

Embedding

Context-aware embeddings. ELMo, BERT

Churakov Igor

Мотивация

Значение зависит от контекста

- He kicked the **bucket**.
- I have yet to cross-off all the items on my **bucket** list.
- The **bucket** was filled with water.

I arrived at the **bank** after crossing the ... (...street? ...river?)

ELMo

Embeddings from Language Models



Language Model

Пусть имеется последовательность из N токенов (t_1, t_2, \dots, t_N)

Прямая языковая модель максимизирует

$$p(t_1, t_2, \dots, t_N) = \prod_{k=1}^N p(t_k | t_1, t_2, \dots, t_{k-1})$$

Language Model

Пусть имеется последовательность из N токенов (t_1, t_2, \dots, t_N)

Прямая языковая модель максимизирует

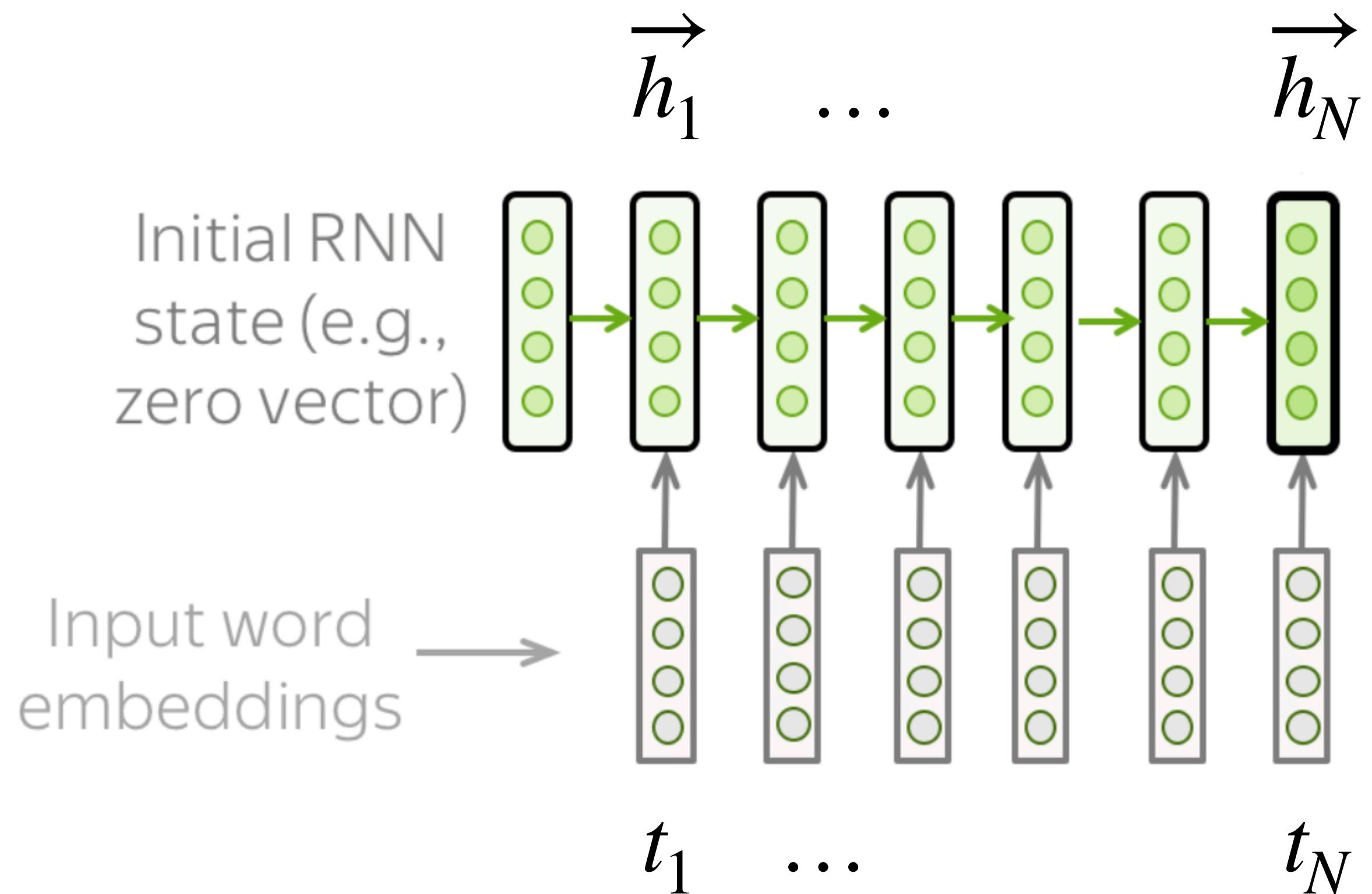
$$p(t_1, t_2, \dots, t_N) = \prod_{k=1}^N p(t_k | t_1, t_2, \dots, t_{k-1})$$

Обратная максимизирует

$$p(t_1, t_2, \dots, t_N) = \prod_{k=1}^N p(t_k | t_{k+1}, t_{k+2}, \dots, t_N)$$

Language Model

Для создания контекстно зависимых репрезентаций токенов воспользуемся уже знакомой нам **LSTM**



biLM

Аналогично получим $\overleftarrow{h}_1, \dots, \overleftarrow{h}_N$ – выходы противоположно направленной LSTM.

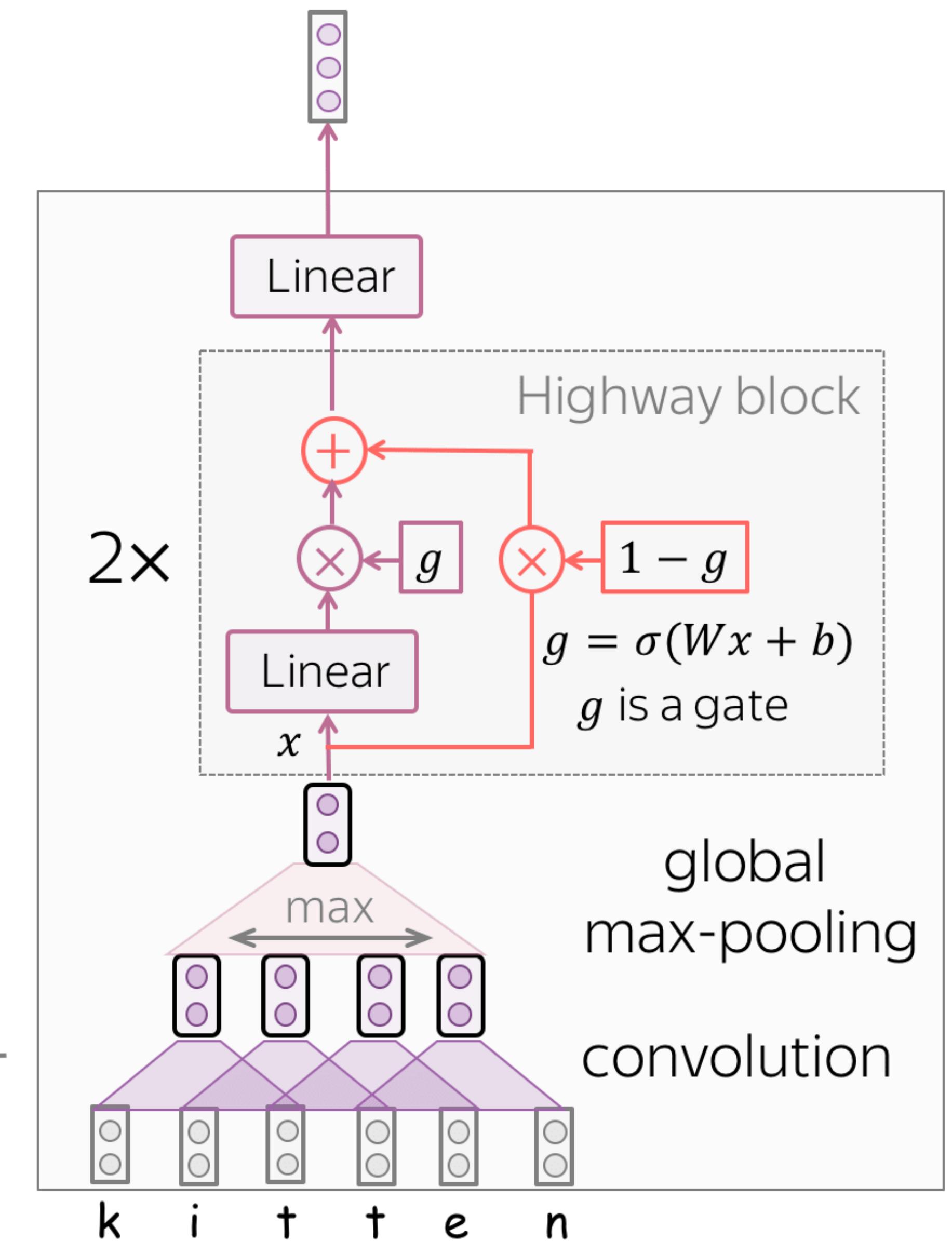
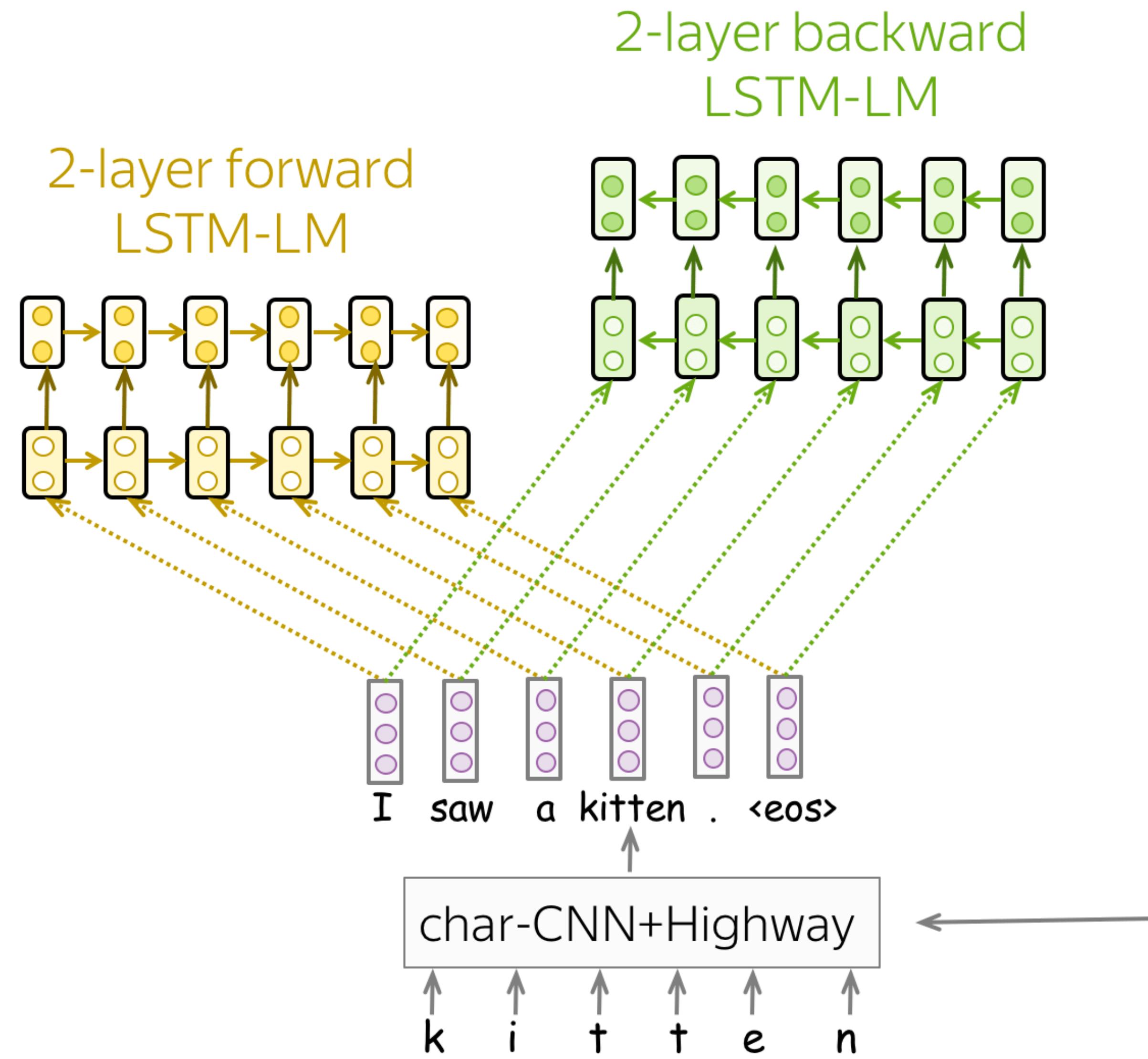
biLM

Аналогично получим $\overleftarrow{h}_1, \dots, \overleftarrow{h}_N$ – выходы противоположно направленной LSTM.

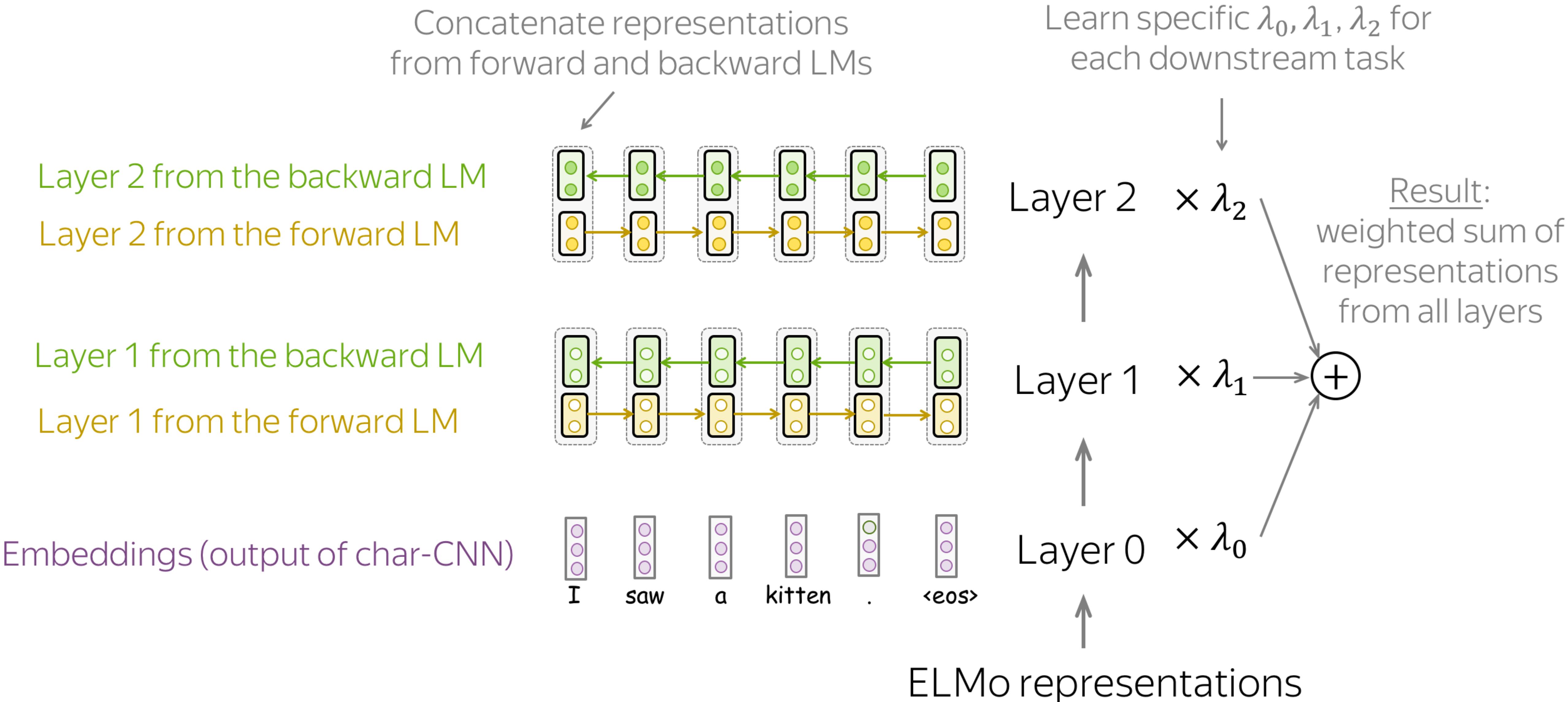
biLM совмещает в себе прямую и обратную языковые модели, мы хотим максимизировать следующий логарифм правдоподобия:

$$\sum_{k=1}^N (\log p(t_k | t_1, \dots, t_{k-1}; \Theta_x, \vec{\Theta}_{LSTM}, \Theta_s) + \log p(t_k | t_{k+1}, \dots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s))$$

Архитектура ELMo



Архитектура ELMo

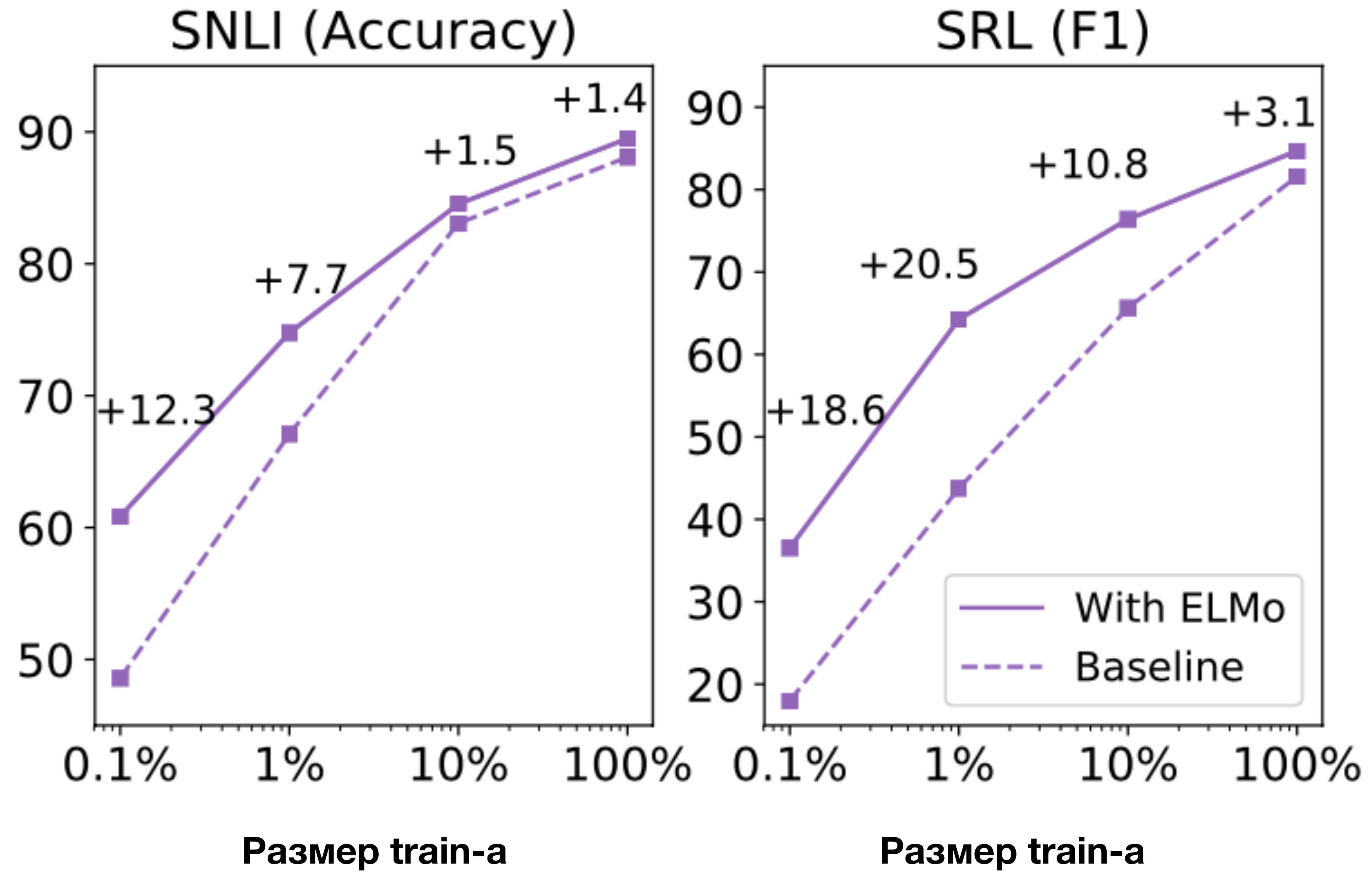


Результаты

TASK	PREVIOUS SOTA	OUR BASELINE	ELMO + BASELINE	INCREASE (ABSOLUTE/ RELATIVE)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17
SRL	He et al. (2017)	81.7	81.4	84.6
Coref	Lee et al. (2017)	67.2	67.2	70.4
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5

Результаты

Более эффективное использование данных



О получившихся эмбеддингах

Source	Nearest Neighbors
GloVe play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM Chico Ruiz made a spectacular <u>play</u> on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent <u>play</u> .
biLM Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful <u>play</u> , and had talent enough to fill the roles competently , with nice understatement .

Table 4: Nearest neighbors to “play” using GloVe and the context embeddings from a biLM.

О получившихся эмбеддингах

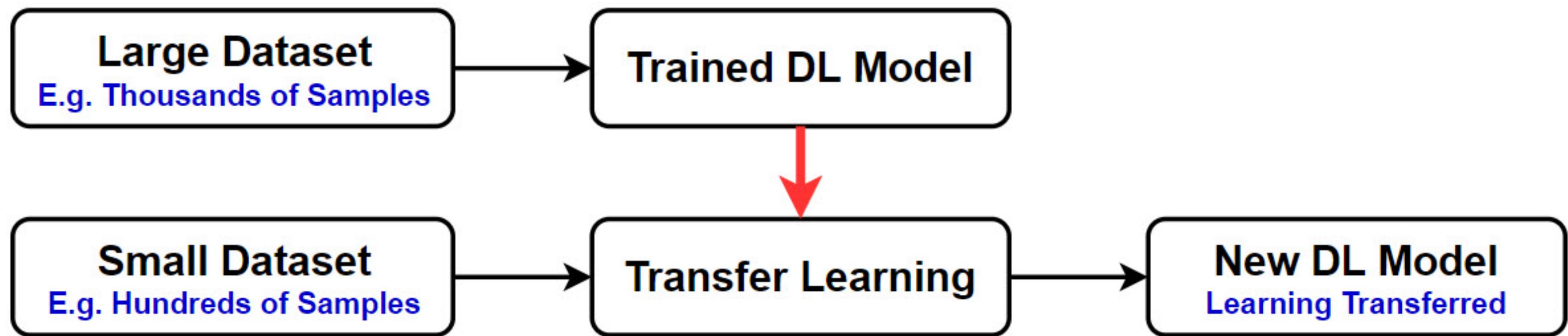
разрешение лексической
многозначности

Определение частей речи

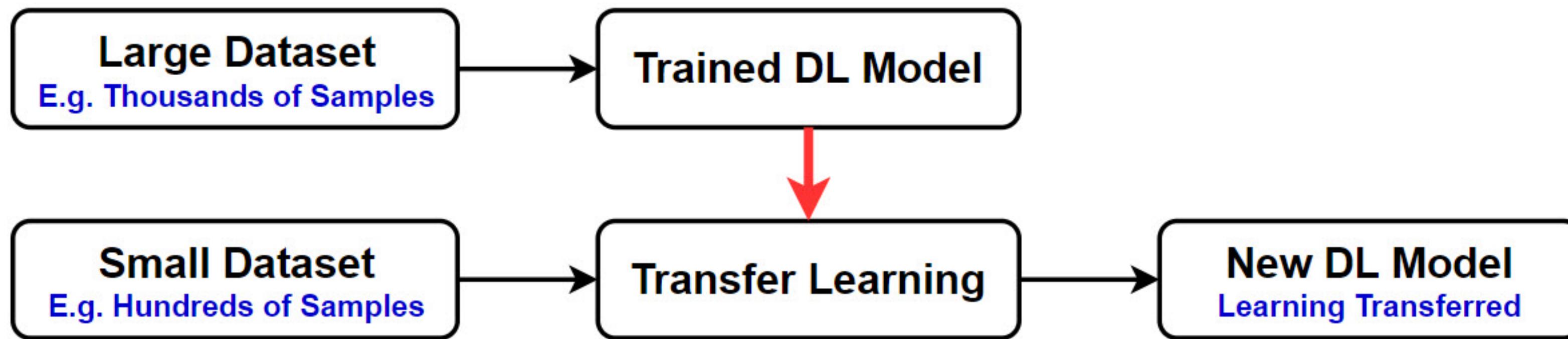
Model	F ₁
WordNet 1st Sense Baseline	65.9
Raganato et al. (2017a)	69.9
Iacobacci et al. (2016)	70.1
CoVe, First Layer	59.4
CoVe, Second Layer	64.7
biLM, First layer	67.4
biLM, Second layer	69.0

Model	Acc.
Collobert et al. (2011)	97.3
Ma and Hovy (2016)	97.6
Ling et al. (2015)	97.8
CoVe, First Layer	93.3
CoVe, Second Layer	92.8
biLM, First Layer	97.3
biLM, Second Layer	96.8

Transfer Learning в NLP задачах



Transfer Learning в NLP задачах



GloVe, Word2Vec → CoVe, ELMo → GPT, BERT

What is encoded

words → words in context

Great idea 1

How it is used for downstream tasks

Input for task-specific models

Input for task-specific models

words in context

Instead of task-specific models

Great idea 2

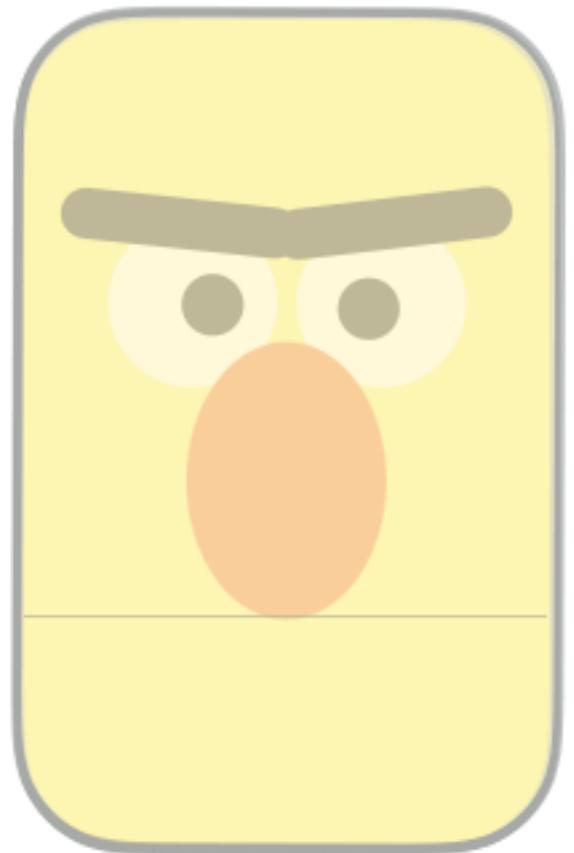
BERT

Bidirectional Encoder Representations from Transformers

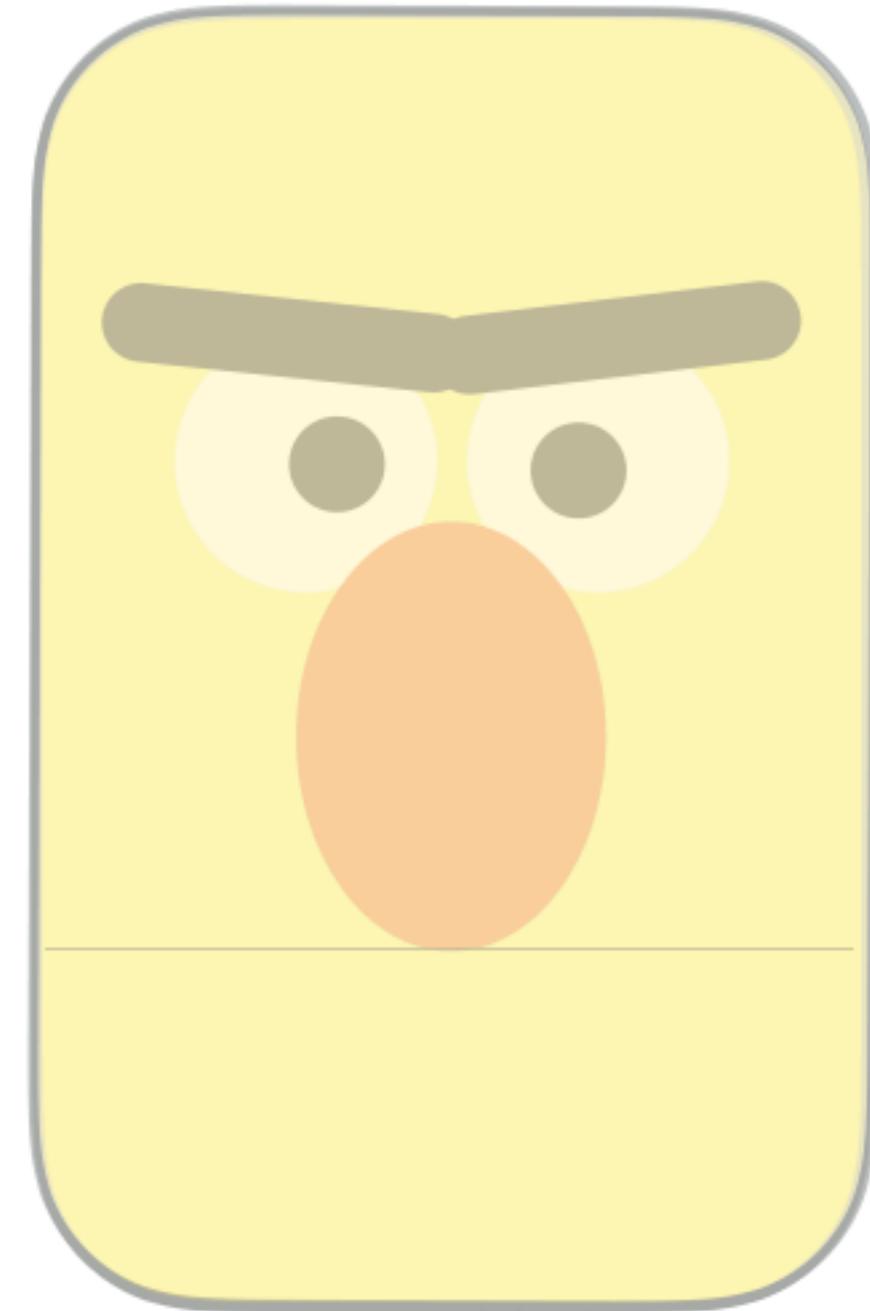


BERT

$BERT_{LARGE}(L = 24, H = 1024, A = 16, \text{Total Parameters} = 340M)$



$BERT_{BASE}$

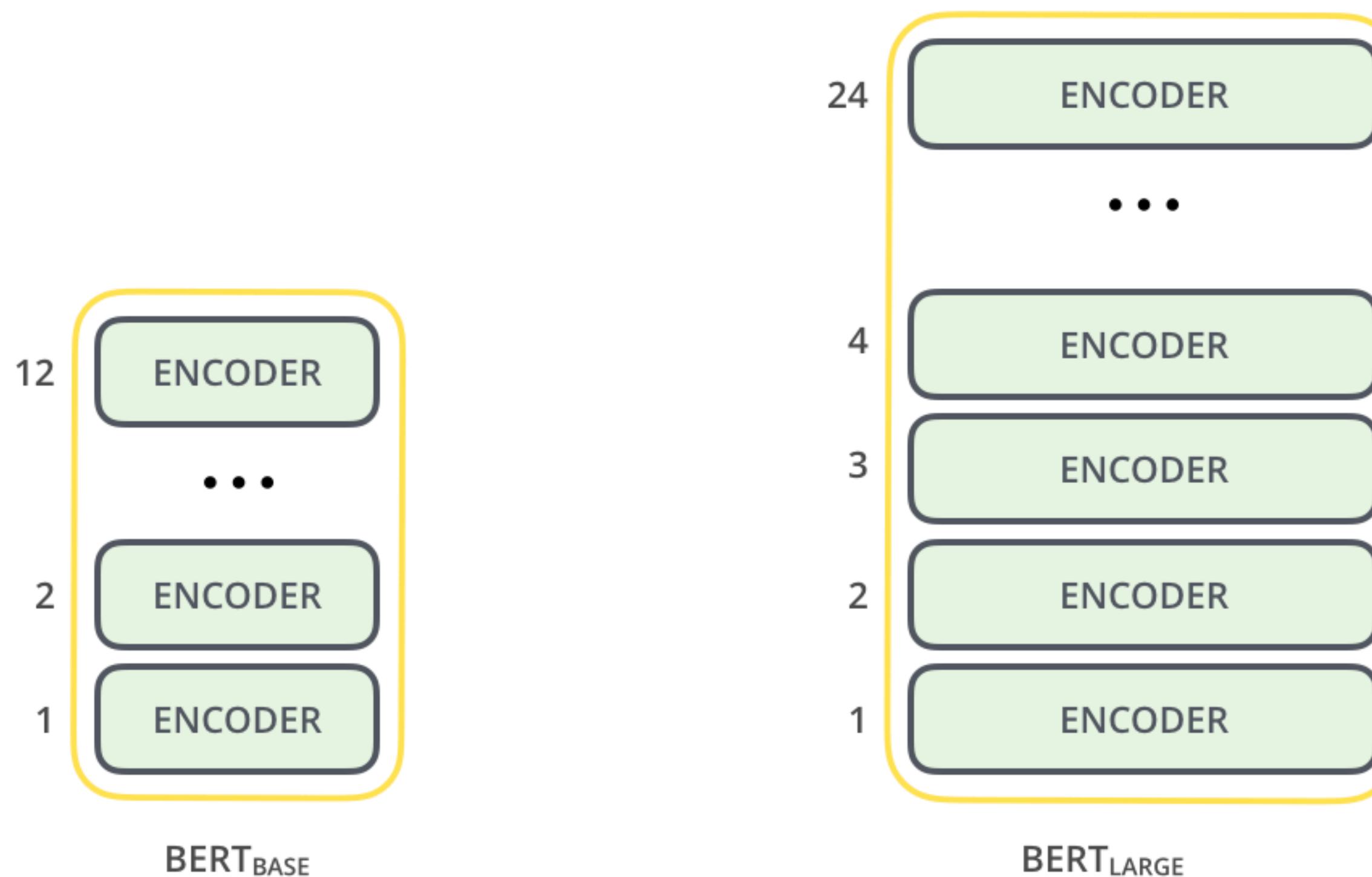


$BERT_{LARGE}$

$BERT_{BASE}(L = 12, H = 768, A = 12, \text{Total Parameters} = 110M)$

BERT

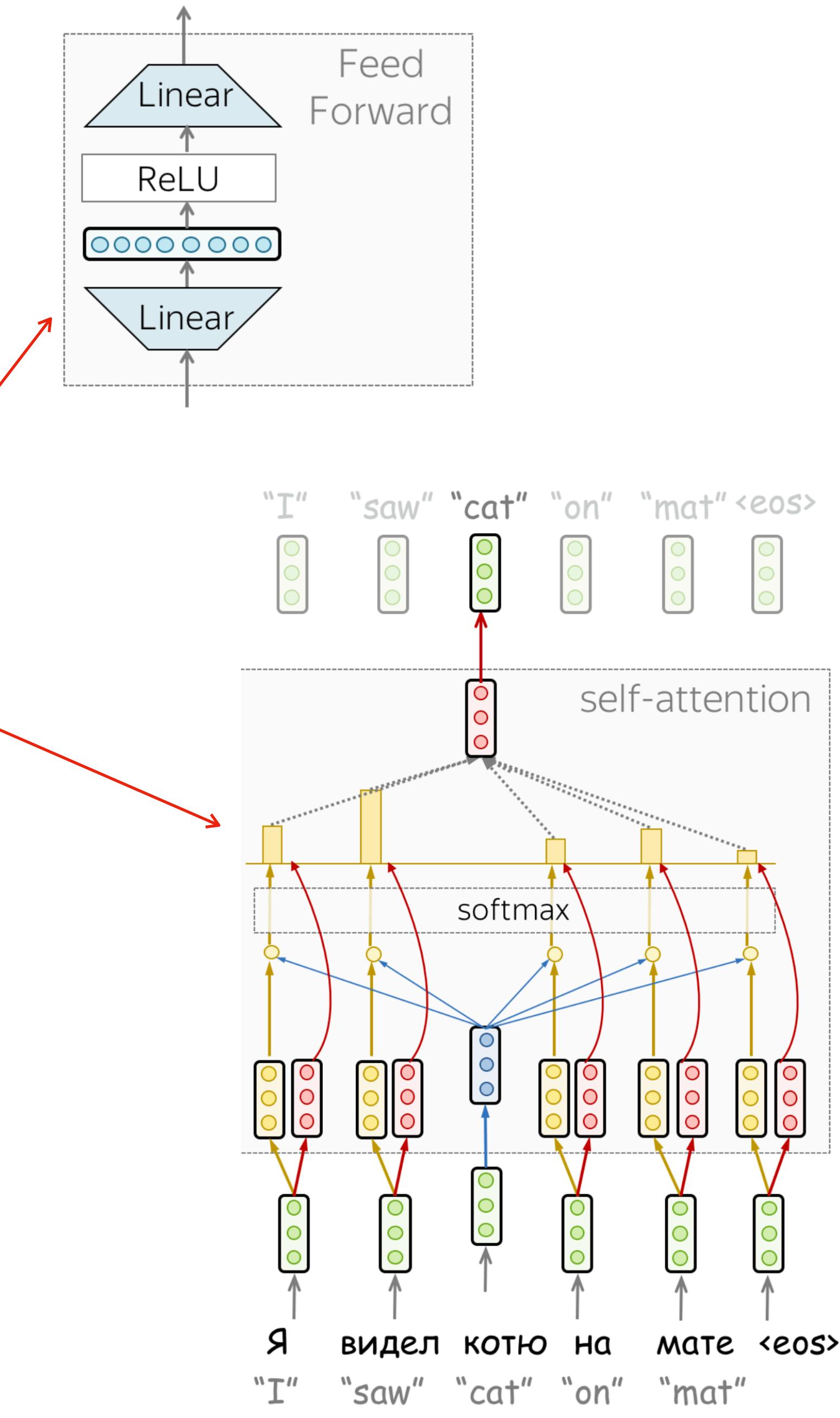
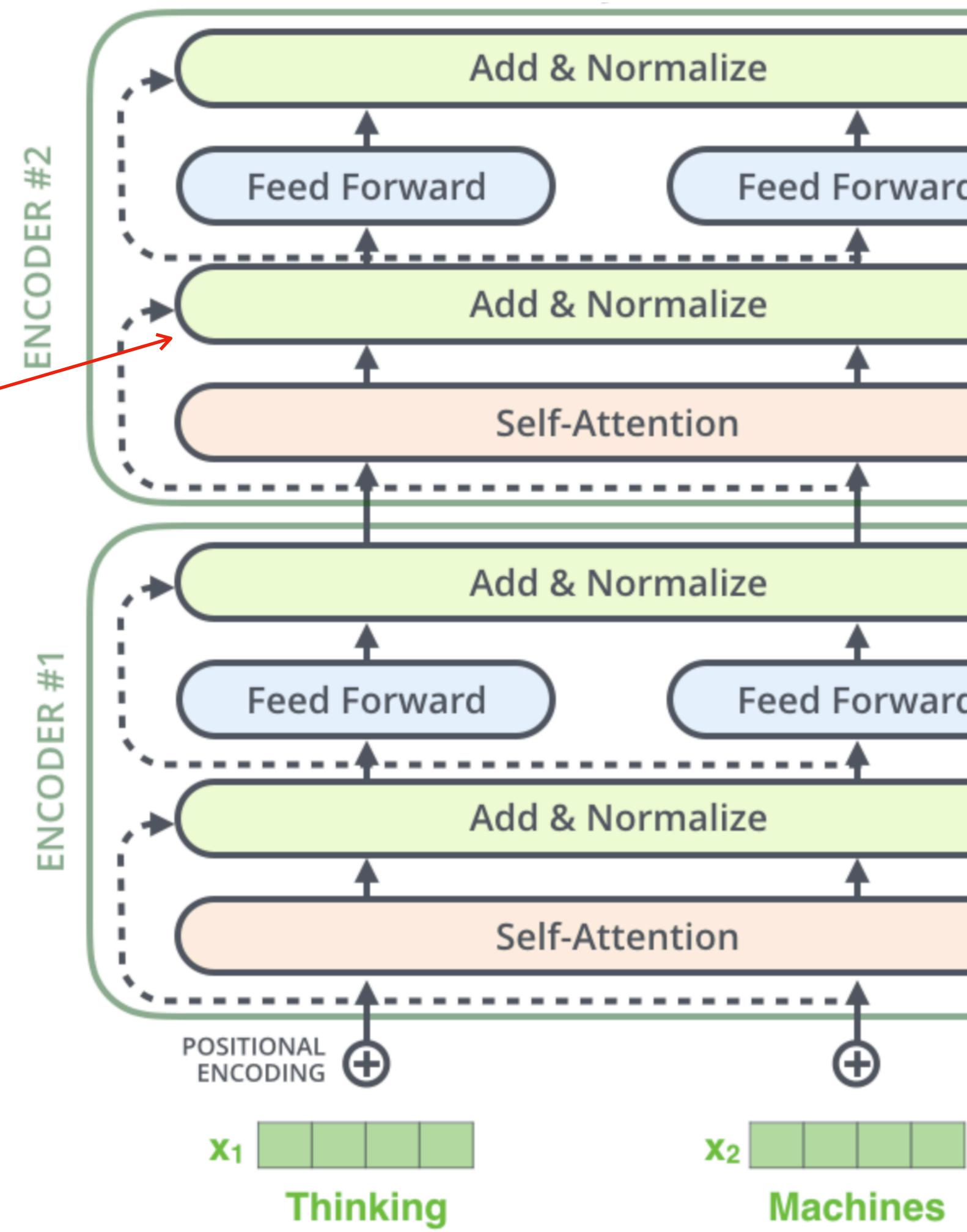
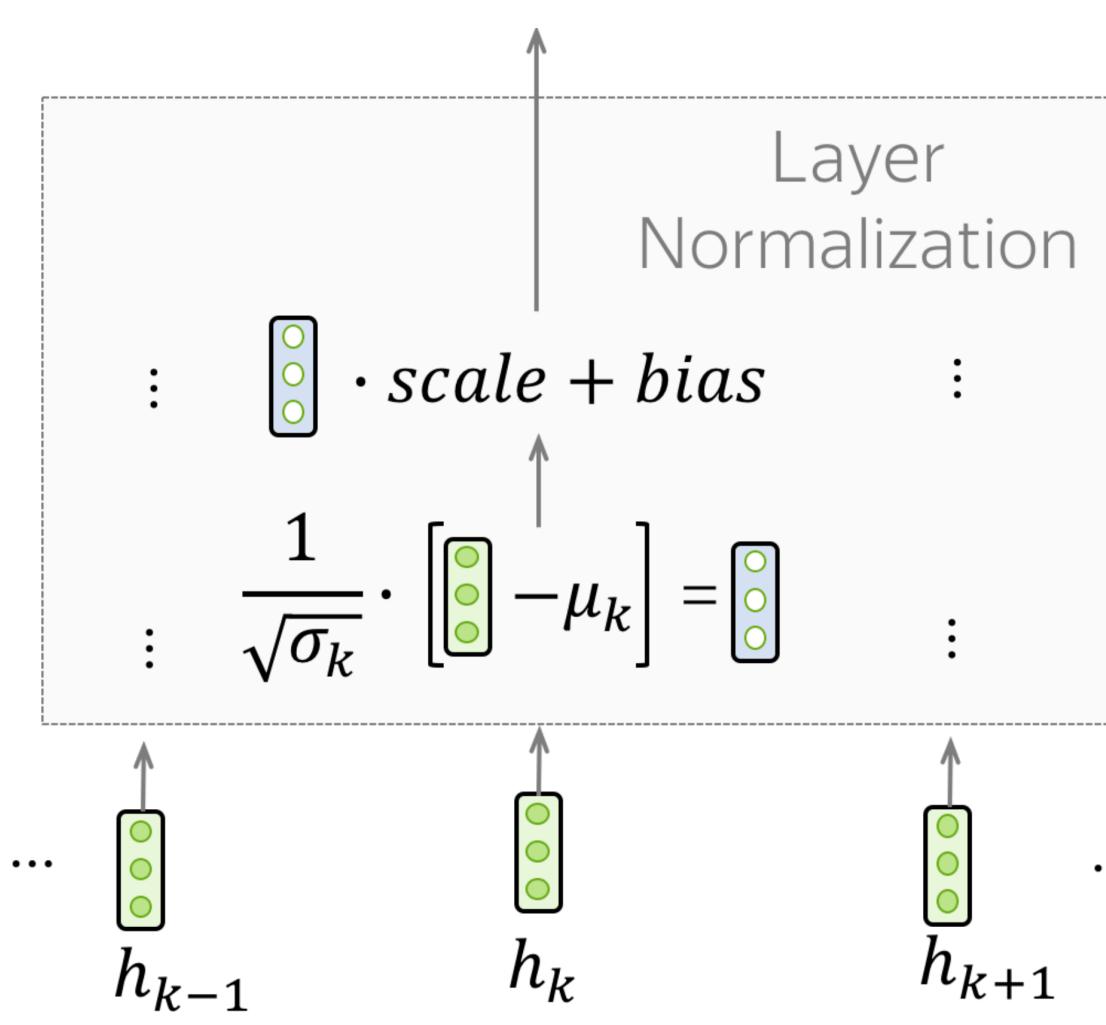
$BERT_{LARGE}(L = 24, H = 1024, A = 16, \text{Total Parameters} = 340M)$



$BERT_{BASE}(L = 12, H = 768, A = 12, \text{Total Parameters} = 110M)$

Трансформер

Небольшое напоминание



BERT

Особенности:

1. Способ обучения:

- Masked Language Modeling
- Next Sentence Prediction

2. Использование:

- Вместо task-specific моделей

WordPieces

1. Инициализировать список ворд-юнитов всеми возможными символами
2. Построить языковую модель на списке из п. 1
3. Сгенерировать новый ворд-юнит комбинируя имеющиеся, выбрав комбинацию, максимально увеличивающую правдоподобие
4. Повторять с п. 2 до тех пор пока увеличение совместного правдоподобия не станет ниже некоторого заранее выбранного значения

WordPieces

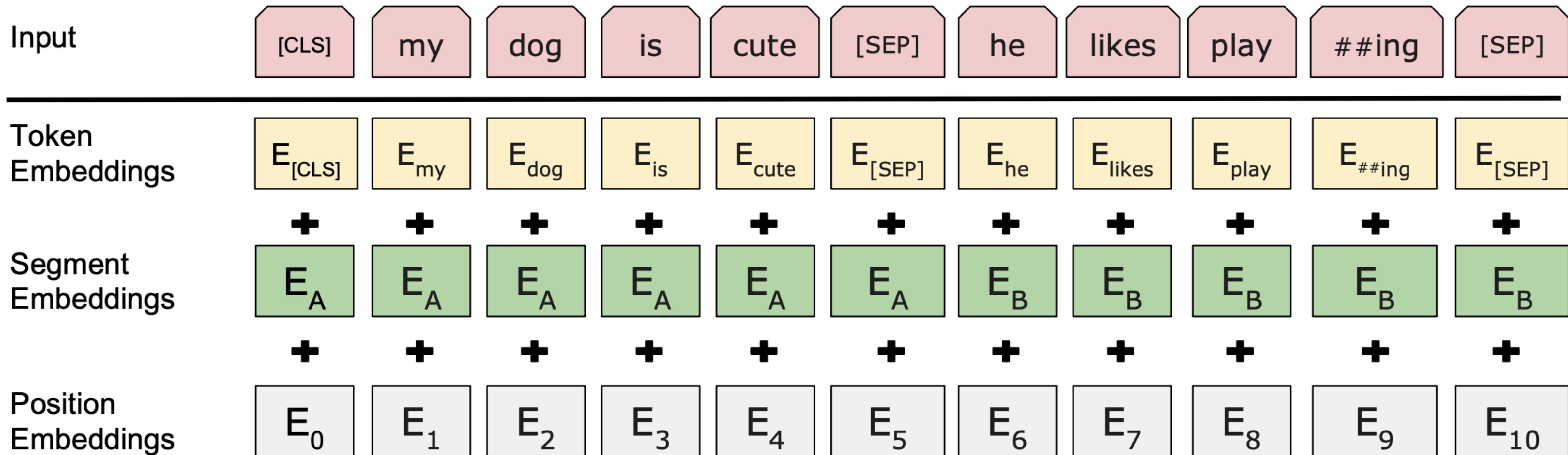
1. Инициализировать список ворд-юнитов всеми возможными символами
2. Построить языковую модель на списке из п. 1
3. Сгенерировать новый ворд-юнит комбинируя имеющиеся, выбрав комбинацию, максимально увеличивающую правдоподобие
4. Повторять с п. 2 до тех пор пока увеличение совместного правдоподобия не станет ниже некоторого заранее выбранного значения

Смысл:

- лучше обрабатывать слова, которые могут иметь разные формы: `walked`, `walker`, `walks`, `walking`.
- Получать смысловые значения очень редких слов, разбивая их на составные части

BERT

Model input



BERT

Next Sentence Prediction

NSP - задача бинарной классификации, половина из обучающих сэмплов это последовательные предложения из текстов, половина – нет.

Используя специальный токен [CLS] модель пытается предсказать лейбл для текущего примера.

Это учит модель понимать отношения между предложениями и помогает понять зависимости, которые пригодятся для задач обработки естественного языка, которые будут рассмотрены позднее.

BERT

Next Sentence Prediction

INPUT: [CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]

LABEL: isNext

BERT

Next Sentence Prediction

INPUT: [CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]

LABEL: isNext

INPUT: [CLS] the man went to [MASK] store [SEP] penguin [MASK] are flight ##less birds [SEP]

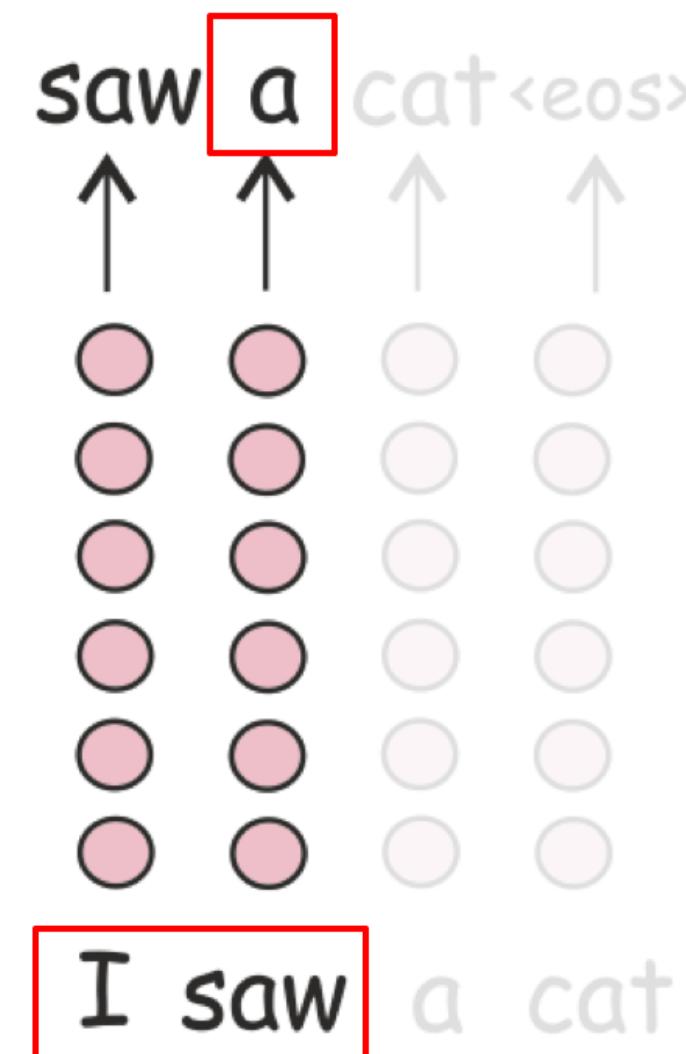
LABEL: notNext

BERT

Masked Language Modeling. Зачем нужны маски?

Language Modeling

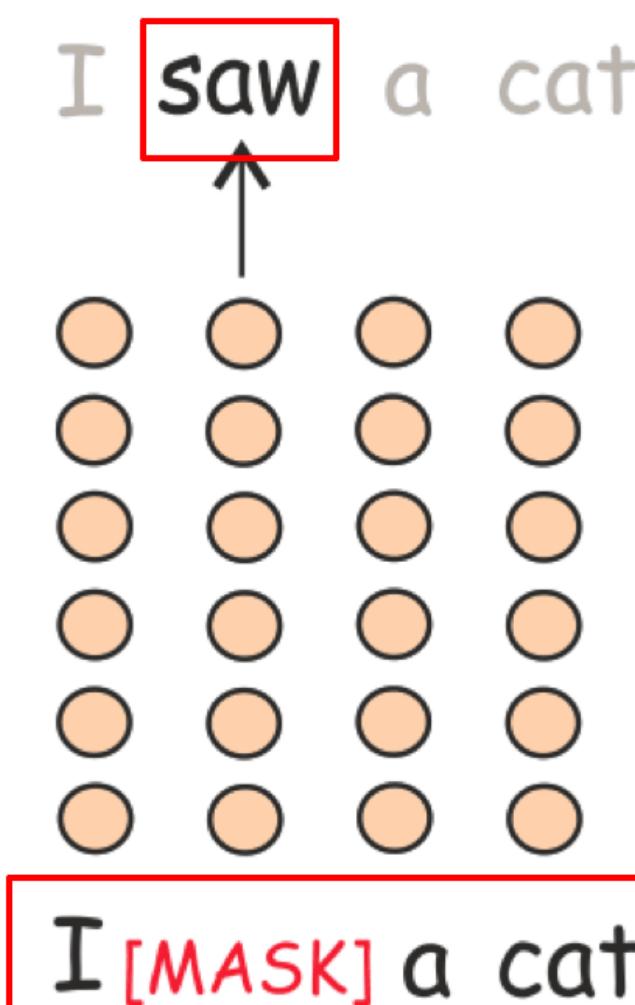
- Target: next token
- Prediction: $P(*) | I \text{ saw}$



left-to-right, does
not see future

Masked Language Modeling

- Target: current token (the true one)
- Prediction: $P(*) | I [\text{MASK}] a cat$

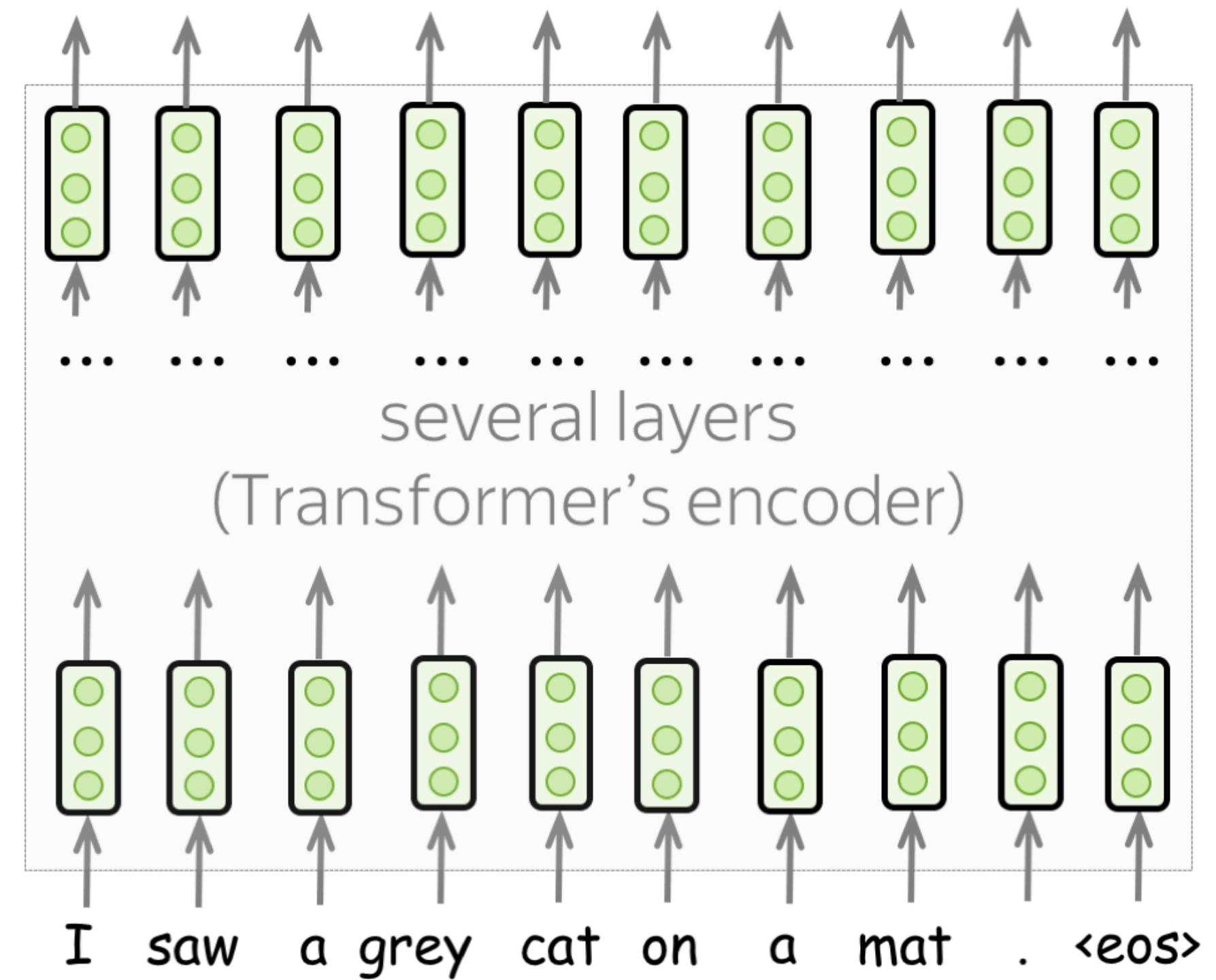


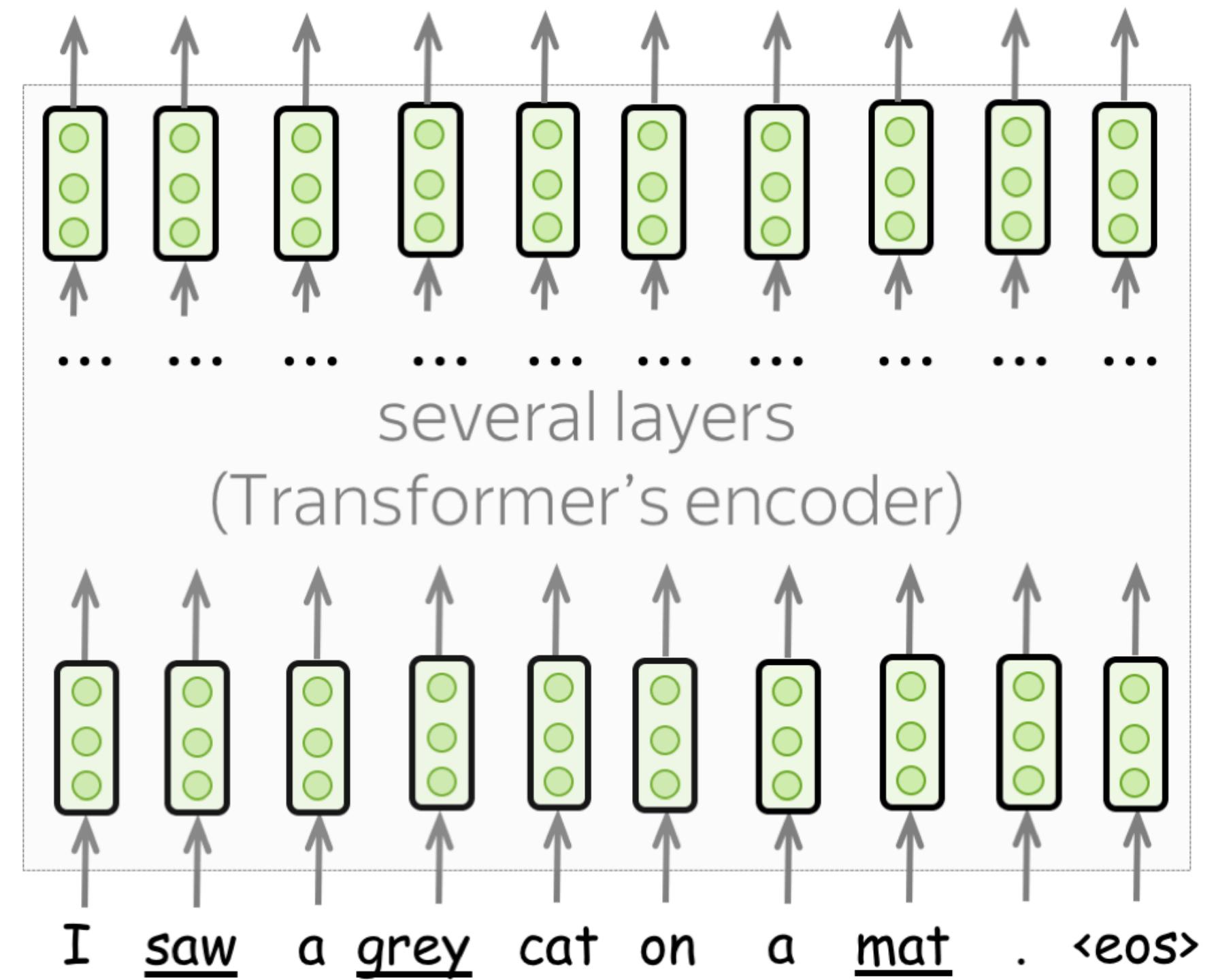
sees the whole text, but
something is corrupted

BERT

Masked Language Modeling

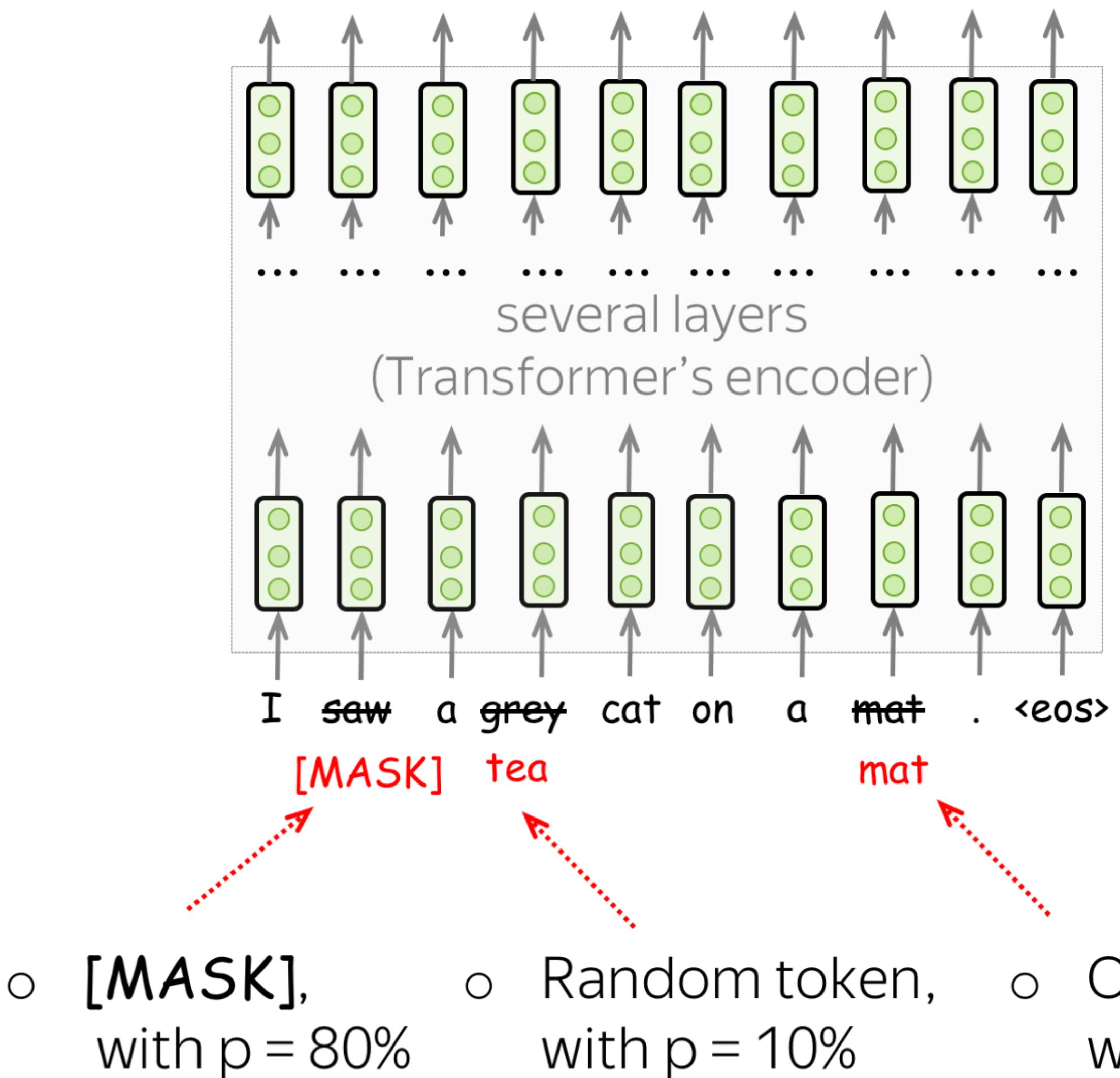
1. Выбрать некоторое количество токенов (15% используют авторы статьи)
2. Заменить выбранные токены
(с вероятностями: 80% на специальный токен [MASK], 10% на случайный токен, 10% - оставить как есть)
3. Предсказание оригинальных токенов (подсчет функции потерь)





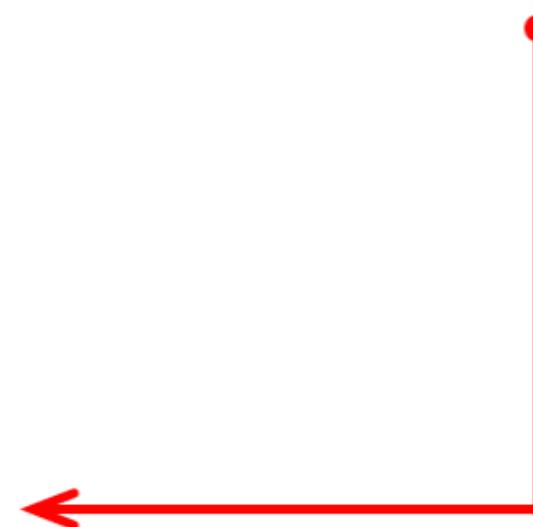
At each training step:

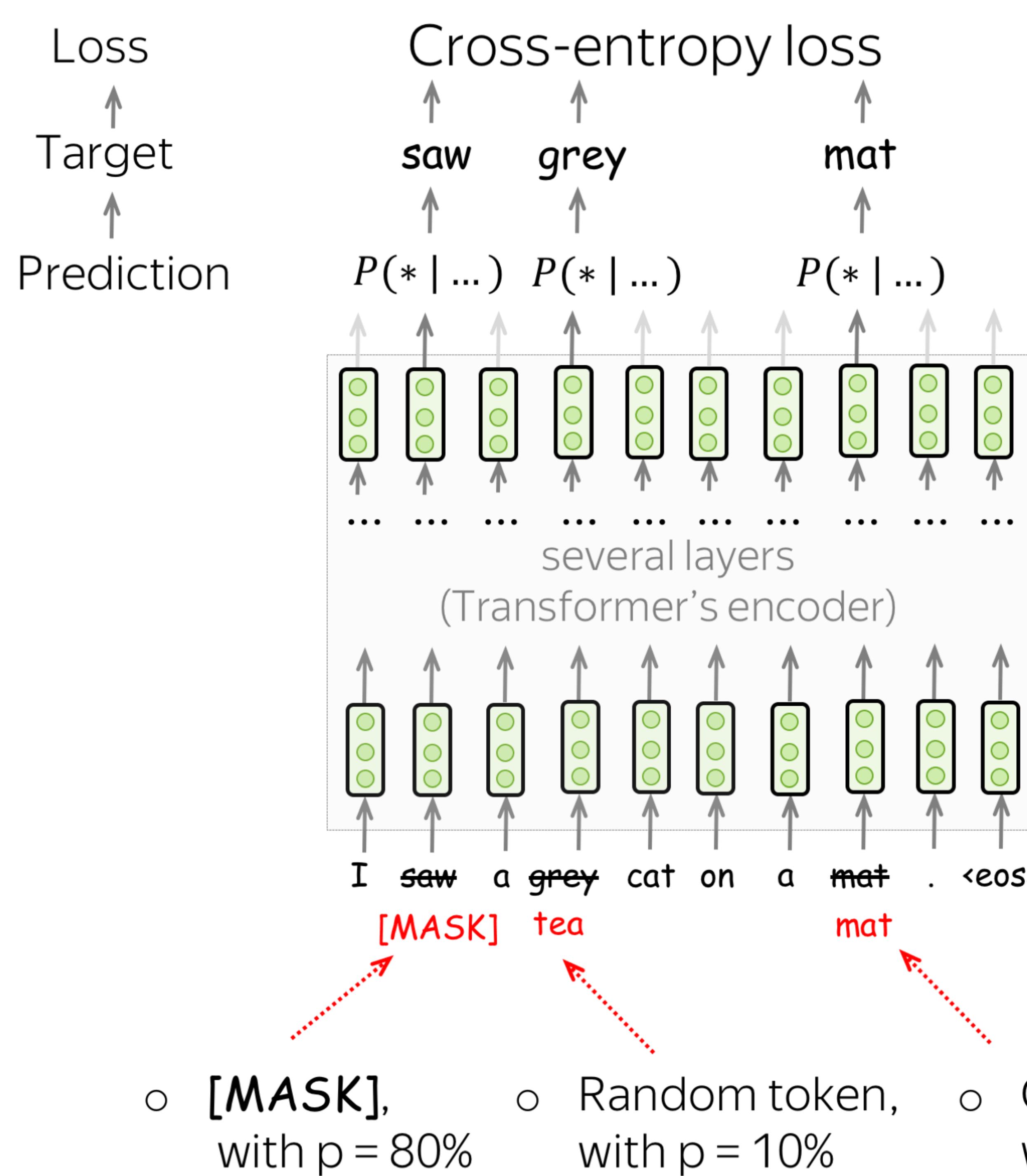
- pick randomly 15% of tokens



At each training step:

- pick randomly 15% of tokens
- replace each of the chosen tokens with something





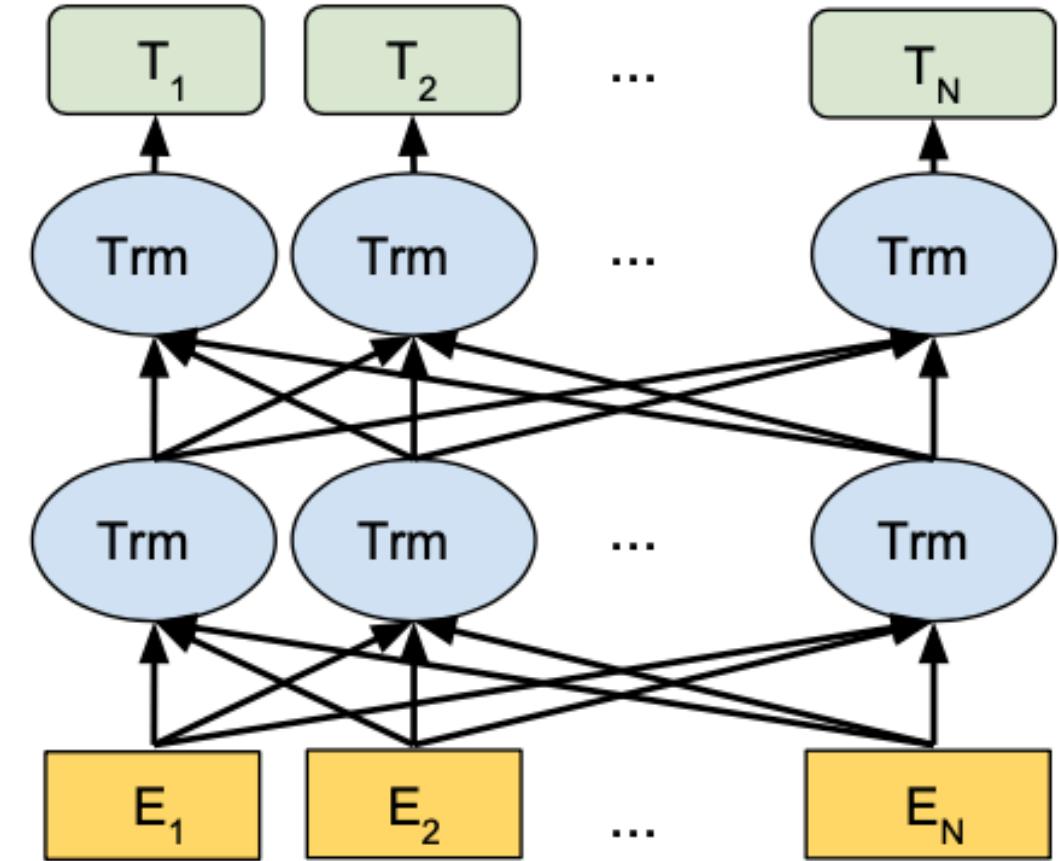
At each training step

- pick randomly 15% of tokens
 - replace each of the chosen tokens with something
 - predict original chosen tokens

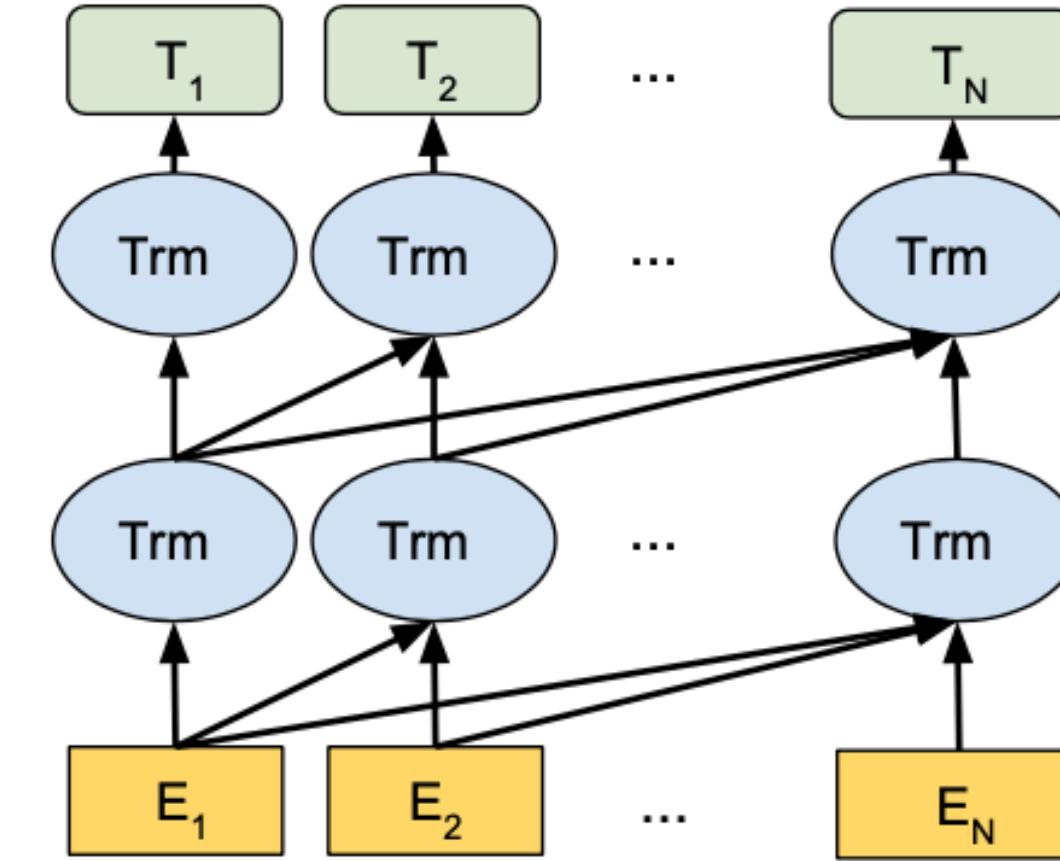
- [MASK], with p = 80%
 - Random token, with p = 10%
 - Original token with p = 10%

Архитектуры

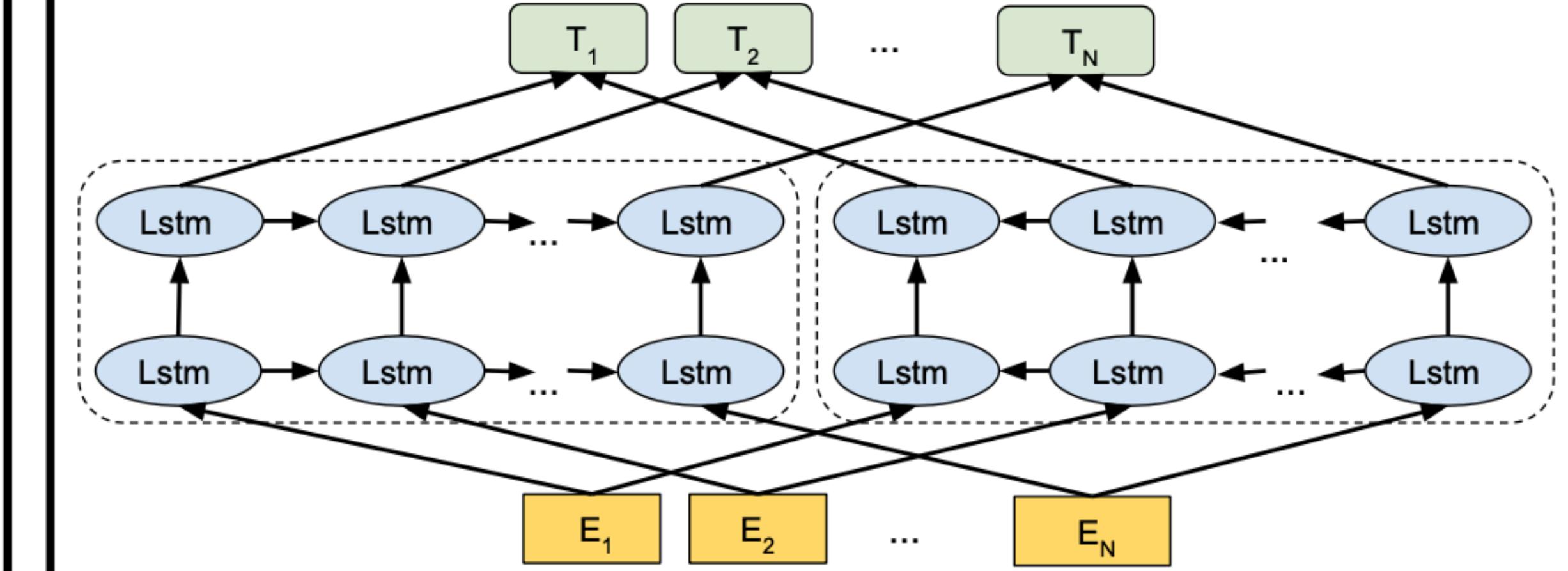
BERT (Ours)



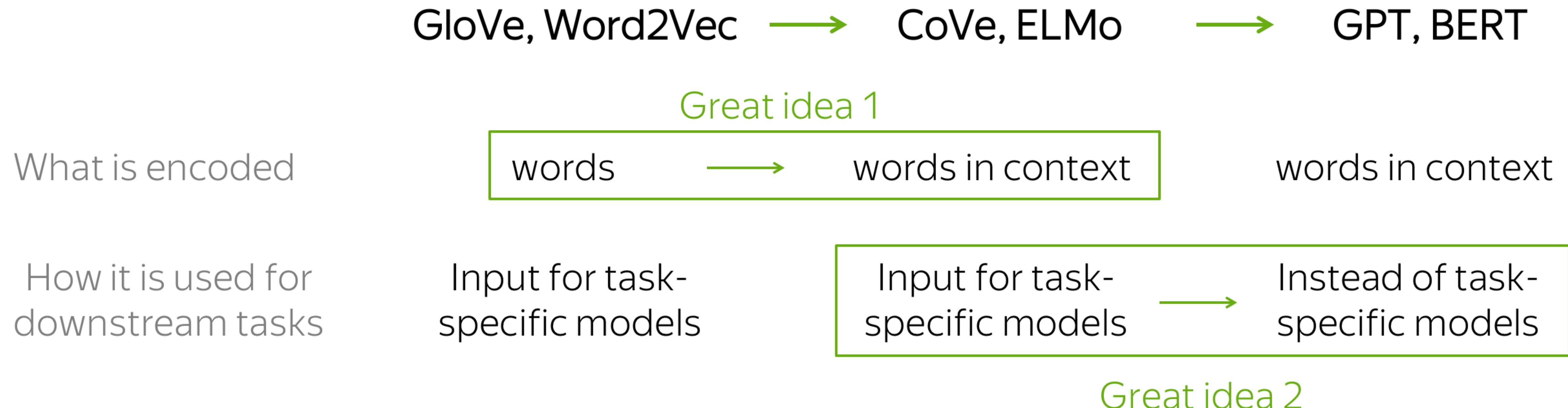
OpenAI GPT



ELMo

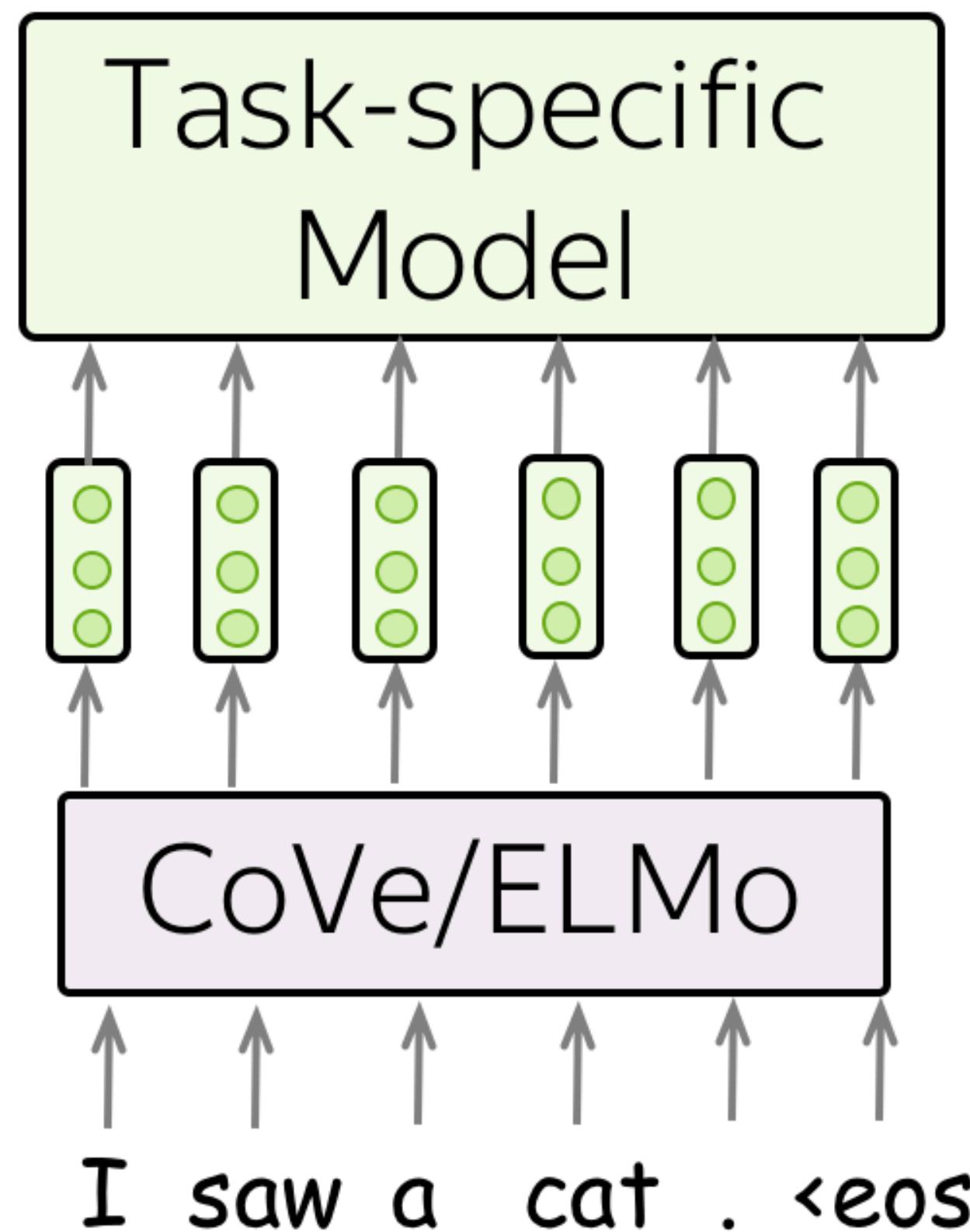


Вспомним о Transfer Learning

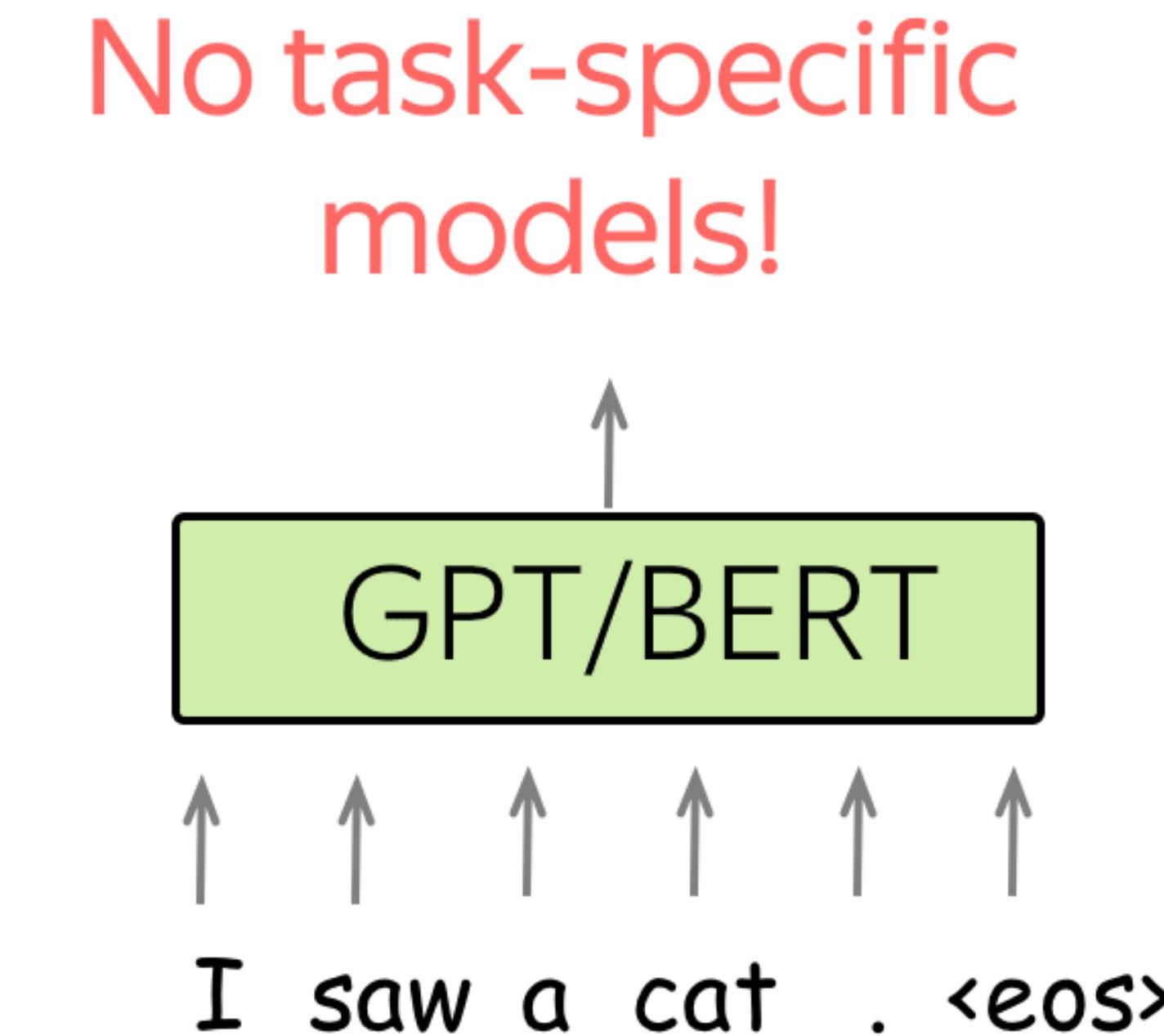


Вспомним о Transfer Learning

Before



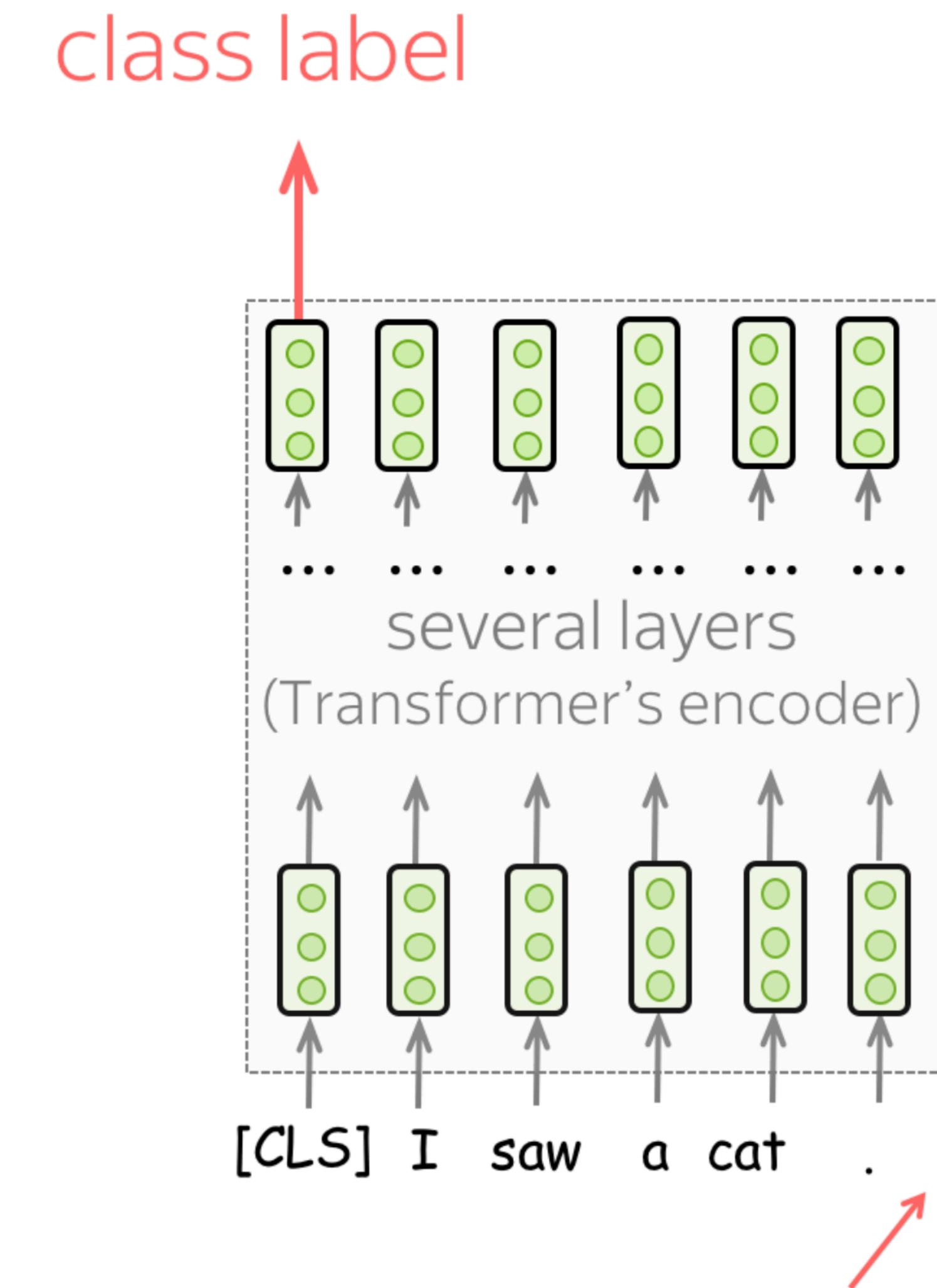
After



Классификация предложений

Чтобы классифицировать отдельные предложения, предлагается использовать представление специального токена [CLS].

На вход подается одно предложение.

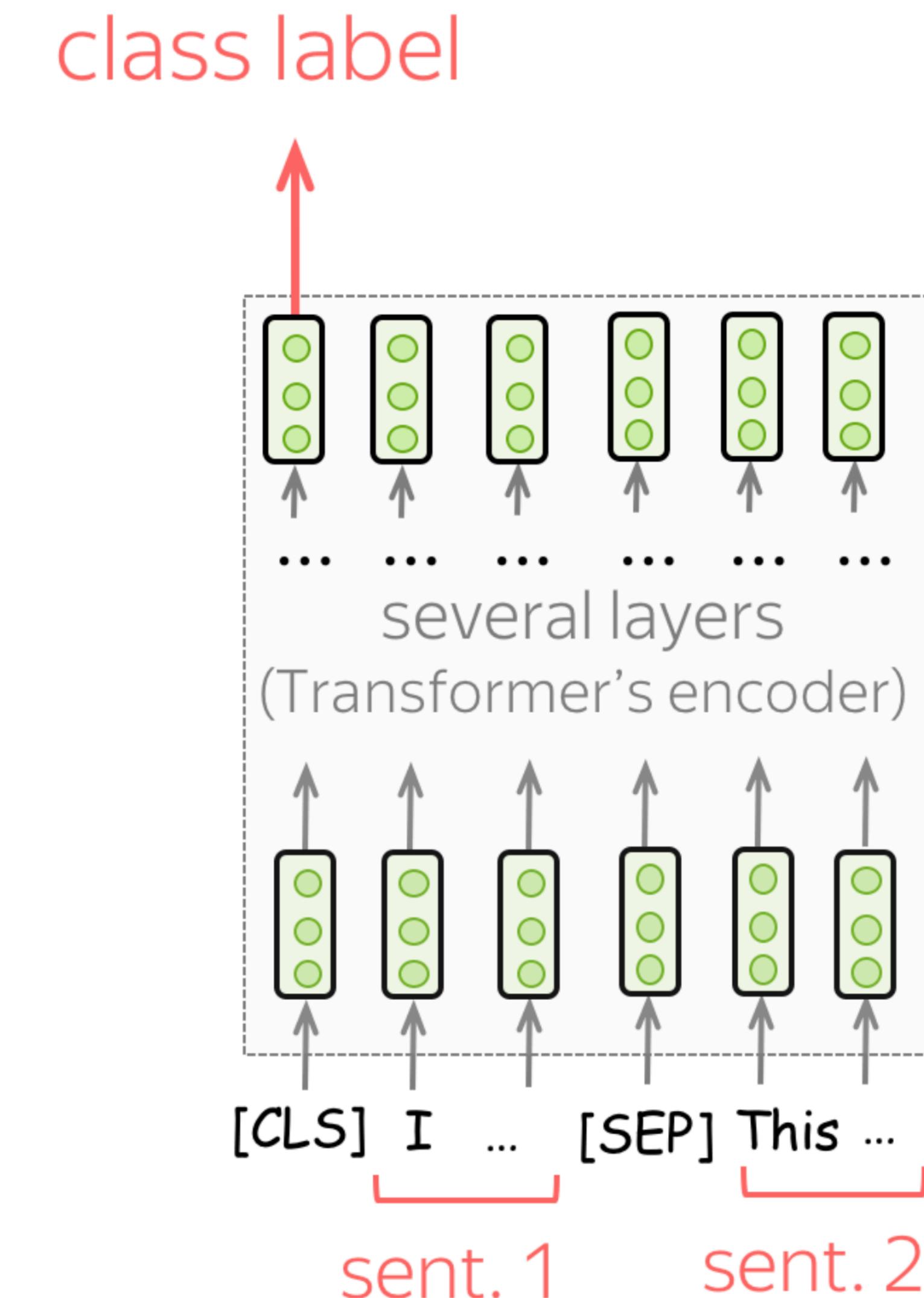


No second sentence!

Классификация пар предложений

Чтобы классифицировать пары предложений, предлагается также использовать представление специального токена [CLS].

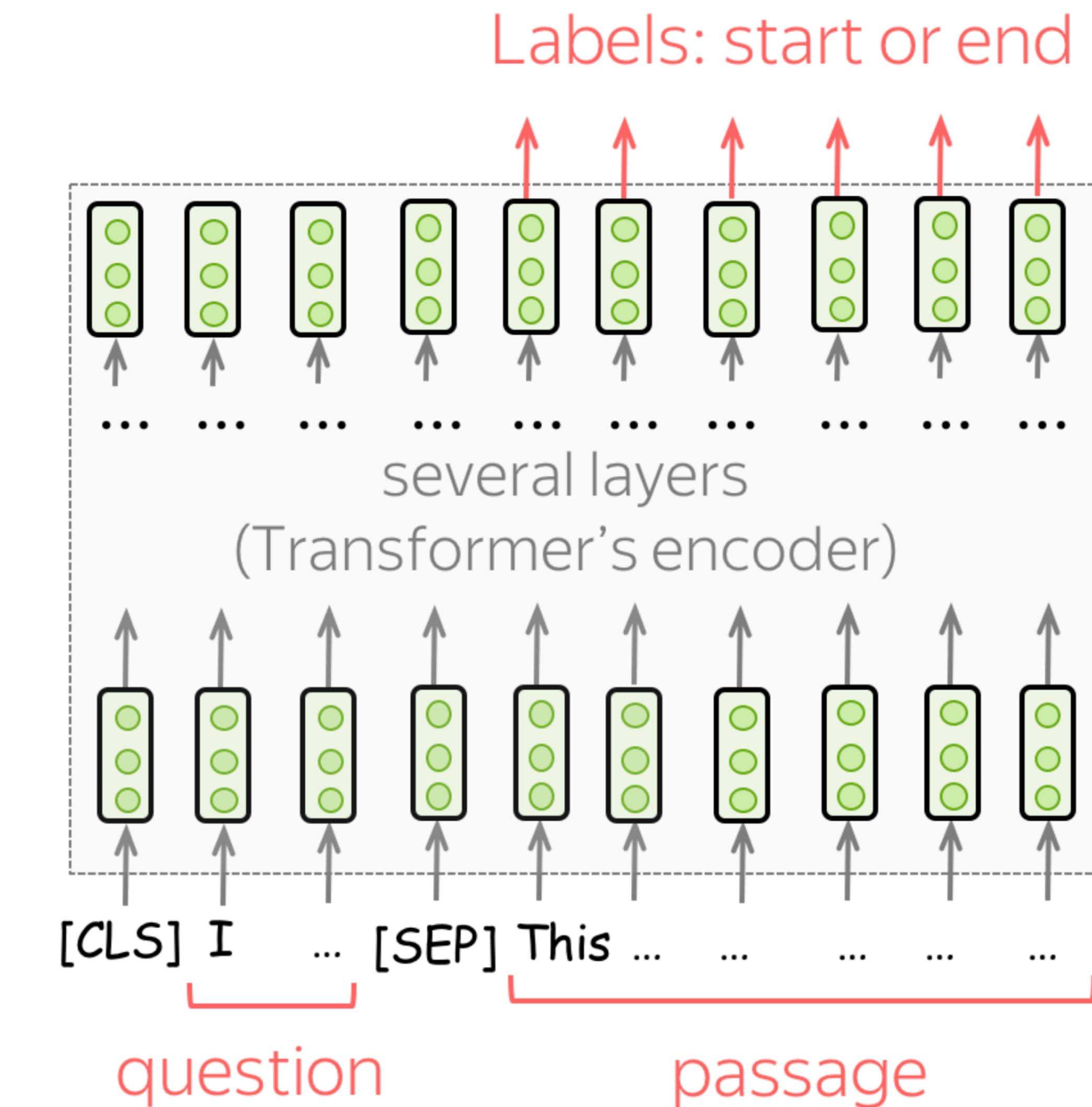
На вход подается оба предложения, прямо как во время обучения.



Нахождение ответов на вопрос

Чтобы решать задачу нахождения ответа на вопрос в отрывке текста, предлагается для репрезентации каждого токена предсказывать является ли он началом или концом требуемой части с ответом на вопрос.

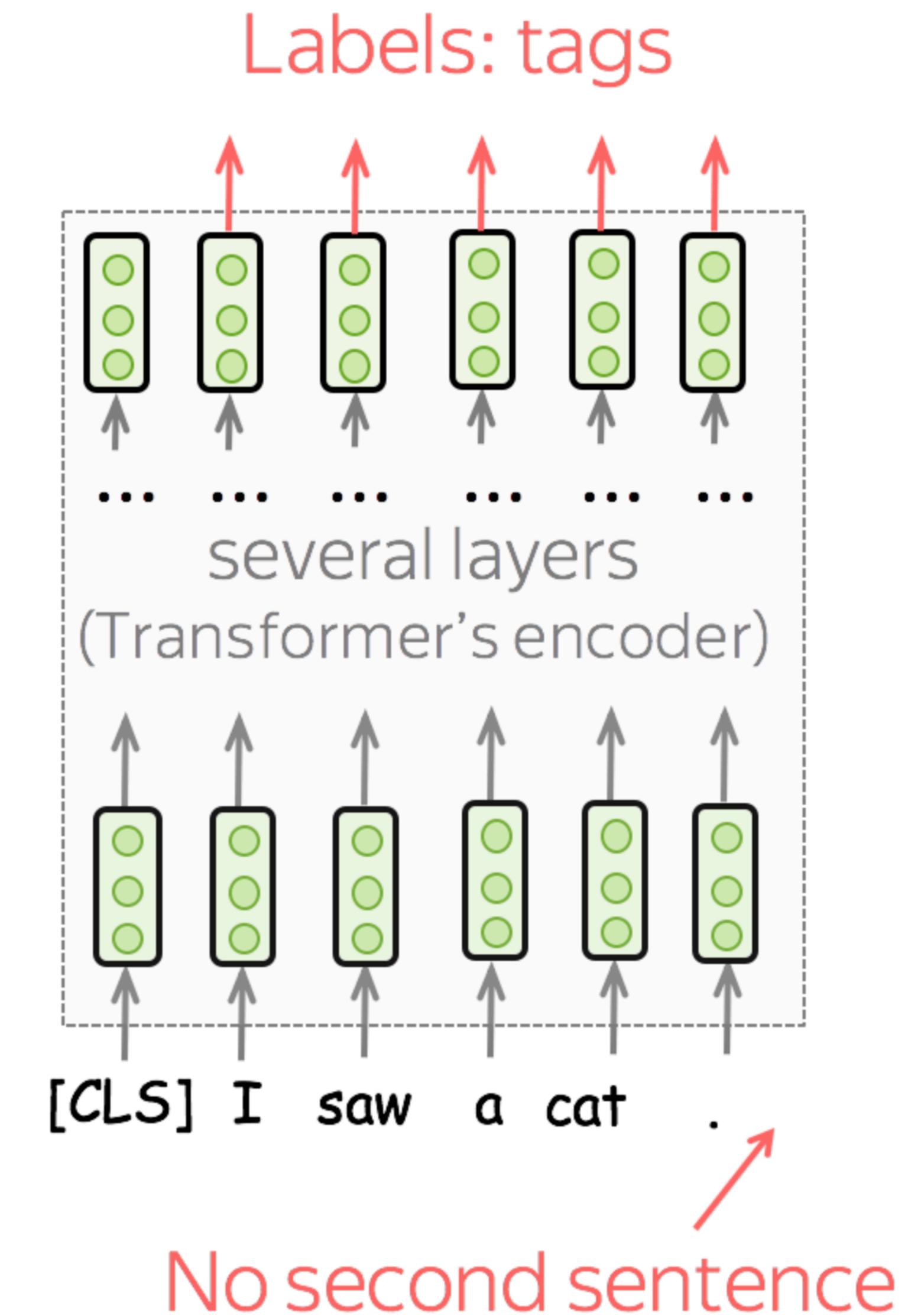
На вход подается вопрос и отрывок через разделитель.



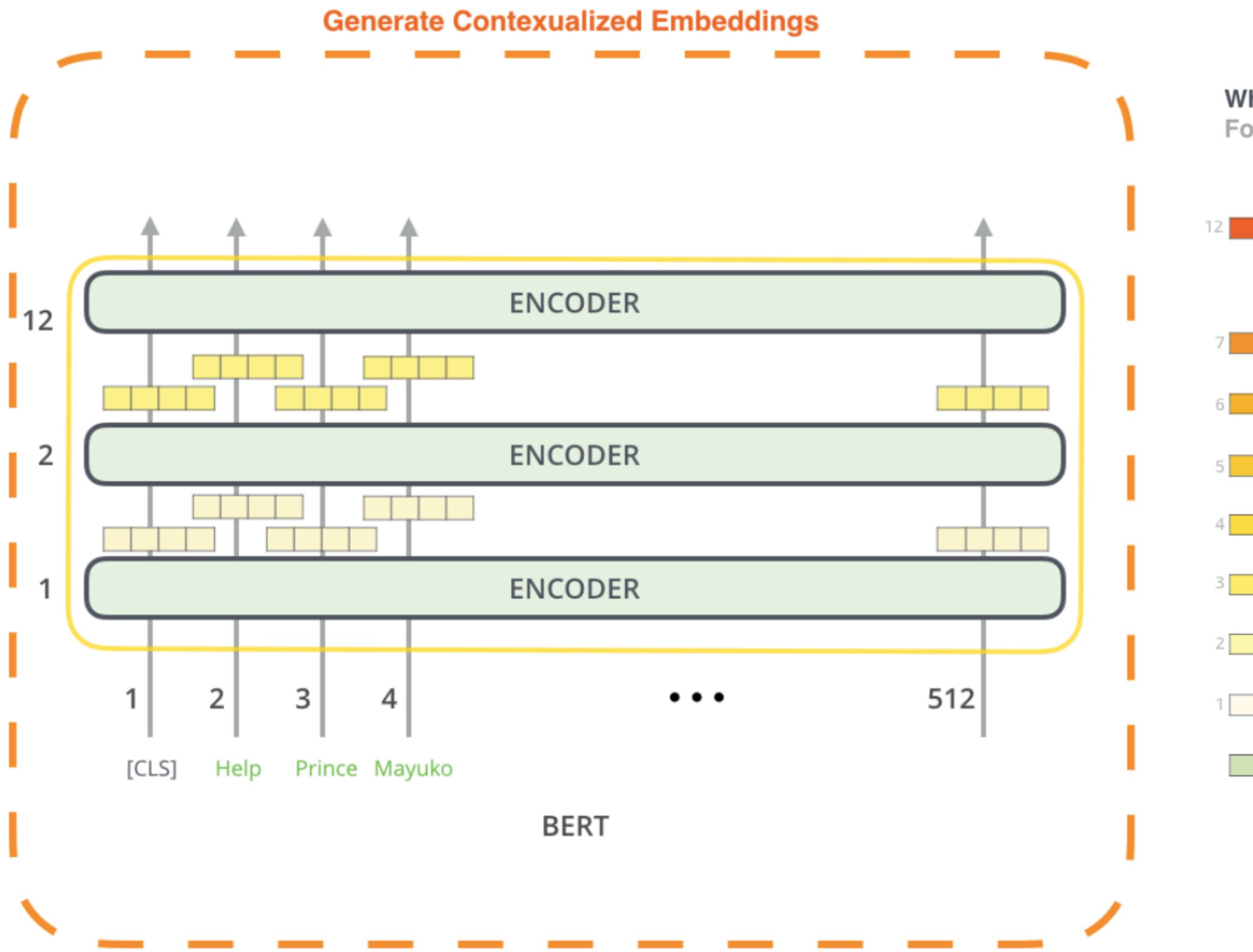
Разметка слов

В задачах разметки слов требуется предсказывать тэги для каждого токена (прим: часть речи или сущность: локация, человек и тд.)

На вход предлагается давать предложение, а по выходам сети на каждом токене строить предсказания тэгов.



Получение эмбеддингов



What is the best contextualized embedding for “Help” in that context?
For named-entity recognition task CoNLL-2003 NER

	Dev F1 Score
First Layer	91.0
Last Hidden Layer	94.9
Sum All 12 Layers	95.5
Second-to-Last Hidden Layer	95.6
Sum Last Four Hidden	95.9
Concat Last Four Hidden	96.1

Legend: Embedding (green), Help (red), Prince (yellow), Mayuko (orange)

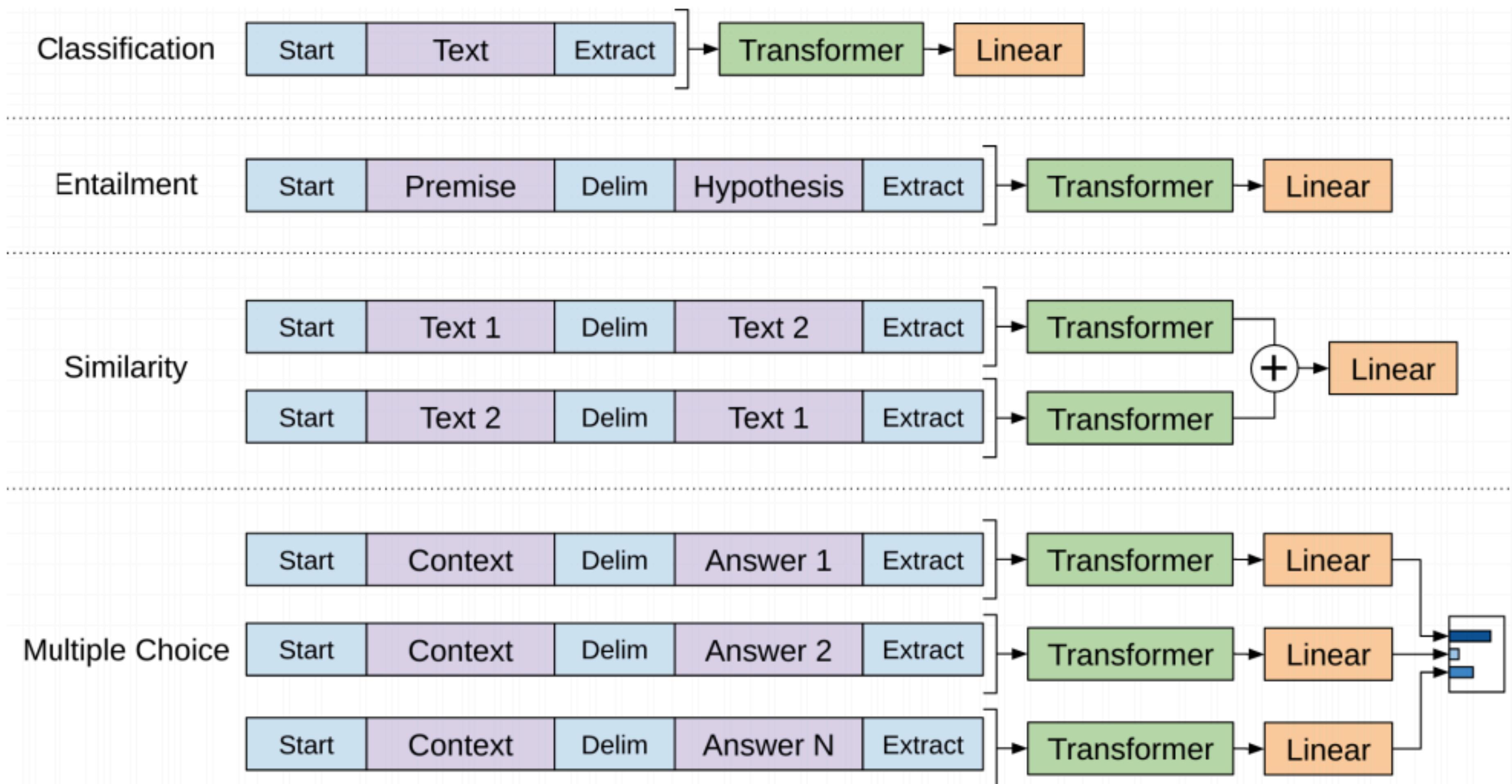
Detailed description of the table rows:

- First Layer:** Shows a green 4x4 matrix labeled "Embedding".
- Last Hidden Layer:** Shows a red 4x4 matrix labeled "Help".
- Sum All 12 Layers:** Shows a 12x4 matrix where each row is a sum of the previous row and a yellow 4x4 matrix labeled "Prince". The final result is a red 4x4 matrix labeled "Help".
- Second-to-Last Hidden Layer:** Shows a red 11x4 matrix labeled "Help".
- Sum Last Four Hidden:** Shows a 4x4 matrix where each row is a sum of the previous row and a red 4x4 matrix labeled "Help". The final result is a red 4x4 matrix labeled "Help".
- Concat Last Four Hidden:** Shows a horizontal stack of four red 4x4 matrices labeled "Help", corresponding to the last four hidden layers.

Результаты

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Average
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
BERT _{BASE}	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT _{LARGE}	86.7/85.9	72.1	92.7	94.9	60.5	86.5	89.3	70.1	82.1

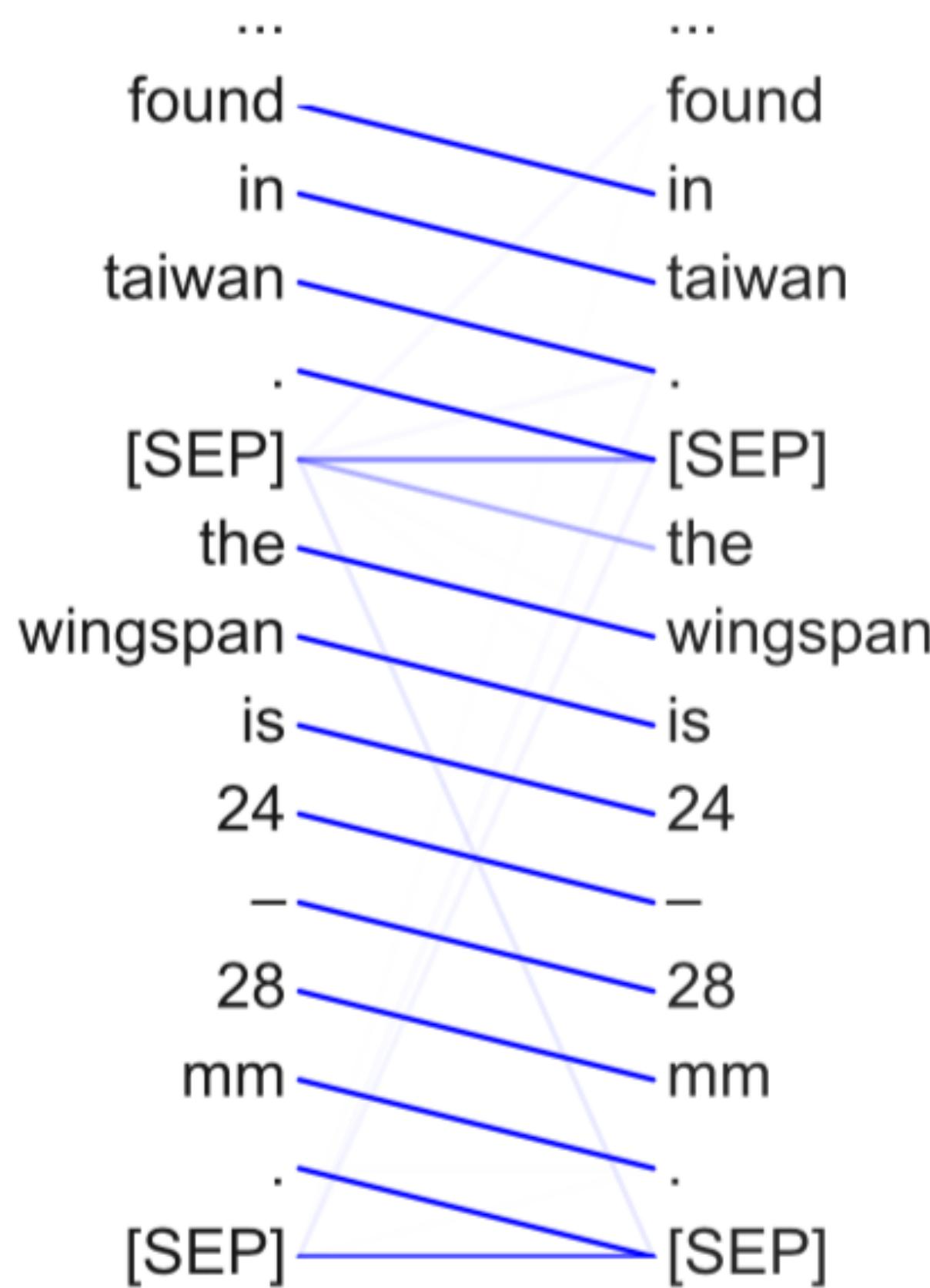
Опять чуть-чуть про GPT



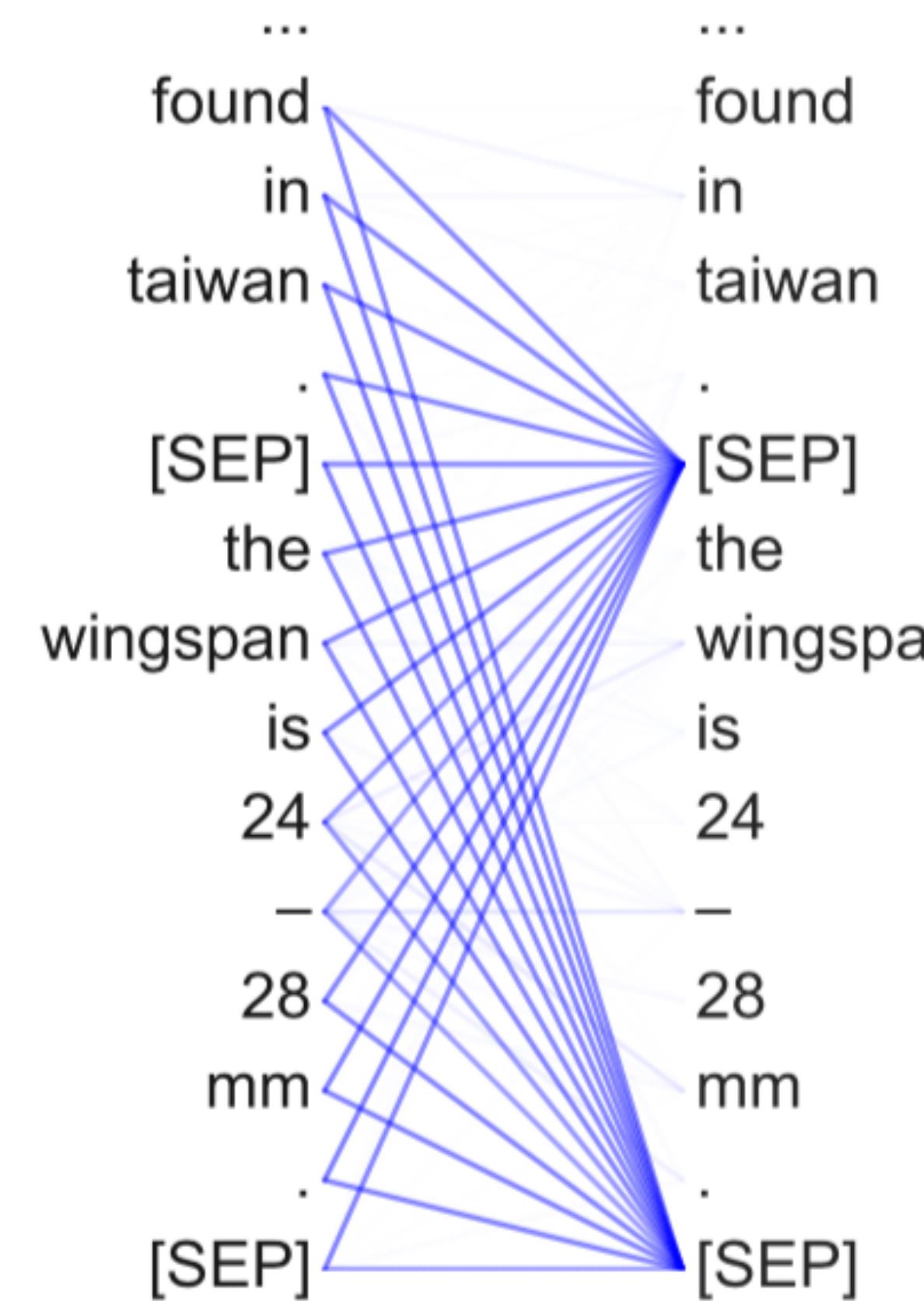
BERT's Attention Heads

(Layer - head number)

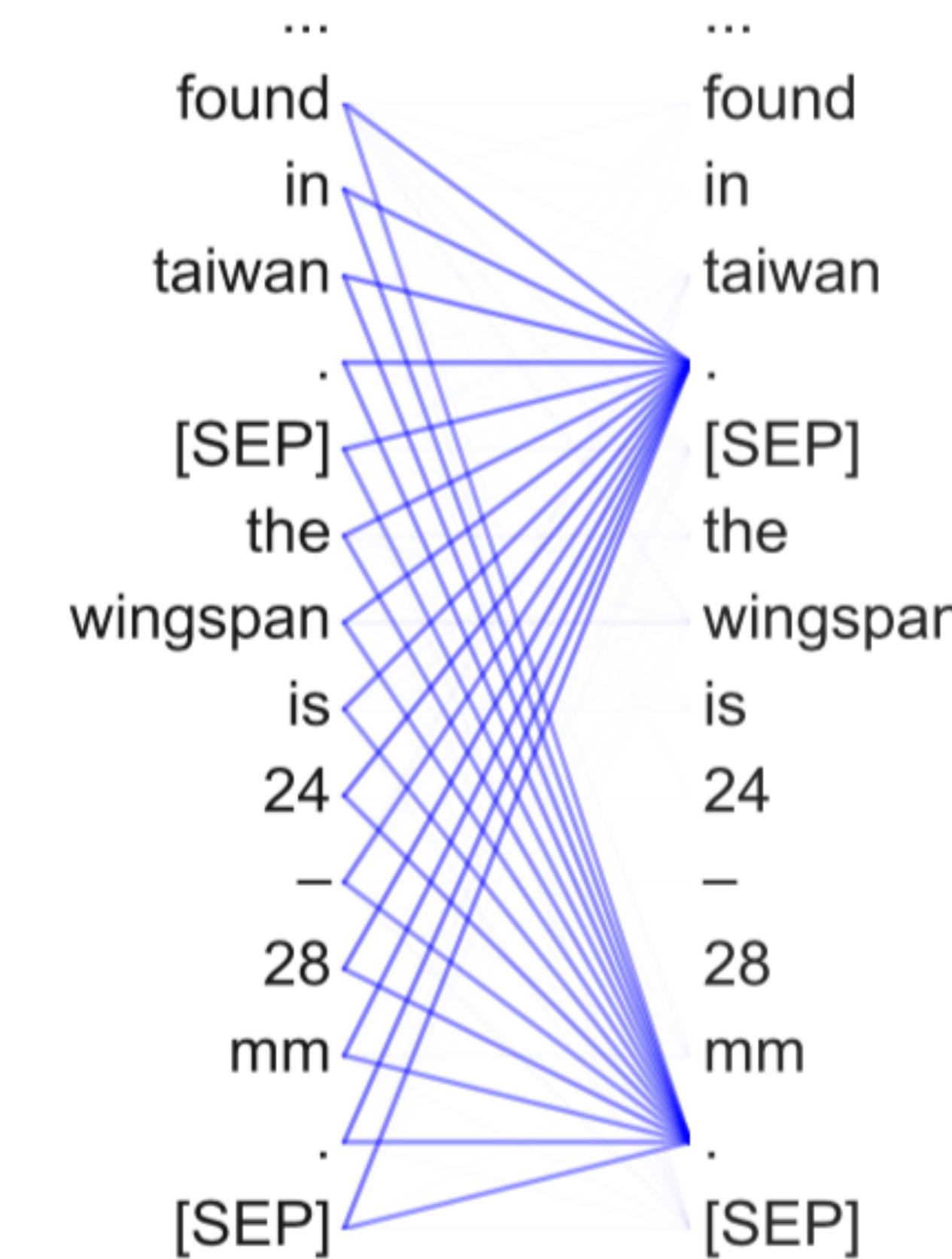
Head 3-1
Attends to next token



Head 8-7
Attends to [SEP]



Head 11-6
Attends to periods



BERT/ELMo знают факты!

Query	Answer	Generation (model log-probability)
Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8] , Florence [-1.8] , Naples [-1.9] , Milan [-2.4] , Bologna [-2.5]
Adolphe Adam died in ____.	Paris	Paris [-0.5] , London [-3.5] , Vienna [-3.6] , Berlin [-3.8] , Brussels [-4.0]
English bulldog is a subclass of ____.	dog	dogs [-0.3] , breeds [-2.2] , dog [-2.4] , cattle [-4.3] , sheep [-4.5]
The official language of Mauritius is ____.	English	English [-0.6] , French [-0.9] , Arabic [-6.2] , Tamil [-6.7] , Malayalam [-7.0]
Patrick Oboya plays in ____ position.	midfielder	centre [-2.0] , center [-2.2] , midfielder [-2.4] , forward [-2.4] , midfield [-2.7]
Hamburg Airport is named after ____.	Hamburg	Hess [-7.0] , Hermann [-7.1] , Schmidt [-7.1] , Hamburg [-7.5] , Ludwig [-7.5]

Query	Answer	Generation (model log-probability)
You are likely to find a overflow in a ____.	drain	sewer [-3.1] , canal [-3.2] , toilet [-3.3] , stream [-3.6] , drain [-3.6]
Ravens can ____.	fly	fly [-1.5] , fight [-1.8] , kill [-2.2] , die [-3.2] , hunt [-3.4]
Joke would make you want to ____.	laugh	cry [-1.7] , die [-1.7] , laugh [-2.0] , vomit [-2.6] , scream [-2.6]
Sometimes virus causes ____.	infection	disease [-1.2] , cancer [-2.0] , infection [-2.6] , plague [-3.3] , fever [-3.4]
Birds have ____.	feathers	wings [-1.8] , nests [-3.1] , feathers [-3.2] , died [-3.7] , eggs [-3.9]
Typing requires ____.	speed	patience [-3.5] , precision [-3.6] , registration [-3.8] , accuracy [-4.0] , speed [-4.1]
Time is ____.	finite	short [-1.7] , passing [-1.8] , precious [-2.9] , irrelevant [-3.2] , gone [-4.0]
You would celebrate because you are ____.	alive	happy [-2.4] , human [-3.3] , alive [-3.3] , young [-3.6] , free [-3.9]
Skills can be ____.	taught	acquired [-2.5] , useful [-2.5] , learned [-2.8] , combined [-3.9] , varied [-3.9]
A pond is for ____.	fish	swimming [-1.3] , fishing [-1.4] , bathing [-2.0] , fish [-2.8] , recreation [-3.1]

Вопросы:

- Чем отличается biLM от LM? Что мы стараемся максимизировать в обоих случаях?
- Что такое Masked Language Modeling? Как он используется для обучения BERT?
- Как решать задачу классификации предложений с помощью BERT?

Источники:

- <https://arxiv.org/pdf/1802.05365.pdf>
- <https://arxiv.org/pdf/1810.04805.pdf>
- https://lena-voita.github.io/nlp_course/transfer_learning.html
- https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html
- <https://jalammar.github.io/illustrated-bert/>