Wyjaśnialne uczenie maszynowe z zastosowaniem konceptów zrozumiałych dla człowieka

W niniejszej pracy przedstawiam serię pięciu publikacji poświęconych zagadnieniom wyjaśnialności modeli uczenia głębokiego.

Traktowanie sztucznych sieci neuronowych jako czarnych skrzynek, powoduje to, że nie ma pewności, czy decyzje modeli są podjęte na podstawie właściwych przesłanek. W moich pracach przedstawiam nowe statyczne metody wyjaśniające modele uczenia głębokiego, gdzie wyjaśnienie globalne jest generowane po wytrenowaniu modelu. Trzy prace dotyczą metody klasyfikatorów diagnostycznych, które badają informacje zawarte w reprezentacjach modeli. Jest to metoda powszechnie stosowana w przetwarzaniu języka naturalnego, jednak do tej pory nie miała ona swojego odpowiednika w widzeniu maszynowym. W moich pracach wprowadzam intuicyjną taksonomię wizualną, która zawiera znaki, słowa i zdania wizualne, analogicznie do liter, słów i zdań języka naturalnego. Dzięki temu definiuję szereg klasyfikatorów diagnostycznych, które pozwalają na badanie różnych cech reprezentacji modeli. Pokazuję przydatność metody klasyfikatorów diagnostycznych na przykładzie wyjaśniania reprezentacji samonadzorowanych. Metoda ta opiera się na obliczeniowej teorii widzenia Marra, dzięki czemu analizujemy reprezentacje za pomocą zrozumiałych dla człowieka cech wizualnych, takich jak tekstury, kolory, kształty i linie. Moje badania pokazują, że relacje między językiem a obrazem są skutecznymi i intuicyjnymi narzędziami do wyjaśniania modeli uczenia głębokiego.

W dwóch pozostałych pracach przedstawiam nową metodę do anonimizacji zbiorów danych oraz metodę wyjaśniającą działanie modeli uczenia głębokiego w diagnostyce raka piersi. Metoda do anonimizacji obrazów działa z wykorzystaniem syjamskich generatywno-przeciwstawnych sieci neuronowych i pozwala na zbadanie, czy reprezentacje modeli uczenia głębokiego zawierają informacje o tożsamości osób na obrazie. Metoda wyjaśniająca w diagnostyce medycznej bada wpływ perturbacji obrazu na decyzję lekarza oraz maszyny, dzięki czemu stwierdzamy, że modele uczenie głębokiego w dużej mierze korzystają z informacji zawartej w składowych obrazu o wysokiej częstotliwości w przestrzeni Fouriera, które to informacje są niedostrzegane przez lekarzy.

Podsumowując, wszystkie powyższe zaproponowane przeze mnie nowe metody wyjaśniające pomagają lepiej zrozumieć modele sztucznej inteligencji. Dzięki tym metodom jesteśmy w stanie zbadać obciążenie modeli, określić ich silne i słabe strony, a także wskazać które pojęcia są dla nich istotne podczas podejmowania decyzji.

Słowa kluczowe: Wyjaśnialna Sztuczna Inteligencja, Klasyfikatory Diagnostyczne, Widzenie Maszynowe, Uczenie Głębokie