Основы практического использования нейронных сетей.

Лекция 9. Трансформеры.

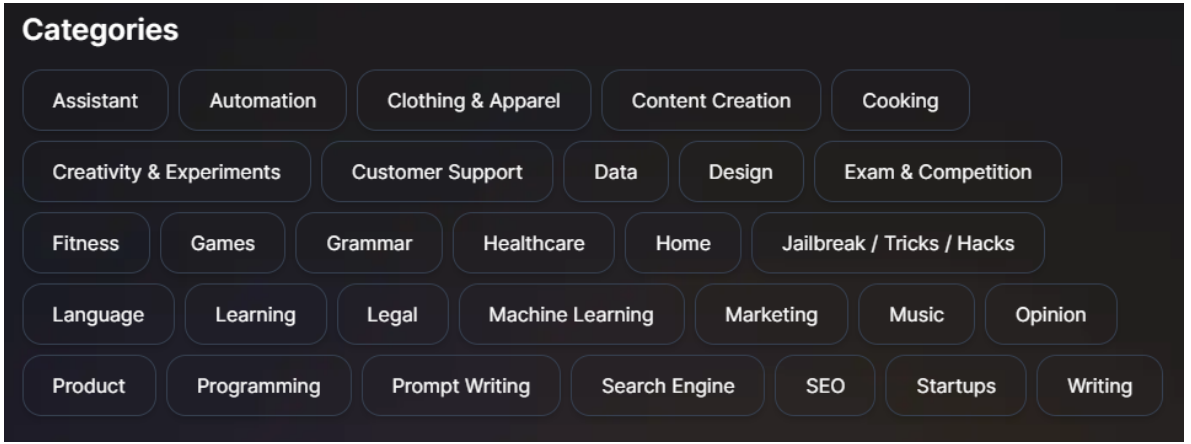
Дмитрий Буряк.
к.ф.-м.н.
dyb04@yandex.ru

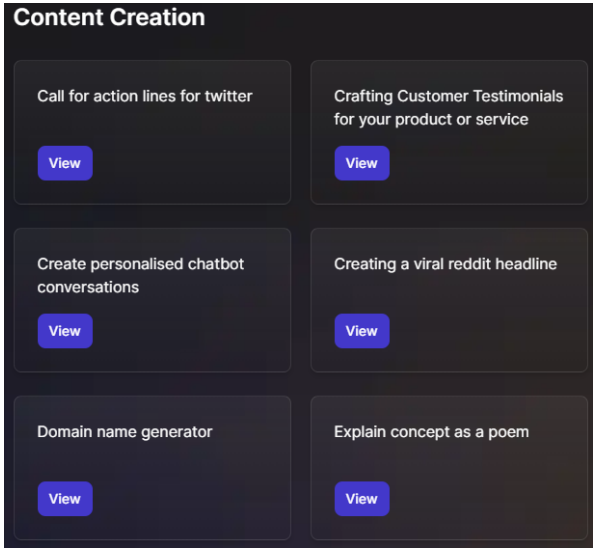
Задачи анализа естественного языка

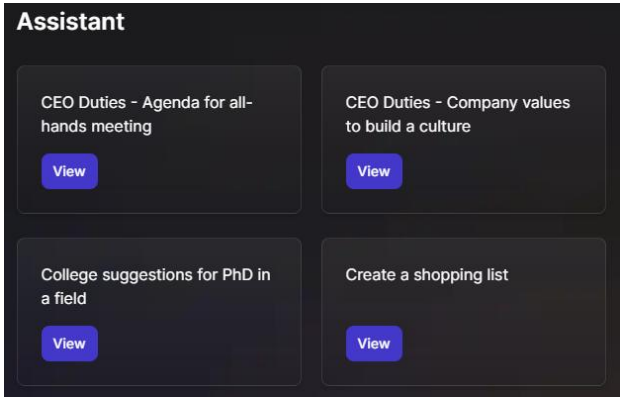
- Natural Language Processing (NLP) – применение методов распознавания образов в словам, предложениям, текстам.
- Виды задач NLP:
  - Классификация текста (text classification)
  - Фильтрация контента (content filtering)
  - Классификация отзывов (sentiment analysis)
  - Моделирование языка (language modeling)
  - Перевод (translation)

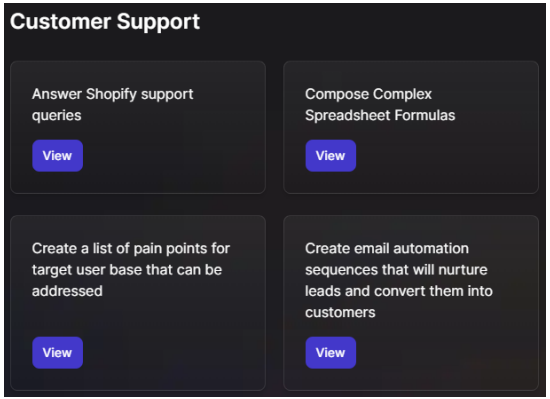
ChatGPT

- ❑ Chat Generative Pre-Trained Transformer
- ❑ Дообученная модель GPT-3
- ❑ 175 млрд параметров
- ❑ Выпущен в ноябре 2022





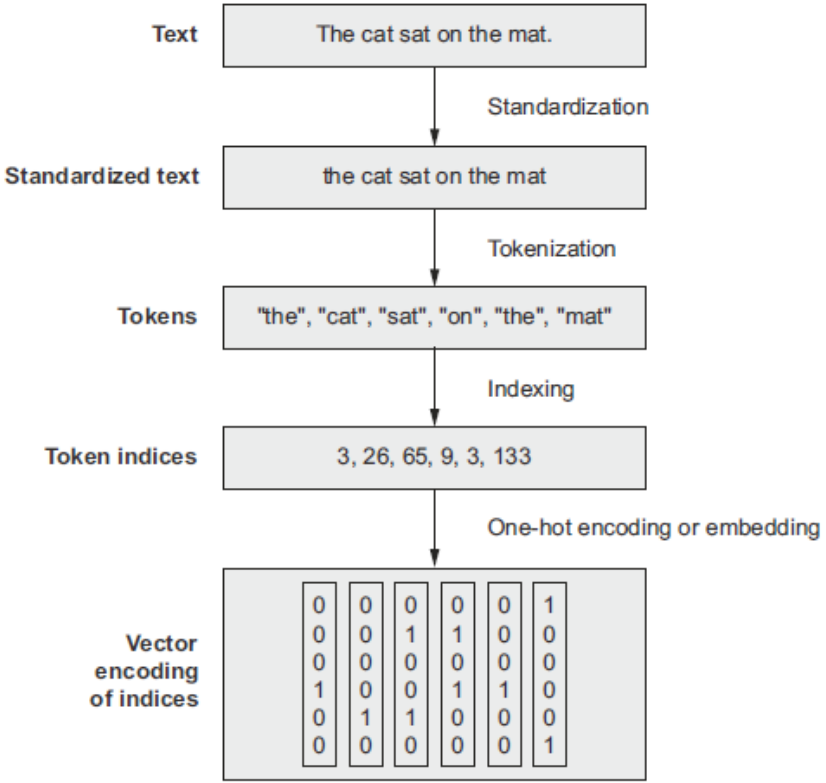




https://cookup.ai/chatgpt/usecases/

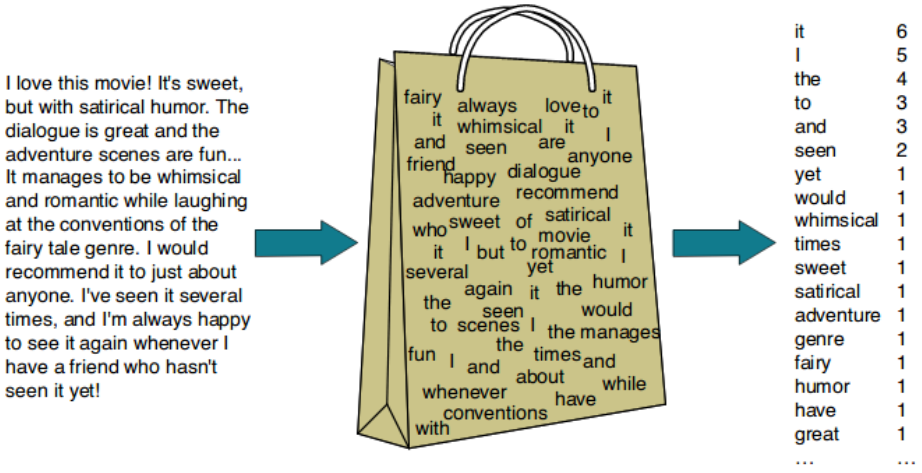
Предобработка текстовых данных

- ❑ Нормализация: приведение к прописным символам, удаление знаков пунктуации.
- ❑ Выделение токенов: символов, слов, групп слов (N-gram).
- ❑ Векторизация: преобразование токена в вектор (словарь → индекс слова в словаре → бинарный вектор/эмбединг).



Представление текста

- Текст – множество слов (N-gram) – bag-of-words.
  - сети прямого распространения, персептроны.
- Текст – последовательность слов
  - Рекуррентные сети (LSTM, GRU)
  - Трансформеры

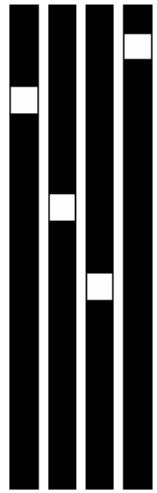


Представление типа bag-of-words
https://koushik1102.medium.com/nlp-bag-of-words-and-tf-idf-explained-fd1f49dce7c4

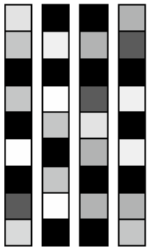
Векторизация. Дескрипторы слов.

- Бинарные вектора (on-hot):
  - высокая размерность
  - ортогональны
  - разрежены.

Не отражают свойств естественного языка

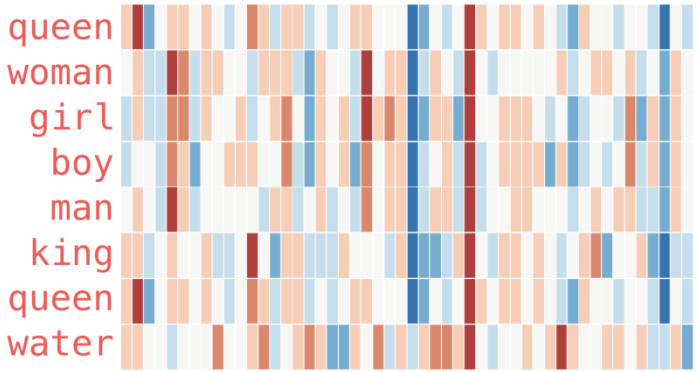


One-hot word vectors:
- Sparse
- High-dimensional
- Hardcoded



Word embeddings:
- Dense
- Lower-dimensional
- Learned from data

- Формирование дескрипторов слов (эмбединг)
  - вектора для близких по смыслу слов должны быть ближе, чем те, которые имеют разное значение
  - получены в результате обучения
  - word2vec
  - GLoVe



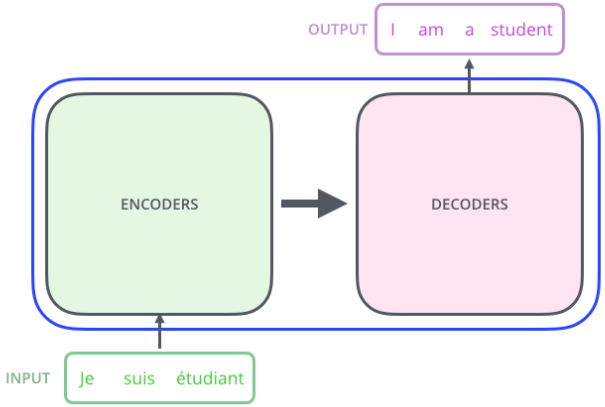
https://habr.com/ru/post/446530/

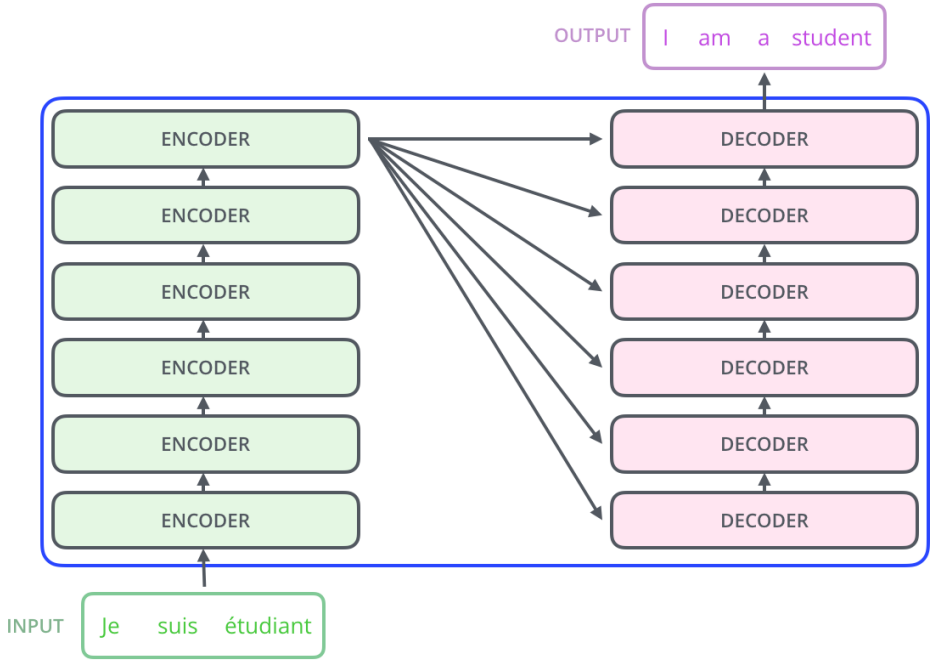
Трансформеры для обработки текстов

- Vaswani et al. Attention is all you need, 2017
- Сеть прямого распространения, частично связанная, несколько параллельных ветвей, без сверточных слоев.
- Изначально предназначена для NLP, сейчас адаптирована для других задач, например, анализ изображений.

Структура трансформера

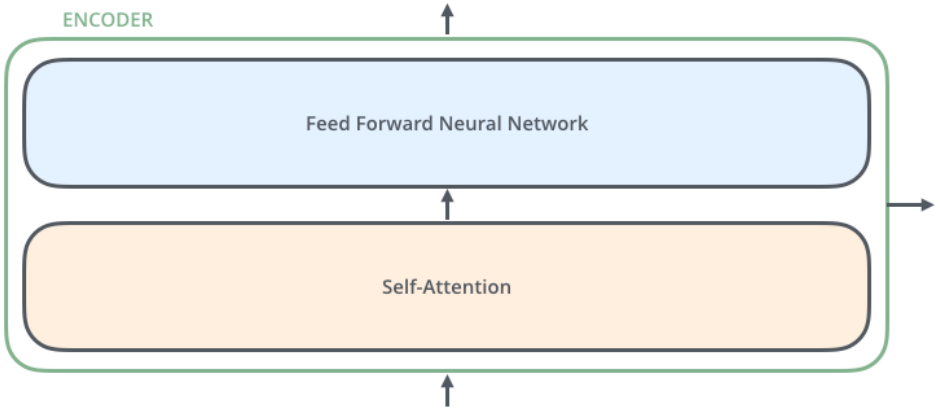
- Энкодер - Декодер





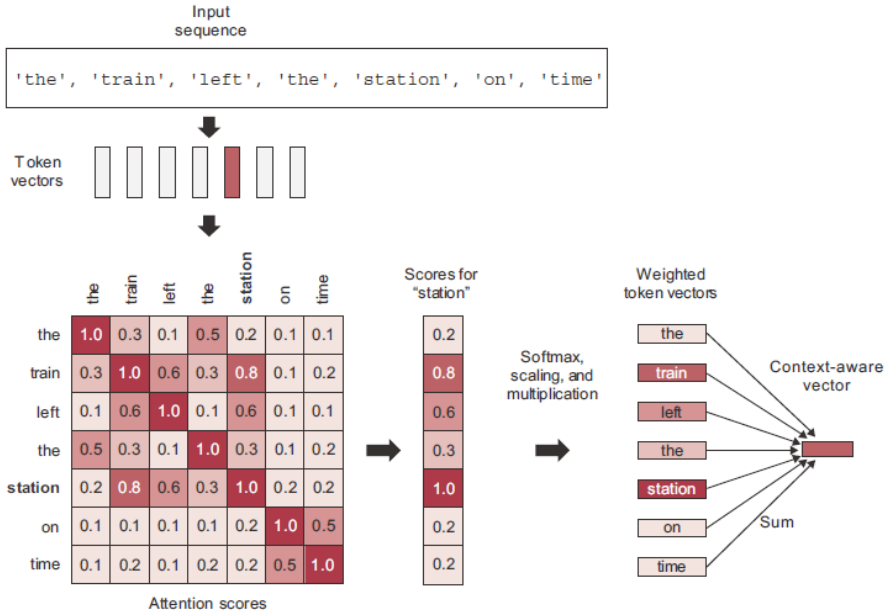
Энкодер

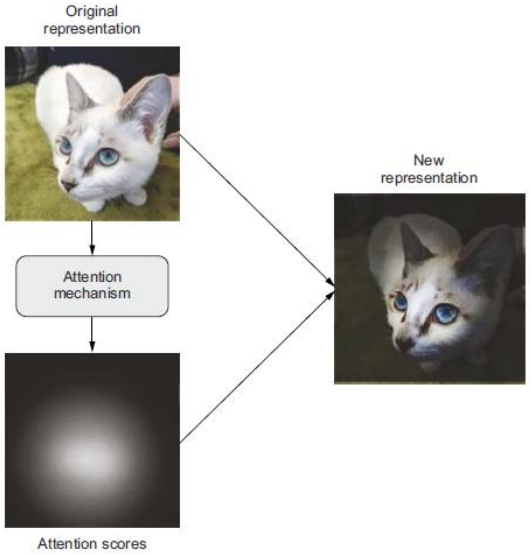
- ❑ Слой внимания: оценить взаимосвязь слов во входящей последовательности.
- ❑ Сеть прямого распространения для отображения результата слоя внимания



Механизм «Внимания»

- Слова могут иметь разное значение в зависимости от контекста
- Выделение признаков, зависящих от контекста.
- Цель – сформировать новое признаковое представление в зависимости от других слов в последовательности.



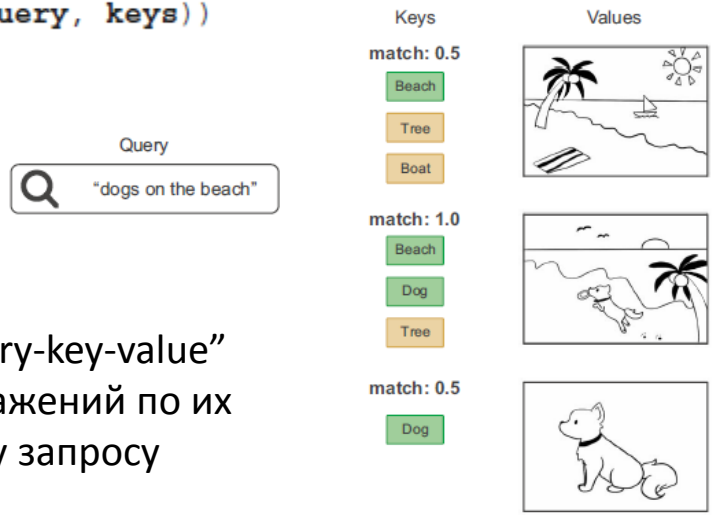


Идея реализации механизма «внимания» в глубоком обучение

Модель “Query-Key-Value”

- ❑ Запрос (query) сравнивается с известными элементами данных (key) → коэффициент схожести.
- ❑ Для каждого известного элемента сопоставлен вектор значений (value).
- ❑ Результат сумма элементов значений, завешенных коэффициентами схожести.

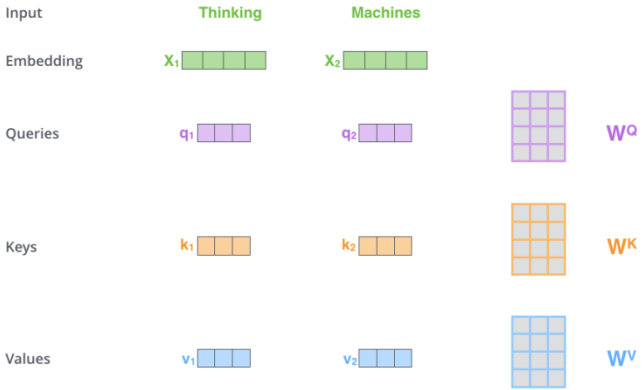
```
outputs = sum(values * pairwise_scores(query, keys))
```



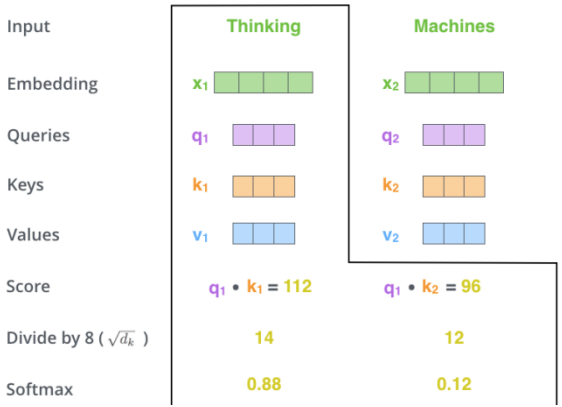
Применение модели “query-key-value” для ранжирования изображений по их релевантности текстовому запросу

Пример “Query-Key-Value” (1)

❑ Элементы запрос, ключ и значение – результат умножения входного эмбединга на соответствующую матрицу

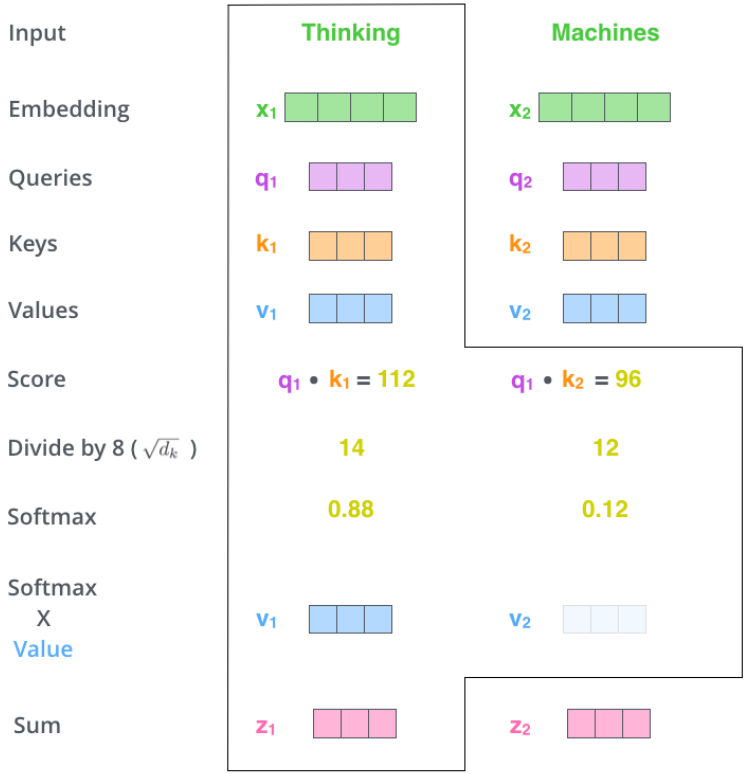


❑ Вычисление коэффициента схожести и взвешивающих коэффициентов
(деление на 8 – для данного примера)



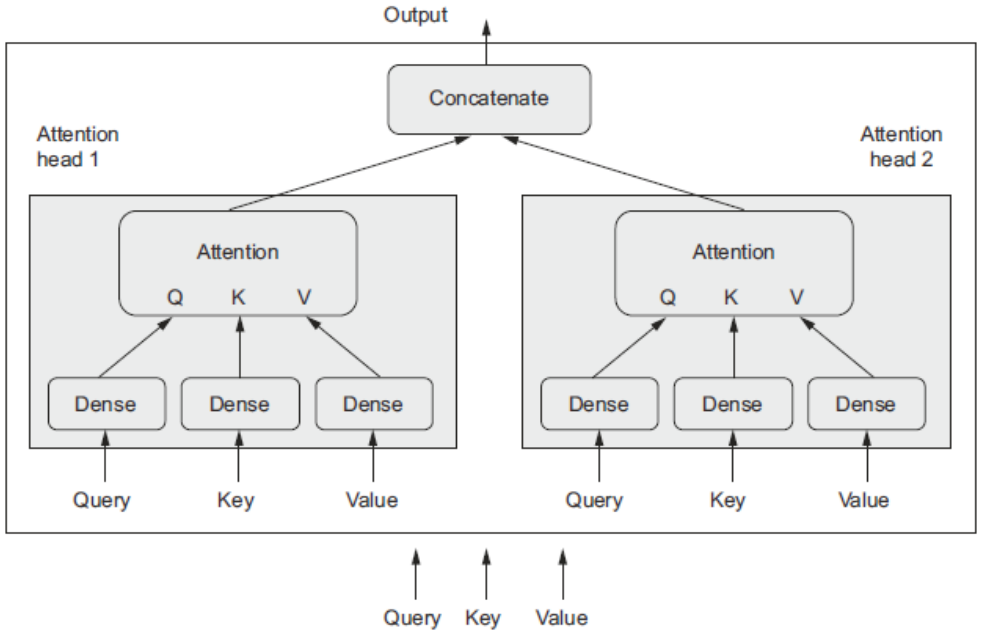
Пример “Query-Key-Value” (2)

- Получение представления с учетом внимания для входного слова

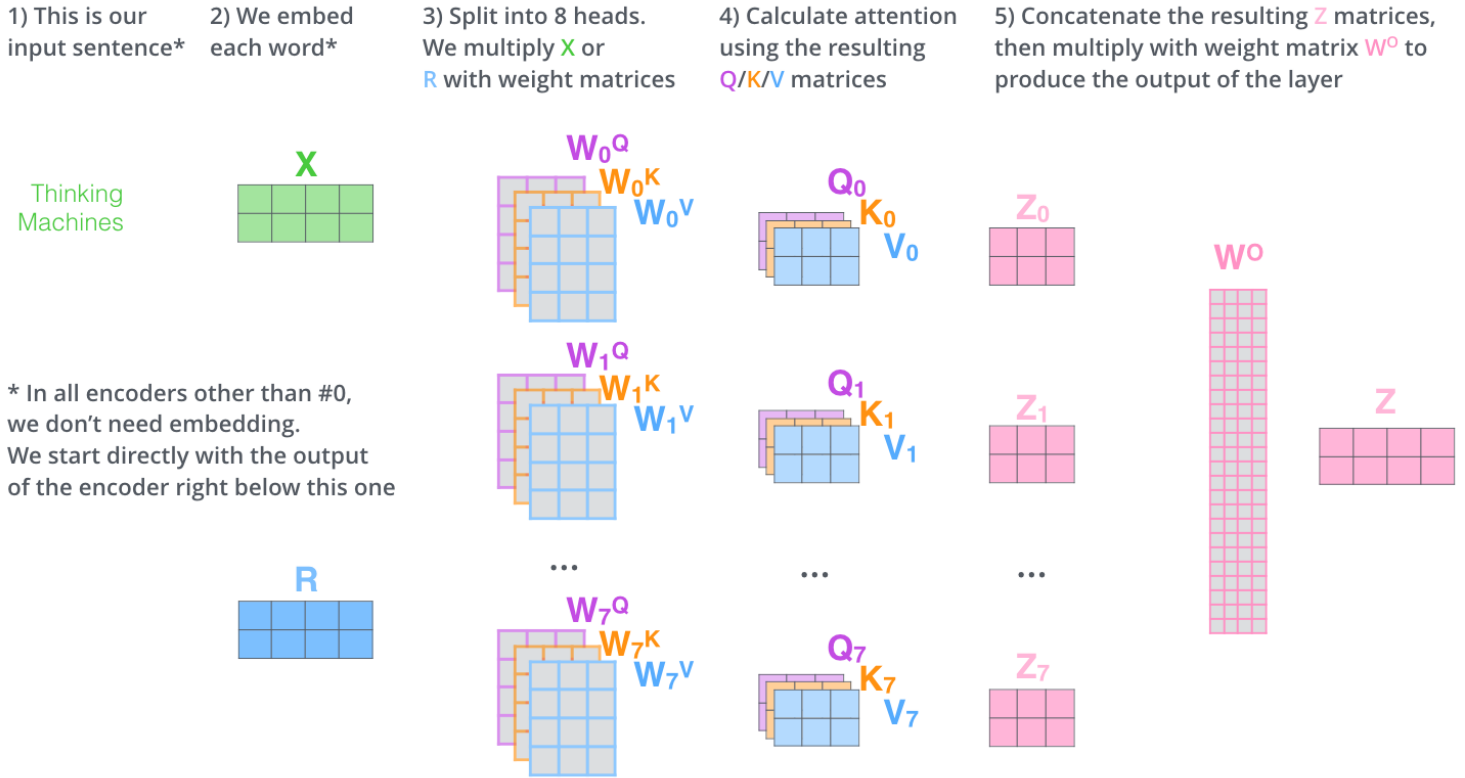


Множественное внимание

❑ Формирование нескольких представлений (параллельно) для входной последовательности с последующей конкатенацией

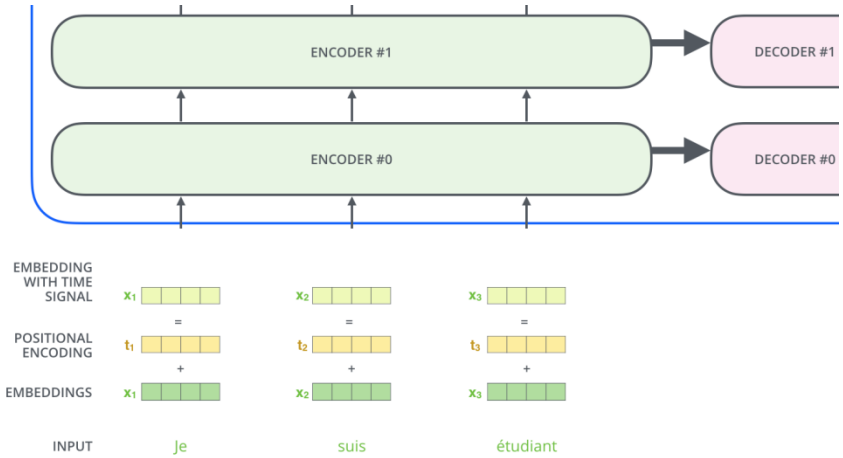


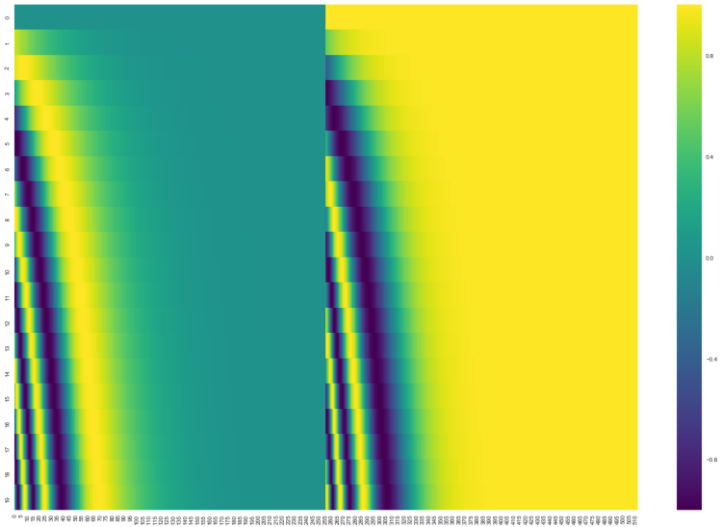
Пример множественного внимания



Позиционное кодирование

- ❑ Необходимо учитывать порядок слов в последовательности.
- ❑ К вектору эмбединг слова добавляется вектор, определяющий его позицию в предложении.

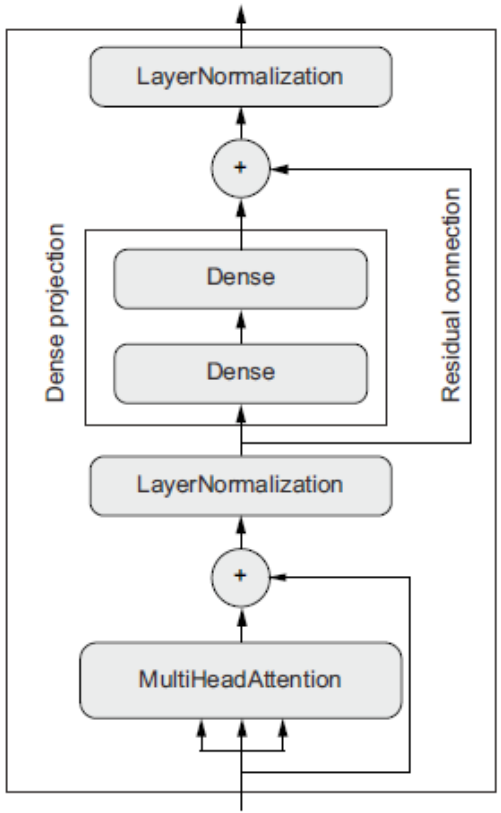




Пример позиционного кодирования для 20 слов. Длина вектора - 512

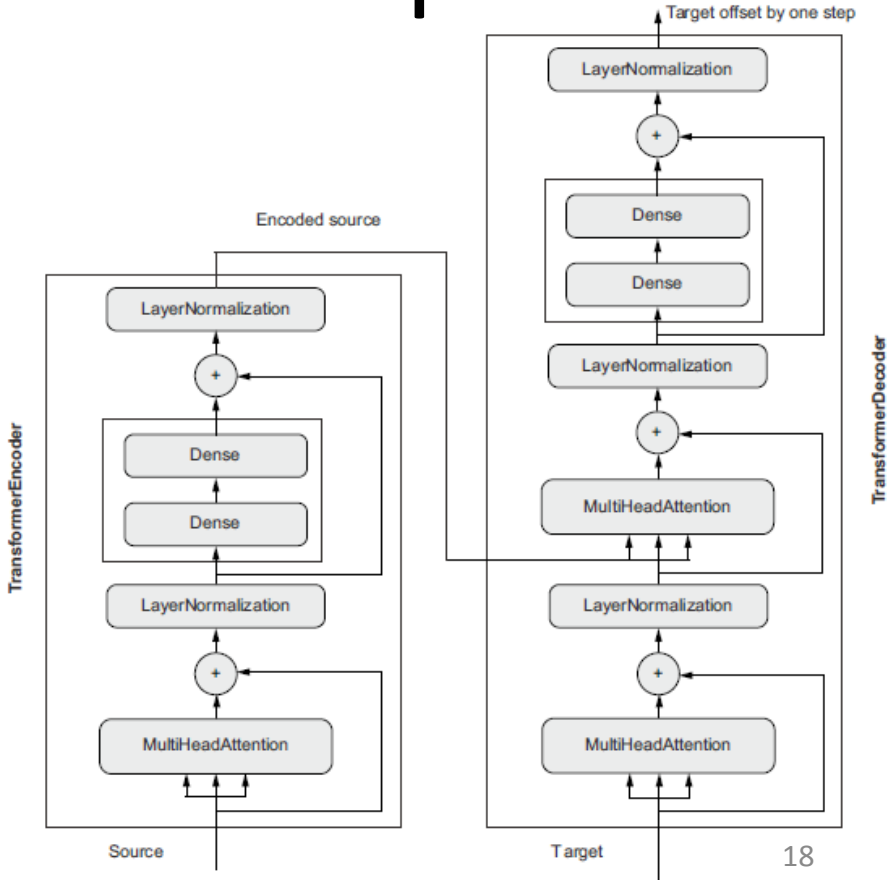
Архитектура энкодера

- Энкодер + классификатор = решение для классификации текстов.
- Энкодер + Декодер = модель для задач последовательность – последовательность (sequence-to-sequence)
  - перевод
  - составление абстракта для текста
  - ответ на вопрос
  - генерация текста.



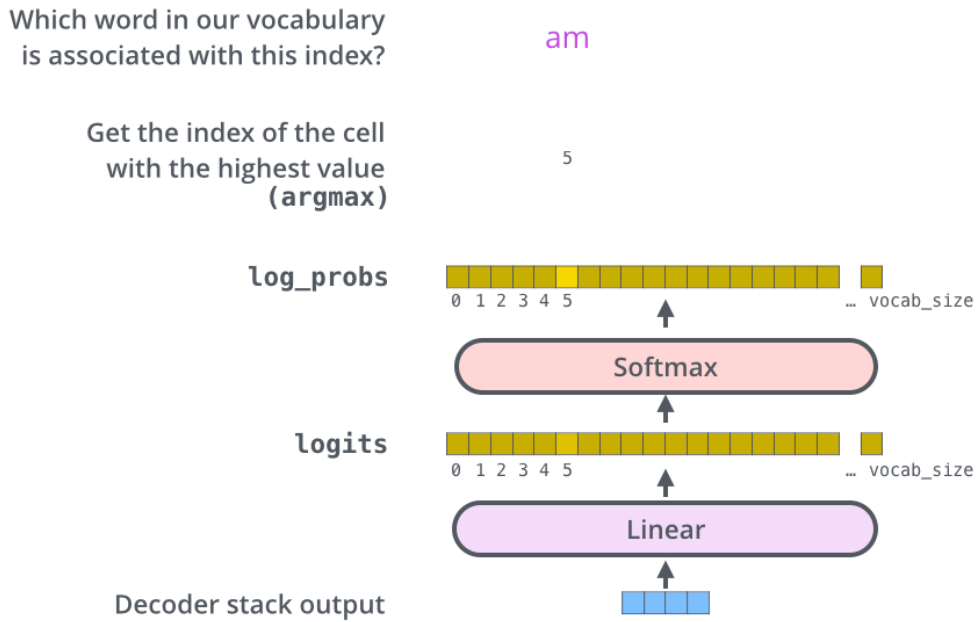
Модель sequence-to-sequence

- Декодер предсказывает выходное слово на основании его взаимосвязи со словами в исходной последовательности:
  - выходное слово – запрос
  - исходная последовательность – ключ, значение



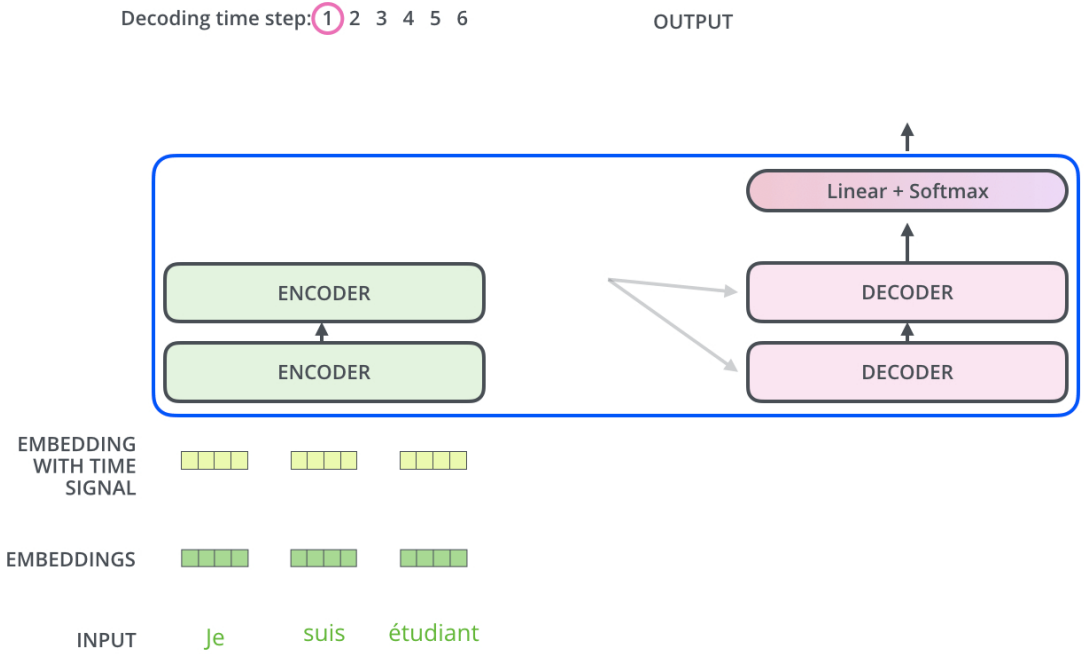
Генерация выходного слоя

- Линейный слой + softmax = вектор (размер словаря)



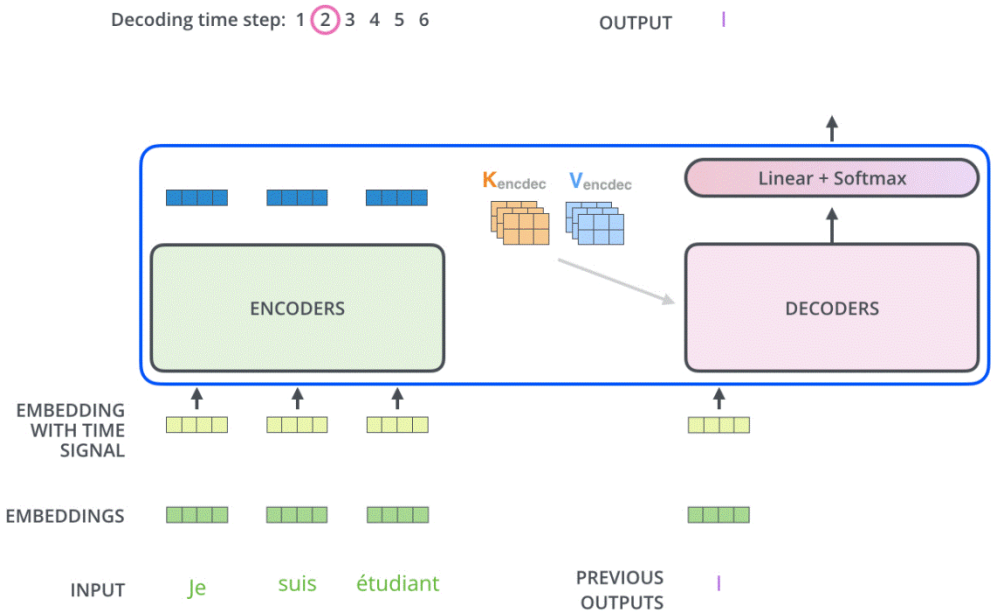
Пример sequence-to-sequence(1)

❑ Кодирование входной последовательности, получение первого выходного слова.



Пример sequence-to-sequence(2)

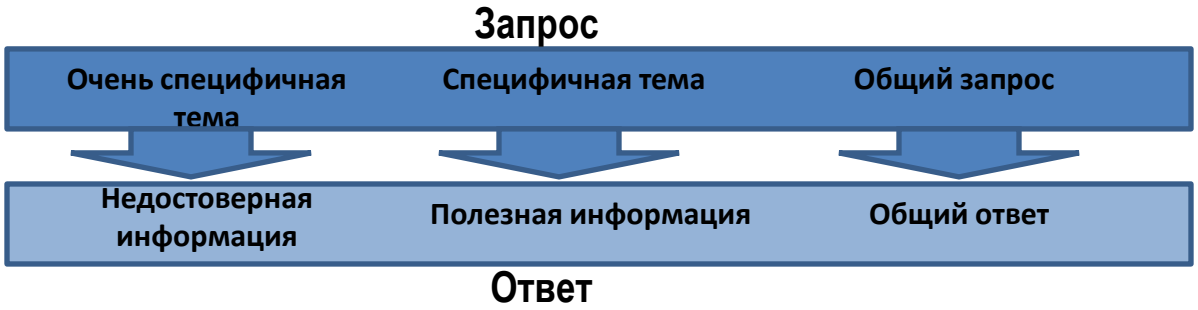
- Вызов декодера с текущим сгенерированным фрагментом пока не появится флаг окончания генерируемой последовательности.



ChatGPT. Недостатки

- ❑ Единственный запрос не эффективен → последовательность детализированных запросов
- ❑ Достоверность ответов:

  - например, запрос цитаты подтверждающей какой-то факт → несуществующий автор, несуществующая книга, цитата относится к другой теме.
- ❑ Вероятностная модель: ответ зависит от контекста
- ❑ Баланс Популярность-Специфичность



https://habr.com/ru/post/713606/

ChatGPT. Возможности

- ❑ Поиск: Можно быстро находить нужную информацию, качество зависит от темы
- ❑ Разделение запроса на части и формулирование новых запросов на основе ответов повышает эффективность.
- ❑ Объяснение сложного концепта на разных уровнях сложности, например, идея книги.
- ❑ Освоение новых навыков, например изучение иностранного языка (написание текстов на заданную тему).
- ❑ Быстрое изучение новых областей знаний: получение основных концепций, идей, взаимосвязей (быстрее, чем через интернет поиск).

https://habr.com/ru/post/713606/

Литература

- François Chollet. Deep Learning with Python, Second Edition, 2021
- Transformer в картинках (https://habr.com/ru/post/486358/)