«Transformer Feed-Forward Layers Are Key-Value Memories»

Работу выполнил

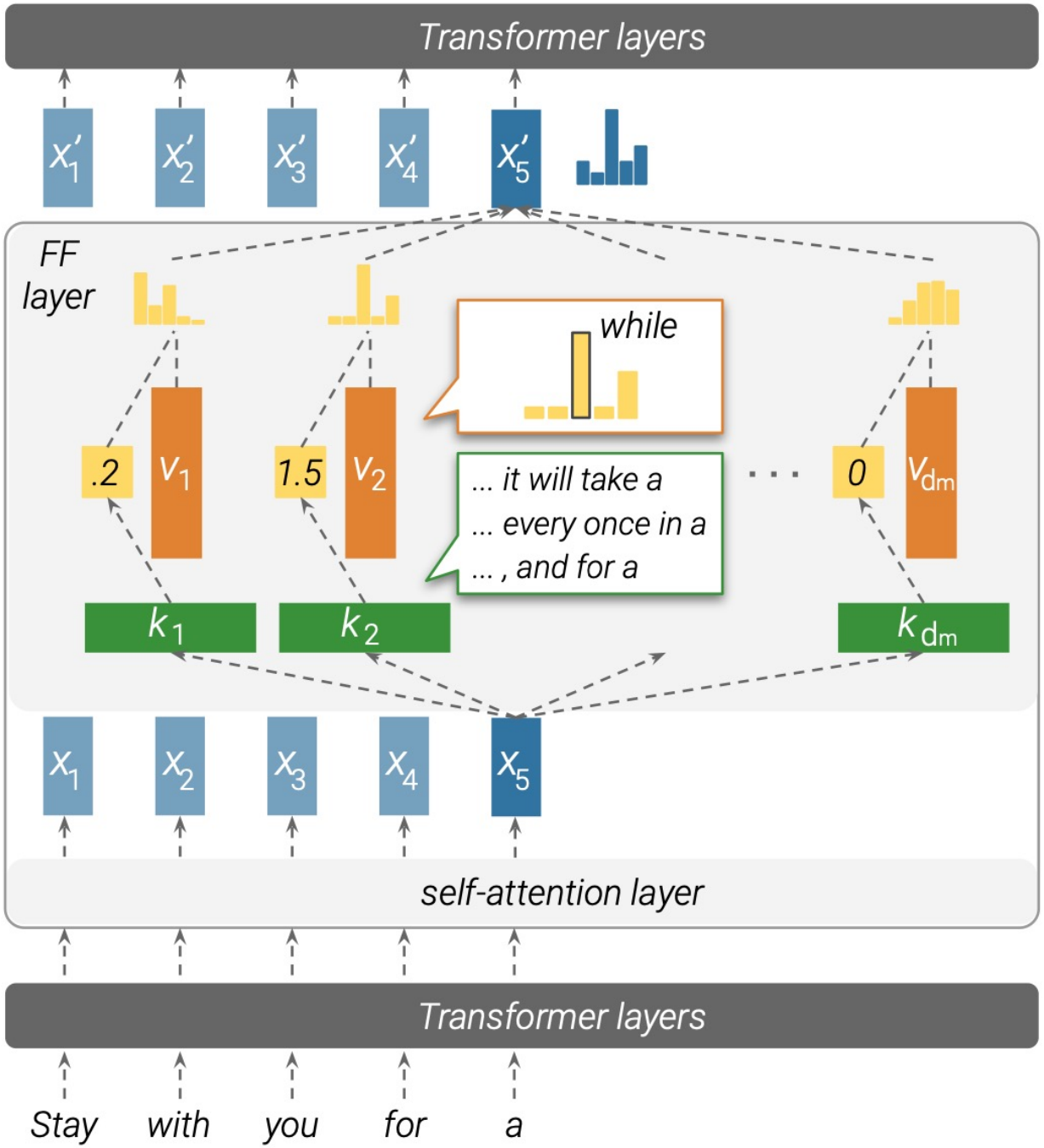
Макаров Г. Н.

Студент 193 группы

«Слои прямого распространения в Трансформерах образуют структуру памяти key-value»

Вступление

1. Основная задача Feed-forward слоёв в Трансформерах
2. Слои прямого распространения являются ненормированной структурой памяти key-value
3. Ключи группируют входные префиксы по set of patterns
4. Values представляют собой распределения токенов
5. Feed-forward слои порождают свою структуру памяти



Структура памяти Key-Value (Unnormalized)

Пусть $\mathbf{x} \in \mathbb{R}^d$ – вектор, соответствующий текстовому префиксу, тогда:

$$\mathrm{FF}(\mathbf{x}) = f(\mathbf{x} \cdot K^\top) \cdot V \qquad (1)$$

Here, $K, V \in \mathbb{R}^{d_m \times d}$ are parameter matrices, and f is a non-linearity such as ReLU.

Где FF, K, V – задают каждый FF слой такой функцией, которая каждый вектор X обрабатывает независимо

$\mathbf{k}_i \in \mathbb{R}^d$, $K \in \mathbb{R}^{d_m \times d}$; $V \in \mathbb{R}^{d_m \times d}$.

Решили оценивать вероятность $\mathbf{k}_i$ при $\mathbf{x}$ как произведение векторов $\mathbf{k}_i$

Перебираем i, ищем самый подходящий ключ

$$p(k_i \mid x) \propto \exp(\mathbf{x} \cdot \mathbf{k}_i)$$

$$\mathbf{MN}(\mathbf{x}) = \sum_{i=1}^{d_m} p(k_i \mid x)\mathbf{v}_i$$

В силу оценки слева, можем сказать, что $\sum_{i=1}^{d_m} p(k_i \mid x)$

Это – $\mathrm{softmax}(\mathbf{x} \cdot K^\top)$

$$\mathbf{MN}(\mathbf{x}) = \mathrm{softmax}(\mathbf{x} \cdot K^\top) \cdot V \qquad (2)$$

Все различия между (1) и (2) состоят в том, что в качестве функции в NM используется Softmax

$\mathbf{m} = f(\mathbf{x} \cdot K^\top)$ – функция активации $\mathbf{m}_i$ *memory coefficient*

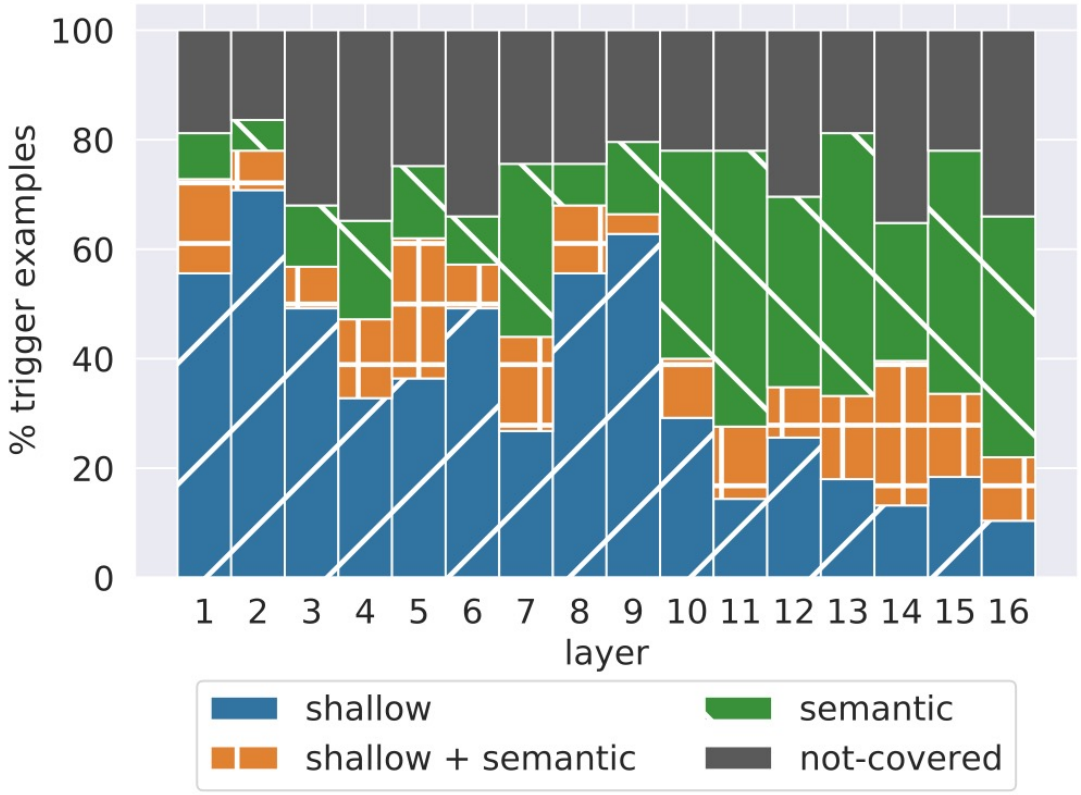
Идентификация шаблонов

Key	Pattern	Example trigger prefixes
$\mathbf{k}^{1}_{449}$	Ends with *"substitutes"* (shallow)	*At the meeting, Elton said that "for artistic reasons there could be no substitutes* *In German service, they were used as substitutes* *Two weeks later, he came off the substitutes*
$\mathbf{k}^{6}_{2546}$	Military, ends with *"base"*/*"bases"* (shallow + semantic)	*On 1 April the SRSG authorised the SADF to leave their bases* *Aircraft from all four carriers attacked the Australian base* *Bombers flying missions to Rabaul and other Japanese bases*
$\mathbf{k}^{10}_{2997}$	a "part of" relation (semantic)	*In June 2012 she was named as one of the team that competed* *He was also a part of the Indian delegation* *Toy Story is also among the top ten in the BFI list of the 50 films you should*
$\mathbf{k}^{13}_{2989}$	Ends with a time range (semantic)	*Worldwide, most tornadoes occur in the late afternoon, between 3 pm and 7* *Weekend tolls are in effect from 7:00 pm Friday until* *The building is open to the public seven days a week, from 11:00 am to*
$\mathbf{k}^{16}_{1935}$	TV shows (semantic)	*Time shifting viewing added 57 percent to the episode's* *The first season set that the episode was included in was as part of the* *From the original NBC daytime version , archived*

Существует 3 типа patterns: shallow, shallow+semantic, semantic

Эксперимент

Взяли префиксы всех предложений и трактуя их как вход X посчитали для них memory coef. [ReLU(x * k)]
И так для всех слоев (у нас есть разные hidden слои, на каждом свои ключи k.
Далее, на каждом таком слое взяли какой-то top таких префиксов с максимальным memory coef. И снова попросили людей выделить в каждой группе таких префиксов с максимальным memory coef (у каждого слоя была одна своя такая группа префиксов с максимальным memory coef) pattern, а также указать природу этого pattern (shallow, semantic или что-то между) на графике это показано разными цветами. Получили, что на первых hidden слоях паттерны в основном shallow, а на последних - semantic.

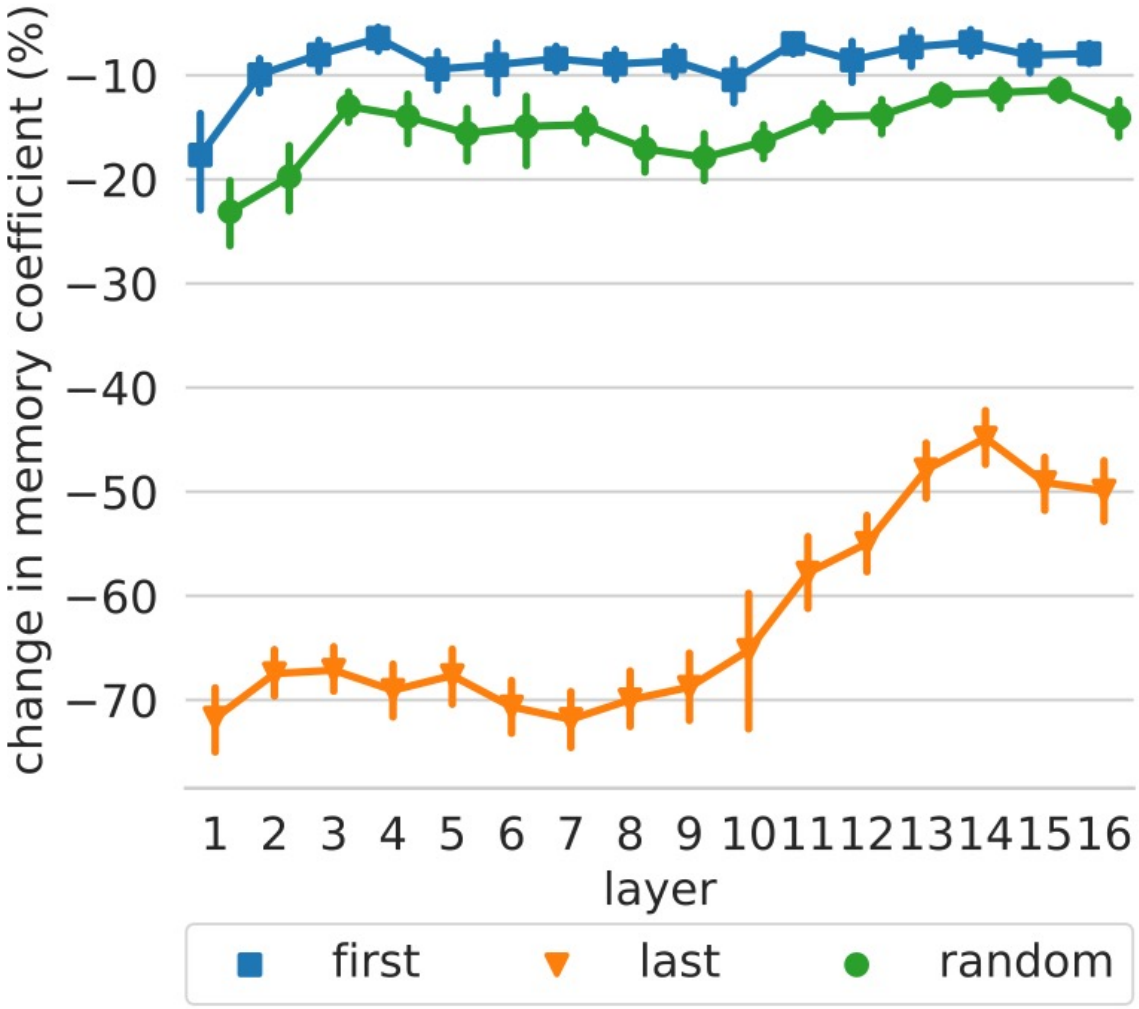


Изменение в memory coef.

График показывает, что на 1-10 слоях после удаления последнего токена memory coef. префиксов из топа падает сильнее всего (на 60%+), в то время как на 11-16 слоях memory coef. префиксов из выделенного топа падает значительно слабее (на 40%)

Под топ префиксами подразумеваются префиксы с максимальным MC (на каждом они свой топ)

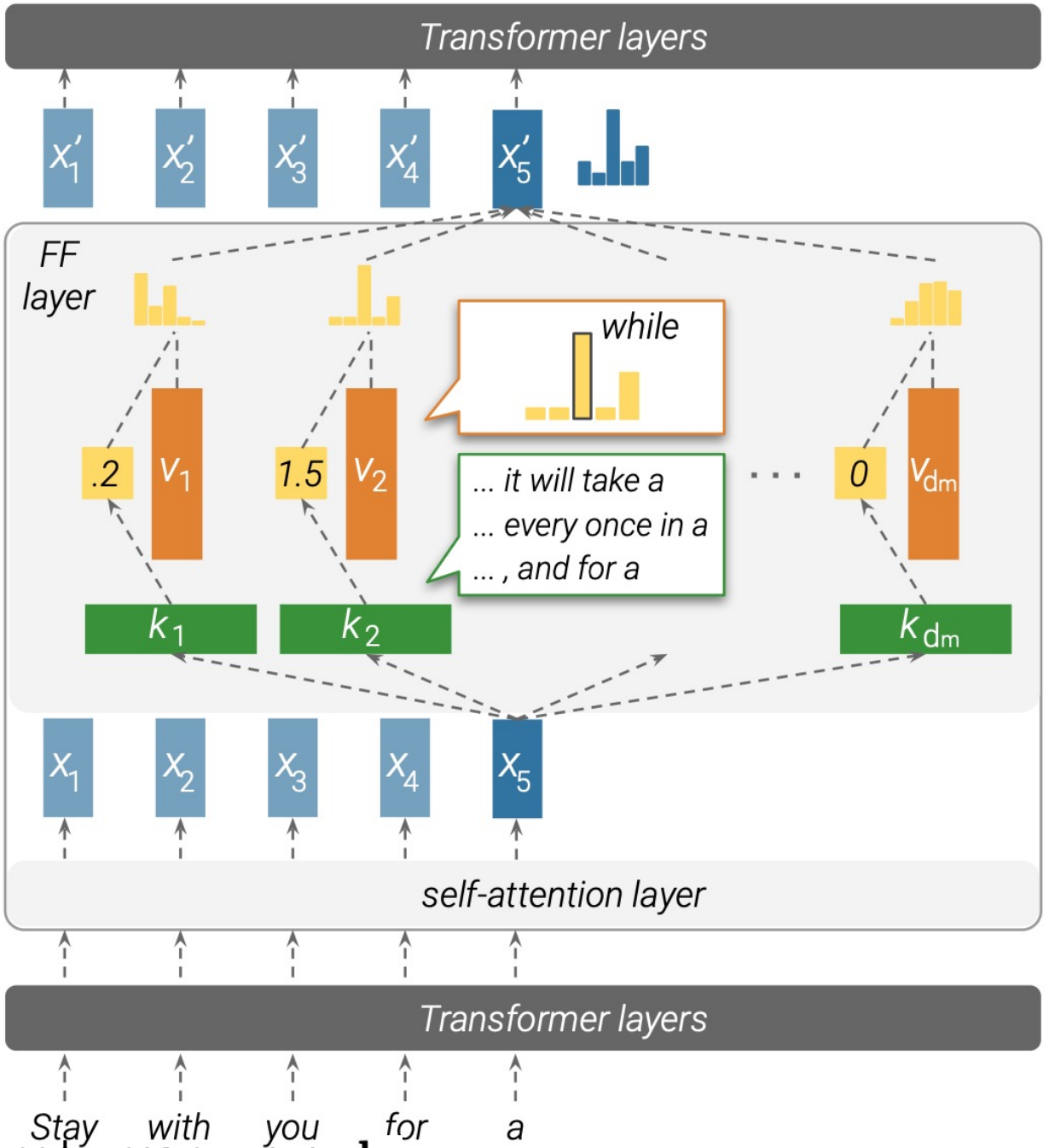
При этом от удаления первого или случайного токена – изменение происходит в меньшей степени (до 25%) Но при этом на низких слоях оно наибольшее.



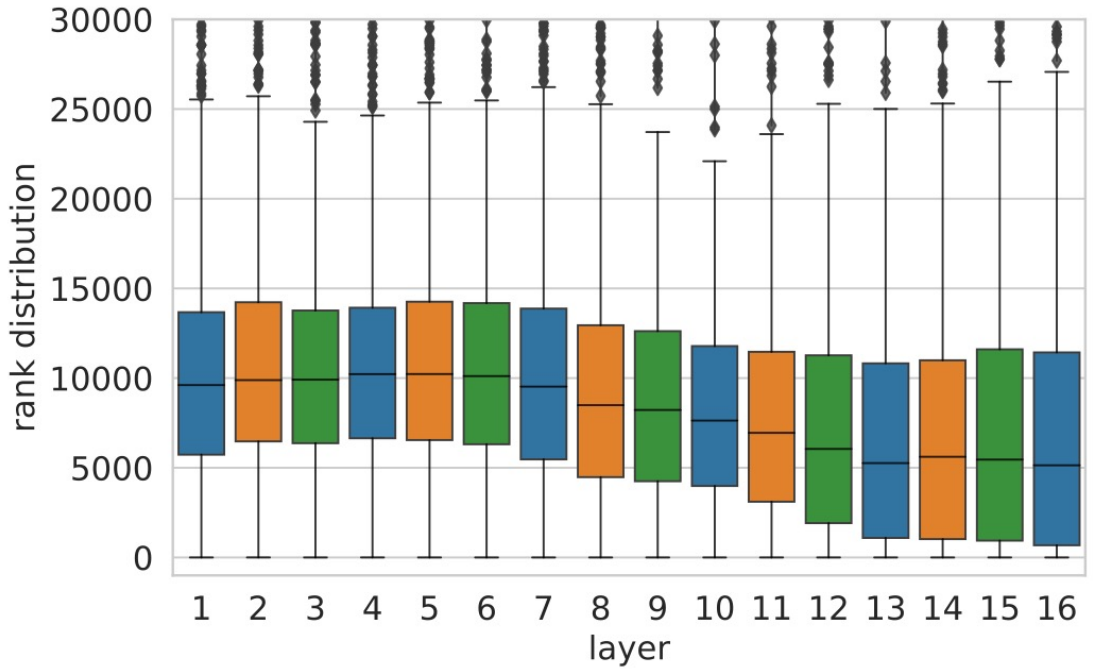
Agreement rate

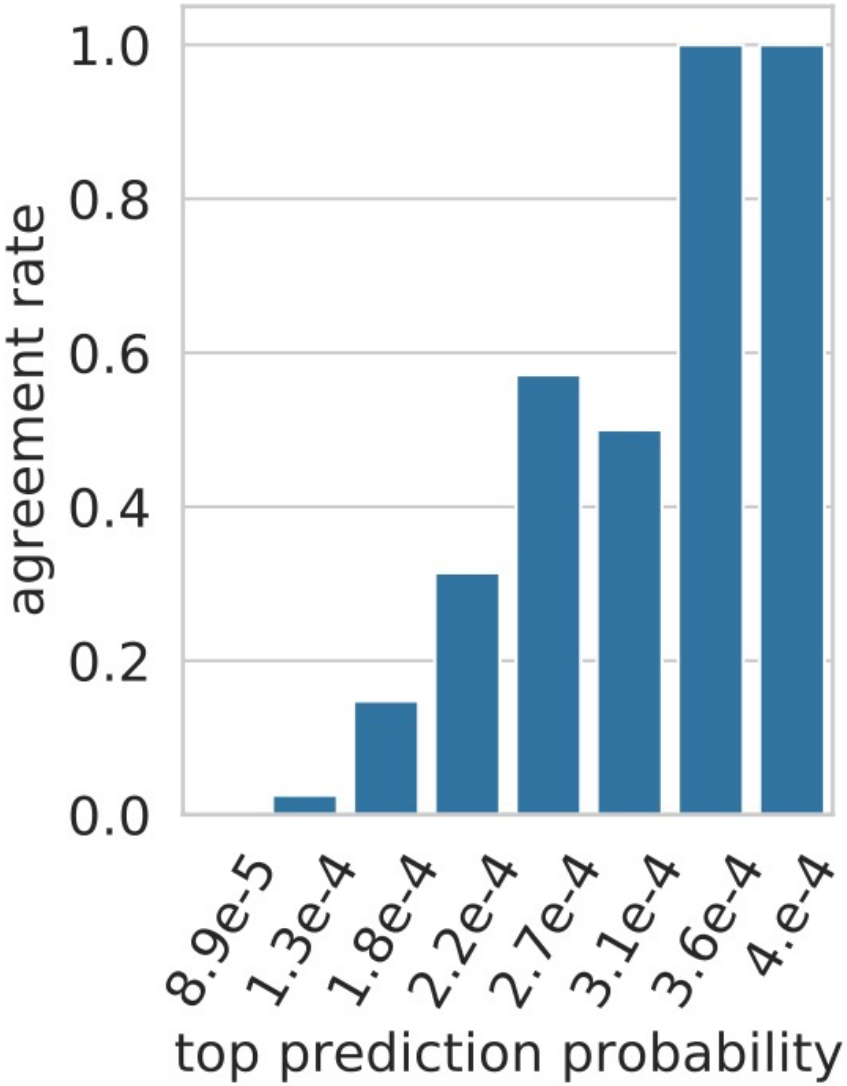
$$\mathbf{p}_i^\ell = \text{softmax}(\mathbf{v}_i^\ell \cdot E) \quad \text{argmax}(\mathbf{p}_i^\ell) = w_i^\ell$$

В конце результат MN(x)/FF(x) умножают на матрицу эмбедингов, чтобы получить слова. Здесь мы умножаем каждый вектор $\mathbf{v}_i^\ell \cdot E$ получаем какое-то распределение токенов для каждого $\mathbf{v}_i^\ell$



Согласованность между w_i^ℓ и следующим токеном у топ префиксов с ключем $\mathbf{k}_i$


rank distribution
30000
25000
20000
15000
10000
5000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
layer


agreement rate
1.0
0.8
0.6
0.4
0.2
0.0
8.9e-5
1.3e-4
1.8e-4
2.2e-4
2.7e-4
3.1e-4
3.6e-4
4.e-4
top prediction probability

Value	Prediction	Precision@50	Trigger example
$\mathbf{v}^{15}_{222}$	*each*	68%	*But when bees and wasps resemble each*
$\mathbf{v}^{16}_{752}$	*played*	16%	*Her first role was in Vijay Lalwani's psychological thriller Karthik Calling Karthik, where Padukone was cast as the supportive girlfriend of a depressed man (played*
$\mathbf{v}^{13}_{2601}$	*extratropical*	4%	*Most of the winter precipitation is the result of synoptic scale, low pressure weather systems (large scale storms such as extratropical*
$\mathbf{v}^{15}_{881}$	*part*	92%	*Comet served only briefly with the fleet, owing in large part*
$\mathbf{v}^{16}_{2070}$	*line*	84%	*Sailing from Lorient in October 1805 with one ship of the line*
$\mathbf{v}^{12}_{3186}$	*jail*	4%	*On May 11, 2011, four days after scoring 6 touchdowns for the Slaughter, Grady was sentenced to twenty days in jail*

Table 2: Example values, their top prediction, the fraction of their key's top-50 trigger examples that agree with their prediction, and a matching trigger example (with the target token marked in blue).

Всем спасибо за внимание!