Review of Bachelor's Thesis

Student: Sarvaš Marek
Title: Interpretability of Neural Networks in Speech Processing (id 24073)
Reviewer: Mošner Ladislav, Ing., DCGM FIT BUT

1. Assignment complexity — **more demanding assignment**

Zadání považuji za obtížnější, neboť zpracování bakalářské práce vyžaduje kombinaci nastudování netriviální literatury a prokázání programátorských schopností při implementaci technik ze článků za použití moderní knihovny pro strojové učení.

2. Completeness of assignment requirements — **assignment fulfilled**

Student nejen splnil zadání, ale dokonce rozšířil 3. bod zadání. Bod 3 žádá aplikaci interpretační metody na nějakou úlohu zpracování řeči s cílem replikování výsledků. Nad rámec replikace bylo využito LRP metody pro modifikaci spektrogramů aktuální datové sady a přetrénování modelu na nově získaných datech. Modifikace spočívala v nulování časově-frekvenčních "binů", které byly vyhodnoceny pomocí LRP jako nejrelevantnější pro rozhodování sítě.

3. Length of technical report — **in usual extent**

Text práce je bohatý na informace.

4. Presentation level of technical report — **92 p. (A)**

Text technické zprávy je vhodně strukturovaný do kapitol a podkapitol. Text začíná teoretickými východisky, tj. popisem neuronových sítí společně s jejich učením a uvedením do tématu interpretace neuronových sítí (kde se dozvídáme o použité metodě LRP, ale i alternativách). Následuje popis použitých datových sad a architektur neuronových sítí. Přes kapitolu týkající se návrhu text směřuje k experimentální části. Rozsahy jednotlivých kapitol jsou v pořádku, navazují na sebe a text se pěkně čte. Teoretická část se zabývá netriviální problematikou, což má za důsledek občasné nepřesnosti (například v popisu učení s a bez učitele nebo LRP-0). Některé rovnice - např. (2.7), (2.9) - obsahují chyby, postrádají závorky (3.19), nebo by zasloužily popis symbolů (3.6), (3.17). U obrázků příznaků (filterbank) se objevují nesprávné popisky osy y.

5. Formal aspects of technical report — **92 p. (A)**

Text technické zprávy je v anglickém jazyce, který je na dobré úrovni. Gramatické chyby a překlepy se objevují zřídka. Typografická stránka je rovněž v pořádku až na drobné nedostatky. Často chybí mezera mezi slovem a levou závorkou či slovem a citací. Obrázek 4.4 není odkazován z textu. Občas se objevují problémy při odkazování rovnic (např. vynechání závorek).

6. Literature usage — **95 p. (A)**

Technickou zprávu doplňuje rozsáhlý seznam souvisejících literárních zdrojů, což poukazuje na obsáhlé studium tématu. Mezi zdroji jsou známé knihy o strojovém učení a neuronových sítích, články z konferencí (např. Interspeech), vědeckých časopisů (např. CSL) a archivu arXiv. Práce (mimo jiné) replikuje článek [4] využívající metodu LRP [20]. Oba jsou svázány se skupinou autorů Binder, Samek, Lapuschkin a kol. Student vhodně čerpal i z dalších článků těchto autorů. Občas se vyskytují citace sekundárních zdrojů (např. u Occlusion Analysis, SmoothGrad nebo u jádra LRP). V některých případech by mohly být citace doplněny (např. pro AlexNet). Vlastní výsledky jsou odlišeny od převzatých. U některých citací chybí bibliografické informace (např. [19], [32]).

7. Implementation results — **90 p. (A)**

Programová část práce je na dobré úrovni. Implementace využívá populární knihovnu PyTorch a je schopna běhu na GPU. Kód je dobře rozdělen do tříd, funkcí a souborů. Obsahuje pochopitelné a vhodně volené komentáře. Součástí přiloženého média je i návod zahrnující přímočarou přípravu Anaconda prostředí. Po přípravě běží kód bez problémů. Kromě implementace LRP, kódu spojeného s modely a datovou sadou obsahuje řešení i "demo skripty" (např. pro vykreslování map relevancí pomocí LRP).

8. Utilizability of results

Práce vychází z publikované literatury. Část experimentů je věnována replikování článku [4] zaměřeného na interpretaci AlexNet sítě pro klasifikaci pohlaví ze spektrogramů. Student došel k podobným závěrům jako autoři [4], což hodnotím kladně. Získané výsledky jsou rozšířeny dvěma směry. 1) Student využil LRP výstupů pro zvýšení robustnosti AlexNet modelu skrze přetrénování. 2) Byly provedeny podobné interpretační experimenty i s větší a robustnější ResNet sítí pro klasifikaci mluvčích z LFBE příznaků. Interpretovatelnost neuronových sítí je aktuálně velmi zajímavé a řešené téma. Proto získané výsledky a jejich rozšíření mají potenciál zaujmout. I

přesto, že již existují implementace LRP, jak student zmiňuje v práci, vlastní řešení může být zajímavé i pro širší komunitu tím, že využívá PyTorch (například oproti LRP Toolbox, který využívá samotný Python nebo Caffe).

9. Questions for defence

- Při interpretaci sítě klasifikující mluvčí jsou v Obrázcích 6.10, 6.12 a 6.14 uvedené i hodnoty relevancí. Jejich dynamický rozsah se diametrálně liší při použití různých LRP pravidel. Komentujte tento rozdíl. Jak se dynamický rozsah liší od rozsahů relevancí z experimentů týkajících se klasifikace pohlaví?
- V textu se několikrát vyskytuje tzv. "Clever Hans predictor". Vysvětlete, co pojem znamená a odkud se vzal právě tento název.

10. Total assessment

95 p. excellent (A)

Vysvětlení rozhodování neuronových sítí není snadná úloha i kvůli omezené možnosti vyhodnocení, potvrzení věrohodnosti a lidské interpretovatelnosti. S těmito nesnázemi se musel potýkat i student. Oceňuji proto odvahu, množství nastudované literatury a kvalitu odvedené práce (jak programové, tak textové). I přes (v některých případech) ne snadno interpretovatelné výsledky považuji práci za zdařilou a hodnotím kladně jako celek.

In Brno 3 June 2021

Mošner Ladislav, Ing.
reviewer