Нейросетевые языковые модели Transformer

Евгений Борисов

Нейросетевые языковые модели

Языковая модель
- предсказываем следующее слово на основе предыдущих
- оценка (вероятность) совместимости цепочки слов

Оценка цепочки слов (биграммная модель):

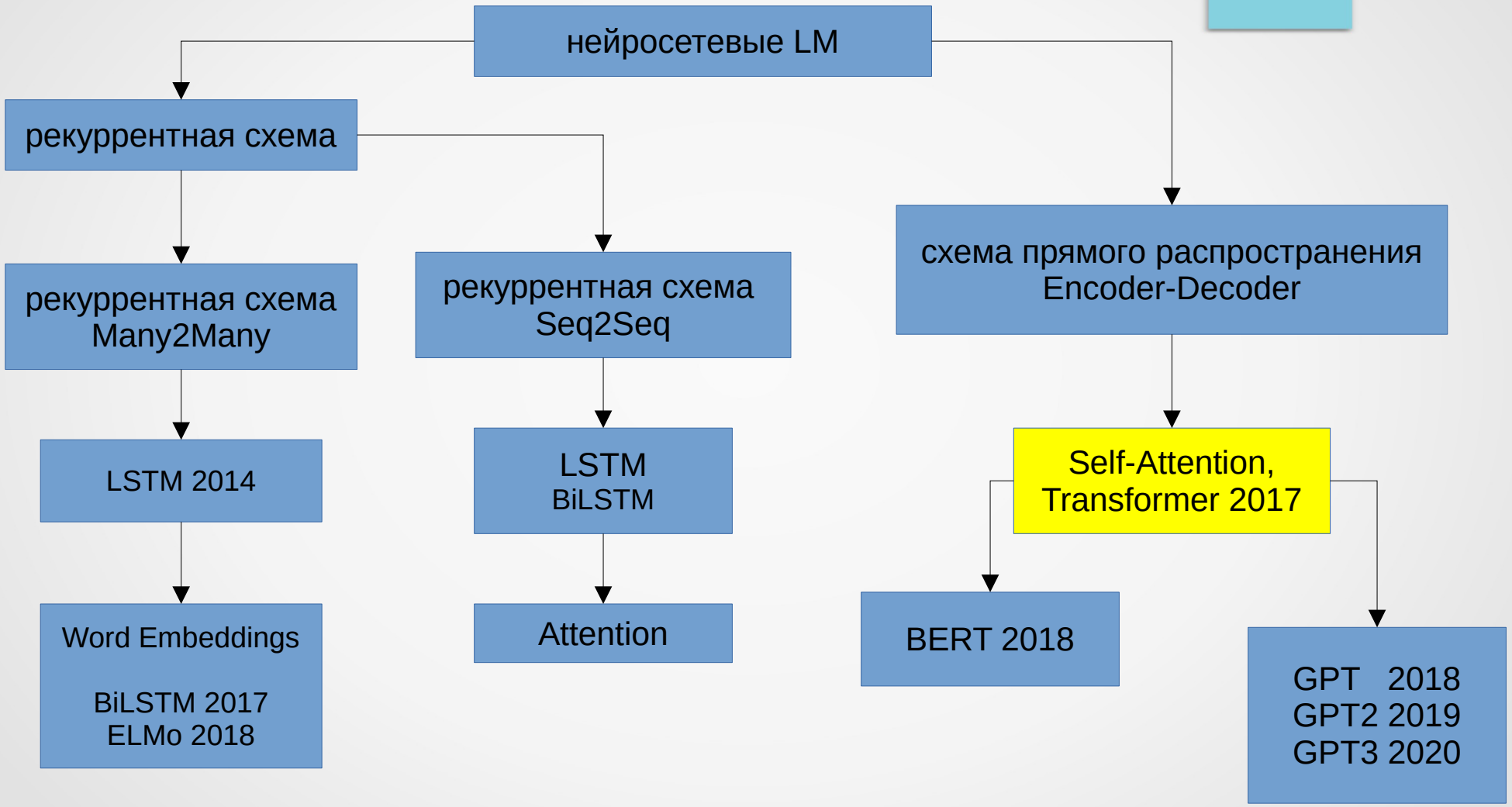
$$p(w_1 \dots w_n) = \prod_{k=1}^{n} p(w_k | w_{k-1})$$

$$p(w_n | w_{n-1}) = \frac{C(w_{n-1} w_n)}{C(w_{n-1})}$$

P - вероятность совместного использования слов

C(w) – количество слов w в тексте

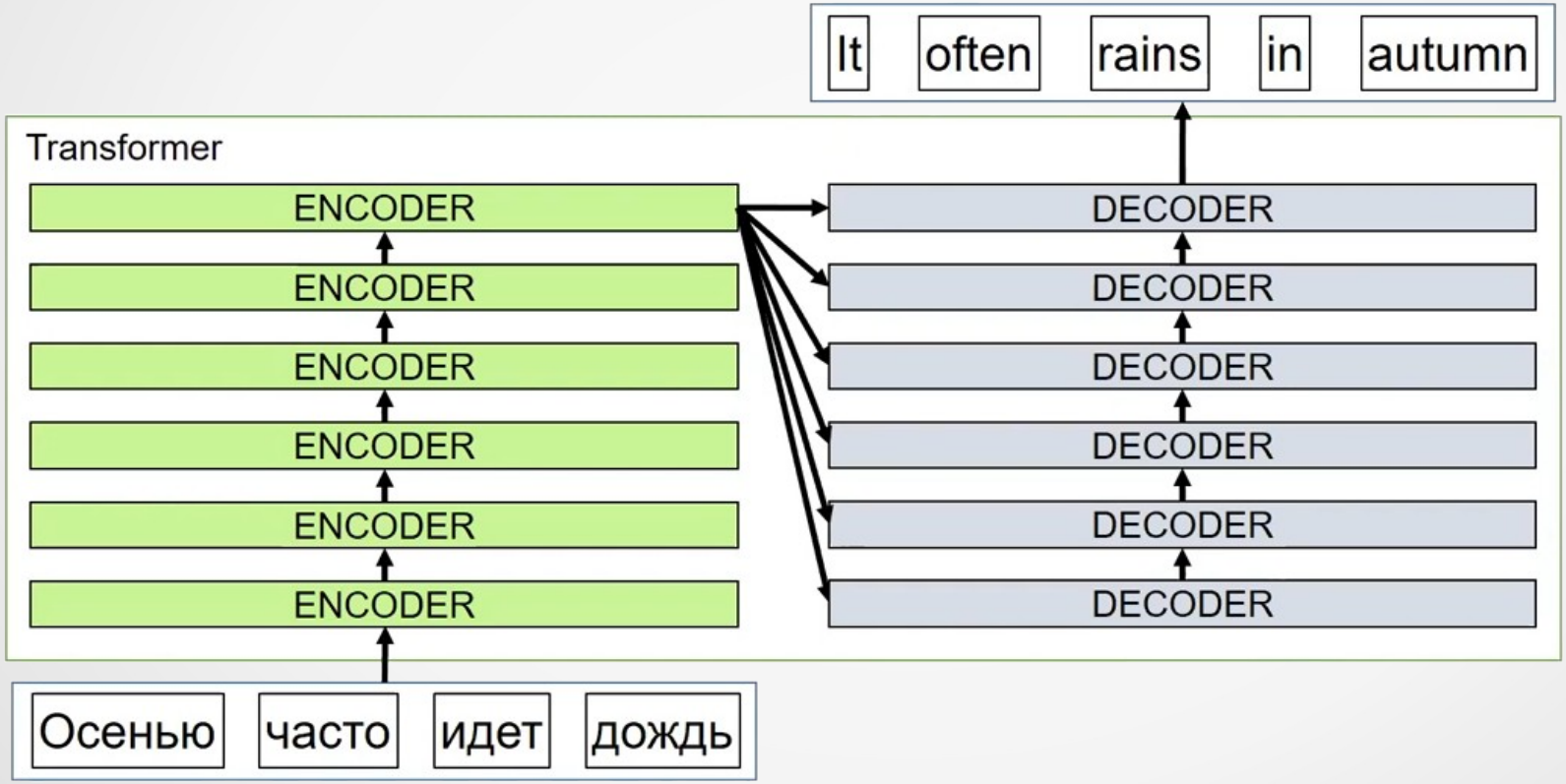
Нейросетевые языковые модели



Нейросетевые языковые модели

Схема Encoder-Decoder прямого распространения

Модель Transformer и механизм Self-Attention
Attention Is All You Need (2017) https://arxiv.org/abs/1706.03762



токенизация BPE (Byte Pair Encoding)
Sennrich R., et al. Neural machine translation of rare words with subword units– 2015.

https://www.youtube.com/watch?v=_8MGdpt4I9M

Нейросетевые языковые модели

Transformer: Энкодер

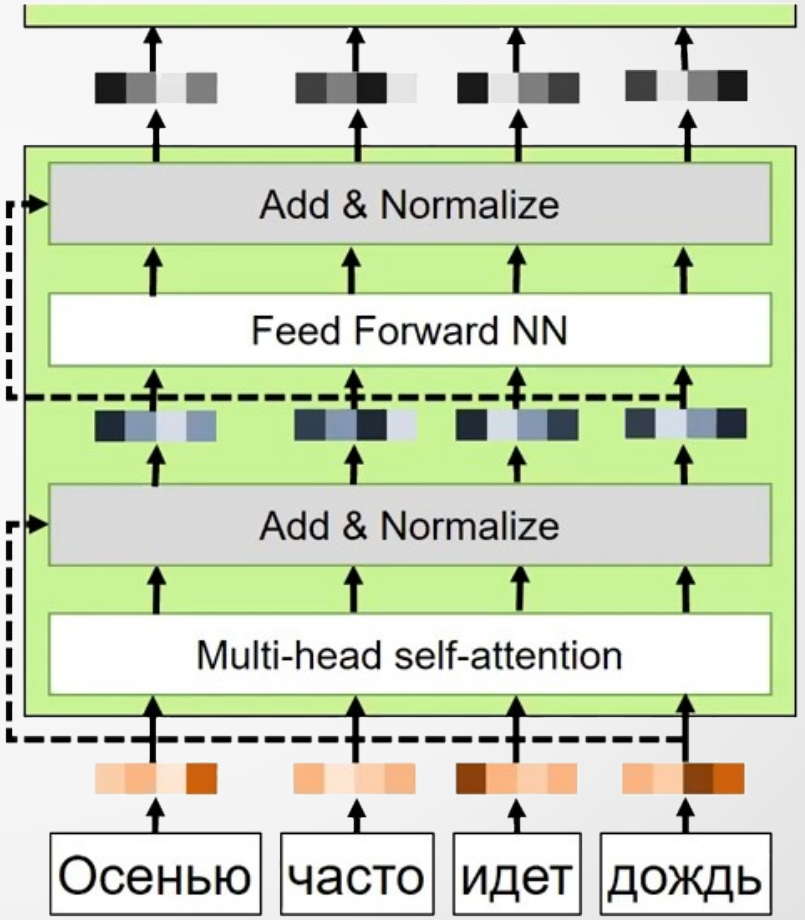
- блок внимания MHSA

- skip connection, normalization

- сеть прямого распространения

все слова подаём в модель одновременно,
они обрабатываются совместно,

схема вычислений хорошо распараллеливается

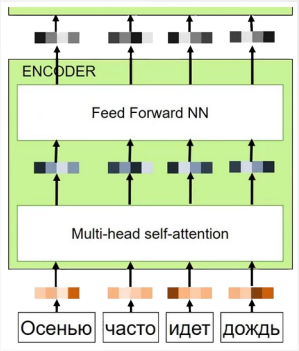


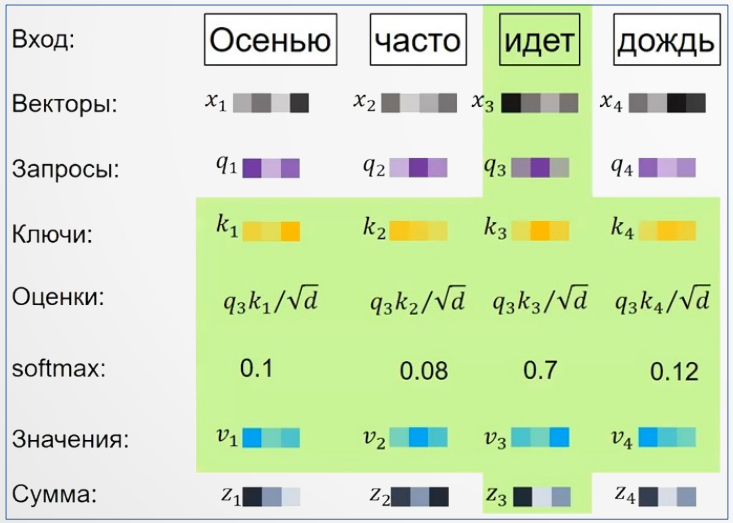
https://www.youtube.com/watch?v=_8MGdpt4I9M

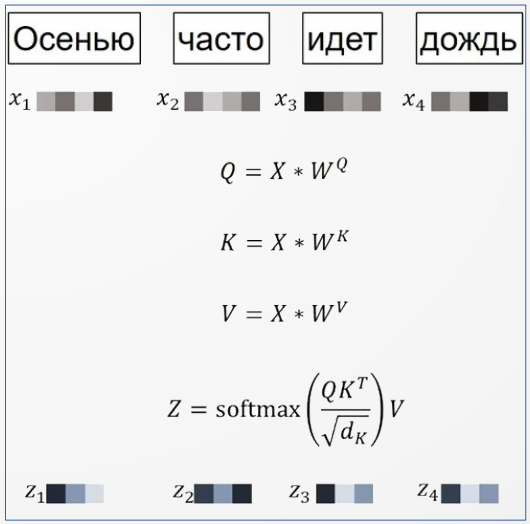
Нейросетевые языковые модели

`Transformer: Encoder : Self-Attention`

query - откуда смотрим (из какого слова)
key - куда смотрим (на какое слово)
value - смысл (условно) слова







все слова подаём в модель одновременно,
они обрабатываются совместно,
схема вычислений хорошо распараллеливается

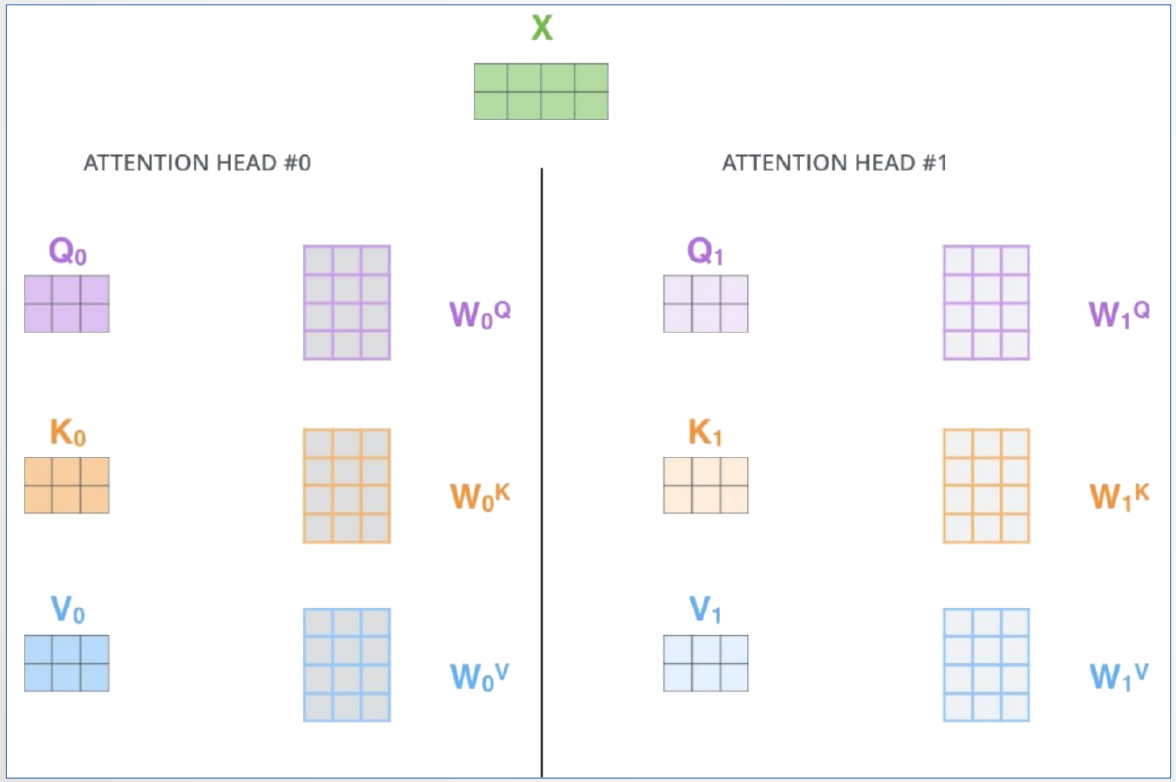
https://www.youtube.com/watch?v=_8MGdpt4I9M

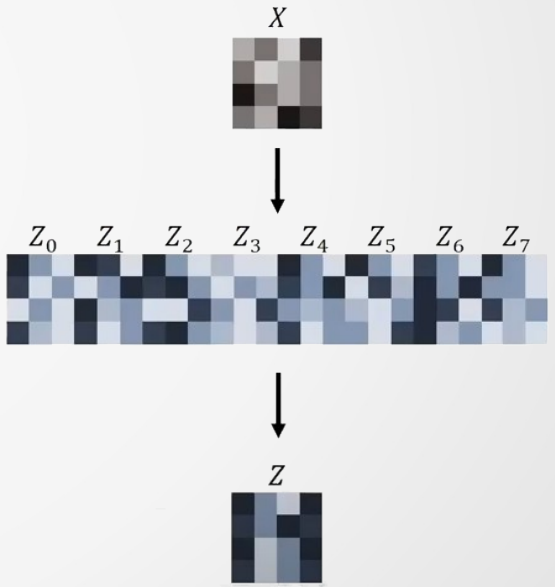
Нейросетевые языковые модели

Transformer: Encoder : Multi-Head-Self-Attention

Используем параллельно несколько блоков Self-Attention с разными весами

Результаты агрегируются в размер входа X для стекирования блоков энкодера





Нейросетевые языковые модели

Transformer: Encoder : Multi-Head-Self-Attention

1) This is our input sentence*

2) We embed each word*

3) Split into 8 heads. We multiply X with weight matrices

4) Calculate attention using the resulting Q/K/V matrices

5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer

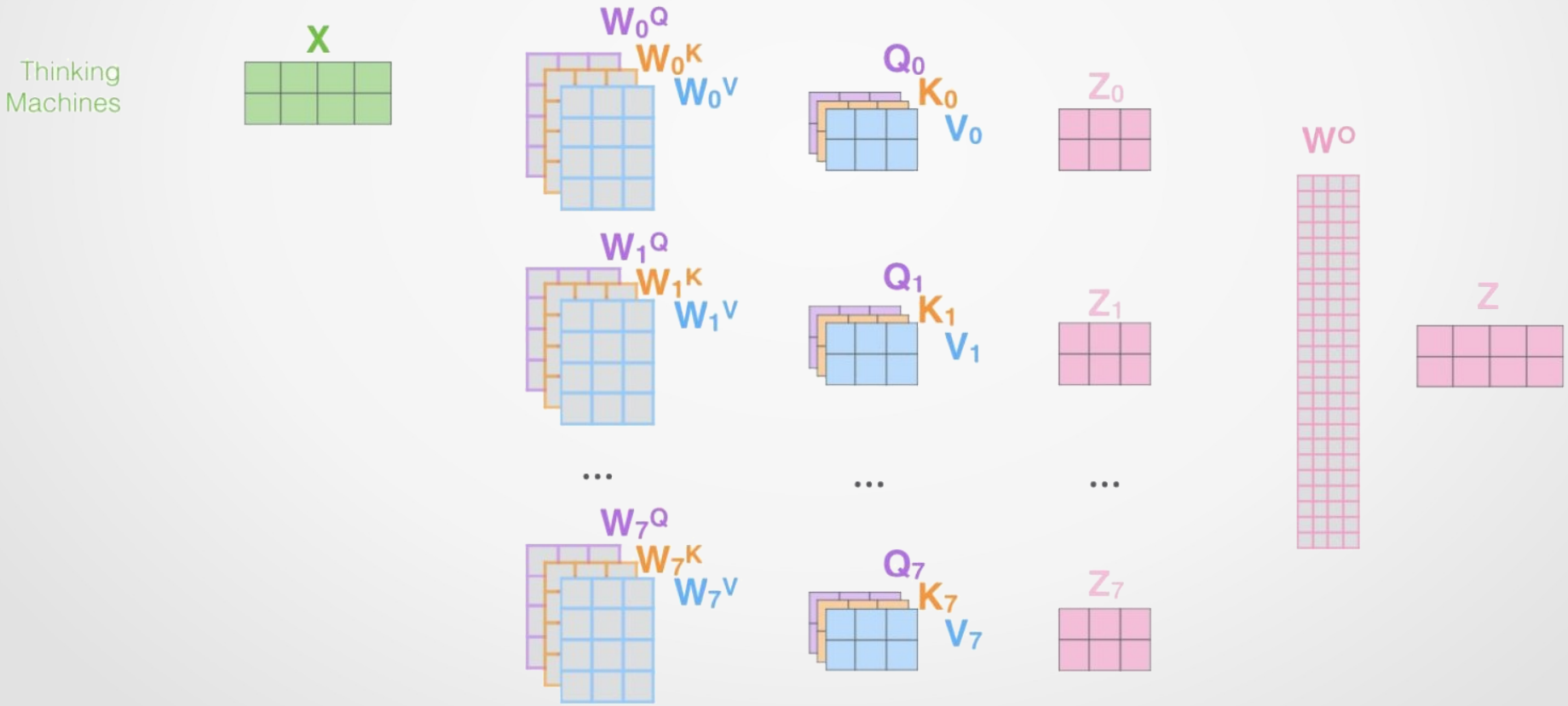


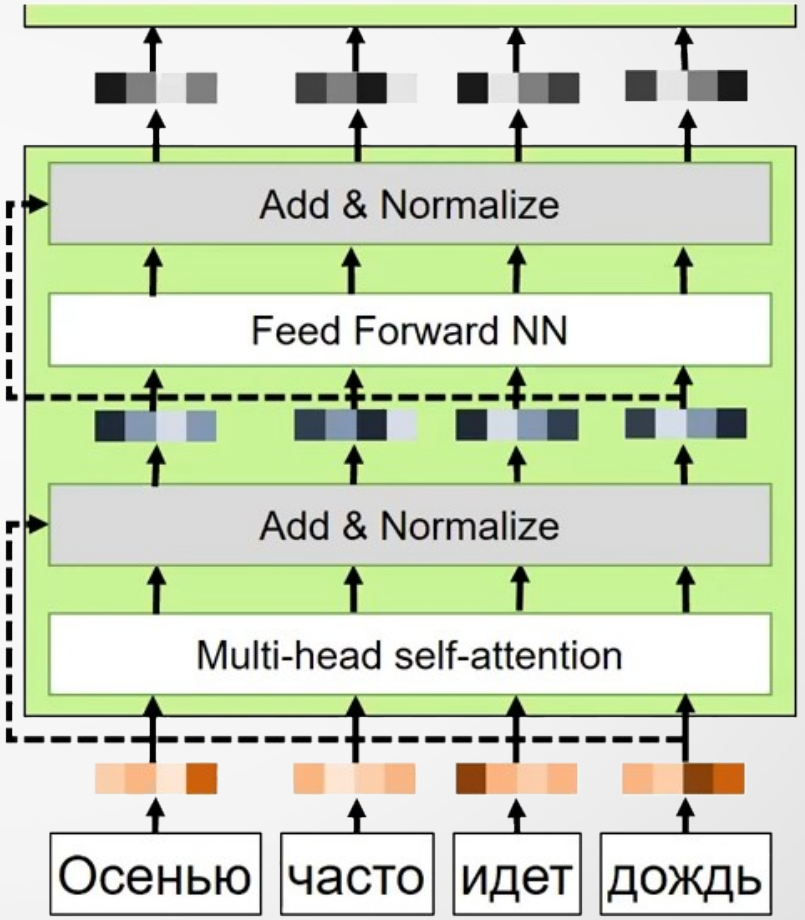
Image source: https://jalammar.github.io/illustrated-transformer/

Нейросетевые языковые модели

Transformer: Encoder

Проблема: не учитывается порядок слов

Решение: positional encoding



https://www.youtube.com/watch?v=_8MGdpt4I9M

Нейросетевые языковые модели

Transformer: Encoder : Positional encoding

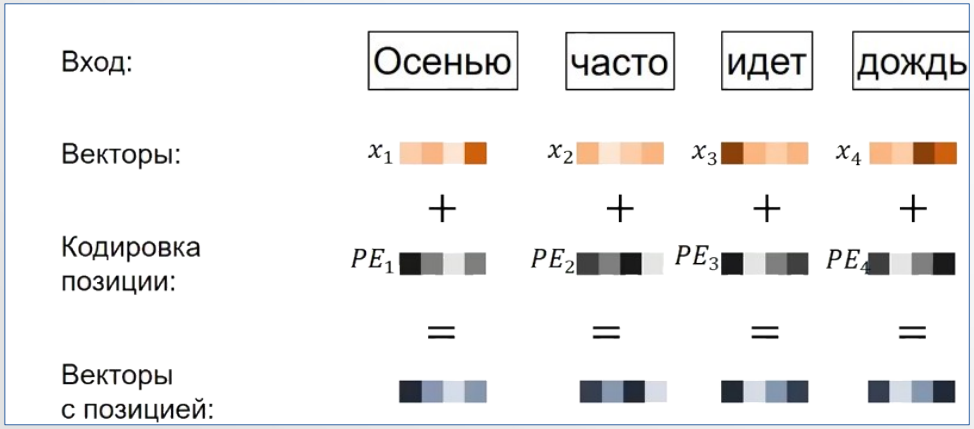
Необходимо обозначить позицию слова, выполняя условия
- уникальность для каждого слова
- не зависит от длинны предложения
- детерминирован (не стохастический)

$$\vec{p_t}^{(i)} = f(t)^{(i)} = \begin{cases} \sin(\omega_k t), & \text{if } i = 2k \\ \cos(\omega_k t), & \text{if } i = 2k+1 \end{cases}$$

$$\omega_k = \frac{1}{10000^{2k/d}}$$

$$\vec{p_t} = \begin{bmatrix} \sin(\omega_1.t) \\ \cos(\omega_1.t) \\ \sin(\omega_2.t) \\ \cos(\omega_2.t) \\ \vdots \\ \sin(\omega_{d/2}.t) \\ \cos(\omega_{d/2}.t) \end{bmatrix}_{d \times 1}$$

t – номер слова в строке
d – размерность входа модели
k – номер элемента в векторе PE



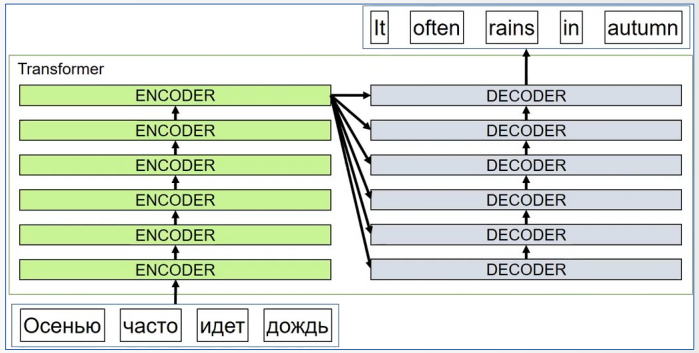
PE – не обучается вместе с моделью, но вычисляется по формуле

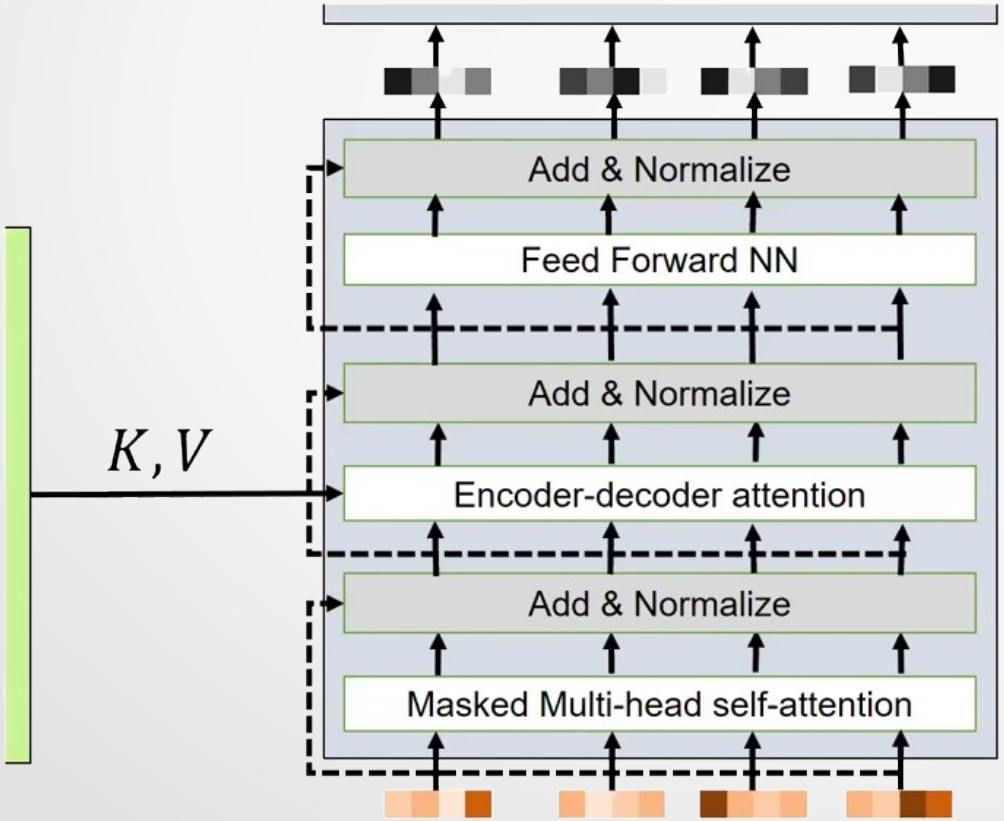
длина последовательности ограничена

Нейросетевые языковые модели

Transformer: Decoder

авторегрессионная модель - выход подаём на вход





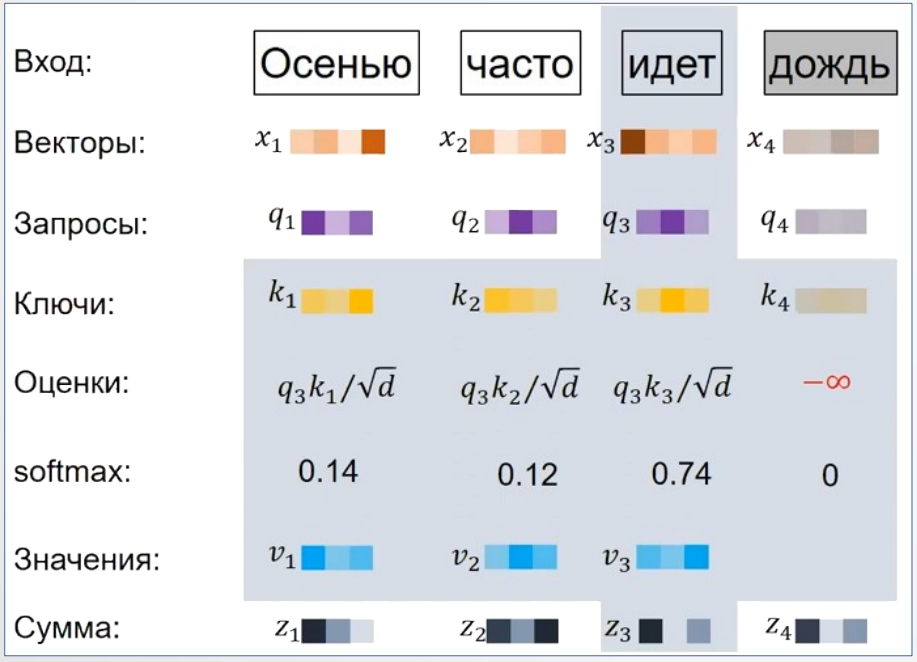
- encoder-decoder attention, используем [K,V] из последнего блока encoder

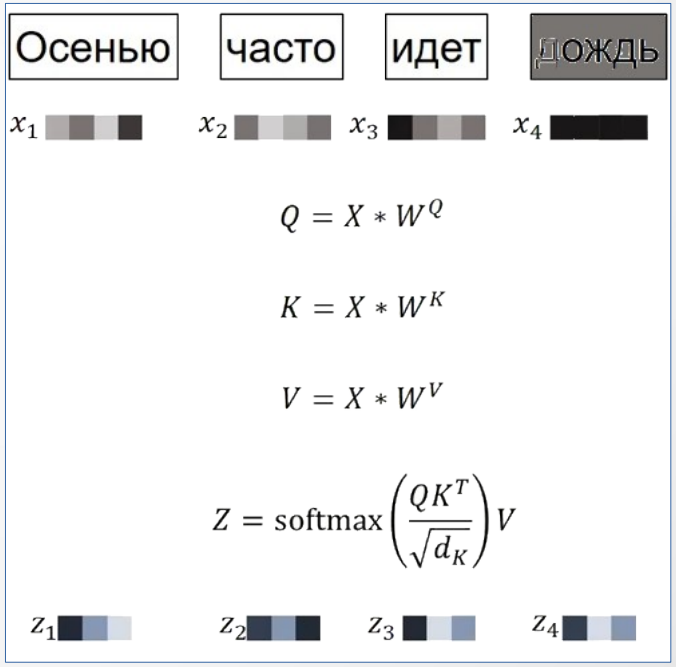
- masked MHSA
при расчёте self-attention используем только левый контекст

Нейросетевые языковые модели

Transformer: Decoder : Musked Multi-Head-Self-Attention

в процессе обучения модели,
при расчёте self-attention используем только левый контекст



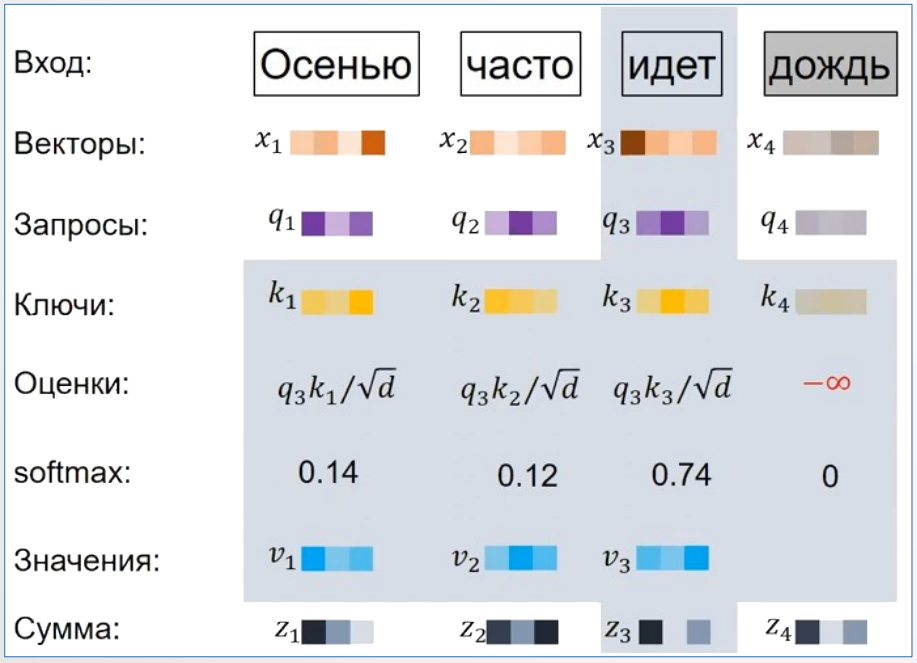


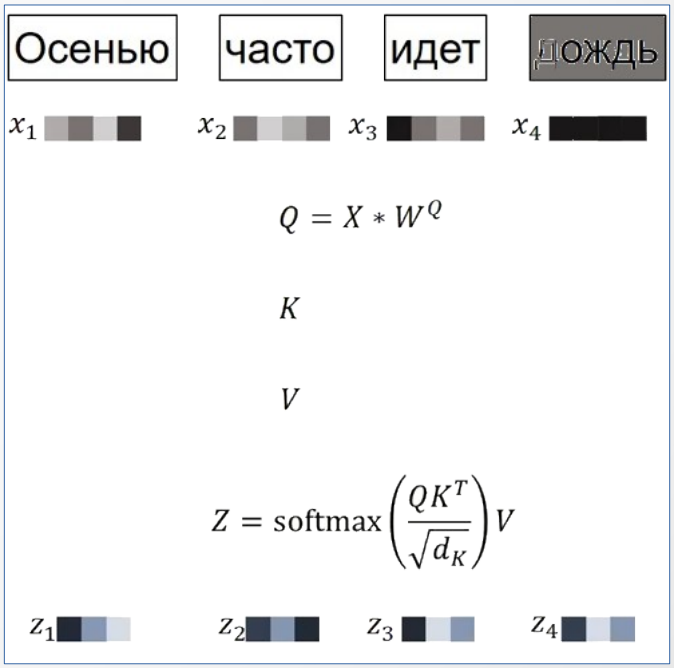
Нейросетевые языковые модели

Transformer: Decoder : Encoder-Decoder Attention

используем [K,V] берём из последнего блока encoder

похоже на Attention из рекуррентных SEQ2SEQ





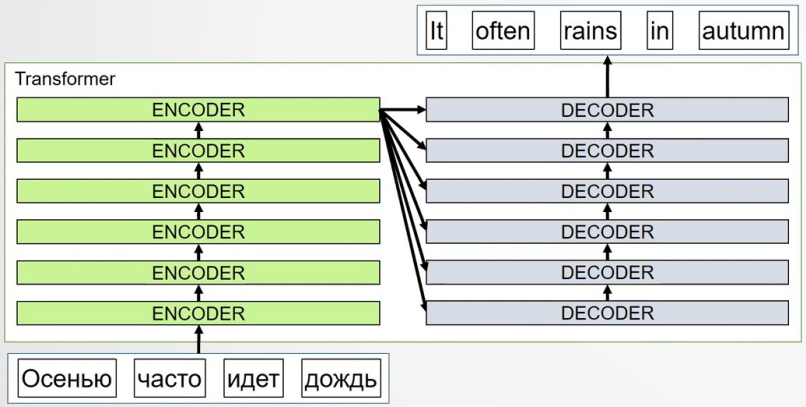
Нейросетевые языковые модели

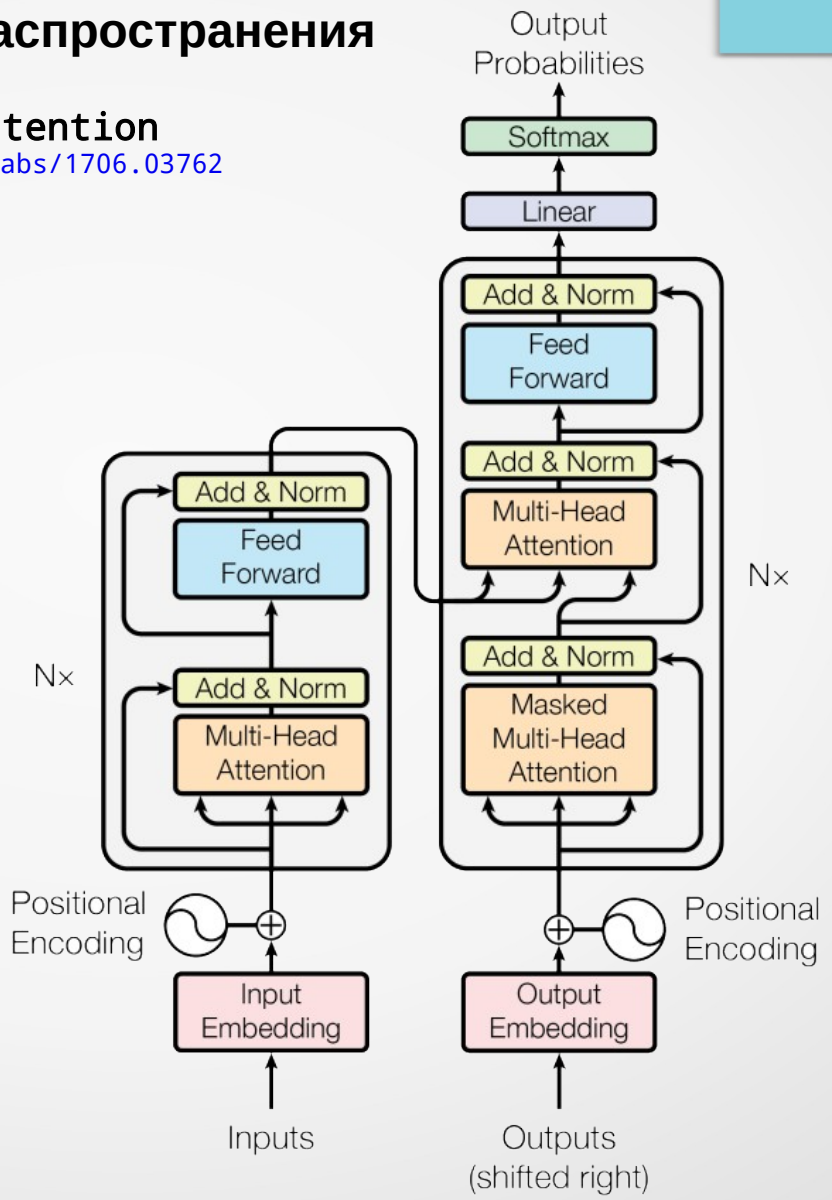
Схема Encoder-Decoder прямого распространения

Модель Transformer и механизм Self-Attention

Attention Is All You Need (2017) https://arxiv.org/abs/1706.03762

токенизация BPE (Byte Pair Encoding)
Sennrich R., et al. Neural machine translation of rare words with subword units– 2015.





https://www.youtube.com/watch?v=_8MGdpt4I9M

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=UETKUIlYE6g

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/