Проблематика следования правилу в контексте машинного обучения

Михаил Сонькин (образовательная программа «Фундаментальная и прикладная лингвистика»)

Аннотация

В статье рассказано о современном положении машинного обучения и том, как новые технологии помогают компьютеру выйти за пределы эксперимента «китайской комнаты» и обрести способность следования правилу. Для этого сравниваются инструменты, используемые в нейронных сетях, и природа следования правилу в интерпретации Людвига Вигтенштейна, Питера Уинча и Чарльза Тэйлора.

Ключевые слова: искусственный интеллект, следование правилу, Витгенштейн

1

Сегодняшнее состояние Искусственного Интеллекта (а если точнее – быстротечное развитие машинного обучения и нейронных сетей в частности) провоцирует переосмыслить множество вопросов на тему философии языка, эпистемологии и даже этики. Так, оно помогает нам с другой стороны взглянуть на проблему следования правилу, которую поставил Людвиг Витгенштейн.

Мы попробуем понять, в чем заключается отличие между тем, как взаимодействуют с правилами (в том смысле, в котором их имеет в виду Витгенштейн) человек и ИИ, и помогают ли в этом контексте нейронные сети уменьшить между ними разницу.

2

Парадокс следования правилу Витгенштейн обсуждает в своих «Философских исследованиях»[1]. В качестве одного из примеров, иллюстрирующих этот парадокс, он приводит ситуацию: представим, что мы научили ребенка писать последовательность натуральных чисел (n+1), а потом попросили сделать то же самое, но прибавлять по 2, начиная с тысячи. Ребенок пишет 1000, 1004, 1008 и т.д. и говорит, что так понимает последовательность.

Витгенштейн приходит к выводу, что в таком случае бессмысленно объяснять ребенку, почему он неправ, поскольку мы сами не знаем правильный ответ, а только подразумеваем его. Подразумеваем же мы его благодаря тому, что у нас есть опыт употребления подобной

[1] *Витгенштейн Л.* Философские исследования. М.: Издательство АСТ. 2018. С. 120.

формулы. Если же мы бы не знали то, как устроена базовая алгебра, мы могли бы проинтерпретировать правило n+2 как угодно.

Мы получаем, что «план действий нельзя определить правилом, поскольку всякий план действий можно привести в соответствие правилу»[2]. За этим следует, что существует нечто, связывающее правило и его выполнение, – то, что Витгенштейн называет подчинением (или следованием, в зависимости от перевода) правилу (Regelfolgen).

3

Для того, чтобы было возможно соотнести этот парадокс с ИИ, необходимо кратко описать главные положения современного машинного обучения.

Идея машинного обучения – использование имеющихся данных для автоматической оптимизации алгоритма. Этот подход отличается от классического rule-based подхода, который рассматривает исключительно правила, которые машина получает без "обучения" на каких-либо данных.

Нейронные сети – это одна из реализаций машинного обучения. В самом начале у программы параметры, которые в идеале должны преобразовывать данные таким образом, чтобы на выходе пользователь получал верный «ответ», заданы случайно. Во время тренировки (training) программа определенным образом подстраивается под новые полученные данные так, чтобы при следующей итерации ее ответы меньше отличались от настоящих.

Естественно, процесс обучения нейронной сети значительно более сложный и нюансированный, однако пока что нам достаточно основных понятий.

4

Важно отметить разницу между темой настоящей статьи и вопросом о способности машины мыслить: для обсуждения последнего нам пришлось бы охватить невероятный объем теории, чтобы хотя бы прийти к пониманию, что конкретно мы определяем как «сознание» или «мыслить».

[2] *Витгенштейн Л.* Философские исследования. М.: Издательство АСТ. 2018. С. 130.

На тему машины и сознания написано множество трудов, среди которых в рамках этой статьи выделим два основных: статья Алана Тьюринга[3] и статья Джона Сёрля[4]. Алан Тьюринг в своей работе сначала задает вопрос: «Могут ли машины мыслить?», однако сразу же решает заменить его на новый. Создавая контекст для вопроса, он впервые описывает игру в имитацию (The Imitation Game): два человека A и B – мужчина и женщина – разговаривают с третьим C, который их не видит. Цель человека C – по ответам на вопросы, которые он задает, определить, кто из двух людей принадлежит какому полу. Тьюринг заменяет изначальный вопрос – «Могут ли машины мыслить?» – на несколько: «Что будет, если A или B заменить на машину? Будет ли в таком случае C ошибаться настолько же часто?».

Вопрос сознания машины оказывается для Тьюринга не настолько релевантным, насколько другой, во многих смыслах более практичный – «Может ли машина имитировать мышление?»

Здесь под имитацией имеется в виду способность непосредственно обмануть человека, в то время как программа машины полностью игнорируется. Если принять как должное, что машина действительно способна имитировать мышление, то можно вывести две противоположные гипотезы, которые были сформулированы Джоном Сёрлем в 1980 году[5]:

1. Машина способна мыслить, вне зависимости от того, какой программой она руководствуется («сильный» ИИ)
2. Машина способна только имитировать мышление («слабый» ИИ)

Сам Сёрль оспаривает гипотезу «сильного» ИИ и приводит в пример мысленный эксперимент, известный как «Китайская комната»: если человека, не знающего китайский, посадить в комнату и просить его генерировать китайский текст из другого китайского текста в соответствии с правилами, прописанными на родном языке человека, то мы не можем говорить, что этот человек понимает китайский язык.

В ответ на идею о том, что человек является частью системы, которая, в отличие от него, понимает китайский, Сёрль развивает эксперимент и предлагает представить, что этот человек

[3] *Turing A*. Computing machinery and intelligence // Parsing the turing test. Springer, Dordrecht, 2009. С. 23-65.

[4] *Серль Д*. Сознание, мозг и программы //Аналитическая философия: становление и развитие/сост. АФ Грязнов. М.: Дом интеллектуальной книги, 1998. С. 376-401.

[5] Там же. С. 377.

выучивает наизусть все китайские символы и все правила и делает все преобразования в уме, таким образом воплощая в себе всю систему. В этом случае этот человек все равно не понимает китайский язык.

Этот ответ спорный, потому что в нем заведомо предполагается, что машинный код – это отдельный язык. В конце концов, язык программирования – это не язык, на котором «говорит» машина, а способ для человека задать правила. В случае с цифровыми компьютерами система при запуске читает не код, а сигналы, подающиеся ей в результате прочтения этого кода другой программой. Возможно, язык программирования и код, написанный на нем, не стоит учитывать в нашем эксперименте, поскольку они на самом деле нужны не для взаимодействия с системой, а для создания ИИ.

5

Применим модель китайской комнаты к контексту следования правилу. Между комнатой и человеком можно выделить одно значимое различие: комната может интерпретировать правило (т.е. алгоритм) только одним образом, в то время как человеку необходим тот самый промежуточный этап – интерпретация и следование правилу. В некотором смысле это можно сформулировать так, что для китайской комнаты формулировка правила равнозначна его следованию.

Однако предположим, что в нашей машине два алгоритма: первый показывает, как преобразовывать текст (назовем его генерация), второй – как менять первый алгоритм (обучение). Изначальные параметры первого алгоритма заданы случайным образом. Так, мы, как создатели машины, абстрагируемся от процесса, которым руководствуется машина при выполнении задания, и ее процесс становится для нас «черным ящиком». Примерно так работает нейронная сеть: мы подаем ей большое количество данных, которые помогают ей с каждым шагом все лучше настроить свои параметры.

Можно ли все еще говорить, что у машины есть только один способ интерпретировать правило? Для этого нужно понять, как мы хотим рассматривать эти два алгоритма: как единый алгоритм в два шага или как два разных алгоритма, следующих друг за другом. В последнем случае мы считаем, что обучение – это просто способ дать машине ресурсы для того, чтобы

она смогла действовать самостоятельно, а генерация – это уже то, что машина делает сама, поскольку это действия, которые мы не прописывали.

Таким образом, мы можем проинтерпретировать архитектуру нейронной сети так, что она, в каком-то смысле, имеет то среднее звено – следование правилу – которое отсутствует у более простых ИИ.

6

В связи с образованием понятия «следование правилу» встает новый вопрос – в чем суть этого понятия? Некоторые искали ответ в том, как индивидуум при следовании правилу взаимодействует с обществом: Питер Уинч[6], дополняя мысль Витгенштейна о невозможности следовать правилу только одному человеку только один раз, утверждает, что следование правилу обязательно включает в себя рассмотрение реакции других людей.

Так, если человек нас попросит закончить ряд «1 3 5 7», но возразит нашему решению «9 11 13 15», говоря, что верный ответ – повторить эти четыре цифры, и будет продолжать так себя вести с каждым добавлением, он тоже будет следовать правилу. Правило заключалось бы в том, чтобы говорить что угодно, противоречащее тому правилу, которое выбрали мы. Из этого Уинч делает вывод, что при определении понятия следования правилу «необходимо не только принимать во внимание действия человека, который является кандидатом на следование правилу, но и реакции других людей на то, что он делает»[7].

Чарльз Тэйлор применяет понятие «фонового понимания», наследуя его у Сёрля, для того, чтобы прояснить понятие следования правилу[8]. Витгенштейн говорит об одной особенности человеческого взаимодействия с правилами: «Я повинуюсь правилу слепо»[9]. Однако Тэйлор замечает, что, если у нас есть необходимость, мы можем обосновать наше действие в соответствии с правилом. Он выводит, что где-то хранится знание о том, как следовать

---

[6] *Уинч П.* Идея социальной науки и ее отношение к философии/пер. с англ //М. Горбачева и Т. Дмитриева. М.: Русское феноменологическое общество, 1996.

[7] *Уинч П.* Идея социальной науки и ее отношение к философии /пер. с англ // М. Горбачева и Т. Дмитриева. М.: Русское феноменологическое общество, 1996. С. 22.

[8] *Волков В. В., Хархордин О. В.,* Теория практик //СПб.: Изд-во Европейского университета в Санкт-Петербурге, 2008. С. 94-95.

[9] *Витгенштейн Л.* Философские исследования. М.: Издательство АС, 2018. С. 136.

правилу, и выдвигает гипотезу, что следование правилам является телесными навыками и изначально существует в человеке.

7

Если мы считаем, что нейронная сеть, благодаря своей архитектуре, имеет способность следовать правилу (или по крайней мере способна это имитировать; здесь вполне возможно разделение, аналогичное «сильному» и «слабому» ИИ), то стоит похожим образом попробовать выделить, в чем ее суть.

Интересным образом, объяснения Уинча и Тэйлора можно так или иначе применить к нейронной сети. Уинч утверждает, что следование правилу учитывает реакцию от общества. Но, предположим, у нас есть кнопка и правило «Не нажимать на кнопку». Если мы нажмем на кнопку, нас ударит током, и мы понимаем, как ему следовать. Для того, чтобы из действия сделать вывод о следовании правилу, нам не понадобилась реакция от общества, нам хватило одного стимула. Соответственно, поощрение/порицание можно приравнять к такому же типу стимулов, которые определяют наше следование правилу. Если так расширить понятие следования правилу, то можно считать, что нейронная сеть учится похожим образом: неправильный ответ заставляет ослаблять те связи в архитектуре, которые к нему привели, правильный ответ заставляет их укреплять.

Идея про «фоновое понимание», т.е. изначальную способность следовать правилу, помогает нам объяснить, почему архитектура нейронной сети, которую задает человек, а не машина, имеет смысл в контексте следования правилу: это такой же изначальный навык машины, который в дальнейшем тренируется на данных и стимулах, которые служат реакцией на действия, и в итоге формирует следование правилу.

8

Наш анализ открывает множество вопросов в сфере ИИ. Стоит ли утверждать, что китайская комната способна обретать опыт самостоятельно? Стоит ли архитектуру нейронной сети считать таким же «телесным навыком», как нашу способность из формулировки правила выводить действие?

Проанализировав «Философские исследования» Витгенштейна и воспользовавшись классическим примером Сёрля, мы можем предположить, что нейронные сети развиваются в сторону образования следования правилу как новой способности ИИ.

Библиография

Turing A. Computing machinery and intelligence // Parsing the turing test. Springer, Dordrecht, 2009. С. 23-65.

Витгенштейн Л. Философские исследования. М.: Издательство АСТ, 2018.

Волков В. В., Хархордин О. В., Теория практик // СПб.: Изд-во Европейского университета в Санкт-Петербурге, 2008.

Сёрль Дж. Сознание, мозг и программы //Аналитическая философия: становление и развитие / сост. АФ Грязнов. М.: Дом интеллектуальной книги, 1998. С. 376-401.

П. Уинч. Идея социальной науки и ее отношение к философии/пер. с англ // М. Горбачева и Т. Дмитриева. М.: Русское феноменологическое общество, 1996.