Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Matěj Kocián		
Název práce	Adversarial examples in machine learning		
Rok odevzdání	2018		
Studijní program	Informatika	**Studijní obor**	Teoretická informatika

Autor posudku	Roman Neruda	**Role**	Oponent
Pracoviště	Ústav informatiky AV ČR		

Text posudku:

Předkládaná práce se zabývá velmi aktuálním tématem robustnosti modelů strojového učení. Klíčovým tématem je studium adversariálních vzorů v klasifikačních problémech, které vykazují podobnost se zástupci určité třídy, ale jsou modelem strojového učení klasifikovány jako zástupci třídy jiné. V současnosti se v odborné literatuře věnuje velká pozornost adversariálním vzorům zvláště v souvislosti s modely hlubokých neuronových sítí použitých na klasifikaci obrazových dat. Matěj Kocián ve své práci provádí rešerši současných metod a navrhuje několik původních metod obrany proti adversariálním vzorům.

Struktura práce je následující. Cíle a struktura práce jsou specifikovány v úvodu. V první kapitole jsou popsány základní pojmy strojového učení a modelů hlubokých neuronových sítí, které jsou dále používány. Druhá kapitola představuje úvod do aktuálního stavu problematiky adversariálních vzorů. Jsou zde popsány hlavní metody jejich generování, jejich vlastnosti i možnosti obrany. Vlastní autorovy výsledky v oblasti adversariálních vzorů jsou obsažené v kapitole tři, která se postupně věnuje použití RBF sítí, diskretizace vstupů, a metody útoku pomocí adversariálních perturbací. Výsledky práce a možnosti dalšího pokračování jsou stručně shrnuty v závěru.

Práce měla naplnit tři postupné cíle, provést zevrubný přehled výsledků v oblasti adversariálních vzorů, navrhnout nové možnosti obrany proti adversariálním útokům, a také přijít s metodami jak adversariální vzory generovat.

První cíl práce byl naplněn v kapitole druhé, která obsahuje podrobný přehled metod i relevantní literatury. Jde o oblast velmi novou a autor práce má dobrý přehled i v nejnovějších článcích publikovaných v preprintech na platformě arXiv.

Za největší přínos práce považuji naplnění druhého cíle, t.j. návrh možností obrany proti adversariálním vzorům. Předně je nutno zmínit, že metoda útoku perturbacemi pomocí FGSM se v praxi ukazuje jako velmi efektivní, takže šlo o nesnadný úkol. Autor se vydal cestou návrhu nové architektury hluboké sítě, kde využil předchozí teoretické i praktické výsledky o rozdílu v odolnosti jednotek, které počítají lineární kombinace vstupů a těch, kde vstupy procházejí nelineární transformací. Autorem navržená architektura kombinuje konvoluční sítě s lokálními jednotkami typu RBF (Radial Basis Functions) a dle experimentálních výsledků vykazuje větší odolnost vůči adversariálním útokům. Druhá původní metoda ochrany spočívá v diskretizaci vstupů sítě, které se zaokrouhlují v extrémním případě až na binární hodnoty. Je intuitivní, že takový přístup by měl zaručit odolnost modelu vůči nepřátelským vzorům za cenu snížení přesnosti klasifikace. Velkým přínosem autora je, že danou metodu

experimentálně prověřil na dvou datových množinách a definoval optimální míru kompromisu mezi klasifikační chybou a robustností modelu.

Třetím cílem práce byl návrh a experimenty s novými metodami adversariálních útoků. Zde se autor zaměřil na generování společných perturbací, tedy bez ohledu na vzor, který se perturbuje. Tuto metodu experimentálně ověřil na cílené i necílené útoky proti konvoluční síti na datové množině MNIST.

Obecně mi naplnění třetího cíle přijde jako nejméně rozsáhlé a zůstává v něm řada možností k pokračování. Zdůvodnění o větším významu obrany proti útokům považuji tedy za subjektivní až diskutabilní. Na druhou stranu, práce mi přijde i tak velmi povedená a svými výsledky nadprůměrná. Všechny experimenty v práci prezentované jsou dobře zdůvodněny, popsány, a jsou pečlivě vyhodnoceny, takže závěry na jejich základě jsou věrohodné a opodstatněné. Celá práce je velmi přehledně a logicky uspořádána a je srozumitelná. Použitá literatura je relevantní a obsahuje velké množství velmi aktuálních článků.

Dle mého názoru autor nadstandardním způsobem naplnil nároky kladené na diplomovou práci. Jako otázku na zamyšlení při obhajobě bych rád slyšel o možnostech využití popsaných metod pro jiná data než obrazová.

Práci doporučuji k obhajobě.

Práci navrhuji na zvláštní ocenění.

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Práce se zabývá významným a velmi aktuálním tématem v oblasti strojového učení a její úroveň je nadstandardní.

Datum 6. června 2018 **Podpis**