

Маркус дю Сотой. Нейросеть разумная (отрывок)

Глава 5. От нисходящего к восходящему

Алгоритмические галлюцинации

Успехи, достигнутые за последние пять лет в области компьютерного зрения, поразили всех. Причем новые алгоритмы способны ориентироваться не только в изображениях человеческого тела. Достижение способности декодировать визуальные образы, сравнимой с той, которой обладает человеческий мозг, было трудной задачей для любого компьютера, претендующего на роль конкурента человеческого разума. Цифровая видеокамера может зафиксировать изображение в таких подробностях, сохранение которых далеко превосходит возможности человеческого мозга, но это не значит, что такая камера способна извлечь из миллионов пикселей связное изложение. Мы все еще далеки от понимания того, как именно мозг обрабатывает данные и объединяет их в единую концепцию, не говоря уже о воспроизведении этого процесса в наших кремниевых друзьях.

Почему, когда мы получаем информацию, поступающую через наши органы чувств, мы способны интегрировать ее в единое целое? Мы не воспринимаем красный цвет и кубическую форму красного кубика как две разные характеристики. Они сливаются воедино. Воспроизведение такого слияния является одной из самых трудных задач в процессе обучения компьютера интерпретации изображений. Считывание изображения пиксель за пикселем мало что говорит об общей картине. Чтобы ощутить это на собственном опыте, возьмите лист бумаги и сделайте в нем маленькую дырочку. А теперь положите этот лист на изображение человеческого лица формата А4. Передвигая дырочку по изображению, почти невозможно понять, чей это портрет.

Еще пять лет назад эта задача казалась неразрешимой. Но это было до пришествия машинного обучения. В прошлом программисты пытались разрабатывать нисходящие алгоритмы для распознавания визуальных образов. Но попытки создания наборов инструкций типа «если... то...» для идентификации изображений так и не увенчались успехом. Все изменила восходящая стратегия, позволяющая алгоритму создавать свое собственное дерево решений, исходя из тренировочных данных. Новым ингредиентом, давшим такую возможность, было огромное количество размеченных визуальных данных, имеющееся теперь в Сети. Любая картинка, опубликованная в инстаграме, в сопровождении наших комментариев дает полезные данные, ускоряющие процесс обучения.

Чтобы оценить силу этих алгоритмов, можно загрузить какое-нибудь изображение на сайт `Coogler` по адресу <https://cloud.google.com/vision>. В прошлом году я загрузил туда изображение нашей рождественской елки, и алгоритм определил с вероятностью 97%, что ему показывают именно изображение рождественской елки. Может показаться, что в этом нет ничего особенно потрясающего, но на самом деле такой результат впечатляет. Однако эти алгоритмы не безошибочны. После первоначальных восторгов у них обнаружились и отрезвляющие недостатки. Взять, например, алгоритмы, которые британская Служба столичной полиции сейчас обучает находить в Сети порнографические изображения с участием детей. В данный момент их очень сильно сбивают с толку изображения пустынь.

«Иногда он выдает картинку с пустыней и считает, что это непристойное или порнографическое изображение, — признал в недавнем интервью Марк Стоукс, руководитель отдела цифровой и электронной криминалистики. — Почему-то многие используют изображения пустыни в качестве экранной заставки, и алгоритм отмечает такие картинки, принимая цвет песка за цвет кожи». К тому же очертания дюн, по-видимому, соответствуют формам, которые алгоритм принимает за изгибы частей тела.

Мы видели множество ярких демонстраций самых затейливых способов, которыми можно обмануть компьютерное зрение, внушив алгоритму, что он видит нечто, чего на самом деле нет. Независимая группа `LabSix`, работающая в области исследований искусственного интеллекта и состоящая из выпускников и старшекурсников Массачусетского технологического института, сумела запутать алгоритмы распознавания визуальных образов так, что они приняли трехмерную модель черепахи за винтовку. Результат оставался неизменным, под каким бы углом ни держали черепаху — ее даже можно было поместить в среду, в которой ожидалось найти черепах, а не оружие.

Алгоритм обманули следующим образом: поверх черепахи наложили изображение, которое кажется человеческому глазу рисунком панциря и кожи черепахи, но на самом деле было хитрым образом построено на основе изображений винтовок. Изображение винтовки постепенно и многократно изменялось до тех пор, пока не стало неразличимым для человека. Однако компьютер по-прежнему извлекает из него информацию о винтовке, даже после ее искажения, а она имеет в его системе распознавания объектов более высокий приоритет, чем информация о черепахе, поверх которой она была нанесена. Алгоритмы также удавалось ввести в заблуждение так, чтобы они приняли изображение кошки за тарелку гуакамоле, но достижение

LabSix состоит в том, что в работе этой группы алгоритм был уверен, что ему показывают винтовку, независимо от того, под каким углом поворачивали черепаху.

Та же группа продемонстрировала, что изображение собаки, постепенно преобразующейся, пиксель за пикселем, в пару лыжников на склоне, по-прежнему распознается как собака, даже когда собака полностью исчезает с экрана. Этот результат был тем более замечательным, что алгоритм, использованный в этой работе, был для хакеров абсолютным черным ящиком. Они не знали, как именно расшифровывается изображение, и тем не менее сумели обмануть алгоритм.

Исследователи из компании Google зашли на один шаг дальше: они создали изображения, настолько интересные для алгоритма, что он не обращает никакого внимания на остальные части картинки. При этом использовалось то обстоятельство, что алгоритмы присваивают более высокий приоритет тем пикселям, которые они считают важными для классификации изображения. Если алгоритм пытается распознать лицо, он игнорирует пиксели, относящиеся к фону — небу, траве, деревьям и так далее. В Google создали психоделические цветовые пятна, которые полностью завладевают вниманием алгоритма: если в обычной ситуации он был способен распознать изображение банана, то при наличии подобного цветового пятна банан совершенно исчезает из поля его зрения. Пятна можно сделать так, чтобы они распознавались как изображения произвольных объектов — например, изображение тостера. Какое бы изображение ни показали алгоритму, если в нем есть такое пятно, алгоритм будет считать, что видит тостер. Этим он несколько похож на собаку, которую можно полностью отвлечь от чего угодно, показав ей мячик: из создаваемого ею мира исчезает всё на свете, и она не видит ничего, кроме этого мячика, и не думает ни о чем другом. В большинстве предыдущих попыток взлома таких алгоритмов нужно было иметь какую-то информацию об изображении, которое алгоритм хотели заставить неправильно классифицировать, но эти вновь открытые пятна обладают тем преимуществом, что работают независимо от того, какое изображение они искажают.

Человека такие уловки не обманывают, но это не значит, что мы не подвержены аналогичным эффектам. Особенности работы нашего мозга используют фокусники: мы обычно отвлекаемся на какой-то один элемент в поле зрения и можем совершенно не замечать, что еще происходит в это же время. Классический пример этого эффекта дает знаменитый видеофильм с двумя командами, пасующими баскетбольные мячи. Если зрителей просят подсчитывать пасы, выполненные одной из команд, внимательно следя за перемещениями мяча, большинство совершенно не замечает человека в костюме обезьяны, который проходит между игроками, бьет себя в грудь, а затем уходит с площадки. Описанные хакерские атаки на компьютерное зрение просто выявляют слепые пятна алгоритмов — но таких слепых пятен полно и у человека.

Учитывая, что в управлении беспилотными автомобилями используются алгоритмы компьютерного зрения, возможность такой атаки явно представляет опасность. Представьте себе знак «Стоп», на который прикрепили наклейку с психоделическим пятном, или систему безопасности, управляемую алгоритмом, который совершенно не замечает пистолета, потому что считает, что это черепаха.

Я решил как следует испытать алгоритм Kinect и попытаться обмануть его, изгибая свое тело необычным образом, но это оказалось не так-то просто. Даже когда я принимал причудливые позы йоги, которые алгоритм не видел в своих тренировочных данных, ему все равно удавалось с высокой точностью идентифицировать части моего тела. Поскольку наши тела вряд ли могут делать что-нибудь принципиально новое, этот алгоритм по большей части заморожен и не будет развиваться дальше. Ему уже незачем изменяться, так как он уже хорошо делает то, для чего он создан. Но другим алгоритмам, возможно, придется продолжать приспосабливаться к новым представлениям и изменениям окружающего их мира. Алгоритмы, рекомендуемые фильмы, которые нам, возможно, захочется посмотреть, книги, которые нам, возможно, захочется прочитать, музыку, которую нам, возможно, захочется послушать, должны быть достаточно расторопными, чтобы реагировать на изменения наших вкусов и на поток новых творческих произведений, порождаемых человеческим кодом.

Именно в этом аспекте и проявляется могущество алгоритма, способного продолжать учиться, изменяться и приспосабливаться к новым данным. Машинное обучение открыло перспективу появления алгоритмов, изменяющихся и взрослеющих так же, как люди.