Transformer Feed-Forward Layers Are Key-Value Memories

Mor Geva, Roei Schuster, Jonathan Berant, Omer Levy

1) Первая версия статьи 29 декабря 2020 года. Вторая (последняя) - 5 сентября 2021 года. Была принята на EMNLP2021 и была показана 8 ноября 2021 года в формате Virtual Poster
2) Авторы статьи:

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:
    - Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies
    - What's in your Head? Emergent Behaviour in Multi-Task Transformer Models
    - Break, Perturb, Build: Automatic Perturbation of Reasoning Paths through Question Decomposition
    - 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:
    - Scene graph to image generation with contextualized object layout refinement
    - Span-based semantic parsing for compositional generalization.
    - Latent compositional representations improve systematic generalization in grounded question answering
    - Scaling laws under the microscope: predicting transformer performance from small scale experiments
    - Unobserved local structures make compositional generalization hard.
    - 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:

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

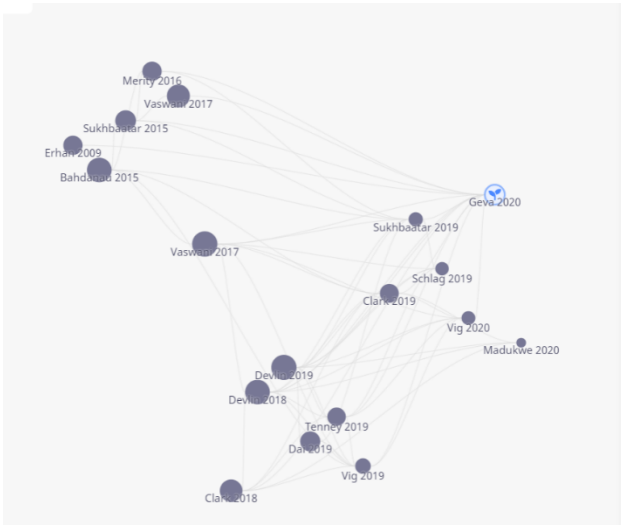
Omer Levy.

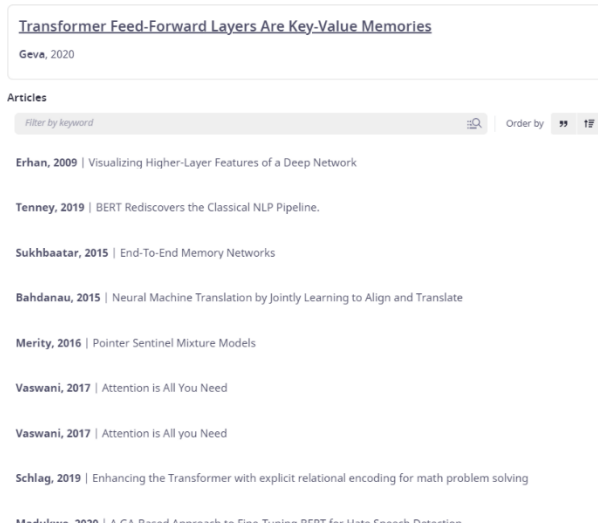
1. Completed a PhD at Bar-Ilan University, and did postdoctoral research at the University of Washington.
2. A senior lecturer at Tel Aviv University's school of computer science
3. A research scientist at Facebook AI Research
4. Research is in the intersection of natural language processing (NLP) and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

6. Was co-authors in articles - Roberta, GLUE

Работа отчасти является прямым продолжением одной из работ-опор от других авторов (While the theoretical similarity between feed-forward layers and key-value memories has previously been suggested by Sukhbaatar et al. (2019), we take this observation one step further, and analyze the "memories" that the feed-forward layers store). также следующие статьи не являются продолжениями от этих авторов.

3)





Knowledge Neurons in Pretrained Transformers *ArXiv* **2021**
Damai Dai, Li Dong, Y. Hao, Zhifang Sui, Furu Wei

4)

Locating and Editing Factual Knowledge in GPT **2022**
Kevin Meng, David Bau, A. Andonian, Y. Belinkov

Analyzing Commonsense Emergence in Few-shot Knowledge Models **2021**
Jeff Da, Ronan Le Bras, Ximing Lu, Yejin Choi, Antoine Bosselut

Attention Approximates Sparse Distributed Memory *ArXiv* **2021**
Trenton Bricken, C. Pehlevan

BERTnesia: Investigating the capture and forgetting of knowledge in BERT *BLACKBOXNLP* **2020**
Jonas Wallat, Jaspreet Singh, Avishek Anand

Consistency and Coherence from Points of Contextual Similarity *ArXiv* **2021**
Oleg V. Vasilyev, John Bohannon

Distilling Relation Embeddings from Pre-trained Language Models *EMNLP* **2021**
Asahi Ushio, José Camacho-Collados, S. Schockaert

5) Конкурентов у статьи нет
6) Завершили исследование слоёв в трансформере, так что воде больше нечего исследовать
7) Его нет