Машинное обучение

Как построить надежные модели искусственного интеллекта

Яда Пруксачаткун
Мэтью Макатир
Субхабрата Маджумдар

УДК 004.8
ББК 32.813
П85

Practicing Trustworthy Machine Learning: Consistent, Transparent, and Fair AI Pipelines

Matthew McAteer, Subhabrata Majumdar, Yada Pruksachatkun

Во внутреннем оформлении использована фотография:
Irina Bg / Shutterstock / FOTODOM
Используется по лицензии от Shutterstock / FOTODOM

Пруксачаткун, Яда.

П85 Машинное обучение. Как построить надежные модели искусственного интеллекта / Яда Пруксачаткун, Мэтью Макатир, Субхабрата Маджумдар: [перевод с английского М. Стефанец]. — Алматы : Астана иностранная пресса, 2026. — 304 с. — (O'Reilly. Книги по программированию).

ISBN 978-601-12-6019-0

Это практическое руководство по созданию устойчивых, безопасных и понятных ML-систем. Авторы рассматривают ключевые аспекты разработки надежных моделей: от выявления уязвимостей и предвзятости до оценки прозрачности алгоритмов, защиты от атак и управления долговыми обязательствами в ML-проекте.

Книга помогает понять, как действуют современные подходы к честности, интерпретируемости и безопасности, и показывает, как применять их в реальных условиях — там, где модели сталкиваются с изменчивой средой, шумными данными и человеческими сценариями использования.

УДК 004.8
ББК 32.813

ISBN 978-601-12-6019-0

Оглавление

Предисловие

В современном мире системы искусственного интеллекта используются в ключевых сферах деятельности: медицине, юриспруденции, обороне и безопасности. Решения, которые мы принимаем на основе ответов нейросетей, приносят как огромные прибыли, так и серьезные убытки. Ставки высоки, как никогда, поэтому крайне важно иметь надежные модели машинного обучения. Мы сталкиваемся с проблемой: нестабильные и непредсказуемые системы ошибаются, выдают принципиально разные результаты в схожих ситуациях и не могут объяснить свои суждения в понятной человеку форме. Эта книга расскажет, как построить модель, которая будет отвечать актуальным требованиям не только цифрового, но и реального мира.

Внедрение машинного обучения в бизнес-процессы

Если вы держите в руках эту книгу, то, вероятно, уже осознали всю важность технологии машинного обучения. Методы ML проникают во все уголки нашей жизни, независимо от области их применения. Сооснователь Google Brain Эндрю Ын метко сравнил ИИ с «новым электричеством» *(https://www.wipo.int/en/web/wipo-magazine/articles/artificial-intelligence-the-new-electricity-55628)*. По сути, мы имеем дело с универсальным аппроксиматором функции. Однако, если не знать, как правильно использовать этот инструмент, он может быть не менее опасен, чем электричество. Попадая на линию электропередачи, фольгированные воздушные шары загораются и повреждают провода. Последствия ошибок моделей могут быть столь же разрушительны.

Развертывание ML-приложении в реальном мире в корне отличается от работы над моделями в изолированной среде. Датасеты для обучения обычно не охватывают весь спектр реальных данных. Информация, с которой модели придется иметь дело в будущем, может отличаться от той, что ей уже знакома, особенно если тренировочная выборка была подготовлена недостаточно тщательно. Если модель обучена на искаженных в ту или иную сторону данных, у ее пользователя возникают этические и юридические риски. Ситуация усугубляется, когда к тому же нет возможности объяснить решения

ML-модели в понятной форме. Даже если эта опасность миновала, надвигается другая. С каждым годом киберугрозы становятся все более изощренными. Возможно, когда-нибудь одного запроса к рабочей модели будет достаточно, чтобы украсть конфиденциальную информацию.

Однако не все прогнозы столь пессимистичны. Нам уже доступны передовые методы управления данными — реальными и сгенерированными. Существуют разные способы измерить, насколько вновь поступившие данные отличаются от уже имеющихся у нас. Можно научиться выявлять и устранять систематическую предвзятость, объяснять и интерпретировать модели искусственного интеллекта. Что касается безопасности и надежности, крупнейшие компании в области машинного обучения выпускают инструменты для защиты моделей от внешних угроз.

Эта книга станет вашим помощником в деле «восстановления линии электропередачи». В ней мы рассматриваем все решения — от стандартных до самых современных.

Распространение моделей на основе трансформеров

В конце 2010-х — начале 2020-х годов, когда мы еще не начали писать эту книгу, трансформеры, представляющие собой разновидность архитектуры моделей глубокого обучения, уже активно использовались для обработки естественного языка. Пока мы работали над книгой, эта практика стремительно распространялась. Трансформеры быстро нашли применение в различных областях, включая компьютерное зрение, обработку табличных данных и даже обучение с подкреплением. Пятнадцать лет назад глубокое обучение выглядело иначе. Каждая задача и сфера применения требовали настолько уникальной архитектуры нейронной сети, что эксперту по компьютерному зрению было сложно до конца понять принципы работы с естественным языком. И наоборот, исследователям в области NLP было трудно разобраться в методах компьютерного зрения, основанных на глубоком обучении.

Трансформеры были представлены в 2017 году в статье Attention Is All You Need («Все, что вам нужно, — внимание»)[1]. Классические рекуррентные (RNN) и сверточные (CNN) нейронные сети оперируют последовательностями входных данных, обрабатывая их небольшими пакетами. Трансформер же, используя механизм внимания, устанавливает связи между всеми элементами данных одновременно. Как следствие, эта нейросеть может делать выводы на основе всего набора данных, на котором обучается.

Ключевое достоинство трансформера — способность устанавливать связи между точками данных во всем датасете. Именно эту модель используют для поиска ответов на вопросы, прогнозирования текста и машинного

перевода. В последнее время интерес к трансформерам вышел за рамки NLP и распространился на область компьютерного зрения, в частности, в задачах классификации изображений[2]. Распространение трансформеров — новое явление, которое продолжит влиять на рынок.

Хотя трансформеры не следует использовать для решения всех задач без исключения — во многих случаях лучше работают менее затратные по вычислительным ресурсам и памяти методы, — именно модель трансформеров легла в основу книги, так как на эту архитектуру опираются почти все достижения в глубоком обучении за последние годы.

Бум больших высокопроизводительных ML-моделей

Трансформеры не просто получили широкое распространение: они позволили людям приобрести доступ к системам искусственного интеллекта, которые десять лет назад казались фантастикой. В 2019 году компания OpenAI представила GPT-3 — языковую модель, способную генерировать тексты, практически неотличимые от написанных человеком. Новые возможности нейросетей продолжают находить даже сейчас, когда компании создают продукты на базе этих моделей[3]. Например, открытием 2022 года стал прием, позволяющий значительно увеличить производительность GPT-3. Способность модели решать сложные задачи сравнили на двух математических бенчмарках — MultiArith и GSM8K. Точность возросла с 17,7 до 78,7% и с 10,4 до 40,7% соответственно. Как удалось добиться таких показателей? С помощью фразы Let's think step by step, предваряющей каждый вопрос[4]. Странности на этом не заканчиваются. Формирование такого промпта может привести к тому, что модель примется выводить цепочку рассуждений, которая не обязательно закончится ответом. Чтобы получить корректное заключение, понадобятся дополнительные промпты и запросы[5, 6].

Во время написания этой книги появилась еще одна модель искусственного интеллекта — StableDiffusion. С ее помощью текстовый запрос можно преобразовать в изображение. Эта модель была обучена аналогично нейросетям DALL·E 2 (OpenAI), Imagen (Google), Parti (Google) и MidJourney *(https://openai.com/index/dall-e-2/)*, поэтому качество генерируемых изображений было сопоставимым. В отличие от предшествующих моделей исходный код и веса StableDiffusion были опубликованы в открытом доступе, что стало проблемой для сообщества, занимающегося вопросами безопасности систем машинного обучения. Публикация противоречила принципу предосторожности в разработке технологий, согласно которому высокопроизводительные ML-модели не должны внедряться в широкую практику, пока не оценено их потенциальное воздействие и не подтверждена безопасность. Одновременно со StableDiffusion разработчики выпустили различные инструменты,

призванные обеспечить безопасность использования модели[7, 8]. Хотя этот последний шаг стоит поддержать, ситуация в целом указывает на недостаточность инициатив по обеспечению безопасности в области машинного обучения, даже для моделей с существенно меньшим уровнем риска.

Мы видим, как конкурирующие компании и команды — Google, DeepMind, OpenAI и Microsoft — активно создают схожие между собой модели для работы с изображениями и текстом. Эти проекты развиваются параллельно, поэтому можно сказать, что прогресс не ограничен недостатком новых идей. Между тем, создаются условия для нечестной игры. Кто-то может попытаться обойти конкурентов, не встраивая разумных ограничений в свою модель. Часто в крупных компаниях вопросы безопасности отнимают время на разработку продуктов, а инженеры из этих компаний могут уйти в стартапы, которые готовы быстрее воплощать идеи в жизнь. Поскольку похожие проекты разрабатывают одновременно, секретность больше не обеспечивает такого уровня защиты, как раньше.

Один из самых перспективных способов обеспечивать безопасность для компаний — поддерживать прозрачность и открыто обсуждать риски и варианты их устранения[9]. Об этом также пойдет речь в книге.

Зачем мы написали эту книгу

Как исследователи в области машинного обучения и инженеры ML-систем для бизнеса мы заметили: между созданием начальной ML-модели для статического датасета и ее внедрением существует большой разрыв. Основная причина — недостаток надежности моделей. Системы хорошо работают в тестовой среде, но на боевом сервере возникают критические ошибки. Многие крупные компании создают отделы по безопасности ИИ-проектов для анализа потенциальных рисков и последствий ошибок ML-систем — как уже существующих, так и разрабатываемых[10]. К сожалению, позволить себе целый отдел могут далеко не все компании. Если такой отдел и существует, средств на его финансирование, как правило, недостаточно, а циклы разработки моделей слишком коротки, чтобы выполнять полномасштабные проверки. Конкурент всегда может первым выйти на рынок с аналогичной моделью.

Мы написали эту книгу, чтобы доступно объяснить, как строить надежные ML-модели. Материал не претендует на значительную новизну. Мы хотели, чтобы книга была понятна людям без опыта исследований в сфере машинного обучения, закладывала основу знаний, помогала определить степень надежности моделей, а также описывала методы ее повышения. В книге вы найдете фрагменты кода, который можно использовать в собственных проектах; ссылки на проекты и ресурсы с открытым исходным кодом; ссылки на примеры с кодом и уроки, многие из которых можно проходить прямо в браузере.

Опыт — бесценный ресурс, но, чтобы его получить, нужно знать, с чего начать. С помощью этой книги вы сможете развернуть ML-приложения в реальном мире — шумном, сложном и порой опасном. Эта книга обязана своим появлением многочисленным исследователям, инженерам и другим специалистам. Надеемся, что совместными усилиями мы поможем командам разработчиков развернуть ML-системы в рабочей среде.

Для кого эта книга

Если вы участвуете в разработке моделей искусственного интеллекта, если вам важно не нанести непреднамеренного вреда людям, обществу или окружающей среде, — эта книга для вас.

Книга предназначена для специалистов по работе с данными и инженеров с базовыми знаниями об искусственном интеллекте. Отдельные разделы заинтересуют персонал и руководителей, которые разбираются в теме в общих чертах. Информация окажется особенно ценной для разработчиков, которые стремятся повысить компетентность и приобрести новые навыки в области надежного машинного обучения. Мы предполагаем, что читатели знакомы с основами глубокого обучения и языком программирования Python.

Одного прочтения инженерам будет достаточно, чтобы получить базовое представление о «надежности» систем машинного обучения. Всегда можно вернуться к материалу, адаптировать фрагменты кода и применить их для проверки надежности собственных моделей.

Выравнивание и безопасность в области ИИ

Существует целое направление, посвященное вопросам выравнивания и безопасности искусственного интеллекта. *Выровнять ИИ под наши потребности* — значит настроить систему так, чтобы она выполняла задачи в соответствии с ожиданиями человека без непредвиденных последствий. Выравнивание (или *согласование*) — часть более широкой проблемы *безопасности ИИ*, которая охватывает ряд потенциальных проблем. К ним относятся: поддержка социальных предрассудков, использование искусственного интеллекта в военных целях, мошеннических схемах или кибератаках, а также нарушение этики и нежелательное поведение, выходящее за рамки установленных социальных норм.

Выравнивание ИИ рассматривается как *способ устранения* этих рисков, поскольку подразумевает, что система в результате будет разделять человеческие ценности. В научной среде существует множество исследований, посвященных вопросам надежного машинного обучения как такового.

Одна из ключевых сложностей в изучении искусственного интеллекта — несоответствие между теретическими концепциями и практической

реализацией. Исследователи используют понятия из психологии (*намерение, желание, цель, мотивация*) и философии (*система ценностей, практическая польза*), но эти категории неприменимы в случае систем искусственного интеллекта. Возможно, когда-нибудь удастся создать ИИ, который точно имитирует работу человеческого мозга на уровне нейронов и синапсов. В таком случае философские и психологические идеи придутся кстати. Однако, как показывает опыт работы одного из авторов в сфере нейробиологии, современные нейросети устроены иначе. Биологические нейроны функционируют не как логические элементы — их поведение описывается тысячами взаимосвязанных нелинейных уравнений. Смоделировать такой нейрон — уже сложная задача, для которой нужна отдельная нейросеть, а не одна весовая переменная[11]. Гарантировать, что ИИ никогда не нанесет вред, пока невозможно, но уже сейчас Cohere, OpenAI, AI21 Labs[12] и другие крупные игроки активно работают над снижением рисков и внедряют лучшие практики.

Еще одна сложность связана с тем, что в исследованиях безопасности ИИ часто обсуждаются гипотетические сценарии. Например, рассматривается искусственный интеллект общего назначения — AGI (Artificial General Intelligence) — и самосовершенствующиеся системы[13, 14]. Однако эти темы уже находят отражение в реальности. Пока мы писали книгу, OpenAI представила модель DALL·E 2, которая создает изображения по текстовому описанию, а DeepMind — Gato, универсальный трансформер для задач обработки языка, компьютерного зрения и обучения с подкреплением[15]. После этих релизов прогнозы о сроках появления общего ИИ сдвинулись в сторону более ранних дат[16]. Теперь гораздо проще представить худшие сценарии, в которых цели сильного ИИ идут вразрез с нашими.

Безопасность искусственного интеллекта — это не вопрос далекого будущего, а серьезная проблема настоящего. Уже сегодня мы сталкиваемся с возможными угрозами: мгновенными обвалами на финансовых рынках из-за торговых ботов[17]; химическим оружием на основе моделей для поиска лекарств (достаточно изменить один знак в формуле)[18]; беспилотными дронами с системой распознавания лиц, которые потенциально могут использоваться для этнических чисток[19]. Рассуждения на высокие философские темы типа поведения сверхумных автономных систем не имеют смысла, пока нет конкретных методов диагностики и исправления ошибок.

Неясно, приведет ли развитие ИИ к сценарию наподобие «собирателя скрепок»[20]. Однако этот мысленный эксперимент иллюстрирует экзистенциальный риск, который система с неограниченными возможностями и отсутствием понимания человеческих ценностей может нести для людей.

В статье Set Sail For Fail? On AI Risk («Беда или победа? О рисках ИИ») в блоге Nintil Хосе Луис Рикон Фернандес де ла Пуэнте дает критическую оценку исследований по безопасности искусственного интеллекта (*https://nintil.com/ai-safety*).

Поскольку дискуссии по теме не утихают, в этой книге мы собрали практические инструменты и примеры с кодом для инженеров. Мы не пытались дать однозначное определение надежности и предоставили читателю самому решить, какой смысл он вкладывает в это понятие. Мы указали ошибки, которые снижают надежность систем искусственного интеллекта, и наметили пути избавления от таких ошибок. Необходимо разрабатывать решения, позволяющие устранить ограничения моделей, которые уже выполняют важные функции в жизни людей.

Хорошая новость: многие из этих проблем можно решить. Задача сложная, но вполне реальная[21].

Модели искусственного интеллекта на PyTorch с платформы HuggingFace

Фрагменты кода в книге написаны на основе библиотеки трансформеров HuggingFace. Большинство примеров описывают модели на PyTorch. Этот фреймворк компании Meta[22, 23] опирается на те же математические формулы, что и TensorFlow или JAX. Хотя код для других фреймворков может отличаться, принципы работы остаются неизменными.

HuggingFace стала востребованным инструментом для обмена параметрами моделей ИИ. Первоначально библиотека была ориентирована на языковые модели, позднее — на модели компьютерного зрения, преобразования текста в изображения, обработки звука и обучения с подкреплением[24].

Перед загрузкой любой модели с платформы HuggingFace убедитесь, что файл безопасен и не содержит вирусов или вредоносного кода. Ранее платформа использовала модуль pickle на Python для загрузки моделей. Видеоблогер Янник Килчер на своем канале показал (*https://www.youtube.com/watch?app=desktop&v=2ethDz9KnLk&feature=youtu.be*), что pickle позволяет хранить произвольный код, который может представлять угрозу для пользователя, его данных или устройств. Возможная схема атаки была продемонстрирована на примере модели totally-harmless model (с англ. «точно не вредоносная модель») *(https://huggingface.co/ykilcher/totally-harmless-model).*

Проблему решили с помощью патча torch-save (*https://github.com/yk/patch-torch-save*). После выхода видео на HuggingFace также появилось предупреждение о возможном выполнении произвольного кода. Всегда проверяйте файлы перед загрузкой и используйте только проверенные источники.

Основные понятия

Чтобы помочь вам извлечь максимум пользы из этой книги, мы собрали ключевые термины с пояснениями и ссылками для дальнейшего изучения.

Эмбеддинги

Эмбеддинг — векторное представление слов. Каждое слово преобразуется в вектор, который отражает семантические свойства такого слова. Основные алгоритмы включают GloVe и Word2Vec.

Языковые модели

Языковые модели — алгоритмы, которые учатся прогнозировать вероятность появления токена в зависимости от контекста. Языковые модели бывают авторегрессионными и маскированными. Первые автоматически предсказывают следующий токен на основе всех предыдущих, вторые учитывают контекст до и после токена[25].

Механизм внимания

Механизм внимания — метод, используемый в различных моделях машинного обучения, который позволяет им динамически фокусироваться на определенных токенах входных данных на каждом шаге генерации[26].

Условные обозначения, принятые в книге

В книге используются следующие способы выделения текста.

Курсив

Курсивом выделены новые термины, URL, электронные адреса, имена и расширения файлов.

`Моноширинный шрифт`

Моноширинным шрифтом выделены фрагменты кода, имена переменных или функций, базы и типы данных, переменные среды, инструкции и ключевые слова.

`Полужирный моноширинный шрифт`

Полужирным моноширинным шрифтом выделены команды или значения, которые вводит пользователь.

`Курсивный моноширинный шрифт`

Курсивным моноширинным шрифтом выделен текст, вместо которого следует подставить пользовательские данные или значения с учетом контекста.

Так оформляются подсказки и рекомендации, а в главах 7 и 8 — задания или промпты.

Так оформляются общие комментарии.

Так оформляются важные фрагменты текста и предостережения.

Фрагменты кода

Скачайте дополнительные материалы, включая примеры с фрагментами кода и задания, с GitHub: *(https://github.com/matthew-mcateer/practicing_trustworthy_machine_learning)*.

Возникли технические вопросы или проблемы при использовании примеров кода? Отправьте письмо на адрес bookuestions@oreilly.com.

Эта книга — ваш инструмент. Фрагменты кода из текста можно свободно использовать в своих проектах и документации без предварительного разрешения, если не воспроизводится значительная часть кода. Для написания программы на основе нескольких фрагментов кода из этой книги специальное разрешение не требуется.

Разрешение требуется для продажи или распространения примеров из книг издательства O'Reilly. Ответ на вопрос путем ссылки на эту книгу и цитирования кода из примера также не требует разрешения. Однако, чтобы включить значительный объем примеров кода в документацию своего продукта, нужно получить разрешение.

При цитировании рекомендуется указать источник, включая название книги, авторов, издательство и ISBN, например: Practicing Trustworthy Machine Learning («Машинное обучение. Как построить надежные модели искусственного интеллекта»), Yada Pruksachatkun, Matthew McAteer, Subhabrata Majumdar (O'Reilly). Copyright 2023 Yada Pruksachatkun, Matthew McAteer, and Subhabrata Majumdar, 978-1-098-12027-6.