Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Adam Lechovský		
Název práce	Restaurování a vylepšování technické kvality zvukových nahrávek metodami strojového učení		
Rok odevzdání	2022		
Studijní program	Informatika	**Studijní obor**	Umělá inteligence

Autor posudku	Ondřej Dušek	**Role**	oponent
Pracoviště	Ústav formální a aplikované lingvistiky		

Text posudku:

Shrnutí obsahu Ve své diplomové práci se Adam Lechovský zabývá zlepšením kvality audionahrávek pomocí neuronových sítí. Navrhuje sedm různých architektur neuronových sítí, které pracují s celkem 10 sekundami surového audia (waveformu). Předpokládají vzorkovací frekvenci 44,1 kHz, tj. počítají se 441 tisíci vstupů a výstupů. Jedná se o čtyři konvoluční sítě (dvě velikosti standardních konvolucí, jedna varianta s dilatovanými konvolucemi a jedna síť typu ResNet), dva typy rekurentních sítí (LSTM, GRU) a jednu kombinaci architektur (konvoluční síť jako předzpracování pro LSTM).

Autor tyto sítě aplikuje na tři různé synteticky vytvořené distorze audia: bílý šum, snížení dynamického rozsahu a redukci vzorkovací frekvence (každý druhý vzorek je vynulován). Všechny architektury jsou tedy trénovány ve třech variantách. Pro trénování se používá kombinace dvou velkých volně dostupných datových sad: Free Music Archive (hudba) a Mozilla Common Voice (hlas). Autor u obou sad používá jen podmnožinu, protože dat je pro jeho potřeby a na jeho výpočetní možnosti příliš mnoho.

Všechny architektury jsou nejprve vyhodnoceny automaticky na testovacích sekcích obou datových sat, se stejnými syntetickými distorzemi jako při trénování. Zde autor používá velikost kvadratické ztrátové funkce (MSE loss) a detaily několika vybraných spektrogramů. Dále autor všechny architektury evaluuje subjektivně na základě několika reálných vzorků zašuměného audia nahraných na vlastní mikrofon. Ty obsahují jeho vlastní hlas a reprodukovanou hudbu z testovací sady. Zde autor postupuje čistě kvalitativně a porovnává vlastní dojmy z poslechu výstupů jednotlivých sítí. Práce obsahuje i několik předem stanovených hypotéz o relativní výkonnosti jednotlivých architektur, které autor posléze porovnává s výsledky.

Z experimentu vychází jako nejlepší architektury ResNet, LSTM a GRU, ale žádná ze sítí není prakticky použitelná. Sítě se sice do určité míry naučí redukovat umělé distorze, ale zároveň je zejména redukce šumu příliš silná; na reálných nahrávkách (bez syntetické distorze) dochází spíše ke zhoršení kvality.

Text práce sestává z úvodu se shrnutím cílů práce, závěru s přehledem výsledků a možných vylepšení, a čtyř číslovaných kapitol. První dvě kapitoly jsou teoretické: Kapitola 1 představuje základy reprezentace audia v neuronových sítích a některé metriky pro měření kvality výstupu. Kapitola 2 obsahuje seznam několika neuronových architektur (plně propojené sítě, konvoluční i rekurentní sítě) včetně detailních popisů jednotlivých neuronových vrstev, letmo je zmíněna (ne)vhodnost architektur pro použití k práci se zvukem. Dále jsou uvedeny základní regularizace neuronových sítí, trénovací ztrátové funkce a představeno je několik volně dostupných zvukových datových sad. Druhé dvě kapitoly pak popisují samotné experimenty provedené v práci. Kapitola 3 vysvětluje celkové nastavení experimentu, detaily použitých architektur, přípravu dat i postup evaluace. Stanovuje pak několik hypotéz o předpokládaných výsledcích. Kapitola 4 je přehledem všech výstupů práce. Shrnuje postup trénování a výpočetní náročnost architektur i jejich výkonnost (pomocí automatického i subjektivního vyhodnocení). Obsahuje i zhodnocení platnosti předem stanovených hypotéz.

Celkové hodnocení Práce v zásadě splňuje všechny formální náležitosti kladené na diplomovou práci, ale zároveň obsahově vykazuje celkem zásadní nedostatky, které v podstatě znemožňují použití výsledků práce v praxi nebo k dalšímu výzkumu.

Samotný text je zpracován poměrně kvalitně, práce je psaná dobře srozumitelnou angličtinou s malým množstvím chyb, které navíc nejsou zásadní a nebrání porozumění. Typy neuronových architektur se pro tento problém zdají zvolené rozumně, postup stanovení hypotéz a jejich ověřování (dokonce za pomoci statistických testů) si zaslouží

uznání. Základní struktura experimentů je také zvolena vhodně a samotné experimenty neobsahují nic, co by reportované hodnoty vyloženě zneplatňovalo.

Jsou tu ale tři zásadní nedostatky, které výrazně snižují relevanci a použitelnost výsledků práce. Zde je uvádím jen stručně, detaily následují níže:

1. Práci naprosto schází návaznost na relevantní související literaturu – ač se zdaleka nejedná o první experiment svého druhu, žádná literatura o odšumění nebo modifikaci audia pomocí neuronových sítí není citována ani zohledněna.

2. Na výsledcích se pravděpodobně negativně podepsalo dost nestandardní provedení experimentů bez jakéhokoliv ladění či zohlednění konvergence trénování jednotlivých architektur. Kromě toho i samotná příprava dat zřejmě není dostatečná pro realistické nasazení výsledných modelů.

3. Použité evaluační metody jsou celkem omezené, evaluace by zasloužila rozšíření.

Celkově se práce příliš soustředí na samotné neuronové architektury a zanedbává ostatní důležité aspekty experimentů se strojovým učením. Práci tedy doporučuji k obhajobě jen s velkými výhradami.

Připomínky k návaznosti na literaturu Práce úplně opomíjí výzkum v oblasti redukce šumu – jak text teoretických kapitol, tak v nich citované práce se zabývají jen buď základní teorií neuronových sítí, nebo základy zpracování audia. V podstatě jediná explicitně citovaná práce o neuronovém zpracování zvuku je síť WaveNet, ta byla ale vyvinuta k syntéze řeči. Jiné práce o neuronovém zpracování audia nebo odrušení šumu (ať už neuronovém nebo ne) citovány nejsou.

Zohledněna není ani zjevně relevantní literatura uvedená v oficiálním zadání diplomové práce, navíc jednoduché hledání „audio denoising neural networks“ na Google Scholaru vrací hned několik na první pohled relevantních výsledků. Důsledkem toho je, že architektury použité v práci neodpovídají současnému stavu výzkumu a není možné je snadno porovnat – dnešní systémy (např. VoiceFixer, N-HANS) mají typicky komplexnější strukturu, ač konceptuálně založenou na architektuře ResNet.

Popis teorie neuronových sítí ani nebylo v práci třeba probírat do takového detailu, jde o celkem standardní a základní architektury. Část z nich navíc neni ani v práci využita nebo vztažena k dalšímu obsahu. V sekcích popisujících základy zpracování audia by se naopak hodilo víc vysvětlování a víc dokladů tvrzení citacemi, ale proti opomenutí neuronových systému se jedná o detail.

Připomínky k experimentům U experimentů nastává první problém už s přípravou trénovacích dat. Typy šumů syntetizované pro trénování modelů jsou poměrně jednoduché a není pak divu, že si natrénované modely neporadí s realistickými nahrávkami. Mnohem výhodnější by bylo k datům přičítat reálný šum, např. vybraný z volně dostupné sady AudioSet. Obecně je pro dobrou generalizaci neuronových sítí potřeba větší variabilita dat. Toto by ale ještě nebyl tak velký problém – řekněme, že to jen danou úlohu dělá poněkud méně realistickou.

Mnohem větší problém je celkový postup experimentální práce. Podle popisu vypadá takto: autor implementoval neuronové architektury a ty pak bez jakéhokoliv nastavování hyperparametrů nechal trénovat po předem stanovenou dobu, bez ohledu na trénovací křivky a rychlost konvergence, a pro evaluaci používá jednoduše poslední checkpoint. Tohle rozhodně není standardní postup při práci s neuronovými sítěmi.

Hyperparametry mohou mít naprosto zásadní vliv na výkon, proto je třeba je nějakým způsobem vyladit nebo aspoň základně ověřit, že na nich výkon příliš nezávisí. Vzhledem k výpočetní náročnosti není nutné k tomu použít celou trénovací sadu (stačí netriviální podmnožina), ale tento krok nelze zcela přeskočit. Určitě to platí pro parametry optimalizátoru, které práce komentuje a kde se možná dá do jisté míry spoléhat na defaultní hodnoty z Tensorflow, ač to není ideální. Mnohem víc to ale platí pro parametry sítě samotné (počet a šířka vrstev), které zřejmě vůbec žádným laděním neprošly a hodnoty jsou nastaveny podle pocitu autora. Nevhodná volba těchto parametrů může celý experiment dost poškodit.

Při trénování neuronových sítí se zpravidla zohledňuje konvergence modelu a finální checkpoint se vybírá podle ztrátové funkce na vývojové datové sadě. Stanovit čas trénování předem, navíc pro všechny architektury stejně, mi absolutně nedává smysl.

S tím souvisí i problémy při srovnávání jednotlivých architektur. Srovnávat je při shodném času trénování a ne při shodném počtu zpracovaných trénovacích příkladu taktéž není standardní postup. Čas trénování je určitě důležité kritérium, ke kterému se přihlíží, ale výkon modelů se standardně srovnává na shodné trénovací sadě.

Z toho vyplývá, že porovnání architektur a vyhodnocení hypotéz autorem stanovených a rozebíraných v sekcích 3.11 a 4.2.2 nemají velkou platnost.

Připomínky k evaluaci Nedostatky v evaluaci zřejmě souvisí s chybějící návazností na literaturu. Z automatických metrik na testovací sadě mohla práce použít mnohem širší škálu než čistě chybovou funkci; minimálně použití poměru signál-šum (SNR) by se přímo nabízelo, protože ho autor používá na výběru vlastního realistického audia. Kromě toho existují další metriky: mel cepstral distortion, segmental SNR, short-time objective intelligibility atp.; výstupy neuronových sítí jde se syntetickým šumem snadno porovnávat s původní, nezašuměnou verzí audia. Práce taktéž mohla zohlednit nebo aspoň zmínit benchmark pro tuto úlohu – Deep Noise Suppression Challenge.

Podobné problémy jako u automatických metrik lze vyčíst i v subjektivním hodnocení. Zde jsme omezeni na autorovy dojmy, ale práce mohla použít standardního postupu lidského hodnocení audia s Mean Opinion Score (např. pomocí frameworku MUSHRA) a hodnocení tak kvantifikovat. Při tomto standardním hodnocení by navíc (oproti postupu použitém v práci) nemohlo dojít k ovlivnění hodnotitele vlastními předpojetími, protože jednotlivé systémy jsou pro účely hodnocení neoznačeny a zamíchány.

Dotazy Na základě čtení práce mám ještě několik dotazů, oproti výše uvedeným připomínkám jsou ale méně zásadní:

- Proč nejsou jednotlivé druhy šumu v práci kombinovány – nedosáhlo by se tak robustnějších modelů?
- Proč sítě pracují s tak dlouhým segmentem audia? Většina typů šumu nevykazuje závislosti v tak dlouhém rozsahu. Nebylo by praktičtější počítat např. s jednou sekundou audia a modelu dát na vstup i nějaký kontext (výstup pro předchozí sekundu audia)?
- K sekci 3.4.3: proč nebylo možné implementaci rozdělit na dva procesy a umožnit tak paralelní zpracování? Python přeci podporuje komunikaci mezi procesy.
- Na obr. 4.1 a 4.2 to vypadá, že při trénování v jednom místě pro některé modely výrazně vzroste ztrátová funkce – je pro toto nějaké vysvětlení?
- Proč není provedena automatická evaluace pomocí SNR na celé testovací sadě – dalo by se to porovnat s výsledky na výběru realistických dat (tabulky 4.5, 4.6)? Jaké jsou v tabulkách 4.5 a 4.6 jednotky?
- Jak byly zvoleny příklady pro obr. 4.4–4.9? Proč se jedná o různé architektury?

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 30. 8. 2022

Podpis: