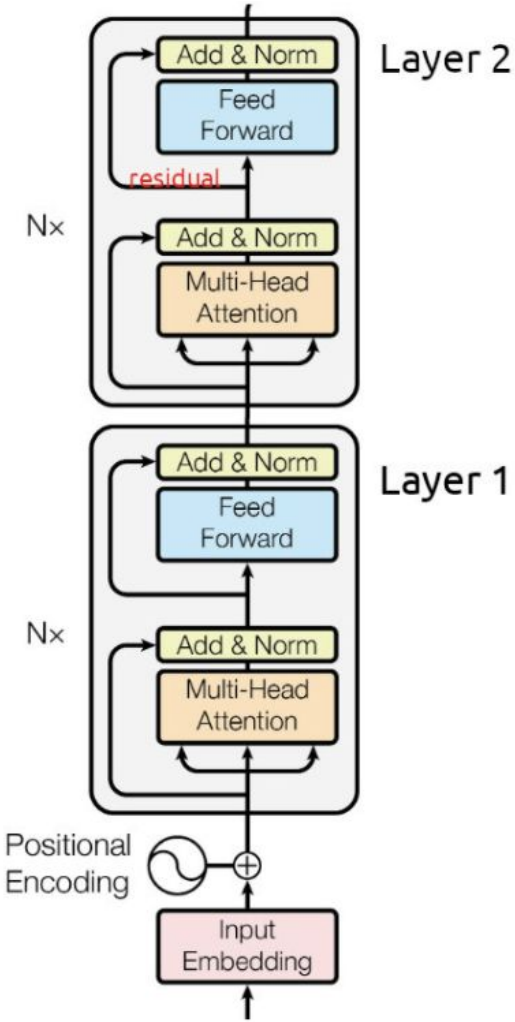
Transformer Feed-Forward Layers Are Key-Value Memories

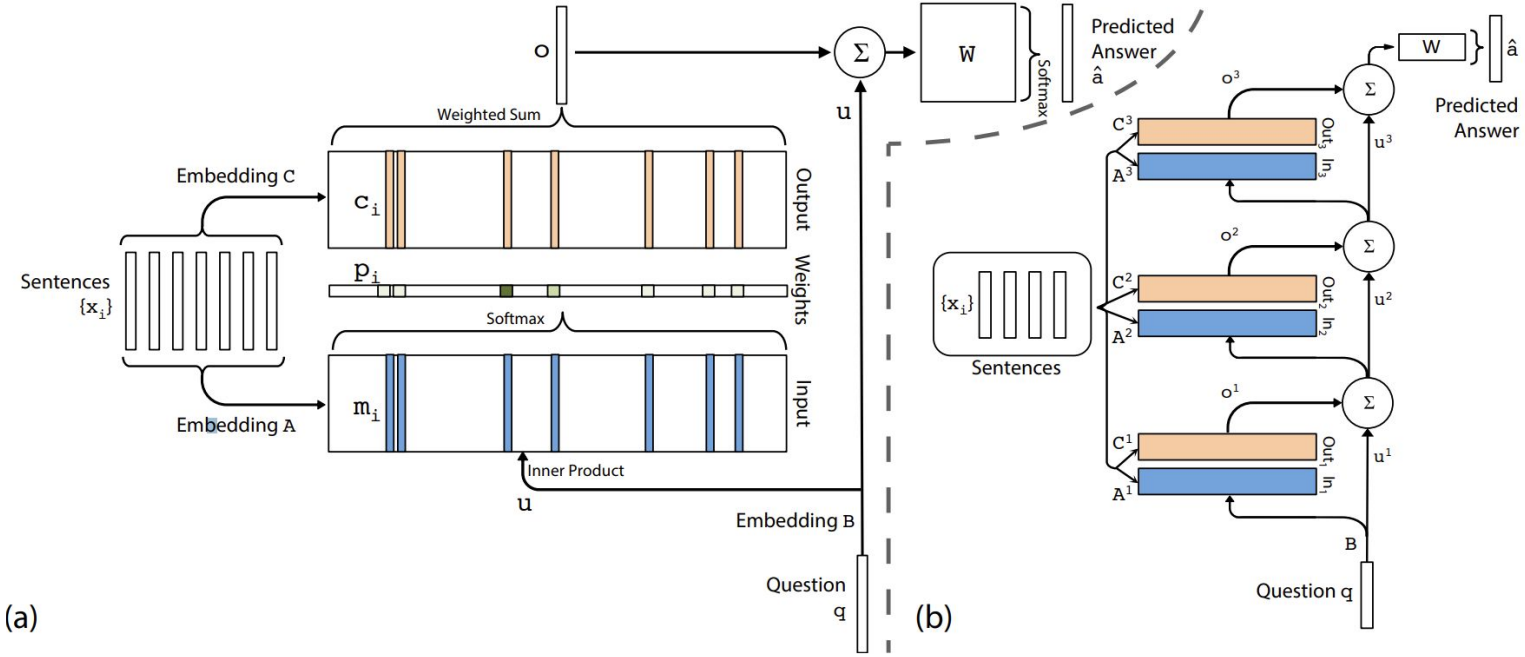
Андреев Никита
Денисенко Наталья
Щербинин Артём
Конодюк Никита

Motivation

- FF слои составляют 2/3 параметров трансформера
- Однако про FF слои довольно мало исследований
- Какова же их функция в трансформерах?

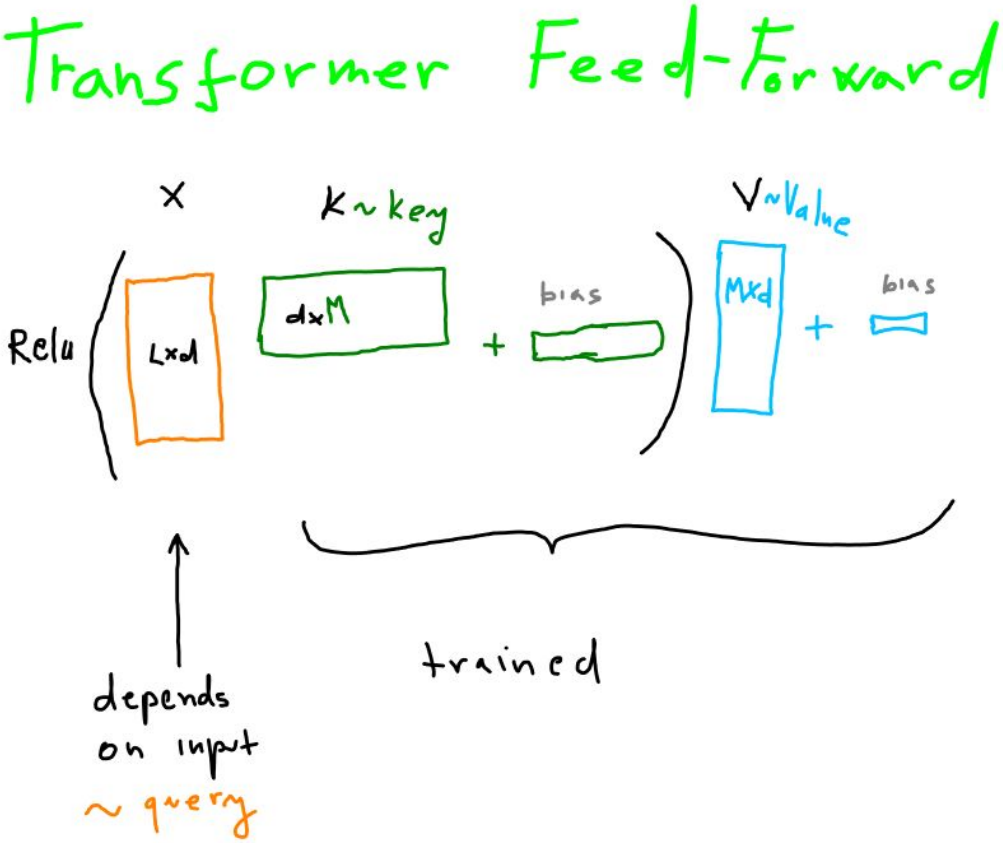


Key-Value Memory



$$p_i = \text{Softmax}(u^T m_i) \qquad o = \sum_i p_i c_i$$

Transformer Feed Forward layer



Almost identical

Feed Forward:

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

Neural Memory:

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

FF слой может вести себя похоже на Key-Value Neural Memory

Но что же он хранят в памяти?

Keys Capture Input Patterns

- Посчитали Memory coefficients для префиксов предложений

$$\mathrm{ReLU}(\mathbf{x}_j^{\ell} \cdot \mathbf{k}_i^{\ell})$$

- Для предложения “I love dogs” префиксами являются “I”, “I love” и “I love dogs”

- Затем взяли top-t префиксов, для которых репрезентации дали максимальное произведение с ключом, и выделили паттерны

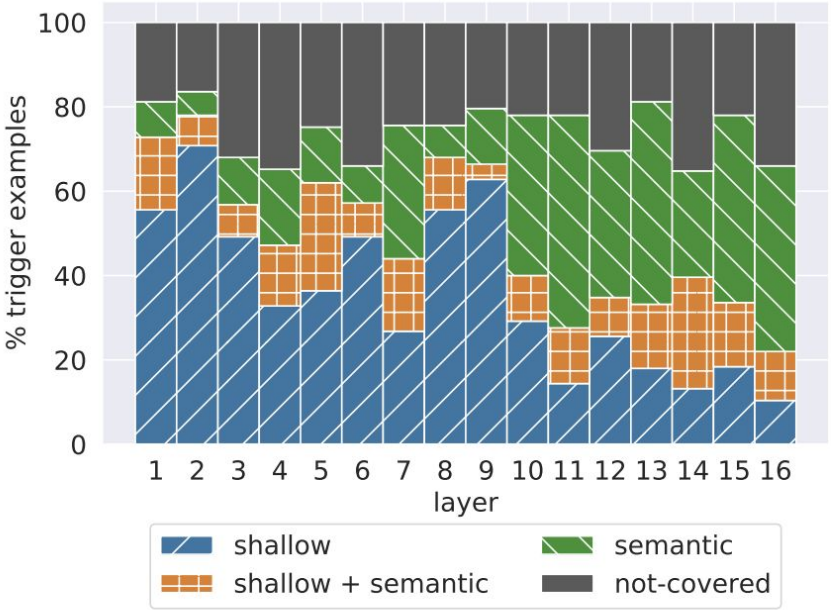
Keys Capture Input Patterns

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*

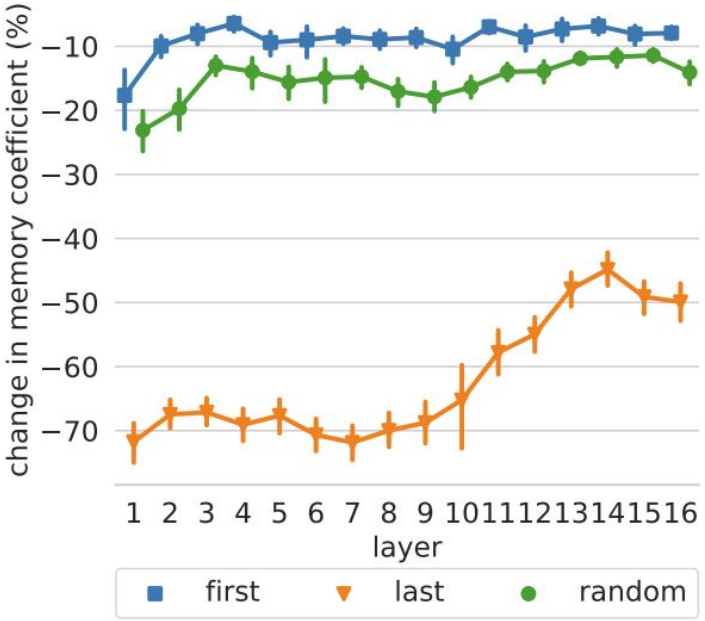
- В среднем 3.6 паттернов для одного ключа
- Большинство (65%-80%) префиксов подходят под хотя бы один паттерн

Keys Capture Input Patterns

Типы паттернов:



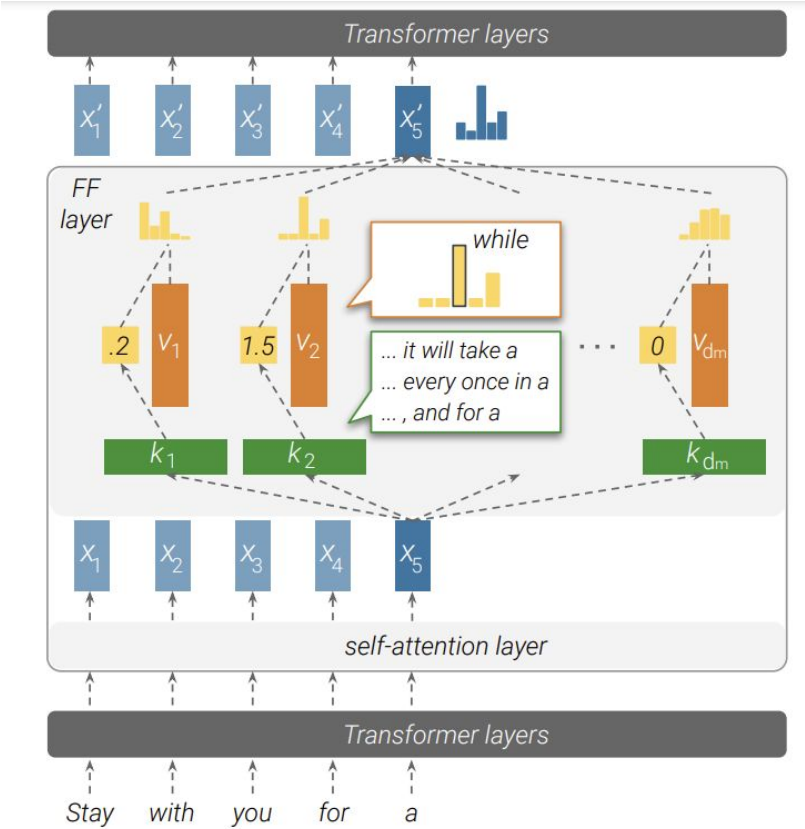
Изменения в коэффициентах если убрать токен из входа



Values Represent Distributions

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

Считаем вероятности через матрицу для выходных эмбеддингов и софтмакс



Values Represent Distributions

Для ключа смотрим на следующий токен для топового префикса

В векторе значений выбираем токен с наибольшей вероятностью

Для сравнения: agreement rate для случайного токена 0.0004%

Также если посмотреть на вероятность следующего токена в векторе вероятностей, то она возрастает на более поздних слоях

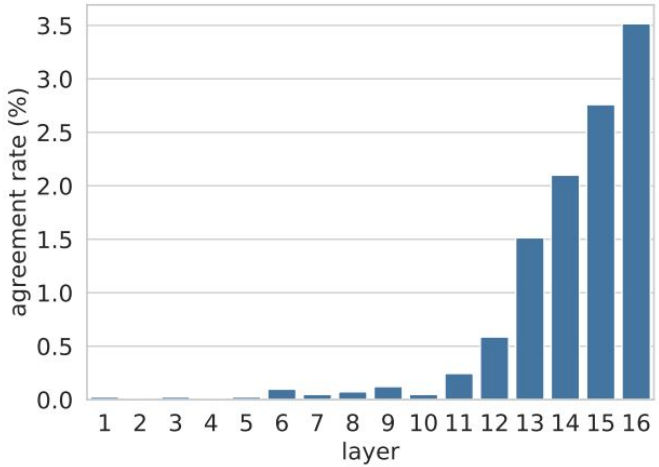
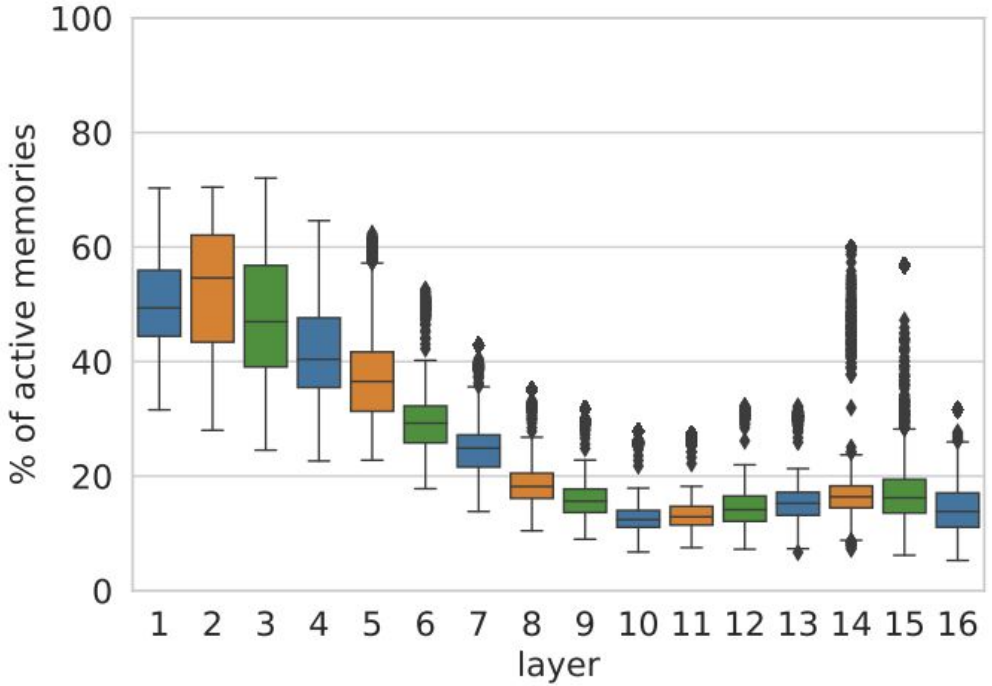


Figure 4: Agreement rate between the top-ranked token based on the value vector $\mathbf{v}_i^\ell$, and the next token of the top-ranked trigger example associated with the key vector $\mathbf{k}_i^\ell$.

Aggregating Memories

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

После 9 слоя наблюдается заметное снижение количества положительных memory coefficients (тогда же когда появляется больше семантических паттернов)

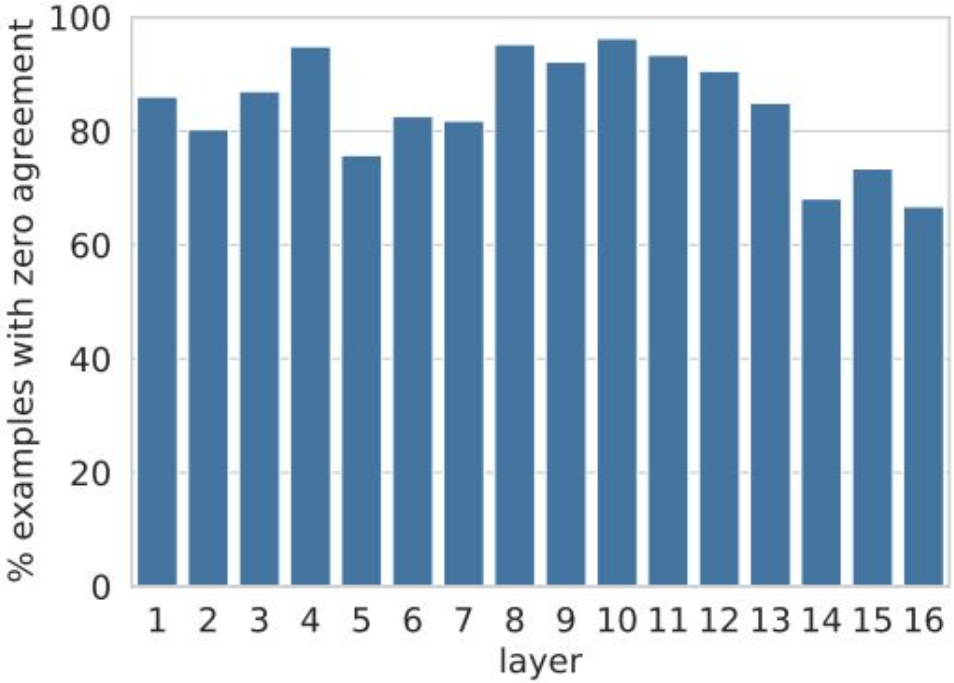


Aggregating Memories

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

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

В случаях совпадения очень часты вспомогательные слова по типу “the”, “of”

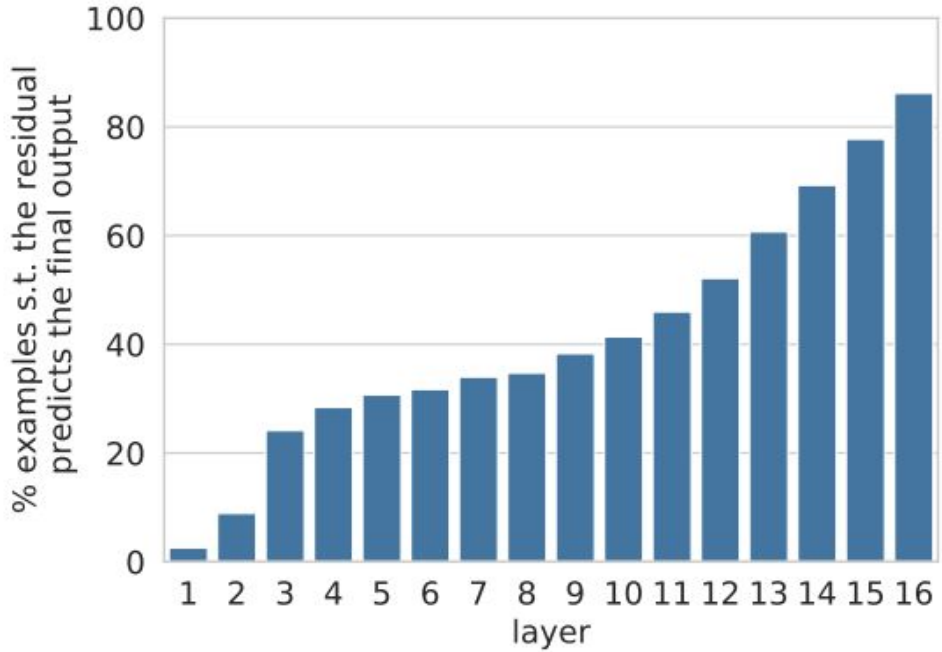


Inter-Layer Prediction Refinement

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

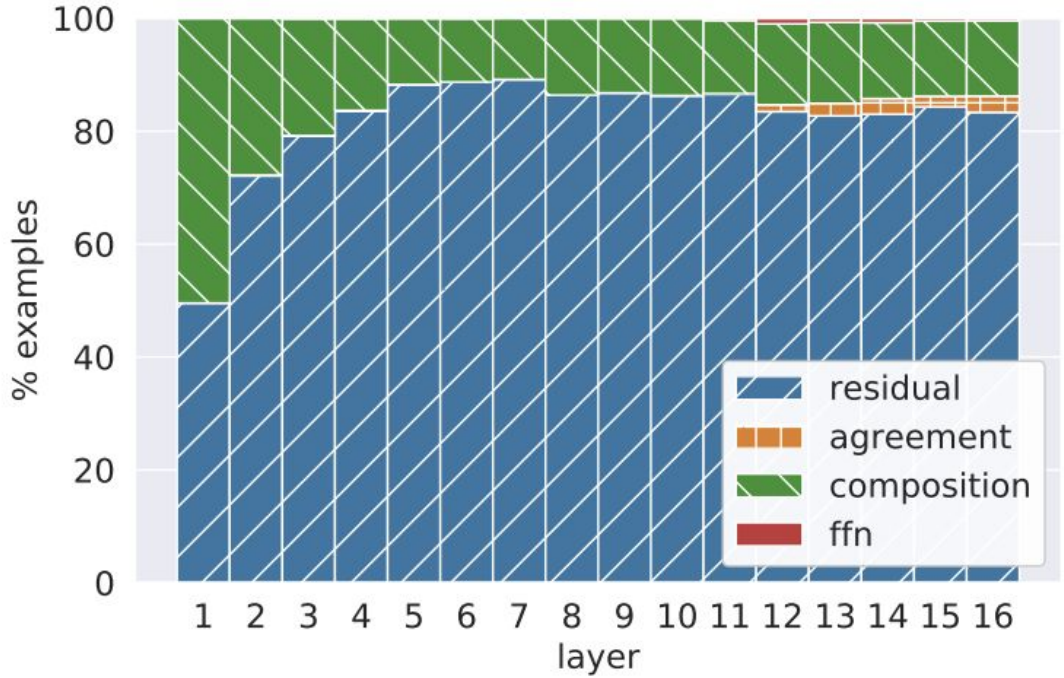
Когда топ предсказание residual-а совпадает с итоговым выходом модели

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



Inter-Layer Prediction Refinement

Случаи совпадения предсказания на выходе слоя с предсказаниями разных типов



Выводы

- Ключи (keys) отвечают за паттерны во входных данных (причем паттерны есть двух типов, которые преобладают в зависимости от слоя модели)

- Значения (values) отражают распределения, которые соотносятся с соответствующими ключами

- Feed Forward слой “проверяет” ответы, полученные из residual предсказаний

Рецензент

Денисенко Наталья

Достоинства и недостатки

- исследован менее изученный участок трансформеров
- было проведено множество экспериментов
- довольно прост в воспроизведении так как присутствует репозиторий на гитхабе

Стиль написания статьи

- написано опрятно, все формулы объяснены, а графики подписаны
- авторы описали все исследуемые модули и используемые свойства
- эксперименты внятно описаны
- авторы сами описали возможные направления работы

- небольшая придирка к тому, что релевантные работы оказались в самом конце статьи

Практик-исследователь

О статье

1) Первая версия статьи - 29 декабря 2020 года

2) Вторая (последняя) версия - 5 сентября 2021 года

3) Была принята на EMNLP2021

4) Была показана 8 ноября 2021 года в формате Virtual Poster

О людях

Mor Geva

Roei Schuster

Jonathan Berant

Omer Levy

Mor Geva

1) Computer Science Ph.D. candidate at Tel Aviv University
2) Researcher at the Allen Institute for AI
3) Jonathan Berant - advisor
4) Had interned at AI2, Google AI and Microsoft Media AI.
5) Work on problems in Natural Language Processing and Machine Learning
6) Wrote 7 articles before that
7) Wrote 4 articles after that, co-authored with Jonathan Berant:
    1. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies
    2. What's in your Head? Emergent Behaviour in Multi-Task Transformer Models
    3. Break, Perturb, Build: Automatic Perturbation of Reasoning Paths through Question Decomposition
    4. SCROLLS: Standardized CompaRison Over Long Language Sequences

Jonathan Berant

1) Completed a PhD in computer science at Tel Aviv University in 2012
2) An associate professor at the Blavatnik School of Computer Science, and a Research Scientist and The Allen Institute for Artificial Intelligence.
3) Field of research is Natural Language Processing
4) Wrote a lot of articles before that
5) Wrote 6 articles after that:

1. Scene graph to image generation with contextualized object layout refinement
2. Span-based semantic parsing for compositional generalization.
3. Latent compositional representations improve systematic generalization in grounded question answering
4. Scaling laws under the microscope: predicting transformer performance from small scale experiments
5. Unobserved local structures make compositional generalization hard.
6. SCROLLS: standardized comparison over long language sequences.

Roei Schuster

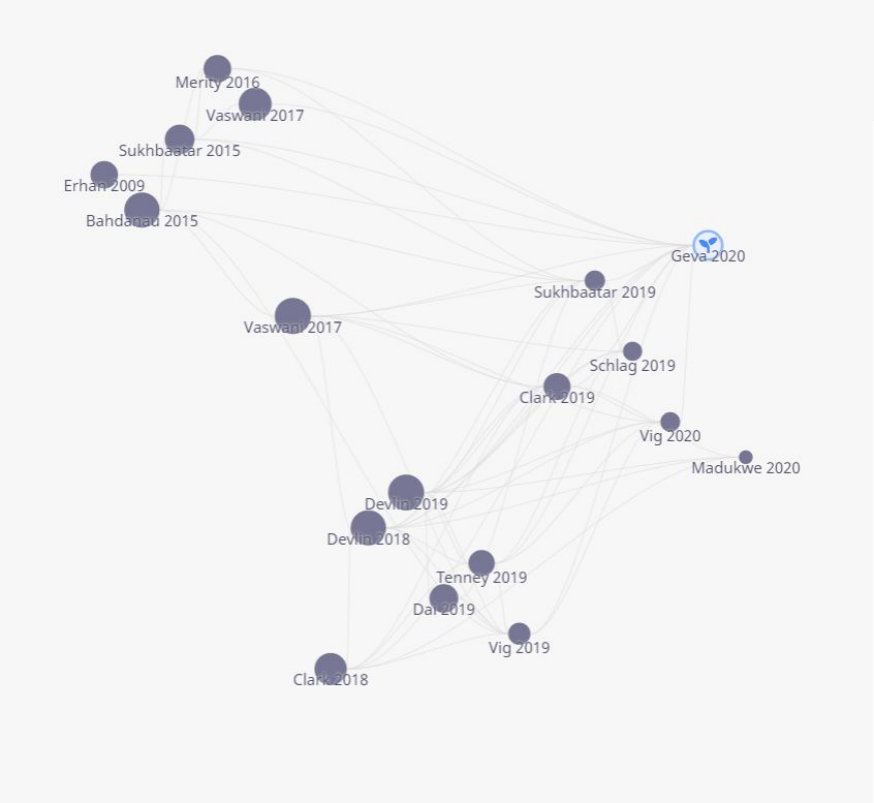
1) Completed a PhD in computer science at Tel Aviv University
2) A Postdoctoral Fellow at the Vector Institute for AI
3) Prof. Nicolas Papernot- advisor
4) Was also a researcher at Cornell Tech
5) Interested in the broad intersection of information security and machine learning
6) Wrote 8 articles before that
7) Wrote 3 articles after that:

1. **Lend Me Your Ear: Passive Remote Physical Side Channels on PCs**
2. **Squint Hard Enough: Evaluating Perceptual Hashing with Machine Learning**
3. **When the Curious Abandon Honesty: Federated Learning is Not Private**

Omer Levy

1) Completed a PhD at Bar-Ilan University, and did postdoctoral research at the University of Washington.
2) A senior lecturer at Tel Aviv University’s school of computer science
3) A research scientist at Facebook AI Research
4) Research is in the intersection of natural language processing (NLP) and machine learning
5) Co-authored 14 papers after that
6) Was co-authors in articles - Roberta, GLUE

Опоры и цитирования



20 цитирований
17 опор:

Madukwe, 2020 | A GA-Based Approach to Fine-Tuning BERT for Hate Speech Detection

Schlag, 2019 | Enhancing the Transformer with explicit relational encoding for math problem solving

Vig, 2020 | Investigating Gender Bias in Language Models Using Causal Mediation Analysis

Sukhbaatar, 2019 | Augmenting Self-attention with Persistent Memory

Vig, 2019 | Analyzing the Structure of Attention in a Transformer Language Model

Tenney, 2019 | BERT Rediscovers the Classical NLP Pipeline.

Clark, 2019 | What Does BERT Look at? An Analysis of BERT's Attention

Erhan, 2009 | Visualizing Higher-Layer Features of a Deep Network

Merity, 2016 | Pointer Sentinel Mixture Models

Dai, 2019 | Transformer-XL: Attentive Language Models beyond a Fixed-Length Context