Neurális hálók robusztusságának vizsgálata Taylor modell használatával

Szabó Tamás
II. évf. programtervező informatikus MSc

Horváth János
I. évf. programtervező informatikus MSc

Témavezető: Dr. Bánhelyi Balázs

SZTE TTIK Számítógépes Optimalizálás Tanszék

Manapság mesterséges intelligencia egyre több területen kerül alkalmazásra, és nem csak tudományos munkában. Az egyszerűbb, ártalmatlanabb használati eset közé sorolható az okostelefonokba épített felismerő rendszerek (milyen állat található a képen?). Egy biztonság-kritikusabb terület lehet azonban a jövőben valószínűleg nagy számban megjelenő önvezető autók. Ezek járművek fontos részét képezi majd egy objektumszegmentáló és -felismerő rendszer is. A neurális hálózatokat pedig tipikusan ilyen feladatokra szokták használni. Ezen hálózatok lényege, hogy megfelelő mennyiségű adat hozzáférhetősége esetén különböző feladatokra taníthatóak be, az egyik elterjedt alkalmazásuknak tekinthető az objektumfelismerés képről.

Viszont pont amiatt, hogy egyre elterjedtebb a neuronhálók használata, fontos, hogy elég pontosak és biztonságosak is legyenek, ne lehessen összezavarni őket. Egy 2014-es cikkben (Szegedy et al.) megmutatták, hogy bizonyos technikákkal van lehetőség olyan adverzális/ellenséges példákat létrehozására kis perturbáció hozzáadásával más képekből, melyek emberi szem számára nem észrevehetőek, azonban a hálózatokat képesek megtéveszteni.

Léteznek már a piacon különböző robusztusságvizsgáló rendszerek, melyekkel van lehetőség neuronhálók ellenőrzésére, ilyen például az ERAN és a MIPVerify. Mi is egy ilyen rendszert hoztunk létre Taylor modellre alapozva MATLAB/INTLAB környezetben Sigmoid, Tanh és ReLU aktivációs függvénnyel rendelkező hálózatok ellenőrzésére.