Neuronhálok serülékenysége és robosztusságának vizsgálatára használt eljárás fejlesztése

Bánhelyi Balázs

Szegedi Tudományegyetem, Informatikai Intézet

banhelyi@inf.u-szeged.hu

A mesterséges neurális hálók számos tudományterületen megjelennek. Megfigyelhető, hogy bizonyos esetekben ezek a hálózatok is tévedhetnek. Gyakran az input kis torzítására már fals eredménnyel térnek vissza [1]. Az ilyen hibák kiküszöbölésére számos módszer létezik. A robusztus tanítás már a tanítási folyamat alatt megpróbálja csökkenteni a háló sebezhetőségét és növelni az ellenállóképességet. Más technikák, a már kész hálókon történő ellenséges példa detektáláson alapulnak. Számos matematikailag korrektnek gondolt rendszer létezik ellenséges példák detektálására, de gyakran ezek implementálásakor a praktikusságra koncentrálnak, a numerikus hibák kiküszöbölése helyett. Ezek a numerikus hibák a hálózat működése közben is megjelennek és a rétegek alatt folyamatos felhalmozódnak, mely szintén hibás osztályozáshoz vezethet. A példák detektálására a MIPVerify az adott input képekhez, a lefixált perturbáció típus és hozzá tartozó korlát mellett keresi az adott korlátokon belüli, legközelebbi ellenpéldát és határozza meg azokat a perturbáció értékeket, melyekkel deformálva az eredeti inputot, már téves eredményt kapunk [2]. A MIPVerify különböző MILP feladatok sorozataként fogalmazza meg a problémát, amelyek megoldására külső solverek alkalmazhatóak. A rendszer működéséből adódó pontatlansági hibák nagy része a lebegőpontos aritmetikából fakad. Az előadásunkban bemutatjuk a MIPVerify sebezhetőségeit [3], illetve mutatunk technológiákat melyek ezen sebezhetőségeket kezelik, miközben a hatékonyságából nem veszít.

Hivatkozások

[1] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In 3rd International Conference on Learning Representations (ICLR), 2015.

[2] Vincent Tjeng, Kai Y. Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. In International Conference on Learning Representations (ICLR), 2019.

[3] Dániel Zombori, Balázs Bánhelyi, Tibor Csendes, István Megyeri, Márk Jelasity. Fooling a complete neural network verifier. In International Conference on Learning Representations (ICLR), 2021.