

Grokking

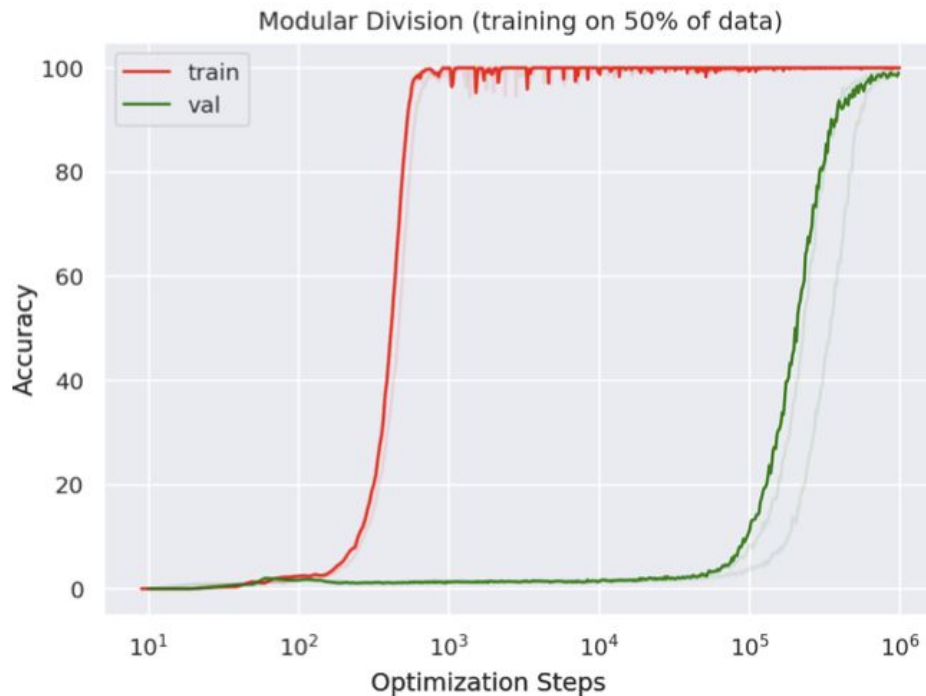
Generalization beyond overfitting on small algorithmic datasets

Докладчик: Козлова Ольга
Рецензент: Княжевский Владимир
Хакер: Иванов Данила

Что такое Grokking?

Феномен, при котором обобщающая способность модели возрастает после продолжительного периода переобучения.































Original: We show that, long after severely overfitting, validation accuracy sometimes suddenly begins to increase from chance level toward perfect generalization. We call this phenomenon 'grokking'.



А какие условия?

- Синтетические данные (таблички, в которых часть клеток закрашена)
- Маленький размер датасета
- Маленькая модель (трансформер)

★	a	b	c	d	e
a	a	d	?	c	d
b	c	d	d	a	c
c	?	e	d	b	d
d	a	?	?	b	c
e	b	b	c	?	a

☞					
			?		
					
	?				
		?	?		
				?	

Подробнее про данные

Табличка с бинарными операциями вида:

$a \circ b = c$, где a, b, c – токены, \circ –

различные варианты бинарных операций

Original: The datasets we consider are binary operation tables of the form $a \circ b = c$ where a, b, c are discrete symbols with no internal structure, and \circ is a binary operation.

Примеры операций:

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

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

$$x \circ y = x^3 + xy \pmod{p} \text{ for } 0 \leq x, y < p$$

$$x \circ y = x^3 + xy^2 + y \pmod{p} \text{ for } 0 \leq x, y < p$$

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

$$x \circ y = x \cdot y \cdot x^{-1} \text{ for } x, y \in S_5$$

Что сделали авторы?

- Обнаружили сам эффект
- Провели эксперименты
 - Выяснили закономерности (какие – дальше)
 - Визуализировали полученные результаты (картинки - дальше)
- Пригласили всех исследовать дальше

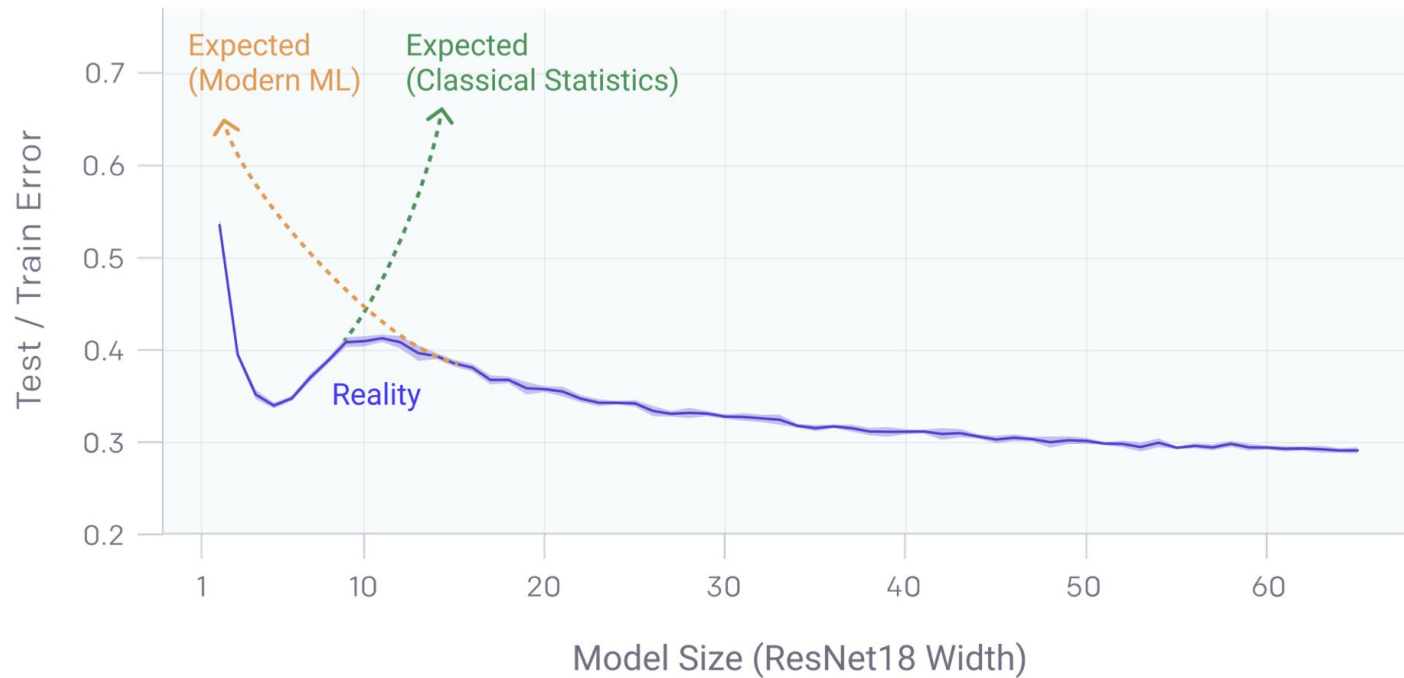
Бонус: как это обнаружили?

Ответ от авторов:

"The team at OpenAI was studying how transformers behave on algorithmic tasks, and they left a training job running over night. When they came back they were surprised to see that the transformer had somehow solved the task."

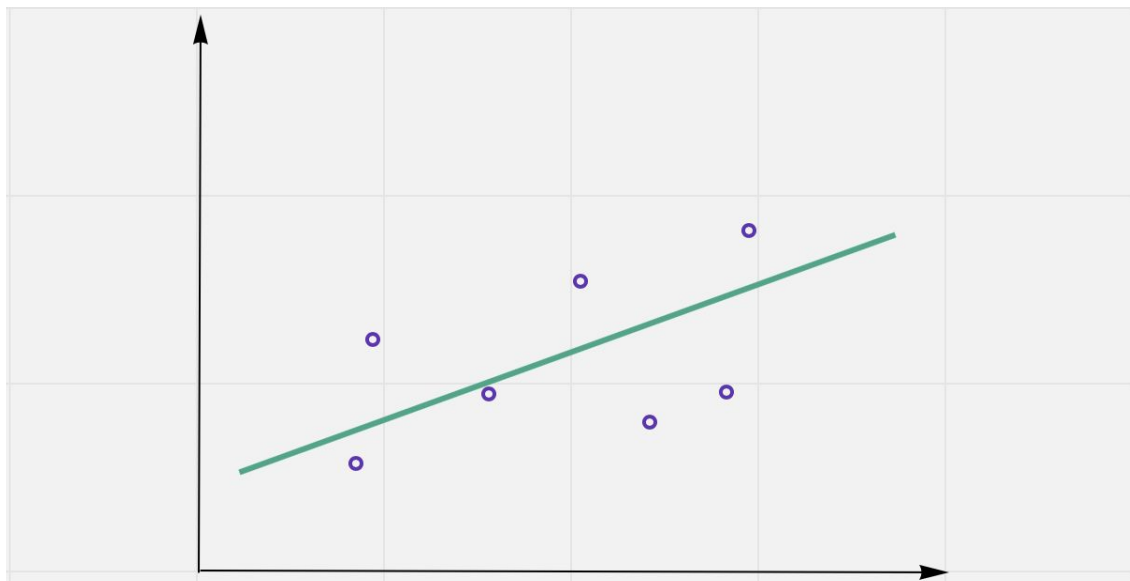
Как это работает?

Double descent



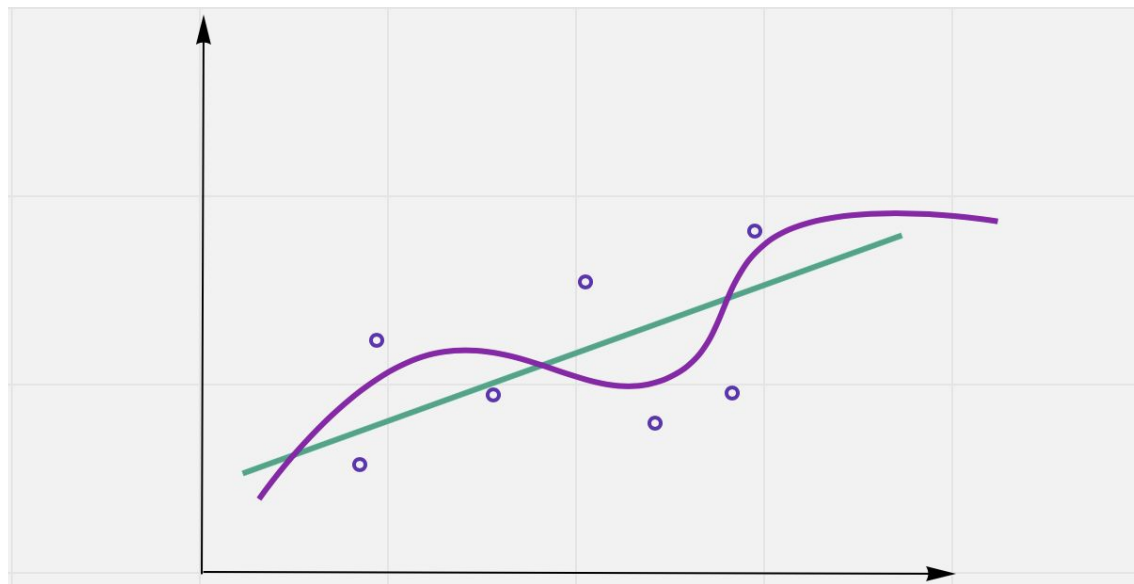
Double descent на примере

Очень простая модель



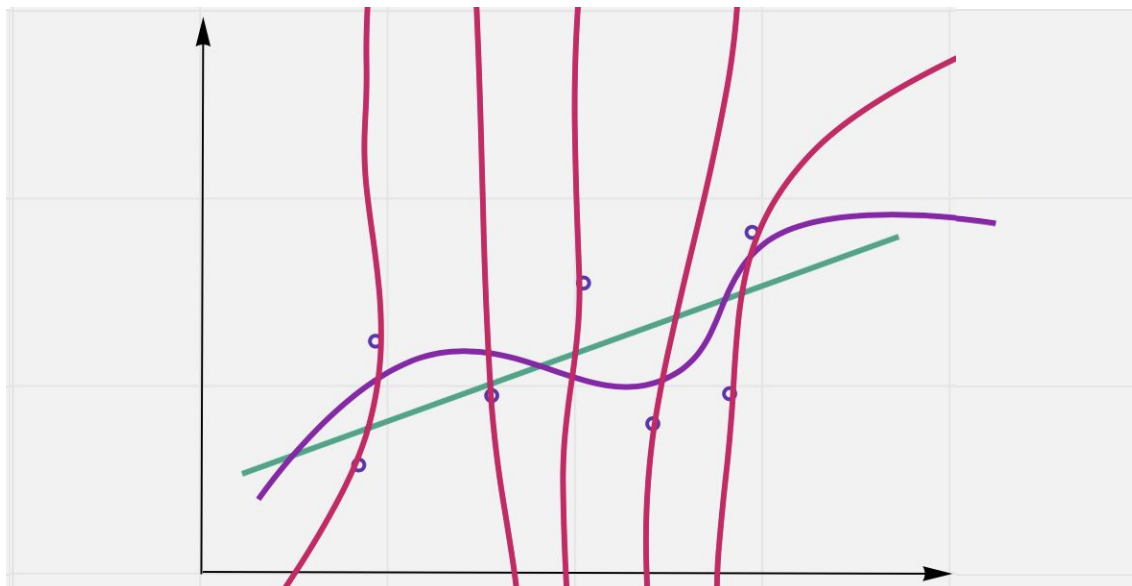
Double descent на примере

Немного усложнили – хорошо, но не идеально



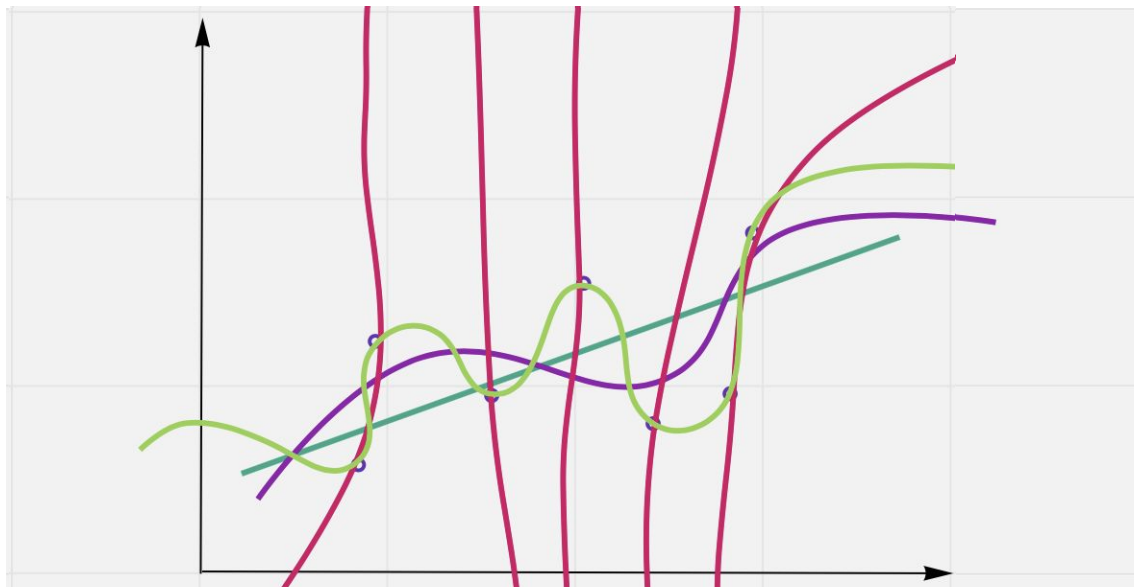
Double descent на примере

Слишком много параметров – переобучение



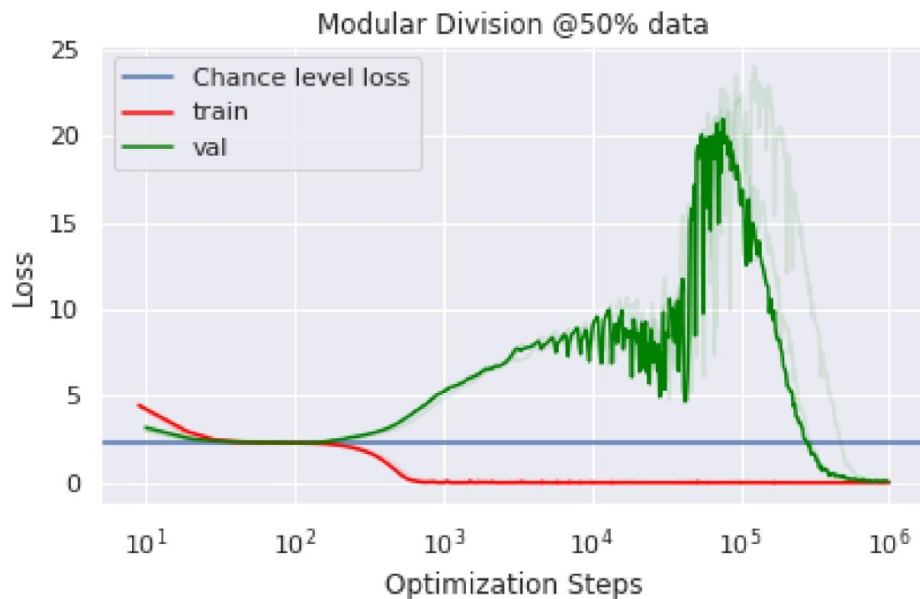
Double descent на примере

Так много параметров, что модель смогла выучить сложную закономерность

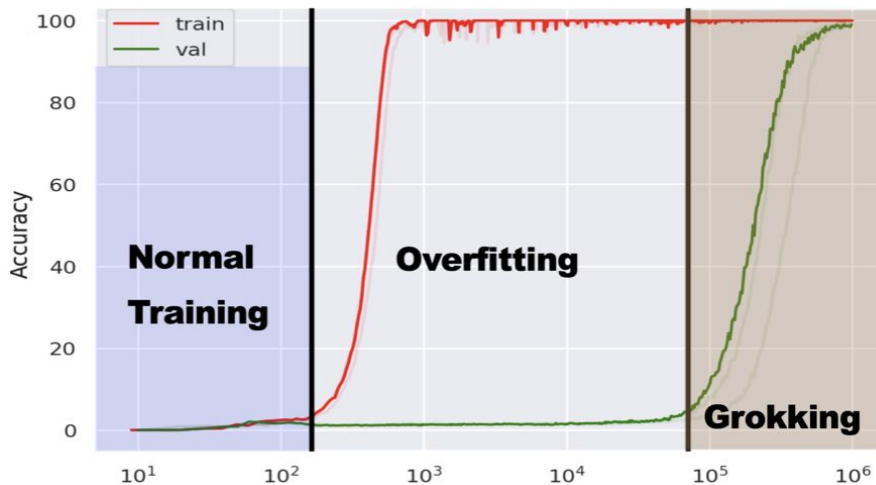
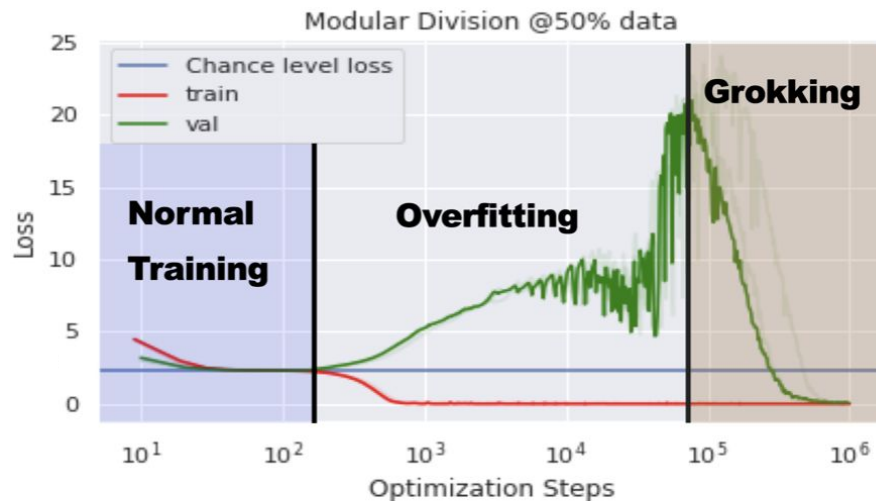


Double descent

Авторы тоже наблюдают такой эффект для функции потерь



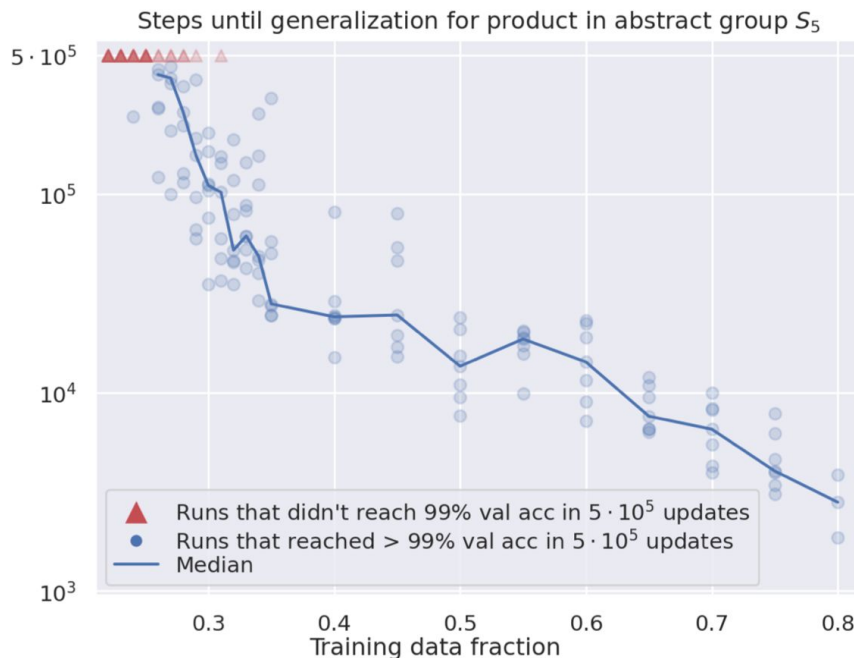
Double descent



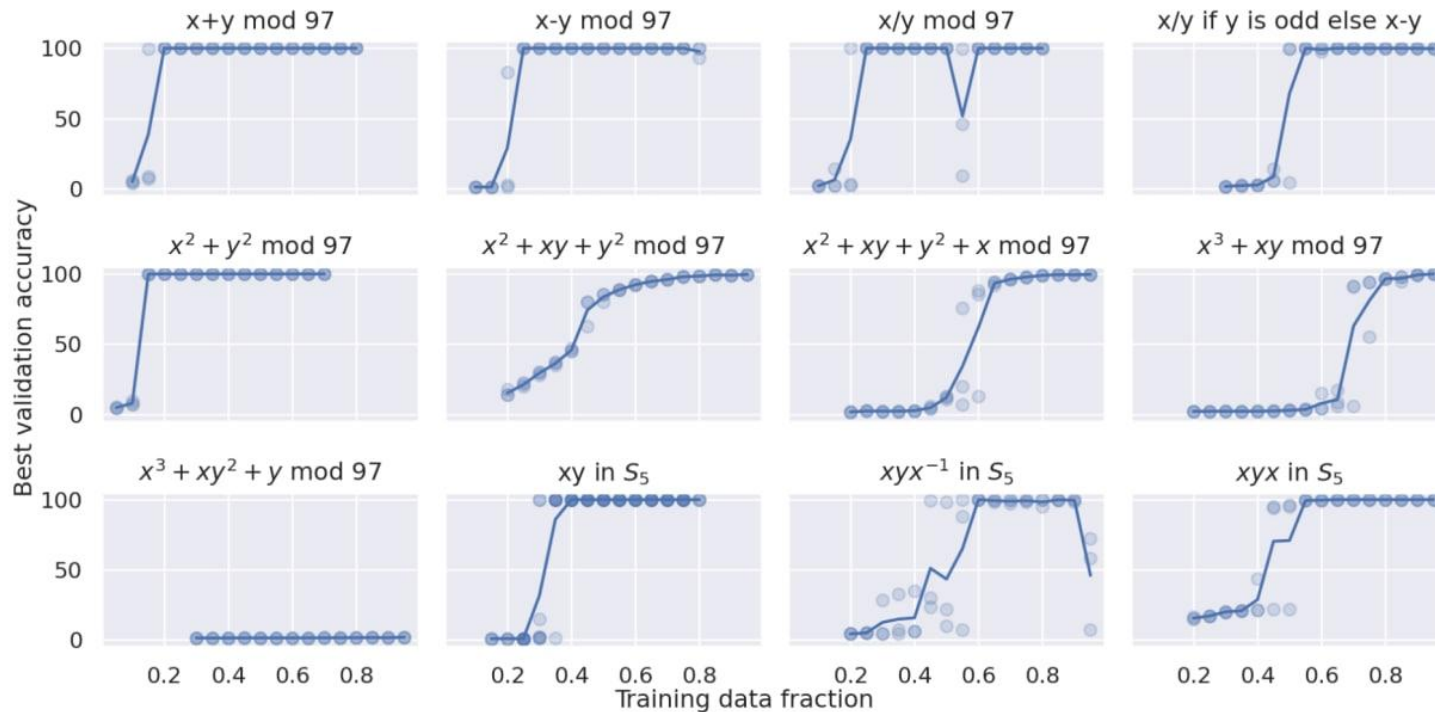
Эксперименты

Объем обучающих данных

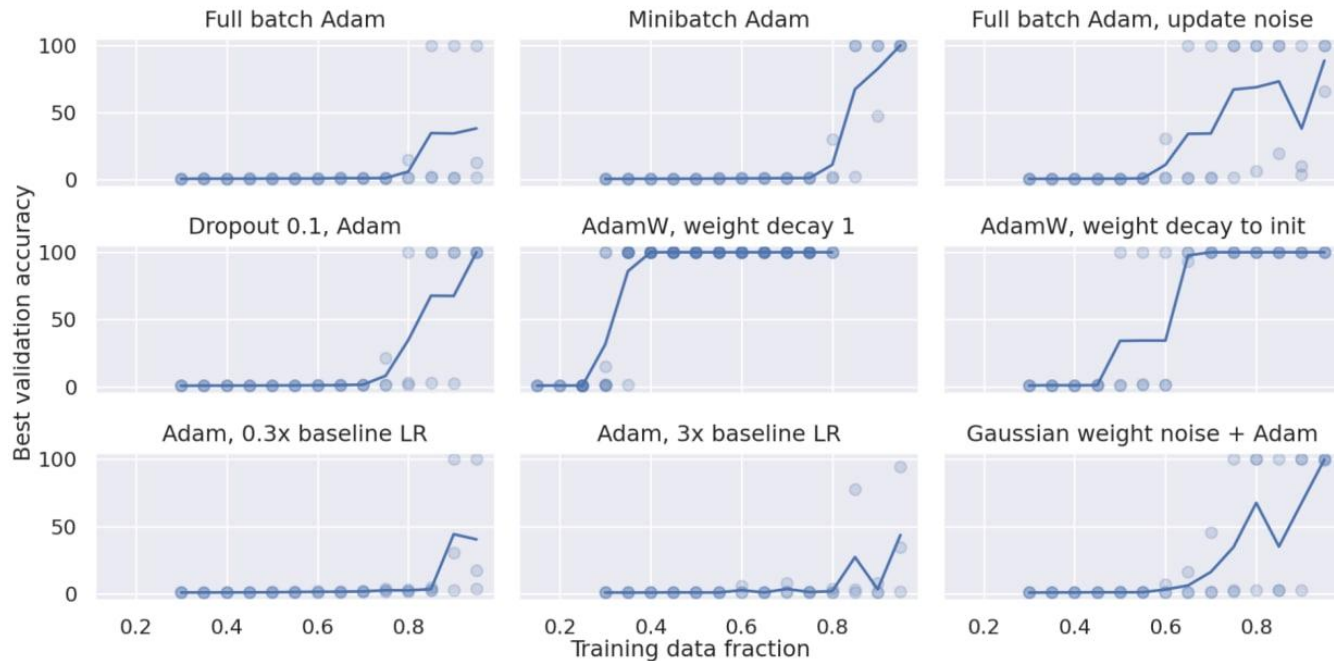
- Пробуем менять объем тренировочных данных
- Чем больше данных даем, тем быстрее обобщается



Различные бинарные операции



Эксперименты с оптимизатором



1000

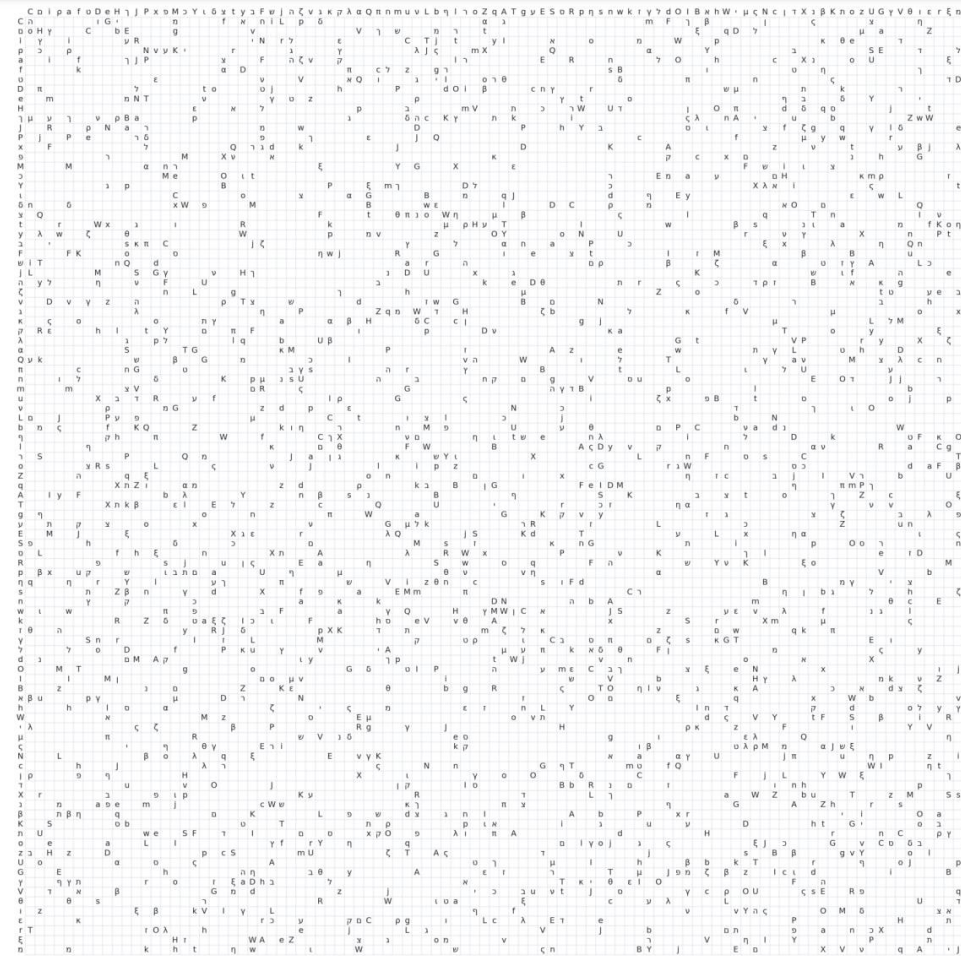


1000

Если вам скучно

One of the binary operation tables presented to the networks that the network can perfectly fill in. Each symbol is represented as a letter in English, Hebrew, or Greek alphabet for reader's convenience.

We invite the reader to guess which operation is represented here.



ИСТОЧНИКИ

- Сама статья на arxiv [en]:
 - <https://arxiv.org/abs/2201.02177>
- Постер с воркшопа [en]:
 - https://mathai-iclr.github.io/papers/posters/MATHAI_29_poster.png
- Обзор на youtube [en]:
 - <https://www.youtube.com/watch?v=dND-7llwrpw>
- Обзор в telegram [ru]:
 - https://t.me/gonzo_ML/831
- Картиночки про double descent (красивые):
 - <https://mlu-explain.github.io/double-descent/>

Некоторые заметки из рецензии

Авторы

Open AI

ICLR (International Conference on Learning Representations), 1st Mathematical Reasoning in General Artificial Intelligence Workshop

Обзор последующих работ

Чаще всего цитирование работы выглядит так:

Predictability and Surprise in Large Generative Models: "Though performance is predictable at a general level, performance on a specific task can sometimes emerge quite unpredictably and abruptly at scale".

Unsolved Problems in ML Safety: "We are better able to make models safe when we know what capabilities they possess".

Обзор последующих работ

Есть и прямые продолжения:

Understanding Grokking: An Effective Theory of Representation Learning:
аналогичные результаты и попытка вывести теоремы.

A Mechanistic Interpretability Analysis of Grokking: reverse engineering, сложение по модулю 97 делается через дискретные преобразования Фурье.

Сильные стороны

Все понятно объяснено, эксперименты описаны подробно.

Само открытие, что grokking происходит стабильно для определенных данных и моделей.

Потенциал для исследований.

Слабые стороны

Неясно, получится ли найти у статьи какое-то применение на практике.

Одна модель.