Название статьи (авторы статьи): Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets (Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, Vedant Misra)
Автор обзора-рецензии: Владимир Княжевский

1. Опишите суть работы в паре предложений, выделите ее основной вклад.

В работе рассматривают эффект под названием grokking, который состоит в неожиданном улучшении качества на валидационной выборке через значительное число эпох после переобучения. Эксперименты на небольших датасетах, модулирующих простые бинарные формулы стабильно показывают этот эффект.

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.

Статья опубликована в мае 2021 на известной конференции ICLR в секции 1st Mathematical Reasoning in General Artificial Intelligence Workshop. Авторы работают в OpenAI. Найти каких-то схожих работ у них не удалось, в основном они занимаюся Code generation, Reinforcemnet learning и Language models. Идея статьи пришла к ним случайно: они изучали работу трансформеров на алгоритмических задачах, оставили трансформер работать на ночь, и он продемонстрировал к утру описанный эффект.

3. Какие из статей в списке ссылок 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).

Поскольку находка случайная, то говорить о каких-то серьезных связях с предшествующими работами трудно. Схожая ситуация (с двоцным спуском) исследуется в работах: Deep double descent: Where bigger models and more data hurt; Triple descent and the two kinds of overfitting: Where & why do they appear?

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

Довольно часто статью цитируют в качестве иллюстрации неких утверждений о том, что поведение нейросетей может быть для нас неожиданным и важно хорошо понимать, почему оно возникает. Например, в статье Predictability and Surprise in Large Generative Models ссылка на статью про grokking является фактически иллюстрацией к фразе: "Though performance is predictable at a general level, performance on a specific task can sometimes emerge quite unpredictably and abruptly at scale" (речь идет о large generative models). Или, скажем, в статье Unsolved Problems in ML Safety есть слова: "We are better able to make models safe when we know what capabilities they possess", и далее описываются различные примеры того, когда эти возможности оказывались нетривиальными, в том числе и grokking.
У статьи есть и несколько прямых продолжений.

Статья Towards Understanding Grokking: An Effective Theory of Representation Learning также изучает grokking для бинарных операций, но на примере совсем маленькой модели. При этом используются результаты из физики, и предлагается теория, которая предсказывает необходимый размер обучающего датасета и траектории обучения.

В статье A Mechanistic Interpretability Analysis of Grokking предлагается похожая на предыдущую статью теория, объясняющая grokking, а также проводится reverse ingeneering (т.е. выясняется что именно выучивает нейросеть) для одной из бинарных операций - сложения по модулю.

5. Есть ли у работы прямые конкуренты? Опишите как соотносится данная работа с этими конкурентами.

В общем-то конкурентов нет, если не считать таковыми все работы, изучающие обобщение нейросетей.

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.

В работе, на мой взгляд, все очень доходчиво объяснено. Эксперименты описаны подробно, всегда понятно, что именно происходит. Наибольший интерес представляет само открытие, что grokking происходит стабильно для определенных данных и моделей – в предшествующих работах double descent не происходил после переобучения. Значимость работы для теоретического понимания того, как устроено обучение нейросетей высока, поскольку описанное в статье поведение очень необычно, и редко встречается.

7. Опишите слабые, на ваш взгляд, стороны работы, обращая внимание на те же моменты, что и в предыдущем пункте?

Неясно, получится ли найти у статьи какое-то прямое применение на практике. Разрыв в accuracy сокращается при увеличении размера train sample, самая сложная из опробованных бинарных функций вообще не показала хороших результатов на validation – и реальная задача вполне может быть очень сложной и при этом иметь большую обучающую выборку.

Есть и мелкие недостатки: например, на рис. 3, справа непонятно, по какому принципу точки окрашены в цвета.

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.

Непонятно, почему не обучается одна из бинарных функций; было бы интересно понять, с чем это связано: просто надо еще пообучать или для нее в принципе нет grokking'a? Насколько сложной должна быть зависимость чтобы не было grokking'a?