Лекция 8
Нейронные сети для обработки естественного языка

Байгушев Данила

2 ноября 2022 г.

Задачи NLP

- Машинный перевод
- Анализ тональности
- Выделение именованных сущностей
- Чат-боты
- Понимание естественного языка
- Понимание изображений

Entity tracking

mary got the milk there john moved to the bedroom sandra went back to the kitchen mary travelled to the hallway john got the football there john went to the hallway john put down the football mary went to the garden john went to the kitchen sandra travelled to the hallway daniel went to the hallway mary discarded the milk where is the milk ?
answer: garden

Visual QA[1]

Who is wearing glasses?

man

woman

Where is the child sitting?

fridge

arms

Is the umbrella upside down?

yes

no

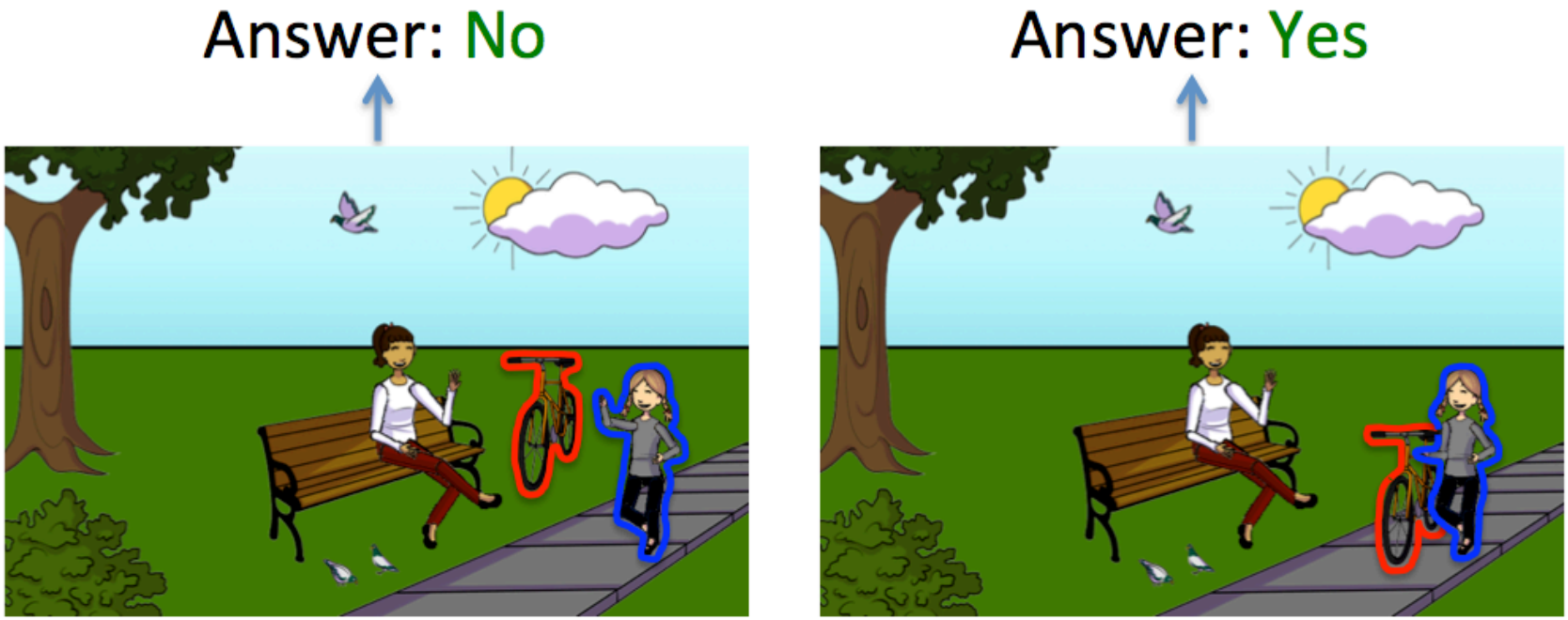
How many children are in the bed?

2

1

[1]Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering (CVPR 2017)

Visual QA[2]



[2]Yin and Yang: Balancing and Answering Binary Visual Questions (CVPR 2016)

Анализ тональности

Отзыв положительный или отрицательный?

y=1 Мне очень понравился этот фильм. Никогда раньше ничего подобного не видел!!!)))

y=0 Ужасно! Ушла с середины фильма, тк больше невозможно было смотреть.

y=0 Ну да, конечно. Просто отличный фильм.

Анализ тональности

Отзыв положительный или отрицательный?

y=1 Мне очень понравился этот фильм. Никогда раньше ничего подобного не видел!!!)))

y=0 Ужасно! Ушла с середины фильма, тк больше невозможно было смотреть.

y=0 Ну да, конечно. Просто отличный фильм.

Простой подход: Bag-of-words + Logistic regression Какие есть проблемы у такого подхода?

Анализ тональности

Отзыв положительный или отрицательный?

y=1 Мне очень понравился этот фильм. Никогда раньше ничего подобного не видел!!!)))

y=0 Ужасно! Ушла с середины фильма, тк больше невозможно было смотреть.

y=0 Ну да, конечно. Просто отличный фильм.

Простой подход: Bag-of-words + Logistic regression Какие есть проблемы у такого подхода?

- Не учитывает сарказм
- Не учитывает схожесть слов (например, кот $\leftrightarrow$ котенок)
- Не учитывает порядок слов

Представление слов

Задача

Сопоставить каждому слову w из словаря V вектор $e(w)$.

Подходы:

- One-hot encoding
- Counts
- CBOW
- Skip-grams

One-hot encoding

Кодируем слово w_i вектором $[0, 0, \ldots, 0, \underbrace{1}_{i}, 0, \ldots, 0]^T$ Плюсы:

- Просто реализовать
- Можно использовать разреженное представление

Минусы:

- Не учитывает близость слов
- Огромная размерность

Counts

... and the cute **kitten** purred and then ...
... the cute **furry** cat purred and miaowed ...
... that small **kitten** miaowed and she ...
... the loud **furry** dog ran and bit ...

Словарь: bit, cute, furry, loud, miaowed, purred, ran, small

kitten: cute, purred, small, miaowed $\Rightarrow [0, 1, 0, 0, 1, 1, 0, 1]^T$
cat: cute, furry, miaowed $\Rightarrow [0, 1, 1, 0, 1, 0, 0, 0]^T$
dog: loud, furry, ran, bit $\Rightarrow [1, 0, 1, 1, 0, 0, 1, 0]^T$

$$sim(w_1, w_2) = \frac{<w_1, w_2>}{||w_1|| \cdot ||w_2||}$$

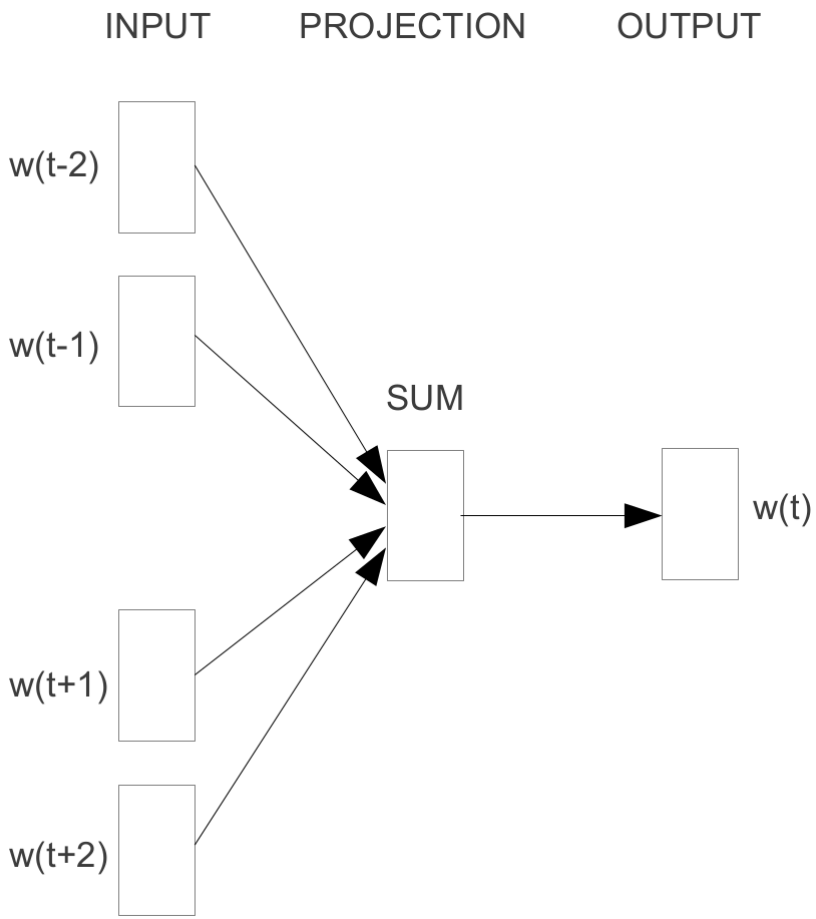
Embedding matrix

Матрица представлений:

$$E = \begin{bmatrix} —e_1— \\ —e_2— \\ \cdots\cdots\cdots \\ —e_{|V|}— \end{bmatrix}$$

Каждая строка — представление одного слова.
Идея: обучим матрицу E при помощи нейронной сети.

Continuous bag of words



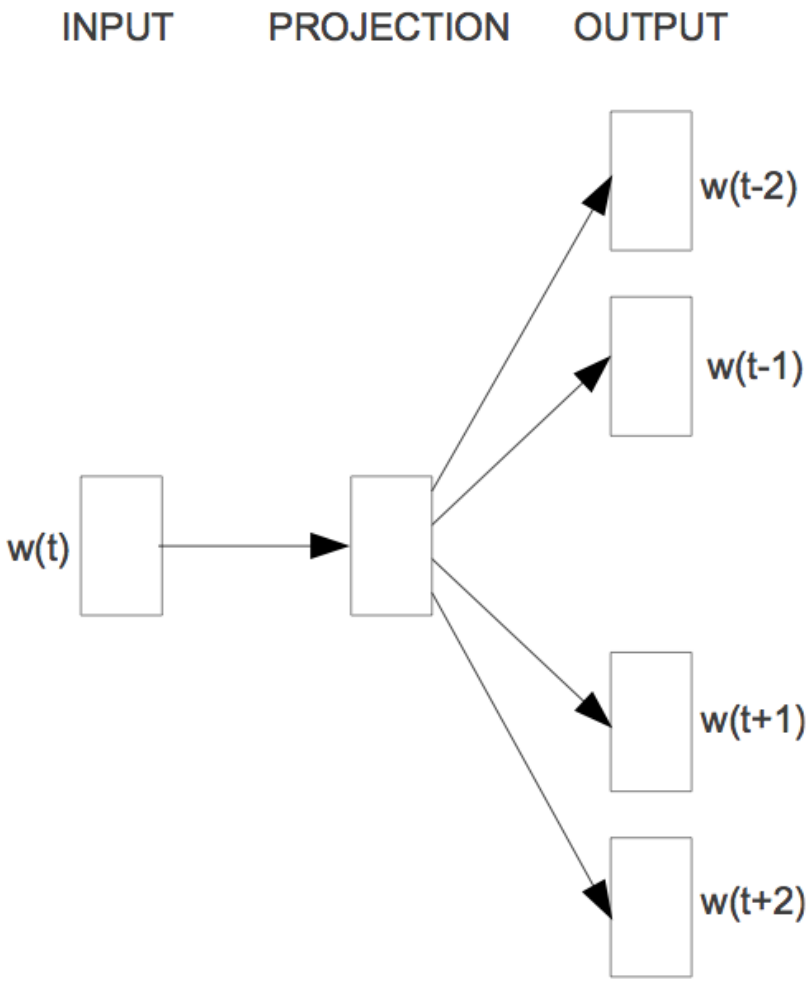
Предсказываем пропущенное слово по контексту.
Представление слова:

$$h_t = W \sum_{w \in \text{context}(w_t)} \text{one\_hot}(w)$$

$$P(w_i|\text{context}(w_i)) = \text{softmax}\left(W'h\right)[w_i]$$

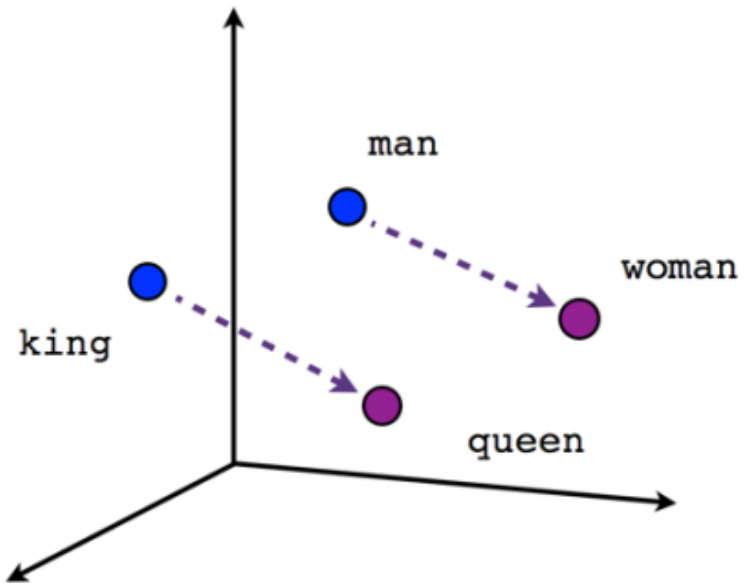
Функция потерь:
$L = -\log P(w_i|\text{context}(w_i))$

Skip-gram

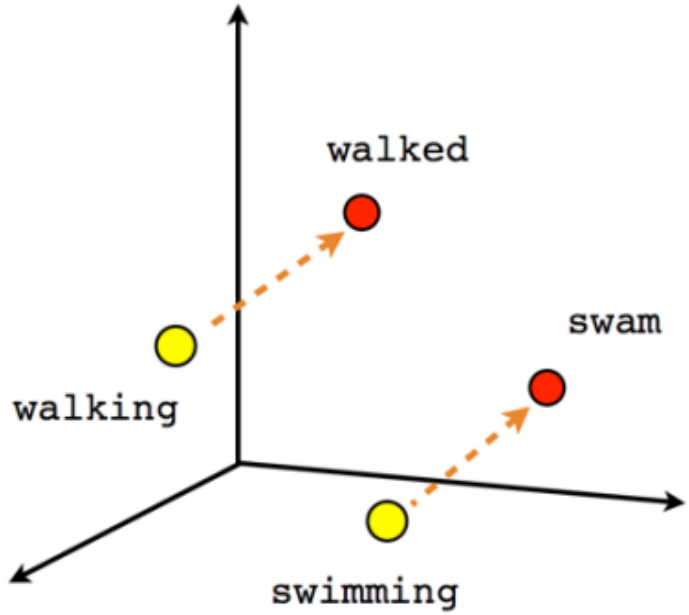


Предсказываем контекст по пропущенному слову

word2vec: арифметика



Male-Female



Verb tense

$$e(\text{«king»}) - e(\text{«man»}) + e(\text{«woman»}) \simeq e(\text{«queen»})$$
$$e(\text{«swimming»}) + e(\text{«walked»}) - e(\text{«walking»}) \simeq e(\text{«swam»})$$

Вероятностная постановка

Как решать такие разные задачи по NLP?

Вероятностная постановка

Как решать такие разные задачи по NLP?

По последовательности слов $w_1, w_2, \ldots, w_n$ надо найти распределение $P(w_1, w_2, \ldots, w_n)$.

Вероятностная постановка

Как решать такие разные задачи по NLP?

По последовательности слов $w_1, w_2, \ldots, w_n$ надо найти распределение $P(w_1, w_2, \ldots, w_n)$.

Chain rule: $P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i | w_1, w_2, \ldots, w_{i-1})$ Требуется научиться генерировать следующее слово по предыдущим.

N-grams

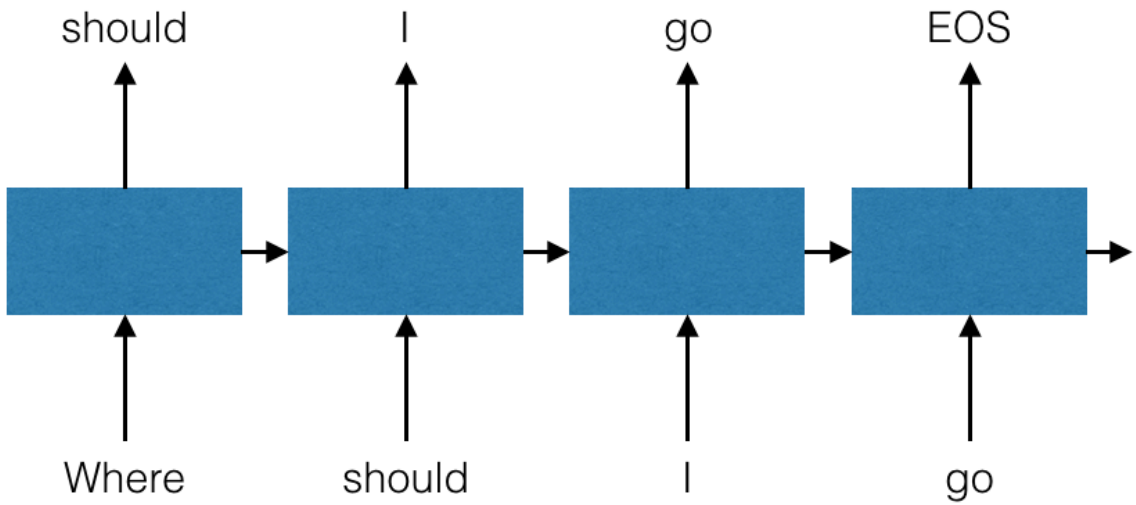
$$P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i | w_1, \ldots, w_{i-1}) \simeq$$

$$\prod_{i=1}^{n} P(w_i | w_{i-k}, \ldots, w_{i-1})$$

Обучение: считаем количество вхождений $w_{i-k}, \ldots, w_{i-1}, w_i$ и нормируем, чтобы получить вероятности.

- Чем больше k, тем более общая/переобученная модель
- Требуется много памяти
- Ограниченная длина контекста

Нейронные сети



$$P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i | w_1, \ldots, w_{i-1})$$

$$P(w_i | w_1, \ldots, w_{i-1}), c_i \simeq f(w_i | c_{i-1})$$

Где f — нейронная сеть

Seq2Seq

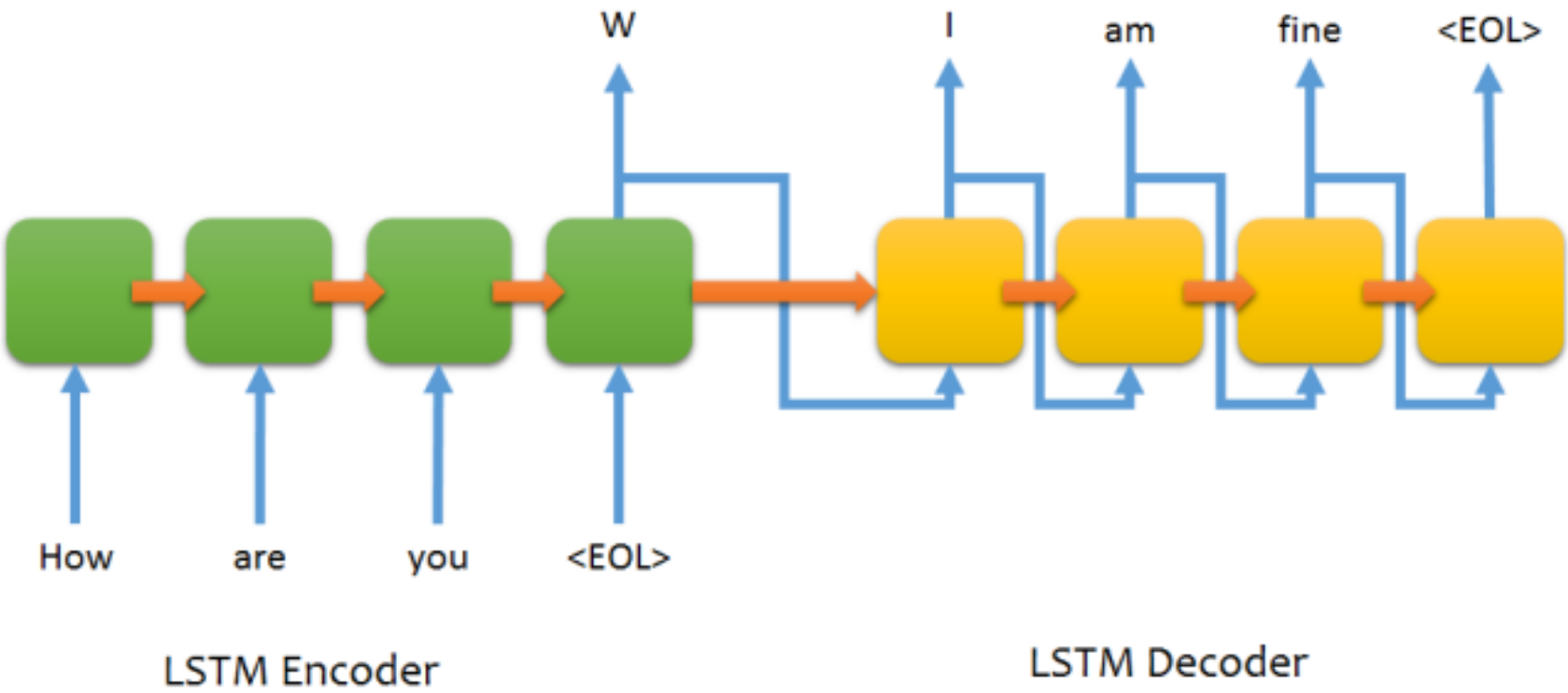


Figure: Перевод последовательностей друг в друга

Современные проблемы NLP

- Поиск ответа в тексте
- Суммаризация
- Генерация продолжения
- Заполнение пропущенных частей
- Ответ на вопросы

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	**90.115**	**92.580**
7 Mar 06, 2020	ELECTRA (single model) *Google Brain & Stanford*	88.716	91.365
8 Feb 24, 2020	ALBERT (Single model) *SRCB_DML*	88.592	91.286
8 Sep 16, 2019	ALBERT (single model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	88.107	90.902
8 Jul 26, 2019	UPM (ensemble) *Anonymous*	88.231	90.713
8 Feb 10, 2020	SkERT-Large (single model) *Skelter Labs*	87.994	90.944
9 Nov 15, 2019	XLNet (single model) *Google Brain & CMU*	87.926	90.689

Пример: SQuAD

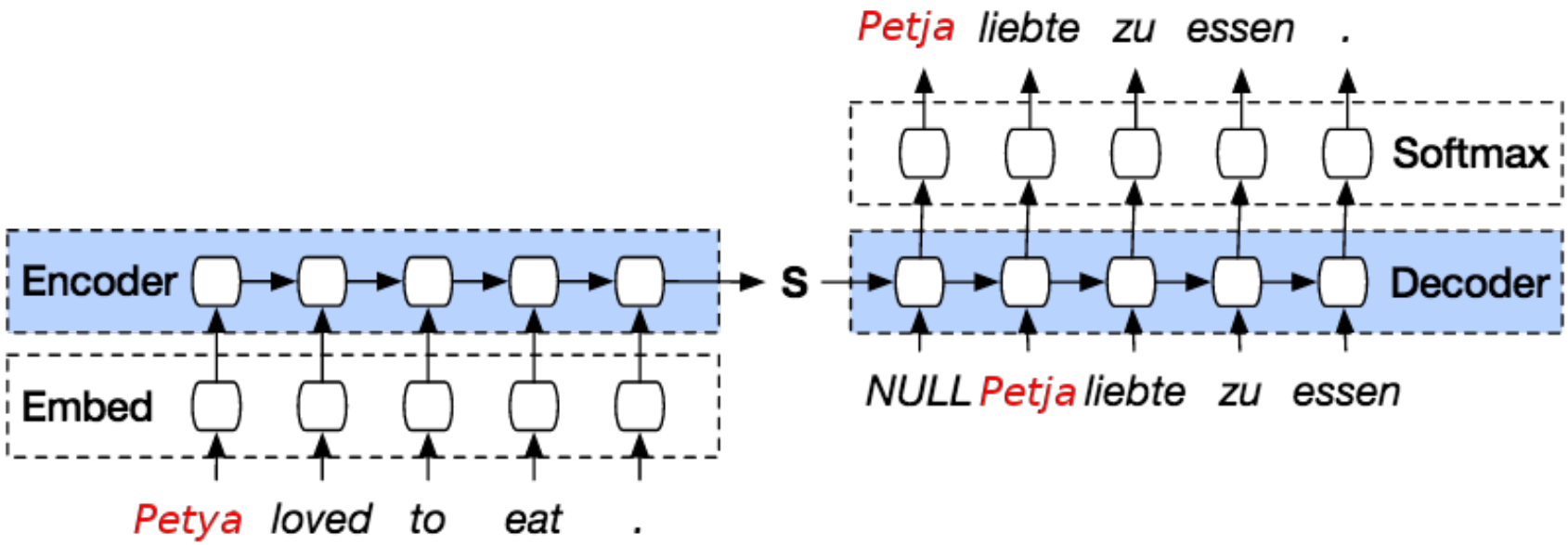
The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

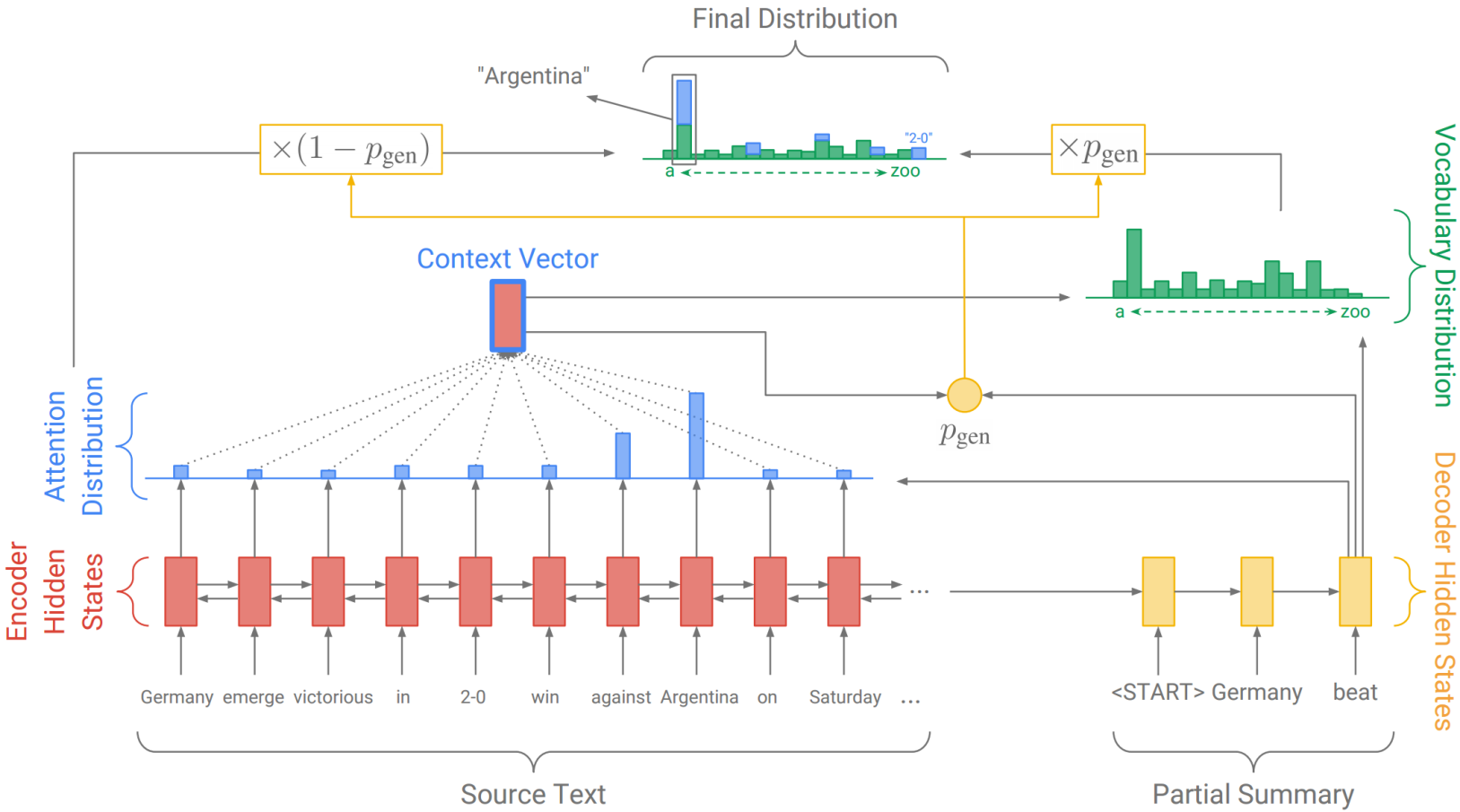
RNN

Проблема: надо запоминать точные сущности из текста, например, имена, названия, ..., также для цитирования надо запомнить точный текст, при ограниченном размере эмбеддинга это невозможно.



Attention[3]

Добавим Attention.



[3] https://arxiv.org/abs/1409.0473

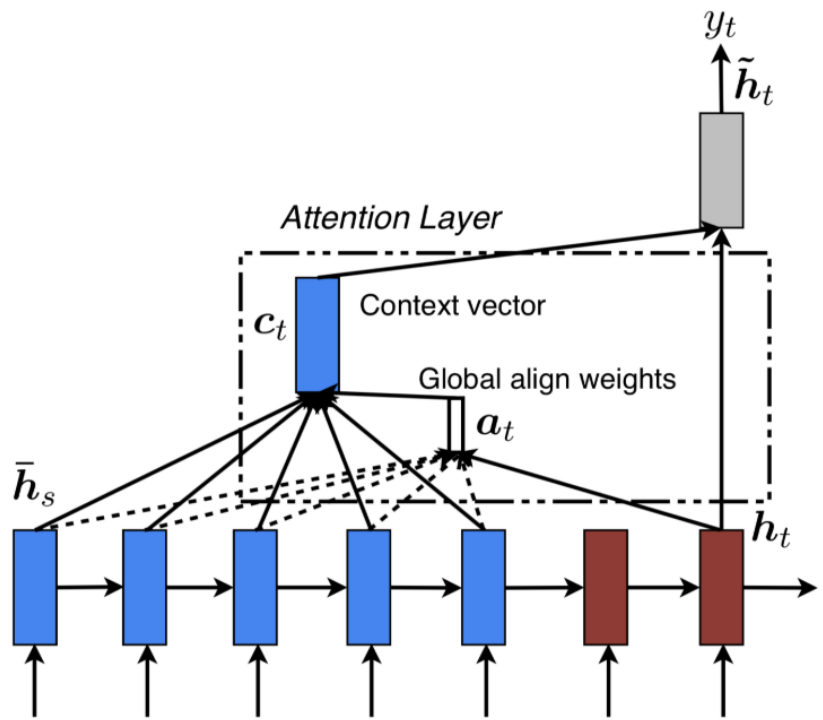
Attention

$$\text{decoder}_i = \text{RNN}(...)$$

$$\text{attention\_score}_{ij} = \text{softmax}_j(\text{attention}(\text{decoder}_i, \text{encoder\_output}_j))$$

$$\text{context}_i = \sum_j \text{attention\_score}_j \cdot \text{encoder\_output}_j$$

$$\text{decoder\_output}_i = \text{softmax}(f(\text{decoder}_i, \text{context}_i))$$



Attention to images

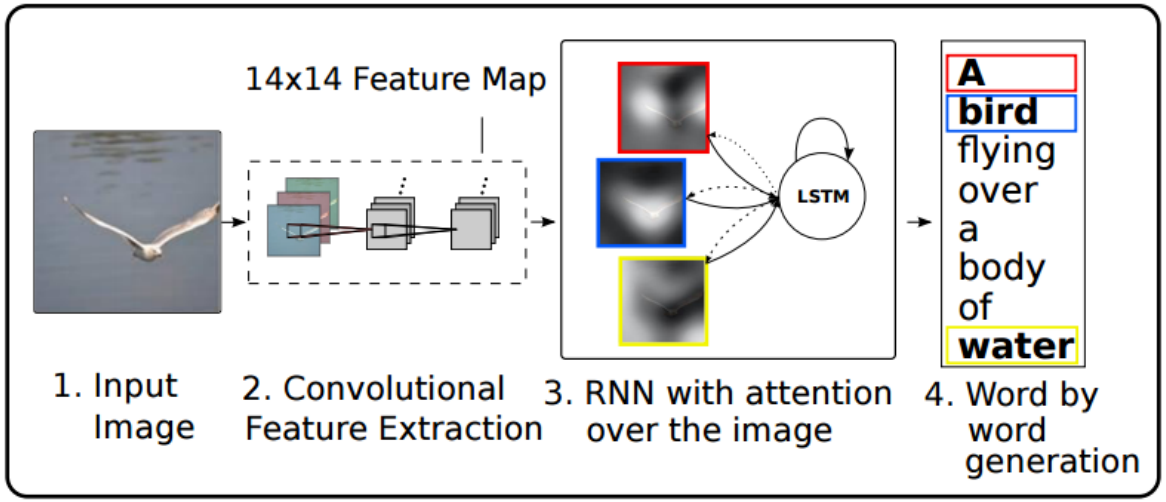
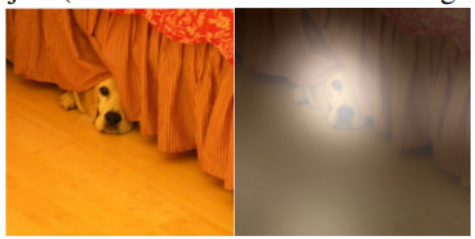


Figure 3. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)

A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

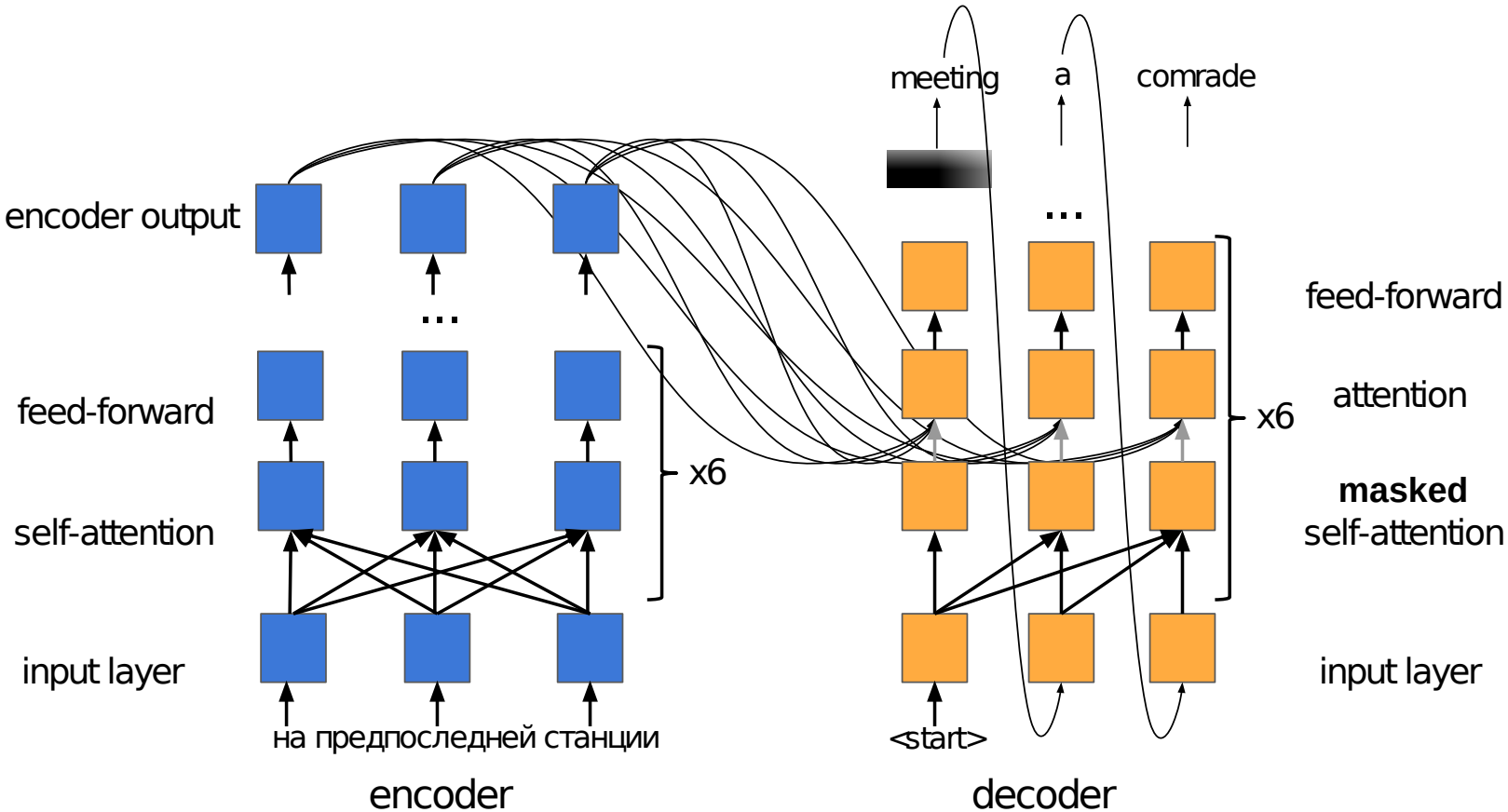
A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

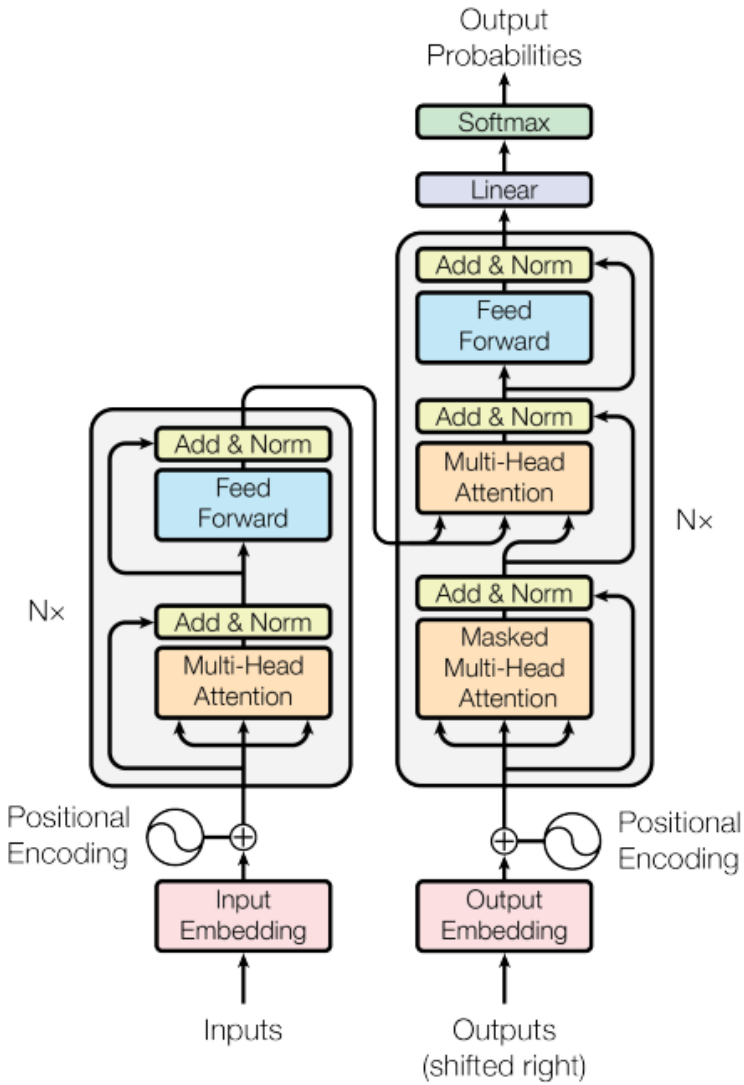
Transformer (Attention is all you need)[4]

Визуализация (gif)

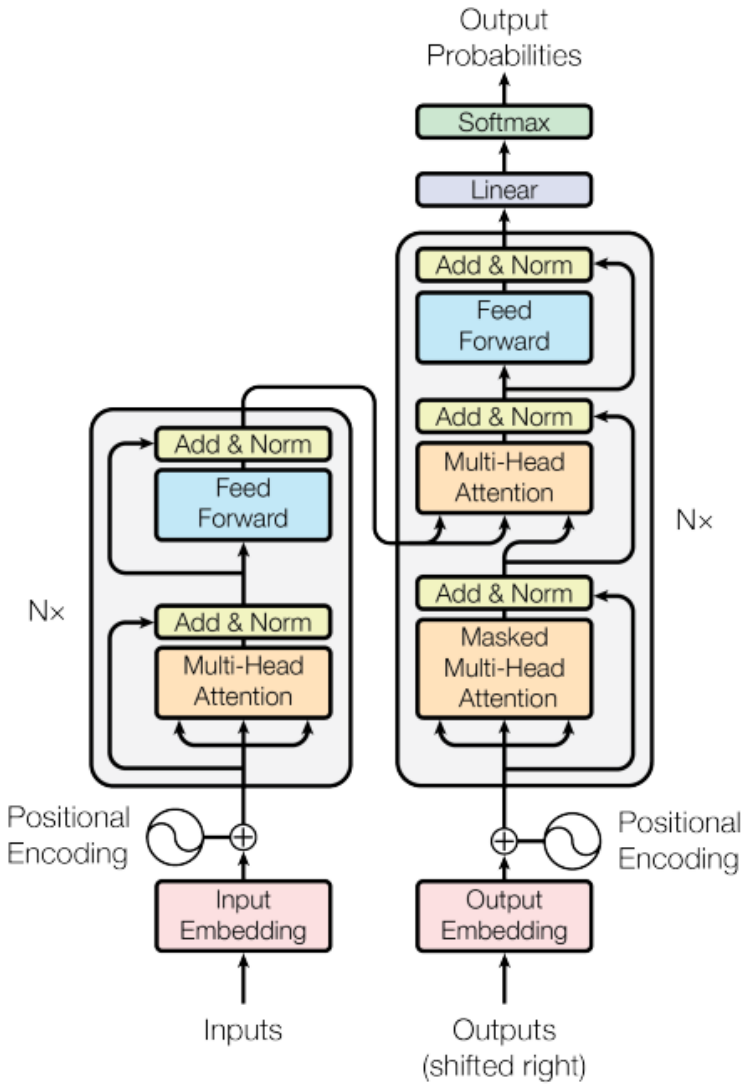


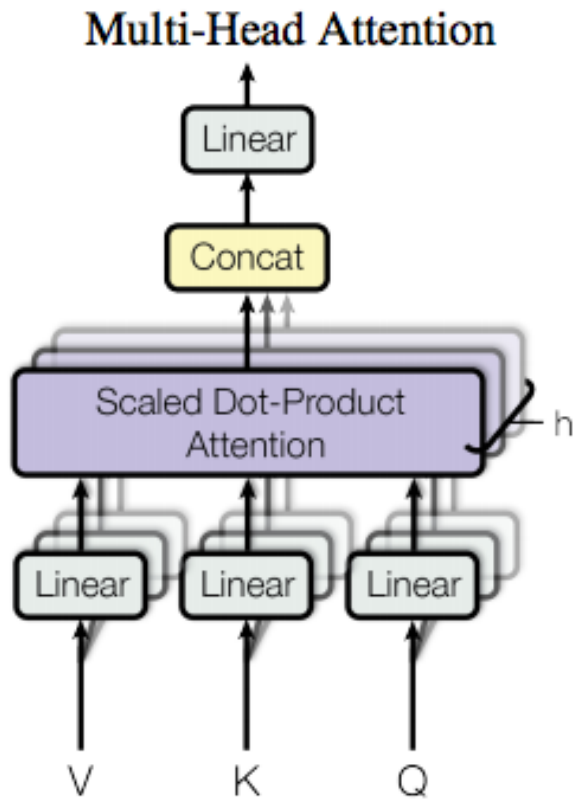
[4] https://arxiv.org/abs/1706.03762

Transformer (Attention is all you need)



Transformer (Attention is all you need)





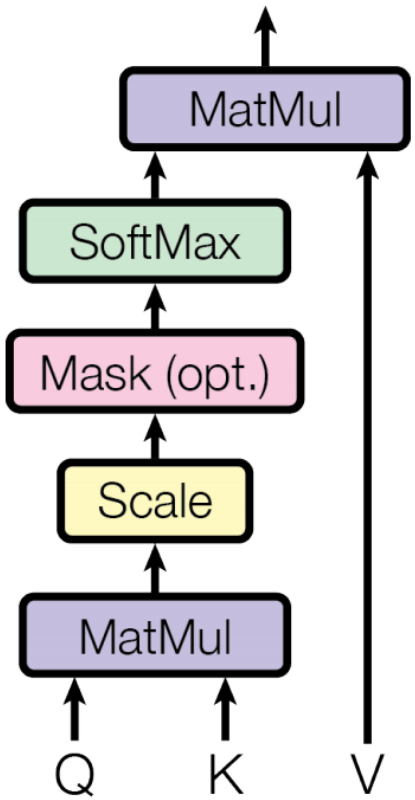
Attention (Attention is all you need)

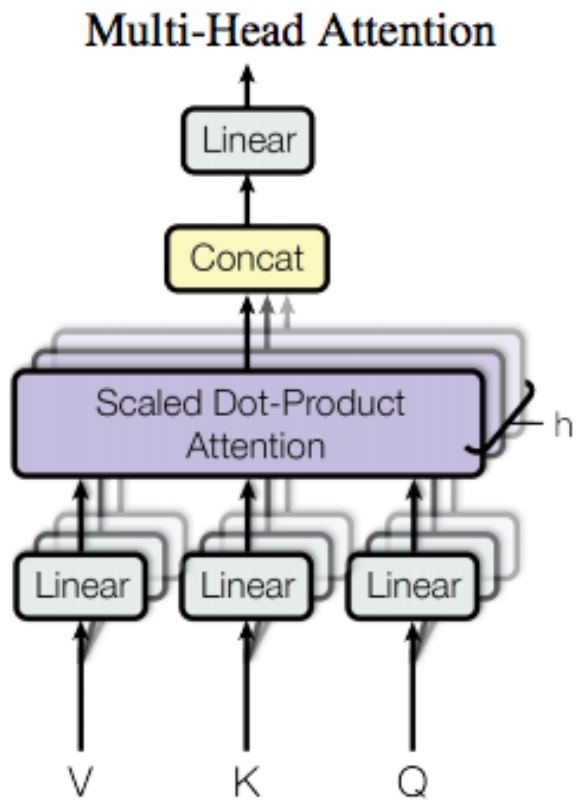
Одна голова: $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$.

Параллелим: $\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$

Где $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$

One Head Attention

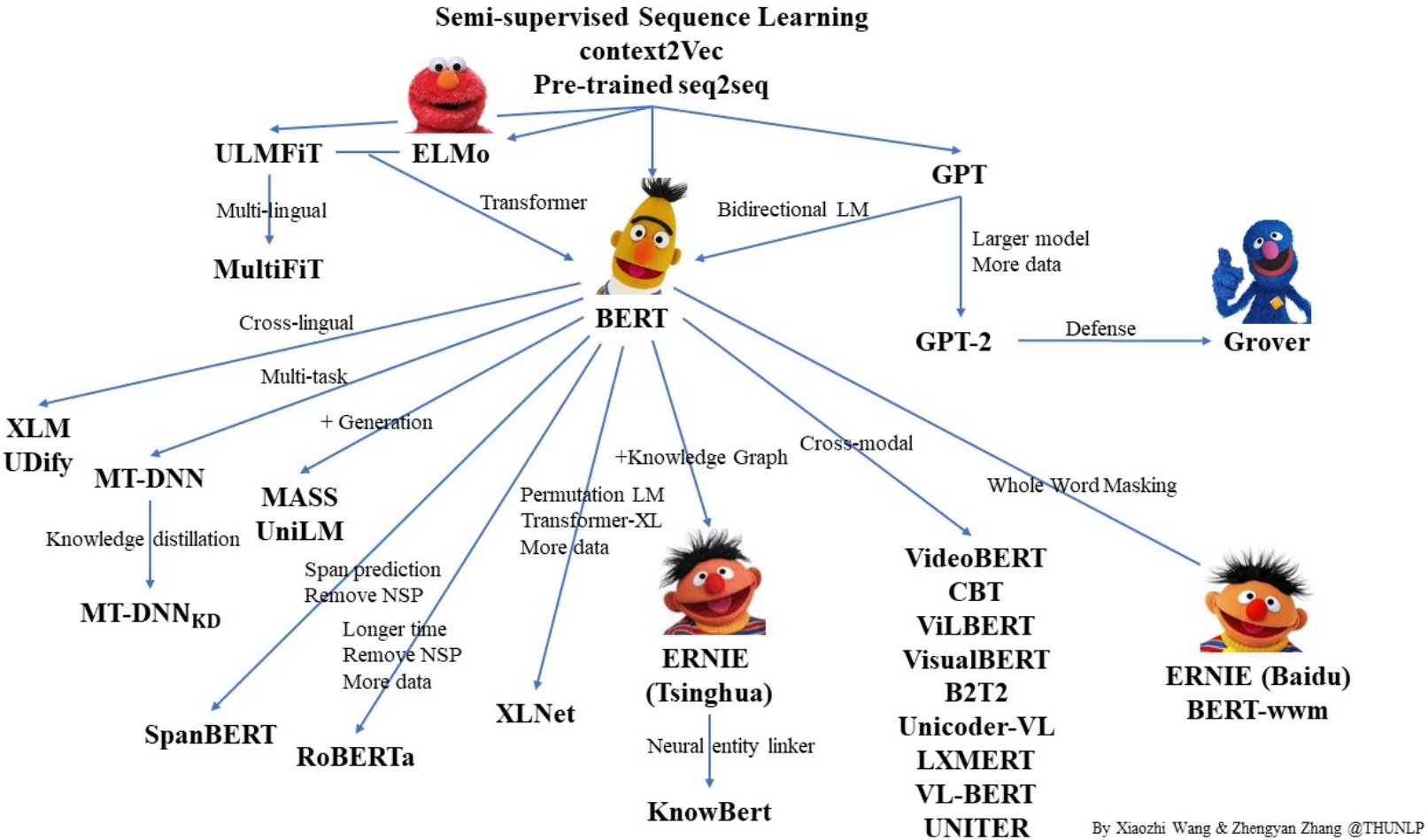




Transformer (Attention is all you need)

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**	$2.3 \cdot 10^{19}$	

Transformer Family

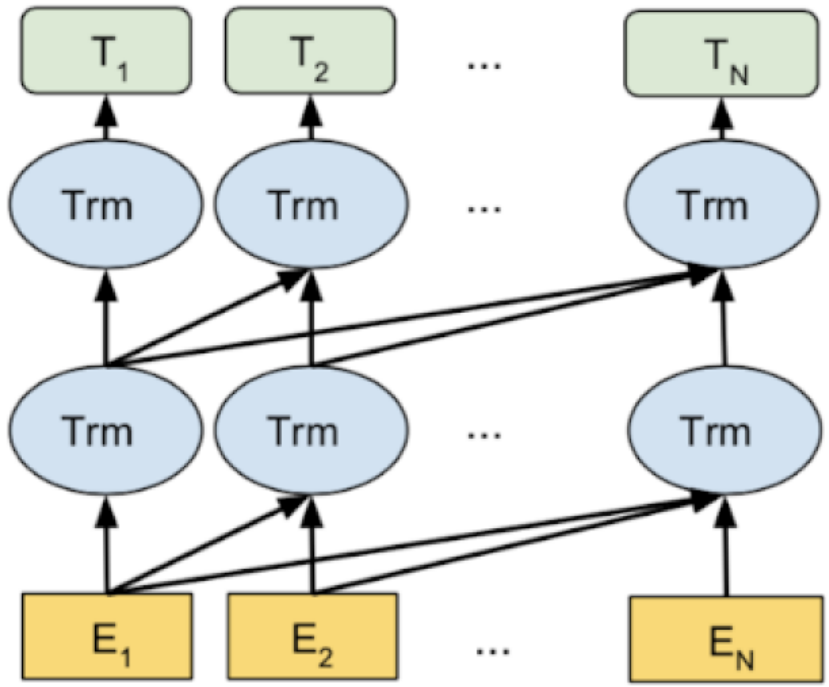


GPT[5]

Language model:

$$P(w_{n+1}|w_1, \ldots, w_n) = f(w_1, \ldots, w_n)$$

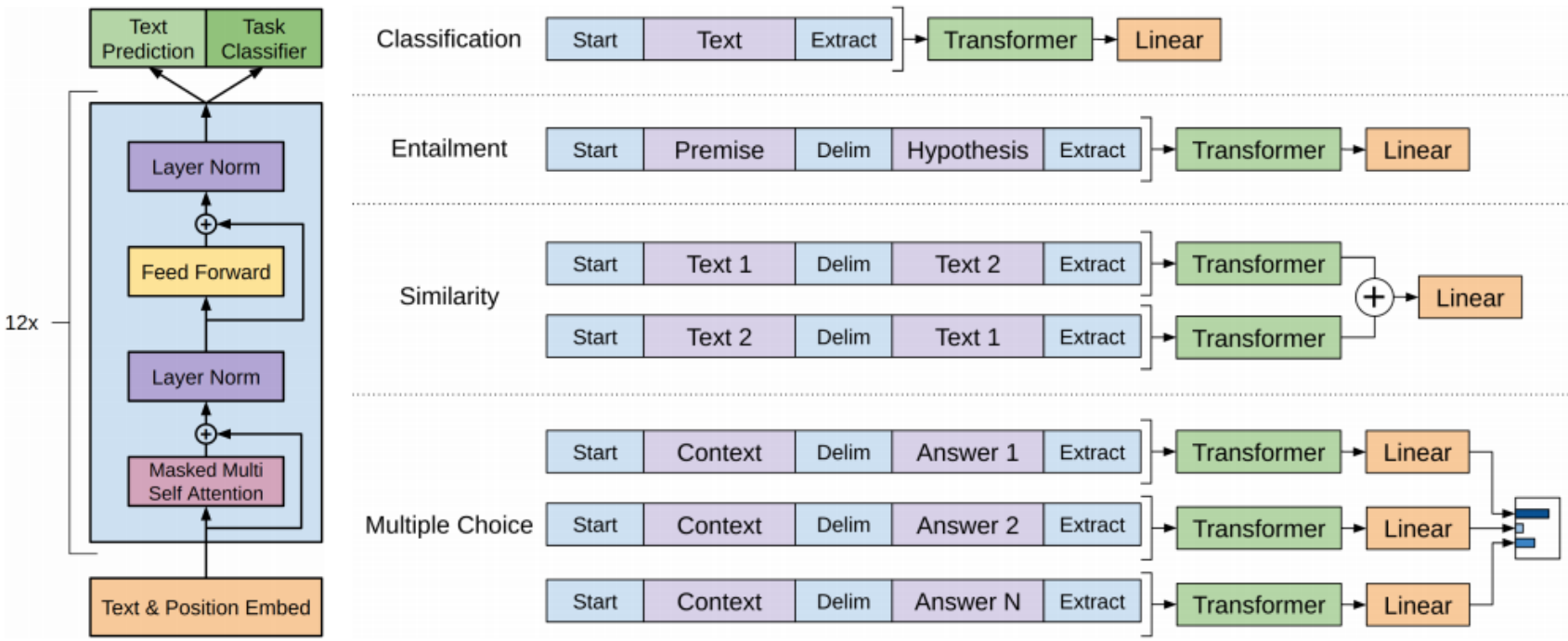
В случае GPT используются слои трансформера.
Также для ускорения обучения мы можем за один проход учить сразу несколько предсказаний за счет масок. (каждая позиция может "смотреть" только назад)



[5] https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

Multitask pretraining

Предобучаем модель на огромном корпусе $\rightarrow$
fine-tune на конкретную задачу.



GPT-2[6] (TL;DR)

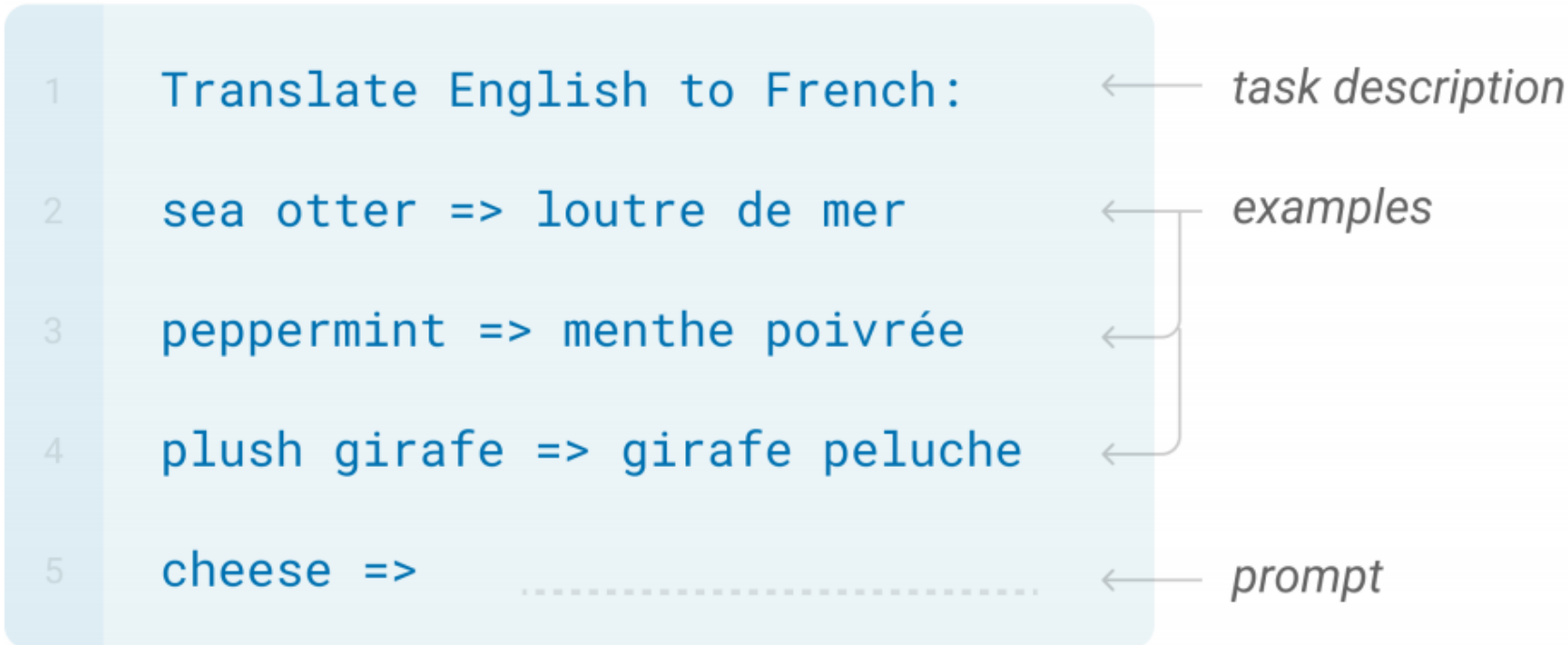
SYSTEM PROMPT (HUMAN-WRITTEN)	*A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.*
MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)	The incident occurred on the downtown train line, which runs from Covington and Ashland stations. In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief. "The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation," said Tom Hicks, the U.S. Energy Secretary, in a statement. "Our top priority is to secure the theft and ensure it doesn't happen again." The stolen material was taken from the University of Cincinnati's Research Triangle Park nuclear research site, according to a news release from Department officials. The Nuclear Regulatory Commission did not immediately release any information. According to the release, the U.S. Department of Energy's Office of Nuclear Material Safety and Security is leading that team's investigation. "The safety of people, the environment and the nation's nuclear stockpile is our highest priority," Hicks said. "We will get to the bottom of this and make no excuses.

[6] https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

GPT-3[7] (zero-shot learning!)

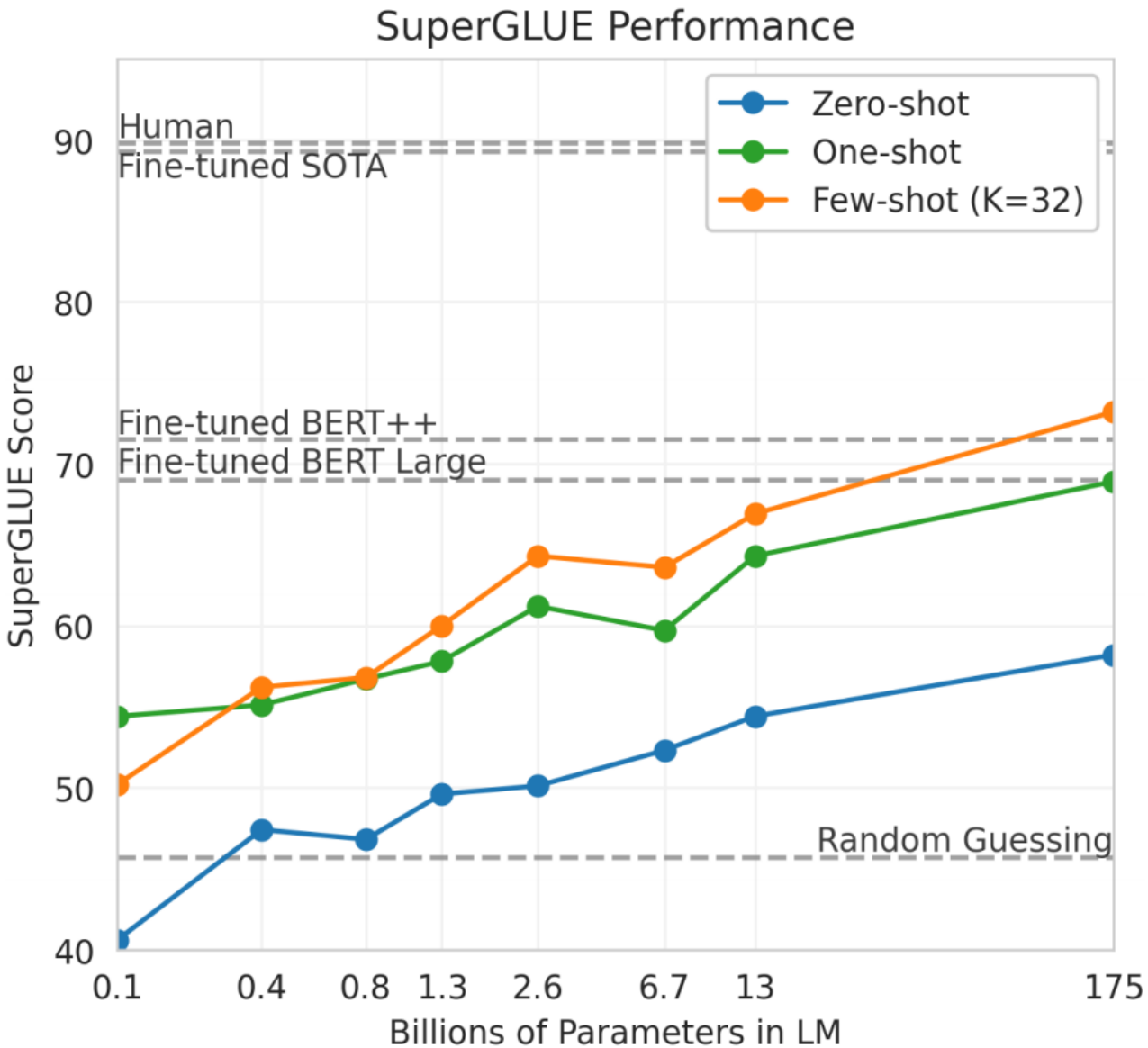
Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

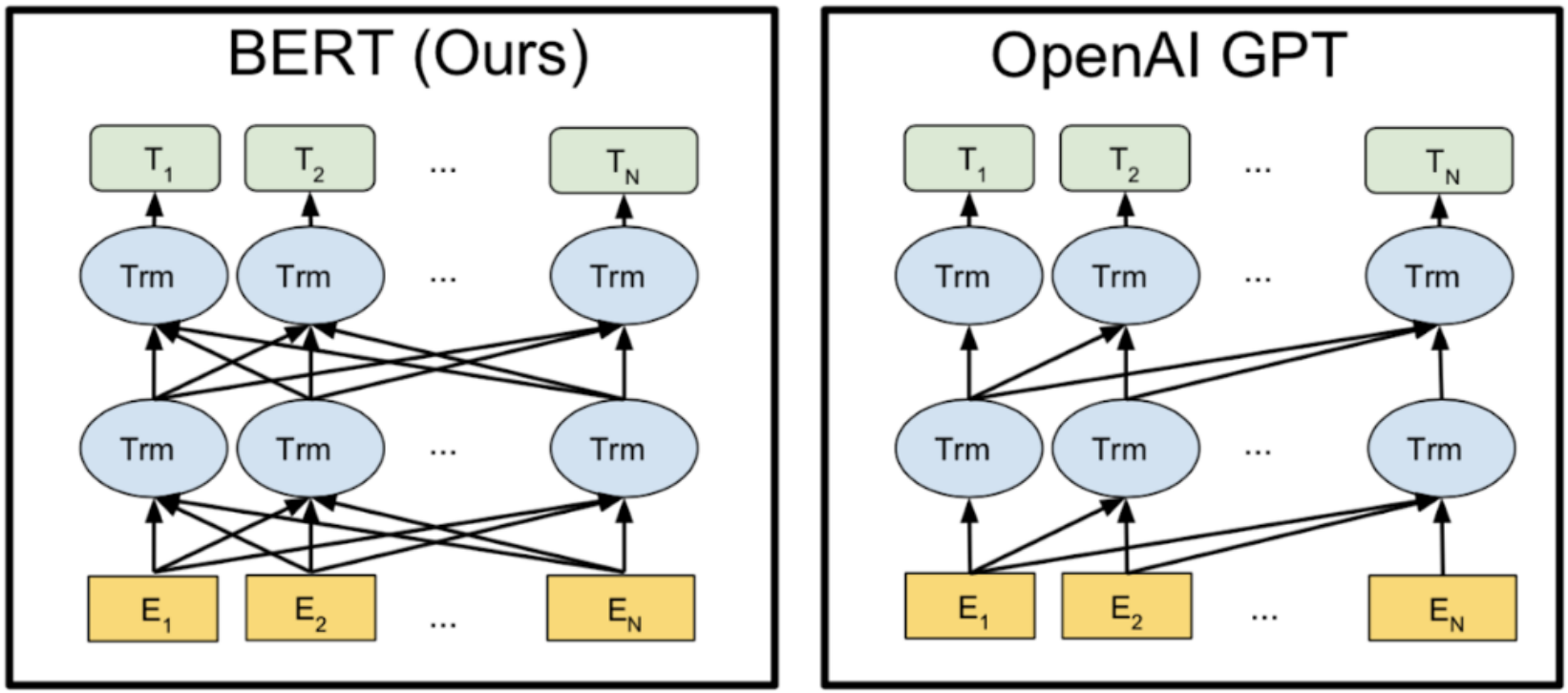


[7] https://arxiv.org/abs/2005.14165

GPT-3 (zero-shot learning!)

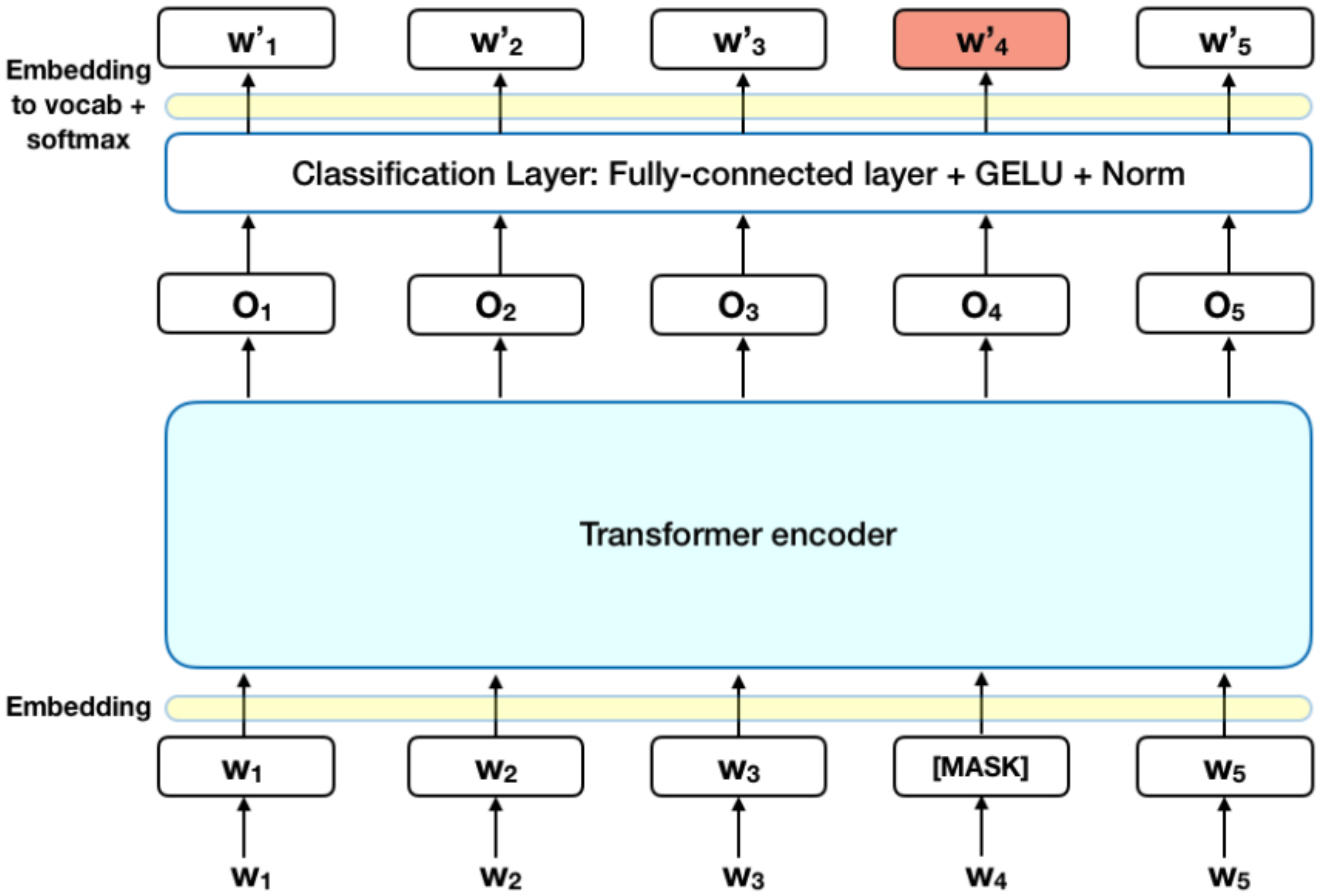


BERT vs GPT


BERT (Ours)
T1
T2
...
TN
Trm
Trm
...
Trm
Trm
Trm
...
Trm
E1
E2
...
EN
OpenAI GPT
T1
T2
...
TN
Trm
Trm
...
Trm
Trm
Trm
...
Trm
E1
E2
...
EN

BERT[8]

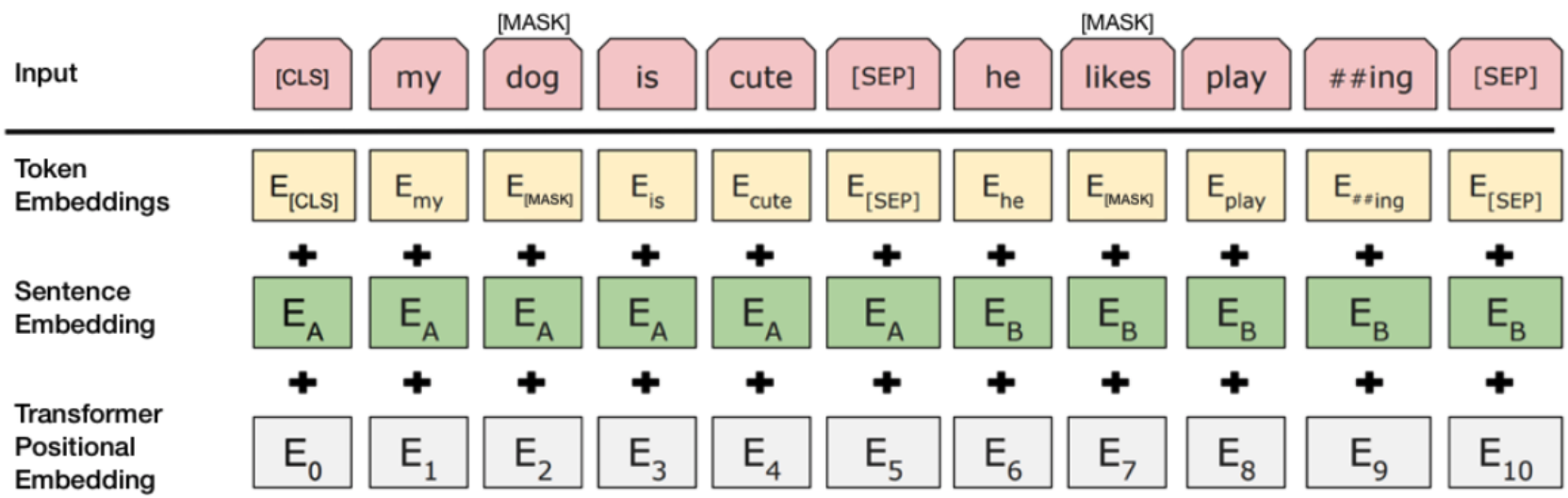
Bidirectional Encoder Representations from Transformers



[8] https://arxiv.org/abs/1810.04805

BERT - детали

- MASK - некоторые слова заменяем на токен неизвестного слова и пытаемся их восстановить.
- NSP - Для пары предложений пытаемся предсказать, правда ли, что B следует за A. (берем B случайно в 50% случаев)
  Нужно для улучшения модели языка и вопросно-ответных задач.
- Обучающее множество включает всю английскую википедию и книги не защищенные авторским правом. Для большой модели надо 4 дня на 16-и cloud TPU.



BERT - SOTA

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
Published				
BiDAF+ELMo (Single)	-	85.6	-	85.8
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

(a) BERT на SQuAD v1.0 (найти сегмент с ответом)

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
OpenAI GPT	-	78.0
$\text{BERT}_{\text{BASE}}$	81.6	-
$\text{BERT}_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

(b) BERT на SWAG (выбор из нескольких вариантов ответа)

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
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

RoBERTa[9]

"We find that BERT was significantly undertrained, and can match or exceed the performance of every model published after it."

[9] https://arxiv.org/abs/1907.11692

RoBERTa[9]

"We find that BERT was significantly undertrained, and can match or exceed the performance of every model published after it."

- Объём - BERT обучался на 16GB текстов, мы будем учить на 160GB (включая датасет "хороших сайтов" GPT-2).
- NSP - учиться лучше на больших отрезках текста (параграфах, а не парах предложений), NSP не нужен! (без него на итоговых задачах не хуже, а иногда и лучше)
- Размер батча - оригинальный BERT учился на 256 примерах за раз, в работе показано, что лучше будет брать намного больший батч, например, 8K. (тут это был предел технических возможностей, есть работы, в которых увеличивали вплоть до 32K)
- RoBERTa - Robustly optimized BERT approach.

[9] https://arxiv.org/abs/1907.11692

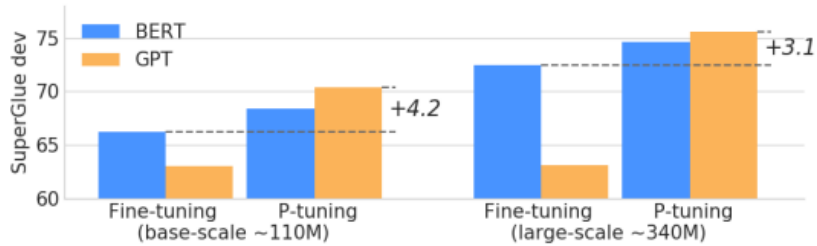
P-tuning[10]

Prompt	P@1
[X] is located in [Y]. *(original)*	31.29
[X] is located in which country or state? [Y].	19.78
[X] is located in which country? [Y].	31.40
[X] is located in which country? In [Y].	51.08

[10] https://arxiv.org/abs/2103.10385, https://arxiv.org/abs/2110.07602

P-tuning[10]

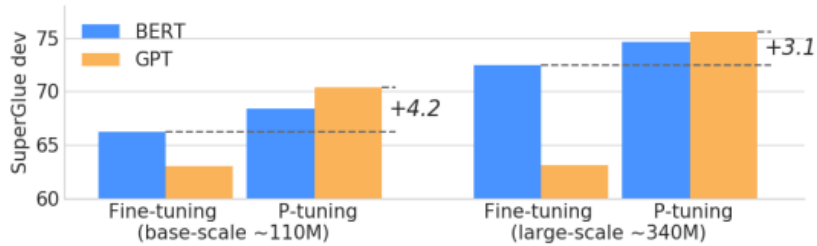
Prompt	P@1
[X] is located in [Y]. *(original)*	31.29
[X] is located in which country or state? [Y].	19.78
[X] is located in which country? [Y].	31.40
[X] is located in which country? In [Y].	51.08

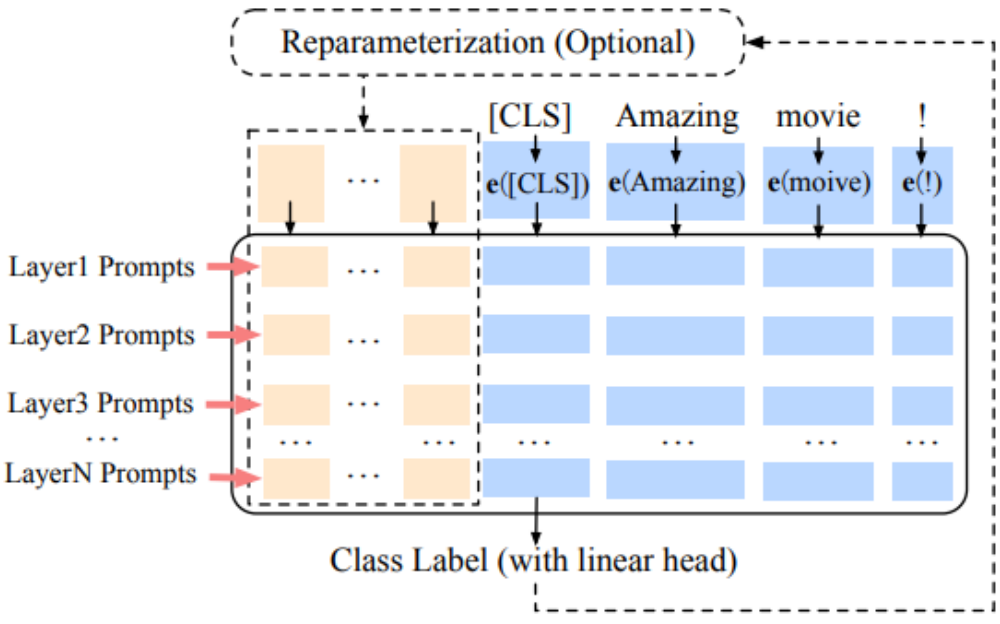


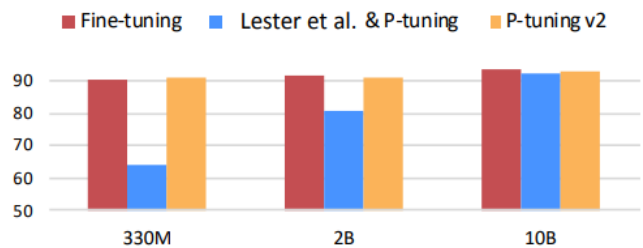
[10] https://arxiv.org/abs/2103.10385, https://arxiv.org/abs/2110.07602

P-tuning[10]

Prompt	P@1
[X] is located in [Y]. *(original)*	31.29
[X] is located in which country or state? [Y].	19.78
[X] is located in which country? [Y].	31.40
[X] is located in which country? In [Y].	51.08

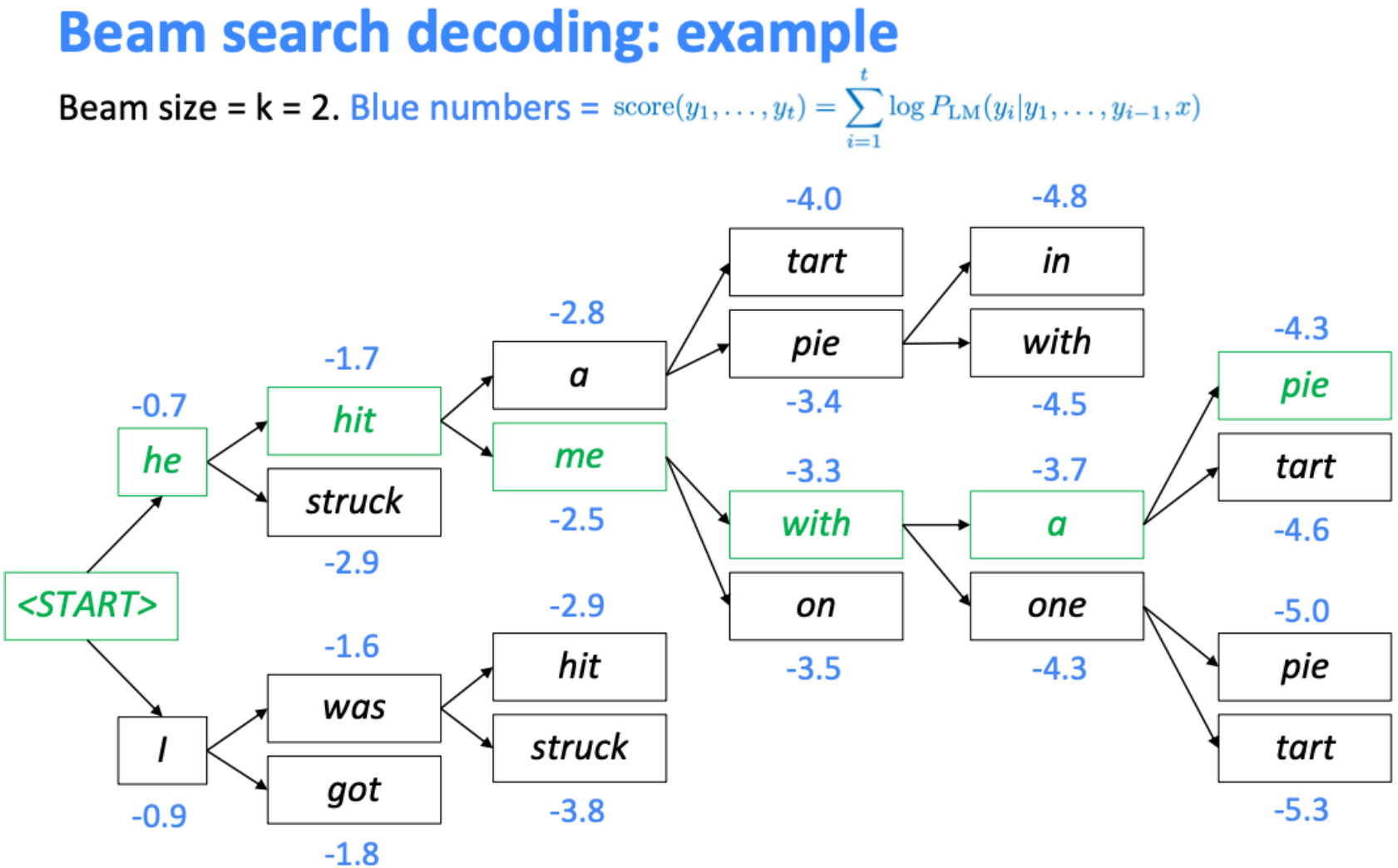




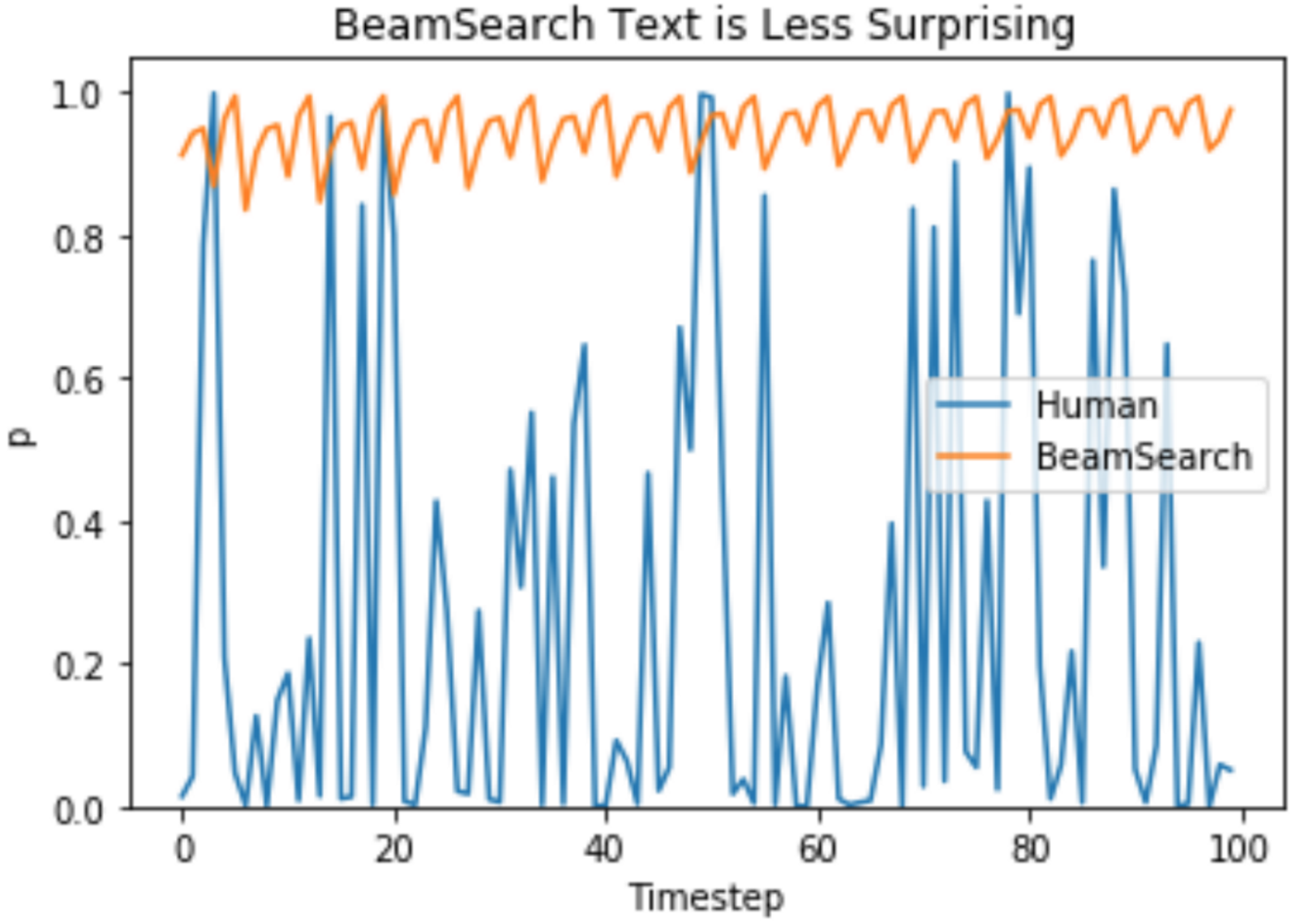


[10] https://arxiv.org/abs/2103.10385, https://arxiv.org/abs/2110.07602

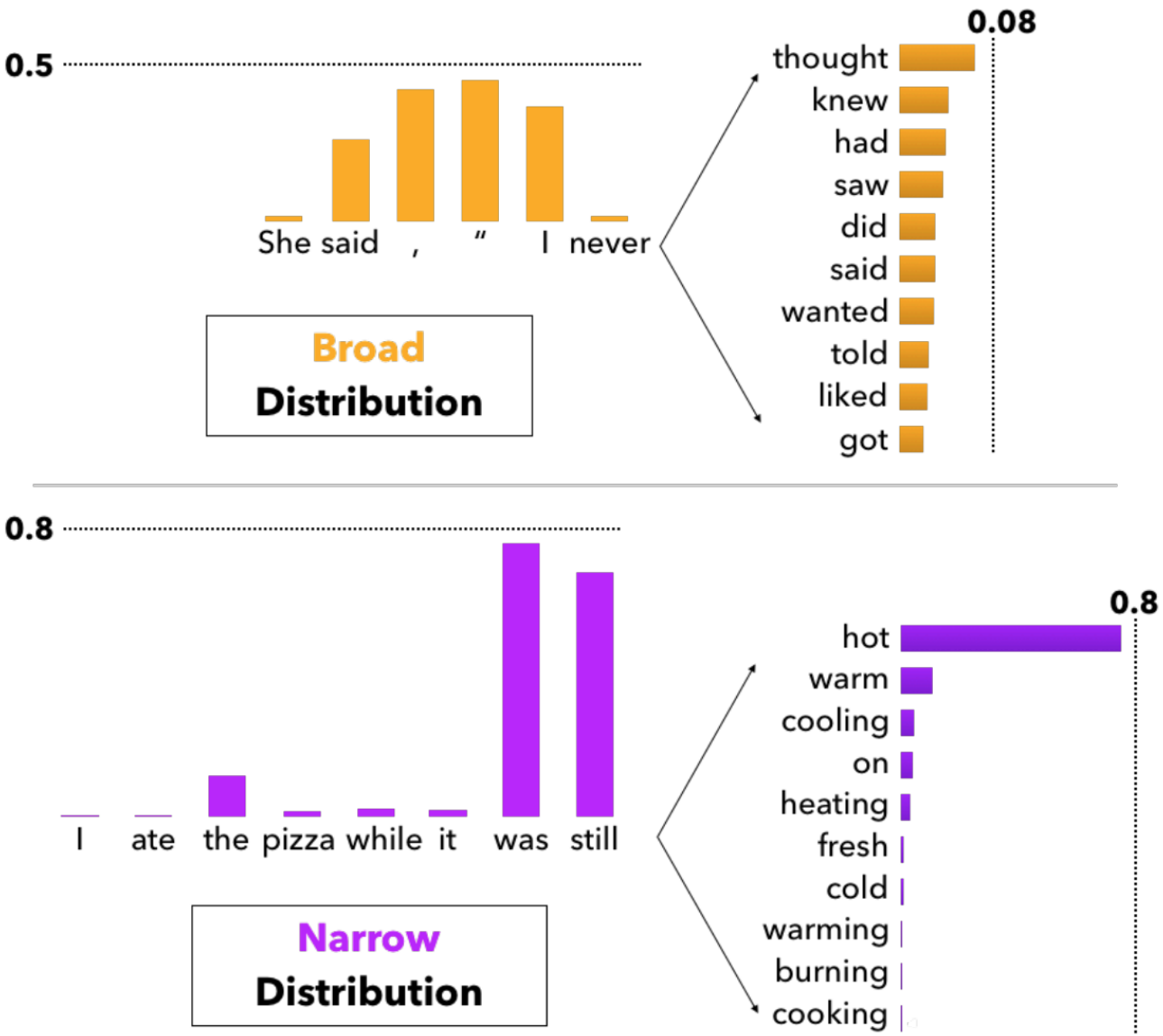
Beam search


Beam search decoding: example
Beam size = k = 2. Blue numbers = $\text{score}(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$
<START>
he
-0.7
I
-0.9
hit
-1.7
struck
-2.9
was
-1.6
got
-1.8
a
-2.8
me
-2.5
hit
-2.9
struck
-3.8
tart
-4.0
pie
-3.4
with
-3.3
on
-3.5
in
-4.8
with
-4.5
a
-3.7
one
-4.3
pie
-4.3
tart
-4.6
pie
-5.0
tart
-5.3

Beam search problems

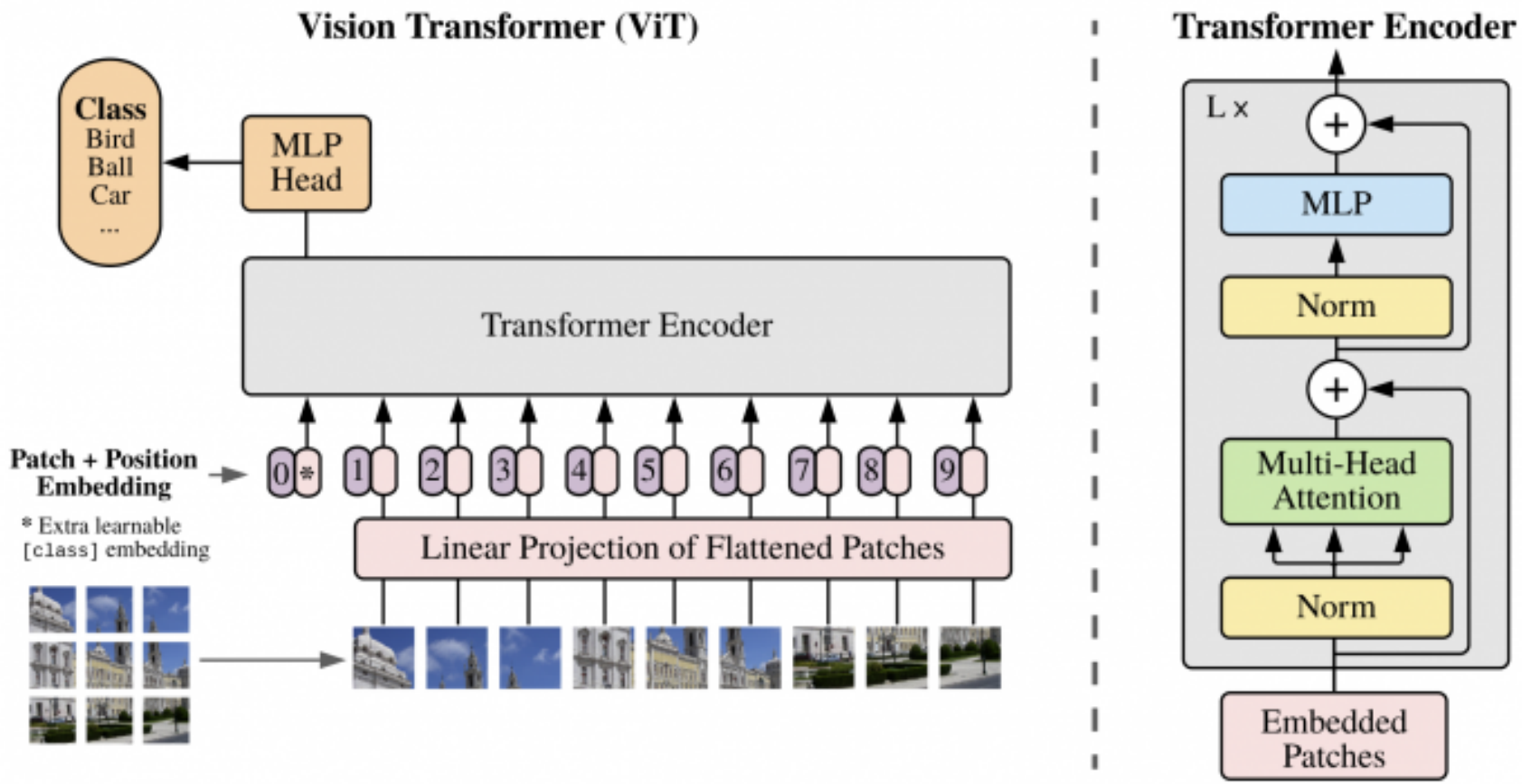


Top p and top k sampling[11], nukleos and typical sampling



[11] https://towardsdatascience.com/how-to-sample-from-language-models-682bceb97277

ViT


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches
Transformer Encoder
L x
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

За рамками лекции

- Различные виды BPE
- Distillation
- Технические подробности обучения огромных моделей
- Некоторые другие улучшения трансформеров
- Модели для длинных последовательностей. Transformer XL, Reformer, Sparse Transformer, BigBird...
- ViT outcomes

Вопросы