УДК 004.89

ОЦЕНКА ПЕРЕНОСИМОСТИ ПРЕДОБУЧЕННЫХ МОДЕЛЕЙ С ТОЧКИ ЗРЕНИЯ ДОВЕРИЯ

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

Предобученные модели становятся неотъемлемым инструментом современной инженерии искусственного интеллекта. Используемые в самых различных доменах – от медицины и промышленной диагностики до автоматического перевода и генерации текстов – они позволяют существенно сократить время разработки и ресурсы, необходимые для построения высококачественных ИИ-систем.

Стоит отметить, что использование предобученных моделей вне контекста их первоначального обучения порождает важный и пока недостаточно проработанный вопрос: насколько допустим и надёжен перенос этих моделей на новые задачи и данные, и как можно оценить переносимость предобученной модели с точки зрения доверия [1]?

Термин «переносимость» (transferability) обычно применяется в техническом смысле и описывает способность модели, обученной на одном распределении данных, эффективно функционировать в другом [2].

В контексте доверенного искусственного интеллекта переносимость обретает расширенное значение: она включает не только снижение ошибки и адаптацию архитектуры, но и сохранение тех свойств, которые формируют доверие – устойчивости, интерпретируемости, соблюдения нормативных ограничений, способности к верификации, согласованности поведения и предсказуемости. В условиях высокой регуляторной ответственности (например, в финансовых или медицинских ИИ-системах) даже незначительное снижение доверительных характеристик может привести к недопустимым последствиям.

Оценка переносимости в терминах доверия требует разработки системных методик, которые выходят за рамки классической валидации на тестовой выборке. Важно не только убедиться, что предобученная модель сохраняет приемлемую точность после дообучения, но и понять, как она изменилась в процессе переноса: какие аспекты её поведения сохранились, какие нарушились, где появились новые зоны риска. Одним из фундаментальных шагов в такой оценке является анализ сдвига распределений между данными, на которых модель обучалась, и теми, на которых она применяется. Если новый набор данных

существенно отличается по плотности признаков, по классовой структуре, по семантическому содержанию – то даже хорошо обобщающая модель может демонстрировать деградацию устойчивости или предвзятости [3].

Следующий уровень оценки переносимости связан с анализом семантической совместимости латентных пространств. Даже при сохранении внешних метрик точности, представления, формируемые моделью, могут быть несовместимы с новой задачей или категорией данных. Это особенно критично для языковых моделей, когда модель, обученная на общественно-доминирующем дискурсе, переносится в сферу медицинских, правовых или культурно маргинальных текстов. Снижение переносимости в этом случае не выражается напрямую в ошибках классификации, но проявляется в снижении интерпретируемости, ухудшении градации уверенности, неадекватной генерации альтернатив и потере доверия пользователя [4].

Важно также учитывать, что при переносе модели в новую прикладную среду меняются и контексты риска. Там, где в исходной задаче ошибки были статистически нейтральны (например, в рекомендательной системе), в новой задаче они могут быть фатальными (например, в системе медицинского триажа). Оценка переносимости модели должна учитывать возможные переходы между классами рисков, а следовательно – проводить стресс-тестирование в новых условиях, моделировать редкие, но критичные случаи, анализировать устойчивость к адверсариальным воздействиям, даже если ранее это не считалось необходимым [5].

Особую роль играет также анализ переносимости доверительных предикторов – таких как calibrated confidence, uncertainty estimation, mechanisms of abstention и др. Предобученные модели часто включают внутренние механизмы, оценивающие уверенность в предсказаниях, либо адаптированы к стратегиям отказа от ответа в случае сомнений. При переносе эти механизмы могут нарушаться – калибровка теряется, шкалы изменяются, уверенность становится некоррелированной с истинностью. Именно такие сбои подрывают доверие пользователей, даже если метрики точности в среднем сохраняются. Следовательно, тестирование переносимости должно включать не только метрики качества, но и метрики стабильности доверительных структур [6].

Важным аспектом становится возможность объяснения решений предобученной модели после переноса. Если в исходной задаче объяснения имели логическую и семантическую согласованность с предметной областью, то при переносе в новый контекст возможна утрата интерпретируемости: значимые признаки могут быть переоценены или игнорированы, логика решения – непрозрачна, а само объяснение – ошибочно убедительно. Подобные ситуации критичны в задачах, где человек-оператор или эксперт опирается на объяснение при принятии окончательного решения. Поэтому оценка переносимости

должна включать экспертизу согласованности объяснений до и после дообучения модели.

Существуют и архитектурные подходы к обеспечению доверенной переносимости. В частности, использование модульных предобученных блоков, где часть модели (например, encoder) остаётся фиксированной, а адаптация происходит за счёт добавления тонких слоёв или attention-масок, позволяет ограничить изменчивость модели и отслеживать зоны, в которых происходит адаптация. Такие схемы, как fine-tuning с замороженными слоями, adapter-based learning или prompting, упрощают контроль за изменениями и облегчают аудит поведения модели.

Наконец, полноценная оценка переносимости требует построения метапроцедур мониторинга и валидации на всём жизненном цикле модели. Это означает, что нельзя ограничиваться однократной проверкой на тестовой выборке после дообучения. Необходимо встроить процедуры верификации и анализа в MLOps-пайплайн, отслеживать поведение модели в реальном времени, фиксировать изменения в латентных пространствах, визуализировать доверительные интервалы и создавать отчётность по доверительным метрикам. Особенно важно обеспечить механизмы возврата (rollback) и отслеживания последствий при ухудшении доверительных характеристик.

Таким образом, оценка переносимости предобученных моделей с точки зрения доверия – это не просто проверка функциональной адаптируемости, но полноценный процесс анализа совместимости, устойчивости, интерпретируемости и сохранения доверительных структур.

Этот процесс требует совмещения инженерных практик, теоретических оснований и этических ориентиров, поскольку именно от доверия к перенесённой модели зависит, будет ли она принята экспертами, пользователями и обществом.

В условиях стремительного распространения моделей общего назначения, таких как крупные языковые модели и мультимодальные ИИ-системы, развитие и стандартизация таких подходов становится задачей первостепенной важности.

Список литературы:

1. Pan, S. J. A Survey on Transfer Learning / S. J. Pan, Q. Yang // IEEE Transactions on Knowledge and Data Engineering. – 2010. – Vol. 22, № 10. – P. 1345–1359. – DOI 10.1109/TKDE.2009.191.

2. Yosinski, J. How Transferable Are Features in Deep Neural Networks? / J. Yosinski, J. Clune, Y. Bengio, H. Lipson // Advances in Neural Information Processing Systems. – 2014. – Vol. 27.

3. Weiss, K. A Survey of Transfer Learning / K. Weiss, T. M. Khoshgoftaar, D. Wang // Journal of Big Data. – 2016. – Vol. 3. – Article 9. – DOI 10.1186/s40537-016-0043-6.

4. Kornblith, S. Do Better ImageNet Models Transfer Better? / S. Kornblith, J. Shlens, Q. V. Le // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. – Long Beach : IEEE, 2019. – P. 2656–2666. – DOI 10.1109/CVPR.2019.00277.

5. Recht, B. Do ImageNet Classifiers Generalize to ImageNet? / B. Recht, R. Roelofs, L. Schmidt, V. Shankar // Proceedings of the 36th International Conference on Machine Learning. – 2019. – Vol. 97. – P. 5389–5400.

6. Hendrycks, D. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations / D. Hendrycks, T. Dietterich // International Conference on Learning Representations. – 2019.