Rozšířený Abstrakt

Využití hlubokých neuronových sítí pro úlohu rozpoznávání řeči

Ing. Lukáš Matějů, Ing. Petr Červa, Ph.D.

Abstrakt

Příspěvek je věnován hlubokým neuronovým sítím a to konkrétně architektuře DNN-HMM, která spojuje skryté markovské modely a hluboké neuronové sítě, pro rozpoznávání spojité řeči. Experimentálně je ověřeno, že tato architektura překonává klasické modely založené čistě na skrytých markovských modelech. Ve zbývající části příspěvku jsou hledány ideální parametry neuronové sítě pro natrénování modelů pro češtinu. Příspěvek je konkrétně zaměřen na šířku a hloubku sítě.

Úvod

Za standardní přístup k akustickému modelování pro různé úlohy zpracování řeči byly po dlouhou dobu považovány skryté markovské modely (HMM) využívající mixtury (GMM), tzv. GMM-HMM modely. Před nedávnem byl objeven nový přístup [1], který kombinuje klasické HMM modely s hlubokými neuronovými sítěmi (DNN). Tato nová hybridní DNN-HMM architektura přináší značné zlepšení úspěšnosti rozpoznávání v celé škále aplikací. Příkladem aplikace je detekce řeči Google Voice [2] na mobilních telefonech, případně na portále YouTube [3]. Microsoft využívá DNN u vyhledávače Bing [4]. Mezi další aplikace patří rozpoznávání spojité řeči s velkým slovníkem [1], detekce klíčových slov, přepis konverzací, detekce mluvčího a další.

Cílem tohoto příspěvku je porovnání klasických GMM-HMM s architekturou DNN-HMM na našich datech pro rozpoznávání spojité řeči – češtiny. Hlavním cílem práce je nalezení ideálních parametrů neuronových sítí, které přinesou nejvyšší úspěšnost rozpoznávání. Tento příspěvek se konkrétně věnuje parametrům šířce a hloubce neuronové sítě.

Experiment a metody

K natrénování hlubokých neuronových sítí je používán framework Torch. Všechny sítě jsou trénovány na 300 hodinách mluvené češtiny. Sítě jsou trénovány po 35 epoch s využitím dávek o velikosti 1024 a rychlostí učení 0,08. Aktivační funkce je ReLu. Příznakové vektory jsou složeny z 11 MFCC vektorů, 5 předcházejících, současného a 5 následujících. DNN přebírá architekturu GMM-HMM, poskytuje odhady věrohodnosti pro fyzické markovské stavy.

K vyhodnocení natrénovaných modelů je použit náš rozpoznávač. Jazyková část je založena na slovníku a jazykovém modelu. Slovník obsahuje 550 000 unikátních záznamů, jazykový model je bigramový. Testovací data jsou rozdělena do několika kategorií podle zdroje – nahrávky vysílání, diktáty soudů, záznamy přednášek a nelineárně zkreslené záznamy. Přehled je v tabulce 1. Pro diktáty je použit menší slovník a tematicky zaměřený jazykový model.

Tabulka 1. Přehled testovacích sad

Sada	Vysílání	Diktát	Přednášky	Zkreslené
Slova	99 858	5 714	81 436	35 132
Hodiny	11,5	1	13	4,5

Základní experiment porovnává DNN-HMM se GMM-HMM architekturou. Následující experiment je zaměřen na porovnání sítí s šířkou 512, 1024, respektive 2048 neuronů. Síť obsahuje 5 skrytých vrstev. Závěrečný test zkoumá hloubku sítě, šířka je 1024. Příznakové vektory u tohoto experimentu jsou FBANK, které poskytují další zvýšení úspěšnosti rozpoznávání.

Výsledky a diskuze

DNN-HMM architektura, jak ukazuje tabulka 2, poskytuje i v základním nastavení výrazné zlepšení na všech testovacích sadách. Největší je u sady zkreslené, což potvrzuje domněnku, že DNN jsou více robustní.

Tabulka 2. Úspěšnost [%] rozpoznávání – porovnání GMM-HMM a DNN-HMM architektur

Sada	Vysílání	Diktát	Přednášky	Zkreslené
GMM-HMM	80,93	86,88	73,23	43,77
DNN-HMM	**85,34**	**88,03**	**78,85**	**65,72**

Vyhodnocení experimentu na šířku sítě je v tabulce 3. Nejvyšší úspěšnosti rozpoznávání poskytují sítě s 1024, respektive 2048 neurony na skrytou vrstvu. První uvedená síť je ale výrazně rychleji natrénovaná a výpočetně efektivnější při dekódování.

Tabulka 3. Úspěšnost [%] rozpoznávání pro experiment se šířkou sítě

Sada	Vysílání	Diktát	Přednášky	Zkreslené
512 neuronů	82,85	87,41	76,02	62,26
1024 neuronů	**85,34**	**88,03**	**78,85**	**65,72**
2048 neuronů	85,19	87,73	78,56	65,62

Podle výsledků z tabulky 4 je ideální síť s 5 skrytými vrstvami. Časově úspornější variantou je díky podobným výsledkům i síť se 4 skrytými vrstvami. Přidání 6. vrstvy naopak nepřináší žádné další zlepšení.

Tabulka 4. Úspěšnost [%] rozpoznávání pro experiment s hloubkou sítě

Sada	Vysílání	Diktát	Přednášky	Zkreslené
4 skryté vrstvy	86,47	87,87	80,35	68,27
5 skrytých vrstev	**86,55**	**88,09**	**80,57**	68,18
6 skrytých vrstev	86,50	87,78	80,43	**68,40**

Závěr

Provedené experimenty dokazují, že architektura DNN-HMM poskytuje značné zlepšení v úspěšnosti rozpoznávání na všech testovacích sadách a podporují tak všeobecné mínění. V rámci celé práce jsou hledány ideální parametry pro naše trénovací data pro češtinu. Příspěvek pokrývá testy na vyhodnocení šířky a hloubky neuronové sítě. 1024 neuronů a 5 skrytých vrstev je experimentálně zjištěná ideální konfigurace. Časově efektivnější variantou je síť s 512 neurony na 4 skrytých vrstvách.

Další experimenty je možné zaměřit například na otestování různých aktivačních funkcí nebo odlišných parametrizací vstupních dat, případně na délku příznakových vektorů. Předtrénování DNN je také další možný směr testování.

Reference

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, *Context-dependent pre-trained deep neural networks for large vocabulary speech recognition*, Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 1, pp. 30-42, 2012.

[2] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, *An application of pretrained deep neural networks to large vocabulary Conversational Speech Recognition*, Tech. Rep. 001, University of Toronto, 2012.

[3] N. Ryant, M. Liberman, and J. Yuan, *Speech activity detection on YouTube using deep neural networks*, in Proc. INTERSPEECH 2013, 2013.

[4] G. E. Dahl, D. Yu, L. Deng, and A. Acero, *Large vocabulary continuous speech recognition with context-dependent DBN-HMMS*, in Proc. ICASSP 2011, 2011.