78M • 3.03k
- ai21labs/Jamba-v0.1**
Text Generation • Updated about 11 hours ago • 293 • 450
- databricks/dbrx-base**
Text Generation • Updated about 16 hours ago • 1.68k • 294
- Nexusflow/Starling-LM-7B-beta**
Text Generation • Updated 2 days ago • 5.7k • 174
- alpindale/Mistral-7B-v0.2-hf**
Text Generation • Updated 4 days ago • 8.35k • 121
- mistralai/Mixtral-8x7B-Instruct-v0.1**
Text Generation • Updated 29 days ago • 965k • 3.5k
- NousResearch/Hermes-2-Pro-Mistral-7B**
Text Generation • Updated about 14 hours ago • 38.7k • 347
- hpcai-tech/grok-1**
Text Generation • Updated 1 day ago • 1.15k • 56
- meta-llama/Llama-2-7b**
Text Generation • Updated 10 days ago • 3.74k
- mlabonne/Beyonder-4x7B-v3**
Text Generation • Updated about 13 hours ago • 739 • 41

Большие языковые модели

предобученные и модифицированные модели для русского языка

<https://huggingface.co/ai-forever>

<https://huggingface.co/IlyaGusev>



Models 79 🔍 ⬆️ Sort: Recently Updated

Model Name	Description	Updated	Downloads	Likes
ai-forever/kandinsky3-diffusers		Updated Feb 20	13	2
ai-forever/ruT5-base	Text2Text Generation	Updated Dec 11, 2023	5.92k	13
ai-forever/FRED-T5-large	Text2Text Generation	Updated Dec 5, 2023	2.46k	17
ai-forever/mGPT	Text Generation	Updated Dec 5, 2023	47k	218
ai-forever/rugpt2large		Updated Dec 5, 2023	411	7
ai-forever/ruT5-large	Text2Text Generation	Updated Dec 28, 2023	659	32
ai-forever/FRED-T5-1.7B	Text2Text Generation	Updated Dec 5, 2023	4.04k	63
ai-forever/mGPT-13B	Text Generation	Updated Dec 5, 2023	1.51k	46
ai-forever/ruGPT-3.5-13B	Text Generation	Updated Dec 5, 2023	3.82k	215
ai-forever/rugpt3small_based_on_gpt2	Text Generation	Updated Dec 5, 2023	11.8k	25

Большие языковые модели

квантование LLM

может уменьшать размер базовой модели на порядок
путём снижения точности (количества бит) вычислений (float32 в int8)
при не критичном понижении качества результата

снижаются требования к производительности аппаратуры



Большие языковые модели

Какие задачи можно решать

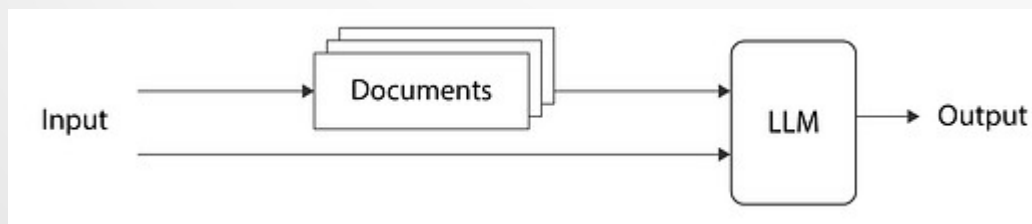
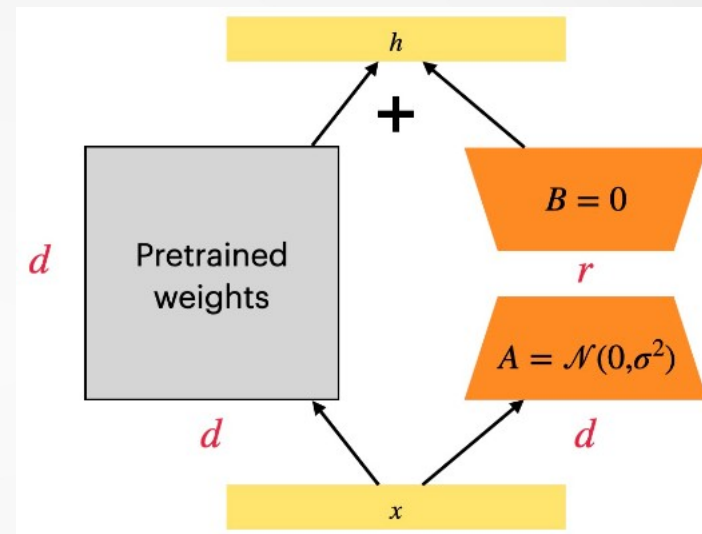
- машинный перевод
- автоматические ассистенты
- автоматическое реферирование больших текстов

....

Большие языковые модели

Тюнинг базовой модели под свою задачу

- прямое дообучение
- частичное дообучение PEFT/LoRA
- RAG (Retrieval-Augmented Generation)



Литература

Литература

Борисов Е.С. Методы машинного обучения. 2024
https://github.com/mechanoid5/ml_lectorium_2024_I

Борисов Е.С. Методы обработки текстов на естественном языке. 2024
https://github.com/mechanoid5/ml_nlp_2024_I

Майоров В.Д. Основы обработки текстов. 10. Языковые модели. ИСП РАН, 2021
https://www.youtube.com/watch?v=_8MGdpt4I9M

Тихомиров М.М. Основы обработки текстов. 14. Большие языковые модели. ИСП РАН, 2023
https://www.youtube.com/watch?v=EC6_rMs1vsY

Нейчев Радослав Self-Attention. Transformer overview. Лекторий ФПМИ, 2020
<https://www.youtube.com/watch?v=UETKUIIYE6g>

Jay Alammar Transformer в картинках. (Перевод - Е.Смирнова, С. Шкарин)
<https://habr.com/ru/articles/486358/>

Jay Alammar GPT-2 в картинках. (Перевод - Е.Смирнова, С. Шкарин)
<https://habr.com/ru/articles/490842/>