V Praze dne 11. 8. 2022

Posudek oponenta na bakalářskou práci

Robustní strojové učení a adversariální vzorky

Autor práce: **Pavel Jakš**

Práce se zabývá problematikou robustního strojového učení. Konkrétně se zaměřuje na fenomén adversariálních vzorků, neboli vstupů speciálně designovaných ke zmatení cílového modelu. Cílem práce bylo seznámení se s problematikou adversariálních vzorků, a to jak z pohledu útočníka tvořícího dané vzorky, tak z pohledu obránce který vytváří model robustní vůči těmto útokům.

Práce je rozdělena do sedmi kapitol. V první kapitole autor popisuje hlavní typy neuronových sítí a jejich komponenty. Ve druhé kapitole autor popisuje problém učení neuronových sítí a srovnává základní algoritmy používané na řešení tohoto problému. Třetí kapitola je věnovaná konceptu adversariálních vzorků a poskytuje popisu hlavních algoritmu pro jejich hledání jako jsou metody FGSM a PGD. Čtvrtá kapitola popisuje problém učení robustní sítě. V páté kapitole autor v empirickém experimentu na klasifikaci rukou psaných číslic srovnává algoritmy učení neuronové sítě. Šestá kapitola je zaměřena na empirické experimenty s adversariální vzorky a autor v ní porovnává útoky různých algoritmu na standardní sítě. V sedmé kapitole aplikuje ty samé útoky na robustní sítě a poskytuje porovnání.

Při zpracování tématu si autor si nastudoval a popsal základní koncepty robustního strojového učení adversariálních vzorku. Popis metod je napsán přehledně s důrazem na korektní matematickou formulaci problémů. Přehledová část práce se dobře čte a jednotlivé části na sebe logicky navazují. Autor samostatně naimplementoval vybrané algoritmy a poskytl experimentální pohled na rozdíly mezi jednotlivými algoritmy adversariálních útoků. Výsledky experimentu, ať už v podobě obrázků nalezených adversariálních vzorku či tabulek nejsou přehledně popsané, což vede k horší orientaci v provedených experimentech. Celkově by výsledky experimentální části zasloužily detailnější popis.

Vzhledem ke splnění všech bodů zadání doporučuji práci k obhajobě. Jinak kvalitní práci brzdí horší experimentální část. Například u kvantitativních výsledků není srovnání metod korektní. Metody FGSM/I-FGSM/PGD vždy splní omezení normou, ale negarantují úspěšnost útoku. Naopak, metoda CW útoku optimalizuje úspěšnost ale negarantuje splnění omezeni normou. Autorem provedené kvalitativní srovnání adversariálních vzorku není také vypovídající, protože autor srovnává útoky omezené různou normou se stejnou tolerancí. Vzhledem k přihlédnutí k těmto nedostatkům **navrhuji známku C**.

K obhajobě bych měl následující dotazy:

- V práci se píše, že FGSM, I-FGSM a PGD metody jsou necílené. Jak lze tyto metody změnit, aby je šlo použít pro cílené útoky?

- V práci používáte fixní počet iteraci při trénování. Podle čeho jste určoval počet iterací algoritmu?

- V experimentální části metoda CW útoku vykazuje v mnoha výsledcích téměř 100% úspěšnost útoku.
    - Znamená to, že metody FGSM/I-FGSM/PGD jsou horší?
    - Jak by jste upravit evaluaci aby bylo srovnání s metodami FGSM/I-FGSM/PGD více vypovídající?

- V práci na několika místech kvalitativně srovnáváte adversariální vzorky pro různé normy se stejnou mírou tolerance. Který adversariální vzorek bude vizuálně blíže původnímu vzorku?
    - Perturbace je omezena euklidovskou normou s tolerancí 0.5
    - Perturbace je omezena maximovou normou s tolerancí 0.5

Mgr. Vojtěch Čermák