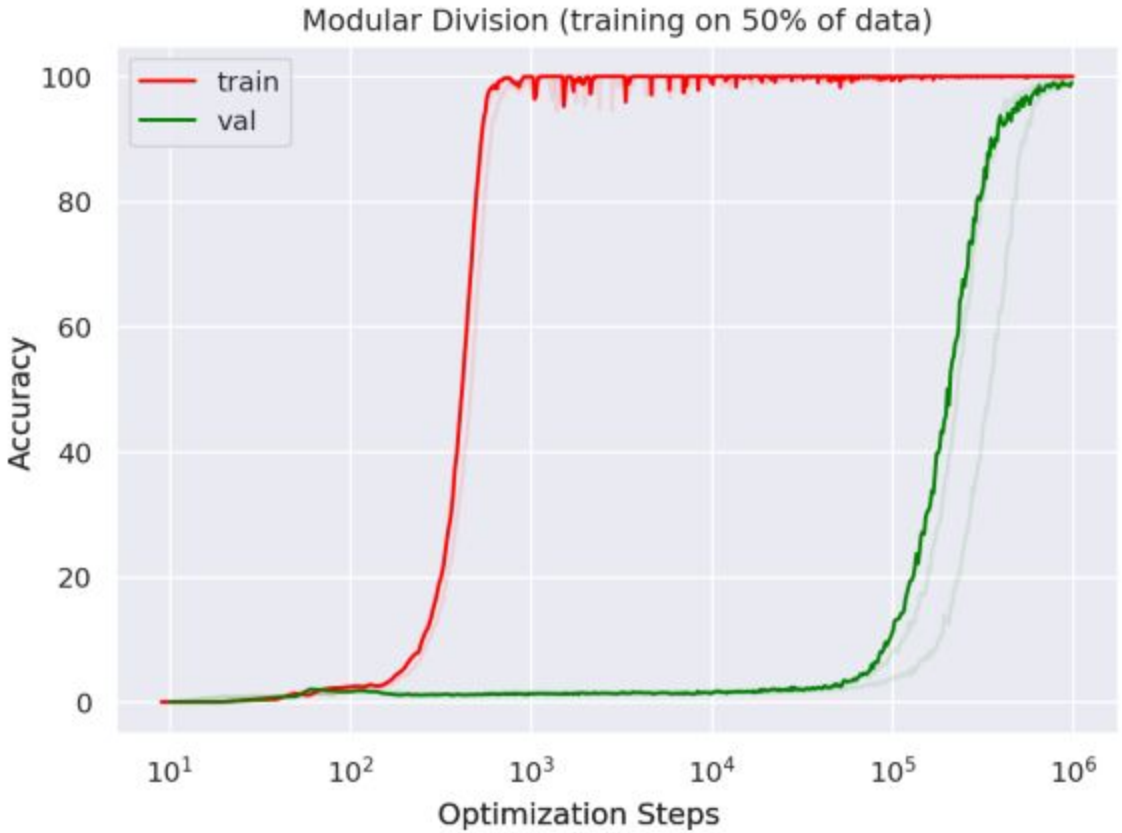
GROKKING: GENERALIZATION BEYOND OVERFITTING ON SMALL ALGORITHMIC DATASETS

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin

Обзор-рецензия

Рецензент: Лишуди Дмитрий


Modular Division (training on 50% of data)
train
val
Accuracy
100
80
60
40
20
0
10^1
10^2
10^3
10^4
10^5
10^6
Optimization Steps

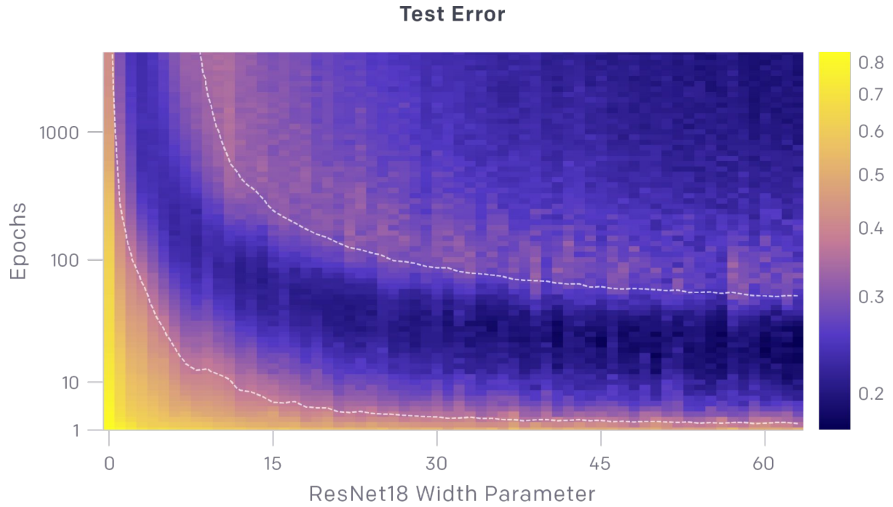
Истоки

- Грокать – интуитивно понимать; схватывать
- Авторы статьи из OpenAI; занимались в основном LM.
- Изначально смотрели на трансформеры для алгоритмических задач.
- Открыли гроккинг случайно, забыли выключить обучение модели.
- Статья совсем маленькая; подали только на воркшоп ICLR 2021, но не саму конференцию.
- Позиция: смотрите какую интересную вещь нашли, подробно не изучили, но хорошо проверили. Дальше пусть сообщество думает.

Похожий феномен: **Double Descent**

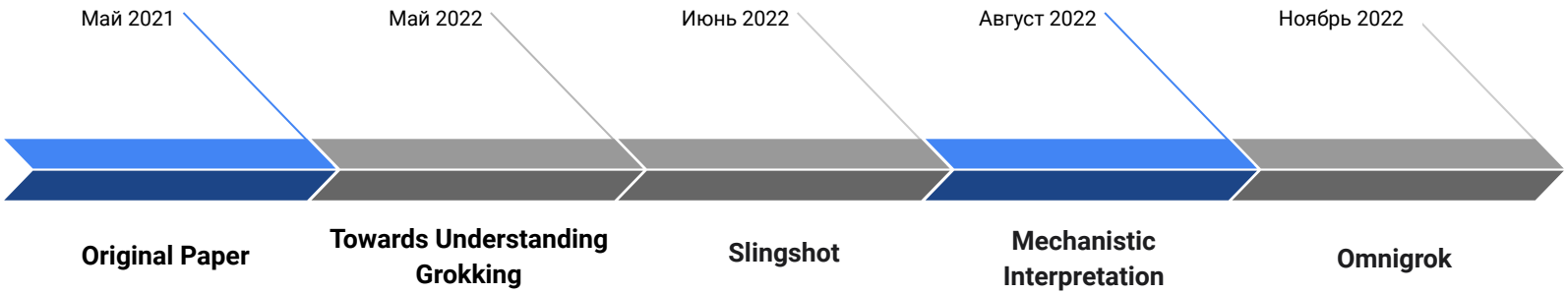
Классический (model-wise) Double Descent

Epoch-wise Double Descent



Источник: https://openai.com/blog/deep-double-descent/

Таймлайн


Май 2021
Май 2022
Июнь 2022
Август 2022
Ноябрь 2022
Original Paper
Towards Understanding Grokking
Slingshot
Mechanistic Interpretation
Omnigrok

Нулевое объяснение: random walk

Самое популярное объяснение на заре гроккинга (например см Grokking "Grokking"):

Веса просто долго блуждали по ландшафту потерь пока не повезло упасть в широкий оптимум.

Аналогия: Долговременный эксперимент по эволюции e. coli.

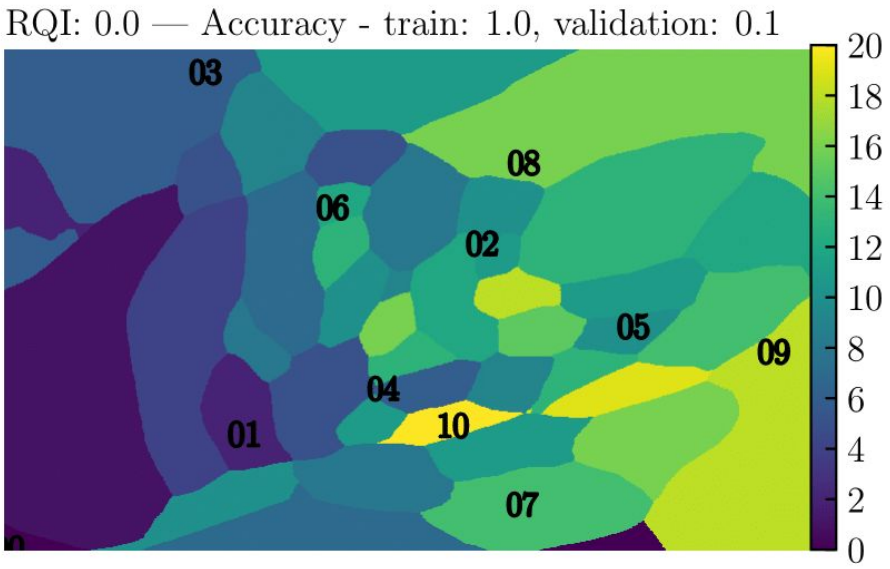
- Эксперимент идёт с 1988 года.
- Бактерии в среде из воды, цитрата натрия и глюкозы.
- Бактерии питаются глюкозой; она основной ограничитель размножения.
- Спустя 20 лет, одна из популяций мутировала и научилась усваивать цитрат натрия.
- Из-за этого получилось в разы больше размножиться, большой эволюционный скачок.

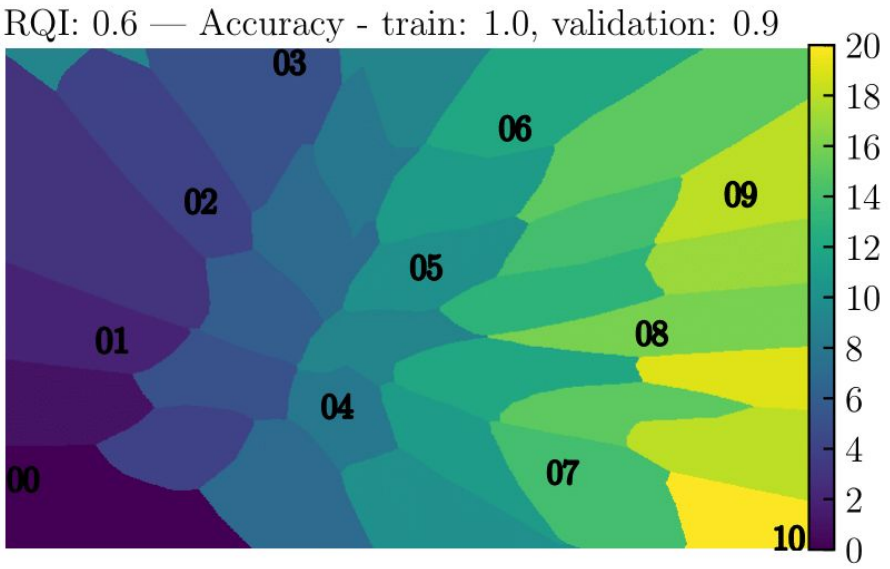
Towards Understanding Grokking: An Effective Theory of Representation Learning

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, Mike Williams

- Авторы из MIT; занимаются физикой/интерпретацией DL.
- Прошли на NeurIPS 2022.
- Анализирует гроккинг с точки зрения математики.
- Рассматривается модель эмбеддинг E_i – декодер Dec.
- $(a, b) \rightarrow D(E_a + E_b)$
- Утверждение: главное – правильно обучить представления E_i.
- Будем анализировать представления без привязки к декодеру.

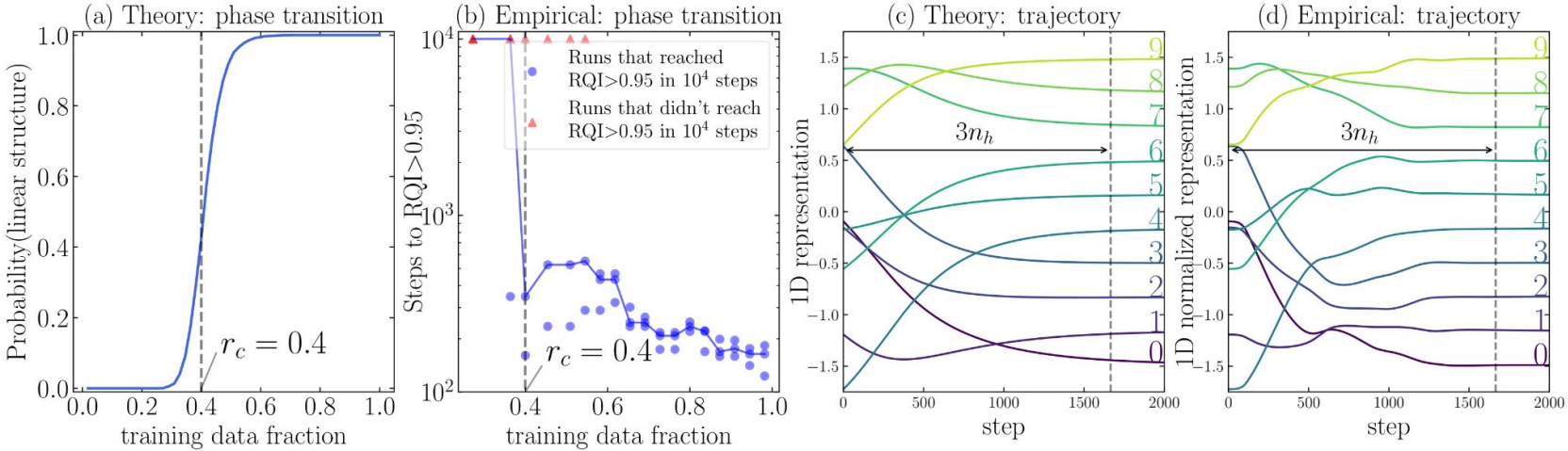
Towards Understanding Grokking: анализ сложения



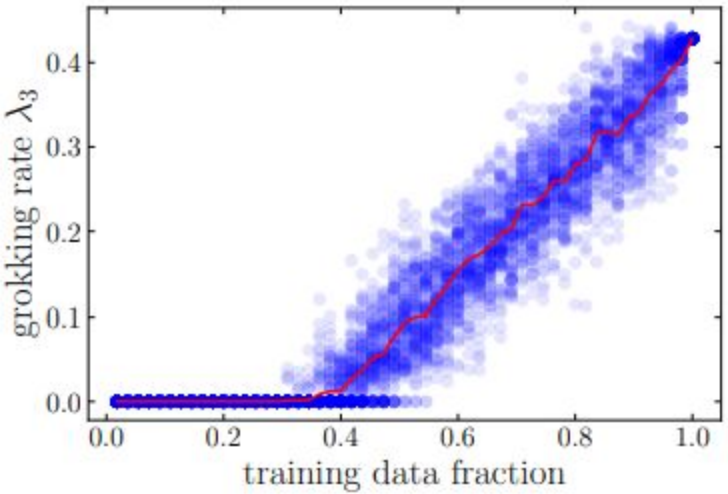


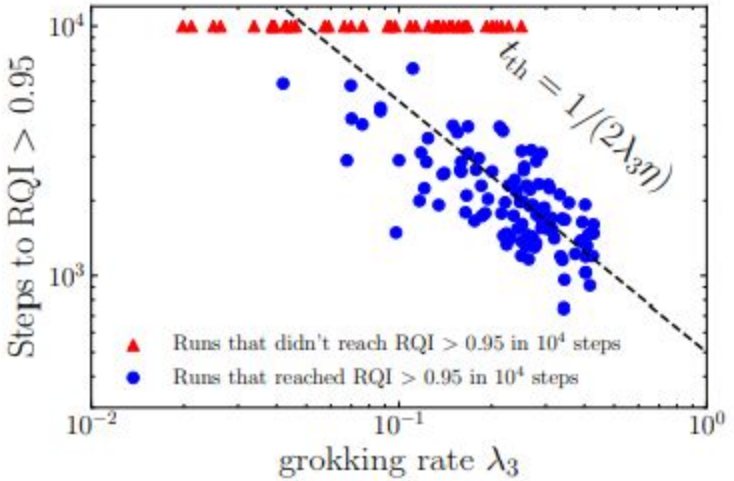
Heatmap для двумерных эмбеддингов в задаче сложения на [0, … 10]

Towards Understanding Grokking: теоретические/эмпирические результаты



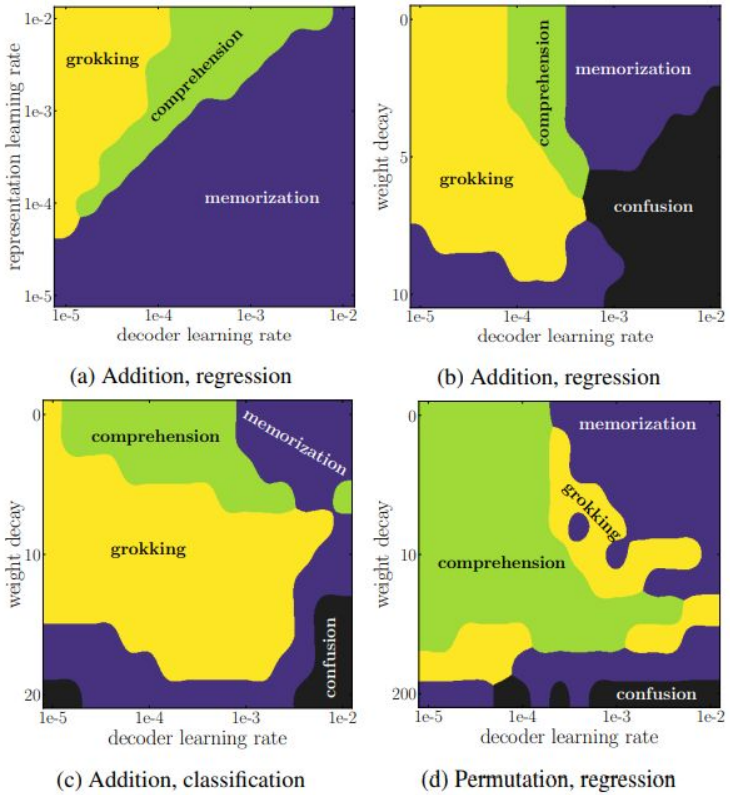
Towards Understanding Grokking: теоретические/эмпирические результаты





Из теории: $E(t) \sim \exp(-\lambda_3 t)$

Towards Understanding Grokking: фазовые диаграммы



Towards Understanding Grokking: выводы

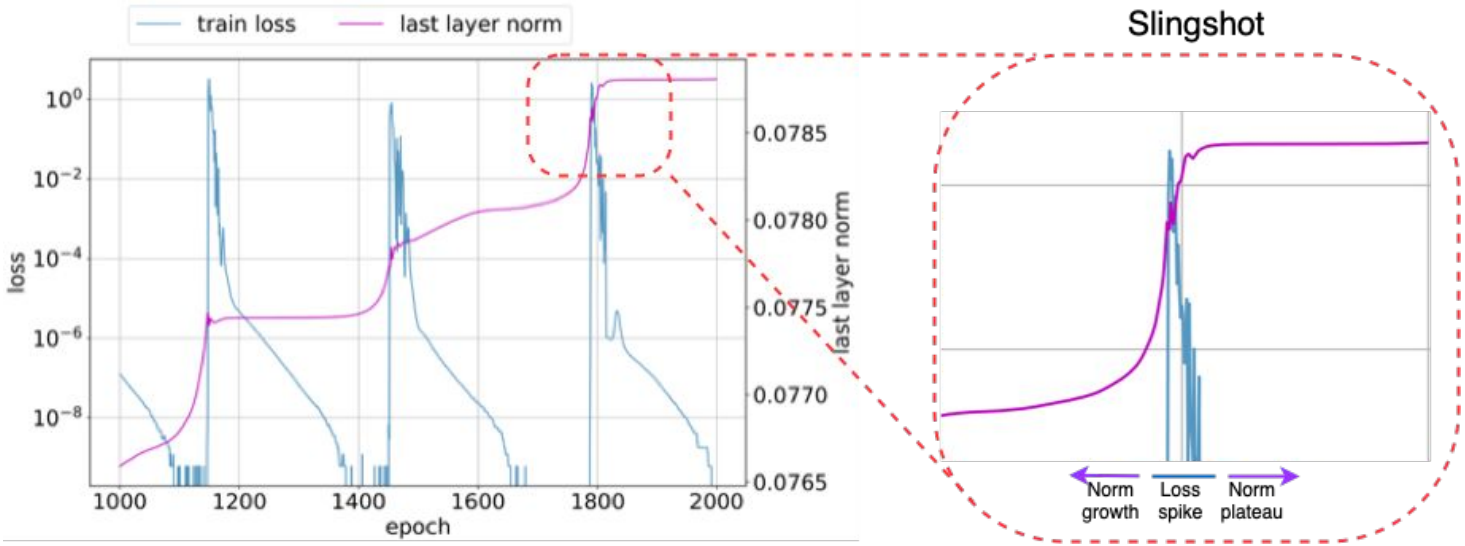
- В алгоритмическом датасете главное – обучение хорошего представления.
- Есть критический размер обучающей выборки, который отличает обобщающее решение от необобщающего
- Гроккинг является фазой между запоминанием и обобщением.

The Slingshot Mechanism: An Empirical Study of Adaptive Optimizers and the Grokking Phenomenon

Vimal Thilak, Etai Littwin, Shuangfei Zhai, Omid Saremi, Roni Paiss, Joshua Susskind

- Прошёл на “Has it Trained Yet?” воркшоп NeurIPS 2022
- Что если в определённый момент мы делаем очень большой прыжок?
- Механизм рогатки (slingshot)

Slingshot: иллюстрация

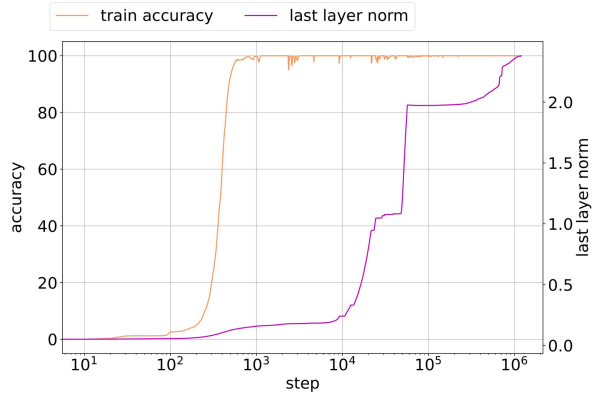


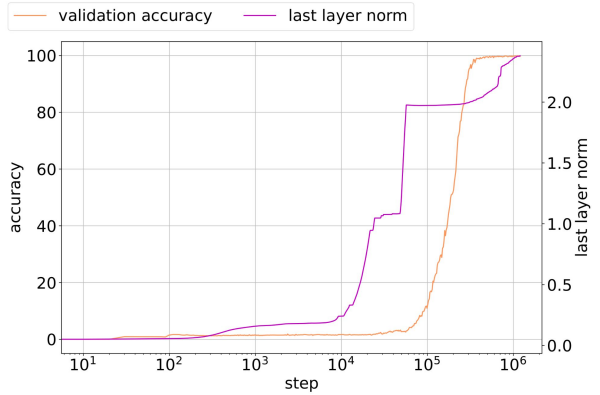
Обучение делится на 2 фазы:

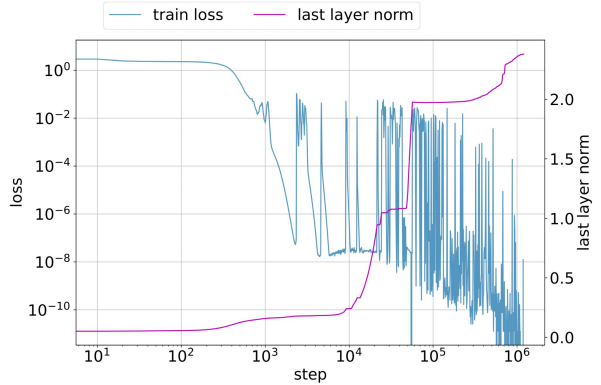
Стабильная: норма последнего слоя на плато, потери падают

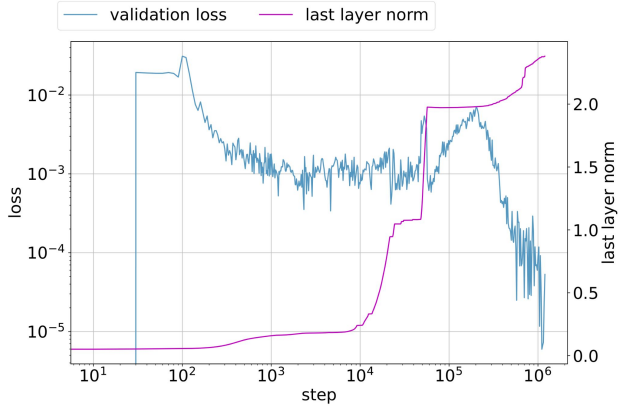
Нестабильная: норма последнего слоя резко растёт, лосс взрывается.

Slingshot: графики потерь и точности









Slingshot: объяснение авторов

- Такой эффект наблюдается только для адаптивных оптимизаторов (Adam/AdamW/RMSProp) и только в поздних стадиях обучения (переобучение).
- Похоже, что норма последнего слоя растёт, пока веса не попадают в минимум с маленькой кривизной.
- В этих минимумах малые направления градиента увеличиваются и оптимизатор “стреляет” веса в другой регион ландшафта потерь.

Slingshot: сомнения

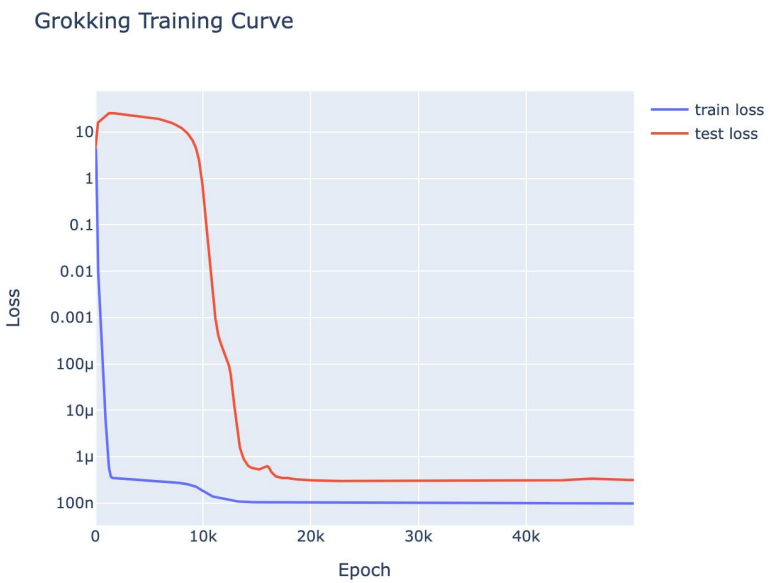


График потерь для fp64 логитов

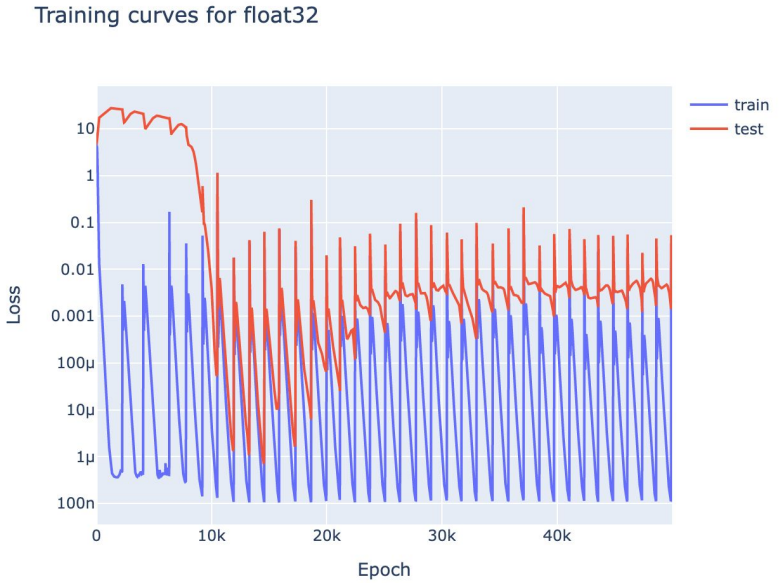


График потерь для fp32 логитов

Минимум log_softmax в fp32 – около 1e-7. Меньшие значения зануляются!

Slingshot: в чем же дело?

- Когда модель достаточно хорошо выучивает некоторые объекты, из-за округления потери на них падают в 0 и они выпадают из выборки.
- После этого модель страшно переобучается на оставшихся точках, моментум протащит это достаточно далеко

A Mechanistic Interpretability Analysis of Grokking

Neel Nanda, Tom Lieberum

- Ранее авторы писали про реверс-инжиниринг трансформеров с помощью *схем* (*circuits*).
- Применили этот подход к трансформеру из оригинальной статьи.
- Оказывается модель предсказывает через DCT.
- Целевая функция переодична, отсюда и тригонометрия.
- Как меняются репрезентации?

A Mechanistic Interpretability: репрезентации

- **Это не случайное блуждание, процесс постепенный!**

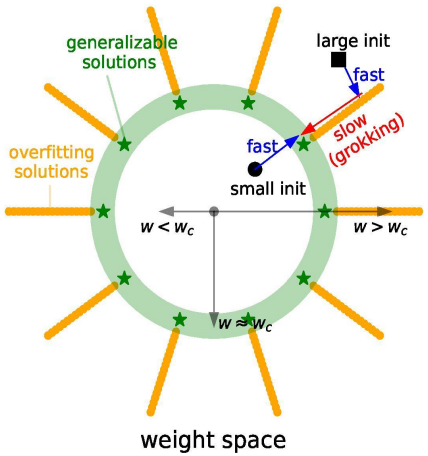
Omnigrok: Grokking Beyond Algorithmic Data

Ziming Liu, Eric J. Michaud & Max Tegmark

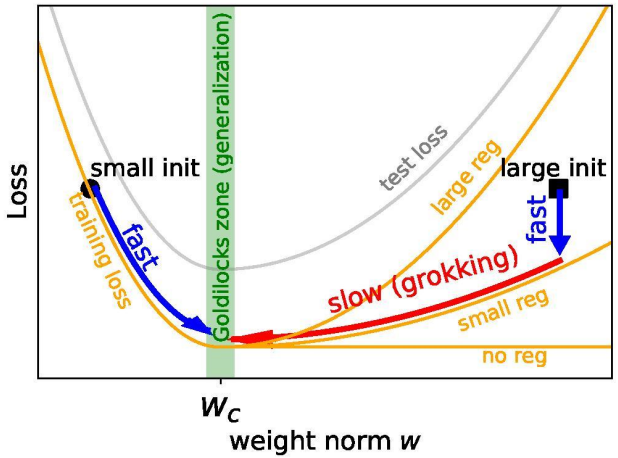
- Очень хорошие рецензии, пройдет на ICLR 2023.
- Изучает гроккинг эмпирически, с точки зрения ландшафта функции потерь.
- Те же авторы, что и у *Towards Understanding Grokking.*
- Ключ к пониманию – норма весов!

Omnigrok: Механизм LU

- Потери на тесте формы “U”: bias-variance tradeoff.
- Потери на трейне формы “L”: большие веса не мешают переобучаться.
- Обобщающая область лежит в сферической оболочке.
- Больше данных - шире оболочка.
- Если w_init слишком большой, то спуск медленный.



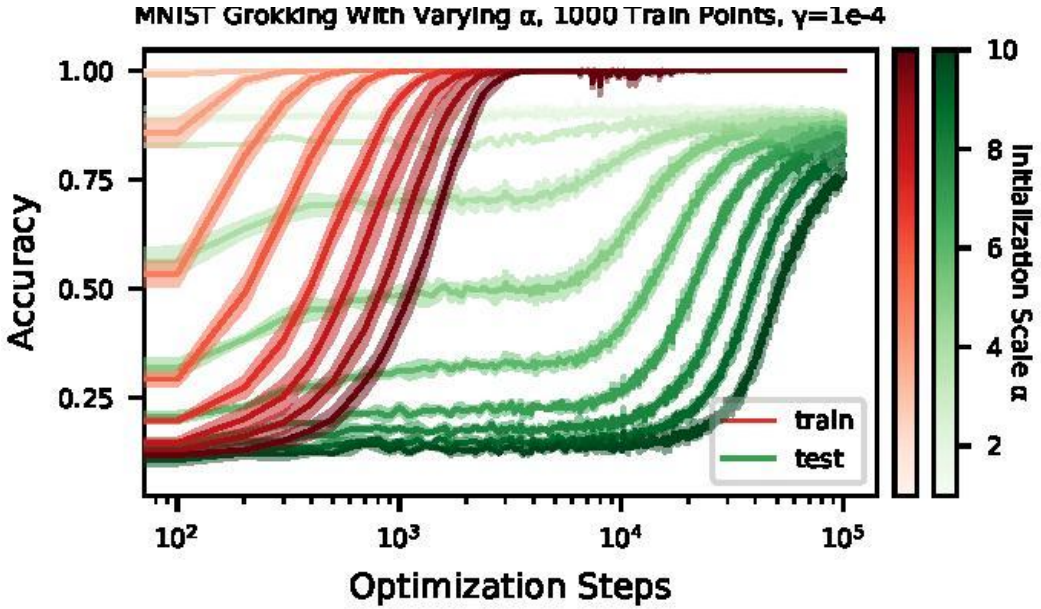
Вид “сверху” на ландшафты потерь.



Разрез ландшафтов потерь.
Потери на трейне формы **L**
Потери на тесте формы **U**.

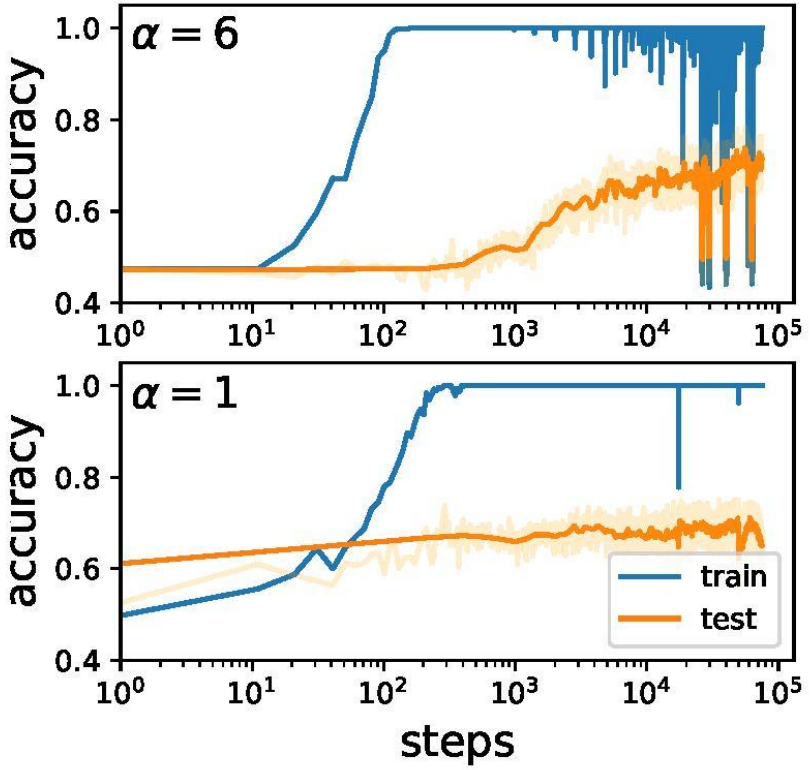
Omnigrok: грокаем MNIST

- Давайте искусственно завысим начальные веса.
- Гроккинг видно даже на MLP классификации картинок!



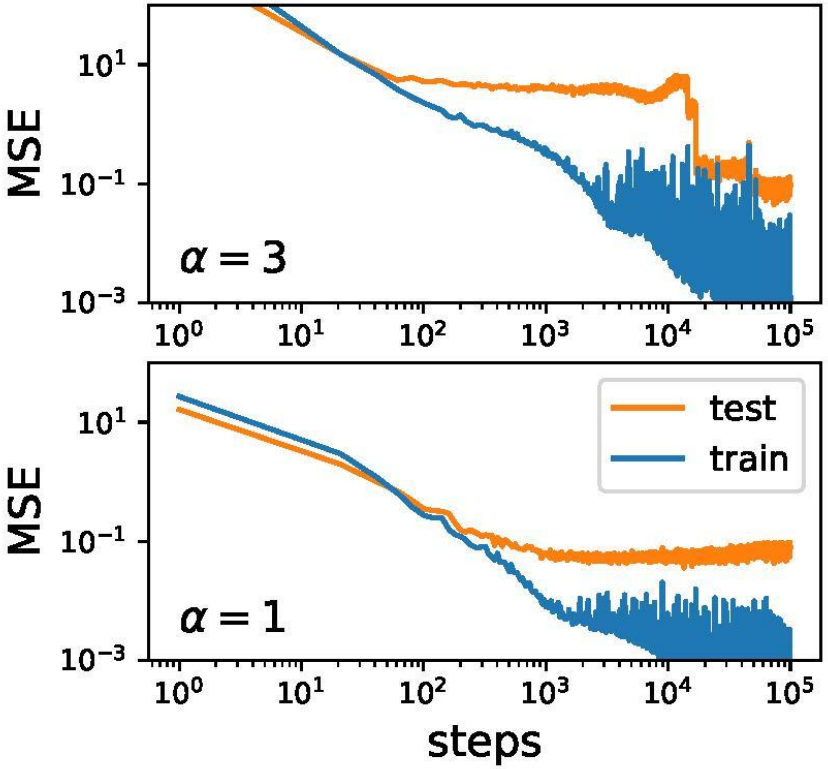
Omnigrok: грокаем тексты

- Давайте искусственно завысим начальные веса.
- Гроккинг видно даже на MLP классификации картинок!
- А ещё для LSTM классификации текстов.



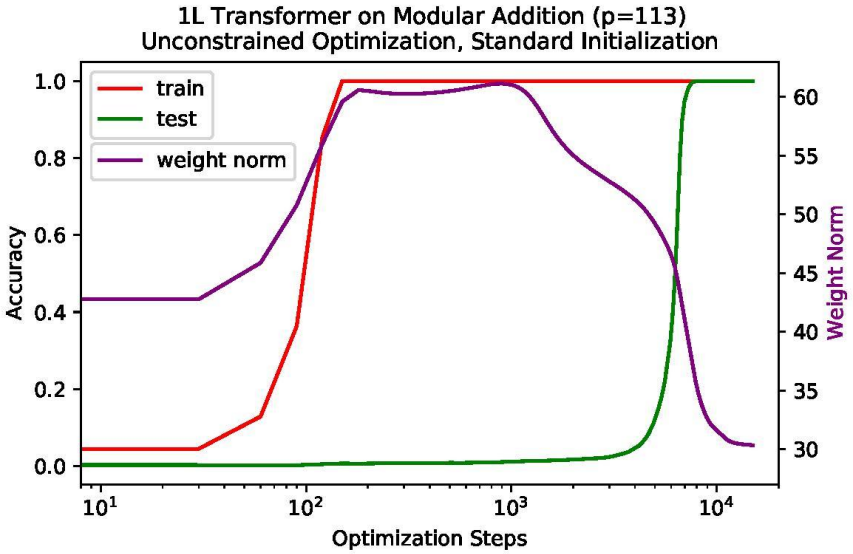
Omnigrok: грокаем графы

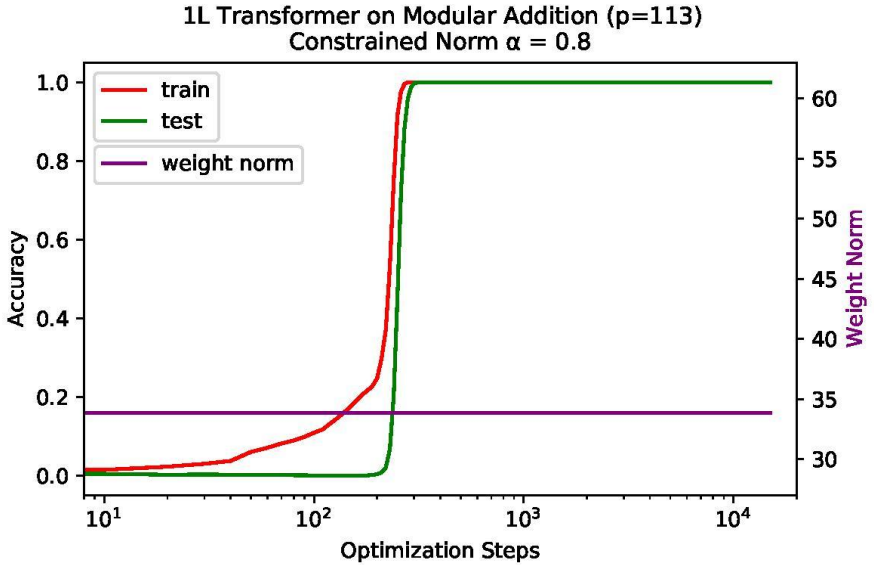
- Давайте искусственно завысим начальные веса.
- Гроккинг видно даже на MLP классификации картинок!
- А ещё для LSTM классификации текстов.
- А ещё для GCNN регрессии свойств молекул.



Omnigrok: побеждаем алгоритмический гроккинг

Попробуем после шага оптимизации нормализовать веса модели.





Характеристика статьи

Сильные стороны:

- Обнаружили новое явление, вдохновили новые исследования.
- Хороший ablation: проверили что феномен реален, а не какая-то ошибка.

Слабые стороны:

- Статья идейно не закончена, большая часть результатов в последующих работах.
- Узкая постановка: только трансформер и алгоритмические данные
- Скорее всего на практике не применимо, просто забавный эффект.

Источники

- Оригинальная статья
- Deep Double Descent
- Towards Understanding Grokking
- Omnigrok
- The Slingshot Effect
- Блогпост про mechanistic interpretation