

Глубокое обучение и вообще

Ульянкин Филипп

8 июня 2022 г.

Посиделка 13: The Mappet show (улица Сезам)

Agenda

- Attention is all you need
- Трансформеры
- BERT ROBERTA ALBERT and co
- GPT

Attention is All You Need!

Attention is All You Need! (2017)

- RNN это очень долго! Всегда, чтобы найти следующий токен, надо знать предыдущий
- Backward pass идёт ещё и через время :(
- Transformer — нейросетевая архитектура для задач seq2seq, основанная исключительно на полно связных слоях
- Превзошла существовавшие seq2seq архитектуры как по качеству, так и по скорости работы
- Основной элемент — multi-head self-attention

Attention is All You Need! (2017)

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Llion Jones*
Google Research
llion@google.com

Illia Polosukhin* [‡]
illia.polosukhin@gmail.com

Noam Shazeer*
Google Brain
noam@google.com

Aidan N. Gomez* [†]
University of Toronto
aidan@cs.toronto.edu

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Lukasz Kaiser*
Google Brain
lukaszkaiser@google.com

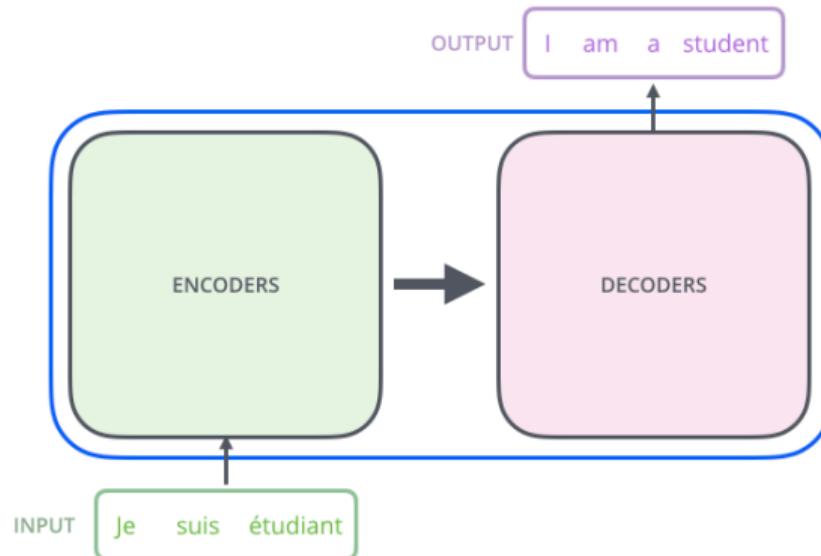
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

<https://arxiv.org/abs/1706.03762>

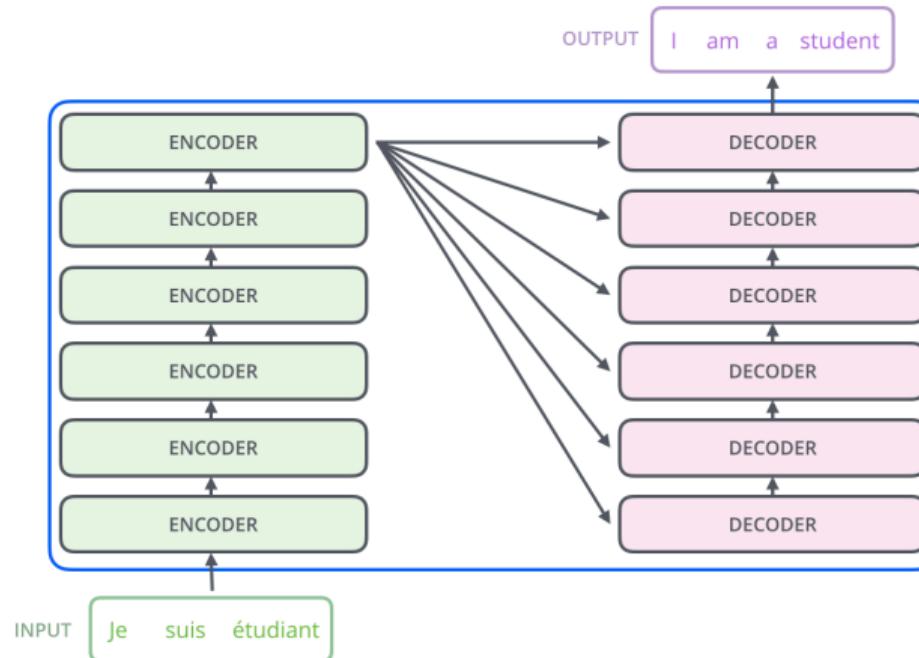
Transformer

Верхнеуровнного - это просто энкодер и декодер



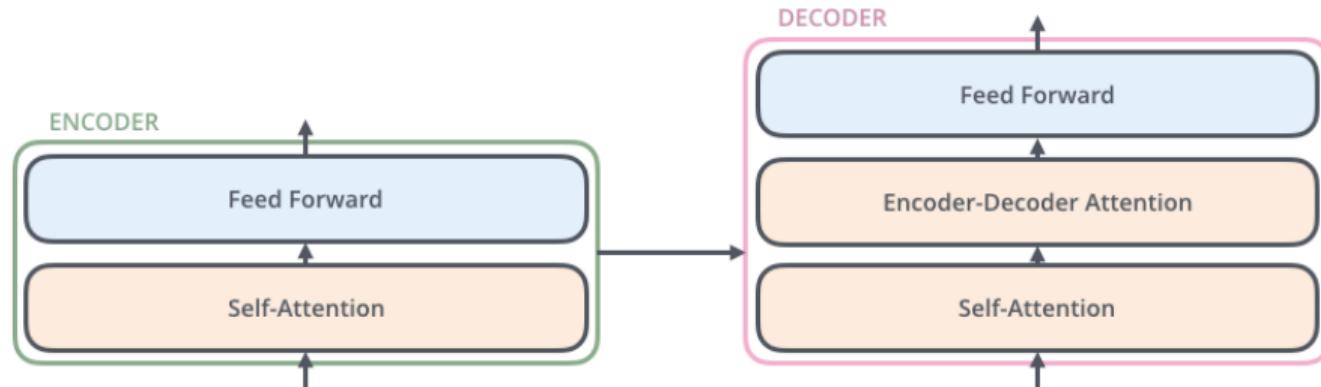
Transformer

Энкодер и декодер состоят из одинаковых блоков; веса во всех блоках разные



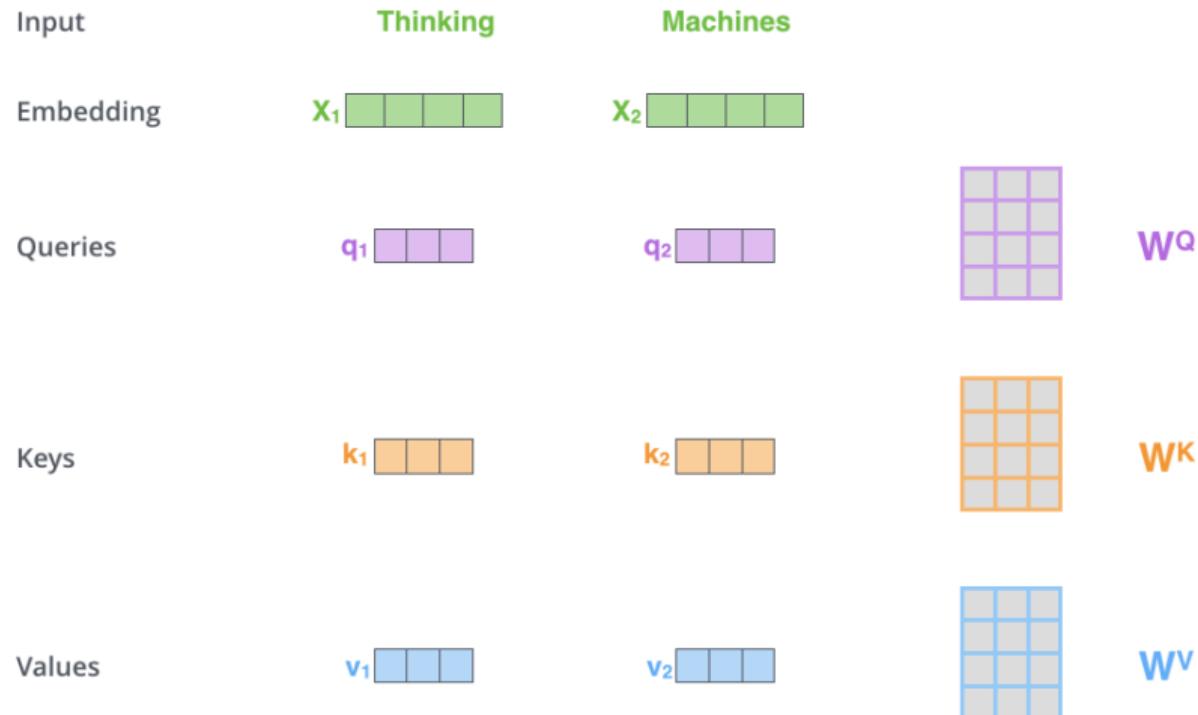
Transformer

В энкодере происходят две вещи: сначала вход прогоняется через self-attention, а затем — через полносвязный слой. В декодере помимо обычного self-attention есть ещё и attention из энкодера.



<http://jamalmar.github.io/illustrated-transformer/>

Self-attention



Абстракции

- Для каждого входного слова считаются три вектора: Query, Key и Value
- Матрицы W^Q , W^K , W^V обучаются вместе с моделью
- Value - то, что мы знаем об этом слове
- Query, Key помогают искать связи между словами, мы ходим по всем словам и пытаемся понять насколько они связаны между собой
- Query - мое текущее слово, Key - мое слово с которым я сравниваю себя

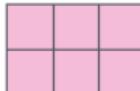
Self-attention

Цель этого слоя — сложить Value с некоторыми весами

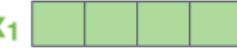
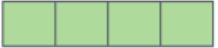
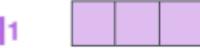
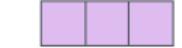
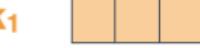
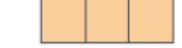
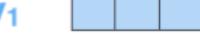
$$\text{softmax} \left(\frac{\begin{matrix} \mathbf{Q} & \times & \mathbf{K}^T \\ \begin{matrix} \text{purple} & \end{matrix} & \times & \begin{matrix} \text{orange} & \end{matrix} \\ \hline \end{matrix}}{\sqrt{d_k}} \right) \mathbf{V}$$

\mathbf{Q} \mathbf{K}^T \mathbf{V}

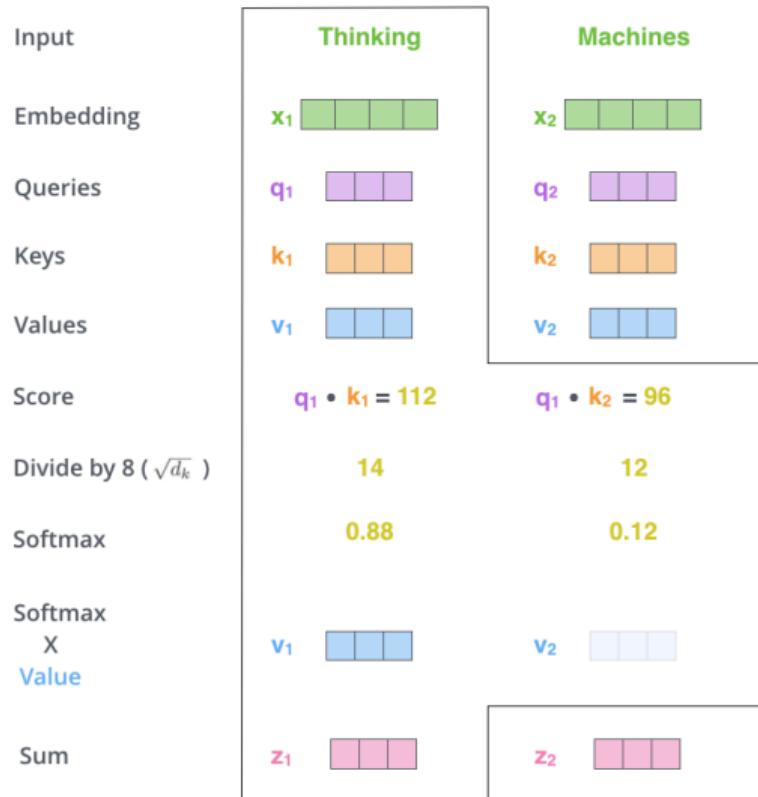
\mathbf{Z}

= 

Более детально

Input	Thinking		Machines	
Embedding	x_1		x_2	
Queries	q_1		q_2	
Keys	k_1		k_2	
Values	v_1		v_2	
Score	$q_1 \cdot k_1 = 112$		$q_1 \cdot k_2 = 96$	
Divide by 8 ($\sqrt{d_k}$)	14		12	
Softmax	0.88		0.12	

Более детально



Query, Key, Value

Each vector receives three representations (“roles”)

$$[W_Q] \times \begin{array}{c} \text{green} \\ \text{green} \\ \text{green} \end{array} = \begin{array}{c} \text{blue} \\ \text{blue} \\ \text{blue} \end{array}$$

Query: vector from which the attention is looking

“Hey there, do you have this information?”

$$[W_K] \times \begin{array}{c} \text{green} \\ \text{green} \\ \text{green} \end{array} = \begin{array}{c} \text{yellow} \\ \text{yellow} \\ \text{yellow} \end{array}$$

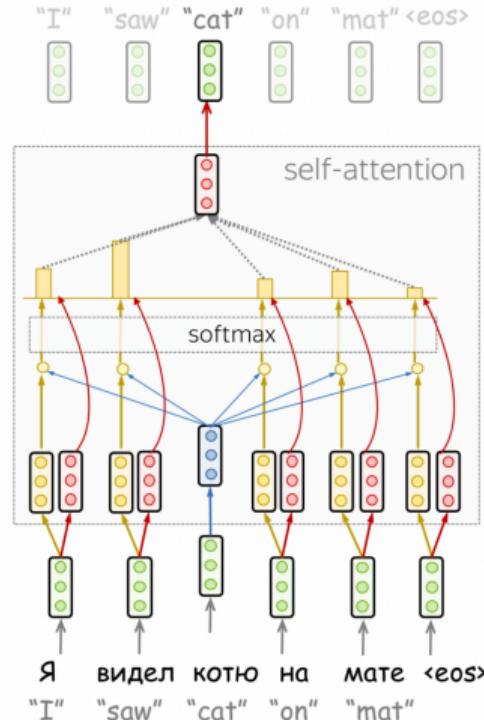
Key: vector at which the query looks to compute weights

“Hi, I have this information – give me a large weight!”

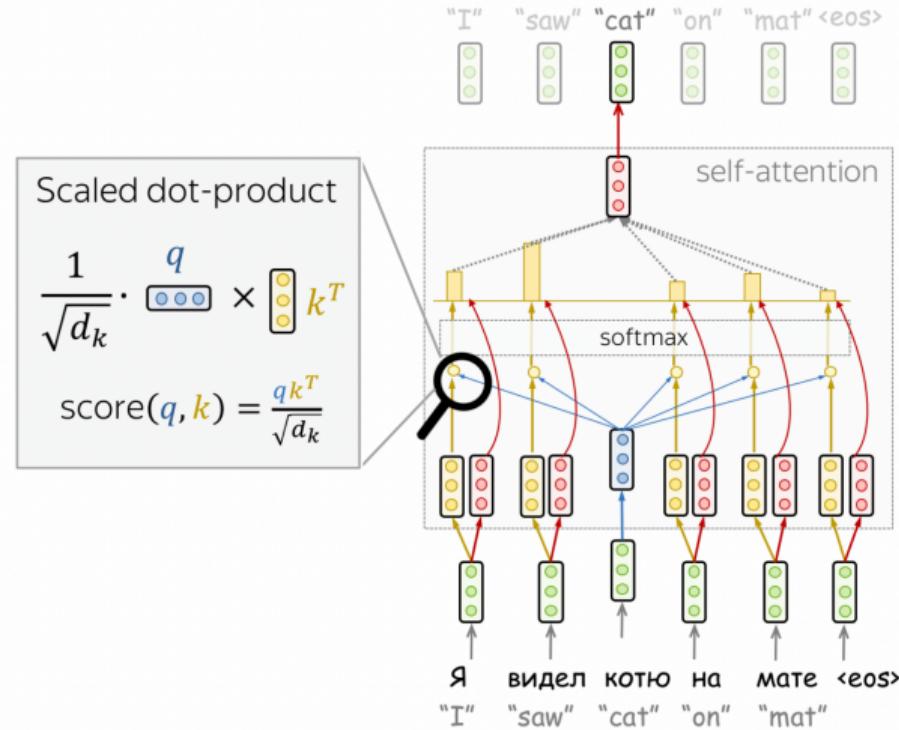
$$[W_V] \times \begin{array}{c} \text{green} \\ \text{green} \\ \text{green} \end{array} = \begin{array}{c} \text{pink} \\ \text{pink} \\ \text{pink} \end{array}$$

Value: their weighted sum is attention output

“Here’s the information I have!”



Query, Key, Value



https://github.com/yandexdataschool/nlp_course/tree/2021/week04_seq2seq

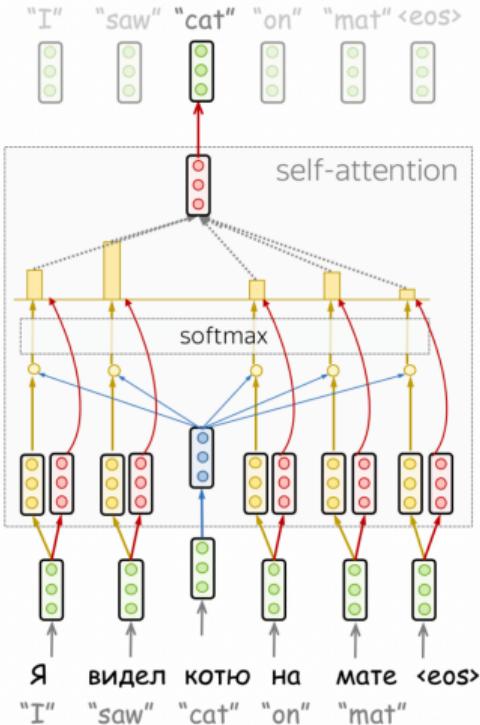
Query, Key, Value

$$\text{Attention}(q, k, v) = \text{softmax}\left(\frac{qk^T}{\sqrt{d_k}}\right)v$$

Attention weights

from to

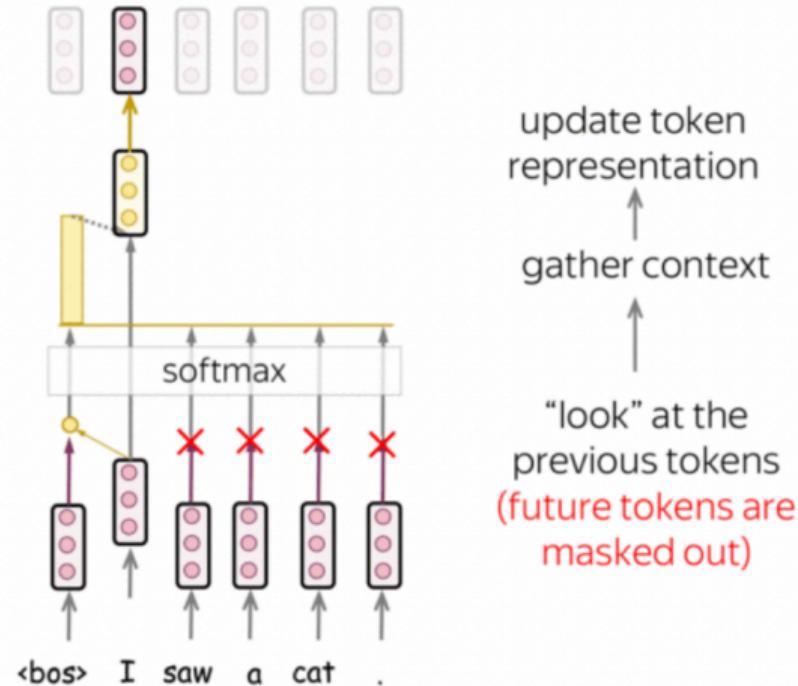
vector dimensionality of K, V



Masked Self-Attention

In the decoder, we forbid looking at future tokens – we don't know them

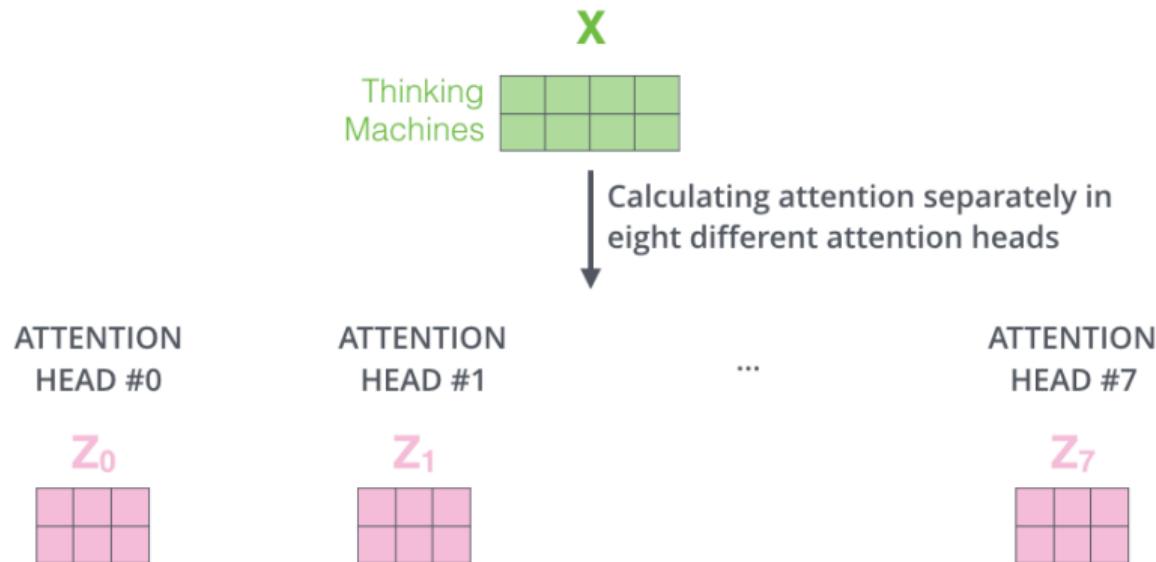
Note: in training, decoder processes all target tokens at once – without masks, it would see future



Зверь с кучей голов

Зверь с кучей голов

Несколько голов обеспечивают разное внимание



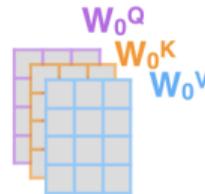
Слой целиком

1) This is our input sentence*
2) We embed each word*

Thinking
Machines



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



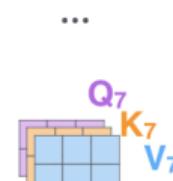
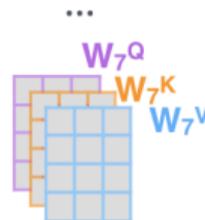
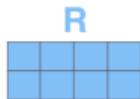
4) Calculate attention using the resulting $Q/K/V$ matrices



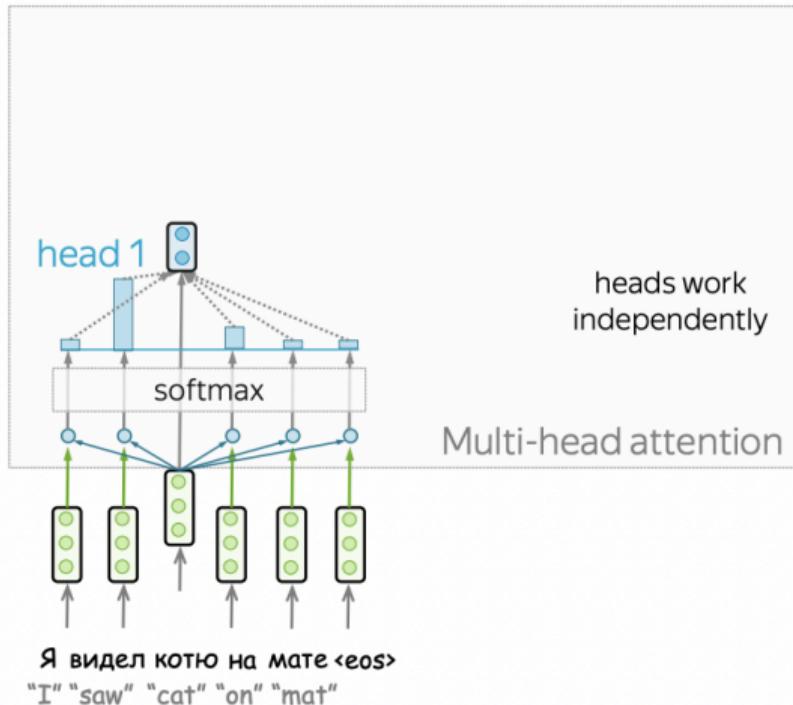
5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer



* In all encoders other than #0, we don't need embedding.
We start directly with the output of the encoder right below this one

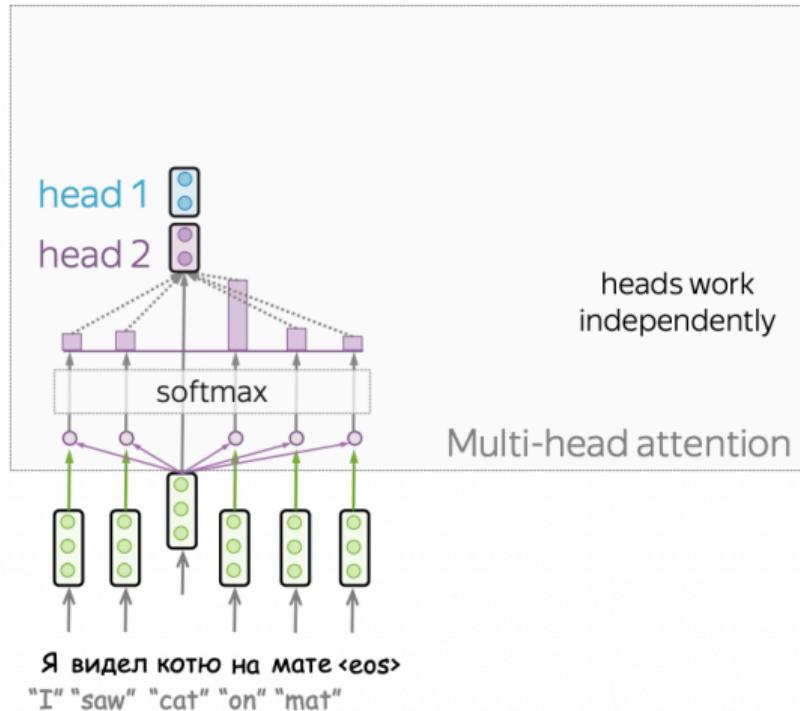


Multi-Head Attention



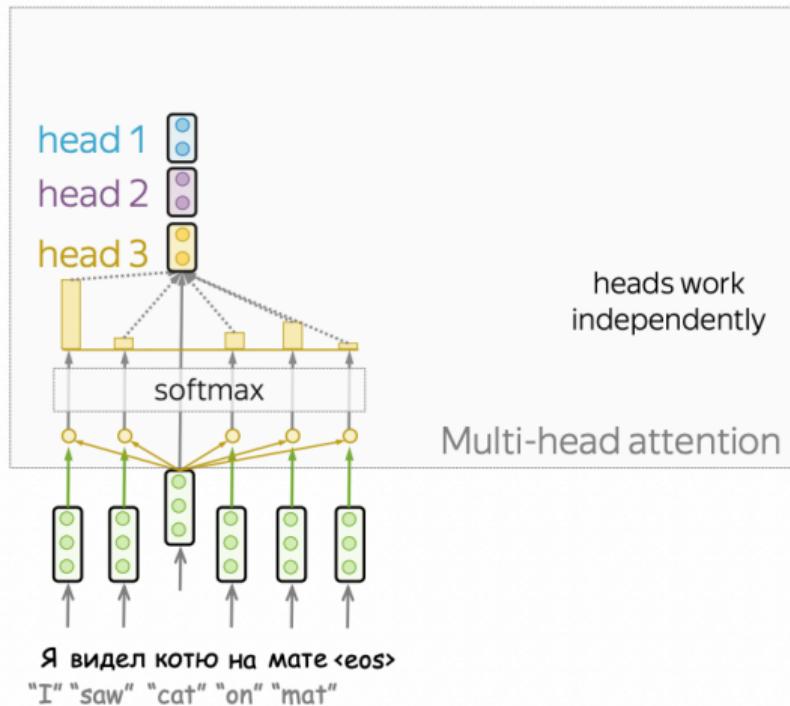
https://github.com/yandexdataschool/nlp_course/tree/2021/week04_seq2seq

Multi-Head Attention



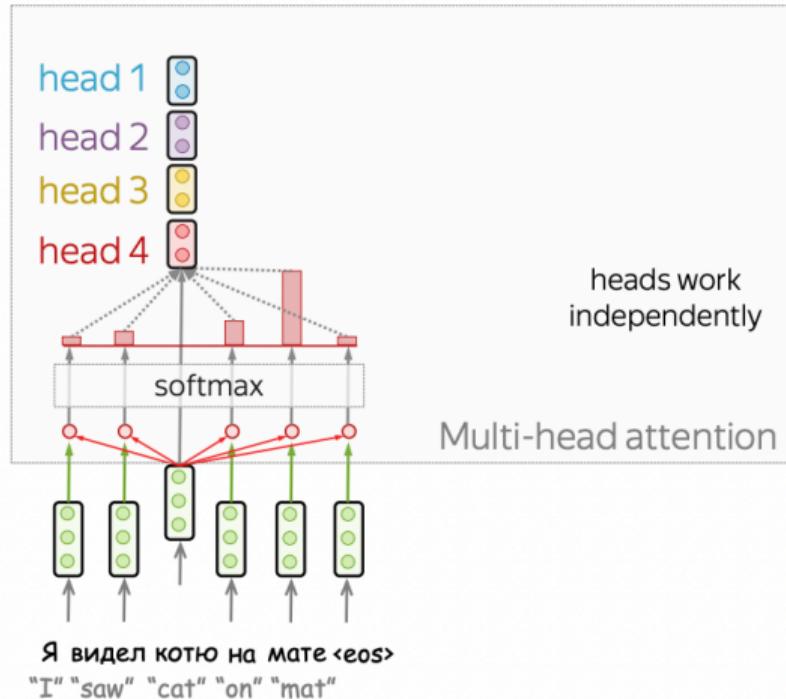
https://github.com/yandexdataschool/nlp_course/tree/2021/week04_seq2seq

Multi-Head Attention

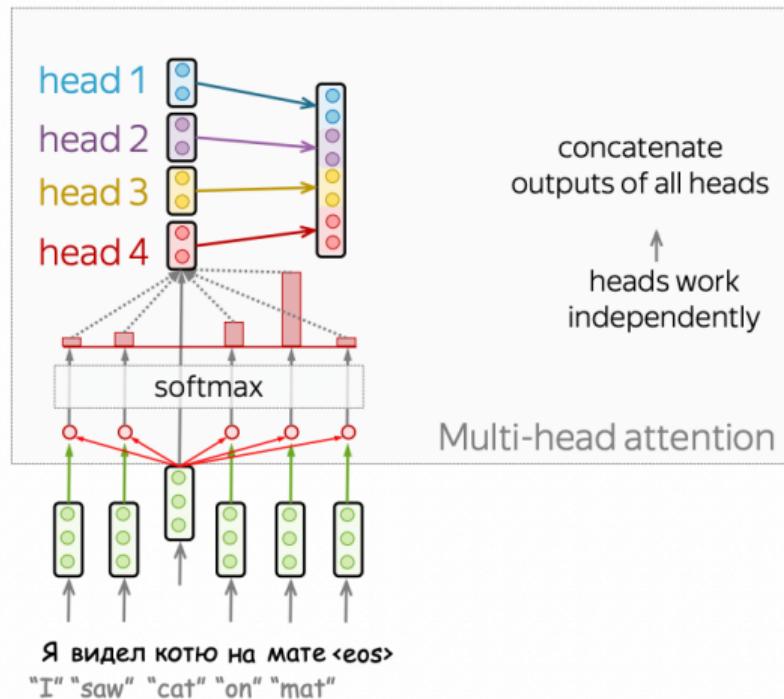


https://github.com/yandexdataschool/nlp_course/tree/2021/week04_seq2seq

Multi-Head Attention

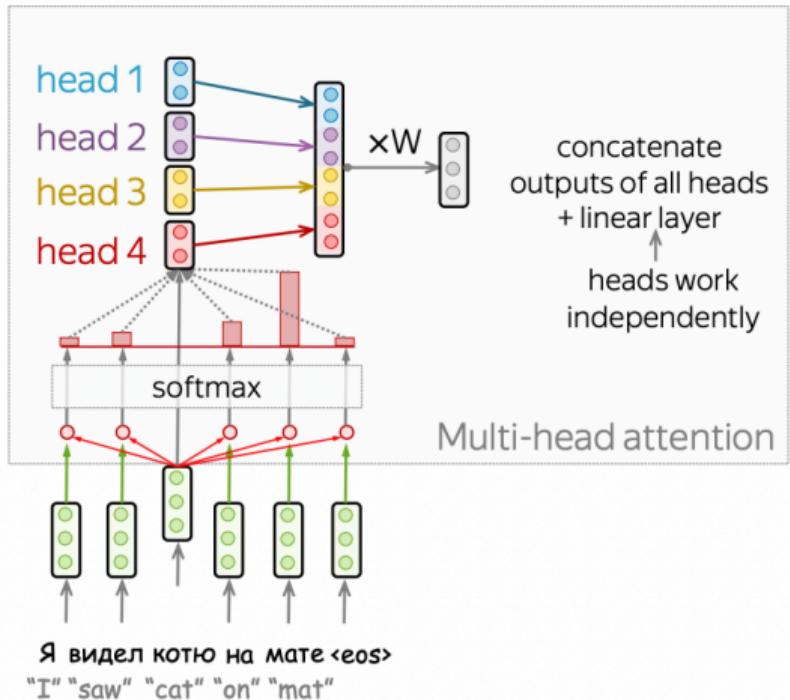


Multi-Head Attention



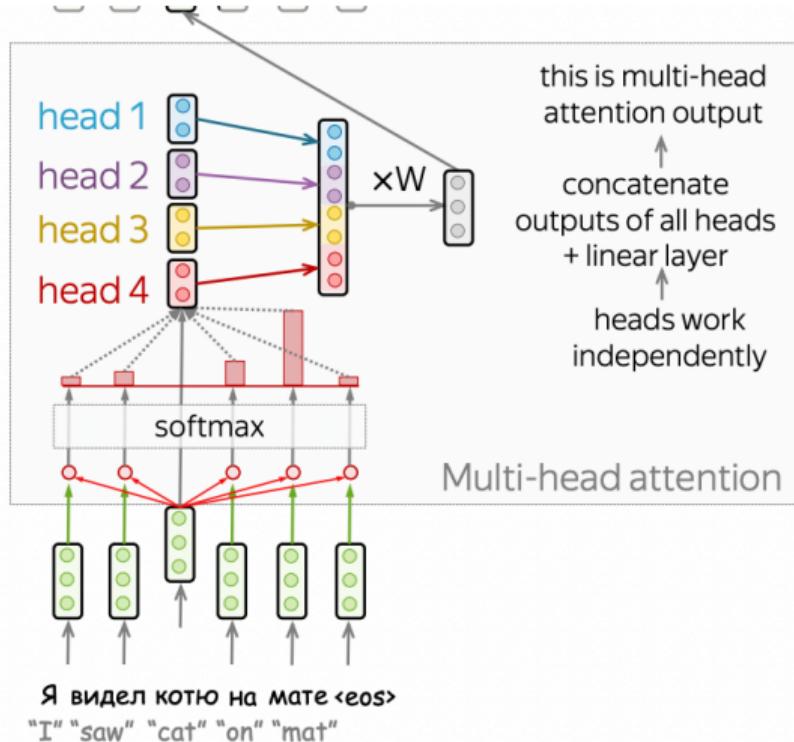
https://github.com/yandexdataschool/nlp_course/tree/2021/week04_seq2seq

Multi-Head Attention



https://github.com/yandexdataschool/nlp_course/tree/2021/week04_seq2seq

Multi-Head Attention

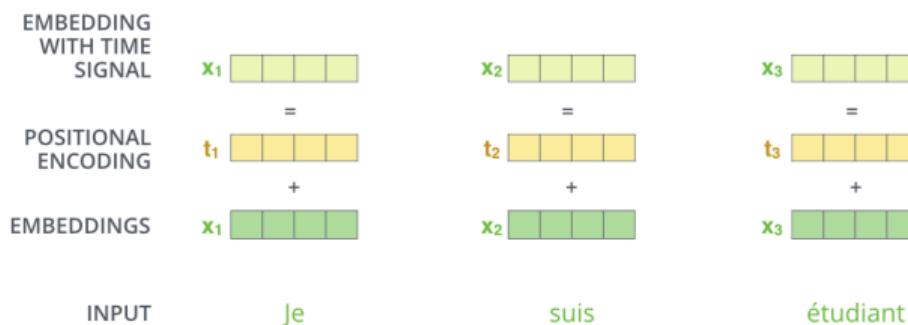
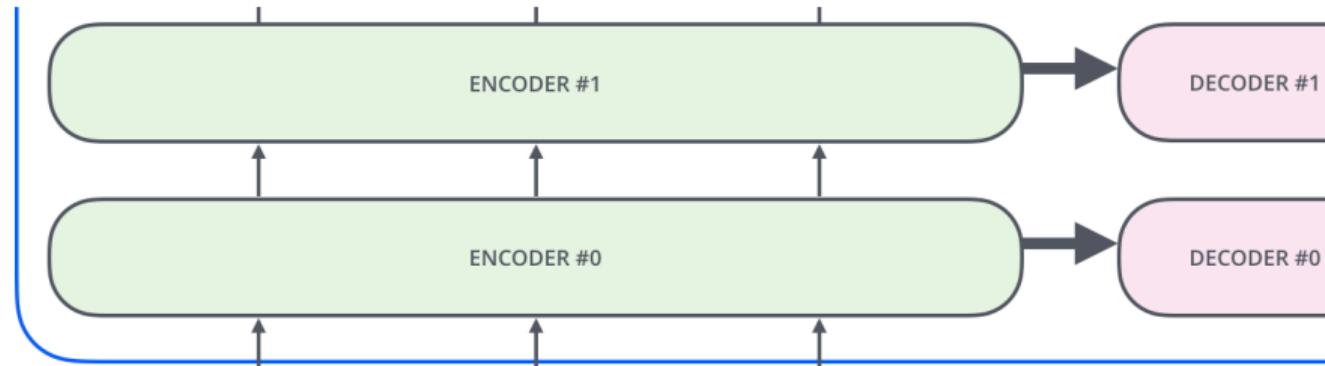


https://github.com/yandexdataschool/nlp_course/tree/2021/week04_seq2seq

Positional Encoding

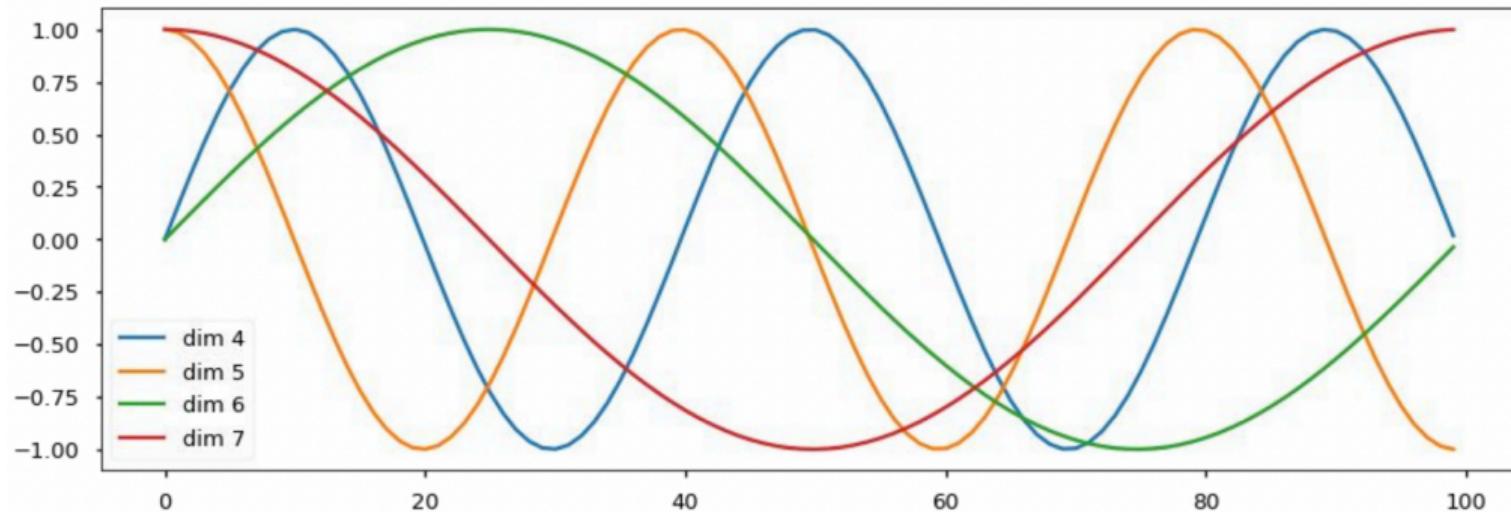
Positional encoding

Для учёта позиции можно приплусовать дополнительный вектор



Positional encoding

Можно закодировать позицию с помощью синусоиды или какого-нибудь ОНЕ-вектора



Positional encoding

Можно обучить под позицию отдельный эмбеддинг

Fixed encodings:

$$\text{PE}_{pos,2i} = \sin(pos/10000^{2i/d_{model}}),$$

$$\text{PE}_{pos,2i+1} = \cos(pos/10000^{2i/d_{model}})$$

pos – position, i - dimension

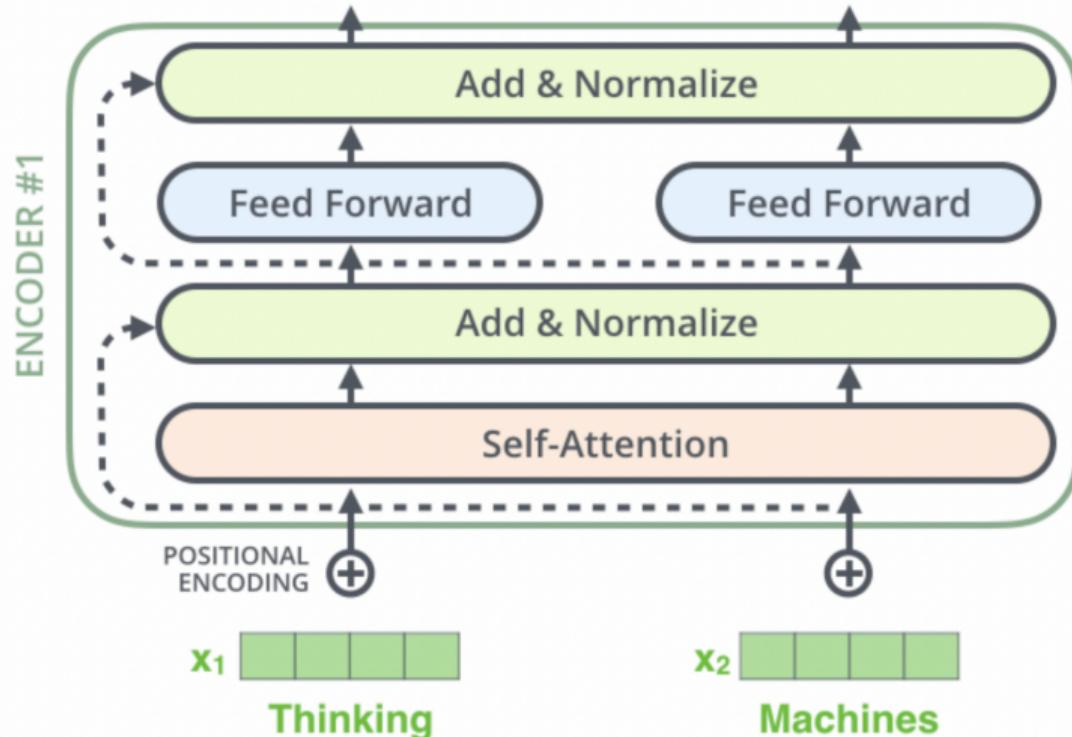
“token x on position k ” →

Input is sum of two embeddings: for token and position

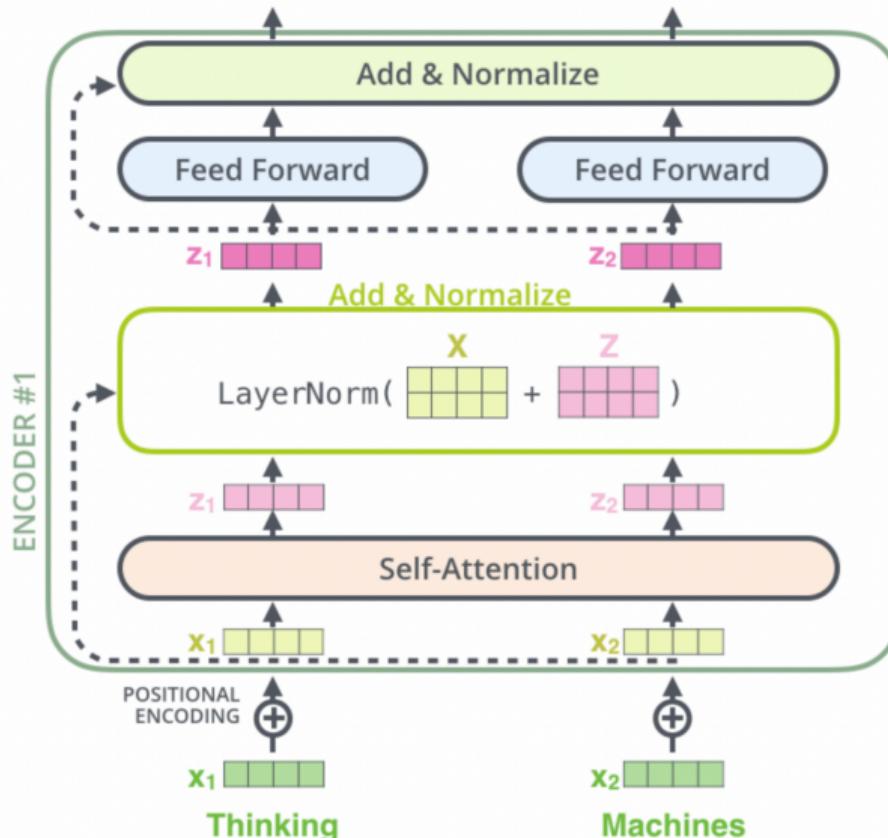


Энкодер и декодер

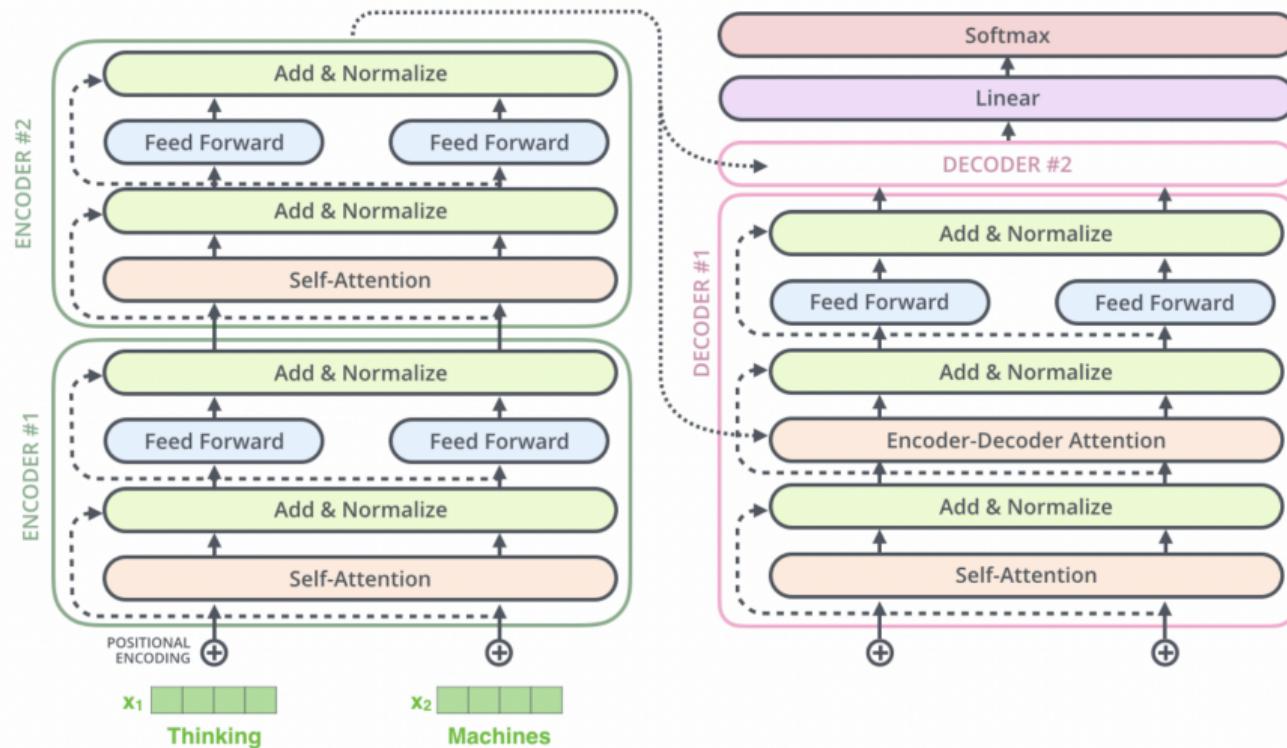
Что происходит в энкодере



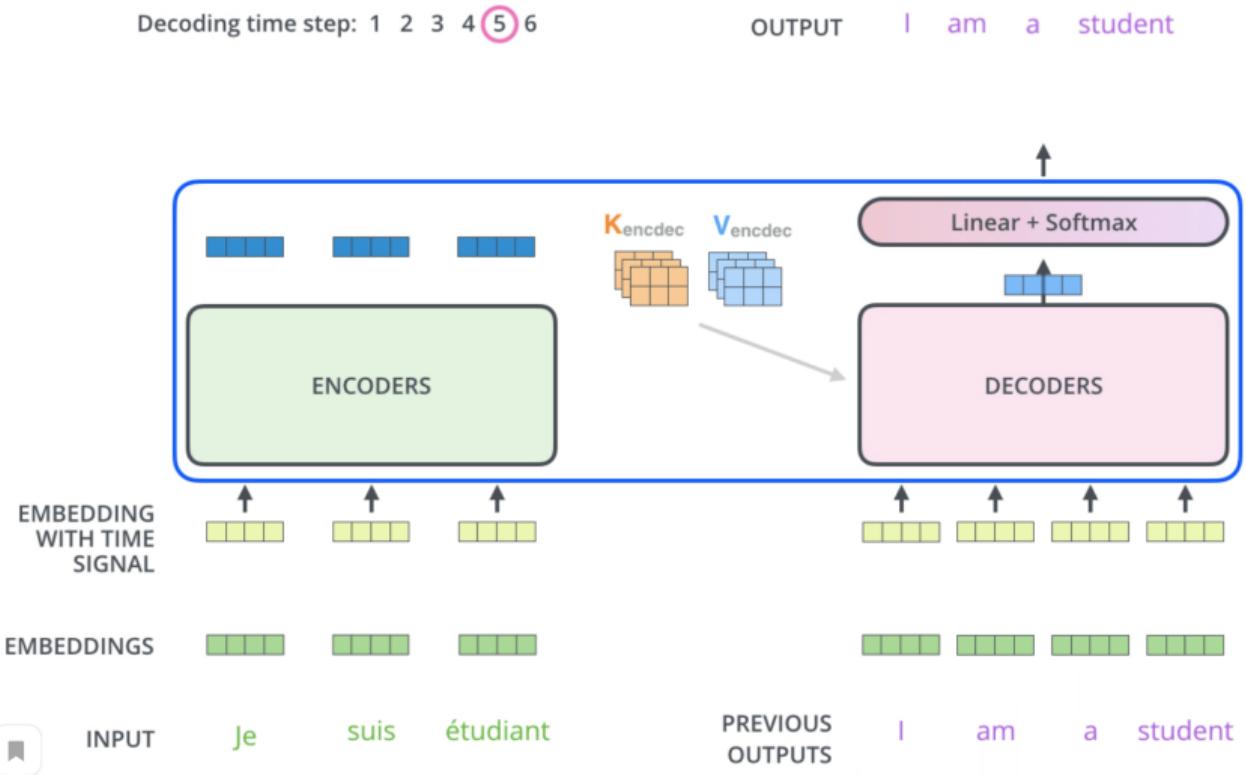
Что происходит в Энкодере



Две модели рядом



Что происходит в декодере?

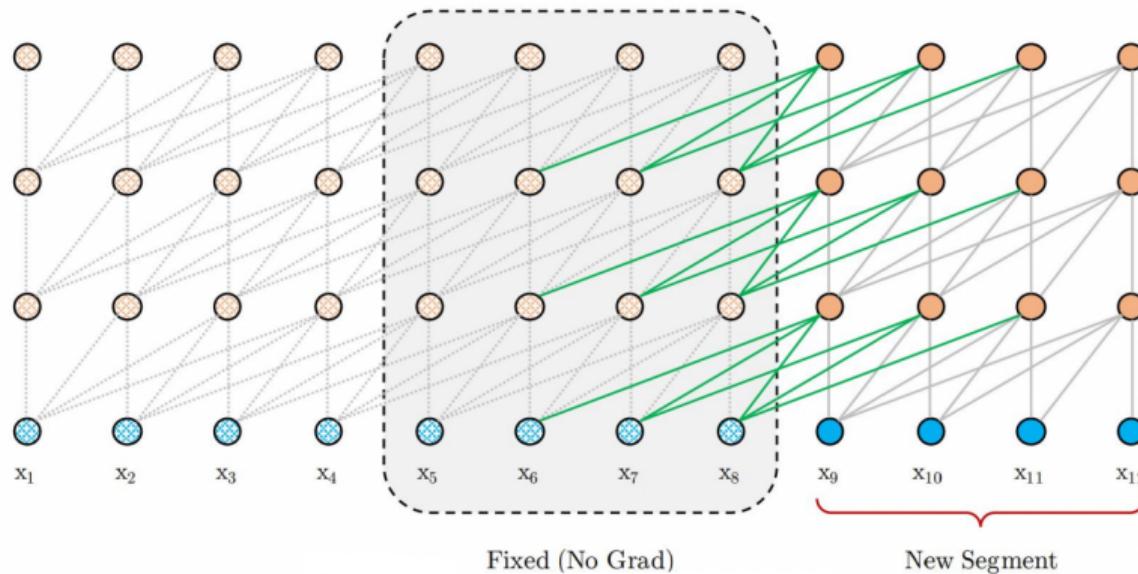


Byte Pair Encoding

Другие виды трансформеров

Transformer-XL (2019)

Добавляется возможность посмотреть на репрезентации предыдущего сегмента \Rightarrow можно выучивать более длинные зависимости.



<https://arxiv.org/abs/1901.02860>

- Большие объёмы неразмеченных данных в интернете в разных доменах (книги, новости, википедия, иные тексты из интернет-страниц)
- Размеченных данных мало. Качественная разметка дорогая и долгая
- Много вычислительных ресурсов, GPU, TPU, фреймворки распределённых вычислений
- Можем ли мы как-то заиспользовать имеющиеся ресурсы?

- Да, можем! Использовать будем semi-supervised learning
- Обучаем большой трансформер на какой-нибудь unsupervised задаче на очень больших данных (очень долго, порядка нескольких недель на 64 гпу);
- Дообучаем трансформер на малом корпусе размеченных данных (очень быстро, порядка 1 часа на одной ГПУ).

BERT (2017)

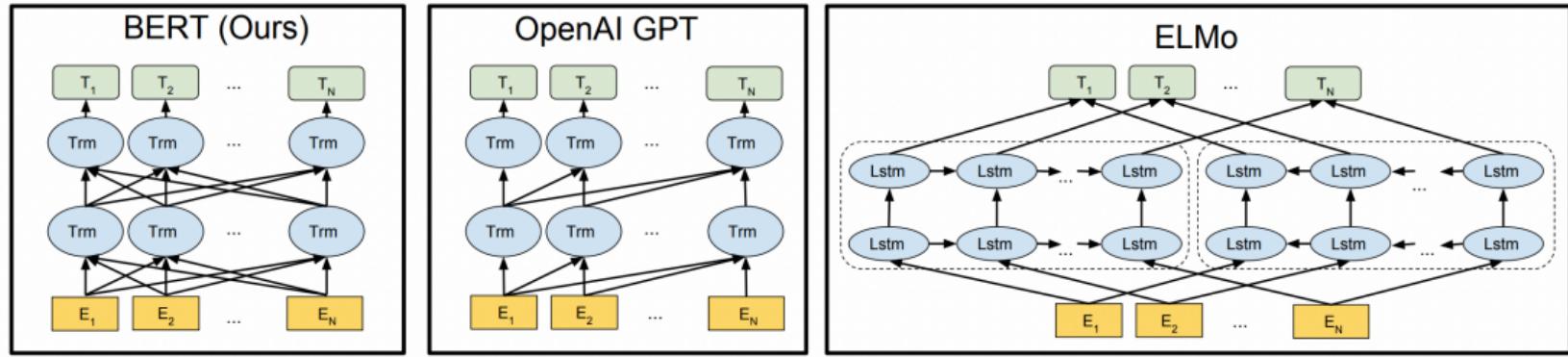


Figure 3: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTMs to generate features for downstream tasks. Among the three, only BERT representations are jointly conditioned on both left and right context in all layers. In addition to the architecture differences, BERT and OpenAI GPT are fine-tuning approaches, while ELMo is a feature-based approach.

<https://arxiv.org/pdf/1810.04805.pdf>

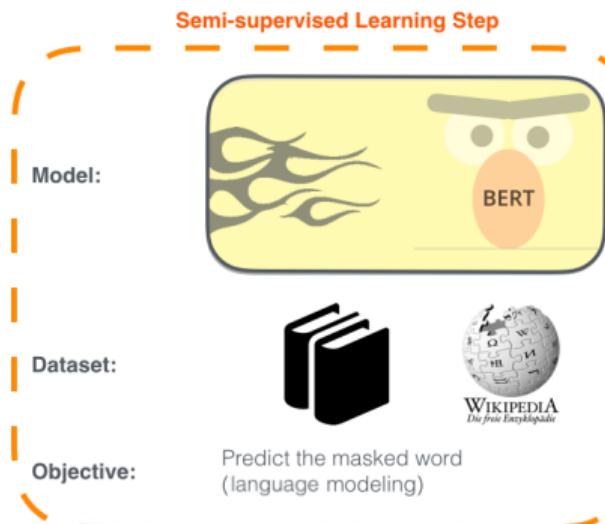
BERT (2018)

- Bidirectional Encoder Representations from Transformers
- Предобученный на искусственно придуманной задаче энкодер
- Его можно файн-тюнить для решения других задач

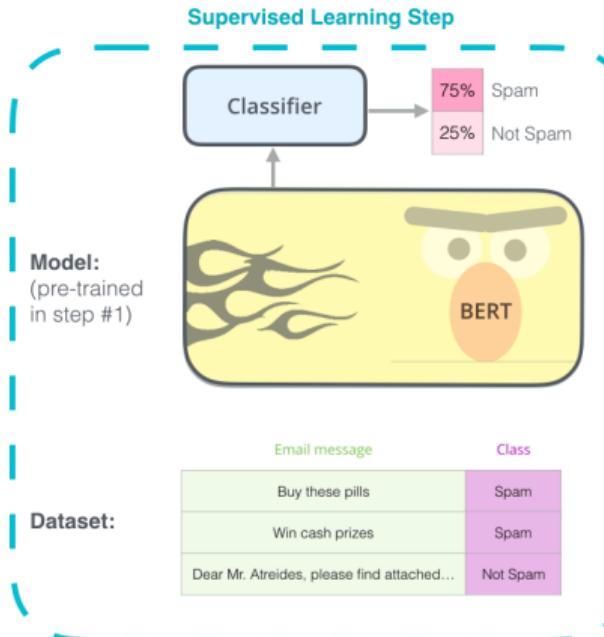
BERT (2018)

- 1 - **Semi-supervised** training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.



- 2 - **Supervised** training on a specific task with a labeled dataset.



BERT (2019)

- **Masked Language Model:** выкидываем часть слов (обычно 15%), модель должна их восстановить
- **Next Sentence Prediction:** подаём на вход два предложения, модель должна понять в каком порядке они идут

BERT (2018)

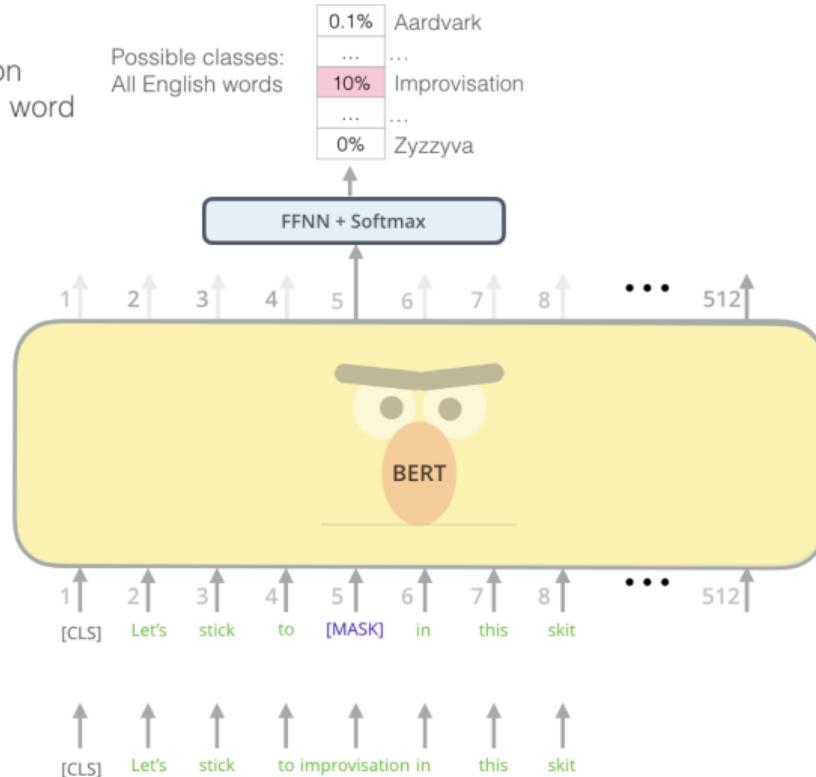
Use the output of the masked word's position to predict the masked word

Possible classes:
All English words



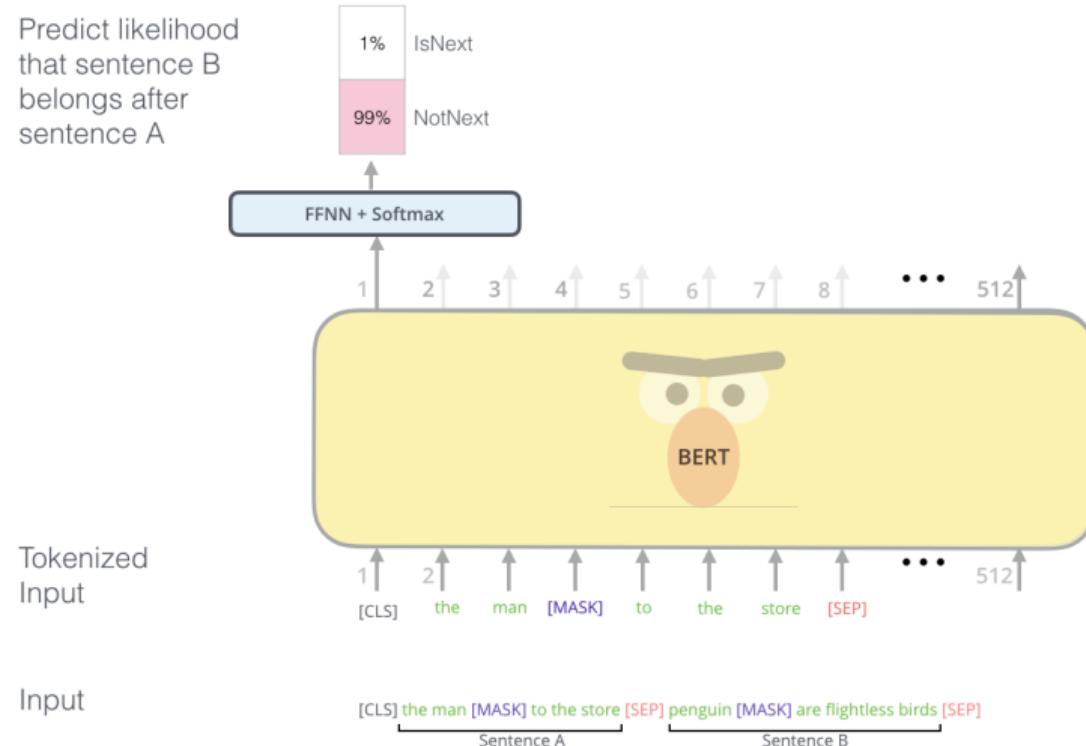
FFNN + Softmax

Randomly mask 15% of tokens



BERT (2018)

Predict likelihood
that sentence B
belongs after
sentence A



Как применять (finetuning)

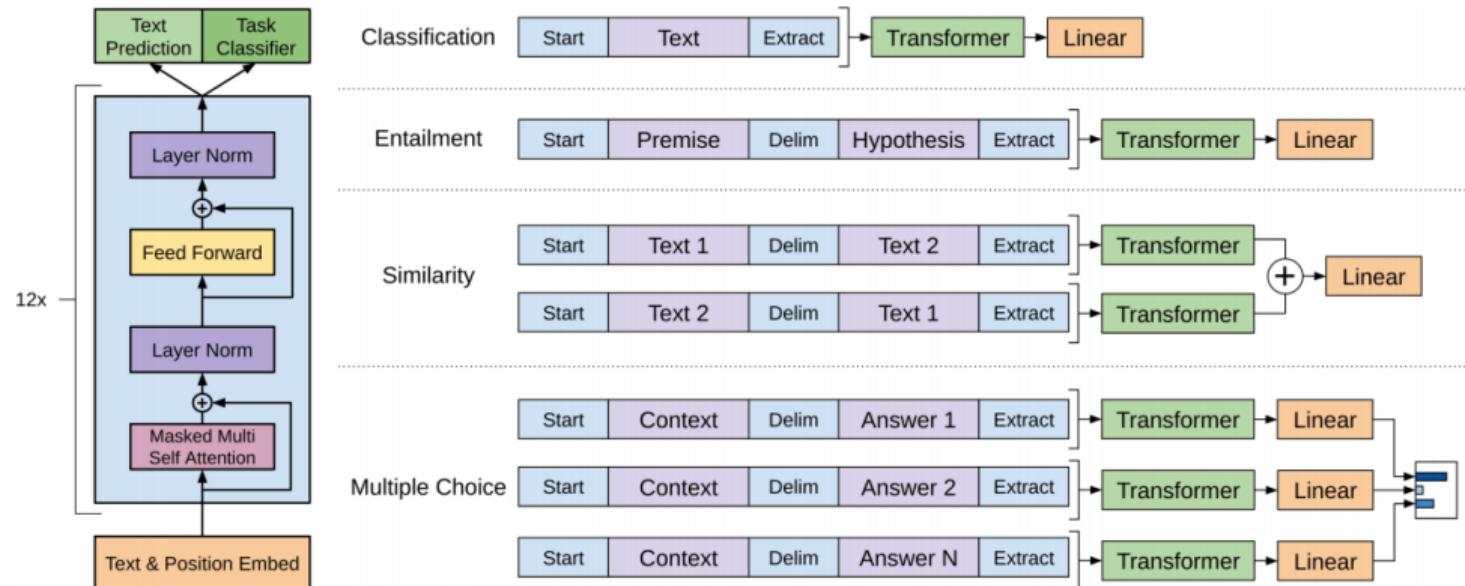


Figure 1: (**left**) Transformer architecture and training objectives used in this work. (**right**) Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

RoBERTa: A Robustly Optimized BERT (2019)

BERT был не дотюнен

- Взять больше данных, тренировать дольше
- Next sentence prediction лишний
- Более длинные предложения
- Большие батчи
- Динамическое маскирование

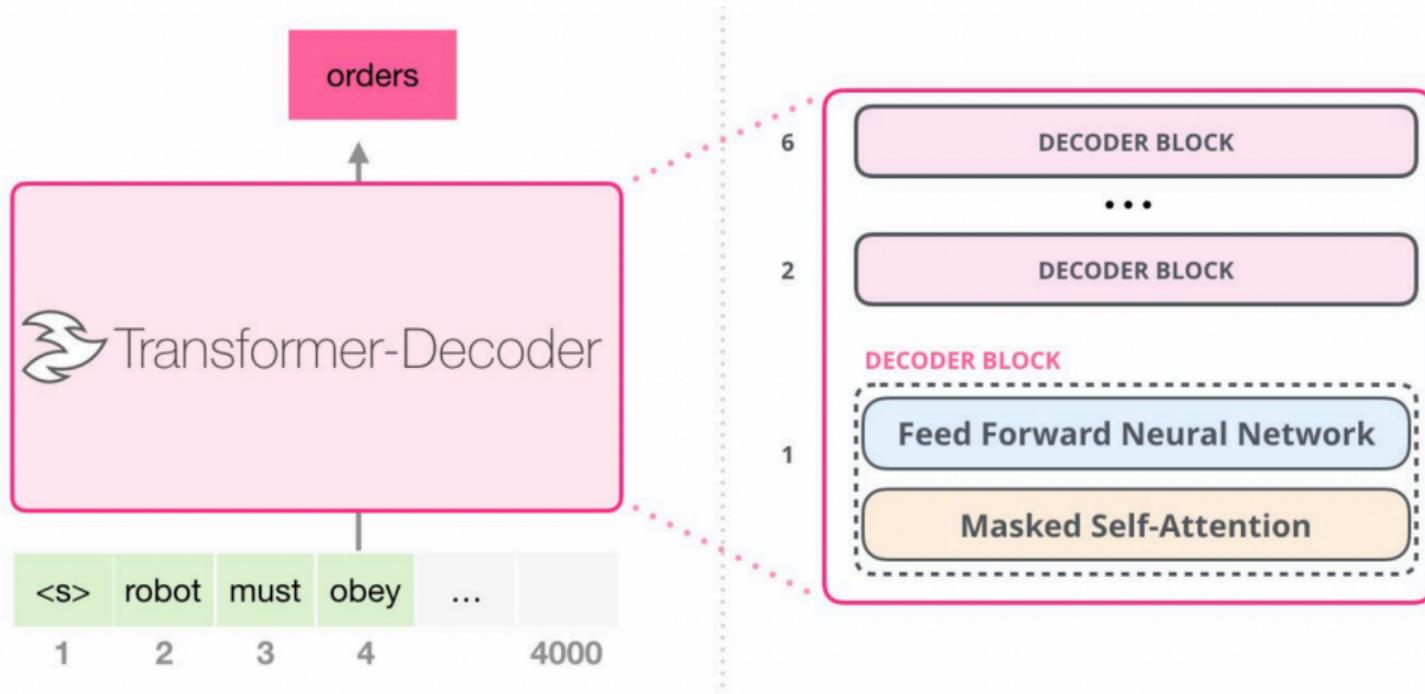
<https://arxiv.org/abs/1907.11692>

GPT-2 (2019)

- Языковая модель на декодере трансформера
- Умеет генерить продолжение текста. Настолько хорошо, что OpenAI отказался её публиковать и устроил мощный PR
- Публикует понемногу, начиная с маленьких моделей
- Разные языковые модели на трансформерах можно попробовать здесь:
<https://transformer.huggingface.co/>

<https://github.com/openai/gpt-2>

GPT-2 (2019)



<https://github.com/openai/gpt-2>

GPT-2 (2019)

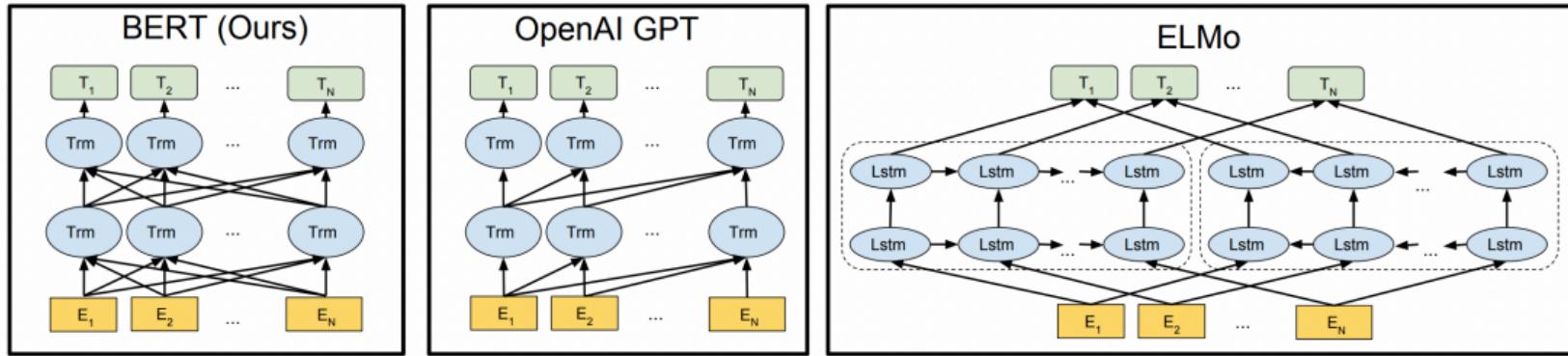
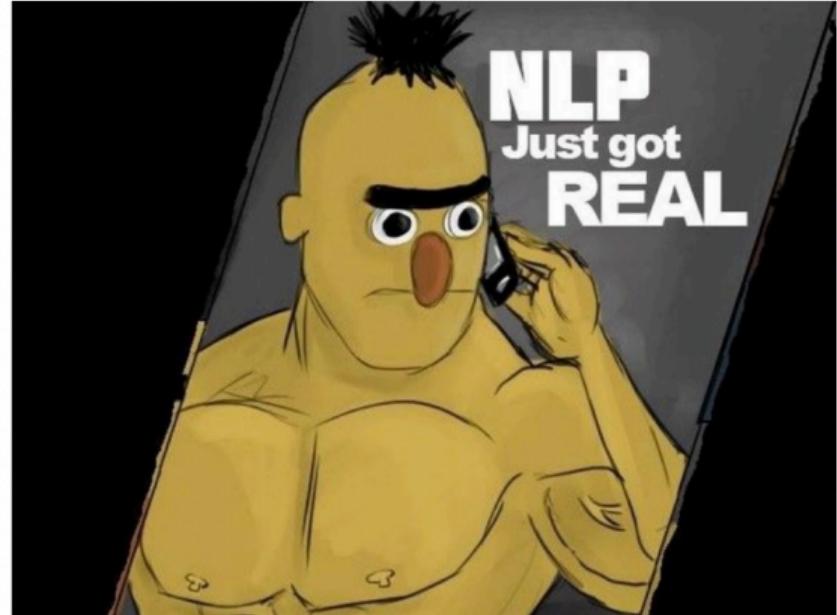


Figure 3: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTMs to generate features for downstream tasks. Among the three, only BERT representations are jointly conditioned on both left and right context in all layers. In addition to the architecture differences, BERT and OpenAI GPT are fine-tuning approaches, while ELMo is a feature-based approach.

<https://arxiv.org/pdf/1810.04805.pdf>

Language Model Zoo

- ELMo
- ULMFiT
- GPT
- BERT (BioBERT, ClinicalBERT, ...)
- ERNIE
- XLNet
- KERMIT
- ERNIE 2.0
- GPT-2
- ALBERT
- ...



<https://arxiv.org/pdf/1810.04805.pdf>