Oponentní posudek na disertační práci Ing. Jakuba Víta

Doktorský studijní program, studijní obor: Kybernetika
Západočeská univerzita v Plzni, Fakulta aplikovaných věd

Generování české řeči pomocí neuronových sítí

Ing. Jakub Vít se ve své disertační práci zabývá jednou z úloh zpracování řeči. Práce se váže k tématu syntézy řeči. Vývoj metody i experimenty se zaměřují na nové architektury pro počítačové generování řeči pomocí neuronových sítí. Tato metodika je dnes již nedílnou součástí moderní informatiky. V posledních letech došlo k obrovskému rozvoji nejenom poznatků o teoretických aspektech, ale dnes především o aplikacích umělých neuronových sítí. Mezi tyto aplikace patří také metody zabývající se zpracováním řeči. V případě předkládané disertační práce se jedná o generování české řeči s vysokou kvalitou a přirozeností.

Předkládaná práce je členěna do 12 kapitol a je doplněna rozsáhlým seznamem prostudované literatury (obsahuje 74 položek), seznamem vlastních publikací nebo publikací, jejichž je spoluautorem (28 položek, z toho jeden U.S. patent), dále abstraktem a klíčovými slovy (v češtině a angličtině), obsahem, seznamem obrázků (71), tabulek (15) a seznamem zkratek a symbolů. Práci dokreslují experimenty popisující výstupy, kterých bylo dosaženo vlastní implementací metod WaveNet a WaveRNN na syntézu českého jazyka, a také experimenty, jejichž cílem bylo navrhnout a vyvinout nový systém TTS pro syntézu řeči s vyšší kvalitou než v té době stávající systém, který byl založen na konkatenační syntéze.

V první polovině práce se autor zaměřuje na stručný úvod do problematiky, vytyčuje si cíle své disertační práce, popisuje její strukturu a časový kontext. Vzhledem k tomu, že na své disertační práci pan Ing. Vít pracoval delší dobu a také s ohledem na budoucí realizaci, bylo nezbytné provést řadu experimentů, včetně poslechových testů. Experimenty byly využity také k průběžné korekci směru výzkumu. Poznatky plynoucí z experimentů, ale i rychlý vývoj nových metod je charakteristický zejména pro aplikace umělých neuronových sítí.

Pan Ing. Vít se nejdříve zabývá obecným rozborem problematiky syntézy řeči, jako je zpracování textu, fonetická transkripce, vlastní syntéza řeči a hodnocení její kvality (přirozenosti a srozumitelnosti) poslechovými testy. Jednu kapitolu věnuje stručnému popisu tradičních metod syntézy řeči, jako je konkatenační syntéza a parametrická syntéza HMM, ale také zmínkám o parametrické syntéze pomocí umělých neuronových sítí (LSTM sítě, GRU sítě a obousměrné rekurentní sítě). V této kapitole se stručně zmiňuje o lingvistických příznacích a základních komponentách neuronových sítí, jako jsou hyperparametry, aktivační a ztrátové funkce. Zbylé kapitoly, počínaje kapitolou 4, jsou pro práci stěžejní. Konkrétně se jedná o popis architektury WaveNet a WaveRNN a jejich funkce, které jsou hlavním tématem disertační práce. V šesté kapitole disertant popisuje původní systém SpeechLab, který vytvořil jako podpůrný nástroj pro vlastní vývoj syntézy řeči. V závěru se disertant velmi stručně věnuje závěrečnému zhodnocení výsledků a zamýšlí se nad možnostmi výzkumu v této oblasti v budoucnu.

Téma zvolené pro disertační práci je velmi aktuální. Generování a syntéza řeči je potřebná v mnoha oblastech lidské činnosti. Velmi užitečná je pak především syntéza z textu (TTS). Vedle dnes už nutného požadavku na srozumitelnost je jedním z hlavních požadavků přirozenost syntetické řeči a její produkce v reálném čase. Jedná se o velmi důležitou oblast zpracování signálu, kterou se zabývají velmi intenzivně mnohá výzkumná pracoviště na celém světě. Obtížnost řešení souvisí s charakterem řečového signálu, který závisí na nestacionaritě

signálu, na lingvistických vlastnostech, ale také na fyziologických vlastnostech člověka. K velkému množství příspěvků na nejrůznějších konferencích, seminářích a workshopech přidal svůj díl i autor předkládané disertační práce.

Hlavním cílem práce je výzkum vlastností a realizace nových architektur pro generování české řeči pomocí umělých neuronových sítí. Základem výzkumu je vytvoření nového TTS frameworku založeného na aplikacích těchto sítí, který je použitelný i v praxi. K řešení přispívá i systém SpeechLab, který vznikl jako vedlejší produkt výzkumu a který usnadňuje práce související s výzkumem a s nasazením syntézy řeči v praxi (ruční segmentace, anotace, nahrávání a sestavování trénovacích dat).

Navržené řešení je komplexní, použité metody jsou moderní a jejich výběr je v souladu se současnými trendy v oblasti počítačového zpracování signálů. Výzkum v této oblasti předpokládá rozsáhlé znalosti v mnoha oborech, které jsou podepřeny velmi dobrým matematickým zázemím a zkušenostmi s řešením otázek zpracování řečového signálu, které měl autor možnost získat na jednom z předních pracovišť v tomto oboru. Za stěžejní výsledek a původní přínos disertanta považuji vytvoření nového způsobu generování české řeči, která má svá specifika oproti, v literatuře převážně citovaným jazykům, kterými jsou angličtina a němčina. Čeština je totiž velmi obtížným jazykem ke zpracování standardními matematicky algoritmizovatelnými metodami díky svým fonetickým a fonologickým vlastnostem. To je také, podle mého názoru, hlavní důvod velkého rozvoje metod založených na umělých neuronových sítích i na pracovištích v naší republice.

Po formální a technické stránce je předkládaná práce na dobré úrovni, i když text disertační práce je nevyvážený, některé kapitoly jsou velmi dobře propracovány, jiné formulace, zřejmě převzaté z uvedené literatury a přeložené do češtiny, jsou poněkud těžkopádné, často působí jako „vytržené z kontextu“. Přesto mohu konstatovat, že disertační práce je psána přehledně, autor prokázal schopnost pracovat tvůrčím způsobem, je schopen orientovat se v literatuře. Také logická stavba práce je na dobré úrovni. Oceňuji používání anglických termínů u pojmů vysvětlovaných v češtině. Tento způsob pomáhá lepší orientaci čtenáře v zahraniční literatuře.

Odpovídající počet prací, jichž je autorem či spoluautorem, publikovaných ve sbornících na prestižních mezinárodních konferencích a v časopisech, dokazuje nejenom autorovu schopnost vědecky pracovat, ale i schopnost informovat o dosažených výsledcích.

Mám několik připomínek a dotazů. Mezi připomínky patří např. konstatování, že:

- Některé formulace jsou vhodné spíše do beletrie, než do technického textu. Např. „poražení metody“, „vysyntetizované“, „geniální“, „nařezání“, „důmyslné triky“ nebo „WaveNet dokáží předehnat metodu HMM“, „signál trpí“.
- Na str. 44 píšete, že rekurentní sítě mají schopnost pracovat bez kontextu. To nepovažuji za výhodu, protože kontext je důležitý pro určení kvality výsledné řeči. To platí zejména u vícejazyčného zpracování.
- Používáte pojem „hluboké sítě“. V práci by měl být alespoň stručný popis tohoto typu sítě.
- Na str. 46 píšete, že „Vstupní hodnoty neuronu jsou nejprve spočítány pomocí vah (weights) a přičtením konstanty (bias), poté jsou předány do aktivační funkce. Aktivační funkce nemusí mít vždy práh (bias). Funkce prahu souvisí s možností ovlivňovat rozsah hodnot vstupních a výstupních signálů.
- Str. 47: MSE není součet kvadrátů rozdílů mezi skutečnou a predikovanou hodnotou. Tím je SSE (sum squares error).

- Z textu vyplývá, že se domníváte, že pouze WaveNet umí generovat řeč bez použití parametrizace. To ale umí většina modelů založených na neuronových sítích.
- Studie pro určení kvality, o kterých se zmiňujete, jsou pro mandarínskou čínštinu a němčinu. Pro slovanskou skupinu jazyků je určování kvality a srozumitelnosti mnohem složitější. To platí nejen pro standardní metody, ale i pro neuronové sítě.
- Nevidím důvod vytvoření umělých vět, která nedávají smysl, pro poslechové testy.
- V textu pod obr. 71 píšete, že první neuronová síť schopná generovat kvalitní řeč byla architektura WaveNet. S tím nesouhlasím. První neuronová architektura pro kvalitní TTS (pro angličtinu) byl NetTalk autorů Terrence J. Sejnowskiho a Charlese R. Rosenberga, publikovaná v Technické zprávě John Hopkins University už v r. 1986.
- Nesouhlasím s tvrzením, že výsledky neukazují závislost mezi úspěšností a tím, o jaký jazyk se při syntéze jednalo (kap. 10.4, str. 89, 3. odstavec). Těžko se srovnává, když nebyl použit stejný nebo alespoň přibližný počet slov.

Dotazy k práci mám následující:

- Jak lze řešit problém „zapomínání“ u neuronových sítí, a to nejen rekurentních.
- Jaké další typy neuronových sítí znáte, kromě rekurentních.
- Vysvětlete obr. 15 a 16. Nevidím žádnou rekurenci.
- V čem jsou výhody a nevýhody hlubokých neuronových sítí?
- Kolik skrytých vrstev má hluboká síť WaveNet? Je to vždy 30?
- Str. 69, 2. odst. pod výčtem: Píšete, že „délka vstupní a výstupní sekvence je počet fonémů ve větě a výstupem je jejich trvání“. Znamená to, že délka sekvencí je proměnná a závisí na délce věty?
- Jaké lingvistické příznaky byly použity v pozičním vektoru (dimenze 4)?
- Věty, na kterých jste testoval metodu WaveNet, byly pouze věty oznamovací nebo také otázky, příkazy apod.?
- Pro jaké jazyky je vámi vyvinutý systém určen?

Na závěr konstatuji, že i přes uvedené připomínky, které nejsou zásadního rázu, pan Ing. Jakub Vít projevil schopnost samostatně vědecky pracovat. Vytčené cíle byly splněny. Proto mohu konstatovat, že **disertační práci doporučuji k obhajobě**.

V Praze, 22. 3. 2023

Prof. Ing. Jana Tučková, CSc.

e-mail: tuckovaj12@gmail.com
tel.: +420721502556

62,4 se směrodatnou odchylkou 26,8, zatímco WaveRNN průměr 79,8 se směrodatnou odchylkou 18,6. To znamená, že tyto směrodatné odchylky jsou obrovské, oba intervaly průměr±odchylka se velmi překrývají a zároveň bylo velmi málo respondentů (navíc skoro polovina z nich byla „biased"). Úplně chybí jakákoli další (byť jakkoli elementární) statistická analýza signifikantnosti a robustnosti, abychom viděli, jaká je pravděpodobnost, že takto získané výsledky nejsou zcela náhodné. V tomto kontextu pak autorovo tvrzení *„Závěry testu jsou zřejmé. Nově navržený [...] systém [...] porazil v poslechovém testu stávající systém [...]"* působí snad až úsměvně naivně. Bohužel v podobné metodologické (ne)kvalitě jsou v podstatě všechny poslechové testy.

Vzhledem k tomu, že cíl práce byl definován jako „navrhnout a vyvinout metodu syntézy řeči, která by dosáhla lepších výsledků než stávající [...] systém", považoval bych cíl za splněný. Metoda byla vyvinuta a implementována. Jen poněkud pokulhává její experimentální vyhodnocení. Nicméně ze zahraniční literatury je zřejmé, že prakticky ve všech kontextech je WaveRNN lepší než stará Unit selection. Můžeme tedy s přijetím určitého rizika předpokládám, že by se to stejné skutečně potvrdilo i zde. Škoda, že si autor nedal během uplynulých 8 let na těchto několika málo poslechových testech více záležet.

c) Stanovisko k výsledkům disertační a k původnímu konkrétnímu přínosu

Toto stanovisko jsem v podstatě již podal v části (a). Zde bych tedy jen zopakoval, že největším původním přínosem práce je patrně vznik softwarového nástroje SpeechLab pro využití v rámci vývoje systému ARTIC. Pro externího hodnotitele je však velmi obtížné zhodnotit, jaká je skutečná kvalita tohoto nástroje a jaký je jeho přínos alespoň v praktické rovině. Z kontextu předpokládám, že nástroj funguje dobře a prokázal určitou užitečnost.

Původní vědecký přínos práce je však bohužel velmi nízký. Sice vznikla vlastní „domácí" implemetace algoritmu WaveNET – k čemu je to ale dobré, když už dávno existuje několik open-source implementací téhož, které jsou navíc patrně efektivnější a mají celou komunitu uživatelů a vývojářů? Navíc mezitím vznikla celá řada dalších (a patrně i lepších) modelů jako Tacotron, DeepVoice, FastSpeech apod. (vesměs již v letech 2017-2020, tedy zcela v průběhu dlouhého vznikání této disertační práce). Jak si vůči nim stojí „domácí" implementace WaveNetu? Nebylo by lepší pro další vývoj ARTICu stejně sáhnout po některé osvědčení open-source implementaci?

Mé výhrady k přínosu poslechových testů jsem zminil již v části (b).

d) Vyjádření k systematice, přehlednosti, formální úpravě a jazykové úrovni

Po formální stránce práce splňuje všechny požadavky. Strukturovaná je přehledně. Jazykově je také zpracována kvalitně.

e) Vyjádření k publikacím studenta

Jakub Vít je hlavním autorem nebo spoluatorem dostatečného množství odborných publikací. Můj intuitivní pocit při pohledu na seznam publikací je, že nejintenzivnější a nejzajímavější publikační období bylo v letech 2017–2019 – tedy v době největšího rozkvětu LSTM-based

Oponentský posudek disertační práce Ing. Jakuba Víta

Téma disertační práce: **Generování české řeči pomocí neuronových sítí**

Posudek vypracoval: Ing. Mgr. Jan Romportl, Ph.D. (CEO at Elin.ai & Deep Tech Partner at Presto Ventures)

a) Zhodnocení významu disertační práce pro obor

Ačkoli je z disertační práce patrné, že pan Vít má opravdu hluboké znalosti tématu syntézy řeči (a to ve zcela praktickém smyslu vývoje softwarových implementací), jde bohužel dle mého názoru a značně promarněnou šanci.

Největší problém vidím v neadekvátně dlouhém čase vzniku celé práce. Sám autor uvádí (v odst. 1.3), že na práci dělal v letech 2015 až 2023, tedy prakticky 8 let. Vzhledem k rychlosti světového vývoje v oblasti AI je až neuvěřitelné, jak málo vědecky relevantních výstupů tato disertační práce nakonec obsahuje.

V roce 2015 šlo skutečně o aktuální téma. Kdyby autor dané výsledky prezentoval třeba v roce 2019, bylo by to sice již poněkud za světovým průměrem, ale pořád by to určitý lokální impakt mohlo mít. Když je ale prezentuje v roce 2023, troufám si říci, že jde o výsledky v podstatě zbytečné. Nadto je z mého pohledu diskutabilní i jejich metodologie – viz dále.

Z tohoto důvodu je význam předkládané disertační práce pro obor z mezinárodního hlediska velmi blízký nule. Z lokálního hlediska by snad bylo možné najít určitý „dokumentární" význam – tedy jak práce mapuje určité epizody vývoje TTS systému ARTIC.

Mohu se domnívat spíše z širšího kontextu (nikoli přímo z textu disertace), že největší význam disertační práce jsou její softwarově-implementační výstupy konkrétních algoritmů a nástrojů pro systém ARTIC. Přepokládám však, že i u nich by bývalo bylo lepší, kdyby vznikly daleko dříve – rozhodně by to více pomohlo systému ARTIC v prostředí jiných konkurenčních TTS systémů.

b) Vyjádření k postupu řešení problému, k použitým metodám a splnění určeného cíle

Samotná softwarová implemetace algoritmů WaveNet a WaveRNN patrně proběhla dobře, ačkoli to čistě jen z textu disertace nelze ověřit. Uvítal bych například nějaké funkční demo nebo GitHub repozitář (ideálně obojí).

Vědecké jádro disertace pak spočívá v prezentaci pěti poměrně triviálních experimentů – vesměs založených na velmi jednoduchých poslechových testech.

Zde mám dvě zásadní výhrady:

1) Reprodukovatelnost: úplně chybí jakékoli materiály, které by umožnily nezávislé zreprodukování a ověření výsledků.
2) Pochybná metodologie a průkaznost: vezměme např. poslechové texty z kapitoly 9 – těch se zúčastnilo pouhých 18 posluchačů, z čehož 8 mělo expertní znalosti v této oblasti. Výsledky testu v Tabulce 10 pak ukazují, že Unit selection má průměrné skóre

technik a WaveNetu. Troufl bych si říci, že kdyby byla bývala disertace publikována už tehdy, mohla být hodnocena v úplně jiném kontextu a mohla mít daleko lepší impakt.

f) Jednoznačné vyjádření oponenta k doporučení či nedoporučení k obhajobě

Disertační práci **doporučuji k obhajobě**.

Komentář: Ačkoli na disertaci spatřuji velké množství negativních věcí a poměrně velkých problémů, domnívám se, že přínos Jakuba Víta k rozvoji TTS systému ARTIC je rozhodně nezanedbatelný. Zároveň mě osobně zaráží určitý překvapující nepoměr mezi skutečnými schopnostmi pana Víta a tím, jak se patrně nedostatečně zrcadlí v samotném textu disertace. Je také možné, že mi není znám celý kontext vzniku práce a důvodů, proč se tak protáhla (což mělo zásadní negativní vliv na její přínos a význam). Jsem tedy přesvědčen, že pan Vít musí mít možnost práci obhajovat a vše náležitě vysvětlit.

V Plzni, 26.6.2023

Ing. Mgr. Jan Romportl, Ph.D.

DocuSigned by:

BE764B6CC471473...