GROKKING: Generalization beyond overfitting on small algorithmic datasets

Иван Фридман 192
14 декабря 2022

Решаемая задача и используемые модели

Задача - предсказание результата бинарного отношения

Примеры отношений:

$$x \circ y = x \cdot y \text{ for } x, y \in S_5$$

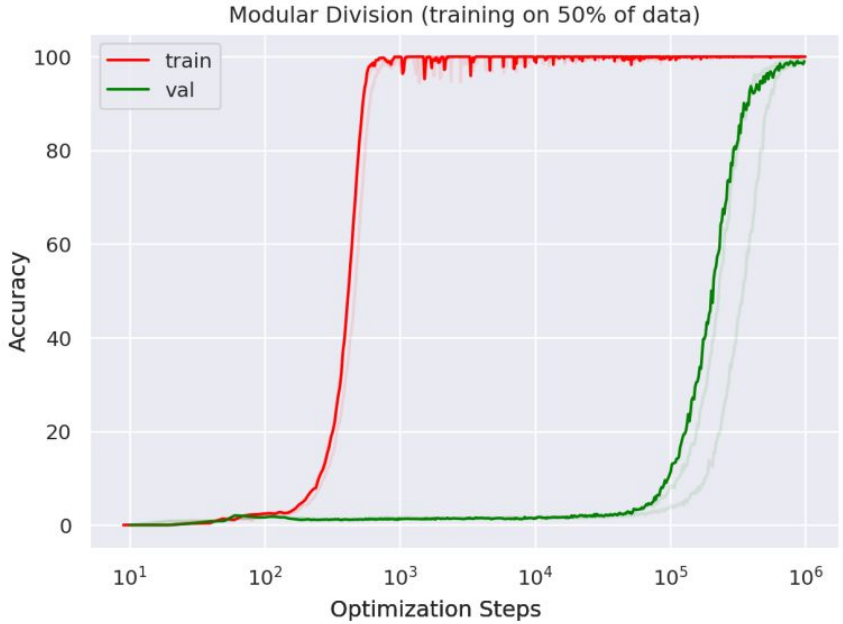
$$x \circ y = x + y \pmod{p} \text{ for } 0 \leq x, y < p$$

$$x \circ y = [x/y \pmod{p} \text{ if } y \text{ is odd, otherwise } x - y \pmod{p}]$$

Используемая архитектура - декодер-only трансформер с каузальным маскированным аттеншеном

Примерно 4 · 10 ^ 5 параметров в модели

Гроккинг



Гроккинг - явление, при котором качество работы на валидационном наборе данных улучшается значительно позже, чем происходит переобучение на тестовом наборе

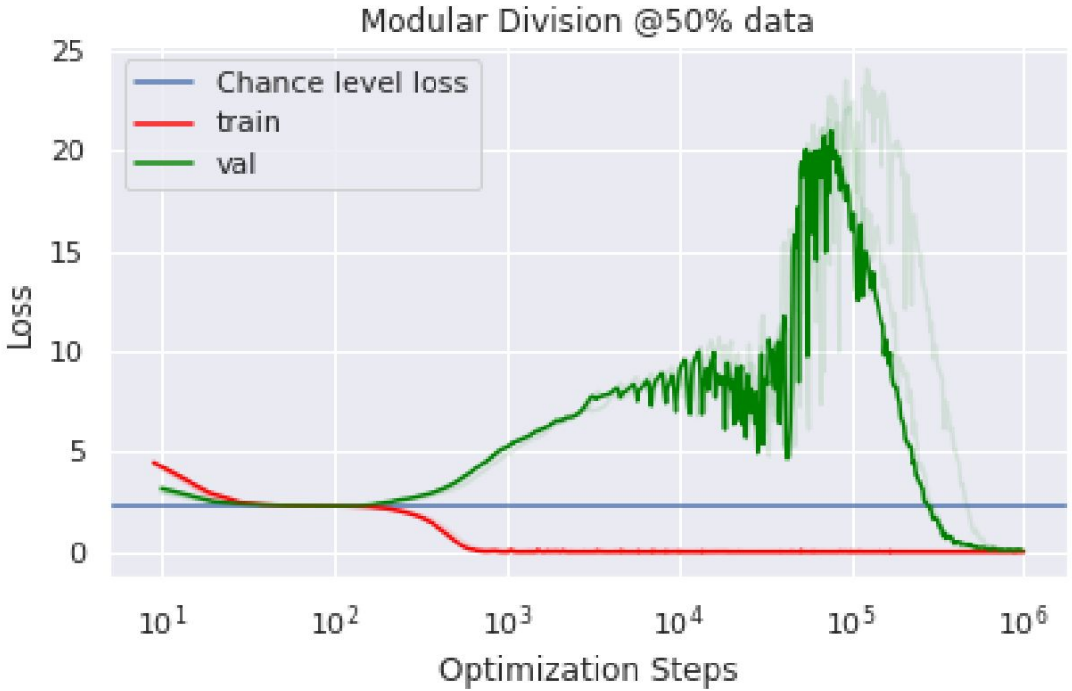
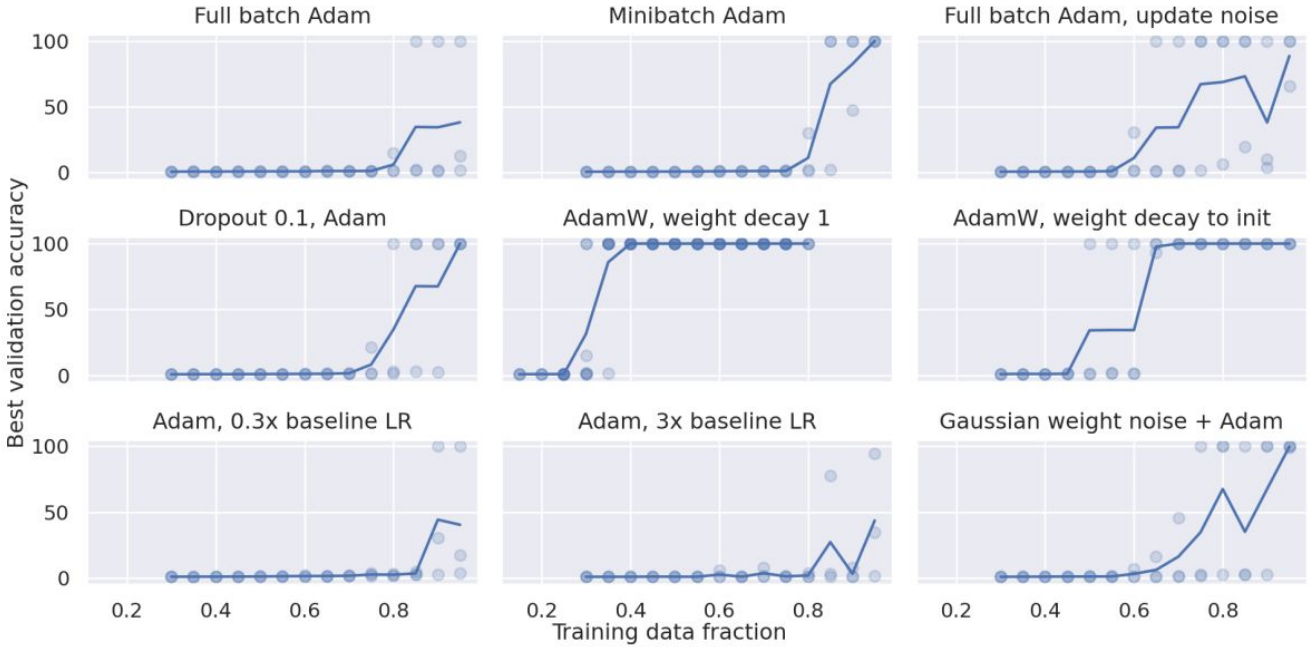
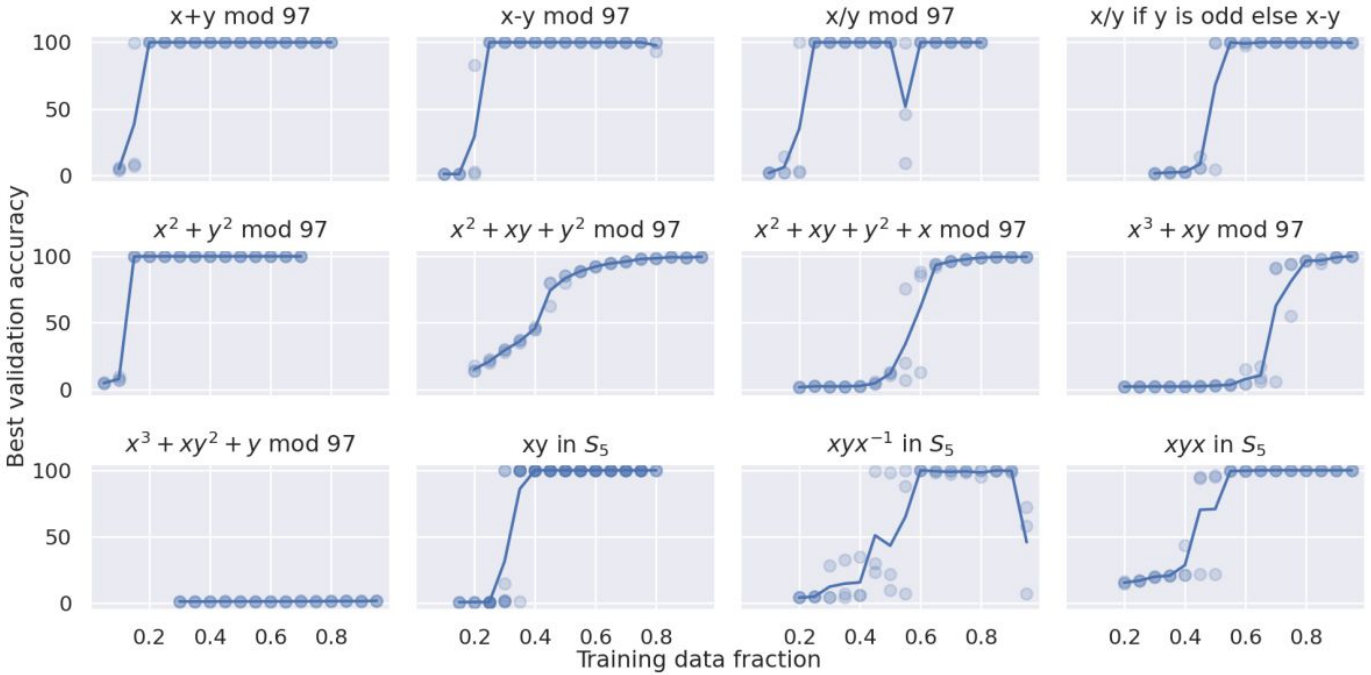


График ошибки для задачи деления по модулю

Уровни генерализации для разных методов оптимизации при разных размерах обучающей выборки за фиксированное количество операций



Уровни генерализации для разных бинарных отношений при разных размерах обучающей выборки за фиксированное количество операций



Основные наблюдения:

- Гроккинг происходит на большей части бинарных операций, но на некоторых более заметен, чем на других
- Гроккинг - типичное поведение модели, когда размер набора данных близок к минимальному, при котором модель в принципе генерализируется
- Чем больше набор данных, тем ближе друг к другу идут кривые ошибки и точности на обучающей выборке и на валидационной выборке
- Кроме гроккинга, при уменьшении размера набора данных увеличивается время, за которое модель генерализуется
- В окрестности 25-30% набора данных для обучения уменьшение его на 1% приводит к увеличению медианного времени до генерализации на 40-50%
- Для симметричных бинарных операций обучение происходит быстрее
- Метод затухания весов позволяет уменьшить необходимое количество данных в обучающем наборе данных примерно в 2 раза

Механистическая интерпретируемость

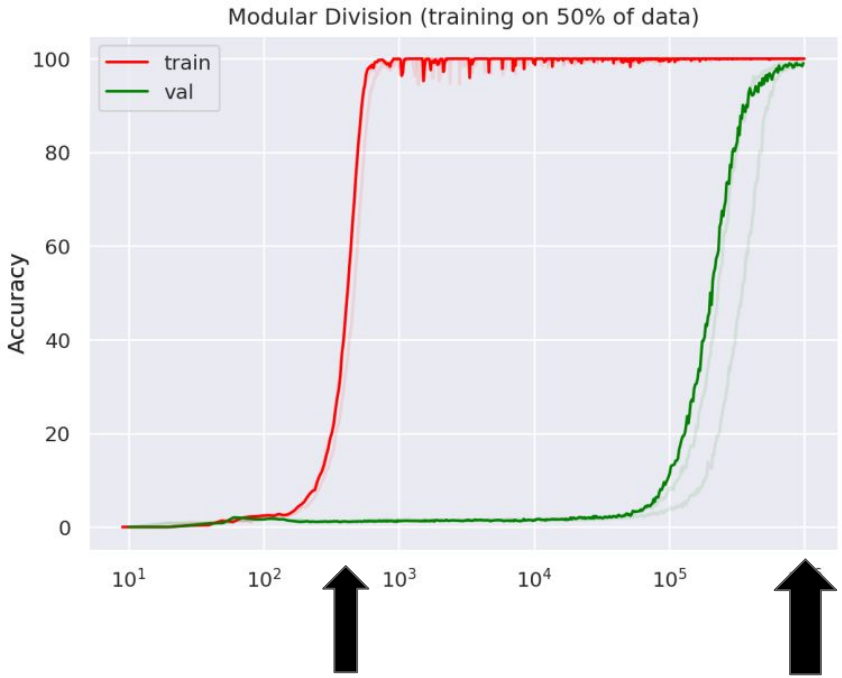
Фазовое изменение - резкое изменение качества работы модели во время обучения. Является общим явлением, которое происходит в том числе в больших моделях, обучаемых для production-задач

Механистическая интерпретируемость - теория, одно из главных утверждений которой заключается в том, что нейронные сети можно подвергнуть реверс-инжинирингу, чтобы интерпретировать алгоритм, которым она пользуется для решения задачи

Основные утверждения про гроккинг

- Гроккинг - это на самом деле фазовое изменение. Если мы выбираем регуляризацию и размер набора данных так, что регуляризованное решение лишь немного превосходит решение с запоминанием всех примеров, то происходит гроккинг
- Обобщающие схемы, используемые для модульного сложения, развиваются в ходе обучения именно тогда, когда происходит гроккинг, то есть когда модель переходит от решения с запоминанием всех примеров к генерализованному решению

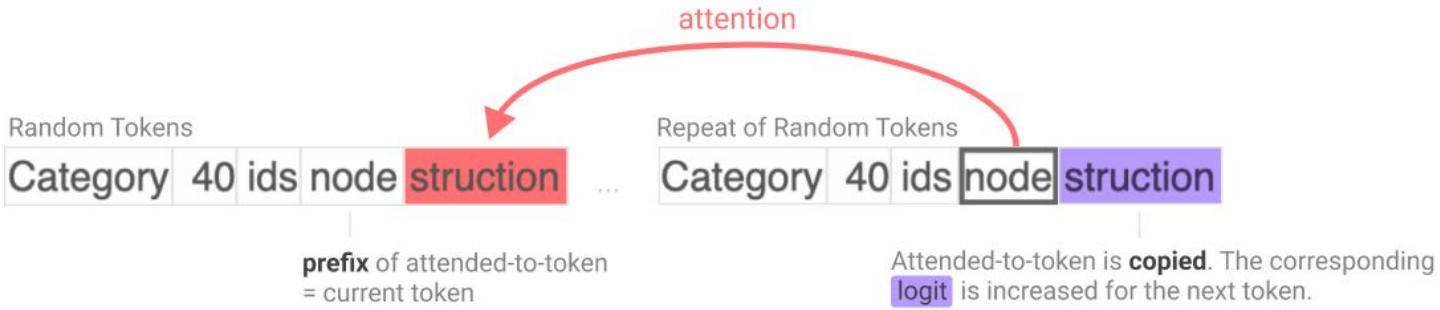
Интуитивное объяснение гроккинга



модель выучивает решение с запоминанием примеров

модель выучивает генерализующее решение

Индукционная схема в трансформерах



Возможные объяснения возникновения схем в нейронных сетях

- Объяснение лотерейного билета - каждый слой изначально состоит из большого количества разнообразных схем, некоторые из которых улучшают работу алгоритма, однако большая часть не улучшает. В процессе оптимизации ошибки полезные схемы отбираются и начинают использоваться с большим весом, а вес бесполезных схем падает
- Объяснение случайного пути - сеть случайно блуждает по разным состояниям, пока ей не повезет и она не наткнется на наполовину сформированные части схемы, которые она уже доведет до конечного состояния
- Эволюционное объяснение - сеть сначала формирует часть схемы, которая выполняет сама по себе какую-то полезную функцию, а потом уже на ее основании достраивает полную схему

Обучение

В начале обучения модель идет к решению с запоминанием пар, так как в генерализованном решении более длинные пути прокидывания градиентов и они затухают сильнее

↓

Лосс такого решения выходит на плато

↓

Начинает образовываться упрощенное запоминание, например, модель понимает, что x + y = y + x и создается специальная схема

↓

После нескольких таких итераций происходит фазовый сдвиг