Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. Matěj Kocián
Název práce	Adversarial Examples in Machine Learning
Rok odevzdání	2018
Studijní program	Informatika **Studijní obor** Umělá inteligence

Autor posudku	Mgr. Martin Pilát, Ph.D. **Role** vedoucí
Pracoviště	KTIML MFF UK

Text posudku:

Ukazuje se, že pro mnoho metod pro klasifikaci obrázků lze jednoduše pozměnit některé vstupy tak, že pro lidi jsou rozdíly nerozeznatelné, ale metody pro ně předpovídají zcela jinou (chybnou) třídu než pro původní vstupy. Takové vzory se nazývají matoucí. Ve své práci se Matěj Kocián zabývá právě takovými matoucími vzory. Jedná se o velmi aktuální a aktivně zkoumané téma.

Práce je rozdělena do třech kapitol. První kapitola představuje úvod do (hlubokých) neuronových sítí a jiných klasifikačních metod. Druhá kapitola potom přináší shrnutí předchozích prací v oblasti hledání matoucích vzorů a hledání modelů odolných proti matoucím vzorům. Vlastní přínos práce (dvě metody pro obranu před matoucími vzory a jedna metoda pro jejich vytváření) je soustředěn ve třetí kapitole.

Obě úvodní kapitoly přináší velmi dobrý úvod do zkoumané problematiky. Druhou kapitolu, věnující se matoucím vzorům, považuji za aktuálně nejlepší shrnutí existujících metod a věřím, že i sama o sobě je významným přínosem pro tuto oblast.

Jak jsem již zmínil, vlastní přínos práce je v kapitole třetí. Zde se student napřed zabývá hledáním metod pro obranu před matoucími vzory. Vychází z myšlenky, která se často objevuje v literatuře o matoucích vzorech, že RBF sítě by mohly být odolnější vůči matoucím vzorům. Tato hypotéza je otestována ve srovnání s běžnými konvolučními sítěmi a ukazuje se, že tomu tak není. Nicméně, student zde navrhuje novou architekturu hlubokých sítí (kterou nazývá RBFoluční), která nahrazuje konvoluce právě RBF jednotkami. Ukazuje se, že tato nová architektura je odolnější než RBF sítě, i než konvoluční sítě. Student potom ještě testuje použití zaokrouhlování jako metody pro obranu před matoucími vzory (po zaokrouhlení by se matoucí vzory mohly zobrazit na stejné vstupy jako původní vzory). Tato metoda je úspěšná na jednom ze dvou testovaných datasetů. V poslední částí této kapitoly potom student testuje možnost najít jednu masku, která by po přičtení k libovolnému obrázku způsobila, že výsledný obrázek bude klasifikován špatně (jde o hledání matoucích vzorů v konstantním čase). Experimenty ukazují, že takový přístup vytváření

matoucích vzorů je možný a snižuje úspěšnost modelu z více než 99 % na 65 %.

Popisy jednotlivých metod jsou dostatečně podrobné a jsou prezentovány v souvislosti s existující literaturou. Experimenty jsou dobře popsané a vyhodnocené. Celá práce je psána velmi dobrou angličtinou. Student se v ní seznámil s velkým množstvím existující literatury a dobře ji shrnul. Jak sepsaná rešerše, tak nově vytvořené metody jsou významným přínosem do zkoumané oblasti. Student ukázal, že je schopen samostatné práce.

Práci doporučuji k obhajobě.

Práci navrhuji na zvláštní ocenění.

Práce přináší velmi dobré a významné výsledky v oblasti hledání matoucích vzorů. Student navrhl vlastní model hluboké neuronové sítě, která je vůči matoucím vzorům odolnější. Celkově práce rozhodně patří mezi nejlepší, které jsem vedl.

V Praze dne 4. června 2018

Podpis: