

Глубинное обучение

Обработка естественного языка: Attention, Transformer

Ирина Сапарина

Sequence-to-sequence model

Приложения: перевод

≡ Google Translate ⋮

Text Documents

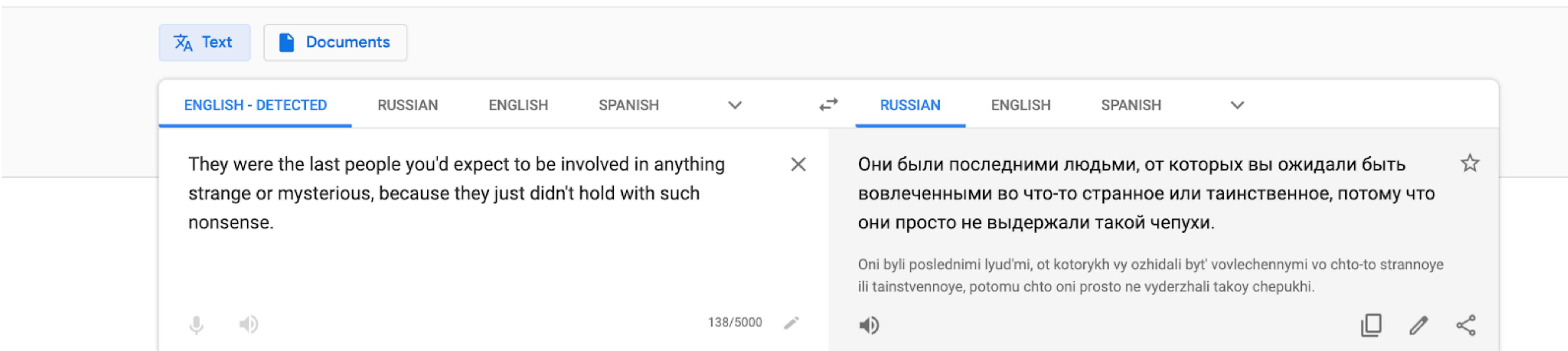
ENGLISH - DETECTED RUSSIAN ENGLISH SPANISH RUSSIAN ENGLISH SPANISH

They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.

Они были последними людьми, от которых вы ожидали быть вовлеченными во что-то странное или таинственное, потому что они просто не выдержали такой чепухи.

Oni byli poslednimi lyud'mi, ot kotorikh vy ozhidali byt' vovlechennymi vo chto-to strannoye ili tainstvennoye, potomu chto oni prosto ne vyderzhali takoy chepukhi.

Send feedback



Приложения: перевод

The screenshot shows the Google Translate interface. At the top, it says "≡ Google Translate" and has a three-dot menu icon. Below that, there are two tabs: "Text" (selected) and "Documents". The main area shows a translation from English to Russian. The English input is: "They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense." The Russian output is: "Они были последними людьми, от которых вы ожидали быть вовлеченными во что-то странное или таинственное, потому что они просто не выдержали такой чепухи." Below the translation, the Russian text is also displayed in its phonetic transcription: "Oni byli poslednimi lyud'mi, ot kotorikh vy ozhidali byt' vovlechennymi vo chto-to strannoye ili tainstvennoye, potomu chto oni prosto ne vyderzhali takoy chepukhi." At the bottom of the translation box, there are icons for microphone, speaker, edit, and share, along with a character count of "138/5000". On the far right, there is a "Send feedback" link.

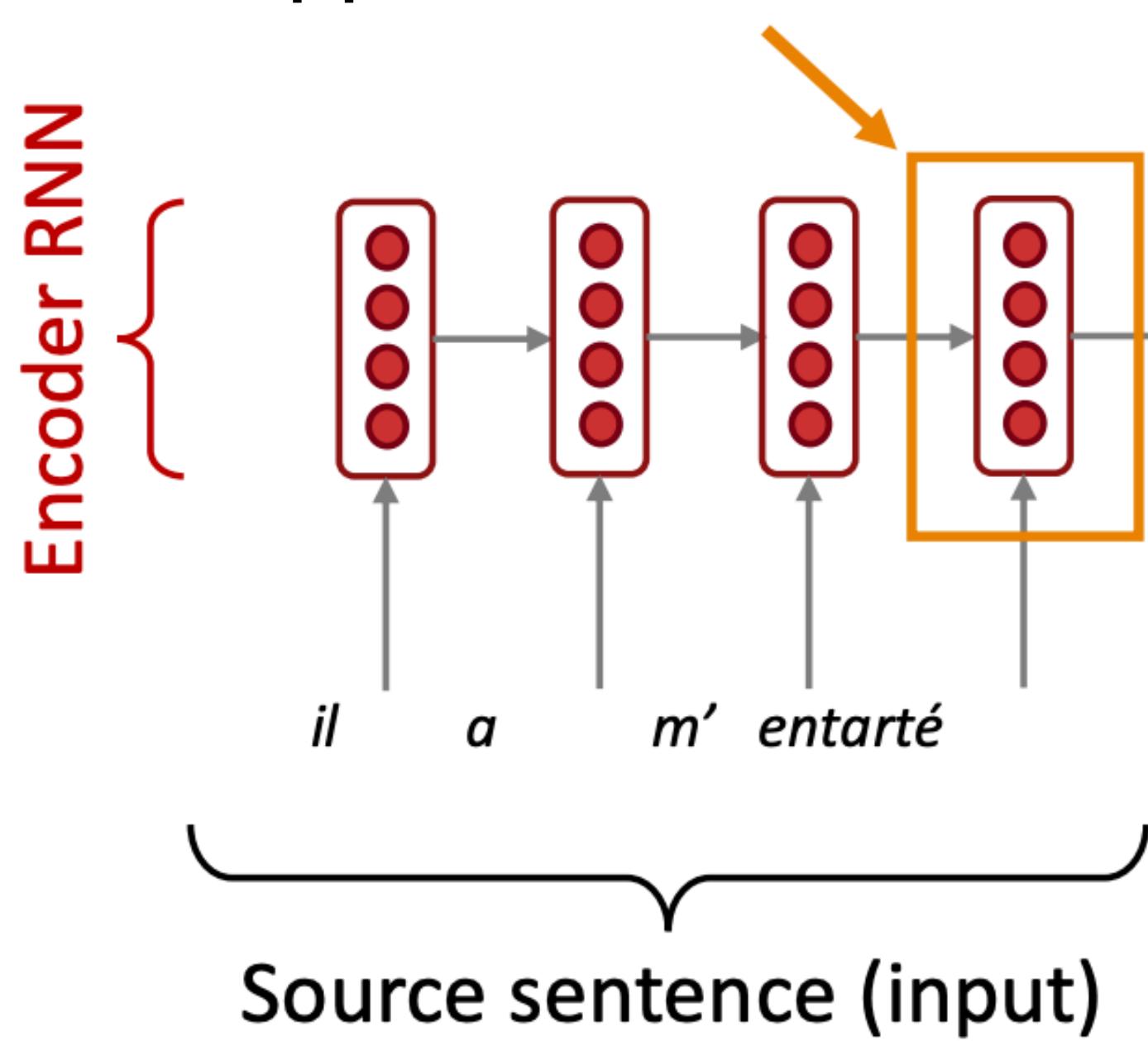
Авторегрессионная модель $p(y|x) = \prod_{i=1}^n p(y_i|y_1, \dots, y_{i-1}, \underline{x})$

x - текст на исходном языке (source)

y - перевод текста на другой язык (target)

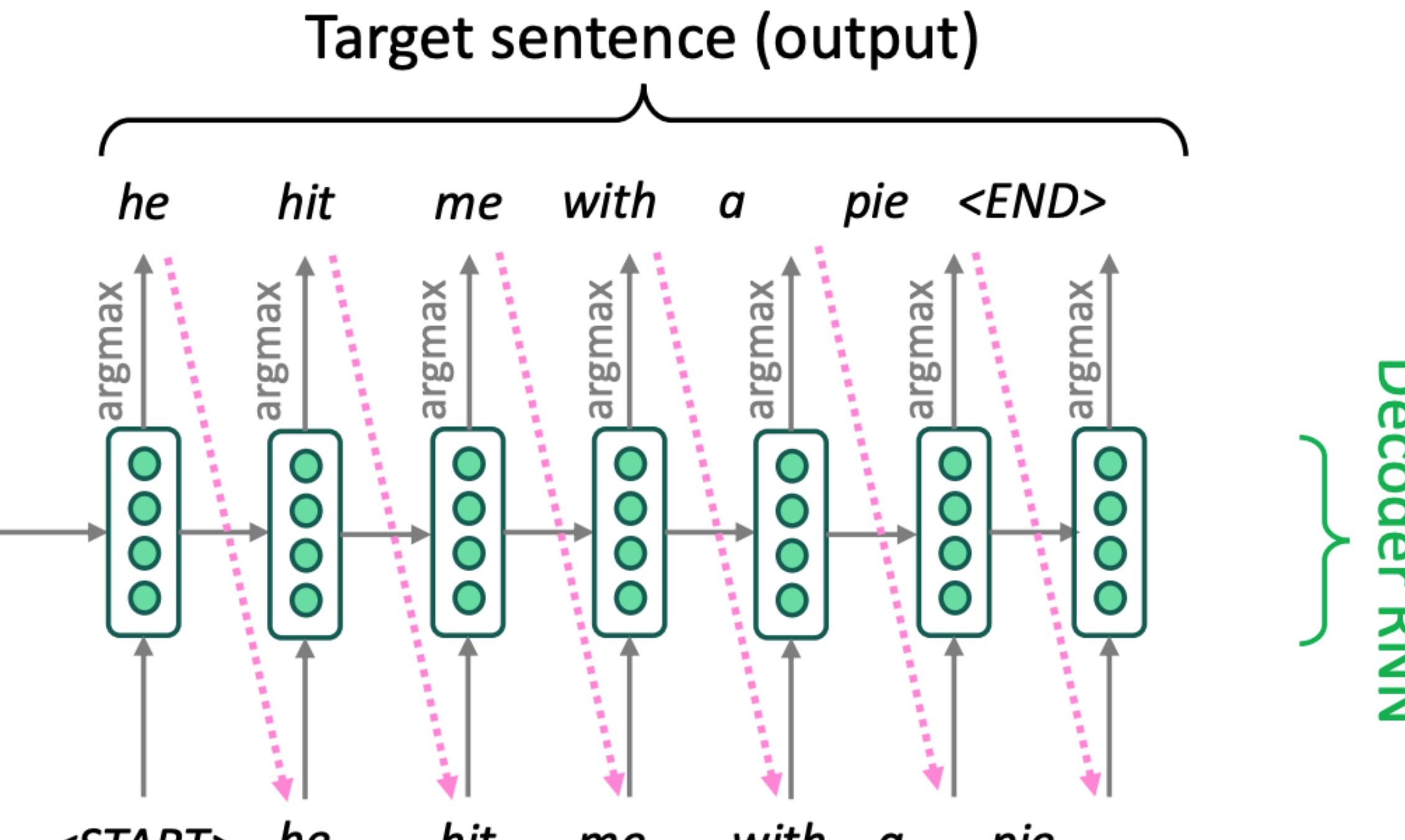
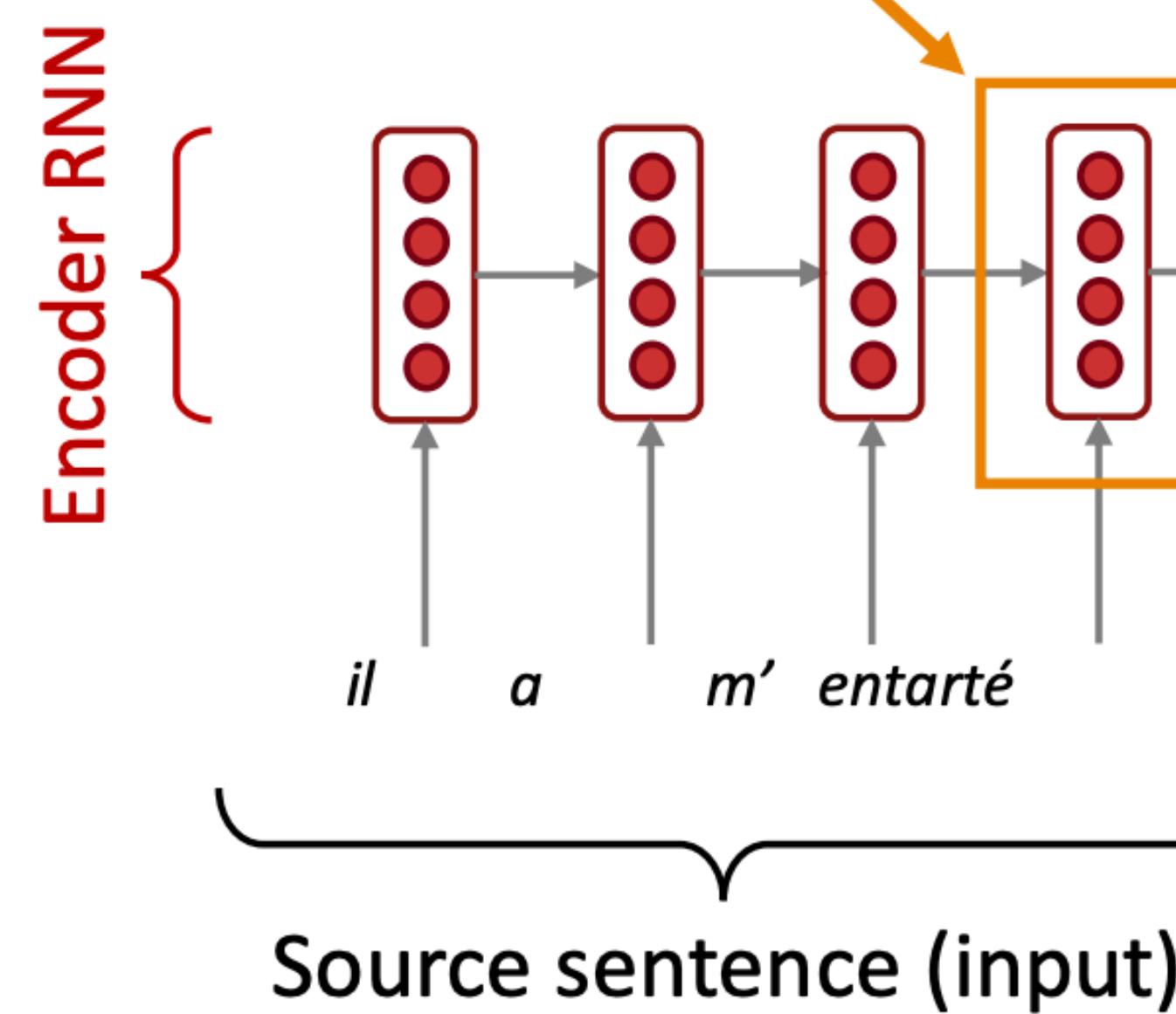
Sequence-to-sequence model

Последний hidden state -
представление всего
исходного текста



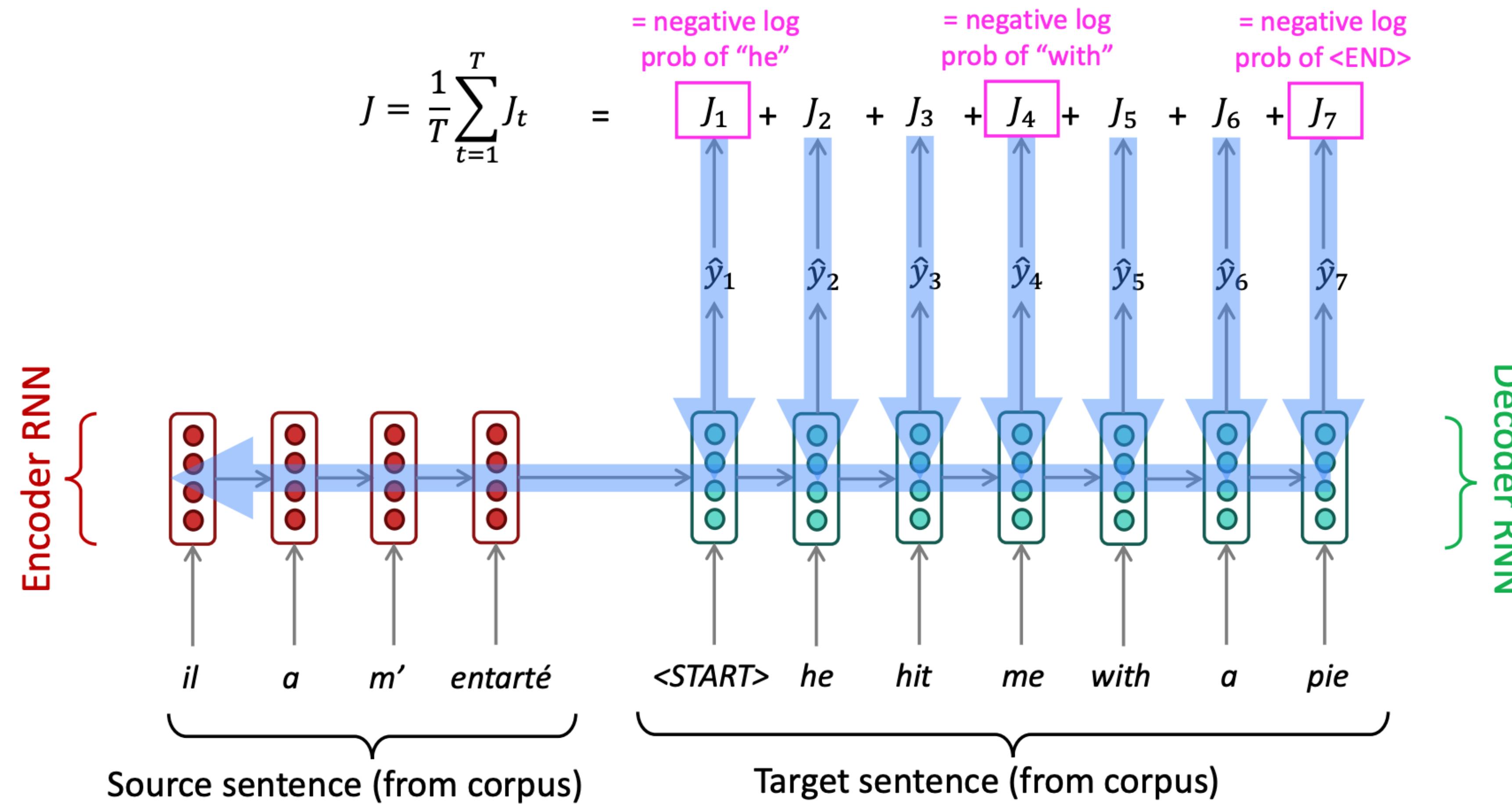
Sequence-to-sequence model

Используем как h^0
для языковой модели
на другом языке



Sequence-to-sequence model

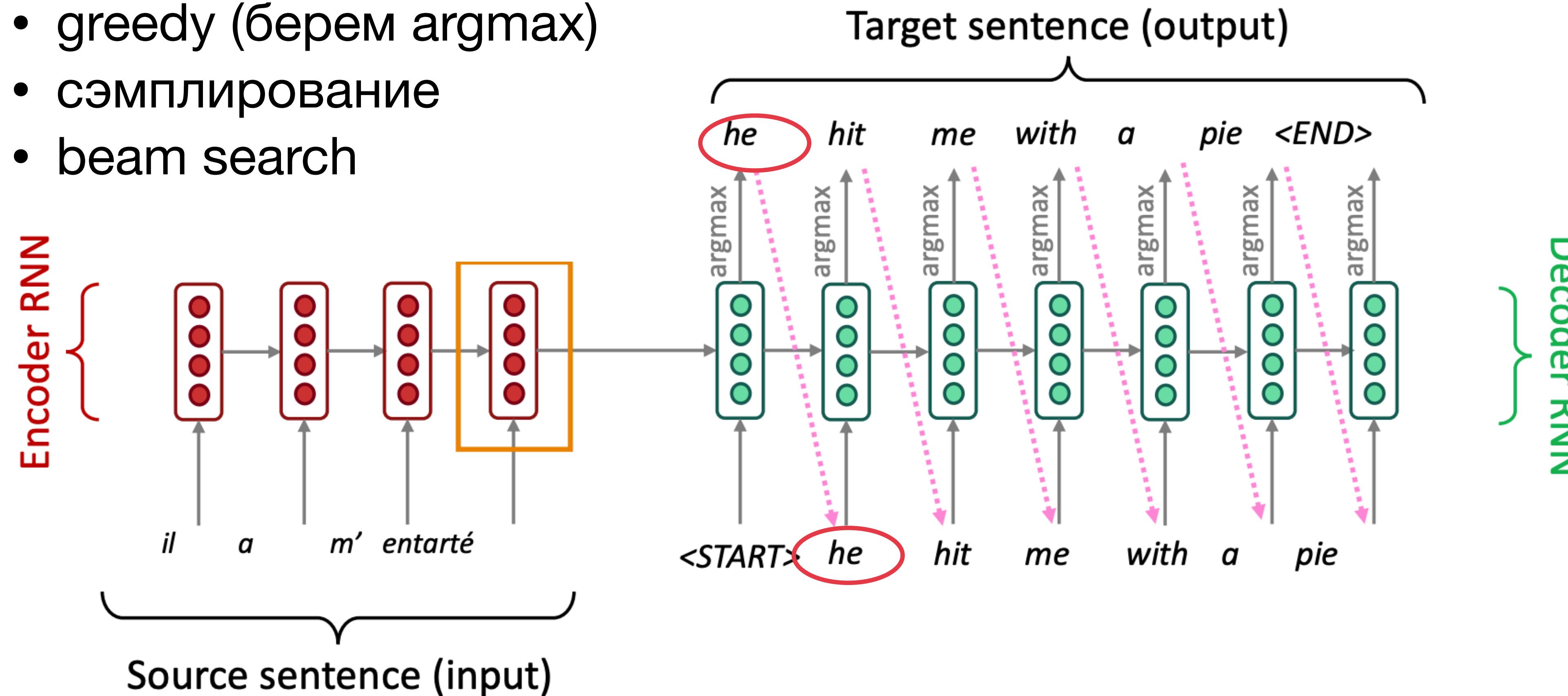
Обучение: нужен параллельный корпус



Sequence-to-sequence model

Генерация:

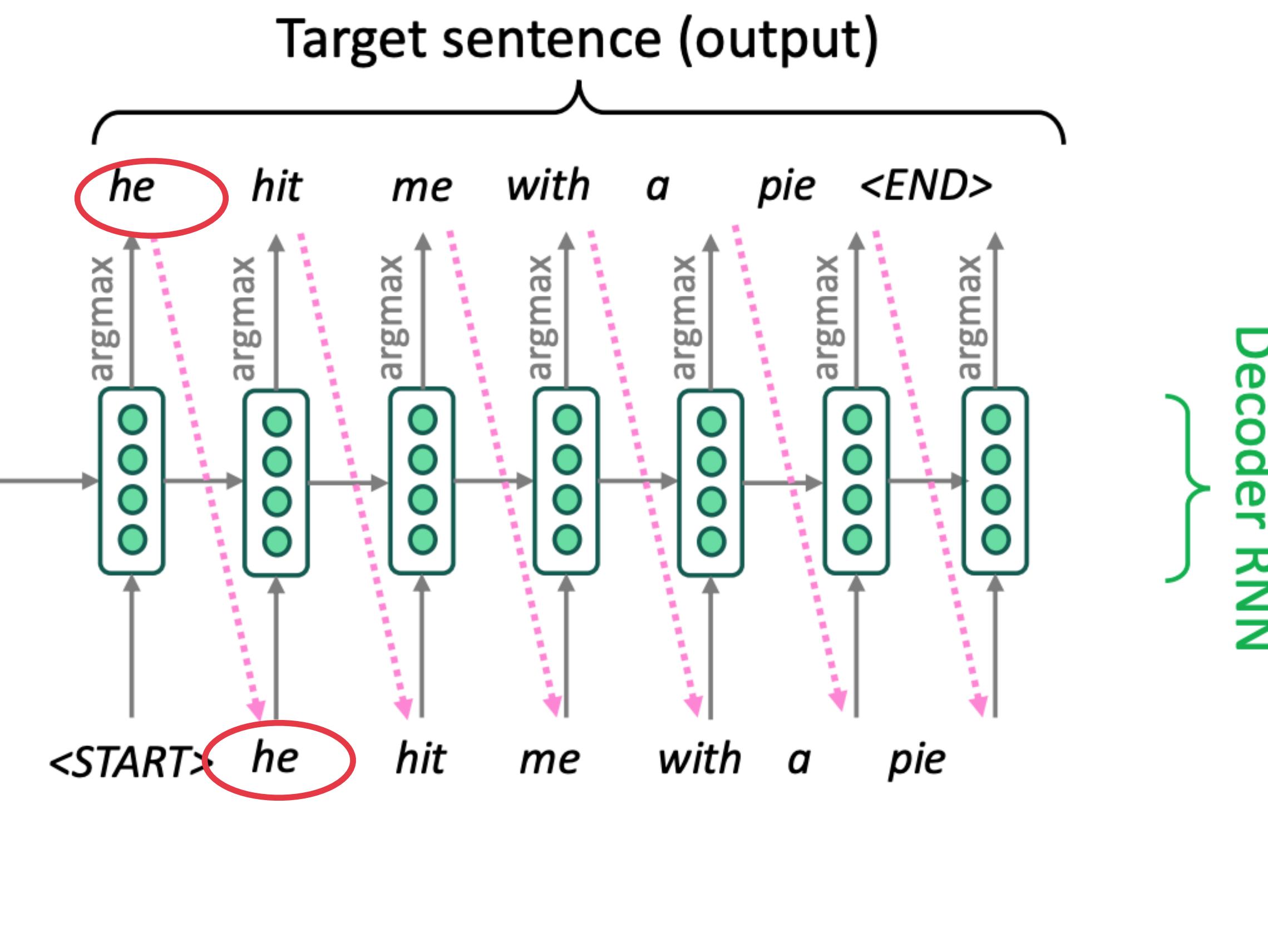
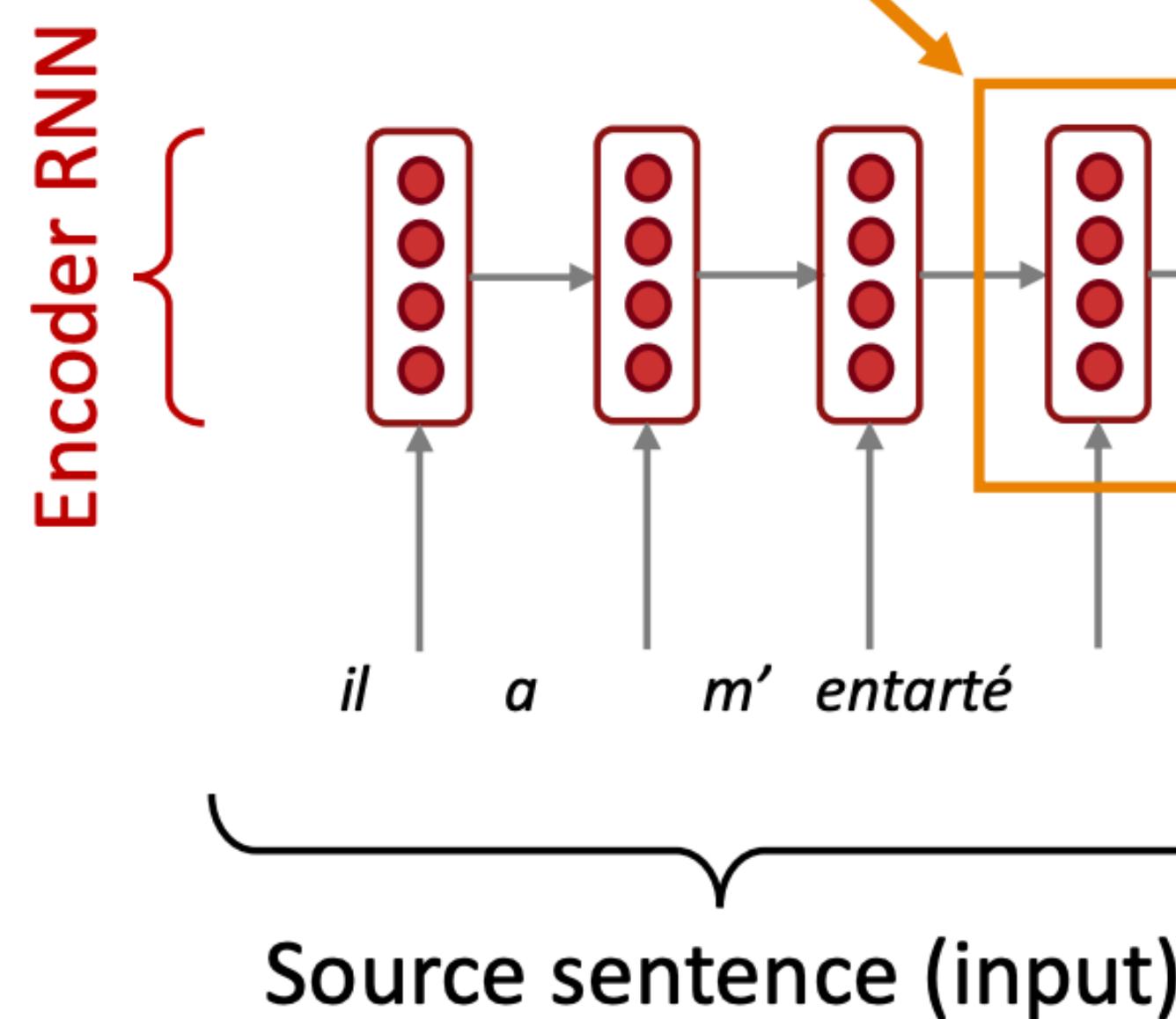
- greedy (берем argmax)
- сэмплирование
- beam search



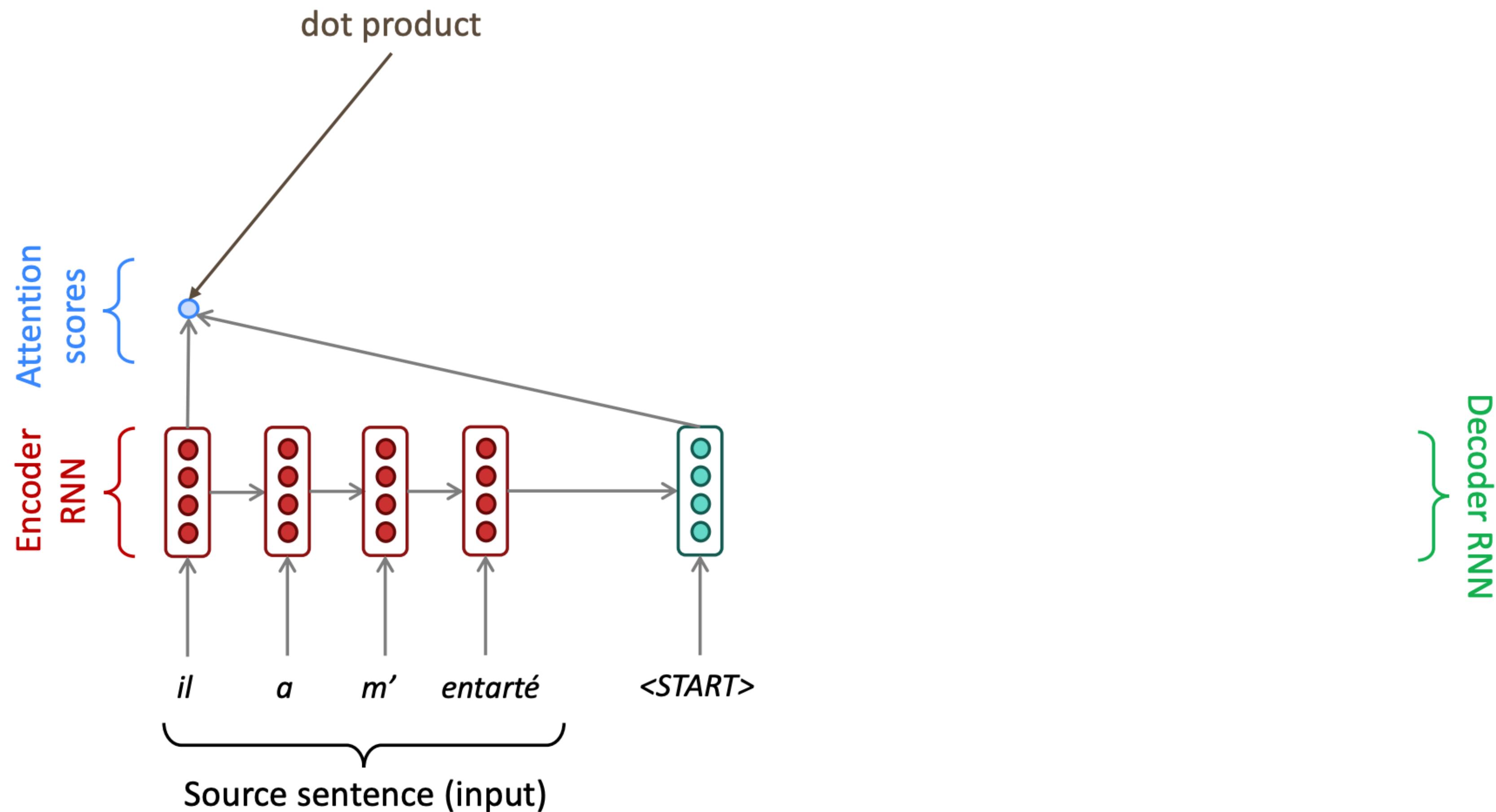
Sequence-to-sequence model

Informational bottleneck:

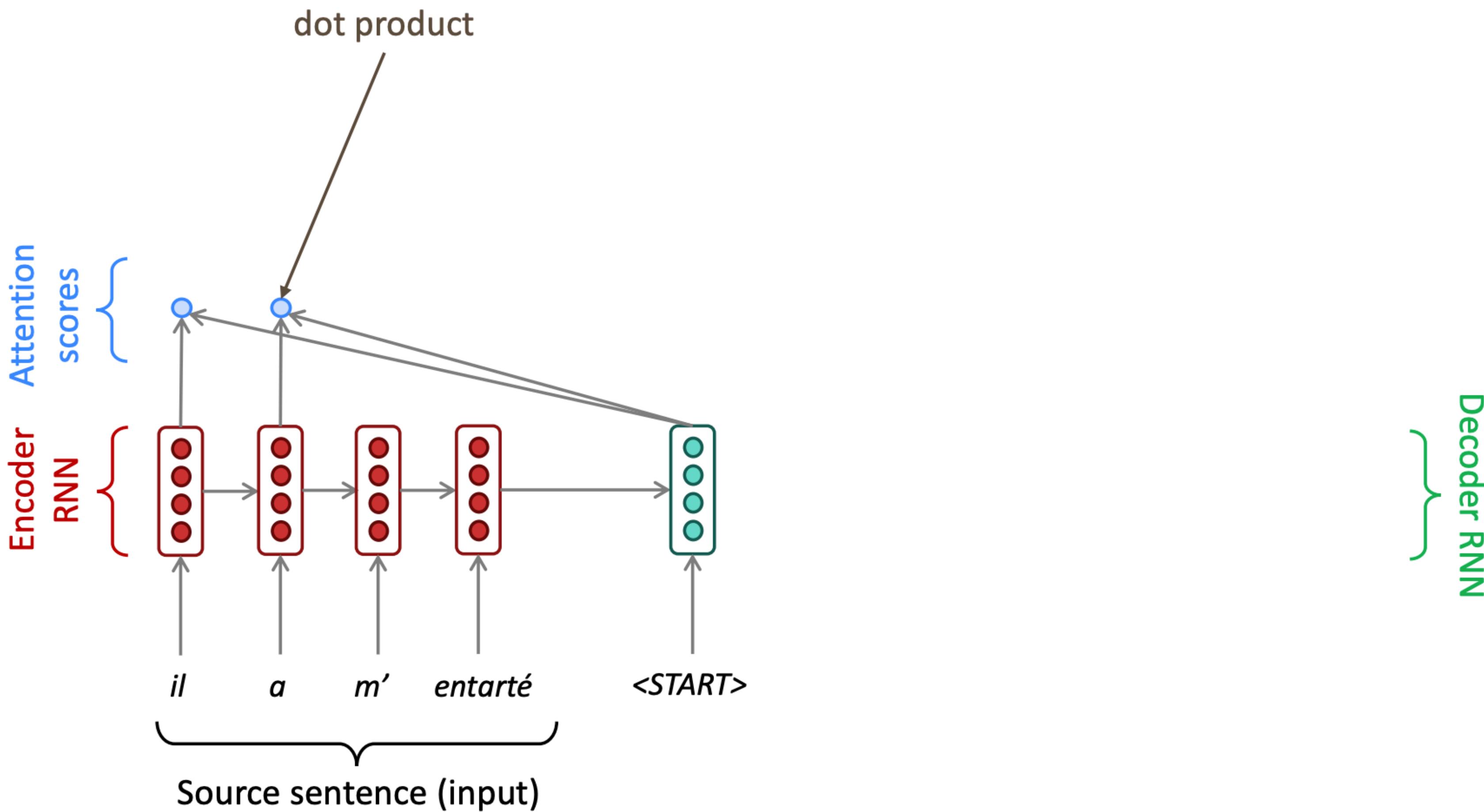
В одном векторе должна быть вся информация



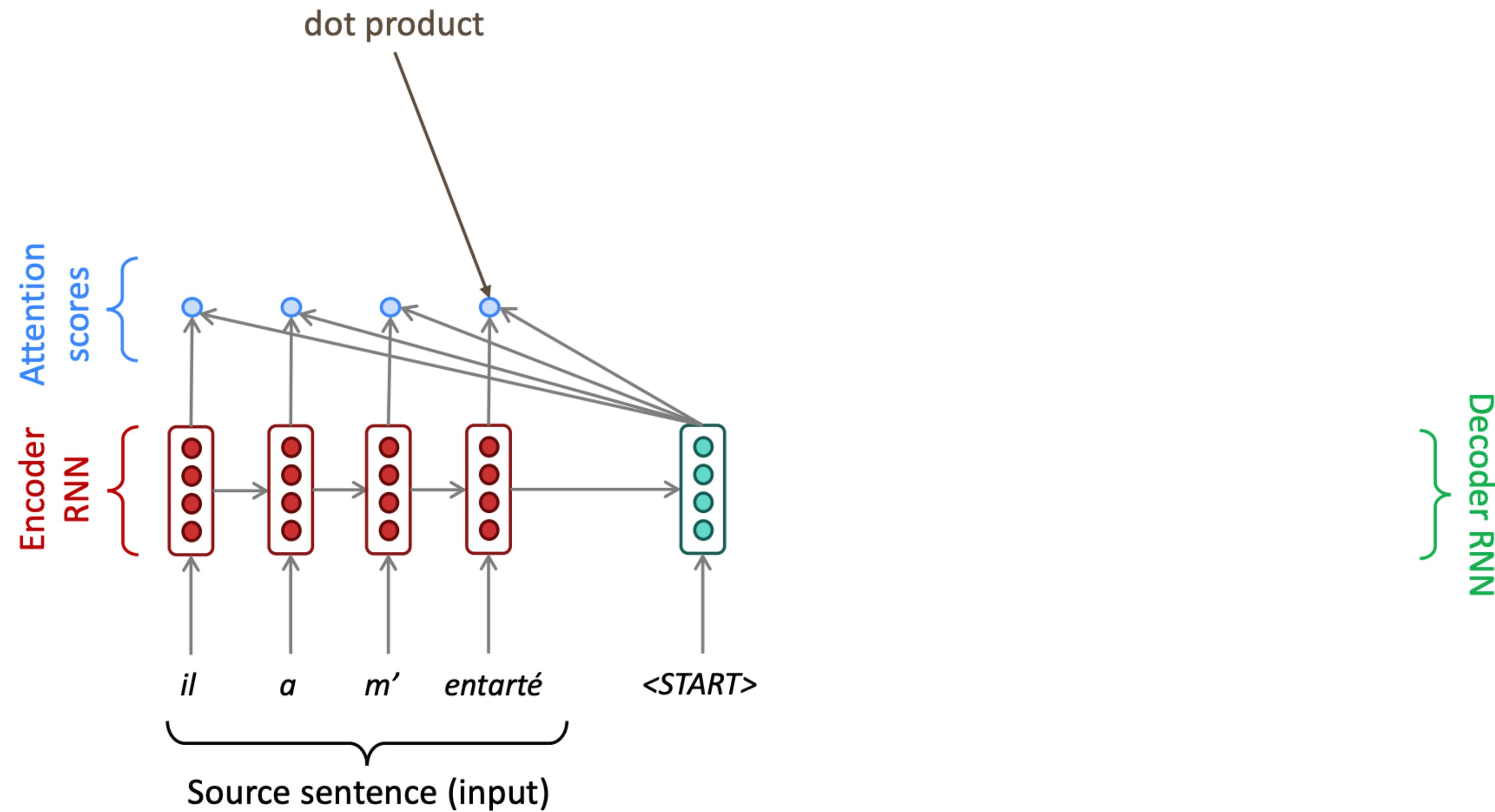
Attention



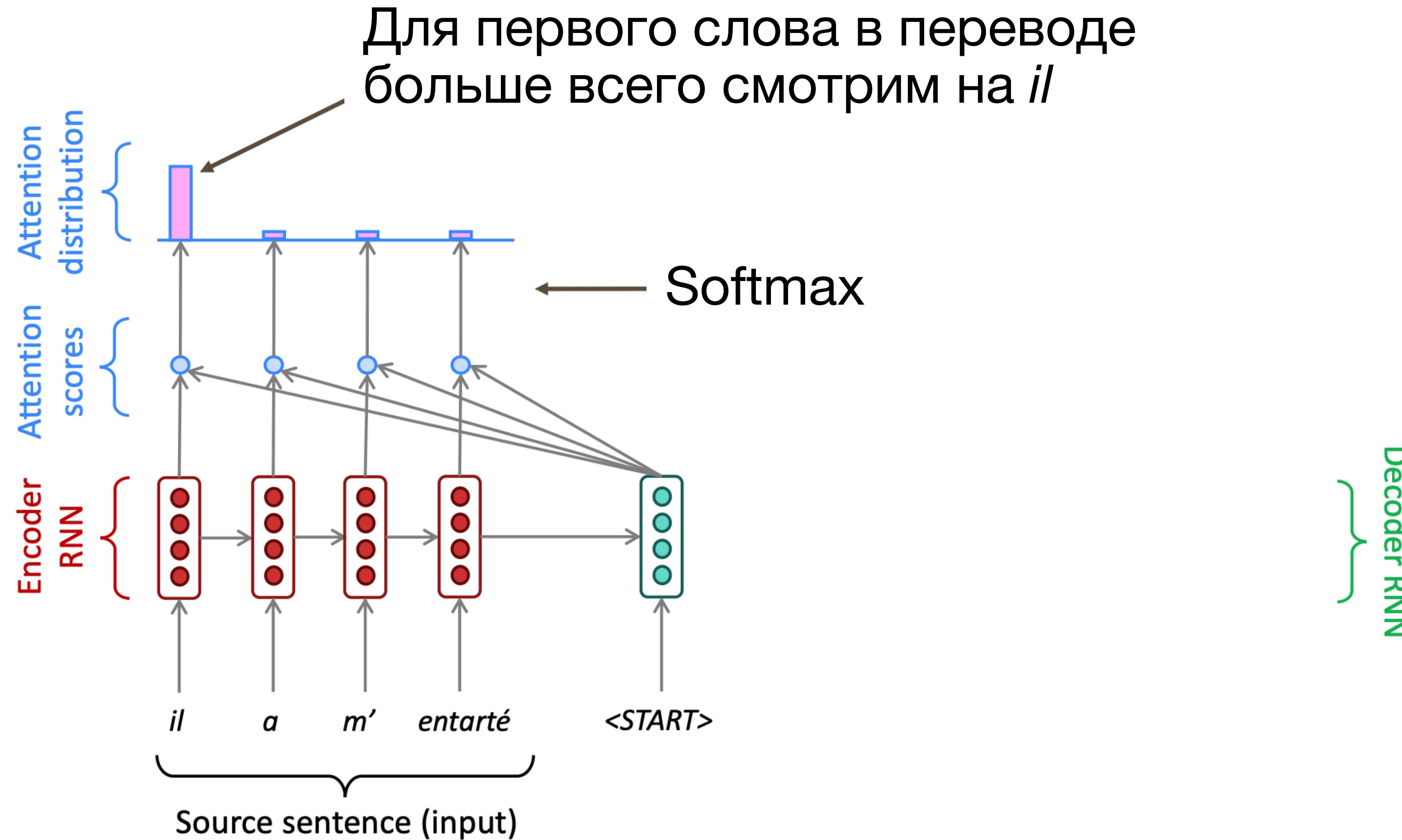
Attention



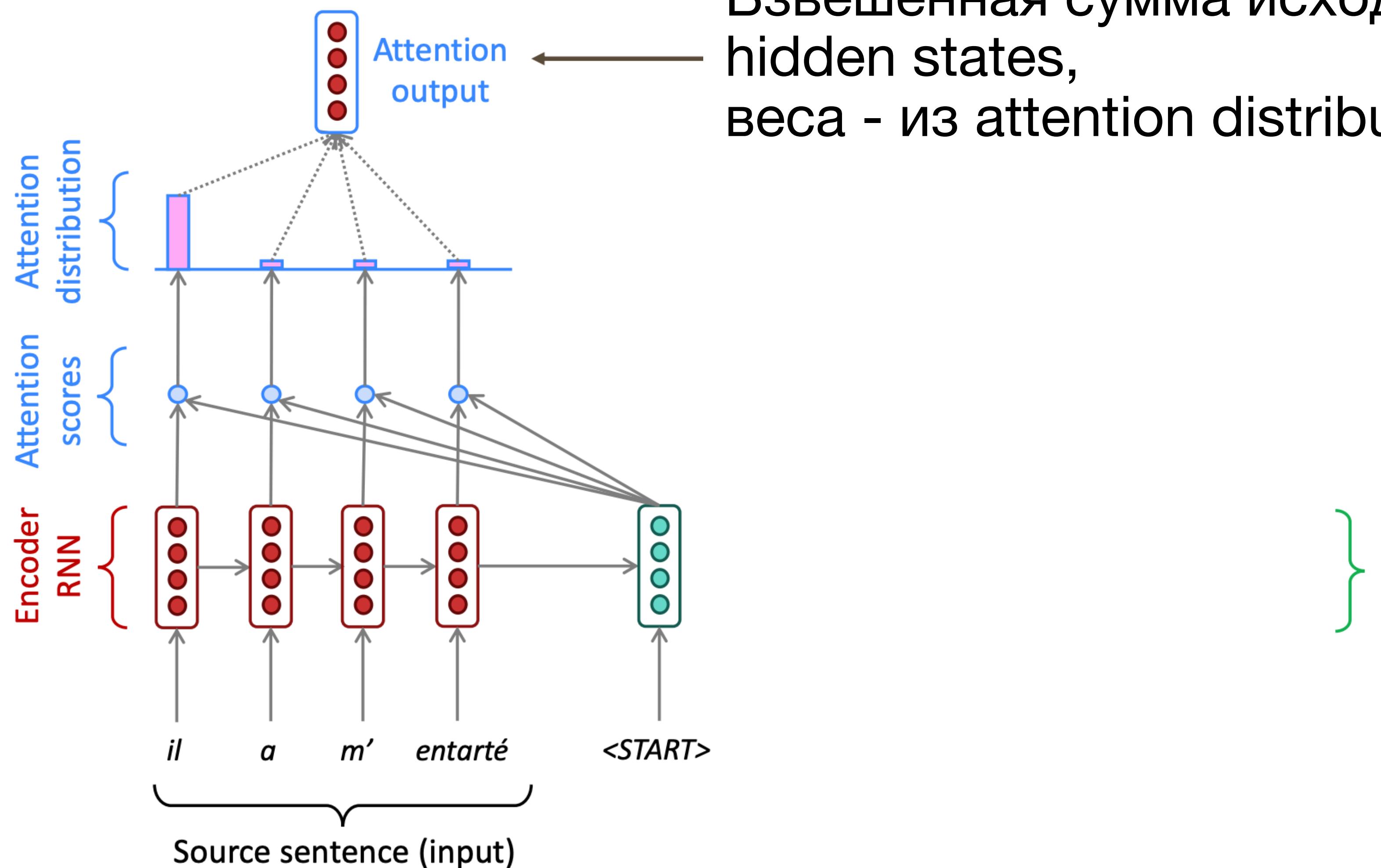
Attention



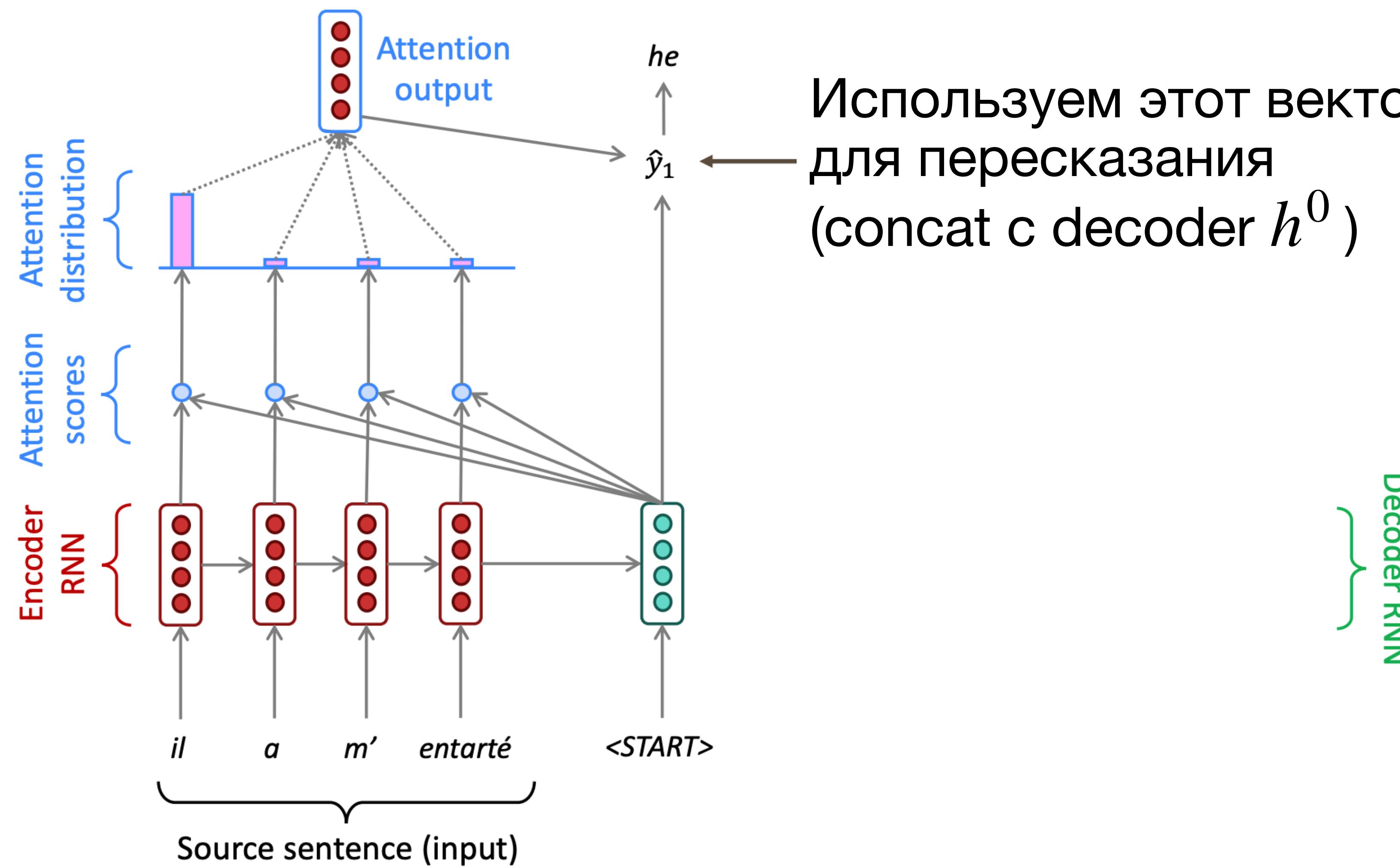
Attention



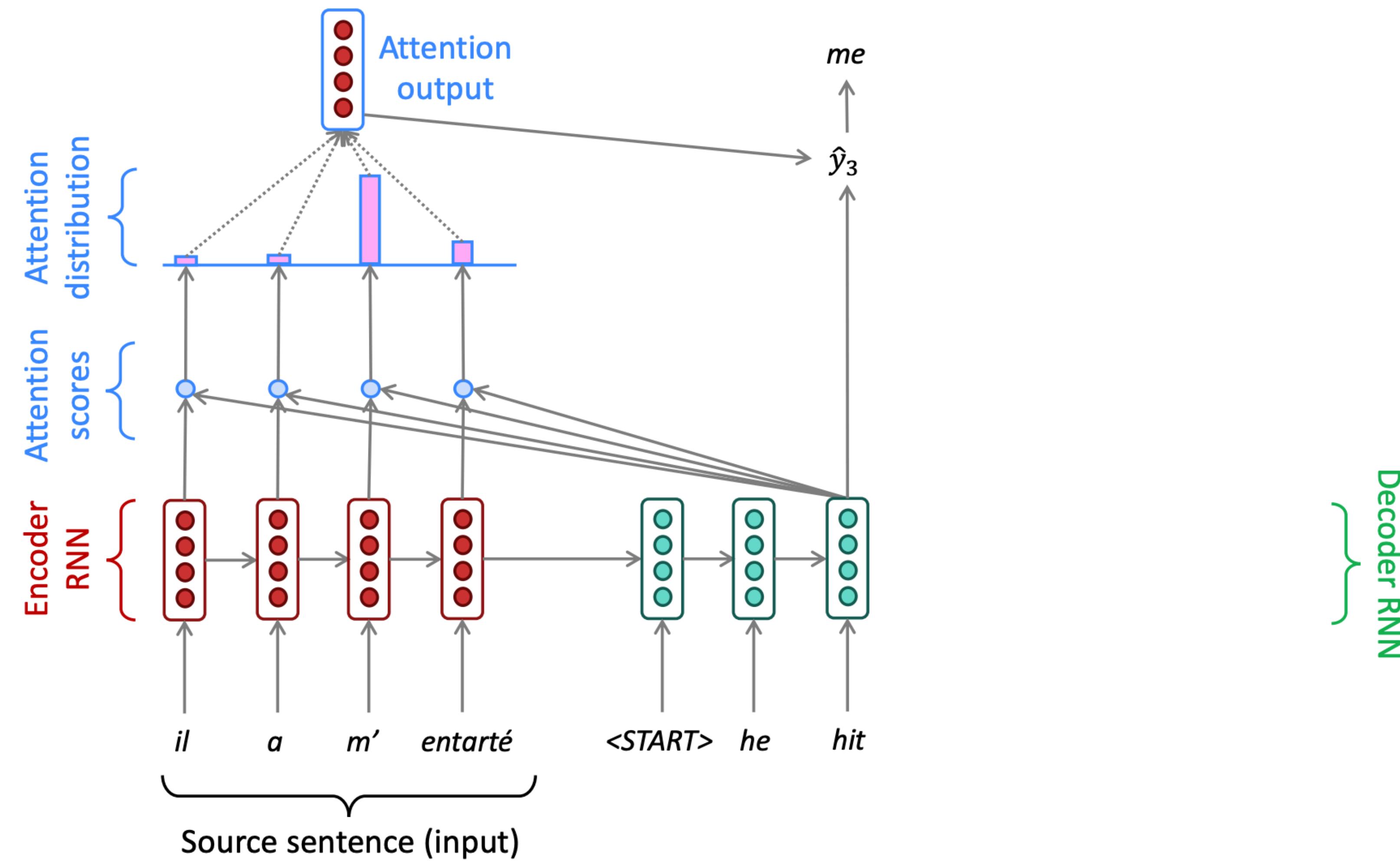
Attention



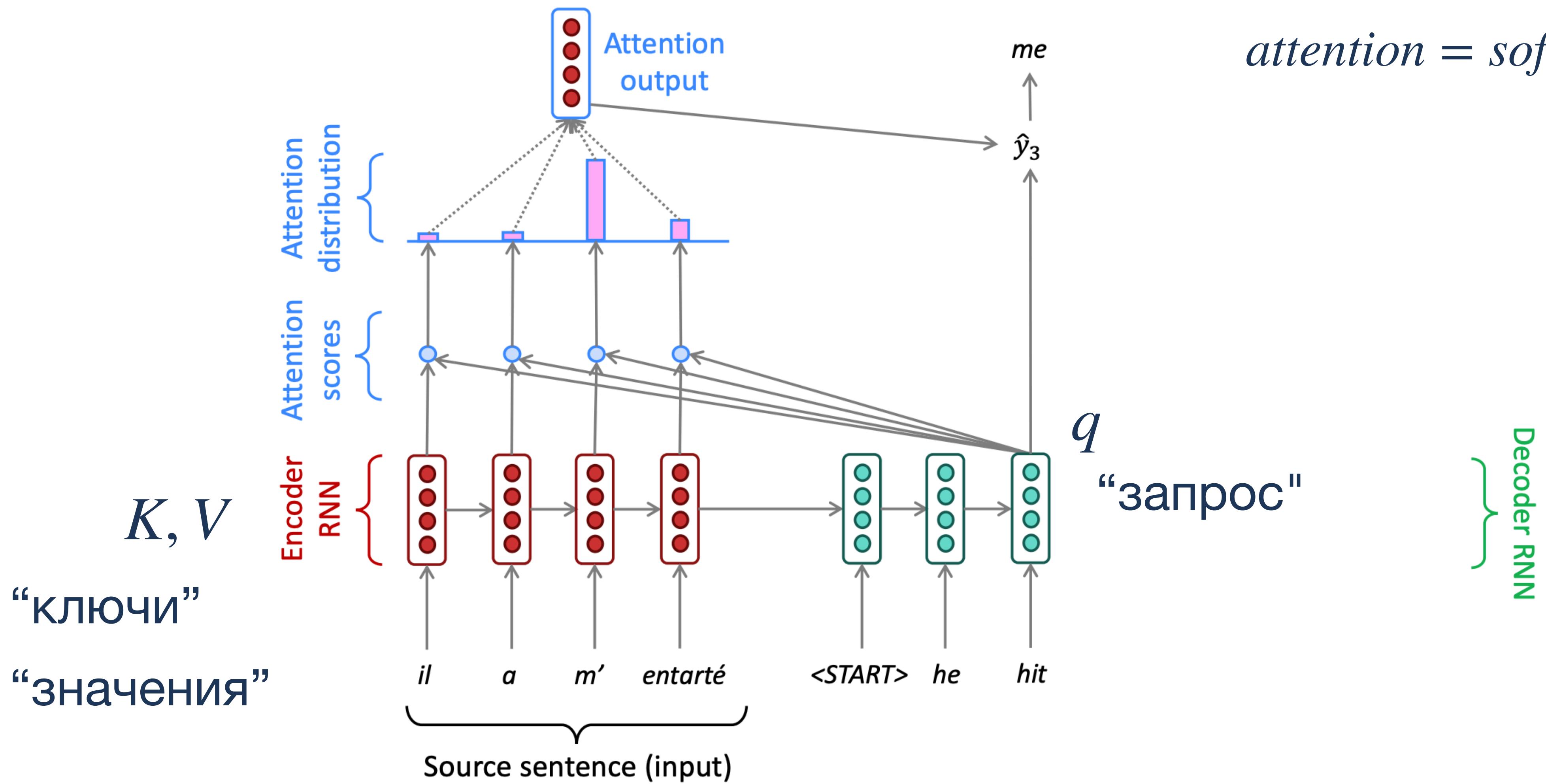
Attention



Attention



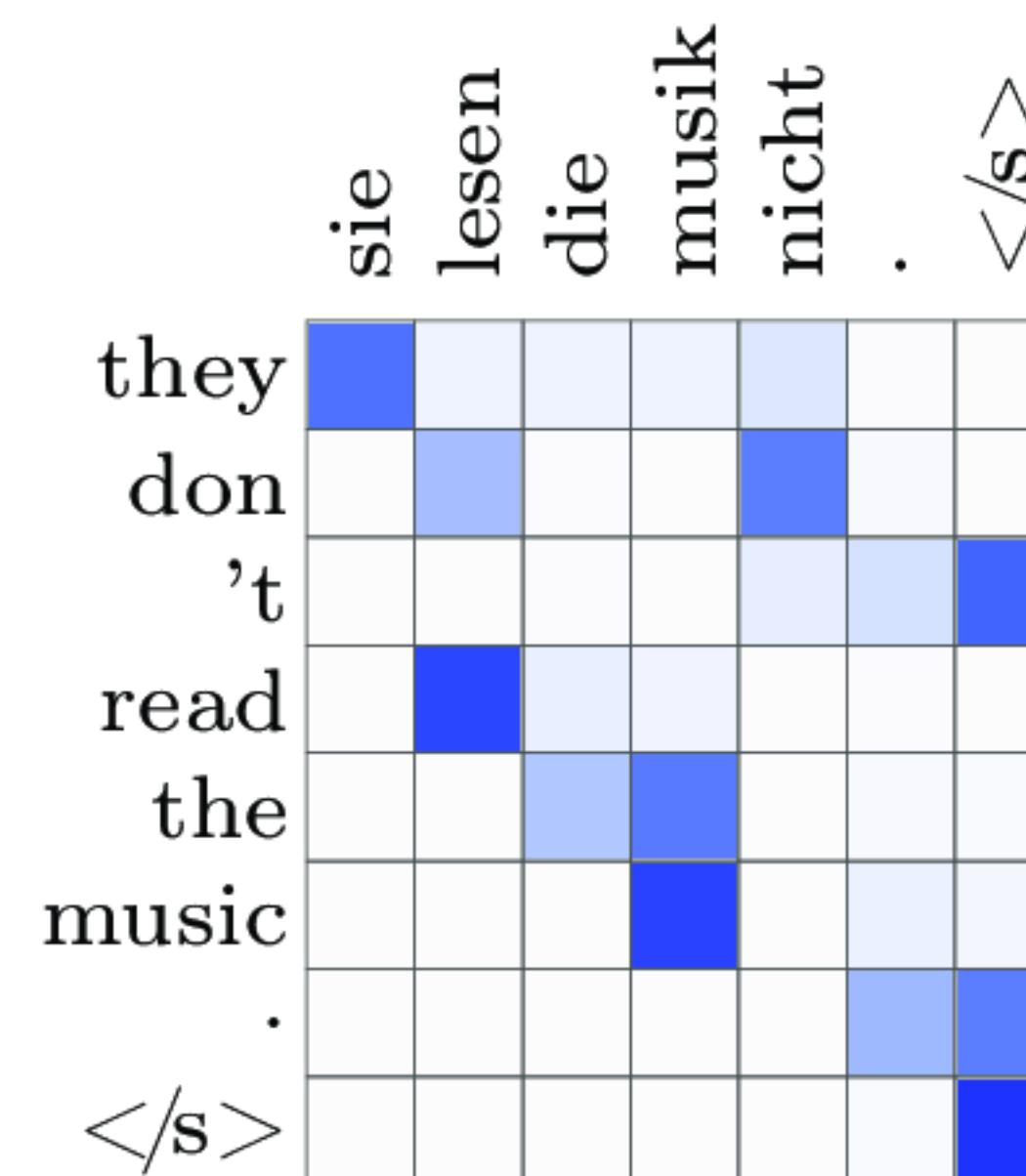
Attention



$$\text{attention} = \text{softmax}(qK^T)V$$

Attention

- Убирает bottleneck, всегда лучше качество
- Дает интерпретируемость (можно визуализировать распределения)



Transformer

Transformer

The screenshot shows the Google Translate interface. At the top, there's a navigation bar with a menu icon, the "Google Translate" logo, and a three-dot menu icon. Below the bar, there are two tabs: "Text" (selected) and "Documents". The main area has two language pairs: English to Russian and Russian to English. The English input field contains the sentence: "They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense." The Russian output field contains the translation: "Они были последними людьми, от которых вы ожидали быть вовлеченными во что-то странное или таинственное, потому что они просто не выдержали такой чепухи." Below the input and output fields are various interaction icons like microphone, speaker, and edit.

≡ Google Translate

⋮

Text Documents

ENGLISH - DETECTED RUSSIAN ENGLISH SPANISH RUSSIAN ENGLISH SPANISH

They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.

Они были последними людьми, от которых вы ожидали быть вовлеченными во что-то странное или таинственное, потому что они просто не выдержали такой чепухи.

Oni byli poslednimi lyud'mi, ot kotorikh vy ozhidali byt' vovlechennymi vo chto-to strannoye ili tainstvennoye, potomu chto oni prosto ne vyderzhali takoy chepukhi.

138/5000

Send feedback

Изначально предложен для перевода

Transformer

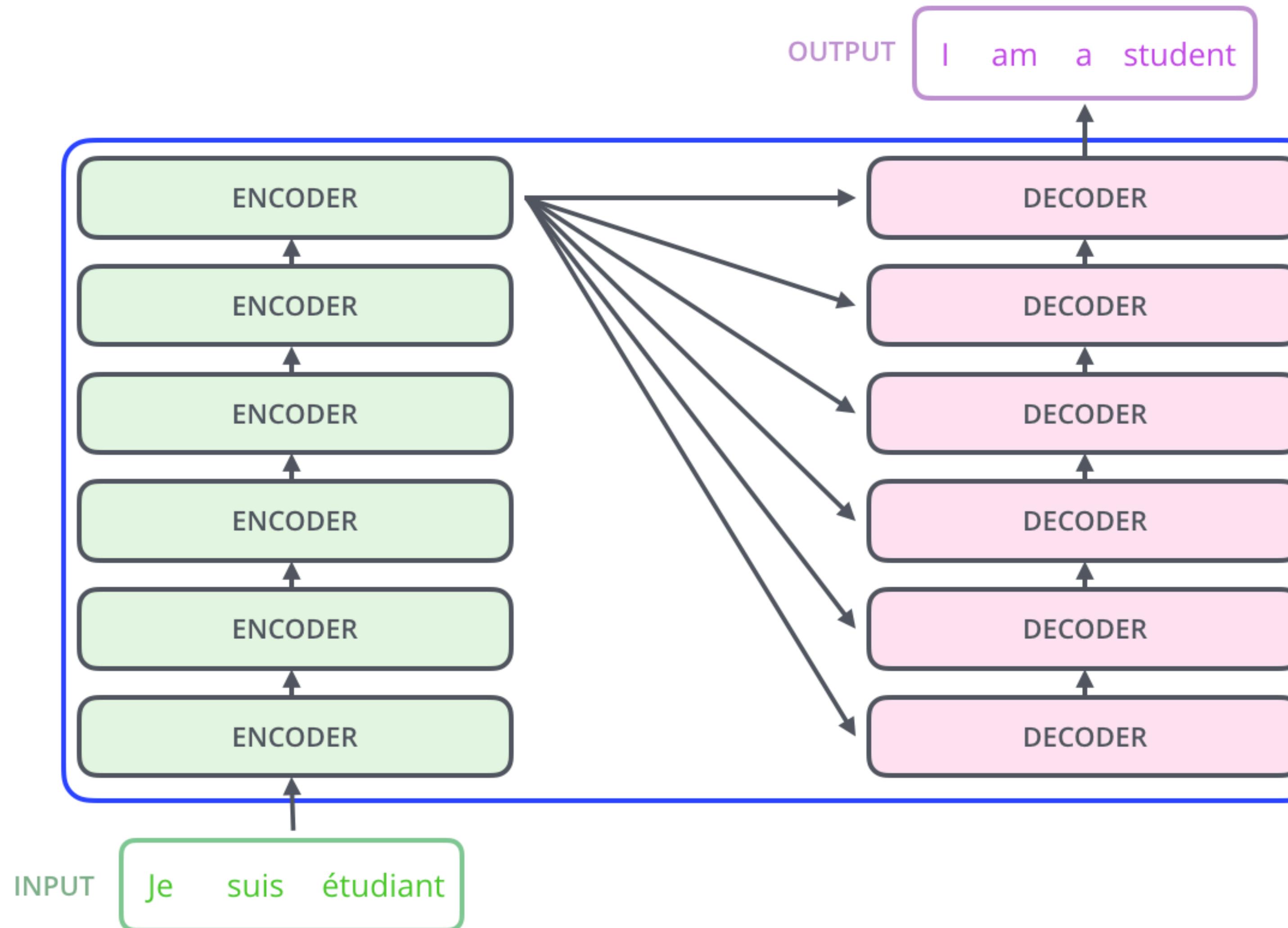
Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	41.29	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1		$3.3 \cdot 10^{18}$
Transformer (big)	28.4	41.8		$2.3 \cdot 10^{19}$

Изначально предложен для перевода

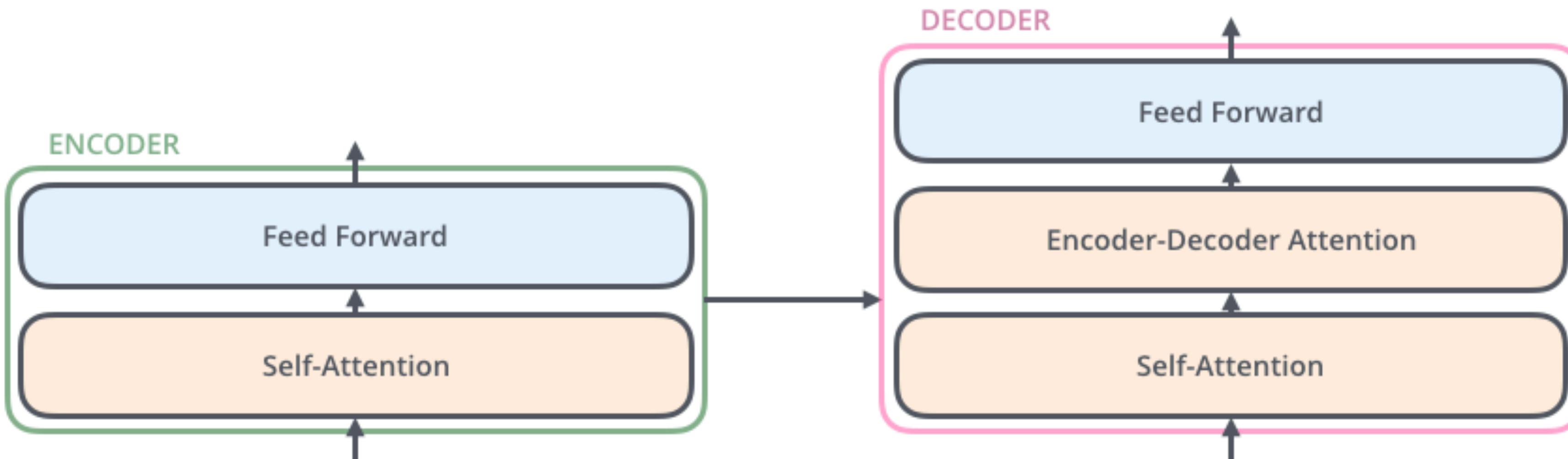
- Лучше качество (в других задачах тоже!)
- Быстрее

Transformer



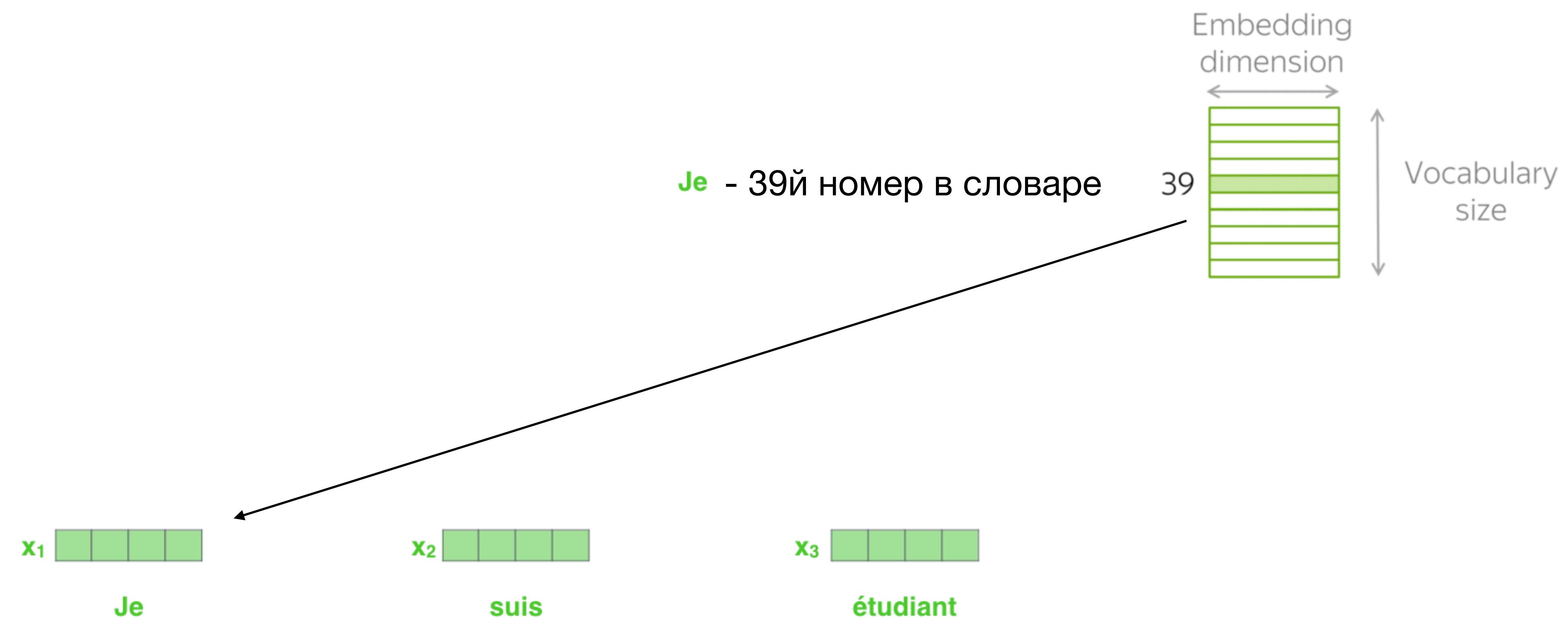
Transformer

Encoder block/ Decoder block

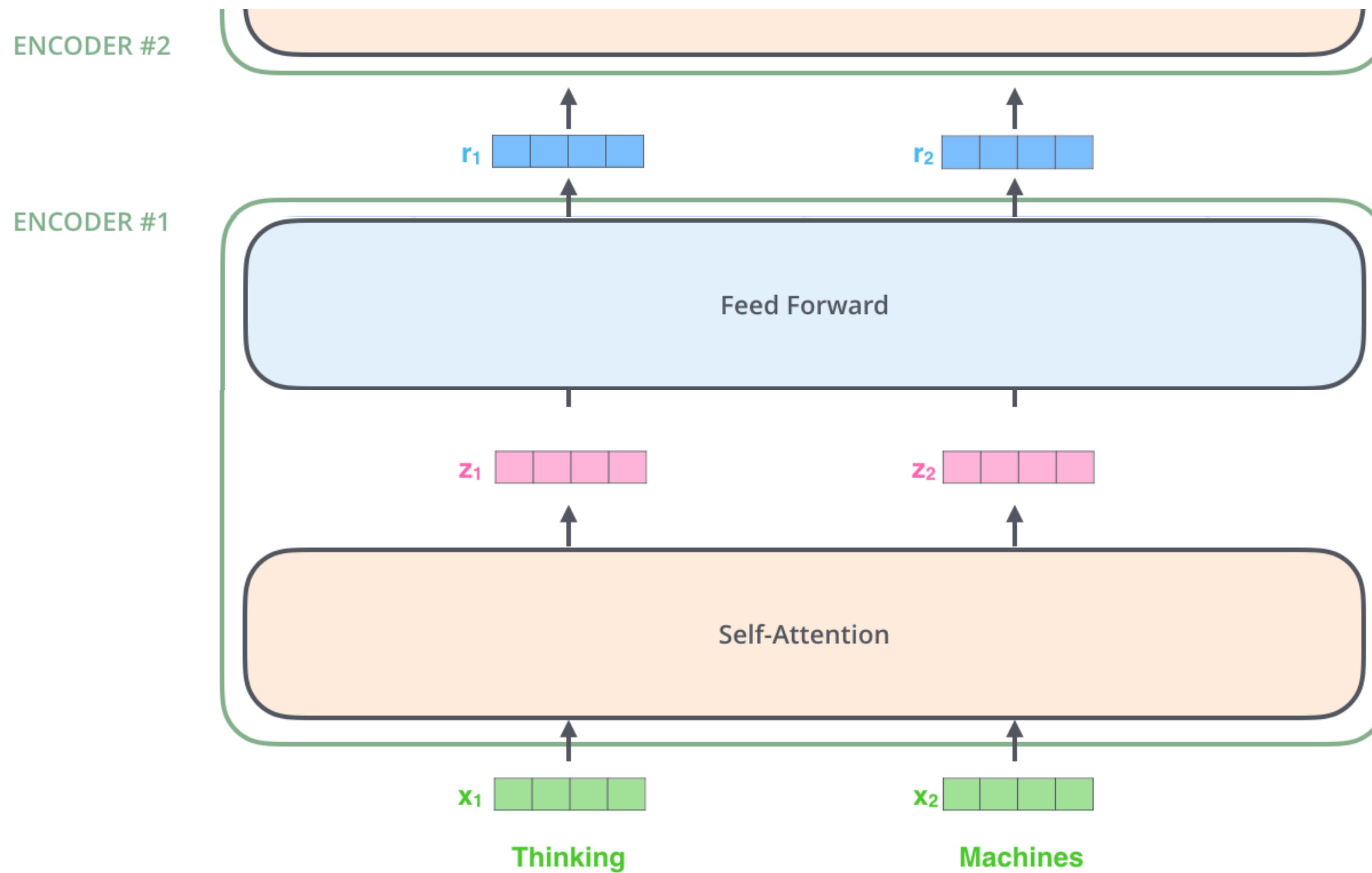


Transformer: encoder block

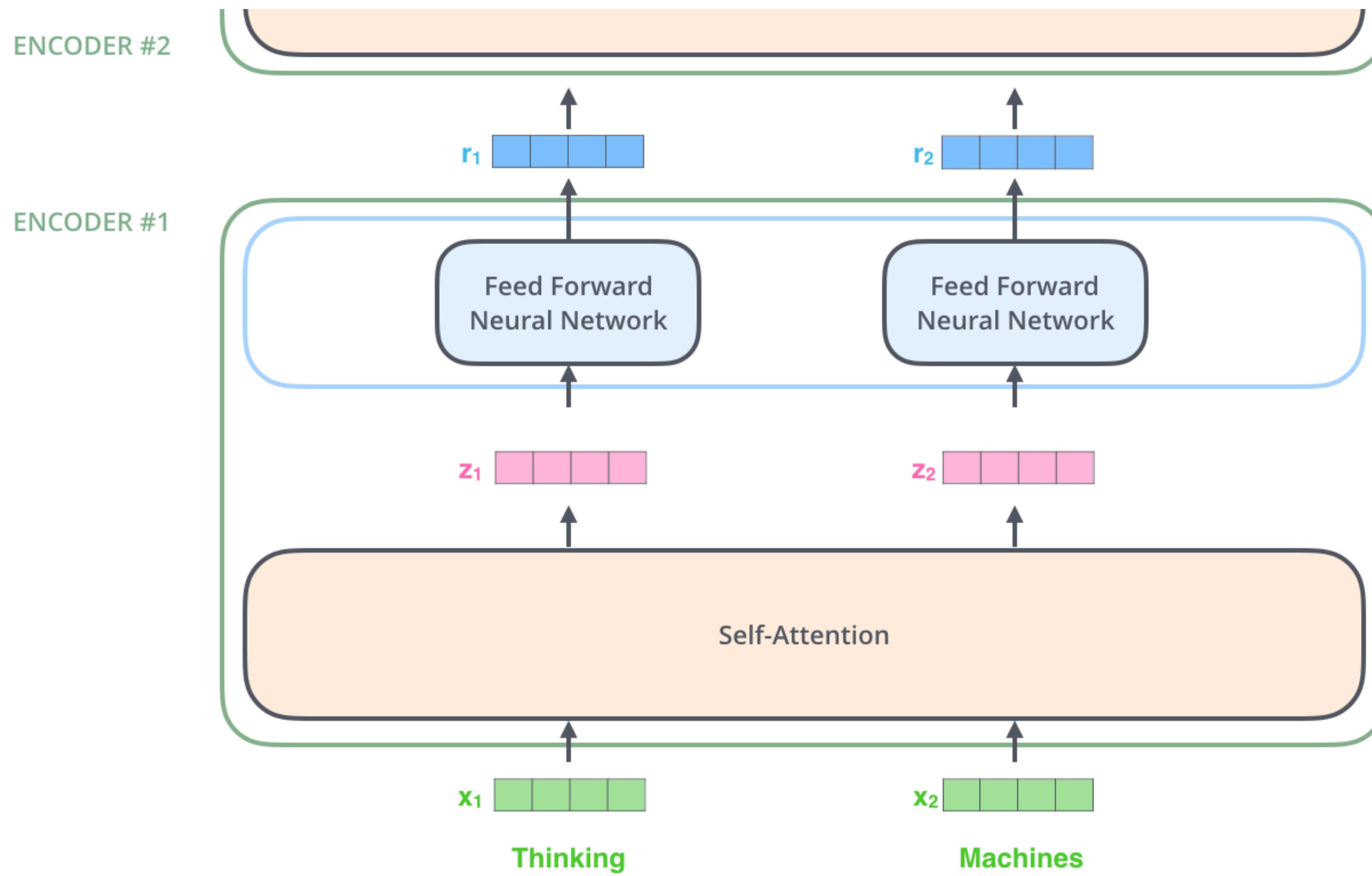
Слой эмбеддингов -
для каждого входного слова достаем из таблицы вектор (обучаемый)



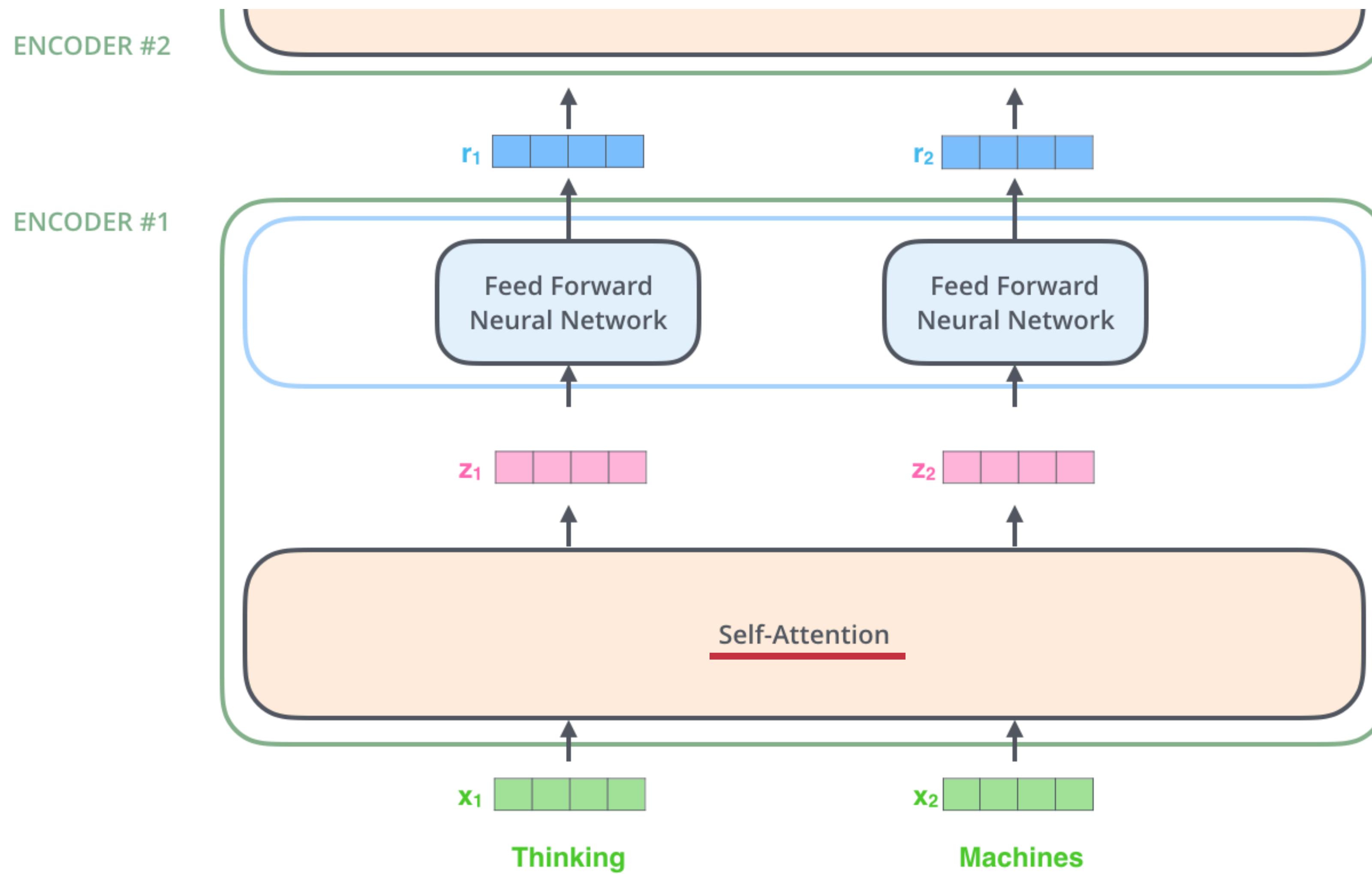
Transformer: encoder block



Transformer: encoder block



Transformer: encoder block



Transformer: self-attention

Seq2seq attention: между decoder vector и encoder vectors

- какие слова в input “важны” для предсказания текущего output

$$\text{attention} = \text{softmax}(qK^T)V$$

Transformer: self-attention

Seq2seq attention: между decoder vector и encoder vectors

- какие слова в input “важны” для предсказания текущего output

$$\text{attention} = \text{softmax}(qK^T)V$$

Self-attention (encoder): между encoder векторами

- какие слова в input как соотносятся между собой

$$\text{attention} = \text{softmax}\left(\frac{QK^T}{d}\right)V$$

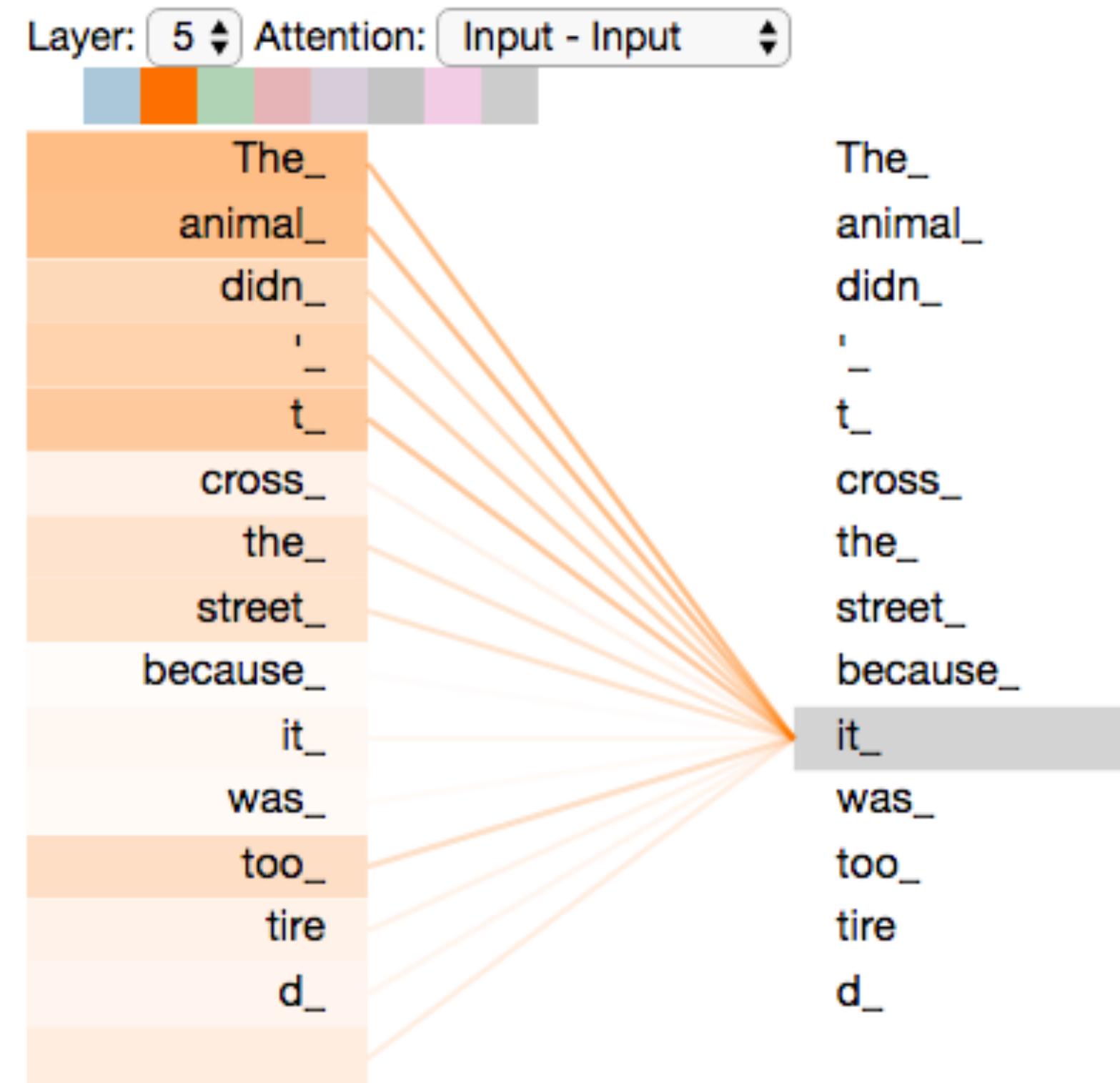
Transformer: self-attention

"The animal didn't cross the street because **it** was too tired"

На что указывает **it**?

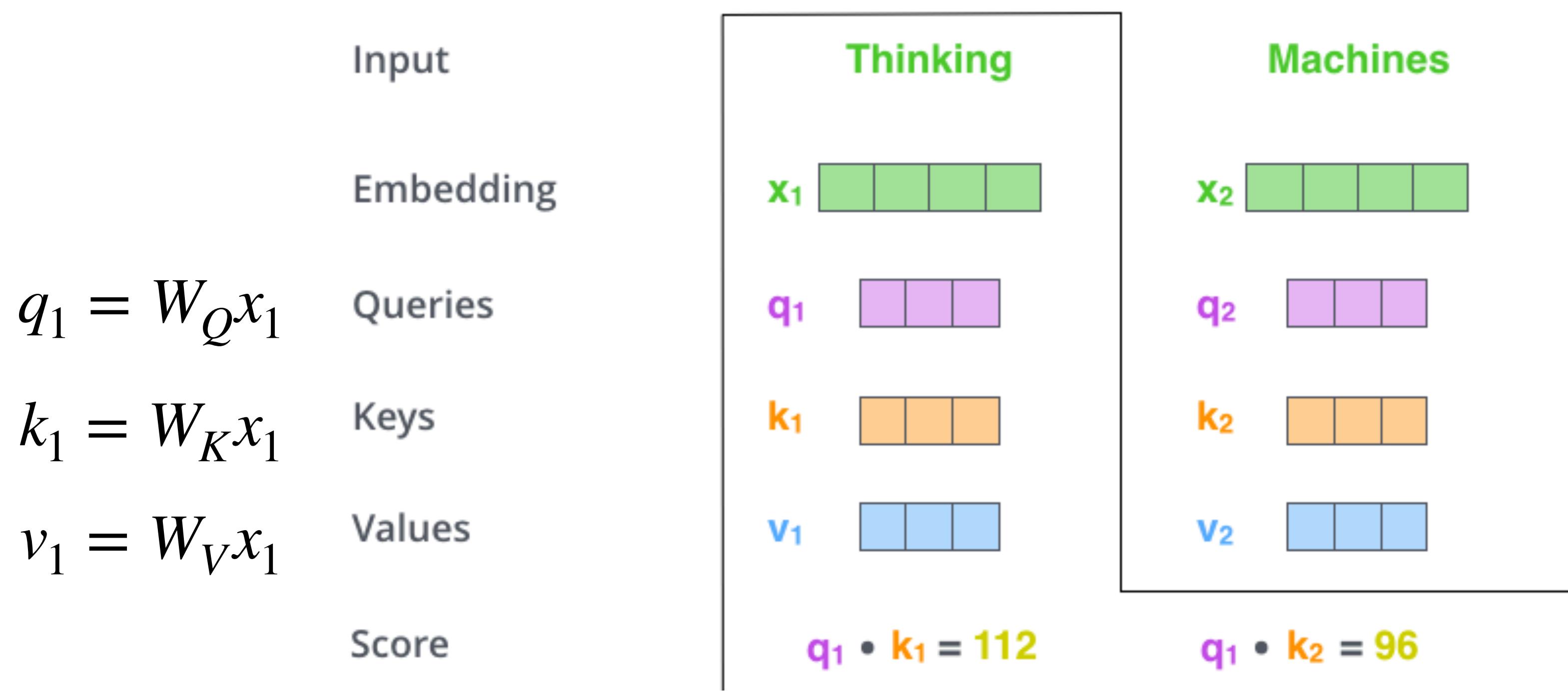
Transformer: self-attention

"The animal didn't cross the street because **it** was too tired"



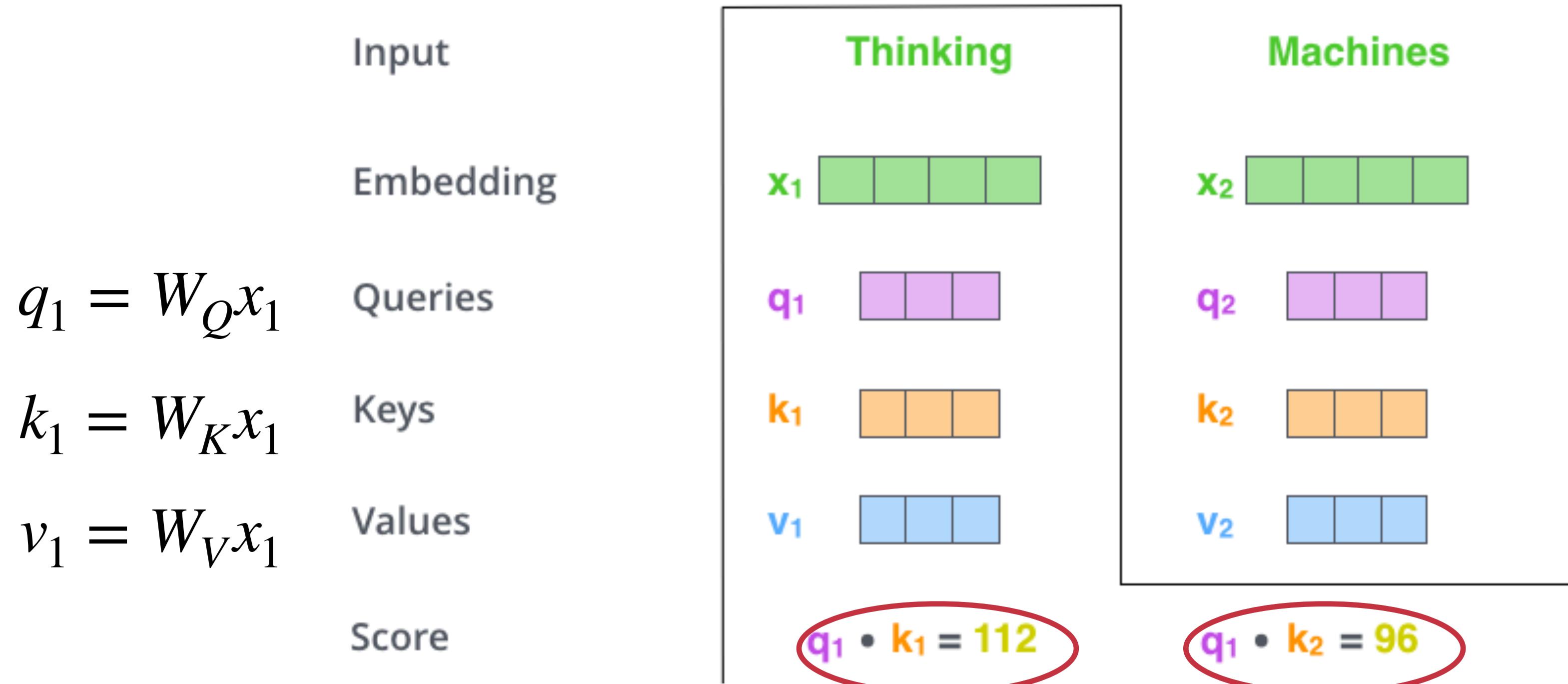
Transformer: self-attention

$$\text{attention} = \text{softmax}\left(\frac{QK^T}{d}\right)V$$



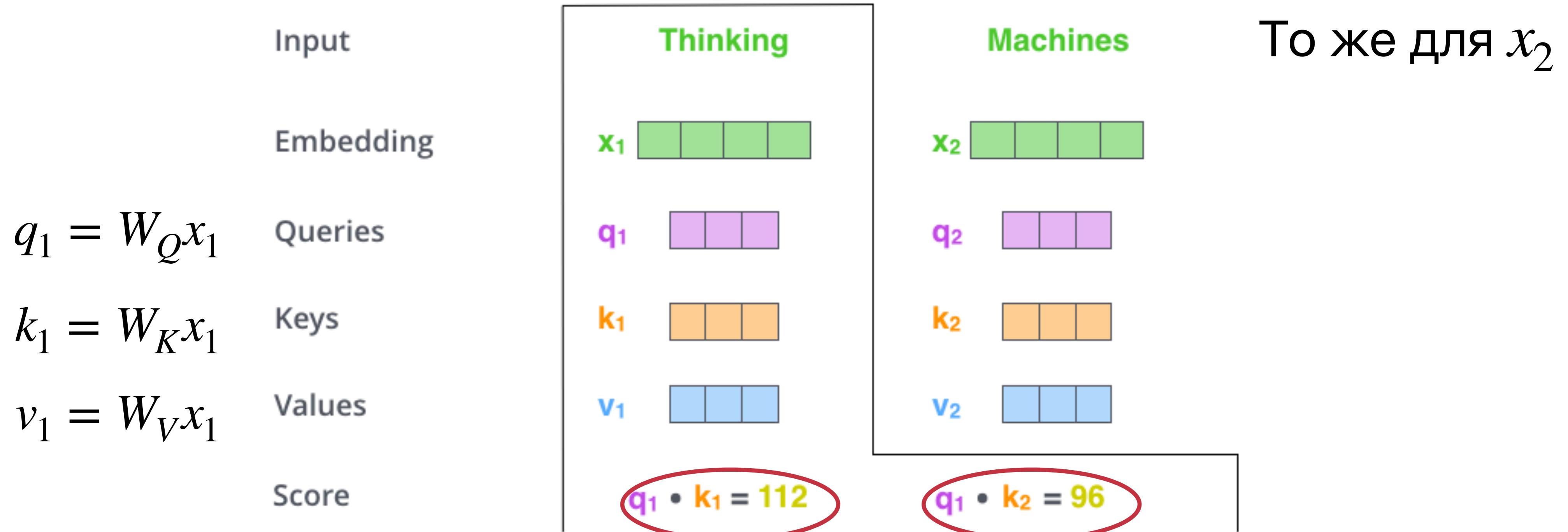
Transformer: self-attention

$$\text{attention} = \text{softmax}\left(\frac{QK^T}{d}\right)V$$



Transformer: self-attention

$$\text{attention} = \text{softmax}\left(\frac{QK^T}{d}\right)V$$



Transformer: self-attention

$$\text{attention} = \text{softmax}\left(\frac{QK^T}{d}\right)V$$

Input

Embedding

$$q_1 = W_Q x_1$$

Queries

$$k_1 = W_K x_1$$

Keys

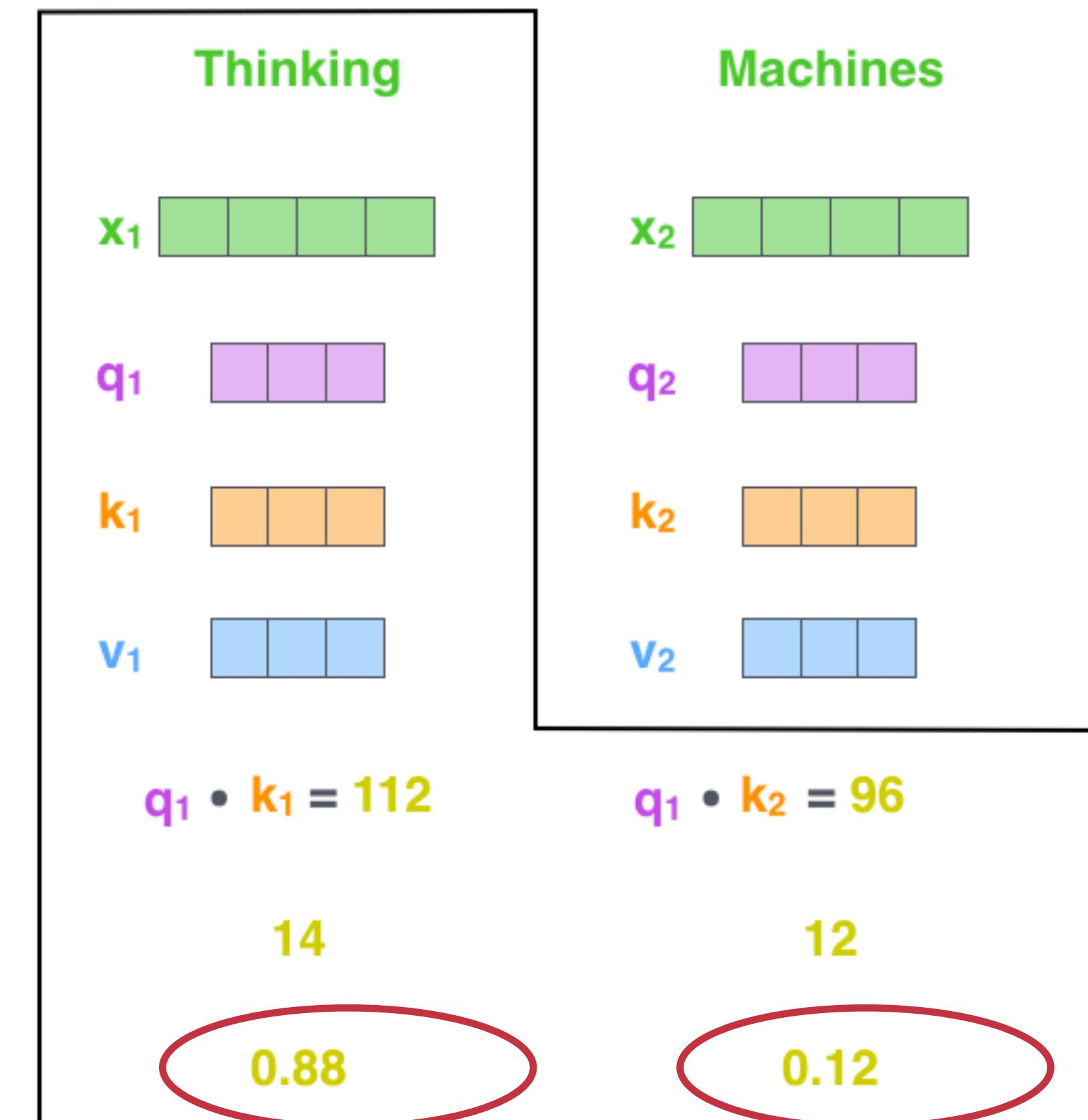
$$v_1 = W_V x_1$$

Values

Score

Divide by 8 ($\sqrt{d_k}$)

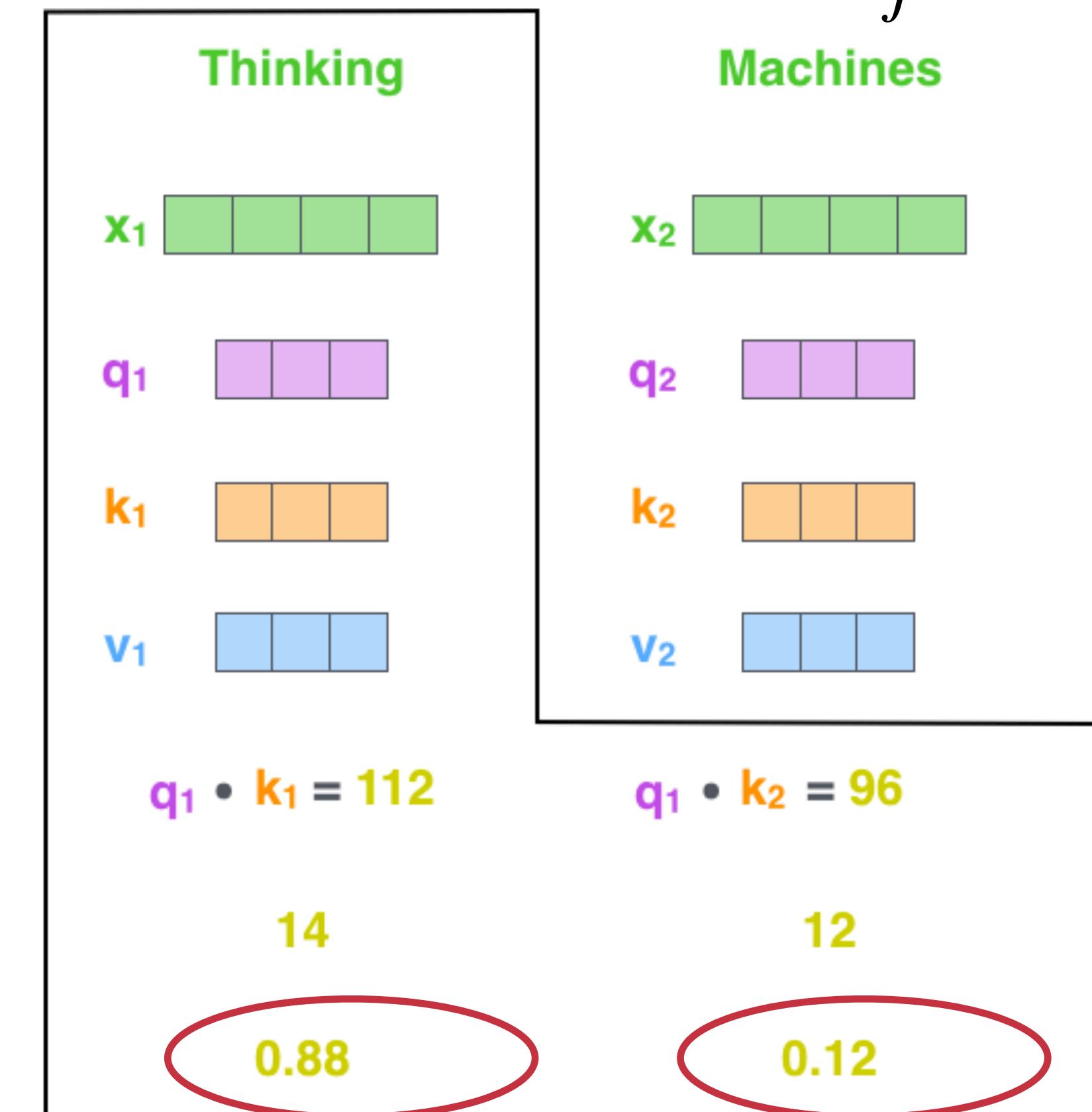
Softmax



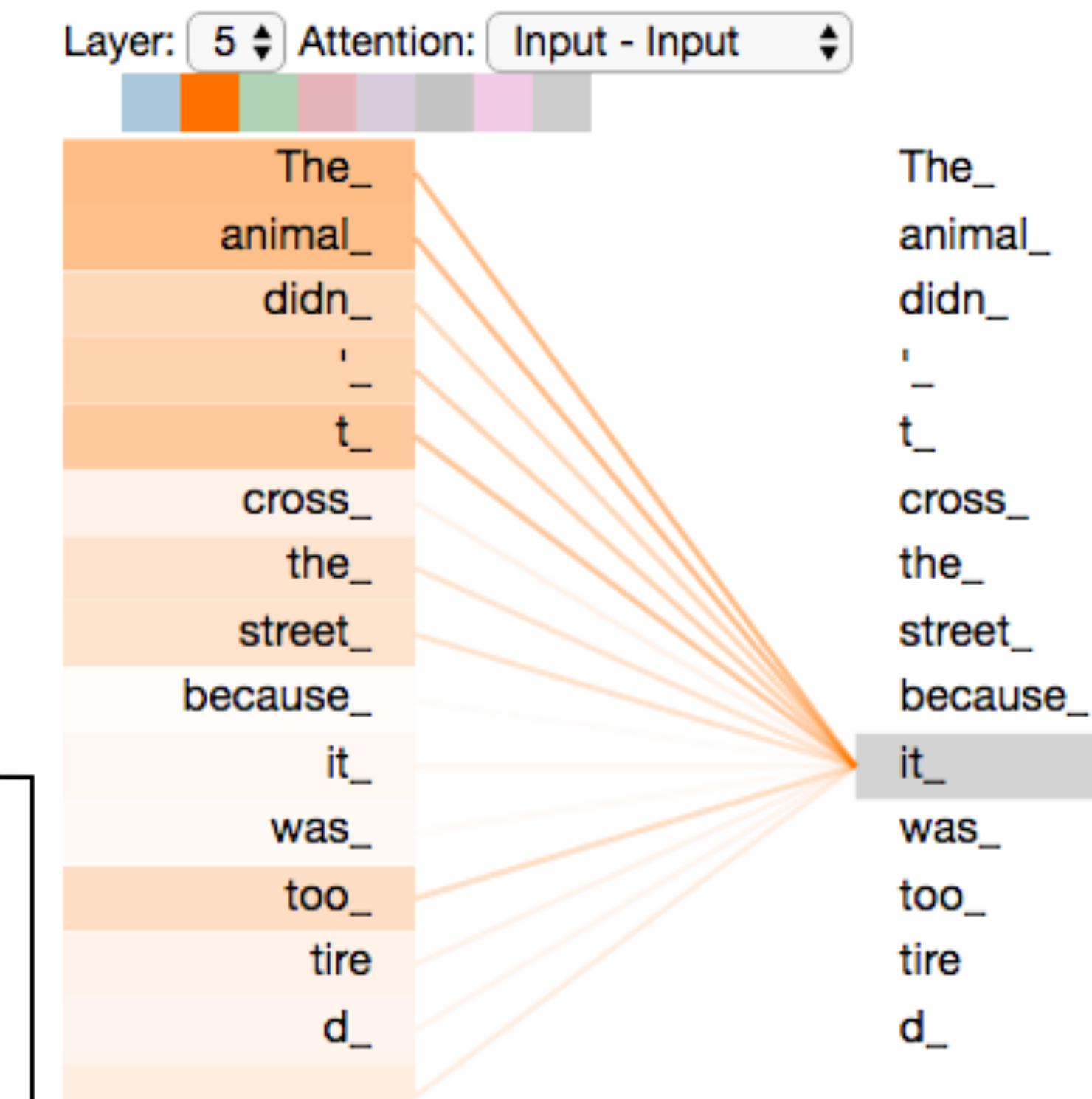
Transformer: self-attention

$$\text{attention} = \text{softmax}\left(\frac{QK^T}{d}\right)V$$

Input
 Embedding
 $q_1 = W_Q x_1$ Queries
 $k_1 = W_K x_1$ Keys
 $v_1 = W_V x_1$ Values
 Score
 Divide by 8 ($\sqrt{d_k}$)
 Softmax



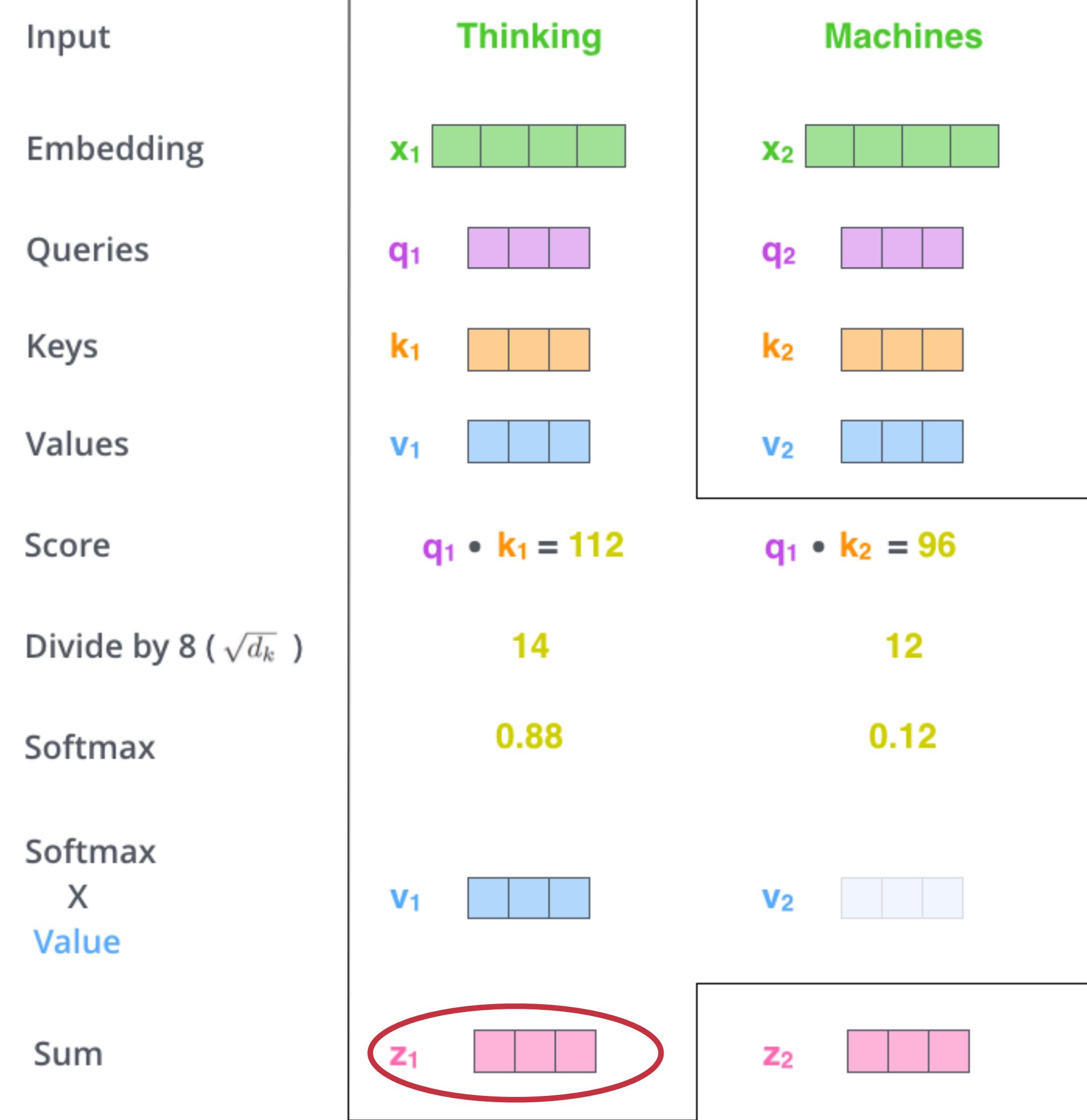
Для каждого x_i получаем веса, насколько он соотносится с x_j - всеми элементами входа



Transformer: self-attention

$$\text{attention} = \text{softmax}\left(\frac{QK^T}{d}\right)V$$

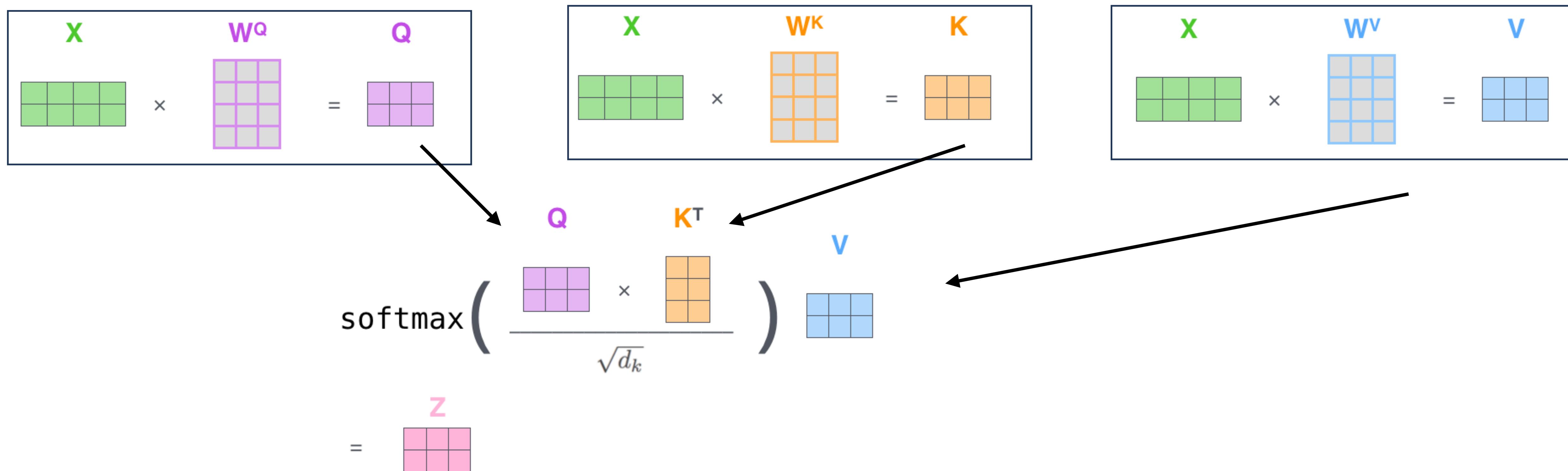
Взвешенная сумма всех v_j ,
веса - из softmax



Transformer: self-attention

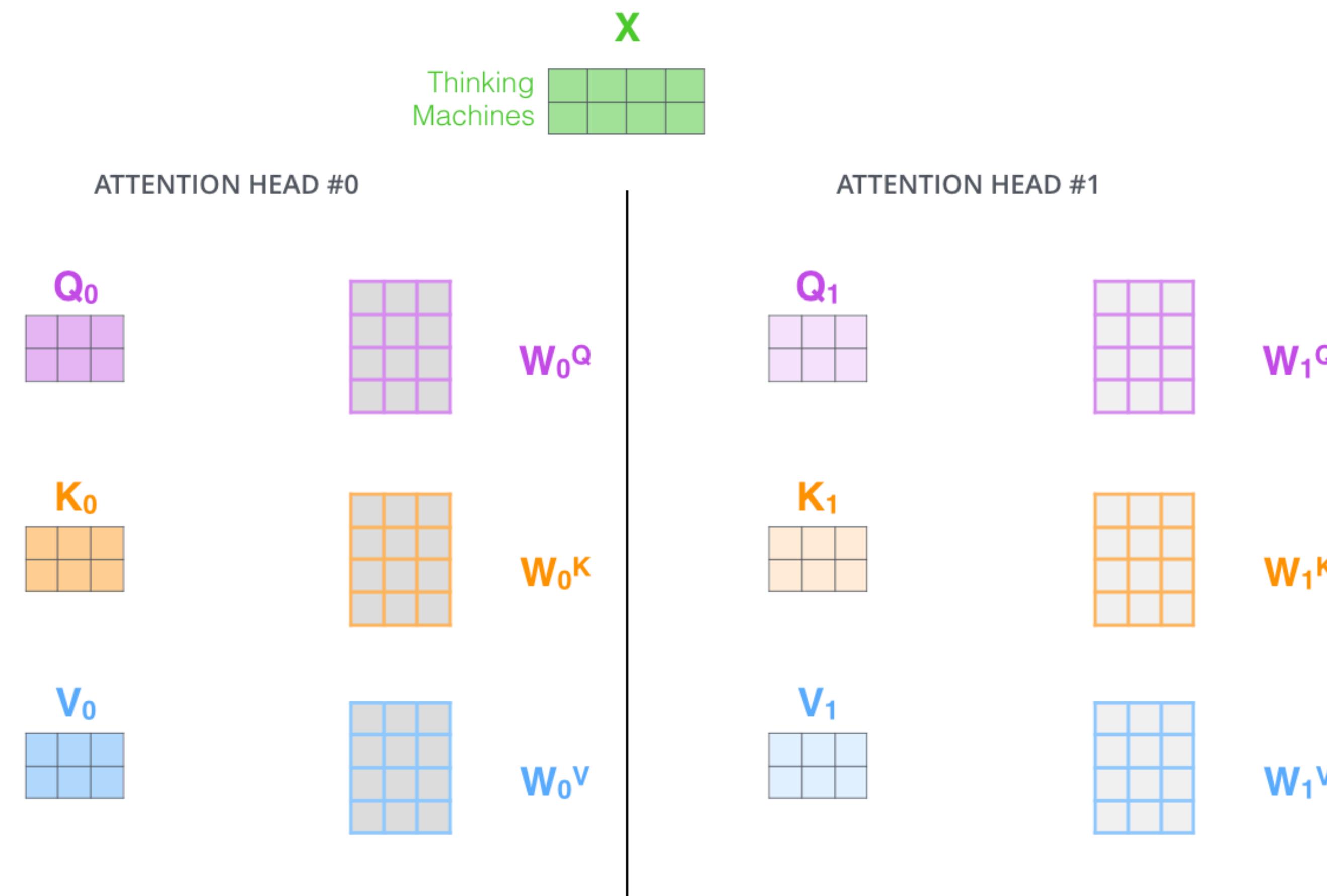
$$\text{attention} = \text{softmax}\left(\frac{QK^T}{d}\right)V$$

Матричная запись



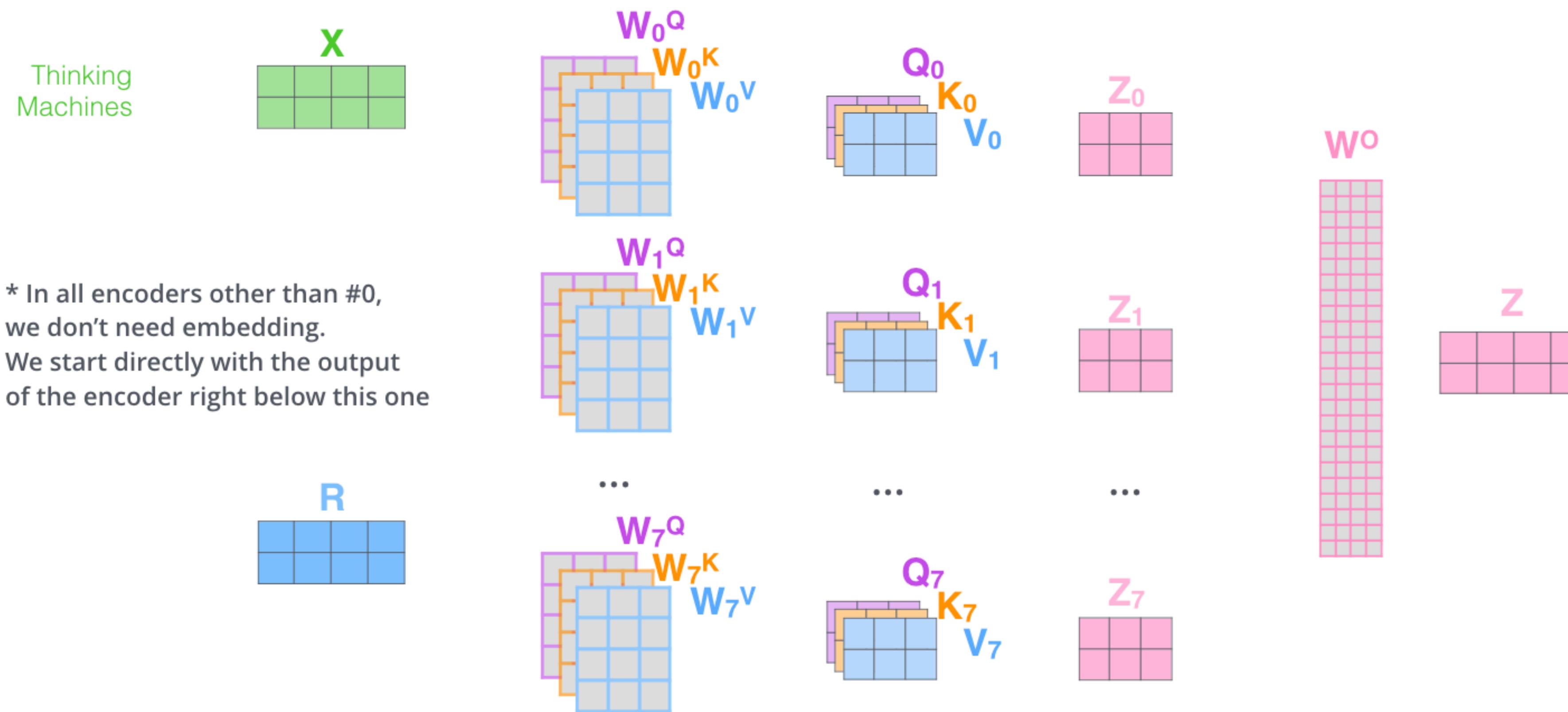
Transformer: multihead self-attention

Используем несколько self-attention слоев сразу - с разными весами



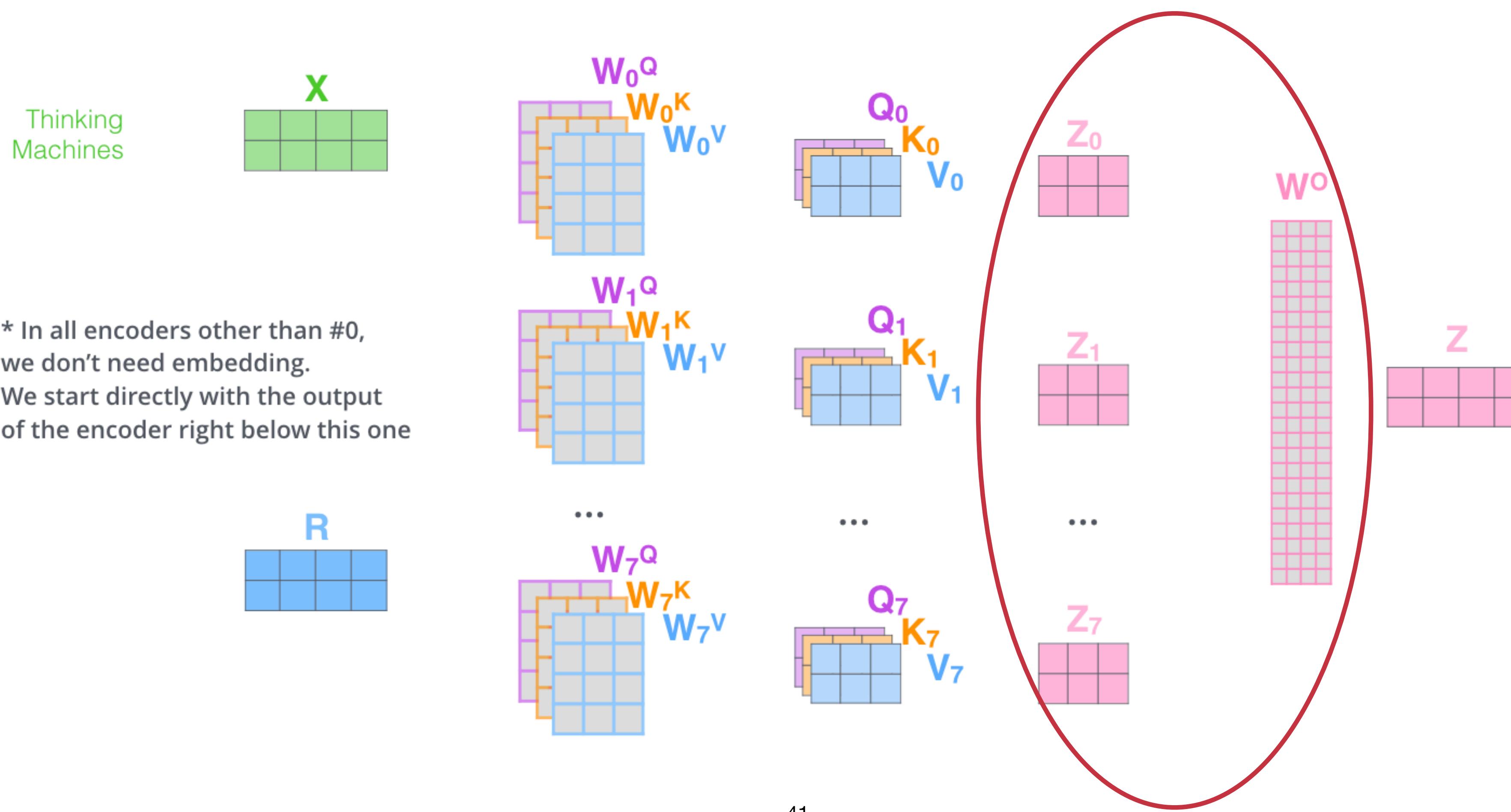
Transformer: multihead self-attention

Используем несколько self-attention слоев сразу - с разными весами



Transformer: multihead self-attention

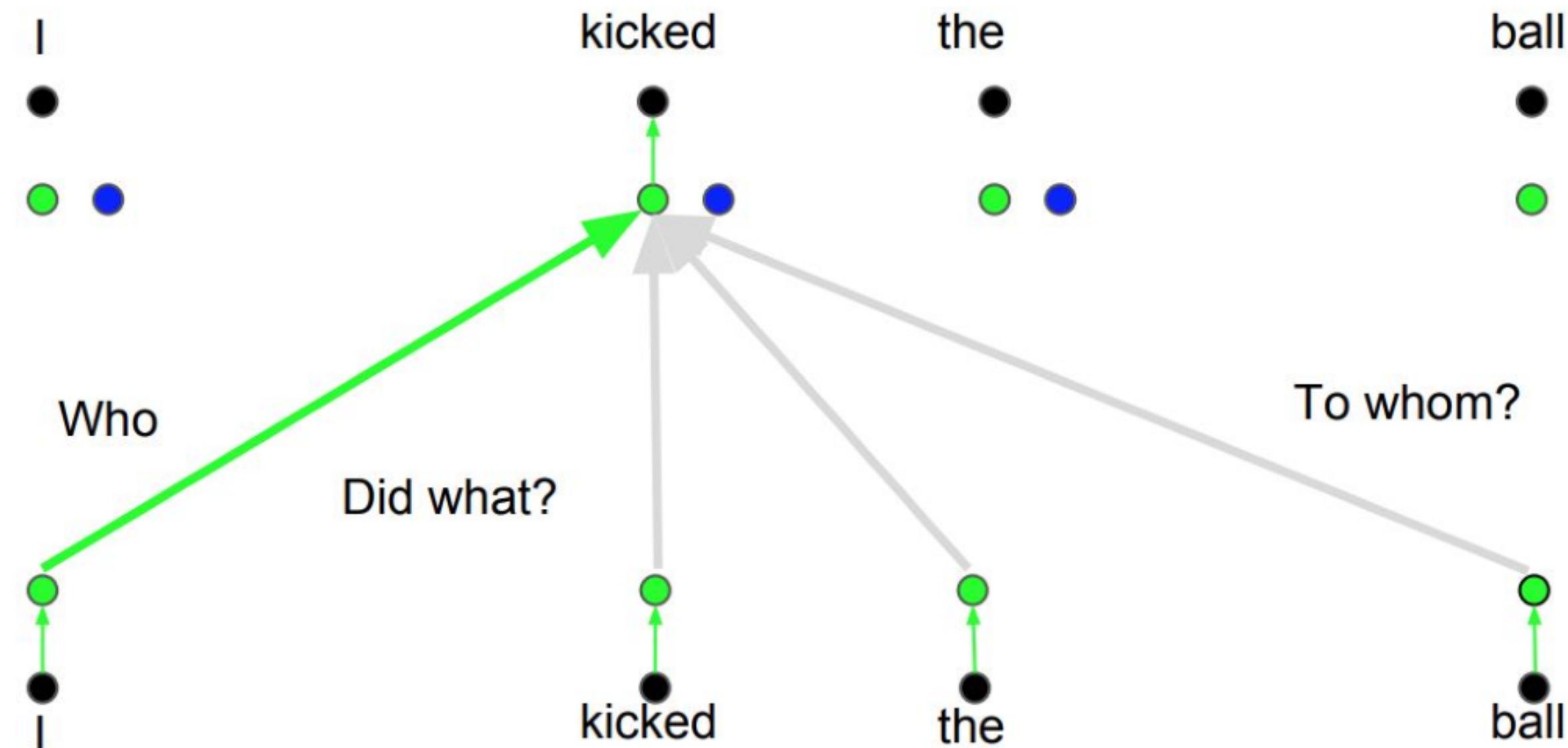
Используем несколько self-attention слоев сразу - с разными весами



Transformer: multihead self-attention

Мотивация:

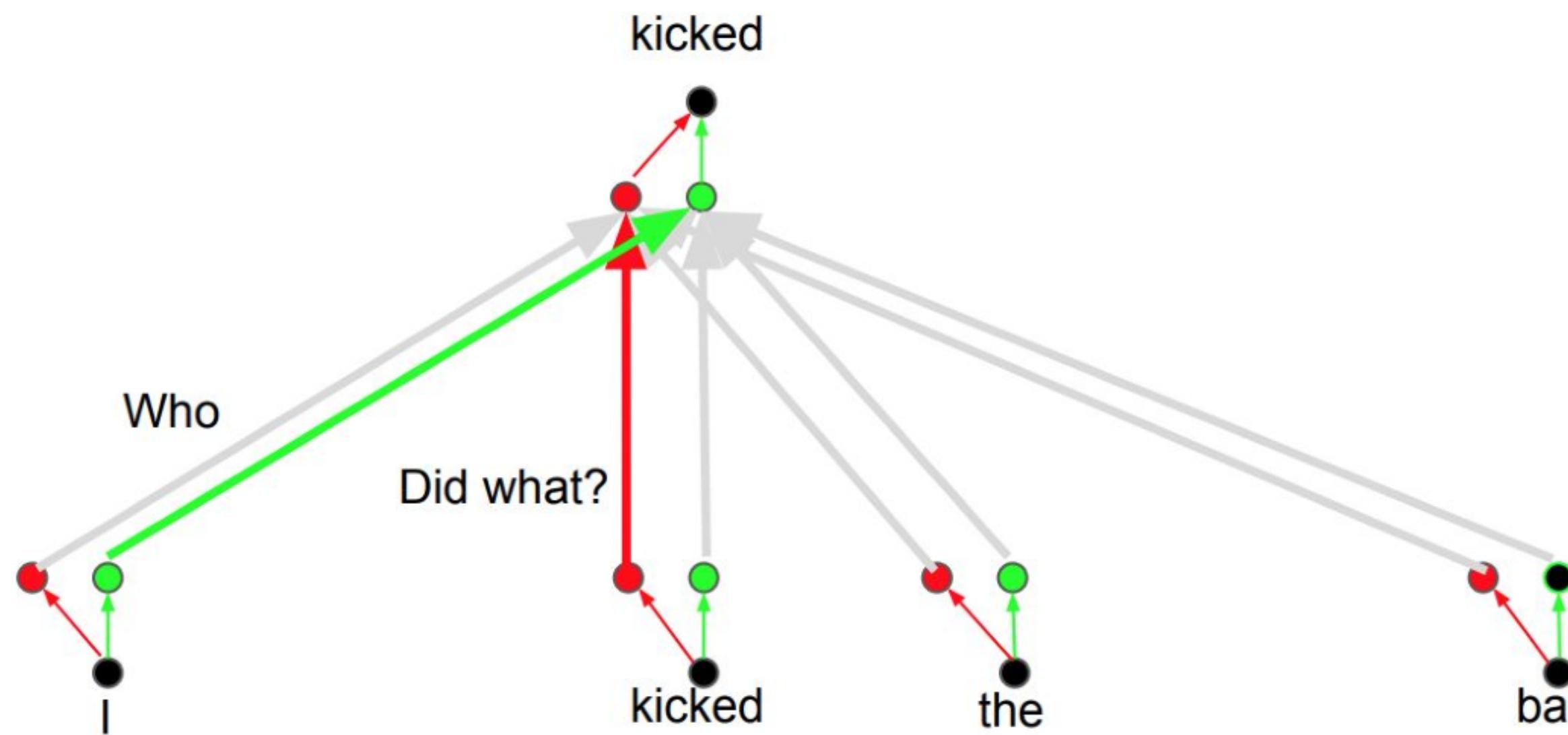
Attention head: Who



Transformer: multihead self-attention

Мотивация:

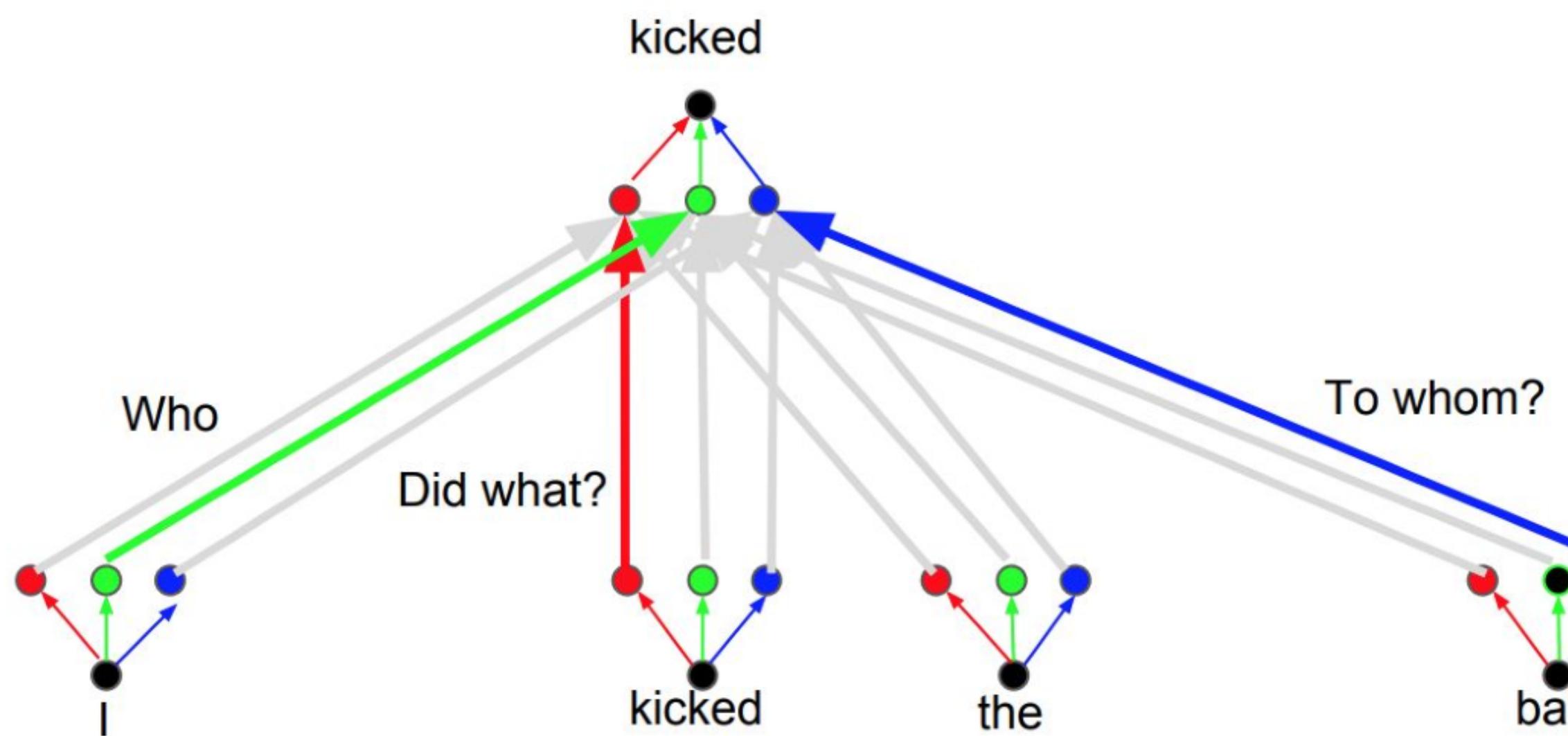
Attention head: Did What?



Transformer: multihead self-attention

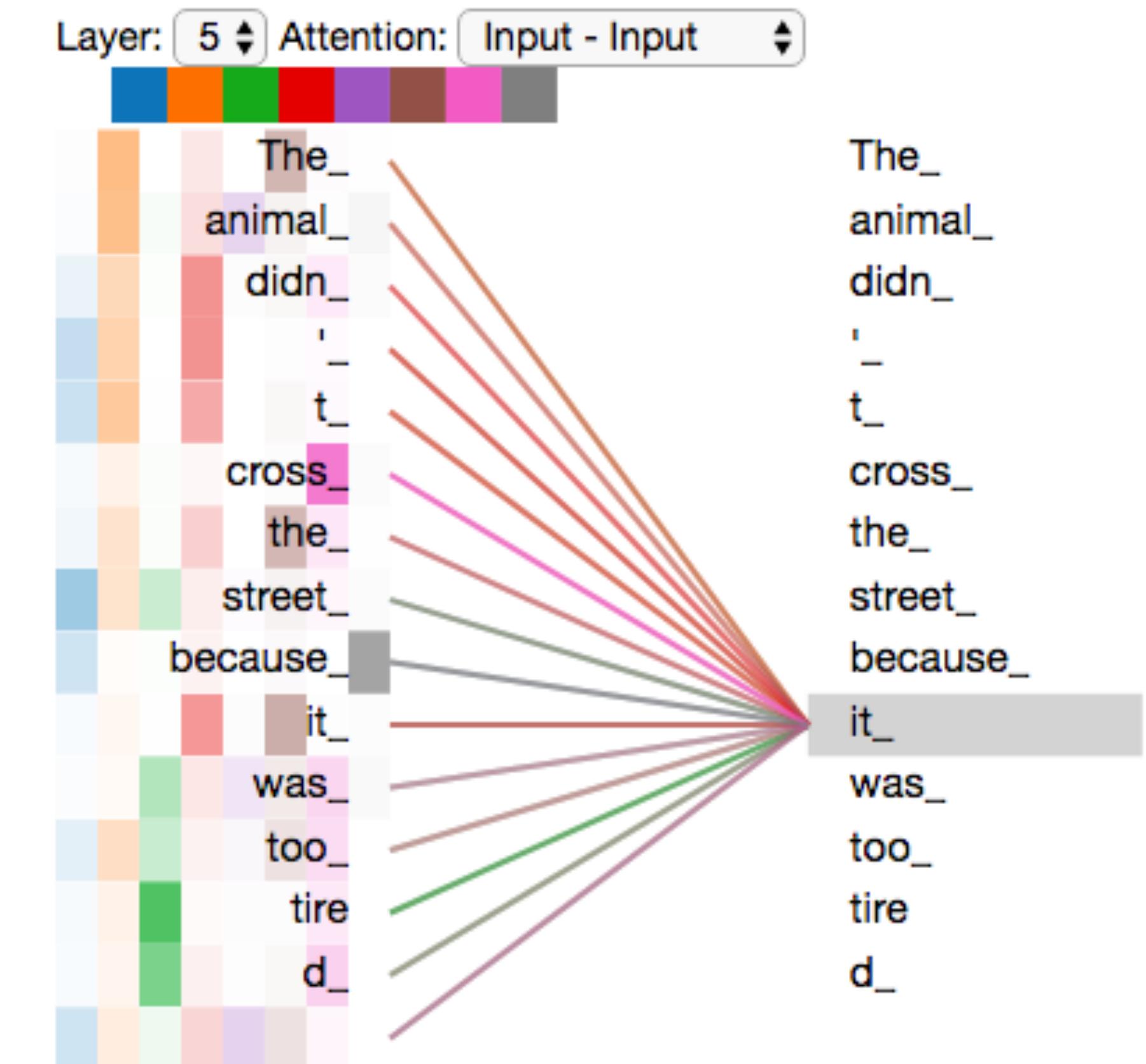
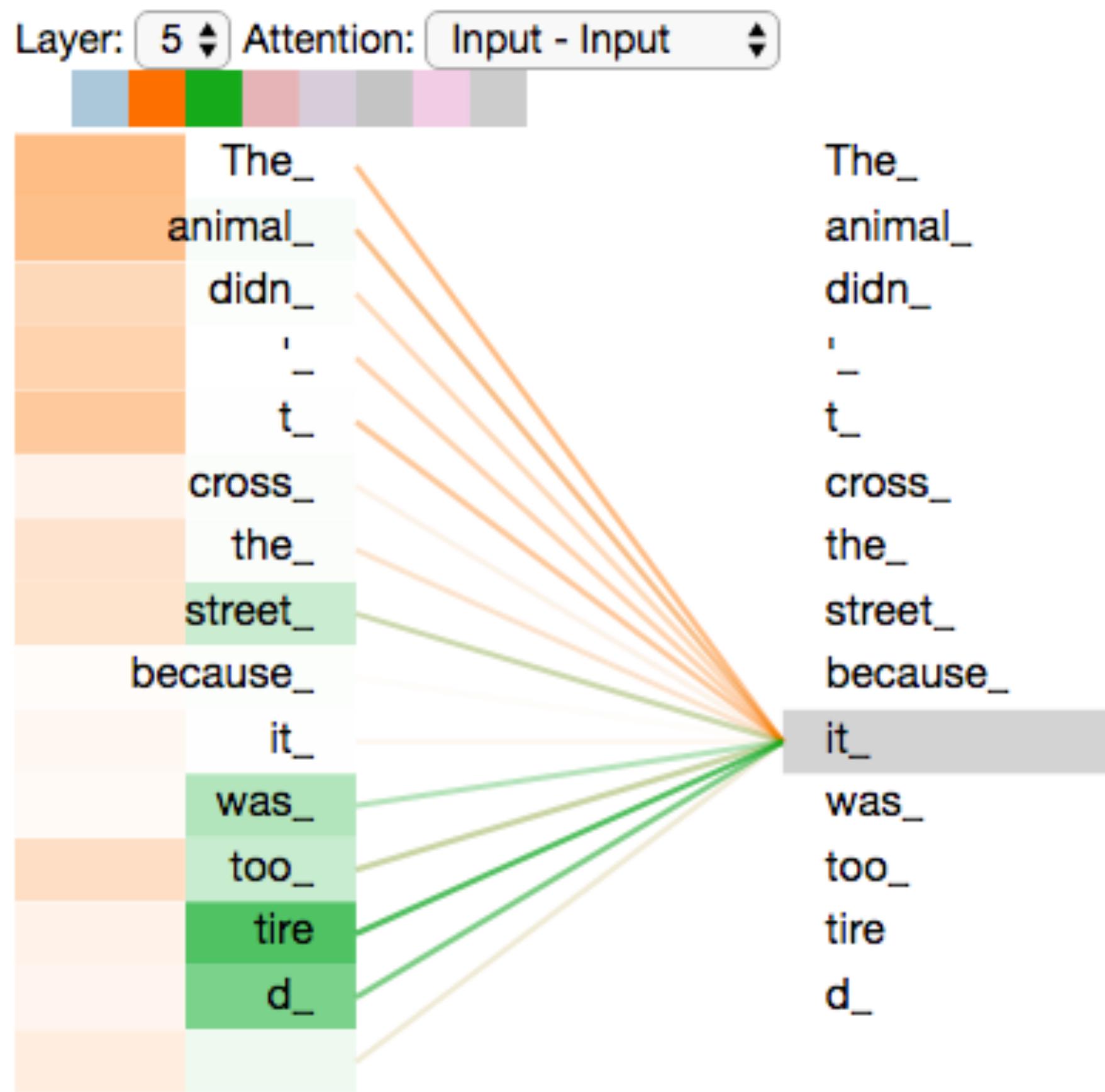
Мотивация:

Attention head: To Whom?



Transformer: multihead self-attention

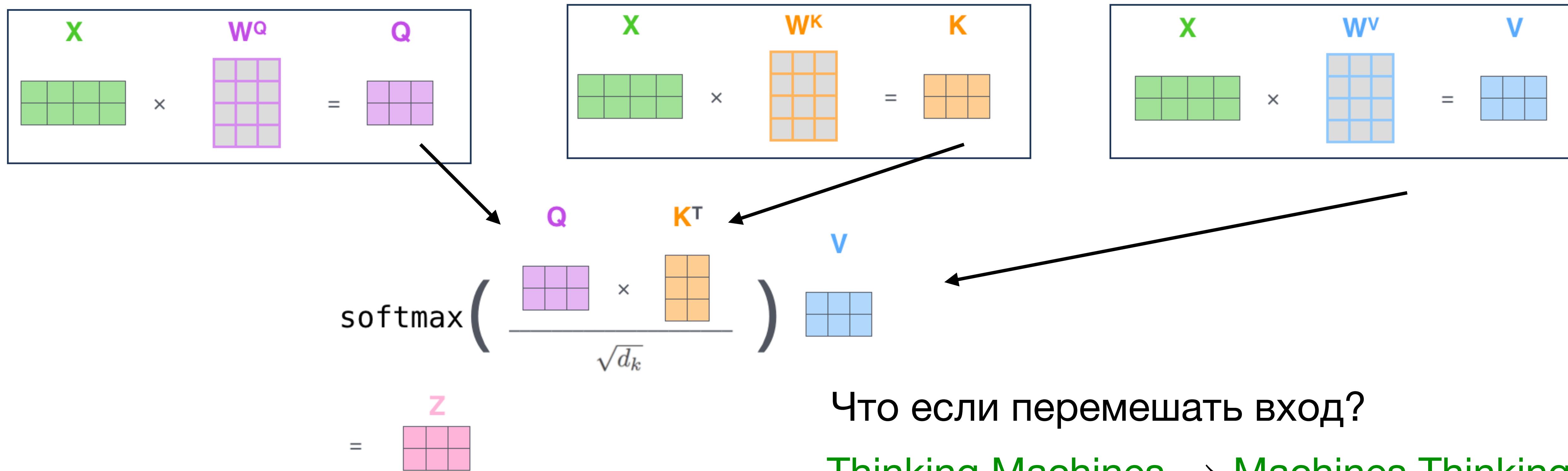
Мотивация: разные self-attention “обращают внимание” на разные признаки



[Image credit](#)

Transformer: multihead self-attention

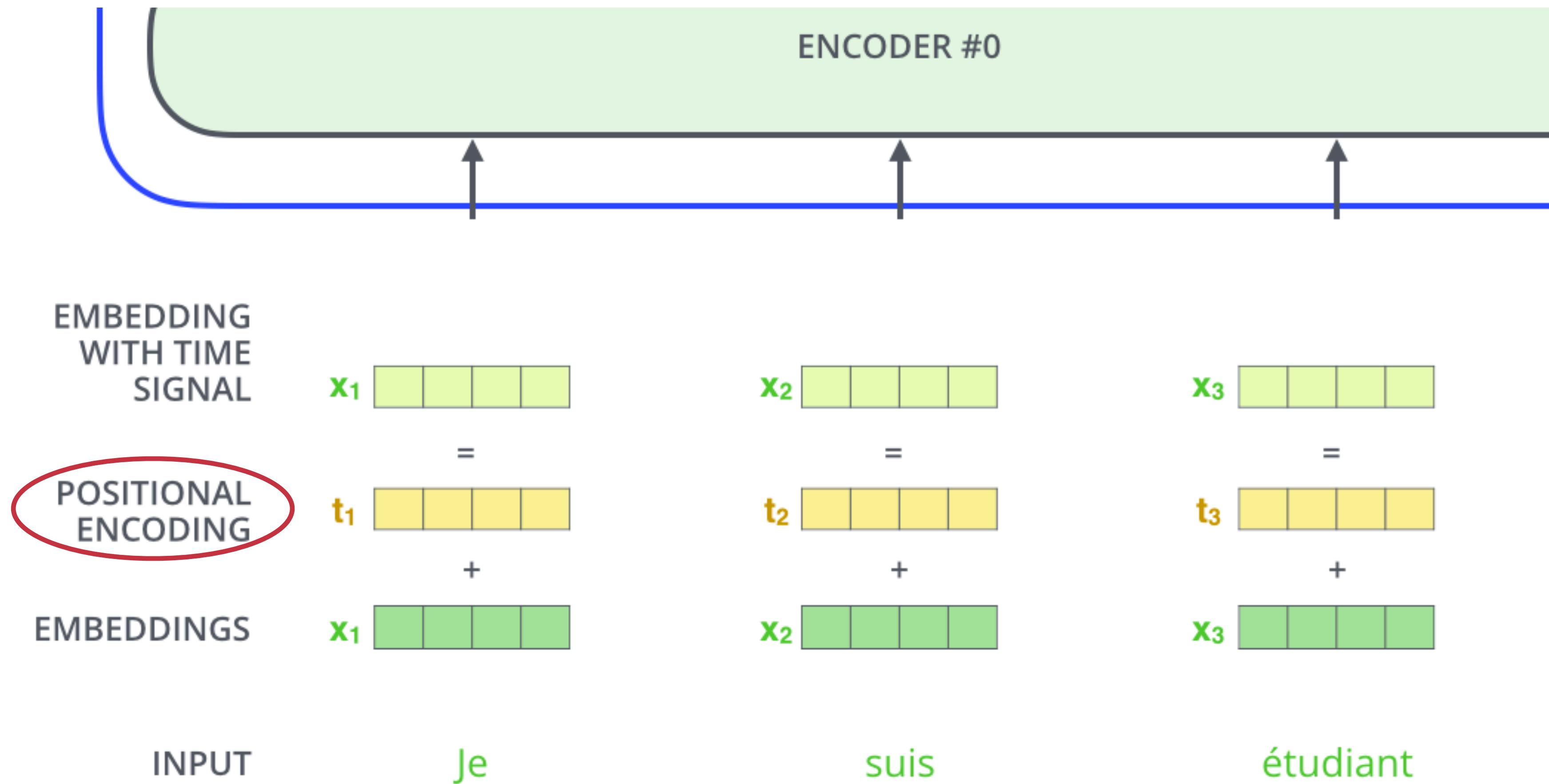
$$\text{attention} = \text{softmax}\left(\frac{QK^T}{d}\right)V$$



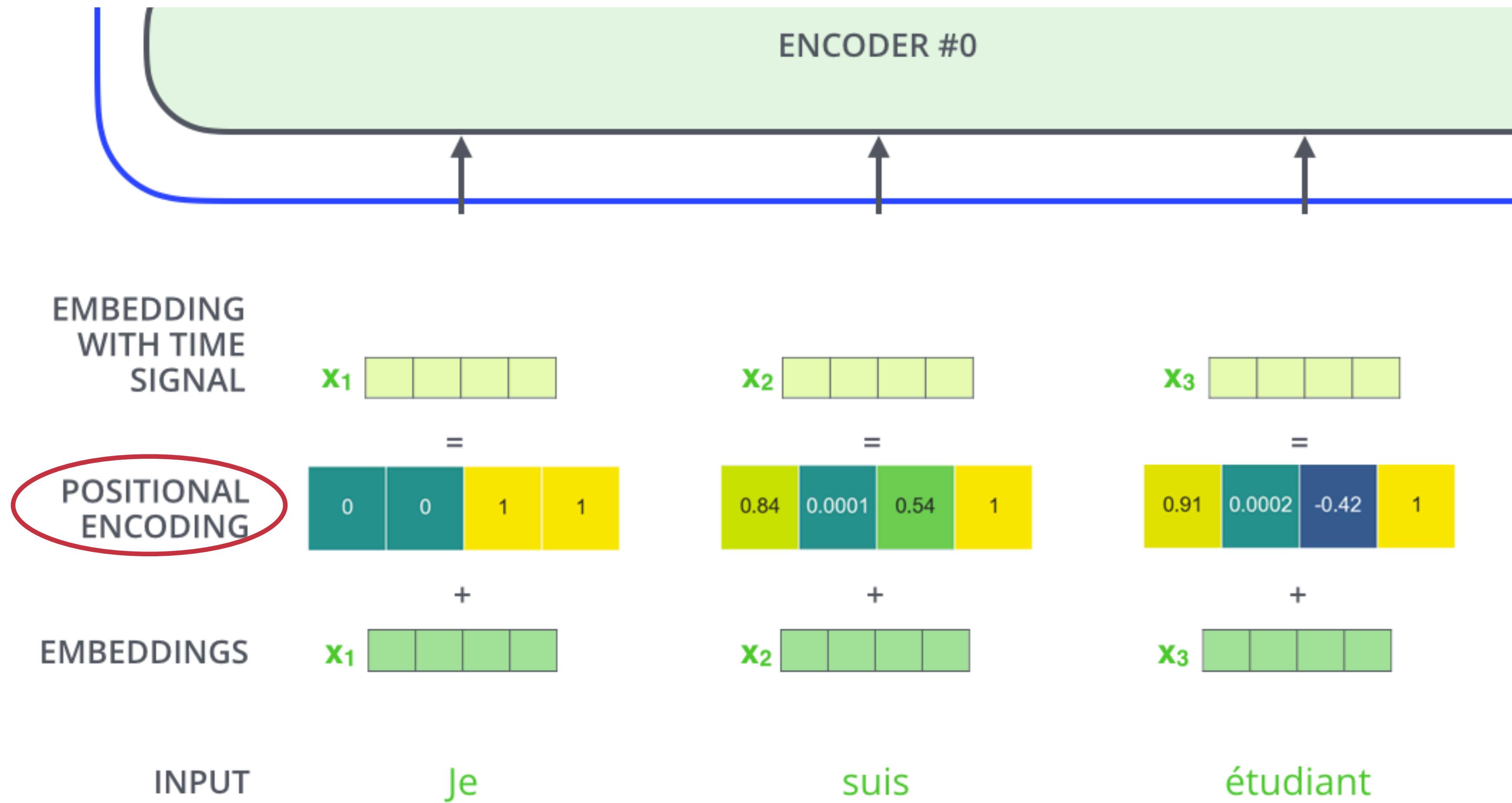
Что если перемешать вход?

Thinking Machines → Machines Thinking

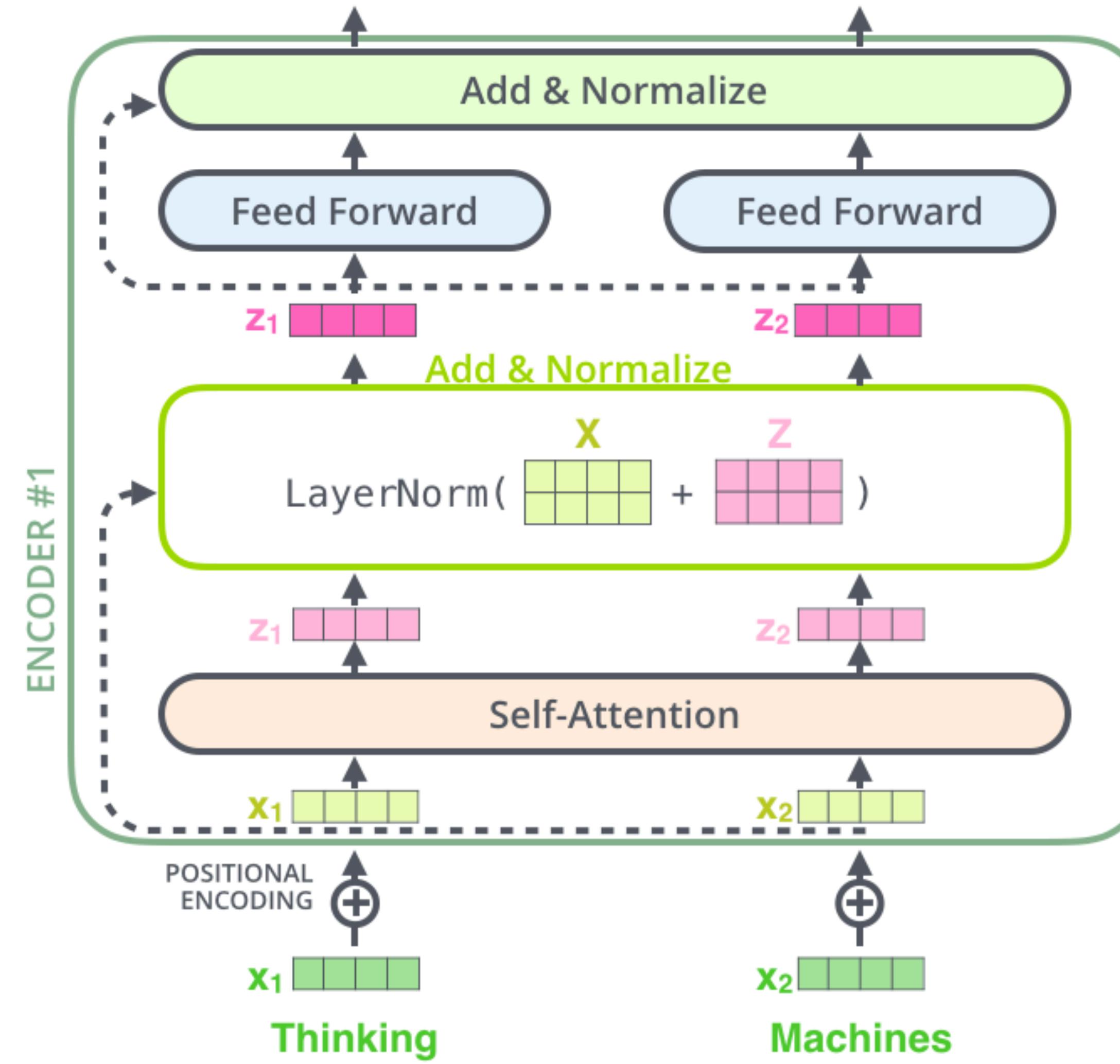
Transformer: positional embeddings



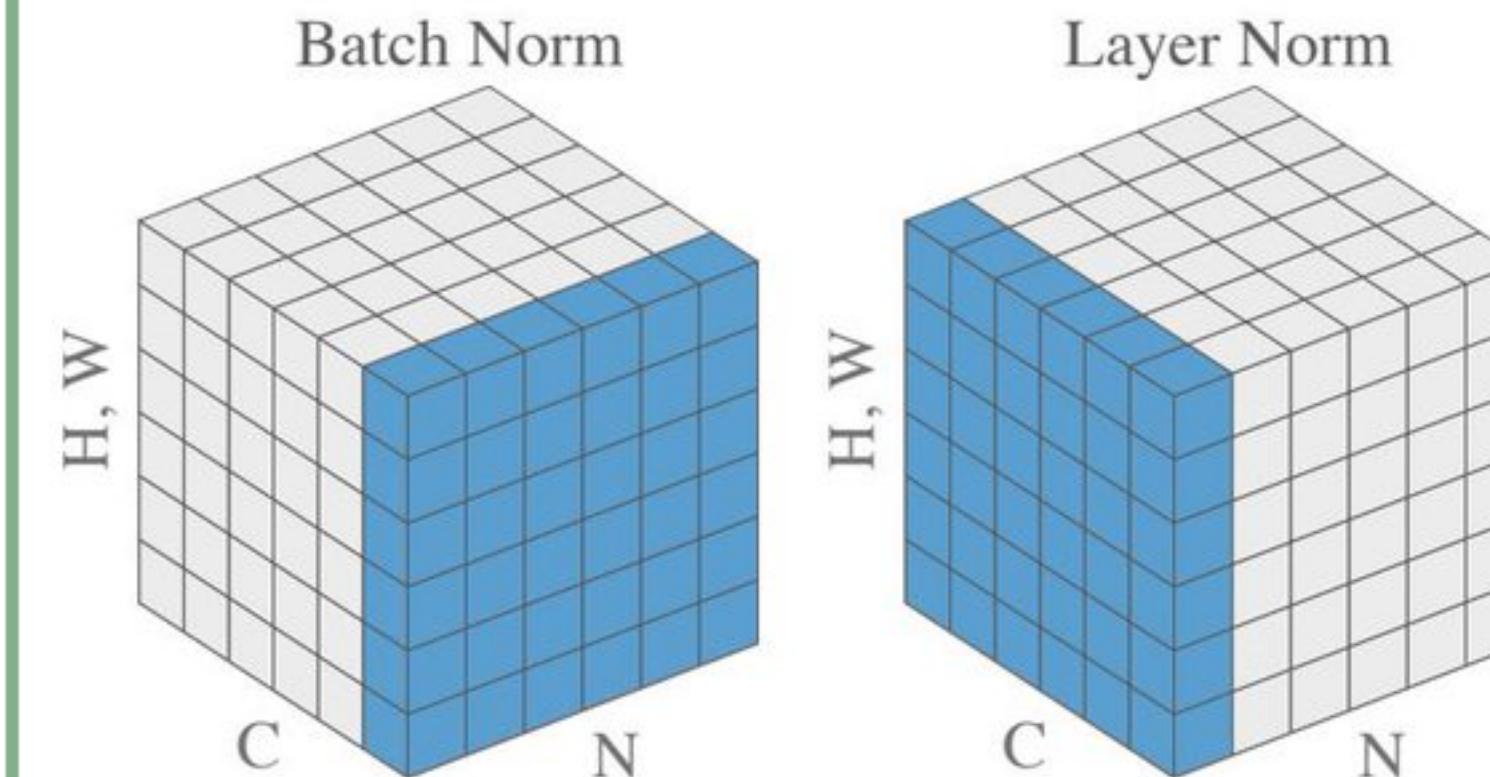
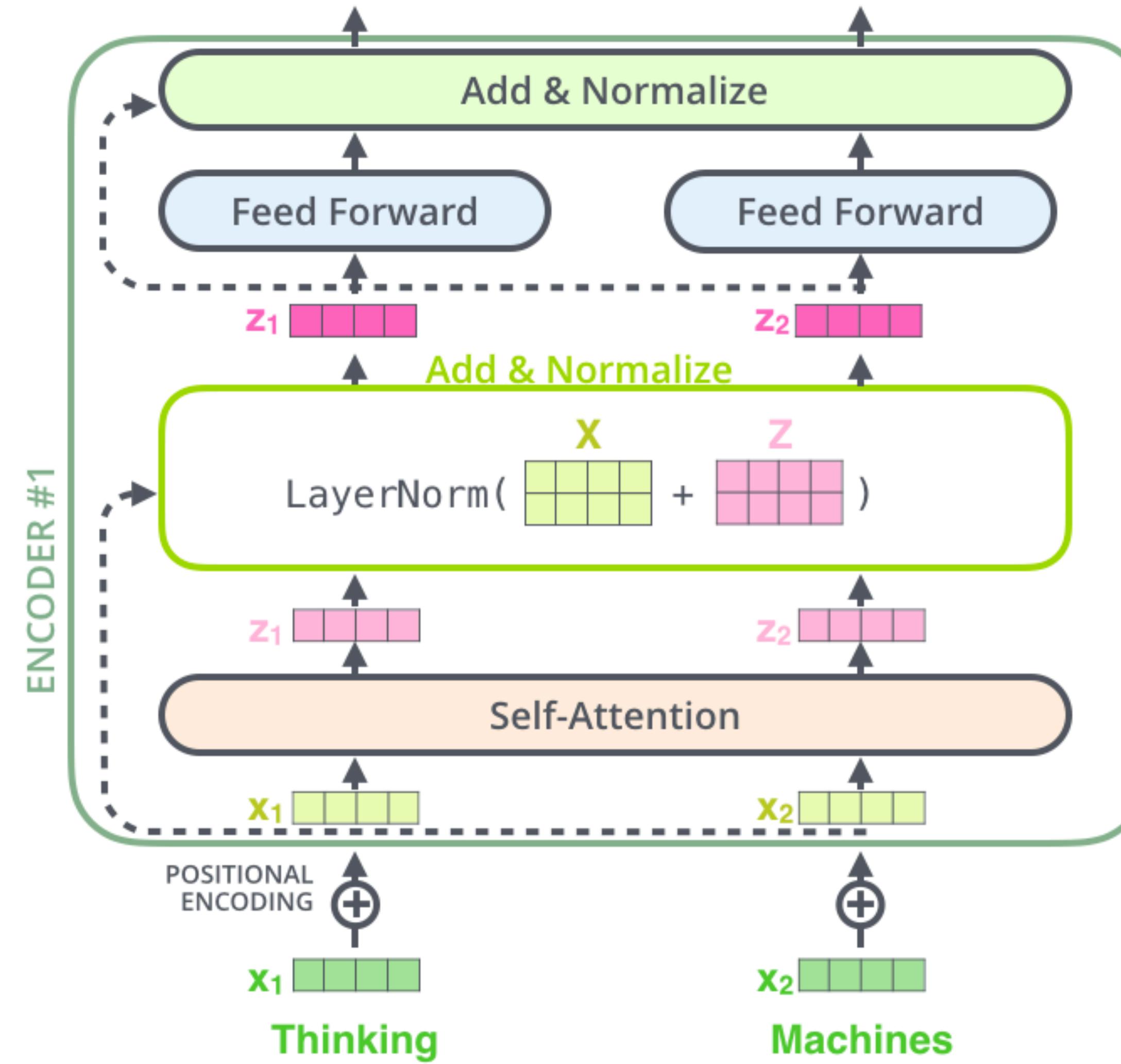
Transformer: positional embeddings



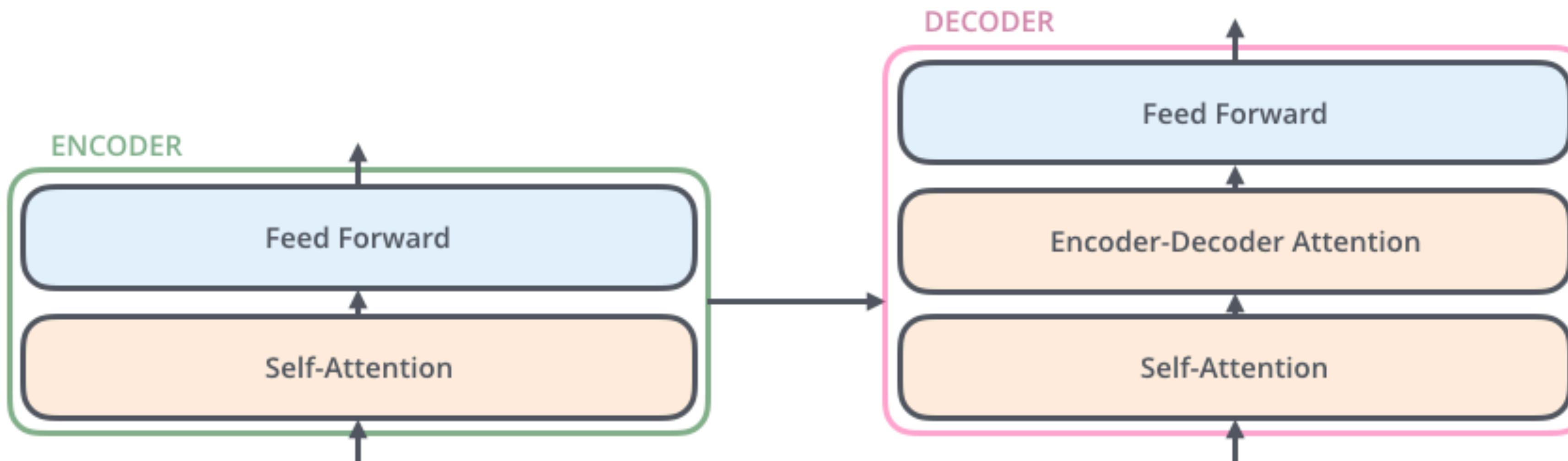
Transformer: Residuals & LayerNorm



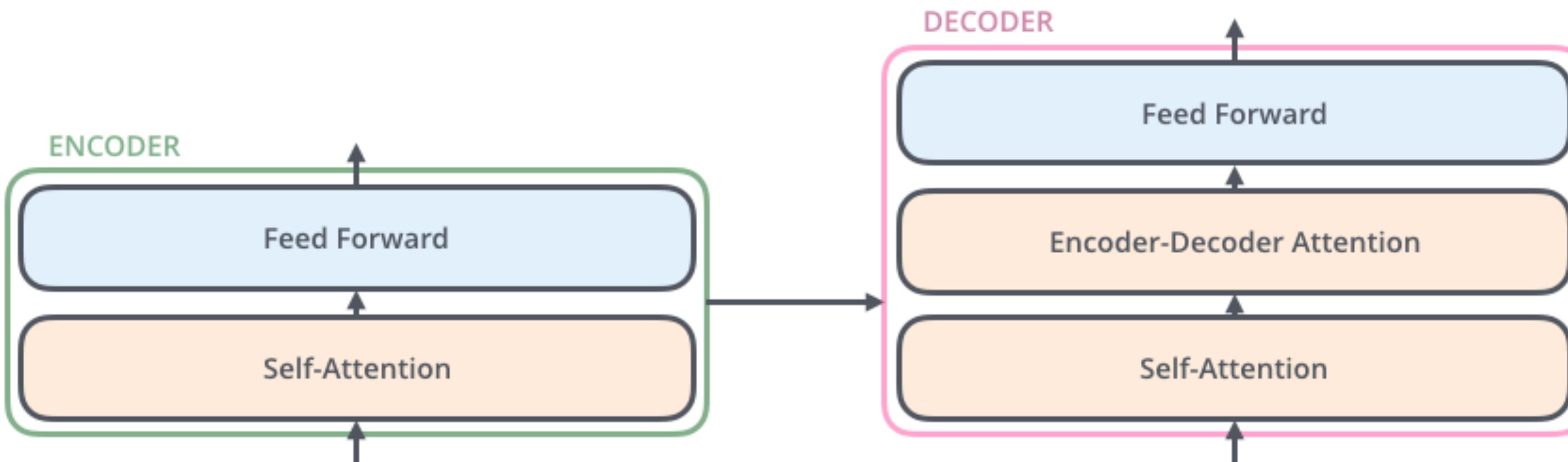
Transformer: Residuals & LayerNorm



Transformer: encoder-decoder attention

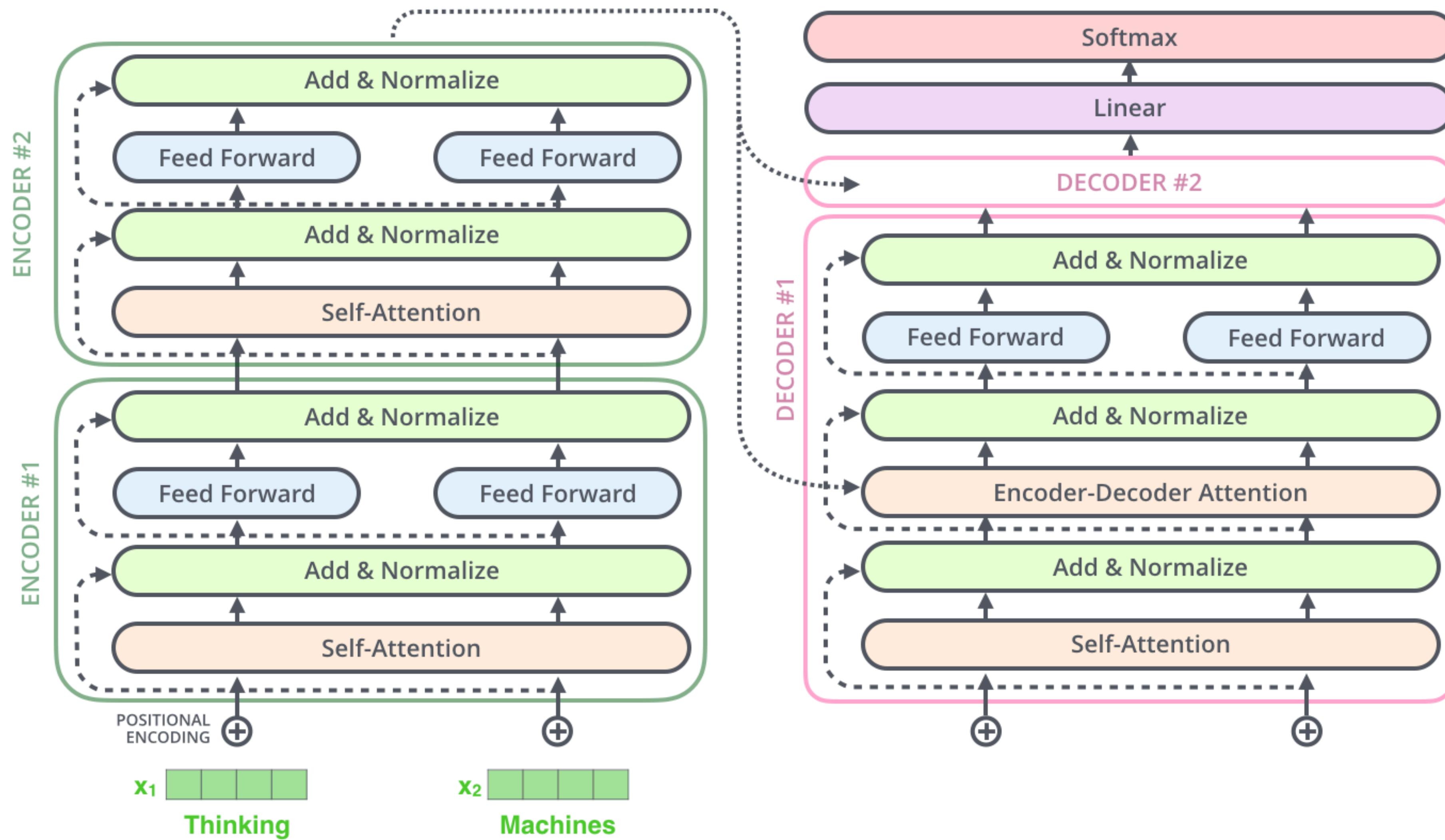


Transformer: encoder-decoder attention



$$\text{Enc-Dec attention} = \text{softmax}\left(\frac{Q_{decoder}K_{encoder}^T}{d}\right)V_{encoder}$$

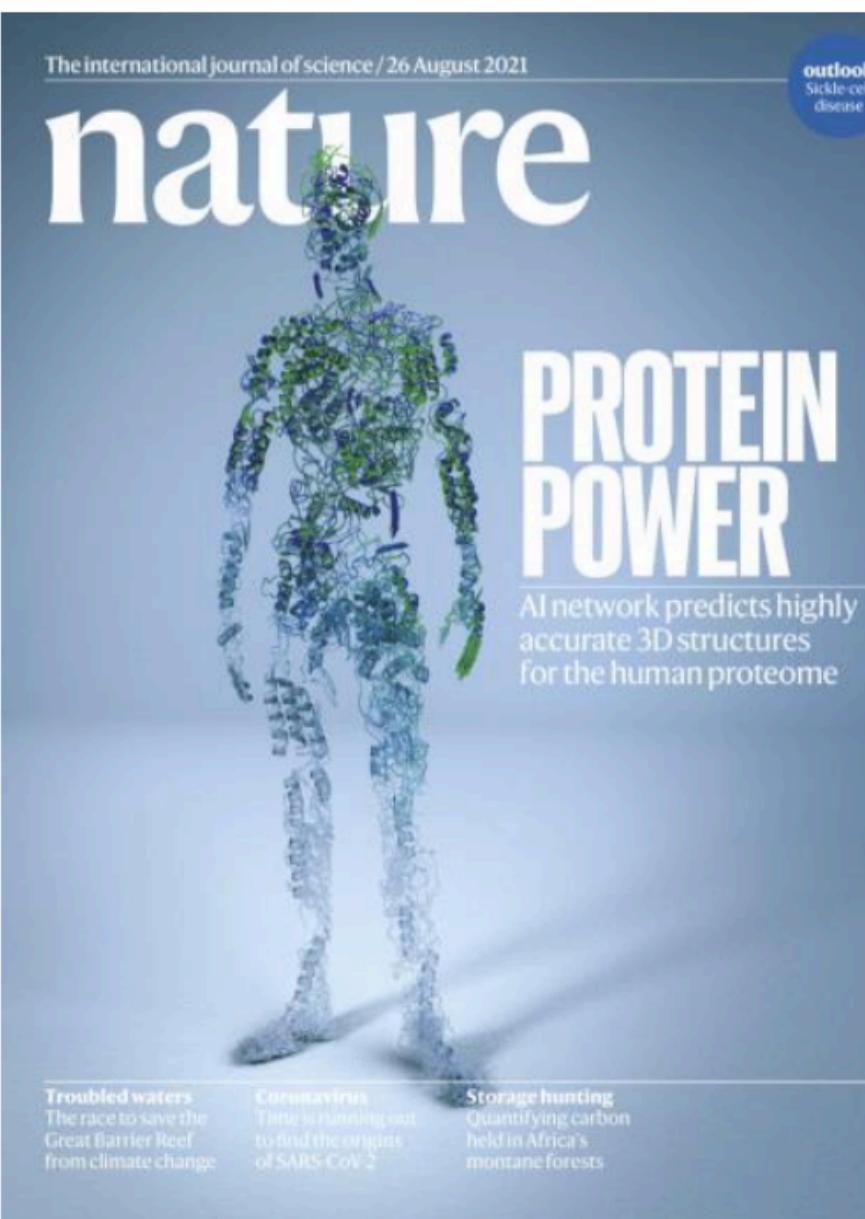
Transformer



Transformer

- BERT, GPT-1,2,3, etc. - **предобученные** трансформеры
- Трансформеры работают не только на текстах!

Protein Folding



[Jumper et al. 2021] aka AlphaFold2!

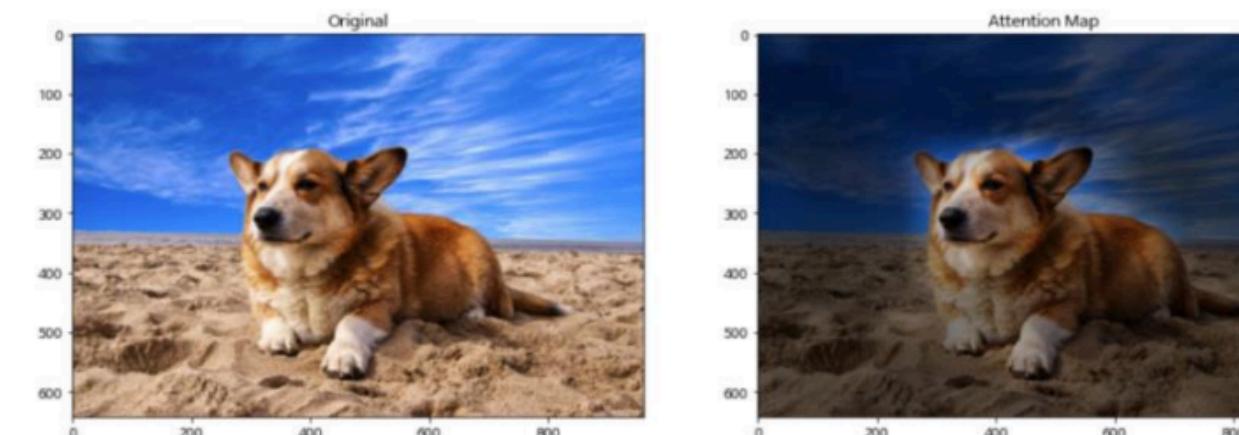
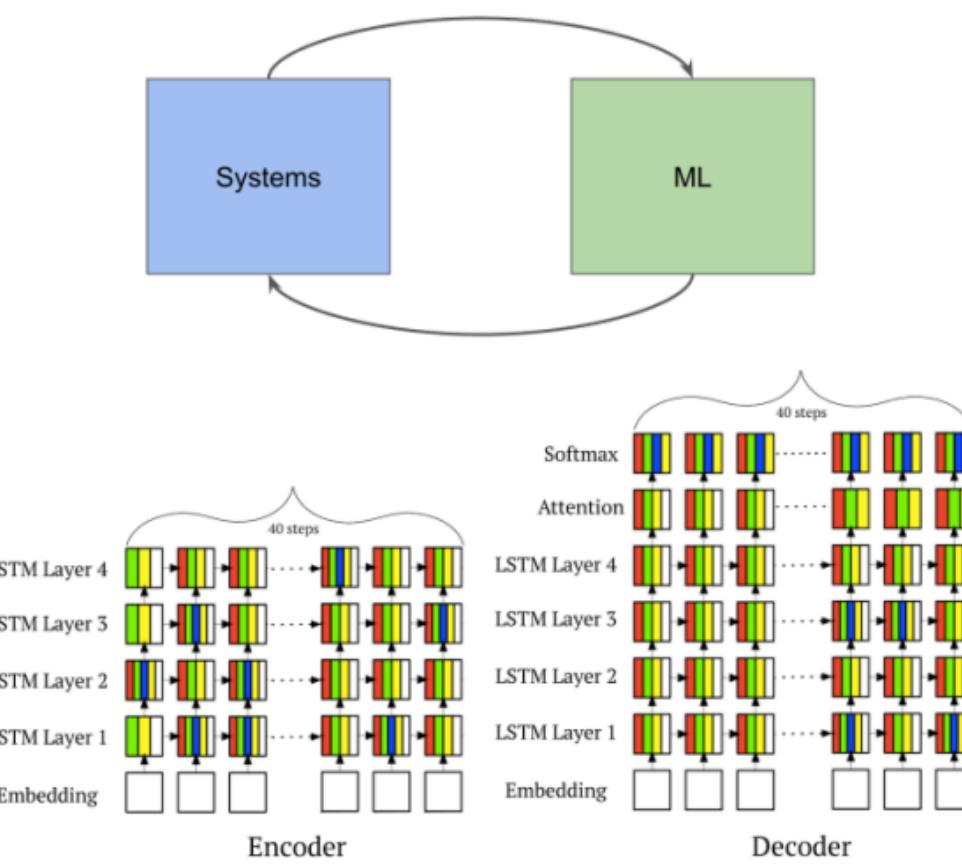


Image Classification

[Dosovitskiy et al. 2020]: Vision Transformer (ViT) outperforms ResNet-based baselines with substantially less compute.

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	88.55 ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4 / 88.5*
ImageNet ReaL	90.72 ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	99.50 ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	—
CIFAR-100	94.55 ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	—
Oxford-IIIT Pets	97.56 ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	—
Oxford Flowers-102	99.68 ± 0.02	99.74 ± 0.00	99.61 ± 0.02	99.63 ± 0.03	—
VTAB (19 tasks)	77.63 ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	—
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k



ML for Systems

[Zhou et al. 2020]: A Transformer-based compiler model (GO-one) speeds up a Transformer model!

Model (#devices)	GO-one (s)	HP (s)	METIS (s)	HDP (s)	Run time speed up over HP / HDP	Search speed up over HDP
2-layer RNNLM (2)	0.173	0.192	0.355	0.191	9.9% / 9.4%	2.9x
4-layer RNNLM (4)	0.210	0.239	0.503	0.251	13.8% / 16.3%	1.76x
8-layer RNNLM (8)	0.320	0.332	OOM	0.764	3.8% / 58.1%	27.8x
2-layer GNMT (2)	0.301	0.384	0.344	0.327	27.6% / 14.3%	30x
4-layer GNMT (4)	0.350	0.469	0.466	0.432	34% / 23.4%	58.8x
8-layer GNMT (8)	0.440	0.562	OOM	0.693	21.7% / 36.5%	7.35x
2-layer Transformer-XL (2)	0.223	0.268	0.37	0.262	20.1% / 17.4%	40x
4-layer Transformer-XL (4)	0.230	0.27	OOM	0.259	17.4% / 12.6%	26.7x
8-layer Transformer-XL (8)	0.350	0.46	OOM	0.425	23.9% / 16.7%	16.7x
Inception v3 (2)	0.229	0.312	OOM	0.301	26.6% / 23.9%	13.5x
Inception (2) b64	0.423	0.731	OOM	0.498	42.1% / 29.3%	21.0x
AmoebaNet (4)	0.394	0.44	0.426	0.418	26.1% / 6.1%	58.8x
2-stack 18-layer WaveNet (2)	0.317	0.376	OOM	0.354	18.6% / 11.7%	6.67x
4-stack 36-layer WaveNet (4)	0.659	0.988	OOM	0.721	50% / 9.4%	20x
GEOMEAN	-	-	-	-	20.5% / 18.2%	15x