Название статьи (авторы статьи): Grokking: Generalisation beyond Overfitting Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, Vedant Misra

Автор обзора-рецензии: Раевская Алеся

При проведении исследования, мы предлагаем ответить вам на следующие вопросы:
1. Опишите суть работы в паре предложений, выделите ее основной вклад.
Статья посвящена эффекту “grokking” - неожиданное улучшение результата на валидационной выборке уже после обучения. Эксперименты проводились на небольших алгоритмических датасетах из бинарных операций (например, сложение по модулю) и стабильно показывают этот эффект. Сам эффект подробно не изучался, но было хорошо показано, что это не просто ошибка и он действительно существует.

2. Когда написана работа? Опубликована ли она на какой-то конференции? Кто ее авторы, есть ли у них другие схожие работы? Подумайте как авторы пришли к идее статьи -- может быть это прямое улучшение их предыдущей работы, может это похоже на случайную находку
Статья опубликована в мае 2021 на воркшопе конференции ICLR. Авторы работают в OpenAI. Эффект был обнаружен случайно, после того как они забыли трансформер на ночь, поэтому схожих работ у них нет. В основном они занимаются Language models, Code generation, Reinforcement learning.

3. Какие из статей в списке ссылок (или почему-то не из списка, hello Mr. Schmidhuber) оказали наибольшее влияние на данную работу? Можно ли выделить какие-то 1-3 статьи, которые можно назвать базовыми для этой работы? Опишите в чем связь с этими работами (без математики, просто суть).
Так как эффект был обнаружен случайно, то у этой статьи нет предшественников

4. Кто цитирует данную статью? Есть ли у этой работы прямые продолжения, которые стоит прочесть тем, кто заинтересовался этой работой?

Часто статья цитируется в качестве доказательства непредсказуемости поведения нейронных сетей, но есть и несколько прямых продолжений.

- Towards Understanding Grokking: An Effective Theory of Representation Learning (Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, Mike Williams)
    - Изучает эффект для бинарных операций, так же как и в оригинале, но предлагается математическая теория, которая предсказывает необходимый размер обучающего датасета и стадиями обучения
- A Mechanistic Interpretability Analysis of Grokking (Neel Nanda, Tom Lieberum)
    - Теория возникновения grokking с помощью reverse-engineering, предсказание через косинусы (периодичность возникает из-за модуля)
- Omnigrok: Grokking Beyond Algorithmic Data (Ziming Liu, Eric J. Michaud, Max Tegmark)
    - Grokking изучается с помощью ландшафта потерь
    - Если инициализировать большими весами, то спуск будет медленный и будет возникать grokking

- Нашли grokking на других моделях (графх, текстах), с помощью инициализации большими весами

5. Есть ли у работы прямые конкуренты (которые, например, вышли одновременно с работой или еще по каким-то причинам не вошли в предыдущие два пункта)? Опишите как соотносится данная работа с этими конкурентами (без математики, просто суть).
Конкурентов нет, но есть другие работы изучающие эффекты нейронных сетей (Double Descent, например)

6. Опишите сильные, на ваш взгляд, стороны работы. Стоит обратить внимание на корректность утверждений в работе, значимость и новизну вклада, актуальность для исследовательского сообщества, понятность текста и воспроизводимость результатов (чтобы хорошо это сделать полезно набраться контекста из предыдущих пунктов: изучить prior work, продолжения, может даже посмотреть код/поговорить с *хакером*).
Сильные стороны работы состоят в открытии нового явления и тщательной проверке того, что явление действительно возникает. По сути целью работы являлось подтвердить существование эффекта и пригласить других к дальнейшему изучению, что и получилось

7. Опишите слабые, на ваш взгляд, стороны работы, обращая внимание на те же моменты, что и в предыдущем пункте.
Статья небольшая, поэтому многие моменты опущены. Проверялось только на одном небольшом типе данных. А также, непонятно, как это применять на практике

8. Предложите как можно было бы улучшить статью: какие дополнительные утверждения/эксперименты стоило бы рассмотреть, какие вопросы остались не закрытыми для вас после прочтения статьи, обсуждение связи с какими работами дополнило бы работу.
Улучшения статьи проводились в последующих работах. Был найден эффект на других моделях, были выдвинуты теория из-за чего он возникает. И хоть, изначальная статья небольшая и неполная, но в последующих статьях основные пробелы были закрыты.

9. Попробуйте на основе результатов статьи предложить исследование, не проведенное к текущему моменту, или идею применение в индустриальных приложениях.
В изначальной статье упоминалась бинарная функция, для которой не был обнаружен эффект. Было бы интересно посмотреть, с чем это связано. Со сложностью функции или с чем-то еще.