Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Vojtěch Čermák
Název práce	Tvorba nepřátelských vzorů hlubokými generativními modely
Rok odevzdání	2021
Studijní program	Informatika **Studijní obor** Umělá inteligence

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

Text posudku:

Práce Vojtěcha Čermáka se zabývá důležitým problémem tzv. nepřátelských vzorů v hlubokém učení. Konkrétně student navrhuje použití generativních modelů k vytváření těchto vzorů, které mohou být libovolně daleko do trénovacích vzorů, ale zároveň jsou lidmi stále snadno rozpoznatelné.

Celá práce je rozdělena do pěti kapitol (včetně úvodu). V úvodu student popisuje nepřátelské vzory jako takové a jejich historii společně se strukturou práce. Další dvě kapitoly potom popisují neuronové sítě obecně a následně zmiňují hluboké generativní modely a podrobněji diskutují i problém nepřátelských vzorů. Tyto kapitoly je napsány relativně čtivě a srozumitelně, nicméně jsou v nich občas drobné nepřesnosti (například zmiňují, že mezi nejčastěji používané aktivační funkce patří arcus tangens, správně by ale měl být hyperbolický tangens). Také mi hlavně na začátku druhé kapitoly chybí odkazy na existující literaturu, jak obecný úvod do neuronových sítí, tak popis VAE žádné neobsahují. Moc zřejmá také není motivace pro rozdělení tohoto popisu do dvou kapitol, kde se opakují ty samé podkapitoly, připadalo by mi přirozenější obě kapitoly sloučit do jedné.

Hlavní přínos práce je ve čtvrté kapitole, kde student popisuje nové přístupy pro vytváření nepřátelských vzorů. K tomu používá latentní prostor generativních modelů, ve kterém postupně mění vstupy tak, aby se přiblížil od jednoho vzoru v jedné třídě k jinému vzoru ve třídě jiné. Ukazuje, že tímto způsobem je možné vytvořit vzor, který pro člověka vypadá, jako že patří do třídy první, ale model ho klasifikuje jako třídu druhou. Tuto základní myšlenku student následně dále rozvíjí a vylepšuje. Navržená metoda je potom otestována v páté kapitole, kde je vyhodnocována na několika modelech a datasetech. Metoda samotná je zajímavá a umožňuje najít nepřátelské vzory nejen pro klasicky trénované modely, ale také pro modely trénované tak, aby byly robustní vůči nepřátelským vzorům. Popis metody je kvalitní, nechybí žádné podstatné detaily. Experimenty jsou také provedeny velmi pěkně a jsou dobře vyhodnoceny.

Celkově se jedná o velmi kvalitní práci, jejíž jedinou slabinou je slabší textová část, hlavně v první polovině. Samotné navržené metody jsou zajímavé a rozhodně přináší nové poznatky do oblasti vytváření nepřátelských vzorů v hlubokém učení. Student prokázal, že je schopný kvalitní práce a podařilo se mu splnit vytyčené cíle. Mám pouze několik otázek k obhajobě.

1. Jak dlouho trvá trénování používaných generativních modelů a jak dlouho potom trvá nalezení nepřátelského vzoru?
2. Co přesně by bylo potřeba udělat, aby popsaná metoda byla aplikovatelná na větší obrázky, např. z ImageNetu?

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 9. června 2021

Podpis: