Posudek oponenta diplomové práce

Student: Zubalík Petr, Bc.
Téma: Odhad obličeje z řečového signálu (id 24862)
Oponent: Mošner Ladislav, Ing., UPGM FIT VUT

1. Náročnost zadání **obtížnější zadání**

Zadání považuji za obtížnější, neboť k jejímu řešení je třeba porozumět několika konceptům. Mezi ně se řadí neuronové sítě a jejich trénování, detailní porozumění knihovně využité pro trénování (model i výpočet objektivní funkce jsou totiž netriviální), efektivní využití výpočetních zdrojů. Složitost modelu s sebou nese značné požadavky na čas k natrénování, a tedy i dobré plánování experimentů je nutností. Komplikaci skýtá zadání i ve vyhodnocení, jelikož není snadné výstupy vyhodnotit kvantitativně. Tím pádem nemusí být přímočaré určit, jestli se experimenty ubírají správným směrem.

2. Splnění požadavků zadání **zadání splněno**

3. Rozsah technické zprávy **je v obvyklém rozmezí**

Text je bohatý na informace.

4. Prezentační úroveň předložené práce **72 b. (C)**

Text zprávy psané v českém jazyce je srozumitelný. Velice často ovšem sklouzává k neformálnosti, nepřesnosti a nekonzistenci v notaci, což je slabinou práce. Z velké části je text členěn logicky do kapitol a sekcí. Domnívám se, že některé části teorie by mohly být upozaděny/odstraněny. Např. nepřesný popis MFCC je nadbytečný, neboť použité modely tento typ koeficientů nepoužívají. Obdobně detekce hran, segmentace či filtry pro rozmazání/zaostření jsou nadbytečné. Ušetřené místo by mohlo být věnováno konkrétnějšímu popisu experimentů - kapitole jimi se zabývající by prospělo lepší členění a hodnocení jednotlivých (i neúspěšných) kroků (např. pomocí metrik využívaných ve validaci). Popis současných přístupů je uveden v kapitole týkající se návrhu. Preferoval bych tuto část v teoretických východiscích. Na druhou stranu zvolená pozice umožňuje přímé porovnání navrženého systému s existujícími.

5. Formální úprava technické zprávy **80 b. (B)**

Z typografického hlediska je práce vesměs v pořádku. Jisté problémy se však vyskytují. Některé obrázky nejsou odkazovány z textu. Zejména v kapitole Experimenty se vyskytují chybné reference. Popisek tabulky se nachází pod ní. Práci by prospěly vektorové obrázky namísto rastrových, kde je to možné. Po jazykové stránce je text dobrý. Občas se vyskytují problémy s gramatikou, čárkami a překlepy. V případech, kdy existuje zažitý překlad do češtiny, je dobré jej využívat (např. datová sada místo použitého "datasetu").

6. Práce s literaturou **95 b. (A)**

Student prokázal dobré zorientování se v problematice a práci s literaturou při důsledném citování. Seznam citované literatury je značný a zahrnuje pro práci relevantní zdroje. Mnohé články byly publikovány na relevantních konferencích (např. Interspeech, CVPR) či v časopisech (např. IEEE Transactions). Několik článků bylo publikováno v rámci archivu arXiv. Podstatná část teorie vychází ze známé knihy Deep Learning od I. Goodfellowa. Mnoho zdrojů je jen pár let starých. Student se v některých oblastech seznámil se SOTA přístupy (např. UniSpeech-SAT). Velmi kladně hodnotím využití znalostí ze stěžejního článku pro tuto práci (Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation) a jejich přenesení do jiné domény.

7. Realizační výstup **95 b. (A)**

Student navrhl zajímavý model pro odhad obličeje z řečové nahrávky. I přes inspiraci v existující literatuře je architektura inovativní. Zejména díky využití předtrénovaného modelu StyleGAN a přístupu k odhadu "stylů" pro tento model z audia v časové doméně. Byly učiněny vhodné kroky k ověření, že je model schopen konvergence. S ohledem na úlohu a komplikovanou objektivní funkci (která zahrnuje propagaci výstupů dalšími předtrénovanými neuronovými sítěmi) není snadné model natrénovat. I přesto se studentovi podařilo iterativně dospět k výsledkům, které lze považovat vzhledem k omezenému času za uspokojivé. Vyhodnocení je spíše strohé, což ale odpovídá tomu, že kvantitativní posouzení není pro daný úkol přímočaré/jednoznačné. Student projevil schopnost přebírat relevantní existující řešení a úspěšně je integrovat (např. části kódů, které byly převzaté z repozitářů stěžejních článků). Kladně hodnotím iniciativu při navržení vlastní objektivní funkce "Style loss", jejíž další zkoumání považuji za zajímavé. Celkově student odvedl spoustu dobré práce.

8. Využitelnost výsledků

Přesto, že konceptem audio kodér/obrázkový dekodér připomíná architektura existující přístupy (zejména

Speech2Face), využitím StyleGAN generátoru a kodéru založeném na ResNet je práce inovativní. Pokud by se podařilo nalézt přístup k trénování, který by vedl na dobře generalizující model s kvalitou výstupů podobnou Speech2Face, bylo by možné uvažovat o publikaci na některé z konferencí.

9. Otázky k obhajobě

- Z kodéru, jenž je založen na architektuře ResNet, jsou extrahovány 3 vnitřní reprezentace využívané pro tvorbu stylů (Obrázek 3.4). Komentujte důvod a důsledky využití sčítání vnitřní reprezentace s nadvzorkovanou reprezentací na vyšší úrovni.
- Normalizace ("frontalizace") obrázků obličejů trvala dle zprávy 1500 hodin (62,5 dne). Využil jste nějaký přístup k paralelizaci výpočtu? Pakliže výpočet neprobíhal sériově, jak dlouho reálně trval?
- Zkuste se zamyslet nad možností předtrénování kodéru pouze pomocí "Style loss". Tento přístup by se koncepčně podobal Speech2Face. Jaké jsou potenciální benefity a problémy spojené s tímto předtrénováním?

10. Souhrnné hodnocení

90 b. výborně (A)

I přesto, že dosažené výsledky nemají nutně kýženou kvalitu, domnívám se, že jsou výstupy vzhledem k obtížnosti úkolu jako takového velmi uspokojivé. Student udělal spoustu práce, provedl zajímavý návrh, pro který sám vybral StyleGAN, implementoval části systému a práci dovedl do rozumného cíle. Slabiny technického textu v kontextu celé práce proto upozaďuji.

Prohlášení: Uděluji VUT v Brně souhlas ke zveřejnění tohoto posudku v listinné i elektronické formě.

V Brně dne: 2. června 2022

Mošner Ladislav, Ing.
oponent