Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Jan Vainer
Název práce	Efficient neural speech synthesis
Rok odevzdání	2020
Studijní program	Informatika **Studijní obor** Umělá inteligence

Autor posudku	Jan Hajič jr. **Role** oponent
Pracoviště	Ústav formální a aplikované lingvistiky

Text posudku:

Diplomová práce se zabývá zrychlením a zmenšením neuronových modelů pro syntézu mluvené řeči. Druhá kapitola je výtečný popis použitých modelů a technik, včetně poměrně detailního popisu toho, jak jsou aplikovány na TTS. Oceňuji též přehledný, široký a informativní popis příbuzných prací ve třetí kapitole. První tři kapitoly práce by se v podstatě daly publikovat jako přehledový článek o TTS. Vzhledem k tomu, že přínosy spočívají v použití metodologie teacher-student sítí, bych býval uvítal část 2.6 věnovanou alespoň stručně tomuto postupu.

Čtvrtá kapitola popisuje studentem navržený model. Ukazuje, že diplomant výtečně ovládá hluboké učení, včetně triků (např. v částech 4.3.2 či 4.3.4, použití zajímavějších ztrátových funkcí v části 4.4.3) jejichž aplikace vyžaduje hluboké porozumění tomu, jak učící proces probíhá. Pátá kapitola je také silná. Diplomant provedl zjevně velké množství experimentů, přesvědčivě diskutuje jejich výsledky a navrhuje zlepšení. Šestá kapitola, evaluace, je důsledná, diplomant věnoval výběru metod evaluace dostatečnou pozornost (např. metodologie MUSHRA pro odhad kvality vygenerovaného audia). Předkládaný systém je vyhodnocován proti state-of-the-art TTS systémům a výsledky jsou přesvědčivě lepší. Vyhodnocení rychlosti je také přesvědčivé a cílům práce adekvátní. Oceňuji však především sekci Future work (6.4), která je promyšlená a nejedná se pouze o výčet nestihnutých experimentů; vyzdvihl bych potřebnost studie srovnávající hlavní implementace STFT, nejen pro výzkum řeči, ale např. i v hudební informatice. Závěry práce (kapitola 7) dle mého soudu zcela opodstatněně prohlašuje cíle práce za splněné.

Bibliografie práce je rozsáhlá a aktuální (32 ze 79 citovaných materiálů není starší než tři roky). Práce je psána perfektní odbornou angličtinou, srozumitelně a logicky. Překlepy v textu (kap. 2, první odstavec: ,,teSt-to-speech", různě: syntesized vs. synthesised, homogeneOus) jsou minimální, srozumitelnost není nikde ohrožena. Především bych vyzdvihl, že text má konzistentně vysokou informační hustotu, žádné zbytečné věty.

Mám jednu kritickou připomínku: text by mohl jasněji formulovat klíčové nápady v návrhu

teacher-student modelu. Jedná se o malé zásahy, které v jinak bezproblémově srozumitelné práci zřetelněji vymezí nápady diplomanta vedoucí ke splnění cílů práce. Konkrétně: na začátku kapitoly 4 by mělo být jasně řečeno, proč je třeba predikovat trvání fonémů, což je to, na co se používá teacher network. Pokud práci dobře chápu, klíčové je, že jakmile máme pro každý foném počet frames ve výstupním spektrogramu, vstupní a výstupní sekvence začnou být stejně dlouhé a lze použít rychlé metody diskutované v sekci 3.2.3 (s. 27), avšak toto by mělo být zrekapitulováno (nejspíše v sekci 4.1) takovým způsobem, aby čtenář pochybnost o svém (ne)pochopení klíčových nápadů práce neměl. Hodilo by se v části 3.2.3 také lépe zavést termín autoregressive flow.

Podobně na začátku sekce 4.3 by bylo vhodné jednou větou říci, co se učitelská síť vlastně učí, tj. co je vstupem a výstupem modelu. Informaci využívanou posléze žákovskou sítí se totiž učitelská síť učí jaksi ,,mimochodem" v rámci mechanismu attention použitého uvnitř sítě. Příslušná informace v textu je (první odstavec sekce 4.3.3), nicméně by se četlo lépe, kdyby se onen odstavec přesunul hned před část 4.3.1. V kapitole 4 také není řečeno, *proč* se délky odvozují právě z alignmentu v sekvenční predikci framů spektrogramu, a jaké by případně mohly být alternativy (toto je diskutováno v části 6.4, bod Alternative alignment models na s. 60). K této diskusi ještě patří klíčový postřeh z části 5.2.6 – za jakých podmínek se učitelský model naučí potřebné informace o zarovnání fonémů, když je vlastně pro predikci výstupního spektrogramu nemusí potřebovat (obzvlášť když se používá teacher forcing). Všechny tyto informace v práci jsou, pouze si je čtenář musí seskládat při čtení kapitoly 4 sám; pomohlo by je usouvstažnit v popisu modelu explicitně.

Tato kritická připomínka však nijak nenarušuje celkový dojem z práce: jedná se o vynikající práci vylepšuje state of the art ve zvolené úloze, jednoznačně je třeba ji doporučit k obhajobě a zároveň doufám, že její výsledky budou také vydány formou recenzované publikace.

Dotazy k obhajobě:

Je pravda, že se v práci postup teacher-student používá jinak než v původní myšlence ,,destilace", tj. zmenšování modelu tím, že se student učí proti měkčím distribucím učitelského výstupu, spíše než aby se učil tvrdé kategorizaci trénovacích labelů?

Jaké jiné (potažmo rychlejší) způsoby, jak se naučit predikovat trvání fonémů, byste považoval za nejperspektivnější? Vidíte nějaký nadějný způsob, jak alignment mezi fonémy a ground-truth spektrogramem hledat přímočařeji?

Práci doporučuji k obhajobě.

Práci navrhuji na zvláštní ocenění.

Práce přináší přesvědčivě state-of-the-art výsledky na otevřeném a podstatném výzkumném problému. Navíc je metodologicky excelentní a replikovatelná (otevřené datasety, zdrojový kód publikován na githubu, transparentní evaluace), takže vedle výborného state-of-the-art výsledku by mohla sloužit jako vzorová práce pro diplomanty zabývající se strojovým učením.

V Praze dne 6. 6. 2020

Podpis: