Mateusz Gniewkowski

Rzetelna ewaluacja: w stronę wiarygodnych reprezentacji w uczeniu maszynowym

Streszczenie

Punktem wyjścia niniejszej rozprawy jest rozbieżność między „jednoliczbową" ewaluacją a realną wiarygodnością modeli, czyli ich zachowaniem w rzeczywistych warunkach. Wysoka trafność na danych należących do tego samego rozkładu co dane uczące (ang. *Independent and Identically Distributed*, IID) nie mówi, czy model opiera decyzję na właściwych przesłankach ani czy jego reprezentacje są użyteczne poza wąskim zadaniem. Nie wiemy też, jak model zachowa się w przypadku przesunięcia rozkładu lub pod wpływem drobnych, semantycznie nieistotnych zmian. W praktyce systemy trafiają w środowiska pełne korelacji pozornych, nowości i bodźców, które potrafią łatwo „oszukać" klasyfikator. Na tę lukę między testem a wdrożeniem składają się w szczególności: niestacjonarność i zmienność domeny; wąskie, niereprezentatywne procedury walidacyjne (brak oceny poza rozkładem uczącym i kalibracji); ryzyka związane z danymi (wycieki, niezrównoważenie klas, artefakty pozyskiwania); brak mechanizmów postępowania z niepewnością (detekcja OOD); a także specyfika uczenia modeli, która sprzyja korelacjom pozornym i uproszczonym regułom decyzyjnym. W odpowiedzi proponujemy dyscyplinę „audyt → pomiar → poprawa", która łączy metody wyjaśnialności (ang. *Explainable Artificial Inteligence*, XAI), metryki jakości reprezentacji oraz testy poza-dystrybucyjne (ang. *Out-of-Distribution*, OOD) w spójny, powtarzalny cykl: najpierw rozpoznajemy, na jakich „dowodach" model opiera decyzje; następnie ilościowo to weryfikujemy; na końcu wprowadzamy małe, ukierunkowane modyfikacje, minimalizujące koszt w dokładności na danych czystych, a jednocześnie podnoszące praktyczną jakość narzędzia.

Poza samą pętlą „audyt → pomiar → poprawa" wprowadzamy *portfolio* metod do audytu i pomiaru jakości oraz odporności reprezentacji: klasteryzację jako wczesny wskaźnik korelacji pozornych i jakości organizacji danych; generalizację OOD oraz detekcję OOD; audyty XAI (lokalne i globalne) ujawniające cechy potencjalnie nieistotne przyczynowo; oraz skalowanie wielowymiarowe wraz z *liczbową* oceną jakości rzutowania. Uzupełniamy to o ewaluację adwersaryjną z wykorzystaniem wyjaśnialności oraz o dwie ścieżki poprawy modeli: (i) dla tekstu – strojenie kontrastowe i destylacja wspierane protokołem *human-in-the-loop*; (ii) dla obrazu – lekka funkcja straty ukierunkowana na podniesienie miary zgodności lokalizacyjnej – rozumianej jako odsetek masy atrybucji przypadającej

na adnotowany obiekt – bez pogorszenia trafności, a przy tym zwiększająca odporność na ataki czarnoskrzynkowe. Całość domyka praktyczny i zwięzły przewodnik, który syntezuje przedstawione metody w spójny, operacyjny schemat oceny i ciągłego doskonalenia modeli uczenia maszynowego w całym cyklu ich życia.

Po zarysowaniu problematyki i wprowadzeniu czytelnika w metodologię, przechodzimy do pierwszego studium przypadku, które ilustruje działanie pętli „audyt → pomiar → poprawa" w praktyce. W rozdziele trzecim, zaproponowano praktyczne narzędzie do wektoryzacji i klasyfikacji ruchu HTTP/URL oparte na modelu językowym RoBERTa. Następnie przeprowadzono systematyczny audyt XAI, który ujawnia słabe strony rozwiązania. Równolegle zmierzono jakość reprezentacji w zadaniach klasteryzacji oraz separacji OOD, co odsłoniło problemy niewidoczne z perspektywy standardowych miar jakości klasyfikacji. W dalszym kroku, kierując się wynikami audytu, przekształcono geometrię osadzeń, aby poprawić jakość reprezentacji, po czym ponownie przeprowadzono pomiary i porównano je z podejściem wyjściowym.

W dalszej części dysertacji proponujemy „zmilitaryzowanie" metod wyjaśnialności, czyli ich celowe wykorzystanie jako narzędzi do konstruowania ataków adwersaryjnych na klasyfikatory tekstowe. Kluczową ideą jest użycie atrybucji jako mechanizmu zawężania przestrzeni poszukiwań przykładów adwersaryjnych: zamiast eksplorować całą kombinatoryczną przestrzeń możliwych modyfikacji tekstu, atak koncentruje się wyłącznie na cechach o największym wpływnie na decyzje modelu. Zaproponowane metody modyfikują – przy zachowaniu poprawności znaczeniowej i gramatycznej – te cechy próbek, które są najbardziej istotne z perspektywy klasyfikacji, w celu wymuszenia błędu. Ataki realizujemy zarówno w trybie nieukierunkowanym (przestawienie do dowolnej innej klasy), jak i ukierunkowanym (do konkretnej, zadanej klasy). Pokazujemy, jak stosować tę procedurę zarówno wobec klasycznych klasyfikatorów tekstowych, jak i wobec modeli generatywnych dostępnych zdalnie, takich jak OpenChat, ChatGPT czy LLaMA, skonfigurowanych do klasyfikacji *zero-shot*. Dodatkowo oceniamy percepcyjną jakość wygenerowanych próbek w badaniu z udziałem ludzi.

Następnie przechodzimy do propozycji ukierunkowanej metody poprawy jakości reprezentacji wizualnych w modelach do klasyfikacji obrazów. Punktem wyjścia jest model o wysokiej trafności, którego decyzje audytujemy pod kątem sensowności używanych „dowodów", rozumianej jako zgodność atrybucji z rzeczywistym obiektem. Wykorzystujemy miarę zgodności lokalizacyjnej – definiowaną jako odsetek masy atrybucji pokrywającej się z adnotowanym obiektem – nie tylko jako narzędzie diagnostyczne, lecz jako sygnał sterujący procesem uczenia. Na tej podstawie proponujemy nową funkcję straty, która ogranicza poleganie na korelacjach kontekstowych i przesuwa reprezentacje w stronę cech obiektowych. Zaproponowana procedura stanowi lekki i architektonicznie agnostyczny mechanizm poprawy reprezentacji. Prowadzi ona do istotnej poprawy modelu z perspektywy miary zgodności lokalizacyjnej i odporności na przeprowadzane ataki adwersaryjne bez

wiedzy o modelu (ang. *black-box adversarial attacks*), przy zachowaniu jakości na danych czystych.

Rozprawa pokazuje, że wiarygodności modeli nie da się ocenić wyłącznie jedną liczbą – na przykład punktem F1 ani samą trafnością testową. Ten, wciąż dominujący w literaturze, paradygmat jest zbyt wąski, by gwarantować niezawodność w praktyce. Zamiast tego modele należy osadzać w cyklu „audyt → pomiar → poprawa", który pozwala zobaczyć działanie systemu w szerszej perspektywie i wprowadzać celowane ulepszenia. Wysokie wartości klasycznych metryk pozostają oczywiście dobrym punktem startowym, lecz niewiele mówią o tym, czy i który model najlepiej sprawdzi się w realnej aplikacji. Modele projektowane z myślą o zachowaniu w warunkach niespodziewanych (w tym poza dystrybucją oraz w pokrewnych zadaniach) rzadziej zawodzą i mogą być obdarzane większym zaufaniem tam, gdzie stawka jest najwyższa.