

УДК 004.85

МАШИННОЕ ОБУЧЕНИЕ

Ерошевич К.В., ассистент кафедры ИиАПС
Кузбасский государственный технический университет
имени Т.Ф. Горбачева
г. Кемерово

Есть такой процесс, он называется автоматизация производства – это история про то, что бы использовать технические средства и математические методы, что бы освободить человека от повседневных рабочих процессов. Если речь идет про рутинные процессы, то возникает вопрос: зачем нам вообще нужны все эти умные машины и искусственный интеллект, и почему нельзя просто взять и свести задачи к базовым алгоритмам: каким то немудреным последовательностям действий? Ну да, в случае чего то простого – это будет десяток или два действий. А в случае чего то сложного пару сотен или тысяч действий. Так сделать конечно можно, но это не универсально.

Представьте что будет с алгоритмом, если в задачи произошло отклонение хотя бы на одно действие? Тут же появятся много новых параметров, которые нужно учитывать вместе с теми, что уже есть. И очень часто бывают процессы настолько сложные, что сам базовый алгоритм действий не понятен. Поэтому нам нужен искусственный интеллект, как некий более точный инструмент. И постановка задачи немного другая: нсть очень много данных, закономерности в них не ясны и не понятны, и необходимо найти алгорит, который поможет их выявить.

И тут самое время провести аналогию между нашим мозгом и искусственной нейронной сетью, но не в мышлении, а в обучении. Именно в том, как мы учимся: на многих примерах, методом проб и ошибок, если что то не получается сразу, то нам показывают какой должен быть результат. И эта крайне жизненная аналогия работает. Продвижение беспилотных автомобилей, прогноз погоды с точностью до минут, работа на бирже в решении вопросов с ценными бумагами и многое другое – это все примеры сложных, не однотипных процессов, в которые искусственный ителлект уже во всю принимает решения.

А можно ли пойти дальше и научить машину выполнять еще и креативные задачи? Есть мнение, что творчество это единственное преимущество, на которое может рассчитывать человечество в конкуренции с машинным обечением. Некоторые эксперты считают, что креативные процессы являются одними из самых тяжело автоматизируемых. Но на сегодняшний день искусственный интеллект сочиняет музыку, пишет картины и литературные произведения.

Что же нужно, что бы обучить машину? Что бы стать профессионалом в той или иной области нужно потратить не менее 10000 часов, что удержать и закрепить новую привычку – 21 день. А с обучение машин все обстоит по другому. И тут важны три основные компонента: данные, характеристики и алгоритм.

Данные. Хотим распознавать котов – нужны фотографии котов, хотим фильтровать спам – нужны примеры спам-писем. Причем фотографий и примеров спам-писем нужно очень много: несколько десятков тысяч. Как собирать данные? Кто то делает это вручную, а кто то использует своих пользователей для этого(хороший пример: каптча на некоторых сайтах). В процессе обучения может скапливаться огромное количество данных. Известно, что за сутки машина может запоминать до ста миллионов новых слов. Кстати таким образом в свое время были оцифрованы все выпуски газета New York Times и многие книги постигла та же участь. Теперь они все доступны на сервисе «Google Книги». Чистые наборы данных, или DataSet – это очень важно.

Признаки, или feature, или свойства, или характеристики. Ими может быть все, что угодно: марка автомобиля, цены на акции, частота употребления определенного слова – машине важно на ей конкретно нужно смотреть. Удобно когда данные уже лежат в таблицах, ведь название колонок это и есть признаки. В случае размеченных данных машине всегда действовать легче. Такой подход называется «обучение с учителем».

Но бывают ситуации, когда машине просто вываливают много данных. Данные не размечены и машина сама пытается найти закономерности. Такой подход в машинном обучении называется «обучение без учителя». В таком случае поиск и отбор признаков занимает намного больше времени, чем все остальное обучение. И то не факт, что все будет работать хорошо. А иногда машине не может найти закономерности в данных.

Обучение с частичным привлечение учителя – золотая середина. Обучающий датасет содержит в себе как размеченные данные, так и не размеченные. Этот метод полезен, если из датасета трудно извлечь все важные признаки или разметить все объекты очень трудоемкая задача. Этот метод используется для анализа медицинских изображений. Например сканы компьютерной томографии или МРТ. Рентгенолог может разметить небольшое количество сканов с опухолями или заболеваниями. Но вручную разметить большое количество сканов трудоемкая и дорогостоящая задача. Однако машина может извлечь информацию из небольшого количества размеченных данных и улучшить точность предсказаний на неразмеченных данных.

Популярный метод обучения, требующий небольшой набор размеченных данных, использует генеративно-состязательной сети или GAN. GAN можно представить как соревнование двух нейронных сетей, где каждая пытается переиграть другую. Одна сеть – это генератор, создает новые объекты данных, имитирующие обучающую выборку. А другая сеть,

дескриминатор, оценивает, реальные эти данные или поддельные. Сети взаимодействуют друг с другом и циклично совершенствуются. Генератор пытается создавать убедительные подделки, а дискриминатор старается лучше отличать подделки от оригинала.

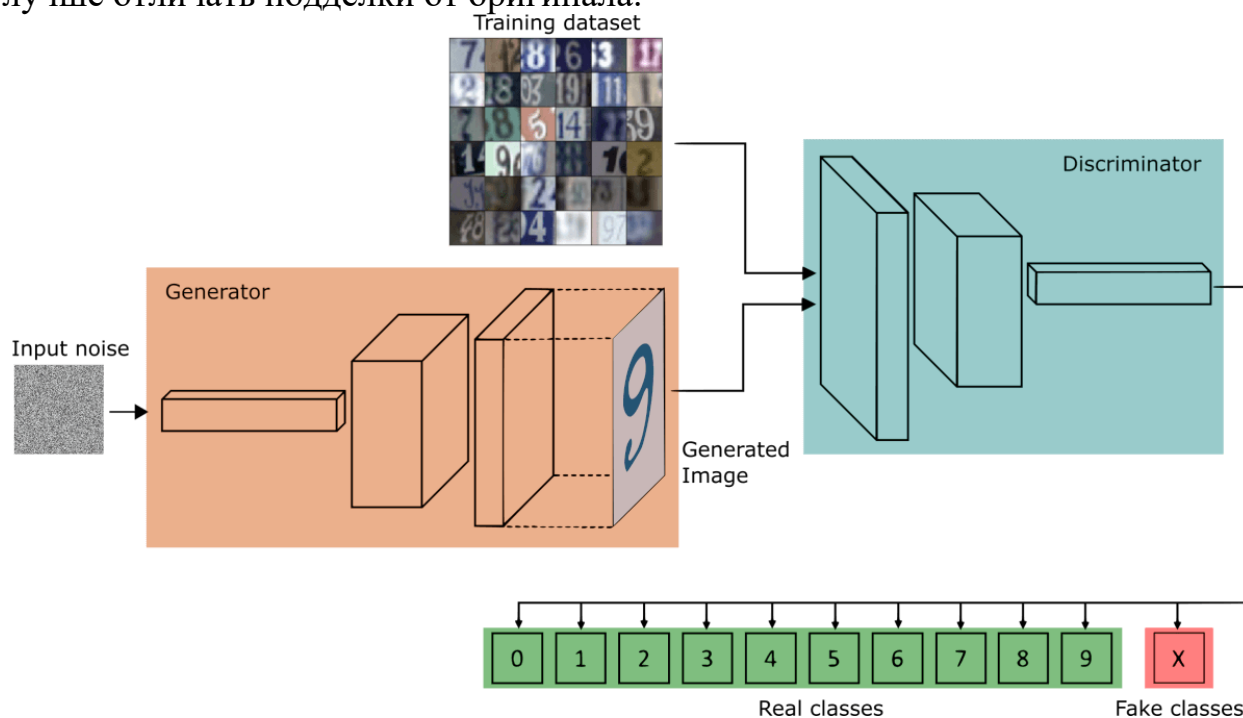


Рисунок 1 Как работает GAN

Машинное обучение и нейронные сети идут семимильными шагами в промышленность, медицину, в повседневную жизнь людей. С их помощью автоматизируются производства, диагнозы больным ставятся с высокой точностью и на ранних стадиях развития, упрощается выполнение повседневных задач.

Список литературы

1. Машинное обучение – это легко [Электронный ресурс] // Хабр: Сообщество IT-специалистов – Режим доступа: <https://habr.com/ru/post/319288/> – (Дата обращения: 01.02.2020).
2. Применение машинного обучения и Data Science в промышленности [Электронный ресурс] // Хабр: Сообщество IT-специалистов – Режим доступа: <https://habr.com/ru/company/mailru/blog/462769/> – (Дата обращения: 01.02.2020).
3. Обзор задач компьютерного зрения в медицине [Электронный ресурс] // Хабр: Сообщество IT-специалистов – Режим доступа: <https://habr.com/ru/post/309152/> – (Дата обращения: 05.02.2020).