Posudek disertační práce Ing. Karla Veselého

Semi-supervised Training of Deep Neural Networks for Speech Recognition

Hlavním tématem předložené disertační práce je trénování hlubokých neuronových sítí v systémech rozpoznávání mluvené řeči s využitím jak anotovaných, tak i neanotovaných trénovacích nahrávek (v práci se tento způsob trénování nazývá přiléhavým anglickým termínem "semi-supervised learning"). Toto téma zcela určitě odpovídá oboru disertace a je bezpochyby vysoce aktuální, a to ze dvou důvodů. Zaprvé, neuronové sítě jsou v posledních letech jednoznačně "nástrojem první volby" pro vývoj systémů rozpoznávání řeči. Zadruhé se v posledním desetiletí výrazně zvýšilo množství dostupných digitálních řečových nahrávek, které je potenciálně možno použít pro trénování systémů rozpoznání řeči. Tradičně však musejí být takové nahrávky před vlastním trénováním nejprve ručně anotovány, což je proces časově a finančně náročný. Snaha autora práce o nalezení co nejefektivnějšího způsobu využití těchto nahrávek bez nutnosti manuální anotace je tedy pro obor velmi prospěšná.

Práce je logicky členěna do 10 kapitol. Po úvodních 2 kapitolách, které čtenáře se seznámí s cíli disertační práce a uvedou jej do problematiky využití hlubokých neuronových sítí (DNN) v rozpoznávání řeči, následuje kapitola 3 popisující autorovu vlastní implementaci DNN technik do softwarového nástroje Kaldi. Popisovaná implementace - jakkoliv jde z velké části o tzv. "inženýrskou práci" - je dle mého názoru také významným autorovým přínosem pro celou komunitu věnující se vývoji systémů pro rozpoznávání řeči, neboť Kaldi je v současnosti zřejmě nejpoužívanějším volně dostupným softwarem pro tyto účely. Kapitola 4 stručně představuje jazykové sady použité pro trénování a testování navržených algoritmů a kapitoly 6 až 9 jsou již věnovány hlavnímu tématu práce - vývoji obecného postupu, který by dokázal co nejlépe využít velké množství dostupných neanotovaných nahrávek pro zlepšení kvality rozpoznávání řeči. Všechny varianty v práci navržených postupů vycházejí z toho, že je nejprve s využitím ručně anotovaných dat natrénován základní (seed) systém, který se poté použije pro automatické vytvoření textových přepisů původně neanotovaných nahrávek. Tyto přepisy však samozřejmě ani zdaleka nejsou stoprocentně správné - klíčovým problémem celého procesu "semi-supervised" trénování je tedy otázka vhodného ohodnocení automatických přepisů pomocí tzv. míry důvěry (confidence measure). Autor provedl velké množství experimentů, převážně věnovaným testování vhodné velikosti jednotek, na kterých by bylo nejpřínosnější zmíněnou míru důvěry vyhodnocovat (věty, slova, fóny a jednotlivé stavy akustického modelu). Kombinace algoritmů, která poskytovala nejlepší výsledky na vietnamštině, byla následně otestována i na standardním anglickém korpusu Switchboard. Výsledky těchto testů naznačují, že se autorovi skutečně podařilo najít univerzální postup pro "semi-supervised" trénování. Tento výsledek považují za originální a dostatečně významný přínos předložené dizertační práce.

Práce je psaná dobrou a srozumitelnou angličtinou. Jazykově nejslabším článkem práce je dle mého názoru český abstrakt, který obsahuje zbytečně velké množství anglicismů a pravopisných chyb. Členění do kapitol je z velké části přehledné a logické - jen kapitola 6 působí v kontextu ostatních kapitol poněkud nekonzistentně. Často je v ní z jednoho experimentu nejprve vyvozen nějaký zdánlivě obecný závěr (například na straně 66 - "*In section 6.1, we saw that repeating the manually transcribed data 3x helps*"), který je vzápětí nahrazen jiným zdánlivě obecným závěrem (tentýž odstavec - "*Hence, it is sufficient to*

include manually transcribed data 1x"), o němž je však zřejmé, že byl stejně jako dřívější úsudek formován na základě jediného experimentu. Samotné shrnutí kapitoly 6 pak v zásadě všechny tyto průběžně formulované "závěry" neguje. Rozumím tomu, že takto ve skutečnosti většinou experimentální přístup ve vědě funguje, jen bych při tvorbě finální práce preferoval pečlivější redakci textu.

Z formálních náležitostí textu mám výhrady především k práci s citacemi - někdy se zkratka konference, na které byla část práce publikována, zbytečně a zcela bez potřebného kontextu objevuje přímo v názvu podkapitoly (*6.1 Frame selection by confidence (ASRU2013)*), jindy je citován neúplný název článku nebo jsou zaměněna jména a příjmení autorů (str. 5 - *[Jan et al., 2010]*). Také jsem při čtení práce nejprve ocenil - ne vždy vídaný - seznam zkratek a použitého značení, ale později se ukázalo, že ne všechny zkratky jsou v seznamu uvedeny a navíc jejich používání často není v práci konzistentní (příklad za všechny - poměrně klíčová metrika "*word error rate recovery*", zavedená ve vztahu (5.1) pod označením R_{WER} a v dalších kapitolách používaná již jen pod označením "*WER recovery*" nebo zcela opominutá).

Autor disertace svoje výsledky hojně průběžně publikoval - je autorem či spoluautorem více než 30 článků, v drtivé většině publikovaných na nejprestižnějších konferencích oboru (ICASSP, Interspeech, ASRU). Kvalitu publikační činnosti tedy považuji za vysoce nadprůměrnou, o vědecké erudici uchazeče není podle mého názoru nejmenších pochyb, což dokládá i jeho opakované zapojení do mezinárodních výzkumných týmů.

Přes výše uvedené drobné výhrady práci považuji za podstatný přínos k oboru, práce odpovídá obecně uznávaným požadavkům k udělení akademického titulu Ph.D. a plně ji doporučuji k obhajobě.

Dotazy k obhajobě:

1. Ve všech experimentech je poměr anotovaných a neanotovaných dat 1:7. Testoval jste nějak, zda jsou Vaše závěry platné i pro jiný poměr dat?

2. Jak jste dospěl k rovnici (5.4) na straně 53 - odvozením z rovnic (2.24) a (2.26) či "heuristickým postupem"?

3. V práci není detailně popsáno, jak probíhala optimalizace meta-parametrů. Pokud byly optimalizovány na vývojové (development) sadě, pak jak byla v jednotlivých případech velká? A jaké pak lze vyvozovat závěry na testovacích datech?

V Plzni 14.12.2017 doc. Ing. Pavel Ircing, Ph.D.