

What Can Transformers Learn In-Context? A Case Study of Simple Function Classes

Shivam Garg, Dimitris Tsipras, Percy Liang, Gregory Valiant

Рецензия

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

Как следует из названия исследование посвящено in-context learning или обучение по контексту. Данное явление впервые заметили после создания текстовых моделей с огромным числом параметров (например, GPT-3). Суть контекстного обучения заключается в способности модели решать определенную задачу, основываясь на последовательности подсказок, состоящей из пар пример-ответ. Эти пары по сути и задают контекст, по которому "обучается" модель. Обучение здесь взято в кавычки не случайно, ведь речь идет только об inference шаге или шаге применения модели, когда отсутствуют вычисление градиентов и обновление параметров. Главной целью работы является изучение контекстного обучения моделей-трансформеров. Авторы вводят различные классы функций и смотрят, удастся ли моделям выучить эти зависимости по подсказкам, представляющим из себя последовательности пар (точка x_i , значение функции f_i из класса в точке x_i). Результаты контекстного обучения трансформеров на различных классах функций по сути и является основным научным вкладом статьи.

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 (пример случайного: grokking, пример последовательного улучшения: stylegan 1-2-3, score matching see <https://yang-song.net/publications>).

Работа опубликована на "архиве" 1 августа 2022 года. В авторах 4 человека – Shivam Garg, Dimitris Tsipras, Percy Liang, Gregory Valiant, при этом у всех в качестве аффилиации указан Стенфордский университет. Предлагаю остановиться на каждом исследователе чуть более подробно:

1. Shivam Garg -- аспирант в Стенфорде, член Machine Learning Group and the Theory Group of Stanford. Работает под руководством Gregory Valiant (4-го автора данной работы). Начинающий исследователь, у него около 40 цитирований в 2022 году. Соавтор 8 статей, причем все довольно разнородные, в том числе нет ни одной статьи, посвященной контекстному обучению.
2. Dimitris Tsipras -- пост-док CS-исследователь в Стенфорде, находящийся под менторством Percy Liang и Gregory Valiant (3-й и 4-й авторы). У него гораздо больше цитирований (4950 в 2022 году), в основном за счет статьи "Towards Deep Learning Models Resistant to Adversarial Attacks". Судя по статьям, в которых он участвует в качестве автора, можно сделать вывод, что его главным интересом являются атаки на модели и всяческие способы улучшения процесса обучения, например, Batch Normalization.
3. Percy Liang -- доцент компьютерных наук. Соавтор более, чем 250 статей. В 2022 году у него 9875 цитирований, большинство за счет соавторства в статье, где впервые был представлен датасет SQuAD.

Достоверно определить его интересы и выявить степень участия в различных работах довольно сложно, однако у него есть статья в соавторстве, которая также посвящена in-context learning. Она называется "An explanation of in-context learning as implicit bayesian inference".

4. Gregory Valiant -- доцент компьютерных наук. 674 цитирования в 2022 году. Также является соавтором большого количества статей (более 100). Среди его работ я не нашел статей, связанных с темой рассматриваемого исследования.

В качестве вывода из анализа авторов можно заметить, что с большой вероятностью данная работа не была логичным продолжением какого-либо другого исследования, в том числе потому, что только у одного из четырех авторов есть статья в области контекстного обучения.

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).

Можно выделить 4 связанных с работой области:

1. In-context learning
2. Transformers
3. Meta learning (learning to learn)
4. Data-driven algorithm design

В области контекстного обучения можно выделить следующие статьи:

- Ключевую роль здесь сыграла статья "Language Models are Few-Shot Learners", в которой было продемонстрировано, что большие модели могут хорошо работать с новыми задачами и данными, получая на вход подсказку. Данная статья стала отправной точкой области "in-context learning", поскольку именно в ней впервые был замечен эффект контекстного обучения.
- Вторая статья, которую можно выделить – это "An explanation of in-context learning as implicit bayesian inference" – Xie et al. 2022. В этой работе также все авторы из Стенфорда, причем один из них – общий доцент. В ней рассматривает контекстное обучение как скрытый байесовский вывод и предлагают фреймворк для него, объясняющий, как работает контекстное обучение.

В области трансформеров можно выделить следующие статьи:

- "Attention is all you need" – Vaswani et al. 2017 – базовая работа, в которой впервые была показана архитектура.
- "Transformers can do bayesian inference" – Müller et al. 2021 – тесно связанная статья, в которой показывают, что трансформеры могут аппроксимировать байесовский вывод. В ней вводят "Prior-data fitted transformer", обучают его аппроксимировать байесовский вывод с гауссовскими процессами и байесовскими нейронными сетями в качестве априорных распределений, а затем полученные знания используются для решения разных задач, например, классификации на табличных данных или few-shot классификации картинок. По сути эта работа очень похожа на TabPFN.

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

Если верить [Google Scholar](#), данная статья цитируется 7 раз. [Semantic Scholar](#) по тому же запросу приводит 17 цитирований. Среди главных продолжений необходимо выделить две статьи:

- "General-Purpose In-Context Learning by Meta-Learning Transformers", 8 декабря 2022 года – в ней авторы пытаются рассмотреть еще более широкий спектр задач, формулируя задачу контекстного обучения,

как обучение для обучения.

- “Why Can GPT Learn In-Context? Language Models Secretly Perform Gradient Descent as Meta-Optimizers”, 20 декабря 2022 года – новое исследование про большие лингвистические модели, которое рассматривает явление контекстного обучения как скрытый градиентный спуск в трансформерах.

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).

Я не смог выделить прямых конкурентов среди опубликованных на данный момент статей. Мне кажется, это довольно логично, поскольку область контекстного обучения довольно новая и других работ, изучающих это явление для моделей-трансформеров пока что нет.

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).

- Одним из больших плюсов, в том числе для воспроизводимости результатов, я считаю наличие исходного кода в открытом доступе. Для данной работы это условие выполняется, код многих экспериментов действительно присутствует в репозитории на гитхабе, однако в после ознакомления с авторским *ipynb*-ноутбуком, я не остался до конца доволен. На мой взгляд, там сильно не хватает структурированности кода, а также хотя бы минимальных комментариев к построенным графикам и проводимым экспериментам.
- Другой сильной стороной работы является основательность, с которой авторы подошли к исследованию и проведению экспериментов, по крайней мере, это точно можно заметить для класса линейных функций. Так, в работе отдельно исследуется зависимость от данных, на которых обучается модель, а также устойчивость контекстно обученных трансформеров к сдвигам распределений подсказок. Последнее по сути может трактоваться и как исследование *inductive biases* трансформер-архитектуры.
- Меня обрадовал тот факт, что у исследователей хватило сил на то, чтобы формализовать и достаточно математично описать свои изыскания (об этом можно судить по большому количеству введенных обозначений, большинство из которых в итоге углубило понимание проводимых экспериментов)

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

- Мне кажется, что можно было бы упростить восприятие темы исследования, которая будет новой для достаточной доли читателей, приведя несколько хороших и понятных примеров обучения и работы модели, которая обучается по контексту. Пусть формализация экспериментов и является хорошей чертой, мне все же показалось, что в некоторых местах она была избыточна. Было и ощущение того, что я слишком много раз перечитываю одну и ту же мысль, переформулированную по-разному, вместо этого мне бы хотелось видеть больше примеров. Также неплохо, на мой взгляд, смотрелось бы и сравнение обычного обучения и обучения по контексту с выделенными ключевыми отличиями.
- Также хочу отметить слабо структурированный демонстрационный *ipynb*-ноутбук с небольшим количеством комментариев.

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.

В качестве ответа на этот пункт я хотел бы, во-первых, предложить исправить указанные мной выше недостатки. Во-вторых, одной из целей работы являлось изучения влияния данных для обучения в широком смысле на возможность модели к контекстному обучению. На мой взгляд, этот вопрос можно было раскрыть более широко.


9. Попробуйте на основе результатов статьи предложить исследование, не проведенное к текущему моменту, или идею применения в промышленных приложениях.

Логичным предложением будет продолжение исследования явления контекстного обучения, поскольку авторы лишь немного продвинулись в понимании этого феномена, протестировав конкретную архитектуру на определенных классах функций. Одно из направлений -- попробовать завести "in-context learning", например, на рекуррентных моделях. Также до сих пор непонятно, за счет чего конкретно возникает этот эффект, поэтому доступна и обширная область для теоретических изысканий.

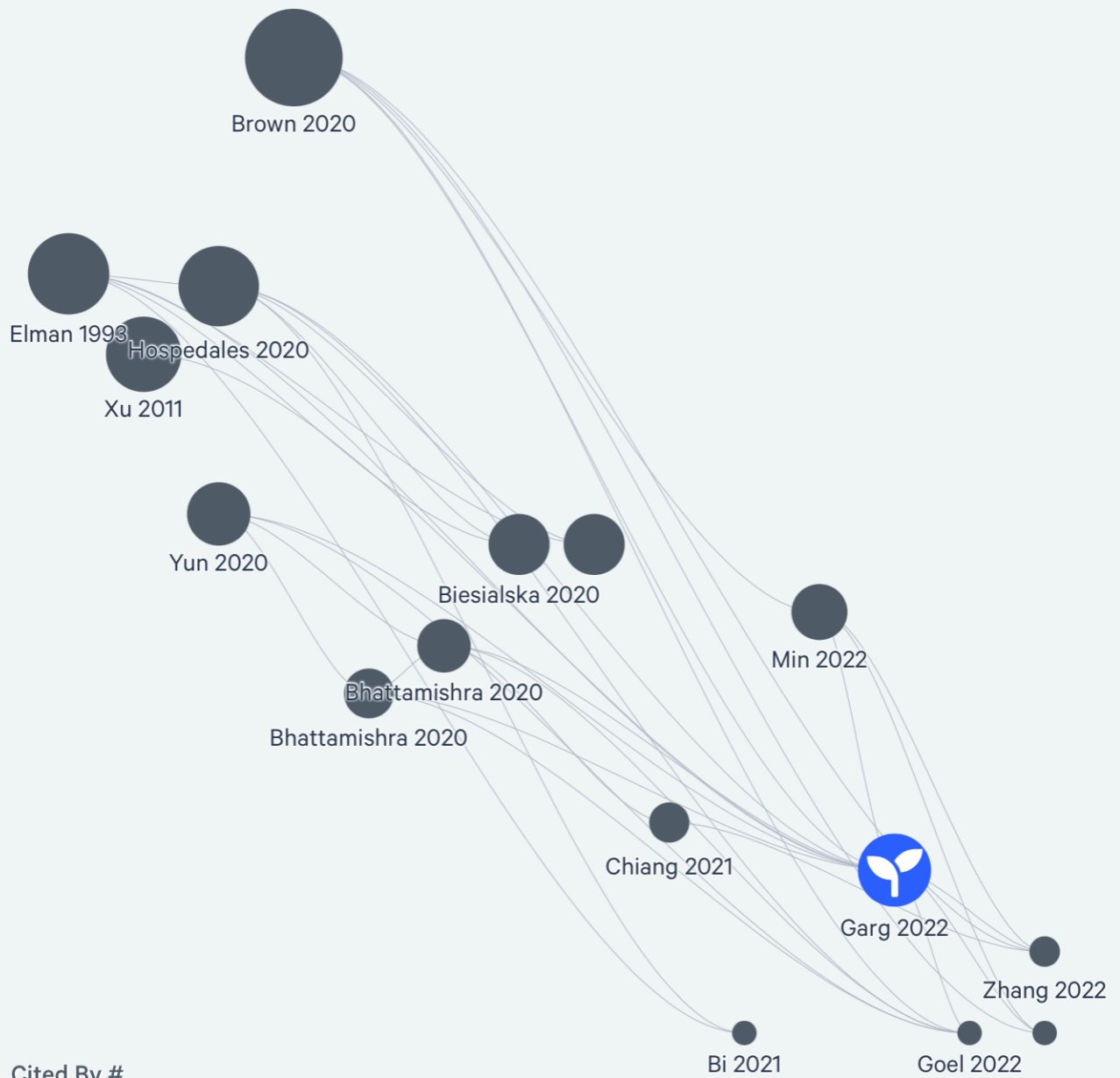
10. Может вы просто еще что-то любопытное про статью найдете =)

Интересно посмотреть на граф связей между статьями, который умеет строить сайт [litmap](#).

Интерактивный оригинал доступен [вот здесь](#).

 Seed Article

 Top Related Articles



 Cited By #

 Cited By #

 Date