Posudek bakalářské práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Jakub Hejhal
Název práce	Výzkum zranitelnosti reálných systémů umělé inteligence vůči adversariálním útokům
Rok odevzdání	2021
Studijní program	Informatika
Studijní obor	Obecná informatika
Autor posudku	Mgr. Martin Pilát, Ph.D. Oponent
Pracoviště	Katedra teoretické informatiky a matematické logiky

K celé práci

	lepší	OK	horší	nevyhovuje
Obtížnost zadání		X		
Splnění zadání	X	X		
Rozsah práce *... textová i implementační část, zohlednění náročnosti*		X		

Práce se zabývá testováním zranitelnosti reálných systémů pro rozpoznávání obrázků vzhledem k tzv. nepřátelským vzorům. To jsou vstupy speciálně upravené tak, aby pro člověka tato úprava nebyla viditelná, ale pro model je matoucí - předpovídá pro ni jinou třídu, než do jaké vzor skutečně patří. Tento cíl se podařilo splnit a byl vytvořen nástroj, který umožňuje vytváření těchto matoucích vzorů. Nástroj byl následně otestován proti Google Vision API.

K práci mám pouze jednu otázku: Při mapování výstupů z cloudového klasifikátoru do tříd organismů a ne-organismů se používá relativně složitý postup. Nestačilo by i zde použít WordNet a podívat se, jestli některá z nadřazených tříd je organismus? Chápu, že zvolený přístup je obecnější, ze zbytku práce se mi ale nezdá, že by tato obecnost byla potřeba.

Textová část práce

	lepší	OK	horší	nevyhovuje
Formální úprava *... jazyková úroveň, typografická úroveň, citace*		X		
Struktura textu *... kontext, cíle, analýza, návrh, vyhodnocení, úroveň detailu*		X		
Analýza	X	X		
Vývojová dokumentace		X	X	
Uživatelská dokumentace		X	X	

Textová část práce je velmi dobře napsána, především potom přehled existujících metod pro vytváření nepřátelských vzorů. Na druhou stranu popis vlastního systému (AdvPipe) by mohl být podrobnější, není úplně jasné, co vše je nakonec implementováno. Toto je navíc umocněno tím, že k práci není přiložena žádná podrobnější dokumentace, kterou se mi nepodařilo najít ani ve zdrojových kódech. Ačkoliv je u tohoto typu práce jistě důležitější samotný text a implementace, dokumentace by neměla být opomenuta.

Implementační část práce

	lepší	OK	horší	nevyhovuje
Kvalita návrhu *... architektura, struktury a algoritmy, použité technologie*		X		
Kvalita zpracování *... jmenné konvence, formátování, komentáře, testování*		X		
Stabilita implementace		X		

Zdrojové kódy bohužel nejsou k práci přímo přiloženy – v práci je pouze odkaz na GitHub. To je sice víceméně standardní postup ve vědecké komunitě, nicméně u bakalářské práce by bylo vhodnější kódy odevzdat i společně s prací v příloze. Nicméně implementace je na velmi dobré úrovni, nechybí ani skripty pro zopakování experimentů, což výrazně zlepšuje replikovatelnost výsledků práce.

Celkové hodnocení Výborně

Práci navrhuji na zvláštní ocenění Ne

Datum 23. srpna 2021 Podpis