

УДК 004.89

РАЗРАБОТКА ИНСТРУМЕНТОВ ДЛЯ ОТСЛЕЖИВАНИЯ ПРОИСХОЖДЕНИЯ И ВЕРСИЙ ПРЕДОБУЧЕННЫХ МОДЕЛЕЙ

Гегельский М.А., Оператор научной роты, III курс
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

В условиях бурного роста масштабов и сложности предобученных моделей машинного обучения – от языковых трансформеров с сотнями миллиардов параметров до мультимодальных архитектур, взаимодействующих с текстами, изображениями и аудио – вопрос отслеживания происхождения, версий и эволюции моделей приобретает фундаментальное значение. Модель, применяемая в реальной системе, больше не является статическим объектом. Она становится динамическим артефактом, который проходит через множество стадий: предобучение, тонкая настройка (fine-tuning), сжатие, адаптация, интеграция с другими компонентами, обновление и постэксплуатационную поддержку. При этом даже незначительное изменение в данных, гиперпараметрах или зависимостях может радикально изменить поведение модели. В таком контексте становится критически важным создание инструментов, позволяющих отслеживать происхождение (provenance), версии и изменения моделей на протяжении всего их жизненного цикла [1].

Под происхождением модели (model provenance) в инженерной практике понимается полный и верифицируемый журнал всех этапов создания, модификации и использования модели. Это включает:

- источник обучающих данных (вплоть до конкретных выборок, версий и условий очистки); [2]
- параметры и конфигурации модели на этапе обучения (включая seed генераторов, архитектурные гиперпараметры, версии фреймворков и аппаратного обеспечения);
- последовательность трансформаций, проведённых после предобучения: адаптация под задачу, удаление признаков, квантизация, прунинг;
- интеграция в пайплайн и окружение деплоя (вплоть до точной версии контейнера и зависимостей);
- история обновлений, дообучений, исправлений и отзывов модели.

Без этой информации невозможно обеспечить репродуцируемость, верификацию поведения, юридическую подотчётность, а также доверие со стороны

пользователей, партнёров и регуляторов. Именно поэтому разработка инструментов для отслеживания происхождения и версий становится не просто инженерной задачей, а частью стратегии построения зрелой, устойчивой и этически ответственной ИИ-инфраструктуры.

Современные практики отслеживания версий модели развиваются в рамках концепции ModelOps – направления, которое расширяет принципы DevOps на сферу машинного обучения. В рамках ModelOps модели рассматриваются как управляемые артефакты, для которых поддерживаются механизмы регистрации, версии, отката, сравнения и верификации. Наиболее известными инструментами в этой области являются MLflow, Weights & Biases, DVC (Data Version Control), Kubeflow Pipelines, Neptune.ai, Guild AI и другие. Эти системы позволяют отслеживать версии моделей, сопоставленные с конкретными конфигурациями обучения, входными данными и результатами. Например, MLflow позволяет сохранять и сравнивать метрики качества, параметры обучения, артефакты (веса, логики вывода, визуализации), и связывать их с системой управления версиями кода (например, Git).

Однако инструменты отслеживания должны выходить за рамки управления артефактами и включать в себя семантический контроль эволюции модели. Это означает, что необходимо не просто зафиксировать тот факт, что модель была обновлена, а понять как и почему её поведение изменилось. Для этого разрабатываются методы сравнения версий моделей на уровне поведения: идентичных входов, распределения предсказаний, чувствительности к признакам, генерализуемости. Такие методы позволяют, например, обнаружить регрессию в определённой подгруппе данных или рост неопределённости, вызванный новым обучающим корпусом. Некоторые платформы, такие как Deepchecks, Evidently AI, Arize или Fiddler, предлагают автоматическую генерацию отчётов об изменениях в поведении модели при смене версии, что особенно важно для систем, подлежащих внешнему аудиту.

Важным аспектом является также прослеживаемость источников данных, поскольку именно они в значительной степени определяют поведение модели. Здесь на первый план выходит концепция Data Provenance и инструменты контроля версий данных, такие как DVC, Pachyderm, LakeFS. Они позволяют связывать конкретную модель с точным срезом обучающего датасета, включая метаинформацию об источниках, условиях очистки, политике приватности. Это критично при использовании персональных, медицинских, юридически защищённых данных, а также при необходимости соблюдения требований нормативов (GDPR, HIPAA, AI Act) [3].

Другим уровнем становятся инструменты, ориентированные на юридическую и этическую прослеживаемость. В рамках новых европейских инициатив (например, AI Act, ISO/IEC 42001) предобученные модели, особенно

"общего назначения" (GPAI), должны сопровождаться документацией, отражающей как этапы предобучения, так и все модификации, производимые конечными пользователями. Это означает необходимость разработки публичных реестров моделей, в которых отображается происхождение, конфигурации, ответственность сторон, истории обновлений, области допустимого применения и ограничения. Некоторые инициативы уже движутся в этом направлении: HuggingFace Hub, OpenRAIL лицензии, Model Cards, Data Sheets, Transparency Logs [4].

На техническом уровне разрабатываются стандарты и протоколы, позволяющие моделям подписывать себя и быть воспроизводимыми верифицируемым способом. Это может быть реализовано с помощью криптографических хэшей весов, конфигурационных файлов, зависимостей, а также механизма цифровой подписи и blockchain-хранения записей. Такие подходы особенно актуальны для моделей, распространяемых как open-source, но применяемых в критических системах: пользователь может убедиться, что модель действительно соответствует той, что была протестирована, верифицирована и лицензирована [5].

Будущее отслеживания происхождения и версий моделей связано с появлением сквозных экосистем, в которых от этапа генерации данных до конечного вывода и обратной связи будет существовать полный и проверяемый граф артефактов. В такой системе любая версия модели – как и любые данные, на которых она обучалась – смогут быть воспроизведены, проверены и объяснены. Это не только технический вызов, но и новая философия доверенного ИИ, в котором прозрачность, подотчётность и ответственность встроены в саму инфраструктуру. Такие решения станут основой для цифровых цепочек доверия, обязательных для государств, корпораций и международных организаций, использующих ИИ в задачах высокой чувствительности и правовой значимости [6].

Таким образом, разработка инструментов для отслеживания происхождения и версий предобученных моделей – это стратегически важное направление, находящееся на пересечении инженерии, управления жизненным циклом ИИ и цифровой этики. Оно создаёт основу для верифицируемости, регуляторной совместимости и воспроизводимой науки, без которых масштабное и безопасное внедрение искусственного интеллекта в общественную и профессиональную жизнь невозможно.

Список литературы:

1. Simmhan, Y. L. A Survey of Data Provenance in e-Science / Y. L. Simmhan, B. Plale, D. Gannon // ACM SIGMOD Record. – 2005. – Vol. 34, № 3. – P. 31–36. – DOI 10.1145/1084805.1084812.

2. Buneman, P. Why and Where : A Characterization of Data Provenance / P. Buneman, S. Khanna, W.-C. Tan // Database Theory – ICDT 2001. – Berlin : Springer, 2001. – P. 316–330. – DOI 10.1007/3-540-44503-X_20.

3. Cheney, J. Provenance in Databases : Why, How, and Where / J. Cheney, L. Chiticariu, W.-C. Tan // Foundations and Trends in Databases. – 2009. – Vol. 1, № 4. – P. 379–474. – DOI 10.1561/19000000006.

4. Vartak, M. ModelDB : A System for Machine Learning Model Management / M. Vartak, H. Subramanyam, W.-E. Lee [et al.] // Proceedings of the Workshop on Human-In-the-Loop Data Analytics. – New York : ACM, 2016. – Article 14. – DOI 10.1145/2939502.2939516.

5. Zaharia, M. Accelerating the Machine Learning Lifecycle with MLflow / M. Zaharia, A. Chen, A. Davidson [et al.] // IEEE Data Engineering Bulletin. – 2018. – Vol. 41, № 4. – P. 39–45.

6. Baylor, D. TFX : A TensorFlow-Based Production-Scale Machine Learning Platform / D. Baylor, E. Breck, H.-T. Cheng [et al.] // Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – New York : ACM, 2017. – P. 1387–1395. – DOI 10.1145/3097983.3098021.