Abstrakt

V priebehu ostatných pár rokov sa ukázalo, že väčšina hlbokých neurónových sietí vykazuje citlivosť voči zákerne vygenerovaným vstupom, zvaným adverzariálne. Tieto vstupy sú zvyčajne vytvorené z originálnych dát pridaním malého, no veľmi špecifického šumu, ktorý spôsobuje veľkú chybu na výstupe siete.

Napriek intenzívnemu výskumu v tejto oblasti, problém stále pretrváva a nie je uspokojivo vyriešený. Početné obrany proti adverzariálnym vstupom, ako aj metódy ich detekcie, boli navrhnuté, nanešťastie, žiadna z nich neposkytuje úplnú robustnosť voči adverzariálnym útokom. Z toho dôvodu sa ťažisko výskumu pomaly presúva k analýze adverzariálnych vstupov a skúmaniu dôvodov prečo a ako spôsobujú chybovosť hlbokých neurónových sietí.

V našej práci sa venujeme viacerým experimentom v doméne klasifikácie obrázkov, za cieľom lepšieho pochopenia vlastností adverzariálnych vstupov a mechanizmov, ktorými spôsobujú chybné výstupy. Konkrétnejšie sa náš prínos dá rozdeliť do dvoch hlavných oblastí: 1) analyzujeme inherentnú robustnosť sietí so sigmoidálnou aktiváciou, a ako súvisí so saturáciou jednotlivých neurónov. Navyše vyšetrujeme robustnosť novo navrhnutej architektúry RecViT (Recurrent Vision Transformer) a skúmame rôzne trénovacie stratégie na zvýšenie jej robustnosti; a 2) vyhodnocujeme teóriu rozbaľovania manifoldov (manifold disentanglement theorem) viacerými metódami, následne nadväzujeme návrhom nového algoritmu na analýzu skrytých reprezentácií siete. Tento algoritmus je založený na analýze blízkosti konkrétnych vstupov k manifoldom originálnych dát, a v našich experimentoch ho používame na porovnanie správania viacerých rôznych typov adverzariálnych vstupov.

Kľúčové slová: adverzariálne útoky, robustnosť, hlboké neurónové siete, klasifikácia obrázkov, vysvetliteľnosť