# Transformers

Сергей Николенко

НИУ ВШЭ — Санкт-Петербург

14 ноября 2020 г.

---

*Random facts*:

- 14 ноября 1901 г. венский врач Карл Ландштейнер разделил все образцы крови на три группы: A, B и 0
- 14 ноября 1925 г. в Париже открылась выставка искусства сюрреалистов, включающая работы Макса Эрнста, Мана Рэя, Хуана Миро и Пабло Пикассо
- 14 ноября 1916 г. Павел Милюков на открытии осенней сессии Думы произнёс знаменитое: «Что это? Глупость или измена?»
- 14 ноября 1981 г. Гамбия и Сенегал создали федеративное государство Сенегамбия; оно просуществовало до 1989 года

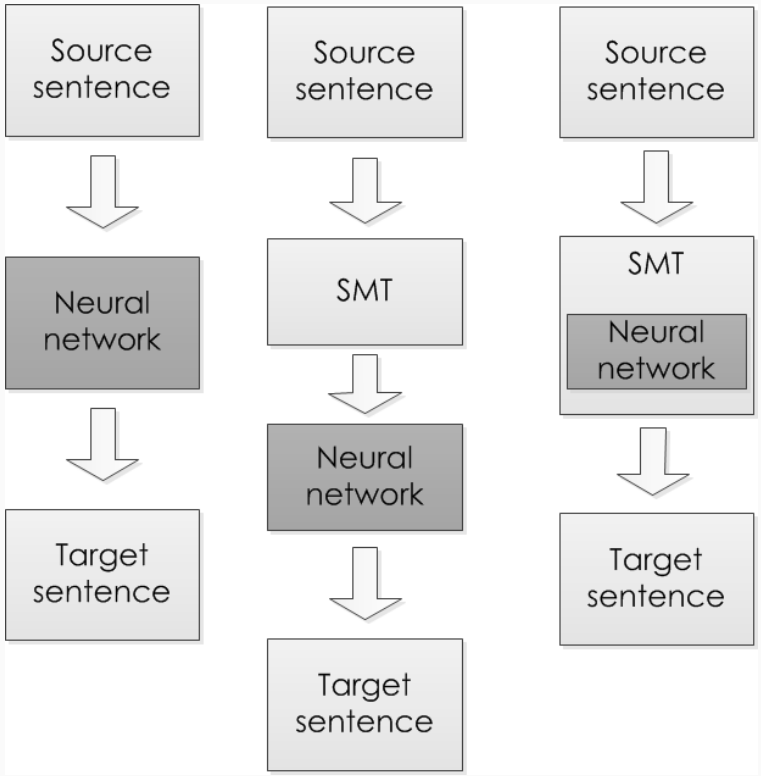
# Машинный перевод: encoder-decoder и внимание

# Машинный перевод

- Перевод – очень хорошая задача:
  - очевидно очень практическая;
  - очевидно очень высокоуровневая, требует понимания;
  - считается довольно неплохо квантифицируемой (BLEU, TER – хотя см. выше);
  - имеет большие доступные датасеты параллельных переводов.

# Машинный перевод

- Статистический машинный перевод (statistical machine translation, SMT): моделируем условную вероятность $p(y \mid x)$ перевода $y$ при условии исходного текста $x$.
- Классический SMT: моделируем $\log p(y \mid x)$ линейной комбинацией признаков, строим признаки.



# Машинный перевод

- Нам больше интересно моделирование sequence-to-sequence:
  - RNN естественным образом моделирует последовательность $X = (x_1, x_2, \ldots, x_T)$ как $p(x_1)$, $p(x_2 \mid x_1)$, …, $p(x_T \mid x_{<T}) = p(x_T \mid x_{T-1}, \ldots, x_1)$, и теперь $p(X)$ – это просто

$$p(X) = p(x_1)p(x_2 \mid x_1) \ldots p(x_k \mid x_{<k}) \ldots p(x_T \mid x_{<T});$$

  - так RNN и в языковых моделях используются;
  - предсказываем следующее слово на основе скрытого состояния и предыдущего слова;
- Как применить эту идею к переводу?

# Метрики качества для sequence-to-sequence моделей

- Дальше будет самое интересное: машинный перевод, диалоговые модели, ответ на вопросы.
- Но как мы будем оценивать NLP-модели, которые генерируют текст?
- Есть метрики качества, которые сравнивают результат с правильными ответами:
  - BLEU (Bilingual Evaluation Understudy): перевзвешенная precision (в т.ч. для нескольких правильных ответов);
  - METEOR: гармоническое среднее precision и recall по униграммам;
  - TER (Translation Edit Rate): число исправлений между выходом и правильным ответом, делённое на среднее число слов;
  - LEPOR: комбинируем базовые факторы и метрики с настраиваемыми параметрами.
- Есть ещё куча метрик, связанных с представлениями слов и предложений (хотим поближе к правильному ответу).
- Одна только проблема...

# Метрики качества для sequence-to-sequence моделей

- ...всё это вообще не работает.

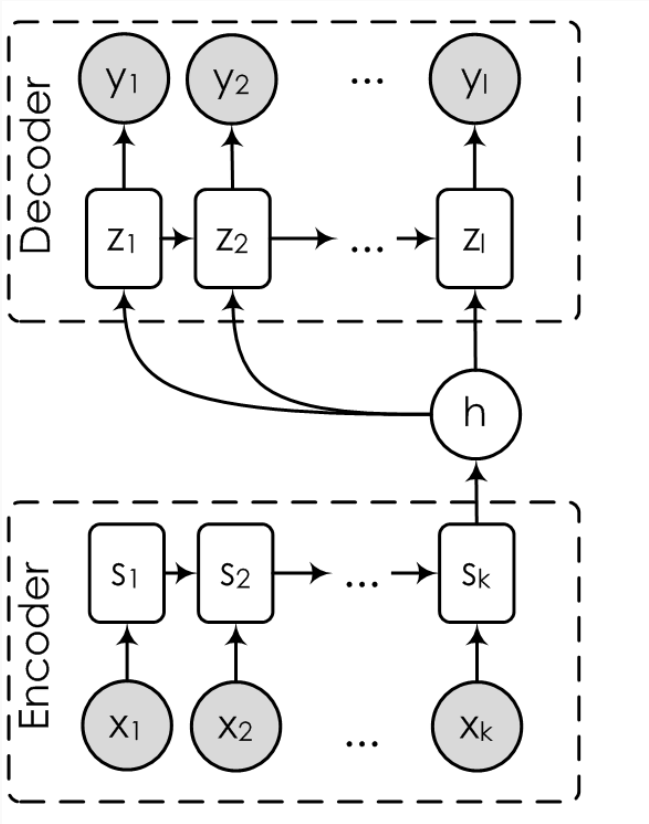
| | Twitter | | | | Ubuntu | | | |
|---|---|---|---|---|---|---|---|---|
| Metric | **Spearman** | p-value | **Pearson** | p-value | **Spearman** | p-value | **Pearson** | p-value |
| Greedy | 0.2119 | 0.034 | 0.1994 | 0.047 | 0.05276 | 0.6 | 0.02049 | 0.84 |
| Average | 0.2259 | 0.024 | 0.1971 | 0.049 | -0.1387 | 0.17 | -0.1631 | 0.10 |
| Extrema | 0.2103 | 0.036 | 0.1842 | 0.067 | 0.09243 | 0.36 | -0.002903 | 0.98 |
| METEOR | 0.1887 | 0.06 | 0.1927 | 0.055 | 0.06314 | 0.53 | 0.1419 | 0.16 |
| BLEU-1 | 0.1665 | 0.098 | 0.1288 | 0.2 | -0.02552 | 0.8 | 0.01929 | 0.85 |
| BLEU-2 | 0.3576 | $< 0.01$ | 0.3874 | $< 0.01$ | 0.03819 | 0.71 | 0.0586 | 0.56 |
| BLEU-3 | 0.3423 | $< 0.01$ | 0.1443 | 0.15 | 0.0878 | 0.38 | 0.1116 | 0.27 |
| BLEU-4 | 0.3417 | $< 0.01$ | 0.1392 | 0.17 | 0.1218 | 0.23 | 0.1132 | 0.26 |
| ROUGE | 0.1235 | 0.22 | 0.09714 | 0.34 | 0.05405 | 0.5933 | 0.06401 | 0.53 |
| Human | 0.9476 | $< 0.01$ | 1.0 | 0.0 | 0.9550 | $< 0.01$ | 1.0 | 0.0 |

Table 3: Correlation between each metric and human judgements for each response. Correlations shown in the human row result from randomly dividing human judges into two groups.

- Тут нужно что-то новое. И пока не совсем ясно, что именно.

# Encoder-decoder архитектуры

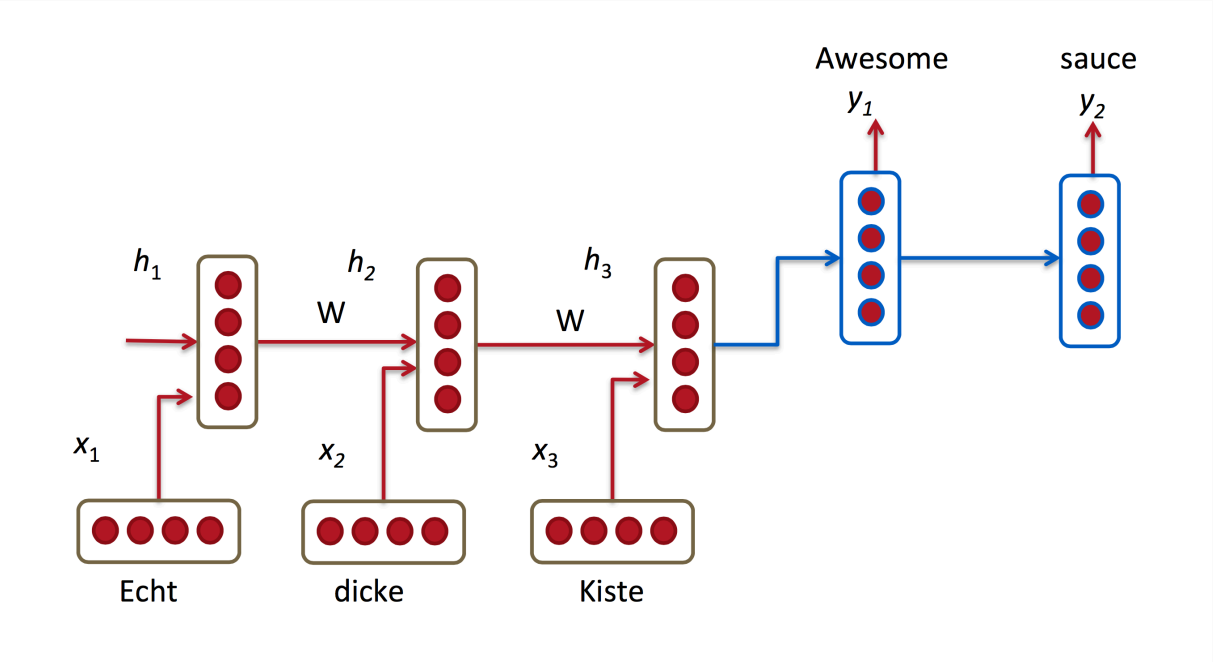
- Encoder-decoder архитектуры (Sutskever et al., 2014; Cho et al., 2014):




- Сначала кодируем, потом декодируем обратно.

# Encoder-decoder архитектуры

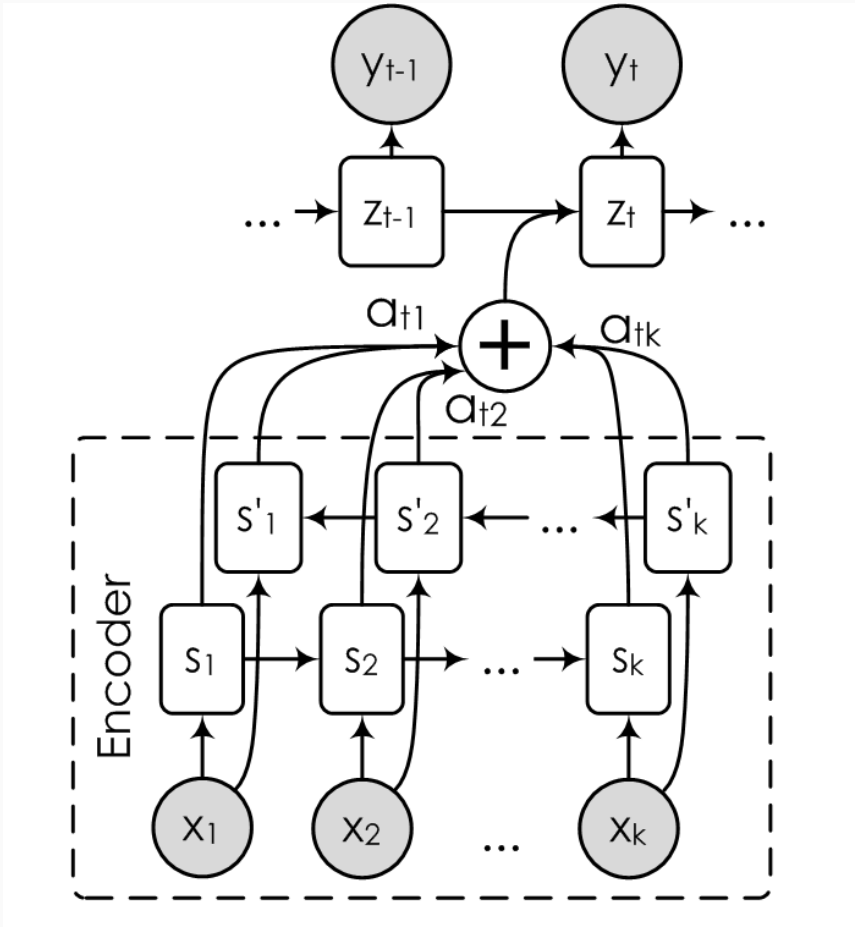
- Так же может работать и с переводом.




- Проблема: надо сжимать всё предложение в один вектор.
- С длинными участками текста это вообще перестаёт работать.

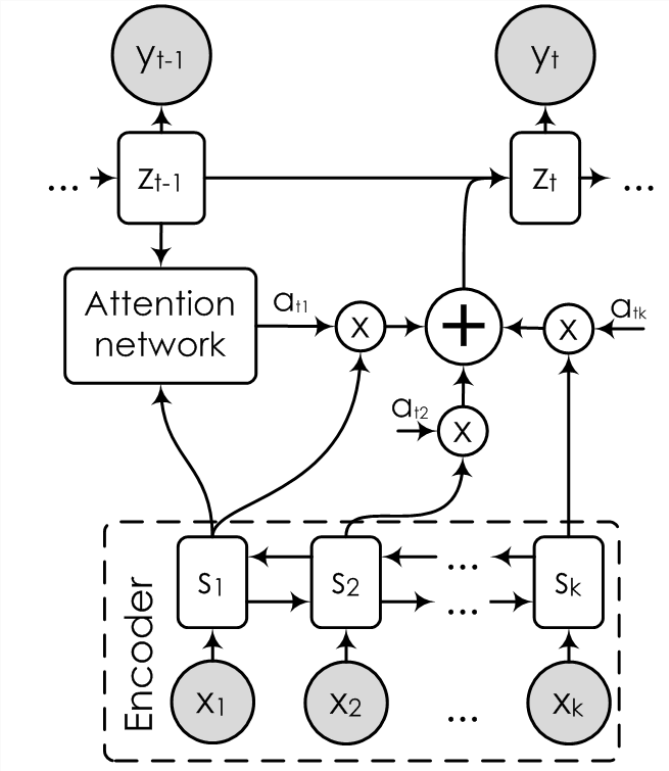
# Внимание в нейронных сетях

- Решение: давайте обучим специальные веса, показывающие, насколько та или иная часть входа важна для текущего выхода.
- Прямое применение – двунаправленный LSTM плюс внимание (Bahdanau et al. 2014):



# Внимание в нейронных сетях

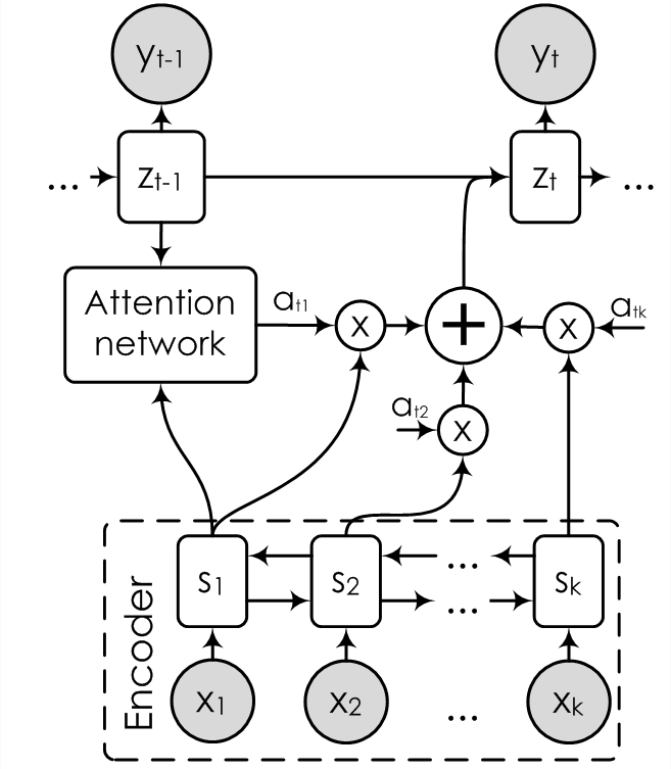
- *Мягкое внимание* (soft attention) (Luong et al. 2015a; 2015b; Jean et al. 2015):
  - encoder – двунаправленная RNN, есть оба контекста;
  - сеть внимания выдаёт оценку релевантности – надо ли переводить это слово прямо сейчас?



# Внимание в нейронных сетях

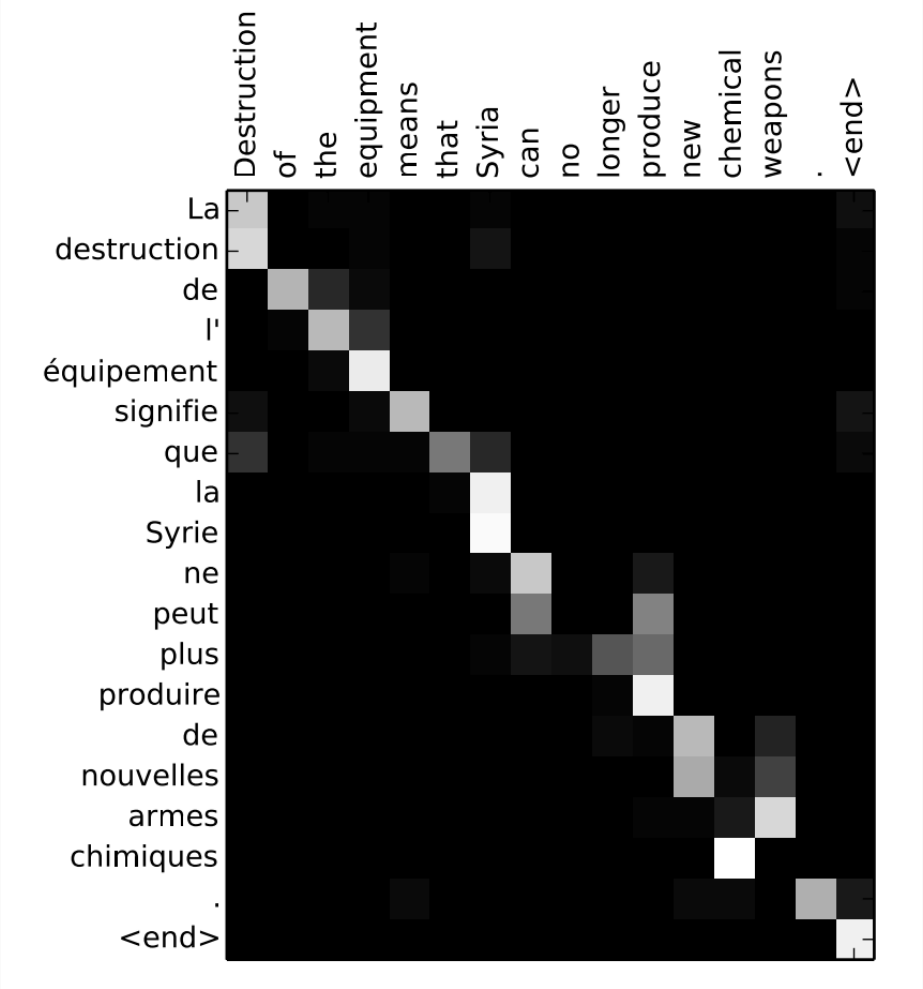
- Формально очень просто: считаем веса внимания $\alpha_{tj}$ и перевзвешиваем векторы контекстов:

$$e_{tj} = a(z_{t-1}, j), \quad \alpha_{tj} = \mathrm{softmax}(e_{tj}; e_{t*}),$$
$$c_t = \sum_j \alpha_{tj} h_j, \text{ и теперь } z_t = f(s_{t1}, y_{t1}, c_i).$$



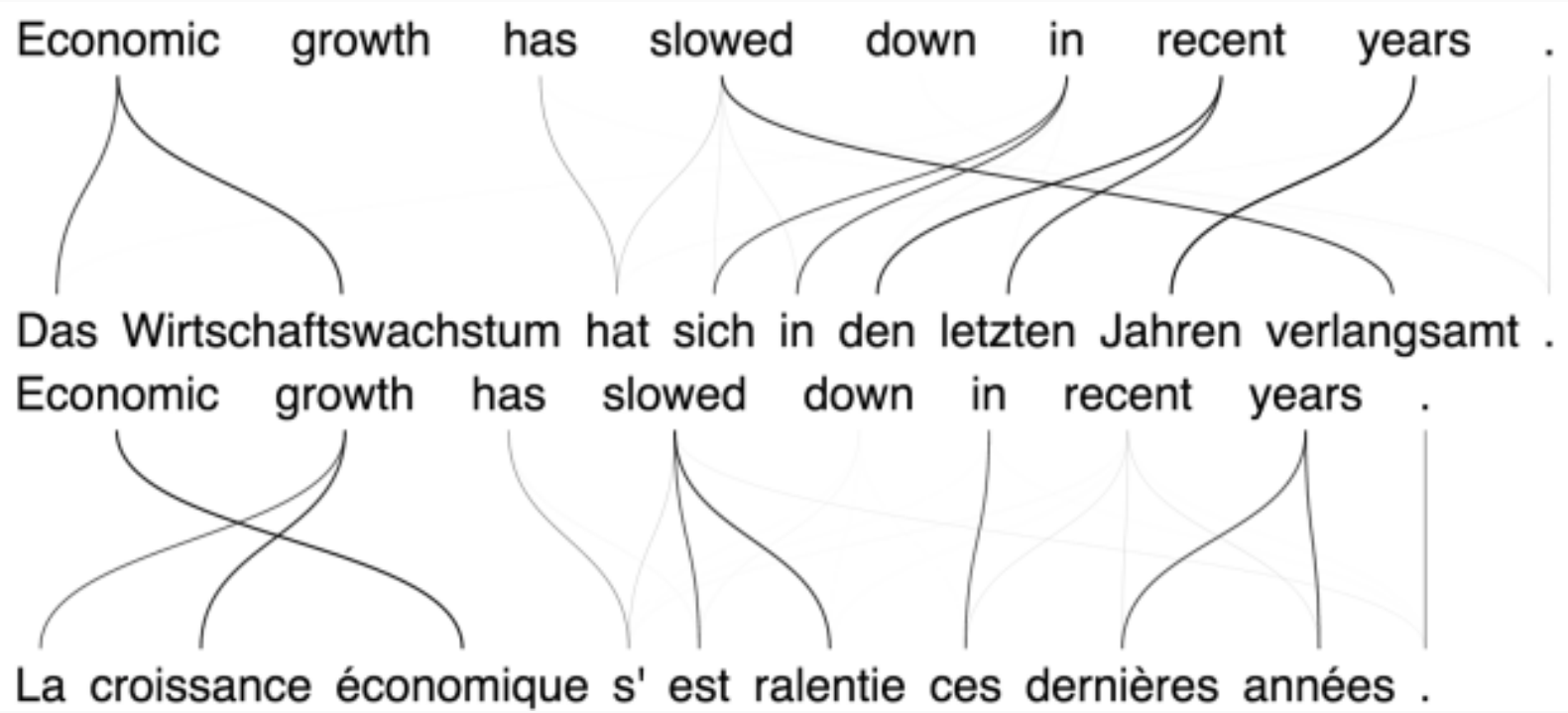
# Внимание в нейронных сетях

- В результате можно визуализировать, на что смотрит сеть:



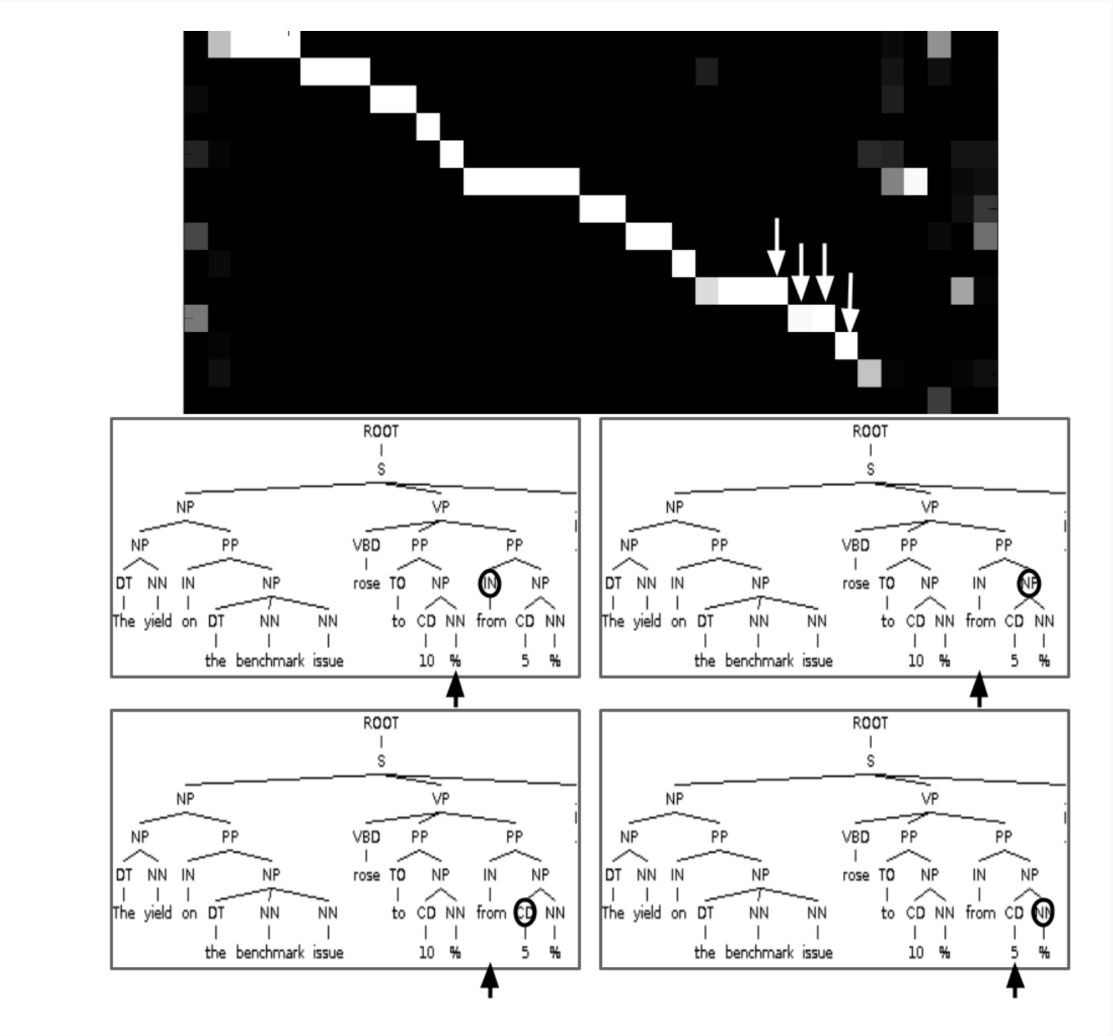
# Внимание в нейронных сетях

- Получается гораздо лучше порядок слов:



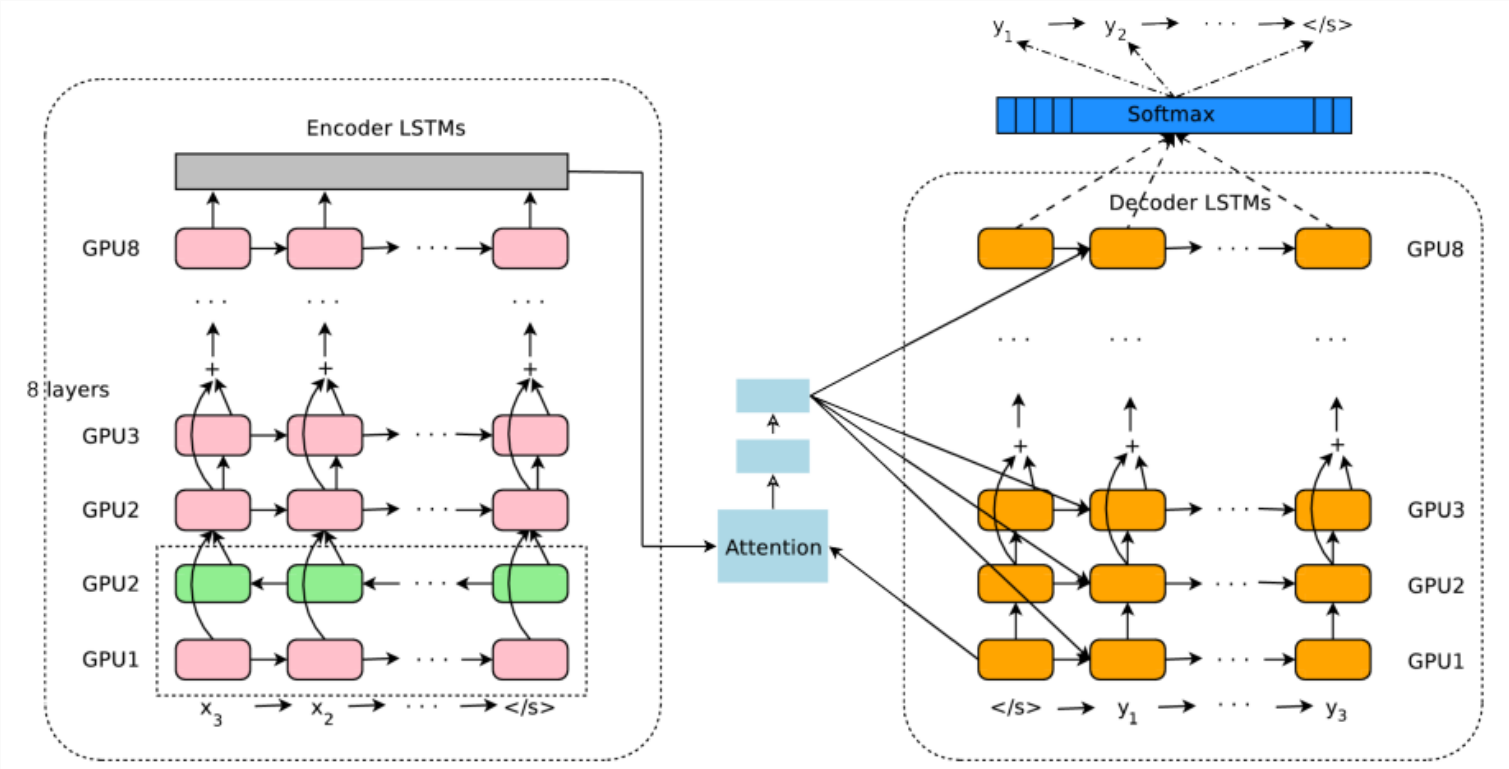
# Внимание в нейронных сетях

- Другая необычная работа – «Grammar as a Foreign Language» (Vinyals et al., 2015)



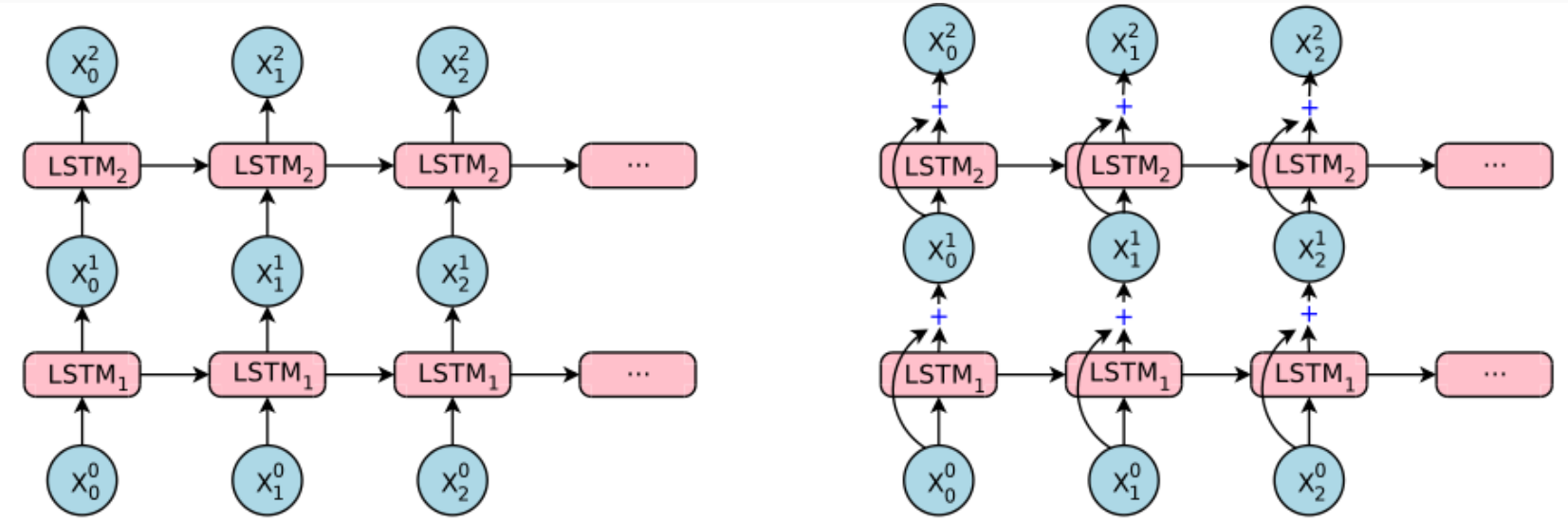
# Google Translate

- Сентябрь 2016: Wu et al., *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*:
  - как на самом деле работает Google Translate;
  - базовая архитектура та же самая: encoder, decoder, attention;
  - RNN глубокие, по 8 уровней в encoder и decoder:



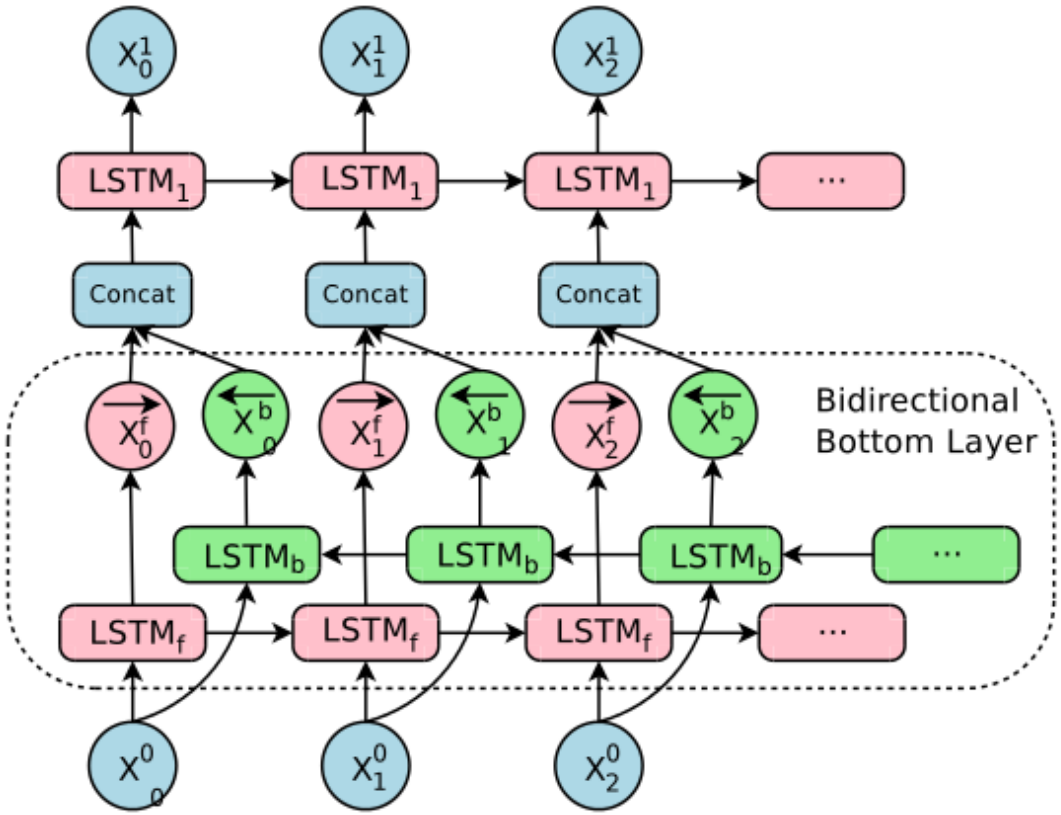
# Google Translate

- Сентябрь 2016: Wu et al., *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*:
  - просто stacked LSTM перестают работать далее 4-5 уровней;
  - поэтому добавляют остаточные связи, как в ResNet:



# Google Translate

- Сентябрь 2016: Wu et al., *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*:
  - нижний уровень, естественно, двунаправленный:



# Google Translate

- Сентябрь 2016: Wu et al., *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*:
- в GNMT ещё две идеи о сегментации слов:
  - *wordpiece model*: разбить слова на кусочки (отдельной моделью); пример из статьи:

    `Jet makers feud over seat width with big orders at stake`

    превращается в

    `_J et _makers _fe ud _over _seat _width _with _big _orders _at _stake`

  - *mixed word/character model*: конвертировать слова, не попадающие в словарь, в последовательность букв-токенов; пример из статьи:

    `Miki` превращается в `<B>M <M>i <M>k <E>i`

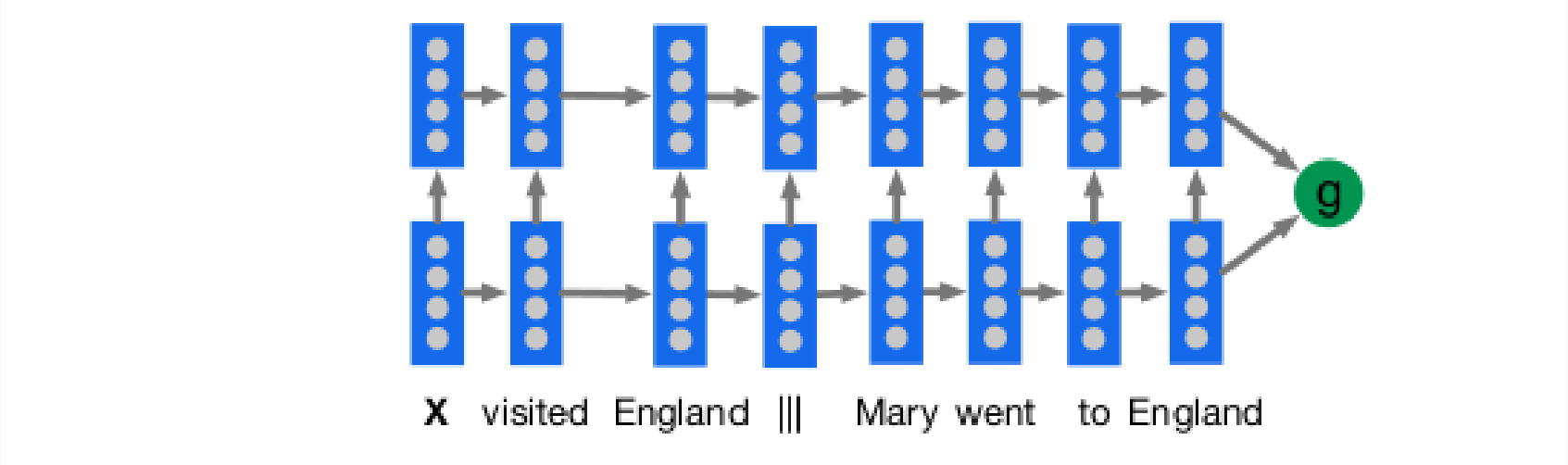
# Teaching machines to read

- (Hermann et al., 2015): «Teaching machines to read and comprehend» (Google DeepMind)
- Предлагают новый способ построить датасет для понимания, автоматически создавая тройки (context, query, answer) из текстов новостей и т.п.

| Original Version | Anonymised Version |
|---|---|
| **Context** | |
| The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." ... | the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " ... |
| **Query** | |
| Producer **X** will not press charges against Jeremy Clarkson, his lawyer says. | producer **X** will not press charges against *ent212* , his lawyer says . |
| **Answer** | |
| Oisin Tymon | *ent193* |

# Teaching machines to read

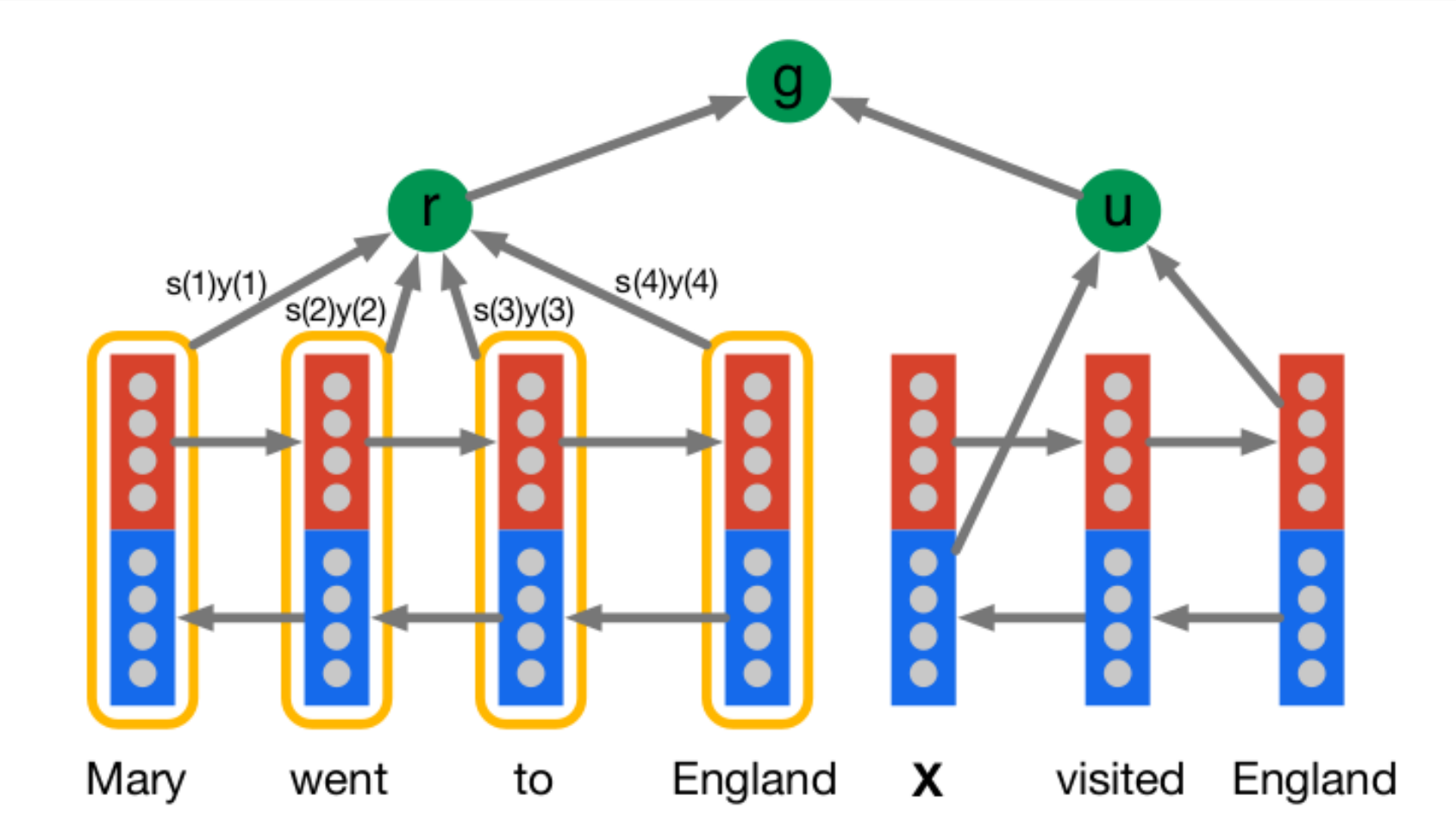
- Базовая модель – глубокий LSTM:




- Но так, конечно, совсем плохо работает.

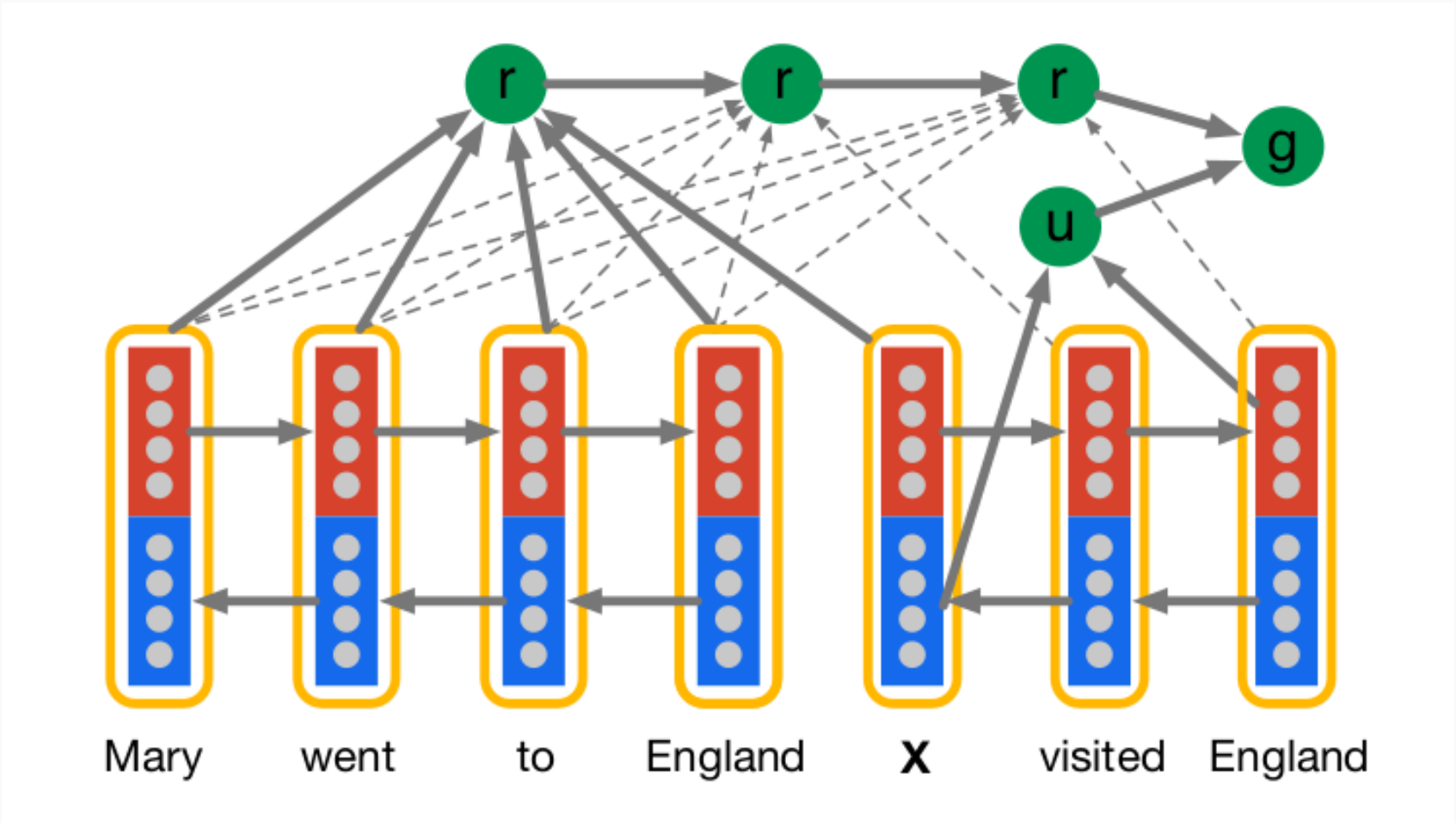
# Teaching machines to read

- Attentive Reader: обучаемся, на какую часть документа смотреть



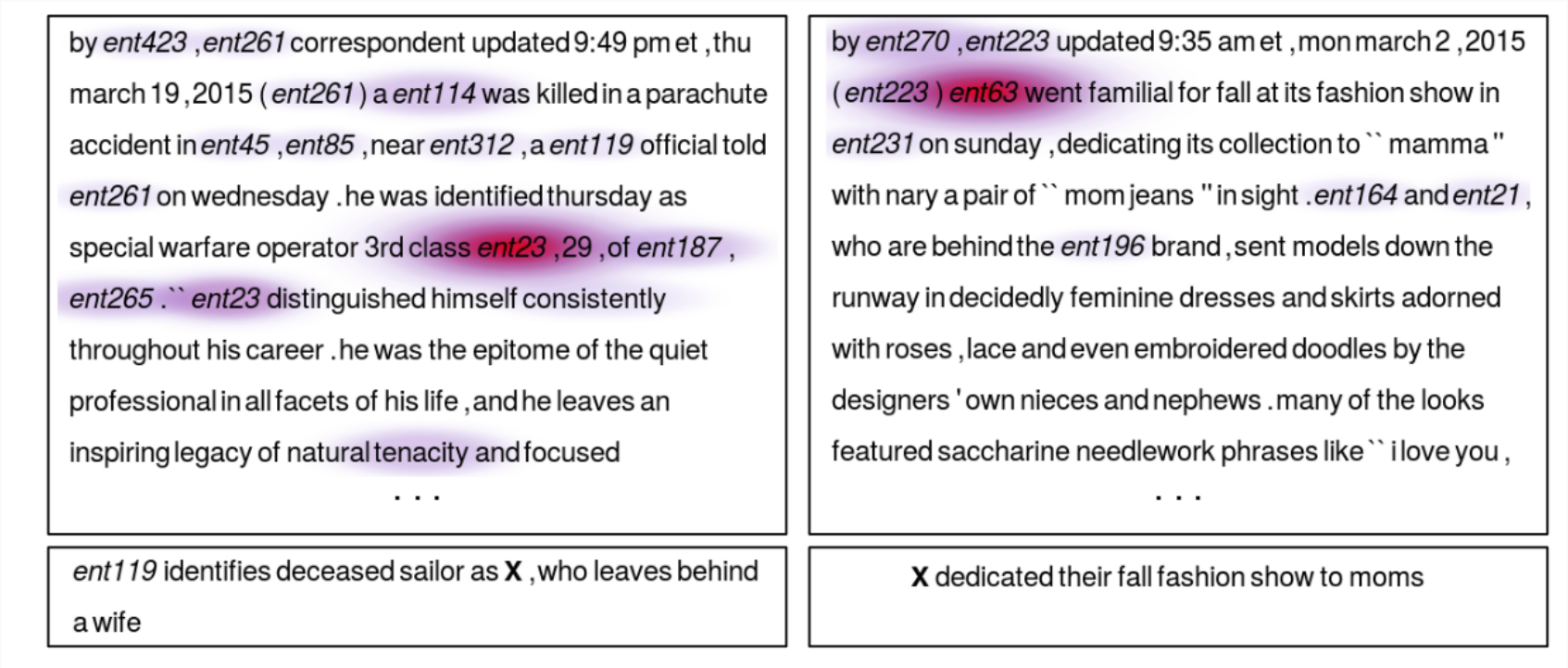
# Teaching machines to read

- Impatient Reader: можем перечитывать нужные части документа по мере прочтения запроса



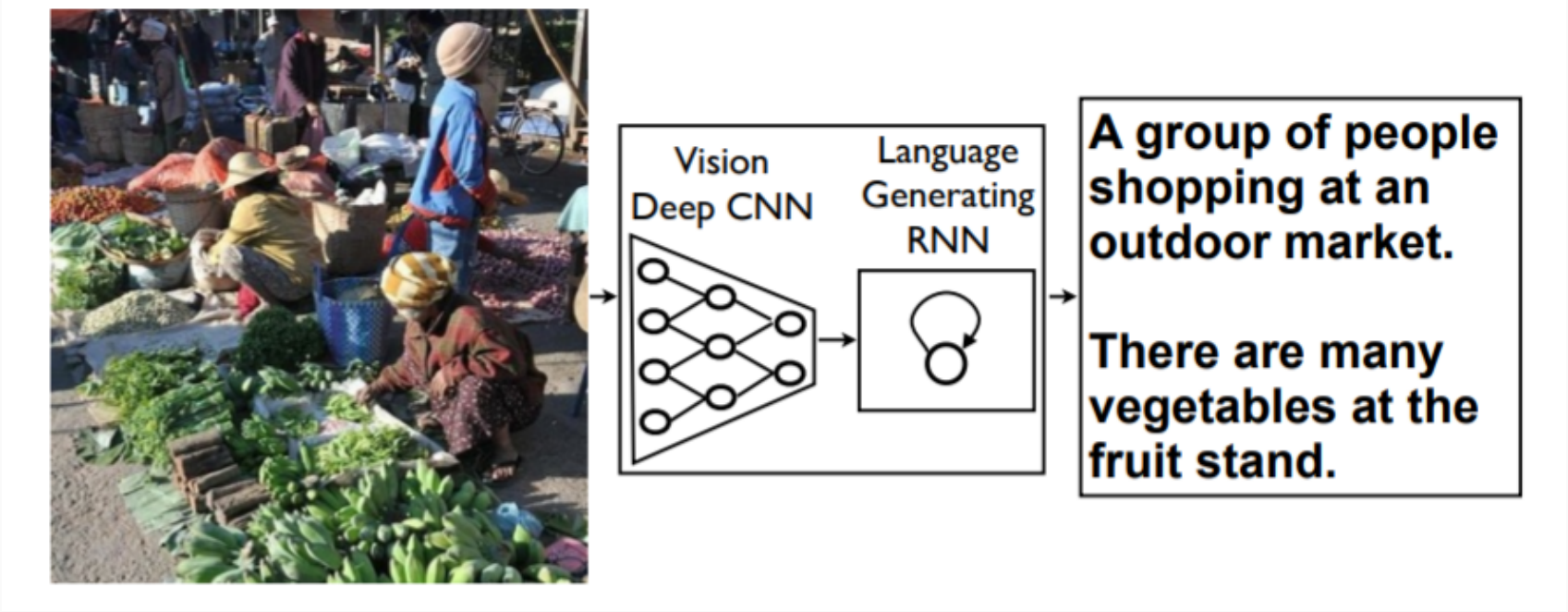
# Teaching machines to read

- Получаются разумные карты внимания:



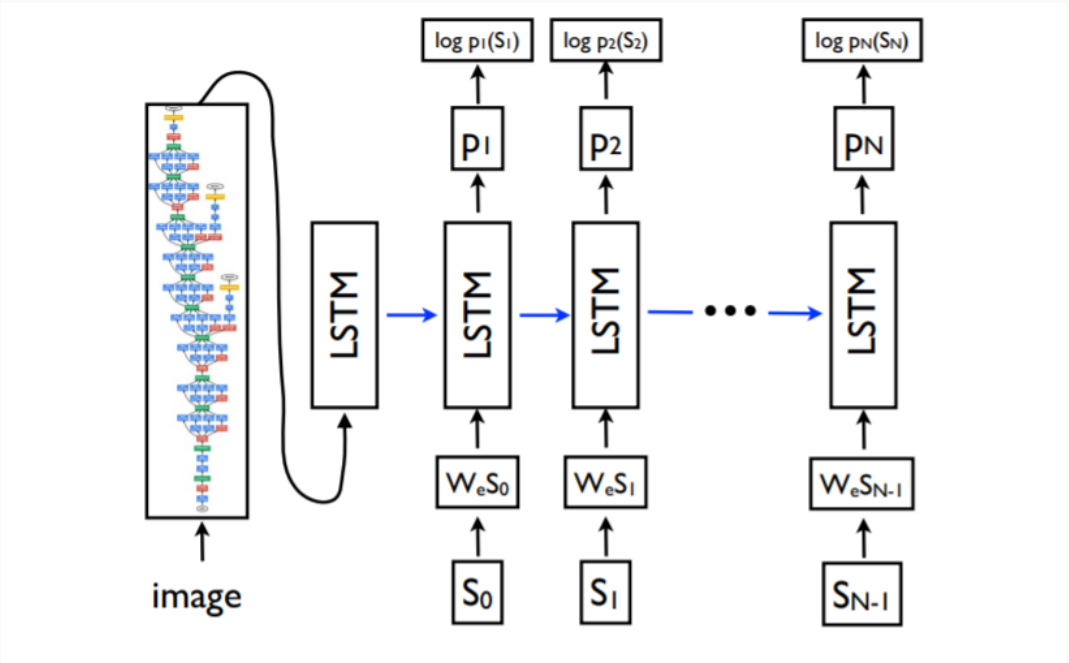
# Show, Attend, and Tell

- Теперь давайте про подписи к картинкам.
- Сначала было «Show and Tell» (Vinyals et al., 2015):



# Show, Attend, and Tell

- Довольно прямолинейная архитектура:
  - целевая функция – это просто $\sum_{(I,S)} \log p(S \mid I; \boldsymbol{\theta})$, где $I$ – картинка, $S$ – описание;
  - раскладываем и моделируем $p(S_t \mid I, S_0, \ldots, S_{t-1})$ рекуррентной сетью с LSTM;
  - а CNN используем, чтобы извлечь признаки.

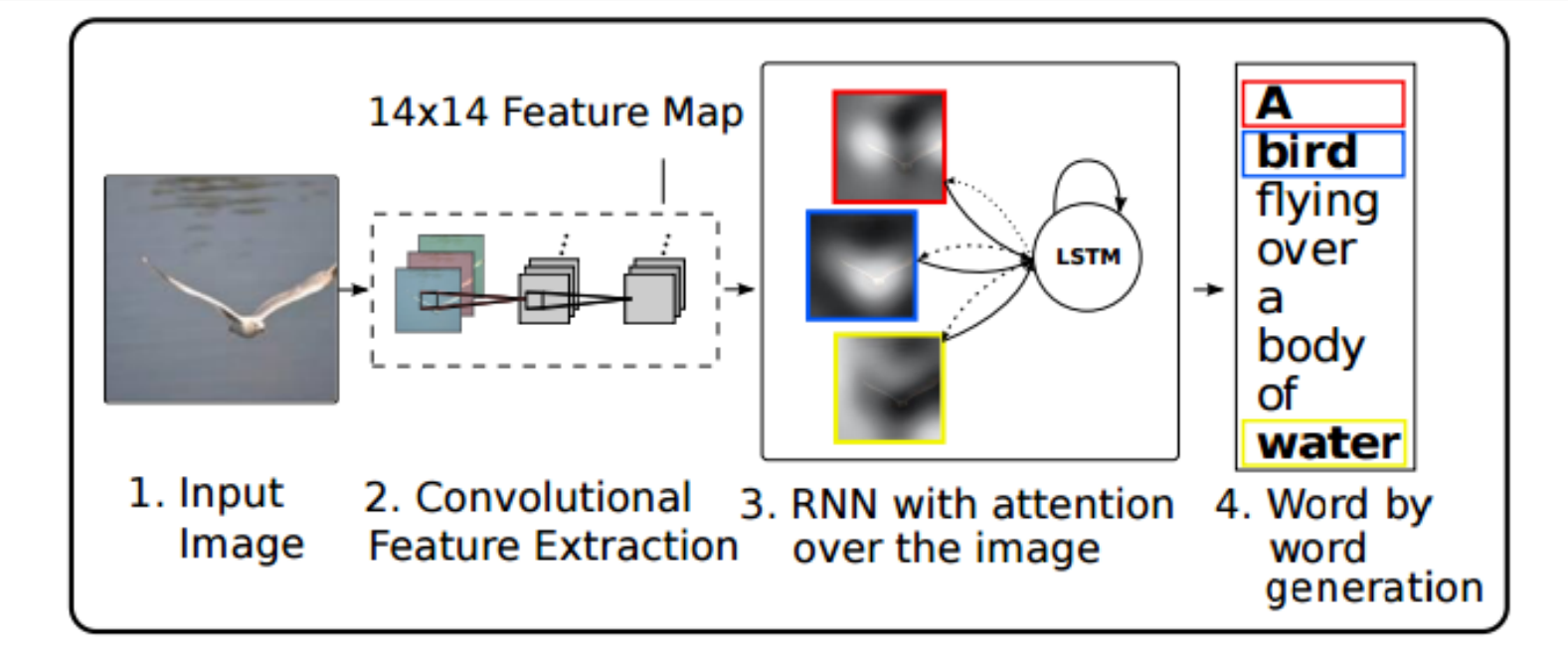


# Show, Attend, and Tell

- Получалось хорошо, но можно лучше:

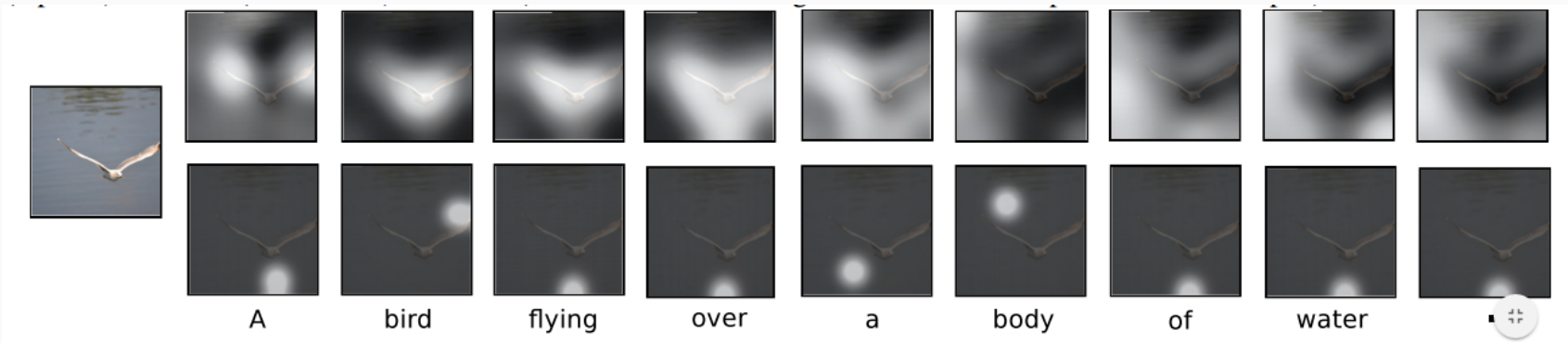
# Show, Attend, and Tell

- Из этого появилось «Show, Attend, and Tell» (Xu et al., 2015)



# Show, Attend, and Tell

- Soft attention vs. hard attention (стохастически выбираем однозначный кусок картинки).




- Soft attention – строим аннотацию с весами

$$\phi(\{\mathbf{a}\}_i, \{\alpha_i\}) = \sum_{i=1}^{L} \alpha_{t,i} \mathbf{a}_i.$$

# Show, Attend, and Tell

- Hard attention обучается максимизацией вариационной нижней оценки

$$L_S = \sum_s p(s \mid \mathbf{a}) \log p(\mathbf{y} \mid s, \mathbf{a}) \leq \log \sum_s p(s \mid \mathbf{a}) p(\mathbf{y} \mid s, \mathbf{a}) = \log p(\mathbf{y} \mid \mathbf{a}).$$

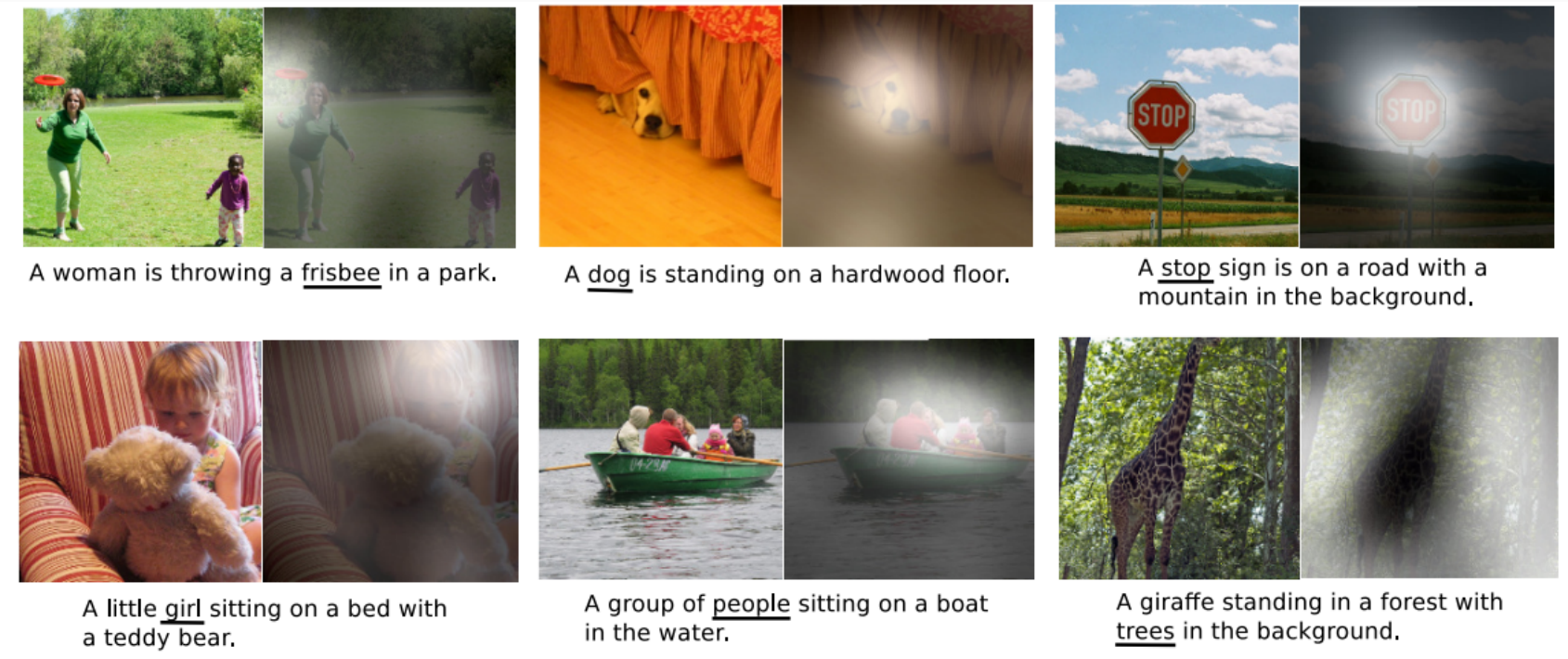
- От $L_S$ можно брать производные:

$$\frac{\partial L_S}{\partial W} = \sum_s p(s \mid \mathbf{a}) \left[ \frac{\partial \log p(\mathbf{y} \mid s, \mathbf{a})}{\partial W} + \log p(\mathbf{y} \mid s, \mathbf{a}) \frac{\partial \log p(s \mid \mathbf{a})}{\partial W} \right].$$

- И дальше сэмплируем $s_t$ с вероятностями $\alpha_i$ и приближаем ожидание выборкой.
- Опять те же трюки, вычитаем baseline, всё такое.

# Show, Attend, and Tell

- Часто получаются очень хорошие результаты:




- А когда плохие, можно посмотреть почему.

# Show, Attend, and Tell

- Примеры – hard attention:


(a) A man and a woman playing frisbee in a field.

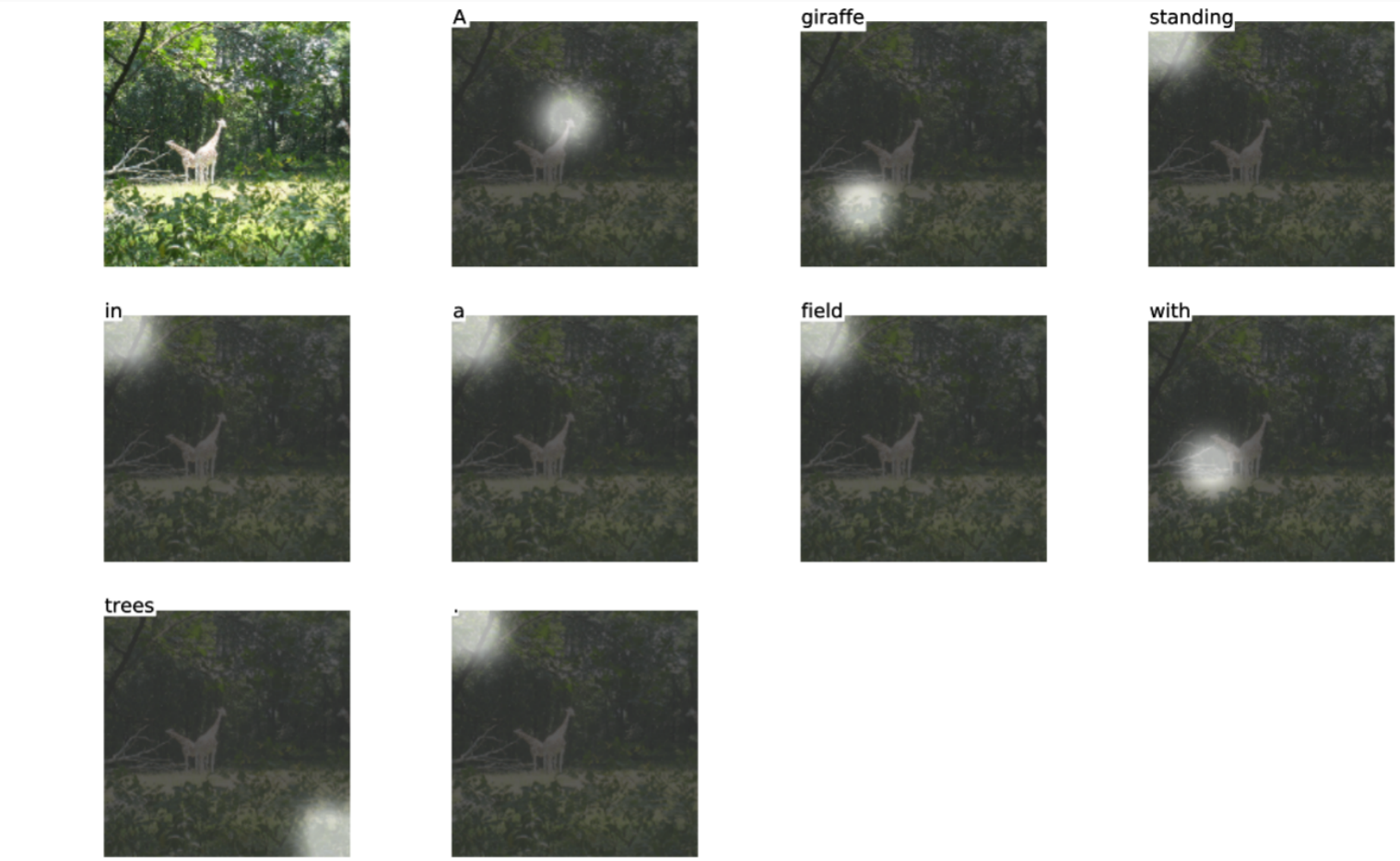
# Show, Attend, and Tell

- Примеры – soft attention:


(b) A woman is throwing a frisbee in a park.

# Show, Attend, and Tell

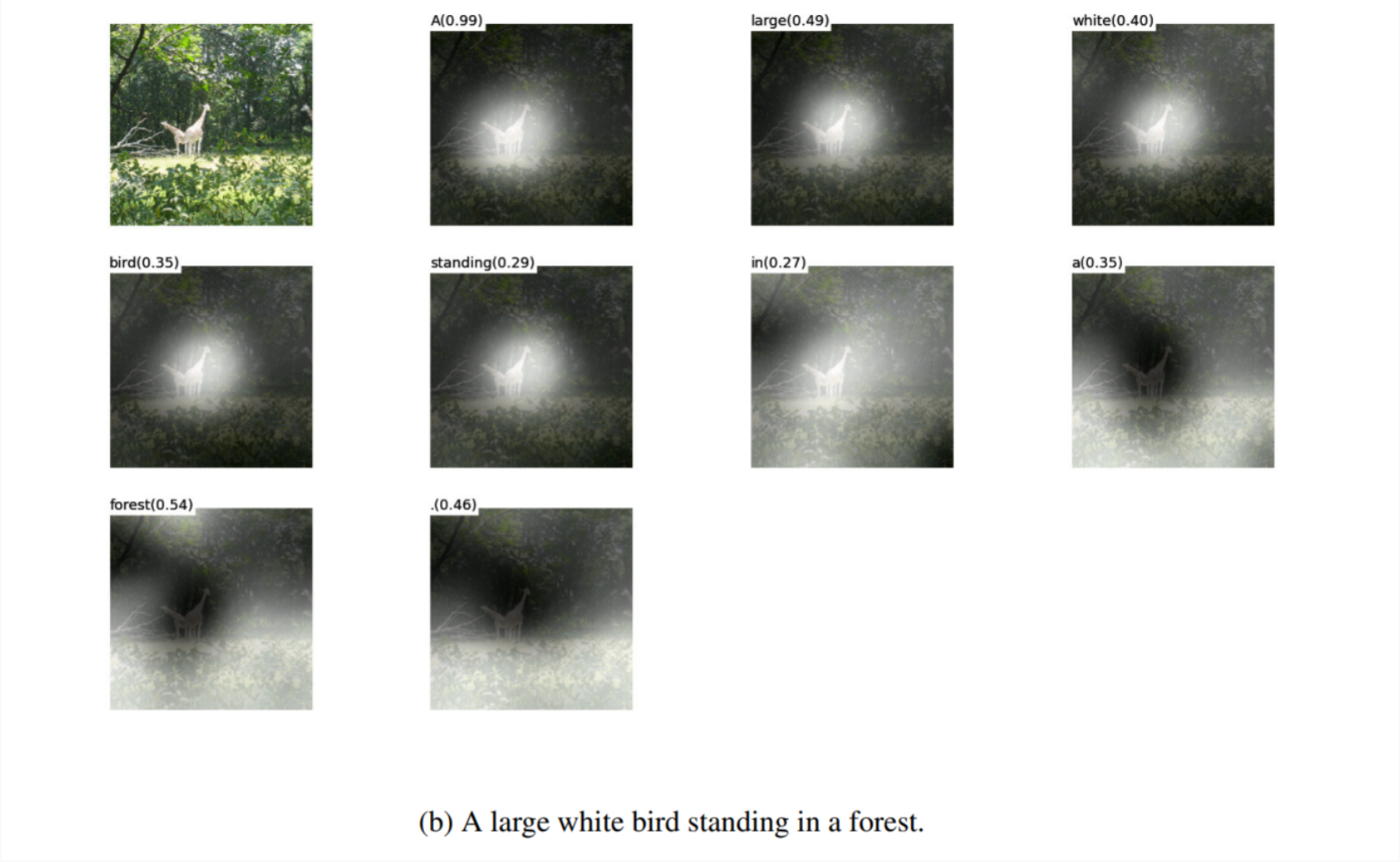
- Примеры – hard attention:




(a) A giraffe standing in the field with trees.

# Show, Attend, and Tell

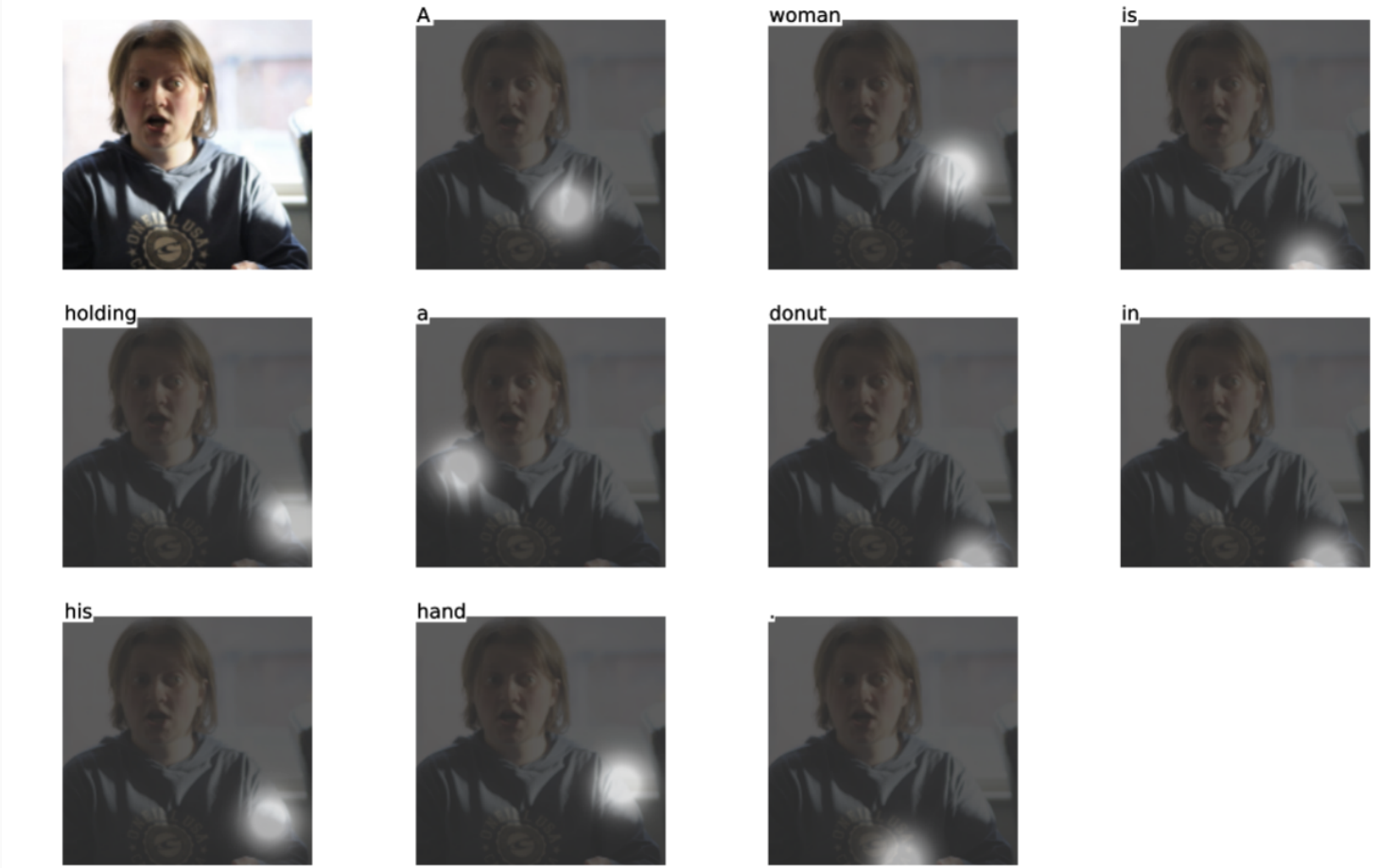
- Примеры – soft attention:




(b) A large white bird standing in a forest.

# Show, Attend, and Tell

- Примеры – hard attention:




(a) A woman is holding a donut in his hand.

# Show, Attend, and Tell

- Примеры – soft attention:


(b) A woman holding a clock in her hand.

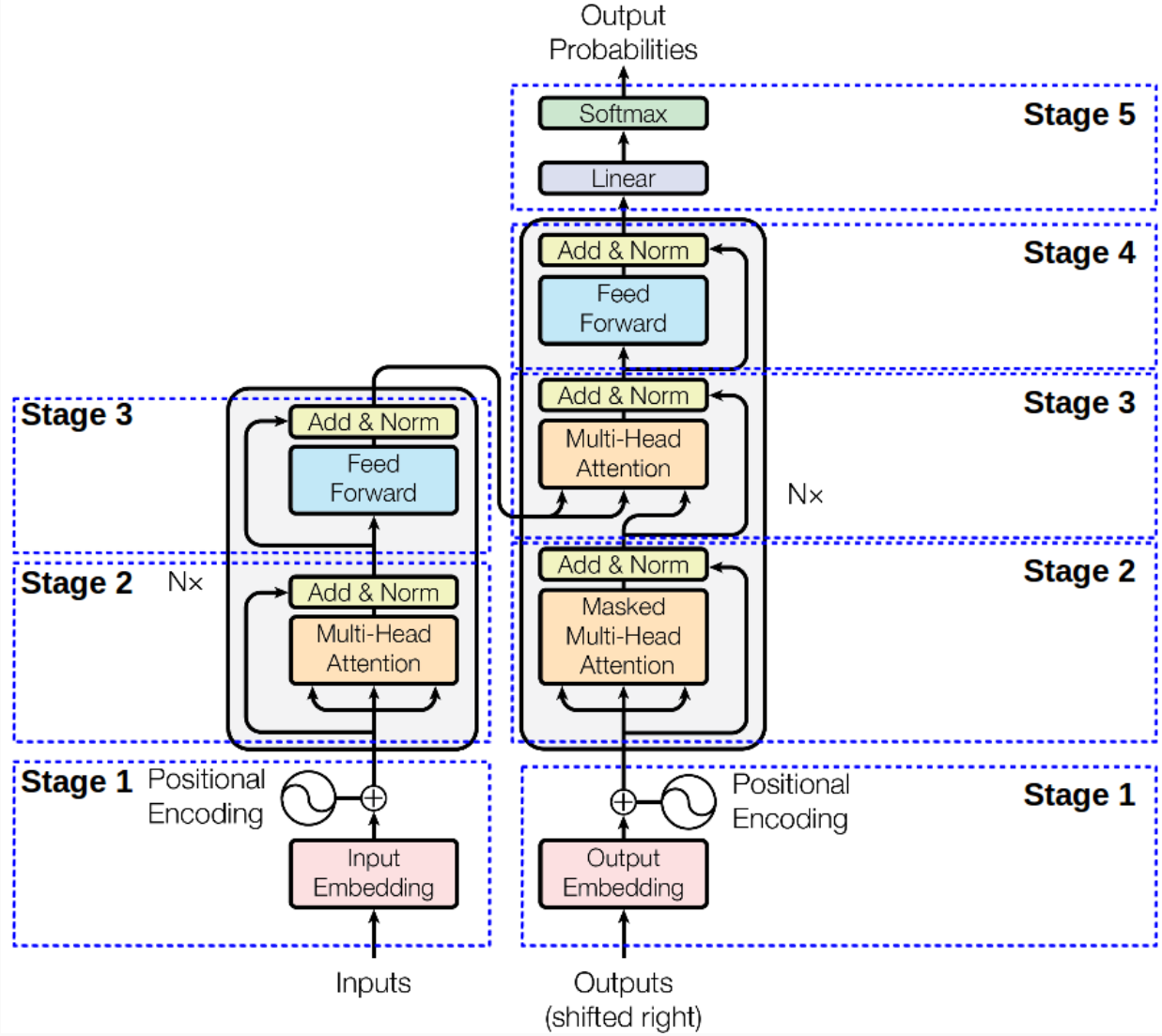
# Transformer и что из него получилось

# Transformer

- Мы изучали перевод на рекуррентных сетях и дошли до архитектуры Google NMT
- Но в 2017 году оказалось, что всё может быть ещё проще и интереснее
- Google: «Attention is all you need» (Vaswani et al., 2017)
- Основная идея – self-attention; оказывается очень плодотворной для всевозможных seq2seq задач
- Главная мотивация – попробовать всё-таки уйти от кодирования вектором постоянной длины

# Transformer

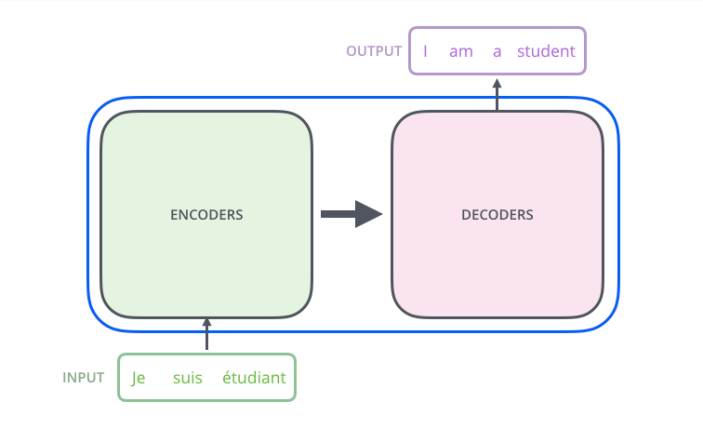
- Общая схема:



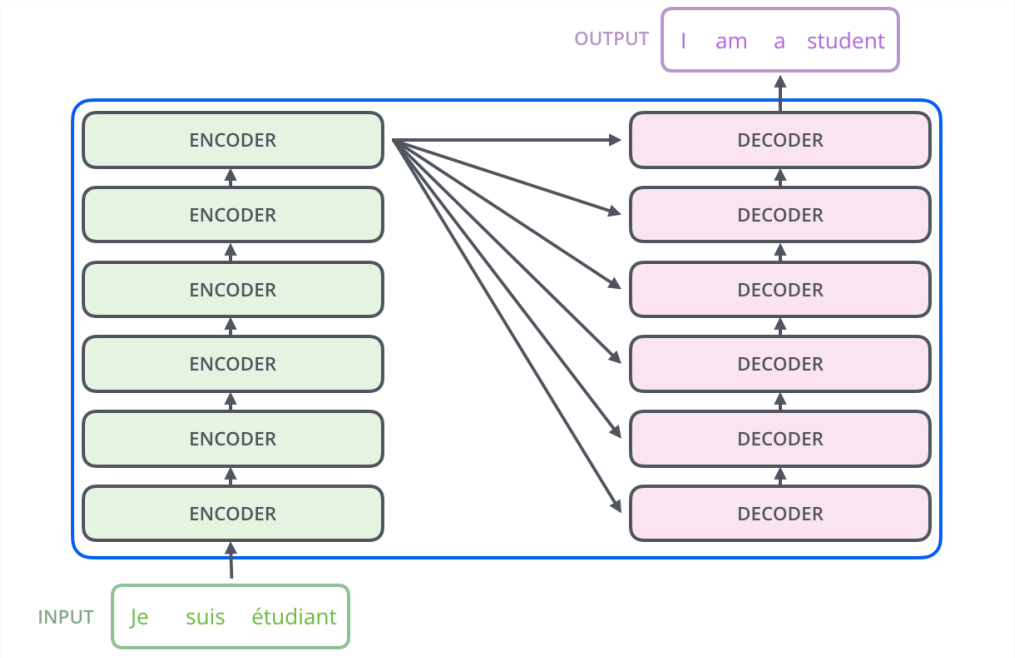

- Теперь подробнее...

# Transformer

- Суть, как и раньше, – encoder-decoder:

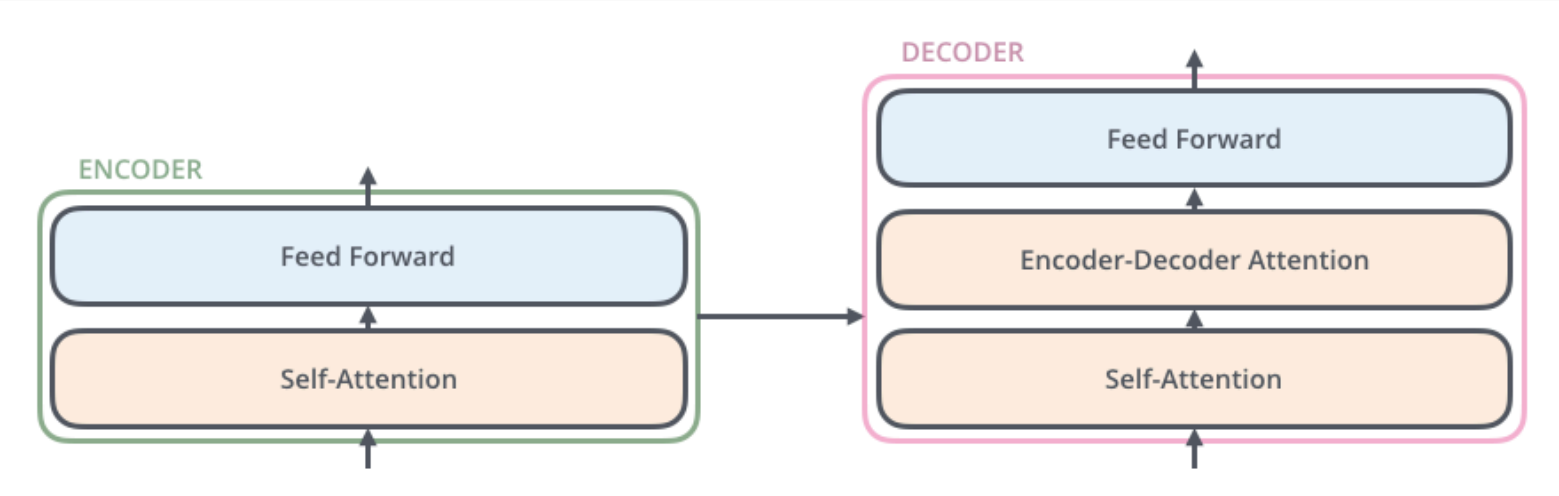



- 6 слоёв encoder’a, результат потом дают 6 слоям декодера:



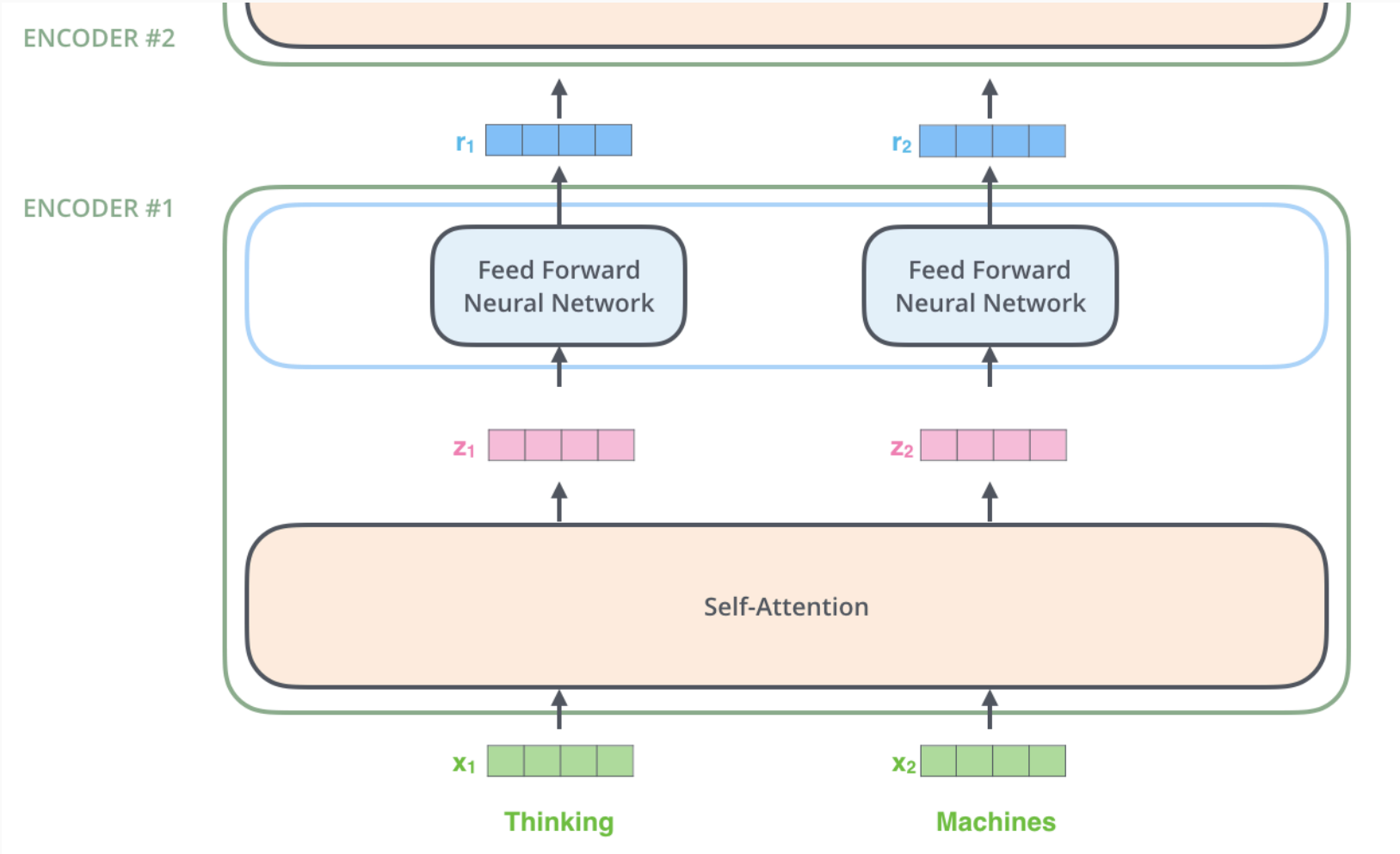
# Transformer

- В каждом слое – слой self-attention, а потом feedforward layer, который независимо применяется к каждой позиции входа. У декодера ещё есть attention между ними:



# Transformer

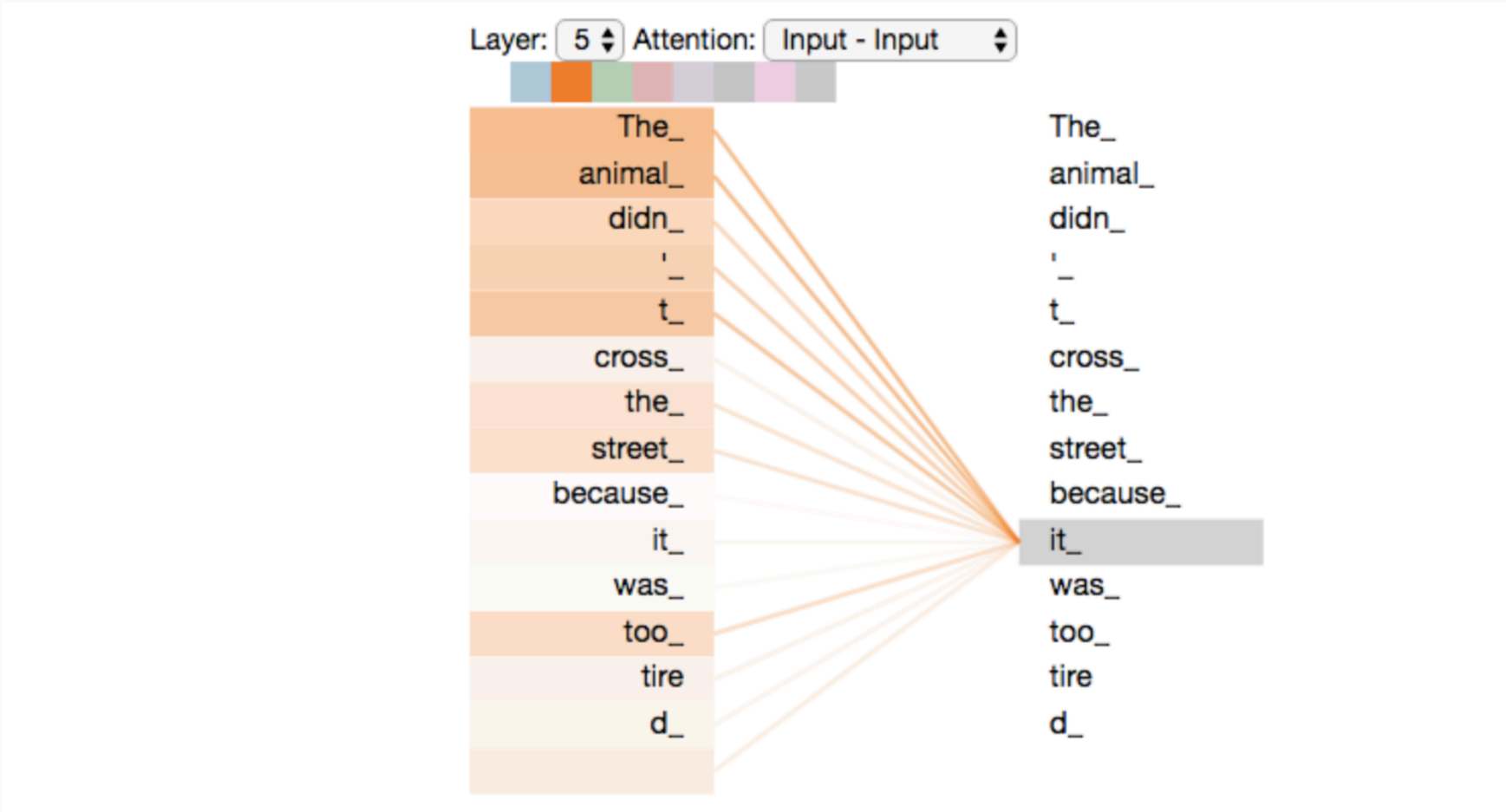
- Слова, естественно, представляются векторами, в feedforward слое всё параллельно:




- Но что же это такое – self-attention?

# Transformer

- Идея в том, чтобы обучить веса, с которыми обработка текущего слова будет учитывать другие слова:

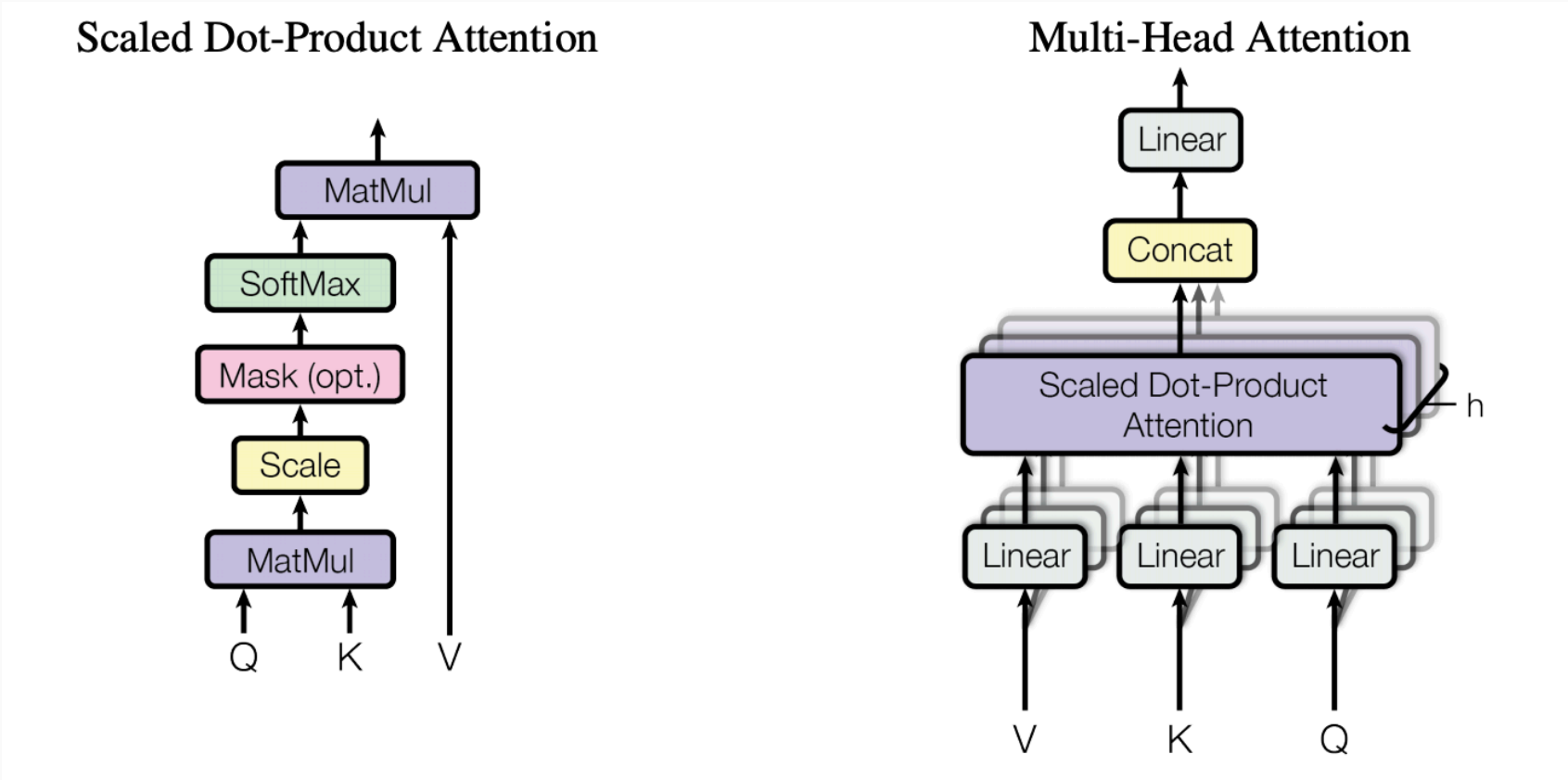



- Теперь детально...

# Transformer

- Scaled Dot-Product Attention состоит из queries, keys и values:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{1}{\sqrt{d_k}} QK^\top\right) V$$

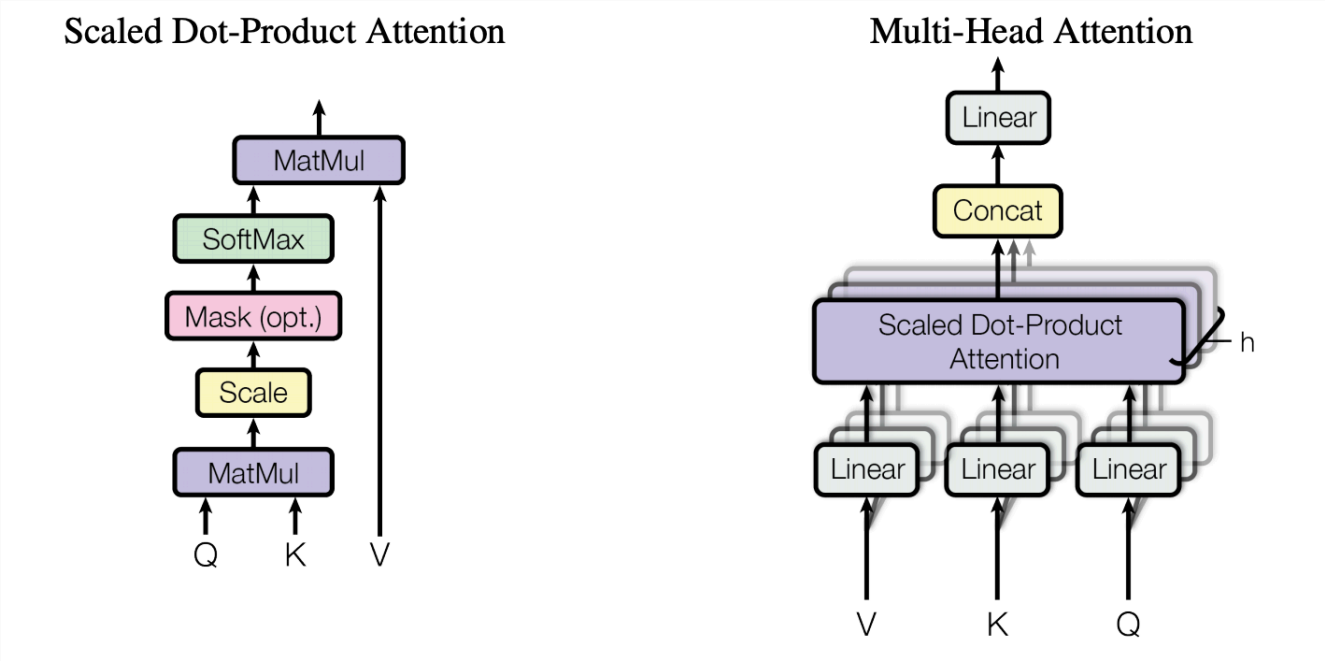


# Transformer

- Multi-head attention объединяет несколько self-attention карт в общие матричные вычисления:

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}\left(\mathrm{head}_1, \ldots, \mathrm{head}_h\right) W^O,$$
$$\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V),$$

где проекции $W_i^*$ – это обучаемые матрицы весов.



# Transformer

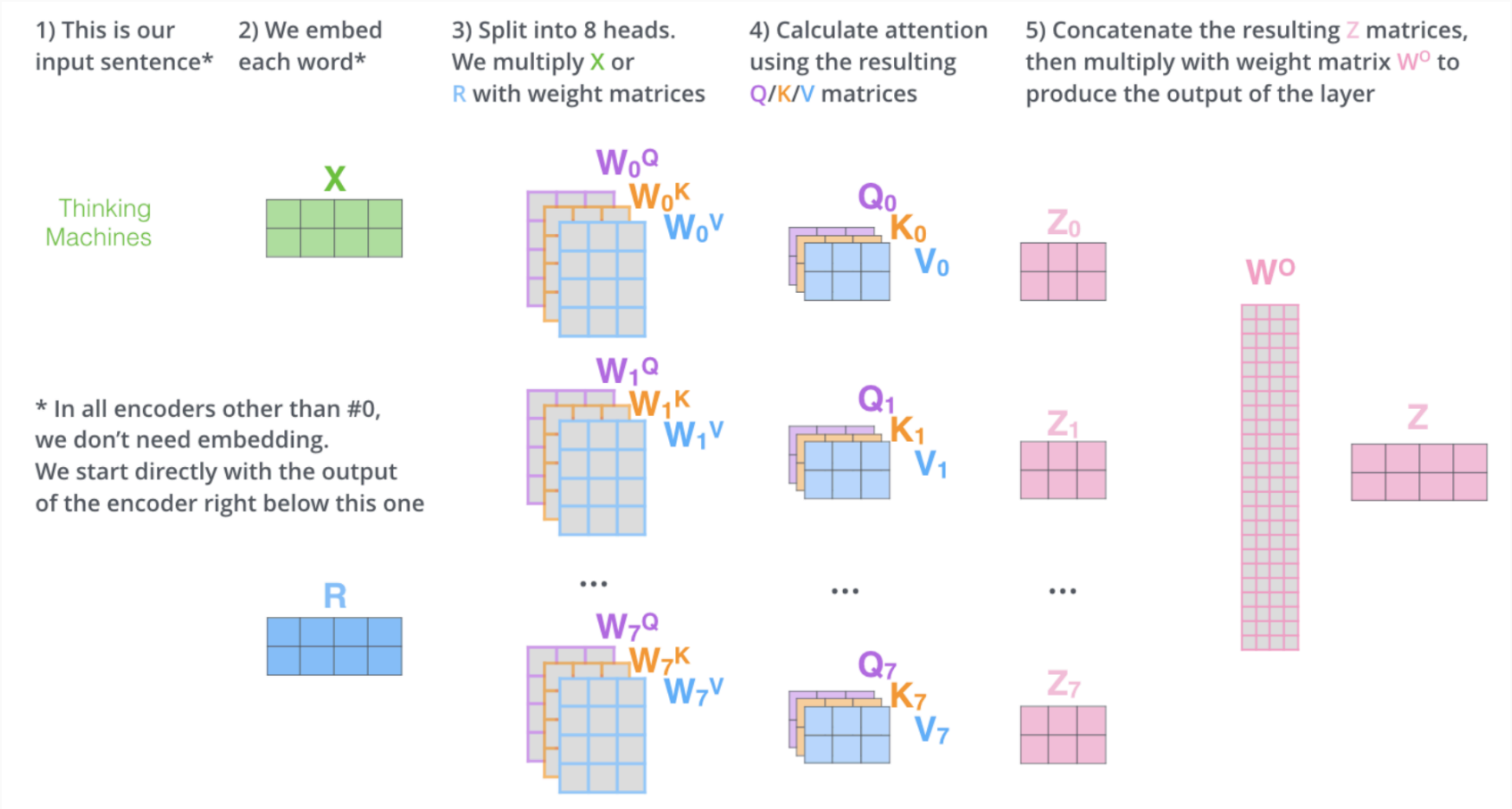
- Последний вопрос – сейчас модель совсем не учитывает порядок слов!
- Для этого добавляем к представлениям слов ещё positional embeddings:

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

т.е. по каждой размерности $i$ идёт своя синусоида; идея sin/cos в том, чтобы для каждого фиксированного $k$ $\mathrm{PE}_{\mathrm{pos}+k}$ было бы линейной функцией от $\mathrm{PE}_{\mathrm{pos}}$, и это облегчило бы обучение того, как смотреть на относительные смещения.

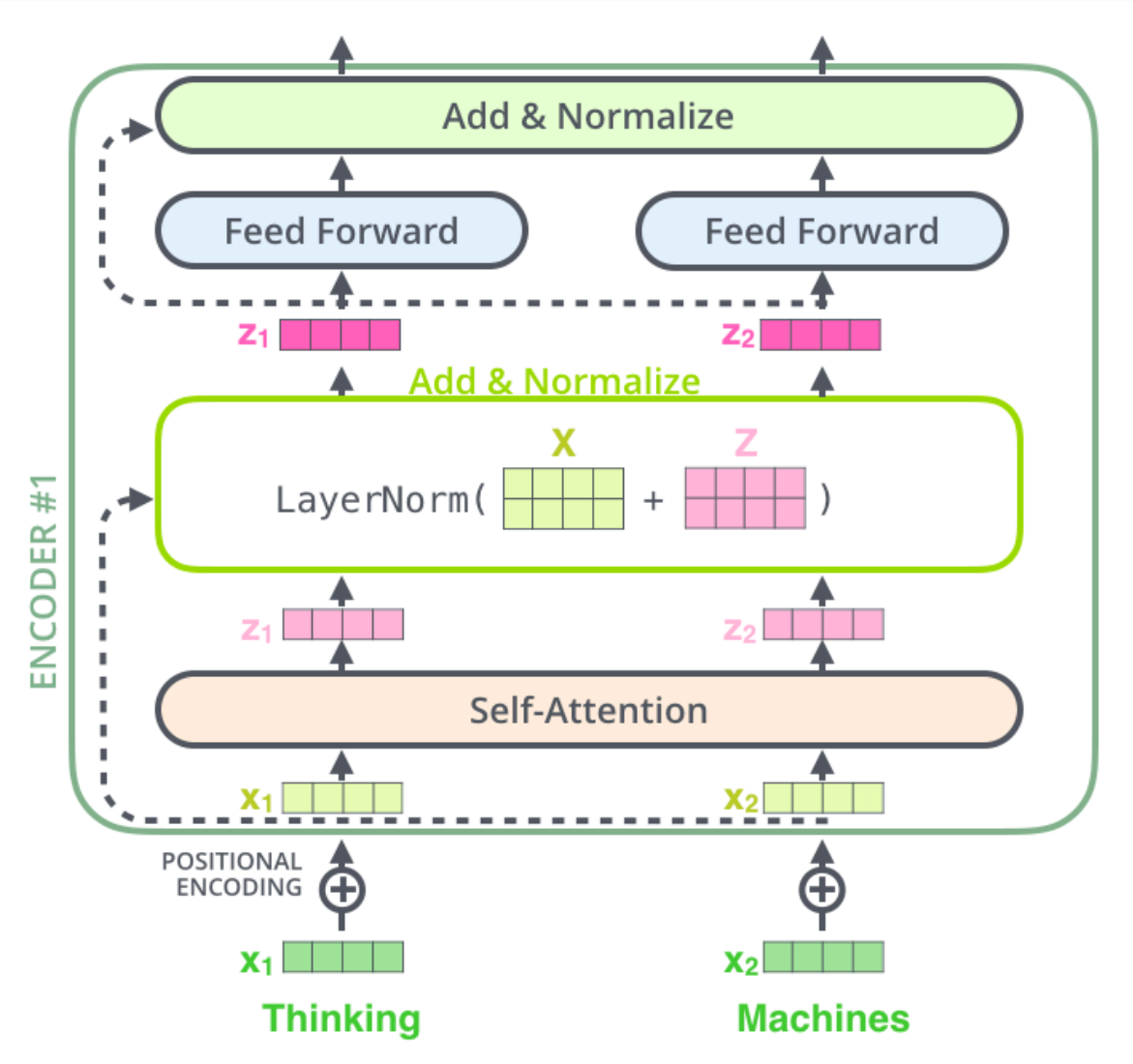
# Transformer

- Putting it all together:



# Transformer

- Putting it all together:



# Transformer

- Бонус от self-attention – во-первых, вычислительный, во-вторых, сокращает пути между словами, в-третьих, потенциальная интерпретируемость.

Table 1: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types. $n$ is the sequence length, $d$ is the representation dimension, $k$ is the kernel size of convolutions and $r$ the size of the neighborhood in restricted self-attention.

| Layer Type | Complexity per Layer | Sequential Operations | Maximum Path Length |
|---|---|---|---|
| Self-Attention | $O(n^2 \cdot d)$ | $O(1)$ | $O(1)$ |
| Recurrent | $O(n \cdot d^2)$ | $O(n)$ | $O(n)$ |
| Convolutional | $O(k \cdot n \cdot d^2)$ | $O(1)$ | $O(log_k(n))$ |
| Self-Attention (restricted) | $O(r \cdot n \cdot d)$ | $O(1)$ | $O(n/r)$ |

# Transformer

- Работает лучше, обучается в сто раз быстрее:

| Model | BLEU | | Training Cost (FLOPs) | |
|---|---|---|---|---|
| | EN-DE | EN-FR | EN-DE | EN-FR |
| ByteNet [15] | 23.75 | | | |
| Deep-Att + PosUnk [32] | | 39.2 | | $1.0 \cdot 10^{20}$ |
| GNMT + RL [31] | 24.6 | 39.92 | $2.3 \cdot 10^{19}$ | $1.4 \cdot 10^{20}$ |
| ConvS2S [8] | 25.16 | 40.46 | $9.6 \cdot 10^{18}$ | $1.5 \cdot 10^{20}$ |
| MoE [26] | 26.03 | 40.56 | $2.0 \cdot 10^{19}$ | $1.2 \cdot 10^{20}$ |
| Deep-Att + PosUnk Ensemble [32] | | 40.4 | | $8.0 \cdot 10^{20}$ |
| GNMT + RL Ensemble [31] | 26.30 | 41.16 | $1.8 \cdot 10^{20}$ | $1.1 \cdot 10^{21}$ |
| ConvS2S Ensemble [8] | 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** | **41.0** | $2.3 \cdot 10^{19}$ | |

# Transformer

- Следующий шаг – OpenAI GPT (Generative Pretrained Transformer) (Radford et al., 2018)
- Используем Transformer в такой последовательности:
  - сначала обучаем обычную языковую модель

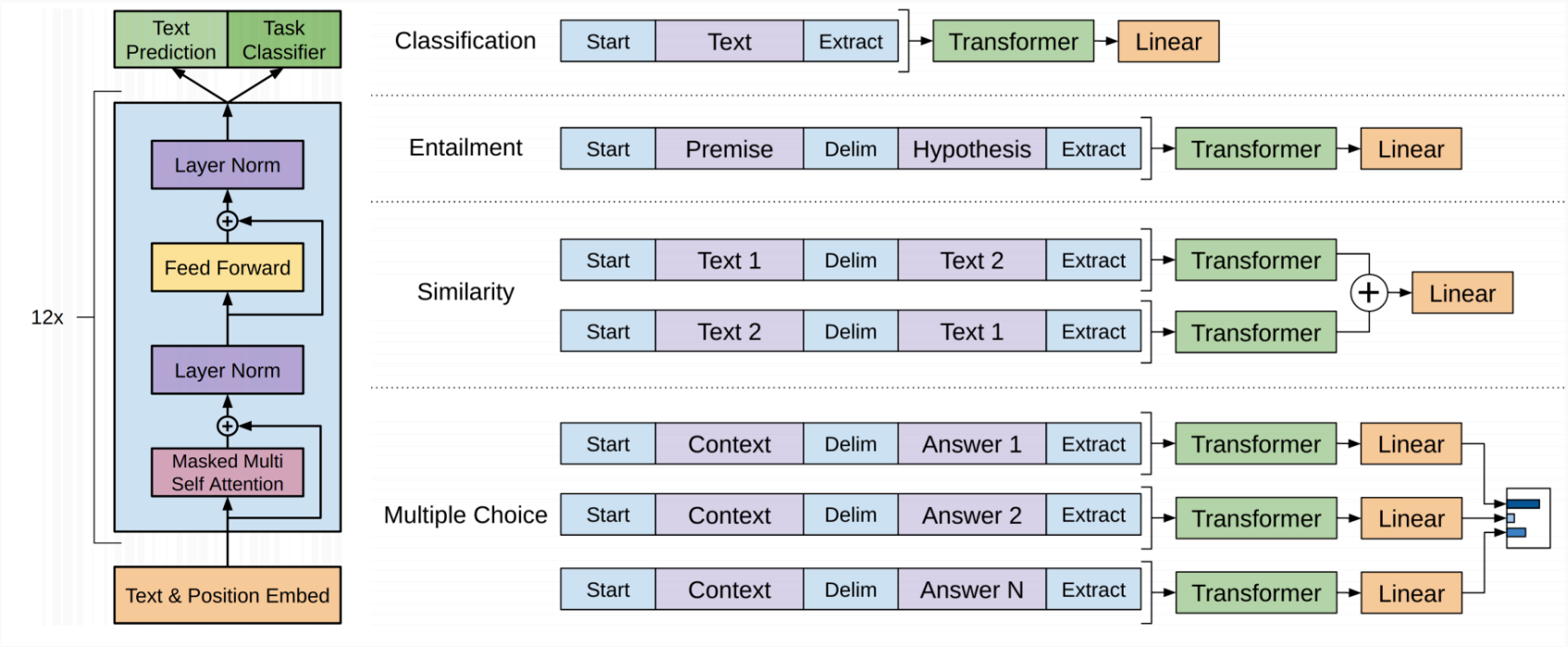
$$L_1(D) = \sum_i \log p(u_i \mid u_{i-k}, \ldots, u_{i-1}; \boldsymbol{\theta});$$

  - потом добавляем новый линейный слой для каждой задачи и делаем fine-tuning уже с учителем:

$$L(C, D) = \sum_{(\mathbf{x}, y)} \log p(y \mid \mathbf{x}) + \lambda L_1(D).$$

# Transformer

- Идея в том, чтобы переиспользовать одну модель на много разных задач:



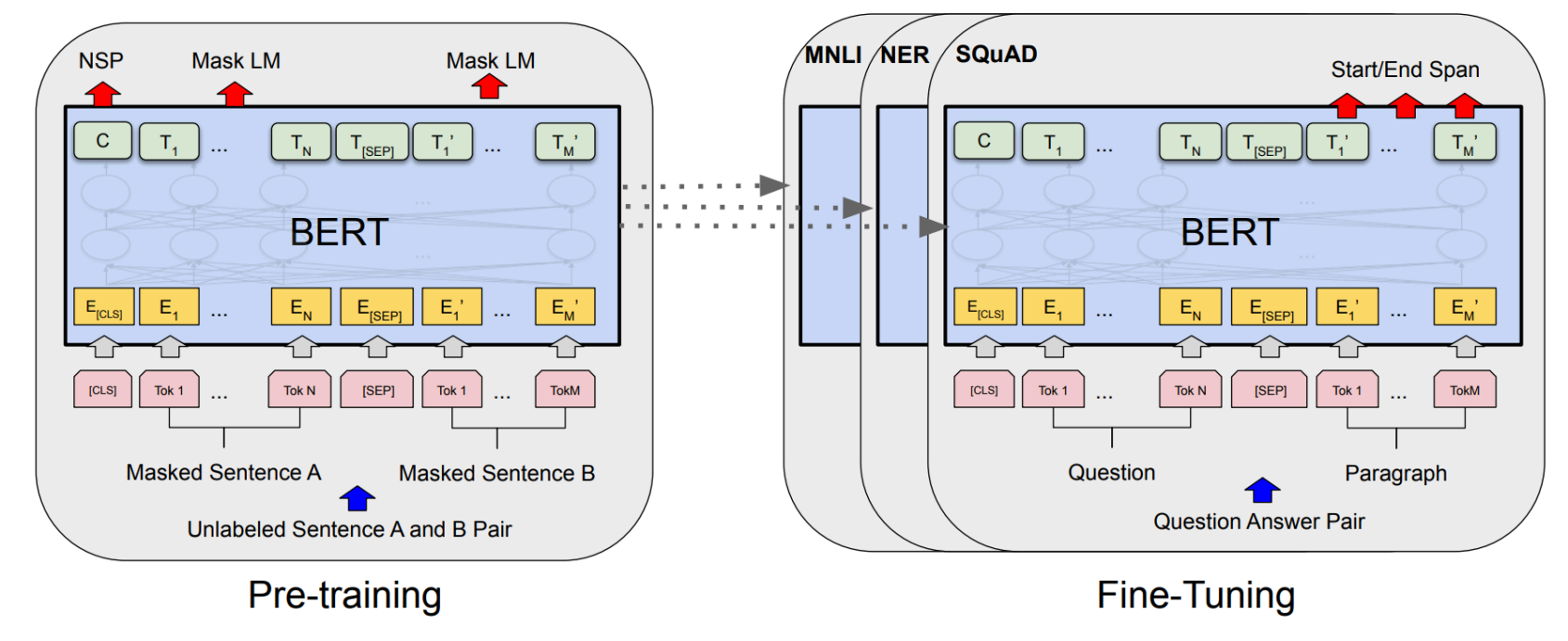

- Получают результаты гораздо лучше, чем были раньше, сразу для нескольких задач.

# BERT-семейство

- (Devlin et al., 2018): BERT – Bidirectional Encoder Representations from Transformers
- Фактически тот же Transformer, но теперь двунаправленный, при условии и левого, и правого контекста во всех слоях.
- Т.е. та же языковая модель, но теперь работает с контекстом и слева, и справа.
- Или так не работает? Что делать?..

# BERT-семейство

- Просто вместо обычной языковой модели будем маскировать случайные слова и пытаться их предсказывать.
- И вторая задача для pretraining – предсказание следующего предложения.



# BERT-семейство

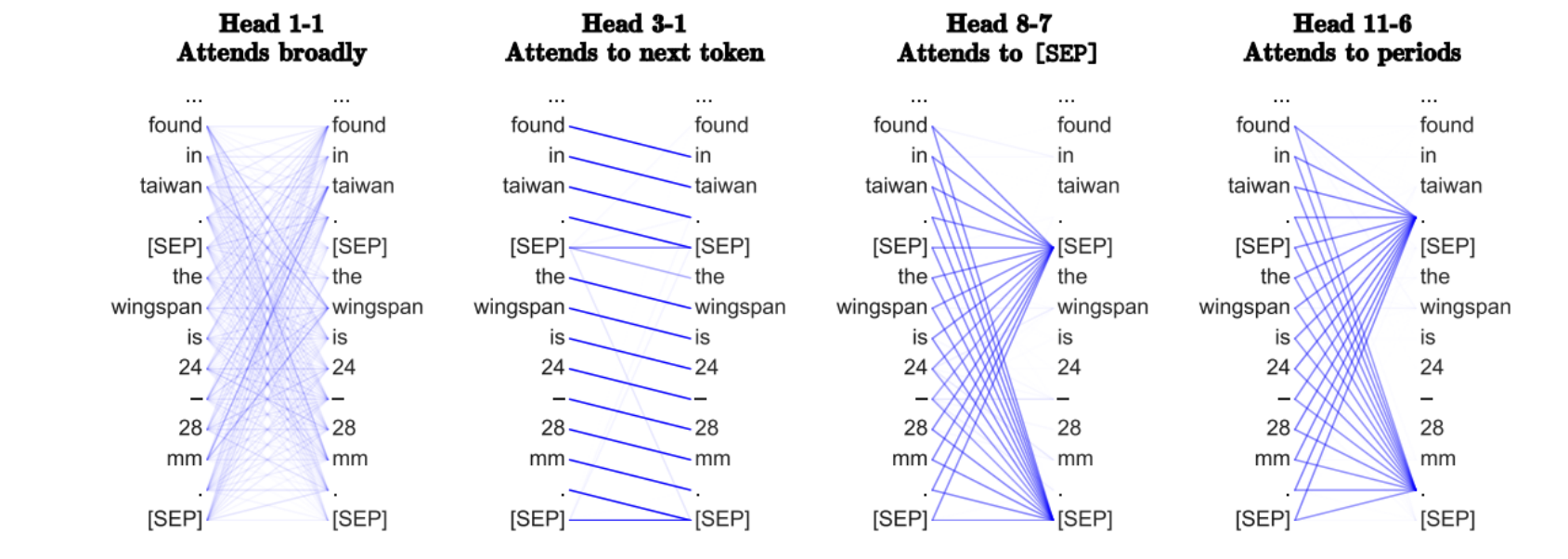
- И всё, в остальном обычный Transformer. Опять побили лучшие результаты для всех задач:

| System | MNLI-(m/mm) 392k | QQP 363k | QNLI 108k | SST-2 67k | CoLA 8.5k | STS-B 5.7k | MRPC 3.5k | RTE 2.5k | **Average** - |
|---|---|---|---|---|---|---|---|---|---|
| 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** |

- Сейчас BERT – стандартная основа для conversational models и вообще чего угодно в NLP.

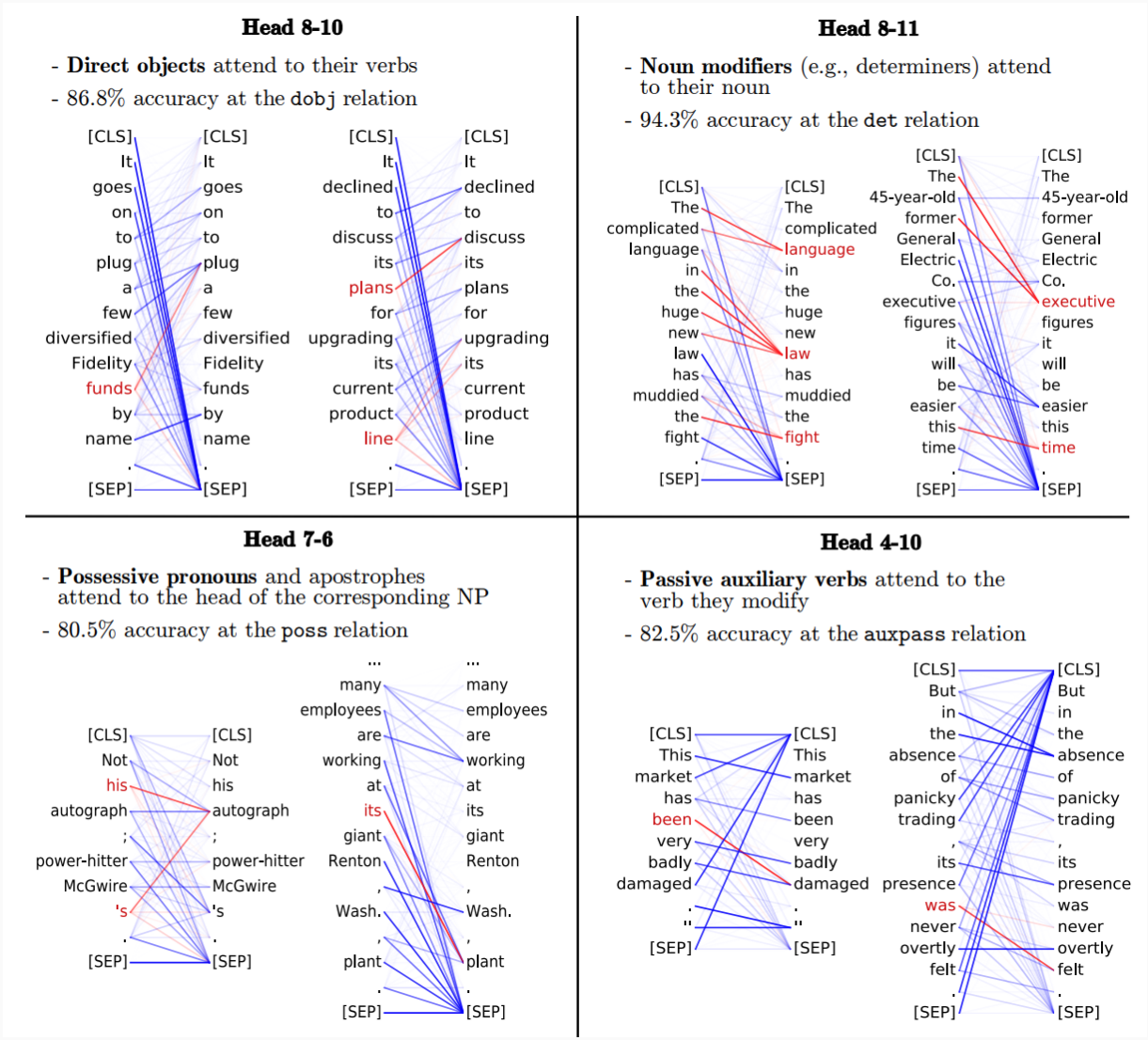
# BERT-семейство

- Ещё важные детали о BERT:
  - wordpiece embeddings: используем фиксированный словарь подслов размером 30K, а слова делим на части: electrodynamics → electro# #dy# #nami# #cs
  - можно найти «головы», которые смотрят по-разному, дают разный attention:



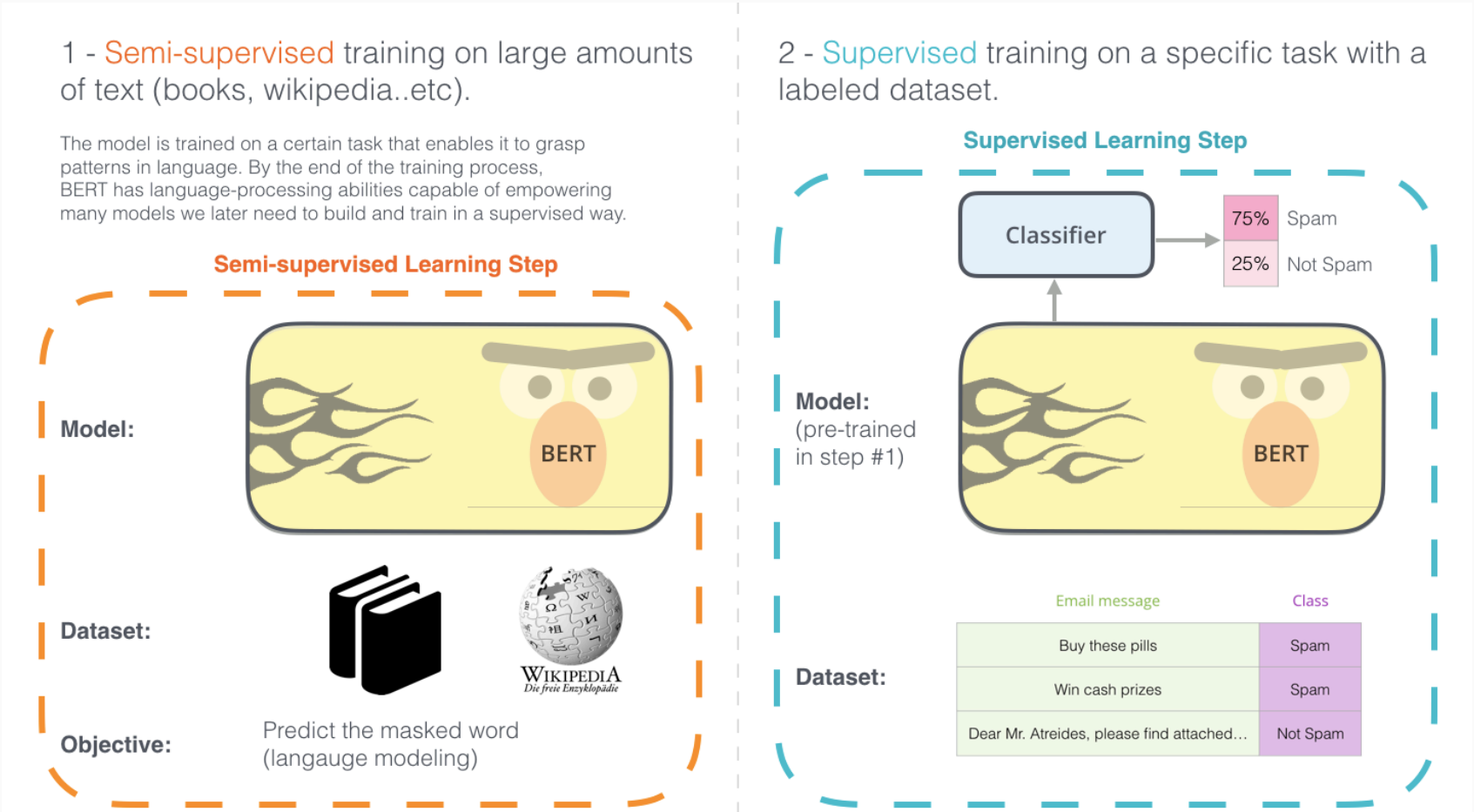
# BERT-семейство

- И даже более грамматически (Clark et al., 2019):



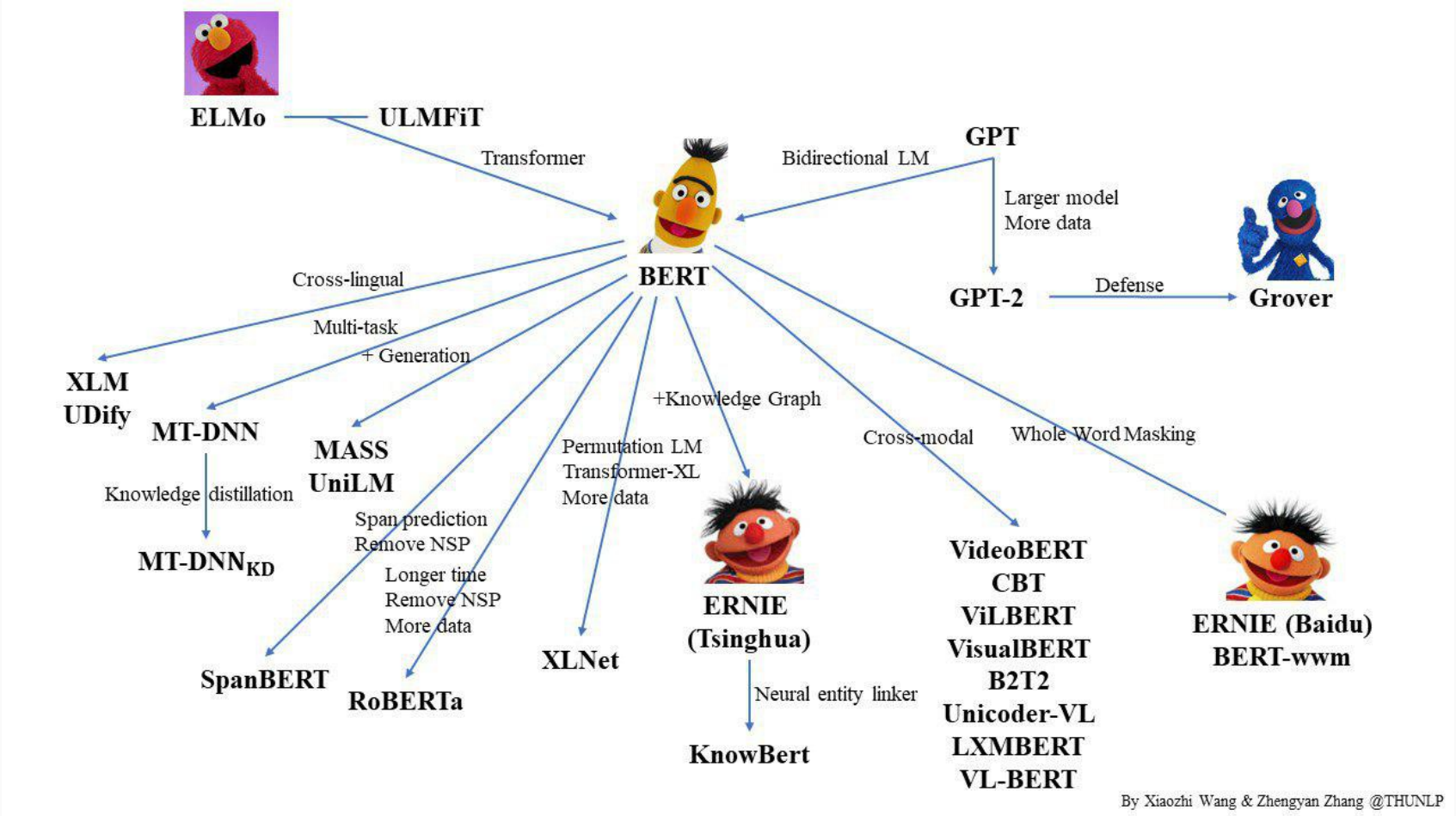
# BERT-семейство

- Процесс обучения: сначала semi-supervised, потом fine-tuning



# BERT-семейство

- BERT-подобных моделей уже очень много:

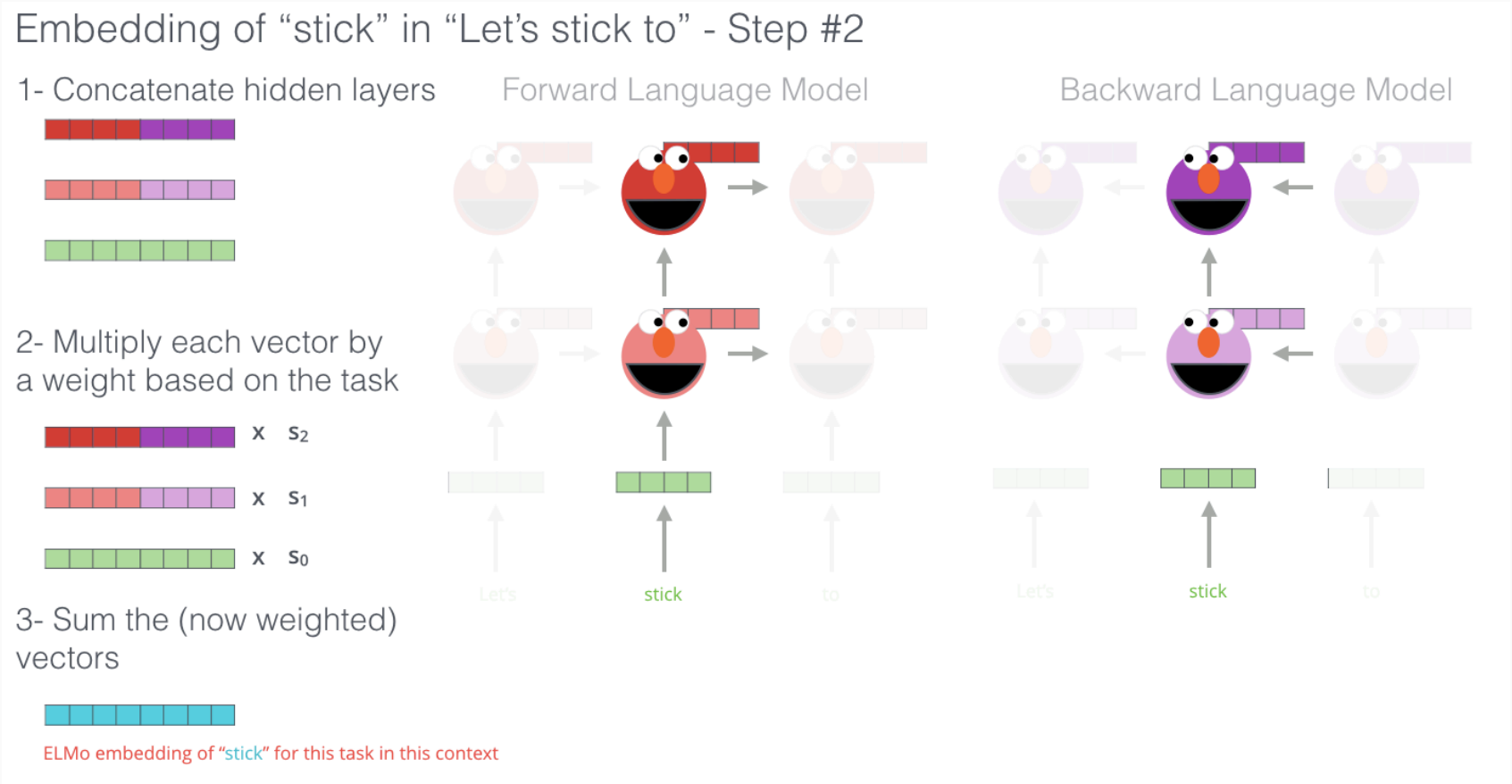


# BERT-семейство

- ELMo: embeddings с контекстом; ведь у слова много смыслов:

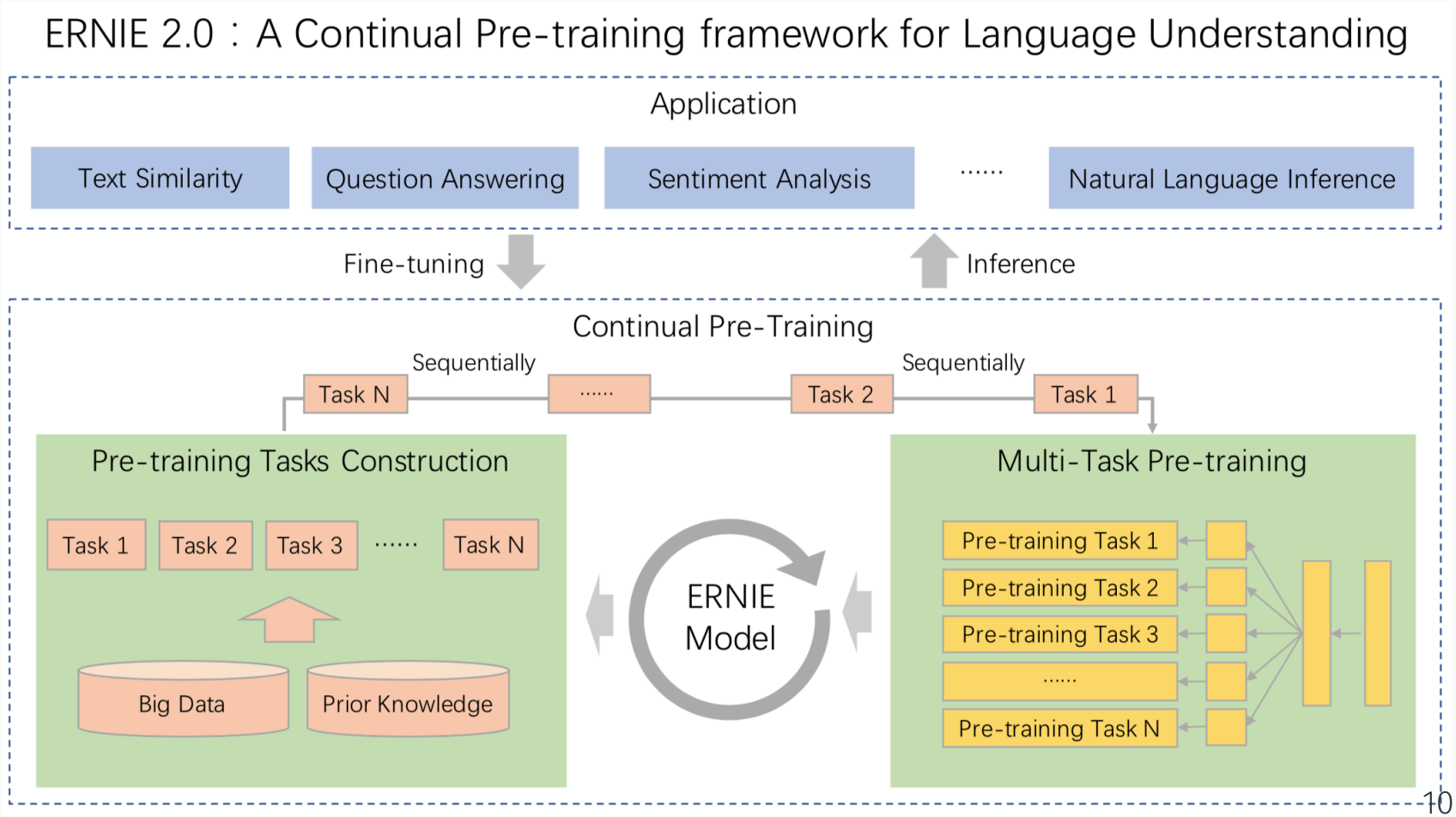
# BERT-семейство

- ELMo берёт линейную комбинацию состояний двунаправленного LSTM с весами, зависящими от конкретной задачи:



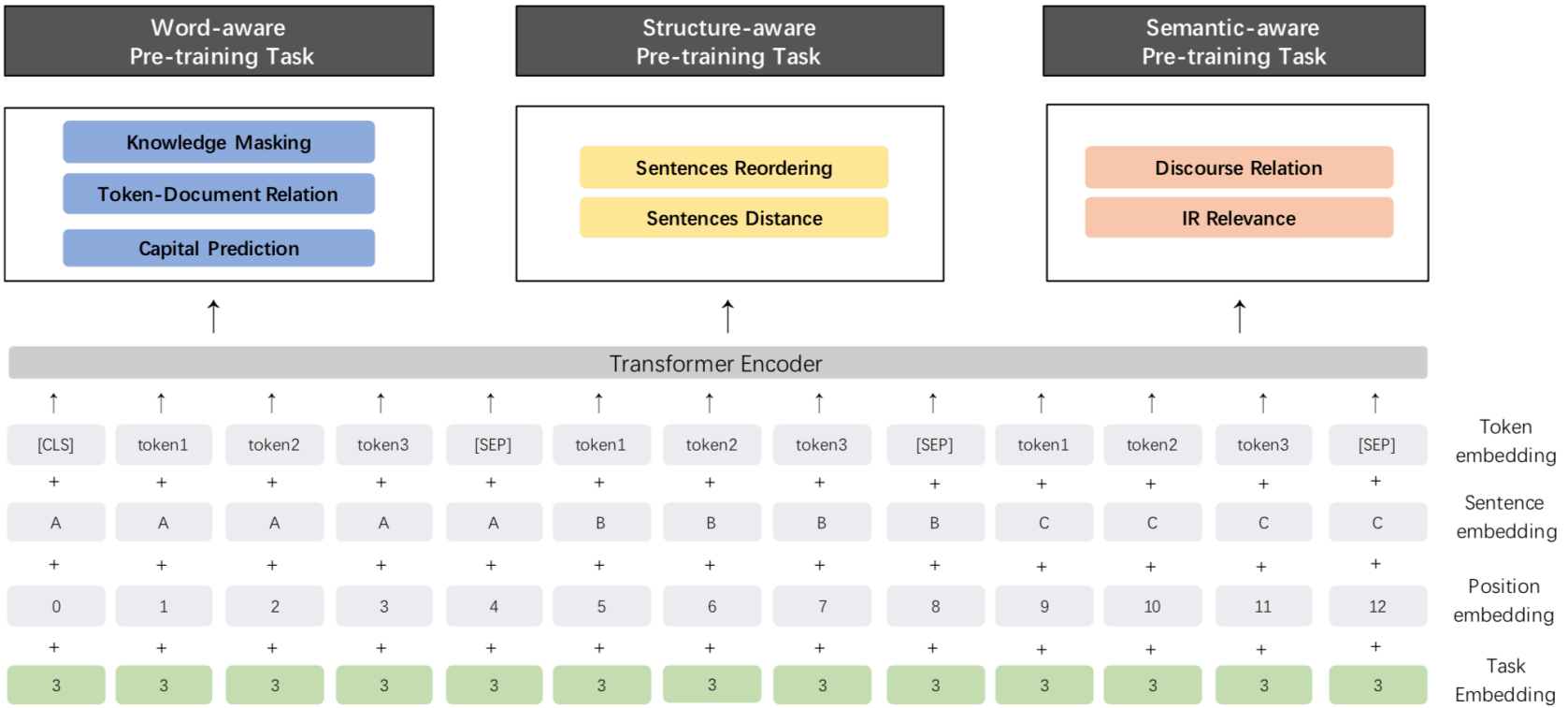
# BERT-семейство

- ERNIE (Zhang et al., 2019) строит пайплайн предобучения на основе разных задач, к которым можно легко породить примеры:



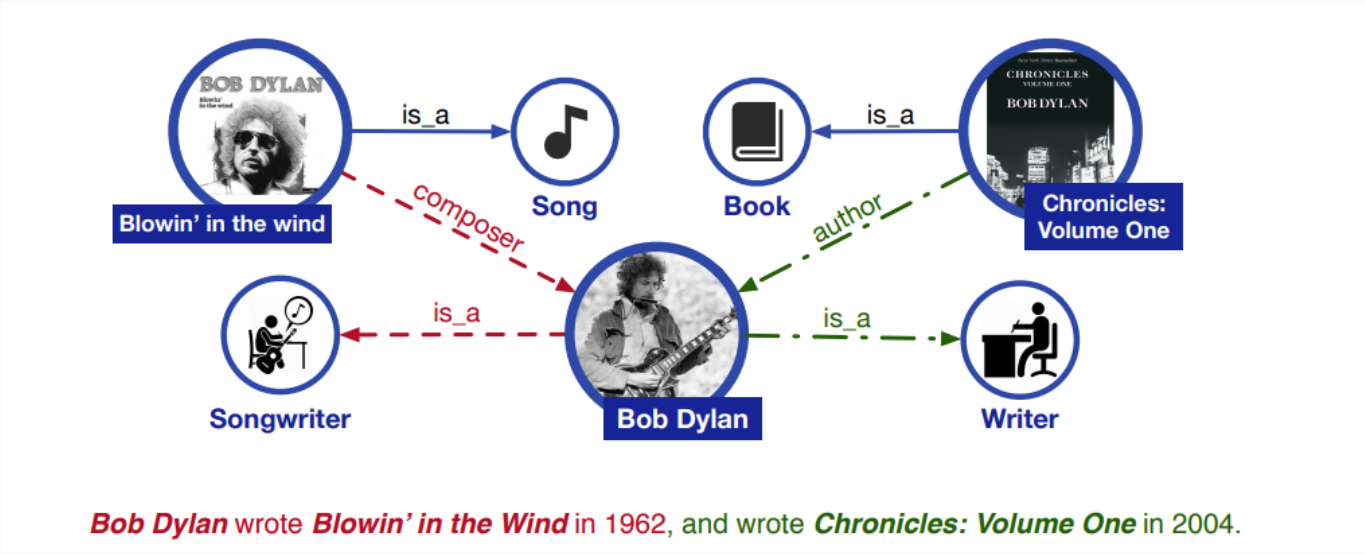
# BERT-семейство

- Задачи типа найти в контексте что-то, предсказать капитализацию:



# BERT-семейство

- Но не только, ещё добавляем знания из knowledge graph:



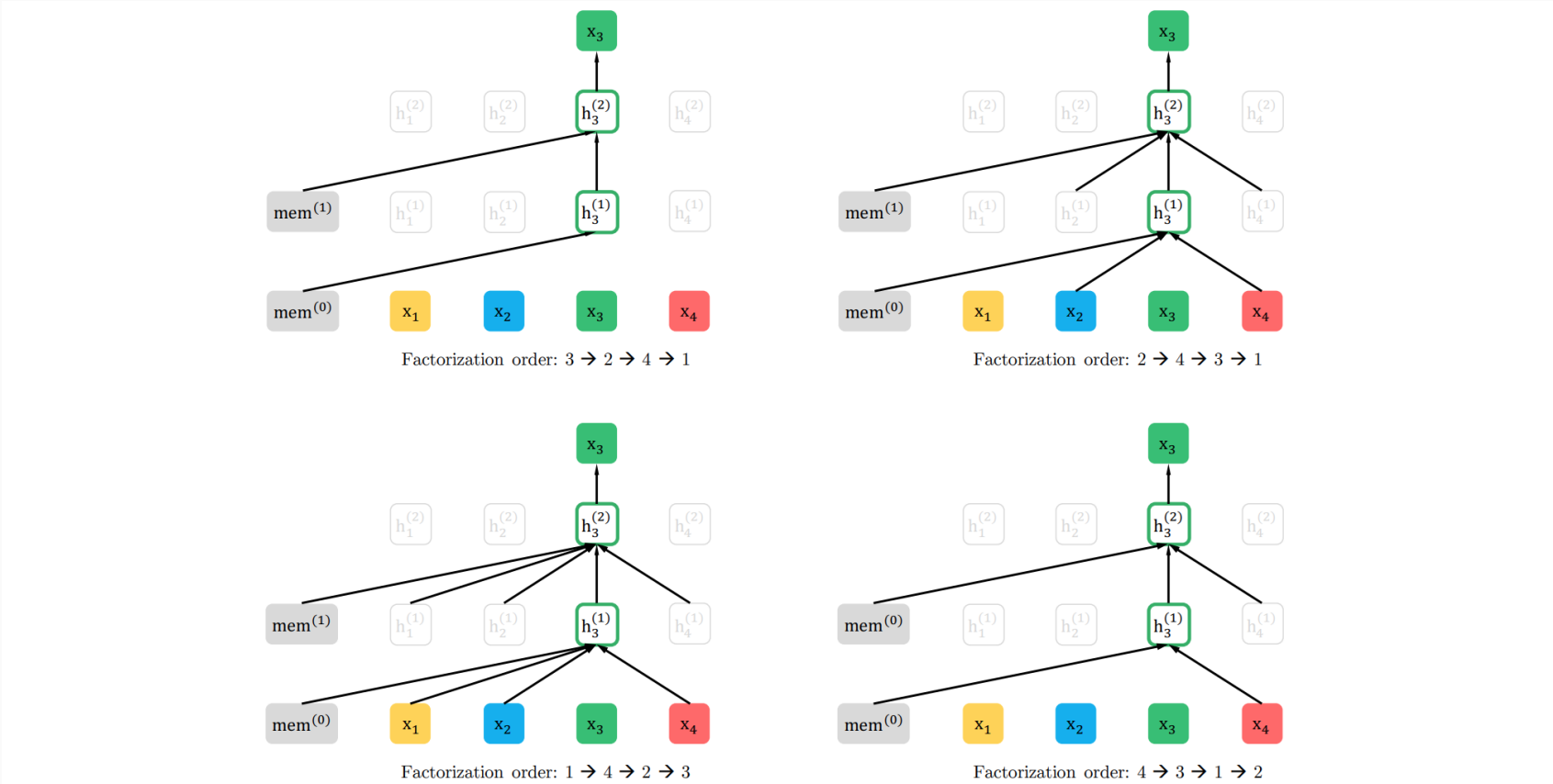

- В реальности это значит, что маски по-умному строятся:

```
- Learned by BERT :[mask] Potter is a series [mask] fantasy novel [mask] by J. [mask] Rowli

- Learned by ERNIE:Harry Potter is a series of [mask] [mask] written by [mask] [mask] [mask
```

# BERT-семейство

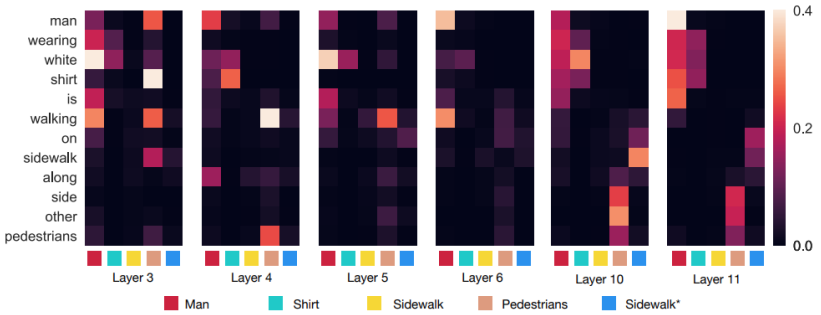
- XLNet (Yang et al., 2019): BERT предсказывает маскированные слова, а XLNet пытается предсказывать данное слово сразу во всех перестановках входного предложения:



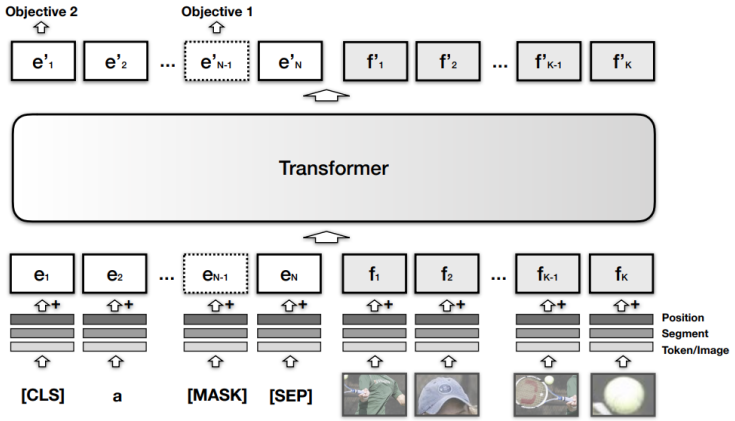

- На самом деле сэмплируют при обучении один случайный порядок, но параметры остаются общими

# BERT-семейство

- Visual BERT (Li et al., 2019) – Show, Attend, and Tell, только через BERT:

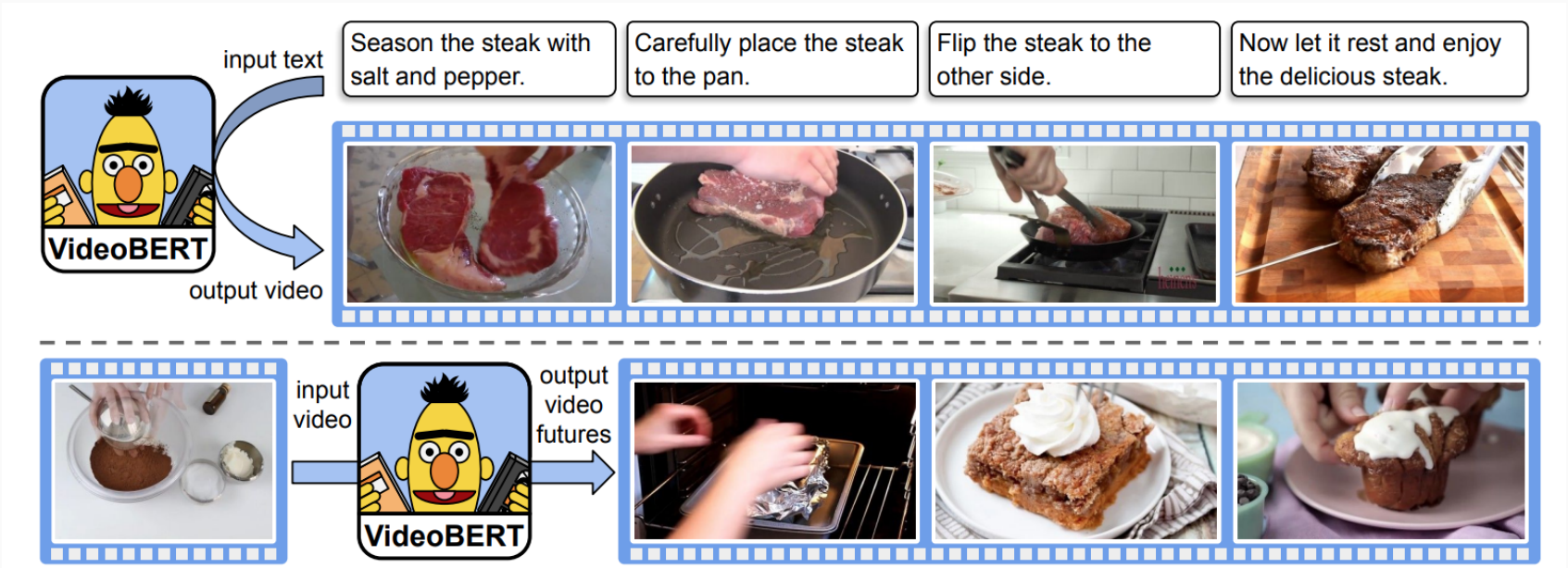



A person hits a ball with a tennis racket



# BERT-семейство

- VideoBERT (Sun et al., 2019) – давайте к предложениям ещё видео приложим:



# BERT-семейство

- В результате получается, например, video captioning:


**GT**: add some chopped basil leaves into it

**VideoBERT**: chop the basil and add to the bowl

**S3D**: cut the tomatoes into thin slices


**GT**: cut the top off of a french loaf

**VideoBERT**: cut the bread into thin slices

**S3D**: place the bread on the pan


**GT**: cut yu choy into diagonally medium pieces

**VideoBERT**: chop the cabbage

**S3D**: cut the roll into thin slices


**GT**: remove the calamari and set it on paper towel

**VideoBERT**: fry the squid in the pan

**S3D**: add the noodles to the pot

# BERT-семейство

- Следующая новость – GPT-2 (Radford et al., 2019).
- Это тот же GPT по сути, но:
  - гораздо больше размером: 1.5B параметров (у GPT было 110M, у BERT 340M);
  - обученный на огромном датасете: WebText – все ссылки с Reddit с кармой $\geq 3$, 40GB текста;
  - без всякого fine-tuning и вообще без supervision, проверялось качество в zero-shot контексте.

# BERT-семейство

- Результаты:

| | LAMBADA (PPL) | LAMBADA (ACC) | CBT-CN (ACC) | CBT-NE (ACC) | WikiText2 (PPL) | PTB (PPL) | enwik8 (BPB) | text8 (BPC) | WikiText103 (PPL) | 1BW (PPL) |
|---|---|---|---|---|---|---|---|---|---|---|
| SOTA | 99.8 | 59.23 | 85.7 | 82.3 | 39.14 | 46.54 | 0.99 | 1.08 | 18.3 | **21.8** |
| 117M | **35.13** | 45.99 | **87.65** | **83.4** | **29.41** | 65.85 | 1.16 | 1.17 | 37.50 | 75.20 |
| 345M | **15.60** | 55.48 | **92.35** | **87.1** | **22.76** | 47.33 | 1.01 | **1.06** | 26.37 | 55.72 |
| 762M | **10.87** | **60.12** | **93.45** | **88.0** | **19.93** | **40.31** | **0.97** | **1.02** | 22.05 | 44.575 |
| 1542M | **8.63** | **63.24** | **93.30** | **89.05** | **18.34** | **35.76** | **0.93** | **0.98** | **17.48** | 42.16 |

- Конечно, гораздо хуже специализированных моделей, но работает прямо само по себе, из небольшого контекста.
- `https://www.reddit.com/user/GPT-2_Bot/`
- `https://openai.com/blog/better-language-models/#sample1`

# BERT-семейство

- Вопросы (есть на BERT хорошая модель, Alberti et al., 2019):

| Question | Generated Answer | Correct | Probability |
|---|---|---|---|
| Who wrote the book the origin of species? | Charles Darwin | ✓ | 83.4% |
| Who is the founder of the ubuntu project? | Mark Shuttleworth | ✓ | 82.0% |
| Who is the quarterback for the green bay packers? | Aaron Rodgers | ✓ | 81.1% |
| Panda is a national animal of which country? | China | ✓ | 76.8% |
| Who came up with the theory of relativity? | Albert Einstein | ✓ | 76.4% |
| When was the first star wars film released? | 1977 | ✓ | 71.4% |
| What is the most common blood type in sweden? | A | ✗ | 70.6% |
| Who is regarded as the founder of psychoanalysis? | Sigmund Freud | ✓ | 69.3% |
| Who took the first steps on the moon in 1969? | Neil Armstrong | ✓ | 66.8% |
| Who is the largest supermarket chain in the uk? | Tesco | ✓ | 65.3% |
| What is the meaning of shalom in english? | peace | ✓ | 64.0% |
| Who was the author of the art of war? | Sun Tzu | ✓ | 59.6% |
| Largest state in the us by land mass? | California | ✗ | 59.2% |
| Green algae is an example of which type of reproduction? | parthenogenesis | ✗ | 56.5% |
| Vikram samvat calender is official in which country? | India | ✓ | 55.6% |
| Who is mostly responsible for writing the declaration of independence? | Thomas Jefferson | ✓ | 53.3% |
| What us state forms the western boundary of montana? | Montana | ✗ | 52.3% |
| Who plays ser davos in game of thrones? | Peter Dinklage | ✗ | 52.1% |
| Who appoints the chair of the federal reserve system? | Janet Yellen | ✗ | 51.5% |
| State the process that divides one nucleus into two genetically identical nuclei? | mitosis | ✓ | 50.7% |
| Who won the most mvp awards in the nba? | Michael Jordan | ✗ | 50.2% |
| What river is associated with the city of rome? | the Tiber | ✓ | 48.6% |
| Who is the first president to be impeached? | Andrew Johnson | ✓ | 48.3% |
| Who is the head of the department of homeland security 2017? | John Kelly | ✓ | 47.0% |
| What is the name given to the common currency to the european union? | Euro | ✓ | 46.8% |
| What was the emperor name in star wars? | Palpatine | ✓ | 46.5% |
| Do you have to have a gun permit to shoot at a range? | No | ✓ | 46.4% |
| Who proposed evolution in 1859 as the basis of biological development? | Charles Darwin | ✓ | 45.7% |
| Nuclear power plant that blew up in russia? | Chernobyl | ✓ | 45.7% |
| Who played john connor in the original terminator? | Arnold Schwarzenegger | ✗ | 45.2% |

# Пример порождения

**In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.**

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

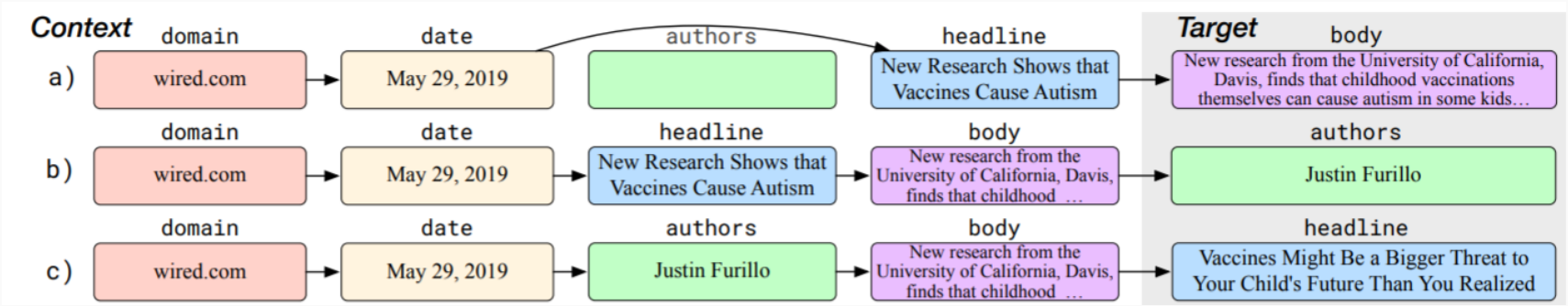
While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. "But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization," said the scientist.

# GROVER

- (Zellers et al., 2019): GROVER, модель для распознавания и порождения фейковых новостей на основе GPT-2.
- Моделируем как

$$p(\text{domain}, \text{date}, \text{authors}, \text{headline}, \text{body}).$$



# GROVER

- Получается, что машинная пропаганда людьми оценивается лучше, чем человеческая:

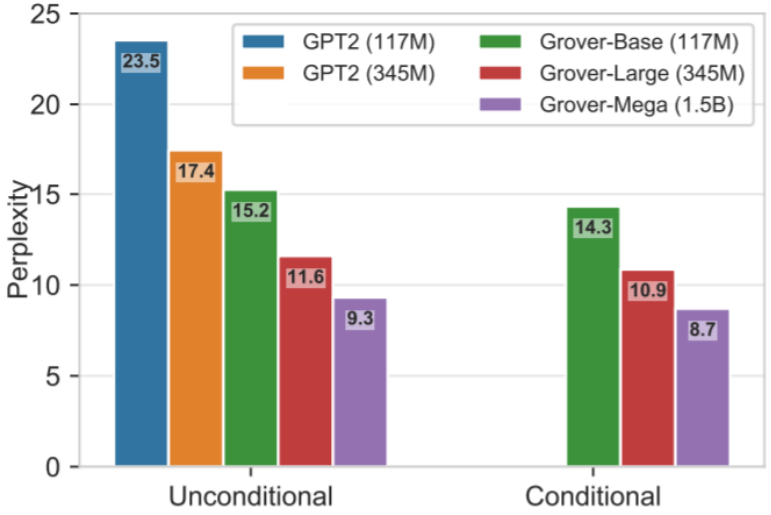



Figure 3: Language Modeling results on the body field of April 2019 articles. We evaluate in the *Unconditional* setting (without provided metadata) as well as in the *Conditional* setting (with all metadata). GROVER sees over a 0.6 point drop in perplexity when given metadata.

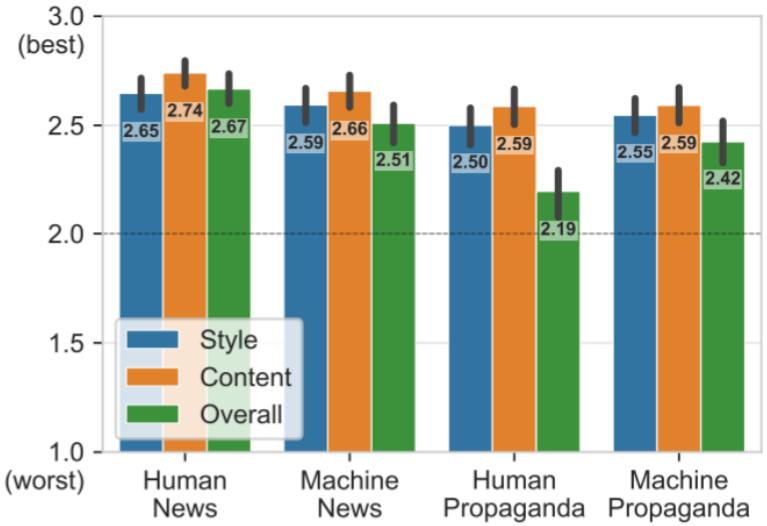



Figure 4: Human evaluation. For each article, three annotators evaluated style, content, and the overall trustworthiness; 100 articles of each category were used. The results show that propaganda generated by GROVER is rated more plausible than the original human-written propaganda.

# GROVER

- Вот наше будущее:

**Original Headline: Don't drink the water: The dark side of water fluoridation**

**Human-written Propaganda**

**Don't drink the water: The dark side of water fluoridation**
March 13. 2019 naturalnews.com

(Natural News) There are 7.7 billion people on this planet (as at March 2019). Only about 5 percent of them drink fluoridated water. Why? Because their governments recognize that fluoride in large amounts becomes a toxic chemical that is not fit for human consumption. The 328,000,000 citizens of the United States drink more fluoridated water than all other countries combined. Why? Because the U.S. government continues to doggedly insist that it is safe and improves dental health.
But what do the facts say? As reported by Waking Times, dozens of peer-reviewed studies published in prestigious journals like The Lancet, have confirmed that fluoride is in fact toxic – especially to the developing brains of children. These chemicals are derived from unprocessed toxic waste which is not purified in any way before being pumped into the water supply. How could it possibly be anything but harmful?
The history of water fluoridation in the United States
So, what prompted the government to start adding something so obviously harmful to our precious water supply?
Waking Times, quoting from an article by The Children's Health Defense Team, explains a little about the history of this practice:
During World War II, fluoride (a compound formed from the chemical element fluorine) came into large-scale production and use as part of the Manhattan Project. According to declassified government documents summarized by Project Censored, Manhattan Project scientists discovered early on that fluoride was a "leading health hazard to bomb program workers and surrounding communities." In order to stave off lawsuits, government scientists "embarked on a campaign to calm the social panic about fluoride…by promoting its usefulness in preventing tooth decay."
The power of the elements: Discover Colloidal Silver Mouthwash with quality, natural ingredients like Sangre de Drago sap, black walnut hulls, menthol crystals and more. Zero artificial sweeteners, colors or alcohol. Learn more at the Health Ranger Store and help support this news site.
To back up its decision, the government embarked on a series of flawed and poorly designed "scientific" studies, which an expert later lambasted as "especially rich in fallacies, improper

**Machine-written Propaganda**

**Fluoride in Your Water Isn't Healthy, Even When It's Recommended**
March 13, 2019 huffingtonpost.com

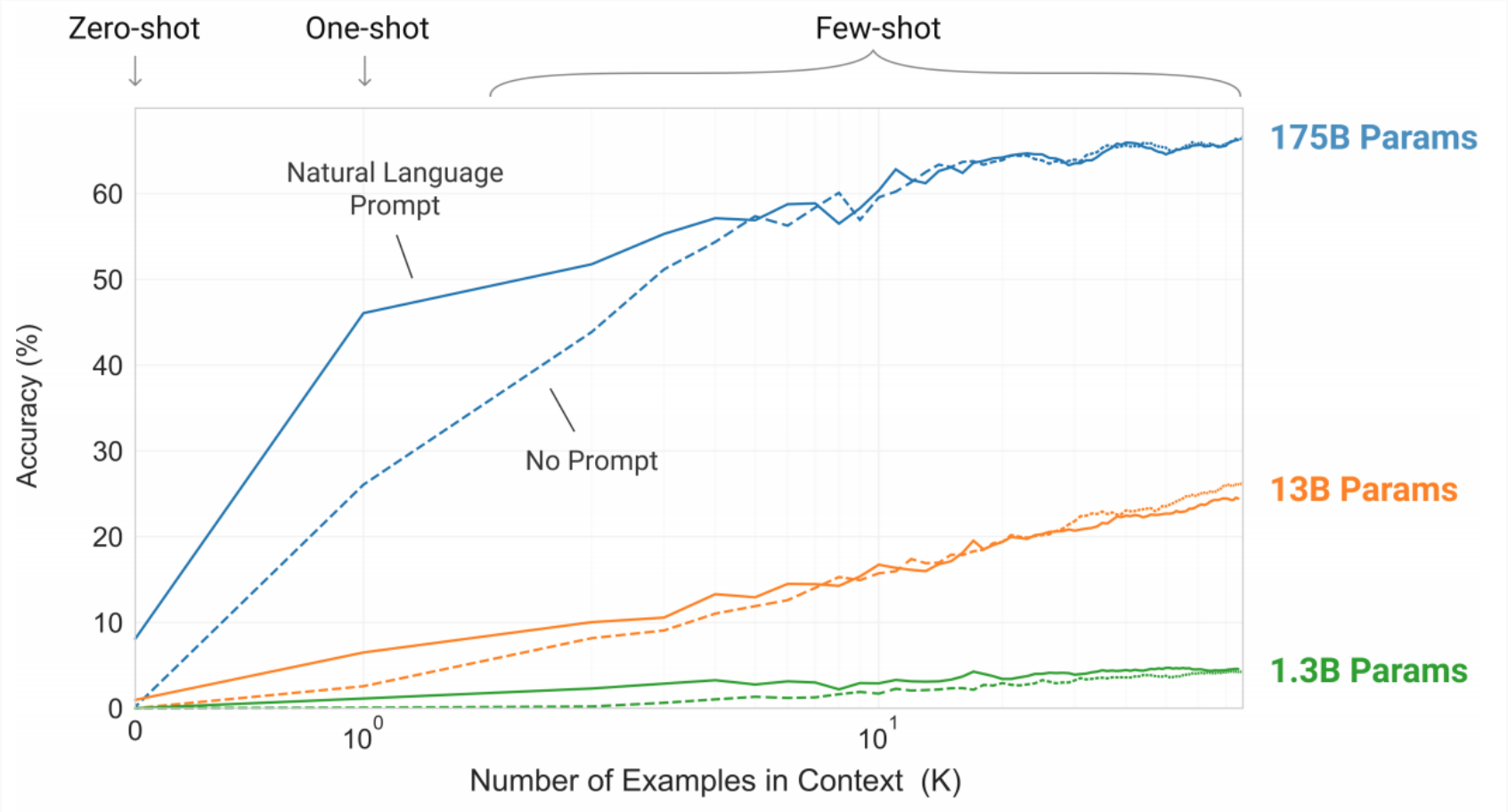
Fluoride is regarded by researchers around the world as the "gold standard" in tooth care, and a safe, common ingredient that has almost been universally found to be safe in past studies of health effects. It's absorbed without interference from the body's natural minerals.
Researchers at the U.S. Geological Survey (USGS) published the results of a multi-state environmental health study last month. It showed that during the first three decades of fluoridation of tap water systems, fluoride produced from the process alone increased rates of dental caries (the biggest contributor to tooth decay) by 16 percent in Mississippi and a whopping 45 percent in Arizona, which implemented fluoridation systems back in 1942. This increase was seen after a decade when fluoride levels didn't change.
USGS also found that fluoridation increased rates of other toxicants and petrochemicals, as well as deaths from brain, lung, kidney and bladder cancer.
It bears noting that there is no clear proof that these specific contaminants were caused by fluoridation, but the USGS study at least hints that this was the case. The epidemic of brain cancers across the U.S. — especially in teenagers — has confounded researchers for decades. The USGS study points to links to numerous studies that have linked water fluoridation with increased risks of cancer.
Even though the majority of studies on water fluoridation have not produced such alarming results, the mainstream medical community is, apparently, still skeptical. Two years ago, doctors from Harvard and Duke universities suggested that fluoride is associated with lower IQ scores and autoantibodies to water. The results of a recent study that followed more than 700 children over a period of four years demonstrated that the kids were more likely to have symptoms of illness, more likely to have higher blood pressure and sleep problems, had higher mean energy expenditure,

# GPT-3

- Ну и, конечно, последние новости – GPT-3 (Brown et al., 2020).
- Это тот же GPT по сути, но:
  - гораздо больше размером: 175B параметров! (у GPT-2 было 1.5B, у самой большой модели на тот момент 17B);
  - обученный на огромном датасете: WebText – все ссылки с Reddit с кармой $\geq 3$, 40GB текста;
  - вместо one-shot и zero-shot learning переходят-таки к few-shot.

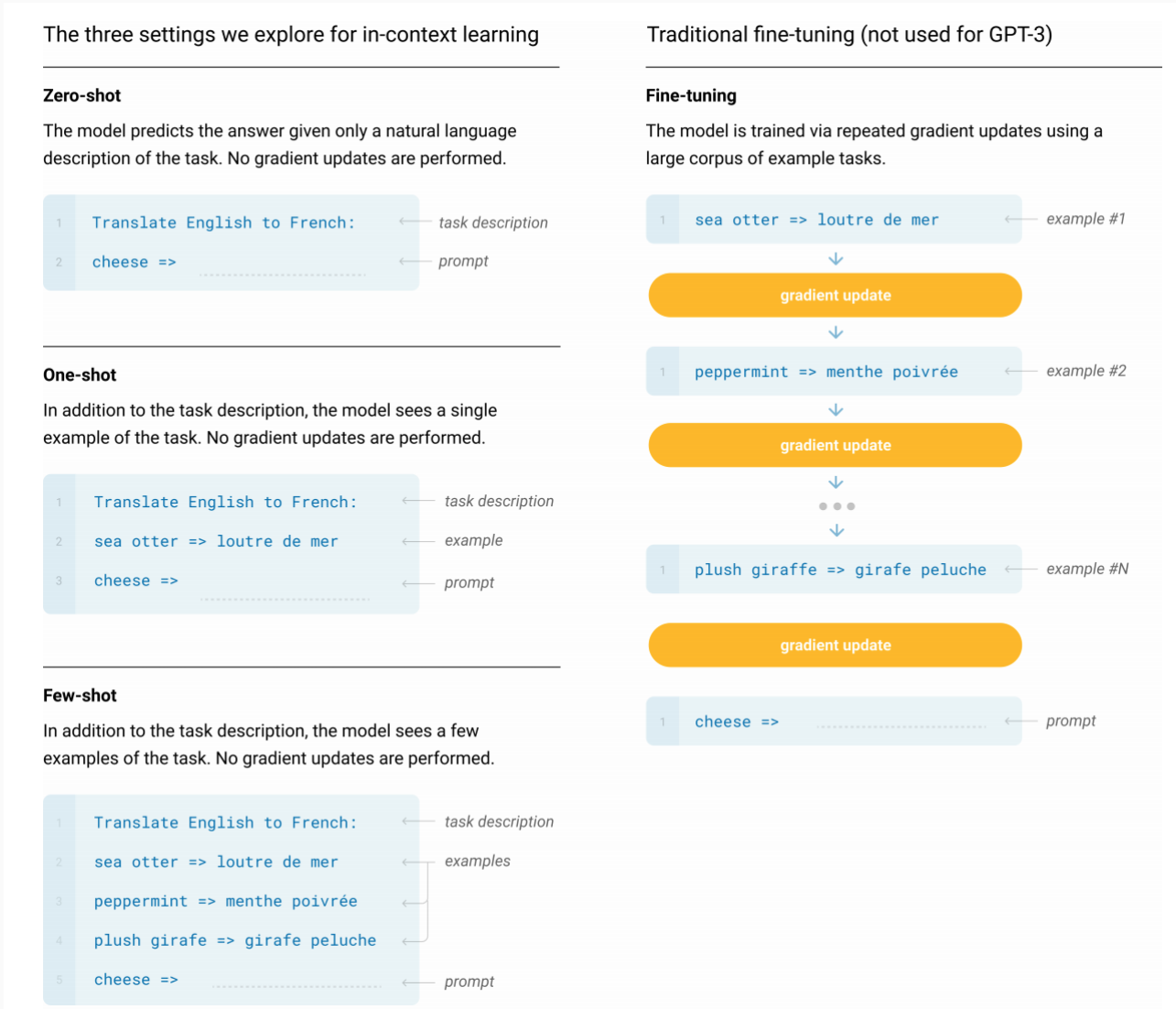
# GPT-3

- Это помогает, и помогает увеличить контекст:



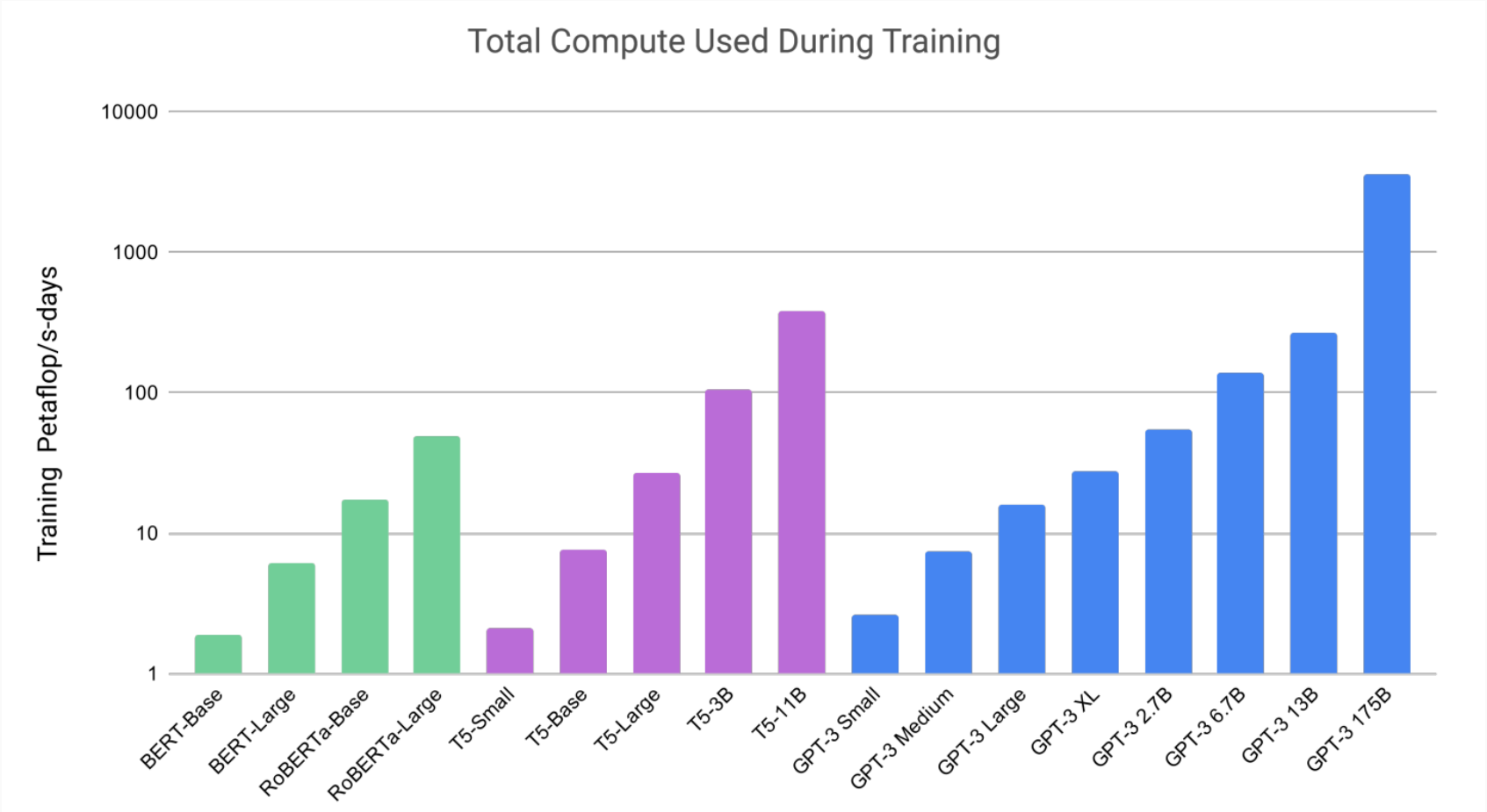
# GPT-3

- Разные типы вспомогательных задач:



# GPT-3

- Очень много compute:



# GPT-3

- Очень большие датасеты:

| Dataset | Quantity (tokens) | Weight in training mix | Epochs elapsed when training for 300B tokens |
|---|---|---|---|
| Common Crawl (filtered) | 410 billion | 60% | 0.44 |
| WebText2 | 19 billion | 22% | 2.9 |
| Books1 | 12 billion | 8% | 1.9 |
| Books2 | 55 billion | 8% | 0.43 |
| Wikipedia | 3 billion | 3% | 3.4 |

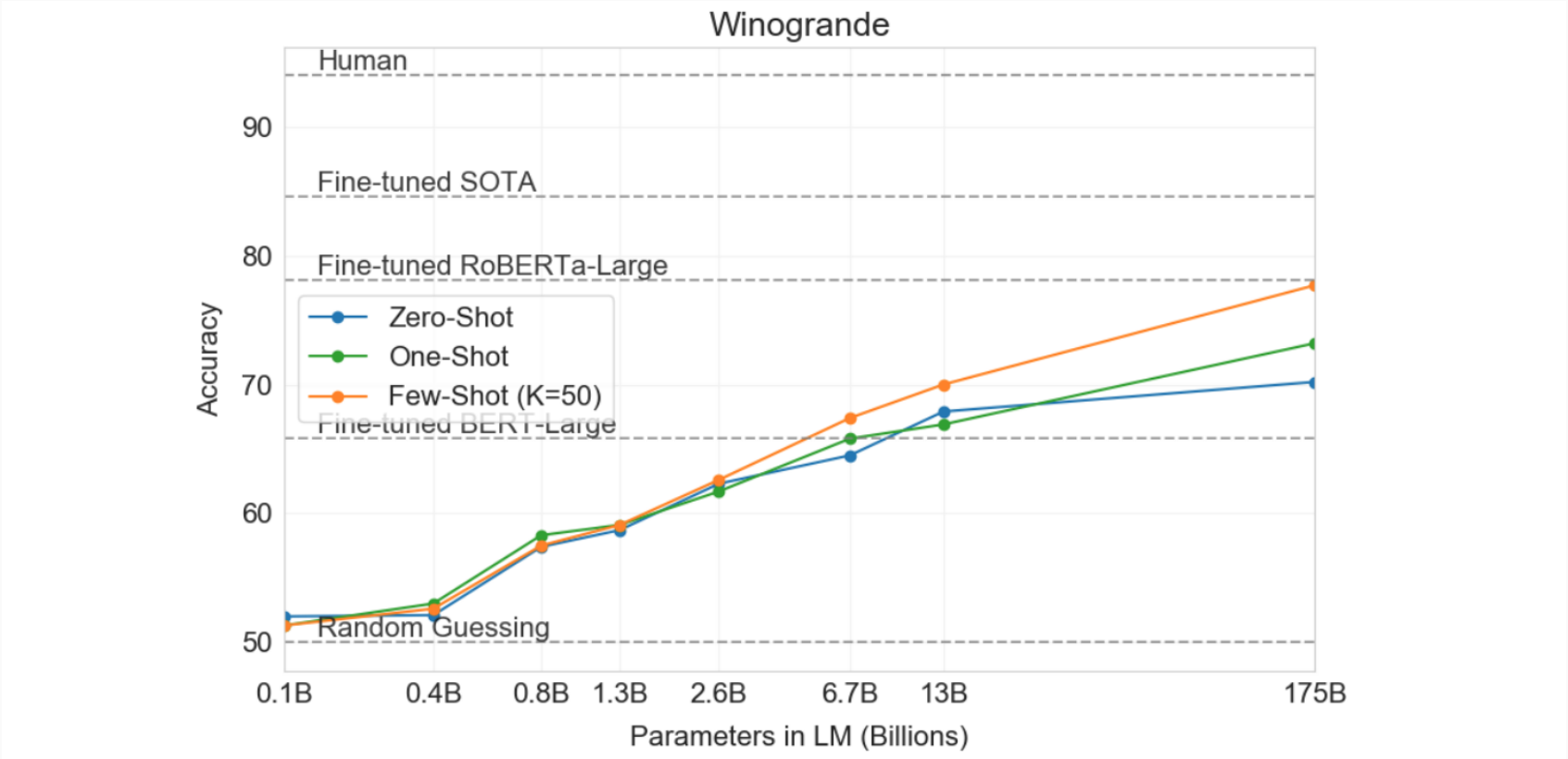
- Улучшенные результаты на стандартных задачах (86.4% на вот таких):

Alice was friends with Bob. Alice went to visit her friend _____. → Bob

George bought some baseball equipment, a ball, a glove, and a _____. →

# GPT-3

- Лучше результаты на задачах на понимание:



# GPT-3

- Примерно таких (это вот и есть Winograd schema):

| | | Twin sentences | Options (**answer**) |
|---|---|---|---|
| ✓ (1) | a | The trophy doesn't fit into the brown suitcase because **it**'s too *large*. | **trophy** / suitcase |
| | b | The trophy doesn't fit into the brown suitcase because **it**'s too *small*. | trophy / **suitcase** |
| ✓ (2) | a | Ann asked Mary what time the library closes, *because* **she** had forgotten. | **Ann** / Mary |
| | b | Ann asked Mary what time the library closes, *but* **she** had forgotten. | Ann / **Mary** |
| ✗ (3) | a | The tree fell down and crashed through the roof of my house. Now, I have to get **it** *removed*. | **tree** / roof |
| | b | The tree fell down and crashed through the roof of my house. Now, I have to get **it** *repaired*. | tree / **roof** |
| ✗ (4) | a | The lions ate the zebras because **they** are *predators*. | **lions** / zebras |
| | b | The lions ate the zebras because **they** are *meaty*. | lions / **zebras** |

Table 1: WSC problems are constructed as pairs (called *twin*) of nearly identical questions with two answer choices. The questions include a *trigger word* that flips the correct answer choice between the questions. Examples (1)-(3) are drawn from WSC (Levesque, Davis, and Morgenstern 2011) and (4) from DPR (Rahman and Ng 2012)). Examples marked with ✗ have language-based bias that current language models can easily detect. Example (4) is undesirable since the word "predators" is more often associated with the word "lions", compared to "zebras"

- Или таких:

| | |
|---|---|
| Context → | Title: The_Blitz<br><br>Background: From the German point of view, March 1941 saw an improvement. The Luftwaffe flew 4,000 sorties that month, including 12 major and three heavy attacks. The electronic war intensified but the Luftwaffe flew major inland missions only on moonlit nights. Ports were easier to find and made better targets. To confuse the British, radio silence was observed until the bombs fell. X- and Y-Gerät beams were placed over false targets and switched only at the last minute. Rapid frequency changes were introduced for X-Gerät, whose wider band of frequencies and greater tactical flexibility ensured it remained effective at a time when British selective jamming was degrading the effectiveness of Y-Gerät.<br><br>Q: How many sorties were flown in March 1941?<br><br>A: 4,000<br><br>Q: When did the Luftwaffe fly inland missions?<br><br>A: |
| Target Completion → | only on moonlit nights |

**Figure G.28:** Formatted dataset example for SQuADv2

# Спасибо!

## Спасибо за внимание!