Исследование контекста статьи “**Transformer Feed-Forward Layers Are Key-Value Memories**”

Автор исследования: Малафеев Михаил

Основные сведения.

Статья выложена на arxiv 29.12.2020. Была обновлена 05.09.2021 с уточнениями для воспроизведения экспериментов статьи.

Авторы статьи:

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

Работа выполнялась в ходе получения Ph.D Mor Geva.

Цитирований: 17.
В основном связаны с прикладными изучениями в области интерпретируемости и дообучения в трансформерах.

Область интересов - NLP, интерпретируемость, обобщающая способность и дообучение, прикладные исследования.

Изучает особенность Feed-Forward слоев хранить воспоминания работу End-To-End Memory Networks и теоретическую гипотезу из статьи Augmenting Self-attention with Persistent Memory про то, что **Transformer Feed-Forward Layers Are Key-Value Memories.**

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