Robusztus modellvarrás: neurális reprezentációk hasonlósága a pontosságon túl

Balogh András
II. évf. Programtervező informatikus MSc

Témavezető: Jelasity Márk

SZTE TTIK Számítógépes Algoritmusok és Mesterséges Intelligencia Tanszék

Napjainkban a neurális hálózatok a gépi tanulás legelterjedtebb modelljei, azonban magas teljesítményüket árnyalják a működésükkel kapcsolatos nyitott, aktívan kutatott kérdések. Ilyen kérdéskör a hálózatok támadhatósága, amely során pl. képosztályozó hálózatok bemenetéhez emberi szemmel érzékelhetetlen perturbációk hozzáadásával a modellek kimenetei drasztikusan megváltoztathatók, ezáltal a hálózatok félrevezethetők.

Az ilyen támadás ellen védett (ún. robusztus) és nem robusztus hálózatok közötti különbségek feltárása fontos probléma, mivel ezzel felfedhető, hogy a támadások a hálózaton belül hol fejtik ki a hatásukat. Ezekből lehet következtetni a nem robusztus hálózatok gyenge pontjaira, amelyek ismerete elősegítheti robusztusabb hálózatok előállítását. A hálózatok összehasonlításának egyik meghatározó módszere a hálózatokat alkotó rétegek és az általuk kinyert reprezentációk elemzése. A dolgozat célja a robusztus és nem robusztus hálózatok reprezentációi közötti funkcionális hasonlóságok és eltérések vizsgálata.

Reprezentációk hasonlóságának vizsgálatára több módszer is ismert, a dolgozatban ezek közül a modellvarrást használom. Ennek segítségével megmutatom, hogy a robusztusság a reprezentációk funkcionális tulajdonsága, ezáltal a reprezentációk funkcionális hasonlóságát szükséges az osztályozási pontosság mellett a robusztusság szempontjából is vizsgálni. A robusztus funkcionális hasonlóság pontosságtól elkülönített vizsgálatának céljából bevezetem a robusztus modellvarrás módszerét.

A dolgozatban standard és robusztus modellvarrás használatával megmutatom, hogy robusztus és nem robusztus reprezentációk az osztályozási pontosság szempontjából funkcionálisan hasonlók, azonban a robusztusság szempontjából a funkcionális eltérések a bemenethez közeli rétegekben keresendők. Az eredményeim továbbá mutatják, hogy robusztus reprezentációk a pontosság megtartása mellett a robusztusságot megőrző és elvesztő módon egyaránt varrhatók. Ezen felül megállapítom, hogy robusztus és nem robusztus hálózatok az osztályozási feladatot azonosan, a reprezentációik klaszterezésével oldják meg.