Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Tomáš Nekvinda		
Název práce	Vícejazyčná syntéza řeči		
Rok odevzdání	2020		
Studijní program	Informatika	**Studijní obor**	Umělá inteligence

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

Text posudku:

Shrnutí obsahu Diplomová práce Tomáše Nekvindy se zabývá vícejazyčnou syntézou řeči (TTS), tj. modely založenými na neuronových sítích, které jsou schopny syntetizovat řeč z psaných textů ve více jazycích (jediný natrénovaný model podporuje více jazyků současně). Výzkum v této oblasti má několikerou motivaci:

- možnost natrénovat stabilní model syntézy řeči i v případě, že pro daný jazyk není dostatek trénovacích dat (použití dat z několika jazyků současně),
- možnost syntetizovat řeč shodným hlasem ve více jazycicích (např. pro konzistenci prezentace firmy v mezinárodním kontextu),
- syntézu řeči přepínající mezi jazyky v rámci jedné věty (code-switching) – to je důležité v případech, kdy se v jednojazyčném textu vyskytují místní názvy a osobní jména z cizích jazyků, např. u syntézy navigačních pokynů nebo čtení zpravodajství.

Neuronové modely v posledních letech výrazně zlepšily kvalitu výstupu TTS a v podstatě vůbec otevřely možnost trénování vícejazyčných modelů bez nutnosti použití nahrávek téhož bilingvního mluvčího. Oblast vícejazyčné syntézy však dosud nebyla velmi dobře prozkoumána, dosavadní vícejazyčné systémy používaly typicky max. 2-3 jazyky a vyžadovaly velké množství trénovacích dat ve všech jazycích.

Experimenty v této diplomové práci jsou založeny na state-of-the-art neuronovém modelu TTS Tacotron 2, který ale autor pro účely vícejazyčné syntézy velmi výrazně upravil. Tacotron 2 je založen na architektuře typu enkodér-dekodér, kde obě tyto komponenty pracují nad daty sekvenčně. Protože sekvenční střídání více enkodérů pro jednotlivé jazyky by bylo nepraktické a neefektivní, nahradil autor rekurentní sítě v enkodéru konvolučními, které dovolují paralelní zpracování. Největší inovací je použití meta-learningu – model má speciální komponentu, která se učí upravovat parametry enkodéru v závislosti na vstupním jazyce (tj. část parametrů neuronové sítě je generovaná jinou částí sítě). To umožňuje nalézt kompromis mezi shodnými parametry enkodéru pro všechny jazyky a zcela oddělenými enkodéry. Pro větší stabilitu trénování autor zavedl použití dávek dat zahrnujících všechny vstupní jazyky najednou. Aby model zvládl code-switching a dovedl udržet stejný hlas i při změně jazyka uprostřed věty, zahrnuje nová architektura i embeddingy řečníků (pro odlišení vlastností různých řečníků v trénovacích datech) a adversariální klasifikátor řečníka (model se trénuje tak, aby klasifikátor nepoznal, který řečník mluví).

Autor pro své experimenty připravil trénovací datové sady vhodné pro natrénování modelu TTS na 10 jazycích a pro code-switching v 5 jazycích (zde je potřeba více mluvčích pro každý jazyk, což nebylo ve všech

jazycích k dispozici). Použil k tomu existující sadu CSS10 (s jedním mluvčím pro 10 jazyků) a část dat z projektu Mozilla CommonVoice, přičemž oba zdroje částečně automaticky a částečně ručně profiltroval. Zejména u dat CommonVoice toto znamenalo velké množství práce, protože původní data jsou určena zejména k trénování systémů rozpoznávání řeči a obsahují velké množství šumu.

Svoji architekturu autor extenzivně vyhodnocuje ve srovnání s několika silnými baseline modely – originální architekturou Tacotron 2 (natrénovanou pro každý jazyk zvlášť) a dvěma variantami nově navrženého modelu, které neobsahují meta-learning – jedna používá separátní enkodéry pro každý jazyk, druhá jeden sdílený enkodér se shodnými parametry pro všechny jazyky.

V prvním experimentu se testuje samotná schopnost natrénování a stabilita při použití dat z více jazyků – zde jako hlavní metrika slouží chybovost na úrovni znaků (character error rate, CER), kde jsou výsledky syntézy dále zpracovány externím rozpoznávačem řeči a kontroluje se, zda modely jsou schopny správně vyslovit vstupní text. Jedná se o aproximaci omezenou kvalitou dostupného rozpoznávače, ale pro vyhodnocení modelů určitě postačuje. Evaluaci pomocí CER doplňuje i metrika mel cepstral distortion, která porovnává vygenerované spektrogramy s referenčními lidskými nahrávkami. Autorem navržený meta-learningový model na většině jazyků statisticky významně poráží všechny baseline modely.

Druhý experiment testuje schopnost modelů provádět code-switching (zde jednojazyčný Tacotron 2 použít nelze). Protože tato úloha je pro modely náročnější, je i evaluaci věnováno více pozornosti a používá se subjektivní porovnání modelů ze dvou hledisek – plynulost (stabilita hlasu napříč jazyky) a přesnost (správné vyslovení zadaného textu). Studie se účastnilo 50 informantů nalezených pomocí crowdsourcingu, po 10 rodilých mluvčích každého z 5 testovaných jazyků, přičemž všichni měli dobrou znalost alespoň jednoho z dalších jazyků. Autor jako testovací data pro tuto studii posbíral věty obsahující code-switching z Wikipedie. I zde se meta-learningový způsob ukázal jako výrazně lepší než dvě ostatní varianty modelu, což bylo potvrzeno i statistickými testy a manuální chybovou analýzou.

Text práce obsahuje úvod, 6 číslovaných kapitol a závěr. V úvodu autor stručně vysvětluje cíle práce a shrnuje obsah následujícího výzkumu. 1. kap. je teoretickým úvodem do zpracování signálu a menším slovníčkem pojmů z oblasti hlubokého učení. 2. kap. obsahuje jednak obecný úvod do metod TTS, jednak přehled nejdůležitějších moderních neuronových modelů pro TTS. 3. kap. shrnuje potřebné evaluační metriky používané v další práci. 4. kap. představuje dostupné datasety pro TTS a motivuje další autorovu práci s daty. V 5. kap. najdeme přehled dosavadní práce ve vícejazyčné syntéze řeči, tj. shrnutí dosavadních přístupů ke konkrétnímu řešenému problému. 6. kap. je zdaleka nejdelší a představuje všechnu experimentální práci autora – architekturu nového modelu a způsob trénování, přípravu nových datových sad, experimenty s vícejazyčným trénováním i code-switchingem, jejich evaluaci i diskusi výsledků. Závěrečná kapitola ještě krátce shrnuje dosažené výsledky a zároveň uvádí možnosti rozšíření práce do budoucna.

Průběh prací Autor na tomto tématu pracoval přibližně rok, z toho nejméně posledních 10 měsíců velmi intenzivně. Veškeré experimenty jsme spolu probírali na pravidelných schůzkách. Autor při práci projevoval velmi iniciativní přístup k problému – sám vyhledával a studoval relevantní literaturu a sám přicházel s nejdůležitějšími řešeními, které výrazně vylepšily výsledky práce. Výborně se sám a aktivně vypořádal také s problémem nedostatku trénovacích dat. I samotné psaní textu probíhalo v aktivní diskusi se mnou, celou práci jsme spolu před jejím

odevzdáním detailně probrali. Se spoluprací s autorem jsem nadmíru spokojený.

Hodnocení Po obsahové stránce má práce velmi výrazný vědecký přínos – posouvá současné poznání v oblasti vícejazyčné syntézy řeči. Výsledný model dovoluje trénovat vícejazyčné modely schopné code-switchingu z mnohem menšího množství dat, než bylo dříve zapotřebí, a dosahuje lepších výsledků než předchozí přístupy. Zadání práce bylo beze zbytku splněno; její celkový přínos rozhodně převyšuje nároky kladené na diplomovou práci. Proto jsme také s autorem zkrácenou verzi práce zaslali do recenzního řízení na prestižní mezinárodní konferenci Interspeech. Věřím, že se práce stane základem dalšího výzkumu v oboru, čemuž napomáhá zveřejnění veškerého zdrojového kódu, ukázek výstupů a vyčištěných trénovacích dat na GitHubu i živá demonstrace na platformě Google Colab. Jak jsem zaznamenal z komentářů na GitHubu projektu, kód už začínají používat další vývojáři.

Text práce je psán dobrou angličtinou. Uspořádání do kapitol je sice na první pohled trochu neobvyklé, ale má vnitřní logiku a výborně dovoluje oddělit přehledovou část od originálního přínosu autora. Veškeré moje komentáře jsme už s autorem prodiskutovali před odevzdáním práce, takže k textu nemám žádné výhrady.

Celkově práci velmi silně doporučuji k obhájení, žádné dotazy k obhajobě nemám.

Práci doporučuji k obhajobě.

Práci navrhuji na zvláštní ocenění.

Práce si jednoznačně zaslouží soutěžit o zvláštní ocenění. Jak jsem již napsal v hodnocení, její vědecký přínos je nezpochybnitelný – jedná se o výzkumnou práci, která snese srovnání ve světovém měřítku. Zatím jsem nevybral, do které konkrétní soutěže ji přihlásit, ale chystám se tomu věnovat po obhajobě.

V Praze dne 30. 6. 2020

Podpis: