Механизм самовнимания в трансформерах

Self-attention — база современных трансформеров для параллельной и контекстной обработки

Выполнили:
Королёва Дарья
Кан Игорь

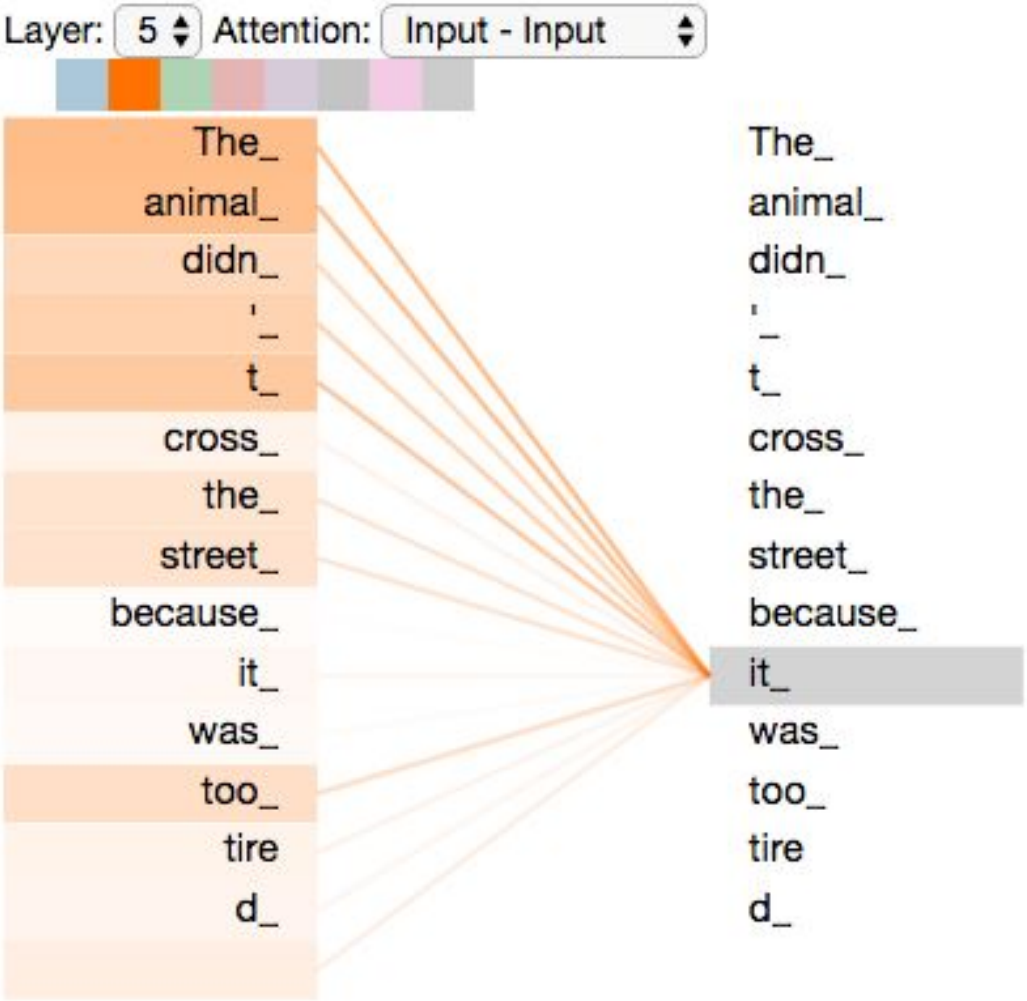
Что такое трансформеры?

Определение трансформеров

Трансформеры — это архитектура нейронных сетей, разработанная для обработки последовательностей данных. Они позволяют эффективно моделировать зависимости между элементами без рекуррентной обработки.

Основные особенности архитектуры

Ключевым элементом трансформеров является механизм self-attention, который оценивает взаимосвязи всех позиций во входной последовательности параллельно, обеспечивая высокую скорость и точность обучения.



Как работает Self-Attention

Например, в фразе *“it was too tired”* модель понимает, что “it” относится к *animal*, а не к *street*.

Введение в self-attention

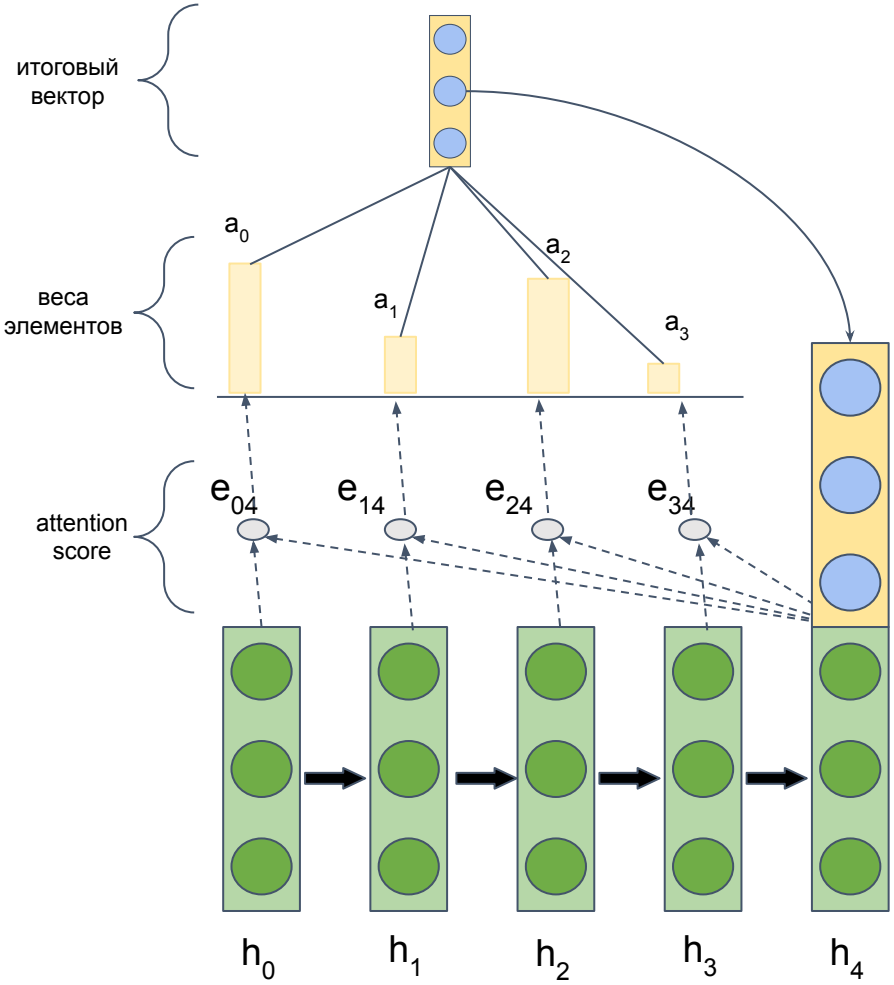
01 Self-attention формирует матрицы Q, K, V из одних и тех же эмбеддингов последовательности, позволяя моделям учесть все пары зависимостей между элементами.

02 Матрица внимания A размером n×n вычисляется с учётом позиционного кодирования для сохранения порядка и создаёт распределение весов для агрегации информации. **Каждый элемент этой матрицы — это вес, показывающий, насколько слово *j* важно для слова *i*.**

Q — Query
«Что я ищу в контексте?»

K — Key
«Какое ключевое содержание есть у других слов?»

V — Value
«Какую информацию я могу дать?»



$$a_i = \mathrm{softmax}_i([e_{1j}, e_{2j}, \dots e_{j-1,j}])$$

$$e_{ij} = h_i^T h_j$$

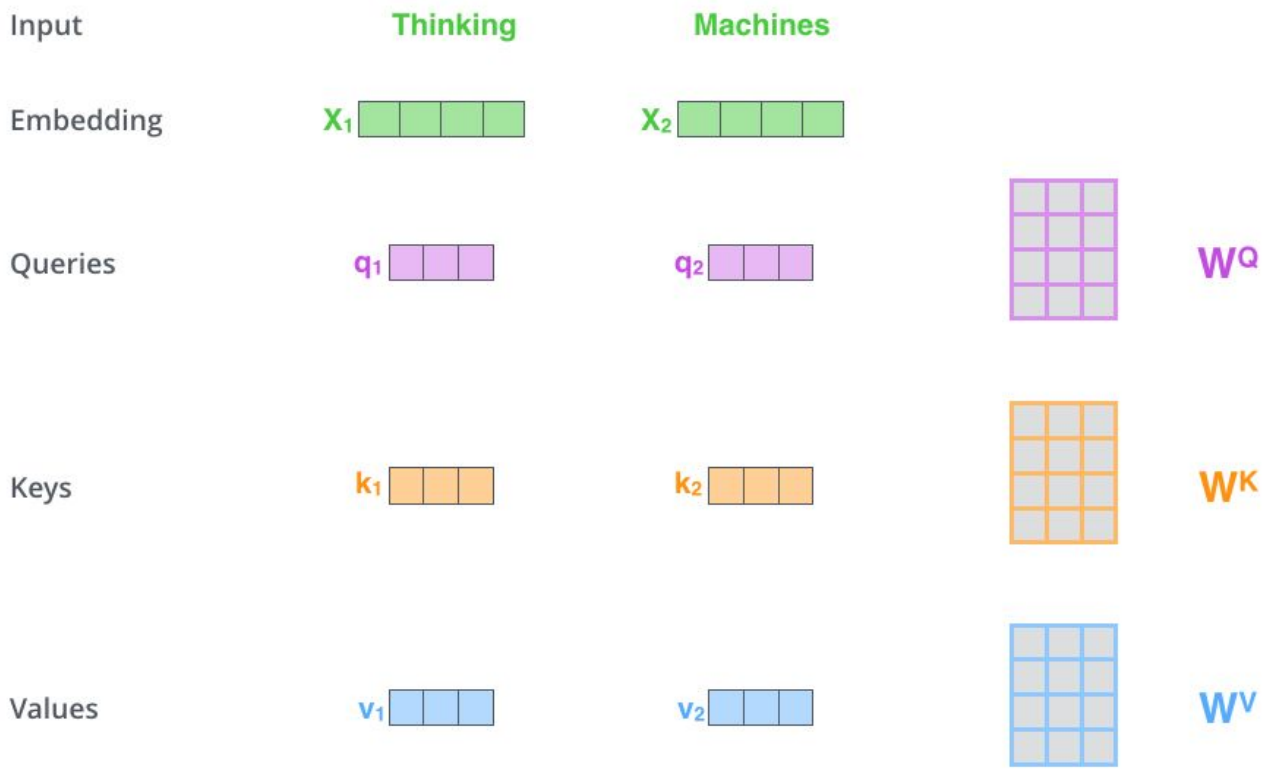
Алгоритм Self-Attention

Для каждого слова формируются три вектора: Query, Key и Value.

$$Q_i = x_i W^Q \qquad K_i = x_i W^K \qquad V_i = x_i W_V$$

Query сравнивается со всеми Key, чтобы вычислить «важность» связей между словами.

Эти веса умножаются на Value, что даёт новое представление слова с учётом контекста.



Алгоритм внимания

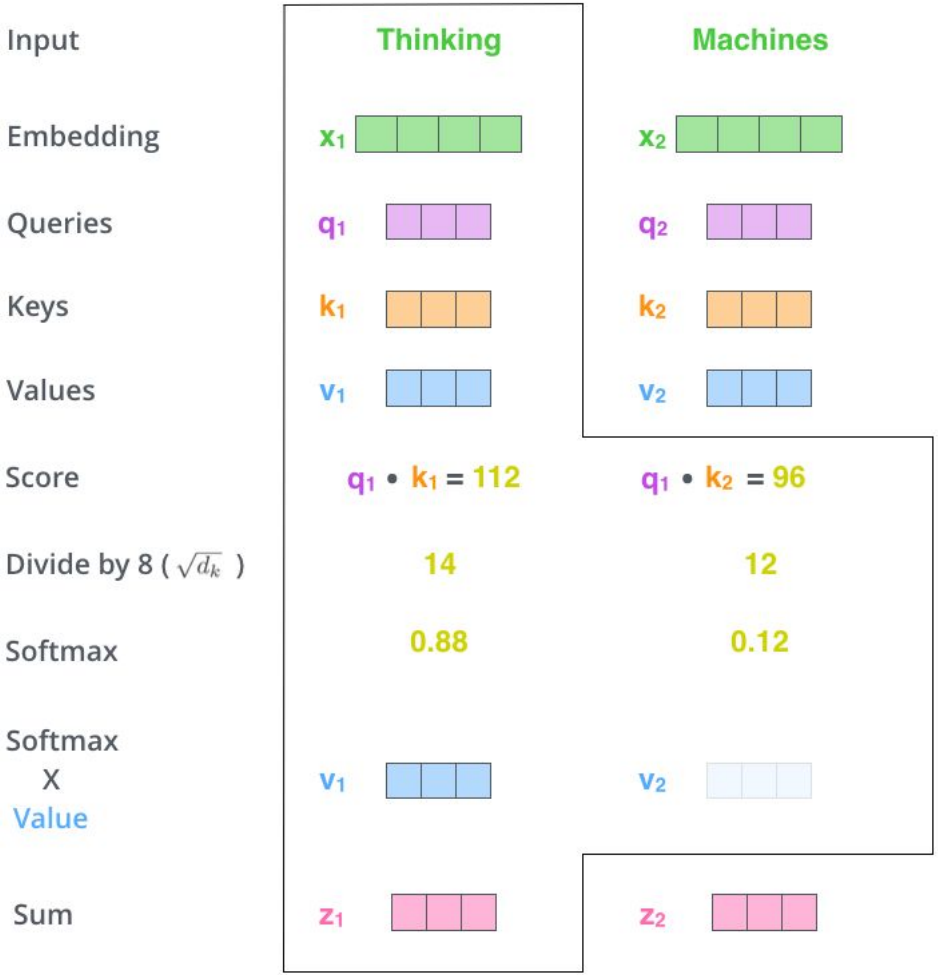
1) Query сравнивается с Key других слов через скалярное произведение, формируя оценки важности.

2) Масштабирование: эти оценки делятся на √d

3) Нормализация: применяется функция softmax.

4) Умножить каждое Value на соответствующий softmax-вес.

5) Просуммировать взвешенные Value — это и есть новое представление слова.



Список литературы

Vaswani A. et al. Attention is All You Need. 2017. Основополагающая статья, вводящая архитектуру трансформеров и механизм self-attention.

Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. Важная работа по контекстным эмбеддингам на базе трансформеров.