Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Jan Vainer		
Název práce	Efektivní neuronová 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 Jana Vainera se zabývá návrhem, implementací a vyhodnocením systému pro výpočetně efektivní syntézu řeči (TTS) založeného na neuronových sítích. Konkrétně se zabývá úlohou samotného jádra TTS – syntézou spektrogramu z fonetického zápisu (fonetický přepis z textu se řeší výslovnostním slovníkem a jednoduchými pravidly, o vytvoření zvukových vln ze spektrogramu se stará externí vokodér).

Motivace pro zvýšení efektivity TTS je jednoznačná: Nové neuronové modely v posledních letech výrazně zlepšily kvalitu TTS a jsou schopny generovat skutečně velmi přirozenou řeč. Jejich nevýhodou však zůstává velká výpočetní náročnost, a to jak při trénování systému, tak při samotné inferenci (syntéze výstupů). Žádný dosavadní neuronový systém TTS neposkytoval současně rychlé trénování, rychlou syntézu (v reálném čase) a vysokou kvalitu syntetického hlasu.

Autor proto navrhl novou neuronovou architekturu, která vychází z několika existujících neuronových modelů TTS a přináší několik vlastních nových vylepšení či zjednodušení. Architektura sestává z dvou neuronových sítí:

- *učitelské sítě*, která se pomocí modelu attention naučí zarovnat řečová data na úrovni jednotlivých hlásek, tj. získat délku trvání každé hlásky v trénovacích datech;
- *studentské sítě*, která se učí predikovat nejdřív délku trvání jednotlivých hlásek na základě informací z učitelské sítě a nakonec samotný spektrogram na základě předpovězených délek hlásek.

Obě sítě jsou založeny zejména na architektuře konvolučních sítí. Protože oproti předchozím architekturám redukují potřebu modelů attention (je přítomna jen jedna vrstva attention v učitelské síti, studentská síť nemá žádné), umožňují výrazné zrychlení trénování i inference. Obojí lze totiž zpracovávat paralelně (tj. generovat všechny časové body spektrogramu současně). Autor uvádí několik technik použitých pro zrychlení trénování a zlepšení robustnosti sítě, jako jsou normalizace nebo úpravy dat pomocí náhodného šumu nebo šumu pocházejícího přímo z výstupu aktuální verze sítě. Práce obsahuje také popis velkého množství experimentů, které autor provedl a které ho vedly k finální volbě architektury.

Výsledný TTS systém je schopen syntetizovat výstup cca 5x rychleji než v reálném čase na běžném CPU, většinu tohoto času navíc zabírá externí vokodér (konverze spektrogramu na audiosignál), navrhovaná architektura pro generování spektrogramu je tedy velice rychlá. Na GPU je možné v reálném čase syntentizovat paralelně několik vět. Oproti předchozím architekturám TTS tedy není třeba žádných zvláštních optimalizačních technik. Kvalita syntetizovaného hlasu je přitom hodnocena jako lepší než u dvou jiných moderních neuronových TTS

modelů (vyhodnocení bylo provedeno pomocí subjektivního porovnání s jinými modely, jejichž implementace jsou dostupné na webu, včetně velmi kvalitního Tacotronu 2, studie se účastnilo 40 dobrovolníků).

Text práce začíná stručným úvodem, který představuje cíle a motivaci práce. Následuje přehled teoretického pozadí, který vysvětluje problém syntézy řeči, zpracování signálu i strojové učení a neuronové sítě, přičemž se soustředí zejména na neuronové architektury používané v další práci. 3. kap. podává přehled relevantní literatury k TTS, který se soustředí zejména na neuronové modely; speciální sekce je věnovaná přístupům, na nichž je práce přímo založena. 4. kap. v detailu popisuje zvolenou finální architekturu včetně motivace, jednotlivých komponent i jejich způsobu trénování. 5. kap. jednak popisuje open-source řečový korpus LJ Speech, na kterém autor prováděl experimenty, jednak dokumentuje průběh prací – ukazuje, které experimenty autor provedl a na základě čeho zvolil finální architekturu sítě. Nejsou tu vynechány ani experimenty, které nevedly ke zlepšení výkonu. 6. kap. představuje evaluaci modelu z hlediska kvality výstupu i z hlediska rychlosti trénování a inference. Autor nakonec uvádí i možnosti dalšího rozšíření a vylepšení modelu do budoucna. Závěrečná 7. kap. je jen krátkým shrnutím vysledků práce.

Průběh prací Autor na své diplomové práci pracoval intenzivně a soustavně cca posledních 10 měsíců; byli jsme přitom pravidelně v kontaktu a veškeré experimenty spolu pravidelně konzultovali. Přitom však hlavní iniciativa vycházela od autora – aktivně si dohledával literaturu, řešil problémy vzniklé během úvodních experimentů a přicházel s vlastními inovacemi. Stejně aktivní byl i při psaní samotného textu, jehož podobu jsme průběžně a velmi detailně konzultovali. Průběh práce považuju za příkladný.

Hodnocení Vědecký obsah práce považuji za výborný, jednoznačně převyšující nároky kladené na diplomovou práci. Jedná se de facto o samostatnou výzkumnou práci na nejvyšší úrovni, což dokládá i fakt, že výsledek práce jsme společně s autorem počátkem května zpracovali do článku a odeslali k recenznímu řízení na prestižní konferenci Interspeech. Výsledek práce má potenciál zlepšit dostupnost vysoce kvalitních TTS hlasů pro další jazyky díky snadnějšímu trénování. Dalšímu výzkumu jistě napomůže i kompletní zdrojový kód zveřejněný na GitHubu a ukázky výstupů na doprovodné webové stránce.

Text práce je psán srozumitelně a přehledně, veškeré moje připomínky jsme už probrali v průběhu práce a autor je do textu zapracoval, proto k němu nemám žádné výhrady.

Celkově tedy práci velmi silně doporučuji k obhajobě; žádné dotazy k obhajobě nemám.

Práci doporučuji k obhajobě.

Práci navrhuji na zvláštní ocenění.

Práci určitě doporučuji do soutěže o zvláštní ocenění. O konkrétních soutěžích jsem zatím neuvažoval, ale mám to v plánu. Důvody jsem v podstatě popsal už v hodnocení práce – jedná se o samostaný výzkum, který obstojí ve srovnání s nejnovějšími vědeckými pracemi v oboru.

V Praze dne 30. 6. 2020

Podpis: