НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

In-context Learning and Induction Heads

Обзор-рецензия

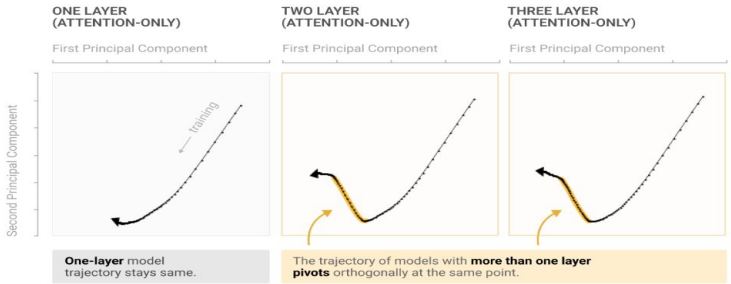
Докладчик: Лишуди Дмитрий
Рецензент: Воробьев Николай
Хакер: Мошков Иван

Москва, 2022

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

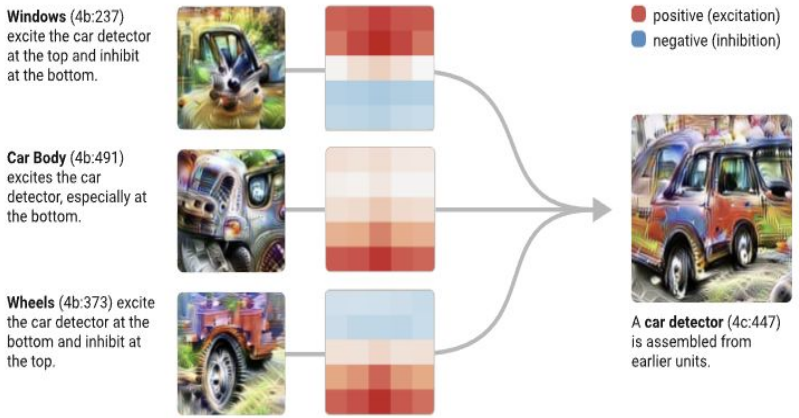
Anthropic: *Building Reliable, Interpretable, and Steerable AI Systems*

Смежные работы

Смежные работы

Thread: circuits [march 2020]

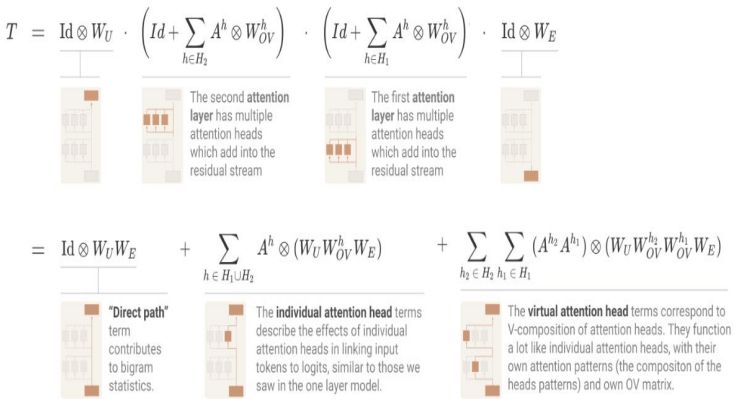
Интерпретация CV



Смежные работы

A Mathematical Framework for Transformer Circuits [dec 2021]

- Интерпретация трансформеров
- Много формул
- Маленький пул моделей

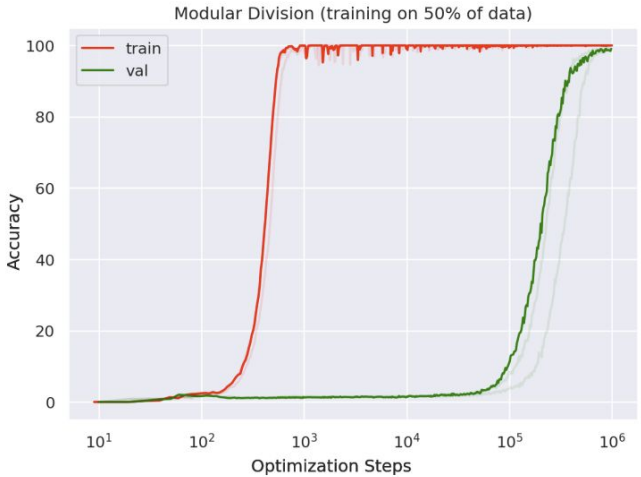


Смежные работы

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