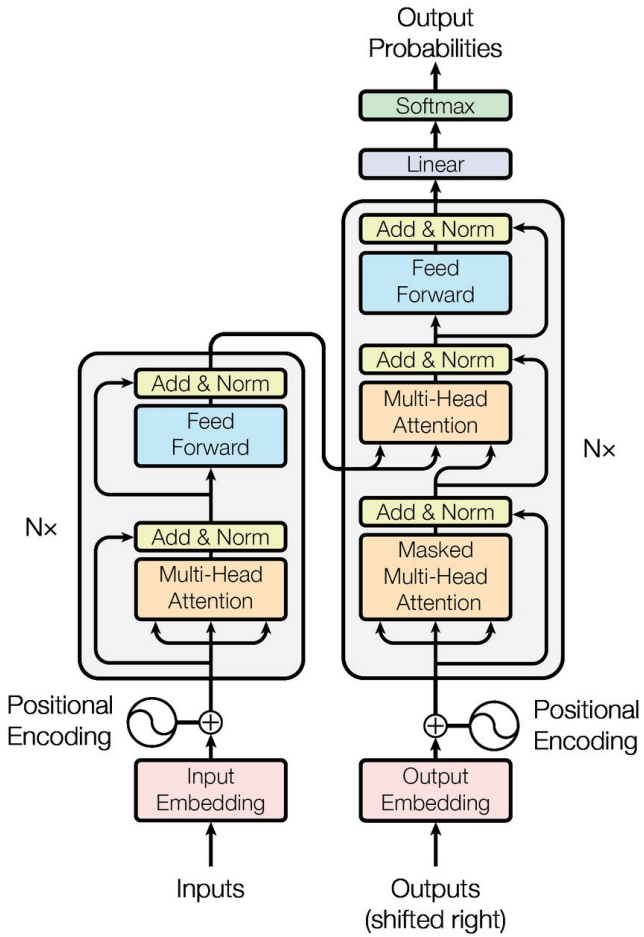
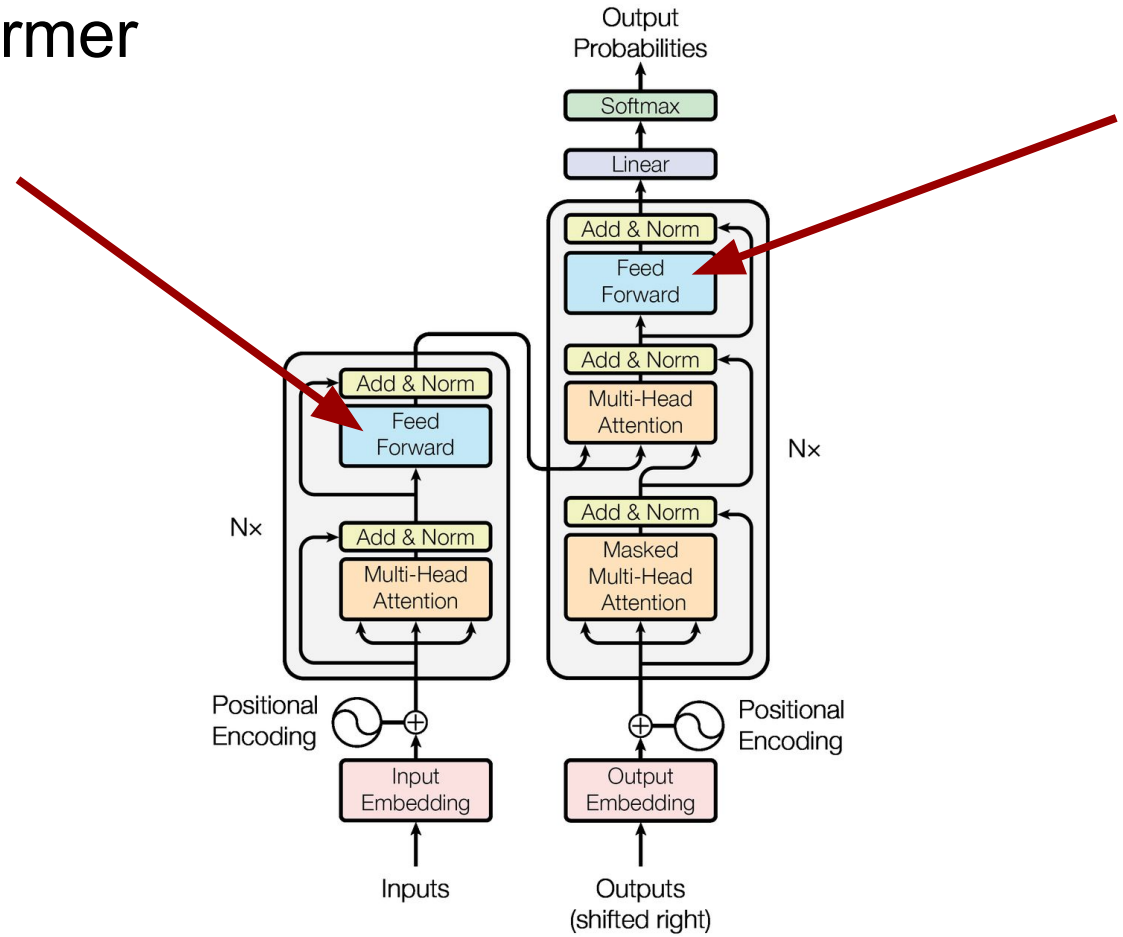
Transformer Feed-Forward Layers Are Key-Value Memories

Болотин Арсений
Еленик Константин
Малафеев Михаил
Семерова Елена

Transformer



Transformer



Параметры

Multi-Head Attention:

$$h = 8$$

$$d_{model} = 512$$

$$d_k = d_v = \frac{d_{model}}{h} = 64$$

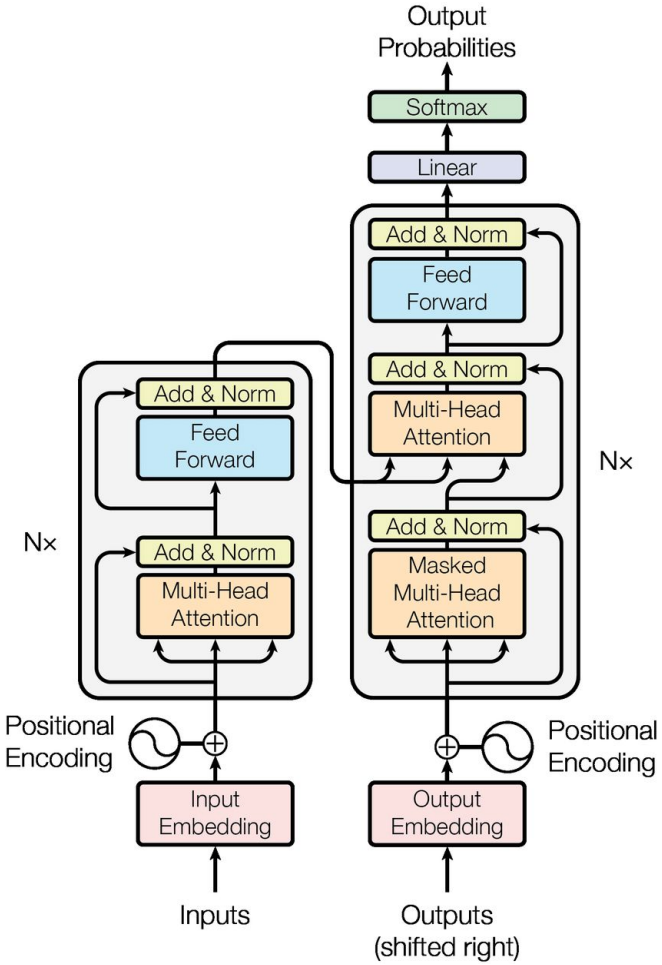
$$P : h \cdot 3 \cdot d_{model} \cdot d_k + d_{model} \cdot d_{model}$$

$$P : 4 \cdot d_{model} \cdot d_{model}$$

Feed Forward:

$$d_{ff} = 2048$$

$$P : 2 \cdot d_{model} \cdot d_{ff}$$



Feed-forward as neural memory

Feed-forward layer

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

$K, V \in \mathcal{R}^{d_m \times d}$ - parameter matrices

f - non-linearity (ReLU)

Neural Memory

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

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

End-to-end memory networks. (2015)
Sukhbaatar, S., Weston, J., & Fergus, R.

Feed-forward as neural memory

Feed-forward layer

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

$K, V \in \mathcal{R}^{d_m \times d}$ - parameter matrices

f - non-linearity (ReLU)

Neural Memory

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

End-to-end memory networks. (2015)
Sukhbaatar, S., Weston, J., & Fergus, R.

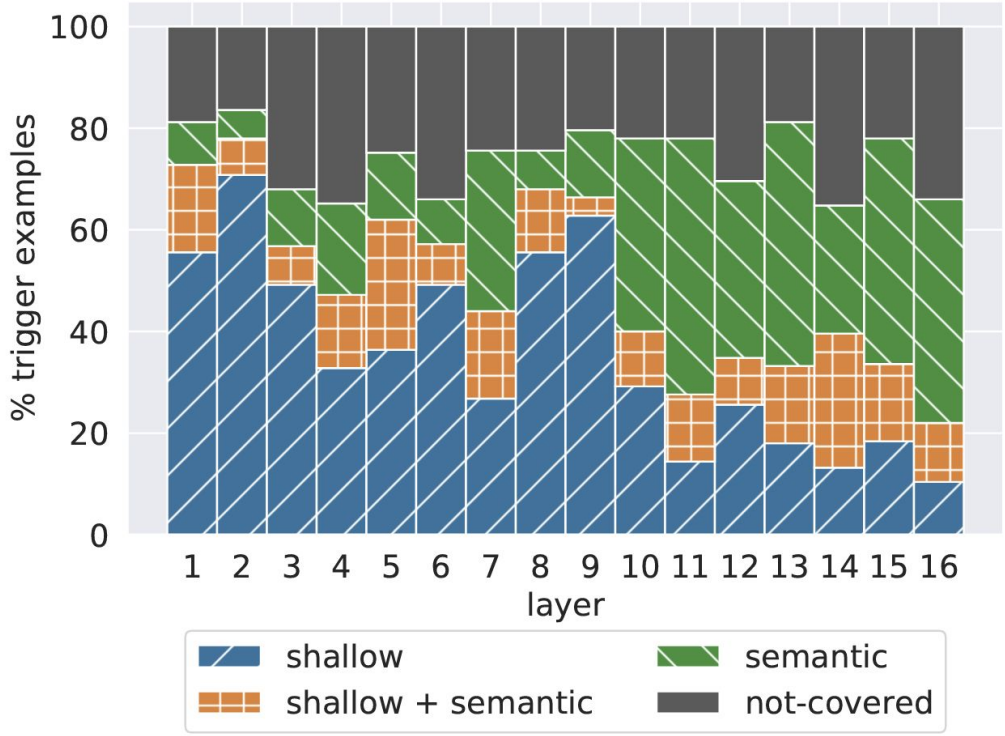
Trigger examples

Вычисляем для каждого префикса каждого предложения $m_i^\ell = ReLU(x_j^\ell \cdot k_i^\ell)$

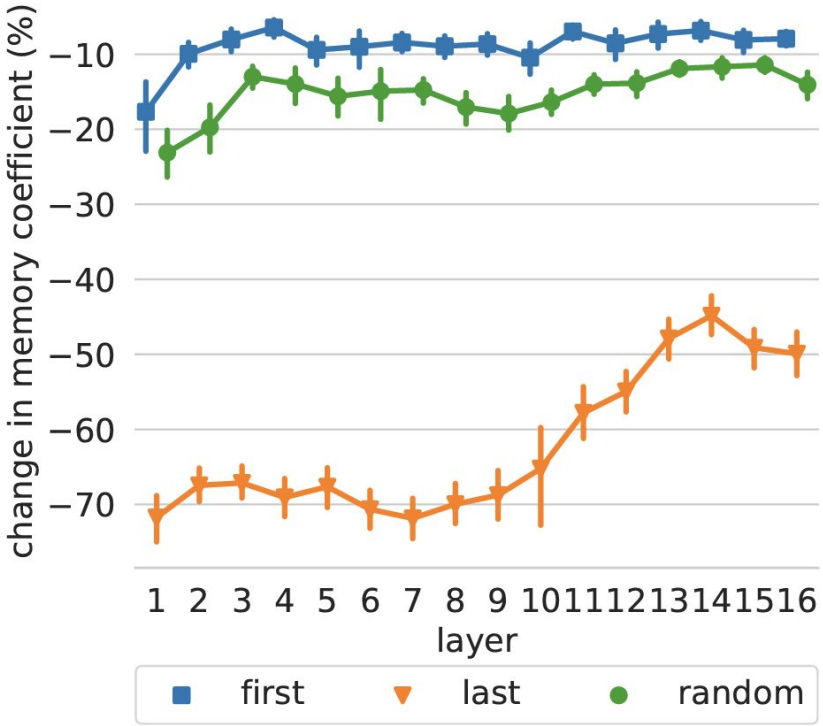
x_j^ℓ - вектор соответствующий последнему токену префикса

Key	Pattern	Example trigger prefixes
$\mathbf{k}_{449}^{1}$	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*
$\mathbf{k}_{2546}^{6}$	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*
$\mathbf{k}_{2997}^{10}$	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*
$\mathbf{k}_{2989}^{13}$	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*
$\mathbf{k}_{1935}^{16}$	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*

Trigger examples: types



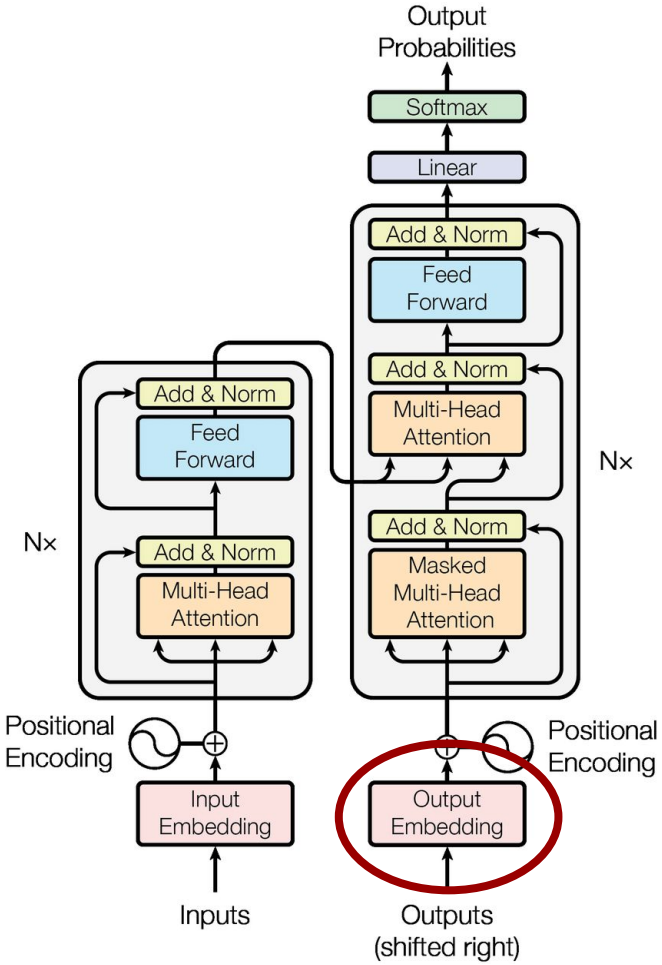
Trigger examples: local modifications



Value as a distribution

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

E - output embedding matrix

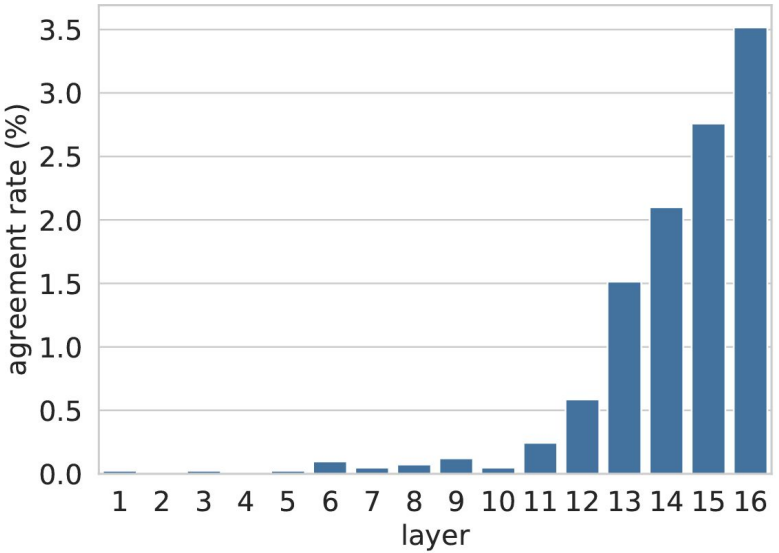


Key-value agreement

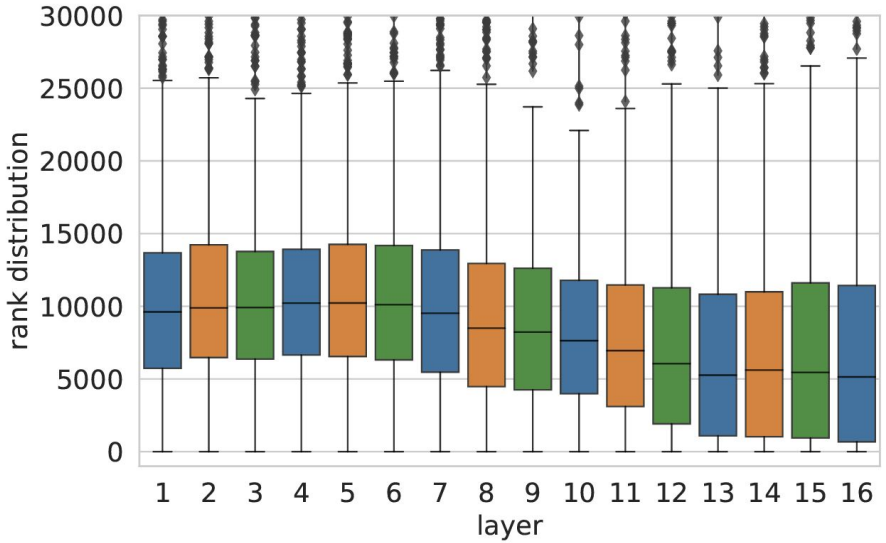
w_i^ℓ - следующий токен в примере, который соответствует наибольшему m_i^ℓ

Key-value agreement:

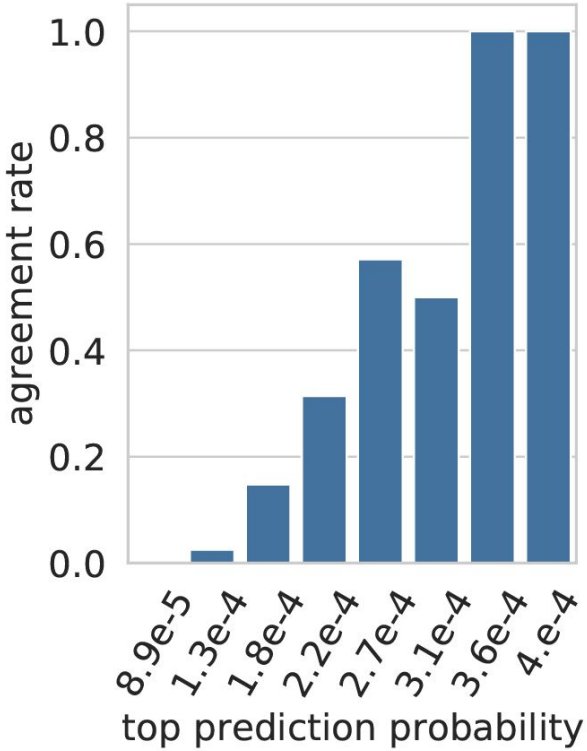
$$argmax(p_i^\ell) = w_i^\ell$$



Key-value agreement

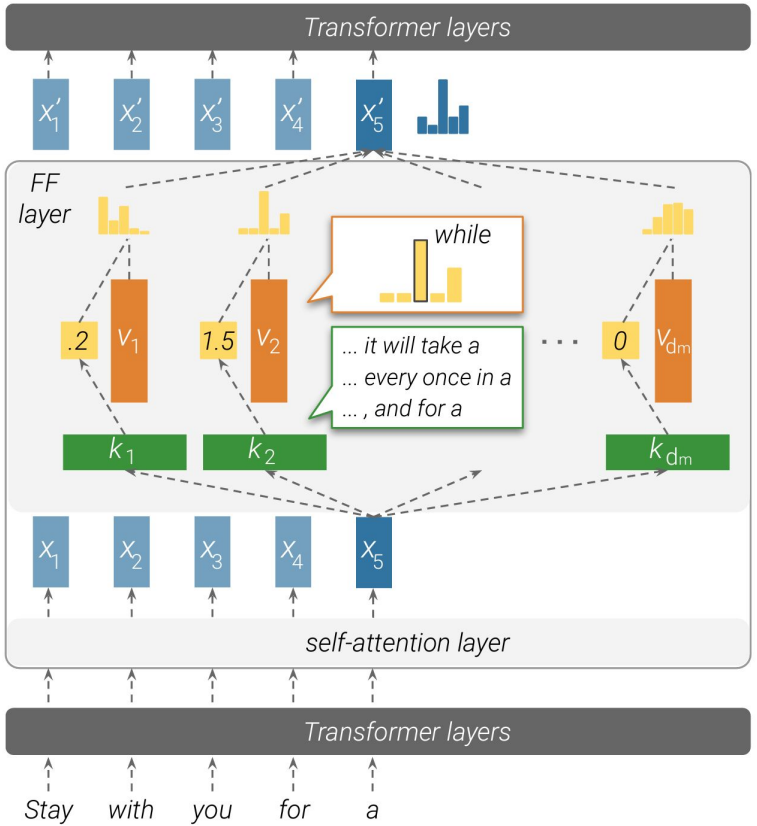


$$\text{rank } w_i^\ell \text{ in } p_i^\ell$$

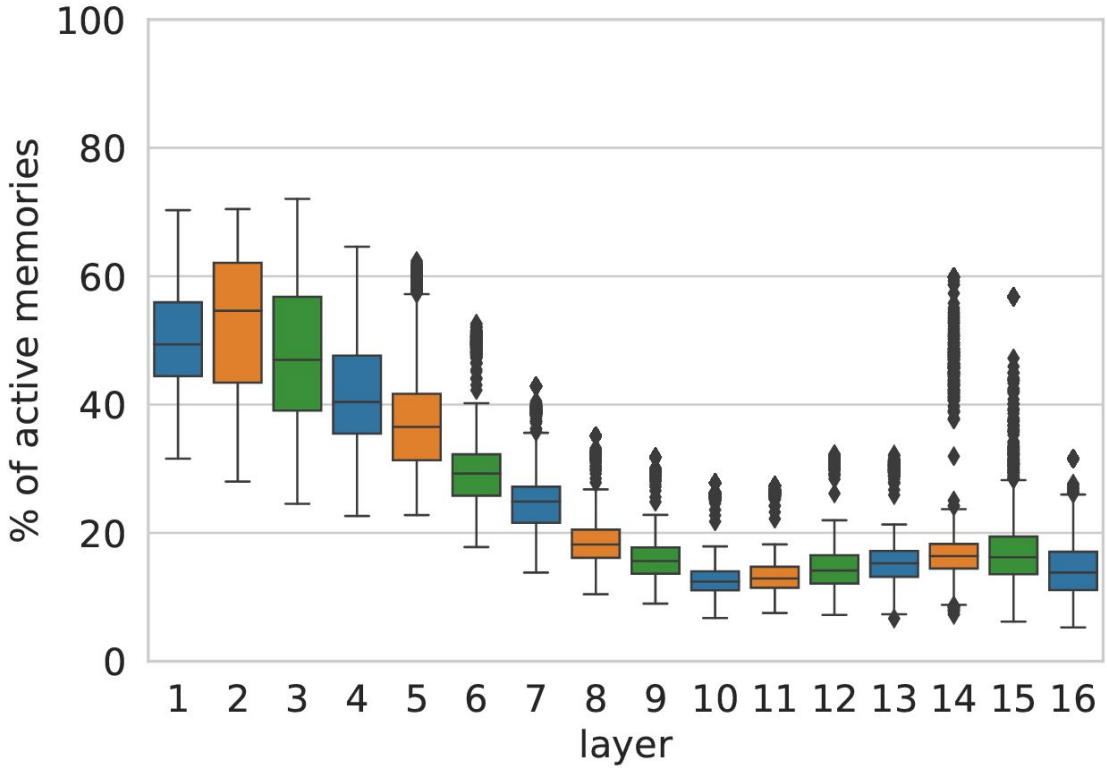


Aggregating Memories

$$\mathbf{y}^{\ell} = \sum_{i} \mathrm{ReLU}(\mathbf{x}^{\ell} \cdot \mathbf{k}_i^{\ell}) \cdot \mathbf{v}_i^{\ell} + \mathbf{b}^{\ell}$$



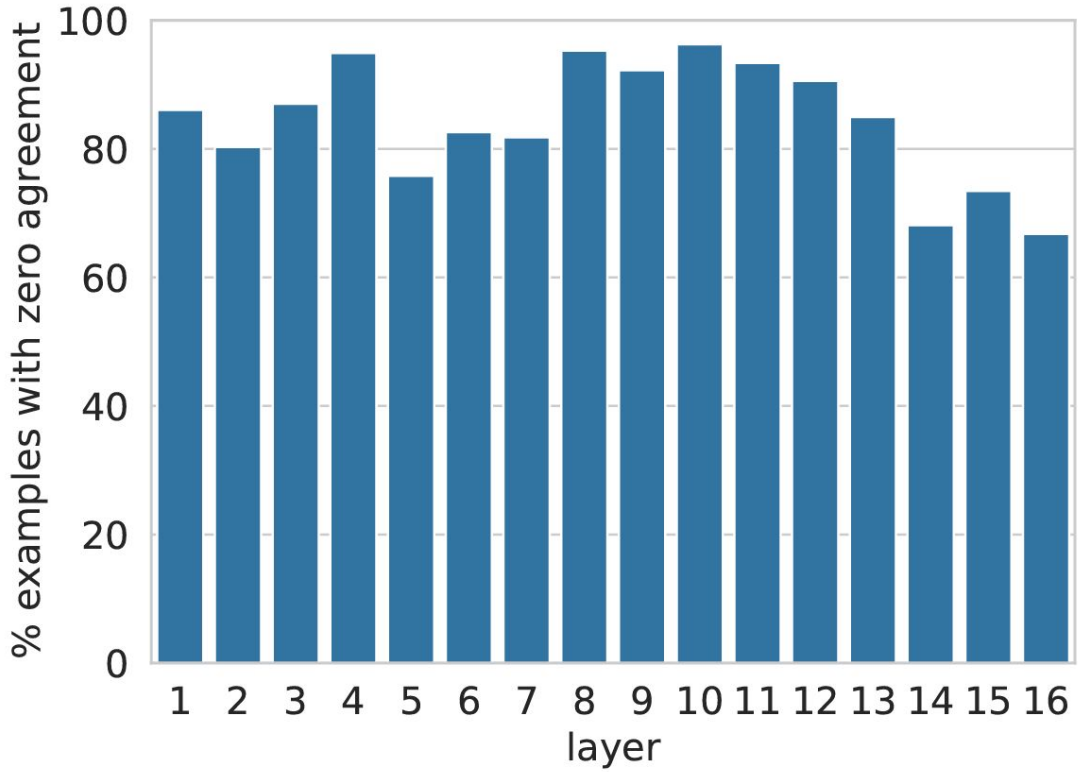
Aggregating Memories: non-zero activations



Aggregating Memories: zero agreement

$$\text{top}(\mathbf{h}) = \text{argmax}(\mathbf{h} \cdot E)$$

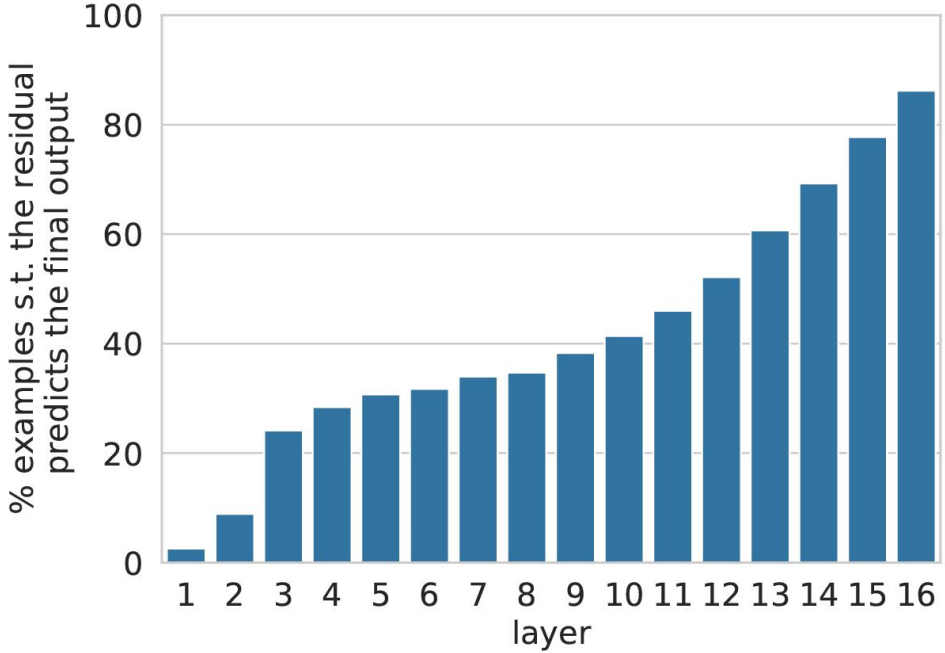
$$\forall i : \text{top}(\mathbf{v}_i^\ell) \neq \text{top}(\mathbf{y}^\ell)$$



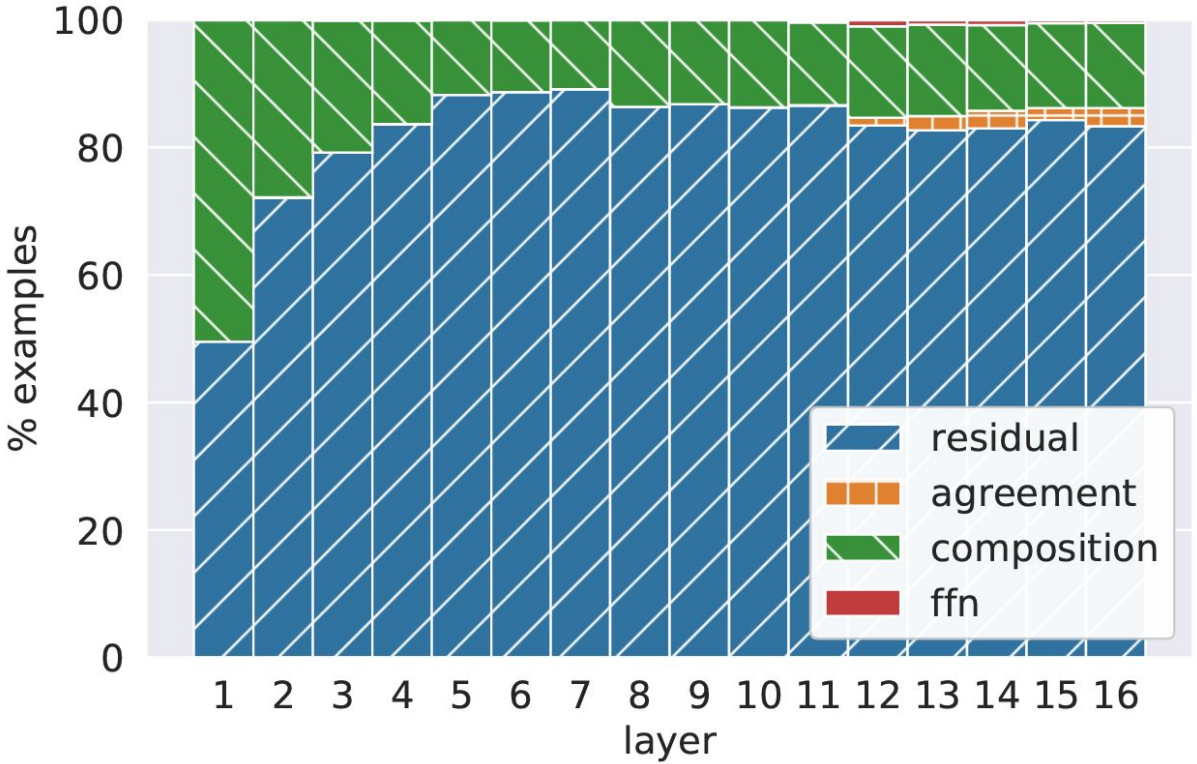
Aggregating Memories: residual connections

$$\mathbf{x}^\ell = \text{LayerNorm}(\mathbf{r}^\ell)$$
$$\mathbf{y}^\ell = \text{FF}(\mathbf{x}^\ell)$$
$$\mathbf{o}^\ell = \mathbf{y}^\ell + \mathbf{r}^\ell$$

$$\text{top}(\mathbf{r}^\ell) = \text{top}(\mathbf{o}^L)$$



Aggregating Memories: elimination mechanism



- Интерпретируемость FF в трансформерах.
- Keys - связаны с понятными паттернами. В первых слоях - элементарные паттерны, в более глубоких - семантические.
- Value - распределение над пространством output embedding.
- Следующий токен с наибольшей активацией в Key коррелирует с argmax токеном из распределения, индуцированного Value.
- Результат всей модели получается комплексно из распределений.
- FF слои уточняют результат, накладывая “вето” на какой-то вариант.

Рецензент

TL;DR

В статье изучают feed-forward слои языковых трансформеров. Авторы показывают, что эти слои ведут себя как key-value memories, где каждый ключ отвечает за определенный паттерн во входных данных, а значения за распределения над словарем. Эксперименты показывают, что усвоенные паттерны могут быть интерпретированы человеком.

Теоретическая обоснованность	+
Эксперименты	++
Новизна	+
Актуальность и значимость	+-

Воспроизводимость	+
Доходчивость	+
Оценка	8
Уверенность	3

Практик-исследователь

Авторы

- Mor Geva – Tel Aviv University(Ph.D), Allen Institute for AI
- Roei Schuster – Tel Aviv University(Ph.D), Vector Institute for AI
- Jonathan Berant – Tel Aviv University(Associate Professor, руководитель), Allen Institute for AI
    - Most cited: Semantic parsing on freebase from question-answer pairs
- Omer Levy – Tel Aviv University, Meta AI
    - Most cited: Roberta: A robustly optimized BERT pretraining approach

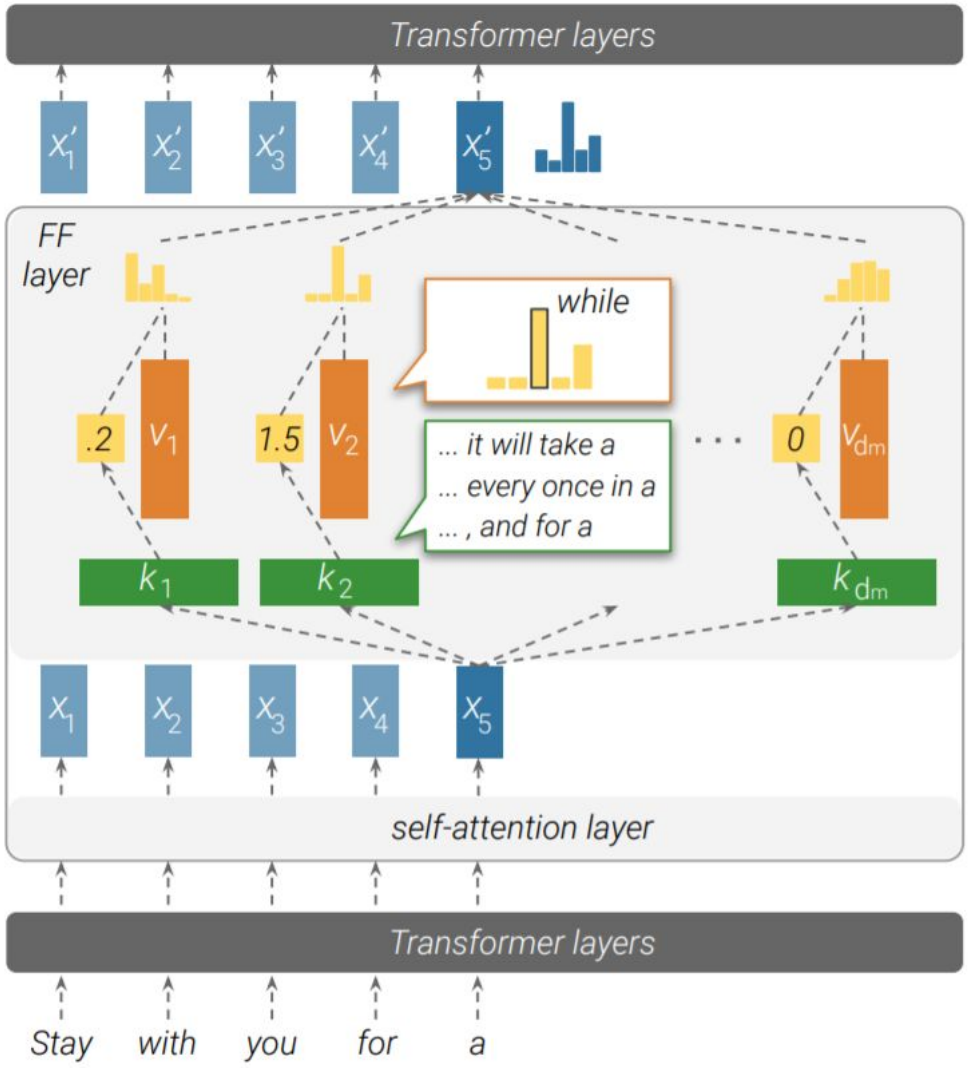
Контекст

Конференция:

- Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing

Опорные работы:

- End-To-End Memory Networks
- Augmenting Self-attention with Persistent Memory



Дальнейшие возможные исследования:

- Обобщение на трансформеры не только в языковых моделях, но и вообще
- Изучение роста корреляции распределения между выходами и ключевыми признаками в feed-forward

Применение и практическое знание:

- Возможность понимания решений модели с точки зрения человека
- Сохранение приватности данных в ходе обучения