Streszczenie

W ostatnich latach zaobserwowano dynamiczny rozwój metod sztucznej inteligencji, który spowodował automatyzację wielu zadań. W tych zadaniach systemy oparte o sztuczną inteligencję uzyskują skuteczność na poziomie zbliżonym, badź nawet wyższym niż ludzki. Osiągnięto to dla wielu problemów w domenach takich jak rozpoznawanie obrazów czy przetwarzanie języka naturalnego. Stało się to możliwe głównie dzięki rozwojowi mocy obliczeniowej (w szczególności kart graficznych) i metod opartych o sztuczne sieci neuronowe (SSN). Jednakże modele oparte o SSN mimo wielkiej skuteczności mają też szereg wad. Istotną wadą jest ich czarnoskrzynkowy charakter, który rozumie się jako brak wyjaśnienia zwracanych przez model predykcji i sposobu jego rozumowania. Brak zrozumienia decyzji modelu jest szczególnie niepożądany w dziedzinach takich jak medycyna, gdzie decyzje mają znaczący wpływ na ludzkie życie.

W związku z brakiem transparentności SSN opracowano wiele metod wyjaśniających ich decyzje. Można podzielić je na dwie grupy: metody post-hoc oraz metody samowyjaśnialne. Pierwsza grupa metod zakłada opracowanie dodatkowego modelu analizującego decyzje podejmowane przez model czarnoskrzynkowy. Zaletą tego podejścia jest możliwość zastosowania go do wytrenowanego już modelu bez zmian w jego architekturze. Natomiast częsta wadą jest nieprecyzjność i niewiarygodność zwracanych wyjaśnień. W przypadku metod samowyjaśnialnych, mechanizm interpretowalności jest zaszyty w ich architekturę, więc wraz z predykcją zwracane jest jej wyjaśnienie. Dzięki temu zapewniają one większą wiarygodność wyjaśnienia, ale są jednocześnie trudniejsze do wytrenowania, a osiągana przez nie skuteczność jest niższa niż ta uzyskiwana przez metody czarnoskrzynkowe.

Niniejszy doktorat skupia się na metodach samowyjaśnialnych opartych na sztucznych sieciach neuronowych, ze szczególnym uwzględnieniem mechanizmu poolingu atencyjnego oraz części prototypowych. Mechanizm poolingu atencyjnego stosuje się do uczenia ze słabym nadzorem, zwłaszcza do uczenia wieloinstancyjnego (gdzie jedna etykieta przypisana jest do zbioru instancji). Natomiast modele oparte o części prototypowe w trakcie treningu uczą się konceptów wizualnych pochodzących z danych treningowych. Na etapie inferencji zapamiętane przez model koncepty wizualne są porównywane do danych wejściowych w celu dokonania ostatecznej predykcji.

W ramach doktoratu, bazując na modelu opartym o części prototypowe ProtoPNet [11], opracowano trzy nowe architektury, które niwelują ograniczenia modelu bazowego. Pierwszy z nich, ProtoPShare [I], współdzieli części prototypowe pomiędzy klasami poprzez ich łączenie w uprzednio wytrenowanym modelu. Łączenie to odbywa się w oparciu o nową metrykę potrafiącą wykryć podobieństwo semantyczne pomiędzy prototypami, nawet gdy są one odległe w przestrzeni ukrytej. Aby wykluczyć potrzebę trenowania modelu bazowego, wprowadzono ProtoPool [II], który uczy się zbioru części prototypowych wraz z ich przypisaniem do

poszczególnych klas, pozwalając na ich współdzielenie. Jest to możliwe dzięki zastosowaniu technik regularyzacyjnych opartych o Gumbel-Softmax oraz wprowadzeniu podobieństwa focal similarity, które wykrywa bardziej charakterystyczne części prototypowe. Uogólnienie tych metod do problemu regresji w przewidywaniu właściwości molekuł zaprezentowano wraz z dedykowanym modelem ProGReST w pracy [III].

Poza rozwojem metod opartych na częściach prototypowych, w ramach doktoratu rozwinięto metodykę poolingu atencyjnego stosowanego w problemach uczenia wieloinstancyjnego i zaproponowano dwa nowe podejścia. Pierwsze z nich, SA-AbMILP [IV] korzysta z mechanizmu self-attention do nauki zależności pomiędzy instancjami w zbiorze. Dzięki temu lepiej sprawdza się dla bardziej złożonych założeń uczenia wieloinstancyjnego, a jednocześnie wyjaśnia jak dana cecha wizualna wpłynęła na decyzję modelu. Drugie z nich, ProtoMIL [V], pozwala na lokalną i globalną interpretowalność dla problemu uczenia wieloinstancyjnego poprzez połączenie ze sobą części prototypowych i poolingu atencyjnego.

Podsumowując, prezentowany doktorat skupia się na metodach interpretowalności dla uczenia głębokiego, a w szczególności na modelach opartych o części prototypowe i mechanizm atencji. W ramach przeprowadzonych badań opublikowanych zostało pięć prac na konferencjach naukowowych posiadających kategorię A* [I, II] i kategorię A [III, IV, V] według rankingu CORE. Doktorant jest pierwszym autorem wszystkich tych publikacji. Ponadto był on kierownikiem grantu NCN Preludium oraz grantu w ramach Inicjatywy Doskonałości Uniwersytetu Jagiellońsiego. Doktorant odbył również staż naukowy w Centrum Wizji Komputerowej Autonomicznego Uniwersytetu Barcelońskiego w grupie badawczej prof. Joosta van de Weijera, oraz jest współautorem aplikacji patentowej złożonej do Europejskiego Biura Patentowego.

Słowa kluczowe: interpretowalność, wyjaśnialna sztuczna inteligencja, uczenie głębokie.