Poprawa własności interpretacji wyników głębokich sieci neuronowych w oparciu o części prototypiczne

W ostatnich latach nastąpił szeroki rozwój metod uczenia głębokiego i ich aplikacja w wielu dziedzinach. Głębokie sieci neuronowe zrewolucjoniwały obszary badawcze takie jak wizja komputerowa czy przetwarzanie języka naturalnego uzyskując wyniki podobne lub lepsze od człowieka.

Nawet jeśli model uzyskuje wysoką skuteczność, brak interpretowalności wyniku jest często wskazywany jako jego główna wada, tak jak w przypadku systemu firmy Google do rozpoznawania zmian nowotworowych na skanach z tomografii komputerowej płuc. Co więcej, prawo UE wymaga by systemy oparte o sztuczną inteligencję były samowyjaśnialne. W związu z tym współcześnie prowadzi się badania nad wyjaśnialnością wyników głębokich sieci neuronowych w obrębie dwóch gałęzi:

- wyjaśnialnej sztucznej inteligencji (XAI), która buduje dodatkowy model objaśniający wynik,
- interpretowalnych modeli głębokich, które ze względu na swoją architekture są wyjaśnialne.

Celem projektu jest opracowanie wysoce interpretowalnych modeli uczenia głębokiego, które będą przedstawiać powody swoich decyzji za pomocą części prototypicznych pozyskanych z treningowego zbioru danych. W tym celu opracowany zostanie model pozwalający na współdzielenie prototypów już od poczatku treningu, dzięki czemu zmniejszy się skomplikowanie otrzymywanego wyjaśnienia. Oraz model pozwalający na zaprezentowanie informacji użytkownikowi o wzajemnym położeniu prototypów, które ma wpływ na klasyfikację.

Planowane jest użycie publicznie dostępnych zbiorów danych takich jak: LIDC-IDR czy Stanford Cars, oraz zbiorów uzyskanych i upublicznionych w ramach współpracy z grupą badawczą prof. Moniki Brzychczy-Włoch takich jak zbiór DIFaS. Opracowane modele zostaną wytrenowane w oparciu o te zbiory danych a następnie przeprowadzona zostanie analiza interpretowalności wyników. Dodatkowo, przeprowadzony zostanie eksperyment sprawdzający czy uzyskane wyjaśnienia predykcji pozwalają na zdefiniowanie różnic pomiędzy klasami danych.

Rozwiązanie przedstawione w tym projekcie może zostać zintegrowane w systemach takich jak: komputerowo-wspomagana diagnostykia, komputerowe wspomaganie decyzji, analiza zdjęć satelitarnych, przetwarzania języka naturalnego, czy autonomicznej jazdy.