МАШИНА, БУДЬ ЧЕЛОВЕКОМ!

Искусственный интеллект переживает второе рождение: теперь его создают на основе информации о том, как учатся дети

Элисон Гопник

Illustration by Simon Prades

ОБ АВТОРЕ

Элисон Гопник (Alison Gopnik) — профессор психологии и философии в Калифорнийском университете в Беркли, изучает проблемы познания ребенком окружающего мира.

Если вы проводите много времени с детьми, то, наверное, вам доводилось удивляться, как эти маленькие человечки могут так быстро столько всего выучить. Философы, начиная еще с Платона, тоже задумывались над этим вопросом, но так и не нашли подходящего ответа. Мой пятилетний внук Оджи уже умеет пользоваться часами, знает названия растений и зверей, не говоря о динозаврах и космических кораблях. Кроме того, он понимает, чего хотят другие люди, что они думают и чувствуют. Он может использовать свои знания для систематизации всего, что он видит и слышит, и строить на их основе новые прогнозы. Например, на днях он заявил, что недавно открытый новый вид тираннозавра, которого показывали в Американском музее естественной истории, был травоядным, и поэтому он не страшный.

При этом все, что получил Оджи от окружающего мира, — это поток фотонов, падающих на сетчатку, и колебания воздуха, контактирующие с его барабанной перепонкой. Нейронный компьютер, находящийся позади его голубых глаз, работает таким образом, что, получив эту ограниченную информацию от органов чувств, выдает предположение о растительноядном тираннозавре. Сможет ли электронный компьютер делать так же — пока не понятно.

В течение последних примерно 15 лет программисты и психологи пытались в этом разобраться. Дети приобретают огромное количество знаний почти без помощи родителей и учителей. И, несмотря на огромные успехи в области искусственного интеллекта, даже самые мощные компьютеры не могут обучаться так же хорошо, как пятилетний мальчик.

Основная задача программистов на ближайшие десятилетия — понять, как на самом деле работает мозг ребенка, и создать столь же эффективную электронную версию. Сейчас они уже начинают разрабатывать искусственный интеллект на основе того, что нам известно о механизмах обучения у людей.

Подняться вверх

После первого всплеска в 1950–1960-х гг. в последующие десятилетия интерес к искусственному интеллекту ослаб. Однако за последние несколько лет произошло несколько крупных открытий, особенно в области машинного обучения, и теперь искусственный интеллект стал одной из актуальнейших тем в технологии. О последствиях этих открытий появилось много утопических и апокалиптических прогнозов. Они буквально предвещали бессмертие или конец света, и про обе эти возможности было много написано.

Я подозреваю, что развитие искусственного интеллекта порождает такие сильные чувства из-за нашего глубинного страха перед «почти человеком». Мысль, что некоторые создания, будь

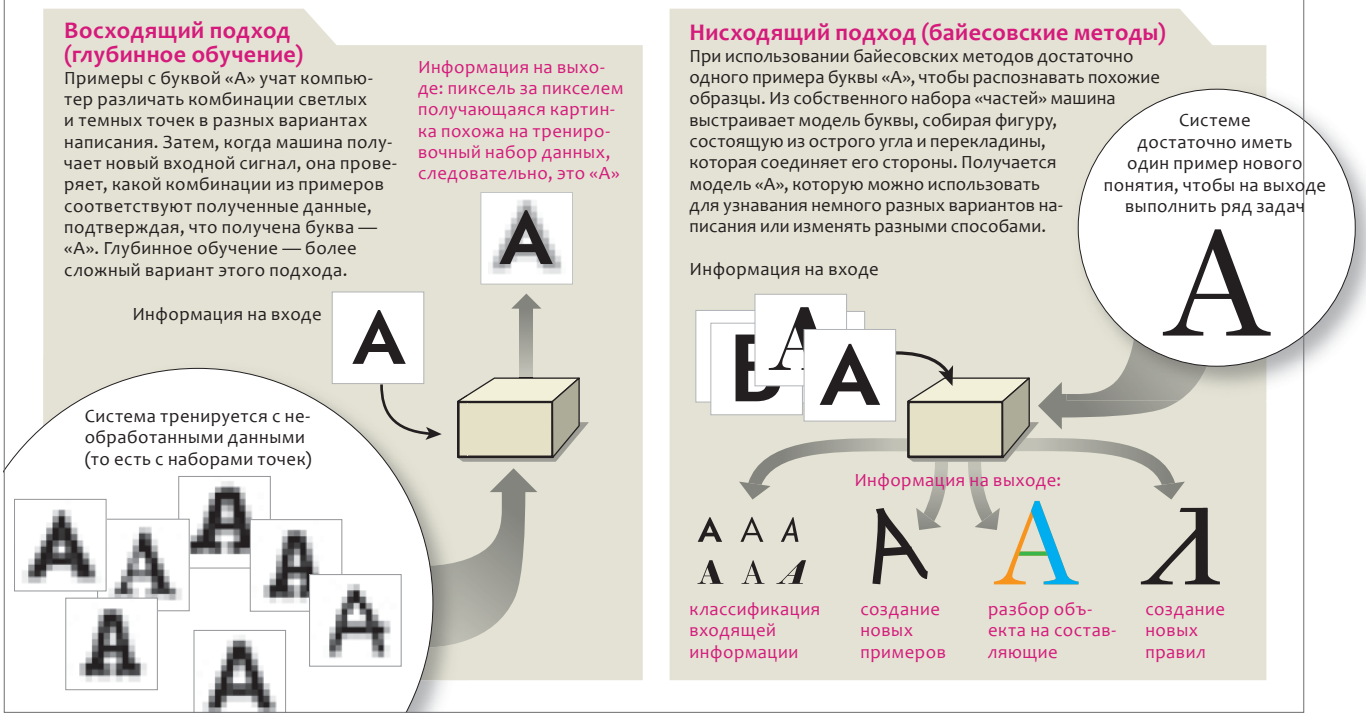
ОСНОВНЫЕ ПОЛОЖЕНИЯ

- Как маленькие дети приобретают свои знания? Этот вопрос уже давно волнует философов и психологов, а теперь им заинтересовались и программисты.
- Специалисты в области искусственного интеллекта изучают умственные способности дошкольников, чтобы разработать способы научить машину познавать мир.
- Две разных стратегии машинного обучения, основанные на попытках скопировать естественные способности детей, изменяют подход к разработке искусственного интеллекта.

ПРОТИВОПОЛОЖНЫЕ СТРАТЕГИИ

Два пути к возрождению искусственного интеллекта

Проблемы, которые среднестатистический пятилетний ребенок решает без затруднений, могут поставить в тупик даже самые мощные компьютеры. Искусственный интеллект в последние годы сделал резкий рывок, и компьютеры могут узнавать информацию о мире примерно так же, как и дети. Машина распознает букву «А» из необработанной сенсорной информации (восходящий подход) или делает предположения на основе существующих знаний (нисходящий подход).



то средневековый голем, чудовище Франкенштейна или сексуальная Ава, роковая женщина-робот из фильма «Из машины» *(Ex Machina)*, могут заполнить брешь между человеческим и искусственным, всегда вызывала большую тревогу.

Но действительно ли компьютеры обучаются лучше, чем люди? Насколько эти жаркие споры отражают истинно революционные достижения? Или это просто шум на пустом месте? На первый взгляд, трудно разобраться в том, как компьютеры распознают, скажем, кошек, произнесенные слова или японские иероглифы. Однако при ближайшем рассмотрении основные идеи, лежащие в основе машинного обучения, оказываются вовсе не такими сложными.

При первом подходе все начинается с тех же фотонов и колебаний воздуха, которые воспринимает Оджи и все мы, — эта информация попадает в компьютер в виде точек цифрового изображения или аудиозаписи звука. Затем в этих цифровых данных компьютер пытается выделить ряд элементов, которые можно объединить и идентифицировать как единый объект окружающего мира. Этот восходящий подход произрастает из идей философов Дэвида Юма и Джона Стюарта Милля, психологов Ивана Петровича Павлова, Берреса Фредерика Скиннера и других.

В 1980-х гг. ученые придумали оригинальный способ использования этого восходящего метода, позволивший компьютерам находить значащие закономерности в данных. Эта коннекционная система, она же нейронная сеть, основывалась на принципах работы нейронов, превращающих падающий на сетчатку свет в образ окружающего мира. Нейронная сеть делает нечто похожее. В ней используются взаимосвязанные элементы, подобные биологическим клеткам, и точки с нижнего уровня сети по мере обработки данных на более высоких уровнях превращаются во все более и более абстрактный образ, такой как нос или целое лицо.

Идея нейронной сети недавно пережила перерождение в связи с появлением новой технологии глубинного обучения, которая сейчас используется в коммерческих целях такими крупнейшими предприятиями, как, например, *Google* или *Facebook*. Кроме того, нынешняя популярность таких систем частично связана и с постоянно растущей мощностью компьютеров — это явление известно как закон Мура. Таким образом, можно создать очень большие массивы данных. С лучшими вычислительными возможностями и с огромным объемом данных для обработки нейросетевые системы могут обучаться значительно более эффективно, чем мы могли когда-то предполагать.

SOURCE: "HUMAN-LEVEL CONCEPT LEARNING THROUGH PROBABILISTIC PROGRAM INDUCTION," BY BRENDEN M. LAKE, RUSLAN SALAKHUTDINOV AND JOSHUA B. TENENBAUM, IN SCIENCE, VOL. 350; DECEMBER 11, 2015 (Bayesian example); illustration by Jen Christiansen

Все эти годы создатели искусственного интеллекта колебались между этим восходящим алгоритмом машинного обучения и альтернативным нисходящим подходом. Нисходящий метод использует то, что машина уже знает, чтобы помочь ей выучить что-то новое. Платон, а также философы-рационалисты, такие как Рене Декарт, были приверженцами нисходящего подхода в обучении. И это сыграло важную роль в первых системах искусственного интеллекта. В 2000-х гг. этот принцип также пережил второе рождение в виде вероятностного, или байесовского, моделирования.

При нисходящем подходе работа начинается с того, что формулируются абстрактные и широкие гипотезы о мире, точно так же как делают ученые. Затем система строит прогнозы, как будут выглядеть данные, если верны ее гипотезы. Как и ученые, такие системы пересматривают свои гипотезы в зависимости от наблюдаемого результата.

Нигерия, виагра и спам

Восходящий метод, вероятно, более понятен, поэтому давайте начнем с него. Представьте, что вы пытаетесь добиться, чтобы компьютер отличал важные сообщения от спама, который попадает в папку «Входящие». Можно заметить, что спам имеет определенные отличительные черты: большой список адресатов, отправлено из Нигерии или Болгарии, предлагает получить приз в миллион долларов или, возможно, в нем говорится о виагре. Однако полезные письма могут выглядеть так же. Вы же не хотите пропустить сообщение о повышении в должности или получении научной премии.

Если вы сравните достаточное число образцов спама с другими электронными письмами, то заметите, что только в спаме можно встретить ряд признаков, скомбинированных определенным образом. Например, сочетание Нигерии и сообщения о выигрыше приза в $1 млн предвещает неприятности. На самом деле есть довольно много признаков более высокого уровня, отличающих спам от полезных сообщений, которые не так очевидны, — например, опечатки или *IP*-адреса. Если вы сумеете выделить их, то сможете точно отфильтровать спам без всякой боязни пропустить сообщение о доставке вашей виагры.

Восходящий подход в машинном обучении позволяет выявлять значимые признаки, чтобы решать задачи такого типа. Для этого нейронная сеть должна пройти процесс самообучения. Она изучает миллионы сообщений из огромных баз данных, где каждое письмо помечено как спам или как обычное письмо. Затем компьютер вычленяет набор идентифицирующих признаков, которые отделяют спам от всего остального.

Аналогичным способом сеть может проверить изображения из интернета, помеченные как «кошка», «дом», «стегозавр» и т.д. Находя общие особенности в каждом наборе изображений, то, что отличает всех кошек от всех собак, сеть затем может узнавать новые изображения кошек, даже если она никогда не видела их раньше.

Один такой восходящий метод называется неконтролируемым обучением, пока он еще только начинает развиваться, но уже может находить разные образы в наборе данных, которые никак не помечены. Он просто выявляет совокупность свойств, характерных для данного объекта: например, носы и глаза всегда находятся вместе, образуя лицо, и отличаются от гор и деревьев на заднем плане. В современных технологиях глубинного обучения выявление объектов происходит за счет разделения задач распознавания между разными уровнями нейронной сети.

В статье, опубликованной в журнале *Nature* в 2015 г., показано, как далеко может продвинуться технология, использующая восходящий подход. Исследователи из компании *DeepMind*, принадлежащей *Google*, применили комбинацию из двух вариантов восходящего подхода (глубинное и стимулированное обучение), чтобы компьютер мог освоить видеоигры на приставке *Atari 2600*. Изначально компьютер ничего не знал о том, как устроены эти игры. Алгоритм глубинного обучения помог системе выделить объекты на экране, а алгоритм стимулированного обучения подкреплял лучшие результаты. В итоге компьютеры достигли высокого уровня прохождения некоторых игр, а в отдельных случаях делали это лучше, чем самые опытные люди-игроки. С другой стороны, в ряде игр, которые человек осваивает так же легко, компьютеры потерпели неудачу.

Возможность обучать искусственный интеллект на огромных выборках, таких как миллионы фотографий *Instagram*, электронных писем или записей голоса, позволяет решить задачи, которые раньше считались недоступными, например, распознавание изображения или речи. И все же надо помнить, что у моего внука нет никаких проблем с тем, чтобы узнать животное или ответить на вопрос, даже при том что у него выборка для обучения была гораздо меньше. Некоторые задачи, элементарные для пятилетнего ребенка, ставят в тупик компьютер — оказалось, что для машины они значительно сложнее, чем научиться играть в шахматы.

Компьютеру, чтобы обучиться распознавать мохнатые и усатые лица, часто нужны миллионы примеров упорядоченных объектов, в то время как нам для классификации достаточно лишь нескольких. После интенсивного обучения компьютер может узнать кошку на изображении, которое он никогда не видел. Однако то, как он это делает, совершенно не похоже на процесс обобщения у человека. И, поскольку программа «рассуждает» по-другому, появляются ошибки. Некоторые изображения кошек не будут отмечены как кошки. Кроме

того, компьютер может неверно увидеть на изображении кошку в случайном пятне, которое никогда не обманет человека.

Спуститься вниз

Другой подход к машинному обучению, изменивший искусственный интеллект в последние годы, ориентирован в ином направлении — «сверху вниз». Он предполагает, что мы можем получить общую информацию из конкретных данных, потому что мы уже знаем многое о мире, а особенно потому, что наш мозг уже способен понимать основные абстрактные понятия. Как и ученые, мы можем использовать эти понятия для формулировки гипотез о мире и составления прогнозов, каковы должны быть факты или события, если наши гипотезы верны. Это прямо противоположно восходящему подходу, когда машина пытается найти какую-либо закономерность в исходных данных.

Эту идею можно проиллюстрировать вновь примером из спама, рассмотрев реальный случай, произошедший со мной. Я получила электронное письмо от редактора журнала с непонятным названием с упоминанием одной из моих работ и просьбой написать статью и опубликовать ее у них. Никакой Нигерии, миллиона долларов или виагры — это сообщение не содержало ничего из обычных признаков спама. Однако с помощью того, что я уже знала, и отвлеченных рассуждений о процессе возникновения спама я смогла понять, что это подозрительное письмо.

Во-первых, я уже знала, что спамеры пытаются получить деньги от людей, играя на их жадности, а ученые так же жаждут публикаций, как обычные люди выигрыша в миллион долларов или улучшенных сексуальных способностей. Еще я знала, что настоящие журналы с открытым доступом покрывают свои издержки за счет авторов, а не подписчиков. Кроме того, моя работа не имела ничего общего с названием этого журнала. Сложив все вместе, я выдвинула правдоподобную гипотезу, что это письмо пытается обмануть ученых, чтобы они заплатили за свою публикацию в этом лжежурнале. Я могу прийти к такому выводу на одном-единственном примере, и я могу протестировать свою гипотезу, проверив добросовестность издателя через поисковые системы.

Специалисты по информатике назвали бы мои рассуждения порождающей моделью, которая способна представить такие абстрактные понятия, как жадность и обман. Эта же модель может описать процесс, который использовался для придумывания гипотезы, процесс рассуждения, приведший к выводу, что это письмо было спамом. Этот прием позволяет мне объяснить, как работает этот тип спама, а также помогает вообразить другие формы спама, даже такие, про которые я ничего не видела и не слышала раньше. Когда я получаю письмо от журнала, эта модель позволяет пройти шаг за шагом в обратном порядке и понять, почему это должно быть спамом.

Порождающие модели имели большое значение во время первой волны работ по искусственному интеллекту и когнитивной науке в 1950–1960-х гг. Однако у них есть недостатки. Во-первых, большинство наблюдаемых явлений может, в принципе, объясняться многими различными гипотезами. В моем случае это письмо могло быть на самом деле честным, даже если это кажется маловероятным. Таким образом, порождающая модель должна включать в себя предположения о вероятности, и это наиболее важное усовершенствование последнего времени в таких методах. Во-вторых, часто бывает непонятно, откуда берутся базовые абстрактные понятия, на которых строится порождающая модель. Такие мыслители, как Рене Декарт и Ноам Хомский, полагали, что эти понятия твердо закреплены у нас уже при рождении. Но неужели мы приходим в этом мир со знанием, что жадность и обман приводят к проигрышу?

Яркий пример современного применения нисходящих методов — это байесовская модель. Она как раз имеет дело с этими двумя вопросами. Названная так в честь статистика и философа XVIII в. Томаса Байеса, она объединяет порождающую модель с теорией вероятности, используя прием, который называется «байесовский вывод». Вероятностная порождающая модель показывает, насколько вероятно то, что вы увидите какую-то структуру в данных, если верна определенная гипотеза. Если это электронное письмо — обман, то оно, вероятно, взывает к жадности читателя. Само собой, письмо может взывать к жадности и при этом не быть спамом. Байесовская модель объединяет ваши представления о потенциальных гипотезах с наблюдаемыми данными, позволяя довольно точно рассчитать, с какой вероятностью перед нами нужное письмо или спам.

Такой нисходящий метод лучше, чем восходящий, соответствует нашим представлениям о том, как обучаются дети. Именно поэтому я со своими коллегами уже 15 лет использую байесовскую модель в наших исследованиях развития детей. И наша лаборатория, и другие ученые применяют этот подход, чтобы понять, как дети познают причинно-следственные связи, и прогнозировать, как и когда дети сформируют новые знания о мире и поменяют уже имеющиеся у них представления.

Байесовский метод отлично подходит и для того, чтобы обучить машину учиться так же, как это делают люди. Сотрудник Массачусетского технологического института Джошуа Тененбаум (Joshua Tenenbaum), с которым у меня была когда-то совместная работа, Бренден Лэйк (Brenden Lake) из Нью-Йоркского университета и их коллеги опубликовали статью в 2015 г. в журнале *Science*. Они

разработали систему искусственного интеллекта, которая могла распознавать незнакомые рукописные буквы. Эта задача проста для людей и требует больших усилий от компьютера.

Подумайте о ваших собственных навыках распознавания. Даже если вы никогда раньше не видели иероглифов на японских свитках, вы скорее всего сможете сравнить два из разных свитков и сказать, одинаковые это или разные символы. Возможно, вы сможете их нарисовать или даже придумать несуществующий японский иероглиф и понять, что эти иероглифы заметно отличаются от корейских или русских символов. Это как раз то, что умеет делать программа, созданная Тененбаумом с коллегами.

При использовании нисходящего подхода компьютер должен был познакомиться с тысячами примеров, выявить закономерности и применить эту информацию для анализа новых изображений. Вместо этого байесовская программа дала компьютеру общую схему, как начертить символ. Например, линия может пойти направо или налево. После того как программа заканчивала с одним символом, она переходила к следующему.

Когда программа видела определенный символ, она могла сделать предположение о последовательности движений, необходимых для его написания. Затем она сама совершала схожий набор действий. Это развивается так же, как моя цепочка рассуждений, запущенная сомнительным письмом из журнала. Вместо того чтобы выяснять, насколько полученное письмо похоже на обман, модель Тененбаума определяет, может ли эта конкретная последовательность действий привести к определенному результату. При использовании одного и того же набора данных нисходящий алгоритм оказывается более эффективным, чем глубинное обучение, кроме того, он больше похож на мышление человека.

Идеальная пара

Эти два ведущих подхода к обучению машин — принцип восходящего анализа и метод нисходящей пошаговой детализации — имеют дополняющие друг друга сильные и слабые стороны. С восходящим алгоритмом компьютеру не нужно понимать что-нибудь про кошек, чтобы начать учиться, но ему нужно очень большое количество данных. Байесовская система может обучиться на малом количестве примеров и делать более широкие обобщения. Однако при использовании этого метода требуется много предварительной работы, чтобы сформулировать правильный набор гипотез. И разработчики обоих типов систем могут столкнуться со схожими затруднениями. Эти два подхода работают только для относительно узких и четко определенных проблем, таких как распознавание письменных знаков или кошек или игра в видеоигры.

Дети не испытывают затруднений при тех же самых ограничениях. Специалисты по психологии развития убедились, что дети каким-то образом комбинируют лучшее от каждого подхода и затем продвигаются гораздо дальше. Оджи может обучиться, используя только один или два примера, как при нисходящем методе. Но он может также выводить новые закономерности из имеющихся данных, подобно системе, работающей по принципу восходящего анализа. Эти закономерности не были ему очевидны с самого начала.

На самом деле Оджи может гораздо больше. Он не только мгновенно узнает кошек и называет разные буквы, он может создавать поразительно новые и творческие рассуждения, которые выходят далеко за пределы его знаний или опыта. Недавно он объяснил, что если взрослый хочет снова стать ребенком, то он не должен есть полезные для здоровья овощи, потому что из-за них дети становятся взрослыми. У нас нет почти никакого представления о том, как появляются творческие рассуждения такого рода.

Всякий раз, когда мы слышим, что искусственный интеллект — это экзистенциальная угроза, мы должны вспоминать про таинственную силу человеческого мышления. Словосочетания «искусственный интеллект» и «машинное обучение» звучат страшно. В некотором смысле так оно и есть. Военные исследуют способы использования этих систем для управления оружием. Однако естественная глупость может нанести больше вреда, чем искусственный интеллект, и людям сейчас надо быть гораздо умнее, чем мы были в прошлом, чтобы правильно контролировать новые технологии. Закон Мура — это серьезная сила, он может иметь важные практические последствия, даже если прогресс в области программирования будет связан только с количественным ростом данных и вычислительной способности, а в понимании механизмов мышления прорыва не произойдет. Тем не менее не надо думать, что мы вот-вот выпустим в мир нового технологического голема. ■

Перевод: М.С. Багоцкая

ДОПОЛНИТЕЛЬНЫЕ ИСТОЧНИКИ

- Бенджо Д. Компьютеры тоже учатся // ВМН, № 8–9, 2016.
- Bayesian Networks, Bayesian Learning and Cognitive Development. Alison Gopnik et al. in Developmental Science, Vol. 10, No. 3, pages 281–287; May 2007.
- Human-Level Concept Learning through Probabilistic Program Induction. Brenden M. Lake et al. in Science, Vol. 350, pages 1332–1338; December 11, 2015.
- The Gardener and the Carpenter: What the New Science of Child Development Tells Us about the Relationship between Parents and Children. Alison Gopnik. Farrar, Straus and Giroux, 2016.