Современные задачи машинного обучения*

М. С. Ананьевский
msaipme@yandex.ru

09 марта 2022 г.

1°. Предпосылки

За последние двадцать лет в компьютерной технике произошел значительный количественный рост в вычислительных мощностях и системах хранения данных. Этот рост обеспечил возможность перехода к практическому решению качественно новых задач.

По сравнению с началом 2000-х годов емкость накопителей на жестких магнитных дисках (HDD) увеличилась в тысячу раз: с десятков гигабайт до десятков терабайт (таблица 1).

Год	Емкость	Модель жесткого диска
2000 г.	0,04 Тбайт	HDD Western Digital Caviar WD400BB
2022 г.	20,00 Тбайт	HDD Western Digital Gold WD201KRYZ

Таблица 1. Емкость наиболее вместительных накопителей на жестких магнитных дисках (HDD) находящиеся в свободной продаже и доступные по цене массовому потребителю.

Рекорд по скорости печати текста на клавиатуре сегодня составляет 1060 знаков в минуту. Если непрерывно печатать с этой рекордной скоростью 100 лет, то совокупный объем созданных данных составит всего 0,052 терабайта. Если учесть реальную скорость печати и возможности по сжатию текстовых данных, то скорее всего, для хранения текстовой части всех электронных писем и всех сообщений в мессенджерах всех россиян за один календарный год будет достаточно пары дисков.

*Семинар по оптимизации, машинному обучению и искусственному интеллекту «O&ML» `http://www.apmath.spbu.ru/oml/`

Одного единственного диска теперь достаточно для хранения всего аудиомира человека, всего что он слышит и произносит на протяжении всей своей жизни, с рождения до самой смерти (таблица 2). Цена такого диска тоже не высока, примерно 60.000 рублей, что соответствует одной средней месячной заработной плате по Санкт-Петербургу.

Хранение телефонных разговоров в масштабах страны теперь тоже возможно: за 2021 год объем разговоров россиян по мобильному телефону составил менее 500 млрд минут, что составляет менее 120.000 терабайт (6000 дисков) при качестве 32 Кбит/сек или менее 480.000 терабайт (24000 дисков) при качестве 128 Кбит/сек (без сжатия данных).

Качество записи	Объем данных за 100 лет
32 Кбит/сек	12 Тбайт
48 Кбит/сек	19 Тбайт
64 Кбит/сек	25 Тбайт
128 Кбит/сек	50 Тбайт

Таблица 2. Объем генерируемых данных в зависимости от качества записи (без сжатия данных).

Вычислительные мощности процессоров общего назначения с начала 2000-х годов выросли в 500 с лишним раз (таблица 3). В разы увеличилась рабочая частота, появилась многопоточность. Изменилась разрядность вычислений: произошел массовый переход с 32-х на 64-х битные архитектуры. Набор инструкций микропроцессоров обогатился большим количеством новых, сложных операций, в том числе с разрядностью 512 бит. Объем физической оперативной памяти адресуемый одним процессором сегодня может составлять восемь терабайт.

Год	Число потоков	Частота	Модель процессора
2000	1	0,60 Ггц	Celeron Pentium 3
2022	128	2,45 Ггц	AMD EPYC 7763

Таблица 3. Наиболее производительные микропроцессоры общего назначения (архитектура x86) находящиеся в свободной продаже и доступные по цене массовому потребителю.

Еще более впечатляющим выглядит рост производительности суперкомпьютеров: с начала 2000-х годов она выросла в 140.000 раз (таблица 4). Это

стало возможным благодаря научно-техническому прогрессу сразу по нескольким направлениям:

- скорость передачи данных: пропускная способность современных сетевых интерфейсов может достигать 200 Гбит/с (200GbE);
- система управления суперкомпьютером: современное программное обеспечение позволяет управлять одновременной согласованной работой более чем 100.000 физических узлов, объединять их в сеть со сложной топологией для эффективного обмена данными, а также создавать среду и предоставлять инструменты для эффективного распределения вычислений (разница между теоретической и практической производительностью за 20 лет сократилась с 50% до 20%, и это при том, что количество узлов выросло в 150 раз);
- производительность процессоров: за 20 лет производительность процессоров общего назначения выросла в 500 раз, появились специализированные вычислительные модули, заточенные под параллельные вычисления и обладающие огромной производительностью в десятки терафлоп в секунду (можно установить несколько в одну материнскую плату).

Самый производительный на сегодняшний день суперкомпьютер Frontier при полной загрузке потребляет 21 мегаватт электроэнергии. Эксплуатировать суперкомпьютеры – дорого, строить и разрабатывать их – очень дорого. Стремительное развитие этой отрасли свидетельствует о мощном экономическом базисе, о том, что в развитых странах промышленность совершает очередной эволюционный переход, и суперкомпьютеры становятся для нее необходимым инструментом, без которого уже невозможно достичь нового технологического уровня производства в самых разных отраслях народного хозяйства.

Год	Число ядер	Ггц	Пфлоп/с	Процессор	Название
2000	8192	0,375	0,012	SP Power3	ASCI White
2021	158976 * 48	2,200	537,21	A64FX	Fugaku
2022	136408 * 64	2,000	1685,65	AMD EPYC, AMD Instinct MI250X	Frontier

Таблица 4. Наиболее производительные суперкомпьютеры.

Отдельные специализированные вычислительные модули, находящиеся в свободной продаже и доступные по цене массовому потребителю, уже превосходят по мощности самые производительные суперкомпьютеры начала 2000-х

годов (таблица 5). Их массовое производство означает, что потребность в вычислительных мощностях есть не только у крупного, но и у мелкого бизнеса.

Модель	Ядер	Ггц	Одинарная точность	Двойная точность
Nvidia H100	16896	1,605	60000 Гфлоп/с	30000 Гфлоп/с

Таблица 5. Специализированные вычислительные модули находящиеся в свободной продаже и доступные массовому потребителю.

Увеличение вычислительной мощности в 140.000 раз означает, что если алгоритм допускает ускорение путем распределения вычислений, то тогда то, для чего 20 лет назад требовалось три месяца счета, сегодня можно посчитать за одну минуту, а за те же три месяца перебрать 140.000 различных вариантов с разными входными данными. Такие возможности существенно расширили круг задач, для которых можно не только теоретически, но и практически найти решение методами численного моделирования или грубым (умным) перебором вариантов.

Изобилие и массовая доступность вычислительных мощностей во многих случаях позволили переложить сложность аналитического нахождения точного решения на численные расчеты, путем построения систем общего вида и их последующего обучения (оптимизации) под решение выбранного класса задач. Этот подход получил название “машинное обучение”.

2°. Влияние машинного обучения на экономику

Прогресс в технологиях машинного обучения создает предпосылки для глубокой автоматизации мировой экономики и значительного прироста производительности труда. Массовое внедрение систем искусственного интелекта в технологические и производственные процессы может оказать в ближайшее десятилетие шоковое влияние на экономики многих стран. Точные экономические последствия предсказать проблематично, тем не менее выделяют пять основных эффектов [1, 2]:

- общий рост производительности труда;
- изменение списка востребованных профессиональных навыков на рынке труда, включая общее увеличение потребности в высококвалифицированных технических специалистах;

- неравномерное распределение эффекта среди разных секторов экономики, уровня зарплат, кадровой подготовки, групп профессий, географического местоположения;
- перестройка рынка труда: некоторые профессии исчезнут, некоторые появятся;
- рост технологической безработицы в краткосрочной, а возможно и в долгосрочной перспективе.

Существует большая неопределенность в том, насколько сильными и быстрыми окажутся эти эффекты. Возможно, что изменения будут медленными, без дестабилизации экономических систем, или наоборот, они будут быстрыми и вызовут экономический шок.

Автоматизация технологических и производственных процессов позволяет наращивать производственные мощности, улучшать качество продукции, ускорять и оптимизировать процессы управления, повысить безопасность труда и отстранить человека от выполнения опасных работ. Для многих задач автоматизация позволяет достигать показателей, которые в принципе невозможно достичь ручным трудом. Исторический опыт свидетельствует, что автоматизация также повышает производительность труда: в США в 1900 году на сектор сельского хозяйства приходилось около 40% рабочих мест, к 2000 году этот показатель снизился до 2%.

Существующие на сегодняшний день технологии (таблица 6) позволяют полностью автоматизировать менее 5% профессий в США. Детальный анализ [3] свидетельствует о широких возможностях для частичной автоматизации: для 60% должностей можно автоматизировать 30% их рабочих обязанностей. Наиболее благоприятными для автоматизации являются профессии связанные со сбором и обработкой данных (потенциал автоматизации до 70%) и физическая работа в предсказуемом окружении (потенциал автоматизации до 80%).

3°. Методы машинного обучения

Общей характерной чертой методов машинного обучения является непрямое решение задачи, то есть система программируется не на решение конкретной задачи, а на обучение решению множества сходных задач. Наиболее активно в последнее десятилетие развиваются подходы основанные на искусственных нейронных сетях.

67%	Распознавание известных образов и категорий	высокий
46%	Формирование сообщений на естественном языке	средний
41%	Сенсорное восприятие	средний
38%	Информационный поиск	высокий
35%	Понимание естественного языка	низкий
17%	Передвижение предметов	высокий
15%	Презентации, выступления	средний
13%	Социальное и эмоциональное распознавание	низкий
13%	Логический вывод и решение задач	низкий
12%	Оптимизация и планирование	высокий
11%	Манипулирование предметами, требующее ловкости и ощущения	средний
10%	Координация в мультиагентном окружении (работа в команде)	низкий
10%	Социальное и эмоциональное окрашивание сообщений	низкий
9%	Социальный и эмоциональный анализ	низкий
4%	Навигация и ориентирование	высокий
4%	Передвижение по местности	низкий
2%	Креативность	низкий
2%	Создание новых образов, категорий, гипотез	низкий

Таблица 6. Процент рабочего времени приходящийся на деятельность, требующую высокого или среднего уровня человечности (для США) и уровень реализации в существующих коммерческих, научно-исследовательских или опытно-конструкторских системах 6.

Математическая составляющая проблемы обучения искусственных нейронных сетей может быть сформулирована как решение задачи оптимизации следующего функционала:

$$Q(w) = \sum_{x \in \mathbb{X}_T} ||f(x,w) - f_*(x)||^2 \to \min_w \tag{1}$$

если отклонение от целевого значение оценивается по методу наименьших квадратов, или

$$Q(w) = \sum_{x \in \mathbb{X}_T} H(f(x,w)|f_*(x)) \to \min_w \tag{2}$$

если выход нейронной сети интерпретируется, как функция распределения вероятностей, а отклонение от целевого значения оценивается по расстоянию Кульбака-Лейблера. Здесь $H(\cdot|\cdot)$ – относительная энтропия, $\mathbb{X}_T$ – обучающая выборка, w – настраиваемые параметры нейронной сети, $f(x,w)$ – значение выхода нейронной сети с параметрами w и входом x, $f_*(x)$ – идеальное (желаемое) значения выхода нейронной сети для входа x.

Особенностью постановки задачи является объем тренировочной выборки $\mathbb{X}_T$. Например, база ImageNet, используемая для сравнения и тестирования алгоритмов компьютерного зрения и классификации изображений, насчитывает 14 миллионов картинок разбитых на 21 тысячу классов. Это означает, что размер выходного вектора равен 21 тысяче, а число слагаемых в функционале превышает 10 миллионов. Количество настраиваемых параметров (размерность вектора w) обычно исчисляется сотнями тысяч. А если учесть, что функция $f(\cdot,\cdot)$ – нелинейная и определяется довольно сложной структурой искусственной нейронной сети, то задача минимизации такого функционала представляет серьезные вычислительные трудности, он даже не влезает в оперативную память сервера.

Наиболее распространенным алгоритмом используемым для минимизации таких функционалов является метод Adam [4], суть которого заключается в использовании алгоритма стохастического градиента с адаптивно настраеваемым шагом. Поиск эффективных алгоритмов обучения искусственных нейронных сетей является важной открытой задачей.

Самая крупная искусственная нейронная сеть 2020 года, GPT-3, была разработана для обработки текстовой информации на естественном языке: генерация текстов, поддержание диалогов в чатах, ответы на вопросы и т.п. Мощность обучающей выборки для этой сети превысила 500 миллиардов токенов, а количество настраиваемых параметров оказалось больше 175 миллиардов.

Сложной теоретической и практической задачей является создание программного обеспечения предназначенного для обучения искусственных нейронных сетей, позволяющее эффективно распределять вычисления между большим количеством узлов со специализированными вычислительными модулями. Доминирующую роль в этом направлении сегодня занимает американская компания Google LLC со своим проектом TensorFlow.

4°. Данные это новая нефть

Для обучения нужна информация, поэтому сбор, подготовка, разметка данных является важной частью построения систем искусственного интеллекта. Подобно тому, как экономика не может работать без энергетики, системы искусственного интеллекта не могут работать без данных, данные это новая нефть, питающая экономику индустрии 4.0, результат четвертой промышленной революции. За обладание данными идет очень жесткая конкурентная война, данные собирают, данные крадут, данные “приземляют”, а государства активно принимают законы ограничивающие возможности для сбора и накопления данных.

Интернет является мощным источником размеченных данных для обучения. Особенно стоит отметить проект reCAPTCHA принадлежащий компании Google LLC. Этот проект, решая задачу защиты сайтов от ботов, фактически осуществляет ручную разметку данных (текстовых и графических), не платя пользователям ни копейки. В 2010 году ежедневно осуществлялось более 100 миллионов показов reCAPTCHA, сложно представить в какую астрономическую сумму обошлась бы разметка такого объема информации классическим ручным способом.

В открытом доступе хорошо подготовленных и размеченных массивов данных очень мало, а те, что есть, в основном предназначены для использования в учебно-научных целях [5]. Отдельно хотелось бы отметить два проекта: MNIST и MovieLens.

База данных MNIST содержит 60 тысяч картинок рукописных цифр, с размером изображений 28 на 28 пикселей. Это очень маленькая база, которая отлично подходит для учебных целей. Классификаторы получаются небольшими, и для их обучения не нужны огромные вычислительные мощности, достаточно старого ноутбука. При этом, она достаточно наглядна, и после обучения классифкатора можно попробовать распознать свои рукописные цифры и проанализировать качество распознавания (часто результаты очень поучительны для студентов).

База данных MovieLens (GroupLens) содержит 26 миллионов оценок и 750 тысяч тегов для 45 тысяч фильмов, поставленных 270 тысячами пользователей. Это прекрасная база для обучения построению рекомендательных систем, так как при работе с ней приходится преодолевать много практических проблем, таких как пропуск данных, несбалансированность данных и т.д.

Некоторые компании проводят открытые конкурсы по построению систем искусственного интеллекта в направлении своих коммерческих интересов. При этом они выкладывают небольшие массивы данных (под разными ограничительными лицензиями), на которых и проводят соревнования, выясняя чье решение покажет лучший результат. Наиболее крупной площадкой для соревнований такого рода является Kaggle.com [6]. Призовой фонд в этих конкурсах может превышать миллион долларов, что на самом деле, является весьма небольшой суммой по сравнению с выгодой, которую компания может получить от нахождения новых эффективных решений или найма высококвалифицированного специалиста.

Малый объем накопленных данных, и в первую очередь малый объем размеченных данных, может оказаться критическим для развития методов машинного обучения и систем искусственного интеллекта. На сегодняшний день, мировым лидером по объему накопленных данных и по умению работать с ними, по-видимому, является американская компания Google LLC.

ЛИТЕРАТУРА

1. John P. Holdren, Megan Smith *Preparing for the Future of Artificial Intelligence* Executive Office of The President of The United States, Washington, D.C., 20502, October 12, 2016.

2. Jason Furman, John P. Holdren, Cecilia Munoz, Megan Smith, Jeffery Zients *Artificial Intelligence, Automation, and the Economy* Executive Office of The President of The United States, Washington, D.C., 20502, December 20, 2016

3. Jacques Bughin, James Manyika, Jonathan Woetzel *A Future that Works: Automation, Employment, and Productivity* McKinsey Global Institute, January 2017

4. Diederik P. Kingma, Jimmy Ba *Adam: A Method for Stochastic Optimization* Advances in Neural Information Processing Systems 28 (NIPS 2015), 2015.

5. *Список открытых баз данных для исследований в области машинного обучения и искусственного интеллекта* https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_research

6. *Соревнования по построению систем искусственного интеллекта* https://www.kaggle.com/competitions