Studentská Konference Fakulty Mechatroniky, informatiky a mezioborových studií
2. červen 2015, Liberec, Česká republika

Využití hlubokých neuronových sítí pro úlohu rozpoznávání řeči

TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

Ing. Lukáš Matějů
Ing. Petr Červa, Ph.D.
Ústav informačních technologií a elektroniky

Abstract

This contribution is dedicated to deep neural networks and DNN-HMM architecture for continuous speech recognition. It's experimentally proven that this architecture yields significant accuracy boost over standard HMM models. Ideal settings of width and depth of DNNs are evaluated.

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 je detekce řeči na portále YouTube [2]. Mezi další aplikace patří rozpoznávání spojité řeči s velkým slovníkem [1], detekce klíčových slov, přepis konverzací, detekce mluvčího a další.

Cíle

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

Metodika

K natrénování DNN je používán framework Torch. Veškeré sítě jsou trénovány na 300 hodinách mluvené češtiny 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 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 kategorií, nahrávky vysílání, diktáty soudů, záznamy přednášek a nelineárně zkreslené záznamy. 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

Výsledky

DNN-HMM architektura poskytuje výrazné zlepšení na všech testovacích sadách. Největší je u sady zkreslené, což potvrzuje domněnku, že DNN jsou robustnější. Viz tabulka 2.

Tabulka 2. Úspěšnost [%] – DNN-HMM vs. GMM-HMM

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**

Ideální šířkou sítě pro naše data je 1024, respektive 2048 neuronů. První je ale výrazně efektivnější při dekódování.

Tabulka 3. Úspěšnost [%] – parametr šířka DNN

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

Ideální hloubkou sítě je 5 skrytých vrstev, viz tabulka 4. Celkové zlepšení úspěšnosti je způsobeno FBANK příznaky.

Tabulka 4. Úspěšnost [%] – parametr hloubka DNN

Sada	Vysílání	Diktát	Přednášky	Zkreslené
4 skr. v.	86,47	87,87	80,35	68,27
5 skr. v.	**86,55**	**88,09**	**80,57**	68,18
6 skr. v.	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. 1024 neuronů a 5 skrytých vrstev je experimentálně zjištěná ideální konfigurace parametrů DNN pro naše data. Další experimenty je možné zaměřit například na otestování různých aktivačních funkcí nebo odlišných parametrizací vstupních dat, případně na délku příznakových vektorů.

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, 2012.

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

Kontakt

Ing. Lukáš Matějů, lukas.mateju@tul.cz

Tato práce byla podpořena z projektu Studentské grantové soutěže (SGS) na Technické univerzitě v Liberci v roce 2015.