Studentská Konference Fakulty Mechatroniky, informatiky a mezioborových studií
27. květen 2014, Liberec, Česká republika

Využití hlubokých neuronových sítí v systémech rozpoznávání řeči

TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky a mezioborových studií

Bc. Martin Paroubek
Ing. Petr Červa, Ph.D.
Ústav informačních technologií a elektroniky

Abstract

The thesis dealt with using a new DNN-HMM system for speech recognition. The goal was to establish the level of influence of the neural network's layout, its pretraining and the size of its training data set on the accuracy of its recognition. The training of this neural network was conducted on a GPU using modified scripts from the Theano library. There was a training data set available with 56 hours of Polish speech and the resulting models were tested on three sets, which contained texts in journalistic and professional style. The so-called "accuracy" was used to compare the results. It was established that using neural networks as acoustic models results in an improvement of several percent over the system currently in use and futhermore that pretraining using the discriminative method has no effect on the network's accuracy. Topology was described with the highest degree of accuracy and it was concluded that the amount of data present in the training data set may be dependent on the context of the testing set.

Úvod

Hluboké neuronové sítě byly označeny v článku NY Times [1] za největší změnu v přesnosti od roku 1979 a vzhledem k množství článku s překonáním současných systémů založených na GMM-HMM bylo rozhodnuto, že je nutné tento přístup otestovat i v laboratoři počítačového zpracování řeči na TUL.
Neuronové sítě v příspěvku byly použity pro vytvoření akustických modelů, které nahradily GMM část současného systému rozpoznávání. Experimentovalo se s různou topologií neuronové sítě, použitým postupem pro trénováním a různým množstvím dat v trénovacím korpusu. Vzhledem k množství různých parametrů a jejich vzájemné závislosti vycházelo se již z úspěšných pokusů publikovaných pro AJ.[2]

Cíl

Cílem bylo prozkoumat vliv uspořádání neuronové sítě, vliv předtrénování a vliv velikosti trénovacího korpusu na přesnost rozpoznávání.

Metodika

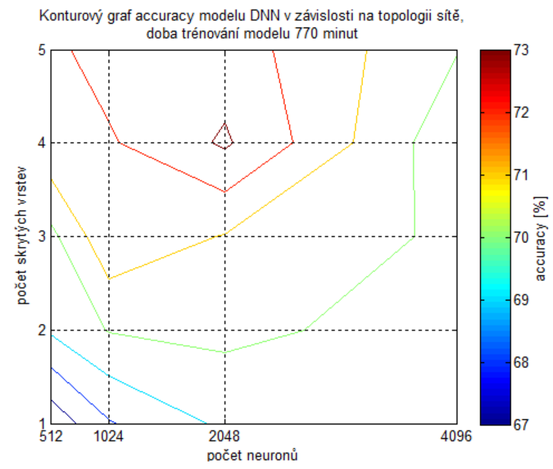
Veškeré experimenty sdílely nastavení těchto parametrů:η=0,08 (parametr učení), α=1 (momentum), i=429 (počet vstupních neuronů resp. příznaků), o=3180 (počet výstupních neuronů resp. stavů HMM). Trénování akustických modelů probíhalo na GPU pomocí algoritmů největšího spádu a zpětného šíření chyby. Trénovací korpus obsahoval 56 hodin spojité polské řeči. U výsledných modelů byla experimentálně ověřena přesnost rozpoznávání na třech testovacích sadách a porovnána s přesností při použití současného systému viz tabulka.

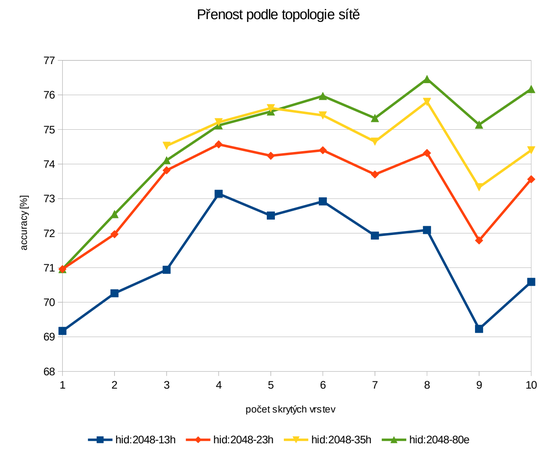
Accuracy baseline systému [%]

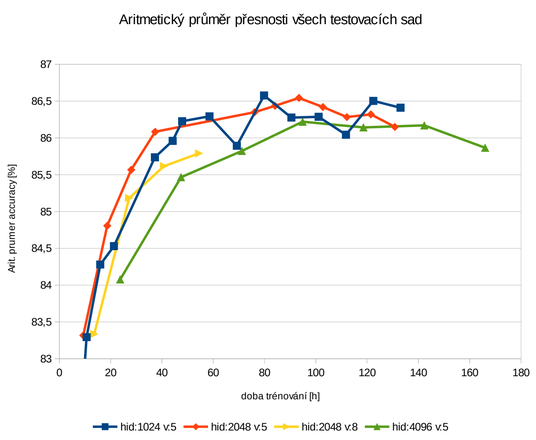
Broadcast	Justice	Pomocne
73,14	88,09	85,16

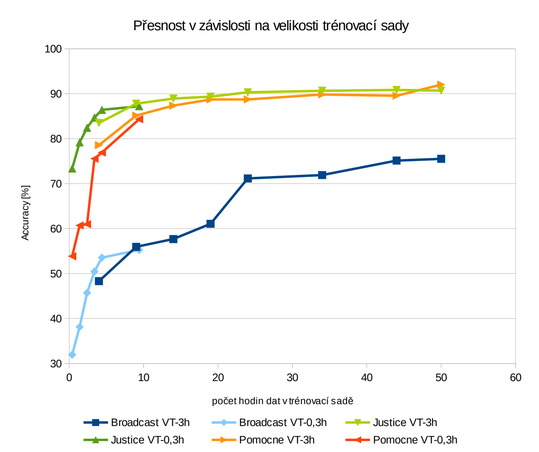
Výsledky

Nejdříve byly odzkoušeny různé topologie neuronových sítí. Doba trénování těchto modelů se pohybovala od 12 hodin až po 7 dní. Bylo zjištěno, že příliš velké sítě (4096 neuronů ve skryté vrstvě) je náročnější natrénovat a nedosahují přesností, které byly natrénovány pro sítě o 1024 nebo 2048 neuronech ve skrytých vrstvách. Zároveň bylo zjištěno, že vyšší počet vrstev přináší užitek do 5 až 6 skrytých vrstev za předpokladu, že byl model trénován do 35 hodin. Při porovnání hodnot přesnosti na dobách trénování 13, 23 a 35 hodin se zdálo, že modely s vyšším počtem vrstev stahují náskok modelů s nižším počtem vrstev. Nejlepší modely byly dotrénovány a bylo zjištěno, že jako nejpřesnější se zdá topologie o šířce 1024 neuronů a 5 skrytých vrstvách a po 80 hodinách trénování. Na testovacích sadách došlo k absolutnímu zlepšení přesnosti o 3% až 7% oproti současně používanému systému.









V další fázích experimentů se zaměřilo na metodu trénování po jednotlivých vrstvách. Byly odzkoušeny dva postupy v prvním byla každá nová vrstva natrénována do co nejlepší přesnosti na testovací části trénovacího korpusu a v závěru proběhlo diskriminativní trénování celé sítě. V druhém pokusu byly jednotlivé vrstvy trénovány po 5 epoch. Ani v jednom případě nedošlo ke zlepšení přesnosti oproti metodě bez předtrénování.
V posledním experimentu bylo úkolem zjistit, o kolik lze přesnost modelu zlepšit přidáním dalších dat. Tento průběh byl sledován na trénovacích korpusech o velikosti 1 hodiny až 56 hodin. Pro testovací sady Justice a Pomocne bylo zjištěno, že dostačující velikost je okolo 30 hodin. Pro testovací sadu Broadcast která měla celkově nižší hodnoty přesnosti ve všech experimentech se zdálo, že by přesnosti přispělo i přidání dalších dat.

Závěr

Výsledný model zaznamenal zlepšení o 3% až 7% oproti současným systémům. Přestože metoda předtrénování nedoznala zlepšení oproti běžnému postupu trénování, mohla by být užitečná v případě menších trénovacích korpusů. Množství dat nutné pro trénování neuronových sítě se zdá být závislé na kontextu použití. Další pokusy by se mohly odvíjet směrem použití jiných aktivačních funkci (např. ReLU) anebo využití znovupoužitelnosti neuronových sítí k natrénování akustického modelu jiného jazyka, pro který třeba není dostatek trénovacích dat.

Reference

[1] Scientists See Promise in Deep-Learning Programs. *The New York Times* [online]. 2012 [cit. 2014-05-10]. Dostupné z:www.nytimes.com /2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html

[2] HINTON, Geoffrey, Li DENG, Dong YU, George DAHL, Abdel-rahman MOHAMED, Navdeep JAITLY, Andrew SENIOR, Vincent VANHOUCKE, Patrick NGUYEN, Tara SAINATH a Brian KINGSBURY. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine.* 2012, s. 82-97. Dostupné z: www.cs.toronto.edu/~hinton/absps/DNN-2012-proof.pdf

Kontakt

Bc. Martin Paroubek martin.paroubek@gmail.com

Tato práce byla podpořena z projektu Studentské grantové soutěže (SGS) na Technické univerzitě v Liberci v roce 2014.