НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

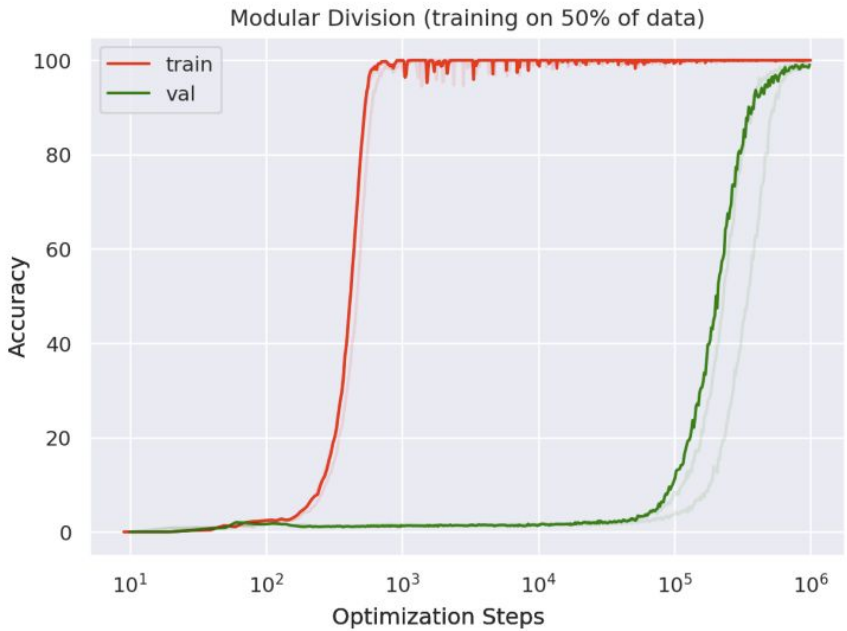
Grokking: Generalization beyond overfitting on small algorithmic datasets

Докладчик: Воробьев Николай
Рецензент: Лишуди Дмитрий
Хакер: Асланов Алишер

Москва, 2022

Введение

Grokking – это феномен внезапного появления у модели обобщающей способности после длительного переобучения



Ограничения и особенности

Данные

- синтетические таблицы
- небольшой размер
- на вход подаем токены

Модель

- трансформер (2 слоя)
- 400к параметров

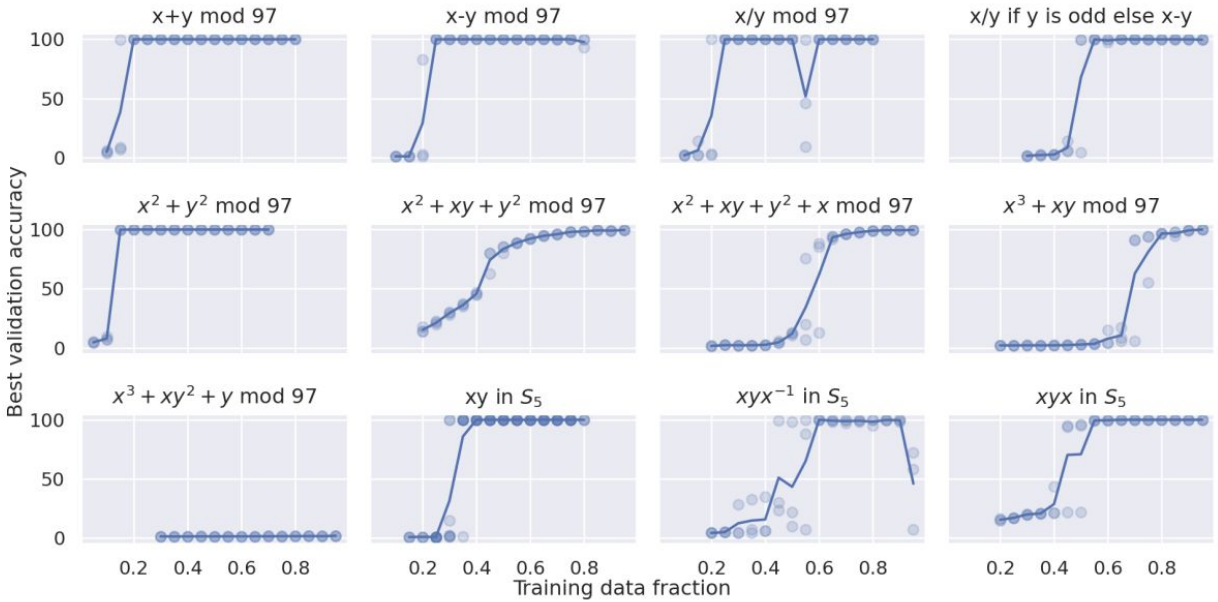
Пример данных

⊛	😁	😇	😌	😍	😎
😁	😁	😍	?	😌	😍
😇	😌	😍	😍	😁	😌
😌	?	😎	😍	😇	😍
😍	😁	?	?	😇	😌
😎	😇	😇	😌	?	😁

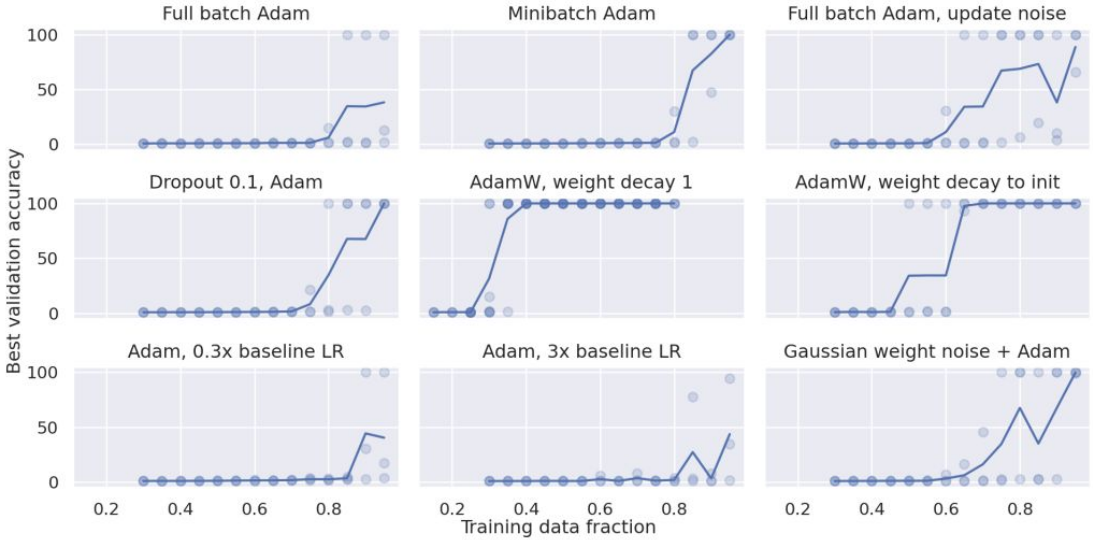
★	a	b	c	d	e
a	a	d	?	c	d
b	c	d	d	a	c
c	?	e	d	b	d
d	a	?	?	b	c
e	b	b	c	?	a

Эксперименты

Бинарные операции



Параметры обучения



Визуализация результатов

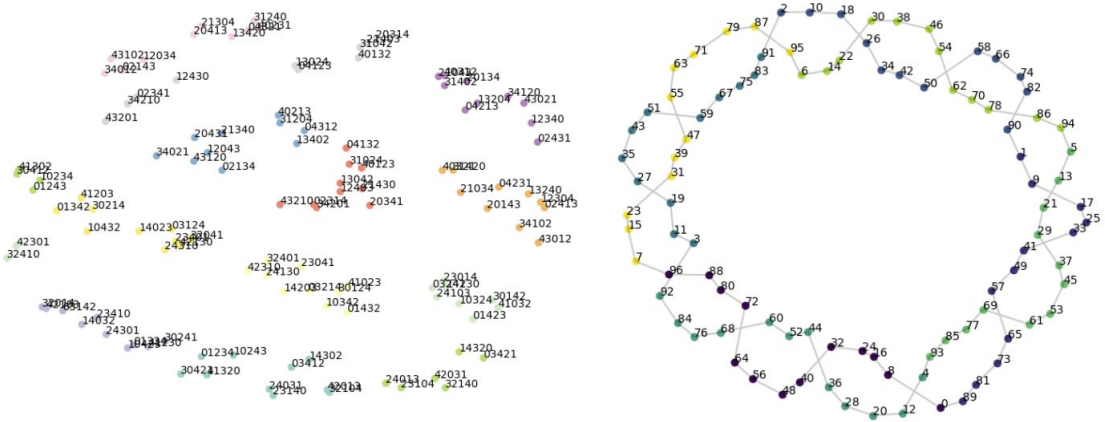
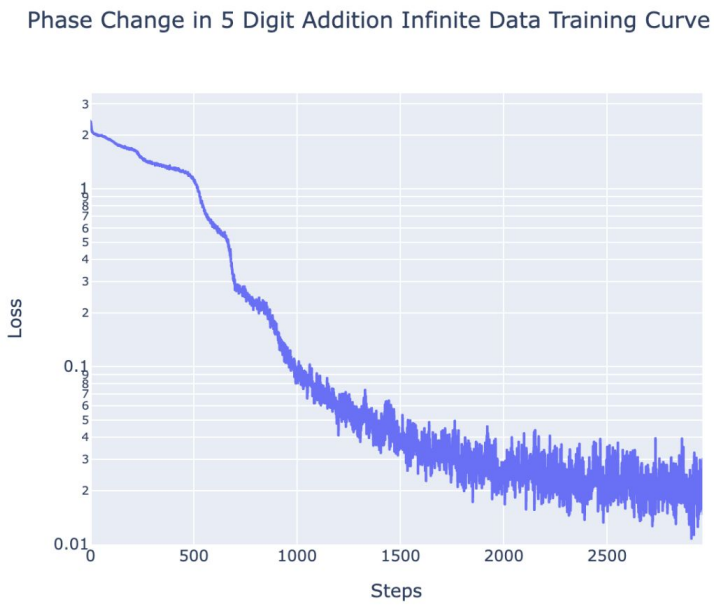
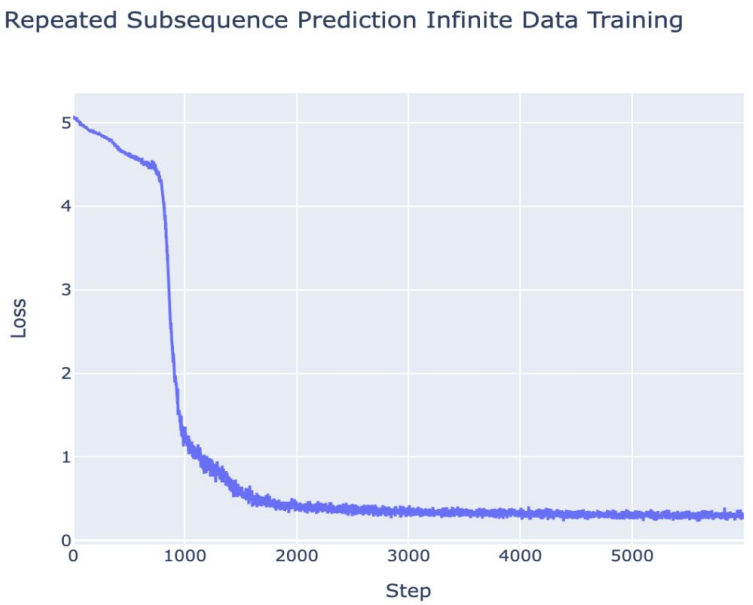


Figure 3: **Left**. t-SNE projection of the output layer weights from a network trained on S_5. We see clusters of permutations, and each cluster is a coset of the subgroup $\langle (0,3)(1,4), (1,2)(3,4) \rangle$ or one of its conjugates. **Right**. t-SNE projection of the output layer weights from a network trained on modular addition. The lines show the result of adding 8 to each element. The colors show the residue of each element modulo 8.

Интерпретация

Phase changing: примеры





Phase changing: наблюдение

Смена фазы на обучающей и тестовой выборке

Modular addition mod 113 loss curve, trained on 95% of the data

Phase changing: наблюдение

Смена фазы на обучающей и тестовой выборке

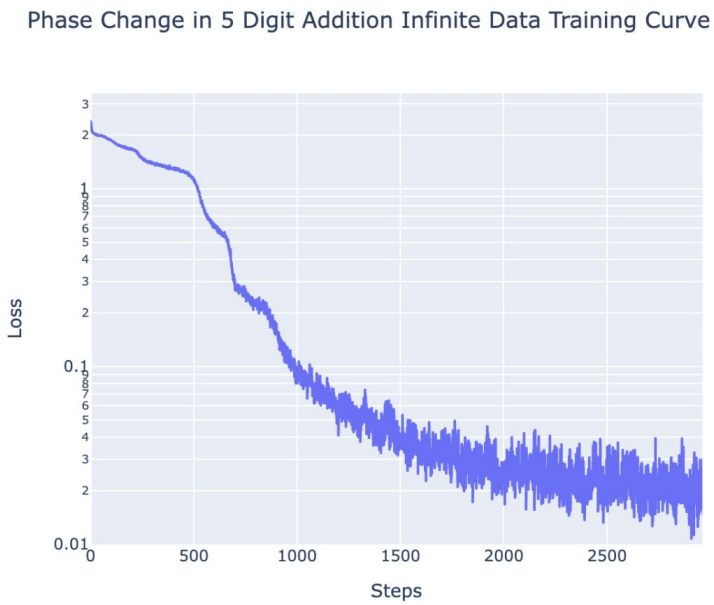
Modular addition mod 113 loss curve, trained on 95% of the data

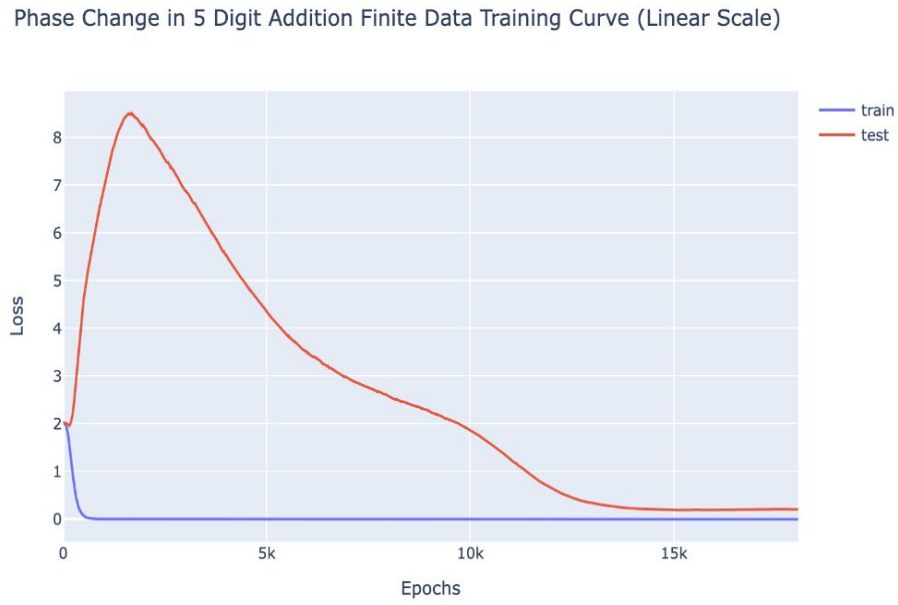
Аналогичное поведение на других задачах!

Идея

Grokking = Phase changing + Regularisation + Limited Data

Grokking = Phase changing + Regularisation + Limited Data



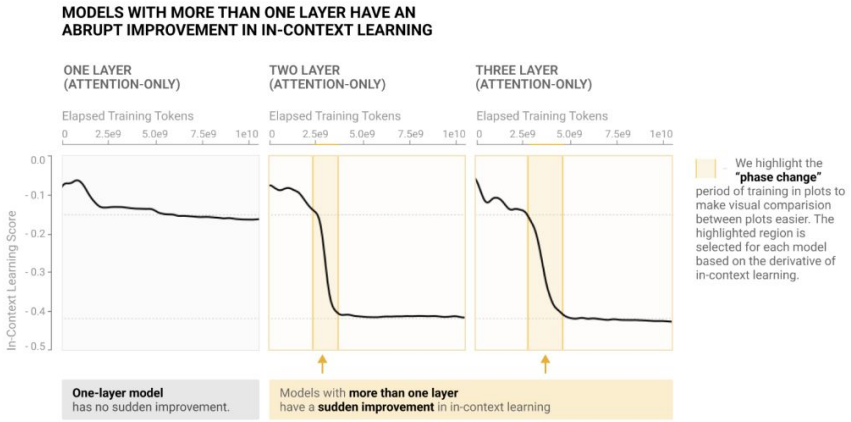


Phase changing: интуиция

Фазовые изменения присущи композициям

Phase changing: интуиция

Фазовые изменения присущи композициям



In-context learning curves for small transformers (figure copied from Anthropic's Induction Heads paper) - 2L and 3L models develop induction heads and show a phase change, 1L models do not

Grokking: интуиция

- Модель хочет запомнить выборку
- Много данных → сильно проще обобщить
- Мало данных → проще запомнить
- Среднее количество данных → grokking

Источники

- Оригинальная статья на arxiv [en]: https://arxiv.org/pdf/2201.02177.pdf
- Постер с воркшопа [en]: https://mathai-iclr.github.io/papers/posters/MATHAI_29_poster.png
- Блог пост [en]: https://www.alignmentforum.org/posts/N6WM6hs7RQMKDhYjB/a-mechanistic-interpretability-analysis-of-grokking