MULTI-TASK NEURAL NETWORKS FOR SPEECH RECOGNITION

Ekaterina Egorova
Master Degree Programme (2. year), FIT BUT
E-mail: xegoro00@stud.fit.vutbr.cz

Supervised by: Martin Karafiát
E-mail: karafiat@fit.vutbr.cz

Abstract: The article covers experiments on TIMIT database exploring the possibility of using multi-task neural networks for speech recognition. Multi-task neural networks are deep neural networks solving several different classification tasks simultaneously. The secondary tasks chosen for the experiments are gender, context, articulatory characteristics and a fusion of some of them. The experiments show that addition of such tasks can enhance the learning and improve recognition accuracy.

Keywords: Speech recognition, neural networks, deep neural networks, multi-task neural networks.

1 INTRODUCTION

In the last few years neural networks have been growing stronger in the field of speech recognition. They are used to train acoustic models, each of the output usually being a phoneme, a tied-state or phone-state. Recent developments in computer power have enabled to train bigger networks, which successfully rivaled and even exceeded traditional Gaussian Mixture Models.

Multi-task neural networks differ from common neural networks in that its output consists of several blocks of nodes, each block representing different classification task. It has been known that for two interconnected tasks which are trained jointly, an accuracy increase is possible [1]. If the tasks do not add to each other's performance, they can still be trained with the same success as separately. Having one network do several connected classifications is not only good in terms of universalizing the training, it can also open new possibilities in multilingual training.

2 MULTI-TASK NEURAL NETWORKS

The idea behind multi-task learning is that it is sometimes more profitable to learn several classification problems simultaneously rather than use separate neural networks for them. In multi-task learning, the network is trained to perform both the primary classification task and one or more secondary tasks using a shared representation in the hidden layers. The network is trained for all the tasks, and the error backpropagates from all of them during the learning. With the addition of the secondary tasks, neural network does not need to change its structure except for the size of the output layer. After training is complete, the portion of the network associated with the secondary tasks is discarded and the classification is performed identically to a conventional single task classifier.

For speech recognition multi-task structure opens uncountable possibilities of usage, as a lot of speech characteristics are interdependent. Multi-task neural networks are not new and have been experimented with since 1989, when classic NETtalk application used one net to learn both phonemes and their stresses [1]. But this is only one of the many possible combinations of tasks that could yield some improvement in speech recognition. Some of the possible settings can include joined learning of segmental and suprasegmental characteristics (e.g. tones in tonal languages), phoneme labels and

phoneme characteristics, or phoneme inventories of different languages in a multilingual task. In a recent paper, [2], the following secondary tasks were explored: the phone label, the phone context, and the state context. The best results were achieved with phone context as a secondary task, with 1.4% decrease in error rate.

3 EXPERIMENTS

The experiments were performed in a Neural Network Trainer TNet framework [1], which was extended to allow multi-task training. The new flavour of objective function goes block by block over the output neurons and calculates cross entropy error function in each block separately. Backpropagation is then performed from all the tasks.

For the experiments, TIMIT database [2] has been chosen. It contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. The TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16kHz speech waveform file for each utterance.

3.1 TEST TASK

The first experiment was intended for the testing of the new scripts. For this experiment, the network was trained to deal with two tasks: classifying phonemes and classifying the gender of the speaker. The network dimensions are 368:500:41, the number of output units equaling the number of phoneme labels (39) plus the number of gender labels (2).

When the resulting weights are generated, the first 39 outputs are taken to produce phoneme hypothesis on test data and the last 2 outputs produce gender hypothesis. The hypothesis setting forbids to change between male and female labels for different frames in one utterance.

On the phoneme task the system yielded the same result as the baseline, the baseline being trained on a neural network with the same dimensions, with only the output layer being of the size of 39. The accuracy rate for both tasks can be seen in table 1. Note that it can differ depending on the initialization (not more than 0.1%).

Baseline	Multi-task
69.5	69.4

Table 1: Scores on phoneme label task (% phoneme accuracy)

On the gender task the multi-task system yielded accuracy 98.0 %, which means it successfully learned both tasks.

3.2 CONTEXT SECONDARY TASKS

After making sure the new setting works, more complex experiments were performed in order to find out which secondary tasks may be helpful for acoustic modeling. For each of the following experiments, including the baseline, a bigger network was trained, with four hidden layers consisting of 2048 units each. The choice of bigger network ensures that it will be able to learn several rather complex tasks.

Initially, the replication of the experiment in [2] was done. First task was phoneme labels (39), second one was left context and the third one was right context (both represented by phoneme labels, of the

[1] http://speech.fit.vutbr.cz/cs/software/neural-network-trainer-tnet

[2] http://catalog.ldc.upenn.edu/LDC93S1

dimension 39 each). The experiment has shown an accuracy increase in comparison with the baseline, as shown in table 2.

3.3 ARTICULATORY SECONDARY TASKS

Phonemes represented in the primary task can be grouped phonetically according to different phonetical classes. The most obvious is the vowel/consonant distinction, but other characterisitcs can also be helpful. For these experiments, the following characteristics were used as secondary tasks: place and manner of articulation, participation of voice in the pronunciation and additional articulatory characteristics, such as rounded/unrounded for vowels. As shown in table 2, throwing all the articulatory secondary tasks together doesn't help, so separate settings have been made for each articulatory characteristic to see which of them help. It turned out that the addition of the information about place and manner of articulation and if the phoneme is a vowel or a consonant yield better results than the baseline. So in the next experiment all the helping articulatory characteristic were thrown together, which produces an even bigger increase in accuracy.

3.4 FUSION OF CONTEXT AND ARTICULATORY SECONDARY TASKS

As both some of articulatory characteristics and context information help the training, the final experiment was made with 5 secondary tasks: context (left and right), place and manner of articulation and vowel/consonant characteristics. The resulting test accuracy is 72.8%, which is half percent better than the baseline.

Task	PhnAcc %	$\triangle$ diff
Baseline	72.3	0
Context	**72.6**	**+0.3**
Articulatory	71.5	-0.8
Place	**72.5**	**+0.2**
Manner	**72.5**	**+0.2**
Vowel/Consonant	**72.4**	**+0.1**
Voice	72.3	0
Additional	72.1	-0.2
Place+Manner+Vowel/Cons	**72.6**	**+0.3**
Context+Place+Manner+Vowel/Cons	**72.8**	**+0.5**

Table 2: Comparison of the baseline and different multi-task settings

4 CONCLUSION

The experiments have shown that multi-task neural networks can be more effective than single-task neural networks if the secondary tasks are chosen wisely. Both context and articulatory tasks have been found helpful, and their combination is the most effective.

REFERENCES

[1] R. Caruana. Multitask learning: A knowledge-based source of inductive bias. Machine Learning, vol. 28, pp. 41-75, 1997

[2] M.L. Seltzer and J. Droppo. Multi-task learning in Deep Neural Networks for improved phoneme recognition. In Proceedings of ICASSP 2013 Vancouver, Canada