Megbízható megoldás keresése a neurális hálók robusztusságának vizsgálatára

Horváth János
III. évf. programtervező informatikus
Szász Attila
II. évf. programtervező informatikus

Témavezetők: Bánhelyi Balázs, Zombori Dániel

SZTE TTIK Számítógépes Optimalizálás Tanszék

A mesterséges neurális hálók nagy kifejezőerővel rendelkező eszközök, azonban jelentős sebezhetőséget rejthetnek magukban. Előfordulhat olyan tanításra használt bemenet, amely megfelelő kimenetet generál, de a vártnál kisebb környezetében előállítható hibás kimenet is. Ilyen ellenséges példák keresésére ad lehetőséget a MIPVerify rendszer, ami ma a szakirodalomban leginkább elfogadott algoritmus erre a problémára. A MIPVerify egy lineáris rétegekből álló neuronhálóban megkeresi a bemenethez legközelebbi ellenséges példát. A feladatot MILP problémák sorozataként fogalmazza meg és külső solver segítségével szolgáltat megoldást. Az ilyen eszközök azonban véthetnek numerikus hibát, melyek a további feladatokra tovább terjedve elrejthetnek bizonyos ellenséges példákat. Ezen numerikus hibákat egy egzakt számábrázolást alkalmazó megoldó nem követi el, viszont a feladatokat jelentősen lassabban oldja meg. Megvizsgáltuk, hogy különböző méretű neurális hálók verifikálása esetén mennyire növekszik az időigény, ha az SCIP egzakt megoldót használjuk.