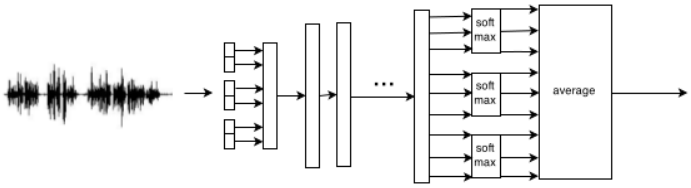
Excel@FIT 2015
http://excel.fit.vutbr.cz

Multiframe neural networks in speech recognition

Kateřina Žmolíková*



Abstract
This paper presents a modification of neural network structure in speech recognition system which leads to improving the accuracy of the system. Deep neural networks are widely used as a part of acoustic model which aims to predict the score of acoustic units given the speech signal. The input of deep neural network is a sequence of speech frames. Typically the network tries to classify the central one of these frames while using the context frames as an additional information.
In the multiframe model the output of the network is extended to predict classes of multiple frames. This modification leads to obtaining multiple predictions for one frame. Combining these predictions results in better accuracy of the network.
The approach was tested on Wall Street Journal dataset. Experimenting with different sizes of the context on the input and output of the network lead to 7% and 12% relative improvement on two testsets.

Keywords: speech recognition — deep neural networks — acoustic modelling

Supplementary Material: *N/A*

*xzmoli02@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Speech recognition is an active research area which aims to automatically transcribe speech into text. In the last decade deep neural networks have become one of the key components of speech recognition systems showing significantly better results than previously used methods. The goal of this project is to improve accuracy of the speech recognition system by simple modification of the neural network.

The neural network is used for acoustic modelling which is core part in the speech recognition system. The task of the neural network is to classify short segments of a speech signal into classes representing phones. In order to classify the segment correctly it is useful to have information about its context. Therefore the input of the network doesn't contain only the segment which should be classified but also a small number of previous and following segments. In paper *Autoregressive product of multi-frame predicition can improve the accuracy of hybrid models* [1] Navdeep Jaitly presented an idea to make use of the additional segments on the input and let the network classify them as well as the central segment. During the training the network has more information about the context and during the decoding we obtain more predictions for one segment of speech. These predictions can be combined which leads to improvement of the accuracy.

This approach is also slightly similar to split-context method used by Petr Schwarz [2]. In this method temporal context is split into two parts and each of them is handled by separate classifier. Predicitions of both classifiers are then merged together.

In this work we implemented the multiframe ap-

proach to Kaldi toolkit [3] and tested its performance on Wall Street Journal dataset [4]. We experimented with several settings, analyzed obtained results and the behaviour of the network during the training.

In this paper we will briefly introduce the speech recognition problem and how deep neural networks are used in it. We will precise the approach of training and decoding with multiframe neural networks and in Section 4 we will present and analyze the obtained results.

2. Neural networks in speech recognition

This section will summarize necessary theory about speech recognition systems and deep neural networks.

2.1 Speech recognition system

The input of speech recognition system is a raw speech signal. At first the signal needs to be transformed into representation that is more suitable for the system. It is split into short 20 milisecond segments which are then converted into frequency domain. This way we extract typically 13 dimensional feature vector for each segment. The task of the system is then to provide the most probable sentence corresponding to this sequence of feature vectors.

To compute the probability of a sentence the system needs to have knowledge about how sentences are formed in the language and how individual words can be decomposed into phones. This is done by language and pronunciation model. The core of the system is an acoustic model which provides the probability that input feature vectors correspond to individual phones.

Acoustic models in almost all speech recognition systems are based on Hidden Markov Models [5]. Each phone is represented by Hidden Markov Model which is composed of three states. The role of deep neural network is to compute probability that feature vector coresponds to a state of HMM.

2.2 Deep neural networks

Neural network is model representing a mapping of its inputs to its outputs. It is composed of mutually connected neurons that compute weighted sum of their inputs followed by simple nonlinearity. By modifying the weights of this sum the function computed by the network can be changed. With sufficient training data we can train the network to compute the function we need. We do this by computing the error function on the output of the network and minimizing this function with respect to the weights. The optimization is most commonly done by stochastic gradient descent.

In speech recognition system the neural network is used to assign the feature vectors to the states of Hidden Markov Model. The input of the network consists of this feature vector together with a short context. The outputs are the probabilities representing how well each state of each HMM fits the feature vector. To get proper probability distribution the final layer typically implements softmax function.

3. Multiframe approach

Although the input of the neural network contains not only one frame of speech but also its context, on the output only the central frame is classified. If we force the network to classify also the context frames we obtain more predictions for each frame which we can combine together.

3.1 Training with multiframe targets

For the training the output of the neural network is extended by adding units predicting probabilities of HMM states for context feature vectors. Figure 1 shows how the architecture of the network changes. In the baseline case the output layer of the network contains number of neurons corresponding to number of HMM states. In the multiframe model the output layer is K-times larger, where K is the size of the output context.

Network can be trained using classic backpropagation algorithm, only the targets of the training need to be properly extended.

3.2 Decoding with multiframe targets

During the test phase we obtain n prediction for each frame. Each of this predicition comes from different contexts on the input. We can combine these predicitions by applying geometric average.

4. Experiments

This section will introduce experiments testing performance of the multiframe method and their results. We will also specify used architecture of the deep neural network, method of training and the data which were used. All experiments were done using Kaldi speech recognition toolkit.

4.1 Dataset specification

For experiments we used subset of Wall Street Journal dataset [4] which includes 14 hours of training data and two testsets — *test_dev93* and *test_eval92*. This dataset contains read paragraphs from Wall Street Journal. It is therefore english corpus with large vocabulary. We

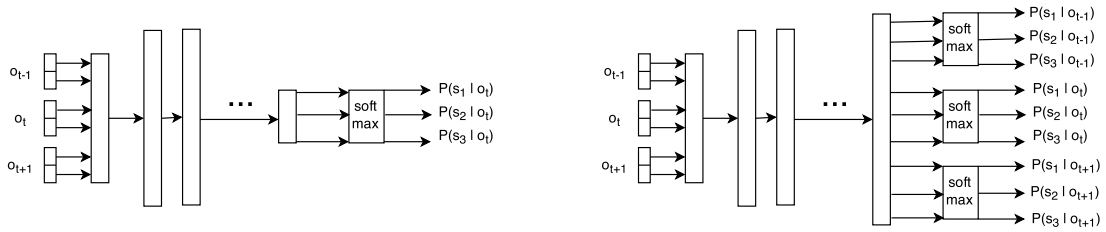


Figure 1. Difference between baseline and multiframe structure of deep neural network. The input of both networks contains 3 frames o_{t-1}, o_t and o_{t+1}. The baseline network predicts probabilities of HMM states for the central frame o_t. In the multiframe model the output of the network includes probabilities of context frames o_{t-1}, o_{t+1}.

used trigram language model provided with the dataset. Baseline system was build using standard Kaldi recipe.

4.2 System parameters

Detailed description of the speech recognition system is out of scope of this paper. This section will therefore describe basic parameters of the system accompanied by references to relevant sources.

We used deep neural network with six hidden layers each containing 2048 neurons with sigmoid activation function. As input features we used MFCC features transformed with LDA [6], MLLT [7] and fMLLR [8] transforms. Mean and variance normalization was applied on the input of the network. We tested various sizes of input and output context. Output layer of baseline network contained 3514 neurons.

Network was initialized by generative pretraining using Restricted Boltzmann Machines [9] [10]. Discriminative training was done by standard backpropagation algorithm with learning rate 0.008 gradually lowered during the training.

4.3 Results

Experiments aimed to find the best size of contexts on the input and output of the network. Input sizes ranged from 5 frames to 15 frames, output sizes from 7 to 17 frames. Table 1 shows best results for every input context. The metric used for evaluating the performance of the system is Word error rate (WER) which incorporates three kinds of errors - substitutions, insertions and deletions. The final score is sum of number of these errors normalized by number of words in the reference.

In most of the cases the best result occured when output context was 4 frames larger than the input context. This may be explained by properties of the input features. The features were preprocessed by doing LDA+MLLT transforms which already compute one feature using 3 left and 3 right context frames. Therefore when we use features transformed this way on the input, the network has information about slightly larger context than the number of frames on the input.

Table 1. Results

IN	OUT	WER dev93 [%]	WER eval92 [%]
15	1	8.39	4.86
3	7	7.98	4.52
5	9	7.77	4.24
7	11	7.87	4.45
9	13	7.98	4.45
11	15	7.97	4.54
13	17	8.29	4.68
15	15	8.20	4.73

Overall best result occured using combination of input context 5 and output context 9. On the testset test_dev93 the result was 7.77% and test_eval92 4.24%. Best result without using multiframe method was with input context 15 - 8.39%, 4.86%. Relative improvement on both testsets was therefore respectively 7% and 12%.

Best size of input context for multiframe case was much smaller than best size for baseline case. However in the multiframe case the network combines probabilities predicted from different contexts. The network therefore effectively uses much larger context for prediciting one frame.

5. Conclusions

In this paper we showed how extending the output of the neural network to do multiple predictions at once can improve its accuracy. We used this technique for speech recognition task where deep neural networks play important role.

We tested the method on Wall Street Journal database where it achieved 7% and 12% relative improvement on two available testsets.

6. Acknowledgments

I would like to thank Karel Vesely and Lukas Burget for lots of advice and help.

References

[1] Navdeep Jaitly, Vincent Vanhoucke, and Geoffrey Hinton. Autoregressive product of multi-frame predictions can improve the accuracy of hybrid models. In *Proceedings of Interspeech 2014*, 2014.

[2] Petr Schwarz, Pavel Matějka, and Jan Černocký. Towards lower error rates in phoneme recognition. *Lecture Notes in Computer Science*, 2004(3206):465–472, 2004.

[3] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December 2011. IEEE Catalog No.: CFP11SRW-USB.

[4] Douglas B. Paul and Janet M. Baker. The design for the wall street journal-based csr corpus. In *DARPA Speech and Language Workshop*. Morgan Kaufmann Publishers, 1992.

[5] Mark Gales and Steve Young. The application of hidden markov models in speech recognition. *Found. Trends Signal Process.*, 1(3):195–304, January 2007.

[6] R. Haeb-Umbach and H. Ney. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, pages 13–16, Washington, DC, USA, 1992. IEEE Computer Society.

[7] Mark J. F. Gales. Semi-tied covariance matrices for hidden markov models. *IEEE Transactions on Speech and Audio Processing*, 7(3):272–281, 1999.

[8] Mark JF Gales. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer speech & language*, 12(2):75–98, 1998.

[9] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, July 2006.

[10] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *J. Mach. Learn. Res.*, 11:625–660, March 2010.