TRANSFORMERS & THEIR OFFSPRINGS

Андреева Дарья
Data Scientist, X5 Tech

AI Run

attention is...

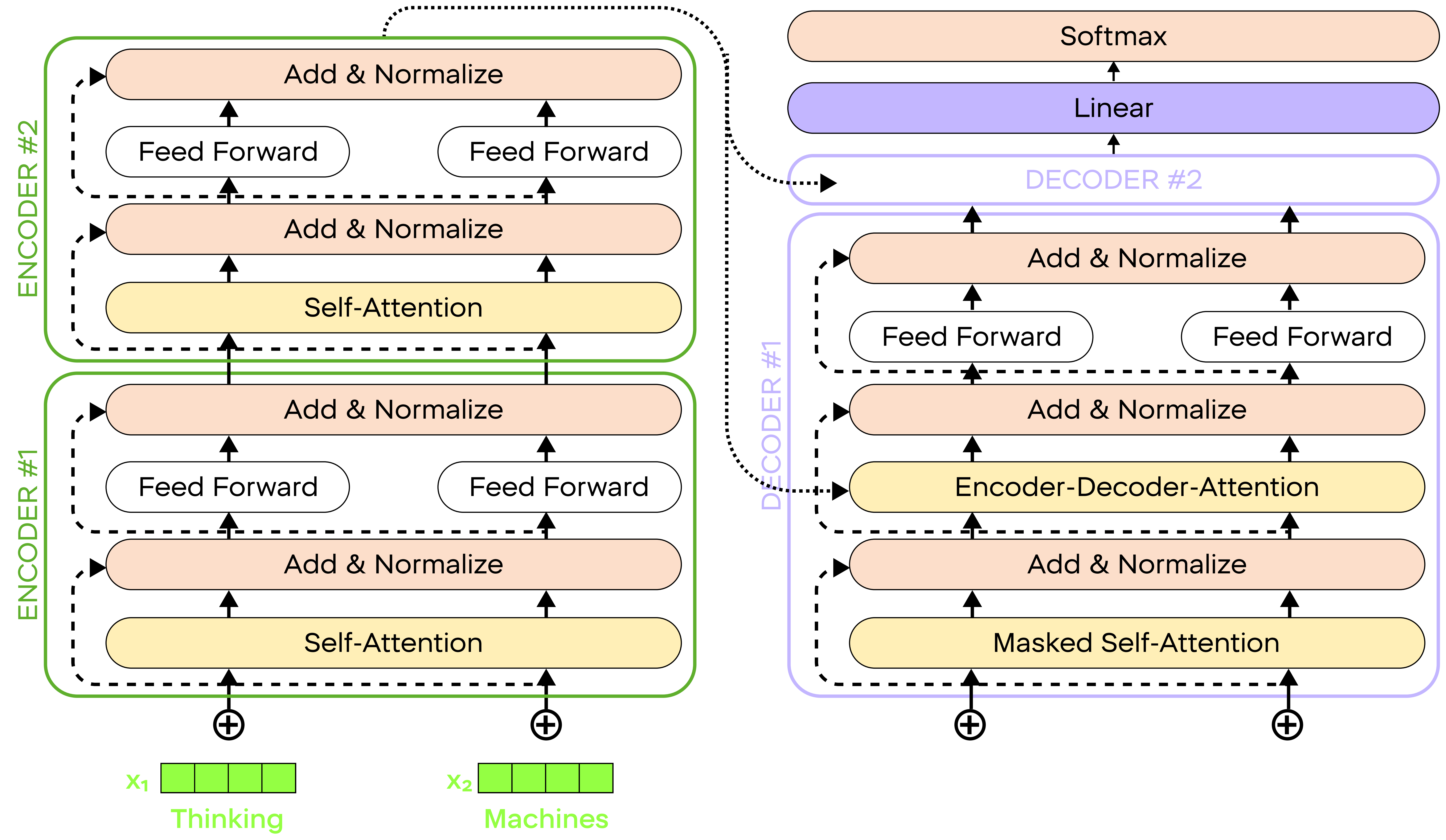
$$Attention(Q, V, K) = Softmax(\frac{QK^T}{\sqrt{d}})V$$

возможность отловить более сложные зависимости и концепции

позиционные эмбеддинги

чуть удобнее параллелить вычисления

attention is...



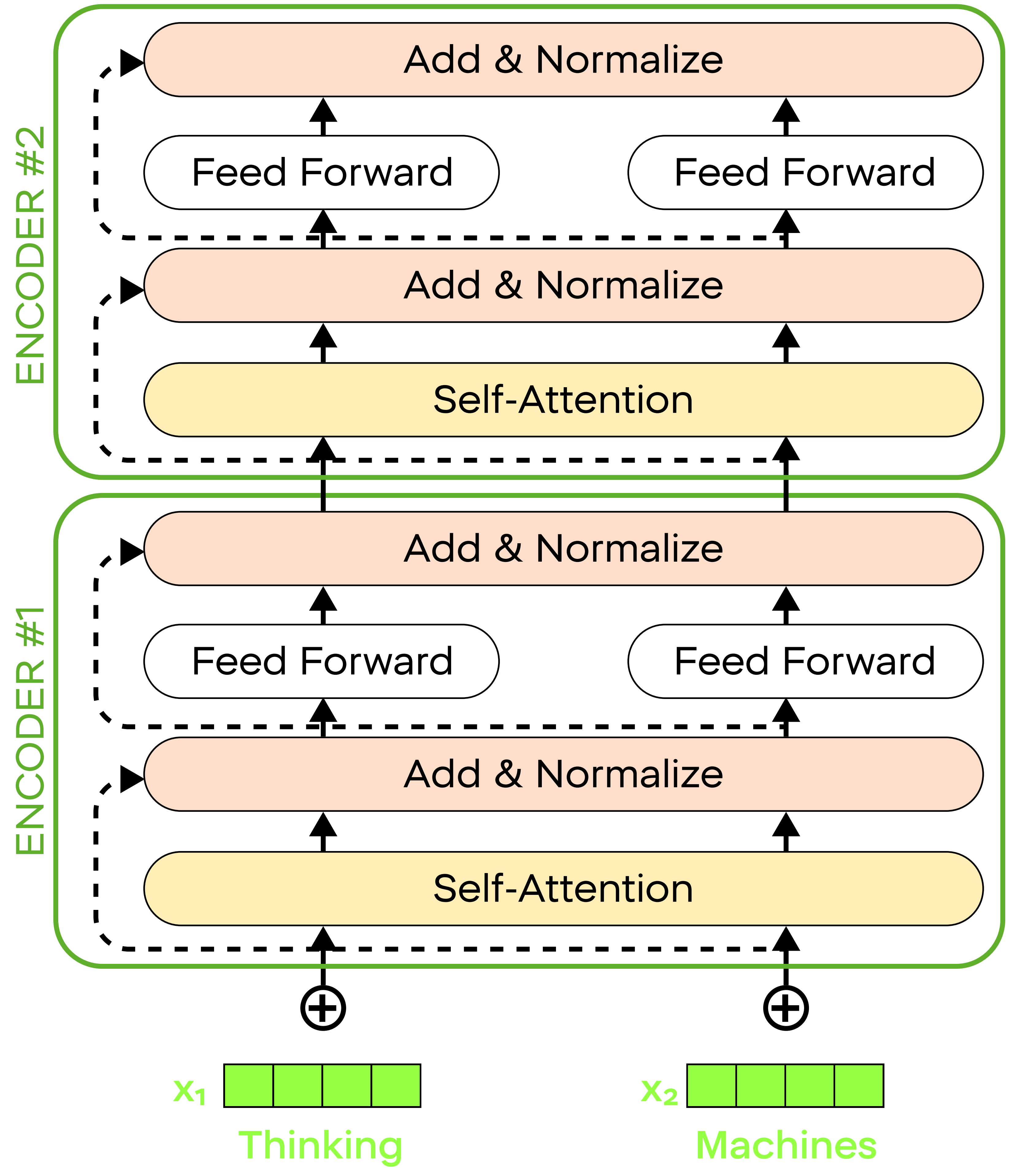
X5 Tech

BERT

BIDIRECTIONAL
ENCODER
REPRESENTATIONS
FROM
TRANSFORMERS

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding: https://arxiv.org/abs/1810.04805

bert

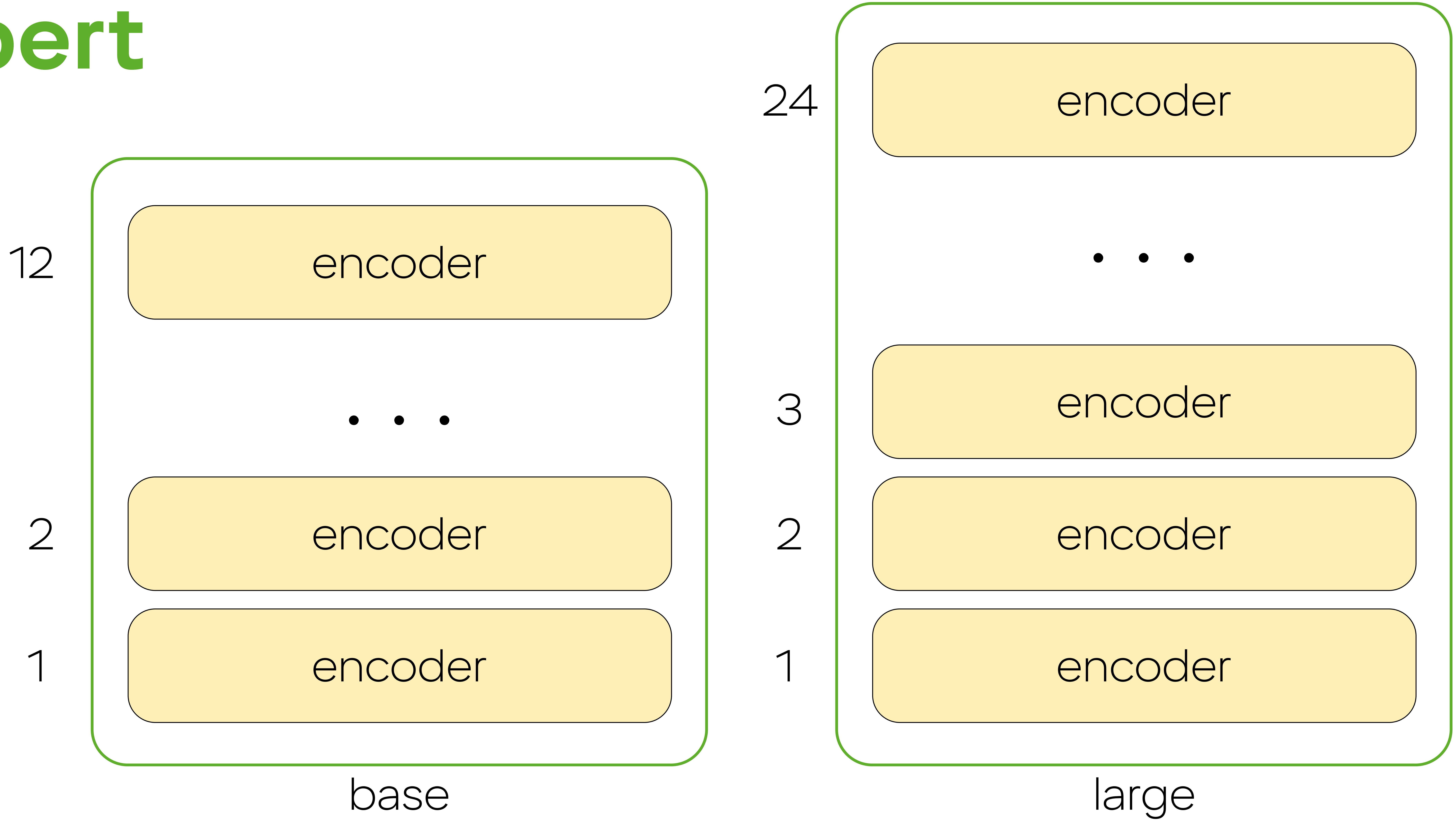


Ключевые отличия:

- Два направления (использует контекст и справа, и слева)
- Состоит только из энкодеров

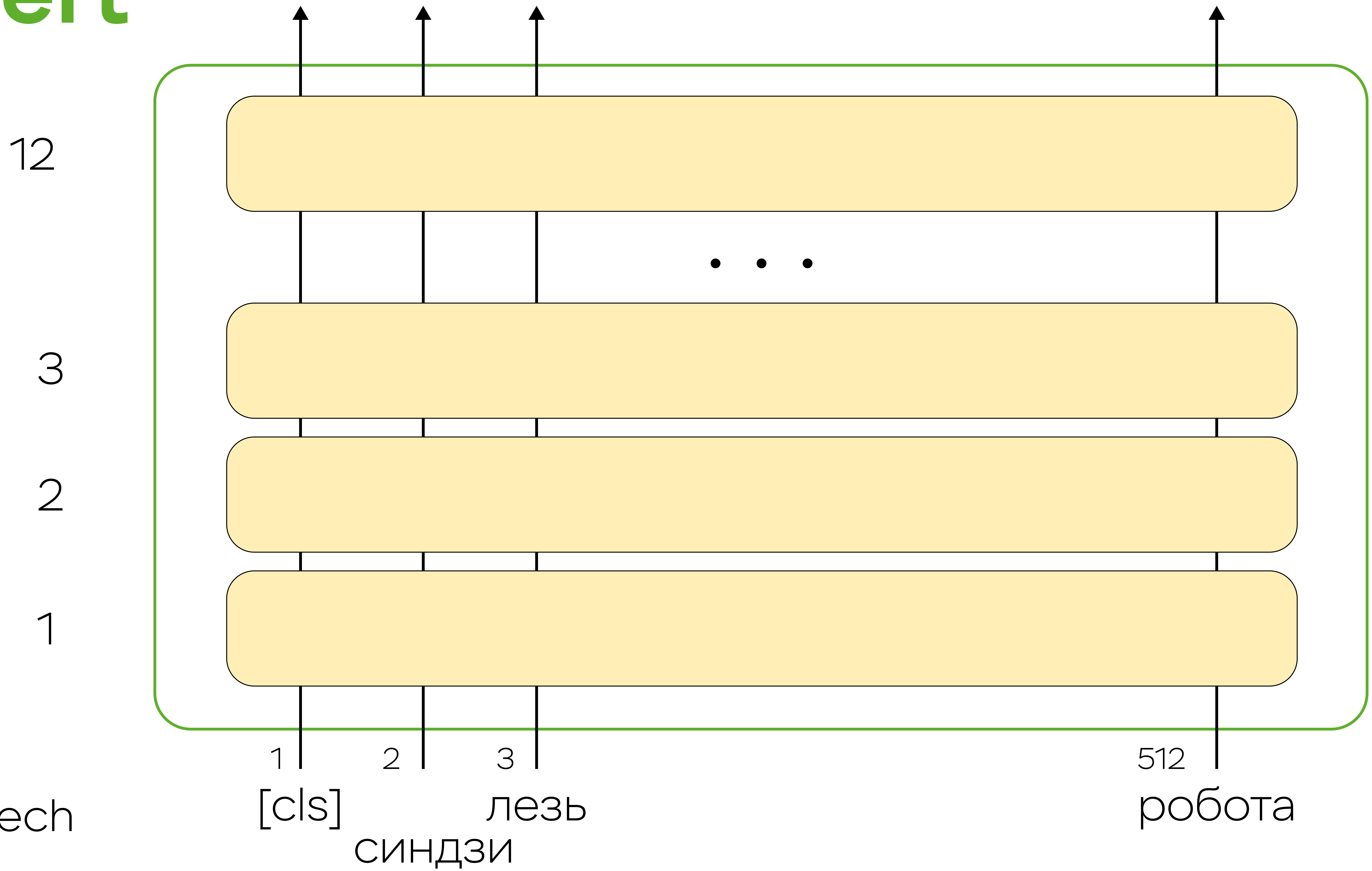
X5 Tech

bert


12
encoder
2
encoder
1
encoder
base
24
encoder
3
encoder
2
encoder
1
encoder
large

X5 Tech

bert

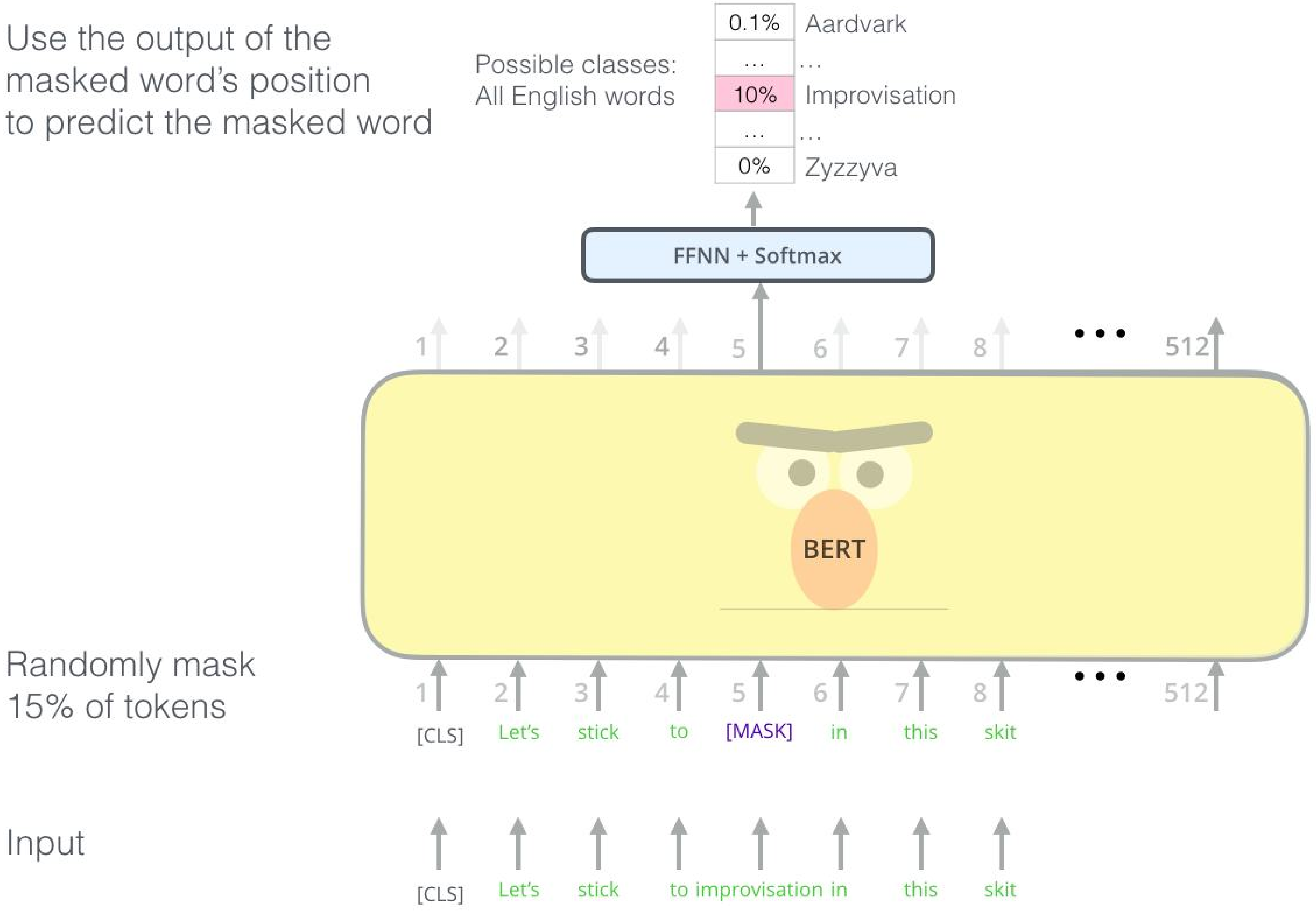


X5 Tech

bert

два подхода к обучению:

1. mlm: masked language model

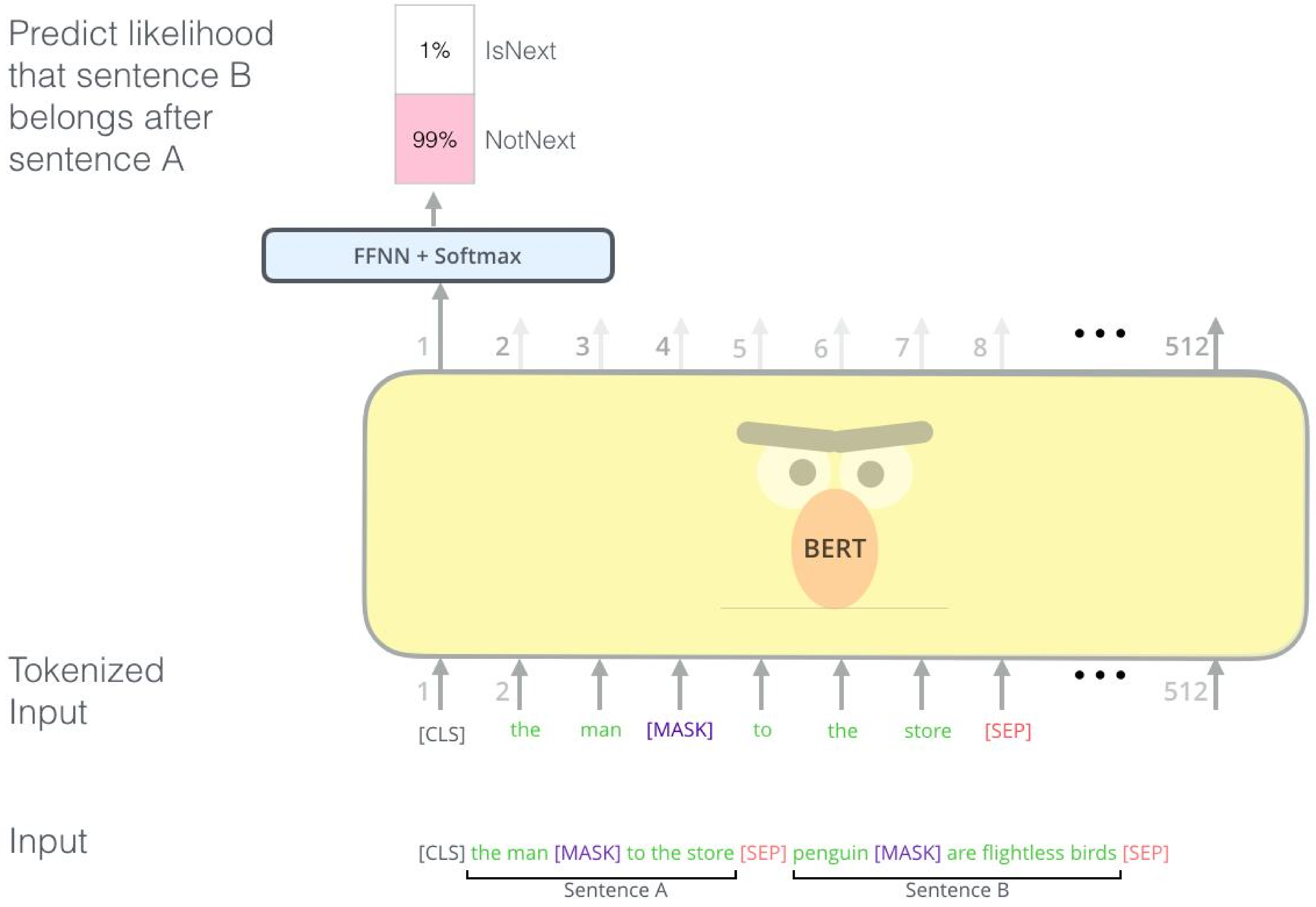


X5 Tech

bert: train

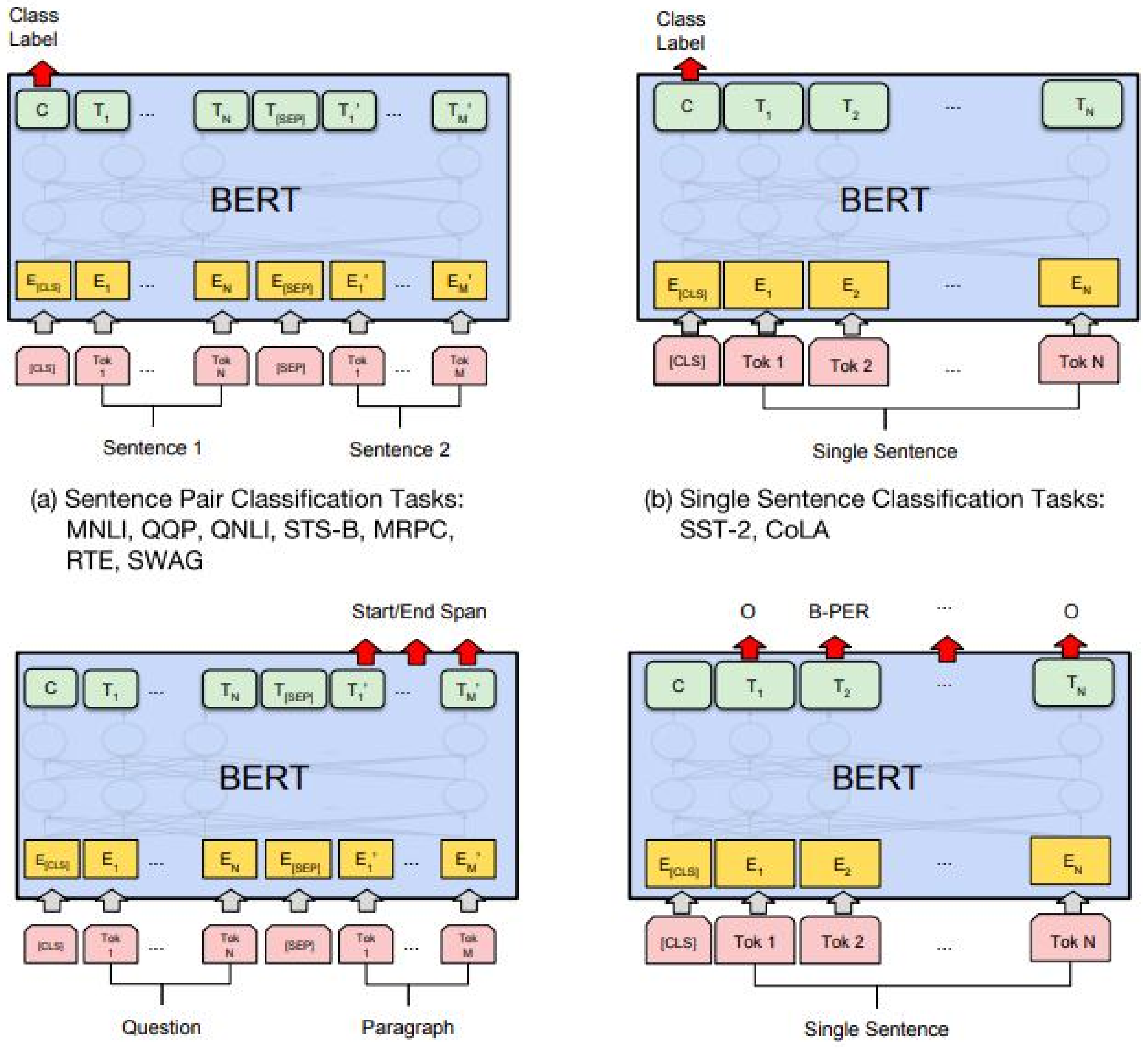
два подхода к обучению:

2. two-sentence task



X5 Tech

bert: различные задачи



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

WORD PIECE —

$$score = \frac{freq\_of\_pair[w_1, w_2]}{freq[w_1] * freq[w_2]}$$

“мама мыла раму, маму мыла рама”

1. vocab = [“м”, “а”, “р”, “у”...]
2. vocab += [“ма”, “му”, ...] пары с наивысшим скором
3. vocab.size == n: stop

X5 Tech

BERT:
THE OFFSPRINGS

X5 Tech

DISTILBERT

$$L(X,W) = \alpha \cdot H(y, \sigma(z_s, T=1)) + \beta \cdot H(\sigma(z_t, T=\tau), \sigma(z_s, T=\tau))$$

1. Запустим две модели: исходную и меньшую
2. При обучении меньшей будем использовать выходы большей модели и учить меньшую их повторить

X5 Tech

DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter: https://arxiv.org/abs/1910.01108

distilbert

Table 1: **DistilBERT retains 97% of BERT performance.** Comparison on the dev sets of the GLUE benchmark. ELMo results as reported by the authors. BERT and DistilBERT results are the medians of 5 runs with different seeds.

Model	**Score**	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI
ELMo	68.7	44.1	68.6	76.6	71.1	86.2	53.4	91.5	70.4	56.3
BERT-base	79.5	56.3	86.7	88.6	91.8	89.6	69.3	92.7	89.0	53.5
DistilBERT	77.0	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9	56.3

Table 2: **DistilBERT yields to comparable performance on downstream tasks.** Comparison on downstream tasks: IMDb (test accuracy) and SQuAD 1.1 (EM/F1 on dev set). D: with a second step of distillation during fine-tuning.

Model	IMDb (acc.)	SQuAD (EM/F1)
BERT-base	93.46	81.2/88.5
DistilBERT	92.82	77.7/85.8
DistilBERT (D)	-	79.1/86.9

Table 3: **DistilBERT is significantly smaller while being constantly faster.** Inference time of a full pass of GLUE task STS-B (sentiment analysis) on CPU with a batch size of 1.

Model	# param. (Millions)	Inf. time (seconds)
ELMo	180	895
BERT-base	110	668
DistilBERT	66	410

ROBERTA

1. Теперь маскируем текст не на этапе препроцессинга, а прямо при обучении
2. Не обучаемся на предсказании следующего предложения

X5 Tech

RoBERTa: A Robustly Optimized BERT Pretraining Approach: https://arxiv.org/abs/1907.11692

roberta

Model	SQuAD 1.1/2.0	MNLI-m	SST-2	RACE
Our reimplementation (with NSP loss):				
SEGMENT-PAIR	90.4/78.7	84.0	92.9	64.2
SENTENCE-PAIR	88.7/76.2	82.9	92.1	63.0
Our reimplementation (without NSP loss):				
FULL-SENTENCES	90.4/79.1	84.7	92.5	64.8
DOC-SENTENCES	90.6/79.7	84.7	92.7	65.6
$\text{BERT}_{\text{BASE}}$	88.5/76.3	84.3	92.8	64.3
$\text{XLNet}_{\text{BASE}}$ (K = 7)	–/81.3	85.8	92.7	66.1
$\text{XLNet}_{\text{BASE}}$ (K = 6)	–/81.0	85.6	93.4	66.7

Table 2: Development set results for base models pretrained over BOOKCORPUS and WIKIPEDIA. All models are trained for 1M steps with a batch size of 256 sequences. We report F1 for SQuAD and accuracy for MNLI-m, SST-2 and RACE. Reported results are medians over five random initializations (seeds). Results for $\text{BERT}_{\text{BASE}}$ and $\text{XLNet}_{\text{BASE}}$ are from Yang et al. (2019).

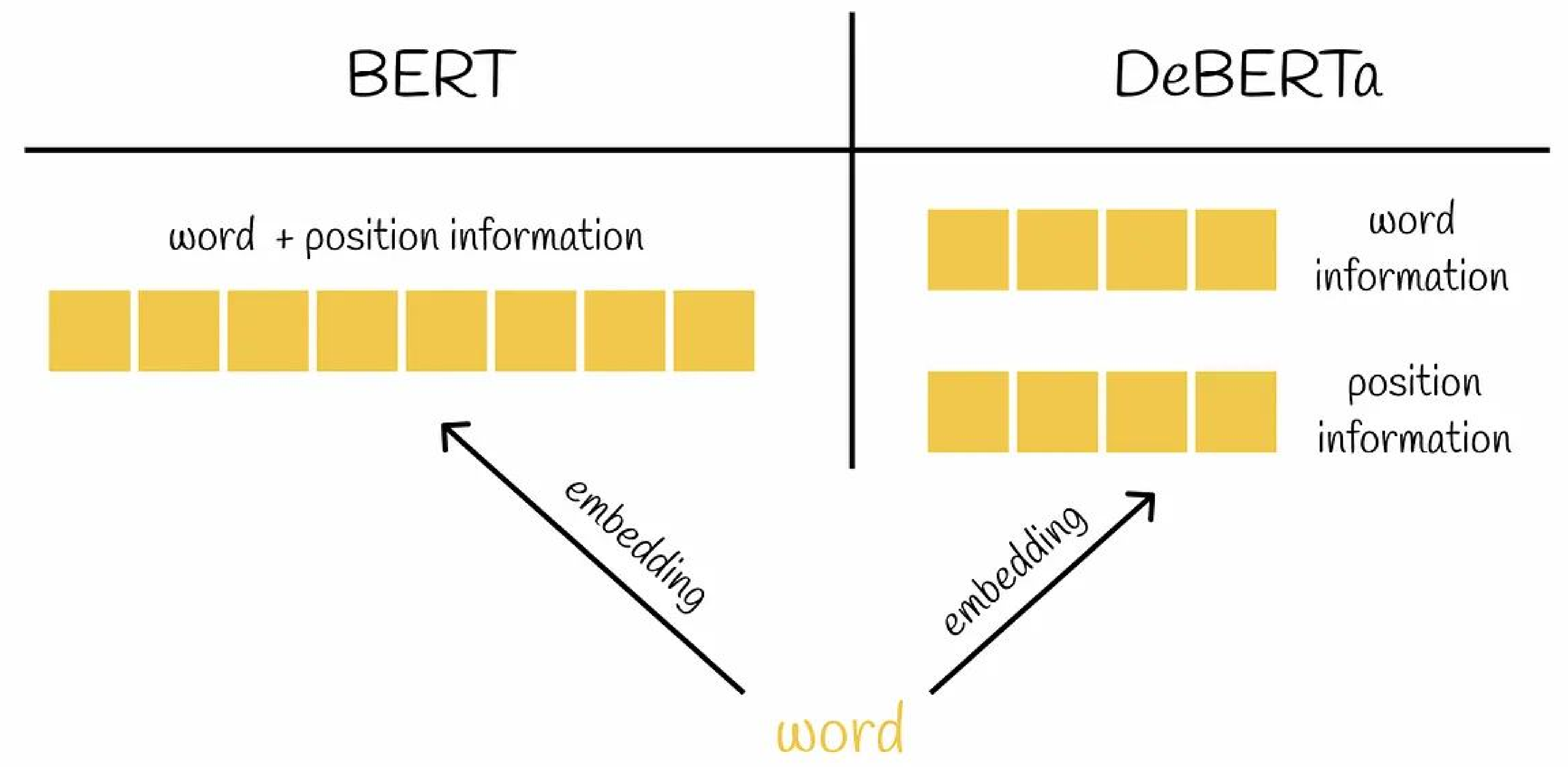
roberta

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7
$\text{XLNet}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	94.0/87.8	88.4	94.4
+ additional data	126GB	2K	500K	94.5/88.8	89.8	95.6

Table 4: Development set results for RoBERTa as we pretrain over more data (16GB → 160GB of text) and pretrain for longer (100K → 300K → 500K steps). Each row accumulates improvements from the rows above. RoBERTa matches the architecture and training objective of $\text{BERT}_{\text{LARGE}}$. Results for $\text{BERT}_{\text{LARGE}}$ and $\text{XLNet}_{\text{LARGE}}$ are from Devlin et al. (2019) and Yang et al. (2019), respectively. Complete results on all GLUE tasks can be found in the Appendix.

X5 Tech

DEBERTA



1. Вместо одного эмбеддинга для слова теперь два: отдельно позиционный, отдельно слово
2. Немного изменена архитектура
3. Считывает относительную позицию, а не абсолютную

deberta

Model	MNLI-m/mm Acc	SQuAD v1.1 F1/EM	SQuAD v2.0 F1/EM	RACE Acc	ReCoRD F1/EM	SWAG Acc	NER F1
BERT_{large}	86.6/-	90.9/84.1	81.8/79.0	72.0	-	86.6	92.8
ALBERT_{large}	86.5/-	91.8/85.2	84.9/81.8	75.2	-	-	-
RoBERTa_{large}	90.2/90.2	94.6/88.9	89.4/86.5	83.2	90.6/90.0	89.9	93.4
XLNet_{large}	90.8/90.8	95.1/89.7	90.6/87.9	85.4	-	-	-
Megatron_{336M}	89.7/90.0	94.2/88.0	88.1/84.8	83.0	-	-	-
DeBERTa_{large}	**91.1/91.1**	**95.5/90.1**	**90.7/88.0**	**86.8**	**91.4/91.0**	**90.8**	**93.8**
$\text{ALBERT}_{xxlarge}$	90.8/-	94.8/89.3	90.2/87.4	86.5	-	-	-
$\text{Megatron}_{1.3B}$	90.9/91.0	94.9/89.1	90.2/87.1	87.3	-	-	-
$\text{Megatron}_{3.9B}$	91.4/91.4	95.5/90.0	91.2/88.5	89.5	-	-	-

Table 2: Results on MNLI in/out-domain, SQuAD v1.1, SQuAD v2.0, RACE, ReCoRD, SWAG, CoNLL 2003 NER development set. Note that missing results in literature are signified by "-".