Nr rejestracyjny: 2022/47/B/ST6/03397; Kierownik projektu: dr hab. Bartosz Michał Zieliński

Interpretowalne metody zrównoważonej sztucznej inteligencji tłumaczące decyzje w sposób intuicyjny

W codziennym życiu człowieka coraz więcej czynności wspomaganych jest przez algorytmy sztucznej inteligencji (SI), choćby w trakcie zakupów czy wytyczaniu trasy przejazdu samochodem. Co więcej, nawet w dziedzinach takich jak medycyna, gdzie decyduje się o losie ludzkiego życia, wykorzystywane są metody SI. Trend ten pokazuje, że coraz więcej aspektów życia jak i gospodarki będzie wspomaganych przez rozwiązania cyfrowe.

Coraz szersze stosowanie rozwiązań SI wiążę się jednak ze zwiększonym zapotrzebowaniem na energię elektryczną, z uwagi na kosztowny proces trenowania i eksploatacji modeli w środowisku produkcyjnym, a także z pytaniem o to czy decyzje tych rozwiązań są godne zaufania. Dlatego w ramach projektu podejmujemy dwa zagadnienia badawcze: Jak projektować metody SI by w pełni wykorzystywały wiedzę uzyskaną wcześniej, dzięki czemu zaoszczędzone zostaną zasoby w procesie dotrenowania modelu? Oraz: Jak trenować te modele, aby ich decyzje były interpretowalne dla użytkowników końcowych, np. lekarzy? Pytania te są szczególnie istotne w przypadku głębokich sieci neuronowych, które uzyskują wysoką skuteczność w wielu aplikacjach, ale nie dostarczają wyjaśnień swoich predykcji z uwagi na ich czarnoskrzynkowy charakter.

W związku z tym, w ramach tego grantu planowane jest opracowanie i rozwijanie modeli, które wykorzystują głębokie sieci neuronowe, ale równocześnie są w stanie zwrócić interpretacje uzyskanego wyniku i zaprezentować pozyskaną przez siebie wiedzę. W tym celu zostaną przeprowadzone badania, które po pierwsze odpowiedzą na pytania jak należy oceniać interpretowalność modelu oraz jak bardzo te interpretacje są zrozumiałe bądź mylące dla końcowego użytkownika. By ograniczyć koszty badań użytkownika, prowadzone będą pracę nad metryką liczbową pozwalającą ocenić właściwości modeli interpretowalnych przy ograniczonym zaangażowaniu człowieka.

W kolejnym kroku planowane są pracę nad odpowiednim przedstawianiem interpretacji użytkownikowi, z wykorzystaniem teorii kognitywistycznych, metod wizualizacyjnych i modeli generujących opis obrazu. Dodatkowo, przeprowadzone zostaną prace nad ujednoliceniem semantyki uzyskiwanej przez model z tą postrzeganą przez człowieka. Na przykład, w przypadku problemu rozpoznawania gatunków ptaków, charakterystyczne cechy wróbli powinny być dla modelu bardziej podobne do siebie niż do cech charakterystycznych albatrosów.

Aby zaadresować problem zrównoważonego wykorzystania zasobów obliczeniowych, a więc także energii elektrycznej, rozwinięte zostaną metody zapobiegające katastroficznemu zapominaniu, tak aby podczas nauki nowej wiedzy, stara nie była zapomina. Również możliwości nauki modelu bez potrzeby etykietowania ogromnych zbiorów danych, co może być kosztowne, w szczególności w danych biomedycznych, jest pożądaną cechą metod SI. W związku z tym, planowane jest opracowanie metody uczenia ciągłego dla głębokich modeli interpretowalnych jak i uczenia kontrastowego, pozwalającego na identyfikację istotnych cech zbioru danych, które później będzie można prezentować użytkownikowi w celu wyjaśnienia predykcji.

Ostatnim krokiem projektu jest zastosowanie opracowanych technik do problemów, które nie są standardowo stosowane przy ewaluacji metod SI, takich jak nauki o życiu. Dzięki temu pokażemy, że stosowanie metod interpretowalnej SI w innych dziedzinach jest równie efektywne jak dla obrazów naturalnych. W szczególności istotnym krokiem jest zastosowanie tychże metod do zdjęć medycznych i wyczerpujące testy zwracanych interpretacji, które określą jakiego typu informacje pozyskuje model oraz czy są one w stanie zwiększyć zaufanie specyficznych użytkowników takich jak lekarze.

Podsumowując, projekt ma na celu przeprowadzenie badań z zakresu sztucznej inteligencji, która uzasadnienia podejmowane przez siebie decyzje i pozwala na ekologiczny i zrównoważony rozwój dziedziny.