Современное состояние NLP. Базовые методы векторизации текстов.

First Step in NLP: 2.0

План рассказа

- Современное состояние NLP
- Базовые методы векторизации текстов
- Практика

Современное состояние NLP

Задачи обработки языка (NLP)

- Классификация текстов/документов
- Распознавание именных сущностей (NER)
- Машинный перевод
- Вопрос-ответные системы (QA)
- Извлечение информации (IR)
- Суммаризация текстов/документов
- Генерация текста

и множество других

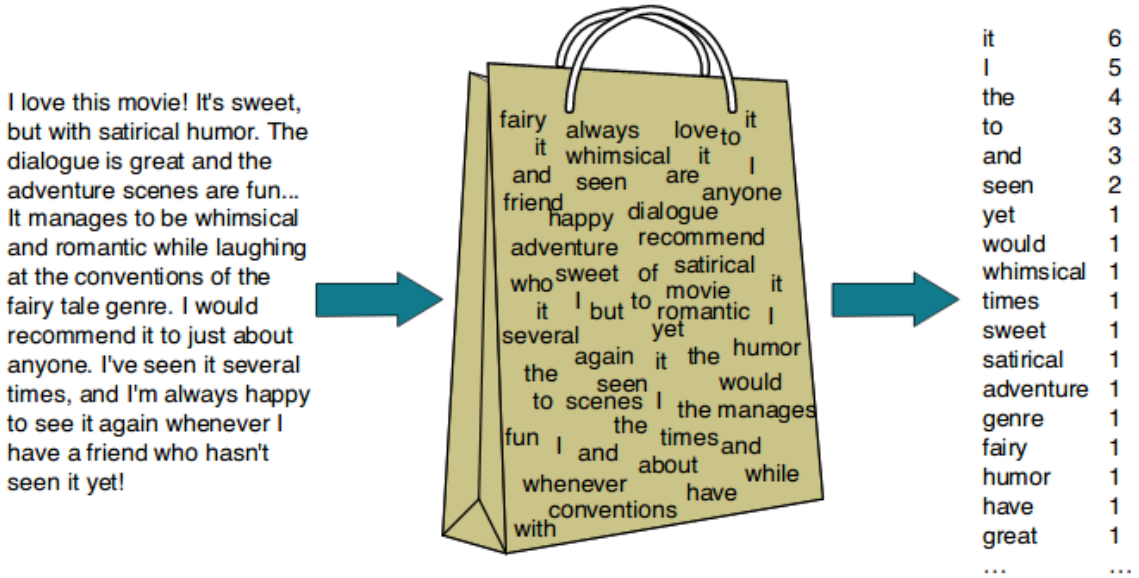
Реальные кейсы применения NLP

- Поисковые системы (Google, Yandex, etc.)
- Машинный перевод (Google Translate, Abbyy Lingvo, Linguee, etc.)
- Виртуальные ассистенты, голосовые помощники etc.
- Фильтр спама (e-mail / телефонный / etc.)
- Дополнение текста, автокоррекция
- Авторазметка отзывов пользователей
- Чат-боты
- Автосуммаризация текста

Подходы к решению NLP-задач

Классическое машинное обучение:

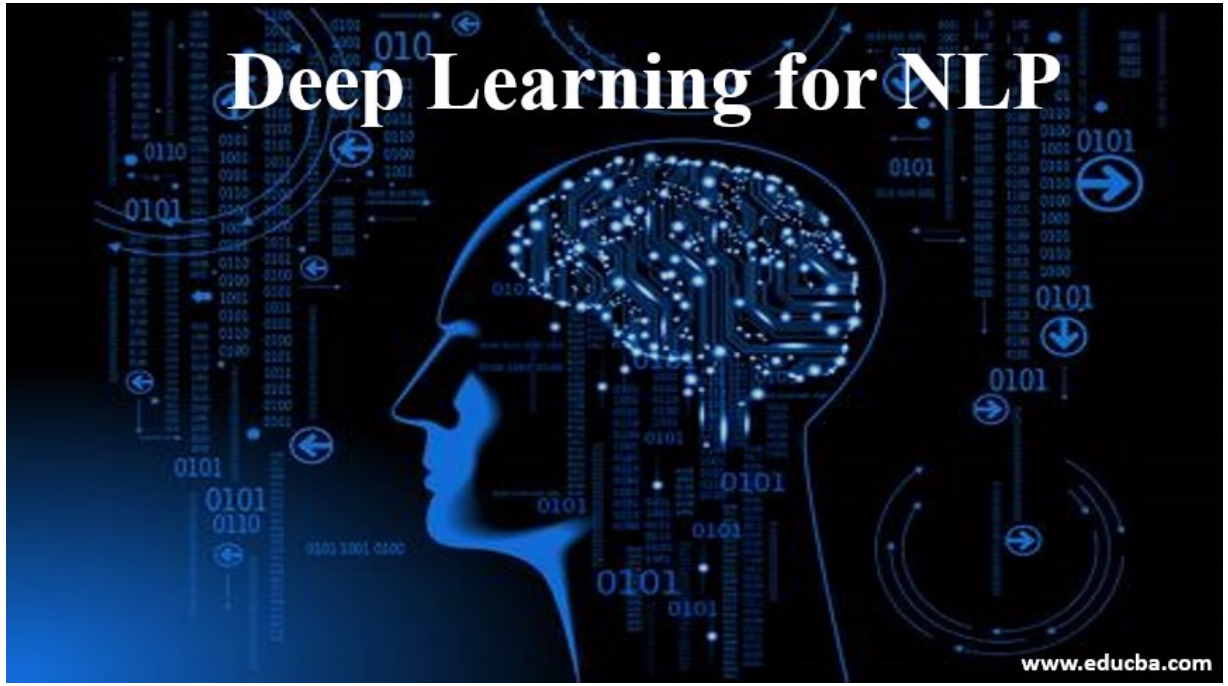
- Извлекаем признаки из текстов (bag of words, tf-idf)
- На этих признаках обучаем ML-модель



Подходы к решению NLP-задач

Глубинное обучение:

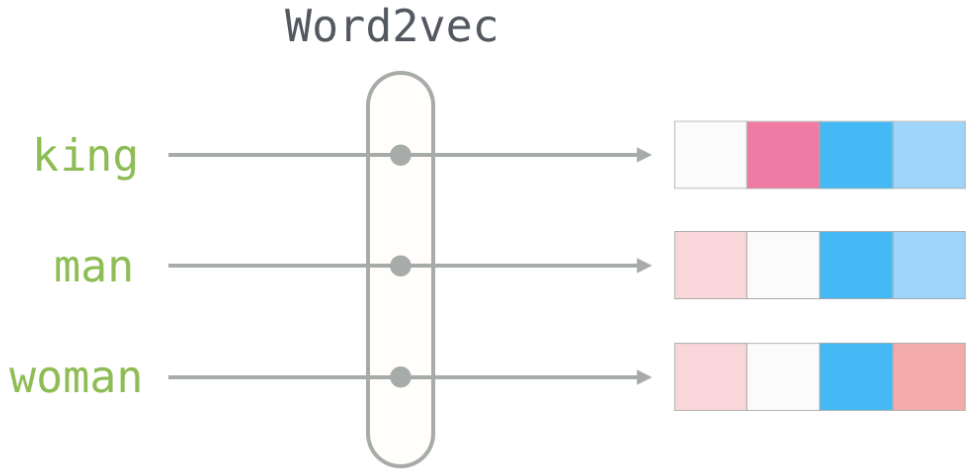
- Нейронные сети самостоятельно извлекают необходимую информацию из текстов



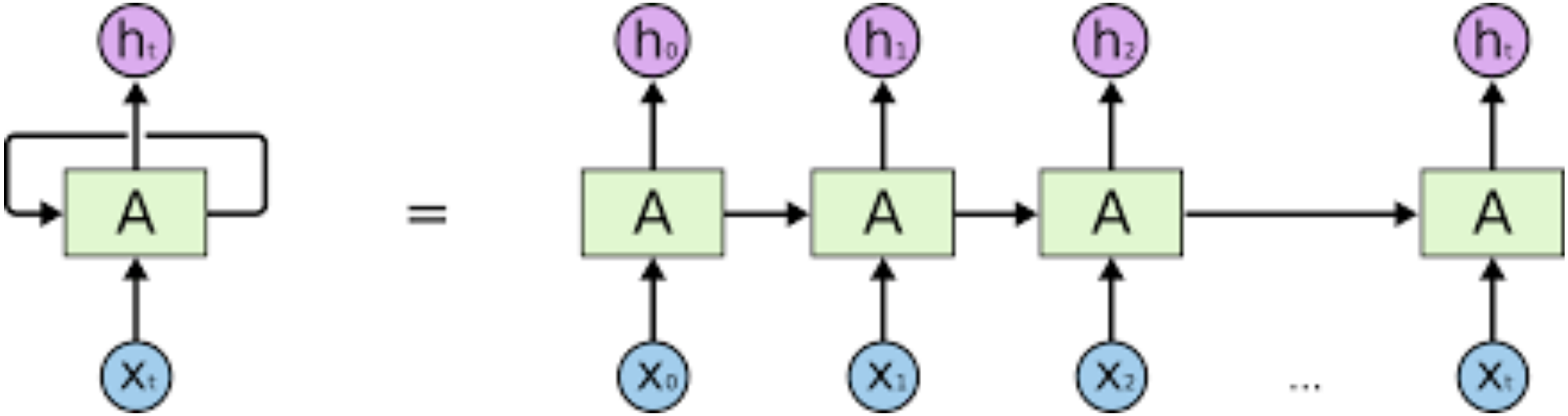
Глубинное обучение в NLP

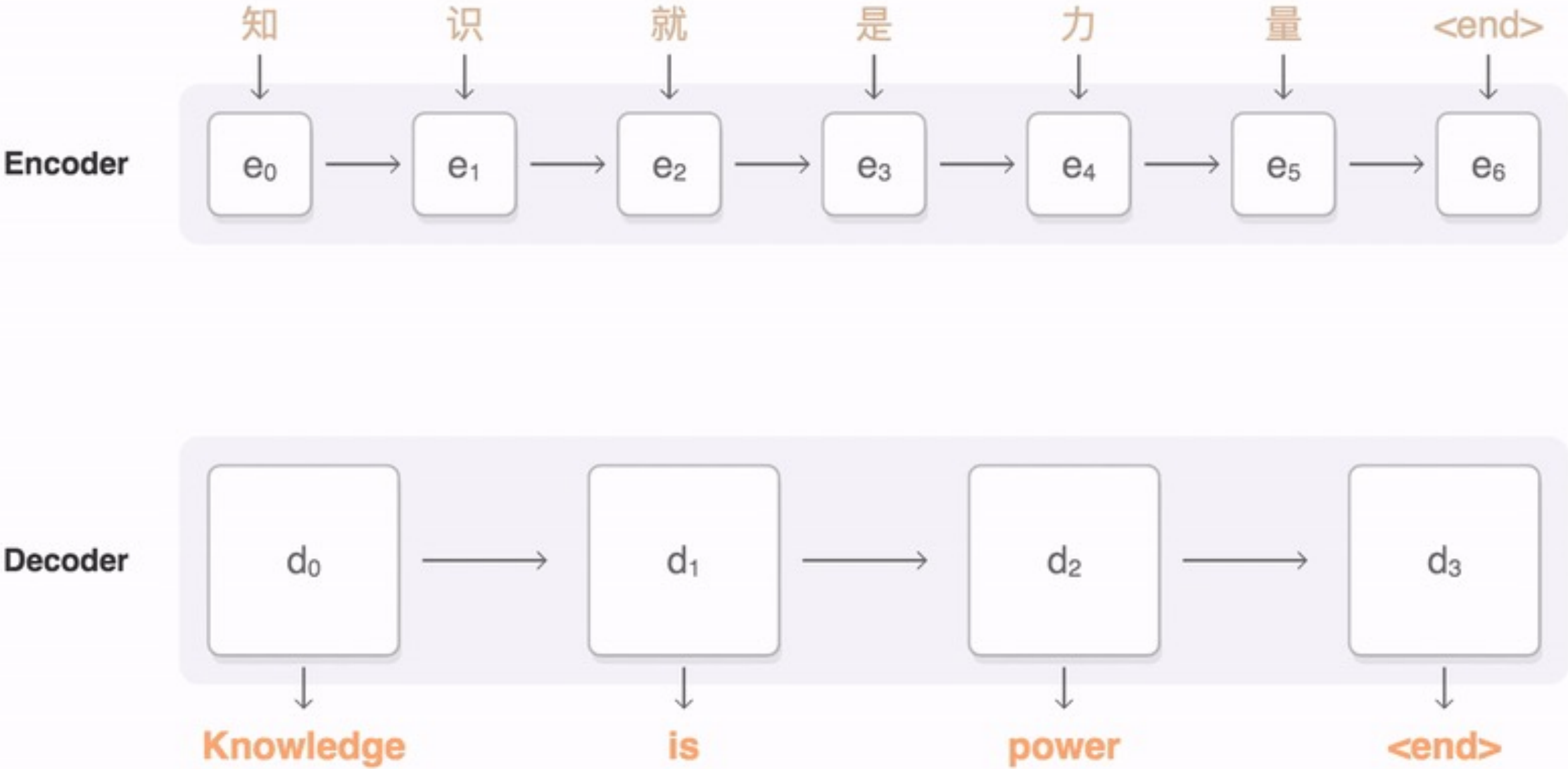
Виды нейронных сетей:

1. Полносвязные нейронные сети (основа основ) – word2vec (2013), fasttext, GloVe
2. Рекуррентные нейронные сети
3. Attention и трансформеры (с 2017)



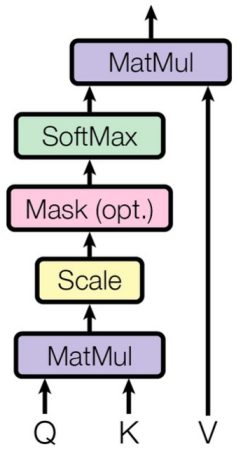
2. Рекуррентная нейронная сеть



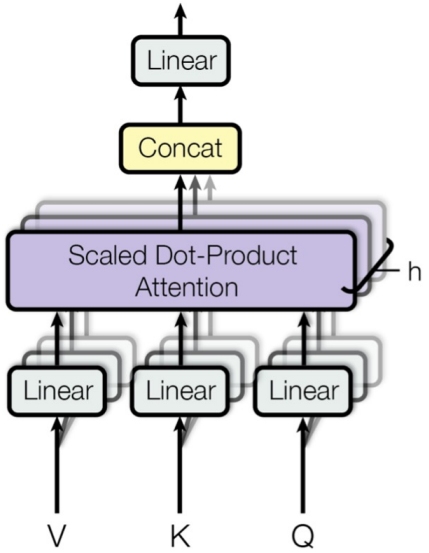

知
识
就
是
力
量
<end>
Encoder
e0
e1
e2
e3
e4
e5
e6
Decoder
d0
d1
d2
d3
Knowledge
is
power
<end>

3. Transformer и Attention

Scaled Dot-Product Attention

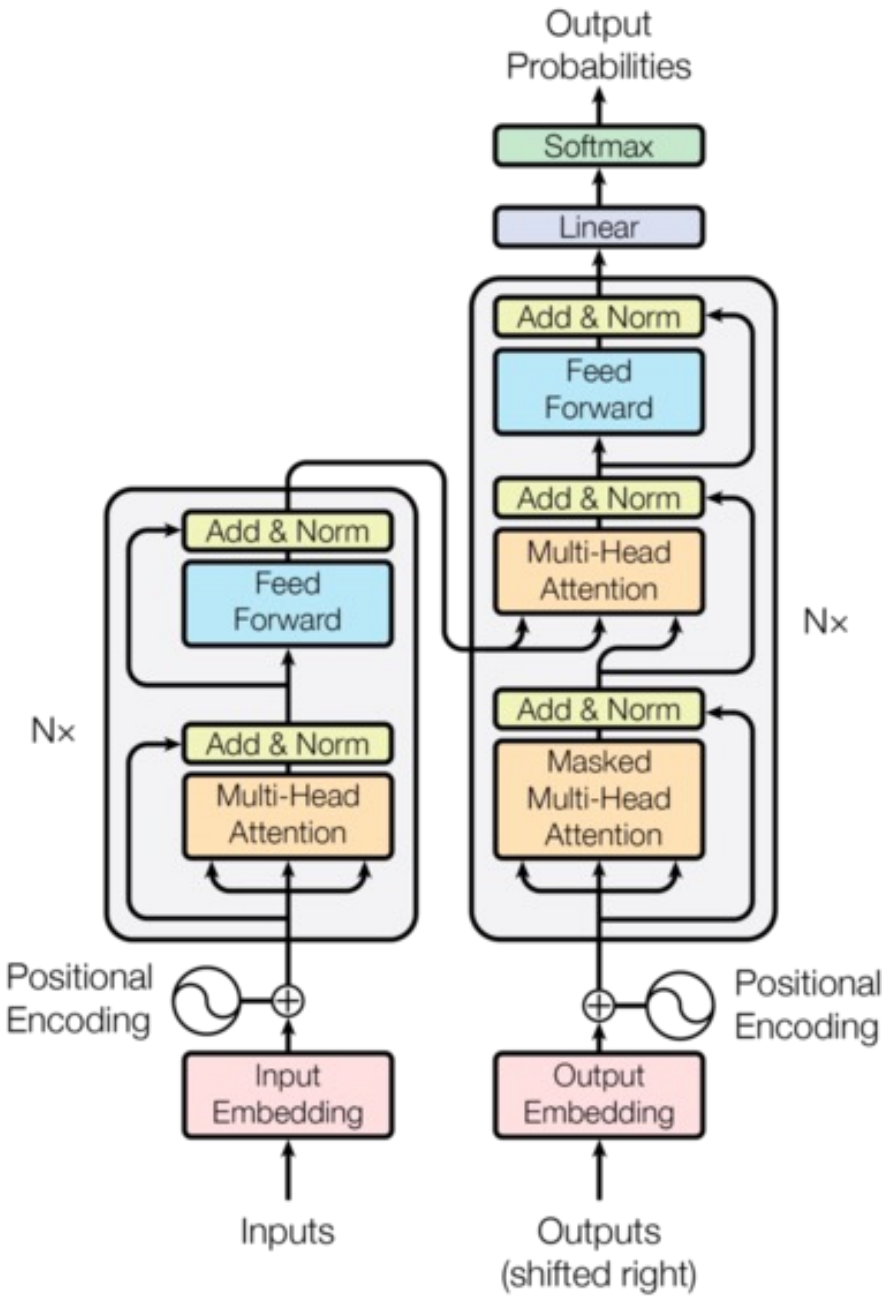


Multi-Head Attention



$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

TRANSFORMER



Зачем нужны простые методы, если есть трансформеры?

- Скорость (векторизации, обучения и применения модели)
- Простота и прозрачность интерпретации
- Достаточно высокое качество моделей, обученных на простых векторизациях

Зачем нужны простые методы, если есть трансформеры? Томаш Миколов (Dec. 2023)

Word2Vec – самый простой DL-подход к векторизации текстов. Однако он остается популярным и активно применяется в задачах уже более 10 лет!

14 декабря 2023 года Томаш Миколов опубликовал сообщение по поводу получения награды за свою статью про word2vec. Вот интересные мысли оттуда:

- *"Yesterday we received a Test of Time Award at NeurIPS for the word2vec paper from ten years ago*
- *In fact, the original word2vec paper was rejected at the first ICLR conference in 2013*
- *The code ended up being over-optimized because I was waiting for many months for approval to publish it*
- *We did show that word2vec is much better than GloVe when trained on the same data.*
- *I published the first ever study showing that neural nets beat n-gram language models increasingly more with more training data when everything is done correctly (today this sounds obvious, but back in the days this was widely considered impossible - even most Google guys did think that the more data you have, the more futile is to work on anything besides n-grams and smoothing techniques).*

Полный текст сообщения.