On the loss landscape in grokking: mode connectivity

A Preprint

Грозный С.В
Кафедра ММП, факультет ВМК
МГУ им. М.В. Ломоносова
groznysv@my.msu.ru

Южаков Т. А.
ФКН, НИУ ВШЭ
Исследовательская группа Байесовских методов

Ветров Д.П.
кандидат ф.-м. наук
Профессор НИУ ВШЭ
Исследовательская группа Байесовских методов

Abstract

В данной работе исследован эффект гроккинга в глубоком обучении, представляющий собой интересную тенденцию, при которой нейронные сети достигают 100% точности на обучающей выборке, сохраняя при этом случайный уровень точности на отложенной выборке, а затем неожиданно достигают 100% точности на ней. Было предложено анализировать ландшафт функции потерь, используя пути, соединяющие оптимумы, для объяснения этого явления. В результате работы было получено, что сложная структура ландшафта функции потерь в гроккинге объясняется блужданием по многообразию функции потерь с нулевой ошибкой, начиная с узкого минимума и переходя к более широкому, что способствует улучшению обобщающей способности модели. Кроме того, было показано, что использование определенных кривых для соединения оптимумов может существенно снизить ошибку на отложенной выборке. Эти результаты могут иметь важное значение для оптимизации и эффективного обучения глубоких нейронных сетей.

1 Введение

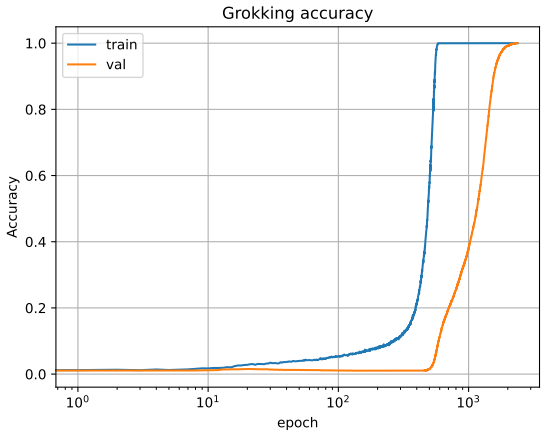
Глубокие нейронные сети стали мощным инструментом для обработки сложных данных и решения разнообразных задач: от классификации изображений до генерации текста. Однако, оптимизация и эффективное обучение таких моделей остаются актуальными проблемами.

Также, в процессе развития глубокого обучения обнаруживают всё больше и больше необычных явлений, которые трудно объяснить. Сложность заключается в том, что модели в последнее время имеют миллионы, а то и миллиарды параметров, и поэтому высокая размерность функции потерь влечет за собой трудности в ее интерпретации. Понимание геометрических свойств поверхности функции потерь может помочь в поиске причин возникновения данных явлений и построении более эффективных процедур обучения.

В данной работе исследование ландшафта функции потерь будем проводить на примере т.н. эффекта гроккинга[4]. Его суть заключается в том, что даже при достижении нулевой ошибки на обучении в сети продолжают происходить процессы упорядочивания информации. Попробуем объяснить это через анализ ландшафта функции потерь с помощью построения путей, связывающих оптимумы.

2 Эффект гроккинга

Гроккинг - это феномен неожиданного обобщения в нейронных сетях, обученных на определенных алгоритмических задачах. При гроккинге график точности (и потерь) сети показывает две фазы (Рис. 1). В начале обучения точность обучения достигает 100%, в то время как точность на отложенной выборке остается близкой к случайной (т.е. запоминание обучающей выборки). Значительно позже в процессе обучения точность обобщения внезапно подскакивает до 100%. Отметим, что если обучать модель без коэффициента регуляризации, то эффекта гроккинга не будет.



★	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

Рис. 1: Слева: Точность: однослойный decoder-only трансформер, оптимизатор SGD с параметрами lr=0.1, weight decay=0.001. Справа: Сгенерированные данные, поделенные на train и test(знак ?).

2.1 Описание данных

Этот эффект можно пронаблюдать на данных следующего вида: $a \circ b = c$, где a, b, c - числа, а $\circ$ - некая бинарная операция. Составляется таблица (Рис. 1), где строки и столбцы это всевозможные значения a и b, в ячейках которой хранятся соответствующие этим a и b - c. Далее, случайным образом стираются

некоторые ячейки(то есть разбиваем выборку на train и test(пустые ячейки)). Задача состоит в том, чтобы заполнить пустые ячейки в соответствии с выше описанной операцией.

Для экспериментов в качестве бинарной операции возьмем умножение по модулю 97 : $x * y \pmod{97}$. Пусть $x \in [\,0, 96\,]$, $y \in [\,1, 96]$, а результат операции, очевидно, лежит на $[\,0, 96\,]$.

2.2 Описание модели и метода оптимизации

Авторы оригинальной статьи разбивают $a \circ b = c$ на 5 токенов: $a, \circ, b, =, c$. Для экспериментов будет достаточно только три токена: a, b, c, так как это обеспечивает более стабильное обучение.

Сопоставим каждому числу от 0 до 96 соответствующий вектор эмбеддинга размерности 128, а после применяем однослойный decoder-only трансформер без нормализации из статьи omnigrok[3]. В качестве стратегии оптимизации будем использовать стохастический градиентный спуск с темпом обучения 0.01, коэффициентом регуляризации 0.001 и размером батча 512.

3 Связность оптимумов нейронных сетей

Как известно [1], глобальные оптимумы, в которых достигается нулевая ошибка на обучении, образуют многообразие нулевого трейн лосса. Попробуем построить путь, соединяющий эти оптимумы и исследуем поведение функции потерь по мере его прохождения. Для начала введем несколько определений.

Опр. 1 Пусть w_1, $w_2 \in \mathbb{R}^n$. Отображение вида

$$\phi(t) : [\,0, 1\,] \to \mathbb{R}^n$$

такое, что $\phi(0) = w_1, \phi(1) = w_2$, будем называть путем, соединяющим точки w_1 и w_2.

Самый простой путь, который можно построить - это отрезок. Определим его в контексте нашей задачи.

Опр. 2 Пусть w_1, $w_2 \in \mathbb{R}^n$. Функцию вида

$$\phi(t) = tw_1 + (1 - t)w_2,\ t \in [0, 1]$$

будем называть отрезком, соединяющим точки w_1 и w_2.

Также, попробуем соединить оптимумы такой кривой, чтобы значение функции потерь было минимально.
Метод поиска этой кривой между парой оптимумов интуитивно очень прост: параметризуем путь и минимизируем средние потери на этом пути относительно его параметров. Точнее, минимизируем

$$L(\theta) = \int_0^1 Loss(\phi_\theta(t))\, dt = E_{t\sim U[0,1]} Loss(t)$$

по θ, где $Loss$ - функция потерь.

Также есть способ эффективно вычислять стохастические градиенты $L(\theta)$ по θ. Для этого выбираем точку t из равномерного распределения на отрезке $[0, 1]$, а зачем вычисляем градиент $Loss(\theta)$ по θ, используя правило цепочки:

$$\frac{\partial Loss(\phi(t))}{\partial \theta} = \frac{\partial Loss(\phi(t))}{\partial \phi(t)} \frac{\partial \phi(t)}{\partial t}$$

Используя эту стохастическую оценку градиента, происходит минимизация $L(\theta)$ с помощью стандартного SGD.

В экспериментах в качестве параметризованной кривой будем использовать ломаную с одним изгибом.

Опр. 3 Пусть w_1, w_2, $\theta \in \mathbb{R}^n$. Функцию вида

$$\phi_\theta(t) = \begin{cases} 2\left(t\theta + (0.5 - t)w_1\right), & t \in [0, 0.5] \\ 2\left((t - 0.5)w_2 + (1 - t)\theta\right), & t \in [0.5, 1] \end{cases}$$

будем называть ломаной с одним изгибом.

Исследования показывают [2], что кривизна поверхности функции потерь значительно скоррелирована с обобщающей способностью. Поэтому будем также считать норму стохастичского градиента.

4 Эксперименты

Пусть w_1 и w_2 - веса модели, при которых модель впервые достигла 100% на обучающей и валидационной выборке соответственно. Если смотреть на график 1, то это соответствует моменту запоминания (будем называть это точкой 1) и моменту гроккинга (будем называть это точкой 2).

Попробуем соединить эти точки различными способами и посмотрим на поведение графиков функции потерь и ошибки (один минус точность). Также будем считать среднюю норму стохастического градиента, т.е. вычислять все минибатч градиенты по выборке, считаем их норму и после вычисляем среднее данной статистики.

4.1 Исследование линейной связности оптимумов

Посмотрим, как ведут себя вышеперечисленные метрики, если соединить точку 1 и 2 с помощью отрезка:

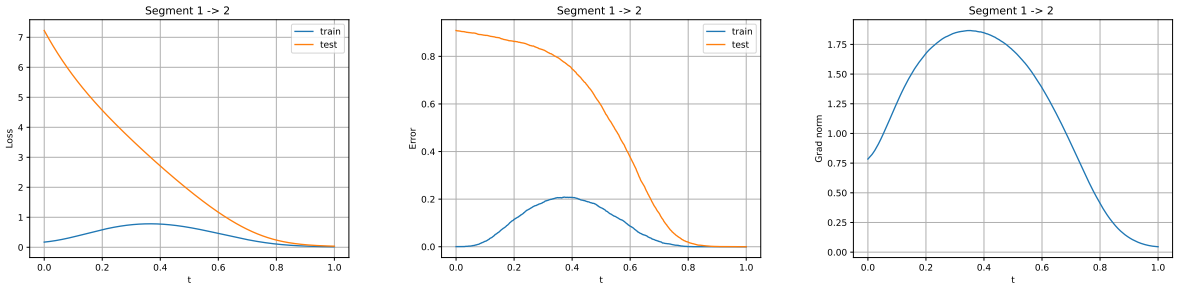


Рис. 2: Кросс-энтропия с l_2 регуляризацией (слева), ошибка (центр) и норма стох. града (справа) как функции от точек на кривой $\phi(t)$ - отрезок.

Как и ожидалось, модель показывает просадку точности на обучающей выборке(пик при $t = 0.35$). Отметим, что несмотря на это, график на отложенной выборке монотонно падает. Норма градиента перетекает из 0.75 к значению близкому к нулю.

Можно предположить, что происходит следующее: модель выскакивает из узкого минимума и, двигаясь по сложному рельефу функции потерь, попадает в широкий минимум.

4.2 Исследование нелинейной связности оптимумов

4.2.1 Соединение последовательными отрезками

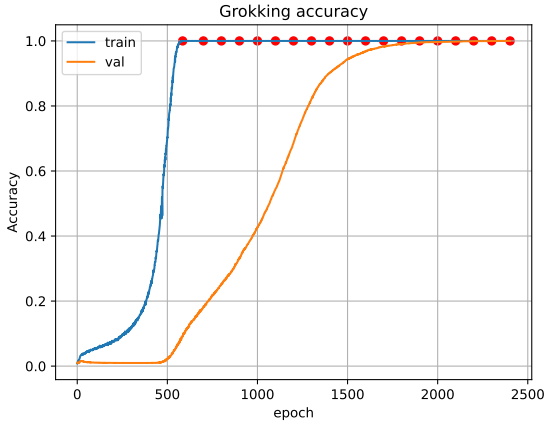


Рис. 3: Однослойный decoder-only трансформер, оптимизатор SGD с параметрами lr=0.1, weight decay=0.001, логирование весов модели каждые сто эпох (красные круги).

В данном эксперименте построим путь от точки 1 до точки 2 с помощью последовательно соединенных отрезков. Для этого фиксируем веса каждые 100 эпох и последовательно соединяем. Продемонстрируем графики только для первого и последнего отрезка:

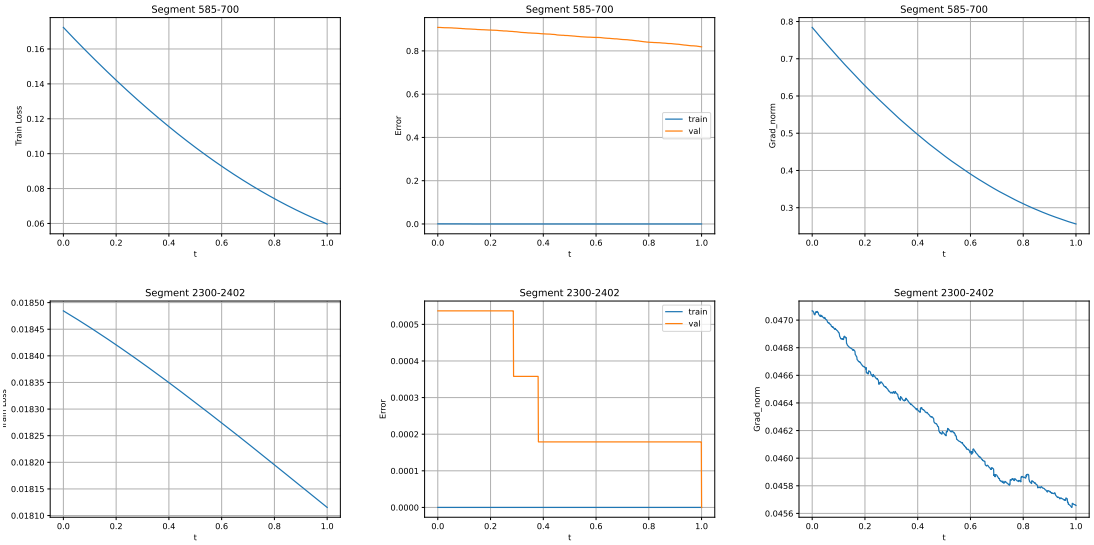


Рис. 4: 1 ряд: отрезок, соединяющий 585 и 700 эпоху. 2 ряд: отрезок, соединяющий 2300 и 2402 эпоху. 585 и 2402 эпохи соответствуют 100% точности на обучающей, низкой на отложенной и 100% на отложенной, 100% на обучающей выборке соответственно.

Как можно заметить, на первом отрезке происходит существенное изменение функции потерь и нормы градиента, а на последнем изменения незначительные. Удивительно, что на протяжении всего пути ошибка на обучающей выборке нулевая.

Таким образом, удалось соединить два оптимума последовательными отрезками с констатно нулевой ошибкой. Из этого можно сделать вывод, что после запоминания обучающей выборки происходит дрейфование по многообразию нулевого трейн лосса.

4.2.2 Соединение ломаной с одним изгибом

Теперь попробуем соединить оптимумы с помощью одной параметризованной кривой, а именно ломаной с одним изгибом. Причем, соединять оптимумы будем не только соответствующие точкам 1 и 2, но и полученных при различных начальных инициализациях. Далее за точки 1 и 1' будем обозначать момент первого достижения 100% точности на обучающей выборке при разных инициализациях соответственно. Аналогично определяются точки 2 и 2', как точки первого достижения 100% на валидации.

Ранее удалось соединить точки 1 и 2 несколькими отрезками с нулевой ошибкой. Посмотрим, как ведут себя метрики на кривой, подобранной с помощью оптимизиации. По графику (Рис. 5) ошибки на обучающей выборке можно убедиться, что оптимизация параметра θ произошла успешна. Также по графику нормы градиента видно, как из узкого минимума переходим в широкий.

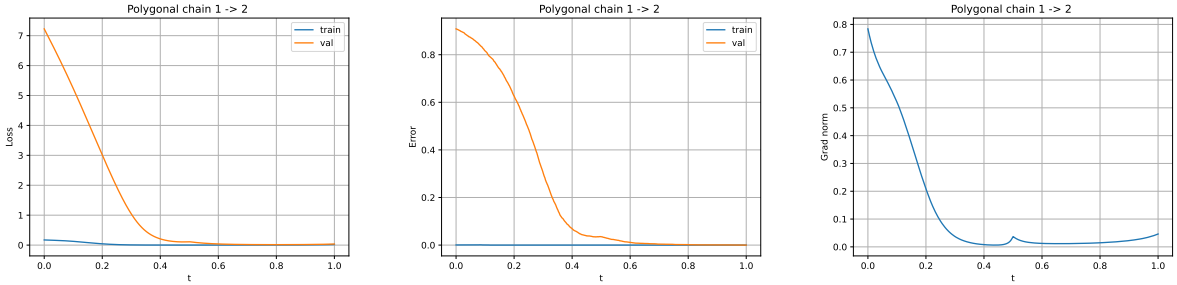


Рис. 5: Кросс-энтропия с l_2 регуляризацией (слева), ошибка (центр) и норма стох. града (справа) как функции от точек на кривой $\phi_\theta(t)$ - ломаная с одним изгибом.

Заметим, что по сравнению с отрезком (Рис. 2), выйти на плато хорошего качества на валидации удалось значительно раньше: с момента $t = 0.6$.

Теперь попробуем построить путь через оптимумы «запоминания» обучающей выборки, полученные через разные инициализации. Оптимизируем кривую по параметру θ и считаем метрики:

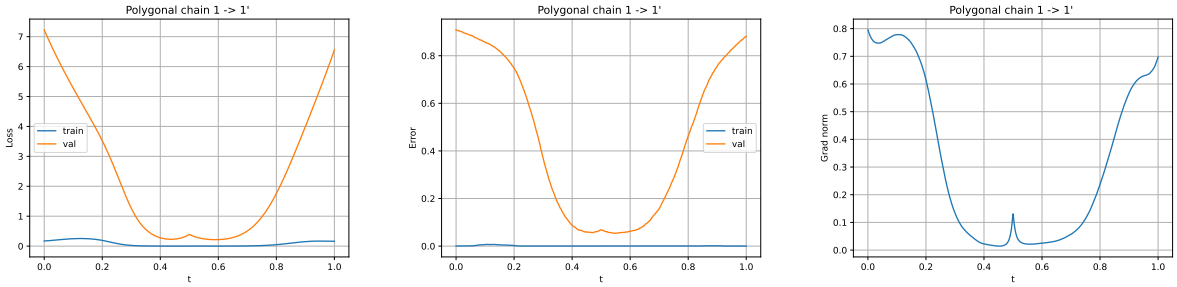


Рис. 6: Кросс-энтропия с l_2 регуляризацией (слева), ошибка (центр) и норма стох. града (справа) как функции от точек на кривой $\phi(t)$. 1 и 1' - точки 100% точности на обуч. выборке при разных инициализациях.

Точность на обучающей выборке показывает, что оптимизация кривой по параметру удалась. Также, наблюдается довольно необычный эффект: при соединении оптимумов с высокой валидационной ошибкой по пути встречается довольно широкий минимум с хорошим обобщением (это особенно заметно в окрестности точки $t = 0.5$).

Проведем аналогичный эксперимент, но для точек, соответствующих 100% точности на валидации:

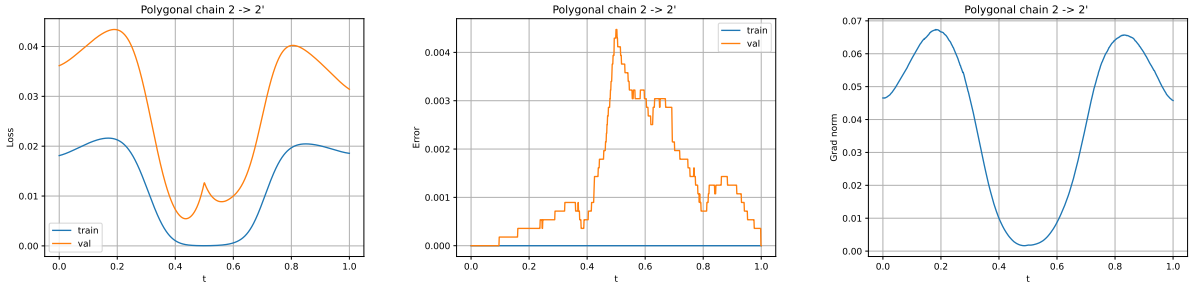


Рис. 7: Кросс-энтропия с l_2 регуляризацией (слева), ошибка (центр) и норма стох. града (справа) как функции от точек на кривой $\phi(t)$. 2 и 2' - точки 100% точности на вал. выборке при разных инициализациях.

Первое, что бросается в глаза, это довольно маленькое изменение значений метрик на кривой. Из этого можно сделать вывод, что два оптимума с хорошим обобщением в задаче гроккинга соединимы с низкой ошибкой на валидации. Также отметим, что график нормы градиента содержит в себе два симметричных холма. Можно сделать предположение, что, двигаясь по кривой, модель сначала выскакивает из минимума, а после в него же и возвращается.

5 Выводы

В данной работе было показано, что ландшафт функции потерь в задаче гроккинга имеет сложную структуру: при соединении двух оптимумов одним отрезком функция потерь показывает просадку в точности. Однако если использовать более сложные кривые, то построение пути с низкой ошибкой на обучающей выборке не составляет труда.

Также, по построению нескольких отрезков, соединяющих оптимумы, можно сделать вывод, что гроккинг возникает в силу блуждания по многообразию функции потерь с нулевой ошибкой. Сначала модель скатывается в узкий минимум, а после переходит в более широкий, где лучше обобщающая способность.

Кроме того, с помощью построения ломаной с одним изгибом удалось достичь довольно низкую ошибку на отложенной выборке, стартуя из точки с низкой валидацией. Если данный эффект удастся повторить на других задачах, то данную оптимизацию можно использовать как менее вычислительно затратную.

Список литературы

[1] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns, 2018.

[2] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima, 2017.

[3] Ziming Liu, Eric J. Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data, 2023.

[4] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022.