УДК 004.048

А. Луцків, к. т. н., доц.; С. Андруньків
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

ІНТЕРПРЕТАЦІЯ 'ЧОРНИХ СКРИНЬОК'. ВИКОРИСТАННЯ МЕТОДІВ SHAP ТА LIME ДЛЯ ВІЗУАЛІЗАЦІЇ ПРОЦЕСУ ПРИЙНЯТТЯ РІШЕНЬ У ГЛИБОКИХ НЕЙРОННИХ МЕРЕЖАХ

UDC 004.048

A. Lutskiv, PhD., Assoc. Prof.; S. Andrunkiv

INTERPRETATION OF 'BLACK BOXES'. USE OF SHAP AND LIME METHODS TO VISUALIZE THE DECISION-MAKING PROCESS IN DEEP NEURAL NETWORKS

Глибокі нейронні мережі досягли надзвичайної точності у складних задачах, таких як розпізнавання зображень, обробка мови та медична діагностика. Однак їхня внутрішня складність, що складається з мільйонів параметрів, перетворює їх на «чорні скриньки». Відсутність прозорості та неможливість пояснити, чому модель прийняла певне рішення, є головною перешкодою для їхнього впровадження у сфери медицини, фінансів, юриспруденції, де довіра та валідація є обов'язковими.

Для досягнення мети було проаналізовано:

– метод LIME, Local Interpretable Model-agnostic Explanations, який будує просту лінійну модель навколо одного прогнозу, щоб пояснити його локально;

– метод SHAP, SHapley Additive exPlanations, який базується на теорії ігор для обчислення точного внеску кожної ознаки у кінцевий прогноз, забезпечуючи локальну та глобальну інтерпретацію.

LIME продемонстрував здатність швидко надавати інтуїтивно зрозумілі візуальні пояснення для окремих прогнозів, виділяючи «суперпікселі» на зображенні, які найбільше вплинули на рішення (див. рис. 1, зліва). На зображенні зеленою маскою виділені суперпікселі, що відповідають за правильну класифікацію «хаскі», а червоною – ті, що заважають, та які модель теж асоціює з «хаскі».

SHAP надав докладніші та стійкіші пояснення. На рисунку 1, праворуч, схожа візуалізація. На зображенні червоні пікселі показують, що очі та морда собаки підвищили ймовірність класу «хаскі», а сині, що вказують на білу шерсть та фон – знизили. Ця візуалізація часто є чіткішою та менш «шумною», ніж LIME.

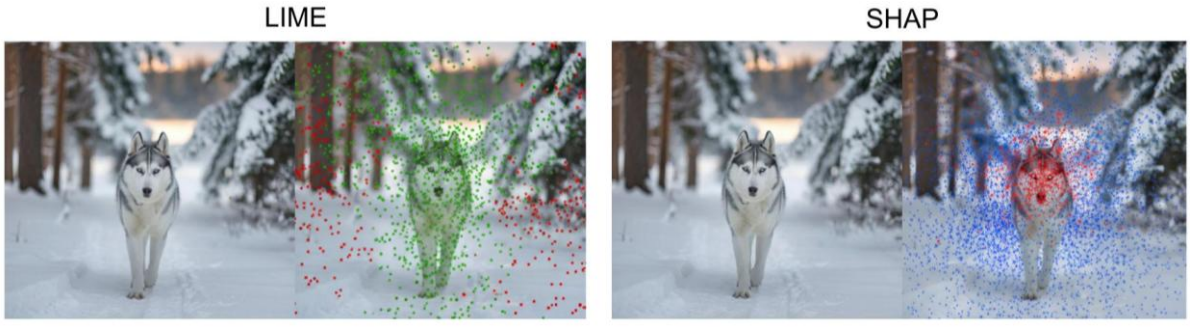

Рисунок 1. Порівняння візуалізацій LIME та SHAP для пояснення класифікації зображення