Streszczenie

Sztuczne sieci neuronowe są obiektami matematycznymi, które – jak każda inna metoda uczenia maszynowego – podejmują decyzje na podstawie zbioru danych treningowych. Praktyka dowiodła, że mogą one być z powodzeniem zastosowane w różnorodnych zadaniach, takich jak tłumaczenie maszynowe, rozpoznawanie obrazów, czy wykrywanie nowotworów.

Te imponujące wyniki początkowo przesłoniły ograniczenia sieci neuronowych i ich potencjalne słabości. Ostatnio jednak wykazano, że sieci neuronowe mogą być zaskakująco niedokładne, szczególnie gdy narażone są na zwodnicze przykłady (ang. adversarial examples). W rezultacie zastosowania sieci neuronowych do podejmowania szczególnie ważnych decyzji zaczęły być kwestionowane, a badania nad ich niezawodnością (ang. robustness) zyskały uwagę.

Niezawodność w uczeniu maszynowym można zdefiniować na wiele sposobów. Intuicyjnie jest to zdolność danego modelu do prawidłowego funkcjonowania pomimo zwodniczych, błędnych lub nieoczekiwanych danych wejściowych. W statystyce niezawodne modele są definiowane jako odporne na błędy spowodowane odstępstwami od założeń. Dlatego też, w szerszym znaczeniu, niezawodność w uczeniu maszynowym może być rozumiana jako zdolność do generalizowania (ang. generalization), ponieważ typowe założenie, że dane treningowe i testowe pochodzą z tego samego rozkładu, zazwyczaj nie jest spełnione w praktyce.

Niezależnie od dokładnej definicji, badania nad niezawodnością sieci neuronowych są wielowymiarowe. Jednym z kierunków jest poprawa ich zdolności do generalizowania poprzez proponowanie nowych lub ulepszonych architektur czy technik regularyzacji. Innym podejściem jest analizowanie sieci neuronowych celem ujawnienia ich niepożądanych właściwości. Wreszcie, niezawodność modeli można poprawić także pośrednio, poprzez zwiększanie ich interpretowalności. Nie zwiększa to niezawodności modelu per se, ale poprawia niezawodność całego procesu decyzyjnego.

W niniejszej rozprawie, która składa się z serii sześciu opublikowanych artykułów, przedstawiamy metody, które realizują wszystkie wspomniane powyżej podejścia.

Konrad Tomasz Żołna

Digitally signed by Konrad Tomasz Żołna
Date: 2020.06.23 09:00:19 +02'00'