ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

POSUDEK VEDOUCÍHO ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	**Anomaly Detection Methods for Log Files**
Jméno autora:	**Martin Koryťák**
Typ práce:	diplomová
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Katedra počítačů
Vedoucí práce:	Ing. Jan Drchal, PhD.
Pracoviště vedoucího práce:	Katedra počítačů

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání **náročnější**

Hodnocení náročnosti zadání závěrečné práce.

Práce se soustředí na aplikace neuronových autoenkodérů pro detekci anomálií nad sekvenčními daty logů. V návaznosti na předchozí práce a v souladu se současným odklonem od tradičních rekurentních architektur, jsme se rozhodli zaměřit na architektury založené na konvolučních vrstvách a self-attention přístupu. To od studenta vyžadovalo nastudování a dobré pochopení nových a netriviálních metod strojového učení.

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 bylo bez výhrad splněno. Rozsah práce je nadstandardní.

Aktivita a samostatnost při zpracování práce **A - výborně**

Posuďte, zda byl student během řešení aktivní, zda dodržoval dohodnuté termíny, jestli své řešení průběžně konzultoval a zda byl na konzultace dostatečně připraven. Posuďte schopnost studenta samostatné tvůrčí práce.

Student byl po celou dobu mimořádně aktivní, na konzultace byl vždy perfektně připraven. Rád bych ocenil jeho samostatnost při vyhledávání relevantních zdrojů i samotném řešení problémů (zejména návrhů různých architektur neuronových sítí). Student si rovněž dobře plánoval jednotlivé kroky a tak byla diplomová práce ve finálním stavu se značným předstihem, před cílovým datem odevzdání.

Odborná úroveň **A - výborně**

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

Práce je z odborného hlediska na vysoké úrovni. Velmi oceňuji velké množství architektur a přístupů, které student zvládl navrhnout a experimentálně ověřit. Všechny kroky návrhu i analýzy experimentů jsou vždy pečlivě motivovány. Velmi oceňuji přístup, který student uplatnil pro ladění meta-parametrů – v případě neuronových přístupů se, jak bývá zvykem, neomezil pouze na ladění parametrů optimalizační metody, ale hledal i detailně parametrizovanou optimální architekturu sítí.

Jako nedostatečné se mohou jevit experimenty na jediné datové sadě HDFS. Faktem však je, že se jedná o jedinou vhodnou dostupnou sadu – ostatní datasety buď postrádají anotace, které jsou potřeba pro evaluaci metod, případně jsou pouze triviálně anotovány na úrovni jednotlivých řádek souborů.

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

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

Práce je psána výbornou angličtinou. Text je dobře strukturovaný, čtivý a dobře srozumitelný. Velmi pěkná teoretická část podrobně popisuje moderní metody pro zpracování sekvencí, temporální autoenkodéry, související typy vrstev neuronových sítí a metody detekce anomálií se zaměřením na detekci anomálií nad logy. Práce dále obsahuje návrhovou a implementační část a nosnou experimentální kapitolu, zakončenou podrobnou diskusí a shrnutím všech dosažených

výsledků. Text práce je z typografického hlediska na vysoké úrovni, navíc je doplněn řadou přehledných, originálních ilustrací.

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.

Zdroje jsou citovány správně. Rozsah citací je nadprůměrný.

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.

Výsledky dosažené studentem považuji za mimořádně přínosné a plánujeme na nich založit následnou publikaci.

Nejlepší metoda založená na číselných reprezentacích (embeddings) se kvalitativně vyrovná starším typům metod, pracujících s daty předzpracovanými parserem, přináší však velikou výhodu v možnostech zpracování logů, které se nevyskytly v trénovacích datech. Není také nutné pracně ladit parametry parseru.

Po domluvě se mnou, se student zaměřil na specifický typ semi-supervizovaného přístupu, kdy je příslušný autoenkodér trénován nad většími daty normálního provozu (s anomáliemi odfiltrovanými s použitím dostupných anotací) a práh rozhodovací funkce pro detekci anomálií následně kalibrován na menší, plně anotované části datové sady (v textu označena jako validační množina). Postup trénování těchto metod bývá v praxi, vzhledem k typicky nízkému zastoupení anomálií, relaxován: anomálie nejsou z dat normálního provozu odstraňovány, což signifikantně snižuje nároky na anotace. Experimenty s tímto přístupem plánujeme navázat v rámci dalšího výzkumu.

III. CELKOVÉ HODNOCENÍ A NÁVRH KLASIFIKACE

Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení.

Předloženou závěrečnou práci považuji za nadstandardní a hodnotím ji klasifikačním stupněm **A - výborně.**

Datum: 3.6.2021 Podpis: