Wykrywanie i zmniejszanie wpływu tendencyjności danych za pomocą objaśnialnej sztucznej inteligencji

Głębokie uczenie jest obecnie najszybciej rozwijającą się dziedziną sztucznej inteligencji. Szybki rozwój głębokich sieci neuronowych, uważany jest przez część ekspertów jako rewolucja AI (ang. Artificial intelligence), podczas gdy sama sztuczna inteligencja jest coraz częściej nazywana „nową elektrycznością". Jedną z najpopularniejszych modeli stosowanych w głębokim uczeniu w wizji komputerowej są konwolucyjne sieci neuronowe, o skomplikowanej budowie, opisane milionami parametrów. Te paramenty są dobierane automatycznie w trakcie procesu uczenia. Sztuczne sieci neuronowe uczą się na podstawie zestawów danych wejściowych (czyli tego co będzie analizowane przez system, np. zdjęcie znamienia skórnego) oraz danych wyjściowych (czyli wyniku pożądanej ekspertyzy, np. stwierdzenie, że na zdjęciu jest nowotwór skóry). Dawniej, to eksperci w danej dziedzinie wskazywali co podczas uczenia jest najważniejsze, jednak aktualnie stosowane głębokie sieci neuronowe samodzielnie analizują dane i szukają powiązań między nimi. Do tego celu zazwyczaj wykorzystywane są duże ilości specjalnie przygotowanych danych, których jakość ma ogromny wpływ na skuteczność modeli. Dane są często zaszumione, tendencyjne , a czasem nawet zawierają nieprawidłowe etykiety. Jednym z często omawianych problemów jest występowanie tendencyjności danych (ang. bias) – czyli niewystraczająco reprezentatywnych danych, których analiza bez znajomości ukrytych przyczyn, może powodować powstawanie błędów logicznych. Taki problem pojawił się między innymi w predykcji możliwych powikłań po zapaleniu płuc w 1990 roku w Pittsburgu, gdzie inteligentny system wywnioskował, że osoby chore na astmę zyskały dzięki niej dodatkową odporność. Jest to typowy przykład błędnego rozumowania, ponieważ w przeszłości astmatycy o wysokim zagrożeniu powikłań byli pod czujną opieką lekarzy, dzięki czemu rzadko mieli jakiekolwiek powikłania. Tendencyjność danych może przyjmować także inne postaci, np. w jednym z systemów sztucznej inteligencji sieć neuronowa nauczyła się rozpoznawać wyścigi konne ze zdjęć tylko i wyłącznie na podstawie podpisów występujących na wszystkich fotografiach, kompletnie przy tym ignorując konie i jeźdźców.

Co więcej, nawet projektanci tych systemów nie są pewni w jaki sposób wytrenowane sieci podejmują decyzje oraz na jakich cechach się najczęściej skupiają. Pomimo powyższych problemów, głębokie systemy są coraz częściej wykorzystywane do rozwiązywania bardzo ważnych i krytycznych zadań takich jak transport (samochody autonomiczne), medycyna, systemy prawe, bankowość i militaria.

Aby sprostać tym wyzwaniom, projekt ma na celu opracowanie metod objaśnialnej sztucznej inteligencji (ang. Explainable Artificial Intelligence - XAI), które mogą pomóc w wykrywaniu i zmniejszeniu problemu tendencyjności danych. Projekt obejmuje badanie i integrację metod XAI z nowymi i istniejącymi systemami sztucznej inteligencji, w szczególności z głębokimi sieciami neuronowymi w dziedzinie wizji komputerowej. Jednym ze sposobów kategoryzowania metod XAI jest podzielenie ich na lokalne i globalne wyjaśnienia. Analiza lokalna ma na celu wyjaśnienie pojedynczej predykcji modelu, podczas gdy analiza globalna objaśnia, w jaki sposób działa cały model. Projekt ma na celu opracowanie zarówno metod lokalnych, jak i globalnych, w celu zwiększenia interpetowalności systemów typu czarna skrzynka, opartych na głębokich sieciach neuronowych, w celu: ich uzasadnienia, kontrolowania ich procesu decyzyjnego oraz odkrywania nowej wiedzy. Pierwszym zaplanowanym krokiem jest rozwijanie globalnie świadomych objaśnień lokalnych (ang. Globally aware local explanations) dla uzasadnienia poprawności predykcji. Kolejnym etapem będzie, rozwijanie objaśnień globalnych (ang. Global explanations) dla celów detekcji niepożądanych tendencyjności danych. Ostatecznie, zostanie podjęta próba mająca na celu rozwijanie trenowalnego pola uwagi (ang. trainable attention), którego zadaniem jest eliminacja wpływu niepożądanej tendencyjności na pracę modelu.