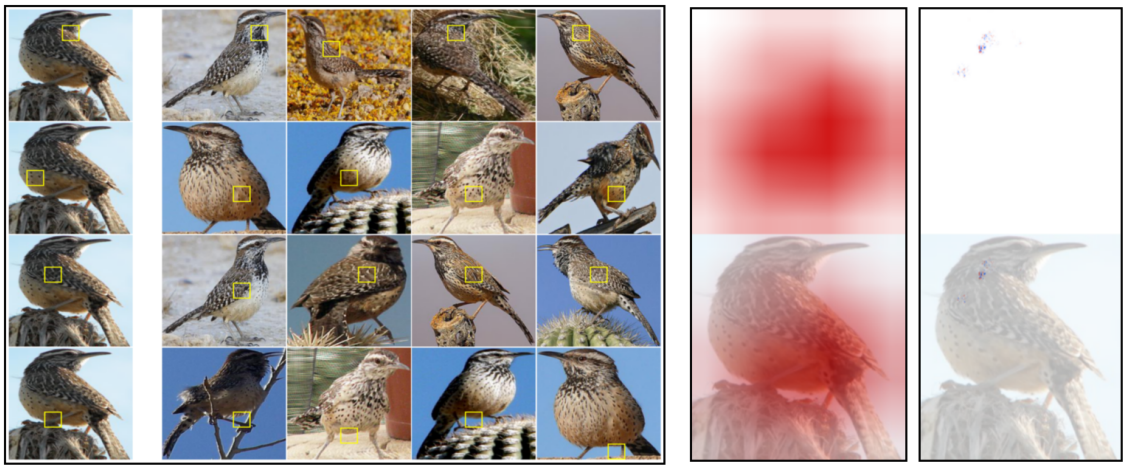
Nr rejestracyjny: 2025/57/N/ST6/04389; Kierownik projektu: lic. Piotr Borycki

Wyjaśnianie pretrenownych modeli głębokiego uczenia poprzez prototypy

Wraz z rosnącym wykorzystoniem metod sztucznej intelgiencji w praktycznych zastosowaniach, rośnie potrzeba zrozumienia w jaki sposób podjęta została każda decyzja. Wiele modeli płytkich takich jak drzewa decyzyjne stworzone są od początku z myślą o interpretowalności, nie jest to jednak prawdą w przypadku wielu złożonych modeli takich jak na przykład głębokie sieci neuronowe. W związku z tym w ostatnich latach coraz większe znaczenie zyskują tzw. metody wyjaśnialnej sztucznej inteligencji (ang. explainable AI, XAI), które mają na celu zrozumienia decyzji podejmowanych przez takie systemy.

Dotychczas stosowane metody XAI dzielą się na dwa główne podejścia. Pierwsze to metody post-hoc, czyli takie, które próbują wytłumaczyć decyzje już wytrenowanego systemu który może być wykorzystany w praktyce. Metody z tej kategorii najczęściej opierają się na analizie śladów po podjętej decyzji – na przykład w przypadku obrazów wskazaniu, które elementy zdjęcia miały największy wpływ na podjętą decyzję. Jednym z największych problemów tego typu metod jest fakt, że często oferują one zbyt ogólne lub trudne do interpretacji wyniki.

Drugim z podejść są metody ante-hoc, czyli algorytmy sztucznej inteligencji, które od samego początku tworzone są z myślą o tym, by były zrozumiałe. Jednymi z najczęściej występujących w tej kategorii metod, są metody oparte o koncept prototypów. Prototypy są przykładami, które algorytm sztucznej inteligencji widział już wcześniej, a uzasadnienie decyzji oparte o nie, polega na rozumowaniu w postaci "wybrałem tę opcję, ponieważ przypomina ona wcześniej widziany przypadek". Taki sposób myślenia jest znacznie bliższy temu, jak podejmują decyzje ludzie. W tym przypadku problemem jest natomiast fakt, że stworzenie takich systemów jest trudne, kosztowne i często nie można ich zastosować do już istniejących algorytmów.



Rysunek 1: Po lewej stronie widoczne są wyjaśnienia metody inspirowanej prototypami. Po prawej stronie widoczne są natomiast wyjąsnienia oferowane przez istniejący metody post-hoc.

Celem opisywanego projektu badawczego jest połączenie obu podejść, to znaczy stworzenie nowej metody, która będzie kompatybilna z już działającymi systemami AI, a jednocześnie będzie w stanie oferować intuicyjne i zrozumiałe wyjaśnienia, podobne do tych stosowanych w metodach opartych o pojęcie prototypów.