Growing bottleneck features for tandem ASR

Joe Frankel, Dong Wang, Simon King

Centre for Speech Technology Research, University of Edinburgh, UK
joe@cstr.ed.ac.uk; dwang2@inf.ed.ac.uk; Simon.King@ed.ac.uk

Abstract

We present a method for training bottleneck MLPs for use in tandem ASR. Experiments on meetings data show that this approach leads to improved performance compared with training MLPs from a random initialization.

Index Terms: tandem ASR, bottleneck MLP

1. Introduction

Tandem automatic speech recognition (ASR) [1] extends acoustic parameters such as Mel frequency cepstral coefficients (MFCCs) to include features derived from phone-classification multi-layer perceptrons (MLPs). Rather than interpreting the outputs as phone posteriors, they are subject to a logarithm transformation and dimensionality reduction, then used as features in a standard hidden Markov model (HMM) ASR system.

A common configuration is to use 9 frames of perceptual linear prediction (PLP) cepstra as input to a 3-layer MLP. Principal components analysis (PCA) is used for dimensionality reduction, and whilst there is no clear optimal final dimension, work at ICSI has found that retaining around 95% of the variance, which yields a feature vector of dimension 25 works well. An alternative approach is to use a bottleneck MLP [2] (a.k.a. autoencoder network), and make the dimensionality reduction integral to the network. Multiple hidden layers are used, one of which is low-dimension and is known as the 'bottleneck layer'.

Hinton and Salakhutdinov [3] suggest that autoencoder network training is particularly sensitive to initial conditions, and that when initial weights close to a good solution, gradient descent works well. In order to improve the initial estimates, they propose "pretraining" in which the learned activations from one layer are used as the input to the following layer. We apply similar ideas in the context of training of MLPs for tandem ASR.

2. Tandem MLPs

We present the results of experiments using 3 different MLPs. In all cases, the input layers have 351 input units (9 frames of 39 PLPs) and 46 output units (one for each phone class) with a softmax activation function. The baseline 3-layer MLP has size 351, 5000, 46. We then trained a 5-layer MLP of size 351, 5000, 25, 5000, 46 units from a random initialization. The activations of the bottleneck layer of 25 units are subjected to a logarithm transform, and used as the tandem features.

The third MLP also had 5 layers, though rather than being trained from scratch the MLP was "grown". A 4-layer MLP of size 351, 5000, 25, 46 was first trained using the weights and biases of the input to first hidden layer of the 3-layer MLP above, with the remainder of weights randomly initialized. A 5-layer MLP of size 351, 5000, 25, 5000, 46 was then trained with the weights and biases connecting the input and first two hidden layers taken from the 4-layer MLP, and the remainder of weights randomly initialized. The cross validation accuracy for the grown 5-layer MLP was higher than that of either the randomly-initialized 5-layer and the 3-layer MLP.

3. Experiments and discussion

Experiments use data from the meetings domain, with speech recorded on headset microphones. HMMs and MLPs were trained on a little over 100 hours of data, and results are presented on test data from the NIST Spring 2004 RT04s evaluation. A trigram language model was used during decoding.

The acoustic waveform was parameterized as 12 Mel-frequency cepstral coefficients (MFCCs) with energy, and 1st and 2nd order derivatives. The final feature dimension was 64 once the 25 MLP features were added. The 5-layer MLPs can be used to provide tandem features either from the bottleneck layer, or from the output layer (in which case they are subjected to dimensionality reduction via PCA.

Features	WER
MFCC baseline	40.8%
MFCC + 3-layer tandem	38.8%
MFCC + 5-layer tandem (output)	41.6%
MFCC + 5-layer tandem (bottleneck)	40.9%
MFCC + 5-layer tandem (bottleneck), grown	37.5%

Table 1: ASR results for each of the 3 and 5-layer MLPs

The results presented in Table 1 show that using tandem features from the 3-layer MLP leads to improved performance compared to the MFCC baseline. The 5-layer MLP which was trained from a random initialization gives an increase or no change in WER compared with the MFCC baseline when the tandem features are taken from the output and bottleneck layers respectively. However, the bottleneck features from the tandem MLP which has been grown leads to a reduction in WER compared with the 3-layer MLP result. A paired t-test shows that this result is significant with $p < 0.001$.

We conclude that the integral dimensionality reduction of bottleneck MLPs leads to improved performance over 3-layer tandem MLPs and PCA, and that bottleneck MLPs should be grown rather than trained from a random initialization.

4. References

[1] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature stream extraction for conventional hmm systems," in *Proc ICASSP*, vol. III, Istanbul, Turkey, 2000, pp. 1635–1638.

[2] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for lvcsr of meetings," in *Proc. ICASSP*, Honolulu, April 2007.

[3] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 28, pp. 504 – 507, July 2006.

[1] This work is supported by EPSRC Advanced Research Fellow and Marie Curie Early Stage Training Site EdSST (MEST-CT-2005-020568).