



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

In-context Learning and Induction Heads

Обзор-рецензия

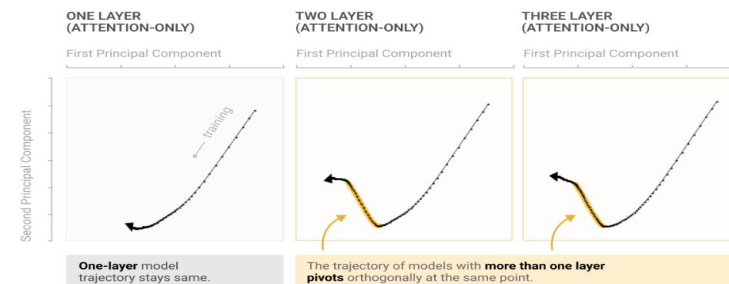
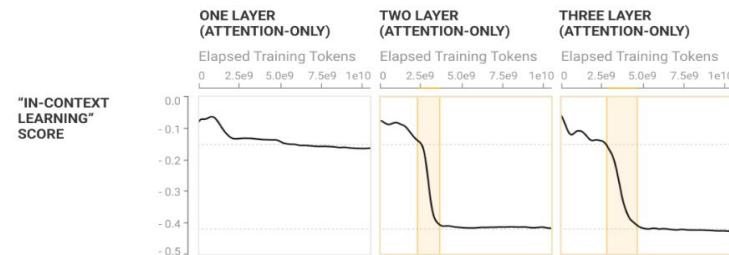
Докладчик:	Лишуди Дмитрий
Рецензент:	Воробьев Николай
Хакер:	Мошков Иван

Идея работы

Поиск взаимосвязей между:

- induction heads
- phase change
- in-context learning

Попытки обобщения на большие модели



Основные авторы



Catherine Olsson



Neel Nanda



Nelson Elhage

Организация

Anthropic: *Building Reliable, Interpretable, and Steerable AI Systems*

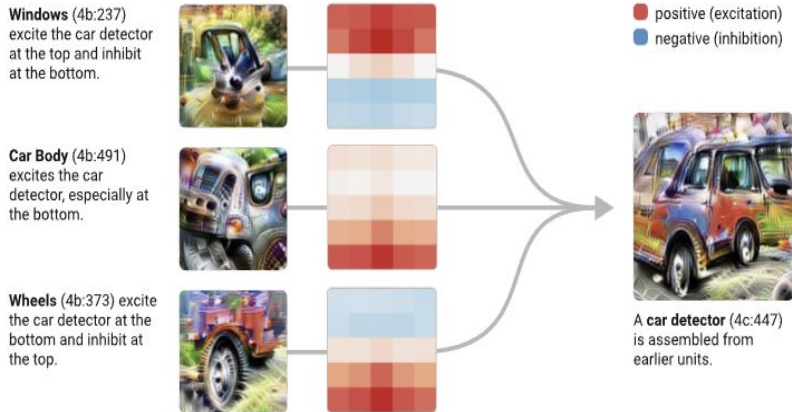


Смежные работы

Смежные работы

Thread: circuits [march 2020]

Интерпретация CV



Смежные работы

A Mathematical Framework for Transformer Circuits [dec 2021]

- Интерпретация трансформеров
- Много формул
- Маленький пул моделей

$$T = \text{Id} \otimes W_U \cdot \left(\text{Id} + \sum_{h \in H_2} A^h \otimes W_{OV}^h \right) \cdot \left(\text{Id} + \sum_{h \in H_1} A^h \otimes W_{OV}^h \right) \cdot \text{Id} \otimes W_E$$

The second attention layer has multiple attention heads which add into the residual stream

The first attention layer has multiple attention heads which add into the residual stream

$$= \text{Id} \otimes W_U W_E + \sum_{h \in H_1 \cup H_2} A^h \otimes (W_U W_{OV}^h W_E) + \sum_{h_2 \in H_2, h_1 \in H_1} (A^{h_2} A^{h_1}) \otimes (W_U W_{OV}^{h_2} W_{OV}^{h_1} W_E)$$

"Direct path" term contributes to bigram statistics.

The individual attention head terms describe the effects of individual attention heads in linking input tokens to logits, similar to those we saw in the one layer model.

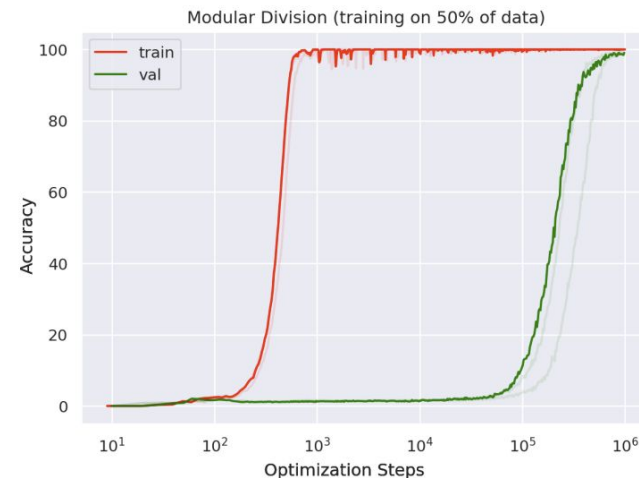
The virtual attention head terms correspond to V-composition of attention heads. They function a lot like individual attention heads, with their own attention patterns (the composition of the heads patterns) and own OV matrix.

Смежные работы

A Mechanistic Interpretability Analysis of Grokking [aug 2022]

Neel Nanda, Tom Lieberum

- Интуиция гроккинга
- Связь с фазовыми изменениями



Сильные стороны

- Последовательность работы
- Наличие визуализаций
- Весомые эксперименты
- Более широкий спектр моделей

Слабые стороны

- Эмпирические выводы
- Плавающие определения