Grokking и вокруг него.

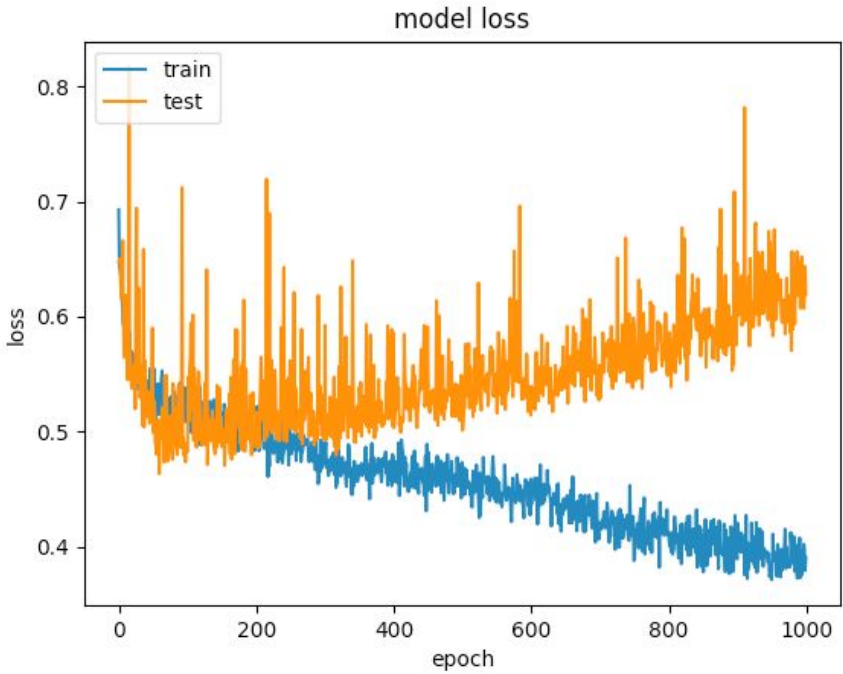
Подготовил Бельский Антон

Введение.

Классическое переобучение:
С чрезмерным уменьшением потерь на тестовой выборке начинает ухудшаться качество предсказаний
на тестовой(валидационной) выборке.

Смысл:
Модель начинает начинает ‘запоминать’ элементы тренировочной выборки - путем подбора коэффициентов модели.

Пример решения: можно остановиться до того как модель переобучится.

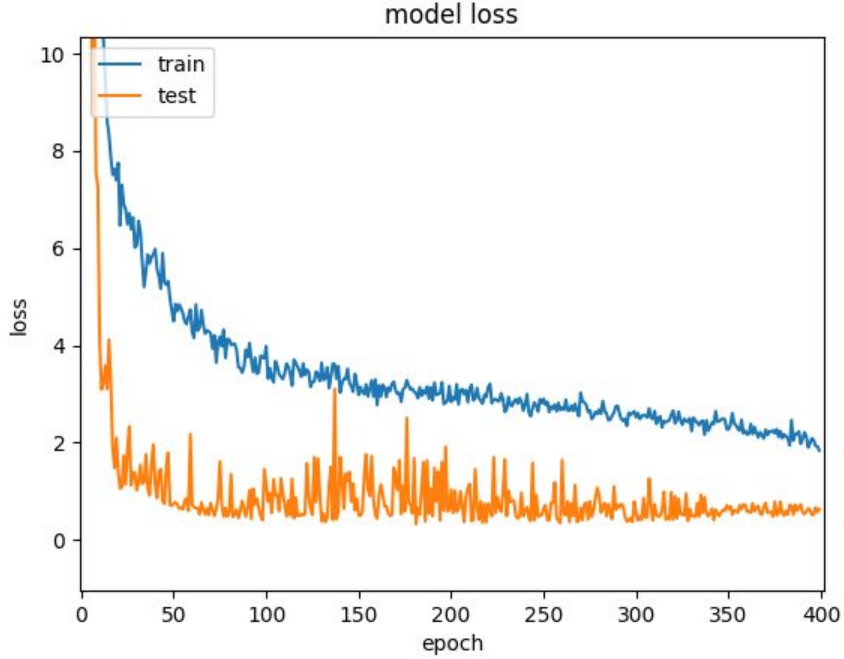


Запоминание(memorization)

Часто причиной может становиться малая тестовая выборка.
С запоминанием уменьшается обобщающая способность модели.

Обобщение(generalization):

Классическое решение: weight decay, например L2 регуляризация - следует из наблюдения: при переобучение получаются большие веса модели.



Постановка задачи - Grokking

Гипотеза: при некоторых факторах решение путем запоминания хуже решения путем обобщения.

Идея: продолжим обучение после того как заметили эффект переобучения.

Постановка исходных экспериментов.

- Обучим нейронную сеть классифицировать сумму чисел по модулю простого числа - сопоставить $z = x + y \pmod p$

- Второй пример эксперимента обучение 1L трансформера по модульному сложение цифр из 5-ти знаковых чисел (11111 + 13579 = 24680) каждая цифра складывается по модулю 10

Наблюдение.

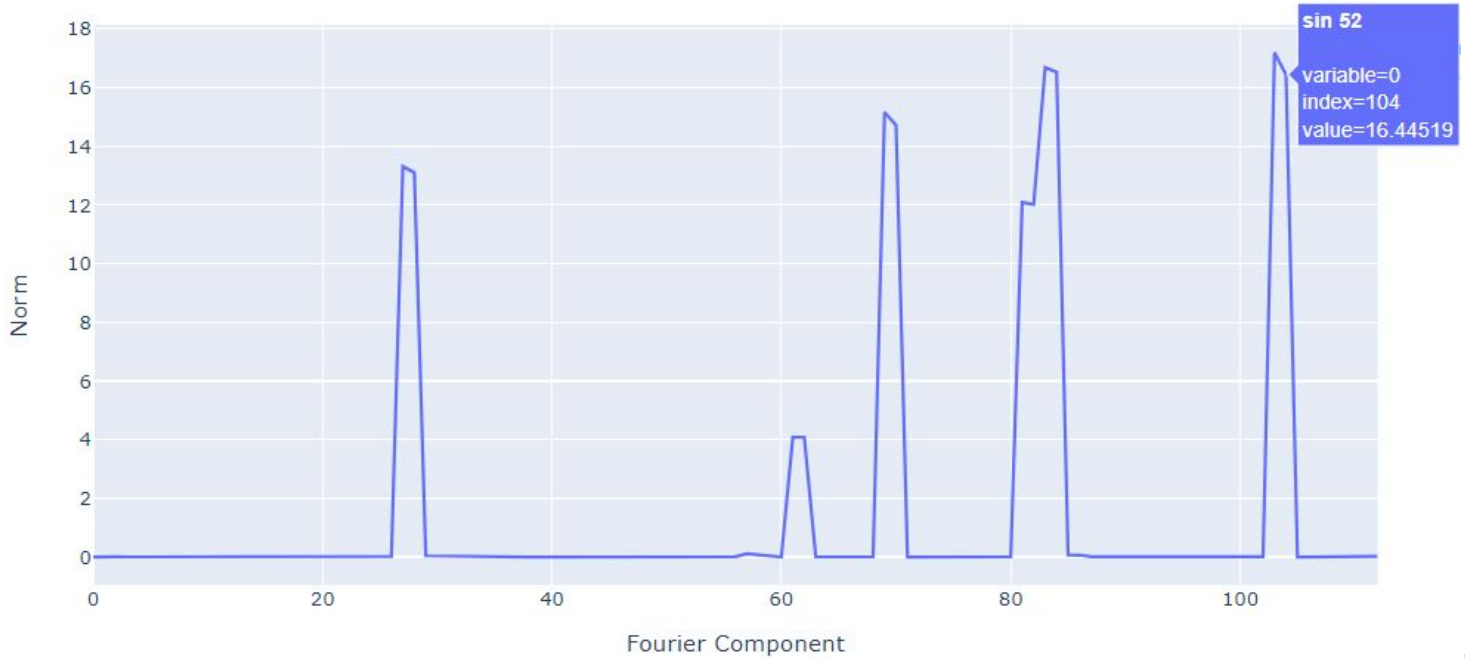
Как видно из графика справа - при небольшом количестве эпох уже виден эффект переобучения.

Но далее при более продолжении обучении ошибка резко уменьшается и на тесте.

Этот эффект зовется **Grokking** - глубокое(интуитивное) понимание.

Описание эффекта.

Промежуточно вычисляет cos(wx), sin(wx) для некоторых фаз,
а также вычисляет их произведения и суммы (для преобразования фурье).
Для сложения по модулю 113 используются частоты: 14,35,41,42,52



нормы вложения
для каждой частоты

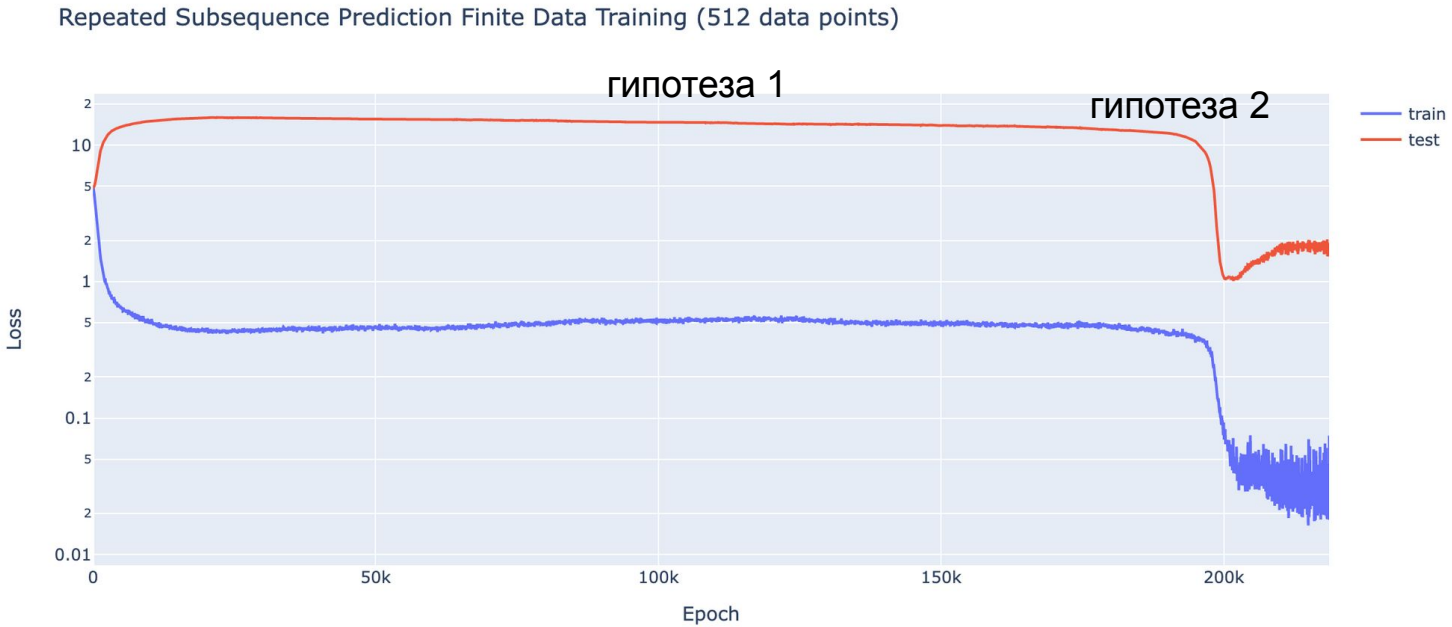
Гипотезы возникновения гроккинга

1. Главная гипотеза: полный выходной сигнал сети - это среднее значение для многих разных схем. Некоторые из этих схем систематически полезны для снижения потерь, а большинство - нет. Градиентный спуск усилит полезные схемы и подавит бесполезные, поэтому модель будет постепенно формироваться.

2. Случайное блуждание - через некоторое количество эпох случайно натыкаемся на сформированный кластер являющийся частью обобщающего решения, а затем градиентный спуск берет верх и формирует решение характеризующееся резким спадом потерь на тестовой выборке.

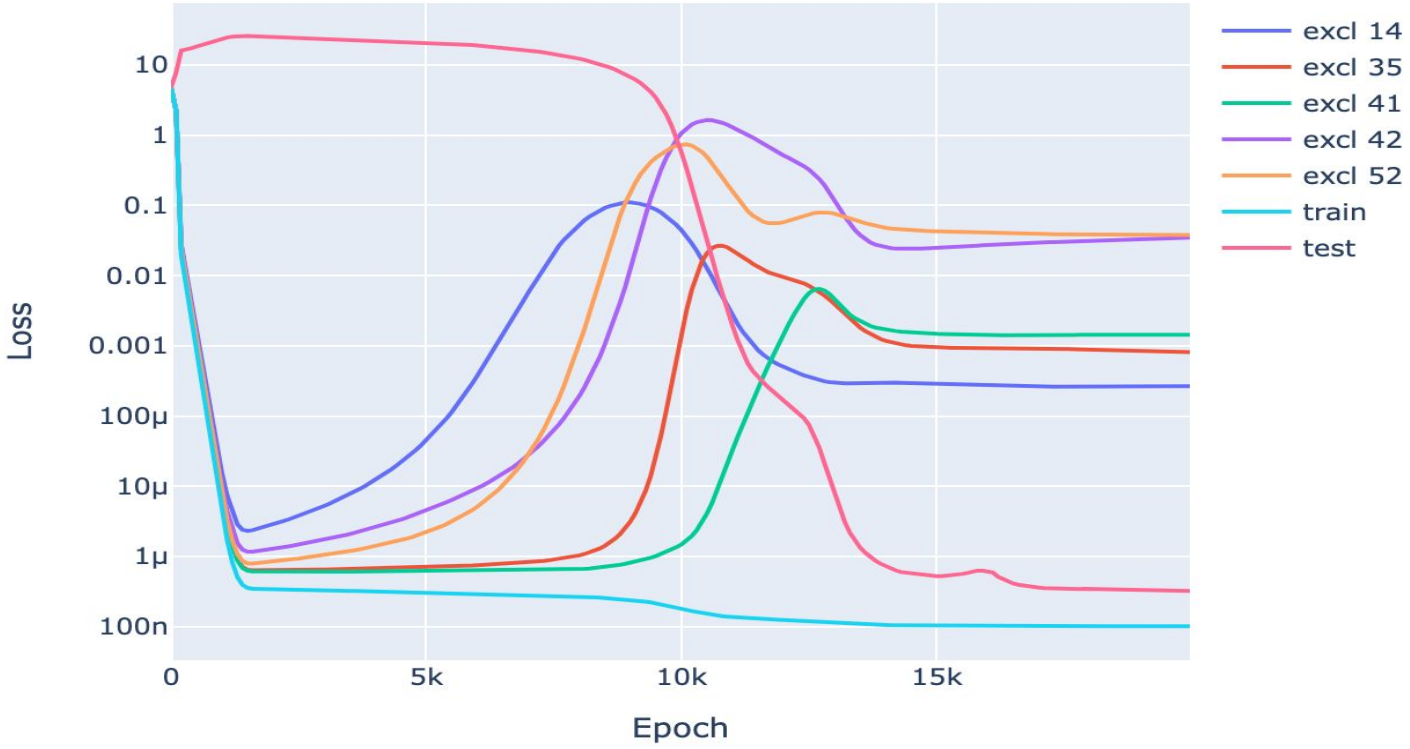
Гипотезы возникновения гроккинга

Основное отличие этих гипотез в том когда именно начинает зарождаться обобщающее решение.



Гипотезы возникновения гроккинга

Аргумент против гипотезы случайного блуждания -
на данном графике мы намеренно удаляем логиты соответствующие вычисленным $\cos(w(x+y))$, $\sin(w(x+y))$для конкретных w и рассматриваем значение потерь.



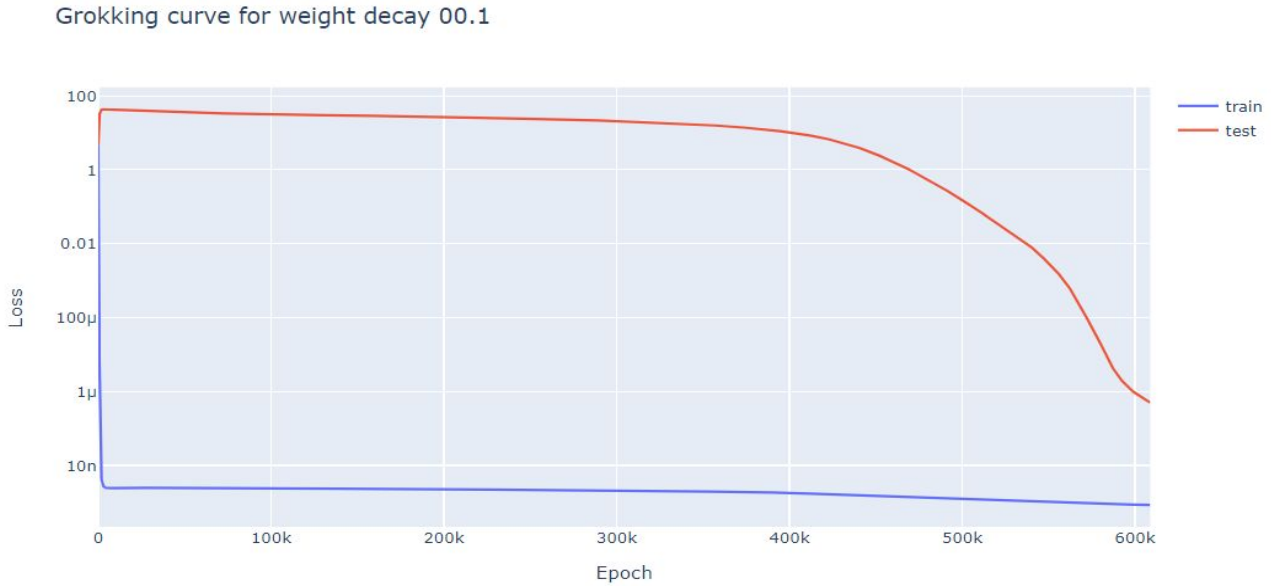
Гипотезы возникновения гроккинга

Вывод: формирование структуры соответствующей “глубокому пониманию” начинается задолго до того как это графически отразится на тестовой выборке.

Влияние задачи на гроккинг

Регуляризация: значительная - не находит лучшее решение, жесткая регуляризация не дает лучше подстроиться под данные. Умеренная - проявляется эффект гроккинга.

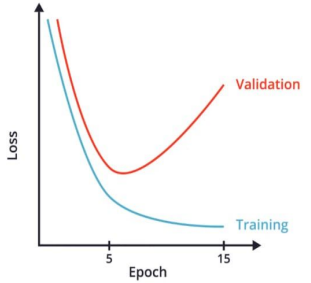
Интуиция эффекта: вначале модель предпочитает запоминающее решение, так как то более легкодостижимо, но затем регуляризация все же заставляет предпочесть обобщающее решение (более труднодоступное).



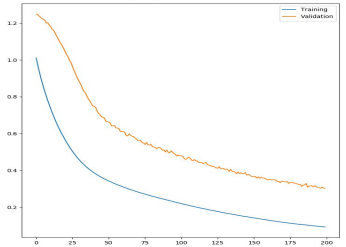
Влияние задачи на гроккинг

Размер входных данных
Эффекта гроккинга - не проявлялся при некотором размере входных данных, а именно:

1. если входных данных слишком мало - модель запоминает.

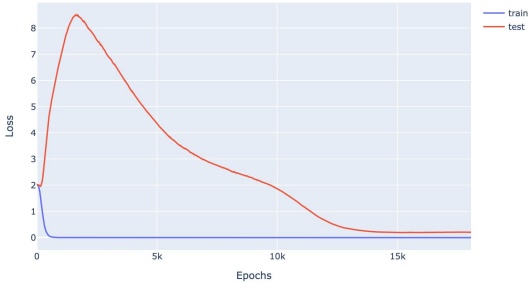


2. если данных слишком много - модель легко обобщает.



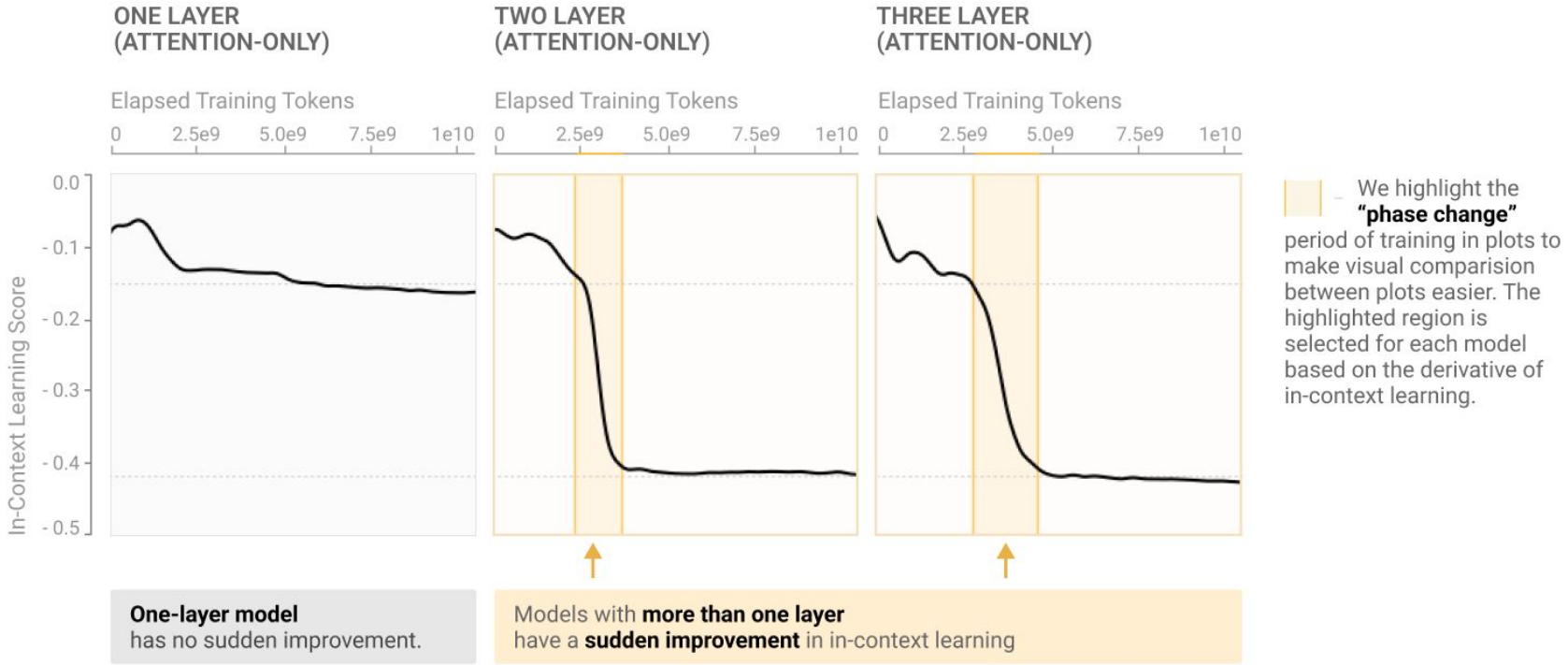
Ручным бинарным для 5-ти знакового сложения - при тестовой выборки около 700 элементов появляется эффект гроккинга.

Интуиция эффекта: гроккинг является эффектом на перес между запоминанием и обобщением.



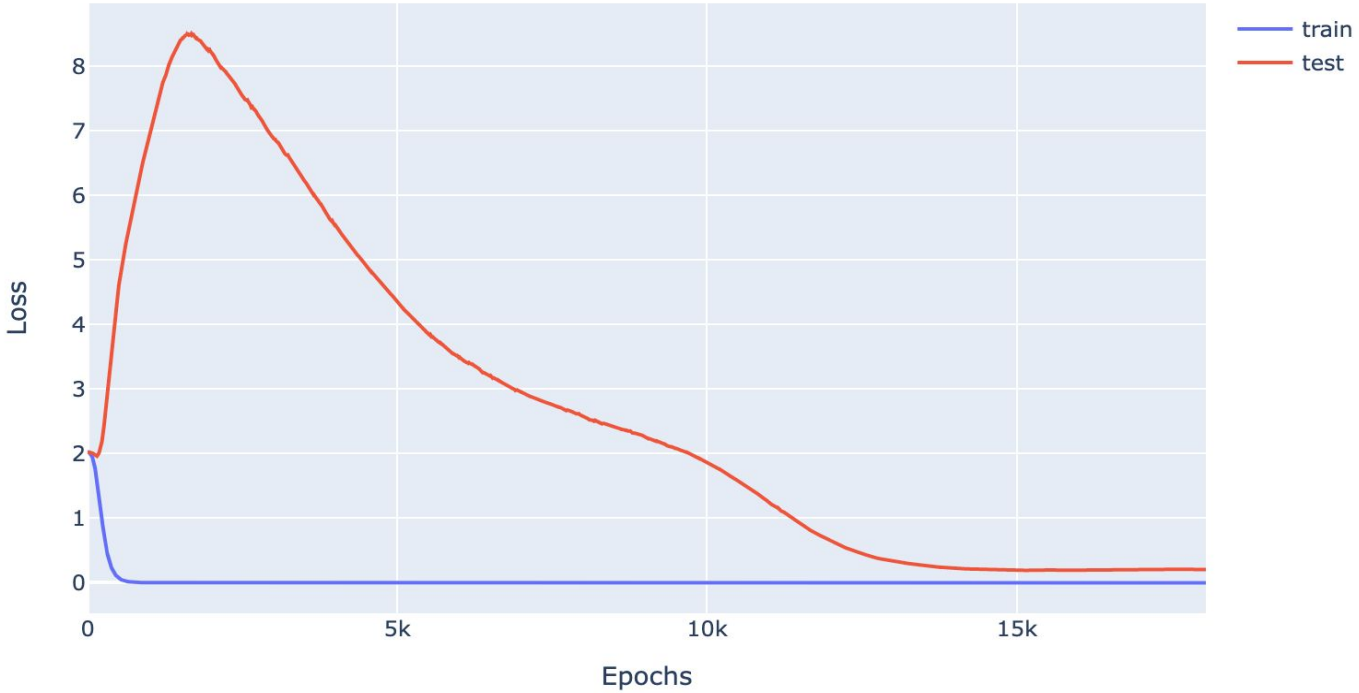
Фазовые переходы.

Рассмотрим эффект возникающий в обучении нейросетей с несколькими слоями, называемый фазовым переходом.
Плато или ухудшение качества-> производительность быстро улучшается, качество резко повышается -> качество выравнивается.



Фазовые переходы.

Рассмотрим подробнее задачу 5-ти знакового сложения, преимущество в огромном размере различных входных данных.



Фазовые переходы.

Предположение: фазовые изменения присуще композиции, возникают они при возникновении схем внутри общей структуры.

Гроккинг также является фазовым переходом.
Изучение эффекта гроккинга сводится к изучению фазовых переходов.

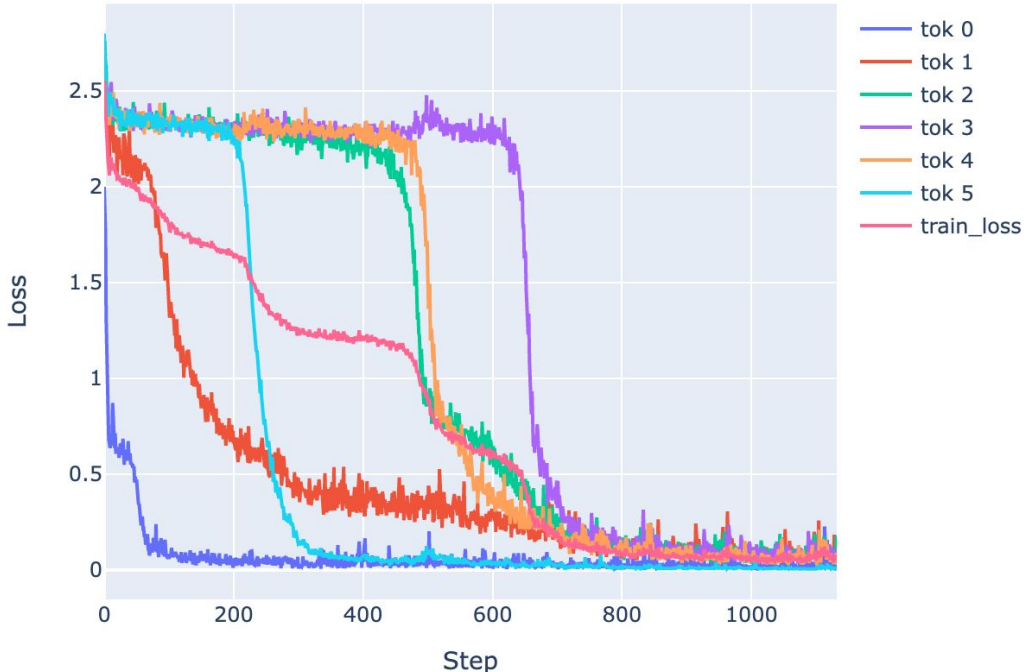
Гипотеза интерпретируемости: большинство вещей которые делают модели построены из интерпретируемых схем.

Основная гипотеза гроккинга - интерпретируемые схемы
Скачок гроккинга - проявление фазового перехода.

Гладкие кривые потерь - результат множества небольших фазовых переходов.

Фазовые переходы.

График получаемый в задаче о 5-ти значном сложении. На графике отмечены потери по отдельным разрядам (tok), а также общая кривая потерь. График сделан в случае большого размера трэйна - для рассмотрения фазовых переходов другого рода.



Характер части спуска не поменялся, на большой выборке отчетливее заметны результаты фазовых переходов по отдельным разрядам.

Выводы

- Гроккинг - эффект возникающий после переобучения, тогда когда обобщающее решение хуже запоминающего.
  Характеризуется тем, что после продолжения обучения модель начинает иметь более высокую обобщающую способность.
- Эффект гроккинга является переходным между запоминанием и обобщением, таким образом и появляется зависимость от размера данных и регуляризации.
- Вопросы гроккинга вероятно можно свести к вопросам фазовых переходов и их связи с интерпретируемыми схемами, а сам гроккинг будет являться таким фазовым переходом.

Список источников

- Изначальная статья google, openAI https://arxiv.org/pdf/2201.02177.pdf
- Основная статья с попыткми продвижений в понимании эффекта и обратным инжинирингом: https://www.alignmentforum.org/posts/N6WM6hs7RQMKDhYjB/a-mechanistic-interpretability-analysis-of-grokking
- Другая статья, где проявляются похожие эффекты но с альтернативным обьянением https://openreview.net/pdf?id=zDiHoIWa0q1
- статья с фазовыми переходами https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html