



Improving language models by retrieving from trillions of tokens

Подготовил: Стамбеков Алмас, ПМИ-213



Предпосылки. Проблемы LLM.

- Transformer (2017)
- BERT (2018)
- GPT - 2 (2019)
- GPT - 3 (2020)
- Огромное кол-во параметров
- Масштабирование
(GPT - 2 -> 3 - 185 миллиардов)
- Фактическая и языковая информации

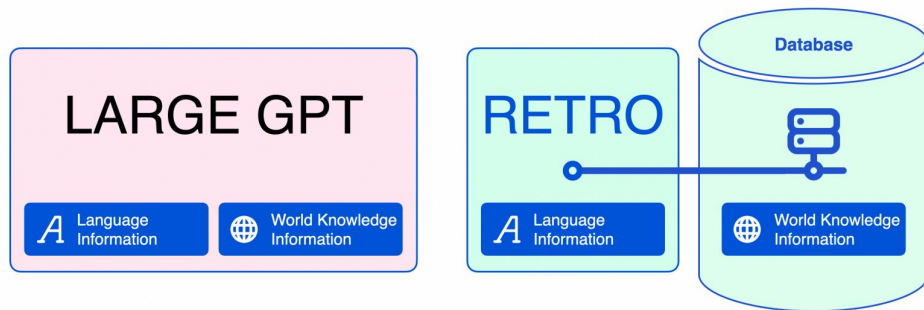


RETRO

- RETRO (**R**etrieval-**E**nhanced **T**Ransf**O**rmer) от DeepMind (2022)
- 7.5 миллиардов параметров (4% от GPT - 3)
- Расширение объема текстовых данных
- Механизм извлечения информации (механизм поиска)

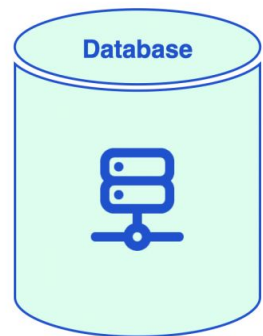
Нейронная база данных (neural database)

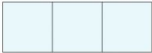

- Ответ на вопрос с использованием содержимого базы данных и структуры языка.
- База данных — это хранилище ключей-значений



База данных

The Dune film was released ...

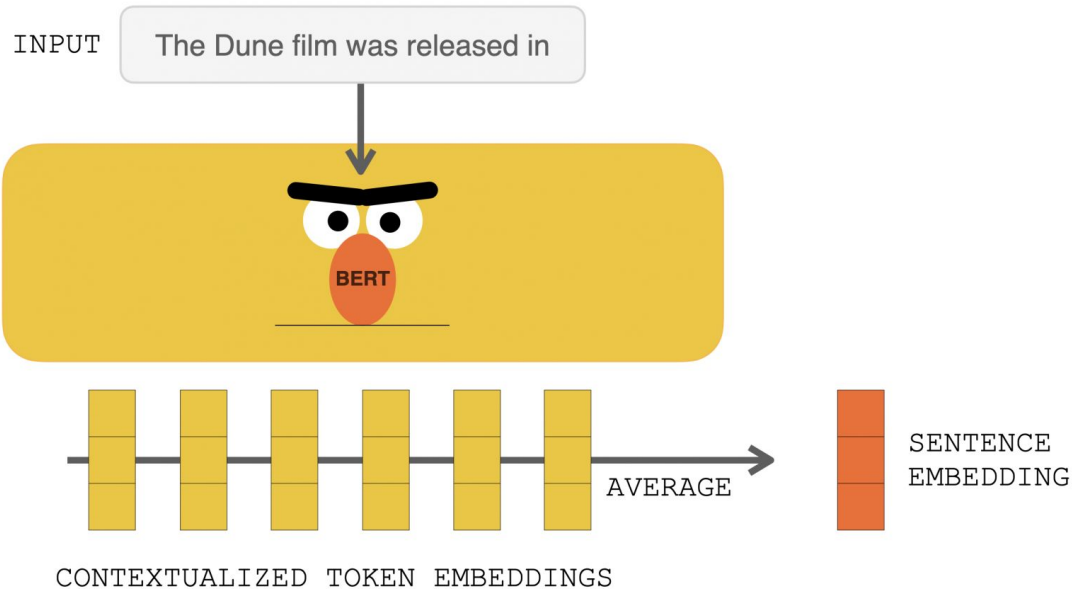


Key (BERT sentence embedding)	Value (text, neighbor and completion chunks. Each up to 64 tokens in length)	
	Dune is a 2021 American epic science fiction film directed by Denis Villeneuve	NEIGHBOR
	It is the first of a planned two-part adaptation of the 1965 novel by Frank Herbert	COMPLETION
	Dune is a 1965 science fiction novel by American author Frank Herbert	NEIGHBOR
	originally published as two separate serials in Analog magazine	COMPLETION
...	...	

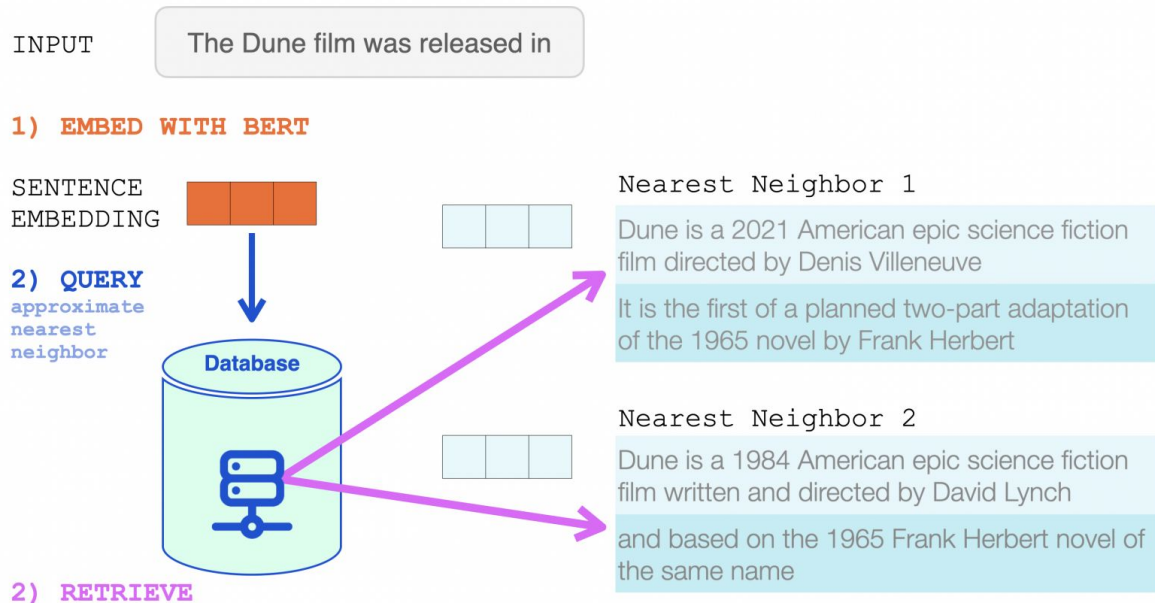
- **Ключ** - эмбединг (BERT)
- **Значение** - предложение из двух частей:
 - **Neighbor** - вычисление ключа
 - **Completion** - продолжение

RETRO содержит 2 триллиона многоязычных токенов на основе набора данных MassiveText.

Механизм поиска в базе данных.



Механизм поиска в базе данных.



Эмбединг предложения
->

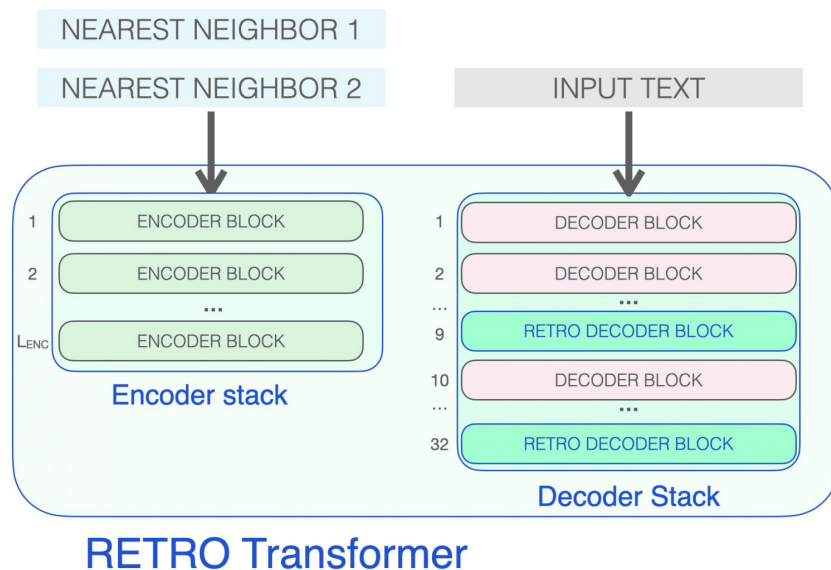
-> Приближенный поиск
ближайшего соседа

Итоговый вход в модель:

1. **Входная последовательность**
2. **Два её ближайших соседа из базы данных**

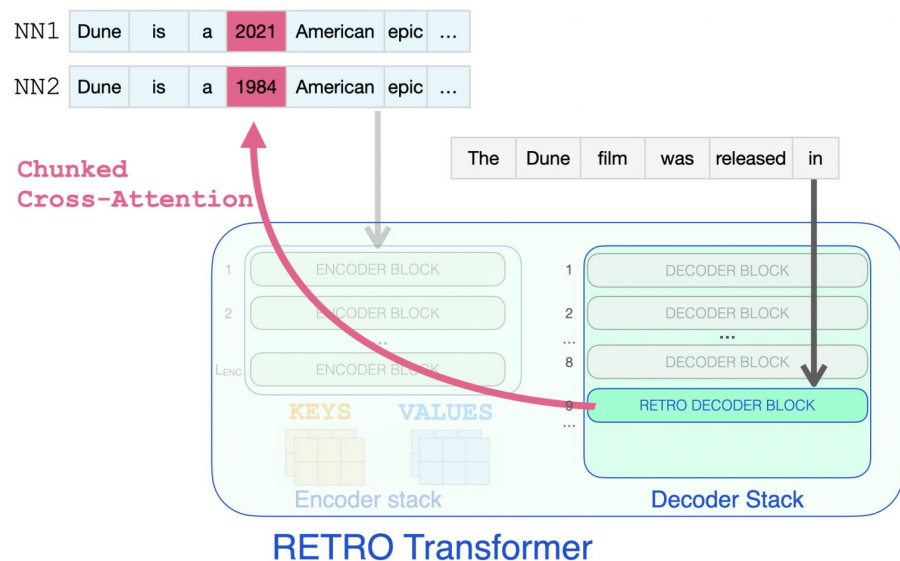
Архитектура RETRO

- **Энкодер** - стандартные блоки Трансформера
- **Декодер**:
 - Стандартный блок декодера
 - Декодер RETRO

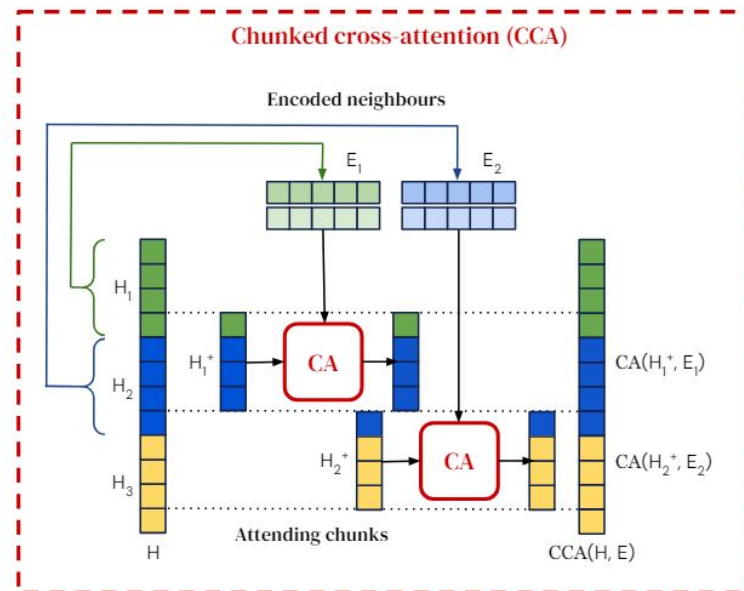
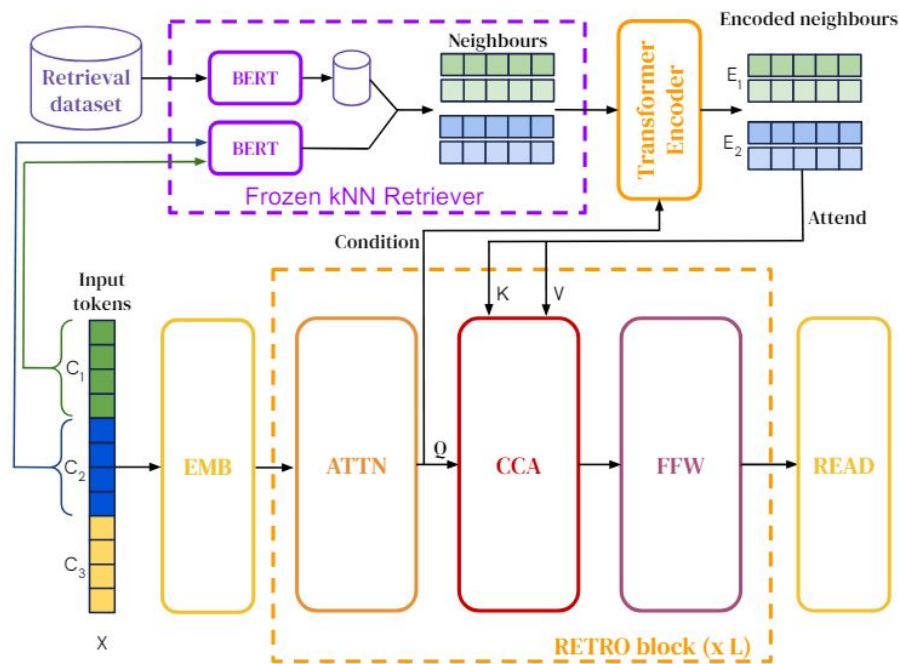


Декодер RETRO

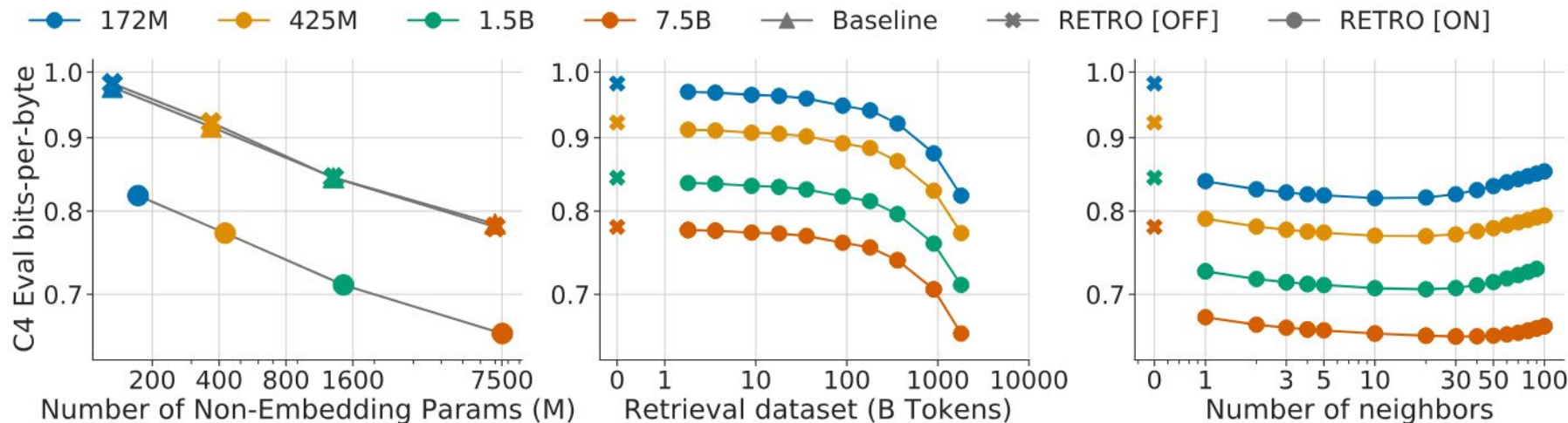
Блок декодера RETRO - извлечение информации из ближайших соседей с использованием **Chunked Cross-Attention**.



Архитектура RETRO



Масштабирование RETRO



Метрика C4 Eval bits-per-byte - оценка эффективности моделей языкового прогнозирования, измеряемой в битах на байт



MassiveText

- Частный датасет от DeepMind
- Содержит 2.35 миллиарда документов или примерно 10.5 ТБ текста

Source	Language	Token count (M)	Documents	Sampling weight
Web	En	483,002	604,938,816	0.314
	Ru	103,954	93,004,882	0.033
	Es	95,762	126,893,286	0.033
	Zh	95,152	121,813,451	0.033
	Fr	59,450	76,612,205	0.033
	De	57,546	77,242,640	0.033
	Pt	44,561	62,524,362	0.033
	It	35,255	42,565,093	0.033
	Sw	2,246	1,971,234	0.0044
	Ur	631	455,429	0.0011
Books	En	3,423,740	20,472,632	0.25
News	En	236,918	397,852,713	0.1
Wikipedia	En	3,977	6,267,214	0.0285
	De	2,155	3,307,818	0.003
	Fr	1,783	2,310,040	0.003
	Ru	1,411	2,767,039	0.003
	Es	1,270	2,885,013	0.003
	It	1,071	2,014,291	0.003
	Zh	927	1,654,772	0.003
	Pt	614	1,423,335	0.003
	Ur	61	344,811	0.0001
	Sw	15	58,090	0.0004
Github	-	374,952	142,881,832	0.05
Total	-	5,026,463	1,792,260,998	1



Материалы

1. [Improving language models by retrieving from trillions of tokens](#)
2. [Retrieval Transformer в картинках](#) (habr)
3. [Retrieval Transformer Enhanced Reinforcement Learning](#) (Medium)
4. [GitHub](#)