POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	**Few-Shot Learning of a Deepfake Detector**
Jméno autora:	**Bc. Vojtěch Brejtr**
Typ práce:	diplomová
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Katedra kybernetiky
Oponent práce:	Ing. Jan Čech, Ph.D.
Pracoviště oponenta práce:	Katedra kybernetiky, VRG

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání **průměrně náročné**

Hodnocení náročnosti zadání závěrečné práce.

Zadání hodnotím jako průměrně náročné. Deep fake detektor je otevřený problém, ale byly použity poměrně standardní techniky.

Splnění zadání **splněno**

Posuďte, zda předložená závěrečná práce splňuje zadání. V komentáři případně uveďte body zadání, které nebyly zcela splněny, nebo zda je práce oproti zadání rozšířena. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

Zadání splněno.

Zvolený postup řešení **správný**

Posuďte, zda student zvolil správný postup nebo metody řešení.

Postup řešení hodnotím jako správný. Využití embeddingu pomocí velkých modelů a nad ním natrénovaný jednoduchý klasifikátor je elegantní řešení problému s málo daty. Experimentální testování je extensivní.

Odborná úroveň **A - výborně**

Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.

Odborná úroveň práce je adekvátní.

Formální a jazyková úroveň, rozsah práce **B - velmi dobře**

Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.

Práce je psaná velmi dobrou angličtinou, všiml jsem si jen nepatrného počtu jazykových chyb. Typografická úprava je bez připomínek. Jedinou výhradu bych měl, že popis souvisejících metod je občas poněkud povrchní a málo informativní.

Výběr zdrojů, korektnost citací **A - výborně**

Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení závěrečné práce. Charakterizujte výběr pramenů. Posuďte, zda student využil všechny relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Citace korektní.

Další komentáře a hodnocení

Vyjádřete se k úrovni dosažených hlavních výsledků závěrečné práce, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, publikačním výstupům, experimentální zručnosti apod.

Práce obsahuje několik zajímavých výsledků. Závěry jsou příznivé a potvrzují, že jednoduchý detektor naučený nad embeddingem z modelu FaRL (pro obličeje) dosahuje podobných výsledků jako specializované modely na detekci deepfake obsahu. Velmi zajímavý je experiment s umělými influencerkami, které je možné s vysokou přesností rozpoznat od reálných.

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Uveďte případné otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.

Elegantní metoda, mnoho pečlivě provedených experimentů, příznivé výsledky.

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **A - výborně.**

Otázky:

1. V poslední době se objevuje spousta fotorealistických generativních modelů. Nejdůležitější a zároveň nejtěžší je schopnost detektoru generalizovat na obrázky vygenerované na novými modely/způsoby, které nebyly v trénovací sadě. Práce obsahuje leave-one-out experiment pro vlastní model a také srovnání konkurenčních metod (ale na plných datech, kde se generalizace netestuje). Je možné, že by detektor naučený tímto způsobem, tzn. tenký klasifikátor nad embeddingem generalizoval lépe?

2. Vysoká přesnost detektoru syntetických influencerek může mít několik vysvětlení. Například v datech je nějaký bias, kde jsou například a) syntetické ženy zaostřenější, nebo s vyhlazenější texturou, jinak nasvícené. Nebo b) obrázky jsou generované velmi podobnou technikou (stejnými generátory, stejný upsacaling nebo postprocessing). Anebo třeba více influencerek je vygenerováno stejnými autory a některá je v trénovací a další v testovací sadě. Proč by to mohlo být? Viditelné shluky na Fig. 7.9 mezi reálnými a syntetickým emebeddingy naznačuje spíš možnost (a).

3. Co se týká MRI skenů, příprava falešného obsahu je vlastně jen kopírování ze stejného snímku. Tyto tzv. copy-move edity se požívají k retušování obrázků a existovaly metody jejich detekce ještě před deep learning. To není příliš realistický scénář. Realističtější by bylo případný nádor včlenit do snímku zdravé osoby, což by mohlo třeba simulovat pokus o pojišťovací podvod. Bylo by to těžší detekovat?

Datum: 31.5.2024 Podpis: