УДК 004.4

Штойко М.С., Радюк П.М., Петровський С.С., Вознюк Л.О.

Хмельницький національний університет

МЕТОД ПОЯСНЕННЯ РЕЗУЛЬТАТІВ ЗАДАЧ КЛАСИФІКАЦІЇ ЗА МОДЕЛЯМИ ГЛИБОКОГО НАВЧАННЯ ЗАСОБАМИ МАШИННОГО НАВЧАННЯ

Розглянуто прикладні аспекти розробки методу пояснення результатів класифікаційних задач для моделей глибокого навчання, який використовує сучасні інструменти машинного навчання. Запропонований метод дозволяє точно та оперативно надавати інтерпретації рішень моделей, таких як CNN, RNN та трансформери, що сприяє кращому розумінню ключових факторів, які впливають на результати класифікації. Інтеграція технік, таких як LIME та SHAP, у рамках єдиної системи надає можливість користувачам аналізувати вплив різних характеристик на рішення моделі, що підвищує прозорість і довіру до отриманих результатів.

The practical aspects of developing a method for explaining classification results for deep learning models using modern machine learning tools are examined. The proposed method provides accurate and efficient interpretations of model decisions, including CNN, RNN, and transformer models, enhancing the understanding of key factors influencing classification outcomes. Integrating techniques such as LIME and SHAP within a unified system enables users to analyze the impact of different features on model decisions, thus increasing transparency and trust in the results obtained.

У сучасному світі глибокі нейронні мережі, такі як CNN, RNN та трансформери, стали основними інструментами для вирішення складних задач класифікації. Ці моделі вражають своєю здатністю обробляти великі обсяги даних і досягати високої точності в прогнозах. Проте їхня природа "чорних ящиків" ускладнює розуміння механізмів прийняття рішень, оскільки навіть експерти можуть не мати чіткого уявлення про те, чому модель приймає певні рішення. Це викликає серйозні ризики в критичних сферах, таких як медицина, де помилка в діагностиці може загрожувати здоров'ю пацієнтів, або у фінансових системах, де необґрунтовані рішення можуть призвести до великих фінансових втрат. Відсутність прозорості у цих моделях також ставить під сумнів етичність їх використання, оскільки результати можуть бути упередженими або непрозорими для користувачів. Таким чином, існує нагальна потреба у розробці методів, які дозволять пояснювати рішення глибоких нейронних мереж, що сприятиме підвищенню довіри до цих технологій і забезпечить їх більш етичне застосування [1].

Різноманітні науковці активно досліджують методи інтерпретації результатів глибоких нейронних мереж, зокрема через інструменти, такі як LIME

(Local Interpretable Model-agnostic Explanations) та SHAP (SHapley Additive exPlanations). Ці методи розроблені для пояснення впливу різних факторів на прийняття рішень моделями, що допомагає користувачам краще розуміти, які ознаки впливають на результати. Однак, їхнє застосування у реальних сценаріях виявляє ряд проблем, зокрема неузгодженість результатів, яка може варіювати залежно від контексту, а також складність інтеграції цих методів у виробничі системи. Ці виклики підкреслюють необхідність подальшого розвитку інтерпретаційних методів, щоб забезпечити їх більш надійне та ефективне використання в різних сферах, включаючи критичні області, такі як медицина та фінанси. Потреба в більш адаптивних і прозорих рішеннях залишається актуальною, оскільки інтеграція цих технологій може сприяти зниженню ризиків, пов'язаних із прийняттям рішень [2].

Метою дослідження є розробка нового методу пояснення результатів класифікаційних задач, що базується на сучасних підходах до машинного навчання. Цей метод повинен забезпечити високий рівень прозорості моделей, що сприятиме підвищенню довіри користувачів до результатів, отриманих за допомогою глибоких нейронних мереж. Для досягнення цієї мети планується інтеграція різних технік Explainable AI у єдину систему, що дозволить комбінувати переваги існуючих підходів. Це включатиме використання методів, таких як LIME і SHAP, в поєднанні з новими підходами для більш глибокого розуміння механізмів прийняття рішень. Таким чином, дослідження прагне не лише покращити інтерпретацію результатів, а й адаптувати методи до специфіки різних задач класифікації [3].

У процесі роботи над темою використовуються різні методи Explainable AI для розробки моделі, здатної надавати інтерпретації результатів класифікацій. Для цього проводиться всебічний аналіз існуючих методів, таких як LIME і SHAP, з метою визначення їхніх сильних та слабких сторін.

Таблиця 1 ілюструє різні методи інтерпретації результатів глибоких нейронних мереж:

Таблиця 1 – Різні методи інтерпретації результатів глибоких нейронних мереж

Метод	**Опис**	**Сильні сторони**	**Слабкі сторони**
LIME	Місцеві інтерпретовані пояснення	Швидкість, простота	Неузгодженість результатів
SHAP	Додаткові пояснення Шеплі	Теоретична основа, точність	Складність у масштабуванні

Експериментальне тестування на бенчмарк-датасетах дозволяє оцінити ефективність нового методу в реальних сценаріях [4]. Попередні результати експериментів свідчать про те, що запропонована модель забезпечує зрозумілі пояснення рішень, що підвищує її прийнятність у практичних застосуваннях, таких як медицина та фінанси. Цей підхід має потенціал для розширення використання глибоких нейронних мереж у критичних сферах, де прозорість є необхідною [5].

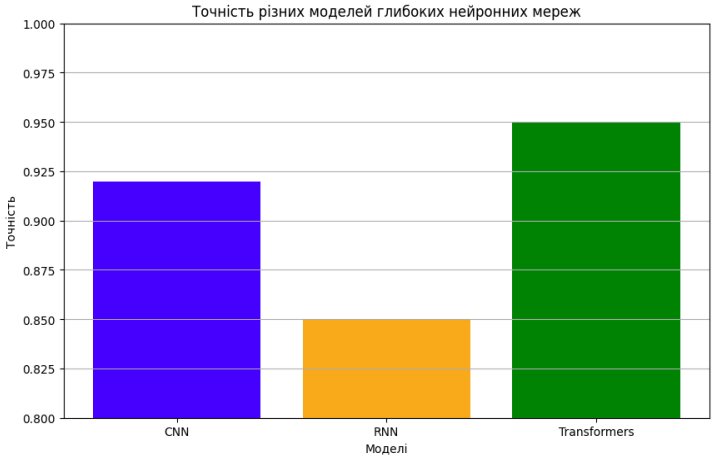

Рисунок 1 – Точність різних моделей

Розроблений метод демонструє значний потенціал у подоланні проблеми "чорної скриньки" в глибоких нейронних мережах, пропонуючи нові механізми для кращого розуміння прийнятих рішень. Це відкриває нові перспективи для вдосконалення інтерпретації результатів у реальному часі, що є особливо важливим для динамічних та критичних застосувань. У майбутніх дослідженнях слід зосередитися на інтеграції цього методу з AutoML системами, що дозволить автоматизувати процеси створення прозорих моделей, підвищуючи ефективність та доступність рішень на основі глибокого навчання.

Перелік посилань

1. Ribeiro M. T., Singh S., Guestrin C. Why Should I Trust You? Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. Pp. 1135-1144.
2. Lundberg S. M., Lee S. I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems. 2017. Vol. 30. Pp. 4765-4774.
3. Chen J., Song L., Wainwright M. J., Jordan M. I. Learning to explain: An information-theoretic perspective on model interpretation. Proceedings of the 34th International Conference on Machine Learning. 2017. Vol. 70. Pp. 1289-1298.
4. Doshi-Velez F., Kim P. Towards a rigorous science of interpretable machine learning. Proceedings of the 34th International Conference on Machine Learning. 2017. Vol. 70. Pp. 2961-2970.
5. Chatzimparmpas A., Mavridis P., Mavridis D. Interpretable Machine Learning: Definitions, Methods, and Applications. Proceedings of the 11th International Conference on Machine Learning and Data Engineering. 2021. Pp. 156-160.