Bezpieczne i wiarygodne głębokie sieci neuronowe

Paweł Morawiecki
Instytut Podstaw Informatyki PAN
pawel.morawiecki@gmail.com

1. Opis tematyki

Głębokie sieci neuronowe w ostatnich kilku latach z coraz większymi sukcesami pokonują kolejne wyzwania z zakresu grafiki komputerowej czy przetwarzania języka naturalnego. Mimo spektakularnych sukcesów algorytmy te nie są wolne od ograniczeń. Jednym z problemów stanowią tzw. zwodnicze przykłady (ang. adversarial examples), które oszukują klasyfikator mimo pozornie nieznaczących zmian w danych wejściowych [1]. Zmiany te są niewidoczne dla ludzkiego oka lecz zmieniają przewidywania nawet dobrze wytrenowanej sieci. Kolejnym wyzwaniem są ataki podczas treningu sieci, gdzie odpowiednio spreparowane obrazy, wprowadzają tzw. tylną furtkę (ang. backdoor) do modelu, która może być wykorzystana później w już wdrożonym systemie [2].

Zagadnieniem badawczym będzie opracowanie i analiza sieci neuronowych, które dawałyby pewne gwarancje w kontekście bezpieczeństwa i wiarygodności działania. Możliwe jest rozszerzenie tematyki o zagadnienie interpretowalności - kluczowe w zastosowaniach takich jak przetwarzanie danych medycznych.

2. Wymagania

a) ukończone studia drugiego stopnia z informatyki, matematyki lub fizyki
b) umiejętność programowania w języku wysokiego poziomu (np. Python)
c) podstawowa wiedza dotycząca maszynowego uczenia
d) dobra znajomość języka angielskiego
e) mile widziana znajomość biblioteki PyTorch lub Tensorflow

[1] Wenqi Wei, Ling Liu, Margaret Loper, Stacey Truex, Mehmet Emre Gursoy, Yanzhao Wu, Yanzhao Wu: Adversarial Examples in Deep Learning: Characterization and Divergence, https://arxiv.org/pdf/1807.00051.pdf

[2] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, Biplav Srivastava: Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering, https://arxiv.org/abs/1811.03728