Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Vojtěch Čermák		
Název práce	Adversarial examples design by deep generative models		
Rok odevzdání	2021		
Studijní program	Informatika	**Studijní obor**	Umělá inteligence

Autor posudku	Roman Neruda	**Role**	Vedoucí
Pracoviště	ÚI AV ČR		

Text posudku:

Předkládaná práce se zabývá generováním matoucích (adversariálních) vzorů, které způsobí chybnou klasifikaci modelu hlubokých neuronových sítí pro rozpoznávání obrazu. Matoucí vzory mají tu vlastnost, že je model klasifikuje do špatné třídy, ale pro lidského pozorovatele by měli působit jako věrohodný reprezentant třídy původní. Adversariální vzory a možnost obrany vůči nim jsou velmi aktuální výzkumné téma, které přímo souvisí s bezpečnostními aspekty použití modelů strojového učení v praxi. Autor v práci navrhuje dva algoritmy generování matoucích vzorů v prostoru latentních reprezentací a na experimentech ukazuje jejich účinnost na tradičních datasetech MNIST, SVHN a CIFAR-10.

Práce je členěna do pěti kapitol a přílohy. První kapitola uvádí do kontextu matoucích vzorů a ukazuje jejich význam pro pochopení mechanismů obrany klasifikátoru obrazových dat. V úvodu je také podrobněji stanoven cíl práce a představena její struktura.
Druhá kapitola zavádí modely hlubokých sítí, které se používají v dalším textu. Autor představuje hluboké generativní modely, které se učí bez učitele distribuci dat a jsou schopny generovat nové vzorky z naučené distribuce (kapitola 2.2). Důraz je kladen na variační autoenkodéry a generativní adversariální sítě (GAN). V kapitole 2.3 se formálně definují matoucí vzory a představí se tradiční gradientní metoda jejich generování FGSM.
Třetí kapitola shrnuje existující literaturu v oblasti adversariálních útoků a obrany proti nim, a v oblasti hlubokých generativních modelů.
Jádrem teoretické části práce je kapitola 4, která zavádí koncept perturbací latentního prostoru generativních modelů. Hlavní výsledky jsou obsaženy v kapitolách 4.3 a 4.4, kde autor ukazuje dva postupy hledání matoucích vzorů v latentním prostoru. První přístup je založen na lineární interpolaci (Algoritmus 3) a druhý přístup na odhadnu gradientu (Algoritmus 4).
Pátá kapitola na experimentech demonstruje účinnost algoritmů na kombinaci různých generativních modelů a tří datasetů. V sekci 5.2.4 se vygenerované matoucí vzory použijí pro úspěšný útok na robustní klasifikátor datasetu MNIST.
Závěr obsahuje shrnutí práce, komentuje dosažené výsledky a naznačuje další možnosti pokračování v práci.

Za hlavní přínos práce považuji následující body:

- Princip hledání adversariálních vzorů perturbacemi v latentním prostoru je originální a umožňuje generovat vzory na základě změn příznaků, které mohou reprezentovat obecnější vlastnosti obrazových dat. Tyto vzory mohou být úspěšnější při útoku než klasické přímé perturbace obrazových dat. Obě navržené metody umožňují jemný průzkum vzorů poblíž hranice klasifikátoru.

- Gradientní metoda v latentním prostoru se inspiruje přístupem FGSM pro odhad gradientu a umožňuje efektivnější určení směru perturbací vzhledem k hranici klasifikátoru a tím nalezení lepších matoucích vzorů.

- Experimentální část ukazuje, že matoucí vzory generované v latentním prostoru mají pravděpodobně jiné vlastnosti než vzory generované klasickým algoritmem FGSM. I útok na robustní klasifikátor učený augmentací s FGSM vzory se ukázal jako velmi úspěšný, takže dosavadní metody obrany vůči tomuto typu matoucích vzorů selhávají.

- Z hlediska obrany je důležité hledat nové typy matoucích vzorů, které pak mohou sloužit jako trénovací sada pro robustnější modely. Z tohoto hlediska je předkládaná práce důležitým krokem k hledání lepších obranných mechanismů.

Závěrem bych rád shrnul, že ve své práci autor navrhl novou metodu generování matoucích vzorů, která se ukazuje jako velmi úspěšná i proti state-of-the-art robustním klasifikátorům. Jde o hodnotný výsledek, který si zaslouží publikaci. Rád doporučuji tuto nadprůměrnou práci k obhajobě.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

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

Datum 11. června 2021 **Podpis**