Review of Bachelor's Thesis

Student: Gaňo Martin
Title: Improving Robustness of Neural Networks against Adversarial Examples (id 22999)
Reviewer: Matyáš Jiří, Ing., DITS FIT BUT

1. Assignment complexity — **more demanding assignment**

Zadání považuji za náročnější vzhledem k tomu, že se student musel zorientovat v rychle se rozvíjející problematice klasifikace pomocí neuronových sítí a způsobů jejich napadení pomocí protipříkladů. Toto téma přesahuje rámec znalostí běžně probíraný v bakalářském studiu.

2. Completeness of assignment requirements — **assignment fulfilled**

Student úspěšně splnil všechny body zadání.

3. Length of technical report — **in usual extent**

4. Presentation level of technical report — **80 p. (B)**

Technická zpráva je přehledně strukturovaná a členěná do logicky navazujících kapitol. Pojmy ze zkoumané oblasti definuje a vysvětluje srozumitelně, navíc s použitím ilustračních příkladů a obrázků. Úvod a závěr práce jasně vymezují záměr práce, její přínos a dosažené výsledky.

5. Formal aspects of technical report — **65 p. (D)**

Práce je psaná anglicky a její jazyková stránka bohužel kvalitou výrazně zaostává za technickým přínosem a prezentační úrovní. Text obsahuje množství překlepů, chyb, chybějících členů, špatných tvarů sloves a nevhodně použitých výrazů. Popisy os v grafech 5.3 a 5.4 jsou příliš malé vzhledem k okolnímu textu. Úroveň textu by se dala výrazně vylepšit pozorným průchodem a korekturou.

6. Literature usage — **75 p. (C)**

Student použil zdroje relevantní k tématu práce.

7. Implementation results — **90 p. (A)**

Realizační výstup práce poskytuje možnost napadat libovolný model neuronové sítě pro obrazovou klasifikaci pomocí různých typů útoků a také trénovat klasifikační modely proti těmto útokům. Existující nástroje neposkytovaly takovou škálu dostupných útoků a možných způsobů obrany. Navíc dříve dostupné nástroje byly často specializované pouze pro určitý typ neuronové sítě a nebylo je možné použít pro obecné modely -- výstup práce toto umožňuje. V budoucnu je možné nástroj dále rozšiřovat o nové techniky útoků a zvyšování robustnosti sítí.

8. Utilizability of results

Výstup práce umožňuje vývoj robustních modelů neuronových sítí, odolných vůči útokům pomocí protipříkladů. V nástroji je implementováno několik typů útoků a obran dostupných v literatuře a může být dále rozšiřován o nové typy útoků a metody obrany.
Experimentální vyhodnocení robustnosti neuronových sítí proběhlo na datových sadách MNIST a CIFAR10, které jsou řešitelné pomocí poměrně jednoduchých neuronových sítí. Experimenty s většími datovými sadami vzhledem k výpočetní náročnosti nebyly možné.

9. Questions for defence

- Jaké struktury (hloubky a použité vrstvy) neuronových sítí byly použity v rámci experimentů - vlastní navržené nebo dostupné z literatury?
- Jak by se změnila robustnost vůči útokům, pokud by byly použity sítě s více či méně vrstvami?

10. Total assessment — **80 p. very good (B)**

Jedná se o práci s obtížnějším zadáním, při jejímž vypracování vznikl kvalitni realizační výstup. Výsledný nástroj umožňuje trénování za účelem zvýšení robustnosti obecných modelů neuronových sítí proti několika typům útoků. Nástroj může být v budoucnu dále rozšiřován.
Slabinou práce je jazyková a typografická stránka výsledného dokumentu, který byl nejspíše dokončován na poslední chvíli.
S přihlédnutím k náročnosti zadání a kvalitnímu realizačnímu výstupu hodnotím práci stupněm 80B.

In Brno 23 June 2020

Matyáš Jiří, Ing.
reviewer