Neurális hálózatok paraméterrobusztussága

Szász Attila

Szegedi Tudományegyetem, Informatikai Intézet

szasz@inf.u-szeged.hu

A neurális hálózatok kiemelkedő figyelmet kaptak az elmúlt években, mind a felhasználói, mind a kutatási oldalról. Napjainkban számos területen alkalmazzák őket, ilyen például a számítógépes látás és a beszédfelismerés. A kutatások döntő többsége az egyre pontosabb és megbízhatóbb hálózatok előállításának irányába mozdult el. A megbízható, más néven robusztus hálózatok tanításának érdekében számos tanítási technikát javasoltak a kutatók. A módszerek döntő többsége két csoportba sorolható. Az adverzális tanítás [1] alapú algoritmusok ellenséges példák felett optimalizálják a hálózatok paramétereit. A minősített [2] (*certified*) tanítás alapú módszerek pedig befoglalást számítanak a hálózat kimeneteire, majd a korlátok alapján feltételezhető legrosszabb esetet minimalizálják. A bemenet fókuszú támadások mellett, a hálózatok paraméter támadása is előtérbe került. Az ilyen típusú támadások a hálózatok paramétereit módosítják és ezáltal váltják ki az ellenséges viselkedést. Ezek alapján kidolgoztak olyan tanítási módszereket, melyek a hálózatok paraméterstabilitásának elérését is beépítik a tanítási folyamatba, melyek közül a szakirodalomban legelterjedtebb az *Adverzális Súly Perturbáció (AWP)* [3] módszere lett. Az algoritmus legfőbb hátránya, hogy a legrosszabb esetet erősen alulbecsüli, így nem nyújt kellő védelmet a paramétertámadások ellen. Kutatásunk során feltártuk az AWP algoritmus legjelentősebb gyenge pontját, és javasoltunk egy certified alapú tanítóalgoritmust, amely számos esetben növelte a hálózatok ellenállóképességét a paramétertámadásokkal szemben.

Hivatkozások

[1] Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. (2019)

[2] Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T. & Kohli, P. On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models. (2019)

[3] Wu, D., Shu-Xia & Wang, Y. Adversarial Weight Perturbation Helps Robust Generalization. (2020)