DOI: 10.25728/mlsd.2023.1315

ПРОБЛЕМЫ ПРИ ПРИМЕНЕНИИ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В АВИАЦИИ

Кулида Е.Л., Лебедев В.Г.
Институт проблем управления им. В.А. Трапезникова РАН, Москва, Россия
elena-kulida@yandex.ru, lebedev-valentin@yandex.ru

Аннотация. Рассматриваются потенциальные области применения методов машинного обучения в авиации. Исследуются проблемы, препятствующие быстрому внедрению методов машинного обучения в авиации, такие как состязательные атаки, проблемы сдвига данных и утечки данных, непрозрачность моделей «черного ящика».

Ключевые слова: модели глубокого обучения, состязательные атаки, объяснимый искусственный интеллект.

Введение

В настоящее время в области технологий искусственного интеллекта (ИИ) достигнуты значительные успехи, связанные, в частности, с применением методов машинного обучения и глубоких нейронных сетей в области компьютерного зрения и обработки естественного языка.

Достижения в области технологий ИИ основаны на возросшей вычислительной мощности суперкомпьютеров, разработанных эффективных библиотеках программного обеспечения для глубокого обучения, больших объемах накопленных данных. Новые технологии начинают играть все более важную роль во многих областях человеческой деятельности, в том числе в авиации.

ИИ и автоматизация использовались в различных областях авиации в течение многих десятилетий. К ним относятся: диагностика происшествий, например, обнаружение беспилотных воздушных систем и предотвращение столкновений, компьютерное обучение пилотов на тренажерах, диагностика компонентов и узлов воздушных судов, автоматизация управления, решение боевых задач и помощь в полете, например, при принятии оперативных решений экипажем, интеллектуальный интерфейс экипажа, сбор/обработка/анализ данных воздушного движения для систем управления воздушным движением, оптимизация структуры воздушного пространства и планирование потоков воздушных судов, оптимизация маршрутов и очередей воздушных судов в зоне аэропорта [1].

Наиболее обсуждаемым перспективным применением ИИ в авиации является автономный полет. На первом этапе ИИ может помогать экипажу при помощи систем поддержки принятия решений при высокой нагрузке, предсказывая и предотвращая аварийные ситуации. От помощи экипажу предполагается постепенно переходить к замене второго пилота виртуальным пилотом, затем к автономным полетам под наблюдением человека и затем к полностью автономным полетам. В дорожной карте искусственного интеллекта Агентства авиационной безопасности Европейского союза [2] говорится:

«В области коммерческого воздушного транспорта временная шкала, связанная с тремя вышеописанными этапами, может быть следующей:

- Первый этап: помощь/усиление экипажа (2022-2025 гг.)
- Второй этап: взаимодействие человека и машины (2025–2030 гг.)
- Третий шаг: автономный коммерческий воздушный транспорт (2035+)».

Однако внедрение перспективных методов ИИ в авиацию тормозится тем обстоятельством, что человеку не понятно, каким образом системы ИИ формируют решения и вследствие этого нет уверенности в правильности предлагаемых решений. Безопасность в авиации является главным приоритетом и традиционно она достигается с помощью человека, который несет за нее ответственность. Для внедрения перспективных систем ИИ в авиацию необходимо, чтобы принимаемые системой решения были понятны пользователю и адаптировались к его состоянию. Повысить доверие человека к системам ИИ призван объяснимый ИИ (Explainable AI, XAI).

1. Области применения технологий искусственного интеллекта в авиации

1.1. Интеллектуальная поддержка экипажа

Системы автоматизации полета и поддержки принятия решений экипажем уже позволили уменьшить экипаж гражданского ВС с 4 до 2 человек. Компьютерные программы имеют три важных преимущества по сравнению с человеком:

- Высокая скорость обработки данных.

- Возможность обработки больших объемов данных.
- Отсутствие физиологических и психологических ограничений, присущих человеку.

По мере развития технологий ИИ компьютерные программы на их основе смогут лучше человека справляться со многими задачами, которые до сих пор решались исключительно людьми. В будущем компьютерные программы смогут быстрее и правильнее человека обнаруживать и распознавать объекты, реагировать на изменяющуюся ситуацию, генерировать варианты необходимых действий.

На рис.1 представлена тенденция изменения сравнительных возможностей человека и интеллектуальных систем в задаче автоматического обнаружения объектов из работы [3]. С некоторого момента доля правильных решений глубоких конволюционных нейронных сетей превосходит долю правильных решений человека-оператора, количество ошибок распознавания нейронных сетей меньше количества ошибок распознавания, допускаемых человеком.

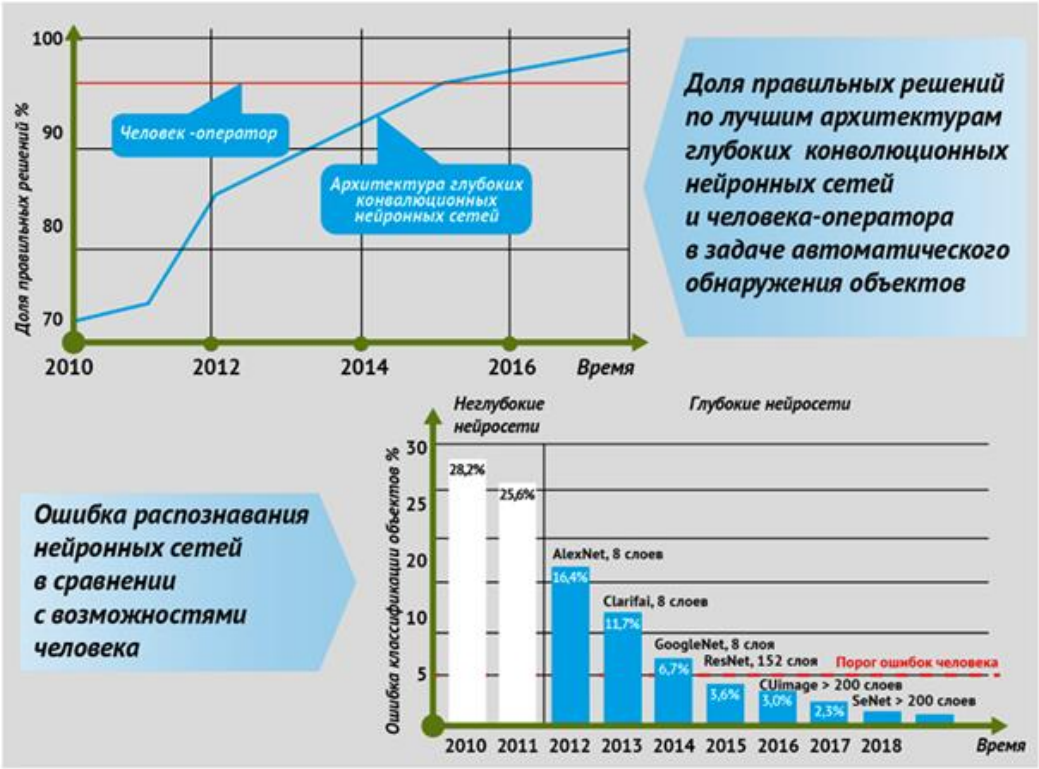

Рис. 1. Сравнение возможностей человека и интеллектуальных систем

1.2. Производство и техническое обслуживание ЛА

В настоящее время для контроля работы различных систем широко используются датчики, в результате накапливаются большие объемы записанных полетных данных. Эти данные содержат скрытые закономерности, которые могут быть обнаружены при помощи методов интеллектуального анализа данных (ИАД). В последние годы много научных работ посвящено применению методов ИАД для обнаружения нештатных ситуаций в работе различных технических систем, в частности, в работе авиационного реактивного двигателя. Выявление нештатных ситуаций в авиационных двигателях имеет большое значение для улучшения работ по своевременному техническому обслуживанию двигателей, что приводит повышению безопасности полетов и повышению эффективности авиакомпаний.

Превентивное выявления неисправностей технических систем основано на поисках в больших наборах аномалий в данных, не соответствующих нормальному функционированию системы.

Различают три группы методов обнаружения аномалий:

- Методы обучения с учителем основаны на наборах данных для обучения и тестирования. В этих методах предполагается, что все аномалии известны заранее и отражены в наборах данных для обучения и тестирования, поэтому такие методы не позволяют обнаруживать новые аномалии. Кроме того, подготовка набора обучающих данных для этих методов очень сложна и проблематична.

- Методы обучения без учителя используют критерии расстояния или плотности в наборе данных. Нормальные точки принадлежат большим и плотным кластерам. Аномальные точки не принадлежат ни одному кластеру или принадлежат очень маленькому кластеру.
- Наиболее перспективными выглядят методы с классификацией одного класса, в которых требуются наборы обучающих данных для нормального класса, которые могут быть получены в реальных полетах, а аномальные данные находятся при отклонении от нормального класса.

Обнаружение аномалий с помощью методов интеллектуального анализа позволяет своевременно диагностировать неполадки в работе технических систем.

1.3. Управление воздушным движением

Управление воздушным движением – это совокупность многих видов деятельности по управлению различными ресурсами, такими как воздушное пространство, воздушные суда, аэропорты, взлетно-посадочные полосы и т.д.

Во многих исследовательских проектах рассматриваются методы ИИ для управления воздушным движением [2]:

- Улучшение стратегического планирования потоков ВС: в проекте INTUIT разработаны методы визуальной аналитики и машинного обучения для понимания компромиссов между безопасность и эффективностью.
- Улучшение прогнозирования траектории: в проектах DART и COPTRA используется машинное обучение для прогнозирования траектории и оценки характеристик самолета до или во время полета. Результаты основаны на моделях, предварительно обученных по записанным траекториям.
- Помощь диспетчеру при разрешении конфликтов: Сингапурский научно-исследовательский институт разработал приложение ИИ, которое использует записи о стратегиях человека-оператора для разрешения конфликтов.
- Оптимизация эшелонирования при заходе на посадку. Решение, уже развернутое в Хитроу, в настоящее время дополнительно усовершенствовано алгоритмами машинного обучения, которые уточняют минимумы эшелонирования в следе на основе параметров, передаваемых от самолетов.

В последнее время появились работы по стратегическому планированию траекторий ВС при помощи глубокого обучения с подкреплением [4]. В основе обучения с подкреплением лежит обратная связь, получаемая моделью от окружающей среды. В будущем это позволит продолжать обучение в процессе эксплуатации и на этой основе адаптировать модель к изменяющимся условиям.

1.4. Организация работы БПЛА в городской среде

В настоящее время БПЛА широко применяются в мире для решения военных и для гражданских задач: для поисково-спасательных работ, аэросъемки, доставки товаров, мониторинга и контроля территорий, охраны объектов и других. Использование БПЛА дешевле и экологичнее использования других видов транспорта.

Однако при возникновении всевозможных непредвиденных ситуаций при организации работы БПЛА, в том числе в перегруженной городской среде со множеством препятствий, требуется принятие большого количества решений в процессе полета, что без применения технологий ИИ вряд ли возможно.

2. Подходы к решению проблем машинного обучения

2.1. Защита моделей глубокого обучения от состязательных атак

Несмотря на феноменальные достижения технологий глубокого обучения нейронных сетей во многих практических приложениях, оказалось, что они уязвимы к малозаметным незначительным возмущениям входных данных, получившим название состязательных возмущений. Состязательные атаки способны обманывать модели глубокого обучения и приводить к получению неправильных результатов, при этом модель делает вывод о высокой достоверности неправильного результата.

В литературе описано большое количество различных методов построения состязательных примеров, т.е. модифицированных входов для классификаторов, позволяющих обмануть модель обучения глубокой сети. *Одношаговые* методы генерируют состязательное возмущение, выполняя одноэтапное вычисление, например, однократно вычисляя градиент потери модели. *Итерационные* методы выполняют одни и те же вычисления несколько раз, чтобы получить одно возмущение, что требует более значительных вычислительных ресурсов. *Универсальное* возмущение способно с высокой вероятностью обмануть модель на разных входах. *Переносимость* относится к способности

состязательного примера оставаться эффективным даже для моделей, отличных от той, которая использовалась для его создания. Атаки методом *«белого ящика»* предполагают полное знание целевой модели, включая значения ее параметров, архитектуру, метод обучения, а в некоторых случаях и данные обучения. Атаки методом *«черного ящика»* предполагают, что злоумышленник не имеет знаний о модели или имеет ограниченные знания, но точно не знает о параметрах модели (*«получерный ящик»*). В работе [5] представлен всесторонний обзор литературы по состязательным атакам на модели глубокого обучения, который убедительно демонстрирует, что атаки со стороны злоумышленников представляют реальную угрозу для практического применения методов глубокого обучения в задачах компьютерного зрения в реальном физическом мире.

В литературе существуют различные несовпадающие точки зрения на причины уязвимости нейронных сетей, это свидетельствует о том, что требуются дальнейшие исследования в этом направлении.

В настоящее время защита от атак противника развивается по трем основным направлениям:

1) Использование состязательного обучения, при котором используются состязательные примеры во время обучения или во время тестирования.

2) Модификация сетей, например, путем добавления дополнительных слоев/подсетей, изменения функций потери/активации и т. д.

3) Использование внешних моделей в качестве сетевого дополнения при классификации невидимых примеров.

Исследования, посвященные проблеме уязвимости глубоких нейронных сетей и методам их защиты, развиваются очень активно в последнее время. С одной стороны, разрабатываются методы защиты нейронных сетей от известных атак, с другой стороны, разрабатываются более мощные атаки.

Для того чтобы технологию глубокого обучения нейронных сетей можно было использовать в области авиации, в которой безопасность является одним из основных требований и ошибки могут привести к катастрофическим последствиям, должны быть разработаны адекватные меры противодействия состязательным атакам. Высокая активность научных исследований в этом направлении позволяет надеяться, что методы глубокого обучения в будущем станут достаточно надежными и устойчивыми против атак злоумышленников.

2.2. Объяснимый искусственный интеллект

В основе многих научных достижений в последние годы лежат технологии машинного обучения, в которых модели являются «черными ящиками». Это означает, что причины, по которым компьютерная система получает тот или иной результат, скрыты не только от пользователей, но даже от разработчиков. Известны проблемы машинного обучения – сдвиг данных и утечка данных, т.е. различные распределения в наборах тестовых данных, на которых модель обучалась и тестировалась, и в реальных наборах данных, на которых модель используется. Это может привести к тому, что на практике модель работает хуже, чем ожидалось. Это тормозит внедрение и сертификацию методов машинного обучения во многих важных сферах человеческой деятельности, включая авиацию, в которых критически важно обеспечить безопасность. Пользователи должны быть уверены, что система будет правильно работать на реальных данных, иначе они не будут использовать эту систему, поскольку при ошибках последствия могут быть катастрофическими.

Для преодоления этого препятствия большие усилия прилагаются в области разработки так называемого объяснимого ИИ. Этой проблемой занимаются в США, Великобритании, Норвегии и других странах.

Одна из идей заключается в том, чтобы построить более простую «модель-объяснитель», локально верную для конкретного входного вектора, которая позволяет выявить признаки, которые наиболее сильно повлияли на принятие решения. Объяснение конкретного решения важно для принятия или отклонения этого решения пользователем. Объяснение репрезентативного набора конкретных решений может повысить понимание работы модели в целом. Многочисленные примеры из литературы свидетельствуют, что после получения объяснений даже неспециалисты в области машинного обучения могут выяснить, что есть серьезные проблемы и классификатору нельзя доверять, и каким образом его необходимо скорректировать.

Были предложены методы объяснений, использующие знания об анализируемой модели, такие как DeepLIFT [6], Integrated Gradients [7]. Широкую известность получили методы, которые рассматривают анализируемую модель классификации как черный ящик: LIME (Local Interpretable Model-agnostic Explanations) [8] и SHAP (SHapley Additive exPlanations) [9]. В этом случае «модель-

объяснитель» не зависит от анализируемой модели, что позволяет сравнить объяснения для разных моделей, оценивать разные модели и выбирать лучшую.

В [8] методе LIME объяснение формально определяется как модель $g \in G$, где G – это класс потенциально интерпретируемых моделей, таких как линейные модели или деревья решений, которая может быть объяснена пользователю с помощью визуальных средств или текстовых артефактов. Подход LIME заключается в подборе интерпретируемой модели, локально верной в окрестности объясняемого примера *x*. Для этого примера выбираются *M* конкретных интерпретируемых изменений Δ_i, которые могут присутствовать или отсутствовать, $z_i \in \{0,1\}$ – индикатор наличия изменения Δ_i. Таким образом вектор $z \in \{0,1\}^M$ определяет 2^M различных вариантов $h(z) = x + \Delta x$.

$f(x) -$ вероятность (или бинарный индикатор) того, что *x* принадлежит к определенному классу. Используется $\pi_x(z)$ – мера близости между экземпляром *z* и *x*, чтобы определить локальную область вокруг *x*.

Обучающая выборка для объясняющей модели:

$$\left\{z^{(i)}, f\left(h\left(z^i\right)\right)\right\}_{i=1}^{N}. \quad (1)$$

Веса примеров для обучающей выборки:

$$\left\{\omega^{(i)} = \pi_x\left(z^{(i)}\right)\right\}_{i=1}^{N}. \quad (2)$$

Наконец, пусть $\mathcal{L}(f, g, \pi_x)$ – это мера того, насколько неверно g аппроксимирует f в локальной области π_x. Для того чтобы обеспечить интерпретируемость и локальную верность, необходимо минимизировать $\mathcal{L}(f, g, \pi_x)$, Объяснение LIME определяется следующим образом:

$$\xi(x) = \underset{g \in G}{argmin}\, \mathcal{L}(f, g, \pi_x) + \Omega(g), \quad (3)$$

где $\Omega(g)$ – мера сложности объяснения g, при этом $\Omega(g)$ должно быть достаточно малым, чтобы модель была понятной пользователю. Например, для деревьев решений $\Omega(g)$ может быть глубиной дерева, а для линейных моделей $\Omega(g)$ может быть количеством ненулевых весов.

Подход LIME не зависит от исходной объясняемой модели и может применяться к разным моделям, при этом многие детали метода могут варьироваться. Однако очевидны проблемы такого подхода:

- Удастся ли подобрать интерпретируемые изменения в малой окрестности вектора *x*?
- Удастся ли построить достаточно простую модель g, которая будет понятна пользователям.
- И, самое главное, не сильно ли исказит упрощение исходную модель?

3. Заключение

Методы машинного обучения все больше используются в различных областях окружающей действительности, в том числе во многих областях авиации. Однако при решении таких задач, в которых критически важным является обеспечение безопасности, существующий уровень исследований и имеющиеся проблемы не позволяют пока говорить о возможности применения методов машинного обучения в реальной обстановке и необходимы дополнительные научные исследования, которые позволят гарантировать стабильный надежный безошибочный и понятный результат.

Известные в настоящее время методы объяснимого ИИ больше ориентированы на разработчиков, чем на пользователей. Для преодоления предубеждений при принятии решений система должна адаптироваться к пользователю, взаимодействовать с пользователем, предоставлять ему необходимую информацию на трех взаимосвязанных уровнях, необходимых при управлении воздушным движением, [10]:

- Описательном – описывающем алгоритм ИИ и его результаты в понятной пользователю форме.
- Прогностическом – прогнозирующем поведение системы при определенном вводе или модификации.
- Предписывающем – предлагающем способ преодоления ошибок или нежелательного поведения системы.

В обзоре [10] утверждается, что в настоящее время разработки в области объяснимого ИИ сосредоточены на описательном уровне, хотя некоторые исследования предоставляют прогностические характеристики — в основном чувствительность предсказаний для некоторых переменных. Однако необходимы дополнительные исследования на прогностическом и

предписывающем уровнях, чтобы оценить потенциальную ценность, которую объяснимый искусственный интеллект в целом может принести авиационному сообществу.

Литература

1. *Кулида Е.Л., Лебедев В.Г.* Перспективы использования методов искусственного интеллекта в авиации // Труды 13-й Международной конференции «Управление развитием крупномасштабных систем» (MLSD'2020, Москва). – М.: ИПУ РАН, 2020. – С. 1535–1541.
2. *European Union Aviation Safety Agency*, Artificial Intelligence Roadmap: A human-centric approach to AI in aviation, (2020), https://www.easa.europa.eu/sites/default/files/dfu/EASA-AI-Roadmap-v1.0.pdf
3. *Лукашов А., Панферов О., Максимов В., Башкиров А.* Искусственный интеллект для авиационных систем // Арсенал Отечества. № 4 (54). – 2021.
4. *Кулида Е.Л., Лебедев В.Г.* Методы решения задач планирования и регулирования потоков воздушного движения Ч. 2. Применение методов глубокого обучения с подкреплением // Проблемы управления. – 2023. № 2. – С. 3–18.
5. *N. Akhtar and A. Mian* Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey // https://doi.org/10.48550/arXiv.1801.00553.
6. *Shrikumar, A.; Greenside, P.; Kundaje, A.* Learning Important Features through Propagating Activation Differences // In Proceedings of the 34th International Conference on Machine Learning – Sydney, Australia, 6–11 August 2017. – Vol. 70, – P. 3145–3153.
7. *Sundararajan, M.; Taly, A.; Yan, Q.* Axiomatic attribution for deep networks // In Proceedings of the 34th International Conference on Machine Learning – Sydney, Australia, 6–11 August 2017 – Vol. 70. – P. 3319–3328.
8. *Ribeiro, M.T.; Singh, S.; Guestrin, C.* "Why should I trust you?" Explaining the predictions of any classifier // In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – San Francisco, CA, USA – 13–17 August 2016. – P. 1135–1144.
9. *Lundberg, S.M., Lee, S.I.* A unified approach to interpreting model predictions // In Proceedings of the Advances in Neural Information Processing Systems – Long Beach, CA, USA, 4–9 December 2017. – P. 4765–4774.
10. *Degas, A.; Islam, M.R.; Hurter, C.; Barua, S.; Rahman, H.; Poudel, M.; Ruscio, D.; Ahmed, M.U.; Begum, S.; Rahman, M.A.; et al.* A Survey on Artificial Intelligence (AI) and eXplainable AI in Air Traffic Management: Current Trends and Development with Future Research Trajectory // Applied Sciences. – 2022, 12, 1295. https://doi.org/10.3390/app12031295