ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	**Deep Learning Based Malware Detection**
Jméno autora:	**Bc. Vít Zlámal**
Typ práce:	diplomová
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Katedra počítačů
Oponent práce:	Ing. Martin Svatoš
Pracoviště oponenta práce:	Katedra počítačů

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání **průměrně náročné**

Hodnocení náročnosti zadání závěrečné práce.

Ke splnění zadání bylo potřeba nastudovat část literatury strojového učení se zaměřením na *imbalaced* dataset.

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.

Všechny body zadání byly splněny.

Zvolený postup řešení **správný**

Posuďte, zda student zvolil správný postup nebo metody řešení.

Autorův přístup k řešení problému je správný.

Odborná úroveň **B - velmi dobře**

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

Sekce 5.3 popisuje hyperparametry (# kernelů, velikost kernelů…) představovaného modelu. Velikost kernelů je zdůvodněna jednou větou bez popisu metodologie (a výstupů příslušných experimentů), které k tomuto rozhodnutí vedly. Rozhodnutí pro počet kernelů je podloženo výsledky v sekci 8.2.1, respektive tabulce 8.3, která obsahuje metriky na testovacím datasetu (validační není použit). Není zřejmé jestli jsou výsledky v tabulce 8.3 výstupem z pouze jednoho běhu nebo vícero (jak by mělo být v případě experimentů s nedeterministickým algoritmem).

Formální a jazyková úroveň, rozsah práce **A - výborně**

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

Rozsah práce je odpovídající. Práce je psaná čtivou angličtinou s pouze několika málo překlepy (např. *thread* namísto *treat*, *claud* namísto *cloud*) k nimž patří několikrát opakovaná chyba ve slovosledu (např. „*In [35] are researchers testing"*). Zápisu ztrátové funkce (např. sekce 2.1, strana 4) má být zřejmě y_i pokračováním dolního indexu.

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 jsou použity správně. V kapitole 3, která popisuje neuronové sítě, bych uvítal citování článků o batch normalizaci, RNN a CNN i když se z dnešního pohledu mohou zdát býti obecnou znalostí.

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.
V některých částech práce je odkazování na jiné části práce nebo se míchá více věcí dohromady, což může čtenářovu orientaci snížit. Například v druhé kapitole (obecný popis strojového učení) se píše která hypotéza bude testována v experimentech. V sekci 5.2 (popis architektury neuronové sítě) je napsáno „*We did not use batch normalization because our convergence during epochs is fast enough...*“, což pravděpodobně vychází z výsledků v kapitole 8 (kapitola experimentů). Sekce 5.3 odkazuje na výsledky experimentů, které jsou až v kapitole 8. V sekci 7.2.3 (kapitola o implementaci) se píše zvolená ztrátová (učící) funkce. Některé termíny by bylo vhodné podložit daty, které jsou na řešení kladeny, například „*relatively hard to train*“, či „*fast enough*“. V kapitole 7 se píše o implementaci v produkčním prostředí, která ale není přiložená a nemohl jsem jí tedy zhodnotit.

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.

Z práce je cítit, že autor problematice rozumí a během řešení postupoval správně, jelikož našel odpovídající metody v literatuře a upravil je pro svoji potřebu. Celkově tak práce působí dobrým dojmem až na drobné nedostatky popsané výše. Studentovi navrhuji položit následující otázky:

- V kapitole 7 je kladen důraz na fungující klasifikátor uvnitř produkčního prostředí (které používá Javu a Scalu). Zajímalo by proč byl upřednostněn MXNetu s učením v pythonu a inferencí v Javě před nějakým frameworkem, který obě fáze zvládne v Javě, např DL4J.
- Jak byla časově náročná implementace infrastruktury (Java, Scala, kap. 7) v poměru učící části naimplementované v Pythonu? (Popřípadě seznámení se s produkčním prostředím.)
- Jaká byla metodologie hledání hodnot v sekci 8.1. Bylo zkoušeno i více hodnot než jen 100 a 120?
- Můžete dát příklad škodlivého patternu v URL adrese? Jsou to statické patterny, nebo kontextově závislé na zbytku URL adresy?
- Jaké jsou nároky na finální klasifikátor: Jak rychlá musí být inference? Jak rychlé musí být učení jednoho modelu (např. na datech za měsíc) a jak často by bylo vhodno klasifikátor přetrénovat?

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **B - velmi dobře.**

Datum: 12.6.2020 Podpis: